Official Statistics at the Crossroads: Data Quality and Access in an Era of Heightened Privacy Risk

John M. Abowd
Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau
Joint Statistical Meetings
Panel: Tuesday, July 30, 2019 2:00-3:50pm

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. Statistics from the 2010 Census for Rhode Island authorized under DRB release CBDRB-FY19-054.

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

Generous estimate: 100GB of data from 2020 Census

Less than 1% of worldwide mobile data use/second

(Source: Cisco VNI Mobile, February 2019 estimate: 11.8TB/second, 29EB/month, mobile data traffic worldwide https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html#_Toc953327.)

The Census Bureau's data stewardship problem looks very different from the one at Amazon, Apple, Facebook, Google, Microsoft, Netflix …

… but appearances are deceiving.

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 correctly, although the attacker may still have uncertainty

Almost everyone in this room knows that:

Comparing common features allows highly reliable entity resolution (these features belong to the same entity)

Machine learning systems build classifiers, recommenders, and demand management systems that use these amplified entity records

All of this is much harder with provable privacy guarantees for the entities!

The Census Bureau's 150B tabulations from 15GB of data …

…and tech industry's data integration and deep-learning AI systems

are both subject to the fundamental economic problem inherent in privacy protection.

Traditional SDL is broken.

That’s not the same as failure.

More than vigilance is required.

Rethinking is essential.

The status quo is no longer an option going forward.

Privacy protection is an economic problem.

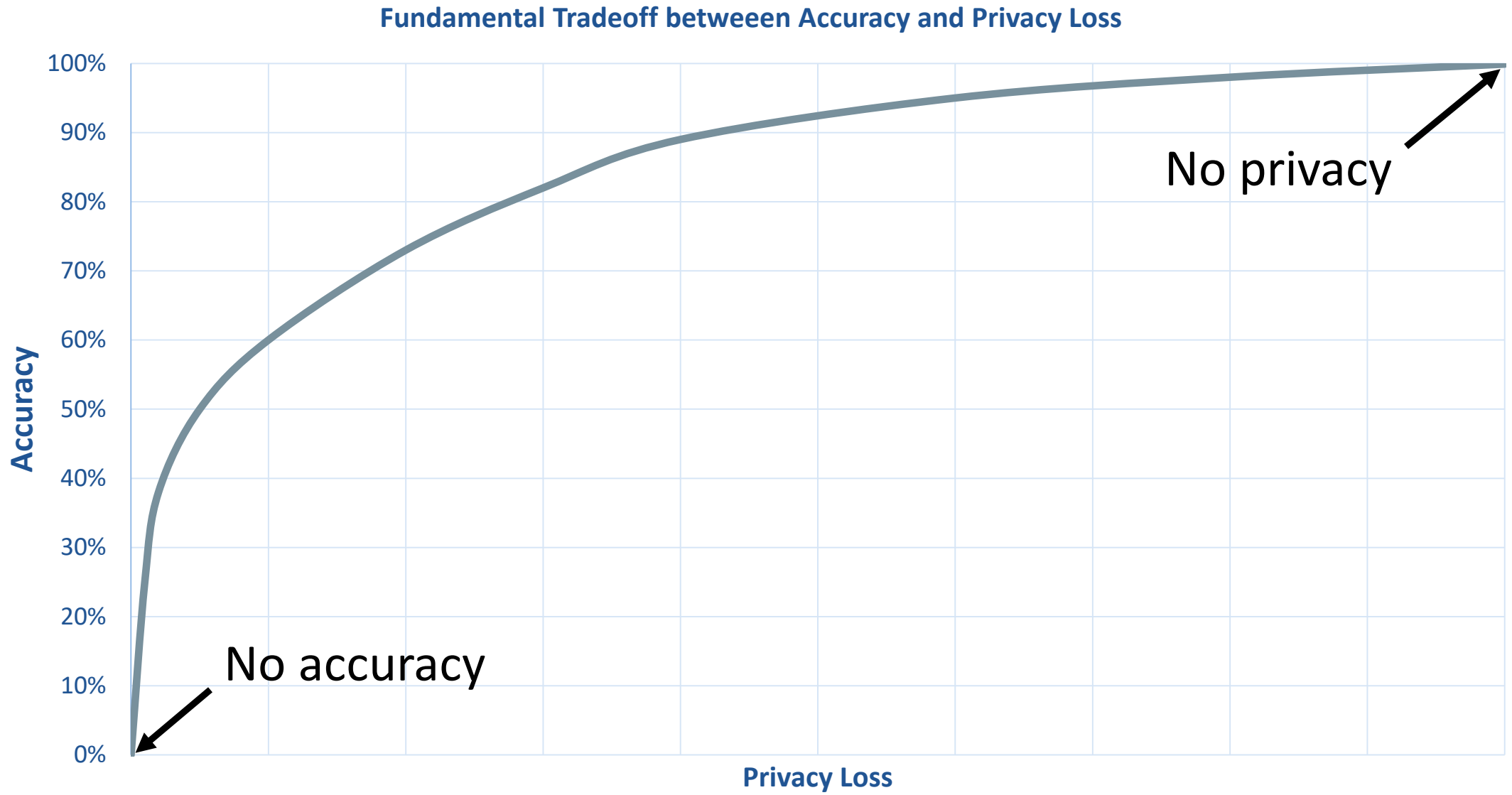
Not a technical problem in computer science or statistics.

