Title 13, Differential Privacy, and the 2020 Decennial Census

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

Metropolitan Washington Council of Governments
October 8, 2019

Shape your future START HERE >

United States® Census 2020

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**

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 necessarily represent the opinions or viewpoints of the U.S. Census Bureau.

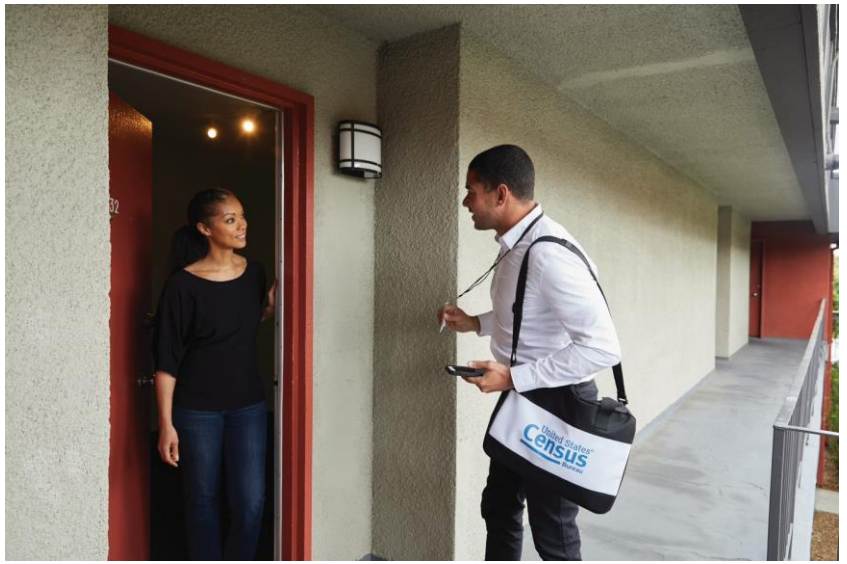
Shape your future START HERE >

United States Census 2020

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.



Shape your future
START HERE >

United States Census 2020

It's the Law

"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)

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!

Shape your future START HERE >

United States Census 2020

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.

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 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

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.

Shape your future START HERE >

United States® Census 2020

Suppression

Removing sensitive values from the data.

A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P

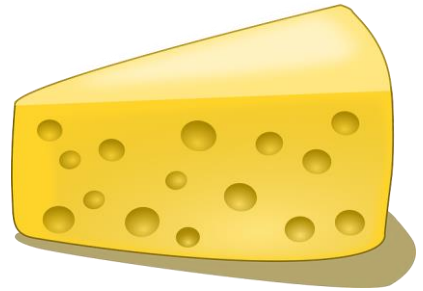
→

A	B	*	D
E	F	G	H
*	J	K	L
M	N	O	*

→

*	B	*	*
E	F	G	H
*	J	*	L
M	N	*	*

=



Shape your future START HERE >

United States Census 2020

Coarsening

Reducing the amount of detail in the data.

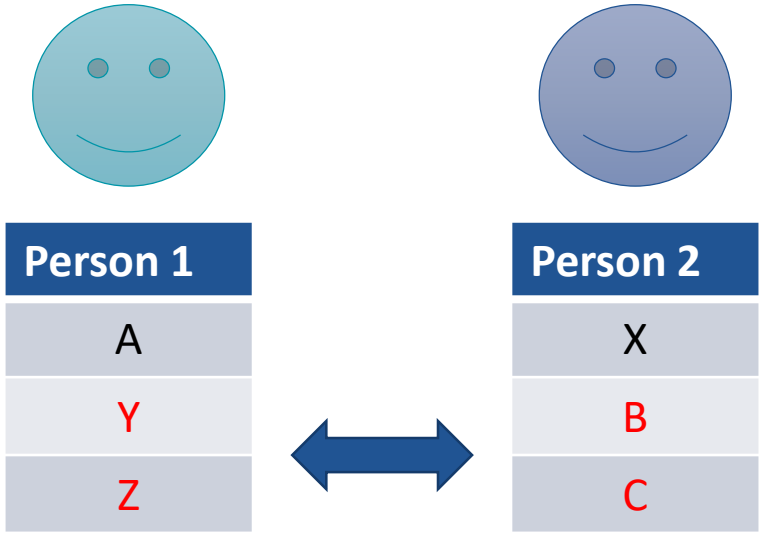
- Geographic aggregation
- Collapsing categories
- Rounding
- Reporting in ranges, etc.

