Formal Privacy:
Making an Impact at Large Organizations

Deploying Differential Privacy for the 2020 Census of Population and Housing

Simson L. Garfinkel
Senior Scientist, Confidentiality and Data Access
U.S. Census Bureau

July 31, 2019
JSM 2019

The views in this presentation are those of the author, and not those of the U.S. Census Bureau.

The views in this presentation are those of the author, and not those of the U.S. Census Bureau.

Acknowledgments

This presentation incorporates work by:

Dan Kifer (Scientific Lead)

John Abowd (Chief Scientist)

Tammy Adams, Robert Ashmead, Aref Dajani, Jason Devine, Nathan Goldschlag, Michael Hay, Cynthia Hollingsworth, Meriton Ibrahimi, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Brett Moran, Ned Porter, Anne Ross, William Sexton, Lars Vilhuber, and Pavel Zhuravlev

Key points about the 2020 Census

"Count everyone once, only once, and in the right place."

World's longest-running statistical program.

First conducted in 1790 by Thomas Jefferson

Must be an "actual Enumeration" (US Constitution)

Data collected under a pledge of confidentiality

Disclosure Avoidance in the 2010 Census: Swapping

2010 Census used household swapping

- Swapping was limited to households within a state
- Swapping was limited to households the same size
- Swapping rate is confidential.

We performed a reconstruction attack and re-identified data from 17% of the US population.

- We did not reconstruct families.
- We did not recover detailed self-identified race codes

Disclosure Avoidance and the 2020 Census: Differential Privacy

USCB first adopt differential privacy in 2008 for OnTheMap

John Abowd became Chief Scientist in 2016 with the goal of modernizing disclosure avoidance

Data products include:

- Decennial Census of Population and Housing
- Economic Census
- American Community Survey
- Ad hoc research in Federal Statistical Research Data Centers
- +100 other major data products

Despite its Size, the Decennial Census is the *Easiest* US Census Bureau Product to Make Differentially Private

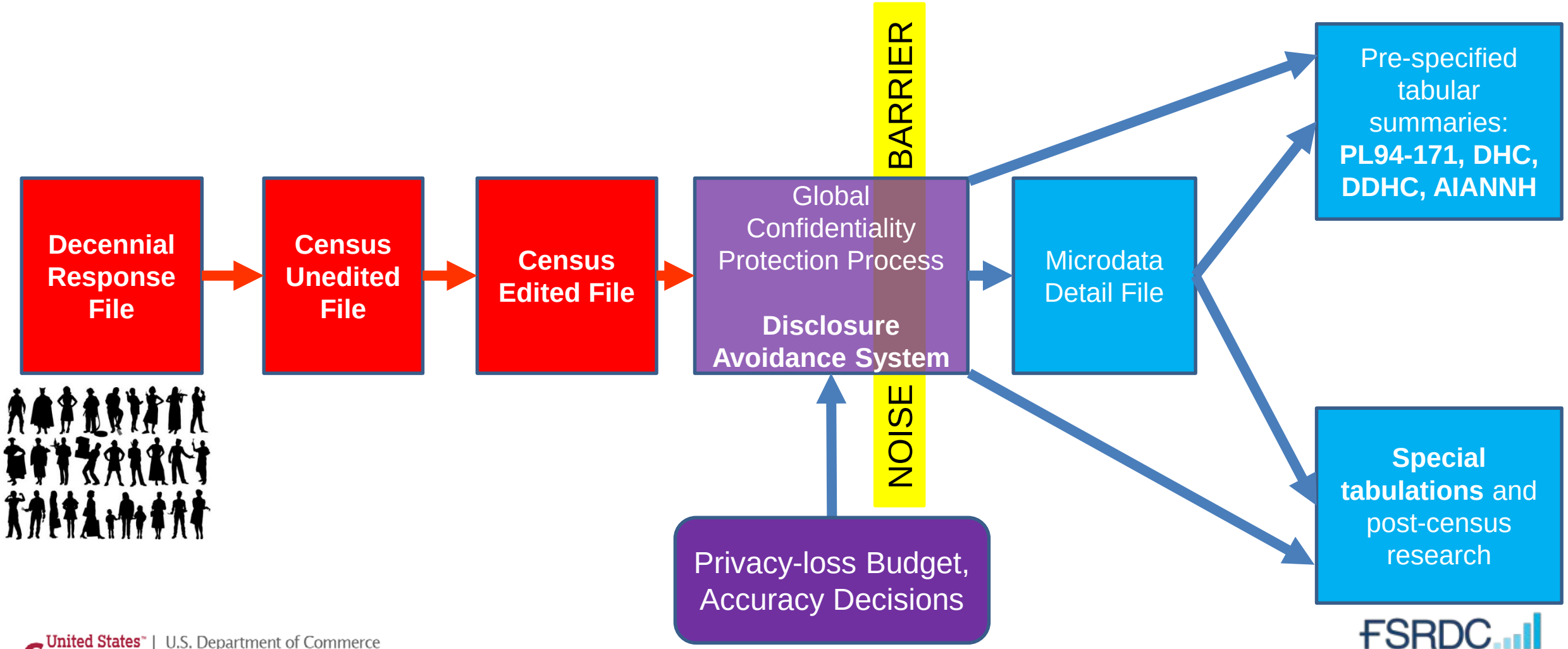
Only 5 tabulation variables collected per person:

Age, Sex, Race, Ethnicity, Relationship to Householder, Location

It's a census — no weights!

National Priority ➔ well-funded

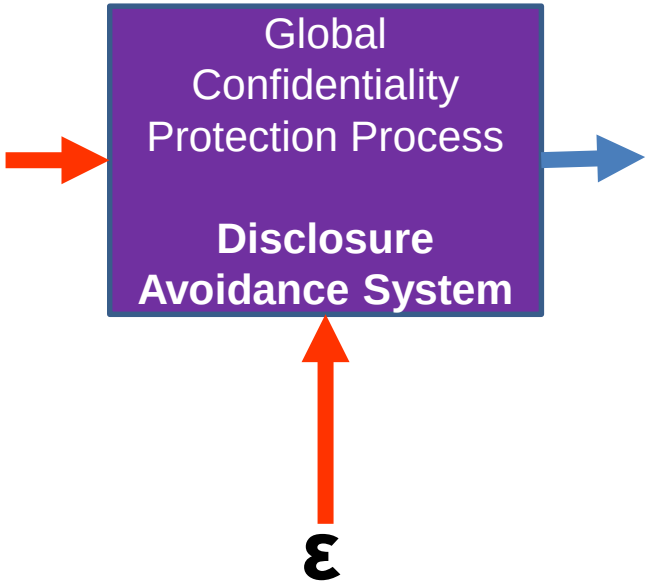
DAS allows the Census Bureau to enforce global confidentiality protections



The Disclosure Avoidance System relies on injects formally private noise

Advantages of noise injection with formal privacy:

- Transparency: the details can be explained to the public
- Tunable privacy guarantees
- Privacy guarantees do not depend on external data
- Protects against accurate database reconstruction
- Protects every member of the population



Challenges:

- Entire country must be processed at once for best accuracy
- Every use of confidential data must be tallied in the *privacy-loss budget*

There was no off-the-shelf system for applying differential privacy to a national census

We had to create a new system that:

- Produced higher-quality statistics at more densely populated geographies
- Produced consistent tables

We created new differential privacy algorithms and processing systems that:

- Produce highly accurate statistics for large populations (e.g. states, counties)
- Create protected microdata that can be used for any tabulation without additional privacy loss
- Fit into the decennial census production system

