

Differential Privacy: Theory to Practice for the 2020 US Census

Simson L. Garfinkel
Chief Scientist, BasisTech, LLC
and
(former) Senior Computer Scientist for Confidentiality and
Data Access, US Census Bureau

Computer Laboratory Security Seminar Series
December 18, 2023

NOTE: The views in this presentation are those of the author(s), and do not necessarily represent those of the U.S. Government, the U.S. Census Bureau, or any other U.S. Government agency.

Abstract

From 2016 through 2021, statisticians and computer scientists at the US Census Bureau worked on the largest and most complex deployment of differential privacy to date: using the modern mathematics of privacy to protect the census responses for more than 330 million residents of the United States as part of the 2020 Census of Population and Housing.

This talk presents a first-hand account of the challenges that were faced trying to apply the still young and evolving theory of differential privacy to the world's longest running statistical program. These challenges included the need to complete and deploy scientific research on a tight deadline, working in complex deployment environments that had been intentionally crippled to achieve cybersecurity goals, working with a hostile data community of data users who did not want formal privacy protections applied to census data, and periodic interference from state and federal officials.

Moving scientific breakthroughs into practice is usually harder than we anticipate. Bigger breakthroughs are usually harder.

Outline for this talk:

Optional!

- What is differential privacy (DP), and why is it a scientific breakthrough?

Optional!

- What is the US Census and why does it matter?
 - How we brought DP to the 2020 Census
 - Internal Challenges
 - External Challenges
 - How the DP deployment timeline compares with public key cryptography

What is differential privacy, and why is it a scientific breakthrough?

(Please raise your hand if you have an expert understanding of differential privacy.)

Differential privacy protects confidential data used for public statistics.

Example:

- You are in a class with 9 other students.
- The teacher announces that the average score is 98%.
- You look at your test and you got an 80%.



ChatGPT



ChatGPT

- Now you know the grades for everyone in the class...

Statistical Disclosure Limitation (aka Disclosure Avoidance) protects confidential information used in statistics

*Student Scores
(Hidden variables)*

S1	S6
S2	S7
S3	S8
S4	S9
S5	S10



*Published Statistics
(Constraints)*

Class Average = 98%

Statistical Disclosure Limitation

Published statistics are constraints on hidden variables

Student Scores (Hidden variables)

S1 S6
S2 S7
S3 S8
S4 S9
S5 S10

Your Score

Published Statistics (Constraints)

Class Average = 98%

Implies:

$$S1+S2+S3+S4+S5+ \\ S6+S7+S8+S9+S10=100\%$$

If
S10=80%
and
 $0 \leq S_n \leq 100\%$
then:
S1..S9 = 100%

Statistical Disclosure Limitation (aka Disclosure Avoidance) protects confidential information used in statistics

*Student Scores
(Hidden variables)*

S1 S6
S2 S7
S3 S8
S4 S9
S5 S10

*Published Statistics
(Constraints)*

Class Average = 98%

What about count? $n = 10$

What about median? $\hat{x} = 100$

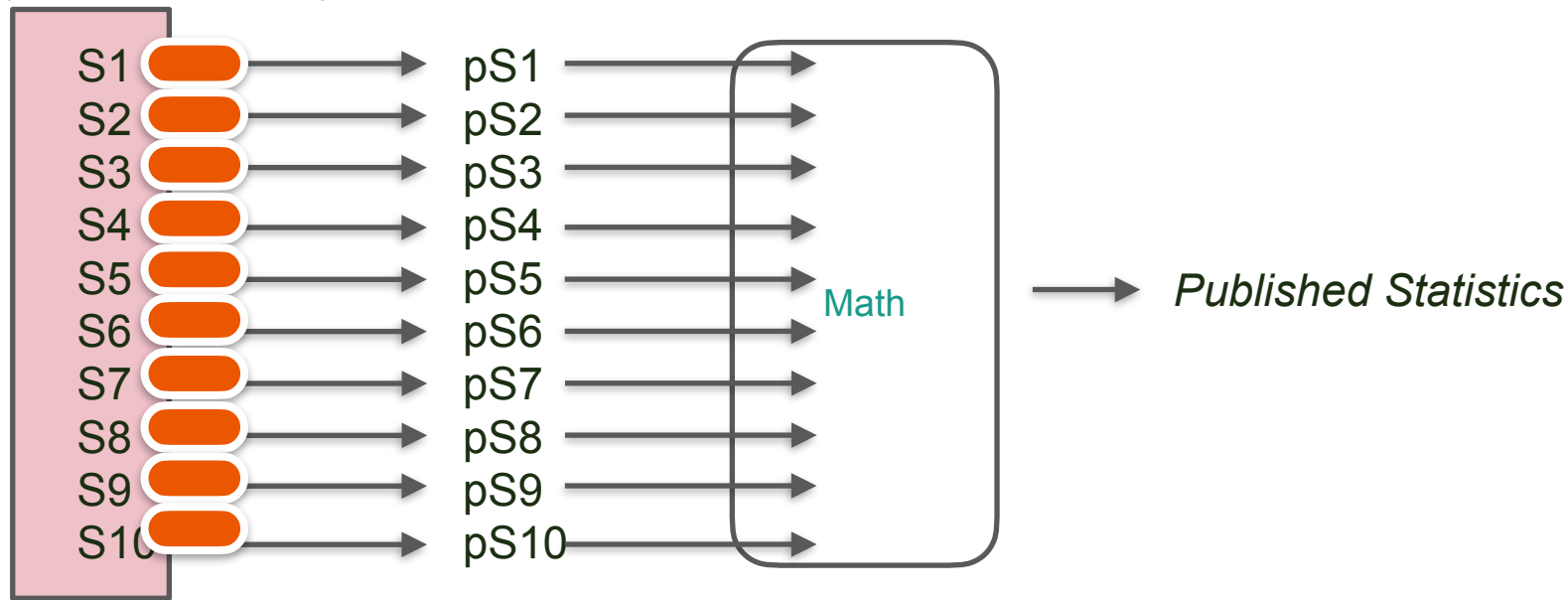
Could we publish that the class median is 100% ?

- These are policy questions!
- Does your policy prevent publishing the grade for half the class without identifying who got top grades.

SDL can be applied on inputs or outputs of a computation.

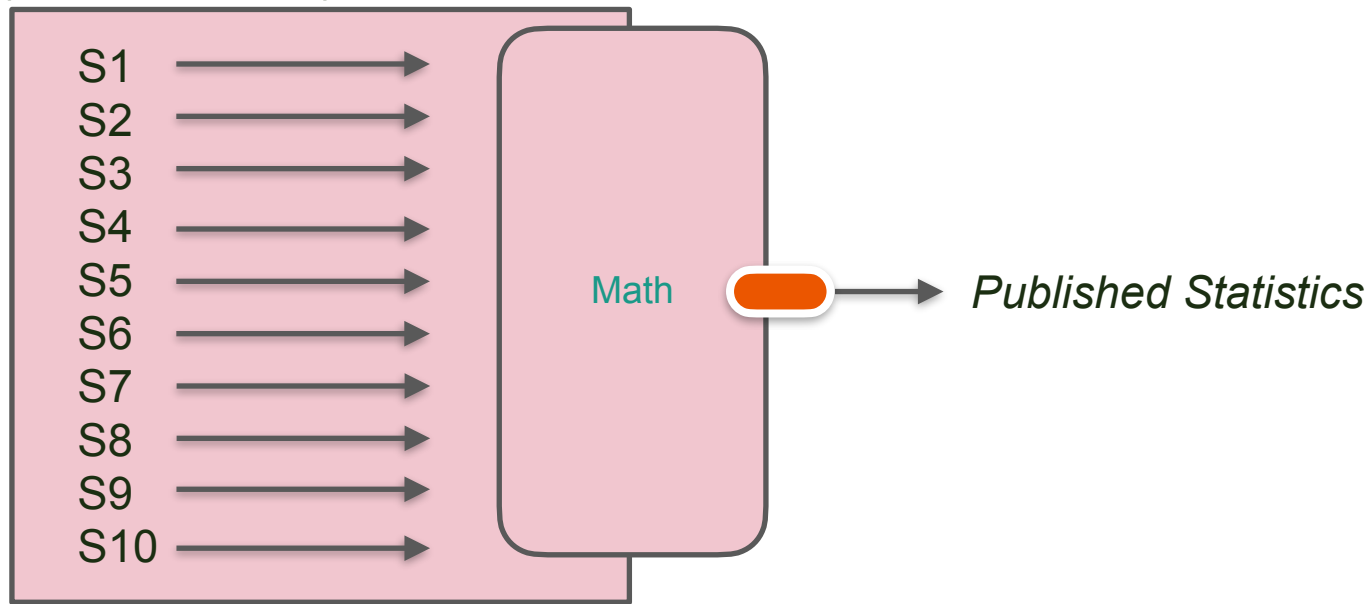
Input protection applies to each variable *before* it is used in the computation

Hidden variables
(Student Scores)



SDL can be applied on inputs or outputs of a computation.
Output protection applies during or after the computation.

Hidden variables
(Student Scores)



There are many SDL approaches.

Protect 98%

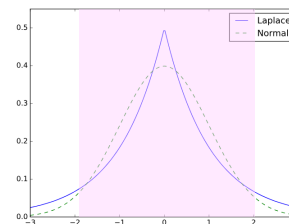


The class average is...

“is between 95% and 100%”

“not reportable due to the small class size”

“97%” (± 0.2 with 95% probability)

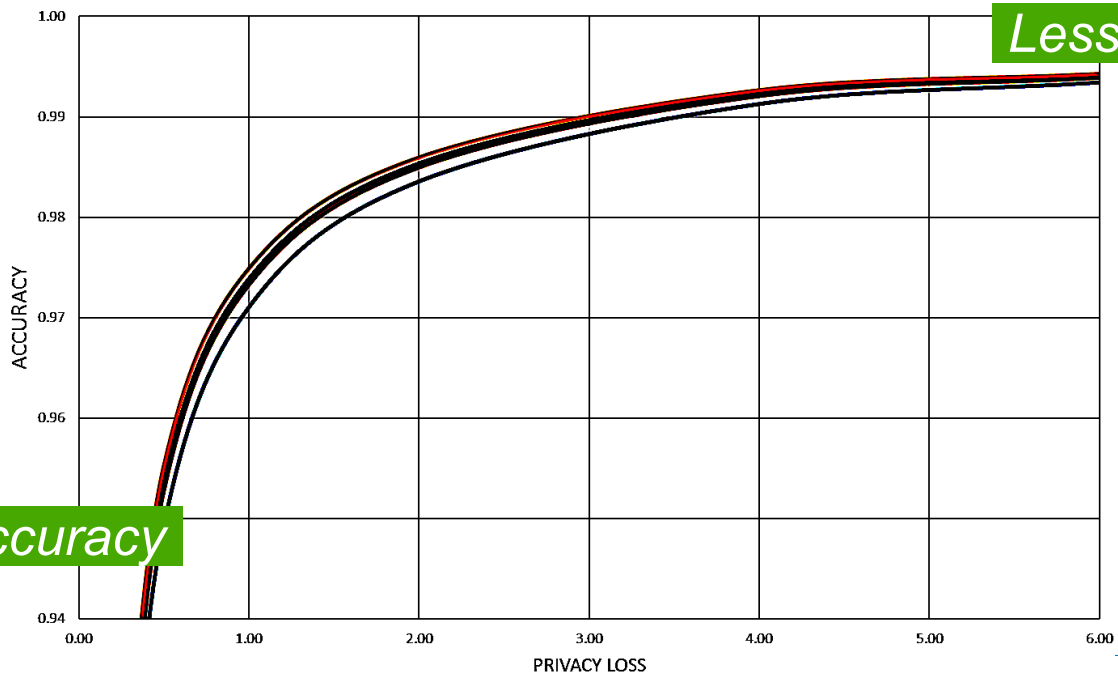


Noise infusion makes it possible to balance accuracy/utility with privacy protection.
More noise → more privacy, less accuracy



Differential privacy is based on the concept of “Privacy Loss” rather than privacy protection.

Privacy loss:
 $0 \leq \epsilon \leq \infty$



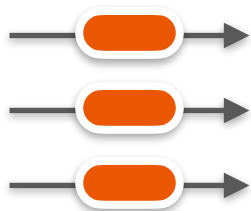
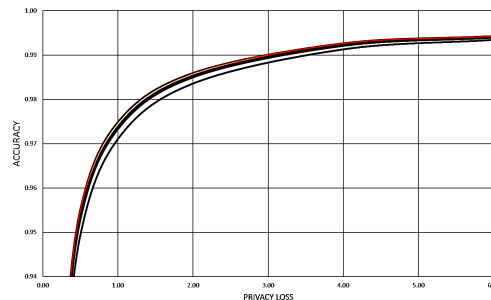
“Privacy bookkeeping” is the differential privacy breakthrough.

