Title

An Examination and Survey of Data Confidentiality Issues and Solutions in Academic Research Computing

Author

Peisert, Sean

Publication Date

2023-12-13

Peer reviewed

An Examination and Survey of Data Confidentiality Issues and Solutions in Academic Research Computing

November 11, 2020

v1.0 — Public Report

Distribution: Public

Sean Peisert[1]

[1] Community engagement/report lead, sppeisert@lbl.gov

About Trusted CI

The mission of Trusted CI is to lead in the development of an NSF Cybersecurity Ecosystem with the workforce, knowledge, processes, and cyberinfrastructure that enables trustworthy science and NSF's vision of a nation that is a global leader in research and innovation.

Acknowledgments

This report would not have been possible without the engaged community team that contributed thoughts and ideas to this report. We sincerely thank that entire team, including the following members, who indicated willingness to be identified as contributors:

- Thomas Barton, University of Chicago, and Internet2
- Sandeep Chandra, Director for the Health Cyberinfrastructure Division and Executive Director for Sherlock Cloud, San Diego Supercomputer Center, University of California, San Diego
- Mercè Crosas, University Research Data Management, Harvard University IT & Chief Data Science and Technology Officer, IQSS, Harvard University
- Erik Deumens, Director of Research Computing, University of Florida
- Robin Donatello, Associate Professor, Department of Mathematics and Statistics, California State University, Chico
- Carolyn Ellis, Regulated Research Program Manager, Purdue University
- Bennet Fauber, University of Michigan
- Forough Ghahramani, Associate Vice President for Research, Innovation, and Sponsored Programs, Edge, Inc.
- Ron Hutchins, Vice President for Information Technology, University of Virginia
- Valerie Meausoone, Research Data Architect & Consultant, Stanford Research Computing Center
- Mayank Varia, Research Associate Professor of Computer Science, Boston University

In addition, Reinhard Gentz, Lawrence Berkeley National Laboratory, participated in and contributed significantly to the community conversations that surfaced the material presented in this report. Jinyue Song, also of Lawrence Berkeley National Laboratory, contributed to the background research on privacy-preserving analysis methods discussed in Section 3.6.

Our sincere thanks to members of the community that offered feedback on this document, including Timothy Wright (University of Notre Dame), and the members of the Trusted CI team that contributed ideas to this document, including Andrew Adams, Robert Cowles, Jason Lee, and Anurag Shankar.

This document is a product of Trusted CI, the NSF Cybersecurity Center of Excellence. Trusted CI is supported by the National Science Foundation under Grants 1234408, 1547272, and 1920430. For more information about Trusted CI please visit: http://trustedci.org/. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Using & Citing this Work

Cite this work using the following information:
Sean Peisert, "An Examination and Survey of Data Confidentiality Issues and Solutions in Academic Research Computing," Trusted CI Report, initial release September 2020, revised November 2020. https://escholarship.org/uc/item/7cz7m1ws

This work is available on the web at the following URL:
https://www.trustedci.org/reports

Table of Contents

Executive Summary

In research academic computing, may be natural to emphasize data integrity over confidentiality. However, there are numerous categories of academic research that do have data confidentiality requirements, from research that is simply embargoed until a future publication date to research that contains industry-owned proprietary information or is subject to government regulation. The contents of this report are based on numerous community conversations with leaders in academic research IT, data librarians, computer science researchers, computer security professionals, and others with roles involving using or enabling the use of sensitive data in academic research. The report discusses challenges to conducting research on data that is in some way sensitive, and solutions that are being

used or could be used to address those challenges and enable the research to take place. Those solutions include technical solutions as well as administrative and procedural issues. The report concludes with recommendations to campuses on issues to consider in order to enable research on sensitive data while ensuring security and privacy as well as usability and usefulness of the environment hosting that data.

1 Introduction and Background

There are many reasons why confidentiality issues are present in scientific research, even in "open" (unclassified) science. These issues range from scientists seeking protections against being "scooped" on career-changing research; to commercial proprietary ownership issues; to legal and regulatory restrictions on exposing personally identifiable information, protected health information (PHI), student information, or national security; to cultural or social norms about the exposure of data. Recent increases in restrictions in the United States for national security purposes, such as Controlled Unclassified Information [CUI], and in Europe and a handful of U.S. States, due to the General Data Protection Regulation (GDPR) [GDPR], and laws such as the California Consumer Privacy Act (CCPA) [CCPA] have only increased the need for ensuring proper computing controls regarding data confidentiality. In cases where data is regulated or considered proprietary, scientists may be unable to obtain access to the data and/or even receive certain types of funding without the ability to ensure proper confidentiality protections.

Nonetheless, sensitive data is important and even prevalent in many scientific disciplines. Implementing security and privacy-preserving methods to enable use of that data is vital to advancing science and medicine and improving public policy [HHL+19]. In some of these cases, it is sufficient for scientists to rely on a well-maintained computing system with appropriate authentication mechanisms and access controls. Examples of where this is employed include many institutional high-performance computing resources and National Science Foundation HPC resources, such as members of the XSEDE consortium, and Department of Energy Office of Science HPC resources, such as NERSC, the ALCF, and the OLCF. It is also how commercial cloud computing environments (e.g., Amazon Web Services, Google Cloud Platform, Microsoft Azure) are typically set up and used for scientific computing.

In other cases, more stringent requirements are necessary, which is the focus of this report. These requirements include the information that many people would expect to be protected, including protected health information, such as that collected by academic medical centers, but also includes:

- social science information, notably including financial data;
- student data, protected by the Family Educational Rights and Privacy Act (FERPA);

energy-related data, including advanced metering information (AMI), electrical topology maps, and supervisory control and data acquisition (SCADA);
- underwater acoustical information;
- chemical and biological data, notably that related to genomic analysis, pharmaceuticals, advanced manufacturing techniques, and materials production; engineering analyses such as airflow dynamics or explosion analysis; and
- computer science and data analysis techniques, including advanced artificial intelligence research (notably for computer vision), quantum computing, and even computer and network traces showing end-user activity.

For data that is regulated, this may be due to needing to adhere to particular NIST standards relating to FISMA [Fism] or DFARS [Dfar], such as SP 800-53 [N53], for operating Federal computing systems, SP 800-66 [N66] for implementing the HIPAA Security Rule [HHSS], or SP 800-171 [N171] for protecting "controlled unclassified information" [CUI] on non-Federal networks. In many situations, scientists use data that may not be regulated but may be considered proprietary by the organization that collected the data, perhaps because it is key to the business model. In these situations, it is common to employ computing resources designed explicitly with increased security properties. It is common to have campus security and privacy officers involved in evaluating such computing resources, the protocols around using them, and vetting and signing data use agreements (DUAs).

Privacy-preserving computational techniques and technologies may also address the sensitive data issue: including data anonymization, differential privacy, secure multiparty computation, or homomorphic or functional encryption, which can also be useful in certain situations. Rather than securing the data more heavily, these techniques reduce the sensitivity of the data returned to the end-user, thereby potentially serving as either an alternative or an adjunct to building and using computing environments that implement rigorous NIST-like standards.

