2020 Disclosure Avoidance System (DAS)

Simson L. Garfinkel
Senior Computer Scientist for Confidentiality and Data Access

John M. Abowd
Chief Scientist
Associate Director for Research and Methodology (ADRM)

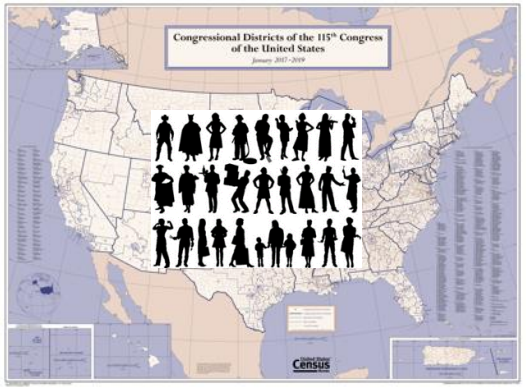
Disclosure Avoidance System

Purpose

- The Disclosure Avoidance System (DAS) assures that the 2020 Census data products meet the legal requirements of Title 13, Section 9 of the U.S. Code.

- The DAS is designed to prevent **improper disclosures of data about individuals and establishments** in the 2020 census data products.

- Stakeholders: All users of data from the 2020 Census.



Disclosure Avoidance System

Agenda

- Project purpose — Why do we need a new DAS?

- Noise injection and differential privacy — A brief tutorial

- State of the project

- Looking forward and conclusion

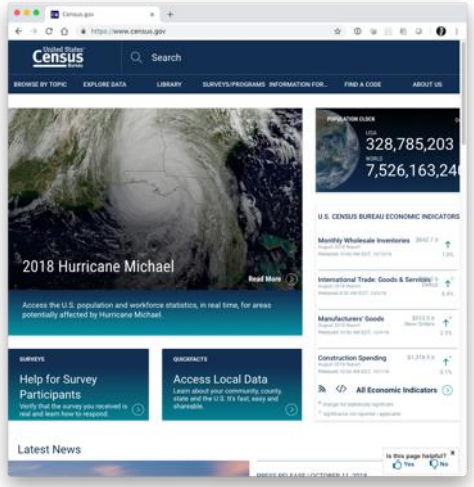
United States Census 2020

Project purpose:

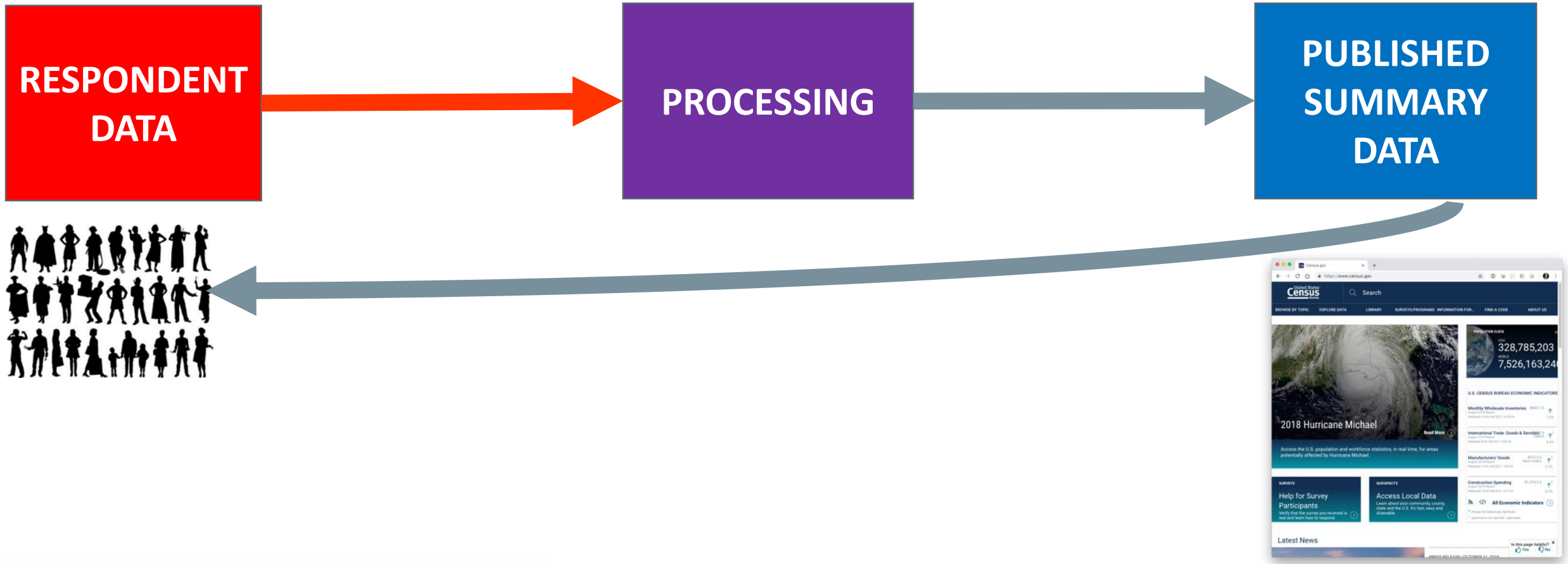
Why we need a new disclosure avoidance system

We create statistics by collecting data, processing and publishing

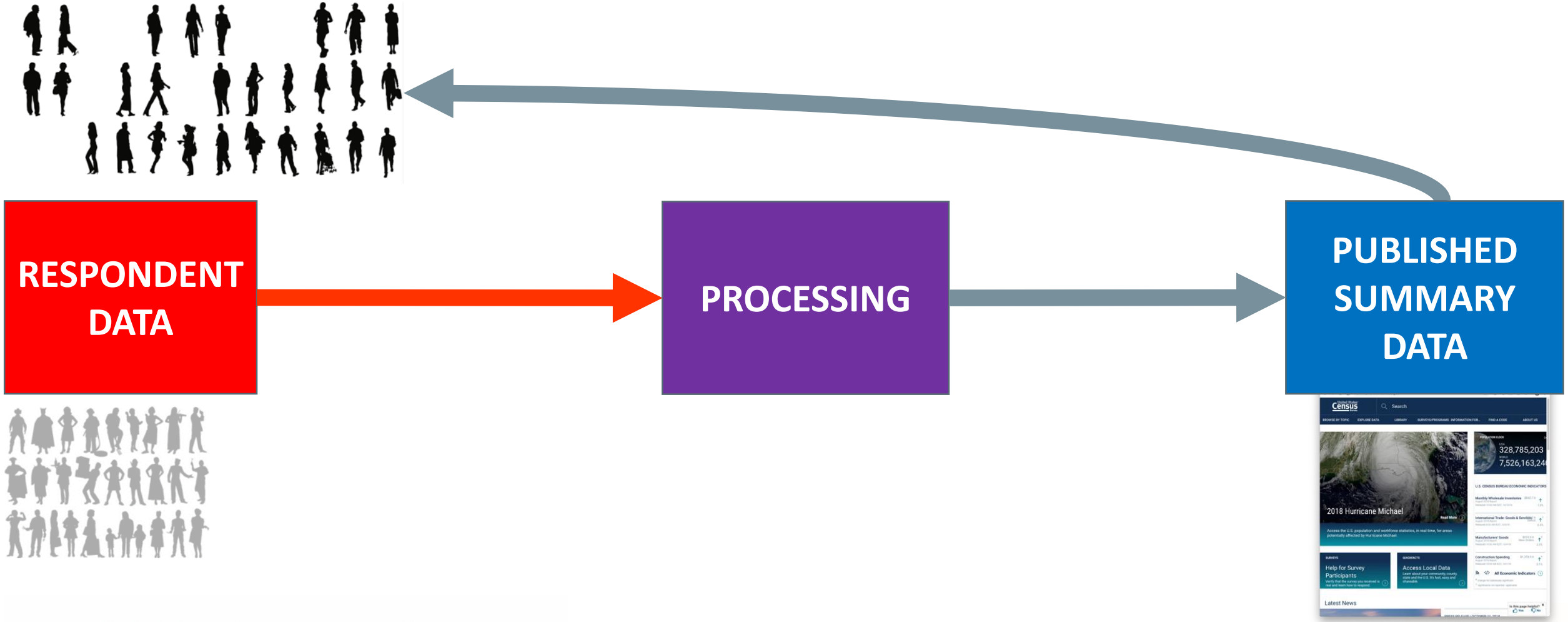
RESPONDENT DATA → **PROCESSING** → **PUBLISHED SUMMARY DATA**



Database reconstruction is a mathematical process that reverses this process.



