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

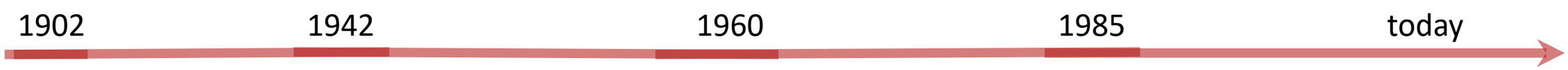
confidentiality of statistical agency data

- "… 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…"

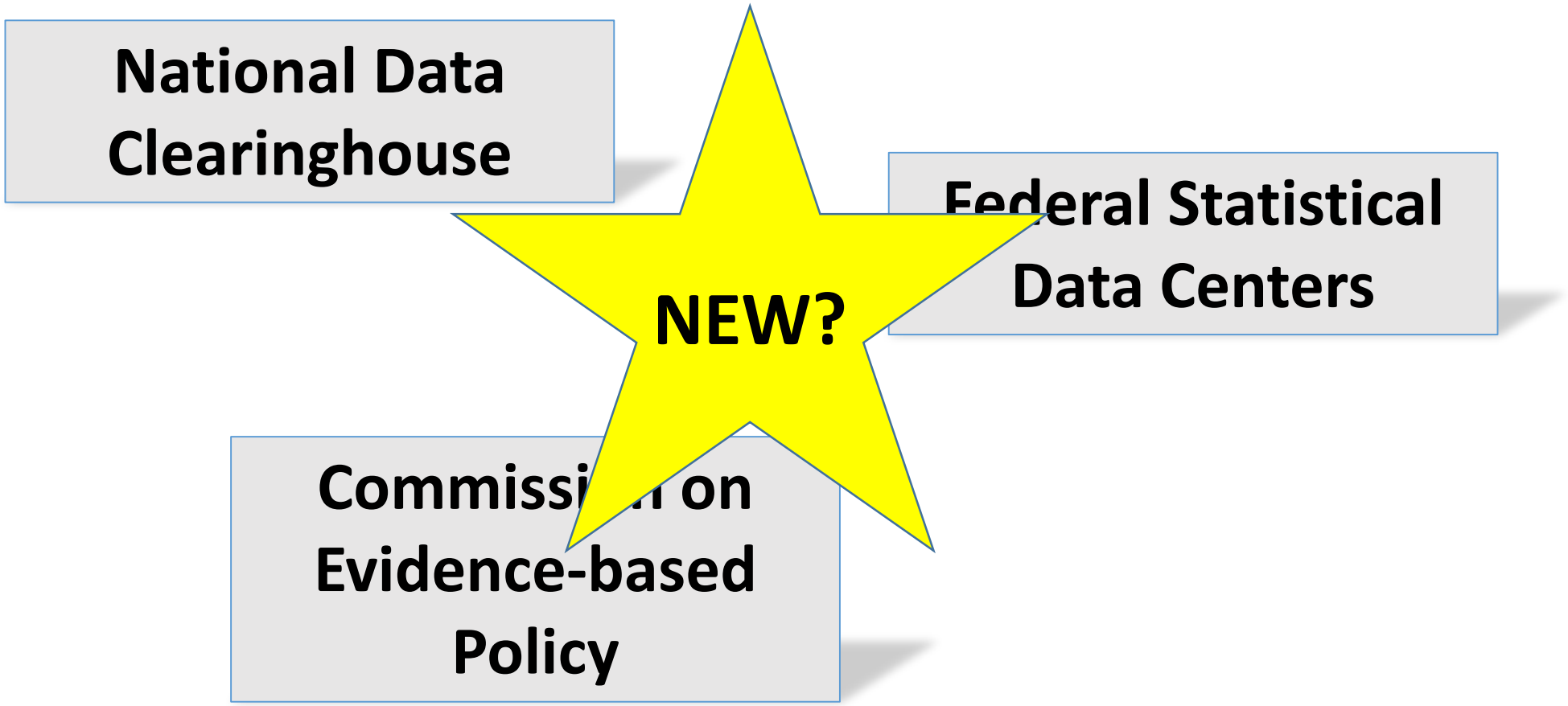
(Walter Wilcox, 1914)

rich new analysis and publications

held back by concerns of citizens and businesses about privacy



Making data (more) accessible

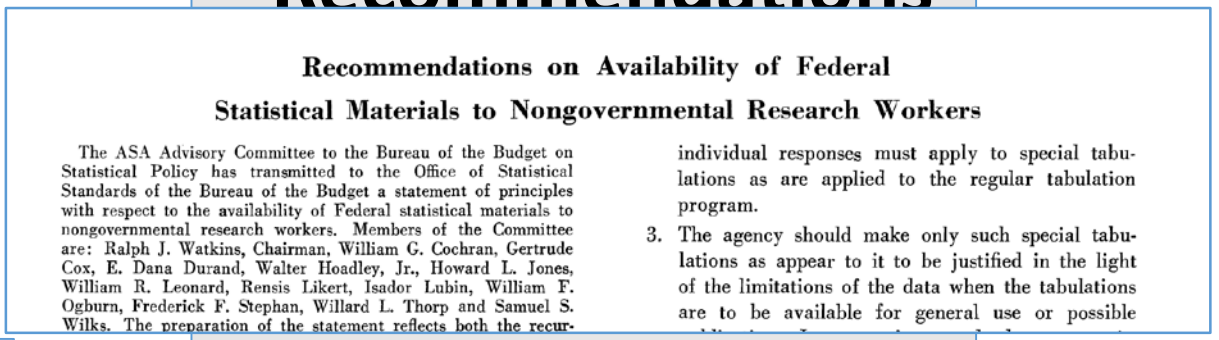


Making data (more) accessible

National Data Bank

Committee on the Preservation and Use of Economic Data

"Recommendations



Recommendations on Availability of Federal Statistical Materials to Nongovernmental Research Workers

The ASA Advisory Committee to the Bureau of the Budget on Statistical Policy has transmitted to the Office of Statistical Standards of the Bureau of the Budget a statement of principles with respect to the availability of Federal statistical materials to nongovernmental research workers. Members of the Committee are: Ralph J. Watkins, Chairman, William G. Cochran, Gertrude Cox, E. Dana Durand, Walter Hoadley, Jr., Howard L. Jones, William R. Leonard, Rensis Likert, Isador Lubin, William F. Ogburn, Frederick F. Stephan, Willard L. Thorp and Samuel S. Wilks. The preparation of the statement reflects both the recur-

individual responses must apply to special tabulations as are applied to the regular tabulation program.

3. The agency should make only such special tabulations as appear to it to be justified in the light of the limitations of the data when the tabulations are to be available for general use or possible

Nongovernmental Research Workers"

MAKING DATA (MORE) ACCESSIBLE

NATIONAL DATA BANK (1965)

1959-1965

"RECOMMENDATIONS ON AVAILABILITY OF [F]EDERAL STATISTICAL MATERIALS TO NONGOVERNMENTAL RESEARCH WORKERS" (1959)

COMMITTEE O[N] THE PRESERVAT[I]ON AND USE OF ECONOMIC DATA (1965)

Src: Kraus (2011) https://www.census.gov/history/pdf/kraus-natdatacenter.pdf

driven by advances in technology…

"These improvements resulted not only in time and space savings, but cost savings as well, enabling researchers to do **more detailed research** and respond more quickly to pressing social issues. […] **government programs** designed to address social issues […] **called for more information** and data on those issues. […] As research needs grew and research capabilities expanded, […] **increasing the demand for data**." 1960s!

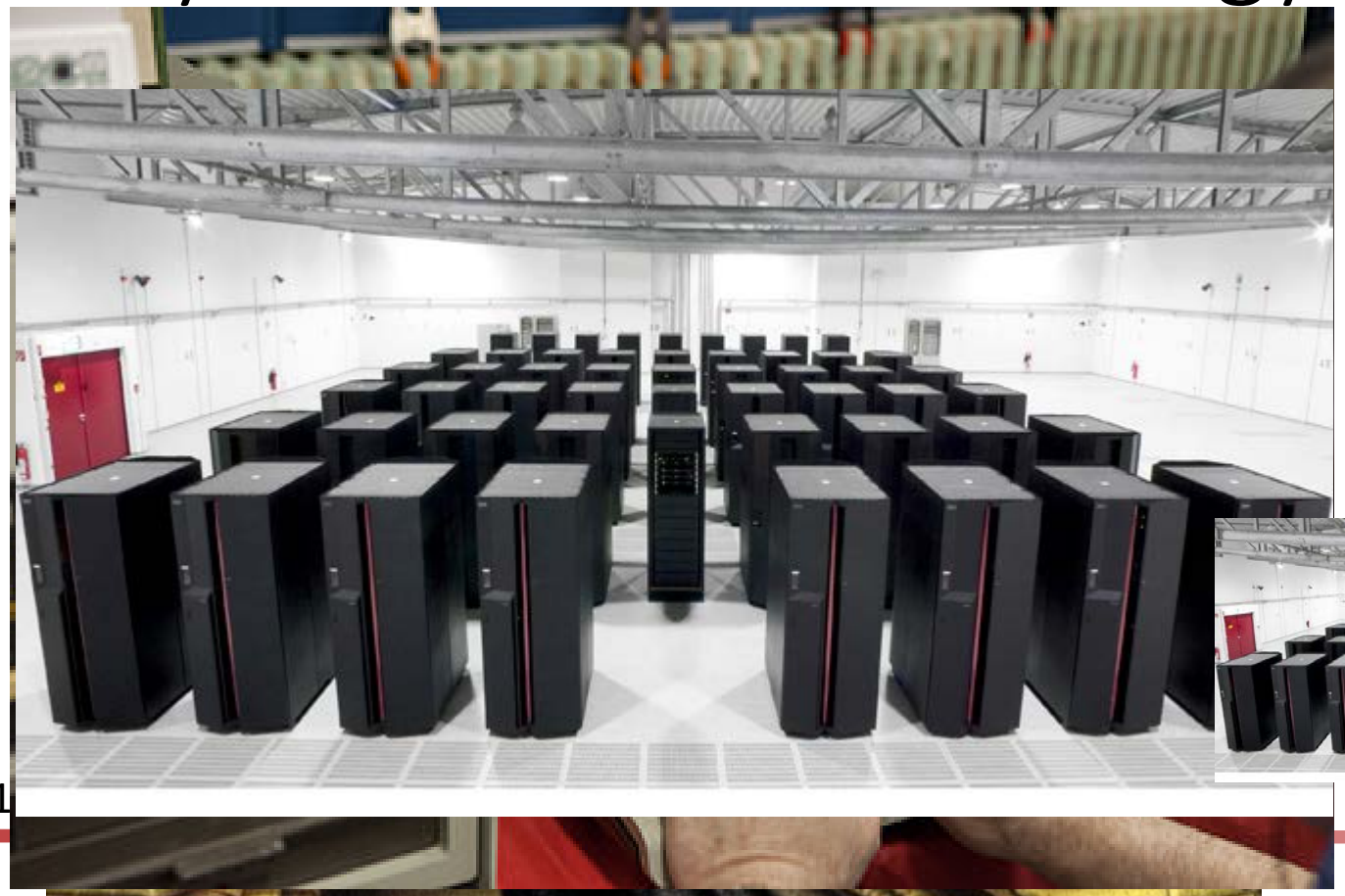
Src: Kraus (2011) https://www.census.gov/history/pdf/kraus-natdatacenter.pdf

professional associations

- At the 1959 annual meeting of the **American Economic Association**, members of the executive committee discussed the need for access to social and economic data for research purposes
  → **Ruggles Report** in **April 1965**
- **American Statistical Association** (ASA) Advisory Committee assisted Bureau of the Budget (pre-OMB)
  → "**Recommendations** on Availability of Federal Statistical Materials to Nongovernmental Research Workers," The American Statistician, vol. 13, no. 4 (**October 1959**) DOI: 10.1080/00031305.1959.10482600

driven by advances in technology

1902



today

researchers knocking on the door

researcher access and privacy concerns

- 1960s in the US: proposal for “National Data Bank” with the goal of combining survey and administrative data to make available to researchers
  - Instead, and partially as a consequence, privacy laws were formalized in the 1970s (“Privacy Act 1974” (Public Law 93-579, 5 U.S.C. § 552a)) specifically prohibited “matching” programs, linking data from different agencies.
- More recently: 2016 Australian Census elicited substantial controversy
  - Identifiable data with explicit goal of enabling linkages between the census and administrative data, as well as linkages across historical censuses

