The unintended disparate effects of privacy in decision tasks


Jacob Christopher

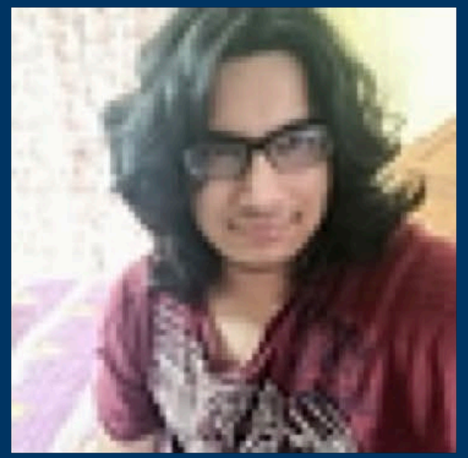

James Kotary


My Dinh


Vincenzo Di Vito


Michael Cardei


Jinhao Liang


Saswat Das


Key Zhu


Cuong Tran

Ferdinando Fioretto University of Virginia

https://nandofioretto.com

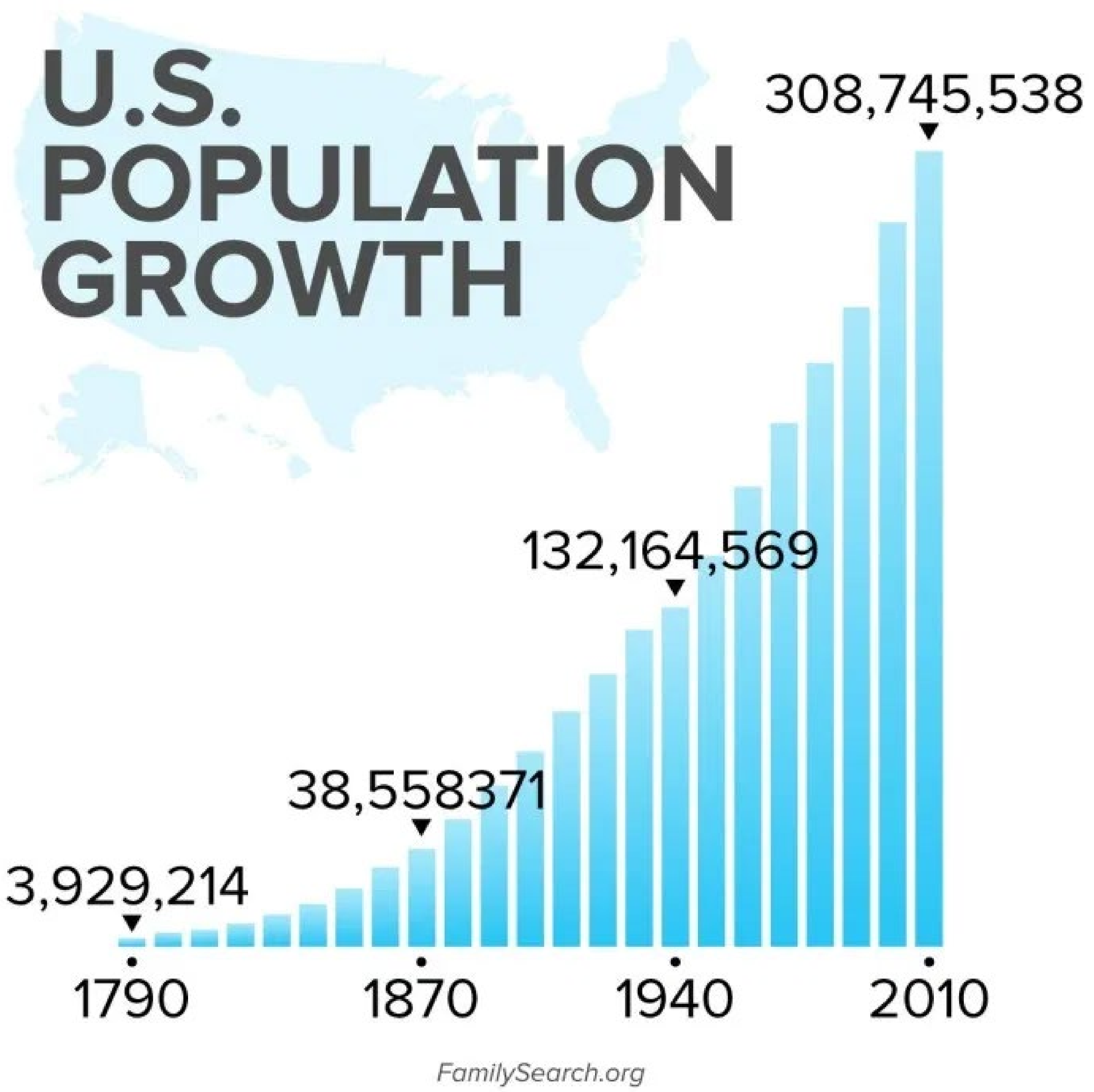
@nandofioretto

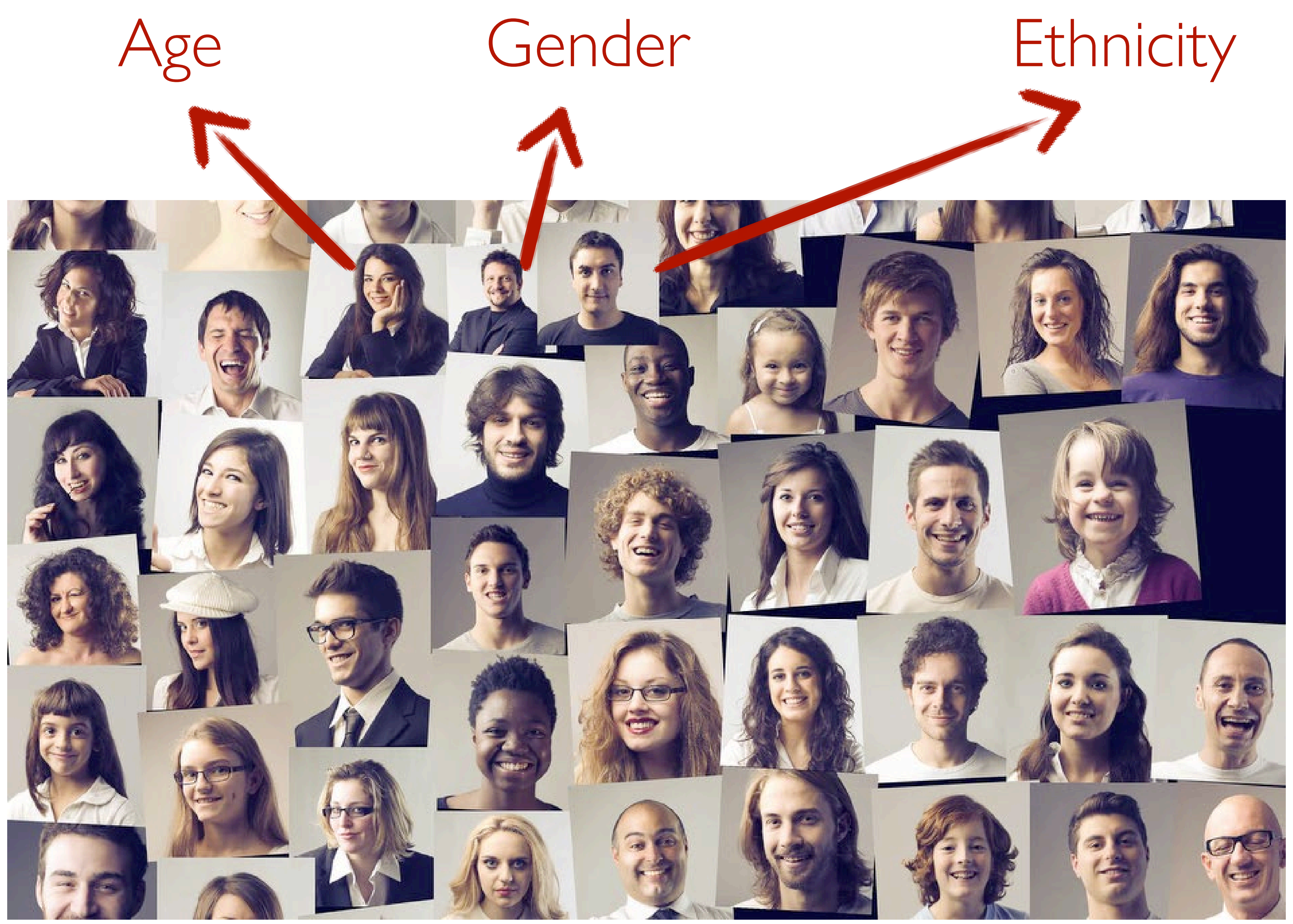
nandofioretto@gmail.com

@FCSM-24

US Census data collection

Enumeration of the total population living the US





US Census data collection

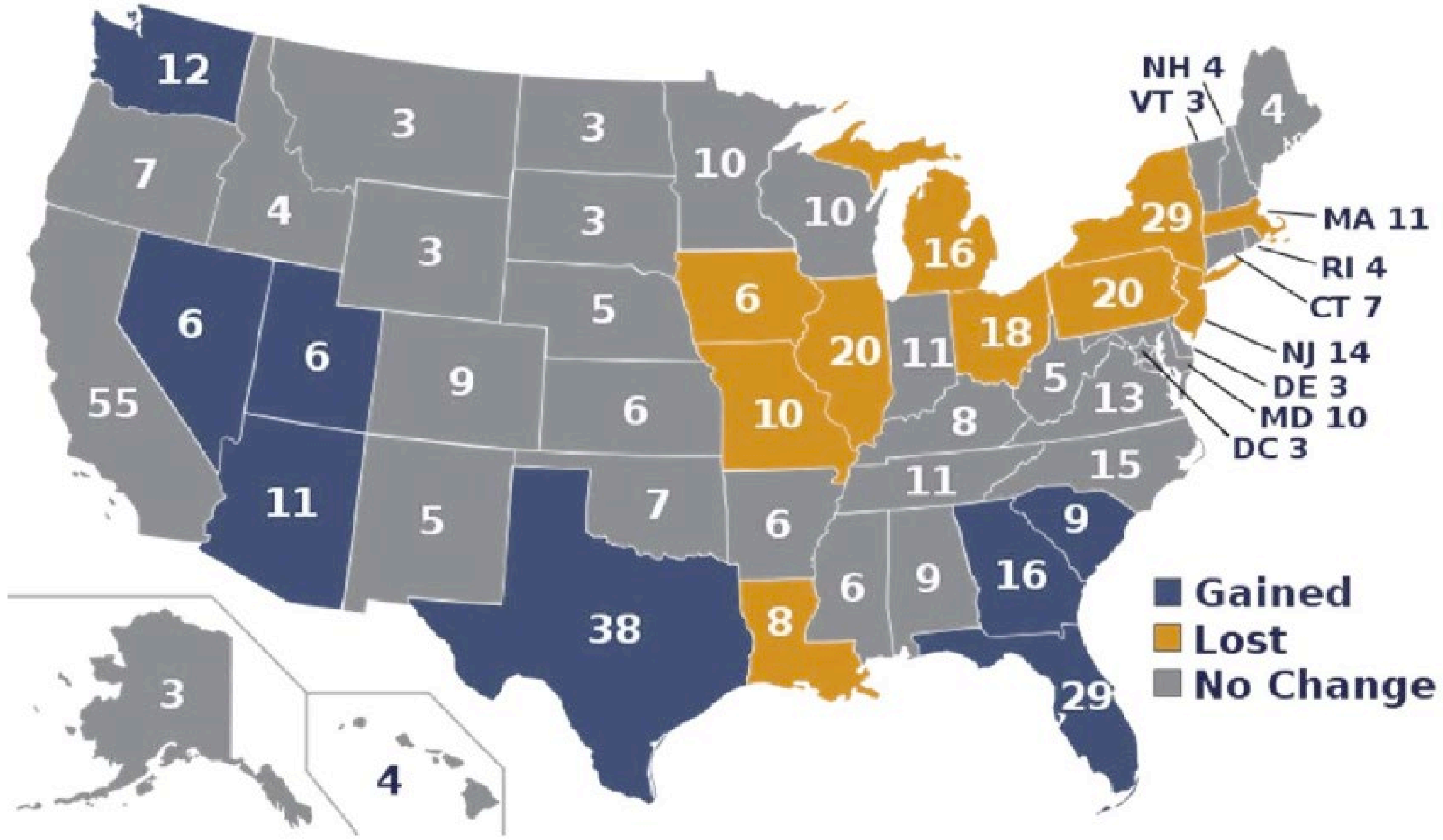
Accurate count is important

- Used to apportion multiple federal funding streams.
- $665 billions allocated to 132 economic security programs (2022) other than health insurance or social security benefits.

SNAP
Supplemental
Nutrition
Assistance
Program

Highway Planning and Construction

U.S. DEPARTMENT OF EDUCATION

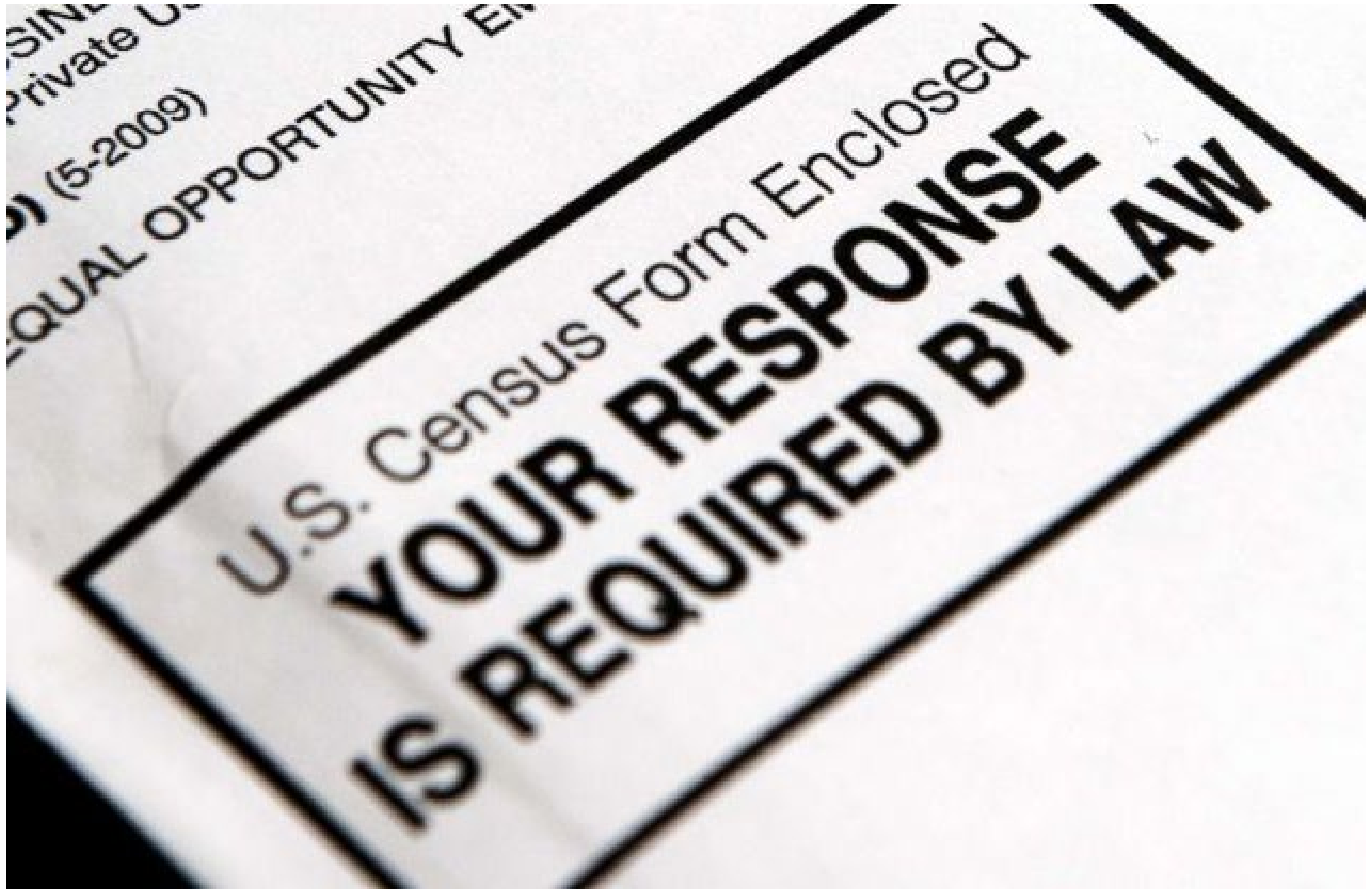


Determine the number of seats that states get in the US House of Representatives.