Database reconstruction is a mathematical process that reverses this process.

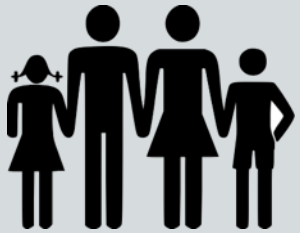
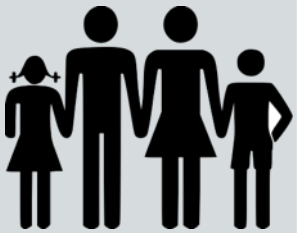
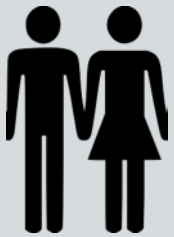


Consider a census block:

PUBLISHED DATA

	Counts
Age < 18	4
Age >= 18	6
Race 1	4
Race 2	4
Race 3	2

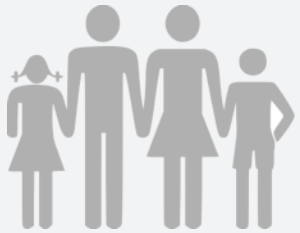
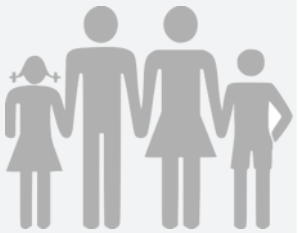
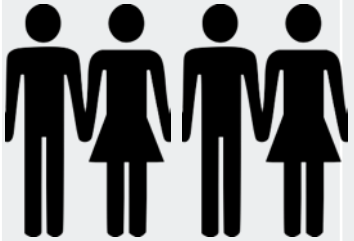
RECONSTRUCTED DATA

