Protecting the Confidentiality of the 2020 Census Statistics

Simson L. Garfinkel
Senior Scientist, Confidentiality and Data Access
U.S. Census Bureau

Presented to the Trust and Confidentiality Working Group
California Complete Count—Census 2020
Friday, March 1, 9:30am Pacific Time

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

Questions we've been asked to address

1. How is the U.S. Census Bureau's messaging concerning privacy and confidentiality different from 2010?
2. What is the U.S. Census Bureau's messaging to explain differential privacy to the public?
3. How does the U.S. Census Bureau's differential privacy safeguard both individual and block-level data?
4. Which industry professionals are conducting penetration tests and what are the results of those tests? How are the results of those tests being communicated to help build public trust in data confidentiality?
5. Who is hosting the U.S. Census Bureau's data? Which cloud service is being used?

Privacy and Confidentiality: 2010 vs. 2020

Traditional Methods: the 2000 and 2010 censuses:

DRF

Raw data from respondents: **Decennial Response File**

CUF

Selection & unduplication: **Census Unedited File**

CEF

Edits, imputations: **Census Edited File**

HDF

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

Confidential data

Public data

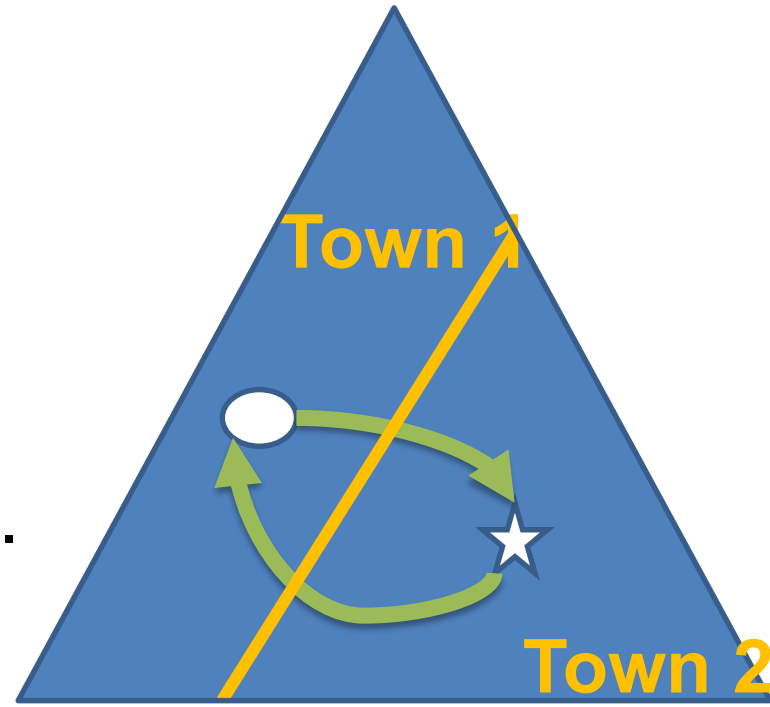
The protection system used in 2000 and 2010 relied on swapping households.

Some households were swapped with other households

- Swapped households had the same size.
- Swapping limited to within each state.

Disadvantages:

- Swap rate and details of swapping not disclosed.
- Privacy protection was not quantified.
- Impact on data quality not quantified.



State “X”

After we swapped, we tabulated the 2010 Census of Population and Housing

Basic results from the 2010 Census:

Total population	**308,745,538**
Household population	300,758,215
Group quarters population	7,987,323

Households	**116,716,292**

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

Today's reality:

Too many statistics published too accurately from a confidential database exposes the entire database with near certainty.

The Database Reconstruction Theorem
Dinur & Nissim, 2003

In 2010 and earlier censuses:

- We released exact population counts at the block, tract and county level.
- We released counts for age in years, OMB race/ethnicity, sex, relationship to householder, in Summary File 2: detailed race data based on the swapped data.
- We released 25 statistics per person, but only collected six pieces of data per person:

  Block • Age • Sex • Race • Ethnicity • Relationship to Householder

We tested how well 2010 Census data could withstand a simulated database reconstruction and re-identification attack using today's data science

1. Reconstructed all 308,745,538 microdata records.
2. Linked the reconstructed records to commercial databases to acquire PII
   - Successful linkages to commercial data = "putative re-identifications"
3. Compared putative re-identifications to confidential data
   - Successful linkages to confidential data = "confirmed re-identifications"
4. Harm: attacker can learn self-response race and ethnicity

Results: traditional privacy protection methods now too vulnerable; no longer acceptable.

