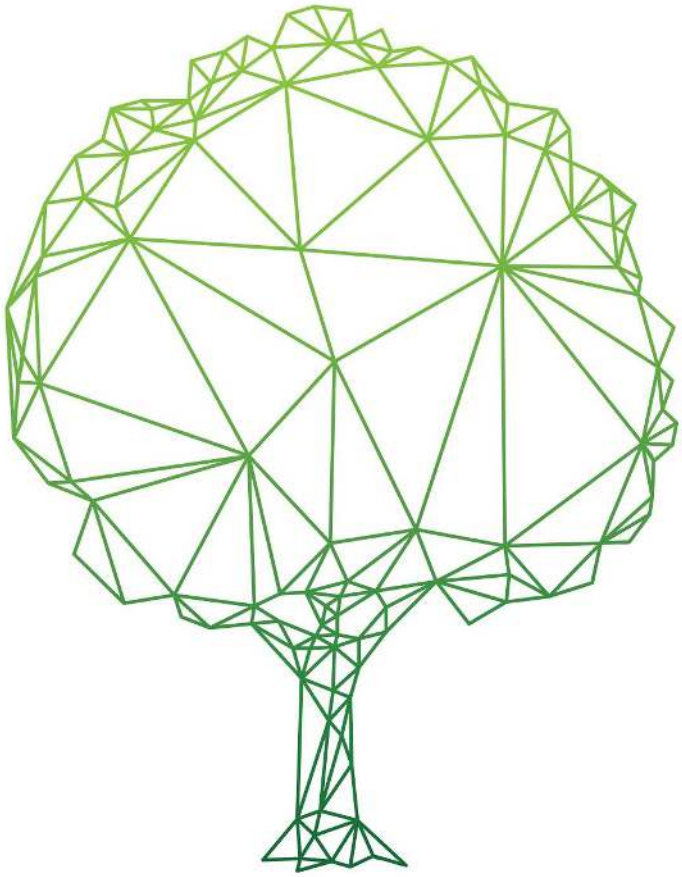
DETERMINING THE SHARED POPULATION BETWEEN SERVICE PROVIDERS

# How Tulsa Is Preserving Privacy and Sharing Data for *Social Good*



asemio

**Authors**

Aaron Bean, Asemio, LLC

Rev. Jeff Jaynes, Restore Hope Ministries

TJ Sexton, CAP Tulsa

7/15/2019

asemio

# Abstract

Data sharing between organizations addressing social risk factors has the potential to amplify impact by increasing direct service capacity and efficiency. Unfortunately, the risks of and restrictions on sharing personal data often limit this potential, and adherence to regulations such as HIPAA and FERPA can make data sharing a significant challenge. Through the development of new technologies such as secure multi-party computation (MPC), communities have an opportunity to share individual-level data without exposing personally identifiable information (PII). In this white paper, we describe the process of using MPC technology to answer questions that can aid service providers in exploring the barriers that underserved populations may be facing. The first question we asked: what is the overlap of populations served by two distinct organizations? The results of the overlap analysis confirmed that a significant opportunity exists to increase access to services for a subset of individuals through better outreach.

Lessons learned in this project are already directly informing the technical work of other initiatives in the Tulsa community. Included in this white paper is an example of using MPC for shared data analysis, with lessons and insights that can be applied to other collaboratives and communities.

# Problem Statement

All communities grapple with complex social problems that impose detrimental effects on the well-being of their populations. One shared goal among problem-solving organizations is to improve outcomes for affected populations by increasing interoperability, refining resource alignment, and streamlining community services. Access to data is often looked to as a panacea for achieving these systems-level advancements. Data has become a central component of almost every community-focused strategy, often acting as the linchpin of outcomes-improvement work, systems-level change initiatives, pay-for-success projects, and social impact strategies.[1]

[1] Baker, M., Baldwin, M., Bean, A., Davis, P. Stout, M. Vanderlip, E. R. (2019). Architecting resilient and adaptive communities through technological innovation. Asemio.

