Differential Privacy: Theory to Practice for the 2020 US Census

Simson L. Garfinkel
Chief Scientist, BasisTech, LLC
Lecturer, Harvard John A. Paulson School of Engineering and Applied Sciences

(Former Senior Computer Scientist for Confidentiality and Data Access, US Census Bureau)

AIT-Budapest
Aquincum Institute of Technology
November 14, 2024

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 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:

- What is differential privacy (DP), and why is it a scientific breakthrough?
- What is the US Census and why does it matter?
- How we brought DP to the 2020 Census
- Internal Challenges
- External Challenges
- Personal Reflectionds

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 confidential data.

Student Scores
(Hidden variables)

S1	S6
S2	S7
S3	S8
S4	S9
S5	S10

Your Score

Class Average = 98 (published)

Implies:

$$\frac{(S1 + S2 + S3 + S4 + S5 + S6 + S7 + S8 + S9 + S10)}{10} = 98$$

If

S10=80

and

0 ≤ Sn ≤ 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%

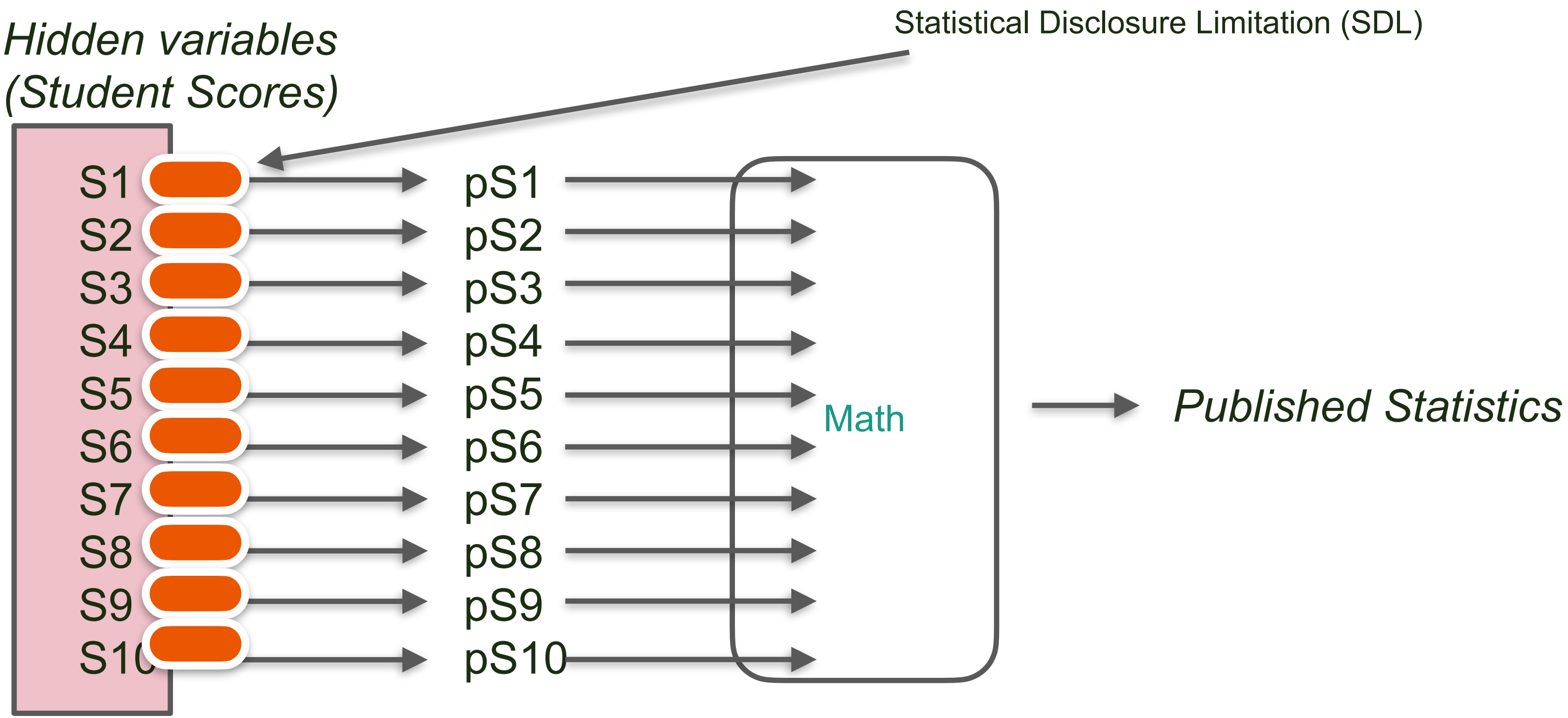
What about count? $n = 10$

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

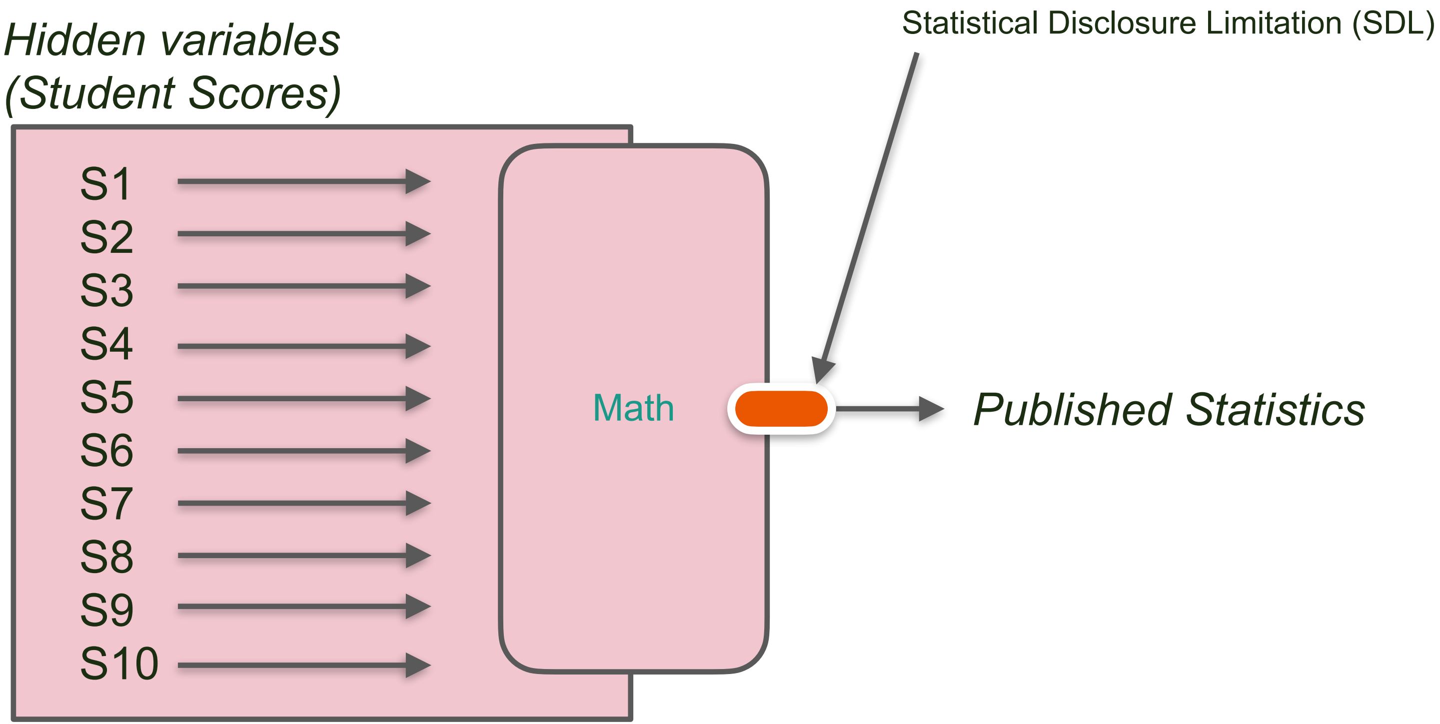
Could we publish that the class median is 100% ?

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

Statistical Disclosure Limitation (SDL) can be applied on inputs or outputs of a computation. Input protection applies to each variable *before* it is used in the computation.



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



There are many SDL approaches.

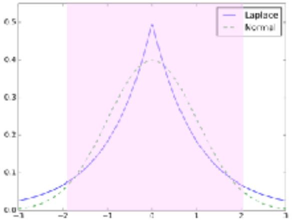
Protect 98%

The class average is…

98% → *Generalization* → *"is between 95% and 100%"*

98% → *Suppression* → *"not reportable due to the small class size"*

98% → *Noise infusion* → *"97%" (± 0..2 with 95% probability)*



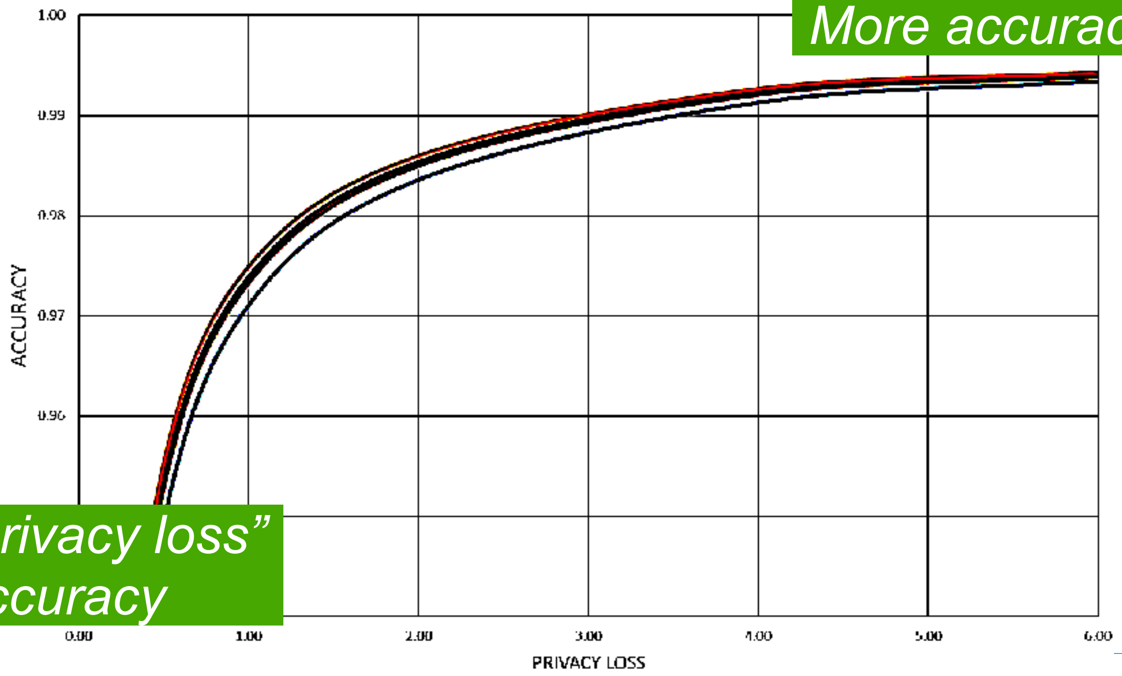
Differential privacy is a form of noise infusion