This report describes lessons learned from our experiences in speaking to people involved in decisions around the acquisition and use of sensitive data. These people include domain scientists who use scientific computing resources, the operators of campus-level secure enclaves, and security and privacy officers and policy experts in campus research information technology (IT), legal, and university library type environments.

Some of the points that we addressed with scientists and computing operators included discussing security and privacy constraints that currently exist with research data, why those constraints exist (e.g., regulated data, privacy concerns, national security concerns, and proprietary concerns), and which entity imposes those requirements. We also

discussed where constraints exist — on data analysis inputs, outputs of data analysis, or both, and restrictions on publishing research; and ways in which security and privacy constraints on sensitive data hinder science, e.g., by limiting access to necessary data, or in preventing access entirely (e.g., can use data only in a particular location/ or only through a limited set of tools)

We also dug into scientific computing workflows themselves to understand the interaction model with the data, the degree to which data is clean and curated, or dirty and messy, and the processes for cleaning messy data before real work can get done. Vitally for much of modern science, we also discussed collaboration external to scientists' institutions leveraging sensitive data and ways in which current environments, tools, and other restrictions hinder sharing and collaboration.

Finally, concerning specific advanced privacy-preserving solutions, we examined places where techniques other than secure enclaves might enable privacy-preserving analysis of data. Such analysis could reduce usability issues with accessing sensitive data, lower certain barriers to data access, and lower costs to institutions for maintaining specialized, secure computing environments.

This document is organized as follows: In Section 2, we present an overview of our findings. In Section 3, we describe a set of technical solutions currently employed by a variety of scientists conducting research on data with higher requirements to ensure the control and confidentiality of that data. These solutions range from operational security controls prescribed by NIST to a set of state of the art best practices for data privacy approaches. In Section 4, we describe administrative and procedural issues almost always paired with technical solutions. In Section 5, we conclude this document with a summary and a set of recommendations for scientists, research computing professionals, campus security and privacy officials, and other related personnel to consider the appropriate paths forward for their situations.

2 Overview of Findings

Our community conversations show that many areas within academic research involve sensitive data of some kind, requiring higher-than-normal protections. To that end, it is clear that the university's critical role is to provide standardized technologies, policies, and procedures at scale for the university environment to enable the use of sensitive data. Most universities have expertise in developing DUAs and also have expertise either in provisioning on-campus computing environments or in helping to identify alternative environments elsewhere.

At the same time, an overriding theme among those institutions who provide secure data computing and storage environments, or oversee legal agreements to accept data is that there exist many scientists engaged on their own in research with sensitive data, without involvement by university IT or legal staff but for which the university is legally responsible. In these situations, researchers often sign their DUAs on their own and set up computing environments that are often not appropriately secured.

This report discusses both the technical aspects of secure computing environments and the legal and procedural issues involved in accepting and using sensitive data.

3 Technical Solutions

In this section, we describe a set of technical solutions currently employed by a variety of scientists for conducting research on data with a higher set of requirements for ensuring the control and confidentiality of that data. These solutions range from operational security controls prescribed by NIST to a handful of state of the art data privacy approaches.

3.1 Campus or College-Operated Secure Enclaves

Most campus computing clusters, be they small departmental or college-level systems to campus-wide high-performance computing environments, are not designed with an eye to regulatory compliance typically required for processing legally-regulated data.

Campuses that do have computing environments designed for regulated or otherwise "sensitive" data have come about those environments in a variety of ways — sometimes campuses have made a strategic investment to build such environments, and other times, such environments have evolved out of smaller solutions built by individual PIs with a need for secure computing, or a collection of PIs who have pooled their resources to set up such an environment. Unlike "traditional" computing, however, most campuses do not subsidize computing systems for regulated data. Given this, despite the need of many PIs to have such environments, comparatively few exist in academic settings.

A non-exhaustive list of well-known universities with "secure computing enclaves" include Duke University [Duk], Indiana University [Ind], Purdue University [Pur], Stanford University [Stan], University of Chicago [Chi], University of Connecticut [Con], University of Florida [Flo], University of California, Berkeley, and University of California, San Diego [She]. Trusted CI has evaluated such environments, such as REED+ at Purdue University [Ada19] and UC Berkeley's Secure Research Data and Compute (SRDC) Platform [Sha20]. While many of these environments exist, they are still not especially common due to the

cost involved in both standing up and maintaining them, and the expertise required to do so. We note that most secure computing environments are distinct from more general-purpose campus computing environments, some of the environments, notably Indiana University's, use the same high level of security regardless of the data involved to avoid duplicative hardware, software, and the necessary duplicate administration of multiple sets of hardware. In other cases, there are even more than two tiers, to enable controls specific to certain regulations. For example, one university has a campus computing environment for "fully public" data, a second computing environment for data covered by the HIPAA Security Rule, GDPR, and CCPA, and other personally identifiable information, and a third computing environment for Controlled Unclassified Information (CUI), which includes data that is "official use only" or is covered by U.S. International Traffic in Arms Regulations (ITAR) and is therefore export controlled.

While most of these campus-level computing environments are run within the auspices of a campus-level Research IT team, given the particular security requirements derived from the laws protecting the sensitive data that make these systems distinctive from "fully public" environments, these systems often have dedicated teams of people who build, maintain, and administer these specialized systems, often making them quite expensive to operate — dedicated hardware, dedicated IT staff with a U.S. Citizenship requirements and a high degree of specialization, often significant documentation and audit requirements pertaining to the regulations that govern the use of the systems, and even stronger-than-normal physical protection requirements for the buildings and rooms in which these systems reside.

Common threads of these environments includes additional physical security of the computing machinery, virtual private networks with end-to-end encryption over networks, encryption of data at rest, two-factor authentication (typically using physical tokens, not just "soft" tokens), remote desktop to access the compute environment itself, including disabling cut/copy/paste operations, "airlocks" with two-person rules to get data or software into or out of the system, rigorous access controls, data deletion policies, and strict isolation or separation of processes — that is, two different users cannot execute processes on the same nodes at the same time. Furthermore, it goes without saying that all transfers of sensitive data in or out of the enclave must be encrypted as well. This can be enabled at small scale with scp / sftp or at large scale leveraging encrypted data transfer functionality in Globus's GridFTP. This list is not intended to be exhaustive by any means, but to provide a flavor of the kind of expertise that is needed. In some cases, additional techniques, including secure boot, or encrypted virtual machines were used.

For cases requiring maximum security, some “secure compute” environments were described as being entirely offline, locked in an electronically keyed and tracked room, and not connected to any network.