Shape your future START HERE >

United States Census 2020

Swapping

Exchanging sensitive values between records

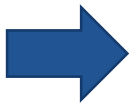


Shape your future START HERE >
United States Census 2020

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

Shape your future START HERE >
United States Census 2020

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*.

Shape your future START HERE >

United States® Census 2020

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.

Shape your future START HERE >

United States Census 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			

Shape your future START HERE >

United States Census 2020

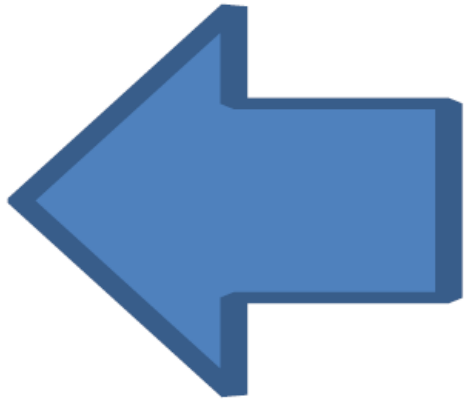
Reconstruction: An Example

	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 F	3	36	36.7

Shape your future START HERE >
United States Census 2020

Reconstruction: An Example

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



	Count	Median	Mean
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 F	3	36	36.7

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

Shape your future START HERE >
United States Census 2020

Re-identification

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

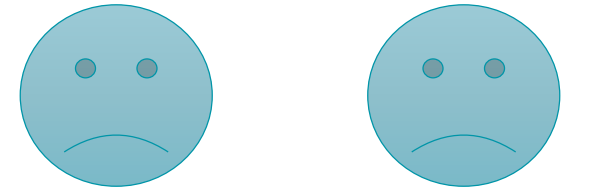
Dataset A

Person 1	Person 2
A	X
B	Y
C	Z

+

External Dataset B

Joe Smith	Jane Doe
C	Z



Joe Smith	Jane Doe
A	X
B	Y
C	Z

Shape your future START HERE >

United States Census 2020

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…

Shape your future START HERE >
United States Census 2020

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.

Shape your future START HERE >

United States Census 2020

Reconstructing the 2010 Census: How did we do it?

- Performed database reconstruction for all 308,745,538 people enumerated in the 2010 Census from public 2010 data products.
- Linked reconstructed records to commercially available databases.
- Successful record linkage to commercial data = “putative re-identification”
- Compared putative re-identifications to confidential data.
- Successful linkage to confidential data = “confirmed re-identification”
- Potential harm to individuals: can learn self-response race and ethnicity.

Shape your future START HERE >
United States Census 2020

Reconstructing the 2010 Census: What did we find?

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

Shape your future START HERE >
United States Census 2020

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

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!

Shape your future START HERE >

United States® Census 2020

Differential Privacy is a Promise

"You will not be affected, adversely or otherwise, by allowing your data to be used…no matter what other…information sources are available."

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

Shape your future START HERE >

United States® Census 2020

Assessing Privacy Risk

Traditional Disclosure Avoidance Considers Absolute Privacy Risk

Can an individual be re-identified in the data, and can some sensitive attribute about them be inferred?

Evaluates risk given a particular, defined mode of attack, asking: What is the likelihood, at this precise moment in time, of re-identification and inferential disclosure by a particular type of attacker with a defined set of available external information?

Formal Privacy is about Relative Privacy Risk

Does not directly measure re-identification risk (which requires specification of an attacker model).

Instead, it defines the maximum privacy "leakage" of each release of information compared to some counterfactual benchmark (e.g., compared to a world in which a respondent does not participate, or provides incorrect information).

