# Disclosure Avoidance At-Scale

William Sexton, Mathematical Statistician
Center for Enterprise Dissemination - Disclosure Avoidance
United States Census Bureau
william.n.sexton@census.gov

JSM, Denver, CO
July, 2019

*This presentation is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the author and not those of the U.S. Census Bureau.*
*The Census Bureau's Disclosure Review Board and Disclosure Avoidance Officers have reviewed this data product for unauthorized disclosure of confidential information and have approved the disclosure avoidance practices applied to this release. (DRB Approval # CBDRB-FY19-446)*

# Acknowledgements

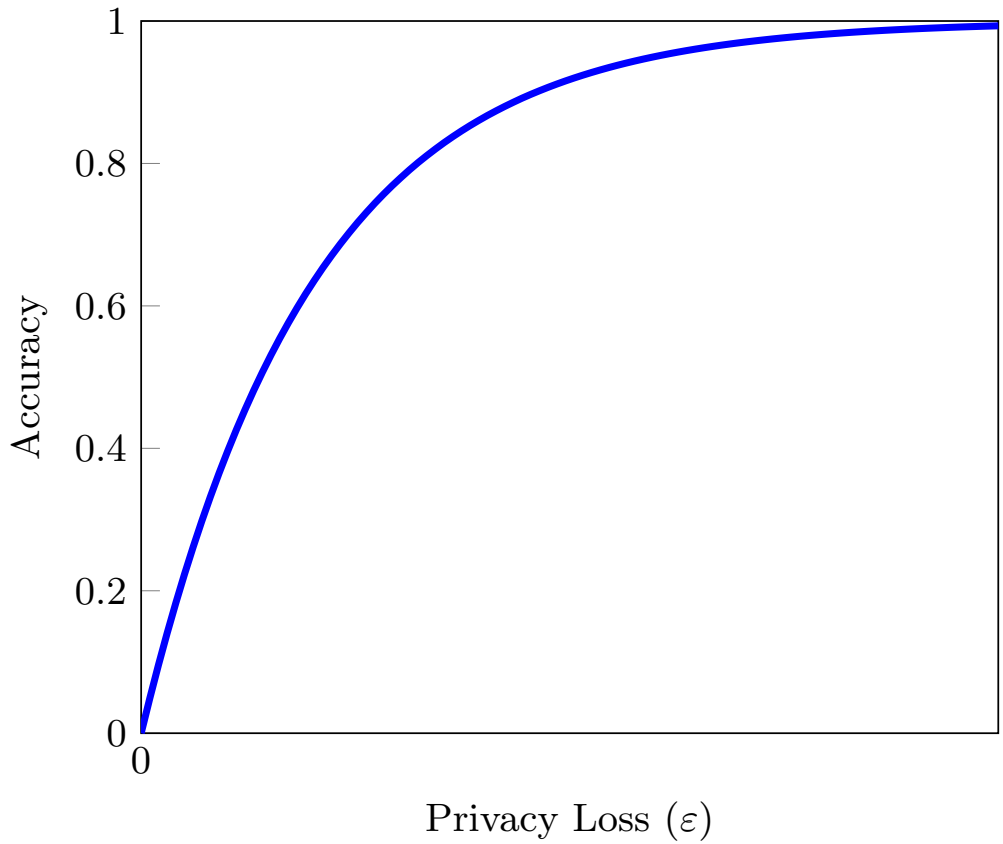
**2020 Disclosure Avoidance System (DAS) Project Lead:**
John Abowd; U.S. Census Bureau & Cornell University

**2020 DAS Scientific Lead:**
Daniel Kifer, Pennsylvania State University

**2020 DAS Team:**
Robert Ashmead (former), Simson Garfinkel, Phil Leclerc, Brett Moran, Pavel Zhuravlev; U.S. Census Bureau

# The Dual Mandate of the Census Bureau

- Collect data and disseminate accurate statistics about the US population.
- Protect the privacy of individual data.

