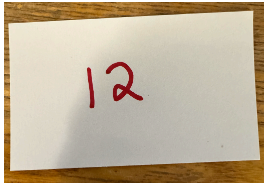
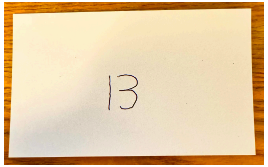
Please take 2 index cards.

Write a number on each such that *together* the two integers between -99 and 99 that add to your age.

Example — If you are 25, you could write:



10 and 15
-5 and 30
12 and 13



Please write each card with a different pen.

How Privacy Enhancing Technologies (PETs) let governments and business share sensitive data while protecting privacy.

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)

MIT 2-105 • 11am • January 14, 2025

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

Tax returns and financial filings, health records, education records, and crime data are just some of detailed and highly sensitive data that governments have about people.

Businesses also have huge archives of sensitive data, including consumer purchases, cellphone mobility traces, and video surveillance.

Today a tiny fraction of these data are released as “open data” or sold as “de-identified data.” The rest are locked up, unable to benefit society or promote new economic activity. Worse, much of that allegedly de-identified data can actually be re-identified, as happened when journalists at The Pillar used de-identified data to identify Catholic priests who were going to gay bars and using hookup apps.

Privacy Enhancing Technologies (PETs) use advanced mathematics and computational techniques to let organizations analyze and publish sensitive data while protecting the privacy or individuals and sensitive data from organizations. Although these techniques have existed for decades, they are increasingly being deployed by governments and businesses.

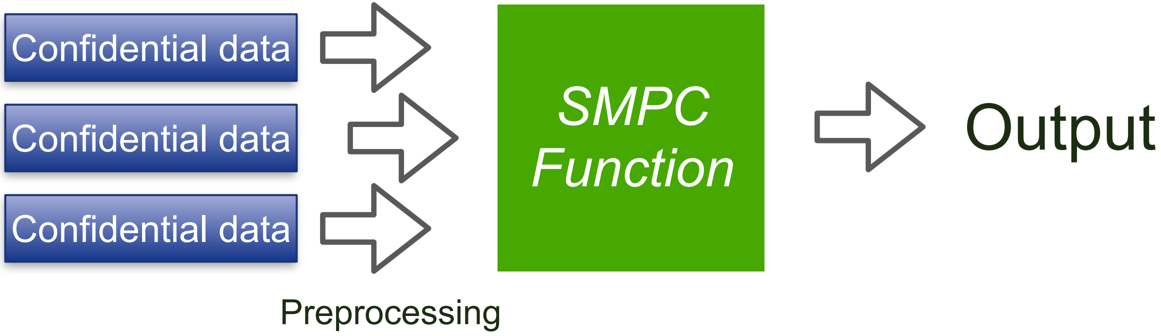
PETs are not without controversy: when the US Census Bureau adopted a PET called “differential privacy” for the 2020 Census, more than 4000 academics signed an open letter voicing their opposition: they were concerned that differential privacy would do such a good job protecting privacy that the resulting data would be useless for academic research.

This talk presents the case for PETs, explains popular PETs for a non-technical audience, and discusses the specific controversy of deploying differential privacy for the 2020 US Census.

Let’s compute the average age of people here with secure multiparty computation.

Secure multiparty computation (SMPC) is form of *secure computation*.

SMPC can compute the output of a function without having access to its inputs!



Confidential data

Confidential data

Confidential data

Preprocessing

SMPC Function

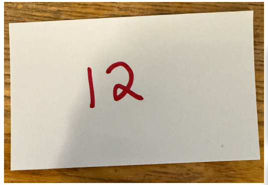
Output

Please pass one card to the left and one to the right!

Please pass all cards forward!

We can now compute the average age… without learning anyone’s age.

In a hypothetical room that turned in these eight cards:



10

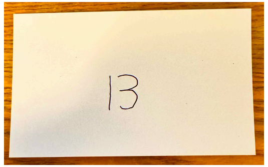
15

-7

8

4

43



30 + 68 = 98 ÷ 4 = 24.5

There are 4 people in the room, and their average age is 24.5

SMPC is a (PET) for Privacy Preserving Data Analytics

PETS — Technologies that by their design inherently protect privacy

- Examples: encryption, secret ballot voting, numbered PO Boxes, anonymous email

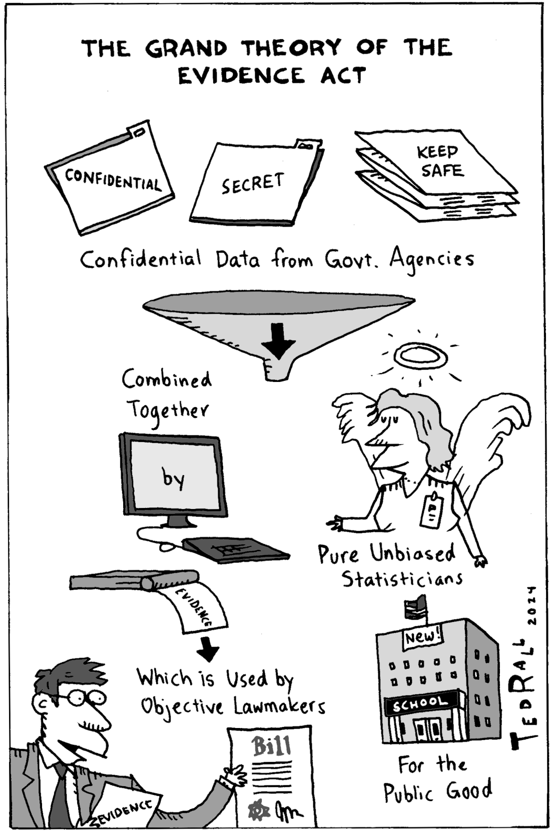
Privacy Preserving Data Analytics (PPDA)

- Also called Privacy Preserving Data Mining (PPDM)
- Produce a data product while protecting the input data
- Example: The average age of people in the room is 24.5

PPDA can separate the *confidential data* from the data that are *collected* or *stored*

- Better protection for the *confidential data*
- Reduced security requirements for *collected* and *stored data*
- Improved acceptance for data providers

On January 15, 2018 President Trump signed the Foundations for Evidence-Based Policymaking Act.


THE GRAND THEORY OF THE EVIDENCE ACT
CONFIDENTIAL
SECRET
KEEP SAFE
Confidential Data from Govt. Agencies
Combined Together
by
Pure Unbiased Statisticians
EVIDENCE
Which is Used by Objective Lawmakers
Bill
EVIDENCE
New!
SCHOOL
For the Public Good
TEDRALL 2024

The problem: the input “confidential data” are tremendously sensitive!

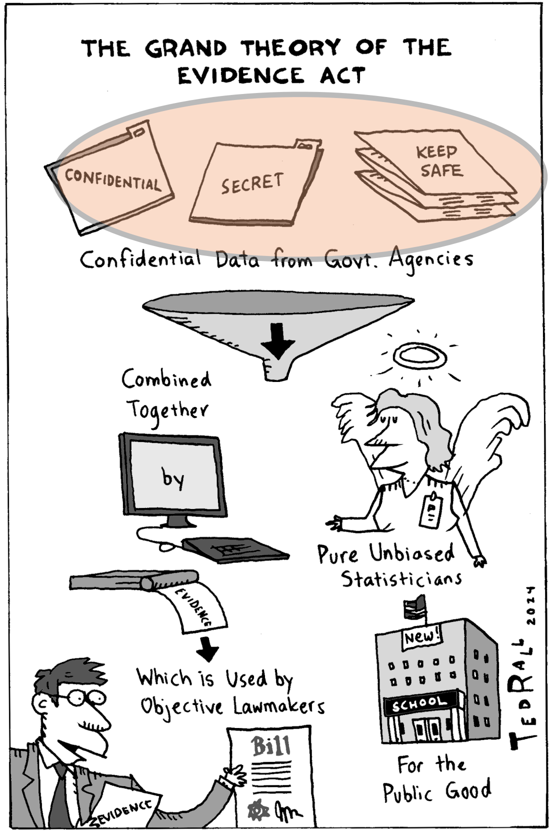
Many efforts to use confidential data are stalled:

- Negotiating data agreements
- Ensuring security for the confidential data
- Background investigations for the statisticians to work with the sensitive data
- Ensuring that the publications won’t compromise privacy

Today many respondents *do not provide data* due to confidentiality concerns

Many businesses cannot make full use of their data

PETs can make new data products possible

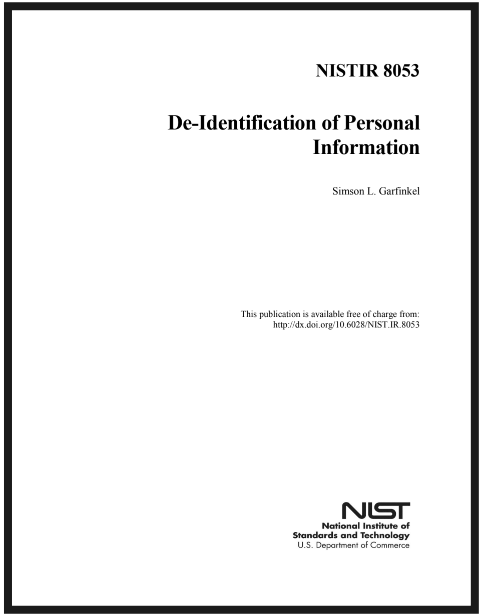

THE GRAND THEORY OF THE EVIDENCE ACT
CONFIDENTIAL
SECRET
KEEP SAFE
Confidential Data from Govt. Agencies
Combined Together
by
Pure Unbiased Statisticians
EVIDENCE
Which is Used by Objective Lawmakers
Bill
EVIDENCE
New!
SCHOOL
For the Public Good
TEDRALL 2024

This talk introduces PETs and shows how they can benefit governments and business

Outline for this talk:

- A secure multiparty computation (SMPC) example ✓
- Output controls — why secure computation is not enough
- Introduction to the PETs Zoo
- How we brought DP to the 2020 Census
- The Census 2020 DP backlash!
- Find out more!

