Research infrastructure for the safe analysis of sensitive data

Ian Foster
Department of Computer Science, The University of Chicago
Mathematics and Computer Science Division, Argonne National Laboratory
5735 South Ellis Avenue, Chicago, IL 60637
Tel: 630 400 0426
Email: foster@uchicago.edu

Abstract
The use of administrative and other new data sources for the scientific study of human beings and their interactions is often hindered by the legal, technical, and operational difficulties inherent in connecting analysts not only to data but also to the new analysis methods required to link and analyze those data. We examine how new digital research infrastructures can be used to reduce such barriers. Experience in other domains shows that appropriate infrastructure can enable the efficient, secure, and collaborative integration of domain expertise, data, and analysis capabilities. We review the state of the art in these areas and argue for the use of cloud hosted enclaves as a safe interaction point for analysts, data, and software, and as a means of automating and thus professionalizing data stewardship processes.

Keywords: Safe data, enclaves, cloud computing, data stewardship

1 Introduction

Access to new types of data has revolutionized much of science (Hey, Tansley, and Tolle, 2009). That revolution has yet to fully make its way to the scientific study of human beings and their interactions in such areas as program management, policy development, and

scholarly research (Lane, Heus, and Mulcahy, 2008; Metzler, 2016). Progress has been hindered by the legal, technical, and operational difficulties inherent in connecting *analysts* (domain experts with questions) to the multiple sources of sensitive *data* from which answers are to be extracted and the analysis *methods* required to link and analyze those data. Indeed, the high cost of assembling these three essential elements of data-driven scholarship on human systems has often prevented its large-scale application, at least outside the narrow realms of commercial data mining and national security.

We examine here how digital *research infrastructure* can be used to reduce barriers to connecting analysts, data, and methods, and thus to accelerate data-driven investigations of human systems. A research infrastructure encompasses both technology and process. As *technology*, it is a collection of computer hardware, software, and networks designed and operated to support research activities. As *process*, it implements and enforces policies, conventions, and rules to ensure that the technology is applied in ways that meet user needs in such areas as security and trust. Experience in other scientific domains (Atkins et al., 2003; Finholt, 2002; Foster, 2002) shows that appropriate research infrastructure can enable the efficient, secure, and collaborative integration of domain expertise, data, and analysis capabilities.

We focus here on the technological aspects of research infrastructure. We are not ignorant of the profound legal and ethical challenges associated with the analysis of sensitive human data that no technology or process, however sophisticated, can ever fully address. However, we believe that well-designed technologies, when subject to appropriately controlled processes, can reduce barriers to secure data access, use, and reuse. They can, for example, provide safe environments for data cleaning, ensure secure auditing of data accesses, and protect against attack vectors that individual research labs, statistical agencies, and other institutions are hard pressed to counter. And they can package these solutions in ways that can be be replicated rapidly and easily in different jurisdictions.

We proceed as follows. First, in Section 2, we elucidate technology and process requirements for a research infrastructure for sensitive human data. In Sections 3 and 4, we review approaches taken and lessons learned from previous work in other domains. In Section 5, we explore how cloud computing can enable new approaches to research

infrastructure that may transform how data on human subjects are handled. We conclude in Section 6.

2 Problem statement

Our overarching goal is to accelerate data-driven research and policy around human beings and their interactions so as to support a range of program management, policy development, and scholarly purposes. To this end, we want to enable efficient, effective, and secure access to sensitive data about societal systems. To give just one example, analysis of detailed data about the life histories of ex-offenders and on factors such as educational and employment opportunities, housing programs, and health services in the locales to which ex-offenders are released can suggest new approaches to reducing recidivism. But to answer such questions, analysts need the ability to link highly sensitive data from multiple sources.

In examining approaches to this problem, we distinguish the interests of three classes of actors: data providers, analysts, and method developers.

Data providers are those who collect data, such as federal agencies, state agencies, municipalities, companies, and universities. A data provider may be prepared to make their data available to external analysts, but typically only if they see the benefits as outweighing the costs and risks. *Benefits* can include new information or insights that result from their data being analyzed from fresh perspectives and/or combined with data from other data providers; *costs* the effort required to organize data for external access; and *risks* the legal, reputational, or other negative consequences if legal, regulatory, or other constraints on data access and use are not followed. We need research infrastructures that maximize benefits for data providers while minimizing associated costs and risks.

Analysts are those who work with data for such purposes as program management, policy development, or scholarly research. It is analysts who ultimately (we hope) will deliver benefits to data producers. Thus a second set of requirements for a research infrastructure relate to analyst productivity, which may be compromised by difficulties in data discovery, access, linkage, and analysis. While analysts may have deep knowledge about specific domains of social science (e.g., criminal justice, employment, education), they will not

necessarily be conversant with datasets from sources with which they were previously unfamiliar, making both discovery and use difficult. Steps taken by data providers to reduce risk, such as de-identification, can hinder subsequent analysis. As the volume and variety of data grow, analysts may find themselves requiring new methods and tools: for example, new linkage methods and high-performance computing tools to process large datasets. Increasingly, also, they find themselves under pressure to document the steps that were followed to reach their conclusions.

Method developers are social scientists, statisticians, computer scientists, and others who develop new methods for linkage, analysis, visualization, etc., especially of large datasets. These individuals can have much to offer analysts working with large datasets, but historically they have had limited or no access to realistic test data (Metzler, 2016), reducing their ability to tackle problems that really matter to data providers and analysts. By treating them as stakeholders in research infrastructure, we recognize that data providers and analysts alike may have an interest in facilitating their work.

