# The Unintended Effects of Privacy in Decision and Learning


James

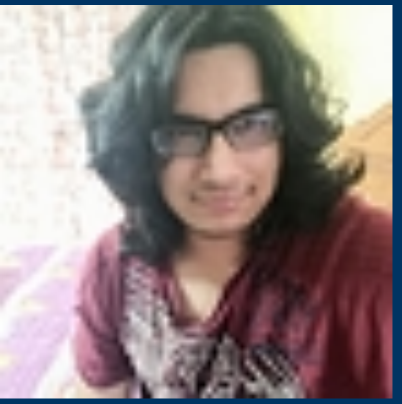

My


Vincenzo Vito


Jacob


Michael


Jinhao


Saswat


Key


Cuong

**Nando Fioretto** @S-HPC, 2024

https://nandofioretto.com

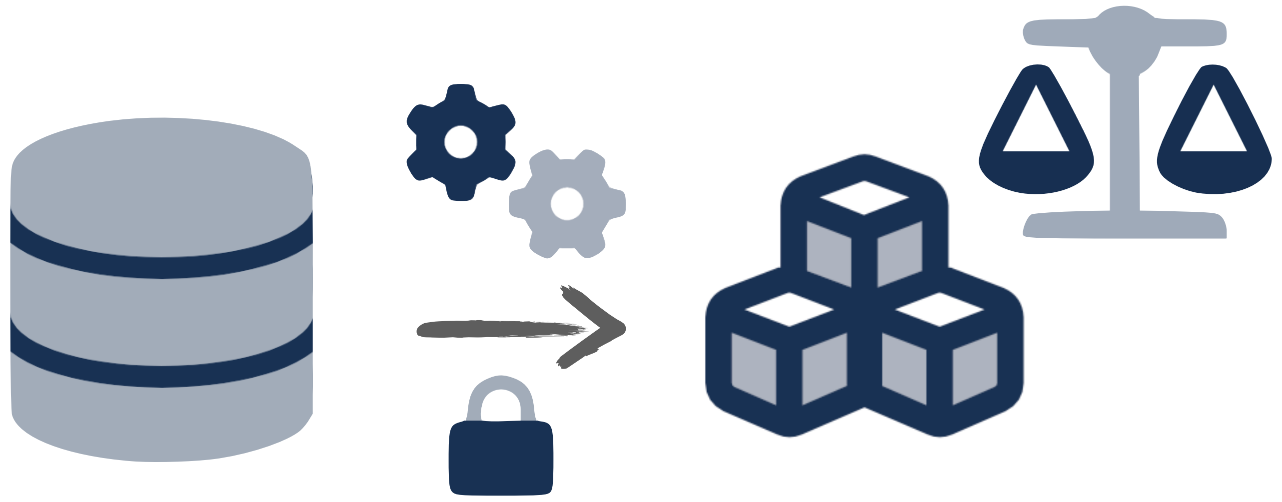
@nandofioretto

nandofioretto@gmail.com

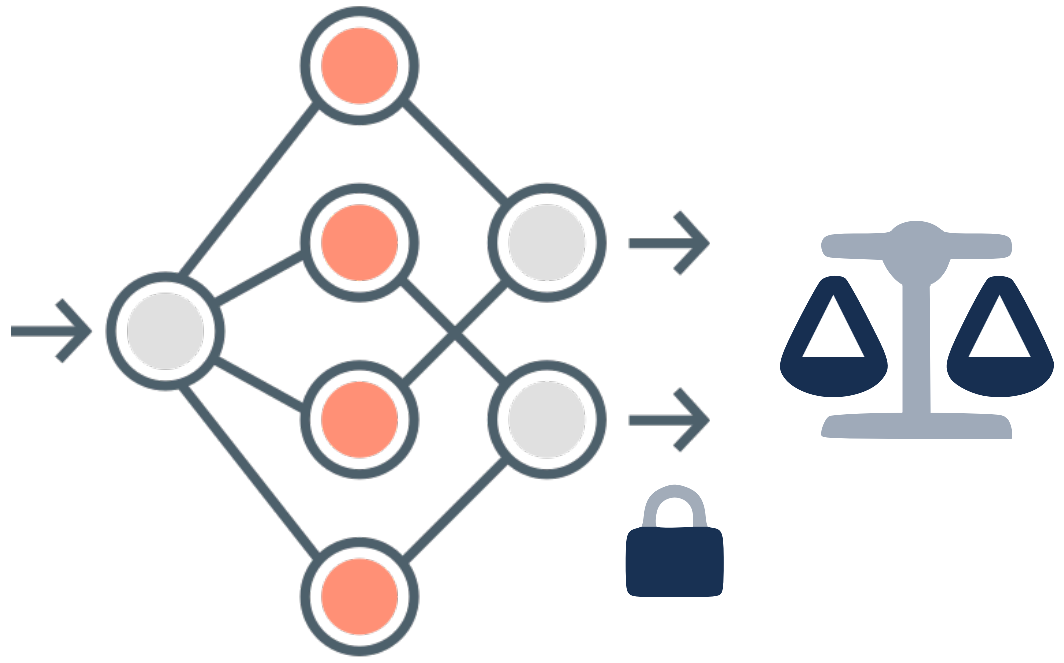
# Agenda

**Preliminaries**

**Fairness impacts of DP in decision making**



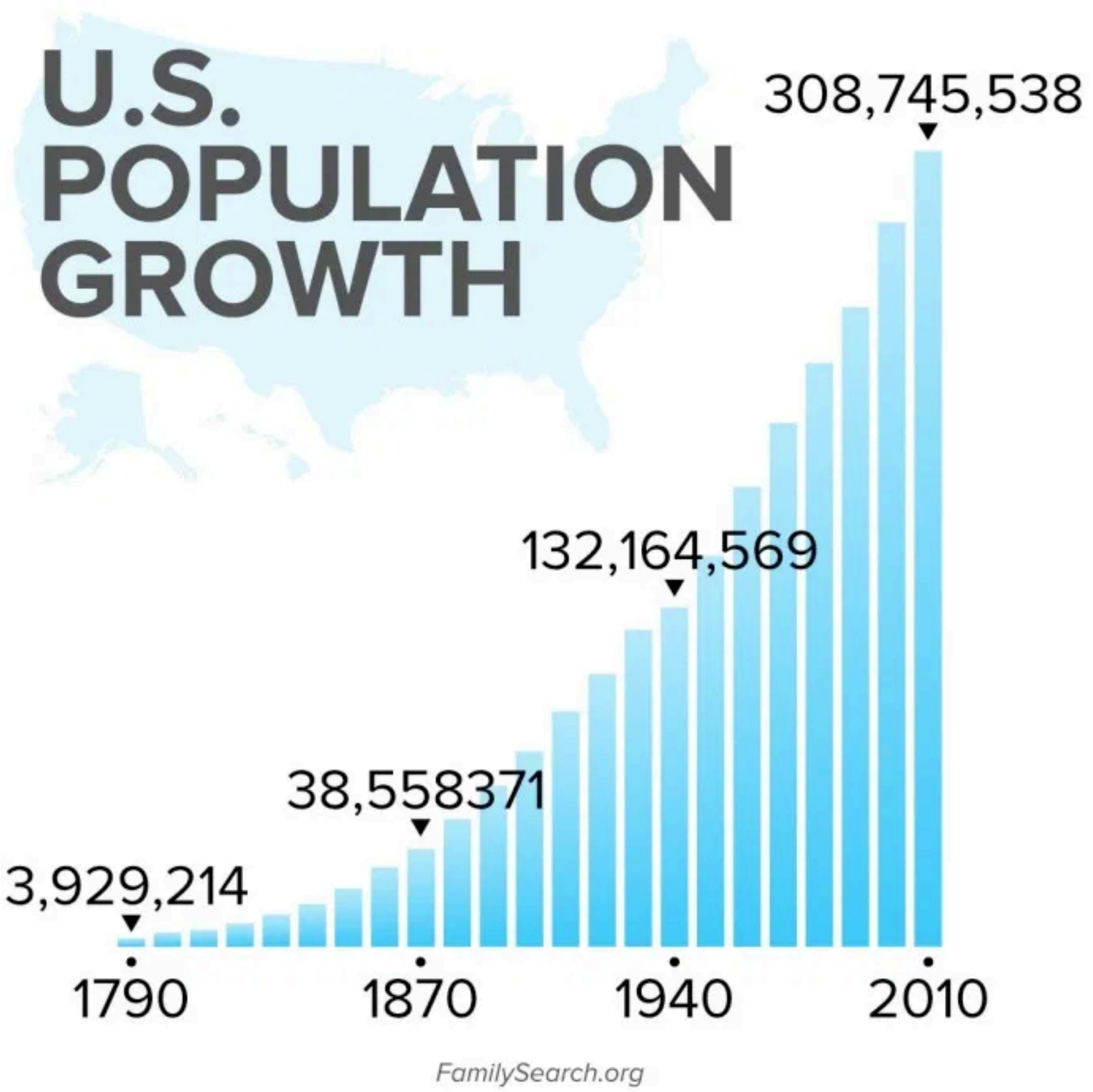
**Fairness impacts of DP in learning**

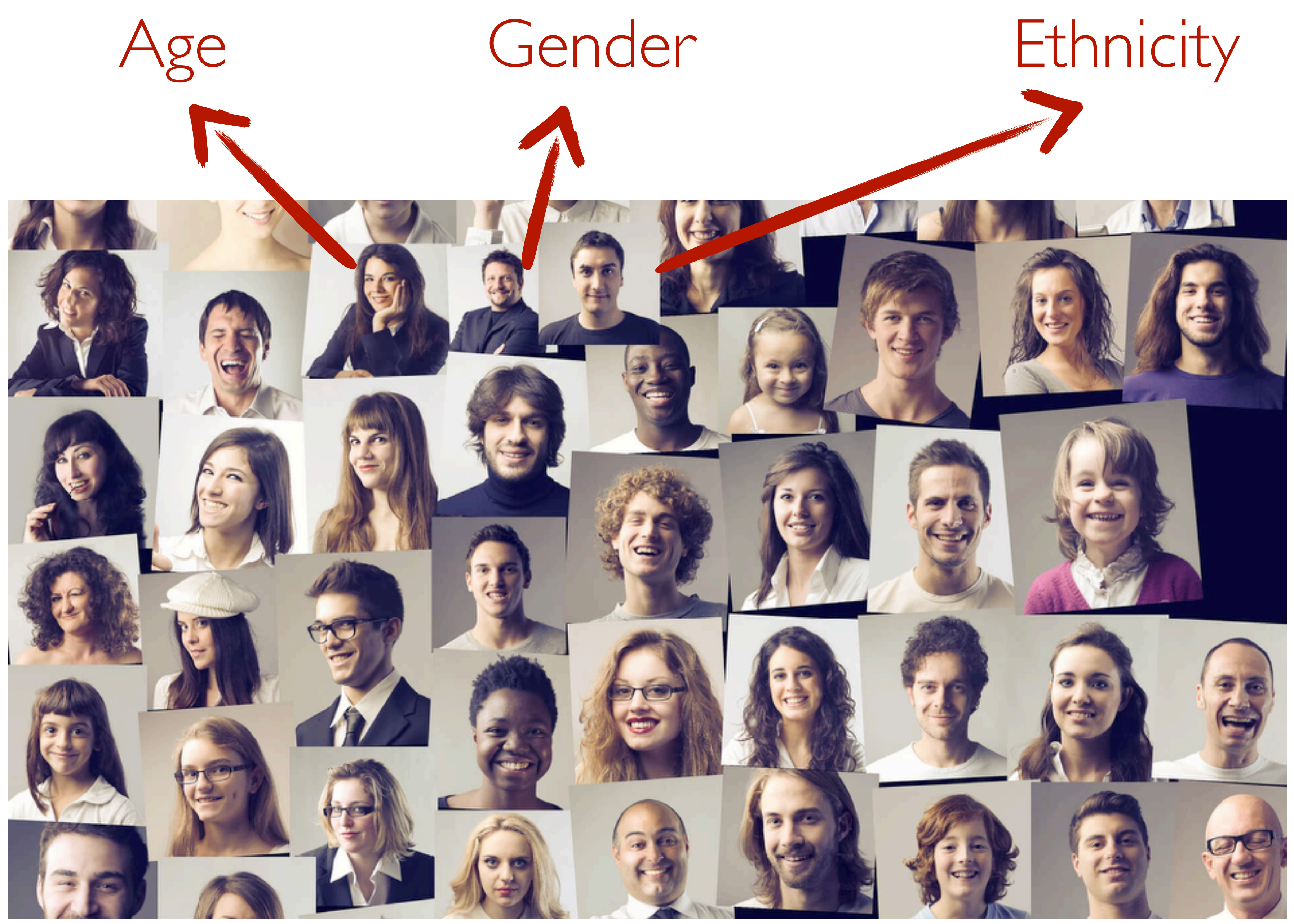


