INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER
PANEL STUDY OF INCOME DYNAMICS
UNIVERSITY OF MICHIGAN

© 2014 Ted Goff

"I'm Big Data, and this is my friend No Privacy."

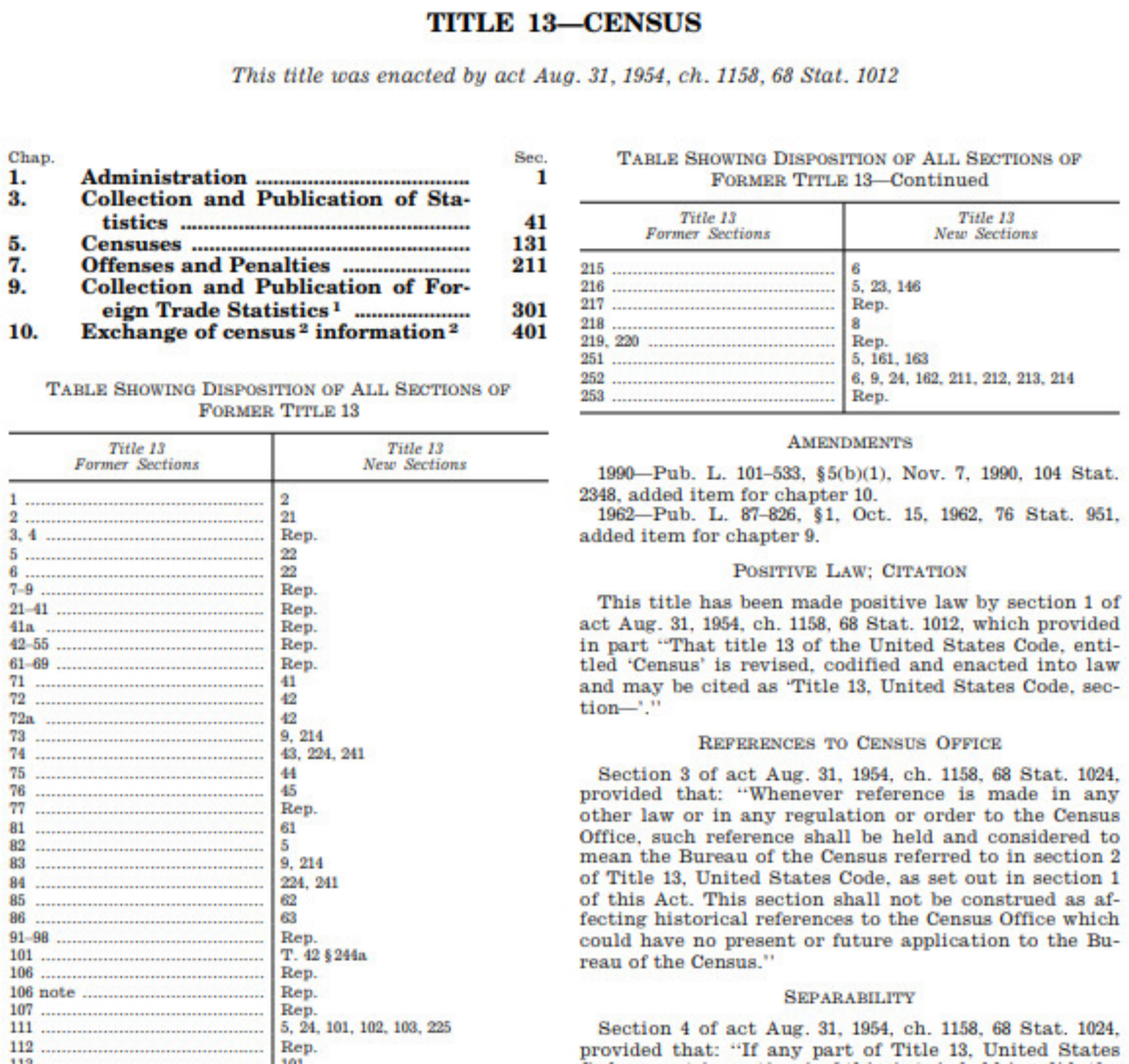
Privacy and Science

David Johnson
University of Michigan

Why we need disclosure techniques and why differential privacy

- Title 13 requires non-disclosure
  - Reconstruction of data cannot be possible
  - Rules cannot be released
- Fundamental Law of Information Recovery
- Commission on Evidence-based policymaking recommendations
- NAS report, Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy, recommendations
- “Big data” companies use DP

Title 13 and Privacy



TITLE 13—CENSUS

This title was enacted by act Aug. 31, 1954, ch. 1158, 68 Stat. 1012

TABLE SHOWING DISPOSITION OF ALL SECTIONS OF FORMER TITLE 13

Title 13 Former Sections	*Title 13 New Sections*
1	2
2	21
3, 4	Rep.
5	22
6	22
7–9	Rep.
21–41	Rep.
41a	Rep.
42–55	Rep.
61–69	Rep.
71	41
72	42
72a	42
73	9, 214
74	43, 224, 241
75	44
76	45
77	Rep.
81	61
82	5
83	9, 214
84	224, 241
85	62
86	63
91–98	Rep.
101	T. 42 § 244a
106	Rep.
106 note	Rep.
107	Rep.
111	5, 24, 101, 102, 103, 225
112	Rep.
113	101

TABLE SHOWING DISPOSITION OF ALL SECTIONS OF FORMER TITLE 13—Continued

Title 13 Former Sections	*Title 13 New Sections*
215	6
216	5, 23, 146
217	Rep.
218	8
219, 220	Rep.
251	5, 161, 163
252	6, 9, 24, 162, 211, 212, 213, 214
253	Rep.

AMENDMENTS

1990—Pub. L. 101–533, §5(b)(1), Nov. 7, 1990, 104 Stat. 2348, added item for chapter 10.

1962—Pub. L. 87–826, §1, Oct. 15, 1962, 76 Stat. 951, added item for chapter 9.

POSITIVE LAW; CITATION

This title has been made positive law by section 1 of act Aug. 31, 1954, ch. 1158, 68 Stat. 1012, which provided in part "That title 13 of the United States Code, entitled 'Census' is revised, codified and enacted into law and may be cited as 'Title 13, United States Code, section—'."

