Trying to Be a Good Data Steward in the 21st Century

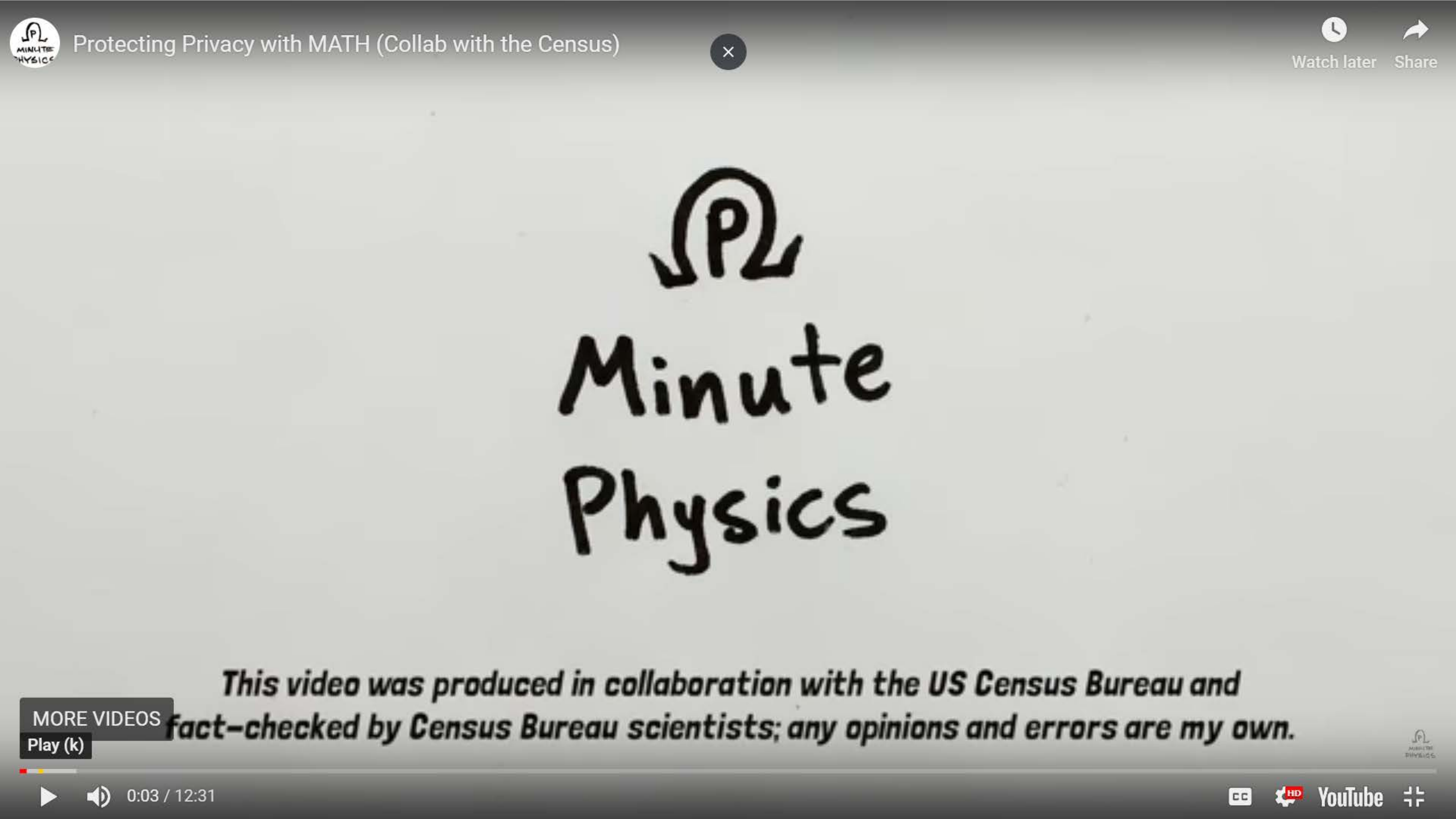
John M. Abowd
Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau

Joint Statistical Meetings 2020
SPAIG Lunchtime Speaker: Thursday August 6, 2020 12:00PM

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

Shape your future START HERE >

United States® Census 2020



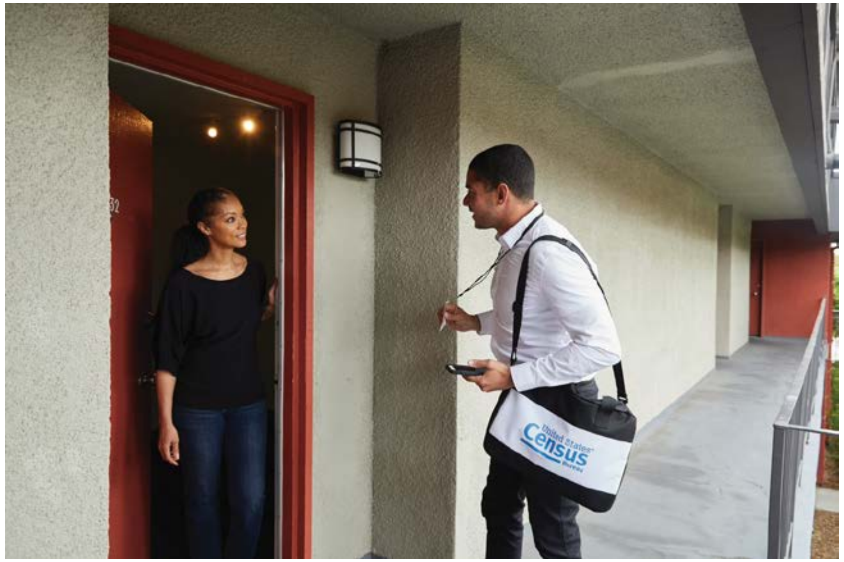
Protecting Privacy with MATH (Collab with the Census)

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.

The Commitment to Data Stewardship

Data stewardship is central to the Census Bureau's mission to produce high-quality statistics about the people and economy of the United States.

Our commitment to protect the privacy of our respondents and the confidentiality of their data is both a legal obligation and a core component of our institutional culture.



Upholding the Promise: Today and Tomorrow

A 21st Century data steward cannot merely consider privacy threats that exist today.

A good steward must ensure that disclosure avoidance methods are also sufficient to protect against the threats of tomorrow!

