Confidentiality protection and physical safeguards

Lars Vilhuber

Cornell University

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

confidentiality and access

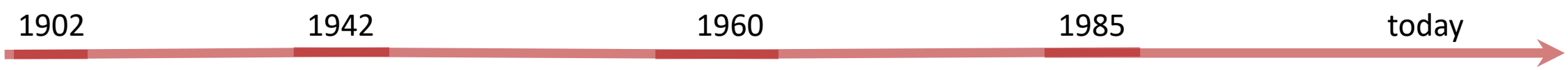
confidentiality of statistical agency data

- "… when the secretary of [Commerce and Labor] directed that the census schedules of manufacturing establishments should be open to the inspection of officials belonging to another bureau within the same department […] and the director [of the Census Bureau] refused […] because of the pledge of secrecy…"

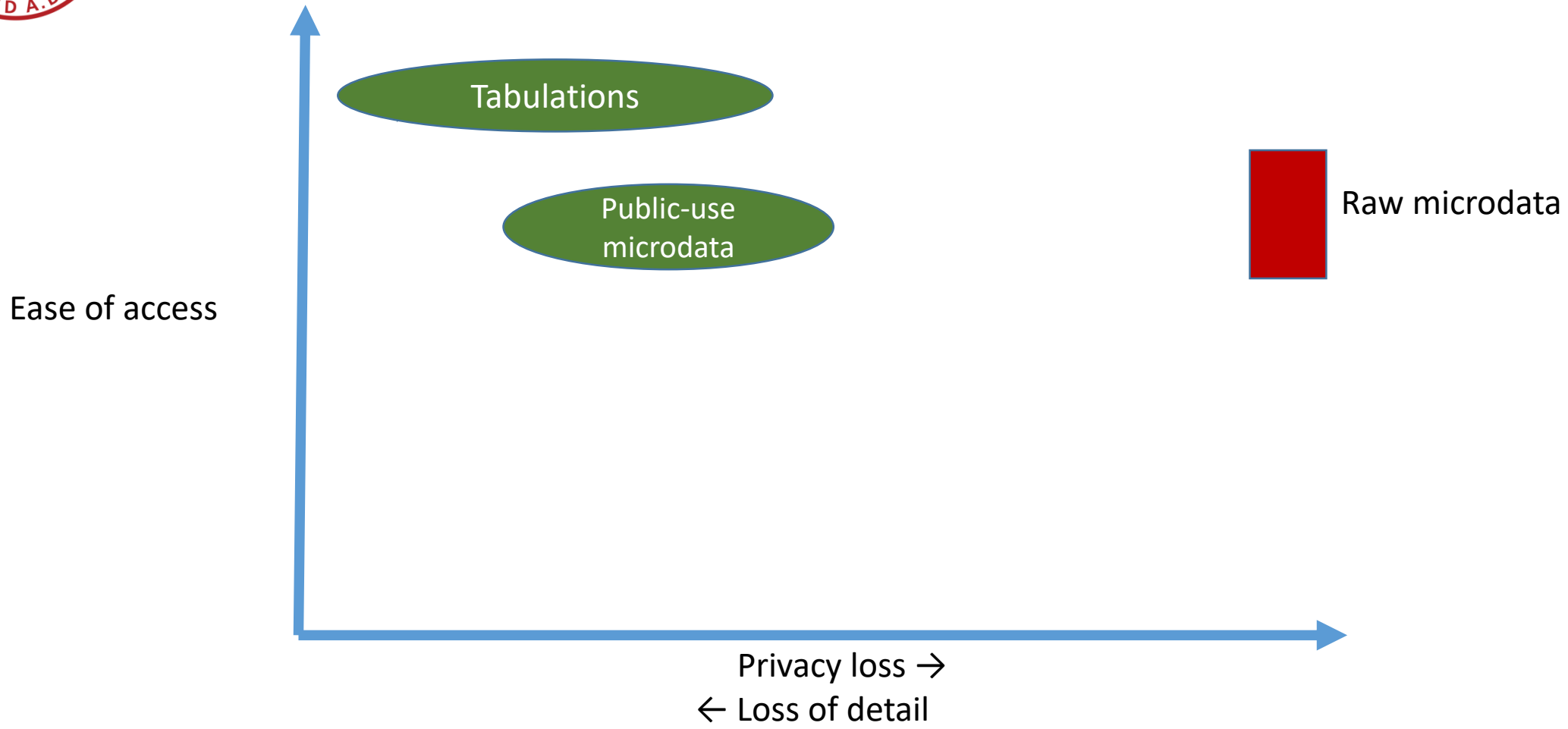
(Walter Wilcox, 1914)

rich new analysis and publications

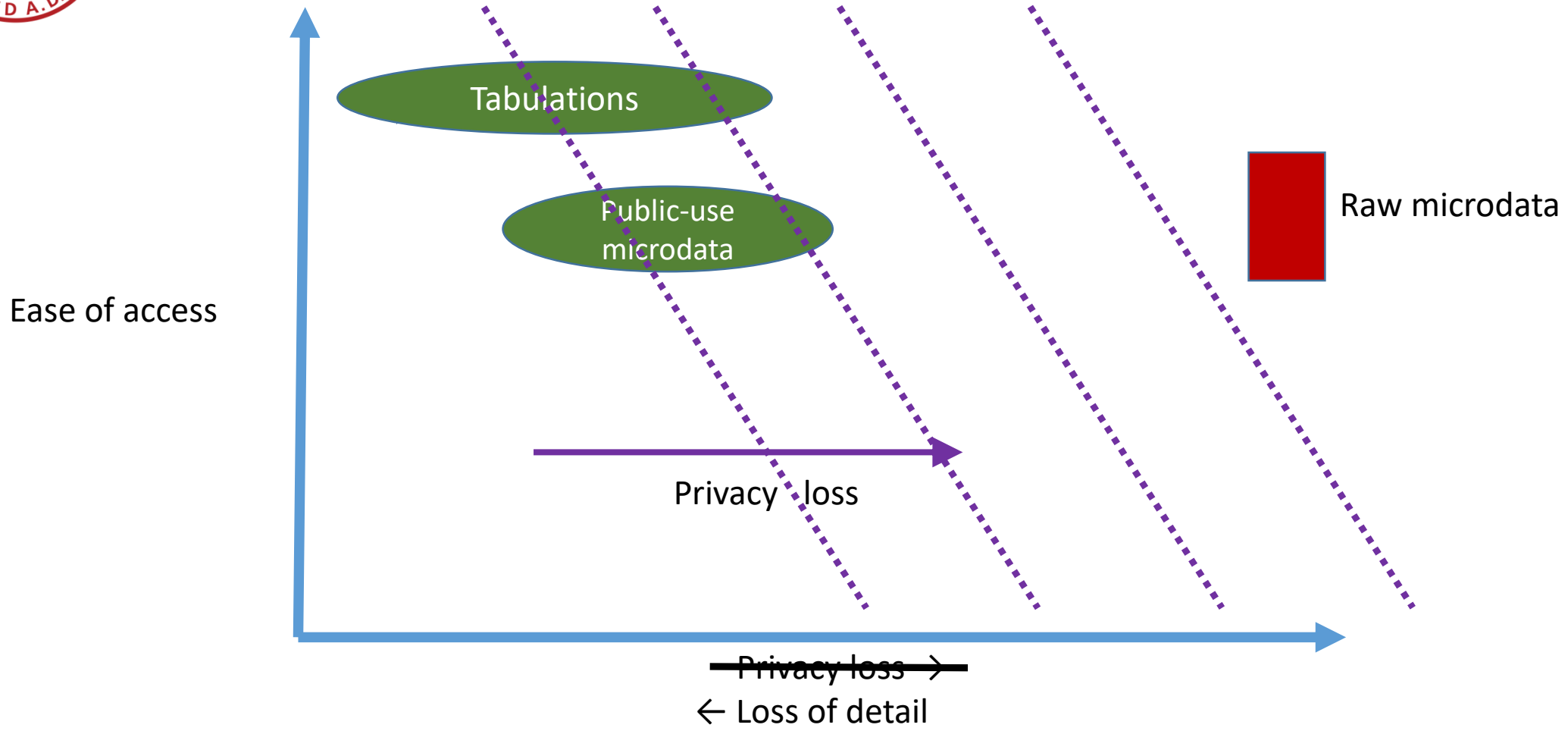
held back by concerns of citizens and businesses about privacy



