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

Michael Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology Directorate
U.S. Census Bureau

October 19, 2022

Shape your future START HERE >

United States® Census 2020

Acknowledgements

This presentation includes work by the Census Bureau's 2020 Disclosure Avoidance System development team, Census Bureau colleagues, and our collaborators, from the following Census Bureau divisions and outside organizations: ADCOM, ADDC, ADRM, CED, CEDDA, CEDSCI, CES, CSRM, DCMD, DITD, ESMD, GEO, POP, TAB, CDF, Econometrica Inc., Galois, Knexus Research Corp, MITRE, NLT, TI, and Tumult Labs.

We also acknowledge and greatly appreciate the ongoing feedback we have received from external stakeholder groups that has contributed to the design and improvement of the Disclosure Avoidance System.

For more information and technical details relating to the issues discussed in these slides, please contact the author at michael.b.hawes@census.gov.

Any opinions and viewpoints expressed in this presentation are the author's own, and do not represent the opinions or viewpoints of the U.S. Census Bureau.

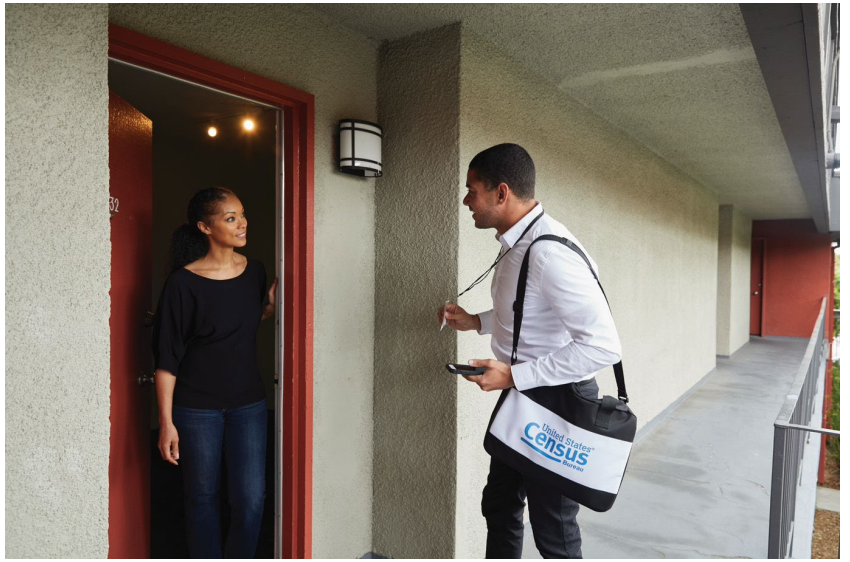
Shape your future START HERE >

United States Census 2020

Our Commitment to Privacy and Confidentiality

Data stewardship is central to the Census Bureau's mission to produce high-quality statistics about the people and economy of the United States.

Our commitment to protect the privacy of our respondents and the confidentiality of their data is both a legal obligation and a core component of our institutional culture.



Shape your future
START HERE >
United States Census 2020

The Census Bureau's Privacy Protections Over Time

Throughout its history, the Census Bureau has been at the forefront of the design and implementation of statistical methods to safeguard respondent data.

Over the decades, as we have increased the number and detail of the data products we release, so too have we improved the statistical techniques we use to protect those data.

Shape your future START HERE >
United States Census 2020

The Privacy Challenge

Every time you release any statistic calculated from a confidential data source you “leak” a small amount of private information.

If you release too many statistics, too accurately, you will eventually reveal the entire underlying confidential data source.

Dinur, Irit and Kobbi Nissim (2003) “Revealing Information while Preserving Privacy” PODS, June 9-12, 2003, San Diego, CA

Shape your future START HERE >

United States® Census 2020

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

Re-identification

Linking public data to external data sources to re-identify specific individuals within the data.

Name	Block	Age	Sex		Block	Age	Sex	Race	Relationship
Jane Smith	1234	66	Female	+	1234	66	Female	Black	Married
Joe Public	1234	84	Male		1234	84	Male	Black	Married
John Citizen	1234	30	Male		1234	30	Male	White	Married

