Disclosure Avoidance for the 2020 Census

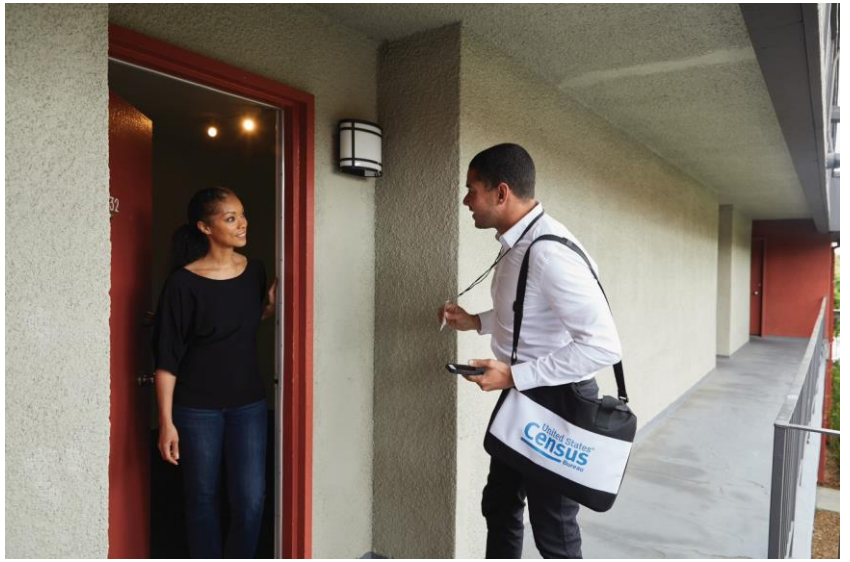
Michael Hawes
Cynthia Hollingsworth

September 28, 2021

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.



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.

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 can perform complex, large-scale calculations with increasing ease.

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

Re-identification

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

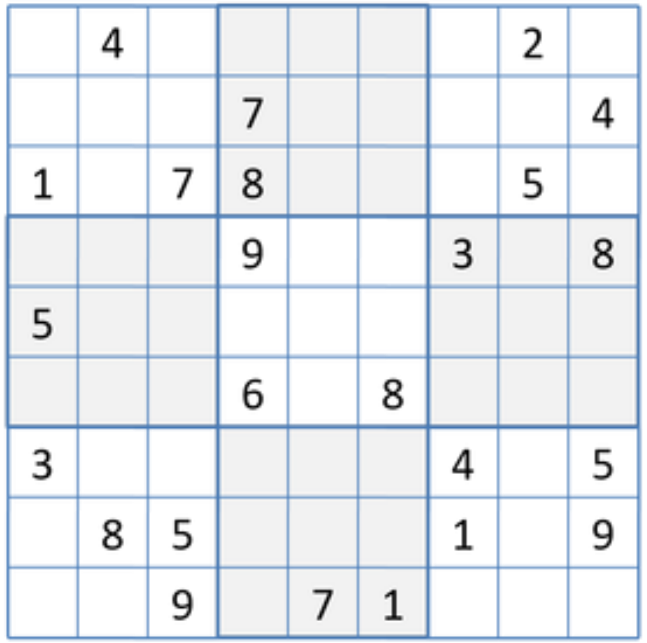
Confidential Data

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			

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.

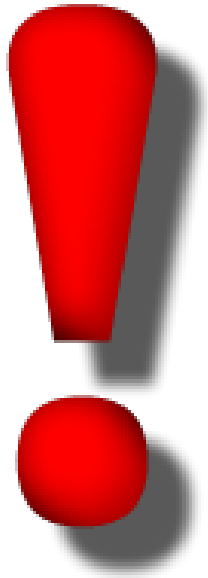
Reconstructing the 2010 Census: What Did We Find?

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 reconstructed microdata are a close approximation of the Hundred Percent Detail (HDF) file and violate the disclosure avoidance rules for microdata in place for the 2010 Census.

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.



