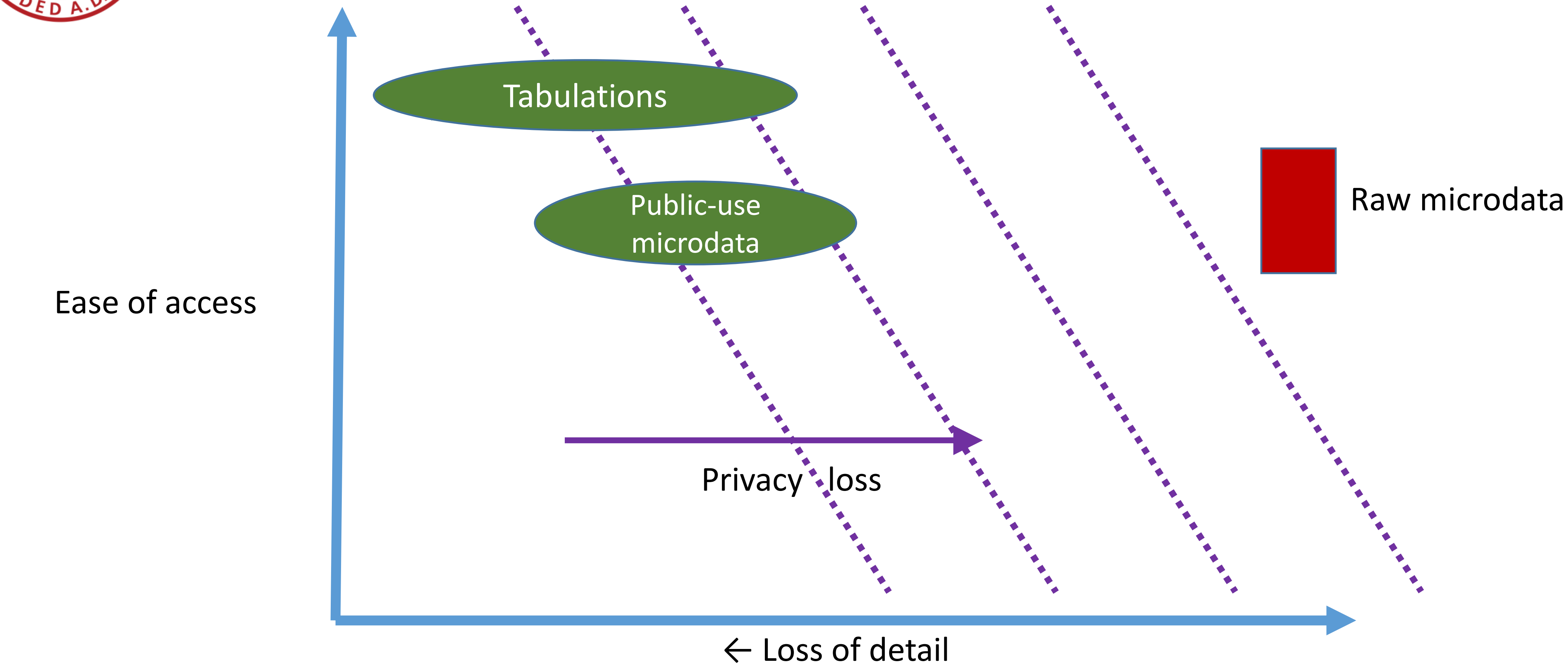
Confidentiality protection and physical safeguards

Lars Vilhuber

Cornell University

Funding acknowledged under NSF-#1131848 (NCRN) and a grant from the Alfred P. Sloan Foundation