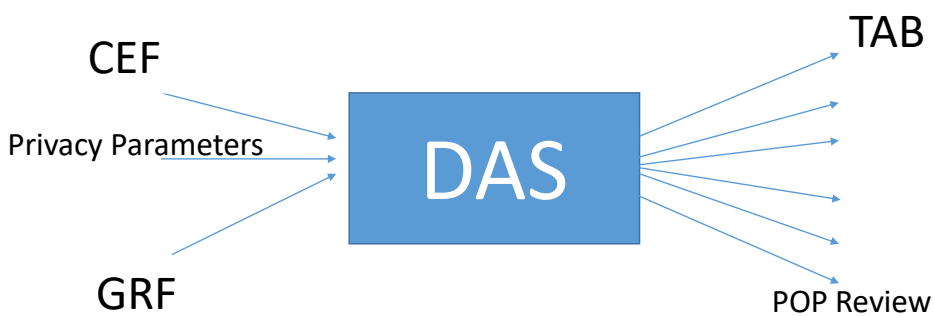


# Overview of TopDown and Trade-off

- The 2020 Disclosure Avoidance System (DAS), in particular its core TopDown algorithm, provides a production technology. The production activity is up to policy makers, however the choice is constrained by the given technology.
- Empirical results are useful for modeling the production possibility frontier (PPF). Understanding the privacy-loss, accuracy trade-off specifically associated with the technology at hand is necessary for informed decision making.
- Empirical results measure fitness-for-use against important use-cases such as redistricting.

# Disclosure Avoidance System (DAS)

- The DAS is a small component of the entire 2020 Census operation.
- The codomain of the DAS needs to lie in the domain of any system it left composes with.
- Historically, this implied DAS must output microdata.



# Rethinking the Microdata Requirement

- Historically, all the major data products were sourced by microdata.
- Improperly constrains production technology that defines the PPF. Simple example: Laplace Mechanism vs NoisyMax.
- The end-goal is high utility data products. Allowing for greater flexibility in the algorithm design will lead to better production technologies.
- Don't need to abandon microdata completely though. What can be produced well as microdata?

# Redesign of Data Products

- Microdata supported products (PL94, DHC-Persons and DHC-Households, Demo Profile).
- Other Products (Detailed Race/AIAN, Person-Household Joins) - Out of scope for TopDown.
- Table classification:
  - by geography.
  - by universe: what is the item being counted (person or household)?

# Rethinking the Microdata Detail File (MDF) Specifications

- Historically, wide variety of variables were preserved through the DA process (date of birth, mafid, allocation flags).
- Reverse engineer microdata schema to meet demands of revised data products.
- Attributes/attribute domains should have as much detail as necessary but no more.
- Microdata will consist of two disjoint files (one for each universe: person and household).
  - Each record universe is the Cartesian product of its attribute domains, after eliminating structural zeros.

# MDF Person

| | | | | |
|---|---|---|---|---|
| 13 | QAGE | Edited Age | INT(3) | 0-115 |
| 14 | CENHISP | Hispanic Origin | CHAR(1) | 1 = Not Hispanic |
| | | | | 2 = Hispanic |
| 15 | CENRACE | Census Race | CHAR(2) | 01 = White alone |
| | | | | 02 = Black alone |
| | | | | 03 = AIAN alone |
| | | | | 04 = Asian alone |
| | | | | 05 = NHPI alone |
| | | | | 06 = SOR alone |
| | | | | 07 = White; Black |
| | | | | 08 = White; AIAN |
| | | | | 09 = White; Asian |
| | | | | 10 = White; NHPI |
| | | | | 11 = White; SOR |
| | | | | 12 = Black; AIAN |
| | | | | 13 = Black; Asian |
| | | | | 14 = Black; NHPI |
| | | | | 15 = Black; SOR |
| | | | | 16 = AIAN; Asian |
| | | | | 17 = AIAN; NHPI |
| | | | | 18 = AIAN; SOR |
| | | | | 19 = Asian; NHPI |
| | | | | 20 = Asian; SOR |
| | | | | 21 = NHPI; SOR |
| | | | | 22 = White; Black; AIAN |
| | | | | 23 = White; Black; Asian |
| | | | | 24 = White; Black; NHPI |
| | | | | 25 = White; Black; SOR |
| | | | | ... |
| | | | | 63 = White; Black; AIAN; Asian; NHPI; SOR |

- Naive cardinality of Person Record Universe (excluding block id) $= 83{,}311{,}200 \approx 83$ million.

