Actuate

DataSafes
Data Sharing with End-To-End Privacy

Actuate is developing a program called DataSafes to research and demonstrate a privacy-preserving, multiparty data sharing system that enables researchers to analyze multiple private datasets while allowing data owners to maintain privacy guarantees. The program's objective will be an end-to-end framework allowing for rigorous protection of privacy during the entire cycle of statistical analysis from data cleaning/linkage to model discovery to the generation of results from discovered models. Our approach is to adapt automated model discovery and data linkage methods to privacy-preserving computation and couple the execution of discovered models with management of output privacy.

Actuate will execute the DataSafes program by funding and managing expert project teams at universities, companies, and nonprofits to perform the applied research, prototyping, and application demonstrations required to achieve the program's goals.

The opportunity

Private health, business, and government records hold information with the potential to drive critical new discoveries in the economic, social and life sciences. Data-informed research may show us how to better educate children, avoid disease, reduce crime and incarceration, improve public services, and much more. In settings where such data is accessible, we have learned about the potential impacts of social context during childhood on college attendance and long-term incomes [1], the long-term economic impacts of universal broadband internet access [2], outcomes associated with universal child care programs on crime and well-being [3], long-term cardiac and diabetes outcomes associated with bariatric surgery [4] [5] and the effects of government funding on doctoral research education [6].

In all of the cases described above, privileged access to administrative data was granted to a small set of highly vetted researchers. Unfortunately, this kind of research is rarely possible because administrative data, collected from individuals or entities, are rarely available to researchers due to privacy and sharing restrictions. More generally, data containing the identifiable information of individuals and their behaviors are usually controlled by multiple owners (government, business, health providers, etc.) who cannot share with each other, let alone with researchers due to privacy risks.

Access to administrative data is rare and the use of these data alone cannot address many important research questions in the economic, social and life sciences. Many of the most important economic, social and health questions facing our society require understanding of both publicly held administrative data and privately owned data about individuals or businesses. Consider three examples of research questions that we cannot easily address today due to limited data sharing:

1. The effect of startup tax and incentives policy on long-term small business outcomes: while a number of incentive and tax policies have been proposed to help create small businesses, it is hard to estimate their long-term impact. This is largely due to the inability to track and measure businesses who have received these incentives over time as the data on their performance is held by private data aggregators (e.g. sales and debt ratio metrics owned by companies like Intuit and credit rating agencies) or in federal tax records while data on incentives is often owned by state and local governments.

2. The effect of economic incentives on long-term productivity and health outcomes: while short-term survey studies have assessed employee wellness incentives through survey [7], the ability to track

incentives and long-term outcomes requires linkage of employment and health records owned by private and public entities.

3. Controlled study comparison of credit availability schemes on long-term poverty outcomes: while microfinance measures have been assessed via randomized controlled trials (RCTs) [8], direct comparisons of micro-credit, tax incentives, and other mechanisms have not. Ideally, these experiments would be conducted with multi-arm RCT-like designs that make use of long-term financial data for follow-up. Unfortunately, these data are often held by financial institutions and are not linked to intervention data from governments and charities.

Our program attempts to make answering these kinds of research questions possible by reducing the risk of sharing sensitive data. Because we propose to automate large parts of the data preparation, annotation, cleaning and modeling processes, we believe that DataSafes could also result in more rigorous findings that are easier to replicate.

In such economic, social and life science applications, data is primarily used for statistical modeling and machine learning purposes. These purposes require a specific kind of computation to determine linkages between datasets, to transform/select/clean/extract data, and, ultimately, to perform statistical analyses on views of that data.

Existing challenges

While methods to perform privacy-preserving computation exist [9] [10], they have strong limitations and demand significant computational overhead. Extensions to handle multiple datasets owned by different parties, known as secure multiparty computation (SMPC), allow privacy-preserving computation without requiring trust between the different owners [11] [12] [13]. However, these methods also have strong restrictions on what can be computed and how taxed that computation is.

Furthermore, these methods require users to know exactly what they want to compute and, when that computation occurs on multiple datasets, how those data are linked. These methods assume that the linking and model discovery process can be done by people and on compute infrastructure trusted by all parties (i.e. a shared enclave with trusted data scientists to curate joins and analyses). In practice, finding both people and facilities trusted by all parties is the fundamental limiter for researchers accessing the data they need [14]. Typically, the process of discovering a modeling pipeline and data linkages requires experts-in-the-loop iteration with both subject matter and statistical/data expertise.

