What Does It Really Mean to Be a 21st Century Statistical Agency?

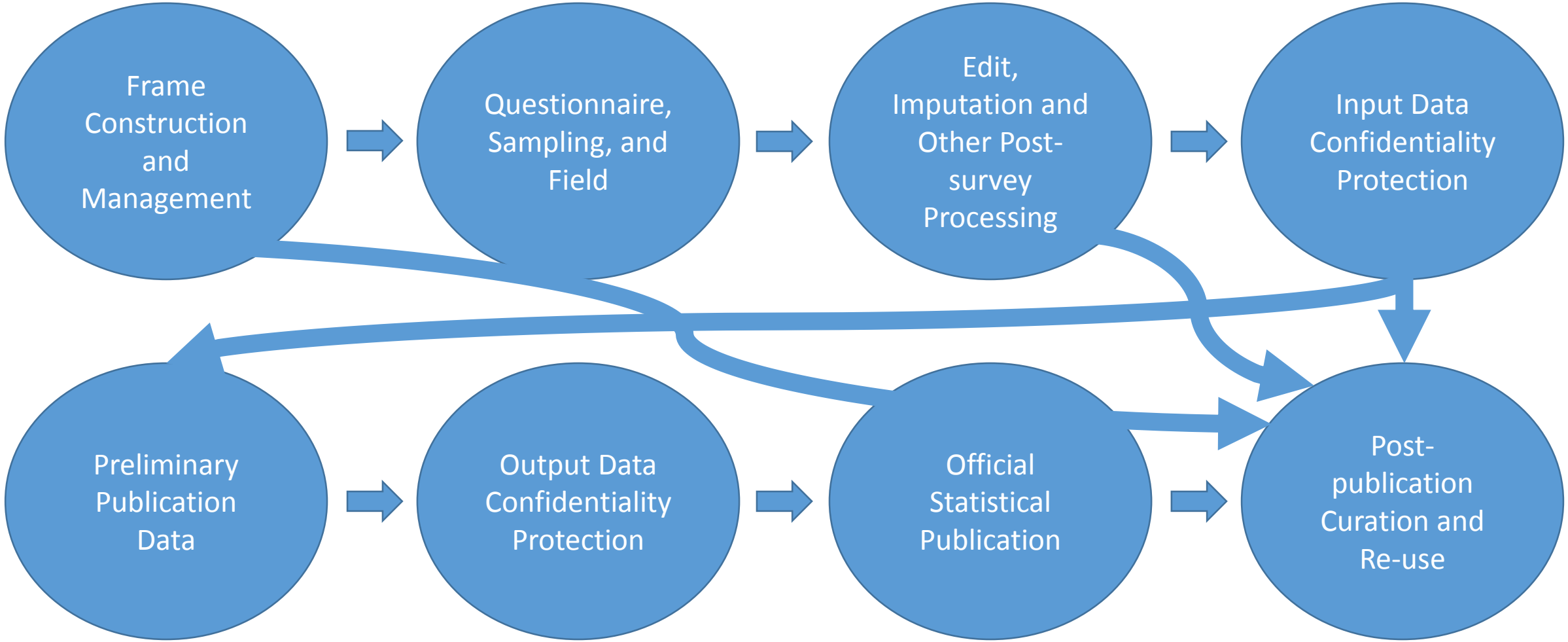
John M. Abowd
U.S Census Bureau
Federal Computer Assisted Survey Information Collection Workshops
April 11-12, 2017

Acknowledgements and Disclaimer

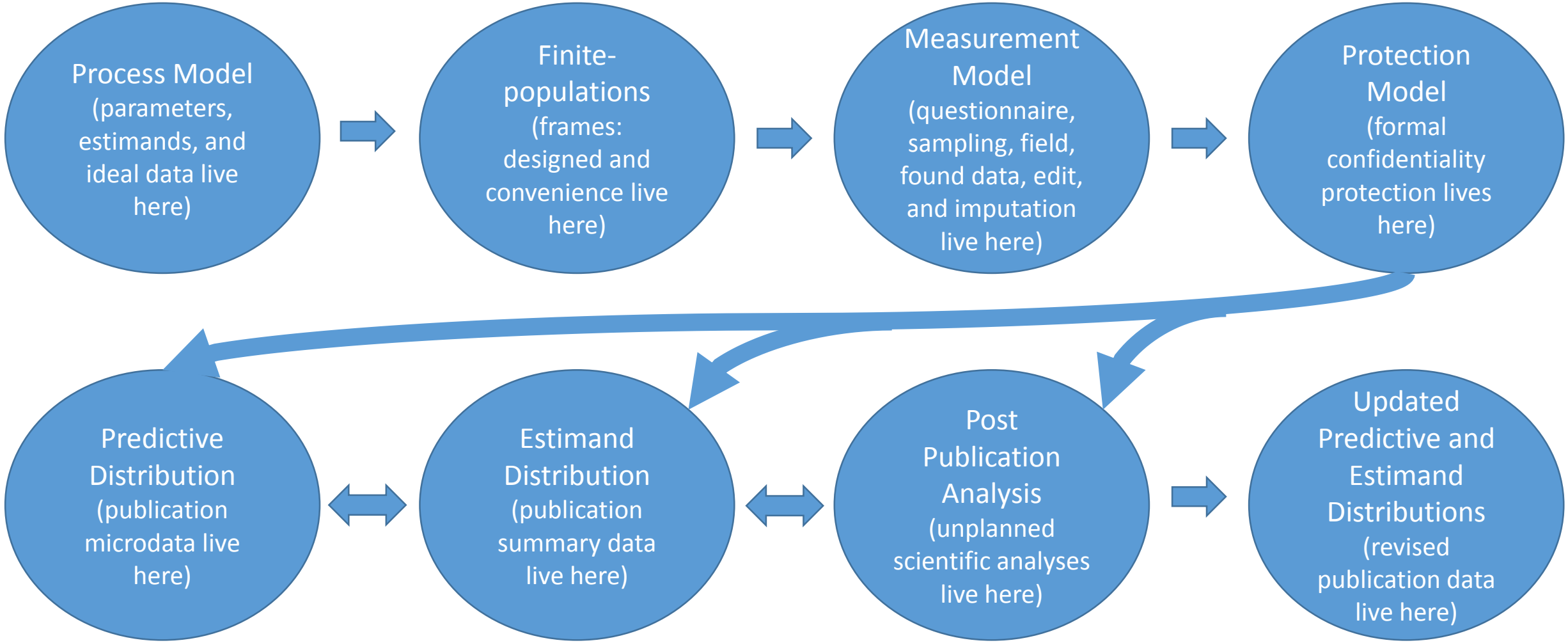
- Many of the research ideas in this talk were produced by collaborations in the National Science Foundation-Census Research Network
- Support from the NSF and the Alfred P. Sloan is gratefully acknowledged

- The opinions expressed in this talk are my own, and not those of the Census Bureau or any of the research sponsors

Designed Survey Data


Frame Construction and Management
Questionnaire, Sampling, and Field
Edit, Imputation and Other Post-survey Processing
Input Data Confidentiality Protection
Preliminary Publication Data
Output Data Confidentiality Protection
Official Statistical Publication
Post-publication Curation and Re-use