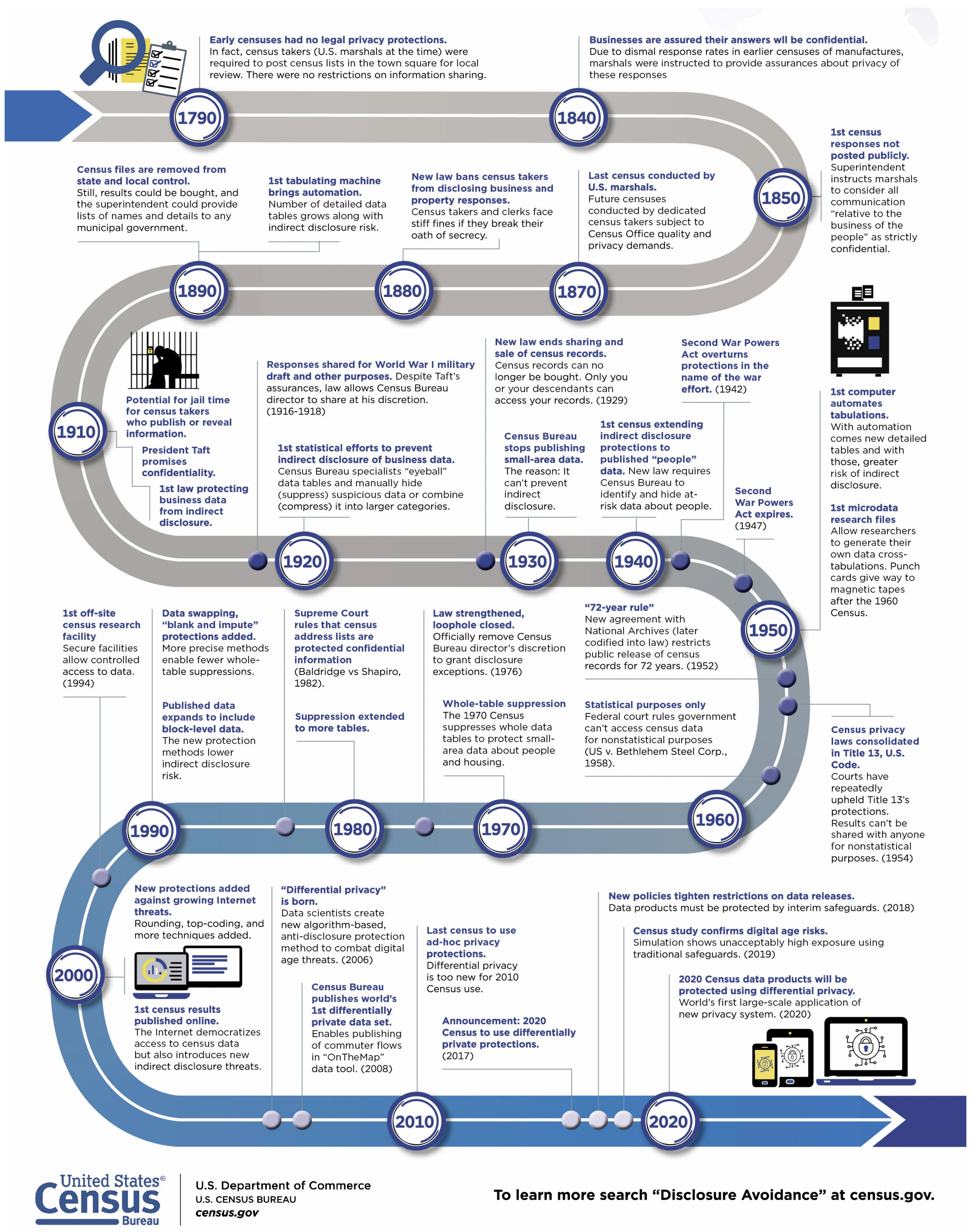
US Census data collection

Privacy is required by law

Because of the importance to have accuracy count congress makes the data collection mandatory.



Title 13: Census is required to retain data confidentiality.



Ferdinando Fioretto | UVA

Reconstruction Attacks

United States Census Bureau

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

308,745,548 people in 2010 release which implements some “protection”

Commercial databases

Linkage Attacks — Results from UC Census:

- Census blocks correctly reconstructed in all 6,207,027, inhabited blocks.
- Block, sex, age, race, ethnicity reconstructed:
  - Exactly: 46% of population (142M).
  - Allowing age +/- 1 year: 71% of population (219M).
- Name, block sex, age, race, ethnicity:
  - Confirmed re-identification: 38% of population.

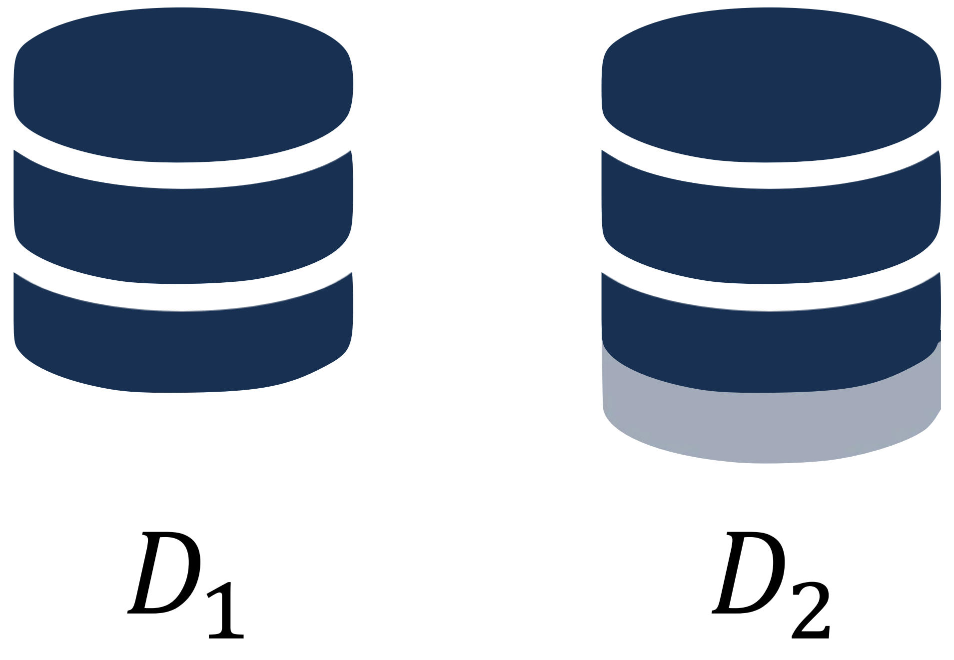
Ramachandran et al. 2012 McKenna et al. 2018

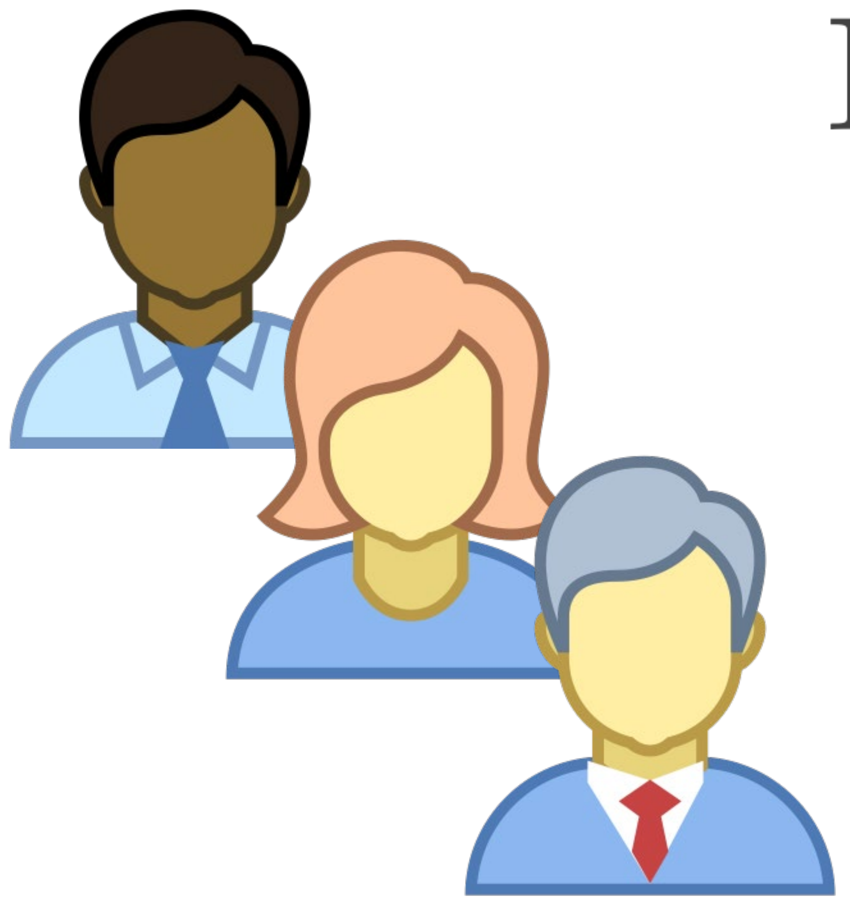
Differential Privacy

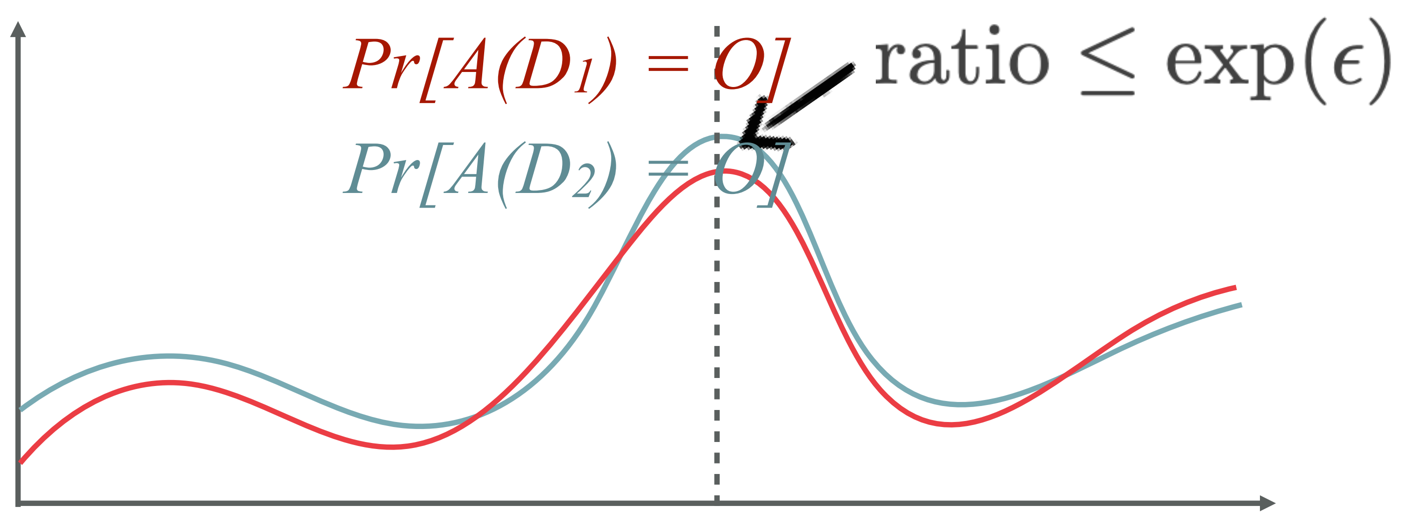
Definition

A randomized algorithm $\mathcal{A}$ is ε-differentially private if, for all pairs of inputs D_1, D_2, differing in one entry, and for any output O:

$$\frac{\Pr[\mathcal{A}(D_1) = O]}{\Pr[\mathcal{A}(D_2) = O]} \leq \exp(\epsilon)$$







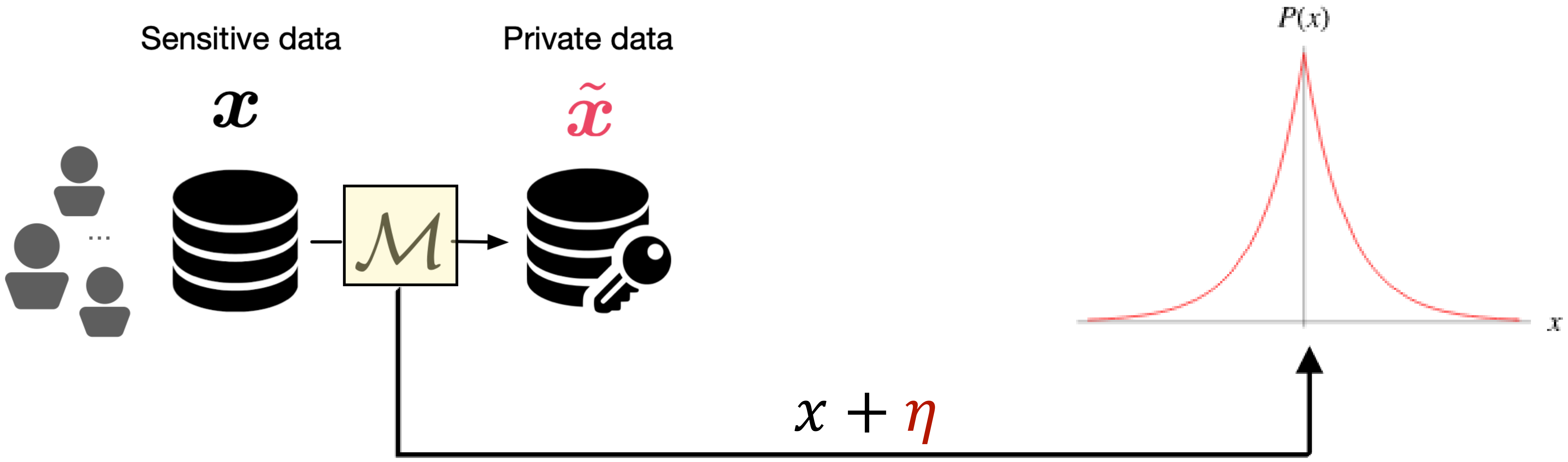
Intuition: An adversary should not be able to use output O to distinguish between any D_1 and D_2

