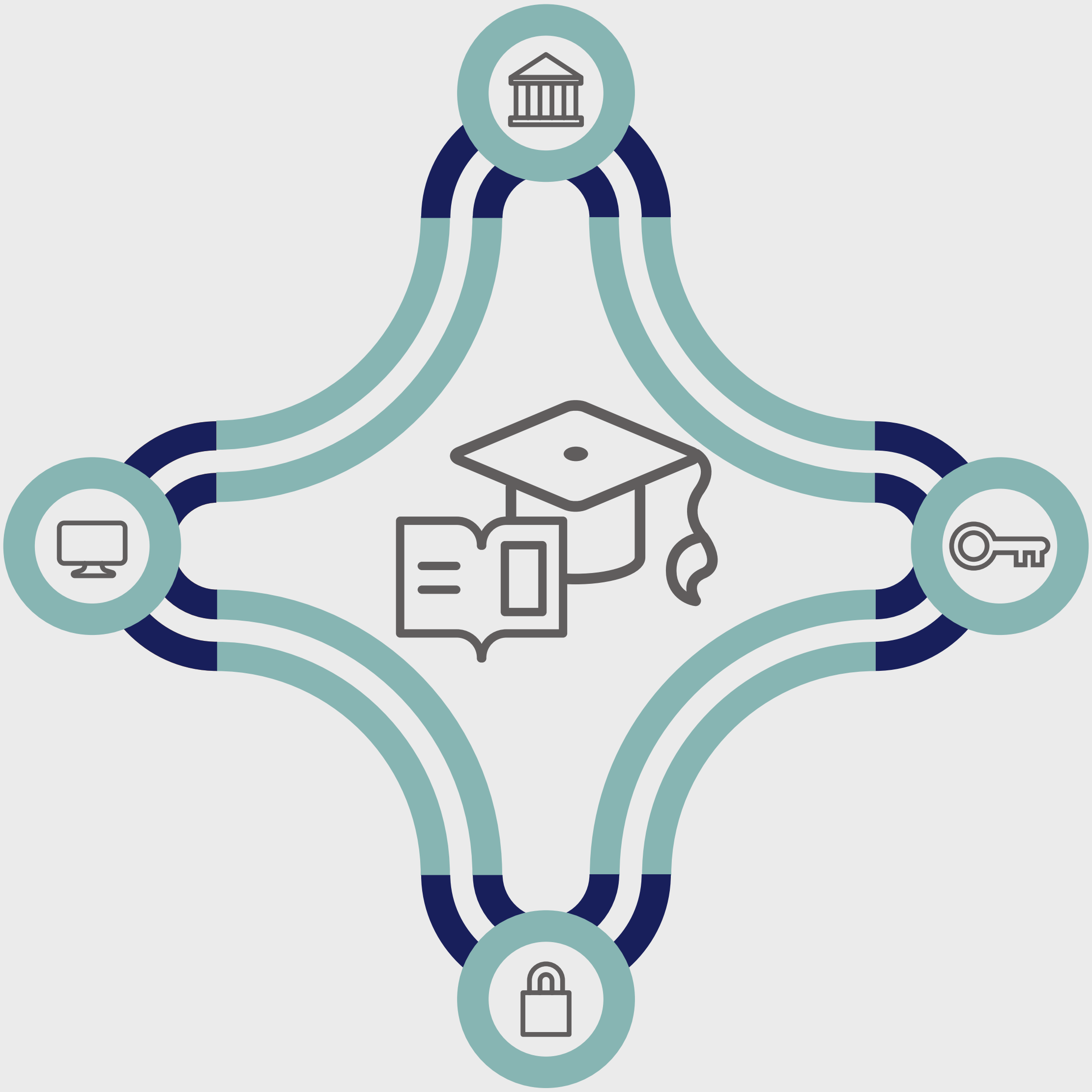
Postsecondary Data Infrastructure: What is Possible Today

AUTHOR: AMY O'HARA, GEORGETOWN UNIVERSITY

JUNE 2019

MDI
GEORGETOWN UNIVERSITY
MASSIVE DATA INSTITUTE
McCourt School of Public Policy

POSTSEC DATA



Dr. Amy O'Hara is a Research Professor in the Massive Data Institute at Georgetown University. Her areas of expertise include data governance, data integration, and privacy.

This paper is part of the larger series *Protecting Students, Advancing Data: A Series on Data Privacy and Security in Higher Education.* In August 2018, the Institute for Higher Education Policy (IHEP) first convened a Privacy and Security Advisory Board of privacy and security experts and higher education experts to explore some of the most pressing data privacy and security issues of the day. The resulting paper series serves as a resource for policymakers as they develop sound postsecondary data policy and centers privacy and security as a top priority. This report is based on research with support from Arnold Ventures. The findings and conclusions contained within are solely those of the author.

Introduction

Data sharing across government agencies allows consumers, policymakers, practitioners, and researchers to answer pressing questions. Creating a data infrastructure to enable this data sharing for higher education data is challenging, however, due to legal, privacy, technical, and perception issues. To overcome these challenges, postsecondary education can learn from other domains to permit secure, responsible data access and use. Working models from both the public sector and academia show how sensitive data from multiple sources can be linked and accessed for authorized uses.