External Data

Confidential Data

Shape your future START HERE >

United States Census 2020

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

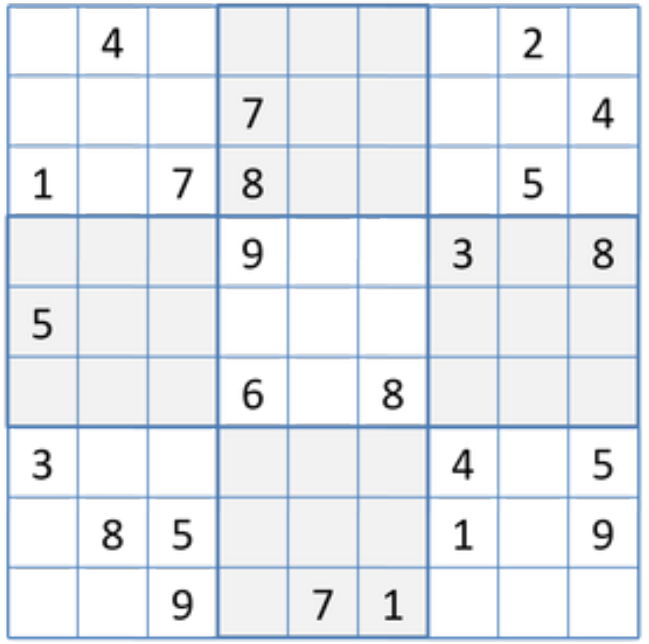
United States® Census Bureau

Reconstruction

The recreation of individual-level data from tabular or aggregate data.

If you release enough tables or statistics, eventually there will be a unique solution for what the underlying individual-level data were.

Computer algorithms can do this very easily.



	4						2	
			7					4
1		7	8				5	
			9			3		8
5								
			6		8			
3						4		5
	8	5				1		9
		9		7	1			

Shape your future START HERE >

United States Census 2020

Reconstructing the 2010 Census

- The 2010 Census collected information on the location, age, sex, race, ethnicity, and relationship (to householder) status for ~309 Million individuals. (1.9 Billion confidential data points)
- The 2010 Census data products released over 150 billion statistics
- We conducted an internal experiment to see if we could reconstruct and re-identify the 2010 Census records.

Shape your future START HERE >
United States Census 2020

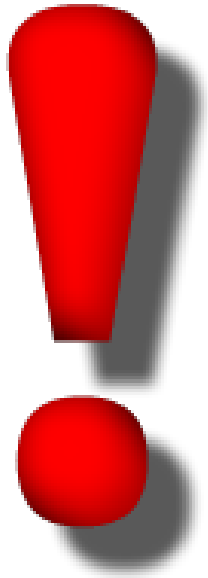
Reconstructing the 2010 Census: What Did We Find?

1. On the 309 million reconstructed records, census block and voting age (18+) were correctly reconstructed for all individuals in all 6,207,027 inhabited blocks.
2. Block, sex, age (in years), race (OMB 63 categories), and ethnicity were reconstructed:
    1. Exactly for 46% of the population (142 million individuals)
    2. Within +/- one year for 71% of the population (219 million individuals)
3. Block, sex, and age were then linked to commercial data, which provided presumed re-identification of 45% of the population (138 million individuals).
4. Name, block, sex, age, race, ethnicity were then compared to the confidential data, which yielded confirmed re-identifications for 38% of the presumed re-identifications (52 million individuals).
5. For the confirmed re-identifications, race and ethnicity are learned correctly, though the attacker may still have uncertainty.

The reconstructed microdata are a close approximation of the Hundred Percent Detail (HDF) file and violate the disclosure avoidance rules for microdata in place for the 2010 Census.

The Census Bureau's Decision

- Advances in computing power and the availability of external data sources make database reconstruction and re-identification increasingly likely.
- The Census Bureau recognized that its traditional disclosure avoidance methods are increasingly insufficient to counter these risks.
- To meet its continuing obligations to safeguard respondent information, the Census Bureau has committed to modernizing its approach to privacy protections.



Shape your future START HERE >
United States Census 2020

Disclosure Avoidance

Disclosure avoidance methods seek to make reconstruction and re-identification more difficult, by:

- Reducing precision
- Removing vulnerable records, or
- Adding uncertainty

Commonly used (legacy) methods include:

- Primary/complementary suppression
- Rounding
- Top/bottom coding of extreme values
- Sampling
- Record swapping
- Noise injection

Shape your future START HERE >
United States Census 2020

Problem #1 – Impact on Data

All statistical techniques to protect privacy impose a tradeoff between the degree of privacy protection and the resulting accuracy of the data.

Swap rates, noise injection parameters, cell suppression thresholds, etc. determine this tradeoff.

Shape your future START HERE >

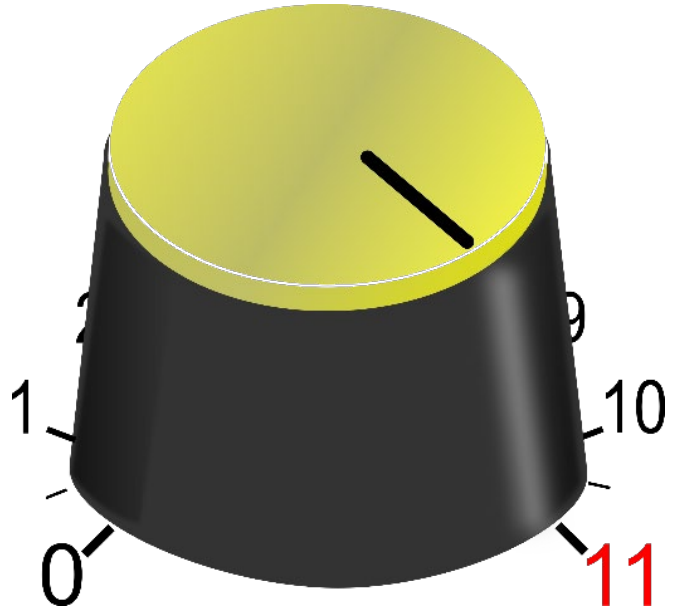
United States Census 2020

Problem #2 – How much is enough?

