

Implementing a Secure Data Enclave with Columbia University Central Resources

Presentation Notes, May 29, 2013, IASSIST 2013, Cologne, Germany

Rajendra Bose

Manager, Research Computing Services

Introduction

These slides were presented as part of IASSIST 2013 Session C, Facilitating Access to Sensitive Data, chaired by David Schiller, IAB, Nuremburg, Germany. We plan to submit an article providing more details about this presentation topic to IASSIST Quarterly within the next month.

Columbia's approach builds on IASSIST presentations and panels during the last three years, and we thank our colleagues for sharing their experiences and technical approaches.

A related IASSIST 2013 session was the Session A Panel: Managing Access to Restricted Data in Universities, chaired by Katherine McNeill, MIT Libraries.

As mentioned in the panel, some institutions (NORC, ICPSR, others) act as repositories or clearinghouses for restricted data and are building secure data infrastructures so that researchers other institutions can use their services remotely. However, some universities like Columbia and Cornell are moving forward and developing their own secure data enclaves, partly because of the growing need for such services in research disciplines beyond the social sciences, and partly because researchers might create their own restricted data and would like to store it at their home institution. Dialogue and discussion about these two trends by the IASSIST community will be important in the months and years ahead.

Slide 2/10

Planning for the current Columbia Secure Data Enclave pilot service began in early 2012. The SDE pilot has been configured for a maximum of ten users (this was defined by the number of available free Citrix XenDesktop licenses) over a year timeframe to assess whether the system meets the needs of social science researchers. A decision point has been set for early Fall 2013 for the faculty and central IT division to decide whether to continue and expand the service or to not proceed with the service. The pilot system has limited storage of 1 TB. The system has been designed to scale up if expanded. The SDE pilot targeted the acceptance of the SDE computing environment in applications for restricted data submitted by researchers to the US Bureau of Labor Statistics and the custodians of the Add Health datasets commonly used by the Columbia social science community.

The user experience of our SDE is straightforward: the screenshot shows that a researcher with an SDE account logs into the system using a Citrix plugin for a web browser and gains access to a limited Windows environment, pre-populated with office (Microsoft Word, Excel and Powerpoint) data analysis (R and Stata) and utility applications. Users have access only to special directories set up by a Data Security Officer who acts as data custodian, and the

virtual Windows desktop in the browser includes limitations such as no cut and paste capabilities, no access to files or drives from the local desktop or laptop machine, and so on.

The SDE service pilot has been certified by an internal Columbia group at the Medical Center to be compliant with HIPAA, PHI and PII security standards.

Slide 3/10

The presentation focuses on several points:

- Success at implementing the SDE service pilot was an iterative procedure, and required working with the unique organizational structure of the university and having senior social science faculty communicate the need for the SDE service and engaging directly with central IT and university leadership.
- Columbia's approach uses university central IT resources, including teams of network and security specialists and a virtual machine infrastructure located in the university data center.
- The SDE infrastructure includes staff fulfilling specialized roles.
- A brief technical overview of the system is provided at the end of the presentation.

Slide 4/10

Faculty sponsors of the SDE include social science faculty affiliated with the Columbia Population Research Center (<http://cupop.columbia.edu>). Roughly 50 faculty affiliates and an equal number of additional research staff and students span multiple campuses, schools, departments and centers at the university.

Slide 5/10

At Columbia the Provost is the chief academic officer and oversees all academic schools, departments, institutes and centers, as well as the libraries which most people recognize as an important part of the university's research support infrastructure. Other administrative officials oversee other parts of the research infrastructure including the various units under the Executive VP for Research and the central IT division.

Slide 6/10

The SDE service pilot infrastructure makes use of central IT (CUIT) resources and several EVPR groups. The libraries are not yet involved in the SDE.

Slide 7/10

The roughly 300 person CUIT organization at Columbia includes several divisions and groups; for the SDE the involvement of the Information Security Office, Systems, Storage and Network Engineering groups in the design and implementation of the system was critical. The Research Computing Services group interacts with faculty sponsors and acts as overall coordinator of the service with the other CUIT groups.

