Decisions with Privacy-Protected Data

Ian M. Schmutte
Department of Economics
Terry College of Business
University of Georgia

NBER Methods Lecture
July 17, 2020

Outline

This Talk

- Trading off privacy and data accuracy
- Learning from privacy protected data
- Accuracy as improved decision making

Next Talk

- Implementing formal privacy in Census data

Dual Mandate

accurate statistical summaries

privacy guarantee

Choice of ε

Each calculation based on the data consumes the privacy budget

$$\epsilon_1 + \epsilon_2 + \epsilon_3 + \cdots \epsilon_n = \epsilon_{Total}$$

2018 Decennial Census End to End Test:	$\epsilon = .25$
2010 Demonstration Data Products	$\epsilon = 4 + 2$
On The Map	$\epsilon = 8.9$
Opportunity Atlas	$\epsilon = 8.0$

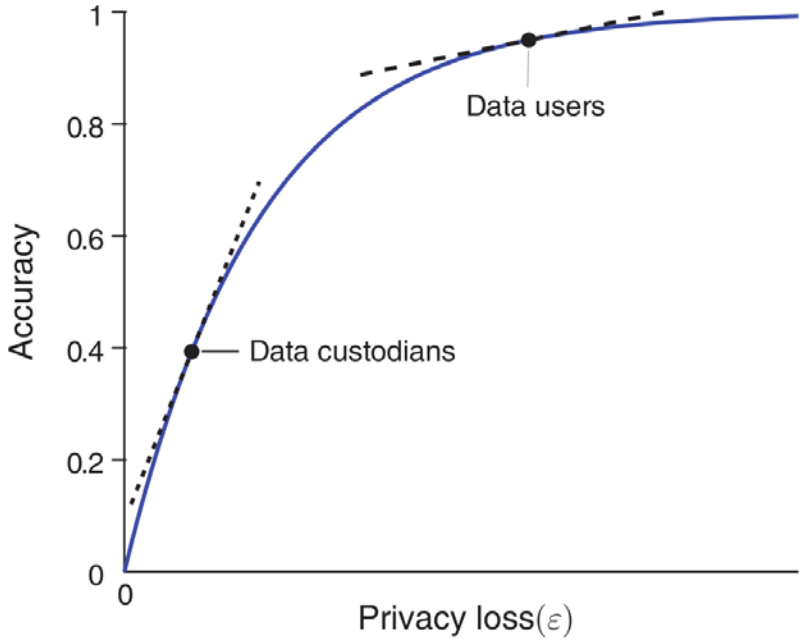
Economic Perspective (Abowd and Schmutte 2019)

1. finite resource: information in an existing database

2. competing uses
   - accuracy
   - privacy

3. an optimal allocation should equate
   - Marginal rate of transformation
   - Willingness to pay

4. accuracy and privacy are public goods



Learning from Privacy Protected Data

General model of privacy protection

Population: N

Complete data matrix:

$$\mathrm{D}, (N \times K)$$

Process parameter: θ_p

Distributions

Data model: $p_D\left(D \middle| \theta_p\right)$

Prior: $p_{\theta_p}(\theta)$

Estimands of interest

Functions of D(finite-population)

Functions of θ_p (super-population)

Ignorable Privacy Protection

- Published data: Z
- Privacy model:

$$p_{Z|D}(Z|D,\theta_M)$$

- Privacy parameter: θ_M, with prior $p_{\theta_M|\theta_p}(\theta|\theta_p)$
- Likelihood for published data

$$L_{\theta}^{pub}(\theta_p,\theta_M) = \int p_{Z|D}(Z|D,\theta_M)p_D(D|\theta_p)dD$$