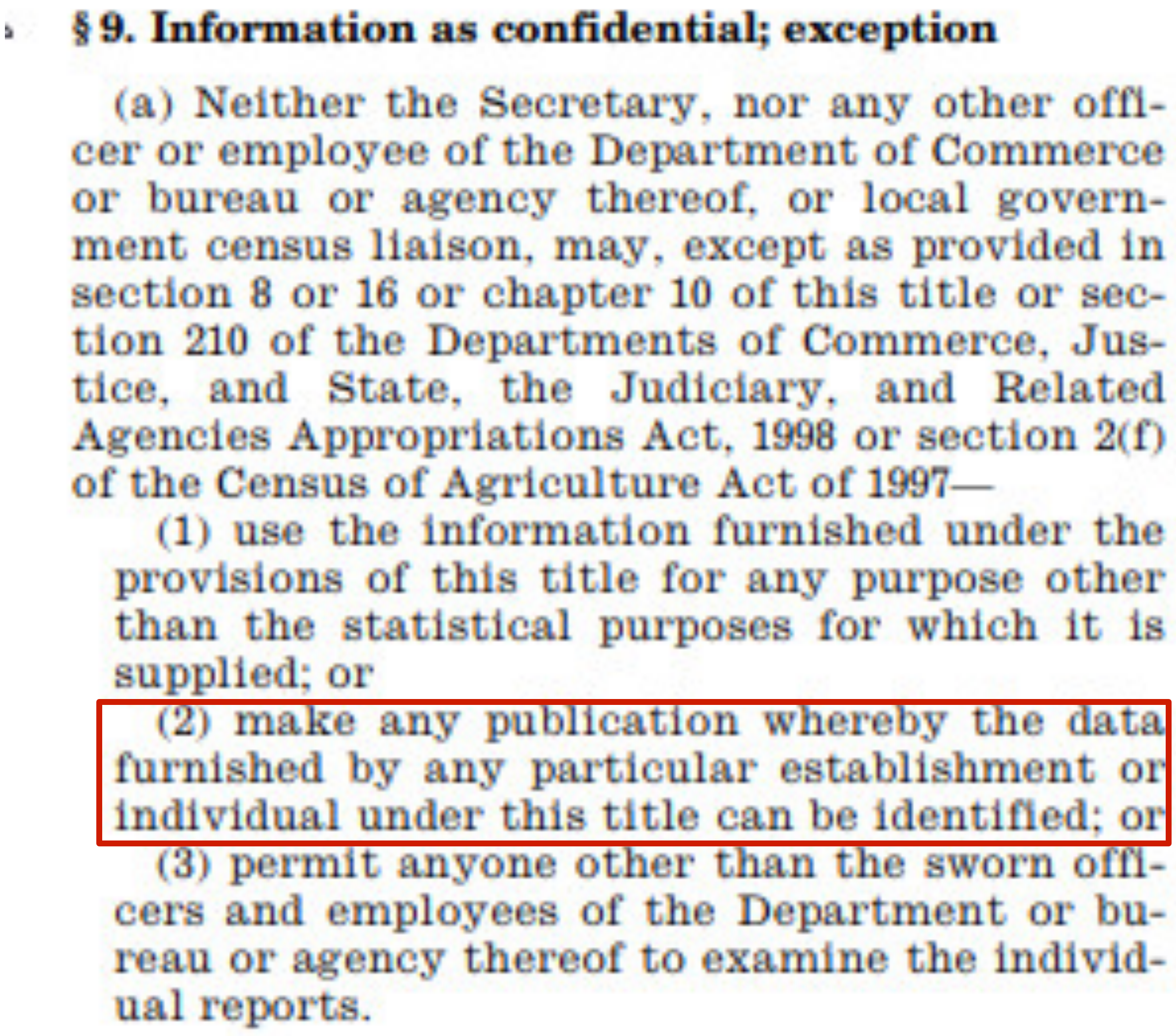
REFERENCES TO CENSUS OFFICE

Section 3 of act Aug. 31, 1954, ch. 1158, 68 Stat. 1024, provided that: "Whenever reference is made in any other law or in any regulation or order to the Census Office, such reference shall be held and considered to mean the Bureau of the Census referred to in section 2 of Title 13, United States Code, as set out in section 1 of this Act. This section shall not be construed as affecting historical references to the Census Office which could have no present or future application to the Bureau of the Census."

SEPARABILITY

Section 4 of act Aug. 31, 1954, ch. 1158, 68 Stat. 1024, provided that: "If any part of Title 13, United States

Title 13.9: produce no publication where an individual can be identified



§ 9. Information as confidential; exception

(a) Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, or local government census liaison, may, except as provided in section 8 or 16 or chapter 10 of this title or section 210 of the Departments of Commerce, Justice, and State, the Judiciary, and Related Agencies Appropriations Act, 1998 or section 2(f) of the Census of Agriculture Act of 1997—

(1) use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or

(2) make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or

(3) permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports.

Why we need disclosure techniques and why differential privacy

- Title 13 requires non-disclosure
  - Reconstruction of data cannot be possible
  - Rules cannot be released
- Fundamental Law of Information Recovery
- Commission on Evidence-based policymaking recommendations
- NAS report, Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy, recommendations
- “Big data” companies use DP

UNIVERSITY OF MICHIGAN | PSID

Fundamental Law of Information Recovery

Fundamental Law of Information Recovery states that overly accurate estimates of too many statistics can completely destroy privacy

Need to use state-of-the-art techniques (Commission on Evidence-Based Policymaking)

- RECOMMENDATION 3-2: The President should direct Federal departments, in coordination with the National Secure Data Service, **to adopt state-of-the-art database, cryptography, privacy-preserving**, and privacy-enhancing technologies for confidential data used for evidence building.

Work with academia and improve methods (Groves and Harris-Kojetin, NAS report)

Volume 1, 5-1: Statistical agencies should engage in collaborative **research with academia and industry to continuously develop new techniques** to address potential breaches of the confidentiality of their data

Volume 1, 5-2: Federal statistical agencies should adopt **modern database, cryptography, privacy preserving**, and privacy-enhancing technologies