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 a access mechanism

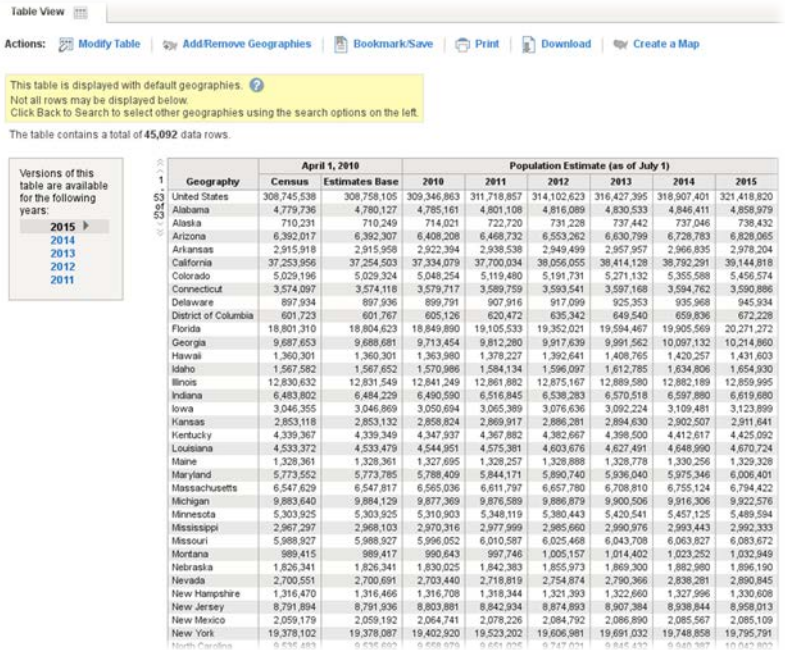


Table View

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.

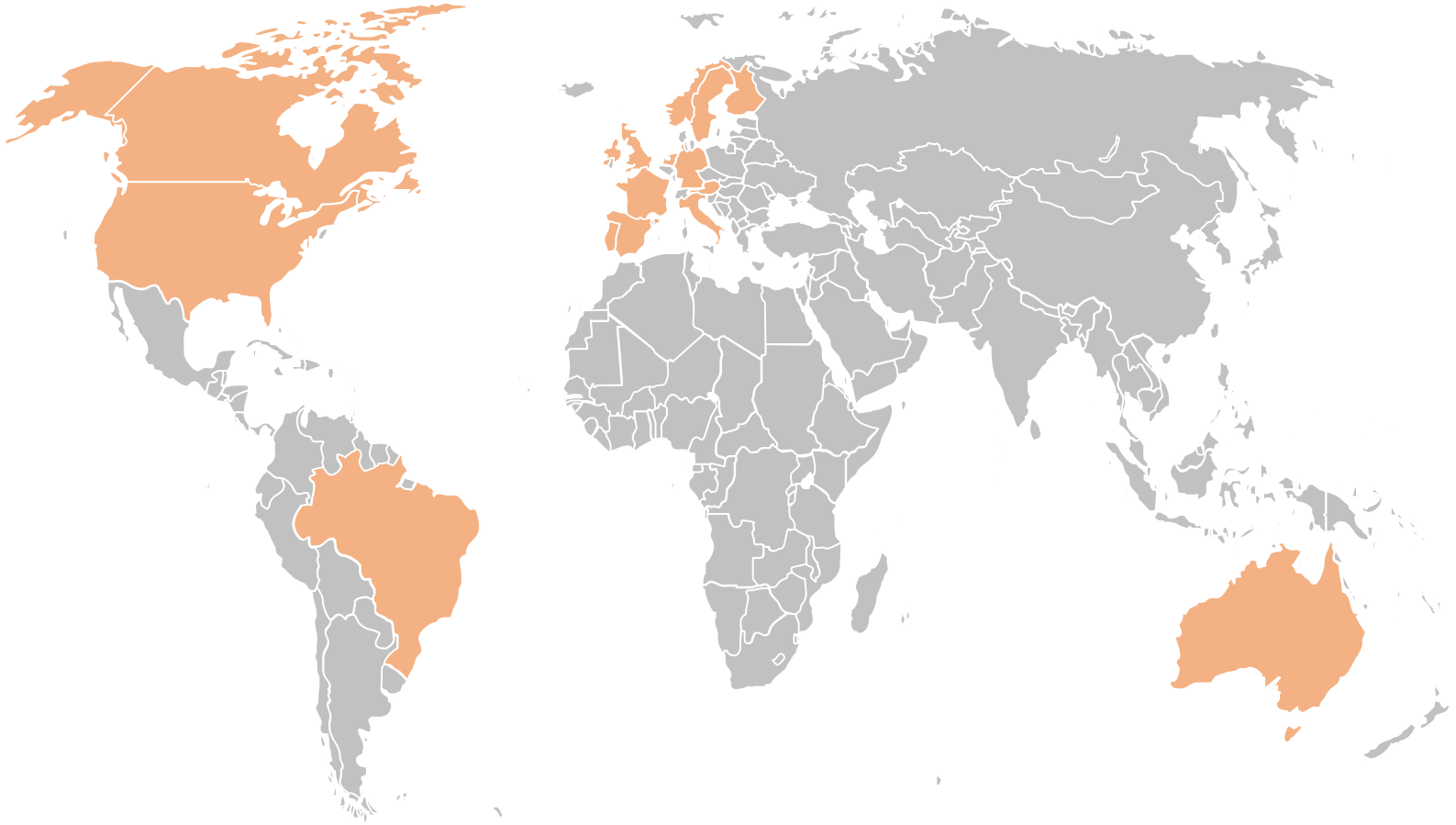
Versions of this table are available for the following years:
2015
2014
2013
2012
2011

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

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***
- Highlight the roles of “community”
  - Training
  - Legal framework
  - Role of institutions

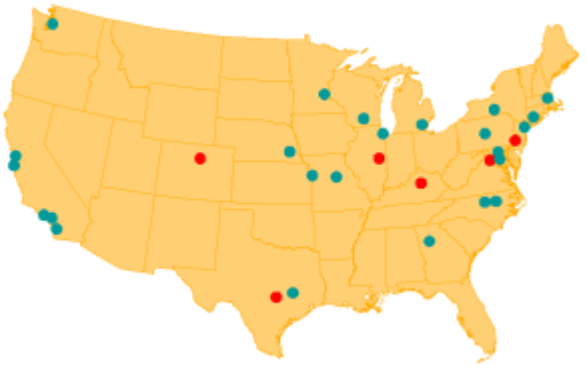
some geographic limitation



history again

really brief history in the US

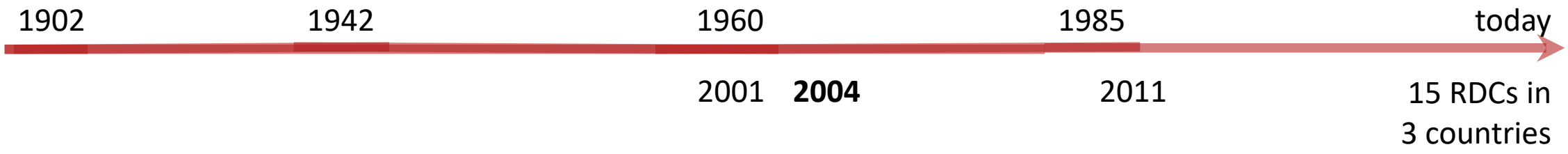
- Starting in the 1960s and 70s, increased use of public-use microdata samples and surveys
- Researcher access at Census Bureau headquarters in the 1970s
- 1990: [Computing power: 3.5 MFLOPS for $9000]
- First RDC at Boston in the 1994
- A small number of RDCs in the 1990s
- Thin clients in the 2000s
- 2016: 24 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 RDC created at University of Michigan (with NSF funding)



other countries: France

- Centre d'accès securisé distant (CASD, France)
  - Note: within same agency that enabled AKM (1999)
  - INSEE recommended implementing a secure center for
  - 2008 modification to Statistics Law made possible pilot
  - Pilot infrastructure becomes permanent in 2009
  - Expansion with per-project cost (invoicing) in 2012