Even when an entire linkage and modeling pipeline (in the form of a program) is known, executing these pipelines is generally not straightforward and often not possible with existing privacy-preserving computation methods due to strong limitations on both the kind of compute and the efficiency with which the compute can be performed.

Privacy-preserving computations of this sort also require management of output privacy because multiple outputs from different programs operating on the same datasets can result in cumulative privacy leakage. While there are methods to prevent deanonymization threats, they require an infrastructure for tracking data queries that is both trusted and tied to data-owner-defined privacy policies. Such systems don't exist today outside of computer science research labs.

Simply put, today's limitations prevent data owners from sharing their data with each other or with researchers easily. Moreover, the added trust needed between data owners makes research requiring multiple datasets very difficult to perform. The best examples of multi-dataset research are supported by secured data lakes where data owners agree to provide their data to a trusted party who either (1)

authorizes trusted researchers to have limited access to perform analysis or (2) provides data scientists who work with researchers separated by a privacy wall (a data intermediary model). Yet the number of legal agreements and trust relationships required to make multiple datasets available in these lakes is very high (exponential in the number of participants). As a result, such systems are limited in scale, and the additional overhead of analyzing the data either through intermediaries or trusted access is inefficient at best.

Ultimately, what is missing today is an end-to-end framework that combines ways to preserve privacy during the model discovery process with both privacy-preserving computation and management of output privacy.

Our program

We propose to build a prototype of a privacy-preserving, multiparty data sharing system called DataSafes that tackles these issues. DataSafes will allow researchers to analyze multiple private datasets while allowing data owners to maintain privacy guarantees. This system will enable:

A. **linkage** of private datasets including complex joins;

B. **discovery** of data cleaning and transformations, featurization operations, and variable selections;

C. **modeling** of complex tabular and unstructured data; and

D. **maintenance of output privacy** for export of analytical results

while never exposing the raw (and potentially sensitive) data to the researcher.

The goal of the DataSafes program is to enable subject matter experts to accomplish these operations efficiently while preserving privacy, thus eliminating the need for trusted data intermediaries who may not have the subject matter expertise to do functions A-C well. To do this we propose to semi-automate the steps described above. Instead of data intermediaries working with subject matter experts separated by a privacy wall, we will use automated data cleaning/linking, machine learning and statistical analysis techniques to propose models that can be curated by subject matter experts using synthesized data to diagnose model performance.

For such a system to be adoptable and maintainable, we will create an open-source platform and a prototype service that allows trustworthy hosting of encrypted data and computation without requiring data owners to acquire highly technical private/distributed computation expertise. If we are successful, data owners will be able to publish encrypted forms of their data for outside parties (including researchers) to analyze and use through a secure computation infrastructure (see Figure 1).

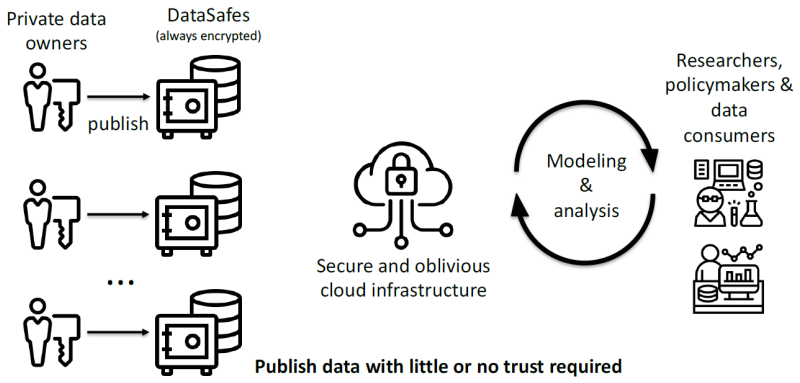

Figure 1: DataSafes computational infrastructure

Research approach

The DataSafes program will be a three-year effort to research, prototype and demonstrate a storage, compute and analytical modeling infrastructure that allows multi-dataset analysis while preserving privacy. The program will engage in applied research activities to develop efficient privacy-preserving compute; safe storage for private data; and algorithms for privacy-preserving linking, analysis and modeling of data. The program will use these component technologies to build a large-scale prototype system that is capable of supporting real-world problems. In order to evaluate and demonstrate DataSafe capabilities, the program will work with a number of application partners to define challenge problems and corresponding datasets for annual testing of the prototype as it develops. To ensure that this testing accounts for real-world threats, the program will employ an evaluation red team to penetration test the prototype implementation. If the prototype system is successful, users will be able to derive valuable information from private data while maintaining privacy protection for data owners. The structure of the program and its notional schedule is shown in Figure 2.

