Privacy-Preserved Data Sharing for Evidence-Based Policy Decisions:

A Demonstration Project Using Human Services Administrative Records for Evidence-Building Activities

TECHNICAL PAPER

March 2019

BIPARTISAN POLICY CENTER

AUTHORS

Nicholas R. Hart, Ph.D.
Director, Evidence Project
Bipartisan Policy Center

David W. Archer, Ph.D.
Principal Scientist, Data Privacy and
Cryptography, Galois, Inc.

Erin Dalton, M.S.
Deputy Director, Office of Analytics,
Technology and Planning Allegheny County,
Pennsylvania, Department of Human Services

ACKNOWLEDGMENTS

The Bipartisan Policy Center thanks Schmidt Futures for their generous financial support of this project. BPC also appreciates the time and attention from the staff of the Allegheny County, Pennsylvania, Department of Human Services for their willingness to partner on the project in pursuit of improved privacy-preserving data-sharing capabilities. The authors also thank 10 anonymous reviewers of this paper for their valuable insights and contributions to the technical accuracy and clarity of the discussion presented, as well as numerous individuals who provided advice on developing a compelling and useful demonstration project.

DISCLAIMER

This technical paper is the product of BPC's Evidence Project and Galois, Inc. The findings and conclusions expressed by the authors do not necessarily reflect the views or opinions of BPC, its founders, its funders, or its board of directors.

Executive Summary

Emerging privacy-preserving technologies and approaches hold considerable promise for improving data privacy and confidentiality in the 21st century. At the same time, more information is becoming accessible to support evidence-based policymaking.

In 2017, the U.S. Commission on Evidence-Based Policymaking unanimously recommended that further attention be given to the deployment of privacy-preserving data-sharing applications. If these types of applications can be tested and scaled in the near-term, they could vastly improve insights about important policy problems by using disparate datasets. At the same time, the approaches could promote substantial gains in privacy for the American public.

There are numerous ways to engage in privacy-preserving data sharing. This paper primarily focuses on secure computation, which allows information to be accessed securely, guarantees privacy, and permits analysis without making private information available. Three key issues motivated the launch of a domestic secure computation demonstration project using real government-collected data:

- **Using new privacy-preserving approaches addresses pressing needs in society.** Current widely accepted approaches to managing privacy risks—like preventing the identification of individuals or organizations in public datasets—will become less effective over time. While there are many practices currently in use to keep government-collected data confidential, they do not often incorporate modern developments in computer science, mathematics, and statistics in a timely way. New approaches can enable researchers to combine datasets to improve the capability for insights, without being impeded by traditional concerns about bringing large, identifiable datasets together. In fact, if successful, traditional approaches to combining data for analysis may not be as necessary.
- **There are emerging technical applications to deploy certain privacy-preserving approaches in targeted settings.** These emerging procedures are increasingly enabling larger-scale testing of privacy-preserving approaches across a variety of policy domains, governmental jurisdictions, and agency settings to demonstrate the privacy guarantees that accompany data access and use.
- **Widespread adoption and use by public administrators will only follow meaningful and successful demonstration projects.** For example, secure computation approaches are complex and can be difficult to understand for those unfamiliar with their potential. Implementing new privacy-preserving approaches will require thoughtful attention to public policy implications, public opinions, legal restrictions, and other administrative limitations that vary by agency and governmental entity.

This project used real-world government data to illustrate the applicability of secure computation compared to the classic data infrastructure available to some local governments. The project took place in a domestic, non-intelligence setting to increase the salience of potential lessons for public agencies.

Data obtained under a confidentiality agreement from Allegheny County's Department of Human Services in Pennsylvania were analyzed to generate basic insights using privacy-preserving platforms. The analysis required merging more than 2 million records from five datasets owned by multiple government agencies in Allegheny County. Specifically, the demonstration relied on individual-level records about services to the homeless, mental health services, causes and incidences of mortality, family interventions, and incarceration to analyze four key questions about the proportion of: (1) people serving a sentence in jail who received publicly-funded mental health services; (2) parents involved in child welfare cases who received publicly-funded mental health services; (3) people serving a sentence in jail who received homelessness services; and (4) suicide victims who previously received publicly-funded mental health services. To BPC's knowledge, this demonstration is the first of its kind completed in the human services field.

To demonstrate and characterize applicability of privacy-preserving computation for these analyses, the project team performed them on two distinct privacy-preserving platforms. The first platform, called *Jana* and developed as part of the Brandeis program for the Defense Advanced Research Projects Agency, achieves secure computation entirely in software. Jana uses a combination of encryption techniques to protect data while at rest and in transit, and uses secure multiparty computation to protect data during computation. Specifically, Jana uses multiple servers to perform computation on *cryptographic secret shares* of data, while assuring that those servers never see the data in decrypted form.

The second platform, called *FIDES* and developed as part of the IMPACT program for the U.S. Department of Homeland Security, achieves secure computation via a hardware-enabled cryptographic enclave. Specifically, FIDES uses an Intel Corporation processor and the Intel Software Guard Extensions to compute in an area of the processor that is restricted from access by other code running on the computer, including the computer's own operating system. No part of the processor or software, aside from that hardware-secured enclave, ever sees the data in decrypted form.

These two privacy-preserving computation platforms offer similar approaches: data arrive at the computation platform already encrypted, analysis is performed in ways that strictly do not reveal anything about the data, and results are securely provided to users. The goal in these experiments was to compare these two approaches with a classic data analysis setting. Successful completion of the demonstration with human services data yielded the following insights:

- **The experiments produced valid, reliable results.** Both platforms generated valid results consistent with traditional data analysis approaches. This outcome suggests that the queries using these privacy-preserving approaches are not subject to diminished quality that would affect the validity or reliability of statistical conclusions. Therefore, multiparty computation models satisfy the demonstration's core criteria for enabling data use and privacy preservation.
- **The efficiency of the experiments presents a trade-off for policymakers.** Different modes of operationalizing the privacy-preserving technologies offer trade-offs for answer timeliness. Analyses with nearly 200,000 records using the software-based approach required nearly three hours to complete, whereas the same queries in the hardware-enabled environment returned results in one-tenth of a second. These times have substantial implications for applications in government operations with rapid decision-making architectures.

These findings suggest that these approaches offer considerable promise for public policy in achieving improved data analysis and tangible privacy protections at the same time. However, effort is still needed to further develop privacy-preserving technologies to make their deployment more time efficient prior to widespread use in government agencies. The scope and scale of such deployments will likely have either substantial cost implications or substantial delays in response times for computation, depending on the desired trade-off for the privacy-preserving approach. In addition to developing technical precision for privacy guarantees, further development of the technologies must also include learning about approaches for deploying the protections within complex organizational or governmental infrastructures and legal frameworks that may not explicitly encourage such activities.

This demonstration project offers a compelling example of how the technologies can be deployed—which can advance consideration of the approach within domestic, non-intelligence agencies at all levels of government.

Introduction

There are significant potential benefits for communities that construct statistical information about their constituents. For example, rigorous data analysis in medical practice can lead to new insights about improving health care and treatment options for vulnerable populations. The use of individual-level data to study addiction and drug treatment success can support efforts to reduce the consequences of the opioid epidemic and improve response times for emergency responders. The list of potential benefits for society is nearly endless.

But the use of individual-level records raises concerns about the risks of having private information being inappropriately accessed and used for unauthorized activities. While data can be responsibly used to generate gains for society without identifying individuals or organizations, the same data can also be used with a different intent to identify individuals, localize their whereabouts, and draw conclusions about behaviors, health, and political or social agendas. This information could then be used for illegitimate, nefarious, or irresponsible purposes. As more information is collected through public and private sources, these risks continually evolve.

For governments and organizations that pledge data will be protected or held in confidence, these evolving risks pose continuing challenges to any privacy guarantees made to the American public. This is particularly the case as a small set of attributes can single out an individual in a population, a small number of location data points can approximate where a person can be found at a given time, and simple analytics over such data can suggest other personal characteristics. Improper use of such localization, identification, and conclusions can result in financial, social, and physical harm to individuals.

While the risks continue to evolve, so too do approaches and frameworks for managing risks. Effectively managing risks is important considering the American public's expectation that information accomplish positive gains and uses in society. Public administrators have a defined need for using government-collected data to inform how they allocate services and benefits. Data subjects, the people who provide information in exchange for government services and benefits, are also afforded some expectations of privacy and confidentiality, including certain legal rights.[1]

Historically, the use of data and the promise of confidentiality were viewed as trade-offs, but emerging approaches increasingly present privacy guarantees that enable data use and privacy improvements simultaneously.

