

2020 Disclosure Avoidance System (DAS)

Presenter: John L. Eltinge

Assistant Director for Research and Methodology

Presenting materials originally from:

Simson L. Garfinkel

Senior Computer Scientist for Confidentiality and Data Access

John M. Abowd

Chief Scientist and Associate Director for Research and Methodology (ADRM)

Acknowledgements and Disclaimer

Almost all of the materials covered in the slides were originally prepared by Simson Garfinkel and John Abowd of the United States Census Bureau.

The views expressed in this presentation are those of the authors and speaker, and do not necessarily represent the policies of the United States Census Bureau.

General Background

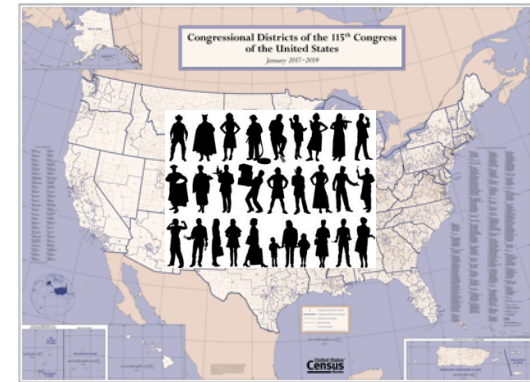
Essentially All Large-Scale Statistical Programs Require a Complex Balance of Multiple Dimensions of:

- **Quality**
- **Risk (Including Disclosure Risk)**
- **Cost**

Disclosure Avoidance System

Purpose

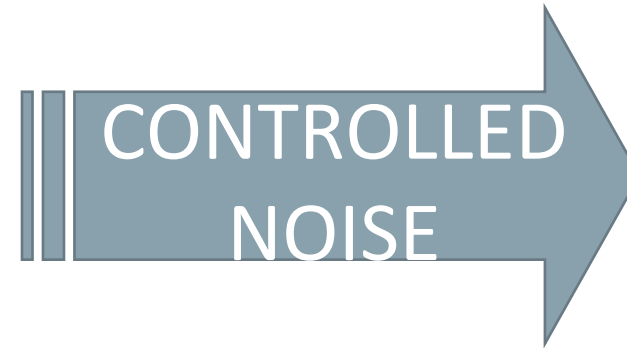
- The Disclosure Avoidance System (DAS) assures that the 2020 Census data products meet the legal requirements of Title 13, Section 9 of the U.S. Code.
- The DAS is designed to prevent **improper disclosures of data about individuals and establishments** in the 2020 census data products.
- Stakeholders: 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



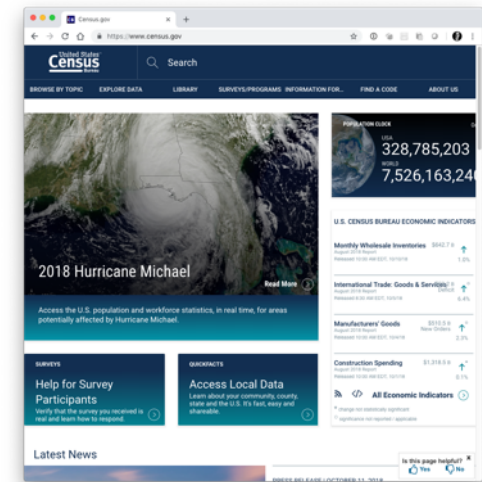
United States
Census
2020

Project purpose:

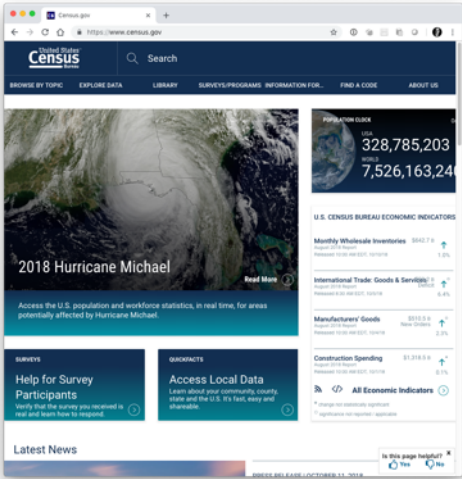
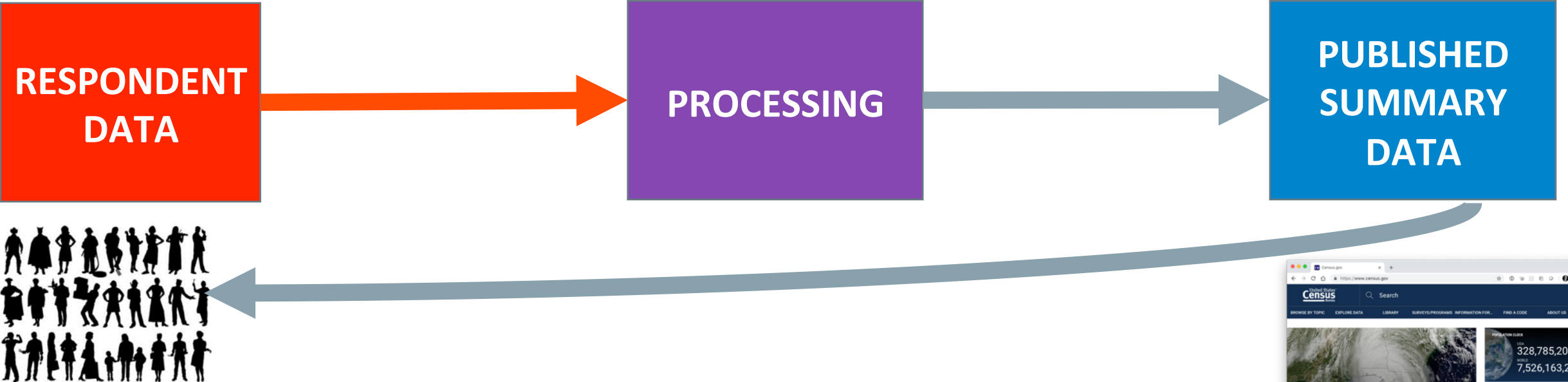
Why we need a new disclosure avoidance system



