Data Protection and Official Statistics

Lessons Learned from the 2020 Census
and a
Proposed Framework for Discussing and Evaluating Disclosure Avoidance Systems

Michael B. Hawes
Senior Statistician for Scientific Communication
U.S. Census Bureau

July 22, 2024

Guest Lecture
STAT S-115
Harvard University

Any opinions or viewpoints are the presenter's own and do not reflect the opinions or viewpoints of the U.S. Census Bureau

United States® Census Bureau

A brief bit about me:

B.A., Duke University (Political Science and History)

Diplôme, Sciences-Po Paris (Political and Social Sciences)

M.A., University of Chicago (International Relations)

michael.b.hawes@census.gov
301.763.1960 (office)

- Senior Statistician for Scientific Communication, U.S. Census Bureau
- Appointed Member, Federal Committee on Statistical Methodology
- Appointed Member, American Statistical Association Committee on Privacy and Confidentiality

A few prior career highlights:

- Senior Advisor for Data Access and Privacy, U.S. Census Bureau
- Director of Student Privacy, U.S. Department of Education
- Statistical Privacy Advisor, U.S. Department of Education
- Privacy Consultant, Federal Commission on Evidence-based Policymaking

United States® Census Bureau

What we'll cover

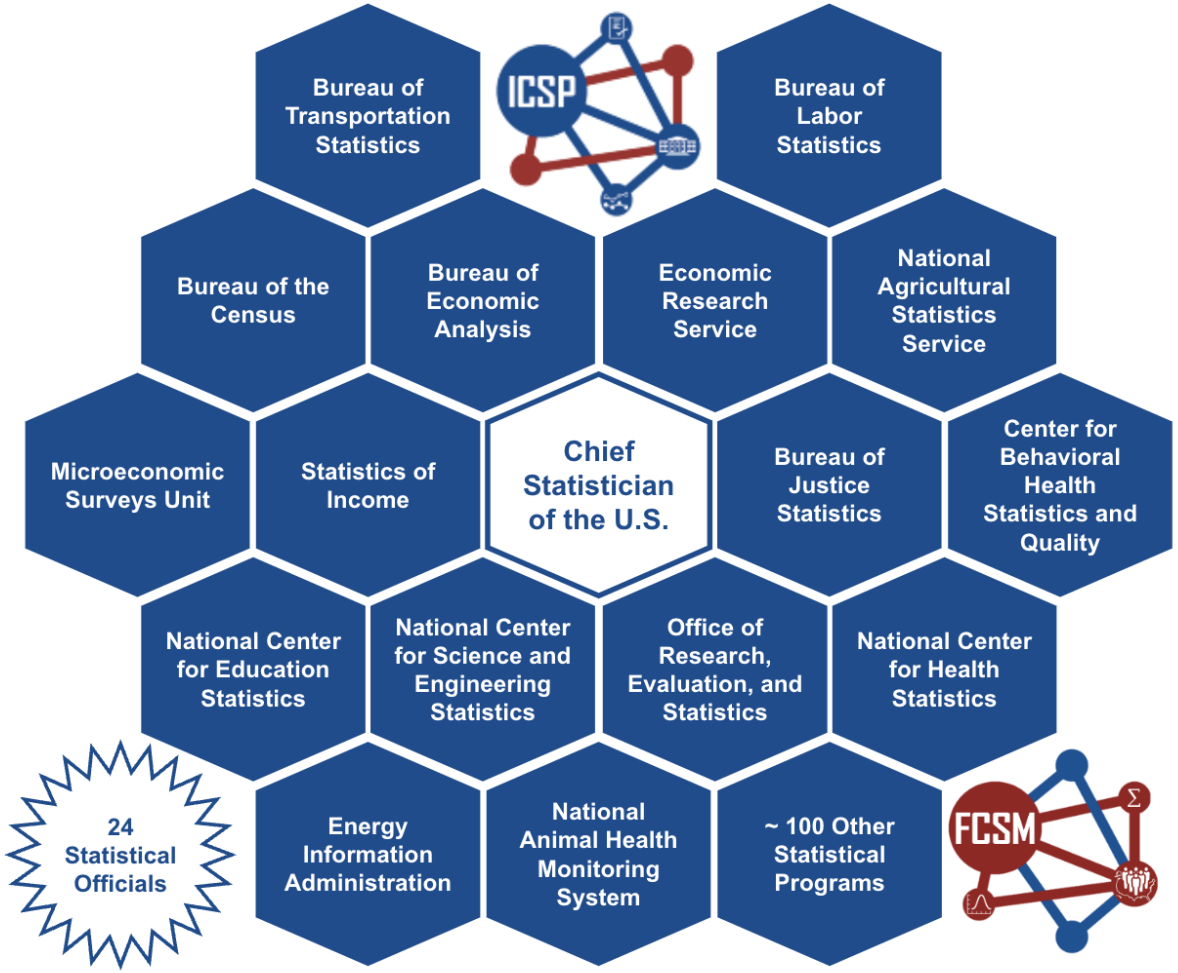
- Background
- The Data Protection Challenge and the Triple Tradeoff of Official Statistics
- The 2020 Census Disclosure Avoidance System
- Lessons Learned from Implementing Differential Privacy for the 2020 Census
- A Principled Framework for Disclosure Avoidance and the Characteristics of an Ideal, Applied, Disclosure Avoidance System

United States® Census Bureau

U.S. Federal Statistical System

Decentralized Statistical System

- 13 Principle Statistical Agencies
- 3 Recognized Statistical Units
- 20 Statistical Officials
- Over 100 Statistical Programs



https://www.statspolicy.gov/

U.S. Census Bureau

The Census Bureau's mission is to serve as the nation’s leading provider of quality data about its people and economy.

- Decennial Census of Population and Housing
- Economic Census
- Census of Governments

…and over 100 demographic and economic surveys on a monthly, quarterly, or annual basis.

United States® Census Bureau

U.S. Census Bureau at a Glance

"Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers…The actual Enumeration shall be made…within every subsequent Term of ten Years, in such Manner as they shall by Law direct."

U.S. Constitution, Article I, Section 2

United States® Census Bureau

Decennial Census of Population and Housing

- First conducted in 1790 under the direction of Thomas Jefferson.
- The 2020 Census was the 24th decennial enumeration of the United States population.
- Counts each resident of the 50 states, District of Columbia, Puerto Rico, and the Island Areas.
- In 2020, hired and deployed approx. 288,000 enumerators to conduct non-response follow up to 56 million addresses.

United States® Census Bureau

Technological Innovation

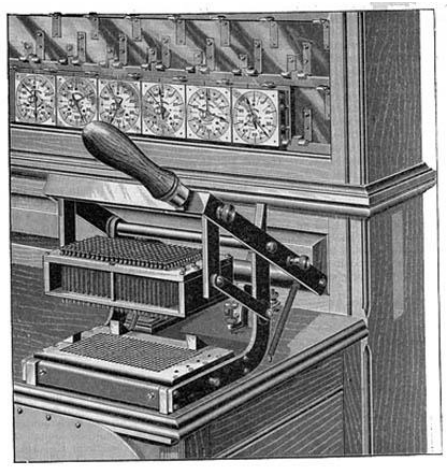
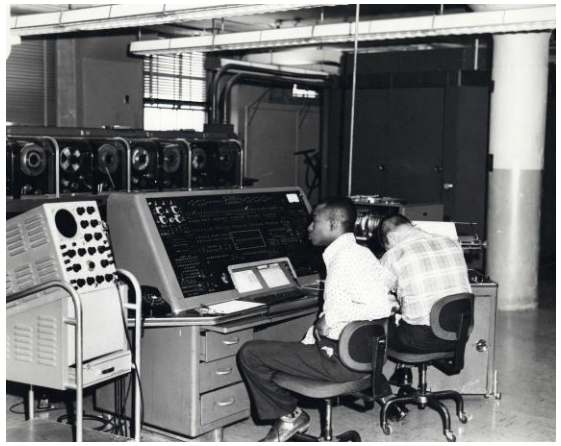


Fig. 8.—Circuit-Closing Press.

Hollerith's Electric Sorting and Tabulating Machine.