Differential Privacy

Notable properties

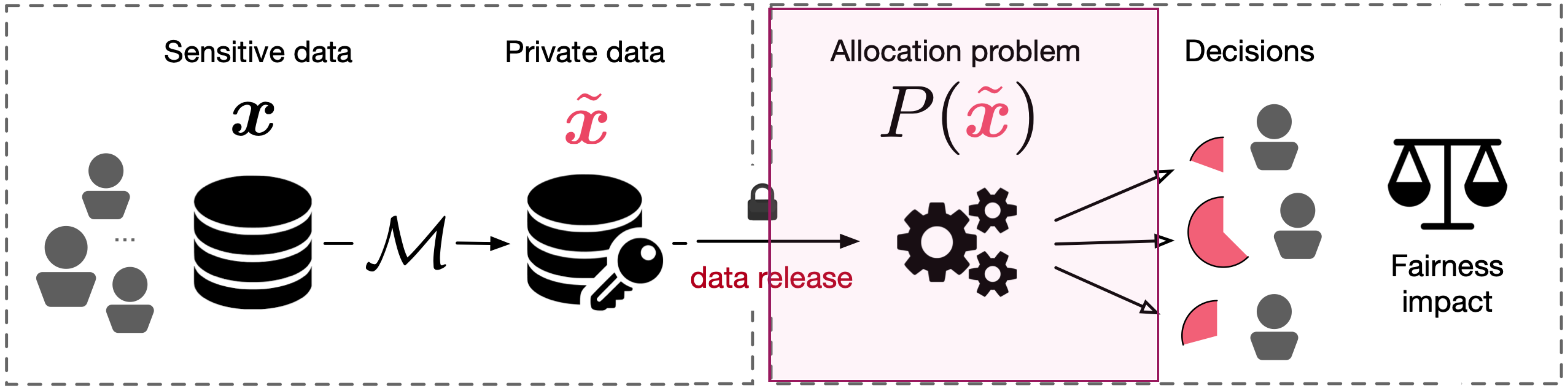
- Immune to linkage attack: Adversary knows arbitrary auxiliary information.
- Composability: If A_1 enjoys ε_1-differential privacy and A_2 enjoys ε_2-differential privacy, then, their composition $A_1(D)$, $A_2(D)$ enjoys $(\varepsilon_1 + \varepsilon_2)$-differential privacy.
- Post-processing immunity: If A enjoys ε-differential privacy and g is an arbitrary data-independent mapping, then $g \circ A$ s ε-differential private.

DP algorithms rely on randomization



Fairness in downstream decisions

Setting



Bias: $B^i_P(M, x) = \mathbb{E}_{\tilde{\boldsymbol{x}} \sim M(\boldsymbol{x})}[P_i(\tilde{\boldsymbol{x}})] - P_i(\boldsymbol{x})$

Definition ($\boldsymbol{\alpha}$-Fairness). A data-release mechanism M is said α-fair w.r.t. a problem P if, for all datasets $x \in \mathcal{X}$ and all $i \in [n]$

$$\xi^i_B(P, \mathcal{M}, \boldsymbol{x}) = \max_{j \in [n]} \left| B^i_P(\mathcal{M}, \boldsymbol{x}) - B^j_P(\mathcal{M}, \boldsymbol{x}) \right| \leq \alpha$$

Disproportionate impacts in decision making

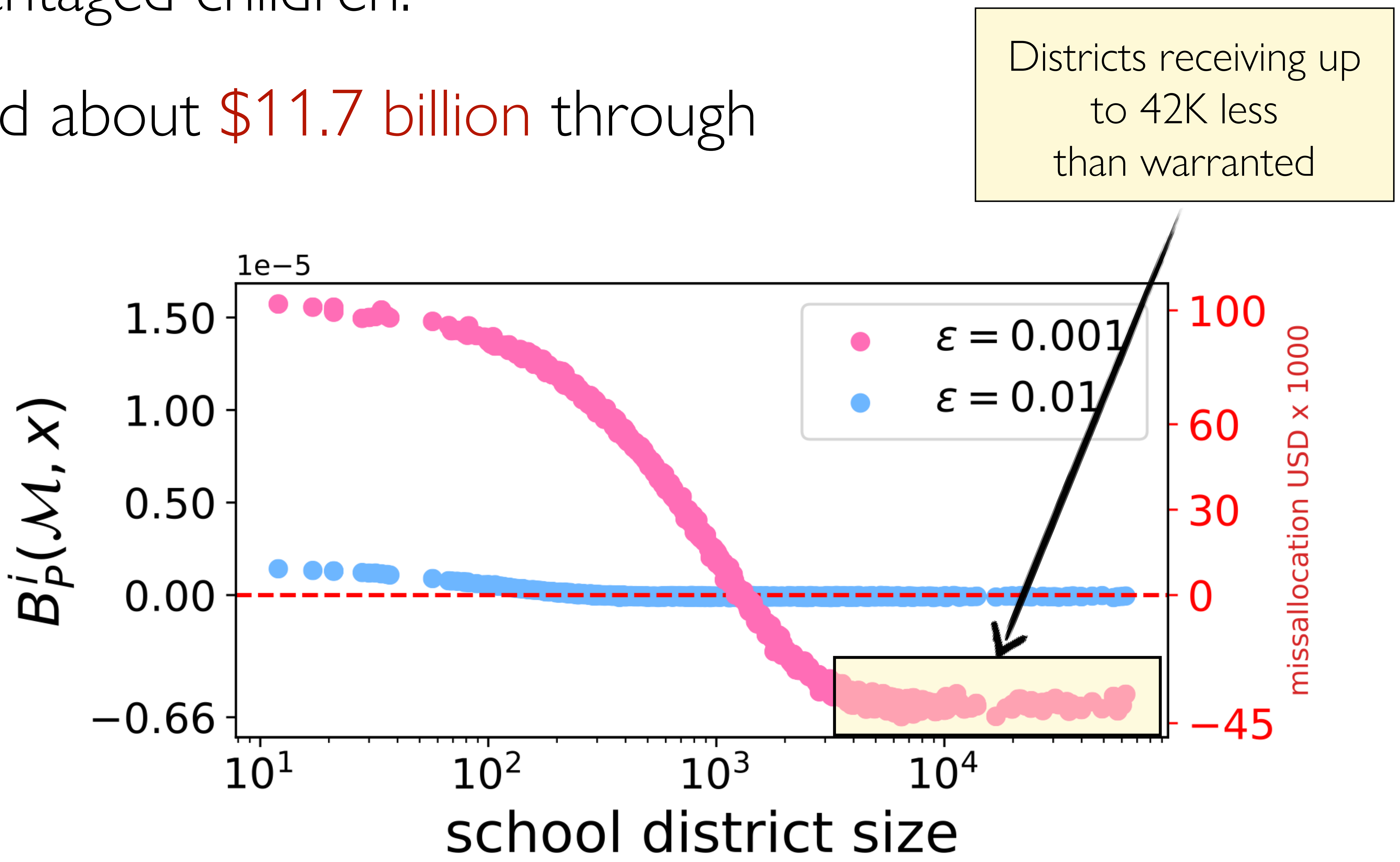
Title 1 allotment

- Title 1 of the Elementary and Secondary Education Act is one of the largest U.S. program offering educational assistance to disadvantaged children.
- In the fiscal year 2021 alone, it distributed about $11.7 billion through several types of grants.
- **Allotment:**

count of children 5 to 17 in district i

$$P_i^F(\boldsymbol{x}) \stackrel{\text{def}}{=} \left(\frac{\boxed{x_i} \cdot a_i}{\sum_{i \in [n]} x_i \cdot \boxed{a_i}} \right)$$

student expenditures in district i



Shape of the decision problem

First key result

- **Theorem (informal):** It is the "shape" of the decision problem that characterizes the unfairness of the outcomes, even using an unbiased DP mechanism.

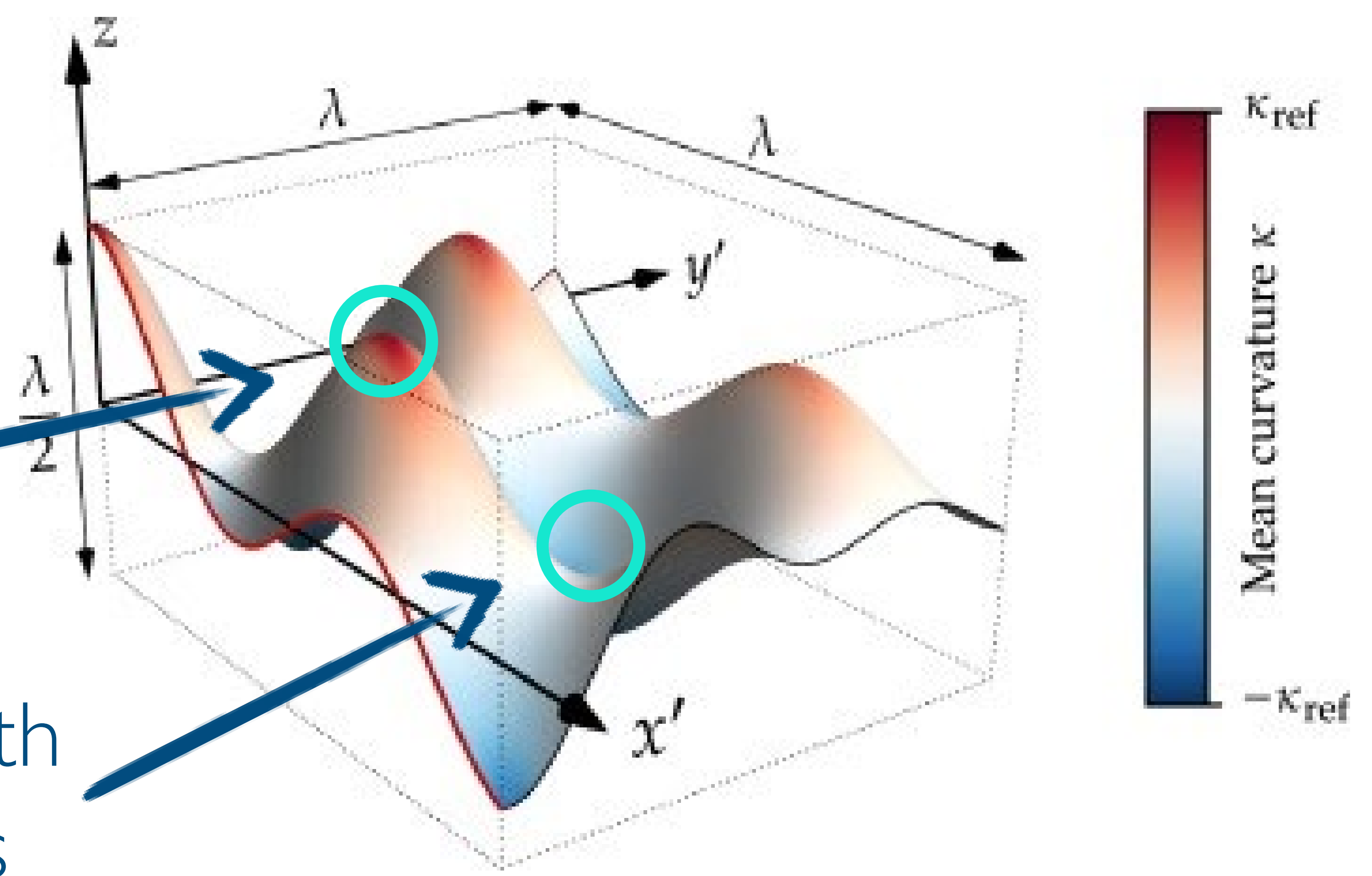
- The problem bias can be approximated as (when P_i is at least twice differentiable):

$$B_P^i(\mathcal{M}, \boldsymbol{x}) = \mathbb{E}[P_i(\tilde{\boldsymbol{x}} = \boldsymbol{x} + \eta)] - P_i(\boldsymbol{x})$$
$$\approx \frac{1}{2} \boldsymbol{H} P_i(\boldsymbol{x}) \times \mathrm{Var}[\eta]$$

Local curvature of problem Pi

Variance of the noisy input (depends on ϵ)

entities with high errors



entities with low errors

- Fairness can be bounded whenever the problem local curvature is constant across entities, since the variance is also constant and bounded.

Shape of the decision problem

First key result

- **Theorem (informal):** It is the "shape" of the decision problem that characterizes the unfairness of the outcomes, even using an unbiased DP mechanism.

- The problem bias can be approximated as (when P_i is at least twice differentiable):

$$B_P^i(\mathcal{M}, \boldsymbol{x}) = \mathbb{E}[P_i(\tilde{\boldsymbol{x}} = \boldsymbol{x} + \eta)] - P_i(\boldsymbol{x})$$
$$\approx \frac{1}{2} \boldsymbol{H} P_i(\boldsymbol{x}) \times \mathrm{Var}[\eta]$$

Local curvature of problem Pi

Variance of the noisy input (depends on ϵ)

A data release mechanism M is α-fair w.r.t. P, for some finite α, if for all datasets x, exists constants $c_{jl}^i \in \mathbb{R}, (i \in [n], j, l \in [k])$

$$(\boldsymbol{H}P_i)_{j,l}(\boldsymbol{x}) = c_{j,l}^i \quad (i \in [n]\ j, l \in [k]).$$

- Corollary: (Perfect)-fairness cannot be achieved if P is any non-linear function, as in the case of the allocations considered.

