Differential Privacy and the 2020 Census in the United States

John M. Abowd
Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau

Association for the Advancement of Artificial Intelligence
Privacy Preserving AI Workshop, February 8, 2021, 9:00am EST USA

United States® Census Bureau

The views expressed in this talk are my own and not those of the U.S. Census Bureau.

Going from suppression to differential privacy is much easier than going from publishing all the microdata to differential privacy.

United States® Census Bureau

Start | Base Map | Selection | Results

Save | Load | Feedback | Previous Extent | Hide Tabs | Hide Chart/Report

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 | Zoom to Selection

Clear Overlays | Animate Overlays

Report/Map Outputs

Detailed Report

Export Geography

Print Chart/Map

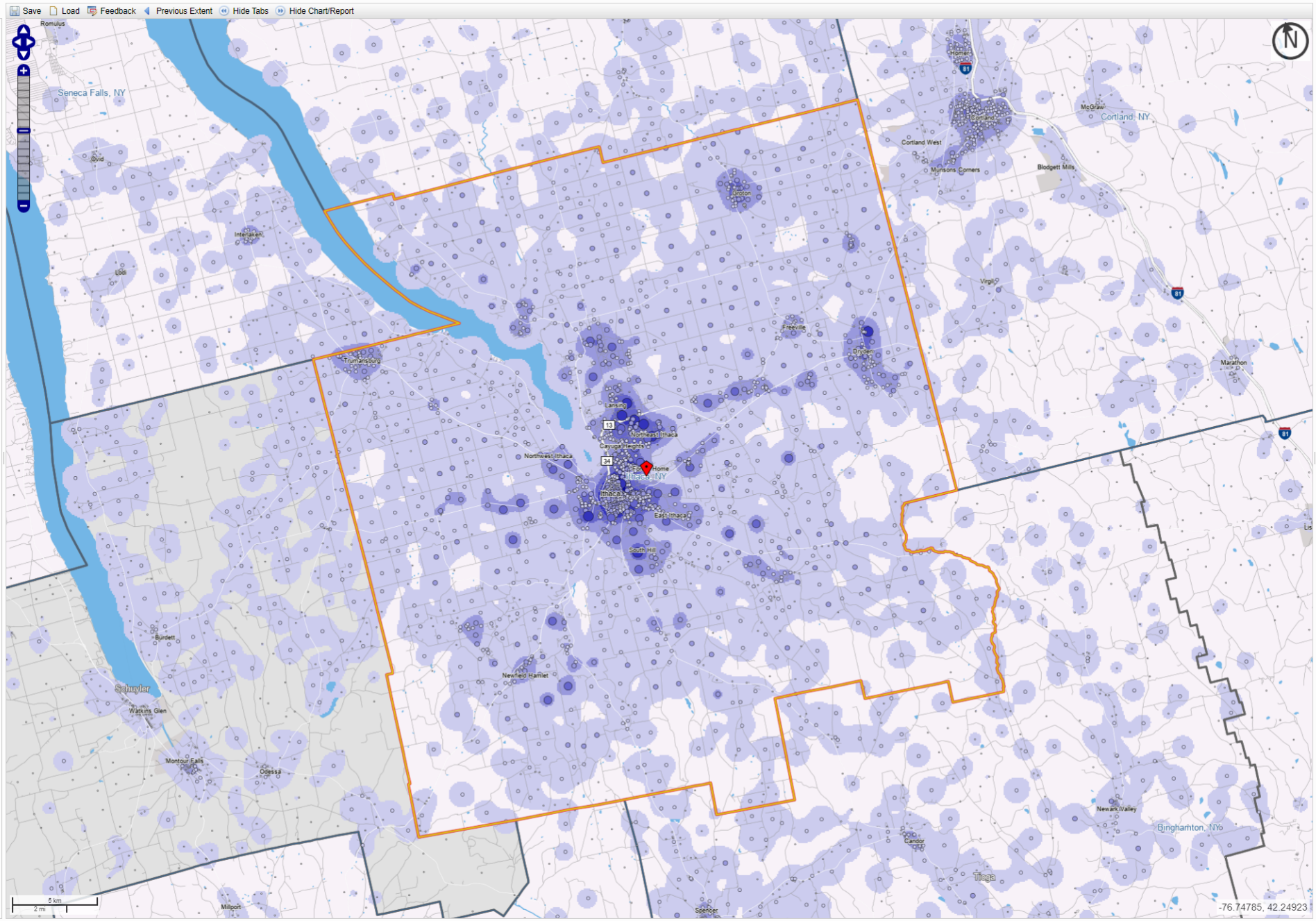
Legends

- 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

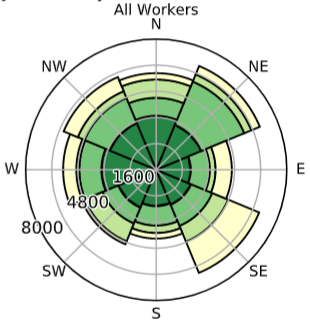
Analysis Settings

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 Date	07/20/2020 16:43 - OnTheMap 6.6
Code Revision	d7f8a300c9f4e458f61bc73d3099ca2cb8f8feaa
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