DP provides:

- The tradeoff between privacy loss and accuracy.

Composition rules:

- Accounting for total privacy loss in complex statistical pipelines



Parallel Composition
(e.g. multiple blocks)



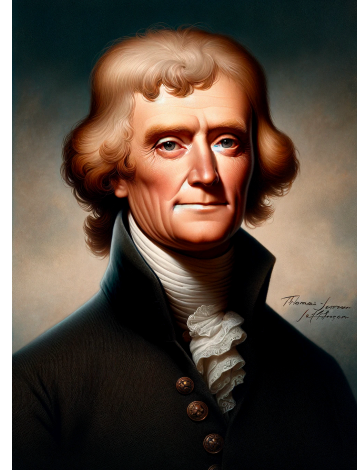
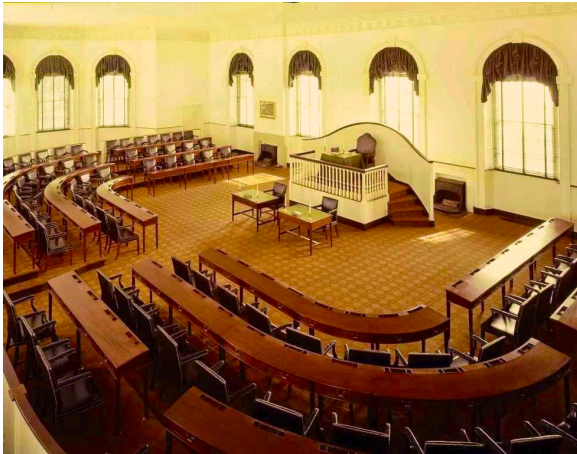
Serial Composition
(e.g. some statistics within a block)

What is the US Census and why does it matter?

The US Census is the world's longest running statistical program.

First US Census:
1790

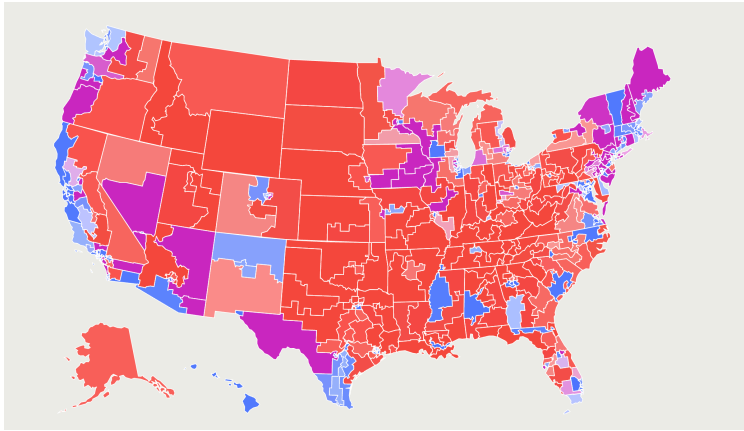
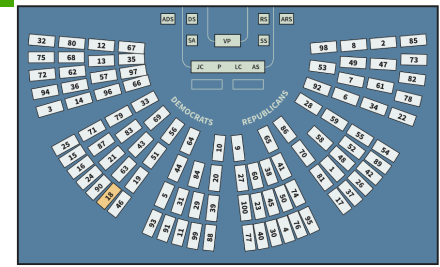
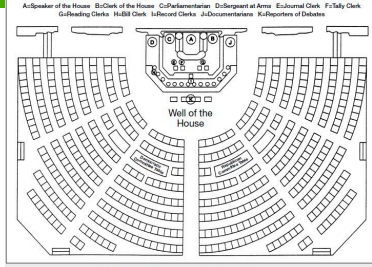
Purpose:
Apportion the US House of Representatives



ChatGPT

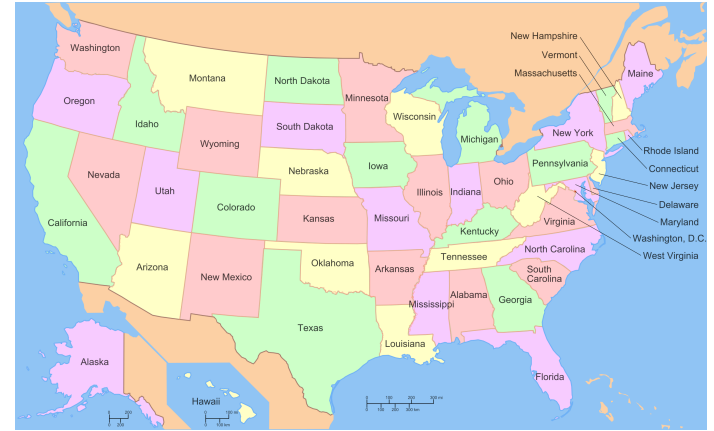
Thomas Jefferson
Primary author, US Declaration of Independence
First US Secretary of State
First US Patent Commissioner (reviewed every patent)
Oversaw First US Census

The US Constitution calls for a census every 10 years.
2020 was the 23rd US census.



Each congressional district elects a member to the House of Representatives.

There have been 435 seats since 1912



Each state elects 2 senators

The 1790 census collected just six pieces of information for each family

The name of the family's head

The number of free white males age 16 and older (including head of family)

The number of free white males under 16

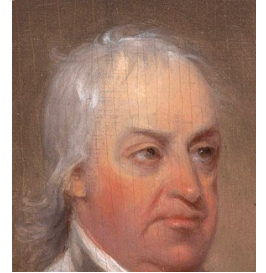
The number of free white females (including head of family)

The number of all other free persons (free African-Americans)

The number of slaves, each of whom contributed $\frac{3}{5}$ th to the final apportionment.



Virginia Representative
James Madison wanted to
collect more information.



New Hampshire Representative
Samuel Livermore argued that
collecting more information
would be costly and might lead
to higher taxes.

By 1860 the Census was collecting a lot of information.
Privacy became an issue.

Census Marshal – “I jist want to know how many of yez is defa, dumb, blind, insat and idiotic—likewise how many convicts there is in the family—what all your ages are especially the old women and the young ladies—and how many dollars the old gentleman is worth!”

Saturday Evening Post,
18 August 1860



The 2010 Census used three approaches to maintain statistical confidentiality

#1 – Record Swapping.

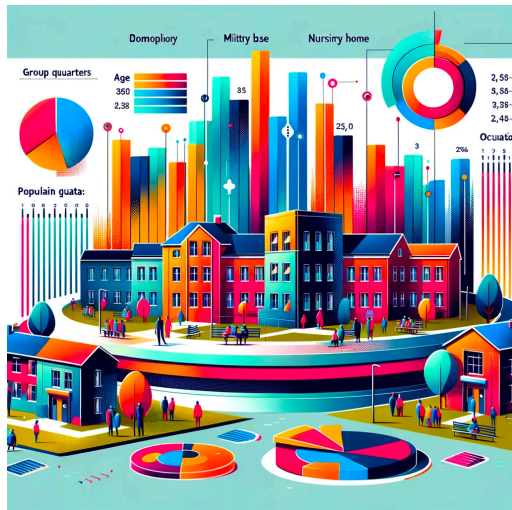
#2 – Synthetic data for group quarters (dorms, barracks, nursing homes, etc.)

#3 – Suppression (tables from 2000 were no longer provided)



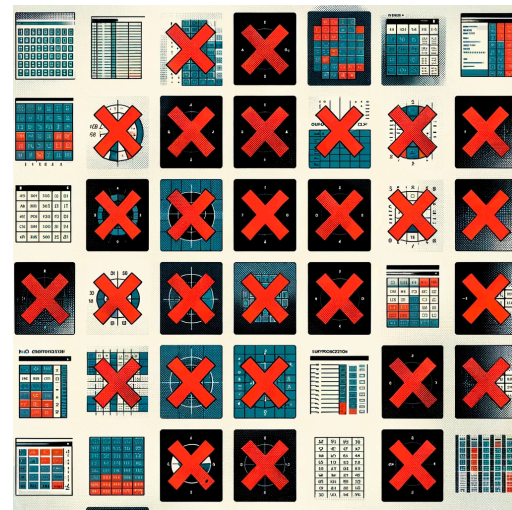
Swapping

ChatGPT



Synthetic Data

ChatGPT

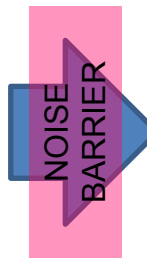


Suppression

ChatGPT

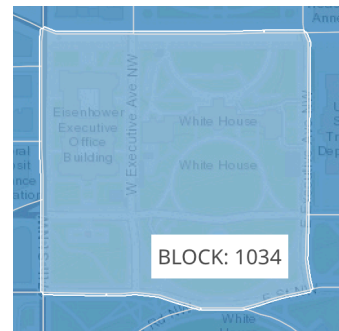
For 2020, we added noise to histograms of every census geography.
The histograms were then used to tabulate the official statistics.

	White	Black	AIAN	Asian	NHOPI
age ≥ 18	?	?	?	?	?
age < 18	?	?	?	?	?



	White	Black	AIAN	Asian	NHOPI
age ≥ 18	6	1	0	0	0
age < 18	0	0	0	0	0

Census Block 7500000US110019800001034
Total population: 10
(3 values not shown)



Explore DC:

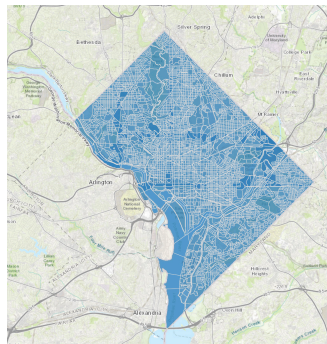
<https://opendata.dc.gov/datasets/DCGIS::census-blocks-in-2020/explore>

Technical Documentation on values:

https://www2.census.gov/programs-surveys/decennial/2020/technical-documentation/complete-tech-docs/summary-file/2020Census_PL94_171Redistricting_NationalTechDoc.pdf

The 2020 US census has millions of different geographies.
 Each has a histogram with thousands of cells. Each cell gets its own noise.

<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	...
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	...
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	...
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	...
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	...
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	<table border="1"> <thead> <tr><th>White</th><th>Black</th><th>AIAN</th><th>Asian</th><th>NHOP</th></tr> </thead> <tbody> <tr><td>age 2-18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> <tr><td>age <18</td><td>?</td><td>?</td><td>?</td><td>?</td></tr> </tbody> </table>	White	Black	AIAN	Asian	NHOP	age 2-18	?	?	?	?	age <18	?	?	?	?	...
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												
White	Black	AIAN	Asian	NHOP																																												
age 2-18	?	?	?	?																																												
age <18	?	?	?	?																																												



Washington DC = 6,507 blocks

US = 8,132,968 blocks
84,414 tracts
3143 counties
50 states
1 capital district
1 nation

How we brought DP to the 2020 Census

2010 Census used swapping and synthetic data for privacy protection

Swapping started in 1960

- First year that Census data were made available on punch cards and magnetic tape for research.

Goal: Protect households that are outliers

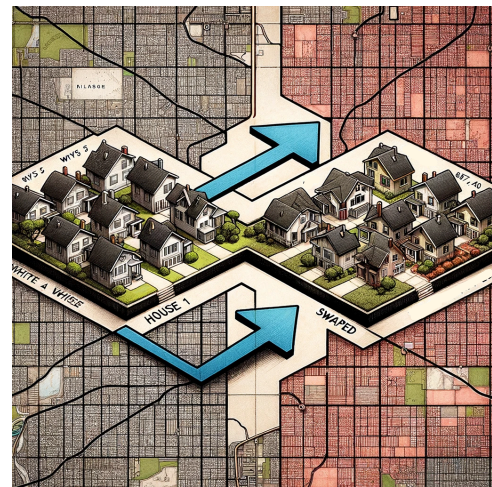
- e.g. the only house on a block, racial outliers, etc.
- The actual swapping mechanism and rate are confidential

Synthetic data – for group quarters

- e.g. dorms, military bases, nursing homes, monasteries, etc.

The Census Bureau had concerns with swapping.

- High rates did obvious damage to the data that were readily apparent.
- Low swap rates did not [really] protect privacy; high rates would result in damage.
- Swapping only protected swapped households.
- But... a non-zero swap rate provides deniability



2016 — The Census Bureau moves to Differential Privacy

2016 — John Abowd becomes Chief Scientist & Dan Kifer joins for his sabbatical.

2016 — Tammy Adams reconstructs microdata for Fairfax County

- Shows that the 2010 Census privacy protection mechanism was vulnerable by applying “database reconstruction” to the published tables.

2017 — I start as Chief of the Center for Disclosure Avoidance Research.