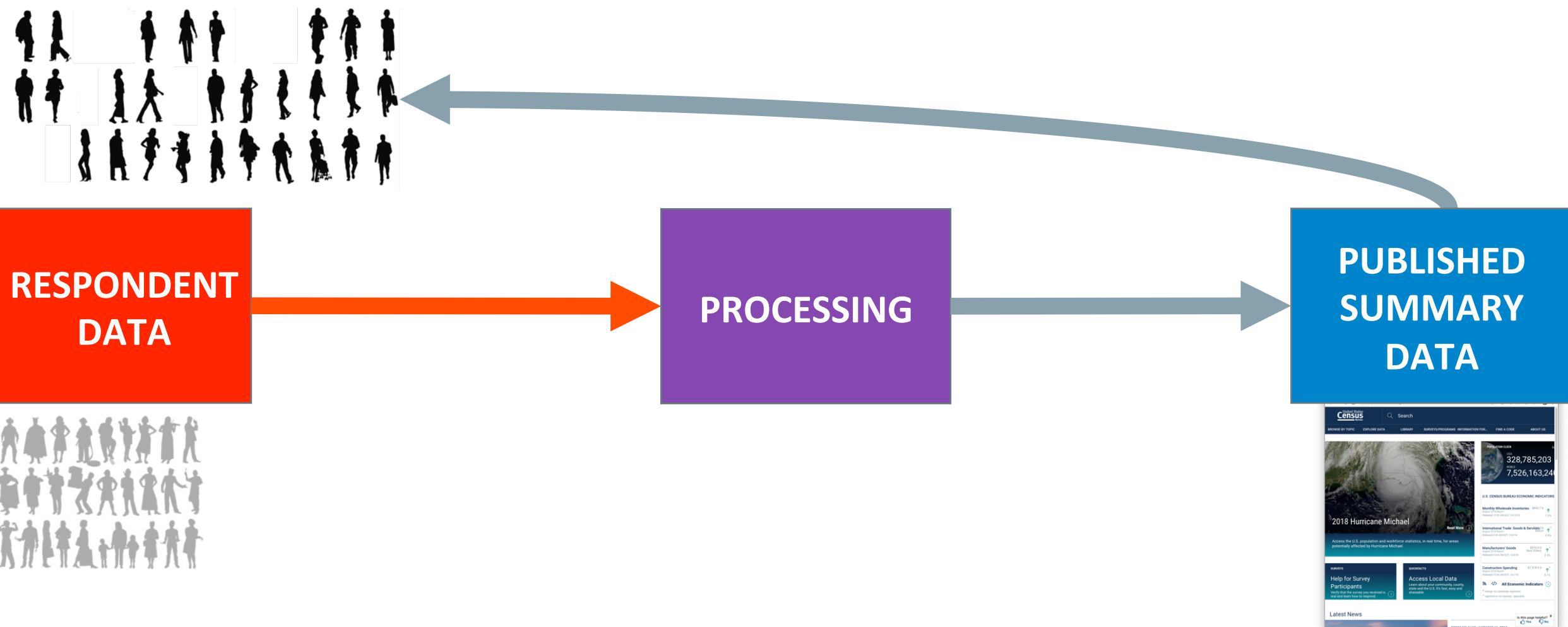
We create statistics by collecting data, processing and publishing



Database reconstruction is a mathematical process that reverses this process.



Database reconstruction is a mathematical process that reverses this process.



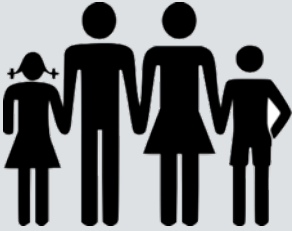
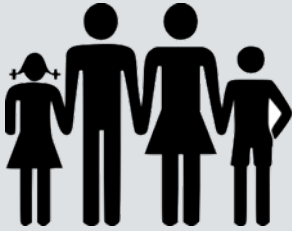
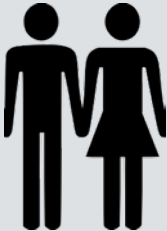
Consider a census block:



PUBLISHED DATA

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

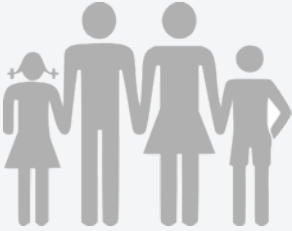
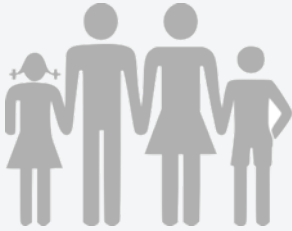

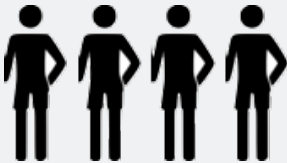
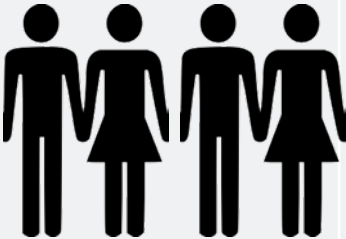