Shape your future START HERE >
United States Census 2020

Sensitivity

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

Impacted by:

- **Type of calculation**
- **Size of population**
- **Range of possible values**

	Age
John	51
Jane	55
Joe	61

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

Shape your future START HERE >

United States® Census 2020

Precise amounts of noise

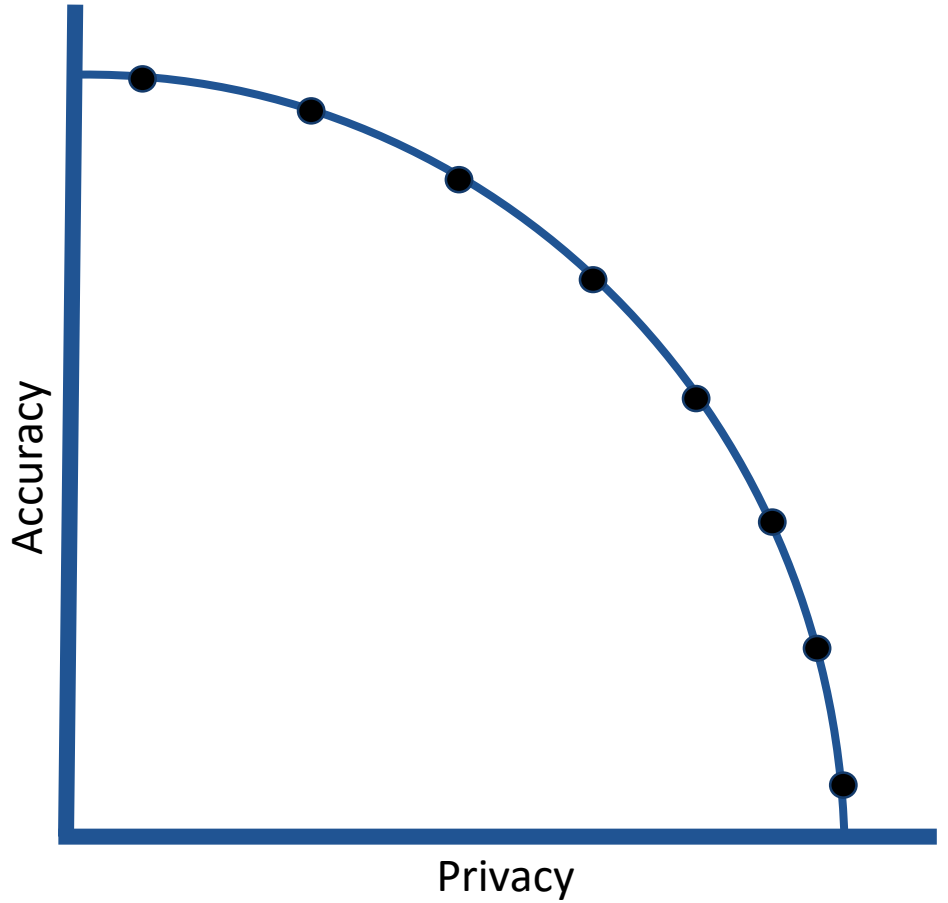
Differential privacy allows us to inject a precisely calibrated amount of noise into the data based on the sensitivity of the calculation being performed.

Shape your future START HERE >

United States® Census 2020

Privacy vs. Accuracy

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



Shape your future START HERE >
United States Census 2020

Establishing a Privacy-loss 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."

- **$\varepsilon=0$ (perfect privacy) would result in completely useless data**
- **$\varepsilon=\infty$ (perfect accuracy) would result in releasing the data in fully identifiable form**

