

# Differential Privacy and the Overall Privacy of Decennial Data

**Michael Hawes**

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

Census Information Center & State  
Data Center Training Conference

Charlotte, NC  
June 12, 2019

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

# First, the fine print...

## Disclaimer:

This presentation is intended to explain the U.S. Census Bureau's plans to utilize formal privacy methods for the 2020 Decennial Census to a primarily non-technical audience. As such, many of the computer science and mathematical concepts underlying formal privacy have been simplified for the sake of clarity. For more information and technical details relating to the issues discussed in these slides, please contact the author at [michael.b.hawes@census.gov](mailto:michael.b.hawes@census.gov).

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

All images used in this presentation are in the public domain.

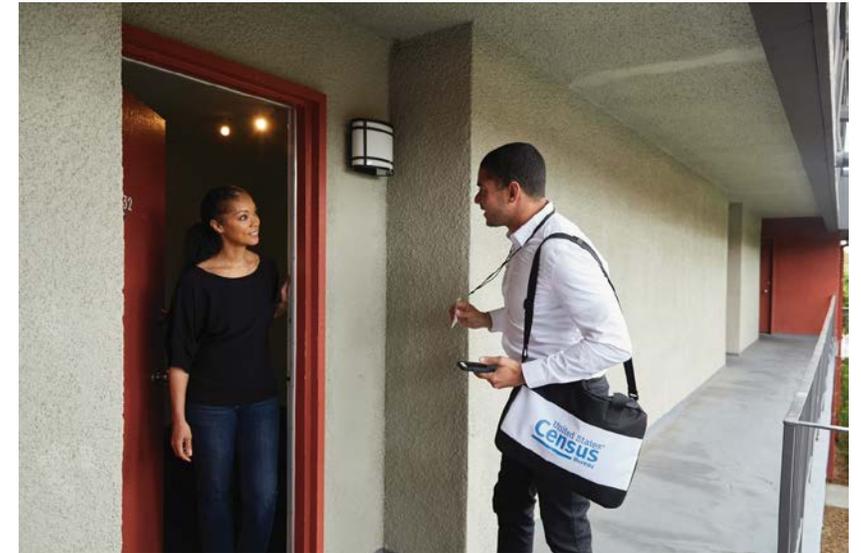
# Presentation Overview

- Title 13 and our commitment to data stewardship
- Where we came from: the Census Bureau's privacy protections over time
- The growing threat of re-identification
- Differential Privacy – what it is, and what it isn't!
- Implications for the 2020 Decennial Census
- Questions

# Our Commitment to Data Stewardship

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.



# It's the Law

Title 13, Section 9 of the United State Code prohibits the Census Bureau from releasing identifiable data “furnished by any particular establishment or individual.”

Census Bureau employees are sworn for life to safeguard respondents' information.

Penalties for violating these protections can include fines of up to \$250,000, and/or imprisonment for up to five years!



# Keeping the Public's Trust

Safeguarding the public's data is about more than just complying with the law!

The quality and accuracy of our censuses and surveys depend on our ability to keep the public's trust.

In an era of declining trust in government, increasingly common corporate data breaches, and declining response rates to surveys, we must do everything we can to keep our promise to protect the confidentiality of our respondent's data.



# 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!

# 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.

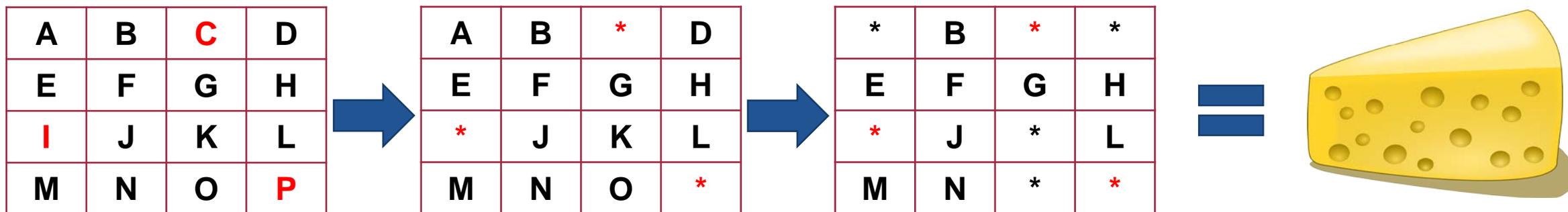
# Privacy and Data Usability

Every disclosure avoidance method reduces the accuracy and usability of the data.

Traditional methods for protecting privacy (suppression, coarsening, and perturbation) can have significant impacts on the usability of the resulting data products, but data users are often not aware of the magnitude of those effects.

# Suppression

Removing sensitive values from the data.