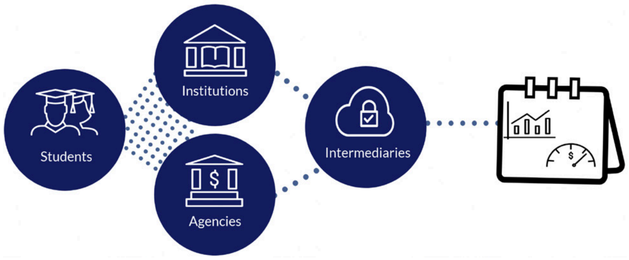
This brief describes best practices in use today and the emerging technology that could further protect future data systems and creates a new framework, the "Five Safes", for controlling data access and use. To support decisions facing students, administrators, evaluators, and policymakers, a postsecondary infrastructure must support cycles of data discovery, request, access, analysis, review, and release. It must be cost-effective, secure, and efficient and, ideally, it will be highly automated, transparent, and adaptable. Other industries have successfully developed such infrastructures, and postsecondary education can learn from their experiences.

A functional data infrastructure relies on trust and control between the data providers, intermediaries, and users. The system should support equitable access for approved users and offer the ability to conduct independent analyses with scientific integrity for reasonable financial costs. Policymakers and developers should ensure the creation of expedient, convenient data access modes that allow for policy analyses.

Who is involved in data sharing?

Data, from genesis to analysis, involves a wide variety of stakeholders. In the education context, students are the owners of their data. Student data is maintained and generated through high frequency interactions with educational institutions and government agencies. From generation, data can flow in a few directions:

- Institutions and agencies may share data, within legal and security constraints, acting as data providers.
- They may securely transmit information to data intermediaries *(defined in Appendix A: Key Terms)* who standardize and match data across sources or over time. Intermediaries can also provide this service for many institutions, increasing the efficiency and security of data processing necessary to produce insights for data consumers.



The conditions by which data are shared and analyzed are strikingly similar across sectors like healthcare, housing, human services, and workforce, though the motivations for providing data or analysis may vary (e.g., original analysis, regulatory/legislative mandate). Some data providers believe their duty is to grant (legal, safe) access,[1] while others may also wish to be (or appear) socially engaged or trustworthy.[2] Regardless of motivation, a provider's desire to share data can easily be overwhelmed by legal and reputational risks, including embarrassment over results or erroneous inferences, discrepancies in their data or errors in previous releases, spills or compromised identities, and negative reactions from data subjects (or their proxies and advocates). A robust data infrastructure must have strong controls in place to mitigate these risks.

Five Safes: A New Framework for Data Access and Use

The "Five Safes" framework describes an approach for controlling data access and use. The five safes are: safe projects, safe people, safe settings, safe data, and safe outputs.[3]

SAFE PROJECTS

Building safe projects requires governance protocols to control project requests, review, and approval processes, and may require institutional board or ethics board review and approval. Clear and thorough data use agreements (DUA) also contribute to safe projects by articulating acceptable uses, linkages, and scopes for analyses.[4]

Federal agencies often have DUAs with other government units (i.e., federal, state, local), universities, intermediaries, for-profit, and not-for-profit organizations. For example, the Census Bureau has agreements with entities, including:[5]

- The Department of Housing and Urban Development (HUD) to obtain data on voucher-assisted renters, public housing units, and Federal Housing Administration-insured loans. These data are used to improve the quality of the American Housing Survey (AHS) and American Community Survey (ACS) and in research projects to understand the cost and adequacy of rent assistance.
- Two cities for Homeless Management Information Systems data to improve population measurement.
- Numerous vendors to buy data extracts including property tax, deed, foreclosure, and multiple listing service data from Corelogic to improve the ACS, AHS, and other address lists.
- 21 state Supplemental Nutrition Assistance Program agencies to obtain case-level, monthly program participation information. These data are used in joint research with the United States Department of Agriculture to produce reports for contributing states. The data also support research on program participants and program efficiency, including studies on work requirements and local labor demand.
- The Institute for Research on Innovation and Science at the University of Michigan to link federal grantee data to individual and firm census data to measure the impact of research funding.
- A county United Way 211 agency to evaluate call data relative to population density and demographics.
- The University of Texas system to study labor market outcomes for college graduates.

Research teams are pursuing ways to automate agreement formation and data usage controls. For example, the Research Data Alliance is working on policy development, and lawyers and computer scientists are collaborating to develop new "smart" contracts that encode rules and permissions to automatically execute pre-defined functions.[6] This algorithmic approach to permissions requires clear knowledge and interpretation of laws, rules, and policies, allowing data-use laws to be translated into "if-this-then-that" terms in the contract. Such a logic-driven approach can define allowable data uses, obligations to regulate data access, allowable linkages, and provisioning and release requirements. Researchers are exploring the privacy requirements in certain laws to draft logic to test this approach.[7]

SAFE PEOPLE

Data users should be screened and trained to become "safe people." Currently, researchers must meet different requirements to obtain access to data systems, depending on the system and agency involved. For instance, obtaining access to a data system may require institutional attachment, proof of research competence (e.g., grants received, curriculum vitae), citizenship or tenure in the country, or mandatory training. Some providers currently require background checks and fingerprinting, while others only require joining a research team. In the future, a user's vetting and approval by one organization could carry over to other associated organizations; this will require durable credentials and agreed upon standards and training.

