Accuracy Disparities and Social Choices in the Deployment of Privacy Mechanisms

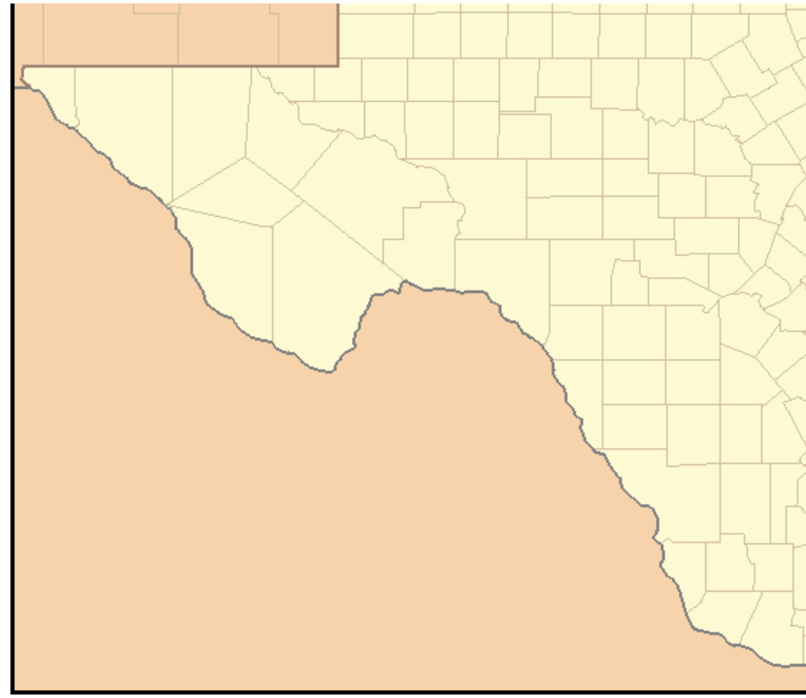
Gerome Miklau
UMass Amherst

Joint work with:
Ashwin Machanavajjhala (Duke), Michael Hay (Colgate), Ryan McKenna (UMass), David Pujol (Duke), Satya Kuppam (UMass)

The opinions expressed in this talk are my own and not those of the U.S. Census Bureau

Assignment

- A: universe of disjoint **assignee populations**
- M: assignment method (deterministic)
- O: outcome space



$$\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \xrightarrow{\text{Query}} \begin{matrix} Q(a_1) = (x_1, y_1) \\ \vdots \\ Q(a_n) = (x_n, y_n) \end{matrix} \xrightarrow{\text{Assign M}} \begin{pmatrix} o_1 \\ \vdots \\ o_n \end{pmatrix}$$

$$o_i = \begin{cases} 1, & \textbf{if } q(a_i) > T \\ 0, & \textbf{otherwise} \end{cases}$$

Assignee population	
a_1	Anderson County
a_2	Andrews County
a_3	Angelina County
…	…
a_n	Zavala County

Statistics
(123, 7483)
(598, 8341)
…
(382, 7937)

Outcome
☒ Qualified
☐ Not Qualified
☐ Not Qualified
…
☒ Qualified

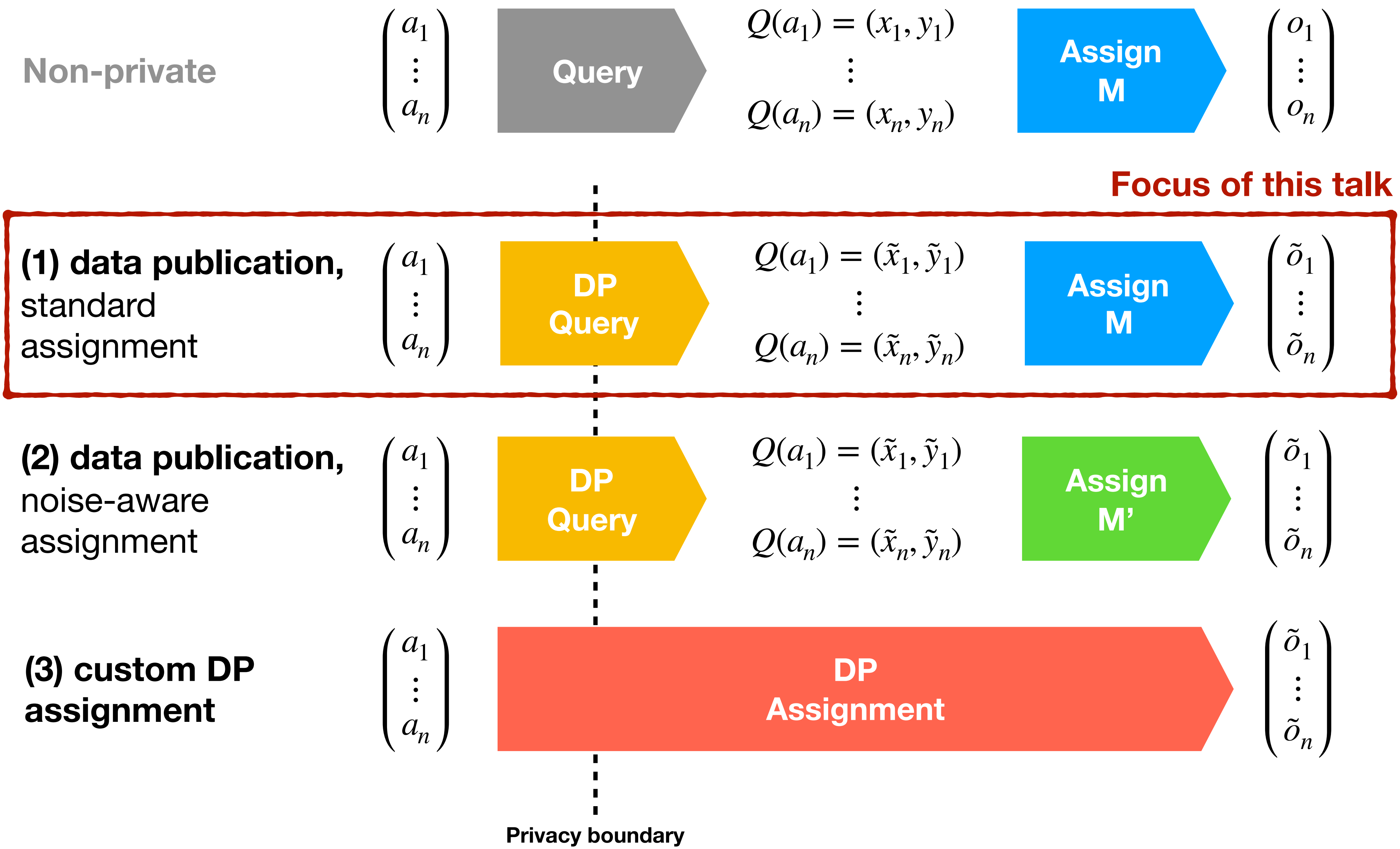
Assignment problems

Problem	Assignee Populations	Population Statistics	Outcome space
Federal funds allocation	states, counties, school districts, …	population counts	$
Congressional apportionment	states	resident counts	seats
Minority language voting rights benefit	voting districts	voting-age citizens, limited English, and illiteracy.	{0,1}
Urban/Rural classification	census tracts	population counts	{0,1}
Redistricting tests	districts	population counts	{0,1}

Consequences of inaccuracy

Problem	Consequence
Federal funds allocation	funds misallocated
Congressional apportionment	seats in house misallocated: unfair representation
Minority language voting rights benefit	minority language voters disenfranchised; or jurisdictions waste money on unnecessary voting materials
Urban/Rural classification	urban benefits misallocated
Redistricting tests	valid district plans rejected; invalid district plans accepted