# MDF Unit

| | | | | |
|---|---|---|---|---|
| 11 | VACS | Vacancy Status | CHAR(1) | 0 = NIU |
| | | | | 1 = Vacant, for rent |
| | | | | 2 = Vacant, rented, not occupied |
| | | | | 3 = Vacant, for sale only |
| | | | | 4 = Vacant, sold, not occupied |
| | | | | 5 = Vacant, for seasonal, recreational, or occasional use |
| | | | | 6 = Vacant, for migrant workers |
| | | | | 7 = Vacant, other |
| 12 | HHSIZE | Population Count | INT(5) | 0 (vacant or NIU), 1, 2, 3, 4, 5, 6, 7+ |
| 13 | HHT | Household/Family Type | CHAR(1) | 0 = NIU |
| | | | | 1 = Married couple household |
| | | | | 2 = Other family household: Male householder |

- Naive cardinality of the Unit Record Universe (excluding block id) $= 7{,}188{,}480{,}000{,}000 \approx 7$ trillion.

# Refining the Record Universes

- Removing structural zeros/merging variables:
  - A person cannot reside simultaneously in a household and GQ.
  - A 1-person household cannot be a married family household.
- Managing corner cases separately:
  - Vacancy status does not cross with occupied household attributes.
- Person and Household histograms are both under 3 million cells.
- For comparison, the 2018 E2E code ran on a 2,000 cell histogram.
- Largest successful runs have been on a subset of the person variables, roughly 467k cells.

# Initializing the DAS

- The core TopDown algorithm will run twice (once for Persons and once for Households).
  - One consequence is that it is difficult to maintain consistency accross universes.
  - We strive for within universe consistency. Stakeholder input is vital here.
- Primary input is the confidential Census Edited File (CEF).
- The CEF is treated as ground truth. That is, our privacy analysis does not account for operations preceeding the DAS including edit and imputation procedures.
- Assumption: Input data are clean. No missing values, out of range values, etc.

# 2010 Test Products: Intro

- The DAS team is generating test products that demonstrate the computational capabilities of the DAS at present.
- The DAS is capable of processing the 2010 CEF and producing protected microdata adhering to (slightly simplified) 2020 MDF specifications.
- The test MDF can be used to tabulate about 70% of the tables in the proposed 2020 DHC data product.

