Research infrastructures for the safe analysis of sensitive data

Ian Foster, foster@uchicago.edu

1 Introduction

Access to new types of data has revolutionized much of science [18]. Yet that revolution has yet to fully make its way to the scientific study of human beings and their interactions, where 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 *microdata* from which answers are to be extracted and the analysis *methods* required to link and analyze data. Indeed, experience suggests that the considerable effort currently required to assemble these three essential elements of data-driven scholarship on human systems has largely prevented its large-scale application, at least outside the narrow realms of commercial data mining and national security.

We examine here ways in which technology can be used to reduce barriers to analyst-data-methods integration and thus accelerate data-driven investigations of human systems in such areas as program management, policy development, and scholarly research. Specifically, we examine prospects for developing new *research infrastructures* that can enable the efficient, secure, and collaborative integration of relevant domain expertise, data sources, and analysis capabilities.

A research infrastructure (sometimes termed cyberinfrastructure [3]) is a collection of computer hardware, software, and networks designed and operated to support the work of researchers. People have been building infrastructures for storing and sharing 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. 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. Inevitably, developers of such systems face many tradeoffs: between scale, cost, the types of questions that can be asked and answered, security, reliability, and other factors.

Our goals here are threefold: to elucidate requirements for a research infrastructure for sensitive human microdata; review approaches taken and lessons learned from previous work in other sciences; and propose specific ideas for new approaches to research infrastructure that we believe can leverage past experiences and new technologies to deal with data on human subjects in ways that improve on those employed in previous systems.

While we adopt a purely technological perspective, we are not ignorant of the profound legal and ethical challenges associated with the analysis of sensitive human data that no technology, however sophisticated, can ever fully address. However, we believe that well-designed technologies, when operated in appropriately controlled 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.

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 specific datasets, 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 [23], 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 discovery, access, linkage, and analysis challenges of "big data." 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 use of online data repositories to aggregate data from many data producers. These systems have proven highly successful, particularly when backed by policies that require data deposit. For example, the Protein Data Bank, established in 1971 with seven protein structures, has grown to more than 120,000 protein structures today, while GenBank, established in 1982, now holds close to two trillion bases of genetic sequence data. However, these systems are data repositories, not data analysis systems: 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.

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 [6, 38], an evolution of the centralized Earth System Grid 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 [21] 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 data from the repository to their local computers for analysis. Thus, some data providers implement data service interfaces that perform computations on data in response to requests from remote users. These computations 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 [15].

Systems that allow for user-initiated 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, and/or running on a cloud that provides for elastic computing capacity, perhaps with accounting to enable recouping of costs.

In addition to avoiding the need to download data for analysis, server-side data analysis can be attractive as a means of accessing complex software. In the biological sciences, for example, data analysis software often requires considerable expertise to install and operate. Thus systems like Galaxy [14] that allow analysts to upload data for processing with standard analysis 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. The SEED [28] and MG-RAST [24] systems exemplify this approach: they permit researchers to upload microbial genomes and metagenomic data, respectively, for analysis, and then retain the uploaded data in their database as a means of 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. 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 [33]. 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 Cancer Genome Atlas (TCGA), which holds genetic sequence data from more than 500 cancer tissue samples. The TCGA's Data Use Certification Agreement [1] requires that researchers agree to maintain the privacy of the patients who provided tissue samples, access the data securely, and follow TCGA publication guidelines.

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. In essence, such systems impose costs on data providers to simplify life for data consumers.

An alternative approach to data repository design focuses on minimizing costs for data providers. In so-called *data lakes* [34], 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 and curation. Some data will be highly documented and carefully curated, 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 those data's metadata over time. Halevy coined the term *data space* [16] for systems that encourage such pay-as-you-go improvements, by for example cataloging all datasets and recording provenance relationships between datasets. Google's GOODS system [17] uses such methods to manage a data lake of more than 20 billion datasets.

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

Dataspace 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.

To address these concerns, we would like 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 easily with others. Various methods have been developed and applied for this purpose in science. Collaborative tagging mechanisms [7] enable researchers to share structured or unstructured annotations on documents, data, and code. Notebook technologies such as Jupyter [20] have become popular as a means of sharing code in understandable ways. Conventions and tools for recording provenance relationships between datasets and code have been developed [25]. The ability to assign persistent identifiers to datasets, data subsets, and code is important for provenance, reuse, and citation. Various identifier schemes have been developed with varying degrees of formality [8, 29].