Alternatives for private assignment



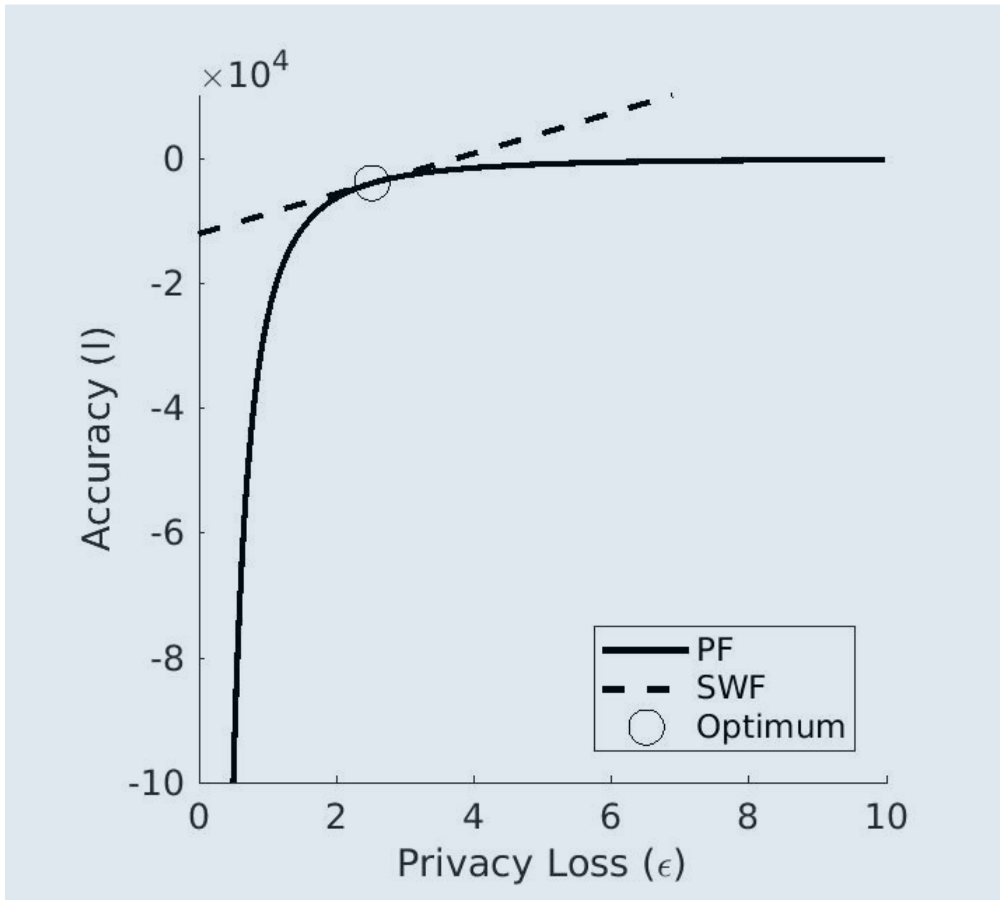
Common statistical agency practices

- Census tables based on **surveys** include estimates of sampling error (not the impact of disclosure limitation)
- Critical assignment problems may receive special treatment:
  - Redistricting and apportionment: no disclosure limitation on some supporting statistics.
  - Voting rights determinations: special variance reduction.
- In general, published tables **treated as true** for assignment problems.

Social choice: accuracy vs. privacy loss

Total Error

Utilitarian social welfare: sum of individual utilities



Abowd and Schmutte. An economic analysis of privacy protection and statistical accuracy as social choices. American Economic Review, 109(1), 2019.

Accuracy disparity

Given:

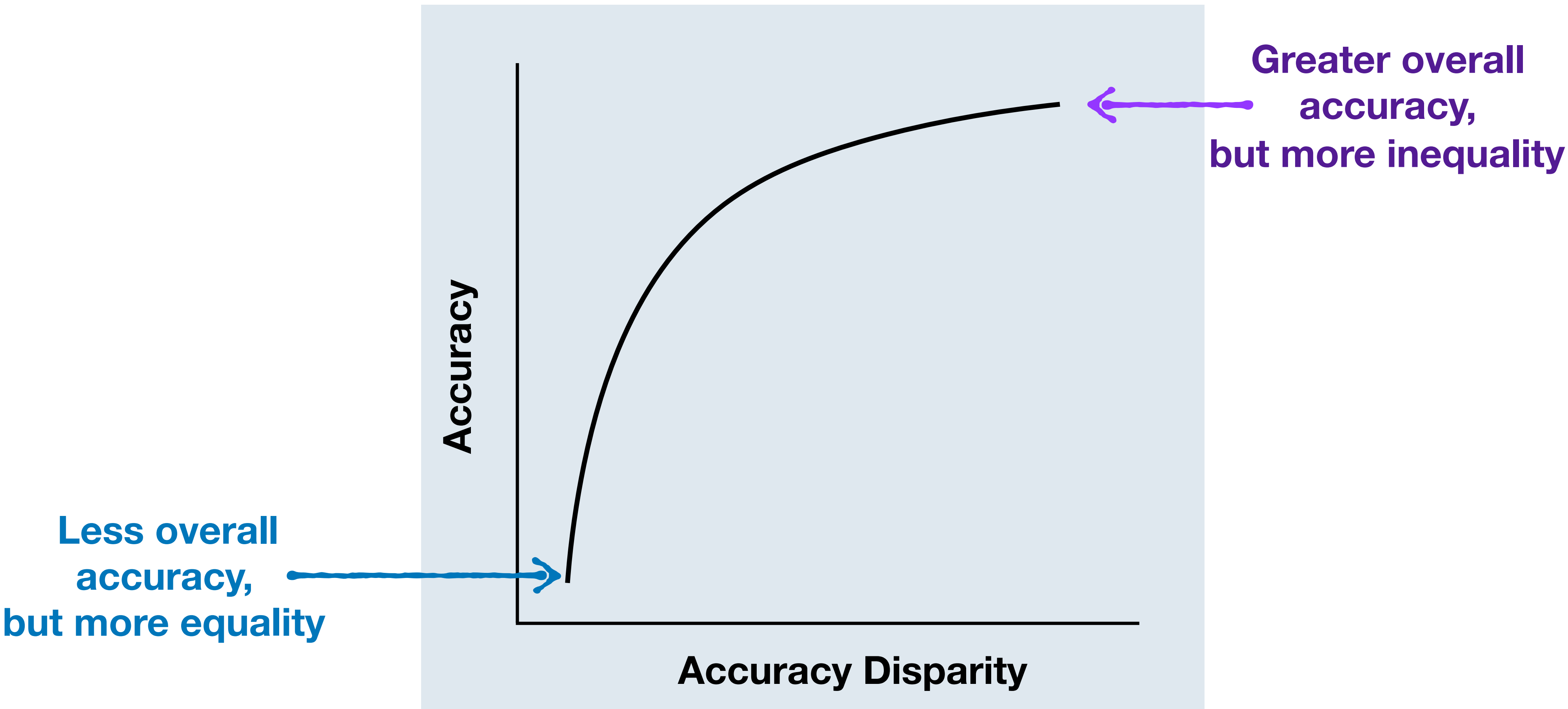
- a fixed privacy loss budget and
- the best available privacy mechanism

Do assignee populations bear the burden of inaccuracy equally?