**What's next?**

# US Census data collection

## Enumeration of the total population living the US



# US Census data collection

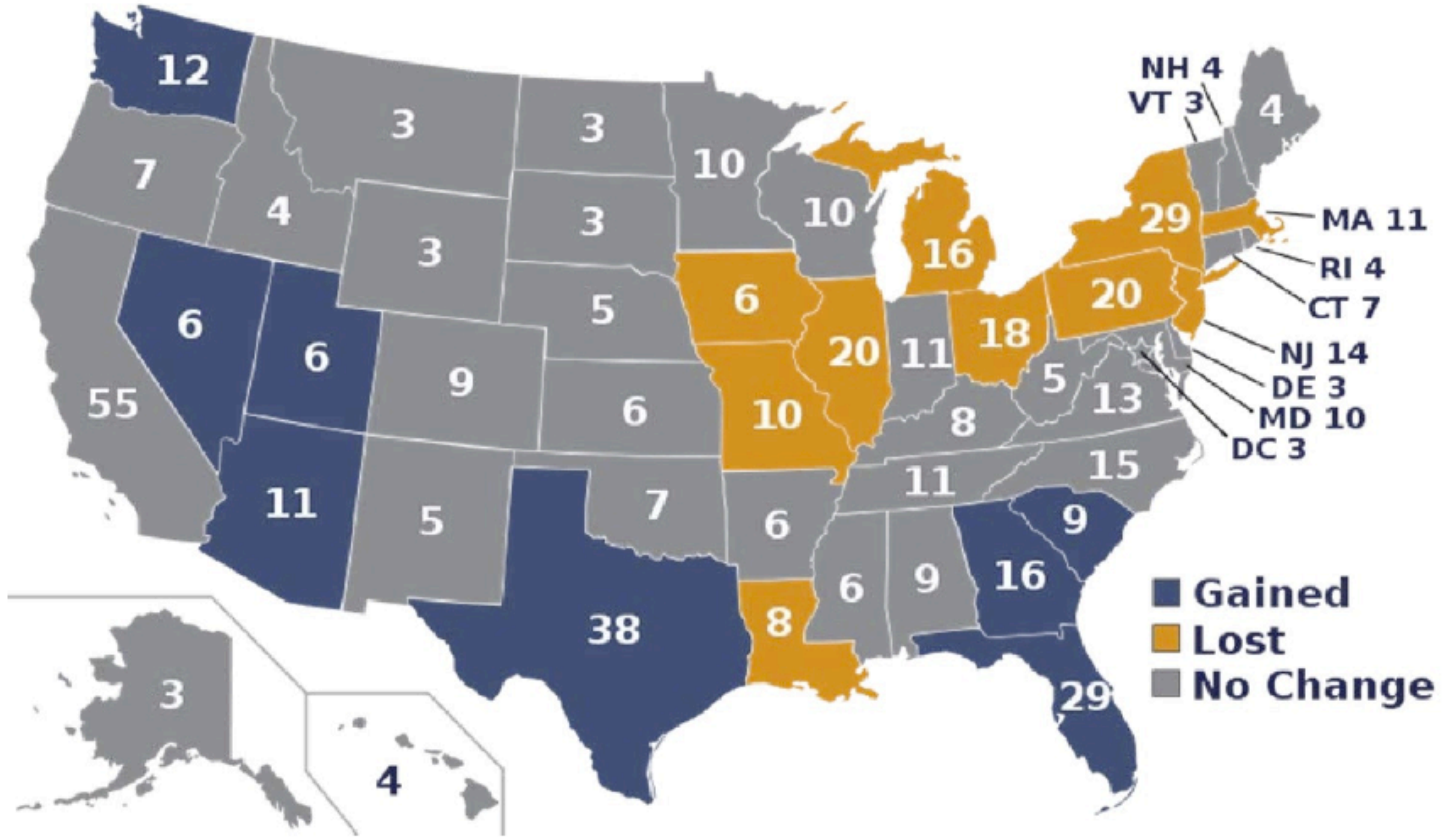
## Accurate count is important

- 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

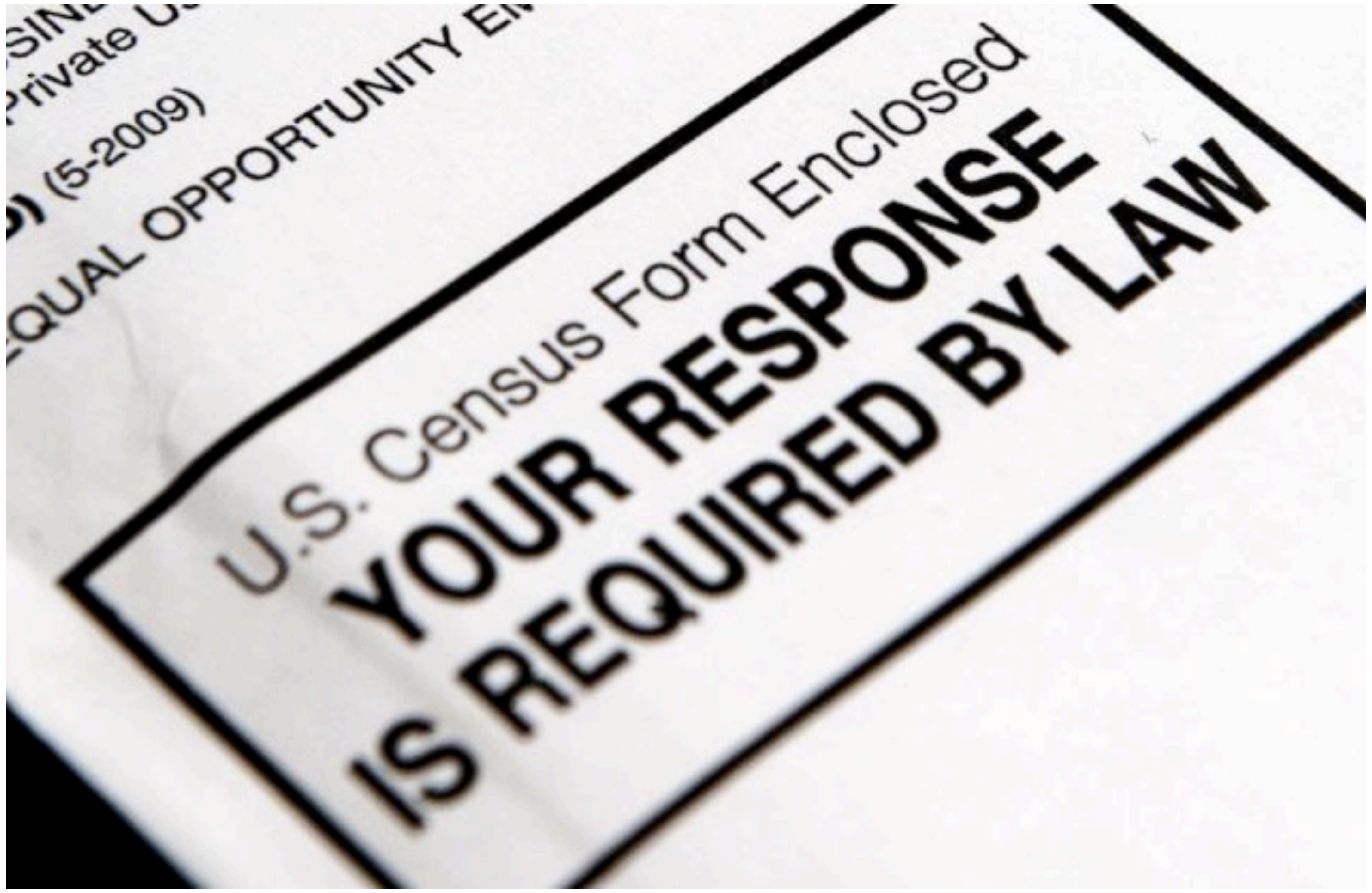




Determine the number of seats that states get in the US House of Representatives.

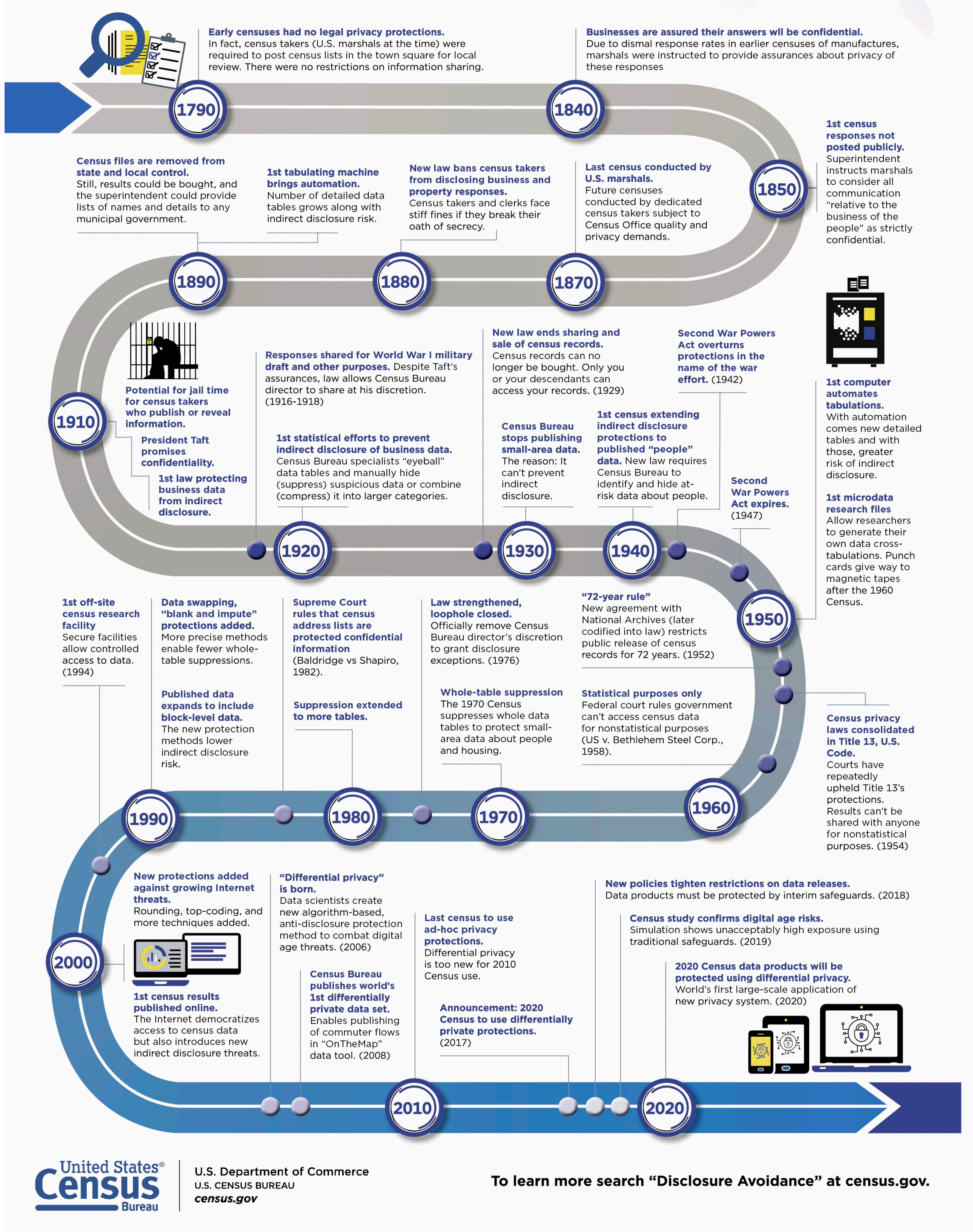
# US Census data collection

## 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.

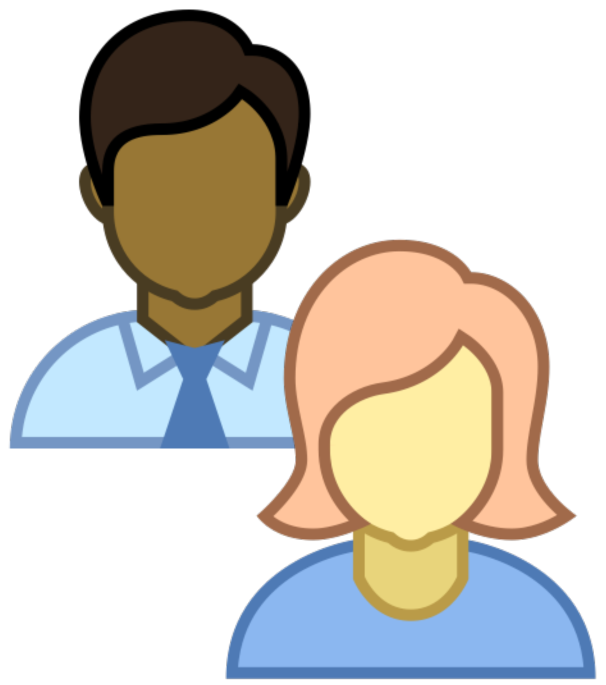


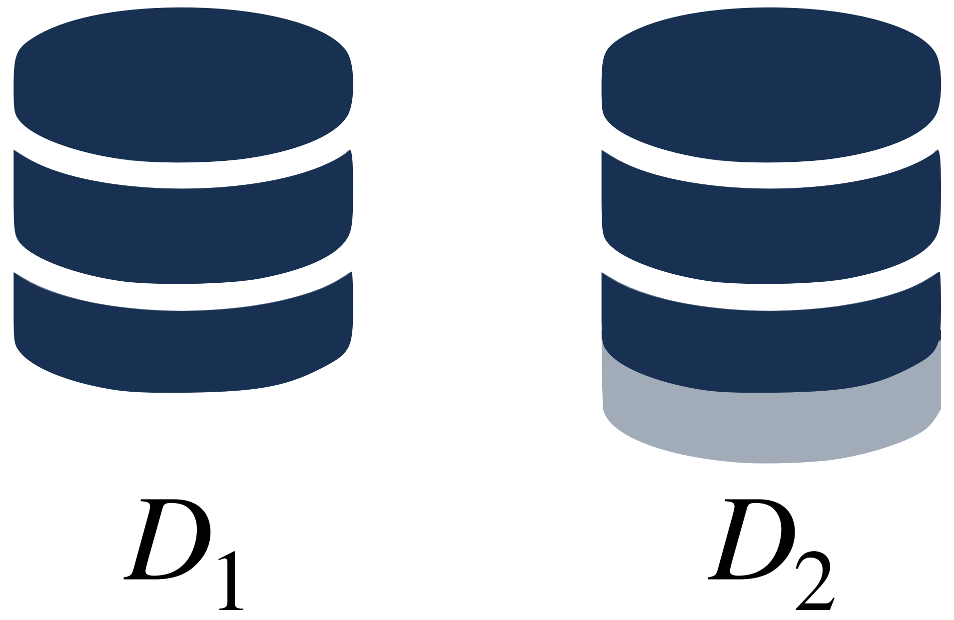
Ramachandran et al. 2012 McKenna et al. 2018