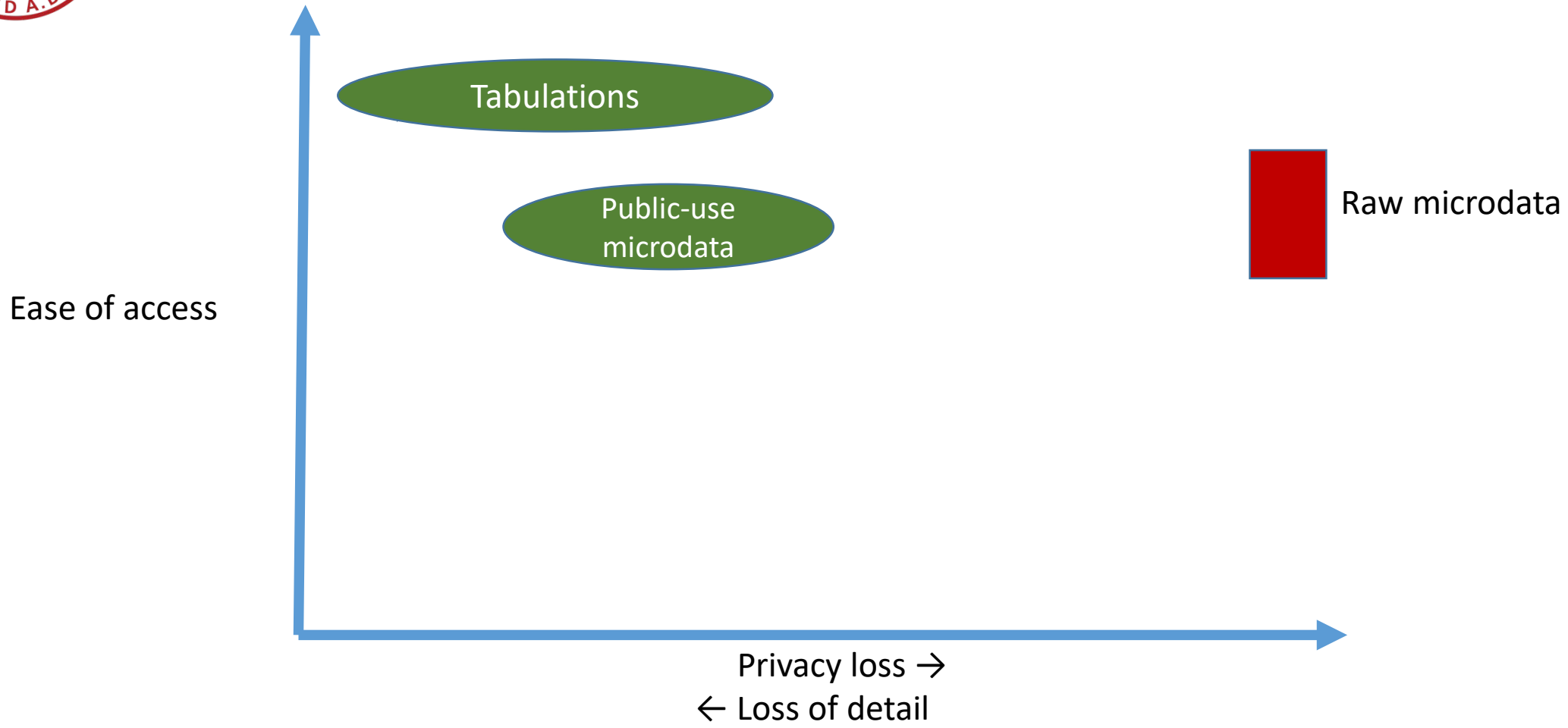
Loi Informatique et Libertés (1978)

1985 1990 1995 1st working group 2007 2008 **2009** 2012 today 371 APs in 8 countries

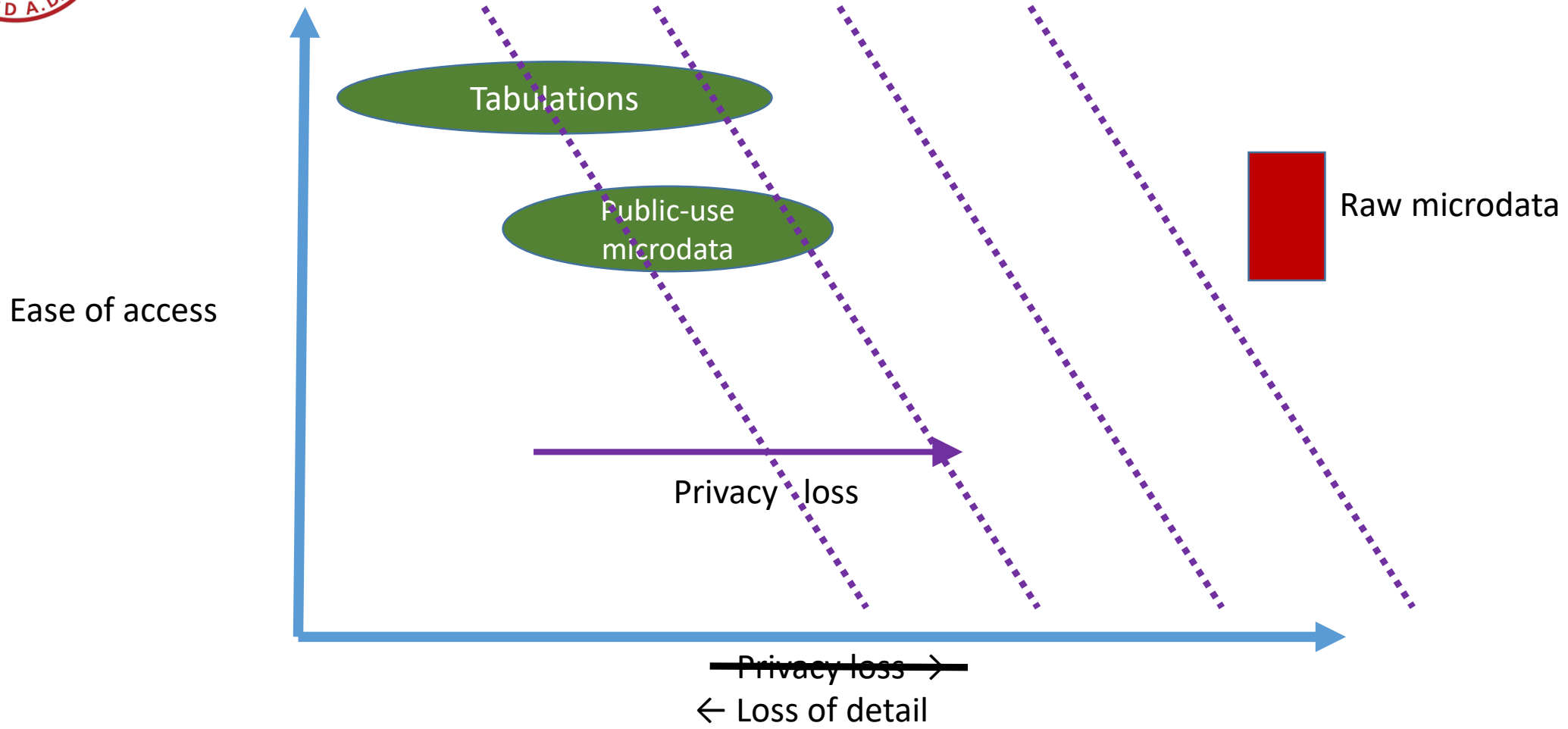
Source: presentation to the CASD Scientific Committee in 2013