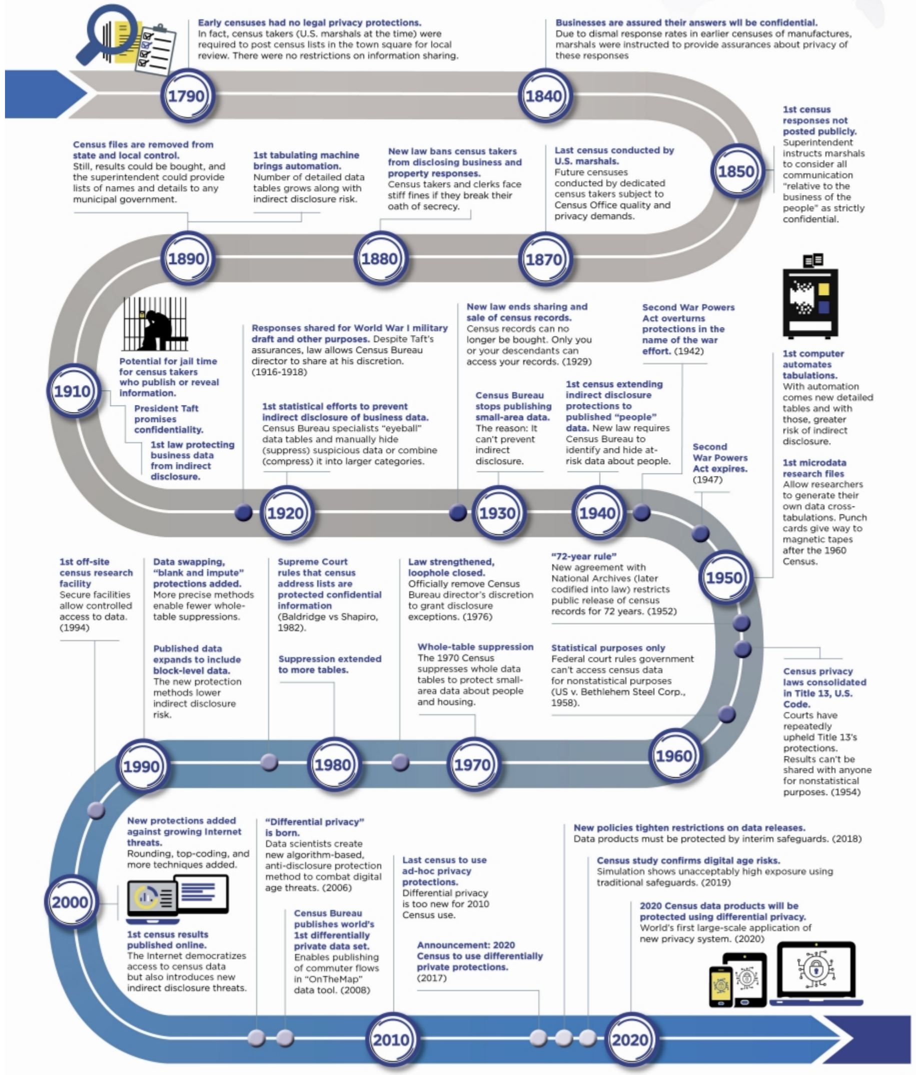
A HISTORY OF CENSUS PRIVACY PROTECTIONS

Today's law is clear: The Census Bureau must keep responses completely confidential. It cannot release identifiable information about an individual, household or business to anyone, including other government or law enforcement agencies.

It wasn't always that way. Public attitudes on privacy have changed since the first census in 1790. Early laws and policies focused on preventing direct disclosure of personal information. Later, laws and policies addressed the growing threat of indirect disclosure—the risk that someone might be able to figure out the identity of a person or business just by analyzing the statistics we publish.

Twenty-first century privacy threats—faster and more powerful computers, new data science, and exponential growth in personal data available online—demand new safeguards to protect against indirect disclosure.

See how the laws and protections have changed from 1790 to the 2020 Census—the first census to use advanced disclosure protections based on the new data science known as "differential privacy."



1790

Early censuses had no legal privacy protections. In fact, census takers (U.S. marshals at the time) were required to post census lists in the town square for local review. There were no restrictions on information sharing.

1840

Businesses are assured their answers wil be confidential. Due to dismal response rates in earlier censuses of manufactures, marshals were instructed to provide assurances about privacy of these responses

1850

1st census responses not posted publicly. Superintendent instructs marshals to consider all communication "relative to the business of the people" as strictly confidential.

1870

Last census conducted by U.S. marshals. Future censuses conducted by dedicated census takers subject to Census Office quality and privacy demands.

1880

New law bans census takers from disclosing business and property responses. Census takers and clerks face stiff fines if they break their oath of secrecy.

1890

Census files are removed from state and local control. Still, results could be bought, and the superintendent could provide lists of names and details to any municipal government.

1st tabulating machine brings automation. Number of detailed data tables grows along with indirect disclosure risk.

1910

Potential for jail time for census takers who publish or reveal information.

President Taft promises confidentiality.

1st law protecting business data from indirect disclosure.

1920

Responses shared for World War I military draft and other purposes. Despite Taft's assurances, law allows Census Bureau director to share at his discretion. (1916-1918)

1st statistical efforts to prevent indirect disclosure of business data. Census Bureau specialists "eyeball" data tables and manually hide (suppress) suspicious data or combine (compress) it into larger categories.

1930

New law ends sharing and sale of census records. Census records can no longer be bought. Only you or your descendants can access your records. (1929)

Census Bureau stops publishing small-area data. The reason: It can't prevent indirect disclosure.

1940

1st census extending indirect disclosure protections to published "people" data. New law requires Census Bureau to identify and hide at-risk data about people.

Second War Powers Act overturns protections in the name of the war effort. (1942)

Second War Powers Act expires. (1947)

1950

1st computer automates tabulations. With automation comes new detailed tables and with those, greater risk of indirect disclosure.

1st microdata research files Allow researchers to generate their own data cross-tabulations. Punch cards give way to magnetic tapes after the 1960 Census.