RECONSTRUCTED DATA

	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

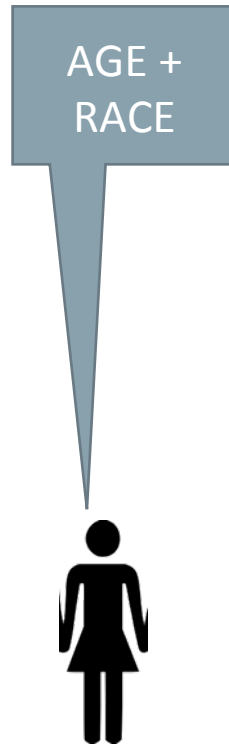
RECONSTRUCTED DATA

	Race 1	Race 2	Race 3
R1			
R2			

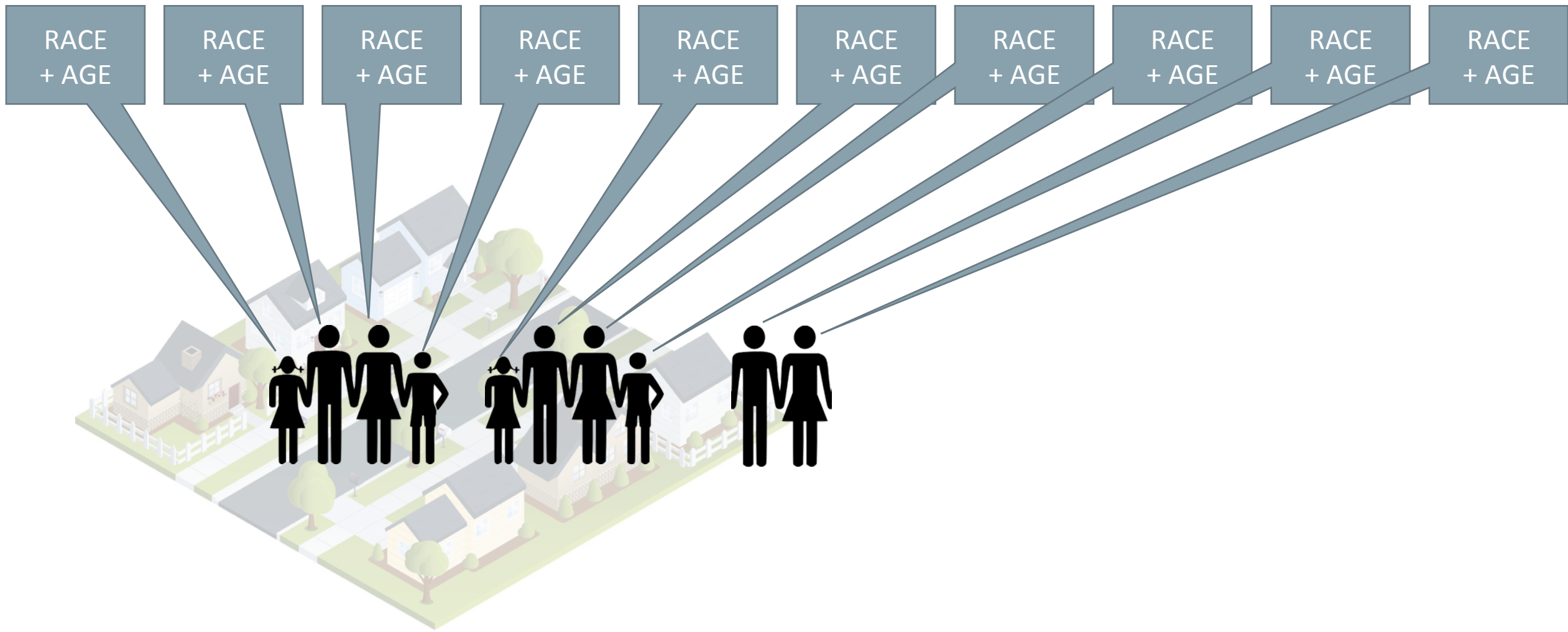
PUBLISHED DATA

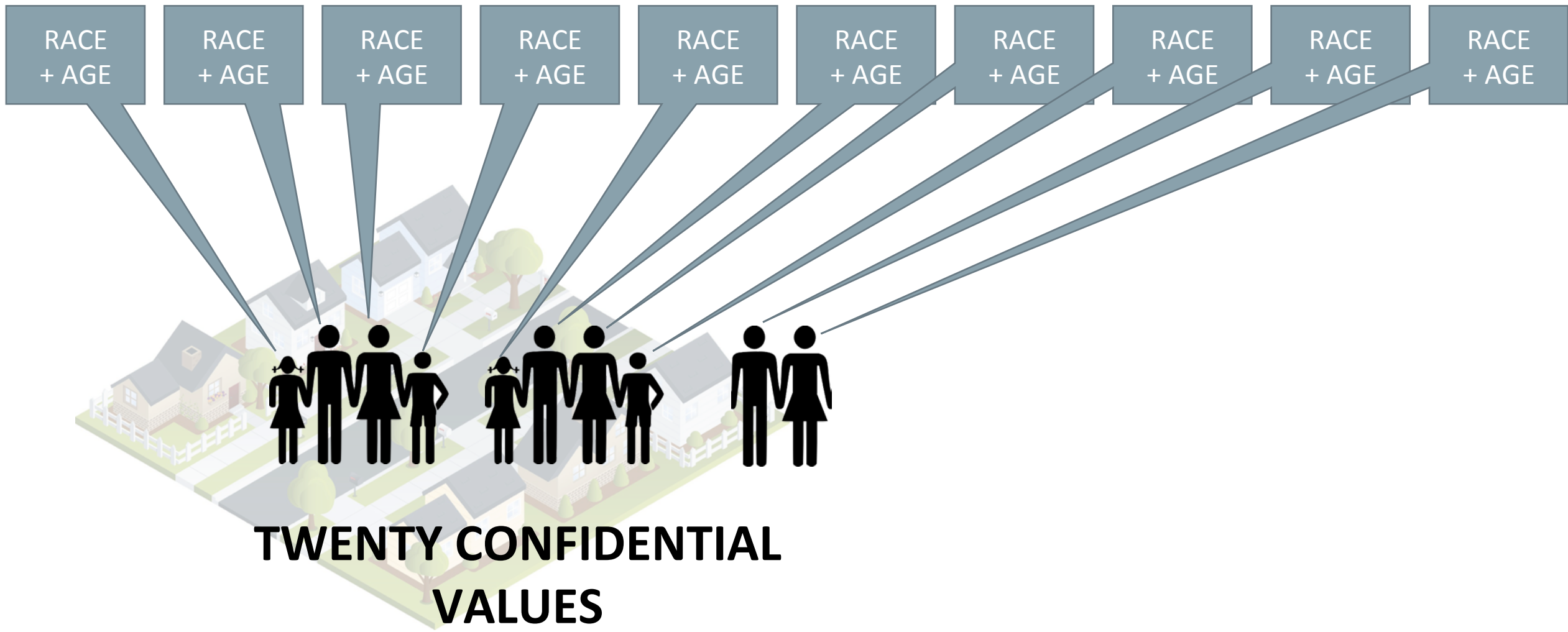
	Counts
Age < 18	4
Age >= 18	6
Race 1	4
Race 2	4
Race 3	2