mechanisms

access methods



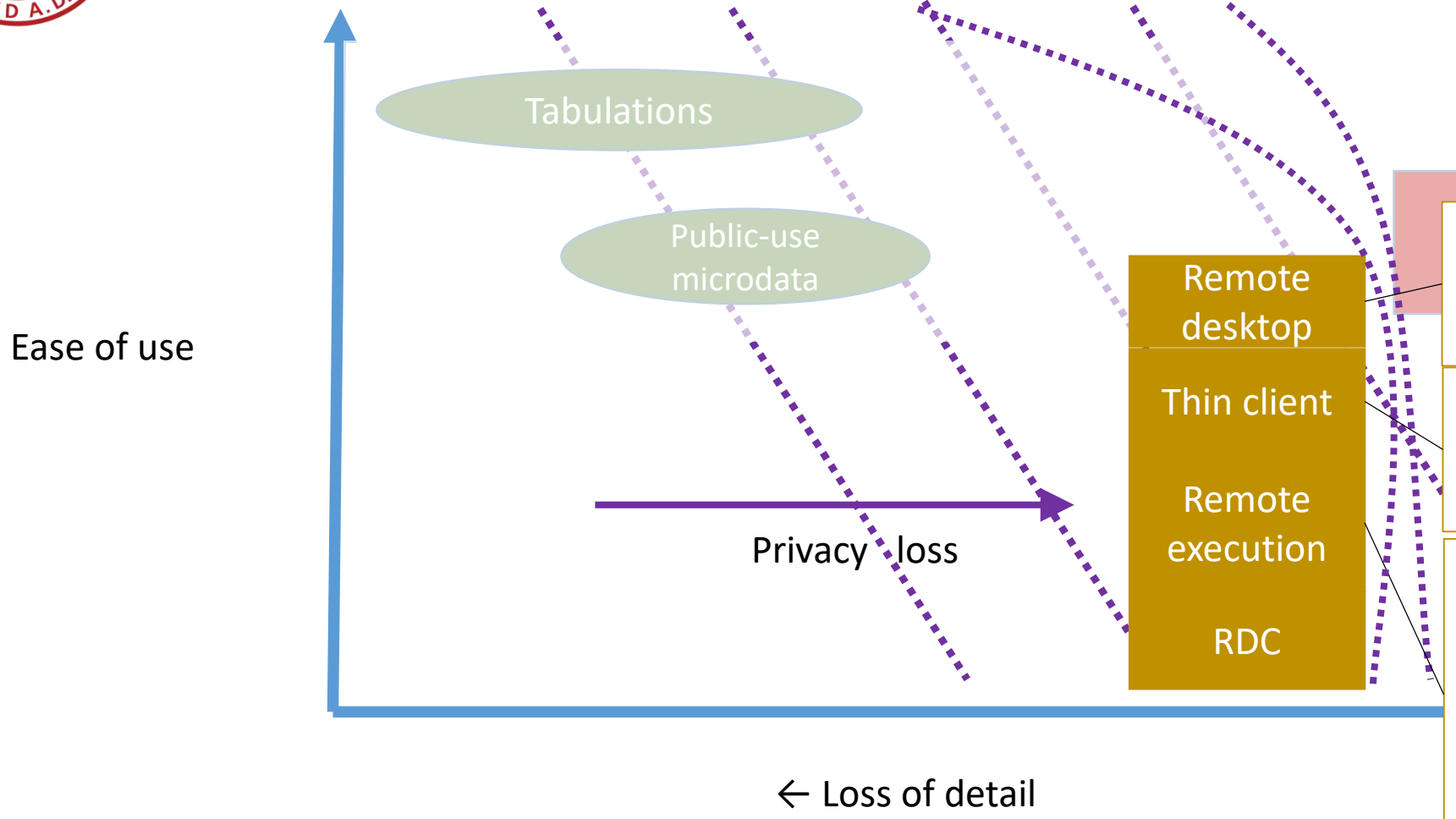
access methods



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

access methods: enclaves



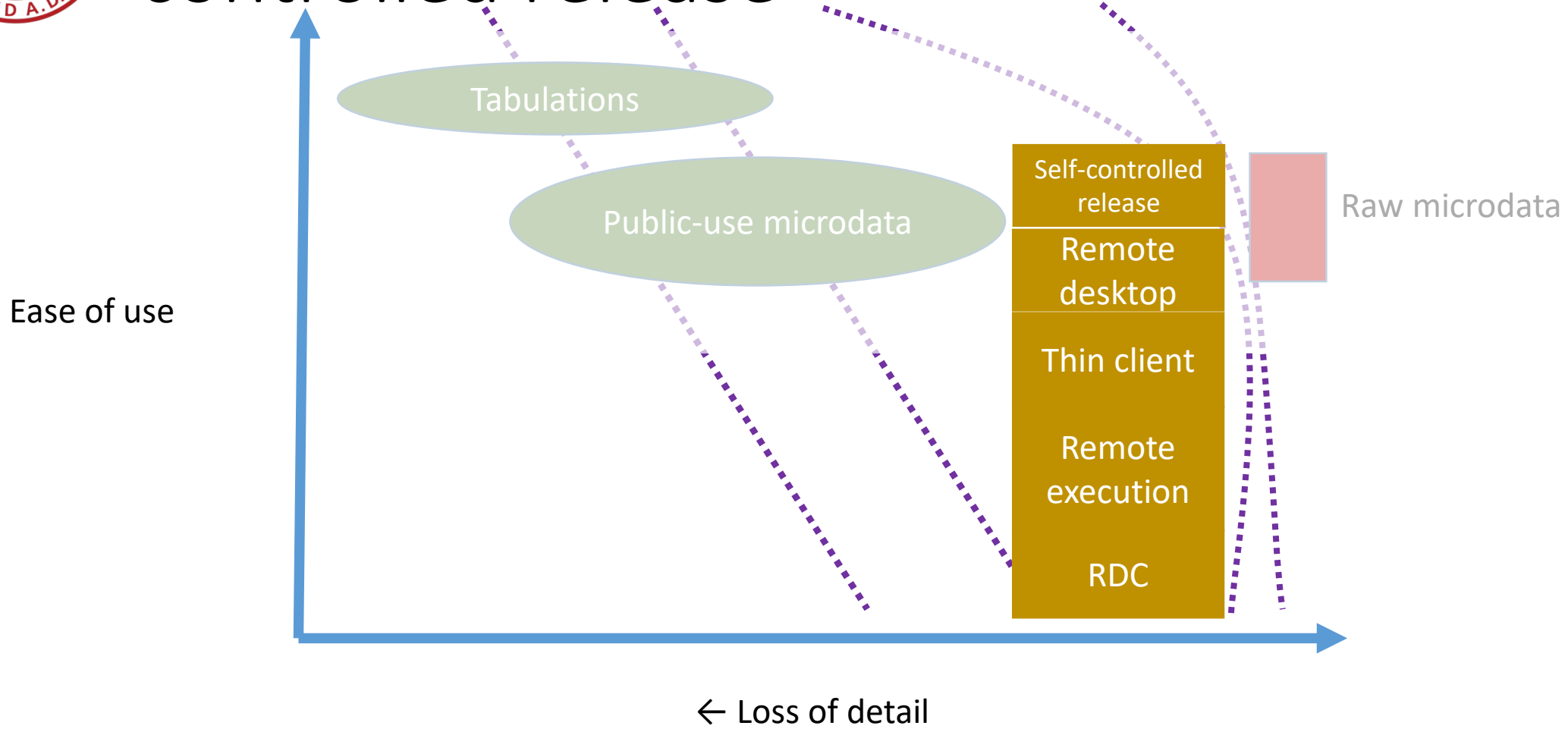
Software on your own PC giving a view onto secure data environment, with manual DA

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

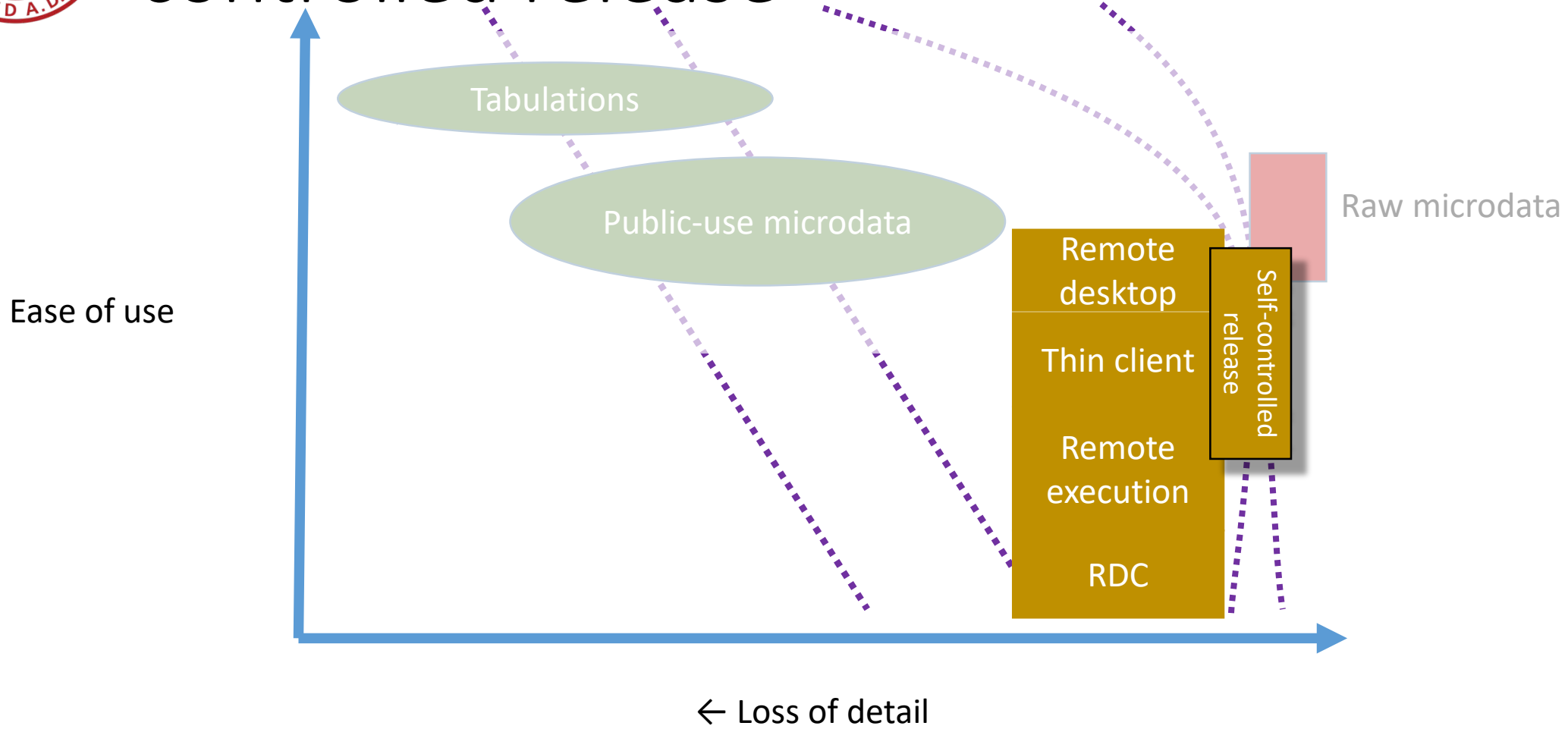
Submitting analysis programs by email or through website, with manual disclosure avoidance (DA)

(possibly combined with synthetic microdata)

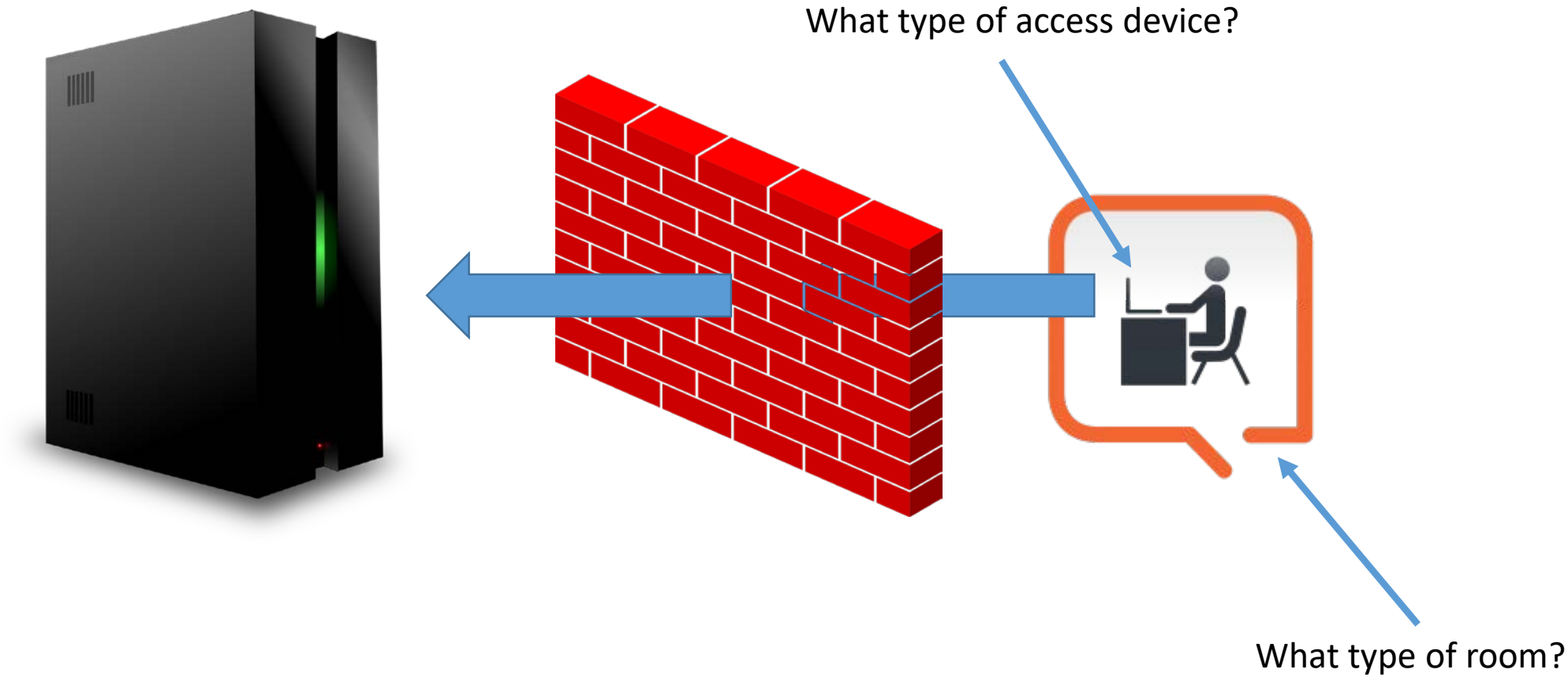
access methods: enclaves with researcher-controlled release



