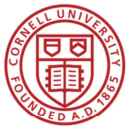


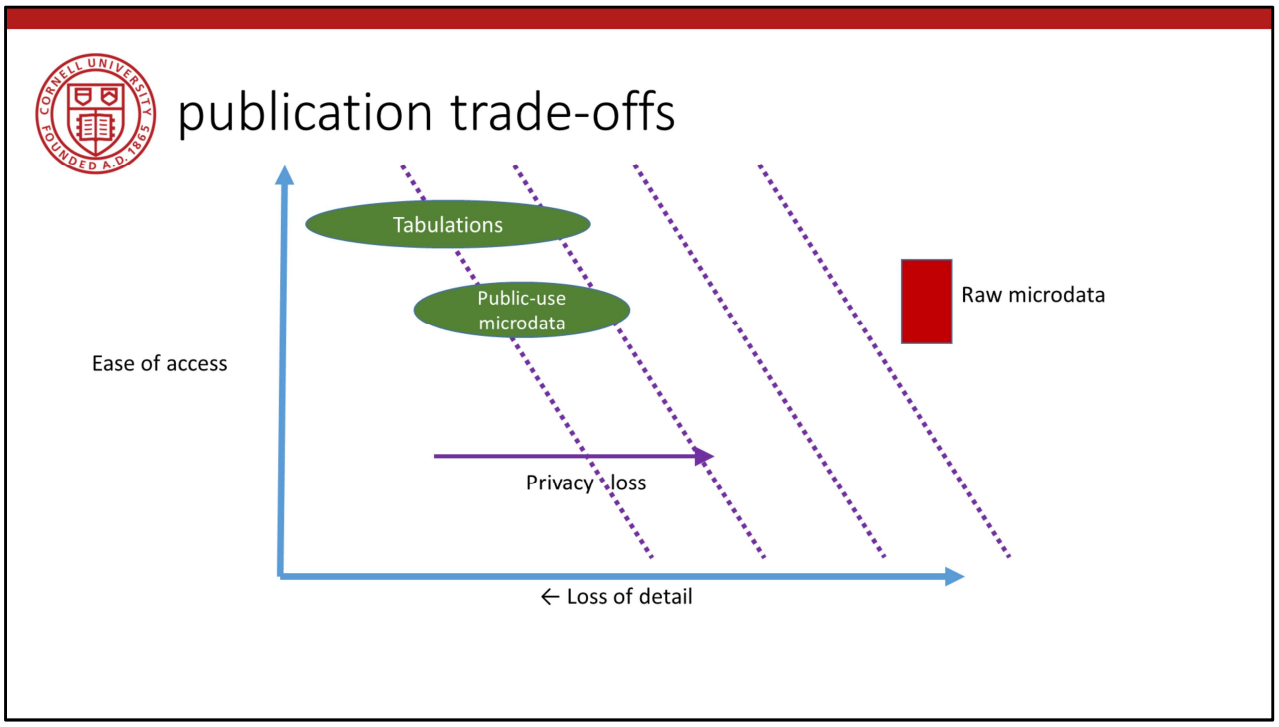
Confidentiality protection and physical safeguards

Lars Vilhuber

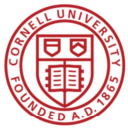
Cornell University

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

What I will be describing here are the physical safeguards, as well as the context – historical, legal, and cultural – of secure access to confidential microdata, as observed in various systems around the world. I will draw on my personal experience with 4 such systems, and my participation in scientific committees and conferences on the topic for a few more.



What will frame the analysis is the tension between undesirable (possibly illegal) publication of raw microdata, and the accessibility to a broader public of information on the contents of that raw microdata. The traditional solution was the publication of public-use microdata, but for a variety of reasons, that is no longer seen as the only possible solution, or not the ideal solution.



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?

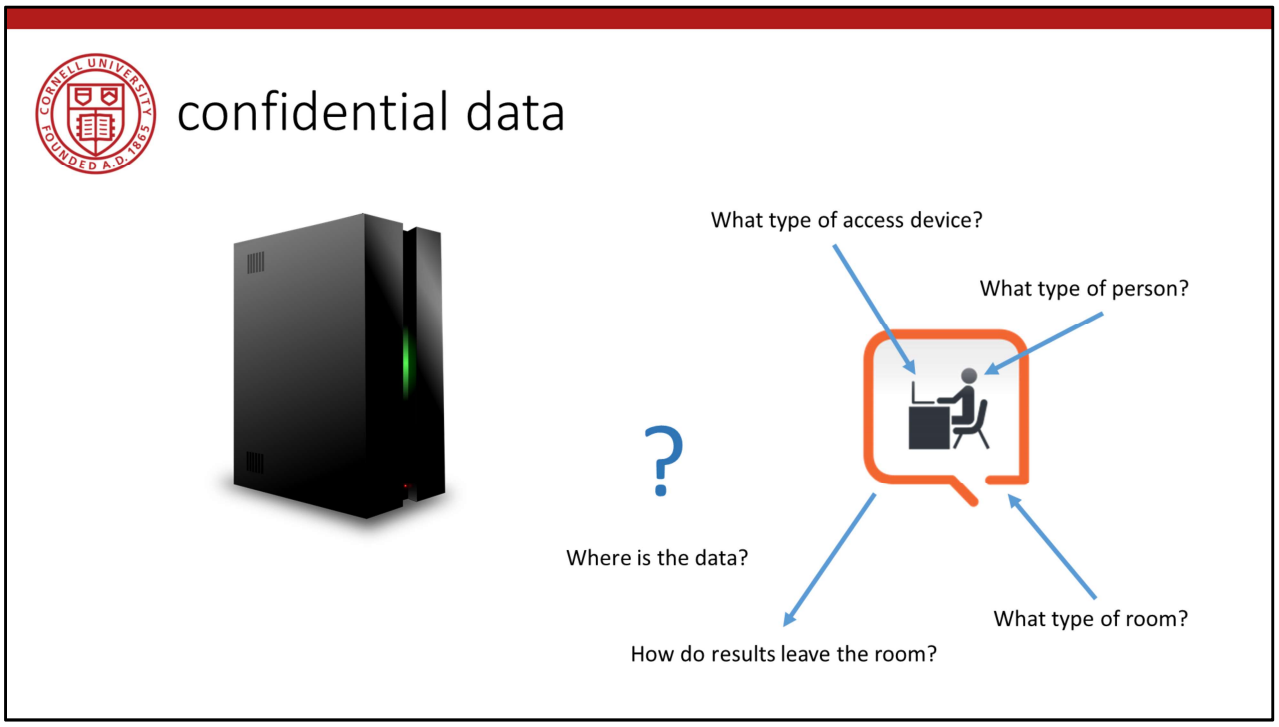
The goal thus of the systems I will describe is….

