

The U.S. 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

9th Annual FDIC Consumer Research Symposium

Distinguished Guest Lecture: Friday, October 18, 2019 1:15-2:00pm

The views expressed in this talk are my own and not those of the U.S. Census Bureau. Examples from the 1940 Census are based on public-use micro-data.

Acknowledgments

The Census Bureau's 2020 Disclosure Avoidance System incorporates work by Daniel Kifer (Scientific Lead), Simson Garfinkel (Senior Computer Scientist for Confidentiality and Data Access), Rob Sienkiewicz (Chief, Center for Enterprise Dissemination), Tamara Adams, Robert Ashmead, Stephen Clark, Craig Corl, Aref Dajani, Jason Devine, 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, and Pavel Zhuralev.



https://www.census.gov/about/policies/privacy/statistical_safeguards.html



Minute Physics

This video was produced in collaboration with the US Census Bureau and fact-checked by Census Bureau scientists; any opinions and errors are my own.

MORE VIDEOS
Play (k)



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

Major data products:

- Apportion the House of Representatives
(due December 31, 2020)
- Supply data to all state redistricting offices
(due April 1, 2021)
- Demographic and housing characteristics
(no statutory deadline, target summer 2021)
- Detailed race and ethnicity data
(no statutory deadline)
- American Indian, Alaska Native, Native Hawaiian data
(no statutory deadline)

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

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.

The Database Reconstruction Vulnerability

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.*

Privacy protection is an economic problem.

Not a technical problem in computer science or statistics.

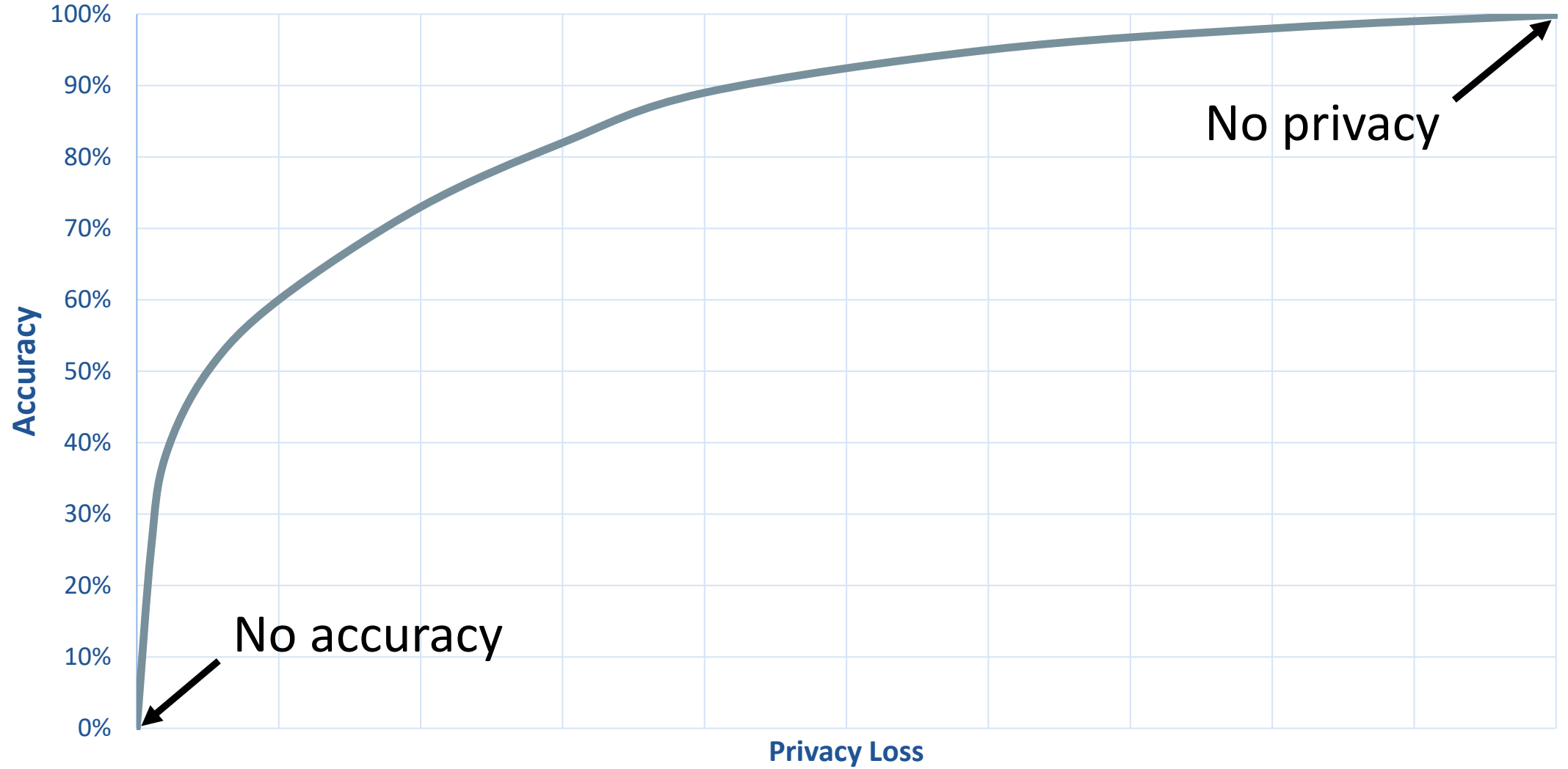
Allocation of a scarce resource (data in the confidential database) between competing uses:

