Facilitating collaboration with restricted-use data

John E. Marcotte
ICPSR
University of Michigan
IASSIST 2017

ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

Why Collaboration

- Team or collaborative research is often better science
- Cross-discipline collaborations take advantage of different expertise
- In the United States, the *National Institutes of Health* are stressing collaborative research

Facilitating collaboration with restricted-use data

- Security: Requirements and Challenges
- Typical security plans
- Cloud computing(with security enhancements) for collaboration
- Discussion

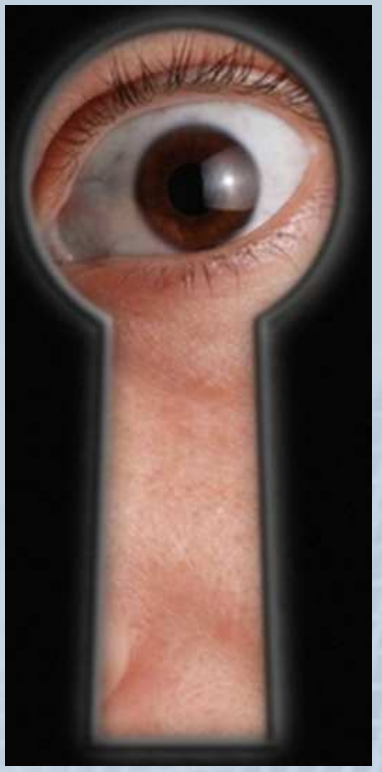
Why Security

- Disclosure Risk
- Sensitivity
- Potential Harm
- Required by Data Provider

Security Purpose

- Prevent data breaches
- Prevent disclosure

Data breaches and disclosure hurt all research



Security Requirements

- **Encryption**: transmittal and storage
- **Internet**: blocked so files cannot be copied directly to Internet sites
- **Output vetting for disclosure**: researcher self reviews or third party
- **Monitored processing**: prevent unauthorized materials; not always required

Typical Security Plans

- Non-networked computer
- Cold room
- Census-type enclave

Impede Collaborations

- Plans make collaboration difficult
- Hinder sharing output and writing

Challenges for Security

- More data analysis leads to better public policy and better science
- Team or collaborative research is often better science
- Provide access while maintaining security

Goals

- Facilitate team research by providing collaboration space
- Sharing working files including output in computing environment that meets security requirements
- Sharing writing

Bring Researchers to Data

- Researchers *come* to the data instead of *sending* data to researchers
- Researchers can collaborate virtually from different locations

Cloud Computing

- Where's the cloud
- Who's cloud
- Cloud security
- Resources rented

Security Enhancements

- Encrypted connection to cloud
- Two factor authentication
- Simulated non-networked computer
- Researchers cannot copy files in or out
- Third party output vetting is possible

Cloud Restrictions

- Specified end-point for connection
- Access not allowed from pubic places such as libraries and cafes
- Researchers agree not to remove unauthorized notes

Cloud Costs

Operating a cloud with sufficient security does have costs

- Setup of cloud
- OS licenses
- Software licenses
- Staff to maintain systems
- Staff to vet output

Data *users* may have to pay, but some data *providers* pay for researchers to be able to access the data

Share Output

Team Writing

Cloud Technology

- ***VMware***
- ***Citrix***
- ***NoMachine*** (NX)
- Other

Software

- Productivity suite
  - Word processor
  - Spreadsheet
  - Presentation
- Quantitative Analysis
  - SAS
  - SPSS
  - Stata
  - R
  - Other
- Qualitative Analysis

Software Limitation

- Researchers are limited to what's available in that "cloud"
- Extras such as Stata ado files and R modules must be vetted before being added
- Researchers have to wait for updates and additions
- Not all software will be available

Virtual Environments Plus

- More powerful computing than available on typical desktop
- Cloud could have multiple computers including multi-node Linux clusters

ICPSR VDE

- **ICPSR** offers a *Virtual Data Environment (VDE)* for restricted-use data
- ICPSR VDE can also serve as collaboration space
- ICPSR VDE is built on *VMware*

Discussion