Are Locations Unique?

Spatial Statistics, Differential Privacy and the 2020 Census

Lee Fiorio, UW Geography

Neal Marquez, UW Sociology

October 10th, 2019

CSDE Computational Demography Working Group

Goals of today’s presentation

- Describe the concerns around privacy and the 2020 Census
    - What is a database reconstruction attack?
- Explain differential privacy *in theory*
- Discuss differential privacy *in practice*
    - How will the Census implementation of differential privacy deal with geography?
    - What are the implications on different kinds of geography accuracy?

Credit Where Credit is Due

This presentation synthesizes two existing presentations, one by John Abowd at Census and the other by Dave Van Riper at IPUMS – please check out these out:

- Abowd, John. (2019). “Stepping-up: The Census Bureau Tries to Be a Good Data Steward in the 21st Century”
- Van Riper, David, Tracey Kugler, José Pacas, & Jonathan Schroeder (2019). “Differential privacy and the decennial census”

Other valuable references:

- Garfinkel, Simson L., John M. Abowd, and Christian Martindale. (2018)."Understanding database reconstruction attacks on public data.”
- Reiter, Jerome (2019). Differential Privacy and Federal Data Releases. *Annual review of statistics and its application*, *6*, 85-101.
- Griffith, David (1984). Reexamining the Question ‘are Locations Unique?’. *Progress in Geography*, *8*(1), 82-94.

Privacy and the 2020 Census

The challenges of a census (Abowd, 2019)

1. Collect all of the data necessary to underpin our democracy

2. Protect the privacy of individual data to ensure trust and prevent abuse

The Database Reconstruction Vulnerability (Abowd, 2019; Garfinkel et al, 2018)

The argument:

- Census Bureau publishes too many statistics
  - Tables (i.e. queries) can be combined and then used to reconstruct underlying confidential microdata → "database reconstruction attack"

- Thus, noise infusion is necessary

- And as such, transparency about noise infusion methods is a benefit

What is a Database Reconstruction Attack? (Garfinkel et al, 2018)

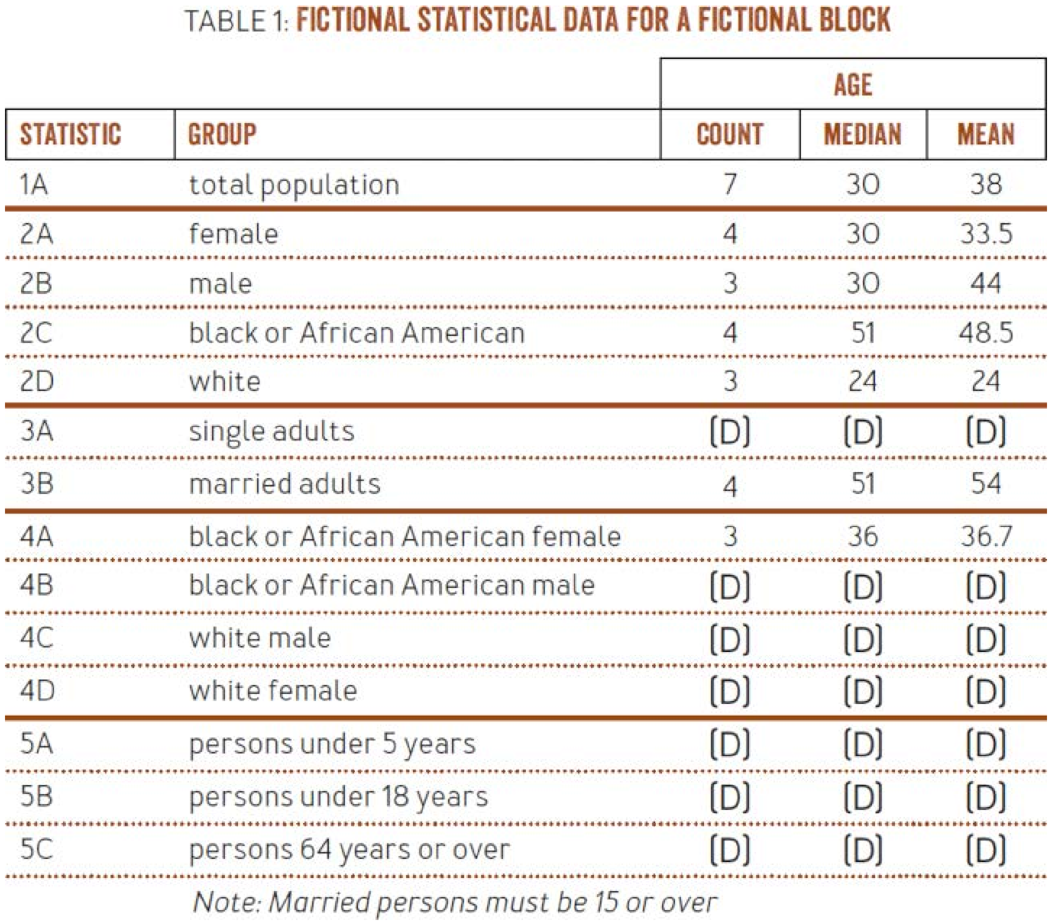


TABLE 1: **FICTIONAL STATISTICAL DATA FOR A FICTIONAL BLOCK**

		AGE		
STATISTIC	**GROUP**	**COUNT**	**MEDIAN**	**MEAN**
1A	total population	7	30	38
2A	female	4	30	33.5
2B	male	3	30	44
2C	black or African American	4	51	48.5
2D	white	3	24	24
3A	single adults	(D)	(D)	(D)
3B	married adults	4	51	54
4A	black or African American female	3	36	36.7
4B	black or African American male	(D)	(D)	(D)
4C	white male	(D)	(D)	(D)
4D	white female	(D)	(D)	(D)
5A	persons under 5 years	(D)	(D)	(D)
5B	persons under 18 years	(D)	(D)	(D)
5C	persons 64 years or over	(D)	(D)	(D)

Note: Married persons must be 15 or over

- Use reported statistics
  - Count
  - Median
  - Mean
  - Others
- Calculate constraints
- Solve for unique solution (i.e. solve for confidential microdata)

What has the Census Bureau done in the past? (Garfinkel et al, 2018)

- Cell suppression
- Complimentary cell suppression
- Top-coding
- Noise-injection
- Swapping

All together these techniques are known at Census as statistical disclosure limitation (SDL)

But it may not be enough (Abowd, 2019)

In a recent experiment using the 2010 Census, the Census Bureau did the following:

- Database reconstruction for all 308,745,538 people the census
- Link reconstructed records to commercial databases: acquire PII
- Successful linkage to commercial data: putative re-identification
- Compare putative re-identifications to confidential data
- Successful linkage to confidential data: confirmed re-identification
- *Harm: attacker can learn self-response race and ethnicity*

What the Census Bureau found (Abowd, 2019)

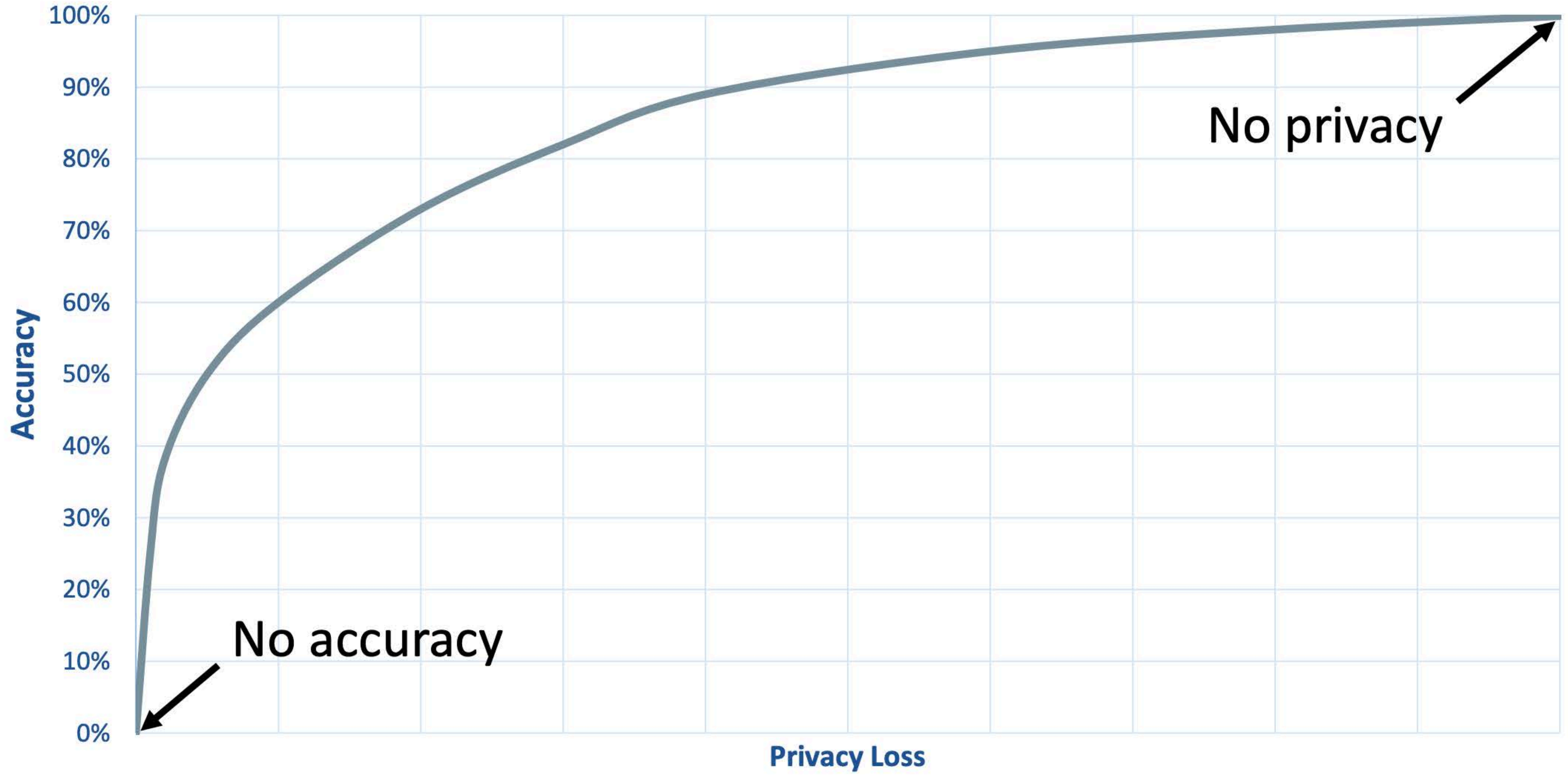
- Census block and voting age (18+) correctly reconstructed in all 6,207,027 inhabited blocks
- Block, sex, age (in years), race (OMB 63 categories), ethnicity reconstructed
  - Exactly: 46% of population (142 million of 308,745,538)
  - Allowing age +/- one year: 71% of population (219 million of 308,745,538)
- Block, sex, age linked to commercial data to acquire PII • Putative re-identifications:
  - 45% of population (138 million of 308,745,538)
- Name, block, sex, age, race, ethnicity compared to confidential data
  - Confirmed re-identifications: 38% of putative (52 million; 17% of population)
  - For the confirmed re-identifications, race and ethnicity are learned exactly, not just statistically

In the words of Abowd (2019)…

“Absolutely the hardest lesson in modern data science is the constraint on publication that the fundamental law of information recovery imposes.”

“I usually call it the death knell for traditional methods of publication, and not just in statistical agencies.”

