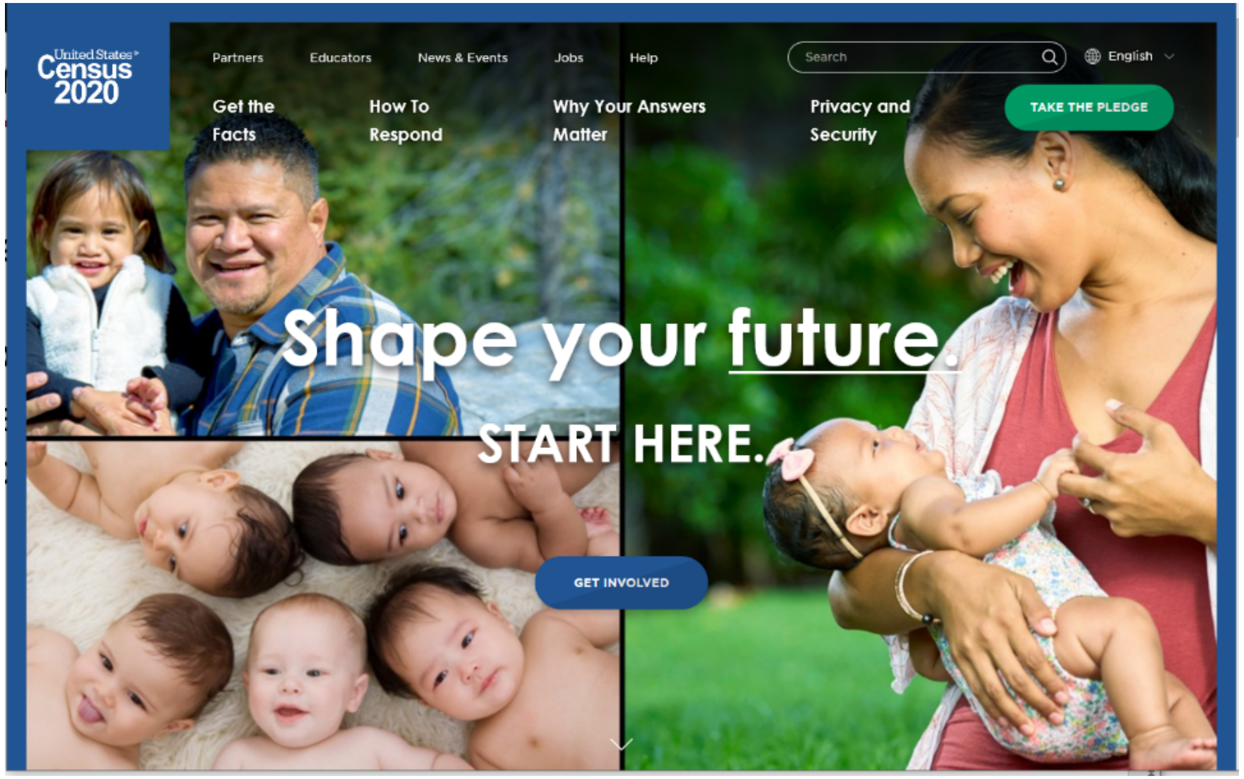
# Differential Privacy and the 2020 Census

**Simson L. Garfinkel**
**U.S. Census Bureau**

March 3, 2020
Presentation at Google




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

Shape
your future
START HERE >

United States Census 2020

# Abstract

The goal of the 2020 Census is to count everyone once, only once and in the right place. The decennial activity, mandated by the US Constitution, was first overseen by Thomas Jefferson in 1790 is the oldest continuously operating statistical program on the planet.

As part of the 2020 Census, each household in the United States will be asked to provide the number of residents, their ages, sex, race, ethnicity, as well as the inter-household relationship. This information will used to apportion the House of Representatives and to distribute more than $675 billion in federal aid to the US states.

The Census Bureau is legally prohibited from making publications in which the data contributed by a specific individual or establishment can be identified. Advances in computer performance, computer science, and the availability of “big data” makes that harder today than ever before.

In 2018-19, the Census Bureau conducted a “red-team” attack against the data that it published from the 2010 census and discovered that it could reconstruct microdata for all 308,745,538 residents, and that it could correctly re-identify data from 52 million.

Differential privacy was created in 2006 to precisely solve this problem. With differential privacy, it is possible to bound the privacy loss that results from a data publication, but doing so decreases the accuracy of the published data. It does this by introducing uncertainty, or error, into the published statistics. While the naïve application of differential privacy can result in substantial error for even modest privacy protection, it is possible to create sophisticated algorithms that do a better job balancing accuracy and privacy loss.

# Acknowledgments

**This presentation incorporates work by:**

John Abowd (Chief Scientist)

Dan Kifer (Scientific Lead)

Simson Garfinkel (Senior Computer Scientist for Confidentiality and Data Access)

Rob Sienkiewicz (Chief, Center for Enterprise Dissemination)

Tamara Adams, Robert Ashmead, Stephen Clark, Craig Corl, Aref Dajani, Jason Devine, Nathan Goldschlag, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Brett Moran, Edward Porter, Sarah Powazek, Anne Ross, Ian Schmutte, William Sexton, Lars Vilhuber, and Pavel Zhuralev.

# Outline

Motivation

Technology change and the US Census Bureau

Privacy protection for the 2020 Census

Challenges deploying differential privacy

The public policy questions

# Motivation


## Article 1, Section 2

**“The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct.”**

# “...in such Manner as they shall by Law direct.” Public Law 94-171

http://uscode.house.gov/statutes/pl/94/171.pdf

**PL94-171 and SF1 Statistics per Census Block:**

P1 – Total population by block x RACE (PL94-171)

P2 – Total population, (Hispanic & Not Hispanic) x RACE (PL94-171)

P3 – Race for Population 18 years and over (PL94-171)

P4 – (Hispanic & Not Hispanic) 18 years and over x RACE (PL94-171)

P12 and P12A-H – Sex By Age (23 age buckets) x RACE (SF1)

P14 – Sex By Age For Population Under 20 (20 age buckets) (SF1)

P22 – Household Type by Age of Householder (5 year buckets) (SF1)

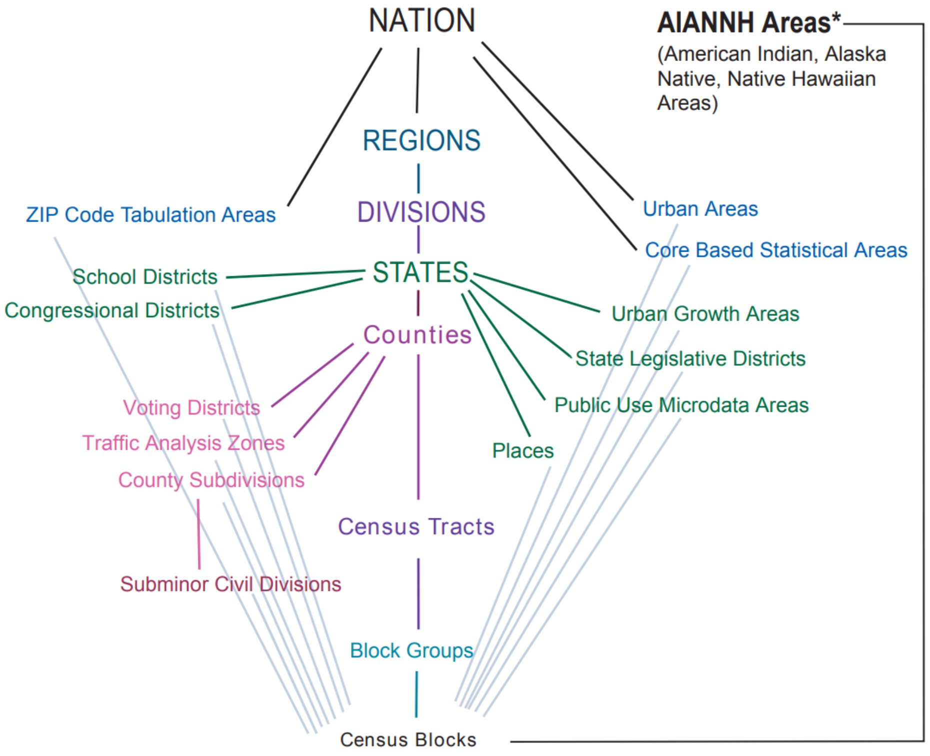
P42 – Group Quarters population by GQ type (PL94-171)

H1 – Occupancy Status (Occupied & Vacant) (PL94-171)

**SF1 Statistics per Census Tract:**

PCT12 – Sex By Age (105 age buckets)

...



# Uses of the Decennial Census Data

## Apportioning the House of Representatives (U.S. Constitution)

50 numbers of total state population as of April 1

## Enforcing Voting Rights Act of 1965 Section 2

Prohibits every state and local government from imposing any voting law that results in discrimination against racial or language minorities.

## Distributing Federal Funds

$675 billion in FY2015

# Privacy and the Decennial Census

**Title 13 Section 9 of the US Code Prohibits the US Census Bureau from making any publication that reveals data provided by a person or an establishment.**

**Respondent data cannot be used for non-statistical purposes.**

**Census Bureau employees are *sworn for life* to protect respondent data.**


https://www.census.gov/about/policies/privacy.html

# Data that we collect:

| Variable | Range | Bits |
|---|---|---|
| Block | 6,207,027 inhabited blocks | 23 |
| Sex | 2 (Female/Male) | 1 |
| Age | 103 (0-99 single age year categories, 100-104, 105-109, 110+) | 7 |
| Race | 63 allowable race combinations | 7 |
| Ethnicity | 2 (Hispanic/Not) | 1 |
| Relationship | 17 values | 5 |
| Total | | 44 |

**<u>2010 values:</u>**

**308,745,538 people x 6 variables = 1,852,473,228 measurements**

**308,745,538 people x 44 bits = 13,584,803,672 bits ≈ 1.7 GB**

# 2010 Census: Summary of Publications
## (approximate counts)

| Publication | Released counts |
|---|---|
| PL94-171 Redistricting | 2,771,998,263 |
| Balance of Summary File 1 | 2,806,899,669 |
| Total Statistics in PL94-171 and Balance of SF1: | 5,578,897,932 |
| | |
| Published Statistics/person | 18 |
| Recall: Collected variables/person: | 6 |
| **Published Statistics/collected variable** | **18 ÷ 6 ≈ 3** |

# (Dinur Nissim 2003) Database Reconstruction

**Publishing too many queries on a confidential database with too much accuracy reveals the contents of the database**

**Today we call this the "fundamental law of information recovery."**

**Dinur & Nissim proposed a generalized solution of adding noise.**

**Revealing Information while Preserving Privacy**

Irit Dinur Kobbi Nissim*
NEC Research Institute
4 Independence Way
Princeton, NJ 08540
{iritd,kobbi}@research.nj.nec.com

**ABSTRACT**

We examine the tradeoff between privacy and usability of statistical databases. We model a statistical database by an $n$-bit string $d_1, .., d_n$, with a query being a subset $q \subseteq [n]$ to be answered by $\sum_{i \in q} d_i$. Our main result is a polynomial reconstruction algorithm of data from noisy (perturbed) subset sums. Applying this reconstruction algorithm to statistical databases we show that in order to achieve privacy one has to add perturbation of magnitude $\Omega(\sqrt{n})$. That is, smaller perturbation always results in a strong violation of privacy. We show that this result is tight by exemplifying access algorithms for statistical databases that preserve privacy while adding perturbation of magnitude $\tilde{O}(\sqrt{n})$.

For time-$T$ bounded adversaries we demonstrate a privacy-preserving access algorithm whose perturbation magnitude is $\approx \sqrt{T}$.

**Keywords**

Integrity and Security, Data Reconstruction, Subset-sums with noise.

**1. INTRODUCTION**

One simple tempting solution is to remove from the database all 'identifying' attributes such as the patients' names and social security numbers. However, this solution is not enough to protect patient privacy since there usually exist other means of identifying patients, via *indirectly identifying* attributes stored in the database. Usually, identification may still be achieved by coming across just a few 'innocuous' looking attributes[1].