Inevitably, the interests of these different parties can conflict, and thus a research infrastructure that is intended to support the interests of all three groups of stakeholders needs to be able to manage tradeoffs.

3 Research data integration and analysis infrastructures

Social scientists are not the first to grapple with the challenges of "big data." Indeed, people have been building infrastructures for storing and sharing large quantities of information in support of research since at least the Library of Alexandria in the third century BCE. Computers have transformed how we work with information by enabling automated management and analysis, and it is now routine for researchers in the physical and biological sciences to create and use research infrastructures that store and enable the analysis of trillions of data elements. Developers of such systems have negotiated tradeoffs between scale, cost, the types of questions that can be asked and answered, security, reliability, and other factors. Here we review some important approaches.

3.1 Repositories and services

Community data repositories. Physical and biological scientists have encountered and addressed a variety of big data challenges as improved instrumentation and computational capabilities produced ever more data. One important innovation was the online data repository that aggregated data from many producers. Such systems have proven highly successful, particularly when backed by policies that require data deposit. For example, the Protein Data Bank (PDB) (Berman et al., 2000), established in 1971 with seven protein structures, has grown to more than 120,000 protein structures; GenBank (Benson et al., 2013), established in 1982, now holds close to two trillion bases of genetic sequence data; the Genomic Data Commons (GDC), established in 2016, holds more than two quadrillion ($2x10^{15}$) bytes (Grossman et al., 2016). However, these systems are primarily data repositories, not data analysis systems (although PDB provides some comparison functions): once researchers identify data of interest, they must download them to perform analysis locally. Such download and local analysis can become impractical as data grow in complexity and size, particularly when analysts want to compute over *all* data rather than simply examining individual data elements.

Federated data repositories. The cost and governance issues associated with a centralized repository can lead communities to develop federated repositories, in which many data providers all implement common protocols and standards so that researchers can query to determine what data are available at any site and then download those data for analysis. For example, the Earth System Grid Federation (Williams et al., 2016), an evolution of the centralized Earth System Grid (ESG) (Bernholdt et al., 2005) established to store and distribute climate simulation results associated with Intergovernmental Panel on Climate Change assessments, links climate data providers at dozens of site worldwide. Federated repositories avoid the cost and governance challenges of a centralized site, but can introduce significant challenges of consistency in protocols. And they still require that researchers download data for analysis.

High energy physics provides an unusual example of an inverse approach. Data is produced in extremely large volumes at a small number of locations, such as the Large Hadron Collider (LHC) in Geneva. The data volumes are too large for the computational capacity

that could be acquired at CERN, and thus the LHC Computing Grid (Lamanna, 2004) distributes data from the central LHC site to many computing centers for analysis.

Data services. When data are large, it can be impractical for analysts to download them to their local computers for analysis. Thus, some data providers implement data service interfaces that perform computations on data in response to user requests. This *server-side computation* (so-called because it is performed on the data server, not the client) may be restricted to predefined operations (e.g., subsetting, certain statistical analyses) or may allow for execution of arbitrary user-supplied code. The ability to run arbitrary user-supplied code is powerful, but can raise challenging security concerns. One partial solution is to restrict user-supplied code to specific programming dialects of execution types, as is done for example in the SkyServer astronomical database: only Structured Query Language (SQL) queries are supported (Szalay et al., 2002).

Systems that allow for user-initiated server-side computation may also need to manage potentially large computational demands, particularly if remote analysts are allowed to request computations over large quantities of data. This problem can be dealt with by restricting the amount of computation that can be performed, implementing queues, leveraging grid computing (Foster, 2002), and/or running on a cloud that provides for elastic computing capacity, perhaps with accounting to enable recouping of costs.

Server-side data analysis can also be attractive as a means of accessing complex software that may require considerable expertise to install and operate. In the biological sciences, systems like Galaxy (Goecks et al., 2010) that allow analysts to upload data for processing with standard software have become popular, particularly for those with limited local infrastructure and expertise. By incentivizing data upload, such systems can also encourage the accumulation of large data collections. SEED (Overbeek et al., 2004) and MG-RAST (Meyer et al., 2008) exemplify this approach: they retain microbial genomes and metagenomic data, respectively, that researchers upload for analysis, increasing their coverage of species and environments.

Federated data services. Data services, like databases, can be federated via the definition and implementation of appropriate protocol and data standards (Foster and Grossman,

2003). For example, in astronomy, different groups have created databases of the sky at different wavelengths. To enable cross-database queries, the virtual observatory community has defined standards that allow a researcher to query a digital sky survey database for objects with certain characteristics (Szalay and Gray, 2001). A user interested in finding, for example, stars that are visible in the infrared but not the optical spectra (possible brown dwarves) would perform queries against both infrared and optical databases.

The systems described so far deal with data that may be completely open (e.g., PDB or GenBank data) or available to anyone who registers their scientific interest (e.g., ESGF). In other cases, access is granted only to researchers who agree to abide by specified policies. This is the case, for example, for the GDC and for the Cancer Genome Atlas (TCGA), which holds genetic sequence data from more than 500 cancer tissue samples. The TCGA's Data Use Certification Agreement ("The Cancer Genome Atlas", 2014) requires that researchers agree to maintain the privacy of the patients who provided tissue samples, access the data securely, and follow publication guidelines. However, such agreements are typically enforced by social norms rather than by automated processes.