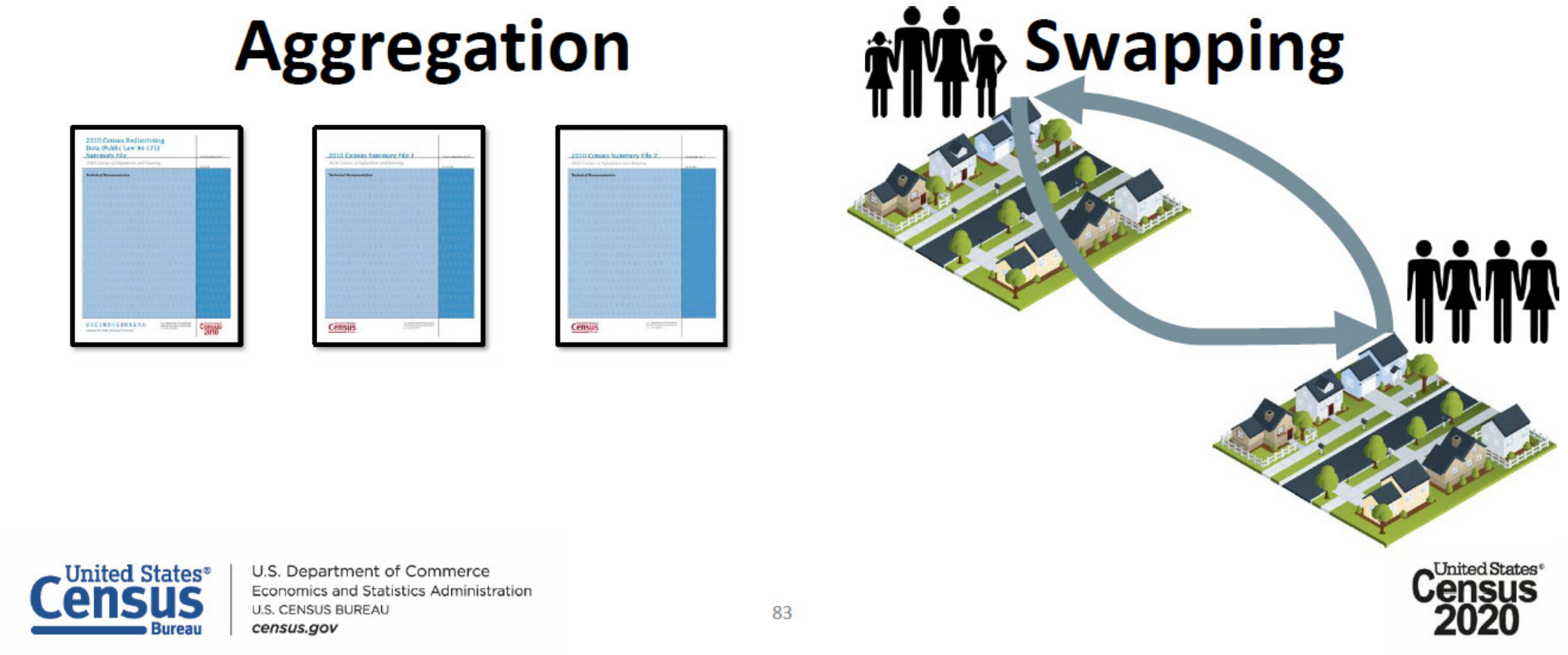
What is differential privacy

- Differential privacy describes a promise, made by a data holder, or curator, to a data subject:
- “You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available.”
- ε-differential privacy says probability changes only by ε.

Current Disclosure Techniques

- only release information at high levels of aggregation;
- swap data, in which variables from pairs of different records are interchanged;
- top- or bottom-code data, e.g., reporting an income of $150,000 for all respondents with incomes higher than that figure;
- add noise to individuals' responses or to computed statistics;
- create synthetic data, in which models fit to the real data are used to generate artificial data that are then released.

Two Privacy Mechanisms for the 2010 Census



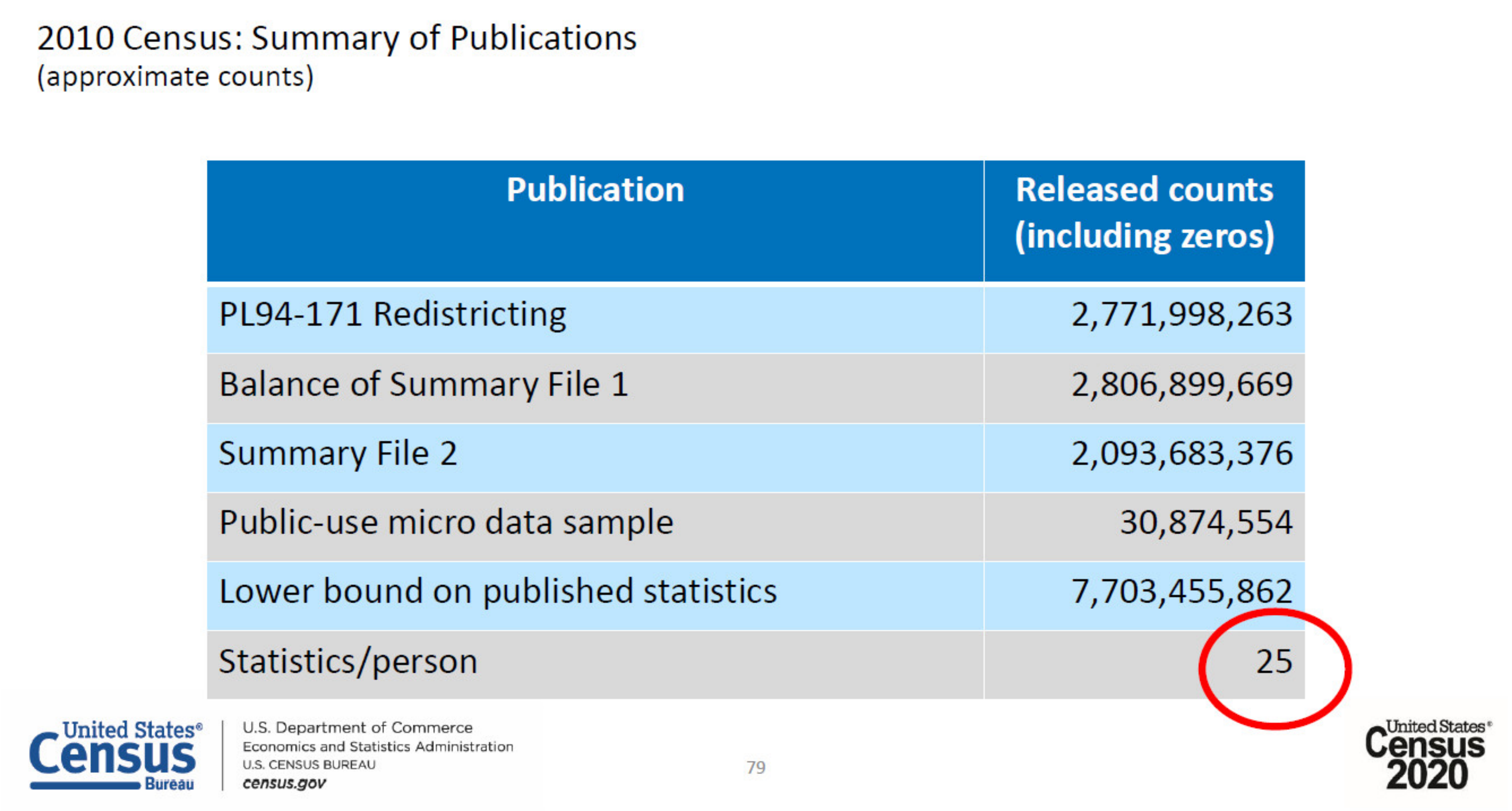
Aggregation

Swapping

2010 Census Redistricting Data (Public Law 94-171) Summary File
2010 Census Summary File 1
2010 Census Summary File 2
United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov
83
United States Census 2020