MOTIVATION FOR A MULTIPARTY COMPUTATION DEMONSTRATION

In 2017, the U.S. Commission on Evidence-Based Policymaking (Evidence Commission) issued its final report to Congress and the president, outlining an expectation that the federal government would use the data it collects to responsibly inform public policy decisions.[2] The commission comprised 15 members, including five individuals with experience in privacy issues who were appointed by elected leaders. After more than a year of study and input from federal agencies and the American public, the Evidence Commission offered unanimous suggestions for how to enable improved accessibility of government data while strengthening certain privacy protections.

The Evidence Commission rejected the idea that increased accessibility to data for statistical purposes must necessarily increase privacy risks. Instead, the commission proposed a series of recommendations that collectively formulated a strategy for ensuring a strong and transparent legal, administrative, and social framework to motivate improvements to government's infrastructure in coming years. When taken together, the Evidence Commission's 22 recommendations offer a compelling approach to achieving a larger vision in which "rigorous evidence is created efficiently, as a routine part of government operations, and used to construct effective public policy."[3] The commission acknowledged that this vision is challenged by declining response rates to surveys and barriers to collecting data for relevant analyses. Thus, the commission encouraged a more efficient use of existing administrative records—the data gathered by government programs and agencies through the course of normal operations.

★ ★ ★

"The Commission [on Evidence-Based Policymaking] believes that improved access to data under more privacy-protective conditions can lead to an increase in both the quantity and the quality of evidence to inform important program and policy decisions. ... Modern technology and statistical methods, combined with transparency and a strong legal framework, create the opportunity to use data for evidence building in ways that were not possible in the past."

—–THE PROMISE OF EVIDENCE-BASED POLICYMAKING[4]

The Evidence Commission's statutory task was to study evidence-building activities, referred to in the report as "statistical activities," which are activities that produce aggregate information about groups. Evidence-building activities lead to insights about groups of individuals or businesses without compromising the identities of any single person or entity. Due to the nature of Congress's charge, the commission prioritized activities to protect confidentiality, envisioning minimal risk of re-identification for data subjects. In the Evidence Commission's final recommendations, one issue was explicit: the federal government must do more to explore and deploy privacy-preserving and privacy-enhancing technologies to protect confidential data.[5]

One privacy-preserving approach—called "secure multiparty computation"—that the Evidence Commission considered and encouraged connects data in multiple locations using cryptography to guarantee privacy while permitting legitimate data analyses.[6] Take, for example, individuals who want to know their average wages without sharing their total wages with a peer. Rather than directly sharing the total wages, everyone shares a smaller piece of information with other individuals and a trusted external consultant.[7] Sharing these smaller pieces of information does not reveal total wages, but does reveal enough information to perform a statistical operation that recreates the total about the group of individuals without revealing sensitive information.[8]

Conceptually, the approach could be useful in many policy areas with sensitive data. A handful of successful tests of secure multiparty computation have been discussed publicly, including those using synthetic data,[9] private-sector wage information,[10] genetics,[11] and crop prices,[12] among others.[13] In the commission's research, however, no known existing demonstrations of the technology were identified using government data in domestic research or evidence-building capacities.[14] In addition, there are no known comparisons of multiparty computation to hardware-enabled secure computation that use such data.

Notably, few applications are present in the federal government. Intelligence agencies and the National Science Foundation have issued contracts, but multiparty computation has not yet emerged in other domestic agencies.[15] Several federal legislative efforts emerged in 2017 and 2018 that called for the incorporation of multiparty computation approaches, such as for educational analyses[16] and fungal-infection research in hospitals.[17] However, the limited presence of multiparty computation in federal legislation or contracts may suggest that public administrators view it as a signal about potential legal or cost implications and are simply waiting for additional explanations of the approach and validation that it achieves goals. In that context, additional demonstrations could motivate public administrators and policymakers to reconsider the approach.

In the spirit of the Evidence Commission's recommendation and recognizing the limited past demonstrations of the multiparty computation approach using government-collected data, this project is the first known of its kind in the human services field within the United States. This paper describes a project conducted by Galois, Inc., sponsored by the Bipartisan Policy Center, with support from the Allegheny County Department of Human Services (Pennsylvania). Beyond developing the capacity to implement the commission's recommendation, the motivation for the project is the confluence of two seemingly contradicting goals: (1) obtaining relevant and cost-effective data for evidence-based policymaking, and (2) assuring the privacy of the subjects' data.

The project specifically sought to (1) validate that multiparty computation can achieve equivalent statistical outcomes as those performed through a classic data analytic infrastructure, and (2) demonstrate the efficiency of multiparty computation techniques while maintaining data encryption. In summary, this demonstration project built and executed a prototype information system that used secure multiparty computation technology to enable useful analysis of individual-level data collected by numerous government agencies and routinely shared among agencies to generate statistics, all while preserving data confidentiality and minimizing the effort required to re-purpose those existing data for these new analytics.

In addition to providing context for the Evidence Commission's recommendation about privacy-preserving technologies and conditions for privacy protection, this technical paper also briefly discusses the limitations of existing approaches to confidentiality protection, the tangible benefit proposition of multiparty computation, and existing platforms for achieving the computation in a trust-based system. The demonstration project in Allegheny County and its results and implications will be discussed, as well as potential next steps for a technical research agenda. Building on the recommendation from the Evidence Commission, a final section outlines policy implications and considerations for further development in government agencies.

PUBLIC AGENCIES AND NEW PRIVACY TECHNIQUES

While privacy-preserving approaches hold substantial conceptual promise for privacy guarantees, in practice the applications have been somewhat limited to date. Research applications from the intelligence, computer science, and statistical communities can offer proof of privacy improvements, yet widespread adoption in government agency settings will require public administrators and policymakers to consider several key questions:

- **Satisfaction of Legal Obligations.** Public administrators and their legal advisors must be convinced that the approach satisfies legal obligations for privacy and confidentiality protections, which will require sufficient demonstrations using real data. Many of the legal and regulatory obligations demand absolute confidentiality.
- **Satisfaction of Stakeholder Privacy Expectations.** Even if legal obligations are achieved, the approaches must pass muster with the population the data describe. Appropriate transparency and accountability mechanisms that ensure responsible and appropriate uses of data may also bolster privacy expectations.[18]
- **Capacity to Develop and Maintain Platforms.** Currently, deployment requires expertise to establish platforms and computational protocols to execute analysis. Widespread adoption and scaling will be contingent on a workforce that has the ability to educate decision-makers in public agencies about the nuances of multiparty computation as well as the expertise to execute the projects.
- **Calibrating Decisions for Computation Time.** Even with sophisticated platforms, running an analysis can be computationally intensive and, in some programs, time-consuming relative to traditional methods. From a public administration perspective, using multiparty computation relative to existing approaches for database analysis, which can also produce near-instantaneous results, may make it challenging to justify the cost of implementation and may delay responsiveness—which, in turn, could affect the availability and timeliness of information for key policy decisions.
- **Justification of the Cost.** In weighing the adoption of privacy-preserving approaches, public administrators will need to consider how the potential privacy benefits relate to the actual costs of implementation.

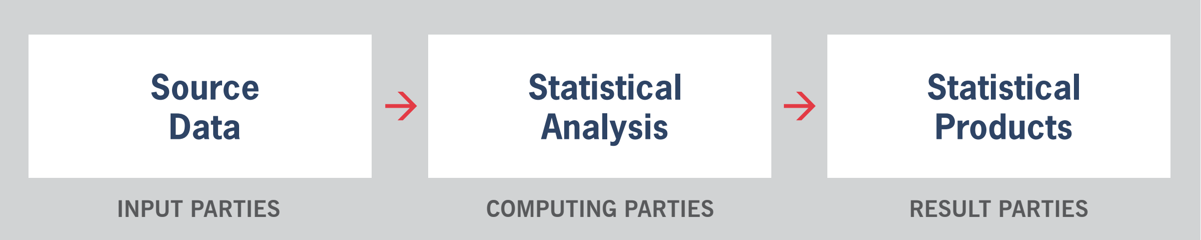
At a practical level, multiparty computation may not be feasible or practical for all government agencies or entities, and certainly not for all conceivable analytical queries. Agencies may lack the resources, expertise, or will to pursue the approach. Limitations to data quality, organization, or infrastructure may impede the capabilities of execution.

Finally, many jurisdictions exercise limited data-sharing or even analytical activities due to a lack of political will or motivation. Privacy preserving approaches, including multiparty computation, are not sufficient to overcome such a gap. However, the increased capabilities and approaches for engaging in data sharing could help overcome these obstacles.