Traditional disclosure avoidance methods provide little ability to quantify privacy protections, *especially across multiple data releases from the same confidential source.*

When faced with rising disclosure risk, disclosure avoidance practitioners adjust their implementation parameters.

BUT, this is largely a scattershot solution that over-protects some data, while often under-protecting the most vulnerable records.



Shape your future START HERE >
United States Census 2020

Differential Privacy

DP is not a disclosure avoidance "method" as much as it is a framework for defining and then quantifying 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.

Shape your future START HERE >
United States Census 2020

Differential Privacy

When combined with noise injection, DP allows you to precisely control the amount of private information leakage in your published statistics.

- Infinitely tunable – parameter "dials" can be set anywhere from perfect privacy to perfect accuracy.
- Privacy guarantee is mathematically provable and future-proof.
- The precise calibration of statistical noise enables optimal data accuracy for any given level of privacy protection.*

*Absent post-processing requirements, which can introduce error independent of that needed to protect privacy.

Shape your future START HERE >
United States® Census 2020

2020 Census Data Products

"Group I Products"

- P.L. 94-171 Redistricting Data Summary File
- Demographic Profile
- Demographic and Housing Characteristics File

"Group II Products"

- Detailed Demographic and Housing Characteristics Files (Detailed DHC-A, Detailed DHC-B, Supplemental DHC)

"Group III Products"

TBD, may include:

- Public Use Microdata
- Special Tabulations
- FSRDC Access
- Out-year uses of 2020 Census data

Shape your future START HERE >

United States Census 2020

Components of the 2020 Census Disclosure Avoidance System (DAS)

"Group I Products"

TopDown Algorithm (TDA)

Produces privacy-protected microdata (Microdata Detail File) that can be ingested by Decennial tabulation systems

"Group II Products"

SafeTab
PHSafe

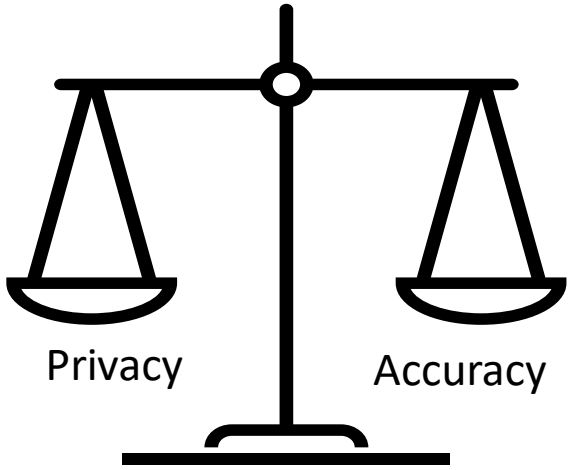
Produce privacy-protected tabulations directly

"Group III Products"