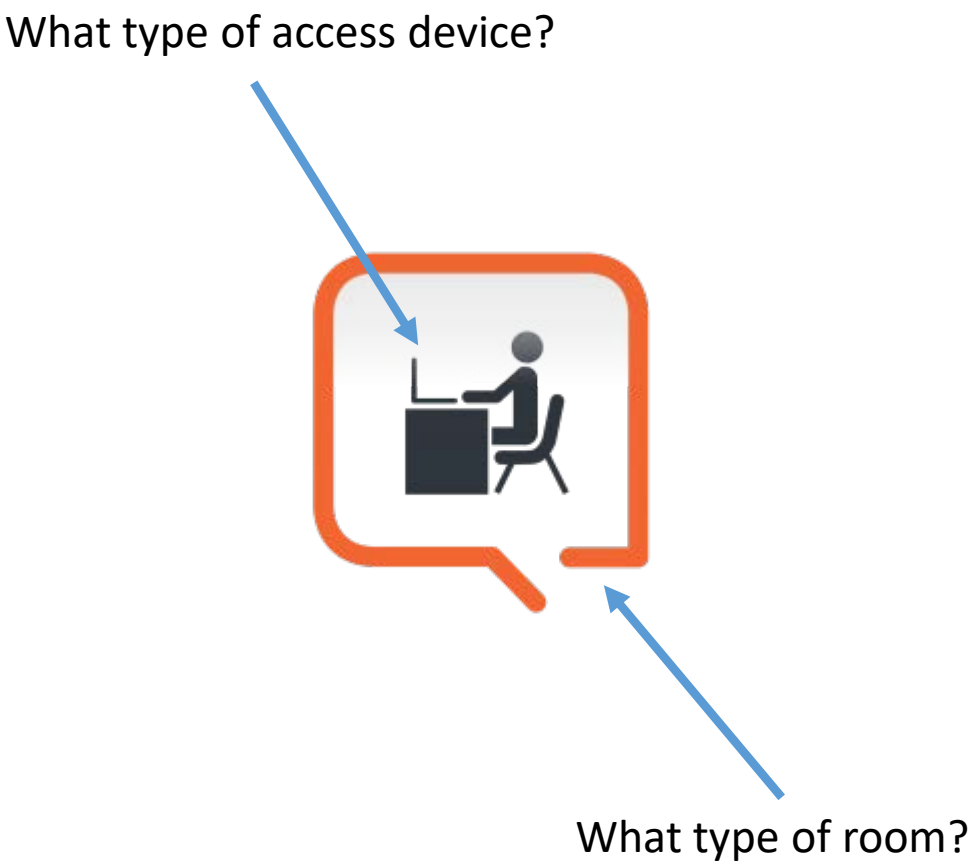
access methods: enclaves with researcher-controlled release



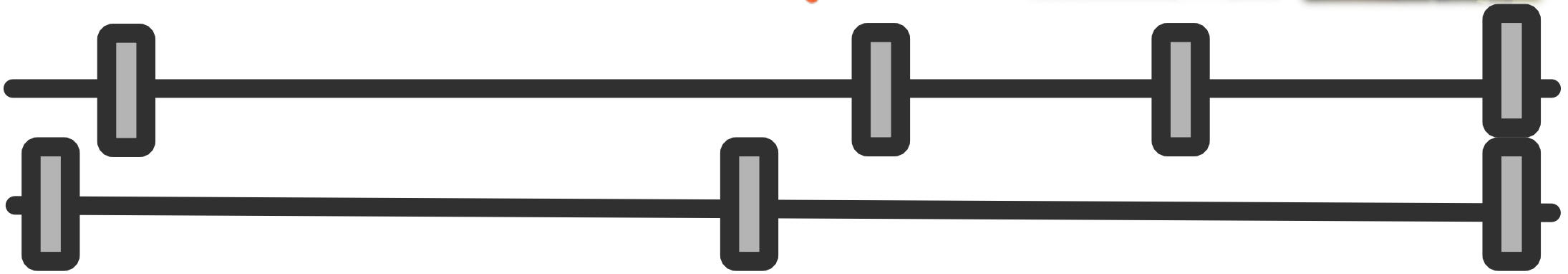
basic paradigm



basic paradigm



basic paradigm



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

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]

that box thing

- 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)

scalability (CASD)

… added in 2016 alone

- 71 access points
- 232 users
- 62 projects

Totals

- 371 access points
- 1402 users
- 472 projects

lessons to be learned?

The very first RDCs were in North America (USA and Canada)

European systems came later

But can they provide new insights for our systems?

Access matrix for confidential data

Control of:	Data access	Analysis computers	Access computers	Access rooms	Analysis methods
FSRDC researcher	**Full**	**Full**	**Full**	**Full (badge access)**	**Some (choice of software)**
Census employee	**Full**	**Full**	**None (VDI)**	**None (VDI)**	**Some (choice of software)**
IAB: RDC researcher	**Full**	**Full**	**Full**	**Full (trusted person)**	**Some (choice of software)**
IAB: JoSuA researcher	**Full**	**Full**	**None (Web application)**	**None (Web application)**	**Smaller (software, whitelist commands)**
IAB employee	**Full**	**Full**	**Full (IAB laptop)**	**None (VDI)**	**Some (choice of software)**
CASD researcher	**Full**	**Full**	**Extra Full (custom-built hardware)**	**Some (university office, EU)**	**Some (choice of software)**

Access matrix for confidential data

Control of:	Data access	Analysis computers	Access computers	Access rooms	Analysis methods
FSRDC researcher	**Full**	**Full**	**Full**	**Full (badge access)**	**Some** (choice of software)
Stat.Denmark (typical EU)	**Full**	**Full**	**None (VDI)** - **Some** (host institution)	**None (VDI)** - **Some** (host institution)	**Some** (choice of software)
RDC Canada	**Some** (demog. only)	**Some** (host institution)	**Some** (host institution)	**Full (badge access)**	**Some** (choice of software)
Stat.Canada (typical of HQ)	**Full** (incl. business)	**Full**	**Full**	**Full**	**Some** (choice of software)
RDC Canada (thin client)	**planned**				

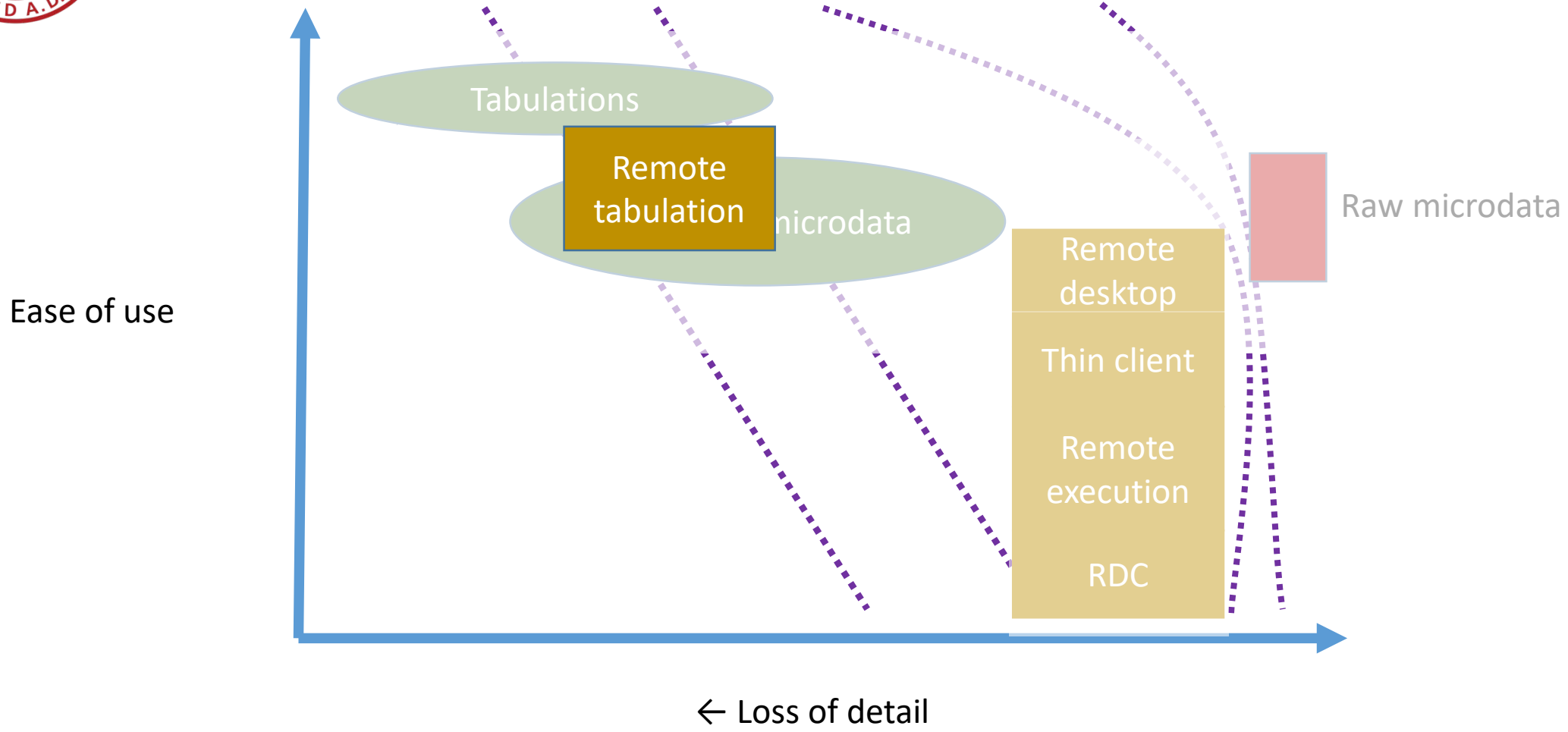
Access matrix for confidential data

Control of:	Data access	Analysis computers	Access computers	Access rooms	Analysis methods
FSRDC researcher	**Full**	**Full**	**Full**	**Full (badge access)**	**Some (choice of software)**
Census employee	**Full**	**Full**	**None (VDI)**	**None (VDI)**	**Some (choice of software)**
IAB: JoSuA researcher	**Full**	**Full**	**None (Web application)**	**None (Web application)**	**Smaller (software, whitelist commands)**
CASD researcher	**Full**	**Full**	**Extra Full (custom-built hardware)**	**Some (university office, EU)**	**Some (choice of software)**
Stat.Denmark (typical EU)	**Full**	**Full**	**None (VDI)** **- Some (host institution)**	**None (VDI) -** **Some (host institution)**	**Some (choice of software)**