4 Sensitive data

Sensitive data are sufficiently confidential that the data provider cannot rely on researcher *declarations* to maintain confidentiality: positive *enforcement* is required to reduce the risk of unwanted disclosure. We consider two classes of such enforcement 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 allow full access but then restrict 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 [37]. Dwork and Smith 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 [13]. 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.

Various techniques have been developed with the goal of enabling access to data statistics without revealing information about individuals. Curators may aggregate microdata, suppress certain information in the microdata set, or perturb the values of microdata variables before publication [11, 31, 32]. The concept of differential privacy provides one 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 [12].

One form of information suppression is de-identification [35], which may involve removing identifiers altogether (anonymization) or replacing each identifier in the dataset with a unique key (pseudonymization or coding) [30]. 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 [4, 27], 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. Data providers typically apply a portfolio approach to data security, 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) [10]. Various such approaches have been applied, with 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 companies, national security organizations, and stewards of public datasets such as the U.S. Census [2], all of whom routinely 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. 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 data integration from multiple sources. The analyses that we aim to support require that many people be able to access 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. 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 [5].

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. Highly 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.

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 operates Amazon Web Services, Google operates the Google Cloud, and Microsoft operates 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 procedure 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.

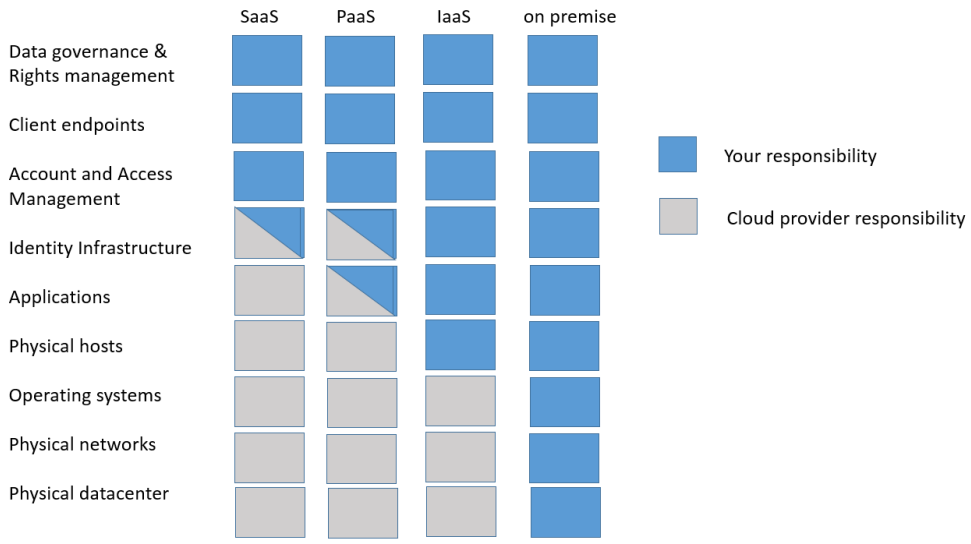

Figure 1: Cloud security responsibilities, as discussed in the text.

Security in a cloud-hosted secure data enclave is thus the joint responsibility of the cloud provider and the user. In the U.S., the federal government has defined an assessment and authorization process, the Federal Risk and Authorization Management Program (FedRAMP) [26], 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

Table 1: Perspectives on the properties of different approaches

	Community repositories	Statistical disclosure control	Private enclave	Virtual enclave
Access full microdata		No	Yes	Yes
Leverage work of others				
Perform large computations				
Integrate data from multiple sources				

5 A safe data platform approach

We have reviewed a variety of approaches to the sharing and analysis of large datasets. Each has distinct advantages and disadvantages for data providers, analysts, and method

providers. We now introduce an approach that combines elements of cloud, data lake, and other approaches described above in ways that we believe has attractive properties for our target stakeholders. The basic idea is to leverage a commercial cloud as a secure, scalable **virtual data enclave** for data sharing, access, and analysis, implementing within that cloud a **safe data platform** that provides automated, and thus fully auditable and replicable, implementations of the various data protection approaches described earlier. In so doing, we believe that we can allow highly 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. These ideas build on earlier discussions of secure research infrastructures for sensitive data by Lane et al. [22].