Assignee population		True outcome	Correct Classification
a_1	Anderson County	☒ Qualified	$\boxed{54\%}$
a_2	Andrews County	☐ Not Qualified	85%
a_3	Angelina County	☐ Not Qualified	95%
…	…	…	
a_n	Zavala County	☒ Qualified	89%

Social choice:
accuracy vs. accuracy disparity

FOR A FIXED EPSILON:



Remainder of the talk

1. Introduction
2. **Causes of accuracy disparities**
3. Cases studies
    - Voting rights benefits
    - Title I education funding
4. Discussion and conclusion

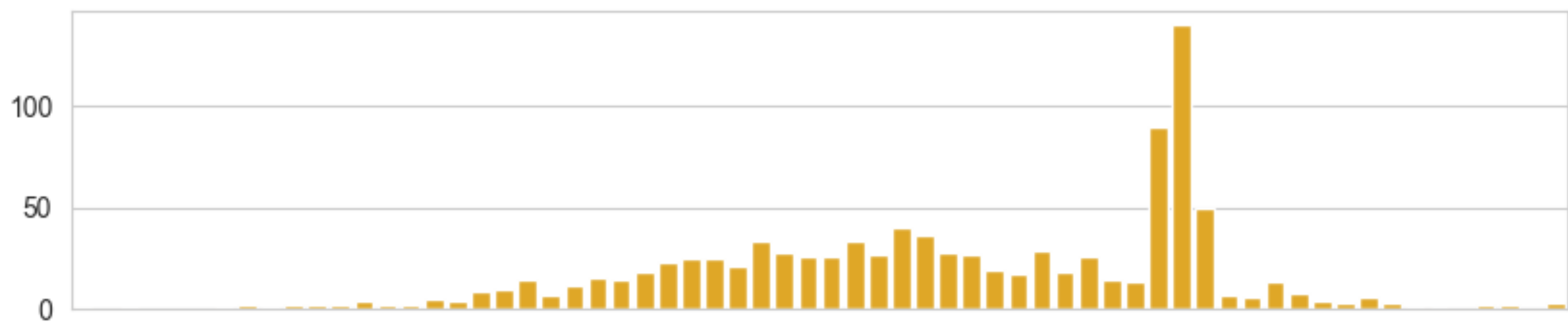
Accuracy disparities

- **Different groups may experience:**
  - **unequal error** rates in estimated counts.
  - **bias** in estimated counts
  - **unequal outcomes**

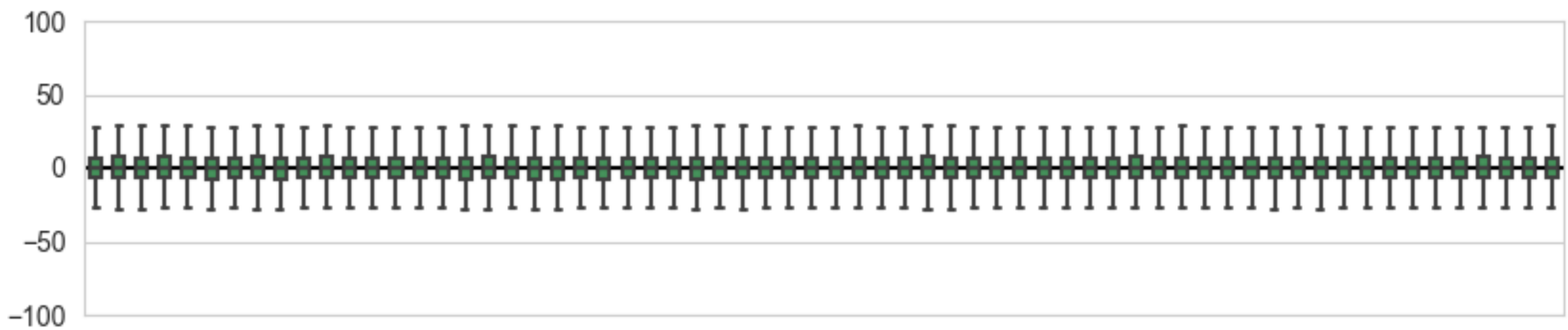
- **Algorithmic techniques that contribute to this:**
  - post-processing → **bias**
  - data-adaptive algorithms → **bias**
  - optimizing total error on a workload → **unequal error**
  - threshold conditions in assignment → **unequal outcomes**

Laplace mechanism

True sensitive data

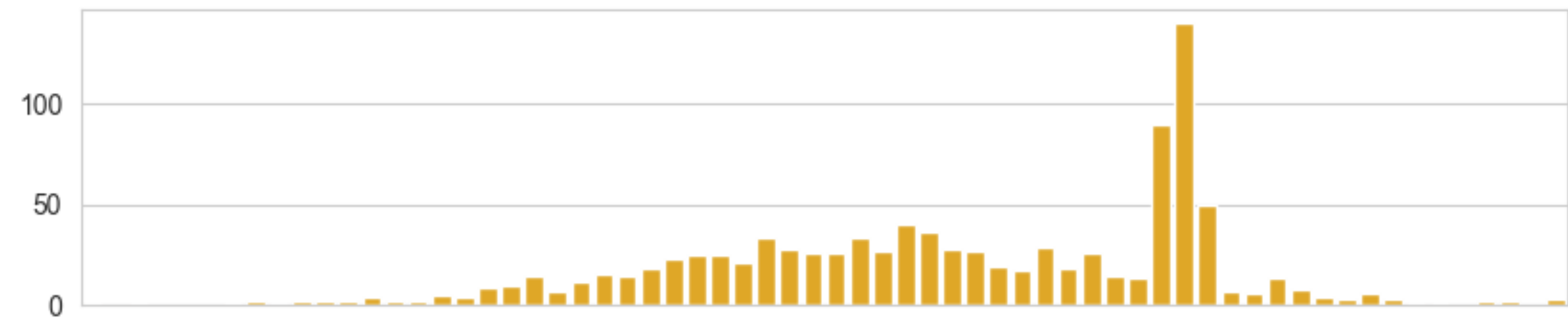


eps=.1   Expected L1 per query error = 9.98



Alternative mechanisms

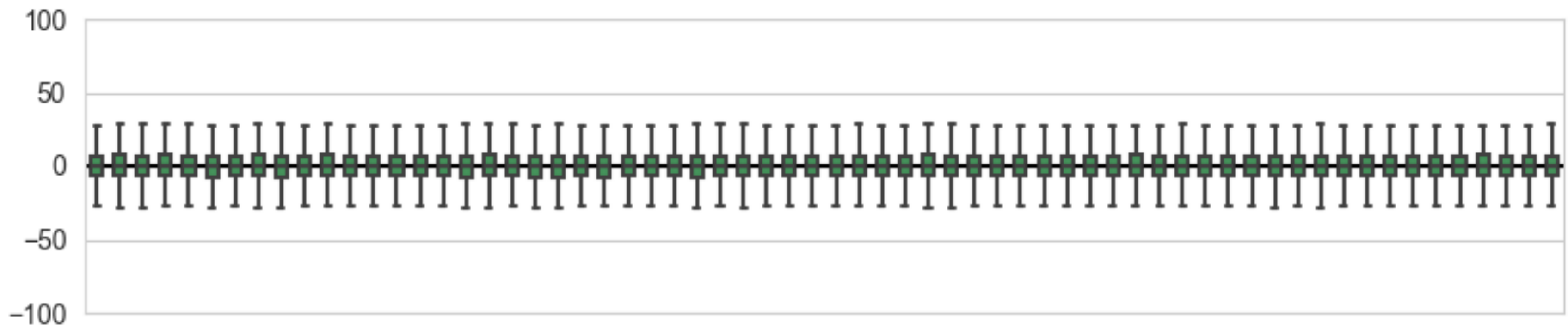
input data:



eps=.1

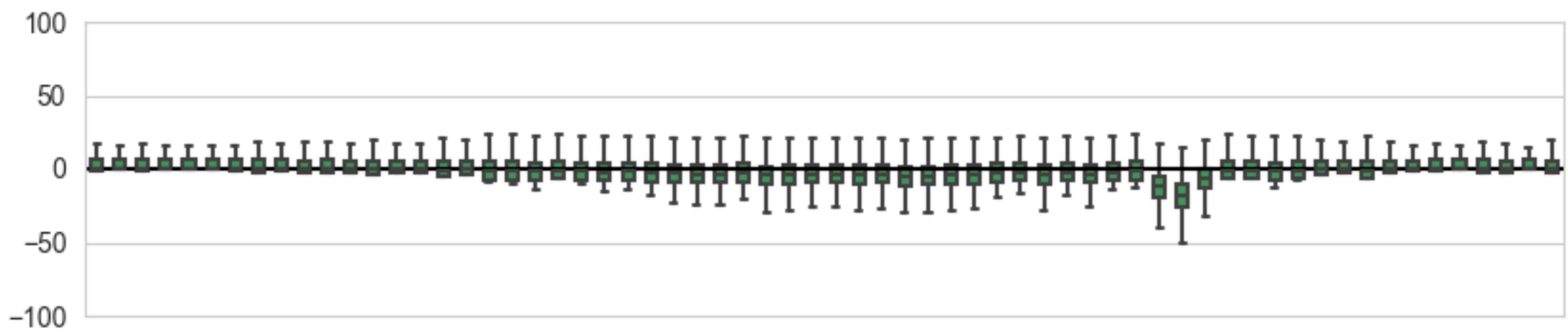
Laplace

error=9.98



Laplace w/
rounding

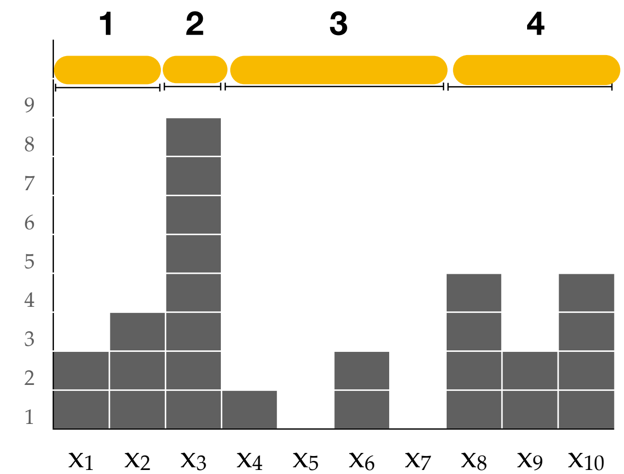
error=7.67

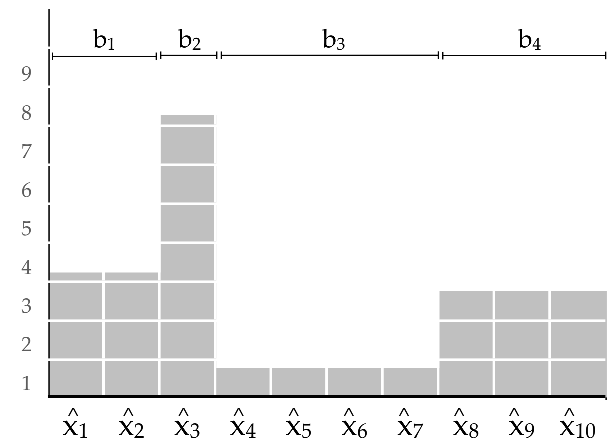


Data-adaptive mechanisms

DAWA

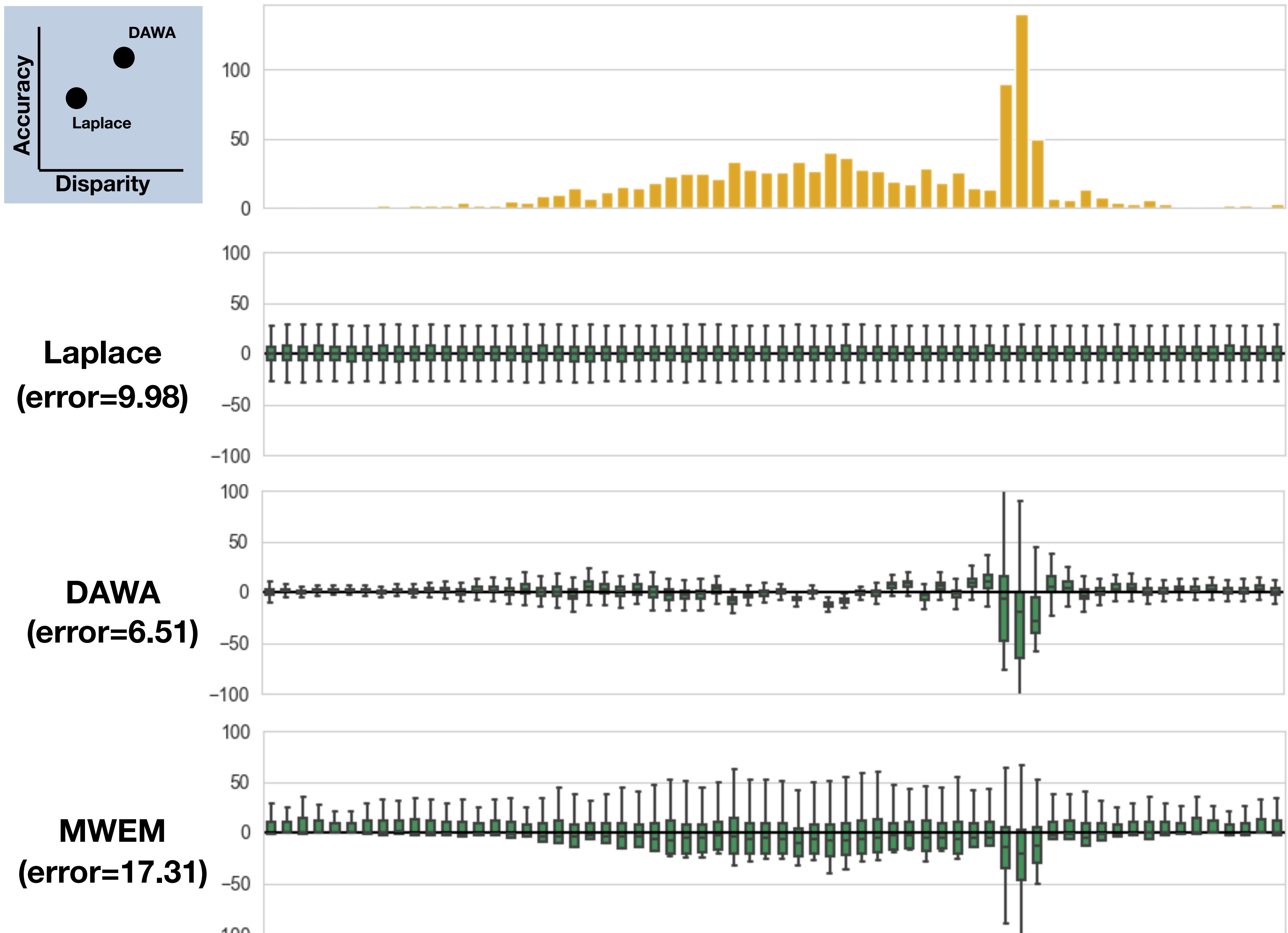
- Private data reduction
- Workload-adaptive measurements
- Least-squares inference