Inference based on

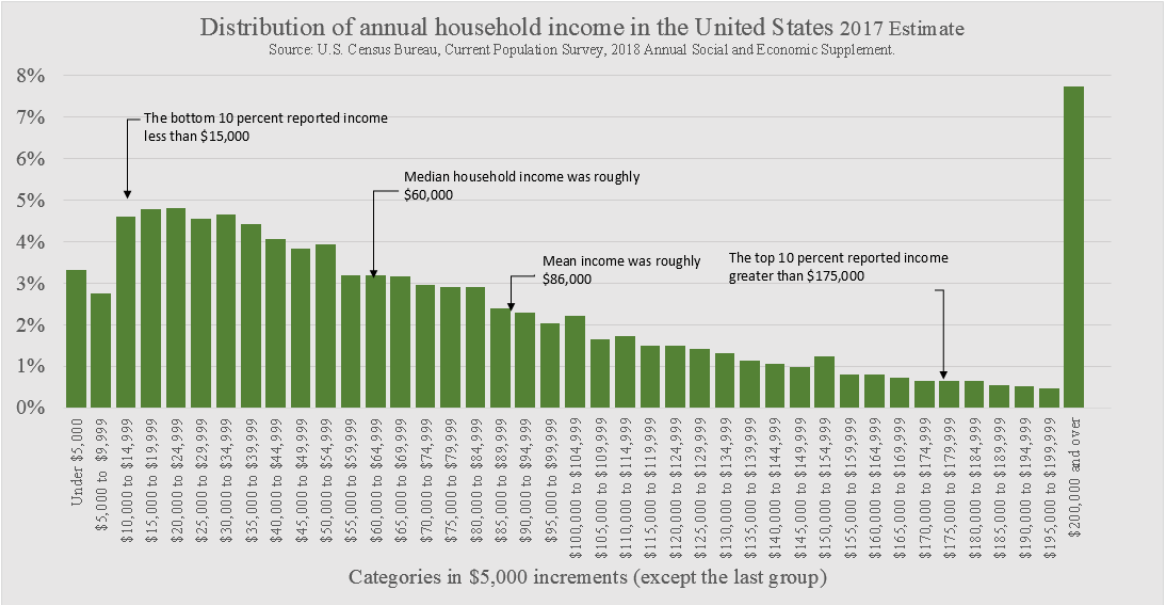
$$p_{\theta_p|Z}(\theta_p|Z) = \int p_{\theta_p|D}(\theta_p|D)p_{D|Z}(D|Z)dD$$

Ignorable privacy protection

$$p_{\theta_p|D}(\theta_p|D = Z) = p_{\theta_p|Z}(\theta_p|Z)$$

- **Is privacy-protection ignorable?**
- **Can privacy-aware analysis be conducted?**
- **Is the privacy model discoverable?**

Topcoding



Ignorable

for inference on quantiles less than quantile of T (e.g. 90-10 ratio in CPS)

Non-ignorable

for quantiles above T

but *privacy-aware analysis* is sometimes possible

Suppress and Impute

- *Non-ignorable*
  - cf. Bollinger and Hirsch (2006)
  - Induces bias
- *No privacy-aware analysis*
  - Unknown model
  - Unknown rate
  - Unknown variables
- *Not discoverable*

14	41	50	58	65
15	24	26	30	25
52	53	66	47	51
68	6	44	17	32
38	26	33	42	64

$$p_{Z|Y}(Z|Y,\theta_M)$$

13	41	51	58	65
15	24	25	30	24
51	54	66	48	51
68	6	44	16	32
38	25	33	42	65

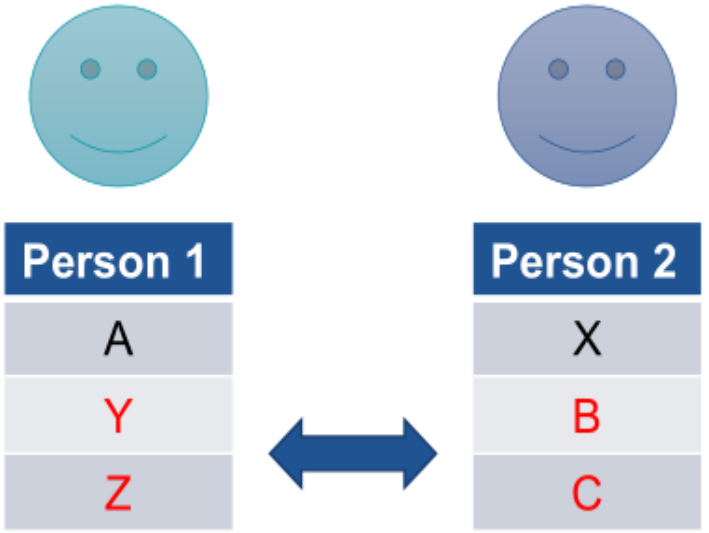
Swapping

High-risk records:

- Matched to a “nearby” record
- .. And swapped

Preserves counts on key characteristics

May prevent disclosure of sensitive attributes



Swapping

Ignorable if..