Some campuses, such as University of Florida and Indiana University, leverage the same high-performance computing systems for both "sensitive" and "non-sensitive" computing. This can have significant advantages both in "raising the bar" for security of the non-sensitive computing environment, and also reducing duplication and complexity involved in maintaining two systems. Indeed, the duplication can be significant, involving not just having two or more sets of computing hardware and the people who administer it, but also requiring additional machine rooms, power, and more. Potential downsides of shared sensitive / non-sensitive infrastructure include usability and cost. The additional layers of security required for computing systems hosting sensitive data add cost. If there is a roughly similar or even greater amount of computing on sensitive data than on non-sensitive data, then a shared system might make sense. If the scope and scale of non-sensitive computing research vastly outstrips computing on sensitive data, then it could cost more to make an HPC system meet the security requirements for sensitive data than it might to add a much smaller, second system. Similar issues arise with usability and whether it makes sense to subject all users to the same (high) requirements.

Campus-provided computing enclaves can help with improving security and, of course, economies of scale over PI-managed solutions. In line with this, they can also enable larger systems that can handle larger size datasets and computations. To build such environments requires significant campus resources beyond a little overhead from small research grants in order to stand up such an environment and operate it. Some campuses resist committing such resources. Others see a critical mass needing such resources or simply “build it” hoping “they will come” — in most institutions, the need seems either to be there or grows into the infrastructure. Institutions with secure computing enclaves seem not to regret constructing and operating them, and indeed, such environments seem typically to end up being the basis for huge amounts of future funding, given that very large new doors are opened through their availability.

However, performance is not always paramount and not all data is “big.” Sometimes all that is needed is a compliant desktop. Indeed, some data transfers do not even use networks at all but move data by DVD. Even in this case, however, campus oversight is critical to track where such environments exist, how they are developed and maintained, and what sensitive data is present.

As we will discuss in later sections, some universities that do not provide campus or college-level solutions for secure computing, and may require researchers to seek alternate

environments. In some cases, this rationale may simply be due to lack of resources. In other cases, the rationale may be due to (or also due to) the simple desire for the university not to assume liability that comes with taking responsibility for hosting the data. Alternate environments that researchers may need to use might include shared cyberinfrastructure, commercial cloud computing environments, solutions stood up by individual PIs, external environments provided by the data provider themselves, or other third-party solutions. In the subsequent sections, we describe these environments and what some of the pros and cons of using them might be.

3.2 Cloud Computing Resources

Given the startup costs involved in acquiring, configuring, and hosting "secure computing" resources for sensitive data, and also the liability involved with hosting data in a university-hosted environment, some campuses leverage cloud computing environments instead. In general, the cloud computing environments referred to in this case are the ones that are rated for storing data covered by HIPAA, DFARS, or FISMA. The file-sharing platform, Box, as well as all three major cloud computing platforms (Amazon AWS, Google Cloud Platform (GCP), Microsoft Azure) have HIPAA-Compliant infrastructure and services, although in order for a customer to leverage those portions of the platforms, specific Business Associate Agreements (BAAs) are required. When FISMA or DFARS compliance is referred to, it is typically in the context of FISMA "moderate" or "high impact" data and likely refers to access via specialized cloud environments, including FedRAMP, including AWS GovCloud, and similar services from Google and Microsoft, that have instances only in the United States and are hardened and audited in a variety of other ways to comply with U.S. government regulations.

On the surface, cloud computing can seem like it would shift liability to the cloud provider for securing data both for storage (Box, AWS, Azure, GCP) and computation (AWS, Azure, GCP). In some cases, this can partially be true. For example, in many Software as a Service (SaaS) applications, responsibility for the operating system and software security patching is delegated to the cloud provider, reducing much of the security configuration that end users are responsible for to the access control settings for the software. However, SaaS is not the most frequent use of cloud computing — many scientific computing users are instead leveraging Infrastructure as a Service (Iaas), where the responsibility for managing the server operating systems and the software running on those systems is the responsibility of the end user. In this situation, the liability for security is nearly entirely still the responsibility of the end user, other than the physical and architectural security for the underlying cloud platform.

Therefore, cloud computing is clearly not a panacea for many reasons. While it can reduce or eliminate the need for personnel with local expertise in operating physical systems, and data center costs in keeping the disks spinning and the lights on, it does not necessarily improve security at all, and itself comes at significant cost of storing and computing data, and, importantly, getting any derived data out of the cloud environment again, after processing has taken place. It also requires specific expertise in using cloud computing environments, which can also be harder to find that expertise in traditional UNIX and Windows server administration, given its relative newness and the speed with which it is evolving and changing.

It is also possible to implement a system that involves a hybrid on-premises / cloud environment to find a balance in the the tradeoffs between each solution other than all or nothing. For example, the cloud environment might be used for long-term data storage and scalability, but the local environment may have superior usability. Such solutions can alter the cost considerations somewhat but then require hiring personnel with expertise in both server administration and also cloud administration.

As we discuss in a subsequent section, there are solutions, including cloud-based solutions, that can enable IaaS cloud computing while reducing liability for the end user. These solutions come with an even higher cost, but can be appropriate in certain circumstances.

Finally, we note that cloud computing is not currently universally permitted as a solution for storing and computing regulated data. For example, the U.S. Department of Defense permits the use of AWS GovCloud, but the U.S. Department of State currently does not.

3.3 PI-Managed Solutions

Sometimes PIs build their own solutions for handling sensitive datasets, particularly when campuses don't provide solutions. It can also come up when campuses do provide computing environments, but PIs either don't know about those environments or choose not to use them, for reasons that might include elements such as cost, control, performance, or usability. In either case, PI-built and managed solutions are particularly common in disciplines that have the traditional capabilities to manage computing systems, such as computer science and other physical sciences and engineering disciplines.

While it is not impossible that individual PIs or academic departments could be capable of managing their own such systems in a responsible way, when PIs do so, it tends to suggest that the campus isn't involved at all. As we discuss further in Section 4, this means that the campus likely has not reviewed and approved any data use agreements, had any review on the design and implementation of the computing environment, nor had any input on the

policies and procedures in place. As a result, the campus is implicitly taking on substantial risk of regulatory penalties should a data breach occur, and without its knowledge. HIPAA even has a category of penalties for "did not know" (that the campus had PHI).

3.4 Data Provider-Managed Solutions

There are situations in which researchers leverage neither a campus solution for sensitive data, nor their own solution. In some cases, campuses don't want any sensitive data on campus at all, and require researchers to find external solutions, such as those provided by the data provider itself. In addition, there are times when some data providers, notably those from private industry, want data analysts to use systems provided by the company providing the data. This situation is often true in the defense, finance, and pharmaceutical industries where the sensitivity relates to intellectual property, and not "just" government regulations.

3.5 Third-Party Secure Enclaves

As we have discussed, not every campus has the resources to set up a secure computing environment on its own -- particularly outside of the R1, PhD granting institutions. One alternative to campus computing solutions, data-provider-provided solutions and typical FedRAMP cloud computing environments, is a managed cloud environment such as that run by the San Diego Supercomputer Center (SDSC) at the University of California, San Diego, called "Sherlock" [She]. While Sherlock began as a local computing environment at SDSC for certain FISMA and HIPAA categories, it moved into the cloud as the desire for reducing local computing footprint and more elastic computing scalability increased. In addition, the Sherlock service is available outside of UCSD — any U.S. academic institution can partner with SDSC/Sherlock to gain access to these services.