Source: Garfinkel and Abowd, Census Program Management Review presentation, October 2018

Reconstruction is feasible: 25 published estimates for every person in 2010 Census



2010 Census: Summary of Publications
(approximate counts)

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 data sample	30,874,554
Lower bound on published statistics	7,703,455,862
Statistics/person	25

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

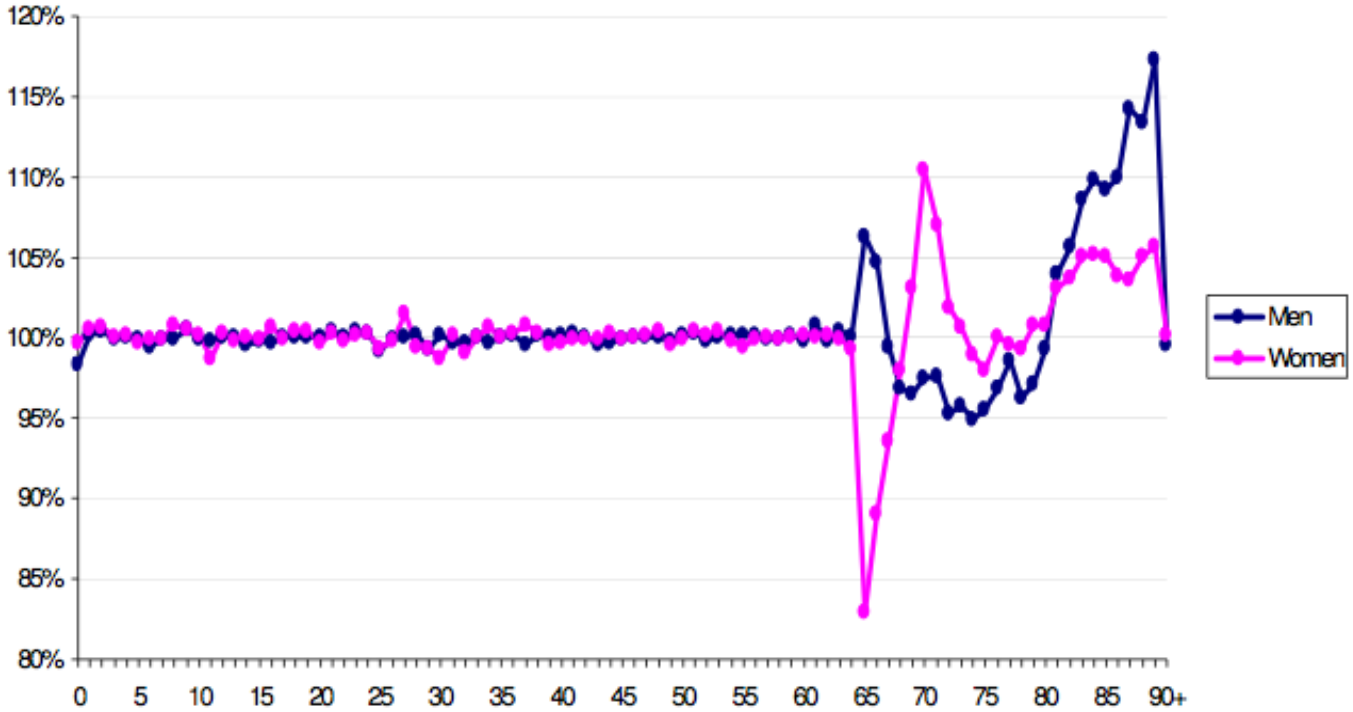
79

Source: Garfinkel and Abowd, Census Program Management Review presentation, October 2018

Example of Disclosure gone wrong

Alexander, Davern, Stevenson (Public Opinion Quarterly 2010)

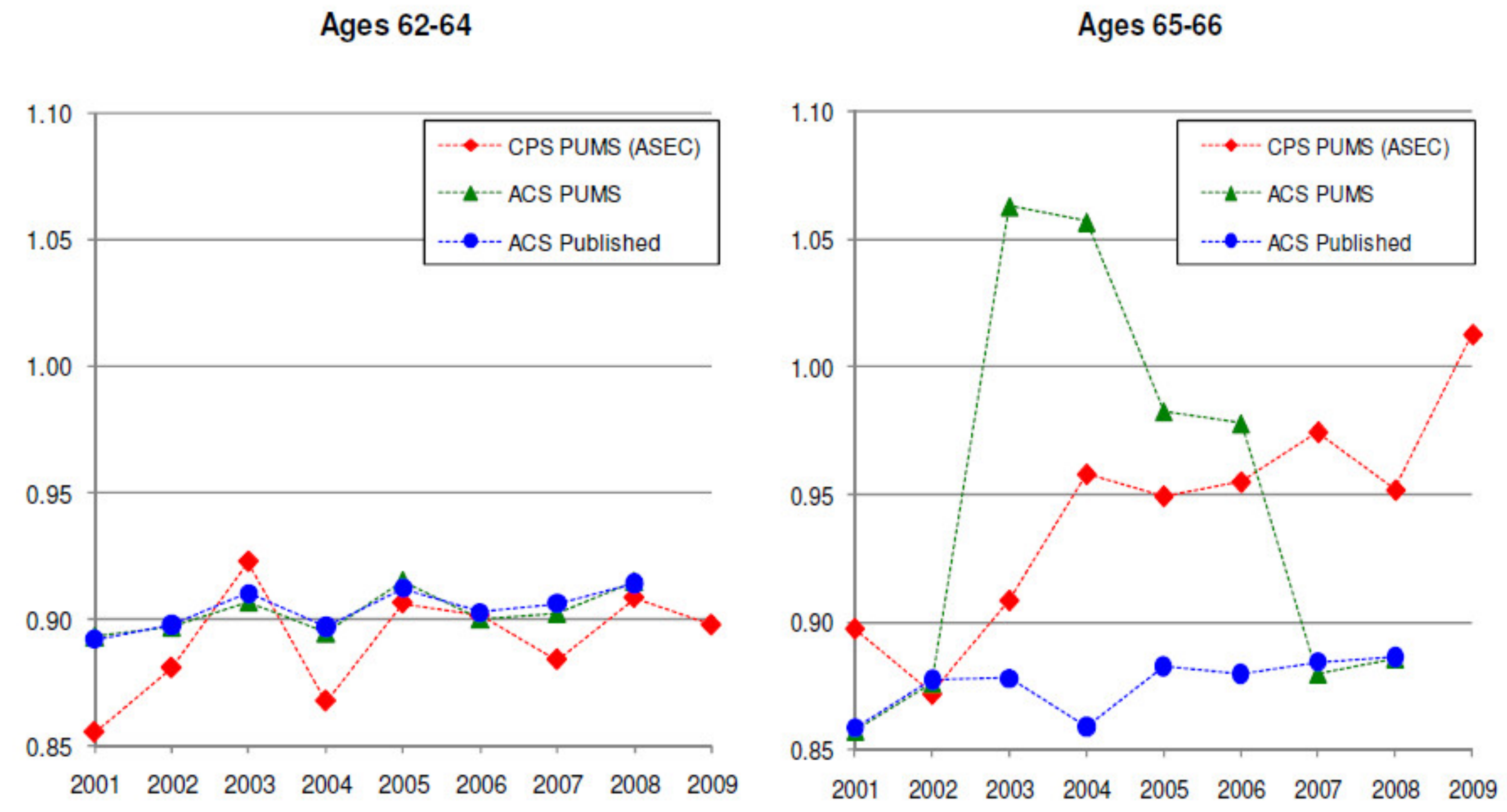
Figure 1. Population estimates from 2000 5% Census PUMS as a percentage of published data