"72-year rule" New agreement with National Archives (later codified into law) restricts public release of census records for 72 years. (1952)

Census privacy laws consolidated in Title 13, U.S. Code. Courts have repeatedly upheld Title 13's protections. Results can't be shared with anyone for nonstatistical purposes. (1954)

Statistical purposes only Federal court rules government can't access census data for nonstatistical purposes (US v. Bethlehem Steel Corp., 1958).

1960

1970

Whole-table suppression The 1970 Census suppresses whole data tables to protect small-area data about people and housing.

Law strengthened, loophole closed. Officially remove Census Bureau director's discretion to grant disclosure exceptions. (1976)

1980

Suppression extended to more tables.

Supreme Court rules that census address lists are protected confidential information (Baldridge vs Shapiro, 1982).

1990

Published data expands to include block-level data. The new protection methods lower indirect disclosure risk.

Data swapping, "blank and impute" protections added. More precise methods enable fewer whole-table suppressions.

1st off-site census research facility Secure facilities allow controlled access to data. (1994)

2000

1st census results published online. The Internet democratizes access to census data but also introduces new indirect disclosure threats.

New protections added against growing Internet threats. Rounding, top-coding, and more techniques added.

"Differential privacy" is born. Data scientists create new algorithm-based, anti-disclosure protection method to combat digital age threats. (2006)

Census Bureau publishes world's 1st differentially private data set. Enables publishing of commuter flows in "OnTheMap" data tool. (2008)

2010

Last census to use ad-hoc privacy protections. Differential privacy is too new for 2010 Census use.

Announcement: 2020 Census to use differentially private protections. (2017)

New policies tighten restrictions on data releases. Data products must be protected by interim safeguards. (2018)

Census study confirms digital age risks. Simulation shows unacceptably high exposure using traditional safeguards. (2019)

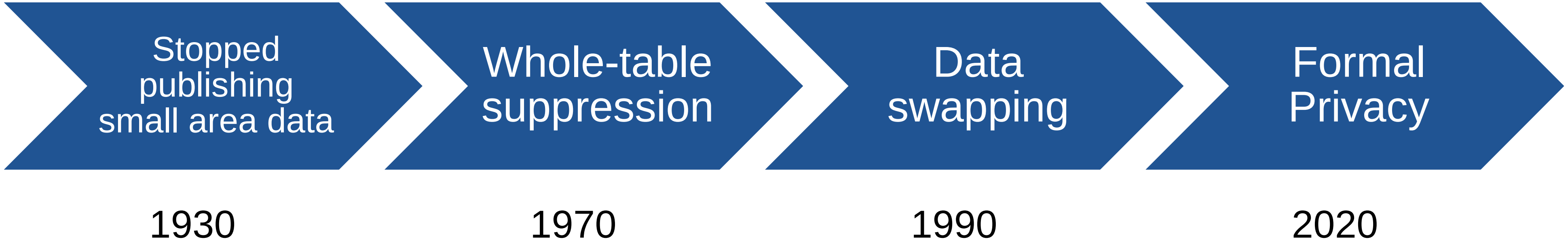
2020

2020 Census data products will be protected using differential privacy. World's first large-scale application of new privacy system. (2020)

Highlight Summary of Privacy Protections Over Time

Throughout its history, the Census Bureau has been at the forefront of the design and implementation of statistical methods to safeguard respondent data.

Over the decades, as we have increased the number and detail of the data products we release, so too have we improved the statistical techniques we use to protect those data.



Stopped publishing small area data

1930

Whole-table suppression

1970

Data swapping

1990

Formal Privacy

2020

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

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.

Privacy protection is an economic problem.

Not a technical problem in computer science or statistics.

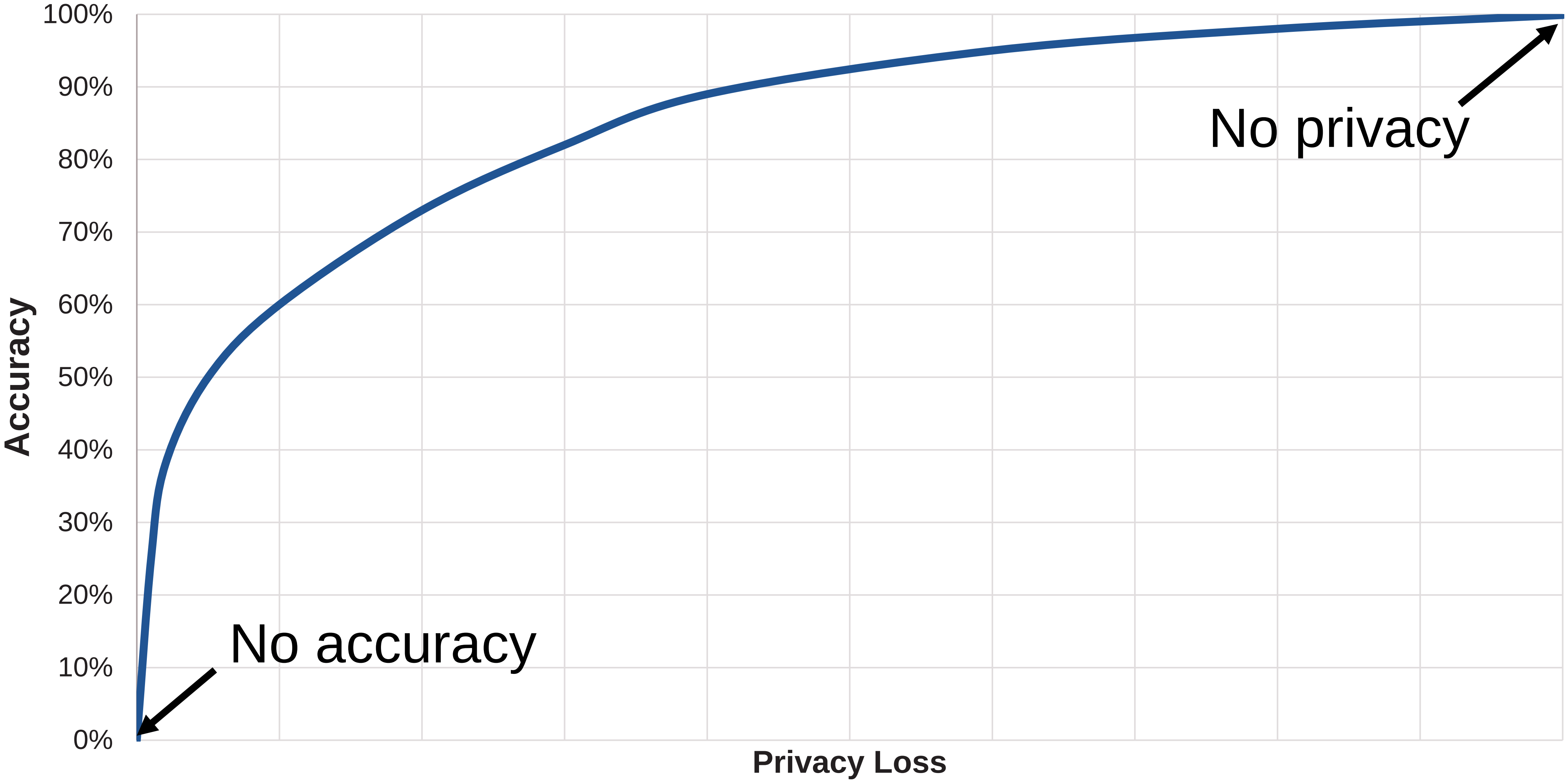
Allocation of a scarce resource (data in the confidential database) between competing uses:

information products

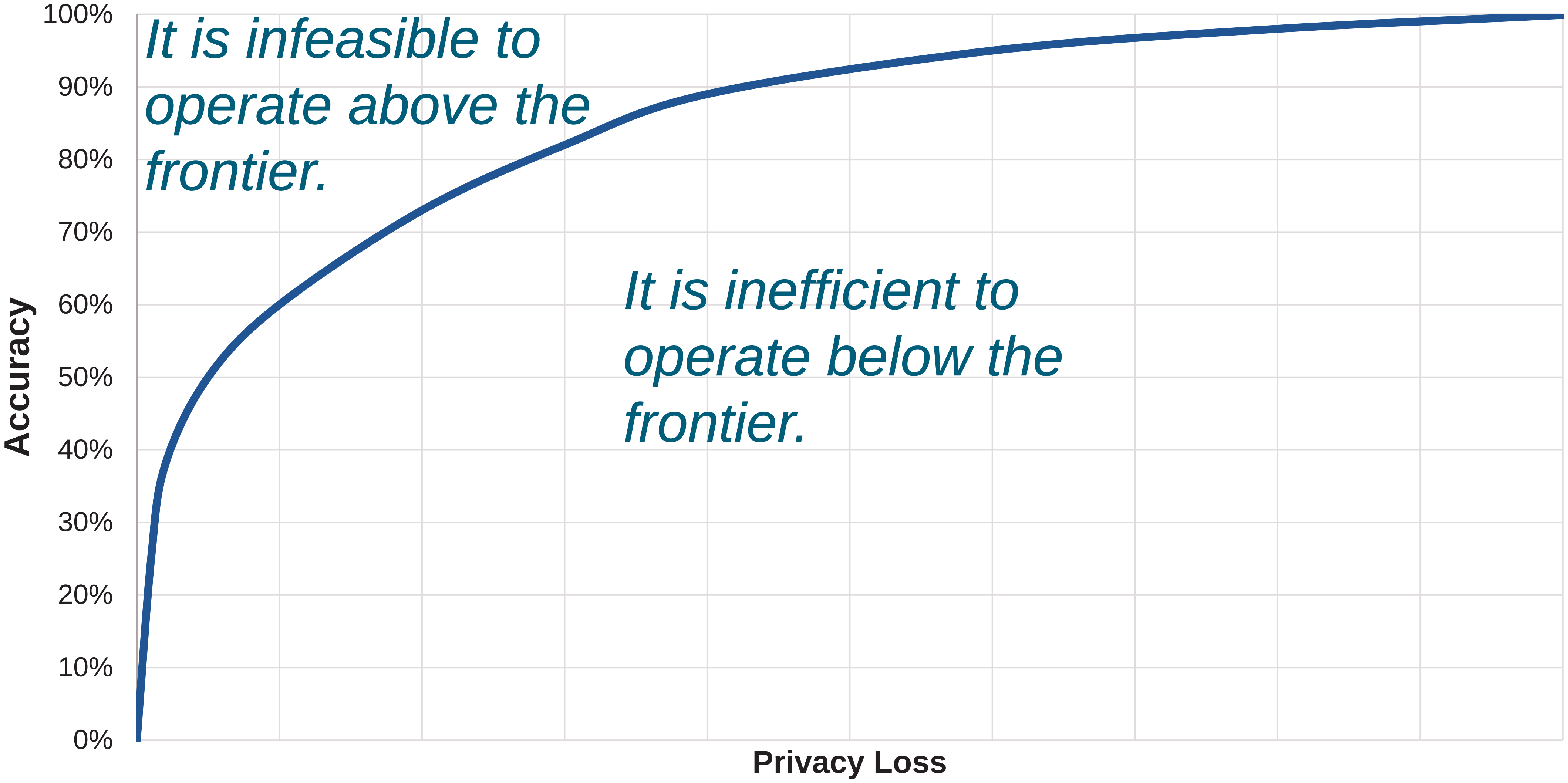
and

privacy protection.