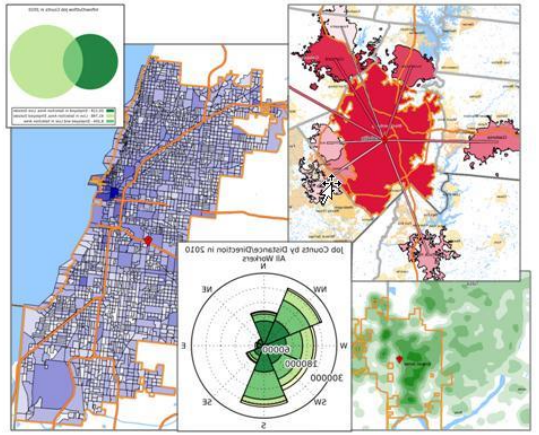
1890 Census

Herman Hollerith's
Punch Card and
Electronic Tabulator



1950 Census

UNIVAC I



2008

First Use of
Differential Privacy for
OnTheMap

United States Census 2020

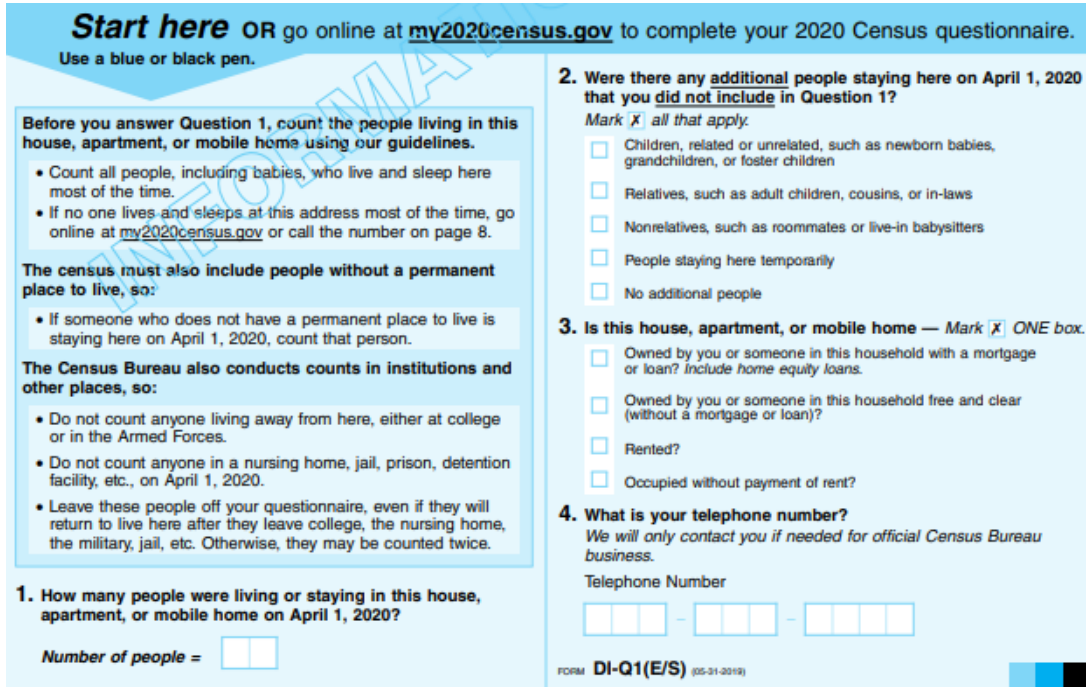
2020

Use of Satellite Imagery
Internet Self Response
Mobile Device Enumeration
2020 Disclosure Avoidance System

United States® Census Bureau

2020 Census Response Information

- Name
- Sex
- Birthdate
- Race (Major Race Groups and Detailed Race Write-in Boxes)
- Ethnicity (Hispanic/Not Hispanic)
- Relationship to Householder
- Household Tenure (e.g., Owned/Rented)



Start here OR go online at **my2020census.gov** to complete your 2020 Census questionnaire.

Use a blue or black pen.

Before you answer Question 1, count the people living in this house, apartment, or mobile home using our guidelines.

- Count all people, including babies, who live and sleep here most of the time.
- If no one lives and sleeps at this address most of the time, go online at my2020census.gov or call the number on page 8.

The census must also include people without a permanent place to live, so:

- If someone who does not have a permanent place to live is staying here on April 1, 2020, count that person.

The Census Bureau also conducts counts in institutions and other places, so:

- Do not count anyone living away from here, either at college or in the Armed Forces.
- Do not count anyone in a nursing home, jail, prison, detention facility, etc., on April 1, 2020.
- Leave these people off your questionnaire, even if they will return to live here after they leave college, the nursing home, the military, jail, etc. Otherwise, they may be counted twice.

1. How many people were living or staying in this house, apartment, or mobile home on April 1, 2020?

Number of people =

2. Were there any additional people staying here on April 1, 2020 that you did not include in Question 1?

Mark ✗ all that apply.

- ☐ Children, related or unrelated, such as newborn babies, grandchildren, or foster children
- ☐ Relatives, such as adult children, cousins, or in-laws
- ☐ Nonrelatives, such as roommates or live-in babysitters
- ☐ People staying here temporarily
- ☐ No additional people

3. Is this house, apartment, or mobile home — *Mark ✗ ONE box.*

- ☐ Owned by you or someone in this household with a mortgage or loan? *Include home equity loans.*
- ☐ Owned by you or someone in this household free and clear (without a mortgage or loan)?
- ☐ Rented?
- ☐ Occupied without payment of rent?

4. What is your telephone number?

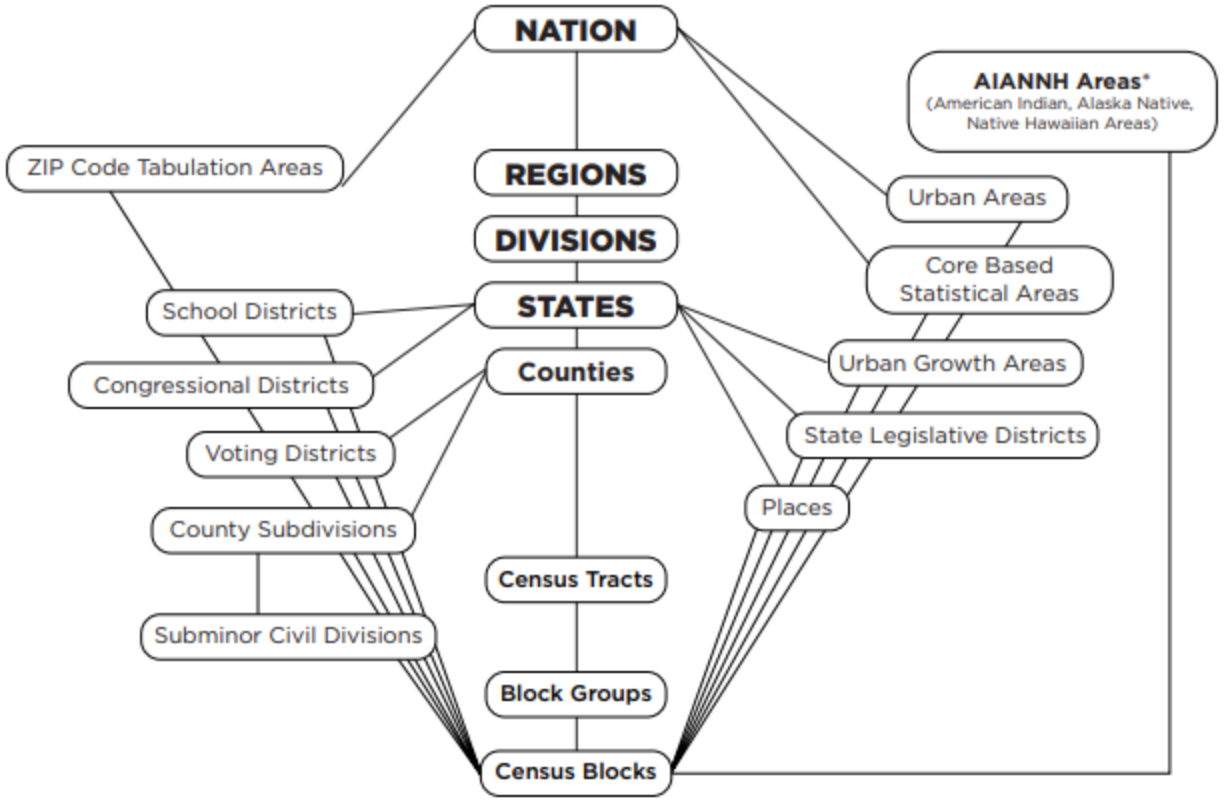
We will only contact you if needed for official Census Bureau business.