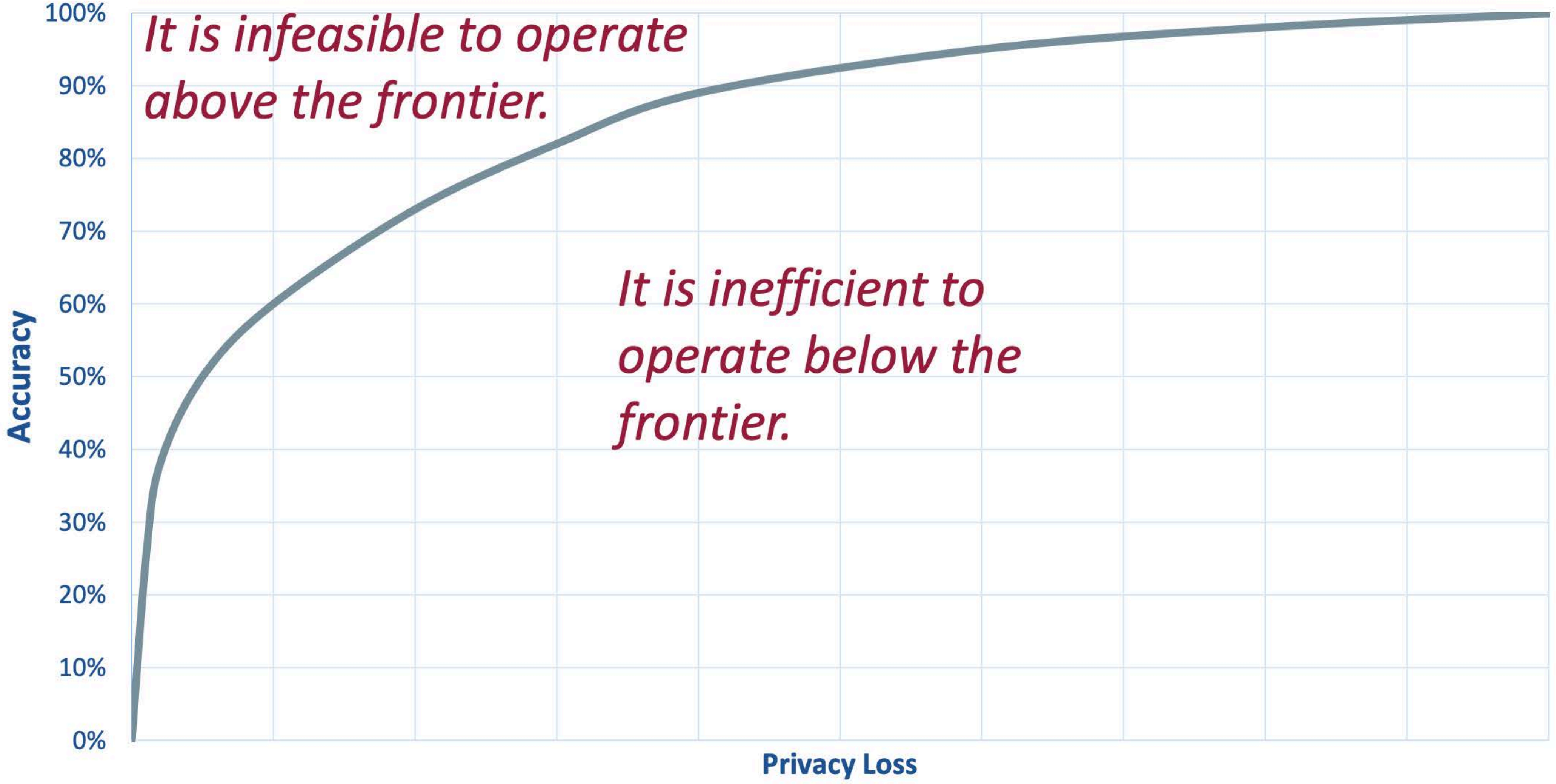
Fundamental Tradeoff betweeen Accuracy and Privacy Loss



No privacy

No accuracy

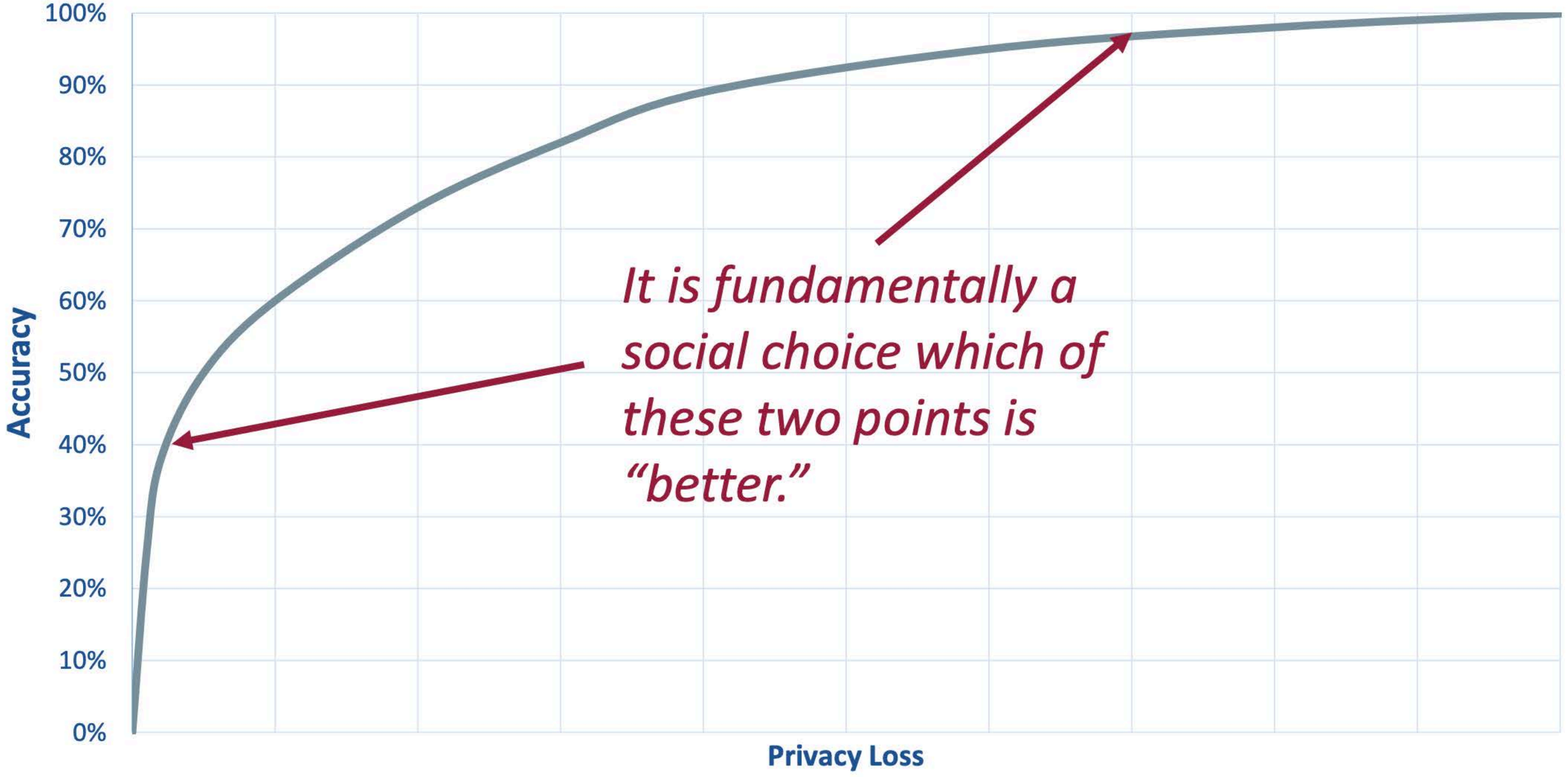
Fundamental Tradeoff betweeen Accuracy and Privacy Loss



It is infeasible to operate above the frontier.

It is inefficient to operate below the frontier.