Disclosure Avoidance

Disclosure avoidance methods seek to make reconstruction and re-identification more difficult, by:

- Reducing precision
- Removing vulnerable records, or
- Adding uncertainty

Commonly used (legacy) methods include:

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

Problem #1 – Impact on Data

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

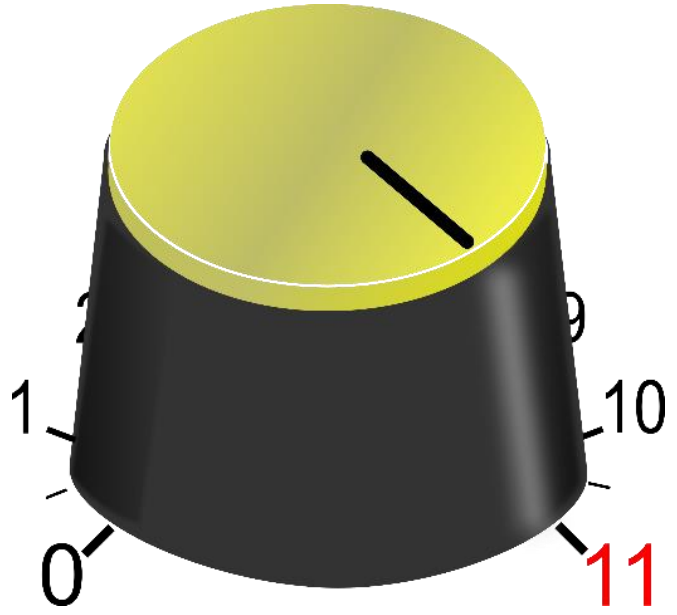
Swap rates, noise injection parameters, cell suppression thresholds, etc. determine this tradeoff.

Problem #2 – How much is enough?

Traditional disclosure avoidance methods provide little ability to quantify privacy protections, *especially across multiple data releases from the same confidential source.*

When faced with rising disclosure risk, disclosure avoidance practitioners adjust their implementation parameters.

BUT, this is largely a scattershot solution that over-protects some data, while often under-protecting the most vulnerable records.



Differential Privacy

DP is not a disclosure avoidance "method" as much as it is a framework for defining and then quantifying 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.

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 data accuracy for any given level of privacy protection.*

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

2020 Census Data Products

"Group I Products"

- P.L. 94-171 Redistricting Data Summary File
- Demographic Profile
- Demographic and Housing Characteristics File

"Group II Products"

- Detailed Demographic and Housing Characteristics File

"Group III Products"

TBD, may include:

- Public Use Microdata
- Special Tabulations
- FSRDC Access
- Out-year uses of 2020 Census data

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"

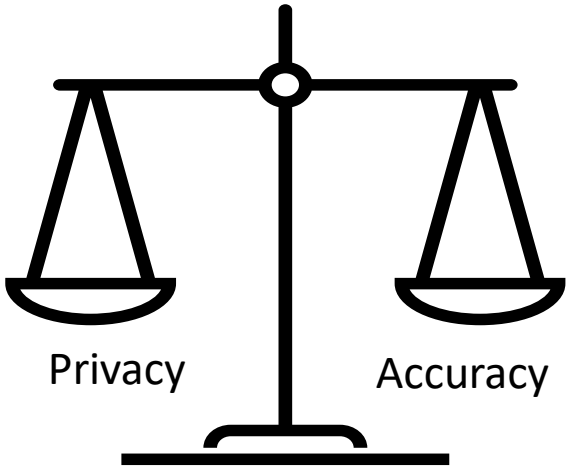
SafeTab
PHSafe

Produce privacy-protected tabulations directly

"Group III Products"

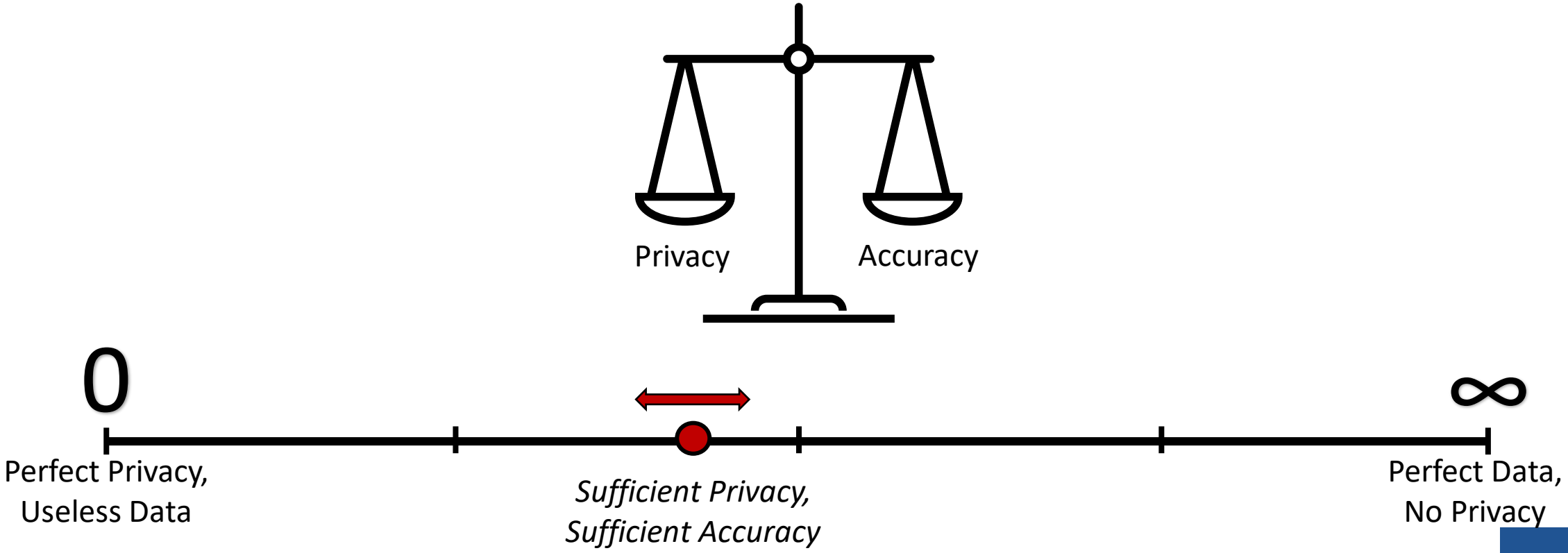
TDA
SafeTab
PHSafe
or other formally privacy solutions

