Disclosure Avoidance and the 2020 Census Demographic and Housing Characteristics (DHC) File

Michael Hawes
Senior Survey Statistician for Scientific Communication
Research and Methodology Directorate
U.S. Census Bureau

April 27, 2023

United States Census Bureau

2020 Census Data Products

Released

Apportionment
April 26, 2021

Redistricting File
(Public Law 94-171)
August 12, 2021
September 16, 2021

Demographic Profile

Demographic and Housing Characteristics File (DHC)
May 25, 2023

Congressional District Summary Files
Planned August 2023

Detailed DHC-A
Planned September 2023

Detailed DHC-B
Release Date TBD

Supplemental DHC (S-DHC)
Release Date TBD

Future Effort

Public Use Microdata File

Special Tabulations

More information about the products is available on the About 2020 Census Data Products webpage.

United States Census Bureau

Apportionment Release

- Apportionment is the process of dividing the 435 memberships, or seats, in the U.S. House of Representatives among the 50 states. At the conclusion of each decennial census, the results are used to calculate the number of seats to which each state is entitled.

- **Results were released on April 26, 2021**
- **Subjects include:**
  - Resident population
  - Overseas population
  - Apportionment population
- **Geography:** 50 states, the District of Columbia (DC), and Puerto Rico
- **Disclosure avoidance:** Results do <u>not</u> undergo disclosure avoidance

United States® Census Bureau

Redistricting File (Public Law 94-171)

- Public Law 94-171 directs the Census Bureau to provide data to the governors and legislative leadership in each of the 50 states for redistricting purposes. This product is the first file released that includes demographic and housing characteristics.

- **Results were released on August 12, 2021 (Summary Files) and September 16, 2021 (data.census.gov)**
- **Subjects include:**
  - Voting age
  - Race
  - Hispanic or Latino origin
  - Housing occupancy
  - Group quarters (GQ) population by major GQ type
- **Lowest level of geography:** Census Block
- **Disclosure avoidance:** Differentially private TopDown Algorithm (TDA)

United States Census Bureau

Demographic Profile

- This product will provide select demographic and housing characteristics about local communities in a streamlined, easy to use format.

- **Expected release date:** May 2023
- **Subjects include:**
  - Sex by 5-year age groups
  - Median age by sex
  - Race
  - Hispanic or Latino origin
  - Relationship to householder
  - GQ population
  - Household type
  - Housing occupancy
  - Housing tenure
- **Lowest level of geography:** Tract
- **Disclosure avoidance:** Differentially private TDA

United States® Census Bureau

Demographic and Housing Characteristics File (DHC)

- The DHC will include many of the demographic and housing tables previously included in 2010 Summary File 1 (2010 SF1). Some tables are repeated by race and ethnicity.

- **Expected release date:** May 2023
- **Subjects include:**
  - Sex by single year-of-age
  - Hispanic or Latino origin of householder by race of householder
  - GQ population by sex by age
  - Relationship by age for population under 18 years
  - Household type by relationship and presence of people of specific ages
  - Multigenerational households
  - Family type by presence of children
  - Tenure by household size
  - Tenure by household type by age of householder
  - Vacancy Status
- **Lowest level of geography:** Varies with many tables at Census Block
- **Disclosure avoidance:** Differentially private TDA

United States Census Bureau

Detailed Demographic and Housing Characteristics File A (Detailed DHC-A)

- Detailed DHC-A includes population counts repeated by approximately 370 detailed racial and ethnic groups and 1,200 detailed American Indian and Alaska Native (AIAN) tribal and village population groups
- **Expected release date:** Sept 2023
- **Subjects are repeated by detailed racial and ethnic groups:**
  - Total population
  - Sex by Age for Selected Age Categories
- **Proposed levels of geography:** Nation, State, County, Tract, Place, AIANNH areas
- **Disclosure avoidance:** Differentially private SafeTab-P algorithm

United States® Census Bureau

Detailed Demographic and Housing Characteristics File B (Detailed DHC-B)

- Detailed DHC-B includes household counts repeated by approximately 370 detailed racial and ethnic groups and 1,200 detailed American Indian and Alaska Native (AIAN) tribal and village population groups

- **Expected release date:** TBD
- **Subjects are repeated by detailed racial and ethnic groups:**
  - Household Type
  - Tenure
- **Proposed levels of geography:** Nation, State, County, Tract, Place, AIANNH areas
- **Disclosure avoidance:** Differentially private SafeTab-H algorithm

