Stepping-up: The Census Bureau Tries to Be a Good Data Steward in the 21st Century

John M. Abowd
Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau
Simons Institute *Data Privacy: From Foundations to Applications*
March 4, 2019 9:30-10:30

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

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

The challenges of a census:

1. collect all of the data necessary to underpin our democracy
2. protect the privacy of individual data to ensure trust and prevent abuse

The Database Reconstruction Vulnerability

These are the lessons from cryptography

- Too many statistics
- Noise infusion is necessary
- Transparency about methods is a benefit

What we did

- Database reconstruction for all 308,745,538 people in 2010 Census
- Link reconstructed records to commercial databases: acquire PII
- Successful linkage to commercial data: putative re-identification
- Compare putative re-identifications to confidential data
- Successful linkage to confidential data: confirmed re-identification
- Harm: attacker can learn self-response race and ethnicity

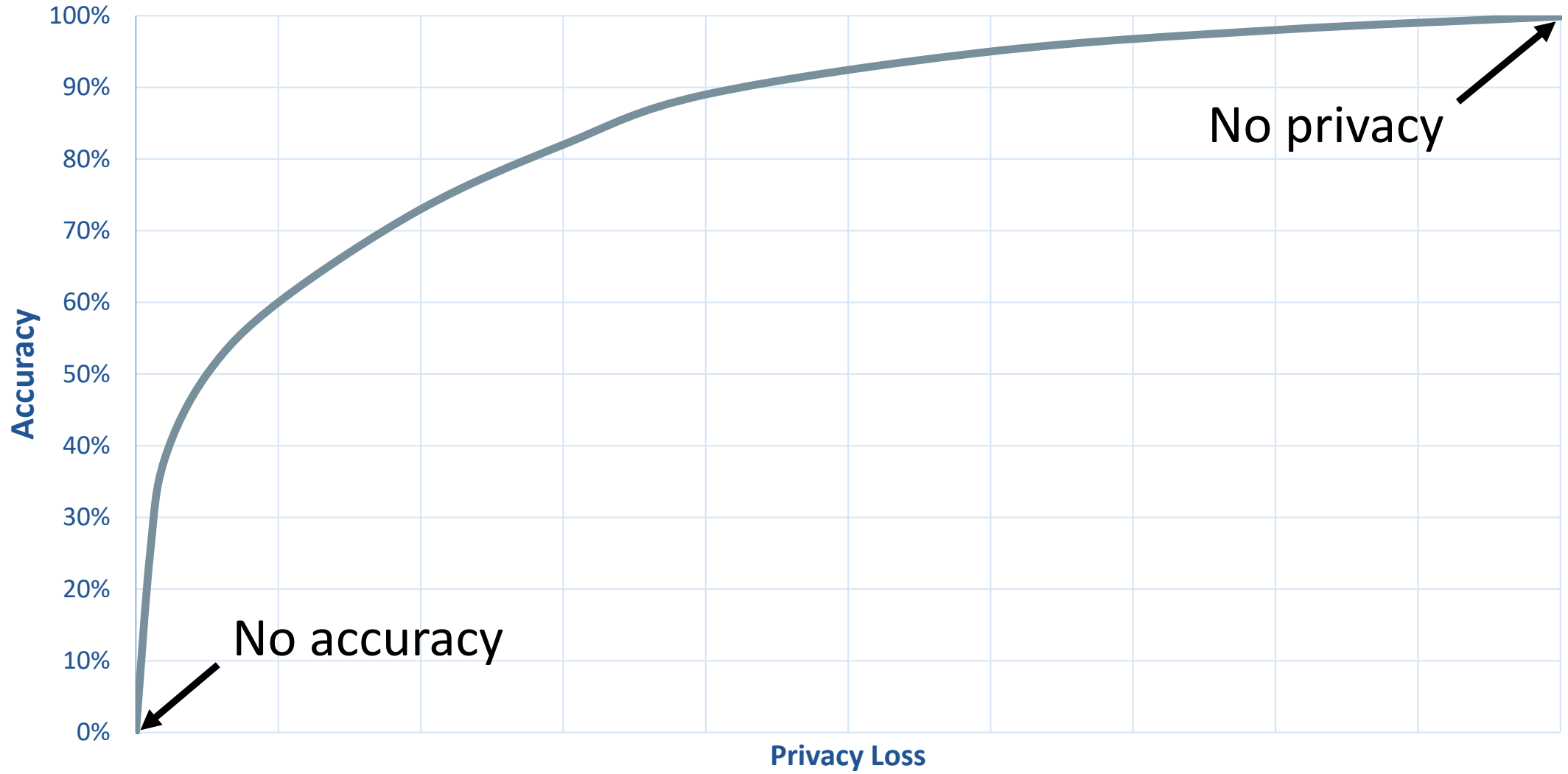
What we found

- Census block and voting age (18+) correctly reconstructed in all 6,207,027 inhabited blocks
- Block, sex, age (in years), race (OMB 63 categories), ethnicity reconstructed
  - Exactly: 46% of population (142 million of 308,745,538)
  - Allowing age +/- one year: 71% of population (219 million of 308,745,538)
- Block, sex, age linked to commercial data to acquire PII
  - Putative re-identifications: 45% of population (138 million of 308,745,538)
- Name, block, sex, age, race, ethnicity compared to confidential data
  - Confirmed re-identifications: 38% of putative (52 million; 17% of population)
- For the confirmed re-identifications, race and ethnicity are learned exactly, not just statistically

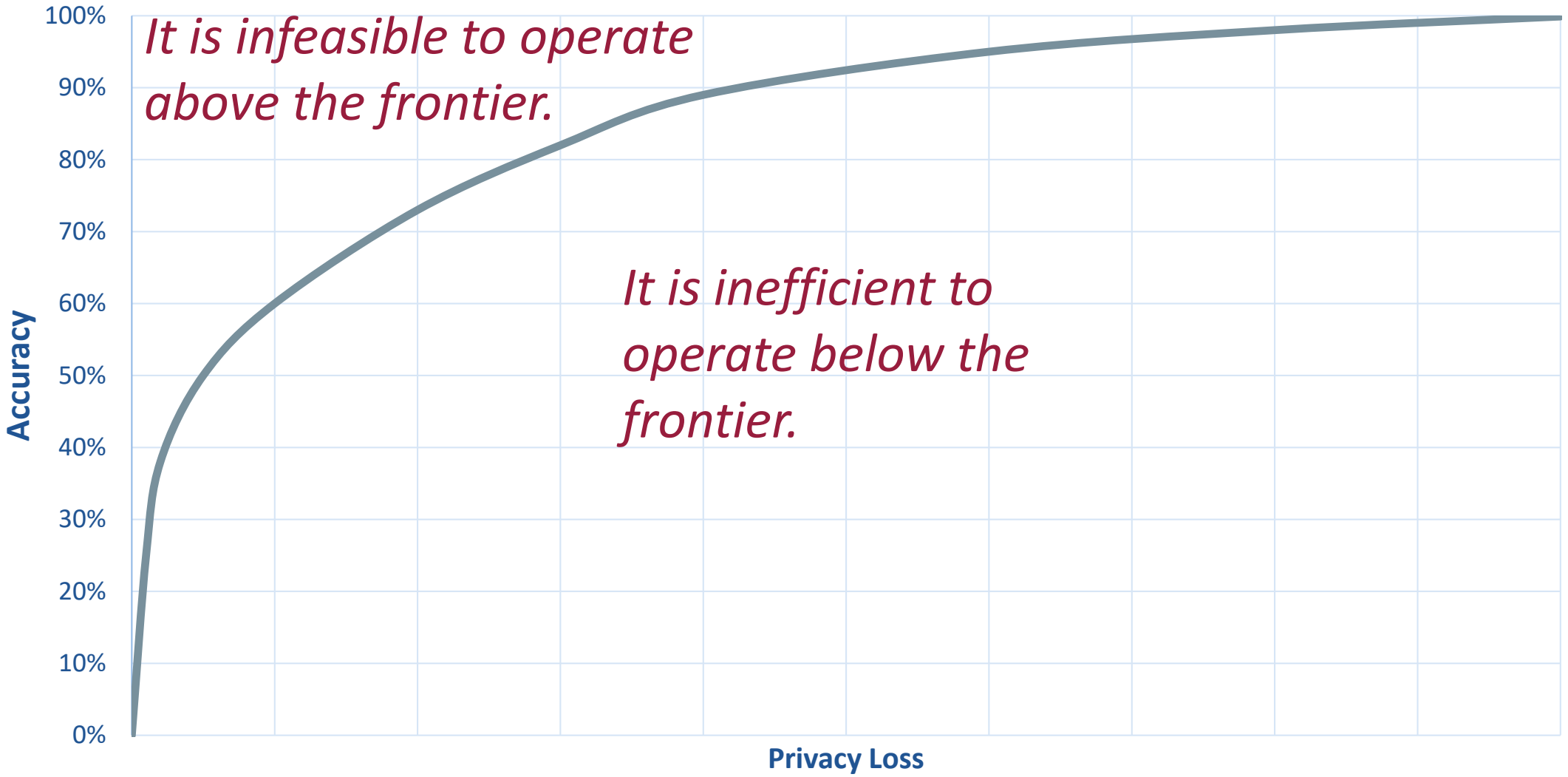
Absolutely the hardest lesson in modern data science is the constraint on publication that the fundamental law of information recovery imposes.