The topic of this work is to explore the conditions under which such a privacy preserving database access mechanism can exist.

**A Threshold for Noisy Reconstruction.** Viewing query-answer pairs as an 'encoding' of the bits $d_1, .., d_n$, the goal of the privacy-breaking adversary is to efficiently 'decode' this encoding i.e. to obtain values of some $d_i$s. In our setting, the 'decoding' algorithm is given access to subset sums of the $d_i$s perturbed by adding some random noise of magnitude $\leq \mathcal{E}$. We show an interesting threshold phenomenon where either almost all of the $d_i$s can be reconstructed, in case $\mathcal{E} \ll \sqrt{n}$, or none of them, when $\mathcal{E} \gg \sqrt{n}$.

**1.1 A Brief Background**

The problem of protecting sensitive information in a database while allowing *statistical* queries (i.e. queries about

# (Dwork, McSherry, Nissim & Smith 2006) Differential Privacy

**Differential Privacy tells us how much noise to add!**

**Key features:**

Lower bound for the amount of noise that needs to be added

Upper bound for privacy loss

Mechanisms are composable

## Calibrating Noise to Sensitivity in Private Dat Analysis

Cynthia Dwork[1], Frank McSherry[1], Kobbi Nissim[2], and Adam Smith[3]*

[1] Microsoft Research, Silicon Valley. {dwork,mcsherry}@microsoft.com
[2] Ben-Gurion University. kobbi@cs.bgu.ac.il
[3] Weizmann Institute of Science. adam.smith@weizmann.ac.il

**Abstract.** We continue a line of research initiated in [10, 11] on privacy-preserving statistical databases. Consider a trusted server that holds a database of sensitive information. Given a query function $f$ mapping databases to reals, the so-called *true answer* is the result of applying $f$ to the database. To protect privacy, the true answer is perturbed by the addition of random noise generated according to a carefully chosen distribution, and this response, the true answer plus noise, is returned to the user.

Previous work focused on the case of noisy sums, in which $f$ =

# Technology and the Decennial Census

Shape your future START HERE >

United States® Census 2020

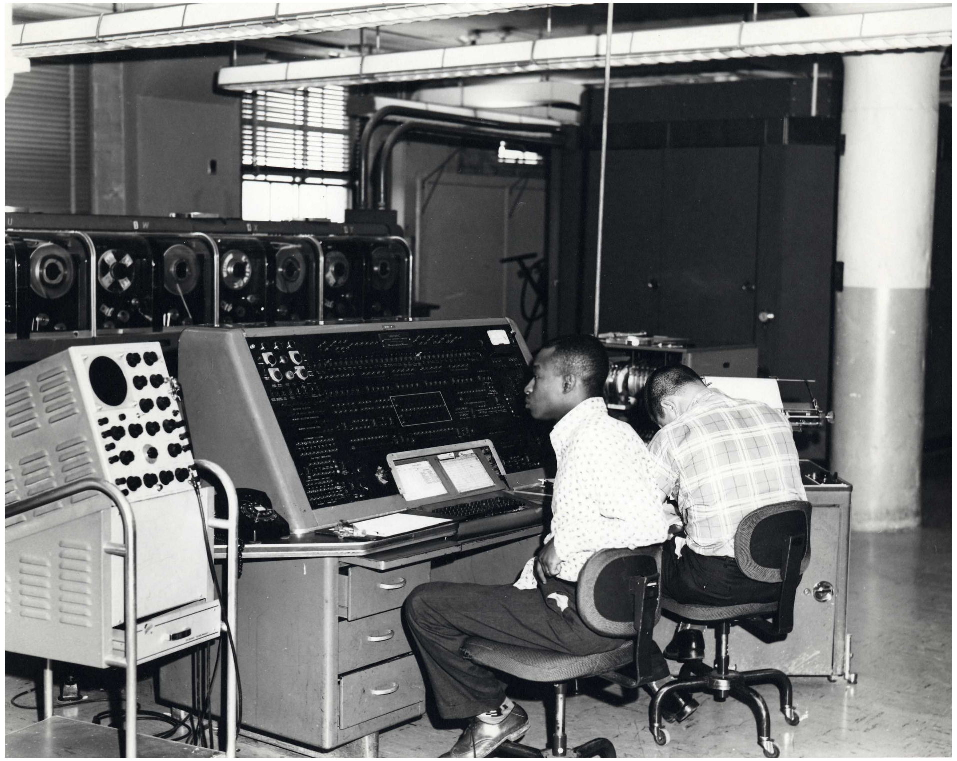
# Punch Cards were invented for the 1890 Census

https://www.census.gov/history/www/innovations/technology/the_hollerith_tabulator.html

# The Census Bureau bought UNIVAC 1, the world's first commercial general-purpose electronic digital computer

https://www.census.gov/history/www/innovations/technology/univac_i.html

June 14, 1951

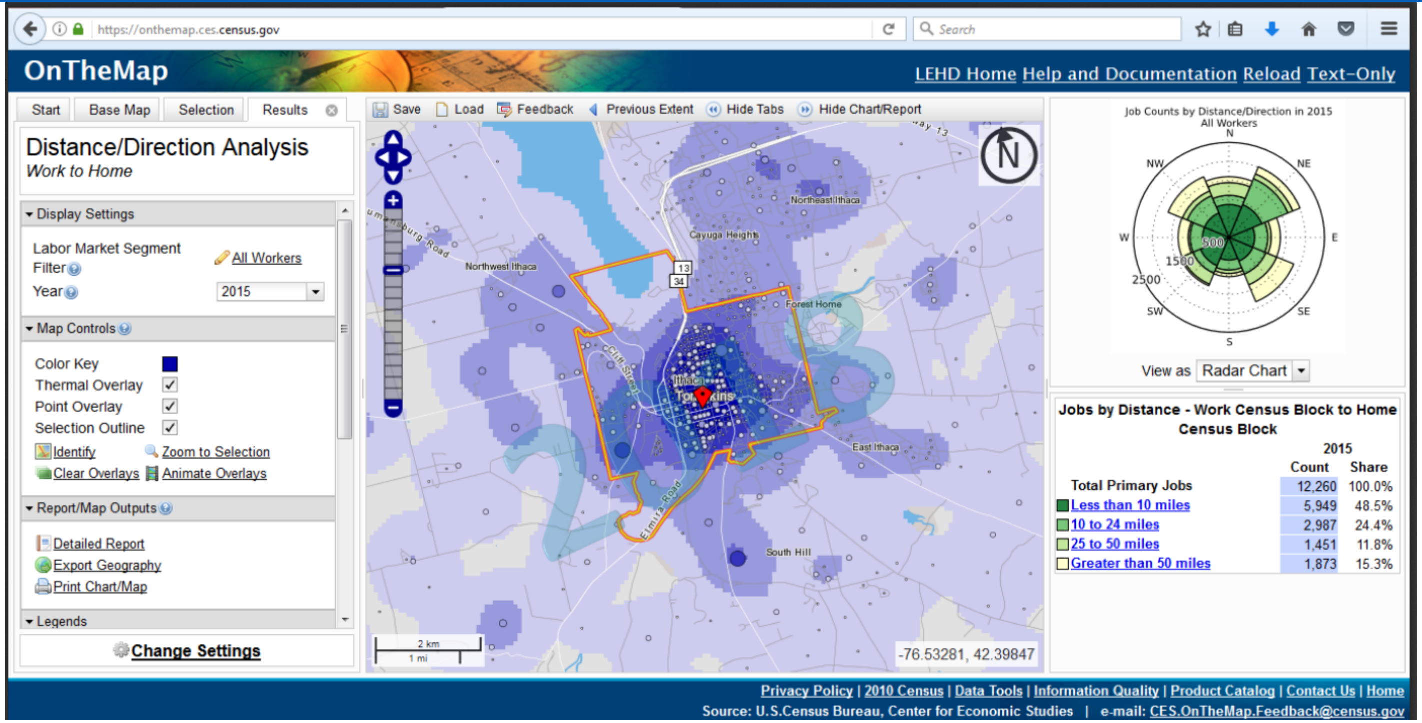


# Differential Privacy was invented for the US Census


**Cynthia Dwork at the Harvard Data Initiative Conference, October 25, 2019**

# The Census Bureau deployed DP “OnTheMap” in 2008!

https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf



# Differential privacy was not ready for the 2010 census.

**September 26, 2005** – Census Bureau awarded $500+ million contract to Lockheed Martin Corporation for the 2010 Census Decennial Response Integration System (DRIS)

**March 30, 2009** – Census Bureau launches a massive operation to verify and update more than 145 million addresses as it prepares to mail out the 2010 census questionnaire.

**March 1, 2010** – 2010 census questionnaires begin arriving in mailboxes throughout the United States and Island Areas

**April 1, 2010** – Census Day.

**December 21, 2010** – The Census Bureau announces the 2010 population counts and delivers the apportionment counts to the president.

https://www.census.gov/history/www/through_the_decades/overview/2010_overview_1.html

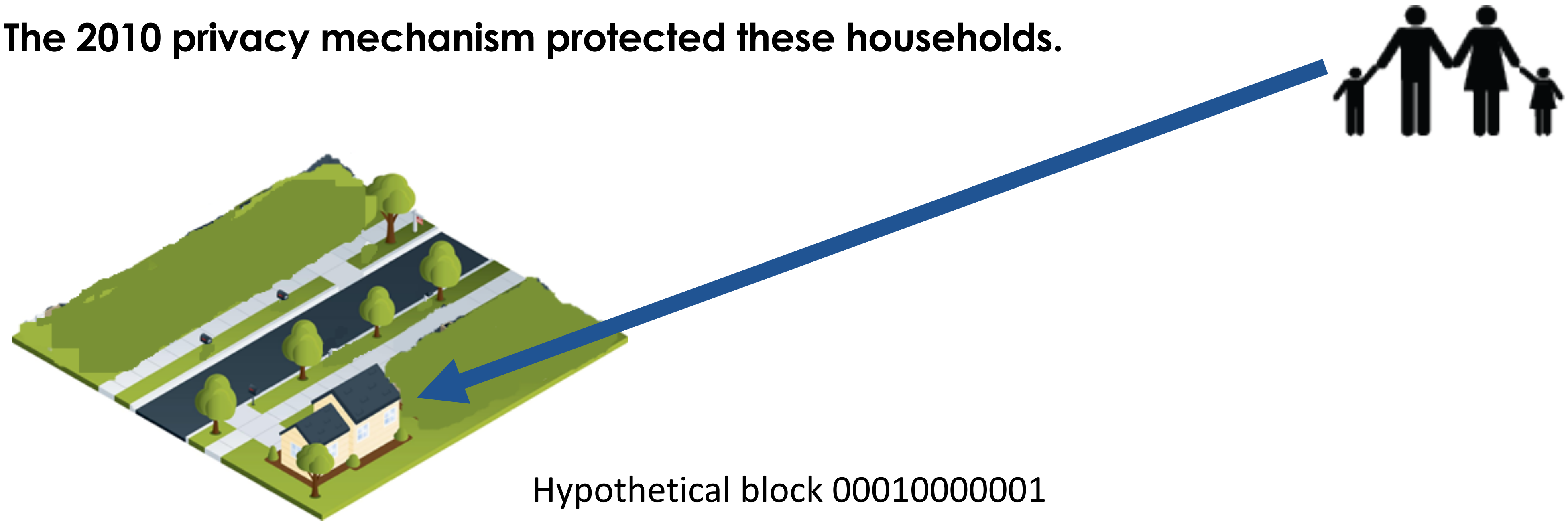
IT'S IN OUR HANDS
United States Census 2010