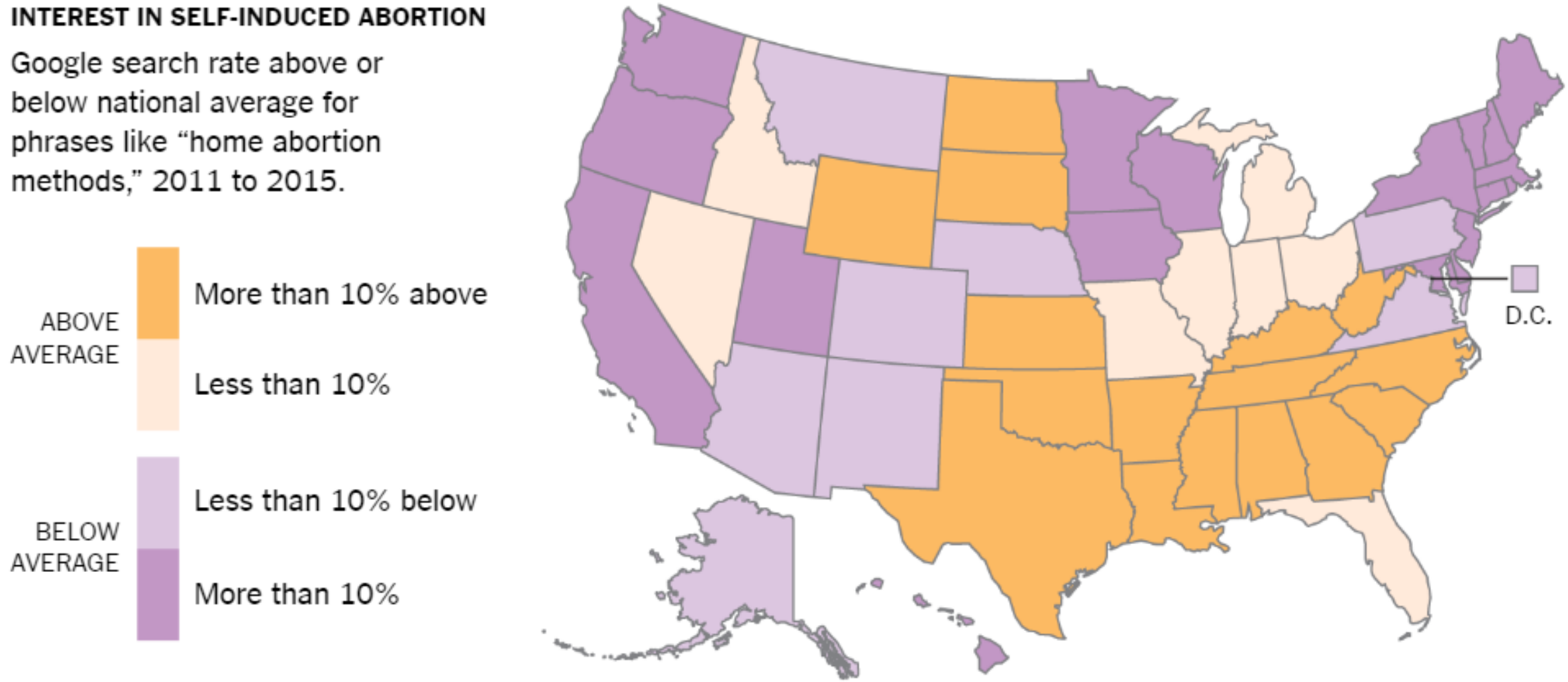
Designed Data


Process Model (parameters, estimands, and ideal data live here)
Finite-populations (frames: designed and convenience live here)
Measurement Model (questionnaire, sampling, field, found data, edit, and imputation live here)
Protection Model (formal confidentiality protection lives here)
Predictive Distribution (publication microdata live here)
Estimand Distribution (publication summary data live here)
Post Publication Analysis (unplanned scientific analyses live here)
Updated Predictive and Estimand Distributions (revised publication data live here)

The Data Bloggers

INTEREST IN SELF-INDUCED ABORTION

Google search rate above or below national average for phrases like “home abortion methods,” 2011 to 2015.



Sources: Guttmacher Institute (state laws); analysis of Google data by Seth Stephens-Davidowitz (searches)

By Bill Marsh/The New York Times

How Would You Assess the Suitability for Use of These Data?

- How large a survey would you need to validate charts like these?
- What is the required sample size where the MSE of the simple random sample is lower than the MSE of the found data?
- If the incidence is low, and the frame bias is 20%?
- If the incidence is low, and the frame bias is 5%?

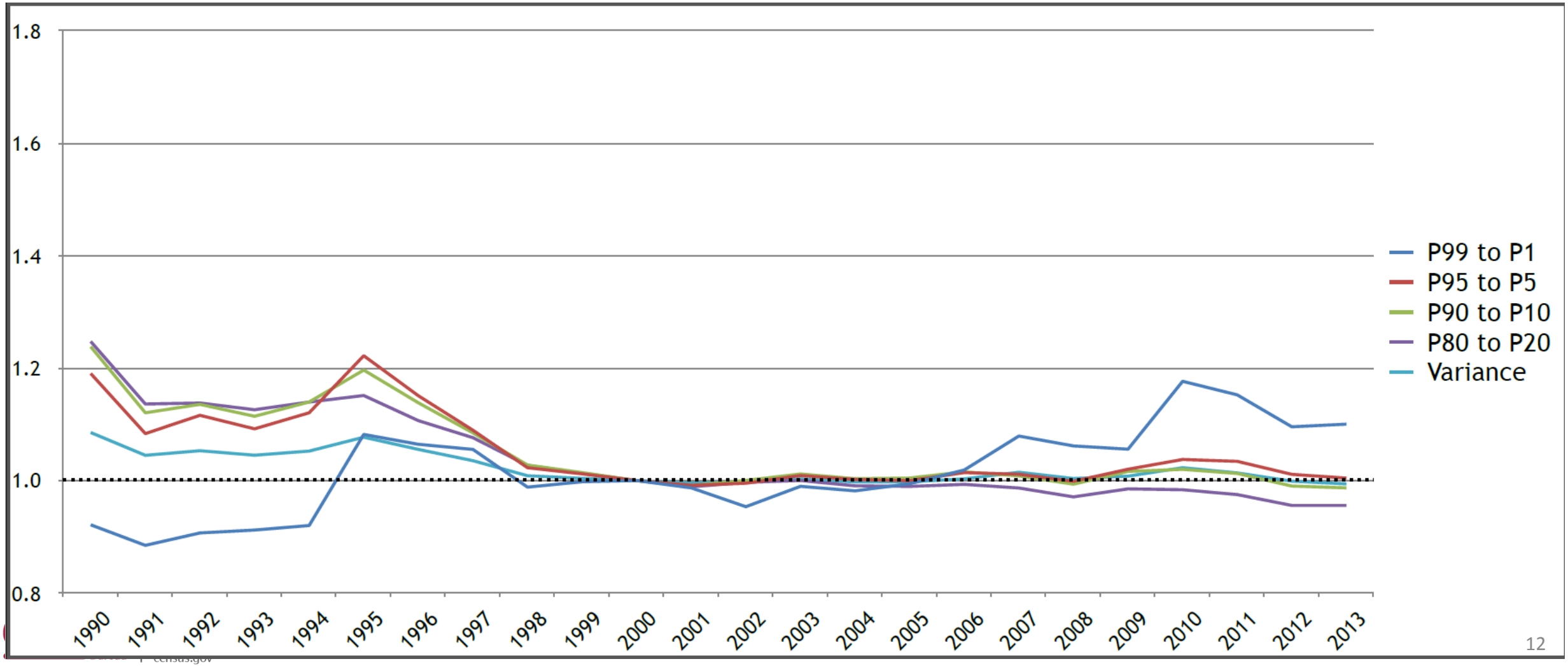
Every Official Statistician Should Quote This Slide!