information products

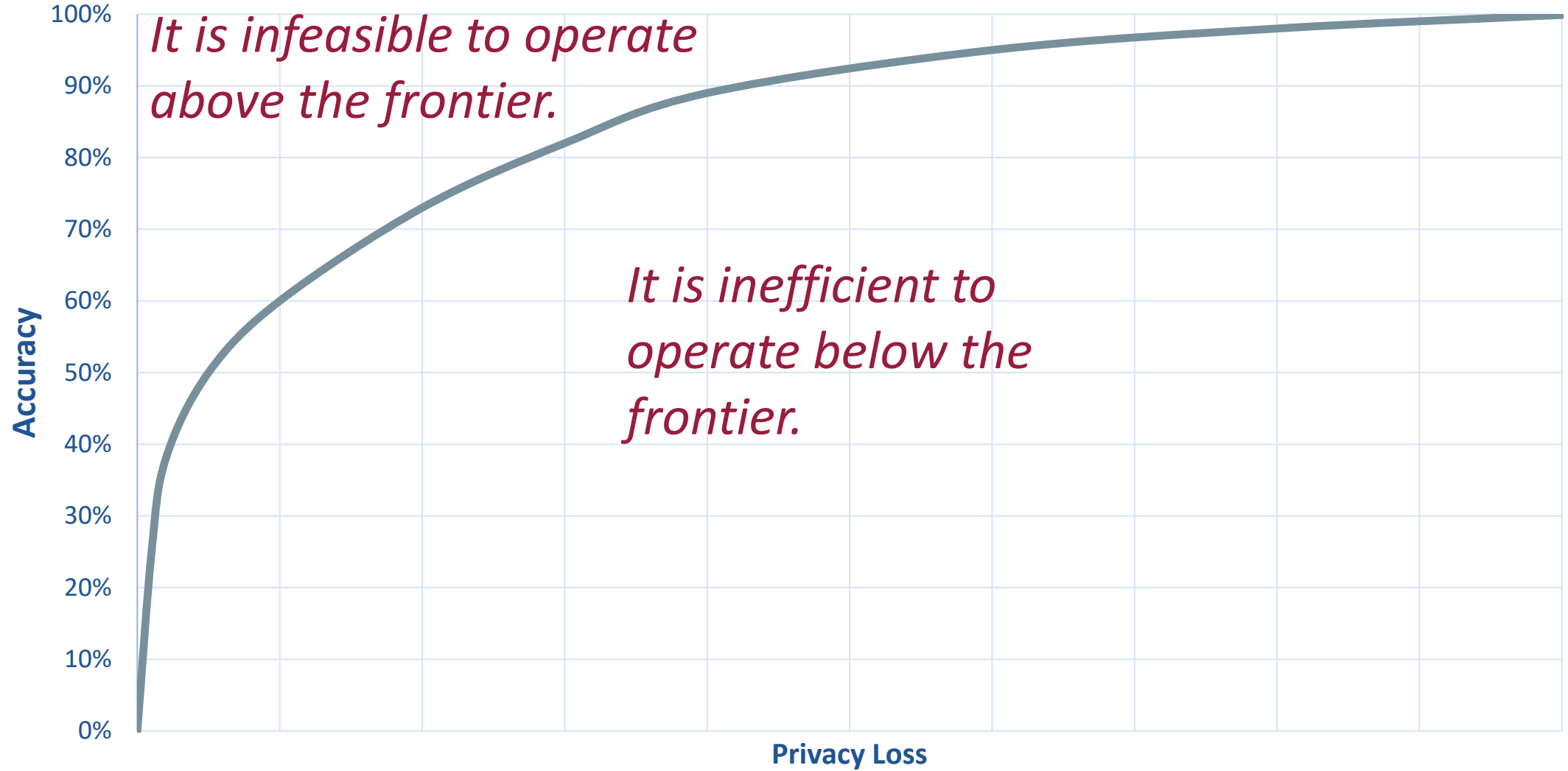
and

privacy protection.