	Race 1	Race 2	Race 3
			



AGE \geq 18







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

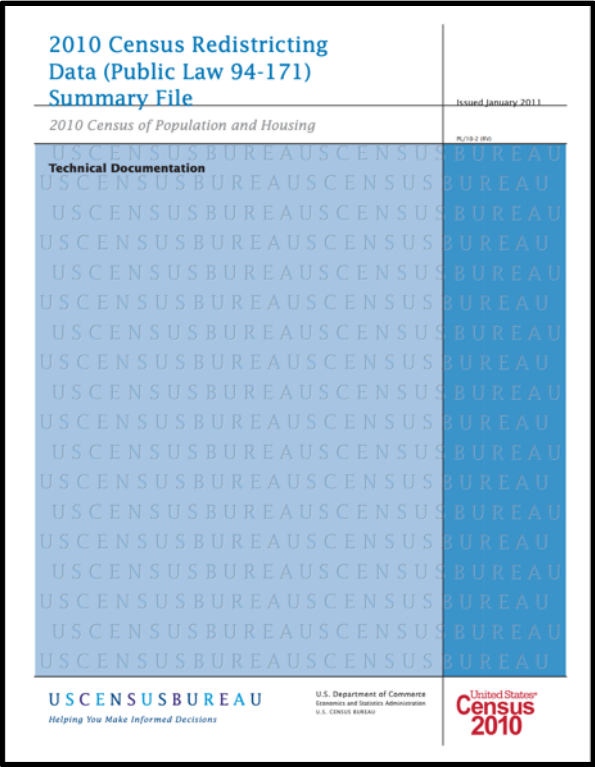
United States[®]
Census
2010

United States™
Census
Bureau

2010 Census of Population and Housing

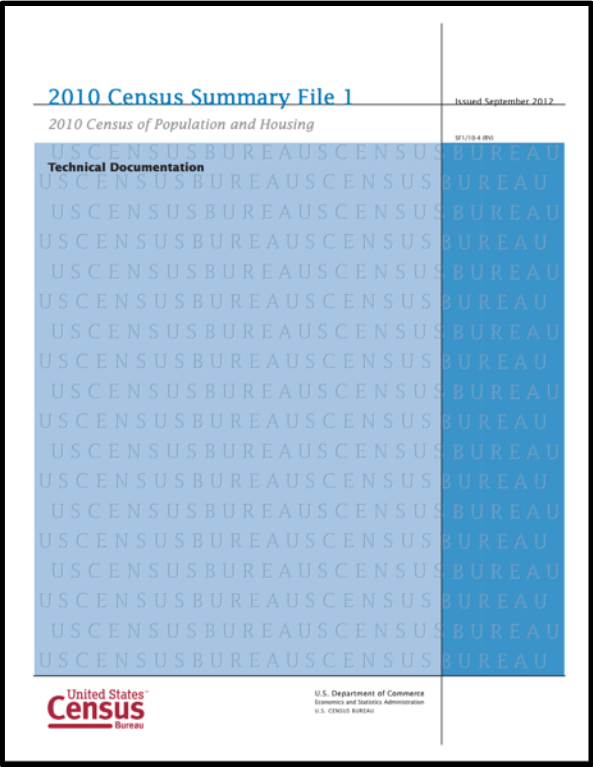
Total population	308,745,538
Pieces of information per person:	6
Total pieces of information:	1,852,473,228

2010 Census Publication Schedule



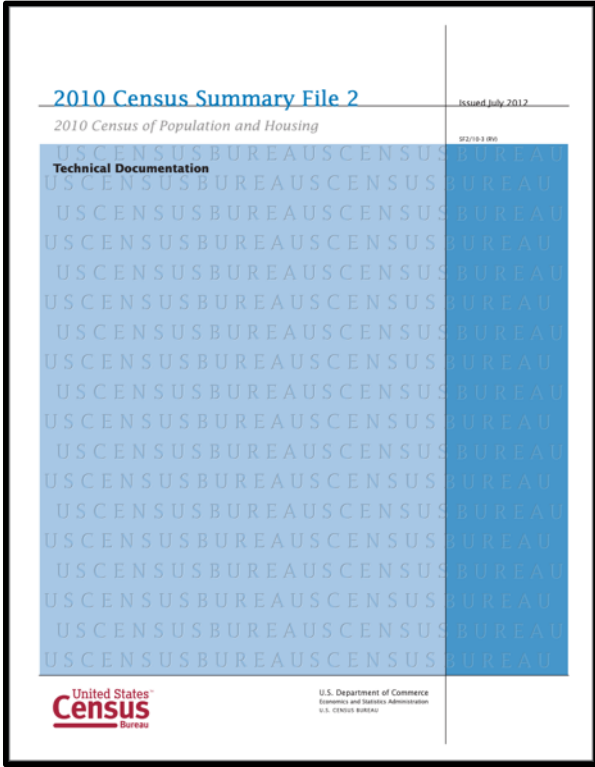
PL94-171 Redistricting

2,771,998,263



Balance of Summary File 1

2,806,899,669



Summary File 2

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

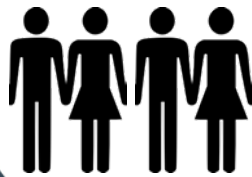
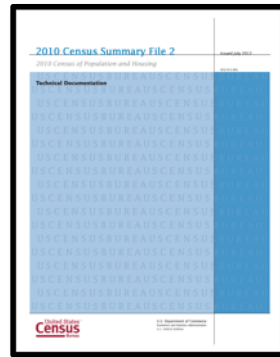
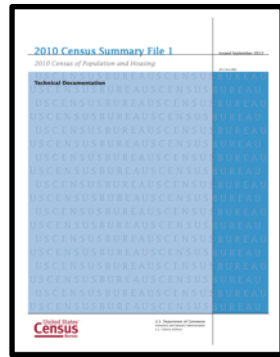
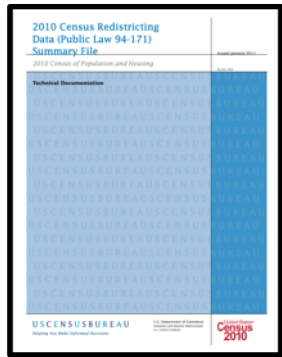
Two privacy mechanisms for the 2010 Census