Telephone Number

FORM **DI-Q1(E/S)** (05-31-2019)

Census Geographies

Figure 2-1.
Standard Hierarchy of Census Geographic Entities



* Refer to the "Hierarchy of American Indian, Alaska Native, and Native Hawaiian Areas."

United States® Census Bureau

Uses of Decennial Census Data

- To apportion the U.S. House of Representatives
- To draw federal, state, and local legislative and voting districts
- To allocate over $2.8 Trillion in federal funds annually
- To inform federal, state, tribal, and local planning and policymaking
- To support research, development, and investment
- To serve as a benchmark for public and private surveys, projections, and estimates throughout the decade

…and much more.

United States® Census Bureau

2020 Census Data Products

"Group I Products"

- P.L. 94-171 Redistricting Data Summary File
- Demographic Profiles
- Demographic and Housing Characteristics File (DHC)

"Group II Products"

- Detailed DHC-A
- Detailed DHC-B
- Supplemental DHC

"Group III Products"

- Public Use Microdata
- Research-based Statistical Products
- Researcher Access
- Out-year uses of 2020 Census data

United States® Census Bureau

13 U.S. Code

Section 8(b): "…the Secretary may furnish copies of tabulations and other statistical materials **which do not disclose the information reported by, or on behalf of, any particular respondent**…"

Section 9: "[The Census Bureau may not] make any publication **whereby the data furnished by any particular establishment or individual under this title can be identified**"

The Challenge of Disclosure Avoidance

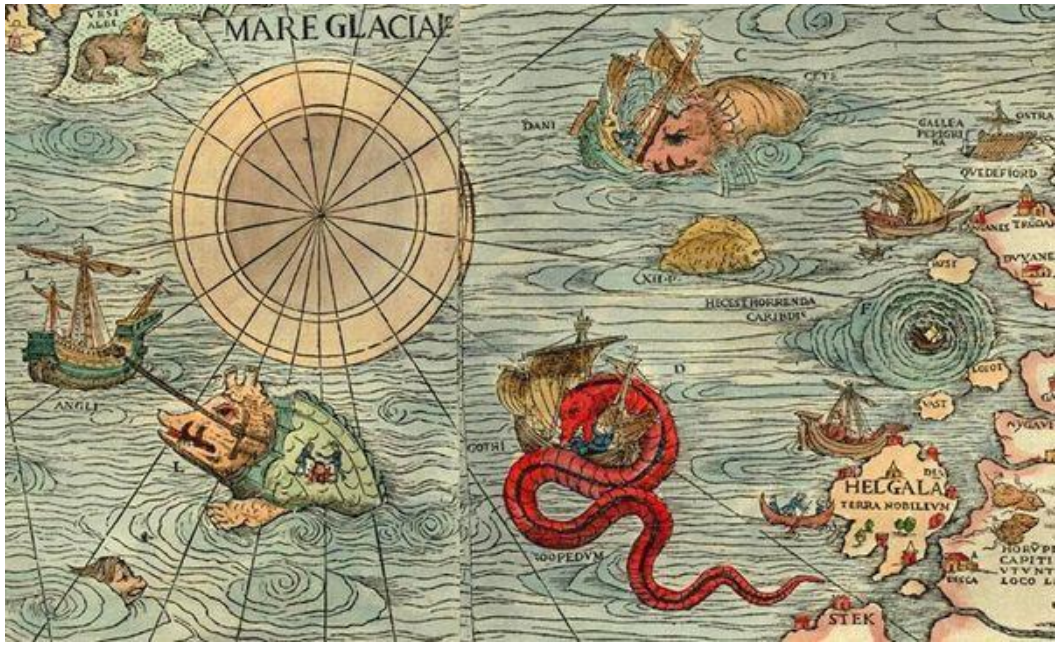
"It is the responsibility of Federal statistical agencies and recognized statistical units to produce and disseminate relevant and timely information; conduct credible, accurate, and objective statistical activities; and protect the trust of information providers by ensuring confidentiality and exclusive statistical use of their responses"

-OMB Statistical Policy Directive No.1 (2014)

"It has long been recognized that any available tabulation of the characteristics of a population is likely to narrow the range of uncertainty about the characteristics of specific individuals known to be members of that population…The release of any data usually entails at least some element of risk. A decision to eliminate all risk of disclosure would curtail statistical releases drastically, if not completely…"

-FCSM Statistical Policy Working Paper #2 (1979)

HIC SUNT DRACONES



United States® Census Bureau

The Challenge of Disclosure Avoidance

- 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

Re-identification Attacks

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

Name	Block	Age	Sex
Jane Smith	1234	66	Female
Joe Public	1234	84	Male
John Citizen	1234	30	Male

External Data

Block	Age	Sex	Race	Relationship
1234	66	Female	Black	Married
1234	84	Male	Black	Married
1234	30	Male	White	Married

Anonymized Agency Data

United States® Census Bureau

Reconstruction Attacks

- 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			

United States® Census Bureau

Disclosure Avoidance Methods

Disclosure avoidance methods seek to safeguard confidentiality, by:

- Reducing precision (Coarsening)
- Removing vulnerable records (Suppression), or
- Adding uncertainty (Perturbation)

Commonly used methods include:

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

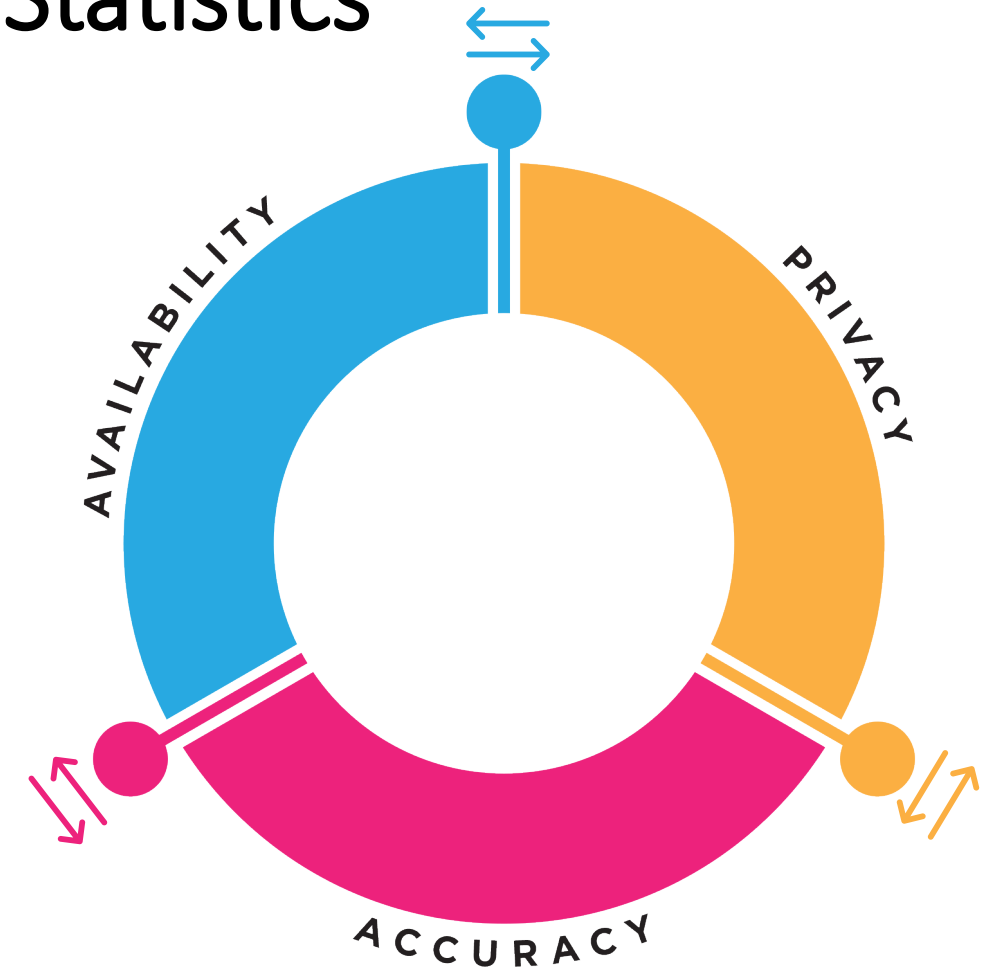
United States® Census Bureau

The Triple Trade-Off of Official Statistics

The more statistics you publish, and the greater the granularity and accuracy of those statistics, the greater the disclosure risk.