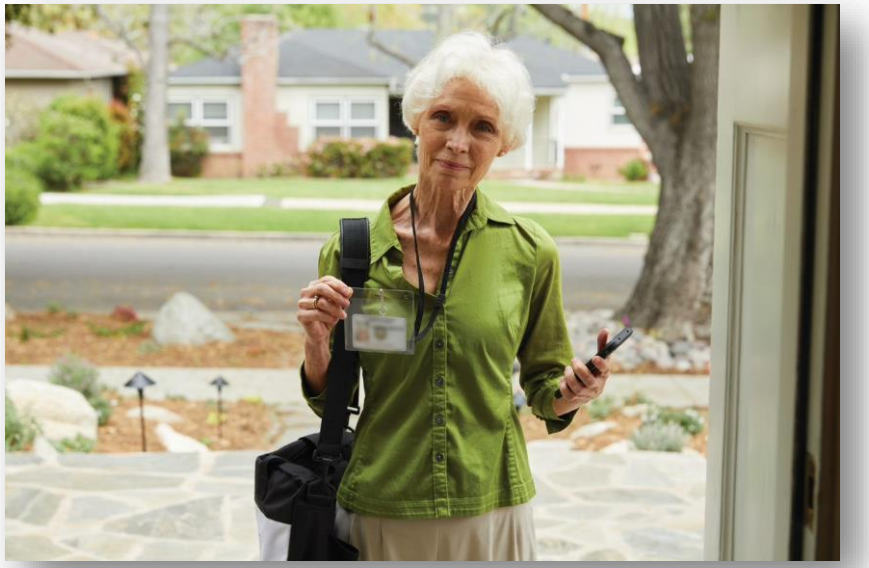
United States® Census Bureau

Background on Confidentiality Protections for the 2020 Census Data Products

United States Census Bureau

Keeping the Public's Trust: Title 13

"To stimulate public cooperation necessary for an accurate census…Congress has provided assurances that information furnished by individuals is to be treated as confidential. ***Title 13 U.S.C. §§ 8(b) and 9(a)*** *explicitly provide for nondisclosure of certain census data, and* ***no discretion is provided to the Census Bureau on whether or not to disclose such data****…"* (U.S. Supreme Court, Baldrige v. Shapiro, 1982)



To safeguard the public's confidential census responses, the Census Bureau has long employed a variety of statistical techniques to mitigate disclosure risk in our published data products.

United States® Census Bureau

Disclosure Avoidance for Past Censuses

1970-1980 Censuses

	528	
		794
	581	
137	941	189
931		
	250	
		590

SUPPRESSION

1990-2010 Censuses

668	178	779
91	8	159
809	112	811
518	424	955
989	352	765
237	411	686
77	820	590

SWAPPING

United States Census Bureau

The Ever-rising Risk of Disclosure

- Any data release carries some risk of disclosure.
- Improvements in computing power and the explosion of third-party data mean that disclosure risk has increased significantly.
- Protecting confidentiality means adapting and responding to these increasing threats

United States Census Bureau

Disclosure Avoidance for the 2020 Census

The 2020 Census improves on the noise injection methods of the 1990-2010 Censuses by employing a mathematical framework known as Differential Privacy (DP) to assess and quantify disclosure risk and confidentiality protection.

Every individual that is reflected in a particular statistic contributes towards that statistic's value.

Every statistic that you publish "leaks" a small amount of private information.

DP as a framework allows you to assess each individual's contribution to the statistic, and to measure (and thus, limit) how much information about them will leak.

United States Census Bureau

The 2020 Census Disclosure Avoidance System (DAS)

TopDown Algorithm (TDA)

Produces privacy-protected microdata (Microdata Detail File) that is ingested by Decennial tabulation system

- Redistricting Data (P.L. 94-171) Summary File
- Demographic Profile
- Demographic and Housing Characteristics File (DHC)
- Congressional District Summary Files