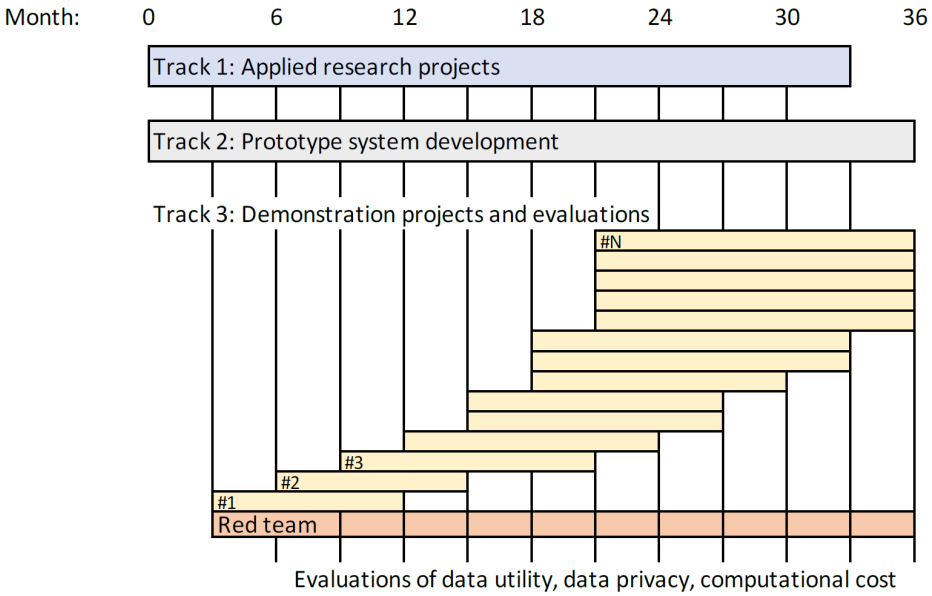

Figure 2: DataSafes program design and schedule

Privacy-preserving data and compute infrastructure

At the core of DataSafes is the ability to host private data and enable secure computation on that data. DataSafes will build a distributed storage infrastructure that encrypts data at rest and without provider access to private encryption keys. This allows storage of data on potentially untrustworthy systems (like commercial cloud infrastructure). Storage is coupled with secure compute infrastructure that allows efficient computation on encrypted or partially encrypted data through fully homomorphic encryption (FHE), partially homomorphic encryption (PHE), federated/private learning and secure MPC methods. Since no single method for secure computation is optimal for all algorithms that DataSafes may need to support, different processes (such as record linking and modeling) will select the optimal implementation. Storage and compute must also be orchestrated in secure ways to ensure that the exchange of encryption keys (when necessary) is secure and auditable [15]. Similarly, the underlying compute infrastructure requires that all user-facing operations maintain output privacy.

In order to facilitate the scalability and adoption of DataSafes, storage and compute infrastructure will be deployable to commercial clouds that support hardware-rooted trust [16]. Also, to this end, we anticipate licensing of DataSafes to be open-source and non-viral. If the full program is successful, commercial cloud providers will be able to host DataSafes to offer privacy-preserving computation/storage as a service.

Automated, privacy-preserving linkage/join of multiple datasets

To compute over multiple datasets, it is often necessary to align records through a linkage (or join) between datasets. Consider two datasets, A and B, each containing records. A join is a function mapping each record of A to zero or more records in B where the records of A and B share some common key (the join key). Often the join key is directly expressed in the records of A or B but is computed as shown in **Figure 3**. The manual process of discovering these join key generating functions requires extensive knowledge of both datasets and their mapping. Hence, the process requires owners of A and B to trust a data scientist and possibly subject matter experts with access to data from A and B. In practice, this trust becomes the critical bottleneck to data sharing as the number of datasets expands.