This talk is based in part on:



NISTIR 8053

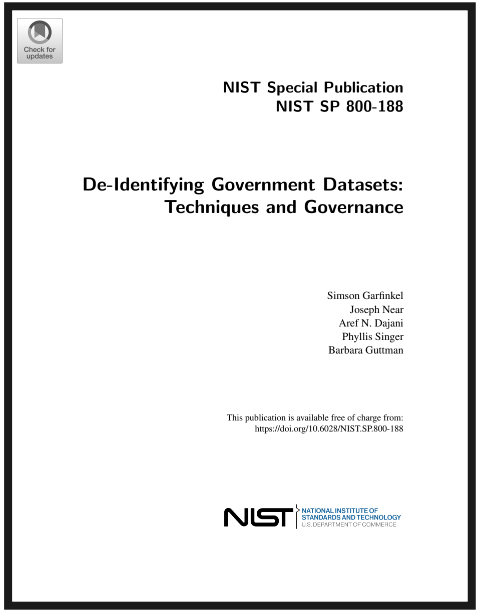
De-Identification of Personal Information

Simson L. Garfinkel

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8053

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 8053 (2015)



Check for updates

NIST Special Publication
NIST SP 800-188

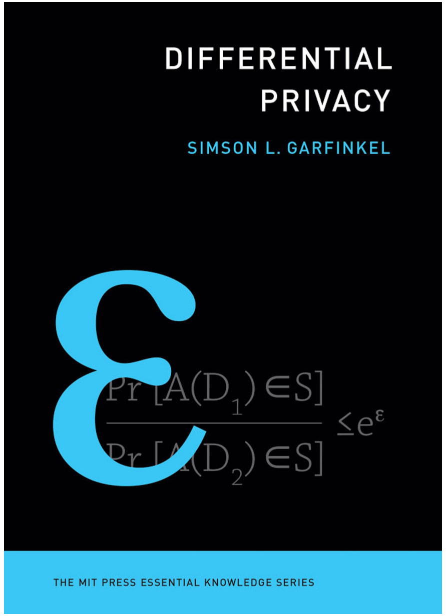
De-Identifying Government Datasets: Techniques and Governance

Simson Garfinkel
Joseph Near
Aref N. Dajani
Phyllis Singer
Barbara Guttman

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.800-188

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST SP 800-188 (2015)



DIFFERENTIAL PRIVACY

SIMSON L. GARFINKEL

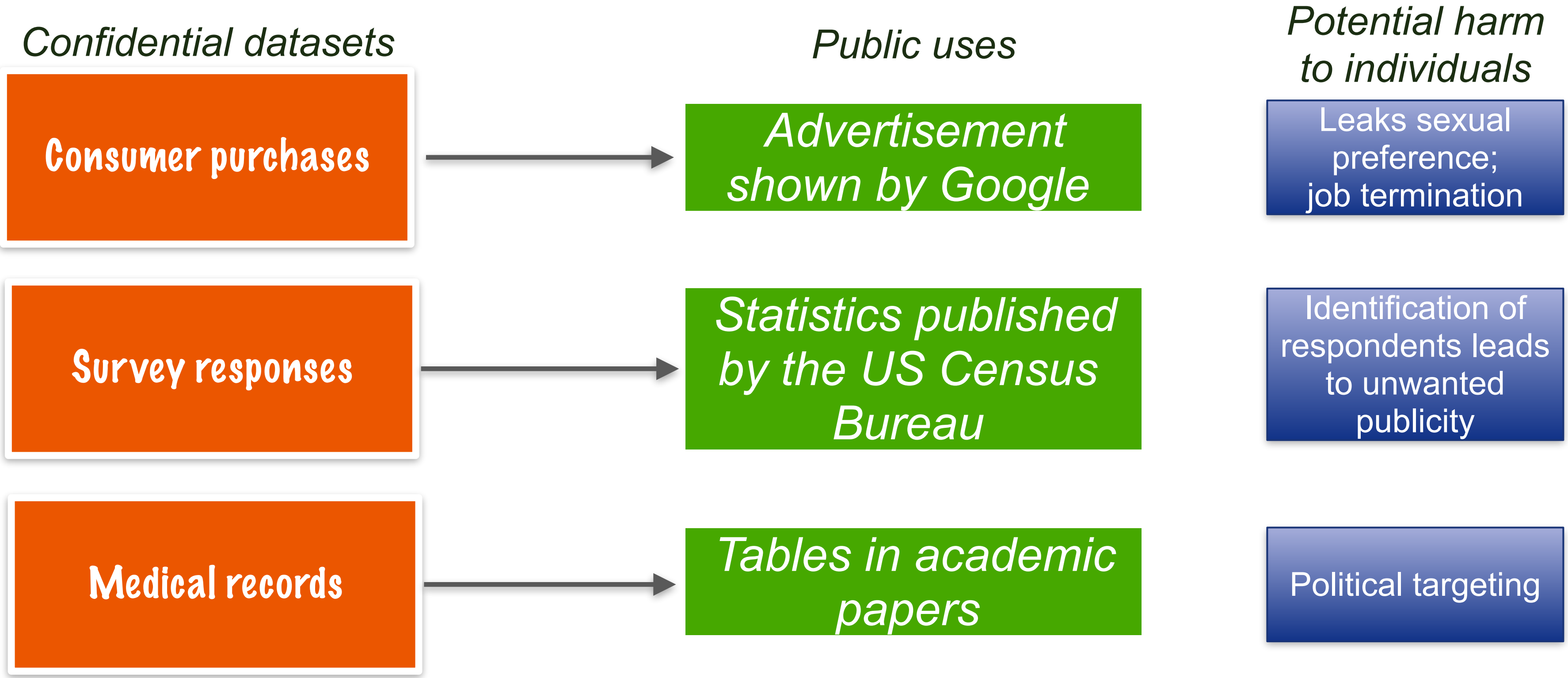
$$\varepsilon \quad \frac{\Pr[A(D_1) \in S]}{\Pr[A(D_2) \in S]} \leq e^{\varepsilon}$$

THE MIT PRESS ESSENTIAL KNOWLEDGE SERIES

MIT Press 2025!

Output Controls:
Why Secure Computation Isn't Enough

Many governments and businesses make *public use* of *confidential data*.



Confidential datasets

Public uses

Potential harm to individuals

Consumer purchases → *Advertisement shown by Google* — Leaks sexual preference; job termination

Survey responses → *Statistics published by the US Census Bureau* — Identification of respondents leads to unwanted publicity

Medical records → *Tables in academic papers* — Political targeting

Output controls can limit data privacy damage caused by data

Privacy Preserving Data Analytics *require* output controls.

You are in a class with 9 other students.

You look at your test and you got an 80%

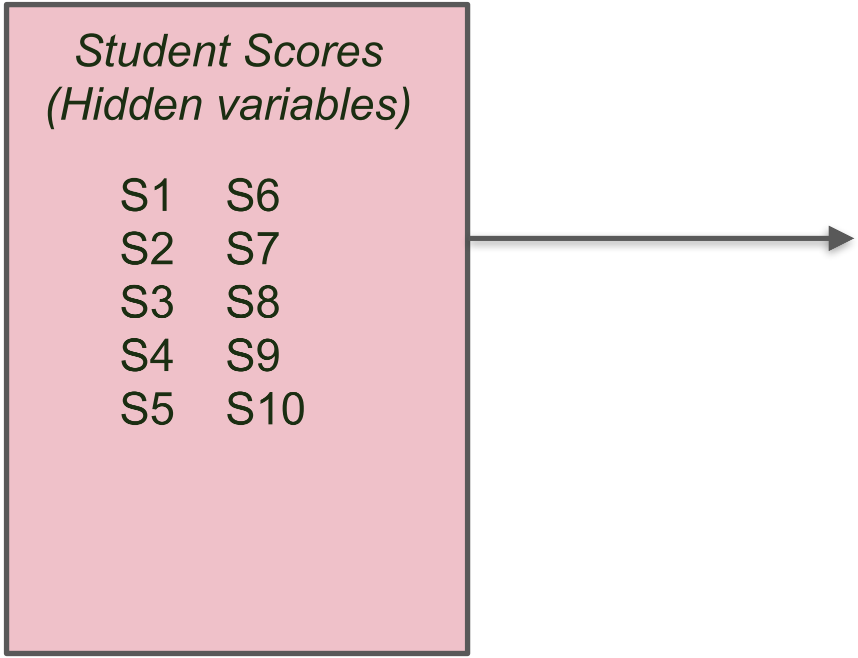
ChatGPT

ChatGPT

The teacher announces that the average score is 98%

Now you know the grades for everyone in the class! (They all scored higher.)

Every published statistic is a *constraint* on the confidential data

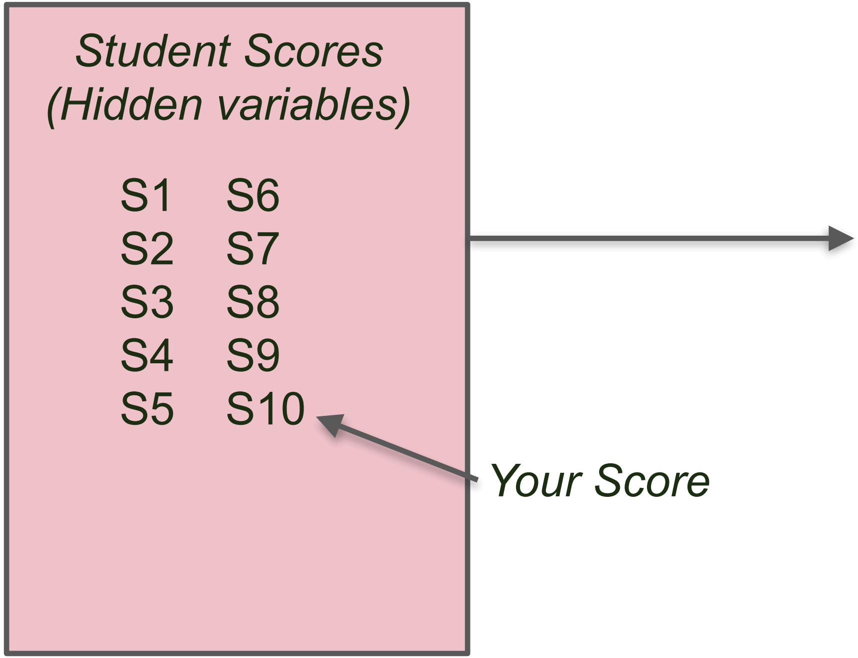


Student Scores
(Hidden variables)

S1 S6
S2 S7
S3 S8
S4 S9
S5 S10

Class Average = 98 (published)

These constraints can be mathematically *solved*.