However, the tension between the need to obtain private data for increased efficacy of community-level analytics and the need to protect the identity of vulnerable populations continues to grow. Community data sharing initiatives typically involve expensive governance cycles and intricate legal frameworks that are designed to control risk.[2] The sharing of certain types of personal data is governed by laws and regulations such as FERPA[3] and HIPAA[4], which necessarily prioritize protecting sensitive data over enabling data sharing for community-level benefit. Organizations serving vulnerable populations are particularly wary of privacy breaches, which not only violate federal law but also transgress the ethical mandate of protecting the privacy and trust of those populations. Leaders are understandably cautious about investing time, energy, money, and trust in data sharing initiatives when the risks often appear to outweigh the potential gains.

There is a clear need to establish a model that can serve our communities better by enabling community analysis of integrated data more quickly, at a lower cost, and in a manner that enhances privacy and security protection for individuals contributing, and organizations collecting, sensitive personal data.

# Solution

## High-Level Solution

To help address the need for faster, more secure analysis of integrated data, we evaluated modern cryptosystems to determine if they could offer a viable solution. Through this research, we identified advances in cryptography (e.g., homomorphic encryption, secure multi-party computation) that would allow for the creation of new technologies that meet our needs by protecting input values while permitting the sharing of aggregate results.

[2] Allen, C. (2014). *Data Governance and Data Sharing Agreements for Community-Wide Health Information Exchange: Lessons from the Beacon Communities.* Washington, DC: EGEMS.
[3] The Family Educational Rights and Privacy Act. 20 U.S.C. § 1232g; 34 CFR Part 99.
[4] The Health Insurance Portability and Accountability Act of 1996. Pub. L. 104-191. Stat. 1936.

MPC can be described as "a subfield of cryptography with the goal of creating methods for parties to jointly compute a function over their inputs while keeping those inputs private."[5] In other words, MPC facilitates the encryption of data from disparate source systems to be shared, analyzed, and returned as aggregated results while maintaining the privacy of PII and other forms of sensitive information. One formative MPC case used the technology to analyze gender and ethnicity wage gaps among employers within the greater Boston area.[6] The researchers collected individuals' compensation data from privately held companies in order to calculate an aggregate statistic (a sum, the computed function) over the data points. Researchers could view the employee earnings totals aggregated across all companies, but the individual company aggregates remained private and were never revealed to any single party.

In the case of service providers, inputs are typically demographic or service information related to individuals. The organizations that collect this data are legally liable for its privacy. The computed functions that organizations could utilize include measures related to individuals' data—for example, average age or total counts of shared records, such as the overlap in the number of children between service providers. By using MPC technologies, social service organizations may discover answers to long-standing questions about their served population (e.g., what other services they are receiving).

## Solution Details

### VALUE PROPOSITION

This project sought to test the viability of using MPC technology to provide community-level population analysis to local nonprofit stakeholders, who envision using the results to effect social change and increase service delivery across otherwise distinct sectors. Our aim was to provide a meaningful answer to one question that we found pertinent to various nonprofit organizations in the community: what is the overlap of populations served by two disparate organizations?

[5] Secure Multi-Party Computation. (n.d.). In Wikipedia: The Free Encyclopedia. Retrieved November 9, 2018, from https://www.en.wikipedia.org/wiki/Secure_multi-party_computation
[6] Lapets, A., Volgushev, N., Bestavros, A., Jansen, F., Varia, M. (2016). *Secure Multi-Party Computation for Analytics Deployed as a Lightweight Web Application.* Technical Report BU-CS-TR 2016-008. Boston University: Computer Science Department.