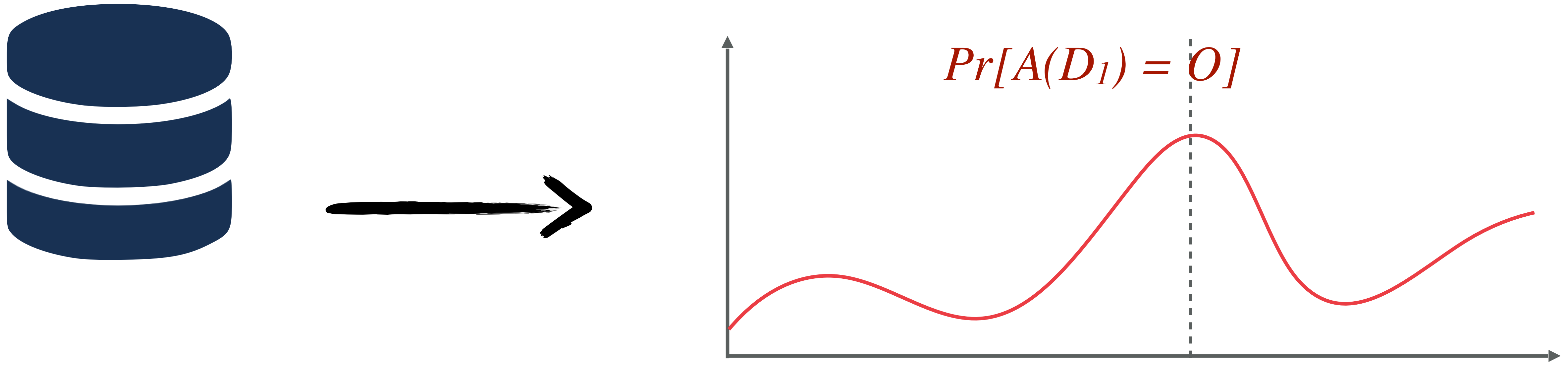
# Differential Privacy
## Definition

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

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





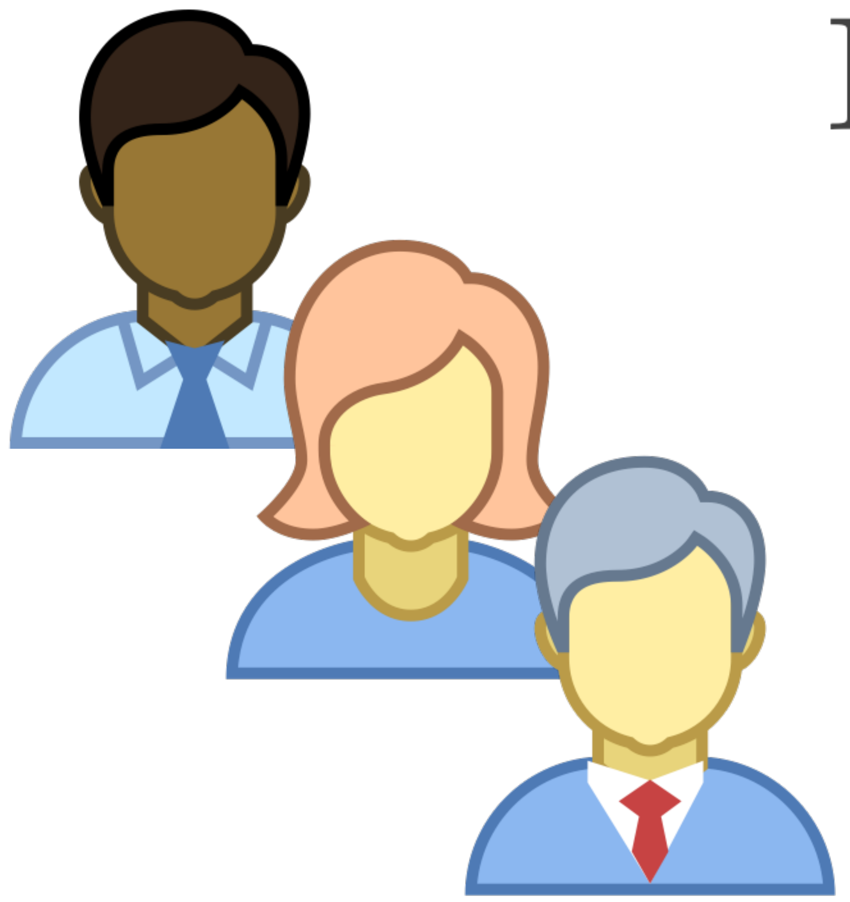
# Differential Privacy

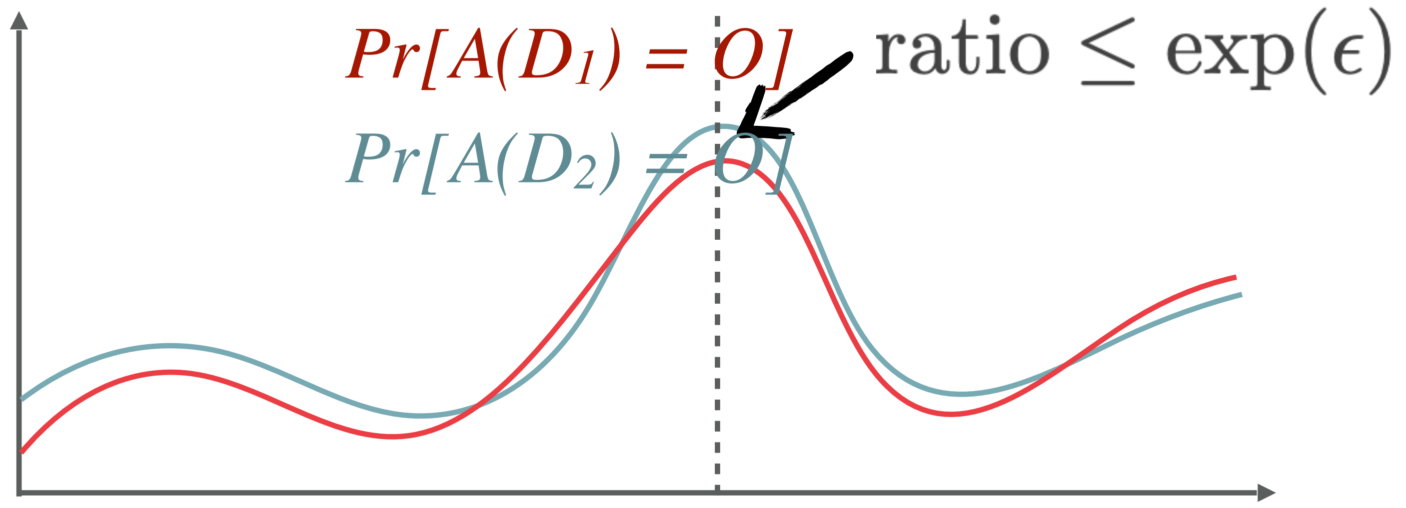
## Definition

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

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

$D_1$ $D_2$



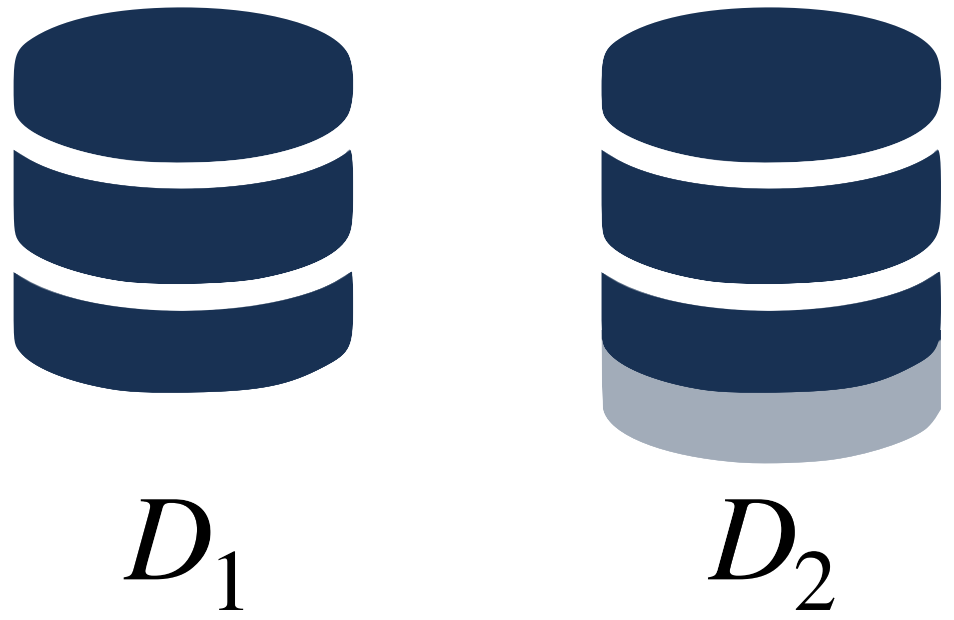


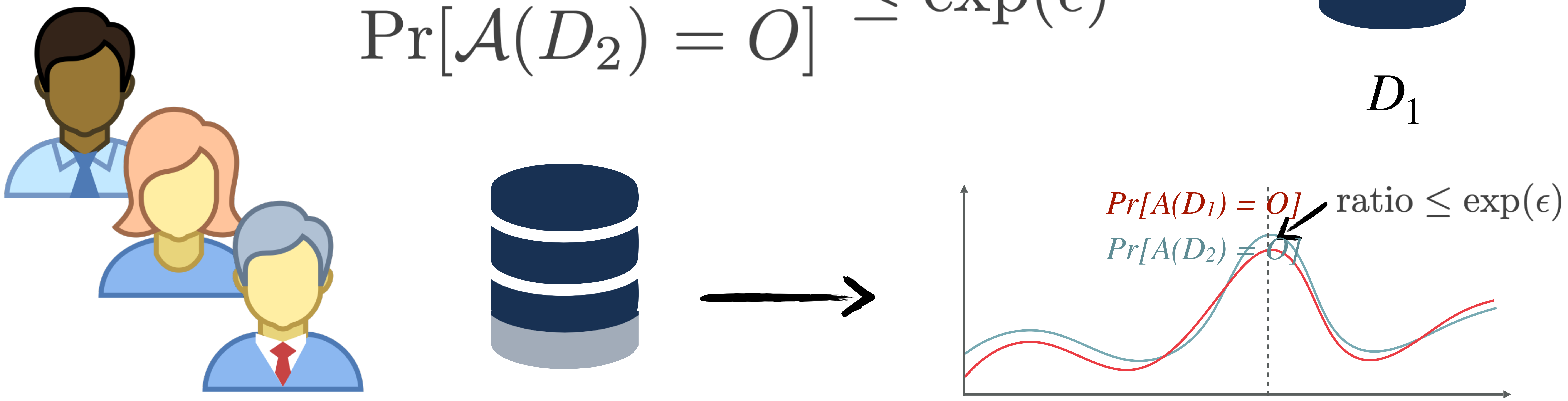
# Differential Privacy

## Definition

A randomized algorithm $\mathscr{A}$ 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

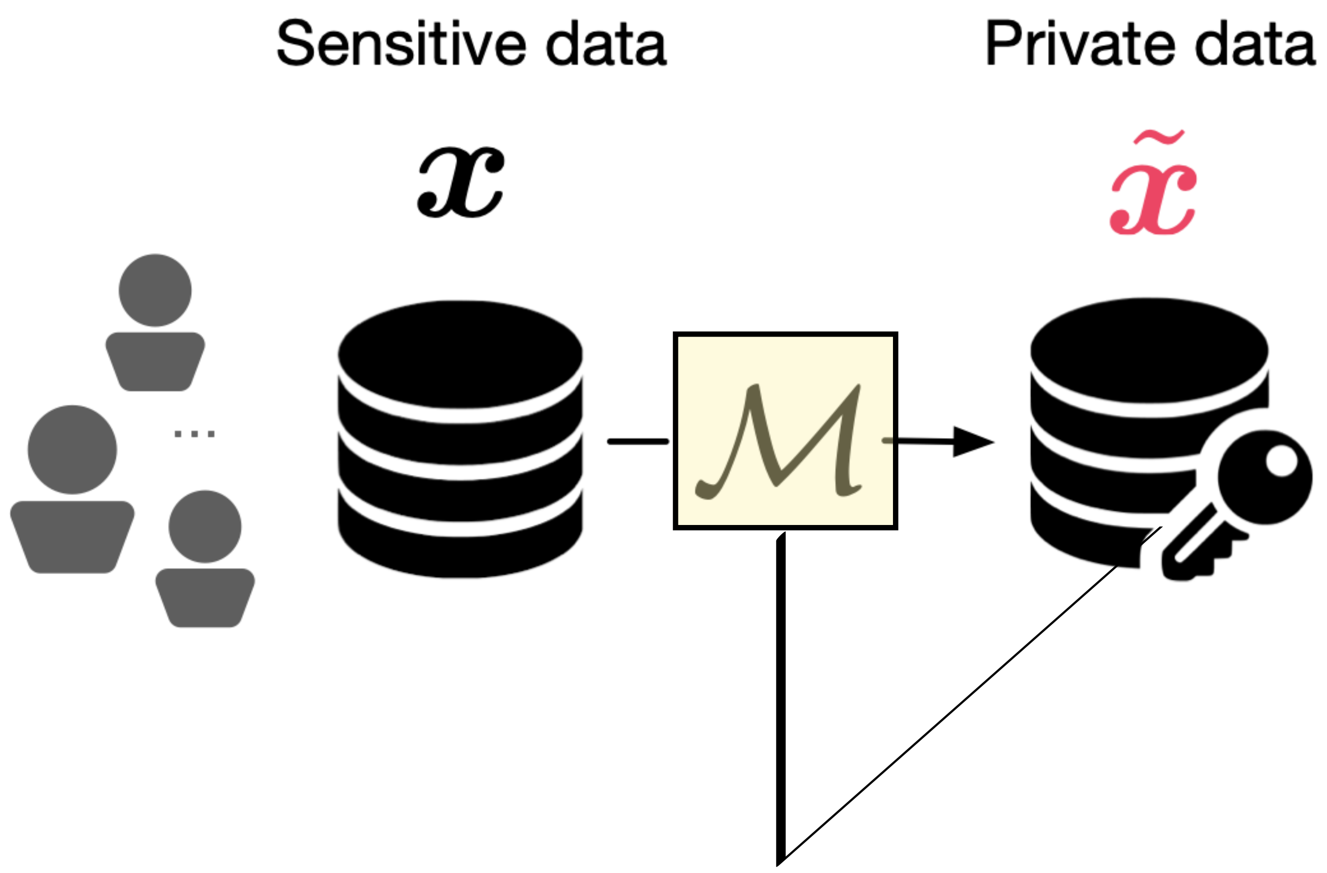
- Immune to linkage attack: Adversary knows arbitrary auxiliary information.
- Composability: If $A_1$ enjoys $\varepsilon_1$-differential privacy and $A_2$ enjoys $\varepsilon_2$-differential privacy, then, their composition $A_1(D), A_2(D)$ enjoys $(\varepsilon_1 + \varepsilon_2)$-differential privacy.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

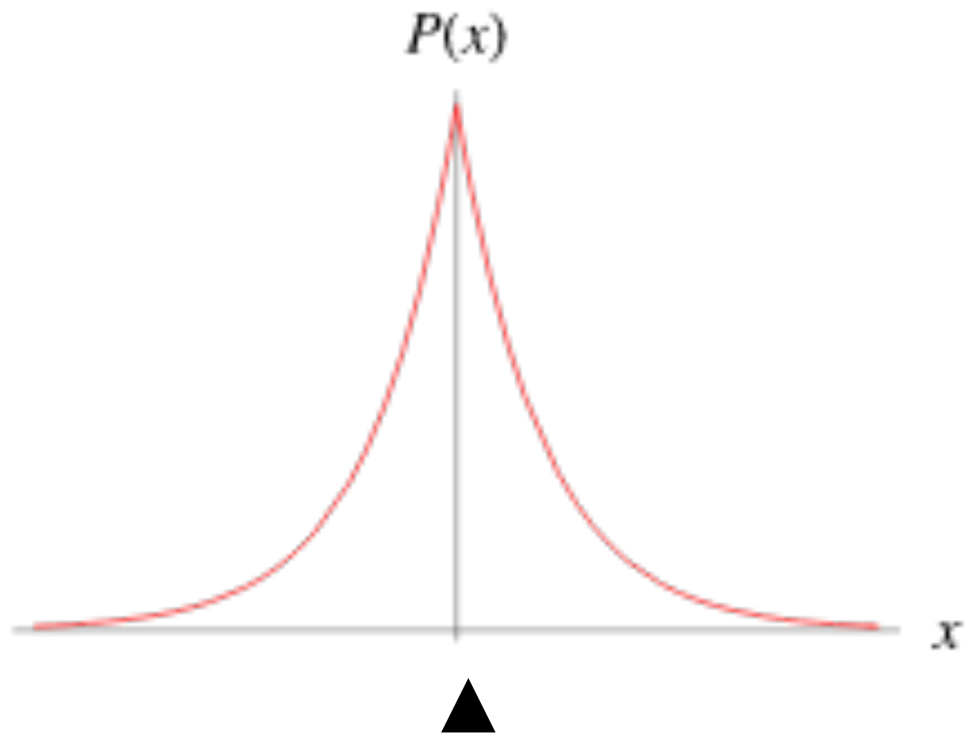
# Differential Privacy

## Notable properties

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

DP algorithms rely on randomization

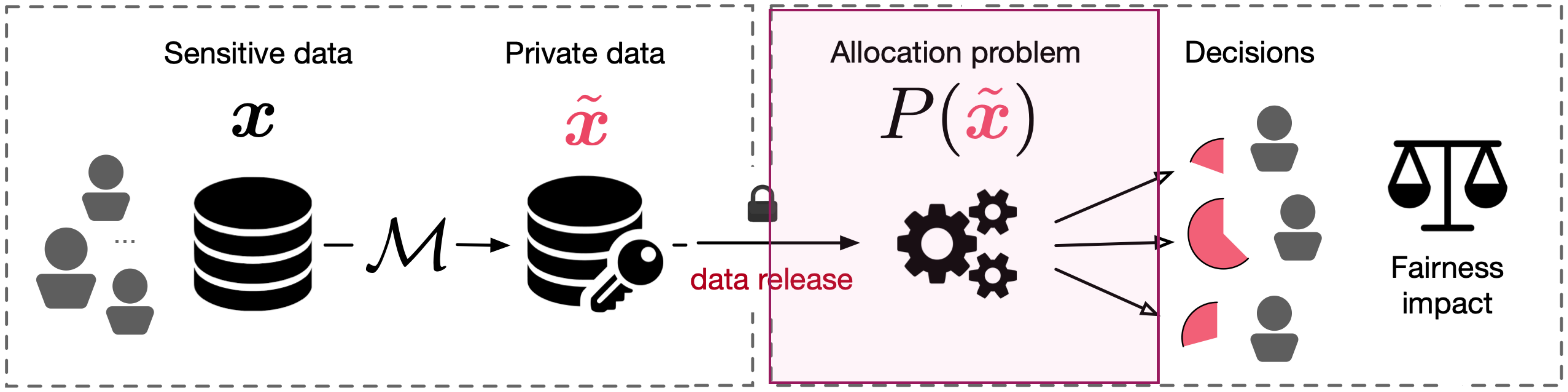



$x+\eta$

# Fairness in downstream decisions

## Setting




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

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

$$\xi_B^i(P, \mathcal{M}, \boldsymbol{x}) = \max_{j \in [n]} \left| B_P^i(\mathcal{M}, \boldsymbol{x}) - B_P^j(\mathcal{M}, \boldsymbol{x}) \right| \leq \alpha$$