publication trade-offs



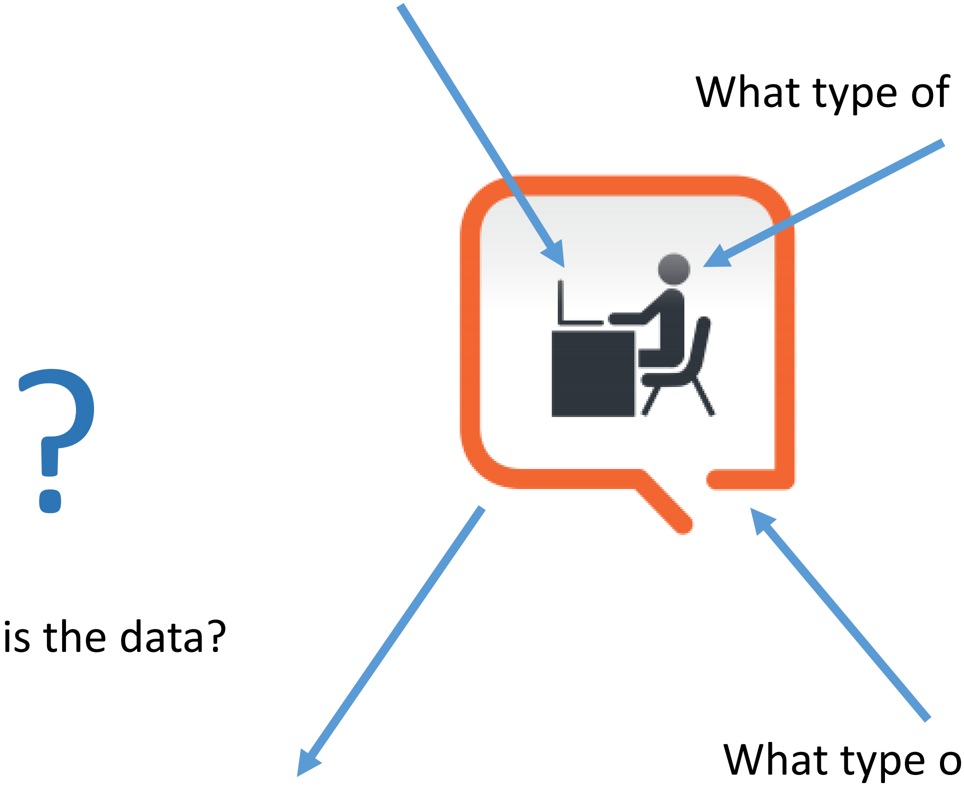
How to provide **easy and convenient** access to data with **more detail** than public-use microdata, **less privacy loss** than direct publication of raw data?

public use data

Data provider/
custodian

Data user/
researcher

confidential data



What type of access device?

What type of person?

Where is the data?

What type of room?

How do results leave the room?

basic paradigm

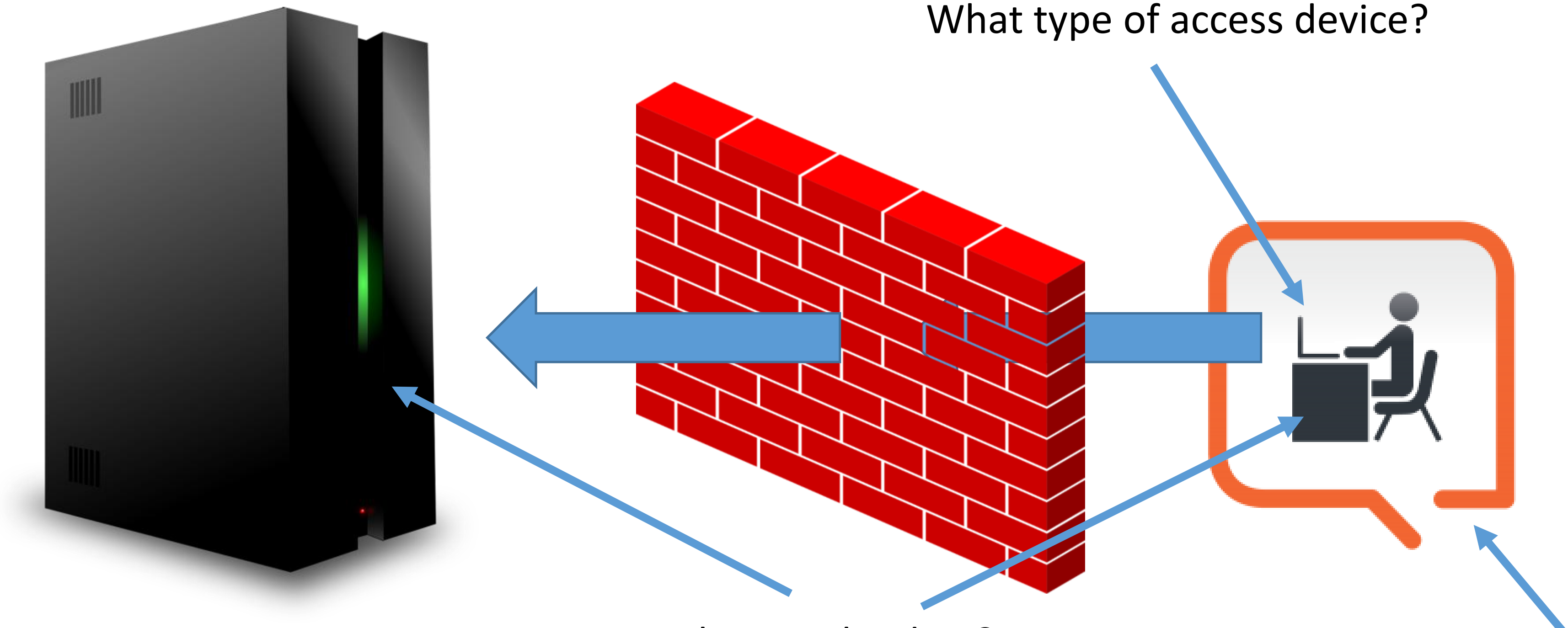
“Data Enclave” or “Secure Room”

Where is the data?