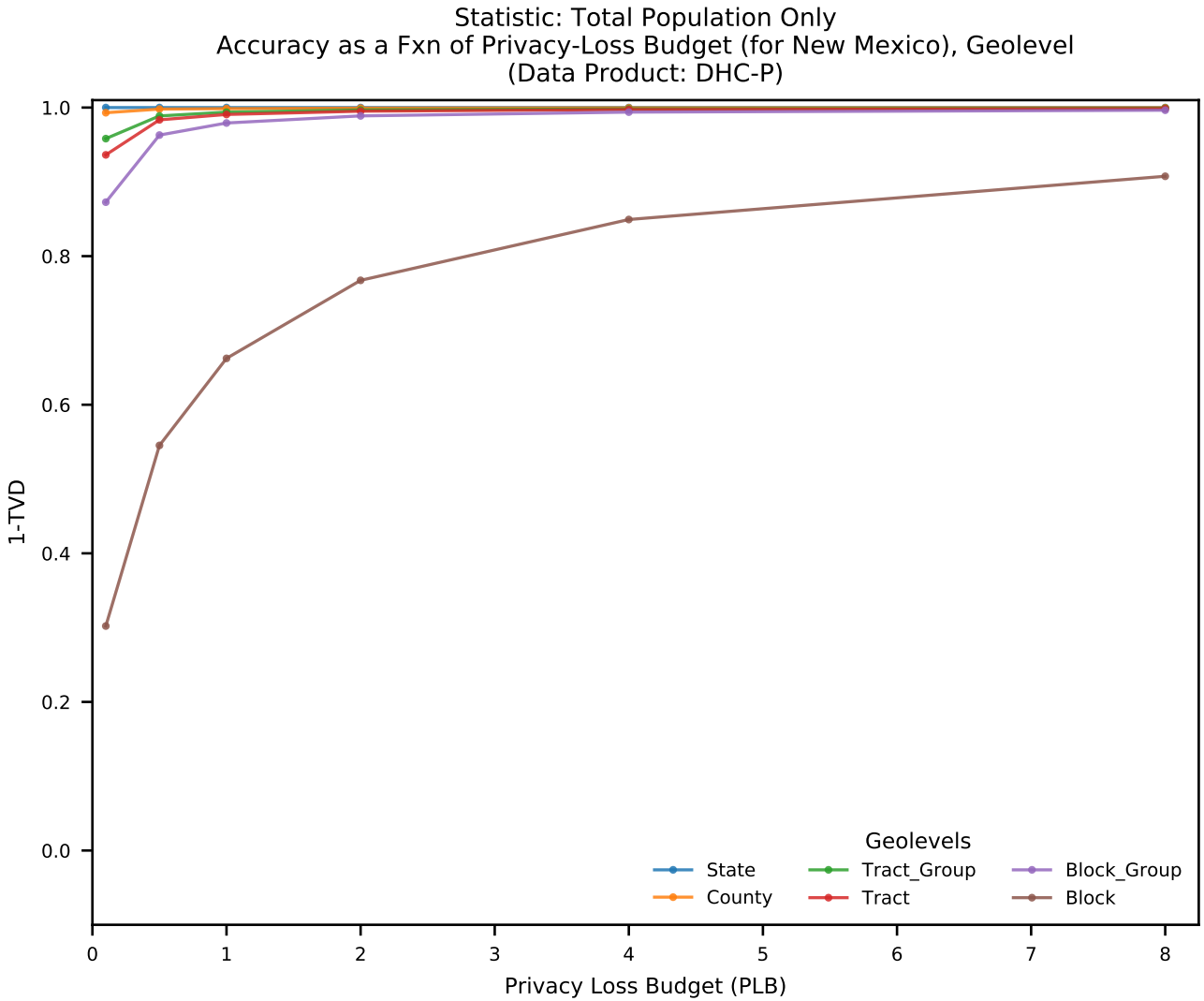
# 2010 Test Products: Scale

- The DAS is capable of processing the entire nation ($\sim$ 310 million person records and $\sim$ 120 million household records) rather than a small test area (such as Providence, RI in the 2018 End-to-End (E2E) test).
- The "slightly simplified" 2020 MDF specification translates to roughly 200 times the scale of the 2018 E2E test in terms of histogram size.
- The DAS can produce microdata for persons and households. Households characteristics were essentially non-existent in the 2018 E2E test.