access methods

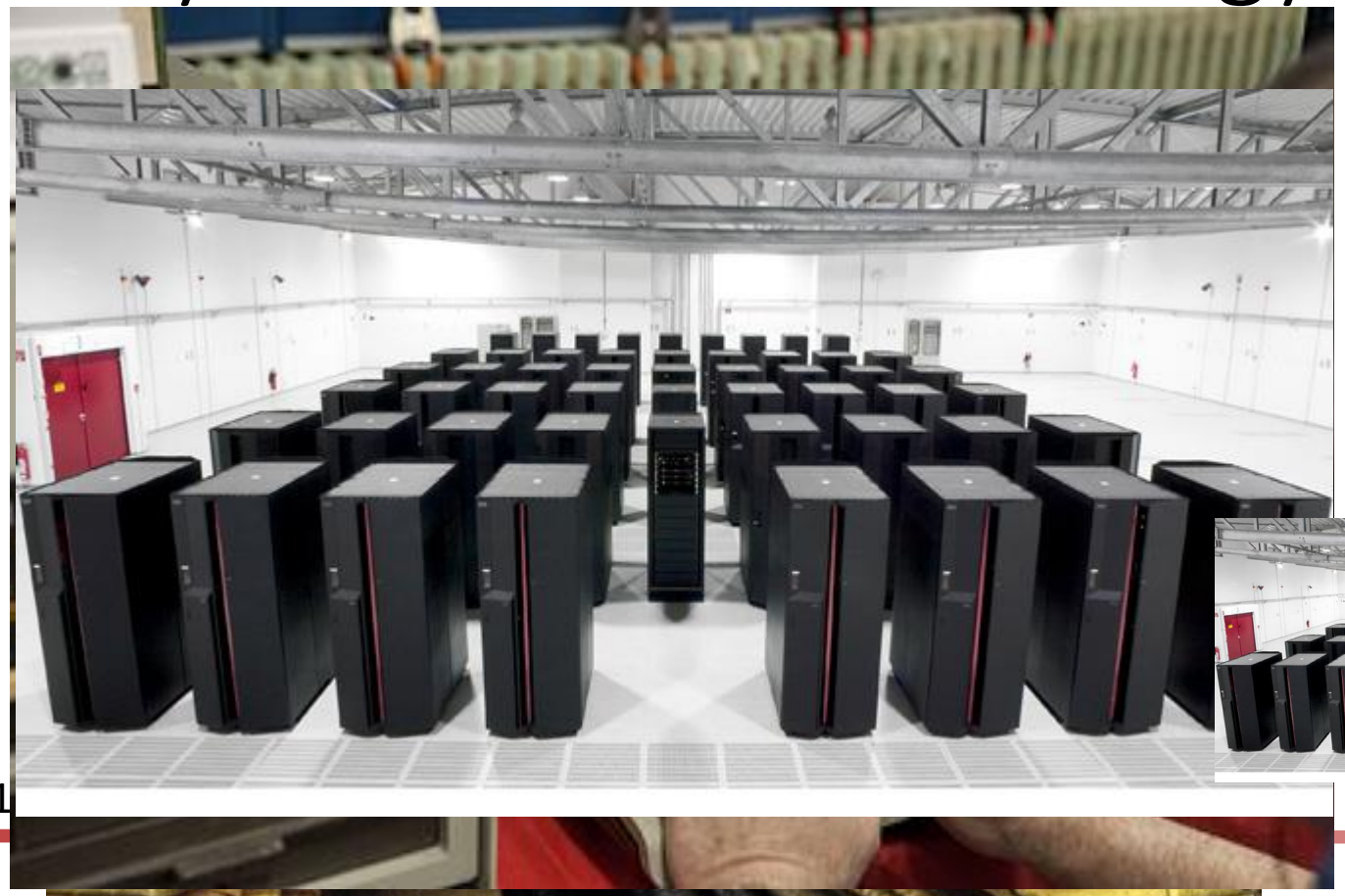


access methods



driven by advances in technology

1902



today

researchers knocking on the door

my talk today

focus

I will focus on access mechanisms for **researchers**

Source: (Fox News/REUTERS/Kacper Pempel/Files/https://goo.gl/ZHMkog)

I will exclude

- Newer mechanisms to create <u>tabular</u> data (synthetic data, differentially-private data)

I will include

- Use of analytically-valid synthetic data as an access mechanism

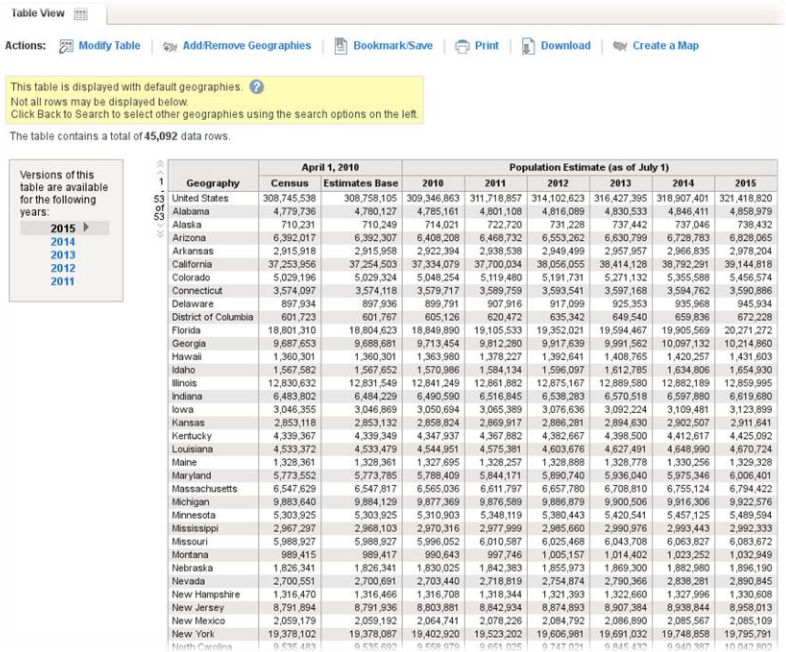


Table View

Actions: Modify Table | Add/Remove Geographies | Bookmark/Save | Print | Download | Create a Map

This table is displayed with default geographies.
Not all rows may be displayed below.
Click Back to Search to select other geographies using the search options on the left.