What type of room?

making things virtual

"Virtual Data Enclave"



What type of access device?

Where is the data?

What type of room?

virtual data enclaves

Synonyms:

VDI

(virtual desktop infrastructure)

Thin clients

Remote desktop

Examples in 1990s

Physical data enclaves

- BLS HQ
- BJS data access
- Department of Education data
- Census Bureau RDCs
- Canadian RDCs
- HRS restricted-access data
- and many more

Virtual data enclaves

(data remains in secure data center)

Examples in 2017

Physical data enclaves

- BLS HQ ?
- BJS data access
- Department of Education data
- ~~Census Bureau RDCs~~
- Canadian RDCs
- ~~HRS restricted-access data~~
- and many more

Virtual data enclaves

(data remains in secure data center)

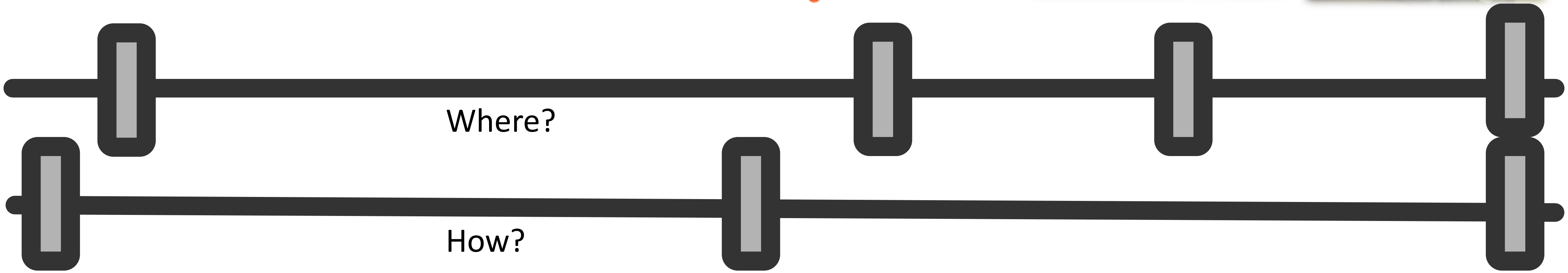
- Census Bureau/Federal Statistical RDCs (since early 2000s)
- German IAB RDCs (since mid 2000s)
- French CASD (since late 2000s)
- Cornell's CRADC, NORC (early 2000s)
- HRS restricted access data (2015)
- and many many more

