

Understanding the 2020 Census Disclosure Avoidance System:

Simulated Reconstruction-Abetted Re-identification Attack on the 2010 Census

Michael Hawes

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

May 7, 2021

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, including: John Abowd, Tammy Adams, Robert Ashmead, Craig Corl, Ryan Cummings, Jason Devine, John Fattaleh, Simson Garfinkel, Nathan Goldschlag, Michael Hawes, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Kyle Irimata, Dan Kifer, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Claudia Molinar, Brett Moran, Ned Porter, Sarah Powazek, Vikram Rao, Chris Rivers, Anne Ross, Ian Schmutte, William Sexton, Rob Sienkiewicz, Matthew Spence, Tori Velkoff, Lars Vilhuber, Bei Wang, Tommy Wright, Bill Yates, Rolando Rodriguez, and Pavel Zhuravlev.

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.

”[Computer scientists] have demonstrated that they can often ‘reidentify’ or ‘deanonymize’ individuals hidden in anonymized data with astonishing ease.”

- [Professor Paul Ohm](#) (2010)

“There is significant evidence demonstrating that technological advances and the ability to combine disparate pieces of data can lead to identification of a consumer...even if the individual pieces of data do not constitute PII. Moreover, not only is it possible to re-identify non-PII data through various means, businesses have strong incentives to actually do so.”

- [Federal Trade Commission](#) (2012)

“Anonymization...is not robust against near-term future re-identification methods.”

- [President’s Council of Advisors on Science and Technology](#) (2014)

The emerging threat of reconstruction-abetted re-identification attacks

Traditional assessments of disclosure risk have focused on re-identification from microdata products, or simple deduction/subtraction attacks on tabular data products.

Advances in computing power and the availability of powerful optimization algorithms have introduced a new vector of privacy attacks: reconstruction-abetted re-identification attacks on tabular data.

The Database Reconstruction Theorem

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



Reconstruction

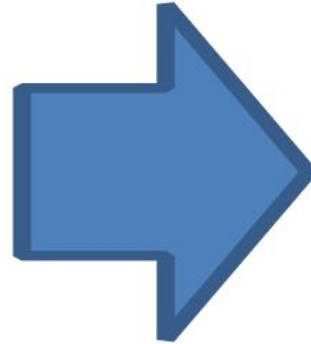
The recreation of individual-level data from tabular or aggregate data.

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			

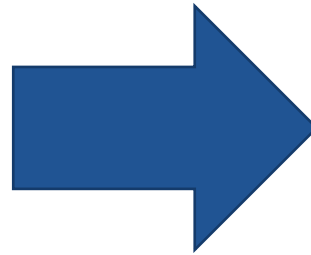
Reconstruction: A Toy Example



Block 1234	Count	Median Age	Mean Age
Total	7	30	38
Female	4	30	33.5
Male	3	30	44
Black	4	51	48.5
White	3	24	24
Married	4	51	54
Black Female	3	36	36.7

Reconstruction: An Example

Block 1234	Count	Median Age	Mean Age
Total	7	30	38
Female	4	30	33.5
Male	3	30	44
Black	4	51	48.5
White	3	24	24
Married	4	51	54
Black Female	3	36	36.7



Block	Age	Sex	Race	Relationship
1234	66	Female	Black	Married
1234	84	Male	Black	Married
1234	30	Male	White	Married
1234	36	Female	Black	Married
1234	8	Female	Black	Single
1234	18	Male	White	Single
1234	24	Female	White	Single

This table can be expressed by 164 equations.
Solving those equations takes 0.2 seconds on a 2013
MacBook Pro.