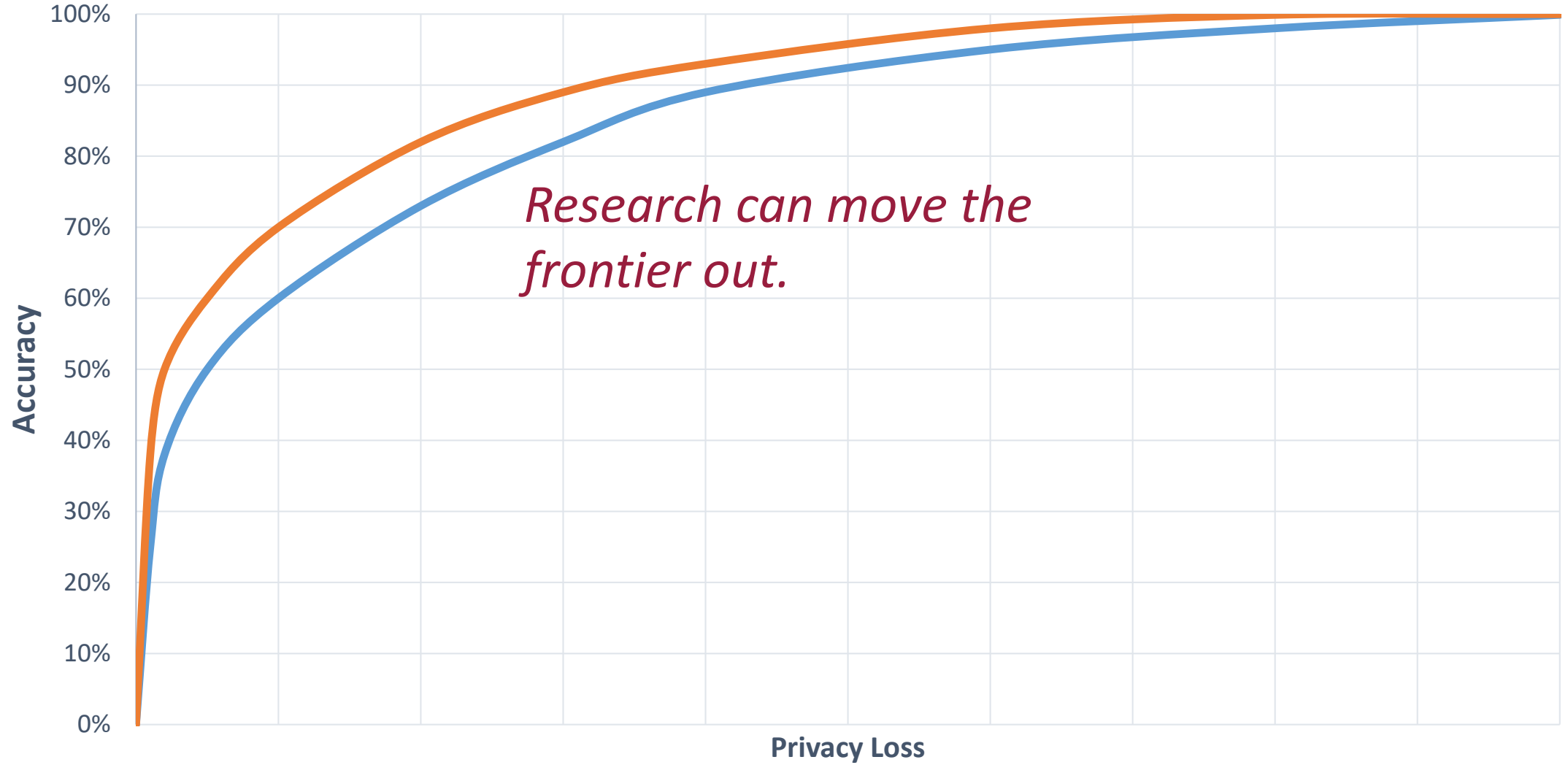
Fundamental Tradeoff between Accuracy and Privacy Loss