Fundamental Tradeoff betweeen Accuracy and Privacy Loss



It is fundamentally a social choice which of these two points is "better."

Defining Differential Privacy

What is Differential Privacy? (Reiter, 2018)

The “ε” in “ε-DP”

Differential privacy is a *criterion* for privacy protection

$$\Pr[M(D) \in S\,] \leq e^{\varepsilon} \cdot \Pr[M(D') \in S\,] + \delta$$

A guarantee “on the incremental disclosure risks of participating in D over whatever disclosure risks the data subjects face even if they do not participate in D…

“When attackers cannot tell whether an output derives from **D** or any neighboring database, they cannot tell whether any particular individual participated in **D** or not. Thus, data subjects can be assured that the choice to participate or not does not affect their privacy.”

Differential Privacy is NOT (Van Riper et al, 2019):

- An algorithm for disclosure control
- An absolute guarantee against disclosure risk

“True” microdata

Sex **School**

Male | Never
Male | Never
Male | Never

x12
Male | Attending
Male | Attending
⋮
Male | Attending

x33
Male | Past
⋮
Male | Past

Sex **School**

x4
Female | Never
⋮
Female | Never

x17
Female | Attending
⋮
Female | Attending

x31
Female | Past
⋮
Female | Past

Source: Van Riper et al (2019)

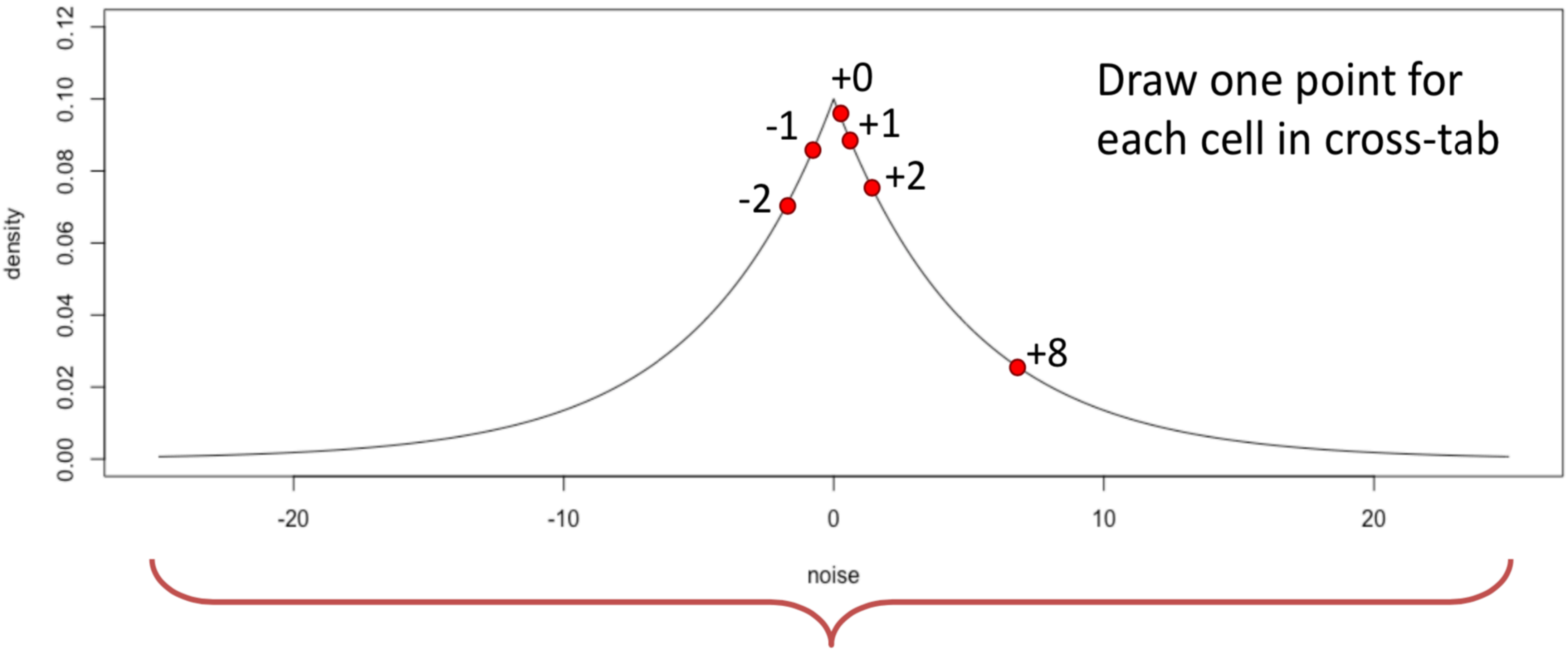
Construct cross-tabs from “true” data

	School Attendance		
	Never	Attending	Past
Male	3	12	33
Female	4	17	31

Source: Van Riper et al (2019)

Population = 100

Draw noise from Laplace distribution



Draw one point for each cell in cross-tab

spread is determined by **ε**

Source: Van Riper et al (2019)

