

Differential Privacy: What GIS Users Need To Know

John M. Abowd

U.S. Census Bureau

Esri Webinar July 22, 2020

The views expressed in this presentation are those of the speaker not 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.



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.

Privacy protection out of the shadows

- Certain privacy practices for previous censuses depended upon obfuscation
- DAS demonstration data are the most transparent view into Census Bureau privacy practices ever
- We appreciate and are excited to assess feedback from our external partners

Distance/Direction Analysis

Work to Home

Display Settings

Labor Market Segment Filter: All Workers

Year: 2017

Map Controls

Color Key

- Thermal Overlay
- Point Overlay
- Selection Outline
- Identify
- Clear Overlays
- Zoom to Selection
- Animate Overlays

Report/Map Outputs

- Detailed Report
- Export Geography
- Print Chart/Map

Legends

- 5 - 99 Jobs/Sq.Mile
- 100 - 382 Jobs/Sq.Mile
- 383 - 854 Jobs/Sq.Mile
- 855 - 1,514 Jobs/Sq.Mile
- 1,515 - 2,364 Jobs/Sq.Mile
- 1 - 3 Jobs
- 4 - 18 Jobs
- 19 - 59 Jobs
- 60 - 139 Jobs
- 140 - 271 Jobs

Analysis Selection

Analysis Settings

Analysis Type: Distance/Direction

Selection area as: Work

Year(s): 2017

Job Type: Primary Jobs

Selection Area: Ithaca, NY from Metropolitan/Micropolitan Areas (CBSA)

Selected Census Blocks: 3,082

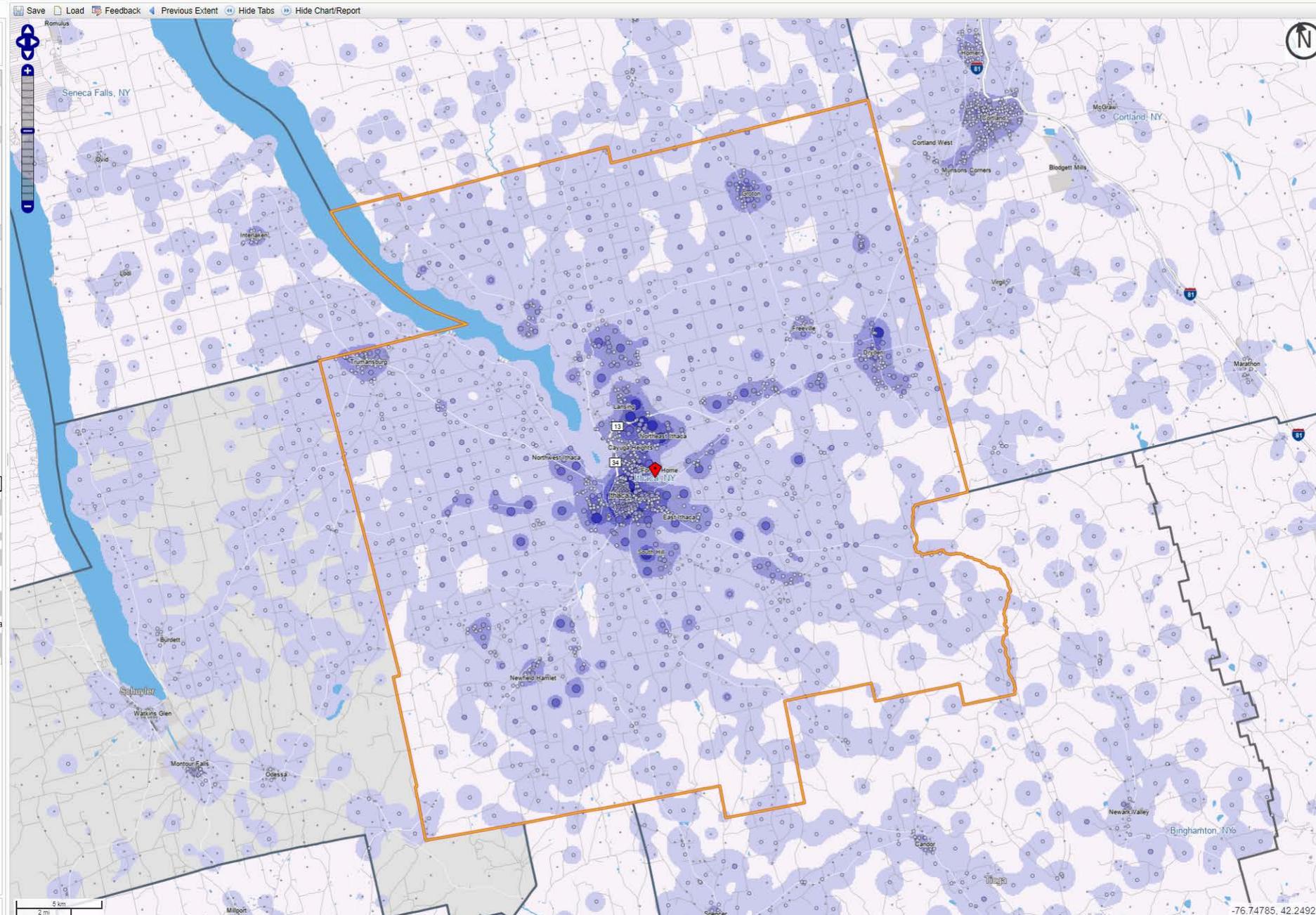
Analysis Generation: 07/20/2020 16:43 - OnTheMap 6.6