# Disproportionate impacts in decision making

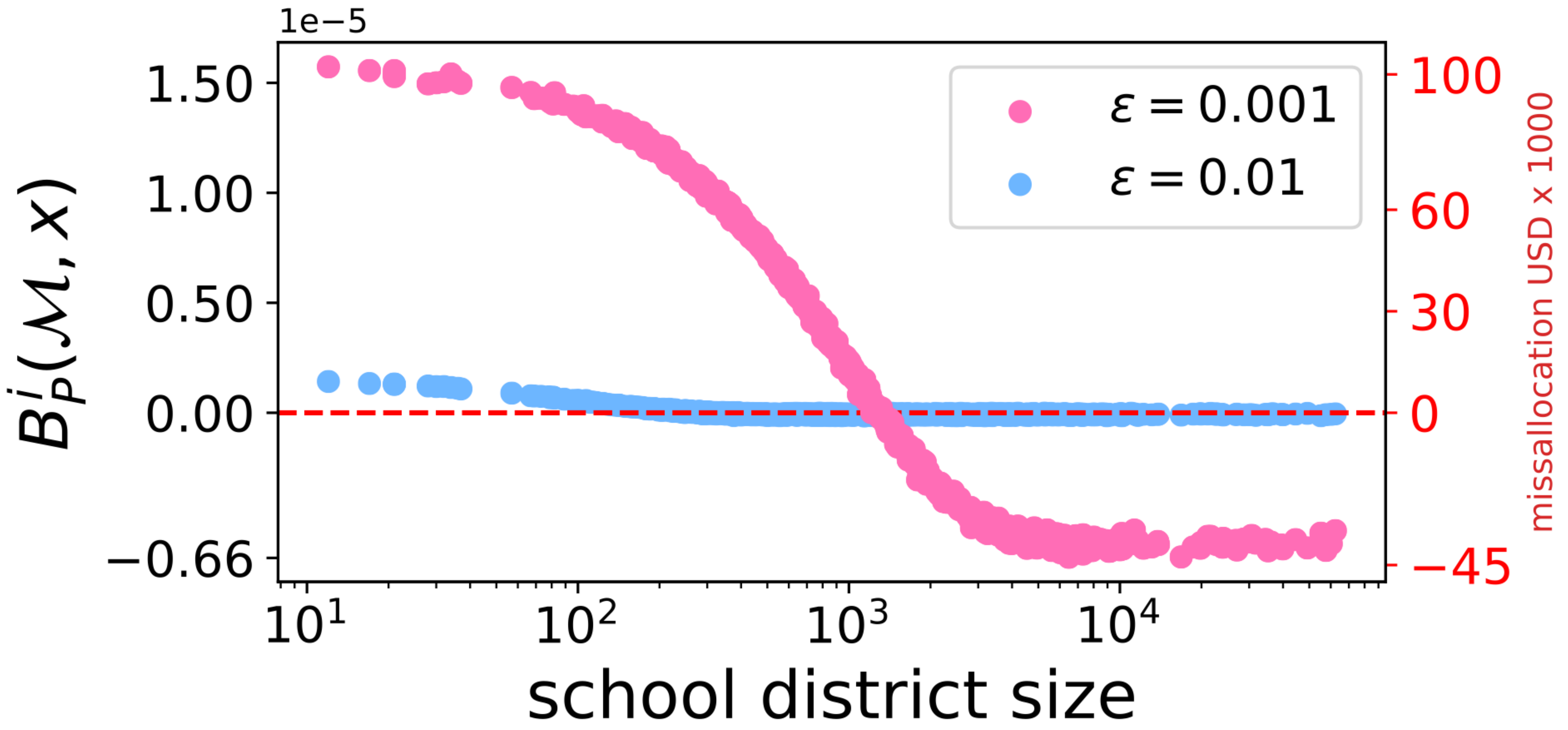
## Title 1 allotment

- 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






Fioretto et al. IJCAI:2021

# Disproportionate impacts in decision making

## Title 1 allotment

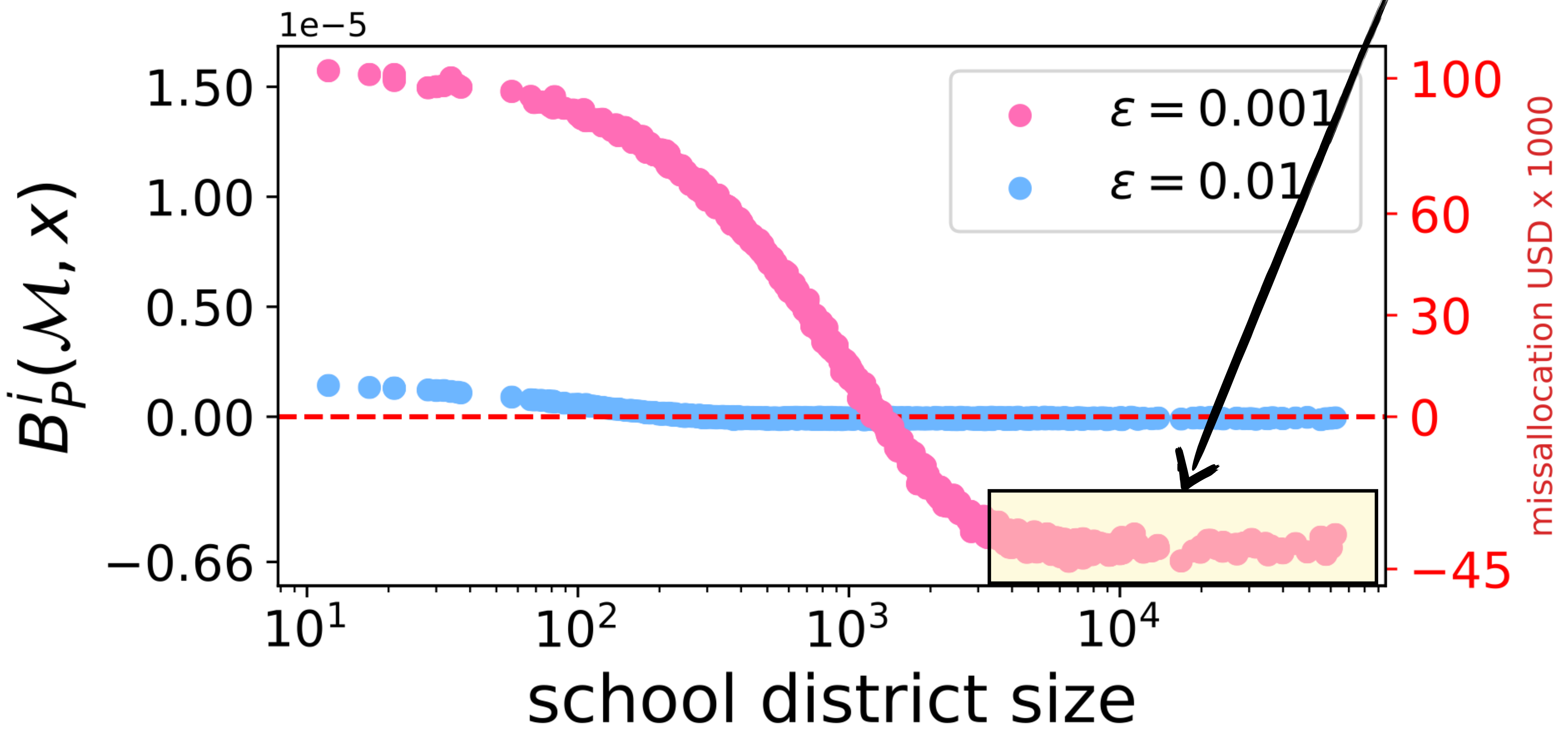
- Title 1 of the Elementary and Secondary Education Act is one of the largest U.S. program offering educational assistance to disadvantaged children.
- In the fiscal year 2021 alone, it distributed about \$11.7 billion through several types of grants.
- **Allotment:**

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

Districts receiving up to 42K less than warranted



# 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):

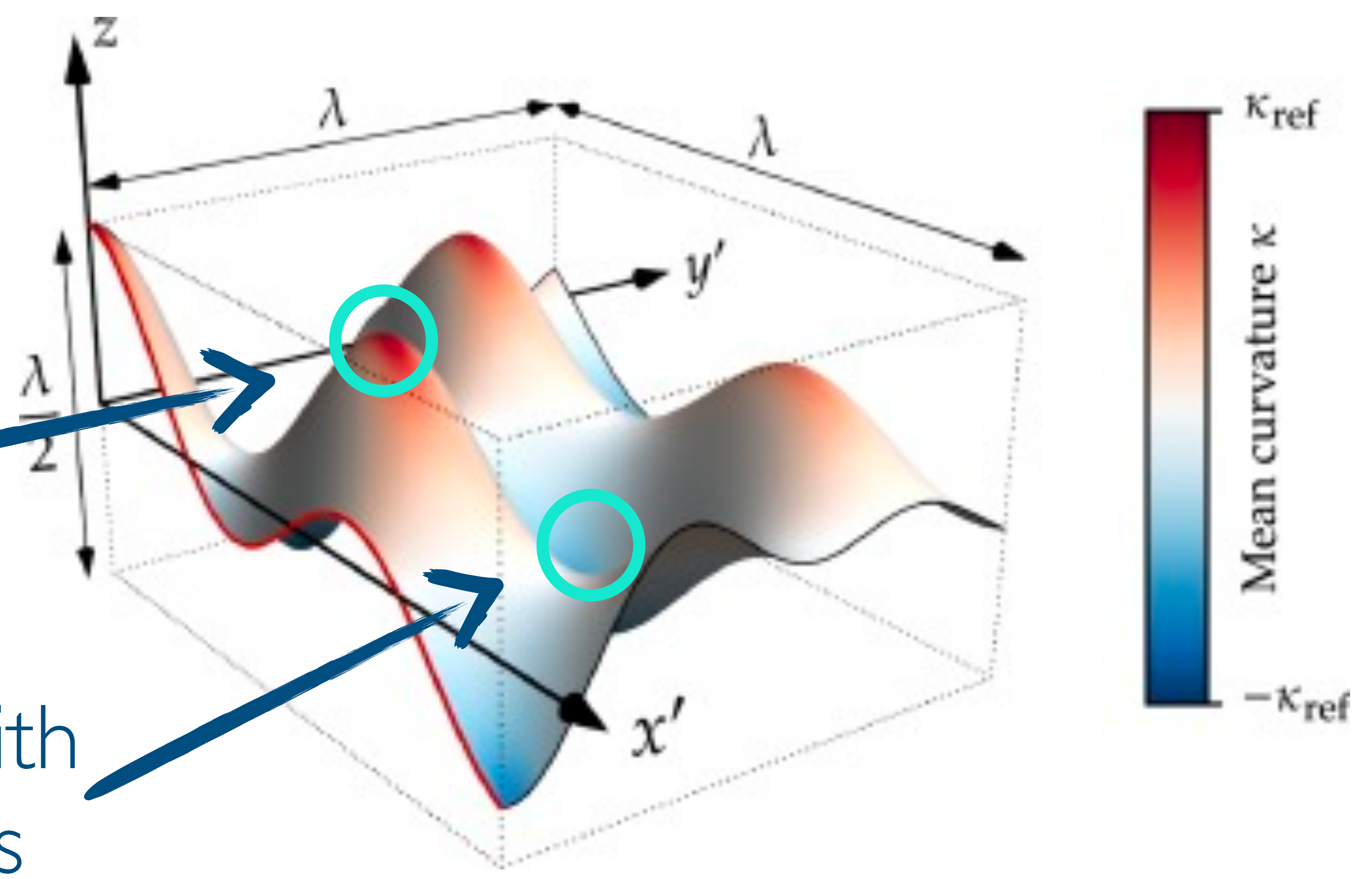
$$B_P^i(\mathcal{M}, \boldsymbol{x}) = \mathbb{E}[P_i(\tilde{\boldsymbol{x}} = \boldsymbol{x} + \eta)] - P_i(\boldsymbol{x})$$
$$\approx \frac{1}{2} \boldsymbol{H} P_i(\boldsymbol{x}) \times \mathrm{Var}[\eta]$$

Local curvature of problem Pi

Variance of the noisy input (depends on $\epsilon$)

entities with high errors

entities with low errors



# Shape of the decision problem

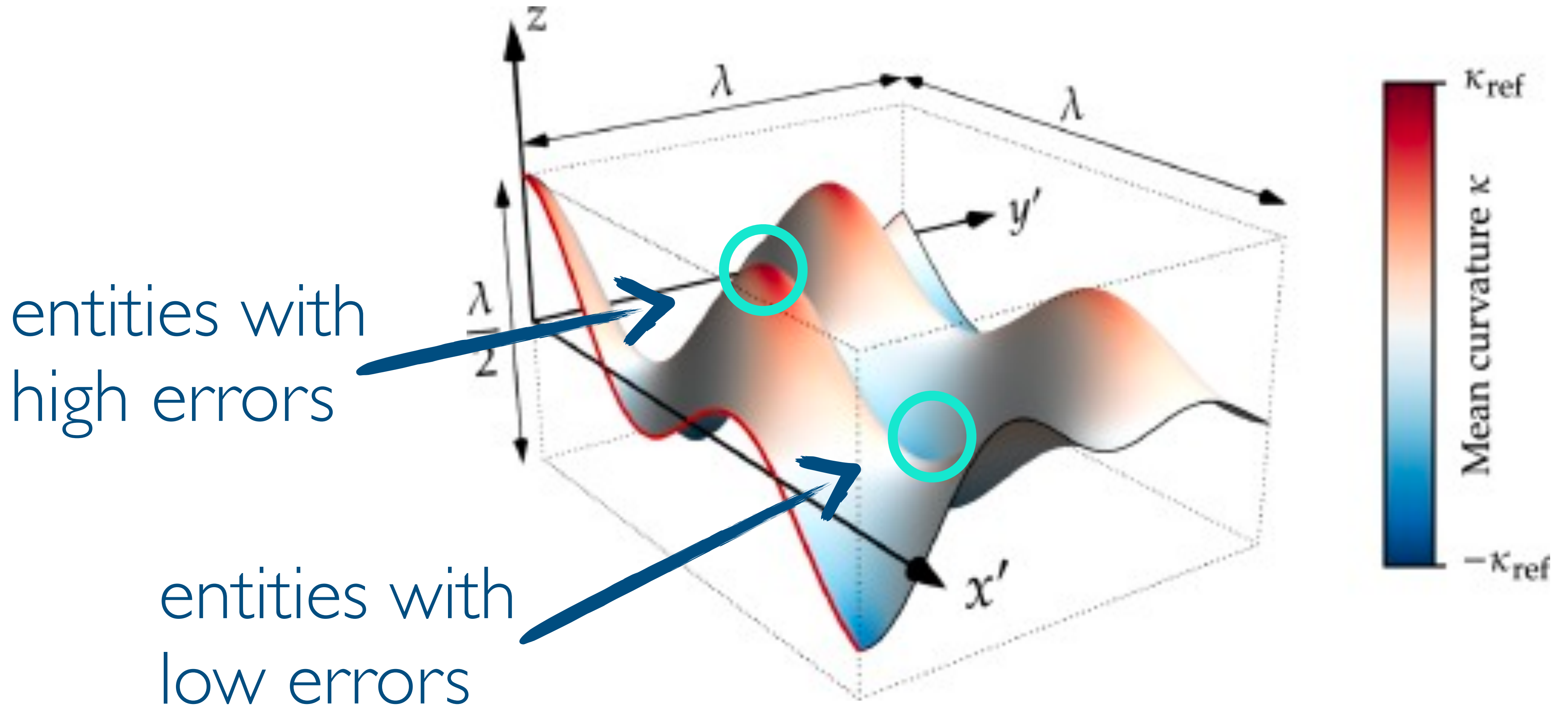
## First key result

- **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 $\epsilon$)




- 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 $\epsilon$)

A data release mechanism M is $\alpha$-fair w.r.t. P, for some finite $\alpha$, 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.

# Disproportionate impacts in downstream decisions

## Minority language voting rights

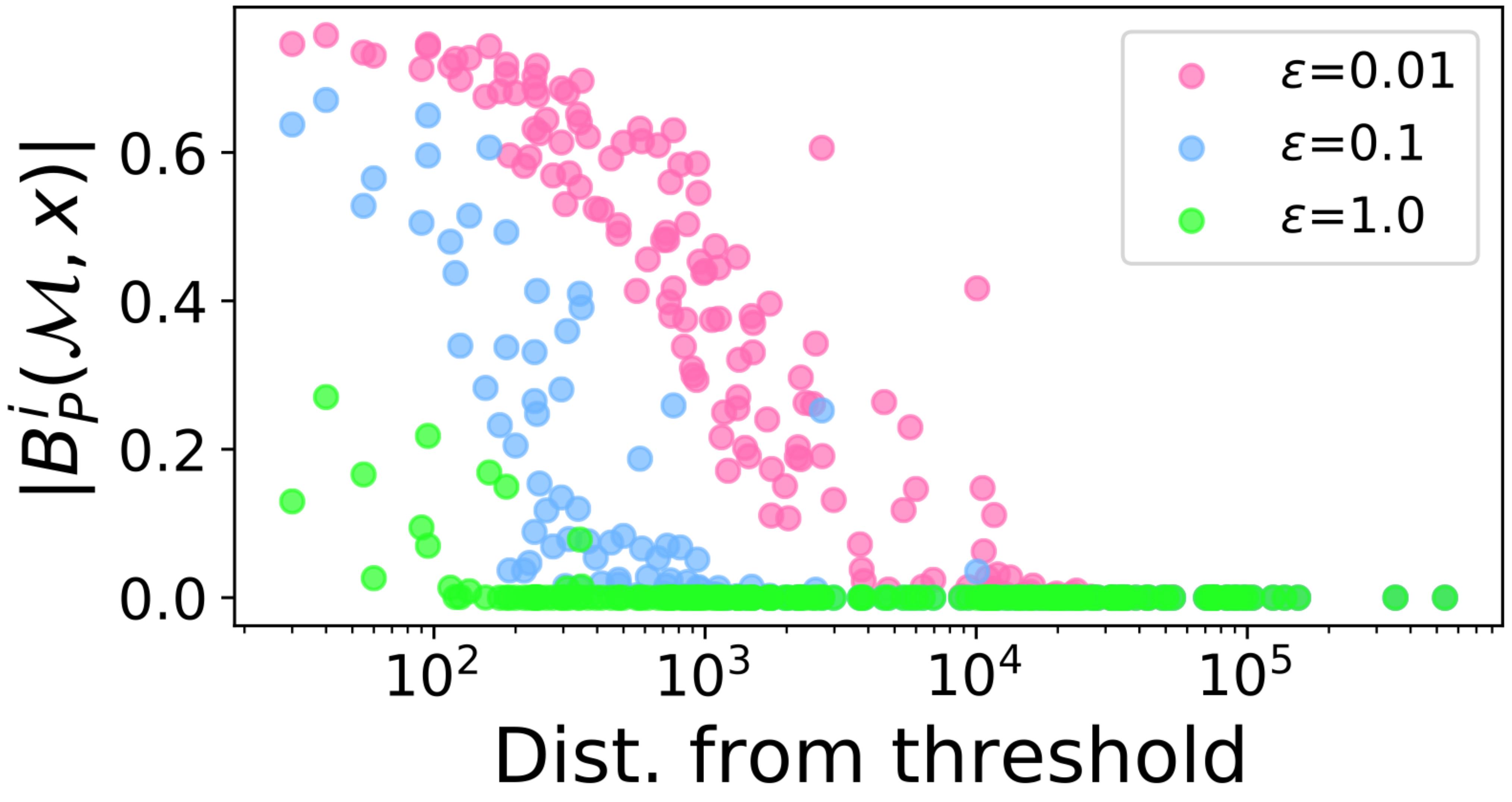
- The *Voting Rights Act* of 1965 provides a body of protections for racial and language minorities.
- Section 203 describes the conditions under which local jurisdictions must provide minority language voting assistance during an election.
- Jurisdiction i must provide language assistance (including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