Sources: Census 2000 Summary File 4, Table PCT3 (http://factfinder.census.gov); Census 2000 5% sample, IPUMS-USA (http://usa.ipums.org/).

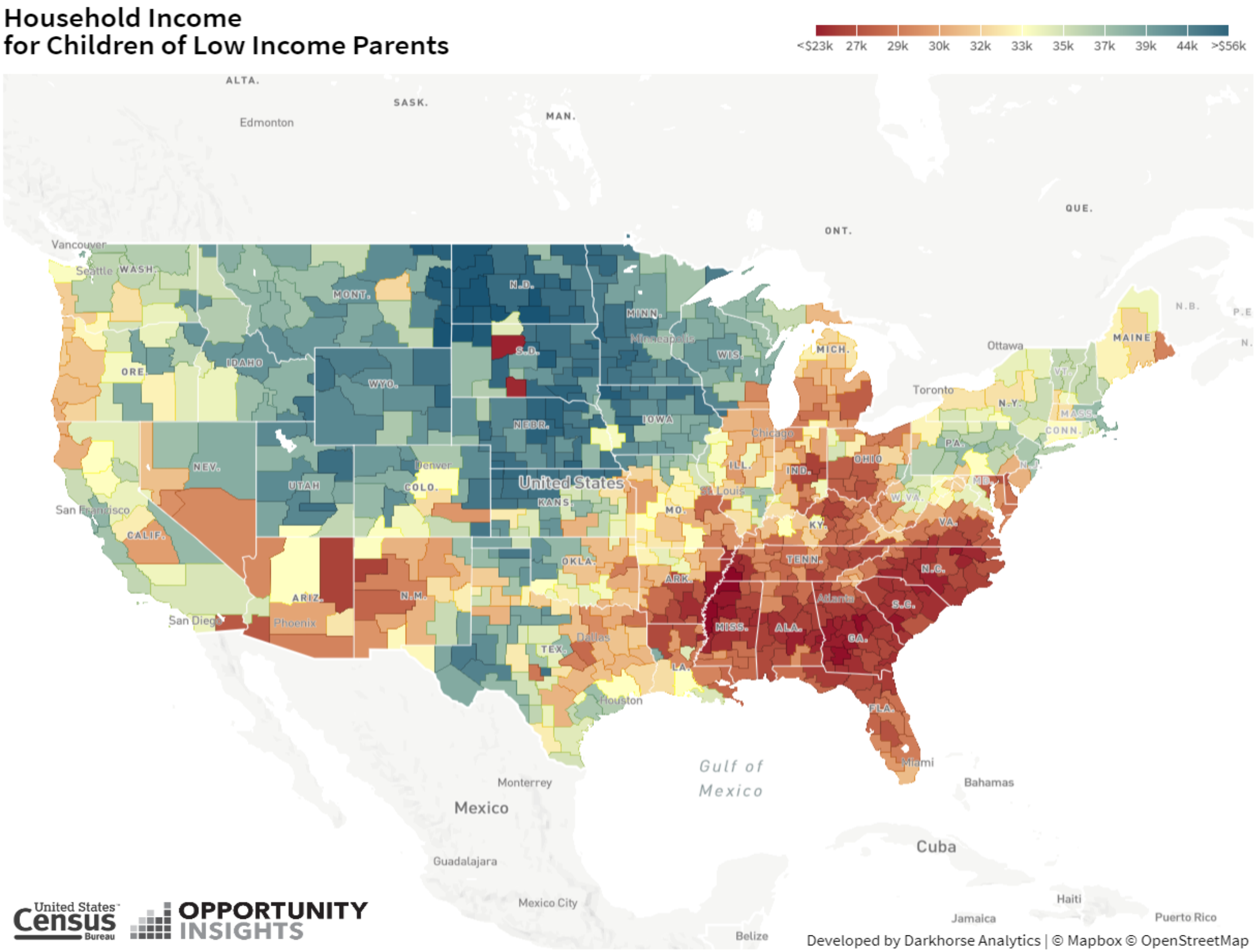
And even worse in ACS and CPS (CPS elderly poverty rates had to be revised)

Figure 4. Ratio of men to women in CPS PUMS, ACS PUMS, and published data



Source: Alexander, Davern, Stevenson, Public Opinon Quarterly Fall 2010.

Example of DP success: Opportunity Atlas Chetty, Hendren, Jones, Porter (NBER 2018)

