Privacy Enhancing Technologies (PETs) 101

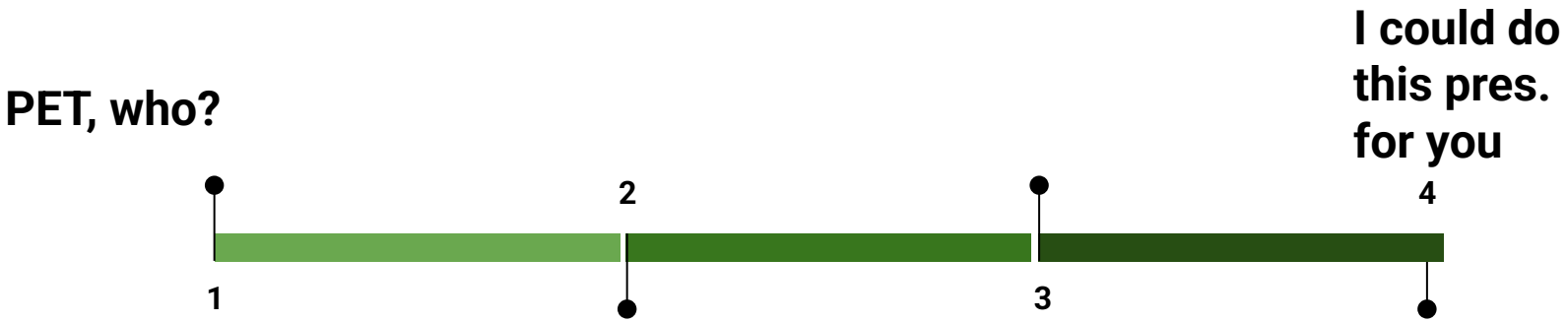
Stephanie Straus

Massive Data Institute
McCourt School of Public Policy

A4L Annual Meeting
March 18, 2025

GEORGETOWN UNIVERSITY

Pulse Check: Knowledge of PETs



Common Use Cases of PETs for Education

Linking across agencies/institutions

Training data for development teams/new applications

Responding to microdata requests from researchers

Responding to data requests from legislators, media, public

What am I missing?

Current Privacy Protection Methods

- Lockdowns
- Data use:
  - Rely on contracts (licenses, NDAs)
    - Share full copy of data with requestor
    - Trusted third parties
- Data release:
  - Statistical disclosure controls
    - Rounding, swapping, suppression

Current Privacy Protection Methods → Not Enough!

- Lockdowns

Full privacy, but no utility!

- Data use:
  - Rely on contracts (licenses, NDAs)
    - Share full copy of data with requestor
    - Trusted third parties

Addtl pair of eyes on raw data + addtl copies of data in new locations

Trusting parties will adhere to contract terms

- Data release:
  - Statistical disclosure controls
    - Rounding, swapping, suppression

Computer power can possibly re-ID

Full privacy, but questionable utility!

What Are Privacy Enhancing Technologies (PETs)?

- Cryptographic techniques that increase data protection while allowing for greater data utility
- Also known as Privacy Preserving Technologies
- Can enhance how data are analyzed and/or published

PETs may not be compliant, and do not guarantee complete privacy.

PETs complement, but don't replace, DSAs or good governance protocols.

PETs are safer and more secure ways to analyze, link, and share data

How PETs Address These Issues

Input Privacy


Privacy preserving record linkage (secure hashing)


Trusted execution environments / enclaves

Intermediaries


Secure multiparty computation


Homomorphic encryption

Output Privacy

Traditional SDL


Differential privacy


Private query server


Synthetic data

Common PETs and Education Examples

- PPTs in Education project
  - Focus on K-12 and higher education
  - Conduct stakeholder interviews to:
    - Identify existing and abandoned PPT projects
    - Elucidate the barriers
    - Find entry points to implementation