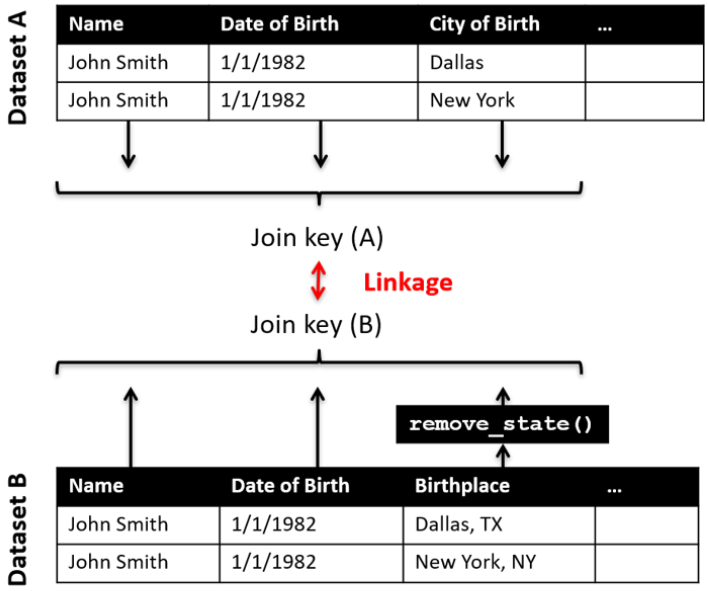

Figure 3: Linkage between datasets

We propose to semi-automate and fully protect this process through automated join discovery methods coupled with privacy-preserving curation, as shown conceptually in Figure 4. Given two datasets A and B, automated linkage systems first attempt to discover potential linkages between A and B. These linkages are presented to users with anonymized data in small batches to help users assess the quality of the linkage. If none of the proposed join functions is satisfactory, users can author their own based on samples of anonymized data from both A and B. Crucially, the linkage and anonymization computations are done within a secure multiparty computation system that protects A and B from each other and from the linkage and anonymization algorithm. Moreover, the anonymized samples are measured in information leakage terms using differential privacy methods to prevent privacy loss during the iterative process.