Basic approach for a DP Census

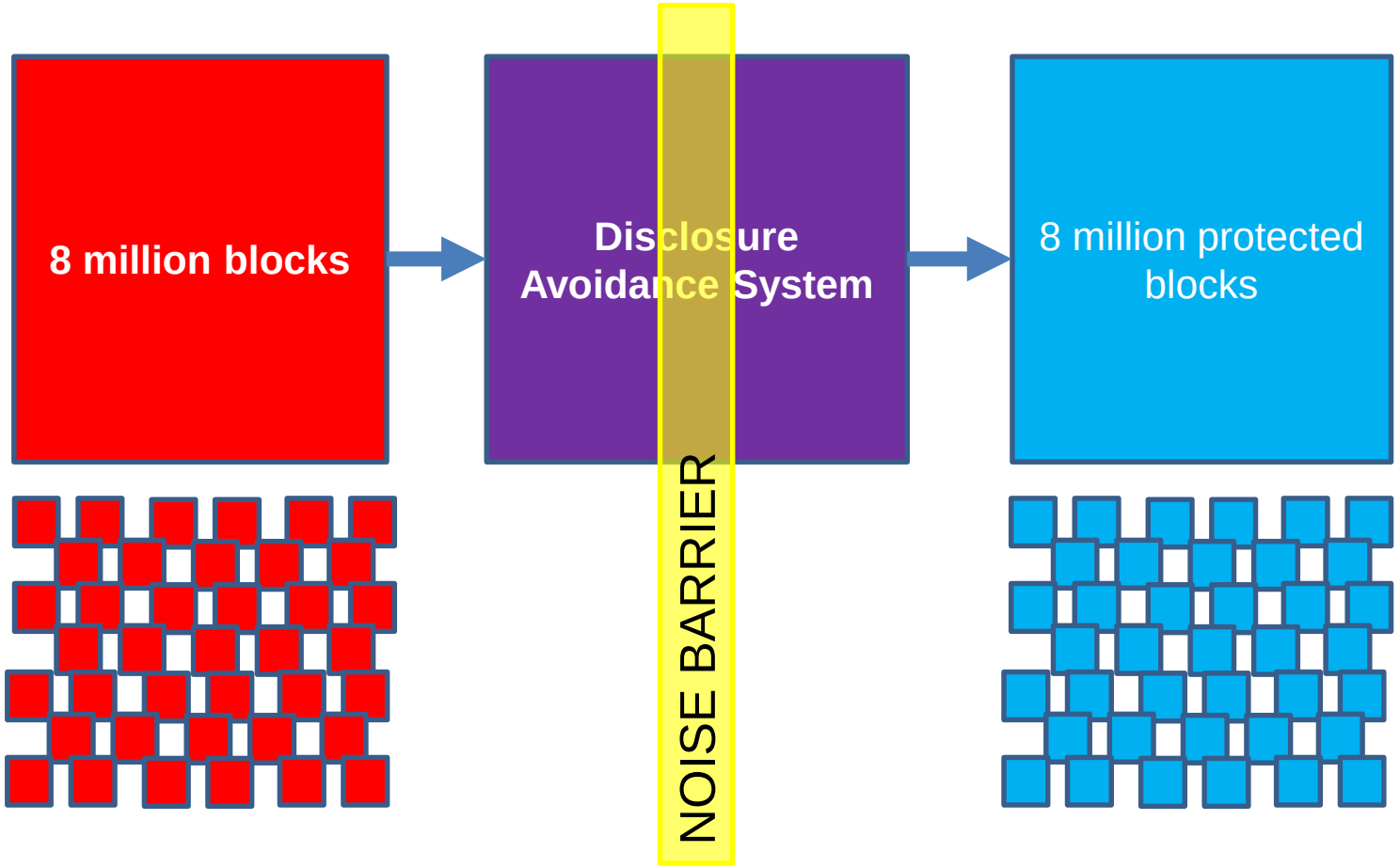
Treat the *entire census* as a set of queries on histograms.

Select the specific queries to measure

- Six *geolevels* (nation, state, county, tract, block group, block)
- Thousands of queries per *geounit*
- Billions of queries overall
- Histogram has billions of cells

First effort: The block-by-block algorithm

Independently protect each block (parallel composition)



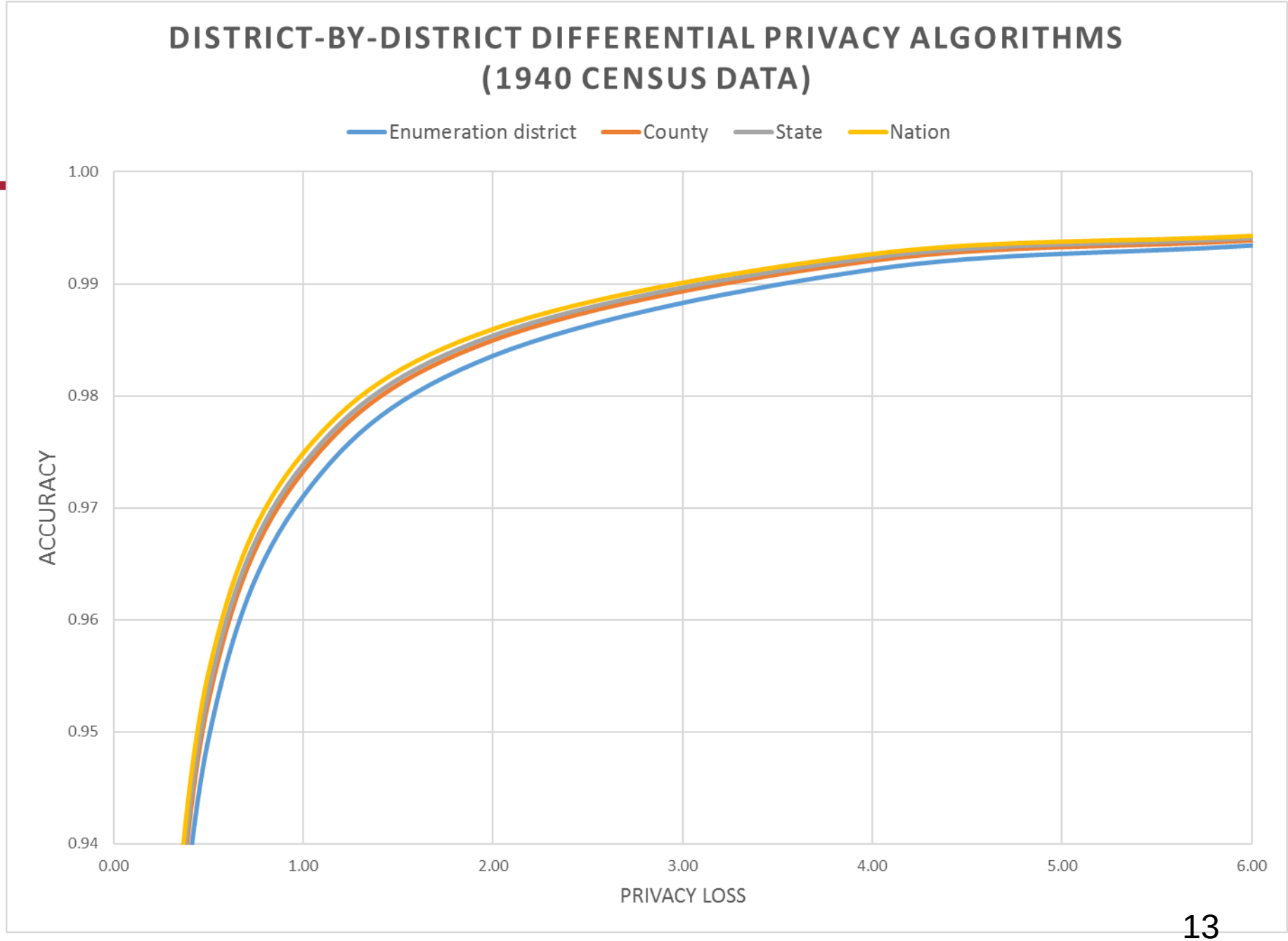
Measure queries for each block; privatize queries; convert results back to microdata

Tested with data from 1940

1940 hierarchy:

- Nation
- State
- County
- Enumeration District

Download from usa.ipums.org



Block-by-block algorithm (also called bottomUp)

Mechanism:

- Select, Measure, Reconstruct separately on each block

Advantages:

- Simple and easy to parallelize
- Privacy cost does not depend on # of blocks
- Releasing DP for one block has same cost as releasing for all

Disadvantages

- Significant error at higher level
- Error adds up
- Variance of each geounit is proportional to the number of blocks it contains

New algorithm: the top-down mechanism

Step 1: Generate national histogram without geographic identifiers.