Fundamental Tradeoff between Accuracy and Privacy Loss



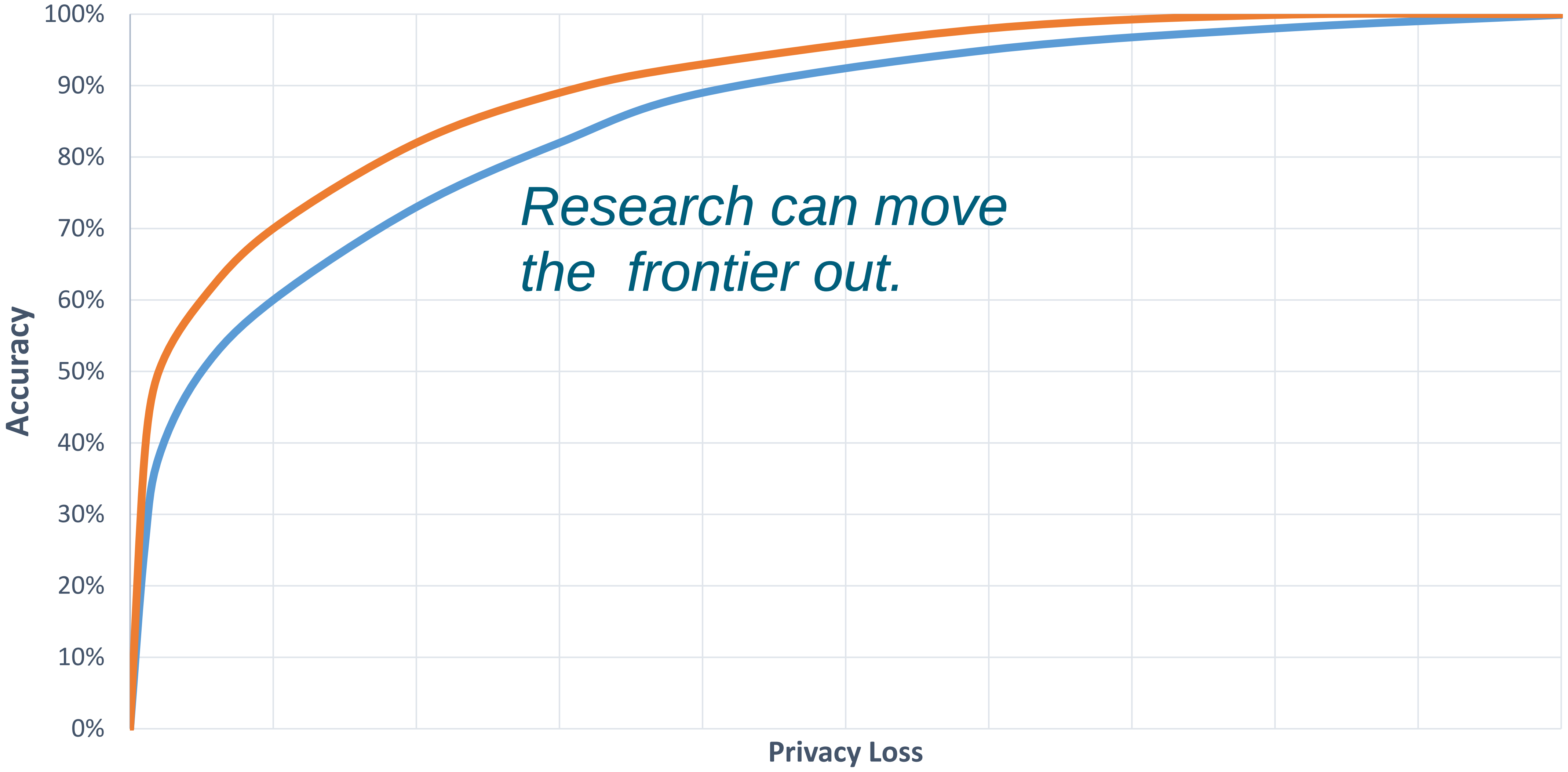
Fundamental Tradeoff between Accuracy and Privacy Loss



It is infeasible to operate above the frontier.

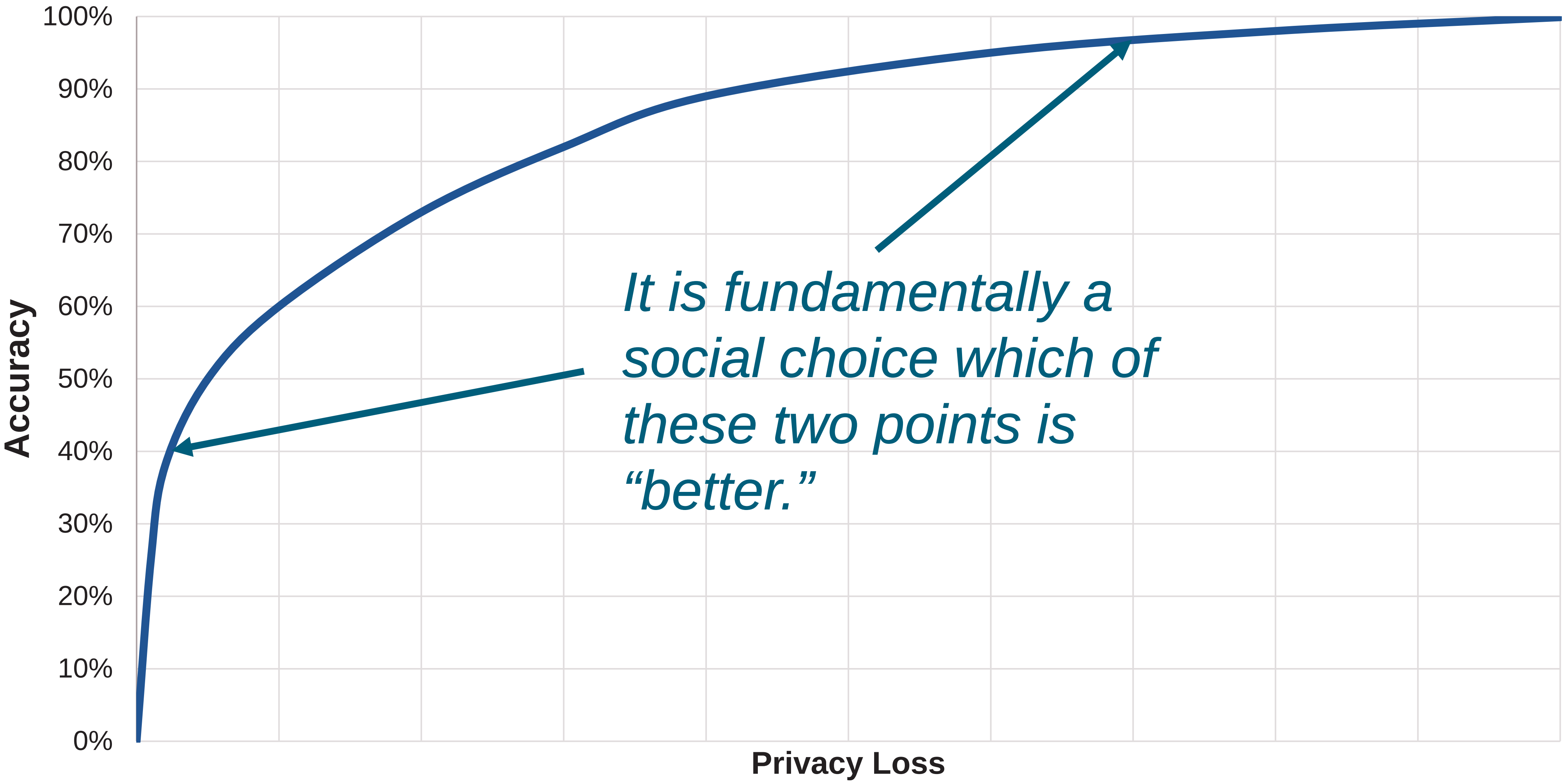
It is inefficient to operate below the frontier.