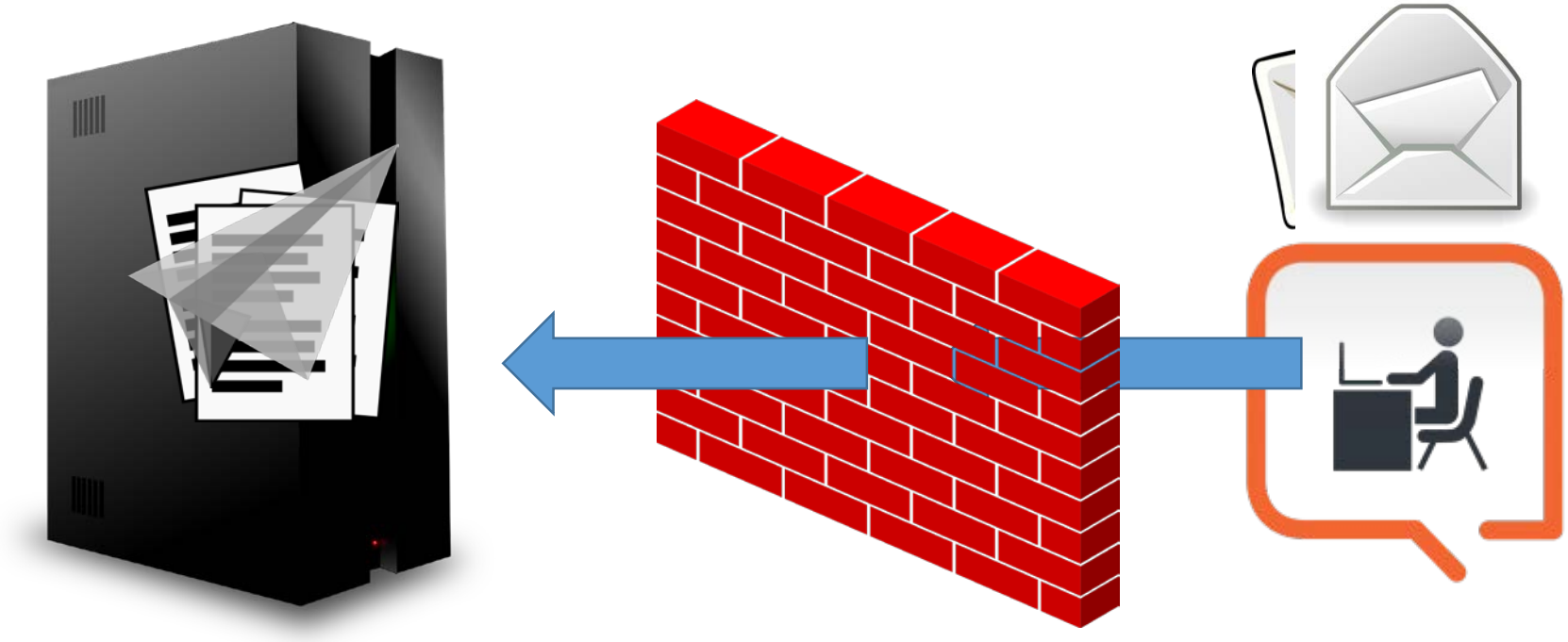
Access matrix for confidential data

Control of:	Data access	Access computers	Access rooms	Analysis methods	Disclosure avoidance
FSRDC researcher	**Full**	**Full**	**Full (badge access)**	**Some (choice of software)**	**Manual/ variety of rules**
Census employee	**Full**	**None (VDI)**	**None (VDI)**	**Some (choice of software)**	**Manual/ self/ variety of rules**
IAB: JoSuA researcher	**Full**	**None (Web application)**	**None (Web application)**	**Smaller (software, whitelist commands)**	**Manual/ variety of rules**
CASD researcher	**Full**	**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)	**Full**	**None (VDI)** - **Some (host institution)**	**None (VDI)** - **Some (host institution)**	**Some (choice of software)**	**Manual/ self/ variety of rules**

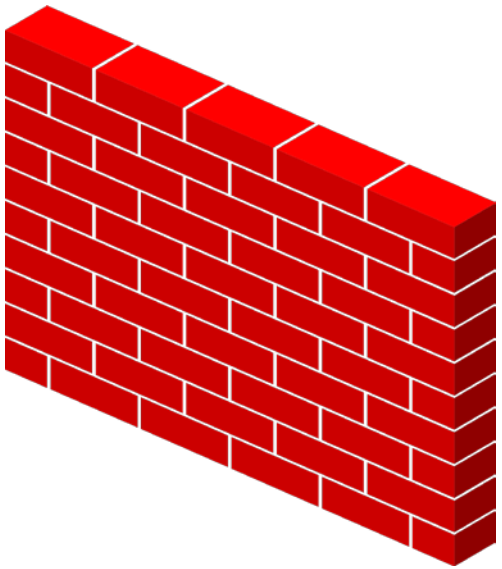
access methods: remote tabulation



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
  - IAB's JoSuA starts with a regular "on-site" access
  - NCHS has a regular proposal process, billing is involved
  - StatCan's RTRA has a proposal process, review, etc.
- Testing in order to process remotely
  - Dummy or test files (IAB)
  - "Synthetic" files (StatCan)
  - Pre-defined data dictionaries (NCHS)

synthetic data and remote submission

- StatCan: “synthetic” data = univariate draws, no analytic validity
  - IAB also creates these types of files, but calls them “test” files
- Census Bureau: “synthetic data” = analytically valid, conditional on congeniality of the model
  - Model is “verbally” described, but not formally

synthetic data: verification model

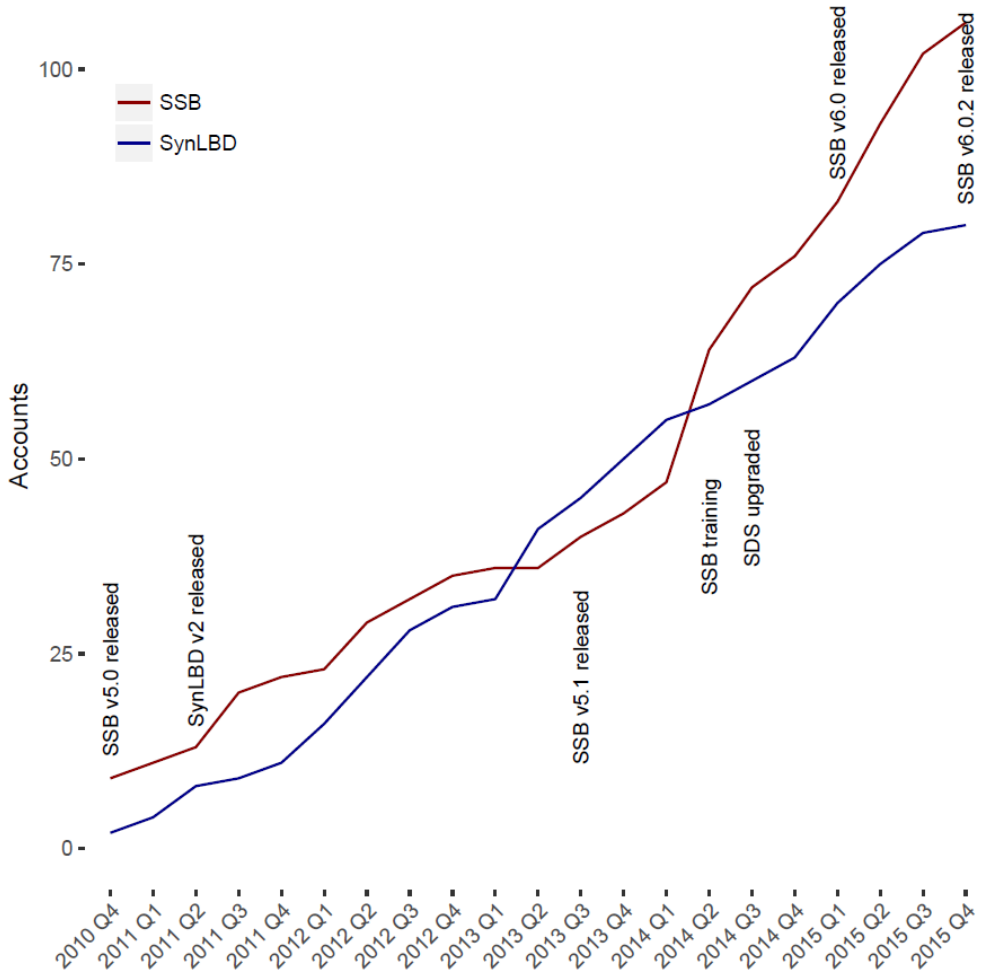
- Under development (Reiter, Machanavajjhala)
- Applied to OPM data
  - Researcher develops model on synthetic data
  - Assessment through submission of programs for “verification”
  - Researcher obtains (DP) indication of proximity to actual results
  - Restrictions on possible models (?)
- Under development
  - Come back in April for NCRN workshop

synthetic data: validation model

- Used for SIPP Synthetic Beta, Synthetic LBD
  - Researcher develops model on synthetic data
  - Assessment through submission of programs for “validation”
  - Researcher obtains actual results from model run on confidential data, subject to traditional disclosure avoidance rules
  - No restrictions on types of models
- In progress since 2011
  - Approx. 200 users
  - Approx. 6-8% of users request validation
  - Some unknown fraction “self-validate” through full FSRDC project

some results from Synthetic Data Server

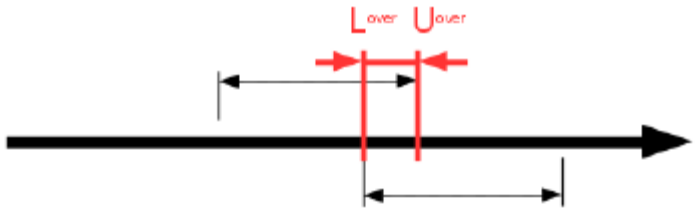
6 years, 5 (versions of) synthetic datasets, over 180 users



validation

- About 6-8% request validation
- Metric: confidence interval overlap J_k (Karr et al, 2006)

$$J_k^* = \frac{1}{2}\left[\frac{U^{over} - L^{over}}{U - L} + \frac{U^{over} - L^{over}}{U^* - L^*}\right]$$