5.1 A safe data platform

The core of the approach that we outline here is a Safe Data Platform (SDP) that systematizes data stewardship and use. Historically, the stewardship of sensitive data has rested on a portfolio approach: *safe people* (restricted access), *safe projects* (project audits), *safe settings* (secure environments), and *safe outputs* (screening). We envision this SDP replacing current ad hoc, manual, incomplete implementations of these proven data protection approaches with automated, and thus fully auditable and replicable, implementations. Specifically, it will employ *identity management* to restrict access and thus provide safe people; *technical security procedures* to provide safe settings; *structured and traceable project review and audit* to ensure safe projects; and *data export controls* to provide safe outputs, all under the control and oversight of the data stewards who have ultimate responsibility for each dataset.

We envision the SDP implementing two major abstractions to:

- **Safe collections**, sets of data and associated metadata, plus policies governing, for example, where data within the collection must be stored, the approval process that must be followed to request access, the monitoring required for access and use, and data export policies.
- **Safe workspaces**, sets of data and code that can be used to analyze those data.

And two major services:

- A **safe search service**, which will analysts to discover datasets that meet research goals and that they are allowed to access.
- A **safe stewardship service**, which will implement and enforce policies concerning data access and export, such as requiring steward review and approval of derived data before sharing.

By thus institutionalizing major elements of the data curation process, we believe that this SDP approach can fundamentally change how researchers access, find, use, and reuse data on human subjects. It will straightforward for a wide variety of data providers to import datasets into access-controlled Safe Collections on which provider-specified policies are enforced. It will allow analysts to perform Safe Search on Safe Collections to discover datasets that meet research goals, request access to restricted data, and import selected datasets into Safe Workspaces, comprising code and data, for analysis. Analysts will also be

able to export derived datasets, subject to data stewardship policies (e.g., requiring steward review and approval of derived data before sharing), as enforced by the Safe Stewardship service.

5.2 The safe data platform in use

Figure 2 depicts important SDP workflows. In this figure, (1) an analyst searches across multiple collections (e.g., data from Census and Justice) 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. 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 SDP users for reuse. The latter sharing process involves assigning a persistent identifier, assembling metadata for discovery, and organizing the data for easy loading into workspaces.

A steward can also use SDP 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 environment, including data, tools, etc., as they would any other dataset, to support reuse.

5.3 A note about the word "platform"

We use the word *platform* in referring to SDP deliberately. As explicated by van Astyne et al. [36], "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 SDP will enable many researchers and communities to prepare and analyze sensitive data, share data and code, and engage in collaborative research. It itself will not be a sole provider of service "solutions"—an approach that would be anathema to the overarching goal of accelerating research by facilitating safe innovation by many contributors. Rather, by 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.

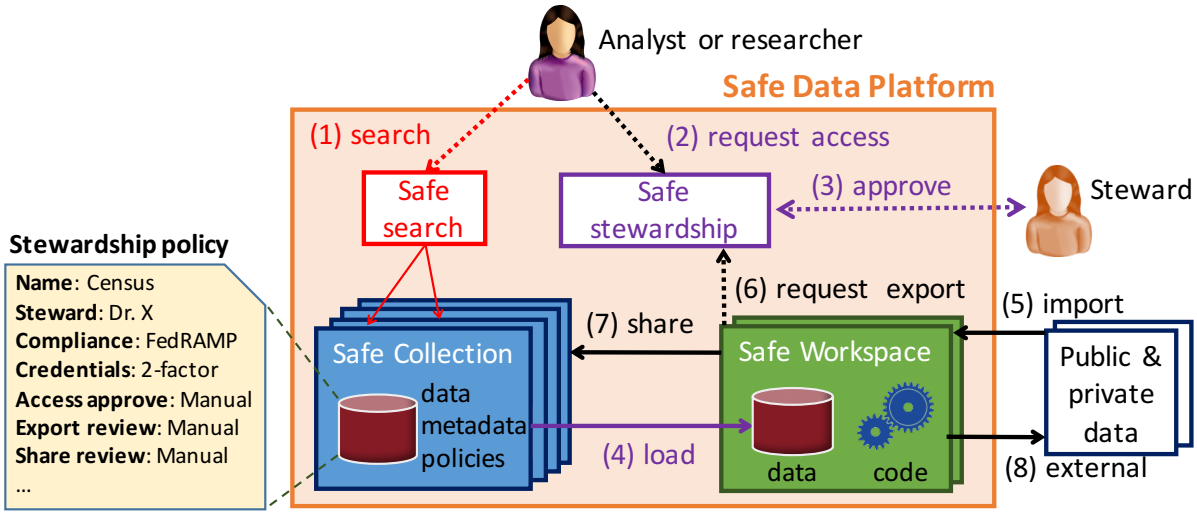

Figure 2: The cloud-hosted safe data platform, showing the various actors and actions.