Fundamental Tradeoff between Accuracy and Privacy Loss



Research can move the frontier out.

Fundamental Tradeoff between Accuracy and Privacy Loss



It is fundamentally a social choice which of these two points is “better.”

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.

Accurate, but for Whom?

- The 2020 Census Disclosure Avoidance System operates under interpretable formal privacy guarantees, given privacy-loss budgets
- Accuracy properties depend upon the output metric
- Different data users will have a particular sets of analyses they wish to be accurate
- Tuning accuracy to a given analysis can reduce accuracy for other analyses
- Policy makers must consider reasonable overall accuracy metrics for privacy tradeoff
- Knowing how overall metrics correspond to user results could help optimize DAS

Privacy Protection out of the Shadows

- Certain privacy practices for previous censuses depended upon obfuscation
- 2020 Census Disclosure Avoidance System is the most transparent view into Census Bureau privacy practices ever
- Constructive feedback has and will continue to benefit Census Bureau and external partners
- The Census Bureau appreciates and is eager to assess feedback from external partners

Algorithms Matter

Implemented TopDown Algorithmic Summary

Take differentially private measurements at every level of the hierarchy

At each level of TopDown post-process:

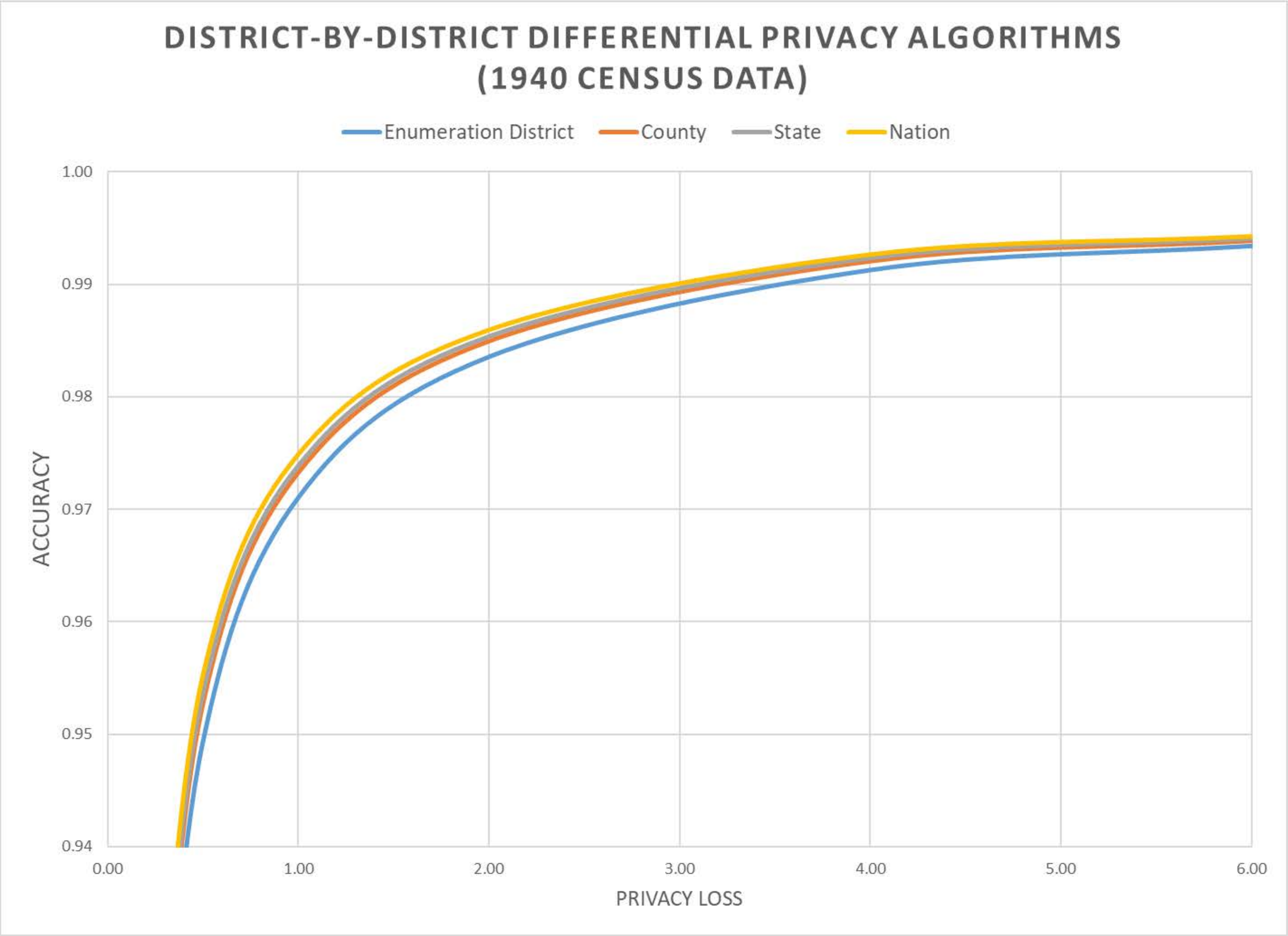
- Solve an optimization problem to produce non-negative tables
- Solve a second optimization problem to round to integer tables
- Generate micro-data from the integer tables