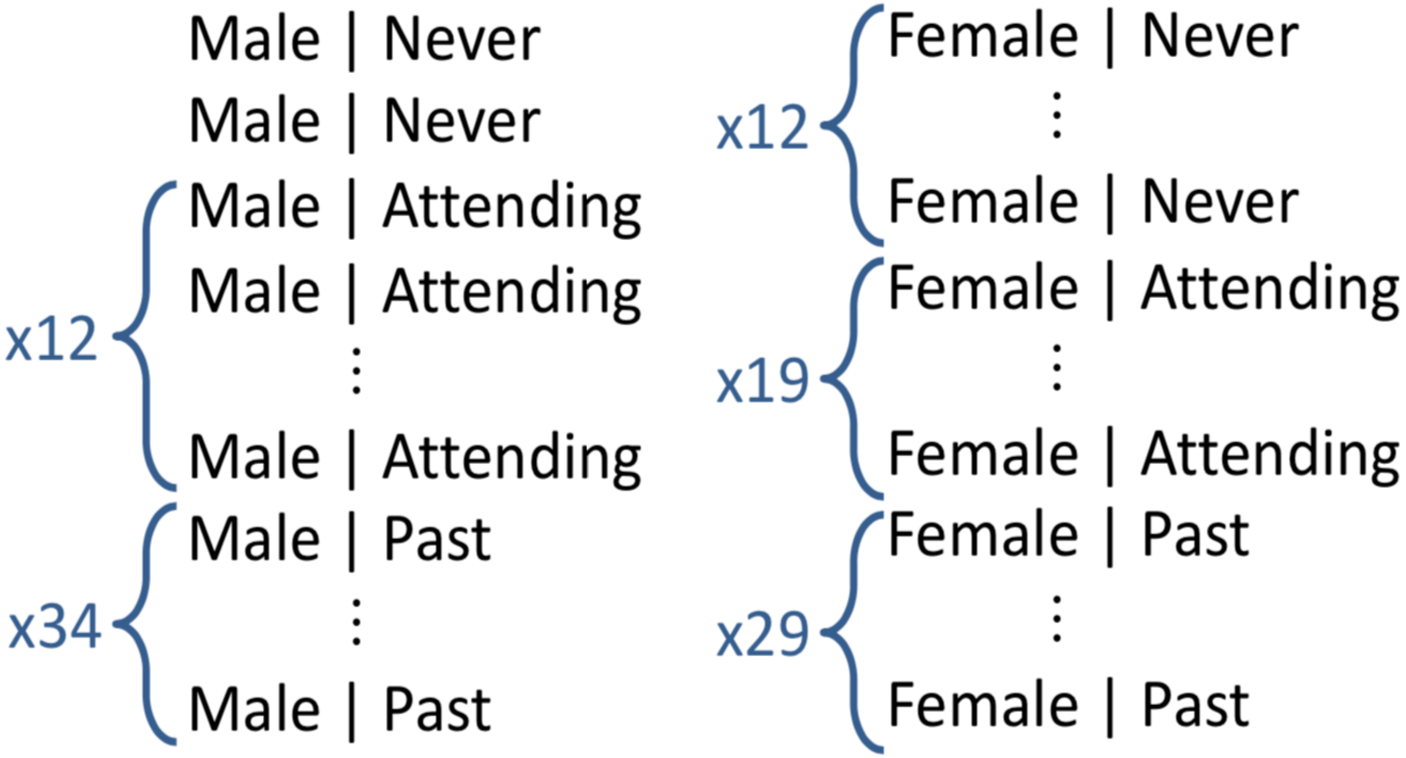
Add noise to cross-tab

	School Attendance		
	Never	Attending	Past
Male	3 - 1 = **2**	12 + 0 = **12**	33 + 1 = **34**
Female	4 + 8 = **12**	17 + 2 = **19**	31 - 2 = **29**

Sum = 108

Source: Van Riper et al (2019)

Construct synthetic microdata



Male | Never

Male | Never

x12

Male | Attending

Male | Attending

⋮

Male | Attending

x34

Male | Past

⋮

Male | Past

x12

Female | Never

⋮

Female | Never

x19

Female | Attending

⋮

Female | Attending

x29

Female | Past

⋮

Female | Past

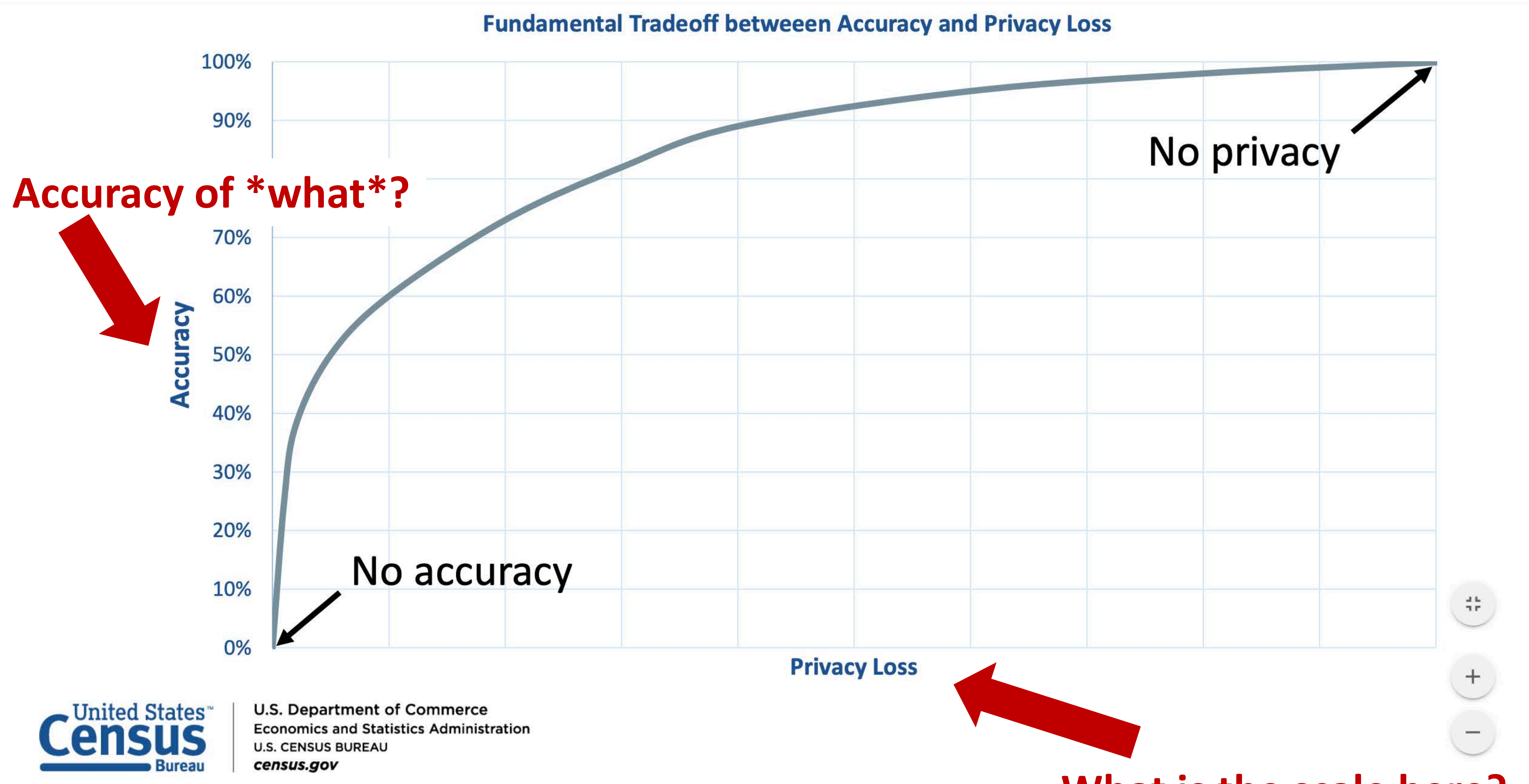
Source: Van Riper et al (2019)

(True) vs. Noise infused table

	School Attendance		
	Never	Attending	Past
Male	(3) 2	(12) 12	(33) 34
Female	(4) 12	(17) 19	(31) 29