Slide 8/10

In this project we gained a better understanding and appreciation of the various roles that a data custodian and other staff fulfill during the research process. This includes assisting the

researchers as they interact with the data-granting agency and/or the university to apply for restricted data, and helping the researchers use the SDE if the application proves successful. The DSO creates the Data Protection Plan used in the restricted data application, and authors the SDE Use Agreement that the researcher signs before using the system. For the purposes of the SDE pilot we aggregate several different roles into one Data Security Officer (DSO) role. If the system is ultimately expanded than multiple staff will likely be needed to fill the different roles.

Slide 9/10

The DSO creates accounts for SDE users, sets up a predefined directory structure for each account, acts as data custodian and loads restricted data into the appropriate directory, and responds to user requests to move files off the system. At this point the DSO does not vet system output, and the DSO must make manual home directory backups. These topics will be revisited if the pilot system is expanded.

Slide 10/10

Only a brief technical overview is provided in this presentation: SDE pilot users log into the system over the campus network. SDE pilot users who are faculty or staff can log into the system using a VPN connection when off-campus. The SDE pilot system consists of three components: (1) the SDE Server, (2) the SDE Gateway, and (3) the Isolated SDE Network. These components are located in the secure university data center. The SDE Server and Gateway are virtual machines (VMs) within the CUIT VM infrastructure. Users access the system through Citrix XenDesktop software on the SDE Gateway. When a user logs in, a temporary VM is created for the SDE session. The Isolated SDE Network is a closed network with no access to the internet (except through the SDE Gateway), and is regularly scanned for vulnerabilities by the Information Security Office using standard procedures. The DSO performs administrative functions from an office location outside the data center. CUIT system administrators do not have accounts with access to restricted data; only the DSO and individual SDE users have such accounts.