- Matched *about half* of the people enumerated in the 2010 Census to commercial and other online information.
- *But:* more than half of these matches are incorrect.
- *And:* an external attacker has no means of confirming them.
- For the confirmed re-identifications, race and ethnicity are learned exactly, not statistically

1. “How is the U.S. Census Bureau's messaging concerning privacy and confidentiality different from 2010?”

“How is the U.S. Census Bureau's messaging concerning privacy and confidentiality different from 2010?”

For the 2020 Census, we are adopting *Formal Privacy* to protect privacy and confidentiality.

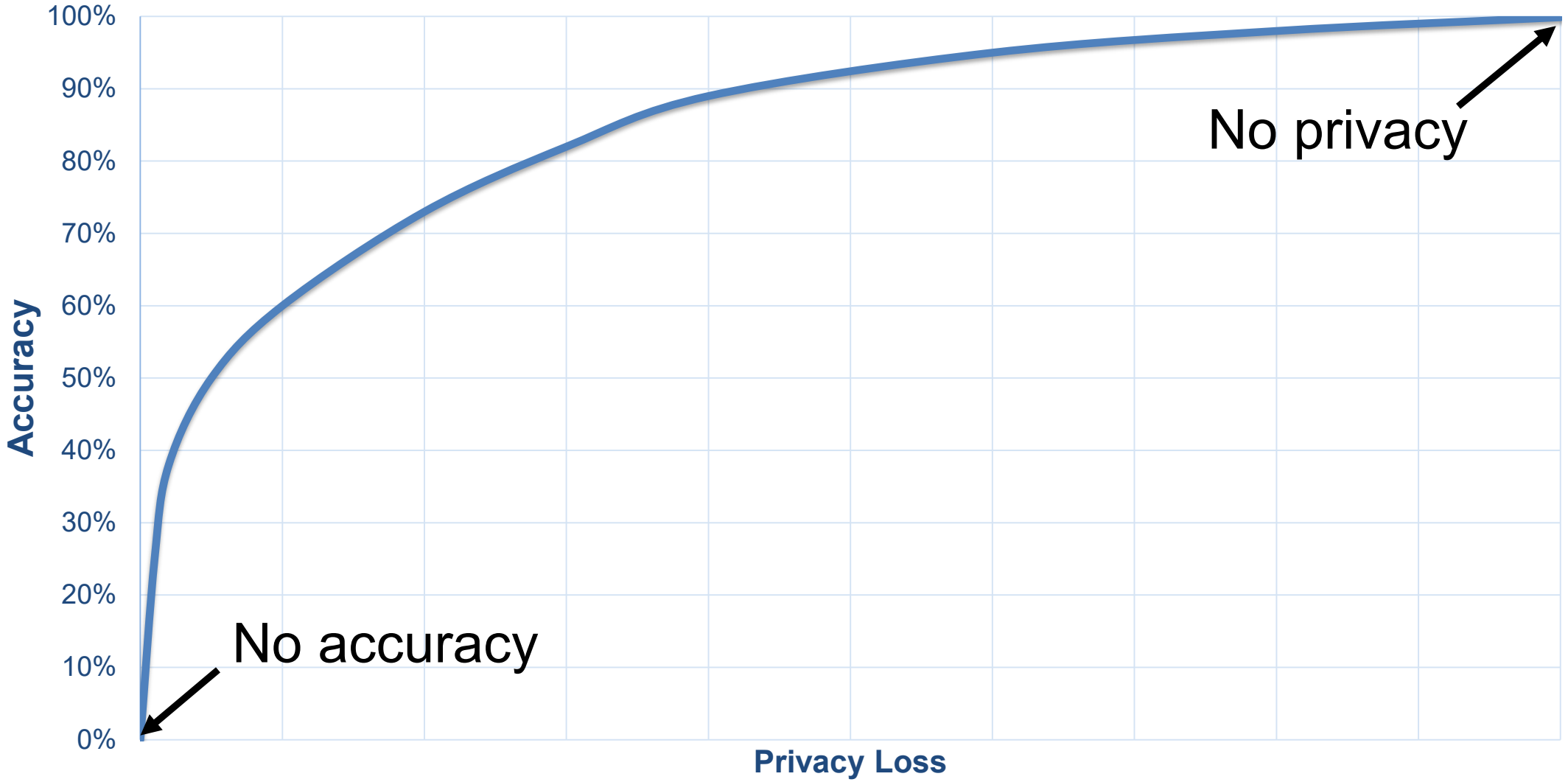
Formal Privacy gives us:

- Techniques for protecting confidentiality that have been subjected to rigorous, mathematical proofs.
- The ability to directly manage the accuracy/privacy-loss trade-off.

Note: Application of formal privacy currently limited to the 2020 Census.

Formal Privacy and the 2020 Census

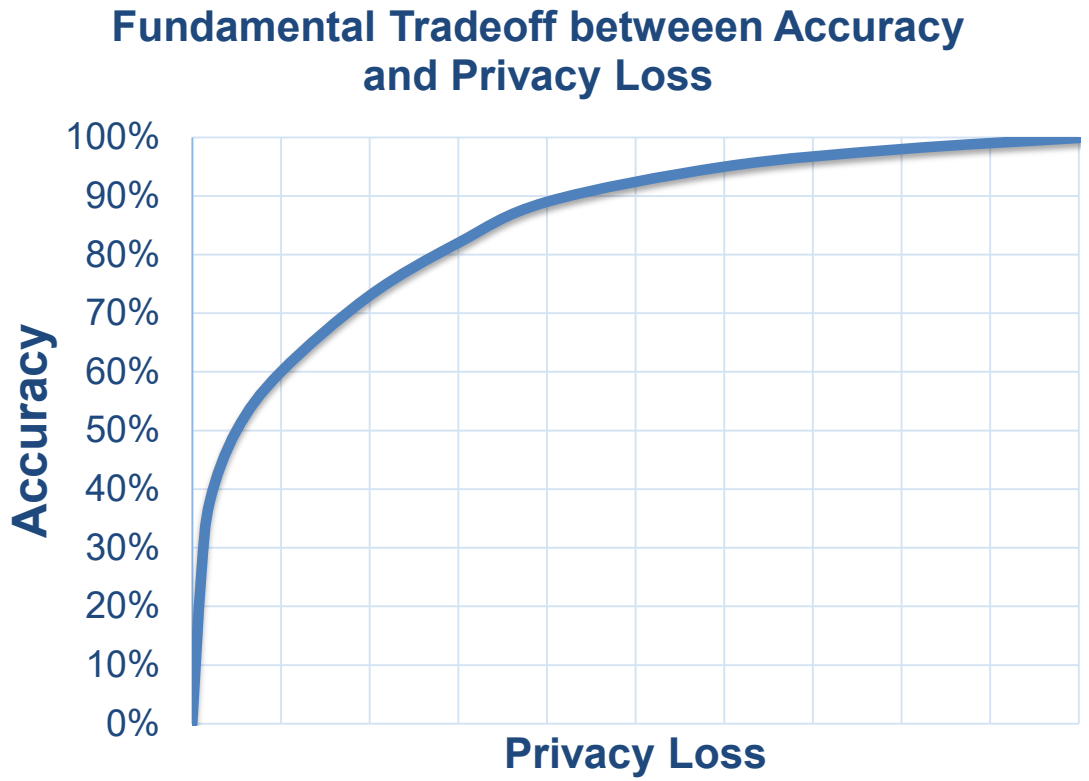
Fundamental Tradeoff betweeen Accuracy and Privacy Loss



“Differential Privacy” is the formal privacy system that we are using.