Fioretto al. IJCAI:2021   Tran al. NeurIPS:2022

Disproportionate impacts in downstream decisions

Minority language voting rights

- The ***Voting Rights Act*** of 1965 provides a body of protections for racial and language minorities.
- Section 203 describes the conditions under which local jurisdictions must provide minority language voting assistance during an election.
- Jurisdiction i must provide language assistance (including voter registration, ballots, and instructions) iff decision rule $P_i^M(x)$returns true with:

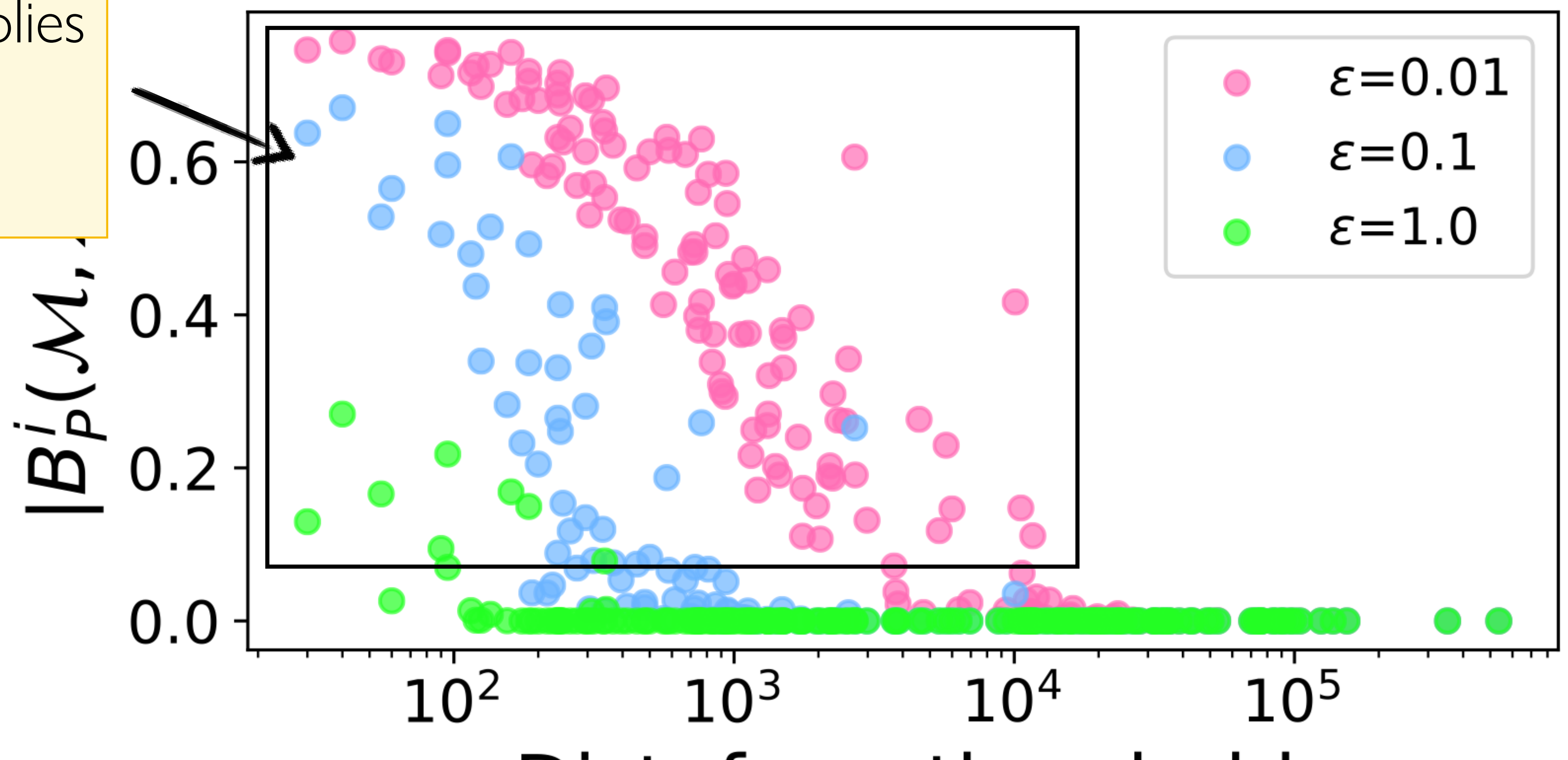
Misclassification implies potentially disenfranchising

+ < 5th grade education

$$P_i^M(\boldsymbol{x}) \stackrel{\text{def}}{=} \left(\frac{x_i^{\boxed{sp}}}{x_i^{\boxed{s}}} > 0.05 \vee x_i^{sp} > 10^4 \right) \wedge \frac{x_i^{\boxed{spe}}}{x_i^{sp}} > 0.0131.$$

no. of ppl in i speaking minority language s

+ limited English proficiency



sorted x^s

Fairness composition

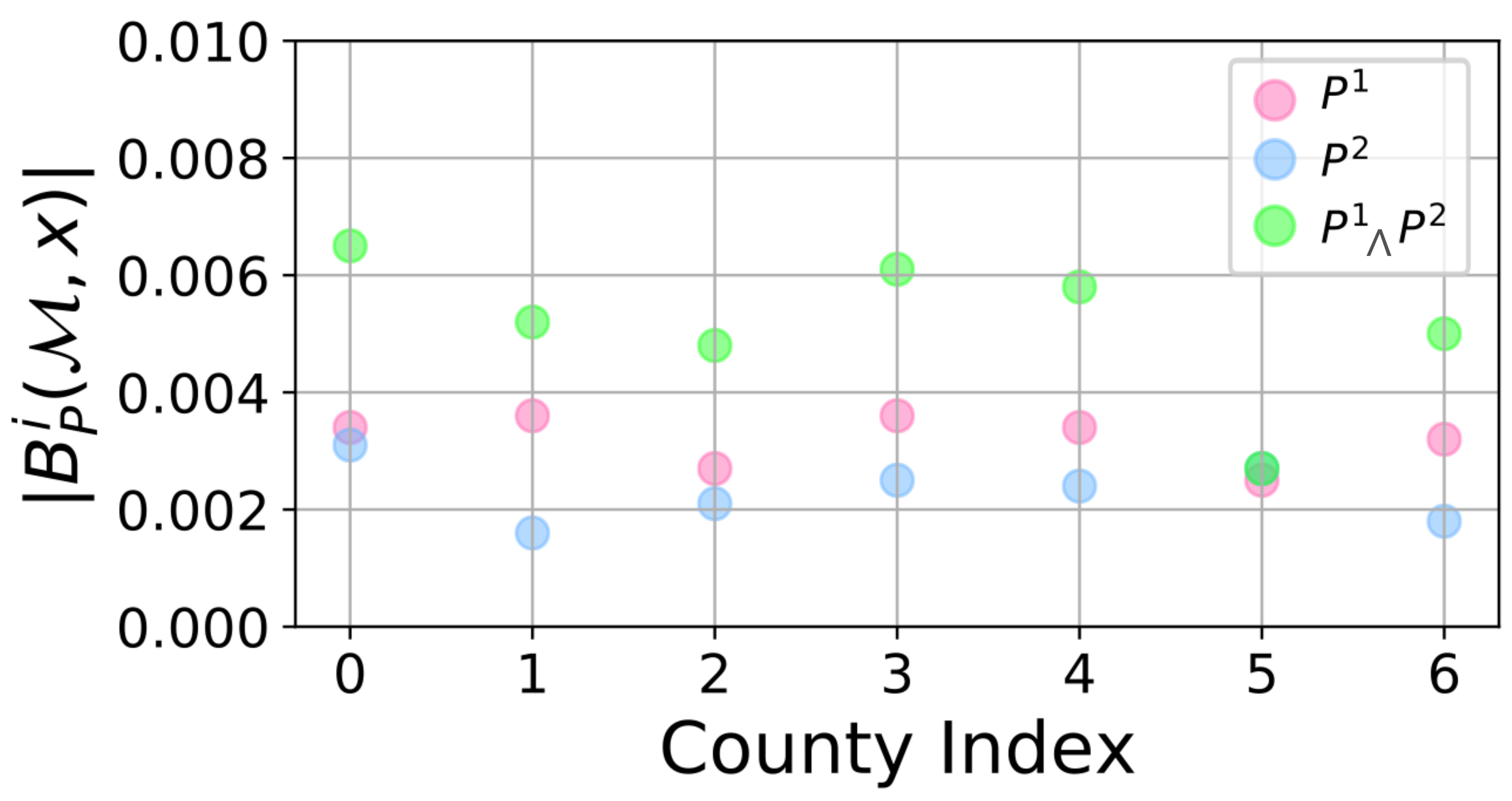
Second key result

$$P_i^M(\boldsymbol{x}) \stackrel{\text{def}}{=} \left(\frac{x_i^{sp}}{x_i^{s}} > 0.05 \lor \boxed{x_i^{sp} > 10^4} \right) \land \boxed{\frac{x_i^{spe}}{x_i^{sp}} > 0.0131.}$$

$$P^1(x^{sp}) = \mathbb{1}\{x^{sp} \geq 10^4\}$$

$$P^2(x^{sp}, x^{spe}) = \mathbb{1}\{\frac{x^{spe}}{x^{sp}} > 0.0131\}$$

Minority Language Voting Rights



- Small bias when considered individually
- However, when they are combined using logical connector ∧, the resulting absolute bias increases substantially, as illustrated by the associated green circles.

- **Theorem (informal)**: The logical composition of two α_1- and α_2-fair mechanisms is α-fair with $\alpha \geq max(\alpha_1, \alpha_2)$.
- The unfairness induced by "composing" predicates is no smaller than that of their individual components.

Shape of the decision problem

Important conclusion

*Using DP to generate private inputs of decision problems commonly adopted to make policy determination **will necessarily introduce fairness issues**, despite the noise being unbiased!*

Mitigation solution

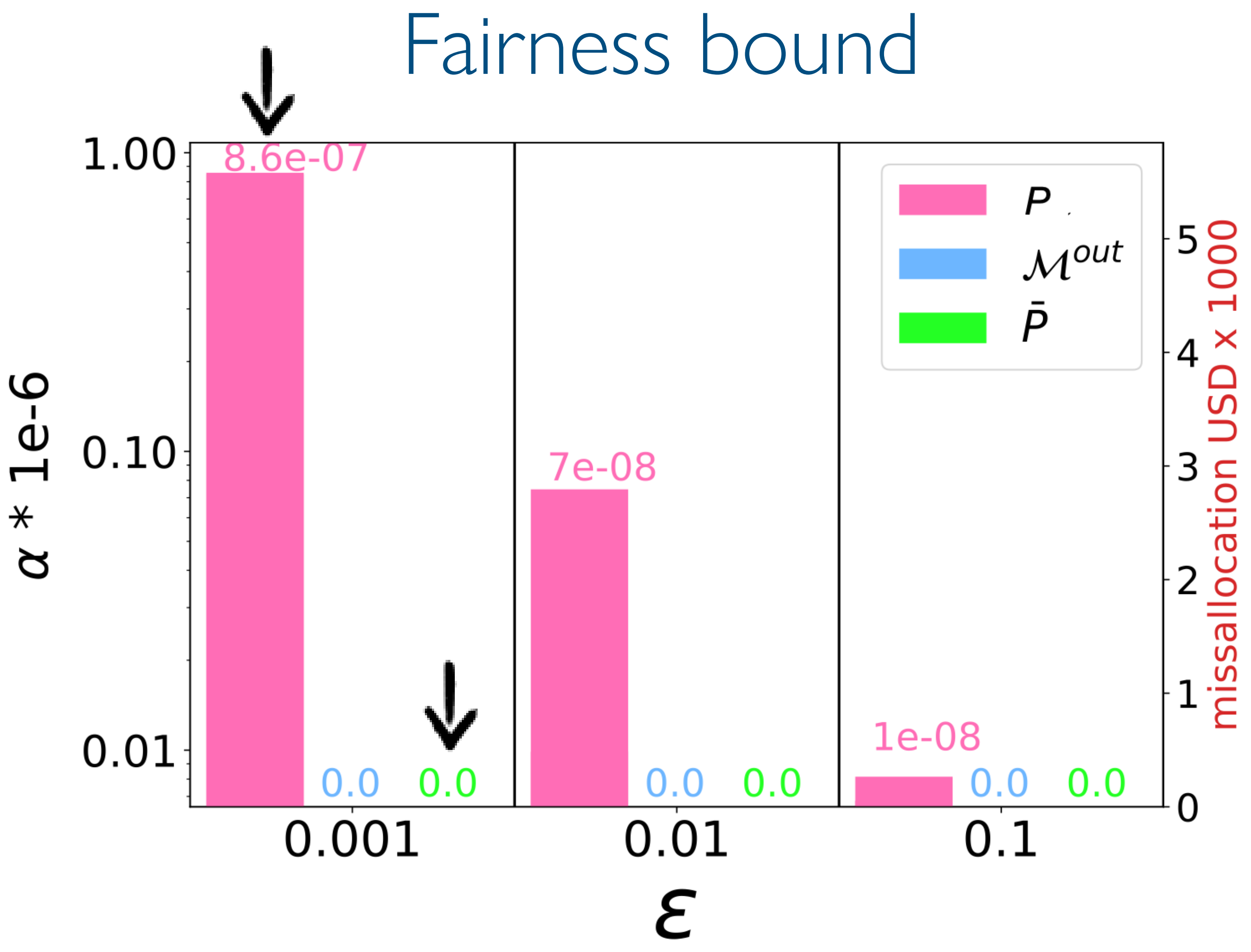
Fair allocations

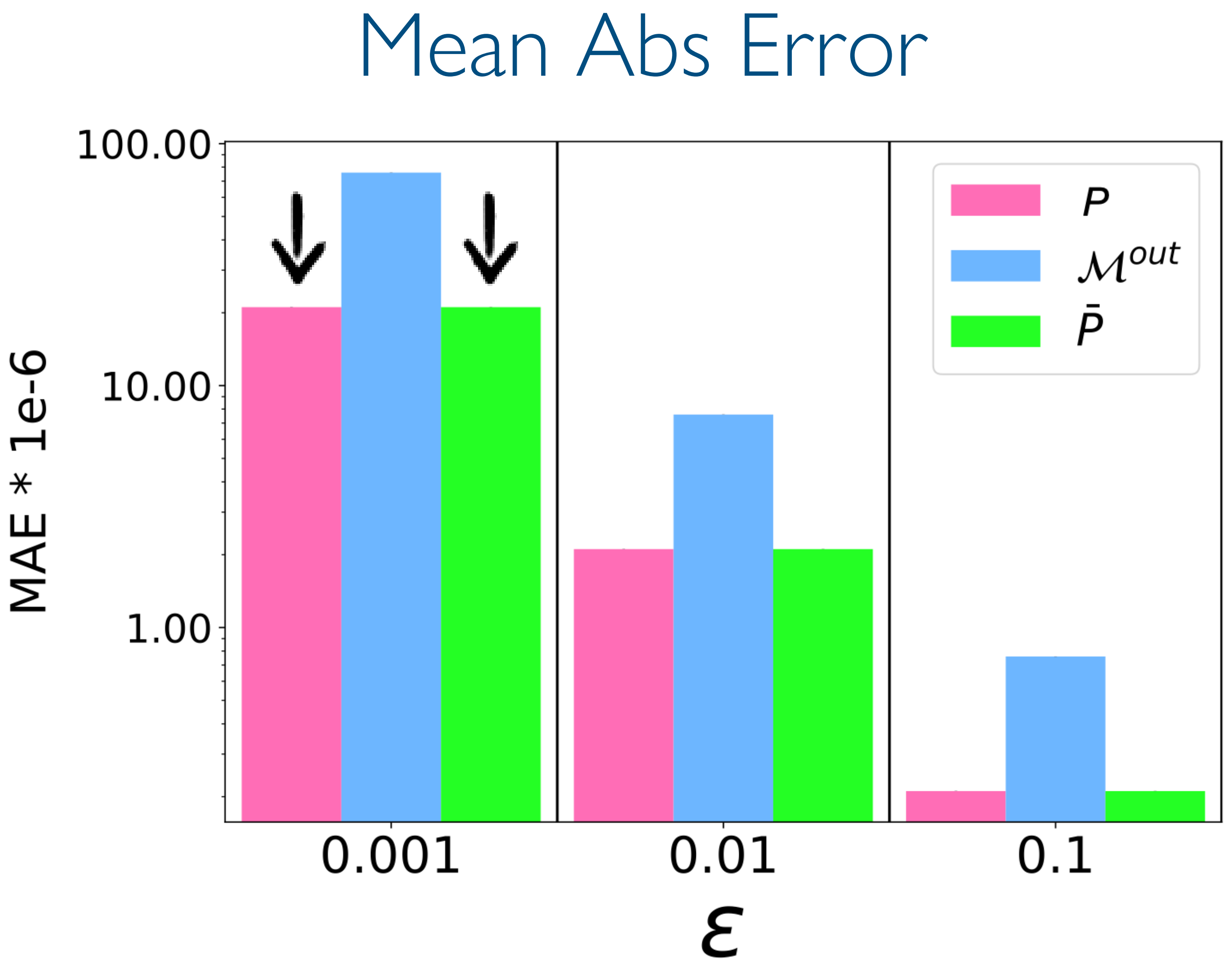
- Note that the observed issues are not data-driven, but problem-driven.
- **Corollary**: If P is a linear function, then mechanism M is fair w.r.t. P.
- Linearizing the allotment problem — General idea: Given a problem P_i derive a linear approximation $\tilde{P}_i$ of P_i