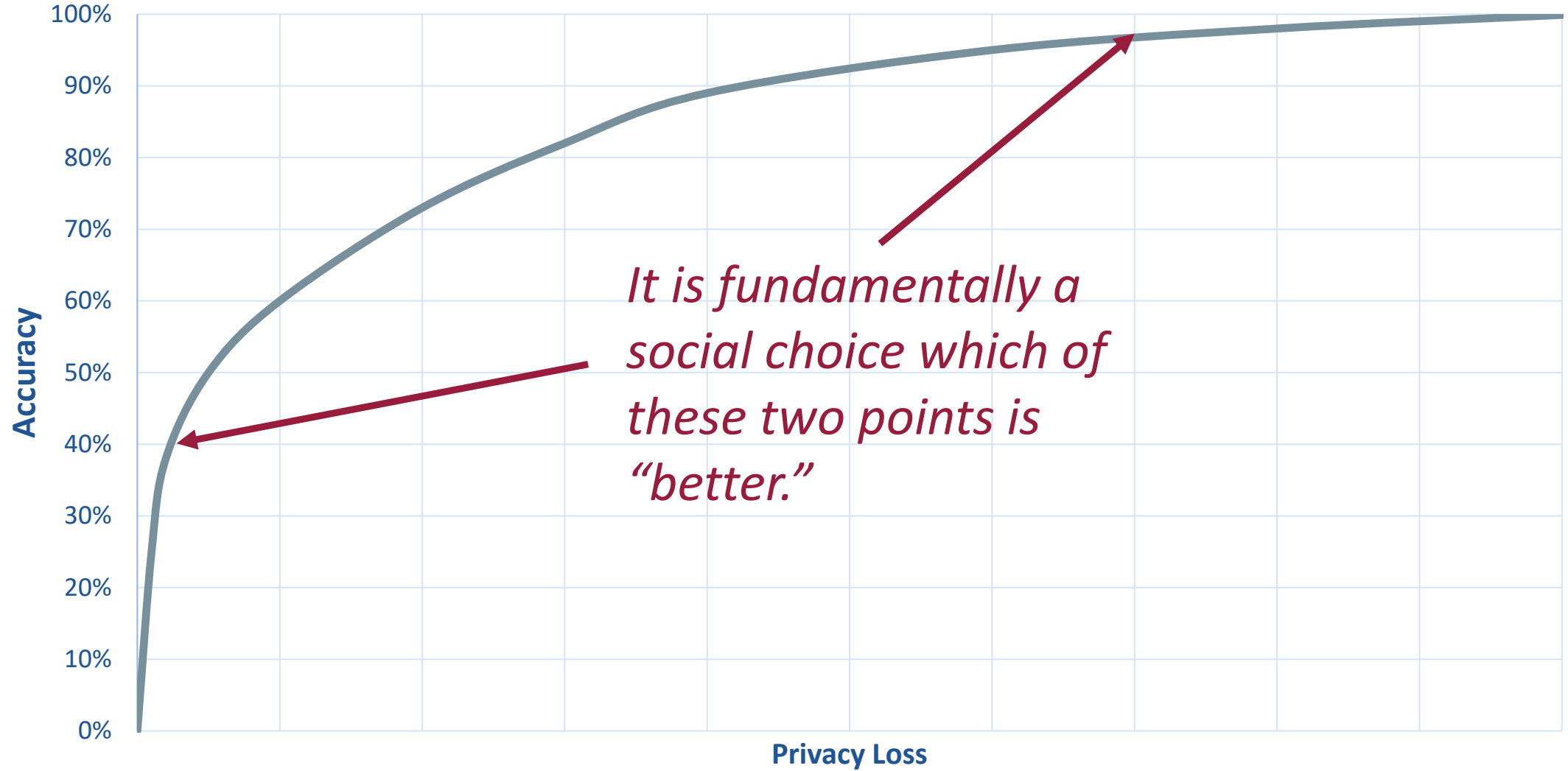
Fundamental Tradeoff between Accuracy and Privacy Loss



Fundamental Tradeoff between Accuracy and Privacy Loss



Fundamental Tradeoff between Accuracy and Privacy Loss



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.

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 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 the demographic and housing characteristics.

But there are more than **100 billion** other queries published from the 2010 Census that are not easy to make 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.

Now, let's see how that system works.

Algorithms Matter

The TopDown Algorithm

National table of US population

$2 \times 126 \times 24 \times 115 \times 2$

Spend ϵ_1 privacy-loss budget

National table with all 1.5M cells filled, structural zeros imposed with accuracy allowed by ϵ_1
 $2 \times 126 \times 24 \times 115 \times 2$

Sex: Male / Female
Race + Hispanic: 126 possible values
Relationship to Householder/GQ: 24
Age: 0-114