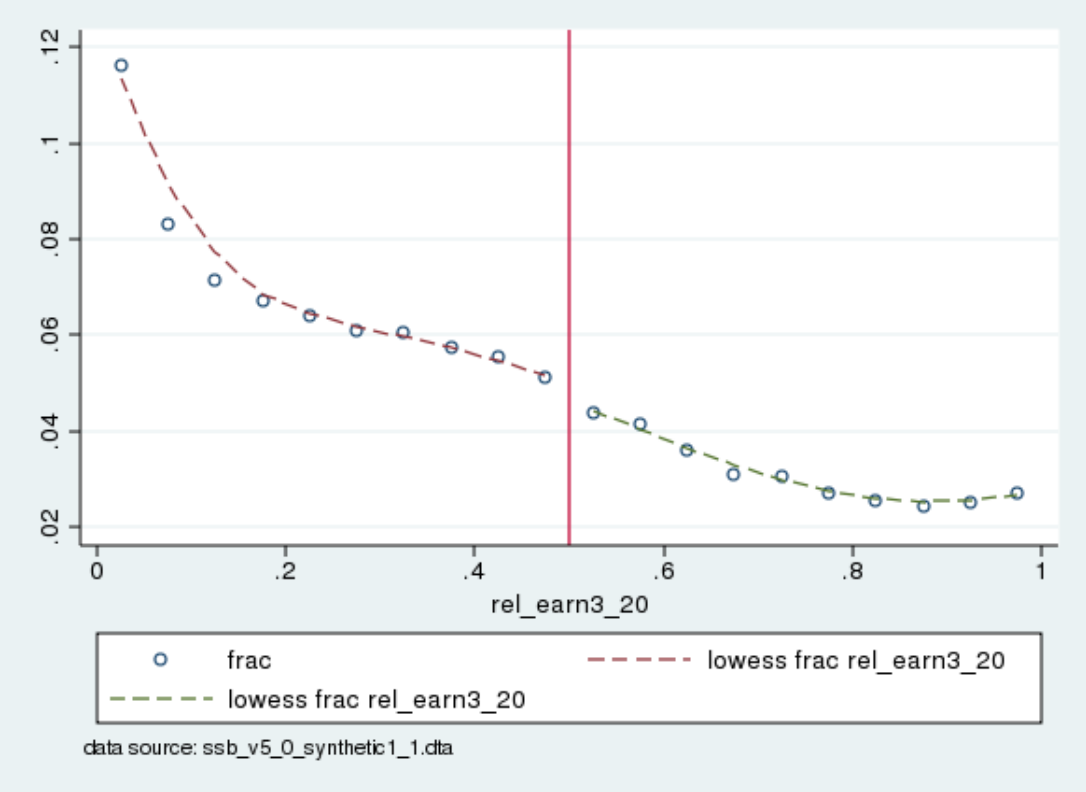


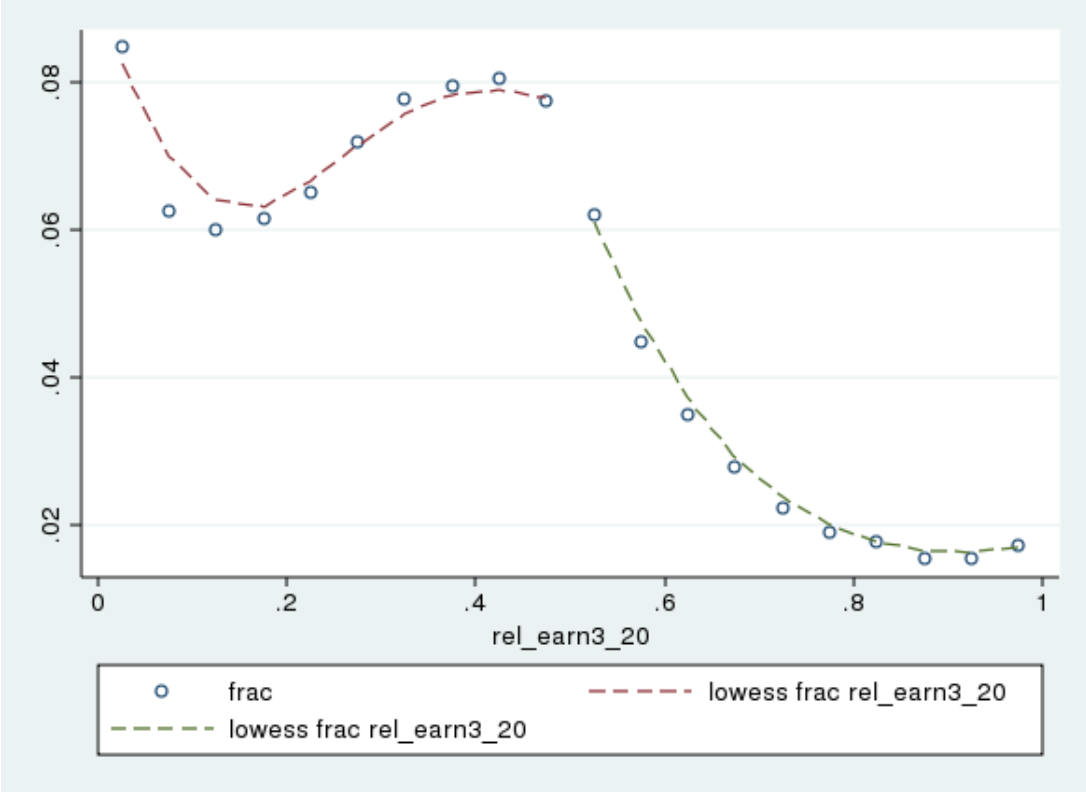
some results: SynLBD

Table 1: Confidence interval overlap $J^*_{k,m}$

User	Request	Mean	75th	90th	Max
A	1	0.160	0.246	0.725	0.889
A	2	0.101	0	0.523	0.924
B	1	0.869	1.000	1.000	1.000
C	1	0.219	0.509	0.725	0.995

an illustrative example (Bertrand et al, 2015)





some results: SSB

	Mean	Median	75%	95%	Max	PctGrtThan0
1	0.49	0.54	0.79	0.91	0.98	82.38
2	0.39	0.52	0.56	0.71	0.94	73.20

synthetic data takeaways

- Allow for modelling on the level of public-use data
  - More sophisticated than the usual remote submission system
- Allow for faster validation on confidential data
  - Faster than RDC proposal process
- Is limited in terms of analytic validity
  - But that may not be a bad thing
- Can accelerate disclosure avoidance process
  - All tables that are to be released can be created beforehand
  - In theory…

Access matrix for confidential data

Control of:	Data access	Access computers	Access rooms	Analysis methods	Disclosure avoidance
FSRDC researcher	Full	Full	Full (badge access)	Some (choice of software)	Manual/ variety of rules
Census employee	Full	None (VDI)	None (VDI)	Some (choice of software)	Manual/ self/ variety of rules
IAB: JoSuA researcher	Full	None (Web application)	None (Web application)	Smaller (software, whitelist commands)	Manual/ variety of rules
CASD researcher	Full	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)	Full	None (VDI) - Some (host institution)	None (VDI) - Some (host institution)	Some (choice of software)	Manual/ self/ variety of rules

Access matrix for confidential data

Control of:	Data access	Access computers	Access rooms	Analysis methods	Disclosure avoidance
FSRDC researcher	Full	Full	Full (badge access)	Some (choice of software)	Manual/ variety of rules
Census employee	Full	None (VDI)	None (VDI)	Some (choice of software)	Manual/ self/ variety of rules
IAB: JoSuA researcher	Full	None (Web application)	None (Web application)	Smaller (software, whitelist commands)	Manual/ variety of rules
CASD researcher	**Full**	**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)	Full	None (VDI) - Some (host institution)	None (VDI) - Some (host institution)	Some (choice of software)	Manual/ self/ variety of rules

trust and access

trust and access

- Frequent discussion
  - Security measures are for (malevolent) **intruders**/opponents
  - Researchers are **trusted** collaborators…
  - … who **know** what they are doing
- A corollary:
  - Protect against the bad guys
  - But let the "good" guys do their thing
- Examples:
  - Network-moderated access
  - Contracts with disclosure avoidance rules

How do you know who the good guys are?

Also known as the "old boys' network"

theory: culture

- Laws set the ground rules
  - The way penalties and contracts are set up are important
- Researchers and agencies create the communities in which these rules are applied and enforced
  - Training and “indoctrination”
  - Common forums
  - More or less tight binding of researchers into the community

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.
- 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**

penalties

- 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.

penalties

Note:

No system admits to ever having had to enforce the rules.

(rumors and videos notwithstanding)

training

hidden element: how is Disclosure Avoidance done?

- Most access methods:
  - Enforcing minimum count of entities in a statistic (coefficient, mean, stddev)
  - Prohibiting creation of tabular data (or making it very expensive)
  - (Vain) attempt at tracking overlapping releases
- Automated systems
  - Tracking of cells, implementation of (randomized) rounding, suppression, (output) noise infusion (StatCan, ABS)
  - Similar in CB's Microdata Analysis System/Automated Query System
- Newer mechanisms
  - Noise infusion upon computation
  - Differentially-private output perturbation (of model-based statistics, incl. coefficients and expected counts)

hidden cost: how to train the users?

- Programming
  - DP-safe programming is hard for computer scientists → lost cause with social scientists until incorporated into SAS, Stata, etc.
- Tools
  - Mostly lacking (in all of the environments that I have experienced)
- Concept
  - Researchers have a hard time understanding confidentiality constraints
  - Researchers have a hard time accepting confidentiality constraints

results from a survey of FSRDC users

- Survey run in October 2015, 145 respondents

FSRDC user experience: DA protocols

In order to obtain results from the analysis of restricted-access data, disclosure avoidance is applied, either to the analysis itself, or to the results from the analysis. Regarding your experience with disclosure avoidance protocols, please select the statement that best matches your experience prior to your connection with the FSRDC:

- ○ I or my team members had no prior experience with disclosure avoidance protocols
- ○ I or my team members had some experience with disclosure avoidance protocols
- ○ I or my team members are quite familiar with disclosure avoidance protocols

- 39% no prior experience
- 30% some experience
- 31% quite familiar

FSRDC user experience training

Please assess your agreement with the following statement: "After we applied for a FSRDC project, we were well informed about the disclosure avoidance protocols and process."

- ○ Strongly Disagree
- ○ Disagree
- ○ Neither Agree nor Disagree
- ○ Agree
- ○ Strongly Agree

- 14% disagree or strongly disagree
- 73% agree or strongly agree
- 13% do neither

results from a survey of FSRDC users

- Disclosure avoidance:
  - Users (NCHS) complained that “Disclosure avoidance personnel declined to approve output because they were not familiar with the software” despite pre-approval of generic output.
  - Other users grudgingly acknowledged that they “cannot avoid disclosure review” (on a NCHS project).

training content, method, and frequency

	Frequency	Access rules?	Disclosure rules?	Disclosure avoidance tools?	Method
FSRDC	yearly	Initial	As needed	No	Online
IAB RDC	Initial	Initial	Yes	No	PDF (Contract, other)
CASD (France)	Initial	Initial	Yes	No	In person (3h)
Denmark	Initial	Yes	No	No	PDF (Contract)

Responsible Conduct of Research (RCR)