SAFE SETTINGS

The most important control factors involve the data user's interface and environment. Many current practices regulate data inputs, computation, and outputs, creating safe settings and safe data.[8]

SAFE DATA

Aligned with "safe settings," data users should create "safe data". The practices for both impose restrictions on what an analyst can use, what an analyst can do, the analyst's computing environment, and the analyst's physical location. Considerations for safe settings and data include:

1. What data can the analyst use?
   a. *Actual data.* Analysts typically use extracts that only include the data required to address their questions; or
   b. *Synthetic data*, containing information that looks like the actual data but contains none of the real cases. The analyst never sees the actual data; or
   c. *Encrypted data*, using methods like secure multi-party computation, discussed in more detail later. The analyst never sees the actual data.
2. What can the analyst do with the data?
   a. *No restrictions* on software, methods, or analyses, only adherence to the agreed-upon scope; or
   b. *Restrictions*, such as using specific statistical software or only submitting certain queries.
3. What computing environment will the analyst use?
   a. *Analyst has control*, using their own equipment after obtaining data through a license or contract; or
   b. *Data provider has control* and never gives the analyst a copy of the data. The provider removes the analyst's ability to take data out of the server, to print, or to make copies of the data. Approaches may include:
      i. Lending a laptop or server to the analyst for the duration of the project;
      ii. Using a device like an SD-Box that connects an analyst's computer to the provider's infrastructure;[9] or
      iii. Virtual access to the provider's data environment through an analyst's own equipment or at the provider's location.
4. Where does the analyst work?
   a. *No restrictions.* The analyst can work anywhere; or
   b. *In a designated space*, such as in their office or in a certain room; or
   c. *In physical enclave*, including government or academic research data centers, that may include alarms and cameras.

In the future, smart contracts (described above in "Safe projects") may specify and regulate what the analyst can see and do within a secure virtual environment. Public and private investments in synthetic data and validation servers[10] and use of biometric authentication devices may also improve remote access options.

SAFE OUTPUTS

The fifth safe protects the privacy of data subjects by reducing the risk of individuals being re-identified using the results in the data outputs. Such protection occurs through statistical disclosure limitation methods such as rounding, aggregating, and suppressing results to obscure unique observations in tables, figures, or maps. Some methods to safeguard outputs alter the data by swapping or noise injection. These techniques work by changing the ages or races of individuals in a sparsely populated area or changing income dollar amounts by a small amount, for example. These are common practices today, which make it more difficult for someone to learn which observation in the dataset is which.

The future of safeguarding outputs will involve formal, mathematical techniques that recognize that some data sets—and individuals—may require more privacy protection than others, and sometimes only certain attributes need to be protected.[11] These stronger privacy guarantees, achieved by slightly altering the underlying data, are necessary to adapt to the growing volume of data generated and discoverable today.[12] New privacy-enhancing techniques are maturing; how they will be applied, and to which standards, is unclear. Technical advances in this space are likely to expand data provisioning options in the future.

Safe data access models: Examples across government

Throughout government and the private sector, experts have built data linkages that allow for robust analyses, reflecting a variety of solutions that keep projects, people, settings, data, and outputs safe. A postsecondary data infrastructure could learn from these efforts and adopt similar approaches today.

SAFE DATA ACCESS IN ACTION: FEDERAL STATISTICAL RESEARCH DATA CENTERS (FSRDC)

The Census Bureau manages the Federal Statistical Research Data Center (FSRDC) network, offering secure access to census, survey, federal and state administrative data, and commercial data for approved research projects. The Census Bureau acts as a data intermediary, harmonizing and linking data from many sources for agency staff and the FSRDC labs. Other agencies, including the National Center for Health Statistics (NCHS), Agency for Healthcare Research and Quality, and Bureau of Labor Statistics also host data through the FSRDC.

FSRDCs provide a practical example of the five safes in practice.

- **Safe projects:** FSRDCs only allow statistical analyses—not enforcement, marketing, or surveillance. Access is granted on a need-to-know basis for each analysis. Each project using Census Bureau data must have a benefit to the agency's mission.
- **Safe people:** To qualify to use data in an FSRDC, analysts must work for a government agency or not-for-profit organization, be citizens or have been in the United States 3+ years, and pass a background check including fingerprinting.
- **Safe settings:** Work occurs in one of 29 FSRDC physical labs located at universities or federal agencies, which have badge access, alarms, cameras, and a Census Bureau administrator on site. Computer terminals within the labs permit virtual access to Census Bureau servers, and no data may be removed from the computing environment by the analyst.
- **Safe data:** Census Bureau staff can link datasets for analysis.
- **Safe outputs:** Data output passes through the administrator and an extensive disclosure avoidance review process before results can be published.