5.4 Analysis of stakeholder needs.

We argue that our proposed approach can address stakeholder needs 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.
- *Minimize risks* by enforcing data provider-specified policies for data access, analysis, and output.
- *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.
- *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.

5.5 Discussion

We argue that this proposed Safe Data Platform will enable sensitive data from different sources to be discovered, integrated, and analyzed in a carefully controlled manner, and will, furthermore, allow researchers to share analysis methods, results, and expertise in ways not easily possible today. We believe that by providing such a platform, we can enable a fundamental change in how data on human beings from governments, statistical agencies, research institutions, and other organizations, are made available for academic research. The work will thus both accelerate research and enable a flowering of new methods for studying human subjects.

We do not propose these technological approaches to safe data challenges naively: we recognize that 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 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.

We also believe that the proposed platform will 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.

6 Conclusions

New data sources present fascinating opportunities for new new understanding of human beings and their interactions, and thus better policies, programs, and science. But to seize those opportunities, new research infrastructures are required that can enable analysts to access, integrate, and analyze datasets from multiple sources. This requirement in turn motivates needs for data sharing methods that can minimize costs and risks and maximize benefits 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 methods providers.

We have reviewed approaches taken in the science community for large-scale data sharing and analysis of both general science data and sensitive data about human subjects. We argue that developments in cloud computing present the opportunity for new approaches to research infrastructure for sensitive data, in which cloud-based platforms are used for collaborative discovery on highly sensitive data. Seizing this opportunity will require that methods for reusable code, reproducible research, collaborative knowledge curation, and automated inference be adapted for use in this environment. If we are successful in this endeavor, and develop appropriate methods for indexing, labeling, and organizing derived

data, barriers to access to sensitive data and advanced methods will be reduced, and social science researchers will be able to build on the shoulders of giants.

Acknowledgments

We acknowledge many helpful conversations with, among others, Rachana Ananthakrishnan, Dan Black, Charlie Catlett, Kyle Chard, Ron Jarmin, Frauke Kreuter, Julia Lane, and Steve Tuecke. This research was supported in part by XXX.

References

1. The Cancer Genome Atlas (TCGA) Data Use Certification Agreement. August 20, 2014; Available from: https://cancergenome.nih.gov/pdfs/Data_Use_Certv082014.
2. Federal Statistical Research Data Centers. [Accessed Visited December 16, 2016]; Available from: https://www.census.gov/fsrdc.
3. Atkins, D.E., Droegemeir, K.K., Feldman, S.I., Garcia-Molina, H., Klein, M.L., Messina, P., Messerschmitt, D.G., Ostriker, J.P. and Wright, M.H. Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure, 2003.
4. Barocas, S. and Nissenbaum, H. Big data's end run around anonymity and consent. Lane, J., Stodden, V., Bender, S. and Nissenbaum, H. eds. Privacy, big data, and the public good: Frameworks for Engagement, Cambridge University Press, NY, 2014, 44-75.
5. Bender, S. and Heining, J., The Research-Data-Centre in Research-Data-Centre Approach: A First Step Towards Decentralised International Data Sharing. IASSIST Conference, https://ec.europa.eu/eurostat/cros/system/files/S10P3.pdf_en, 2011.
6. Bernholdt, D., Bharathi, S., Brown, D., Chanchio, K., Chen, M., Chervenak, A., Cinquini, L., Drach, B., Foster, I., Fox, P., Garcia, J., Kesselman, C., Markel, R., Middleton, D., Nefedova, V., Pouchard, L., Shoshani, A., Sim, A., Strand, G. and Williams, D. The Earth System Grid: Supporting the Next Generation of Climate Modeling Research. *Proceedings of the IEEE*, 93(3):485-495, 2005.
7. Cattuto, C., Loreto, V. and Pietronero, L. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences*, 104(5):1461-1464, 2007.
8. Chard, K., D'Arcy, M., Heavner, B., Foster, I., Kesselman, C., Madduri, R., Rodriguez, A., Soiland-Reyes, S., Goble, C., Clark, K., Deutsch, E.W., Dinov, I., Price, N. and Toga, A., I'll Take That to Go: Big Data Bags and Minimal Identifiers for Exchange of Large, Complex Datasets. IEEE International Conference on Big Data, Washington, DC, USA, 2016.
9. Crosas, M., King, G., Honaker, J. and Sweeney, L. Automating open science for big data. *The ANNALS of the American Academy of Political and Social Science*, 659(1):260-273, 2015.
10. Desai, T., Ritchie, F. and Welpton, R. Five Safes: designing data access for research. University of the West of England, http://eprints.uwe.ac.uk/28124/1/1601.pdf, 2016.
11. Domingo-Ferrer, J. and Mateo-Sanz, J.M. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and data Engineering*, 14(1):189-201, 2002.
12. Dwork, C. Differential Privacy: A Cryptographic Approach to Private Data Analysis. Lane, J., Stodden, V., Bender, S. and Nissenbaum, H. eds. Privacy, Big Data, and the Public Good: Frameworks for Engagement, Cambridge University Press, 2014, 296.
13. Dwork, C. and Smith, A. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2):2, 2010.
14. Goecks, J., Nekrutenko, A., Taylor, J. and The Galaxy Team Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010.