Date Code: d7f8a300c9f4e458f61bc73d3099ca2cb8f8fea

Revision: 20170818

LODES Data Version: 20170818

[Change Settings](#)



Job Counts by Distance/Direction in 2017

All Workers

View as: Radar Chart

Jobs by Distance - Work Census Block to Home Census Block

Distance	Count	Share
Total Primary Jobs	45,225	100.0%
Less than 10 miles	21,905	48.4%
10 to 24 miles	11,790	26.1%
25 to 50 miles	5,255	11.6%
Greater than 50 miles	6,275	13.9%

Census TopDown Algorithm (TDA): Requirements and Properties I

TDA is the primary formally private 2020 Census disclosure limitation algorithm under development

Inputs:

- Post-edits-and-imputation microdata records (Census Edited File – CEF)
- Required structural zeros and data-dependent invariants

Processing:

- Convert CEF to an equivalent histogram
- Apply DP measurements and perform mathematical optimizations
- Create noisy histogram; convert back to microdata

Output:

Return the Microdata Detail File (the MDF; microdata with same schema as CEF)

Example:

- Schema: Geography × Ethnicity × Race × Age × Sex × HHGQ
- This product yields a “histogram” (fully saturated contingency table)
- With shape: $\approx 8M \times 2 \times 63 \times 116 \times 2 \times 43 = \approx 8M \times 1.25M$

Census TDA: Requirements and Properties II

Data-dependent invariants:

Properties of true data that must hold exactly (*no noise*)

Current data-dependent invariants:

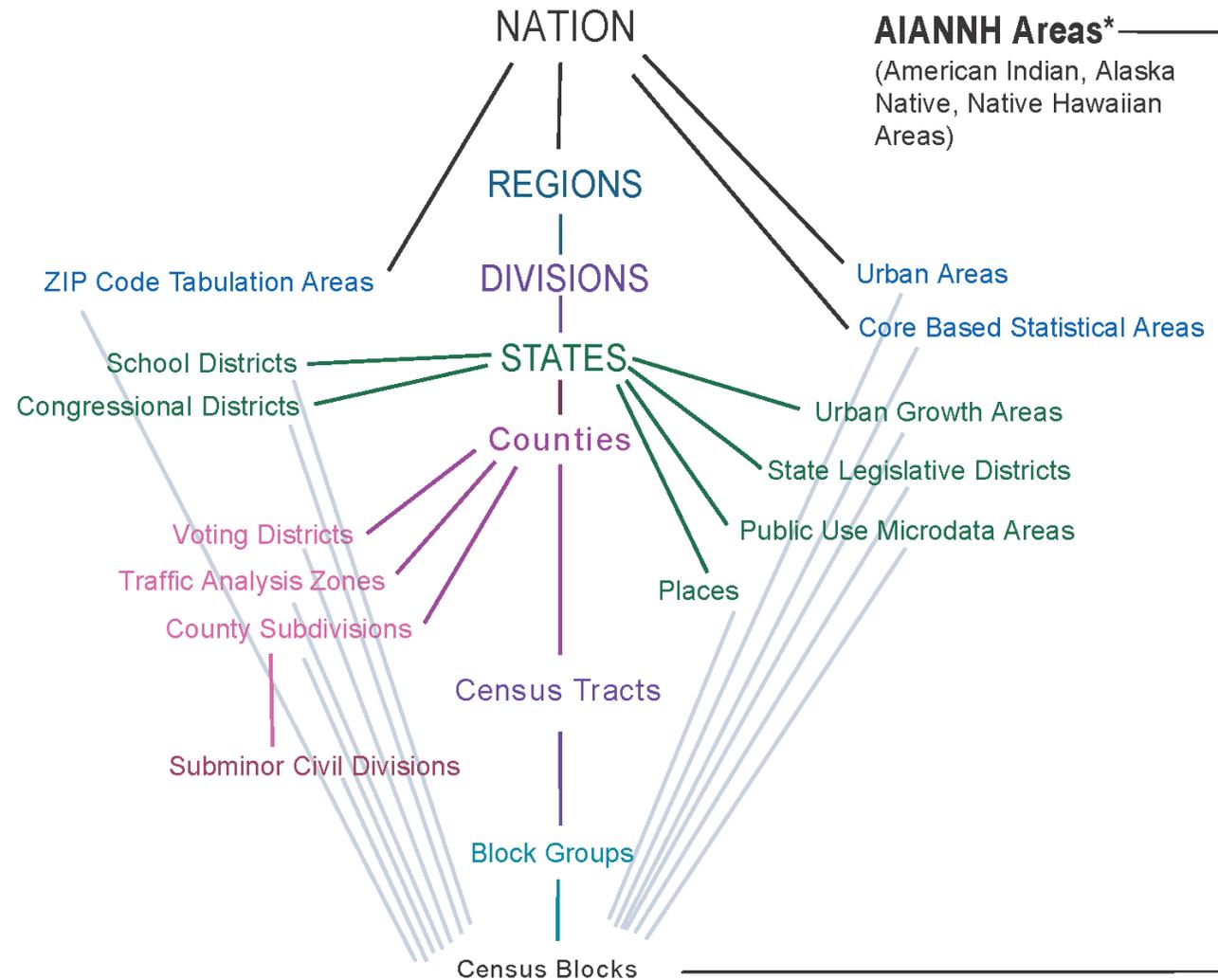
- State population totals
- Count of occupied GQ facilities by type by block (not population)
- Total count of housing units by block (not population)