	2017 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%

Occupation
2 Occupations Selected
Select Multiple Characteristics

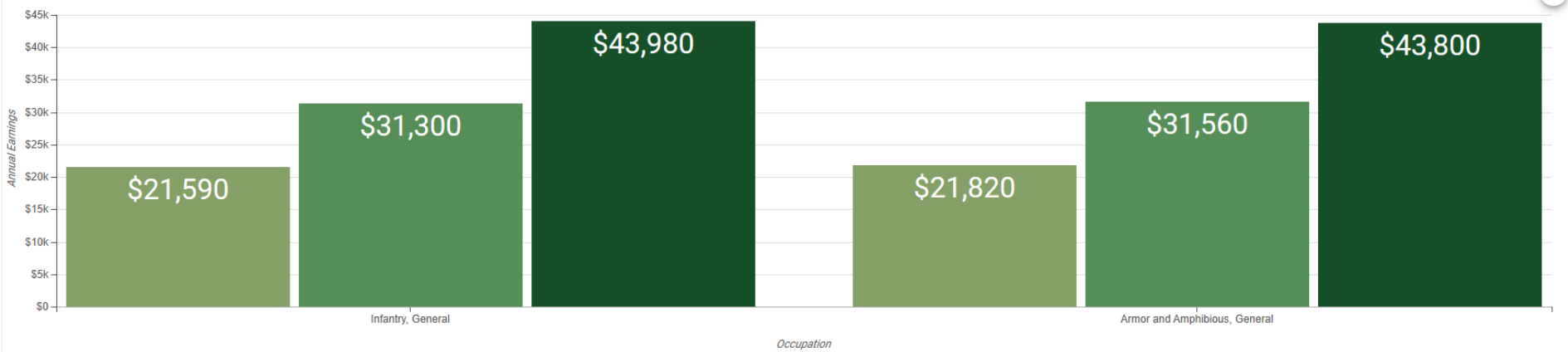
Cohort
2000-2007

Earnings for Employed Veterans by Occupation

Data Type: Earnings

Percentiles: 25% 50% 75%

Years Post Discharge: 1



Major data products from the 2020 Census:

- Apportion the House of Representatives
  (expected release between April 16 and April 30, 2021)
- Supply data to all state redistricting offices
  (TBD)
- Demographic and housing characteristics
  (TBD)
- Detailed race and ethnicity data
  (TBD)
- American Indian, Alaska Native, Native Hawaiian data
  (TBD)

For the 2010 Census, this was *more than 150 billion* statistics from 15GB total data.

United States® Census Bureau

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

United States® Census Bureau

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.

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.

United States® Census Bureau

The incredibly diverse user community is now engaged in the debate about accuracy v. privacy loss that publication of census results at this scale demands.

United States® Census Bureau

Title 13 U.S.Code delegates to the Census Bureau arbitrating these diverse use cases. Short of publishing the micro-data again, many users, maybe even the majority, will be uncomfortable with any choice. How does a statistical agency balance these interests?

United States® Census Bureau

All 2020 Census Publications

- Will all be processed by a collection of differentially private algorithms
- Using a total privacy-loss budget set as policy, not hard-wired, determined by the Data Stewardship Executive Policy Committee
- Preliminary code base, technical documents, and extensive demonstration products based on the 2010 Census confidential data have all been released to the public
- More information: https://www.census.gov/newsroom/blogs/research-matters/2019/10/balancing_privacyan.html