I usually call it the death knell for traditional methods of publication, and not just in statistical agencies.

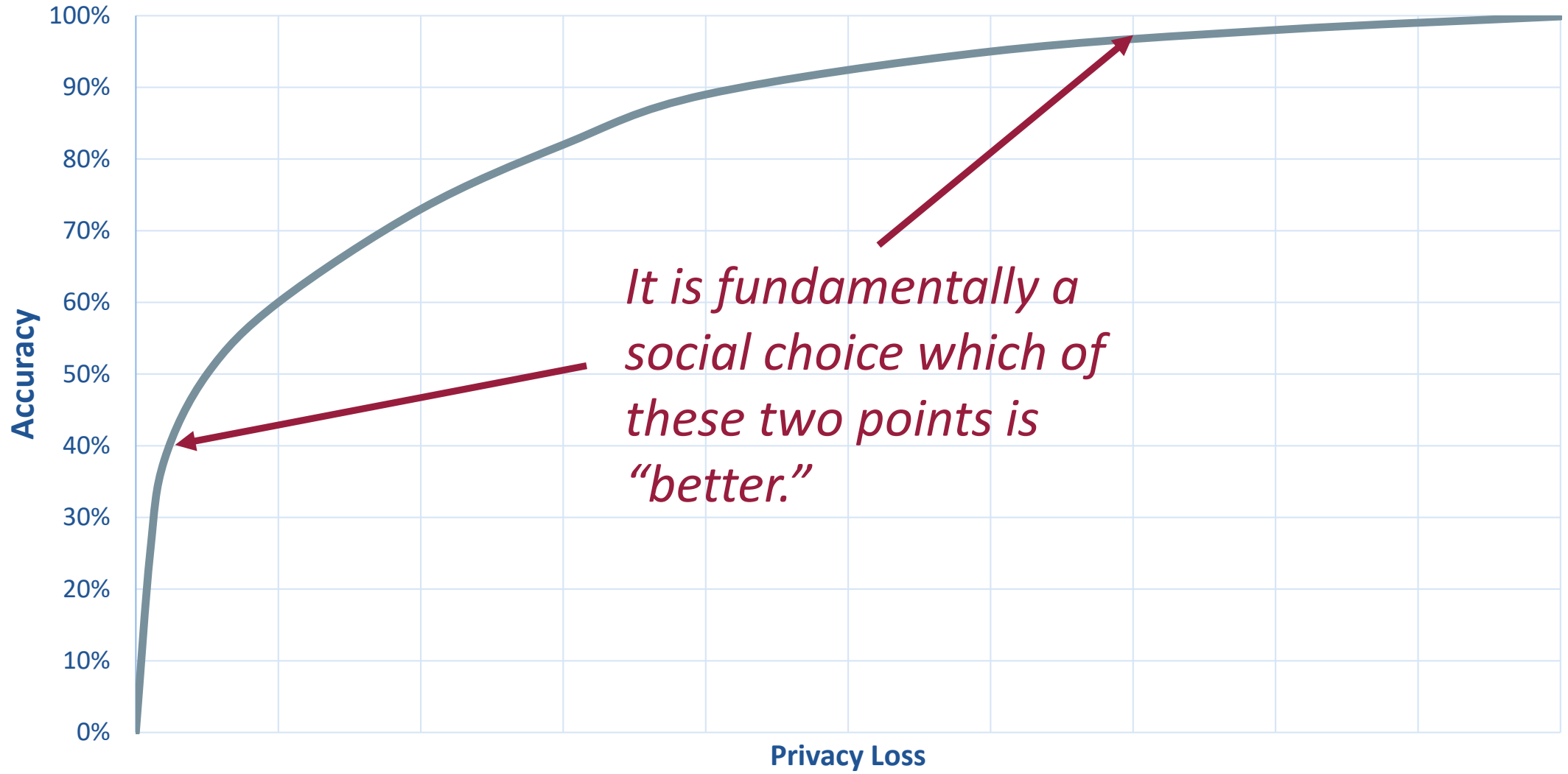
Fundamental Tradeoff betweeen Accuracy and Privacy Loss