Aggregation

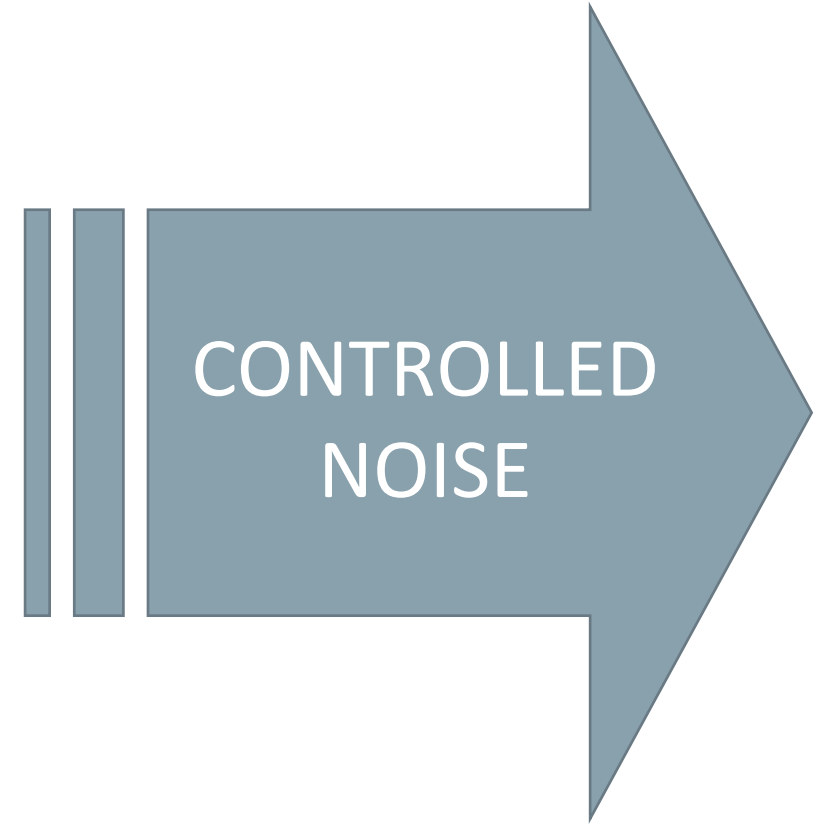


Two privacy mechanisms for the 2010 Census

Aggregation

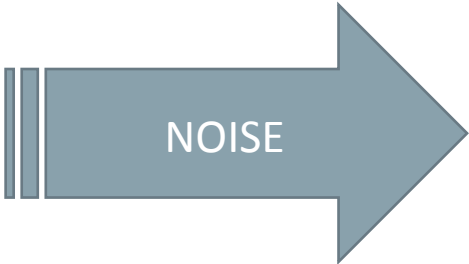


Noise injection and differential privacy



Database reconstruction and noise injection

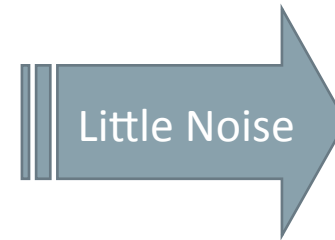
	Counts
Age < 18	4
Age >= 18	6
Race 1	4
Race 2	4
Race 3	2



Counts
5
5
3
5
2

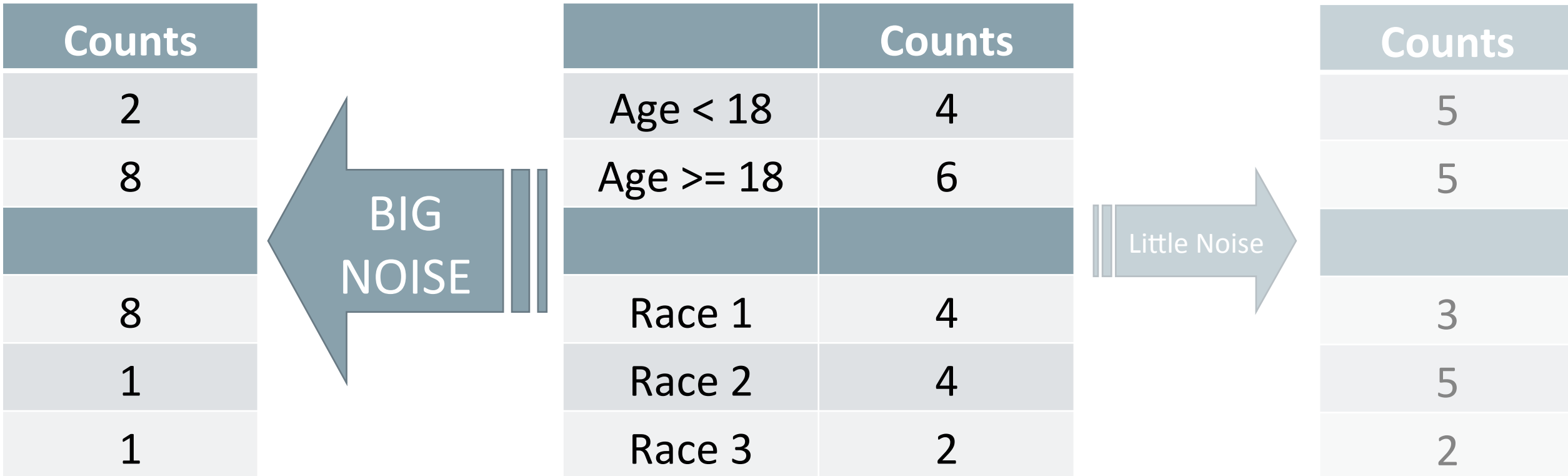
The more noise, the more privacy — and the less accuracy

