Disclosure limitation and confidentiality protection in linked data

Lars Vilhuber

Based on joint work with John M. Abowd and Ian M. Schmutte

administrative data in the infrastructure of official statistics

- Back to the 1960s – frames, if not surveys
- Motivation behind Fellegi's original work
- Today not just frame but data source
  - European censuses based on administrative data
  - US business registers used for Business Dynamics Statistics (BDS), County Business Patterns
- New sources emerging (health, education, law)

rich new analysis and publications

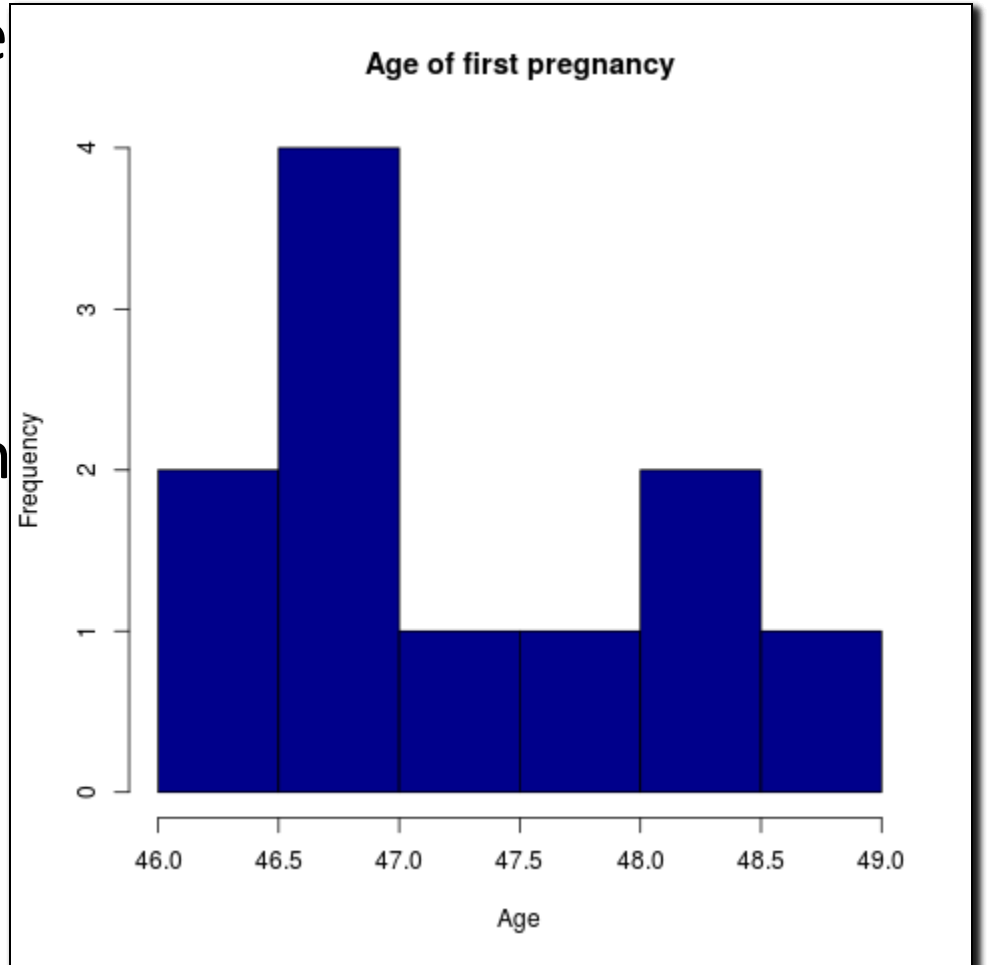
held back by concerns of citizens and businesses about privacy