Reconstruct individual micro-data without geography

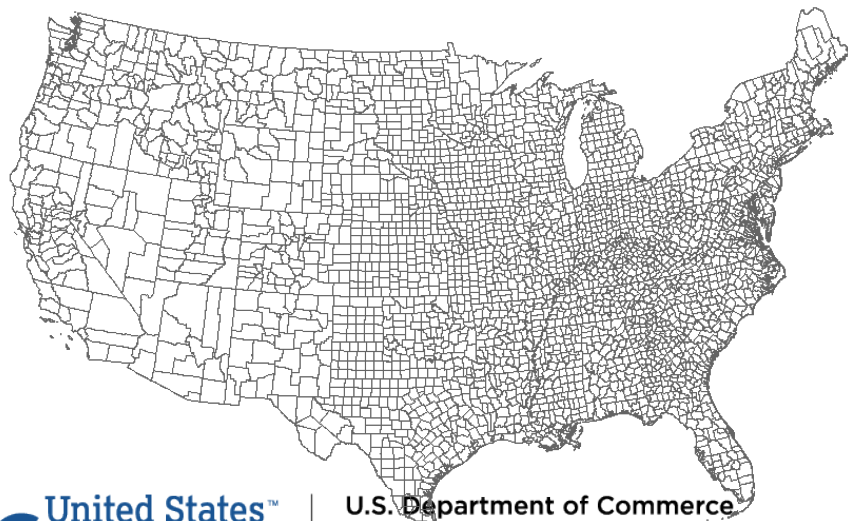
330,000,000 records

County-level

County-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and DHC-P

Spend ϵ_3 privacy-loss budget

Target county-level tables required for best accuracy for PL-94 and DHC-P



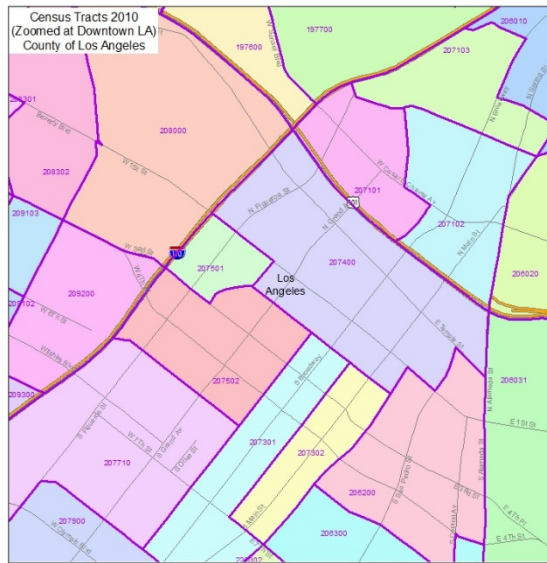
Construct best-fitting individual micro-data with state and county geography
330,000,000 records now including state and county identifiers

Census tract-level

Tract-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and DHC-P

Spend ϵ_4 privacy-loss budget

Target tract-level tables required for best accuracy for PL-94 and DHC-P


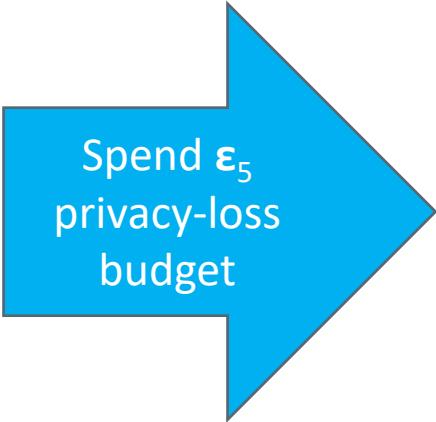


Construct best-fitting individual micro-data with state, county, and tract geography


330,000,000 records now including state, county, and tract identifiers

Block-level

Block-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and DHC-P



Target **Block** tables required for best accuracy for PL-94 and DHC-P



Construct best-fitting individual micro-data with **state, county, tract and block** geography

330,000,000 records now including **state, county, tract, and block** identifiers



Tabulation micro-data

Construct best-fitting individual micro-data with
state, county, tract and block geography