CONTEXT FOR ADOPTING PRIVACY-PRESERVED DATA-SHARING APPROACHES

When it comes to data analysis, privacy protections can be applied at different stages of an analytical process. Indeed, typically multiple protections are used in combination to assure or maximize confidentiality guarantees. When considering these stages, a simplified conceptual model (see Figure 1) includes one or more *Input Parties*, who provide potentially sensitive data to one or more *Computing Parties*, who conduct statistical analysis. Their analysis is then provided to one or more *Result Parties*, who can use that information to inform decisions.

Figure 1: Simplified Conceptual Model of Analytical Stages Affecting Privacy Goals



At each stage of the conceptual model, privacy goals may vary slightly according to the party. Consequently, there are at least three goals for assuring data subjects' privacy: *Input Privacy*, *Output Privacy*, and *Access Control*.

- **Privacy Goal #1: Input Data Privacy** means that no Computing Party can access any input value provided by Input Parties, nor derive any such input from intermediate values available during processing, unless the value has been specifically selected in advance for disclosure.
- **Privacy Goal #2: Output Privacy** means that the results available to Result Parties do not contain, reveal, or allow derivation of identifiable input data beyond what is acceptable by Input Parties. In other words, they do not allow for identification of single records in the analysis.
- **Privacy Goal #3: Access Control** means that the system includes a mechanism for Input Parties to exercise positive control over which computations can be performed by Computing Parties on sensitive input data (thus affecting Input Privacy and Output Privacy) and which results can be published to Result Parties (thus affecting Output Privacy). Such positive control is typically expressed in a formal Access Control language that authenticates participants and the rules by which they participate in the system.

In practice, all three goals are often societally desirable for robust protections. The Commission on Evidence-Based Policymaking intended to address all three goals in its suite of findings and recommendations. While the commission made recommendations about approaches such as multiparty computation that focus primarily on Input Privacy, it also offered conclusions about Access Controls. For example, the commission viewed tiered access as one approach to data sensitivity and preventing the unauthorized use of data, thereby reducing the risk of harm to individuals.[19] Similarly, the commission described the need to conduct risk assessments for Output Privacy to guard against incomplete or inadequate disclosure-avoidance techniques, the approaches used to de-identify data.[20] Each of these approaches can also be compatible with implementation of Input Privacy and multiparty computation techniques. In fact, the prototypes used in this demonstration are jointly compatible and include Access Control capabilities.

To achieve trust, each of the privacy goals should be fulfilled and also accompanied by elements of transparency and accountability. Processes and systems can be put in place to maximize or guarantee Input Privacy, Output Privacy, and Access Controls. But, data subjects may need additional reassurance, including routine knowledge about the protocols and approaches for independent oversight to maintain trust that the privacy goals are achieved at desired levels.

Policymakers in the federal government have routinely reinforced the need for such trust. For example, the Foundations for Evidence-Based Policymaking Act of 2018 (P.L. 115-435) addresses core aspects of trust for statistical activities in the federal government by establishing tiered access systems based on data sensitivity (Access Controls), requiring certain agencies and their employees to take steps to limit certain uses of identifiable records (Input Privacy), and ensuring confidentiality of released information (Output Privacy). Taken together, the Evidence Act's provisions reflect policymakers' general desire for a coordinated approach to fulfilling multiple privacy goals rather than a deliberate intention to address the goals in isolation.

In recent years, some state and local agencies have also taken steps to build trust with constituents through direct actions, such as being open about the mechanisms necessary for privacy. For example, Allegheny County, Pennsylvania, operates a public, open, and transparent process for using government-collected data at the county level. In addition to making information available about the types of activities underway, county officials engage in frequent meetings and dialogues with community members and program stakeholders. These types of engagements vary by jurisdiction, but optimally they serve as a way to understand and address concerns about how data are used and to hold officials accountable for responsible uses.

CURRENT APPROACHES FOR ACHIEVING PRIVACY GOALS

Although policy often restricts sensitive data sharing among organizations and government agencies today, some sharing does take place. Attempts to assure privacy during such sharing generally fall into four categories: (1) de-identification, (2) synthetic substitution, (3) Input Party calculation, and (4) contractual control. Federal statistical agencies pioneered and developed strategies for deployment of many of these approaches within government.[21] However, each has weaknesses that make such sharing a non-zero risk activity.

Data De-identification

One approach to achieving both Input Privacy and Output Privacy while sharing data is to *de-identify* the data prior to sharing. De-identification removes or obfuscates input data that might be used to associate data with individuals and, thus, enable harming those individuals. Several techniques may be used to achieve de-identification either before or after analysis and are commonly used within some federal agencies, such as the U.S. Census Bureau.[22]

Unfortunately, de-identification as a form of disclosure avoidance can be expensive, does not offer an absolute guarantee of confidentiality,[23] requires a high level of technical expertise and precision to be properly implemented, and often restricts data utility, which defeats the purpose of combining source data to produce useful, statistical insights. De-identification may need to be repeated for each new use of data—depending on the kind of research to be done—an expensive effort that must be borne by the owner of the data. Finally, successive de-identification becomes more complex because prior data releases must be taken into consideration and, thus, also carry the additional risk of re-identification.

De-identification approaches are standard for many federal government agencies, and the practice is encouraged in some federal laws. For example, the Health Insurance Portability and Accountability Act's Privacy Rule requires removal of specific fields—including names, small geographic subdivisions, specific dates, Social Security numbers, etc.—to nominally de-identify datasets.[24] Other laws, such as the Confidential Information Protection and Statistical Efficiency Act of 2018, recognize that numerous data fields can be examined in combination, which must be considered when determining appropriate techniques for de-identification and disclosure avoidance.[25]

Synthetic Data Substitution

Another approach is to first learn the statistical relationships among the variables in a dataset and then create one or more plausible datasets that produce approximately the same statistics as those from the original dataset.

By creating synthetic data, statistical relationships such as correlations can be approximately replicated for existing data. Unfortunately, synthetic substitution can typically only reproduce known relationships. For example, an expert preparing synthetic substitute data may be aware that the real data show a correlation between two variables, but he or she may not be aware of other correlations in the data. In many cases, de novo research seeks to discover such new relationships, which is impossible when those relationships are lost during the synthesis process simply because they were not already known. In addition, synthetic data construction is an expensive process that also must be borne by the data owner.

Synthetic data may, however, be useful in conjunction with cryptographic privacy-preserving methods such as multiparty computation. Researchers often first explore data to develop the analyses they eventually use to attain useful results. However, cryptographic privacy-preserving methods actively prevent researchers from seeing data. Synthetic substitutes of sensitive data may thus be useful for exploration and query formulation, resulting in queries that can then be run securely on original sensitive data to attain those results.

Input Party Computation

Another approach to privacy preservation is for the Input Parties to perform the computations needed for analytics and pass the results directly to the Output Parties, obviating the need for intermediary Computing Parties. In this setting, Input Parties never need to reveal sensitive, identifiable data and can control which results are released to Output Parties. Input Parties may also bear responsibility for performing risk assessment about disclosure risk avoidance and also for performing de-identification activities.

Unfortunately, this approach is not scalable because Input Parties must run all analytical queries, regardless of how many Output Parties request multiple analyses. The logistical and computation load to complete all such requests is untenable for many data providers.

Contractual Controls and Agreements

Perhaps today's most popular approach to privacy preservation while sharing data is the application of contractual controls. In this approach, Computing Parties or Output Parties sign *data-sharing agreements* with Input Parties that require users to accept liability in order to assure data confidentiality and to destroy data at an agreed upon time frame, usually after the completion of a project. The contractual controls enable Input Parties to screen eligible users of their data as well as the purposes for which projects are conducted ex ante. In exchange, once approved through a contract, parties are then allowed access to data based on the contractual conditions.

Contractual controls are subject to several practical limitations. First, parties may be unwilling to accept liability for assuring data confidentiality, which could lead to legal exposure for even unintentional errors. Second, data-sharing agreements can establish parameters on user access but may not fully address threats that arise from attempts to exfiltrate data without permission. Third, complete data deletion is often not possible. Fourth, data providers often (rightly) fear that the cybersecurity stances of parties to which they provide data are not ready to adequately protect that data. Finally, contractual controls also necessitate the negotiation of contracts, which requires legal discussions about data uses and purposes. The introduction of attorneys into time-sensitive data requests, particularly in organizations without a routine process for doing so, can extend the time frame for project completions that may not satisfy the goals for Output Parties or eventual users.

WHAT IS MULTIPARTY COMPUTATION?

