J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

ILR LDI

Physically Protecting Sensitive Data

Jim Shen and Lars Vilhuber

February 8, 2021

(cc) BY-NC

Five Safes: In combination, stronger

ONS, The Five Safes: Ensuring Safe Use of Data, http://www.bris.ac.uk/media-library/sites/cmpo/documents/mcivor2018.pdf (Accessed 2021-02-07)

Five Safes: In combination, stronger

- Focus here: SAFE SETTINGS

ONS, The Five Safes: Ensuring Safe Use of Data, http://www.bris.ac.uk/media-library/sites/cmpo/documents/mcivor2018.pdf (Accessed 2021-02-07)

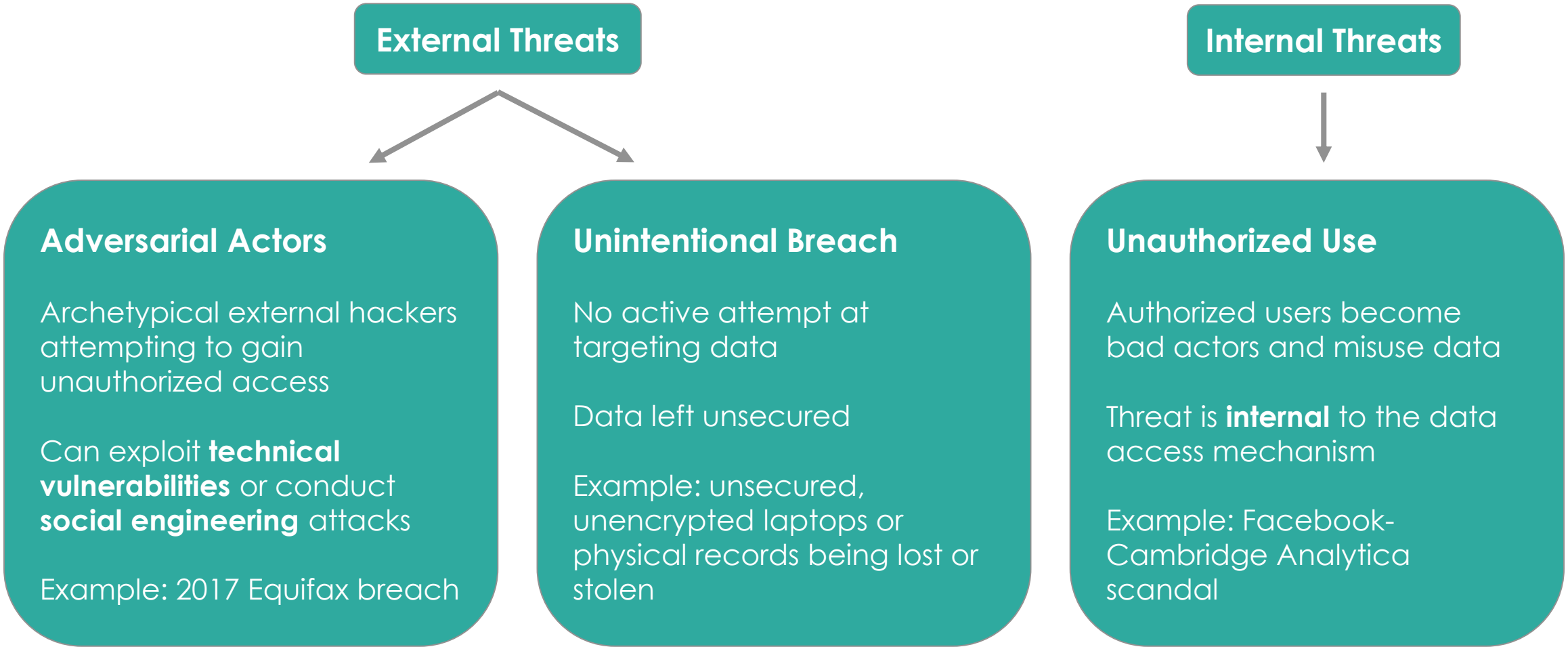
Physically Protecting Sensitive Data

- The physical protection of sensitive data is one of the key parameters that data custodians can and do influence
- Within the Five Safes Framework, "safe settings" are heavily influenced by how data are physically protected
- However, it is also a parameter that is very dependent on current state of technology, the types of threats, and interactions with the other Safes
- **Knowledge of the technological possibilities is of importance for negotiating access to administrative data that does not have an existing access mechanism**

Physically?