PETs in Education Landscape Analysis

Table 3. PPTs in Education

PPT Type	Project	Ongoing?
secure hashing	Birth through Eight Strategy for Tulsa (BEST) with Tulsa Public Schools and Oklahoma Policy Institute Asemio	Yes
secure hashing	Oregon Integrated Client Services Data Warehouse (ICS) Oregon Departments of Education, Health, Human Services, and others	Yes
secure hashing	SILK hash encoding with Administration for Children and Families and education agency partners Georgia Policy Labs	Yes
secure hashing	Hashed matching algorithm for Virginia Longitudinal Data System* University of Virginia Social and Decision Analytics Lab	No
SMC	Boston Women's Workforce Council gender pay gap study Boston University	Yes
SMC	National Post-secondary Student Aid Study-Federal Student Aid linkage demonstration National Center for Education Statistics (NCES) and Georgetown University	No
SMC	Virginia Longitudinal Data System - Defense Advanced Research Projects Agency (DARPA) demonstration State Council of Higher Education for Virginia	Yes

Table 3. PPTs in Education

PPT Type	Project	Ongoing?
TEE	Federated data model with joins on demand for Education Providers Silicon Valley Regional Data Trust	No
TEE	Secure data enclave for research access to student school district records on student social, emotional, academic, and physical well-being Character Lab, University of Pennsylvania	Yes
TEE	Education Research Data Center, WA Departments of Children, Youth, and Families, State Board of Education, and others State of Washington	Yes
TEE	LearnLab's DataShop, world's largest repository of learning interaction data, spun off from Cognitive Tutors program LearnLab, Carnegie Mellon University	Yes
TEE	Secure virtual data enclave for research access to NCES Restricted Use Files Institute of Education Sciences, Coleridge Initiative	Yes
TEE	Secure virtual enclave for research access to safety, health, and outcomes data on children Children's Data Network, University of Southern California	Yes
DP	Post-Secondary Employment Outcomes (PSEO) Census Bureau, multiple post-sec. institutions	Yes
DP	College Scorecard IRS, Statistics of Income Division, Department of Education, Tumult Labs	Yes

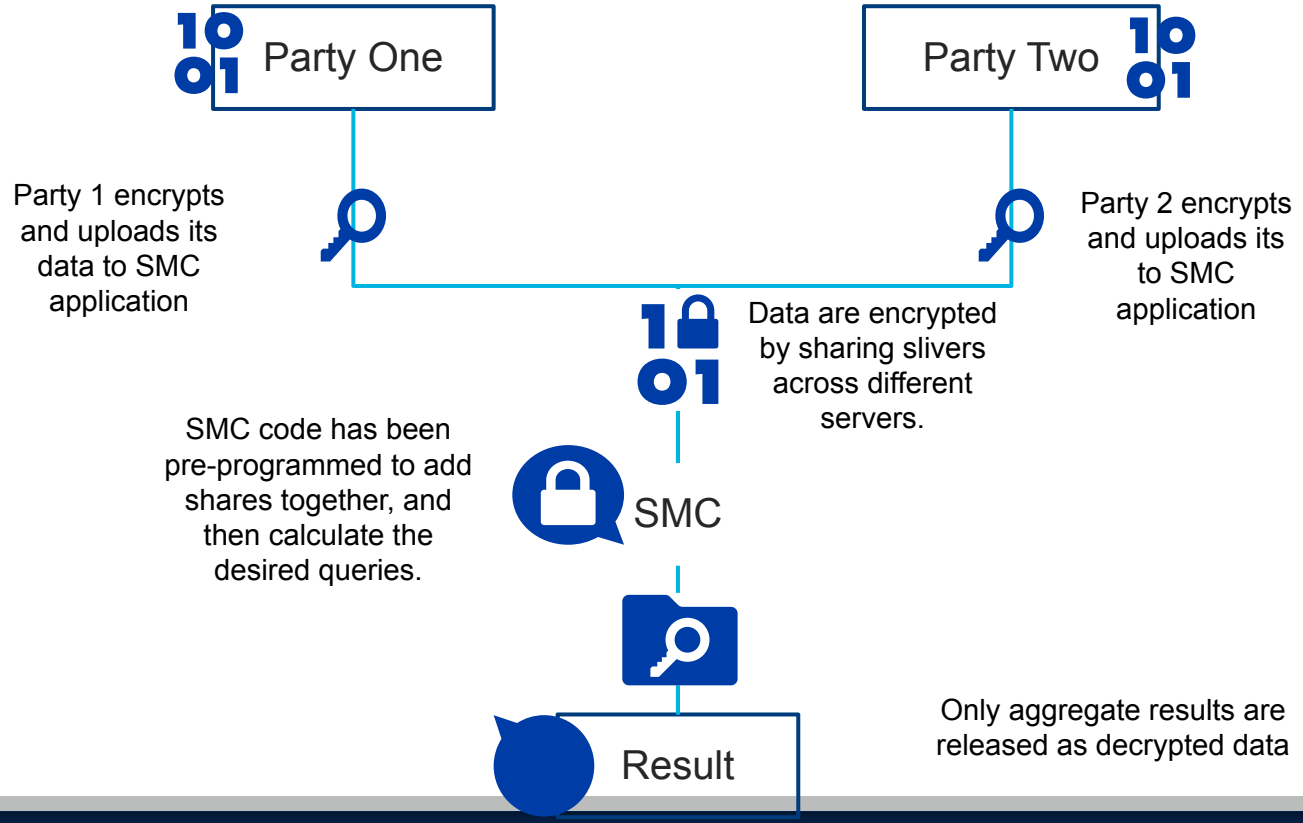
O'Hara, A. & Straus, S. (2022, Jul). Privacy Preserving Technologies in Education. Georgetown University, Washington, D.C.