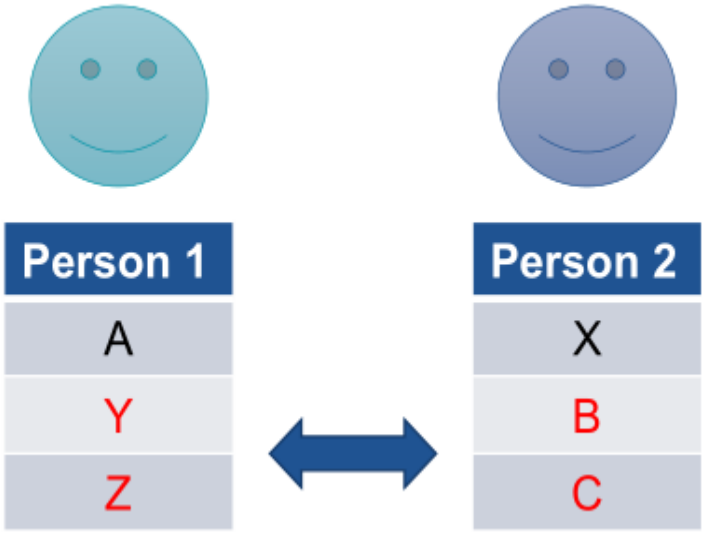
only care about matching variables

Non-ignorable for

covariance between matching and other variables

Parameters are secret

- Swap rate
- Sensitive chars
- Swap domain
- Etc.