public use data

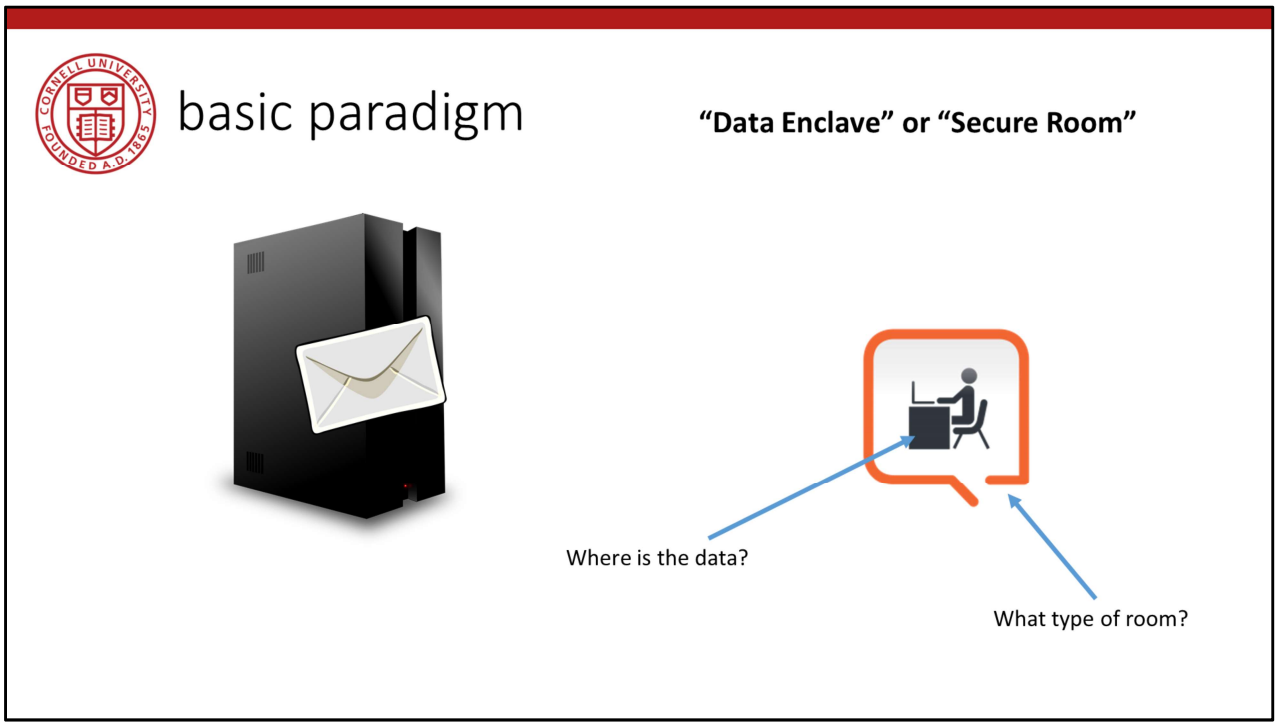
Data provider/
custodian

Data user/
researcher

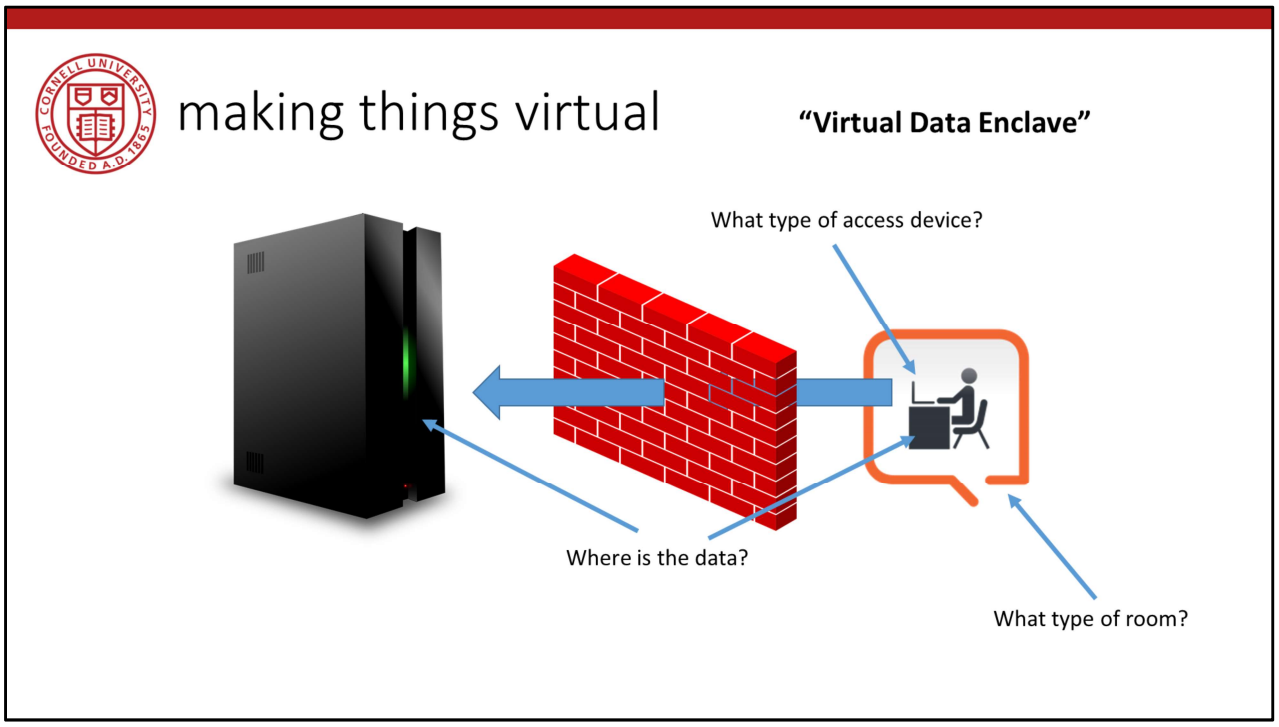
The reference to public-use microdata is useful because it defines the researchers' expectations about how to access data in the United States – but not in other countries, which have not had the extensive access to public-use microdata that US researchers have enjoyed over the decades.



