International Journal of Population Data Science

Journal Website: www.ijpds.org

Research Ready Data Lakes: Protecting Privacy in Relatable Datasets

McMillan, R^{1*} and Reeves, M¹

¹Georgia State University, Andrew Young School of Policy Studies, Georgia Policy Labs

Background with rationale

The Georgia Policy Labs' mission is to improve outcomes for children and families by producing rigorous research with long-term government partners. A key component of this model is having secure access to research-ready, individual level data from multiple sources to answer government agencies' questions within policy windows. Obtaining sensitive data from our partners requires significant relationship building, demonstrations of value, and assurances of our ability to mitigate all security and privacy concerns.

Objectives

1. Securely transfer and de-identify disparate individual level datasets with personally identifiable information from public entities.
2. Clean data and store in a pristine data lake, made available for fast turn-around research.
3. Ensure individual records can be matched across disparate organizations' datasets.

Approach

Our practices, infrastructure, data sharing agreements and security are built to support the intersection of data availability for researchers and security standards that give our partners ease. We highlight two solutions addressing security concerns while supporting our researchers, which can be used by other researchers using sensitive data. First, we discuss our multiple tiers of transfer and access that remove risk from identifiable data. Second, we share the double hash solution created for a partner who was not willing to share PII. We share the source code for our SHA3-512 double hash solution, which allows for matching of records across disparate datasets without receiving PII sensitive elements.

*Corresponding Author:

Email Address: rmcmillan@gsu.edu (R McMillan)

Results

We created reliable matching values without the need for the actual social security numbers or other PII values on our side, enabling a large school district to share its student-level data with us.

Conclusion

The balance of security and easy access for researchers is a common area of friction. Our security set-up and hashing solution allows others to remove this barrier for applied policy research.