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

98%  *“97%” (± 0..2 with 95% probability)*

Noise Infusion

More “privacy loss”
More accuracy



Less “privacy loss”
Less accuracy

Differential privacy

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

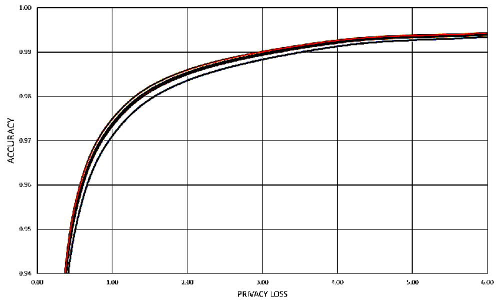
Privacy loss:

$0 \leq \varepsilon \leq \infty$

“Privacy bookkeeping” is the differential privacy breakthrough.

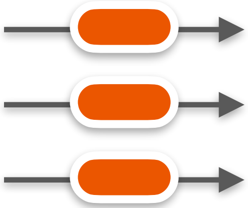
DP provides:

- The tradeoff between privacy loss and accuracy.



Composition rules:

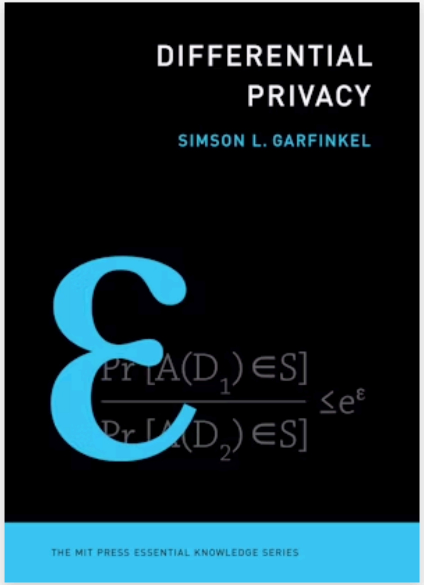
- Accounting for total privacy loss in complex statistical pipelines



Parallel Composition
(e.g. multiple blocks)

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

Differential Privacy, Garfinkel, MIT Press
March 25, 2025. (Open Access)



MIT Press Essential Knowledge series

Differential Privacy

By Simson L. Garfinkel

Paperback -

Paperback
ISBN: 9780262551656
Pub date: March 25, 2025
Publisher: The MIT Press
244 pp., 5 x 7 in, 22 b&w illus.

MIT Press Bookstore | **Penguin Random House** | **Amazon**

Barnes and Noble | **Bookshop.org** | **Indiebound** | **Indigo**

Books a Million

What is the US Census and why does it matter?

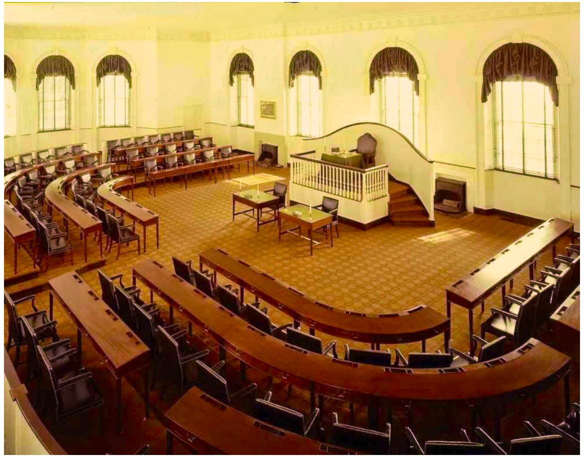
The US Census is the world’s longest running statistical program.

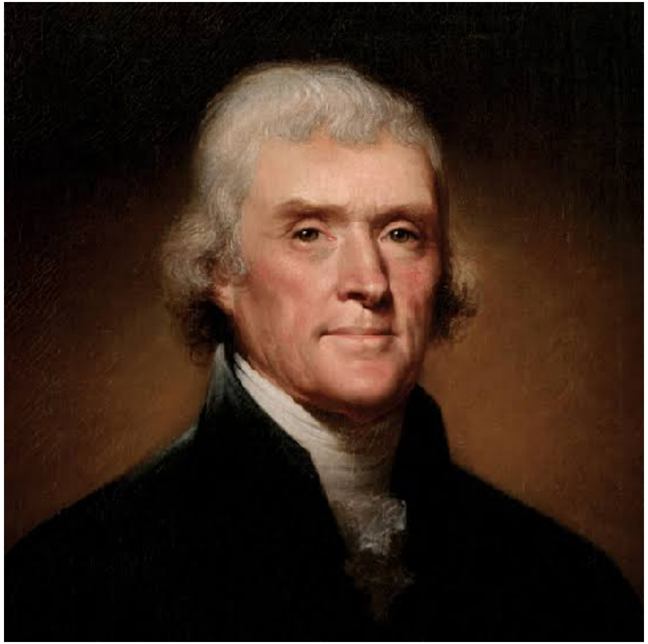
First US Census:

1790

Purpose:

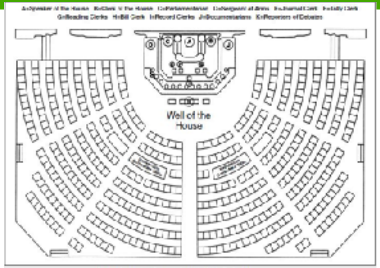
Apportion the US House of Representatives

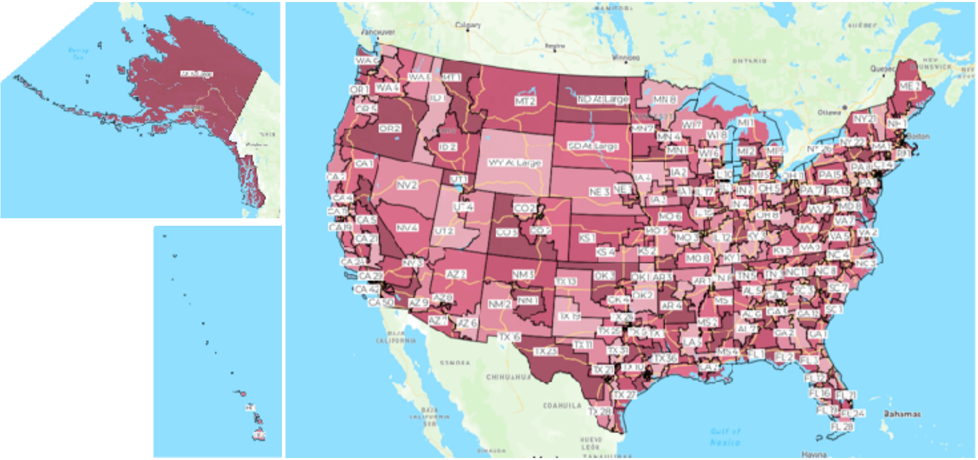




Thomas Jefferson
Primary author, US Declaration of Independence
First US Secretary of State
First US Patent Commissioner (reviewed every patent)
Oversaw first US Census

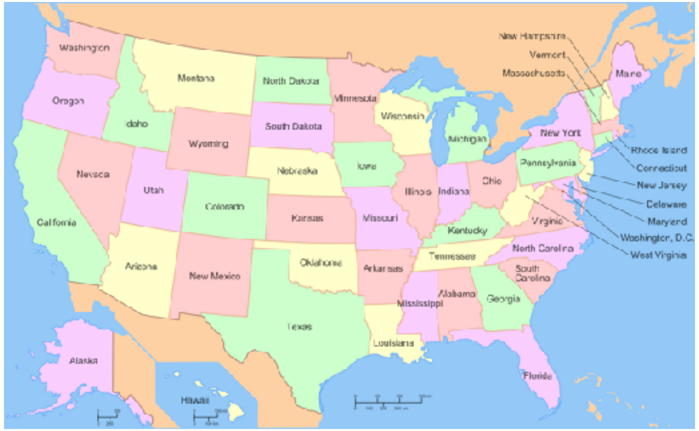
The US Constitution calls for a census every 10 years.
2020 was the 23rd US census.





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

There have been 435 seats since 1912



Each state elects 2 senators

The 2010 Census used three approaches to maintain statistical confidentiality.

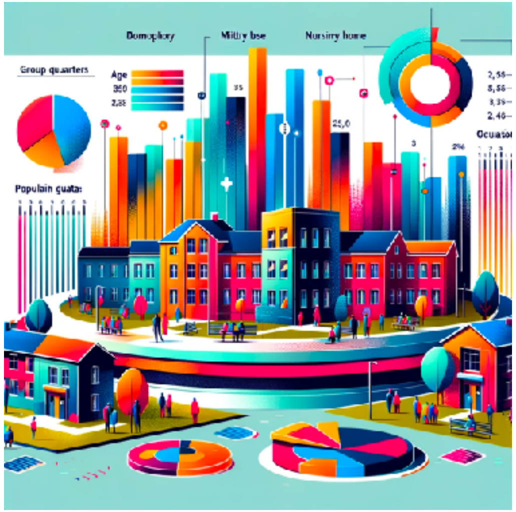
#1 — Record Swapping.