Naïve Method: BottomUp or Block-by-Block

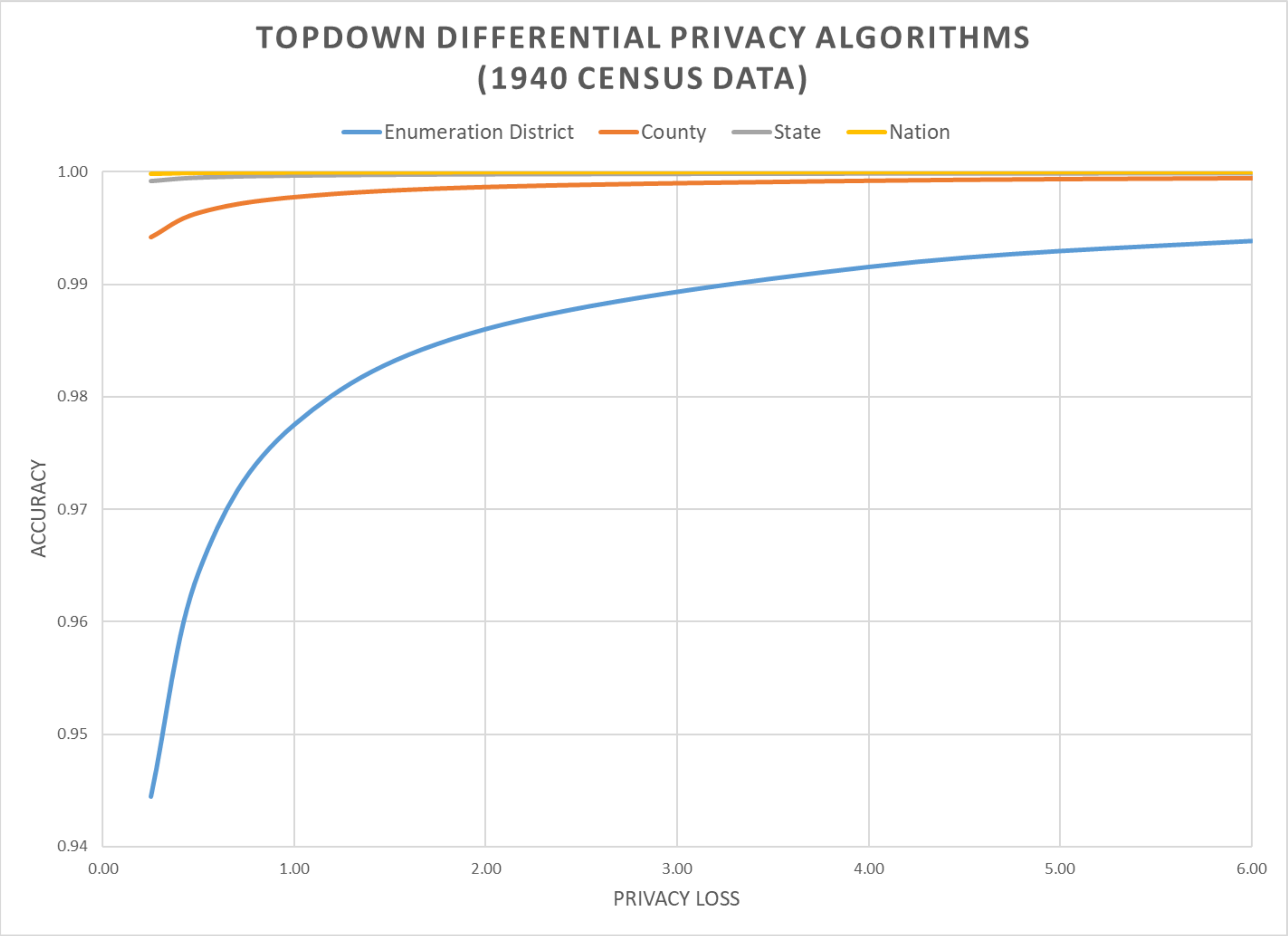
Apply differential privacy algorithms to the most detailed level of geography

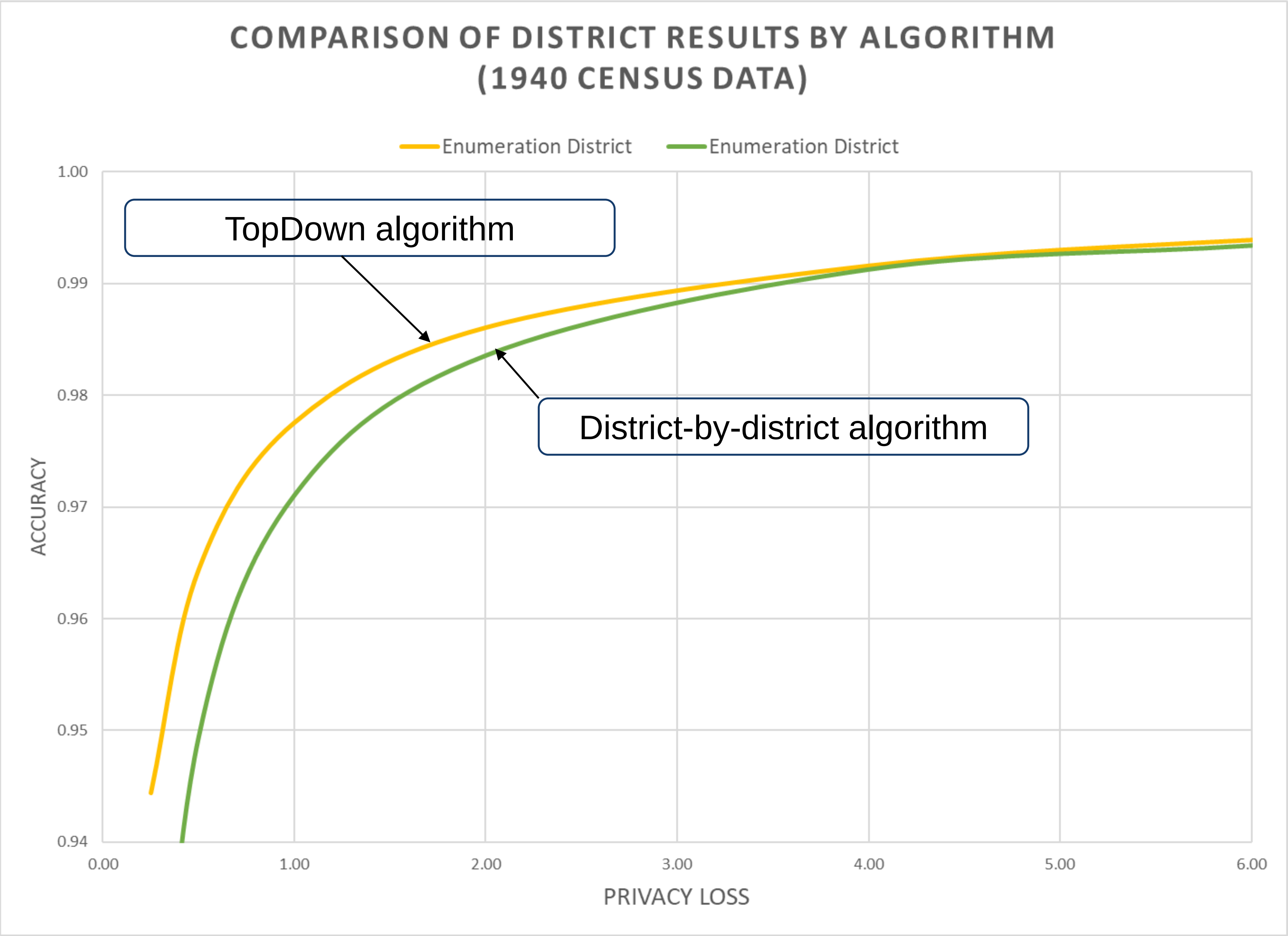
Build all geographic aggregates from those components via post-processing