Allocation of a scarce resource (data in the confidential database) between competing uses:

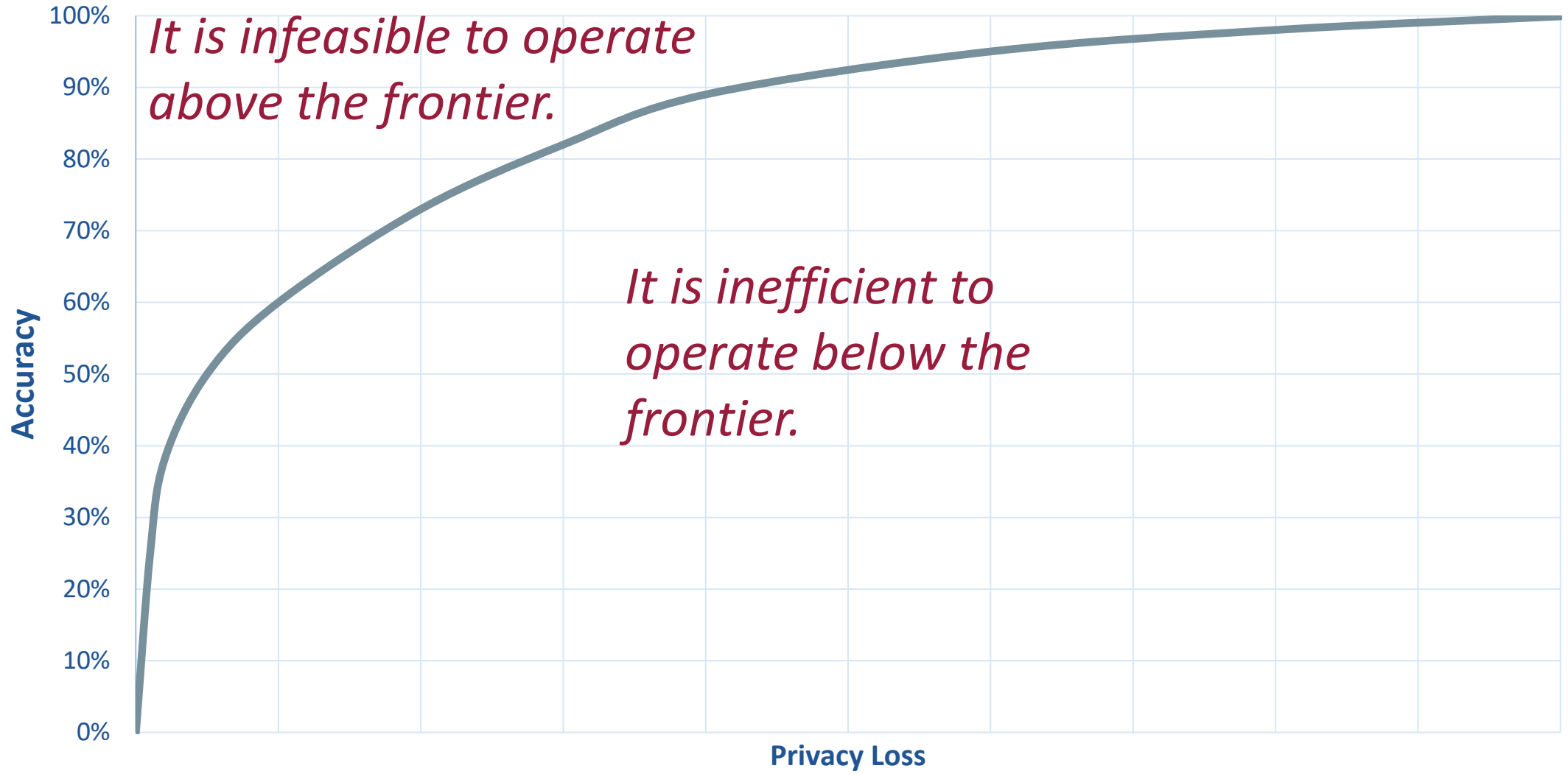
information products

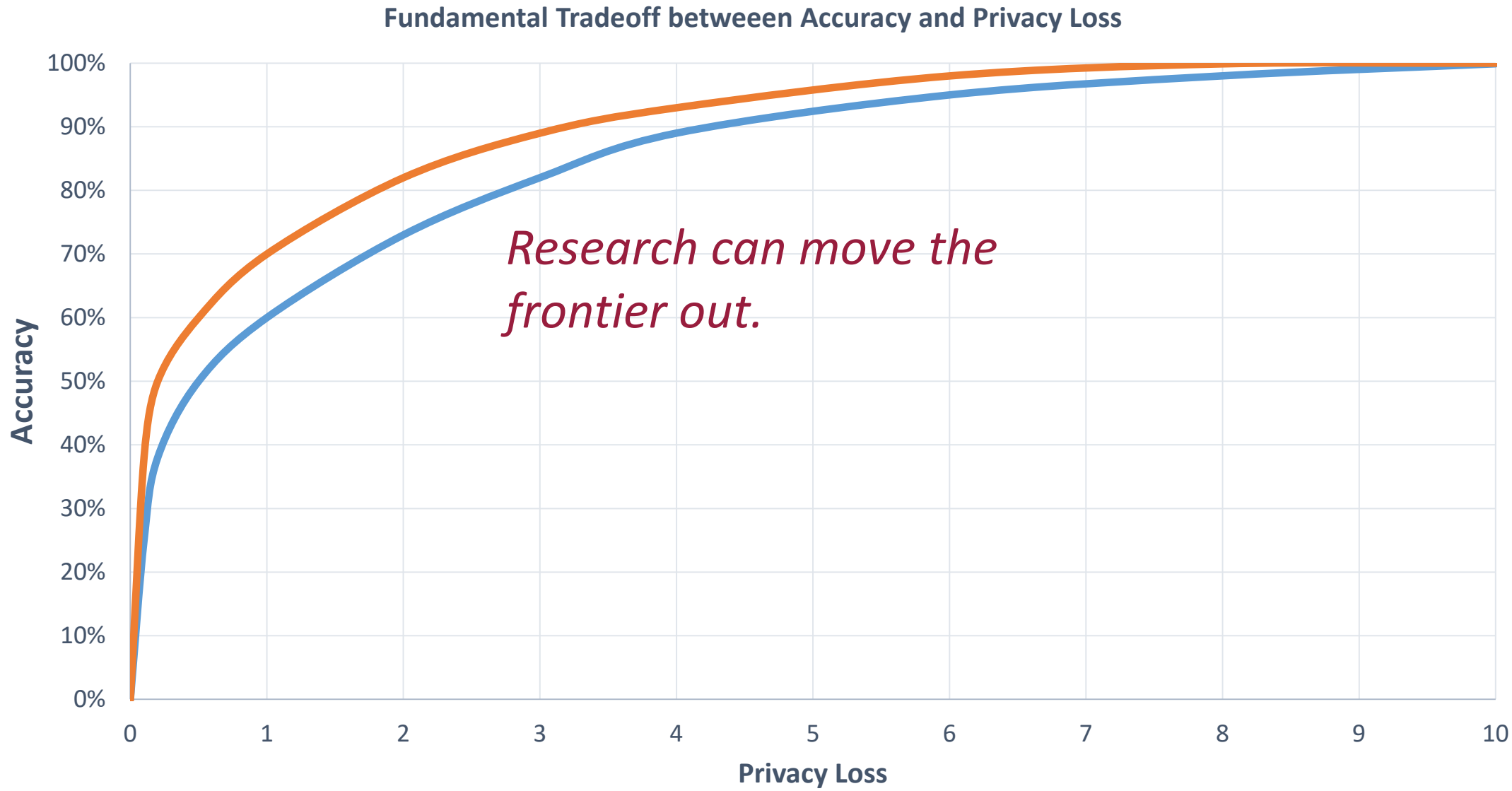
and

privacy protection.


Fundamental Tradeoff betweeen Accuracy and Privacy Loss
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Accuracy
No privacy
No accuracy