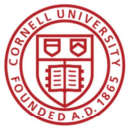
When facing the need to provide access to confidential data, the approach has been to put the researcher into a closed room, and there provide her with access to the data. Questions then arising, and which I will talk about, involve defining the type of room, the type of computer used to access the data, the location of the data itself, who has the rights to access the data, and how results leave the room.



For instance, if the data provider ships the data directly to the secure room, we call it a "data enclave" or a "secure room"



If on the other hand, the data provider retains custody of the data, and provides a remote access to the data through a firewall, we have virtualized the data enclave: the data enclave no longer physically houses the data. We may still walk into a well-defined space – a RDC – but as we will see, that is not always the case.



virtual data enclaves

Synonyms:

VDI

(virtual desktop infrastructure)

Thin clients

Remote desktop

You may have seen synonyms for Virtual Data Enclave, such as these.

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)

Physical data enclaves were widely used in the 1990s, when direct provision of the data to end users was not feasible, including at the pioneering Census and Canadian RDCs

Examples in 2017

Physical data enclaves

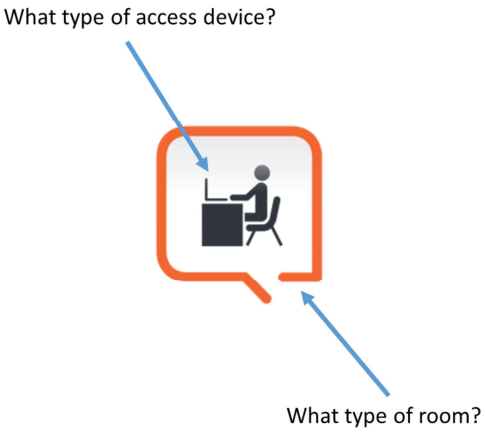
- BLS HQ ?
- BJS data access
- Department of Education data
- ~~Census Bureau RDCs~~
- Canadian RDCs
- ~~HRS restricted-access data~~
- and many more

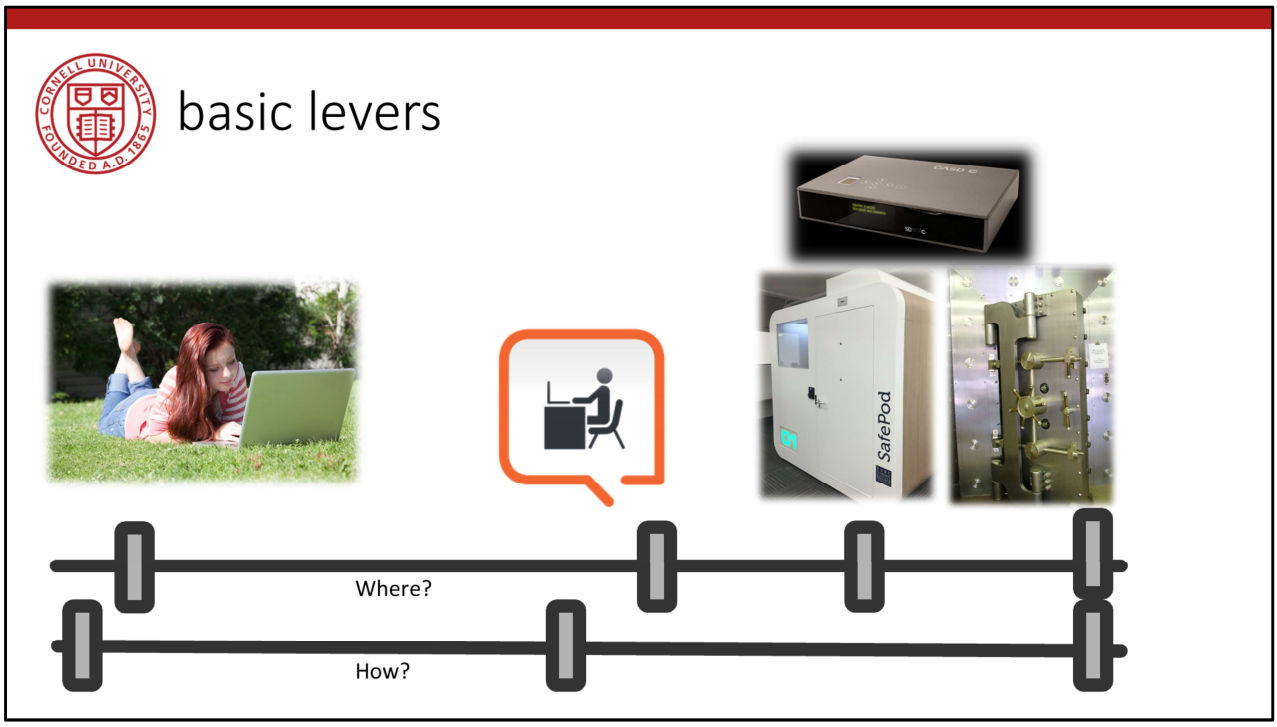
Virtual data enclaves
(data remains in secure data center)