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

“The Fundamental Law of Information Recovery”

Every statistical release based on confidential data leaks some aspect of the confidential data on which it is based.

If the results of enough queries are released, the entire confidential database is eventually revealed.

or more simply…

“Overly accurate answers to too many questions will destroy privacy in a spectacular way.”

> — *The Algorithmic Foundations of Differential Privacy by Cynthia Dwork and Aaron Roth. Foundations and Trends in Theoretical Computer Science. Vol. 9, no. 3–4, pp. 211-407, Aug. 2014. DOI:10.1561/0400000042*

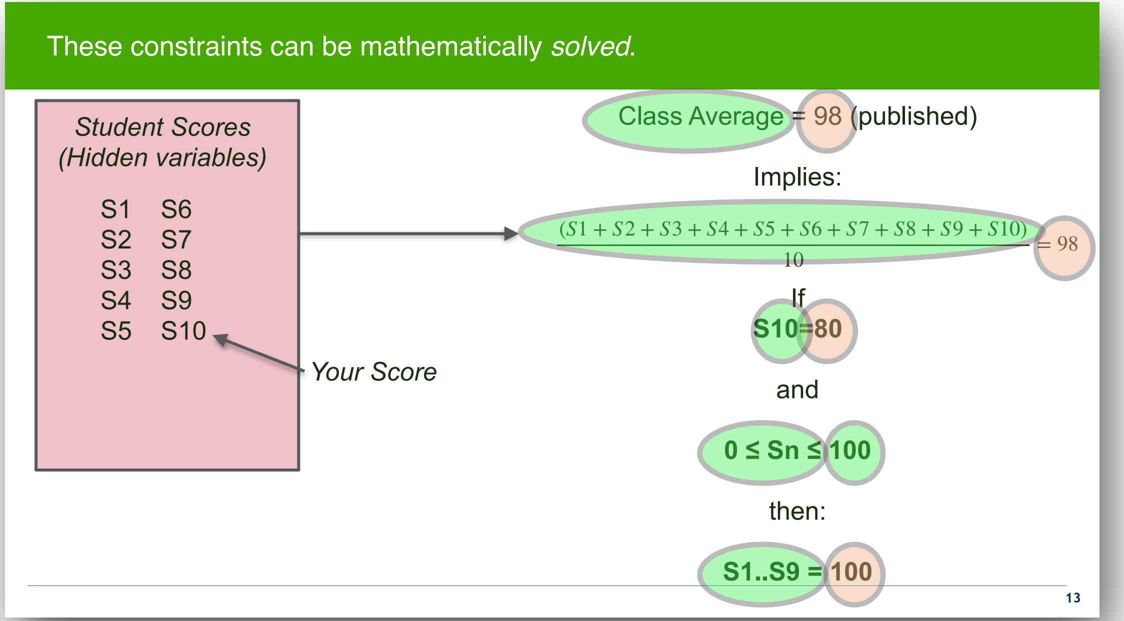
Sometimes “too many questions” is “1 question” (as evidenced here)

The only way to avoid this privacy problem is to break the constraints

There are two kinds of constrains in this system:

- data dependent constraints
- structural constraints

We can’t break the structural constraints



These constraints can be mathematically *solved*.

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

What if the teacher said the class average is between 98 and 100?

This is called *generalization* or *coarsening*

- Frequently used to protect statistical data

Class Average = 98 .. 100

But wait —

- Those other 9 students all got 100!
- They are all friends and get together at lunch.

Now they can compute *bounds* on your grade:

90 ≤ 98 + (your grade) / 10 ≤ 100

They know you got between an 80 and 100!

Generalization seems to do a good job — but the privacy guarantees don't compose. If the class average is 98..100, that sets a *hard limit* on your possible scores.

The other 9 students in the class know that you got between 80 and 100.

They can combine this with other information to learn your grade.

- They are all West Side kids!

Generalization seems to do a good job — but the privacy guarantees don't compose. If the class average is 98..100, that sets a *hard limit* on your possible scores.

The other 9 students in the class know that you got between 80 and 100.

They can combine this with other information to learn your grade.

- They are all West Side kids!

West Side kids

Your class

East Side kids

You

Generalization seems to do a good job — but the privacy guarantees don’t compose. If the class average is 98..100, that sets a *hard limit* on your possible scores.

The other 9 students in the class know that you got between 80 and 100.

They can combine this with other information to learn your grade.

- They are all West Side kids!

West Side kids

Your class

East Side kids

You

Your school publishes a report looking at math scores of all students in the grade.

- The school is proud of its East Side scores!

# of students in grade with math scores in range

	0-60	61-70	71-80	81-90	91-100
West Side kids	10	15	75	95	65
East Side kids	15	45	125	0	0

Your school reveals your grade!

Differential privacy *composes*.

The privacy guarantees of differential privacy cannot be undone by:

- Combining the results with other data releases based on the same confidential data
- Post-processing

DP is also “future proof.” It’s protected against:

- Advances in computing power
- Advances in mathematics

DP does this by:

- Having a mathematical definition of “privacy loss” that results from publishing data
- Proving that every DP “release mechanism” follows the definition
- Proofs can’t be proven wrong by advances in computing power or new math

A common form of differential privacy:
Add a random value (x) to each statistic before it is published.

What we compute:

Class Average = 98 (published) + **x** = 96.5

What we publish:

Class Average = [redacted] = 96.5

The math of differential privacy computes **x**

How do we pick **x**?

What we compute:

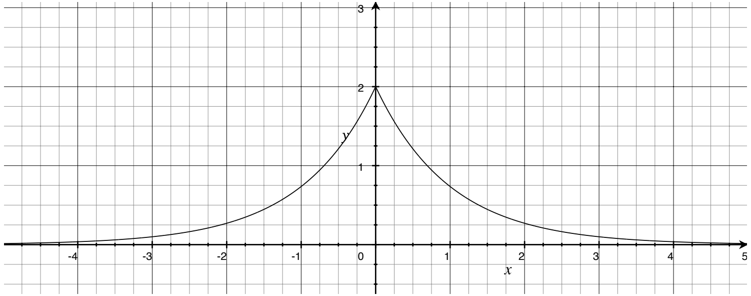
Class Average = 98 (published) **+ x** = 96.5

If **x** is between -3 and 3:

- The reported class average will be between 95 and 101 — this is still a hard limit!
- This is what publishing a range does. It is not DP.

DP draws **x** from a range of random numbers that goes from -∞ to +∞

- 95% chance of being between -3 and 3 (tunable)
- 5% chance of being < -3 or > +3 (tunable)



In theory, the class average might be reported as anything from -∞ to +∞

- In practice, the class average will typically be reported between 97 and 99
- The class average will almost never be reported <95
- You can also "clip" the output and not report anything over 100. This is an example of "post processing"

Differential Privacy is tunable

The “privacy loss” of differential privacy is a tunable parameter: 0 ≤ ε ≤ ∞

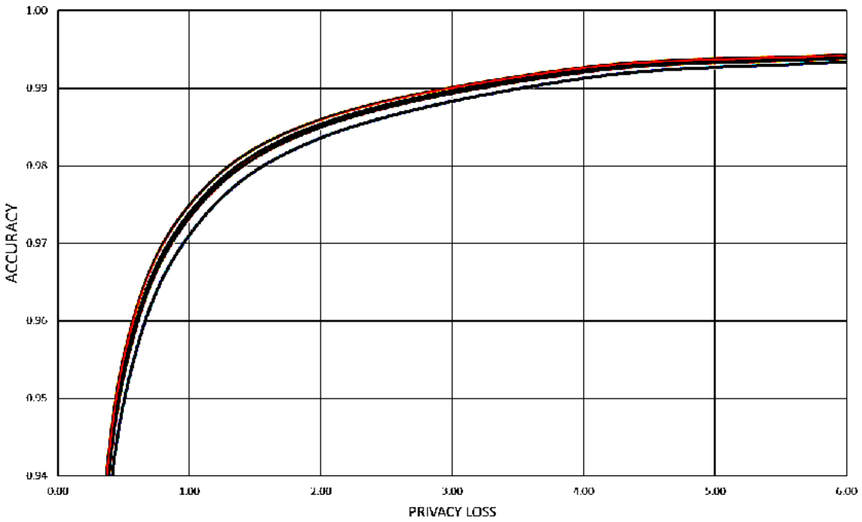
There’s no “right” value for ε — it’s a policy decision

- More privacy loss = more accurate statistics.
- (Public good) vs. (personal privacy protection)

DP tells data users the accuracy of data.

- Accuracy is hidden by other approaches.

More “privacy loss”
More accuracy



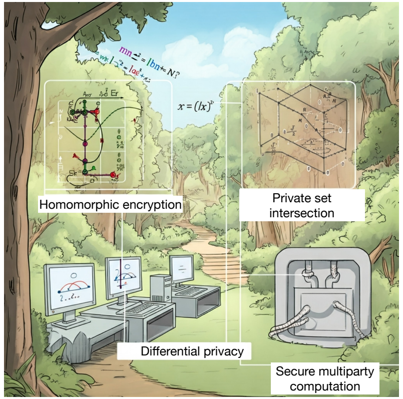
Less “privacy loss”
Less accuracy

This talk introduces PETs and shows how they can benefit governments and business

Outline for this talk:

- A secure multiparty computation (SMPC) example ✓
- Output controls — why secure computation is not enough ✓
- Introduction to the PETs Zoo
- How we brought DP to the 2020 Census
- The Census 2020 DP backlash!
- Find out more!