To achieve this goal, we connected two prominent social impact organizations based in Tulsa that had not previously collaborated, AssistOK and CAP Tulsa. We sought to determine the various ways in which knowing the population overlap between them could increase their ability to provide services to their clients. AssistOK is a consortium of basic needs organizations working within an existing shared data system. CAP Tulsa is a two-generation antipoverty nonprofit and early childhood education provider focused on breaking the cycle of poverty. Both partners are subject to different regulations regarding the privacy and security of their clients' information; CAP Tulsa is FERPA and HIPAA regulated, while AssistOK is not regulated by common privacy and security regulations (e.g., HIPAA, FERPA, or 42 CFR part 2).

## HYPOTHESIS STATEMENT

Together, AssistOK and CAP Tulsa decided to approach this overlap question by focusing on the most vulnerable subset of their respective populations, children. This approach led to the formation of our pilot's hypothesis: There exists a subpopulation of children who have received services from an AssistOK organization but who are not enrolled in CAP Tulsa's early childhood education program. There are many reasons why an individual or a family may seek services with one organization and not the other, but one of the greatest obstacles encountered by vulnerable populations is obtaining information regarding needed services. If disparate service providers can identify the gaps that exist between them, they can begin to research the causes of those gaps and work to improve outreach and increase access to their combined populations.

Answering the shared overlap question above serves the interests of both organizations and the interests of the greater community by allowing for better understanding of population-level gaps in service provision. For CAP Tulsa, this information can help to better target marketing efforts to reach more families with age-eligible children through AssistOK organizations. For AssistOK, the answer to this question could provide insight about additional populations who are likely to benefit from the services they offer, such as food and rent assistance.

DATA SHARING AGREEMENTS (DSAs) AND QUERY PARAMETERS

Existing partnerships between Asemio and CAP Tulsa as well as Asemio and AssistOK provided a strong foundation for expediting the DSA development process that was specific to the purposes of this project. As there was no passage of information shared directly between CAP Tulsa and AssistOK, it was deemed unnecessary for these organizations to enter into a DSA with one another.

To obtain the data we needed to test our pilot's hypothesis, we first queried CAP Tulsa's data sets to find all children under the age of 5 who were enrolled in CAP Tulsa's early childhood education program between 2/1/2017 and 1/31/2019. We then queried AssistOK's data sets to find children under the age of 5 who sought services at one of their sites during the same period. The data was evaluated by traditional plaintext matching as well as the privacy-preserving Sharemind[7] platform, an MPC tool used to perform secure cryptographic computations across disparate data sources. Record matching was completed using exact matching on the following variables for plaintext and MPC approaches: first name, last name, and date of birth.

COMMUNITY ANALYTICS AND MAPPING PORTAL

The aggregate results from the overlap analysis were then exported to a web portal, the Community Analytics Mapping Portal (CAMP), designed by Asemio, to visualize and further explore the data. In the portal, each data set, named by its organization, is given a node that connects to other nodes by a line, or edge. The size of the node is representative of the size of the data set, and the thickness of the edge indicates the size of the overlap as a proportion of the smaller data set (the numerator) to the larger data set (the denominator). Selecting an edge between two nodes allows the user to see a Venn diagram displaying the total record count overlap between the two data sets. An image of this portal can be found in **Figure 1.**

## Results

We encountered two types of results in this pilot. The first result was validation that the behavior and output of the privacy-preserving data ingestion and computation pipeline had the

[7]https://sharemind.cyber.ee/

same output as the traditional non-privacy-preserving evaluation pipeline. The second result was the percentage overlap of the identified subpopulations, to answer our hypothesis.

## RESULT ONE: VALIDATION

After using both MPC privacy-preserving data-sharing and traditional data-sharing methods on the same data sets, we were able to compare the outputs to determine how successful the MPC methods were at linking across records. The traditional analysis served as the control for this pilot. There were 4,133 unique records of children found in CAP's data set and 1,096 unique records of children found across AssistOK's data sets, which resulted from the aforementioned queries. Traditional analysis matched records in 65 instances between these two sources, which means that 65 unique children were enrolled in CAP Tulsa and also sought services at an AssistOK site. Comparatively, MPC analysis also matched 65 records between these two sources. These results demonstrated that MPC technology achieved adequate and accurate results in this pilot.