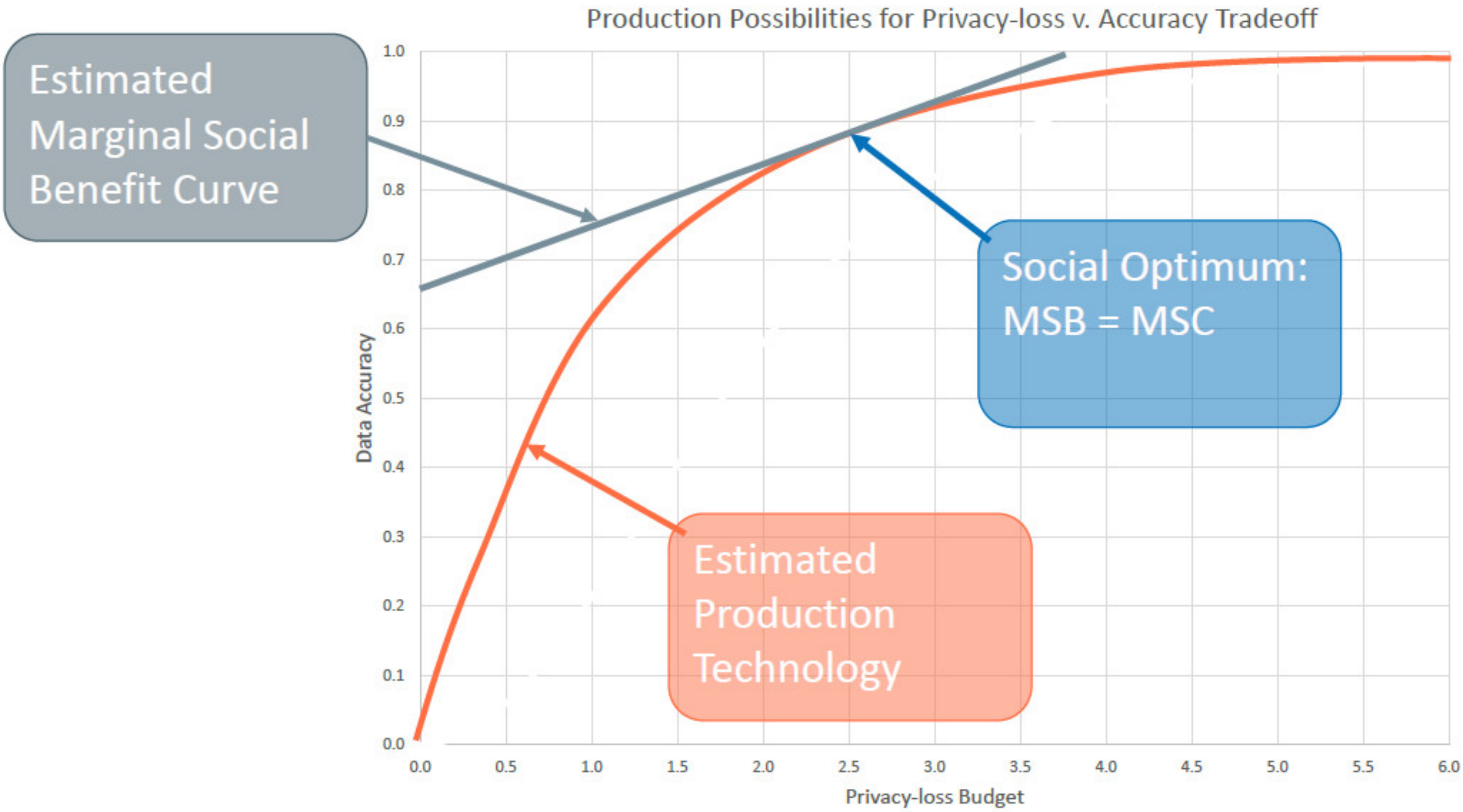


Household Income
for Children of Low Income Parents

Developed by Darkhorse Analytics | © Mapbox © OpenStreetMap

https://www.opportunityatlas.org/

Measuring accuracy, privacy loss and finding ε



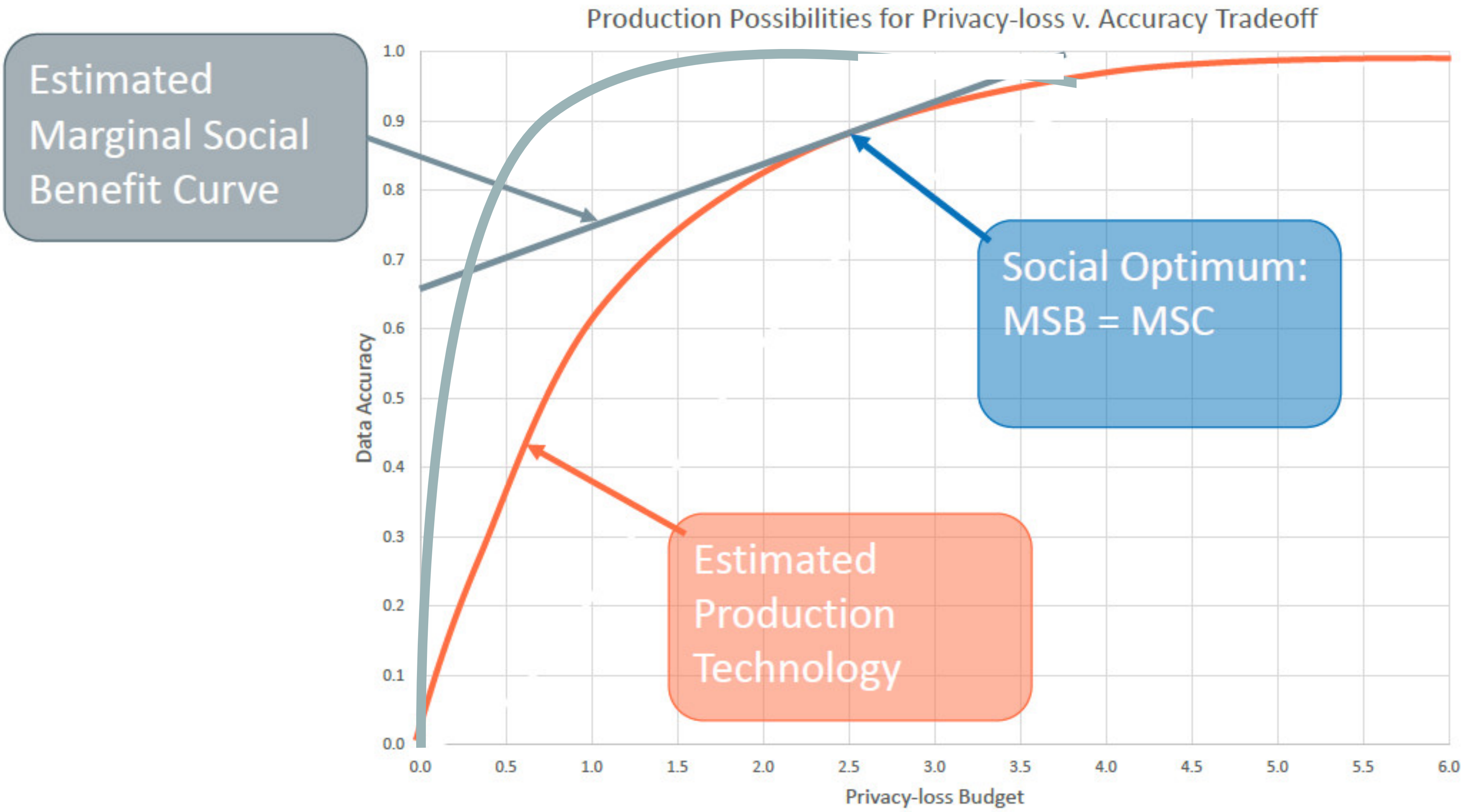
Estimated Marginal Social Benefit Curve

Social Optimum: MSB = MSC

Estimated Production Technology

Source: Abowd, Census Scientific Advisory Committee Presentation, Sept 2018

Different projects may have different trade-offs



Estimated Marginal Social Benefit Curve

Social Optimum: MSB = MSC

Estimated Production Technology

Source: Abowd, Census Scientific Advisory Committee Presentation, Sept 2018

Differential privacy is not a panacea in many ways

- neither differential privacy nor any other technique can circumvent the fundamental law of information recovery
- there is no difference between multiple releases of synthetic datasets and interactive query systems
- differential privacy limits interaction with the dataset, not allowing the analyst to see the raw data
- differential privacy also hides outliers.
- differential privacy may require larger sample sizes in order to learn things about the full population
- differential privacy intentionally introduces statistical noise.