Step 2: Allocate counts in histogram to each geography "top down."

National-level measurements - ε_{nat}

State-level histograms - ε_{state}

County-level histograms - ε_{county}

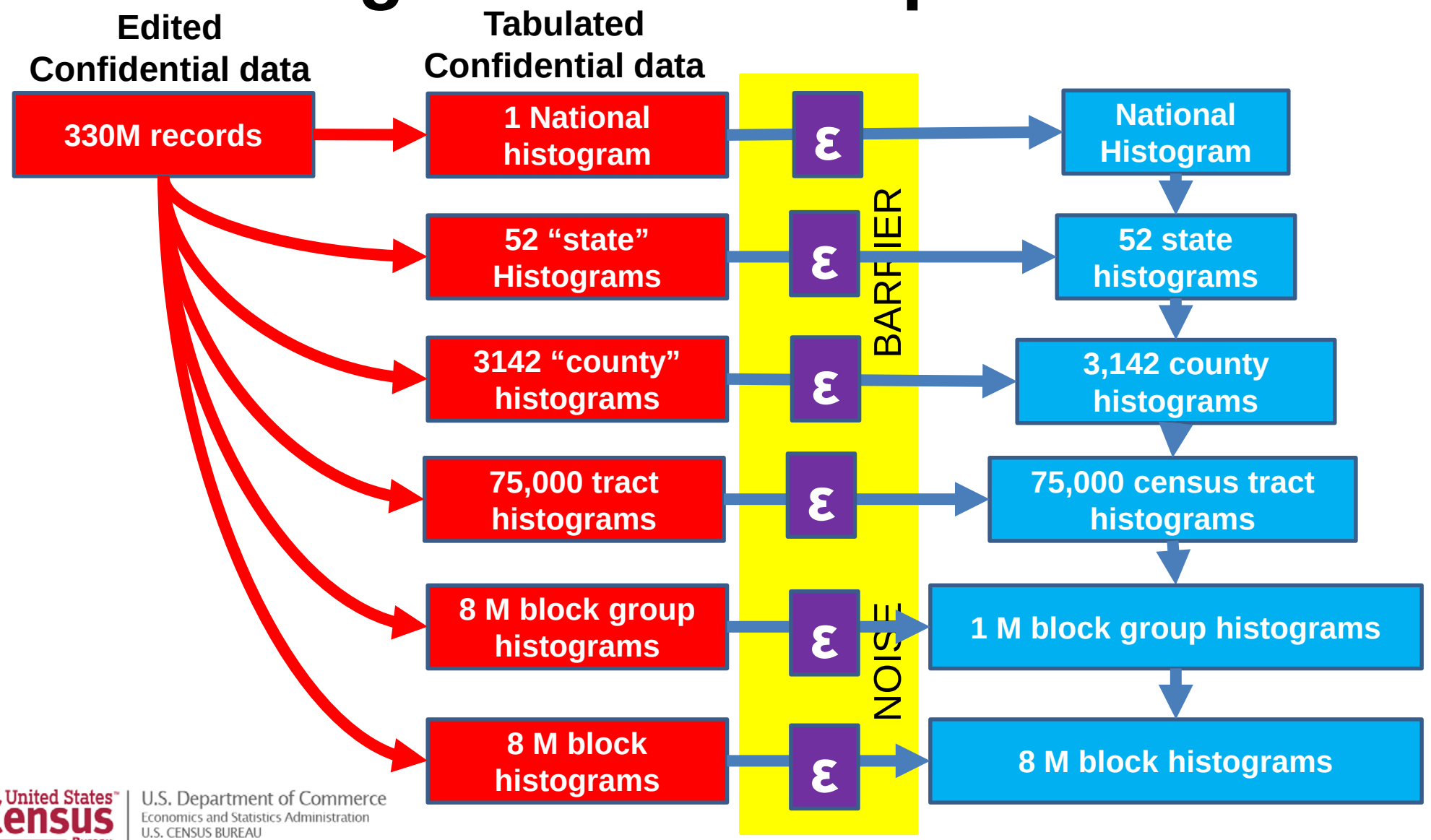
Tract-level histograms - ε_{tract}

Block-group level histograms - $\varepsilon_{blockgroup}$

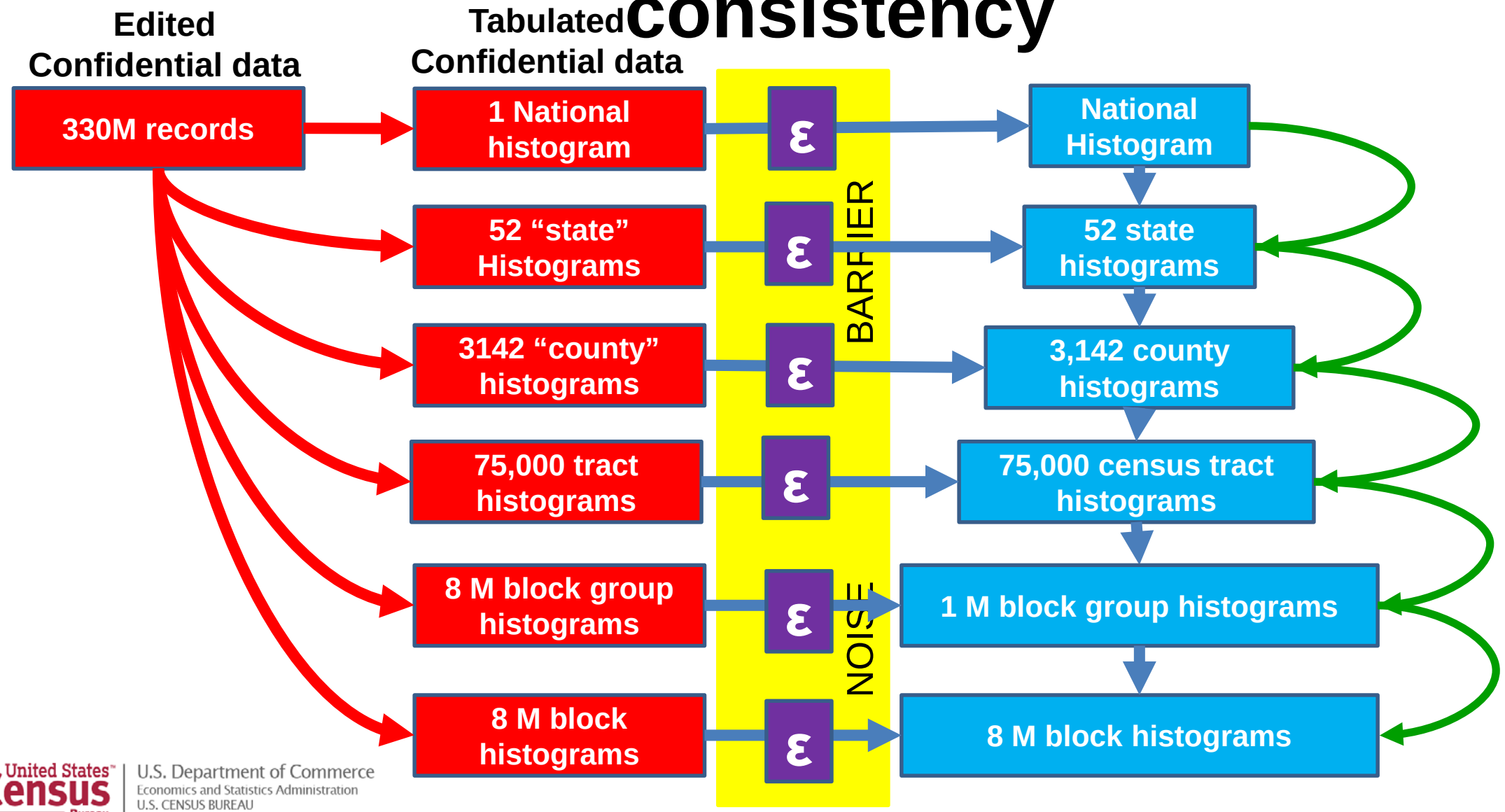
Block-level histograms - ε_{block}

$$\varepsilon = \varepsilon_{nat} + \varepsilon_{state} + \varepsilon_{county} + \varepsilon_{tract} + \varepsilon_{blockgroup} + \varepsilon_{block}$$

New algorithm: the top-down mechanism



Post-process for non-negativity and consistency