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

You can maximize on any two dimensions, but only at profound cost to the third.



United States® Census Bureau

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 risks are constantly increasing.
- Protecting confidentiality means adapting and responding to these increasing threats

United States® Census Bureau

Disclosure Avoidance for Past Censuses

1970-1980 Censuses

	528	
		794
	581	
137	941	189
931		
	250	
		590

SUPPRESSION

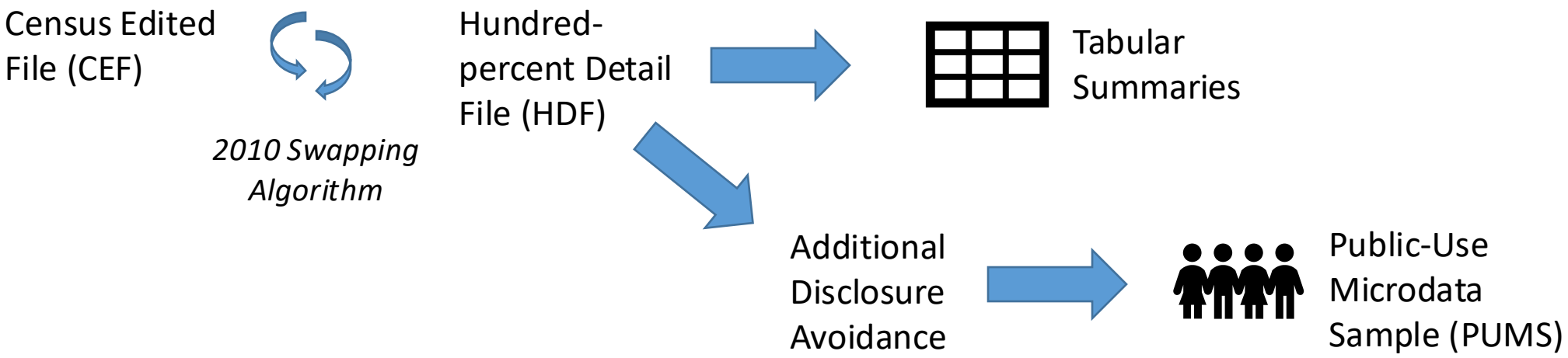
1990-2010 Censuses

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

Disclosure Avoidance Methods for the 2010 Census



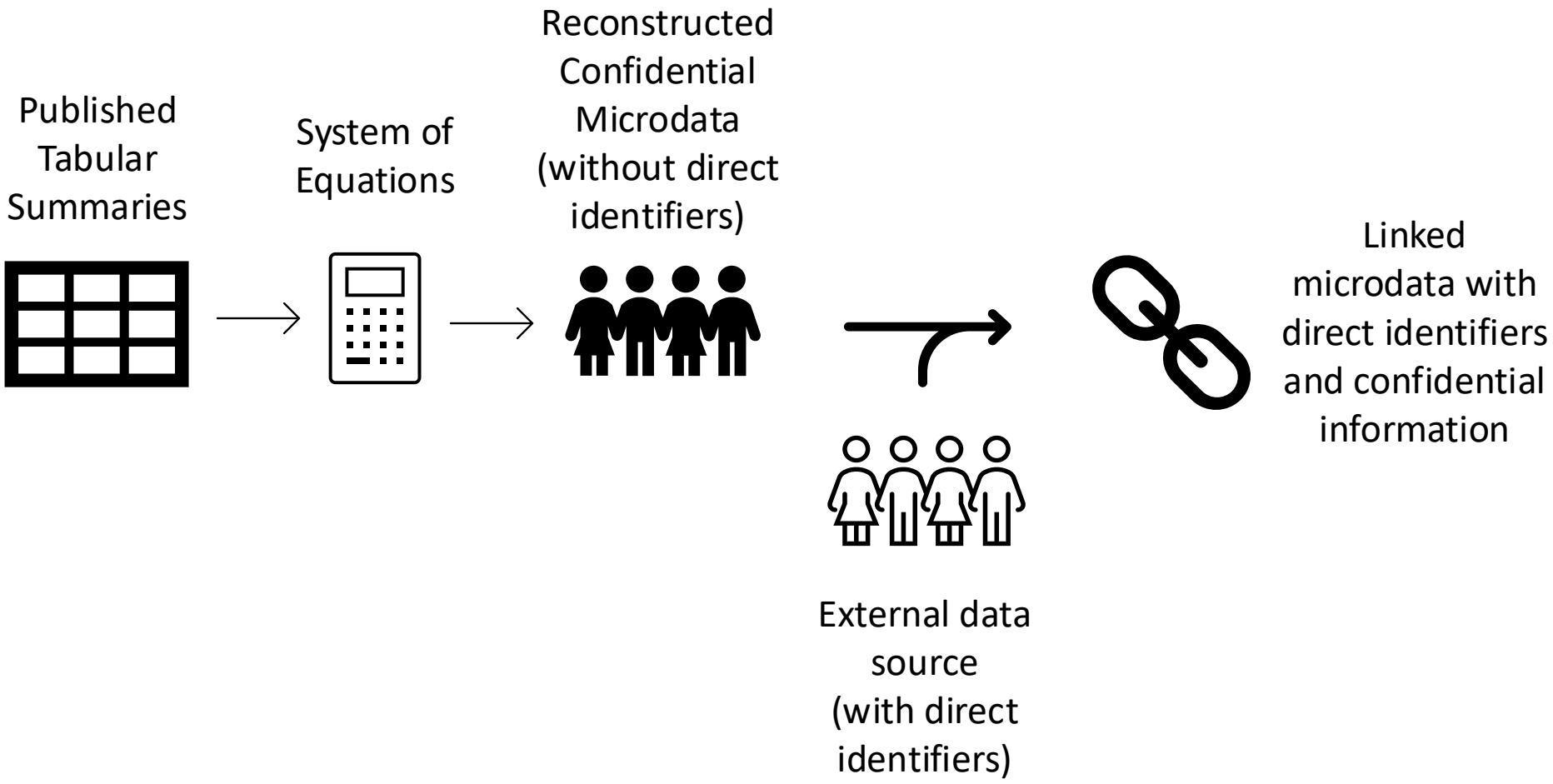
United States® Census Bureau

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.

United States® Census Bureau

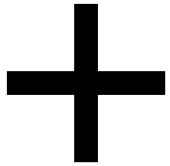
Simulated Reconstruction-abetted Re-identification Attack



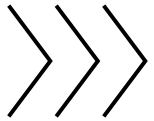
United States® Census Bureau

Implication of 2010 invariants

Exact Total Population Counts
by Block



Exact Voting Age Population Counts
by Block

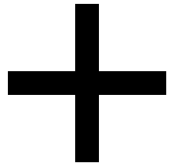


Exact reconstruction of all
308,745,538 records with
correct block and voting age

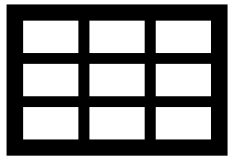
United States® Census Bureau

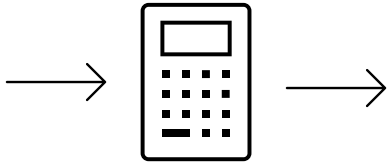
Adding Race, Ethnicity, Sex, and Age to each record

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

United States® Census Bureau

The results were alarming

Disclosure Avoidance for the 2020 Census

The 2020 Census improves on the noise injection methods of the 1990-2010 Censuses by employing a mathematical framework known as Differential Privacy (DP) to assess and quantify disclosure risk and 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.