SafeTab PHSafe

Produce privacy-protected tabulations

- Detailed DHC-A
- Detailed DHC-B
- Supplemental DHC

United States Census Bureau

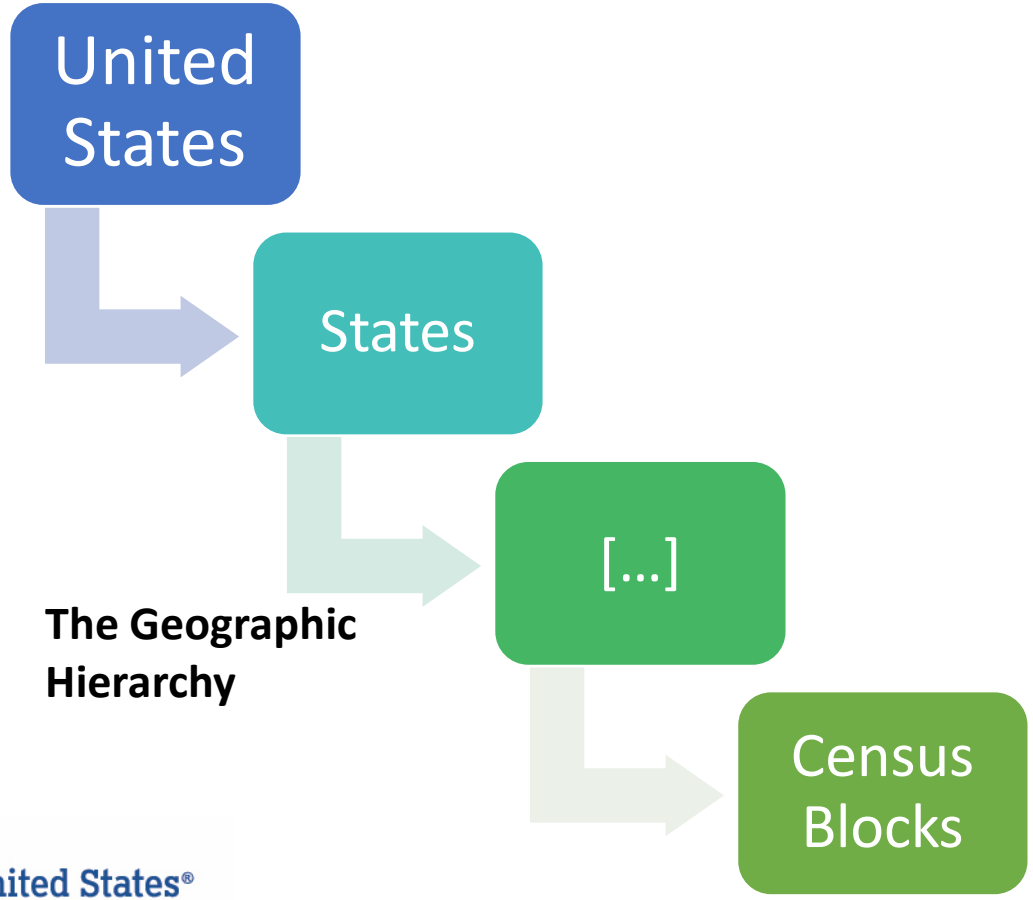
The TopDown Algorithm

*A histogram, in this context, is a tabular representation of the microdata with counts of records for each possible combination of values for each attribute in the microdata.

For complete details see: Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., & Zhuravlev, P. (2022). The 2020 Census Disclosure Avoidance System TopDown Algorithm. *Harvard Data Science Review*. (June) https://doi.org/10.1162/99608f92.529e3cb9