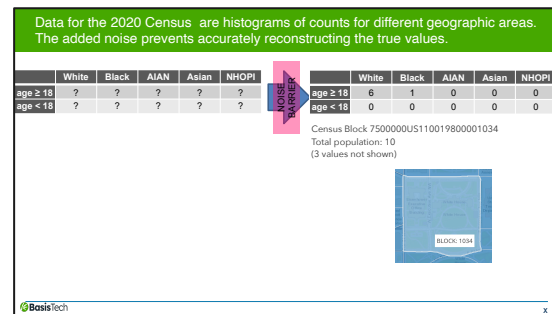
My mission — make formally private:

- 2020 Census — 10 year census of population and housing
- 2022 Economic Census — 5 year survey of establishments
- American Community Survey (ACS) — Annual survey of population and housing
- American Housing Survey — Annual survey of housing units
- Ad hoc disclosure avoidance for research products from

To make the 2020 census differentially private:

1) Tabulate people into histograms. 2) Make histograms private. 3) Tabulate them.

Remember the block with the white house on it?



Complication 1 – How do we make them private?

- We need to understand the entire data pipeline
- Example – Each data product has a different histogram.

–Redistricting products – 2 age categories x 63 race categories x 2 ethnicity = 252 cells

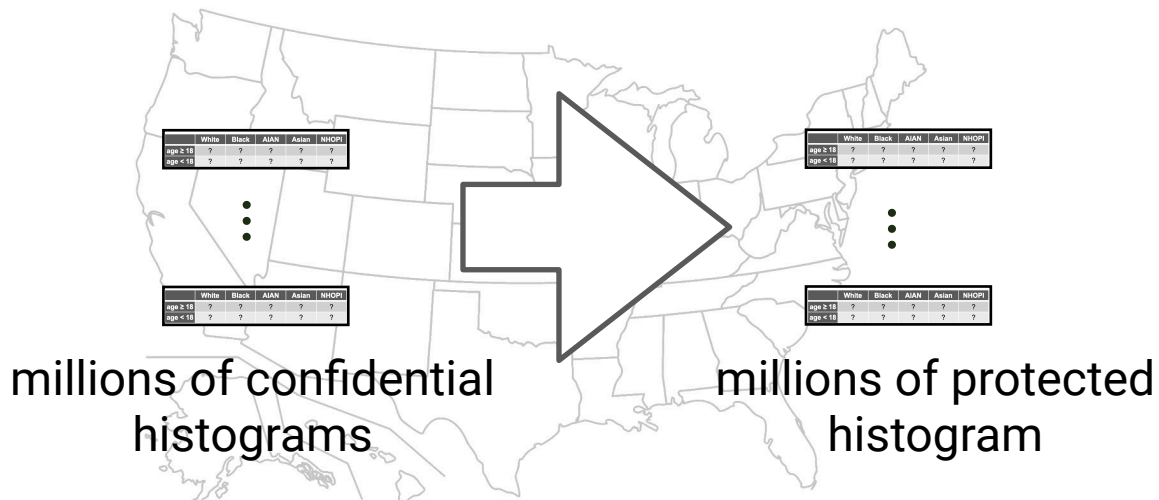
–Demographic – ~100 age categories x 63 race categories x 2 ethnicity x 2 sex x 17 family role ~ = 428,000 cells

Complication 2 – What do we use for test data?

We had to build the mechanisms before we knew the final histograms. How should we make the histograms private?

Naive approach: block-by-block

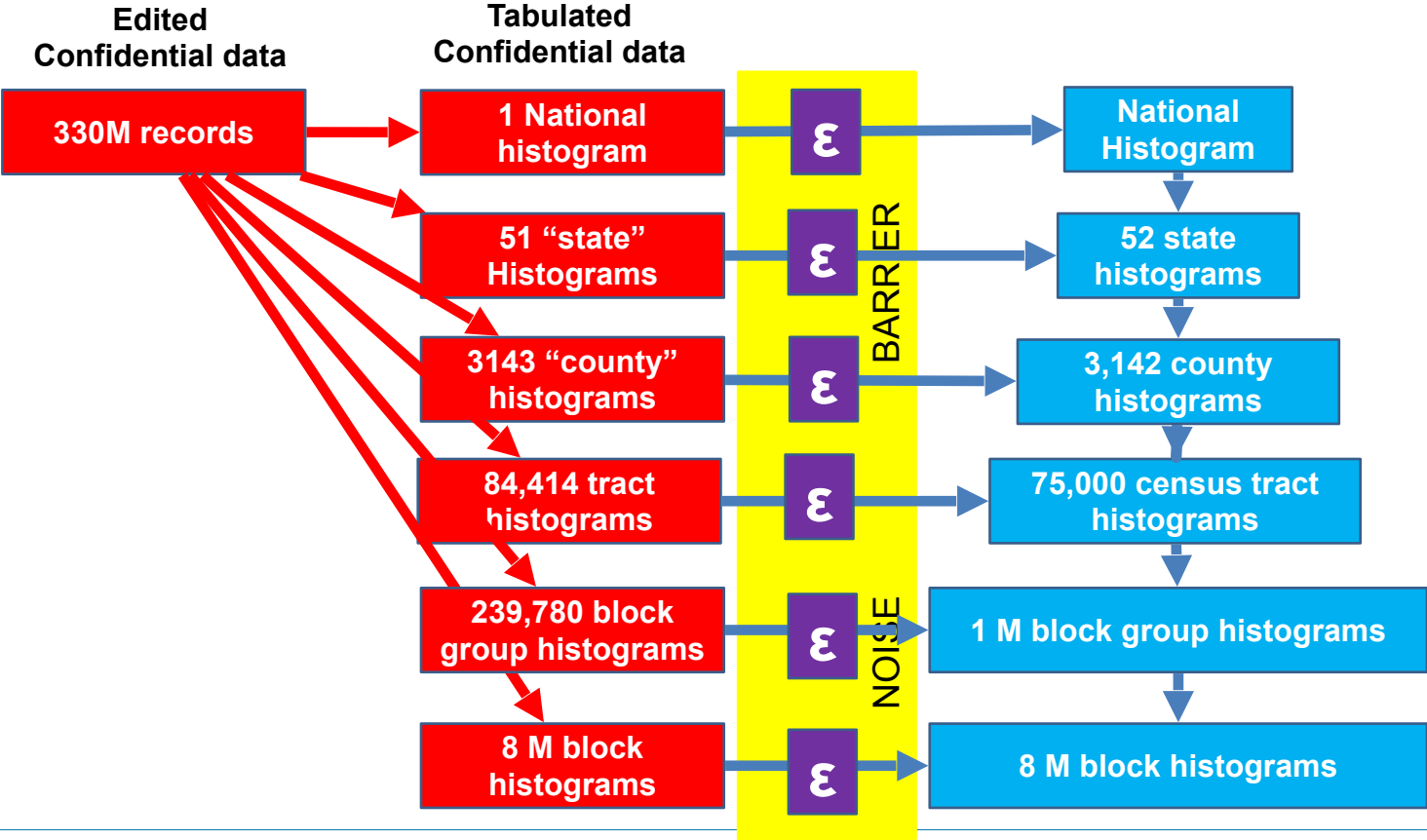
- Add noise to each cell in each histogram.
- Adjust each cell so that it non-negative and integer
- Adjust each histogram so that the total number remains constant. (Is this DP? Discuss!)



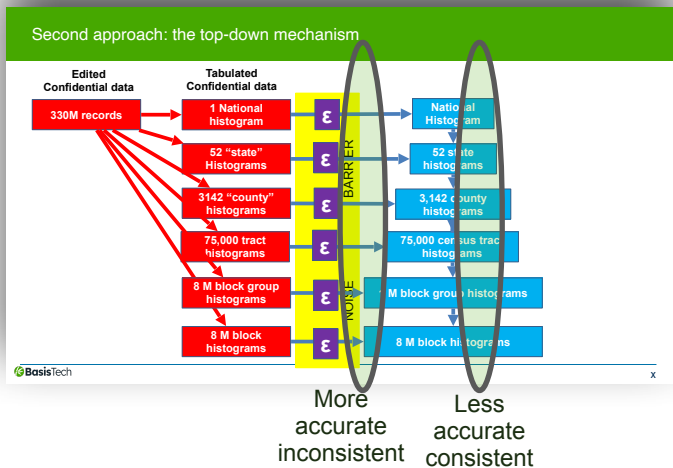
This is “local differential privacy” applied to blocks, rather than people.

Preferred approach: the top-down mechanism

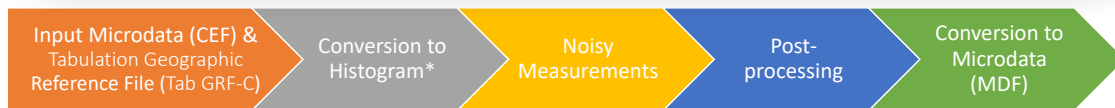
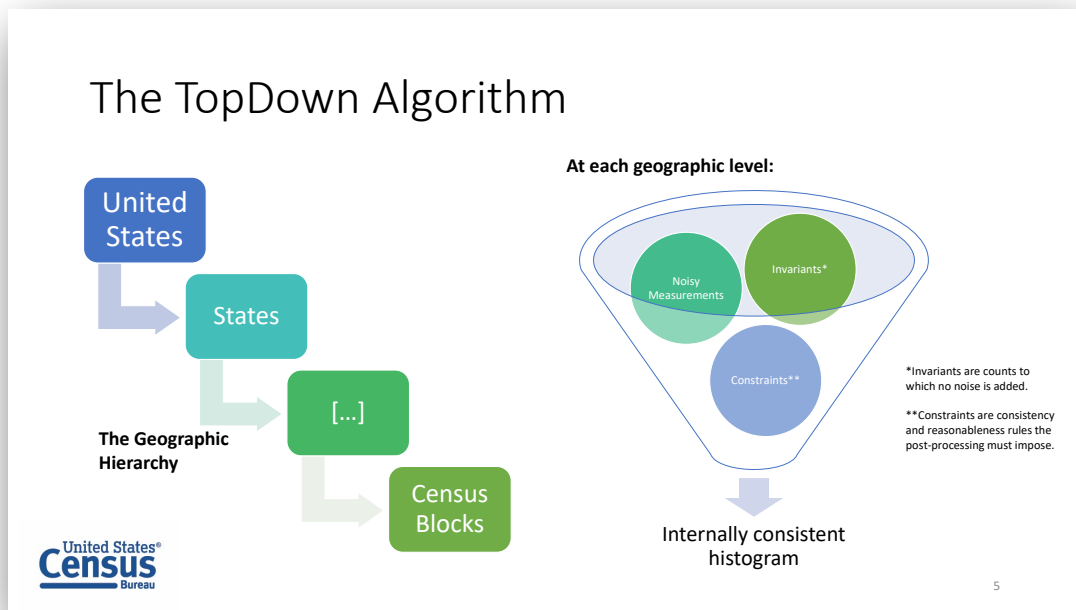
Each histogram provides statistical accuracy to those underneath.



The final visual language

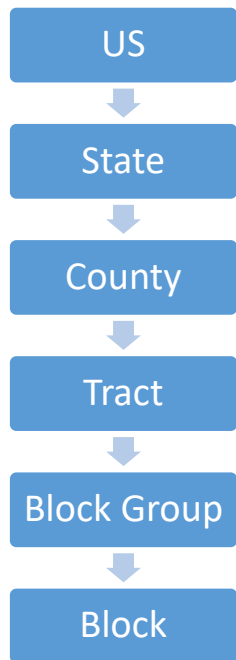


The TopDown Algorithm

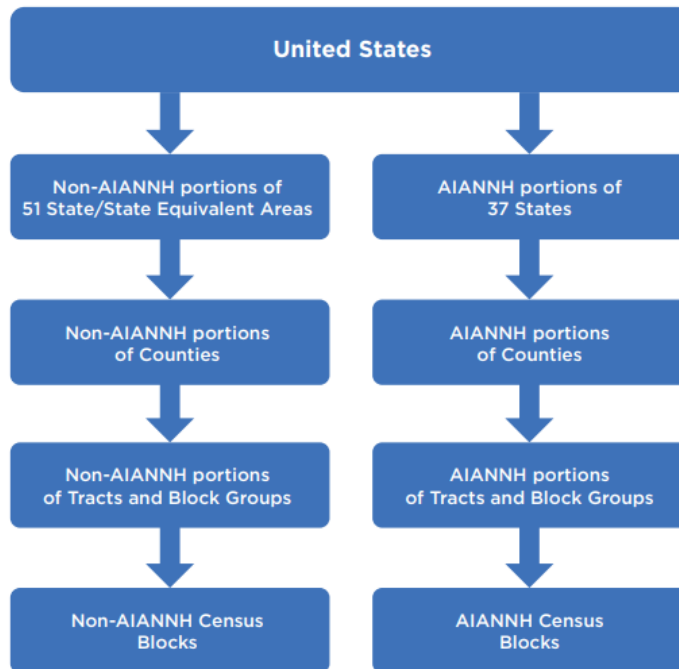


The final geography provides for separate statistics for the US regions designated for American Indians and Native Americans.