privacy concerns

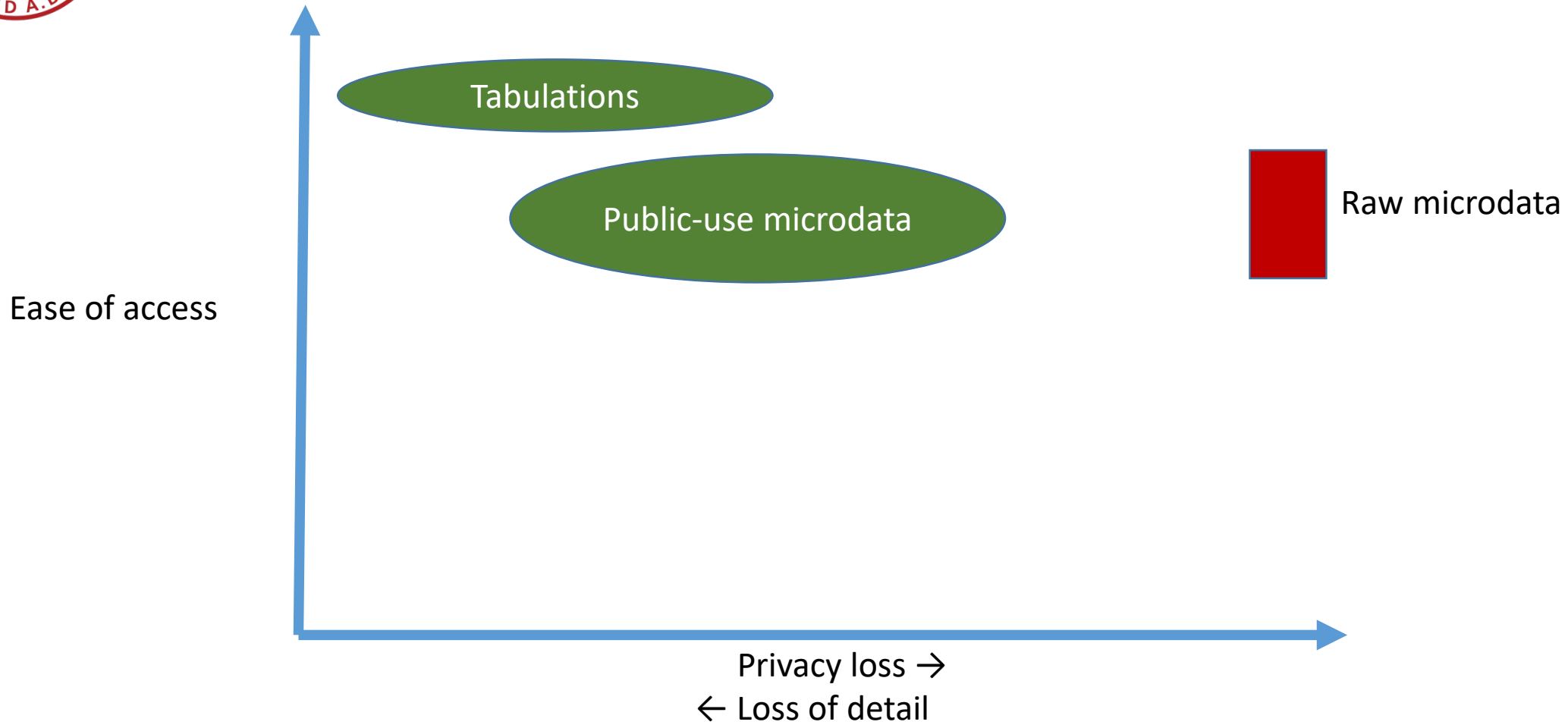
- 1960s in the US: proposal for “National Data Bank” with the goal of combining survey and administrative data to make available to researchers
  - Instead, and partially as a consequence, privacy laws were formalized in the 1970s (“Privacy Act 1974” (Public Law 93-579, 5 U.S.C. § 552a)) specifically 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

current state of protection mechanisms

- Public-use files and tabulations, created using techniques developed for survey file
  - Suppression
  - Coarsening
  - swapping
  - noisy queries (input and output)
- Limited utility for “thin tails” (or thin
  - Business data
  - But more generally “rare data”



access methods



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 tabular data (synthetic data, differentially-private data)