basic levers

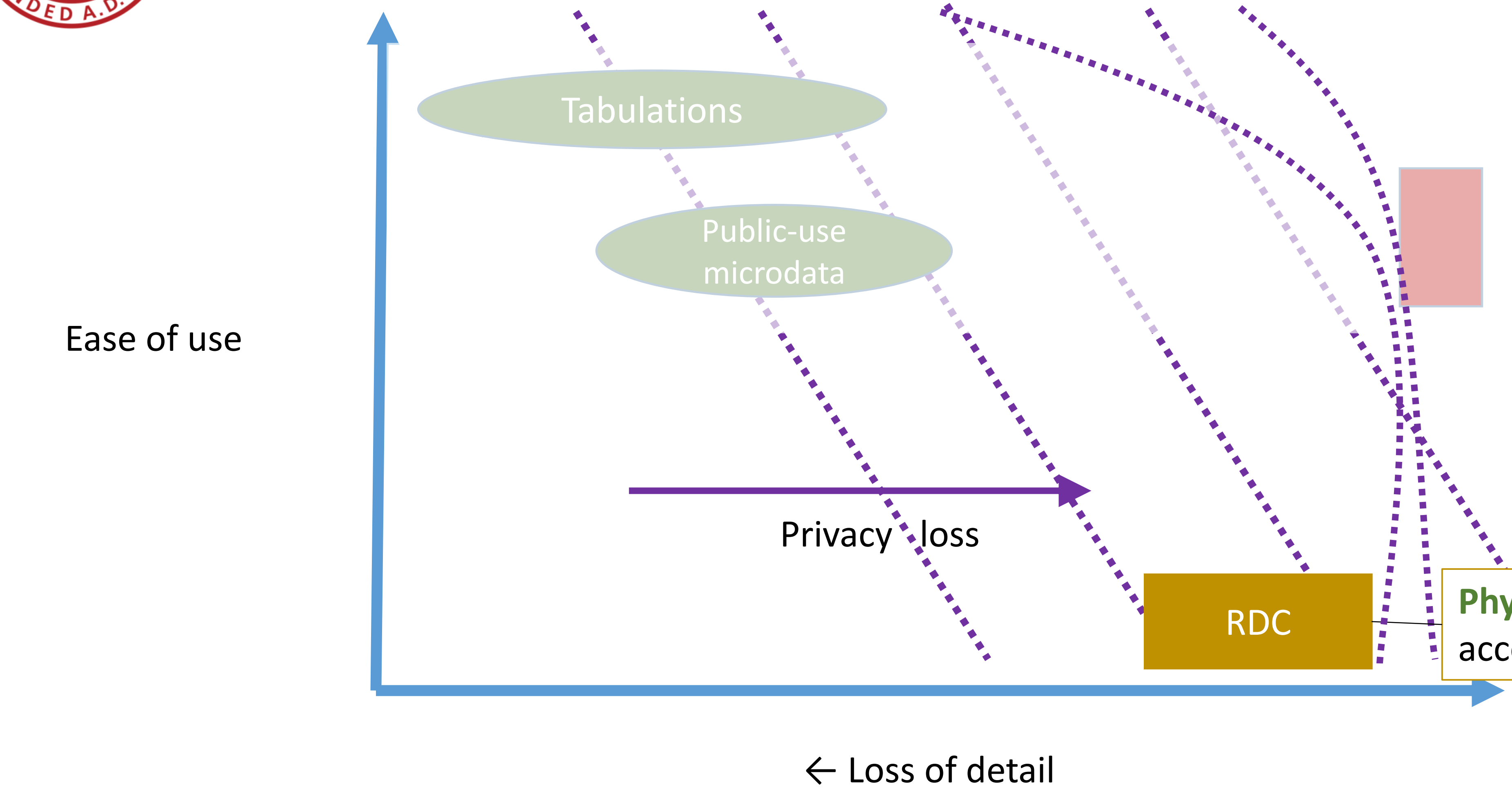
What type of access device?

What type of room?