Takeaway for postsecondary education:

The FSRDC model for data access securely hosts rich postsecondary datasets for analysis. To implement this model, developers will need to manage protocols for details such as processing bottlenecks for data hosting, analyst credentialing, shifting disclosure review requirements, and providing remote access for analysts living far from the current labs.

SAFE DATA ACCESS PRACTICES IN OTHER GOVERNMENT AGENCIES

When developing the data-sharing and access mechanisms for the postsecondary data ecosystem, policymakers can draw on robust examples from other government agencies and sectors like health, defense, and housing. Each exemplifies at least one of the five safes.

- The Centers for Medicare and Medicaid Services (CMS) curates and provisions extracts of administrative data based on an analyst's needs, including personal identifiers when necessary, through a virtual research data center (RDC). CMS can link files for analyses, and researchers can use their own laptop to log into the CMS **safe setting**, a secure environment, from which no data leaves.
- The Defense Manpower Data Center (DMDC) offers access to federated Department of Defense data from more than 60 personnel-related data feeds in their Person Data Environment. Their secure cloud-based enclave offers virtual access to curated microdata in a **safe setting** for agency analysts and external researchers.
- HUD and the National Center for Education Statistics (NCES) both license data for approved research projects. They send data to researchers after a vetting process and non-disclosure agreements are signed. NCES conducts random in-person inspections to monitor data management practices, ensuring that their **safe people** are following rules on their **safe projects.**
- In additional to their participation in the FSRDC network, NCHS offers detailed demographics through its own RDC remote access. NCHS also permits analysis of more sensitive data (e.g., genetic, detailed geography, exact dates, linked files) in their physical RDC, which is separate from the 29 FSRDC labs. Beyond these safe settings, NCHS has a data linkage unit that can match files for analyses, providing **safe data** for their approved users.
- Also separate from the FSRDC, the Census Bureau has piloted the Postsecondary Employment Outcomes (PSEO) program with the Universities of Texas and Colorado systems. Census is expanding this program using data obtained by the Institute for Research on Innovation and Science (IRIS). These partnerships produce post-graduation employment outcomes by campus and degree field.[13] Researchers at IRIS have worked with the Census Bureau on privacy-protecting methods to release the data, assuring **safe outputs.** They infuse noise into the output data in a way that provides "provable differential privacy" and allows users to generate reliable results from queries that would otherwise have high risk of disclosing data for particular students.
- Internal Revenue Service (IRS) and Social Security Administration (SSA) both host statistical research programs that let external analysts use their data through **safe projects.** IRS accepts proposals for its Joint Statistical Research Program that enable researchers to use tax data for tax administration research.[14] SSA sponsors projects through their Retirement and Disability Research Center, permitting use of payroll tax and earnings data, disability, and pension microdata to improve program administration.

Takeaway for postsecondary education:

Multiple government agencies use a variety of setups to securely link and analyze data, supporting access to highly sensitive and highly curated data in safe ways. They have implemented protocols permitting linkages to auxiliary data, access to personally identifiable information, and approaches to protect privacy. Similar linkages and protocols could make postsecondary education data more available for productive analyses.

DATA INTERMEDIARIES TO SUPPORT SAFE AND SECURE DATA

Linkages

Many analyses require data to be linked over multiple time periods or across different datasets. Person-level linkages require access to complete and accurate personally identifiable information. Some linkages can be done by exact matching, as when records with the same exact Social Security Numbers are linked. Alternatively, probabilistic or fuzzy matching is based on the similarity of information, such as name and date of birth, between files. Linkage units are sometimes called Trusted Third Parties that handle restricted information in order to create linkage keys for analysts. Several agencies have data linkage units and will link files on a cost recovery basis.

University and non-profit partners help state and county agencies link program data for administrative and operational uses. The examples described below show how valuable such data linkages can be and demonstrate how intermediaries can support safe and secure data practices.