## RESULT TWO: ANSWERING THE HYPOTHESIS

The percentage of age-eligible children who were served by AssistOK but were not enrolled in CAP Tulsa was 94.07% (1,031 children). The results found using these techniques stress the strategic opportunities for compounding service delivery where eligibility and geography of both programs match or intersect. Even without knowing the identities of the individuals, CAP Tulsa and AssistOK can now work concertedly to educate their clients about the other organization's resources and find ways to minimize the burden of accessing them.

In this case, many families receiving services from AssistOK may live outside of CAP Tulsa's prescribed service area. Despite this potential, many children, their families, and other individuals receiving services from CAP Tulsa may be able to use services across the various AssistOK sites located in the Tulsa metro region. Conducting additional analysis to include the dates of service provision, geographical boundaries, and indirect service recipients may enable AssistOK and CAP Tulsa to narrow their focus to developing marketing and recruiting strategies that reach this subpopulation of eligible children when and where they are seeking services in the greatest volumes.

### LIMITATIONS

MPC is an innovative solution to privacy-preserved data sharing, but it is not without limitations. The computational processing time for MPC can increase substantially depending on the size of the data sets being compared, as well as the methods of comparison themselves. For this pilot, record linking was achieved by using the most simplistic exact matching parameters: first name, last name, and date of birth. Introducing alternative methods, such as fuzzy matching and string-edit distance calculations of record values, could increase the accuracy of the results, provide a greater number of matches, and minimize some pitfalls associated with human error, such as the misspelling of names.[8] Additionally, expanding match criteria to include addresses or phone numbers along with first name, last name, and date of birth, would allow for greater comparison across dissimilar data points.

However, many of these potential improvements come at the price of longer calculation periods. Moreover, while the probabilistic determination of matching records is completed in an encrypted state, it is not impervious to statistical attacks. Small data set sizes and other limiting factors may increase the chances of record-level identification. Thus, there is a tradeoff between the accuracy and value of the results and mitigating the threat of statistical attacks. Finally, other challenges may be less technical in nature but prove more difficult to overcome. MPC is not yet a familiar technology, and its complexity may make it difficult to explain to general audiences and achieve buy-in among stakeholders when there is a lack of trust between organizations.

# Conclusion

### PROJECT OUTCOMES

CAP Tulsa and AssistOK are now working together to host coordinated enrollment and outreach events early this fall, which means that we may see actionable outcomes within a year of the start of our pilot. Overall, this project yielded three major takeaways:

[8] Approximate string matching. (n.d.). In Wikipedia: The Free Encyclopedia. Retrieved from https://en.wikipedia.org/wiki/Approximate_string_matching

1. We were able to iterate and identify areas for improvement to our process by engaging with a small group of trusting, established partners who were willing to try innovative approaches to better understand the populations they serve.
2. Successful implementation and adoption in the social impact space relies on increasing the ability of service providers to improve the lives of their clients and underserved populations. This type of analysis should be approached by developing a deep understanding of the context in which service seekers interact with providers and the context in which providers operate, including how services are provided and accessed. Without this context it is impossible to draw accurate conclusions from the overlap results.
3. Leveraging MPC cannot be considered a magic bullet. To truly combat wicked social problems, there must be focus on both individual care coordination and population research and evaluation. We believe that a combination of MPC and more traditional data sharing pipelines could address both areas. Providing answers to population-level questions will in turn elicit interest among disparate service providers in reaching individuals in those populations directly. Because of the speed and security improvements that MPC provides, we believe that by undertaking initiatives with the methods proven in this pilot, participating partners will be able to more effectively weigh the pros and cons of entering into direct data sharing agreements with each other.