basic levers

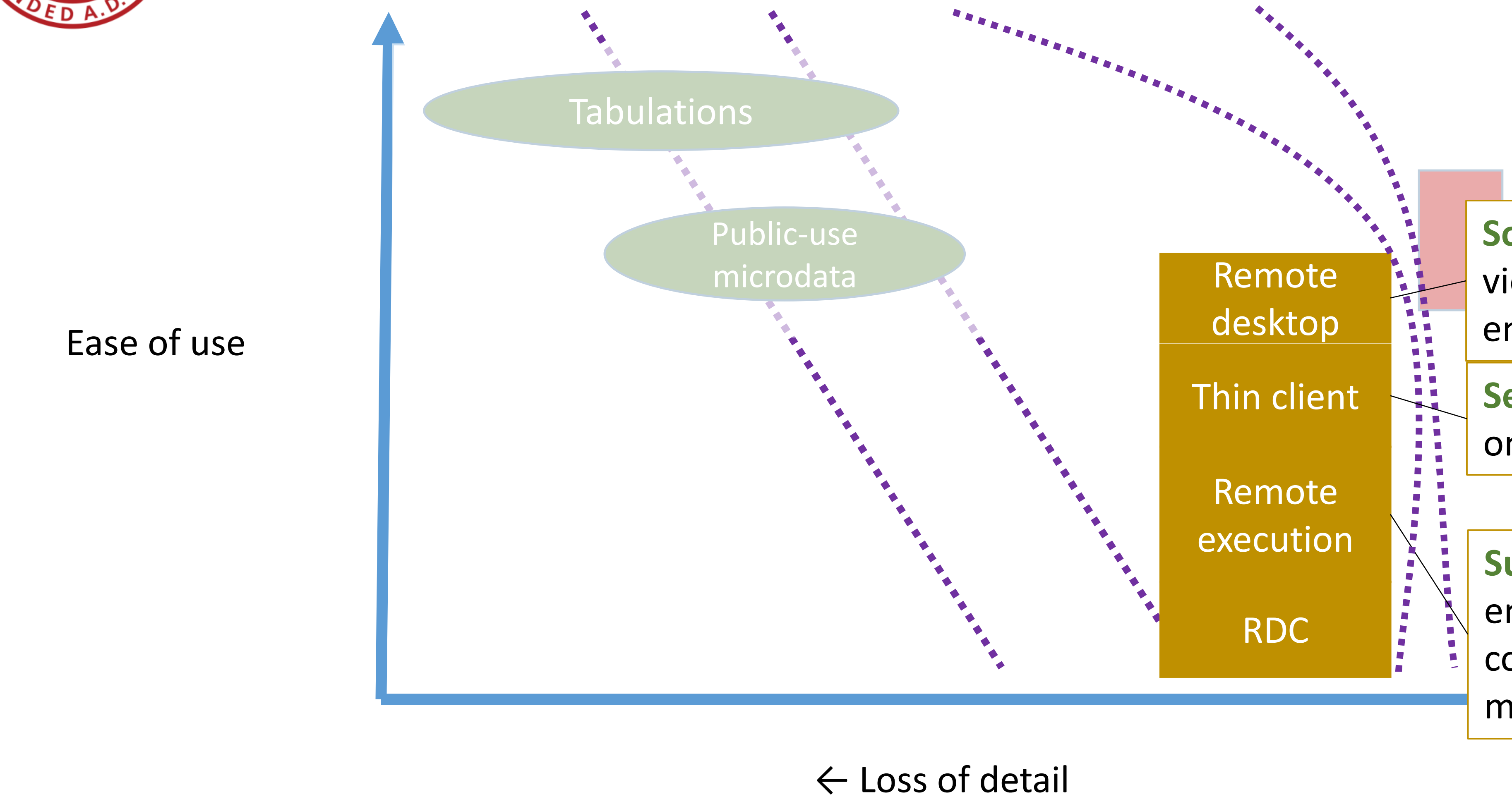


access methods: enclaves



Physically secure room housing access devices and/or data

access methods: enclaves



Software on your own PC giving a view onto secure data environment

Secondary secure PC giving a view onto secure data environment

Submitting analysis programs by email or through website (possibly combined with synthetic microdata)

What type of room?

Access matrix for confidential data

		Control by data provider of:			
	# access points	Access computers	Access rooms	Avail. analysis methods	Type disclosure avoidance
FSRDC researcher	24 sites (~700 users)	Full	Full (badge access)	Some (choice of software)	Manual/ variety of rules
Census staff researcher	n.d.	None (VDI)	None (VDI)	Some (choice of software)	Manual/ self/ variety of rules
IAB: JoSuA researcher	414 users	None (Web application)	None (Web application)	Smaller (software, whitelist commands)	Manual/ variety of rules
CASD researcher	371 sites (1471 users)	Extra Full (custom-built hardware)	Some (university office, EU)	Some (choice of software)	Manual/ variety of rules €300/ pack of 10
Stat.Denmark (typical EU)	?	None (VDI) - Some (host institution)	None (VDI) - Some (host institution)	Some (choice of software)	Manual/ self/ variety of rules

How do results leave the room?

