Issues Encountered Deploying Differential Privacy

Simson L. Garfinkel
Senior Scientist, Confidentiality and Data Access
simson.l.garfinkel@census.gov

John M. Abowd
Chief Scientist and Associate Director for Research and Methodology

Sarah Powazek
MIT

Workshop on Privacy in the Electronic Society
Toronto, Canada — October 15, 2018

Note: talk is 12 minutes with questions. Most slides are backups.

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

Abstract

When differential privacy was created more than a decade ago, the motivating example was statistics published by an official statistics agency. In attempting to transition differential privacy from the academy to practice, and in particular for the 2020 Census of Population and Housing, the U.S. Census Bureau has encountered many challenges unanticipated by differential privacy's creators. These challenges include obtaining qualified personnel and a suitable computing environment, great difficulty accounting for all uses of the confidential data, the lack of release mechanisms that align with the needs of data users, the expectation on the part of data users that they will have access to micro-data, difficulty in setting the value of the privacy-loss parameter, ε (epsilon), and the lack of tools and trained individuals to verify the correctness of differential privacy implementations.

Acknowledgments

This presentation incorporates work by:

- Dan Kifer (Scientific Lead)
- John Abowd (Chief Scientist)
- Tammy Adams, Robert Ashmead, Aref Dajani, Jason Devine, Michael Hay, Cynthia Hollingsworth, Meriton Ibrahimi, Michael Ikeda, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Brett Moran, Ned Porter, Anne Ross and William Sexton

Motivation
Article 1, Section 2:

"…The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct…"

The 2020 Census of Population and Housing

**Count everyone once,
only once, and in the right place.**

United States™ Census Bureau

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

In 2017, the Census Bureau announced that it would use differential privacy for the 2020 Census

There is no off-the-shelf mechanism for applying differential privacy to a national census

Randomized response would introduce far too much noise for any sensible value of ε to be a much statistical value

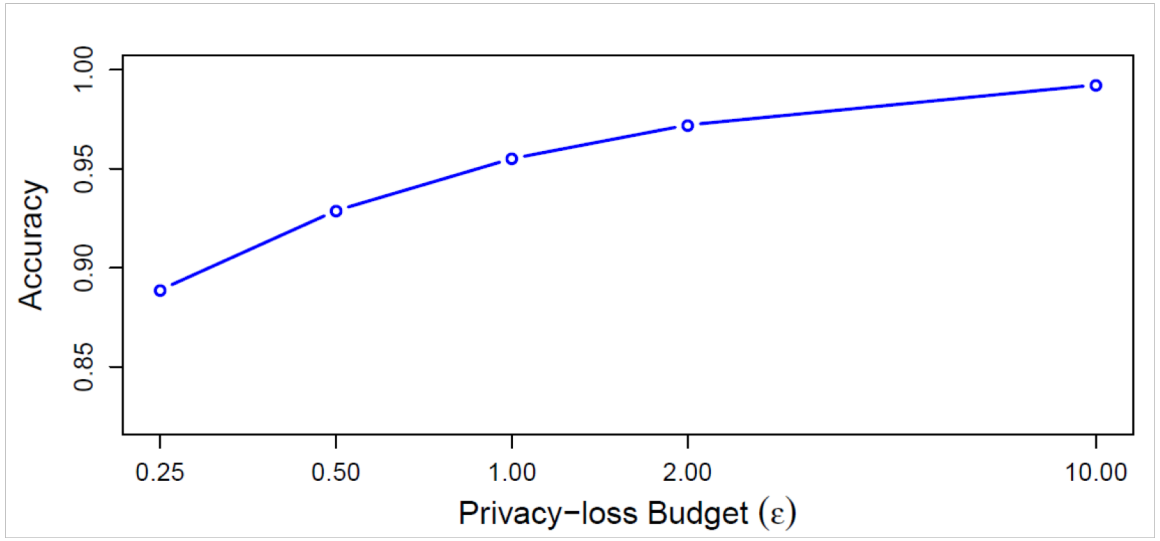
We cannot simply apply the Laplace mechanism to tables:

- Our data users expect consistent tables
- Lack of parallel composition between tables (and sometimes within them)

