# The U.S. Census Bureau Adopts Differential Privacy

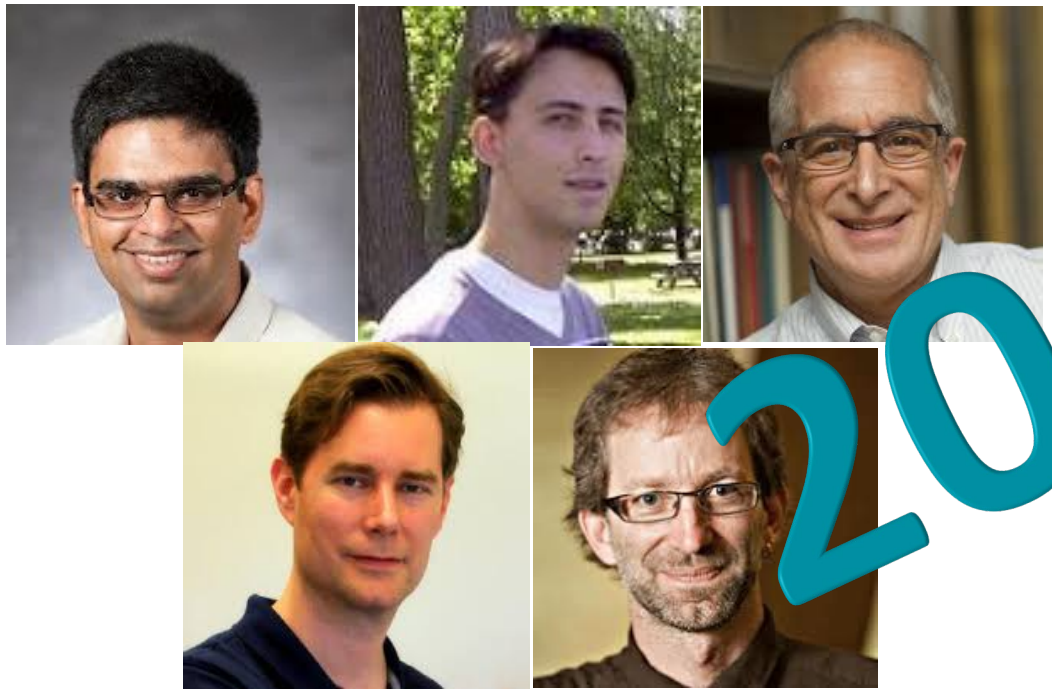
John M. Abowd
Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau
24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining
London, United Kingdom
August 23, 2018

United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

# Acknowledgments and Disclaimer

- The opinions expressed in this talk are the my own and not necessarily those of the U.S. Census Bureau
- The application to the Census Bureau's 2020 publication system incorporates work by Daniel Kifer (Scientific Lead), Simson Garfinkel (Senior Scientist for Confidentiality and Data Access), Tamara Adams, Robert Ashmead, Michael Bentley, Stephen Clark, Aref Dajani, Jason Devine, Nathan Goldschlag, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Gerome Miklau, Brett Moran, Edward Porter, Anne Ross, and Lars Vilhuber [link to the September 2018 Census Scientific Advisory Committee presentation]
- Parts of this talk were supported by the National Science Foundation, the Sloan Foundation, and the Census Bureau (before and after my appointment started)

# A Brief History of Differential Privacy at the U.S. Census Bureau



2008

## Privacy: Theory meets Practice on the Map

Ashwin Machanavajjhala [†1], Daniel Kifer [†2], John Abowd [#3], Johannes Gehrke [†4], Lars Vilhuber [#5]

[†]*Department of Computer Science, Cornell University, U.S.A.*
[#]*Department of Labor Economics, Cornell University, U.S.A.*

[1]mvnak@cs.cornell.edu [2]dkifer@cs.cornell.edu [3]john.abowd@cornell.edu
[4]johannes@cs.cornell.edu [5]lars.vilhuber@cornell.edu

**…t— In … paper, we propose the first formal privacy ana… a da… onymization process known as the synthetic data … nique becoming popular in the statistics …mmuni… …plication for this work is a mapping …gram t… …hows … …mmuting patterns of the population …e Un… …States. T… …urce data for this application were …ted … …e U.S. … …s Bureau, but due to privacy con- …s, t… …sed directly by the mapping program. I… …, we … synthetic data that statistically mimic the o… …l data while providing privacy guarantees. We use these …tic data as a surrogate for the original data.**

**…e find that while some existing definitions of privacy are inapplicable to our target application, others are too conservative and render the synthetic data useless since they guard against privacy breaches that are very unlikely. Moreover, the data in our target application is sparse, and none of the existing solutions are tailored to anonymize sparse data. In this paper, we propose solutions to address the above issues.**

### I. Introduction

In this paper, we study a real-world application of a privacy preserving technology known as synthetic data generation. We present the first formal privacy guarantees (to the best of our knowledge) for this application. This paper chronicles the challenges we faced in this endeavour. The target application is based on data developed by the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD). By combining various Census datasets it is possible to construct a table Commute_Patterns with schema *(id,origin_block, destination_block)* where each row represents a worker. The attribute *id* is a random number serving as a key for the table, *origin_block* is the census block in which the worker lives, and *destination_block* is where the worker works. An origin block $o$ corresponds to a destination block $d$ if there is a tuple with *origin_block* $o$ and *destination_block* $d$. The goal is to plot points on a map that represent commuting patterns for the U.S. to such a mapping application. An anonymized version must be used instead.

The algorithm used to anonymize the data for the above mapping application is known as the synthetic data generation [1], which is becoming popular in the statistical disclosure limitationg community. The main idea behind synthetic data generation is to build a statistical model from the data and then to sample points from the model. These sampled points form the synthetic data, which is then released instead of the original data. While much research has focused on deriving the variance and confidence intervals for various estimators from synthetic data [2], [3], there has been little research on deriving formal guarantees of privacy for such an approach (an exception is [4]).

Much recent research has focused on deriving formal criteria for privacy. These include general notions of statistical closeness [5], variants of the notions of *k-anonymity* [6] and *ℓ-diversity* [7], [8], [9], [10], [11], $(\rho_1, \rho_2)$-*privacy* [12], and variants of *differential privacy* [13], [14]. However, we found that apart from the differential privacy criterion [13], none of the other privacy conditions applied to our scenario.

Picking an off-the-shelf synthetic data generation algorithm and tuning it to satisfy the differential privacy criterion was unsatisfactory for the following reasons. First, in order to satisfy the differential privacy criterion, the generated synthetic data contained little or no information about the original data. We show that this is because differential privacy guards against breaches of privacy that are very unlikely.

Next, no deterministic algorithm can satisfy differential privacy. Randomized algorithms can (albeit with a very small probability) return anonymized datasets that are totally unrepresentative of the input. This is a problem, especially, when we want to publish a single or only a few, anonymous versions of

https://onthemap.ces.census.gov

# OnTheMap

LEHD Home | Help and Documentation | Reload | Text-Only

Start | Base Map | Selection

Save | Load | Feedback | Previous Extent | Hide Tabs

**Welcome to OnTheMap!**

Start an analysis by using one of the tools below (Search, Import Geography, or Load .OTM file). Hover over the Help icons located throughout the application to see Help tips for using specific functionality. Sections in the control panel can be collapsed or opened by clicking the section title.

2015 Data Now Available (09/25/2017)

**Search**

Search

Search All Names

**Import Geography**

Import from KML
Import from SHP
Import from GPS

**Load .OTM File**

*Click the "Load" button b... to load a .OTM file.*

Load

Seattle
San Francisco
San Jose
Los Angeles
San Diego
Phoenix
El Paso
Dallas
Austin
Houston
Memphis
Chicago
Detroit
Boston
New York
Philadelphia
Baltimore
Charlotte
Jacksonville