Sherlock provides what many cloud environments do not provide. That is, it does not just give an environment located in the United States that is rated for compliance with regulations, but actually provides managed solutions and platforms to implement NIST SP 800-53 (e.g., vulnerability scanning, OS management, encryption, physical security, segmentation, documentation), and secure applications running within the environment. At the same time, as one might expect, given that SDSC has "invested millions of dollars on infrastructure, software, and personnel" [Shep], the use of Sherlock comes with a cost, which can run into the tens or hundreds of thousands of dollars per year for the managed service and security compliance alone [Shep] depending on the number of VMs used, complexity of the project scope, and the degree of management and customization desired — and the compute and storage costs for the commercial cloud provider are added on top of the managed services fees.

"Skylab" [Sky,Skyd] is an SDSC solution that provides the core of the Sherlock solution in a customer owned/managed capacity. As with the original Sherlock service, Skylab is also available beyond UCSD as well. As described by the SDSC Sherlock team, end users desiring the ability to manage their own cloud environment, but still wishing to leverage a well-designed, easy to use software framework for cloud computing, may instead wish to use the Skylab software product. End users can either fully manage the deployment or engage SDSC Sherlock team to provide support. As with other aspects of cloud solutions, this choice has tradeoffs between cost, control, and security responsibility and liability.

An alternative to secure computing facilities at every individual campus, without going all the way to leveraging cloud environments (including managed environments, such as Sherlock), is cyberinfrastructure for secure computing that is shared among multiple institutions. An example of such a facility is Virginia ACCORD — an NSF-funded, HIPAA-compliant, secure computing environment managed by the University of Virginia and originally made available to a consortium of eleven public universities in Virginia [ACC]. ACCORD focuses on addressing policy and regulatory issues around data collection, management, and sharing. Most recently, through another NSF grant, and in order to support broader access to secure computing resources for COVID-19 research, UVA established a national ACCORD-COVID program with the goal of making secure computing resources available to all COVID-19 research projects funded by NSF [ACCC]. As of this writing, NCSA is also in the process of seeking security certification for its "Advanced Computational Health Enclave" that would enable processing of PHI [NCS].

NSF has long supported high-performance computing facilities, such as NCSA at UIUC, SDSC at UCSD, TACC at UT, and PSC at CMU, to make computing broadly available to NSF-funded science. However, specialized secure computing facilities for sensitive data, such as ACCORD and ACCORD-COVID, are currently rare, if not even unique. They may provide a model for other regional and national computing consortia going forward, as well.

There are also commercial "Managed Service Providers" (MSP) that provide compliant environments and consulting on how to use them, but are typically quite costly, and, as a result, beyond the reach of many institutions, especially those with budgetary challenges..

3.6 Alternatives to "Higher Walls"

A variety of techniques and technologies exist that can reduce the sensitivity of data and therefore reduce the need to use campus resources to build, protect, and maintain "secure computing enclaves." We discuss several such techniques at a high level in this section. These techniques will also be discussed in greater detail in an appendix to this document, planned to be released later in 2020.

We note that while several of the techniques we discuss in this section are interesting and powerful options that are or are becoming best practices in commercial industry, it is also important to note that they are not silver bullets. Each technique has its own distinct set of tradeoffs, including numerous vital implementation details that are as easy to "get wrong" as configuring and maintaining the secure computing environments we have discussed earlier in this document. Employing these techniques does not absolve institutions of their responsibility for protecting data. At the same time, several of these options represent compelling tools in the toolbox for institutions to consider as options for protecting sensitive data.

3.6.1 Data De-Identification

A variety of approaches exist to reduce the sensitivity of a dataset simply by removing the parts of the dataset deemed to be most sensitive. One way of reducing the sensitivity of data regulated by the HIPAA Privacy Rule [HHSP] and the HIPAA Security Rule [HHSS], known as "Safe Harbor," [HHSG] is to remove the 18 identifiers designated as sensitive. These include the obvious identifiers, such as names, birthdates, and social security numbers and also include "geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes," health plan identification numbers, IP addresses, biometric identifiers, and so on. HIPAA also provides an "expert determination" option where a de-identification expert can analyze the data and provides documented evidence that the data is in fact de-identified, although this method is used much less frequently. One approach to achieving this method is differential privacy, which we will discuss further in Section 3.6.4.

It is worth noting that there are times when reducing the sensitivity of data can also reduce its utility. For example, if some type of research requires finding correlations between relatively precise locations or birthdates or IP addresses, removing that information can limit its usefulness. A great deal of clinical research involves fully-identified patient data, as a result.

A related approach from simply removing data fields is one that enforces generality. One such technique is called k-anonymity [Swe97]. A variety of other techniques in the same "family" of approaches also exists, including ℓ-diversity and τ-closeness. With k-anonymity, sufficient data is masked so that first, all "identifiers" are removed. Identifiers are data fields that uniquely identify an individual on their own, including full name, social security number, drivers license number, street address, etc… In addition, sufficient "quasi-identifiers" are masked so that the quasi-identifiers for each record looks like k other records. Quasi-identifiers are data fields that do not uniquely identify an individual on their own, but collectively can identify an individual if two or more are put together. So, for

example, birthdate, zip code, and gender could be examples of quasi-identifiers. Using *k*-anonymity, depending on the composition of the dataset, we may, for example, wish to mask the gender entirely, and perhaps not eliminate the zip code but at least remove the two least significant digits, and perhaps the day of the month of the birthdate. Only the "sensitive attribute" — that is, the key data field that we wish to protect against association with an individual — is left un-altered.

Simply removing data fields or performing *k*-anonymity clearly can help protect privacy. However, these approaches also reduce utility. And moreover, these approaches have been repeatedly been attacked by using statistical correlation and internal and external linkage attacks in order to de-anonymize data, thus defeating the goals of the anonymization [Swe97, NS08, SAW13]. While this does not mean the de-identification methods have no value, their value is also difficult if not impossible to quantify due to near-impossibility of knowing of all current and future external sources of information that can introduce possible linkage and correlation attacks [NF14].

Given this, de-identification seems risky to entirely rely on, and is perhaps best used in combinations with other technical, legal, and procedural controls. For example, it may be enough to remove the 18 HIPAA identifiers in a way that enables research use of the data (as opposed to simply clinical use) but still requires a certain set of technical controls, security and privacy training, and legal agreements on the use of the system and data.

3.6.2 Secure Multiparty Computation and Fully Homomorphic Encryption

Secure multiparty computation (MPC) is a technique that enables two or more parties that hold sensitive data to enable all that data to be computed collectively without revealing the sensitive data to any other data providers or external parties In its most basic form, framed originally as "Yao's Millionaire Problem," [Yao86], two millionaires are able to determine which millionaire is wealthier without either having to reveal their actual wealth to the other, or even to a trusted third party. This process can be expanded to arbitrary numbers of parties providing data.

