Differential Privacy and the 2020 Census

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

NM Webinar
November 19, 2020

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, and Pavel Zhurlev.

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

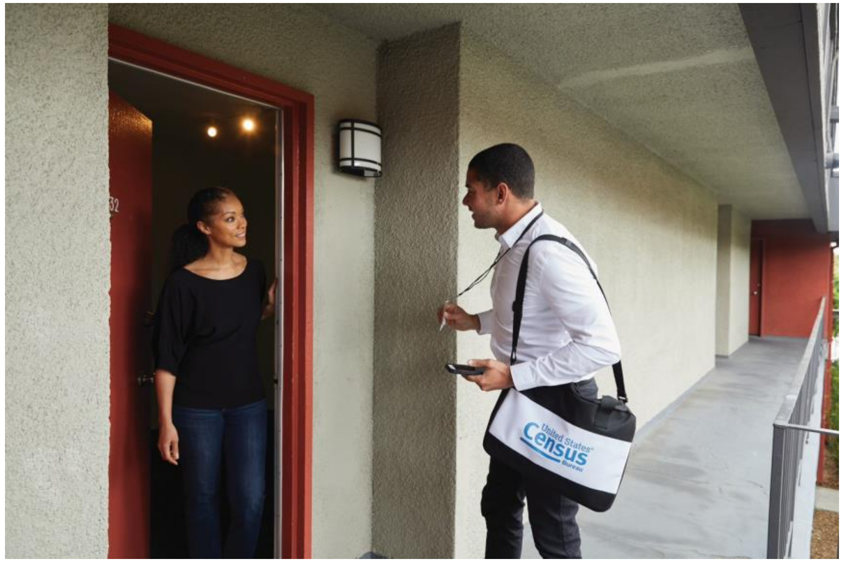
Any opinions and viewpoints expressed in this presentation are the author's own, and do not represent the opinions or viewpoints of the U.S. Census Bureau.

The statistics included in this newsletter have been cleared for public dissemination by the Census Bureau's Disclosure Review Board (CBDRB-FY20-DSEP-001, CBDRB-FY20-281, and CBDRB-FY20-101).

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

Upholding our Promise: Today and Tomorrow

We cannot merely consider privacy threats that exist today.

We must ensure that our disclosure avoidance methods are also sufficient to protect against the threats of tomorrow!

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

The Growing Privacy Threat

More Data and Faster Computers!

In today’s digital age, there has been a proliferation of databases that could potentially be used to attempt to undermine the privacy protections of our statistical data products.

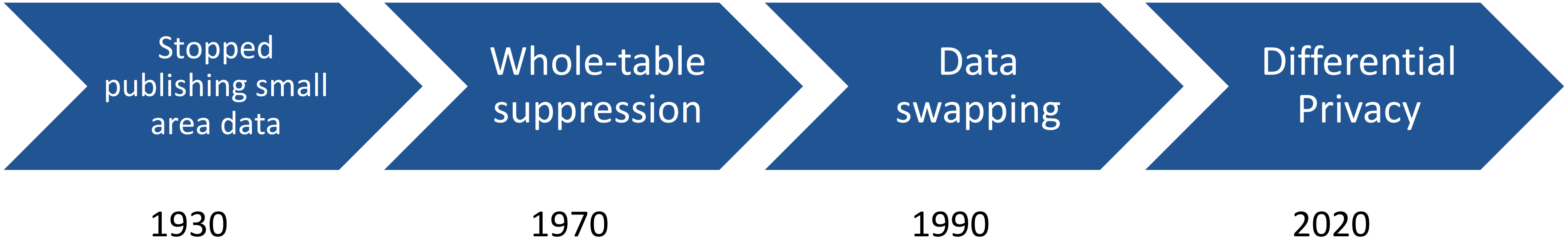
Similarly, today’s computers are able to perform complex, large-scale calculations with increasing ease.

These parallel trends represent new threats to our ability to safeguard respondents’ data.

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.