Re-identification

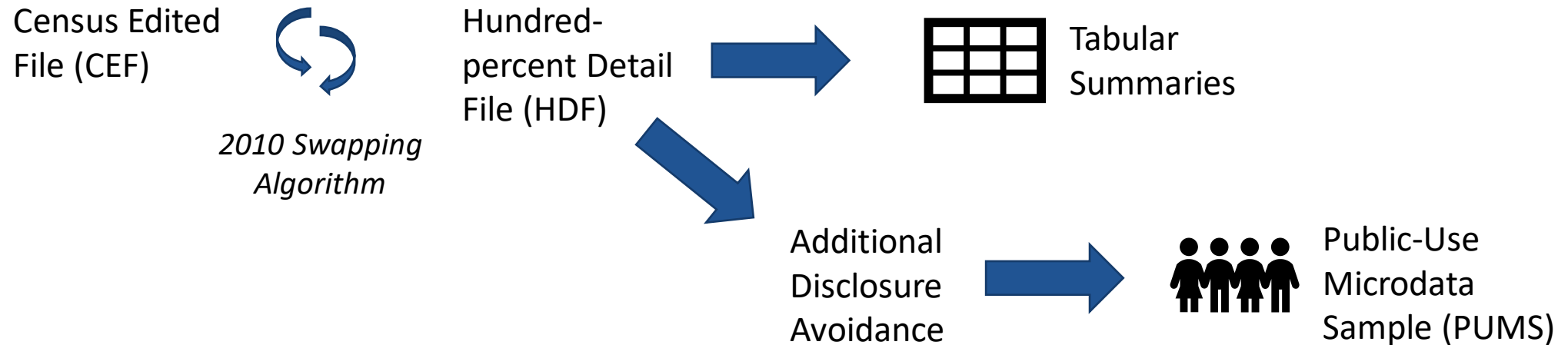
Linking public data to external data sources to re-identify specific individuals within the data.