Top-Down Framework

Advantages:

Easy to parallelize

Each geo-unit can have its own strategy selection

We use High Dimensional Matrix Mechanism [MMHM18]

Parallel composition at each geo-level

Reduced variance for many aggregate regions

Sparsity discovery

- *e.g. very few 100+ aged people who combine 5 races*
- *Once to—down decide a region has no such records in county A, no subregion will have them.*

Evaluating the algorithm

We released runs of the top-down algorithm on data from the 1940 Census.

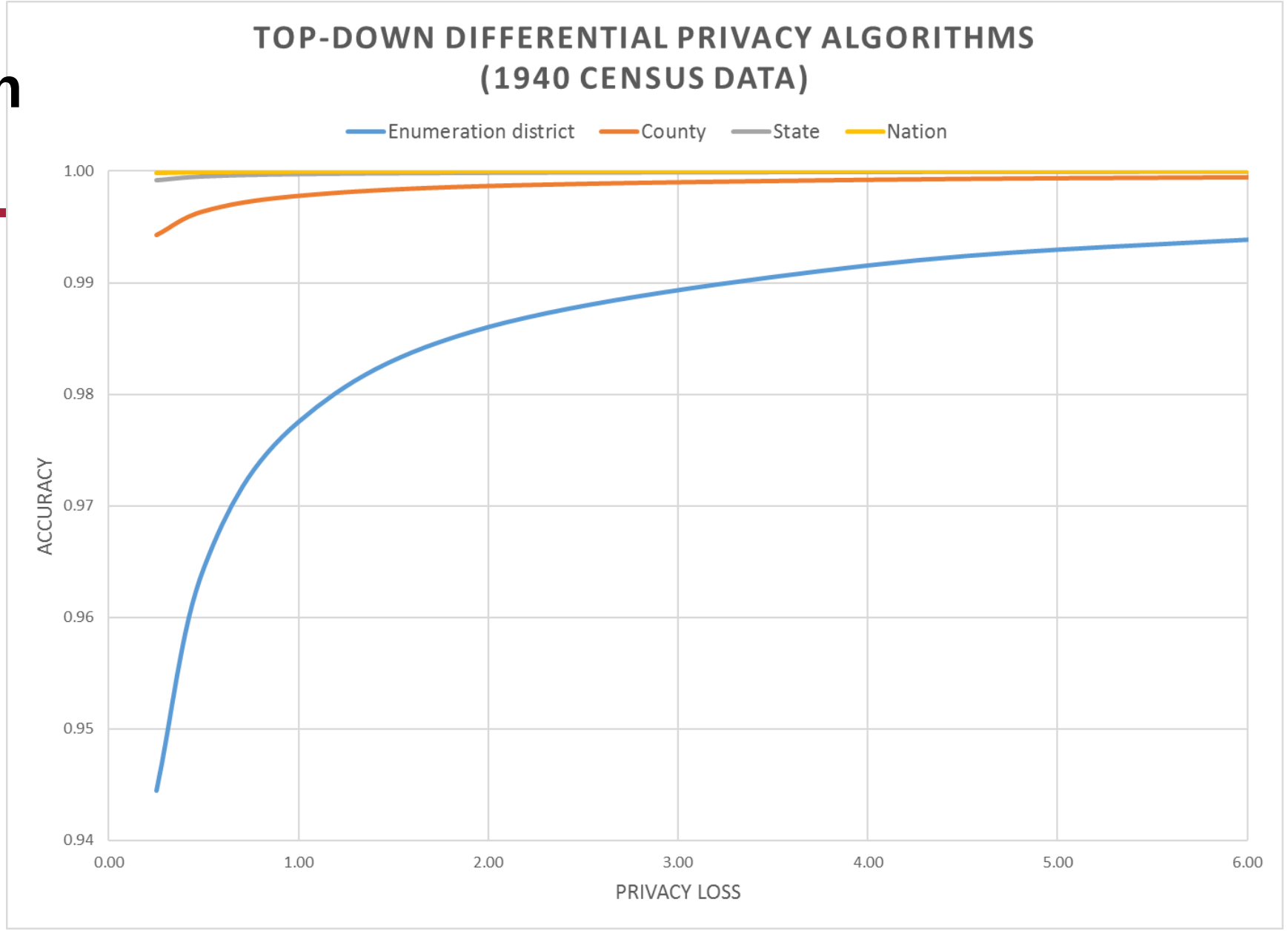
- Epsilon values 0.25 .. 8.0
- Multiple runs at each value of epsilon.