- **Case Western Reserve University's ChildHood Integrated Longitudinal Data system (CHILD)**[15] includes children ages 0-18 who live Cuyahoga County, Ohio, beginning with the 1992 birth cohort. CHILD links data from many sources including birth certificates, publicly subsidized child care, home visiting and early intervention, child abuse and neglect investigations, child welfare placements, juvenile justice filings, Temporary Assistance for Needy Families, Supplemental Nutrition Assistance Program, Medicaid participation, and public-school student records. CHILD data analyses support planning and decision making in the County.
- Similarly, the **Silicon Valley Regional Data Trust (SVRDT)** was established through a partnership between the Santa Clara County Office of Education and the University of California, Santa Cruz. The trust includes data from three northern California counties: San Mateo, Santa Clara, and Santa Cruz. SVRDT curates data from the 66 school districts, juvenile probation, and Health and Human Service agencies in the three counties. With more real-time and current data feeds, the SVRDT can be used administratively and for interventions.
- At the University of Chicago, the **Kilts Center** hosts research on marketing data including Nielsen consumer panel data and scanner data. Subscribers at qualifying institutions can download data subject to data security provisions. Data are de-identified and may only be used for research.
- **Institute for Research on Innovation and Science (IRIS)** at the University of Michigan curates administrative data from academic institutions to measure the return on investment for scientific grants. IRIS demonstrates how a data intermediary can standardize the receipt and harmonization of data from many institutions, link across multiple data systems (e.g., federal grant awards, data on participating faculty and students, patents, and research publications), and control safe research access to the data. IRIS produces value for participating institutions by measuring outcomes resulting from federal grants. IRIS also looks at the outcomes for researchers and students who worked on the grants. Two of its partner universities have piloted an earnings data match with the Census Bureau to assess income levels one, five and 10 years after graduation. IRIS has 35 academic partners to date.
- The **Observational Health Data Sciences and Informatics (OHDSI),** pronounced "odyssey," researches collaborative acts as an intermediary for academia, government, and industry. OHDSI is a non-profit coordinated out of Columbia University, and it supports research in statistics, epidemiology, informatics, and the clinical sciences across twenty counties worldwide. OHDSI uses the Observational Medical Outcomes Partnership (OMOP) common data model that enforces a standard vocabulary on large scale health data, such as claims, electronic health records, and data from registries and longitudinal surveys. Data providers make their data fit what OHDSI requires, making all their encounters, terms, and codes standardized and comparable across sources. This lets users develop standard analytic routines that are effective across providers in the United States and abroad.[16] The approach is "specialized but extendable" and serves as one of many collaborative approaches to federating confidential data in the healthcare domain.[17]
- The **SkyServer** is an astronomical database with a web interface to the Sloan Digital Sky Survey (SDSS), a project "to make a map of the entire universe."[18] The project transforms about 40 terabytes of data read from a telescope in New Mexico into 3 terabytes of processed data. Researchers considered how this overwhelming volume of data could be used effectively and defined 20 typical queries, building the SkyServer

to respond to those queries. A visual query tool allows both experts and novices to explore the universe. More than a billion images and spectra from different celestial objects support queries on galaxies, quasars, and how the universe is expanding.

- The non-profit **Private Capital Research Institute (PCRI)** gathers and standardizes data on private capital activity with global firms and transactions. Data are de-identified by PCRI and hosted for research through NORC at the University of Chicago. The virtual access through the NORC virtual enclave does not allow raw data to be downloaded. NORC has secure physical workspaces that prevent unauthorized access or removal of data, as well as virtual enclaves[19] that allow users to authenticate and interact within a secure network to conduct analyses.

Takeaway for postsecondary education:

These examples demonstrate how postsecondary institutions and the federal government could make better use of data in secure ways. Across industries, these intermediaries are supporting research access through data standardization, linkages, and secure data hosting. University-based and non-profit research and data intermediaries currently process large volumes of data, including confidential student-level data, and have been able to address complex governance and security issues. While complex, examples like these prove that data access challenges are surmountable in other sectors and can be addressed in higher education as well.

DEVELOPING THE FUTURE DATA SHARING ECOSYSTEM

Data preparation

To make comparisons across many institutions, data will need to be clean, standardized, and well-documented. Think of the data infrastructure as a brickyard: a place where people go to obtain materials they can expect to be regularized with predictable dimensions and formats; business processes, documentation, and permissions control the removal of any materials.[20] For a postsecondary education data infrastructure, the field (e.g., agencies and institutions) will need to participate in a common data model, shaped to meet the needs of the user community. A common data model will describe required data elements for the system, required formatting, what valid records look like, valid categories and data labels, and more.

Processing the data to ensure compatibility is an arduous but necessary step. There are thousands of higher education institutions, all with varying degrees of readiness to engage with a more robust data infrastructure. Government agencies, some of which have disjointed, legacy systems, may have readiness issues as well. Sufficient planning time will allow institutions and government agencies to adapt to a new data infrastructure.

Researchers are working on pilots and implementations of technologies that enhance privacy, such as *secure multiparty computation (SMC)* and differential privacy. SMC will allow researchers to analyze data without revealing private information.[21] For example, Estonia used SMC to compute earnings outcomes for Estonian students using only encrypted streams of data. The earnings data remained at the country's revenue agency and the student data remained at each institution, with all computation occurring within an application called Sharemind. Estonian officials determined that no personal data were involved in the computation, since only encrypted data were used to assess outcomes.[22]

The pilot demonstrated that SMC was feasible, but the processing took a long time. With less than one million student records and 10.4 million tax records, the processing took over sixteen days. A recent SMC pilot in Allegheny County, Pennsylvania,[23] used data on homeless and mental health services, causes and incidences of mortality, family interventions, and incarceration across five data systems to address policy questions. This processing on the county's health and human services Integrated Data Systems (IDS) also involved long run times. SMC works, but more research and development is needed to make it scalable and efficient.[24] It is not yet ready to scale for a postsecondary data system that includes thousands of institutions and millions of students.