What is a Privacy-loss Budget?

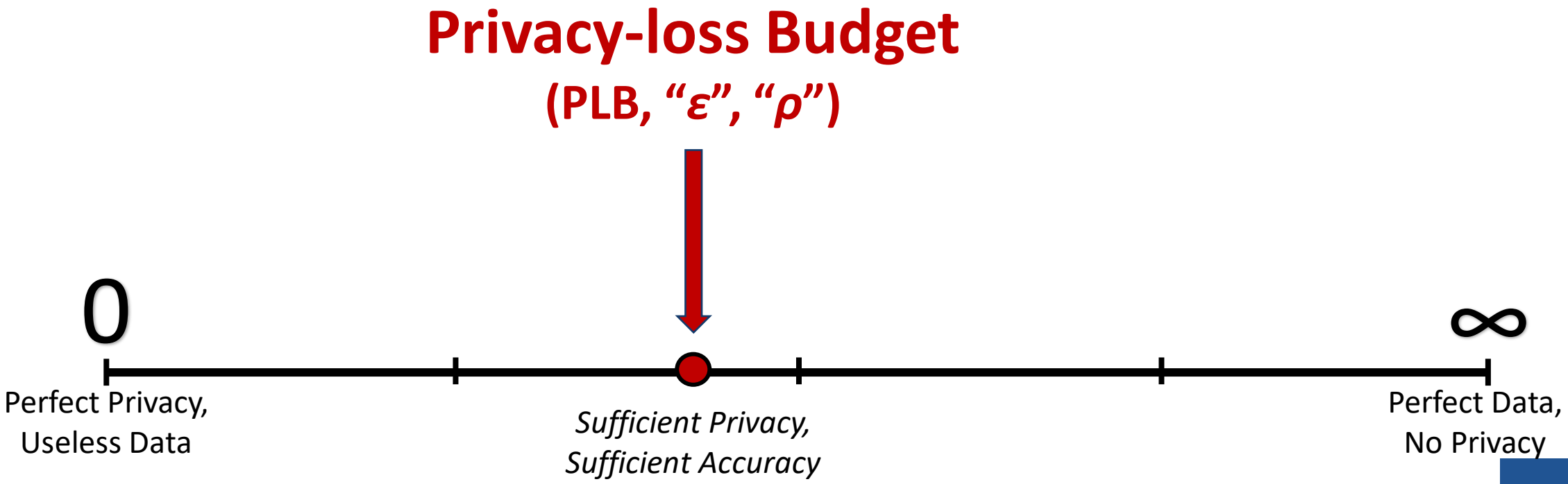


Any disclosure avoidance mechanism imposes a fundamental tradeoff between data protection (privacy/confidentiality) and data accuracy/fitness-for-use.

What is a Privacy-loss Budget?



What is a Privacy-loss Budget?