Standard Spine Tabulation Hierarchy



TDA's American Indian/Alaska Native/Native Hawaiian (AIANNH) Spine for Redistricting



The final privacy loss budget and query set

Global <i>rho</i>	2.56
Global <i>epsilon</i>	17.90
<i>delta</i>	10 ⁻¹⁰

	<i>rho</i> Allocation by Geographic Level
US	2.54%
State	35.13%
County	10.91%
Tract	16.76%
Optimized Block Group*	30.64%
Block	4.03%

Production settings for the
2020 Census Redistricting
Data (P.L. 94-171)
Summary File
(Persons tables P1-P5)

Query	Per Query <i>rho</i> Allocation by Geographic Level					
	US	State	County	Tract	Optimized Block Group*	Block
TOTAL (1 cell)		32.35%	8.32%	6.40%	12.75%	0.00%
CENRACE (63 cells)	0.03%	0.05%	0.03%	0.03%	0.02%	0.01%
HISPANIC (2 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
VOTINGAGE (2 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
HHINSTLEVELS (3 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
HHGQ (8 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
HISPANIC*CENRACE (126 cells)	0.08%	0.10%	0.07%	7.90%	7.89%	0.02%
VOTINGAGE*CENRACE (126 cells)	0.08%	0.10%	0.07%	0.08%	0.07%	0.02%
VOTINGAGE*HISPANIC (4 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
VOTINGAGE*HISPANIC*CENRACE (252 cells)	0.27%	0.29%	0.27%	0.27%	0.18%	0.07%
HHGQ*VOTINGAGE*						
HISPANIC*CENRACE (2,016 cells)	1.99%	1.97%	2.01%	1.97%	9.63%	3.88%

Internal Challenges

Internal challenges were in three main areas:

Census Bureaucratic Challenges

- FISMA (Federal Information Security Modernization Act)

Scientific Challenges

- DP had never been used at this scale before
 - Google's RAPPOR was a large deployment but a simple algorithm*
- We didn't have an algorithm we knew would work!

Engineering Challenges — Build a system that will run reliably, at scale —

- The first time it is run in production (with data collected using a different schema)(
- Without being re-run because of statistical inaccuracy (because of DP guarantees)

Challenge: Finding Data to Develop the Algorithm

January 2017 – Dan Kifer was using the 2010 Census data on a research cluster.

“2010 Census Data” – There were many datasets

- OPS* – Operational File confidential (T13)
- CUF – Census Unedited File confidential (T13)
- HDF – “Hundred percent file” confidential (T13)
- CEF – Census Edited File confidential (T13)
- Published microdata public; swapped; sampled; no addresses (PUMAs)
- Published Tables public; swapped; not record-level

Census 2020 policy prohibited developing operational code with Title 13 data.

I spent months trying to find appropriate synthetic data,
while simultaneously arguing that no such synthetic data existed.

Synthetic data had to:

- Represent the entire US – Rural, Urban, and everything in between
- Be diverse and complex with respect to race, age, households, concentrations, mixing
- Not reveal private, protected information (or else it would be confidential too)

Observation #1 –

- If we could make adequate synthetic data, we wouldn't need to create the DP system!

Observation #2 –

- Making synthetic data *was in fact that we were doing with the DP project!*

Resolving the challenge of developing code with confidential data

We had to transition from the research cluster to the AWS Cloud

- The research cluster was due to be decommissioned
- The cluster didn't have enough compute power
- The 2020 Census had to run in the AWS Cloud

Working in the AWS Cloud with confidential data required:

- ATT – Authority To Test
- ATO – Authority To Operate
- [No such thing as ATD – Authority to Develop]

Required – Documentation, Engineering Plans, Security Plan, etc.

- FISMA – Federal Information Standards Management Act

Challenge: Developing and Auditing a Randomized algorithm

Evaluating the correctness of our runs

- Unit tests

 - What do you test?

 - What are the metrics beyond non-crashing and code coverage?

- Repeatable random numbers

 - “Anyone who considers arithmetical methods for producing random digits is, of course, in a state of sin.” – von Neumann

- Code auditing

 - Galois & MITRE

Evaluating the statistical accuracy of runs...

- What is our definition of accuracy?

- How do we share these results with our outside collaborators?

Evaluating the statistical accuracy of a randomized algorithms: We had two choices.

Choice #1 – Develop a theoretical framework for error injection and propagation.

- Technically difficult to do with the complex TopDownAlgorithm.

Chose #2 – Perform multiple runs of the program and report:

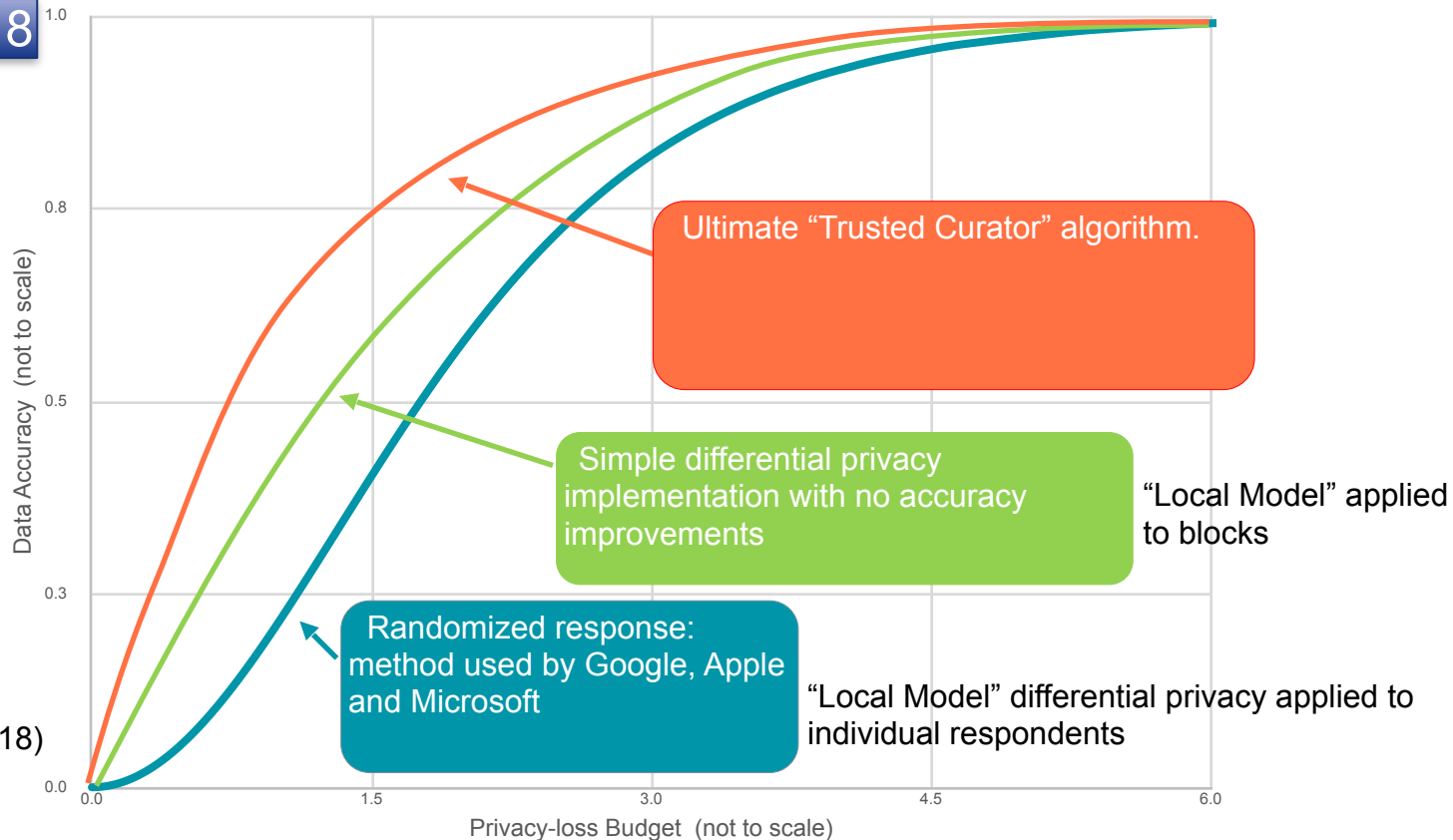
- the variance between runs
- The accuracy of each run.
- the average of the run accuracies.

We could do this for the 2010 data, but not for the 2020 data

- 2010 – not formally private
- 2020 – Each run draws down the privacy budget, even if we only report a single number.

We wanted the most efficient DP algorithm — the best data accuracy for any given privacy loss budget.

Summer 2018



(notional graph c. 2018)

We realized that we could use data from the 1940 Census!

In the US, Census records are only protected for 72-years.

Advantages:

- Microdata downloadable from IPUMS
- No privacy concerns

Disadvantages:

- Different geography
 - Nation - State - County - Enumeration District*
 - vs. Nation - State - County - Track - Block Group - Block*
- Different Races in official Census
- Troubling history of 1940 Census

NATIONAL

The 1940 Census: 72-Year-Old Secrets Revealed

APRIL 2, 2012 · 7:49 AM ET

By [Linton Weeks](#)



An enumerator interviews a woman for the 1940 census. Veiled in secrecy for 72 years because of privacy protections, the 1940 U.S. census is the first historical federal decennial survey to be made available on the Internet initially rather than on microfilm.

National Archives at College Park

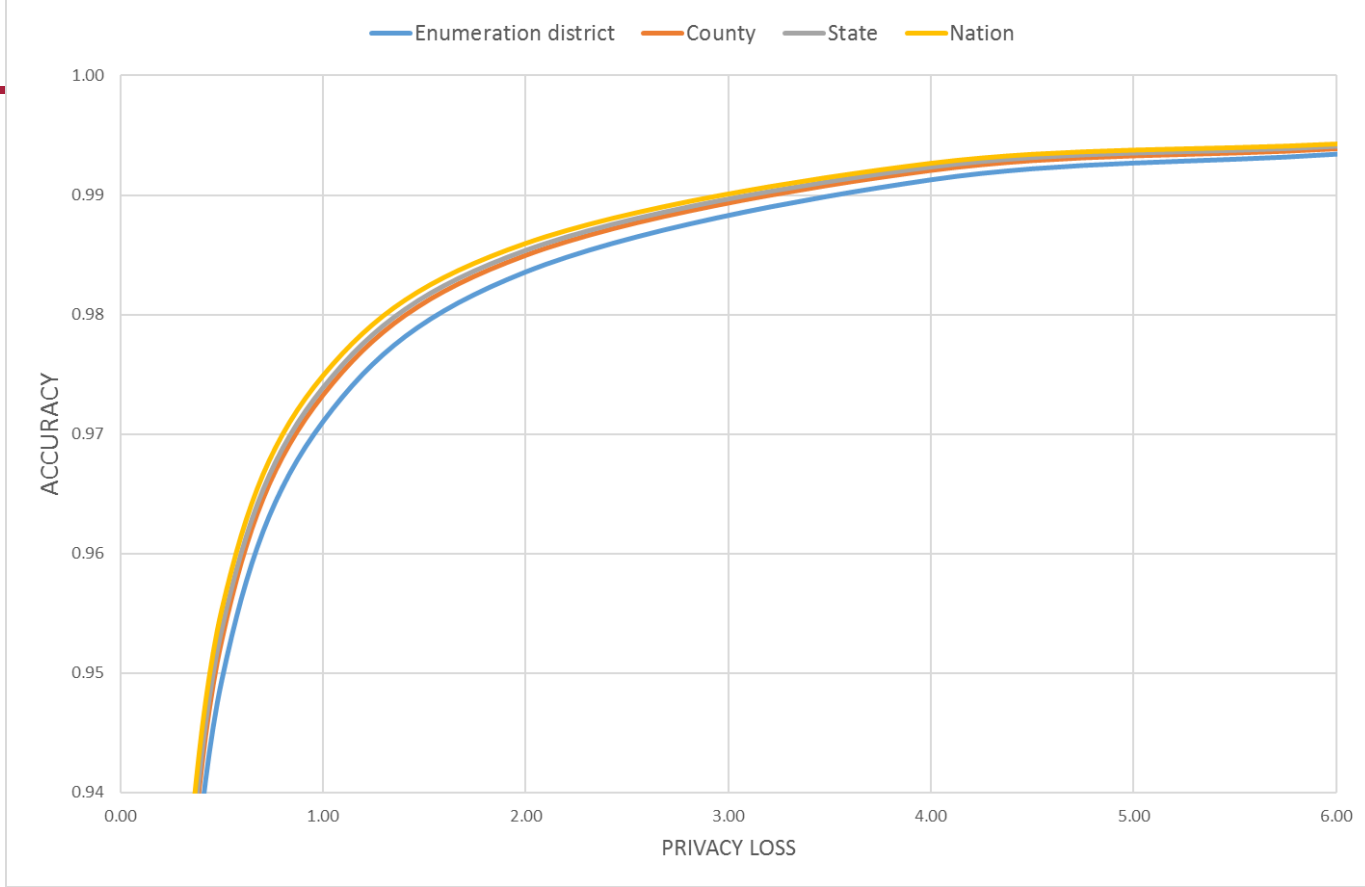
Tested with data from 1940

1940 hierarchy:

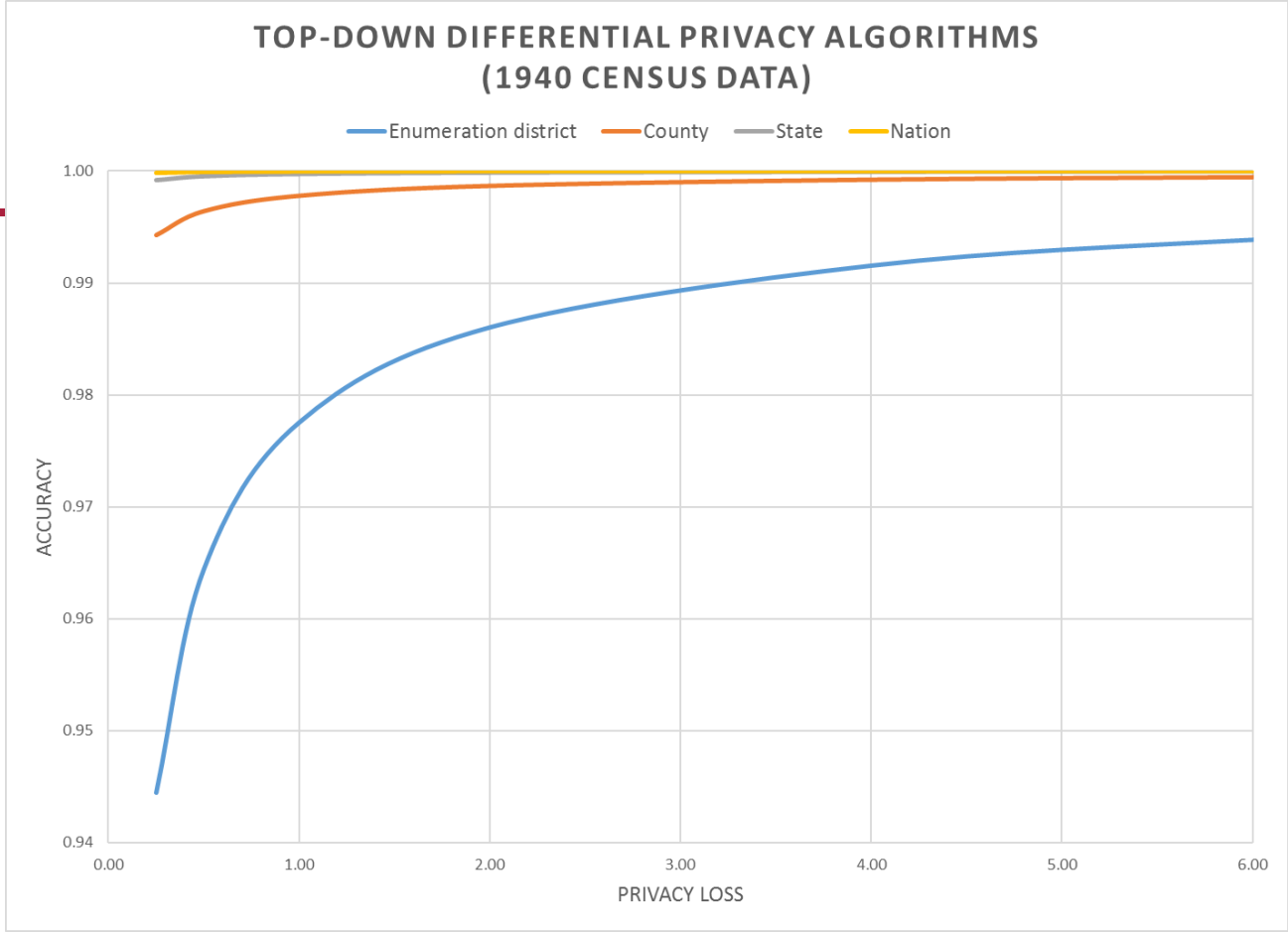
- Nation
- State
- County
- Enumeration District

Download from usa.ipums.org

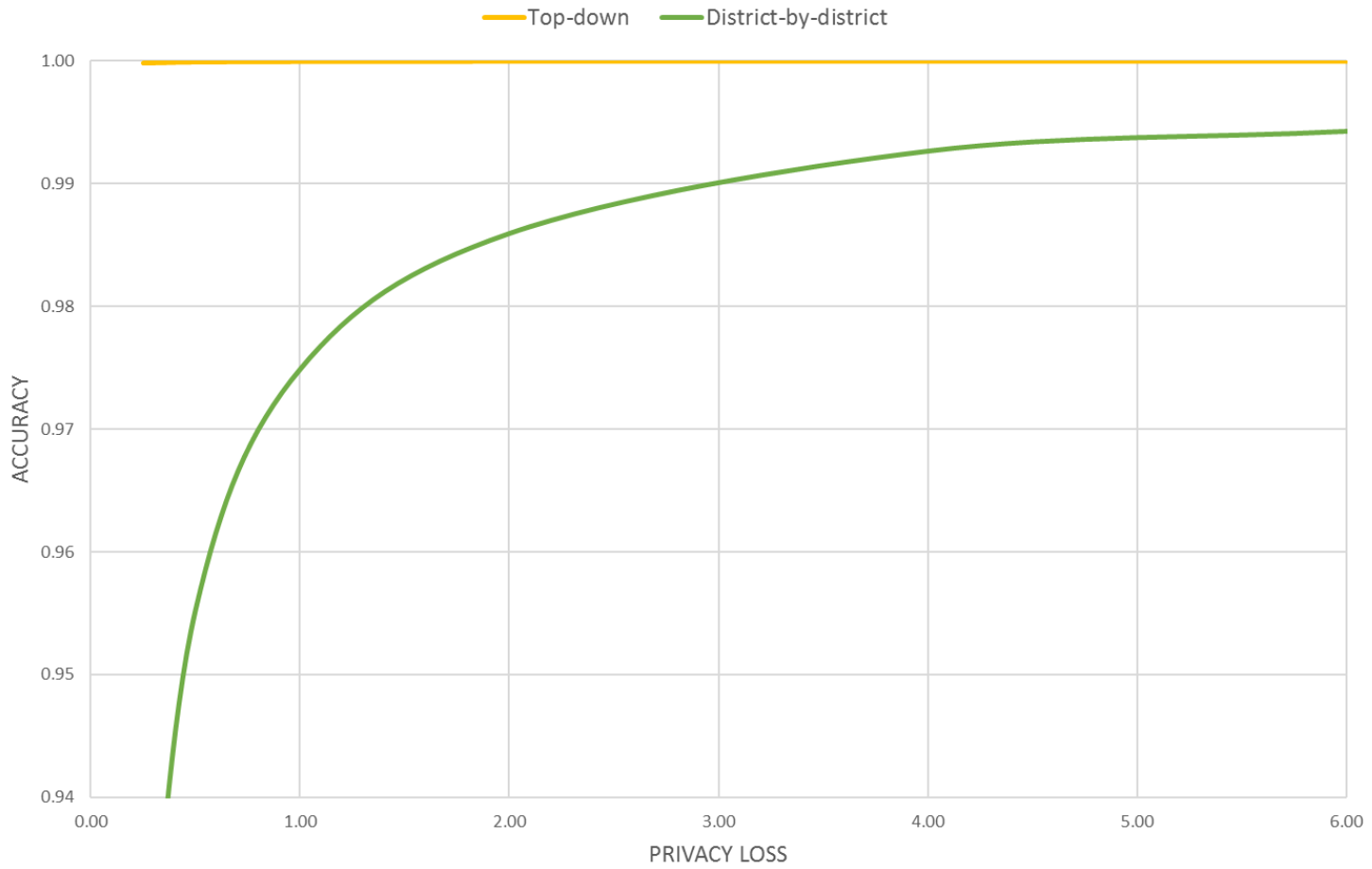
DISTRICT-BY-DISTRICT DIFFERENTIAL PRIVACY ALGORITHMS (1940 CENSUS DATA)



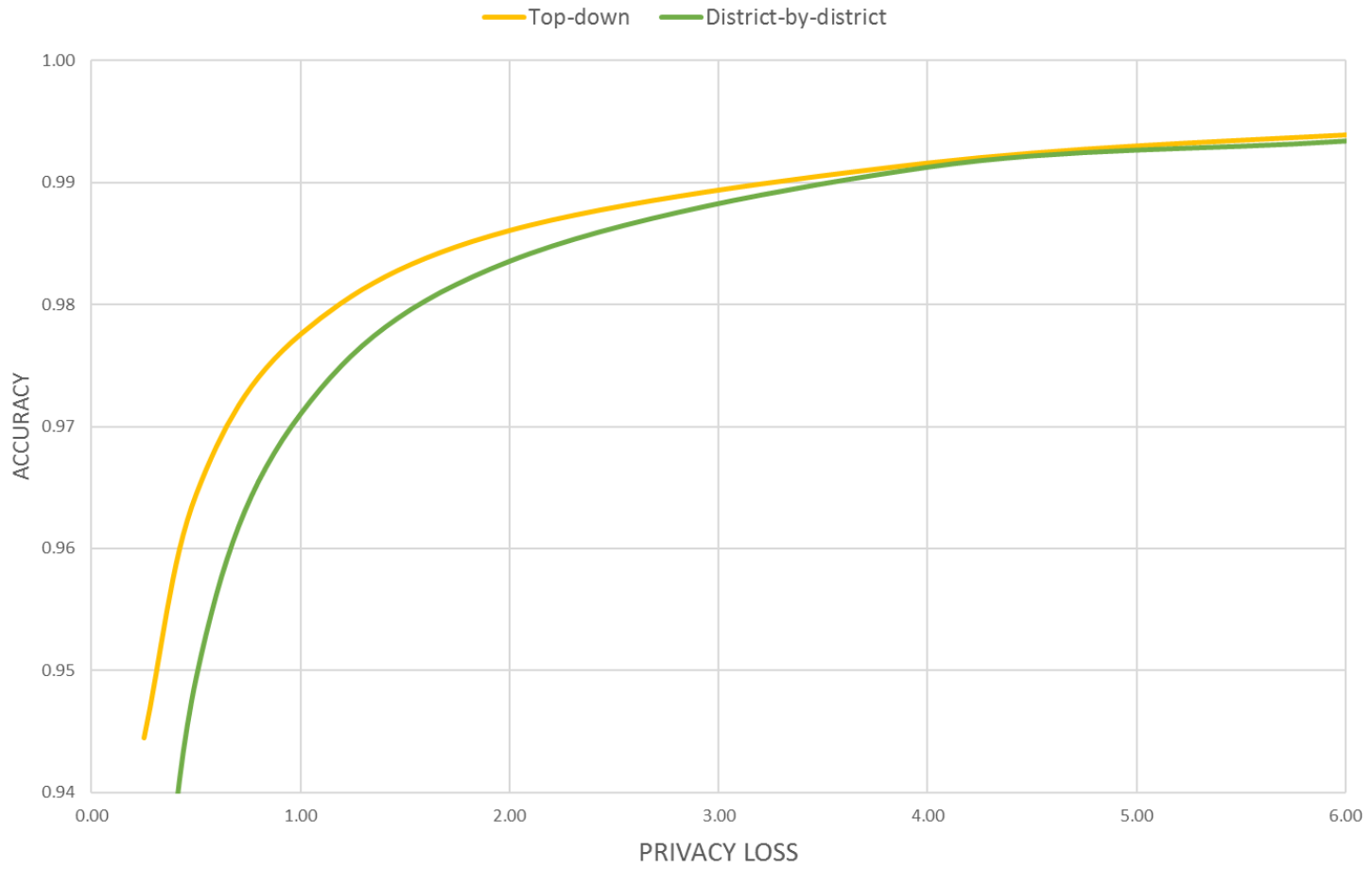
Top-Down: much more accurate!



COMPARISON OF NATIONAL RESULTS BY ALGORITHM (1940 CENSUS DATA)



COMPARISON OF DISTRICT RESULTS BY ALGORITHM (1940 CENSUS DATA)



Policy Issues for the 2020 Census: Invariants

For the 2018 End-to-End test, policy makers wanted exact counts:

- Number of people on each block
- Number of people on each block of voting age
- Number of residences & group quarters on each block

We implemented invariants before we understood their mathematical impact on differential privacy semantics. We then scaled back to four invariants:

- C1: Total population (invariant at the county level for the 2018 E2E)
- ~~C2: Voting age population (population age 18 and older) (eliminated for the 2018 E2E)~~
- C3: Number of housing units (invariant at the block level)
- C4: Number of occupied housing units (invariant at the block level)
- C5: Number of group quarters facilities by group quarters type (invariant at the block level)

Scientific Issues for the 2020 Census: Person-Household Joins

The Census creates two kinds of tables:

Person tables

Household tables

We can create P & H today.