Applications of MPC to enabling scientific research using sensitive data are being explored at a number of institutions. In practical terms, researchers from Boston University recently applied such techniques to track trends in salaries in the City of Boston, to evaluate progress on reducing pay gaps among workers of similar jobs that have traditionally been underpaid [VSG+19]. Using a secure multiparty computation framework, employers could submit data to a central location and statistics on pay disparity could be calculated without violating the privacy of individual workers or specifically calling out the employers who were not doing well — a key criteria, since the

goal was to simply calculate citywide progress, not reveal individual employers not making such progress.

Among other institutions leveraging MPC, the Infrastructure for Privacy-Assured CompuTations (ImPACT) project at RENCI is also exploring the use of MPC for supporting multi-institutional analysis of sensitive data [REN] leveraging the SPDZ/2 software libraries developed by researchers at the University of Bristol [KPR17].

Similar to secure multiparty computation, and sometimes even used as a basis for implementing it, fully homomorphic encryption (FHE) [Gen09, Gen10] is a technique that enables encryption over encrypted data without ever decrypting the data and without exposing that data to any party, including the system upon which it is being computed.

Both fully homomorphic encryption and secure multiparty computation have extraordinary promise for altering the threat models associated with computing. They do also, however, currently have two drawbacks. The first is that programs leveraging them typically need to be rewritten to use MPC or FHE libraries and/or recompiled with specialized compilers that implement the technique. The other drawback is performance. While both MPC and FHE have made significant strides in recent years, and are no longer always many orders of magnitude slower than cleartext computing as they originally were, they both still do have significant performance penalties, which are exacerbated depending on the complexity of the computation being performed, and with larger amounts of data being computed.

3.6.3 Trusted Execution Environments (TEEs)

Trusted execution environments (TEEs) are hardware elements in microprocessors (and sometimes memory controllers) that provide hardware-mediated separation from other processes on the system. Some TEEs can also provide protection against malicious hypervisors, malicious operating systems, and system administrators. In some cases, these systems even enable computing over encrypted data, like homomorphic encryption, but many times faster, due to the hardware support. This increases the ability to protect even against physical attacks against the data being computed on. An example of the simplest form of TEEs is ARM's "TrustZone" technology. Intel's SGX and AMD's Secure Encrypted Virtualization (SEV) family (which builds on AMD's Secure Memory Encryption (SME)) both enable computing while data is encrypted in memory. Each has very different use models and somewhat different threat models that they protect against. However, neither SGX nor TrustZone seem well-suited for scientific computing in the traditional sense, due to the fact that performance overhead is very significant, and neither are designed to run whole operating systems or even whole programs inside of them. Rather significant

program rewrites are typically necessary to leverage the APIs that enable use of the secure aspects of chips supporting SGX and TrustZone.

In contrast, AMD SEV is now widely supported in modern AMD EPYC CPUs and allows programs to run inside whole virtual machines with no program modification at all, and with very small performance overheads [AGA+20]. Amazon [Amz] and Microsoft's [Mic17] cloud environments both support Intel SGX, and Google [Goo20] recently announced AMD SEV support. Most recently, the Confidential Computing Consortium [CCC], currently composed of organizations such as AMD, ARM, Google, Intel, Microsoft, and Red Hat, among others, have issued white papers [CCC20a, CCC20b] outlining key benefits of the use of TEEs over traditional computing environments in terms of data integrity, confidentiality, and testability as well as programmability and performance. These white papers suggest that the need to avoid using cloud systems due to lack of trust of the cloud platform may be changing quickly. One might expect, given the general availability of TEEs that this technology will eventually trickle down to scientific computing at a campus level, though doing so may take a while due to the fact that leveraging TEEs not only requires hardware but the infrastructure to be built up around them.

In the meantime, in leveraging TEEs in the cloud, users and data providers have much stronger protections against malicious co-resident processes, and also no longer have to include Amazon and Microsoft in the threat model of potential malicious actors to protect against, at least with respect to data confidentiality. Particularly given the fact that SEV requires no change in software to leverage, it seems quite possible that the future of computing will broadly leverage SEV-like technology by default, and particularly in environments outside of the direct control of the end user (e.g., the cloud, HPC centers) to acquire the benefits of the stronger security model at very modest performance cost.

3.6.4 Differential Privacy

Releasing statistical information about data, rather than raw data, would seem like a tempting approach for maintaining privacy. However, as shown by Dorothy Denning, Hoffman and Miller [HM70], Schlorer [Sch75], and others, in the 1970s and 1980s, numerous attacks on statistical databases can end up revealing private information [DD79,Den82], including the notoriously thorny issue of trackers [DDS78].

Similarly, much has been made in recent years about "federated machine learning" in which a model is trained on distributed datasets without the raw data ever leaving the computing facilities of the institution hosting the data and training. Such a method has undeniable appeal for institutions wishing to maintain greater control over their data while making it available in some form for use. However, trained machine learning (ML) models

have been shown to be vulnerable, too, to model inversion attacks, external linkage attacks, and other methods [SS15] for revealing information about both the data used to train the model, as well as details about the structure of the model that expose more than the providers of the data on which the model is trained may feel comfortable.

Both of these situations, among others, can potentially benefit from applications of *differential privacy*. Differential privacy [Dwo06] is a statistical technique that guarantees that if an arbitrary calculation is made on a dataset that a data analyst should not be able to determine before or after any individual record has been added to the dataset if that individual record is present. Thus, the technique can provide strong statistical guarantees regarding the privacy of individual records.

Differential privacy is now a mainstream solution, with production use by Apple [App17], Google [ACG+16], Uber [Nea18], the U.S. Census Bureau [Abo18], and the United Nations [Uni19], among others, and even open source distributions from a variety of developers, including Google's general-purpose library for differential privacy [Goo19a], Google's TensorFlow Privacy to enable privacy-preserving machine learning [Goo19b], and IBM's differential privacy library [Dif]. Most recently, Harvard University and Microsoft, with funding from the Sloan Foundation, have embarked upon the creation of a coalition that will build and maintain a set of open-source differential privacy tools called "OpenDP." [OpDP] In addition to enabling privacy-preserving statistical queries, machine learning, and a variety of other analyses, differential privacy can also be used to generate synthetic datasets that enable analysis on data that looks like the original raw data and maintains most of its properties, but contains the same statistical guarantees regarding the ability to identify individual records as other uses of differential privacy does [BLR13].

In addition to the open source tools available that enable interactive differential privacy and non-interactive differential privacy (creation of differentially private synthetic dataset), a number of commercial organizations exist that offer services to perform the currently-extensive process of application and optimization of differential privacy algorithms to a given dataset and analysis applications.

4 Administrative and Procedural Issues

Having "sensitive data" is never simply a technical matter — legal issues are key to sensitive data, and procedures for engaging appropriate personal, as well as procedural methods for managing technologies, are vital. The most high-functioning universities typically have well developed policies and procedures in place that help researchers, campus IT, campus privacy officers, campus risk management, campus legal, and data

providers to all understand the risks involved in accepting data, the methods needed to manage risk, and the responsibilities of all parties involved to optimally manage that risk. For example, a high-functioning campus would never enable individual PIs to sign their own data use agreements, but would require review and approval by the office of the Chief Information Security Officer, and, depending on the campus and the nature of the data, potentially also the Chief Information Privacy Officer, Institutional Review Boards, campus research data leadership (e.g., as part of a campus library), and campus security and risk committee as well.