Source: Van Riper et al (2019)

Differential Privacy *in Practice*



Fundamental Tradeoff betweeen Accuracy and Privacy Loss

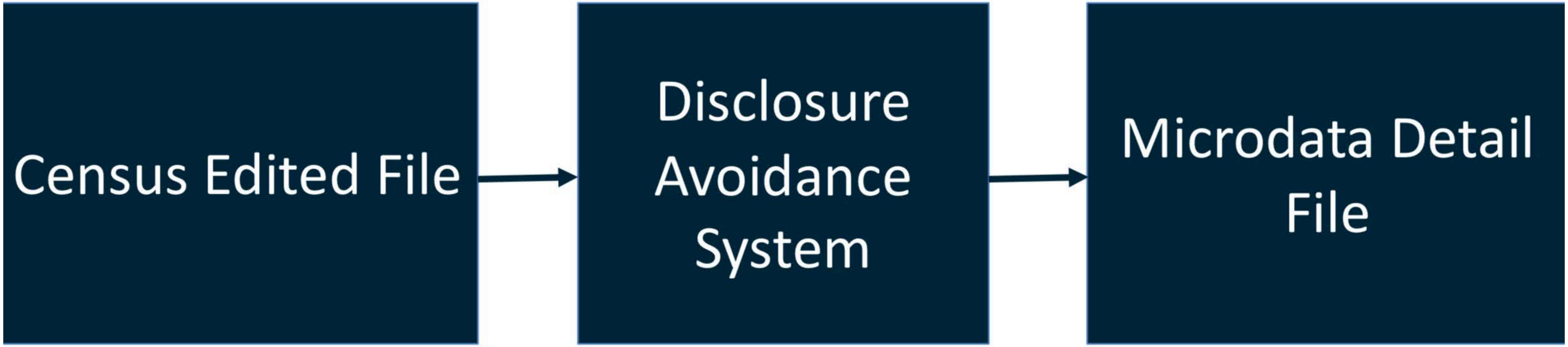
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

No privacy

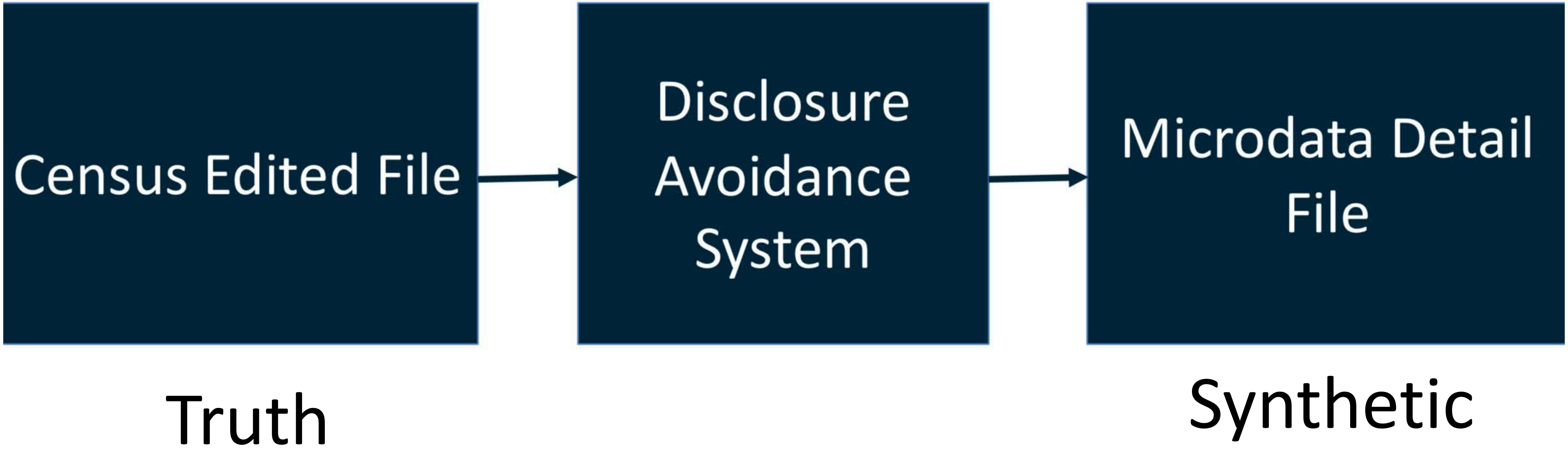
No accuracy

Accuracy of *what*?

What is the scale here?



Source: Van Riper et al (2019)