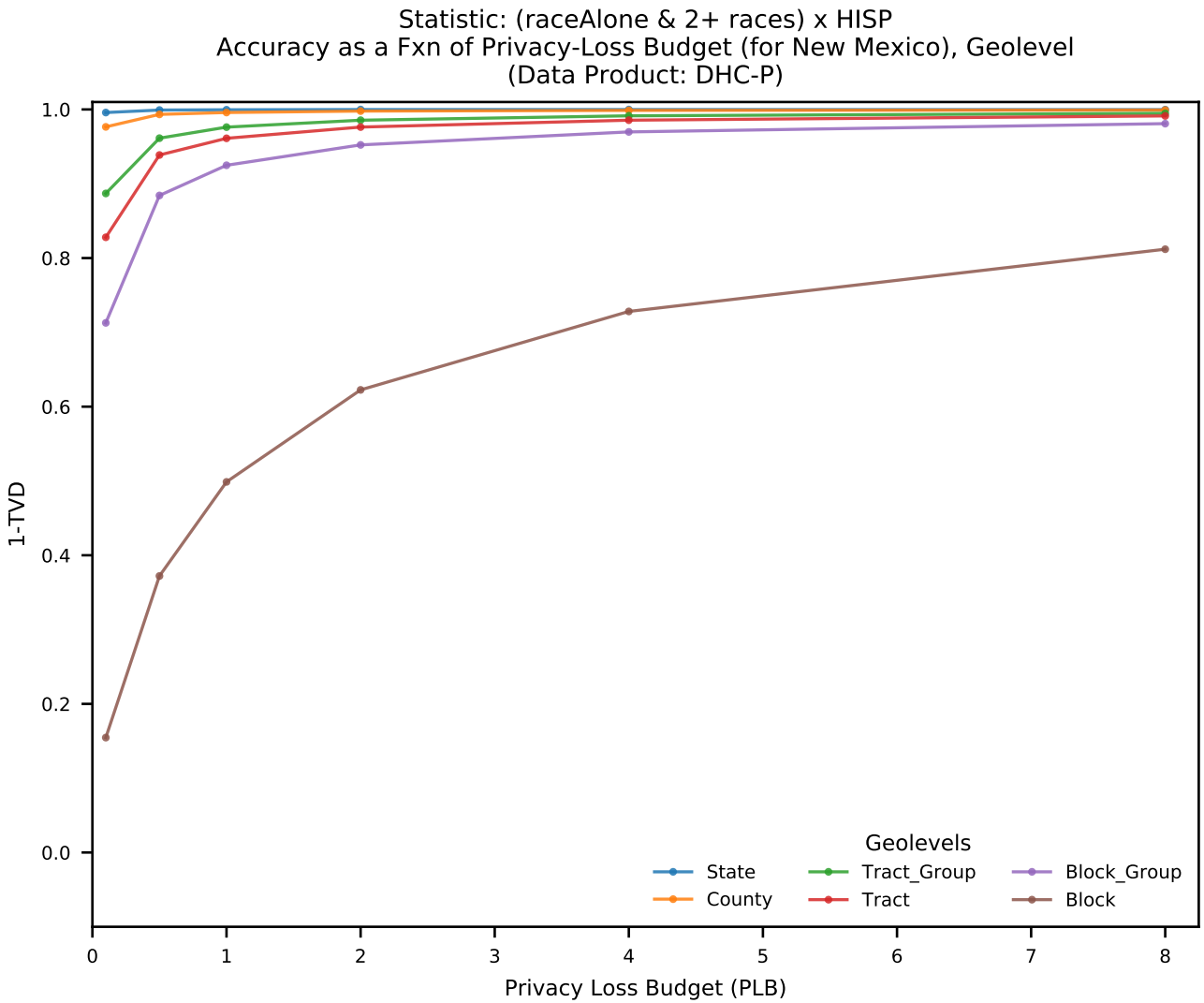
# 2010 Test Products: Resource constraints

- The DAS operates on Amazon Web Services Elastic Map Reduce computer clusters.
- Operating at-scale requires about 18 worker nodes (r4.16xLarge, 64 core - 488 gb RAM)
- Observed run times: $\sim$ 20 hours to produce the household microdata and $\sim$ 60 hours to produce the person microdata.