Redundant data release

$$P_i^F(\boldsymbol{x}) \stackrel{\text{def}}{=} \left(\frac{x_i \cdot a_i}{\boxed{\sum_{i\in[n]} x_i \cdot a_i}}\right)$$

Release its (noisy) version as a constant

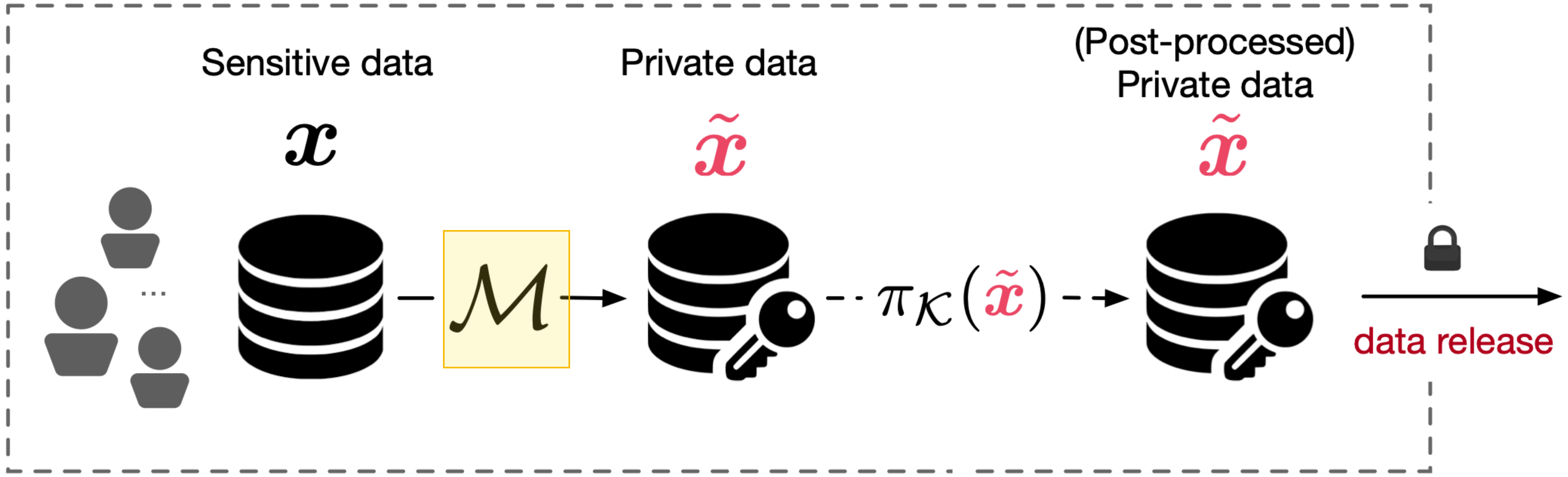




DP Post-processing

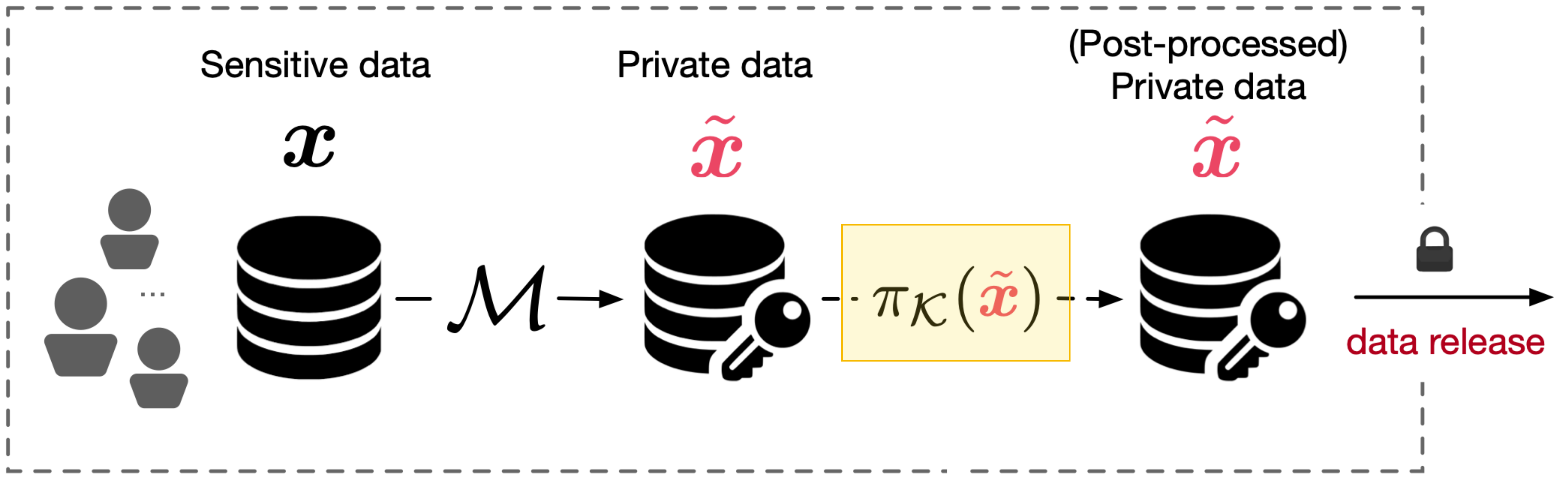
Fairness impact

DP data release with post-processing



1. Apply noise with appropriate parameter $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \text{Noise}$

DP data release with post-processing

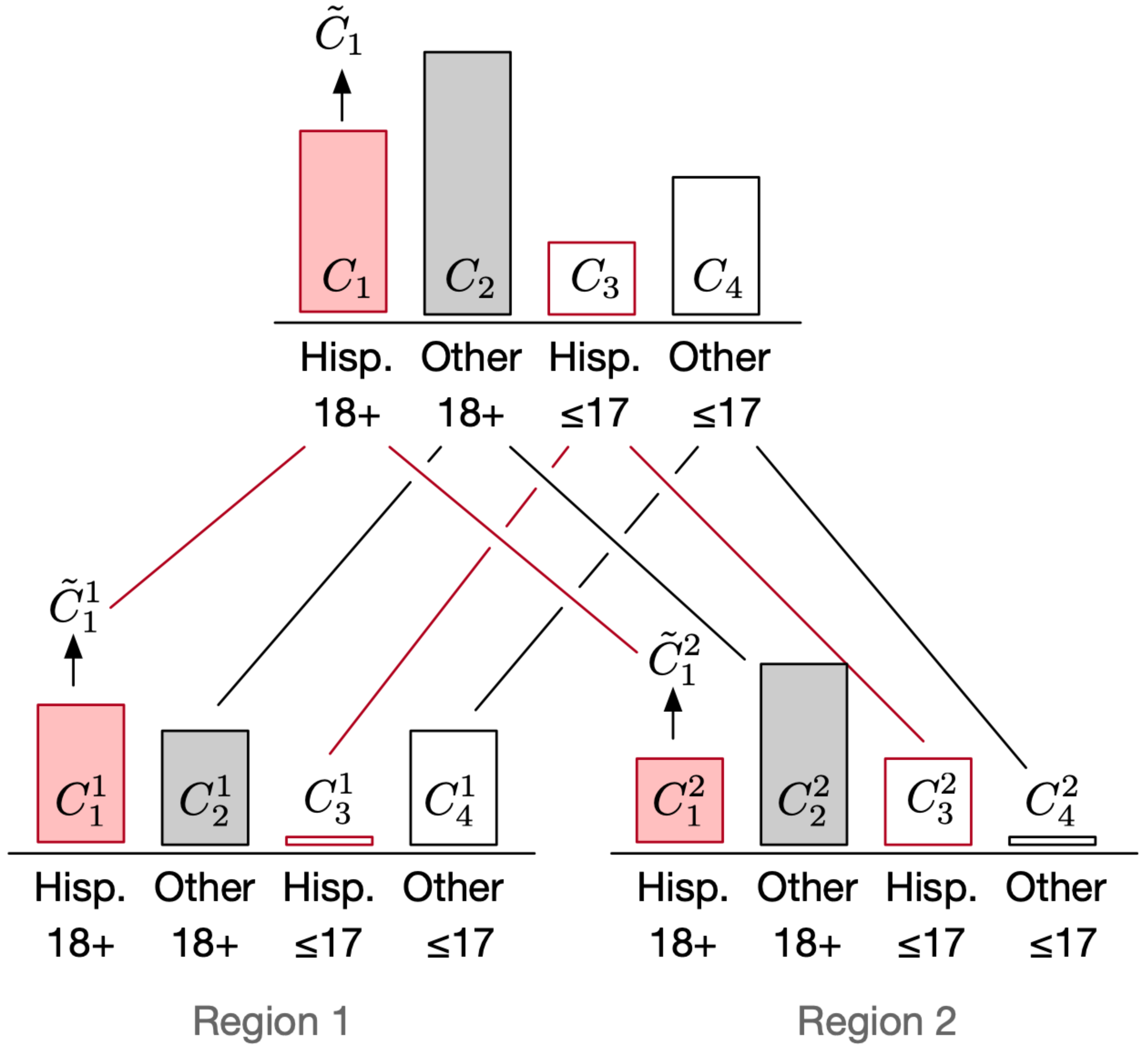


1. Apply noise with appropriate parameter $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \text{Noise}$
2. Post-process output $\tilde{\boldsymbol{x}}$ to enforce consistency

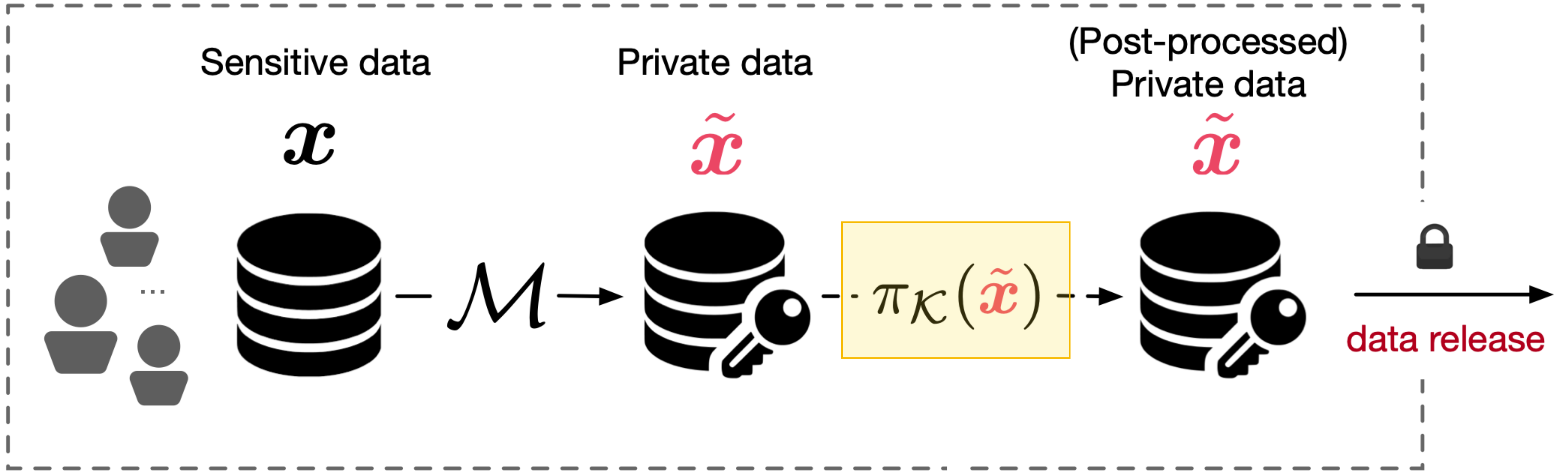
$$\pi_{\mathcal{K}}(\tilde{\boldsymbol{x}}) : \operatorname*{argmin}_{\boldsymbol{v} \in \mathcal{K}} \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2$$

with feasible region defined as

$$\mathcal{K} = \left\{ \boldsymbol{v} \mid \sum_{i=1}^{n} v_i = C, \boldsymbol{v} \geq 0 \right\}$$