Introduction to the PETs Zoo



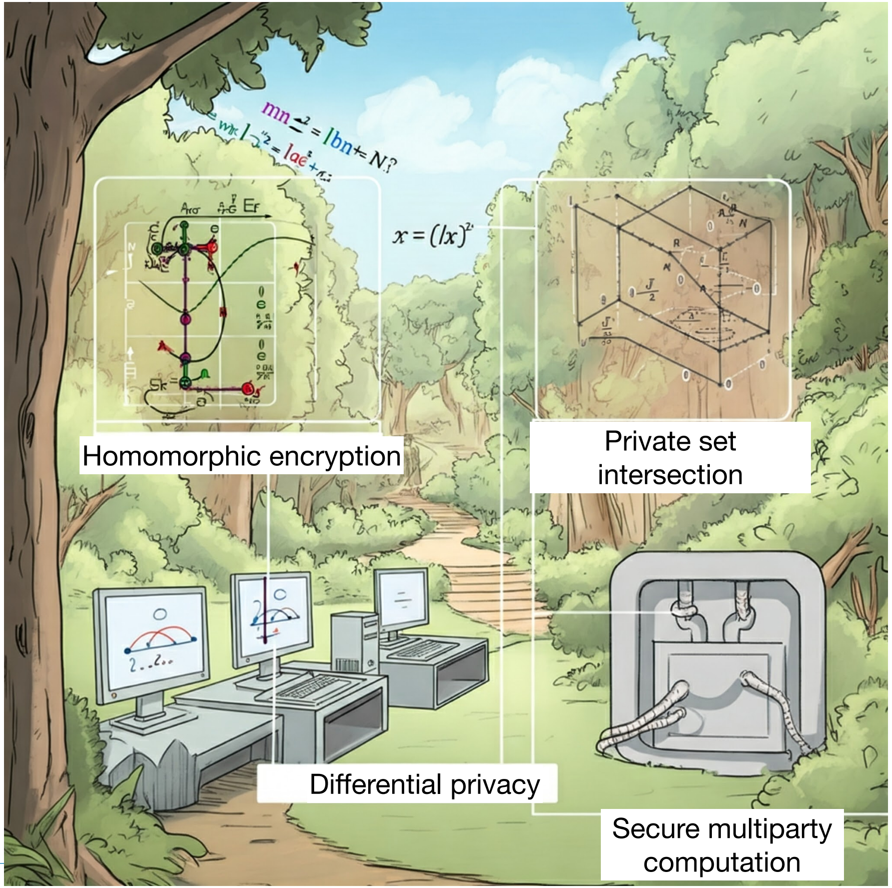
Today there are dozens of PET algorithms.

Differential privacy

- Counting queries (like the US Census)
- Averages and regressions
- Machine learning — protects training data!

Private Information Retrieval

- Encrypted data stored in a database.
- Retrieve it without revealing search keys or which items were accessed



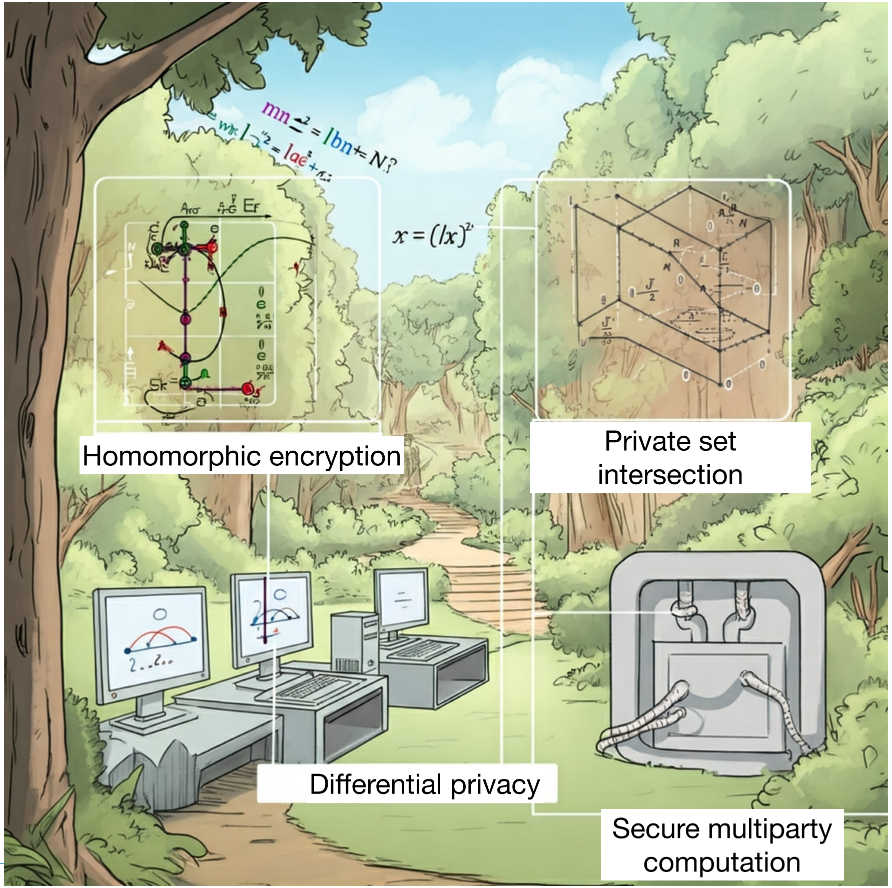
Private Set Intersection

Two organizations can count the number of matching records without revealing any other information.

The organizations can compute *any function* on the intersection records.

Example:

- Two companies count the number of customers they have in common with incomes in the $100k-$200k range without revealing the customers to each other.
- Allows data analysis without data sharing agreements or secure enclaves.



Homomorphic Encryption

Compute any function on encrypted data.

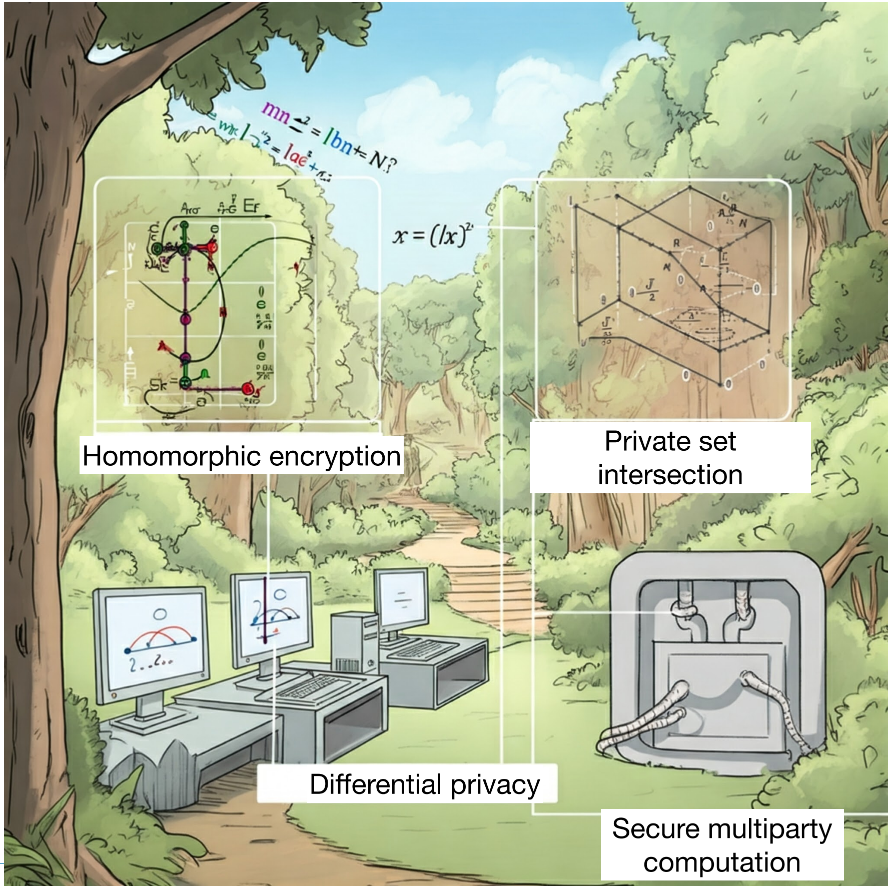
- Without first decrypting the data!
- The result is “encrypted” too!

Applications:

- End-to-end encryption for data pipelines using sensitive data.
- Outsourced computing on sensitive data.

Caveats:

- Dramatically slower than other PETs (currently)



How we brought DP to the 2020 Census

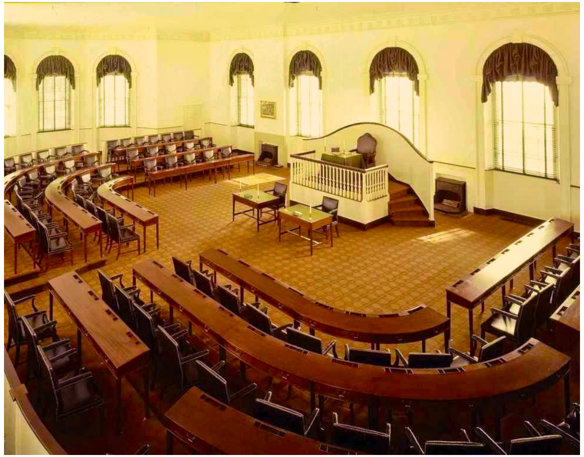
The US Census is the world's longest running statistical program.