1000 km
500 mi

-122.89877, 22.47507

2008: First production implementation of differential privacy worldwide

https://onthemap.ces.census.gov

# OnTheMap

LEHD Home | Help and Documentation | Reload | Text-Only

Start | Base Map | Selection | Results

## Distance/Direction Analysis
*Work to Home*

**Display Settings**

Labor Market Segment Filter: All Workers

Year: 2015

**Map Controls**

Color Key

Thermal Overlay ☑

Point Overlay ☑

Selection Outline ☑

Identify | Zoom to Selection

Clear Overlays | Animate Overlays

**Report/Map Outputs**

Detailed Report

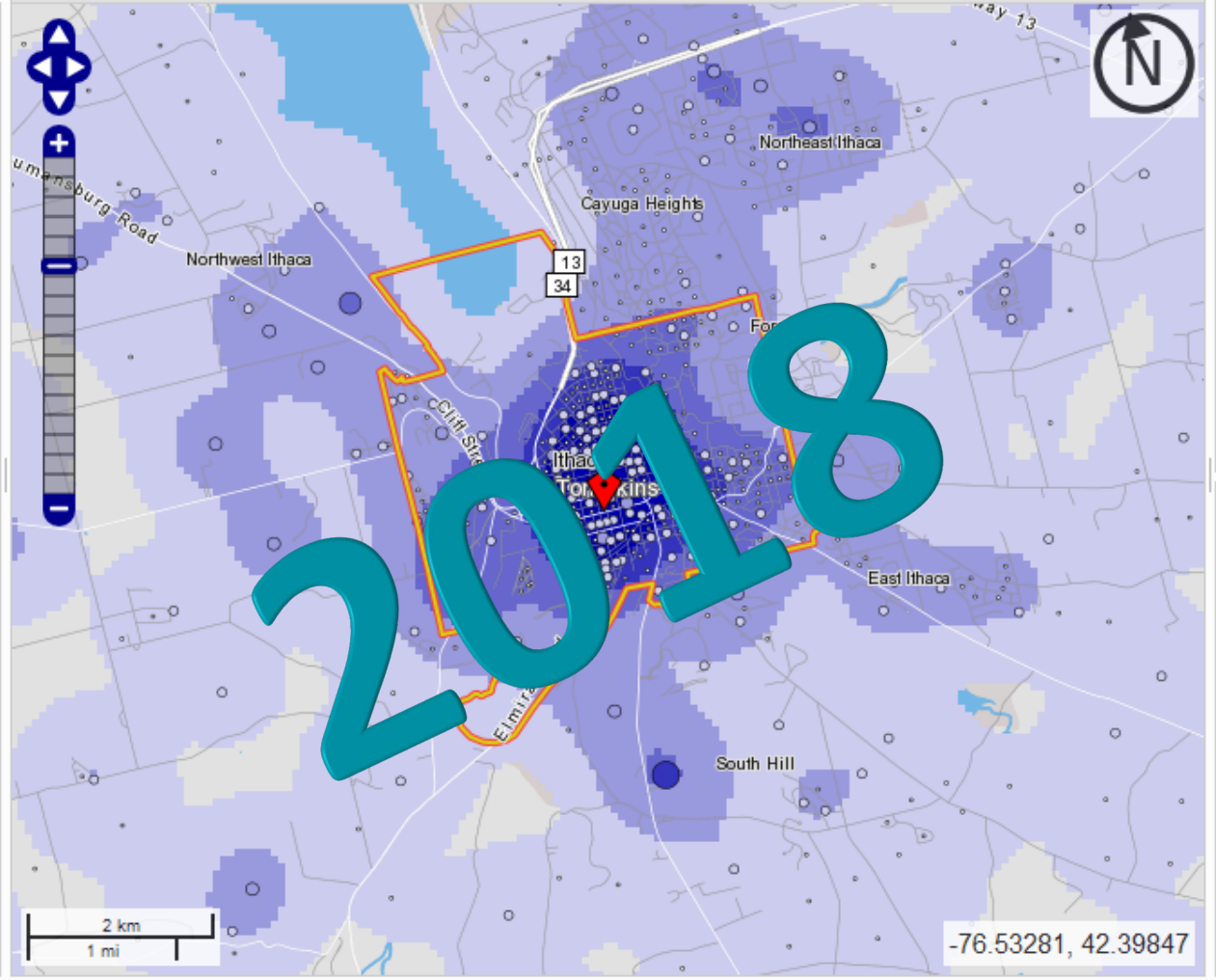
Export Geography

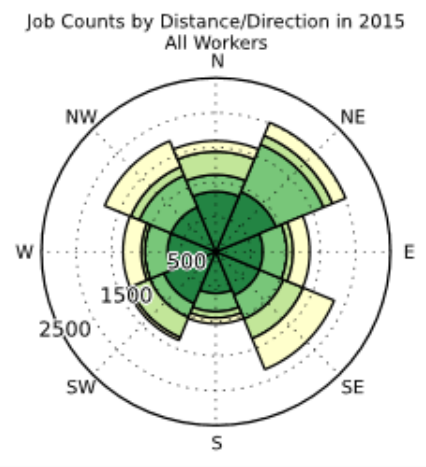
Print Chart/Map

**Legends**

**Change Settings**

Save | Load | Feedback | Previous Extent | Hide Tabs | Hide Chart/Report




View as Radar Chart

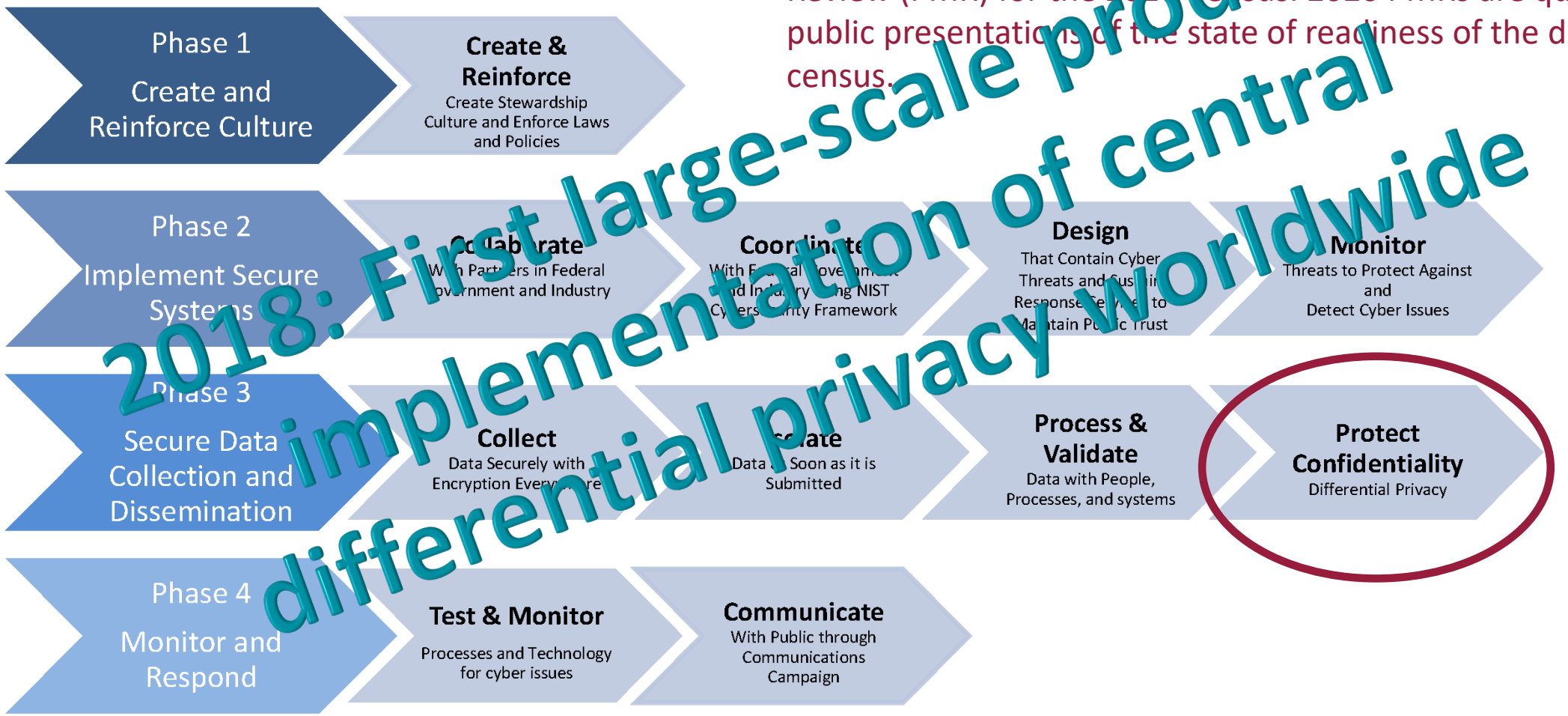
**Jobs by Distance - Work Census Block to Home Census Block**

| | 2015 Count | 2015 Share |
|---|---|---|
| Total Primary Jobs | 12,260 | 100.0% |
| Less than 10 miles | 5,949 | 48.5% |
| 10 to 24 miles | 2,987 | 24.4% |
| 25 to 50 miles | 1,451 | 11.8% |
| Greater than 50 miles | 1,873 | 15.3% |

# Census Data Stewardship

## Our Overall Approach to Maintain Public Trust

This slide is from the August 3, 2018 Program Management Review (PMR) for the 2020 Census. 2020 PMRs are quarterly public presentations of the state of readiness of the decennial census.




2018: First large-scale production implementation of central differential privacy worldwide

# Database Reconstruction

# 2003:
# Database Reconstruction

## ABSTRACT

We examine the tradeoff between privacy and usability of statistical databases. We model a statistical database by an $n$-bit string $d_1, .., d_n$, with a query being a subset $q \subseteq [n]$ to be answered by $\sum_{i \in q} d_i$. Our main result is a polynomial reconstruction algorithm of data from noisy (perturbed) subset sums. Applying this reconstruction algorithm to statistical databases we show that in order to achieve privacy one has to add perturbation of magnitude $\Omega(\sqrt{n})$. That is, smaller perturbation always results in a strong violation of privacy. We show that this result is tight by exemplifying access algorithms for statistical databases that preserve privacy while adding perturbation of magnitude $\tilde{O}(\sqrt{n})$.