Secure Multiparty Computation (MPC)

Secure multiparty computation (MPC): the process by which two distrusting parties jointly compute a research query on their datasets, without ever seeing the other's underlying data.	
☑ only aggregate results released ☑ descriptive statistics ☑ finds overlap in datasets ☑ no trusted third party sees data	☒ time-consuming ☒ limited in operations/statistics ☒ requires careful data preparation ☒ does not address output privacy

- Examples: Estonia, Virginia, our NCES demonstration, Boston Women's Workforce, Allegheny County Department of Human Services demonstration, DARPA and IARPA investments

MPC Overview

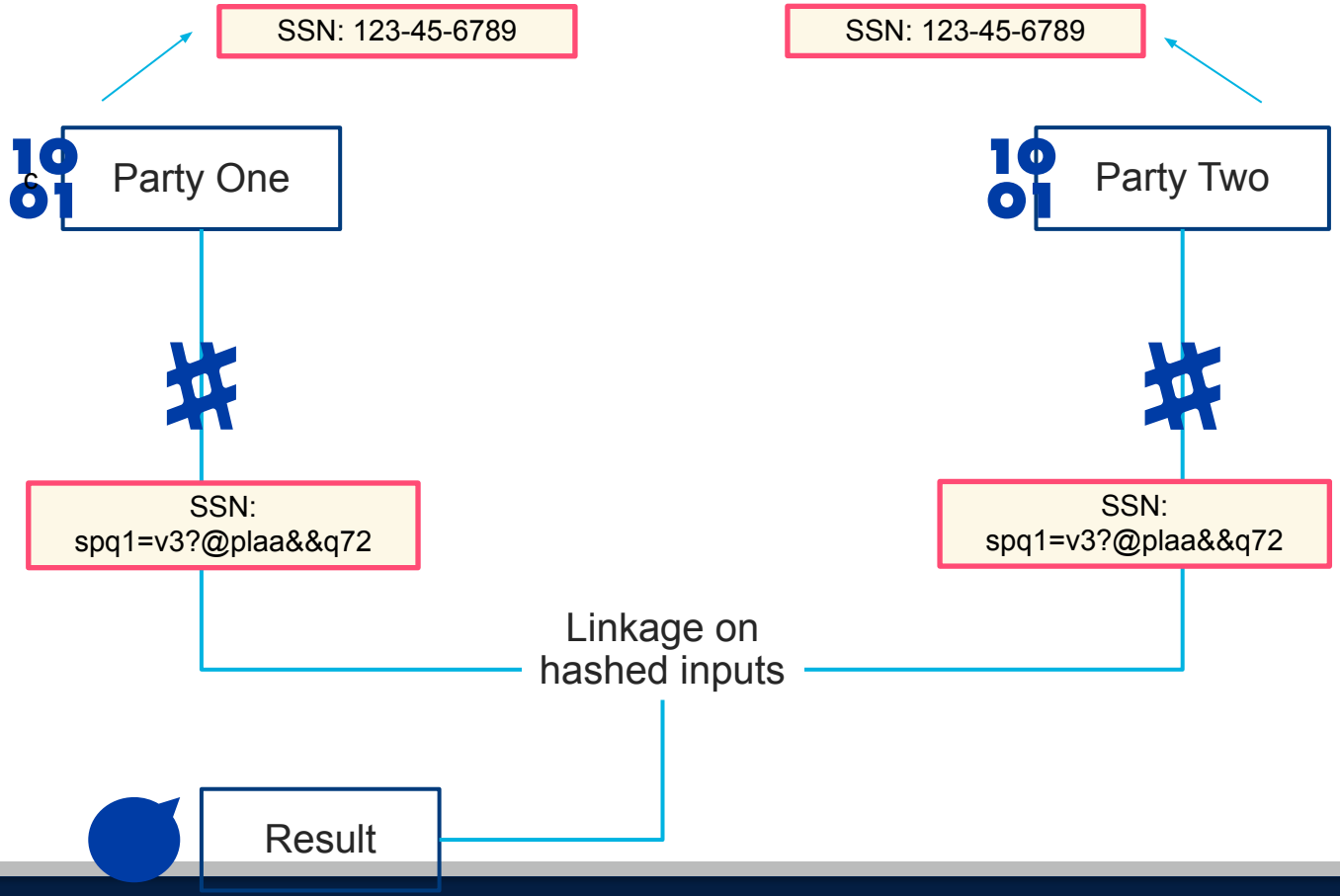


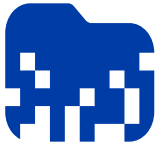
Secure Hashing

Secure hashing: an algorithm that replaces sensitive inputs with a random string of characters (hash) unique to each original record in the data. Can be used to protect passwords or link across datasets.	
☑ generally, not reversible (hash cannot be decrypted, one way) ☑ NIST-approved, developed Secure Hashing Algorithms (SHAs)	☒ protects identifying or linkage fields though other fields may also be disclosive ☒ does not address output privacy

- Examples: Coleridge Initiative's ADRF, Georgia Policy Labs, CAPriCORN in healthcare, N3C system at NCATS, ASPE at HHS