Allocating Privacy Loss Budget (PLB) by Data Product

PL and DHC products were split apart and will be protected using separate privacy loss budgets

"Group I PL Product"

"Group I DHC Product"

"Group II Product"

"Group III Products"

PLB_1 + PLB_2 + PLB_3 + PLB_4

= Global PLB

Allocating PLB *within* Data Products

	***rho* Allocation by Geographic Level**
US	104/4099
State	1440/4099
County	447/4099
Tract	687/4099
Optimized Block Group*	1256/4099
Block	165/4099

Query	**Per Query *rho* Allocation by Geographic Level**					
	US	State	County	Tract	Optimized Block Group*	Block
TOTAL (1 cell)		3773/4097	3126/4097	1567/4102	1705/4099	5/4097
CENRACE (63 cells)	52/4097	6/4097	10/4097	4/2051	3/4099	9/4097
HISPANIC (2 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
VOTINGAGE (2 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HHINSTLEVELS (3 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HHGQ (8 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HISPANIC*CENRACE (126 cells)	130/4097	12/4097	28/4097	1933/4102	1055/4099	21/4097
VOTINGAGE*CENRACE (126 cells)	130/4097	12/4097	28/4097	10/2051	9/4099	21/4097
VOTINGAGE*HISPANIC (4 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
VOTINGAGE*HISPANIC*CENRACE (252 cells)	26/241	2/241	101/4097	67/4102	24/4099	71/4097
HHGQ*VOTINGAGE* HISPANIC*CENRACE (2,016 cells)	189/241	230/4097	754/4097	241/2051	1288/4099	3945/4097

Within each data product, shares of the PLB are then allocated to each statistic calculated.

The PLB shares for each element are set through tuning informed by engagement with our data users.

Engagement on the Demographic and Housing Characteristics (DHC) File

- Engagements will help us make informed decisions about DHC File production
- Engagement plan includes a 3-pronged approach:
  - Engagement/Education (where are we going with differential privacy and the DHC, how do we get there)
  - Feedback/Listening (how DHC tables are used, for what purpose, can DAS tuning support)
  - Demonstration/Implementation (are data fit-for-use, what are the privacy – accuracy tradeoffs)
- Demonstration data using 2010 Census will be released to enable public to assess accuracy and privacy protection of DHC tables
- Demonstration data release and stakeholder engagement will be transparent and timely
  - At least two rounds of demonstration data releases
  - Minimum of 30 days for review and feedback period
  - Clear feedback guidelines

Ongoing Engagement

- Plan to continue external engagements with advisory and stakeholder groups, such as:
  - Census Scientific Advisory Committee (CSAC) and National Advisory Committee on Race (NAC)
  - CSAC and NAC Differential Privacy (DP) Working Groups
  - American Indian and Alaska Native Tribal Leaders
  - Committee on National Statistics (CNSTAT)
  - State Data Center (SDC) and Census Information Center (CIC) networks
  - Federal agency partners
  - Congressional committees and staff
  - And more …
- Internal engagements such as Town Halls and launch of Share Point site
- Plans are still being developed/discussed regarding external engagement on the Detailed DHC Product

Notional Timeline for DHC Development

Now – Fall 2021

- Release Updated 2020 Census Data Products Crosswalk File
- Collect feedback on potential changes to DHC crosswalk (e.g., reduced number of block level tables)
- Collect final feedback on the design of DHC tables
- Confirm DHC use cases
- Complete DAS development/feasibility testing (experiments) based on DHC specifications and accuracy targets
- Conduct internal review of initial DAS implementation
- Reassess timeline and communicate status

Winter 2021/2022

- Create and release first set of demonstration data
- Collect feedback and data fit-for-use assessments
- Assess feedback, incorporate changes to DAS implementation, and conduct internal review
- Reassess timeline and communicate status

Notional Timeline for DHC Development Cont.