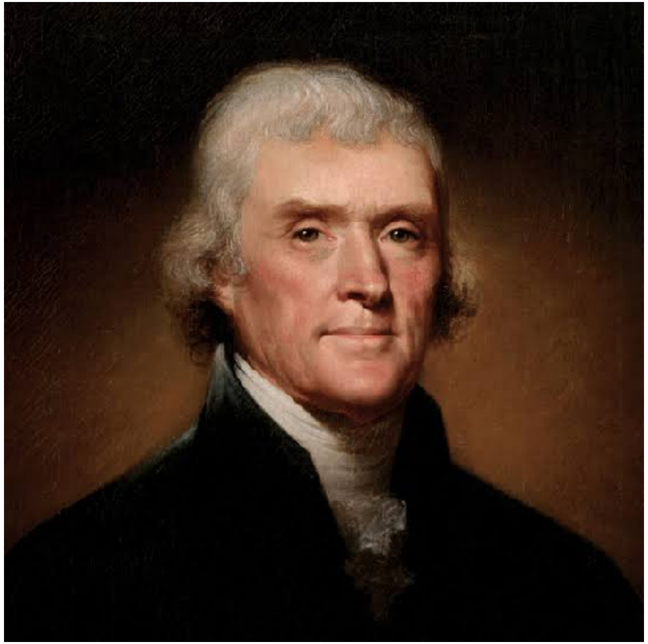
First US Census:

1790

Purpose:

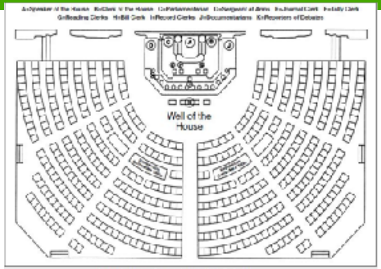
Apportion the US House of Representatives



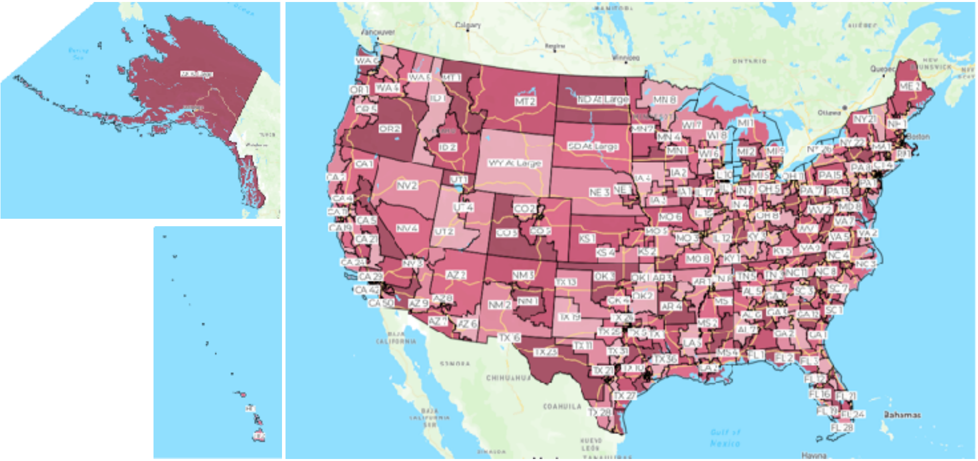


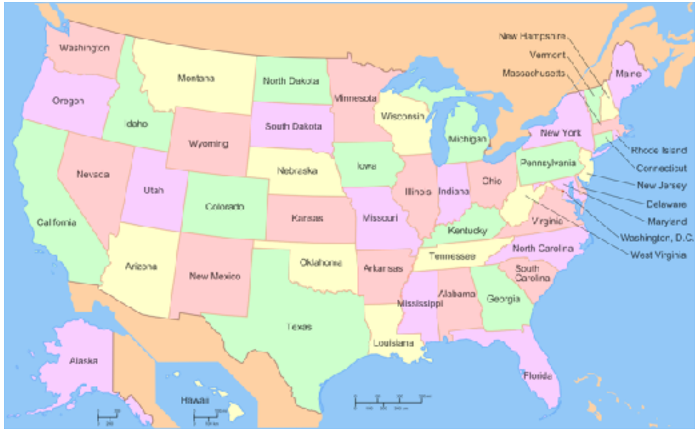
Thomas Jefferson
Primary author, US Declaration of Independence
First US Secretary of State
First US Patent Commissioner (reviewed every patent)
Oversaw first US Census

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



DEMOCRATS
REPUBLICANS
VP





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

There have been 435 seats since 1912

Each state elects 2 senators

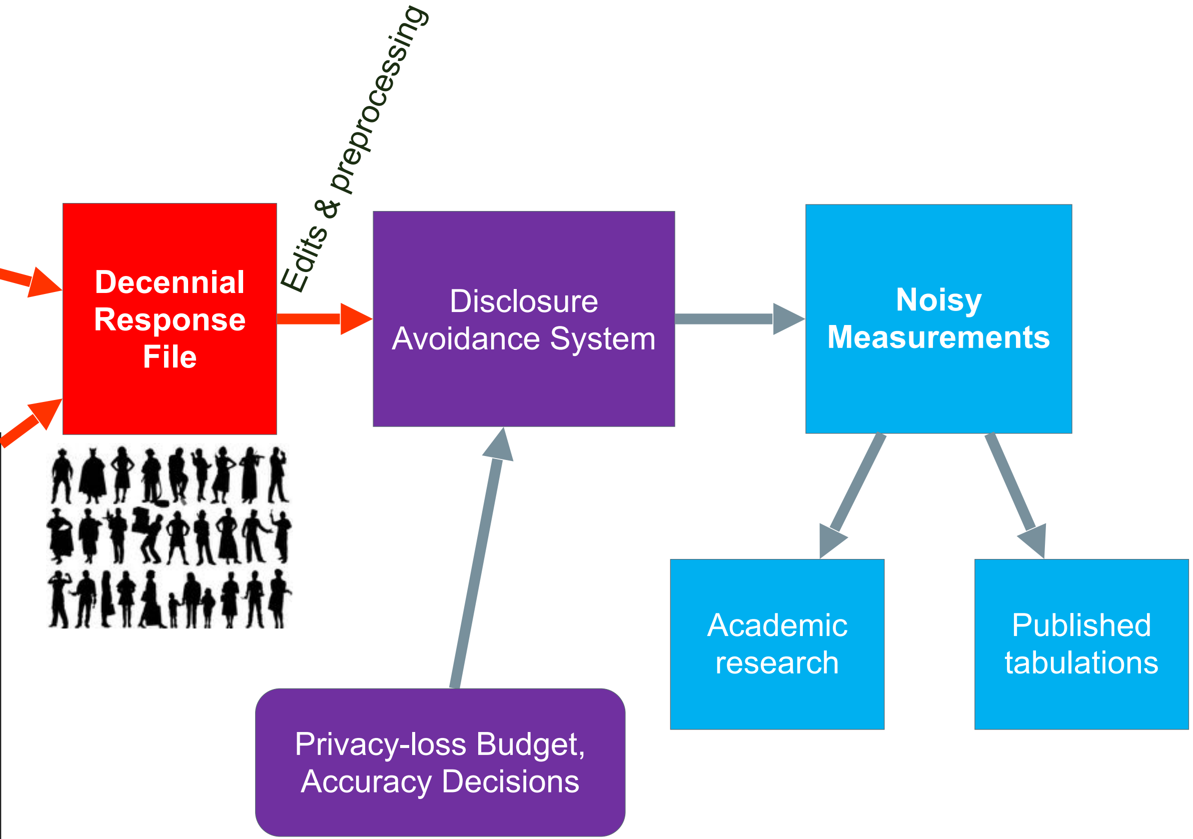
The 2020 Census asked every household in the US to fill out a survey.

Counting everyone in your household can shape your future.

What else do I need to know?

Responding to the 2020 Census is:

- **Easy**
  Every household in the United States will receive a notice to complete the census in early 2020. You can complete the form online, by phone, or by mail.
- **Safe**
  Your personal information is confidential, is protected by law, and can never be used to identify you. It can never be shared with law enforcement agencies or your property manager.
- **Important**
  Businesses and leaders in your community will use the data collected in the census to make decisions about where to build new buildings, revitalize old ones, open stores, create jobs, and more.



Decennial Response File

Edits & preprocessing

Disclosure Avoidance System

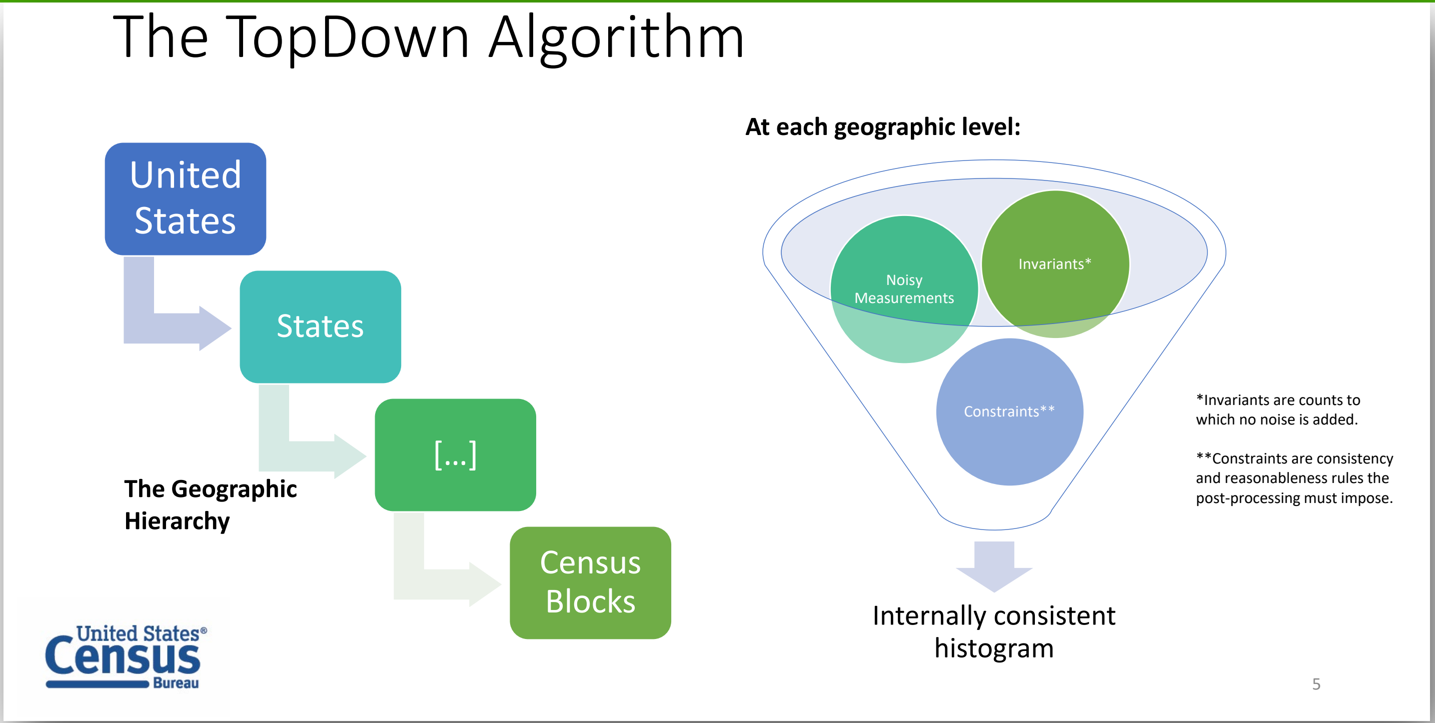
Noisy Measurements

Academic research

Published tabulations

Privacy-loss Budget, Accuracy Decisions

This was implemented using the TopDown Algorithm



The TopDown Algorithm

United States

States

[...]