Utility/Accuracy for pre-specified tabulations

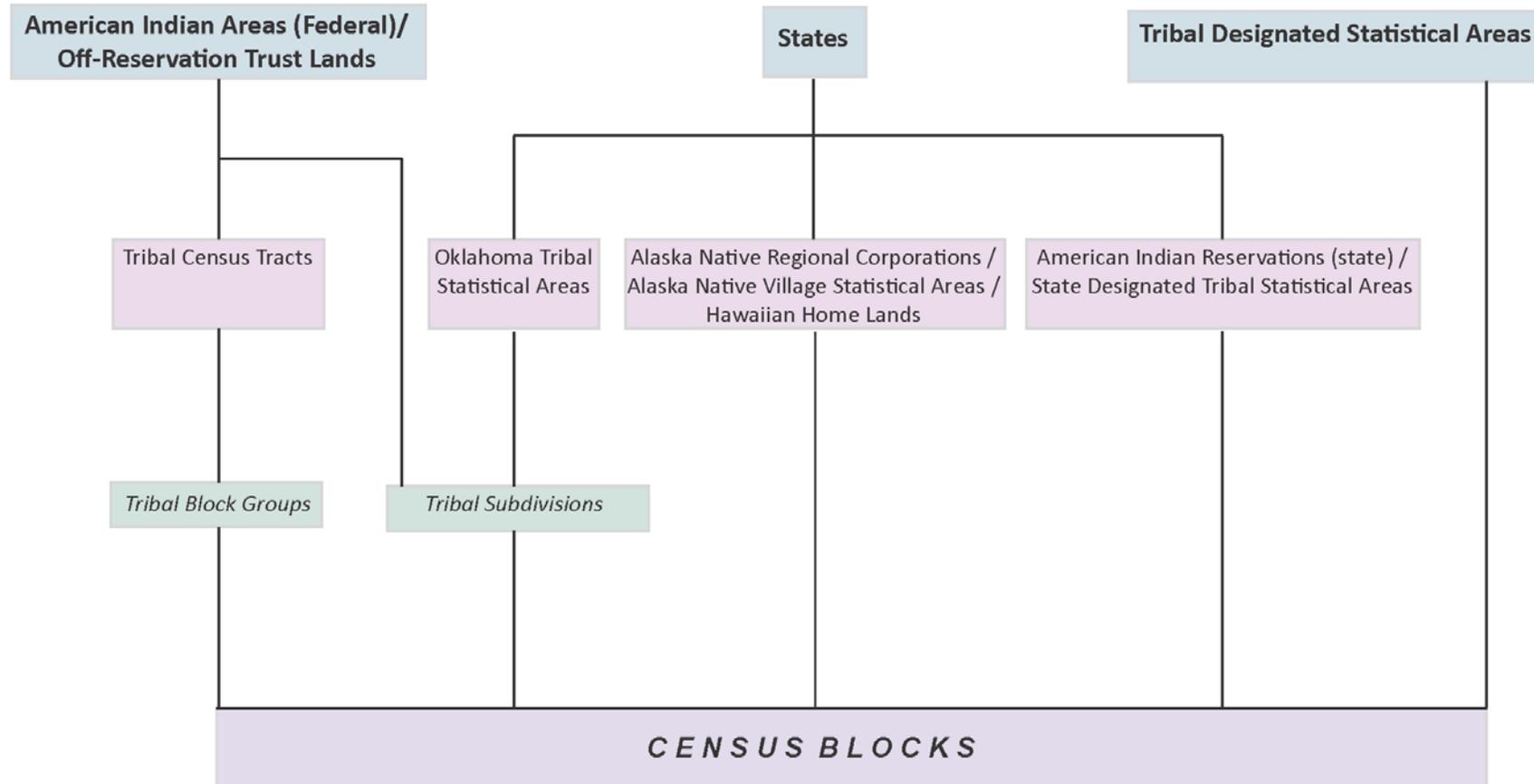
- Full privacy + full accuracy for arbitrary uses = impossible
- P.L. 94-171: tabulations used for redistricting
- Demographic and Housing Characteristics File
 - Principal successor to 2010 Summary File 1
 - TDA creates separate Person and Housing Unit microdata sets

ϵ -consistency: error $\rightarrow 0$ as privacy loss $\epsilon \rightarrow \infty$