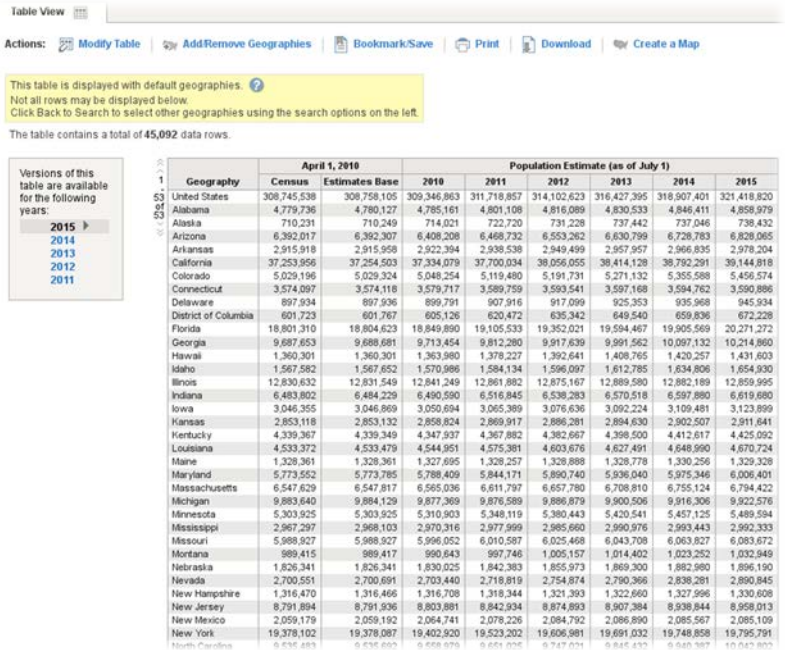


Table View

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.

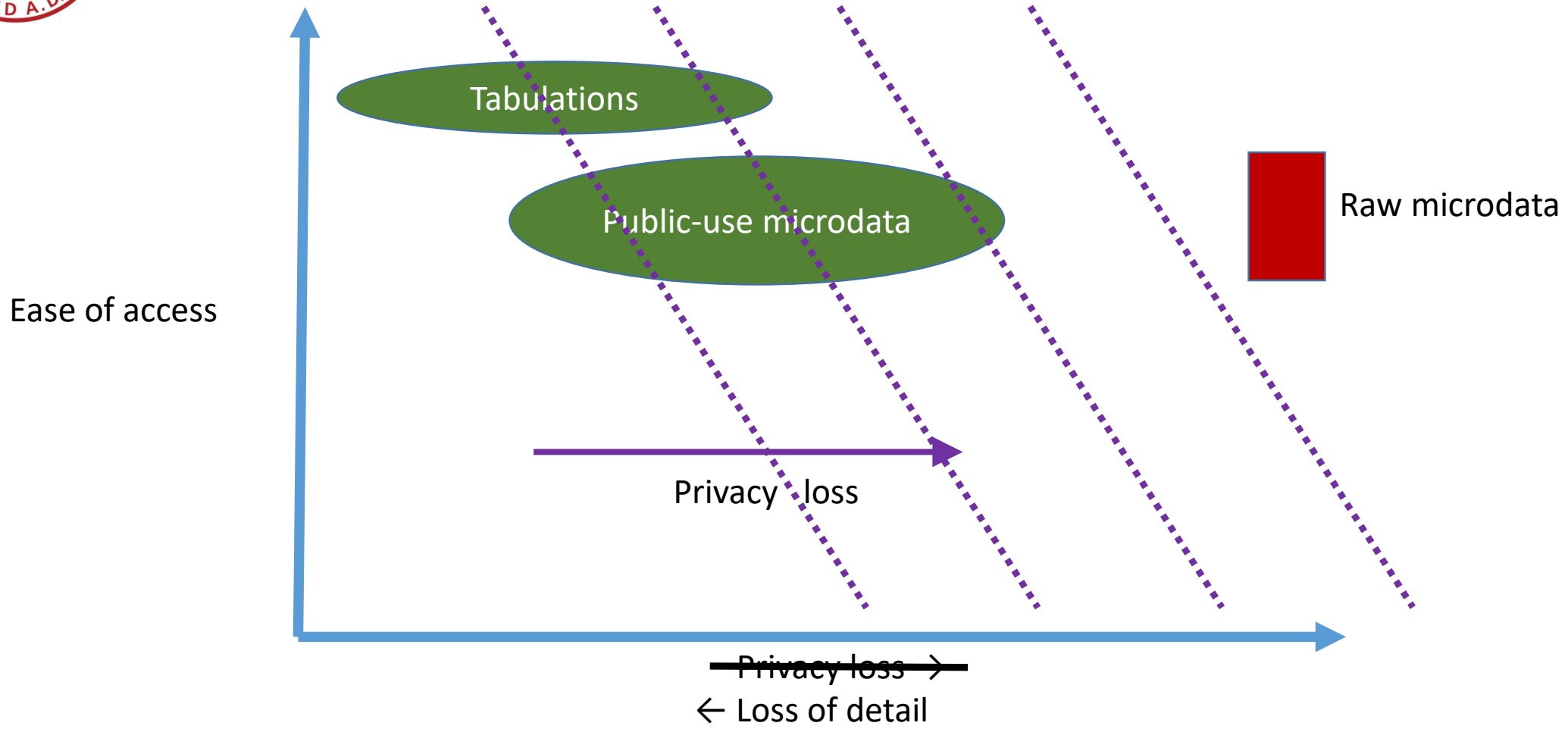
Versions of this table are available for the following years:
2015
2014
2013
2012
2011