Census Blocks

The Geographic Hierarchy

At each geographic level:

Noisy Measurements

Invariants*

Constraints**

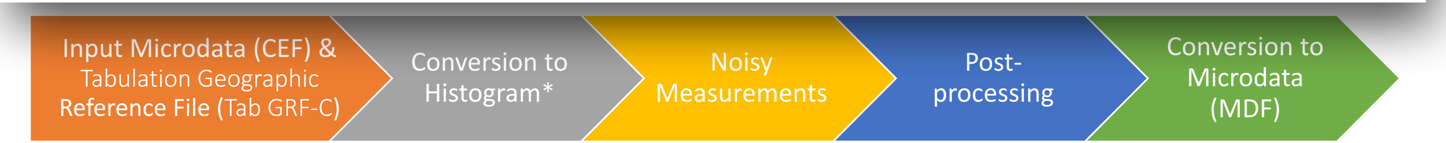
Internally consistent histogram

*Invariants are counts to which no noise is added.

**Constraints are consistency and reasonableness rules the post-processing must impose.

United States® Census Bureau

5


Input Microdata (CEF) & Tabulation Geographic Reference File (Tab GRF-C)
Conversion to Histogram*
Noisy Measurements
Post-processing
Conversion to Microdata (MDF)

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

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:

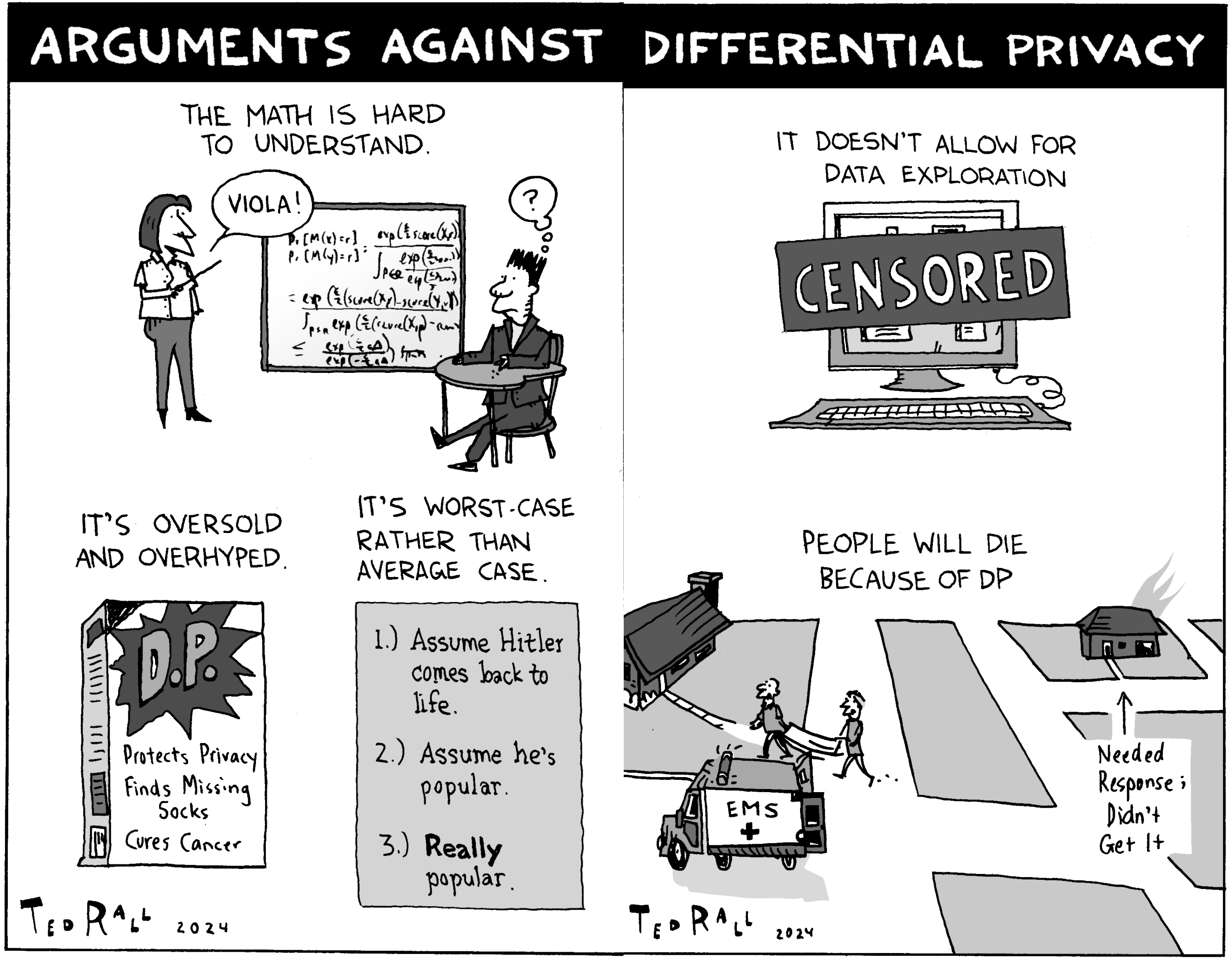
- Python creates ~ 16M mixed integer linear programs solved with Gurobi
- 20-50 AWS 96-core servers with 768GiB RAM

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

The Census 2020 DP Backlash


ARGUMENTS AGAINST DIFFERENTIAL PRIVACY
THE MATH IS HARD TO UNDERSTAND.
VIOLA!
?
IT DOESN'T ALLOW FOR DATA EXPLORATION
CENSORED
IT'S OVERSOLD AND OVERHYPED.
D.P.
Protects Privacy
Finds Missing Socks
Cures Cancer
IT'S WORST-CASE RATHER THAN AVERAGE CASE.
1.) Assume Hitler comes back to life.
2.) Assume he's popular.
3.) Really popular.
PEOPLE WILL DIE BECAUSE OF DP
EMS
Needed Response; Didn't Get It
TED RALL 2024
TED RALL 2024

Data User Challenges - Identified in July 2019

Differential privacy was not widely known or understood.

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

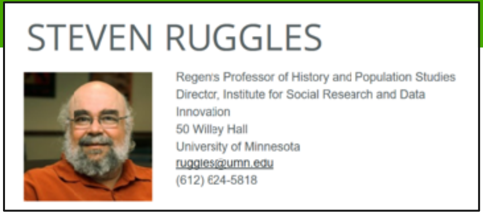
- Some were anxious about the intentional addition of noise.
- Some were concerned that previous studies done with swapped data might not be replicated if they used DP data.

Many data users believed 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.

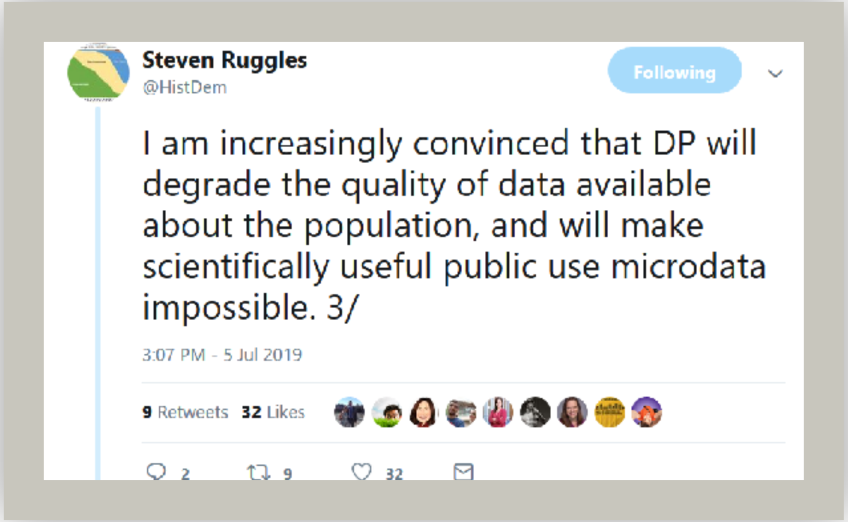
Dr. Steven Ruggles voiced concerns about DP in the 2020 Census.

Ruggles is the Regents Professor of History and Population Studies at the University of Minnesota, and the director of the Institute for Social Research and Data Innovation.