DP data release with post-processing



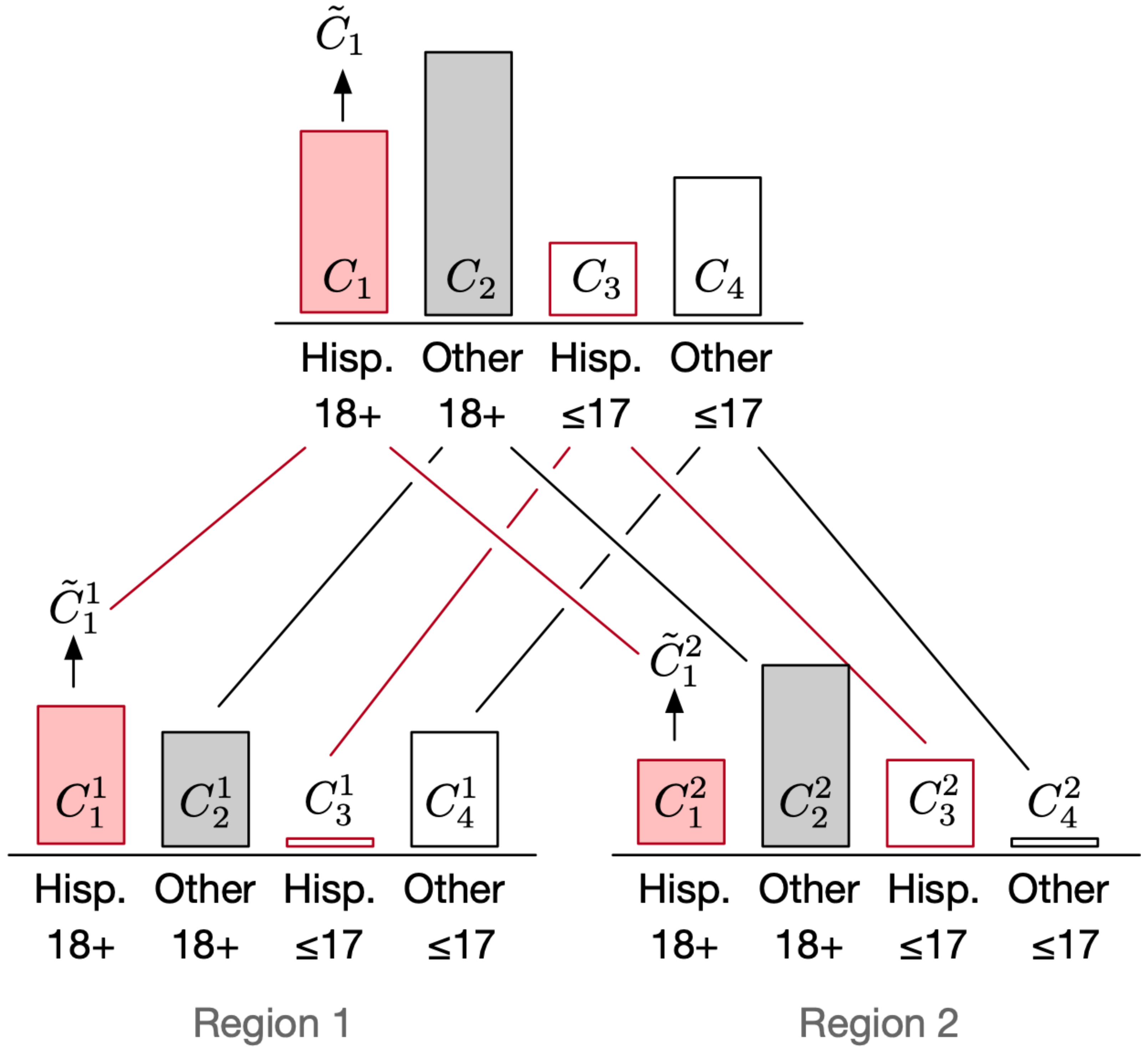
1. Apply noise with appropriate parameter $\quad \tilde{\boldsymbol{x}} = \boldsymbol{x} + \text{Noise}$

2. Post-process output $\tilde{\boldsymbol{x}}$ to enforce consistency

$$\pi_{\mathcal{K}}(\tilde{\boldsymbol{x}}) : \operatorname*{argmin}_{\boldsymbol{v} \in \mathcal{K}} \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2$$

with feasible region defined as

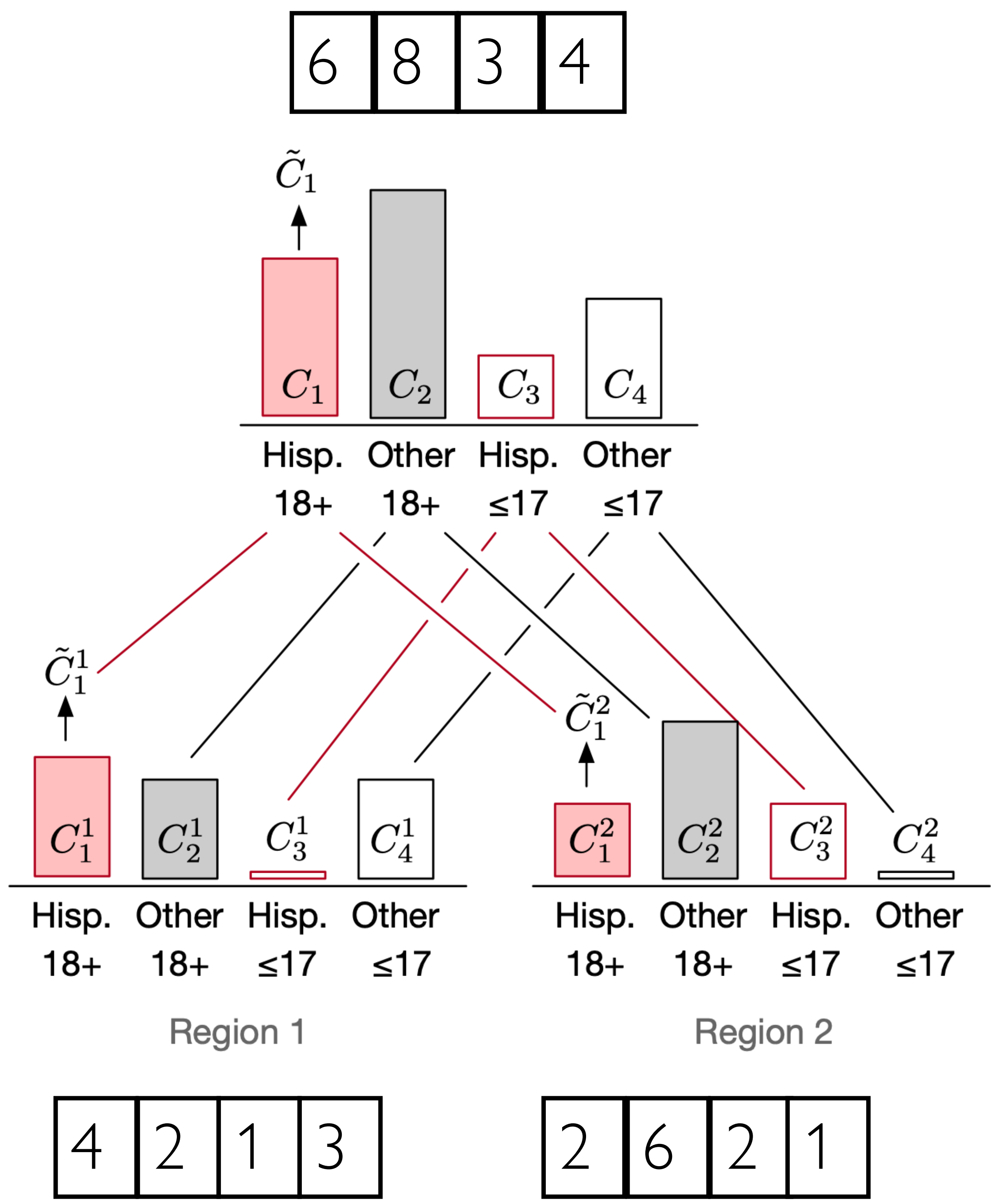
$$\mathcal{K} = \left\{ \boldsymbol{v} \mid \sum_{i=1}^{n} v_i = C, \boldsymbol{v} \geq 0 \right\}$$



Satisfies DP due to post-processing immunity

DP post-processing

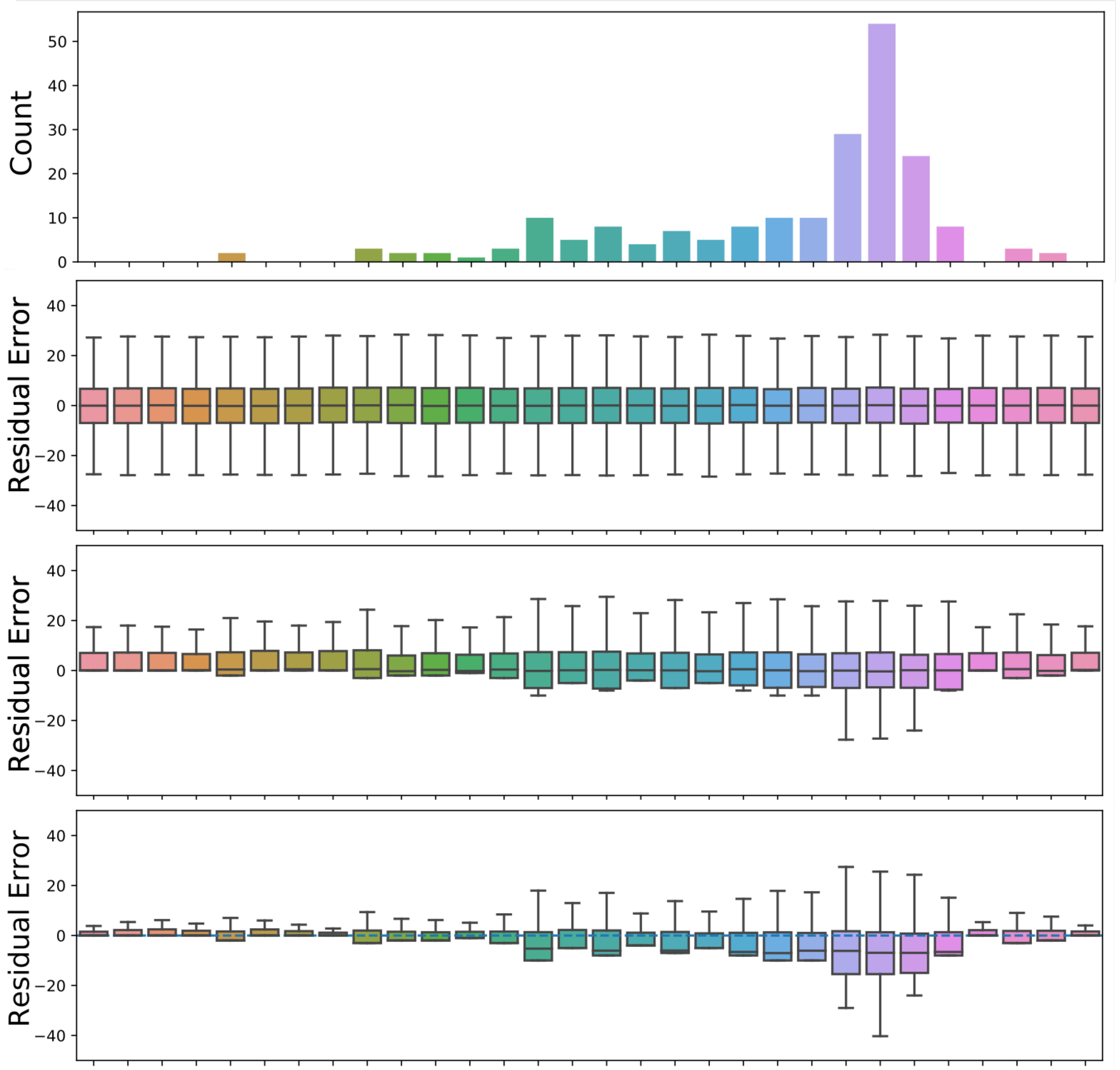
Error and bias



Laplace mechanism

$$\pi_{\geq 0} := \underset{v \geq 0}{\operatorname{argmin}} \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2$$

$$\pi_{\mathcal{K}_S} := \underset{v \in \mathcal{K}_S}{\operatorname{argmin}} \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2 \ , \ \mathcal{K}_S = \{\boldsymbol{v} \in \mathbb{R}^n \mid \sum_i v_i = \tilde{S}, v_i \geq 0\},$$



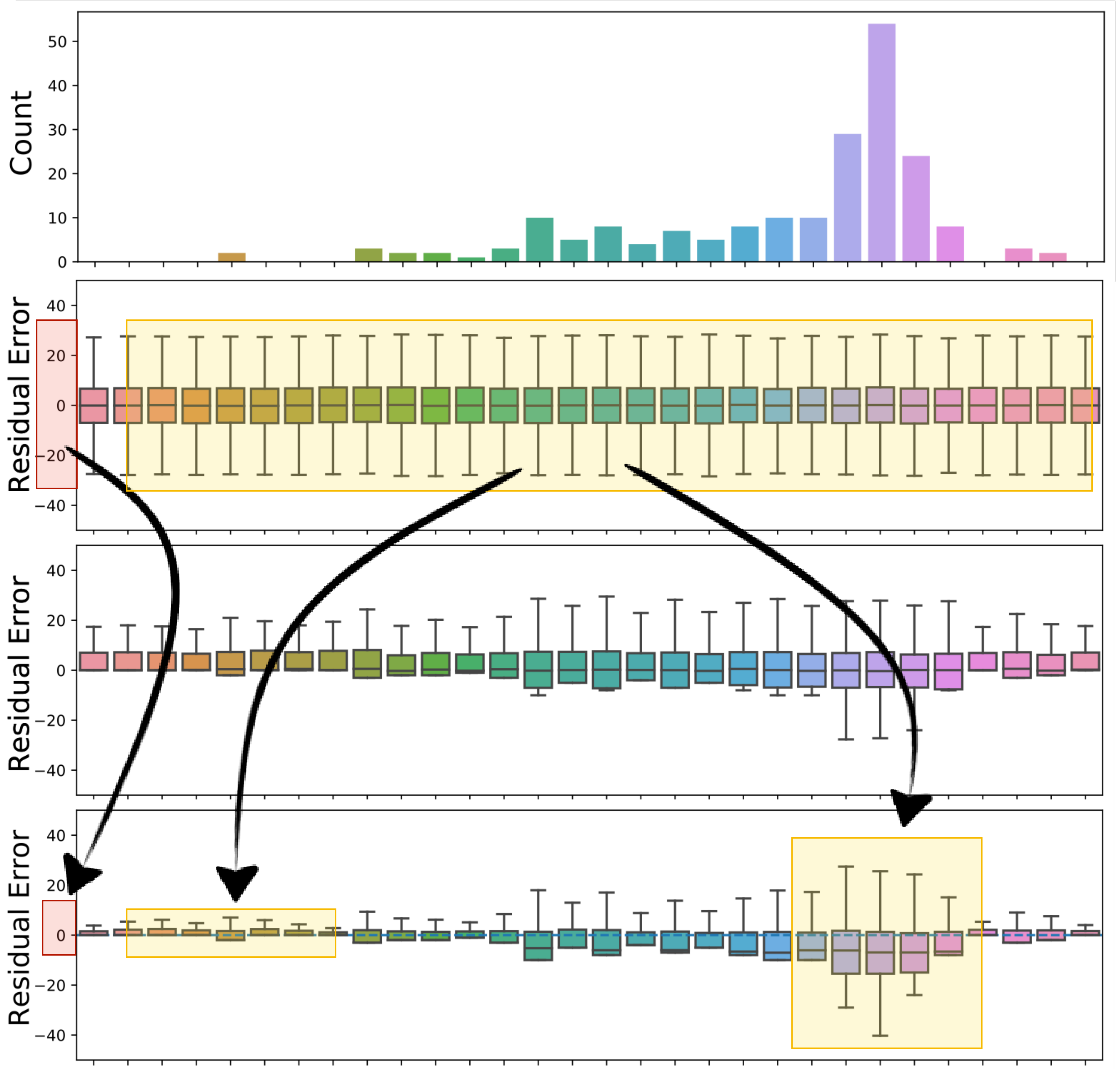
DP post-processing

Error and bias

Observe that post-processing reduces the errors.