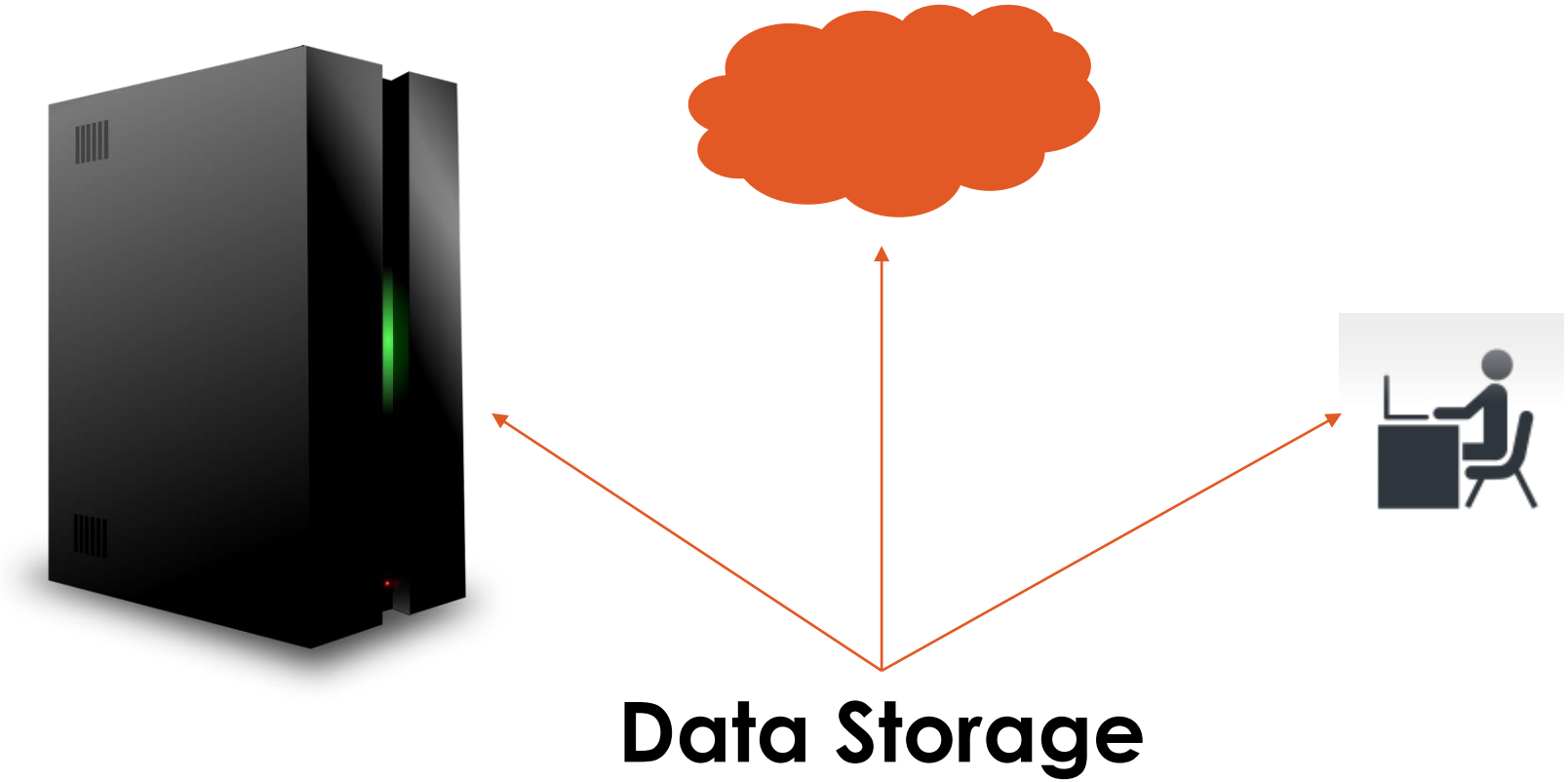
- In contrast to “statistically” or “computationally”
  - See Chapter 5 “Balancing Privacy and Data Usability: An Overview of Disclosure Avoidance Methods” (Ian Schmutte and Lars Vilhuber, Webinar on 2020-11-02)
  - See Chapter 6 “Designing Access with Differential Privacy” (Alexandra Wood and co-authors, Webinar on 2021-02-01)
- Includes
  - IT security measures
  - Building security measures
  - Choice of locations