On the other hand some campuses do not have such policies. While campuses may think that this means that accepting sensitive data is implicitly or even explicitly prohibited, more likely than not, it simply leaves researchers wanting to accomplish their research to sign their own data use agreements and to come up with their own procedures. Taking matters in their own hands may feel like the only choice for researchers whose campuses do not have key policies and procedures, but it also leaves campuses significantly exposed to legal risks in the event of improper data handling, including a data breach. Campuses are ill-advised to simply have a blanket ban on accepting sensitive data, and are also ill-advised to ignore the need for sensitive data as part of scientific research.

Even where policies on sensitive data exist, barriers to efficient and effective handling of questions around sensitive data can vary wildly by institutions. The most effective institutions seem to have broad representation by stakeholders and significant campus visibility. Where there is a disconnect between campus Institutional Review Boards (IRBs), research IT, campus security and privacy officials, and researchers, policies and procedures can cause data issues to become lost in a quagmire. In such situations, not unlike when campuses do not provide leadership on technical solutions, researchers will once again avoid bringing in campus research IT at all, and instead formulate their own solutions, leaving campuses exposed to risk but without a voice in managing it. While campuses should be inclusive regarding representation, campuses must also be nimble — while controls for government-regulated data may follow similar rules year after year, commercial entities providing data may have significantly varying controls and legal provisions to protect sensitive data. Commercial entities will try to minimize their own liability contractually with universities, and the security, privacy, and legal entities can often have to be very creative to address the needs of the data provider while satisfying the needs of the research. It is not always possible to negotiate terms — the U.S. Government often has "flow-down requirements" that commercial entities are unable to relax.

Cost and campus-level will are not the only issue that can prevent campuses from developing computing enclaves for sensitive data. Sometimes the issue is simply political, and authority, including authority over IT resources, is distributed amongst schools or

colleges rather than a single campus-level entity. This is true even among the largest and most well known universities in the country. While there is clearly no single right answer for anything, most would point out that while college-level control can avoid acrimonious political fighting about requirements and costs between colleges, even with competent IT staff at the college level, such solutions can still leave a campus open to a morass of different policies that it is ultimately legally responsible for but has no control over.

Scientists from disciplines that have long dealt in sensitive data, such as medicine and certain social sciences can tend to take security and privacy policies and procedures most seriously. Safety procedures are part of every healthcare environment, and privacy is codified in the Hippocratic Oath. Disciplines in which data can be seen to "want to be free" such as computer science may tend to be less compliant, and have an attitude of "asking for forgiveness, not permission." It is the latter case that may be more inclined to pursue their own solutions and ignore standard procedures. Campuses would benefit by considering the range of expertise and historical attitudes that span the disciplines present in their environments. Few policies are universal across all domains, all disciplines, and all institutions. This also can make it challenging for researchers who move between institutions — what happens when a researcher moves from one university that has well defined security policies for sensitive data to another that does not? Or vice versa — how are researchers who move from lax environments to strict ones properly indoctrinated? How is data deleted after the researcher leaves? And how should access with more transient individuals (e.g., visitors, students) be handled? Few easy answers exist — this is the nature of a great deal of scientific research and is why these situations involve accepting and managing acceptable risk, not eliminating it entirely. An effective approach to consider is spending some effort understanding and enumerating researcher needs and use cases, and designing and offering solutions that have security baked in from the outset.

A data management plan (DMP) is a document describing the data that is acquired or produced during the course of a research project, and the ways in which that data will be handled both during the project and after the project has concluded. A DMP is theoretically an agreement between the PI and the sponsor of the research (in contrast to DUAs, which are agreements between the PI's institution and the data provider). Many or even most sponsored research projects, such as those funded by the National Science Foundation or National Institutes of Health, require DMPs. One might think that the requirement of submitting DMPs along with proposals would address many policies and procedures relating to sensitive data from the outset, before any data or funding ever starts following. However, this frequently turns out not to be the case — most DMPs submitted to funding agencies are either entirely wishful thinking or too high level to be actionable, and are rarely coordinated with campuses until after the proposal is funded, or not at all. Some campuses try to work with scientists to avoid DMP problems preemptively, before the

proverbial clock is ticking in order to accept the terms of grant funding or data access, but such institutions remain in the minority.

5 Discussion and Summary

"Sensitive" data is a part of numerous research domains, and is fundamental to at least some of them. At the smallest scale, at least some degree of campus involvement to enable academic research involving sensitive data is necessary to enable work in fields that otherwise cannot function well, or at all. However, taking this much further, by developing large-scale, campus-level efforts, with robust technical and procedural methods for handling sensitive data can open huge opportunities for campuses and its researchers across domains that may be unavailable without such efforts.[2]

However, there is no universal answer or silver bullet for protecting such data. "Sensitive" data may be stored on campus, in community-owned cyberinfrastructure, in government-sanctioned clouds, in third-party environments, in environments provided by the data provider themselves, or simply be made less sensitive through any of a variety of privacy methods. The criteria as to which of these environments and approaches should be adopted should consider the requirements of the data provider, resources (financial and technical) of the university, and the research methods of individual researchers. Each solution may have its own place, and as technology develops, the best solution for any given organization or researcher within an organization continues to evolve. Campuses must stay alert and nimble to changing needs of researchers, changing rules from sponsors and data providers, and evolving technological solutions.

Traditional de-identification simply by removing certain fields may increasingly be a problem as a standalone technique as data volumes grow. Inherently, machine learning and artificial intelligence algorithms will conflict with the privacy goals of de-identification, given that the whole point of ML/AI is to reveal the biases and structure of data mixed with the real patterns of interest, thus inherently opening up the opportunity for linkage attacks. De-identification may be best used in tandem with other controls, including technical security controls and legal agreements. For example, de-identification can be seen as a method for removing obvious details so that analysts don't stumble upon identifying information accidentally. In addition, strong technical security controls could be seen as methods for preventing leakage of information, and legal controls could be put in place to require that researchers do not attempt to re-identify any of the information.

[2] This conclusion is similar to one of those drawn at the "Enabling Trustworthy Campus Cyberinfrastructure for Science" workshop, hosted jointly by Internet2 and Trusted CI, and held at the University of Maryland in September 2018 at the Quilt/CC* PIs meeting.