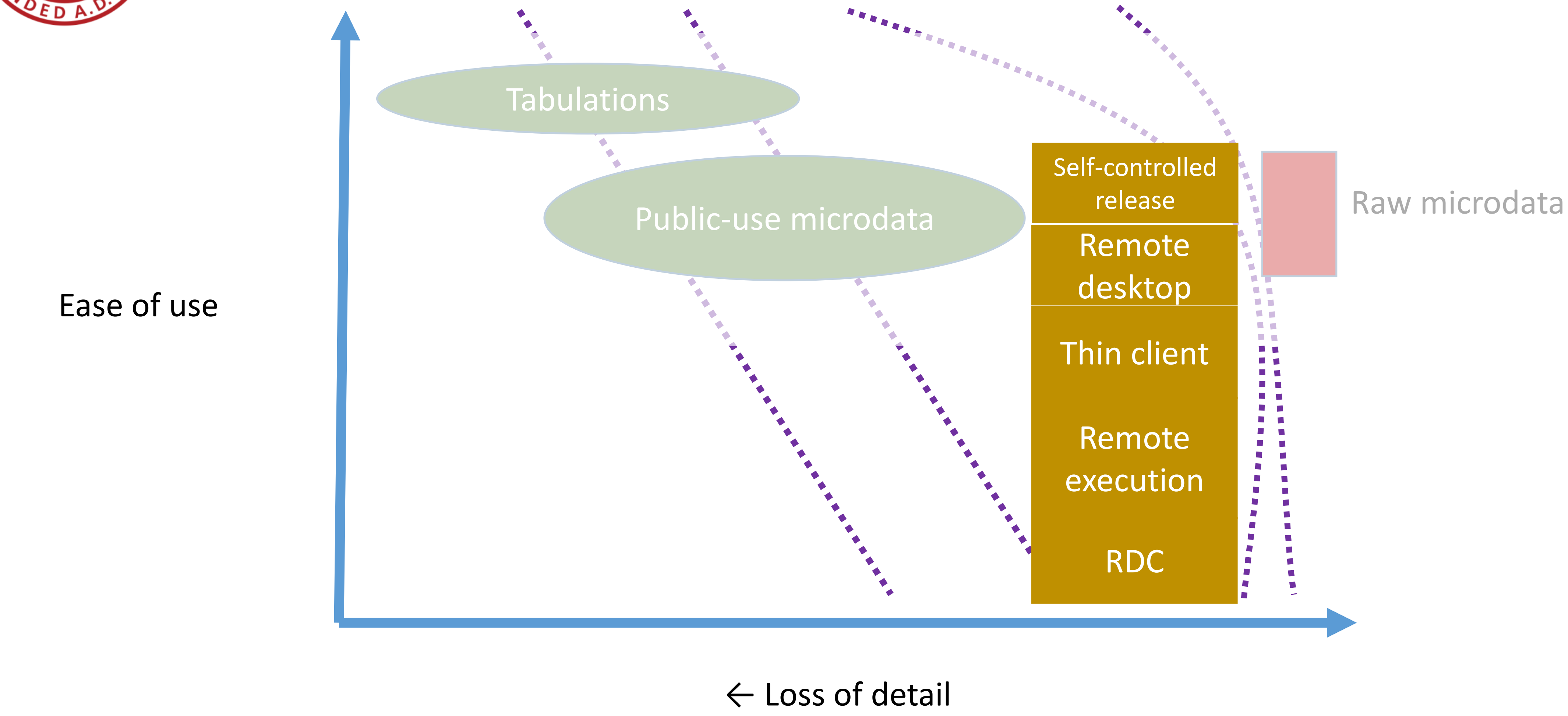
Typically, the **researcher** asks an **authorized agent** of the data provider to **review** the results for **risks of disclosure**, and he will then **send them** to the researcher

What if the “authorized agent” were the researcher?

