



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

NIDDK-CR Resources for Research

Data Science and Meet the Expert Webinar Series



June 26, 2025



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

NIDDK Central Repository Overview

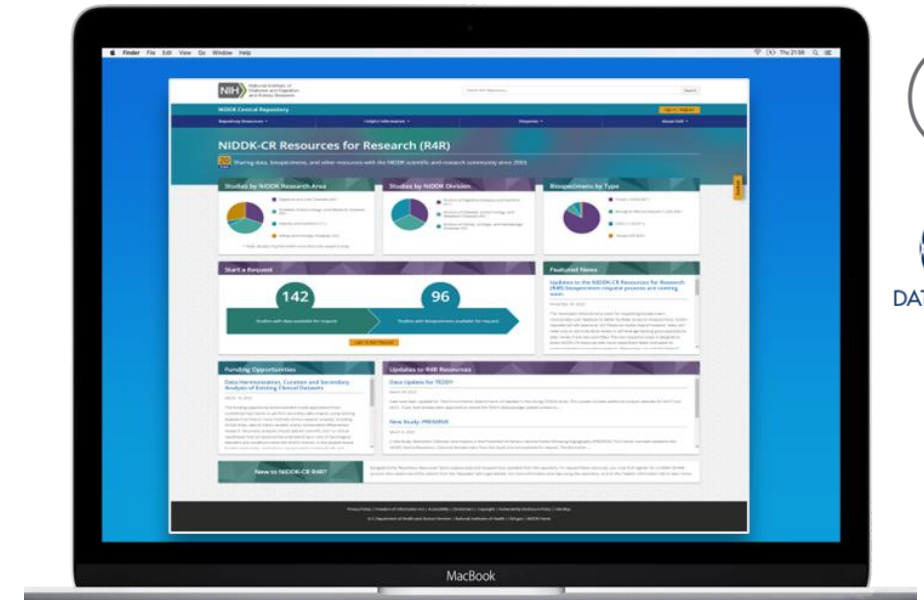
Mission


Established in 2003 to **facilitate sharing of data, biospecimens, and other resources** generated from studies supported by NIDDK and within NIDDK's mission by making these **resources available for request to the broader scientific and research community**.


- Supports receipt and distribution of data and biospecimens in a manner that is ethical, equitable, and efficient
- Enables investigators not involved with the original work to test new hypotheses without the need to collect new data or biospecimens
- Promotes FAIR (Findable, Accessible, Interoperable, and Reusable) and TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) principles





Recorded past tutorials, webinars, and other educational resources can be found on the NIDDK-CR website



Imaging Data Files

15.8 M

Clinical Datasets

>8,400
from 189 clinical studies

Biospecimens

>16 M

Registered Users

6,889

Weekly Users

>5,000

Public Releases

>875

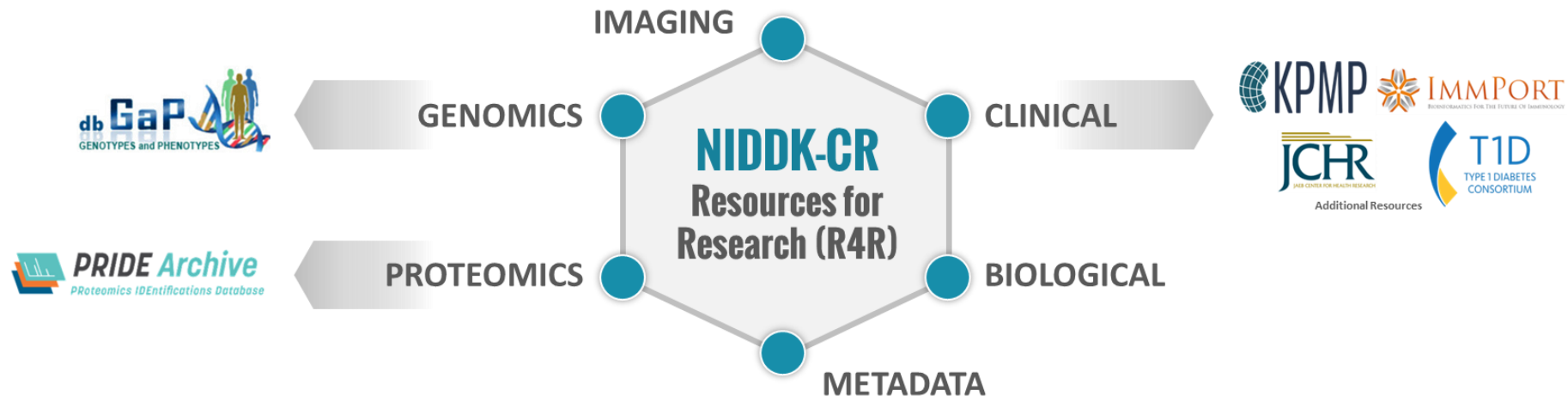


National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

NIDDK Data Sharing Ecosystem

The NIDDK-CR is a part of the broader NIH-funded biomedical data ecosystem and plays a key role in NIH's FAIRness and TRUSTworthiness goals. The NIDDK-CR houses a broad range of data types for secondary research, provides access to biospecimens, and direct links to other repositories with additional resources such as genomics data.



FAIRsharing.org
standards, databases, policies

DataCite
FIND, ACCESS, AND REUSE DATA

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES



Google Dataset Search

Schema.org

NIH U.S. National Library of Medicine
ClinicalTrials.gov

Vivli
CENTER FOR GLOBAL CLINICAL RESEARCH DATA

PLANNING
PHASE

figshare

NIH
HEAL
INITIATIVE



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Future Functionality: Analytics Workbench

Streamlining end-to-end data science lifecycle
and discovery of data-driven biomedical insights.

Innovation and ease of use

A cloud-based analytics environment
where researchers and data scientists
can access a suite of integrated analytics
tools and cloud computing resources to
participate in data challenges and AI
innovation.