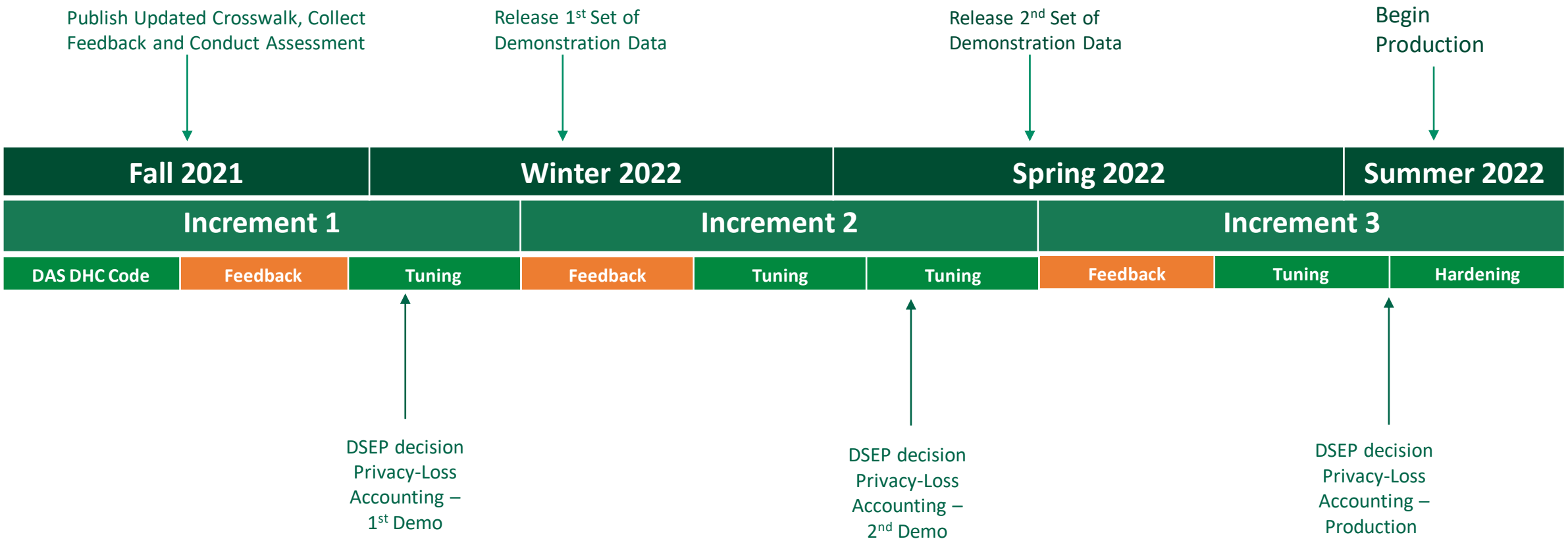
Spring 2022

- Create and release second set of demonstration data
- Conduct Public Workshop
- Collect feedback and assessments
- Assess feedback and determine completion of development

Summer 2022

- DSEP sets final parameters for DHC
- DHC production begins

Notional Timeline for DHC Development Cont.



Questions and Discussion

Supplementary Slides

TDA System Requirements

The 2020 Disclosure Avoidance System's TopDown Algorithm (TDA) will implement formal privacy protections for the P. L. 94-171 Redistricting Data Summary File, Demographic Profiles, Demographic and Housing Characteristics, and Special Tabulations of the 2020 Census.

TDA system requirements include:

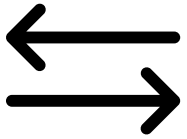
- Input/Output specifications
- Invariants
- Constraints
- Utility/Accuracy for pre-specified tabulations
- ϵ-asymptotic consistency
- Transparency

TDA Process Snapshot

What is a histogram?

Record ID	Block	Race	…	Sex
1	1001	Black	…	Male
2	1001	Black	…	Male
3	1001	Asian	…	Female
4	1001	Asian	…	Female
5	1001	Black	…	Male
6	1001	AIAN	…	Female
7	1001	AIAN	…	Male
8	1001	Black	…	Female
9	1001	Black	…	Female

Microdata: One record per respondent



Attribute Combination (Block/Race/…/Sex)	# of Records
1001/AIAN/…/Male	1
1001/AIAN/…/Female	1
1001/Asian/…/Male	0
1001/Asian/…/Female	2
1001/Black/…/Male	3
1001/Black/…/Female	2
…	…

Histogram: Record count for each unique combination of attributes (including location)

Noisy Measurements

TDA allocates shares of the total privacy-loss budget by geographic level and by query.

Each query of the confidential data will have noise added to its answer.

The noise is taken from a probability distribution with mean=0, and variance determined by the share of the PLB allocated to that particular query at that geographic level.

These noisy measurements are independent of each other, and can include negative values, hence the need for post-processing.

What is noise?

To protect privacy, TDA randomly adds or subtracts a small amount from each statistic it calculates from the confidential data.

Attribute Combination (Block/Race/…/Sex)	# of Records
1001/AIAN/…/Male	1
1001/AIAN/…/Female	1
1001/Asian/…/Male	0
1001/Asian/…/Female	2
1001/Black/…/Male	3
1001/Black/…/Female	2
…	…