Source: Groves and Harris-Kojetin, “Federal Statistics, Multiple Data Sources and Privacy Protection” NAS Report, 2017

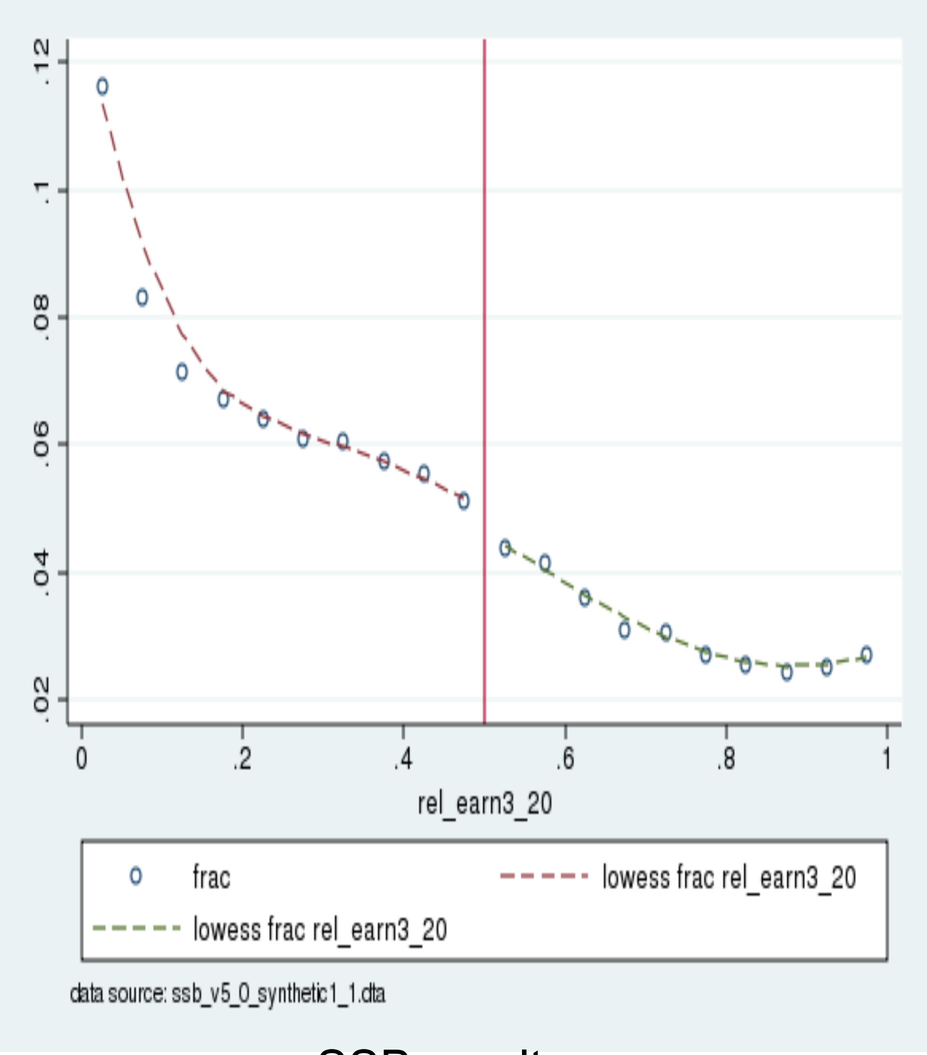
Issue of a Privacy Budget

- Set total privacy-loss budget: **ε**
- Ensure that **ε**1+**ε**2+**ε**3+**ε**4+**ε**5+**ε**A = **ε,**
  - **ε**1 National, **ε**2 state, **ε**3 county, **ε**4 tract, **ε**5 block, **ε**A microdata
- Within each stage, allocate privacy-loss budget between: PL-94, Parts of SF-1 not in PL-94
- Key is how is privacy budget chosen
- How does this budget affect other data sets

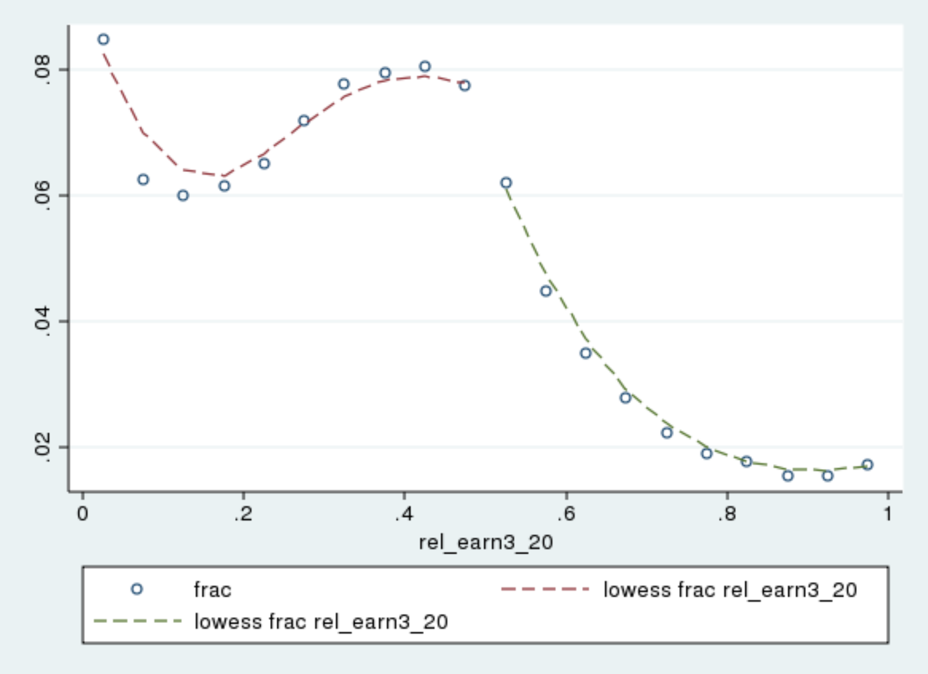
Differential Privacy via Synthetic Data

- OnTheMap
  - https://onthemap.ces.census.gov
- Longitudinal Business Database
  - https://www.census.gov/ces/dataproducts/datasets/lbd.html
- SAIPE – Small Area Income and Poverty Estimates
  - https://www.census.gov/programs-surveys/saipe.html
- SIPP Synthetic Beta File
  - https://www2.vrdc.cornell.edu/news/data/sipp-synthetic-beta-file/