United States® Census Bureau

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 statistical accuracy for any given level of privacy protection.*

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

There is no single way to implement DP

- Because Differential Privacy is a privacy-risk accounting framework, rather than a disclosure avoidance method, there are many different ways you can leverage DP to protect privacy throughout the information lifecycle.
- DP is used in a variety of ways by tech companies (e.g., Apple, Microsoft, Google, Uber) to protect privacy at the point of data collection.
- Similarly, there are many different ways that DP can be used by a trusted data curator (e.g., the Census Bureau) to protect data intended for public dissemination.

2020 Census Data Products

"Group I Products"

- P.L. 94-171 Redistricting Data Summary File
- Demographic Profiles
- Demographic and Housing Characteristics File (DHC)

"Group II Products"

- Detailed DHC-A
- Detailed DHC-B
- Supplemental DHC

"Group III Products"

- Public Use Microdata
- Research-based Statistical Products
- Researcher Access
- Out-year uses of 2020 Census data

United States® Census Bureau

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

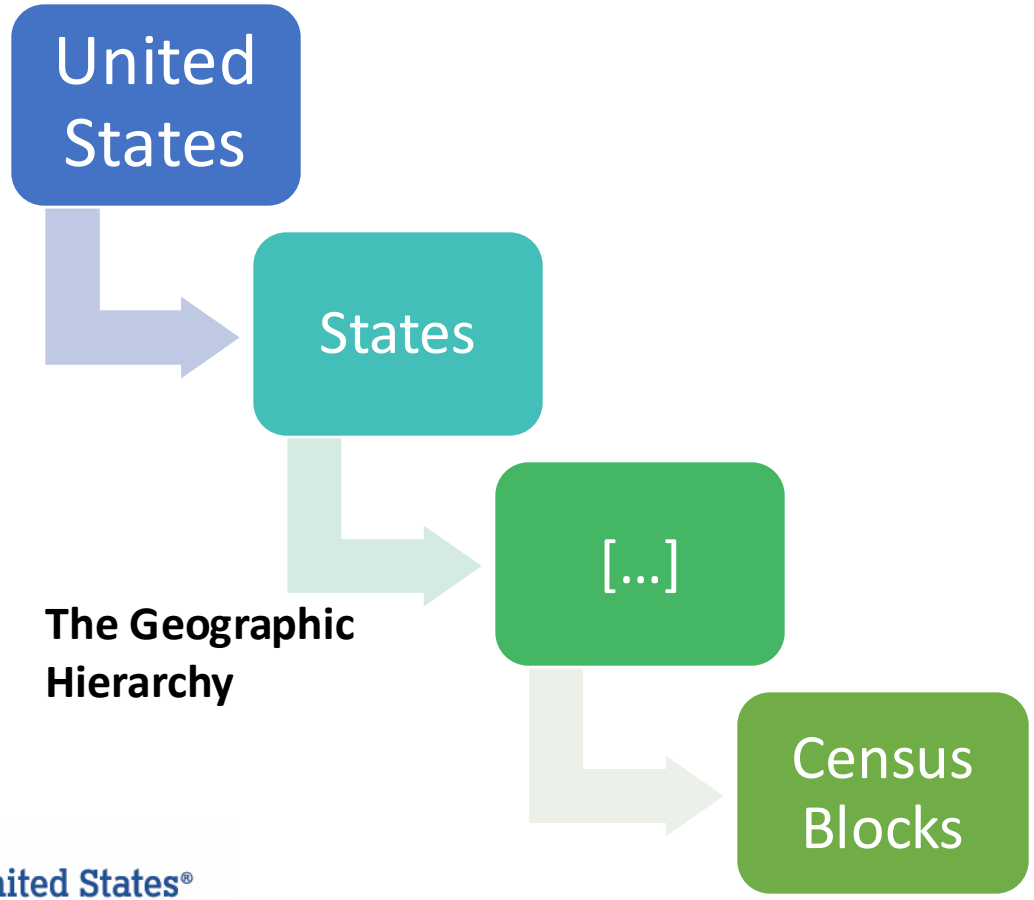
United States® Census Bureau

The TopDown Algorithm

For complete details see: Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., & Zhuravlev, P. (2022). The 2020 Census Disclosure Avoidance System TopDown Algorithm. *Harvard Data Science Review*. (June) https://doi.org/10.1162/99608f92.529e3cb9

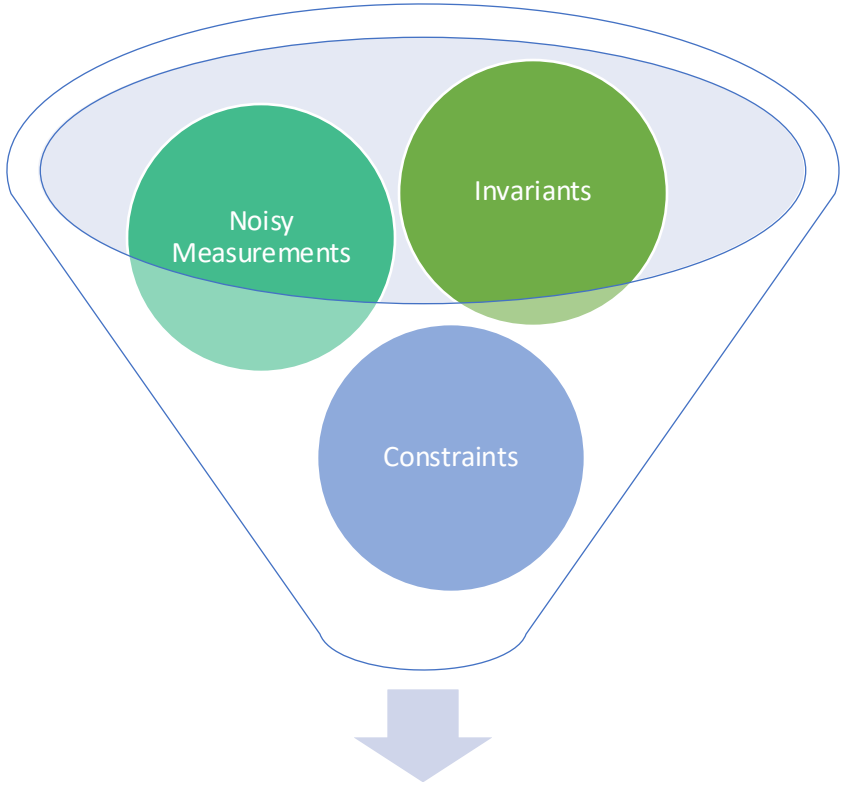
United States® Census Bureau

The TopDown Algorithm



The Geographic Hierarchy

At each geographic level:



Internally consistent histogram

Ensuring Fitness-for-Use

All disclosure avoidance methods, and the parameters of their implementation, impact the resulting data's fitness-for-use in different ways.

Agencies must be deliberate in their selection and implementation of disclosure avoidance methods to ensure they meet the needs of their intended data users.

Requires:

- Subject Matter Expertise
- Research and Evaluation
- Stakeholder Communication and Engagement

United States® Census Bureau

TDA Query Structure

TDA only takes noisy measurements for defined queries (tabulations) at particular geographic levels. Adjusting the queries asked and/or the share of PLB assigned to those queries determine the resulting amount of noise injected into the DHC statistics derived from those queries.