Given the limitations of traditional approaches to data disclosure avoidance in achieving specified privacy, multiparty computation offers a compelling alternative. Multiparty computation is an approach that allows information to be accessed securely, with a privacy guarantee, and analyzed without making individual private information available. The application of multiparty computation relies on cryptographic techniques that mathematically alter information to allow computer code to generate an analysis, without participating computers or humans seeing or being able to access the underlying individual identifiable data. In multiparty computation, Input Parties encrypt sensitive datasets and then transmit encrypted data to Computing Parties. Often, such encryption is performed just prior to transmission. Input Parties secure such data on their own systems prior to transmission independent of the multiparty computation approach. When multiple datasets are involved in a computation, multiple Input Parties may be involved, each contributing one or more datasets to the Computing Parties, who then perform secure computation on the combined datasets.

Multiparty computation offers several distinct benefits relevant for a contemporaneous data-sharing environment. First, multiparty computation can be used with other disclosure avoidance protocols. Indeed, a demonstration project is currently underway to test multiparty computation using synthetic data.[26] The practicality of applying multiparty computation in combination with existing protocols is that the approach both enhances confidentiality protection and provides administrators and data owners with an ability to rely on conventional approaches in the near-term while they develop a stronger understanding of the capabilities of multiparty computation.

Second, multiparty computation can be useful in circumstances where government jurisdictions or organizations have incentives to otherwise not share data with each other when there is a risk of privacy loss. For example, a local jurisdiction may choose to not share program operation data with a state or federal government funder if the data could jeopardize an individual program's operations. However, multiparty computation allows multiple jurisdictions to provide input data without risking organizational re-identification and does so in a way that may still achieve broader public accountability and transparency goals.

Third, data can be useful for multiparty computation when stored in multiple locations. Traditional analysis requires data to be brought together into one place for conducting linkages and analytical operations, which adds risks for data security related to transmittal and storage. Multiparty computation can eliminate the need for additional risks by using data in its original location.

Fourth, multiparty computation can potentially overcome certain legal restrictions to data sharing, such as those that explicitly indicate that personally identifiable information cannot be shared with individuals in a manner that would constitute a disclosure. Thus, multiparty computation may have the potential to overcome certain constraints in place under the income tax code (Title 26) for income and earnings information as well as under certain education restrictions (Title 20) that inhibit or restrict data use.

Finally, multiparty computation systems can be designed to consider a broad array of informational inputs and then make determinations based on various criteria to indicate whether individuals could, for example, be designated for supplemental services without identifying the specific reasons or factors that led to such a determination. In this way, the approach enables data analysis with constraints placed on the output information available to other users (referred to above as *Access Control*).

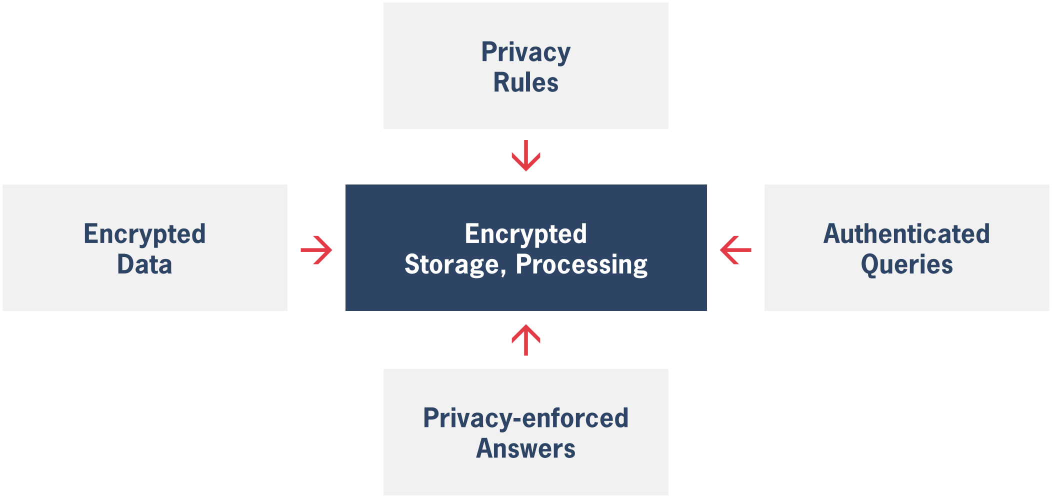
TECHNOLOGY PLATFORMS FOR DEPLOYING TRUSTED MULTIPARTY COMPUTATION

There are several platforms for conducting multiparty computation and other privacy-preserving analytics. Two systems developed by Galois, Inc., are discussed below: the Jana encrypted relational database system and the FIDES Trusted Execution Relational Database System.

The Jana Encrypted Relational Database System

Jana is a relational database research platform originally intended to study trade-offs between assuring privacy and performance. Jana provides many of the same capabilities expected of a relational database, while also supporting Input Privacy, Output Privacy, and Access Control. In this demonstration project, the project team focused on Jana's ability to protect Input Privacy—demonstrating that its privacy-preserving computation prevents the Computing Parties, or any other party that can witness the computation, from learning the input data.

Figure 2: Overview of the Jana Encrypted Database System



Jana's components and data flow are illustrated in Figure 2. When a database administrator creates a Jana database, the structure and format of the data in that database are specified in the typical way, with additional selection by the administrator of the encryptions to use for each attribute (column) and in each table (relation) of the database. Once the database is initialized, data are entered into the system (shown at left in Figure 2) through an interface similar to that used by typical relational databases. Sensitive data are encrypted prior to being entered into the system and are additionally protected while transmitted to the system using typical Transaction Layer Security protocols.

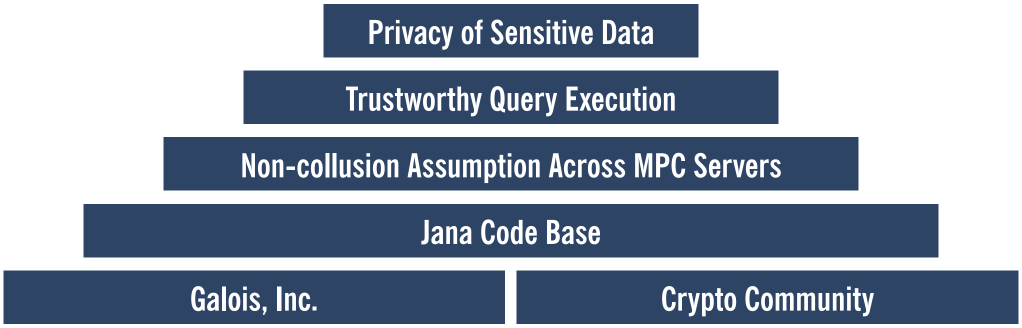
Once received by the Jana system, data are re-encrypted to match the database administrator's specifications. Secure multiparty computation performs the re-encryption so that Jana does not access the input data during the re-encryption process.

Before analyses are accepted by the system, the administrator specifies Access Control rules that restrict what each user may learn about the stored data. This project did not exercise this capability of Jana because those controls have minimal impact on system performance.

The system accepts authenticated analytic queries (as shown at right in Figure 2). Jana modifies each query, implementing specified Access Controls and determining how to execute the queries most effectively given the available data encryptions. Query answers are then sent in encrypted form to the original requester, where those encrypted answers are decrypted into plain text results. Thus, for encrypted data, Jana never learns what the data are or anything about the answers to queries computed on the data (except for the number and size of records returned by a query).

Every system that provides security has a trust basis—a set of assumptions or artifacts that users must trust in order to believe in each claim the system makes about security. The trust basis for Jana is illustrated below in Figure 3.

Figure 3. Trust Basis Diagram for Jana


Privacy of Sensitive Data
Trustworthy Query Execution
Non-collusion Assumption Across MPC Servers
Jana Code Base
Galois, Inc.
Crypto Community

Jana's claim is that the sensitive information provided to the system remains private. In Jana, that privacy depends on query execution that is *trustworthy*—that is, the system does not learn or leak anything about the data or query results. Trustworthy execution in turn relies primarily on a *non-collusion* assumption: that the two (or more) independent computers that process queries in a secure way do not collude to learn the data. That assumption in turn relies on two key underpinnings: the *implementation* of the cryptographic techniques provided by the framework programmers, and the *cryptographic theory* of secure multiparty computation developed by the cryptography research community.