The table contains a total of **45,092** data rows.

Versions of this table are available for the following years:
2015
2014
2013
2012
2011

Geography	April 1, 2010 Census	April 1, 2010 Estimates Base	Population Estimate (as of July 1) 2010	2011	2012	2013	2014	2015
United States	308,745,538	308,758,105	309,346,863	311,718,857	314,102,623	316,427,395	318,907,401	321,418,820
Alabama	4,779,736	4,780,127	4,785,161	4,801,108	4,816,089	4,830,533	4,846,411	4,858,979
Alaska	710,231	710,249	714,021	722,720	731,228	737,442	737,046	738,432
Arizona	6,392,017	6,392,307	6,408,208	6,468,732	6,553,262	6,630,799	6,728,783	6,828,065
Arkansas	2,915,918	2,915,958	2,922,394	2,938,538	2,949,499	2,957,957	2,966,835	2,978,204
California	37,253,956	37,254,503	37,334,079	37,700,034	38,056,055	38,414,128	38,792,291	39,144,818
Colorado	5,029,196	5,029,324	5,048,254	5,119,480	5,191,731	5,271,132	5,355,588	5,456,574
Connecticut	3,574,097	3,574,118	3,579,717	3,589,759	3,593,541	3,597,168	3,594,762	3,590,886
Delaware	897,934	897,936	899,791	907,916	917,099	925,353	935,968	945,934
District of Columbia	601,723	601,767	605,126	620,472	635,342	649,540	659,836	672,228
Florida	18,801,310	18,804,623	18,849,890	19,105,533	19,352,021	19,594,467	19,905,569	20,271,272
Georgia	9,687,653	9,688,681	9,713,454	9,812,280	9,917,639	9,991,562	10,097,132	10,214,860
Hawaii	1,360,301	1,360,301	1,363,980	1,378,227	1,392,641	1,408,765	1,420,257	1,431,603
Idaho	1,567,582	1,567,652	1,570,986	1,584,134	1,596,097	1,612,785	1,634,806	1,654,930
Illinois	12,830,632	12,831,549	12,841,249	12,861,882	12,875,167	12,889,580	12,882,189	12,859,995
Indiana	6,483,802	6,484,229	6,490,590	6,516,845	6,538,283	6,570,518	6,597,880	6,619,680
Iowa	3,046,355	3,046,869	3,050,694	3,065,389	3,076,636	3,092,224	3,109,481	3,123,899
Kansas	2,853,118	2,853,132	2,858,824	2,869,917	2,886,281	2,894,630	2,902,507	2,911,641
Kentucky	4,339,367	4,339,349	4,347,937	4,367,882	4,382,667	4,398,500	4,412,617	4,425,092
Louisiana	4,533,372	4,533,479	4,544,951	4,575,381	4,603,676	4,627,491	4,648,990	4,670,724
Maine	1,328,361	1,328,361	1,327,695	1,328,257	1,328,888	1,328,778	1,330,256	1,329,328
Maryland	5,773,552	5,773,785	5,788,409	5,844,171	5,890,740	5,936,040	5,975,346	6,006,401
Massachusetts	6,547,629	6,547,817	6,565,036	6,611,797	6,657,780	6,708,810	6,755,124	6,794,422
Michigan	9,883,640	9,884,129	9,877,369	9,876,589	9,886,879	9,900,506	9,916,306	9,922,576
Minnesota	5,303,925	5,303,925	5,310,903	5,348,119	5,380,443	5,420,541	5,457,125	5,489,594
Mississippi	2,967,297	2,968,103	2,970,316	2,977,999	2,985,660	2,990,976	2,993,443	2,992,333
Missouri	5,988,927	5,988,927	5,996,052	6,010,587	6,025,468	6,043,708	6,063,827	6,083,672
Montana	989,415	989,417	990,643	997,746	1,005,157	1,014,402	1,023,252	1,032,949
Nebraska	1,826,341	1,826,341	1,830,025	1,842,383	1,855,973	1,869,300	1,882,980	1,896,190
Nevada	2,700,551	2,700,691	2,703,440	2,718,819	2,754,874	2,790,366	2,838,281	2,890,845
New Hampshire	1,316,470	1,316,466	1,316,708	1,318,344	1,321,393	1,322,660	1,327,996	1,330,608
New Jersey	8,791,894	8,791,936	8,803,881	8,842,934	8,874,893	8,907,384	8,938,844	8,958,013
New Mexico	2,059,179	2,059,192	2,064,741	2,078,226	2,084,792	2,086,890	2,085,567	2,085,109
New York	19,378,102	19,378,087	19,402,920	19,523,202	19,606,981	19,691,032	19,748,858	19,795,791
North Carolina	9,535,483	9,535,692	9,558,979	9,651,925	9,747,021	9,845,432	9,940,387	10,042,802

context of my talk today

- Focus on researcher access to authorized data collections
  - Not building new data collections, or enacting new laws
- Focus on the mechanisms for providing access
  - Mostly ***physical***
  - Access to ***microdata***

In light of the “Five Safes Framework”

The "Five Safes framework" (Desai, Ritchie, and Welpton, 2016) provides structure to many aspects of providing secure access to data:

- **Safe projects** - evaluating data analysis projects for appropriateness
- **Safe people** - evaluating the credentials of researchers
- **Safe settings** - how can the data be accessed?
- **Safe data** - how sensitive is the data/ can the data be made?
- **Safe outputs** - how sensitive are analysis results

In light of the “Five Safes Framework”

The "Five Safes framework" (Desai, Ritchie, and Welpton, 2016) provides structure to many aspects of providing secure access to data:

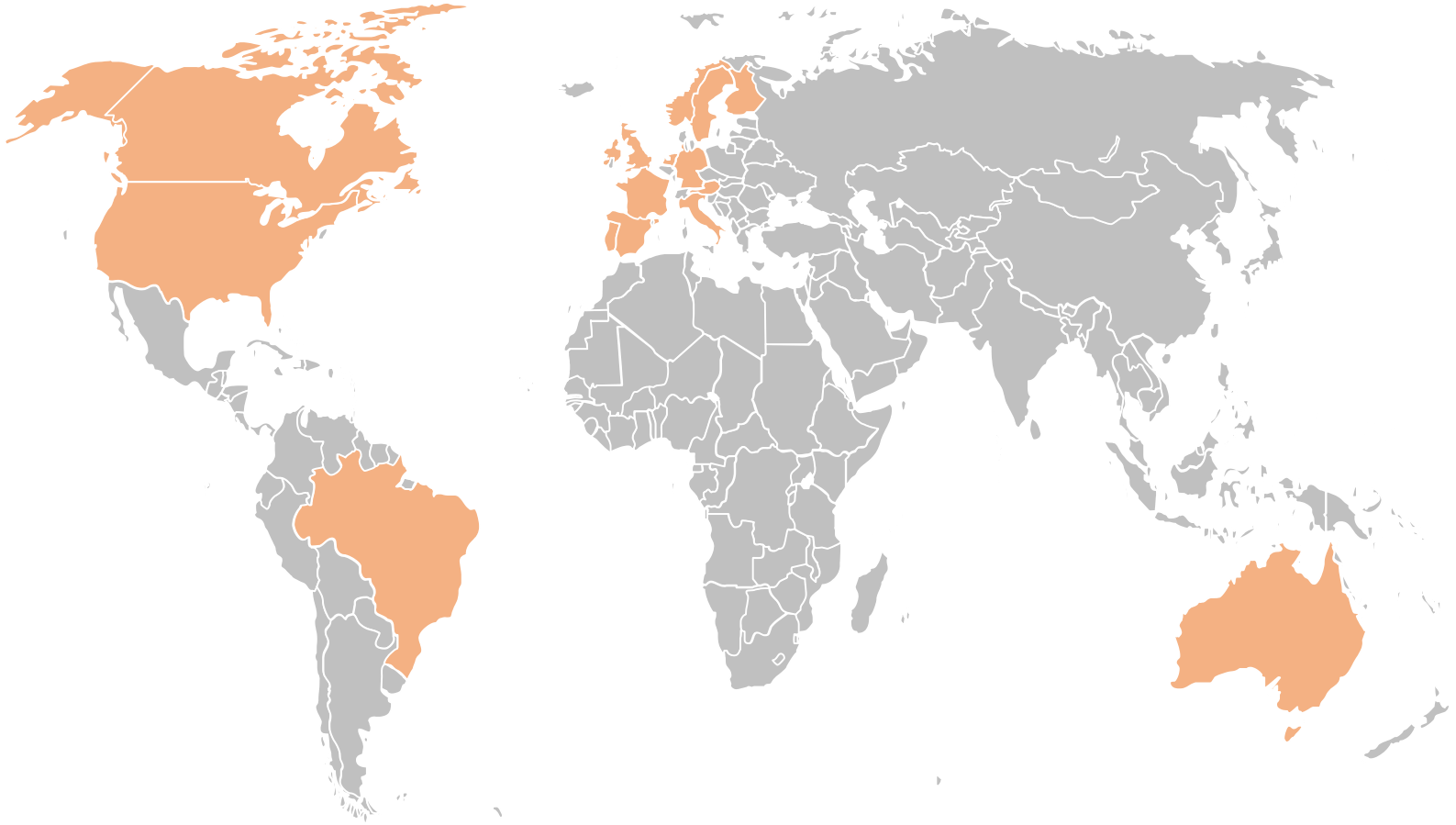
- **Safe projects** - evaluating data analysis projects for appropriateness
- **Safe people** - evaluating the credentials of researchers
- **Safe settings** - how can the data be accessed?
- **Safe data** - how sensitive is the data / can the data be made?
- **Safe outputs** - how sensitive are analysis results

May impact choice of setting

Primary focus

May impact choice of setting

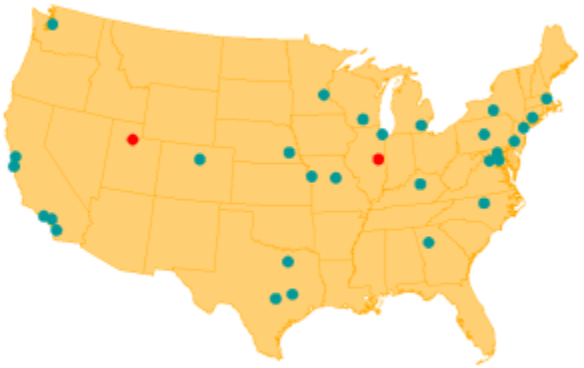
some geographic limitation



history again

really brief history in the US

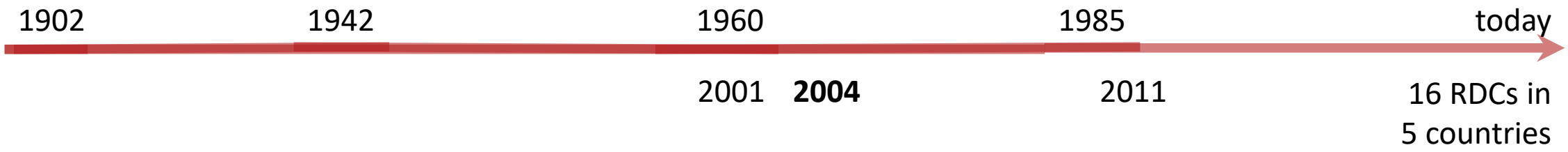
- Starting in the 1960s and 70s, increased use of public-use microdata samples and surveys
- Researcher access at Census Bureau headquarters in the 1970s
- 1990: [Computing power: 3.5 MFLOPS for $9000]
- First RDC at Boston in 1994
- A small number of RDCs in the 1990s
- Thin clients in the 2000s
- 2019: 29 RDCs



other countries: Germany

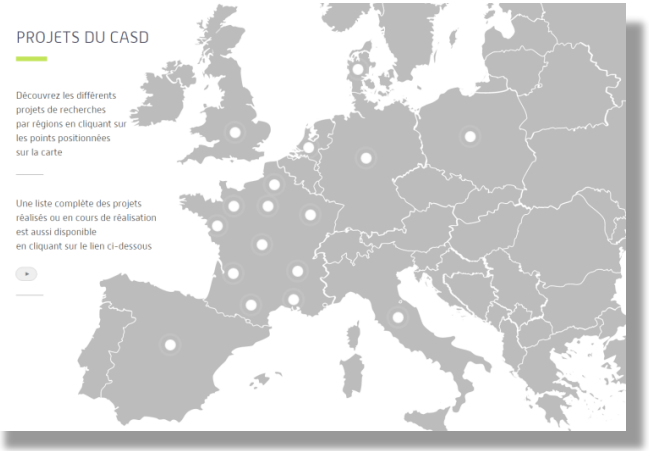
- Institute for Employment Research (IAB), Germany
  - *Commission to improve the informational infrastructure between the scientific community and official statistics* (KVI)

    recommended creation of RDCs by producers of microdata (2001)
  - RDC created in 2004 for "weakly anonymous" data
  - Scientific use files (factually anonymous data) available under licensing agreements to **university data enclaves**
  - 2011 first non-European RDC created at University of Michigan



other countries: France

- Centre d'accès securisé distant (CASD, France)
  - INSEE recommended implementing a secure center for accessing data (2007)
  - 2008 modification to Statistics Law made possible pilot infrastructure
  - Pilot infrastructure becomes permanent in 2009
  - Expansion with per-project cost (invoicing) in 2012



Loi Informatique et Libertés (1978)

1985 1990 1995 2007 2008 **2009** 2012 2018

1st working group

371 APs in 8 countries

Source: presentation to the CASD Scientific Committee in 2013

mechanisms

Variation in access mechanisms

	USA	France	Germany	Canada
Physical enclave	Yes	No	Yes/ No	Yes
Data in same enclave	No	--	No	Yes
Custom hardware	No	Yes	No	No
Standard hardware	Yes	Yes	Yes	No
Multiple data sources/providers	Yes	Yes	No	No/Yes