- Raghunathan (2015)
- If the expected incidence is 10%, and the expected bias in the found data is 20%, then a simple random survey of 290 cases has lower MSE than found data with one billion cases
- If the expected incidence is 10%, and the expected bias in the found data is 5%, then a simple random survey of 4,500 cases has lower MSE than found data with one billion cases
- This is why designed surveys are an essential component of designed data

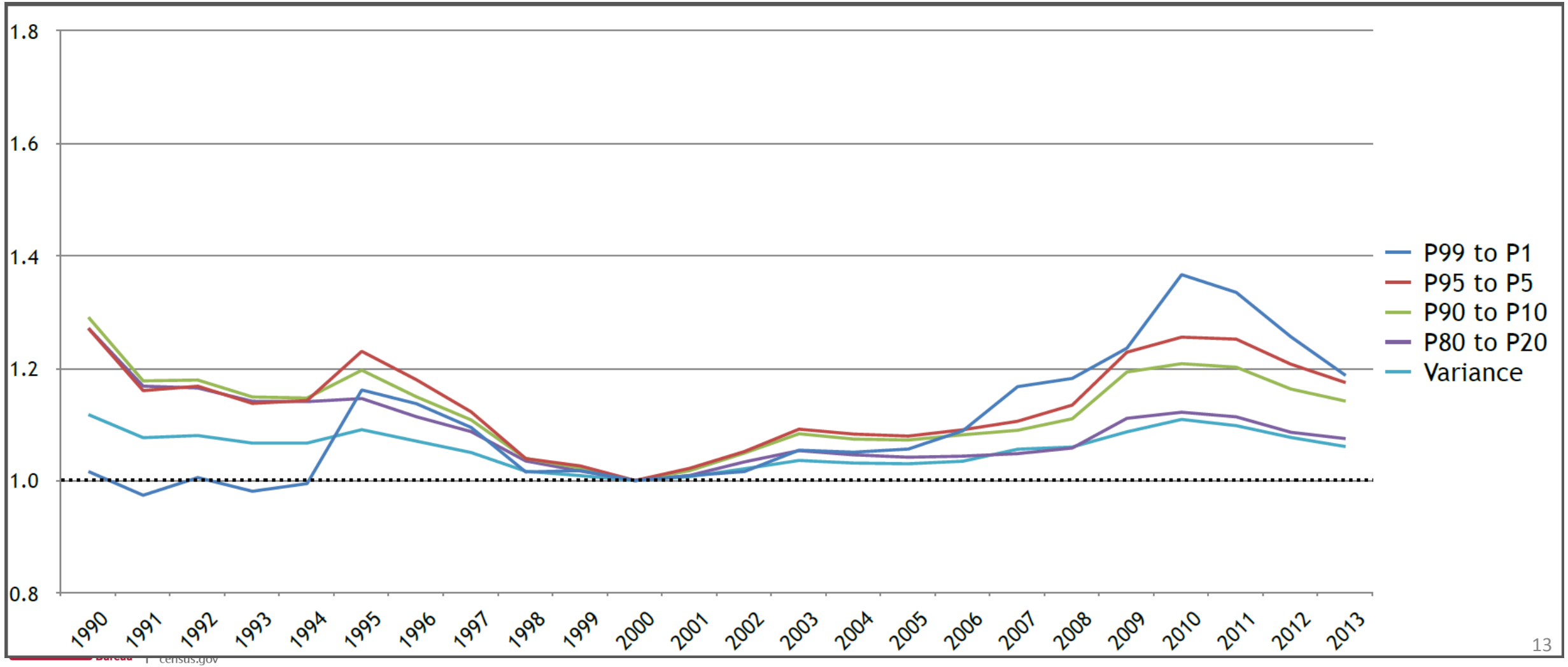
Example of Correcting Frame Bias in Found Data

- Estimating the earnings distribution from administrative data on quarterly earnings
- From Abowd, McKinney and Zhao (2017)

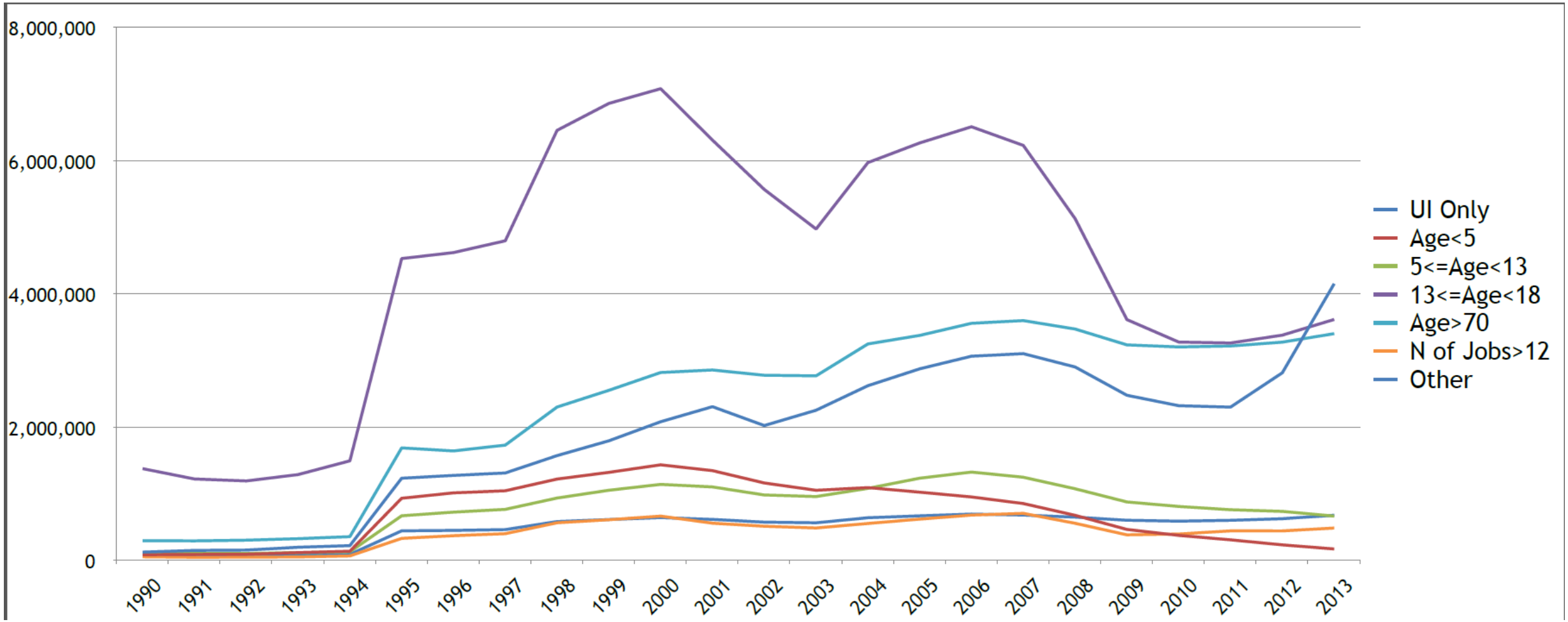
Earnings Inequality in the U.S.-Uncorrected