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

Disclosure Avoidance Methods for the 2010 Census

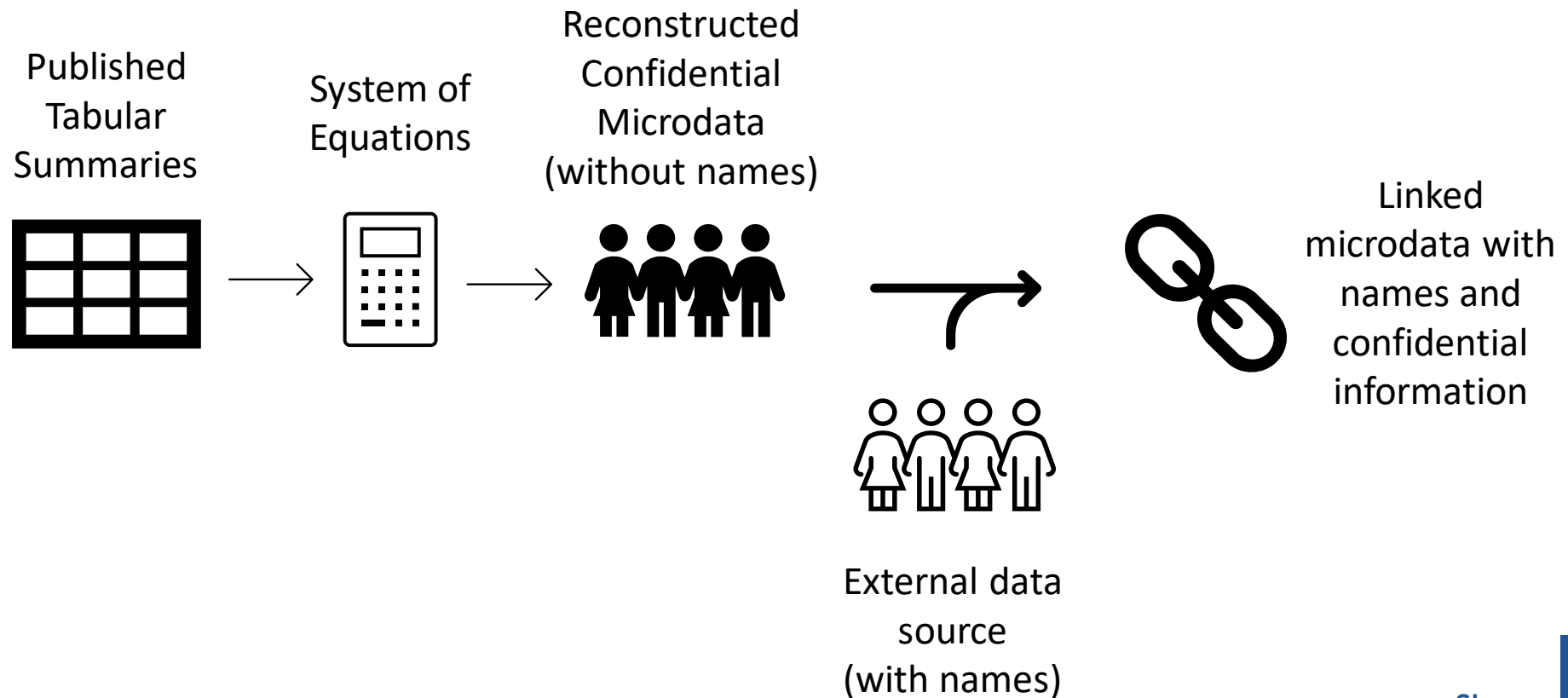


Evaluating the 2010 Disclosure Avoidance Methods

Recognizing the feasibility of reconstruction attacks on published tabular summaries, the Census Bureau decided to conduct a simulated attack on the disclosure avoidance methods used to protect the 2010 Census.

The goal was to evaluate whether the record swapping algorithms used to protect the published 2010 tabular summaries were sufficient to mitigate disclosure risk.

Simulated Reconstruction- abetted Re-identification Attack



Step 1:

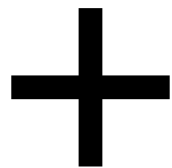
**Reconstruction of block-level
2010 Census records from
published tabular summaries**

Tables used

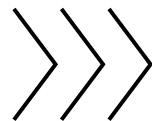
- P001** (Total Population by Block)
- P006** (Total Races Tallied by Block)
- P007** (Hispanic or Latino Origin by Race by Block)
- P009** (Hispanic or Latino, and Not Hispanic or Latino by Race by Block)
- P011** (Hispanic or Latino, and Not Hispanic or Latino by Race for the Population 18 Years and Over by Block)
- P012** (Sex by Age by Block)
- P012A-I** (Sex by Age by Block, iterated by Race)
- P014** (Sex by Single-year-of-age for the Population under 20 Years by Block)
- PCT012A-N** (Sex by Single-year-of-age by Tract, iterated by Race)

Implication of 2010 invariants

Exact Total Population Counts
by Block



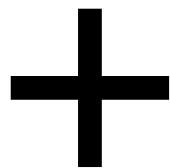
Exact Voting Age Population Counts
by Block



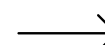
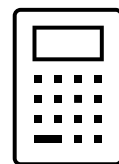
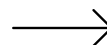
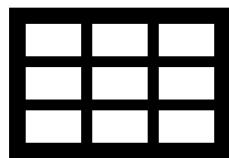
Exact reconstruction of all
308,745,538 records with
correct block and voting age

Adding Race, Ethnicity, Sex, and Age to each record

Exact reconstruction of all
308,745,538 records with
correct block and voting age