Expected Benefits of Analytics Workbench:

Promote
Collaboration

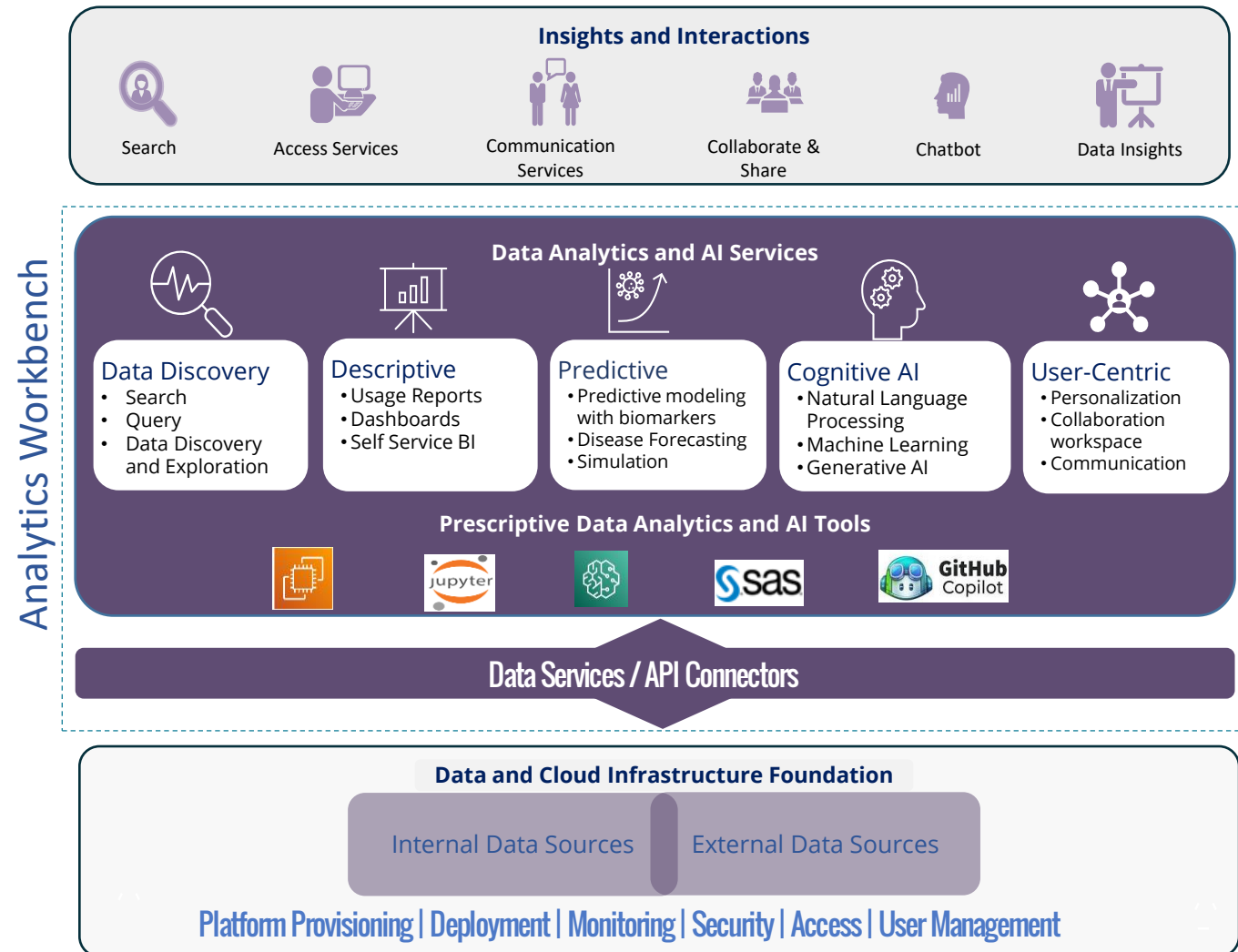
Support AI
Innovation

Minimize Data
Movement

Improve User
Experience

Discover
Data Insights

Advance NIDDK
Research Mission





National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

NIDDK-CR Data Science Centric Challenge Series

Goals of NIDDK-CR Data-science centric challenge series

- Develop tools, approaches, models and/or methods to increase data interoperability and usability for artificial intelligence (AI) and machine learning (ML) applications
- Augment and enhance existing data for future secondary research, including data-driven discovery by AI/ML researchers
- Discover innovative approaches to enhance the utility of datasets for AI/ML applications



Visit our website for more information on our data-centric movement and to learn more about our past data-challenges



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Secondary Data Science and Meet the Expert Webinar Series

About the Series

- Aims to accelerate data science and AI-driven biomedical research by fostering collaboration between biomedical researchers and experts in the field
- Monthly webinar held on the **last Thursday of each month**

Upcoming Webinars

- Today – Different privacy preserving techniques and implications for researchers
- July 31 – Challenges, opportunities, and considerations for secondary researchers using electronic health records and real-world data sources
- August 28 – Impact and innovations realized



Learn more about the webinar series, register for future webinars, and access past webinars materials and recordings



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Meet the Experts



Anya Dabic is a Health Data Scientist at Booz Allen Hamilton that specializes in scientific data management and stewardship. She co-authored two public reports for the NIH NICHD on feasibility of leveraging privacy preserving record linkage (PPRL) for linking pediatric datasets across HHS. She has also supported various data management and sharing programs at the National Institutes of Health to implement the FAIR (findable, accessible, interoperable, reusable) and TRUST (Transparency, Responsibility, User focus, Sustainability and Technology) principles for digital repositories, including the NIDDK Central Repository.



Shruti Gautam is a Bioinformatician and Cloud Platform developer at Booz Allen Hamilton. She supports genomic surveillance and outbreak detection efforts across the CDC account as part of the Advanced Molecular Detection program. As part of this work, she has supported thought-leadership and development efforts around safeguarding privacy while supporting data sharing and data linkage within platform. She has also supported the VA and NIH NIAID agencies as a clinical trial bioinformatician working on epigenetic and vaccine studies.



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Different Privacy- Preserving Methodologies and Implications for Researchers

NIDDK-CR Data Science Meet
the Experts Webinar Series

June 26, 2025

Presented by: Booz Allen Hamilton





Agenda

- Importance of participant confidentiality and data privacy
- Privacy Preserving Record Linkage (PPRL)
 - What is record linkage and PPRL?
 - Benefits and Use Cases of PPRL
 - Essential elements for PPRL
 - Requirements for implementing PPRL
 - PPRL resources
- Differential Privacy within the Privacy Enhancing Technology toolkit
 - Benefits of the methodology
 - Differential privacy in real world use cases
 - Overview of the underlying math at a high level
 - Synthetic dataset generation using differential privacy



Participant Confidentiality and Data Privacy

- **Participant Confidentiality** refers to the **duty of researchers to protect participants' identity and personal information** from unauthorized access or disclosure.
- **Examples:**
 - Using secure systems to store or transmit data
 - Removing names or identifiers before sharing data (de-identification)
 - Certificates of Confidentiality prevent researchers from being forced to disclose data
- **Why Participant Confidentiality Matters?** Protects participants from harm (e.g., discrimination, stigma), Respects autonomy, and Fosters trust in research.
- **Data Privacy** refers to the **right of individuals** (participants) to control **how their personal information is collected, used, and shared**.
- **Examples:**
 - Obtaining informed consent before collecting personal data
 - Participants have right to withdraw anytime
 - Complying with data protection laws like HIPAA Privacy Rule
 - Sharing data with external researchers requires formal agreements
- **Why Data Privacy Matters?** Ensures individuals have control over their data and reduces risk of breaches or misuse.