In contrast, where appropriate, differential privacy is increasingly seen as the gold standard for preserving privacy, as has been increasingly demonstrated through industry and U.S. government use, and as the the Sloan Foundation-funded Harvard and Microsoft-led OpenDP effort has shown, and may be a more secure, private, useful, and usable approach for preserving privacy where its techniques are applicable. That said, calibrating and applying differential privacy is currently non-trivial, and scientists typically don't have tremendous amounts of time to learn technologies and adapt their code to use them. At one extreme, consider how long source code is often used in certain domains --- the high energy physics community often uses Fortran code for decades. Or, more recently, how long it has taken for GPUs to be adopted in genomics. Most researchers typically can't afford the development time for these new approaches. Thus, in the near term, differential privacy is likely to be best applied in situations where the effort is worthwhile. For example, differential privacy may be best used in scientific computing where the given dataset has significant importance (e.g., analyzing data related to a large-scale global health issue), lasting use (e.g., the data will be used over many years), and/or where the data will be used very broadly, rather than in situations where data need only be made available for a very small group of individuals.

Similarly, encrypted computing algorithms clearly have their place, as we have discussed. However, they also have usability constraints, and also performance constraints that might limit their adoption in scientific computing workflows in the near-term. Again, an exception may be in situations in which the effort of modifying and recompiling programs with cryptographic libraries is worth the time of the scientists using the data, and where data is small enough or timing requirements are low enough so that performance issues are reduced.

As we have discussed, trusted execution environments, such as the AMD SEV solution seems likely to provide substantial benefit in removing the requirement of having to trust the data center. In addition, in cloud computing environments, such as Google Cloud Confidential Computing [Goo20], mentioned earlier, the approach can be relatively turn-key. For a campus environment, however, recall that TEEs are not themselves full solutions, as they require whole infrastructures to be built around them. However, in either the cloud case, or in the campus case, once built, whole-VM TEEs can provide significant value in which scientists can more or less simply perform the computing workflows they are used to performing, except with the knowledge they are more secure.

The one universal rule is that sensitive data in research cannot be ignored, or individual PIs will develop their own — and likely unacceptable — environments and procedures to handle such data. Campuses *must* be involved. Ensuring minimal compliance may protect

against certain lawsuits, although true security that actually protects both the data and the institution and its personnel against reputational damage requires robust technical, procedural, political, and legal protections. These protections must be managed at the highest level but must be developed in concert with the end users — the researchers, or the researchers will either not know about them or may feel a lack of ownership in the solution and seek alternative solutions that feel more usable and useful.

Thus the development of such procedures need to involve campus personnel including chief research officers, chief information officers, chief information security officers, campus counsel, chief privacy officers, IRBs, sponsored research, research IT, libraries, and a very broad cross-section of both individual researchers, and also potentially should include certain data providers. Doing so will ensure that solutions have the full weight of campus resources behind them, that the solutions are broadly understood and known to be useful, and that the solutions are acceptable to the organizations whose sensitive data is to be stored.

References

[Abo18] John M. Abowd. Protecting the Confidentiality of America's Statistics: Adopting Modern Disclosure Avoidance Methods at the Census Bureau, August 17, 2018. https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_confi.html

[ACC] The Virginia ACCORD Project. https://www.va-accord.org

[ACCC] HPCwire, "UVA Leads Nationwide Project to Protect Health Data for COVID-19 Research," Aug. 4, 2020. https://www.hpcwire.com/off-the-wire/uva-leads-nationwide-project-to-protect-health-data-for-covid-19-research/

[ACG+16] Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016. https://dl.acm.org/doi/10.1145/2976749.2978318

[Ada19] Andrew K. Adams, "Trusted CI Begins Engagement with REED+," Trusted CI Blog, February 25, 2019. https://blog.trustedci.org/2019/02/trusted-ci-begins-engagement-with-reed.html

[AGA+20] Ayaz Akram, Anna Giannakou, Venkatesh Akella, Jason Lowe-Power, and Sean Peisert. "Performance Analysis of Scientific Computing Workloads on Trusted Execution Environments." *arXiv preprint* arXiv:2010.13216, 25 Oct 2020. https://arxiv.org/abs/2010.13216

[Amz] Amazon, "AWS Nitro Enclaves." https://aws.amazon.com/ec2/nitro/nitro-enclaves/

[App17] Apple, Inc. Learning with Privacy at Scale. Dec. 2017. https://machinelearning.apple.com/research/learning-with-privacy-at-scale

[Bar18] Thomas Barton, *editor*, "Enabling Trustworthy Campus Cyberinfrastructure for Science" workshop, p, hosted jointly by Internet2 and Trusted CI, and held at the University of Maryland in September 2018 at the Quilt/CC* PIs meeting." Dec. 2018. https://www.internet2.edu/blogs/detail/16960

[BLR13] Avrim Blum, Katrina Ligett, and Aaron Roth. A Learning Theory Approach to Non-Interactive Database Privacy. *Journal of the ACM (JACM)*, 60(2):12, 2013. https://dl.acm.org/doi/abs/10.1145/2450142.2450148

[CCC] Confidential Computing Consortium. https://confidentialcomputing.io

[CCC20a] Confidential Computing Consortium, "Confidential Computing: Hardware-Based Trusted Execution for Applications and Data." July 2020. https://confidentialcomputing.io/wp-content/uploads/sites/85/2020/11/confidentialcomputing_outreach_whitepaper-8-5x11-1.pdf

[CCC20b] Confidential Computing Consortium, "Confidential Computing Deep Dive v1.0." Oct. 2020. https://confidentialcomputing.io/wp-content/uploads/sites/85/2020/10/Confidential-Computing-Deep-Dive-white-paper.pdf

[CCPA] State of California Department of Justice, "California Consumer Privacy Act." https://oag.ca.gov/privacy/ccpa

[Chi] University of Chicago, "Secure Data Enclave." https://securedata.uchicago.edu/

[Con] University of Connecticut, "Secured Research Infrastructure (SRI)." https://security.uconn.edu/secured-research-infrastructure/

[CUI] National Archives and Records Administration, "Controlled Unclassified Information (CUI)." https://www.archives.gov/cui

[DD79] Dorothy E. Denning and Peter J. Denning. Data Security. *Computing Surveys*, 11(3), September 1979. https://dl.acm.org/doi/10.1145/356778.356782

[DDS78] Dorothy E. Denning, Peter J. Denning, and Mayer D. Schwartz. The Tracker: A Threat to Statistical Database Security. *ACM Transactions on Database Systems*, 4(1):76–96, March 1978. https://dl.acm.org/doi/abs/10.1145/320064.320069

[Den82] Dorothy Elizabeth Robling Denning. *Cryptography and Data Security*. Addison-Wesley Longman Publishing Co., Inc., 1982. https://faculty.nps.edu/dedennin/publications/Denning-CryptographyDataSecurity.pdf

[Dfar] "Defense Federal Acquisition Regulation Supplement." https://www.acquisition.gov/dfars

[Dif] Diffprivlib: The IBM Differential Privacy Library. https://github.com/IBM/differential-privacy-library

[Duk] Duke University, "Learn about the PRDN [Protected Research Data Network]." https://ssri.duke.edu/learn-about-prdn

[Dwo06] Cynthia Dwork. Differential Privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, part II (ICALP)*, July 2006. https://link.springer.com/chapter/10.1007/11787006_1

[Fism] "Federal Information Security Modernization Act of 2014 (FISMA 2014)." https://www.cisa.gov/federal-information-security-modernization-act

[Flo] Florida University, "UF Research Shield." https://www.rc.ufl.edu/services/restricted-data/researchshield/

[GDPR] General Data Protection Regulation (GDPR). https://gdpr-info.eu/

[Gen09] Craig Gentry. *A Fully Homomorphic Encryption Scheme*. PhD thesis, Stanford University, 2009. https://crypto.stanford.edu/craig/craig-thesis.pdf

[Gen10] Craig Gentry. Computing Arbitrary Functions of Encrypted Data. *Communications of the ACM*, 53(3):97–105, 2010. https://dl.acm.org/doi/10.1145/1666420.1666444

[Goo19a] Google, LLC. Differential Privacy. https://github.com/google/differential-privacy, 2019.