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

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

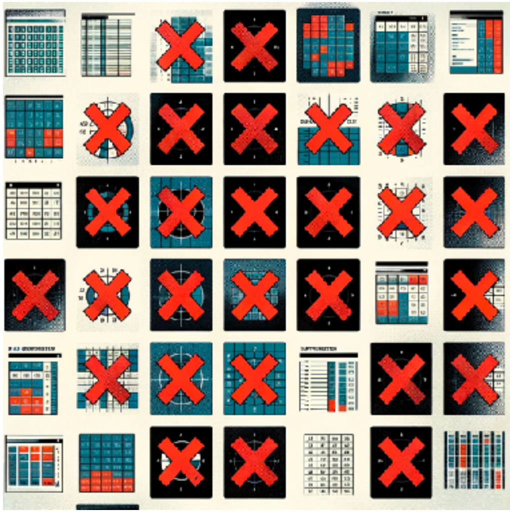
ChatGPT

Swapping



ChatGPT

Synthetic Data



ChatGPT

Suppression

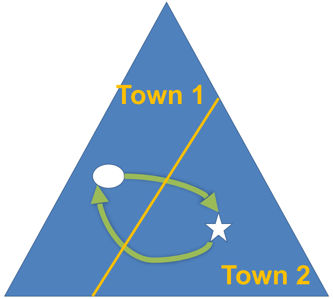
Data flow in the 2010 Census.

Raw data from respondents: Decennial Response File

→ Selection & unduplication: Census Unedited File

→ Edits, imputations: Census Edited File

→ Confidentiality edits (household swapping), tabulation recodes: Hundred-percent Detail File



State “X”

→ Pre-specified tabular summaries: PL94-171, SF1, SF2 (SF3, SF4, … in 2000)

→ **Special tabulations and post-census research**

19

How we brought DP to the 2020 Census

2016 — The Census Bureau moves to Differential Privacy.

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

2016 — Tammy Adams reconstructs micro data for Fairfax County

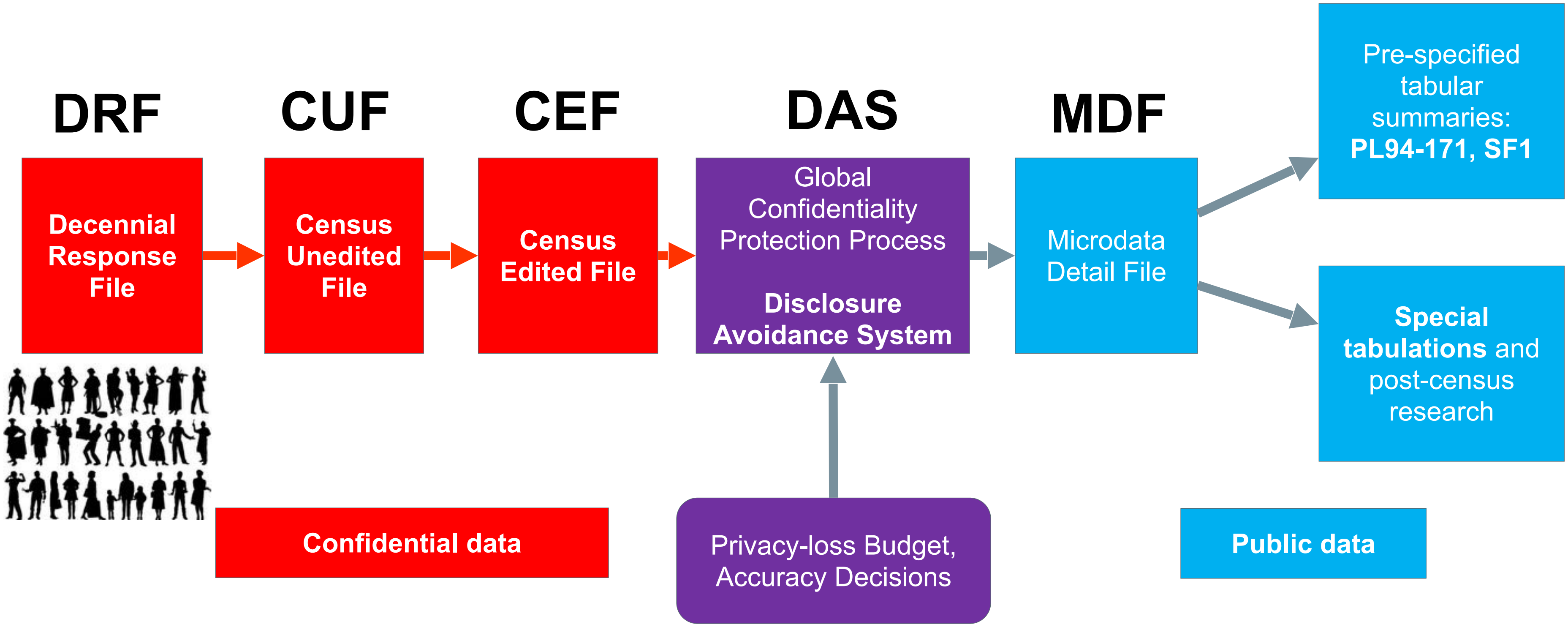
- Shows that the 2010 Census privacy protection mechanism was vulnerable by applying "database reconstruction" to the published tables.

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

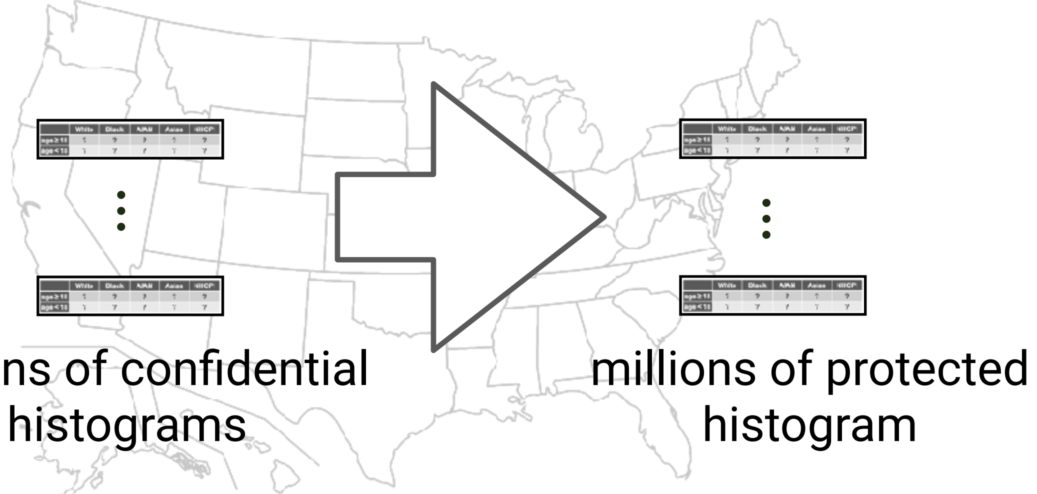
Data flow in the 2020 Census
(Original vision)



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

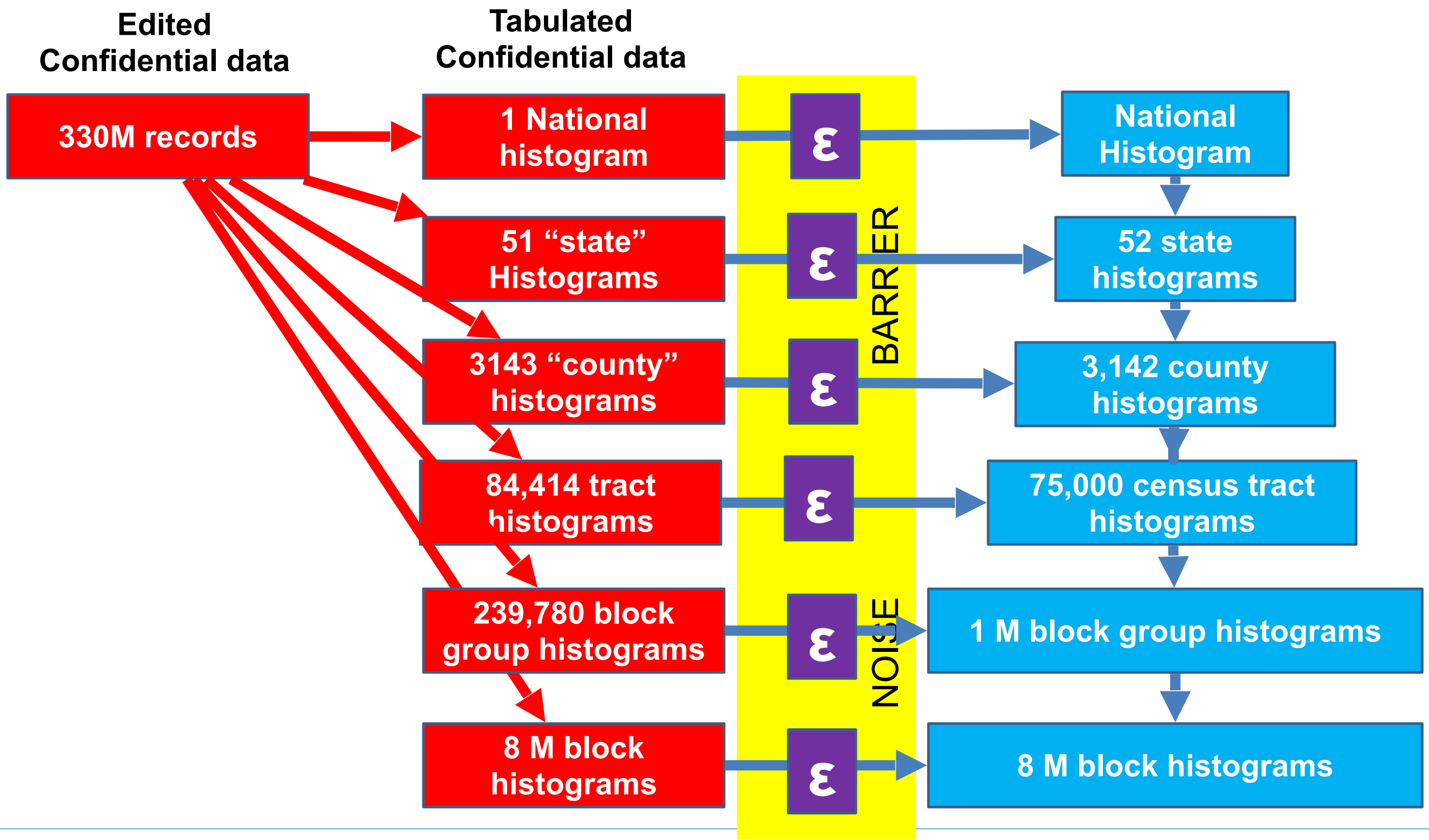
Naive approach: block-by-block

- 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.