Earnings Inequality in the U.S.-Corrected



Records Removed from the Frame and Why

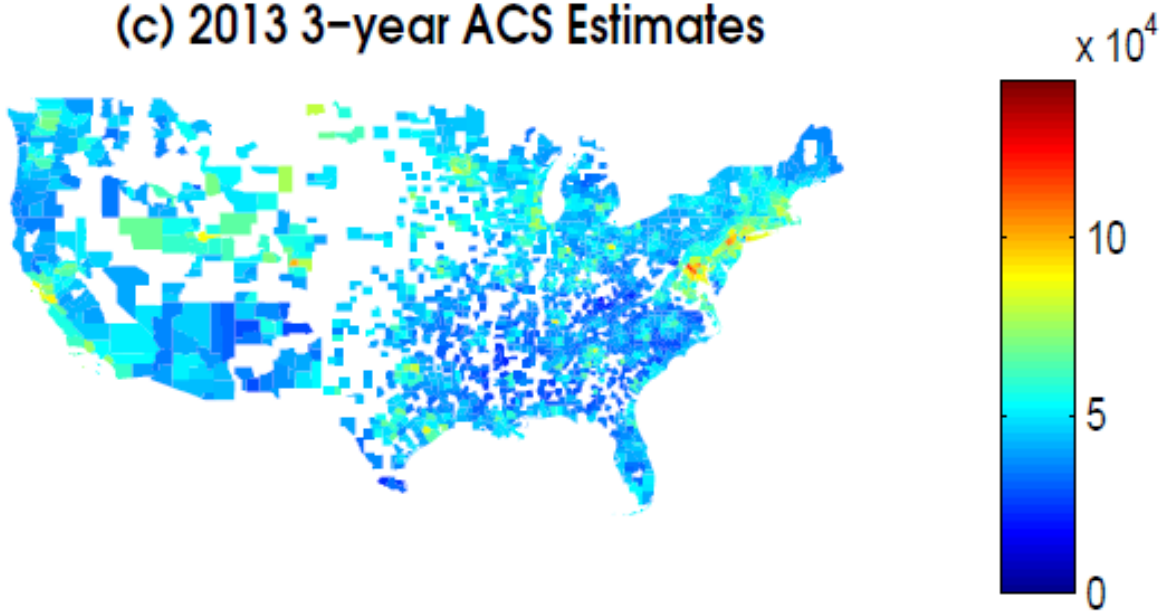


Let Me Model That for You

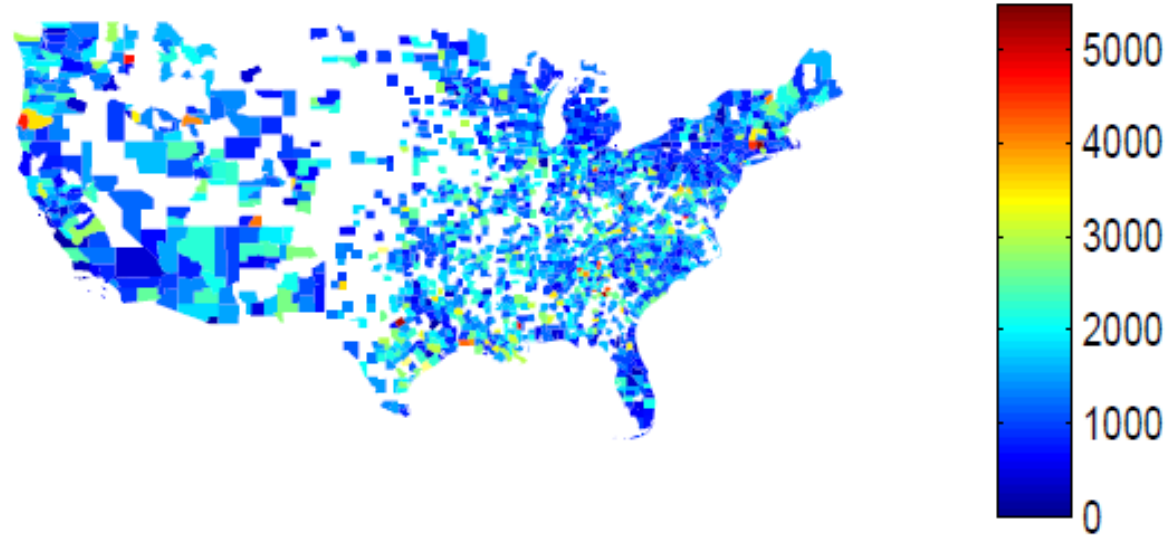
Designed Data Borrow Strength from Multiple Sources

- This example is from Bradley, Wikle and Holan (2015)
- Improving the areal coverage of American Community Survey estimates

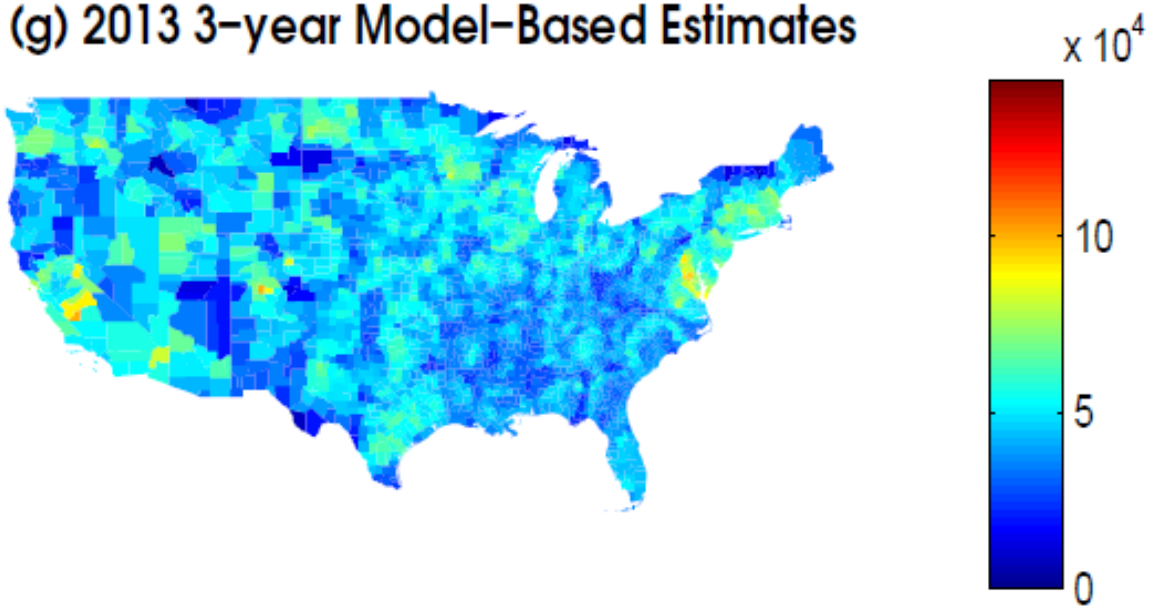
(c) 2013 3-year ACS Estimates



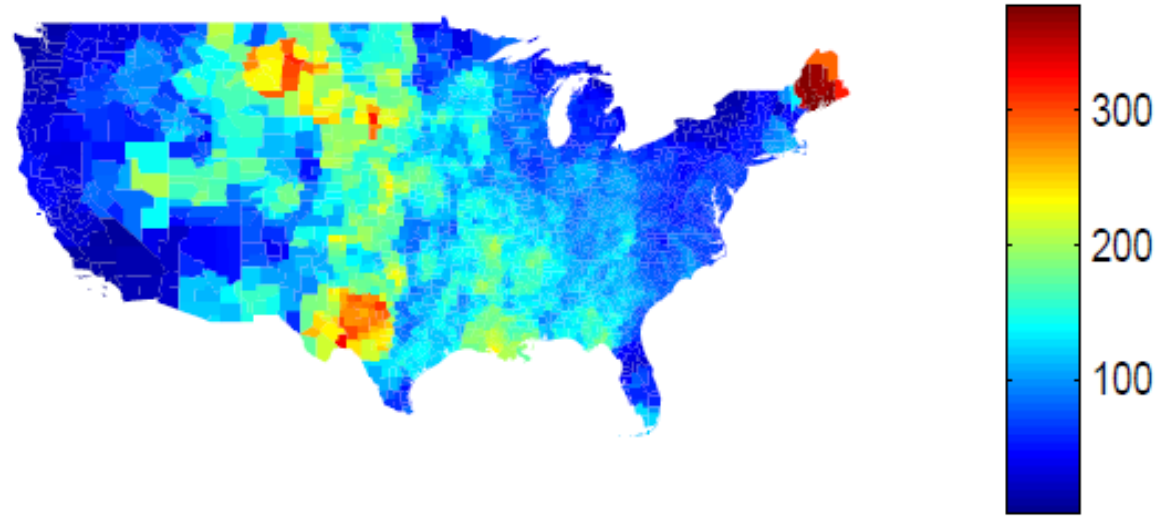
(d) 2013 3-year ACS Estimates of Std.Dev



(g) 2013 3-year Model-Based Estimates



(h) Posterior Standard Deviation



MANHATTAN

1,2 WALL STREET, CIVIC CENTER, GOVERNORS ISLAND, LIBERTY ISLAND, ELLIS ISLAND, TRIBECA, GREENWICH VILLAGE, NOHO, SOHO, LITTLE ITALY
- 42 percent more households with children
- 54 percent more people age 55 to 64
- 137 percent more residents who work in protective services (police, security, etc.)