The FIDES Trusted Execution Relational Database System

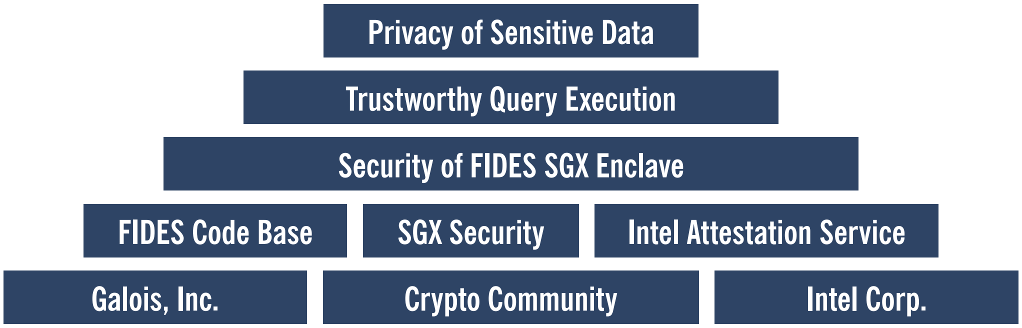
FIDES is a relational database research platform originally intended to study scalability of data access while assuring privacy of database content. FIDES provides many of the same capabilities expected of a relational database, while also supporting Input Privacy, Output Privacy, and Access Control.

At a summary level, FIDES also matches Figure 2. To use FIDES, the user first creates a secure FIDES computation enclave to receive sensitive data. The *enclave* is created from a known software package that has a verifiable *digital signature*. Enclave creation is automatic, requiring no special user expertise.

Once this enclave is created, the user requests access to one or more relevant datasets from Input Parties such as data owners. Tools used by those parties automatically verify the integrity and code content of the enclave using hardware-enabled cryptographic mechanisms included in the Intel Corporation Software Guard Extensions (SGX) standard. Once the Input Parties are each reassured by these means that the user's enclave is genuine and correct, they transmit Advanced Encryption Standard (AES) data in relational database form to the user and then transmit the decryption key for the database *only to the enclave*, to which no user, privileged code, or other entity has access. The enclave then accesses the encrypted database, decrypts it within the enclave's protected memory, and can then perform any necessary data linking among those datasets and then process queries. Those queries are rewritten by enclave code to enforce whatever Access Control policies are required by the administrator providing the encrypted database.

Figure 4 shows the trust basis for FIDES.

Figure 4. Trust Basis Diagram for FIDES


Privacy of Sensitive Data
Trustworthy Query Execution
Security of FIDES SGX Enclave
FIDES Code Base
SGX Security
Intel Attestation Service
Galois, Inc.
Crypto Community
Intel Corp.

In FIDES, the basis of trust for privacy-preserving query execution is first the code loaded into the underlying SGX enclave (similar in many ways to Jana's reliance on the Jana code base shown in Figure 3). FIDES then relies on the integrity of the hardware-enabled computing enclave, as opposed to Jana's reliance on multiparty computation and non-collusion among Computing Parties. The integrity of the SGX enclave relies in turn on the Intel SGX architecture and the Intel remote attestation web service.

Recent research has discovered some security vulnerabilities in the Intel SGX architecture and in similar capabilities provided by other hardware manufacturers. In response, Intel Corporation has offered both short-term mitigations and medium-term permanent solutions to these issues. Such discovery and mitigation is typical for both hardware and software systems, and thus is not overly concerning, though the situation remains worth monitoring.

Explaining Differences Between FIDES and Jana

The FIDES system differs from the Jana system in several ways:

1. FIDES relies on hardware-enabled cryptographic mechanisms to assure that the Input Parties' data remain private; Jana relies on software-enabled secure multiparty computation. Thus, the Jana platform may be more portable than the FIDES platform.
2. In FIDES, the entire database is encrypted in AES and then decrypted only once inside the FIDES secure enclave; in Jana, each item in the database is encrypted independently, and the secure query processor processes data while it remains in an encrypted representation. FIDES relies on hardware features (Intel SGX) to assure that no user, software, or system has access to data or query processing, while Jana relies on cryptographic data representations.
3. FIDES uses a relatively simple security setting: the FIDES enclave acts as a "secure glovebox" that allows data manipulation while preventing unauthorized access. In contrast, Jana relies on multiple servers that must not collude with each other.
4. FIDES is, in general, far faster than Jana for many database operations.

Given that FIDES is easier to understand and deploy than Jana, and that it is much faster and has a more intuitive security model, why is Jana interesting? The answer lies in Jana's portability and independence from computing hardware features. The premise of Jana is that it can be deployed on any general-purpose computing hardware, rather than requiring platforms to support features such as Intel SGX. Jana also does not require access to an Intel-provided web service for verifying security integrity. However, as shown in the Demonstration Project Results section below, that independence comes at a high computational price.

Allegheny County Demonstration Project

Recognizing the potential for multiparty computation as a benefit to privacy-preserving data analytics, as acknowledged by the Commission on Evidence-Based Policymaking, a demonstration project was developed using real government-collected data. The demonstration was designed to (1) provide a real demonstration of the efficiency and capability of multiparty computation techniques for maintaining data encryption of individual-level records, and (2) validate that the techniques achieve answers identical to computations performed through a traditional data analytic infrastructure.

SITE SELECTION

Several governmental entities were explored as potential candidates for the project based on the following selection criteria:

1. A government agency at the federal, state, or local level in the United States;
2. An organization with a robust existing data infrastructure capable of operating complex analysis with data from multiple agencies;
3. An organization with detailed data documentation, or metadata;
4. An organization with a defined data-sharing system that would enable secure access from a third party; and
5. An organization with administrators willing to learn about multiparty computation and support a test of the approach.

Of the multiple government agencies considered at different levels of government, Allegheny County, Pennsylvania satisfied all five selection criteria. Allegheny County is widely recognized as an exemplar for curating and providing analysis frameworks for a broad selection of data across many of its agencies.

Counties, along with government agencies at other levels of government, must choose how to allocate scarce resources in support of marginalized or vulnerable populations within their jurisdictions. Wise allocation decisions depend on analysis of trends and current situations, described in data often held disjointly by diverse agencies. For example, the level of investment in certain homeless services might be adjusted upward or downward based on the objectively measurable impact those services have in terms of stable housing, employment, and other measures of health and family stability.

RESEARCH QUESTIONS WITHIN THE DEMONSTRATION

Initial outreach to Allegheny County personnel occurred in April 2018, and the project team first met with county staff in June 2018 to describe the nature of the proposed project and garner conceptual support. A data-sharing confidentiality agreement was approved in August 2018, and the project was publicly announced in September 2018.[27]

In collaboration with Allegheny County staff, the project team formulated four questions relevant for decisions made by public administrators and policymakers that could be answered using existing data collected by the county:

1. What was the proportion of people serving a sentence in the county jail during 2017 or 2018, who in that same period received publicly-funded mental health services?
2. What was the proportion of parents with open child welfare cases who received publicly-funded mental health services during 2017 and 2018?
3. What was the proportion of people serving a sentence in the county jail during 2017 and 2018, who also received county homelessness services during that period?

4. What was the proportion of suicide victims reported by the medical examiner's office during the 2017 and 2018 period who also previously received publicly-funded mental health services?

The answers to each question could be used by a decision-maker to understand the nature of problems or solutions in a policymaking context.

To study the questions, Allegheny County provided approved members of the project team with secure access to de-identified records in five distinct datasets:

1. Provision of homelessness services to resident populations;
2. Publicly funded mental health services provided to resident populations;
3. Causes and incidences of autopsied death in resident populations;
4. Child welfare cases (for children, youth, and family); and
5. Jail bookings and sentences.

The data are collected by multiple agencies in Allegheny County, but collectively are relevant for addressing human services research questions. The data included records from 2017 and partial-year information from 2018. The data obtained were pre-cleaned by county staff, and personnel were available to answer questions about the meaning and quality of the data fields. The data structure was relational, with a total of five tables and over 2 million total records (see Table 1).

Table 1. Characteristics of Data for Experiments

COMPARING THE THREE APPROACHES

Dataset Description	Collecting Agency	Number of Records
Homeless Services and Supports[28]	Allegheny County Department of Human Services	14,437
Mental Health Services and Supports[29]	Allegheny County Department of Human Services	2,050,000
Medical Examiner Records[30]	Allegheny County Medical Examiner's Office	4,018
Child Welfare Case Data[31]	Allegheny County Department of Human Services	14,118
Jail Booking Data (limited to people serving sentences)[32]	Allegheny County Jail	2,063

Between September 2018 and December 2018, the project team analyzed de-identified records. In collaboration with Allegheny County staff, the project team created eight key Structured Query Language (SQL) queries (two for each question) to answer the four identified research questions within the demonstration. Most questions involved two sets of queries: one to obtain results for calendar year 2017, and one to obtain similar results for 2018. Most queries also included complex operations, such as linking among datasets, aggregate statistical computations over that linked data, and nested (hierarchical) query structures. Notably, such general-purpose computation capabilities are unavailable in other secure computation paradigms, such as Private Information Retrieval, making performance results between this work and other such work incomparable.