Geography	April 1, 2010		Population Estimate (as of July 1)					
	Census	Estimates Base	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

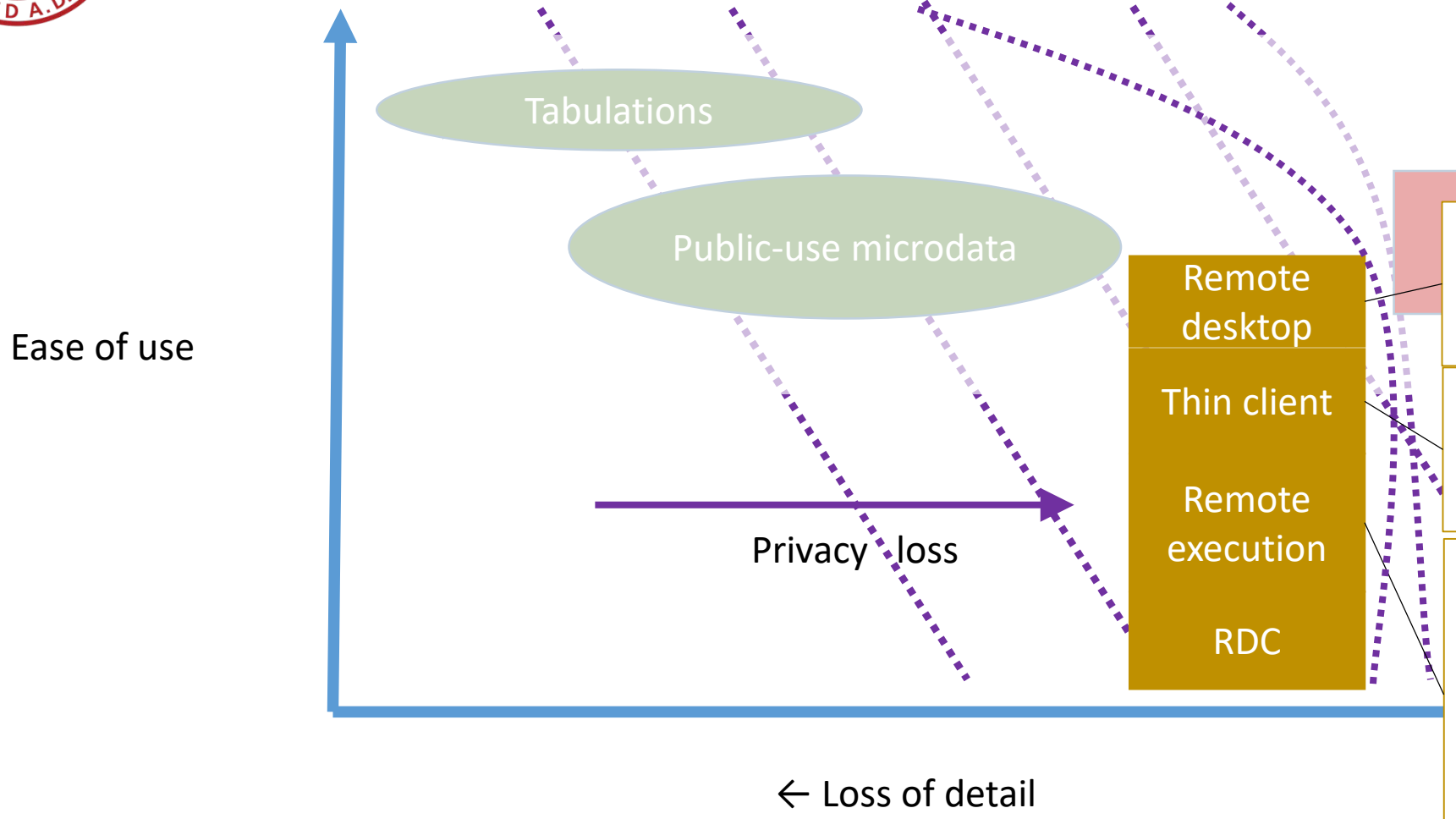
newer methods: **Data Enclaves**

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