This is similar to the local differential privacy implementations in the Chrome browser, iOS, Windows 10, Facebook Social Science One, Uber, ...



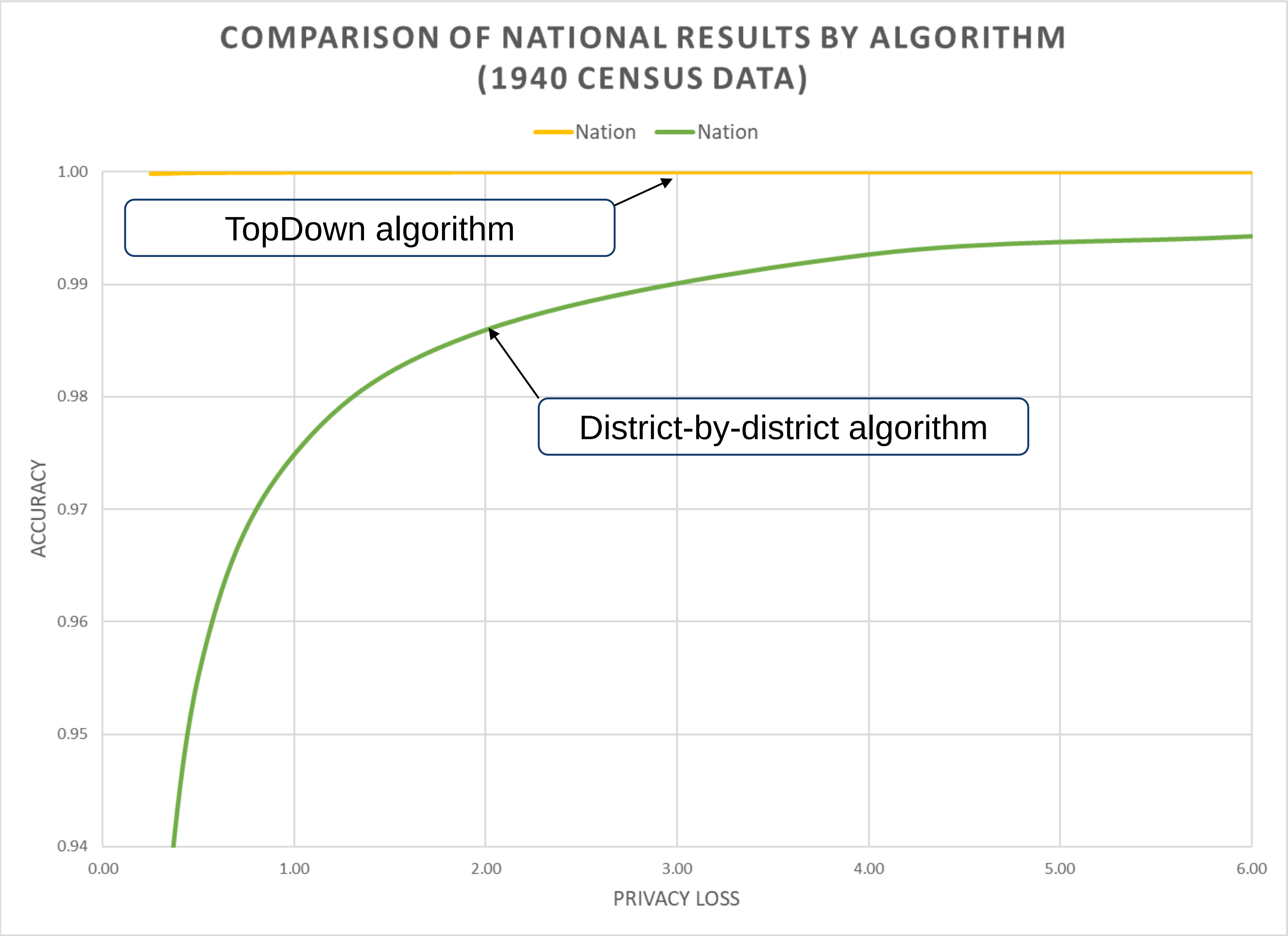
DISTRICT-BY-DISTRICT DIFFERENTIAL PRIVACY ALGORITHMS (1940 CENSUS DATA)


TOPDOWN DIFFERENTIAL PRIVACY ALGORITHMS
(1940 CENSUS DATA)
Enumeration District
County
State
Nation
1.00
0.99
0.98
0.97
0.96
0.95
0.94
ACCURACY
0.00
1.00
2.00
3.00
4.00
5.00
6.00
PRIVACY LOSS



TopDown algorithm

District-by-district algorithm



TopDown algorithm

District-by-district algorithm

But this 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 current 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)

October 25, 2019 Harvard Data Science Review Symposium: https://hdsr.mitpress.mit.edu/pub/h7kdirec/release/5

November 14, 2019 Privatar In:Confidence (video) https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html?wvideo=uugsds68pj

December 11-12, 2019 CNSTAT Workshop: https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518

General information: 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

Bi-weekly newsletters: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/newsletters.html

Thank you

John.Maron.Abowd@census.gov