Differential privacy gives us:

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

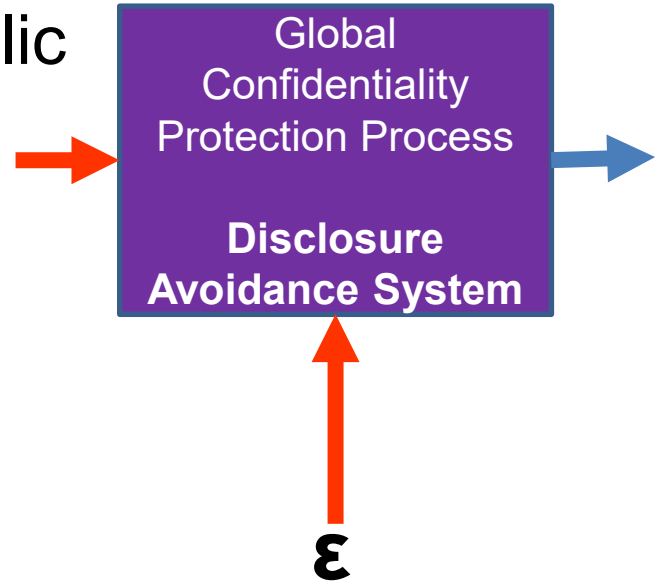


It’s called “differential privacy” because it mathematically models the privacy “differential” that each person experiences from having their data included in the Census Bureau’s data products compared to having their record deleted or replaced with an arbitrary record.

The Disclosure Avoidance System relies on injects formally private noise.

Advantages of noise injection with formal privacy:

- **Transparency:** the details can be explained to the public
- Tunable privacy guarantees
- Privacy guarantees do not depend on external data
- Protects against accurate database reconstruction
- Protects every member of the population



Challenges:

- Entire country must be processed at once for best accuracy
- Every use of confidential data must be tallied in the *privacy-loss budget*

2. What is the U.S. Census Bureau's messaging to explain differential privacy to the public?

What is the U.S. Census Bureau's messaging to explain differential privacy to the public?

- Differential privacy requires that *statistical noise* be added to *every data product from a data set*.
- The noise will balance the requirements for accuracy and confidentiality.
- Some data products will be *more accurate* with differential privacy than they were with swapping.

Consider a census block:

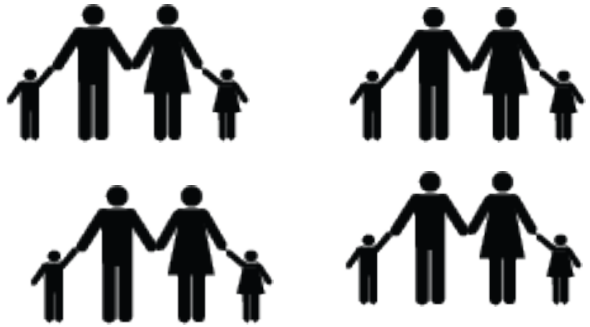
As collected:

As Reported

	Male	Female
Age < 18	10	20
Age >= 18	40	30

Consider a census block:

As collected:



	Male	Female
Age < 18	4	4
Age >= 18	4	4

Pop=16

	Male	Female
Age < 18	5	0
Age >= 18	2	8

More accurate sex distribution

Pop=15

	Male	Female
Age < 18	6	2
Age >= 18	2	7

More accurate age distribution

Pop=17

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

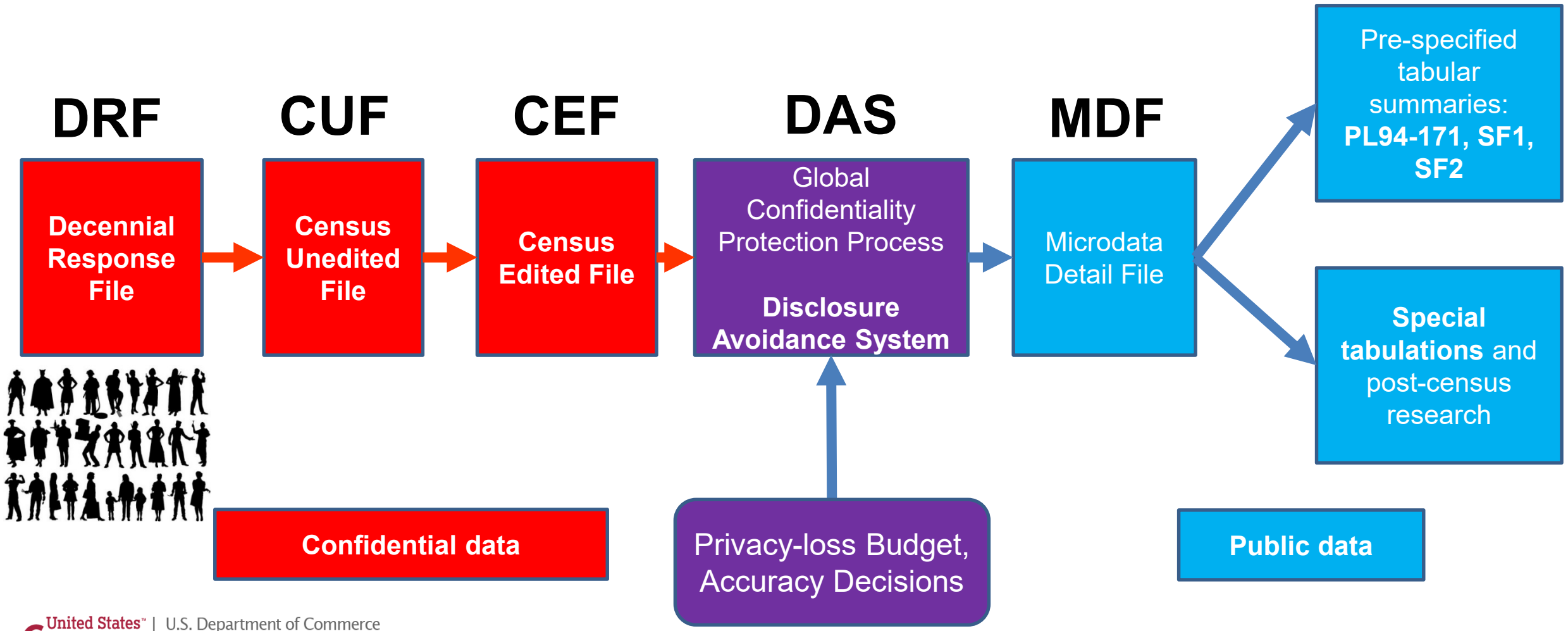
We had to create a new system that:

- Produced higher-quality statistics at more densely populated geographies
- Produced consistent tables

We created new differential privacy algorithms and processing systems that:

- Produce highly accurate statistics for large populations (e.g. states, counties)
- Create privatized microdata that can be used for any tabulation without additional privacy loss
- Fit into the decennial census production system

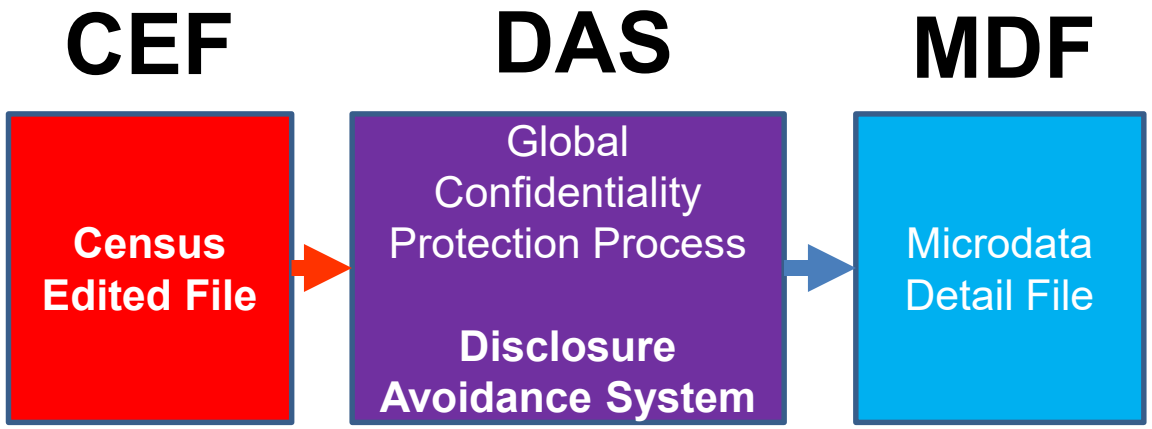
The Disclosure Avoidance System allows the Census Bureau to enforce global confidentiality protections.