For time-$\mathcal{T}$ bounded adversaries we demonstrate a privacy-preserving access algorithm whose perturbation magnitude is $\approx \sqrt{\mathcal{T}}$.

United States Census Bureau

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

## Revealing Information while Preserving Privacy

Irit Dinur  Kobbi Nissim*
NEC Research Institute
4 Independence Way
Princeton, NJ 08540
{iritd,kobbi}@research.nj.nec.com

**ABSTRACT**

We examine the tradeoff between privacy and usability of statistical databases. We model a statistical database by an $n$-bit string $d_1, .., d_n$, with a query being a subset $q \subseteq [n]$ to

the hospital is obliged to keep the privacy of its patients, i.e. leak no medical information that could be related to a specific patient. The hospital needs an access mechanism to the database that allows certain 'statistical' queries to be answered, as long as they do not violate the privacy of any single patient.

*Work partly done when the author was at DIMACS, Rutgers University, and while visiting Microsoft Research Silicon Valley Lab.


*PODS* 2003, June 9-12, 2003, San Diego, CA.


One simple tempting solution is to remove from the database all 'identifying' attributes such as the patients' names and social security numbers. However, this solution is not enough to protect patient privacy since there usually exist other

oaches taken into three main categories: (i) query restriction, (ii) data perturbation, and (iii) output perturbation. We give a brief review of these approaches below, and refer the reader to [2] for a detailed survey of the methods and their weaknesses.

**Query Restriction.** In the query restriction approach, queries are required to obey a special structure, supposedly to prevent the querying adversary from gaining too much information about specific database entries. The limit of this approach is that it allows for a relatively small number of queries.

A related idea is of query auditing [7], i.e. a log of the queries is kept, and every new query is checked for possible compromise, allowing/disallowing the query accordingly.

[1]A patient's gender, approximate age, approximate weight, ethnicity, and marital status – may already suffice for a complete identification of most patients in a database of a thousand patients. The situation is much worse if a relatively 'rare' attribute of some patient is known. For example, a patient having Cystic Fibrosis (frequency $\approx 1/3000$) may be uniquely identified within about a million patients.

# The Database Reconstruction Theorem

- Powerful result from Dinur and Nissim (2003) [link]
- *Too many statistics published too accurately from a confidential database exposes the entire database with near certainty*
- How accurately is "too accurately"?
  - Cumulative noise must be of the order $\sqrt{N}$
- At this conference, I don't need to explain what this means

# The 2010 Census of Population and Housing

# 2010 Census of Population: Summary

| | |
|---|---|
| Total population | 308,745,538 |
| Household population | 300,758,215 |
| Group quarters population | 7,987,323 |
| Households | 116,716,292 |

# 2010 Census: High-level Database Schema

| Variables | Distinct values |
|---|---|
| Habitable blocks | 10,620,683 |
| Habitable tracts | 73,768 |
| Sex | 2 |
| Age | 115 |
| Race/Ethnicity (OMB Categories) | 126 |
| Race/Ethnicity (SF2 Categories) | 600 |
| Relationship to person 1 | 17 |
| National histogram cells (OMB Ethnicity) | 492,660 |

# 2010 Census: Published Statistics

| Publication | Released counts (including zeros) |
|---|---|
| PL94-171 Redistricting | 2,771,998,263 |
| Balance of Summary File 1 | 2,806,899,669 |
| Summary File 2 | 2,093,683,376 |
| Public-use micro sample | 30,874,554 |
| Lower bound on published statistics | 7,703,455,862 |
| Statistics/person | 25 |

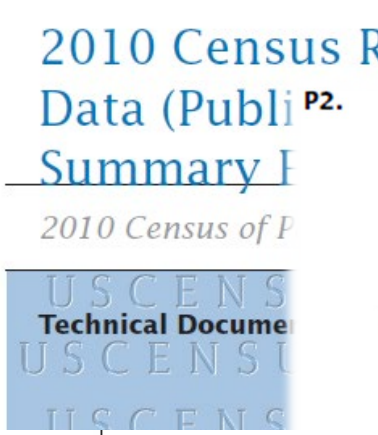
The database reconstruction theorem is the death knell for traditional data publication systems from confidential sources.

# Internal Experiments Using the 2010 Census

- Confirm that the confidential micro-data from the confidential hundred percent detail file can be reconstructed quite accurately from PL94 + balance of SF1
- While there is a vulnerability, the risk of re-identification is small
- Experiments are at the person level, not household
- Experiments have led to the declaration that reconstruction of Title 13-sensitive data is an issue, no longer a risk
- Strong motivation for the adoption of differential privacy for the 2018 End-to-End Census Test and 2020 Census

# Reconstruction Equatio

Collect more than 5 billion statistics from official 2010 Census tables.




2010 Census Summary File 1

Issued September 2012

P1. TOTAL POPULATION [1]

P2. HISPANIC OR LATINO, AND NOT HISPANIC OR LATINO BY RACE [73]
*Universe: Total population*
Total: P00
Hispanic or Latino P00
Not Hispanic or Latino: P00
Population of one race: P00
White alone P00

P4. HISPANIC OR LATINO, AND NOT HISPANIC OR LATINO BY RACE FOR THE POPULATION 18 YEARS AND OVER [73]
*Universe: Total population 18 years and over*
Total:
Hispanic or Latino
Not Hispanic or Latino:
Population of one race:
White alone
Black or African American alone

From the sample space at the block and tract level (2 x 115 x 2 x 63 = 28,980), write the linear equations for each sample statistic, including zeros.

United States Census Bureau

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

| | | | |
|---|---|---|---|
| 18 and 19 years | P012A031 | 07 | 9 |
| 20 years | P012A032 | 07 | 9 |
| 21 years | P012A033 | 07 | 9 |
| 22 to 24 years | P012A034 | 07 | 9 |
| 25 to 29 years | P012A035 | 07 | 9 |
| 30 to 34 years | P012A036 | 07 | 9 |

# Properties of the Solution

- This is the Dinur-Nissim reconstruction equation system for exact statistics
- Can’t be overdetermined (known to come from a real person table)
- Usually underdetermined: potentially many solutions
- But, all solutions share some exact images
  - For example, block and voting age variables are the same in every solution
- Full details will be released this fall

# Formal Privacy

# 2006: Differential Privacy

**Abstract.** We continue a line of research initiated in [10, 11] on privacy-preserving statistical databases. Consider a trusted server that holds a database of sensitive information. Given a query function $f$ mapping databases to reals, the so-called *true answer* is the result of applying $f$ to the database. To protect privacy, the true answer is perturbed by the addition of random noise generated according to a carefully chosen distribution, and this response, the true answer plus noise, is returned to the user.

Previous work focused on the case of noisy sums, in which $f = \sum_i g(x_i)$, where $x_i$ denotes the $i$th row of the database and $g$ maps database rows to $[0, 1]$. We extend the study to general functions $f$, proving that privacy can be preserved by calibrating the standard deviation of the noise according to the *sensitivity* of the function $f$. Roughly speaking, this is the amount that any single argument to $f$ can change its output. The new analysis shows that for several particular applications substantially less noise is needed than was previously understood to be the case.

The first step is a very clean characterization of privacy in terms of indistinguishability of transcripts. Additionally, we obtain separation results showing the increased value of interactive sanitization mechanisms over non-interactive.

United States Census Bureau

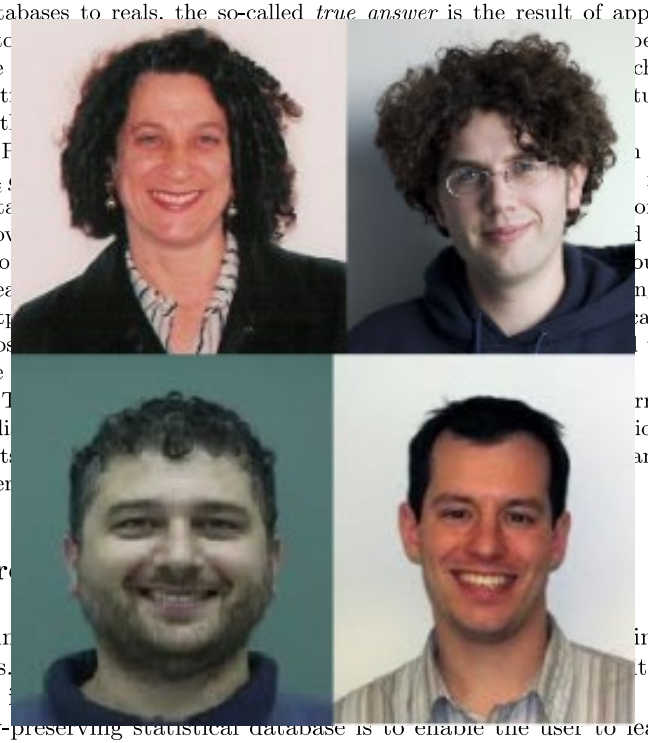
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