Total: 9+0=9

Male: 4+0=4

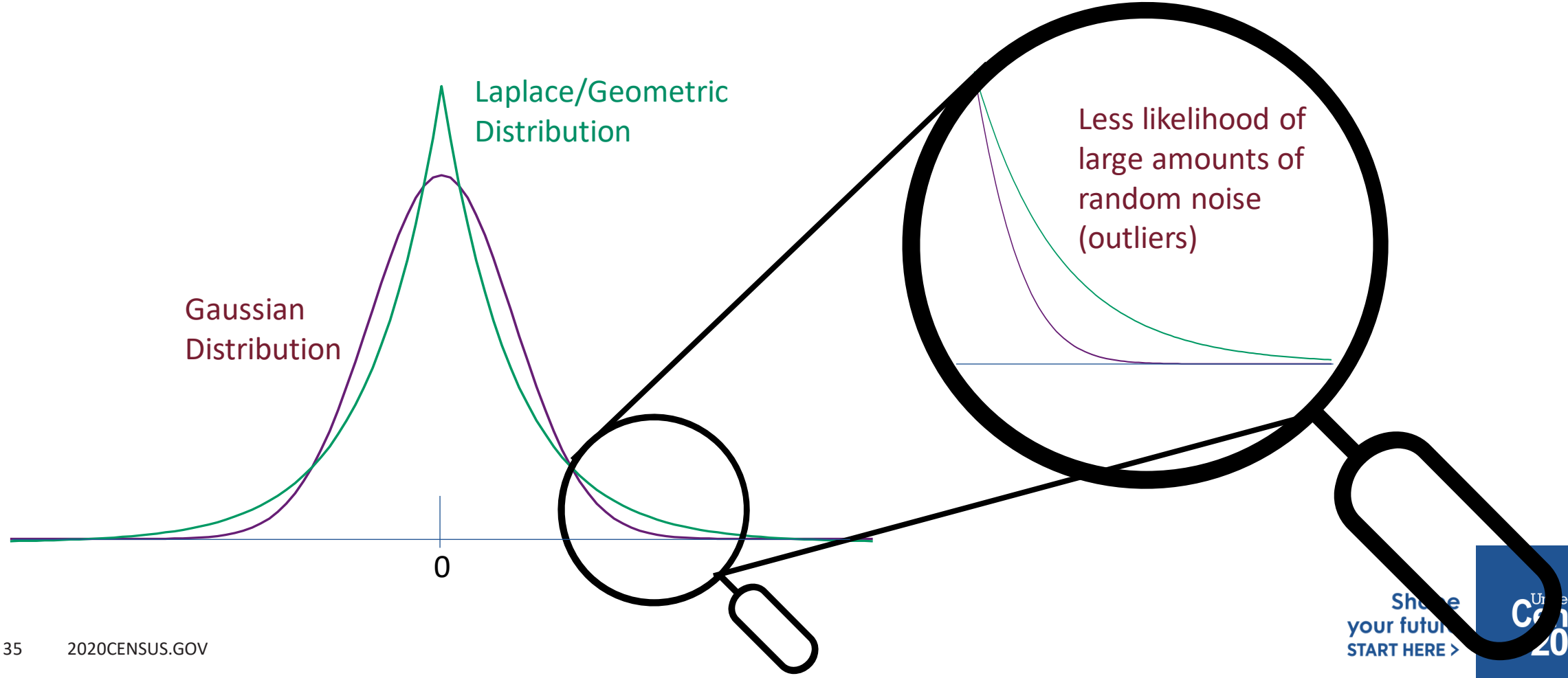
Female: 5-1=4

#AIAN: 2+0=2

#Asian: 2+2=4

#Black: 5-1=4

Zero-Concentrated Differential Privacy (zCDP)



Understanding *epsilon*, *delta* and *rho*

In traditional (ε,0) differential privacy:

The privacy-loss parameter ε (*epsilon*) sets the upper-bound on how much information leakage can occur.

Shares of ε are allocated to each query and sum to the global value of ε.

In zero-concentrated differential privacy (zCDP):

Privacy loss is quantified by the paired parameters ε and δ (delta).