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

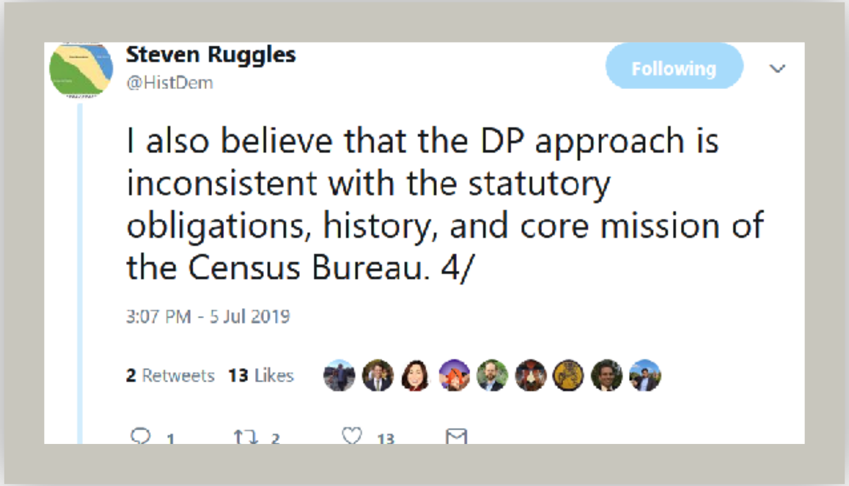


Steven Ruggles
@HistDem

I am increasingly convinced that DP will degrade the quality of data available about the population, and will make scientifically useful public use microdata impossible. 3/

3:07 PM - 5 Jul 2019

9 Retweets 32 Likes



Steven Ruggles
@HistDem

I also believe that the DP approach is inconsistent with the statutory obligations, history, and core mission of the Census Bureau. 4/

3:07 PM - 5 Jul 2019

2 Retweets 13 Likes

Steven Ruggles 5 Jul 2019

Ruggles organized data users against DP.

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:

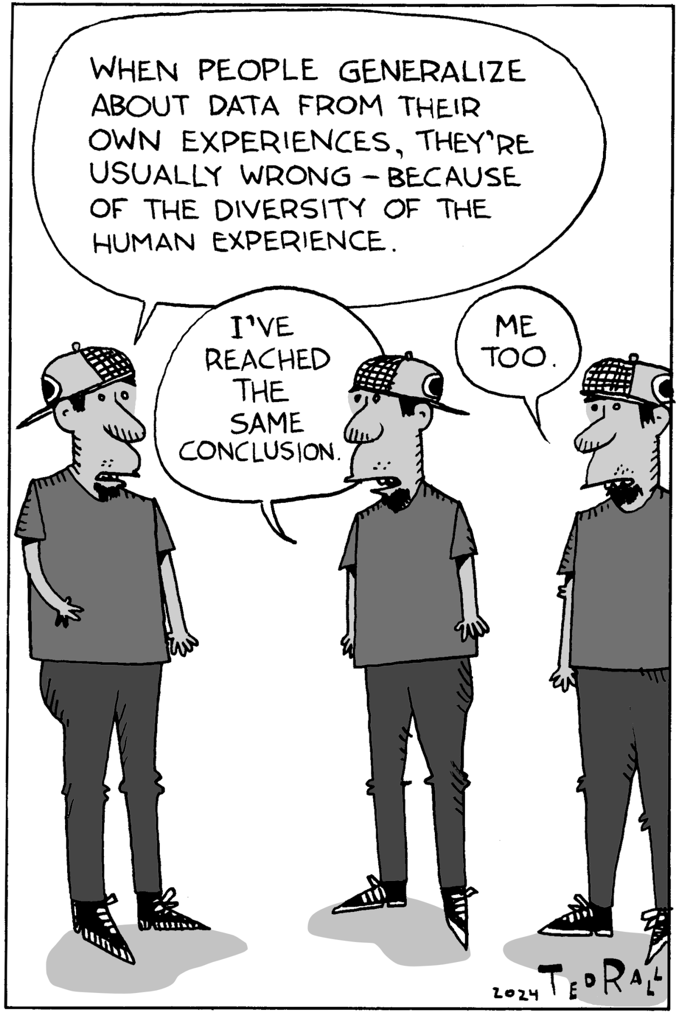
- Organized petition with 4000+ signers asking for no DP in 2020 Census.

Results:

- The US Census Bureau organized events and adjusted the algorithm and parameters to address man data user concerns.
- Plans were shelved to rapidly deploy DP for the American Community Survey.

The problem:
There is a lot more diversity 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).

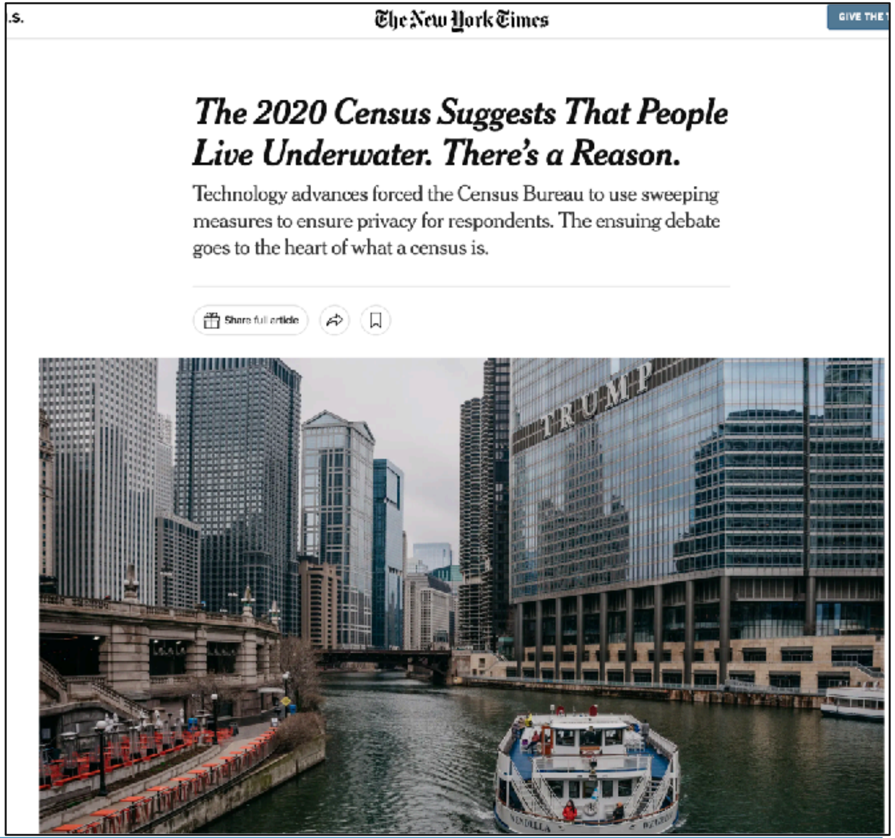

WHEN PEOPLE GENERALIZE ABOUT DATA FROM THEIR OWN EXPERIENCES, THEY'RE USUALLY WRONG – BECAUSE OF THE DIVERSITY OF THE HUMAN EXPERIENCE.
I'VE REACHED THE SAME CONCLUSION.
ME TOO.
2024 TED RALL

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



.S.

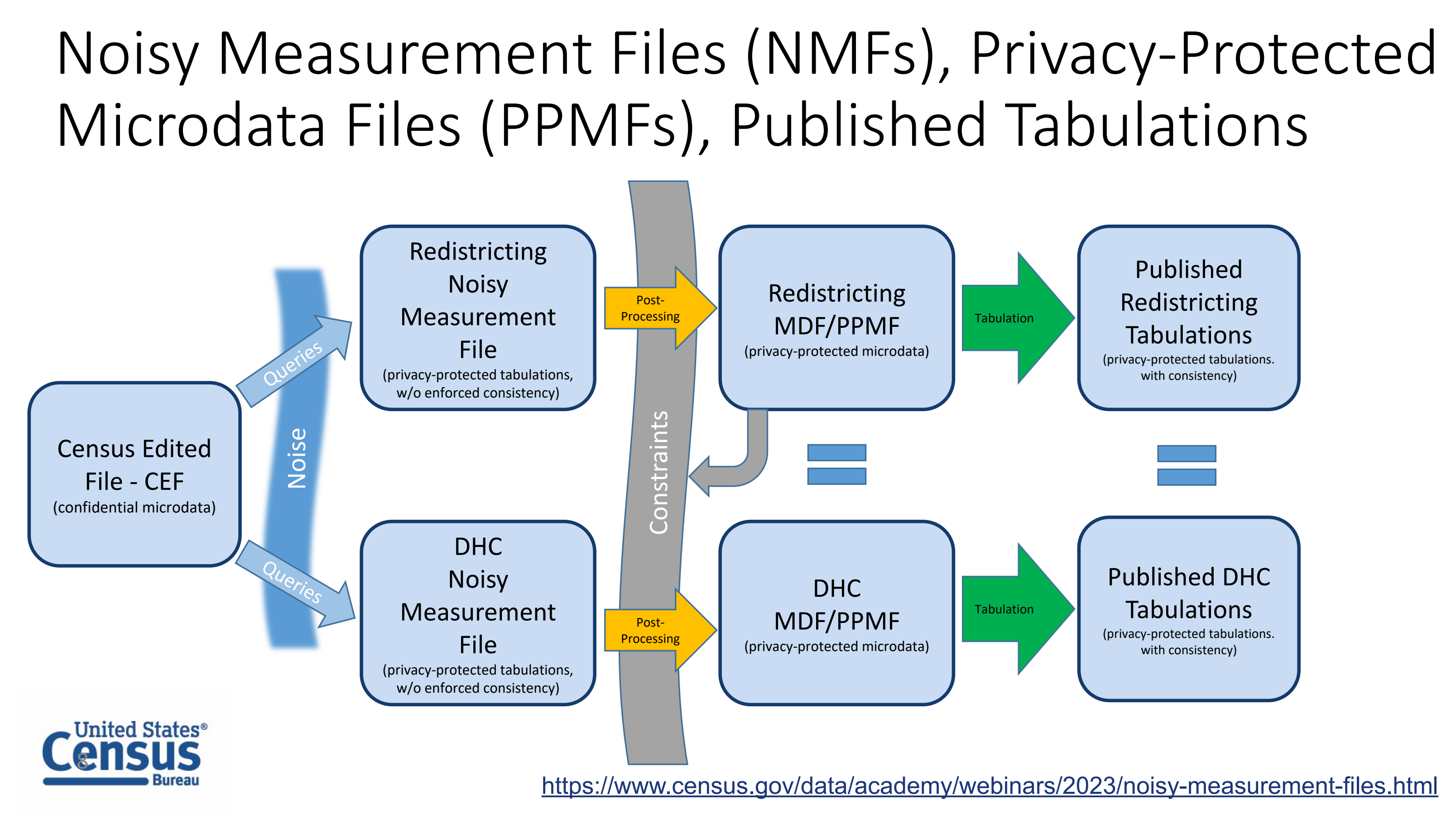
The New York Times

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.