Fundamental Tradeoff betweeen Accuracy and Privacy Loss



Fundamental Tradeoff betweeen Accuracy and Privacy Loss



We fixed the database reconstruction vulnerability for the 2020 Census by implementing differential privacy.

The intention is to demonstrate that statistical data, fit for their intended uses, can be produced when the entire publication system is subject to a formal privacy-loss budget.

To date, the team developing these systems—many of whom are in this room—has demonstrated that bounded ε-differential privacy can be implemented for the data publications from the 2020 Census used to re-draw every legislative district in the nation (PL94-171 tables).

And many of the person and household level tables in Summary File 1.

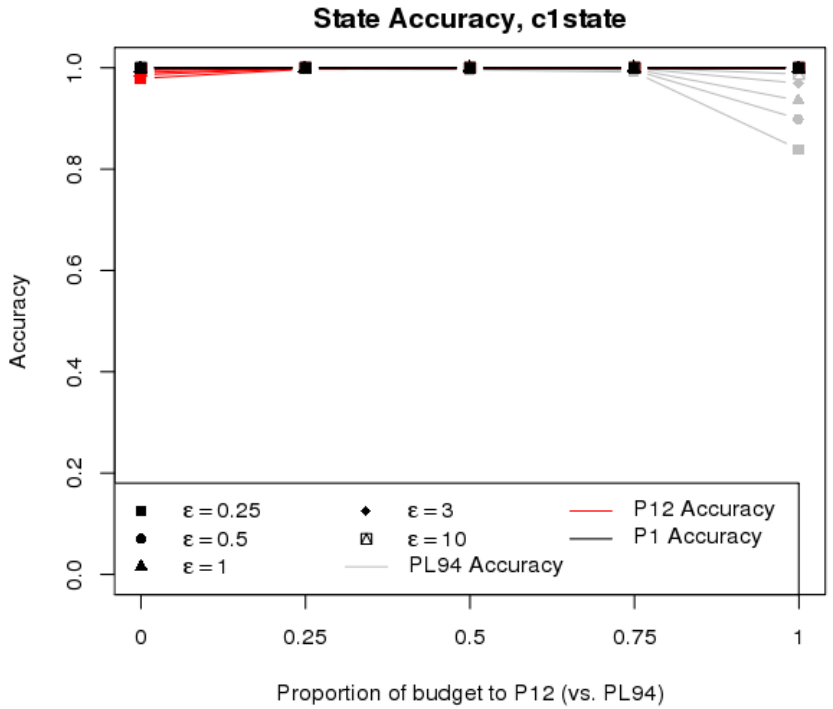
There are close to 100,000,000,000 other queries published from the 2010 Census that are not consistent with a finite privacy-loss budget.