Types of Security Threats



Connecting Researchers with Data

Connecting Researchers with Data



Data Storage

- Physical Media
  - Attached storage (e.g. hard drives, solid state drives)
  - Removable storage (e.g CD's, USB drives)
- Cloud Services
  - Proprietary (e.g. AWS, Google Drive, Dropbox, OneDrive)
  - Open Source (e.g. Nextcloud)
- Reliability and security
  - Prevent data loss and system uptime
  - Prevent unauthorized access to data
- Encryption!

Connecting Researchers with Data

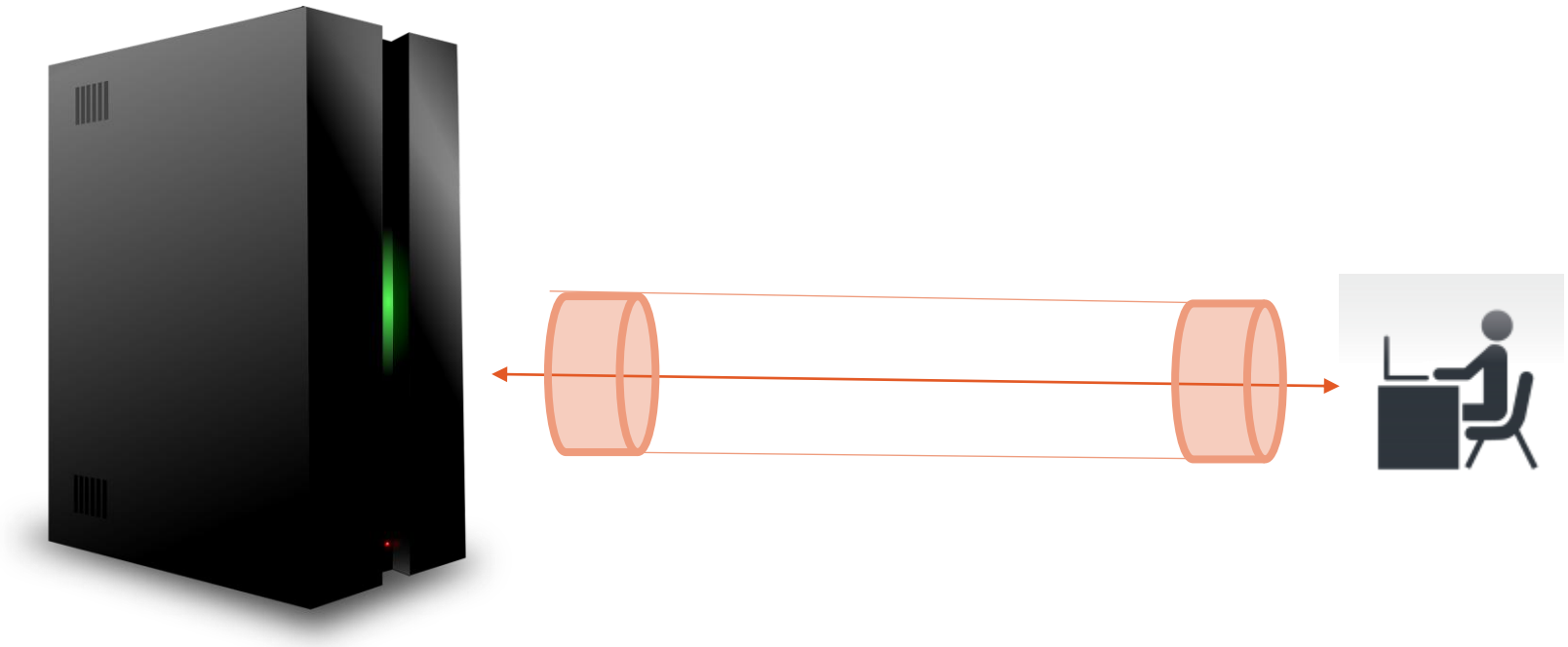
Data transfers

Data Transfer

- Physical Media
  - Removable media can be used to transfer data
- Electronic Transfer
  - Encrypted network protocols (SFTP, HTTPS, VPN)
  - Cloud services

Never send data unencrypted!

Connecting Researchers with Data



Data Access

Electronic Access (network security)

- Virtual Private Networks
  - Exchange data over public networks as if directly connected on a private network
- IP Address Restrictions
  - Restrict allowed IP addresses with allow list or deny list

Encryption

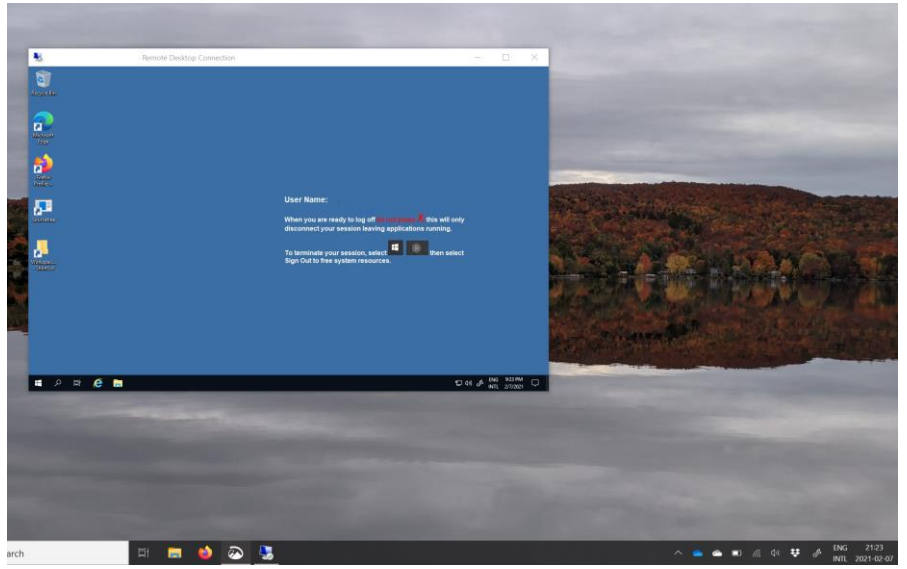
- Minimum security requirement for any data access mechanism
- Full Disk Encryption
  - Software-based (Filevault, Bitlocker, various Linux options)
  - Hardware-based (requires specialized hardware, removes memory as attack vector)
- File Level Encryption
  - Encrypt individual files, only decrypt when in use
  - Examples: GnuPG, VeraCrypt
- Cloud services

Electronic Access (local security)

- System isolation
  - Researcher accesses only as allowed/trusted
  - Research system separate from administrative systems
  - User access isolated from each other
- Technical means of achieving isolation
  - Data Access Controls: Regulate what users can view or use in a computing environment
  - Physical system isolation: Stand-alone computer, dedicated researcher machine
  - Virtual system isolation:
    Virtual machines/ Virtual Desktop Infrastructure/ Docker/ chroot

Electronic Access (Connecting)

- Remote Desktop
  - Enable users to connect to another computer over a network
  - Avoids need to
    - Transfer data to researcher
    - Store data at researcher site
  - Subject to network issues (slow, lag, down)