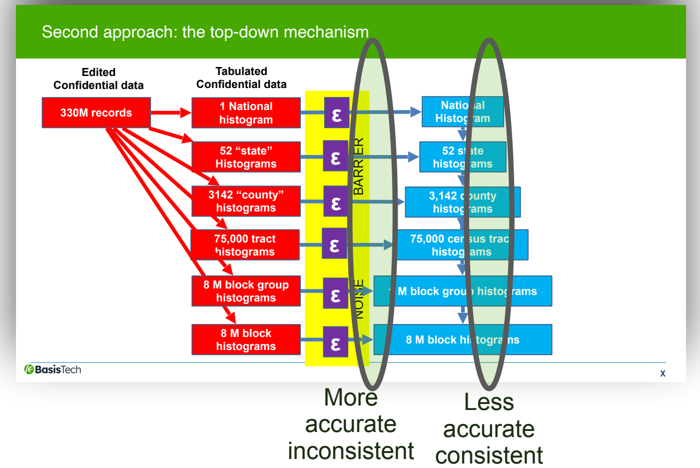


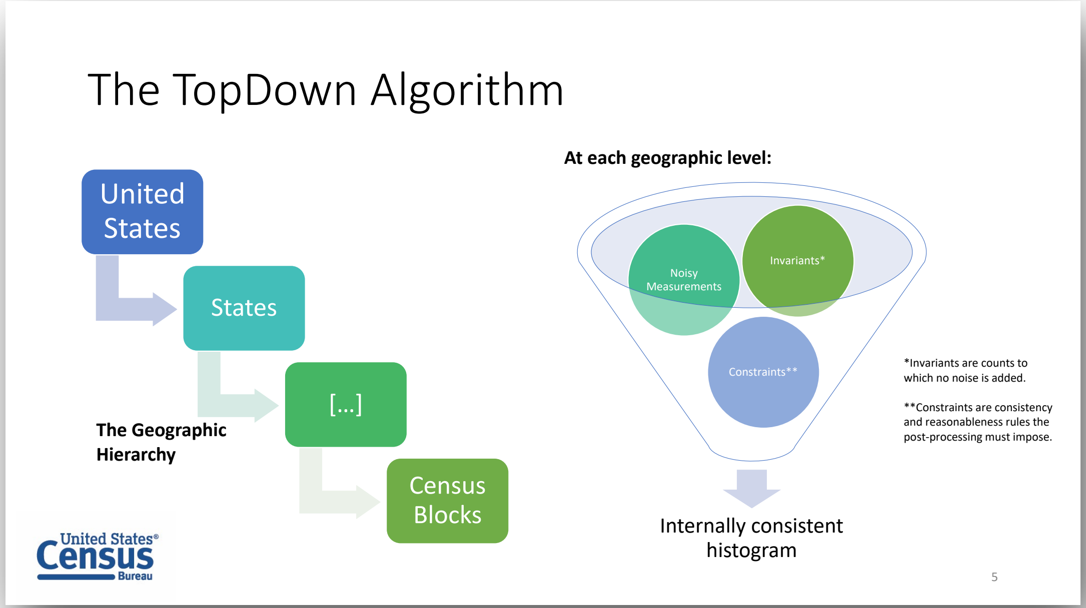
This is “local differential privacy” applied to blocks, rather than people.

Preferred approach: the top-down mechanism
Each histogram provides statistical accuracy to those underneath.



The final visual language.





https://www2.census.gov/about/training-workshops/2023/2023-06-15-nmf-presentation.pdf

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!*

We resolved the challenge of developing code with confidential data.

We transitioned 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

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 TopDownAlorithm.

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.

Technical Challenges

Project management challenges.

Challenges we expected:

- Obtaining qualified personnel and tools
- Obtaining a suitable computing environment
- We didn’t know what the right answer was

Challenges we didn’t expect:

- Desire for “repeatable random numbers”
  - —*For regression tests….*
- Policy that prohibited developing software with “Title 13” data
  - —*They wanted us to use synthetic data for software development*
  - —*If we had realistic synthetic data, we wouldn’t have needed to develop the DAS! (~4 months of arguments)*
- Large amount of system administration required
  - —*Maintaining the “bootstrap script” for the servers*
  - —*Maintaining our own Python distribution*
  - —*Building our own python module repository & managing dependencies over the course of the 5 year project*

The 2020 Census Disclosure Avoidance System: Technical Overview.

~100,000 line program written in Python 3.6

Batch processing with Apache Spark

Input file: 16GB files in Amazon S3

- Sparse data representing 1.3T integers
- Represented as ~ 8M scikit sparse histograms

Processing:

- Python creates ~ 16M mixed integer linear programs solved with Gurobi
- 20-50 AWS 96-core servers with 768GiB RAM

Output file: 1.7 GB sparse (microdata) saved to Amazon S3

Typical cost per run: $1000 - $10,000

Typical time per run: 8-36 hours

Writing the DAS required improving software development.

Initial TopDownAlgorithm was written on a single Linux server with Spark in “local mode.”

We needed to:

- Migrate to AWS and Amazon Elastic Map Reduce.
- Develop tools for managing Amazon S3 as if it were a file system.
- Migrate to “git” as our source-code control system.
- pylint and pytest as a pre-commit hooks to prevent pushes that were problematic.
- pytest for unit tests and pytest-cov for code coverage metrics. (Run by Jenkins)
- Monitoring of each run using a home-grown monitoring system
    - *We were denied access to AWS console due to “security” concerns.*

We built a system for monitoring each run of the TopDownAlgorithm.

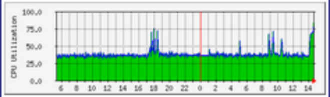
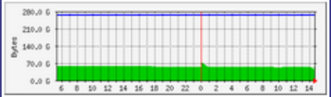
The algorithm computes and protects a histogram for various geographical units at various geographical levels

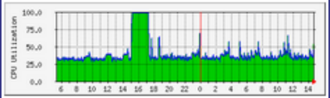
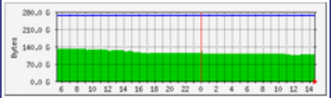
Level	Count	Integers per histogram	Total Histogram storage (bytes)
National	1	217,124	869 KB
State	51	217,124	44 MB
Counties & Equivalents	3143	217,124	2.7 GB
Census Tracts	73,057	217,124	63 GB
Inhabitable Blocks	6.2* M	217,124	5 TB

We actually need two histograms per node! (input & output)

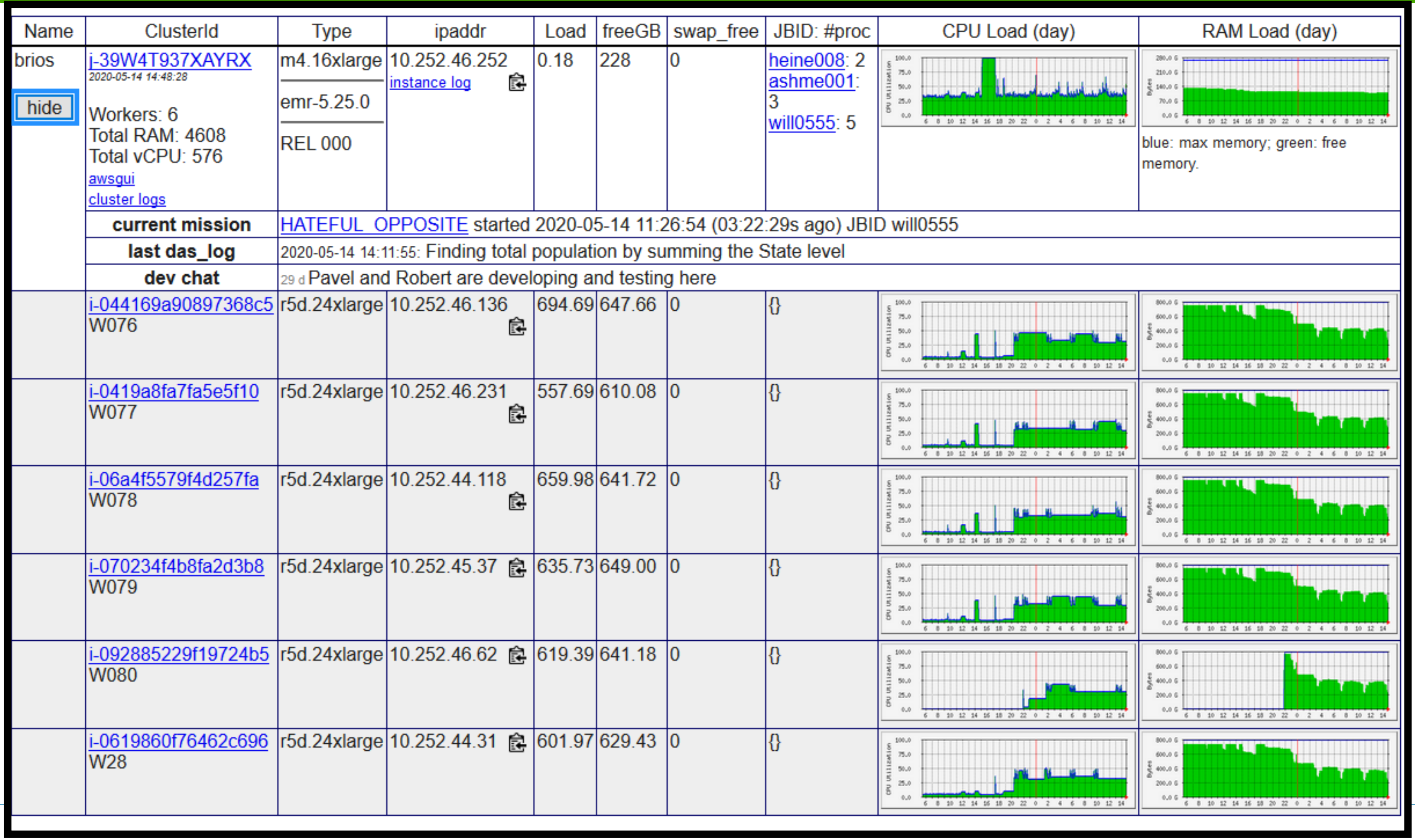
Source: https://www.census.gov/geographies/reference-files/time-series/geo/tallies.html
2010 inhabited block count: 6.2 M; 2020 block count: 8 M (estimated)

The system monitored multiple clusters we created in AWS GovCloud.