## Calibrating Noise to Sensitivity in Private Data Analysis

Cynthia Dwork[1], Frank McSherry[1], Kobbi Nissim[2], and Adam Smith[3]*

[1] Microsoft Research, Silicon Valley. {dwork,mcsherry}@microsoft.com
[2] Ben-Gurion University. kobbi@cs.bgu.ac.il
[3] Weizmann Institute of Science. adam.smith@weizmann.ac.il

**Abstract.** We continue a line of research initiated in [10, 11] on privacy-preserving statistical databases. Consider a trusted server that holds a database of sensitive information. Given a query function $f$ mapping databases to reals, the so-called *true answer* is the result of applying



1 Intr

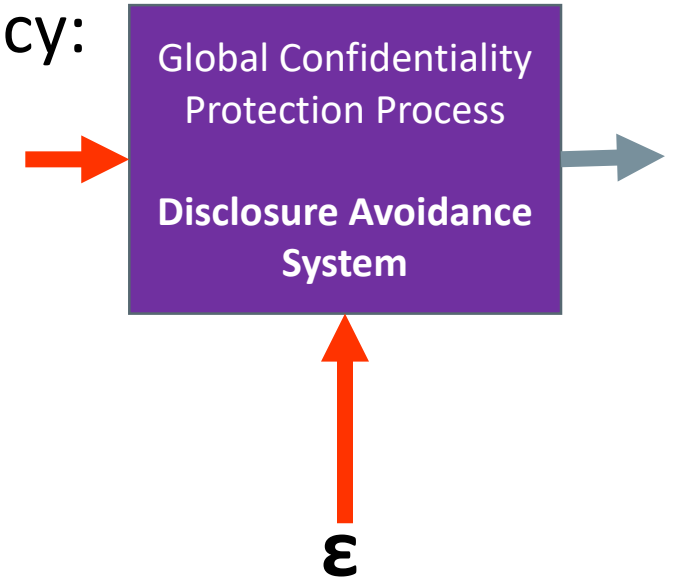
We conti in *statistica* databases. tively, if th database the goal o a privacy-preserving statistical database is to enable the user to learn proper ties of the population as a whole while protecting the privacy of the individua contributors.

We assume the database is held by a trusted server. On input a query functio $f$ mapping databases to reals, the so-called *true answer* is the result of applying $f$ to the database. To protect privacy, the true answer is perturbed by the additio

* Supported by the Louis L. and Anita M. Perlman Postdoctoral Fellowship.

# The Disclosure Avoidance System Relies on Injecting Noise with Formal Privacy Rules

- Advantages of noise injection with differential privacy:
  - Privacy operations are *closed under composition*
  - Privacy guarantees are *robust to post-processing*
  - Privacy guarantees are *future-proof*
  - Privacy guarantees are *provable and tunable*
  - Privacy guarantees are public and explainable
  - Protects against database reconstruction attacks

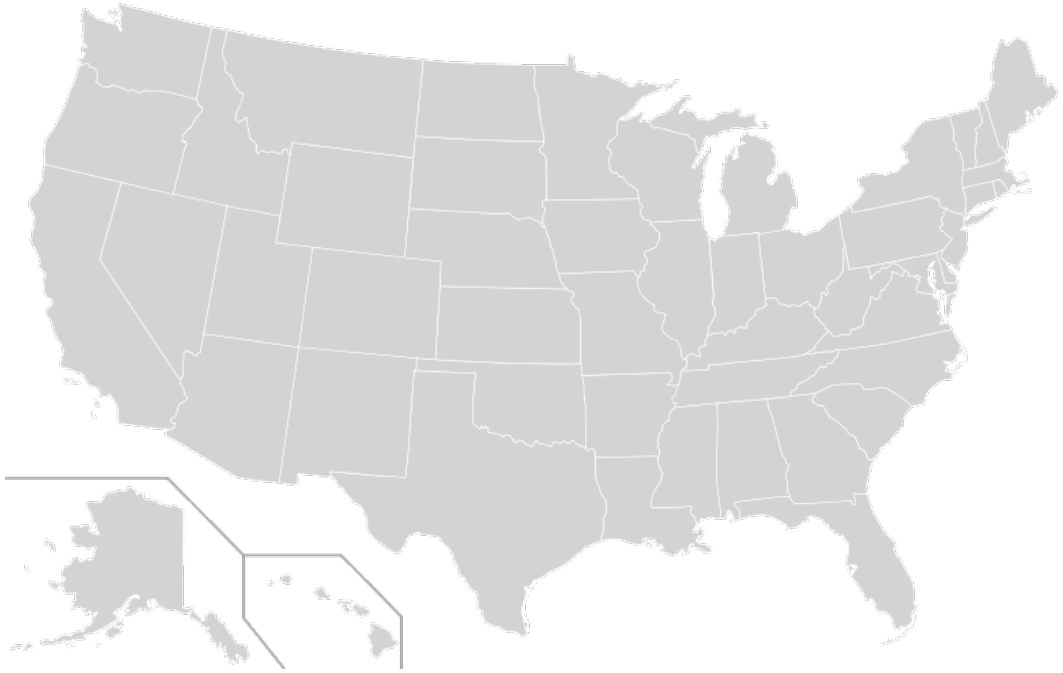



- Disadvantages:
  - Entire country must be processed at once for best accuracy
  - Every use of the private data must be tallied in the *privacy-loss budget*

# Additional Technical Details

- Central differential privacy implementation with a controlled total privacy-loss budget
- Relevant definition is bounded $\varepsilon$-differential privacy (total population of the United States is public)
- Semantic privacy guarantee is $[-2\varepsilon, 2\varepsilon]$ by properties of bounded differential privacy
- Other semantic guarantees, as they affect implemented invariants will be published later this year
- All algorithms, code, and parameter values will be released with the test files for the 2018 End-to-End Census Test

# 2020 Census of Population and Households



United States Census 2020

United States™ Census Bureau

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

# The Top-Down Algorithm

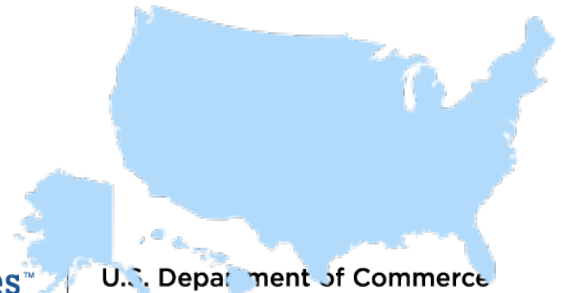
National table of US population

2 x 126 x 17 x 115

Spend $\varepsilon_1$ privacy-loss budget

National table with all 500,000 cells filled, structural zeros imposed with accuracy allowed by $\varepsilon_1$
2 x 126 x 17 x 115

Sex: Male / Female
Race + Hispanic: 126 possible values
Relationship to Householder: 17
Age: 0-114



Reconstruct individual micro-data without geography

330,000,000 records

United States Census Bureau

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

# State-level

State-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1

Spend $\varepsilon_2$ privacy-loss budget

Target state-level tables required for best accuracy for PL-94 and SF-1

Construct best-fitting individual micro-data with state geography

330,000,000 records now including state identifiers


United States Census Bureau

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

# County-level

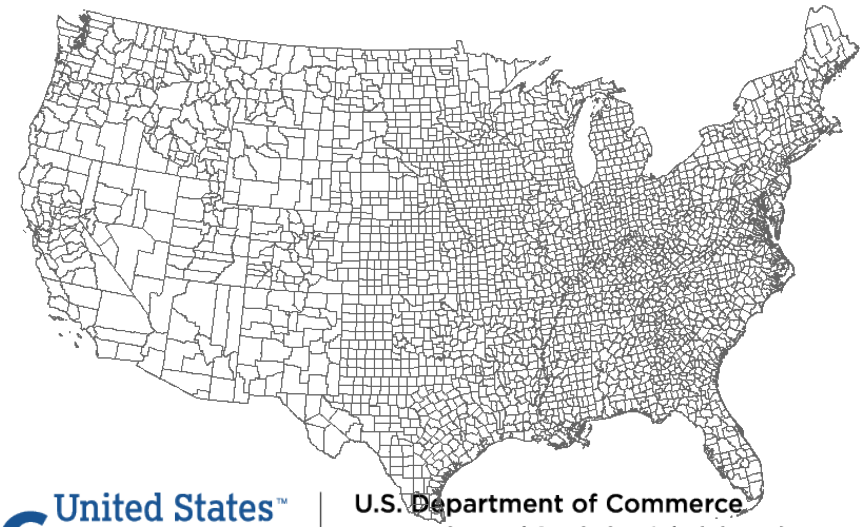
County-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1

