Background

Researchers in a wide range of campus disciplines face an increasingly complex set of requirements to protect their research data, with often overlapping requirements coming from an increasing number of directions: from funders, data providers, state and federal agencies, and from the campus and university itself. This problem is further complicated by disciplinary differences across the campus. Social scientists have long worked with human subjects research data, but they are exploring larger data sets, new methodologies, and new research opportunities. Other fields are perhaps newer to data-driven research (e.g., the humanities) or are exploring human subjects data for the first time or in new and expanded ways (e.g., in computer science and engineering). Simultaneously, even data that do not involve human subjects are increasingly subject to data use agreements and other requirements from data providers. Not surprisingly, much of this regulatory aspect is driven by cybersecurity threats and the potential impact of a data breach, a phenomenon and trend that is rapidly and chaotically evolving. In addition, data is increasingly a valuable commodity in our data-driven world, and many parties are taking an interest in issues of licensing, sharing, and even selling data.

In the summer of 2016, a joint team from Research IT, the Library, the D-Lab, and IST evaluated UC Berkeley’s service offerings for securing research data. Using the methodology developed for the RAE (Research and Academic Engagement) Redo Benchmarking project, the team defined Securing Research Data as:

Services in support of research data sets that have restrictions set by campuses, state or federal law, or contractual agreements. Includes policy and guidance; infrastructure services for storage and computation; consulting and training; cross-campus coordination for managing restricted data; and support for coordinated development to build resources and manage relationships with data providers.

The team identified the following criteria for evaluating service offerings at UC Berkeley and at thirteen peer universities:

- Discoverable and useful policy and guidance documentation
- Secure research environments and infrastructure
- Active training and consulting
- Life-cycle process support and cross-campus coordination
- Active development of restricted data resources

Following the comparative analysis, the team ranked UC Berkeley’s service offerings as a 3 (with 1 being the highest score and 4 the lowest). This three-level ranking was characterized as follows:

Accessible policy guidance with limited services, few/inflexible research environments for secure data, environments limited in terms of compute or storage, limited service support and training, limited coordination among campus stakeholders and service providers, ad-hoc development of discipline specific resources.

In short, researchers working with data that require protections of any kind most often must perform this activity with very little campus support.
Another campus-wide effort that demonstrated broad needs in this area was an audit of Research Data Management conducted by Audit and Advisory Services. The final report to campus, dated June 24, 2016, determined a) that the process for identifying required security protections was driven by many offices that do not necessarily work together in a coordinated way and b) that the scope of the problem and the campus-wide risk are very difficult to ascertain. That is, the campus does not know how much risk it faces, nor does it have the right policies, governance mechanisms, or even relationships with researchers that would facilitate a more accurate assessment.

Other campuses are not only doing more to help researchers, but they are also treating the ability to manage restricted data as a strategic differentiator. As data-intensive research including data science grows on campus, the importance of providing services that allow researchers to work with sensitive data of various kinds becomes strategically valuable and even critical. Indeed researchers at Berkeley are often at the vanguard of working with data that require protections of various kinds, and yet the services campus offers are not adequate. Campus data policies and enterprise services are geared towards administrative data, and most research systems have emphasized data sharing and collaborative access. As a result, researchers often struggle to develop solutions, taking valuable time and energy away from primary research.

Importantly, securing research data is not just a technical challenge. There are numerous policy issues, and the political, technological, and research opportunity landscapes are evolving rapidly. Currently experts on campus are distributed across different organizations, and while they are beginning to coordinate their work better, there is a growing demand for consultations as well as demand for improved service offerings for storing sensitive data, for performing computations in a secure environment, for transferring data with different levels of sensitivity, for collaboration among team members both on campus and beyond, and for preserving, protecting and sharing data sets for the future.

Project and process

Given the complex set of issues and the different offices involved, Research IT and the Library launched a six-month project through the Research Data Management Program with partners in the D-Lab, Information Security & Policy (OCIO), the Office of the Vice Chancellor for Research, and other units within Information Services & Technology.

The project charter and team members are documented in Appendix A.

Project Outcomes

The project team developed a work plan knowing that a six-month project could only be the beginning of a process to address the complex set of issues that the campus faces. At the same time, the team believed that a broad view was needed in order to lay the groundwork for later success. Therefore, team members explicitly targeted the various dimensions of the securing research data challenge: policy, process, partnerships, and technology.

- This whitepaper
- Formation of a D-Lab Working Group pulling together the larger community of stakeholders: researchers, service providers, compliance offices, and others.
- An analysis of the current state on campus and demand for solutions and guidance
- A draft Research Data Classification Guidelines document
● Findings and recommendations regarding broader needs for research data policy
● Benchmarking and case studies from other institutions
● Existing solution designs that researchers use now to store and analyze restricted data
● Technology service recommendations for environments for storage and analysis

These outcomes are presented below and in appendices attached to this whitepaper.

**Current State**

**Assessment of current demand**

In order to estimate demand for improved and new services, the project set out to quantify and classify the current state of sensitive-data use on campus. Team members aggregated information from several sources:

- Consultation requests to the Research Data Management program
- Data security review requests to the Information Security and Policy unit within the Office of the Chief Information Officer, and
- Case notes from the D-Lab program and from offices such as the Industrial Alliances Office.

The resultant set of 68 cases provides a glimpse of the broader need at UC Berkeley. The list comprises only those research projects that sought help from these campus units tasked with overseeing data security. It is not a scientifically representative sample of the research done using sensitive data at UC Berkeley. Yet, it offers rich insight into the scope and nature of that research. It is important to note that these cases represent a subset of the total demand.