Stopped publishing small area data — 1930

Whole-table suppression — 1970

Data swapping — 1990

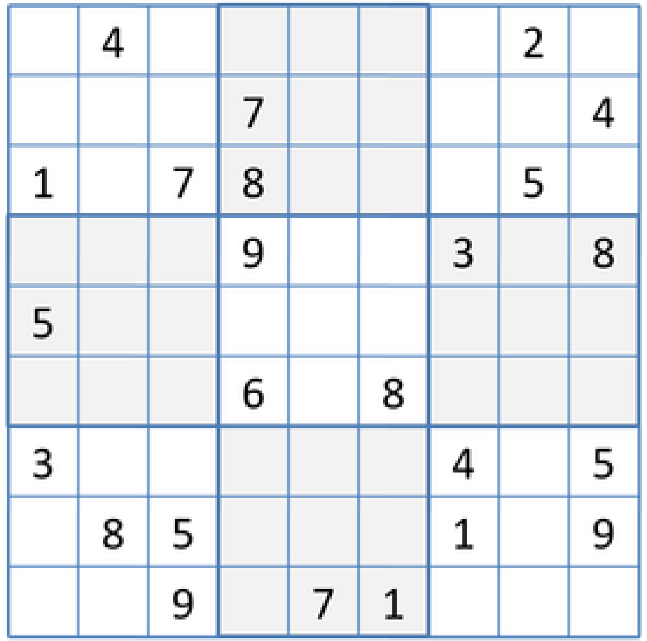
Differential Privacy — 2020

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: An Example

	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

	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

Age	Sex	Race	Relationship
66	Female	Black	Married
84	Male	Black	Married
30	Male	White	Married
36	Female	Black	Married
8	Female	Black	Single
18	Male	White	Single
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	Age	Sex
Jane Smith	66	Female
Joe Public	84	Male
John Citizen	30	Male

External Data

+

Age	Sex	Race	Relationship
66	Female	Black	Married
84	Male	Black	Married
30	Male	White	Married

Confidential Data

Reconstructing the 2010 Census

- The 2010 Census collected information on the 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)

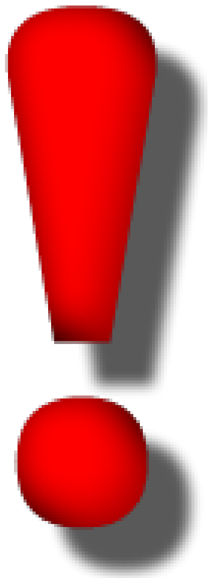
3. Block, sex, and age were then linked to commercial data, which provided presumed re-identification of 45% of the population (138 million individuals).

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 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.



Differential Privacy

-quantifies the precise amount of privacy risk…

-for all calculations/tables/data products produced…

-no matter what external data is available…

-now, or at any point in the future!

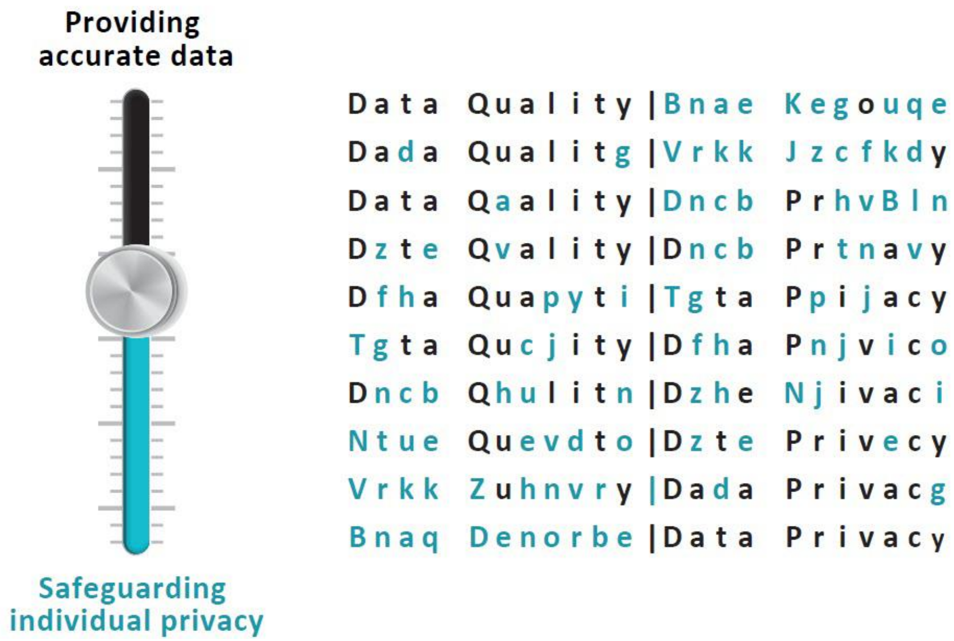
Precise amounts of noise

Differential privacy allows us to inject a precisely calibrated amount of noise into the data to control the privacy risk of any calculation or statistic.

Privacy vs. Accuracy

The only way to absolutely eliminate all risk of re-identification would be to never release any usable data.