3.2 Marts, lakes, and spaces

Another important dimension along which research data infrastructures vary is the degree to which their contents are harmonized to simplify discovery, access, and analysis.

Many scientific data repositories are organized as highly structured *data marts*, with all data and metadata being converted to standardized formats and schema before being uploaded. Access then occurs via standardized protocols and APIs. This is the case with systems such as PDB, GenBank, and ESGF, for example: each defines file format and metadata conventions that must be followed by data providers. In essence, such systems impose costs on data providers to simplify life for data consumers. The question of where costs should be incurred in a data publication pipeline is often a hot topic of debate when federating data. Changing business processes "upstream" can simplify life for data aggregators and also improve data quality. But the associated costs may be unacceptable.

An alternative approach to data repository design focuses on minimizing costs for data providers. In so-called *data lakes* (Terrizzano et al., 2016), data of potentially many types and from potentially many sources can be deposited without concern for conventions. Thus, for example, raw data from experimental apparatus may be found alongside more processed data that have undergone further processing, for example to transform them into common formats. Some data will be highly documented and standardized, while other data may have no associated descriptive metadata. Data lakes work well when analysts who work with important or popular data improve the quality of associated metadata over time. Franklin et al. (2005) coined the term *data space* for systems that encourage such pay-as-you-go improvements, by for example cataloging all datasets and recording provenance relationships between initial and derived datasets. Google's GOODS system (Halevy et al., 2016) uses such methods to manage a data lake of more than 20 billion datasets; however, these datasets are not subject to the access controls required for the human subjects considered here.

Todd Harbour proposes the term "brickyard" to denote a place where people go to obtain materials that they can expect to be regularized, with predictable dimensions and formats: for example, bricks and patio slabs. An effective data sharing system must incorporate methods or incentives for such regularization. A brickyard is also like a research data enclave in another respect: people cannot take whatever they want. Business processes, documentation, and permissions must accompany any removal of materials.

3.3 Collaboration, provenance, data and code reuse, and reproducibility

Data space concepts illustrate one of the many ways in which collaboration can facilitate productive data analysis. All too often in science, individual researchers work independently to understand, correct, and analyze source data. In the process, they may duplicate work that has already been performed by others. The result can be not only wasted effort but also poor science, if for example subtly different, but undocumented, assumptions made by different analysts lead to different results.

These concerns can be addressed by mechanisms that allow work performed by one analyst (e.g., documenting a dataset, creating a derived data product, or developing code for a specific analysis) to be shared with others. Various methods have been developed and

applied for this purpose in science. Collaborative tagging mechanisms (Cattuto, Loreto, and Pietronero, 2007) enable researchers to share structured or unstructured annotations on documents, data, and code. Notebook technologies such as Jupyter (Kluyver et al., 2016) are used to share code in understandable ways. Conventions and tools have been developed for recording provenance relationships between datasets and code (Moreau et al., 2010). The ability to assign persistent identifiers to datasets, data subsets, and code is important for provenance, reuse, and citation (Paskin, 2005; Chard et al., 2016).

4 Sensitive data

Sensitive data are sufficiently confidential that data providers cannot rely on researcher *declarations* to maintain confidentiality: data providers instead need to maintain *positive control* over data access and use, in order to reduce the risk of unwanted disclosure. We consider two classes of such control mechanisms: the *curator model* and *secure enclaves*. The first approach limits the data that analysts can access, the operations that they can perform on data, and/or the results that can be obtained from analyses, to prevent them from ever seeing sensitive data. Secure enclaves, in contrast, allow full access to data but then restrict who is allowed that access and what data can be exported.

4.1 Statistical disclosure control and the curator model

Statistical disclosure control approaches seek to allow analysts to operate on data without ever obtaining access to information about individuals (Willenborg and De Waal, 2012). Dwork and Smith (2010) formalize the problem by defining a *curator model* in which a trusted and trustworthy curator (e.g., the Census Bureau) gathers sensitive information from many respondents (the sample) and then works to release to the public statistical facts about the underlying population, in such a way as not to compromise the privacy of the individual respondents. They distinguish between *noninteractive* access, in which the set of statistics to be computed is predefined, and *interactive* access, in which the curator responds to requests from individual analysts. (The latter approach can introduce scalability challenges.)

Various techniques have been developed to enable access to data statistics without revealing information about individuals. Curators may aggregate data, suppress certain

information data, or perturb the values of variables before publication (Domingo-Ferrer and Mateo-Sanz, 2002; Seastrom, 2010; Skinner, 2009). The concept of *differential privacy* provides a formal framework for thinking about such issues, stating that the results of a query against a dataset with data on a specific individual removed should not be distinguishable from the results when data on that individual are present (Dwork, 2014; Dwork et al., 2006). Related models allow for reasoning about the probability that personal information will be revealed by a series of information releases, such as responses by a curator to multiple interactive requests.

One form of information suppression is de-identification (Uzuner, Luo, and Szolovits, 2007), which may involve removing identifiers altogether (anonymization) or replacing each identifier in the dataset with a unique key (pseudonymization or coding) (Phillips and Knoppers, 2016). The effectiveness of such approaches is vigorously debated, with some arguing that essentially any data can be re-identified via linkage with other datasets or knowledge (Barocas and Nissenbaum, 2014; Ohm, 2010), and others arguing that such steps are often noisy and thus may not be revealing for more than a few individuals.