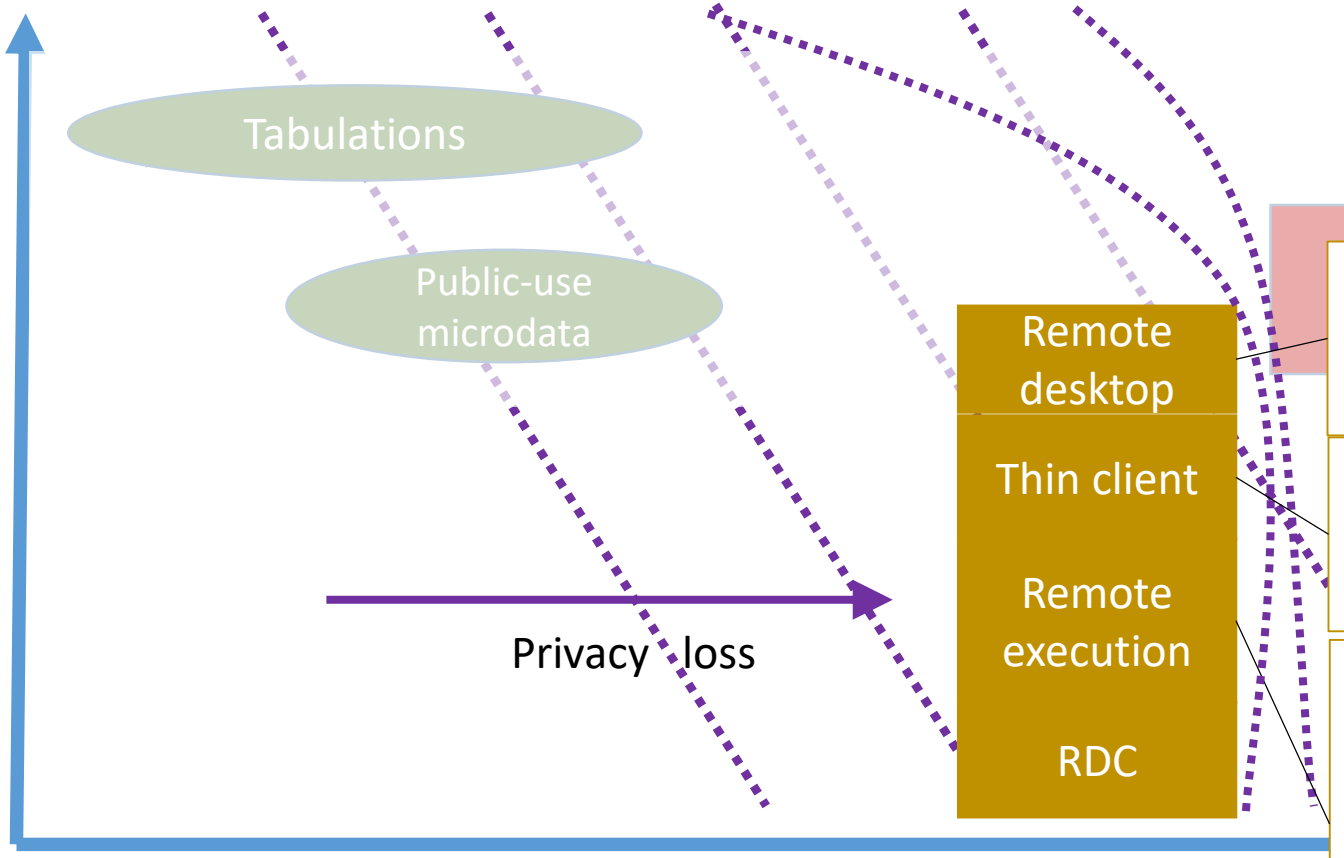
newer methods: **Data Enclaves**

- custom tabulations (by staff) became too onerous
- tabulation and analysis work offloaded onto researchers by providing them with access to protected microdata

what is a data enclave?

- Secure and authorized room
- May be under control of statistical agency but possibly outside the agency ("embassy" model)
  - In some cases may be under contractual control (university)
- May be used to describe
  - The room housing the data
  - The room which the researcher enters
- When the two differ: "Virtual Data Enclave"

access methods: enclaves



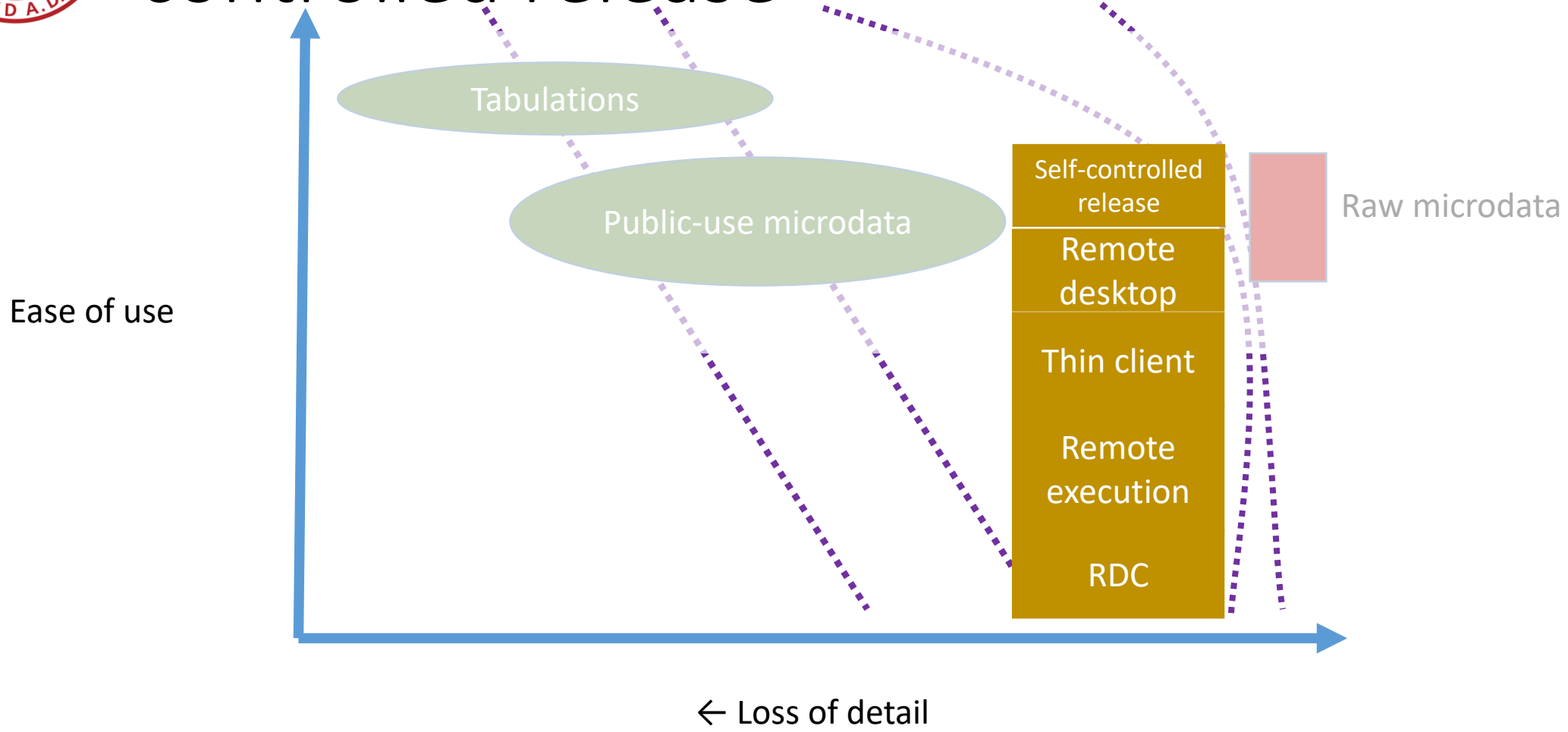
Software on your own PC giving a view onto secure data environment, with manual DA