DHC-P PLB allocations by geographic level and query as reflected in the 2022-03-16 Demonstration Data Product

Global *rho*	3.325
Global *epsilon*	20.01
delta	10^{-10}

	***rho* Allocation by Geographic Level**
US	1.95%
State	27.07%
County	8.42%
Population Estimates Primitive Geography†	12.93%
Tract Subset Group‡	12.93%
Tract Subset‡	23.46%
Optimized Block Group◊	12.93%
Block	0.30%

	Per Query *rho* Allocation by Geographic Level							
Query	US	State	County	Population Estimates Primitive Geography†	Tract Subset Group‡	Tract Subset‡	Optimized Block Group◊	Block
AGE (3 bins) * HHGQ (4 Levels) (12 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
AGE (3 bins) * SEX (6 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
AGE (13 bins) * SEX (26 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
HISPANIC * SEX (4 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
SEX * HHGQ (4 levels) (8 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
HISPANIC * SEX * AGE (13 bins) * HHGQ (8 levels) * CENRACE (26,208 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
HHGQ (8 levels) * AGE (23 bins) * HISPANIC * CENRACE * SEX (46,368 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
RELGQ * AGE (23 bins) * HISPANIC * CENRACE * SEX (243,432 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%
RELGQ * SEX * AGE (116 bins) * HISPANIC * CENRACE (1,227,744 cells)	0.22%	3.01%	0.94%	1.44%	1.44%	2.61%	1.44%	0.03%

United States® Census Bureau

Harvard Data Science Review: Issue 2.2, Spring 2020 (mit.edu)

Here are my some of my lessons learned from the experience…from the spring of 2020.

United States® Census Bureau

Lesson One:

The Emerging Public Policy Debate About Privacy and Accuracy

Lesson Two:

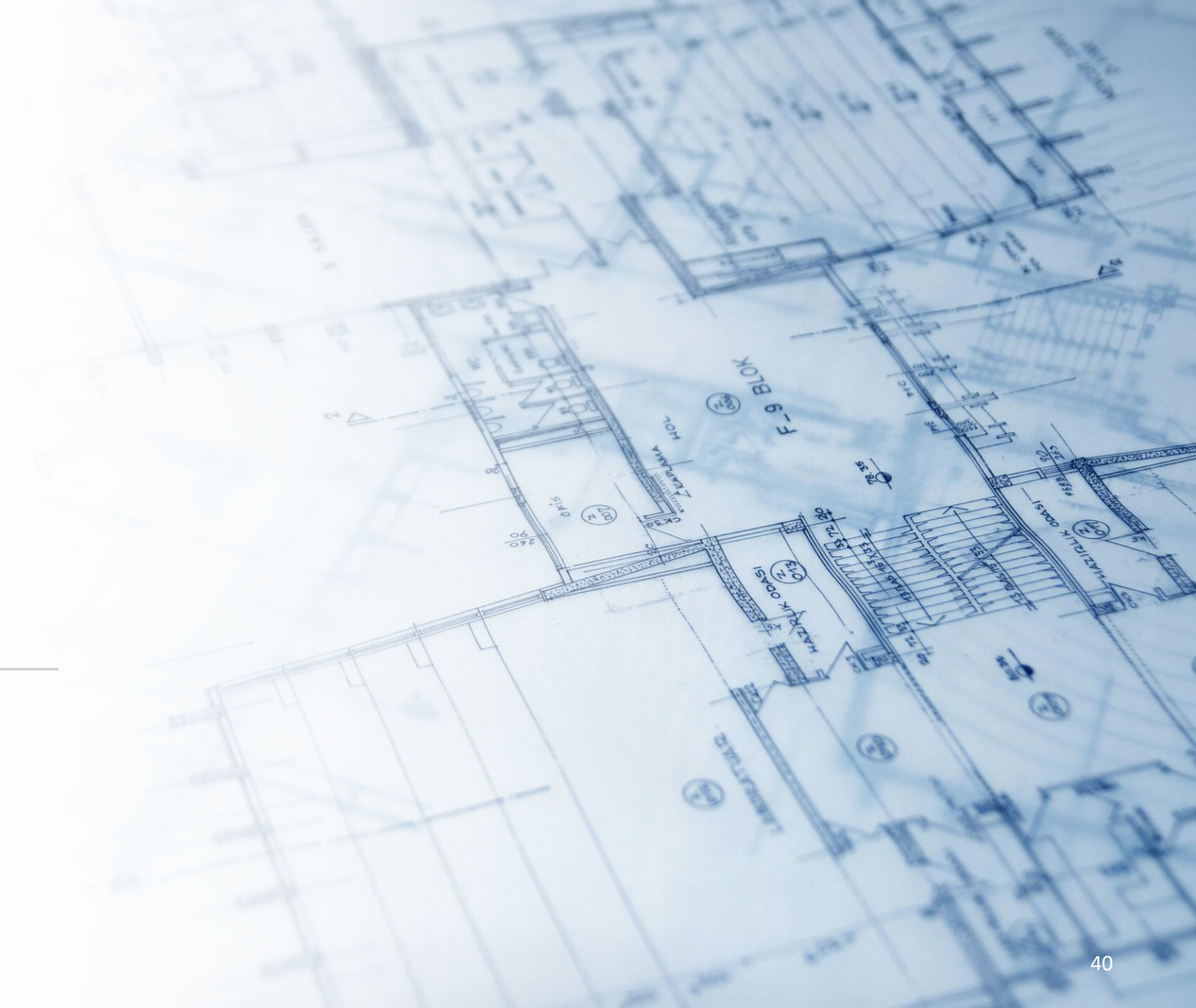
Prioritizing Accuracy* for Diverse Use Cases

United States® Census Bureau

(…and, as we'll discuss later, availability)

Lesson Three:

Choose the Right Design



Lesson Four:

The Best Laid Plans…

Lesson Five:

Rethinking Tabular Consistency and Integrality

Lesson Six: Explore Alternatives

Lesson Seven:

Remember Why We're Doing This

Reflections on disclosure protections for the 2020 Census and looking forward to 2030

United States® Census Bureau

The 2020 DAS

The 2020 DAS is a remarkable technological achievement that leveraged innovative and science-driven solutions to protect the confidentiality of America's flagship statistical product.

The Census Bureau is justifiably proud of this accomplishment.

But, the process by which we got there was not easy…

Timing

The comparatively late decision to develop a new disclosure avoidance system resulted in numerous challenges.

Responsiveness

Design, evaluation, and production should ideally run in a well-ordered sequence.

Key stages of 2020 DAS design and evaluation happened in parallel.

This led to inefficiencies and impacted the responsiveness of Census Bureau staff to stakeholder inquiries and feedback.

New and Challenging Conversations

The Census Bureau was not accustomed to public engagement on disclosure avoidance issues.

This led to inefficiencies and missteps in our stakeholder engagement, especially early in the process.

Towards a Principled Framework for Disclosure Avoidance

Michael B Hawes[1], Evan M Brassell[1], Anthony Caruso[1], Ryan Cumings-Menon[1], Jason Devine[1], Cassandra Dorius[1,2], David Evans[1,3], Kenneth Haase[1], Michele C Hedrick[1], Scott H Holan[1,4], Cynthia D Hollingsworth[1] Eric B Jensen[1], Dan Kifer[1,5], Alexandra Krause[1], Philip Leclerc[1], James Livsey[1], Roberto Ramirez[1], Rolando A Rodríguez[1], Luke T Rogers[1], Matthew Spence[1], Victoria Velkoff[1], Michael Walsh[1], James Whitehorne[1], and Sallie Ann Keller[1,3]

[1]U.S. Census Bureau*, [2]Iowa State University, [3]University of Virginia, [4]University of Missouri, [5]Penn State University

Draft, June 28, 2024

Manuscript under review by the Harvard Data Science Review

An earlier version of this manuscript was presented at the May 2024 NBER Workshop on Data Privacy Protection and the Conduct of Applied Research

Any opinions or viewpoints are the authors' own and do not reflect the opinions or viewpoints of the U.S. Census Bureau.