Differential privacy aims to tell you as much as possible about a group while telling you as little as possible about any individual in the group. It is an evolving field in Statistics and Computer Science that protects personal data by adding a little bit of "noise" in the dataset. In other words, differential privacy slightly alters data to protect individual's privacy, striking a balance between data accuracy and privacy. In one case, healthcare records were federated across many hospitals, allowing each hospital to explore the combined set while protecting the identities of the individual patients.[25] The Census Bureau will apply this method to 2020 Census data releases, and Google, Apple, Microsoft, Facebook, and others are also involved in differential privacy research and implementations.

Takeaway for postsecondary education:

Experts are developing new and innovative privacy enhancing techniques and tools, which could be useful for postsecondary education data systems in the future. The field of data privacy and security needs more investment in research to test how the techniques can scale and be layered.[26] While these emerging techniques are not yet ready to be scaled, the field of higher education should monitor demonstration projects across disciplines and domains and evaluate the expected improvements in runtime, implementation costs, and privacy-preserving effectiveness.

APPLICATIONS TO THE POSTSECONDARY DATA INFRASTRUCTURE

Today, data infrastructures exist that handle sensitive and high-volume data across domains, including healthcare, human services, housing, and workforce. Parallel characteristics between postsecondary education and healthcare are especially notable. Both have individuals using lots of providers, services that are hard to pay for (often with hidden or misunderstood fees), the unclear value of one provider versus another, possible limits on enrollment, and the exclusion of some people from the system. In addition, it is difficult to obtain data on social determinants and outcomes. Answering important questions in both postsecondary education and healthcare requires access to highly personal information from multiple places.

Data infrastructures in healthcare, such as OHDSI, rely on a robust data model managed by a trusted intermediary that handles the data and paperwork flows. The same can be done for a postsecondary education infrastructure. It could rely on an intermediary, such as NCES, to ensure that data are standardized, accessed, and used securely and responsibly in an enclave meeting federal data protection protocols.

Roles for a Postsecondary Infrastructure Intermediary

An intermediary, such as NCES, will be necessary to establish and implement a more robust postsecondary data infrastructure. That intermediary would hold a variety of responsibilities, including to:

- Identify and pursue data sources, sponsoring data collection where necessary.
- Manage agreements after negotiation.
- Enforce negotiated terms of use.
- Ingest and harmonize data.
- Regularly assess the adequacy of the common data model.
- Set norms for private sector data use.
- Act as a trusted third party to link data.
- Coordinate screening, training, and monitoring of researchers.
- Coordinate output review.
- Gather tools and models that make analysis more efficient.
- Provide technical assistance to providers for data harmonization.

When strengthened, the postsecondary data infrastructure could improve how students and parents view institutions and programs, giving them better information when making decisions. The infrastructure could also facilitate new channels of discovery, enabling data joins and cross-school, cross-cohort, and longitudinal analyses that measure student outcomes to see what works, ultimately improving outcomes for students.

Practices and tools exist today to handle postsecondary data access, analysis, and analytic results securely and responsibly. Indeed, those practices are in place at many federal agencies, allowing data to be shared and used in safe ways. Higher education does not need to wait for new and emerging techniques like smart contracts, differential privacy, or SMC. As these tools emerge, they can be incorporated into the postsecondary data infrastructures to make them even more secure and robust. But, in the meantime, current tools and practices can and should be used to provide answers to critical questions facing today's students, families, policymakers, and educators.

APPENDIX A: KEY TERMS IN DATA INFRASTRUCTURE AND GOVERNANCE

Authentication
Verifying that a user is who they claim to be by requiring something they know (passwords, answers to personal questions), something they have (token or key), and/or something they are (biometrics)

Common data model
Information model that accommodates data from many sources by standardizing structure, content, and analytics; should be manageable for data providers and useful for data users; is extensible across domains and evolving based on new data and needs

Data intermediary
Entities that facilitate sharing and access of restricted data; handle multiple sources and multiple users; and have standard request and review processes

Differential privacy
A concept of privacy that hides the effect of each individual in a dataset, trading accuracy for privacy by injecting a small amount of noise dependent on the data and query being run

Enclave
A secure network that lets approved users access restricted data at a specific site (physical enclave) or through credentialed remote access (virtual enclave)

Encrypted data
Data translated into an unreadable form that can only be read by people with a secret key or password; can be encrypted in transit (being sent between institutions) and at rest (inactive copies stored at an institution)

Federated data system
Data management approach where data remain siloed (e.g., at each institution) but have been harmonized to permit queries across the federation or network; distinct from approaches where data are aggregated in a centralized system (like a data lake, where data retain their source format, or a data mart/hub, where data are harmonized prior to entry)

Governance
The combination of people, processes, and information technology controlling data access and use; clarifies ownership, security and risk management, and compliance reporting

Harmonization
Combination of steps that make data consistent and comparable over time, across programs, or across systems; can include parsing, standardizing, and recoding data