15. Gray, J., Szalay, A.S., Thakar, A.R., Kunszt, P.Z., Malik, T., Raddick, J., Stoughton, C. and vandenBerg, J., The SDSS SkyServer - Public Access to the Sloan Digital Sky Server Data. ACM SIGMOD, 2002, 1-11.
16. Halevy, A., Franklin, M. and Maier, D., Principles of dataspace systems. 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2006, ACM, 1-9.
17. Halevy, A., Korn, F., Noy, N.F., Olston, C., Polyzotis, N., Roy, S. and Whang, S.E., GOODS: Organizing Google's Datasets. International Conference on Management of Data, 2016, ACM, 795-806.
18. Hey, T., Tansley, S. and Tolle, K. (eds.). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, 2009.
19. King, G. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods and Research*, 2007.
20. Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J. and Corlay, S., Jupyter Notebooks—a publishing format for reproducible computational workflows. 20th International Conference on Electronic Publishing, IOS Press, 87.
21. Lamanna, M. The LHC computing grid project at CERN. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 534(1-2):1-6, 2004.
22. Lane, J., Heus, P. and Mulcahy, T. Data Access in a Cyber World: Making Use of Cyberinfrastructure. *Transactions on Data Privacy*, 1(1):2-16, 2008.
23. Metzler, K. "The Big Data rich and the Big Data poor": the new digital divide raises questions about future academic research, The Impact Blog, London School of Economics and Political Science, November 22, 2016.
24. Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J. and Edwards, R. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(1):386, 2008.
25. Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E. and Bussche., J.V.d. The open provenance model core specification (v1.1). *Future Generation Computer Systems*, July, 2010.
26. Office of Management and Budget Enhancing the Security of Federal Information and Information Systems, https://www.whitehouse.gov/sites/default/files/omb/memoranda/2014/m-14-03.pdf, 2013.
27. Ohm, P. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57:1701, 2010.
28. Overbeek, R.A., Disz, T. and Stevens:, R.L. The SEED: a peer-to-peer environment for genome annotation. *Communications of the ACM*, 47(11):46-51, 2004.
29. Paskin, N. Digital Object Identifiers for scientific data. *Data Science Journal*, 4:12-20, 2005.
30. Phillips, M. and Knoppers, B.M. The discombobulation of de-identification. *Nature Biotechnology*, 34(11):1102-1103, 2016.
31. Seastrom, M.M. Statistical methods for protecting personally identifiable information in aggregate reporting (NCES No. 2011-603). *National Center for Education Statistics SLDS Technical Brief*, 3, 2010.
32. Skinner, C. Statistical disclosure control for survey data. *Handbook of statistics*, 29:381-396, 2009.
33. Szalay, A. and Gray, J. The World-Wide Telescope. *Science*, 293:2037-2040, 2001.
34. Terrizzano, I., Schwarz, P.M., Roth, M. and Colino, J.E., Data Wrangling: The Challenging Yourney from the Wild to the Lake. Conference on Innovative Data Systems Research.

35. Uzuner, Ö., Luo, Y. and Szolovits, P. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550-563, 2007.
36. van Astyne, M.W., Parker, G.G. and Choudary, S.P. Pipelines, platforms, and the new rules of strategy. *Harvard Business Review*, 94(4):16, 2016.
37. Willenborg, L. and De Waal, T. *Elements of statistical disclosure control*. Springer Science & Business Media, 2012.
38. Williams, D.N., Balaji, V., Cinquini, L., Denvil, S., Duffy, D., Evans, B., Ferraro, R., Hansen, R., Lautenschlager, M. and Trenham, C. A Global Repository for Planet-Sized Experiments and Observations. *Bulletin of the American Meteorological Society*, 97(5):803-816, 2016.