3 LOWER EAST SIDE, CHINATOWN
- 55 percent more adults with bachelor's degrees but no higher degrees
- 43 percent more households of men living alone
- 24 percent fewer Hispanic residents

4,5 CHELSEA, HELL'S KITCHEN, HERALD SQUARE, MIDTOWN, TIMES SQUARE
- 30 percent more residents age 35 to 44
- 21 percent fewer households headed by women
- 52 percent fewer homes owned and occupied by Hispanics

6 MURRAY HILL, EAST MIDTOWN, STUYVESANT TOWN
- 33 percent fewer adults with some college education but no four-year degrees
- 42 percent fewer residents who work in transportation
- 15 percent fewer residents who are widowed, divorced or separated

7 UPPER WEST SIDE, LINCOLN SQUARE
- 34 percent fewer Hispanic families
- 24 percent more married residents
- 105 percent more children under 5

- 46 percent fewer residents who work in construction and manufacturing

2 SUNNYSIDE, WOODSIDE
- 29 percent more residents age 55 to 64
- 39 percent fewer residents who work in construction and manufacturing
- 17 percent more residents who are widowed, divorced or separated

3 JACKSON HEIGHTS, EAST ELMHURST, NORTH CORONA
- 29 percent fewer black households

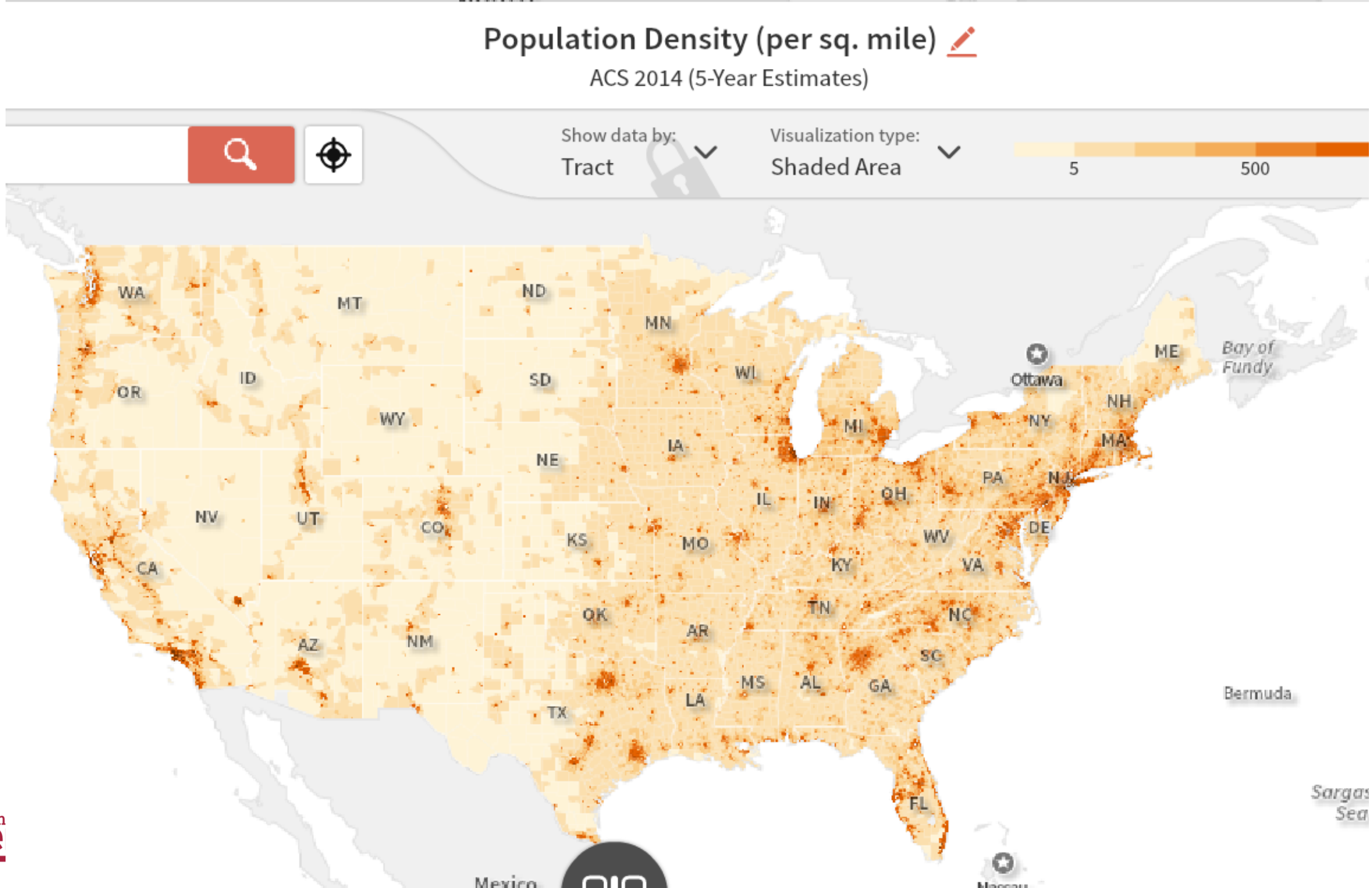
4 ELMHURST, CORONA
- 36 percent fewer blacks
- 24 percent fewer households of women living alone

5 MASPETH, RIDGEWOOD, MIDDLE VILLAGE, GLENDALE
- 56 percent more residents who work in health care
- 14 percent fewer households of women living alone
- 26 percent fewer residents with less than a high school education

6 REGO PARK, FOREST HILLS
- 47 percent more residents who work in building services (janitors, superintendents, etc.)
- 41 percent more residents who work in restaurants and food services
- 48 percent fewer black families

7 FLUSHING, WHITESTONE, COLLEGE POINT
- 45 percent more employed workers
- 50 percent more residents who work in personal services

8 FRESH MEADOWS, JAMAICA HILLS, KEW GARDENS HILLS
- 24 percent more households of men living alone


Population Density (per sq. mile)
ACS 2014 (5-Year Estimates)
Show data by:
Tract
Visualization type:
Shaded Area
5
500
WA
MT
ND
MN
OR
ID
SD
WI
MI
WY
NY
NE
IA
PA
NJ
NV
UT
CO
IL
IN
OH
WV
KS
MO
DE
CA
KY
VA
TN
NC
OK
AR
AZ
NM
SC
MS
AL
GA
LA
TX
FL
ME
NH
MA
Ottawa
Bay of Fundy
Bermuda
Mexico

What If All Data Were Private?

RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response

Úlfar Erlingsson
Google, Inc.
ulfar@google.com

Vasyl Pihur
Google, Inc.
vpihur@google.com

Aleksandra Korolova
University of Southern California
korolova@usc.edu

ABSTRACT

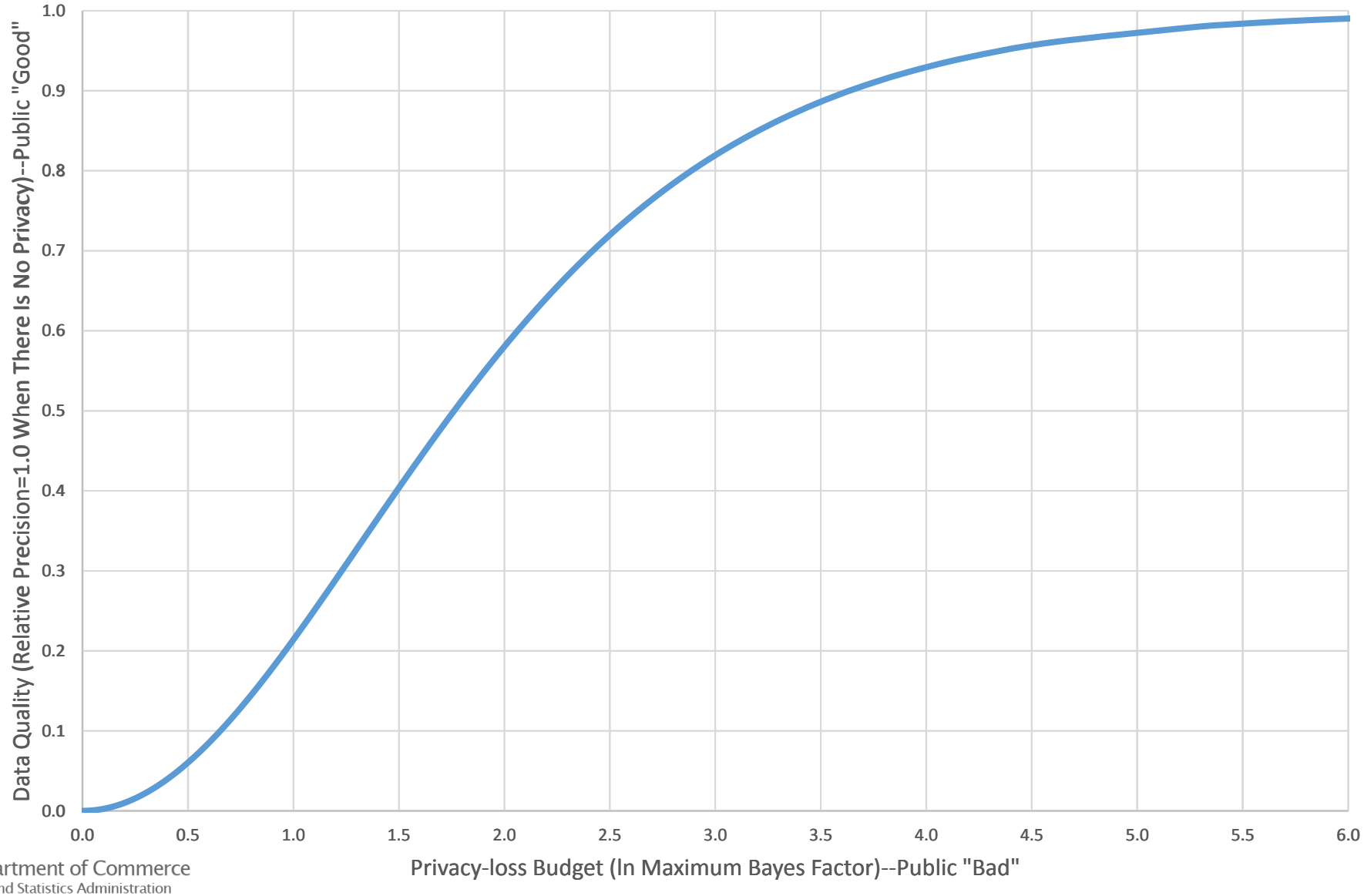
Randomized Aggregatable Privacy-Preserving Ordinal Response, or RAPPOR, is a technology for crowdsourcing statistics from end-user client software, anonymously, with strong privacy guarantees. In short, RAPPORs allow the forest of client data to be studied, without permitting the possibility of looking at individual trees. By applying randomized response in a novel manner, RAPPOR provides the mechanisms for such collection as well as for efficient, high-utility analysis of the collected data. In particular, RAPPOR permits statistics to be collected on the population of client-side strings with strong privacy guarantees for each client, and without linkability of their reports.

This paper describes and motivates RAPPOR, details its differential-privacy and utility guarantees, discusses its practical deployment and properties in the face of different attack models, and finally, gives results of its application to both

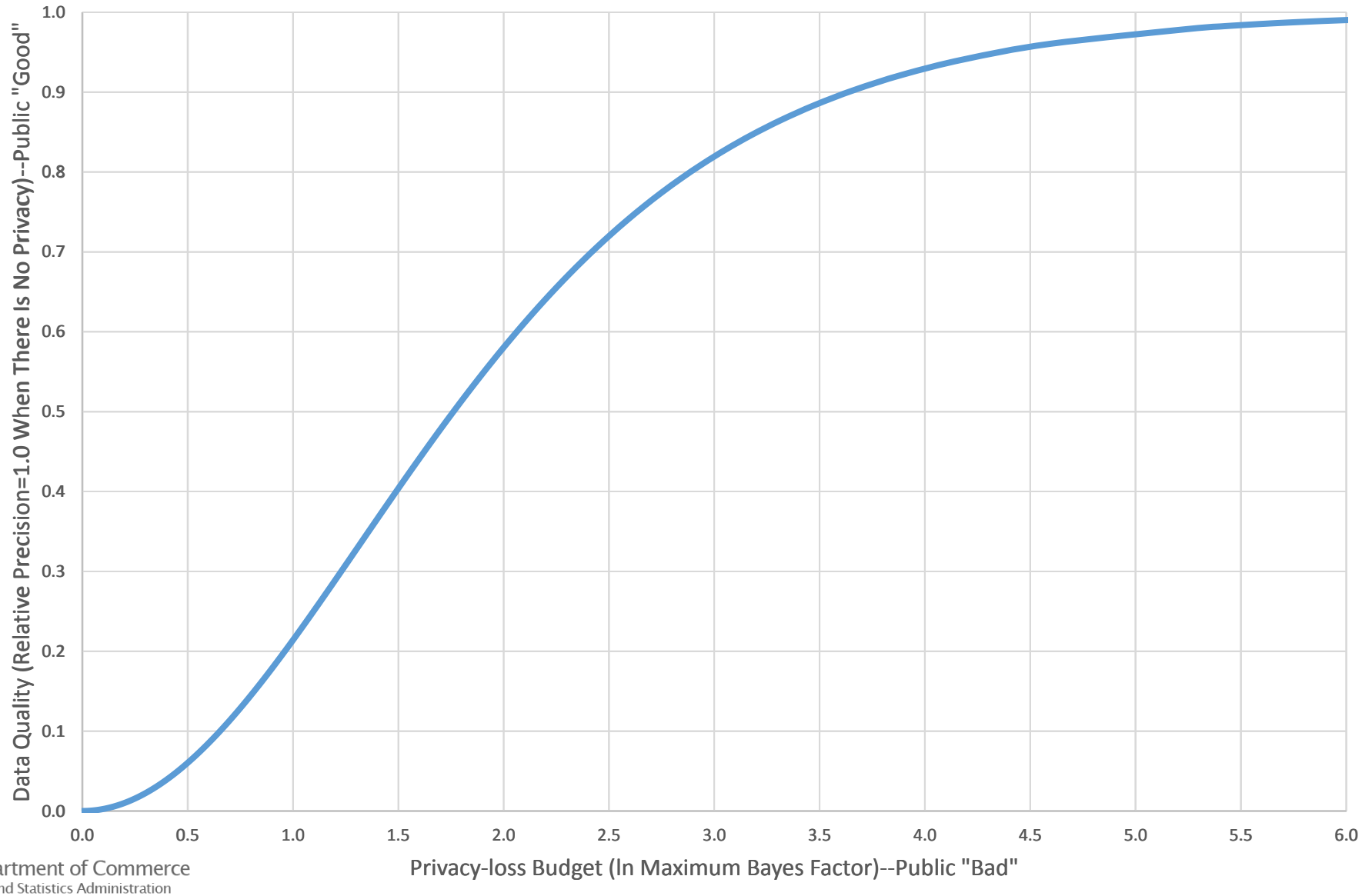
asked to flip a fair coin, in secret, and answer "Yes" if it comes up heads, but tell the truth otherwise (if the coin comes up tails). Using this procedure, each respondent retains very strong deniability for any "Yes" answers, since such answers are most likely attributable to the coin coming up heads; as a refinement, respondents can also choose the untruthful answer by flipping another coin in secret, and get strong deniability for both "Yes" and "No" answers.