Acknowledgements

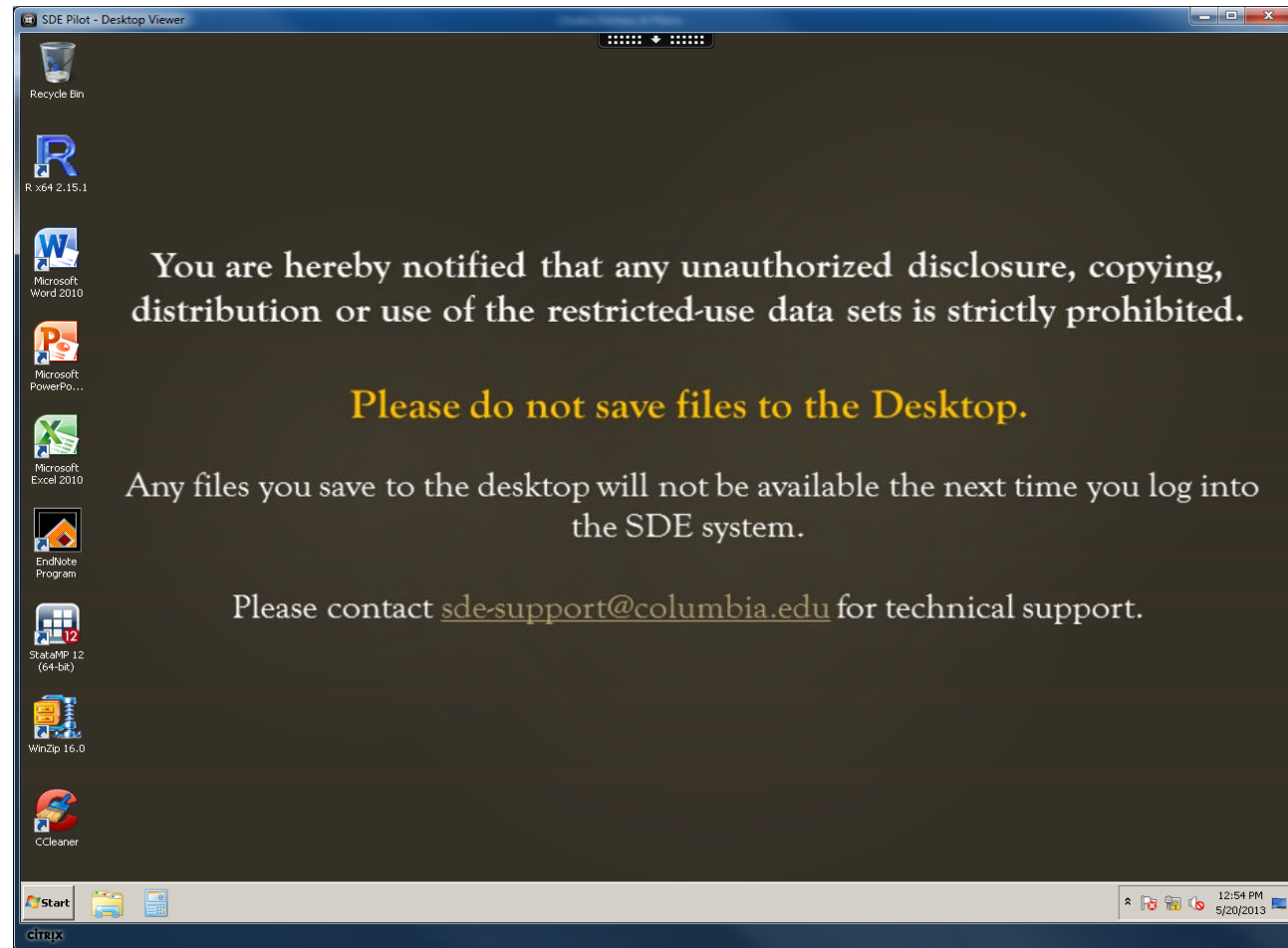
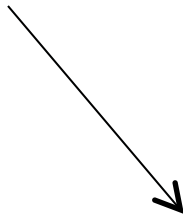
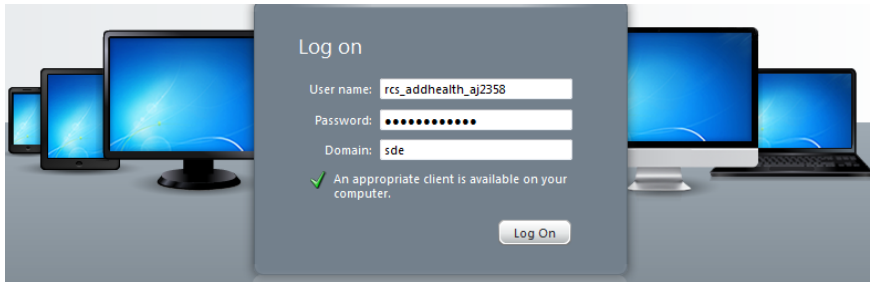
We would like to acknowledge members of the technical design team within CUIT: Terrence Ford, Michael Higgins, Abhishek Joshi, Rob Lane, Patrick Rausch, Jonathan Reams, Joel Rosenblatt, Aziz Usmani.

Implementing a Secure Data Enclave with Columbia University Central Resources

Rajendra Bose
Manager, Research Computing Services

May 29, 2013
IASSIST 2013, Cologne, Germany

Secure Data Enclave Service Pilot 2/10



Implementing an SDE with Central Resources

- Iterative process, social science researchers lead effort
- Use of centrally managed infrastructure
- SDE infrastructure includes specialized staff with different roles
- SDE system overview