United States® Census Bureau

Naïve Method: BottomUp or Block-by-Block

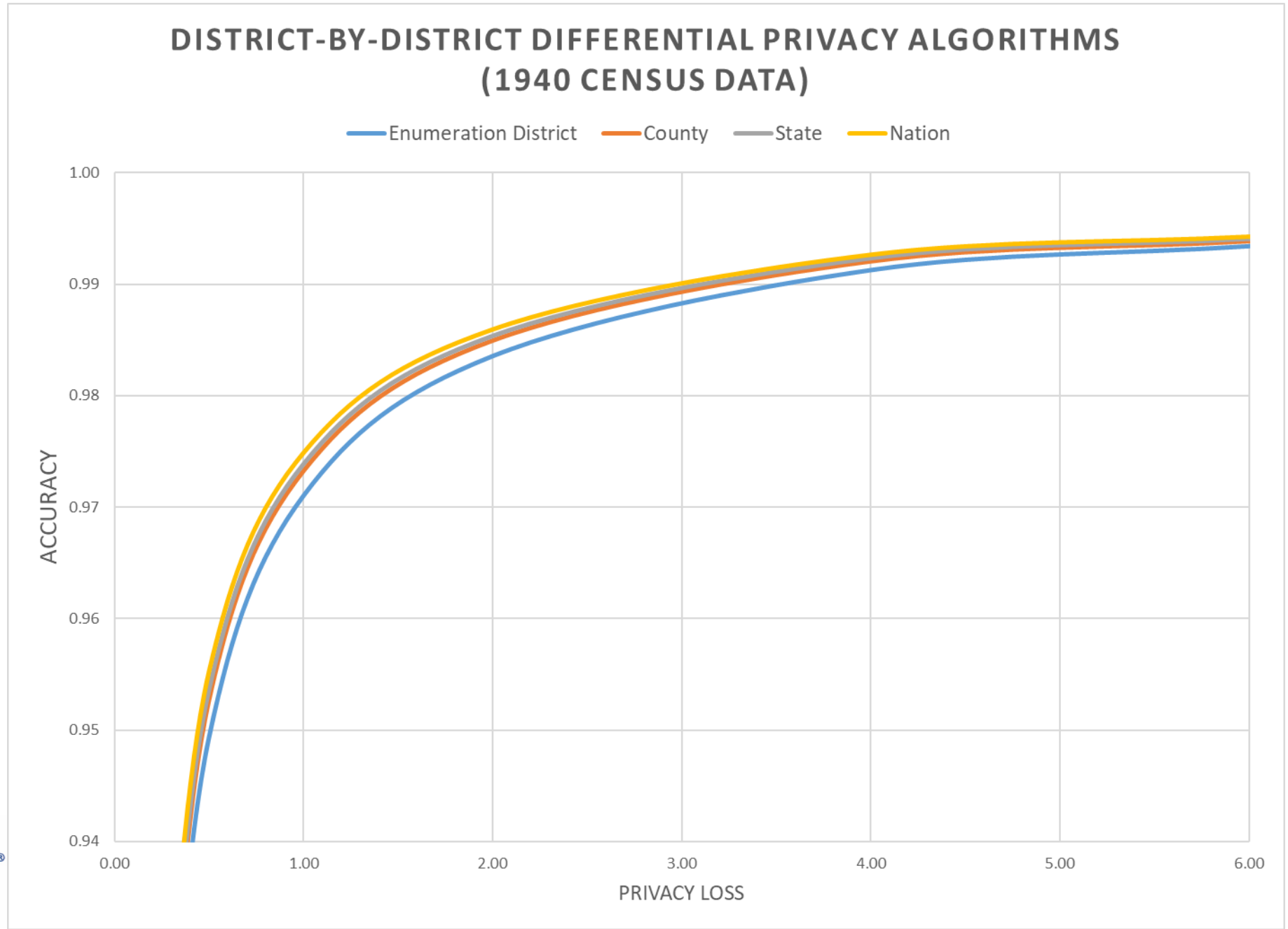
- Apply differential privacy algorithms to the most detailed level of geography
- Build all geographic aggregates from those components as a post-processing
- This is similar to the local differential privacy implementations in the Chrome browser, iOS, and Windows 10.

United States® Census Bureau

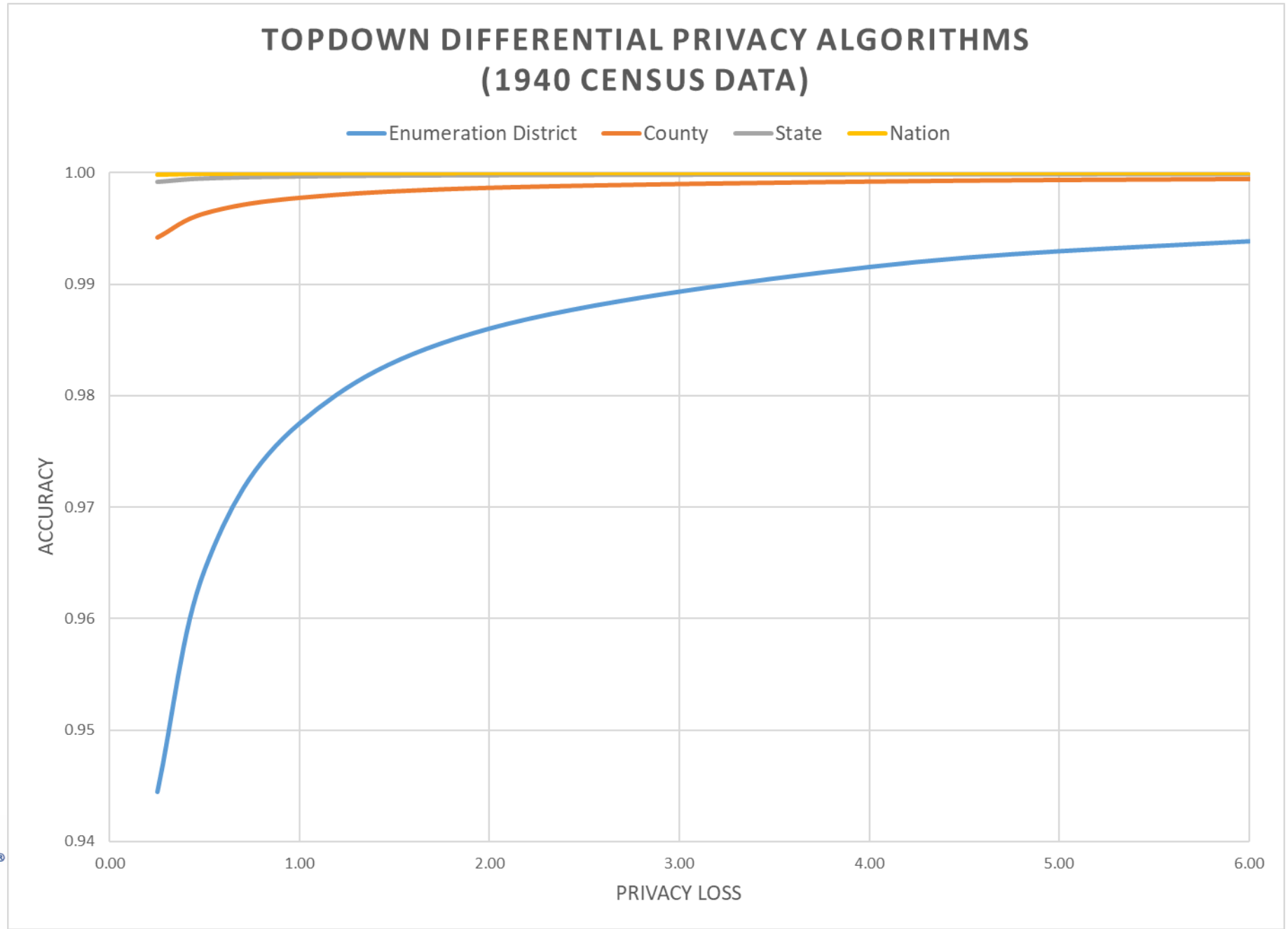
The Census TopDown Algorithm (TDA)