Differential privacy allows us to control accuracy vs. privacy loss

Accuracy concerns — “Fitness for use”

- Fit for voting rights act (VRA) section 2 determinations
- Fit for Census Bureau Population Estimates program
- Fit for distributing $670 million in federal funds



Privacy-loss concerns:

- Meet requirements of Title 13 §9

Final privacy-loss budget determined by
Data Stewardship Executive Policy Committee (DSEP)
with recommendation from Disclosure Review Board (DRB)

The 2020 Disclosure Avoidance System uses Differential Privacy to enforce global confidentiality protections

Decennial Response File → **Census Unedited File** → **Census Edited File** → Global Confidentiality Protection Process **Disclosure Avoidance System** → Microdata Detail File → Pre-specified tabular summaries: **PL94-171, SF1, SF2**

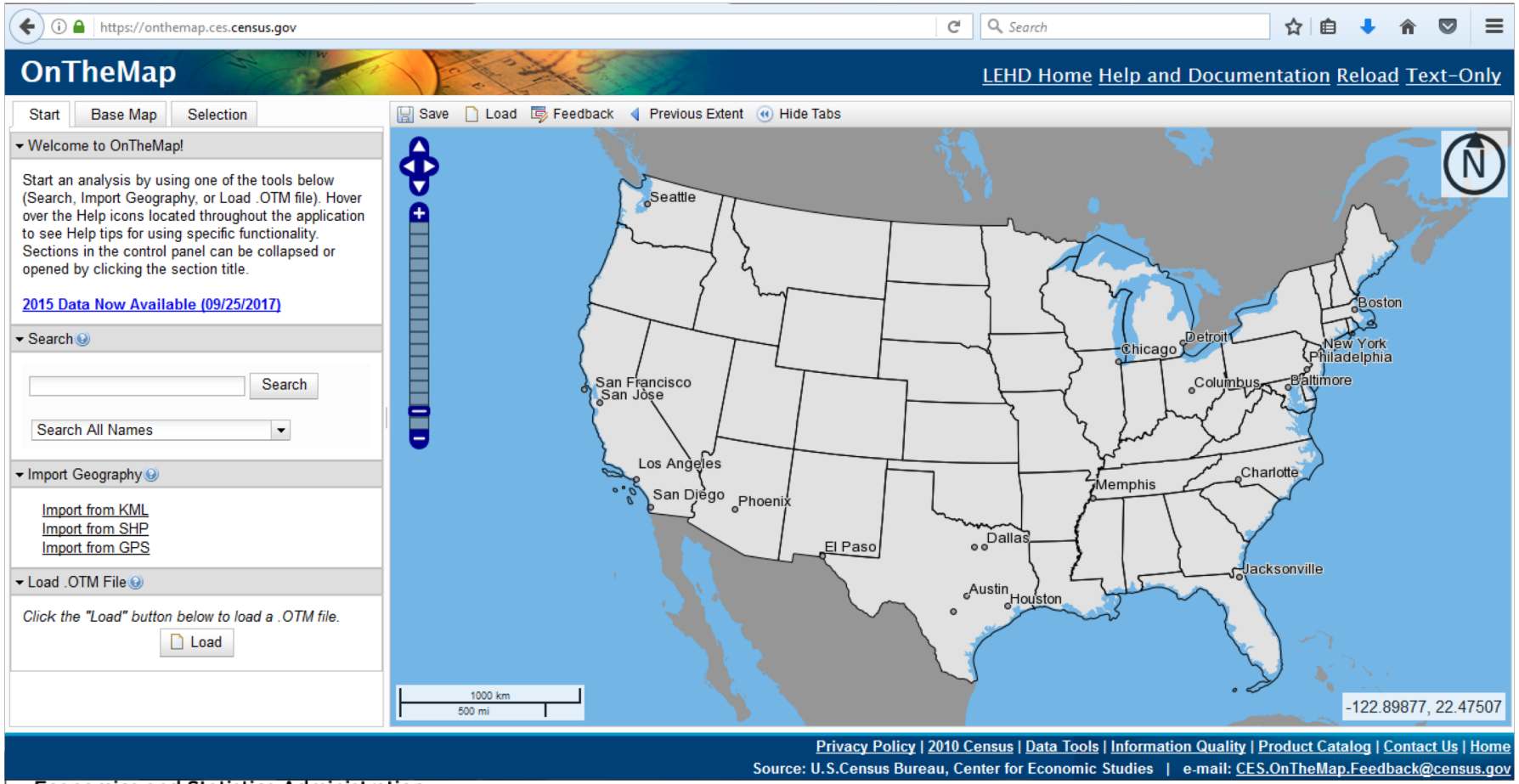
Microdata Detail File → Special tabulations. Post-census research