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



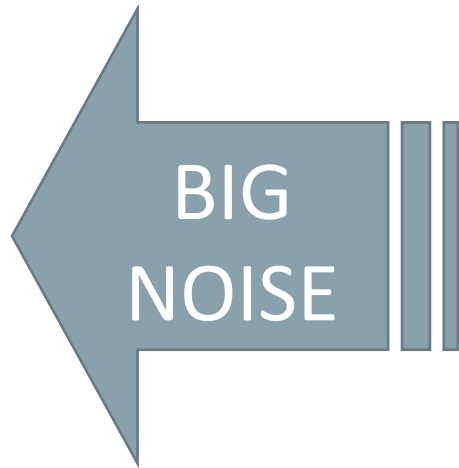
Counts
5
5
3
5
2

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



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

Counts
2
8
8
1
1



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

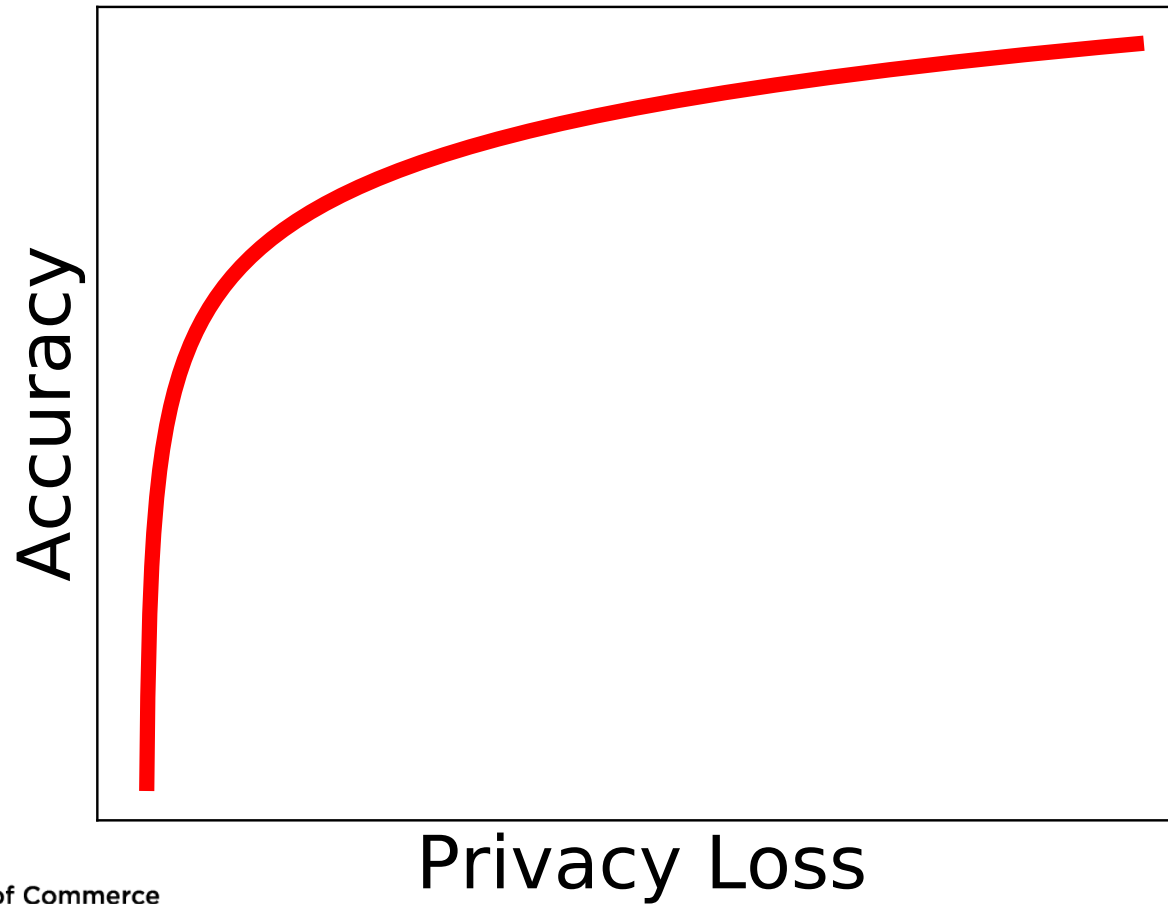
POSSIBILITY 1

POSSIBILITY 2	
	Counts
Age < 18	8
Age >= 18	2

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

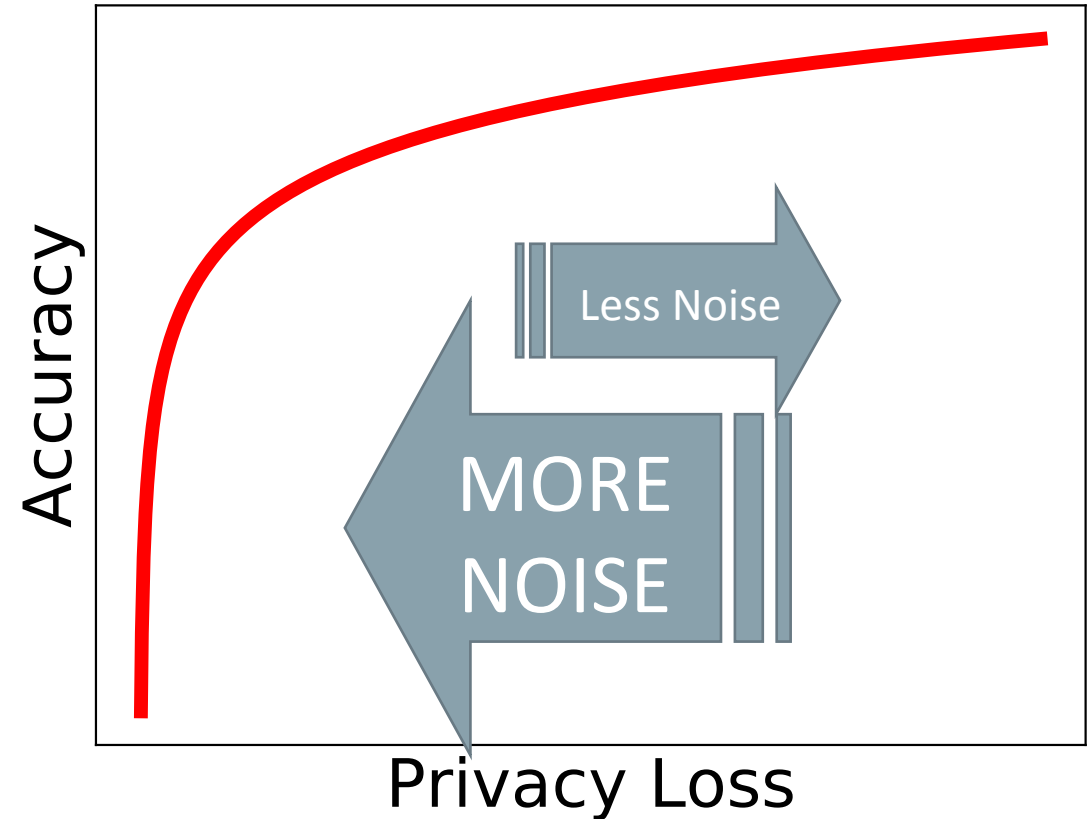