Race, Ethnicity, Sex, and Age
Tables



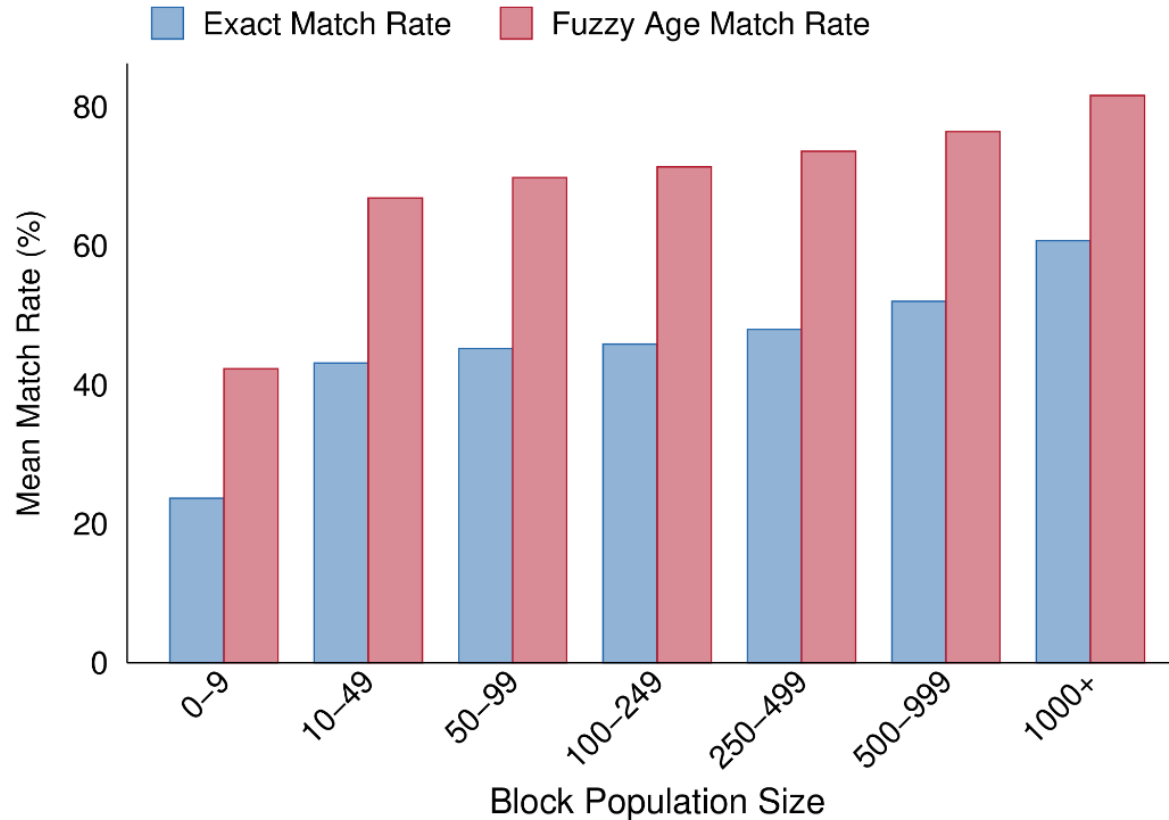
Reconstructed
Data with Block,
Race, Ethnicity,
Sex, and Age



Assessing the accuracy of the reconstruction

Agreement Rates between the Reconstructed Microdata and the 2010 Census Edited File and Hundred-percent Detail File					
	Record Counts		Agreement Rates		
Left file	In Left	In Reconstructed	Exact	Fuzzy Age	One error
CEF	308,745,538	308,745,538	46.48%	70.98%	78.31%
HDF	308,745,538	308,745,538	48.34%	73.33%	80.39%
DRB clearance number CBDRB-FY21-DSEP-003					

Assessing the accuracy of the reconstruction



Block-level agreement rates between the reconstructed 2010 Census microdata and the 2010 Census Edited File by population in the block

DRB clearance number CBDRB-FY21-DSEP-003.

Shape
your future
START HERE >

United States[®]
Census
2020

Population Uniques

Distribution of Population and Population Uniques by Block Population Size

Block Population Bin	Number of Blocks in Bin	2010 Census Population in Bin	Cumulative Population	Percent of Population in Bin	Cumulative Percent of Population	Population Uniques (block, sex, age) in Bin	Percent of (block, sex, age) Uniques in Bin
TOTAL	11,078,297	308,745,538				135,432,888	43.87%
0	4,871,270	0	0	0.00%	0.00%		
1-9	1,823,665	8,069,681	8,069,681	2.61%	2.61%	7,670,927	95.06%
10-49	2,671,753	67,597,683	75,667,364	21.89%	24.51%	53,435,603	79.05%
50-99	994,513	69,073,496	144,740,860	22.37%	46.88%	40,561,372	58.72%
100-249	540,455	80,020,916	224,761,776	25.92%	72.80%	27,258,556	34.06%
250-499	126,344	42,911,477	267,673,253	13.90%	86.70%	5,297,867	12.35%
500-999	40,492	27,028,992	294,702,245	8.75%	95.45%	1,051,924	3.89%
1000+	9,805	14,043,293	308,745,538	4.55%	100.00%	156,639	1.12%

DRB clearance number CBDRB-FY21-DSEP-003.