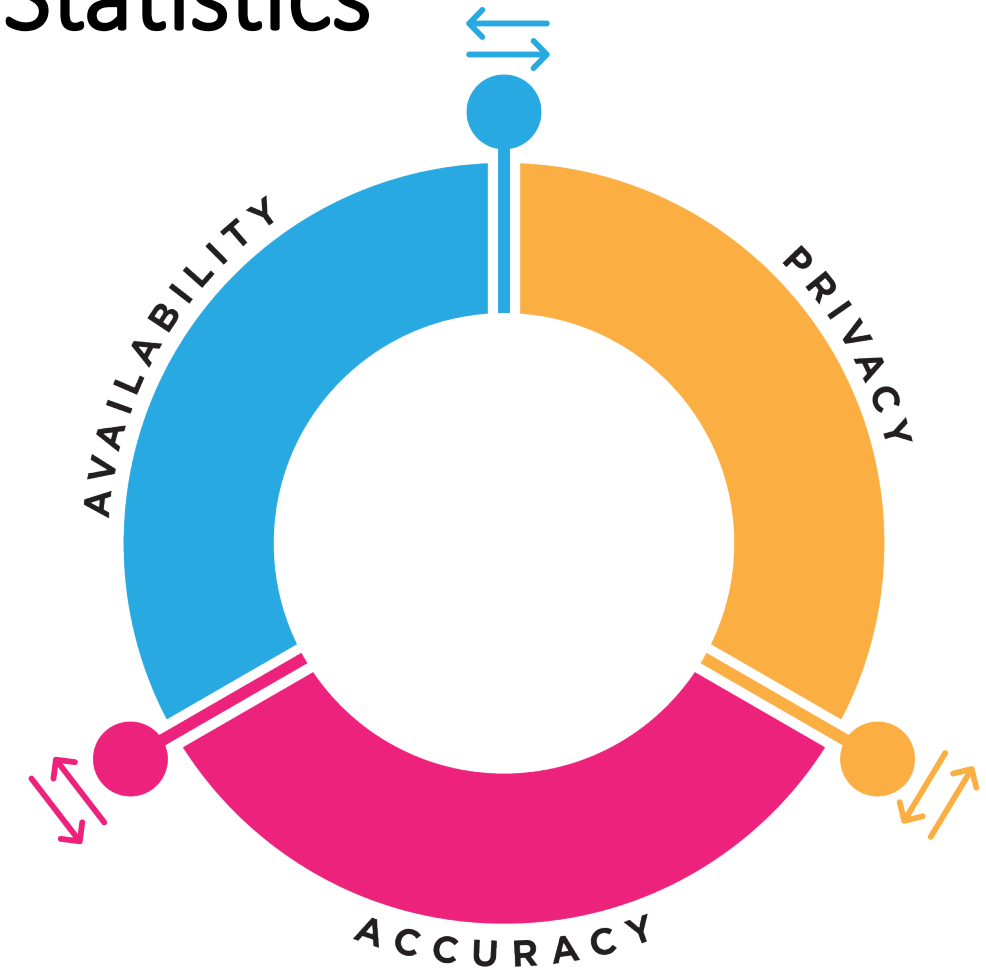
United States® Census Bureau

The Triple Trade-Off of Official Statistics

The more statistics you publish, and the greater the granularity and accuracy of those statistics, the greater the disclosure risk.

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

You can maximize on any two dimensions, but only at profound cost to the third.



United States® Census Bureau

Disclosure Avoidance Techniques and the Triple Tradeoff

The selection of a particular disclosure avoidance (DA) technique does not directly impact agency decision-making within the context of the triple tradeoff. Nearly any DA technique can be applied to implement very different balances along these three dimensions, depending on the implemented parameters selected.

Example DA Techniques	Examples of Parameters that Implement the Triple-Tradeoff
Suppression	Cell size thresholds, p% rules
Coarsening	Rounding rules (e.g., 3, 10, 1000)
Swapping	Swap keys, rates, geographies
Differential Privacy	Privacy-loss budgets and allocations

United States® Census Bureau

Objective

We need a set of overarching principles that an ideal, applied disclosure avoidance system should meet…

…while distinguishing those principles from any choices relating to the implementation of that system.

United States® Census Bureau

What is a Disclosure Avoidance System?

A Disclosure Avoidance System is a set of one or more statistical methods that transform confidential information (or data derived from confidential information) from or about individual data subjects into statistics that describe, estimate, or analyze the characteristics of groups, without identifying the data subjects that comprise such groups.

Disclosure Avoidance Systems accomplish this through the application of statistical disclosure limitation techniques to reduce (but not eliminate) disclosure risk in the statistical products being produced.

United States® Census Bureau

What is an Applied Disclosure Avoidance System?

An applied Disclosure Avoidance System is one that performs within the **operational realities and production cycles of a national statistical office**. As such, it acknowledges and is **adaptable to requirements stemming from the legal, policy, scientific, resource, and stakeholder environments** within which it is operating.

United States® Census Bureau

What is an Ideal, Applied, Disclosure Avoidance System?

An ideal, applied, Disclosure Avoidance System is one that **conforms to a set of overarching principles or features relating to the efficiency, effectiveness, and flexibility** of the system as it transforms confidential information from (or about) data subjects into **quality statistics** for public release.

United States® Census Bureau

Distinguishing these Principles from Implementation Choices

Principles

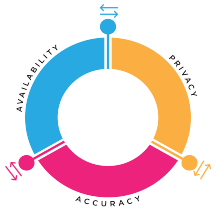
Reflect characteristics that all DA systems should ideally have, regardless of the specific technology or disclosure limitation mechanism being employed.

Should be universally applicable regardless of the type, format, or context of data being protected.

Implementation

Reflect specific choices about the appropriate balance between data availability, utility, and confidentiality.

Should be informed by the characteristics of the data, the context of the statistical product, and the intended objectives and requirements of the agency and its stakeholders.



United States® Census Bureau

IMPORTANT

The Census Bureau's mission is to produce high quality statistics and statistical products.

Title 13's confidentiality protections were established in support of that mission.

We must remember that disclosure avoidance is a legal obligation and a necessary activity, but it is one that we undertake in tandem with, and in support of, our primary mission to produce quality statistics.