TDA
SafeTab
PHSafe
or other formally privacy solutions

Shape your future START HERE >
United States Census 2020

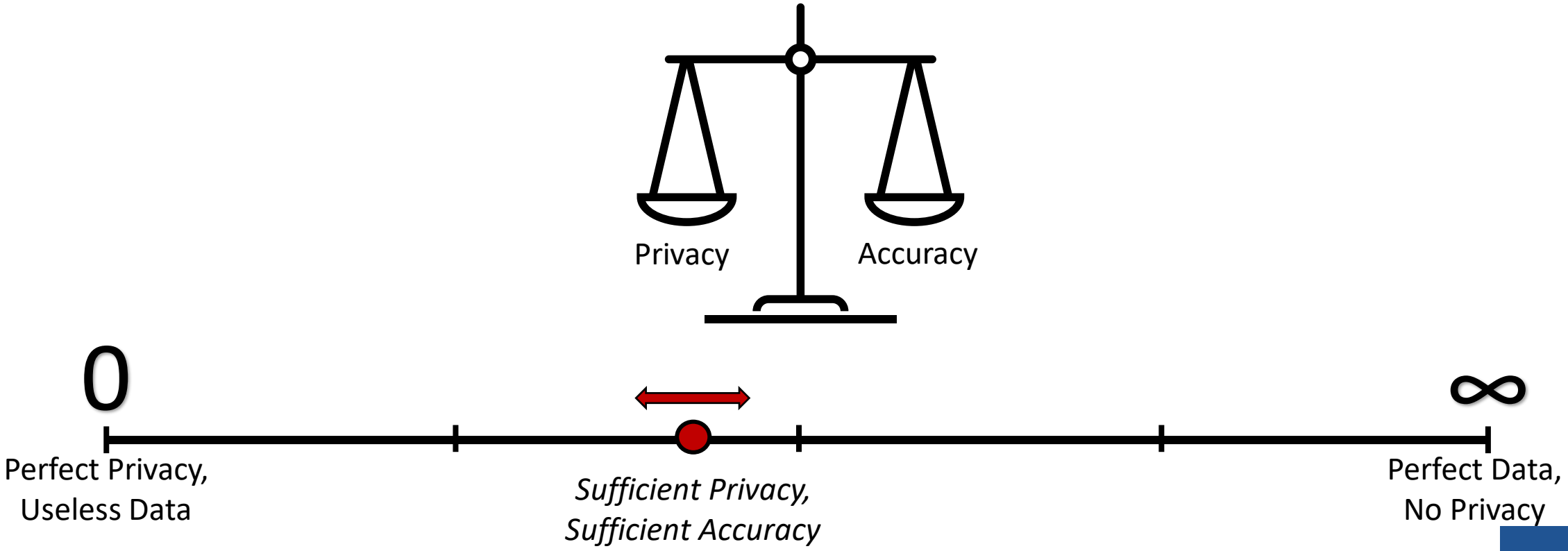
What is a Privacy-loss Budget?



Any disclosure avoidance mechanism imposes a fundamental tradeoff between data protection (privacy/confidentiality) and data accuracy/fitness-for-use.

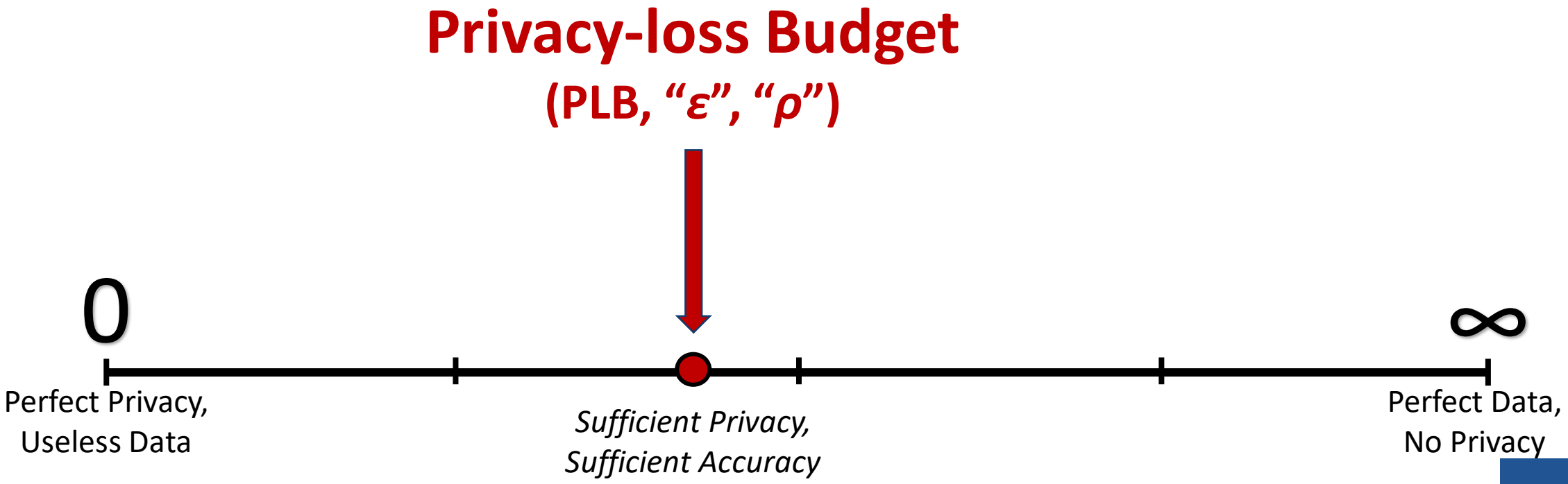
Shape your future START HERE >
United States Census 2020

What is a Privacy-loss Budget?



Shape your future START HERE >
United States Census 2020

What is a Privacy-loss Budget?