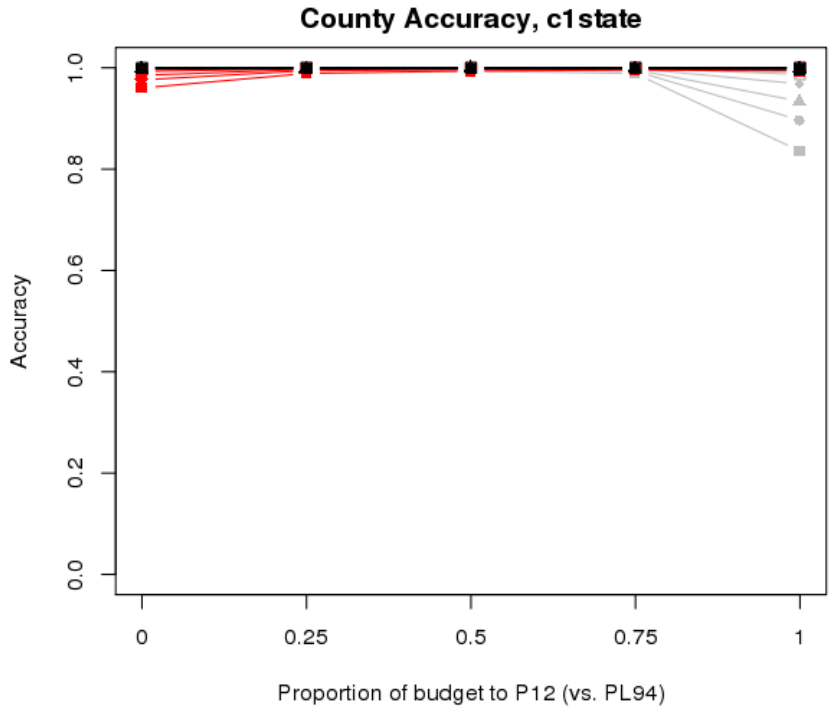
The 2020 Disclosure Avoidance team has also developed methods for quantifying and displaying the system-wide trade-offs between the accuracy of the decennial census data products and the privacy-loss budget assigned to sets of tabulations.

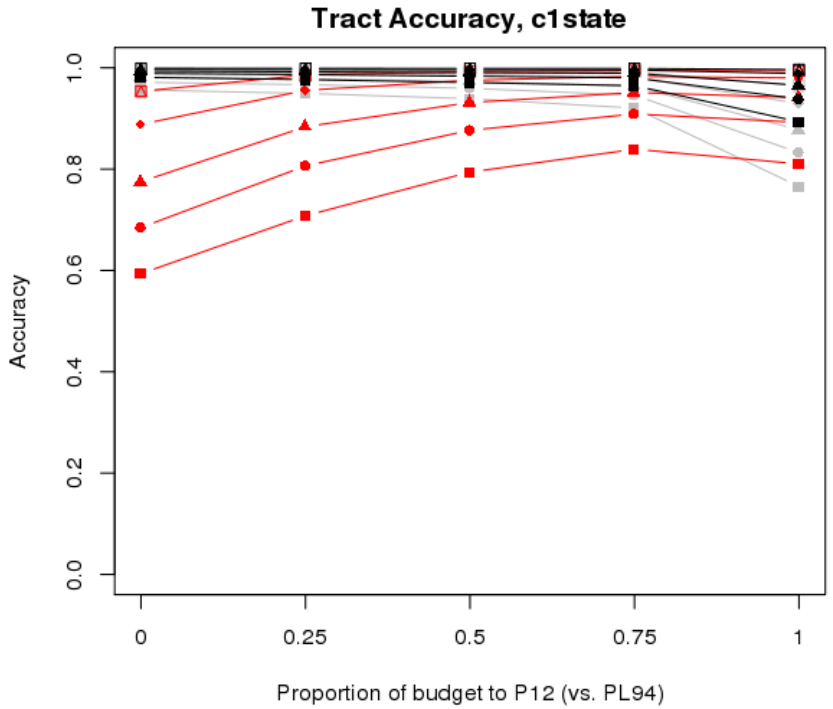
Considering that work began in mid-2016 and that no organization anywhere in the world has yet deployed a full, central differential privacy system, this is already a monumental achievement.