	Race 1	Race 2	Race 3
R1			

PUBLISHED DATA

	Counts
Age < 18	4
Age >= 18	6
Race 1	4
Race 2	4
Race 3	2

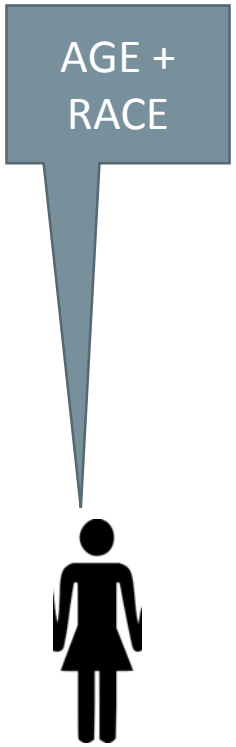
RECONSTRUCTED DATA

	Race 1	Race 2	Race 3
R1			
R2			

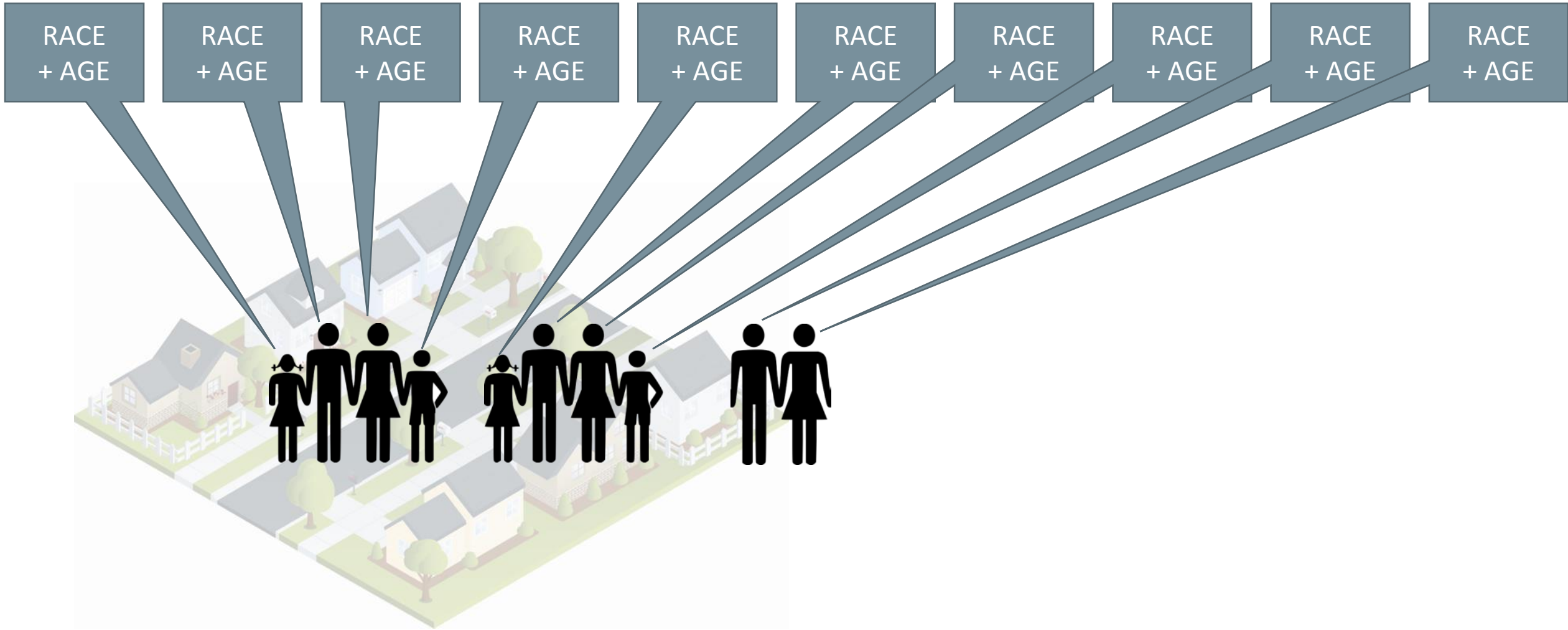
PUBLISHED DATA

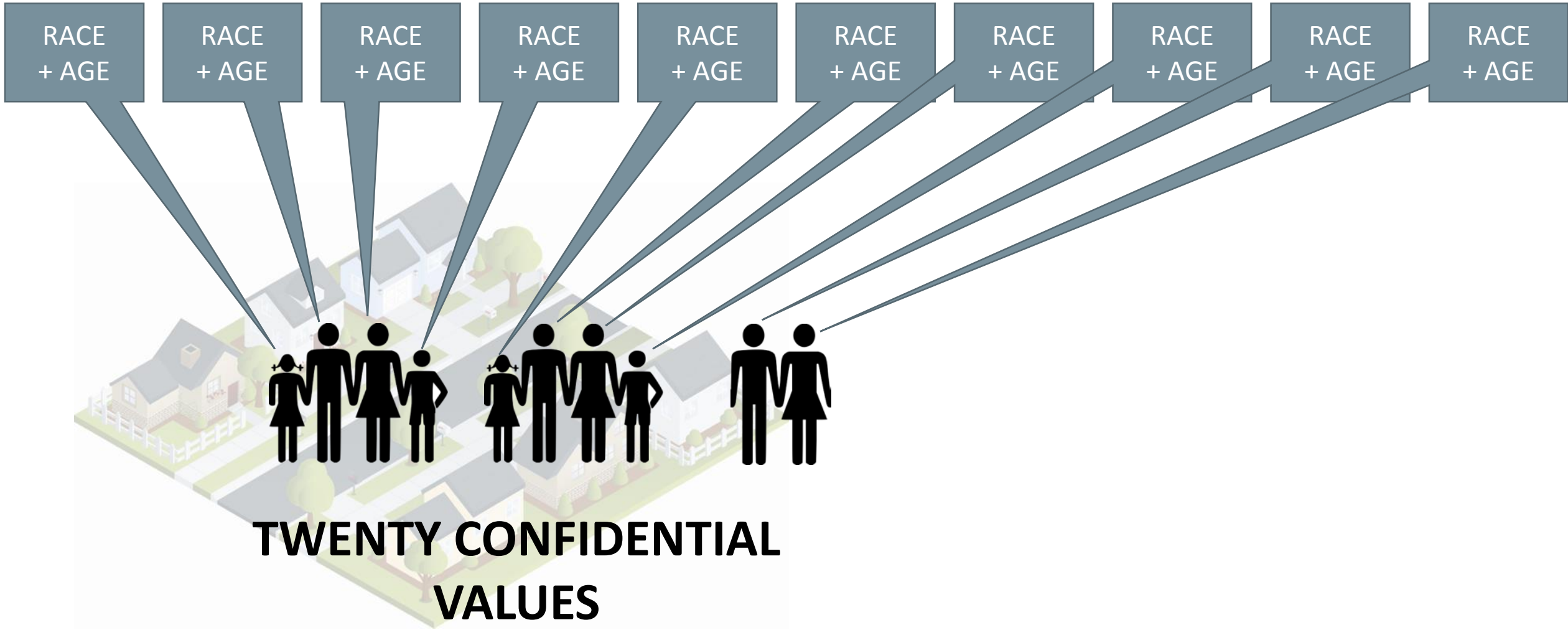
	Counts
Age < 18	4
Age >= 18	6
Race 1	4
Race 2	4
Race 3	2

	Race 1	Race 2	Race 3



AGE >=18


RACE + AGE
RACE + AGE
RACE + AGE
RACE + AGE
RACE + AGE
RACE + AGE
RACE + AGE
RACE + AGE
RACE + AGE
RACE + AGE


RACE + AGE
RACE + AGE
RACE + AGE
RACE + AGE
RACE + AGE
RACE + AGE
RACE + AGE
RACE + AGE
RACE + AGE
RACE + AGE
TWENTY CONFIDENTIAL VALUES

TWENTY CONFIDENTIAL VALUES

PUBLISHED DATA

	Counts
Age < 18	4
Age >= 18	6
Race 1	4
Race 2	4
Race 3	2

FIVE PUBLISHED STATISTICS

IT'S IN **OUR HANDS**

United States®
Census
2010

“This is the official form for all the people at this address.”

United States Census 2010

This is the official form for all the people at this address.
It is quick and easy, and your answers are protected by law.

U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU

Use a blue or black pen.

Start here

The Census must count every person living in the United States on April 1, 2010.

Before you answer Question 1, count the people living in this house, apartment, or mobile home using our guidelines.

- Count all people, including babies, who live and sleep here most of the time.

The Census Bureau also conducts counts in institutions and other places, so:

- Do not count anyone living away either at college or in the Armed Forces.
- Do not count anyone in a nursing home, jail, prison, detention facility, etc., on April 1, 2010.
- Leave these people off your form, even if they will return to live here after they leave college, the nursing home, the military, jail, etc. Otherwise, they may be counted twice.

The Census must also include people without a permanent place to stay, so:

- If someone who has no permanent place to stay is staying here on April 1, 2010, count that person. Otherwise, he or she may be missed in the census.

1. How many people were living or staying in this house, apartment, or mobile home on April 1, 2010?

Number of people =

2. Were there any additional people staying here April 1, 2010 that you did not include in Question 1? *Mark ☒ all that apply.*

- ☐ Children, such as newborn babies or foster children
- ☐ Relatives, such as adult children, cousins, or in-laws
- ☐ Nonrelatives, such as roommates or live-in baby sitters
- ☐ People staying here temporarily
- ☐ No additional people

3. Is this house, apartment, or mobile home — *Mark ☒ ONE box.*

- ☐ Owned by you or someone in this household with a mortgage or loan? *Include home equity loans.*
- ☐ Owned by you or someone in this household free and clear (without a mortgage or loan)?
- ☐ Rented?
- ☐ Occupied without payment of rent?

4. What is your telephone number? *We may call if we don't understand an answer.*

Area Code + Number

OMB No. 0607-0919-C: Approval Expires 12/31/2011.

Form **D-61** (9-25-2008)

5. Please provide information for each person living here. Start with a person living here who owns or rents this house, apartment, or mobile home. If the owner or renter lives somewhere else, start with any adult living here. This will be Person 1.

What is Person 1's name? *Print name below.*

Last Name

First Name MI

6. What is Person 1's sex? *Mark ☒ ONE box.*

☐ Male ☐ Female

7. What is Person 1's age and what is Person 1's date of birth?
Please report babies as age 0 when the child is less than 1 year old.
Print numbers in boxes.

Age on April 1, 2010 | Month | Day | Year of birth

➜ **NOTE: Please answer BOTH Question 8 about Hispanic origin and Question 9 about race. For this census, Hispanic origins are not races.**

8. Is Person 1 of Hispanic, Latino, or Spanish origin?

- ☐ No, not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin — *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

9. What is Person 1's race? *Mark ☒ one or more boxes.*

- ☐ White
- ☐ Black, African Am., or Negro
- ☐ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*
- ☐ Asian Indian
- ☐ Japanese
- ☐ Native Hawaiian
- ☐ Chinese
- ☐ Korean
- ☐ Guamanian or Chamorro
- ☐ Filipino
- ☐ Vietnamese
- ☐ Samoan
- ☐ Other Asian — *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*
- ☐ Other Pacific Islander — *Print race, for example, Fijian, Tongan, and so on.*
- ☐ Some other race — *Print race.*

10. Does Person 1 sometimes live or stay somewhere else?

☐ No ☐ Yes — *Mark ☒ all that apply.*

- ☐ In college housing
- ☐ In the military
- ☐ At a seasonal or second residence
- ☐ For child custody
- ☐ In jail or prison
- ☐ In a nursing home
- ☐ For another reason