Shape your future START HERE >
United States Census 2020

Allocating Privacy Loss Budget (PLB) by Data Product

PL and DHC products were split apart and will be protected using separate privacy loss budgets

"Group I PL Product"

"Group I DHC Product"

"Group II Product"

"Group III Products"

PLB_1 + PLB_2 + PLB_3 + PLB_4

= Global PLB

Shape your future START HERE >

United States Census 2020

Allocating PLB *within* Data Products

	rho Allocation by Geographic Level
US	104/4099
State	1440/4099
County	447/4099
Tract	687/4099
Optimized Block Group*	1256/4099
Block	165/4099

Query	Per Query *rho* Allocation by Geographic Level					
	US	State	County	Tract	Optimized Block Group*	Block
TOTAL (1 cell)		3773/4097	3126/4097	1567/4102	1705/4099	5/4097
CENRACE (63 cells)	52/4097	6/4097	10/4097	4/2051	3/4099	9/4097
HISPANIC (2 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
VOTINGAGE (2 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HHINSTLEVELS (3 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HHGQ (8 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HISPANIC*CENRACE (126 cells)	130/4097	12/4097	28/4097	1933/4102	1055/4099	21/4097
VOTINGAGE*CENRACE (126 cells)	130/4097	12/4097	28/4097	10/2051	9/4099	21/4097
VOTINGAGE*HISPANIC (4 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
VOTINGAGE*HISPANIC*CENRACE (252 cells)	26/241	2/241	101/4097	67/4102	24/4099	71/4097
HHGQ*VOTINGAGE* HISPANIC*CENRACE (2,016 cells)	189/241	230/4097	754/4097	241/2051	1288/4099	3945/4097

Within each data product, shares of the PLB are then allocated to each statistic calculated.

The PLB shares for each element are set through tuning informed by engagement with our data users.

Shape your future START HERE >

United States Census 2020

Working with our stakeholders

Over the past three years, the Census Bureau has released eight sets of demonstration data products generated by running 2010 Census data through the 2020 Census Disclosure Avoidance System.

This allowed our data users to compare the differentially private results to the results published after the 2010 Census (which were protected using "swapping").

The feedback and analyses we received have helped us to tune and adjust our algorithms to ensure that the resulting data will meet our data users' needs.

Shape your future START HERE >

United States® Census 2020

Ongoing Engagement

- Ongoing external engagements with advisory and stakeholder groups, such as:
  - Census Scientific Advisory Committee (CSAC) and National Advisory Committee on Race (NAC)
  - CSAC and NAC Differential Privacy (DP) Working Groups
  - American Indian and Alaska Native Tribal Leaders
  - Committee on National Statistics (CNSTAT)
  - State Data Center (SDC) and Census Information Center (CIC) networks
  - Federal agency partners
  - Congressional committees and staff
  - And more …
- Internal engagements such as Town Halls and launch of Share Point site