Produce Quality Statistics

Protect Confidentiality

Disclosure Avoidance System

Implementation Choices

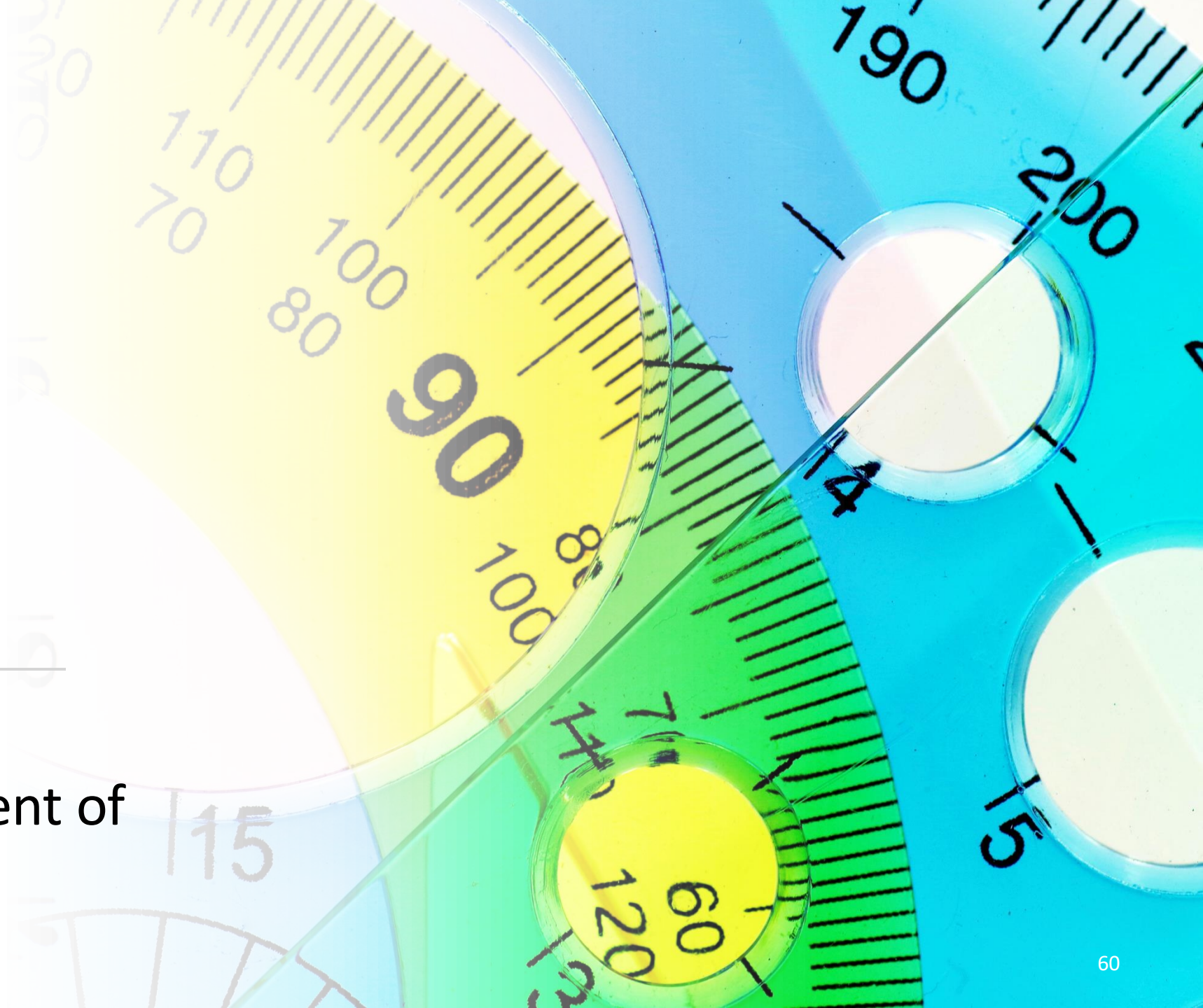
United States® Census Bureau

Characteristics of an Ideal, Applied Disclosure Avoidance System

United States® Census Bureau

Principle #1

It should support meaningful assessment of disclosure risk

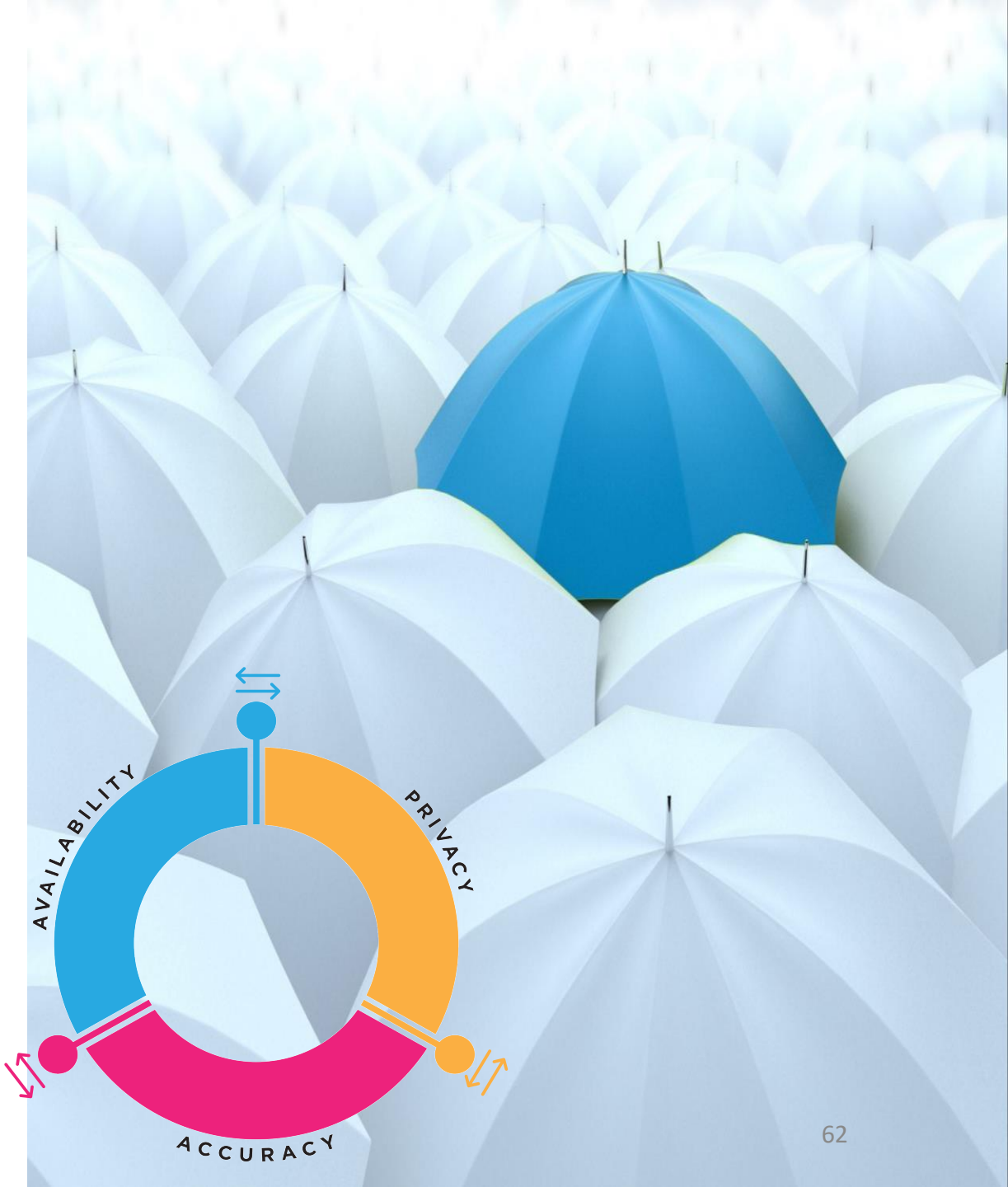


Principle #2

It should support meaningful assessment of the impact of data protections on the quality of statistics

Principle #3

It should be able to target protection



Principle #4

It should be able to target data quality

Principle #5

It should track cumulative disclosure risk over time

United States® Census Bureau

Principle #6

It should be transparent

Principle #7

It should be feasible

Distinguishing these Principles from Implementation Choices

Principles

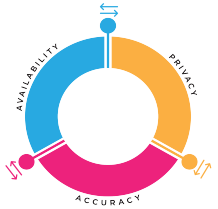
Reflect characteristics that all DA systems should ideally have, regardless of the specific technology or disclosure limitation mechanism being employed.

Should be universally applicable regardless of the type, format, or context of data being protected.

Implementation

Reflect specific choices about the appropriate balance between data availability, utility, and confidentiality.

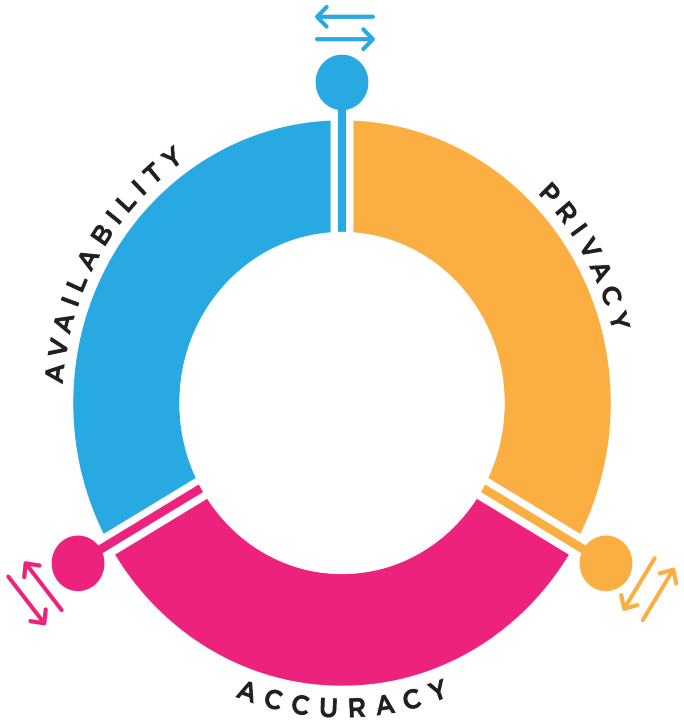
Should be informed by the characteristics of the data, the context of the statistical product, and the intended objectives and requirements of the agency and its stakeholders.



United States® Census Bureau

Examples of implementation choices independent of the selection of a DAS

Desired balance of accuracy, confidentiality, and availability of statistics	Targeting and prioritizing of confidentiality protections	If any data elements should be excluded from protection
Prioritization of use cases	Statistical product design requirements	Operational considerations



The ideal DAS should:

- support meaningful assessment of disclosure risk;
- support meaningful assessment of impact on quality of statistics;
- be able to target protection;
- be able to target data quality;
- track cumulative disclosure risk over time;
- be transparent;
- be feasible.

Using the Principles

- There can be tension between these principles

United States® Census Bureau

Questions and Discussion