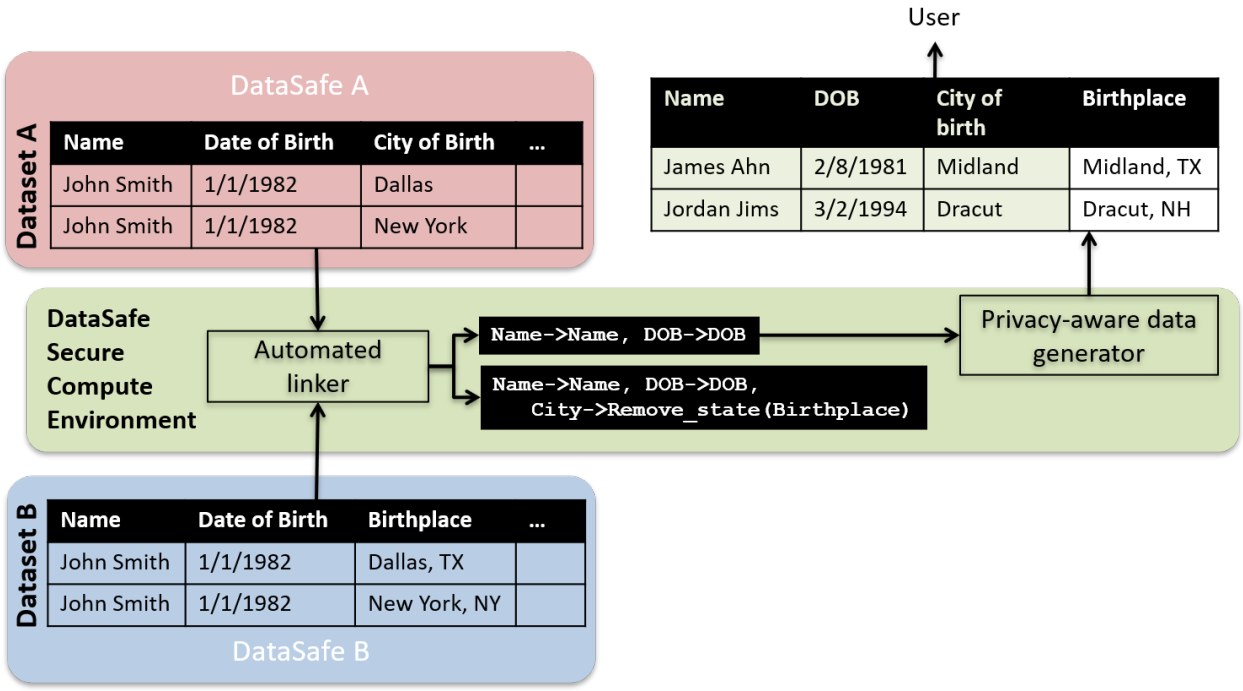

Figure 4: Privacy-preserving, semi-automated data linkage

The technological components of this system involve automated join, fully and partially homomorphic encryption, secure multiparty computation, differential privacy and data anonymization/generation. Recent research advances are encouraging and will need to be extended for the privacy-preserving context of DataSafes.

DARPA's recent program in Data Driven Discovery of Models (D^3M) was able to demonstrate automated clean-up and linking of datasets from noisy data sources. (Actuate's Wade Shen was the D^3M program manager at DARPA.) But this and other automated linking efforts assume that the inference of a join function or entity map has full access to each dataset being linked. In a privacy-preserving context, this is generally not true. Instead, limited correlations are computable and exposable across datasets. Methods that compute models of entities either through embedding [17], clustering [18] or via latent generative [19] approaches have the potential to allow linkage discovery without full access to data. Because of the restrictions of secure computation, it is necessary to develop methods for automated

join discovery and curation that can work with the communication and computation overhead of SMPC and FHE/PHE-based computation.

DataSafes will drive additional research that will make it feasible to achieve:

1) Fast and approximate correlation analysis for secure multiparty computation systems: The heart of most entity resolution and join systems is a correlation between datasets (and subsets of their columns). Linkage systems assume that these operations are tractable and cheap, but in privacy-preserving compute environments, they are not. Hence, we need to extend the class of efficient privacy-preserving compute kernels to include fast and approximate correlation. We hypothesize that this is possible through statistical sketching techniques as described by Mueen et al. [20].

2) Automated join techniques that are aware of secure multiparty circuit limitations: Current techniques assume that the cost of cross-dataset computations is free, and they make no effort to minimize communication during computation. Unfortunately, SMPC and PHE/hybrid realization of multiparty compute makes cross-dataset queries extremely expensive. While some SMPC compilers can minimize costly communications [21], they cannot do this fully automatically. Optimization is required to make these algorithms tractable for large datasets.

3) Generalized data generation: While techniques based on generative adversarial networks (GANs) have been successfully used for privacy-preserving synthesis, they are currently limited to binary, count, categorical and time series data [22]. Extensions to GAN- and DBM-based methods to wider population data types will be required to enable curation of possible joins.

Automated, privacy-preserving statistical modeling of multiple datasets

As described above, to realize end-to-end privacy, research insights from multiple datasets must be obtained without exposing researchers to identifiable information contained in these datasets. The prior section explores the problem of linking records across datasets in privacy-preserving ways. Here we describe how to statistically model linked data without compromising privacy using similar human-in-the-loop techniques.

To do this we use automated machine learning methods that are able to automatically propose statistical models given a modeling objective and a set of data. These approaches have proven to be effective on many different types of problems [23], and DARPA's D^3M program demonstrated that automated modeling and automated data cleaning can be coupled with human curation to improve modeling outcomes as summarized in Figure 5 [24] [25].

Problem	% accurate or root mean square error		
	Expert model - baseline	Fully automated model	Non-expert curation of automated model
Thyroid cancer diagnosis	85%	94%	94%
Baseball hall of fame prediction (5 year)	69%	65%	71%
Infrared spectroscopy star type prediction	17%	31%	30%
Classification of urban sounds	49%	82%	87%
Facebook relationship prediction	63%	67%	92%
Crop yield prediction	9.94 RMSE	7.52 RMSE	6.6 RMSE

Figure 5: Results from the DARPA Data-Driven Discovery of Models (D^{3}M) program