We are working on P x H and Detailed P, H

Q(P): # of men living on a block.

Q(H): # of occupied houses on a block.

Q(P x H): # of children in houses headed by a single man.

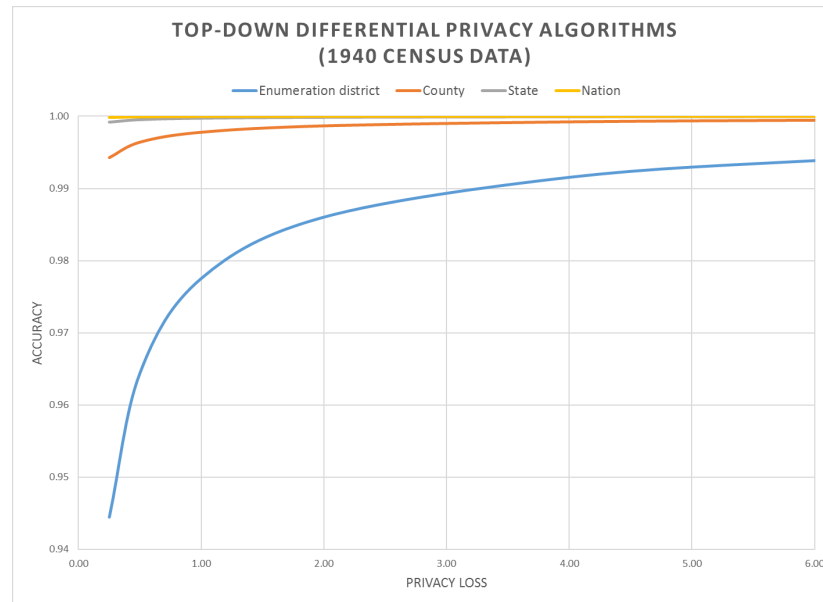
Scientific Issue for any use of DP: Quality Metrics

What is the measure of “quality” or “utility” in a complex data product?

Options:

- L1 error between “true” data set and “protected” data set

- Impact on an algorithm that uses the data (e.g., redistricting and Voting Rights Act enforcement)



The Choice Problem for Redistricting Tabulations Is More Challenging

In the redistricting application, the fitness-for-use is based on :

Supreme Court one-person one-vote decision (All legislative districts must have approximately equal populations; there is judicially approved variation)

Is statistical disclosure limitation a “statistical method” (permitted by *Utah v. Evans*) or “sampling” (prohibited by the Census Act, confirmed in *Commerce v. House of Representatives*)?

Voting Rights Act, Section 2: requires majority-minority districts at all levels, when certain criteria are met

The privacy interest is based on:

Title 13 requirement not to publish exact identifying information

The public policy implications of uses of detailed race, ethnicity and citizenship

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.

We decided to release multiple datasets and hold a workshop

External Challenges

External Chronology (Hotz and Salvo 2022)

2016 – Sept

- John Abowd “presented a case for a new approach to protecting the privacy of respondents to the Census Scientific Advisory Committee (CSAC)”

2017 – Garfinkel presents to CSAC - DP is the plan.

2018 – DP is implemented for the 2018 End-to-End test

- DP is justified because of the reconstruction attack.
- July – Notice in federal register “Soliciting Feedback from Users on 2020 Census Data Products. “This request engendered a sense of bewilderment on the part of data users and triggered a litany of concerns about 2020 Census content that was clearly at risk.”
- Dec – DP incorporated into 4.0 “2020 Census Operational Plan”

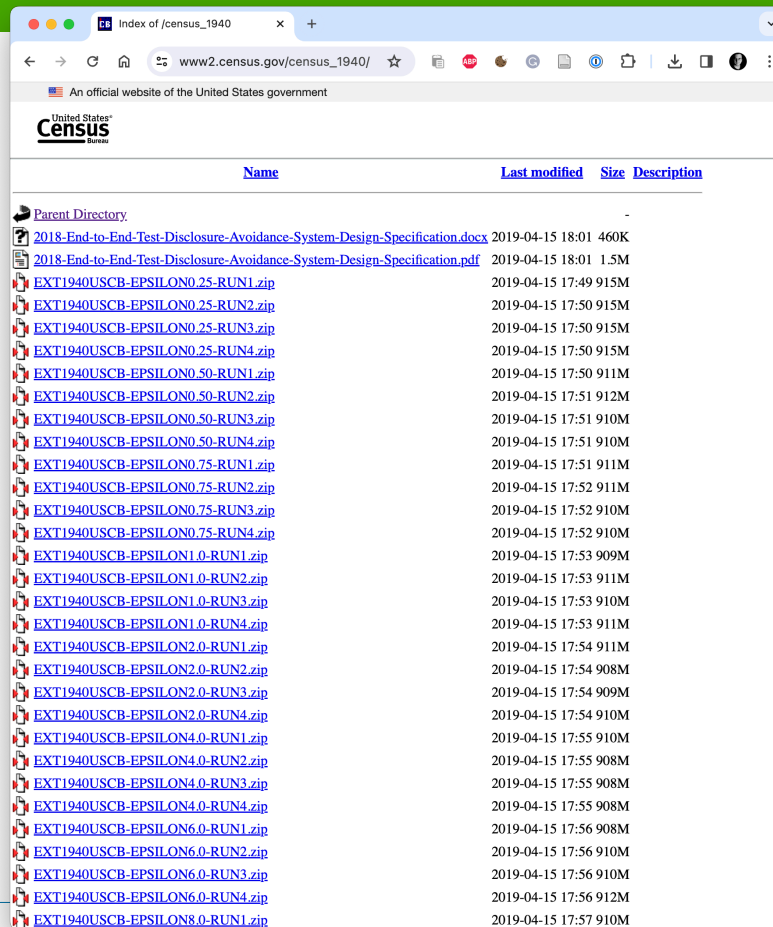
2019 – Dept. Dir. Ron Jarmin announces 2020 will use DP

- Dec 11-12 – CNSTAT workshop, “2020 Census Data Products: Data Needs and Privacy Considerations,”



<https://hdsr.mitpress.mit.edu/pub/ql9z7ehf/release/8>

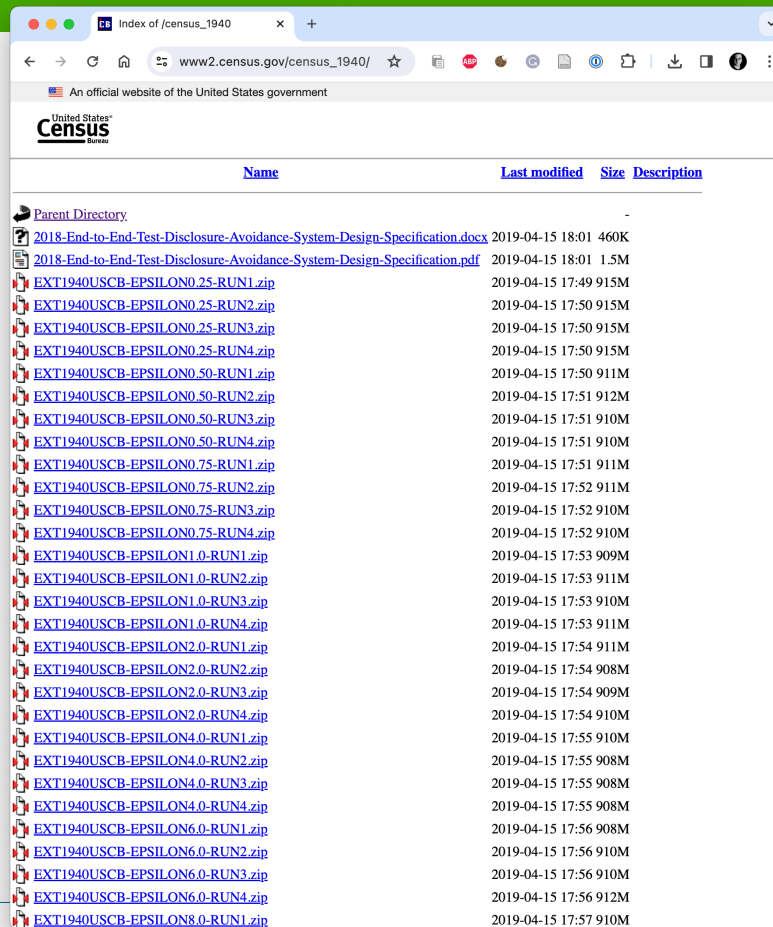
Multiple releases of 1940 data run through the DAS



The screenshot shows a web browser window displaying the directory listing for the 1940 census data on the Census Bureau website. The browser address bar shows the URL www2.census.gov/census_1940/. The page header includes the United States Census Bureau logo and the text "An official website of the United States government".

Name	Last modified	Size	Description
Parent Directory	-	-	-
2018-End-to-End-Test-Disclosure-Avoidance-System-Design-Specification.docx	2019-04-15 18:01	460K	
2018-End-to-End-Test-Disclosure-Avoidance-System-Design-Specification.pdf	2019-04-15 18:01	1.5M	
EXT1940USCB-EPSILON0.25-RUN1.zip	2019-04-15 17:49	915M	
EXT1940USCB-EPSILON0.25-RUN2.zip	2019-04-15 17:50	915M	
EXT1940USCB-EPSILON0.25-RUN3.zip	2019-04-15 17:50	915M	
EXT1940USCB-EPSILON0.25-RUN4.zip	2019-04-15 17:50	915M	
EXT1940USCB-EPSILON0.50-RUN1.zip	2019-04-15 17:50	911M	
EXT1940USCB-EPSILON0.50-RUN2.zip	2019-04-15 17:51	912M	
EXT1940USCB-EPSILON0.50-RUN3.zip	2019-04-15 17:51	910M	
EXT1940USCB-EPSILON0.50-RUN4.zip	2019-04-15 17:51	910M	
EXT1940USCB-EPSILON0.75-RUN1.zip	2019-04-15 17:51	911M	
EXT1940USCB-EPSILON0.75-RUN2.zip	2019-04-15 17:52	911M	
EXT1940USCB-EPSILON0.75-RUN3.zip	2019-04-15 17:52	910M	
EXT1940USCB-EPSILON0.75-RUN4.zip	2019-04-15 17:52	910M	
EXT1940USCB-EPSILON1.0-RUN1.zip	2019-04-15 17:53	909M	
EXT1940USCB-EPSILON1.0-RUN2.zip	2019-04-15 17:53	911M	
EXT1940USCB-EPSILON1.0-RUN3.zip	2019-04-15 17:53	910M	
EXT1940USCB-EPSILON1.0-RUN4.zip	2019-04-15 17:53	911M	
EXT1940USCB-EPSILON2.0-RUN1.zip	2019-04-15 17:54	911M	
EXT1940USCB-EPSILON2.0-RUN2.zip	2019-04-15 17:54	908M	
EXT1940USCB-EPSILON2.0-RUN3.zip	2019-04-15 17:54	909M	
EXT1940USCB-EPSILON2.0-RUN4.zip	2019-04-15 17:54	910M	
EXT1940USCB-EPSILON4.0-RUN1.zip	2019-04-15 17:55	910M	
EXT1940USCB-EPSILON4.0-RUN2.zip	2019-04-15 17:55	908M	
EXT1940USCB-EPSILON4.0-RUN3.zip	2019-04-15 17:55	908M	
EXT1940USCB-EPSILON4.0-RUN4.zip	2019-04-15 17:55	908M	
EXT1940USCB-EPSILON6.0-RUN1.zip	2019-04-15 17:56	908M	
EXT1940USCB-EPSILON6.0-RUN2.zip	2019-04-15 17:56	910M	
EXT1940USCB-EPSILON6.0-RUN3.zip	2019-04-15 17:56	910M	
EXT1940USCB-EPSILON6.0-RUN4.zip	2019-04-15 17:56	912M	
EXT1940USCB-EPSILON8.0-RUN1.zip	2019-04-15 17:57	910M	

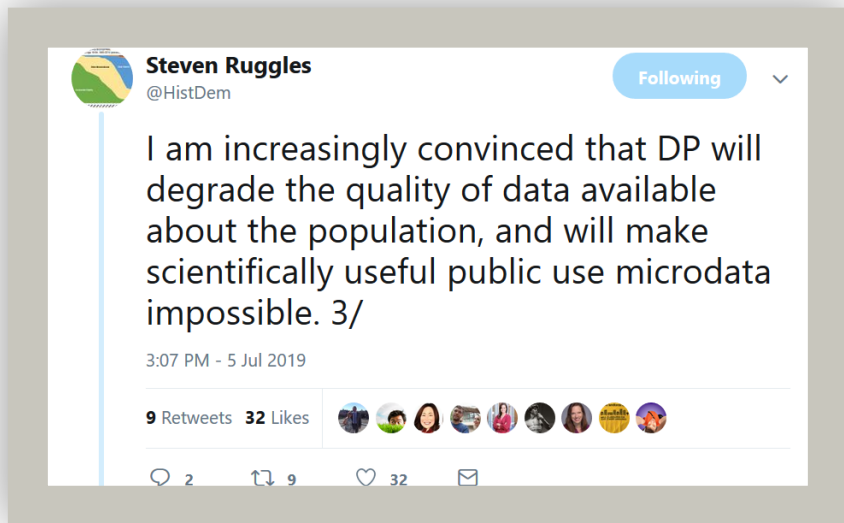
Multiple releases of 1940 data run through the DAS