Using the five datasets described above, the project team performed the four statistical analyses on three platforms. The first platform was a non-secure control experiment, the SQLite relational database system. The second platform was the Jana encrypted database system, currently under development as part of the Brandeis program for the Defense Advanced Research Projects Agency. The third platform was the FIDES encrypted database system, currently under development for the U.S. Department of Homeland Security's Science and Technology Directorate IMPACT program. Each of these platforms supports relational database capabilities and the SQL relational query language; and each produces results typical of relational database outputs. Relational data capability is a crucial requirement: most government agencies and non-governmental organizations use relational data, and the most efficient and well-understood way to achieve cross-organizational data linkage is via the relational model.

Each of the two secure platforms accept previously encrypted data; transmit, store, and process data while it remains cryptographically protected from all unauthorized access; and is equipped with Access Control decision and enforcement capability to control what analysts can access in analytic outputs.

In all cases, dataset linkages were deterministically conducted in the experiments using a common identifier value. A deterministic linkage was performed in exactly the same way across all three experiment sets. Additional data-cleaning and normalization were required prior to query processing.

RESULTS

The multiparty computation demonstration generated valid results, consistent in all cases with the control experiments conducted without privacy preservation. Each of the queries represents a question of particular relevance for public administrators. But for the purposes of the demonstration project, the specific details of each query are not critical for fulfilling the project goals.

Table 2. Results of Demonstration Research Questions

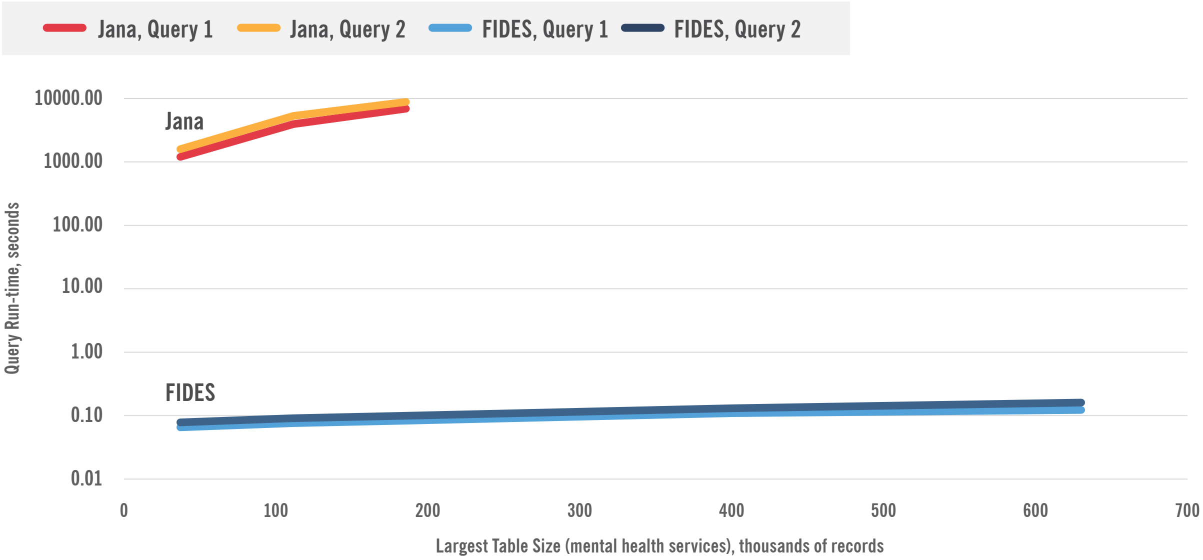
Demonstration Research Question	Traditional	Jana	FIDES
1. Share of people serving a sentence at the Allegheny County Jail who received publicly-funded mental health services[33]	2017: 37.9% Partial 2018: 40.7%	2017: 37.9% Partial 2018: 40.7%	2017: 37.9% Partial 2018: 40.7%
2. Share of parents with open child welfare cases who received publicly-funded mental health services[34]	2017: 26.6% Partial 2018: 26.2%	2017: 26.6% Partial 2018: 26.2%	2017: 26.6% Partial 2018: 26.2%
3. Share of people serving a sentence at the Allegheny County Jail who received county homeless services[35]	2017: 5.1% Partial 2018: 4.6%	2017: 5.1% Partial 2018: 4.6%	2017: 5.1% Partial 2018: 4.6%
4. Share of suicide victims who received publicly-funded mental health services[36]	2017: 11.5% Partial 2018: 15.3%	2017: 11.5% Partial 2018: 15.3%	2017: 11.5% Partial 2018: 15.3%

Both the Jana and FIDES models returned consistent responses to queries to align with a classic analytical process, and there were no statistically significant differences in responses across the approaches (Table 2). Because both secure multiparty computation and secure enclave computation perform the same analytic functions as are performed without privacy preservation, this result is not surprising. However, it does confirm that privacy-preserving approaches are not subject to any diminished quality that would affect the validity or reliability of statistical conclusions. Therefore, the multiparty computation models satisfy the demonstration's core criteria for enabling data use and privacy preservation.

In terms of the computational performance of the different platforms for multiparty computation, the results suggest different modes of privacy-preserving technologies offer trade-offs for answer timeliness (see Figure 5). Using a mental health services table with 110,000 entries took 100 milliseconds to run in FIDES, but it took roughly 7,000 seconds to run in Jana. These times have substantial implications for applications within government operations with rapid decision-making architectures.

The sample result shown in Figure 5 is representative of all queries run during the experiments. FIDES performed well in terms of scalability and exhibited acceptable run times at all data sizes attempted. FIDES performed comparably to running the same queries without any privacy protection, using the same commercial database as used in the FIDES system. Jana, on the other hand, was notably slower by about four orders of magnitude. Because Jana queries in excess of 10,000 seconds (about three hours) were stopped due to long run times, the results are less complete for that system but are expected to continue to see performance declines as the number of records analyzed increases.

Figure 5. Summary Results from Project Query Set on Mental Health Services



Based on the results, FIDES may offer acceptable performance for datasets in the multiyear category at the county level. More work remains to be done to evaluate how well FIDES scales to substantially larger datasets, such as those encountered at the national level or for longitudinal studies. Additional experience from the project team from SGX systems suggests that enclave-based solutions may also scale well up to millions of streaming transactions per second.

In contrast, Jana faces difficulties producing rapid results with large datasets. Indeed, other Jana deployments show that Jana is best suited for datasets up to perhaps a few tens of thousands of records. Jana's performance—four orders of magnitude slower than FIDES—is roughly what the project team expected to see in terms of performance of current software-based secure multiparty computation technology applications. While highly optimized, specifically designed analytics using multiparty computation may achieve a somewhat higher performance. Jana's results in this study are representative of what multiparty computation can achieve today on general-purpose computations.

In sum, the demonstration project suggests satisfaction of the goals at the outset: that privacy-preserving data analytics platforms can generate correct answers in comparison to a typical non-secure analysis and that the privacy guarantees of those platforms are robust. While the performance attained in the hardware-based solution (FIDES) exceeds the performance of the software-based solution (Jana), both performances could be acceptable in certain real-world applications given their success in producing valid results that maintain data privacy.

POTENTIAL PROJECT EXTENSIONS AND FUTURE RESEARCH

While the demonstration project was successful in terms of achieving its goals, there are numerous opportunities for extensions and additional insights that could be gained with additional research support capabilities.

Use of Access Control Mechanisms

Although both FIDES and Jana have the ability to interpret and enforce Access Controls expressed in an access policy language, the demonstration did not make use of these features. In brief, both platforms can assign attributes to users and can then interpret and enforce Access Control rules that determine which results are viewable by users with which attributes and under what conditions. Options for control of result visibility include: no access, differential privacy-protected aggregate access only (currently only in Jana), aggregate access only, and full access. A prototype field deployment of one or both platforms would allow an evaluation of these Access Control features in a realistic environment.

Data-Cleaning

Data from Input Parties often arrive at the Computing Parties with inconsistencies, heuristic assumptions about values of certain attributes, and gaps that require manual adjustments and "cleaning." For example, when NULL values are used, they can be inadvertently inconsistent within a database. In addition, normalizing data across multiple datasets from different Input Parties requires substantial effort to resolve mismatches among datasets. Because these problems are often semantic rather than technical, this project did not address data-cleaning beyond essential elements needed for input into the analytics infrastructure. Such data-cleaning is required for all analytics that share data across organizations, regardless of whether the privacy-preserving techniques described in this project are employed or not. Data-cleaning is best done by the Input Parties before any data are provided to Computing Parties, which was a factor in the selection of Allegheny County as a project site. However, an area for future study would be to examine how to perform data-cleaning that is consistent across datasets when sensitive data come from multiple providers yet must be processed together, and must be cleaned by a party such as a Computing Party unable to access the data in unencrypted form.

