Session I-2: Deploying PETs within Integrated Data Systems: Learnings and Challenges

Amy O'Hara, Ph.D., Georgetown University, Massive Data Institute
Stephanie Straus, Georgetown University, Massive Data Institute

Federal Committee on Statistical Methodology Research and Policy Conference

October 24, 2024

Agenda

- Our Privacy Enhancing Technologies (PETs) work
- Synthetic Data and Differential Privacy
- Determining use cases
- Considerations before deploying
- Next steps

Our Work on PETs at MDI

- Gates Foundation-funded work since 2021
  - PETs in Statewide Longitudinal Data Systems and P20W systems
- PET pilot demonstrations
- PETs resources on our website
  - PET 101 Trainings, PET Primer, PET Zotero library

→ Focus on synthetic data and DP, but PETs in general

3

How do you know which PET is right for you?

- Determining use cases first
  - Linking, access, disclosure avoidance?
- Value the PET could provide to your agency
- Considering data quality, format, and schema
  - Ex. Longitudinal? Relational, integrated database? Small cell sizes?

Example: How NSWERS realized they needed synthetic data

- Use cases
  - External researchers
  - EdTech vendors
  - Internal users
  - Developers
- Value the data can provide
  - Reduce administrative hurdles: Stakeholders can see correctly formatted data before signing DSA
  - Reduce legal burden: DSA execution struggles
  - Reduce privacy and security concerns: less PII being shared

What to consider before implementing these PETs?

- *Have need or use case ID'd first.*
- Laws, statutes, and regulations governing data sharing and use.
- Leadership and their strategic goals.
- Institutional sign-offs.
- Staff capacity, funds, and available resources.

How I've Come to These Conclusions…

NCES Secure Multiparty Computation Demonstration

NATIONAL POSTSECONDARY STUDENT AID STUDY (NPSAS)

IES Institute of Education Sciences

United States Department of Education

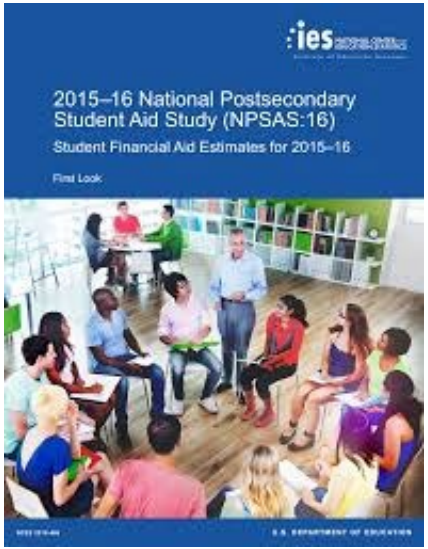
(institution-level data)

+

Federal Student Aid — An OFFICE of the U.S. DEPARTMENT of EDUCATION | NSLDS

(federal grant and loan data)

=



Archer, D., O'Hara, A., Issa, R., and Straus, S. (2021, May). Sharing Sensitive Department of Education Data Across Organizational Boundaries Using Secure Multiparty Computation. Georgetown University, Washington, D.C.

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.

PET Pilot Advising

- Nebraska Statewide Workforce & Educational Reporting System → Synthetic data (see our report)
- ARData (Arkansas) → PPRL
- DC Office of Education Through Employment Pathways → Schema alignment + PPRL

Consideration: Laws, statutes, and regulations governing data sharing and use.

- No current guidance on how PETs interact with existing data privacy statutes
  - Up to individual jurisdictions
  - District, state, and federal laws and regs
- Some questions to consider:
  - Is this PET use considered a disclosure? Is this PET application considered ‘reasonable’ enough privacy protection? Are the resulting data ‘de-identified’ or ‘pseudonymized’?
- Who needs to see/’see’ the data to deploy the PET? Can they?

Consideration: Leadership and their strategic goals.

- Advocate for your PET deployment across departments and up the chain of command
  - See Staff capacity, funds, and available resources slide
  - See Institutional sign-offs slide
- Build PET integration into budget, strategic plans, learning agendas

Consideration: Institutional Sign-Offs

- Information Technology staff, information and security officers, procurement staff
- Using service provider, technical assistance provider, or using open source software?
  - Possible to contract / pass funds to provider?
- PET deployment location
  - On premise, or in cloud / virtual machines?
  - Available storage?
- Procurement processes for proprietary software and code
- Procurement processes for open-source software and code

Consideration: Staff capacity, funds, and available resources.

- Staff and skill sets needed
  - Data scientists, statisticians, database administrators?
- FTE / time commitment for project duration
- Funds
  - Staff time, service provider, new technology, and compute/storage space
  - Can you do this without a TA provider?
- Assess sustainability

Next Steps for our PETs work

- Complete ARData and DC ETEP PET Pilots
  - Produce reports such as the NSWERS publication
- Refine our lessons learned from the pilots into screening tools and how-to guides
- Advocate for legal and regulatory guidance
- Stakeholder outreach to keep understanding data gaps

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.

amy.ohara@georgetown,edu
stephanie.straus@georgetown.edu