4.2 Secure enclaves

Another approach to preventing disclosure of sensitive information is to place physical constraints on data access and export. Desai et al. argue for a portfolio approach to data security (Desai, Ritchie, and Welpton, 2016), with processes defined to ensure *safe people* (i.e., restrictions on who is allowed to access the enclave), *safe projects* (i.e., audits of the purposes for the data is to be used), *safe settings* (i.e., secure environments), and *safe outputs* (e.g., via manual review of data outputs before they are released). Variants of this approach make different tradeoffs between security and convenience.

Air-gapped enclaves. In this first approach, all analysis must be performed in a secure enclave with no Internet connection. This approach is frequently employed by national security organizations and stewards of public datasets such as the U.S. Census' Federal Staistical Research Data Centers, which create "air-gapped" data infrastructures comprising computers that are not connected to the Internet and that users have to visit and use in person, with tight control over what data, if any, they can take with them when they leave. The unfettered access to data provided by these enclaves is invaluable, allowing data to be

exploited to its fullest potential. However, we do not view air-gapped enclaves as an adequate solution for the data sharing and analysis use cases considered here due to their inconvenience, cost, and lack of support for importing data from other sources for purposes of data integration. The analyses that we aim to support require that many people be able to access, integrate, and analyze multiple sensitive datasets.

Secure remote access. The inconvenience inherent in air-gapped enclaves has led various groups to develop systems in which the analyst connects remotely, for example over a virtual private network, to the data enclave (Lane et al., 2008). The identity of the analyst is established via secure authentication and the analyst can then interact with software running at the enclave to perform analyses, review results, and ultimately download outputs (perhaps after review by data enclave staff). This approach is far more convenient for the remote analyst, but introduces risk as the data enclave has little control over the remote analyst's computing environment. To counter that risk, some enclaves require that remote access be allowed only from dedicated secure sites, under the supervision of qualified staff (Bender and Heining, 2011). In another related approach, analysts construct data capsules—virtual machine images with analyst-provided and configured code—that can then be deployed and run on data in an enclave, with controls applied on what data can be released while the capsule runs (Borders et al., 2009; Zeng et al., 2014).

Note that secure enclaves, whether they support remote access or not, do not directly address the need for analysts to integrate data from multiple sources, which may require that data from one source be transported to the other for linkage and analysis.

4.3 Cloud computing

Internet search and social networking companies such as Amazon, Facebook, Google, and Microsoft represent another approach to large-scale data aggregation and analysis. Each of these companies has created an enormous computational infrastructure—a *cloud*—that they use to store and analyze large quantities of data, both public (e.g., public web sites and social media postings) and sensitive (e.g., searches performed, private messages between individuals, books purchased). Importantly, these systems are structured to permit rapid analyses of large fractions of those data, reliably and cost effectively. Professional systems management, supported by the massive revenues of these companies and their large

economies of scale, make them highly reliable and, it seems, also secure, although large disclosures of cloud-hosted information are reported periodically.

Few external researchers and analysts can access the data that these infrastructures contain. However, several companies also operate *public clouds* that anyone with a credit card can access: for example, Amazon Web Services, Google Cloud, and Microsoft Azure. Each of these systems allows interested parties to acquire storage, computing, and other resources and services in an on-demand, pay-as-you go manner. It thus becomes straightforward to instantiate a private data enclave by allocating cloud storage, loading data into that storage, and allocating cloud computers to run analyses on that storage.

Can one reasonably use such a private data enclave to store, share, and analyze sensitive data? The answer to this question varies with geographical location and the data in question, but in the U.S., the federal government has defined policies and procedures that can be followed to satisfy government regulations. To understand the nature of these policies, it is helpful to study the nature of the software components that go into creating a cloud-based service. Figure 1 provides a perspective on this question, showing how responsibilities may be divided according to whether one relies on the cloud provider just for infrastructure (IaaS) or also for platform services (PaaS). (The case in which the cloud provider operates application software, software-as-service or SaaS, is also shown, but is not relevant here.) The cloud provider is responsible for securing the low-level infrastructure that you yourself would have to secure if you established a secure data enclave at your own institution ("on premise"), but that still leaves you responsible for the security of at least some higher-level components.

Security in a cloud-hosted secure data enclave is thus the joint responsibility of the cloud provider and the user, which in the case of human subjects data might be an individual institution or agency—or, as we discuss in Section 5, a *cloud data enclave operator* who provides services for many users and purposes.

Figure 1: Cloud security responsibilities, as discussed in the text. Adapted from Simorjay (2016).

In the U.S., the federal government has defined an assessment and authorization process, the Federal Risk and Authorization Management Program (FedRAMP) (Office of

Management and Budget, 2013), for determining whether a particular combination of cloud provider and user software and procedures can be used for sensitive data from federal agencies. Becoming FedRAMP certified is an onerous process that involves not only substantial engineering but also documentation, assessment by a FedRAMP-accredited third-party assessment organization, and finally review by the FedRAMP Joint Assessment Board. Thus, it is not a process to be taken lightly. However, once completed, the associated documentation can be relied upon by many potential consumers of the cloud service in question: another example of the economies of scale provided by cloud computing (Barnard, 2016).