[Goo19b] Google, LLC. TensorFlow Privacy. https://github.com/tensorflow/privacy, March 6, 2019.

[Goo20] Google, LLC. Introducing Google Cloud Confidential Computing with Confidential VMs. https://cloud.google.com/blog/products/identity-security/introducing-google-cloud-confidential-computing-with-confidential-vms

[HHL+19] Justine S. Hastings, Mark Howison, Ted Lawless, John Ucles, and Preston White. Unlocking Data to Improve Public Policy. *Communications of the ACM*, 62(10):48–53, September 2019. https://cacm.acm.org/magazines/2019/10/239676-unlocking-data-to-improve-public-policy/fulltext

[HHSG] U.S. Department of Health and Human Services, "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule." https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

[HHSP] U.S. Department of Health and Human Services, "HIPAA Privacy Rule." https://www.hhs.gov/hipaa/for-professionals/privacy/index.html

[HHSS] U.S. Department of Health and Human Services, "HIPAA Security Rule." https://www.hhs.gov/hipaa/for-professionals/security/index.html

[HM70] Lance J Hoffman and William F Miller. Getting a Personal Dossier from a Statistical Data Bank. Datamation, 16(5):74–75, 1970.

[KPR17] Marcel Keller, Valerio Pastro, Dragos Rotaru, "Overdrive: Making SPDZ Great Again," Cryptology ePrint Archive, Report 2017/1230, 2017. https://eprint.iacr.org/2017/1230

[Mic17] Microsoft, Introducing Azure confidential computing, September 14. 2017. https://azure.microsoft.com/en-us/blog/introducing-azure-confidential-computing/

[N53] National Institute of Standards and Technology, "Special Publication 800-53 Rev. 4: Security and Privacy Controls for Federal Information Systems and Organizations," January 2015. https://nvd.nist.gov/800-53

[N66] National Institute of Standards and Technology, "Special Publication 800-66 Rev. 1: An Introductory Resource Guide for Implementing the Health Insurance Portability and Accountability Act (HIPAA) Security Rule," October 2008. https://csrc.nist.gov/publications/detail/sp/800-66/rev-1/final

[N171] National Institute of Standards and Technology, "Special Publication 800-171 Rev. 2: Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations," February 2020. https://csrc.nist.gov/publications/detail/sp/800-171/rev-2/final

[NCS] National Center for Supercomputing Applications, "NCSA announces SOC 2 Blog Series," September 1, 2020.

[Nea18] Joe Near. Differential Privacy at Scale: Uber and Berkeley Collaboration. In *Engima*. USENIX, January 16, 2018. https://www.usenix.org/conference/enigma2018/presentation/ensign

[NF14] Arvind Narayanan and Edward W. Felten. No Silver Bullet: De-identification Still Doesn't Work, July 9, 2014. https://www.cs.princeton.edu/~arvindn/publications/no-silver-bullet-de-identification.pdf

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust De-Anonymization of Large Sparse Datasets. In *Proceedings of the 29th IEEE Symposium on Security and Privacy*, pages 111–125, Oakland, CA, May 2008. https://ieeexplore.ieee.org/document/4531148

[Ind] Indiana University, "Research Database Complex (RDC)." https://kb.iu.edu/d/amuw

[OpDP] Harvard University, "Open DP." https://projects.iq.harvard.edu/opendp

[Pur] Purdue University, "REED+ Ecosystem." https://www.rcac.purdue.edu/services/reedplus/

[REN] RENCI, "Infrastructure for Privacy-Assured CompuTations (ImPACT)," https://cyberimpact.us/

[SAW13] Latanya Sweeney, Akua Abu, and Julia Winn. Identifying Participants in the Personal Genome Project by Name. Available at SSRN 2257732, 2013. https://arxiv.org/abs/1304.7605

[Sch75] Jan Schlörer. Identification and Retrieval of Personal Records from a Statistical Data Bank. *Methods of Information in Medicine*, 14(01):7–13, 1975.

[Sha20] Anurag Shankar, "Trusted CI Begins Engagement with UC Berkeley," Trusted CI Blog, February 21, 2020. https://blog.trustedci.org/2020/02/trusted-ci-begins-engagement-with-uc.html

[She] San Diego Supercomputer Center, University of California, San Diego, "Sherlock." https://sherlock.sdsc.edu/

[Shep] San Diego Supercomputer Center, University of California, San Diego, "Sherlock Pricing." https://sherlock.sdsc.edu/pricing

[Sky] San Diego Supercomputer Center, University of California, San Diego, "Skylab." https://sherlock.sdsc.edu/skylab

[Skyd] San Diego Supercomputer Center, University of California, San Diego, "Introducing 'Skylab.'" https://sherlock.sdsc.edu/s/Sherlock-Skylab-Deck-June-2020.pdf

[SS15] Reza Shokri and Vitaly Shmatikov. Privacy-Preserving Deep Learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1310–1321, 2015. https://dl.acm.org/doi/abs/10.1145/2810103.2813687

[Stan] Stanford Nero Computing. https://med.stanford.edu/nero.html

[Swe97] Latanya Sweeney. Weaving Technology and Policy Together to Maintain Confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110, 1997. https://doi.org/10.1111/j.1748-720X.1997.tb01885.x

[Uni19] United Nations Global Working Group (GWG) on Big Data. UN Handbook on Privacy-Preserving Computation Techniques, March 2019. http://publications.officialstatistics.org/handbooks/privacy-preserving-techniques-handbook/UN%20Handbook%20for%20Privacy-Preserving%20Techniques.pdf

[VSG+19] Nikolaj Volgushev, Malte Schwarzkopf, Ben Getchell, Mayank Varia, Andrei Lapets, and Azer Bestavros. Conclave: Secure Multi-Party Computation on Big Data. In *Proceedings of the Fourteenth EuroSys Conference*, 2019. https://dl.acm.org/doi/pdf/10.1145/3302424.3303982

[Yao86] Andrew Chi-Chih Yao. How to Generate and Exchange Secrets. In *Proceedings of the 27th Annual IEEE Symposium on Foundations of Computer Science*, pages 162–167, 1986. https://ieeexplore.ieee.org/document/4568207

Version History

September 28, 2020, v0.8 — Preliminary Public Report
November 11, 2020, v1.0 — Public Report