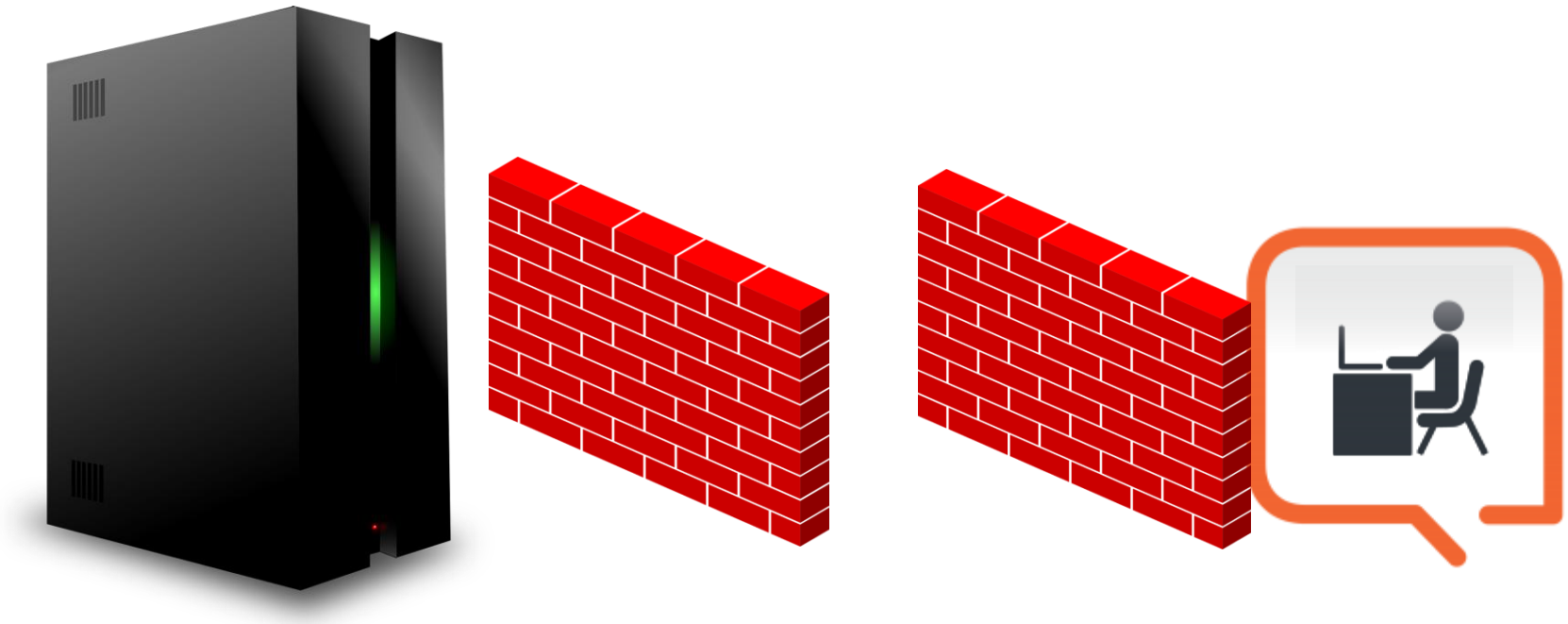
- Thin Clients
  - Optimized for utilizing remote desktop software

Electronic Access (Connecting… sort of)

- Remote Processing / Query system
  - Only code is sent
  - No interactive work
  - May have job limits

```
#PBS -N MyJobName
#PBS -P MyJobProject
#PBS -q queue_1
#PBS -l instance_type=c5.18xlarge

# CD into current working directory
cd $PBS_O_WORKDIR

# Prepare the job environment, edit the current PATH, License Server etc
export PATH=/apps/softwarename/v2020/
export LICENSE_SERVER=1234@licenseserver.internal

# Run the solver
/apps/softwarename/v2020/bin/solver --cpus 36 \
     --input-file myfile.input \
     --parameter1 value1

# Once job is complete, archive output to S3
BACKUP=1
if [[ "$BACKUP" -eq 1 ]];
  then
     aws s3 sync . s3://mybucketname/
fi
```