5 The cloud data enclave: A transformative new approach?

We have reviewed a variety of approaches to large-scale data sharing and analysis. Each has distinct advantages and disadvantages for data providers, analysts, and method providers, and each has a place in the data universe. However, we believe that the emergence of secure, scalable, reliable, and inexpensive public clouds represents an opportunity for transformative change in how data about human subjects are organized, shared, and analyzed. In this section, we explain why we make this statement and outline steps that can be taken to realize its potential.

The basic idea is to leverage a commercial cloud as a secure, scalable **cloud data enclave** for data sharing, access, and analysis, implementing within that cloud a **safe data platform** that provides automated implementations of the various processes associated with data usage. In so doing, we can allow sensitive data from many providers to be discovered, linked, and analyzed in a controlled manner—and to permit, furthermore, analysts and method developers to share data, analysis methods, results, and expertise in ways not easily possible today. The approach thus combines elements of the data enclave, cloud, data lake, and data space approaches described above to meet the stakeholder requirements identified in Section 2.

5.1 Use public cloud to implement a secure and scalable virtual data enclave

The emergence of high-quality public clouds over the past decade has transformed how people, companies, and institutions work with information technology. Anyone with a

credit card can now obtain cost-effective access not only to essentially unlimited storage and computing, but also to many higher-level services that would be prohibitively expensive to implement within a single institution. A growing number of startups, research institutions, and research projects now take advantage of these capabilities.

The requirements identified in Section 2 for working with data about human subjects are distinctive but no less amenable to cloud hosting and automation. Appropriately configured cloud services can be used to provide a secure location for datasets and software; enable packaging of software in reusable forms via virtual machines and containers; permit secure, controlled sharing of data and code; scale computation elastically as required; and enable monitoring and auditing of activity for reliability and security. In each case, we can leverage the enormous investment made by cloud providers in these areas.

The use of public cloud to host a data enclave also enables other economies of scale. We can implement standardized, reusable implementations of automated data stewardship processes. And, as noted above, FedRAMP standards provide a framework for security reviews of the entire system and permit the sharing of review results and authority to operate packages across different data providers (Barnard, 2016). A single location also facilitates collaboration, as data, code, results, and best practices can be shared more easily (subject of course to stewardship: see Section 5.2) when implemented in a uniform manner.

5.2 Provide automated safe stewardship mechanisms

As noted in Section 1, a research infrastructure comprises not only technology but also process. We assert that the public cloud's uniform environment and economies of scale can make it feasible to replace current ad hoc, manual, incomplete implementations of the data protection approaches listed in Section 4.2 with automated, and thus fully auditable and replicable, implementations, as we now describe.

Safe people: We can use secure authentication to validate the identity of people seeking access to the enclave, and require strong methods such as two-factor authentication for action that involve work on sensitive data. Modern identity federation methods can enable reuse of existing credentials (Barnett et al., 2011; Tuecke et al., 2016). To ensure that users

of the enclave operate from a common understanding of roles, responsibilities, and authorities, we can also require that potential users participate in an educational program that teaches these principles.

Safe settings: The detailed security specifications and reviews defined by FedRAMP are one part of the technical security procedures required to provide safe settings. We can provide supporting mechanisms such as *safe collections*, sets of data and associated metadata plus policies governing, for example, where data within a collection must be stored, the approval process that must be followed to request access, and the monitoring required for access and use. Other platform capabilities can include secure logging of all user actions for audit and forensic purposes.

Safe projects: We can implement structured and traceable project review and audit to ensure safe projects. We can provide supporting mechanisms for users such as *safe search*, to allow analysts to discover datasets that meet research goals and that they are allowed to access, and *safe workspaces*, sets of data plus code that can be used to analyze the data. Other mechanisms can allow a user to discover sensitive datasets, request access to those datasets, copy them into a workspace upon access being granted, run analyses on the data within the workspace, and then export both data and code subject to approval processes.

Safe outputs: We can implement data export controls, such as processes that require review of a data export request by an authorized data steward, using suitably secure mechanisms to validate the identity of both the requesting user and the steward and to log the authorization decision.

Data providers will need to be able to specify which processes and policies should apply to each dataset that they make available in the enclave. Alternatively, it may be possible to reach agreement across data providers as to data classifications, in a manner analogous to the information security marking metadata defined within the national security context to enable "interagency access control, automated exchanges, and appropriate protection of shared intelligence" (*Information Security Marking Metadata*, Visited January 1, 2017). Given such marking metadata, data stewardship tools can then determine without further

user input where data is to be placed, what access policies apply, what policies apply to derived data products, etc.

We note that a system that streamlines the processes by which derived datasets are first created and then described, discovered, and reused does not necessarily ensure that data reuse will become commonplace. If users typically stop at the creation step, then the data enclave is likely to accumulate large quantities of "dark data" (Heidorn, 2008): data that are meaningful only to their creator and thus are never reused by others. Potential solutions include incentives (e.g., for producing high-quality datasets that are reused, for reusing data), visibility (e.g., by notifying users of datasets that have been produced with a particular set of inputs), and automation (for example to detect unused and duplicate data).