these distortions might matter a lot
…but we don't *really* know

INACCURATE AGE AND SEX DATA IN THE CENSUS PUMS FILES: EVIDENCE AND IMPLICATIONS

J. TRENT ALEXANDER
MICHAEL DAVERN
BETSEY STEVENSON*

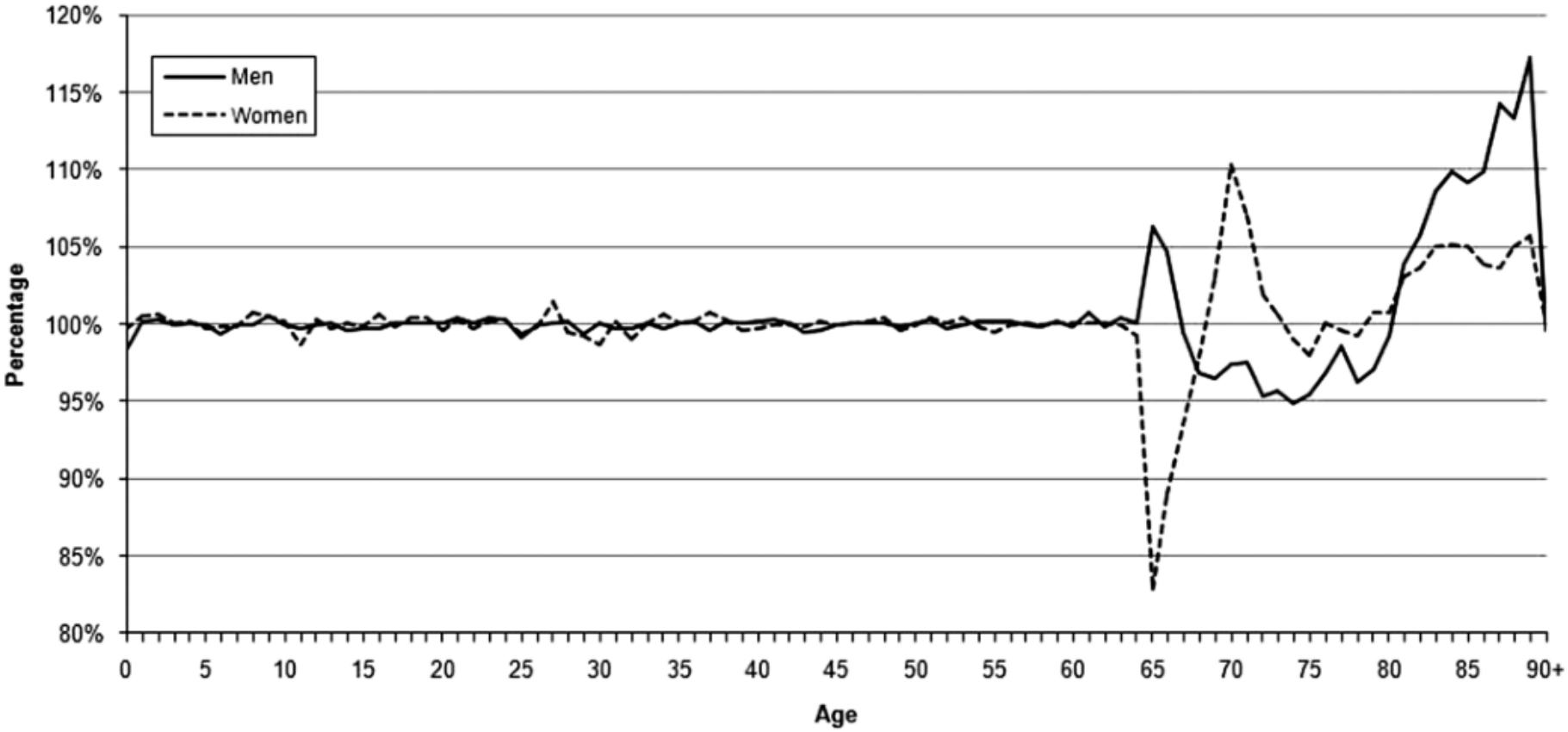


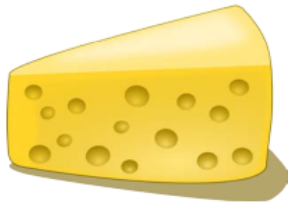
Figure 1. Population estimates from 2000 five-percent Census PUMS as a percentage of Census 2000 published data. Sources: Published population counts are from Census 2000 Summary File 4, Table PCT3 (http://factfinder.census.gov); population estimates are calculated using Census 2000 five-percent sample, IPUMS-USA (http://usa.ipums.org/usa/).

Cell Suppression

A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P

A	B	*	D
E	F	G	H
*	J	K	L
M	N	O	*

*	B	*	*
E	F	G	H
*	J	*	L
M	N	*	*



- "Blank out" cells to protect outliers
  - i.e., where one large firm dominates
- Then "blank out" more cells to prevent subtraction attack
- e.g., Economic Census, County Business Patterns

Cell Suppression

A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P

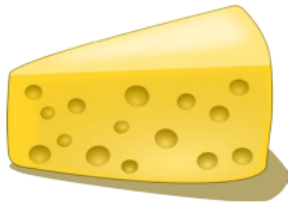
→

A	B	*	D
E	F	G	H
*	J	K	L
M	N	O	*

→

*	B	*	*
E	F	G	H
*	J	*	L
M	N	*	*

=



Not ignorable unless

…suppression was random with respect to your estimand of interest

…or you really only care about the unsuppressed data.

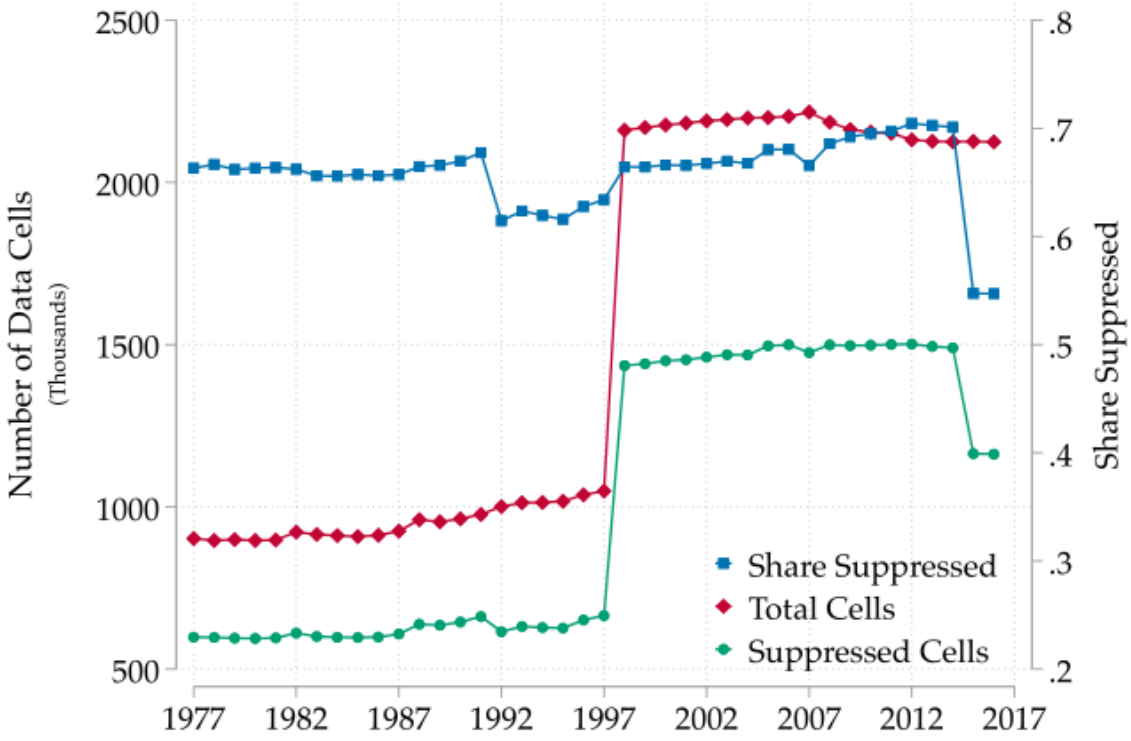
So then what?

Hack the protection!?