Name	ClusterId	Type	ipaddr	Load	freeGB	swap_free	JBID: #proc	CPU Load (day)	RAM Load (day)
abdat show	i-17IN8KRIVZ04I 2020-05-14 14:47:12 Workers: 8 Total RAM: 6144 Total vCPU: 768 awsgui cluster logs	m4.16xlarge emr-5.25.0 REL 003	10.252.44.211 instance log	0.38	232	0	lecle301: 2		 blue: max memory; green: free memory.
	current mission								
	last das_log	2020-05-14 14:35:14: Producing DAS output							
	dev chat	2 d Phil is using abdat for small-scale tests							

Name	ClusterId	Type	ipaddr	Load	freeGB	swap_free	JBID: #proc	CPU Load (day)	RAM Load (day)
brios show	i-39W4T937XAYRX 2020-05-14 14:47:56 Workers: 6 Total RAM: 4608 Total vCPU: 576 awsgui cluster logs	m4.16xlarge emr-5.25.0 REL 000	10.252.46.252 instance log	0.20	228	0	heine008: 2 ashme001: 3 will0555: 5		 blue: max memory; green: free memory.
	current mission	HATEFUL_OPPOSITE started 2020-05-14 11:26:54 (03:21:25s ago) JBID will0555							
	last das_log	2020-05-14 14:11:55: Finding total population by summing the State level							
	dev chat	29 d Pavel and Robert are developing and testing here							

Each cluster could be expanded to identify inefficiencies with the algorithm.



Name	ClusterId	Type	ipaddr	Load	freeGB	swap_free	JBID: #proc	CPU Load (day)	RAM Load (day)
brios hide	j-39W4T937XAYRX 2020-05-14 14:48:28 Workers: 6 Total RAM: 4608 Total vCPU: 576 awsgui cluster logs	m4.16xlarge emr-5.25.0 REL 000	10.252.46.252 instance log	0.18	228	0	heine008: 2 ashme001: 3 will0555: 5		blue: max memory; green: free memory.
	current mission	HATEFUL_OPPOSITE started 2020-05-14 11:26:54 (03:22:29s ago) JBID will0555							
	last das_log	2020-05-14 14:11:55: Finding total population by summing the State level							
	dev chat	29 d Pavel and Robert are developing and testing here							
	i-044169a90897368c5 W076	r5d.24xlarge	10.252.46.136	694.69	647.66	0	{}		
	i-0419a8fa7fa5e5f10 W077	r5d.24xlarge	10.252.46.231	557.69	610.08	0	{}		
	i-06a4f5579f4d257fa W078	r5d.24xlarge	10.252.44.118	659.98	641.72	0	{}		
	i-070234f4b8fa2d3b8 W079	r5d.24xlarge	10.252.45.37	635.73	649.00	0	{}		
	i-092885229f19724b5 W080	r5d.24xlarge	10.252.46.62	619.39	641.18	0	{}		
	i-0619860f76462c696 W28	r5d.24xlarge	10.252.44.31	601.97	629.43	0	{}		

Each DAS run was a “mission.”

Currently Executing / Recently Crashed DAS Runs

Export as CSV Show 25 entries

Search:

Id	Mission Name	Start	Exit Code	Num Geolevels	Num Geounits	ApplicationId	Cost	Seconds	JBID	Gurobi V
14334	LITIGIOUS_WELCOME	2020-05-13, 06:05:16 PM		6		application_1580837820168_1257		72496	ashme001	9.0.0
14348	WEIGHTLESS_UNIQUE	2020-05-14, 09:05:14 AM		7		application_1589461714248_0001		19698	lecle301	9.0.0
14357	PUNCTUAL_SHARE	2020-05-14, 10:05:43 AM		7		application_1589461881774_0001		16729	lecle301	9.0.0
14397	INTRAVENOUS_BREAD	2020-05-14, 02:05:37 PM	1	6		application_1586953709588_0158		115	will0555	9.0.0

Showing 1 to 4 of 4 entries

Previous 1 Next

The Mission Report showed details of each mission.
You never know what might be important when debugging a huge program.

SHIMMERING_SPITE mission report 2
10.252.45.99
(Completed, exit_code = 0)

Total cost: $7,489.01

Total nodes: 59

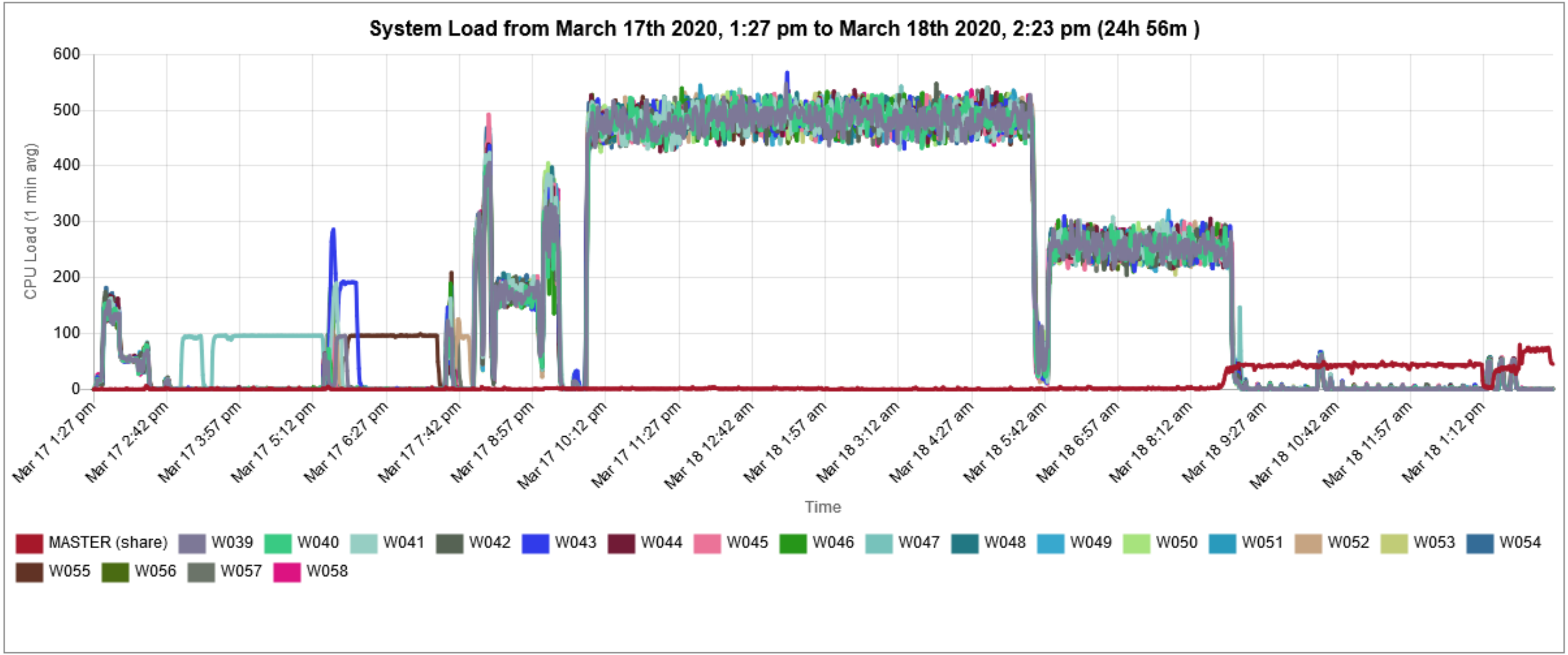
view config file in another tab

Search Results

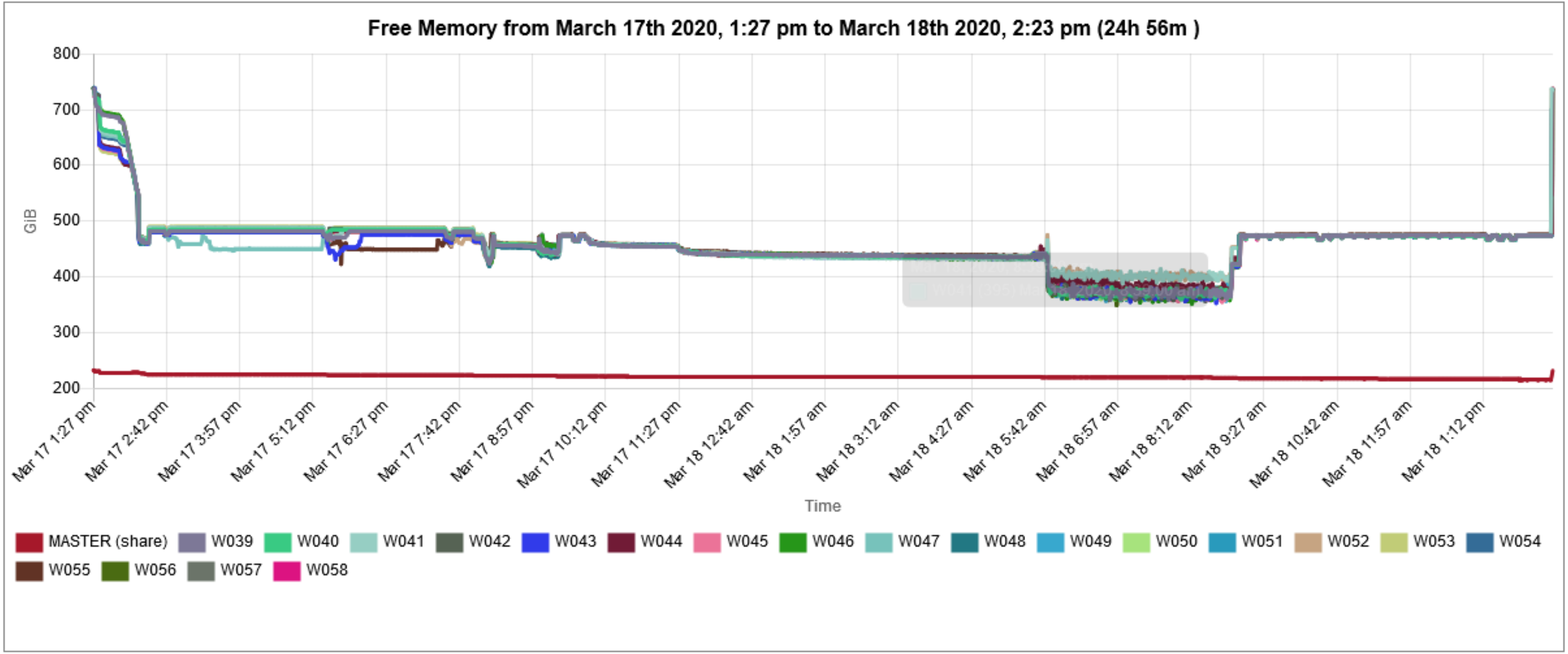
☑ Show Search Results

Key	Value
id	9383822
t	2020-05-14 14:53:04
modified_at	2020-03-18 14:27:55
campaign_name	(None)
mission_name	SHIMMERING_SPITE
mission_url	(None)
das_instance_id	9383822
das2020_version	(None)
git_commit	(None)
das_release	000
jbid	heine008
start	2020-03-17 13:28:27
stop	2020-03-18 14:22:40
exit_code	0
notes	(None)
spark_submit	spark-submit --driver-memory 10g --num-executors 20 --executor-memory 250g --executor-cores 48 --driver-cores 10 --conf spark.driver.maxResultSize=0g --conf s
fname_logfile	(None)
fname_stdout	(None)

System load during a 24 hour run.