Census Edited File

Disclosure Avoidance System

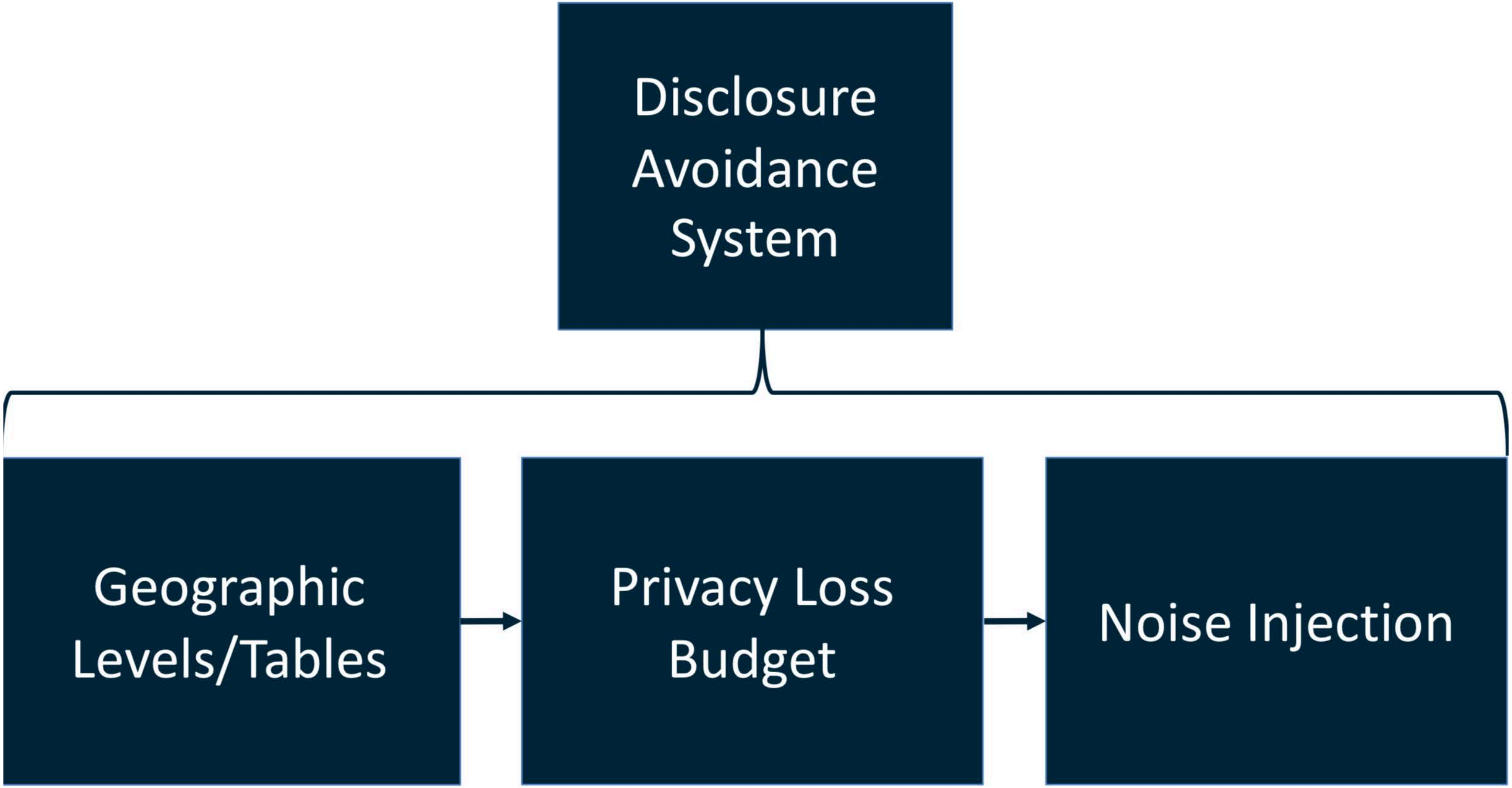
Microdata Detail File

Truth

Synthetic

Source: Van Riper et al (2019)

Source: Van Riper et al (2019)



Disclosure Avoidance System

Geographic Levels/Tables

Privacy Loss Budget

Noise Injection

Policy Decisions (Van Riper et al, 2019)

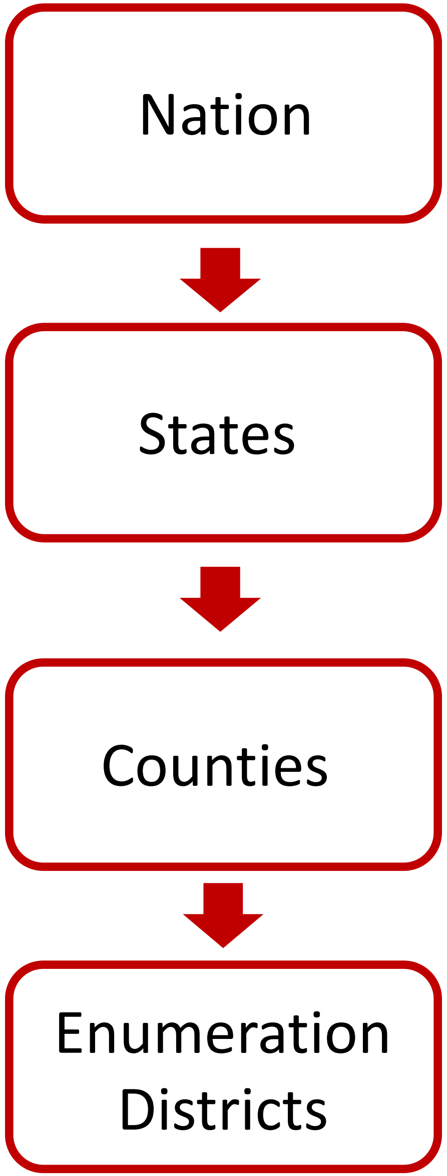
- Global privacy loss budget (ε)
- Geographic levels
  - Fraction of privacy budget allocated to each level
- Tables
  - Fraction of privacy budget allocated to each table

”Top Down” Implementation of DP (Van Riper et al, 2019)

- Two Steps
  1. Generate microdata without geographic identifiers
  2. Assign geographic identifiers to each microdata record
- Step 1:
  - Create national table from “true” data
  - For each cell in table, infuse noise drawn from Laplace distribution
  - Generate microdata with no geographic identifiers from (2) via database reconstruction
- Step 2:
  - Create state table from “true” data
  - For each cell in table, infuse noise drawn from Laplace distribution
  - Use optimization to fit Step 1 microdata to “noisy” state cells
  - Assign state identifier to each Step 1 microdata record
  - Repeat (1) – (4) for remaining geographic levels (counties and enumeration districts)

”Top Down” Implementation of DP (Van Riper et al, 2019)

- Microdata records with state, county, and enumeration district identifiers
- Process does not change characteristics of individual microdata records generated in Step 1
  - But it assigns geographic identifiers so final micro data are “best fitting” to the ”noisy” cross tabs for enumeration districts


Nation

States

Counties

Enumeration Districts

Are Locations Unique?