- 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

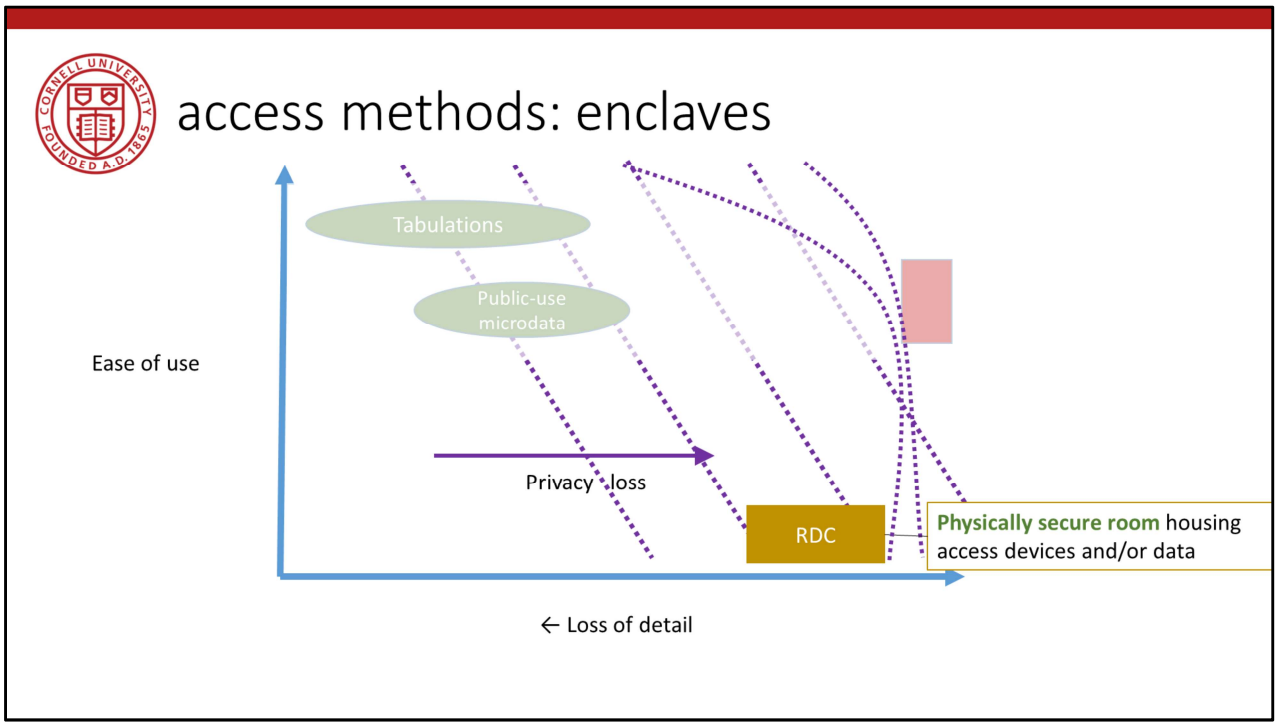
In 2017, some of these have switched to a virtual data enclave structure, most prominent among them the Census Bureau, which switched to a virtual data enclave structure in the early 2000s. Abroad, broader research access to data only began to emerge in the mid 2000s, and in most cases, arose as Virtual Data Enclaves from the start. Of note, Canadian RDC system is still fundamentally a physical data enclave system, but is considering moving part of its infrastructure to Virtual Data Enclave architecture

basic levers

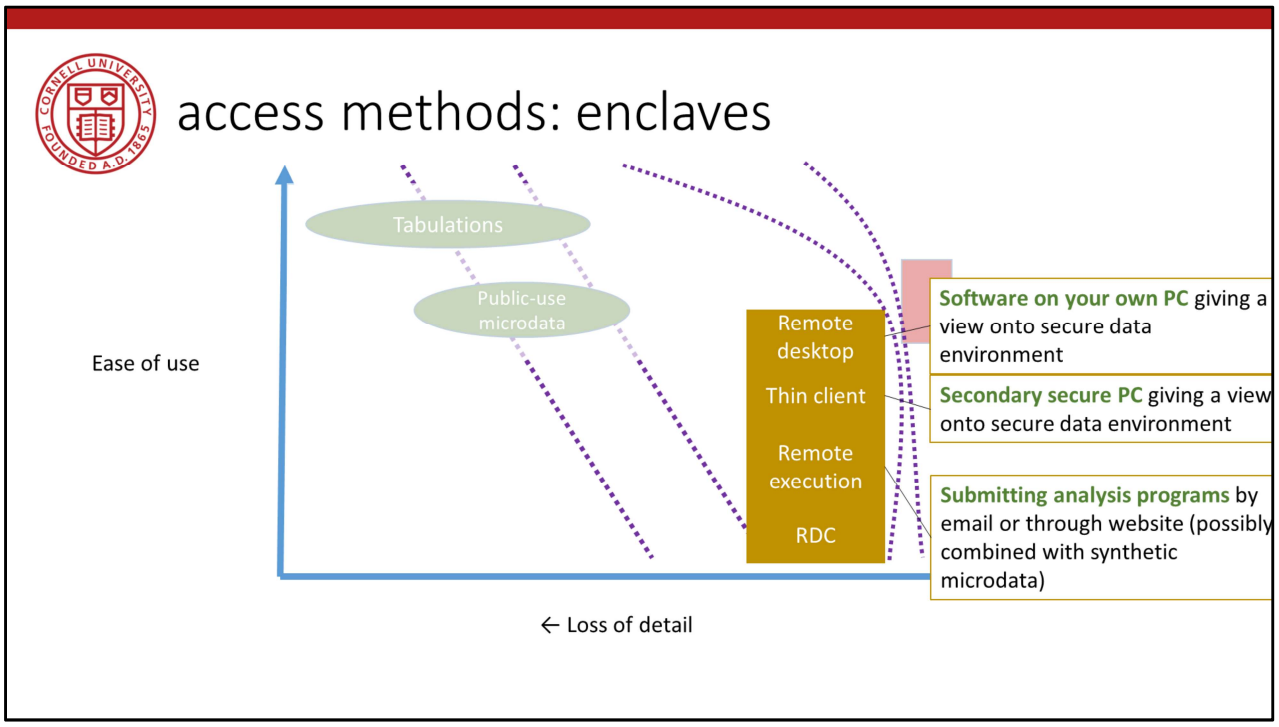




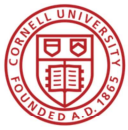
The “where” and “how” of data access are the most obvious levers, ranging from complete freedom to Fort Knox-style security (the BJS HQ data center has been described to me like that). Relaxing the constraints on the where lead to RDCs or even location of thin clients in university offices.



Mapping it back to the tradeoff between ease of use and loss of privacy, these systems tend to have similar loss of privacy – the disclosure avoidance mechanisms are mostly of the same type – but vary in the degree of ease of use.