Shape
your future
START HERE >

United States[®]
Census
2020

Implications of the reconstruction

Existing technology can convert the Census Bureau's traditional tabular summaries into a highly accurate 100% microdata file geocoded to the block level.

This 100% microdata file would not have been considered releasable under the 2010 Census disclosure avoidance rules.

Faced with the new threat of reconstruction attacks, the disclosure avoidance methods used for the 2010 Census no longer meet the acceptable disclosure risk standards that were in place in 2010.

Step 2:

Re-identification attack on the reconstructed 2010 Census microdata

Re-identification Attack

1. Identify a source file (e.g., commercial data)
2. Identify the corresponding census block for each address in source file
3. Identify and link records from source file to the reconstructed data that match exactly on block, sex, and age.
4. Identify and link remaining records in source file that match exactly on block and sex, and match on age plus or minus 1 year.
5. Output the matched records from steps 3 and 4 (*putative re-identifications*)
6. Perform verification (field work or additional linkage) to estimate *confirmation rate*.

Assessing the accuracy of the re-identifications

Record Linkage Summary from Commercial and CEF Record Sources				
PIK, Block, Age, Sex Record Linkage Source	Available Records	Records with PIK, Block, Sex, and Age	Putative Re-identifications using Source	Confirmed Re-identifications
Commercial	413,137,184	286,671,152	137,709,807	52,038,366
CEF	308,745,538	279,179,329	238,175,305	178,958,726
DRB clearance number CBDRB-FY21-DSEP-003.				

Assessing the accuracy of the re-identifications

Confirmation and Recall Rates		
Source	Percentage of U.S. Resident Population (Confirmation Rate)	Percentage of Complete Data Population (Recall Rate)
Commercial	16.85%	18.15%
CEF	57.96%	64.10%
DRB clearance number CBDRB-FY21-DSEP-003.		

Precision Rates	
Source	Confirmed Percentage of Putative Re-identification (Precision Rate)
Commercial	37.79%
CEF	75.14%
DRB Clearance number CBDRB-FY21-DSEP-003.	

Assessing the accuracy of the re-identifications

Disclosure Risk Assessment of Population Uniques by Block Population Size

Block Population Bin	Putative Re-identifications (Source: Commercial Data)	Confirmed Re-identifications (Source: Commercial Data)	Precision (Source: Commercial Data)	Putative Re-identifications (Source: CEF)	Confirmed Re-identifications (Source: CEF)	Precision (Source: CEF)
TOTAL	137,709,807	52,038,366	37.79%	238,175,305	178,958,726	75.14%
0						
1-9	1,921,418	1,387,962	72.24%	4,220,571	4,093,151	96.98%
10-49	25,148,298	13,481,700	53.61%	47,352,910	43,415,168	91.68%
50-99	30,567,157	12,781,790	41.82%	51,846,547	42,515,756	82.00%
100-249	38,306,957	13,225,998	34.53%	63,258,561	45,807,270	72.41%
250-499	21,789,931	6,408,814	29.41%	35,454,412	22,902,054	64.60%
500-999	13,803,283	3,460,118	25.07%	23,280,718	13,514,134	58.05%
1000+	6,172,763	1,291,984	20.93%	12,761,586	6,711,193	52.59%

DRB clearance number CBDRB-FY21-DSEP-003.

Shape
your future
START HERE >

United States[®]
Census
2020

Implications of the simulated attack

The Census Bureau believed in 2010 that it was necessary to coarsen geographic identifiers in microdata such that the minimum population in any published geography was at least 100,000 persons (Public-Use Microdata Areas).