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Best Practices for Protecting Participant Privacy When Sharing Scientific Data

- **NIH has established a set of principles and best practices for protecting the privacy of research participants** when sharing data under the NIH Policy for Data Management and Sharing (DMS)
- **Foundational Principle**
Respecting participant privacy is essential and must align with informed consent, legal, and policy obligations (e.g., Common Rule, HIPAA, NIH policies)
- **Best Practices**
 - ✓ Apply appropriate de-identification while preserving scientific value.
 - ✓ Establish data sharing/use agreements (oversight, responsibilities, restrictions).
 - ✓ Understand and communicate legal protections, including Certificates of Confidentiality for long-term privacy.

For additional information, see: <https://sharing.nih.gov/data-management-and-sharing-policy/protecting-participant-privacy-when-sharing-scientific-data/principles-and-best-practices-for-protecting-participant-privacy>



Data De-Identification Overview

- Per Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule, two de-identification methods: 1) Expert Determination, and 2) Safe Harbor Method
- Data de-identification includes techniques such as redacting, masking, and recoding direct identifiers to anonymize participants from a research study
 - Redacting may include removing or deleting an entire variable from a dataset (e.g., participant phone number), or a participant's data in a study (e.g., due to lack of consent for data sharing)
 - Masking may include replacing data with an anonymized indicator or symbol (e.g., "James Smith admitted to hospital 2 hours after treatment" → "[Name] / XXXXX admitted to hospital 2 hours after treatment")
 - Recoding may include applying a random code to anonymize data (e.g., changing site names from "Montgomery Hospital" to "1", and "York Hospital" to "2")
- It is important to note that there are several terms and techniques for de-identification; the main goal to protect a participant's identity

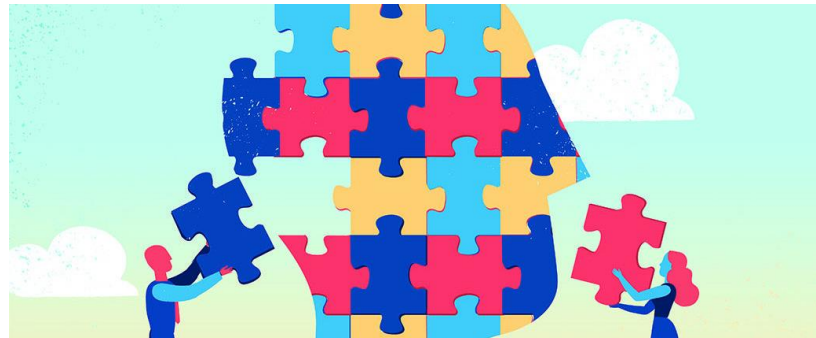


National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Challenges Maintaining Participant Privacy

- Challenge 1: Linking datasets without exposing personal identifiers
- Challenge 2: Cross-institutional collaboration without breaching confidentiality
- Challenge 3: Re-identification risk when linking multiple datasets
- Challenge 4: Protecting individuals in shared or published datasets
- Challenge 5: Balancing utility and privacy
- Challenge 6: Defending against attackers with auxiliary information



99.8% of Americans can be correctly re-identified in any dataset using 15 demographic attributes ([Rocher et al., 2019](#))



Privacy Enhancing Technologies

- **Privacy Enhancing Technologies (PETs)** are tools, methods, and frameworks that protect sensitive information during data collection, storage, processing, linkage, and sharing.
- **Goal:** Enable data use for research and innovation without exposing individuals to privacy risks.
- **Examples:**

PET (* Today's Webinar)	Key Approach	How It Protects
PPRL *	Encrypts or transforms identifiers for linkage	Links data without revealing or sharing PII
Differential Privacy *	Adds statistical noise to data or outputs	Prevents detection of any individual's data
Homomorphic Encryption	Allows computation on encrypted data	Keeps data secure even during analysis
Secure Multiparty Computation	Distributes computation across parties without sharing data	No one party sees the whole dataset
Federated Learning	Trains models locally, shares only updates	Data stays at its source



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Privacy Enhancing Technologies

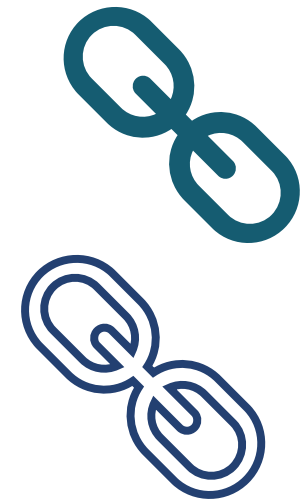
Privacy preserving record linkage (PPRL)





What is record linkage?

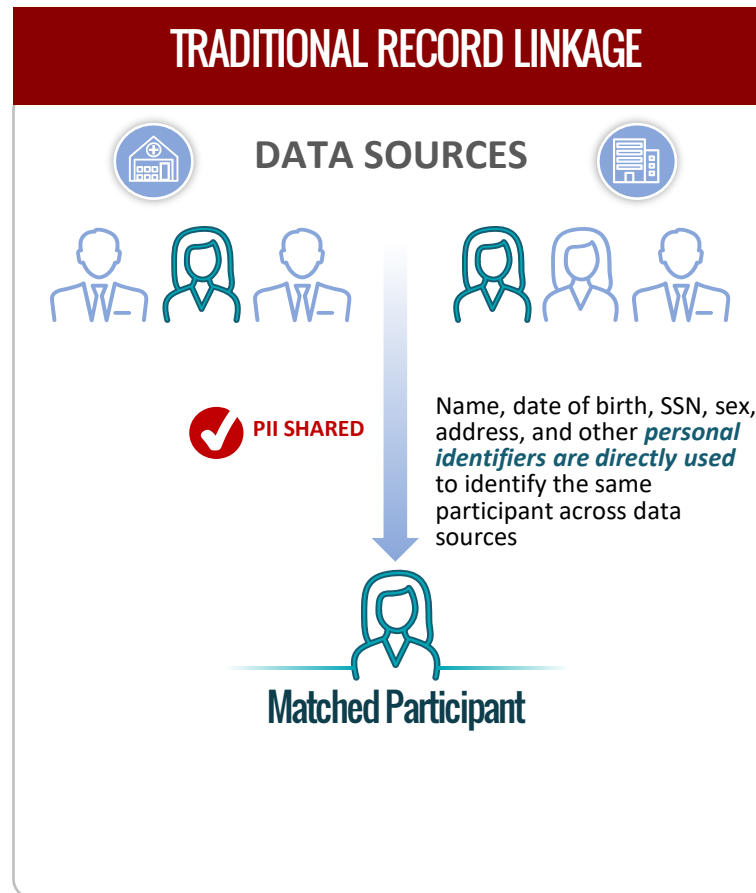
- Record linkage is combining (or bringing together) two or more records that correspond to the *same* individual.
 - Term first introduced in 1946 by Halbert Dunn of the U.S. National Bureau of Statistics: “Each person in the world creates a Book of Life. This Book starts with birth and ends with death. Record linkage is the name of the process of assembling the pages of this Book into a volume.”
- Many U.S. agencies including the Census Bureau, CDC, CMS, Agency for Healthcare Research and Quality (AHRQ), and Administration for Children and Families (ACF) have been using record linkage for decades.