MWEM

- Uniform starting estimate
- Iterate:
  - measurement selection using Exponential Mechanism
  - Multiplicative weights inference


DAWA
Laplace
Accuracy
Disparity
Laplace
(error=9.98)
DAWA
(error=6.51)
MWEM
(error=17.31)
100
50
0
-50
-100

Matrix mechanism: workload adaptivity

Matrix Mechanism (MM)
[Li et al, PODS 2010]

Given workload W, find strategy A that minimizes total squared error on W

$\mathbf{x}$	$\leftarrow$	vectorize(R)
$\mathbf{W}$	$\leftarrow$	vectorize(W)
$\mathbf{A}$	$\leftarrow$	$\text{OPT}_{\text{MM}}(\mathbf{W})$
Δ_A	$\leftarrow$	$\lVert\mathbf{A}\rVert_1$
$\mathbf{a}$	$\leftarrow$	$\mathbf{Ax}$
$\mathbf{y}$	$\leftarrow$	$\mathbf{a}$ + Lap(Δ_A/ϵ)
$\underline{\mathbf{x}}$	$\leftarrow$	$\mathbf{A}^{+}\mathbf{y}$
ans	$\leftarrow$	$\mathbf{W}\underline{\mathbf{x}}$

select a "good" $\mathbf{A}$ for $\mathbf{W}$

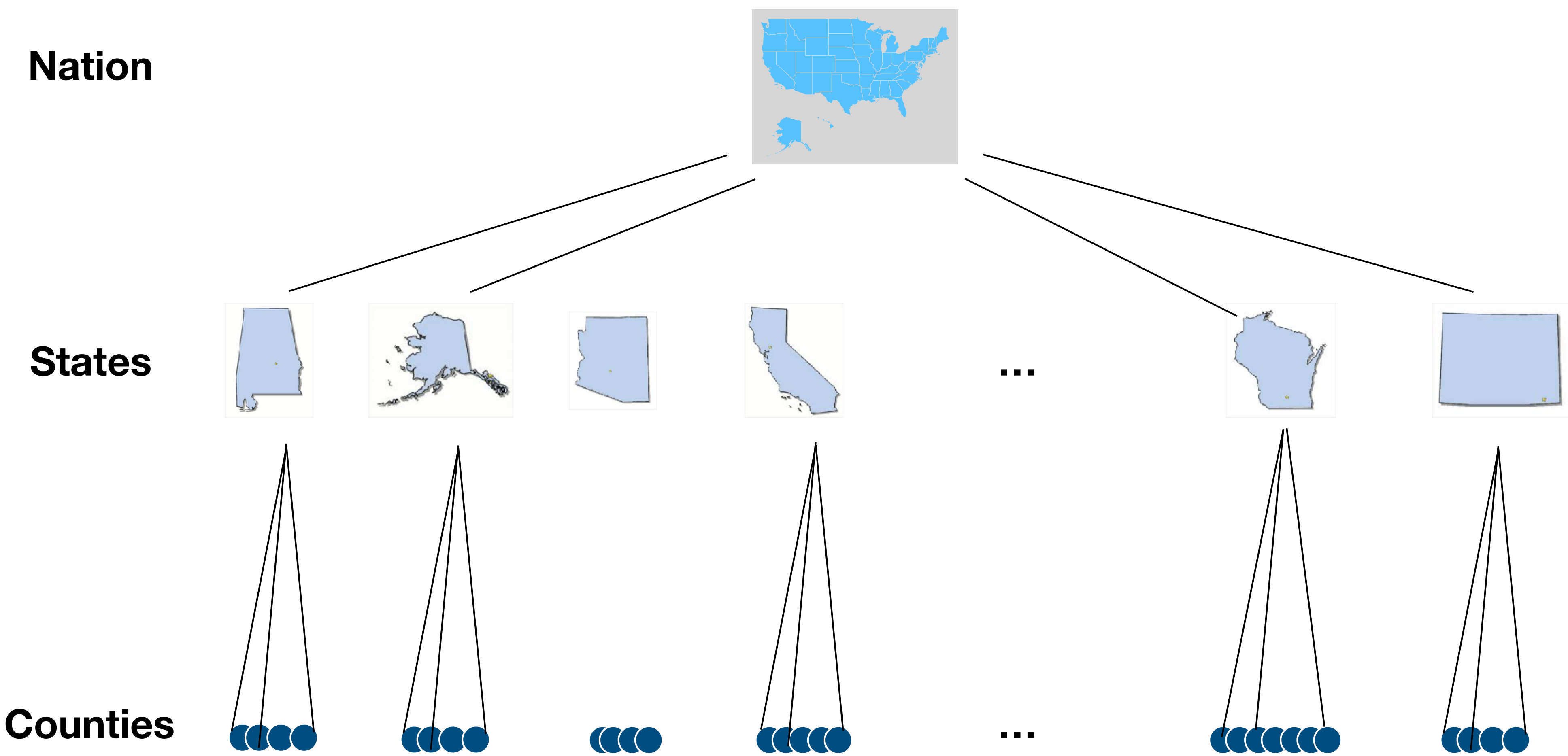
Laplace mechanism on $\mathbf{A}$

Reconstruct answers to $\mathbf{W}$ from noisy $\mathbf{A}$ answers

Key properties:

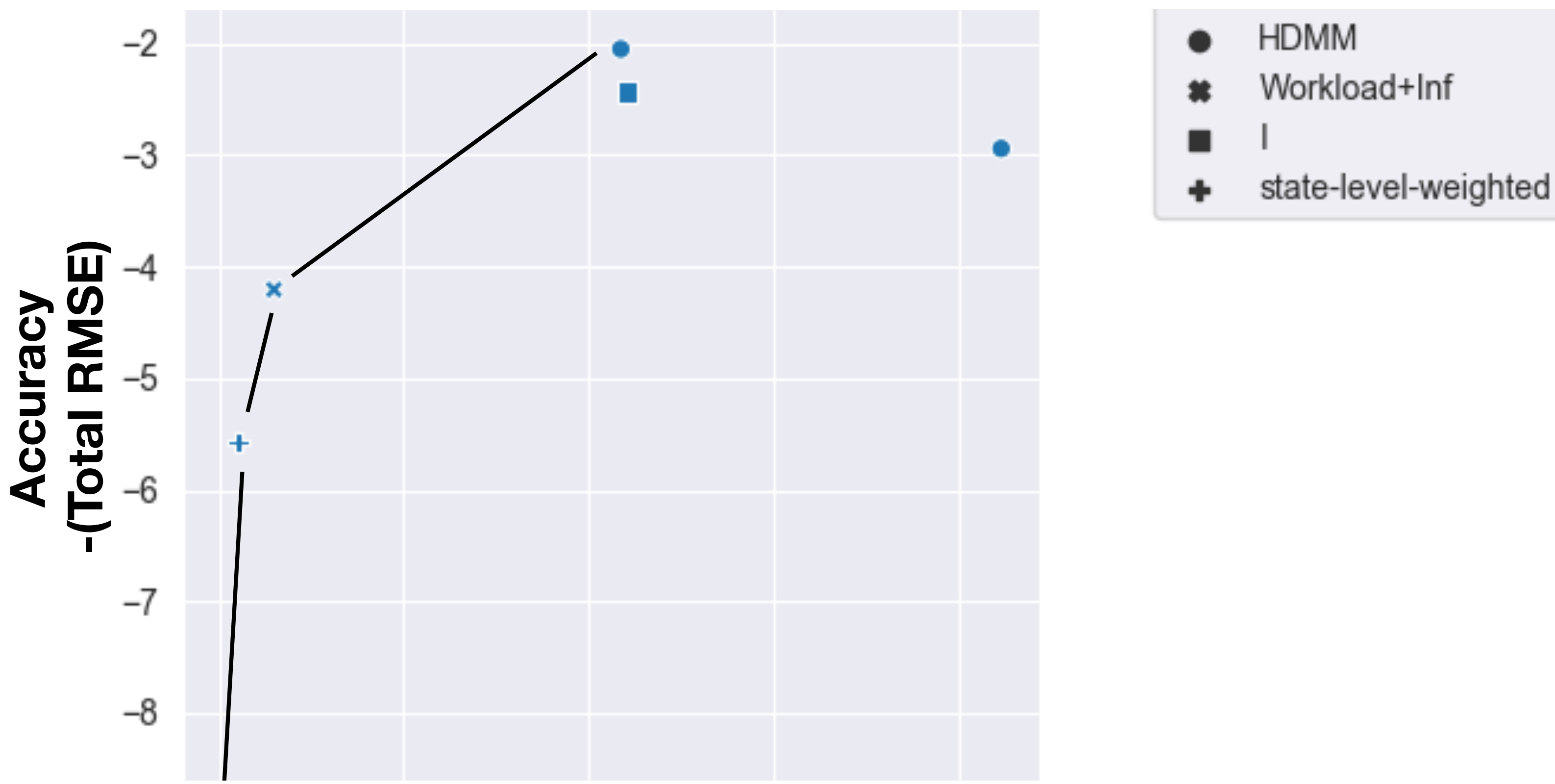
- ➡ **Unbiased answers to workload queries**
- ➡ **Data-independent expected error**
- ➡ **Expected error varies across workload**