Memory usage during 24-hour run.



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

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

I realized that we could demonstrate the algorithm with data from the 1940 Census!

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

Advantages:

- Micro data 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

July 2019

Tested with data from 1940

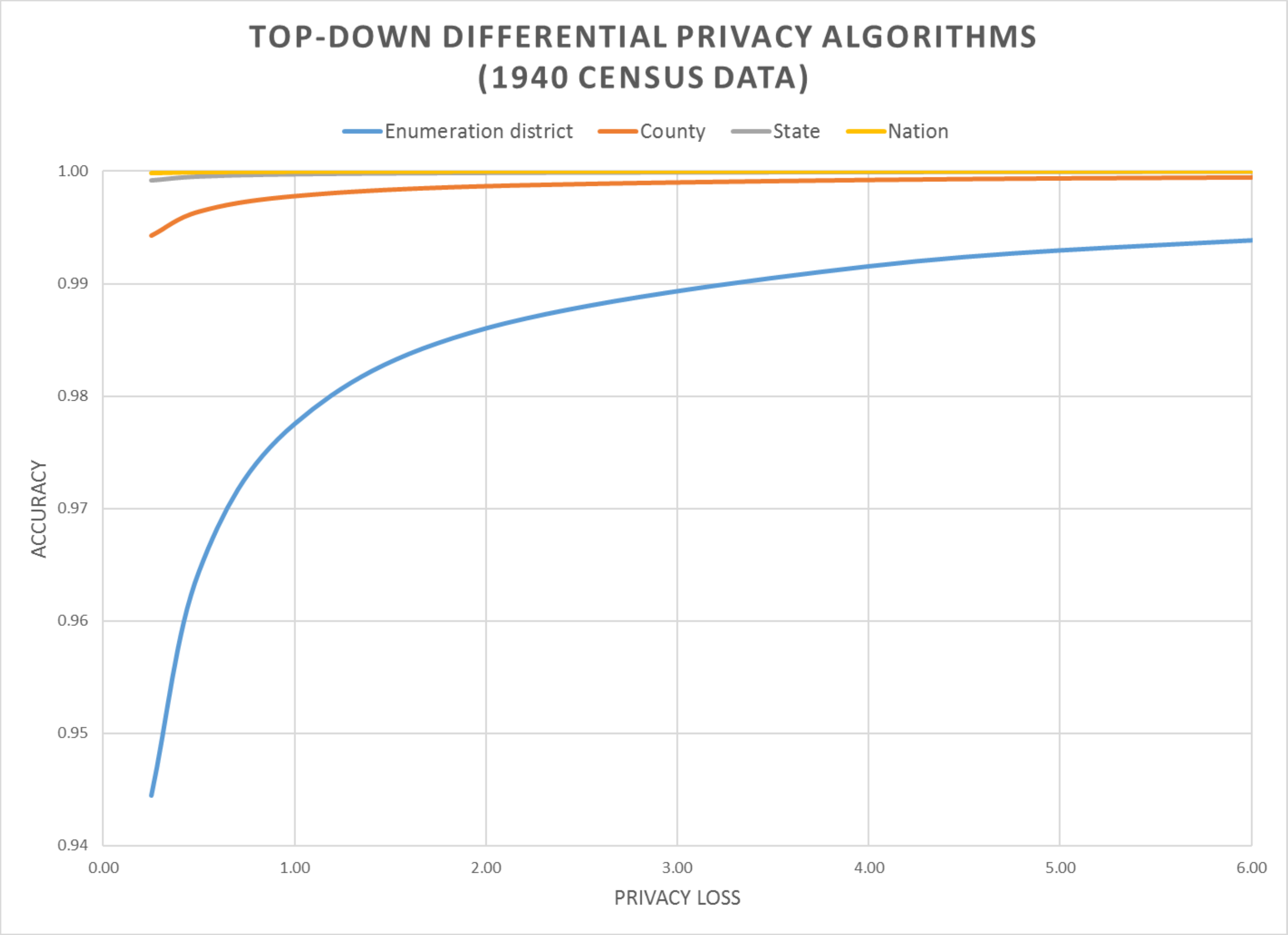
1940 hierarchy:

- Nation
- State
- County
- Enumeration District

Download from usa.ipums.org

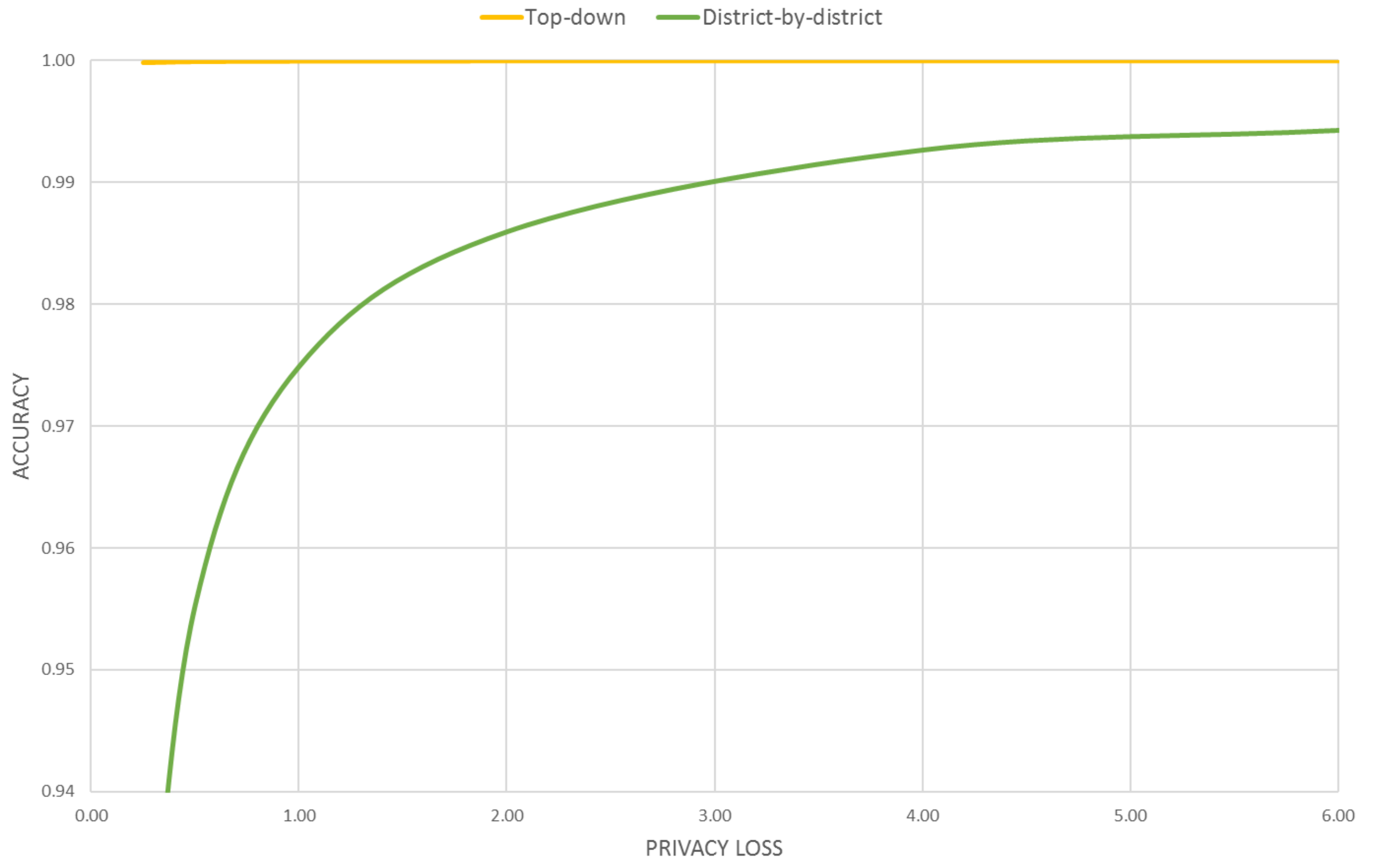
July 2019

Top-Down: much more accurate!