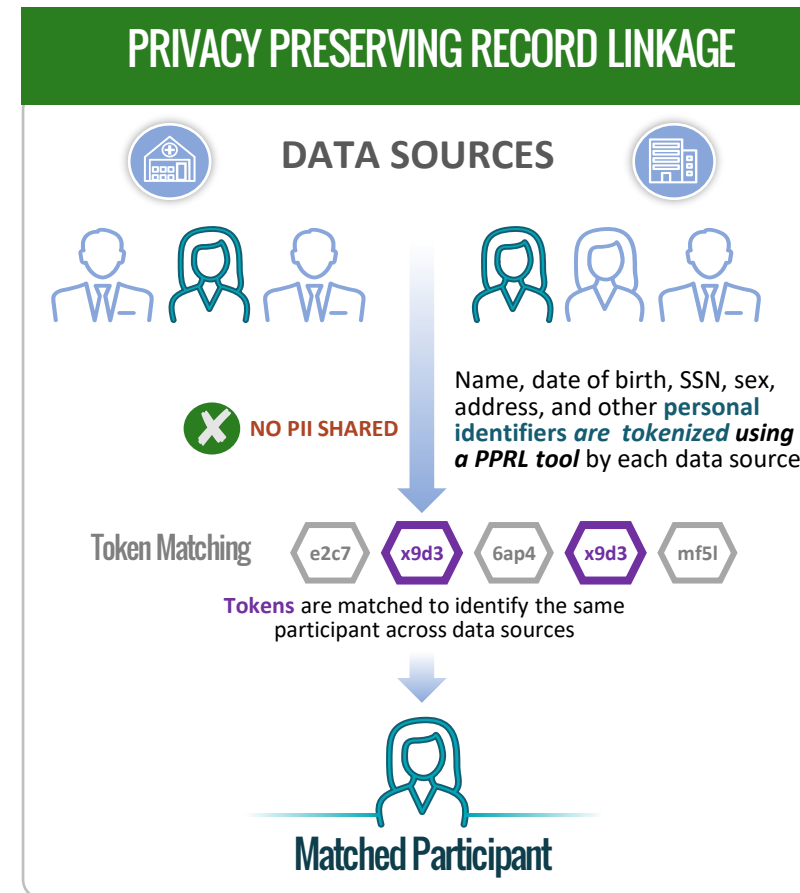


What is PPRL?

Traditional linkage uses direct PII such as name, date of birth, social security number, address, etc. to match records of an individual

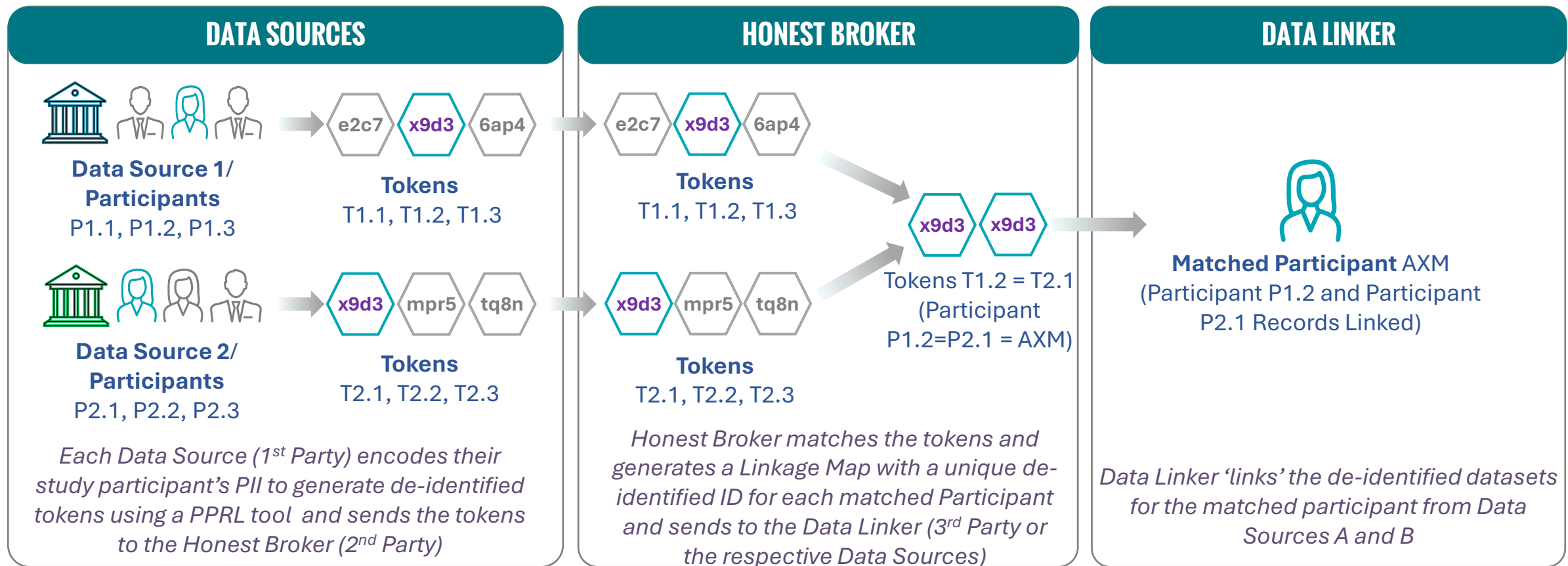


Privacy preserving record linkage (PPRL) encodes the PII to create one-way hashed codes (tokens), which are then matched to identify records of an individual.



PPRL Methodology

- Tokens generated on the same set of PII using the same hashing algorithm will be identical across data sources allowing them to be matched
- A three-party PPRL model where an honest broker performs token matching provides an additional layer of security between PII and the actual data





PPRL Benefits and Use Cases

DATA SOURCES



Electronic
Health Records



Claims Data



Mortality Data



-Omics Data



Pharmacy Data



Registry Data



Images



Wearable Data



Laboratory
Data



Patient
Reported
Outcomes

BENEFITS

Broader sharing of datasets of
a particular individual without
sharing their PII



Address or track a patient's or
participant's journey
(longitudinally) through health
care and research



Enrich data by linking
multimodal (different types of)
data collected



Avoid costly duplication of
data, such as genomic data

USE CASES

Link registry data across state/local jurisdictions
EX: NCI SEER

National Patient Registries

Link - omics and EHR data to get more comprehensive
dataset on an individual

Precision Medicine

Link outcome and EHR data for product surveillance
EX: FDA Sentinel

Medical Device & Drug Monitoring

Link EHR and wearable data across a research cohort
EX: NIH AOU Program

Clinical Research

Link case and vaccination data for infectious diseases
EX: CDC, N3C

Disease Surveillance



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Essential elements for PPRL

- Five essential elements must be addressed when considering PPRL – participants, data, PPRL tool, data governance, and data platform
- These components intersect at various points to address real world use cases effectively and efficiently





Requirements for Implementing PPRL



Participants

- Do the participants you are engaging for your study understand the benefits of linking their data with other data?
- Have they consented to linking? If not, can you get the consent, a waiver of consent, or approval from an institutional review board (IRB) or an equivalent human subjects privacy board to link their data?