3. How does the U.S. Census Bureau's differential privacy safeguard both individual and block-level data?

How does the U.S. Census Bureau's differential privacy safeguard both individual and block-level data?

Individual data is in the CEF but not in the MDF:

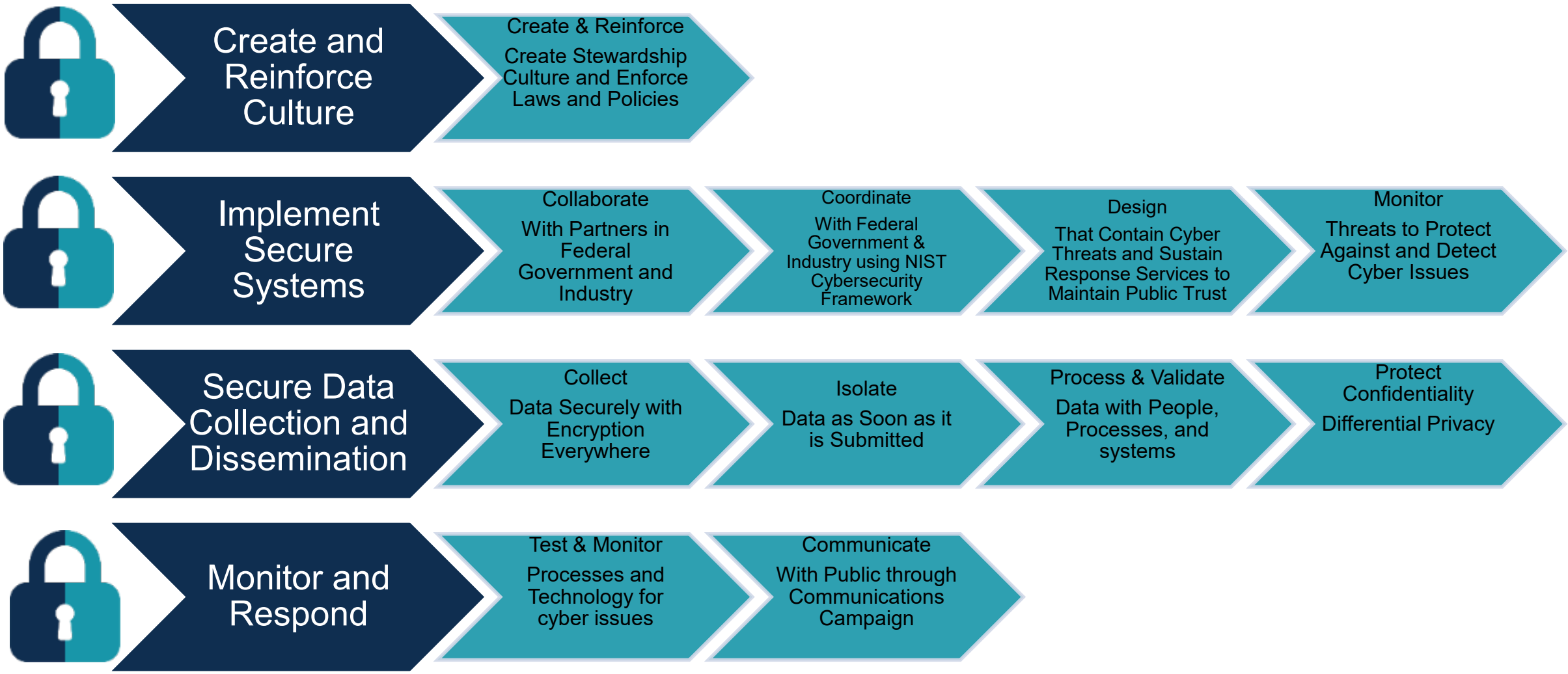


Block-level data has added noise:

- The noise makes database reconstruction much less accurate.
- The noise frustrates using block-level data to find specific individuals.

Census Cyber Security

Our Overall Approach to Maintain Public Trust



More Background on the 2020 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