Privacy Loss

Fundamental Tradeoff betweeen Accuracy and Privacy Loss




Fundamental Tradeoff betweeen Accuracy and Privacy Loss
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Accuracy
0
1
2
3
4
5
6
7
8
9
10
Privacy Loss
Research can move the frontier out.

As with many economic problems, the use of comparative advantage is essential to define the production possibilities, but a social welfare function is required to assess optimal solutions.

But the *CS technology does define the feasible trade-offs*.

It is not a blunt instrument.

It permits estimation of the production possibilities frontier.

Just like guns and butter from your intro micro class.

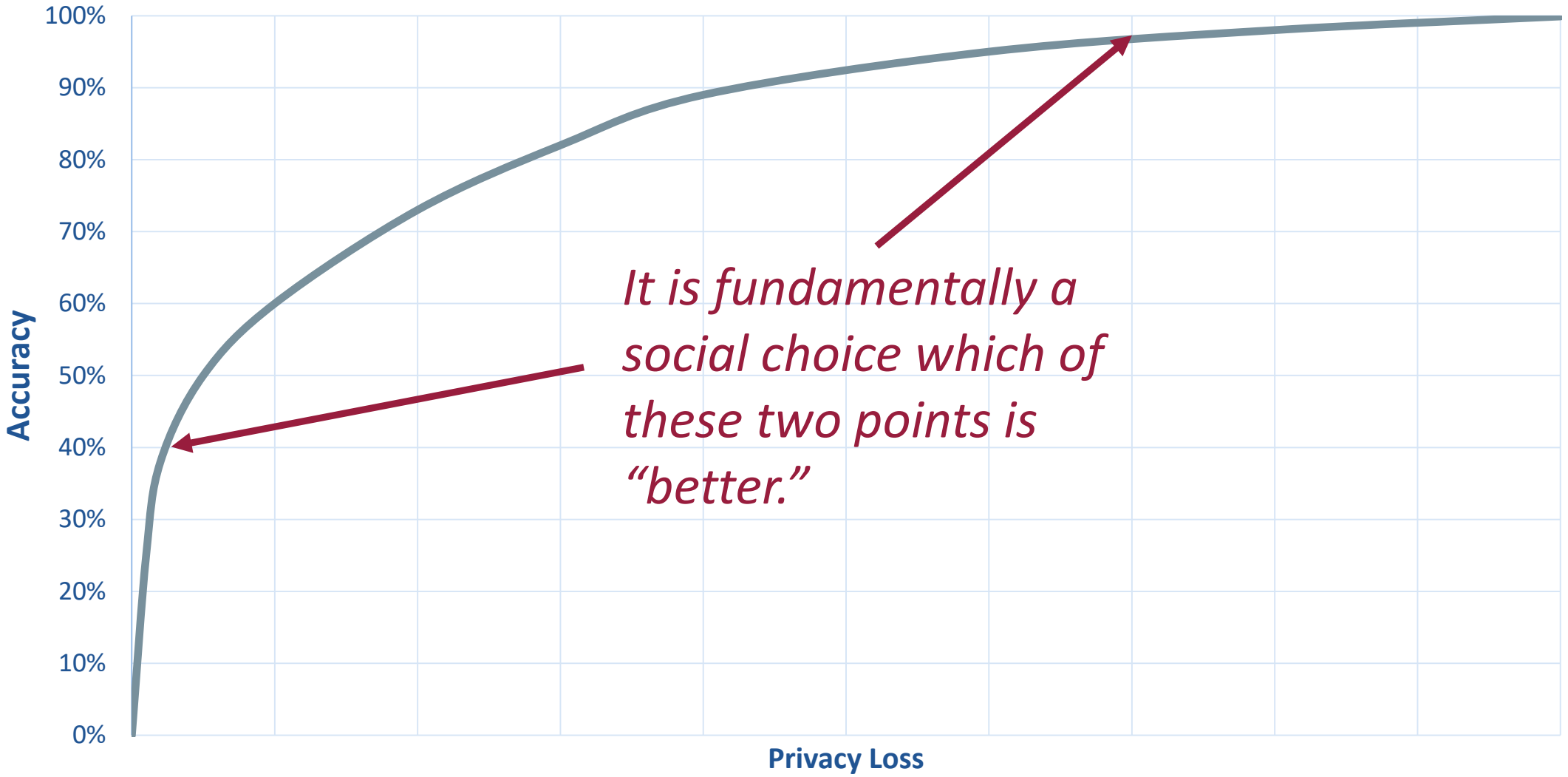
Look yourself in the mirror.