access methods



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

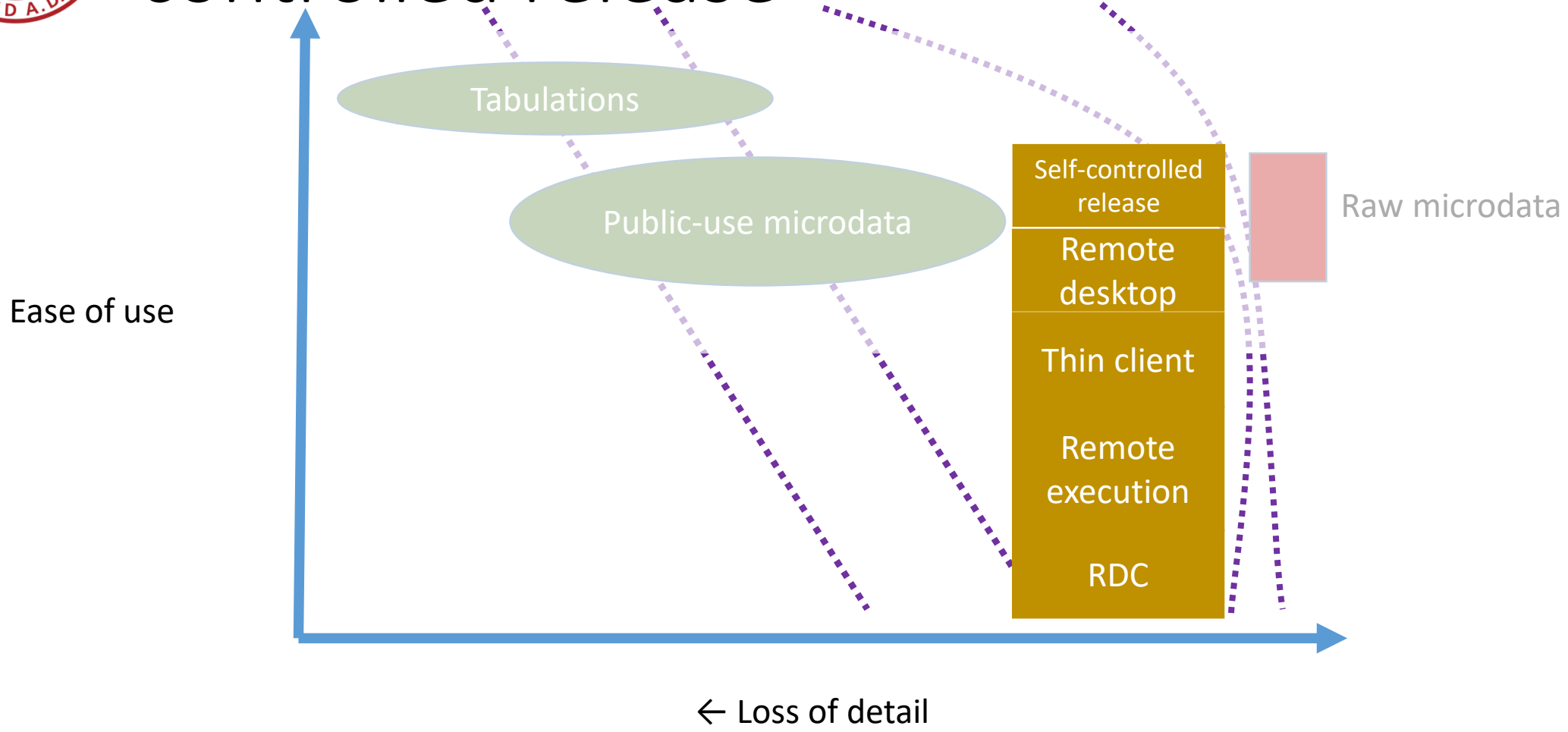
Submitting analysis programs by email or through website, with manual disclosure avoidance (DA)

(possibly combined with synthetic microdata)