Data

- What is the scope of linkage (i.e., what types of data do you need from other sources for the linkage)?
- Who has that data and can you gain access to the data?
- What is the quality of the data? Is the PII and data standardized for linkage?



PPRL Tool

- Which PPRL tool can you use with the PII you have in your data?
- Is the tool freely available (open source) or proprietary? If proprietary, what are the licensing costs?
- Can the tool scale up to accommodate increasing volumes of data?



Governance

- What governance (policies, terms, and conditions for use) do you need to comply with for accessing and linking the data and sharing and/or using the linked data?
- What data disclosure methods will you use to mitigate re-identification risk of the linked data?



Data Platform

- Do you have a data platform to store the data and provision the linked data to end users to access and use?
- Does the platform have the appropriate security controls in compliance with federal data and system policies?
- Is the platform scalable and adaptable to growing needs of users, including data types and volume ?



Open Source PPRL Tools

Tool Name	Description
Carduus (from DataBricks)	The Open Privacy Preserving Record Linkage (OPPRL) protocol is a free and open standard for replacing personally identifiable information (PII) with encrypted tokens in order to preserve the privacy of data subjects.
ANONLINK	Anonlink is an open source (Apache 2.0) suite of technologies that allows organizations to carry out PPRL. It uses cryptographic hashes, blocking, and Bloom filters. Anonlink is written in C++ and provides an interface to Python. Anonlink is modular, consisting of different libraries for generating hashes, calculating similarity scores, and offering an entity service from which the clients can request mappings.
GRLC /PPRL R Package	German Record Linkage Center (GRLC) PPRL R Package is an open source (available for free on CRAN, GPL-3) toolbox developed by GRLC for deterministic, probabilistic, and privacy-preserving record linkage techniques using R. It combines Merge ToolBox with current privacy-preserving techniques
PRIMAT	PRIMAT (Private Matching Toolbox) is an open source (Apache 2.0) toolbox developed by the Database Group of the University of Leipzig, Germany, for the definition and execution of PPRL workflows. It offers several components for data owners and the central linkage unit that provides state-of-the-art PPRL methods, including Bloom-filter-based encoding and locality-sensitive hashing-based blocking, metric space filtering, post-processing, and more.



PPRL resources

1. PPRL for Pediatric COVID-19 Studies, Final Report, NICHD 2022
 - Provides a Record Linkage Implementation Checklist to guides users who are interested in linking data through governance and technical considerations for designing and implementing a record linkage strategy.
2. Patient-Centered Outcomes Research Trust Fund Pediatric Record Linkage Governance Assessment, Final Report, NICHD 2023
 - Provides a data governance information framework to assess two or more datasets can be linked and if so, what governance applies to linking and the linked dataset
3. White House Office of Science and Technology Policy Report on the National Strategy to Advance Privacy-preserving Data Sharing and Analytics



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Privacy Enhancing Technologies

Differential Privacy





Privacy Enhancing Technologies: Differential Privacy

Building trust in a way that is demonstrable to users...

Differential Privacy is a rigorous mathematical definition of privacy, formally guaranteeing that individual-level information about participants in a database is not identifiable.

Example: An algorithm that analyzes a dataset and computes statistics about it – the algorithm is differentially private if by looking at the output, one cannot tell whether any individual's data was included in the original dataset or not.

- **Accomplished by injecting *noise*** (random statistical variations) into the data.
- The programming code can be made public, meaning differential privacy is **transparent to users**. The only information generally not published is the exact value of *noise* added to given data points.
- Enables organizations to **aggregate and analyze data** to learn trends, but **in a way that protects the privacy** of individuals who contributed their data.

Taken one step further → differential privacy can be used to generate **differentially privatized synthetic datasets**



Benefits of Differential Privacy

"Differential privacy allows the [US Census] Bureau to protect against [increasingly sophisticated](#) reconstruction and reidentification attacks that threaten the confidentiality of individual census responses."

– Prof Cynthia Dwork considered one of the founders of Differential Privacy

Benefits:

- **Resiliency** – Unlike prior methods of table suppression or record swapping, differentially private data can be published, analyzed, and linked to other data without any increased risk of disclosure.
 - **Transparency** – The programming code and decisions for differential privacy can be made available to the public and/or users.
 - **Protection for All Data** – Assumes all information is identifying information, eliminating the challenge of flagging only identifying elements.
 - **Trust:** Informing users that differential privacy is in use alleviates confidentiality concerns and encourages participation (in use by Apple, Google, Census Bureau).
- Record swapping
 - Records with similar characteristics but different geographic identifiers (or other identifying attributes) are matched. The values of non-key attributes (e.g., age, race, household characteristics) are then swapped between these matched records.
 - Table suppression
 - Table suppression is a data privacy technique where information, specifically from individual cells or rows/columns in a table, is removed to protect confidential information and prevent the identification of individuals.



Differential Privacy Out in the Real World

	Local Differential Privacy	Central Differential Privacy
Integration of statistical noise	On the user's device	After raw data has been collected
Trust assumptions	No trust required – the data collector does not see real data	Users must trust collectors
Utility	Lower Noise is at the individual level	Higher Noise integrated at aggregated data table level
Deployed?	Apple, Google	US Census

Traditional use of Differential Privacy at Census

The Census Act requires the U.S. Census Bureau to protect respondent confidentiality at every stage of the data lifecycle. Information about specific individuals, households, or businesses *cannot* be revealed, in published statistics or otherwise.

- **Disclosure avoidance** is the process used by the Census Bureau to protect the confidentiality of respondents' personal information.
- The Census Bureau **balances** the **need to collect and report** the data with the statutory **obligation to protect** it.
- The Census has deployed different methods to fulfill this requirement, most recently with **differential privacy**.

Research shows **respondent concern for privacy is among the top reasons for unwillingness** to participate in censuses.

Making demonstrating privacy a mission imperative for the Census Bureau.