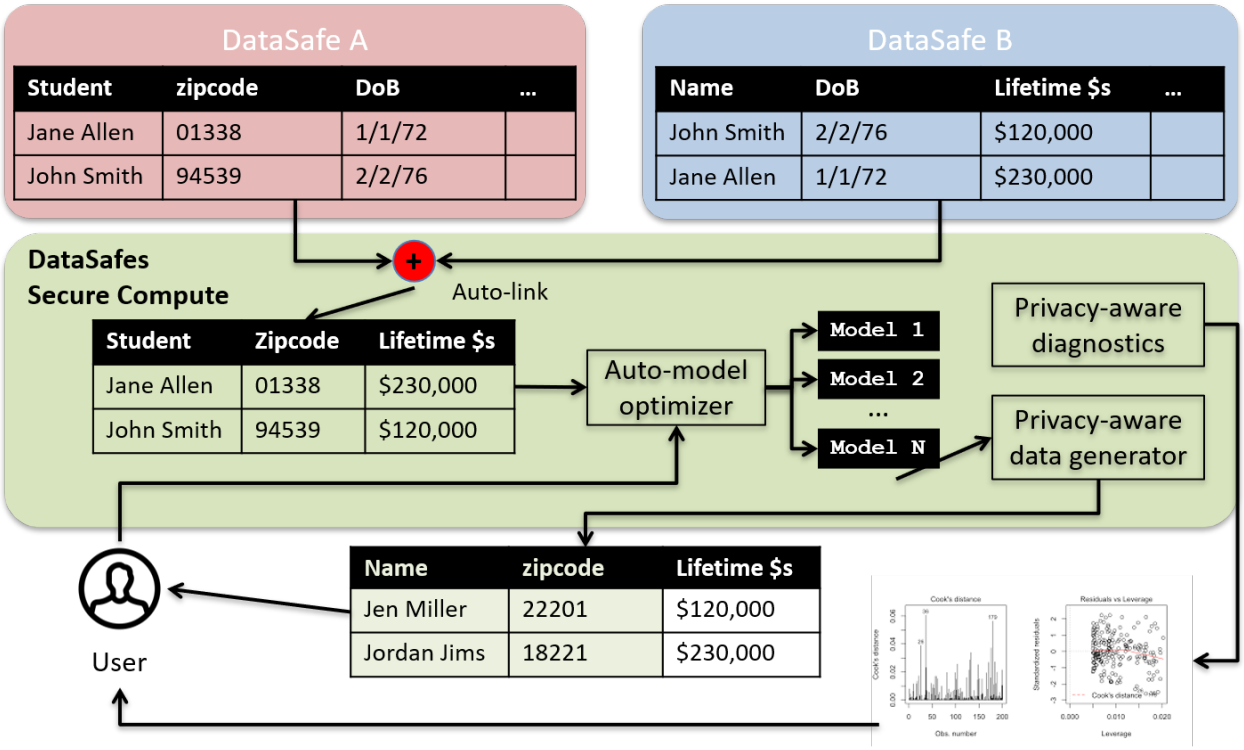

Figure 6: Privacy-preserving statistical modeling

For DataSafes we employ D^{3}M-like modeling approaches but coupled with secure computation, synthetic data generation and privacy-preserving model diagnostics as shown in Figure 6. This approach requires that:

1) Model discovery and optimizations algorithms can be efficiently executed on secure compute infrastructure. Given an objective, automated machine learning systems search for the best models by iteratively hypothesizing improvements. This search is made tractable by learning metamodels, i.e. models that predict the kinds of models and cleaning operations that will be best suited for a given dataset [26]. Metamodels are then used to hypothesize good candidates which are then evaluated

against held out data and the process is repeated. Unfortunately, this process is extremely computationally expensive. For small datasets (100s-1000s of datapoints), these approaches often require many compute hours and for larger datasets these algorithms can take ~10^6 compute hours to converge [27]. In order to adapt these methods to a privacy-preserving secure computation platform, we need to maintain computational efficiency for both meta-learning inference and model optimization processes. We hypothesize that optimized implementations of gradient-based optimization methods on secure multiparty computation [28] and federated learning methods [29] could make this possible.

2) Model diagnostics can be computed while maintaining output privacy. Automated learning systems can learn to optimize spurious correlates [30] [31], but subject matter expertise is needed to validate and curate automatically derived models. During the D^3M program, we showed that human curation can result in better models. However, in the context of DataSafes, output privacy requirements place strong limitations on the kinds of diagnostics a user might have access to during the model optimization process. Given that these diagnostics are often computed as aggregate or group-wise statistics, we hypothesize that many of these diagnostics can be computed with low leakage cost. Recent work by Chen et al [32] suggest that statistical model diagnostics can be computed within an ε-differential privacy framework for regression models with minimal leakage.