+ < 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 $x^s$

# 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:

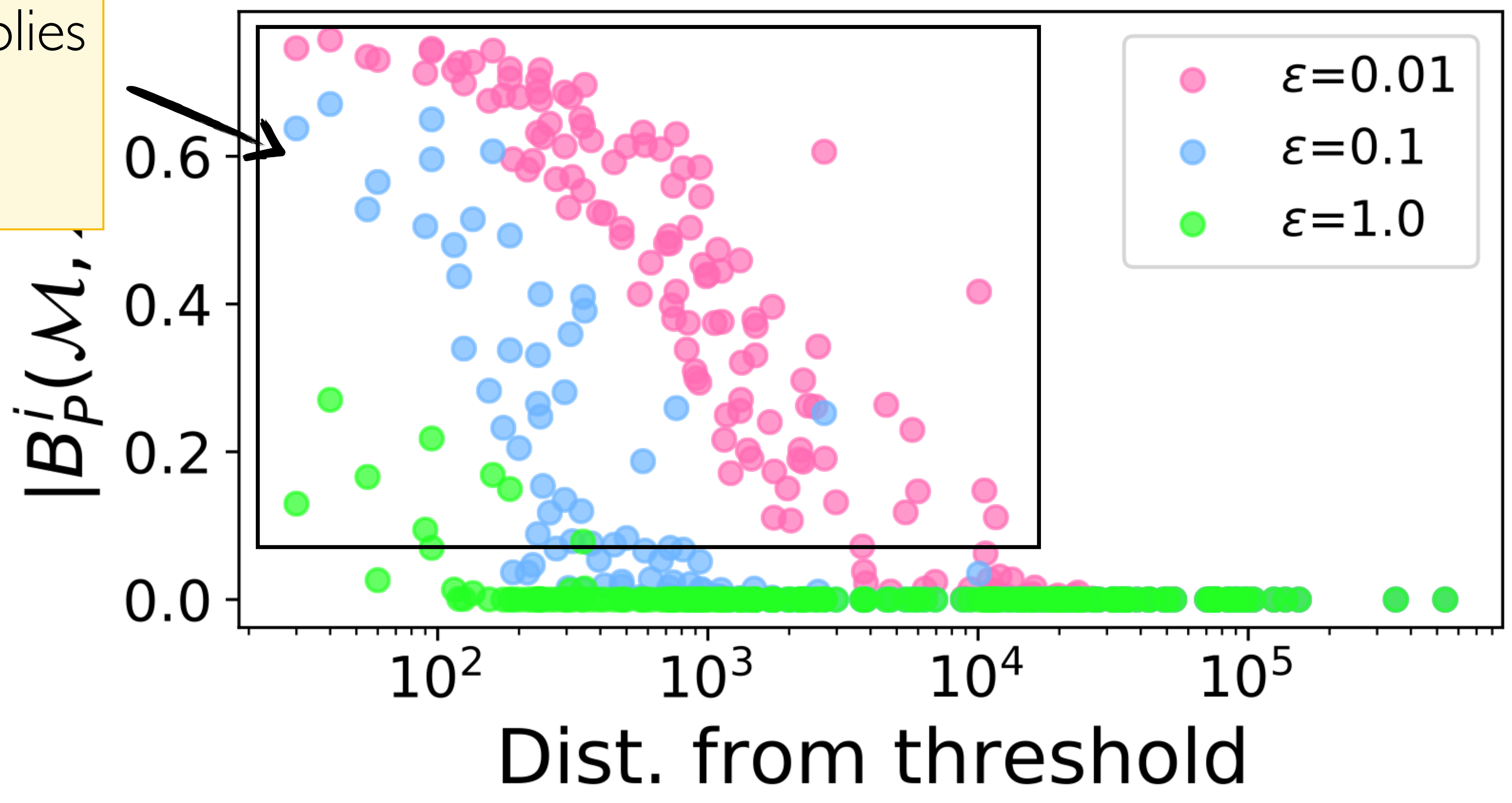
+ < 5th grade education

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

no. of ppl in i speaking minority language s

+ limited English proficiency

Misclassification implies potentially disenfranchising




sorted $x^s$

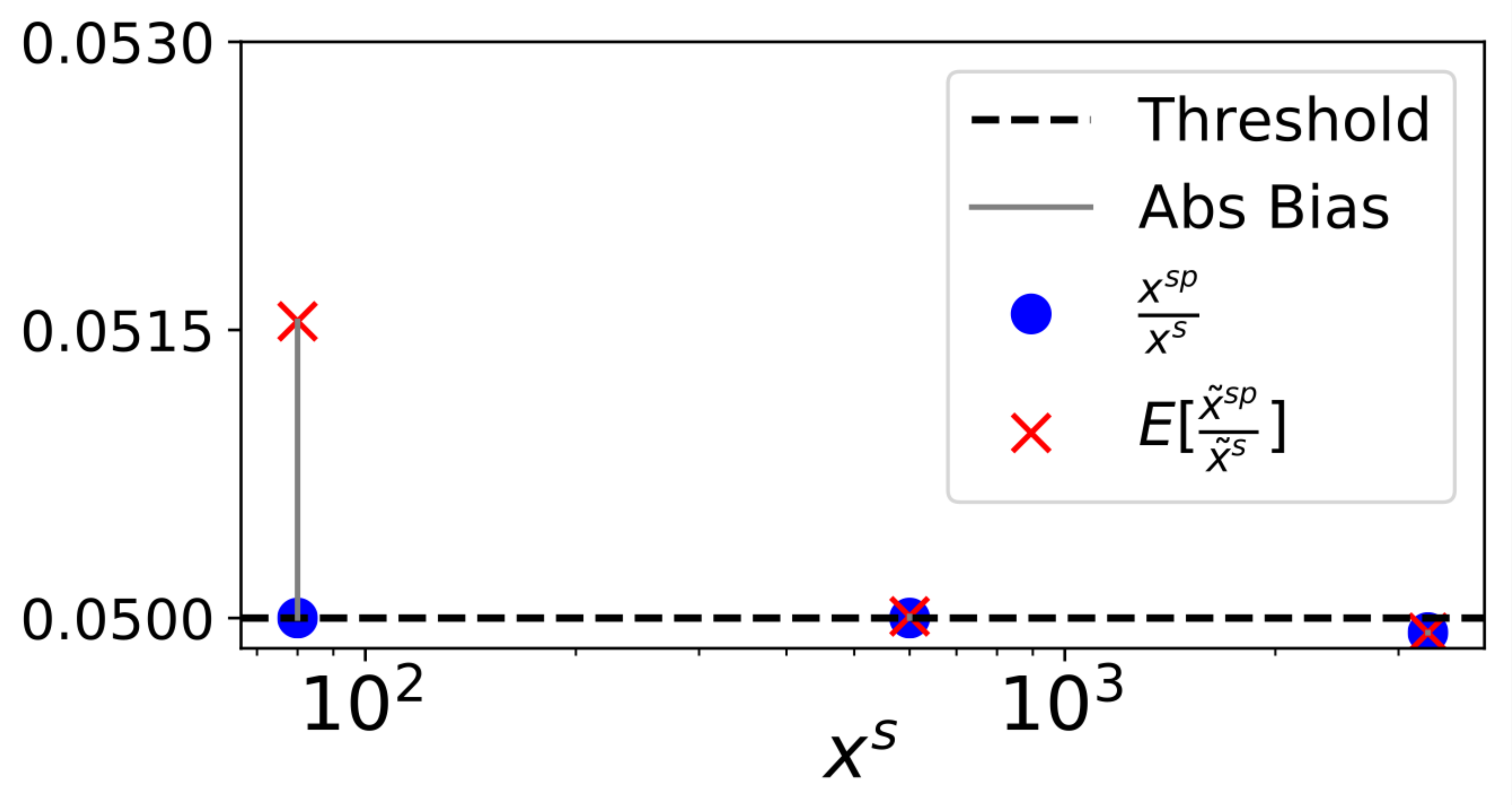
# Fair Decision Rules

## Ratio Functions

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

- *Loving county, TX*, where $x^{sp}/x^s = 0.05$
- *Terrell county, TX*, where $x^{sp}/x^s = 0.05$
- *Union county, NM*, where $x^{sp}/x^s = 0.049$

Minority Language Voting Rights



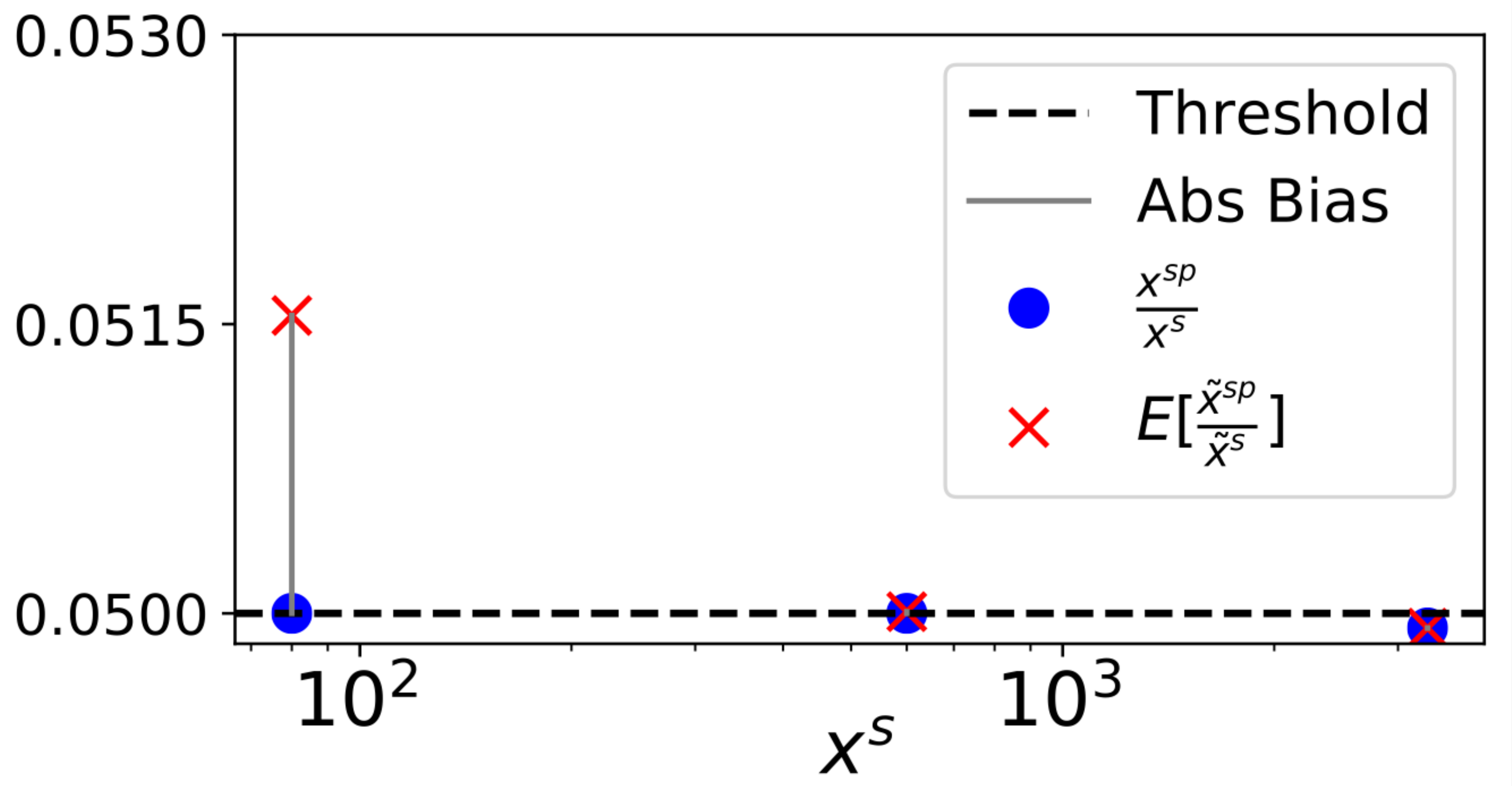
# Fair Decision Rules

## Ratio Functions

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

- *Loving county, TX*, where $x^{sp}/x^s = 0.05 = \frac{4}{80}$
- *Terrell county, TX*, where $x^{sp}/x^s = 0.05 = \frac{30}{600}$
- *Union county, NM*, where $x^{sp}/x^s = 0.049 = \frac{160}{3305}$

Minority Language Voting Rights




- **Theorem (informal):** The perturbation induced by the DP mechanism affects more the county with lower numerator / denominator.