➜ **If more people were counted in Question 1, continue with Person 2.**

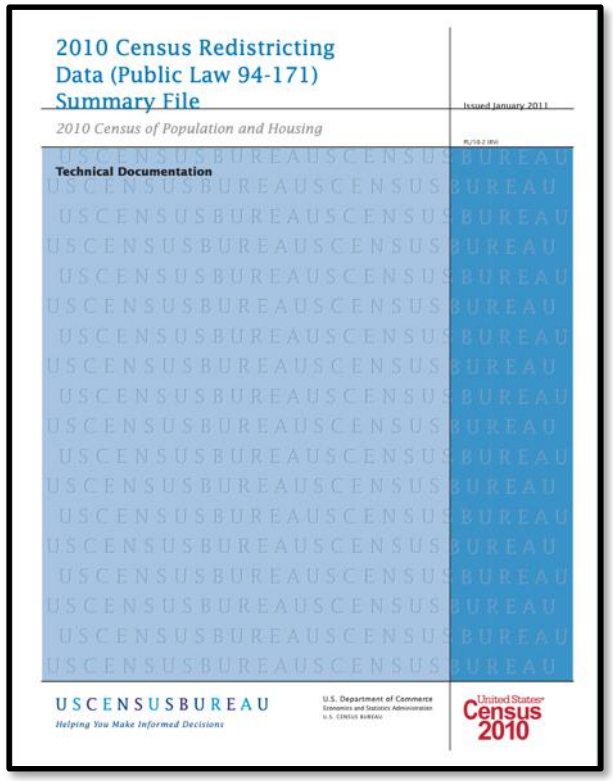
INFORMATIONAL COPY

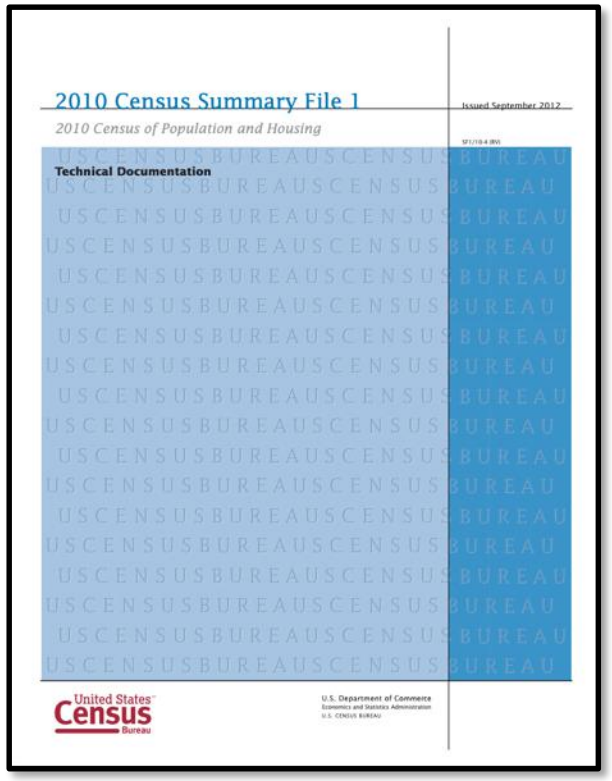
“It is quick and easy, and your answers are protected by law.”

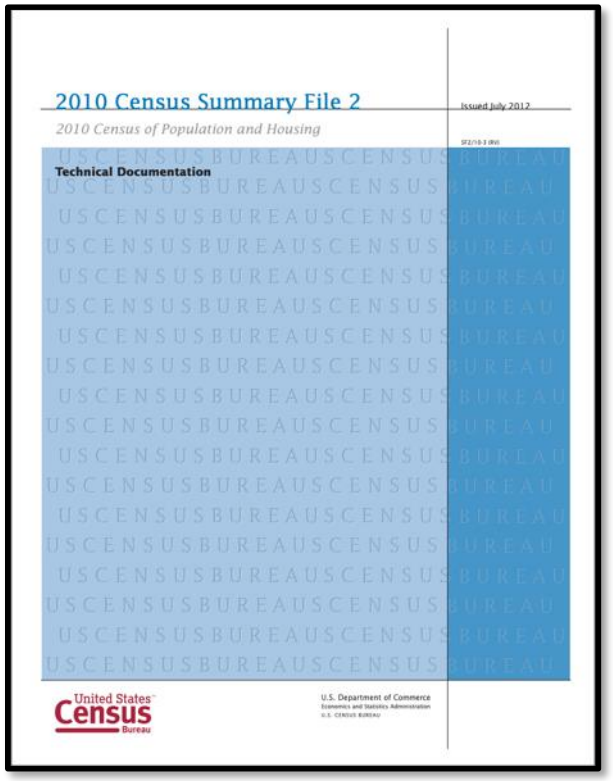
2010 Census of Population and Housing

Total population	**308,745,538**
Pieces of information per person:	6
Total pieces of information:	1,852,473,228

2010 Census Publication Schedule







PL94-171 Redistricting	Balance of Summary File 1	Summary File 2
2,771,998,263	2,806,899,669	2,093,683,376

2010 Census: Summary of Publications
(approximate counts)

Publication	Released counts (including zeros)
PL94-171 Redistricting	2,771,998,263
Balance of Summary File 1	2,806,899,669
Summary File 2	2,093,683,376
Public-use micro data sample	30,874,554
Lower bound on published statistics	7,703,455,862
Statistics/person	25

The threat of database reconstruction

2010 Census Statistics/person collected:	**6**
2010 Census Statistics/person published:	25
Lower bound on collected statistics: (308,745,538 x 6)	1,852,473,228
Lower bound on published statistics (25 statistics per person)	7,703,455,862