Differential privacy allows you to quantify a precise level of "acceptable risk," and to precisely calibrate where on the privacy/accuracy spectrum the resulting data will be.


Providing accurate data
Data Quality|Bnae Kegouqe
Dada Qualitg|Vrkk Jzcfkdy
Data Qaality|Dncb PrhvBln
Dzte Qvality|Dncb Prtnavy
Dfha Quapyti|Tgta Ppijacy
Tgta Qucjity|Dfha Pnjvico
Dncb Qhulitn|Dzhe Njivaci
Ntue Quevdto|Dzte Privecy
Vrkk Zuhnvry|Dada Privacg
Bnaq Denorbe|Data Privacy
Safeguarding individual privacy

Establishing a Privacy-loss Budget

This measure is called the “Privacy-loss Budget” (PLB) or “Epsilon.”

- **ε=0** (perfect privacy) would result in completely useless data
- **ε=∞** (perfect accuracy) would result in releasing the data in fully identifiable form

ε

Epsilon

Shape
your future
START HERE >

United States®
Census
2020

Comparing Methods

Data Accuracy

Differential Privacy is not inherently better or worse than traditional disclosure avoidance methods.

Both can have varying degrees of impact on data quality depending on the parameters selected and the methods' implementation.

Privacy

Differential Privacy is substantially better than traditional methods for protecting privacy, insofar as it actually allows for measurement of the privacy risk.

Implications for the 2020 Census

The switch to Differential Privacy does not change the constitutional mandate to apportion the House of Representatives according to the actual enumeration.

As in 2000 and 2010, the Census Bureau will apply privacy protections to the PL94-171 redistricting data.

The switch to Differential Privacy requires us to re-evaluate the quantity of statistics and tabulations that we will release, because each additional statistic uses up a fraction of the privacy-loss budget (epsilon).

Privacy-loss Budget Allocation

The Census Bureau's Data Stewardship Executive Policy Committee (DSEP) will be making decisions about the privacy-loss budget for the 2020. This includes allocation across different 2020 Census data products, including:

"Group I Products"

- PL94-171 Redistricting data
- Demographic and Housing Characteristics files

"Group II Products"

- Detailed race, ethnicity, and tribal data
- Household characteristics

…and other uses of Decennial Census data.

Privacy-loss Budget Allocation

DSEP will also be deciding how to allocate the privacy-loss budget across different sets of tabulations within each data product.

By Geographic Level:

Nation	20%
State	20%
County	12%
Tract Group	12%
Tract	12%
Block Group	12%
Block	12%

By Query Set:

DHC-Person Allocations	
total	30%
hhgq	15%
votingage * hispanic * cenrace	29%
age * sex * hispanic * cenrace	25%
detailed	1%

(sample allocation for DHC-P, as implemented in 5/27/2020 Detailed Summary Metrics)

Impact of Operational Schedule

Compressing production schedule to meet statutory deadlines.

Focusing all attention on identifying the activities and schedules needed to produce accurate and complete Redistricting (P.L. 94-171) data by March 31, 2021.

Planning and DAS development work for the production and release of the remaining 2020 Census data products will restart immediately following completion of the apportionment and redistricting data planning activities.

Recent DAS Development Priorities

“Hardening” the system for integration testing and production

Bringing data accuracy under the direct control of PLB
(reducing post-processing error and ensuring accuracy improves as PLB increases)

Improving accuracy for legal/political entities
(especially for AIAN tribal areas and populations in “off-spine” geographies)

Ensuring security of random-number generator

New Demonstration Data

Release of 2010 Demonstration Privacy-Protected Microdata Files 2020-11-16

and

Detailed Summary Metrics 2020-11-16

Limited to data necessary to support PL94-171 Redistricting data
(Tables P1-P5 and H1)

Tabulations have also been produced by CNSTAT and IPUMS National Historical Geographic Information System (NHGIS)

PLB of ε=4 for PPMF-Person and ε=0.5 for PPMF-Unit runs

Major DAS Changes included in November 16 PPMF

- Correction of defect found in September PPMF
- Bringing AIAN Tribal Areas “onto the spine”
- Inclusion of a new state-level AIAN tribal area population invariant
- Improvements to DAS Rounder to decrease post-processing error
- Switch from Geometric Distribution to Discrete Gaussian Distribution for DP noise injection
- Stability and reliability enhancements

Additional Resources

Disclosure Avoidance and the 2020 Census Website
https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

Questions? Suggestions?
Email them to 2020DAS@census.gov

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

301-763-1960 (Office)
michael.b.hawes@census.gov