330,000,000 records now including state,
county, tract, and block identifiers



Micro-data used for
tabulating PL-94 and DHC-P

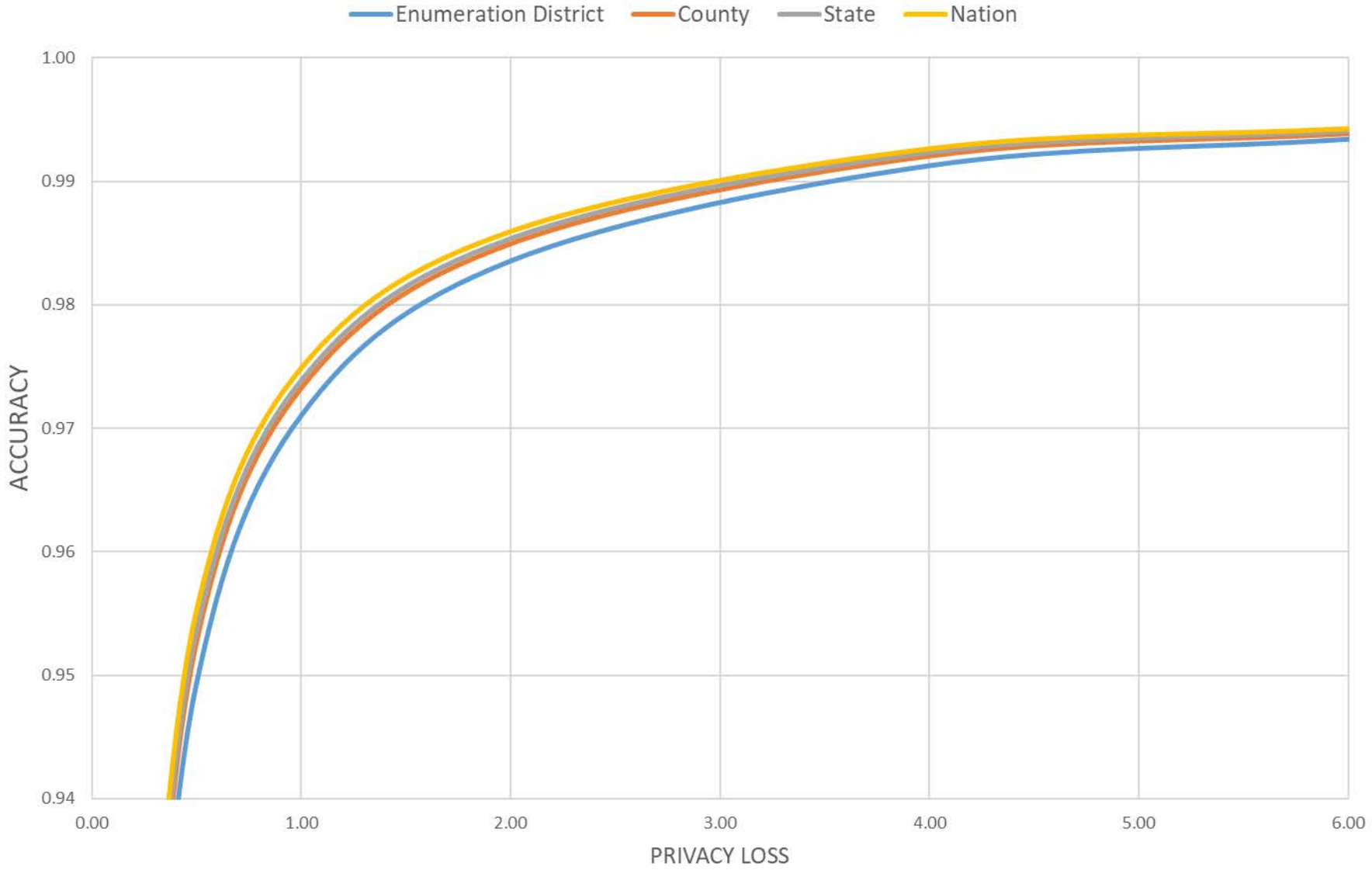
Method Summary

- Take differentially private measurements at every level of the hierarchy
- At each level of TopDown post-process:
 - Solve an L2 optimization to get non-negative tables
 - Solve an L1 optimization to get non-negative, integer tables
 - Generate micro-data from the post-processed tables

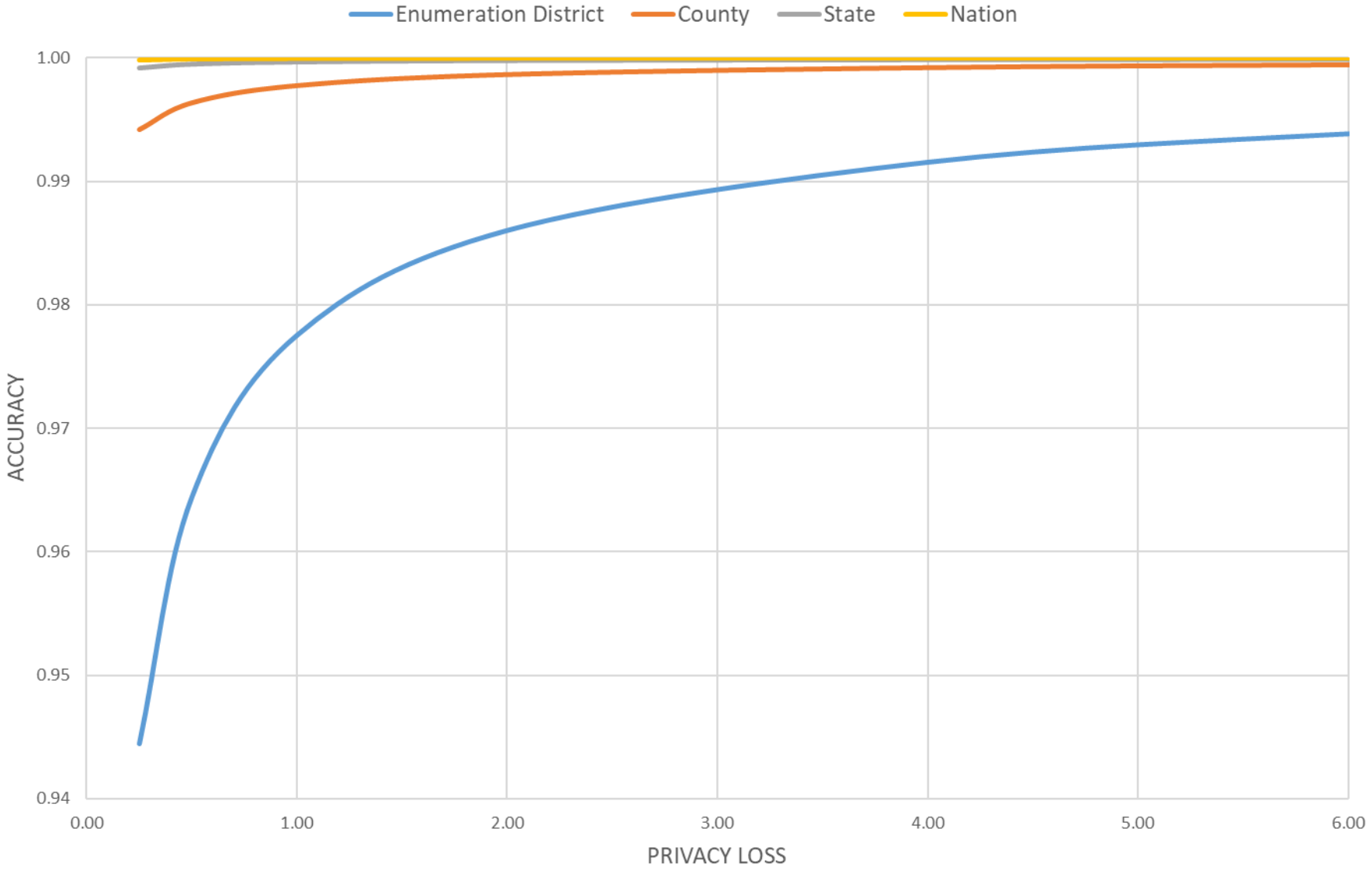
Naïve Method: BottomUp or Block-by-Block