Two privacy mechanisms for the 2010 Census

Aggregation

Two privacy mechanisms for the 2010 Census

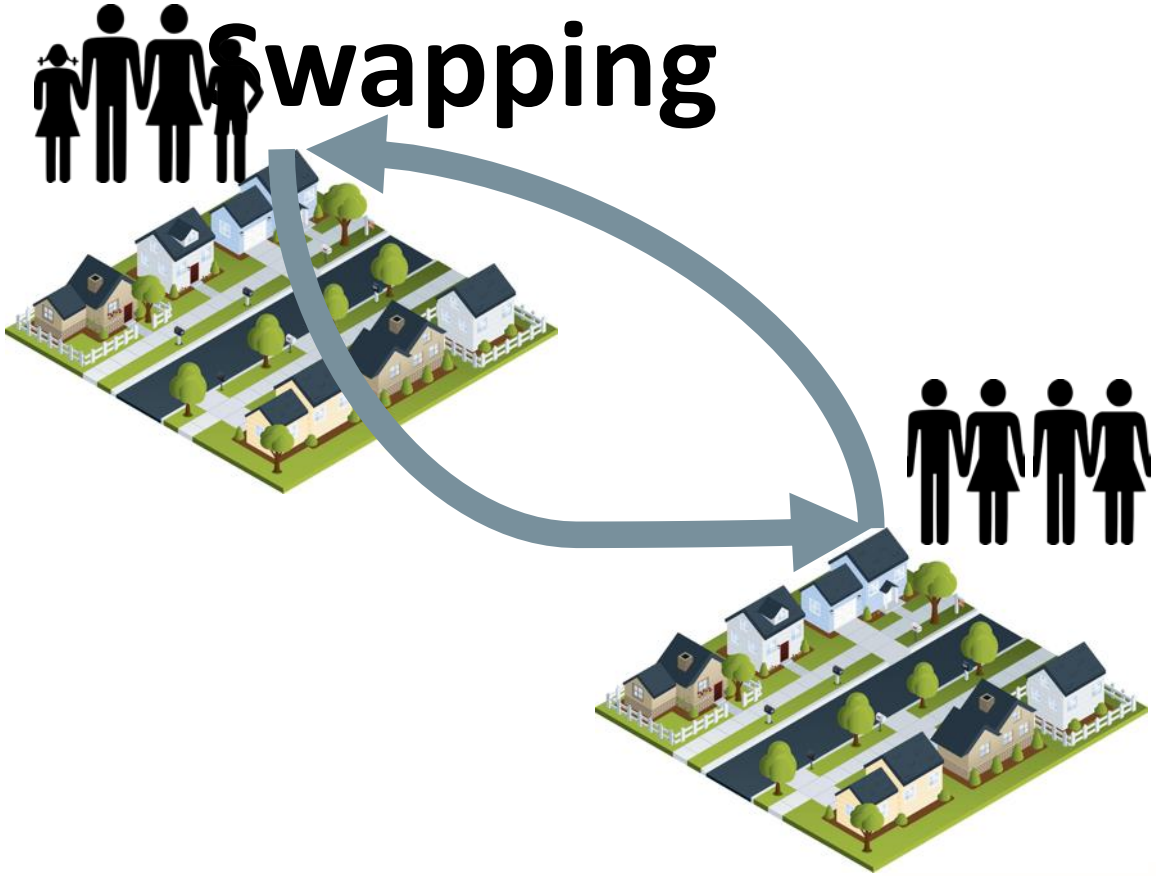
Aggregation

Swapping

Two privacy mechanisms for the 2010 Census

Aggregation

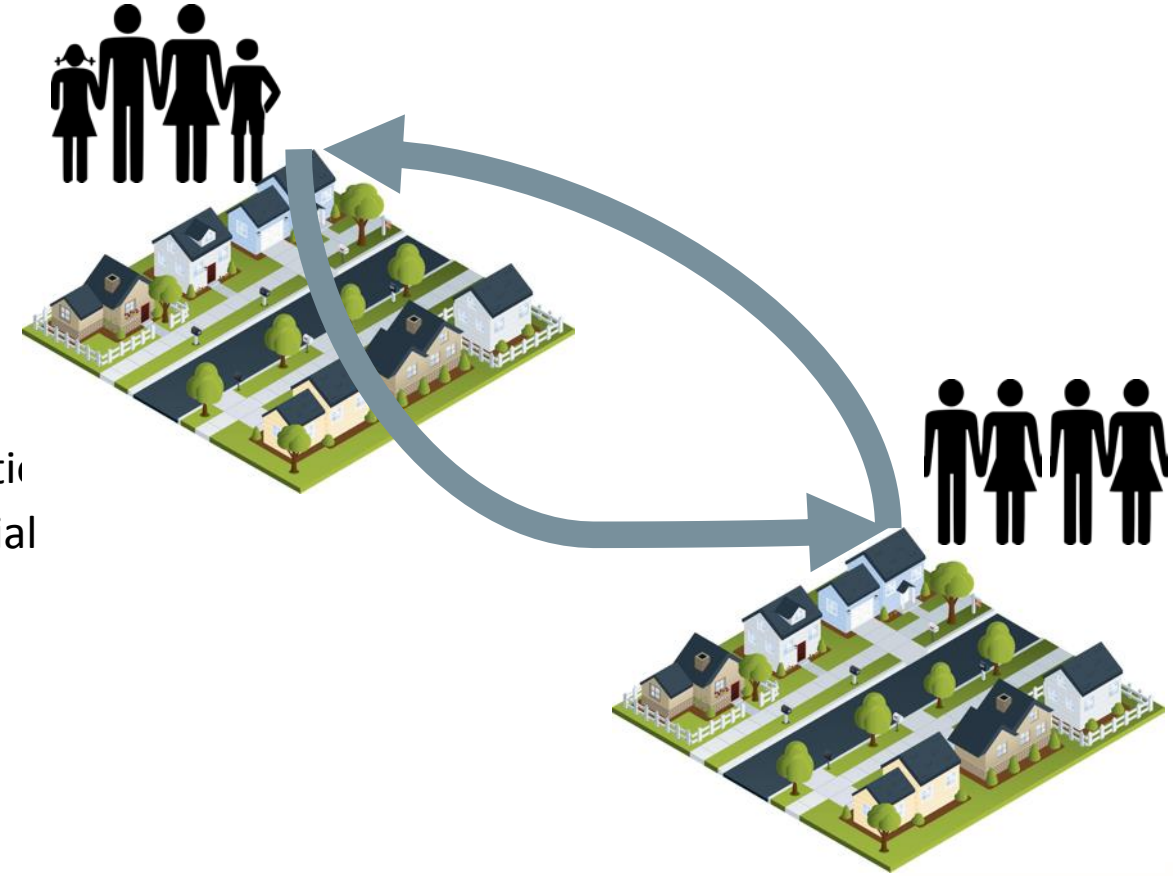
Swapping



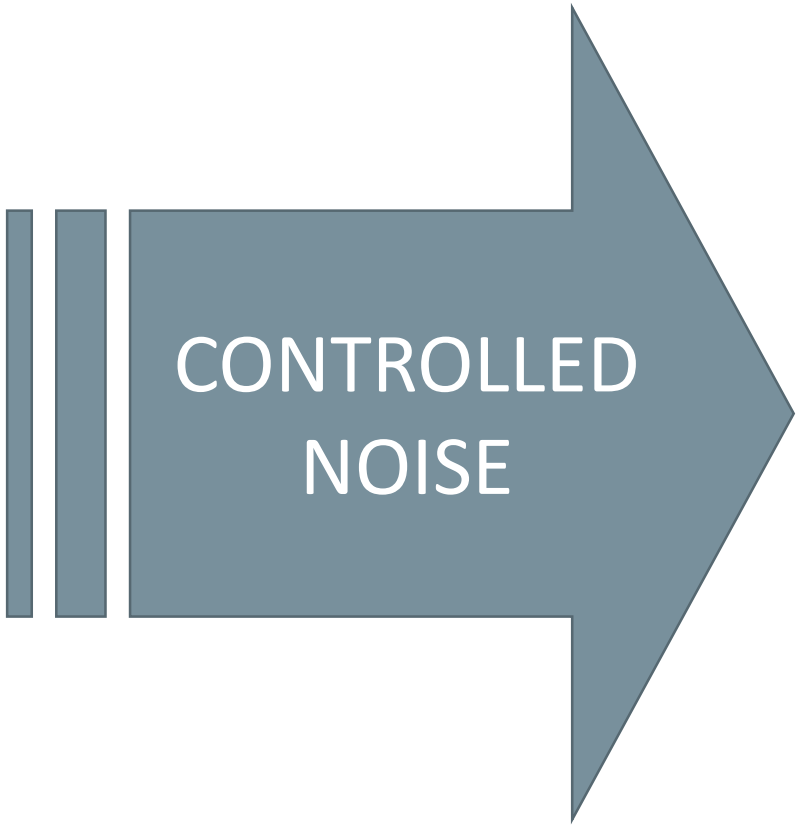
Household swapping was used in 2000 and 2010.

- Advantages of swapping:
  - Easy to understand
  - Does not affect state counts
  - Can be run state-by-state
  - Operation is “invisible” to the rest of census processing

- Disadvantages:
  - Does not consider or protect against database reconstructi
  - Swap rate and details of swapping must remain confidential
  - Privacy guarantee based on the lack of external data