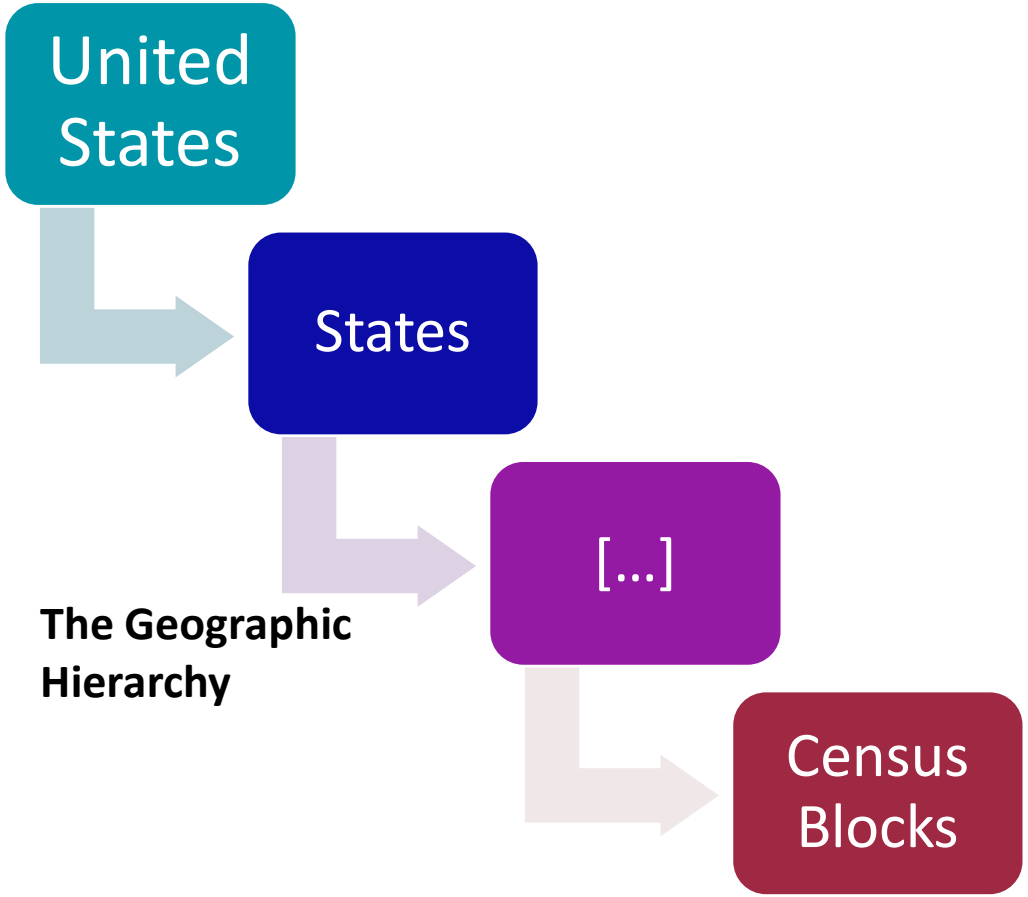
δ is a probabilistic term that establishes the likelihood that privacy loss might exceed the upper bound represented by a particular value of ε.

Within the mechanics of zCDP, privacy-loss budget is allocated to queries by shares of a third parameter, ρ (*rho*).

The global ρ can then be used to calculate the global ε for any given level of δ.

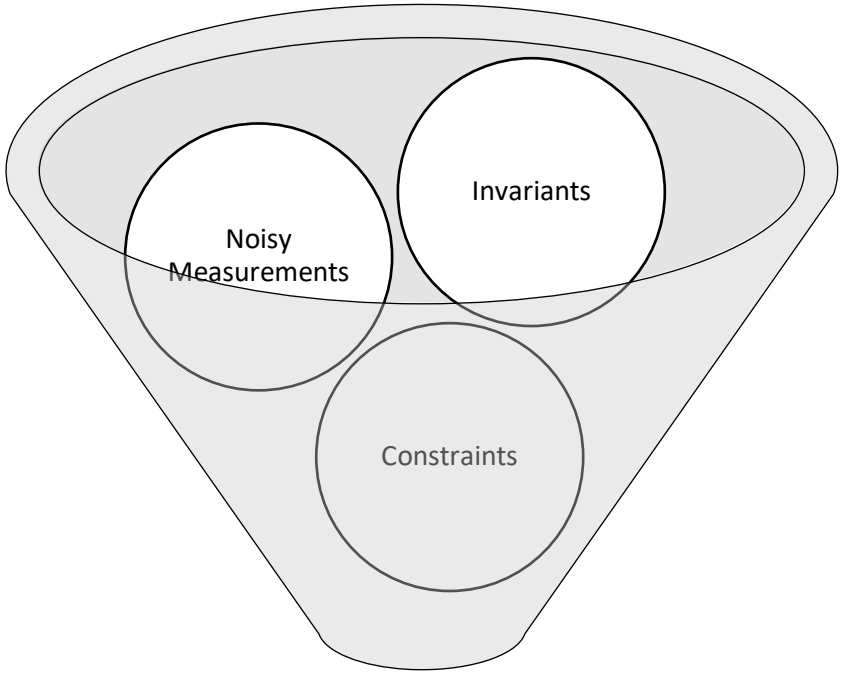
The Census Bureau's privacy accounting uses a value of $\delta=10^{-10}$ so our published values of ε should be interpreted accordingly.