Connecting Researchers with Data



Physical Access

Physical Access

- Secure Rooms
  - Hardened location for data storage and access
  - Various specifications for physical and electronic protections

Physical Access

- Secure Rooms
  - Hardened location for data storage and access
  - Various specifications for physical and electronic protections
- Physical Access Cards
  - Electronic cards that identify the card bearer
  - Card reader validates card with central database

Physical Access

- Secure Rooms
  - Hardened location for data storage and access
  - Various specifications for physical and electronic protections
- Physical Access Cards
  - Electronic cards that identify the card bearer
  - Card reader validates card with central database
- Biometric Authentication
  - Physical and biological features unique to individuals
  - Can be used to authenticate users

Technical Features of Data Access Mechanisms

- Data Storage
  - Physical Media
  - Cloud Services
- Data Transfer
  - Physical Media
  - Electronic Transfer
- Encryption
- Electronic Access
  - Data Access Controls
  - Virtual Private Networks
  - IP Address Restrictions
  - Remote Desktop
  - Thin Clients
- Physical Access
  - Secure Rooms
  - Physical Access Cards
  - Biometric Authentication

Typical Access Mechanisms

Typical Access Mechanisms

- Remote Execution
  - Researcher submits a request to the data custodian
  - Data custodian runs analysis with automated service or staff executing by hand
- Physical Data Enclave
  - Researcher travels to secure location to access and analyze data
  - Data custodian maintains infrastructure and full control of data
- Virtual Data Enclave
  - Researchers remotely access and analyze data
  - More flexible than a physical data enclave
- Researcher Provided Infrastructure
  - Researcher provides infrastructure for storage and analysis

Five Aspects of Physical Security

- The level of **researcher agency over analysis computers**
- The **location of analysis computers and data**
- The **location of access computers**
- The **level of security of access locations**
- The **range of analysis methods available** to researchers

- For each aspect, data access mechanism is classified into three categories
- Weakly aligned with how restrictive it may be on the researcher and how much control the data provider exerts

Researcher Agency over Analysis Computers

- Analysis computers hold and analyze researcher accessible data
- Data custodians determine the level of control that researchers are allowed
- Low Agency
  - Limited to the software that the data provider allows
- Medium Agency
  - May allow some choice or limited configuration options
- High Agency
  - Few restrictions, researchers may own the computer or have administrative privileges

Location of Analysis Computers and Data

- Each data location comes with its own requirements, tradeoffs, and special considerations for the researcher and data provider.
- The location of the data on its own does not define how researchers access the data, or the type of analysis a researcher can conduct.
- Data Provider
  - Retains custody of analysis computer and data, acting as data custodian
- Third-Party
  - A third party acts as the data custodian, potentially serving multiple researchers and data providers
- Researcher
  - Researchers hold data, reducing costs on data providers but relying on enforcement of data use agreements

Location of Access Computers

- When data are not in the same location as the researcher, access computers are distinct from analysis computers
- Ownership is not necessarily aligned with location
- Non-researcher data custodian
  - Researchers must travel to the data custodian to access data
- Third-Party
  - Data custodians and access providers can see efficiency gains
- Researcher
  - Access computers are located with researchers