Noise injection and differential privacy



Database reconstruction and noise injection

	Counts
Age < 18	4
Age >= 18	6
Race 1	4
Race 2	4
Race 3	2

NOISE

Counts
5
5
3
5
2

The more noise, the more privacy — and the less accuracy

	Counts
Age < 18	4
Age >= 18	6
Race 1	4
Race 2	4
Race 3	2

Little Noise

Counts
5
5
3
5
2

The more noise, the more privacy — and the less accuracy

Counts
2
8
8
1
1

BIG NOISE

	Counts
Age < 18	4
Age >= 18	6
Race 1	4
Race 2	4
Race 3	2

Little Noise

Counts
5
5
3
5
2

The more noise, the more privacy — and the less accuracy

Counts
2
8
8
1
1

BIG NOISE

	Counts
Age < 18	4
Age >= 18	6
Race 1	4
Race 2	4
Race 3	2

POSSIBILITY 1

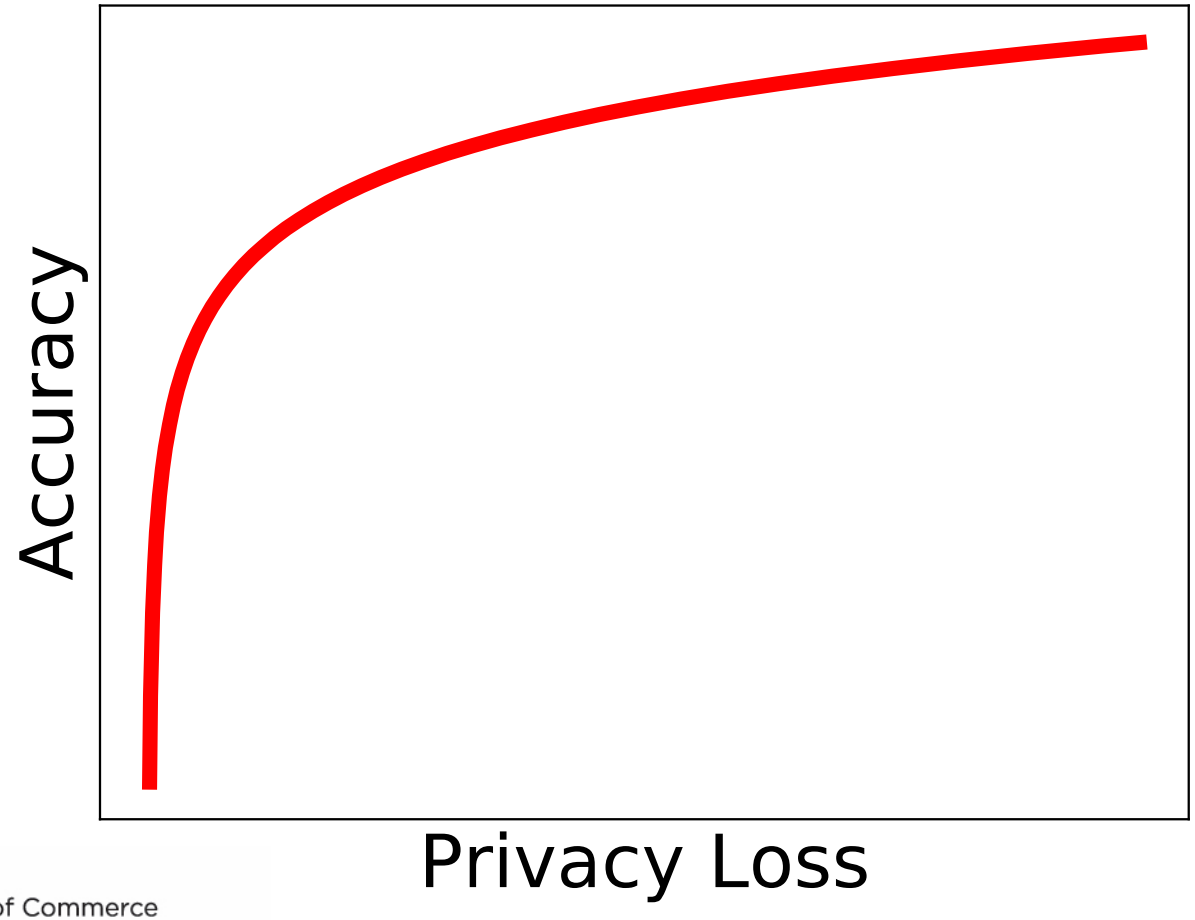
POSSIBILITY 2

	Counts
Age < 18	8
Age >= 18	2
Race 1	3
Race 2	2
Race 3	5

	Counts
Age < 18	3
Age >= 18	7
Race 1	5
Race 2	2
Race 3	3

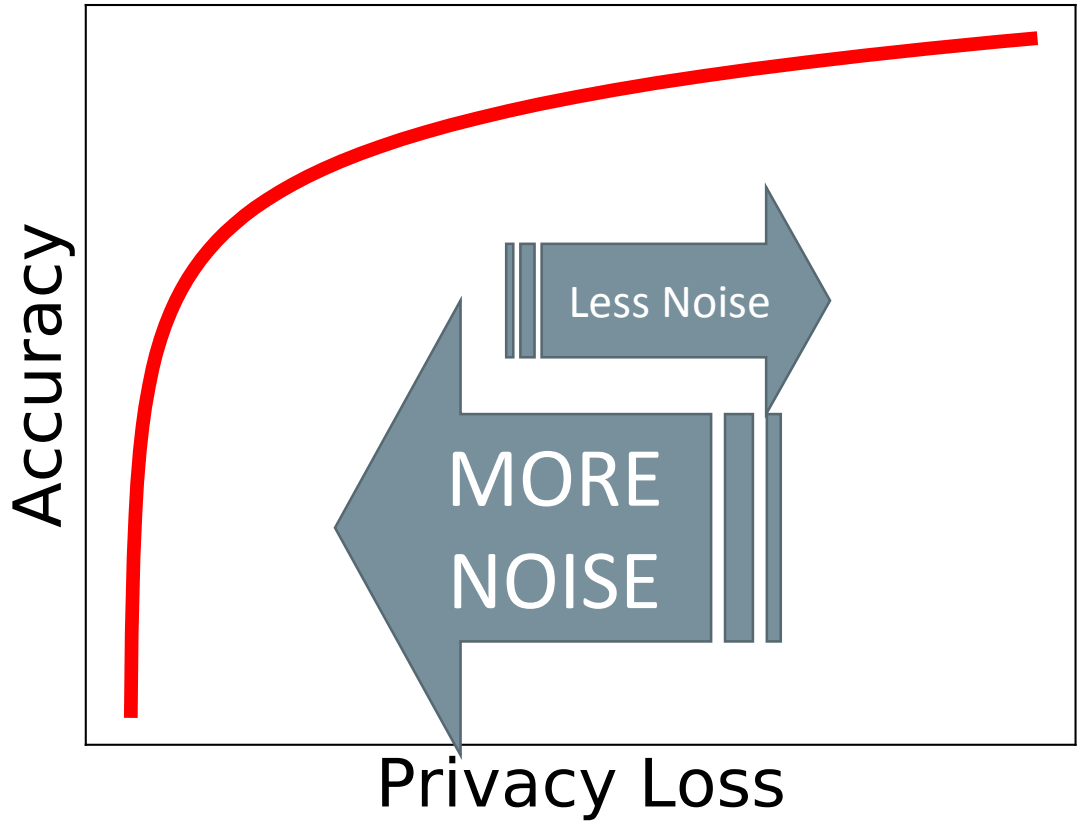
POSSIBILITY 3

Differential privacy is a tool for controlling the noise/accuracy trade-off



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

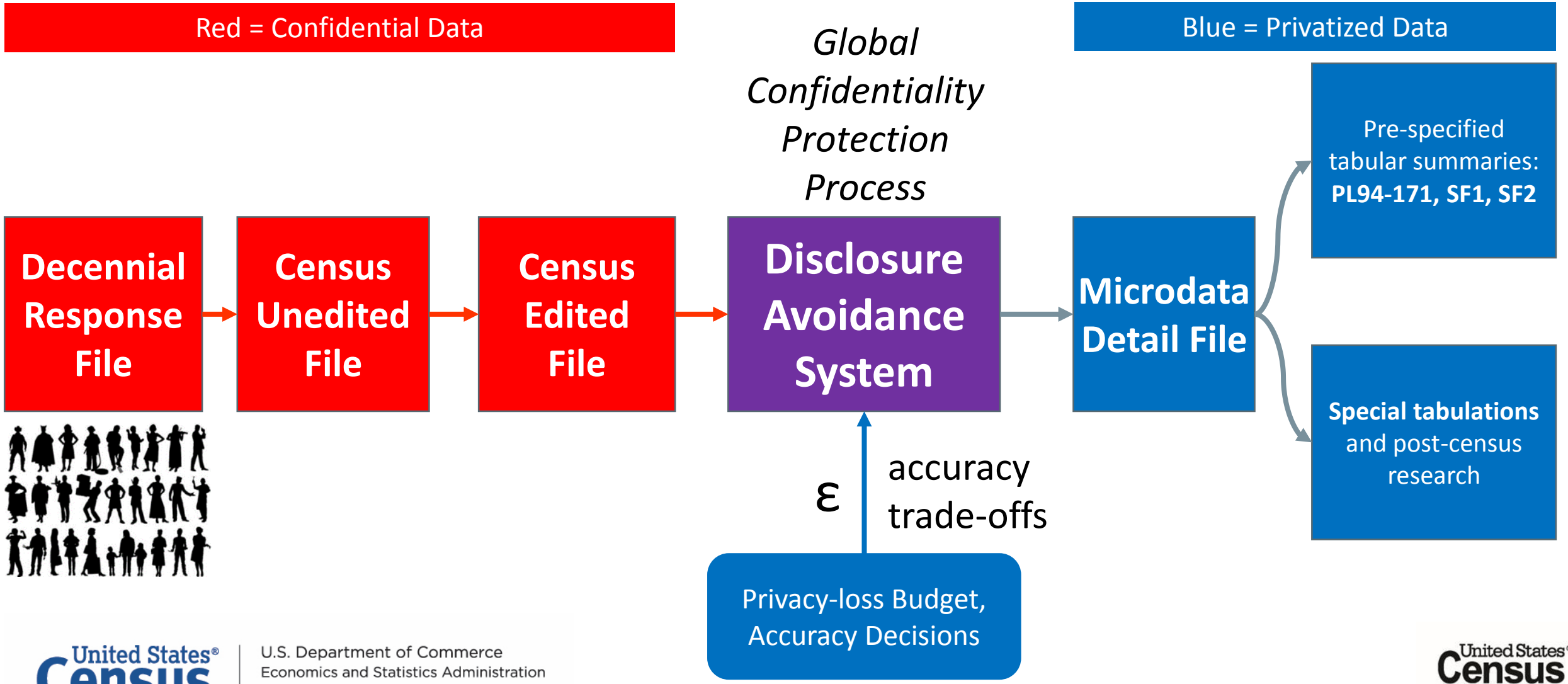
- Differential privacy provides:
  - **Provable bounds** on the maximum privacy loss
  - **Algorithms** that allow policy makers to manage the trade-off between accuracy and privacy loss.