# The 2010 Privacy Mechanism

**Some statistics were published at the block level.**

**A single household on a block might be highly identifiable!**

**The 2010 privacy mechanism protected these households.**



Hypothetical block 00010000001

# The 2010 privacy mechanism swapped households with others the same size.

## Advantages of swapping:

Easy to understand

Does not affect state counts if swaps are within a state

Can be run state-by-state

Operation is “invisible” to rest of Census processing

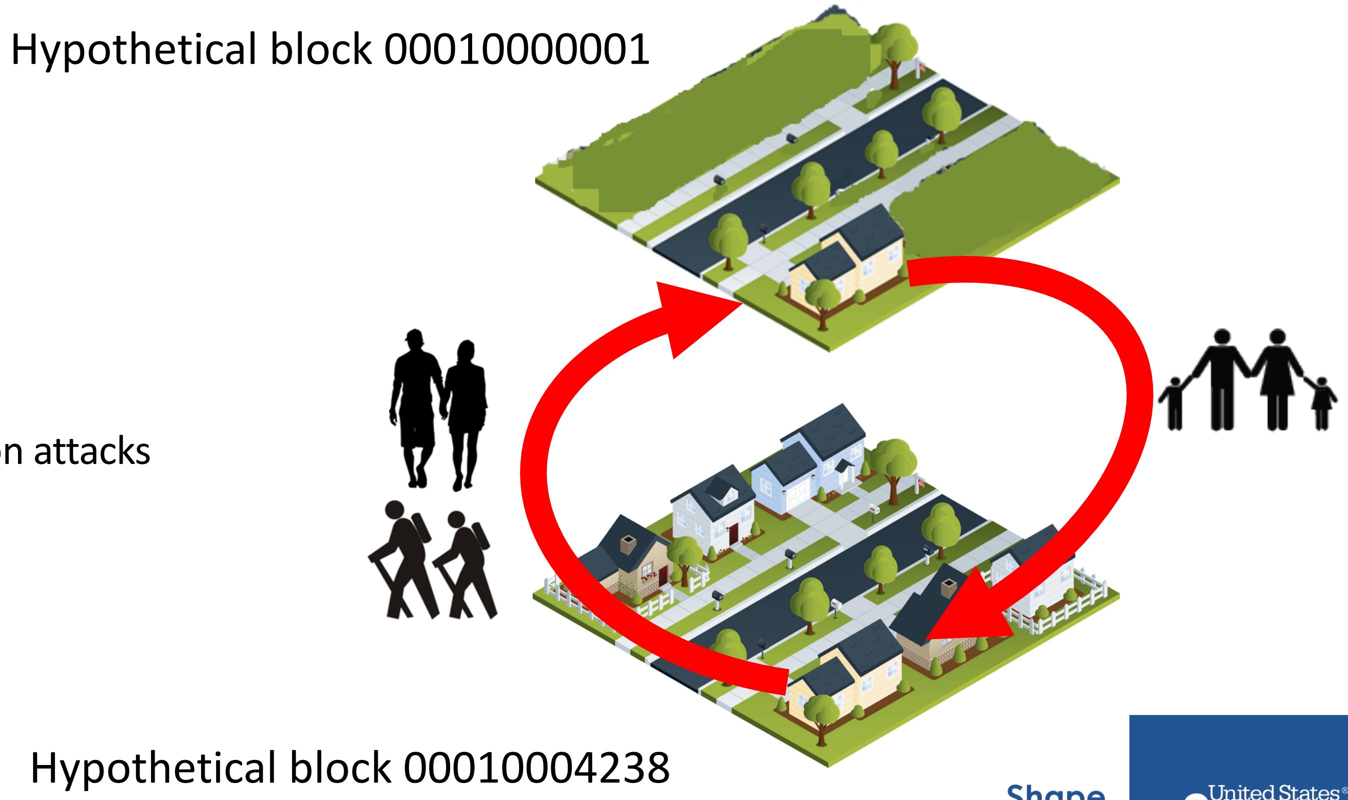
## Disadvantages:

Does not consider or protect against database reconstruction attacks

Privacy protection is not quantified

Swap rate and details of swapping must remain secret

Privacy guarantee based on the lack of external data



# Household–level swapping was applied after editing, before tabulation.

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**

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

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

# We now know that the privacy techniques we used in the 2010 Census were flawed.

These were the best available techniques at the time!

Assumed that disclosure avoidance modifications made for two products from the same confidential data are compatible

Released exact counts at the block, tract and county level.

Released exact counts for age in years,
OMB race/ethnicity, sex, relationship to householder,
in Summary File 2: detailed race data

# Statistical agencies aggregate data from many households together into a single publication.

| | Count | Median | Mean |
|---|---|---|---|
| Total | 7 | 30 | 38 |
| # female | 4 | 30 | 33.5 |
| # male | 3 | 30 | 44 |
| # black | 4 | 51 | 48.5 |
| # white | 3 | 24 | 24 |
| # married | 4 | 51 | 54 |
| # black F | 3 | 36 | 36.7 |

# We now know how to take many aggregate publications and “solve” for the original microdata.

66 FBM & 84 MBM

30 MWM & 36 FBM

8 FBS

18 MWS

24 FWS

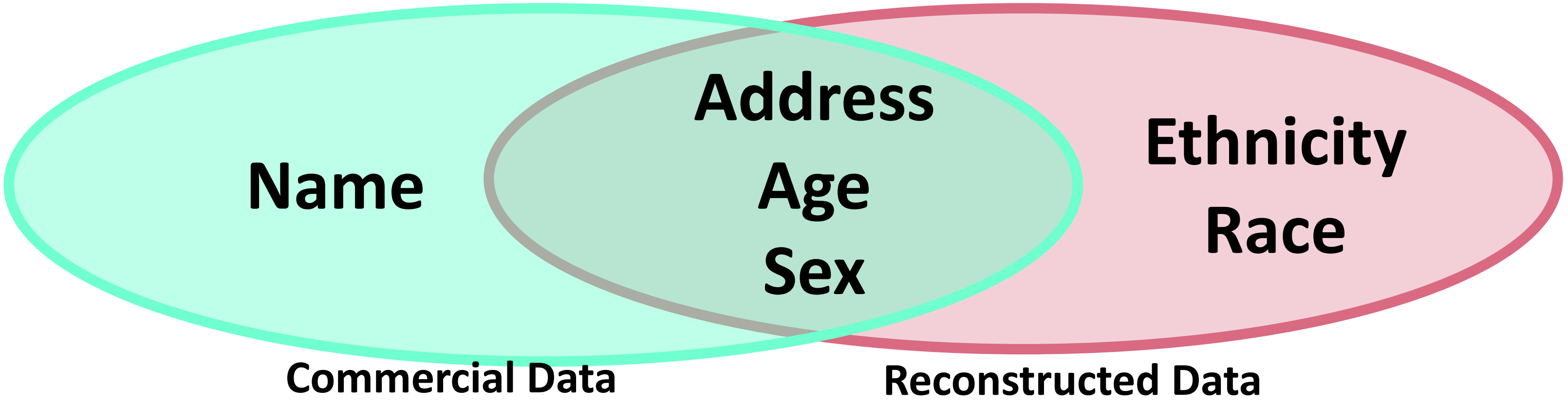
| | Count | Median | Mean |
|---|---|---|---|
| Total | 7 | 30 | 38 |
| # female | 4 | 30 | 33.5 |
| # male | 3 | 30 | 44 |
| # black | 4 | 51 | 48.5 |
| # white | 3 | 24 | 24 |
| # married | 4 | 51 | 54 |
| # black F | 3 | 36 | 36.7 |

This table can be expressed by 164 equations. Solving those equations takes 0.2 seconds on a 2013 MacBook Pro.

# We performed a database reconstruct and re-identification attack for all 308,745,538 people in the 2010 Census

Reconstructed all 308,745,538 microdata records

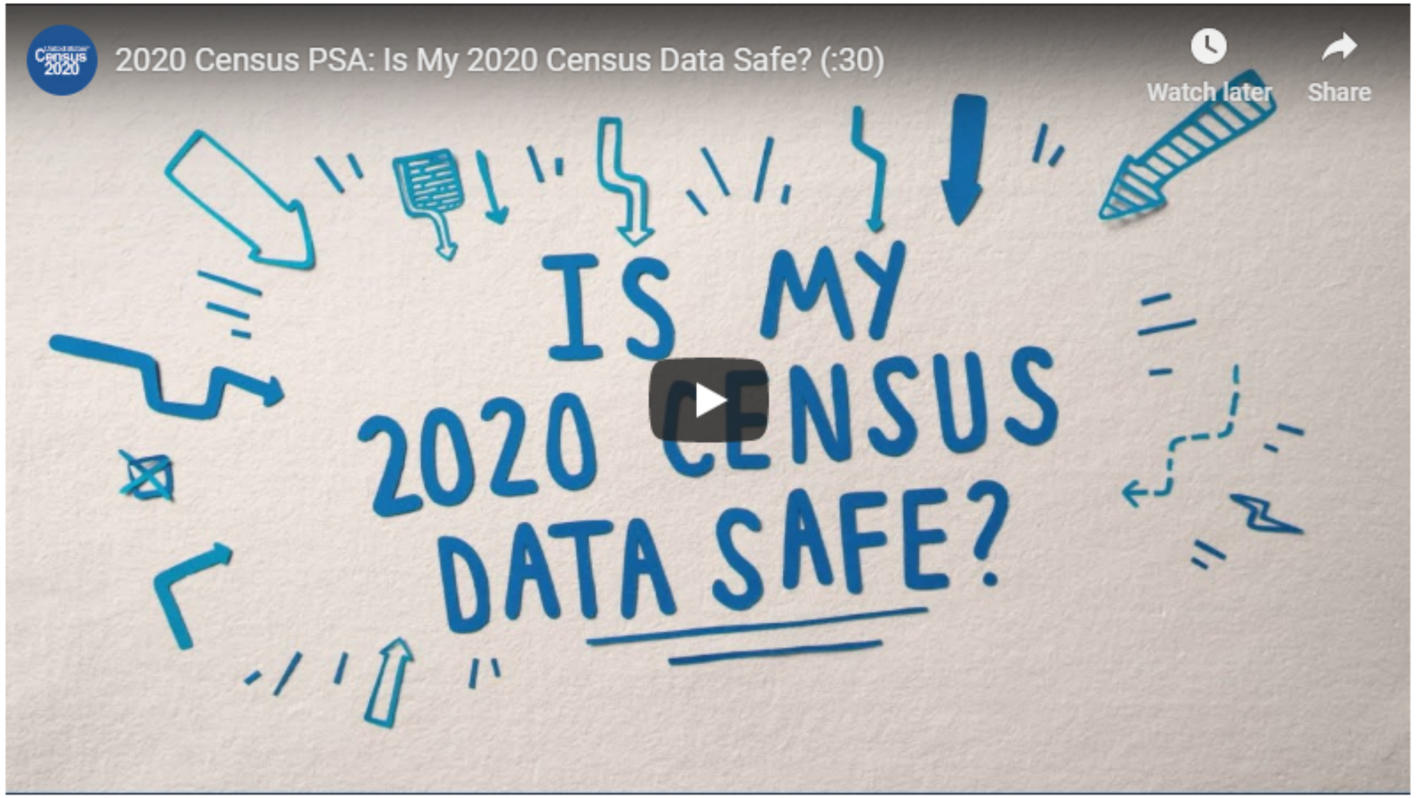
Used four commercial databases of the 2010 US population acquired 2009-2011 in support of the 2010 Census.