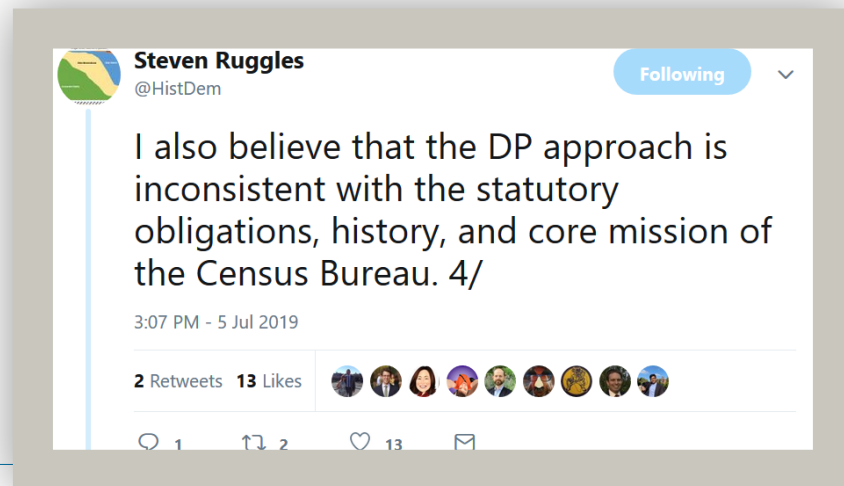
The screenshot shows a web browser window displaying a directory listing of files on the website www2.census.gov/census_1940/. The browser's address bar shows the URL and the page title is "Index of /census_1940". The page content includes the United States Census Bureau logo and a table of files. The table has columns for "Name", "Last modified", "Size", and "Description". The files listed are primarily zip files representing different runs of the 1940 census data, organized by epsilon level (0.25, 0.50, 0.75, 1.0, 2.0, 4.0, 6.0, 8.0) and run number (1, 2, 3, 4). There are also two PDF files at the top of the list.

Name	Last modified	Size	Description
Parent Directory	-	-	-
2018-End-to-End-Test-Disclosure-Avoidance-System-Design-Specification.docx	2019-04-15 18:01	460K	
2018-End-to-End-Test-Disclosure-Avoidance-System-Design-Specification.pdf	2019-04-15 18:01	1.5M	
EXT1940USCB-EPSILON0.25-RUN1.zip	2019-04-15 17:49	915M	
EXT1940USCB-EPSILON0.25-RUN2.zip	2019-04-15 17:50	915M	
EXT1940USCB-EPSILON0.25-RUN3.zip	2019-04-15 17:50	915M	
EXT1940USCB-EPSILON0.25-RUN4.zip	2019-04-15 17:50	915M	
EXT1940USCB-EPSILON0.50-RUN1.zip	2019-04-15 17:50	911M	
EXT1940USCB-EPSILON0.50-RUN2.zip	2019-04-15 17:51	912M	
EXT1940USCB-EPSILON0.50-RUN3.zip	2019-04-15 17:51	910M	
EXT1940USCB-EPSILON0.50-RUN4.zip	2019-04-15 17:51	910M	
EXT1940USCB-EPSILON0.75-RUN1.zip	2019-04-15 17:51	911M	
EXT1940USCB-EPSILON0.75-RUN2.zip	2019-04-15 17:52	911M	
EXT1940USCB-EPSILON0.75-RUN3.zip	2019-04-15 17:52	910M	
EXT1940USCB-EPSILON0.75-RUN4.zip	2019-04-15 17:52	910M	
EXT1940USCB-EPSILON1.0-RUN1.zip	2019-04-15 17:53	909M	
EXT1940USCB-EPSILON1.0-RUN2.zip	2019-04-15 17:53	911M	
EXT1940USCB-EPSILON1.0-RUN3.zip	2019-04-15 17:53	910M	
EXT1940USCB-EPSILON1.0-RUN4.zip	2019-04-15 17:53	911M	
EXT1940USCB-EPSILON2.0-RUN1.zip	2019-04-15 17:54	911M	
EXT1940USCB-EPSILON2.0-RUN2.zip	2019-04-15 17:54	908M	
EXT1940USCB-EPSILON2.0-RUN3.zip	2019-04-15 17:54	909M	
EXT1940USCB-EPSILON2.0-RUN4.zip	2019-04-15 17:54	910M	
EXT1940USCB-EPSILON4.0-RUN1.zip	2019-04-15 17:55	910M	
EXT1940USCB-EPSILON4.0-RUN2.zip	2019-04-15 17:55	908M	
EXT1940USCB-EPSILON4.0-RUN3.zip	2019-04-15 17:55	908M	
EXT1940USCB-EPSILON4.0-RUN4.zip	2019-04-15 17:55	908M	
EXT1940USCB-EPSILON6.0-RUN1.zip	2019-04-15 17:56	908M	
EXT1940USCB-EPSILON6.0-RUN2.zip	2019-04-15 17:56	910M	
EXT1940USCB-EPSILON6.0-RUN3.zip	2019-04-15 17:56	910M	
EXT1940USCB-EPSILON6.0-RUN4.zip	2019-04-15 17:56	912M	
EXT1940USCB-EPSILON8.0-RUN1.zip	2019-04-15 17:57	910M	

Early attacks against differential privacy in the 2020 Census



Steven Ruggles 5 Jul 2019



Organized attack on the move to differential privacy

Ruggles:

- “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”

Action:

- Organized petition with 4000+ signers asking for no DP in 2020 Census.

Results:

- ... ?

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

Analysis of population variances

David Van Riper & Tracy Kugler, IPUMS (APDU 2019)

Note:

- Epsilon 0.25 .. 8.0
- Highly accurate when $n > 1000$
- Less accurate when $n < 1000$
- accuracy \sim size \sim ethnicity

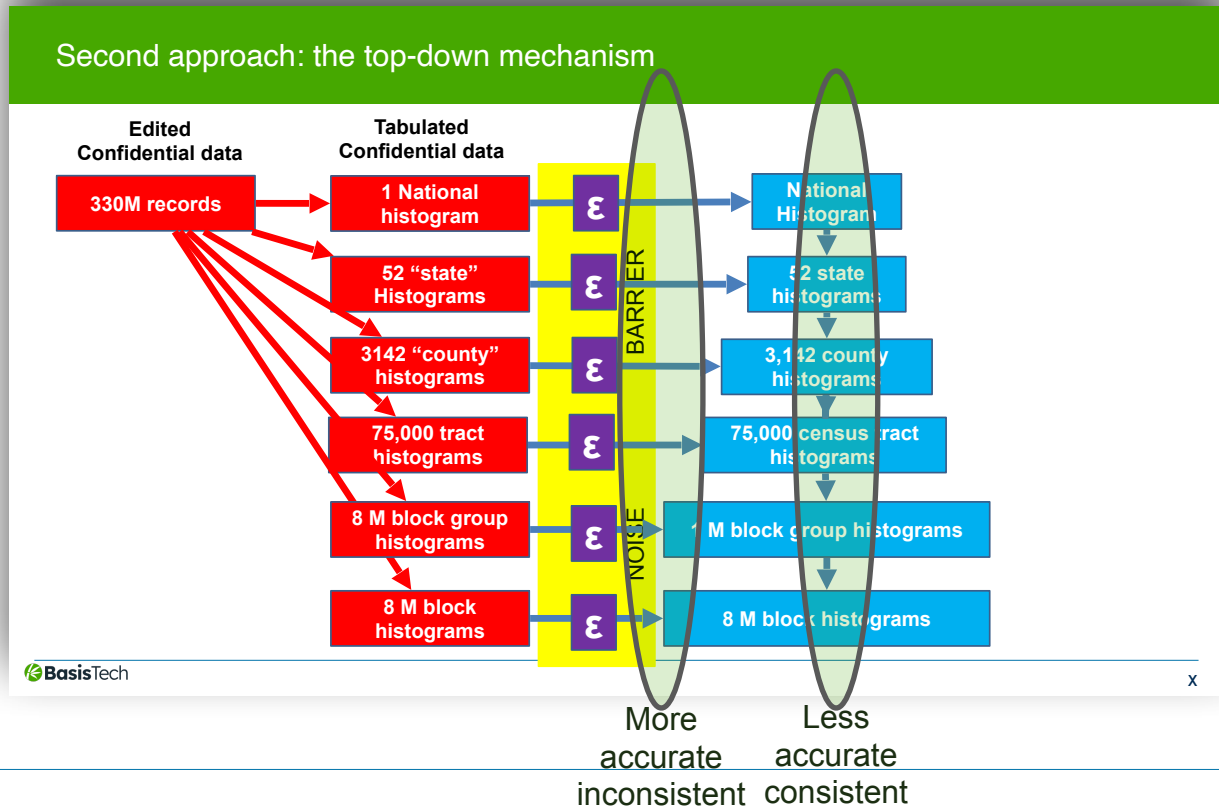


Analysis of population variances

David Van Riper & Tracy Kugler, IPUMS (APDU 2019)

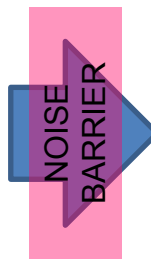


The source of the inaccuracy: integer non-negative constraints

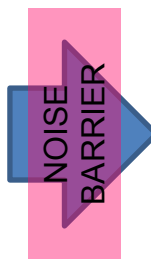


The error comes from enforced consistency

	White	Black	AIAN	Asian	NHOPI
age \geq 18	2	0	0	0	0
age < 18	1	0	0	0	0



	White	Black	AIAN	Asian	NHOPI
age \geq 18	6.8	0.13	-0.025	-0.308	-0.665
age < 18	0.002	-0.177	0.141	-0.107	-0.700



	White	Black	AIAN	Asian	NHOPI
age \geq 18	2.744	-0.901	-0.075	0.627	1.102
age < 18	1.975	-0.207	-1.516	-0.838	-1.892



	White	Black	AIAN	Asian	NHOPI
age \geq 18	3.223	-0.901	-0.753	0.627	-0.590
age < 18	0.148	1.975	-0.207	-1.516	-0.838

Single release of 2010 data for the CNStat meeting

Key observations from 2019 CNSTAT Workshop (Hotz and Salvo)

- “(a.) Population counts for some geographic units and demographic characteristics were not adversely affected by differential privacy.
- “(b.) Concerns with data for small geographic areas and population groups.
- “(c.) The absence of a direct allocation of privacy-loss budget for political and administrative geographic areas, such as places and county subdivisions, or to detailed race groups, such as American Indians.
- “(d.) Problems for temporal consistency of population counts.
- “(e.) Unexpected issues with the postprocessing of the proposed DAS.
- “(f.) Difficulties estimating error.
- “(g.) The importance of protecting privacy.”

Progress towards improving accuracy (Hotz and Salvo)

2020 – Extensive work in “tuning” the algorithm

2021 –

- “It was not until the **spring of 2021**, however, that **major improvements in the DAS were seen**, largely as a result of a **big increase in the privacy-loss budget**, with a goal of optimizing the data for use in redistricting.
- “Using the 2010 Census data, the bureau announced the adoption of an ‘accuracy target’ that optimized the DAS for areas of 500 persons or more, which would provide estimates that were within 5 percentage points (of the 2010 published counts) 95% of the time for the largest race group.
- “In addition, methods were improved for the allocation of the increased PLB to legal, administrative, and political geographic areas, which was a big problem with earlier demonstration files.

(Also – Move to zero-Concentrated Differential Privacy)

Lawsuit by Alabama and other states (Hotz and Salvo)

“State of Alabama v. Department of Commerce, sought to prohibit the Census Bureau from delaying the release of the 2020 PL 94-171 data past the congressionally mandated date of March 31, 2021.”

- “plaintiffs took issue with the Census Bureau’s interpretation of Title 13, namely, that it necessitates the use of differential privacy as the mechanism to protect the confidentiality of census responses.”
- “they posited that the application of differential privacy amounts to manipulation of the data used for redistricting purposes, a situation that will bring “significant harm to Alabama.”
- “Finally, the plaintiffs argued that “the Bureau did not provide notice in the Federal Register of its decision to adopt differential privacy for the 2020 census. Nor did it otherwise seek public comment before the decision was made.” ”

“The U.S. District Court ultimately ruled against the plaintiffs in June of 2021 on the grounds that the case lacked merit based on “ripeness,” or the ability of the plaintiffs to demonstrate harm.”