≈330M people

ε

Privacy-loss budget, Accuracy Decisions

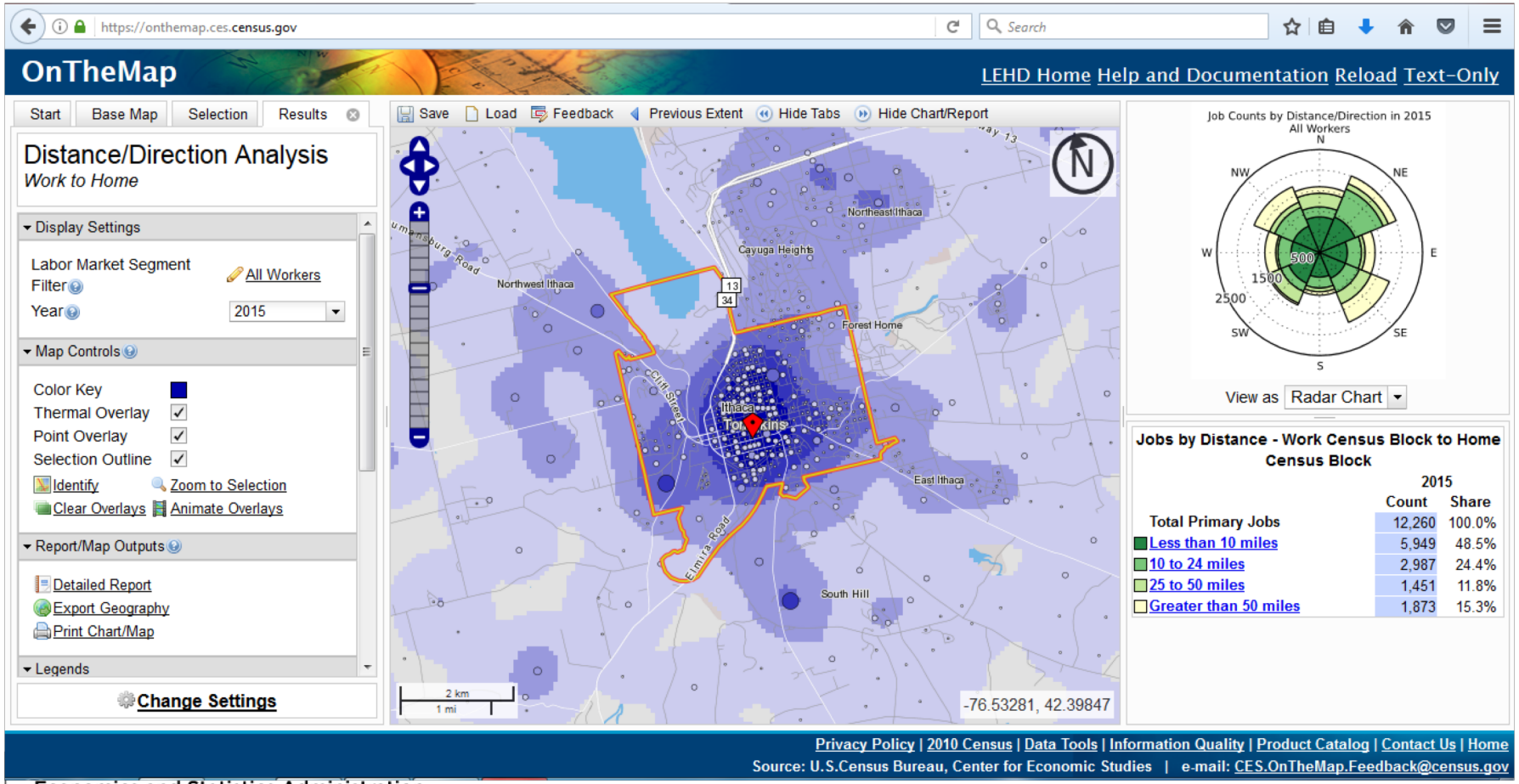
2008: First production implementation of differential privacy in the world