Link rate: 45%

Validated Re-identification Rate: 38% (17% of the US population)

# Privacy protection for the 2020 Census



# In 2017, the Census Bureau announced that it would use differential privacy for the 2020 Census.

**DP is a tool for controlling privacy-loss/accuracy trade-off**

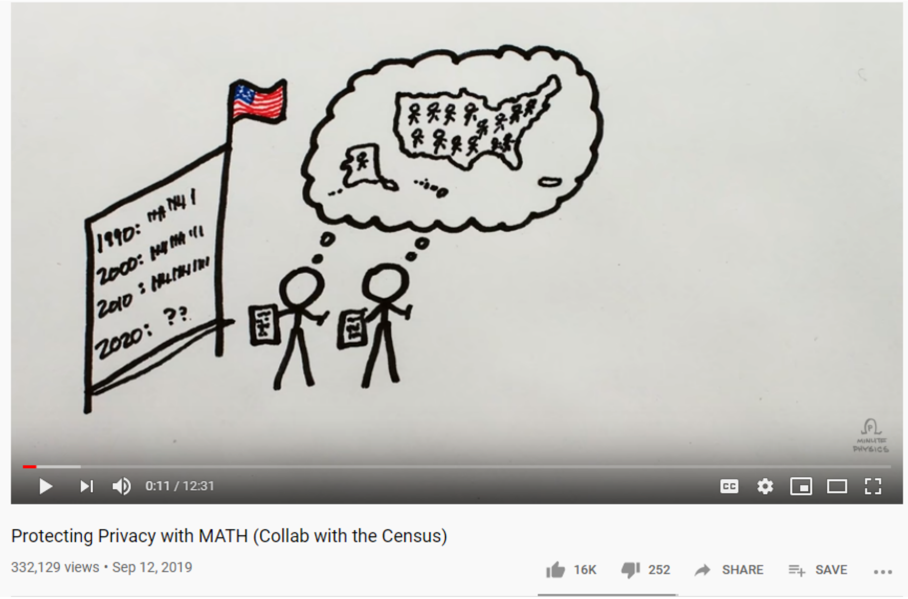
**DP lets us put the accuracy where it is needed.**

**DP privacy is "future-proof"**

**Records in the tabulation data will have no exact counterpart in the confidential data**

**Explicitly protected tabulations have provable, public accuracy levels**

- PL 94-171
- Demographic and Housing Characteristics (DHC)



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

## We had to create a new system that:

Produced higher-quality statistics at more densely populated geographies

Produced consistent tables

## We created a new differential privacy algorithm and system that:

Produces statistics from the top-down

- e.g. National Level -> State Level -> County Level -> Tract Level -> Block Level
- Creates protected microdata that can be used for any tabulation without additional privacy loss

Fits into the decennial census production system

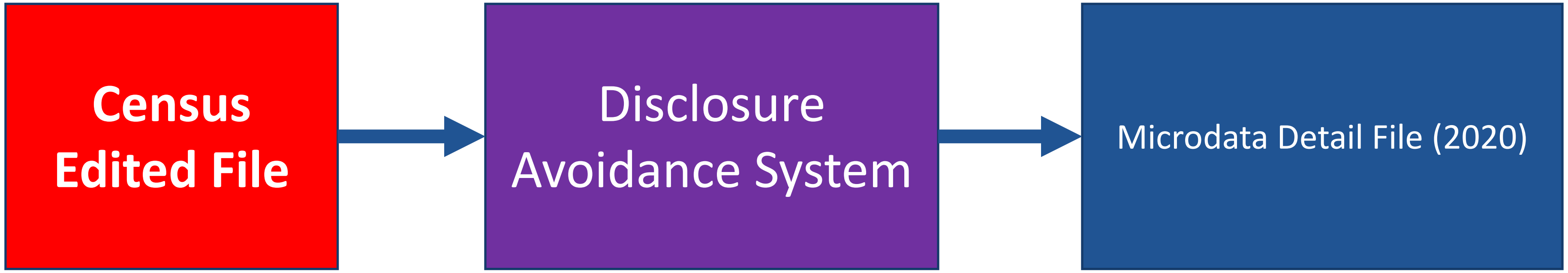
# We planned to create a "Disclosure Avoidance System" that dropped into the Census production system.

**Features of the DAS:**

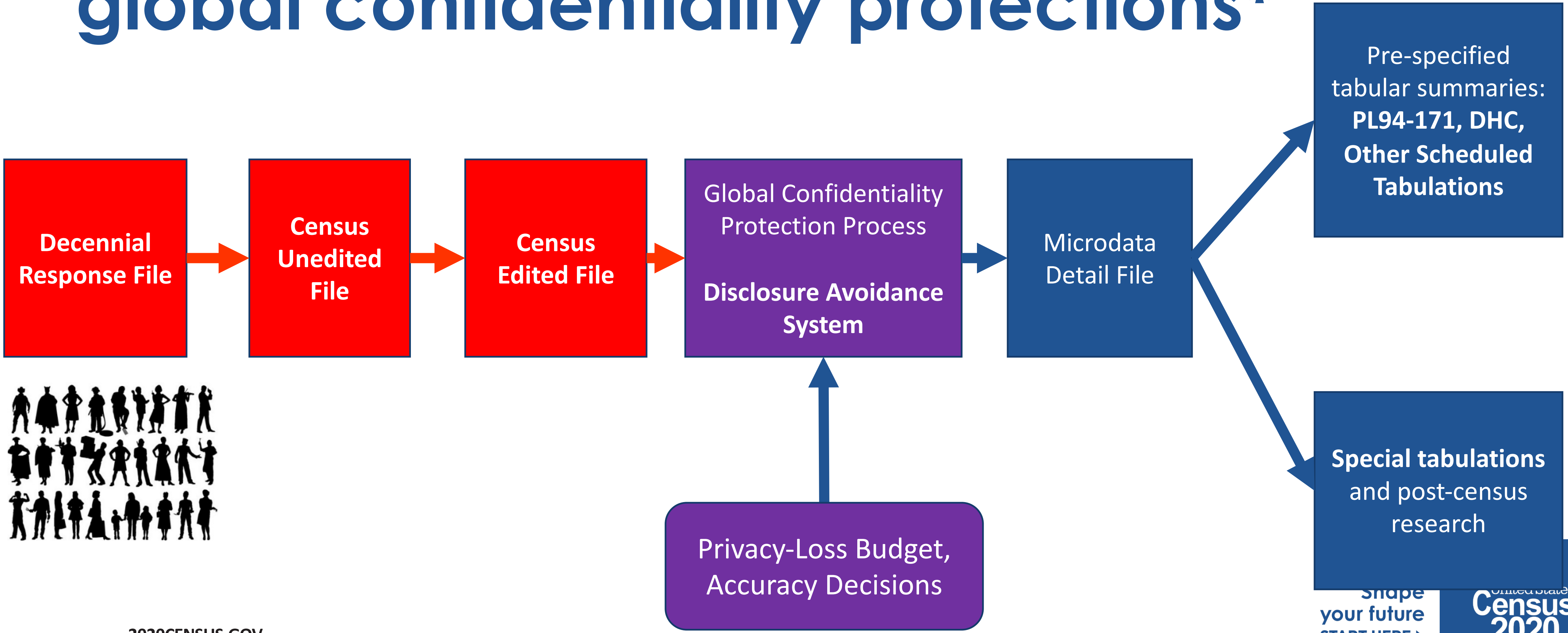
Operates on the edited Census records

Create microdata that would be "safe to tabulate."

Capture all necessary statistics in the microdata.



# The Disclosure Avoidance System allows the Census Bureau to enforce global confidentiality protections*




*Note: See later in this presentation for the design that includes "Group II" data products

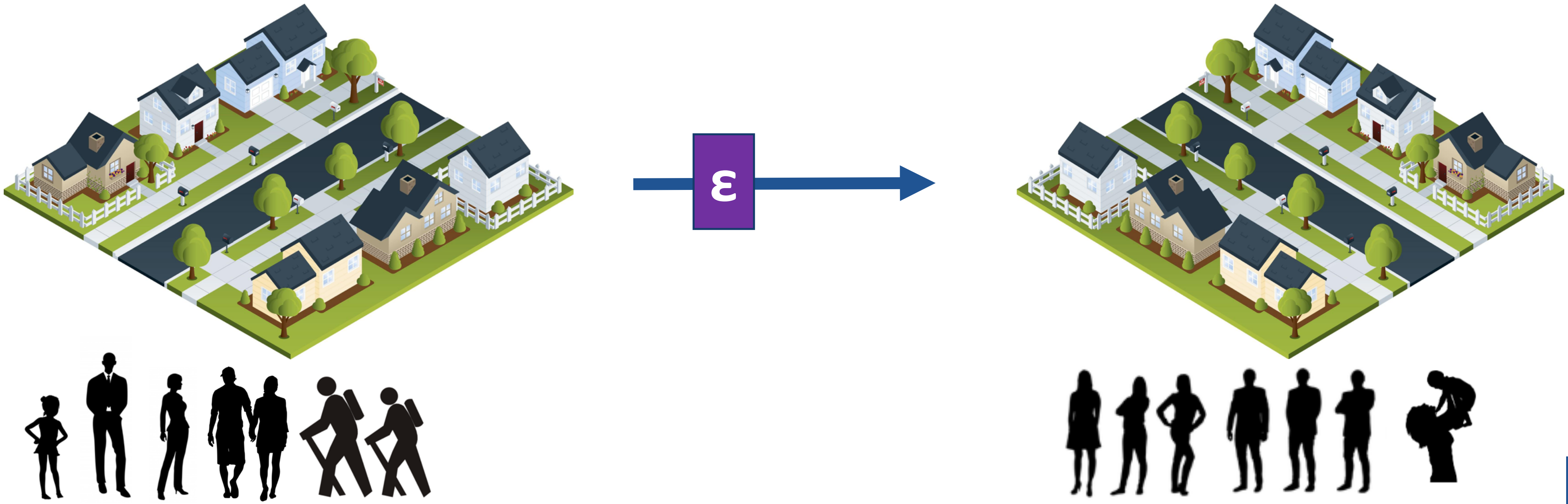
# Our DP mechanism protects histograms of person types.

## Census "block"

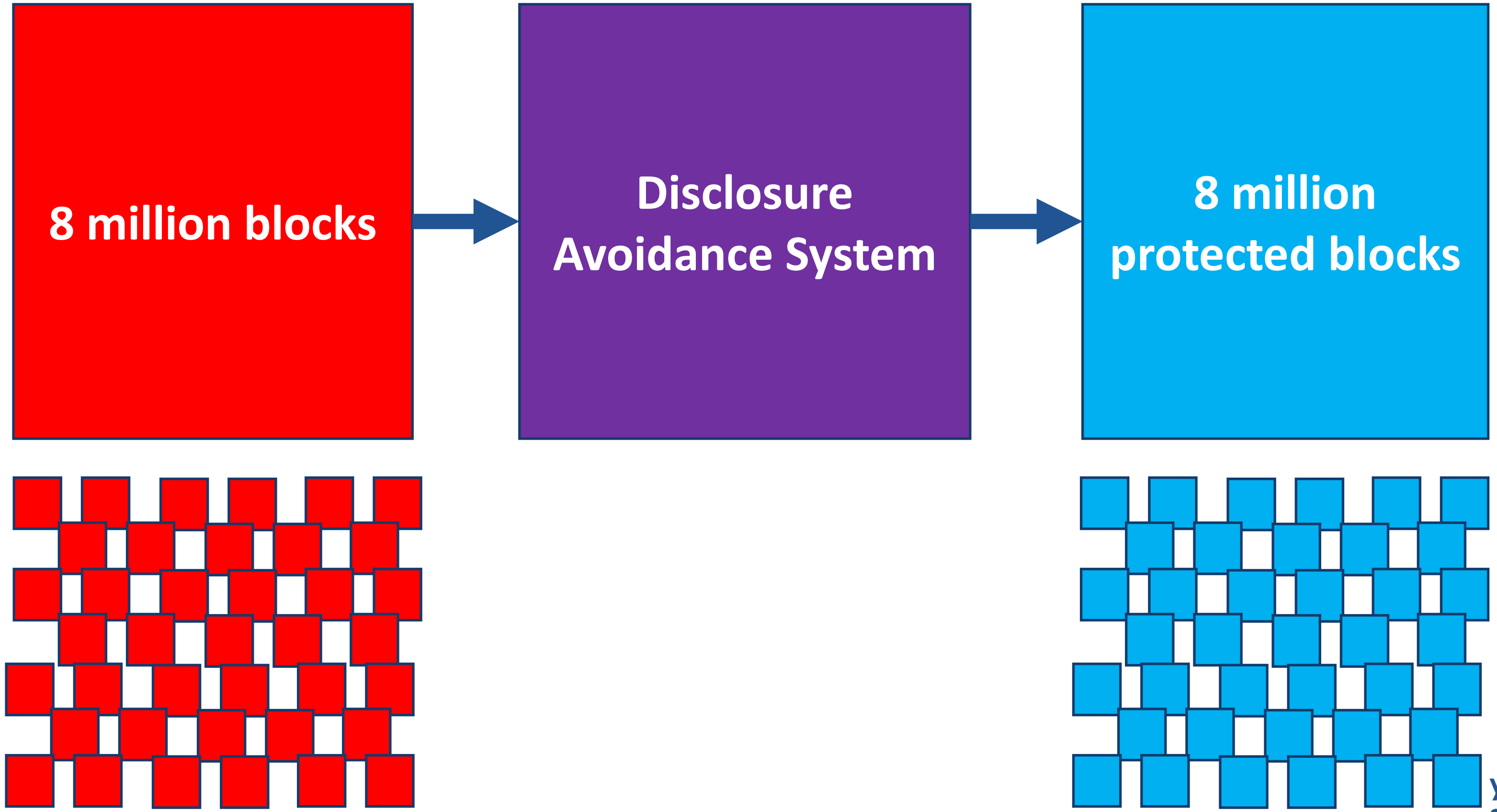
## Census "block" histogram

| Count | Age | Sex | Race | Ethnicity | REL |
|---|---|---|---|---|---|
| 1 | 8 | F | B | - | Child |
| 1 | 18 | M | W | H | Single |
| 1 | 24 | F | W | H | Single |
| 1 | 30 | M | W | - | HH |
| 1 | 36 | F | B | - | Spouse |
| 1 | 66 | F | B | - | HH |
| 1 | 84 | M | B | - | Spouse |