Thinking about differential privacy and ***geographic*** *accuracy*

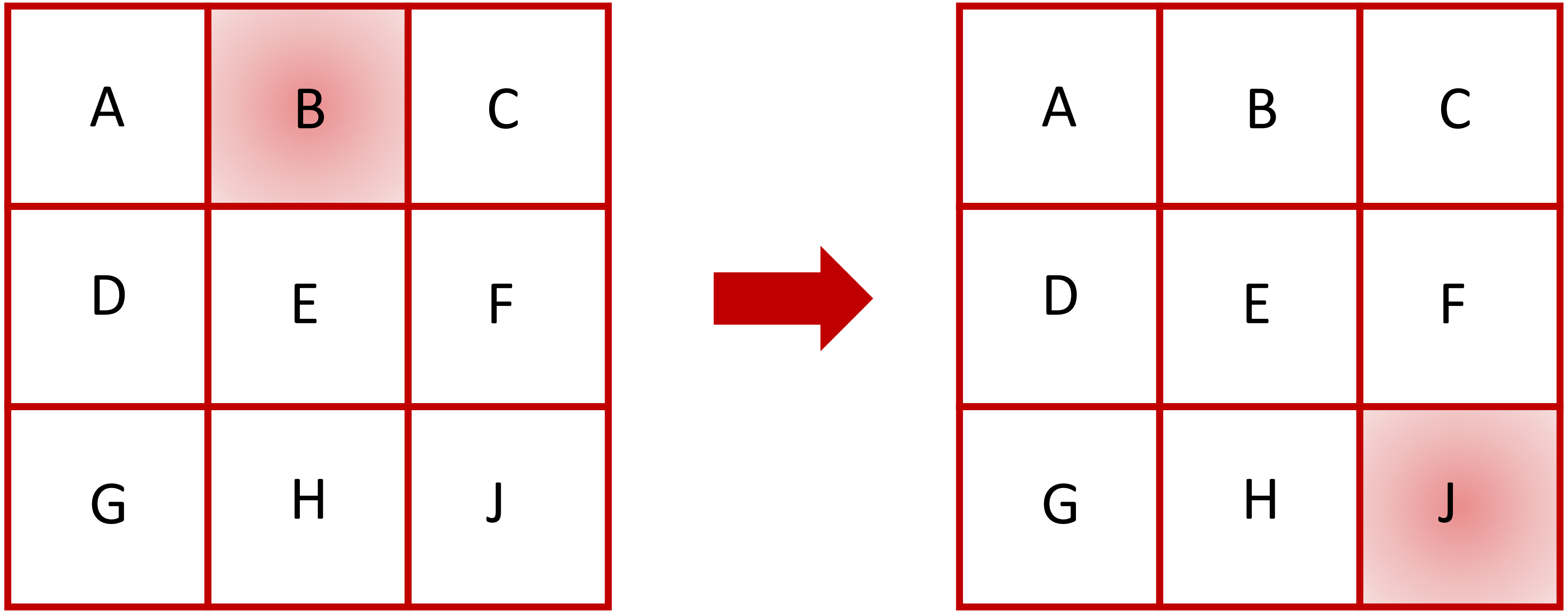
Are locations unique?

- YES
  - Areal units are nothing more than containers of phenomena
  - Space is absolute and its measures Euclidian
  - Idiographic

- NO
  - Areal units are related through juxtaposition of their phenomena
  - Space is relative
  - Nomothetic

So, if locations *ARE* unique, what kinds of accuracy are desired?

Observing phenomena in the ***correct container***



A	B	C
D	E	F
G	H	J

A	B	C
D	E	F
G	H	J

What are the potential implications?

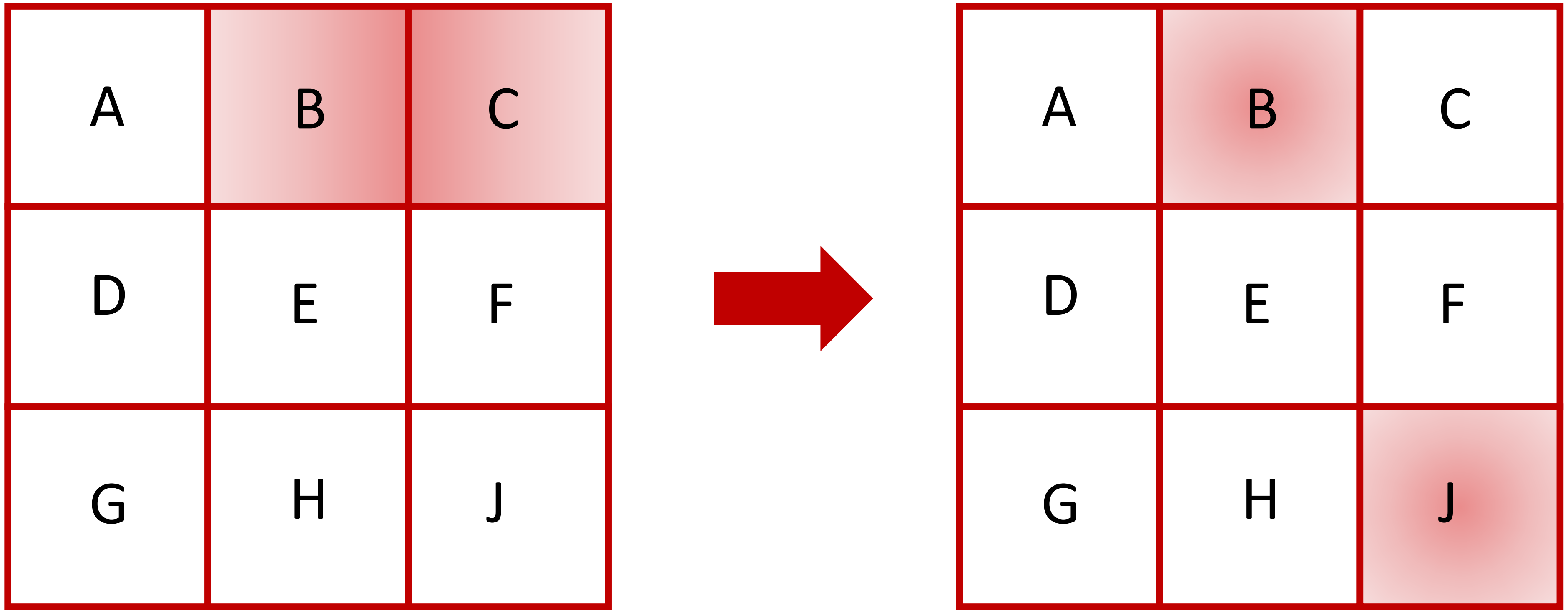
- Political:
  - “Communities of Interests” may become difficult to observe
  - From Justin Levitt’s All About Redistricting website:
    - Preserving "communities of interest" is a redistricting mandate in 24 states
    - A "community of interest" is “a group of people with a common interest (usually, a common interest that legislation might benefit)” -- but often has an idiographic geographical meaning (emphasis mine):
      - “Kansas' 2002 guidelines offered a fairly typical definition: ‘[s]ocial, cultural, racial, ethnic, and economic interests ***common to the population of the area***, which are probable subjects of legislation.’”
      - “Alabama adds the helpful reminder that ‘[i]t is inevitable that some interests will be recognized and others will not, [but] the legislature will attempt to accommodate those felt most strongly by the people ***in each specific location.’”***

What are the potential implications?

- Analytical:
  - Geographic linkages may be compromised
    - Measures of access which rely on **distance**
      - E.g. how far to the closest hospital?
    - Measures of **exposure** which rely on confluence
      - E.g. was the population in a given tract exposed to violent crime?
  - How can we measure change over time?
    - Comparing population in a unique neighborhood over time

So, if locations *ARE NOT* unique, what kinds of accuracy are desired?

Observing the ***correct juxtaposition*** of phenomena



A	B	C
D	E	F
G	H	J

A	B	C
D	E	F
G	H	J

What are the potential implications?

- Analytical/political
    - ‘True’ spatial structure lost
    - Perhaps even a ‘synthetic’ spatial structure imposed
        - i.e. if there is a geography to the (in)accuracy