1930
Stopped publishing
some small-area
data

1970
Whole-table
suppression

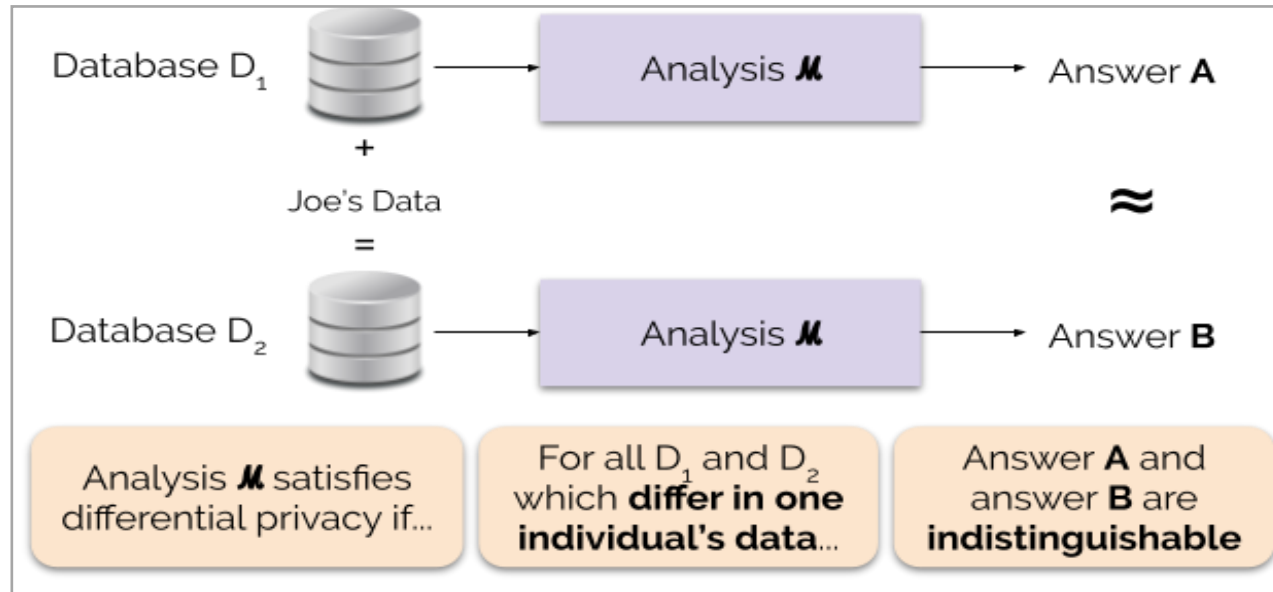
1990
Data
swapping

2020
Differential
privacy

The Census Bureau's recently adopted **Disclosure Avoidance System** employs a two-step process within a framework known as the **TopDown Algorithm** to protect respondent information:

- 1. Differential Privacy Algorithms:** These inject noise into the data.
 - The level of noise introduced is guided by a "privacy loss budget" that defines the upper bound of privacy loss that can occur. (In this case $\epsilon = 2.47$)
- 2. Post-Processing:** This step imposes certain consistencies onto the data (ex. Ensuring that the population totals for counties within a state sum to the state's total population) to make it more usable.

How does this work?



*Graphic from NIST; Differential Privacy for Privacy Preserving Data Analysis

Differential Privacy guarantees to users that the output of a differentially private analysis will be roughly the same, whether they contribute their individual data or not.

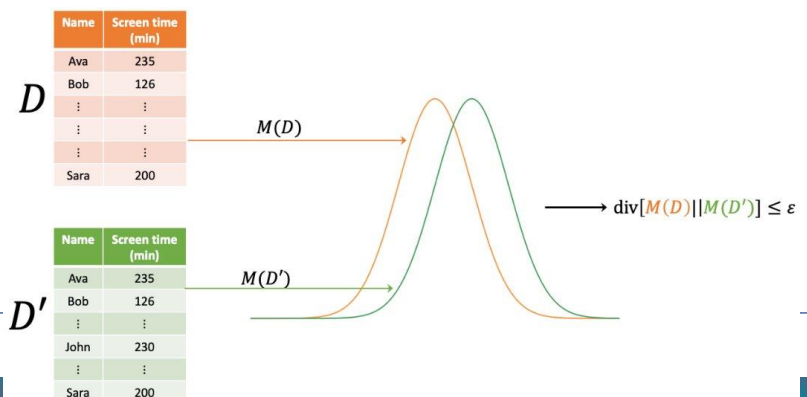
A **mechanism**, denoted as \mathcal{M} above, is a target function (ex: mean) + noise

The goal is to force outputs to be “noisy” so that no individual has much influence.

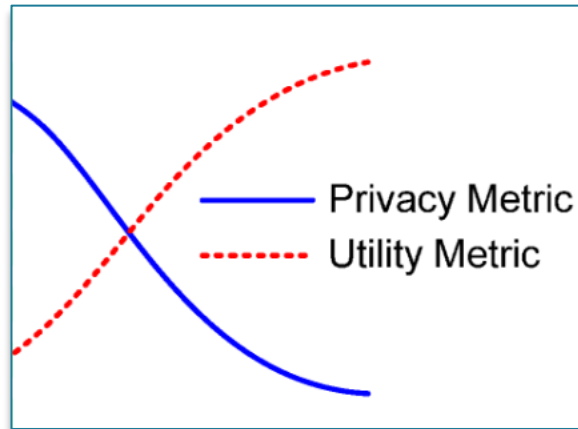
There is a trade-off between privacy and utility

Example: Calculate the hazard of patients experiencing kidney failure or death in individuals with chronic kidney disease

- Use real world data to compute
 - $d(t)$: number of events at time t
 - $n(t)$: number still at risk before time t
- Add laplace noise to $d(t)$ and $n(t)$ \rightarrow **Analysis**
 - $d'(t) = d(t) + \text{laplace}(1/\epsilon)$
 - $n'(t) = n(t) + \text{laplace}(1/\epsilon)$
- Use privatized counts to estimate the hazard \rightarrow **Answer**
 - $h'(t) = \frac{d'(t)}{n'(t)}$
- Compare the hazard distribution between 2 neighboring datasets ie $D_1 = a + \text{Joe}$ vs. $D_2 = a$
 - To achieve complete privacy the hazard estimate would not change with the presence or absence of Joe's data



A look into the tuning parameters



- There is a trade off between data privacy and data utility
- Lower epsilon, increased noise, stronger privacy, but decreased data utility in terms of generating insights from this data
- Important to balance introducing privacy guarantees while ensuring insights generated from the data are scientifically sound.
- We can reduce the impact of introducing noise by collecting larger datasets allowing key patterns to emerge despite noise integration

Epsilon

- Factor that limits the deviation of the output ie regulates how much the output of the mechanism can vary between two neighboring databases
- If person A's data is in the dataset vs. removed from the dataset, the probability of getting any output changes at most by a factor of e^ϵ
- Small epsilon = higher privacy but potentially worse accuracy of output
- US Census 2020: $\epsilon = 4$
- Every new query weakens overall privacy and increases ϵ as attackers can combine information across outputs