# First system applied DP to every block. This was the “block-by-block” system.



# There are roughly 8 million blocks



# We released public test results using data from the 1940 Census


# In 2018 we adopted the TopDown Algorithm (TDA)

**Computes and protects a histogram for various geographical units at various geographical levels**

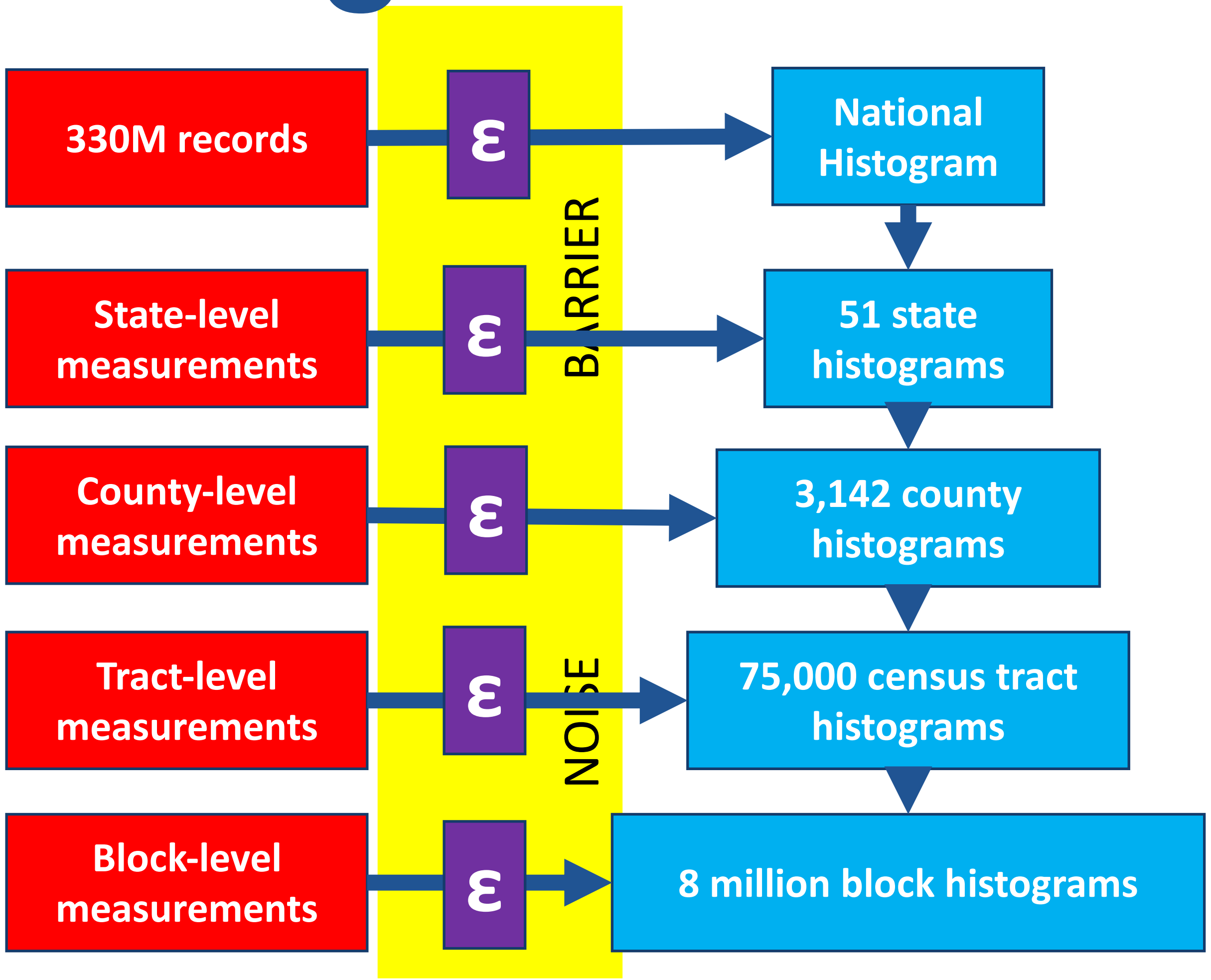
**Illustrated for the current specification of the Demographic and Housing Characteristics Person tables with proposed, approximate 2020 tabulation geography**

| Level | Count | Histogram size |
|---|---|---|
| National | 1 | 1.2M |
| State | 51 | 1.2M |
| County | 3142 | 1.2M |
| Census Tract | 75,000 | 1.2M |
| Block group | 275,000 | 1.2M |
| Blocks | 8M | 1.2M |
| | | |

Shape your future START HERE >

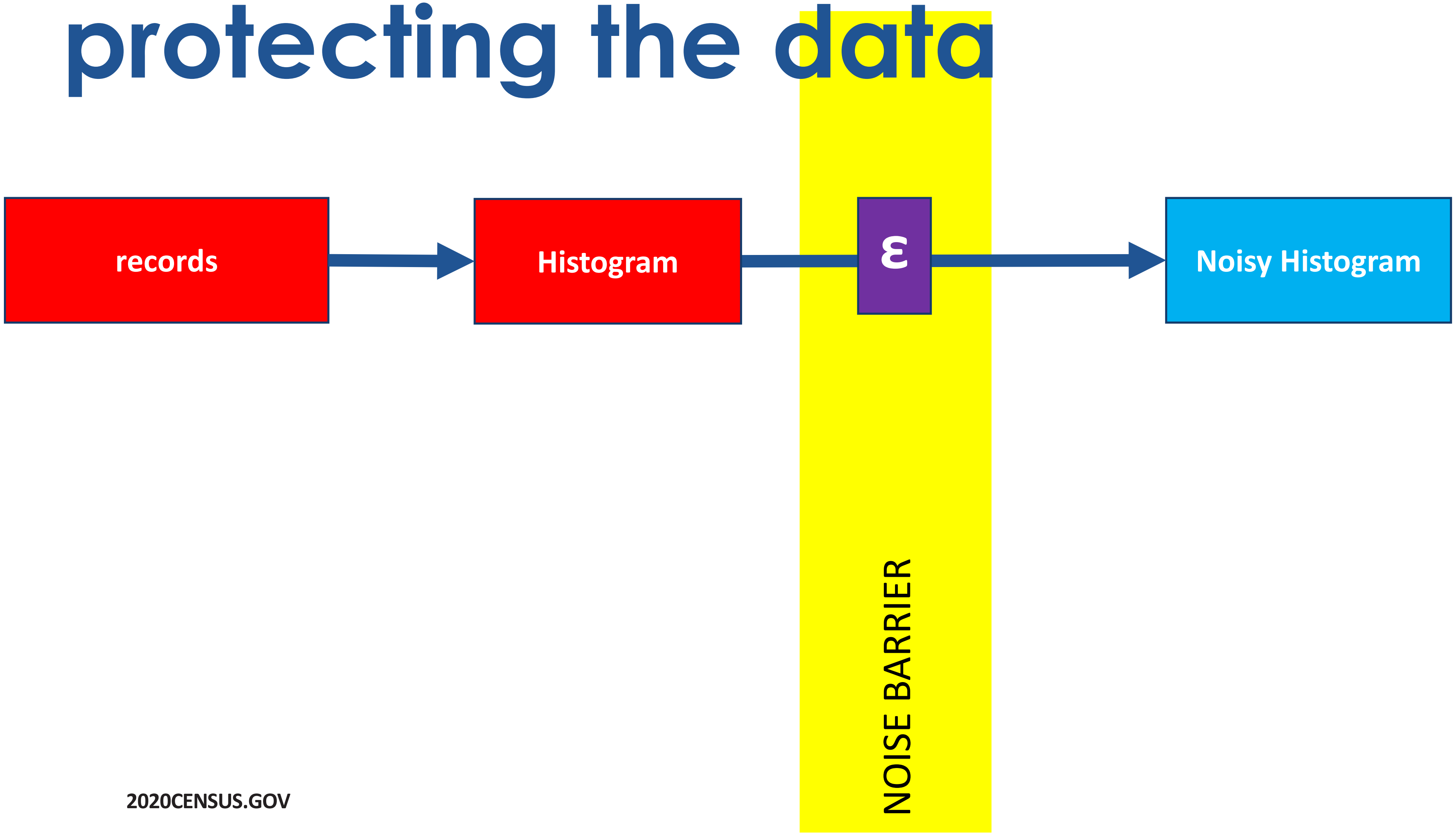
United States® Census 2020

# The TopDown algorithm*




*v1 geographies

# TDA part 1: protecting the data



# TDA part 2: post-processing

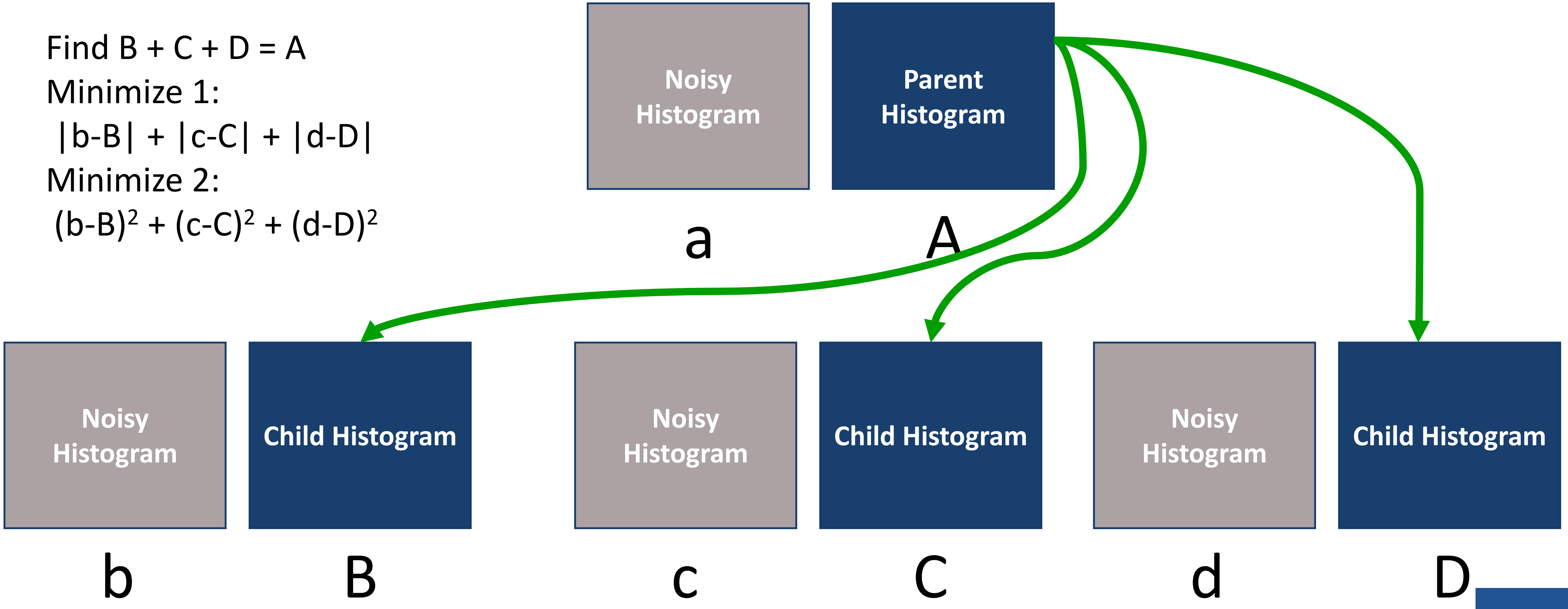
Find B + C + D = A

Minimize 1:

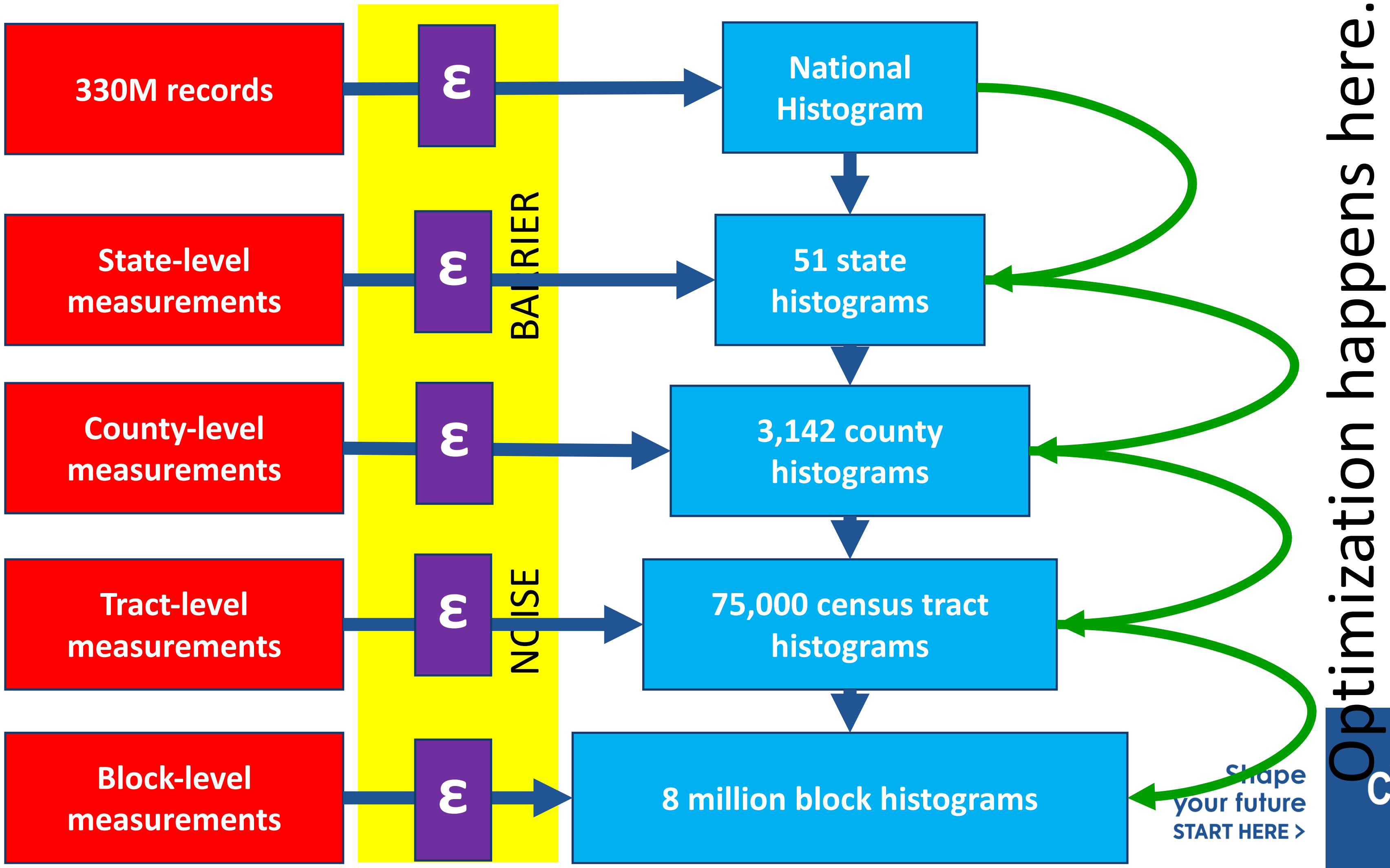
$|b-B| + |c-C| + |d-D|$

Minimize 2:

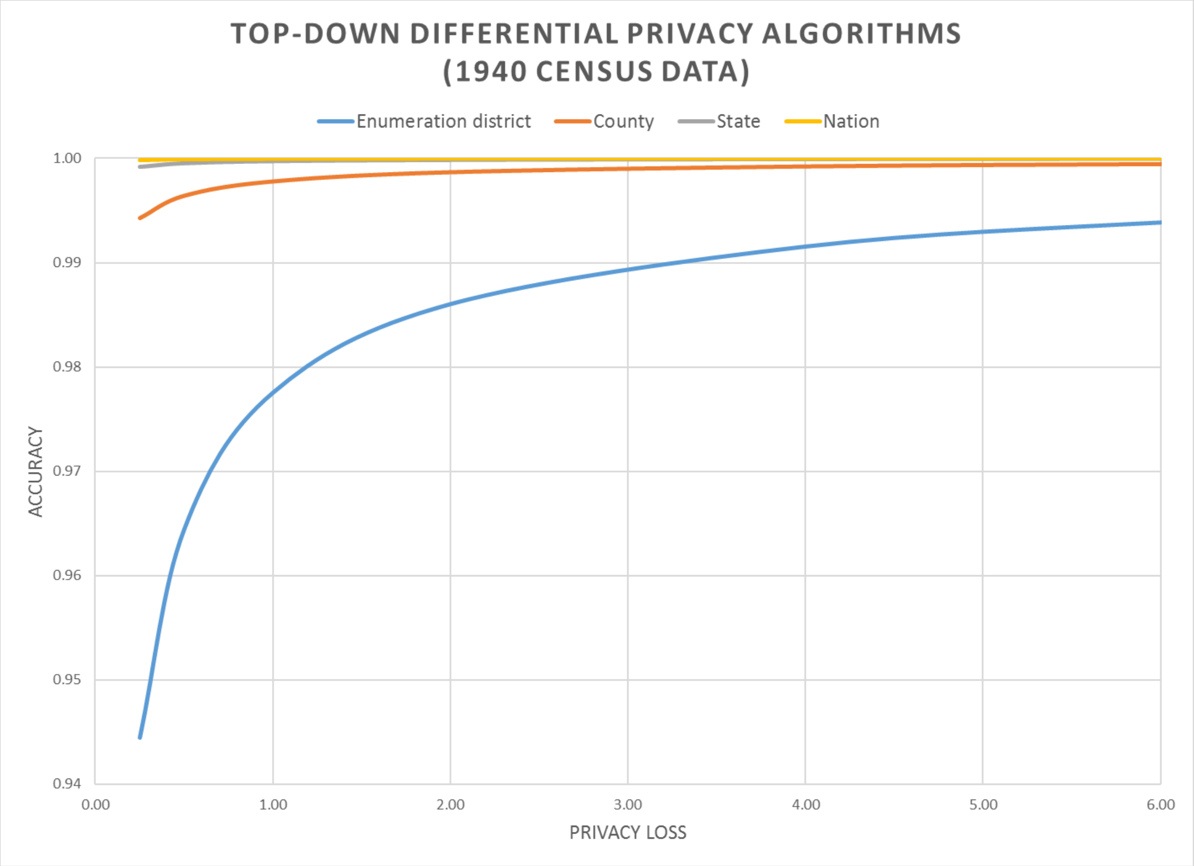
$(b-B)^2 + (c-C)^2 + (d-D)^2$



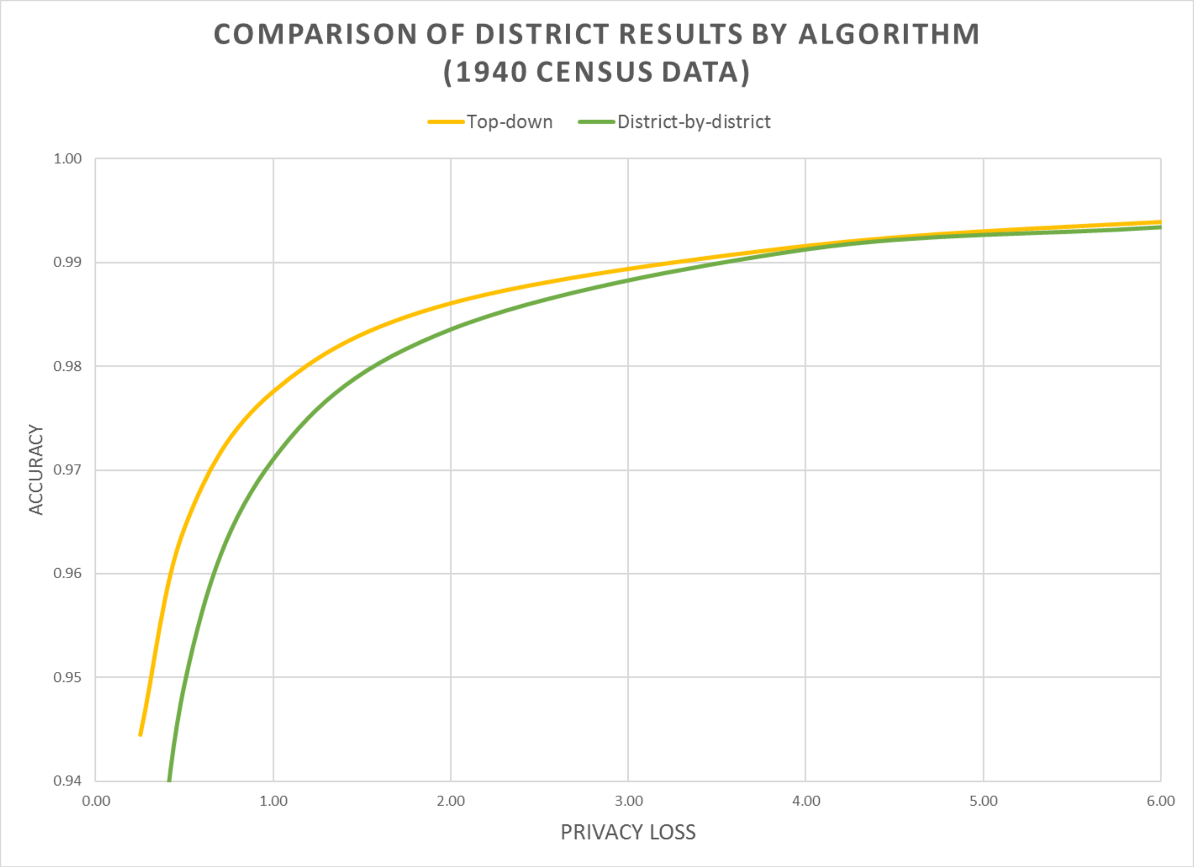
# TDA part 2: post-processing

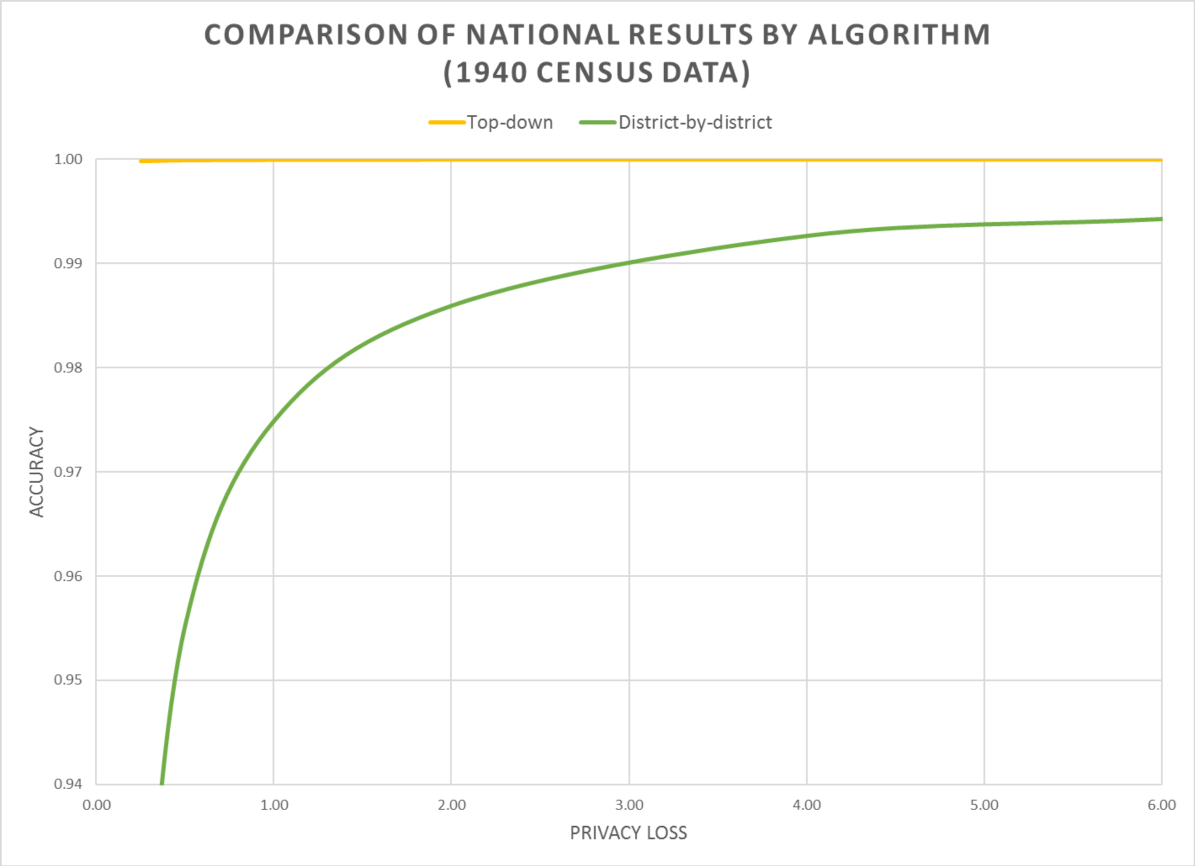


# TDA statistics for runs on the 1940 Census data.



# TDA produces better statistics at all geography levels!



# Challenges deploying differential privacy at the US Census Bureau

**Oh, no—not another learning experience!**

# Our experience with OnTheMap did not prepare the organization for the challenge.

**OnTheMap was a new product, designed from the start to be DP on the residential side.**

(Haney et al. (2017) extends to the employment side)

**The decennial Census of Population and Housing, first performed under the direction of Thomas Jefferson in 1790, is the oldest and most expensive statistical undertaking of the U.S. government.**

**Transitioning existing data products has revealed:**

The limits of today's formal privacy mechanisms

The difficulty of retrofitting legacy statistical products to conform with modern privacy practice

# Managing the Tradeoff

**We thought that one of our primary problems would be managing the privacy loss-accuracy tradeoff**

**The DP research community has not created any significant theory or tools in this area.**

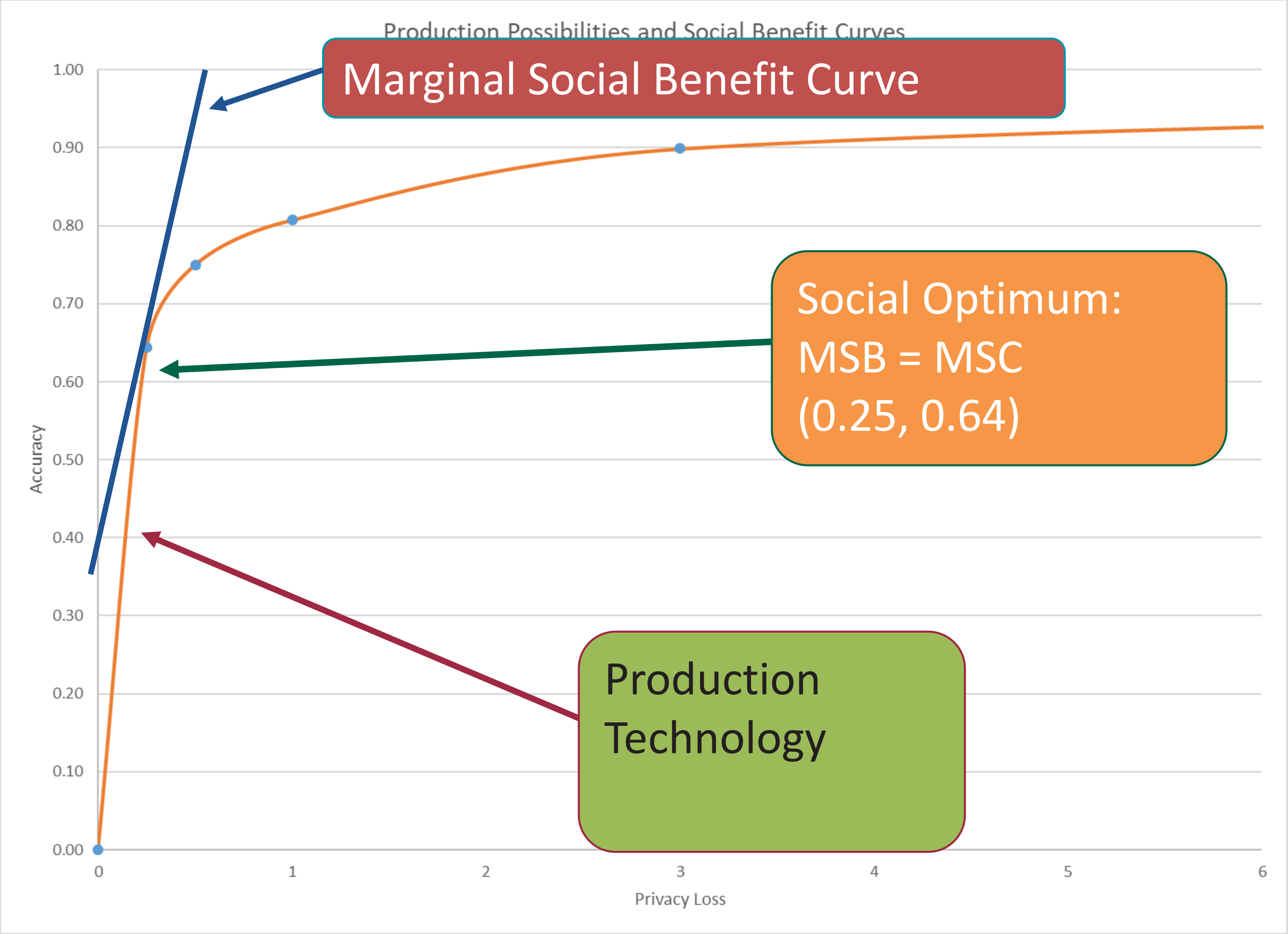
# Basic Principles

**Based on recent economics (2019, *American Economic Review*)**
**https://digitalcommons.ilr.cornell.edu/ldi/48/ or https://arxiv.org/abs/1808.06303**

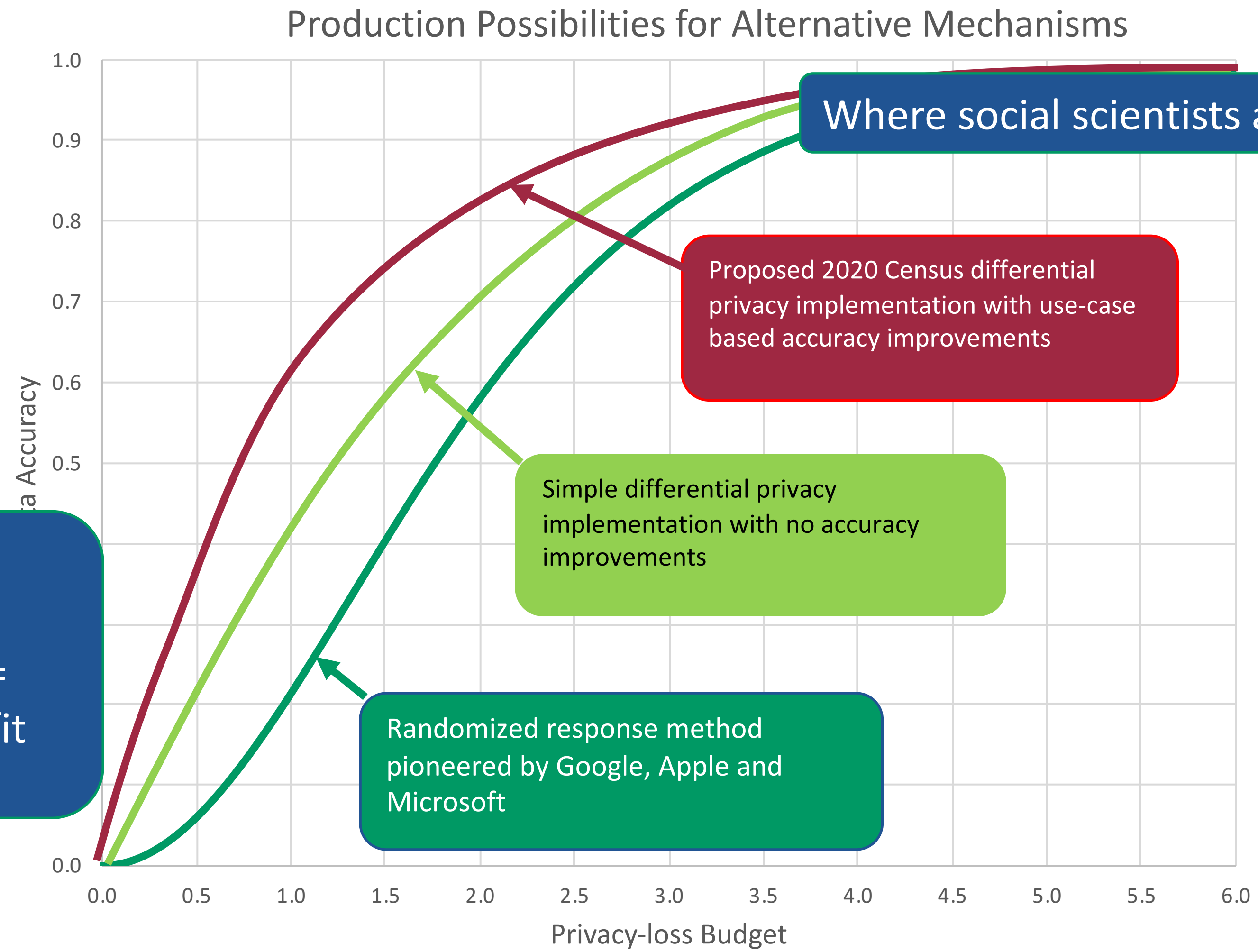
**The marginal social benefit is the sum of all persons' willingness-to-pay for data accuracy with increased privacy loss**

**The marginal rate of transformation is the slope of the privacy-loss v. accuracy graphs we have been examining**

**This is exactly the same problem being addressed by Google in RAPPOR or PROCHLO, Apple in iOS 11, and Microsoft in Windows 10 telemetry**


Production Possibilities and Social Benefit Curves
Marginal Social Benefit Curve
Social Optimum:
MSB = MSC
(0.25, 0.64)
Production
Technology
Accuracy
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0
1
2
3
4
5
6
Privacy Loss

# Scientific Issue: Setting Epsilon




Where social scientists act like MSC = MSB

Proposed 2020 Census differential privacy implementation with use-case based accuracy improvements

Simple differential privacy implementation with no accuracy improvements

Randomized response method pioneered by Google, Apple and Microsoft

Where computer scientists act like Marginal Social Cost = Marginal Social Benefit

# In theory, practice and theory are the same. In practice, they aren't.

# Initial Operational Issues

Obtaining Qualified Personnel and Tools

Recasting high-sensitivity queries

Identifying Structural Zeros

Obtaining a Suitable Computing Environment

Accounting for All Uses of Confidential Data

# Scientific Issues for the 2020 Census

**Hierarchical Mechanisms**

We needed a novel mechanism that:

Assured consistent statistics from US->States->Counties->Tracts

Provided lower error for larger geographies.