Epsilon

Shape your future START HERE >

United States® Census 2020

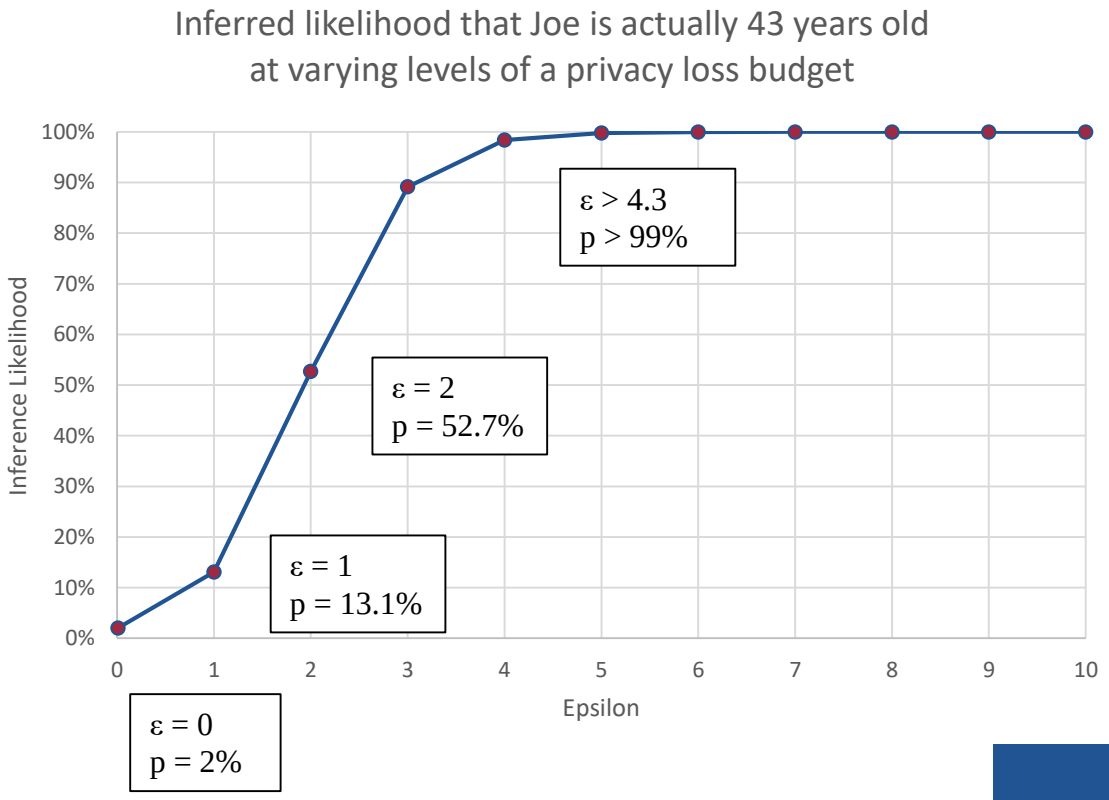
The Formal Guarantee

Can Sara determine Joe's exact age?

Suppose Joe submitted erroneous information for the Census, and the best Sara could otherwise do to determine Joe's exact age (from other available information) is to predict that there is a 2% chance that Joe is 43 years old.

If Joe instead provides accurate information for the Census, then a small amount of information about him will "leak" through the publication of data products. This new information can improve Sara's estimate.

The Privacy-loss Budget determines the amount of that leakage and the corresponding maximum possible improvement to Sara's prediction.



Assumes that Sara has infinite computing resources, infinitely powerful algorithms, and allows her to have arbitrary side knowledge.

Shape your future START HERE >
United States Census 2020

Allocating the Privacy-loss Budget

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

($\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 \ldots + \varepsilon_n = \varepsilon_{Total}$)

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

Shape your future START HERE >

United States® Census 2020

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 range of possible values;**
- **The overall privacy budget (epsilon); and**
- **The allocation of the privacy budget across calculations/tables.**

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.

Shape your future START HERE >

United States Census 2020

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.

Shape your future START HERE >

United States® Census 2020

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.

Shape your future START HERE >

United States® Census 2020

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 products?

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.

Shape your future START HERE >
United States Census 2020

2010 Demonstration Products

- Census Bureau plans to release a set of data products that demonstrate the computational capabilities of the DAS. The current version of the DAS will be run on the 2010 internal data to produce two products:
  - PL 94-171
  - Demographic and Housing Characteristics File (selected tables)

- Allows data users to assess the impacts of the DAS implementation
- Data will be publicly released (Target date: October 2019)
- Table shells/structure will be released in advance

Shape your future START HERE >
United States Census 2020

Questions?

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

Shape your future START HERE >
United States Census 2020