5.3 Build on a safe data platform to encourage contributions

The various mechanisms introduced above can be viewed as constituting a cloud-hosted safe data platform that automates important elements of the sensitive data stewardship process. We use the word *platform* here deliberately. As explained by van Astyne et al. (van Astyne, Parker, & Choudary, 2016), "a platform provides the infrastructure and rules for a marketplace that brings together producers and consumers." A successful platform, like the iPhone, Android, Python, or R, substantially eliminates the friction associated with developing, sharing, finding, and consuming solutions. Similarly, a successful safe data platform will enable many researchers and communities to prepare and analyze sensitive data, share data and code, and collaborate. By thus enabling a rich ecosystem of methods and tools, it will allow research communities to continuously contribute and test new approaches to research and policy questions as the nature of data on human subjects changes. To this end, it will need to incorporate an exchange where communities can deposit and discover reusable code and recipes that they can use to build their own solutions and solution environments—thus enabling the dissemination of ideas and methods.

Figure 2 depicts how a safe data platform might work in practice. In this figure, (1) an analyst searches across multiple collections (e.g., data from different federal or state agencies) to find data that meet specified criteria. This search is performed based on metadata that the analyst is authorized by the appropriate data steward(s) to see. (2) The

analyst requests access to a dataset identified via search, which triggers an approval workflow as specified by associated policy. (The figure shows an example policy: "any workspaces to which data are loaded must be located in FedRAMP-certified storage; researchers must present two- factor credentials; and all approvals are manual.") In this case, the policy requires (3) manual sign off by the associated data steward.

Figure 1: A cloud-hosted safe data platform, showing the various actors and actions.

If approval is granted, then (4) the dataset can be loaded into a workspace with the analyst's desired analytics environment. The analyst may also (5) import additional open or restricted data and code into the workspace. The analyst can now use the data by working in the workspace, eventually creating new data that (6) they can request permission to export for (7) external use (e.g., to create a table in a research article) and/or (8) publication to an existing or new collection for sharing with other platform users for reuse. The latter *release process* involves assigning a persistent identifier, assembling metadata for discovery, and organizing the data for easy loading into workspaces. The definition of standardized release processes that can be adopted by different data providers is another area that could facilitate the smooth operation of a cloud data enclave.

A steward can also use platform capabilities to facilitate use. For example, she can create a workspace for data ingest, (5) import a restricted access dataset to that workspace for preparatory clean-up and de-identification, and then (7) publish the processed dataset to a collection for access. Specialized workspace instance can be provided to support data stewards in this work. Note that a workspace itself can become data: a user can publish an entire computing environment, including data, tools, etc., as they would any other dataset, to support reuse. The platform can also incorporate mechanisms designed to simplify or even enforce the implementation of the data management plans that federal institutions are increasingly requiring the researchers that they fund to follow.

5.4 Analysis of stakeholder needs.

We suggest that the approach that we have just described can address the stakeholder needs of Section 2 as follows. For data providers, it can:

- *Minimize data contribution costs* by implementing simple data upload protocols and APIs, supported by standardized data ingest methods where feasible.
- *Improve releasability of data and publications*, by providing automated, secure, validated pipelines for extracting releasable data and associated metadata, and thus simplify adherence to data publication policies.
- *Minimize risks* by enforcing data provider-specified policies for data access, analysis, and output—ideally supported by standardized data classifications.
- *Maximize benefits* by integrating their data into a rich ecosystem of other data providers, analysis, and method providers, in which controlled sharing of data, analysis methods, and expertise is encouraged.

For analysts, it can:

- *Simplify access to data* by providing standardized methods for discovering, requesting access to, and accessing sensitive data.
- *Enable integration* of datasets from different sources.
- *Educate them in best-of-breed methods*, by providing access to common analysis procedures/techniques and examples of how tools are used by other researcher.
- *Encourage collaboration* by providing standardized methods for creating and sharing annotations on datasets and code.
- *Facilitate reproducible research* by automating the capture of the steps followed to obtain a particular result.
- *Enable big data analysis* by allowing analysts to scale computational resources to meet computational needs—as long as they (or someone else) can pay for the cloud computing time.

For method providers, it can:

- *Provide access to data* required to design, test, and evaluate new methods.

- *Provide access to communities* of individuals who share interests and experience in problems and methods.

6 Conclusions

New data sources present fascinating opportunities for new understanding of human beings and their interactions, and thus better policies, programs, and science. But seizing those opportunities requires new research infrastructures that enable *analysts* to easily access, integrate, and analyze datasets from multiple sources, while also maximizing benefits and minimizing costs and risks for *data providers*. The fact that new data sources are often larger, noisier, and less structured than data conventionally studied in social science leads to a third set of requirements for research infrastructures, relating to scale and access by *method providers*.

We have reviewed approaches for large-scale sharing and analysis of both general science data and sensitive data about human subjects. We argue that developments in cloud computing present opportunities for transformative new approaches to working with sensitive data, in which public clouds are used to create a secure data enclave in which currently manual and ad hoc data stewardship processes are standardized and automated. Such a system can enable sensitive data from different sources to be discovered, integrated, and analyzed in appropriately controlled manners. It can, furthermore, allow researchers to share analysis methods, results, and expertise in ways not easily possible today. The development of such a system thus has the potential to botjh accelerate research and enable a flowering of new methods for studying human subjects.

The analysis of sensitive human data raises profound legal and ethical challenges that no technology, however sophisticated, can ever fully address. However, we believe that well-designed technologies, when operated in appropriately controlled cloud environments, can reduce barriers to secure data access, use, and reuse. They can, for example, provide safe environments for data cleaning, ensure secure auditing of data accesses, and protect against attack vectors that individual research labs, statistical agencies, and other institutions would be hard pressed to counter.