- 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.

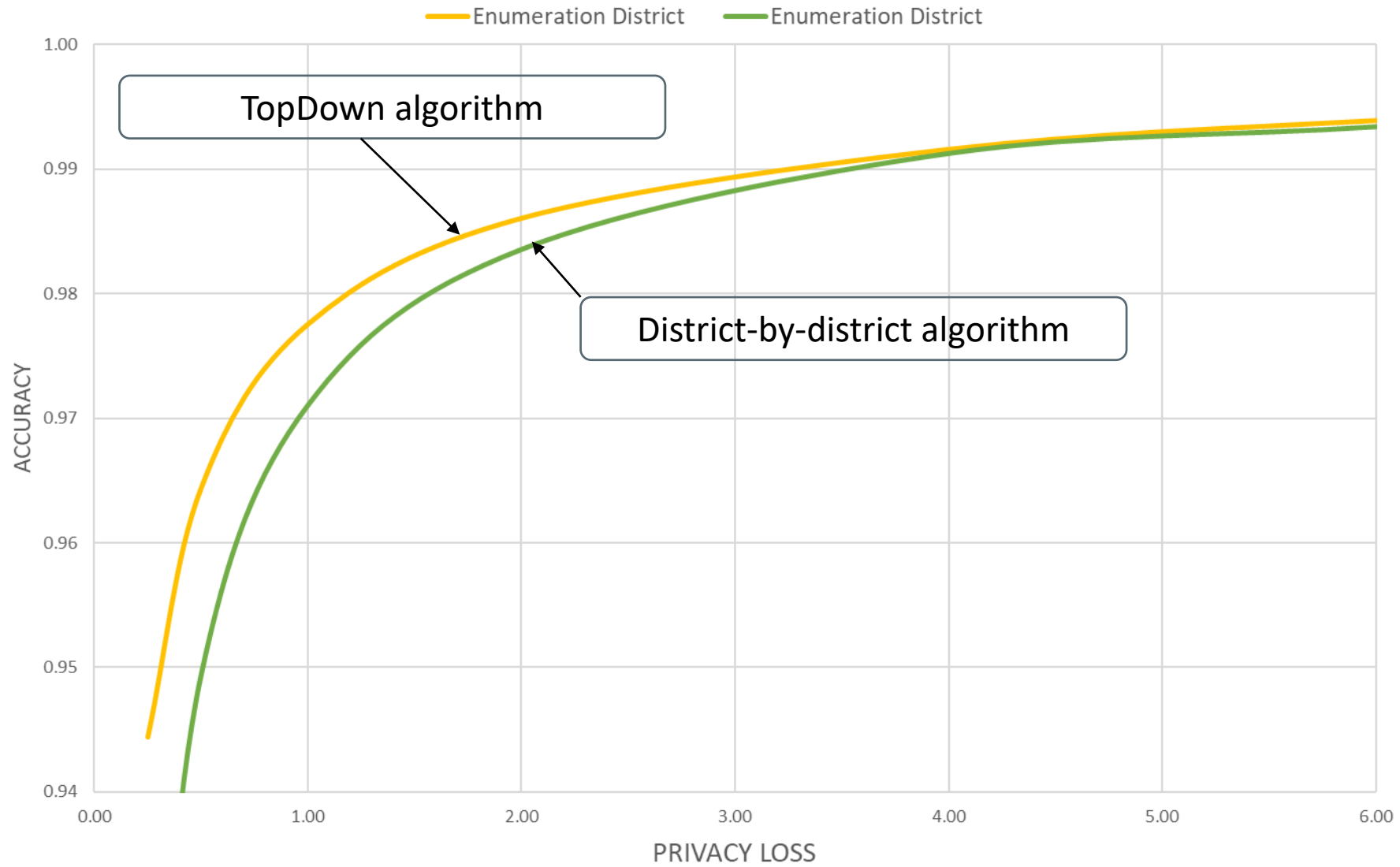
DISTRICT-BY-DISTRICT DIFFERENTIAL PRIVACY ALGORITHMS (1940 CENSUS DATA)