United States Census Bureau

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

OnTheMap lets users learn population, earnings, and commuting patterns for arbitrary areas



United States Census Bureau

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

(2018 screenshot)

Our experience with OnTheMap did not prepare the organization for the challenge

OnTheMap was a new product, designed from the start to be DP on the residential side. Haney et al. (2017) extends to the employment side.

The decennial Census of Population and Housing, first performed under the direction of Thomas Jefferson in 1790, is the oldest and most expensive statistical undertaking of the U.S. government.

Transitioning existing data products has revealed:

- The limits of today's formal privacy mechanisms
- The difficulty of retrofitting legacy statistical products to conform with modern privacy practice

Scientific Issues for the 2020 Census

Hierarchical Mechanisms

We needed a novel mechanism that:

Assured consistent statistics from US->States->Counties->Tracts->Blocks

Provided lower error for geographies with greater populations

Invariants

For the 2018 End-to-End test, policy makers wanted exact counts for:

Number of people on each block

Number of people on each block of voting age

Number of residences and group quarters on each block

These may, however be removed based on what we have learned to-date

Scientific issues for the American Community Survey (and many others!)

ACS replaced the decennial census long form in 2005

ACS uses a *stratified probability sample* and *skip logic/branching*

Current differential privacy algorithms do not handle these features well

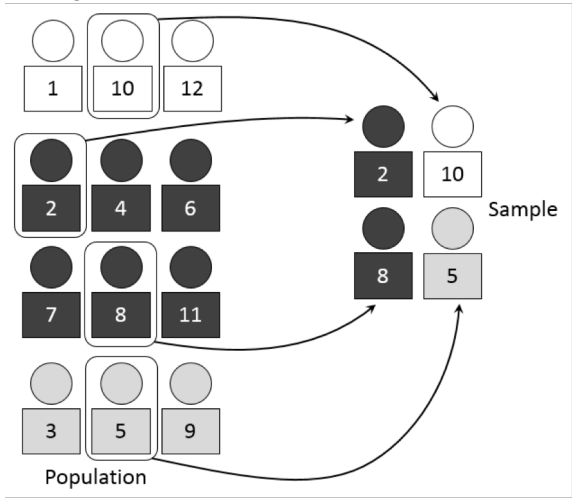


Diagram illustrating stratified sampling, Wikipedia

4 **How many acres is this house or mobile home on?**

- ☐ Less than 1 acre ➔ *SKIP to question 6a*
- ☐ 1 to 9.9 acres
- ☐ 10 or more acres