Shape your future START HERE >
United States Census 2020

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)
Planned May 2023

Detailed DHC-A
Planned Aug 2023

Detailed DHC-B
Release Date TBD

Supplemental DHC (S-DHC)
Release Date TBD

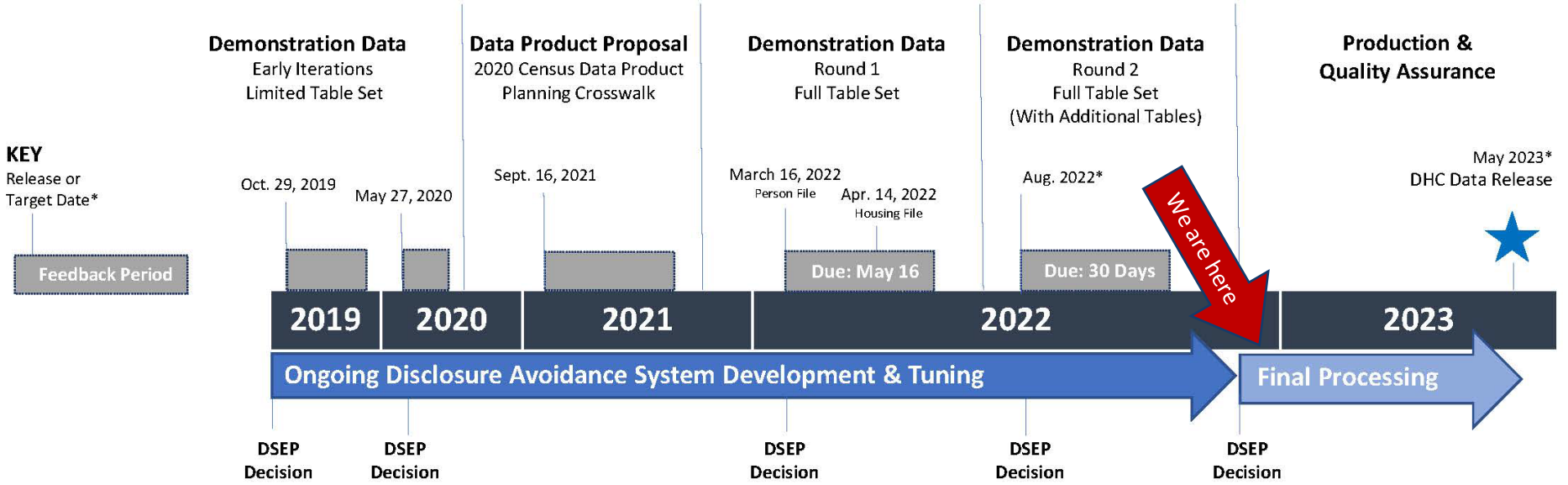
Future Effort

Public Use Microdata Sample (PUMS) File

Special Tabulations

Shape your future START HERE >
United States Census 2020

Demographic and Housing Characteristics File (DHC) Development and Production Timeline



United States® Census Bureau

Demonstration Data

The demonstration data apply the latest version of the disclosure avoidance system (DAS) to 2010 Census data. This allows a side-by-side comparison of DAS impact, quantified by detailed summary metrics. All demonstration data include the person and housing files, but Round 1 split these files into two releases.

DSEP Decision Points

The Data Stewardship Executive Policy Committee (DSEP), composed exclusively of senior career officials, determines the allocation of privacy-loss budget across geography types and queries based on an analysis of data quality and disclosure risk. DSEP makes these determinations prior to every demonstration data release and prior to final release.

Feedback Periods and Due Dates

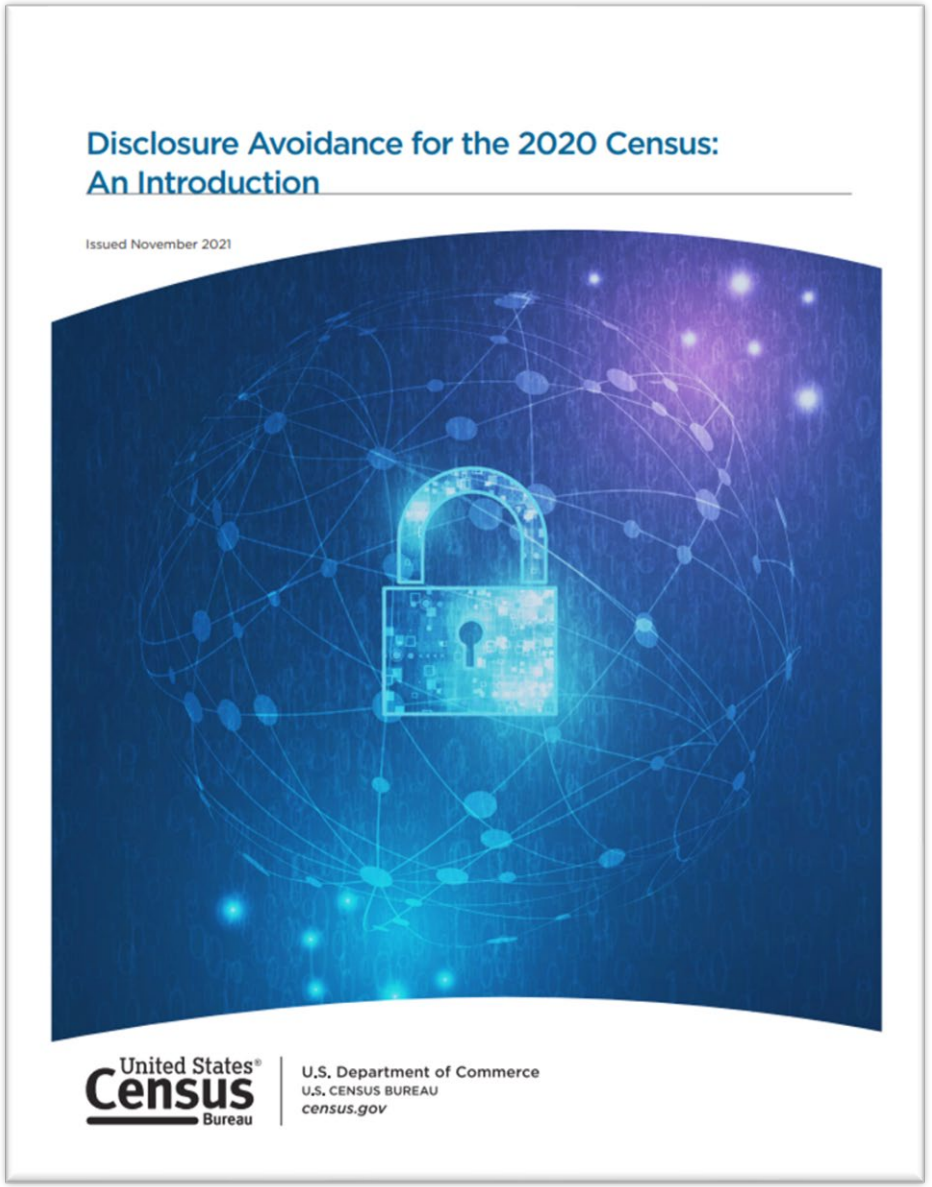
Data users have at least 30 days to evaluate and submit data feedback to help inform further development. Upcoming feedback deadlines are marked.