- Take differentially private measurements at every level of the Census geographic hierarchy
- Moving from the United States to the census block post-process:
  - Solve a non-negative least squares optimization to get non-negative tables at each level of the geographic hierarchy, always consistent with the levels above
  - Solve a mixed integer linear program to get non-negative, integer tables at each level of the geographic hierarchy, always consistent with the levels above
- Generate micro-data from these post-processed tables from the block-level detailed tables (an exact image of a unique Microdata Detail File)

United States® Census Bureau



United States® Census Bureau



United States® Census Bureau

If you feed TDA 11 billion differentially private measurements, it will do a good job that satisfies no one.

United States® Census Bureau

Accurate, but to whom?

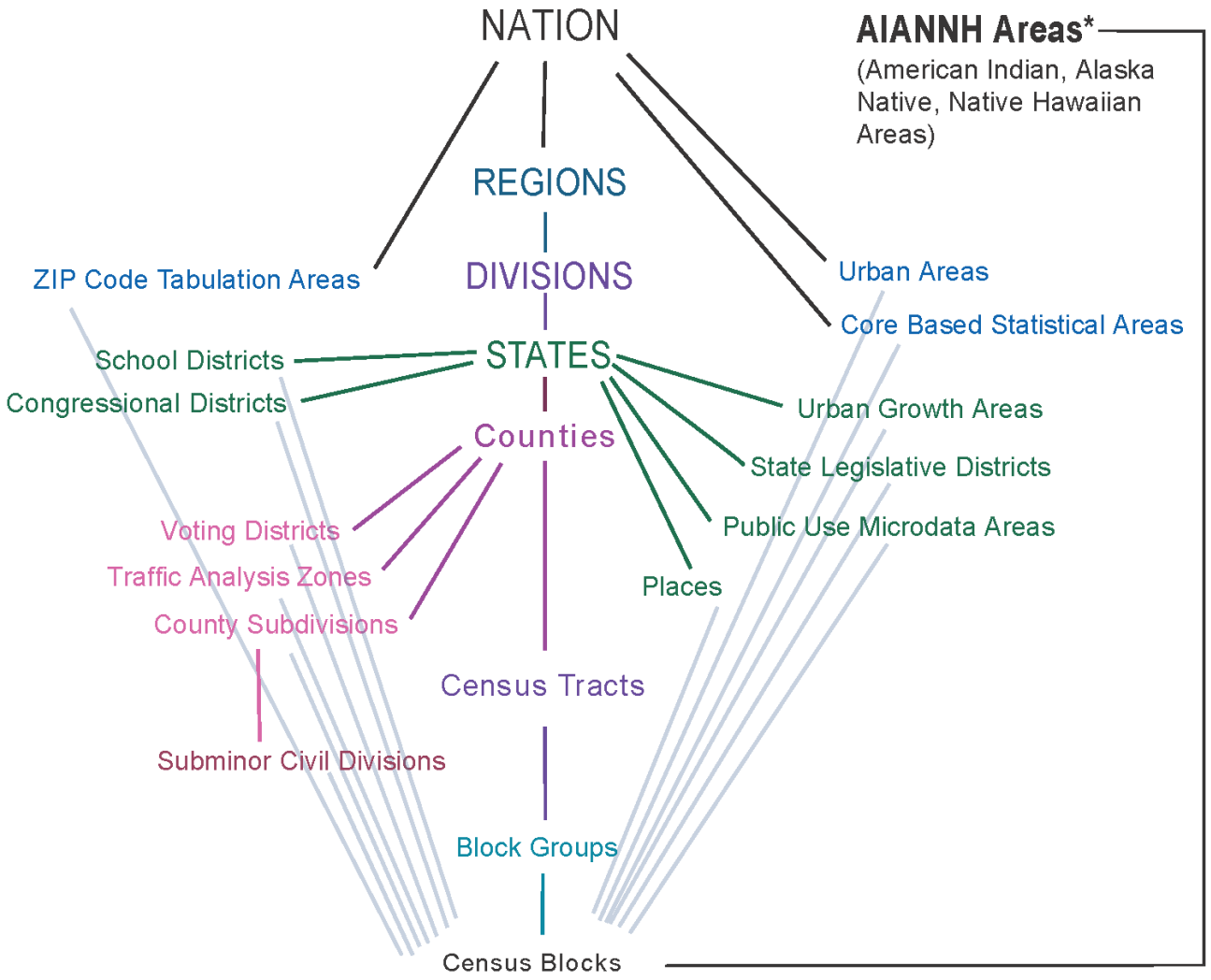
- DAS operates under interpretable formal privacy guarantees, given privacy-loss budgets
- Accuracy properties depend upon the output metric (use case)
- Distinct groups of data users will have a particular analyses they wish to be accurate
- Tuning accuracy for a given analysis can reduce accuracy for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