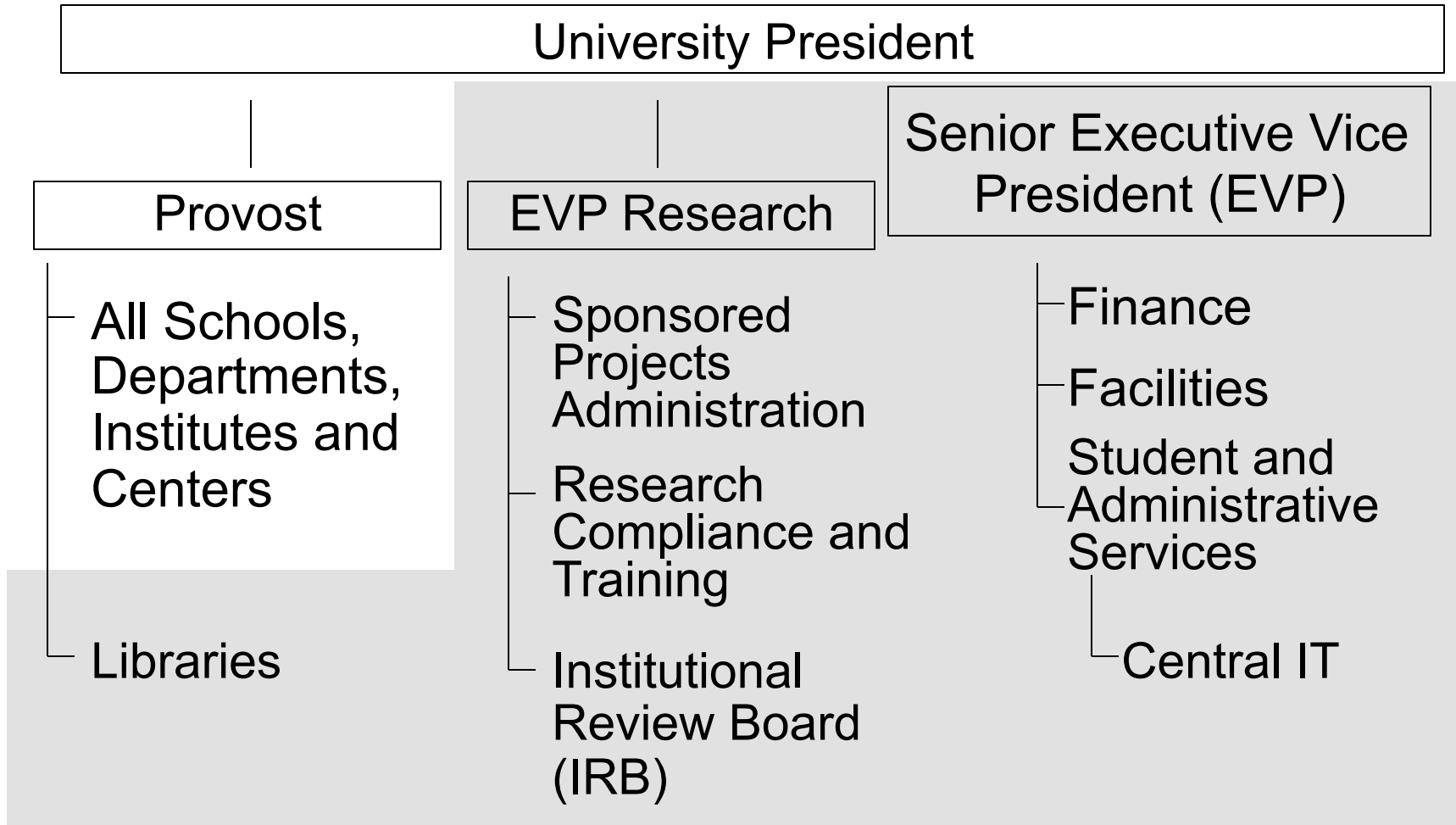
**Columbia University
Medical Center**

**Columbia University
Morningside +
Manhattanville
campuses**



University Structure

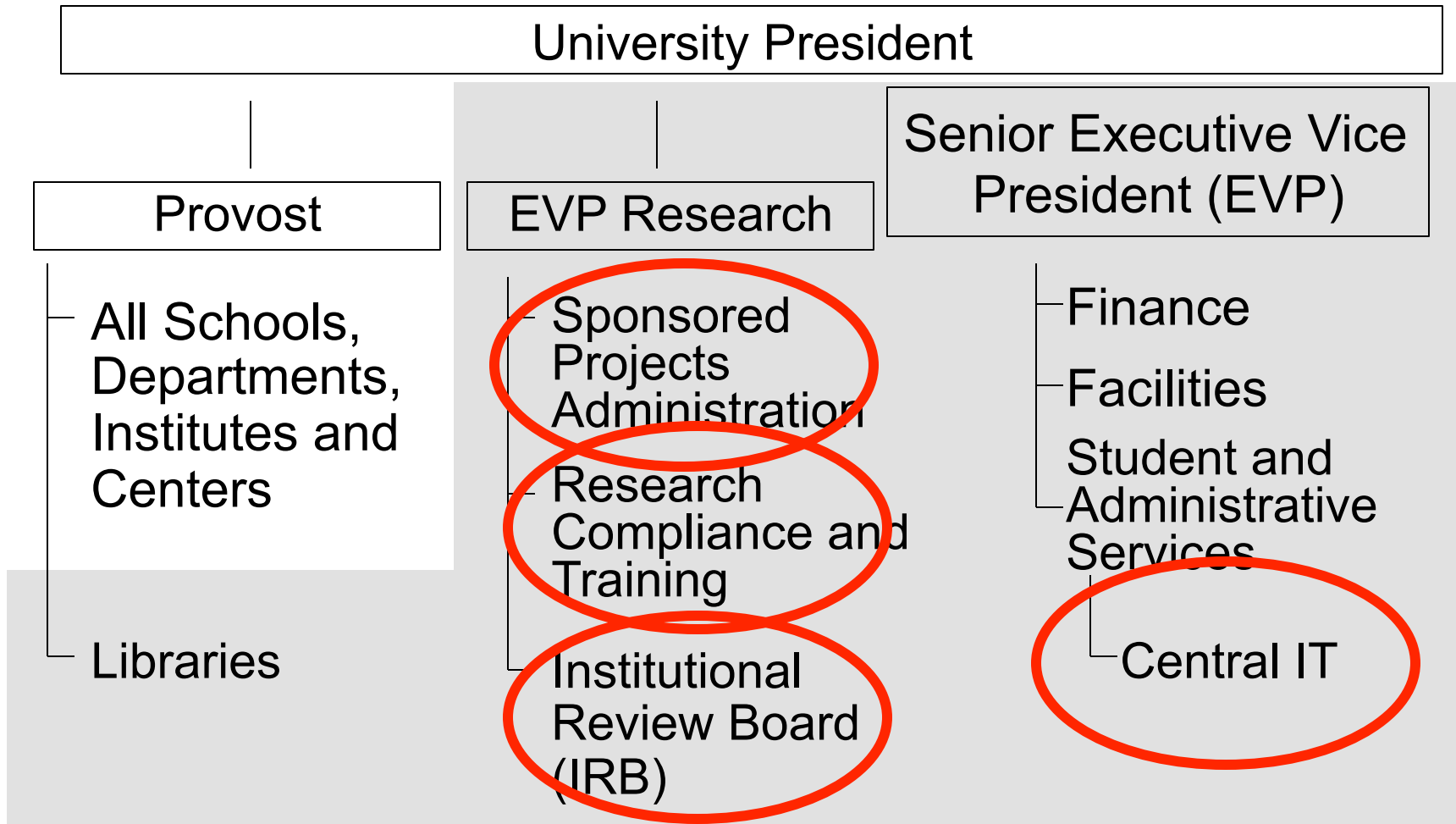
5/10



Research Infrastructure

University Structure