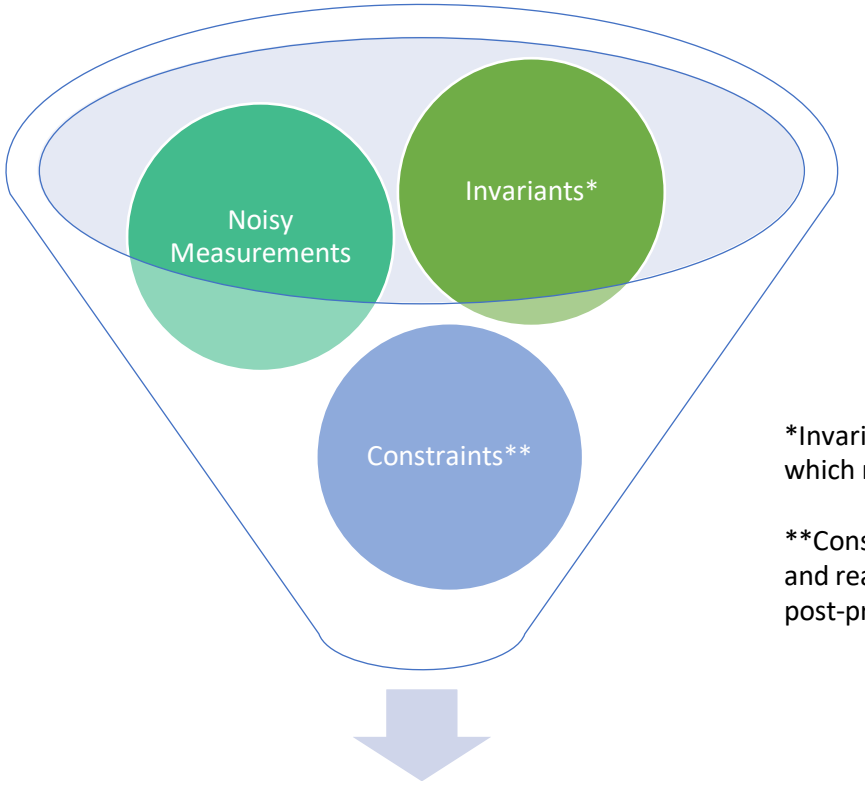
United States Census Bureau

The TopDown Algorithm



The Geographic Hierarchy

At each geographic level:



*Invariants are counts to which no noise is added.

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

United States Census Bureau

Queries and Privacy-loss Budget Allocation

Global *rho*	2.56
Global *epsilon*	17.90
delta	10^{-10}

	***rho* Allocation by Geographic Level**
US	2.54%
State	35.13%
County	10.91%
Tract	16.76%
Optimized Block Group*	30.64%
Block	4.03%

Production settings for the 2020 Census Redistricting Data (P.L. 94-171) Summary File (Persons tables P1-P5)

Query	**Per Query *rho* Allocation by Geographic Level**					
	US	State	County	Tract	Optimized Block Group*	Block
TOTAL (1 cell)		32.35%	8.32%	6.40%	12.75%	0.00%
CENRACE (63 cells)	0.03%	0.05%	0.03%	0.03%	0.02%	0.01%
HISPANIC (2 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
VOTINGAGE (2 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
HHINSTLEVELS (3 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
HHGQ (8 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
HISPANIC*CENRACE (126 cells)	0.08%	0.10%	0.07%	7.90%	7.89%	0.02%
VOTINGAGE*CENRACE (126 cells)	0.08%	0.10%	0.07%	0.08%	0.07%	0.02%
VOTINGAGE*HISPANIC (4 cells)	0.02%	0.05%	0.03%	0.02%	0.02%	0.00%
VOTINGAGE*HISPANIC*CENRACE (252 cells)	0.27%	0.29%	0.27%	0.27%	0.18%	0.07%
HHGQ*VOTINGAGE* HISPANIC*CENRACE (2,016 cells)	1.99%	1.97%	2.01%	1.97%	9.63%	3.88%