# Coarsening

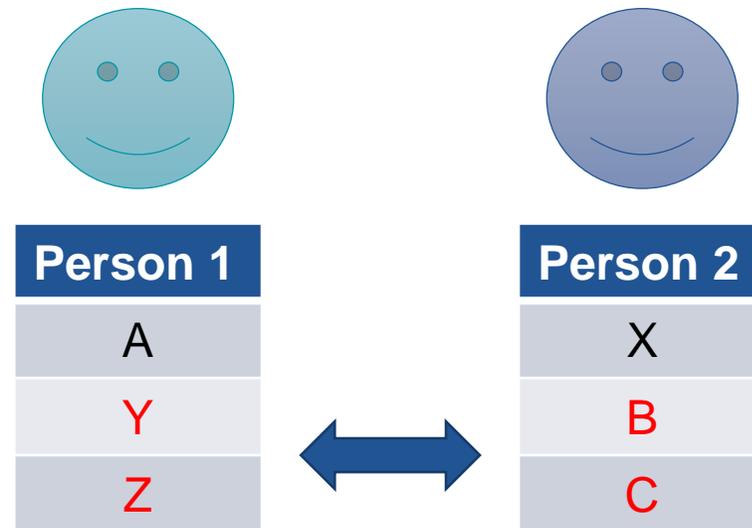
Reducing the amount of detail in the data.

- Geographic aggregation
- Collapsing categories
- Rounding
- Reporting in ranges, etc.



# Swapping

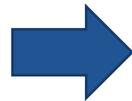
Exchanging sensitive values between records



# Noise

Inserting error to increase uncertainty.

14	41	50	58	65
15	24	26	30	25
52	53	66	47	51
68	6	44	17	32
38	26	33	42	64



13	41	51	58	65
15	24	25	30	24
51	54	66	48	51
68	6	44	16	32
38	25	33	42	65

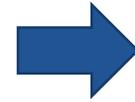
# Synthetic Data

Estimating new data based on patterns in the original data.

10	4	72	38	5
6	29	40	27	8
19	6	23	14	2
13	9	41	18	1
22	11	62	15	2



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots$$



12	4	68	35	7
5	25	46	27	10
22	8	24	17	1
12	10	38	19	3
21	9	65	18	5

# How much is enough?

**Quantifying the privacy risk associated with traditional disclosure avoidance methods is difficult.**

Practitioners of traditional disclosure avoidance techniques rely heavily on expert judgement and personal experience.

Quantifying the remaining privacy risk of a data product protected using traditional methods is, for all intents and purposes, impossible.

Consequently, as the risks of re-identification have risen over time, agencies have had to increase their suppression thresholds, coarsening rules, and swapping rates, to keep pace.

But, as these trends were not being objectively measured, there was no concrete way to determine how much protection was necessary...*until now.*

# 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.

# 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		

# Re-identification

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



# In the News

Reconstruction and Reidentification are not just theoretical possibilities...they are happening!

- **Massachusetts Governor's Medical Records** (Sweeney, 1997)
- **AOL Search Queries** (Barbaro and Zeller, 2006)
- **Netflix Prize** (Narayanan and Shmatikov, 2008)
- **Washington State Medical Records** (Sweeney, 2015)
- and many more...

# 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 7.7 Billion statistics.

Internal Census Bureau research confirms that the confidential 2010 Census microdata can be accurately reconstructed from the publicly released tabulations.

# 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

aka “Formal Privacy”

- quantifies the precise amount of re-identification risk...

  - for all calculations/tables/data products produced...

    - no matter what external data is available...

      - now, or at any point in the future!

# Differential Privacy is a Promise

**“You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available.”**

*Dwork and Roth, Foundations and Trends in Theoretical Computer Science,  
Volume 9, Numbers 3-4, 2014*

# Sensitivity

How much would a calculation be affected by removing any particular individual?

Impacted by:

- Type of calculation
- Size of population
- Diversity (heterogeneity) of values

	Age
John	51
Jane	55
Joe	61

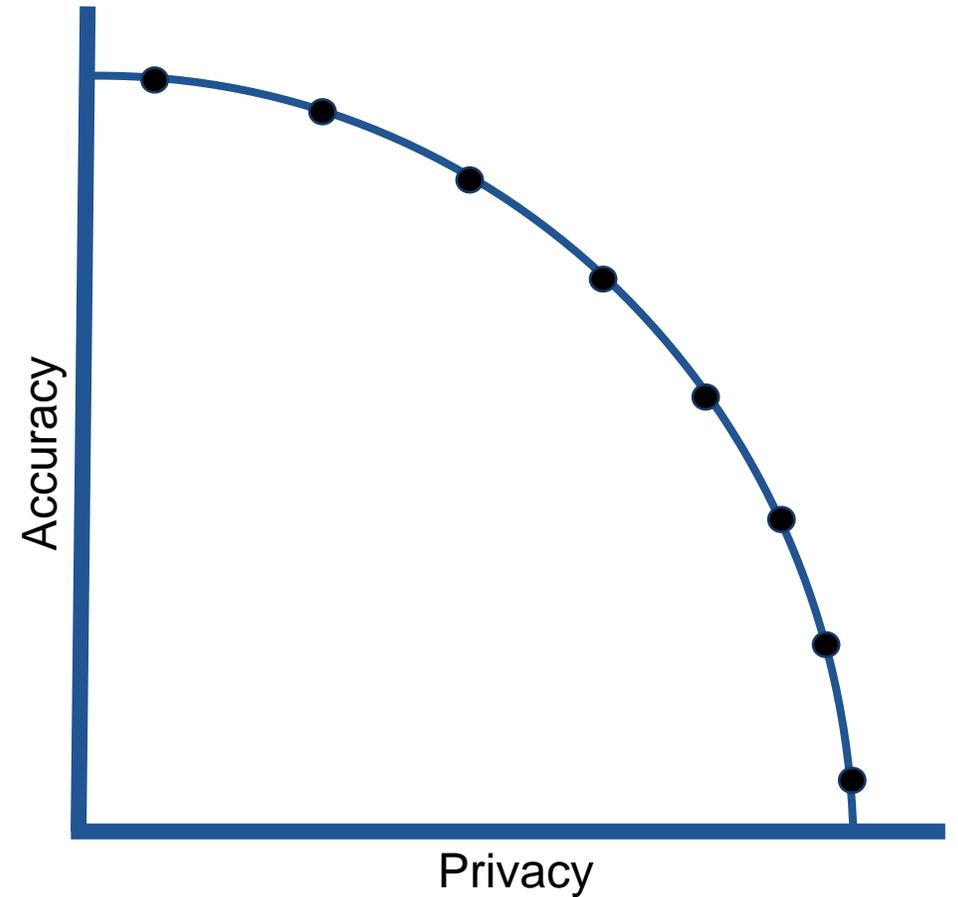
Mean (without John)	58
Mean (without Jane)	56
Mean (without Joe)	53

# Precise amounts of noise

Differential privacy injects precise amounts of noise into the data based on the sensitivity of the calculation being performed.

# Privacy vs. Accuracy

Differential Privacy also allows policymakers to precisely calibrate where on the privacy/accuracy tradeoff curve the resulting data will be.



# Establishing a Privacy Budget

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” of re-identification.

This measure is called the “Privacy Budget” or “Epsilon.”

$\epsilon=0$  (perfect privacy) would result in completely useless data

$\epsilon=\infty$  (perfect accuracy) would result in releasing the data in fully identifiable form



Epsilon

# Allocating the Privacy Budget

Each calculation, query, or tabulation of the data consumes a fraction of the privacy budget.

$$(\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 \dots + \epsilon_n = \epsilon_{\text{Total}})$$

Calculations/tables for which high accuracy is critical can receive a larger share of the overall privacy budget.

# Keeping Accuracy High

When Differential Privacy is applied, the accuracy of the resulting data will be affected by:

- The number of calculations being performed or tables being generated;
- The type of calculation being performed (e.g., count vs. mean);
- The size of the underlying populations for each calculation or table;
- The uniformity/diversity of the population;
- The overall privacy budget (epsilon); and
- The allocation of the privacy budget across calculations/tables.

# 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 Decennial Census

The switch to Differential Privacy will not change the constitutional mandate to reapportion 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 budget (epsilon).

In order to maximize the accuracy of the data, the Census Bureau is carefully evaluating what tabulations will be released at different levels of geography.

# Quantifiable Promise to the Public

Differential Privacy allows us to give a formal, verifiable, and measurable guarantee to the public that we are safeguarding their data.

Differential Privacy is future proof. Neither the future availability of data nor future advancements in computer capabilities can diminish the privacy guarantees.

# You Can Help Us to Help You!

**Senior Census Bureau policymakers will be making important decisions – and they need your input!**

The actual impact of Differential Privacy on the usability and accuracy of the 2020 Census data products will ultimately depend on the following factors:

- What will the overall privacy budget (epsilon) be?
- What statistics will the Census Bureau release at which levels of geography?
- How will the overall privacy budget be allocated across different geographies, tables, and statistics?

In order for the Census Bureau's senior leadership to make the most informed decisions on these questions, they need to know how you plan to use the 2020 Census data.

You can send us your input at [michael.b.hawes@census.gov](mailto:michael.b.hawes@census.gov)

# Questions?

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

301-763-1960 (Office)

[michael.b.hawes@census.gov](mailto:michael.b.hawes@census.gov)

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020