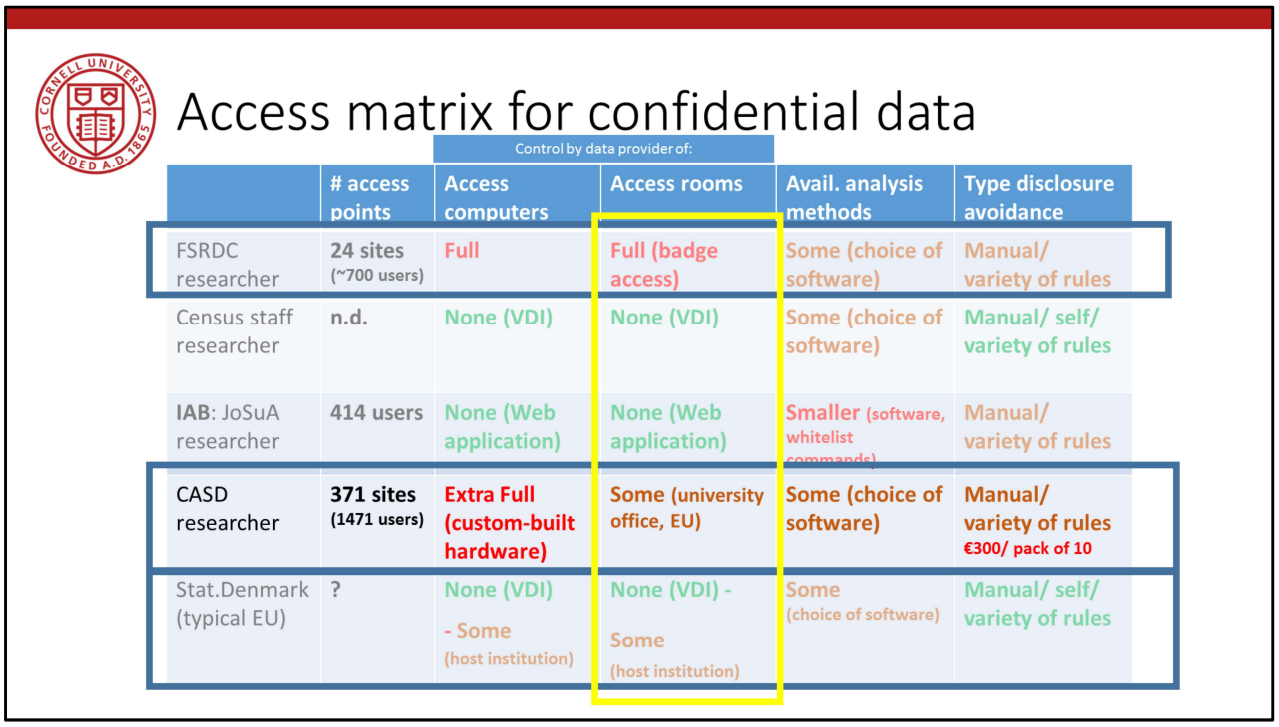


All but the RDC system are necessarily “virtual” data access modes, but not all of them require the researcher to be in an enclave, i.e., a secure room. Rather, it is the data that is secured.

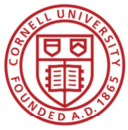


What type of room?

So let's consider what type of constraints are put on the location of access

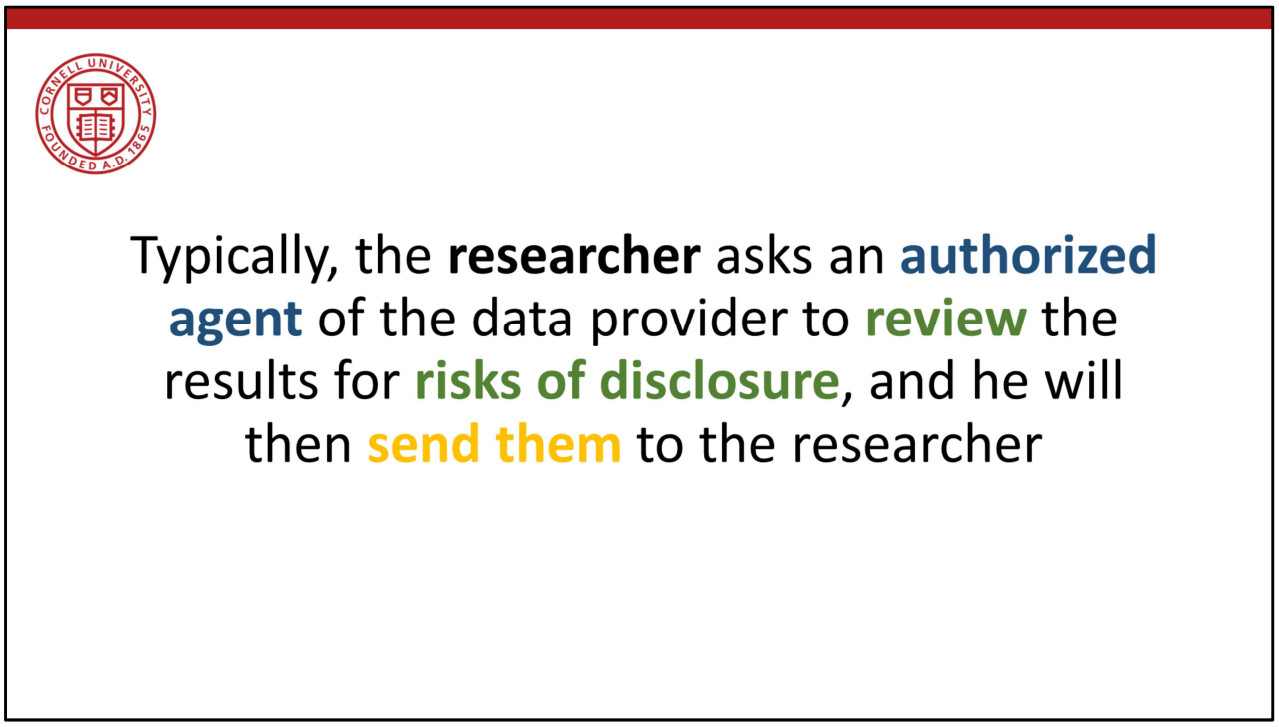


This table has a select set of systems that highlight some key elements of the variation in access modalities. You have heard from representatives describing all of these systems in this commission. I would like to point out the middle-of-the-road access restriction of the French system, which limits access by a custom thin client to university offices (and needs to determine what are legitimate universities and research institutions). In contrast, Statistics Denmark does not impose strong limits on the VDI access it grants – access by any PC is allowed, as long as the access occurs from the university offices or the home of the researcher. In contrast, the FSRDC is a classic "virtual data enclave", sending researchers to secure rooms under full access control of the Census Bureau, using dedicated thin clients.

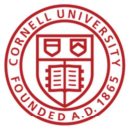


How do results leave the room?

The next question is how results are made available for later publication. As it turns out, there is surprisingly little variation in most of these systems in terms of the disclosure avoidance methods.