Geographic hierarchy

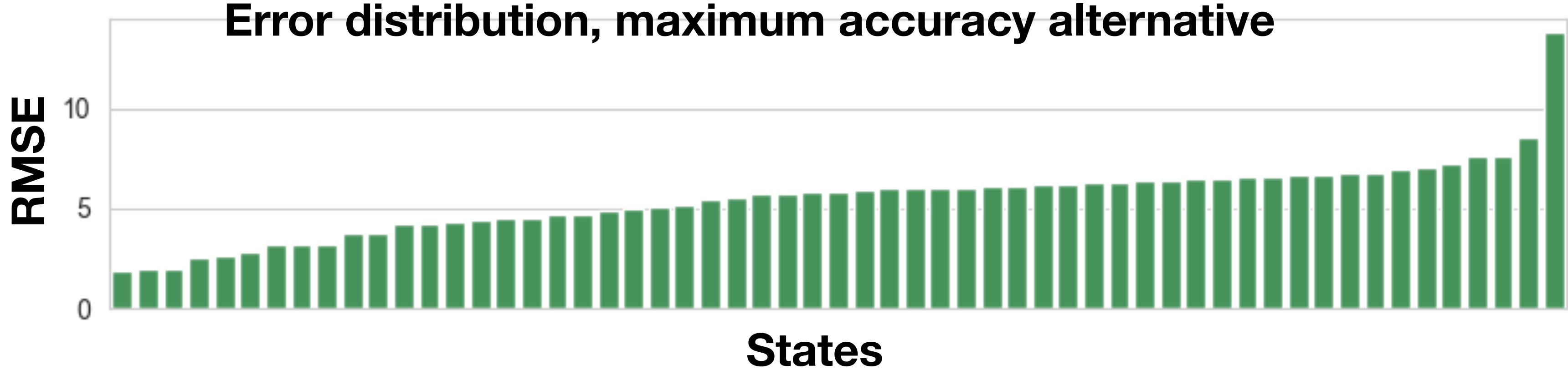


Workload: counts (of some predicate) at county, state, and national level.

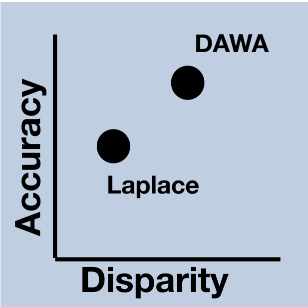
Accuracy on state counts

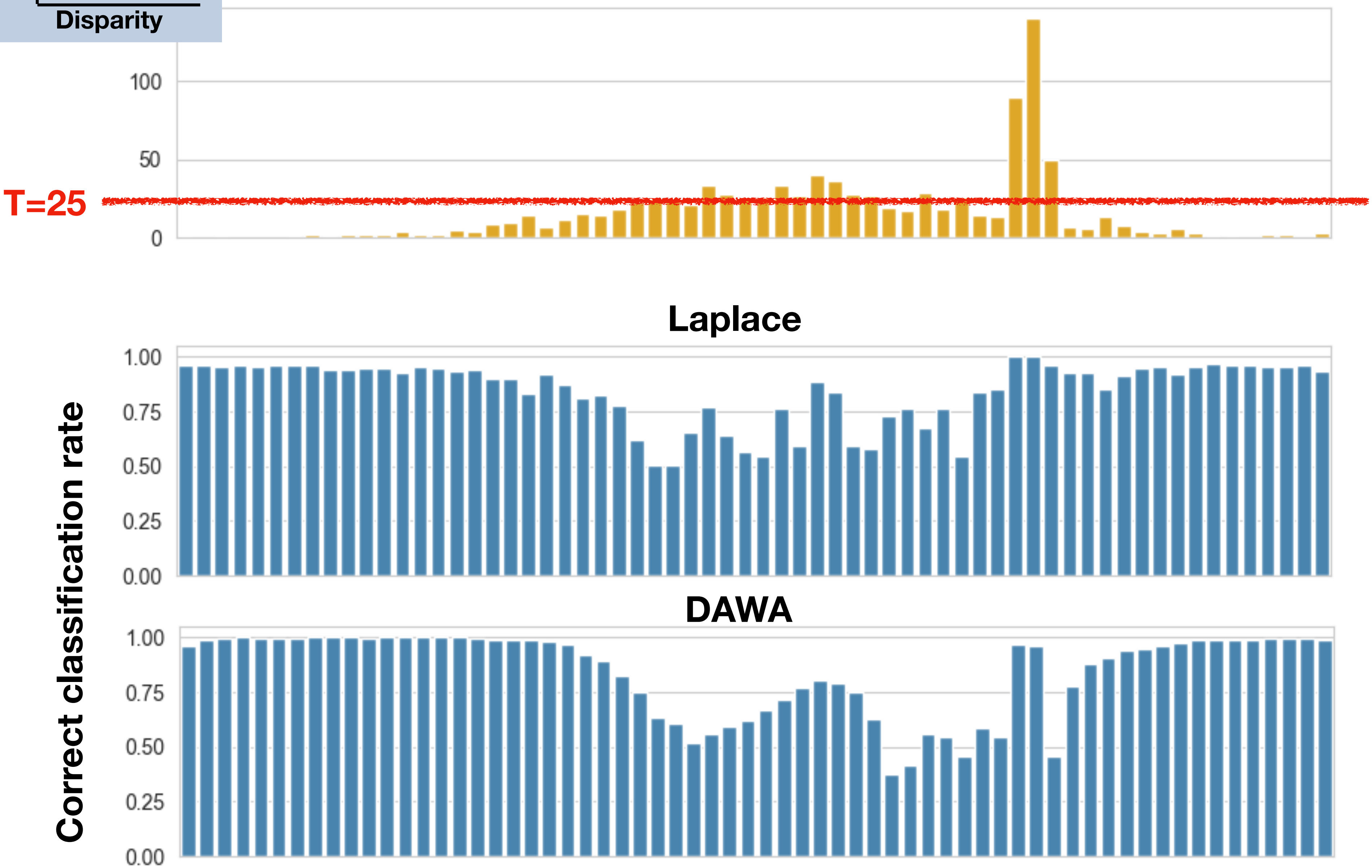


Error distribution, maximum accuracy alternative



Threshold assignment





Outline

1. Introduction
2. Causes of accuracy disparities
3. **Cases studies**
   - **Voting rights benefits**
   - **Title I education funding**
4. Discussion and conclusion