Example: "The distribution of the share of household income earned by the wife exhibits a sharp cliff at 0.5…" - Bertrand, Kamenca, Pan *(QJE 2015)*



SSB results



Internal results

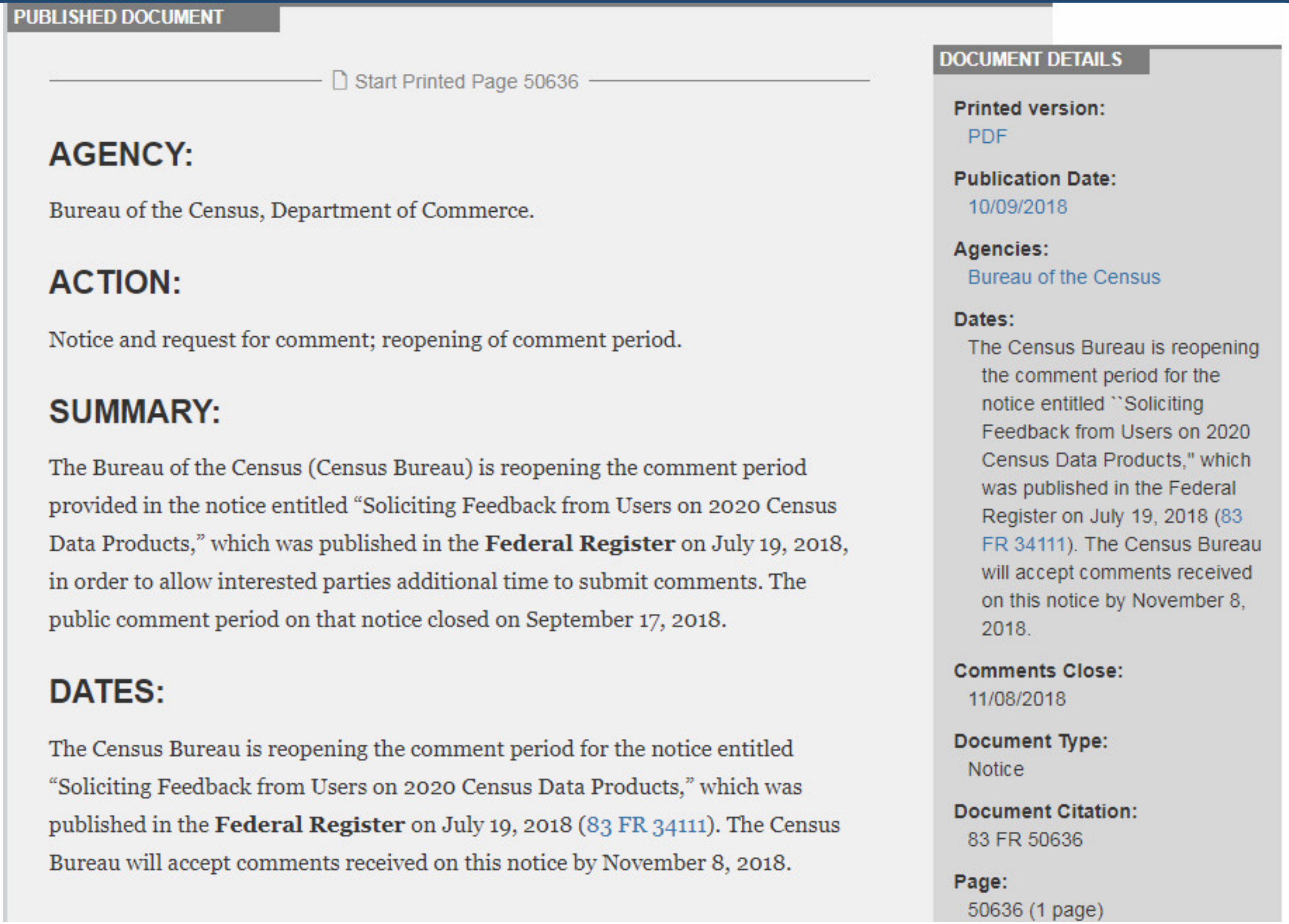
Possible Fundamental Theorem of Synthetic Data

For a finite set of models (or research questions) and a micro data set, there exists a synthetic data set that yields similar results as the original data set

What can we do?

- Respond to all Federal Register Notices
- Participate in Advisory Meetings and other user groups (COPAFS, APDU)
- Don't Complain; Provide Suggestions
- Talk to Census staff – and do it early and often
- Be specific about the effects on accuracy
- Train Census staff

Answer the FR Notice



PUBLISHED DOCUMENT

Start Printed Page 50636

AGENCY:

Bureau of the Census, Department of Commerce.

ACTION:

Notice and request for comment; reopening of comment period.

SUMMARY:

The Bureau of the Census (Census Bureau) is reopening the comment period provided in the notice entitled “Soliciting Feedback from Users on 2020 Census Data Products,” which was published in the **Federal Register** on July 19, 2018, in order to allow interested parties additional time to submit comments. The public comment period on that notice closed on September 17, 2018.

DATES:

The Census Bureau is reopening the comment period for the notice entitled “Soliciting Feedback from Users on 2020 Census Data Products,” which was published in the **Federal Register** on July 19, 2018 (83 FR 34111). The Census Bureau will accept comments received on this notice by November 8, 2018.

DOCUMENT DETAILS

Printed version:
PDF

Publication Date:
10/09/2018

Agencies:
Bureau of the Census

Dates:
The Census Bureau is reopening the comment period for the notice entitled ``Soliciting Feedback from Users on 2020 Census Data Products," which was published in the Federal Register on July 19, 2018 (83 FR 34111). The Census Bureau will accept comments received on this notice by November 8, 2018.

Comments Close:
11/08/2018

Document Type:
Notice

Document Citation:
83 FR 50636

Page:
50636 (1 page)

https://www.federalregister.gov/documents/2018/10/09/2018-21837/soliciting-feedback-from-users-on-2020-census-data-products-reopening-of-comment-period

What can we do?

- Respond to all Federal Register Notices
- Participate in Advisory Meetings and other user groups (COPAFS, APDU)
- Don’t Complain; Provide Suggestions
- Talk to Census staff – and do it early and often
- Be specific about the effects on accuracy
- Train Census staff

We all need training in new methods

Final, 5-1: Federal statistical agencies should ensure their technical staff receive appropriate training in modern computer science technology including but not limited to database, cryptography, privacy-preserving, and privacy enhancing technologies

Groves and Harris-Kojetin, “Federal Statistics, Multiple Data Sources and Privacy Protection” NAS Report, 2017