Security of Access Computers

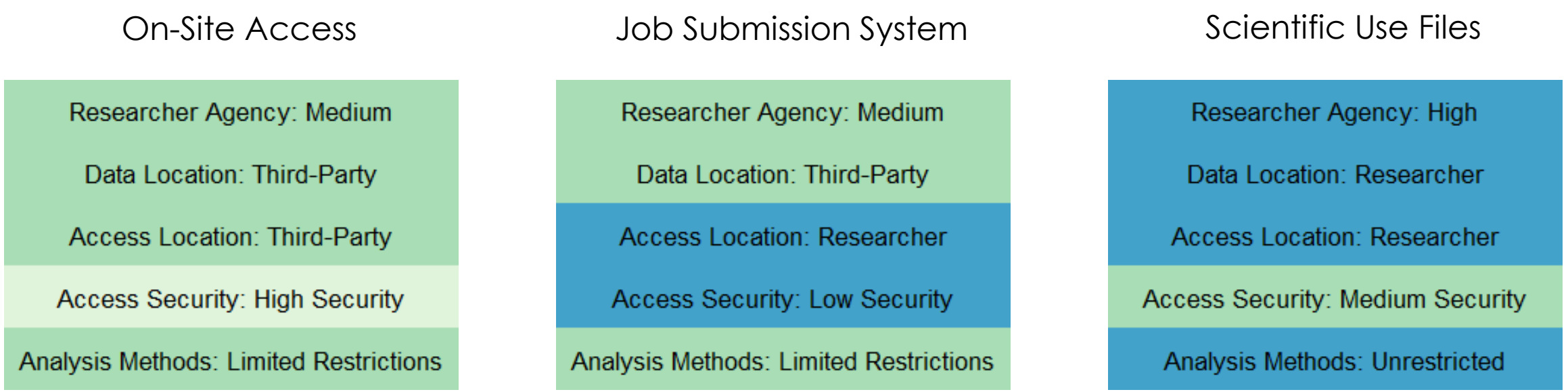
- These are not concrete distinctions between different mechanisms but broad classifications of the overall rigor of physical security regimes
- High Security
  - Strong specifications of physical security such as secure rooms with hardening beyond standard locked doors
- Medium Security
  - Defined location with access restricted to approved researchers, with some security features
- Low Security
  - Few or no physical controls, relying on enforcement of DUA's or no restrictions at all

Range of Analysis Methods Available

- Researchers may be able to leverage a wide range of analysis methods, ranging from simple tabulations to complex machine learning tasks.
- In other cases, they may be limited to a small set of methods, defined by the data custodian for technical or security reasons
- Highly Restricted
  - Strong limitations such as only whitelisted commands or running tabulations
- Limited Restrictions
  - Software elements may be censored, such as inability to inspect individual records
- Unrestricted
  - Researchers can use the full set of methods available provided on analysis computer

Examples Along the Five Aspects: RDC-IAB

- Acts as internal third-party for German Federal Employment Agency
- Provides three different access mechanisms for labor economic data
- RDC-IAB holds most sensitive data for on-site access and remote execution, makes less sensitive data available for researcher provided infrastructure



Examples Along the Five Aspects: OLDA

- Third-party data custodian that transfers de-identified, individual level data to researchers on behalf of Ohio
- Researchers provide local infrastructure for storage and analysis of the data
- Note: "Low security" does not mean "no security"