Thus, one area for additional study is the implementation of realistic data-cleaning in a FIDES Computing Party enclave on data first from one Input Party and later from multiple Input Parties. Such cleaning might involve imputation for missing data, normalization of "special values" used by different Input Parties, and transformation of incompatible data types. For example, this could include fixed-point data in some datasets and floating-point datasets in others. Currently, complex data-cleaning is impractical in a Jana platform due to the computation being performed directly on encrypted data representations.

Additional Calculation Complexity

This demonstration project focused on relational queries with relatively simple numerical or statistical processing: simple mathematics such as counts were used to generate averages, but no complex statistical formulas such as regressions or distributional analyses were included. One extension of interest would be straightforward statistical analysis of data without relational queries—for example, logistic regressions or quartile regressions over large datasets. These approaches would provide insights about the capabilities for highly complex calculations that are often deployed for human services research and evaluation.

Data-Streaming Capability

Another interesting area of study would be to examine the processing of streaming data in near real-time. Galois, Inc., recently demonstrated a proof-of-concept prototype able to do simple mathematics such as tabulations at up to millions of transactions per hour for economics statistics. A more fully featured pilot of this capability would be useful to federal statistical agencies, such as the U.S. Census Bureau.

Probabilistic Record Linkage Across Datasets

This demonstration project relied solely on deterministic record linkage among datasets: a single key between datasets, used with a simple matching algorithm, was used to perform data linkage. However, realistic datasets are often less capable of such precise linkages because data may be corrupted, incorrectly entered, or missing, requiring reliance on other approaches for linking. While probabilistic match algorithms are commonly used for traditional data linkages outside privacy-preserving approaches, such matches have not yet been deployed in combination with privacy-preserving technologies. A pilot using such probabilistic linking—for example, in matching airline passenger manifests with federal no-fly lists—might demonstrate the capability to preserve individual privacy for non-violators and maintain corporate information confidentiality while providing accurate matching across datasets.

Policy Implications and Next Steps

The prototypes reported here demonstrate that technologies such as hardware-enabled privacy preservation are practical and currently perform as expected for useful queries that combine sensitive datasets. They also demonstrate that hardware-independent approaches based on advanced secure computation can achieve the same results, albeit at reduced performance in general computation settings.

The results of this demonstration have some immediate public policy implications. First, much more work is likely needed before multiparty computational approaches could be scaled at a national level or for large state or local government operations. For example, use of a software-based computational approach within the context of the decennial census at current levels would likely result in untenable computing time frames. Extrapolating to just 500,000 records would likely result in computation times measured in days. Thus, the millions of records in a decennial census in the United States would quickly become unwieldy, with compute times in weeks, or longer.

Second, while the hardware-based solution may be timelier for returning results for decision-makers, the cost of identifying and using the infrastructure for millions of records may be prohibitive for recurring or large-scale calculations. However, cases that allow for analyses performed over hours and days should be further explored to determine if such approaches could be justified within operational settings. Third, as noted above regarding potential extensions, computational capabilities have not yet been built into existing platforms to operationalize complicated statistical calculations, including those that result in distributional or stratified statistics to provide information about subgroups.

Fourth, and at a very practical level, the number of programmers steeped in the techniques is currently limited. A larger workforce is needed to implement the approach at scale across local, state, and federal government agencies. The existing workforce within government agencies will likely need to be enhanced to deploy and monitor the effectiveness of new privacy-preserving approaches at scale.

Fifth, at least within the federal government, deployment of privacy-preserving technologies may require additional clarification in the legal and policy framework to enable their use beyond targeted demonstration and research activities. At the state and local levels of government, some frameworks may exist, though it is likely additional legal clarification will be desirable for public administrators and policymakers alike. Minimally, such legal clarification would address any ambiguities that may impede future adoption and could also help plan for the need to provide suitable transparency and accountability mechanisms to protect public trust in statistical or evidence-building activities as well as in privacy-preserving technologies.

Finally, agencies must establish data governance processes and structures to enable improved trust for conducting analyses, even in settings that rely on privacy-preserved analytical approaches. Entities like Allegheny County, for example, could likely adopt such approaches increasingly in the future while jurisdictions without an existing data culture or infrastructure could still face challenges initiating activities in the first place. That said, multiparty computation could help establish the conditions for trust in public agencies over time. Local governments are increasingly moving toward using data to inform better decision-making in all types of activities and, in some ways, may offer insights for federal agencies in the future. This case is but one example.

Privacy-preserving approaches have the strong potential to serve as productive tools in the realization of the full value of government-collected data. These approaches can give decision-makers access to statistical information that helps them understand trends and inform policy decisions, while also maintaining privacy expectations. The natural next steps for advancing privacy-preserving analytics fall along several general themes:

- **Engage in continued research and development of privacy-preserving technologies.** In line with the recommendation from the Evidence Commission, continued effort and resources could be allocated to further explore improvements to the approaches. In addition, the functionality of prototypes could be extended, as outlined in the "Potential Project Extensions" section above. Efforts to continue this research will likely require a combination of government, foundation, and university research support.
- **Develop additional demonstration projects and pilot programs in other settings.** Continued attention to testing privacy-preserving approaches in real-world settings using a range of data types and policy domains will help calibrate the precision and efficiency of the approach, as well as scalability and usability of the capabilities. Further demonstrations will also support the case to public administrators and policymakers that the techniques can truly achieve both privacy and data analysis goals. Future demonstrations could be advanced with support from foundations in controlled settings for state and local governments. Congress or the executive branch could also identify targeted pilots within federal agencies. For example, Congress could use the appropriations process or an authorization to direct agencies to engage in more demonstration projects. The executive branch could also use the emerging Federal Data Strategy to signal interest and attention to more demonstration projects.[37]
- **Plan for future deployment of privacy-preserving technologies.** Even agencies and policymakers who are not yet ready to endorse the approaches could still enable a data infrastructure capable of future deployment. This demonstration relied on the well-organized data infrastructure created by Allegheny County over decades of planning and prioritization, including substantial investments of resources in the infrastructure. Other governments and agencies at the local, state, and federal levels could similarly take steps to develop data documentation, data standards, and data inventories to support successful implementation in the future. For federal agencies, these attributes are required under the Foundations for Evidence-Based Policymaking Act of 2018, though additional work will be needed to ensure that even the data infrastructure of federal agencies is well-suited for the approach. Prioritizing efforts to ensure implementation of the Evidence Act would certainly support future efforts and likely lower the cost of implementation for privacy-preserving technologies in coming years.

In the short-term, this demonstration clearly suggests that pursuing hardware-enabled privacy preservation is ideal where general-purpose computation is required. The application of software-based secure computation approaches will not be useful for general computation activities for years given the need for additional research and development.

In sum, secure computation technology offers the promise for achieving substantial gains in privacy protections for the American public, but there are trade-offs associated with computational cost and timeliness of information that will need to be carefully weighed by policymakers. Nonetheless, the demonstration project's success suggests this privacy-preserving technology will be useful for human services systems and that the approach has the potential to unlock novel insights not otherwise available to government and policymakers today.

Endnotes

[1] N. Hart and K. Wallman. *Transparency, Accountability, and Consent in Evidence Building: How Government Ethically and Legally Uses Administrative Data for Statistical Activities.* Washington, D.C.: Bipartisan Policy Center, 2018.
Available at: https://bipartisanpolicy.org/library/transparency-accountability-and-consent-in-evidence-building/.

[2] U.S. Commission on Evidence-Based Policymaking. *The Promise of Evidence-Based Policymaking: Final Report from the Commission on Evidence-Based Policymaking.* Washington, D.C.: Government Printing Office, 2017.
Available at: https://bipartisanpolicy.org/wp-content/uploads/2018/07/Full-Report-The-Promise-of-Evidence-Based-Policymaking-Report-of-the-Comission-on-Evidence-based-Policymaking.pdf.

[3] Ibid, *i*.

[4] Ibid, 1-2.

[5] Ibid, Recommendation 3-2.

[6] Ibid.

[7] In this context, a "smaller piece" means a piece that is intelligible only when combined with the other related small pieces.