Spend $\varepsilon_3$ privacy-loss budget

Target county-level tables required for best accuracy for PL-94 and SF-1

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

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



United States Census Bureau

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

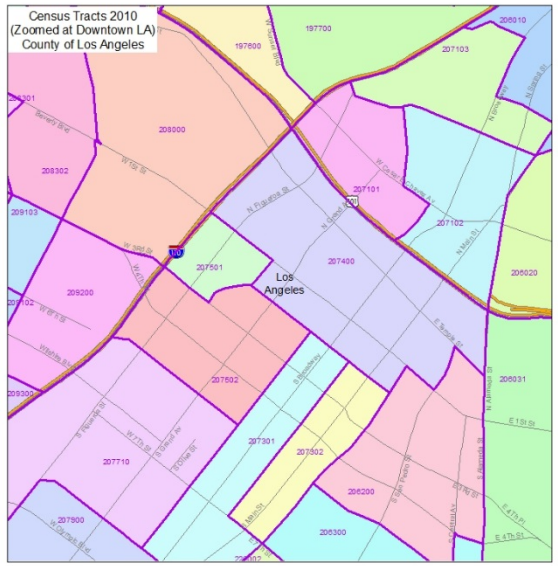
identifiers

# Census tract-level

Tract-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1

Spend $\varepsilon_4$ privacy-loss budget

Target tract-level tables required for best accuracy for PL-94 and SF-1




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

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

'ce

Unit Census Bureau

Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

tract identifiers

# Block-level

Block-level tables for only certain queries; structural zeros imposed; dimensions chosen to produce best accuracy for PL-94 and SF-1

Spend $\varepsilon_5$ privacy-loss budget

Block tract-level tables required for best accuracy for PL-94 and SF-1

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

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


United States Census Bureau

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

# 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, SF-1

# Tabulation micro-data

- How accurate are the tabulation micro-data?


Disclosure Avoidance Certificate

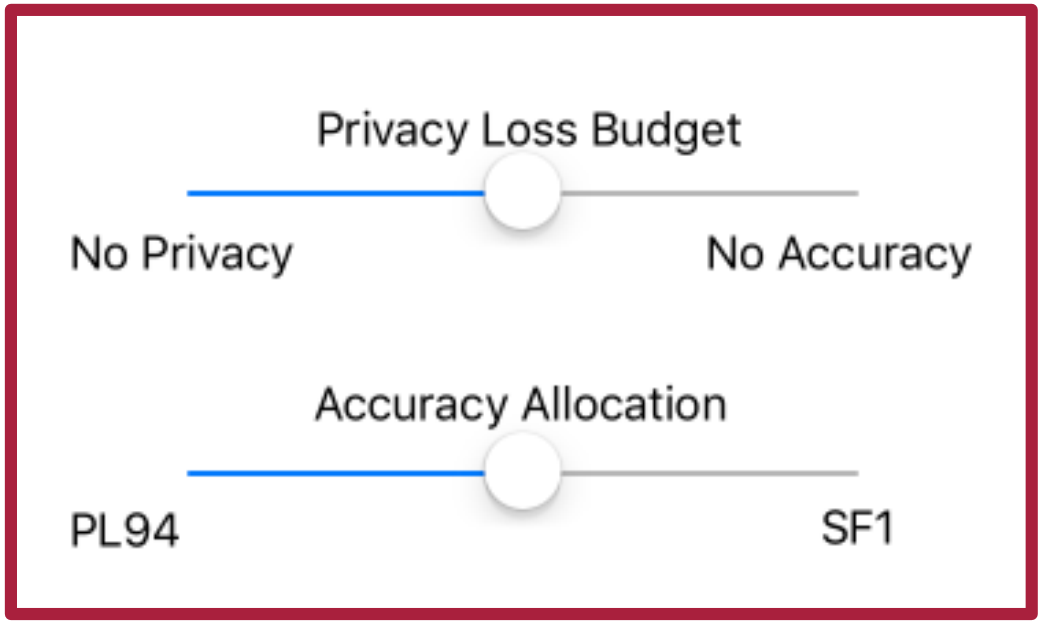
- Certifies that the disclosure avoidance system passed all tests
- Reports the accuracy of the micro-data used for tabulation
- Requires $\boldsymbol{\varepsilon}_A$

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

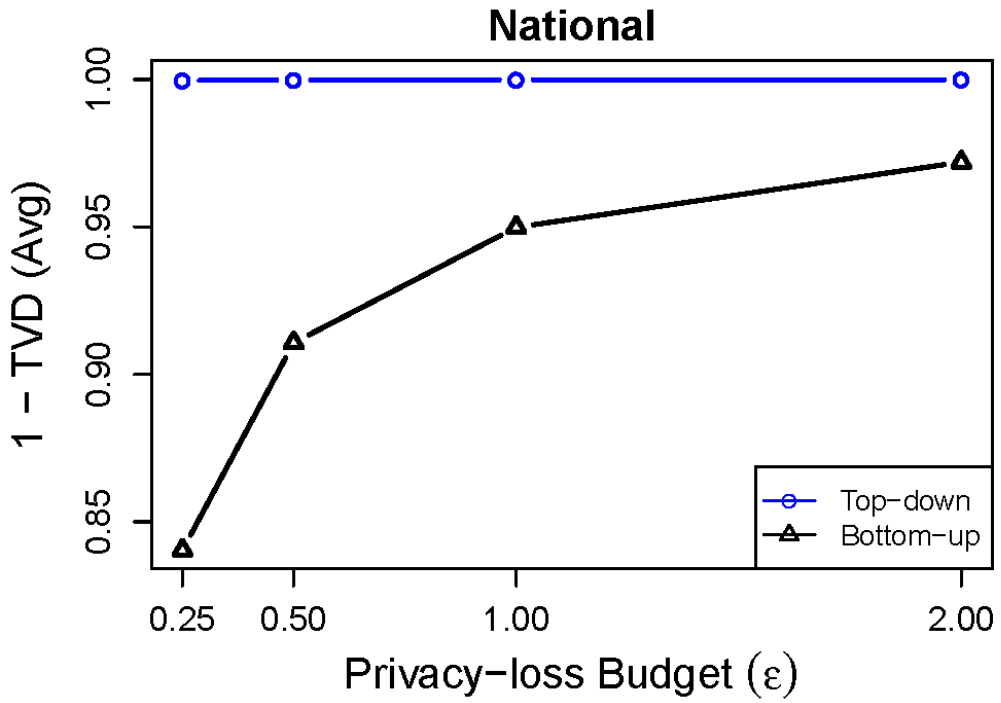
Micro-data used for tabulating
PL-94, SF-1