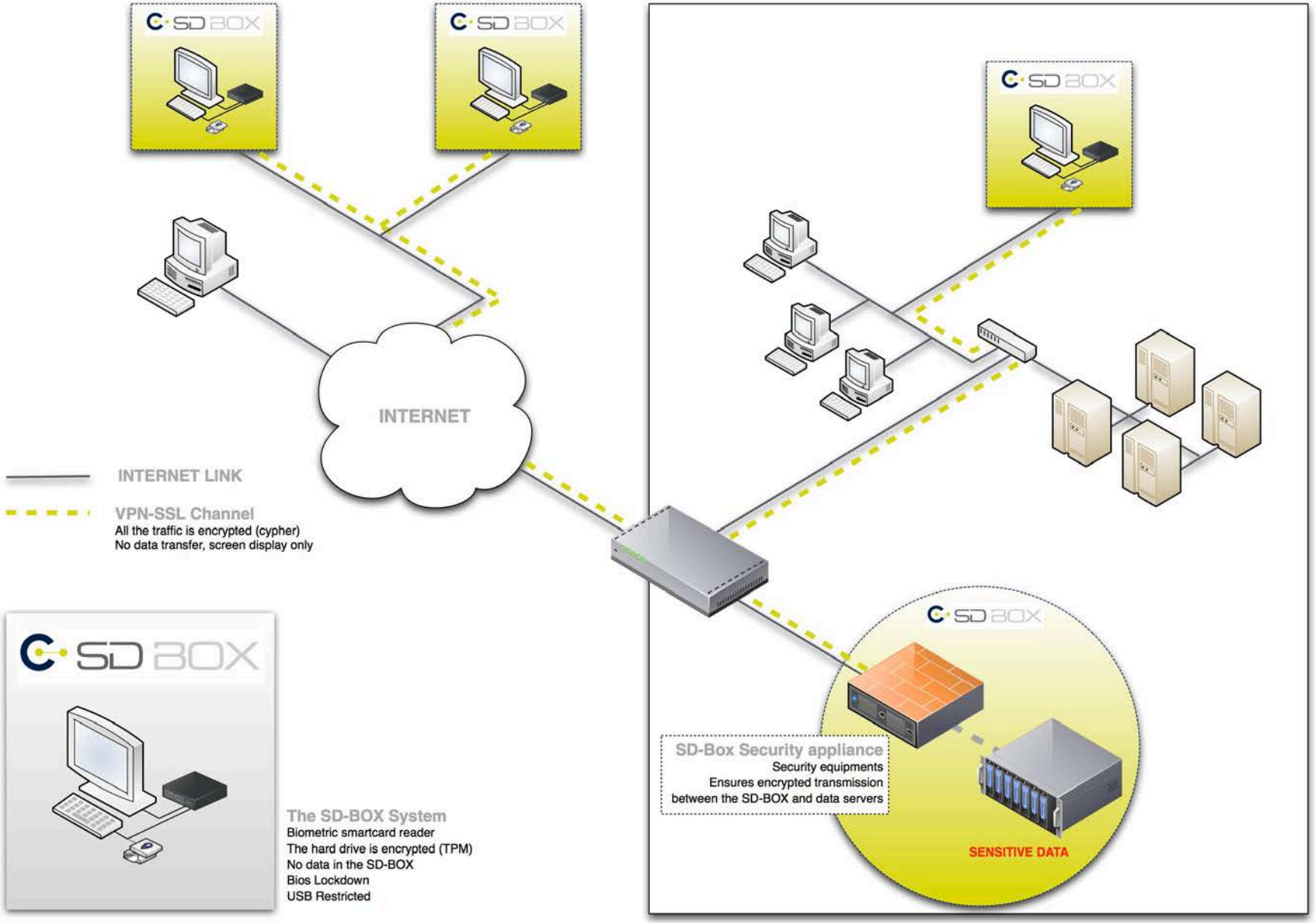
- Course Introduction
- Research Misconduct
- Research Involving Human Subjects
- Plagiarism
- Authorship
- Collaborative Research
- Conflicts of Interest
- Data Management
- Mentoring
- Peer Review

Approximately 3 hours

Note:
No discussion of practical disclosure avoidance, etc.

French training

- One-time initial training
- 3 (three) hours on-site classroom training
- Travel to Paris required
- First slide: **legal penalties**
- Quarter of slides: **disclosure rules**
  - Mostly cell-count rules
  - Mostly (numeric) examples of primary/secondary suppression
  - Examples of what confidential supplementary files should look like
- Half of slides: **technical system** with live demo

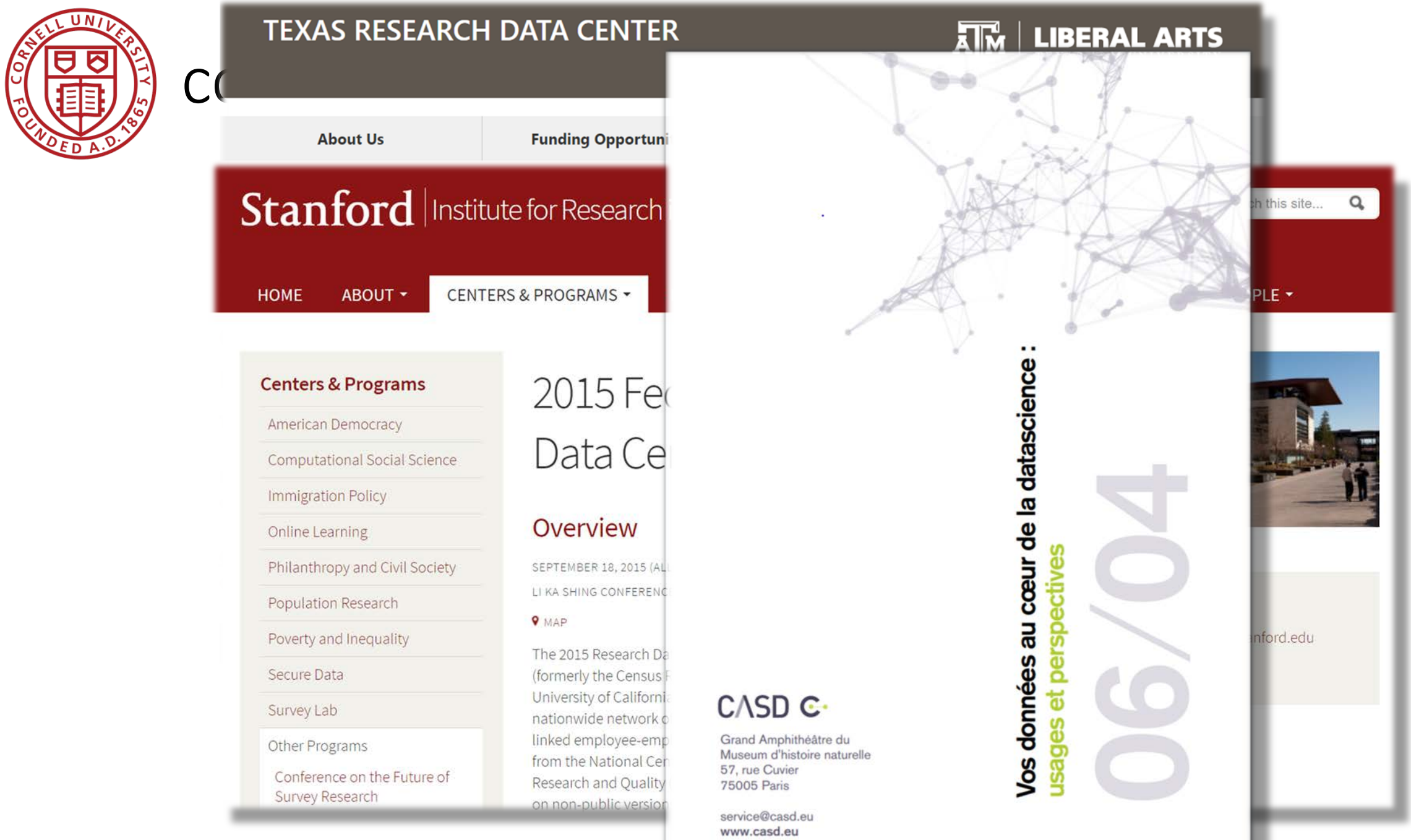


- Source: CASD training materials 2017-01-12

- Source: CASD training materials 2017-01-12

CASD security

- Four-factor authentication
  - Device
  - Card
  - Fingerprint
  - Password
- Loss of card (+fingerprint data)
  → trip back to Paris to create a new one


TEXAS RESEARCH DATA CENTER
LIBERAL ARTS
About Us
Stanford | Institute for Research
HOME
ABOUT
CENTERS & PROGRAMS
Centers & Programs
American Democracy
Computational Social Science
Immigration Policy
Online Learning
Philanthropy and Civil Society
Population Research
Poverty and Inequality
Secure Data
Survey Lab
Other Programs
Conference on the Future of Survey Research
Overview
MAP
Vos données au cœur de la datascience :
usages et perspectives
06/04
CASD
Grand Amphithéâtre du
Museum d'histoire naturelle
57, rue Cuvier
75005 Paris
service@casd.eu
www.casd.eu

community

- Census Bureau
  - CES seminars featuring FSRDC presenters
  - CES graduate mentorship
  - FSRDC conference
- CASD
  - Initial training plays a role
  - So far, only one conference, but very high-profile
    - Piketty video [URL]
    - Minister of State for Digital Affairs Axelle Lemaire

community (cont.)

- IAB
  - (past) visits to Nürnberg suggested and funded to connect with researchers
  - So far, one international conference (at Michigan, organized by Maggie Levenstein)
  - Overseas RDCs are a community-building exercise
  - Flooding the job market with economics graduates who have worked with German data…

training content, method, and frequency

	Frequency	Access rules?	Disclosure rules?	Method	User conference
FSRDC	yearly	General (+ site specific)	General	Online	Yearly
IAB RDC	Initial	No	Yes	PDF (Contract, other)	Irregular
CASD (France)	Initial	Yes	Yes	In person (3h)	2016
Denmark	Initial	Yes	No	PDF (Contract)	?

summary

summary

- **Remote** access of some type is the **standard** practice around the world
- **Access locations** and ease of releasing results vary substantially
- Disclosure avoidance process is still **quite pedestrian** in almost all cases, and DA methods are "**old-fashioned**"
- **Remote submission methods** remain quite limiting
- Newer access mechanisms (synthetic data) successfully combine ability to estimate **arbitrary models** with **robust (provable) protection mechanisms**, but remain at an early stage

access and trust

- Legal obligations matter
  - Criminal vs. contractual obligations
  - Obligating the institutions more strongly may help relax other constraints
- Community matters
  - Pulling researchers close to the statistical agency through training
  - Creating a community through conferences, mentoring, etc.

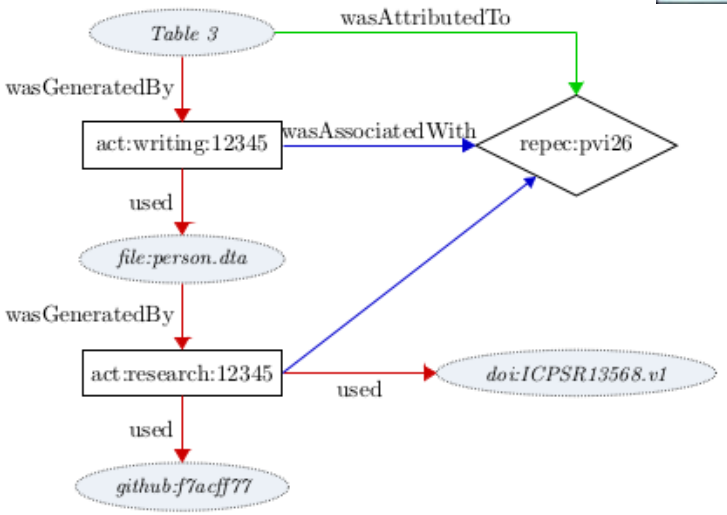
the ultimate achievement

What kind of community, training, legal environment would the FedStat system need to implement to allow researches to access confidential data the same way Census employees do?

p.s. one last thing

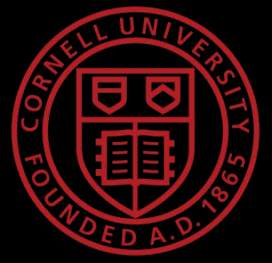
- **Replicability** is a nascent problem
  - More and more journals require provable replicability
  - Cannot be satisfied with **idiosyncratic** access mechanisms
  - Some research with confidential files will **lose** (reputable) publication outlets

- Transparency critical
  - Need capability to be able to **archive** research files within secure enclaves
  - Need ability to **publically identify** such files (documentation) *[DDI, DOI]*



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)

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