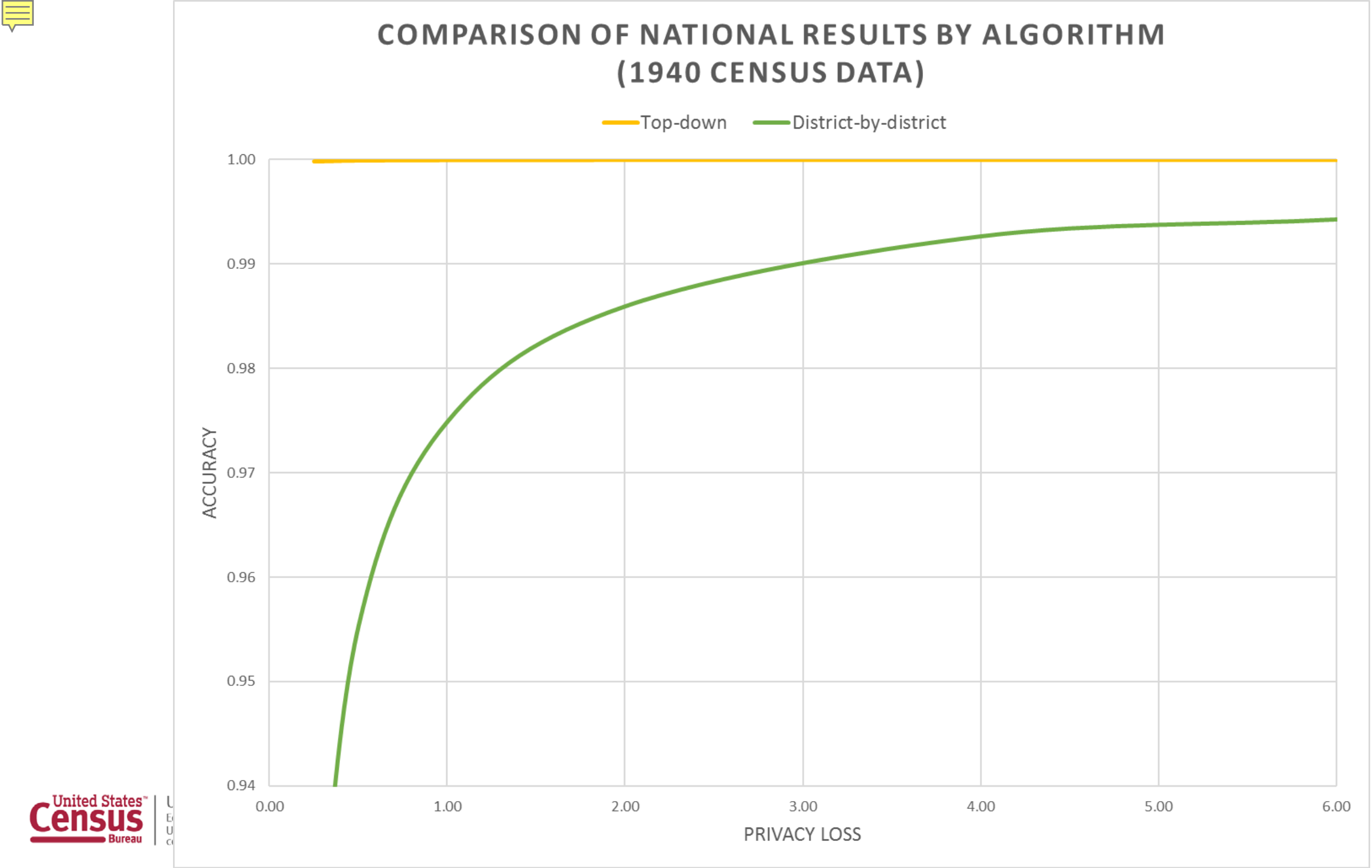
Caveats:

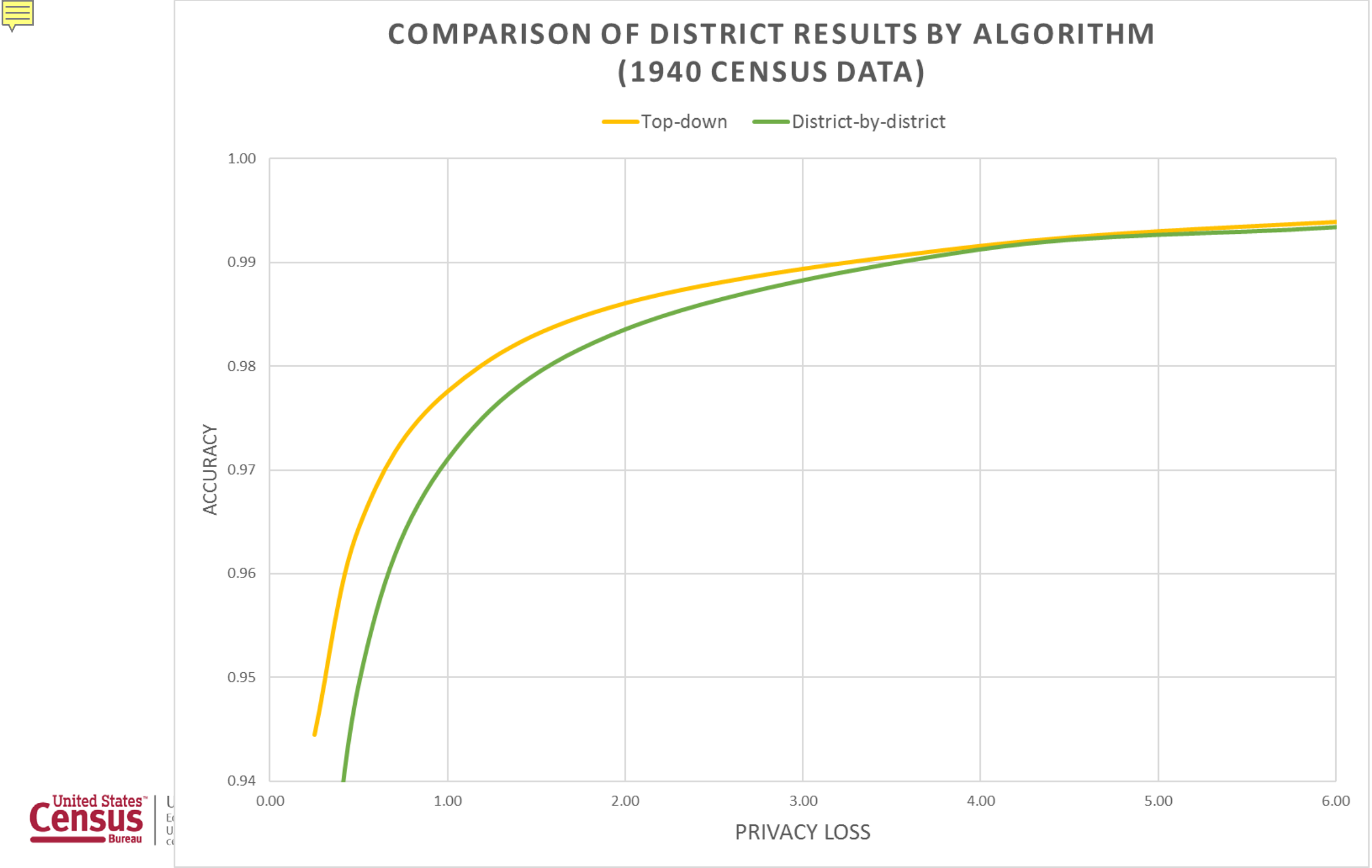
- 1940 data had 4 geography levels: Nation, State, County, Enumeration District.
  2020 data has 6 levels: Nation, State, County, Tract, Block Group and Block.

- 1940 data has 6 races / 2020 data has 63 race combinations
- 1940 data has no citizenship (Citizen or non-Citizen)

Top-Down: much more accurate!




COMPARISON OF NATIONAL RESULTS BY ALGORITHM
(1940 CENSUS DATA)
Top-down
District-by-district
1.00
0.99
0.98
0.97
0.96
0.95
0.94
ACCURACY
0.00
1.00
2.00
3.00
4.00
5.00
6.00
PRIVACY LOSS


COMPARISON OF DISTRICT RESULTS BY ALGORITHM
(1940 CENSUS DATA)
Top-down
District-by-district
1.00
0.99
0.98
0.97
0.96
0.95
0.94
ACCURACY
0.00
1.00
2.00
3.00
4.00
5.00
6.00
PRIVACY LOSS

https://dpwiki.org/demo-top-d

https://dpwiki.org/demo-top-down/dp-demo/sim-top-down/demo.html

Confidential database:

Tiny County pop: 642
f: 318
m: 324

Ruralland pop: 21 f: 9 m: 12			Urbanville pop: 621 f: 309 m: 312		
RBlock pop: 3 f: 1 m: 2	RBlock pop: 7 f: 3 m: 4	RBlock pop: 11 f: 5 m: 6	UBlock pop: 203 f: 101 m: 102	UBlock pop: 207 f: 103 m: 104	UBlock pop: 211 f: 105 m: 106

Noise Barrier

privatize!