3) Diagnosis of modeling and data issues can be done through some combination of these diagnostics and through synthetic data generation and automated anonymization. While model diagnostics are useful for diagnosing failures in the aggregate, there are often data anomalies that skew statistics and mislead automated machine learning approaches. These anomalies often require subject matter expertise to diagnose and repair. While many of the interactive systems for automated modeling allow for repair, they assume that subject matter experts have access to raw data for diagnosis which is not possible with DataSafes. To address this issue, we propose to use the same class of privacy-preserving synthetic data-generation techniques as described in the prior section to assist in diagnosis. We hypothesize that it will be possible to develop synthesis/anonymization methods that preserve anomalies, allowing subject matter experts to design mitigations/corrections.

Large-scale evaluation and demonstration of DataSafes with real problems

Making meaningful progress in the full program will require real-world challenge datasets that test these capabilities at scale. Actuate is engaging with researchers in social and life sciences and data owners from government, healthcare, and commercial sectors. From these engagements we intend to curate two or three testbeds where experimental questions and multiple private datasets intersect. Actuate is working with existing research data enclave owners to form partnerships that allow safe testing of nascent technologies to be developed in the full program. We are currently working with a network of administrative research data facilities (ARDFs) organized by Georgetown's Massive Data Institute/RDC and we anticipate development of other partnerships with both government and private sector data owners. During the full program, we will work with selected testbed sites to conduct human-in-the-loop experiments with subject matter experts using DataSafes systems.

Risks and potential outcomes

Primary technical risks for this program are rooted in uncertainty over: (1) the degree of possible improvement in computational overhead, data utility and privacy protection, and (2) the generalizability of the privacy techniques. Another risk relates to the program's ability to change minds about the widespread use of privacy technologies and the broader sharing of sensitive data with these protections in place. The program's numerous demonstrations and standards and certification processes are designed to address these adoption issues.

If successful, this program will make valuable personal data analyzable and private at the same time. This is the key to enabling pervasive applications and informing laws, regulations and practices for data and privacy. As privacy-protected data analysis expands, individuals, companies and agencies will be able to provide access to more and more valuable data with confidence that it will remain private while helping to solve major challenges.

Acknowledgment

The Alfred P. Sloan Foundation has generously supported the design of the DataSafes program.

References

[1] R. Chetty, N. Hendren and L. F. Katz, "The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment," *American Economic Review,* vol. 106, pp. 855-902, 2016.

[2] A. A., G. I. and M. M., "The Skill Complementarity of Broadband Internet," *Institute for the Study of Labor (IZA),* no. 7762, 2013.

[3] M. Baker, J. Gruber and K. Milligan, "Universal child care, maternal labor supply, and family well-being," *Journal of political Economy,* vol. 116, pp. 709-745, 2008.

[4] M. Sundbom, J. Hedberg, R. Marsk, L. Boman, A. Bylund, J. Hedenbro, A. Laurenius, G. Lundegårdh, P. Möller, T. Olbers and others, "Substantial decrease in comorbidity 5 years after gastric bypass: a population-based study from the Scandinavian Obesity Surgery Registry," *Annals of surgery,* vol. 265, pp. 1166-1171, 2017.

[5] L. Sjöström, K. Narbro, C. D. Sjöström, K. Karason, B. Larsson, H. Wedel, T. Lystig, M. Sullivan, C. Bouchard, B. Carlsson, C. Bengtsson, S. Dahlgren, A. Gummesson, P. Jacobson, J. Karlsson, A.-K. Lindroos, H. Lönroth, I. Näslund, T. Olbers, K. Stenlöf, J. Torgerson, G. Agren, L. M. S. Carlsson and S. O. S. Study, "Effects of bariatric surgery on mortality in Swedish obese subjects.," *The New England journal of medicine,* vol. 357, no. 8, pp. 741-752, 8 2007.

[6] W.-Y. Chang, W. Cheng, J. Lane and B. Weinberg, "Federal funding of doctoral recipients: What can be learned from linked data," *Research Policy,* vol. 48, pp. 1487-1492, 2019.

[7] D. Jones, D. Molitor and J. Reif, "What Do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study," *The Quarterly Journal of Economics,* vol. 134, no. 4, pp. 1747-1791, 2019.