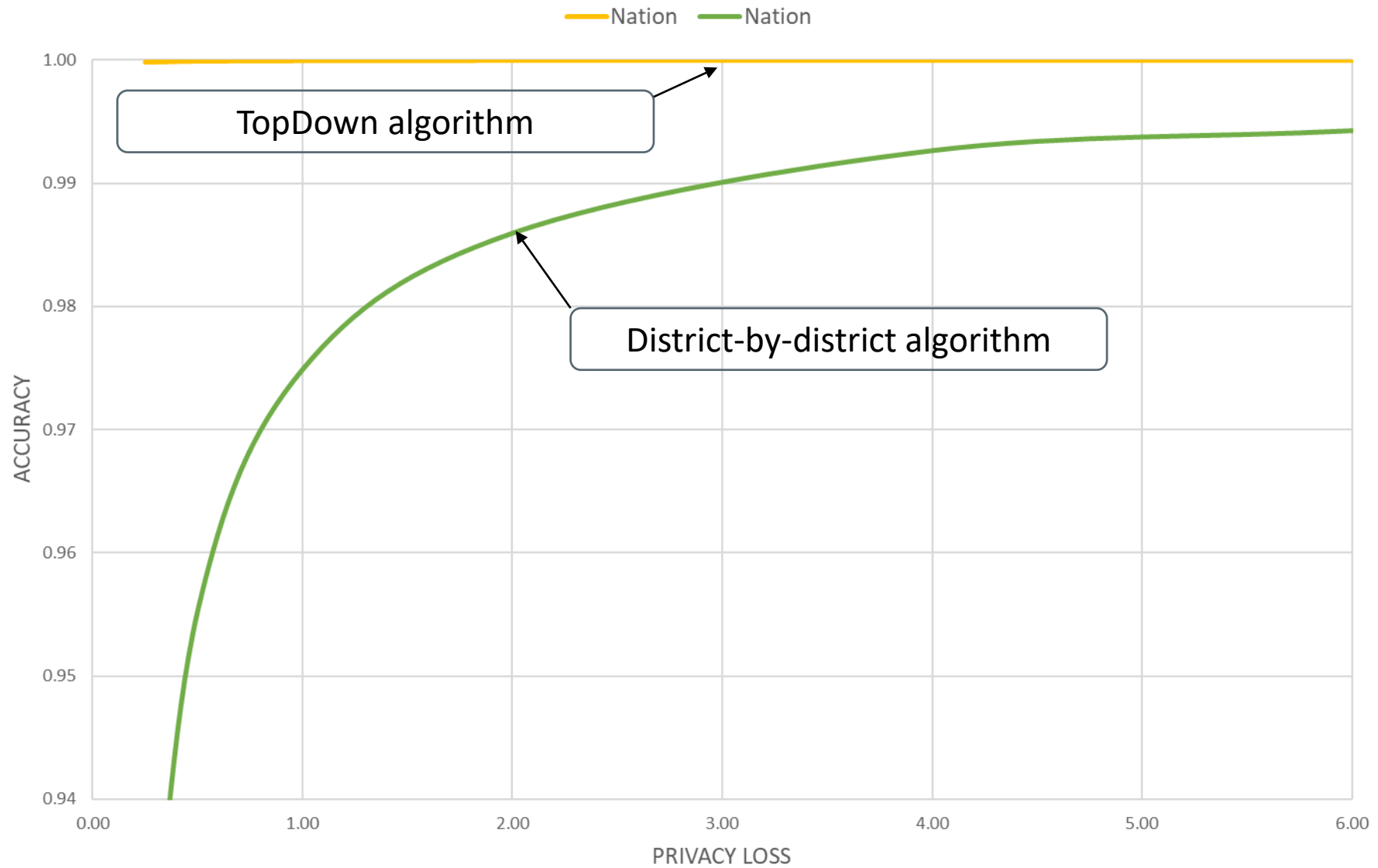
TOPDOWN DIFFERENTIAL PRIVACY ALGORITHMS (1940 CENSUS DATA)



COMPARISON OF DISTRICT RESULTS BY ALGORITHM (1940 CENSUS DATA)



COMPARISON OF NATIONAL RESULTS BY ALGORITHM (1940 CENSUS DATA)



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?#>
- April 15, 2019 Code base and documentation for the 2018 End-to-End Census Test (E2E) version of the 2020 Disclosure Avoidance System <https://github.com/uscensusbureau/census2020-das-e2e>
- June 6, 2019 Blog explaining how to use the code base with the 1940 Census public data from IPUMS https://www.census.gov/newsroom/blogs/research-matters/2019/06/disclosure_avoidance.html
- June 11, 2019 Keynote address “The U.S. Census Bureau Tries to Be a Good Data Steward for the 21st Century” ICML 2019 [abstract](#), [video](#)
- June 29-31, 2019 Joint Statistical Meetings [Census Bureau electronic press kit](#) (See talks by Abowd, Ashmead, Garfinkel, Leclerc, Sexton, and others)

Thank you.

John.Maron.Abowd@census.gov