Branching instruction in the 2018 ACS

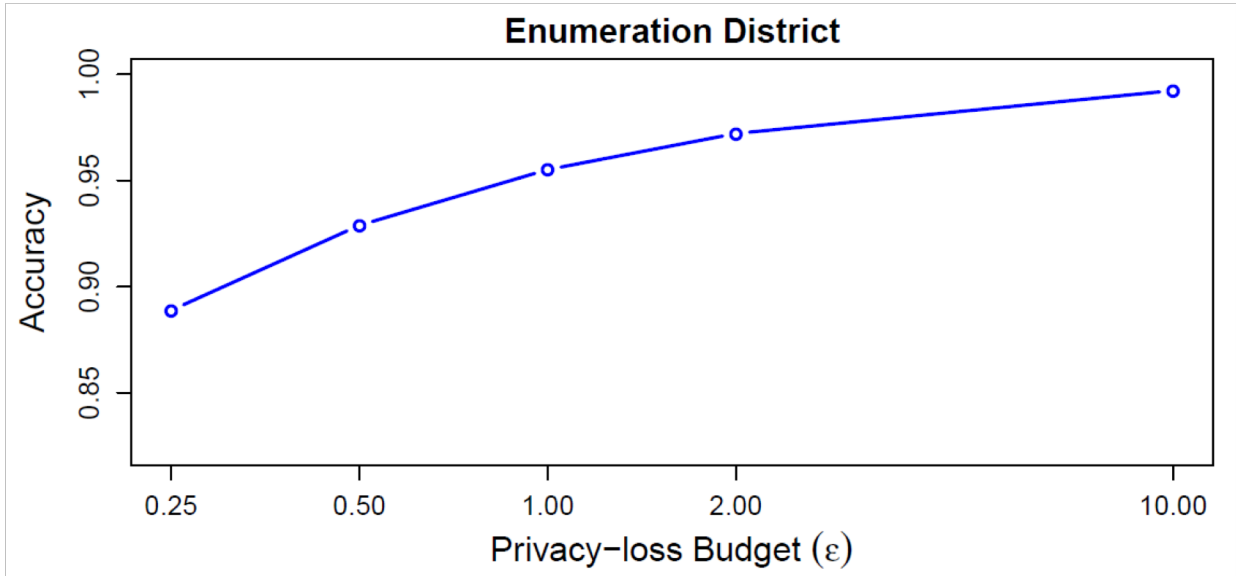
United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Scientific Issues: Quality Metrics

What is the measure of “quality” or “utility” in a complex data product?

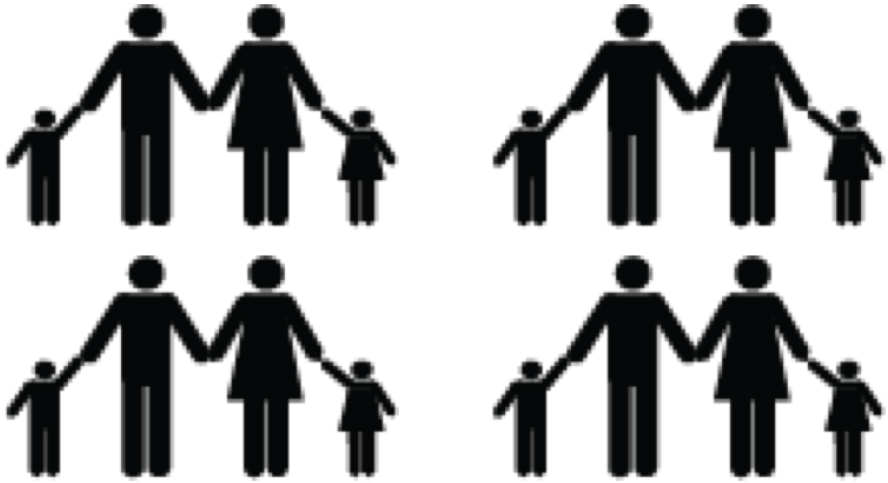
Options:

- L1 error between “true” data set and “privatized” data set
- Impact on an algorithm that uses the data (e.g. voting rights enforcement)



Scientific issues: Accuracy trade-offs

As collected:



As Reported

	Male	Female
Age < 18	4	4
Age >= 18	4	4

Version 1: High privacy loss

	Male	Female
Age < 18	3	2
Age >= 18	5	6

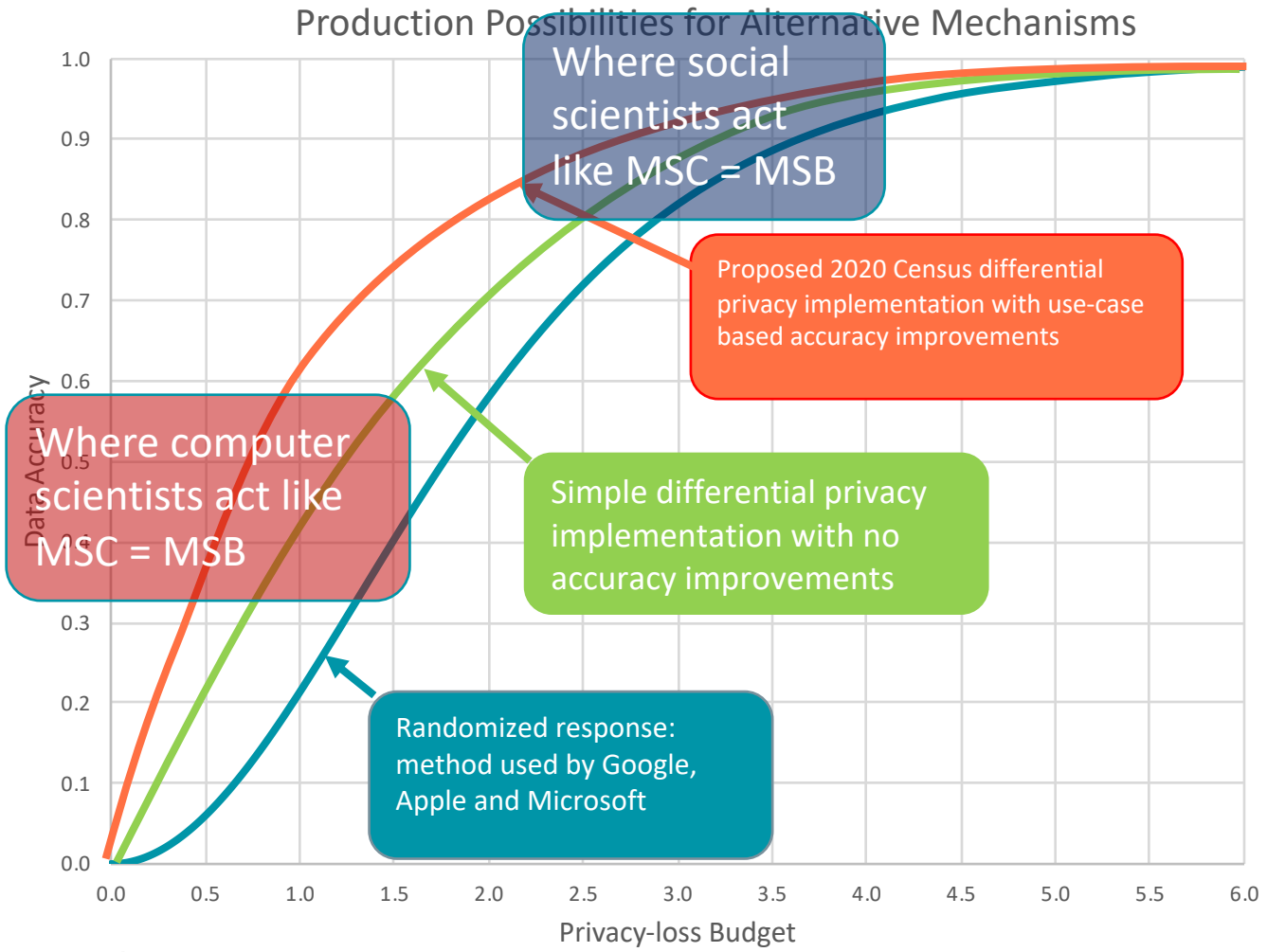
Version 2: More accurate sex distribution

	Male	Female
Age < 18	6	2
Age >= 18	3	5

Version 3: More accurate age distribution

Scientific Issue: Setting Epsilon

Where does the Marginal Social Cost (MSC) equal the Marginal Social Benefit (MSC)?



Computer scientists have spent their efforts improving the accuracy/privacy-loss trade-off….

…not in developing tools for helping policy makers to set epsilon

Operational Issues

Obtaining qualified personnel and tools

Recasting high-sensitivity queries

Identifying structural zeros

Obtaining a suitable computing environment

Accounting for all uses of confidential data

Issues Faced by Data Users

Access to micro-data

- Many users expect access to micro-data (especially internal users)

Difficulties arising from increased transparency

- Many users were not aware of prior disclosure avoidance practices
- The swap rate was never made public

Misunderstandings about randomness and noise injection

Recommendations

Repeated discussions with decision makers

Controlled vocabulary

Integrated communications and outreach strategy

Questions? Looking for an internship? Feel free to email me at: simson.l.garfinkel@census.gov