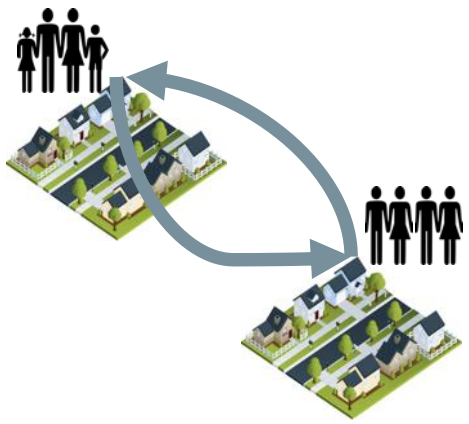
Final privacy-loss budget determined by the
Data Stewardship Executive Policy Committee (DSEP)
with recommendations from the Disclosure Review Board (DRB)

State of the project

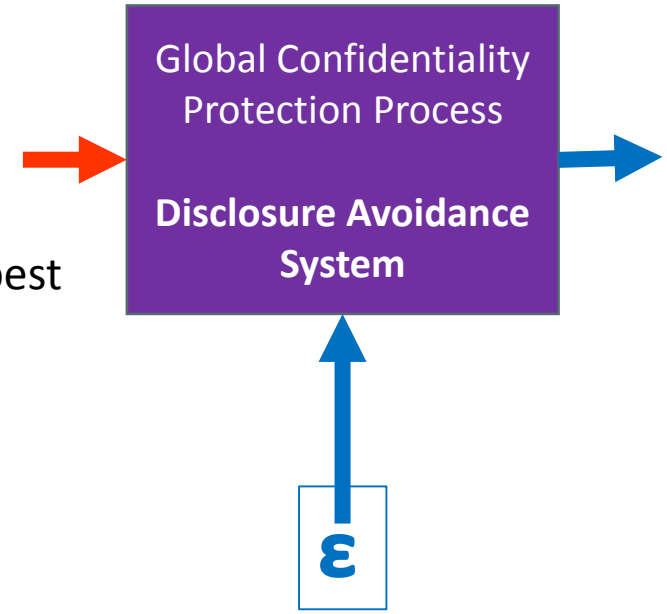
The “Disclosure Avoidance System” is part of the Census data processing pipeline



Differential privacy has many advantages to swapping

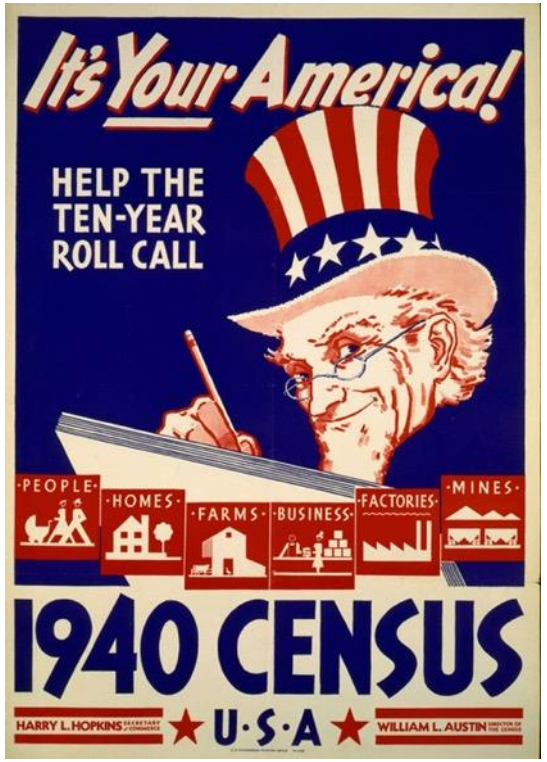


- **Advantages**:
  - Privacy guarantees are *tunable and provable*
  - Privacy guarantees are *future-proof*
  - Privacy guarantees are *public and explainable*
  - Protects against *database reconstruction*
- **Disadvantages**:
  - Entire country must be processed at once for best accuracy
  - Every use of private data must be tallied in the *privacy-loss budget*