Secondary secure PC giving a view onto secure data environment, with manual DA

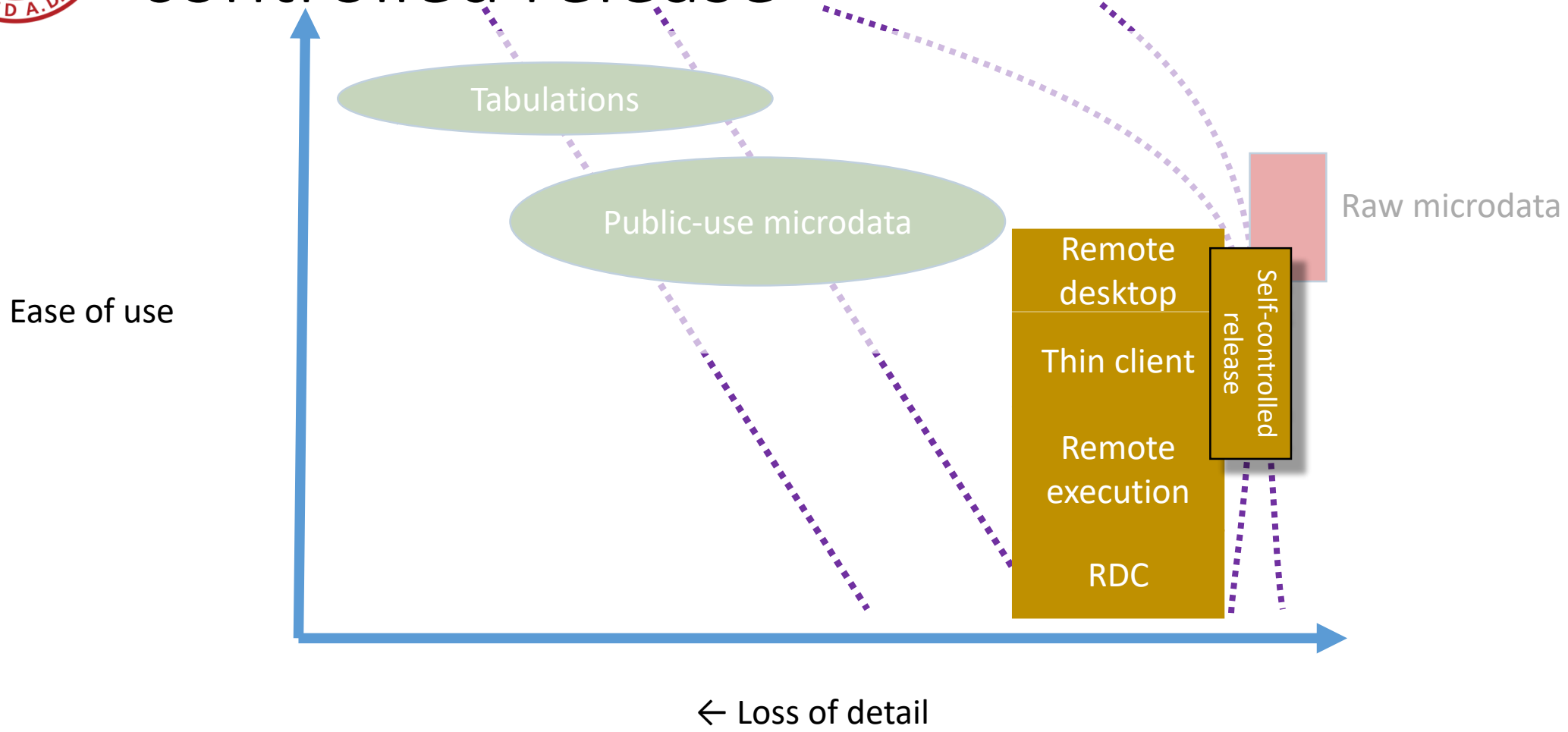
Submitting analysis programs by email or through website, with manual disclosure avoidance (DA)

(possibly combined with synthetic microdata)

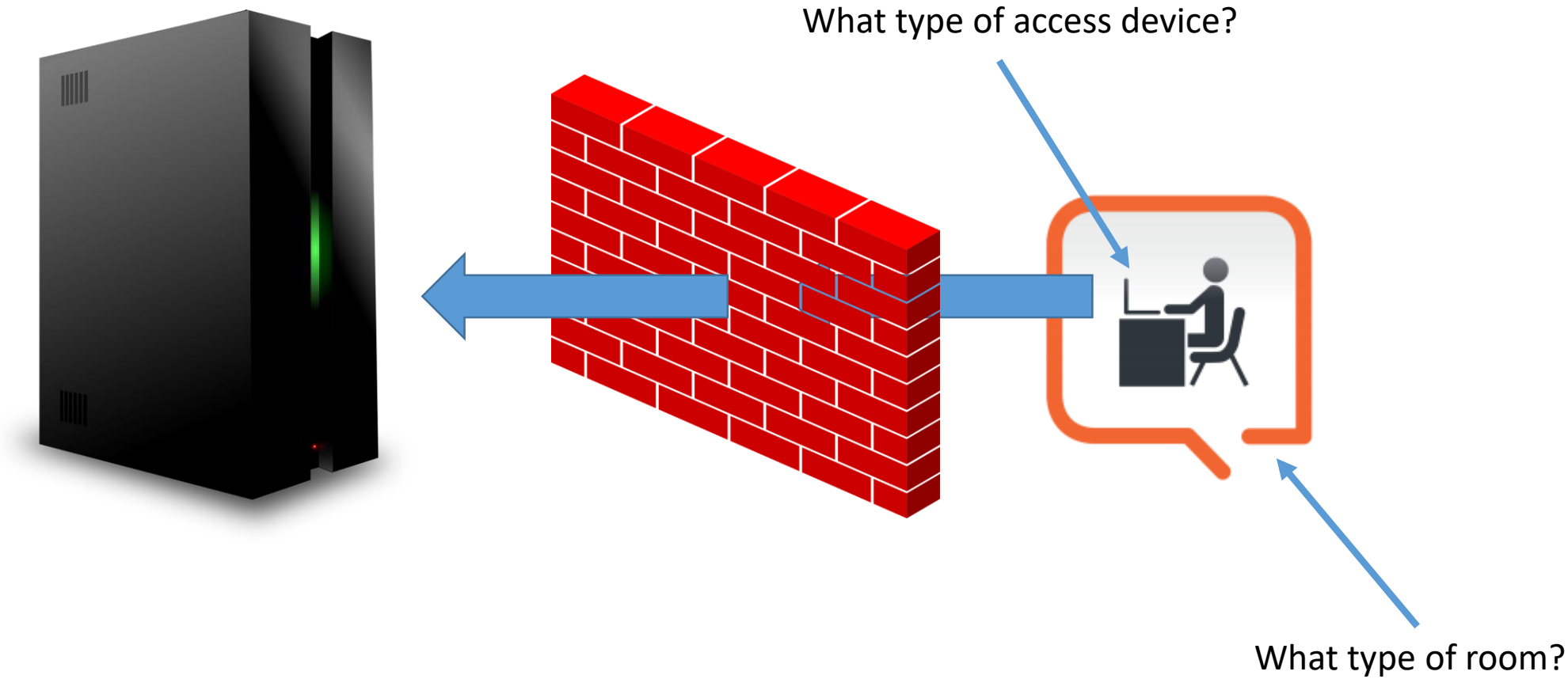
access methods: enclaves with researcher-controlled release