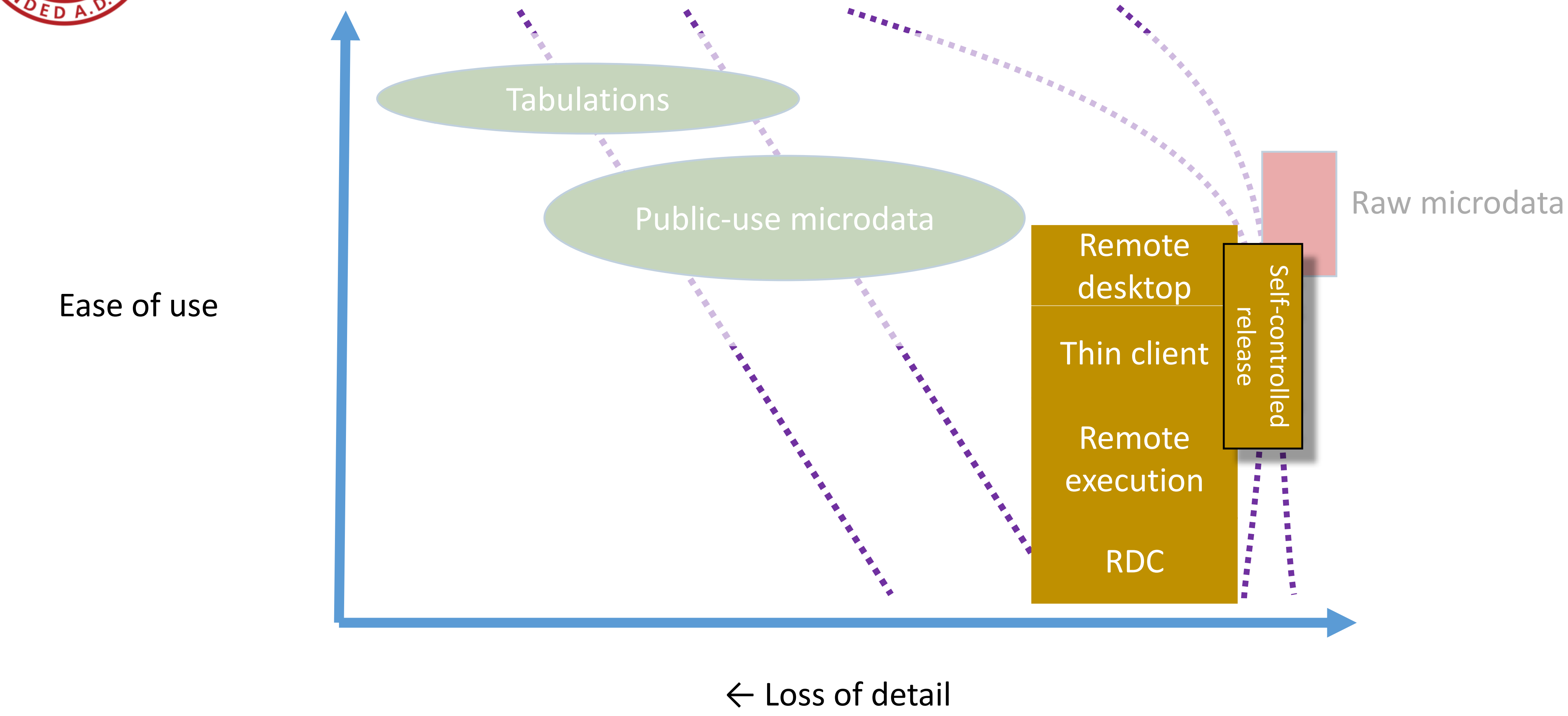
self-controlled release of results

- Researcher controls release of results
  - Prepares results herself
  - According to certain prescribed rules
  - Sends them through a system
  - Automatically receives results typically per email
- Used
  - Most often by contractually-controlled non-enclave data
  - Data in some university- or faculty-controlled enclaves (HRS, Dept. of Ed)
  - Danish researcher access system

access methods: enclaves



access methods: enclaves



Access matrix for confidential data

		Control by data provider of:			
	# access points	**Access computers**	**Access rooms**	**Avail. analysis methods**	**Type disclosure avoidance**
FSRDC researcher	**24 sites** **(~700 users)**	**Full**	**Full (badge access)**	**Some (choice of software)**	**Manual/ variety of rules**
Census staff researcher	**n.d.**	**None (VDI)**	**None (VDI)**	**Some (choice of software)**	**Manual/ self/ variety of rules**
IAB: JoSuA researcher	**414 users**	**None (Web application)**	**None (Web application)**	**Smaller (software, whitelist commands)**	**Manual/ variety of rules**
CASD researcher	**371 sites** **(1471 users)**	**Extra Full (custom-built hardware)**	**Some (university office, EU)**	**Some (choice of software)**	**Manual/ variety of rules** **€300/ pack of 10**
Stat.Denmark (typical EU)	**?**	**None (VDI)** **- Some** **(host institution)**	**None (VDI) -** **Some** **(host institution)**	**Some** **(choice of software)**	**Manual/ self/ variety of rules**

penalties

penalties

- FSRDC and federal employee:
  - federal prison sentence of up to **five (5)** years, a fine of up to **$250,000**, or both.
- France:
  - prison sentence of up to **one (1)** year, a fine of up to **€15,000**, or both.

penalties

- IAB:
  - Loss of data access for up to **two (2)** years for researcher and institution
  - Contractual penalty up to **€60,000** paid by the **institution**
- Denmark:
  - Researcher: Loss of data access **for life**, or up to **three (3)** years for "minor breaches"
  - **Institution**: Loss of access for a positive but limited (undefined) period
  - No financial or penal penalties

Of Note: the FSRDC contract explicitly excludes a responsibility of the university for the actions of its employees, though university remains bound by FWA/IRB.

penalties

- Does **ease of application** matter (penal vs. contractual rules)?
- Is it conducive to more strongly **engage** the researcher’s **employer** (typically but not exclusively a university)?

trust and access

What type of person?