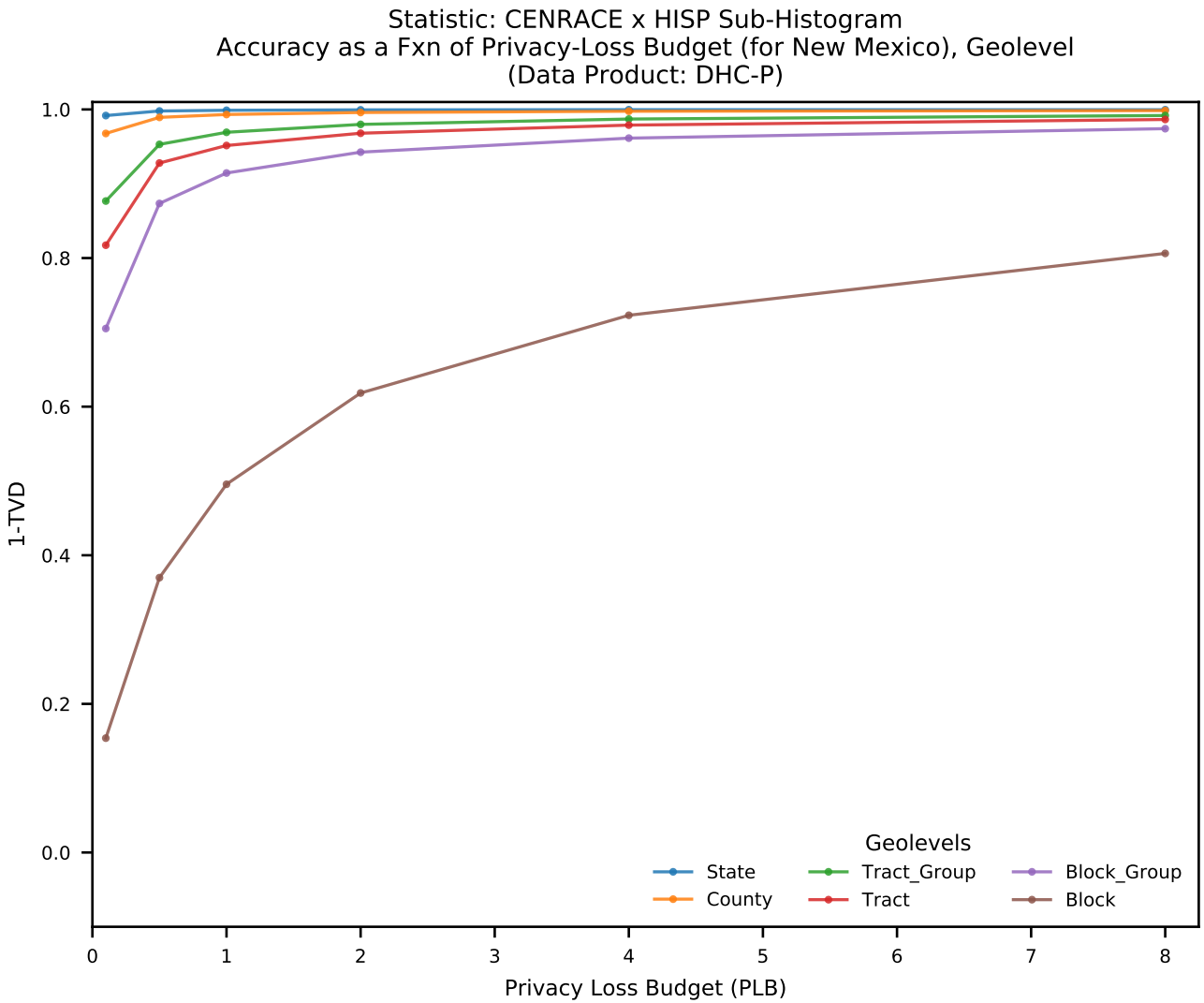
# 2010 Test Products: Privacy-loss, Accuracy Tradeoff