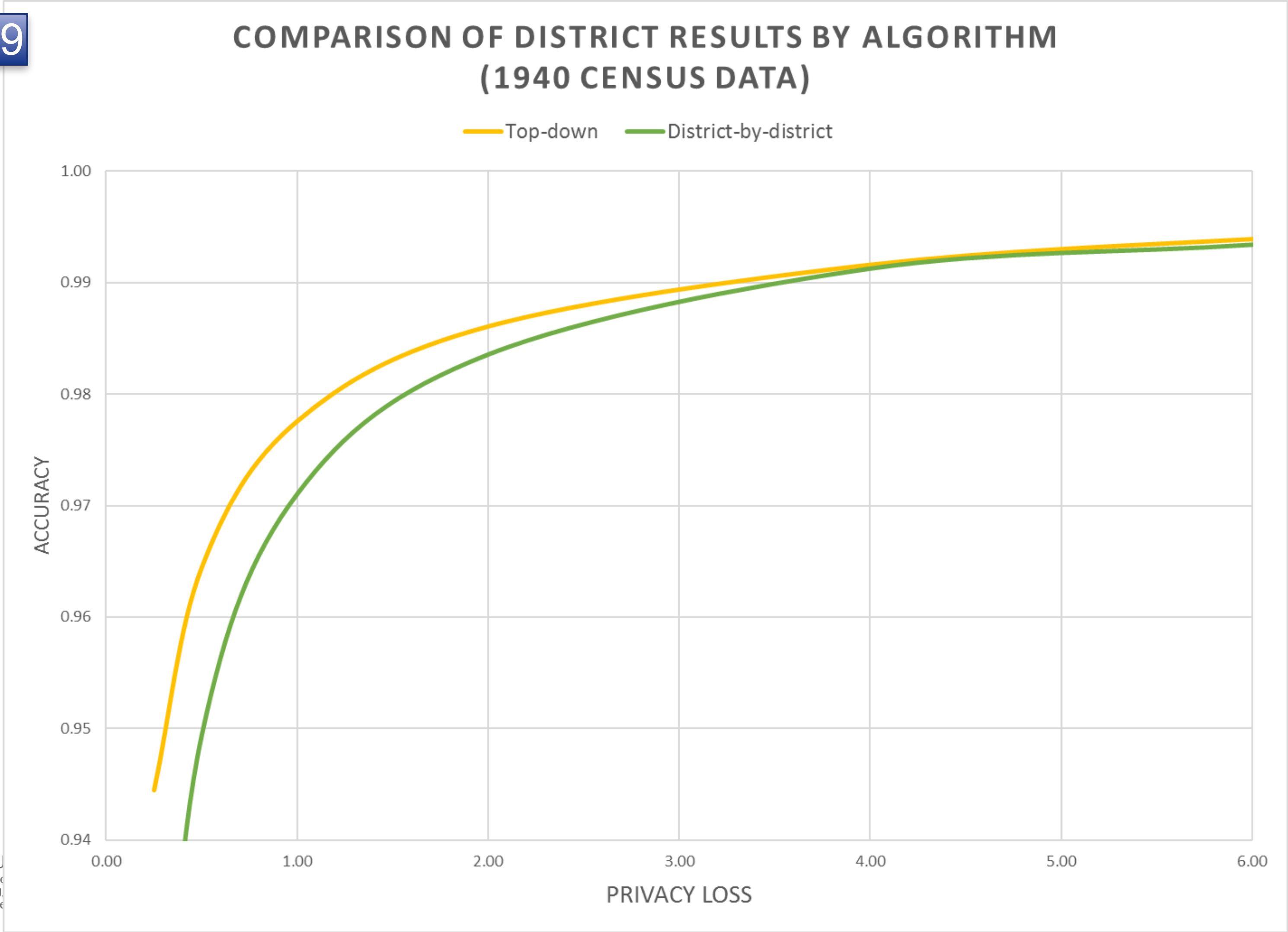


July 2019

COMPARISON OF NATIONAL RESULTS BY ALGORITHM (1940 CENSUS DATA)

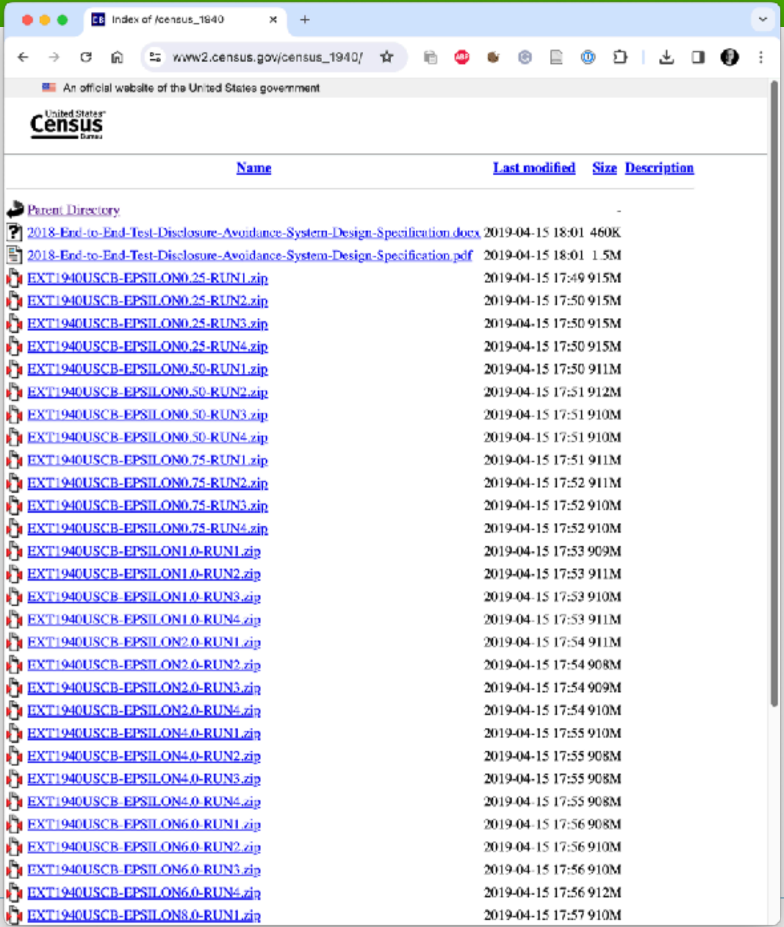


July 2019



United States Census Bureau

Multiple releases of 1940 data run through the DAS.



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	

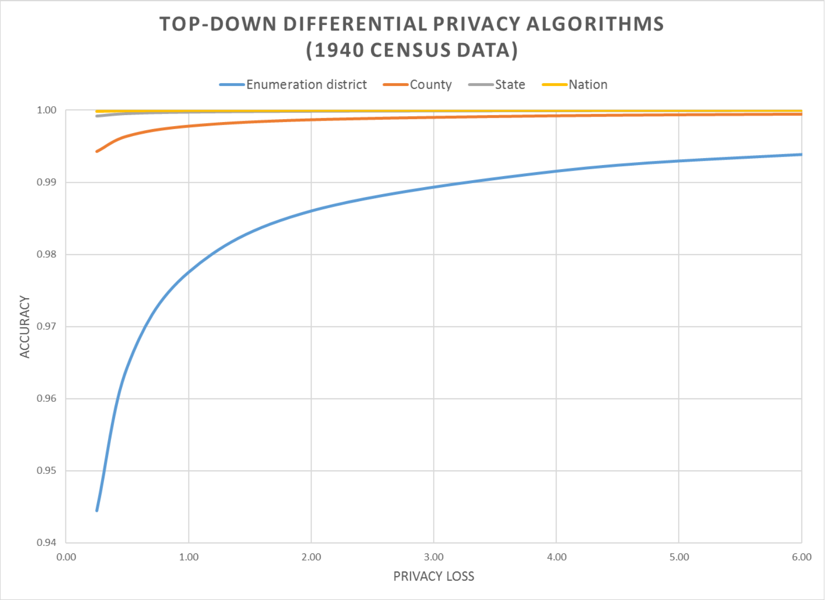
July 2019

Scientific Issue for any use of DP: Quality Metrics

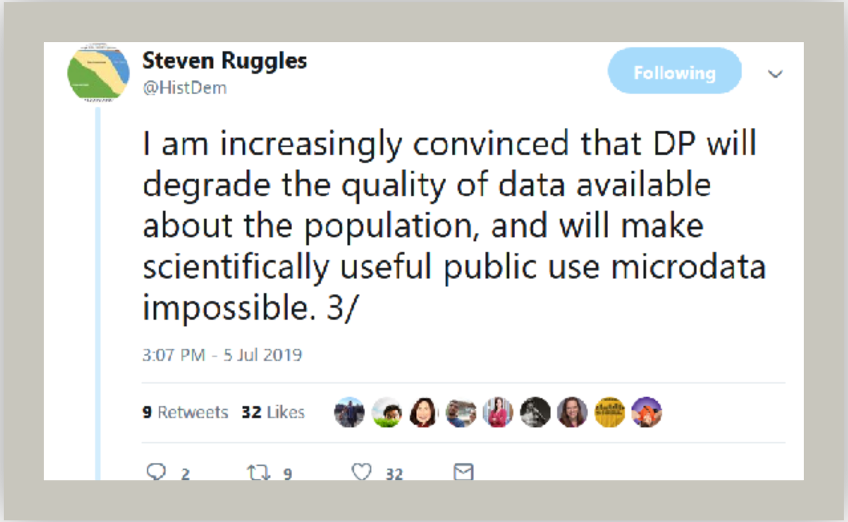
What is the measure of "quality" or "utility" in a complex data product?

Options:

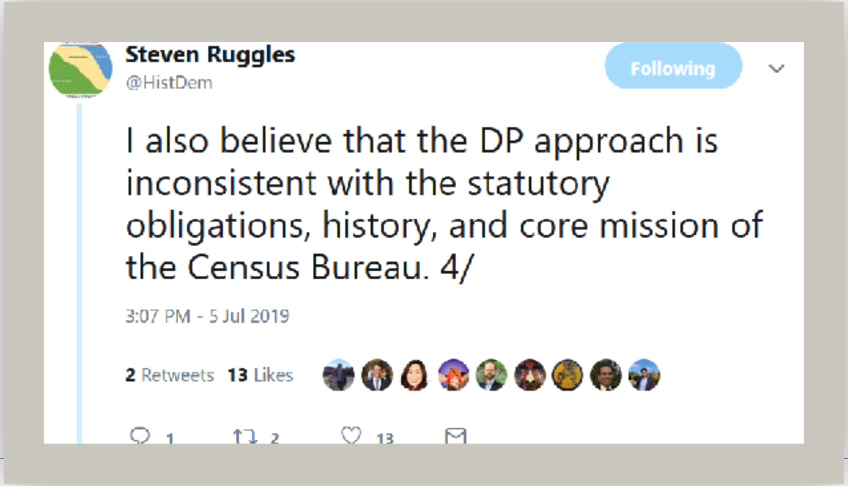
- L1 error between "true" data set and "protected" data set
- Impact on an algorithm that uses the data (e.g., redistricting and Voting Rights Act enforcement)



Early attacks against differential privacy in the 2020 Census.



Steven Ruggles 5 Jul 2019



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

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.