POSSIBILITY 3

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.

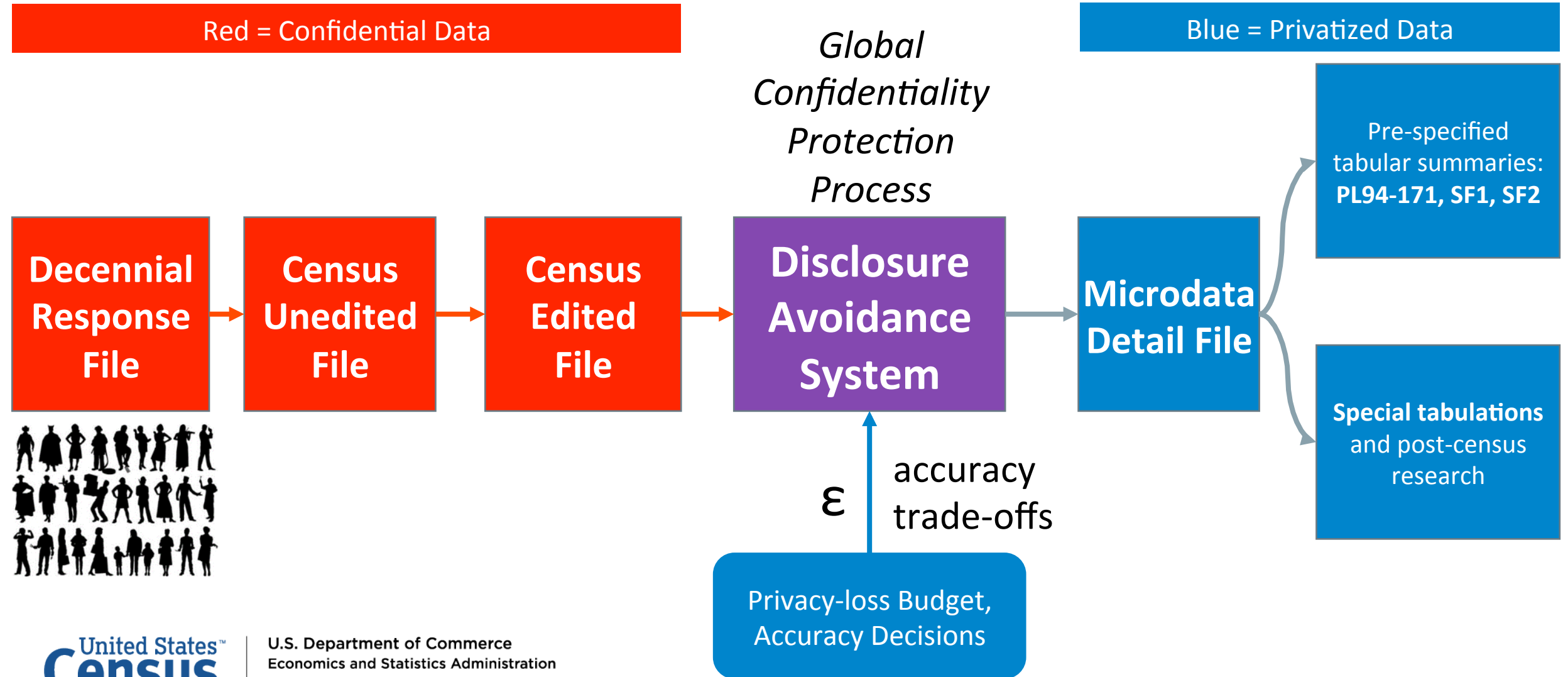
- Differential privacy provides:
 - **Provable bounds** on the maximum privacy loss
 - **Algorithms** that allow policy makers to manage the trade-off between accuracy and privacy loss.



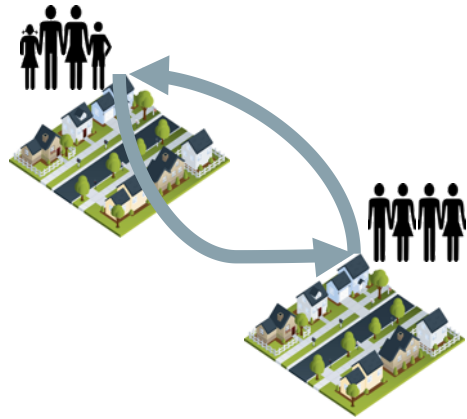
Final privacy-loss budget determined by the
Data Stewardship Executive Policy Committee (DSEP)
with recommendations from the Disclosure Review Board (DRB)