Imputing Missing Values in the US Census Bureau's County Business Patterns

Fabian Eckert, Teresa C. Fort, Peter K. Schott, Natalie J. Yang

NBER Working Paper No. 26632
Issued in January 2020



Source: 1977 to 2017 CBP files and authors' calculations. Figure displays the number of cells in the county files in each year, the number of those cells that are suppressed, and the share of cells that are suppressed. Suppressed cell counts do not include the addition of noise infusion to all cells starting in 2007. Industry classification switches from SIC to NAICS in 1998.

Suppression is not ignorable

A PRACTICAL METHOD TO REDUCE PRIVACY LOSS WHEN DISCLOSING STATISTICS BASED ON SMALL SAMPLES

Raj Chetty
John N. Friedman

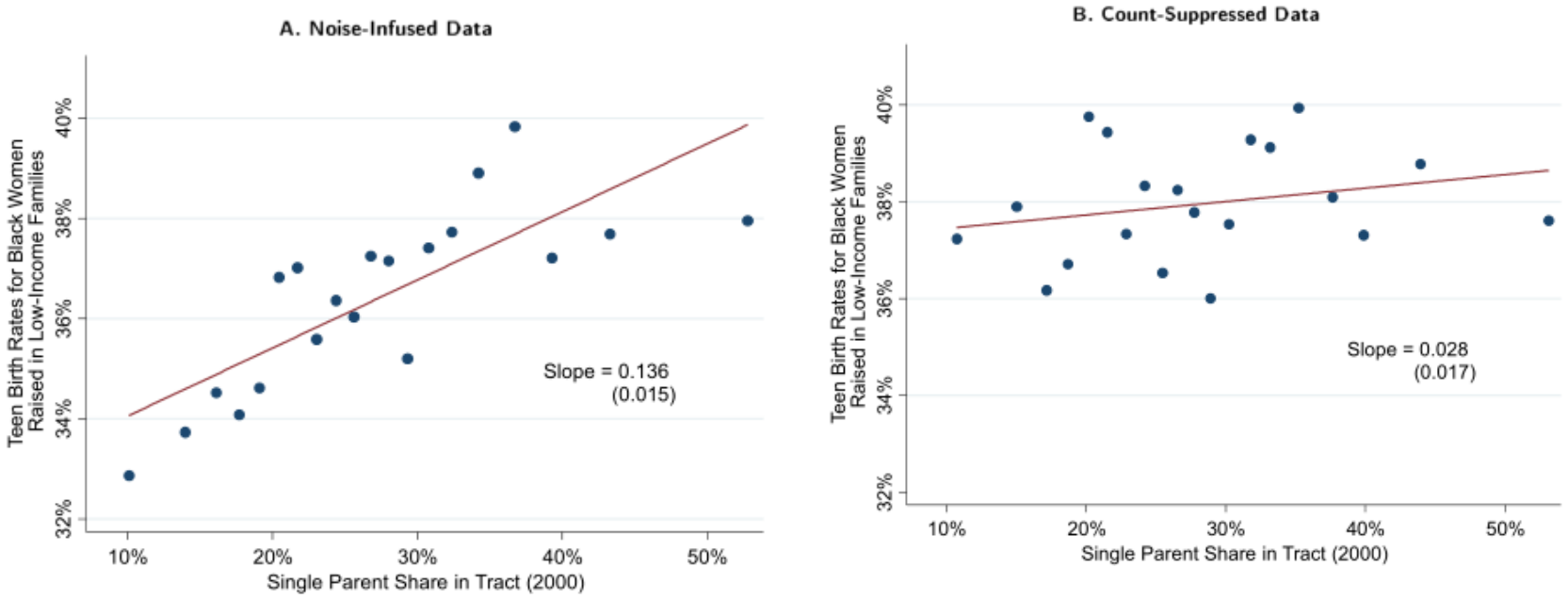


FIGURE 3: Association between Teenage Birth Rates and Single Parent Shares Across Census Tracts