Transparency: source code and parameters made public



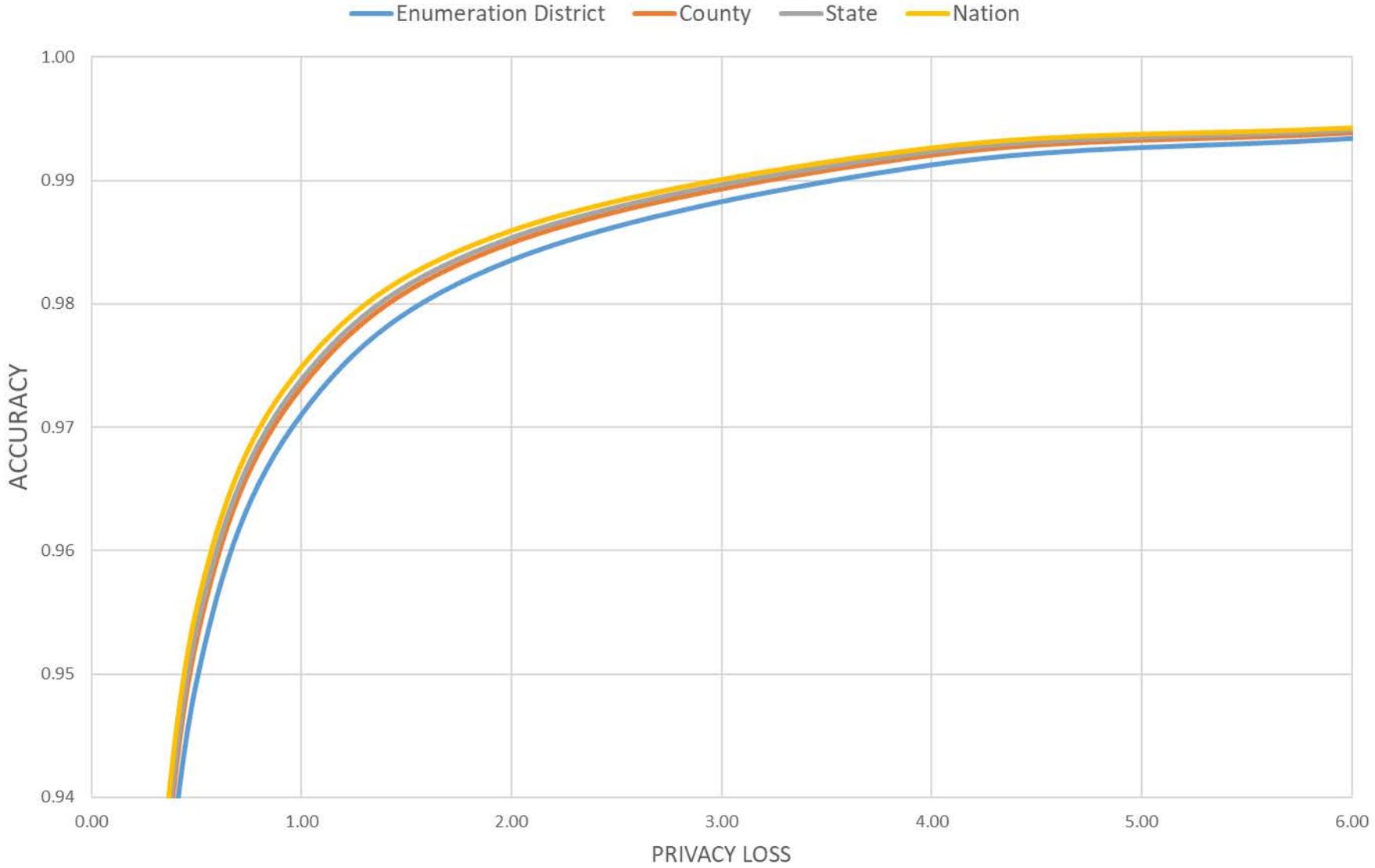
Hierarchy of American Indian, Alaska Native, and Native Hawaiian Areas