Read our blog to learn more:

"2020 Census Data Products: Next Steps for Data Releases"

Available at:

https://www.census.gov/library/publications/2021/decennial/2020-census-disclosure-avoidance-handbook.html



Shape your future START HERE >
United States Census 2020

Additional Guidance for Data Users

Topics for Briefs 1-3

Brief #1 (Target release - Winter 2022)

A summary of key points from the previously released Redistricting Data (P.L. 94-171) Summary File handbook.

Brief #2 (Target release - Spring 2023)

Explain how the Census Bureau selected differential privacy over other disclosure avoidance systems.

Brief #3 (Target release - Spring 2023)

Describe the disclosure avoidance system and framework including:

- A focus on the TopDown Algorithm (TDA).
- A concise definition of differential privacy.
- The information that was collected in 2020, what is published, and the published data's risk of disclosure.
- Plus, a “stay tuned” for the next brief on DHC and Demographic Profile.

United States® Census Bureau

Additional Guidance for Data Users

Topics for Briefs 4-6

Brief #4 (Target release - Spring 2023)

DHC – relevant content, including:

- Explain some of the ways stakeholders use DHC data generally.
- Discuss use cases.
- Demonstrate the potential impact of disclosure avoidance methods on various priority use cases/case studies for DHC.

Brief #5 (Target release - Summer 2023)

Detailed DAS methodology around Detailed DHC-A Product with a focus on other differentially private disclosure avoidance algorithms (non-TDA).

Brief #6 (Target release - Winter 2023)

Explain the Total Uncertainty framework including:

- Highlights from the variability paper.
- The degree of error that's introduced with noise infusion.
- Describe the multiple sources of error with examples for each.

United States® Census Bureau

Additional Guidance for Data Users

Topics for Briefs 7-8

Brief #7 (Target release - TBD)

Detailed DAS methodology around Detailed DHC-B Product with a focus on other differentially private disclosure avoidance algorithms (non-TDA).

Brief #8 (Target release - TBD)

Detailed DAS methodology around Supplemental DHC Product with a focus on other differentially private disclosure avoidance algorithms (non-TDA).

United States® Census Bureau

Stay Informed:

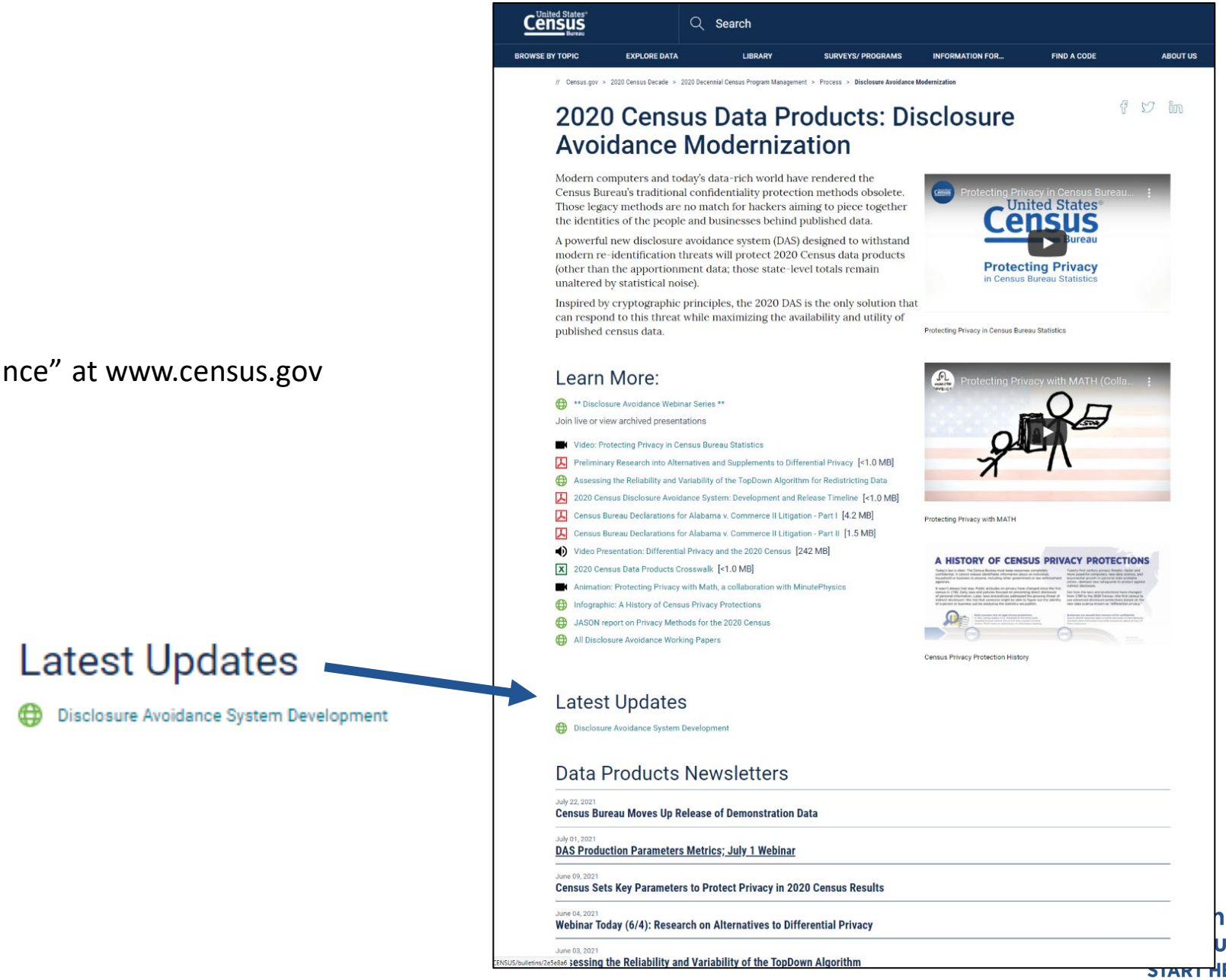
Subscribe to the 2020 Census Data Products Newsletters