However, it increases unfairness!

Laplace mechanism



$$\pi_{\geq 0} := \underset{v \geq 0}{\mathrm{argmin}} \, \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2$$

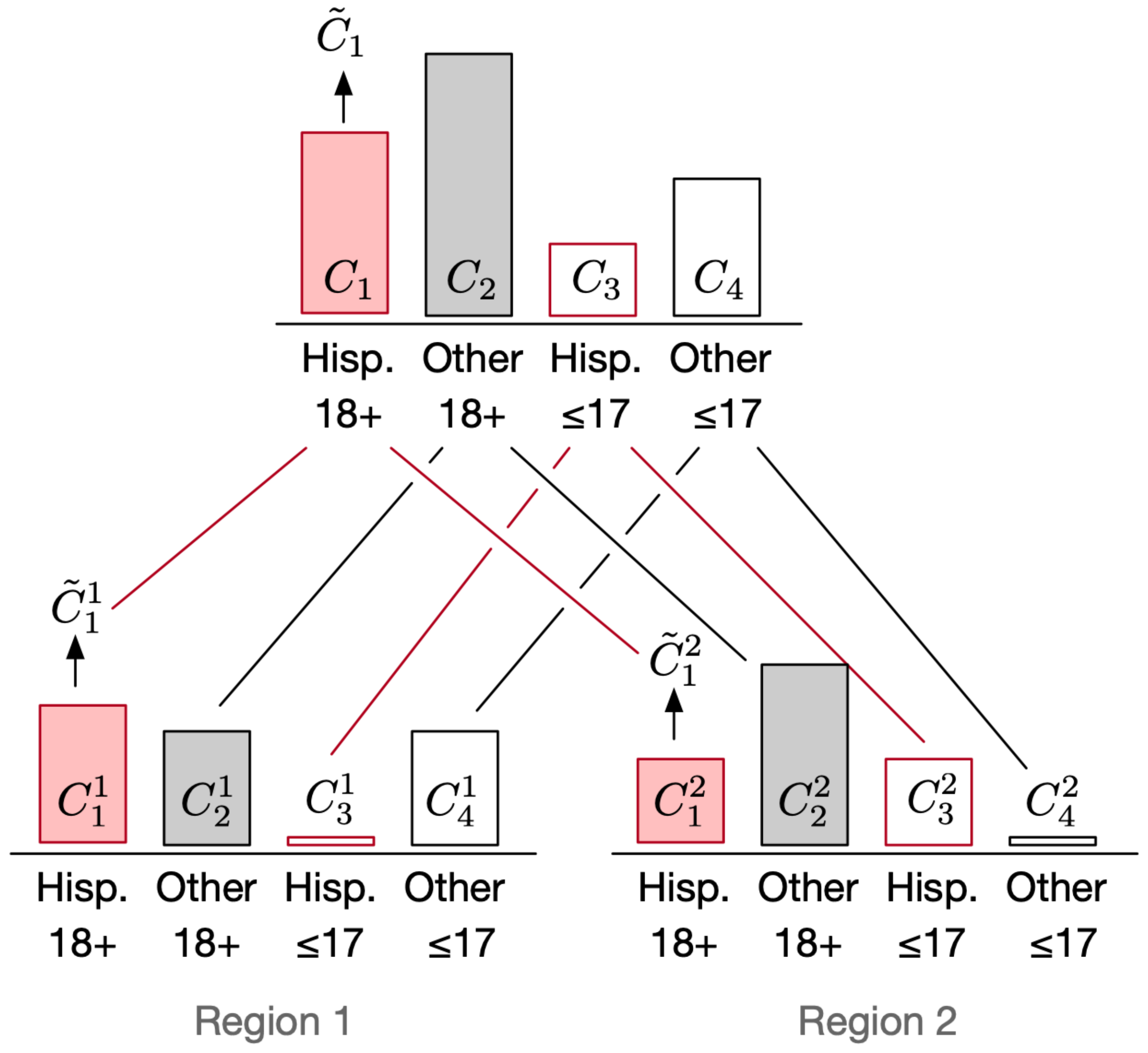
$$\pi_{\mathcal{K}_S} := \underset{\boldsymbol{v} \in \mathcal{K}_S}{\mathrm{argmin}} \, \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2 \, , \; \mathcal{K}_S = \{\boldsymbol{v} \in \mathbb{R}^n \mid \sum_i v_i = \tilde{S}, v_i \geq 0\},$$

Zhu et al. AAAI:2021 Zhu et al. IJCAI:2022, Fioretto et al. AAAI:2024

Bias of post-processing

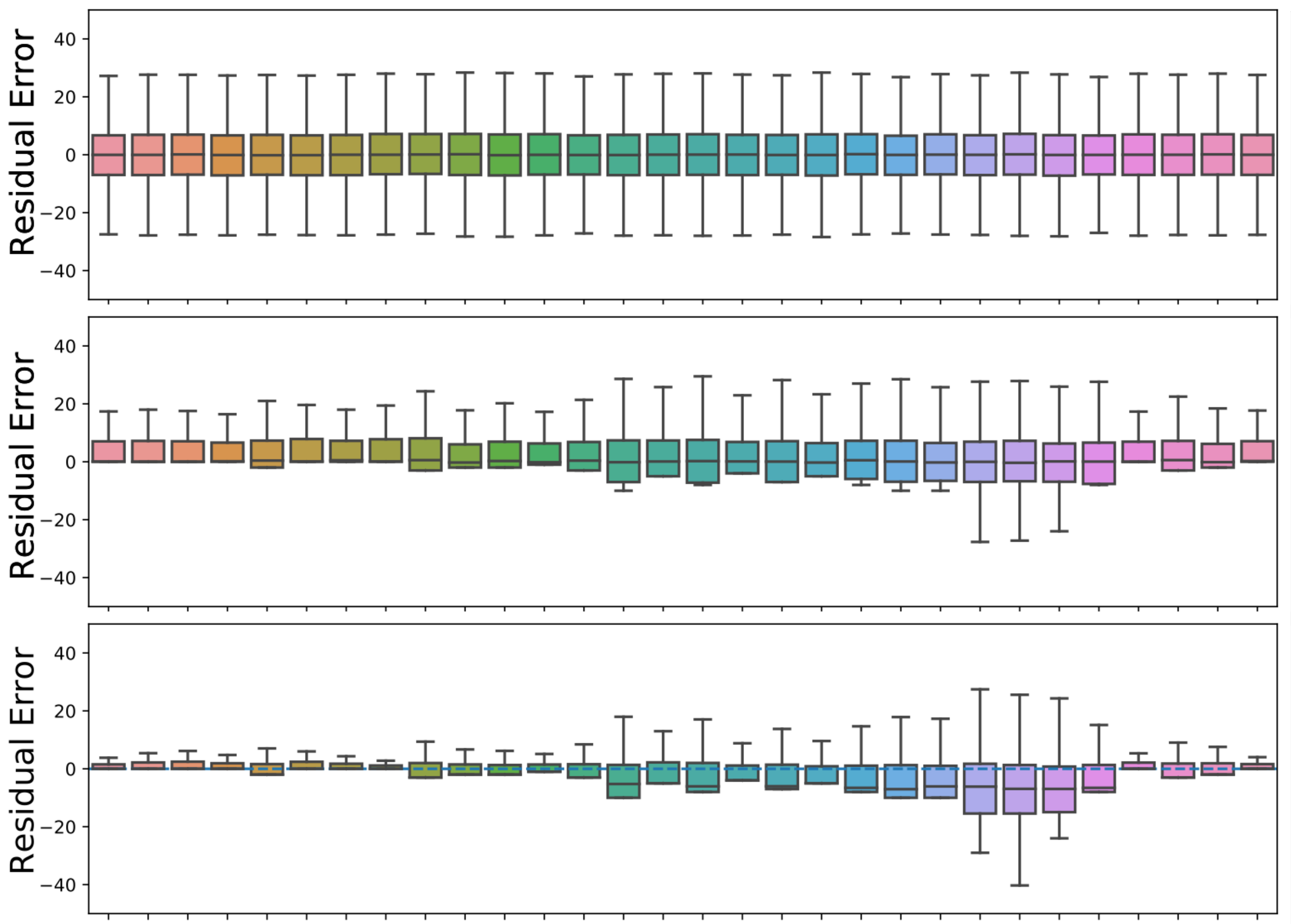
Key result

- **Thm (informal):** The bias is caused by the presence of non-negativity constraints!



$$\pi_{\geq 0} := \underset{v \geq 0}{\operatorname{argmin}} \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2$$

$$\pi_{\mathcal{K}_S} := \underset{\boldsymbol{v} \in \mathcal{K}_S}{\operatorname{argmin}} \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2 \text{ , } \mathcal{K}_S = \{\boldsymbol{v} \in \mathbb{R}^n \mid \sum_i v_i = \tilde{S}, v_i \geq 0\},$$



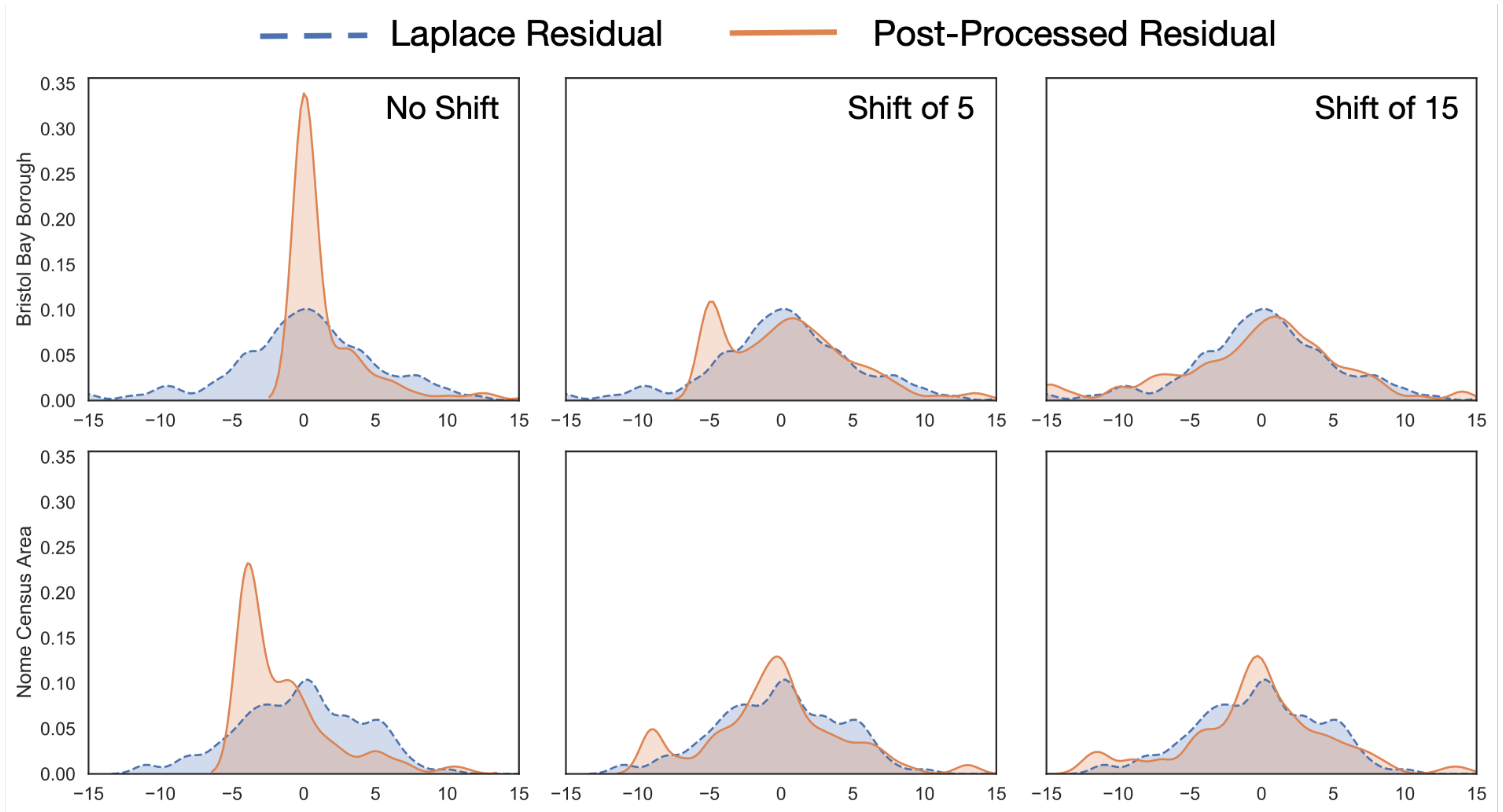
Zhu et al. AAAI:2021 Zhu et al. IJCAI:2022, Fioretto et al. AAAI:2024

Quantifying bias in post-processing

Theorem: *Suppose that the noisy data $\tilde{x}$ is the output of the Laplace mechanism with scale λ. The bias of the post-processed solution $\pi_{\mathcal{K}^+}$ of program (L^+) is bounded, in l_∞ norm, by*

$$\|B_{L^+}(\mathcal{M}, \boldsymbol{x})\|_\infty = \left\|\mathbb{E}_{\tilde{\boldsymbol{x}} \sim \mathcal{M}(\boldsymbol{x})}\left[\pi_{L^+}(\tilde{\boldsymbol{x}}) - \boldsymbol{x}\right]\right\|_\infty \leq C' \cdot \exp\left(\frac{-r_m}{\lambda}\right) \cdot \sum_{i=0}^{n-1} \frac{(\boxed{r_m})^i}{i! \cdot \lambda^i},$$

where C' represents the value $\sup_{\boldsymbol{v} \in \mathcal{K}^+} \|\boldsymbol{v} - \boldsymbol{x}\|_\infty$, which is finite due to the boundedness of the feasible region $\mathcal{K}^+$.