Takeaways

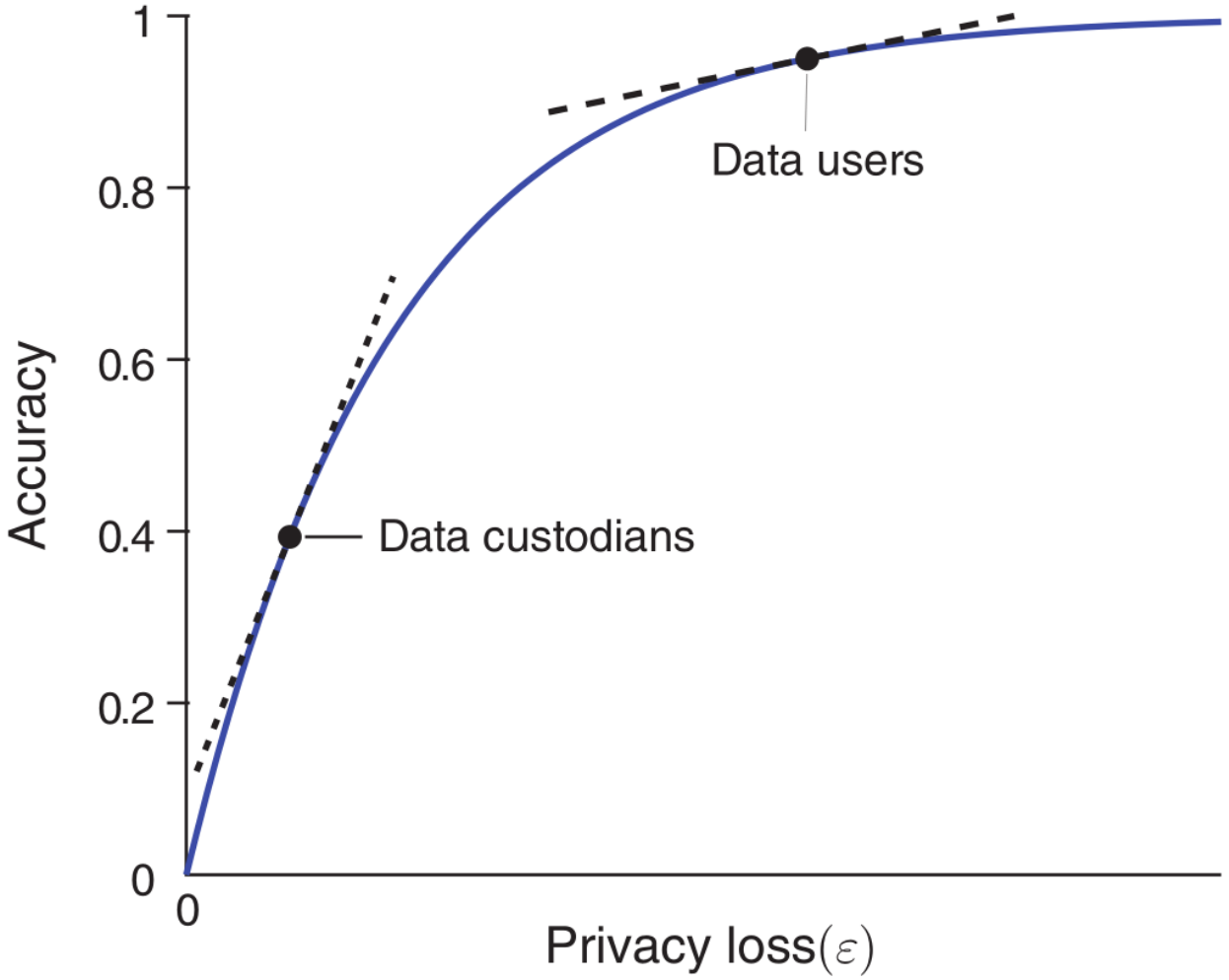
- We know analysis needs to account both for
  - The phenomenon of interest
  - The measurement of that phenomenon

- Accounting for traditional privacy models either
  - Can’t be done
  - Actively undoes privacy protection

- Privacy-aware analysis requires transparent formal privacy systems

Accuracy for What?

What are we buying with privacy loss?



Decision-making

D: a population-level dataset

$q(D)$: some population statistic

a: the published output

Accuracy based on some loss function

$$L(q(D), a)$$

Proposed Accuracy Measures for 2020 Decennial

Use Cases

- Zero-Sum
- Total/Category (Allocation)
- Single Year of Age
- Rates (population shares)
- Percent Threshold
- Numeric Threshold

Accuracy Measures

- Mean/Median Absolute Error (MAE)
- Mean/Median Numeric Error (ME)
- Root Mean-Squared Error (RMSE)
- Mean/Median Absolute Percent Error (MAPE)
- Coefficient of Variation (CV)
- Total Absolute Error of Shares (TAES)
- 90th Percentile Absolute Error

Policy Decision: Minority Language Voting Rights

- Voting Rights Act, Section 203
    - Jurisdictions are evaluated for 68 minority languages
    - Covered if number/shares of speakers surpasses threshold
    - If covered, must provide election information in minority language

What if these decisions must be based on differentially private data?

Policy Decision: Minority Language Voting Rights

Table 1: Voting Rights, Minority Language Determinations

Assignees are all combinations of U.S. voting jurisdictions with each of 68 minority language categories.

- Assignees: $a = (j, l) \in$ Jurisdictions $\times$ Languages
- Outcomes: {Covered, Not-covered}
- $Q = \{vac, lep, lit\}$ where

  $vac(I_a)$: voting age citizens in j speaking language l.

  $lep(I_a)$: voting age citizens in j speaking language l, and limited-English proficient.