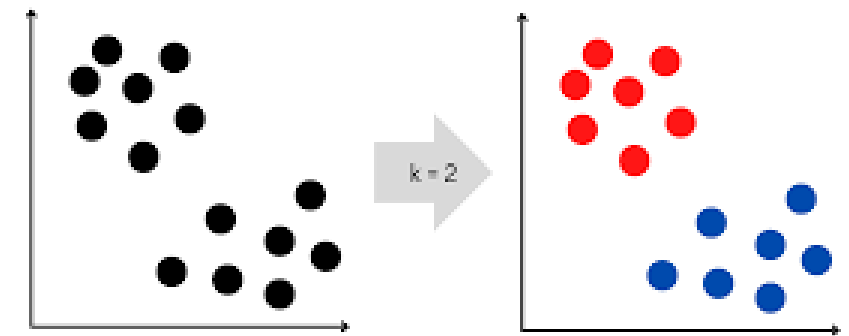
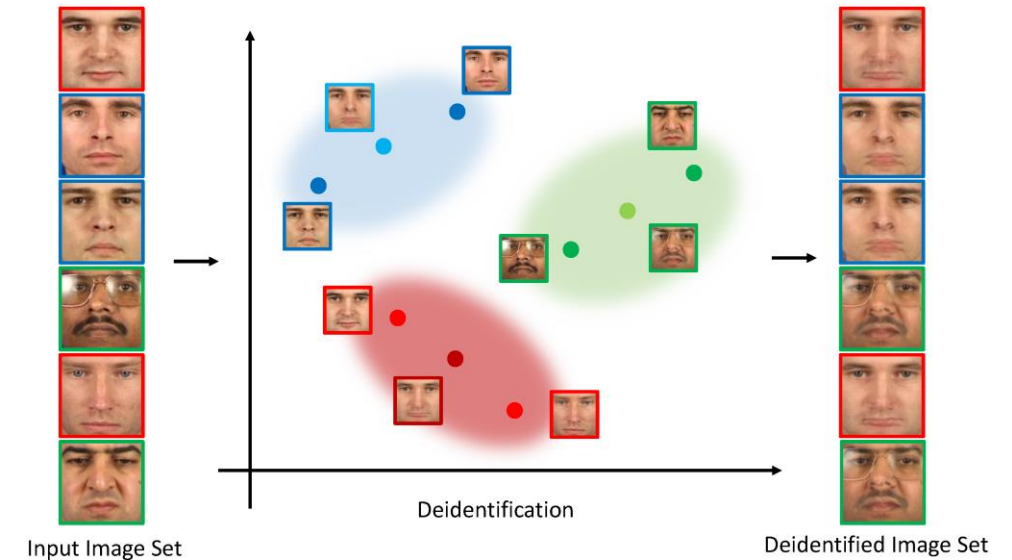
ϵ	Details
0.01	Heavy noise, strong privacy
0.1	Strong protection, strong privacy
1	Moderate privacy practical for many use cases
5-10	Weak privacy; used when data is less sensitive and when utility of data is critical

$$P[\text{Mechanism}(\text{Input}_{\text{dataset } 1}) \in \text{output}] \leq e^\epsilon * P[\text{Mechanism}(\text{Input}_{\text{dataset } 1}) \in \text{output}]$$

Optimizing your privacy budget

Before defining epsilon consider:

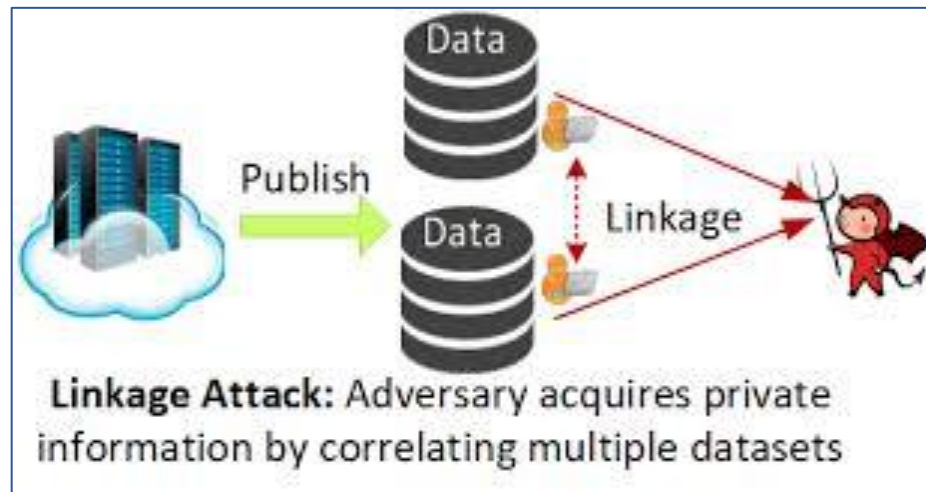
- How sensitive is your raw dataset to any particular individual/data point?
- High sensitivity \rightarrow larger impact of a row on the output of a mechanism \rightarrow higher epsilon
- Who are your data users? External use vs. internal controlled access
- What is the intended use of query results?
- What is the sample size of your study?
- Methods to gauge sensitivity:
 - K-anonymity: any person's records cannot be distinguished from at least $k-1$ other records based on quasi-identifiers
 - L-diversity: extension of k-anonymity; for each group of k records that share the same quasi-identifiers there are at least L different values for sensitive attributes



Differentially Privatized Synthetic Datasets

Current State:

Shares Real Patient Data without PHI/PII

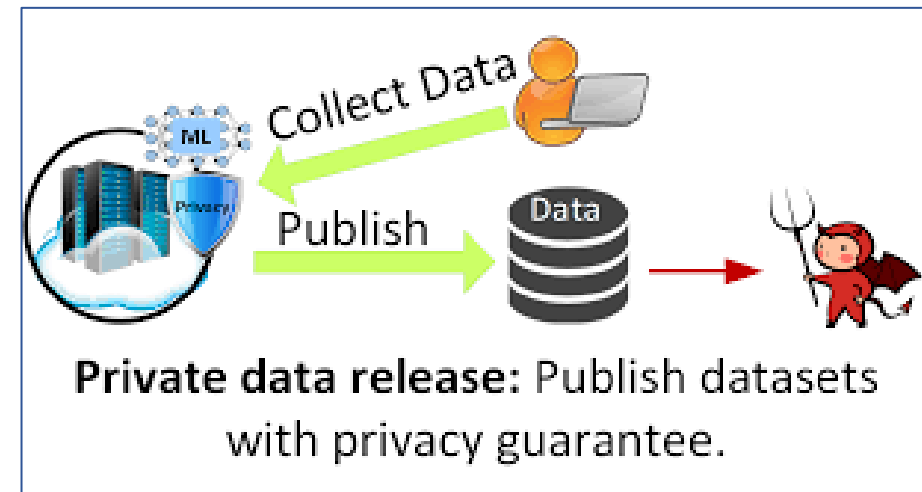


Conventional Anonymization process leaves the door for privacy loss through linkage attack

- Replace traditional identifiable information.
- Unknown level of privacy guarantee.

Future State:

Differentially Privatized Synthetic Data Generation



Differential privatized synthetic data enables NIH to provision datasets with a privacy guarantee.