United States® Census Bureau

Deep Dive: Redistricting Data

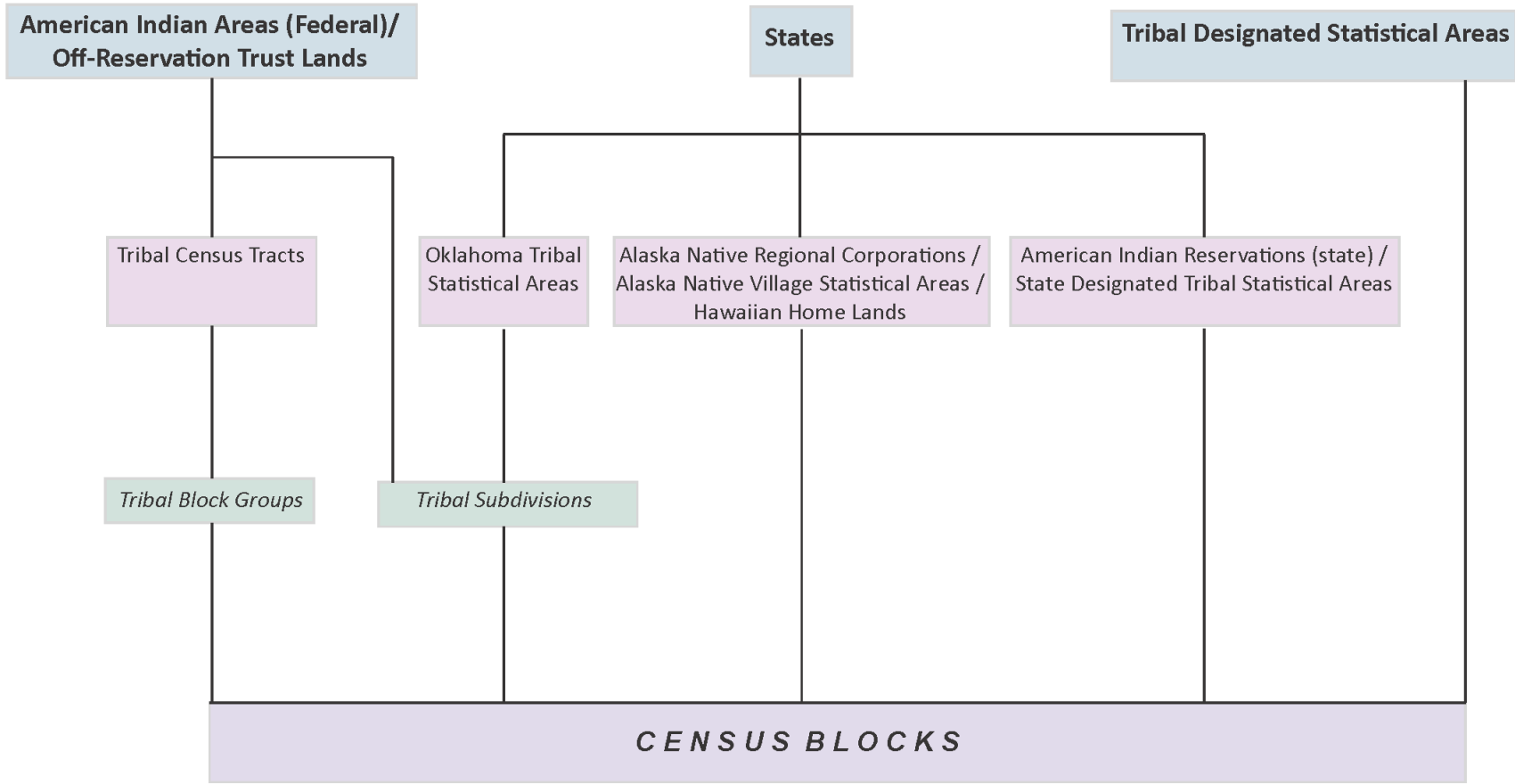
- Legislative districts for politically defined entities of arbitrary size
- Must be (approximately) equal populations in each district
- Districts must be consistent with Section 2 scrutiny under the 1965 Voting Rights Act
  - Large minority populations cannot be clustered into a few districts
  - Majority-minority districts (approximately 50%+ minority population) must be drawn when feasible
- Focus statistics: total population, ratio largest race/ethnic population to total population

United States® Census Bureau



United States® Census Bureau

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

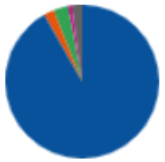


United States® Census Bureau

Senate District 48

In 2010, this district's race/ethnicity breakdown was:

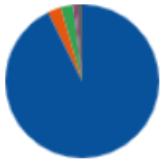
	#	Percent
Total:	292,870	100.0%
White:	268,818	91.8%
Black:	6,268	2.1%
Hispanic:	9,013	3.1%
Asian:	2,501	0.9%
All Other:	6,270	2.1%