Surveys relying on randomized response enable easy computations of accurate population statistics while preserving the privacy of the individuals. Assuming absolute compliance with the randomization protocol (an assumption that may not hold for human subjects, and can even be non-trivial for algorithmic implementations [23]), it is easy to see that in a case where both "Yes" and "No" answers can be denied (flipping two fair coins), the true number of "Yes" answers can be accurately estimated by $2(Y - 0.25)$, where

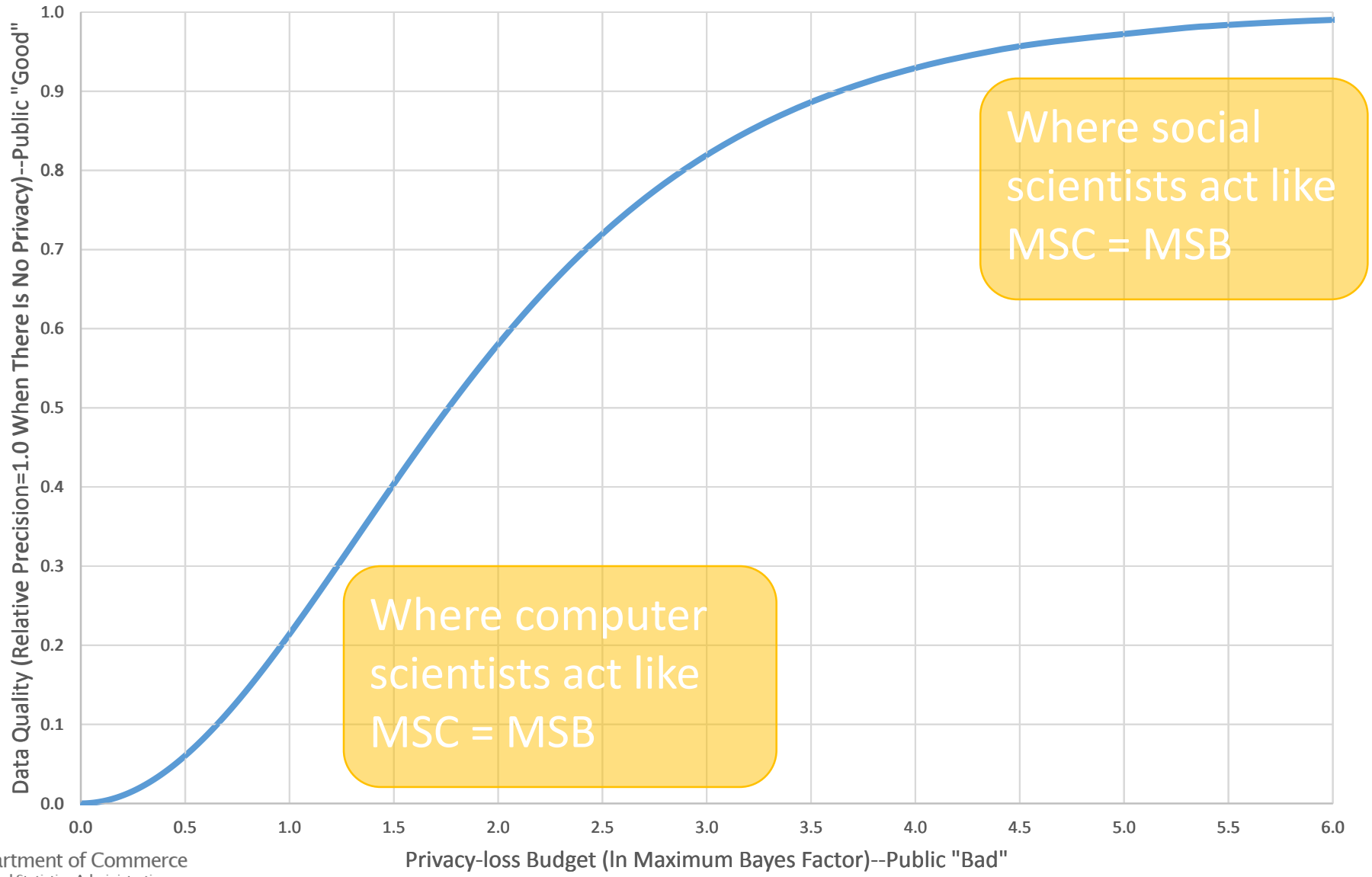
Production Possibilities Frontier/Risk-Utility/Receiver Operation Characteristics for Statistical Disclosure Limitation via Randomized Response



Production Possibilities Frontier/Risk-Utility/Receiver Operation Characteristics for Statistical Disclosure Limitation via Randomized Response

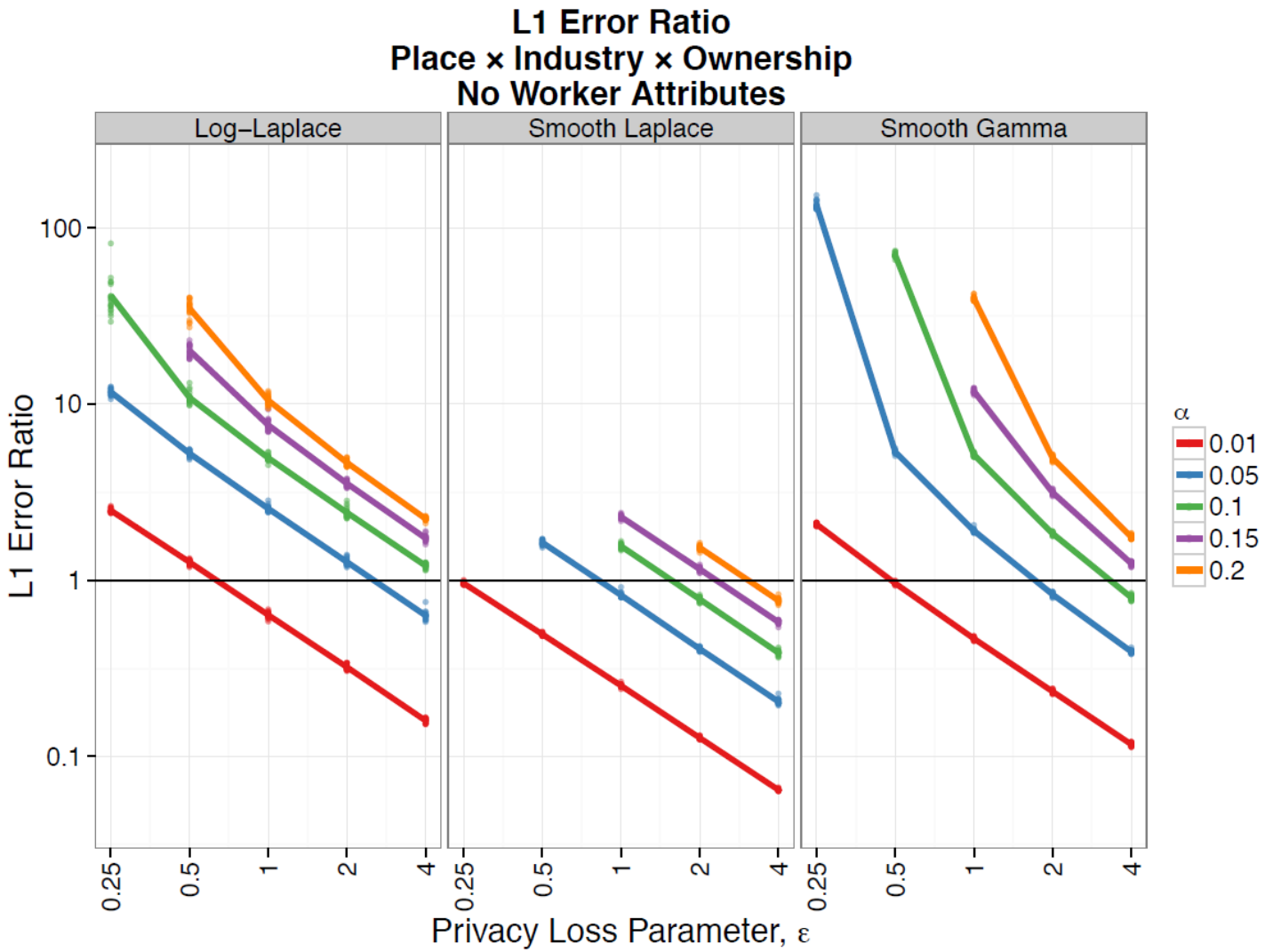


Production Possibilities Frontier/Risk-Utility/Receiver Operation Characteristics for Statistical Disclosure Limitation via Randomized Response



Some Examples

- Dwork (2008): “The parameter e in Definition 1 is public. The choice of e is essentially a social question and is beyond the scope of this paper.” [link, p. 3]
- Dwork (2011): “The parameter e is public, and its selection is a social question. We tend to think of e as, say, 0.01, 0.1, or in some cases, ln 2 or ln 3.” [link, p. 91]
- In OnTheMap, e = 8.9, was required to produce tract-level estimates with acceptable accuracy



Source: Haney et al. 2017

Can We Make Our Science Better and Reproducible?

ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

Survey of Income and Program Participation

About this Survey
Information for Respondents
Data
Events
Guidance for Data Users
DataFerrett
SIPP FTP site
Synthetic SIPP Data
SIPP Users' Guide
Methodology
News
Publications
Technical Documentation
Working Papers

Synthetic SIPP Data

Background on the SIPP Synthetic Beta

The SIPP Synthetic Beta (SSB) is a Census Bureau product that integrates person-level micro-data from a household survey with administrative tax and benefit data. These data link respondents from the Survey of Income and Program Participation (SIPP) to Social Security Administration (SSA)/Internal Revenue Service (IRS) Form W-2 records and SSA records of retirement and disability benefit receip and were produced by Census Bureau staff economists and statisticians in collaboration with researchers at Cornell University, the SSA and the IRS. The purpose of the SSB is to provide access to linked data that are usually not publically available due to confidentiality concerns. To overcome these concerns, Census synthesizes, or models, all the variables in a way that changes the record of each individual so as to preserve the underlying covariate relationships between the variables. Only gender and a link to the first reported marital partner are not altered by the synthesis process and still contain their original values.

Nine SIPP panels (1984, 1990, 1991, 1992, 1993, 1996, 2001, 2004, and 2008) form the basis for the SSB, with a subset of variables available across all the panels selected for inclusion and harmonization of variable definitions across the years covered by the panels. Administrative data are added and some editing is done to correct for logical inconsistencies in the IRS and Social Security earnings and benefits data. Thus, the SSB is a particularly appealing data set for new SIPP users because little data preparation is needed. A complete list of variables included in SSB version 6.0, along with details about the harmonization and editing, is available in our Codebook.

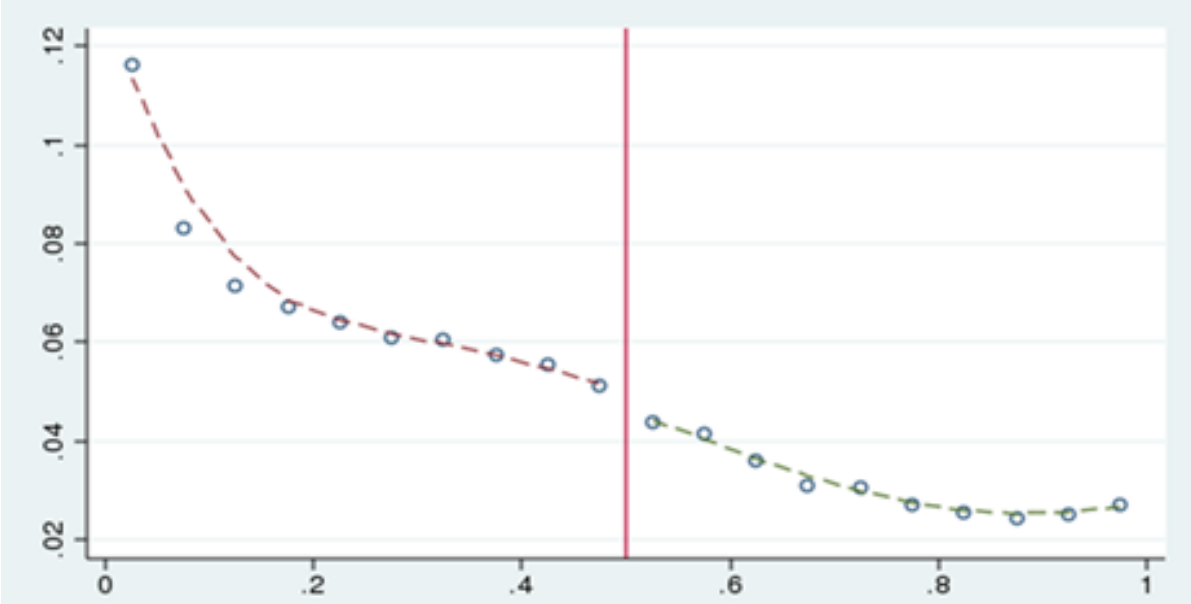
As part of the synthesis process, missing survey data and missing administrative data were multiply imputed. The resulting data sets are called the Completed Gold Standard Files and contain all original, non-missing, confidential values and imputed values in place of originally missing data. These files form the basis for evaluating results from the synthetic data. The goal of the SSB is to produce results that are qualitatively the same as results from the Completed Gold Standard Files.

The Synthetic Longitudinal Business Database

Based on presentations by Kinney/Reiter/Jarmin/Miranda/Reznek[2]/Abowd
on July 31, 2009 at the
Census-NSF-IRS Synthetic Data Workshop
[link] [link]

Kinney/Reiter/Jarmin/Miranda/Reznek/Abowd (2011) "Towards Unrestricted Public Use Microdata: The Synthetic Longitudinal Business Database.", CES-WP-11-04

Work on the Synthetic LBD was supported by NSF Grant ITR-0427889, and ongoing work is supported by the Census Bureau. A portion of this work was conducted by Special Sworn Status researchers of the U.S. Census Bureau at the Triangle Census Research Data Center. Research results and conclusions expressed are those of the authors and do not necessarily reflect the views of the Census Bureau. Results have been screened to ensure that no confidential data are revealed.



Bertrand, Kamenica and Pan (QJE 2015), doi: 10.1093/qje/qjv001

Reproducible science and formal privacy protection are joined at the hip

- Reproducible science:
  - Provenance control and certification
  - Output verification from certified inputs
  - Archiving
  - Curation of data and metadata
- Formal privacy protection:
  - A confidential database contains a finite amount of information
  - Every published use exposes some of this information
  - This privacy loss must be quantified
  - Once quantified, it is public-policy decision how to manage it

Putting the Pieces Together

Suppose we wanted to design a new, continuously updated information system on local labor markets.

Use the ideas from “Data bloggers” to properly combine contemporaneous elements harvested from the data jungle with designed elements produced by the agency.

Use the ideas from “Let me model that for you” to produce local estimates and measures of reliability for all local areas every period, including periods when the designed content is not in the field.

Use the ideas from “What if all data were private?” to provably protect the design-consistent, model-based estimates from all future privacy attacks.

Use the ideas from “Can We Make Our Science Better and Reproducible?” to open a portal to the underlying data that returns safe estimates of hypotheses (i.e., estimates that have a controlled false discovery rate) and incorporates them into future versions of the model.

There are working prototypes of all these pieces running now. That’s where I got the graphics in this talk.

References in order of citation in the talk

Raghunathan (2015) "Statistical Challenges in Combining Information from Big and Small Data Sources" (public version)

Lohr and Raghunathan (forthcoming) "Combining Survey Data with Other Data Sources"

Stephens-Davidowitz and Varian (2015) "A Hands-on Guide to Google Data"

Abowd, McKinney and Zhao (2017) "Earnings Inequality and Mobility Trends in the United States: Nationally Representative Estimates from Longitudinally Linked Employer-Employee Data"

Bradley, Wikle and Holan (2015) "Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates"

Bradley, Holan and Wikle (2015) "Multivariate spatio-temporal models for high-dimensional areal data with application to Longitudinal Employer-Household Dynamics"

Erlingsson, Pihur and Korolova (2014) "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response"

Abowd and Schmutte (2017) "Revisiting the Economics of Privacy"

Haney et al. (2017) "Formal privacy protection for data products combining individual and employer frames" (public version)

American Statistical Association Releases Statement on Statistical Significance and P-values (2016)

Bertrand, Kamenica and Pan (2015) "Gender Identity and Relative Income within Households"

Synthetic SIPP Data (2017)

Synthetic LBD (2017)

Abowd and Schmutte (2015) "Economic Analysis and Statistical Disclosure Limitation"

Thank you.

Contact: john.maron.abowd@census.gov