COMMUNITY IMPACT

An unexpected outcome of this pilot is that it has served as the catalyst for a series of discussions concerning governance and the potential to radically shift privacy and security conversations away from “what can we not do?” to “what *else* can we do?” The project team and additional partners are beginning to explore avenues for making MPC for social good commercially accessible and are evaluating the cost structure required to implement it.

Throughout this project, the participants and other interested parties in the community continuously posed new questions about the possibilities for MPC. Service providers and other stakeholders began to brainstorm use cases and test questions to explore in future pilots. After one brainstorming session, some stakeholders recognized an opportunity to partner with other

organizations in the next Tulsa-based MPC project and are currently securing commitments to begin the next iteration with newly identified data sets.

Community champions and organizations must consider that this technology requires significant technical expertise to implement and that garnering buy-in can be difficult. In the near future, privacy-preserving technology is best suited for communities that are willing to innovate and can be considered “bleeding edge.”

Understanding the potential and limitations of this technology for all involved partners is paramount to successful implementations of MPC for social good. Further, proper data provenance, standardization, and validation of results are imperative to accurately performing exploratory analysis of encrypted data sets. We strongly encourage other communities to assess MPC and other privacy-preserving technologies in order to better understand and address today’s complex social problems.

## Acknowledgements

This white paper was developed with support from Data Across Sectors for Health (DASH), a national program of the Robert Wood Johnson Foundation led by the Illinois Public Health Institute in partnership with the Michigan Public Health Institute. DASH aims to align health care, public health, and other sectors to systematically compile, share, and use data to understand factors that influence health and develop more effective interventions and policies.

DASH is a partner of All In: Data for Community Health, a learning network that provides a space for sharing resources like this one that help communities share data across and beyond traditional health care sectors. With a diverse learning collaborative of 150+ projects that is still growing, the All In offers many technical assistance and networking opportunities to communities across the country.

## Project Background

Restore Hope Ministries was funded by the DASH CIC-START program, which supports short-term activities that help local collaborations take meaningful steps toward planning or implementing multi-sector data systems. Through DASH CIC-START, Restore Hope Ministries worked with Asemio to apply analytics technology to analyze the overlap between individuals who require basic needs assistance (e.g. rent, food, utilities, etc.) and those whose children attend early childhood centers. Asemio developed this white paper to share lessons learned from their use of innovative technology that allows for analysis of personally identifiable information while preserving client privacy.

## References

Actionable Intelligence for Social Policy. (2017). *Towards State-of-the-Art IDS Technology and Data Security Solutions.* Philadelphia, PA: University of Pennsylvania. Retrieved from https://www.aisp.upenn.edu/wp-content/uploads/2016/07/Technology-Data-Security.pdf

Allen, C. (2014). *Data Governance and Data Sharing Agreements for Community-Wide Health Information Exchange: Lessons from the Beacon Communities.* Washington, DC: EGEMS.

Approximate string matching. (n.d.). In Wikipedia: The Free Encyclopedia. Retrieved from https://en.wikipedia.org/wiki/Approximate_string_matching

Baker, M., Baldwin, M., Bean, A., Davis, P., Stout, M., Vanderlip, E. R. (2019). *Architecting resilient and adaptive communities through technological innovation.* Tulsa, OK: Asemio. Retrieved from http://www.asemio.com/wp-content/uploads/2016/10/architectingcommunitieswhitepaperv3.pdf

Hart, N. R., Archer, D. W., Dalton, E. (2019). *Privacy-Preserved Data Sharing for Evidence-Based Policy Decisions: A Demonstration Project Using Human Services Administrative Records for Evidence-Building Activities.* Washington, D.C.: Bipartisan Policy Center. Retrieved from https://bipartisanpolicy.org/report/privacy-preserved-data-sharing-for-evidence-based-policy-decisions-a-demonstration-project-using-human-services-administrative-records-for-evidence-building-activities/

Lapets, A., Volgushev, N., Bestavros, A., Jansen, F., Varia, M. (2016). *Secure Multi-Party Computation for Analytics Deployed as a Lightweight Web Application.* Technical Report BU-CS-TR 2016-008. Boston, MA: Boston University Computer Science Department.