Source: LATFOR, based on 2010 Census modified with prison population counts

The population eligible to vote (citizens of voting age) is estimated as follows:

	#	Percent
Citizen voting age population (CVAP):	219,620	100.0%
White CVAP:	203,505	92.7%
Black CVAP:	6,144	2.8%
Hispanic CVAP:	5,968	2.7%
Asian CVAP:	1,142	0.5%
All Other CVAP:	2,861	1.3%

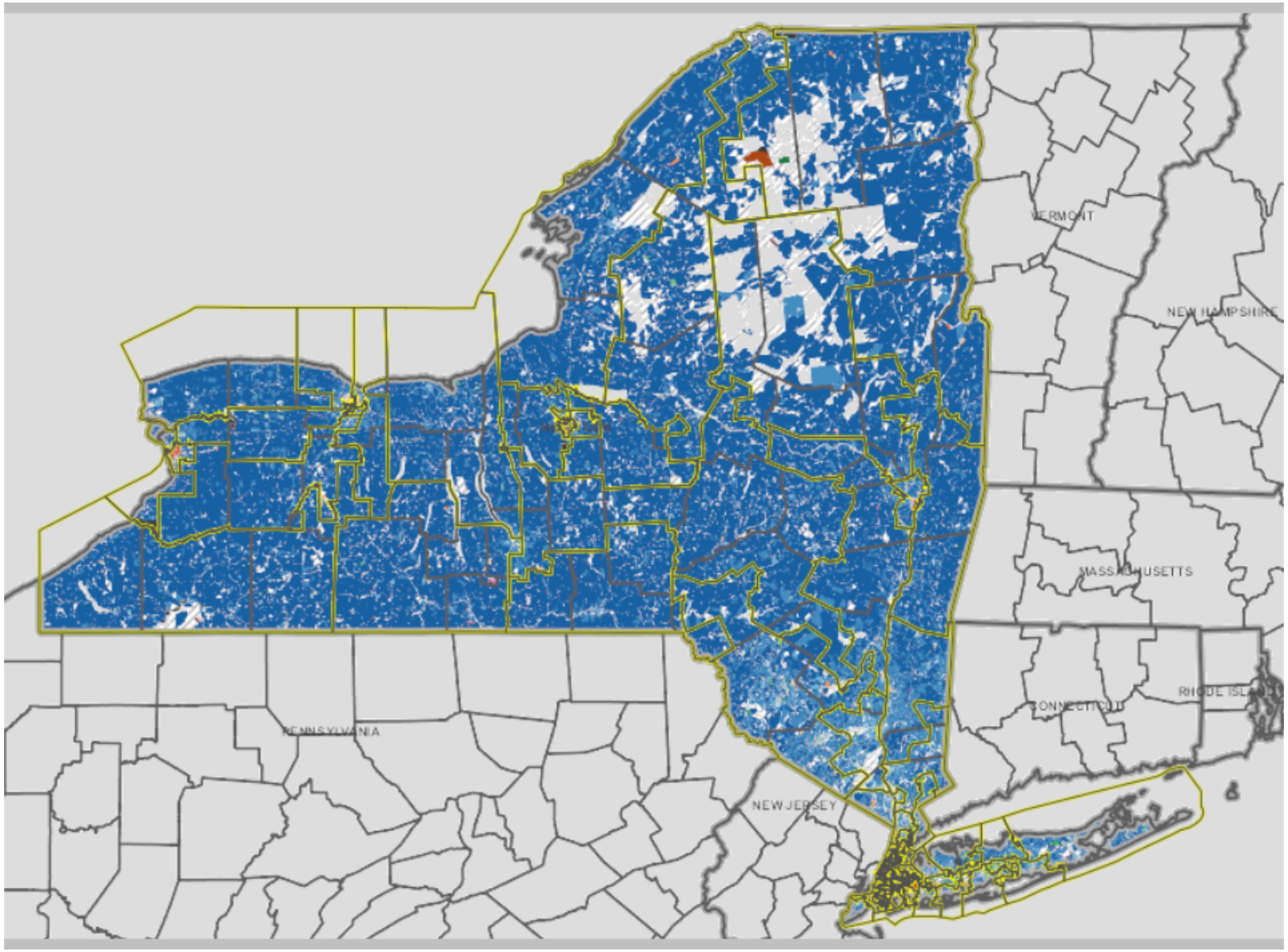


Source: Special Census Bureau tabulation from the 2006-10 5-Year American Community Survey

See the list of candidates for Senate District 48

Compare old and new district maps

Like 105 Share

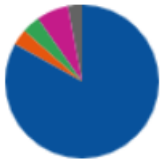


United States® Census Bureau

Assembly District 125

In 2010, this district's race/ethnicity breakdown was:

	#	Percent
Total:	133,580	100.0%
White:	111,052	83.1%
Black:	4,448	3.3%
Hispanic:	5,079	3.8%
Asian:	9,001	6.7%
All Other:	4,000	3.0%



Source: LATFOR, based on 2010 Census modified with prison population counts

The population eligible to vote (citizens of voting age) is estimated as follows:

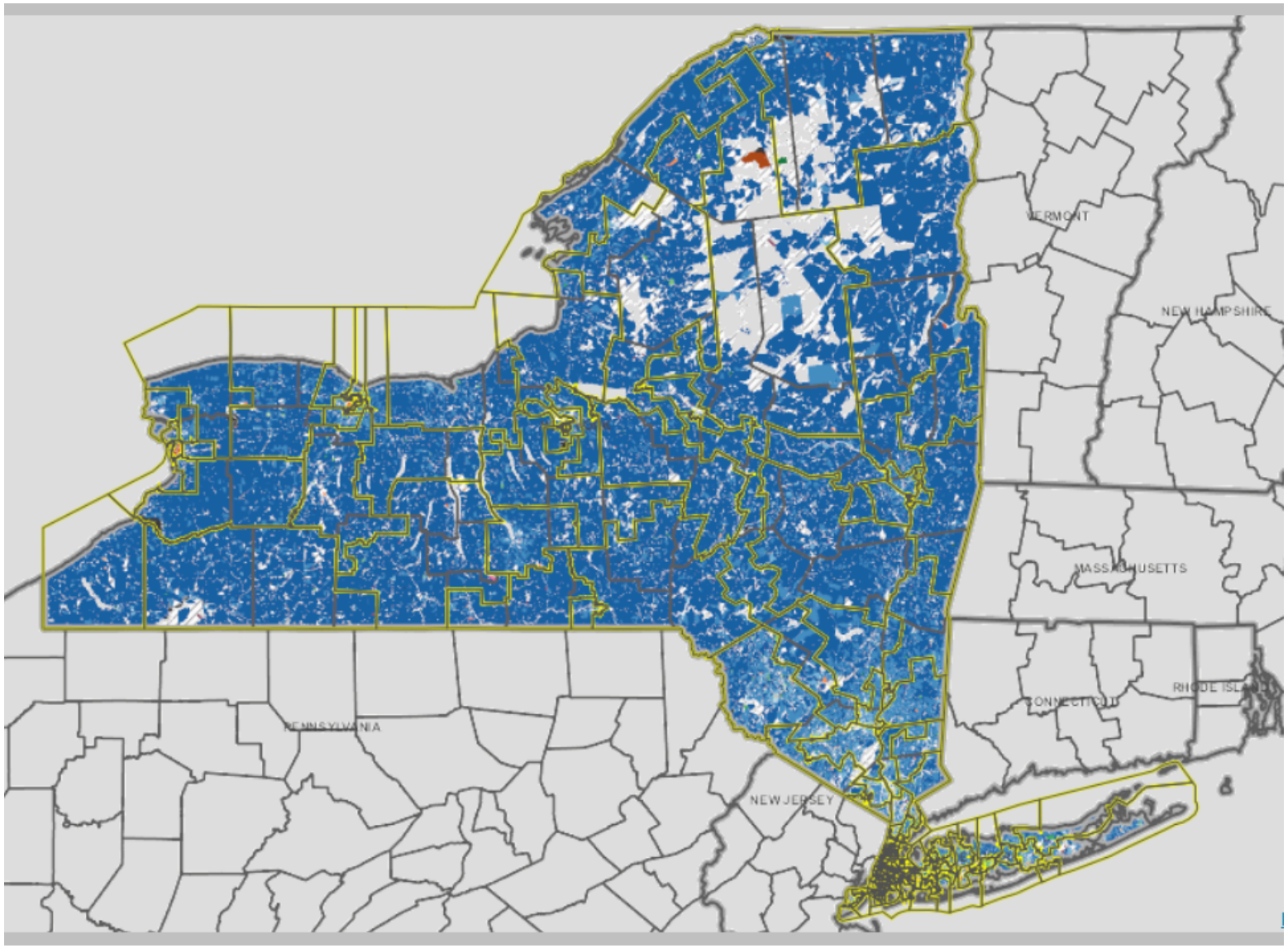
	#	Percent
Citizen voting age population (CVAP):	100,790	100.0%
White CVAP:	89,235	88.5%
Black CVAP:	2,927	2.9%
Hispanic CVAP:	3,221	3.2%
Asian CVAP:	3,710	3.7%
All Other CVAP:	1,697	1.7%

Source: Special Census Bureau tabulation from the 2006-10 5-Year American Community Survey

See the list of candidates for Assembly District 125

Compare old and new district maps

Like 105 Share



United States® Census Bureau

Basic Demographic Accuracy Profile

Table 1.a: Total Population for county size categories - MAE, RMSE, MAPE, CV, MALPE, and outliers

Universe: Total population

Geography: Summary Level 050 - State-County

	Count of Units (N)	MAE	RMSE	MAPE (%)	CV
All counties	3,143	6.57	8.91	0.12	0.01
Counties with total population less than 1,000	35	11.40	28.61	6.38	4.16
Counties with total population 1,000 to 4,999	268	6.08	7.54	0.23	0.25
Counties with total population 5,000 to 9,999	395	6.59	8.50	0.09	0.11
Counties with total population 10,000 to 49,999	1,469	6.20	7.85	0.03	0.03
Counties with total population 50,000 to 99,999	398	6.16	7.80	0.01	0.01
Counties with total population of 100,000 or more	578	7.71	10.37	-	-

Source: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-metrics.html

Workbook: https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ppmf20201116/2020-11-16-data-metrics-tables.xlsx