Typically, … But there is one alternate path for the data to exit by.



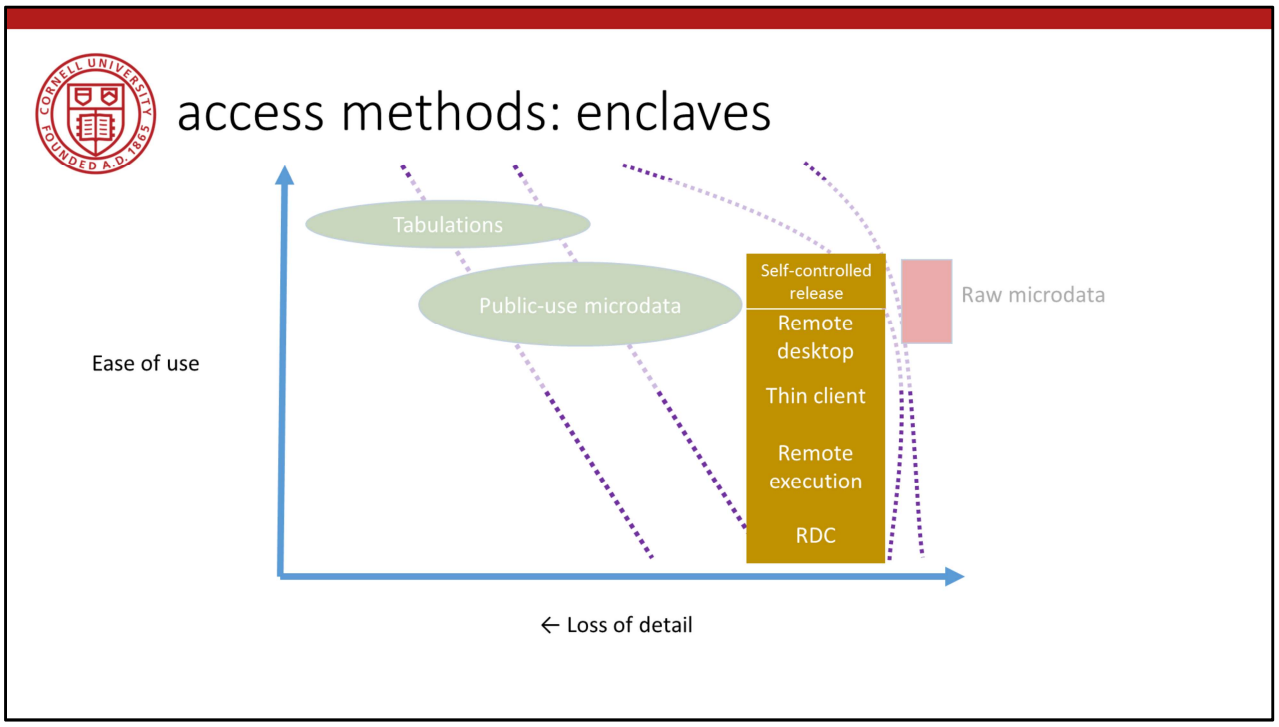
What if the “authorized agent” were the researcher?