# 2010 Test Products: Privacy-loss, Accuracy Tradeoff



# 2010 Test Products: Privacy-loss, Accuracy Tradeoff



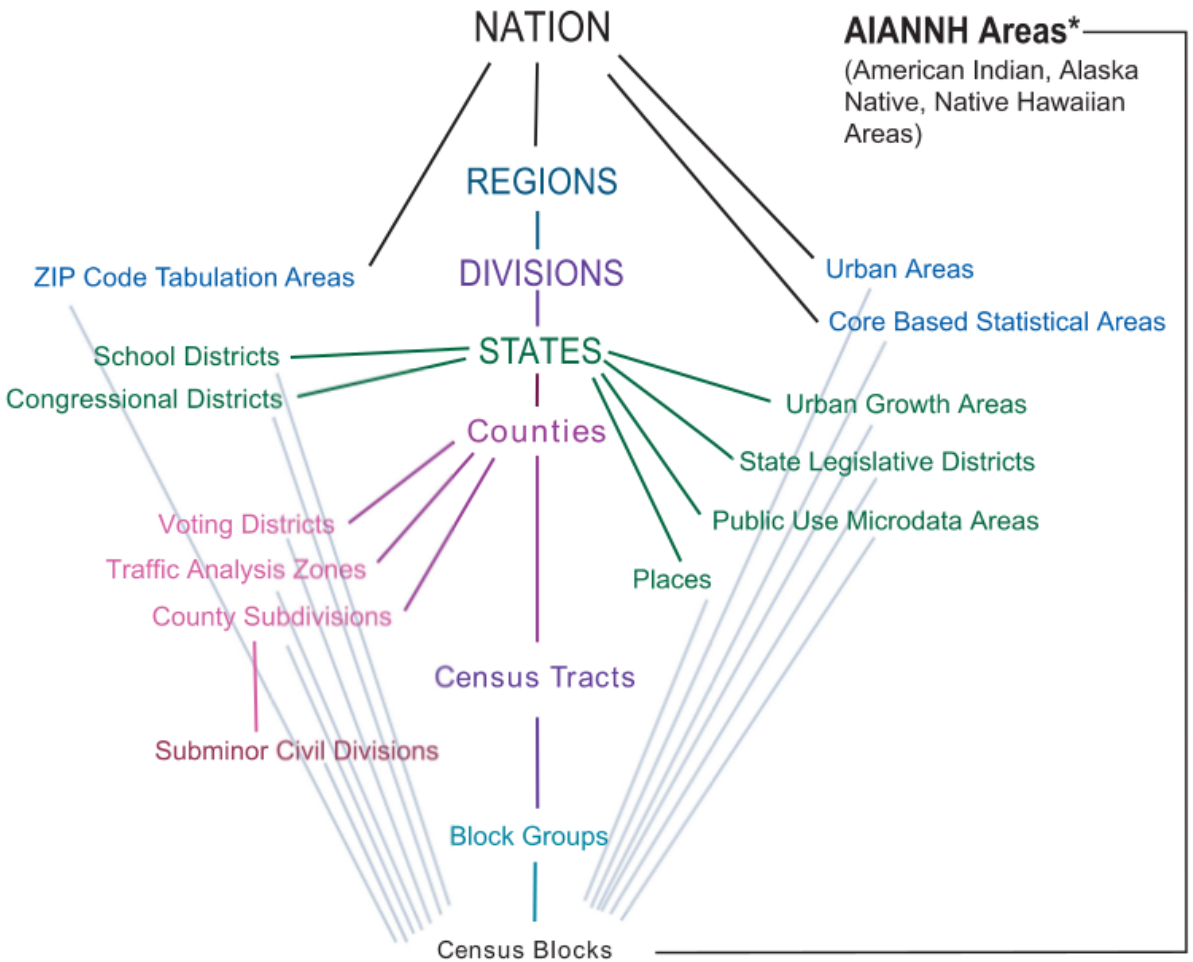
# References

1. Ios Kotsogiannis, Yuchao Tao, Xi He, Maryam Fanaeepour, Ashwin Machanavajjhala, Michael Hay, Gerome Miklau "PrivateSQL: A Differentially Private SQL Query Engine", To appear PVLDB 2019

# Thanks!

william.n.sexton@census.gov

# Backup: Artificial Geolevels



# Backup: Artificial Geolevels

- Introduce artificial geolevels into main hierarchy to help with scaling. Reduces maximum number of children at a given level - a known bottleneck in scaling.
- Prelimary results are promising, especially between county and tracts, with regard to improving tractibility of large scale runs. Full impact on accuracy still being analyzed.

# Backup: Detailed Race/AIAN and Person-Household Joins

- Why not microdata?
  - High sensitivity.
  - Complex consistency requirements.
  - Non-standard geographies.
  - TopDown scalibility.
  - Fundamentally different algorithms required for joins.
  - Small count bias will look even worse.
- For joins, considering PrivateSQL: produces a privatized view from a relational database from which tables can be published [KTHFMHM19].
- For Detailed Races/AIAN, considering a variant of TopDown that relaxes many of the consistency requirements, and only crosses race with select other variables.