Our simulated reconstruction-abetted re-identification attack demonstrated that the tabular summaries from the 2010 Census can be converted into a 100% microdata file with geographic precision to the census block-level.

Our simulated attack demonstrated that, depending on the quality of the external data used, between 52 and 179 million respondents to the 2010 Census can be correctly re-identified from the reconstructed microdata.

Stronger privacy protections, such as those in the 2020 Census Disclosure Avoidance System, are necessary to protect against reconstruction-abetted attacks.

Stay Informed: Subscribe to the 2020 Census Data Products Newsletters

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

2020 Census Population Counts for Apportionment are Now Available

// [Census.gov](#) > [2020 Census Research, Operational Plans, and Oversight](#) > [Process](#) > [Disclosure Avoidance Modernization](#) > [2020 Census Data Products Newsletters](#)



2020 Census Data Products Newsletters

Sign up for news and information about 2020 Census Data Products and the implementation of the new Disclosure Avoidance System.

SIGN-UP FOR NEWSLETTERS

Past Issues:

April 28, 2021

New DAS Update Meets or Exceeds Redistricting Accuracy Targets

April 19, 2021

New Demonstration Data Will Feature Higher Privacy-loss Budget

April 07, 2021

Meeting Redistricting Data Requirements: Accuracy Targets

February 23, 2021

The Road Ahead: Upcoming Disclosure Avoidance System Milestones

February 03, 2021

New DAS Phase: Optimizing Tunable Elements


November 25, 2020

Invariants Set for 2020 Census Data Products

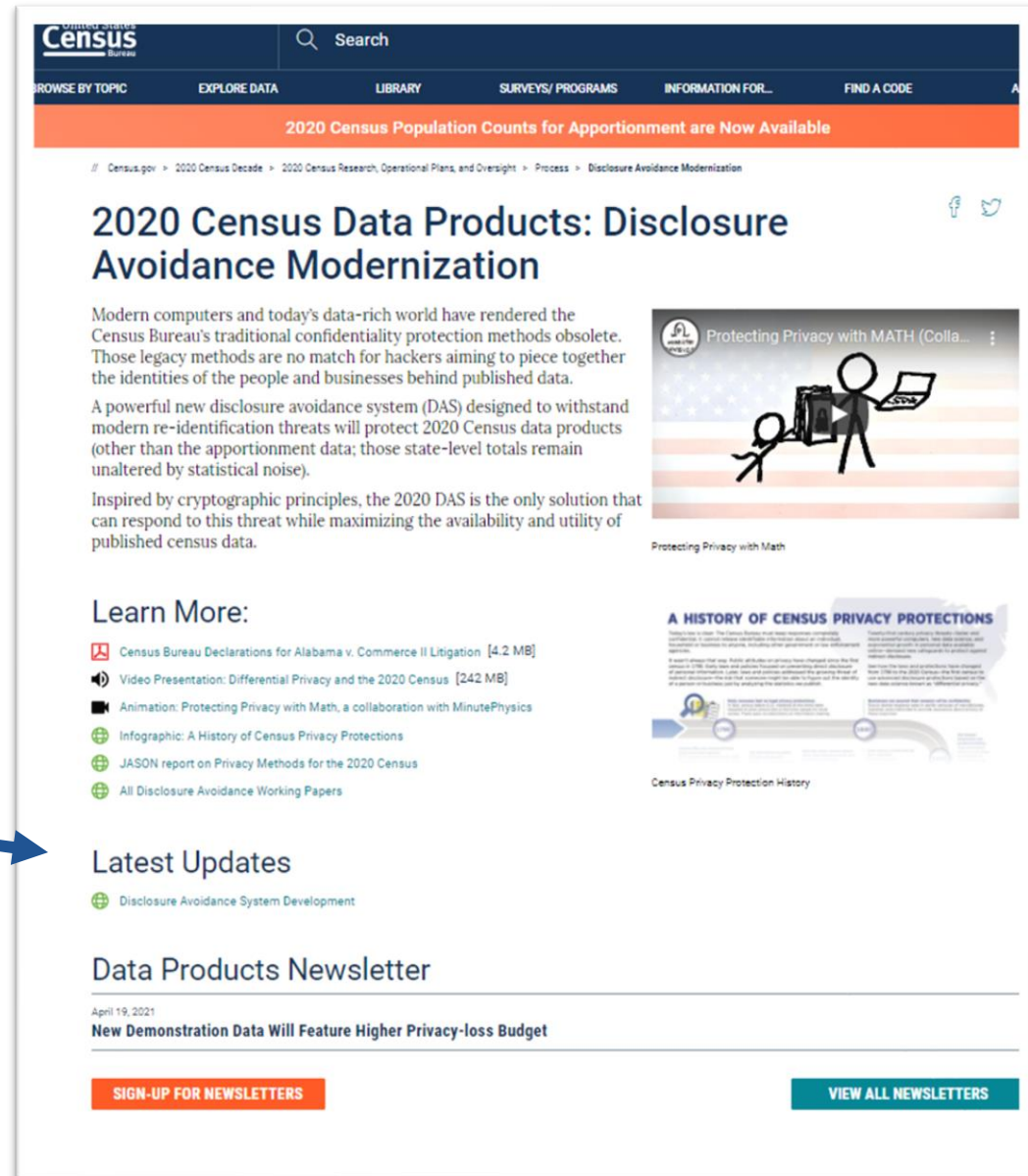
Stay Informed:
Visit Our Website

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

Latest Updates

 [Disclosure Avoidance System Development](#)

Dates and log-In information for
webinar series



The screenshot shows the 2020 Census website with a dark blue header containing the '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...', and 'FIND A CODE'. A prominent orange banner reads '2020 Census Population Counts for Apportionment are Now Available'. The main content area features the article '2020 Census Data Products: Disclosure Avoidance Modernization'. The article