Inputs
Source data entering a data infrastructure whose use, storage, and security requirements are specified in a data use agreement or contract; should include metadata that describe the files, variables, and categories

Linkage
Exact or probabilistic matching that connects data about the same entity, such as a student, institution, or loan; also called entity resolution, record linkage, and data fusion

Metadata
Information about the files, sources, and datasets; can include file layouts, variable descriptions, data dictionaries, value ranges, and notes that aid understanding and appropriate use of data

Noise
Alterations to actual data to protect against reidentification; typically applies to continuous variables by adding or multiplying a randomized number to the original values

Outputs
Products of data analysis and manipulation that include aggregate, statistical, and microdata extracts; most data infrastructures require output review before data leave the computational environment or are publicly released

Parsing
Separating data stings into variables (e.g., "Jane S. Doe" into firstname = Jane, middleinitial = S, lastname=Doe)

Provisioning
Sharing approved data with approved analysts by allowing remote access or securely delivering extracts, in a manner that meets security requirements established by the data provider (e.g., in a virtual or physical enclave)

Recoding
Transforming data into consistent categories (e.g., putting January 31, 2000, 1-31-00, 00-Jan31, 31-01-00 into a MM/DD/YYYY format so they all coded 01/31/2000)

Restricted data
Data that must be protected using the highest level of security, driven by legal or regulatory requirements; have penalties for misuse; are "notice-triggering," meaning that an authority is contacted if access is unauthorized; are distinct from confidential (sensitive) data, which must be controlled according to data provider conditions, and open (non-sensitive) data, which are publicly available

Schema
Documentation of the contents and relationships between files and variables used to manage databases

Secure multiparty computation
Method for two or more parties to jointly compute a function (e.g., determine match/non-match, compute averages or medians, produce regression coefficients) on their inputs using a protocol without revealing anything about the parties' inputs; different than a trusted third-party approach where an intermediary facilitates interactions between the parties, computing the functions on copies of data shared with them

Smart contract
Computer code that automatically transfers or allows access to data when a pre-defined set of terms and conditions are met; relies on distributed ledger technology to record all transactions

Spill
An unintentional data release, also called a breach or a leak

Standardizing
Applying common formats to data (e.g., "123 West Spring Parkway" into 123 W SPRING PKWY); must be sensitive to misspelling, truncation, and language issues

Synthetic data
Simulated data from statistical models that can be made available because they represent imaginary individuals; usually accompanied by a validation server so that results from the synthetic data can be compared to results on the real data to validate findings

Trusted third party (TTP)
A neutral intermediary that receives and protects restricted data; has no stake in the source data or any pending analyses; may conduct linkages, create data extracts, or compute statistics for data providers or analysts

ENDNOTES

[1] For example, the Department of Human Services in Allegheny County, Pennsylvania, hosts analysts in their secure integrated data system.

[2] Private sector firms are determining how to leverage and link their proprietary data, for instance, JP Morgan Chase Institute uses their consumer finance and business data to produce timely analyses to inform policy debates.

[3] For more information on each dimension and implementation issues in the UnitedKingdom, see Desai, Tanvi, Felix Ritchie, and Richard Welpton. "Five Safes: Designing Data Access for Research." University of the West of England, Economics Working Paper Series 1601, Bristol. 2016.http://eprints.uwe.ac.uk/28124/1/1601.pdf

[4] DUA sections typically include: legal authorities, costs/resources, ownership, transmission and storage, access, approval, roles and responsibilities, training, analysis, dissemination, and timelines. A number of guides, checklists, and tools are available to support DUA development. See Petrila, John, Amy O'Hara, and Julie Williamson. "The Feasibility of an Electronic Repository of Data Use and Related Agreements." Report for the Laura and John Arnold Foundation.

[5] See Census Bureau Administrative Data Inventory for additional examples at https://www2.census.gov/about/linkage/data-file-inventory.pdf

[6] See Altman, Micah. "Advancing Robot Lawyers: Towards Interoperable Protected Research Data." https://privacytools.seas.harvard.edu/files/privacytools/files/micah_altman_slides.pdf. 2014. See Karafili, Erisa, and Emil C. Lupu. "Enabling Data Sharing in Contextual Environments: Policy Representation and Analysis." In Proceedings of the 22nd ACM on Symposium on Access Control Models and Technologies - SACMAT '17 Abstracts, 231–38. Indianapolis, Indiana, USA: ACM Press, 2017. https://doi.org/10.1145/3078861.3078876.