The proposed approach can also serve to broaden the community of researchers working to improve the state of the art in safe data management and analysis. Both social scientists and computer scientists will gain from being able to access real data about social problems, without having to create and certify their own trusted data service. The platform should spur the development of many new methods and tools as computer scientists work with different research communities to customize environments; an extensible code exchange will facilitate the transfer and adoption of new methods to many research communities.

Acknowledgments

I thank Kyle Chard, Simson Garfinkel, and Todd Harbor for insightful comments on a draft of this article, and Rachana Ananthakrishnan, Dan Black, Charlie Catlett, Ron Jarmin, Frauke Kreuter, Julia Lane, and Steven Tuecke for many helpful conversations. This research was supported in part by the U.S. Department of Energy under Contract DE-AC02-06CH11357.

Biography

Ian Foster is a Professor of Computer Science at the University of Chicago and a Senior Scientist and Distinguished Fellow at Argonne National Laboratory. His research contributions span high-performance computing, distributed systems, and data-driven discovery. He has published hundreds of research articles and seven books on these and other topics.

References

Daniel E. Atkins, Daniel E., Kelvin K. Droegemeier, Stuart I. Feldman, Hector Garcia-Molina, Michael L. Klein, David G. Messerschmitt, Paul Messina, Jeremiah P. Ostriker, and Margaret H. Wright. 2003. *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon panel on cyberinfrastructure*. Available from www.nsf.gov (accessed June 1, 2017).

Barnard, R. Oct 26, 2016. Think FedRAMP is a bottleneck? Think again. *Federal Computing Week*.

Barnett, William, Von Welch, Alan Walsh, and Craig A. Stewart. 2011. *A roadmap for using NSF cyberinfrastructure with InCommon*. Available from http://hdl.handle.net/2022/13024 (accessed June 1, 2016).

Barocas, Solon, and Helen Nissenbaum. 2014. Big data's end run around anonymity and consent. In *Privacy, big data, and the public good: Frameworks for engagement*, eds. Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 44-75. New York NY: Cambridge University Press.

Bender, Stefan, and Jörg Heining. 2011. The research-data-centre in research-data-centre approach: A first step towards decentralised international data sharing. Paper presented at the IASSIST Conference, 2 June 2011, Vancouver, BC.

Benson, Dennis A., Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2014. GenBank. *Nucleic Acids Research* 42 (D1): D32-D37.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, Talapady N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. The protein data bank. *Nucleic Acids Research* 28 (1): 235-242.

Bernholdt, David, Shishir Bharathi, David Brown, Kasidit Chanchio, Meili Chen, Ann Chervenak, Luca Cinquini, Bob Drach, Ian Foster, Peter Fox, Jose Garcia, Carl Kesselman, Rob Markel, Don Middleton, Veronika Nefedova, Line Pouchard, Arie Shoshani, Alex Sim, Gary Strand, and Dean Williams. 2005. The Earth System Grid: Supporting the next generation of climate modeling research. *Proceedings of the IEEE* 93 (3): 485-495.

Borders, Kevin, Eric Vander Weele, Billy Lau, and Atul Prakash. 2009. Protecting confidential data on personal computers with storage capsules. Paper presented at the 18th USENIX Security Symposium, Ann Arbor, MI.

The Cancer Genome Atlas (TCGA) Data Use Certification Agreement. August 20, 2014. Available at cancergenome.nih.gov (accessed June 1, 2017).

Cattuto, Ciro, Vittorio Loreto, and Luciano Pietronero. 2007. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences, 104*(5):1461-1464.

Chard, Kyle, Mike D'Arcy, Ben Heavner, Ian Foster, Carl Kesselman, Ravi Madduri, Alexis Rodriguez Stian Soiland-Reyes, Carole Goble, Kristi Clark, Eric W. Deutsch, Ivo Dinov, Nathan Price, and Arthur Toga. 2016. I'll take that to go: Big data bags and minimal identifiers for exchange of large, complex datasets. In Proceedings of the IEEE International Conference on Big Data, Washington, DC.

Desai, Tanvi, Felix Ritchie, and Richard Welpton. 2016. *Five Safes: Designing data access for research*. University of the West of England Economics Working Paper Series 1601.

Domingo-Ferrer, Josep, and Josep Maria Mateo-Sanz. 2002. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering* 14 (1): 189-201.

Dwork, Cynthia. 2014. Differential privacy: A cryptographic approach to private data analysis. In *Privacy, big data, and the public good: Frameworks for engagement*, eds. Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 296-322. New York NY: Cambridge University Press

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Theory of Cryptography Conference*, 265-284. Springer Berlin Heidelberg.

Dwork, Cynthia, and Adam Smith. 2010. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality* 1 (2): 135-154.

Finholt, Thomas A. 2002. Collaboratories. *Annual Review of Information Science and Technology*, 36(1):73-107.

Foster, Ian. 2002. The grid: A new infrastructure for 21st century science. *Physics Today*, 55(2):42-47.

Foster, Ian and Robert Grossman. 2003. Data integration in a bandwidth-rich world. *Communications of the ACM*, 46(11):51-57.

Franklin, Michael, Alon Halevy, and David Maier. 2005. From databases to dataspaces: a new abstraction for information management. *ACM SIGMOD Record* 34 (4): 27-33.