text explains that modern computers have rendered traditional confidentiality methods obsolete and introduces a new disclosure avoidance system (DAS) designed to withstand modern re-identification threats. It also mentions that state-level totals remain unaltered by statistical noise and that the 2020 DAS is the only solution that can respond to this threat while maximizing the availability and utility of published census data. To the right of the article is an infographic titled 'Protecting Privacy with MATH (Colla...)' showing stick figures with a laptop and a document. Below the article is a 'Learn More:' section with links to 'Census Bureau Declarations for Alabama v. Commerce II Litigation' (4.2 MB), 'Video Presentation: Differential Privacy and the 2020 Census' (242 MB), 'Animation: Protecting Privacy with Math, a collaboration with MinutePhysics', 'Infographic: A History of Census Privacy Protections', 'JASON report on Privacy Methods for the 2020 Census', and 'All Disclosure Avoidance Working Papers'. To the right of this section is another infographic titled 'A HISTORY OF CENSUS PRIVACY PROTECTIONS' showing a timeline of privacy protections. Below the 'Learn More:' section is a 'Latest Updates' section with a link to 'Disclosure Avoidance System Development'. Further down is a 'Data Products Newsletter' section with the date 'April 19, 2021' and the headline 'New Demonstration Data Will Feature Higher Privacy-loss Budget'. At the bottom are two buttons: 'SIGN-UP FOR NEWSLETTERS' and 'VIEW ALL NEWSLETTERS'.

Webinar Series:

Understanding the 2020 Census Disclosure Avoidance System

All webinars start at **1:00 pm EDT**

No pre-registration necessary. We will archive recordings to the website.

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

Or link: <https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-updates.html>

Day	Date	Title
T	May 4	Differential Privacy 101
F	May 7	The Census Bureau's Simulated Reconstruction-Abetted Re-identification Attack on the 2010 Census
Th	May 13	Differential Privacy 201 and the TopDown Algorithm
F	May 14	Highlights of the April 2021 Detailed Summary Metrics
F	May 21	Analysis of April 2021 Demonstration Data for Redistricting and Voting Rights Act Use Cases

Questions?