ε 1.0

Published official tabulations:

Tiny County pop: 642
f: 319
m: 323

Ruralland pop: 21 f: 10 m: 11			Urbanville pop: 621 f: 309 m: 312		
RBlock pop: 8 f: 4 m: 4	RBlock pop: 2 f: 2 m: 0	RBlock pop: 11 f: 4 m: 7	UBlock pop: 199 f: 98 m: 101	UBlock pop: 212 f: 107 m: 105	UBlock pop: 210 f: 104 m: 106

Each rectangle shows the population statistics for a different geographical area. The top is the total population (pop), followed by the number of females (f) and the number of males (m).

ε specifies the privacy loss budget. Click **privatize!** to re-run the privacy mechanism with a different set of random noises.

Try changng the number of females or males that was counted on a block and see how it changes the official tabulations. Or choose one of the sample scenarios listed below.

select	scenario
	balanced rural and urban blocks
	One rural block with a LOT of males.
	One urban block with a LOT of females.

Note: The simulator uses hypothetical (fake) data provided by the user.

United Stat
Censu
Bur

Two public policy choices:

What is the correct value of epsilon?

Where should the accuracy be allocated?

Organizational Challenges

Process documentation

- All uses of confidential data need to be tracked and accounted.

Workload identification

- All desired queries on MDF should be known in advance.
- Required accuracy for various queries should be understood.
- Queries outside of MDF must also be pre-specified

Correctness and Quality control

- Verifying implementation correctness.
- Data quality checks on tables cannot be done by looking at raw data.

Data User Challenges

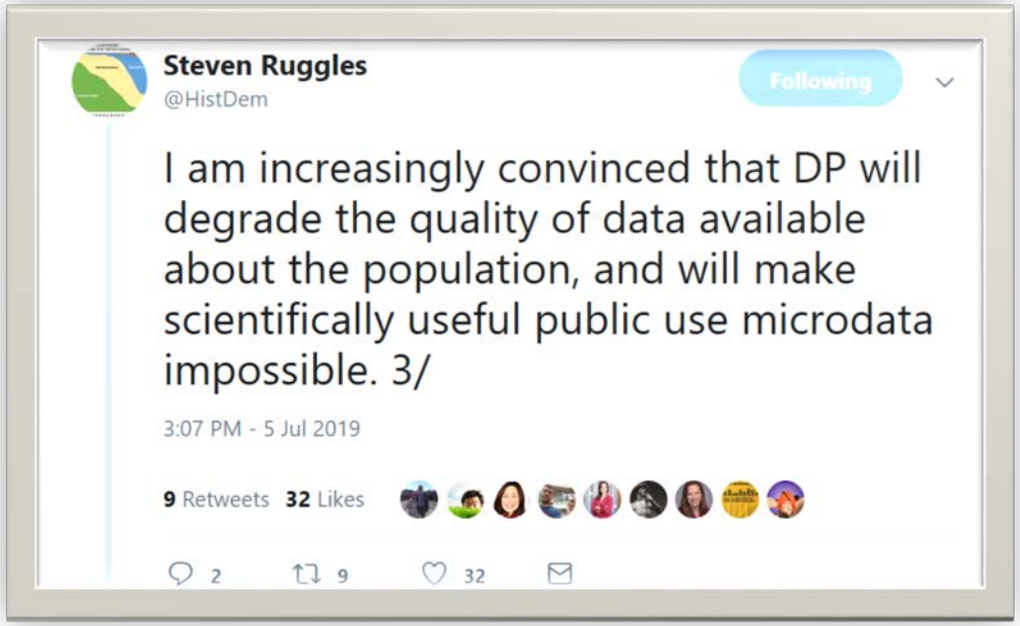
Differential privacy is not widely known or understood.

Many data users want highly accurate data reports on small areas.

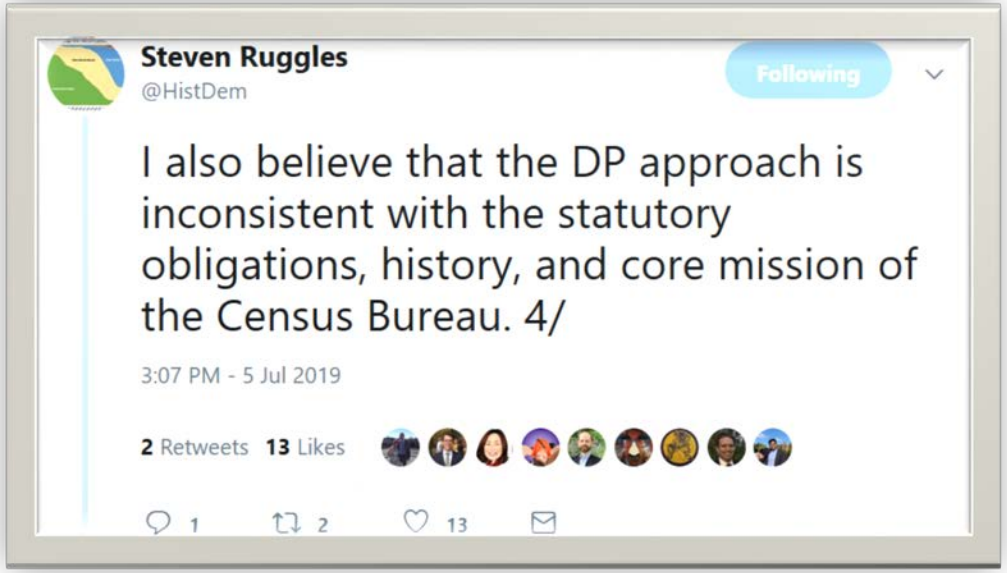
- Some are anxious about the intentional addition of noise.
- Some are concerned that previous studies done with swapped data might not be replicated if they used DP data.

Many data users believe they require access to Public Use Microdata.

Users in 2000 and 2010 didn't know the error introduced by swapping and other protections applied to the tables and PUMS.



Concerns and Responses



Redistricting and Exact Counts

In the US, legislative districts must have equal size.

- Decennial Census counts of each block are the “official counts.”

Some data users are concerned that adding noise to the counts will make them unfit for use.