Goecks, Jeremy, Anton Nekrutenko, James Taylor, and the Galaxy Team. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 11 (8): R86

Grossman, Robert L., Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt. 2016. Toward a shared vision for cancer genomic data. *New England Journal of Medicine* 375 (12): 1109-1112.

Halevy, Alon, Flip Korn, Natalya F. Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, and Steven Euijong Whang. 2016. Goods: Organizing Google's datasets. In *Proceedings of the 2016 International Conference on Management of Data*, 795-806. ACM

Heidorn, P. Bryan. 2008. Shedding light on the dark data in the long tail of science. *Library Trends*, 57 (2): 280-299.

Hey, Tony, Stewart Tansley, and Kristin M. Tolle. 2009. *The fourth paradigm: data-intensive scientific discovery*. Redmond, WA:Microsoft Research, 2009.

Information security marking metadata. Office of the Director of National Intelligence. Available from www.dni.gov (accessed June 1, 2017).

Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. 2016. Jupyter Notebooks—a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, eds Fernando Loizides and Birgit Schmidt, 87-90. IOS Press.

Lamanna, Massimo. 2004. The LHC computing grid project at CERN. *Nuclear*

Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 534 (1): 1-6.

Lane, Julia, Pascal Heus, and Tim Mulcahy. 2008. Data access in a cyber world: Making use of cyberinfrastructure. *Transactions on Data Privacy* 1 (1): 2-16.

Metzler, Kate. November 22, 2016. "The Big Data rich and the Big Data poor": the new digital divide raises questions about future academic research. LSE Impact Blog. Available from blogs.lse.ac.uk/impactofsocialsciences (accessed June 1, 2017).

Meyer, Folker, Daniel Paarmann, Mark D'Souza, Robert Olson, Elizabeth M. Glass, Michael Kubal, Tobias Paczian, Alex Rodriguez, Rick Stevens, Andreas Wilke, Jared Wilkening, and Robert A. Edwards. 2008. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9 (1): 386.

Moreau, Luc, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, Beth Plale, Yogesh Simmhan, Eric Stephan, Jan Van den Bussche. 2011. The open provenance model core specification (v1. 1). *Future Generation Computer Systems* 27 (6): 743-756.

Office of Management and Budget. 2013. *Enhancing the Security of Federal Information and Information Systems*. Available from obamawhitehouse.archives.gov (accessed June 1, 2017).

Ohm, Paul. 2010. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57:1701-1777.

Overbeek, Ross A., Terence Disz, and Rick Stevens. 2004. The SEED: A peer-to-peer environment for genome annotation. *Communications of the ACM* 47(11): 46-51.

Paskin, Norman. 2005. Digital Object Identifiers for scientific data. *Data Science Journal*, 4:12-20.

Phillips, Mark, and Bartha M. Knoppers. 2016. The discombobulation of de-identification. *Nature Biotechnology* 34 (11): 1102-1103.

Seastrom, Marilyn M. 2010. Statistical methods for protecting personally identifiable information in aggregate reporting. National Center for Education Statistics Report 2011-603.

Simorjay, F. 2016. *Shared responsibilities for cloud computing*. Redmond WA: Microsoft. Available from gallery.technet.microsoft.com (accessed June 1, 2017).

Skinner, Chris J. 2009. Statistical disclosure control for survey data. *Handbook of Statistics* 29: 381-396.

Szalay, Alexander, and Jim Gray. 2001. The world-wide telescope. *Science* 293 (5537): 2037-2040.

Szalay, Alexander S., Jim Gray, Ani R. Thakar, Peter Z. Kunszt, Tanu Malik, Jordan Raddick, Christopher Stoughton, and Jan vandenBerg. 2002. The SDSS Skyserver: public access to the Sloan Digital Sky Server data. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, 570-581. ACM.

Terrizzano, Ignacio, Peter M. Schwarz, Mary Roth, and John E. Colino. 2015. Data wrangling: The challenging journey from the wild to the lake. Paper presented at the 7th Biennial Conference on Innovative Data Systems Research, 4 January—7 January, 2015. Asilomar, CA.

Tuecke, Steven, Rachana Ananthakrishnan, Kyle Chard, Mattias Lidman, Brendan McCollam, and Ian Foster. 2016. Globus Auth: A research identity and access management platform. In *Proceedings of the 12th IEEE International Conference on e-Science*, 203-212. IEEE.

Uzuner, Özlem, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association* 14 (5): 550-563.

Van Alstyne, Marshall W., Geoffrey G. Parker, and Sangeet Paul Choudary. 2016. Pipelines, platforms, and the new rules of strategy. *Harvard Business Review* 94 (4): 54-62.

Willenborg, Leon, and Ton De Waal. 2012. *Elements of statistical disclosure control*. Vol. 155. Springer Science & Business Media.

Williams, Dean N., V. Balaji, Luca Cinquini, Sébastien Denvil, Daniel Duffy, Ben Evans, Robert Ferraro, Rose Hansen, Michael Lautenschlager, and Claire Trenham. 2016. A global repository for planet-sized experiments and observations. *Bulletin of the American Meteorological Society* 97 (5): 803-816.

Zeng, Jiaan, Guangchen Ruan, Alexander Crowell, Atul Prakash, and Beth Plale. 2014. Cloud computing data capsules for non-consumptive use of texts. In *Proceedings of the 5th ACM workshop on Scientific Cloud Computing*, 9-16. ACM.