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]

access methods: enclaves with researcher-controlled release



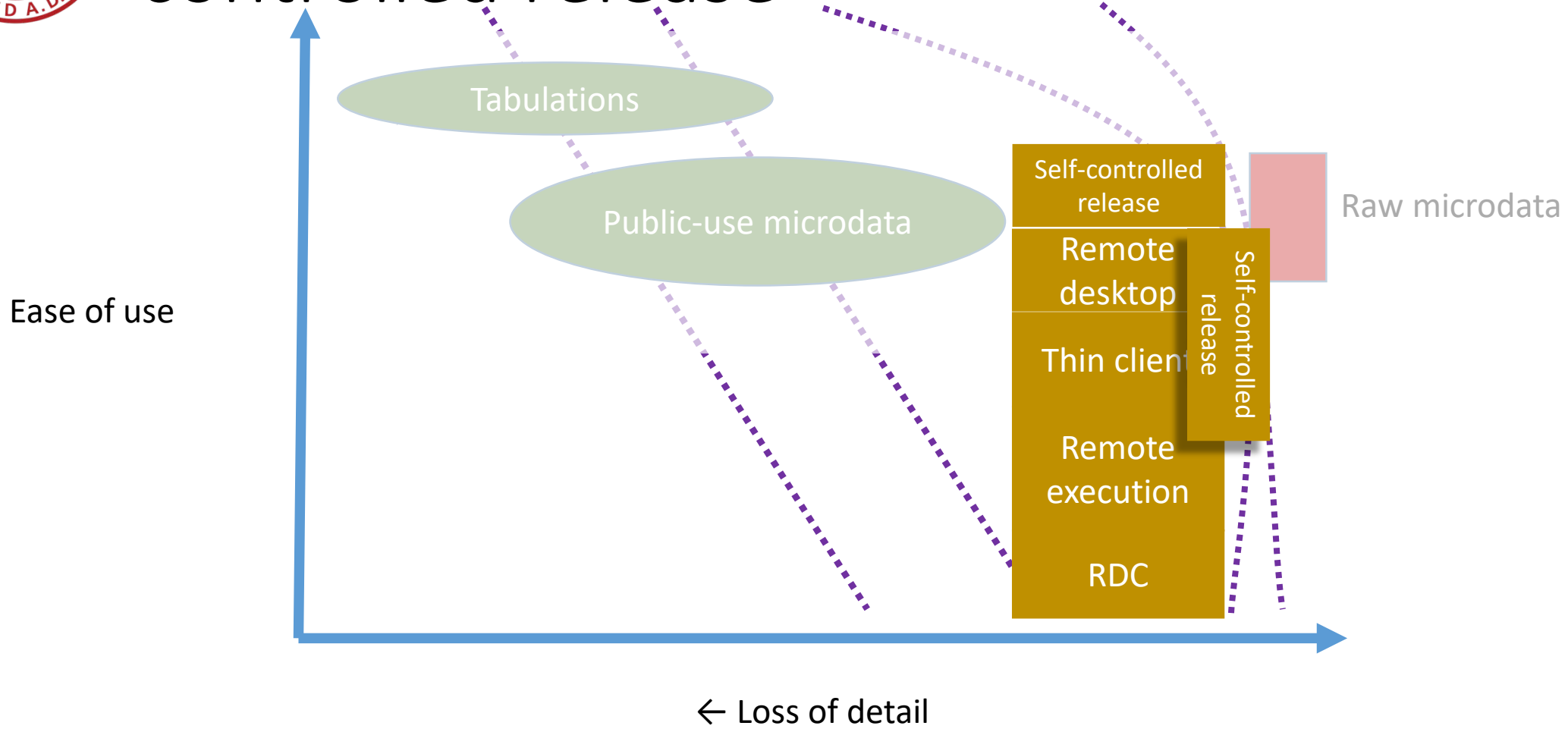
trust and access

- 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
  - Danish remote access with **researcher-controlled release** of results and **authorized establishments**