[8] A. Banerjee, E. Duflo, R. Glennerster and C. Kinnan, "The miracle of microfinance? Evidence from a randomized evaluation," *American Economic Journal: Applied Economics,* vol. 7, no. 1, pp. 22-53, 2015.

[9] C. Gentry, "Computing arbitrary functions of encrypted data," *Communications of the ACM,* vol. 53, pp. 97-105, 2010.

[10] S. Goldwasser, Y. T. Kalai, R. A. Popa, V. Vaikuntanathan and N. Zeldovich, "How to run turing machines on encrypted data," in *Annual Cryptology Conference*, 2013.

[11] A. C. Yao, "Protocols for secure computations," in *23rd annual symposium on foundations of computer science (sfcs 1982)*, 1982.

[12] I. Damgård and J. B. Nielsen, "Scalable and unconditionally secure multiparty computation," in *Annual International Cryptology Conference*, 2007.

[13] Y. Lindell and B. Pinkas, "An efficient protocol for secure two-party computation in the presence of malicious adversaries," *Journal of Cryptology,* vol. 28, pp. 312-350, 2015.

[14] J. Lane, "Building an Infrastructure to Support the Use of Government Administrative Data for Program Performance and Social Science Research," *The ANNALS of the American Academy of Political and Social Science,* vol. 675, pp. 240-252, 2018.

[15] D. Boneh and M. Zhandry, "Multiparty key exchange, efficient traitor tracing, and more from indistinguishability obfuscation," *Algorithmica,* vol. 79, pp. 1233-1285, 2017.

[16] A. Baumann, M. Peinado and G. Hunt, "Shielding applications from an untrusted cloud with haven," *ACM Transactions on Computer Systems (TOCS),* vol. 33, p. 8, 2015.

[17] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute and V. Raghavendra, "Deep learning for entity matching: A design space exploration," in *Proceedings of the 2018 International Conference on Management of Data*, 2018.

[18] I. Bhattacharya and L. Getoor, "Relational clustering for multi-type entity resolution," in *Proceedings of the 4th international workshop on Multi-relational mining*, 2005.

[19] R. C. Steorts and others, "Entity resolution with empirically motivated priors," *Bayesian Analysis,* vol. 10, pp. 849-875, 2015.

[20] A. Mueen, S. Nath and J. Liu, "Fast approximate correlation for massive time-series data," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, 2010.

[21] N. Büscher, D. Demmler, S. Katzenbeisser, D. Kretzmer and T. Schneider, "HyCC: Compilation of hybrid protocols for practical secure computation," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018.

[22] C. Esteban, S. L. Hyland and G. Rätsch, "Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs," 8 6 2017.

[23] M. Feurer, A. Klein, K. Eggensperger, J. T. Springenberg, M. Blum and F. Hutter, "Auto-sklearn: Efficient and Robust Automated Machine Learning," in *Automated Machine Learning*, Springer, 2019, pp. 113-134.

[24] T. Kraska, "Northstar: An interactive data science system," *Proceedings of the VLDB Endowment,* vol. 11, pp. 2150-2164, 2018.

[25] D. Cashman, S. R. Humayoun, F. Heimerl, K. Park, S. Das, J. Thompson, B. Saket, A. Mosca, J. Stasko, A. Endert and others, "Visual analytics for automated model discovery," *arXiv preprint arXiv:1809.10782,* 2018.

[26] C. Finn, P. Abbeel and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017.

[27] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578,* 2016.

[28] *https://github.com/facebookresearch/CrypTen.*

[29] H. B. McMahan, E. Moore, D. Ramage, S. Hampson and others, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629,* 2016.

[30] M. Udell and A. Townsend, "Why Are Big Data Matrices Approximately Low Rank?," *SIAM Journal on Mathematics of Data Science,* vol. 1, pp. 144-160, 2019.

[31] C. S. Calude and G. Longo, "The deluge of spurious correlations in big data," *Foundations of science,* vol. 22, pp. 595-612, 2017.

[32] Y. Chen, A. Machanavajjhala, J. P. Reiter and A. F. Barrientos, "Differentially Private Regression Diagnostics.," in *ICDM*, 2016.

[33] C. A. Mattmann, S. Shah and B. Wilson, "MARVIN: An Open Machine Learning Corpus and Environment for Automated Machine Learning Primitive Annotation and Execution," *arXiv preprint arXiv:1808.03753,* 2018.