However:

- Evaluation of districts is based on official decennial counts; these data are used for 10 years.
- Noise added by DP is significantly less than noise added by other statistical methods currently in use

STEVEN RUGGLES

Regents Professor of History and Population Studies
Director, Institute for Social Research and Data Innovation
50 Willey Hall
University of Minnesota
ruggles@umn.edu
(612) 624-5818

Ruggles Concerns

Differential privacy is not a measure of identifiability

Differential privacy does not measure disclosure risk

"Differential Privacy is not concerned with re-identification of respondents

- "DP prohibits revealing *characteristics* of an individual even if the *identity* of that individual is effectively concealed
- "This is a radical departure from established census law and precedent
- "The Census Bureau has been disseminating individual-level *characteristics* routinely since the first microdata in 1962

Organized attack on the move to differential privacy

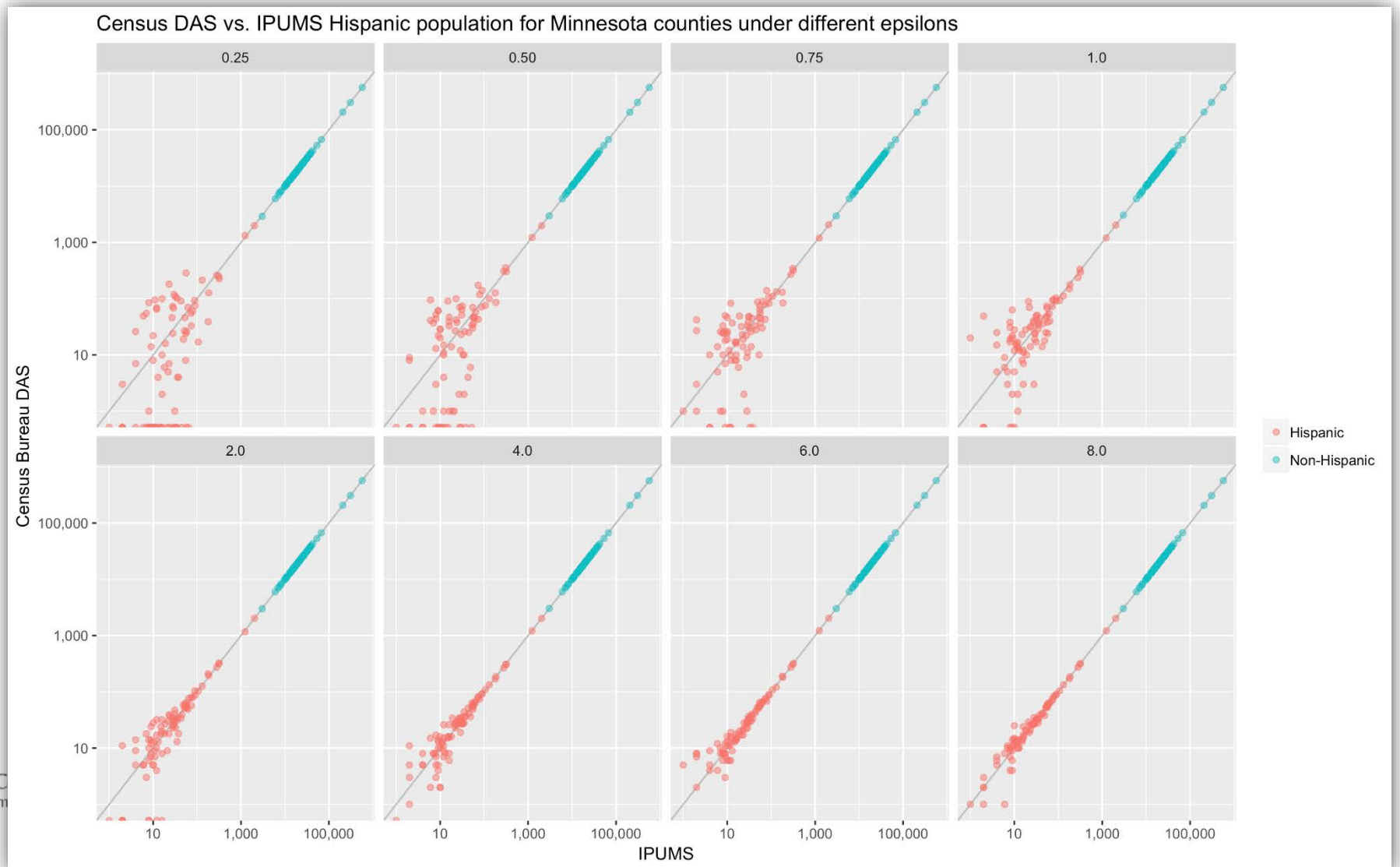
STEVEN RUGGLES

Regents Professor of History and Population Studies
Director, Institute for Social Research and Data Innovation
50 Willey Hall
University of Minnesota
ruggles@umn.edu
(612) 624-5818

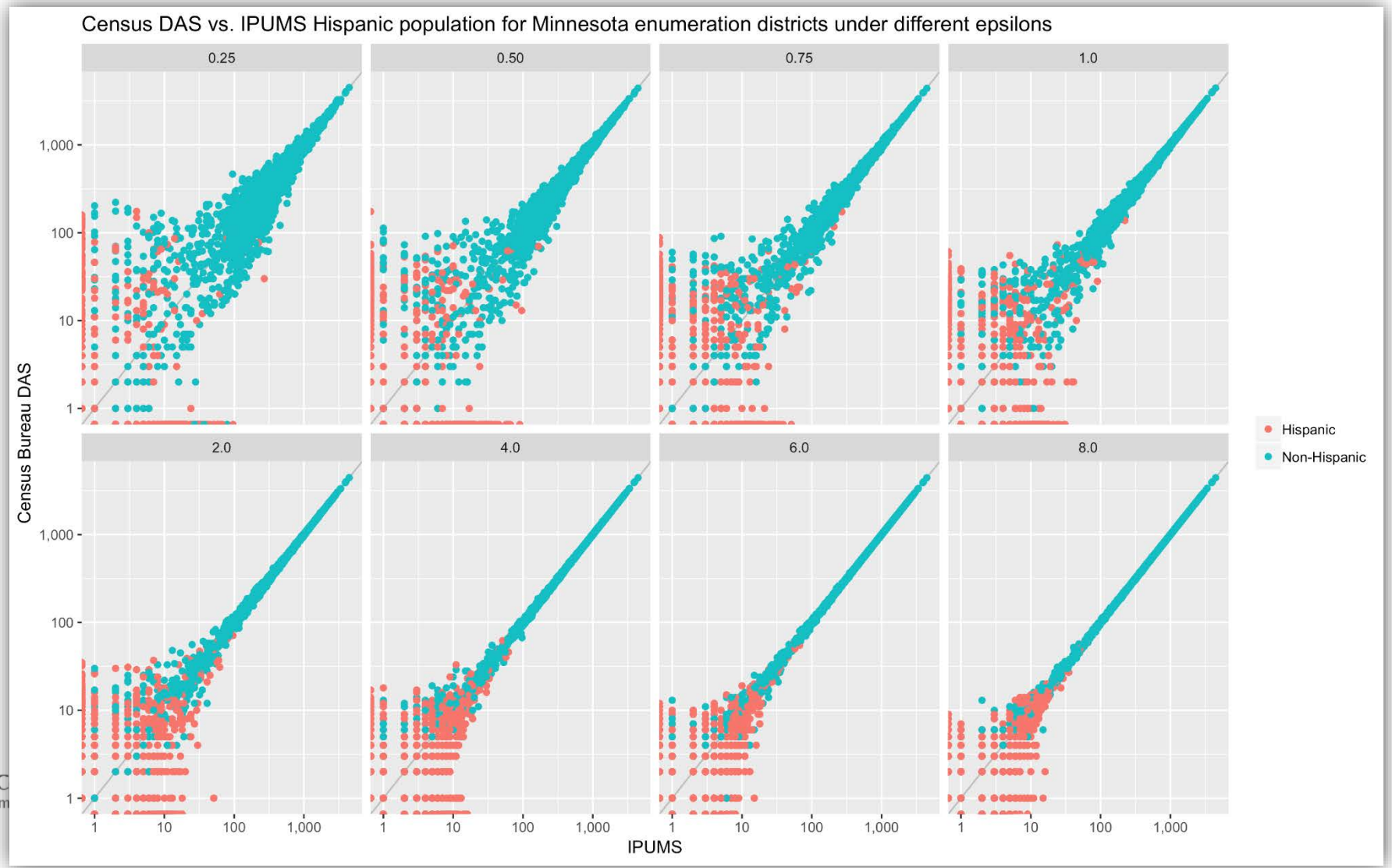
Concerns:

- "Differential privacy will degrade the quality of data available about the population, and will probably make scientifically useful public use microdata impossible
- The differential privacy approach is inconsistent with the statutory obligations, history, and core mission of the Census Bureau"

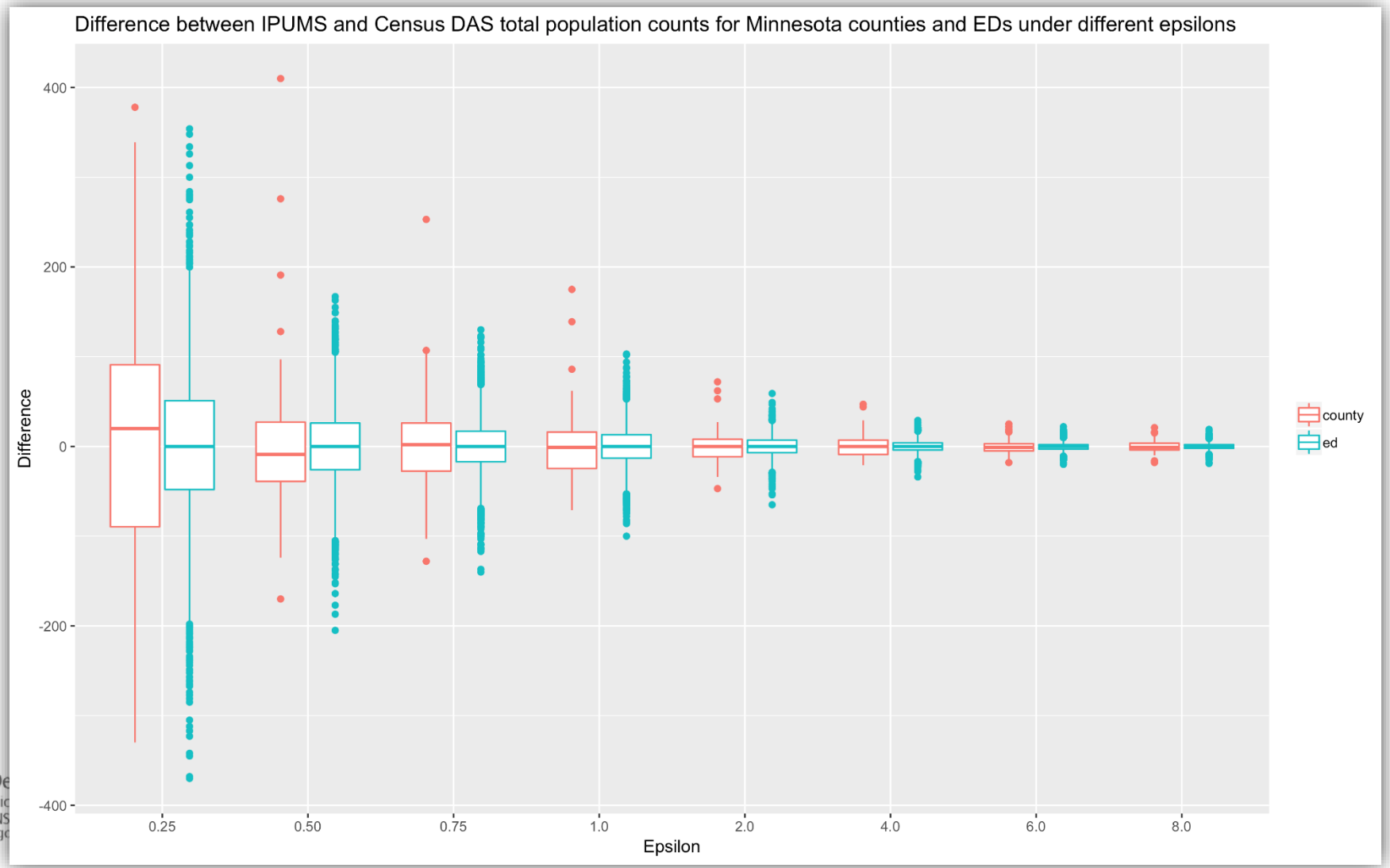
Analysis of population variances
David Van Riper & Tracy Kugler, IPUMS (APDU 2019)



Analysis of population variances
David Van Riper & Tracy Kugler, IPUMS (APDU 2019)



Analysis of population variances
David Van Riper & Tracy Kugler, IPUMS (APDU 2019)



For more information…

practice

DOI:10.1145/3287287

Article development led by acmqueue
queue.acm.org

These attacks on statistical databases are no longer a theoretical danger.

BY SIMSON GARFINKEL, JOHN M. ABOWD, AND CHRISTIAN MARTINDALE

Understanding Database Reconstruction Attacks on Public Data

IN 2020, THE U.S. Census Bureau will conduct the Constitutionally mandated decennial Census of Population and Housing. Because a census involves collecting large amounts of private data under the promise of confidentiality, traditionally statistics are published only at high levels of aggregation. Published statistical tables are vulnerable to *database reconstruction attacks* (DRAs), in which the underlying microdata is recovered merely by finding a set of microdata that is consistent with the published statistical tabulations. A DRA can be performed by using the tables to create a set of mathematical constraints and then solving the resulting set of simultaneous equations. This article shows how such an attack can be addressed by adding noise to the published tabulations, so the reconstruction no longer results in the original data. This has implications for the 2020 census.

The goal of the census is to count every person once, and only once, and in the correct place. The results are used to fulfill the Constitutional requirement to apportion the seats in the U.S. House of Representatives among the states according to their respective numbers.

In addition to this primary purpose of the decennial census, the U.S. Congress has mandated many other uses for the data. For example, the U.S. Department of Justice uses block-by-block counts by race for enforcing the Voting Rights Act. More generally, the results of the decennial census, combined with other data, are used to help distribute more than $675 billion in federal funds to states and local organizations.

Beyond collecting and distributing data on U.S. citizens, the Census Bureau is also charged with protecting the privacy and confidentiality of survey responses. All census publications must uphold the confidentiality standard specified by Title 13, Section 9 of the U.S. Code, which states that Census Bureau publications are prohibited from identifying "the data furnished by any particular establishment or individual." This section prohibits the Census Bureau from publishing respondents' names, addresses, or any other information that might identify a specific person or establishment.

Upholding this confidentiality requirement frequently poses a challenge, because many statistics can inadvertently provide information in a way that can be attributed to a particular entity. For example, if a statistical agency *accurately* reports there are two persons living on a block and the average age of the block's residents is 35, that would constitute an improper disclosure of personal information, because one of the residents could look up the data, subtract their contribution, and infer the age of the other.

Communications of ACM March 2019
Garfinkel & Abowd

Can a set of equations keep U.S. census data private?
By Jeffrey Mervis
Science
Jan. 4, 2019 , 2:50 PM

http://bit.ly/Science2019C1

More Background on the 2020 Disclosure Avoidance System

September 14, 2017 CSAC (overall design)
https://www2.census.gov/cac/sac/meetings/2017-09/garfinkel-modernizing-disclosure-avoidance.pdf

August, 2018 KDD'18 (top-down v. block-by-block)
https://digitalcommons.ilr.cornell.edu/ldi/49/

October, 2018 WPES (implementation issues)
https://arxiv.org/abs/1809.02201

October, 2018 *ACMQueue* (understanding database reconstruction)
https://digitalcommons.ilr.cornell.edu/ldi/50/ or
https://queue.acm.org/detail.cfm?id=3295691

Memorandum 2019.13: Disclosure Avoidance System Design Parameters
https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series/2020-memo-2019_13.html