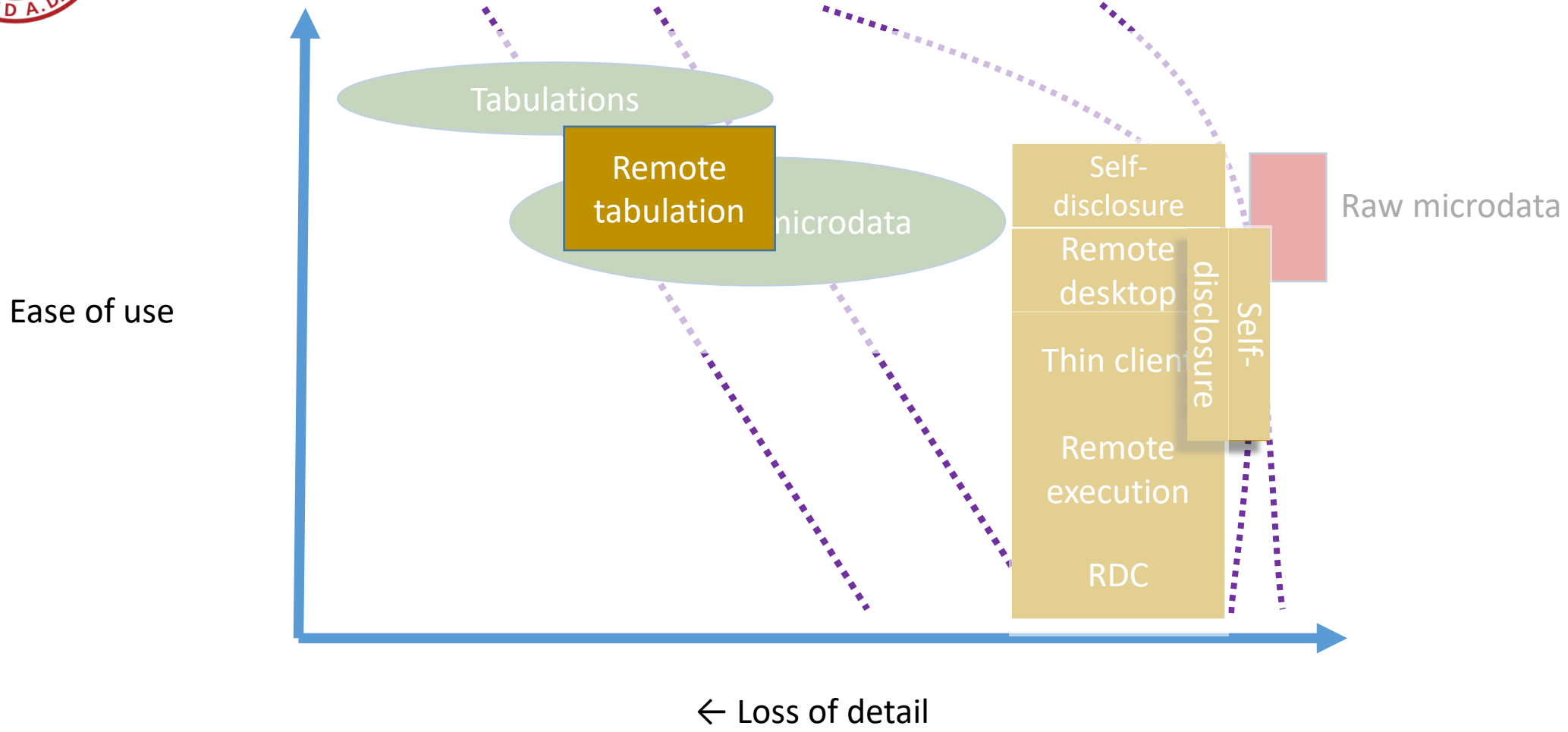
How do you know who the good guys are?

Also known as the "old boys' network"

access methods: enclaves with researcher-controlled release



access methods: remote tabulation



hidden element: how is Disclosure Avoidance done?