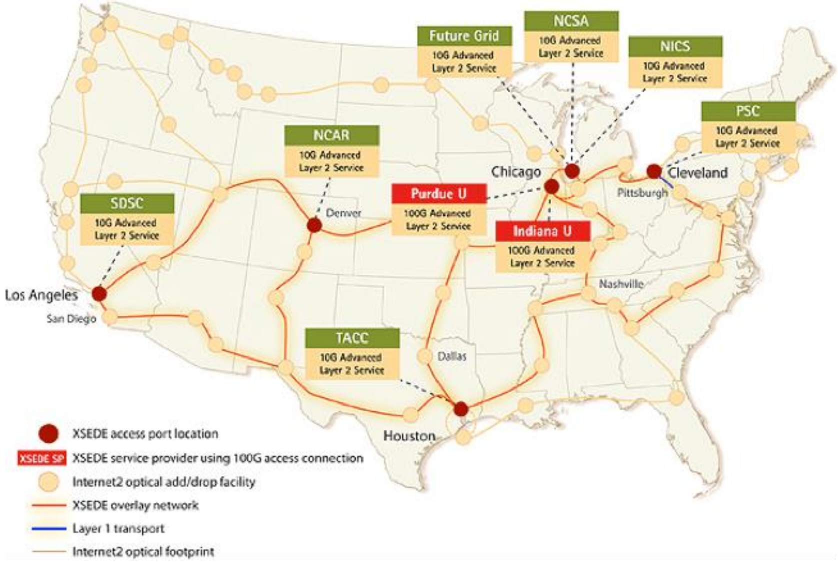
hypothesis: culture matters

- Researchers and agencies create the communities in which rules are applied and enforced
  - Training and “indoctrination”:
    - Training of FSRDC researchers (short, decentralized)
      vs. FedStat employees (≥1 day on-site)
    - 1 full day on-site (in Paris) training for French researchers
  - Common forums:
    - Conferences: Canadian, US (FSRDC, NCHS) yearly RDC conferences
    - Discussion, local groups: users of FSRDC share a common physical space
- More or less tight binding of researchers into a community is important

virtual enclave = centralization

Concerns about centralized compute infrastructure

- Scope
  - FSRDC infrastructure dwarfed by other federal research investments (e.g. XSEDE) that cannot be utilized


Future Grid
10G Advanced Layer 2 Service
NCSA
10G Advanced Layer 2 Service
NICS
10G Advanced Layer 2 Service
PSC
10G Advanced Layer 2 Service
NCAR
10G Advanced Layer 2 Service
SDSC
10G Advanced Layer 2 Service
Purdue U
100G Advanced Layer 2 Service
Indiana U
100G Advanced Layer 2 Service
TACC
10G Advanced Layer 2 Service
Chicago
Cleveland
Pittsburgh
Denver
Nashville
Los Angeles
San Diego
Dallas
Houston
XSEDE access port location
XSEDE SP
XSEDE service provider using 100G access connection
Internet2 optical add/drop facility
XSEDE overlay network
Layer 1 transport
Internet2 optical footprint

Cluster	*Cores*	*Tflops*	*As a multiple of FSRDC*
FSRDC	240	4.36	1x
Wrangler (TACC)	2304	62	14x
Stampede (TACC)	102400	9600	2202x

summary

some concluding thoughts

- How to enable a scalable and secure system
  - Does it require changes in the legal framework?
  - How to build a culture of responsible and secure data access among researchers?
  - What kind of devices or access mechanisms do we want to enable?
  - Who gets to hold the data that researchers actually access?

thank you

lars.vilhuber@cornell.edu

CORNELL UNIVERSITY FOUNDED A.D. 1865

Thanks

- Stefan Bender (formerly IAB and now Bundesbank, Germany)
- Jörg Heining (IAB, Germany)
- Roxanne Silberman (CASD, France)
- Kamel Gadouche (CASD, France)
- Jean Poirier (CIQSS, Canada)

Some References

- Walter Wilcox (1914) cited in Anderson, Margo J., and Seltzer, William. “Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues’.” Journal of Privacy and Confidentiality 1.1 (2009): 7-52, 55-58.
- Kohlmann, Annette (2005): “The Research Data Centre of the Federal Employment Service in the Institute for Employment Research.” In: Schmollers Jahrbuch 125, 437-447
- Allmendinger, Jutta and Kohlmann, Annette (2005) “Datenverfügbarkeit und Datenzugang am Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung”. In: Allgemeines Statistisches Archiv 89, S. 159-182
- Heining, Jörg (2010): “The Research Data Centre of the German Federal Employment Agency: data supply and demand between 2004 and 2009.” In: Zeitschrift für ArbeitsmarktForschung, Jg. 42, H. 4, S. 337-350. http://www.iab.de/389/section.aspx/Publikation/k100128n09
- Kargus, Andrea; Müller, Anne (2014): “Auch in Nürnberg möglich: Von der zweiten Liga in die Champions League - ein Gespräch mit Stefan Bender.” In: IAB-Forum, Nr. 2, S. 38-45. http://www.iab.de/188/section.aspx/Publikation/k141201301
- Kraus, Rebecca S. (2011): “Statistical Déjà Vu: The National Data Center Proposal of 1965 and Its Descendants.” Presentation at JSM 2011. https://www.census.gov/history/pdf/kraus-natdatacenter.pdf