# Fairness composition

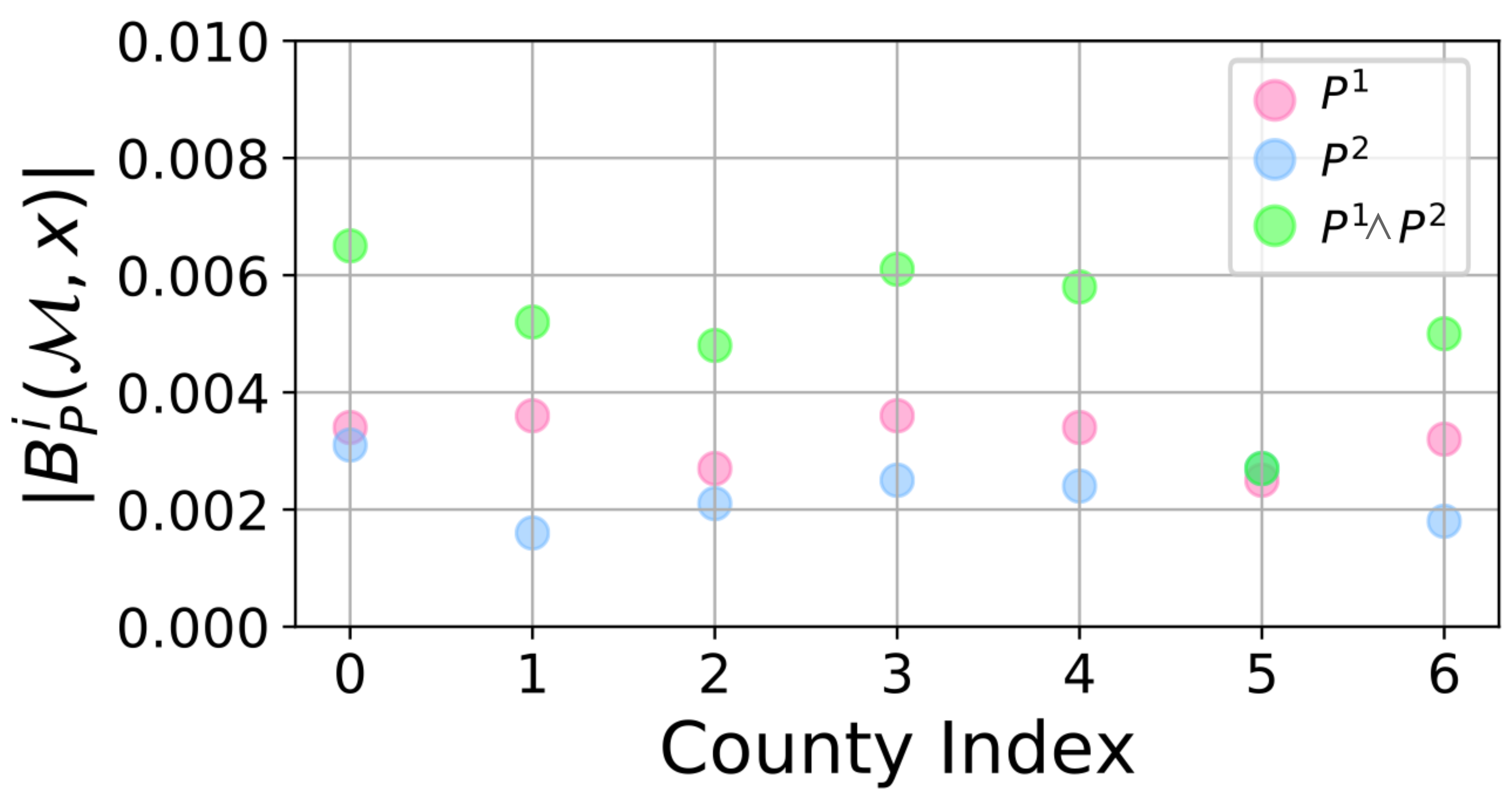
## Second key result

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

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

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

Minority Language Voting Rights




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

# Fairness composition

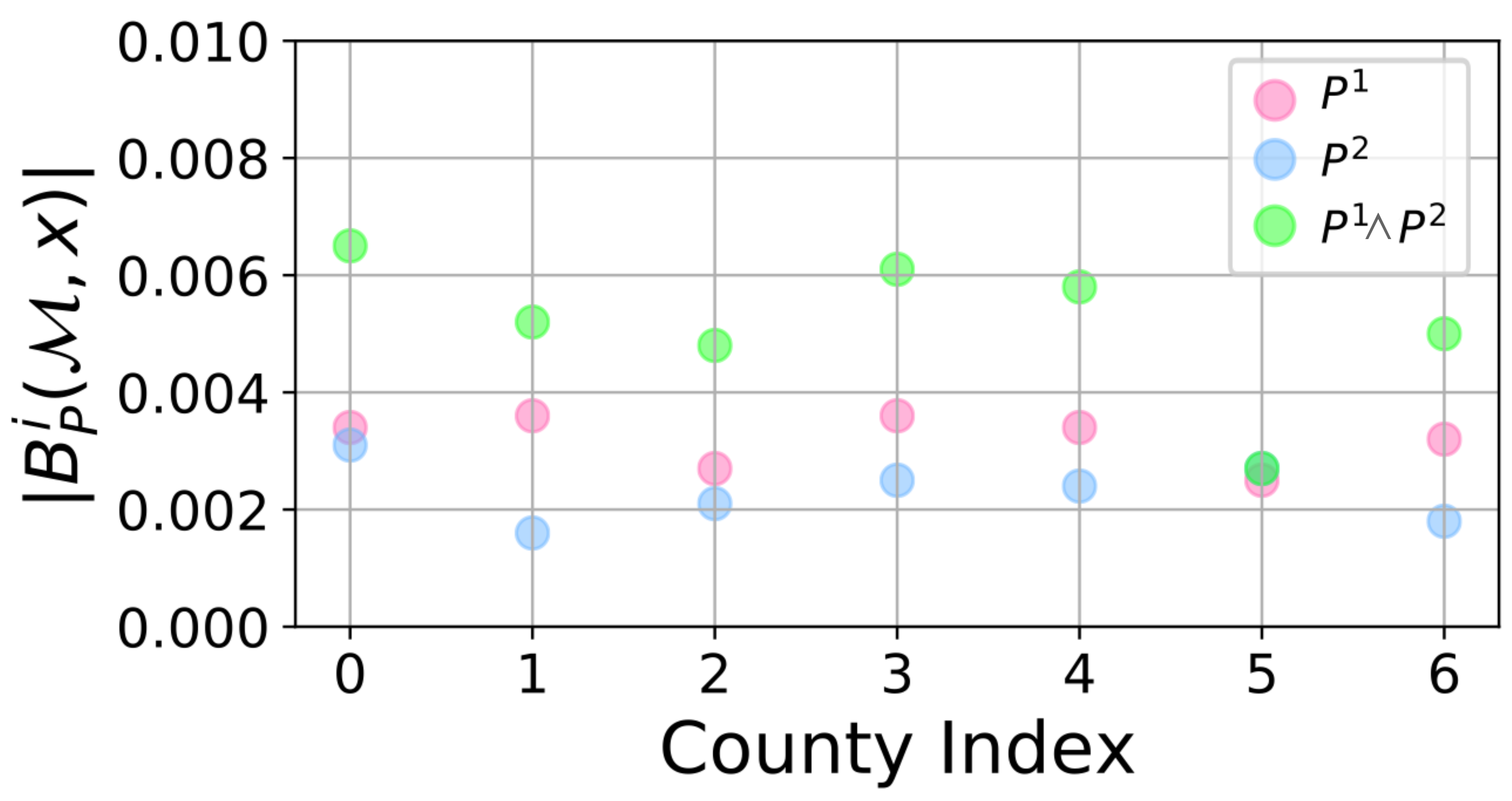
## Second key result

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

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

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

Minority Language Voting Rights




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

- **Theorem (informal)**: The logical composition of two $\alpha_1$- and $\alpha_2$-fair mechanisms is $\alpha$-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

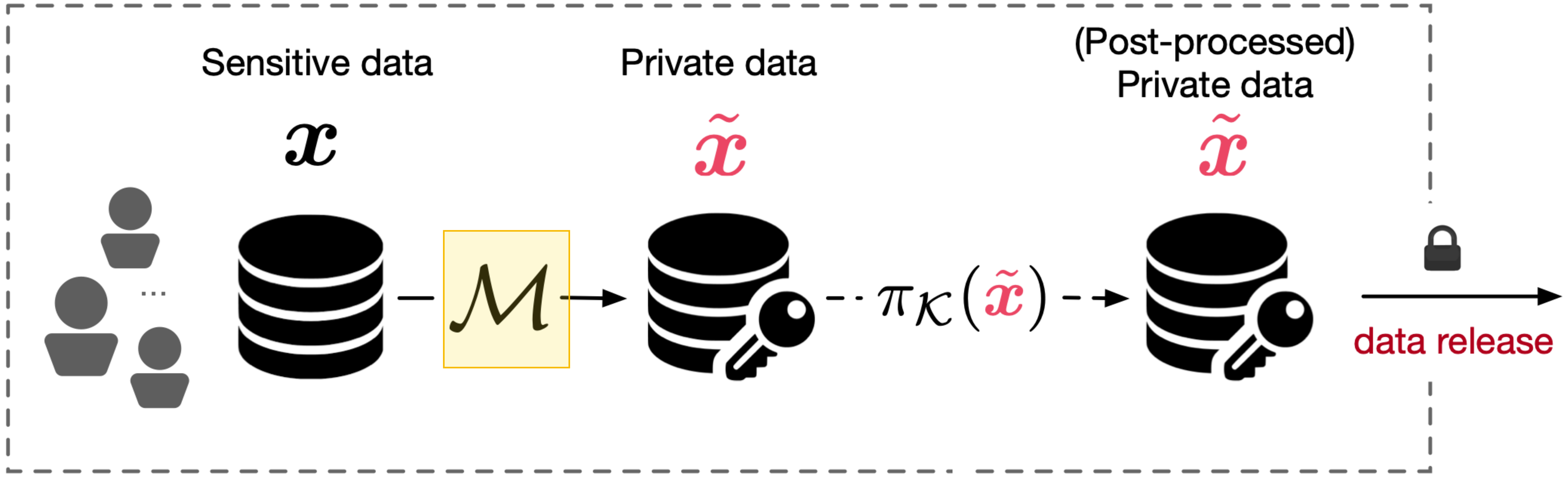
## Important conclusion

*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.*

# DP Post-processing

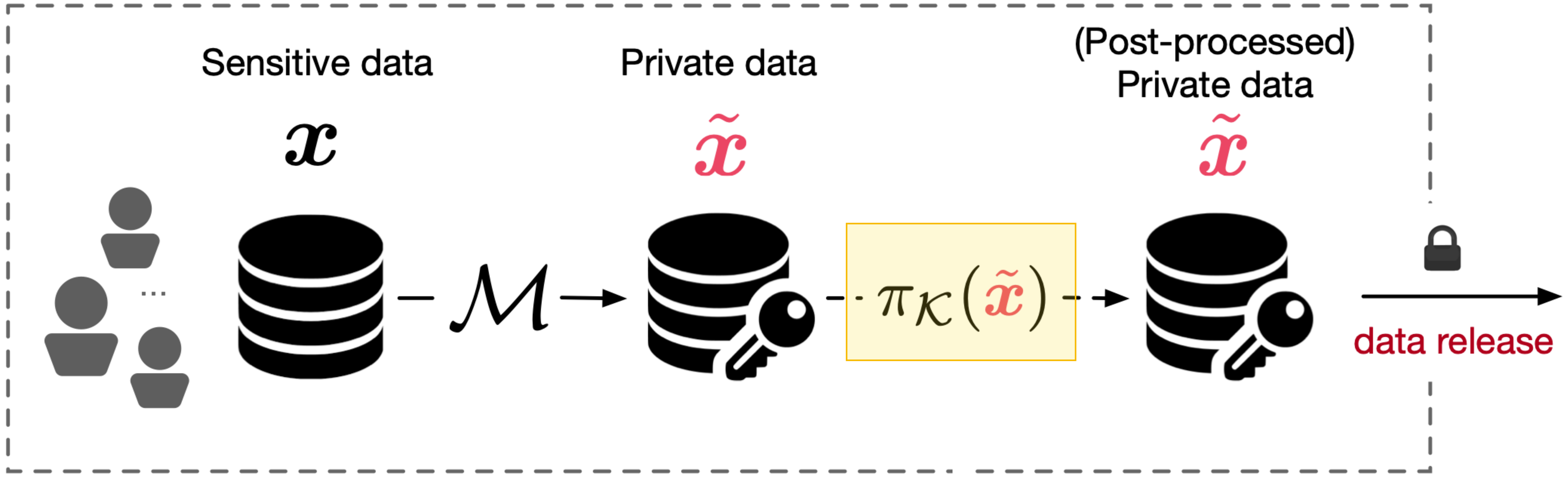
## Fairness impact

# 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

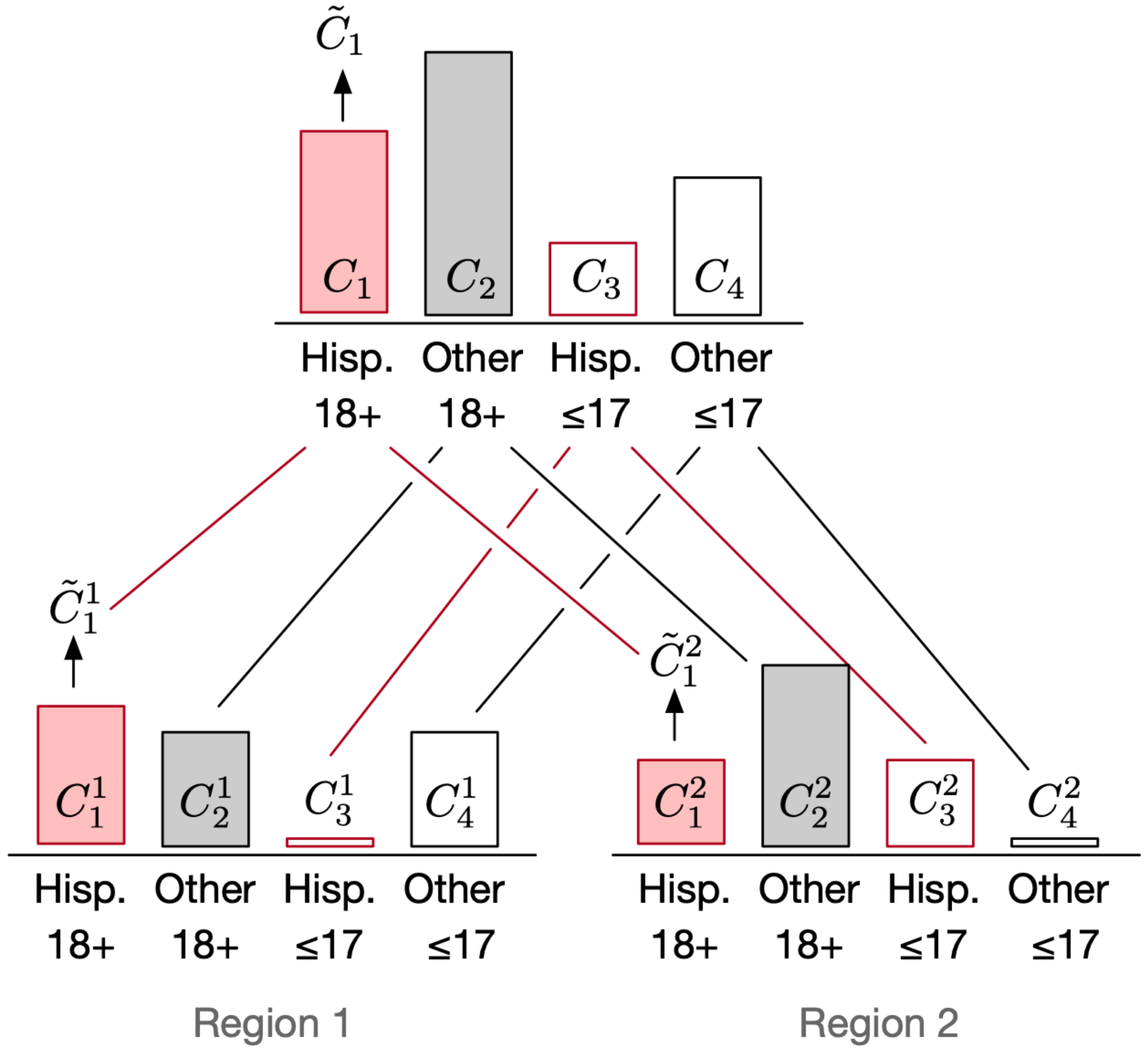



1. Apply noise with appropriate parameter $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \text{Noise}$
2. Post-process output $\tilde{x}$ to enforce consistency

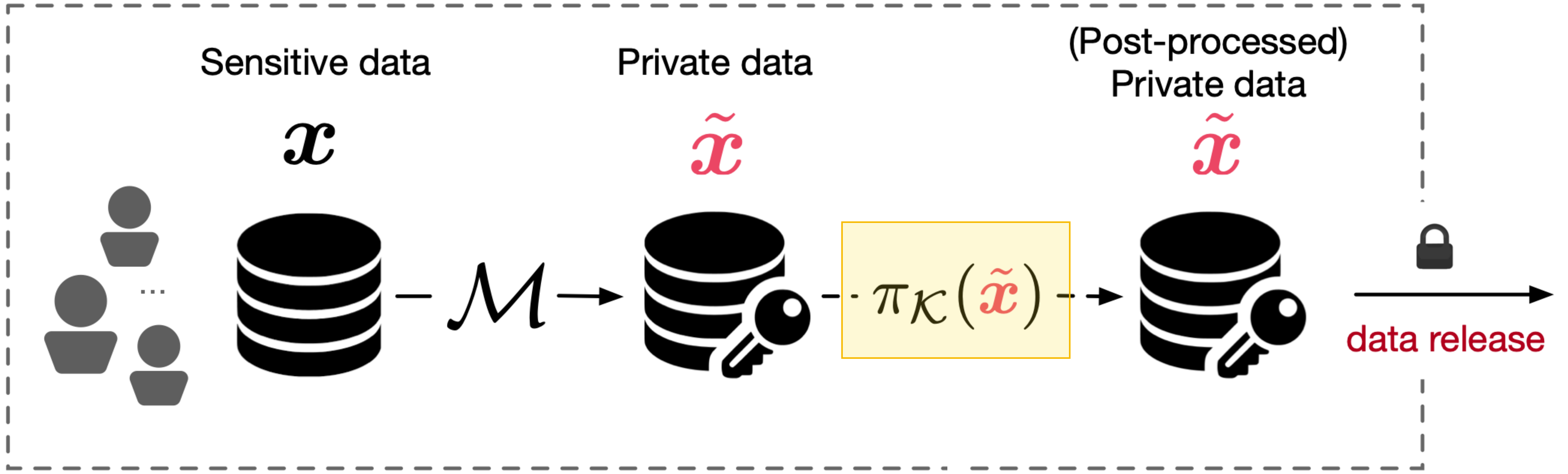
$$\pi_{\mathcal{K}}(\tilde{\boldsymbol{x}}) : \operatorname*{argmin}_{\boldsymbol{v} \in \mathcal{K}} \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2$$