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



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

Census Edited File - CEF (confidential microdata)

Noise

Queries

Queries

Redistricting Noisy Measurement File (privacy-protected tabulations, w/o enforced consistency)

DHC Noisy Measurement File (privacy-protected tabulations, w/o enforced consistency)

Constraints

Post-Processing

Post-Processing

Redistricting MDF/PPMF (privacy-protected microdata)

DHC MDF/PPMF (privacy-protected microdata)

Tabulation

Tabulation

Published Redistricting Tabulations (privacy-protected tabulations. with consistency)

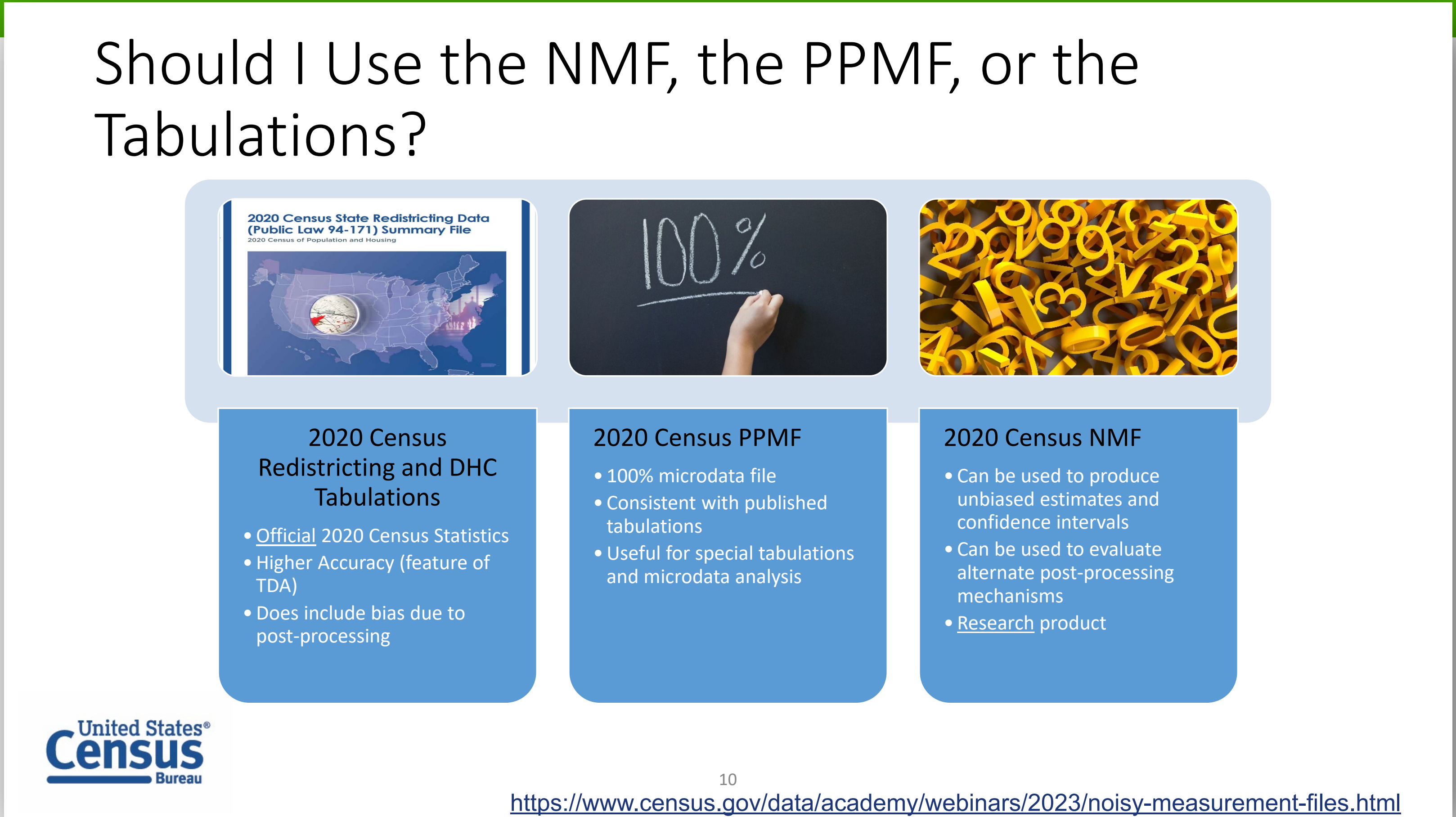
Published DHC Tabulations (privacy-protected tabulations. with consistency)

=

=

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

Multiple data products creates confusion…


2020 Census State Redistricting Data (Public Law 94-171) Summary File
2020 Census of Population and Housing
100%

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

10

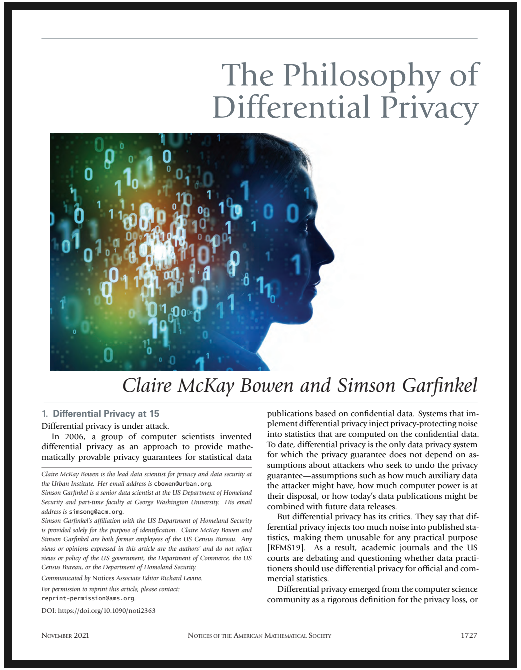
https://www.census.gov/data/academy/webinars/2023/noisy-measurement-files.html

This talk introduces PETs and shows how they can benefit governments and business

Outline for this talk:

- A secure multiparty computation (SMPC) example ✓
- Output controls — why secure computation is not enough ✓
- Introduction to the PETs Zoo ✓
- How we brought DP to the 2020 Census ✓
- The Census 2020 DP backlash! ✓
- Find out more!

There’s a lot more to say…



The Philosophy of Differential Privacy

Claire McKay Bowen and Simson Garfinkel

1. Differential Privacy at 15

Differential privacy is under attack.

In 2006, a group of computer scientists invented differential privacy as an approach to provide mathematically provable privacy guarantees for statistical data publications based on confidential data. Systems that implement differential privacy inject privacy-protecting noise into statistics that are computed on the confidential data. To date, differential privacy is the only data privacy system for which the privacy guarantee does not depend on assumptions about attackers who seek to undo the privacy guarantee—assumptions such as how much auxiliary data the attacker might have, how much computer power is at their disposal, or how today’s data publications might be combined with future data releases.

But differential privacy has its critics. They say that differential privacy injects too much noise into published statistics, making them unusable for any practical purpose [RFMS19]. As a result, academic journals and the US courts are debating and questioning whether data practitioners should use differential privacy for official and commercial statistics.

Differential privacy emerged from the computer science community as a rigorous definition for the privacy loss, or

Claire McKay Bowen is the lead data scientist for privacy and data security at the Urban Institute. Her email address is `cbowen@urban.org`.

Simson Garfinkel is a senior data scientist at the US Department of Homeland Security and part-time faculty at George Washington University. His email address is `simsong@acm.org`.

Simson Garfinkel’s affiliation with the US Department of Homeland Security is provided solely for the purpose of identification. Claire McKay Bowen and Simson Garfinkel are both former employees of the US Census Bureau. Any views or opinions expressed in this article are the authors’ and do not reflect views or policy of the US government, the Department of Commerce, the US Census Bureau, or the Department of Homeland Security.

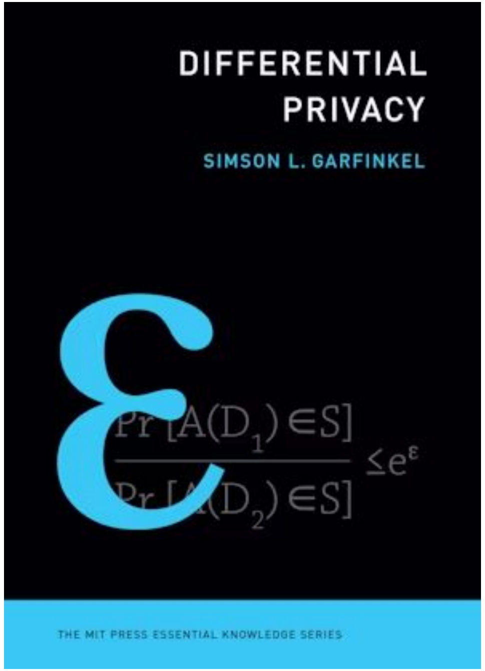
Communicated by Notices *Associate Editor Richard Levine.*

For permission to reprint this article, please contact: `reprint-permission@ams.org`.

DOI: https://doi.org/10.1090/noti2363

November 2021 Notices of the American Mathematical Society 1727

“The Philosophy of Differential Privacy”
Bowen & Garfinkel
Notices of the American Mathematical Society
November 2021



DIFFERENTIAL PRIVACY

SIMSON L. GARFINKEL

$$\frac{\Pr[A(D_1)\in S]}{\Pr[A(D_2)\in S]} \le e^{\varepsilon}$$

THE MIT PRESS ESSENTIAL KNOWLEDGE SERIES

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