**Who uses sensitive data?**

A full quarter of the researchers requesting campus help with their sensitive data hail from a single division within the College of Letters & Science: the Social Sciences division. The majority of these cases involve researchers in the Departments of Sociology and Psychology. Another 30% of the requests come from the School of Public Health and the Vice Chancellor for Research Office, the latter of which represents organized research units, centers, institutes, museums and the UC Botanical Garden. The list, in order of cases reporting, continues with the College of Engineering, the College of Natural Resources and the School of Social Welfare, and extends through a 'long tail' that spans schools such as the Haas School of Business and Berkeley Law, the Department of Statistics and the College of Environmental Design.

*Figure x: Affiliation of researchers using sensitive data (college or school, administrative office, type of program)*
In part this distribution follows naturally from the type of research questions and data addressed by each of these disciplines. The skew in numbers towards social sciences and public health also reflects administrative efforts (spurred by Leon Wong of Information Security and Policy) to increase communications among the campus’s institutional review board, its Sponsored Projects and Industrial Alliances Offices, and its Research Data Management and D-Lab programs. Greater familiarity among the staff of these offices has facilitated the development of a blueprint of sorts for addressing the data security needs of these types of projects. In contrast, a number of the disciplines that comprise the tail of the distribution are probably underrepresented: researchers working with intellectual property such as financial transaction data, market research data, and, perhaps, cutting-edge engineering and physical sciences technologies. Researchers in these areas might be relying on support resources at the college, rather than campus, level.

The distribution of cases handled by these research support offices alludes to different needs across domain and discipline; and different reasons for sensitivity (cf next chart). The breadth of demand complicates support efforts and argues for the need for improved communication and clarity in research administration processes and roles. The efforts accelerated by Leon during his time at Information Security and Policy offer an argument, and a model, for improvements in the way campus organizes itself to support its researchers.

What makes the data sensitive?

Sensitive, or ‘restricted-use’, data can be personally identifiable information (PII), confidential or time-sensitive data, or licensed intellectual property. The most protected forms of personally identifiable information (‘notice-triggering’ data) require that research subjects be notified (and, often, made whole to some extent) when the security of that data is breached. The sensitivity of the data, and the conditions by which it must be handled, are regulated variously by:

- Federal and state law
- Funder guidelines
- Data use agreements
- Licenses
- Campus policy
- Institutional review board (human subjects) approval and informed consent from the subjects themselves
- Ethics.
Among the projects requesting help from campus, more than half are governed by human subjects controls and data use agreements (DUAs) from the data owner -- typically a state or federal agency addressing health, welfare, and educational issues. As noted above, mechanisms for reviewing these types of data are more well-defined and exercised on campus; DUAs due to intellectual property might be handled within the school or college (Haas, engineering, etc.) and therefore underrepresented here. Personal health-related data, whether regulated by the federal Health Information Portability and Accountability Act (HIPAA) or as Research Health Information (more common on the Berkeley campus, which doesn’t have a medical center) comprises another 13% of the cases handled. Research data governed by rules re: export controls, cultural sovereignty, endangered species, etc., are fewer, but important nonetheless.

What are researchers seeking help with re: their sensitive data?

Researchers requesting support often ask for help establishing a secure computing environment OR accessing secure storage. However, it frequently turns out that secure compute AND storage needs go together. Other impacted aspects of the research workflow include data collection, data transfer (as part of a collaborative project or from the researcher’s previous institution), and data sharing/publication. For clinical and health-related research, communications about the data are sensitive and important. Not greatly represented yet, but on the horizon, are issues related to data (records) retention, mandatory annual security reviews (likely to be unfunded by the data owner), and the increasing likelihood of re-identification of individual subjects as more and more personal data becomes available through the Web and through third-party aggregators.
N = 158

[Add final observations / transition to next section?]

Campus policy, guidance, roles and responsibilities
A primary observation of the current state is that policy guiding research data is inadequate. This has resulted in a large degree of uncertainty among researchers and offices responsible for managing and approving agreements of various kinds. …

Current UC Berkeley Secure Research Data Policies
… Rachael’s doc …

Research Data Classification Standard
Across campus, a number of offices and individuals are involved in reviewing data protections required for research projects. Currently, they refer to the campus Data Classification Standard which is published and maintained by Information Security and Policy. The purpose of the Data Classification Standard is defined as follows:

The Berkeley Data Classification Standard is a framework for assessing data sensitivity, measured by the adverse business impact a breach of the data would have upon the campus. This standard provides the foundation for establishing protection profile requirements for each class of data.

The defined Protection Levels range from Protection Level 0 (PL0) for public information to Protection Level 3 (PL3) for systems with an extremely adverse business impact. The examples given in the standard to do not include examples related to research. Because the Data Classification Standard was written primarily for administrative and student data, there are nuances related to research data that are not addressed. First, the primary criterion of “adverse business impact” is one that needs to be translated to be meaningful to research use cases. Similarly, within the current standard, the Protection Level 2 (PL2) is defined as data sets with a high adverse business impact but is characterised as those having a statutory requirement for notification in the case of a breach (e.g., social security numbers, financial account information). However, this notice-triggering requirement is not adequate to identify research data sets that need a high level of protection.
Another shortcoming of the current standard is that the actual protections required for research data sets can vary significantly (as defined by requirements from data providers). Many restricted data sets would be assigned to Protection Level 2. Unfortunately the protections and controls mandated by campus policy in the Minimum Security Standards for Electronic Information (MSSEI) for PL2 data sometimes exceed what is actually required by the data provider. Sometimes these additional controls are desirable from a campus risk performance. Other times, they place a set of constraints on research that are probably not necessary and can be very expensive and difficult to implement.