United States® Census Bureau

Overview of the 2010 DDPS

United States Census Bureau

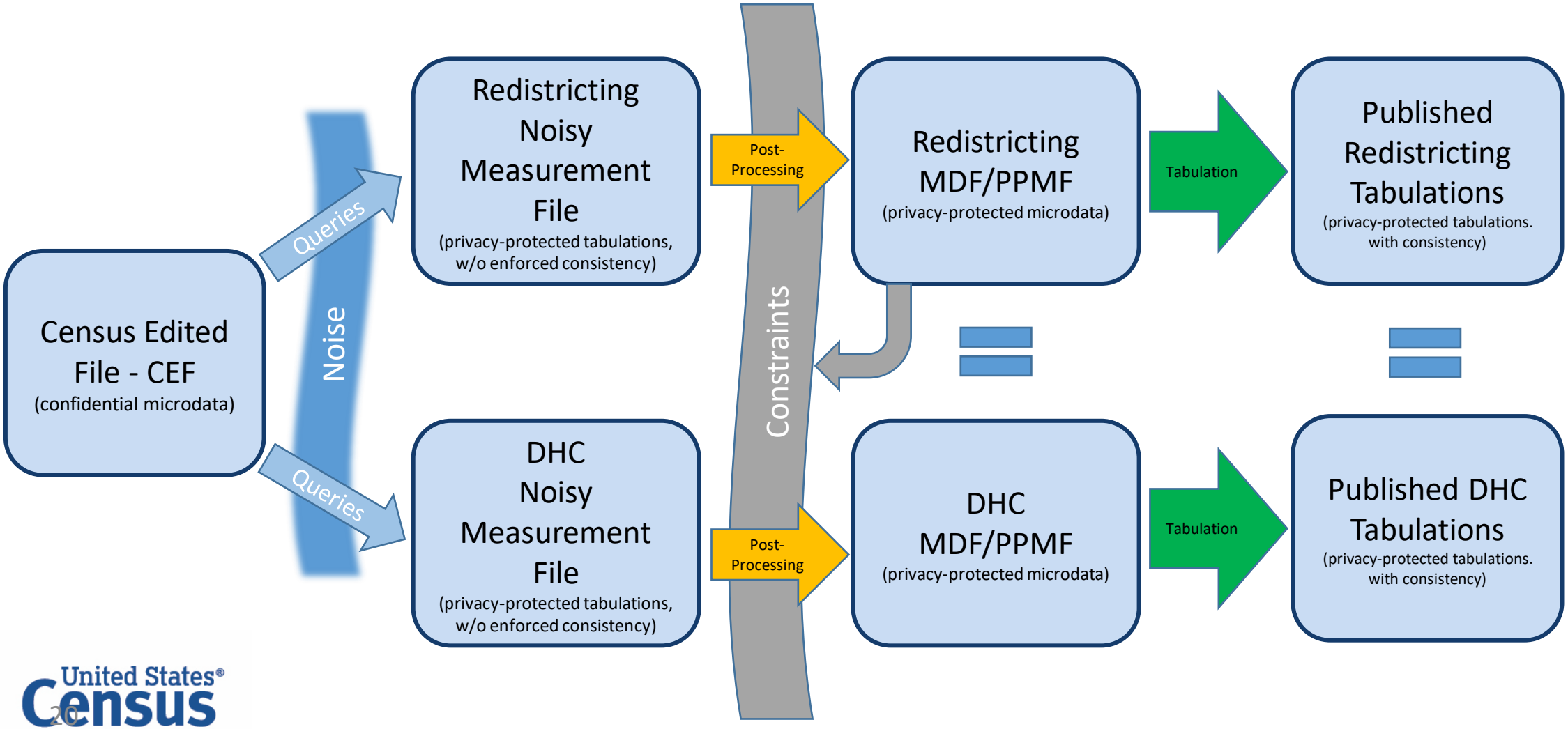
Components of the 2010 Demonstration Data Products Suite – Redistricting and Demographic and Housing Characteristics File – Production Settings (2023-04-03) **(2010 DDPS)**

(2010 Census data processed through the 2020 DAS at production settings)

- 2010 DDPS Fact Sheet
- Detailed Summary Metrics (and Metrics Overview)
- Privacy-Protected Microdata File (PPMF)
- DHC Tabulations (via IPUMS)
- Privacy-loss Budget (PLB) Allocations
- Noisy Measurement File (NMF)