So what if …

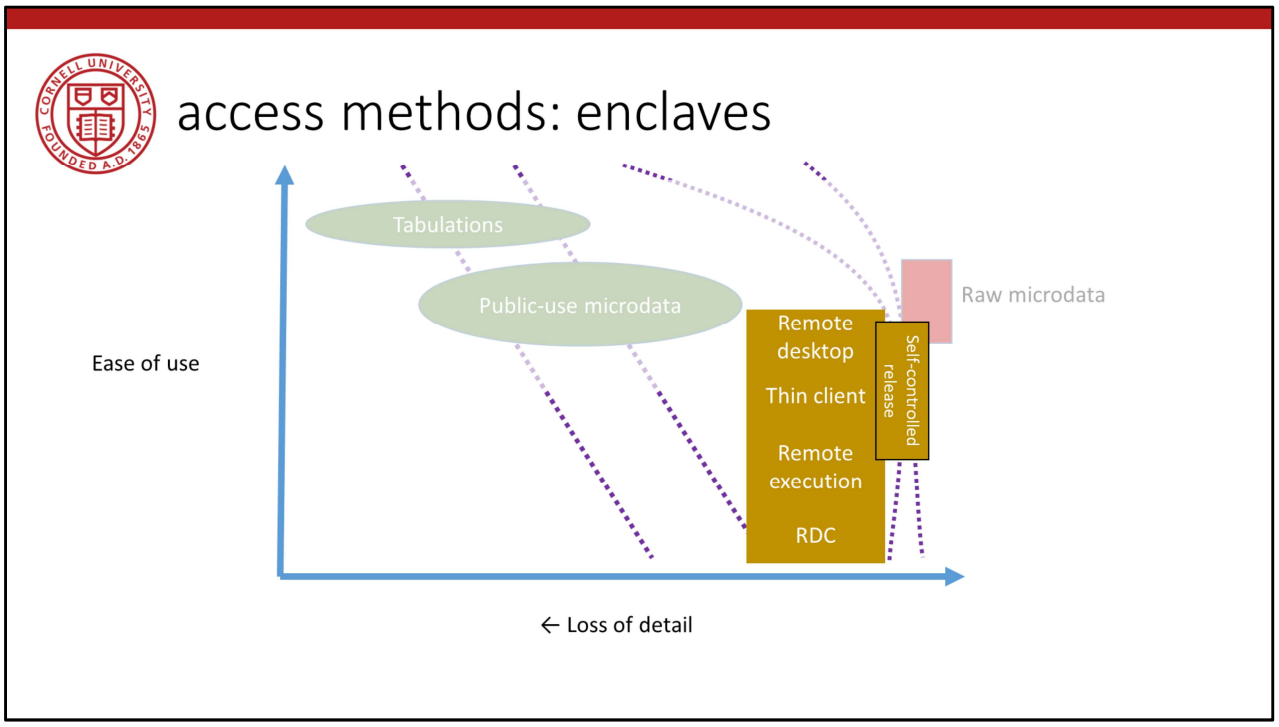
self-controlled release of results

- 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

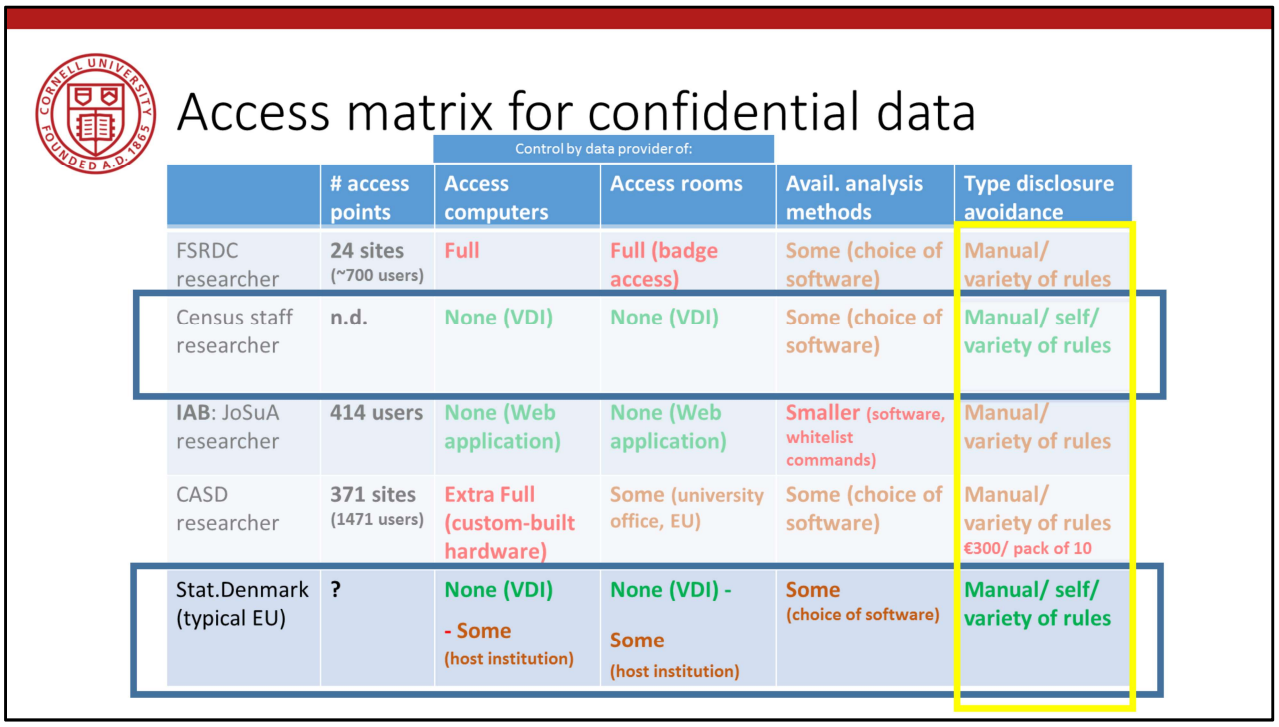
In this case, the researcher has control over the speed and timing of releases, while the agency still controls the rules. This is most often used when data are sent directly to the user or a university data enclave, where final control by the data provider is difficult or impossible. But it is also – radically? – used by the Danish system.



This could be described as the ultimate “ease-of-access” nearly on par with the classic public-use microdata mechanism…

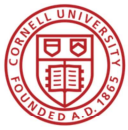


In theory, this could be combined with nearly all of these systems. So do we think that the Danish system is a radical system?



Let's turn back to my little table: The Danish system combines relaxed controls as to the type of computer and location of the access, with very liberal disclosure avoidance mechanism. But as it turns out, such a system already exists in the US case: it is the system that Census staff researchers are subjected too.

That should lead us to consider why we trust certain people with a more open approach, and not others. Two topics come to mind: the penalties applied to failure to comply with rules, and the culture of trust.



penalties

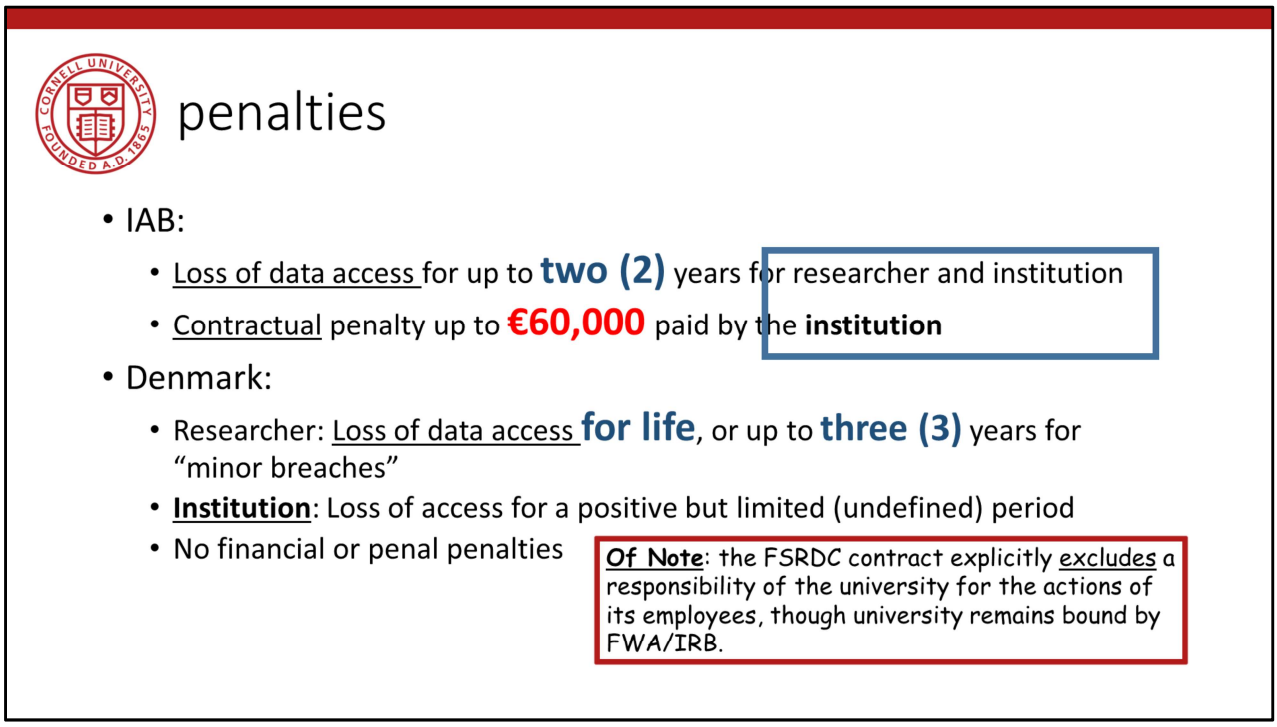
Let’s turn first to some of the penalties.