6/10

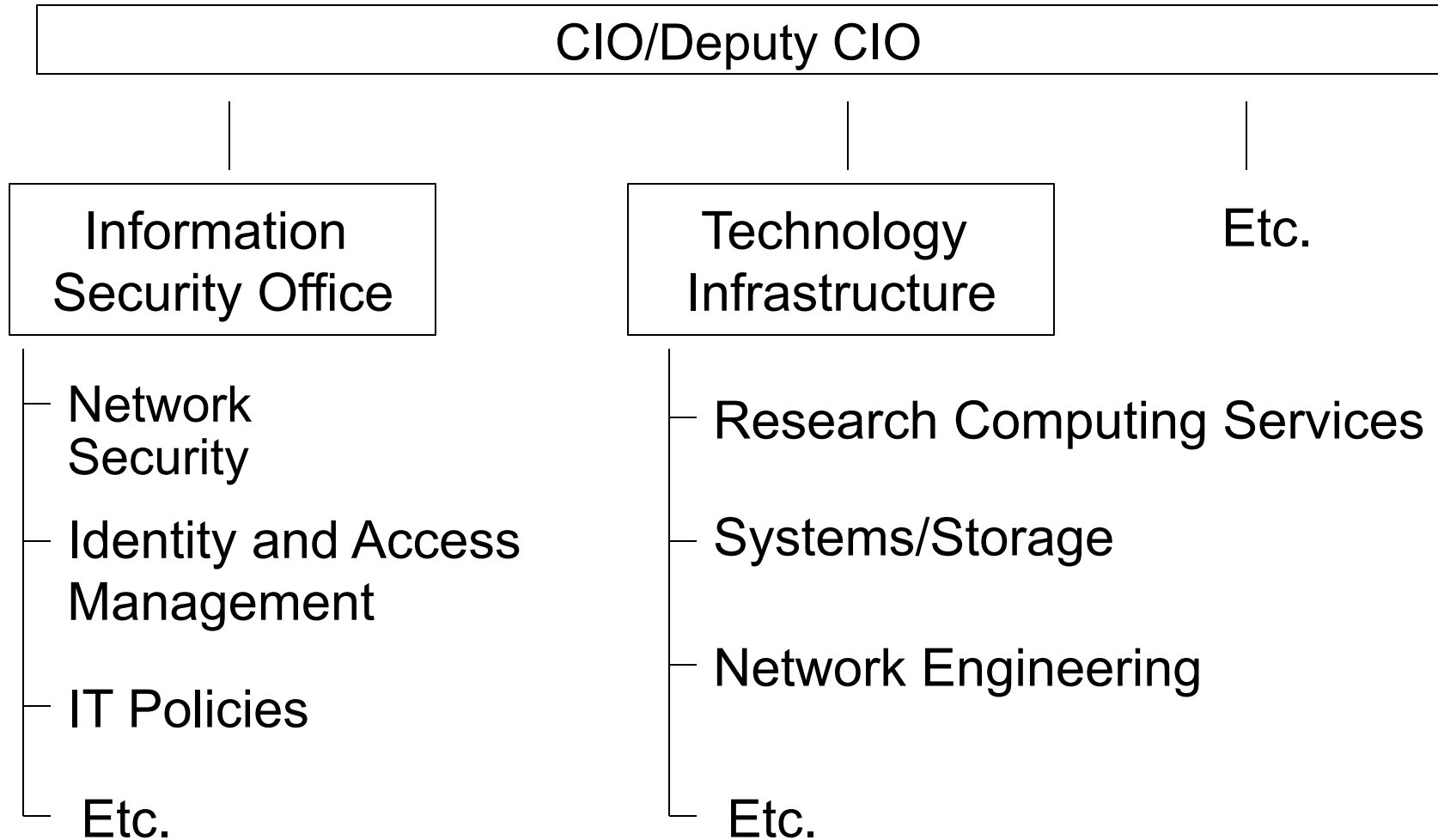


Research Infrastructure

└ **Secure Data Enclave Infrastructure**

Central IT Structure

7/10



Restricted Data
Application

Data Protection Plan

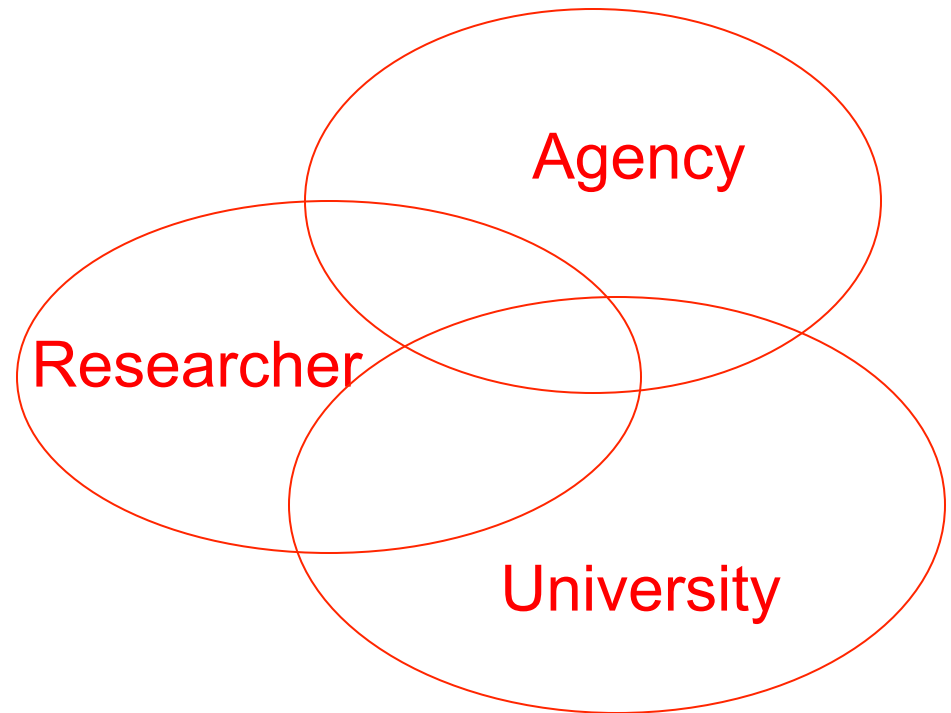