*Search “Disclosure Avoidance” at www.census.gov

Stay Informed:

Visit Our Website

*Search “Disclosure Avoidance” at www.census.gov

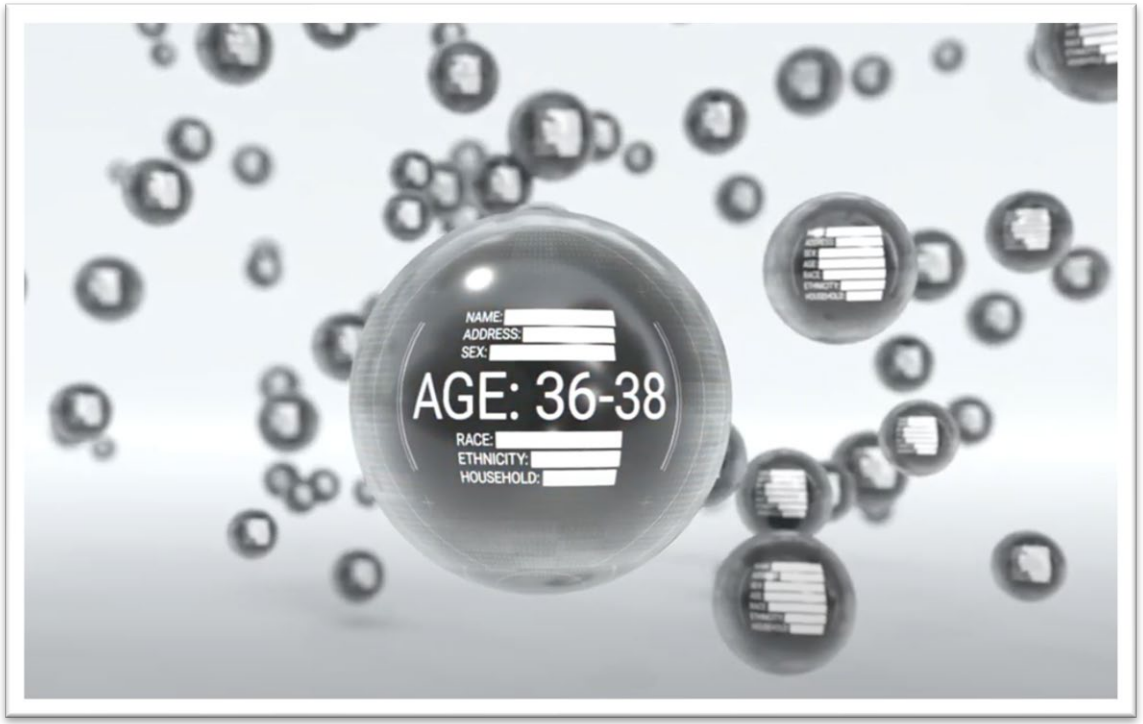


nape
uture
START HERE >
United States Census 2020

Protecting Privacy in Census Bureau Statistics

*Find it on our website and YouTube Page

Search “Disclosure Avoidance” at www.census.gov



Shape your future START HERE >
United States Census 2020

Questions

Send questions and feedback to
2020DAS@census.gov

Shape your future START HERE >
United States Census 2020