State of the project

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



Differential privacy has many advantages to swapping

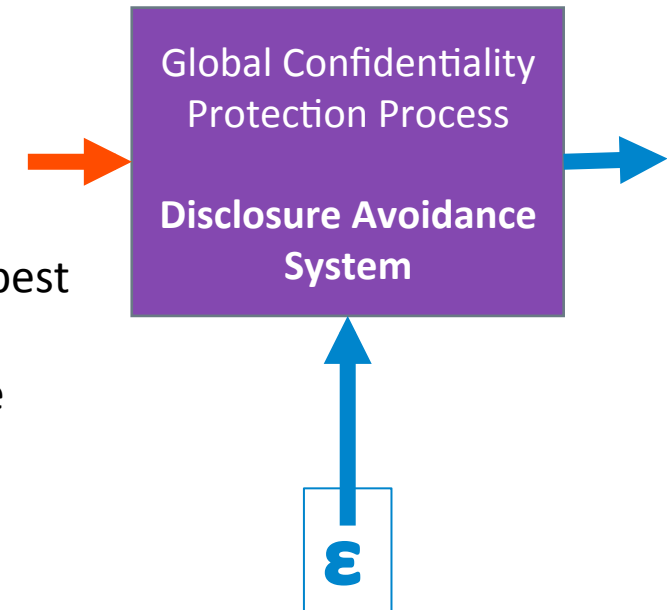


- **Advantages:**

- 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



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

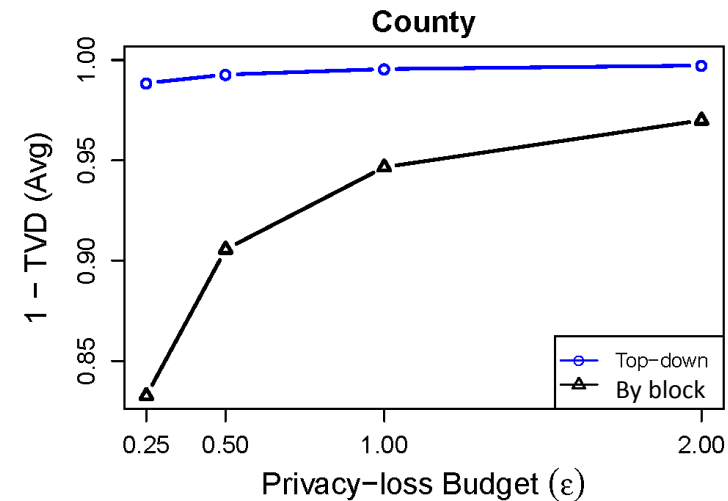
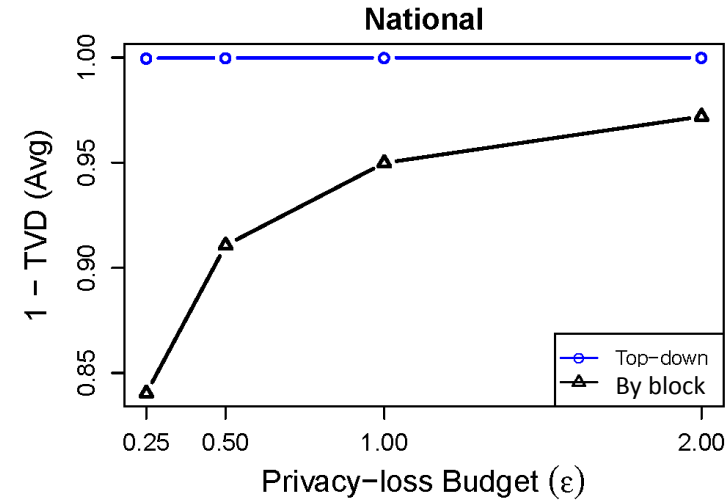
The screenshot shows the 'Data Protection and Privacy' page on the U.S. Census Bureau website. The page has a dark blue header with the 'United States Census Bureau' logo and a search bar. Below the header is a navigation bar with links: 'BROWSE BY TOPIC', 'EXPLORE DATA', 'LIBRARY', 'SURVEYS/PROGRAMS', 'INFORMATION FOR...', 'FIND A CODE', and 'ABOUT US'. The main content area is titled 'Data Protection and Privacy' and includes a sidebar with links: 'Data Stewardship', 'Statistical Safeguards', 'Privacy Impact Assessments (PIA)', 'System of Records Notices (SORN)', 'Online Privacy Policy', and 'Contact Us'. The main text states: 'We are committed to handling your information responsibly. Your information is kept confidential. This commitment applies to the individuals, households, and businesses that answer our surveys, and to those browsing our website.' To the right of this text is a graphic of four stylized human figures with various data points listed next to them, such as 'BANK ACCOUNT', 'CITY OF BIRTH', 'AGE', 'PASSPORT NUMBER', 'CITY OF RESIDENCE', 'SCHOOL', 'TELEPHONE NUMBER', 'JOB', 'ARREST', 'SALARY', 'DRIVERS LICENSE', 'GEOGRAPHIC LOCATION', 'CITY OF BIRTH', 'CITY OF RESIDENCE', 'SCHOOL', 'TELEPHONE NUMBER', 'JOB', 'ARREST', 'SALARY', 'DRIVERS LICENSE', 'GEOGRAPHIC LOCATION'. Below the text are three images: a word cloud with terms like 'INFORMATION', 'PROTECTION', 'ENCRYPTION', 'INTEGRITY', 'SECURITY', 'LOGIC', 'CODE', 'PASSWORD', 'SAFE'; a padlock icon; and a group of four people. Below the images are three links: 'Protecting Online Privacy', 'Protecting Your Data', and 'Our Privacy Principles'. The page also has a 'Survey Information' section with a link 'Are You in a Survey?' and a description: 'If you have received a survey, this site will help you verify that the survey came from us, understand and complete the form, and know how we protect your data.' Below this is a section 'Our Surveys & Programs' with a description: 'Our surveys provide periodic and comprehensive statistics about the nation. This data is critical for government programs, policies, and decision-making.' At the bottom right is a feedback box asking 'Is this page helpful?' with 'Yes' and 'No' buttons. The footer of the page shows a search bar with 'privacy' entered, a 'Highlight All' button, and a match count of '2 of 7 matches'.

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

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](#)



FEDERAL REGISTER

The Daily Journal of the United States Government

0  [Sign in](#) [Sign up](#)

N 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
- Many open questions in mathematical statistics and methodology

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?