Secure Multi-Party Computation. (n.d.). In Wikipedia: The Free Encyclopedia. Retrieved November 9, 2018, from https://www.en.wikipedia.org/wiki/Secure_multi-party_computation

Sharemind. (2019). Customer profile: Asemio. Retrieved from https://sharemind.cyber.ee/customer-profile-asemio/

The Family Educational Rights and Privacy Act. 20 U.S.C. § 1232g; 34 CFR Part 99.

The Health Insurance Portability and Accountability Act of 1996. Pub. L. 104-191. Stat. 1936.

## Figure 1.

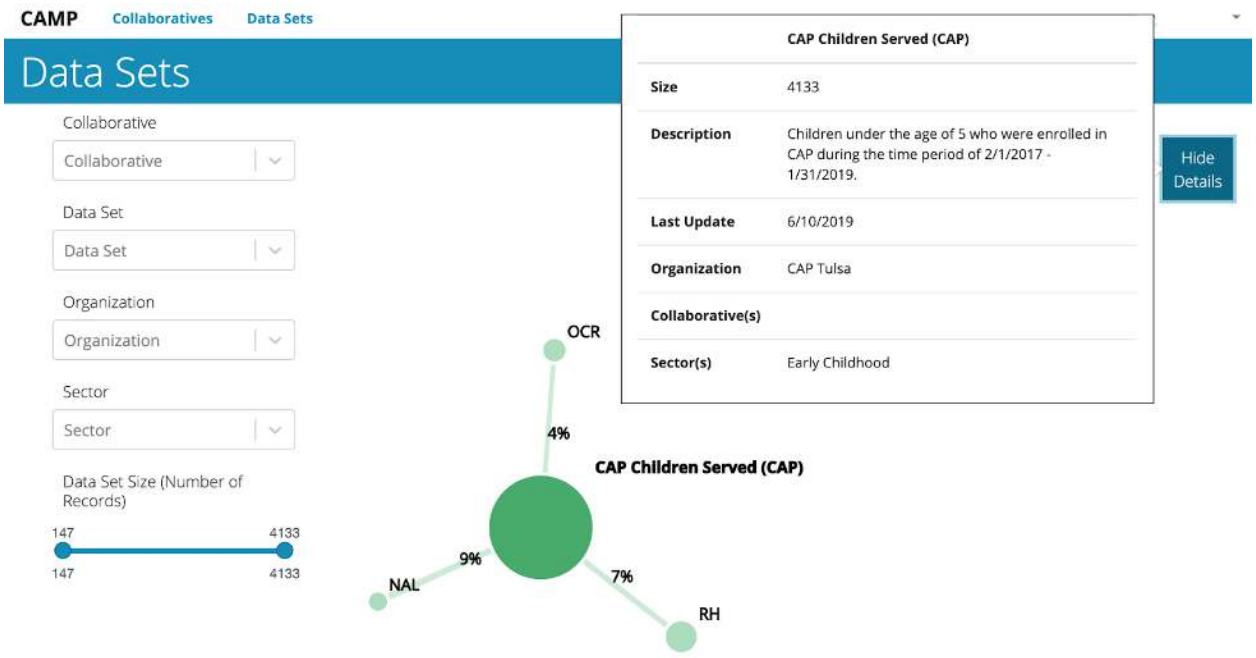



*Figure 1 shows the interactive exploration tool CAMP, created by Asemio for this project. The data sets page displays data sets as nodes, connected to each other by edges that represent the overlap between their populations. The size of the nodes is proportional to the size of the data set. The overlap percentage is calculated directionally, always using the smaller data set as a fraction of the larger data set. The details button shows additional information to describe and categorize the data sets.*