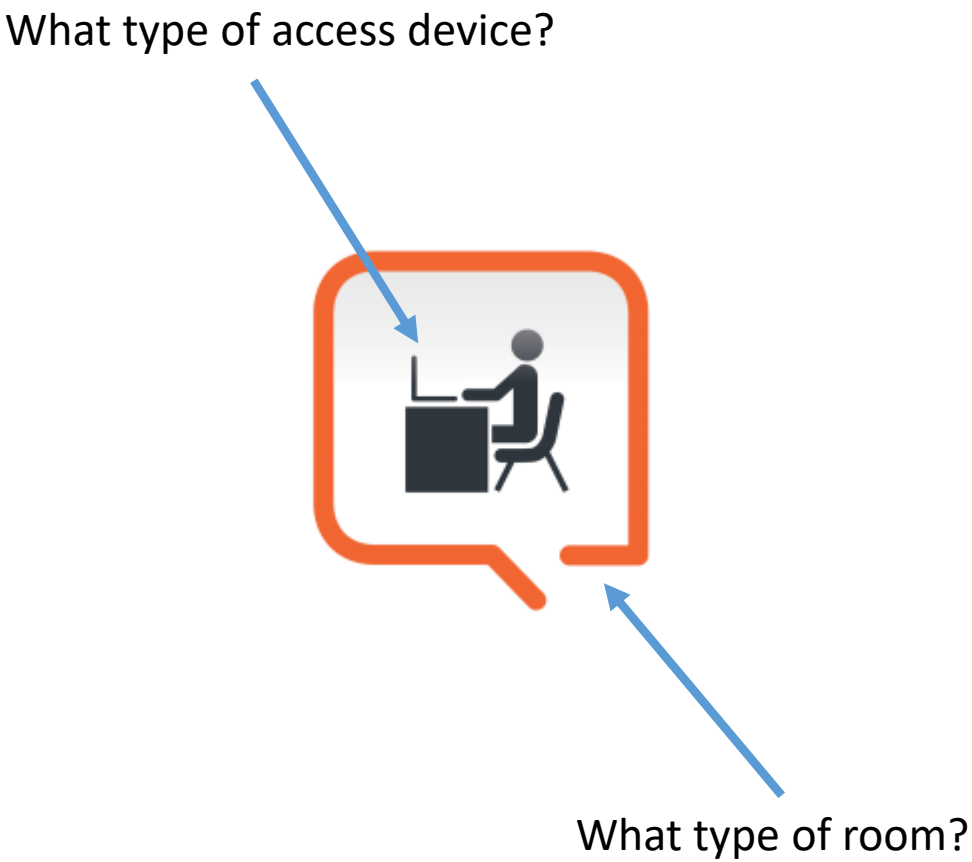
access methods: enclaves with researcher-controlled release



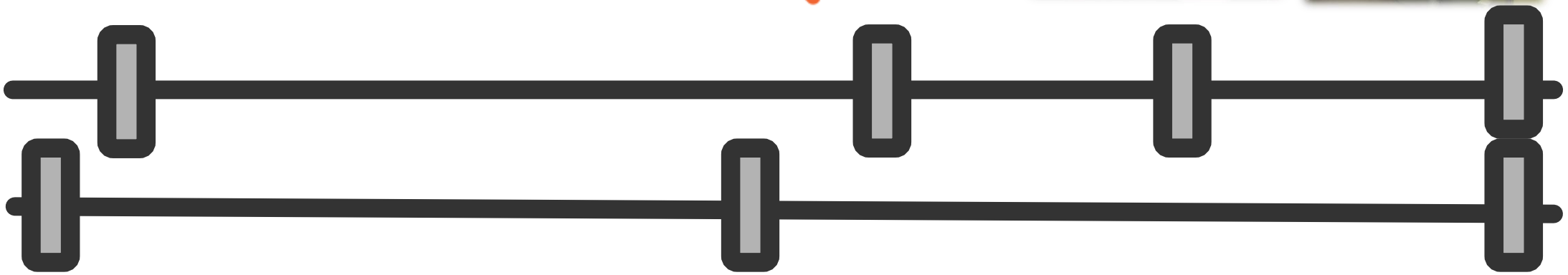
basic paradigm



basic paradigm



basic paradigm



“safe spaces”: that spaceship thing…

- Pre-fabricated secure room
- First one installed in 2015 at University of St. Andrews (Scotland/UK) [For now: EU]
- 2.3m x 1.8m (7’6” x 5’10”)
- Electronic locking, biometric recognition, CCTV, “Smart Glass”
- **£ 25,000 ~ $30,000** incl. installation
- Part of UK ADRN

Source: http://libanswers.st-andrews.ac.uk/faq/93548

access devices: thin clients

- With the notable exception of the Canadian RDCs (for now), **thin clients** are the preferred method of access
  - Surrounded by **walls** = RDC [FSRDC in US, Germany, others]
  - Embedded in a managed device = “thin client” [above, plus France]
  - Software with a managed access token = “remote desktop” or “VDI” [some US agencies; DK, Finland]
- Additional controls may be
  - IP address control [many]
  - Biometric authentication [France]
  - Smart card [France, US]

thin clients: examples

- Often off-the-rack devices
- Custom remote access device used at CASD
- Encrypted storage, biometric smartcard reader, pre-configured VPN
- **€35.00** / month, first user free, additional users €37.00 - €20.00 / month (decreasing)

thin clients: examples

- Software-based VPN
- Software-based remote access software (Citrix, RDP, etc.)
- Combined with various levels of 2FA
  - Software or hardware tokens
  - IP address limits