United States® Census Bureau

Table 1.b: Total Population for incorporated place size categories - MAE, RMSE, MAPE, CV, MALPE, and outliers

Universe: Total population

Geography: Summary Level 160 - State-Place - Incorporated Places Only

	Count of Units (N)	MAE	RMSE	MAPE (%)	CV
All incorporated places	19,540	24.54	34.44	4.98	0.35
Incorporated places with total population less than 500	6,168	19.66	25.73	12.80	10.96
Incorporated places with total population 500 to 999	3,066	23.34	31.16	3.38	4.32
Incorporated places with total population 1,000 to 4,999	5,672	23.53	32.23	1.23	1.38
Incorporated places with total population 5,000 to 9,999	1,664	25.52	35.29	0.37	0.50
Incorporated places with total population 10,000 to 49,999	2,265	33.13	45.73	0.18	0.21
Incorporated places with total population 50,000 to 99,999	432	43.70	58.77	0.07	0.08
Incorporated places with total population of 100,000 or more	273	61.94	84.01	0.03	0.03

Source: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-metrics.html

Workbook: https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ppmf20201116/2020-11-16-data-metrics-tables.xlsx

United States® Census Bureau

Table 5.b: Hispanic or Latino Origin for county size categories - MAE, RMSE, MAPE, CV, MALPE, and outliers

Universe: Total population

Geography: Summary Level 050 - State-County

	Count of Units (N)	MAE	RMSE	MAPE (%)	CV
All counties					
Hispanic or Latino	3,143	32.76	51.39	10.72	0.32
Not Hispanic or Latino	3,143	33.13	51.68	0.27	0.06
Counties with population 0 to 9					
Hispanic or Latino	17	12.76	18.67	503.45	356.54
Not Hispanic or Latino	-	-	-	-	-
Counties with population 10 to 99					
Hispanic or Latino	322	17.59	23.77	40.10	40.94
Not Hispanic or Latino	3	41.00	61.59	47.14	73.91
Counties with population of 100 or more					
Hispanic or Latino	2,804	34.62	53.78	4.36	0.30
Not Hispanic or Latino	3,140	33.12	51.67	0.22	0.06

Source: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-metrics.html

Workbook: https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ppmf20201116/2020-11-16-data-metrics-tables.xlsx

United States® Census Bureau

Table 5.c: Hispanic or Latino Origin for incorporated place size categories - MAE, RMSE, MAPE, CV, MALPE, and outliers

Universe: Total population

Geography: Summary Level 160 - State-Place - Incorporated Places Only

	Count of Units (N)	MAE	RMSE	MAPE (%)	CV
All incorporated places					
Hispanic or Latino	19,540	22.48	43.96	65.67	2.24
Not Hispanic or Latino	19,540	32.16	49.75	5.58	0.63
Incorporated places with population 0 to 9					
Hispanic or Latino	6,241	4.47	7.37	142.15	236.31
Not Hispanic or Latino	44	5.50	7.26	123.58	155.85
Incorporated places with population 10 to 99					
Hispanic or Latino	6,609	14.46	18.94	51.69	52.39
Not Hispanic or Latino	1,217	14.85	19.42	28.69	31.60
Incorporated places with population of 100 or more					
Hispanic or Latino	6,690	47.20	72.38	8.12	1.27
Not Hispanic or Latino	18,279	33.37	51.19	3.76	0.61

Source: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-metrics.html

Workbook: https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ppmf20201116/2020-11-16-data-metrics-tables.xlsx

United States® Census Bureau

Table 6.a: Race Alone for states - MAE, RMSE, MAPE, CV, MALPE, and outliers

Universe: Total population

Geography: Summary Level - 040 State

	Count of Units (N)	MAE	RMSE	MAPE (%)	CV
White alone	51	162.16	190.41	0.01	-
Black alone	51	94.27	109.73	0.37	0.01
AIAN alone	51	199.92	233.78	1.55	0.41
Asian alone	51	103.55	123.79	0.36	0.04
NHPI alone	51	171.24	222.61	21.75	2.10
SOR alone	51	95.49	123.22	0.33	0.03
Two or more races	51	20.71	27.09	0.04	0.02

Source: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-metrics.html

Workbook: https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/ppmf20201116/2020-11-16-data-metrics-tables.xlsx

United States® Census Bureau

How to reconcile these statistics

- Construct error metrics of the form

$$Pr[|TDA - CEF| \leq \alpha] \geq 1 - \beta$$

- Less than α error with probability at least 1-β for a target minimum population
- Statistical interpretation: absolute differences (=RMSE differences) greater than α are outside the 1-β confidence interval
- A single statistic can be used to tune the redistricting application

$$\frac{Population\ of\ Largest\ Race\ or\ Ethnic\ Group}{Total\ Population}$$

- Calculated for the TopDown Algorithm (TDA) output and the 2020 Census (CEF)

United States® Census Bureau

Privacy protection out of the shadows

- Certain privacy practices for previous censuses depended upon obfuscation
- 2020 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

United States® Census Bureau

Selected Resources

- Technical: https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0945_Consistency_for_Large_Scale_Differentially_Private_Histograms.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

United States® Census Bureau

Thank you.

John.Maron.Abowd@census.gov

United States® Census Bureau