- protects the privacy of individuals in datasets
- allows increased and faster access of researchers to health care research data
- addresses the lack of realistic data for software development and testing



Generating Synthetic Data

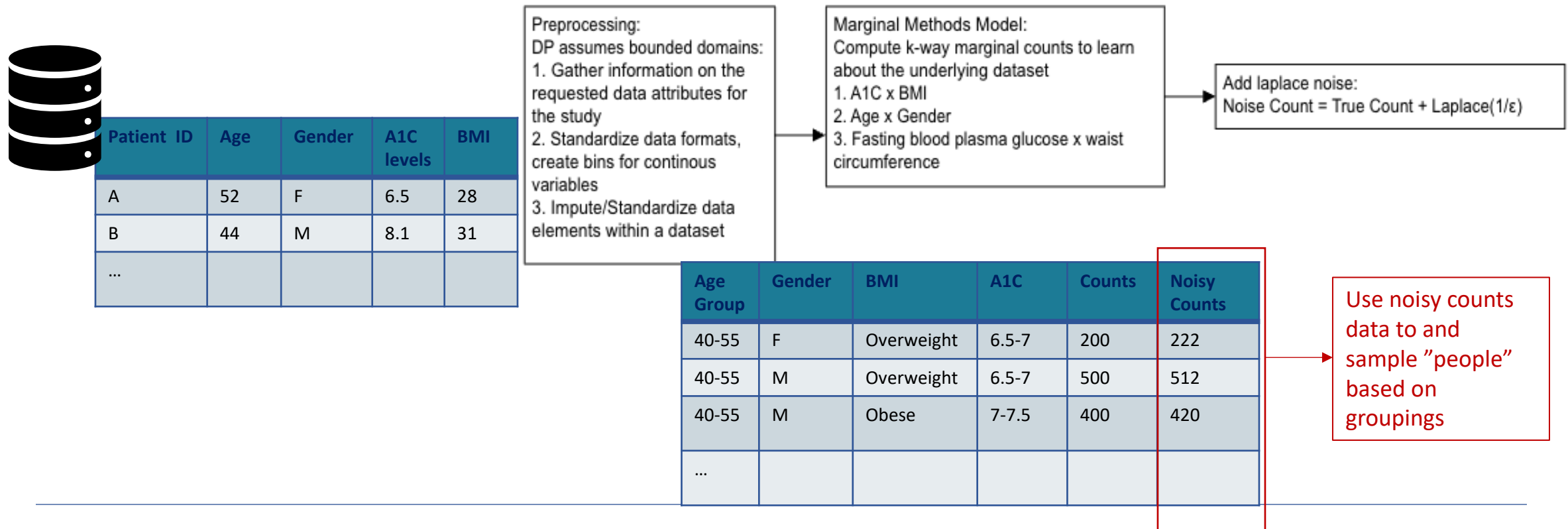
Steps to generate Differentially Privatized Synthetic Data

1. Understand the basics of Differential Privacy
 - Set privacy parameters such as Epsilon to ensure
 2. Data cleaning, imputation, and domain set up:
 - Data is consistent
 - Define the possible range for each variable
 3. Build a privatized model of the overall population
 - Marginals Method: Evaluate the summary of relationships between variables within a dataset with added DP noise
 - Probabilistic Graphical Models (PGMs): learn the dependencies between variables, privatize the model and then sample synthetic data
 - Bayesian Networks
 - Markov Random Fields
 - Differentially Private Generative Adversarial Networks: Train a neural network generator with noise added to gradients or outputs
 4. Create fake individuals sampled from the privatized model
-

Overview of A Simple Example Pipeline

Scenario: You manage a large, NIH-funded cohort study using electronic health record (EHR) data to track patients with Type 2 diabetes across multiple health systems. A collaborating research team requests patient-level data to develop models predicting risk of diabetes-related complications (e.g., kidney disease, retinopathy). Requested variables include lab values (A1C, creatinine), medication history, comorbidities, and demographic factors

How can differential privacy be used in this scenario?



Overview of A Simple Example Pipeline

Scenario: You manage a large, NIH-funded cohort study using electronic health record (EHR) data to track patients with Type 2 diabetes across multiple health systems. A collaborating research team requests patient-level data to develop models predicting risk of diabetes-related complications (e.g., kidney disease, retinopathy). Requested variables include lab values (A1C, creatinine), medication history, comorbidities, and demographic factors

How can differential privacy be used in this scenario?

Patient ID	Age	Gender	A1C levels	BMI
ABCD	41	F	6.7	27
EFGH	55	M	7.2	34
...				

Resulting dataset that
contains synthetic
data based on original
data features

Is your synthetic data reflective of the correlations and features found in the original dataset?

- Compare feature distributions of synthetic vs. real data (correlation, ML model performance)
- Compare noise of synthesis to noise of sampling error
- Benefits:
 - Low risk of individuals being re-identified as the data is synthetic
 - Low risk of drawing incorrect conclusion



Open-Source Differential Privacy Packages

Library	Language(s)	Noise Mechanism(s)	Primary Use Case	Key Features
Google Differential Privacy (+ PyDP wrapper)	C++, Python, Go, Java, Apache Beam	Laplace, Gaussian	Descriptive Statistics, Analytics	Used for releasing aggregate statistics (counts, sums, means) on large datasets while preserving user privacy. Ideal for government, enterprise reporting, or dashboards; Scalable; Widely adopted in data analysis workflows; Configurable ϵ/δ settings
brubinstein/diffpriv	R	Laplace, Gaussian, Exponential, Bernstein	Statistical DP for numeric & categorical data;	Provides multiple DP mechanisms, supports custom function privatization, ideal for research and educational purposes, well-documented with R vignettes.
IBM DiffPrivLib	Python	Laplace, Gaussian	Machine Learning (DP models)	DP-compliant versions of scikit-learn estimator; Easy Python integration for ML and statistics; Enables researchers and practitioners to train and evaluate machine learning models under differential privacy constraints.
TensorFlow Privacy	Python (TensorFlow)	Gaussian	Deep Learning with DP	Used in production ML pipelines to ensure training data remains private, especially in sensitive domains like healthcare, finance, and mobile personalization.



Summary

- Today's best practices (e.g., under NIH DMS policy) emphasize controlled access and ongoing governance that can evolve as risks change — not one-time de-identification + release.
- PETs like differential privacy and PPRL provide stronger, formal privacy protections beyond de-identification.
- There's no one-size-fits-all solution: organizations must evaluate technical feasibility, governance readiness, and legal alignment. Often, combining PETs may offer the best solution for complex privacy challenges.



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository



Q&A and Poll



National Institute of
Diabetes and Digestive
and Kidney Diseases

Central Repository

Contacts:

- Anya Dabic - dabic_andrijana@bah.com
- Shruti Gautam - gautam_shruti@bah.com

Upcoming Webinar: Challenges, Opportunities, and Considerations for Researchers using Electronic Health Records and Real-World Data Sources

- **Date:** July 31st from 2-3pm ET
- **Experts:** Datavant
- **Scan the QR code register**



Thank You!