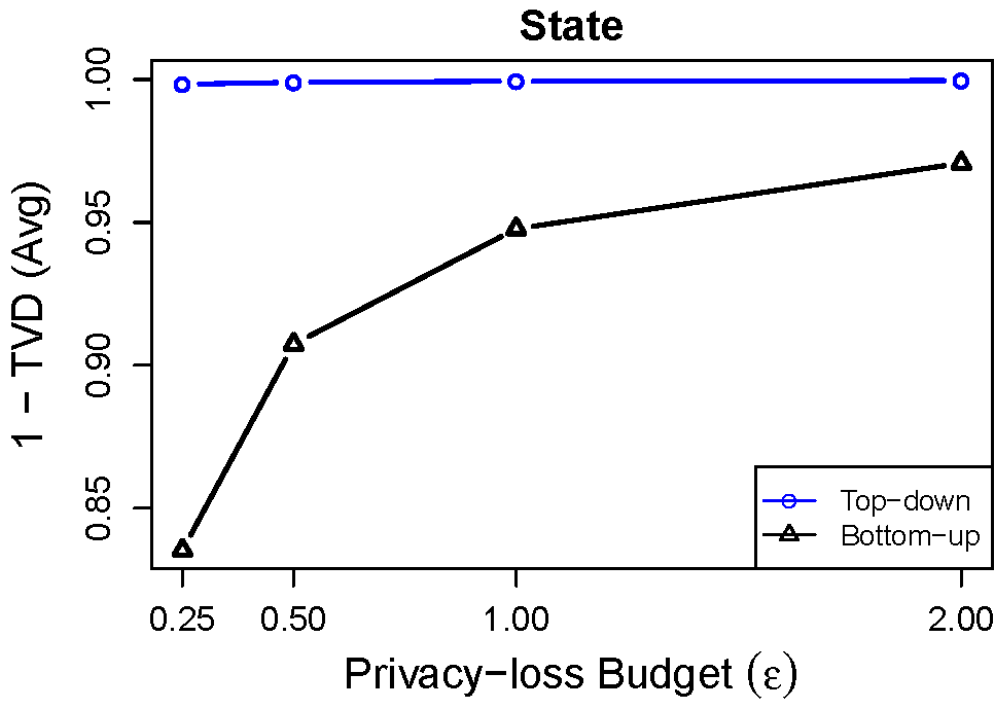
# Operational Decisions

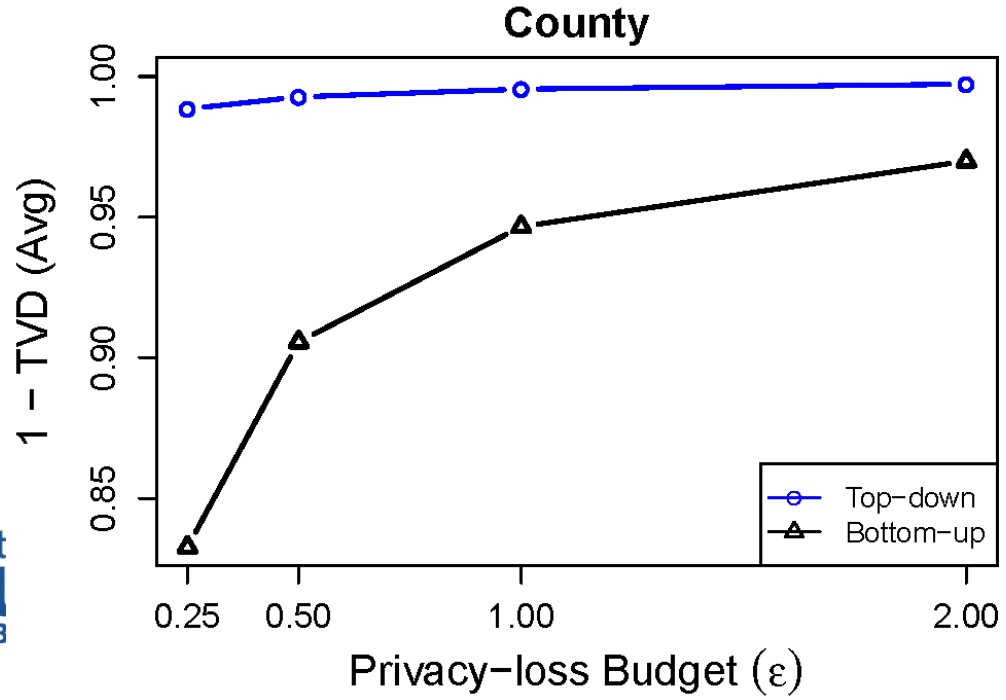


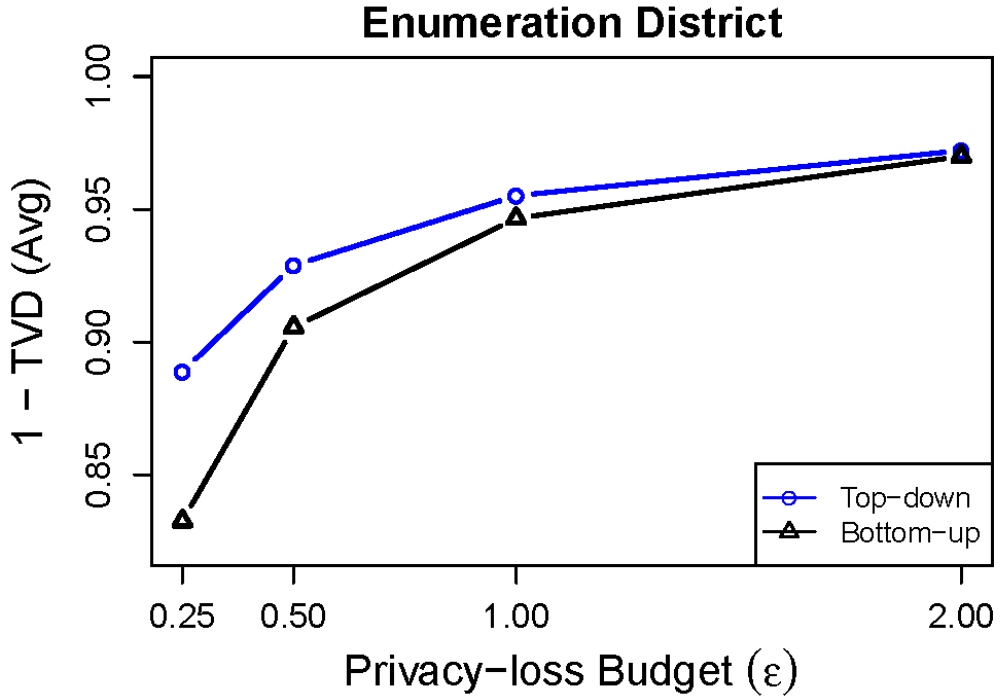

- Set total privacy-loss budget: $\boldsymbol{\varepsilon}$
  - Ensure that $\boldsymbol{\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5 + \varepsilon_A = \varepsilon}$
- Within each stage, allocate privacy-loss budget between:
  - PL-94
  - Parts of SF-1 not in PL-94
- These are policy levers provided by the system
- Levers are set by the Census Bureau's Data Stewardship Executive Policy Committee

# Examples from the 1940 Census of Population


National
1 – TVD (Avg)
1.00
0.95
0.90
0.85
0.25
0.50
1.00
2.00
Privacy–loss Budget (ε)
Top–down
Bottom–up


State
1 – TVD (Avg)
1.00
0.95
0.90
0.85
0.25
0.50
1.00
2.00
Privacy–loss Budget (ε)
Top–down
Bottom–up


County
1 – TVD (Avg)
1.00
0.95
0.90
0.85
0.25
0.50
1.00
2.00
Privacy–loss Budget (ε)
Top–down
Bottom–up


Enumeration District
1 – TVD (Avg)
1.00
0.95
0.90
0.85
0.25
0.50
1.00
2.00
Privacy–loss Budget (ε)
Top–down
Bottom–up

# Managing the Tradeoff

You know what I look like already

# An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices

John M. Abowd and Ian M. Schmutte

August 15, 2018

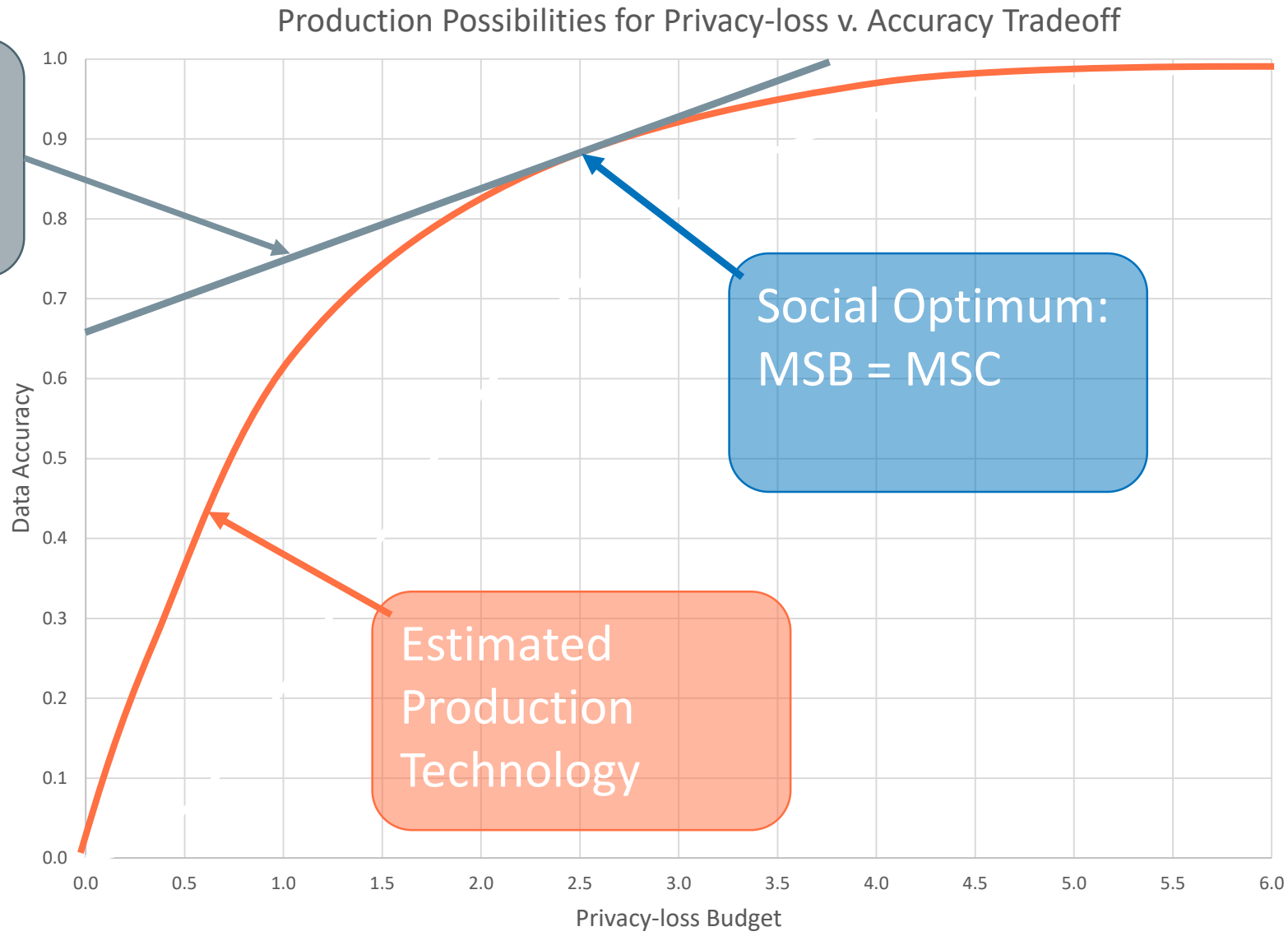
Forthcoming in *American Economic Review*