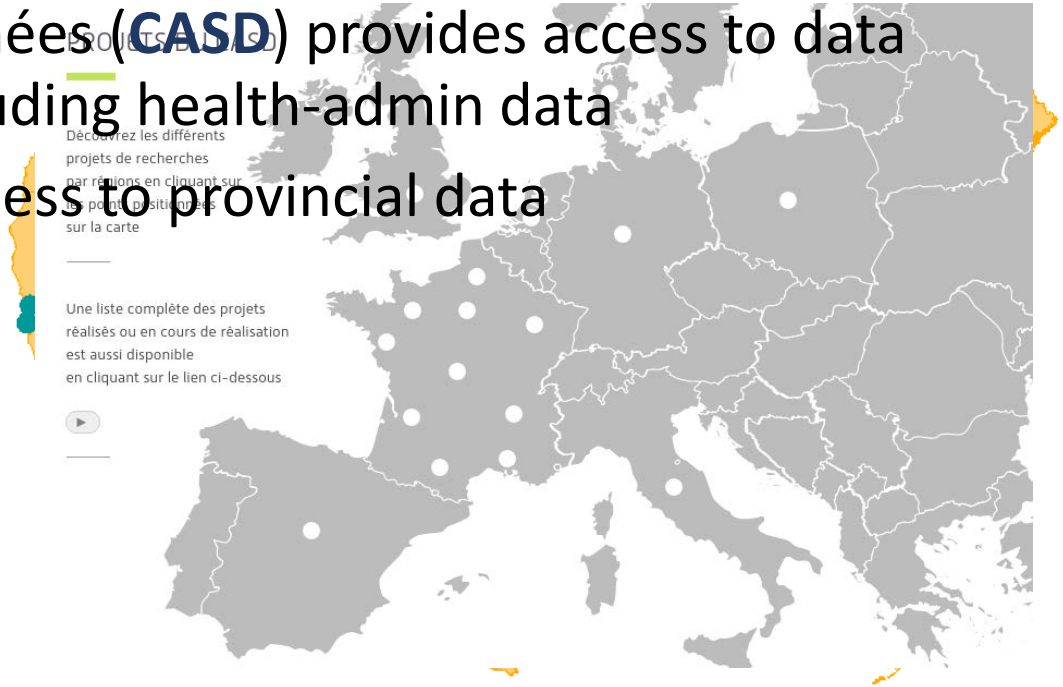
- Most access methods:
  - Enforcing minimum count of entities in a statistic (coefficient, mean, stddev)
  - Prohibiting creation of tabular data (or making it very expensive)
  - (Vain) attempt at tracking overlapping releases
- Automated systems
  - Tracking of cells, implementation of (randomized) 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: managing the network

- Categorizing access requests
  - What is an **authorized** institution? [Eurostat: recognized research entity]
  - Framework **contracts** with each institution
- Managing access **requests**
  - Ideally, that's a problem you would like to have
  - **Training** [France: 3h enrollment sessions], managing **access tokens**, [physical thin clients]
  - Cost? [France: itemized price list]
- Disclosure avoidance
  - Bottleneck – ideally primary function of designated staff
  - Training of users important [Based on our survey of 100 US FSRDC users]
  - Provision of tools important
  - **Enforcement** (in particular for self-disclosure)

natural economies of scale

- Increasing emphasis on consolidation of (national) networks
  - **US** Census Bureau's Research Data Centers (since 1990s) now Federal Statistical Research Data Centers (**FSRDC**) with 8 new federal partners
  - **France** Centre d'accès sécurisé aux données (**CASD**) provides access to data from more than 15 national entities, including health-admin data
  - Some **Canadian** RDCs also providing access to provincial data
- Secondary research benefit
  - Ability to **break out of data silos**



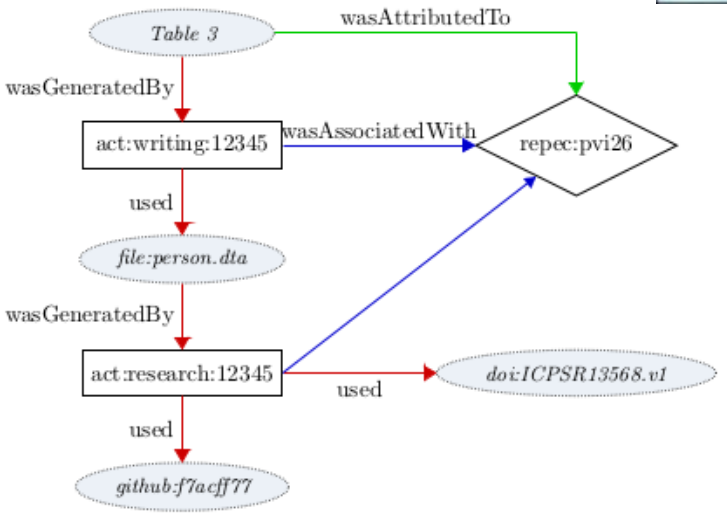
summary

- Many more tools today than in the past
- **Remote** access of some type is the **standard** practice around the world
- Disclosure avoidance is still quite pedestrian in almost all cases
- Newer methods are being developed, but few access mechanisms (proposed or implemented) successfully combine ability to estimate **arbitrary models** with **robust (provable) protection mechanisms**

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