Minority language voting benefits

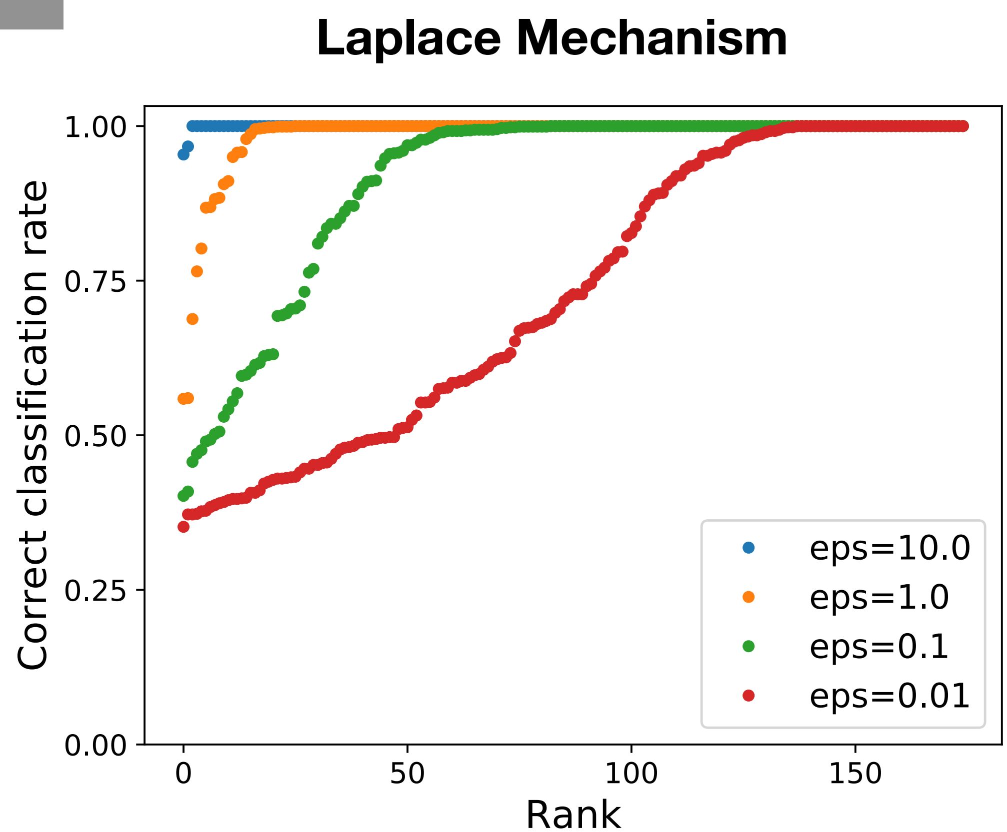
- Section 203 of the 1965 Voting Rights Act (U.S.) specified conditions under which jurisdictions must provide language assistance.
- A jurisdiction determined to be “covered” for language L must provide all election information (voter registration, ballots, and instructions) in the language L.
- Determinations made by the Census Bureau every 5 years, using published data.
- Last determinations in 2016: 263 out of 8000 jurisdictions covered (across all languages). 21 million voters live in these jurisdictions.

Minority language voting benefits

- **For each jurisdiction j:**
  - **For each minority language L:**
    - Define:
      - $q_{vac}(a_j)$ = voting age citizens in **j** speaking language **L**
      - $q_{lep}(a_j)$ = voting age citizens in **j** speaking language **L** and limited-English proficient.
      - $q_{lit}(a_j)$ = voting age citizens in **j** speaking language **L** and limited-English proficient and less than 5th grade education.
    - If $\left(\frac{q_{lep}(a_j)}{q_{vac}(a_j)} > 0.05 \vee q_{lep}(a_j) > 10000\right) \wedge \frac{q_{lit}(a_j)}{q_{lep}(a_j)} > 0.0131$
    - Then a_j is covered for language **L**

Covered jurisdictions

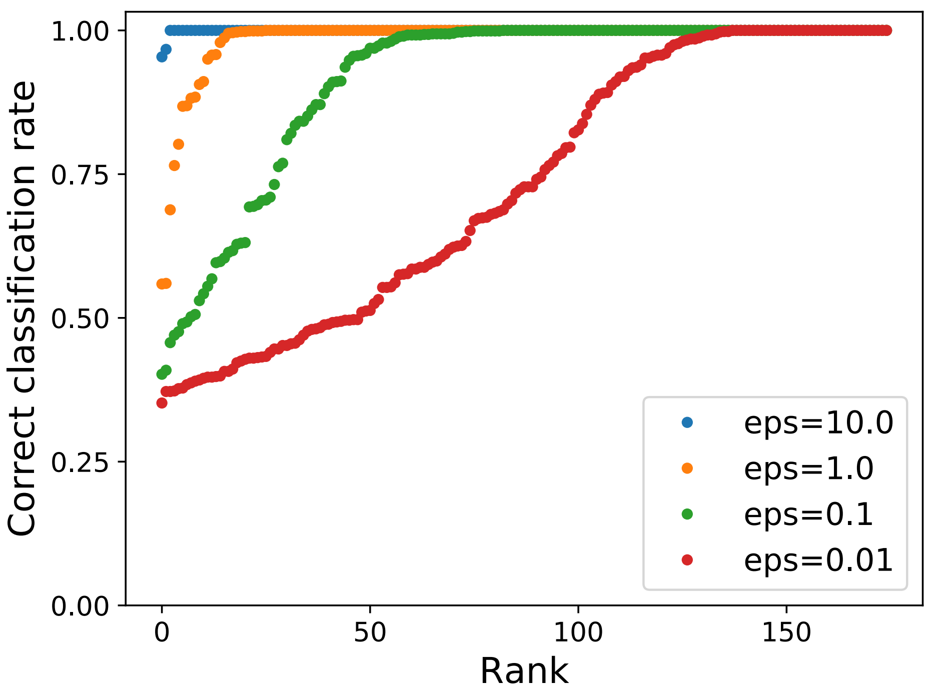
ϵ	% jurisdictions with correct classification >95%	lowest correct classification rate
10.0	100%	95%
1.00	92%	55%
0.10	74%	39%
0.01	33%	37%