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



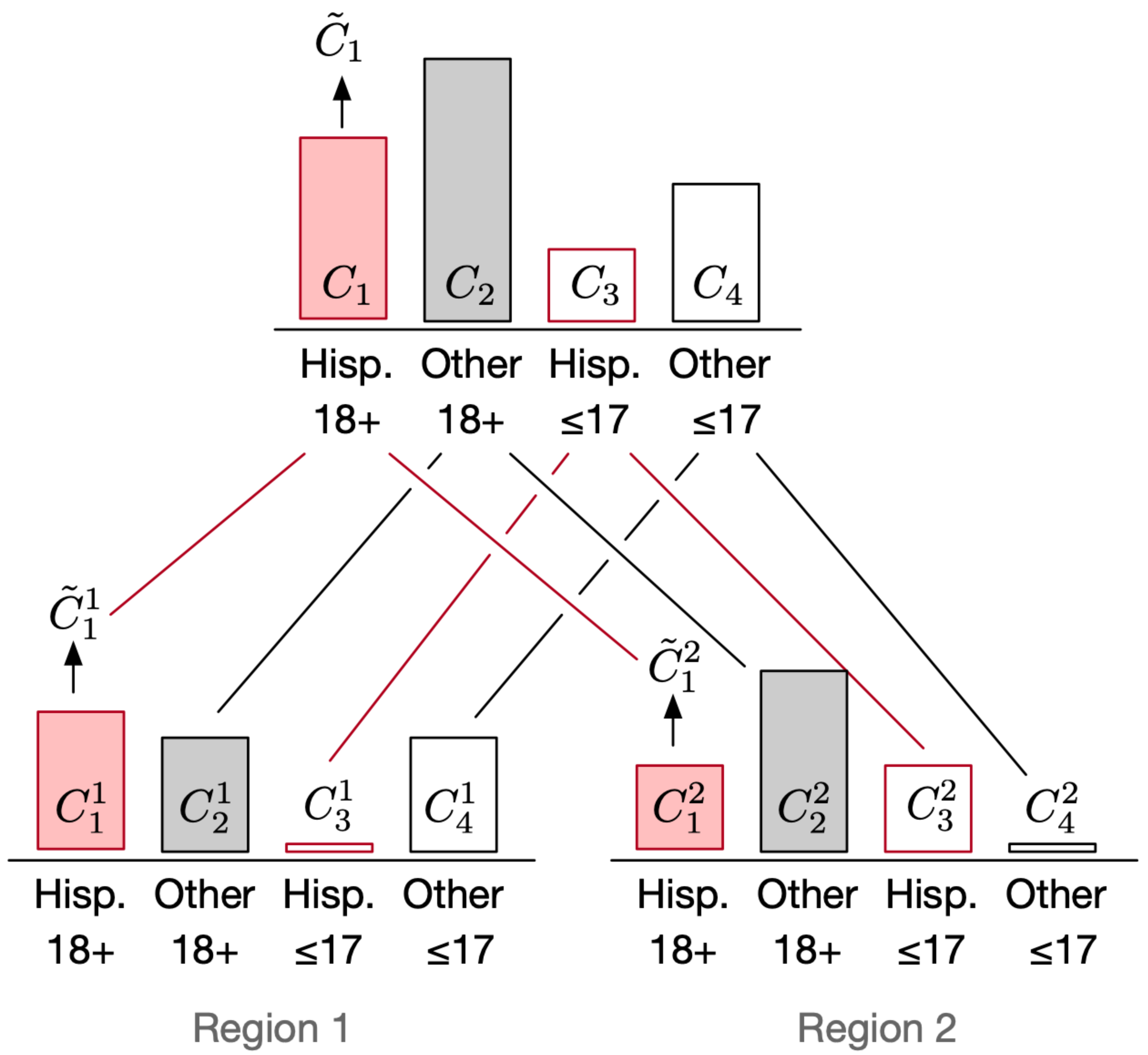

1. Apply noise with appropriate parameter $\boxed{\tilde{\boldsymbol{x}}} = \boldsymbol{x} + \text{Noise}$

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

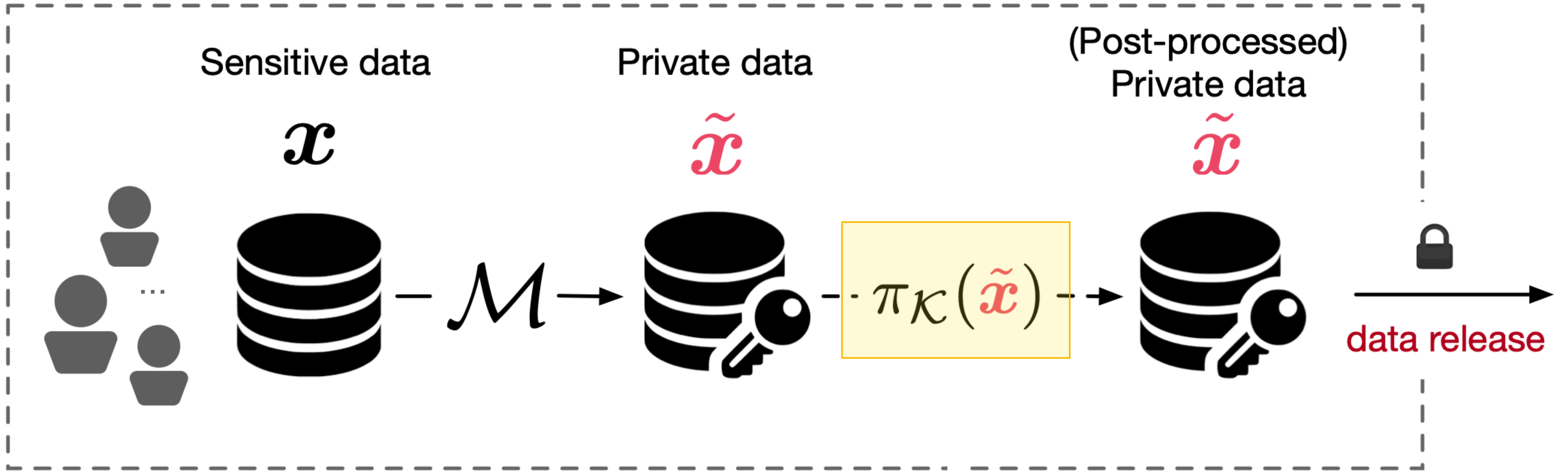
$$\pi_{\mathcal{K}}(\tilde{\boldsymbol{x}}) : \operatorname*{argmin}_{\boldsymbol{v} \in \mathcal{K}} \boxed{\|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2}$$

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



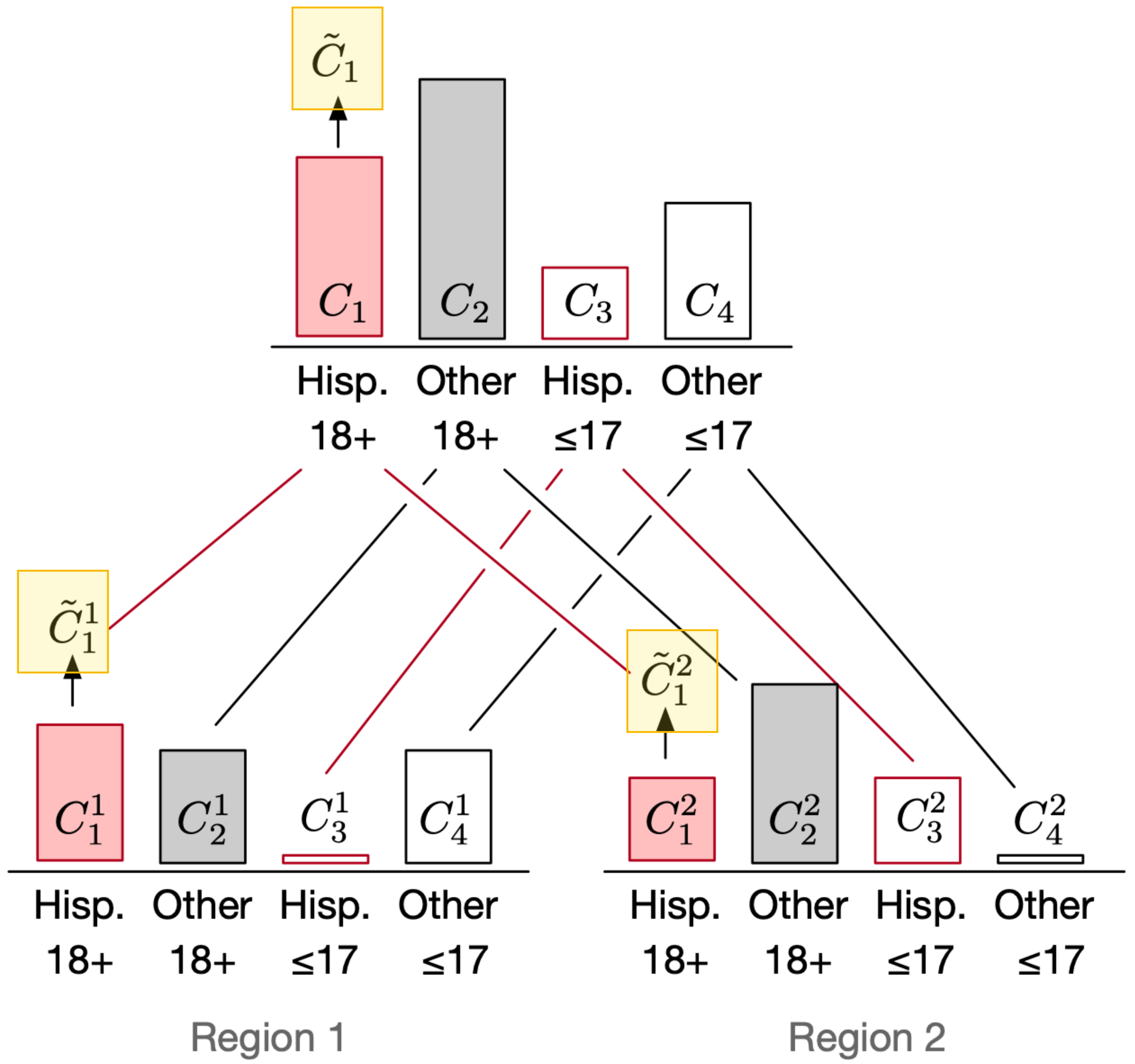

1. Apply noise with appropriate parameter $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \text{Noise}$

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

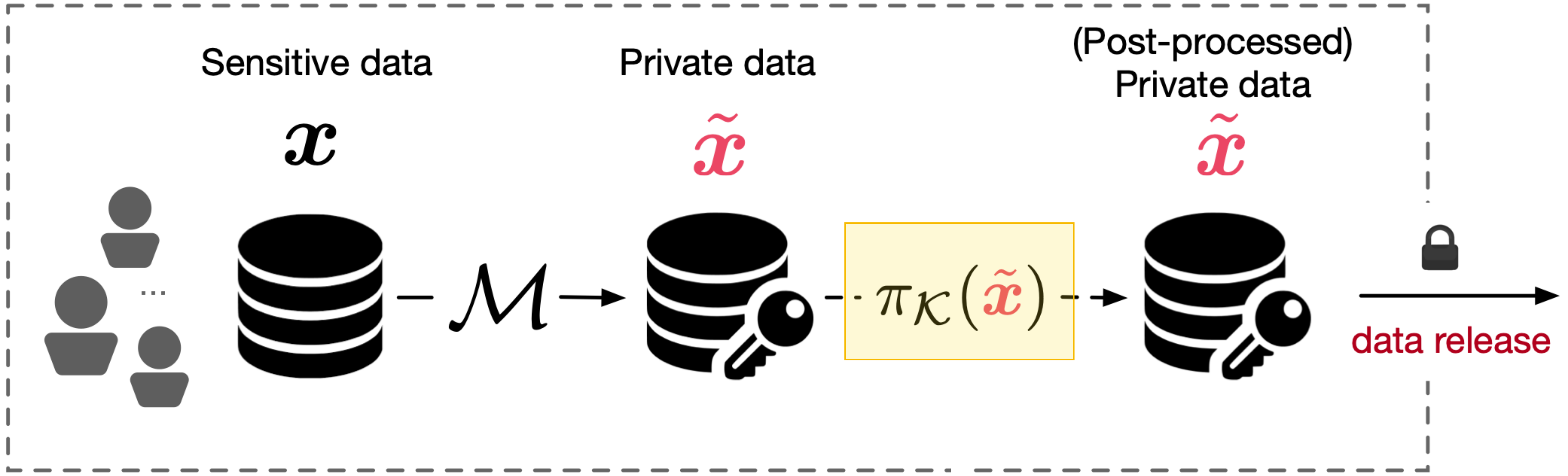
$$\pi_{\mathcal{K}}(\tilde{\boldsymbol{x}}) : \operatorname*{argmin}_{\boldsymbol{v} \in \mathcal{K}} \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2$$

with feasible region defined as

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



# DP data release with post-processing



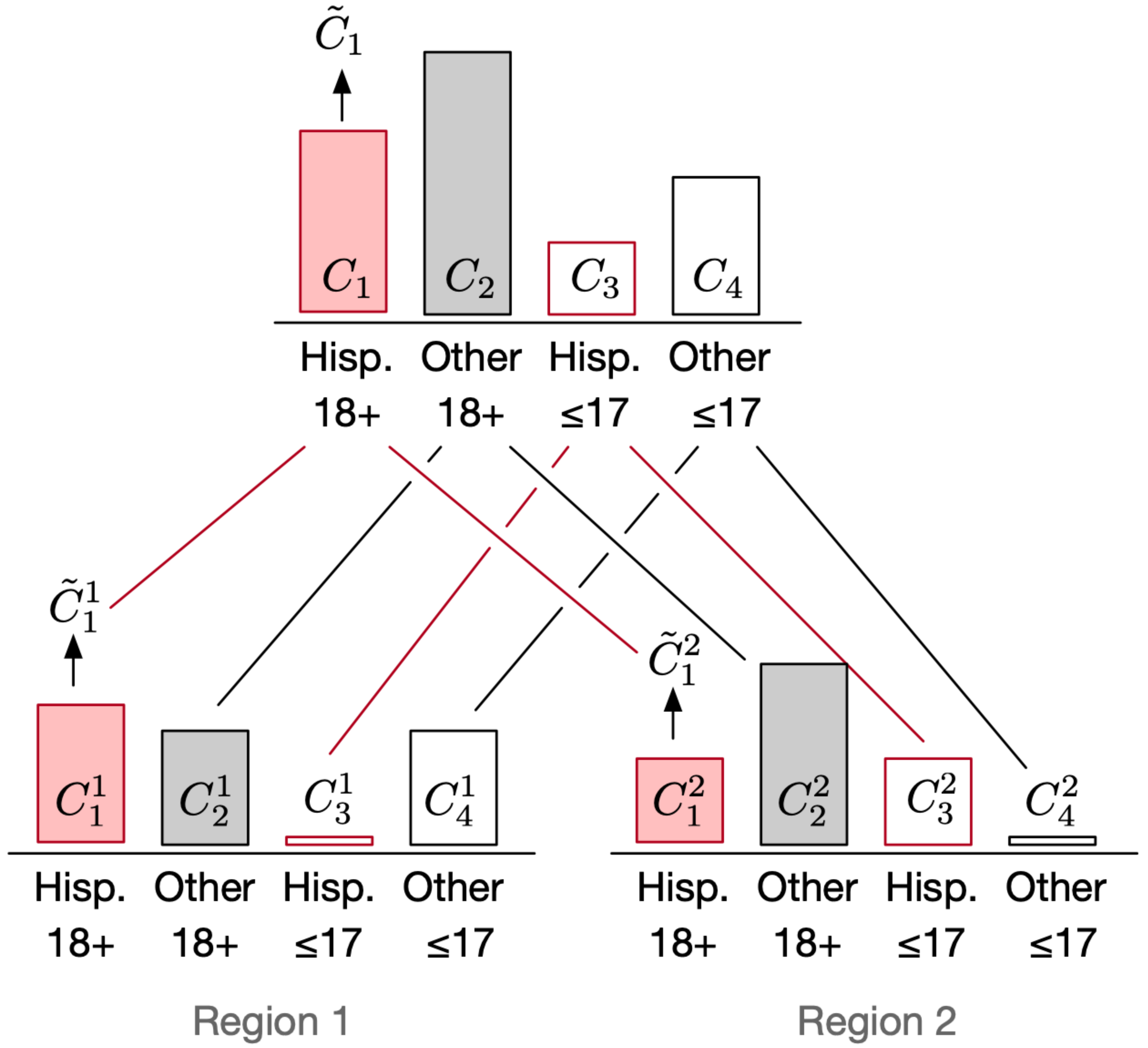

1. Apply noise with appropriate parameter $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \text{Noise}$

2. Post-process output $\tilde{x}$ 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

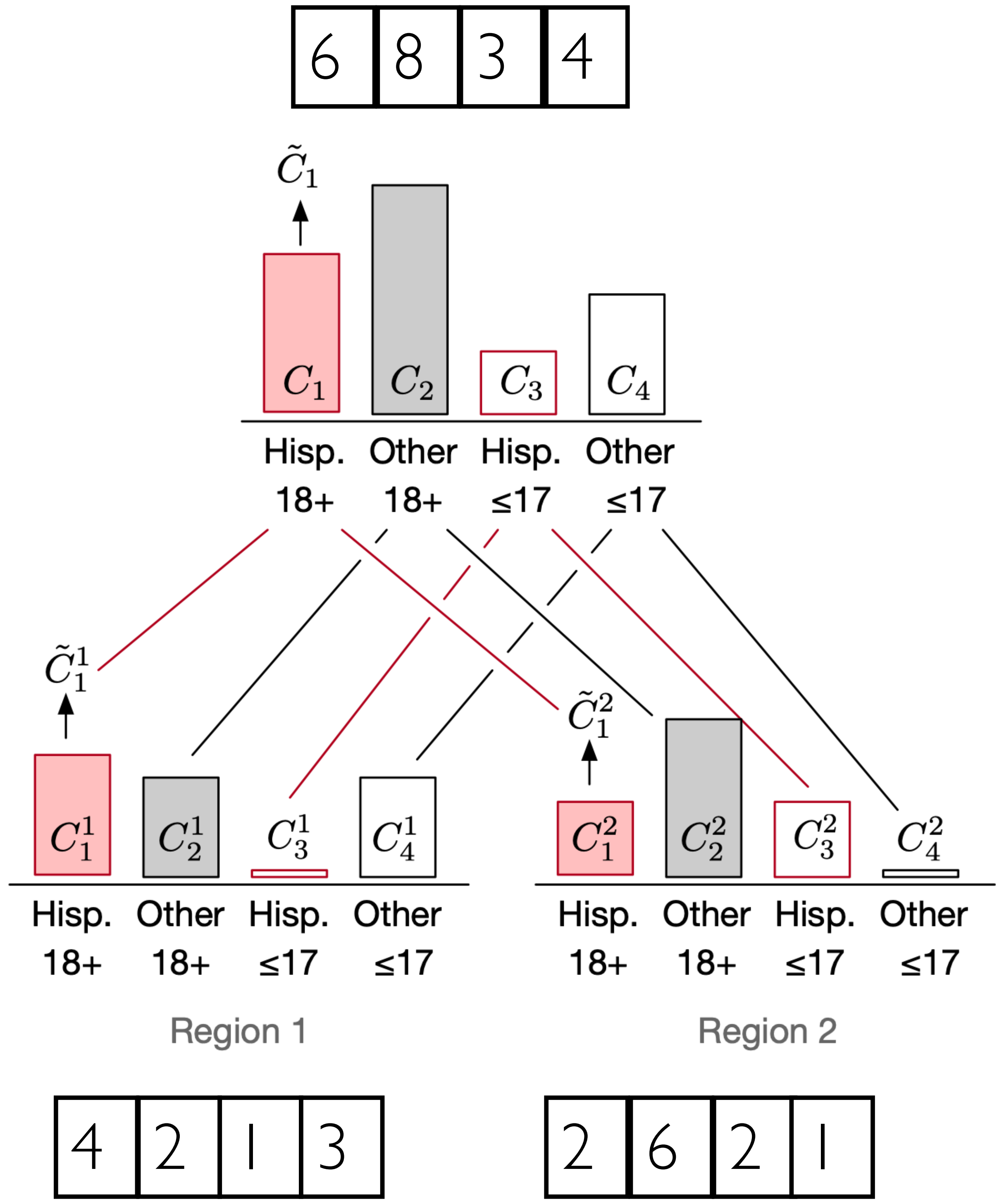
$$\mathcal{K} = \left\{ \boldsymbol{v} \mid \sum_{i=1}^{n} v_i = C, \boxed{\boldsymbol{v} \geq 0} \right\}$$