Abowd: U.S. Census Bureau HQ 8H120, 4600 Silver Hill Rd., Washington, DC 20233, and Cornell University, (email: john.maron.abowd@census.gov); Schmutte: Department of Economics, University of Georgia, B408 Amos Hall, Athens, GA 30602 (email: schmutte@uga.edu).
Abowd and Schmutte acknowledge the support of Alfred P. Sloan Foundation Grant G-2015-13903 and NSF Grant SES-1131848. Abowd acknowledges direct support from NSF Grants BCS-0941226, TC-1012593. Any opinions and conclusions are those of the authors and do not represent the views of the Census Bureau, NSF, or the Sloan Foundation. We thank the Center for Labor Economics

# How to Think about the Social Choice Problem

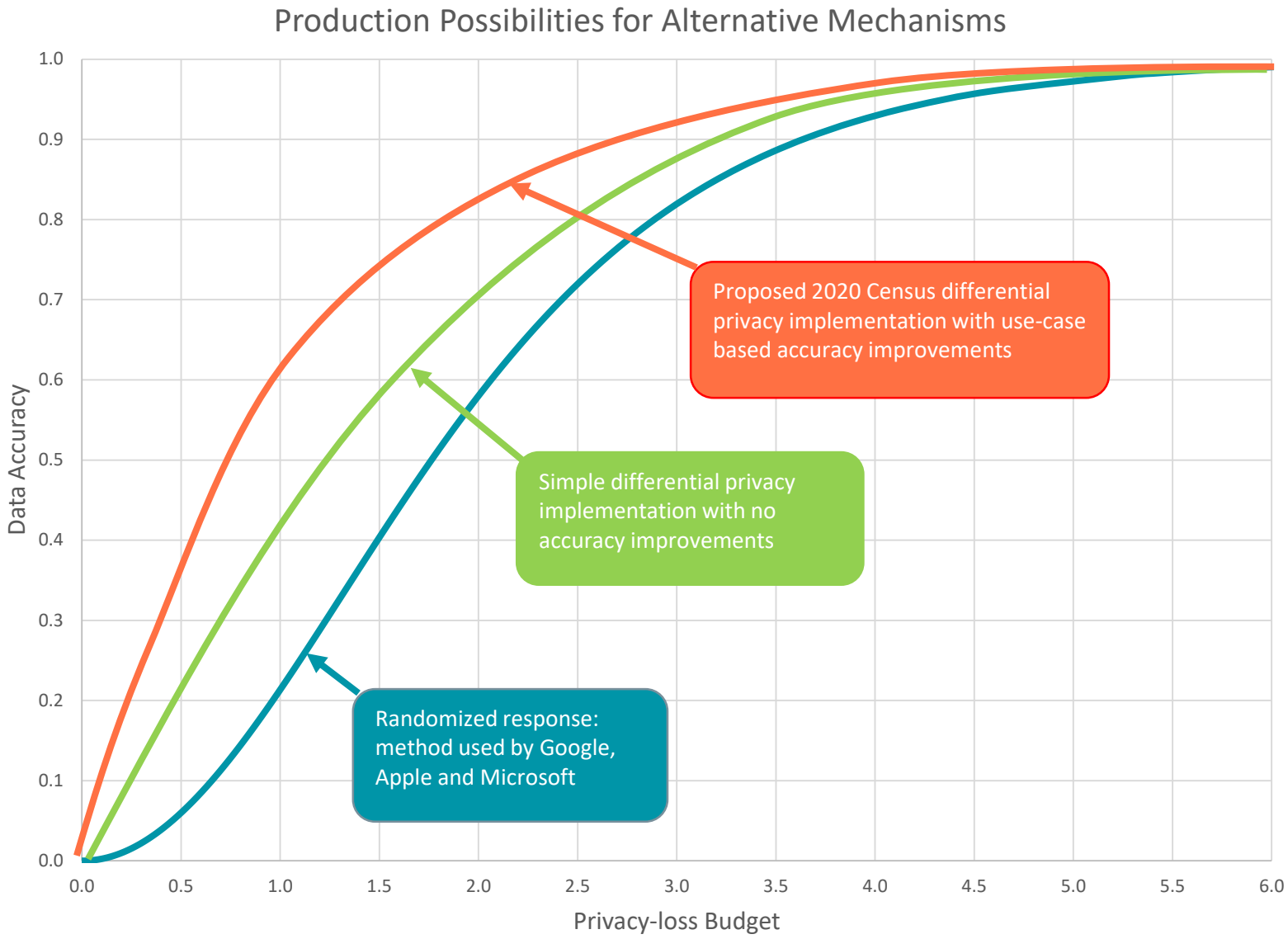
- The marginal social benefit is the sum of all persons' willingness-to-pay for data accuracy with increased privacy loss
- The marginal rate of transformation is the slope of the privacy-loss v. accuracy graphs we have been examining
- This is exactly the same problem being addressed by Google in RAPPOR, Apple in iOS 11, and Microsoft in Windows 10 telemetry