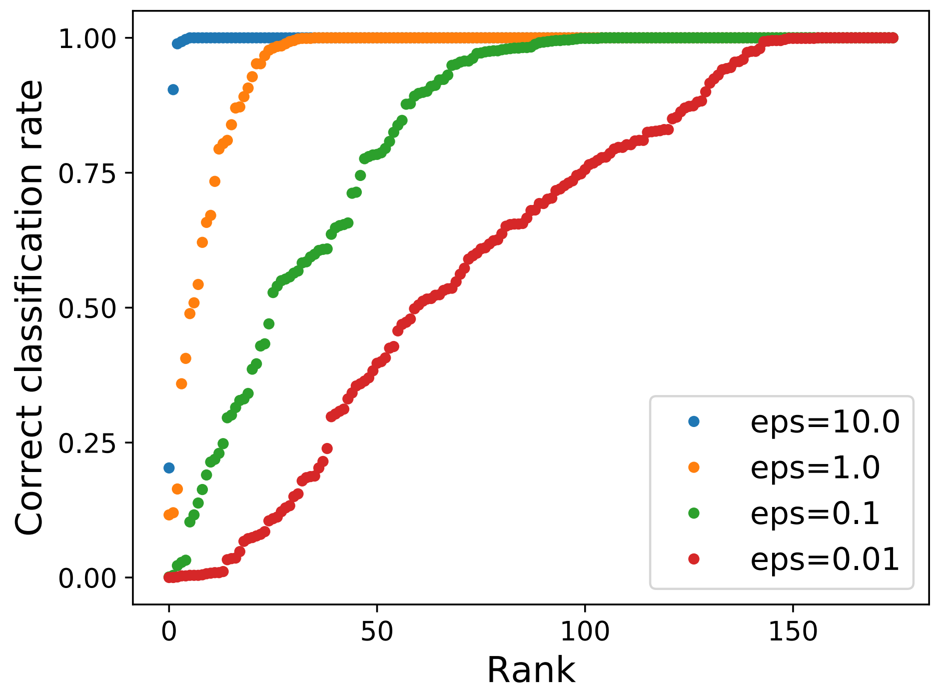
2016 public-use data (treated as ground truth)
"Hispanic" minority language group
175 positively classified jurisdictions

Laplace vs. DAWA

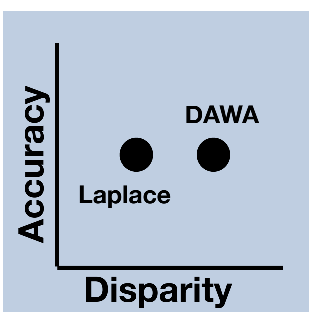
Laplace Mechanism



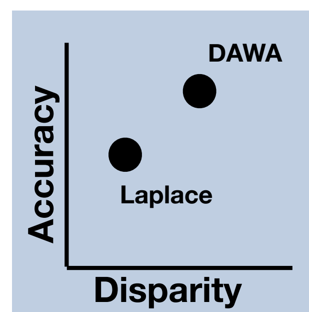
DAWA Algorithm



At $\epsilon = 0.1$ **DAWA and Laplace have equal total error**



At $\epsilon = 0.01$ **DAWA has 30% lower error than Laplace.**



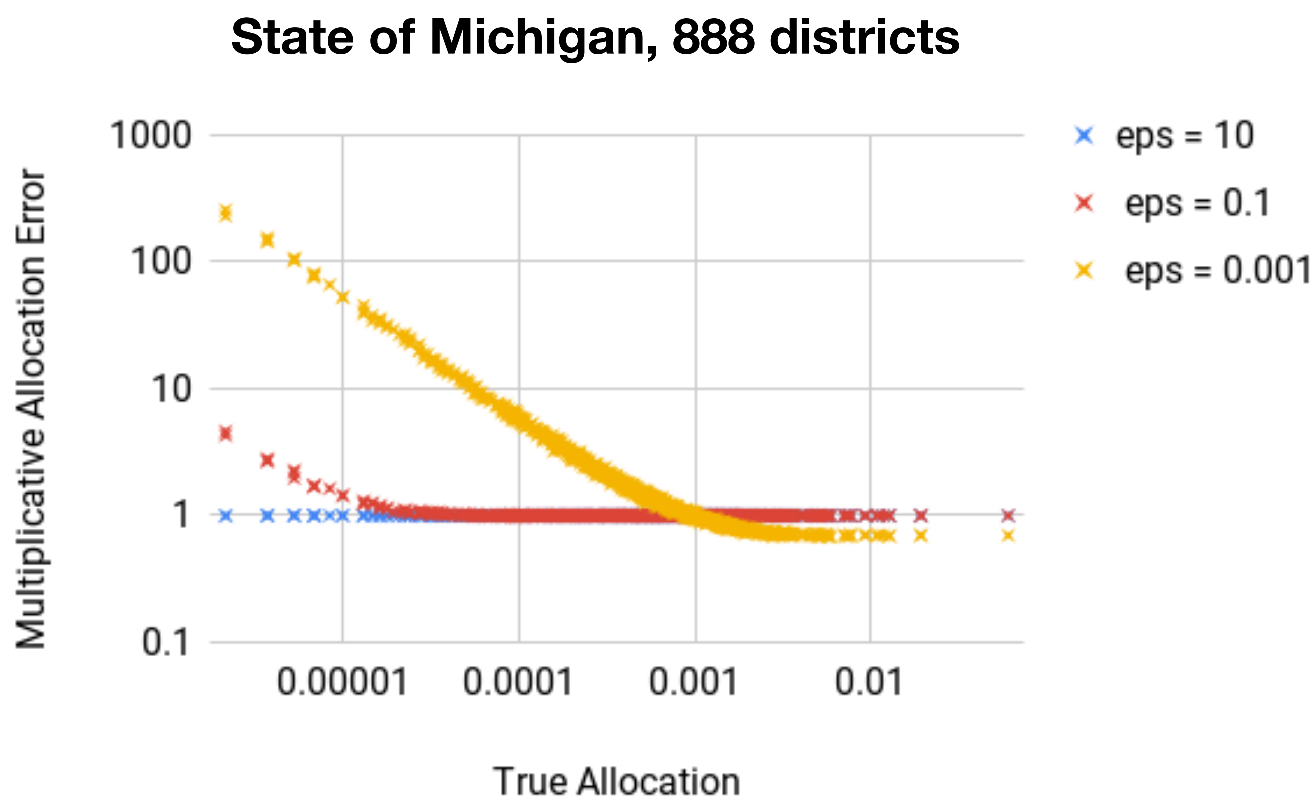
Title I funds allocation

- The allocation of at least $675 billion, annually, relies on Census data.
- Title I of the Elementary and Secondary Education Act of 1965 gives educational funding to school districts in proportion to number of children in financial need.
- In 2015, $6.5 billion was given through Title I “Basic grants”

Title I funds allocation

- Given total allocation **C**
- For each U.S. school district **d**
  - Define:
    - $q_{exp}(a_d)$ = average per student expenditure
    - $q_{eli}(a_d)$ = number of eligible students in district a.
  - Allocate to district **d:** $$\frac{Cq_{exp}(a_d)q_{eli}(a_d)}{\Sigma_i q_{exp}(a_i)q_{eli}(a_i)}$$

Allocation error



eps	small districts	large districts
10.0	1.01x ↑	0.001% ↓
0.10	10x ↑	0.05% ↓
0.001	500x ↑	50% ↓

Outline

1. Introduction

2. Causes of accuracy disparities

3. Cases studies

    - Voting rights benefits

    - Title I education funding

4. Discussion and conclusion

Summary

- Assignee populations do not bear the utility cost of existing privacy mechanisms equally.
- Disparities have a variety of causes:
  - minimizing total error, small counts biased up, counts near a decision boundary, those who get “asked about” less often, outliers biased towards neighbors…

Next steps?

- For what epsilons are disparities small enough to ignore?
- Can we develop privacy mechanisms that allow us to target more complex utility notions?
- Can we remedy disparities through post-processing or by adjusting assignment functions? Is this legally acceptable?
- Should individuals be able to choose how they weigh potential privacy harms against potential utility harms

Thank you

Results in this talk were made with ϵKTELO

https://github.com/ektelo/ektelo