“Moreover, by the plaintiff’s own admission, the issue became moot when the PL94-171 redistricting file was issued in August of 2021.”

More on the legal history

May 2018 — State of Alabama filed suit to block US from using foreign nationals in determining each state's representation in Congress.

- The phrase “differential privacy” does not appear in the original complaint.
- The lawsuit is not successful.

March 2021 — The State of Alabama files suit again, attacking both the use of DP and the decision to delay publishing statistics during the COVID-19 pandemic

- Lawsuit dismissed without prejudice on Sept. 9, 2021 on the grounds that the data products have not yet been published (and therefore Alabama could not demonstrate harm) and because Alabama had known about the intended use of DP *for years* without filing suit.

May 18, 2021 — Fair Lines v. Commerce files suit requesting the release of certain tabular summaries without using the DP framework.

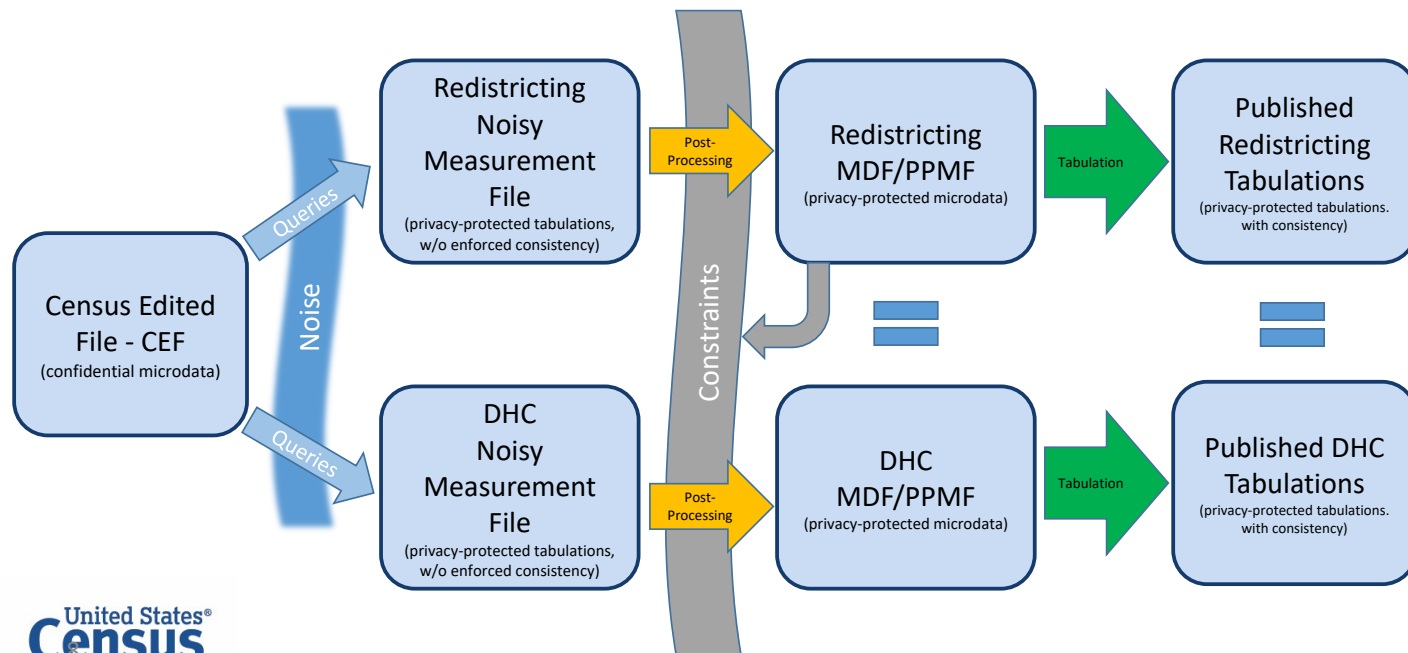
- August 2, 2022 — Fair Lines suit dismissed without prejudice.

One of the distinguishing features of the 2020 Census was the ongoing interference in the census by the President of the United States.

- Executive Order 13880, July 11, 2019 — Link Census data to block-level citizenship data
- Presidential Memorandum of July 21, 2020 to “exclud[e] illegal aliens from the apportionment base following the 2020 Census.”

The Census Bureau released multiple data products for the 2020 census.

Noisy Measurement Files (NMFs), Privacy-Protected Microdata Files (PPMFs), Published Tabulations

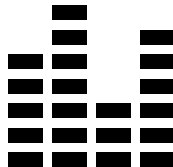


Should I Use the NMF, the PPMF, or the Tabulations?

- There are two sources of error in the published statistics (PPMF and Tabulations):

Differentially private noise

- Unbiased
- Known distribution
- Reflected in the noisy measurements



Post-processing

- Data dependent
 - While the nonnegativity requirement decreases error in the detailed cell counts, it also introduces a positive bias in small counts and an offsetting negative bias in large counts.
 - TDA also reduces the amount of error for many statistics relative to their corresponding noisy measurements.
- Block-level statistics will often have a lower expected variation than you would expect based solely on the amount of PLB assigned to that query at the block level.

Research on the 2020 Census DP project

(Mostly from Radway & Christ 2023, “The Impact of De-Identification on Single-Year-of-age-counts in the US Census.”)

Swapping unique rows caused significant impact on utility – (Kim 2015).

“Low swap rates have essentially no impact on re-identification outcomes” and “high swap rates have only minimal impact” (Hawes and Rodriguez 2021a, 24).

“DP census data is still fit for use in redistricting” (Cohen et al. 2021)

Error from DP small compared to other sources of error. (Steed et al. 2022)

TDA performs poorly for smaller subpopulations and racial minority groups (Kenny et al. 2023).

2020 Census Disclosure Avoidance System Development & Release Timeline (June 30, 2023)

- <https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/das-development-timeline.pdf>

Summary of Public Feedback on the 2010 Demonstration Data Product - Demographic and Housing Characteristics File (August 25, 2022)

- https://www2.census.gov/programs-surveys/decennial/2020/program-management/round_2_feedback.pdf

Empirical study of two aspects of the TopDown Algorithm output for redistricting: Reliability and Variability, Tommy Wright (May 18, 2021)

- <https://www.census.gov/library/working-papers/2021/adrm/SSS2021-02.html>

Personal Reflections

There is a lot more diversity in the data than people realize.

“About 57 percent of the 2010 Census population were ‘unique’ at the smallest census geography, block level, meaning they were the only people in their block with a specific combination of sex, age (in years), race (any of the 63 possible Office of Management and Budget race combinations), and Hispanic/Latino ethnicity” (McKenna 2018).

On May 25, 2021, the Census Bureau released to the Census Scientific Advisory Committee the results of an experiment of applying the suppression rules from the 1980 Census to two of the proposed data releases for the 2020 Census (using the data from the 2010 Census).

- Using only primary suppression, it found that 83.8% of the block-level cells in the P3 table (Race for the population 18 years and over), 95.7% of the block- group level cells, 84.3% of the tract-level cells, and 51.2% of the county-level cells would have needed to be suppressed.
- For the P4 table (Hispanic or Latino, and Not Hispanic or Latino by Race for the Population 18 Years and Over), the suppression numbers are 87.7%, 100.0%, 99.7%, and 84.2% (Hawes 2021a).

Swapping (the legacy mechanism) would not work.

“data swapping as originally proposed by Dalenius and Reiss does not generalize in ways that they thought.” (Fienberg and McIntyre (2005))

“Data swapping in its simplest form, wherein a fraction of households is swapped at random, will ‘normalize’ the strengths of the joint distributions of categorical variables, instead of lowering them.” (Kim 2015)

There were fundamental questions about the purpose of privacy and the availability of auxiliary information.

Many people arguing against DP were white men in positions of power.

- DP protects households that have same-sex parents and are mixed-race.
 - DP makes it harder for hoodlums with baseball bats out to harass mixed-race couples.
- DP protects households that have more than the legal number of residents.
 - “Section 8” (subsidized) housing in the US. (“Council housing” in the UK.)

Q: Should we protect (for example) data for 20 white males age 25 on a block?

- Critics said no. Census says yes.

Critics said the availability of commercial data made census data less important.

- But commercial data has significant gaps – children & race.

Other realizations

Simply making code and data available did not improve transparency.

Critics repeatedly argued that “reconstruction is not re-identification.”

- They neglected that reconstruction itself violated US Code Title 13.

Very few people understood differential privacy.

- “I think I can safely say that nobody really understands quantum mechanics,”

—Richard Feynman.

Critics mischaracterized differential privacy

All epsilons are not equal

- A randomized response epsilon of 1.0 for local model is different than an epsilon of 1.0 in a trusted curator model. There are different accuracy guarantees, and different privacy risks.

The actual privacy threat vs. the theoretical privacy threat is different depending on how epsilon is split up.

- An epsilon of 1.0 to a single question vs. an epsilon of 0.001 over a thousand questions that do not exhibit parallel composition.

Epsilon is the *maximum privacy loss*, but not necessarily the privacy loss.

- A mechanism with an epsilon of 1.0 can also be considered a mechanism with an epsilon of 2.0.
- With better privacy proofs, we can lower the epsilon of some mechanisms.

More misconceptions

Randomized response is a lousy way for thinking about DP.

Critics: “ $\epsilon = 19.61$ translates to binary RR with $p = 0.999999999696$ ”

- But there was no single question with a RR of $\epsilon = 19.61$

There was a reason, but it had nothing to do with DP

An article in The New York Times stated that DP was responsible for allocating 13 adults and one child to Census Block 1002 in downtown Chicago, a block that “consists entirely of a 700-foot bend in the Chicago River” (Wines 2022).

In fact, the TopDown algorithm implements a constraint such that “the number of householders (person one on the questionnaire) cannot be greater than the number of housing units” (J. Abowd et al. 2022).

- Possible reason #1: House boat
- Possible reason #2: Error in geo file

s.

The New York Times

GIVE THE

The 2020 Census Suggests That People Live Underwater. There's a Reason.

Technology advances forced the Census Bureau to use sweeping measures to ensure privacy for respondents. The ensuing debate goes to the heart of what a census is.

Share full article



An article in The New York Times stated that DP was responsible for allocating 13 adults and one child to Census Block 1002 in downtown Chicago, a block that “consists entirely of a 700-foot bend in the Chicago River”(Wines 2022).

In fact, the TopDown algorithm implements a constraint such that “the number of householders (person one on the questionnaire) cannot be greater than the number of housing units” (J. Abowd et al. 2022).

How the deployment of DP compares with public key cryptography

Public Key Cryptography and DP – Similarities

Both are mathematical approaches for protecting data:

- Well-defined protection goals.
- Indefinite time horizon

Implementation Concerns:

- Source of strong random numbers.
- Side channel leakage is a constant threat
- Failures are hidden – it's hard to distinguish working systems from compromised systems.

Security model assumes attacker has:

- Full access to source code
- Unlimited expertise

DP has a different threat model than cryptography.

Crypto threat model has 3 parties:

- The message sender (Alice)
- The message receiver (Bob)
- The eavesdropper (Eve)

DP threat model has 2 parties:

- The message sender
- The message receiver who is also the adversary

You can't even have the goal of being able to deny all data to the adversary!

- DP limits the *information gain* of the adversary to what the sender desires.

DP guarantees are different from crypto guarantees.

DP privacy guarantee is not all-or-nothing. (Similar to property-preserving crypto.)

DP uses a stronger threat model

- Information-theoretic: attackers are not computationally bounded.

Greater flexibility about what constitutes a privacy guarantee:

- That which can't be learned without the data subject's participation
 - the most common form of the guarantee.*
- A relative bound on how much more an attacker can learn about a set of intrinsically private secrets about the data subjects
 - A related form sometimes called 'inferential privacy'.*

Running DP systems inherently involves making and understanding social choices & economics.

Data Usefulness vs. privacy trade off

- What is the cost of the leakage?
- What is the benefit of the leakage?
- Can we find more efficient mechanisms – more benefit for the same cost.

The cost of cryptography disappeared in the 1990s.

- We used to argue about what needed to be encrypted and what didn't.
- Today we have “HTTPS Everywhere.”

Timeline: Public Key Cryptography vs. DP

Year	Public Key Cryptography	Differential Privacy
0	1976 DH / 1977 RSA / 1978 K (PKI)	2003 DN /2006 DMNS
3	1981 - RSA Patent US 4,405,829	2009 - OnTheMap (Census)
8	1986 - ElGamal	2014 - RAPPOR (Google)
13	1991 - PGP	2019 - End-to-End test
15		2021 - Census releases redistricting products
16	1994 - HTTPS	
17	1995 - SSH	2023 - Census releases first Demographic and Housing Products