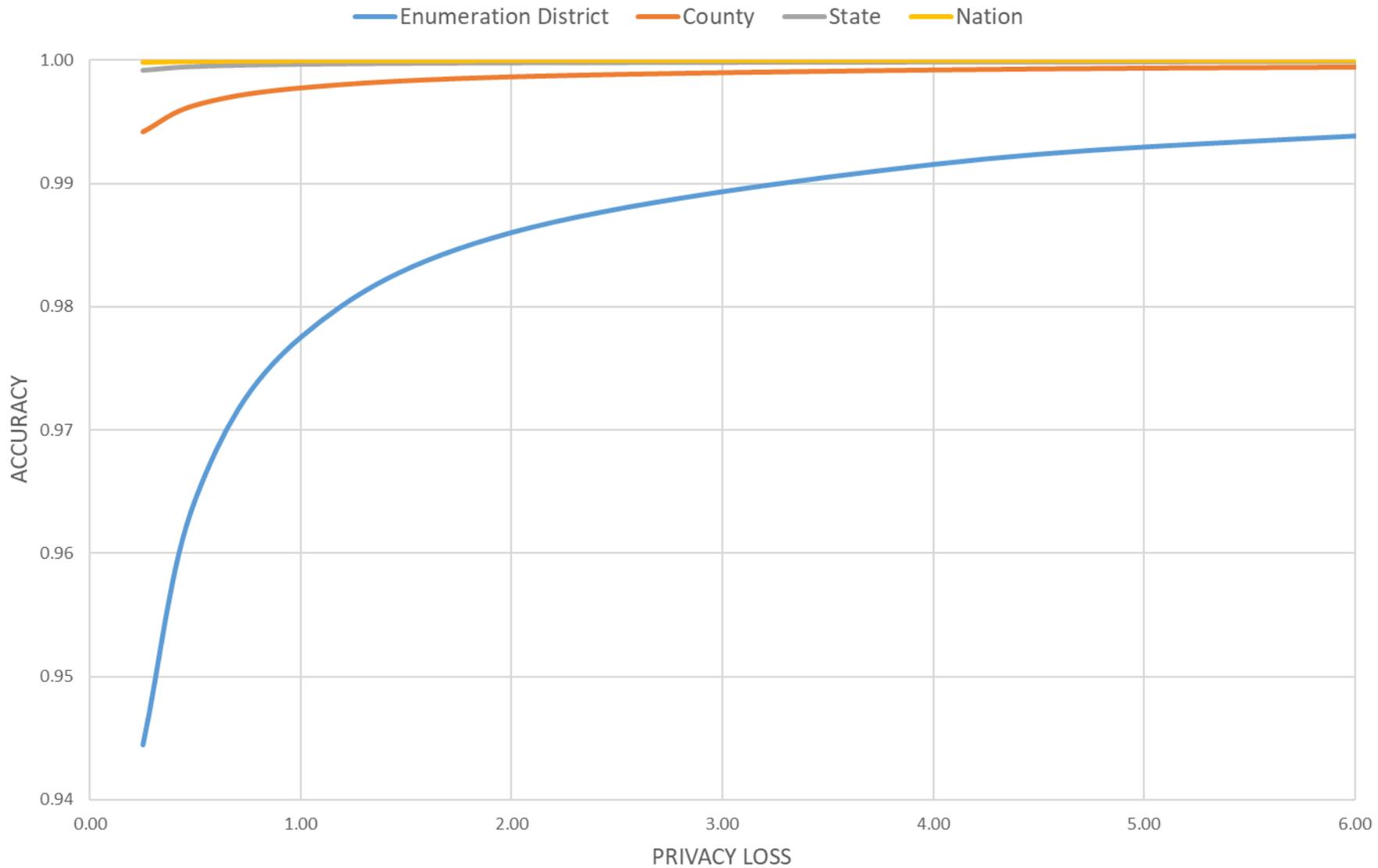
Accurate, but to whom?

- DAS operates under interpretable formal privacy guarantees, given privacy-loss budgets
- Accuracy properties depend upon the output metric (use case)
- Distinct groups of data users will have a particular analyses they wish to be accurate
- Tuning accuracy for a given analysis can reduce accuracy for other analyses
- Policy makers must consider reasonable overall accuracy metrics for privacy tradeoffs
- Knowing how overall metrics correspond to user results could help optimize DAS

DISTRICT-BY-DISTRICT DIFFERENTIAL PRIVACY ALGORITHMS (1940 CENSUS DATA)



TOPDOWN DIFFERENTIAL PRIVACY ALGORITHMS (1940 CENSUS DATA)



Selected Resources

Technical: [https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0945_Consistency for Large Scale Differentially Private Histograms.pdf](https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0945_Consistency%20for%20Large%20Scale%20Differentially%20Private%20Histograms.pdf)

Basics: https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

Updates: <https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html>



Thank you

John.Maron.Abowd@census.gov

Extra Slides

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

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

Reconstructing the 2010 Census: What did we find?

- On the 309 million reconstructed records, census block and voting age (18+) were correctly reconstructed for all records and for 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.