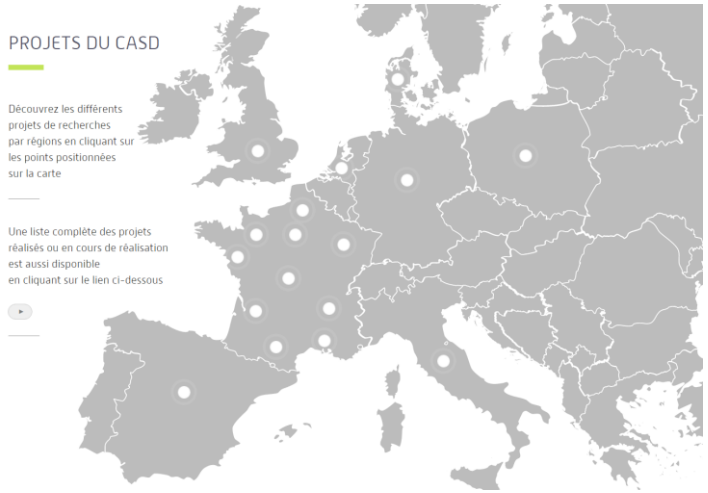
scalability (CASD)

… added in 2016 alone

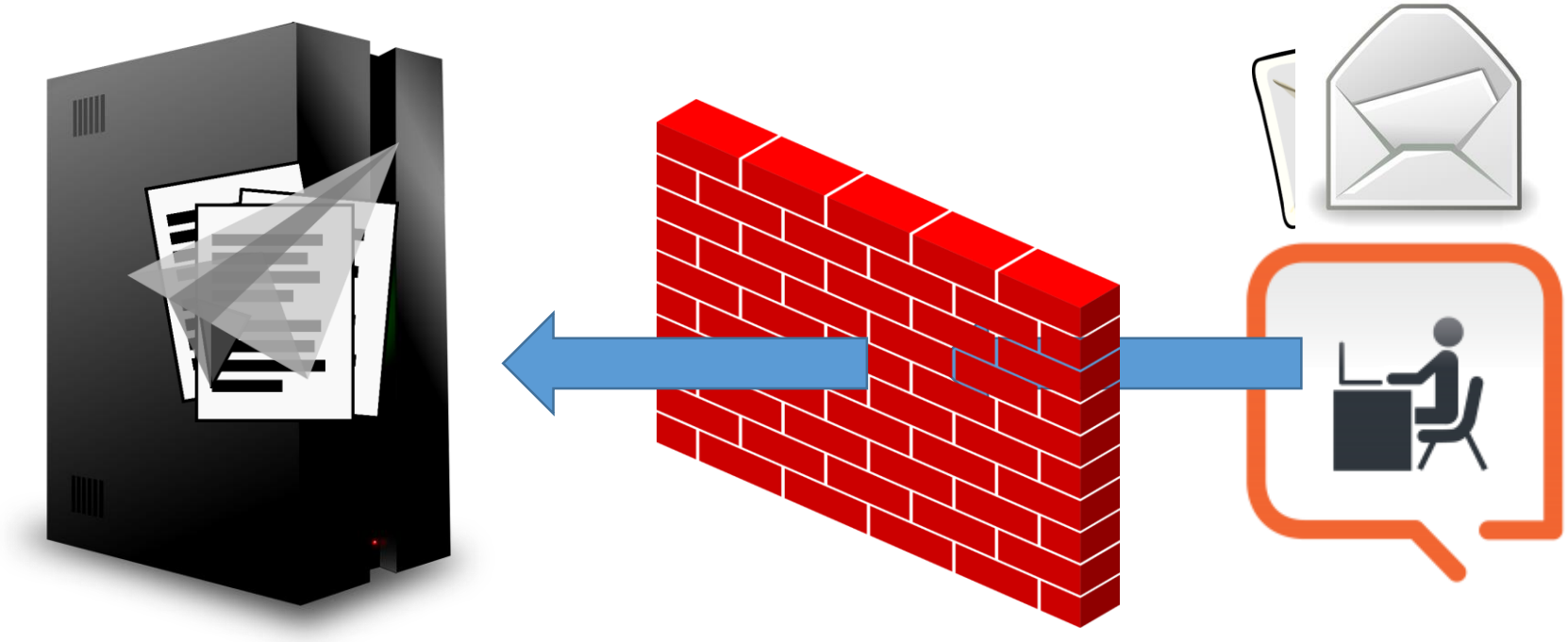
- 71 access points
- 232 users
- 62 projects

Totals

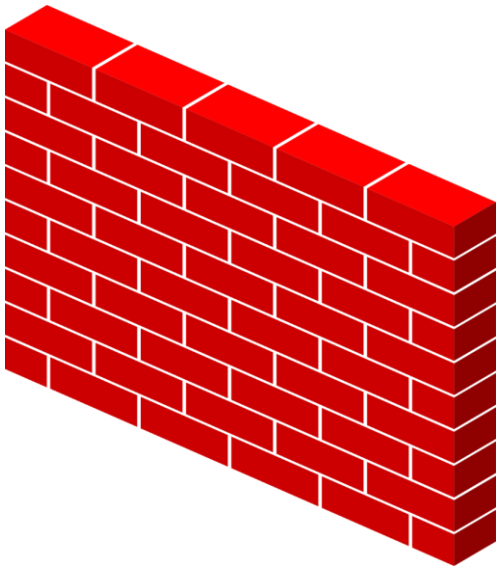
- 371 access points
- 1402 users
- 472 projects



remote processing paradigm



remote processing paradigm



Access matrix for remote submission

Control of:	Access computers	Access rules	Analysis methods	Disclosure avoidance	Cost
CB: Synthetic data	Custom compute cluster	Simplified proposal	Any (SAS, R Stata, Matlab)	Manual/ traditional	$0
IAB: JoSuA researcher	Web application	Full proposal	Smaller (Stata, whitelist commands)	Manual/ traditional	$0
Australian TableBuilder	Web application	Registration	Tables only	Embedded/ tab. noise infusion	$0/ >$0
Canada RTRA	Upload through Web	Simplified proposal + license	Smaller (SAS, whitelist commands)	Automated controlled rounding	$0
NCHS	Upload through FTP	Full proposal	Smaller (SAS, whitelist commands)	Manual/ traditional	$750/mth

The ultimate remote submission

Co-author with an employee of stats agency…

remote access setup

- Some setup required
  - Maybe require some “on-site” access [IAB]
  - May require billing to be set up [NCHS, others]
  - May require custom statistical language/limitations
- Testing in order to process remotely
  - Dummy or test files [IAB], “Synthetic” files [StatCan], Pre-defined data dictionaries [NCHS] **to test syntax**
  - Analytically valid synthetic data [US/Cornell] **to develop models**

data/output release methods

- Manual disclosure avoidance analysis [most]
- Self-disclosure by researcher [Denmark]
- Automated disclosure avoidance
- Verification server [in combination with synthetic data]
- Free [most] or metered [F, NCHS]

summary

To implement “safe spaces”, physical controls matter

- Various degrees of control over the physical space for the researcher
  - [none] - [custom-built secure room in state facility]
- Mostly remote access to data stored in secure computing facilities
- Various degrees of control over physical devices used for access
  - [software only] - [custom-built secure devices]

summary

To implement “safe spaces”, physical controls matter

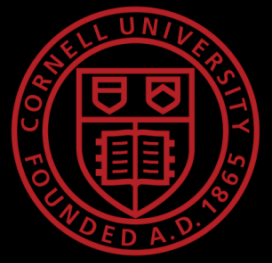
But:

Not independent of the four other safes, in particular

- safe data
- safe people

thank you

lars.vilhuber@cornell.edu



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