IRB protocol approval

SPA Approval

Restricted Data Use
Agreement

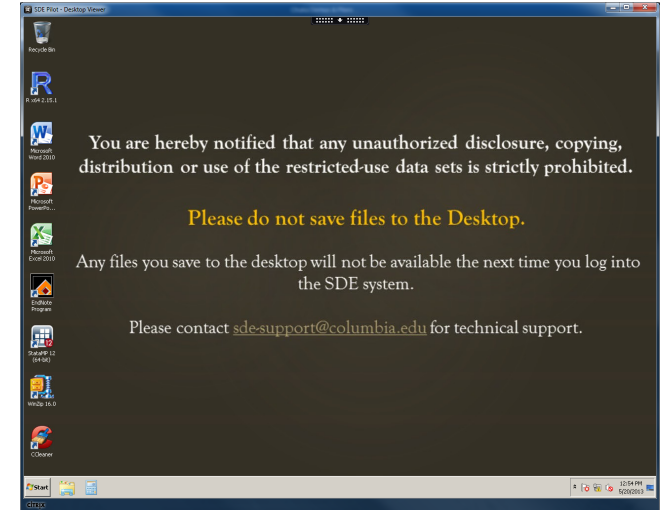
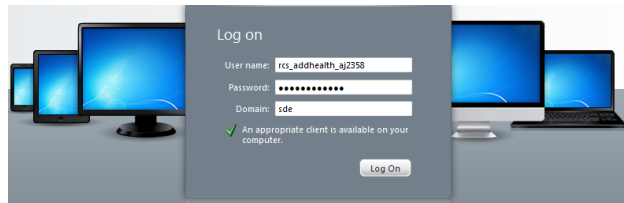
SDE Use Agreement

Data Security Officer (DSO)



SDE Directory Structure

9/10



Data

**Read only:
original data set**

Work

Researcher-generated data subset(s)

Home

**Statistical code files, Word
documents, etc: Backed up**

Output

Holding place for DSO output requests

Group-work

Sharing files and data with research group

SDE System Overview

10/10