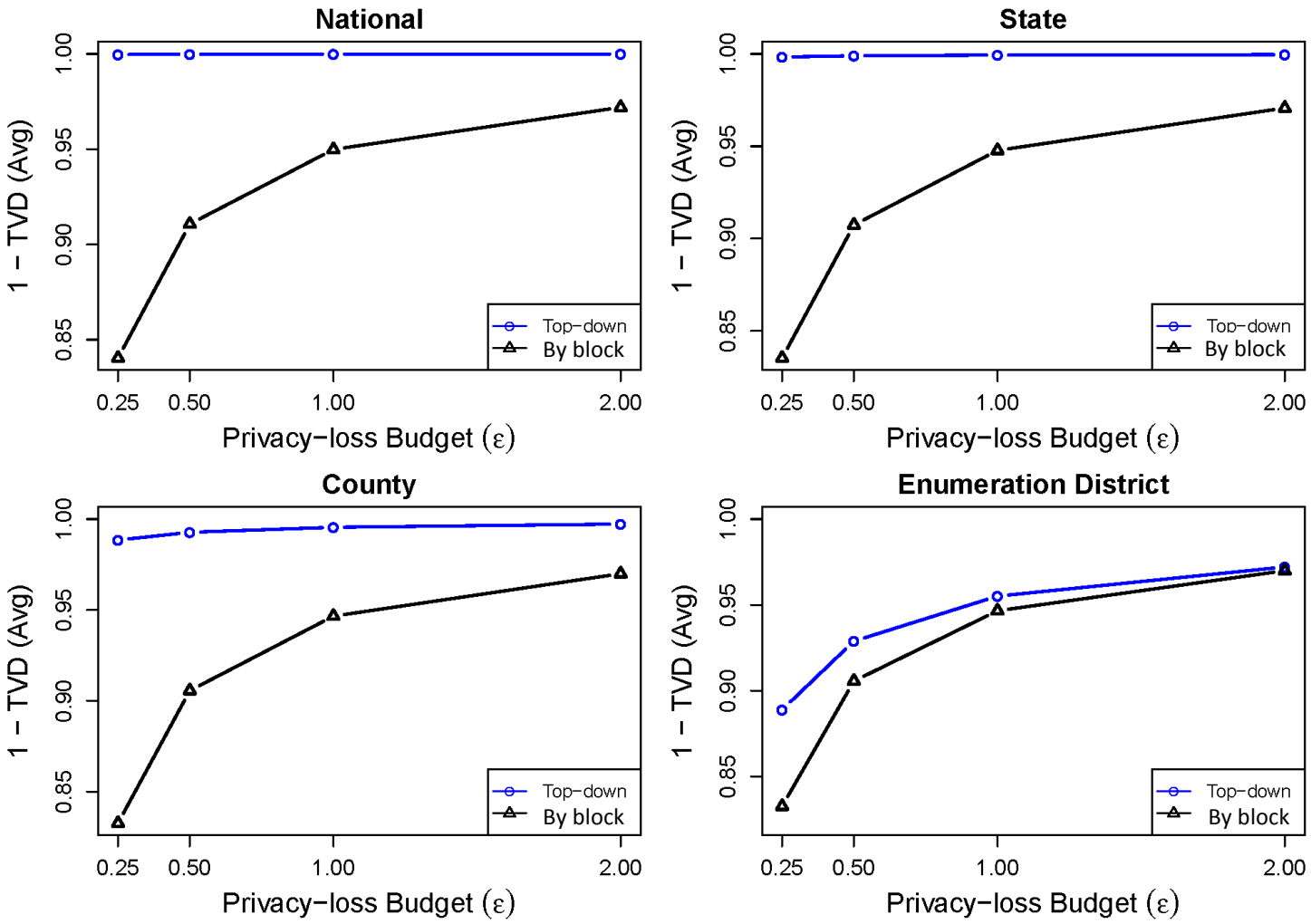


We will make the DAS public!

- Open source system
  - Source code published on the Internet
  - Testable with data from 1940 Census



Run of the 2020 DAS with 1940 Census data



Communications Strategy

- Differential privacy is not widely known or understood outside academia
- Most data users expect the same accuracy regardless of the level of detail
- In 2000 and 2010 we used swapping with an undisclosed swap rate
  - The Census Bureau did not quantify the error rate

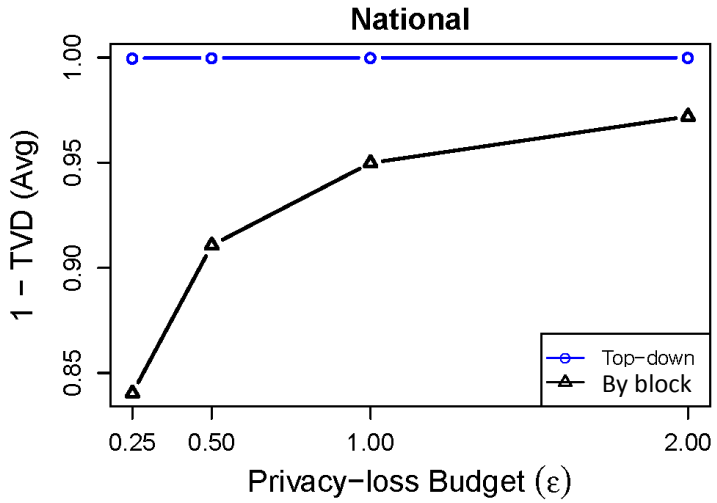
State of the DAS Project(s): Engineering & Science

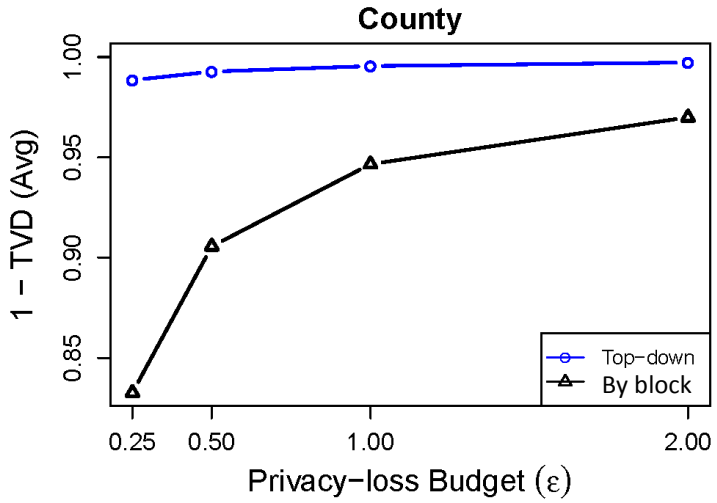
- **ENGINEERING PROJECT – Building a Turnkey Batch-Oriented System**
- Creating a production system that runs within the 2018 End-to-End Census Test and 2020 Census production environments
  - Resource intensive, but only when actively in use
  - Based on Amazon Elastic Map Reduce technology
  - Reads CEF from the Census Data Lake
  - Processes using DAS algorithms and a commercial optimizer
  - Creates the Microdata Detail File
  - Saves results in the Census Data Lake

Status: GREEN

State of the DAS Project(s): Engineering & Science

- **SCIENCE PROJECT — Improving the differential privacy algorithms**
- We are steadily improving the accuracy/privacy trade-off
- Progress requires interactive access to microdata from the 2010 Census, and continued access to high-performance computing on demand.





Looking forward

United States Census 2020

DAS Highlights --- Good news!

- The current “top-down” algorithm handles the PL94-171 queries and generates micro-data that meet the requirements to publish test files.

- We’re sharing tables with **Subject Matter Experts (SMEs)** and discussing possible improvements

- We will soon integrate the **High-Dimensional Matrix Mechanism (HDMM)**, into our top-down algorithm, which will improve accuracy on requested tabulations

- The Census Bureau is collecting “use cases” from our data users

FRN Notice

We want users of 2020 Census Data Products to tell us how they use our data!

First FRN:
83 FR 84111
7/19/2018 -> 9/17/2018

Second FRN:
83 FR 50636
10/09/2018 -> 11/08/2018

0 Sign in Sign up

NATIONAL ARCHIVES

FEDERAL REGISTER
The Daily Journal of the United States Government

Notice

Soliciting Feedback From Users on 2020 Census Data Products; Reopening of Comment Period

A Notice by the Census Bureau on 10/09/2018

This document has a comment period that ends in 33 days. (11/08/2018) SUBMIT A FORMAL COMMENT

PUBLISHED DOCUMENT

Start Printed Page 50636

AGENCY:

Bureau of the Census, Department of Commerce.

ACTION:

Notice and request for comment; reopening of comment period.

SUMMARY:

The Bureau of the Census (Census Bureau) is reopening the comment period provided in the notice entitled "Soliciting Feedback from Users on 2020 Census Data Products," which was published in the **Federal Register** on July 19, 2018, in order to allow interested parties additional time to submit comments. The public comment period on that notice closed on September 17, 2018.

DOCUMENT DETAILS

Printed version:
PDF

Publication Date:
10/09/2018

Agencies:
Bureau of the Census

Dates:
The Census Bureau is reopening the comment period for the notice entitled ``Soliciting Feedback from Users on 2020 Census Data Products,'' which was published in the Federal Register on July 19, 2018 (83 FR 34111). The Census Bureau will accept comments received on this notice by November 8, 2018.

DAS Science Highlights --- Challenges!

- We have not yet addressed **household queries** or **person-household joins**, although we have in-progress research for both
  - Householder queries, e.g. “how many households are headed by someone aged 20-30?”
  - Person-household join, e.g. “how many children are in households headed by someone aged 20-30?”
- Lack of scientists and engineers trained in differential privacy

2020 Disclosure Avoidance System: Conclusions

- We are using **differential privacy** to assure that published statistics do not violate the Census Bureau's Title 13 obligations
- This is a huge step forward for the Census Bureau
- We have a working system and will use it for the 2018 End-to-End Census Test
  - For 2018 we are only producing the PL94-171 redistricting tabulations
- There is a lot of scientific work that remains to be done
- Contact: Simson.L.Garfinkel@census.gov John.M.Abowd@census.gov

QUESTIONS?