  $lit(I_a)$: voting age citizens in j speaking language l, limited-English proficient, and less than 5th grade education.
- $M(a; \mathbf{X}) = \left(\frac{\mathbf{X}_a^{lep}}{\mathbf{X}_a^{vac}} > 0.05 \vee \mathbf{X}_a^{lep} > 10000 \right) \wedge \frac{\mathbf{X}_a^{lit}}{\mathbf{X}_a^{lep}} > 0.0131$

Fair Decision Making Using Privacy-Protected Data

David Pujol
david.pujol@duke.edu
Duke University

Ryan McKenna
rmckenna@cs.umass.edu
University of Massachusetts, Amherst

Satya Kuppam
skuppam@cs.umass.edu
University of Massachusetts, Amherst

Michael Hay
mhay@colgate.edu
Colgate University

Ashwin Machanavajjhala
ashwin@cs.duke.edu
Duke University

Gerome Miklau
miklau@cs.umass.edu
University of Massachusetts, Amherst

Caveat

- Pujol et al. model does not accurately characterize how VRA Section 203 coverage is determined
    - Determination made by Census with model-based small-area estimates that account for sampling variation and other data issues.

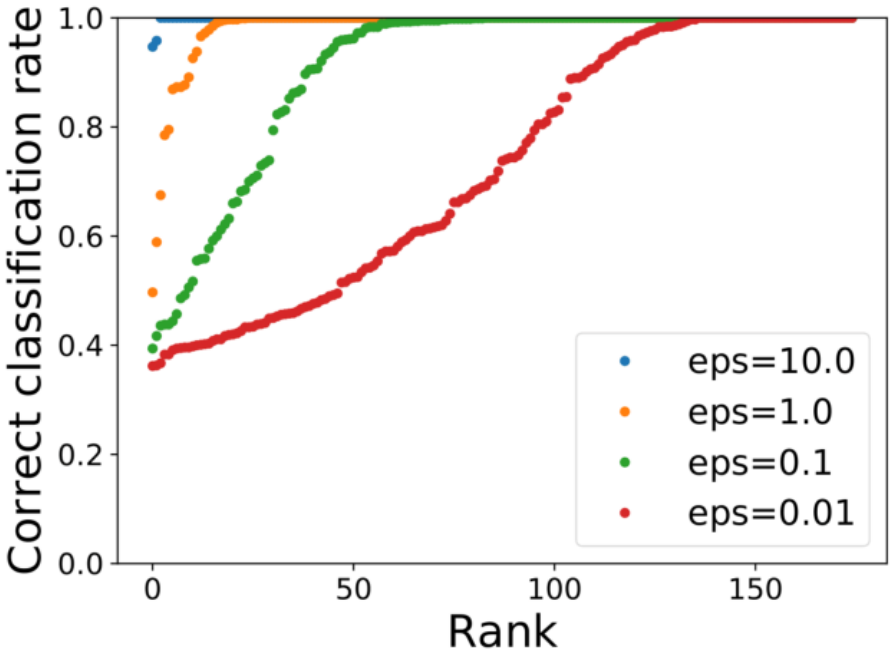
See…(https://www.census.gov/library/working-papers/2018/adrm/RRS2018-12.html)

Statistical Methodology (2016) for Voting Rights Act, Section 203 Determinations

Eric Slud
Robert Ashmead
Patrick Joyce
Tommy Wright

Simulation

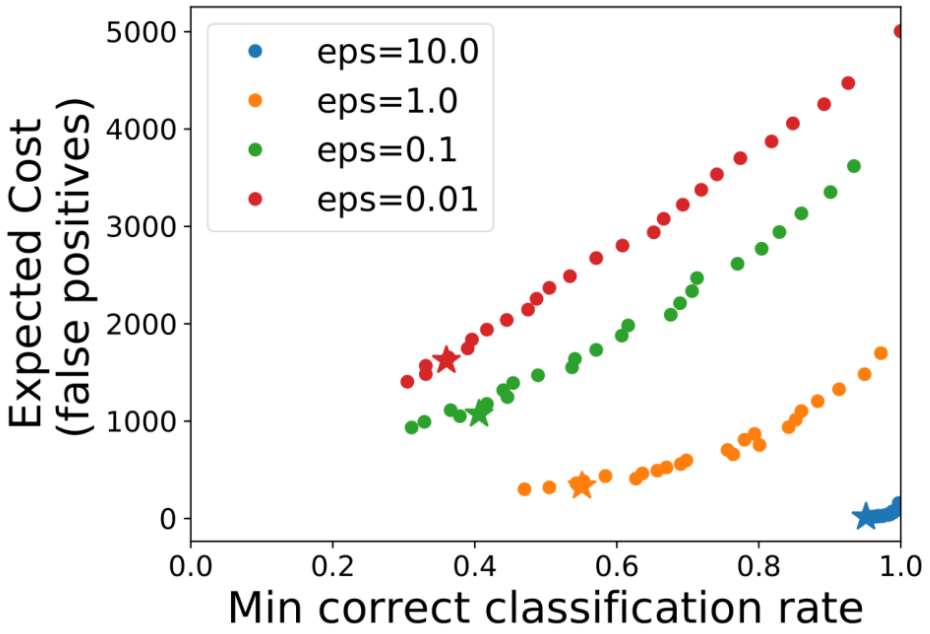
- 2016 ACS treated as “ground truth” for
$$X^{vac}, X^{lep}, X^{lit}$$
- Produce noisy estimates from Laplace mechanism (modified)
$$\tilde{X}^{vac} = X^{vac} + \nu$$
$$\tilde{X}^{lep} = X^{lep} + \nu$$
$$\tilde{X}^{lit} = X^{lit} + \nu$$
- Assume DM *ignores* privacy protection
- How bad?
- Who loses?