**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)

https://github.com/uscensusbureau/census2020-das-e2e/blob/master/etl_e2e/ipums_1940_validator.py

# Invariants in the 2020 Census

**Invariants in the 2020 will be decided by the Census Bureau's Data Stewardship Executive Policy Committee (DSEP).**

**DSEP replaced the policy of holding voting age population invariant because of grave concerns about its effects on the Census Bureau's ability to protect confidentiality, especially in block and block-group level tabulations.**

**The policy was replaced with explicit management of invariants and privacy-loss budgets beginning with the decision memos for Apportionment, the 2018 End-to-End Census Test data products, and the 2010 Demonstration Data Products.**

**Those memos are:**

https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series/2020-memo-2019_12.html

https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series/2020-memo-2019_13.html

https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series/2020-memo-2019_25.html

**DSEP has made no final decisions regarding invariants or the privacy-loss budget for the 2020 Census data publications.**

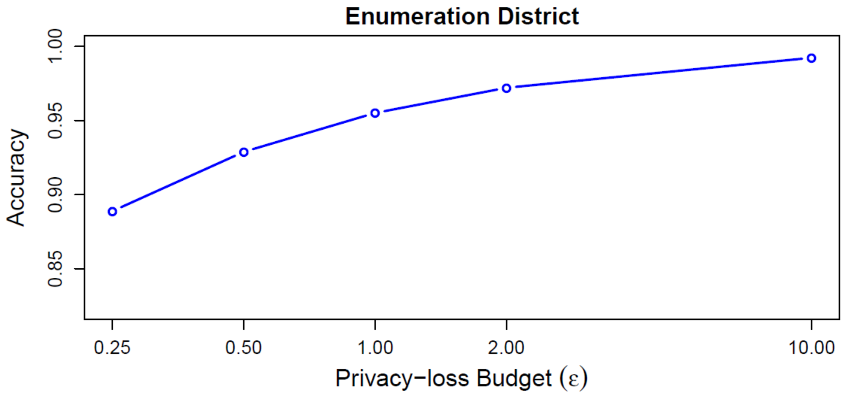
# Scientific Issues: Quality Metrics

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

## Options we considered:

L1 error between "true" data set and "privatized" data set

Impact on an algorithm that uses the data
(e.g. voting rights enforcement)




## Options others considered:

Enumeration District-by-Enumeration District differences of specific demographic subgroups


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

# Many in the data user community are apprehensive about DP.

**Steven Ruggles**
@HistDem
Following

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

2 9 32

**Steven Ruggles**
@HistDem
Following

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

1 2 13

# 2010 Demonstration Data Products

### October 29, 2019 Data and Software Release

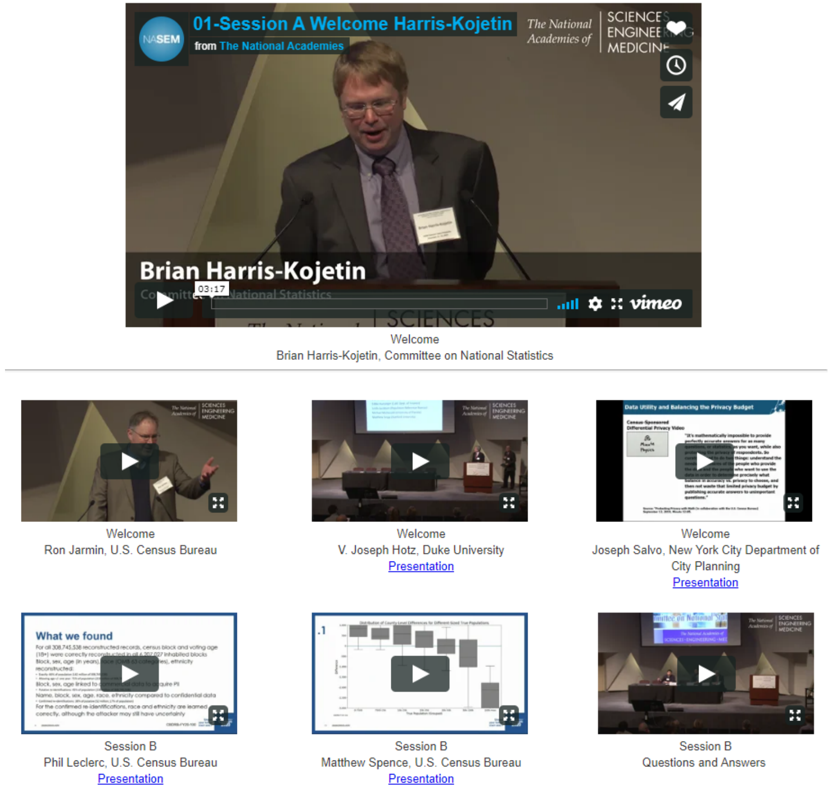
2010 Census Edited File as processed by the 2020 DAS (as of October 2019)

Release of the 2020 DAS Source Code (as of October 2019)

https://github.com/uscensusbureau/census2020-das-2010ddp/

### December 11-12 Workshop
### National Academies of Science

https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518




Shape your future START HERE >

# “2010 Demonstration Data Products” were not well received.


https://www.nytimes.com/interactive/2020/02/06/opinion/census-algorithm-privacy.html

# Issues Faced by Data Users

### Access to Micro-data

Many users expect access to microdata.

### Difficulties Arising from Increased Transparency

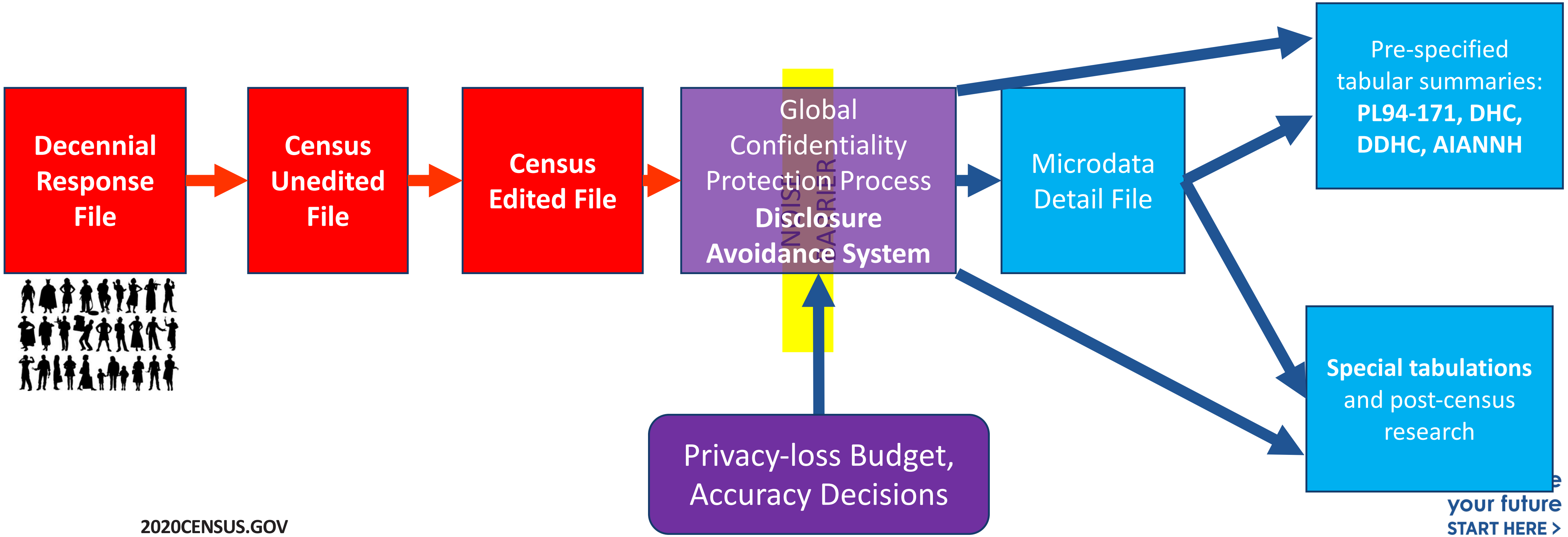
Many users were not aware of prior disclosure avoidance practices.

The swap rate was never made public.

### Misunderstandings about Randomness and Noise Infusion

# By summer 2019, it was clear that we would not realize our microdata goal

**MDF will not include Household-Person Joins and Detailed Race/Ethnicity/ Tribe**



# 2020 DAS: Data Products

| Group I | Group II | Group III |
| --- | --- | --- |
| • PL94-171<br>• Demographic Profiles<br>• DHC-Persons<br>• DHC-Households<br>• CVAP (special tabulation)* | • AIAN<br>• Detailed Race<br>• Person & Household joins<br>• Averages | • PUMS<br>• Special Tabulations |
| ***Supported by MDF*** | ***Generated directly by 3rd Party DB*** | ***Under discussion*** |

*Citizen Voting-Age Population by Race and Ethnicity block-level special tabulation will be produced by March 31, 2021. See: https://www.census.gov/programs-surveys/decennial-census/about/voting-rights/cvap.html, and technical documentation posted therein.

# Recommendations

Keep decision makers “in-the-loop.”

Adopt a controlled vocabulary.

Have an integrated communications strategy.

Minimize confusion from source code releases and make code runnable by outsiders.

Involve data users early and often.

# For more information…

THE WALL STREET JOURNAL.

English Edition ▾ | December 6, 2019 | Print Edition | Video

Politics Economy Business Tech Markets Opinion Life & Arts Real Estat

**Census Overhaul Seeks to Avoid Outing Individual Respondent Data**

***Most Census 2020 results will be adjusted; measures would prevent targeting based on citizenship***

By Paul Overberg

Nov. 10, 2019 7:00 am ET

practice

DOI:10.1145/3287287

Article development led by acmqueue
queue.acm.org

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

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

## Understanding Database Reconstruction Attacks on Public Data

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

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

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

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

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

46 COMMUNICATIONS OF THE ACM | MARCH 2019 | VOL. 62 | NO. 3

Communications of ACM March 2019
Garfinkel & Abowd

Can a set of equations keep U.S. census data private?

**By Jeffrey Mervis**

**Science**

**Jan. 4, 2019 , 2:50 PM**

http://bit.ly/Science2019C1

Shape your future
START HERE >

United States® Census 2020

# The Most Technical 2020 Publications

**Most recent public source code:**

https://github.com/uscensusbureau/census2020-das-2010ddp

**System Design Specification**

https://github.com/uscensusbureau/census2020-das-2010ddp/blob/master/doc/2010-Demonstration-Data-Products-Disclosure-Avoidance-System-Design-Specification%20FINAL.pdf

**Scientific paper describing mechanism:**

https://github.com/uscensusbureau/census2020-das-2010ddp/blob/master/doc/20191020_1843_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf

# More Background on the 2020 Census Disclosure Avoidance System

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

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

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

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

December 6, 2018 CSAC (detailed discussion of algorithms and choices) https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf?#

April 15, 2019 Code base and documentation for the 2018 End-to-End Census Test (E2E) version of the 2020 Disclosure Avoidance System https://github.com/uscensusbureau/census2020-das-e2e

June 6, 2019 Blog explaining how to use the code base with the 1940 Census public data from IPUMS https://www.census.gov/newsroom/blogs/research-matters/2019/06/disclosure_avoidance.html

June 11, 2019 Keynote address "The U.S. Census Bureau Tries to Be a Good Data Steward for the 21st Century" ICML 2019 abstract, video

June 29-31, 2019 Joint Statistical Meetings Census Bureau electronic press kit
(See talks by Abowd, Ashmead, Garfinkel, Leclerc, Sexton, and others)