I have.

How much weight do you put on privacy protection compared to accuracy?

Who is supposed to represent the privacy interests when many decision makers share this tendency to weight accuracy very heavily?

Fundamental Tradeoff betweeen Accuracy and Privacy Loss



If Facebook said …

"If you think you have re-identified someone in public data that we released for research purposes, you can't be sure that you are correct because we used disclosure limitation techniques for which we cannot give you the details."

What would you say?

If a researcher wrote …

"My inferences may not be valid because the agency that provided access did not release details sufficient to correct for bias and variability due to statistical disclosure limitation."

What would you say?

The Census Bureau confronted the economic problem inherent in the database reconstruction vulnerability for the 2020 Census by implementing formal privacy guarantees relying on a core of differentially private subroutines that assign:

the *technology to the 2020 Disclosure Avoidance System* team,

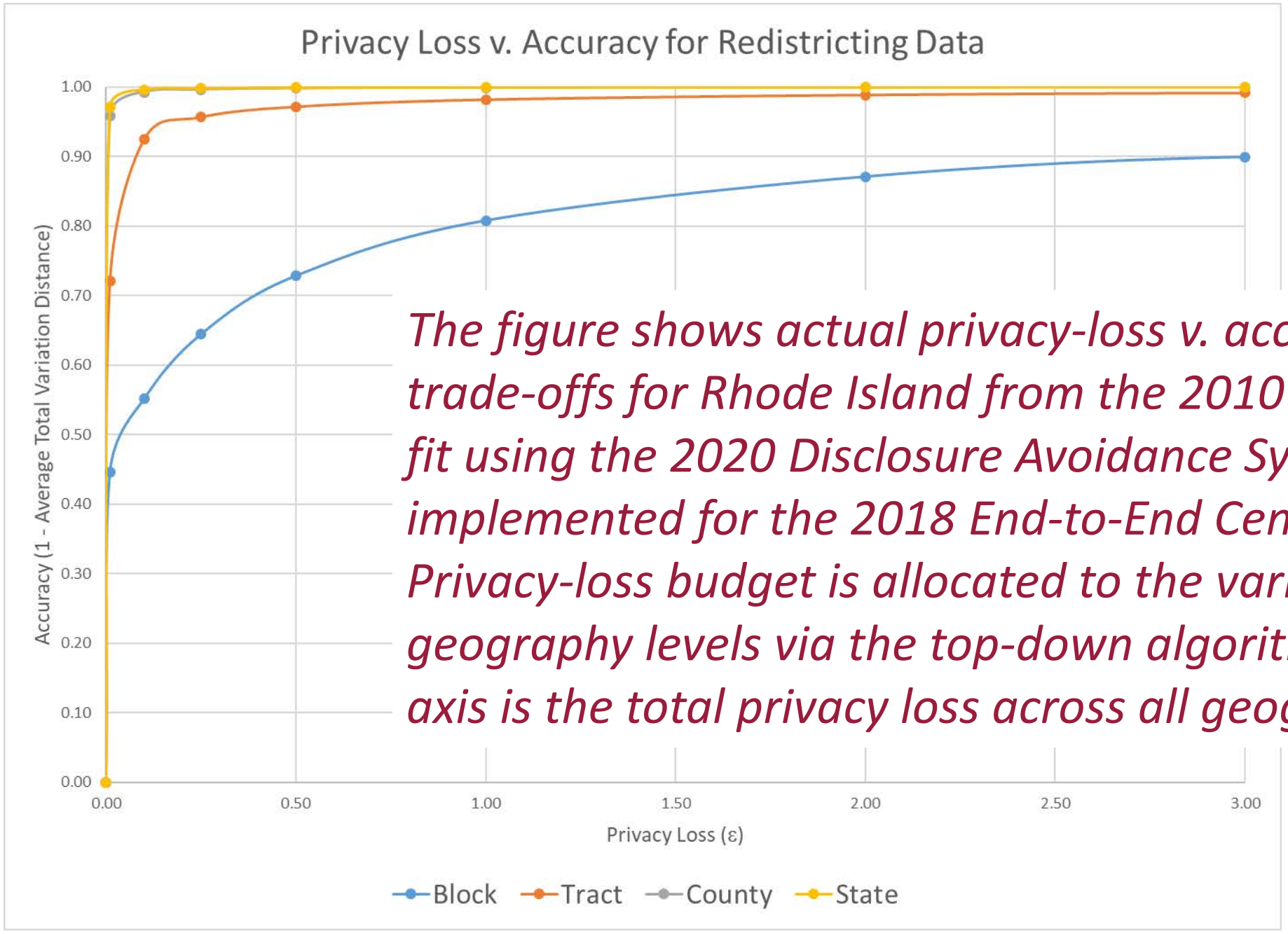
the *policy to the Data Stewardship Executive Policy* committee.

- No final decisions have been made regarding the privacy-loss budget for the 2020 Census
- Test products released from the 2018 End-to-End Census Test used a privacy-loss budget of 0.25 for reasons documented here: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series/2020-memo-2019_13.html
- Harvard Data Science Review workshop on October 25
- More test products (early fall) and CNSTAT workshop (November 21-22) coming

Thank you.