[7] For information on examples involving FERPA, see Nissim, Kobbi, Aaron Bembenek, Alexandra Wood, Mark Bun, and Urs Gasser. "Bridging the Gap Between Computer Science and Legal Approaches to Privacy" 31 (n.d.): 95. For information on examples involving employer-employee data, see Haney, Samuel, Ashwin Machanavajjhala, John M. Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber. "Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics." In Proceedings of the 2017 ACM International Conference on Management of Data, 1339–1354. SIGMOD '17. New York, NY, USA: ACM, 2017. https://doi.org/10.1145/3035918.3035940

[8] For a comprehensive discussion of possible methods, see Goroff, Daniel, Jules Polonetsky, and Omer Tene. "Privacy protective research: facilitating ethically responsible access to administrative data." Ann. Am. Acad. Polit. Soc. Sci. 675(1), 46–66. 2018

[9] Additional information on how the French secure data hub uses SD-Box can be found at https://www.casd.eu/en/technologie/sd-box/

[10] The Office of Personnel Management recently deployed a synthetic data analytic environment. See Barrientos, Andrés F., Alexander Bolton, Tom Balmat, Jerome P. Reiter, John M. de Figueiredo, Ashwin Machanavajjhala, Yan Chen, Charley Kneifel, and Mark DeLong. "Providing Access to Confidential Research Data through Synthesis and Verification: An Application to Data on Employees of the U.S. Federal Government." The Annals of Applied Statistics 12, no. 2 (June 2018): 1124–56. https://doi.org/10.1214/18-AOAS1194. Also,Facebook is developing synthetic data for the Social Science One project (https://socialscience.one/blog/update-social-science-one)

[11] See Kifer, Daniel, and Ashwin Machanavajjhala. "Pufferfish: A Framework for Mathematical Privacy Definitions." ACM Transactions on Database Systems 39, no. 1 (January 6, 2014): 1–36. https://doi.org/10.1145/2514689

[12] See Altman, Micah, Alexandra Wood, David R. O'Brien, and Urs Gasser. "Practical Approaches to Big Data Privacy over Time." International Data Privacy Law 8, no. 1 (February 1, 2018): 29–51. https://doi.org/10.1093/idpl/ipx027

[13] For further information, see https://lehd.ces.census.gov/data/pseo_beta_viz.html

[14] For example, access to tax data through the IRS Joint Statistical Research Program formed the basis for Raj Chetty's Opportunity Insights project (https://opportunityinsights.org/)

[15] Description of system governance and partners can be found at: https://case.edu/socialwork/povertycenter/sites/case.edu.povertycenter/files/2018-09/An-IDS-Resource-for-Cuyahoga-County-12-8.pdf

[16] For example, OMOP creates uniform units of analysis with industry standard measurement units, but also allows unstructured free text for certain fields. See Belenkaya, Rimma, Karthik Natarajan, Mark Velez, and Erica Voss. "OMOP Common Data Model (CDM) & Extract-Transform-Load (ETL) Tutorial," n.d., 152

[17] Healthcare has a massive head start: their models are more mature because of long-time large-scale infrastructure investments by the federal government through National Institutes of Health funding.

[18] See http://skyserver.sdss.org/edr/en/sdss/skyserver/

[19] The Inter-University Consortium for Political and Social Research (ICPSR) also has physical and virtual enclaves to support academic research

[20] See Foster, Ian. "Research Infrastructure for the Safe Analysis of Sensitive Data." The ANNALS of the American Academy of Political and Social Science 675, no. 1 (January 1, 2018): 102–20. https://doi.org/10.1177/0002716217742610

[21] Bater, Johes, Gregory Elliott, Craig Eggen, Satyender Goel, Abel Kho, and Jennie Rogers. "SMCQL: Secure Querying for Federated Databases." eprint arXiv:1606.06808. 2016.https://ui.adsabs.harvard.edu/\#abs/2016arXiv160606808B

[22] For details, see Bogdanov, Dan, Liina Kamm, Baldur Kubo, Reimo Rebane, Ville Sokk and Riivo Talviste. "Students and Taxes: A Privacy-Preserving Study Using Secure Computation." PoPETs 2016 (2016): 117-135

[23] The report is the first to test SMC on human services administrative data. It is located at https://bipartisanpolicy.org/wp-content/uploads/2019/03/Privacy-Preserved-Data-Sharing-for-Evidence-Based-Policy-Decisions.pdf. See Hart, Nicholas R., Erin Dalton, and David Archer. "Privacy-Preserved Data Sharing for Evidence-Based Policy Decisions: A Demonstration Project Using Human Services Administrative Records for Evidence-Building Activities." Bipartisan Policy Center Technical Report. March 2019.

[24] The Intelligence Advanced Research Projects Activity (IARPA) has an active research program called Homomorphic Encryption Computing Techniques with Overhead Reduction (HECTOR) that aims to develop tools like SMC and other secure distributed applications.

[25] For details, see Bater, Johes, Xi He, William Ehrich, Ashwin Machanavajjhala, and Jennie Rogers. "Shrinkwrap: Differentially-Private Query Processing in Private Data Federations." ArXiv:1810.01816 [Cs], October 3, 2018. http://arxiv.org/abs/1810.01816

[26] For example, a federated data system across university systems using SMC (i.e., no identifying information leaves the university), where output data on earnings per program by campus are differentially private

IHEP
INSTITUTE FOR HIGHER EDUCATION POLICY

Protecting Students, Advancing Data: A Series on Data Privacy and Security in Higher Education is a project of the Institute for Higher Education Policy. This report was produced with support from Arnold Ventures. The views expressed in this report are solely those of the authors.