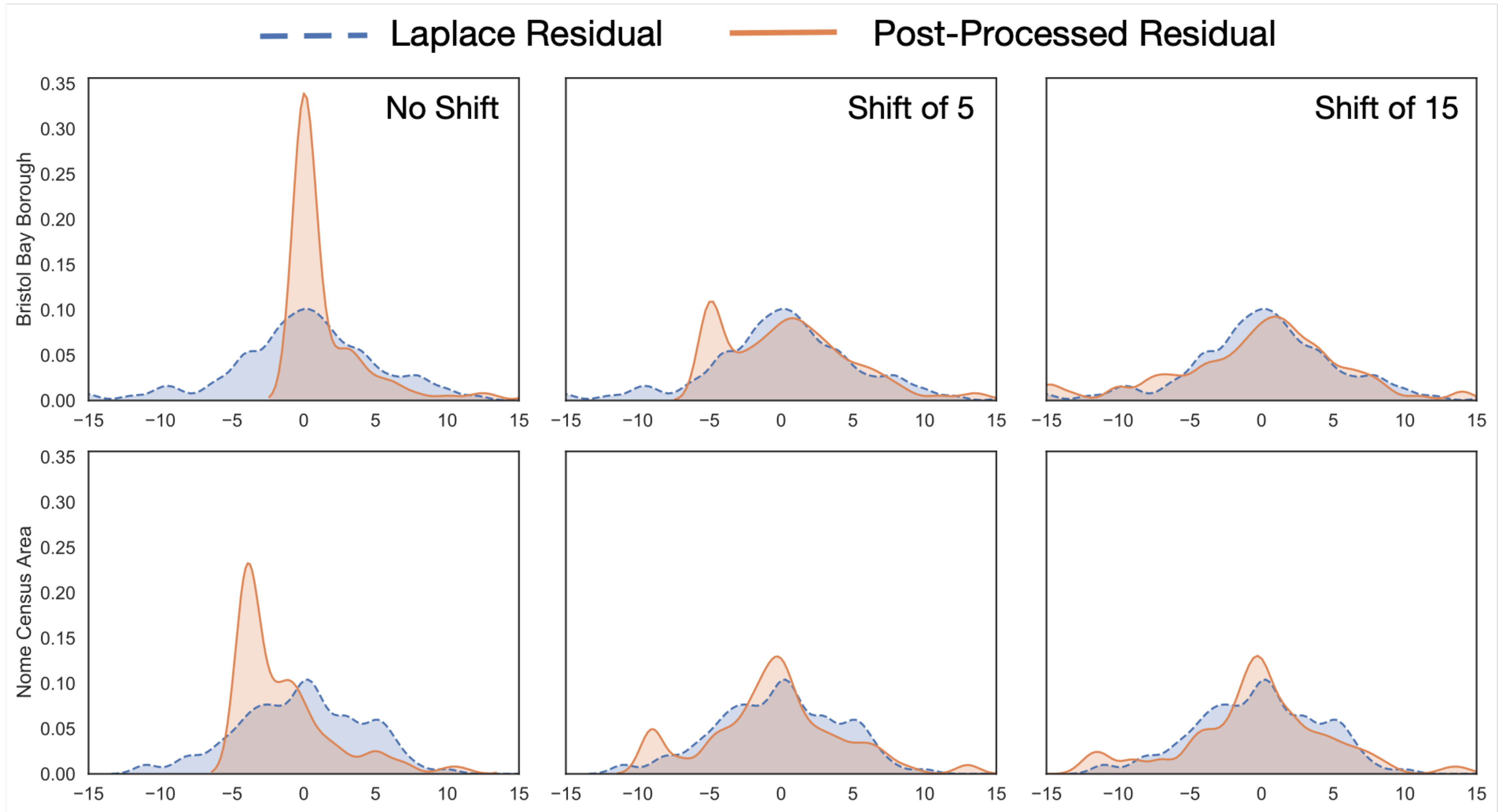
There is an ℓ_1-ball of radius $r_m = \min_i x_i$ and centered in x which is a feasible subspace where there is no bias.

Shifting increases the value of r_m and the bias progressively disappear.

Zhu et al. AAAI:2021 Zhu et al. IJCAI:2022,

Practical considerations

- Post-processing reduces the variance of the noise differently in different "regions". Regions with many subregions (e.g., counties, census blocks, etc.) will have more variance than regions with few subregions.
- It creates situations where counties will be treated fundamentally differently in decision processes.



Aggregating the counts for	Variance
Arizona (pop: 2.37ML in15 counties)	186.67
Texas (pop: 8.89ML in 254 counties)	200.01

~6.5% difference
which may affect allocations!

Zhu et al. AAAI:2021 Zhu et al. IJCAI:2022,

DP post-processing

Important conclusion

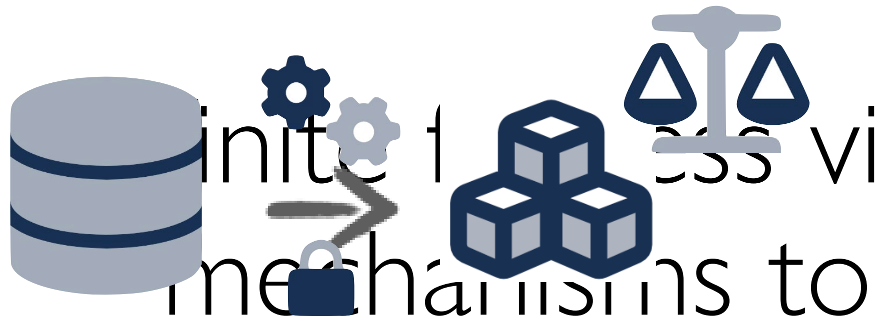
Although post-processing reduces errors, its application to policy determinations should take into account fairness issues.

Conclusions

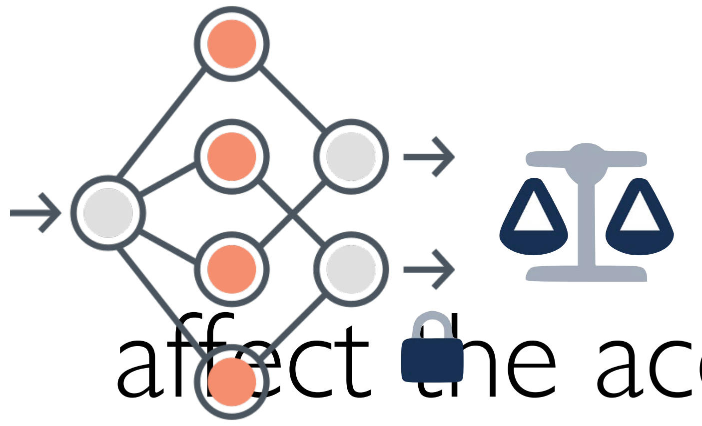
Unintended effects of DP on decisions and learning tasks

- Motivated by the use of rich datasets combined with black-box algorithms

- Proved that several problems with significant societal impacts (allocation of funding, language assistance) exhibit inherent unfairness when applied to a DP release of the census data.



inite ... vic ... mechanisms to r ... ave



affect the accura ... y

Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey

Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, Keyu Zhu

Watch video

IJCAI ECAI

- Exciting research direction that requires close cooperation between multiple areas and can transform the way we approach ML and decision making to render these algorithms more aligned with societal values.

The unintended disparate effects of privacy in decision tasks

Thank you!

Ferdinando Fioretto University of Virginia

https://nandofioretto.com

nandofioretto@gmail.com

@nandofioretto

References (1/2)

"Differential Privacy of Hierarchical Census Data: An Optimization Approach".
Ferdinando Fioretto, Pascal Van Hentenryck, Keyu Zhu. In Artificial Intelligence (AIJ), 2021.

"Differentially Private Empirical Risk Minimization under the Fairness Lens".
Cuong Tran, My H. Dinh, Ferdinando Fioretto. In Conference on Neural Information Processing Systems (NeurIPS), 2021.

"Decision Making with Differential Privacy under the Fairness Lens".
Cuong Tran, Ferdinando Fioretto, Pascal Van Hentenryck, Zhiyan Yao. In International Joint Conference on Artificial Intelligence (IJCAI), 2021.

"Bias and Variance of Post-processing in Differential Privacy".
Keyu Zhu, Pascal Van Hentenryck, Ferdinando Fioretto. In AAAI Conference on Artificial Intelligence (AAAI), 2021.

“A Fairness Analysis on Private Aggregation of Teacher Ensembles".
Cuong Tran, My H. Dinh, Kyle Beiter, Ferdinando Fioretto. CoRR abs/2109.08630 [cs.LG], 2021.

"Post-processing of Differentially Private Data: A Fairness Perspective".
Keyu Zhu, Ferdinando Fioretto, Pascal Van Hentenryck. In International Joint Conference on Artificial Intelligence (IJCAI), 2022.

"Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey".
Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, Keyu Zhu. In International Joint Conference on Artificial Intelligence (IJCAI), 2022.

See https://web.ecs.syr.edu/~ffiorett/publications.html for papers links.

References (2/2)

"A Fairness Analysis on Private Aggregation of Teacher Ensembles".
Cuong Tran, My H. Dinh, Kyle Beiter, Ferdinando Fioretto. In the Workshop of Privacy-Preserving Artificial Intelligence (PPAI)–at AAAI, 2022.

"Fairness Increases Adversarial Vulnerability".
Cuong Tran, Keyu Zhu, Ferdinando Fioretto, Pascal Van Hentenryck CoRR abs/2211.11835 [cs.LG], 2022.

"Pruning has a disparate impact on model accuracy".
Cuong Tran, Ferdinando Fioretto, Jung-Eun Kim, Rakshit Naidu. In Conference on Neural Information Processing Systems (NeurIPS), 2022.

"SF-PATE: Scalable, Fair, and Private Aggregation of Teacher Ensembles".
Cuong Tran, Keyu Zhu, Ferdinando Fioretto, Pascal Van Hentenryck. CoRR abs/2204.05157 [cs.LG], 2022.

"Personalized Privacy Auditing and Optimization at Test Time".
Cuong Tran, Ferdinando Fioretto. CoRR abs/2302.00077 [cs.LG], 2023.

"Privacy and Bias Analysis of Disclosure Avoidance Systems".
Keyu Zhu, Ferdinando Fioretto, Pascal Van Hentenryck, Saswat Das, Christine Task CoRR abs/2301.12204

See https://web.ecs.syr.edu/~ffiorett/publications.html for papers links.

DP Post-processing

Mitigating solution

Definition 4 (Projection onto Simplex Mechanism (PoS)). *The* projection onto simplex mechanism *outputs the allocation as follows.*

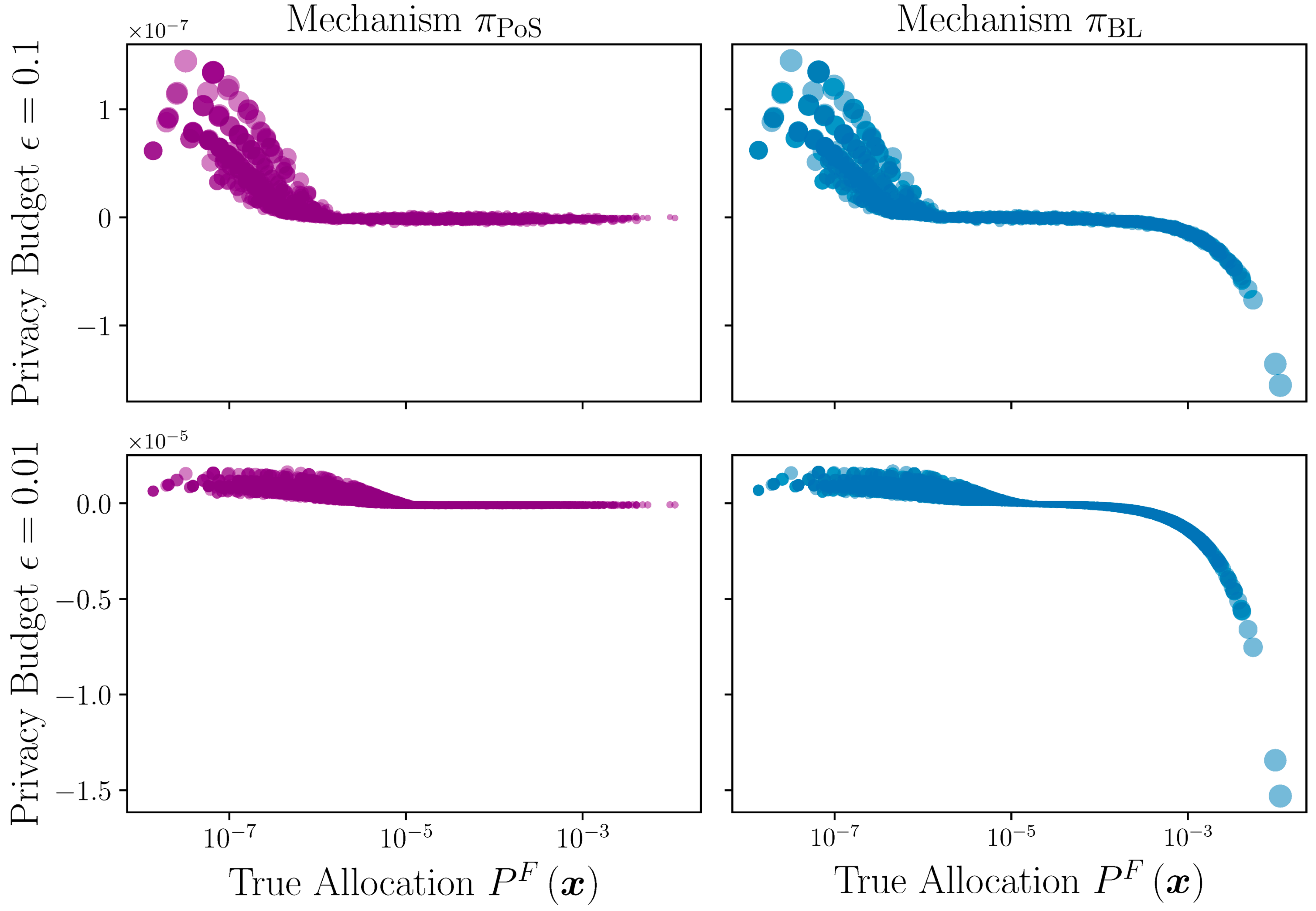
$$\pi_{\text{PoS}}(\tilde{\boldsymbol{x}}) := \underset{\boldsymbol{v} \in \Delta^n}{\arg\min} \left\| \boldsymbol{v} - P^F(\tilde{\boldsymbol{x}}) \right\|_2 \qquad (P_{\text{PoS}})$$

Theorem (informal). For any DP dataset $\tilde{\boldsymbol{x}}$ the PoS mechanism generates the unique optimal solution to program

$$\pi^*_\alpha(\tilde{\boldsymbol{x}}) := \underset{\boldsymbol{v} \in \Delta_n}{\arg\min} \left\| \boldsymbol{v} - P^F(\tilde{\boldsymbol{x}}) \right\|_{\rightleftharpoons} \qquad (P_\alpha)$$

which closely approximate the optimal post-processing mechanism

$$\pi^* := \min_{\pi \in \Pi_{\Delta_n}} \left\| \mathbb{E}_{\tilde{\boldsymbol{x}}} \left[\pi(\tilde{\boldsymbol{x}}) - P^F(\boldsymbol{x}) \right] \right\|_{\rightleftharpoons}$$



Privacy Budgets	$\epsilon = 0.1$		$\epsilon = 0.01$		$\epsilon = 0.001$	
Mechanisms	π_{BL}	π_{PoS}	π_{BL}	π_{PoS}	π_{BL}	π_{PoS}
α-fairness	3.00E-07	**1.50E-07**	1.70E-05	**1.75E-06**	8.06E-04	**2.23E-05**
Cost of Privacy	1.62E-05	**1.41E-05**	1.33E-03	**1.04E-03**	5.90E-02	**3.49E-02**