(a) The D-Laplace algorithm

Privacy-aware decisions

- Inferences and policy assignments should account for the mechanism:
- Decision rule: Covered if
  $\Pr[M(a; x_a) = Covered|\widetilde{x_a}] > p$
- For $\epsilon = 1$, the correct classification rate can be increased to 80 percent and small cost (870 false positives)



Options

1. Bespoke publications tailored to each specific application
   e.g. Just publish a formally private classification

2. Reserve privacy budget to improve inference on particular questions
   e.g. “special tabs” to get improved classifications

3. Use mechanisms that are broadly optimal for a wide range of uses

Universally Optimal Privacy Mechanisms

Basic Setting

- $q(D)$ is a single counting query

- Different data users, i, with preferences,

$$u = u_i(a_i; q(D))$$

For some choice variable a_i

Universally Optimal Privacy Mechanisms

Given published output, $\mathrm{M}(D)$,

Data users make choices based on expected utility

$$\max_{a} E\left[u_i(a_i; q(D))\right]$$

Expectations over posterior beliefs about $q(D)$ given $M(D)$

a *privacy-aware* analysis

Universally Optimal Privacy Mechanisms

Ghosh, Roughgarden, Sundararajan (2012)

geometric mechanism

$$M(D) = q(D) + \nu$$

Where ν is geometrically distributed and scaled to ϵ is

1. Provably ϵ-differentially private
2. Universally optimal for a particular class of information consumers

Good news:
Geometric mechanism is approximated by Laplace, but it is also easy to sample from and discrete

Bad news

Universal optimality result requires

- Actions have the same (finite) domain as outputs (i.e. actions are also counts)
- Payoffs maximized when action "matches" the true count

$$a^* = q(D)$$

- Loss is symmetric around $a = q(D)$

... does not apply to the VRA classification problem (and others like it)

Simplified VRA application

In a simplified VRA, policymaker has preferences

$$u(a; q(D))$$

$q(D)$ is the count speaking Russian with limited English,

$a \in \{0,1\}$ is the VRA classification.

Ideally

$$a^* = M(q(D))$$

Where $M()$ is the classification rule which (for simplicity) only takes the count as its input.

More good news and more bad news

Publication as a constrained information design problem
(Schmutte and Yoder 2020)

- Geometric Mechanism is optimal as long as decision problem is **monotone:**

Choose higher action when beliefs put more weight on higher counts

- This includes classification problems like VRA

- DM would not do better by asking Census to provide classification directly using the same privacy-loss budget!

More good news and more bad news

More Bad News:

GM does not work for non-monotone decision problems

Examples of non-monotone problems:

- Safety Inspections
  - If number of accidents is very high, suspect negligence
  - If number of accidents is very low, suspect fraud

- SARS-CoV-2 antibody tests for my class
  - If positive count is very high, no masks
  - If positive count is very low, no masks
  - Intermediate count: masks!

Takeaways

- For counting queries,
  - Publishing with geometric noise is optimal for a wide range of **monotone** use-cases
    - Requires user post-processing
    - aka *privacy-aware analysis*

- **NO** universally optimal methods for many non-counting queries (Brenner and Nissim 2014)

Statistical decisions with DP population statistics

- Privacy-aware hypothesis testing

  Wang, Li, Kifer (2017); Kifer and Rogers (2017)

- Interactive data analysis

  Good news:

  DP prevents overfitting / generalization bias (Dwork et al. 2016)

  Bad news:

  Point identification for certain estimators (RDD) may be impossible (Komarova and Nekipelov; 2020)

Thank You!

Ian M. Schmutte
http://ianschmutte.org
schmutte@uga.edu