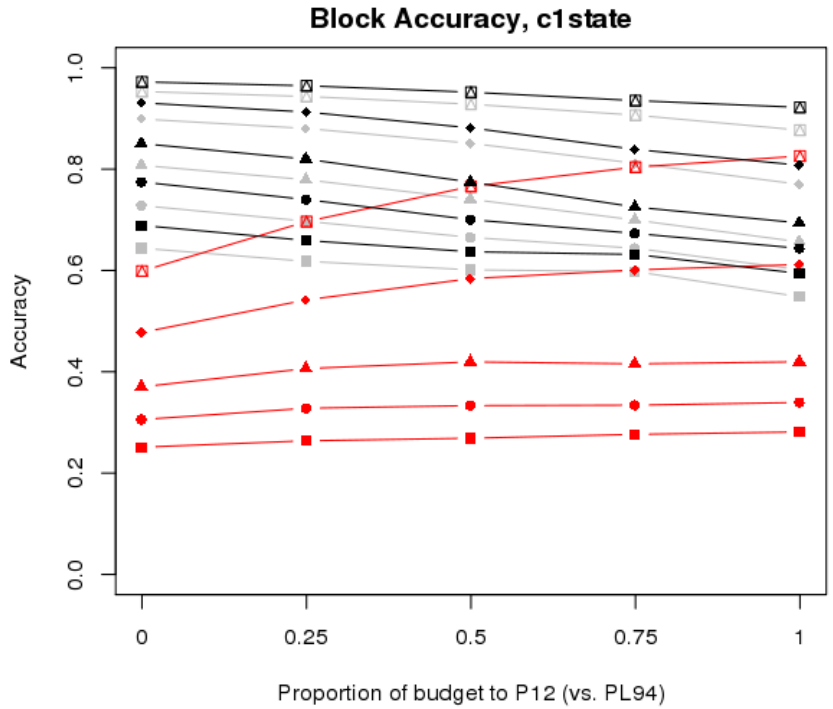
Accuracy v. Privacy Loss for Rhode Island (2010 Census) using the 2018 E2E Test Disclosure Avoidance System

PL94-171: redistricting data, SF1: P12 age x sex data and P1 population data









United States Census Bureau

United States Census Bureau

But it is only the tip of the iceberg.

Demographic profiles, based on the detailed tables traditionally published in summary files following the publication of redistricting data, have far more diverse uses than the redistricting data.

Summarizing those use cases in a set of queries that can be answered with a reasonable privacy-loss budget is the next challenge.

Internet giants, businesses and statistical agencies around the world should also step-up to these challenges. We can learn from, and help, each other enormously.

That’s just the beginning of the story.

What, precisely, should the privacy-loss policy be for all uses of the 2020 Census?

How should we manage invariants?

How should we allocate the privacy-loss budget throughout the next seven decades?

Can we insist that external researchers present their differentially private analysis programs as part of the project review process?

Is so, where do we get the experts to assess the proposals or certify the implementations?

Same process for internal users?

Acknowledgments

The Census Bureau's 2020 Disclosure Avoidance System incorporates work by Daniel Kifer (Scientific Lead), Simson Garfinkel (Senior Scientist for Confidentiality and Data Access), Rob Sienkiewicz (ACC Disclosure Avoidance, Center for Enterprise Dissemination), Tamara Adams, Robert Ashmead, Michael Bentley, Stephen Clark, Craig Corl, Aref Dajani, Nathan Goldschlag, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Brett Moran, Edward Porter, Sarah Powazek, Anne Ross, Ian Schmutte, William Sexton, Lars Vilhuber, Cecil Washington, and Pavel Zhuralev

More Background on the 2020 Census Disclosure Avoidance System

- September 14, 2017 CSAC (overall design) https://www2.census.gov/cac/sac/meetings/2017-09/garfinkel-modernizing-disclosure-avoidance.pdf?#
- August, 2018 KDD'18 (top-down v. block-by-block) https://digitalcommons.ilr.cornell.edu/ldi/49/
- October, 2018 WPES (implementation issues) https://arxiv.org/abs/1809.02201
- October, 2018 *ACMQueue* (understanding database reconstruction) https://digitalcommons.ilr.cornell.edu/ldi/50/ or https://queue.acm.org/detail.cfm?id=3295691
- December 6, 2010 CSAC (detailed discussion of algorithms and choices) https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf?#

Thank you.

John.Maron.Abowd@census.gov