Secure Hashing Overview



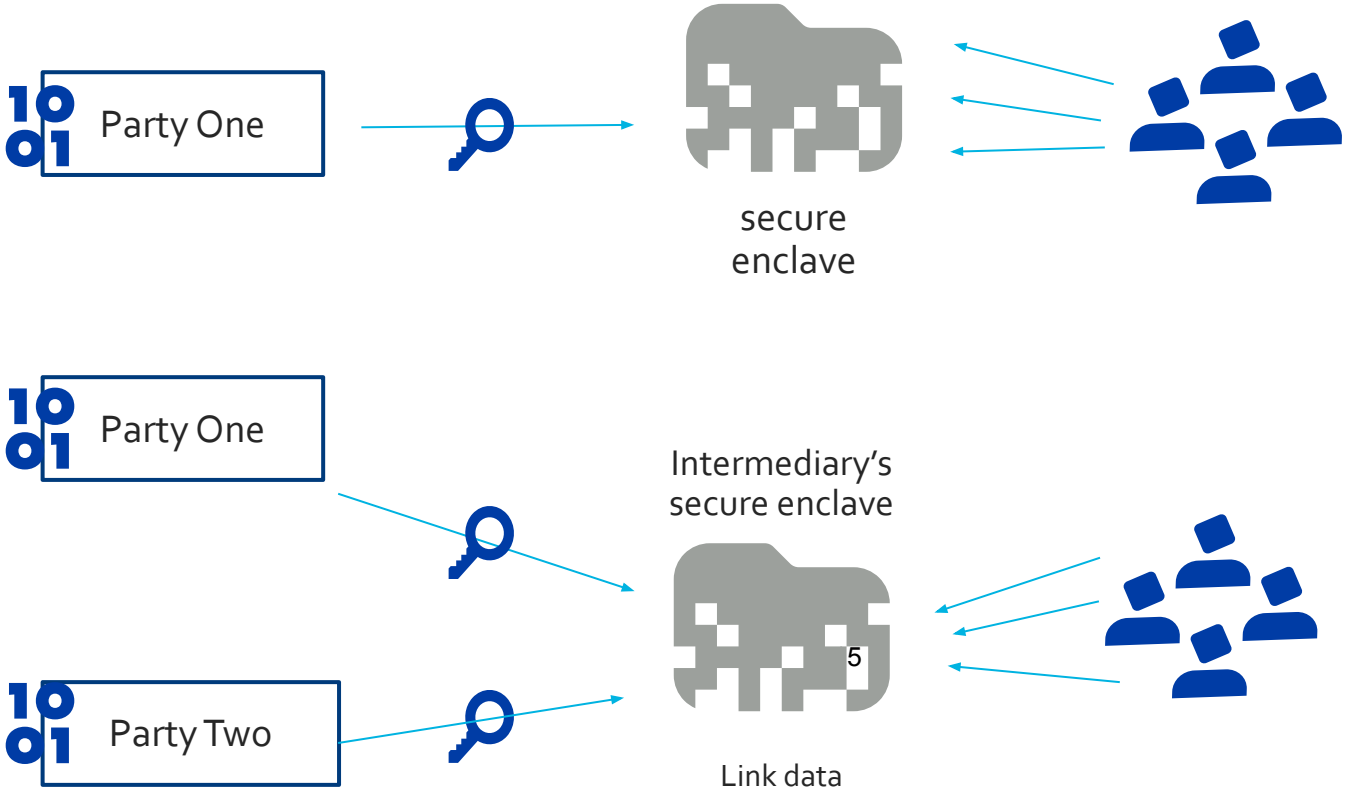


Secure Enclave

Secure enclave: a virtual computing workspace that enables authorized users to access sensitive data and securely conduct analysis.	
☑ supports open-ended queries, not limited in operations/statistics ☑ external users cannot download or extract data without permission	☒ trusted third party sees data (to de-identify, provision, etc.) ☒ external users see potentially sensitive data ☒ must layer output privacy protections

- Education examples: Washington State Education Research Data Center (ERDC), Children's Data Network, Character Lab, UPenn, Coleridge Initiative
- Other examples: Crypto, industry applications that have trusted code on trusted hardware

Secure Enclave Overview



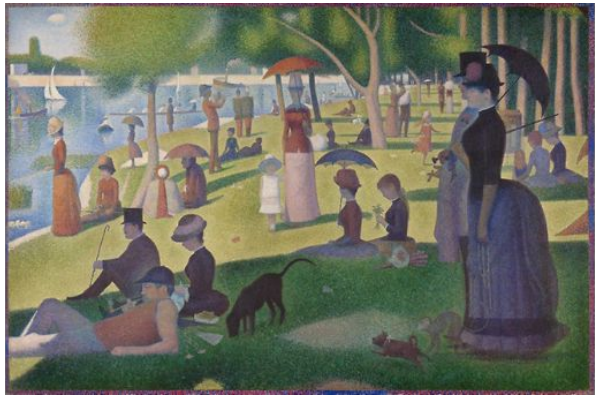
Differential Privacy (DP)

Differential privacy (DP): a method for obscuring identities or attributes in the underlying record-level data by infusing results/statistics with noise.