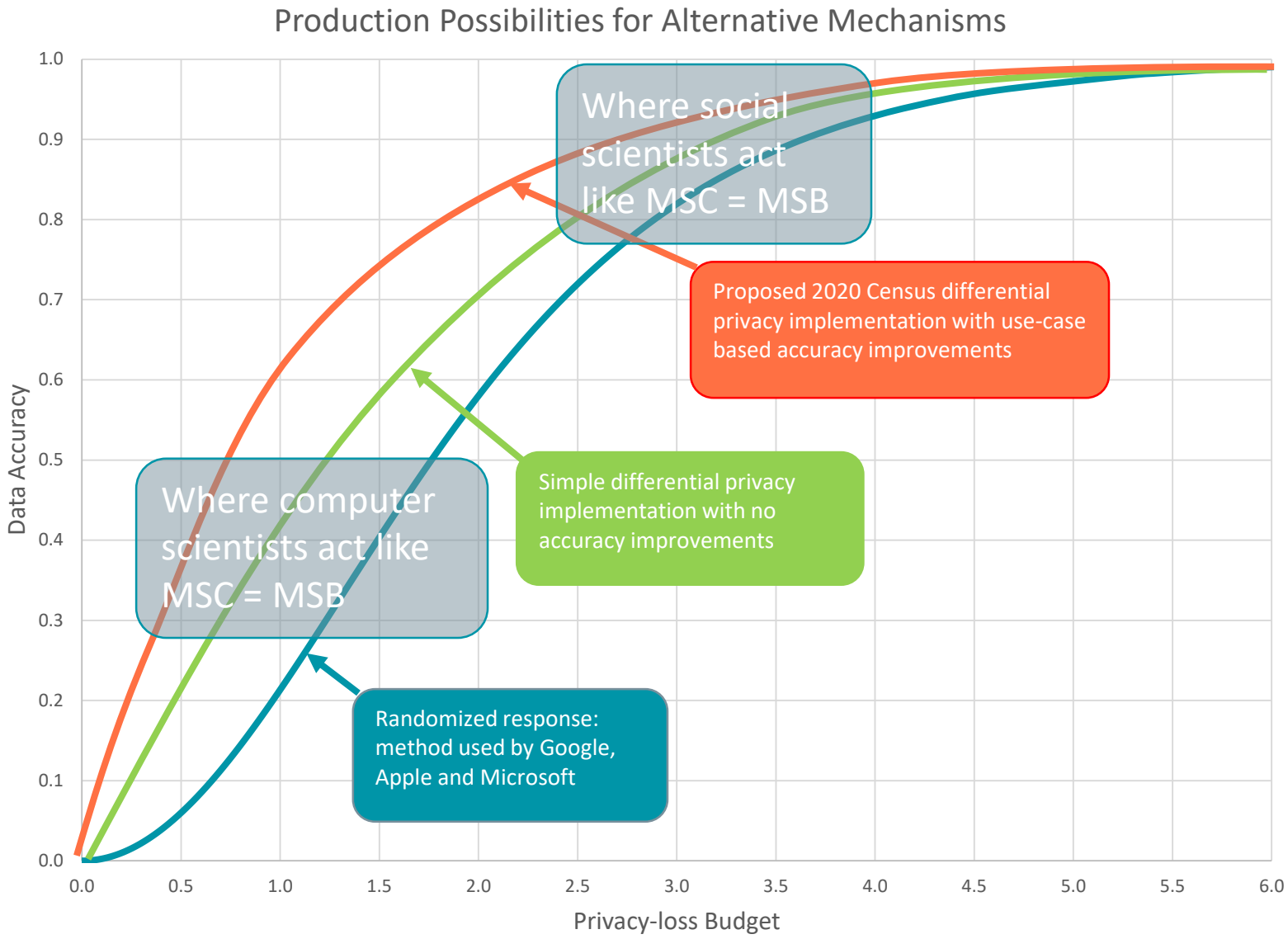
Estimated Marginal Social Benefit Curve



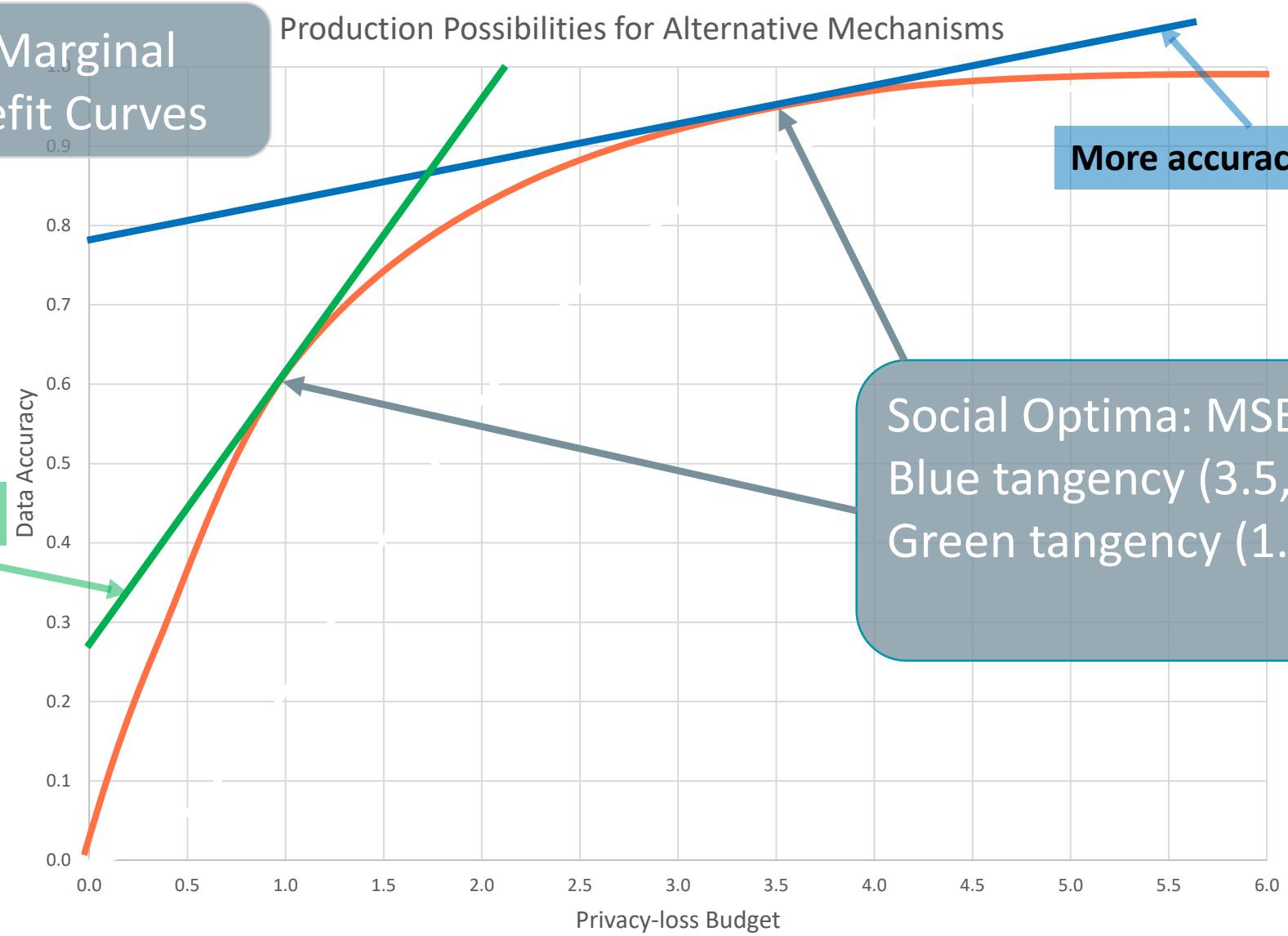
# But the Choice Problem for Redistricting Tabulations Is More Challenging

- In the redistricting application, the fitness-for-use is based on
  - Supreme Court one-person one-vote decision (All legislative districts must have approximately equal populations; there is judicially approved variation)
  - *Is statistical disclosure limitation a "statistical method" (permitted by Utah v. Evans) or "sampling" (prohibited by the Census Act, confirmed in Commerce v. House of Representatives)?*
  - Voting Rights Act, Section 2: requires majority-minority districts at all levels, when certain criteria are met
- The privacy interest is based on
  - Title 13 requirement not to publish exact identifying information
  - The public policy implications of uses of detailed race and ethnicity
- Other use cases: See Federal Register Notice 83 FR 34111 (comments due September 17, 2018)


Production Possibilities for Alternative Mechanisms
Data Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Privacy-loss Budget
Proposed 2020 Census differential privacy implementation with use-case based accuracy improvements
Simple differential privacy implementation with no accuracy improvements
Randomized response: method used by Google, Apple and Microsoft


Production Possibilities for Alternative Mechanisms
Data Accuracy
Privacy-loss Budget
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Where social scientists act like MSC = MSB
Proposed 2020 Census differential privacy implementation with use-case based accuracy improvements
Simple differential privacy implementation with no accuracy improvements
Where computer scientists act like MSC = MSB
Randomized response: method used by Google, Apple and Microsoft

# Estimated Marginal Social Benefit Curves




**More accuracy favoring**

**More privacy favoring**

Social Optima: MSB = MSC
Blue tangency (3.5, 94%)
Green tangency (1.0, 60%)

# Some Other Tools for Managing the Tradeoff

- Machanavajjhala, He and Hay (SIGMOD 2017) Differential Privacy in the Wild: A Tutorial on Current Practices & Open Challenges.
- Harvard Data Privacy Tools Project
  - Kobbi Nissim, Thomas Steinke, Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, David O'Brien, and Salil Vadhan. Forthcoming. Differential Privacy: A Primer for a Non-technical Audience
  - Kobbi Nissim and Alexandra Wood. 2018. Is Privacy Privacy? Philosophical Transaction of the Royal Society A.

# Thank you.

John.Maron.Abowd@census.gov

johnabowd.com

# Selected References


- Dinur, Irit and Kobbi Nissim. 2003. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*(PODS '03). ACM, New York, NY, USA, 202-210. DOI: 10.1145/773153.773173.
- Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. in Halevi, S. & Rabin, T. *(Eds.)* Calibrating Noise to Sensitivity in Private Data Analysis *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings, Springer Berlin Heidelberg*, 265-284, DOI: 10.1007/11681878_14.
- Dwork, Cynthia. 2006. Differential Privacy, *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006), Springer Verlag, 4052*, 1-12, ISBN: 3-540-35907-9.
- Dwork, Cynthia and Aaron Roth. 2014. *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science. Vol. 9, Nos. 3–4. 211–407, DOI: 10.1561/0400000042.
- Dwork, Cynthia, Frank McSherry and Kunal Talwar. 2007. The price of privacy and the limits of LP decoding. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*(STOC '07). ACM, New York, NY, USA, 85-94. DOI:10.1145/1250790.1250804.
- Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd , Johannes Gehrke, and Lars Vilhuber. 2008. Privacy: Theory Meets Practice on the Map, International Conference on Data Engineering (ICDE) 2008: 277-286, doi:10.1109/ICDE.2008.4497436.
- Dwork, Cynthia and Moni Naor. 2010. On the Difficulties of Disclosure Prevention in Statistical Databases or The Case for Differential Privacy, *Journal of Privacy and Confidentiality*: Vol. 2: Iss. 1, Article 8. Available at: http://repository.cmu.edu/jpc/vol2/iss1/8.
- Kifer, Daniel and Ashwin Machanavajjhala. 2011. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data* (SIGMOD '11). ACM, New York, NY, USA, 193-204. DOI:10.1145/1989323.1989345.
- Erlingsson, Úlfar, Vasyl Pihur and Aleksandra Korolova. 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (CCS '14). ACM, New York, NY, USA, 1054-1067. DOI:10.1145/2660267.2660348.
- Abowd, John M. and Ian M. Schmutte. 2017 . Revisiting the economics of privacy: Population statistics and confidentiality protection as public goods. Labor Dynamics Institute, Cornell University, Labor Dynamics Institute, Cornell University, at https://digitalcommons.ilr.cornell.edu/ldi/37/
- Abowd, John M. and Ian M. Schmutte. Forthcoming. An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. American Economic Review, at https://arxiv.org/abs/1808.06303
- Apple, Inc. 2016. Apple previews iOS 10, the biggest iOS release ever. Press Release (June 13). URL=http://www.apple.com/newsroom/2016/06/apple-previews-ios-10-biggest-ios-release-ever.html.
- Ding, Bolin, Janardhan Kulkarni, and Sergey Yekhanin 2017. Collecting Telemetry Data Privately, NIPS 2017.