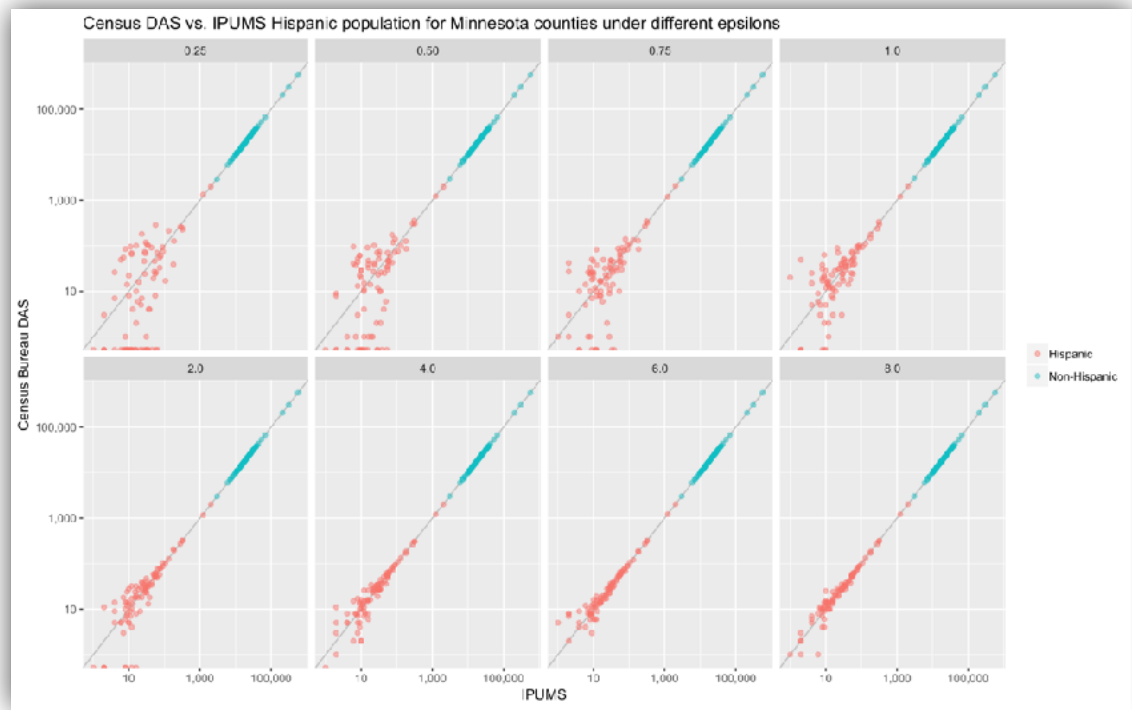
Results:

- The US Census Bureau seriously considered the concerns of the statistician
- (Later, plans were shelved to rapidly deploy DP for the American Community Survey.)

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

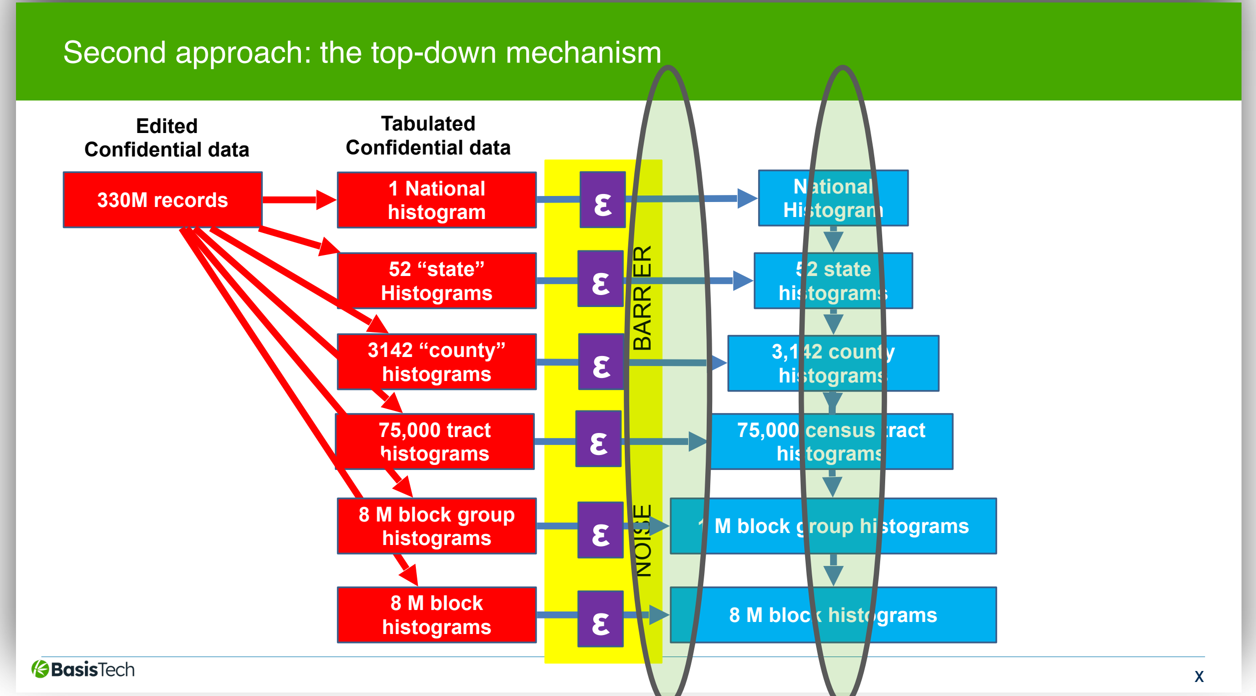
Note:

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



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

The source of the inaccuracy: integer non-negative constraints:



More accurate inconsistent

Less accurate consistent

The error comes from enforced consistency:

	White	Black	AIAN	Asian	NHOPI
age ≥ 18	2	0	0	0	0
age < 18	1	0	0	0	0

NOISE BARRIER

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

NOISE BARRIER

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

NOISE BARRIER

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

We re-released the 2010 data through the DAS for a 2019 special 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.

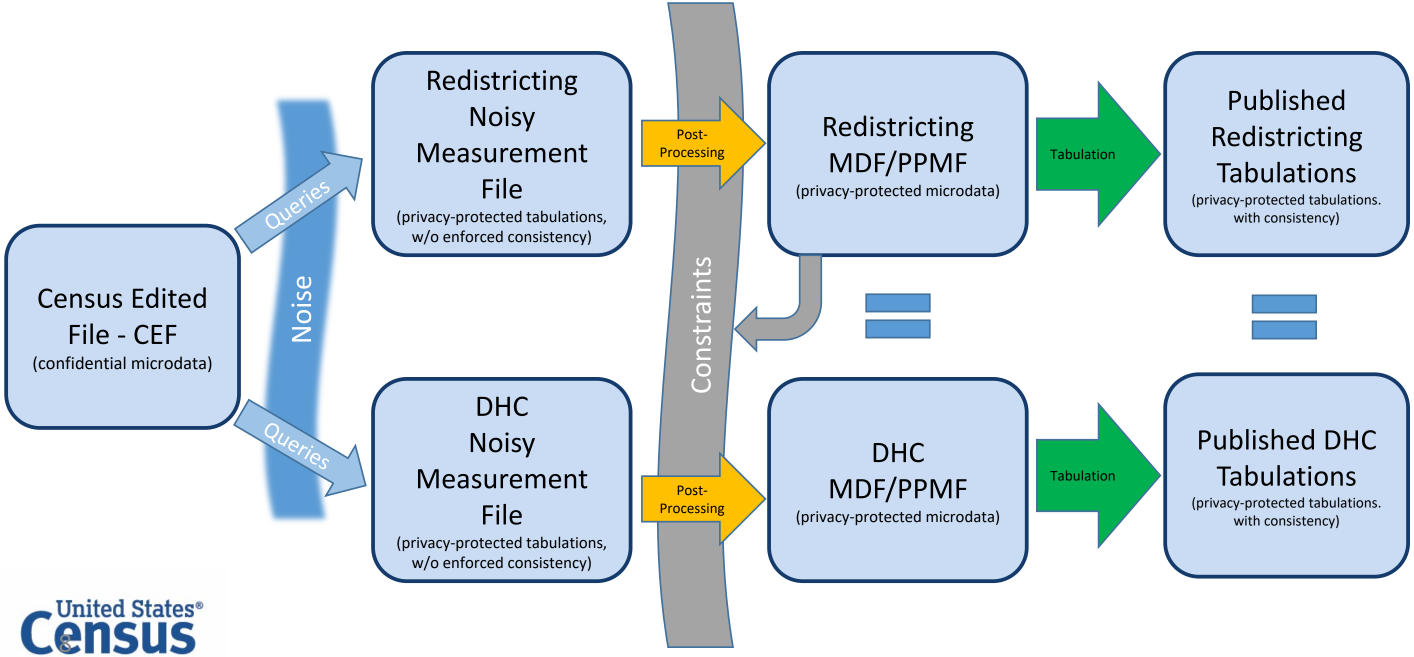
“(f.) Difficulties estimating error.

“(g.) The importance of protecting privacy.”

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

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

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



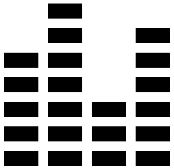
United States® Census Bureau

https://www.census.gov/data/academy/webinars/2023/noisy-measurement-files.html

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

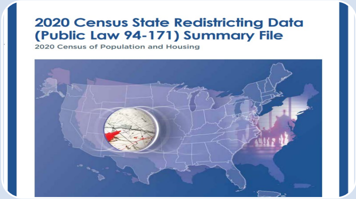


Post-processing
- Data dependent
  - While the nonnegativity requirement decreases error in the detailed cell counts, it also introduces a positive bias in small counts and an offsetting negative bias in large counts.
  - TDA also reduces the amount of error for many statistics relative to their corresponding 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.

9

https://www.census.gov/data/academy/webinars/2023/noisy-measurement-files.html

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



2020 Census Redistricting and DHC Tabulations

- Official 2020 Census Statistics
- Higher Accuracy (feature of TDA)
- Does include bias due to post-processing

2020 Census PPMF

- 100% microdata file
- Consistent with published tabulations
- Useful for special tabulations and microdata analysis

2020 Census NMF

- Can be used to produce unbiased estimates and confidence intervals
- Can be used to evaluate alternate post-processing mechanisms
- Research product

United States® Census Bureau

https://www.census.gov/data/academy/webinars/2023/noisy-measurement-files.html

Errors in the 2020 Census were blamed on 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 implemented 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).

- Likely answer: Error in geography file
- Unlikely answer: house boat

Researchers criticized the 20020 DP data products — and the 2010 products too!

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

Swapping unique rows in the 2010 census 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)

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

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).

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."
- We believed that US law 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.
- Most of the critics were arguing from a position of personal privilege.

Very few people understood differential privacy.

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

  —*Richard Feynman.*

Personal reflections

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: “$\varepsilon = 19.61$ translates to binary RR with $p = 0.99999999696$”

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

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."

Deployment of public key cryptography and DP are similar.

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

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

There’s a lot more to say…

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