☑ reduces re-identification risks for individuals or groups in the data (i.e., students, programs) ☑ provides a formal privacy guarantee (can guard against threats known today and those in the future) ☑ useful for known queries	☒ challenging to implement on low levels of geography or unique population groups without adding a lot of noise ☒ tradeoff between privacy and accuracy – as you add more "noise" (protection) you move further from true values ☒ no input privacy

- Education examples: U.S. Census Bureau's Post-Secondary Employment Outcomes, College Scorecard
- Other examples: Census 2020, Google, Apple, Facebook

Differential Privacy (DP) Overview



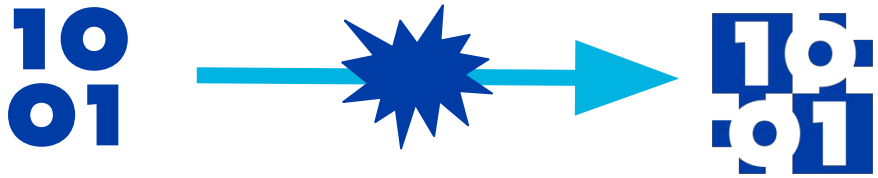
Original

Suppression

Noise Injection

Protecting Your Privacy in a Data-Driven World by Claire McKay Bowen, published by Chapman and Hall/CRC, 2021.
bit.ly/3maky9e

Differential Privacy Overview



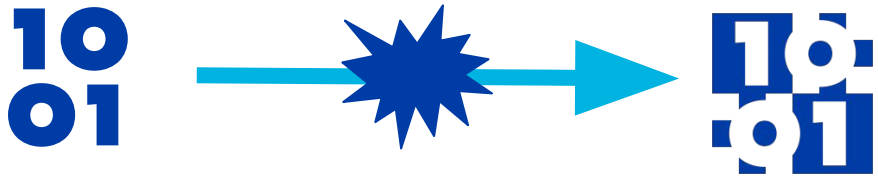
A County	323	+ 27
B County	4,002	- 24
D County	5,275	- 35
E County	427	+ 16
F County	112	+ 14
G County	3,936	- 17
H County	230	+ 8
I County	75	+ 12

A County	350
B County	3,978
D County	5,240
E County	443
F County	126
G County	3,919
H County	238
I County	87

18

- But what about non-numeric fields?

Differential Privacy Overview



Student A	White
Student B	Black
Student C	Asian
Student D	White
Student E	AIAN
Student F	White
Student G	Black
Student H	White
Student I	Asian
Student J	White
Student K	White
Student L	White

change
don't change
change
don't change
don't change
don't change
don't change
change
don't change
don't change
change
don't change

coin toss

Student A	Asian
Student B	Black
Student C	White
Student D	White
Student E	AIAN
Student F	White
Student G	Black
Student H	Asian
Student I	Asian
Student J	White
Student K	White
Student L	White

What It Takes to Implement

- In addition to laws, statutes, and regulations governing data sharing and use:
  - Leadership and strategic goals
    - Build PET integration into budget, learning agendas
  - Institutional sign-offs from IT, information security
    - Security checks
    - Procurement processes for software and code
    - PET deployment location, storage
  - Staff capacity, funds, and available resources.

Common Use Cases of PETs for Education (Again)

Linking across agencies/institutions


PPRL


HE


SMC


synthetic data

Training data for development teams/new applications

Responding to microdata requests from researchers


enclave or TEE


synthetic data


private query server


private query server


differential privacy

Responding to data requests from legislators, media, public

Our PETs Demonstrations

- Completed:
  - U.S. Dept of Education postsecondary data demonstration (report)
  - Nebraska Statewide Workforce & Educational Reporting System (report)
- Ongoing:
  - Arkansas CDO Team
  - DC Office of Education Through Employment Pathways
  - Connecticut Office of Policy and Management
  - Secure query system for the Internal Revenue Service

GEORGETOWN UNIVERSITY
McCourt School of Public Policy

MASSIVE DATA INSTITUTE

For more info, see our:

- PETs Primer
- PETs in Ed paper
- PET 101 Trainings

…and more, on our website.

stephanie.straus@georgetown.edu