The TopDown Approach



The Geographic Hierarchy

At each geographic level:

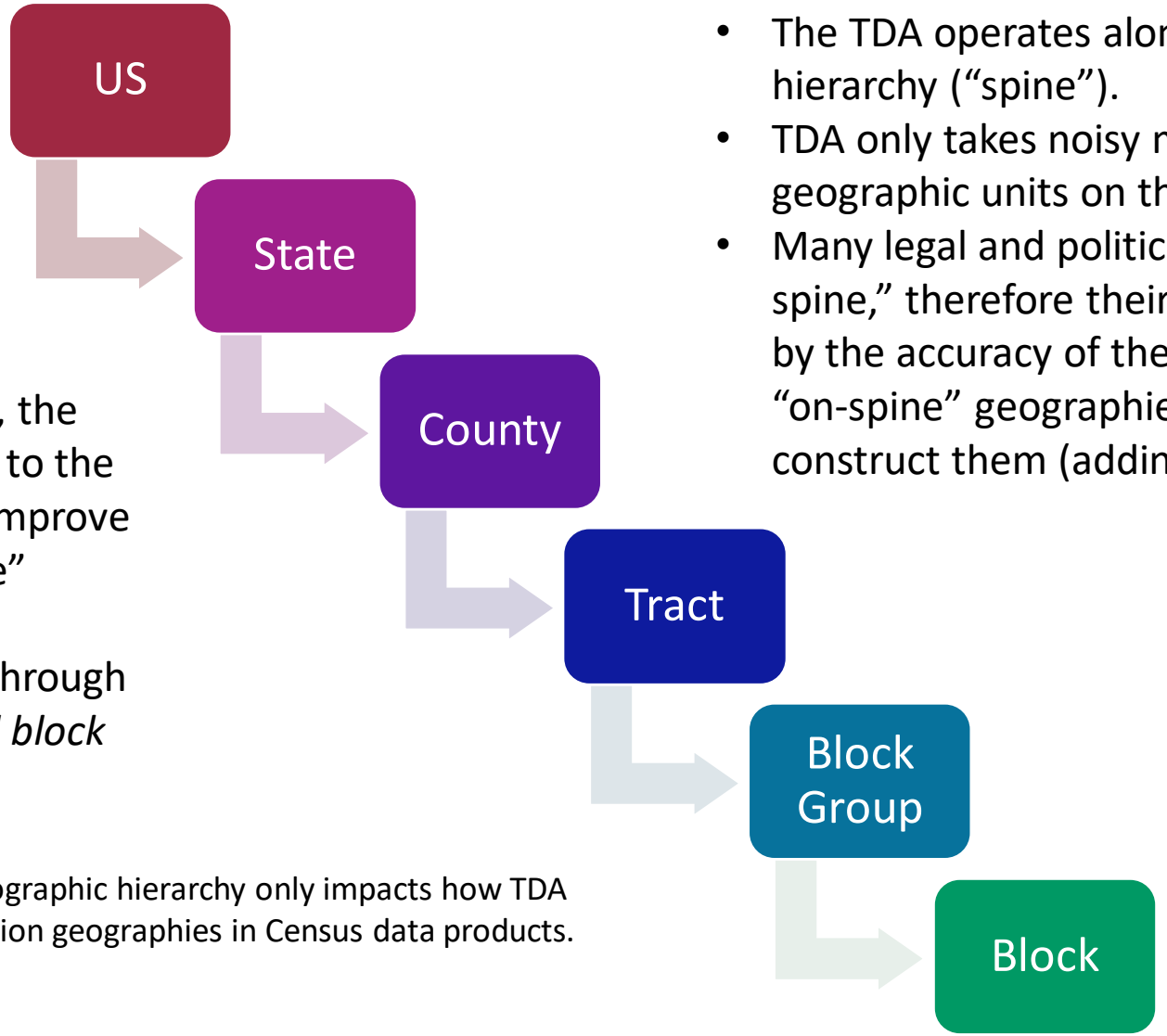


Internally consistent histogram

Benefits of TDA Compared to Block-by-block

- TDA is in stark contrast with naïve alternatives (e.g., block-by-block or bottom-up)
- TDA disclosure-limitation error does not increase with number of contained Census blocks in the geographic entity
- TDA yields increasing relative accuracy as the population being measured increases (in general), and increased count accuracy compared to block-by-block
- TDA “borrows strength” from upper geographic levels to improve count accuracy at lower geographic levels (e.g., for sparsity)

Tabulation Geographic Hierarchy



- The TDA operates along a geographic hierarchy ("spine").
- TDA only takes noisy measurements for geographic units on the hierarchy.
- Many legal and political geographies are "off-spine," therefore their accuracy is impacted by the accuracy of the minimum number of "on-spine" geographies that can be used to construct them (adding or subtracting).

- To address this challenge, the DAS Team made changes to the geographic hierarchy to improve the accuracy of "off-spine" geographies.
- This was done primarily through the creation of *optimized block groups (not shown).*

Note: The optimization of the geographic hierarchy only impacts how TDA operates. It will not affect tabulation geographies in Census data products.