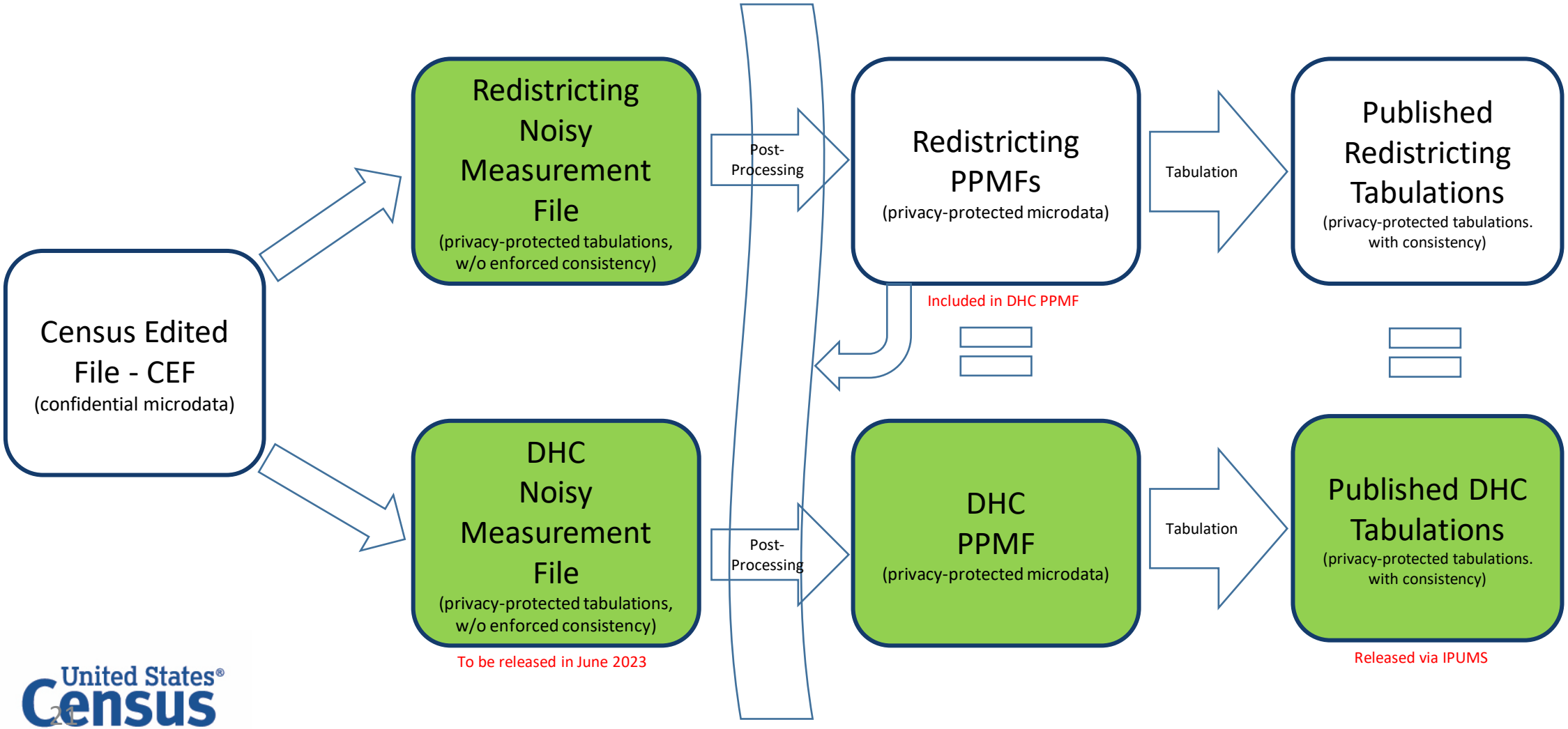
United States Census Bureau

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



United States Census Bureau

2010 DDPS NMFs, PPMF, Tabulations

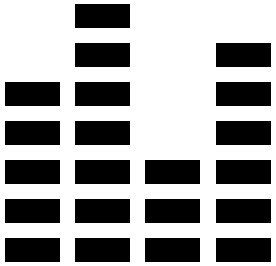


Should I Use the NMF, the PPMF, or the Tabulations?

- There are two sources of error in the published statistics (PPMF and Tabulations):

Differentially private noise

- Unbiased
- Known distribution
- Reflected in the noisy measurements

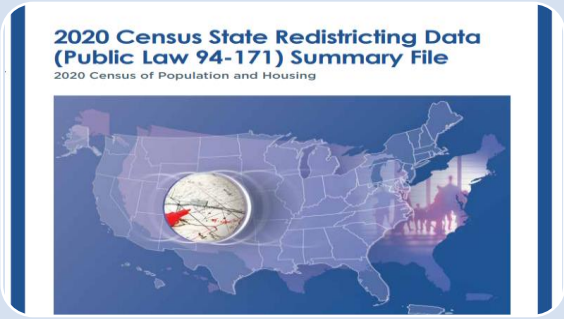


Post-processing

- Data dependent
  - While the nonnegativity requirement decreases error in the detailed cell counts, it also introduces a positive bias in small counts and an offsetting negative bias in large counts.
  - TDA also reduces the amount of error for many statistics relative to their corresponding noisy measurements.
- Block-level statistics will often have a lower expected variation than you would expect based solely on the amount of PLB assigned to that query at the block level.

United States Census Bureau

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

United States Census Bureau

New Resources for Data Users

United States® Census Bureau

Reader-Friendly Disclosure Avoidance Briefs

- Disclosure Avoidance and the 2020 Redistricting Data
- Why the Census Bureau Chose Differential Privacy
- Disclosure Avoidance and the 2020 Census: How the TopDown Algorithm Works

More resources are in development, as well as additional specific guidance and training for using the 2020 Census data.

United States Census Bureau

Coming Soon

- Guidance and examples on how to use the NMF to calculate unbiased estimates and confidence intervals.
- Subscribe to our newsletter to receive the announcement and related webinar info.

United States Census Bureau

Questions?

Or send them to 2020DAS@census.gov

United States Census Bureau