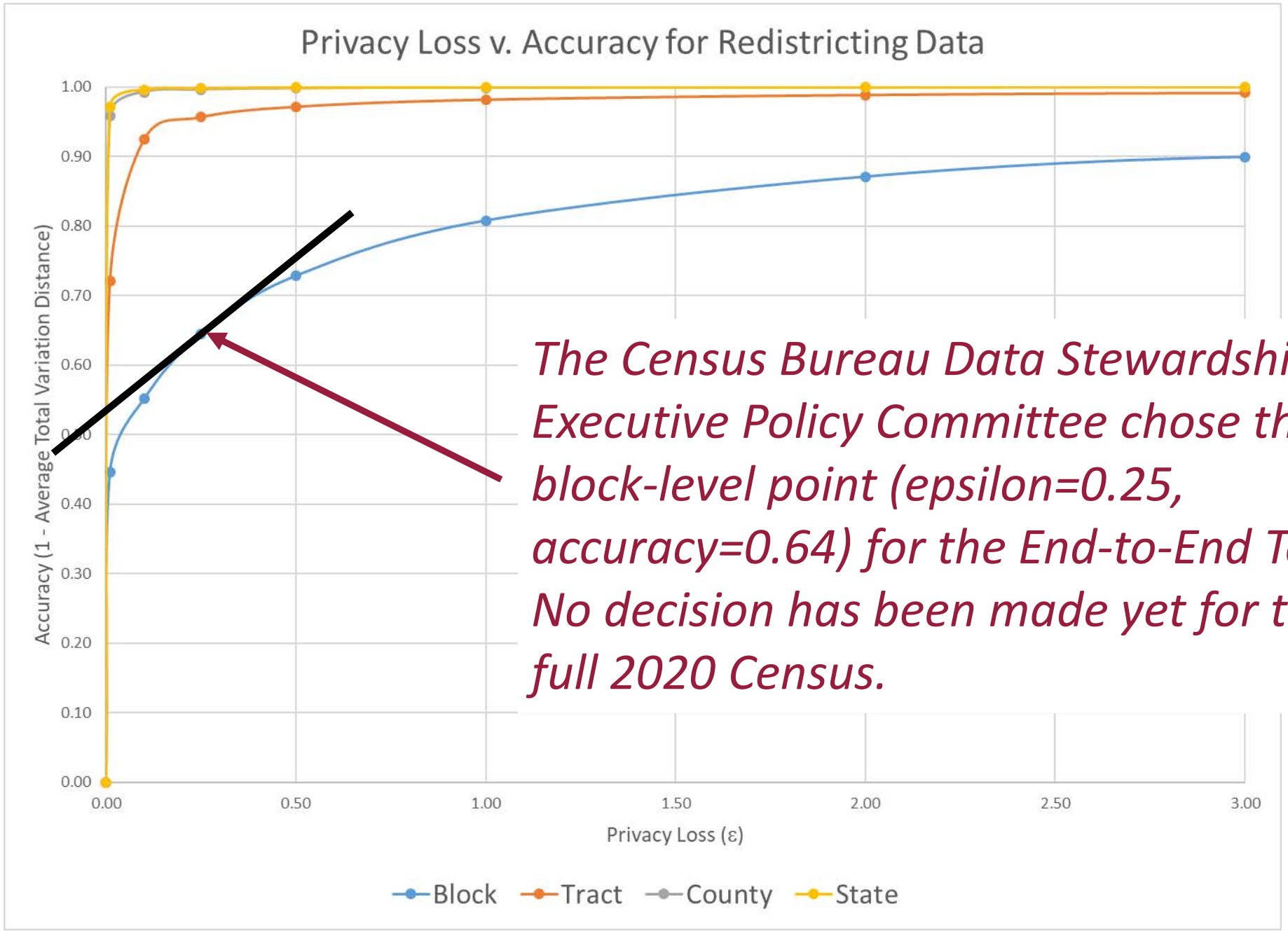
John.Maron.Abowd@census.gov

Backup slides



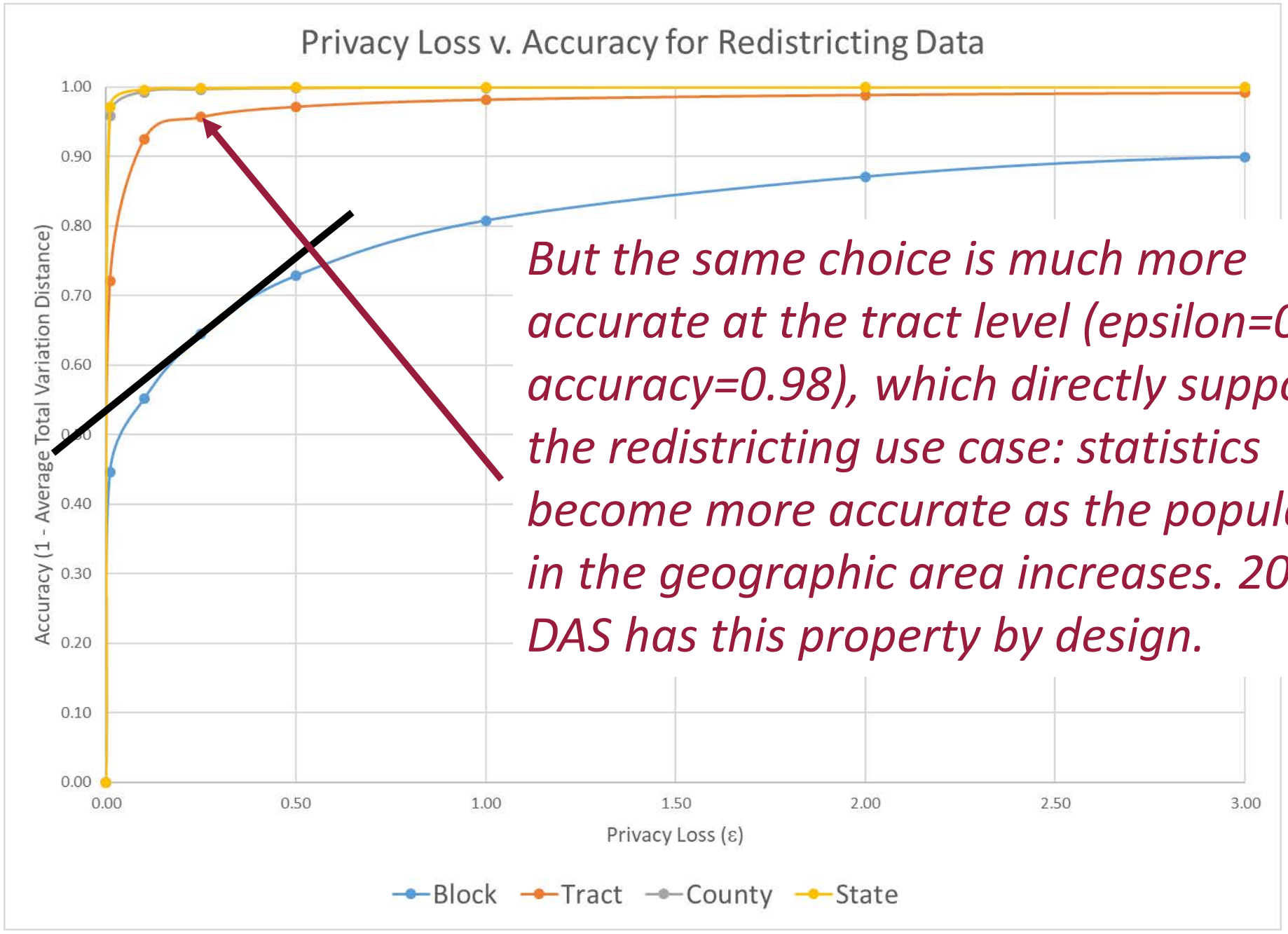
The figure shows actual privacy-loss v. accuracy trade-offs for Rhode Island from the 2010 Census, fit using the 2020 Disclosure Avoidance System as implemented for the 2018 End-to-End Census Test. Privacy-loss budget is allocated to the various geography levels via the top-down algorithm. X-axis is the total privacy loss across all geographies.

United States Census Bureau



The Census Bureau Data Stewardship Executive Policy Committee chose this block-level point (epsilon=0.25, accuracy=0.64) for the End-to-End Test. No decision has been made yet for the full 2020 Census.

United States Census Bureau



But the same choice is much more accurate at the tract level (epsilon=0.25, accuracy=0.98), which directly supports the redistricting use case: statistics become more accurate as the population in the geographic area increases. 2020 DAS has this property by design.

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.

Science and policy must address these questions too:

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

How should the Census Bureau handle management-imposed accuracy requirements?

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

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

If so, where do the experts to assess the proposals or certify the implementations come from?

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, 2018 CSAC (detailed discussion of algorithms and choices) https://www2.census.gov/cac/sac/meetings/2018-12/abowd-disclosure-avoidance.pdf?#
- June 6, 2019 Blog explaining how to use the 2018 End-to-End Census Test version of the 2020 Disclosure Avoidance System with the 1940 Census public data from IPUMS https://www.census.gov/newsroom/blogs/research-matters/2019/06/disclosure_avoidance.html