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.

In the case of the US and France, breaches are punishable under national law with prison and monetary fines.

Note that this cannot explain the difference within the US between staff researchers and FSRDC researchers – they face the same penalties.



IAB and Denmark have no prison sentences, and relatively low or no financial penalties. Their lever is access to the data – do something stupid, and you and the institution lose access to the data for some amount of time, potentially forever.

They do however strongly engage the institution, not just the researcher. Germany (but not Denmark) also rely on the contractual nature – and thus easier enforceability without reliance on specific laws - to allow for transnational access.

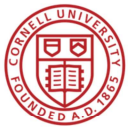
Notice that in the case of the FSRDC, the contract with the university explicitly excludes institutional liability for a researcher's potential misdoings.

Finally, notice that surprisingly the more liberal systems also have more lenient penalties. However, one can imagine that precisely because they are more lenient, they are more easy to impose.

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)?

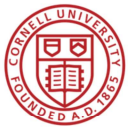
Thus we can ask…



trust

or “what type of person?”

But this did not, for instance, explain the difference in access modes for the US system, between university and staff researchers. Why do we trust certain people and not others?

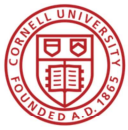


hypothesis: culture matters

- 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

Some of the more liberal systems (Paris, FedStat employees) provide for strong on-site, personal "indoctrination" of the principles of confidentiality – typically a trip to the headquarters and a full-day training session. I will also note that the German system requires the presence of a representative of the data custodian to be personally present each time a researcher logs into the system (has to type a system-level password). While providing extra security, it also is a reminder of the importance of security at each step.

Actually, there is already a lot of community building going on in the US (and Canadian) RDC systems. Regular conferences by and for data users, and the sharing of a common physical space foster a sense of community.



virtual enclave = centralization

One final concern. A virtual data enclave (as currently implemented in the US) implies centralization.

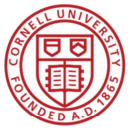
Concerns about centralized compute infrastructure

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

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

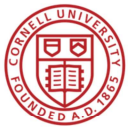
If I compare the computing resources available to the FSRDC with that of other federally funded computational infrastructure, it is abysmal. Just to illustrate: Wrangler is the smallest listed system on XSEDE as of Feb 2017, and has 14x the computational power of the entire FSRDC. It is only one of many XSEDE systems. Stampede is the largest – with 2000x the computational capacity of the FSRDC. We should thus consider the possibility of….

. Information obtained from https://www.xsede.org/web/guest/resources/overview on Feb 18, 2017. FSRDC estimated based on 10 dual-CPU, 12core systems.



many virtual enclaves

= decentralization

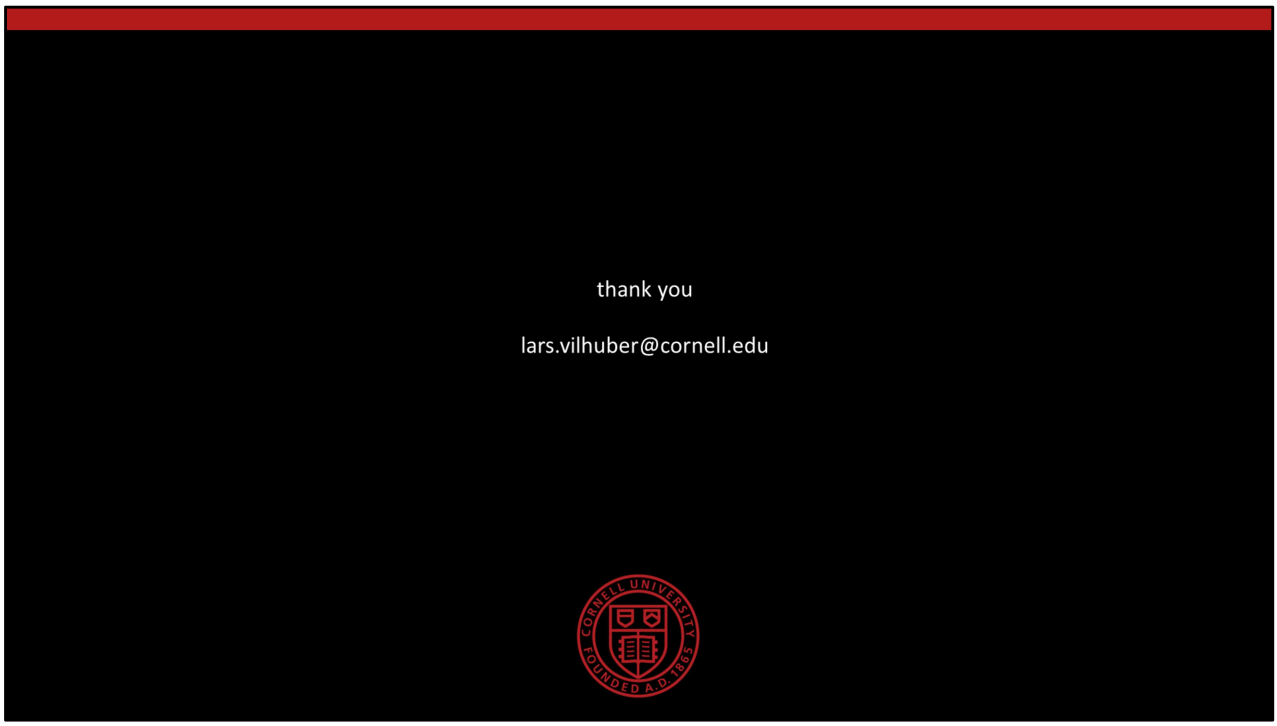


summary

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?

What do I think are the conditions to be able to augment the US system to be scalable and secure? Does it require changes in the legal framework, or just changes in the contracts used to provide access? How to we create a culture of responsible use that is equivalent to the one our FedStat employees already demonstrate every day? Can we rethink what kinds of access devices and physical space researchers should use? And finally, should we rethink who gets to hold the data that researchers actually access – keyword decentralization.


thank you
lars.vilhuber@cornell.edu
CORNELL UNIVERSITY
FOUNDED A.D. 1865

Thank you.

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