[8] N. Hart and K. Fatherree. *A New Technology May Revolutionize Privacy-Preserving Data Analysis: Secure Multi-Party Computation.* Washington, D.C.: Bipartisan Policy Center. Available at: https://bipartisanpolicy.org/blog/secure-multi-party-computation/.

[9] G. Alter. Presentation: "Will Secure Multiparty Computation Reshape Data Privacy?" Washington, D.C.: New America and Bipartisan Policy Center, April 17, 2018. Available at: https://bipartisanpolicy.org/events/will-secure-multiparty-computation-reshape-data-privacy/.

[10] J. Brown. "Tackling the Wage Gap with Code." *BU Today*, 2017.
Available at: http://www.bu.edu/today/2017/tackling-wage-gap-with-code/.

[11] J. Kim, M. Epitropakis, and S. Yoo. Learning Without Peaking: Secure Multi-Party Computation Genetic Programming." *Search-Based Software Engineering*, International Symposium on Search Based Software Engineering: 246-261.
Available at: https://link.springer.com/chapter/10.1007/978-3-319-99241-9_13.

[12] I. Damgard and T. Toft. "Trading Sugar Beet Quotes—Secure Multiparty Computation in Practice." *ERCIM News*, 2008. Available at: https://ercim-news.ercim.eu/en73/special/trading-sugar-beet-quotas-secure-multiparty-computation-in-practice.

[13] D. Archer, D. Bogdanov, L. Kamm, et al. "From Keys to Databases—Real-World Applications of Secure Multi-Party Computation." *The Computer Journal*, 61(12): 1749-1771. December 2018. Available at: International Association for Cryptologic Research ePrint Archive: https://eprint.iacr.org/2018/450.pdf.

[14] After the commission's research phase completed, the Laura and John Arnold Foundation launched a project to pilot multiparty computation using synthetic data to study higher-education outcomes. *See:* G. Alter. Presentation: "Will Secure Multiparty Computation Reshape Data Privacy?" Washington, D.C.: New America and Bipartisan Policy Center, April 17, 2018.
Available at: https://bipartisanpolicy.org/events/will-secure-multiparty-computation-reshape-data-privacy/.

[15] U.S. USA Spending, 2017. Available at USASpending.gov.]

[16] *See:* S. 2169, 115th Congress, Right to Know Before You Go Act.

[17] *See:* HR 6562, 115th Congress, FORWARD Act of 2018.

[18] N. Hart and K. Wallman. *Transparency, Accountability, and Consent in Evidence Building: How Government Ethically and Legally Uses Administrative Data for Statistical Activities.* Washington, D.C.: Bipartisan Policy Center, 2018.
Available at: https://bipartisanpolicy.org/library/transparency-accountability-and-consent-in-evidence-building/.

[19] U.S. Commission on Evidence-Based Policymaking. *The Promise of Evidence-Based Policymaking: Final Report from the Commission on Evidence-Based Policymaking*. Washington, D.C.: Government Printing Office, 2017: 38.
Available at: https://bipartisanpolicy.org/wp-content/uploads/2018/07/Full-Report-The-Promise-of-Evidence-Based-Policymaking-Report-of-the-Comission-on-Evidence-based-Policymaking.pdf.

[20] Ibid, 51.

[21] Federal Committee on Statistical Methodology. *Report on Statistical Disclosure Limitation Methodology.* Statistical Policy Working Paper 22. Washington, D.C.: Federal Committee on Statistical Methodology. Available at: https://nces.ed.gov/FCSM/pdf/spwp22.pdf.

[22] A. Lauger, B. Wisniewski, and L. McKenna. *Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research*. Research Report Series #2014-02. Washington, D.C.: Center for Disclosure Avoidance Research, 2014. Available at: https://www.census.gov/srd/CDAR/cdar2014-02_Discl_Avoid_Techniques.pdf.

[23] Existing research suggests that in certain settings de-identified data can be vulnerable to a variety of re-identification risks, especially when access to other correlated data are available.

[24] U.S. Department of Health and Human Services. Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the HIPAA Privacy Rule. Washington, D.C.: HHS Office for Civil Rights, 2015. Available at: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard.

[25] *See:* Title III of P.L. 115-435, Foundations for Evidence-Based Policymaking Act of 2018.

[26] G. Alter. Presentation: "Will Secure Multiparty Computation Reshape Data Privacy?" Washington, D.C.: New America and Bipartisan Policy Center, April 17, 2018. Available at: https://bipartisanpolicy.org/events/will-secure-multiparty-computation-reshape-data-privacy/.

[27] Bipartisan Policy Center. BPC Partners with Allegheny County on New Privacy-Preserving Data Project. Press Release. Washington, D.C.: Bipartisan Policy Center. Available at: https://bipartisanpolicy.org/press-release/bpc-partners-with-allegheny-county-on-new-privacy-preserving-data-project/.

[28] Defined as individuals and families who are homeless, or at risk of becoming homeless, who are also receiving prevention, support services, and/or housing from Allegheny County Department of Human Services. Prevention services are targeted to individuals who may have a home or apartment but need help with past-due rent, mortgage, or utility bills in order to prevent homelessness. Support services include homeless drop-in centers, day shelters, and case management. Housing includes emergency shelter, bridge and permanent supportive housing, and rapid re-housing.

[29] Defined as individuals who receive any publicly funded (Allegheny County or Medicaid managed care/HealthChoices) mental health service, including both clinical services, such as individual and group therapy, and non-clinical services, such as case management and peer support.

[30] Defined as the number of autopsied deaths occurring within Allegheny County as reported by the Allegheny County Medical Examiner's Office.

[31] Defined as parents of children and youth associated with an open child welfare case in Allegheny County.

[32] Defined as individuals serving a sentence at the Allegheny County Jail who have been assigned a bed.

[33] Two queries were executed. First, we counted the number of distinct individuals who were both serving sentences during the period and who received publicly funded mental health services during the time period from March 4, 2017, to August 20, 2017. Second, we counted the number of distinct people who were serving sentences during the period. Finally, we divide the result of query one by the result of query two. When considering these results from a policy, rather than purely technical, perspective, two issues are of note. First, because people cannot be simultaneously receiving publicly funded mental health services and be incarcerated, these results greatly undercount the portion of people in jail who receive mental health services in the community. For example, if we look at all of the people booked in the Allegheny County Jail in 2017 and examine their previous publicly funded mental health services (back to 2002), we find 68 percent received services Second, we are only counting publicly funded mental health services, so any commercial services remain uncounted.

[34] Two queries were executed. First, we counted the number of distinct parents with open child welfare cases during 2017 who also received publicly funded mental health services during 2017. Second, we counted the number of distinct parents with open child welfare cases during 2017. Finally, we divided the result of query one by the result of query two. These results only include people receiving publicly funded mental health services, so any commercial services remain uncounted.

[35] Two queries were executed. First, we counted the number of distinct individuals who were both serving sentences during the period and received homeless services and supports during the time period from March 4, 2017, to August 20, 2017. Second, we counted the number of distinct people who were serving sentences during the period from March 4, 2017, to August 20, 2017. Finally, we divided the result of query one by the result of query two. When considering these results from a policy, rather than purely technical, perspective, we should note that people cannot be simultaneously receiving homeless services and be incarcerated; these results greatly undercount the portion of people in jail who receive homeless services in the community. For example, if we look at all of the people booked in the Allegheny County Jail in 2017 and examine their homeless services (back to 2013), we find 13 percent received services.

[36] Two queries were executed. First, we counted the number of suicide victims reported by the Medical Examiner's Office during 2017 who also received publicly funded mental health services during 2017. Second, we counted the number of distinct suicides during 2017. Finally, we divided the result of query one by the result of query two. These results only include people receiving publicly funded mental health services, so any commercial services remain uncounted.

[37] *See:* Federal Data Strategy. Available at: http://strategy.data.gov.

BIPARTISAN POLICY CENTER

The Bipartisan Policy Center is a non-profit organization that combines the best ideas from both parties to promote health, security, and opportunity for all Americans. BPC drives principled and politically viable policy solutions through the power of rigorous analysis, painstaking negotiation, and aggressive advocacy.

bipartisanpolicy.org | 202-204-2400
1225 Eye Street NW, Suite 1000
Washington, D.C. 20005

@BPC_Bipartisan
facebook.com/BipartisanPolicyCenter
instagram.com/BPC_Bipartisan

BPC POLICY AREAS

Economy

Education

Energy

Evidence

Finance

Governance

Health

Housing

Immigration

Infrastructure

National Security

BIPARTISAN POLICY CENTER

1225 Eye Street NW, Suite 1000 | Washington, D.C. 20005
202-204-2400| bipartisanpolicy.org