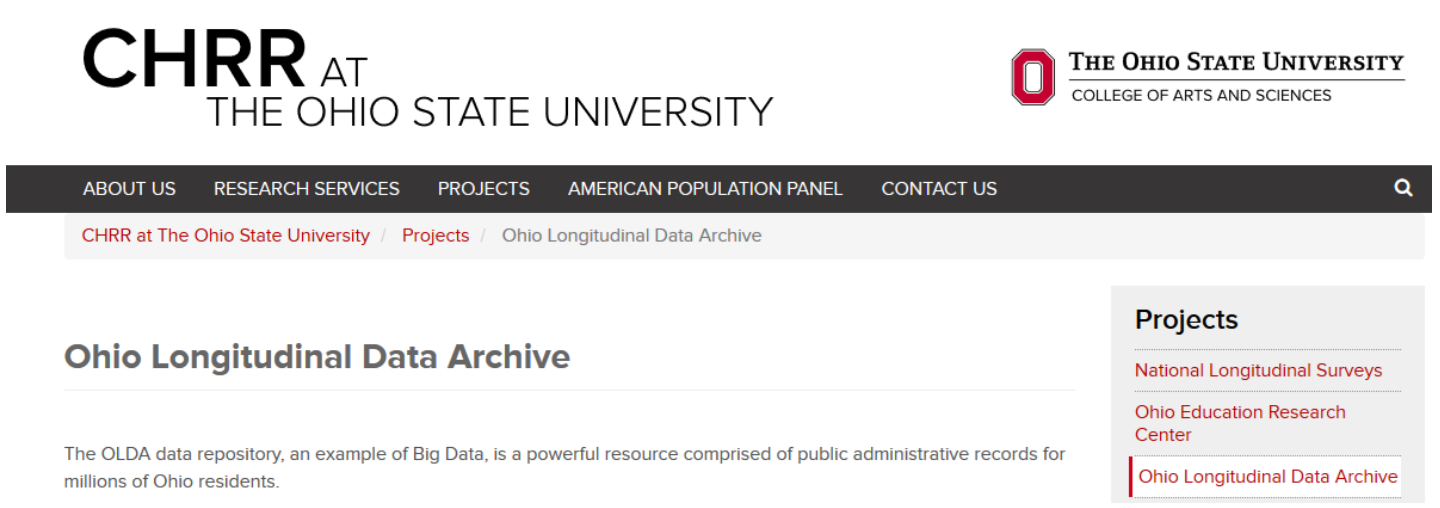
Researcher Agency: High

Data Location: Researcher

Access Location: Researcher

Access Security: Low Security

Analysis Methods: Unrestricted



Examples Along the Five Aspects: NB-IRDT

- Third-party data custodian for Province of New Brunswick
- Makes de-identified personnel and health data available to researchers
- Data held at, and researcher access at, secure NB-IRDT facilities

Researcher Agency: Medium
Data Location: Third-Party
Access Location: Data Custodian
Access Security: High Security
Analysis Methods: Unrestricted

Other Examples

- A wide range of examples from both the Handbook and selected outside examples
- Many options available for data providers and researchers when setting up new data access mechanisms

Data Access Mechanism	Researcher Agency Over Analysis Computer	Location of Data and Analysis Computer	Location of Access Computer	Access Security	Range of Analysis Methods Available
IAB RDC (chapter 7)	Medium	Third-Party	Third-Party	High Security	Limited
IAB JoSuA (chapter 7)	Medium	Third-Party	Researcher	Low Security	Limited
IAB SUF (chapter 7)	High	Researcher	Researcher	Medium Security	Unrestricted
OLDA (chapter 8)	High	Researcher	Researcher	Low Security	Unrestricted
NB-IRDT (chapter 9)	Medium	Third-Party	Data Custodian	High Security	Unrestricted
PCRI (chapter 10)	Medium	Third-Party	Researcher	Low Security	Limited
Aurora (chapter 11)	High	Researcher	Researcher	Low Security	Unrestricted
Stanford-SFUSD (chapter 12)	High	Researcher	Researcher	Low Security	Unrestricted
Cape Town (chapter 13)	High	Researcher	Researcher	Low Security	Unrestricted
DIME (chapter 14)	High	Researcher	Researcher	Low Security	Unrestricted
FSRDC	Medium	Data Provider	Data Custodian	High Security	Unrestricted
NCES	High	Researcher	Researcher	Medium Security	Unrestricted
RTRA	Low	Data Provider	Researcher	Low Security	Highly Restricted
SPN	Low	Third-Party	Third-Party	Medium Security	Unrestricted

Guidance and Examples

Guidance for Data Providers and Researchers

- There are many solutions that balance high security with relatively broad accessibility and convenience for researchers
  - RDC-IAB, NB-IRDT
- There are many examples of relatively simple but effective data access mechanisms with typically lower costs
  - OLDA, Stanford-SFUSD
- Data providers can allow researchers more flexibility in various aspects while maintaining the overall security of the system
  - RDC-IAB
- Necessary aspects of data access mechanisms and restrictions placed on researchers should be considered in the context of the other Five Safes
- Capacity for enforcing the DUA is an important factor for the flexibility of data access mechanisms

Thank you