Satisfies DP due to post-processing immunity

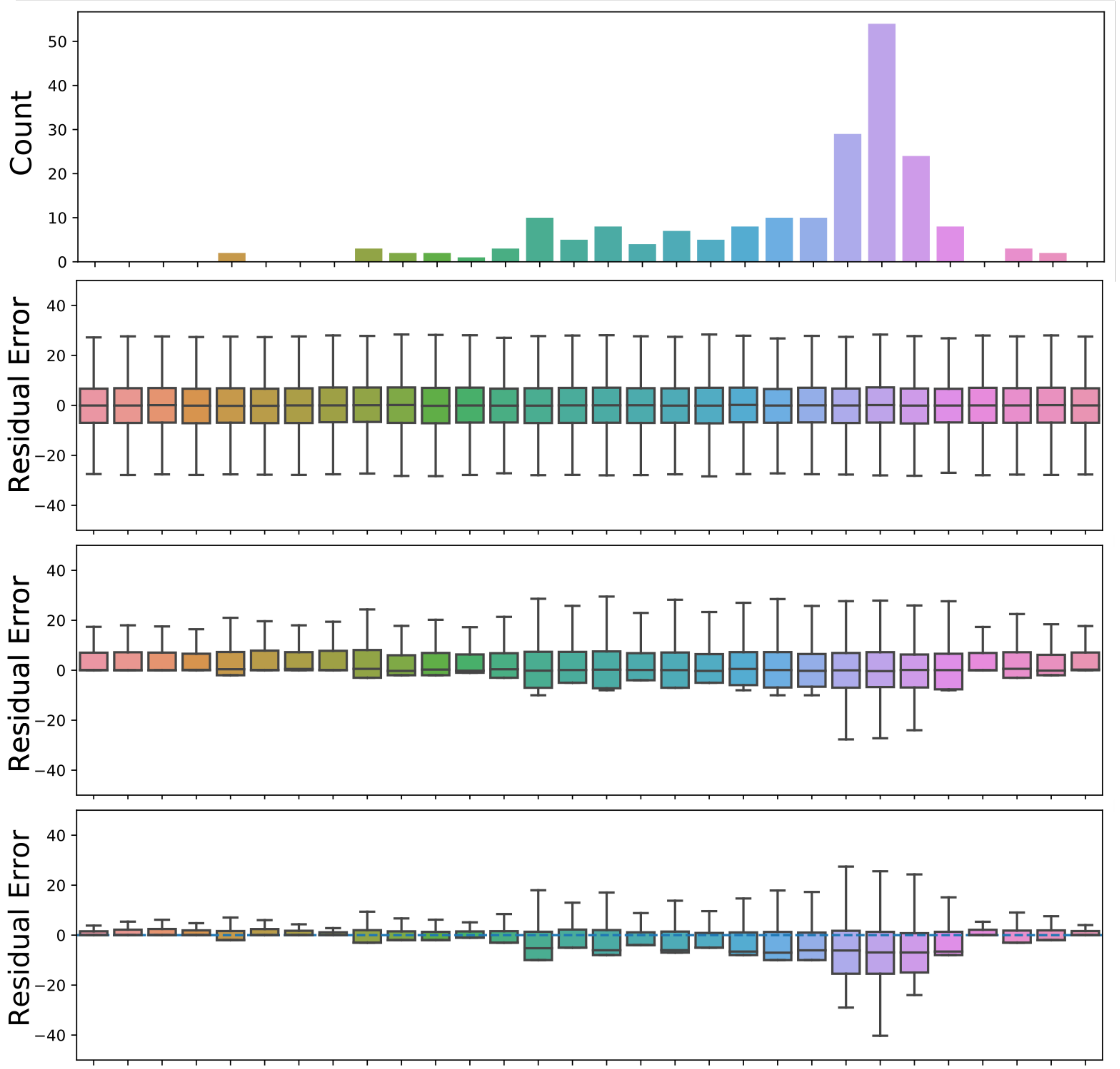
# DP post-processing

## Error and bias




Laplace mechanism

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




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

# DP post-processing

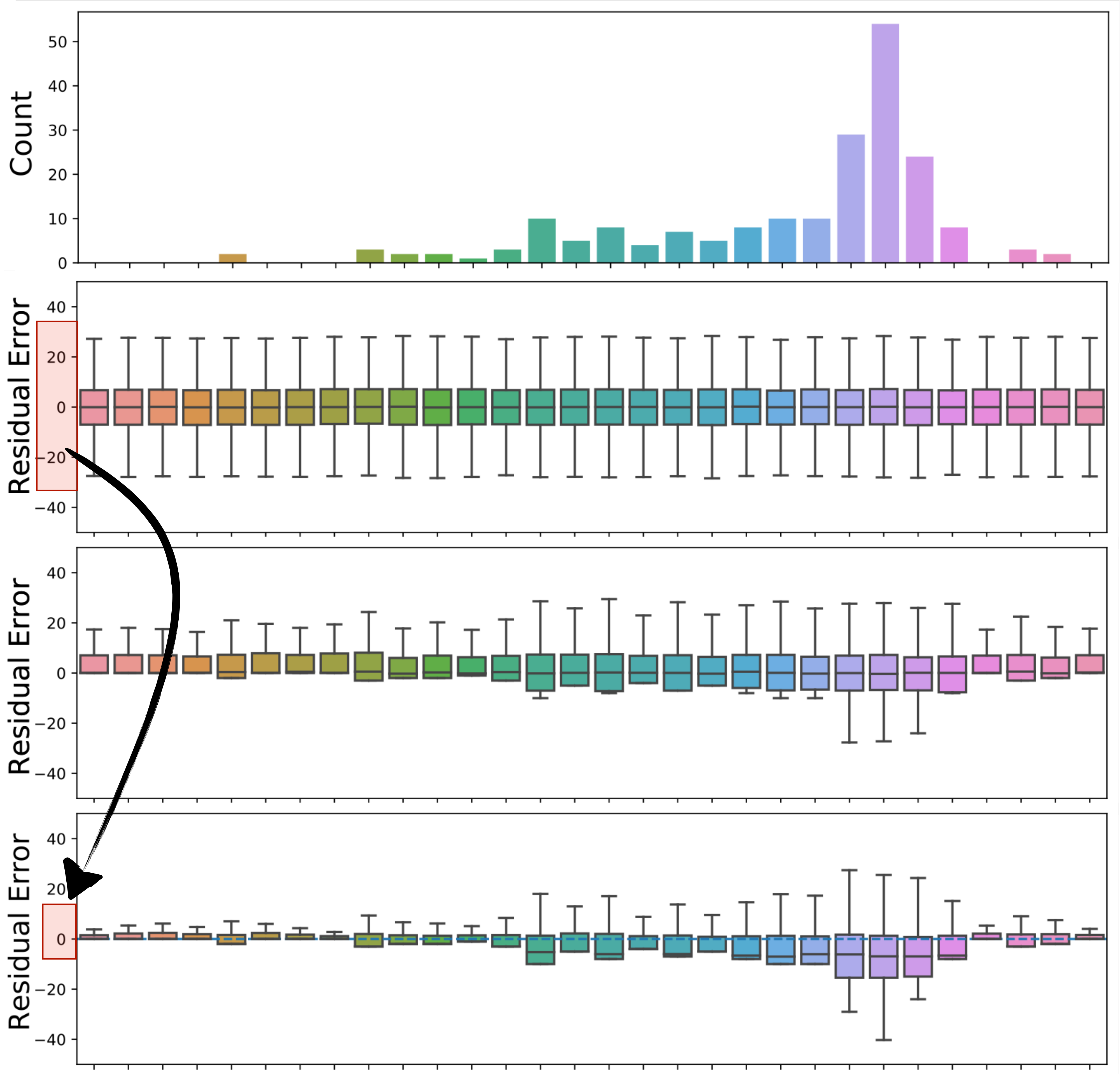
## Error and bias

Observe that post-processing reduces the errors.

Laplace mechanism

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

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



# DP post-processing

## Error and bias

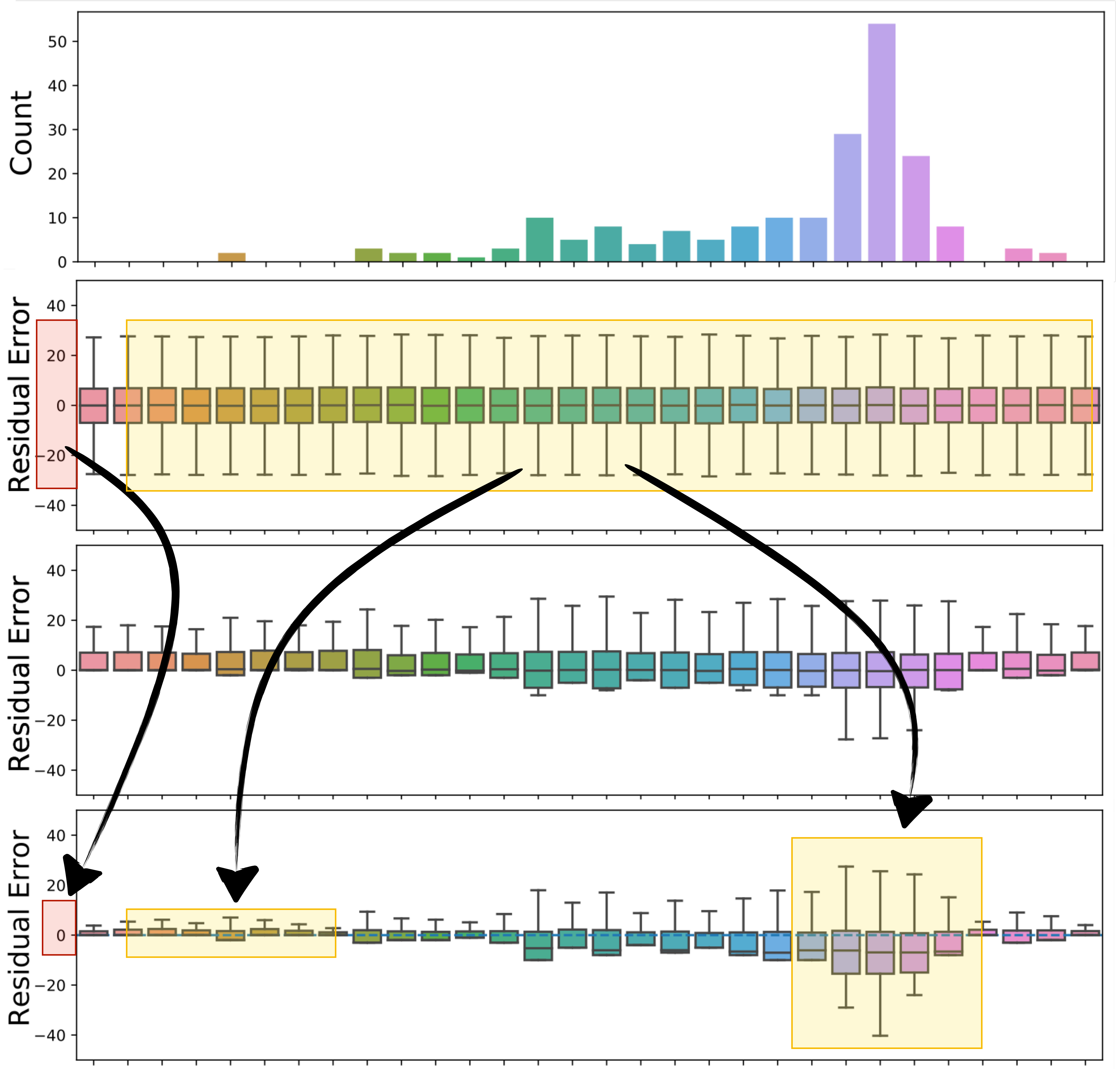
Observe that post-processing reduces the errors.

However, it increases unfairness!

Laplace mechanism

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

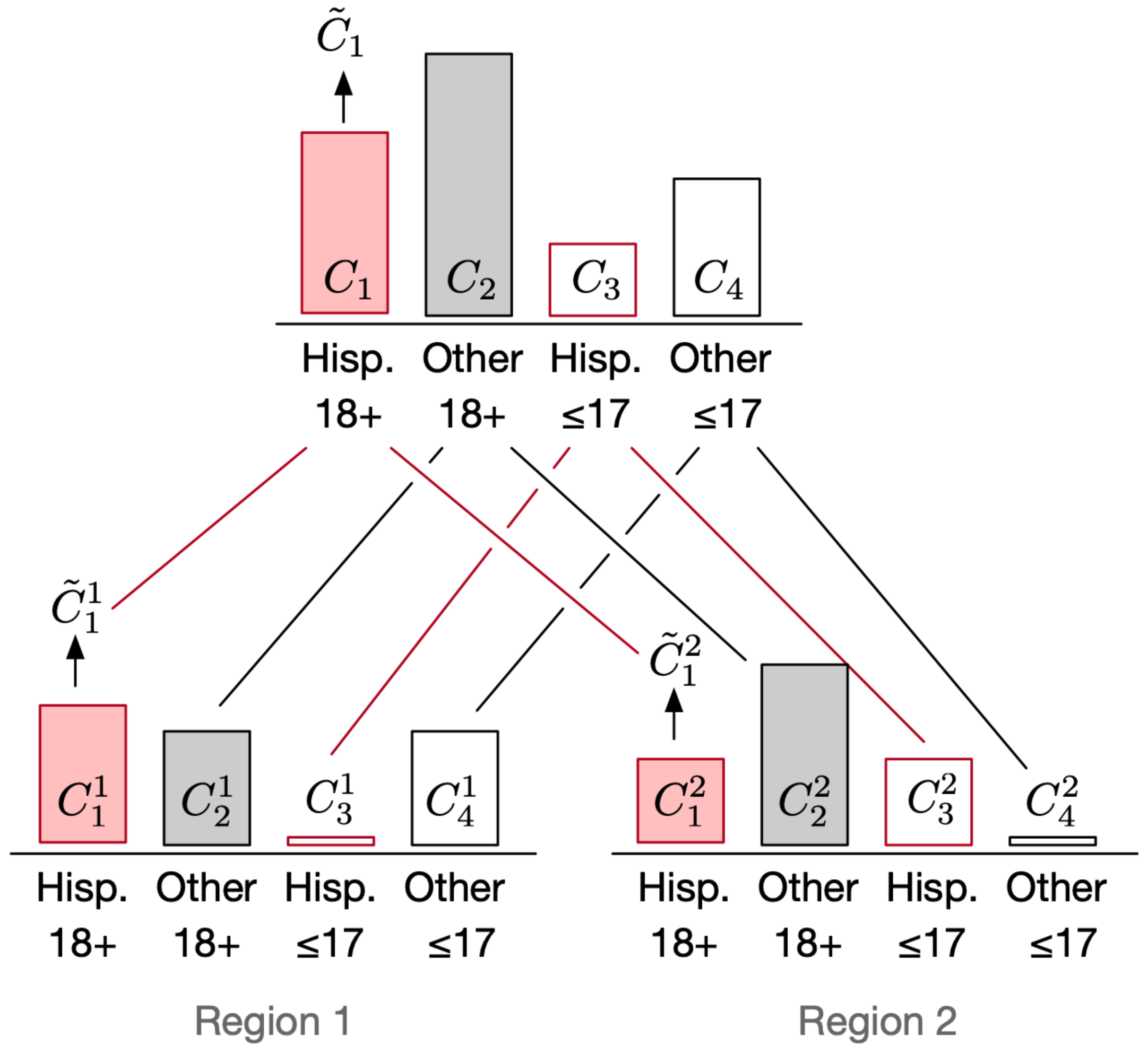
$$\pi_{\mathcal{K}_S} := \operatorname*{argmin}_{\boldsymbol{v} \in \mathcal{K}_S} \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2 \ , \ \mathcal{K}_S = \{\boldsymbol{v} \in \mathbb{R}^n \mid \sum_i v_i = \tilde{S}, v_i \geq 0\},$$



# Bias of post-processing

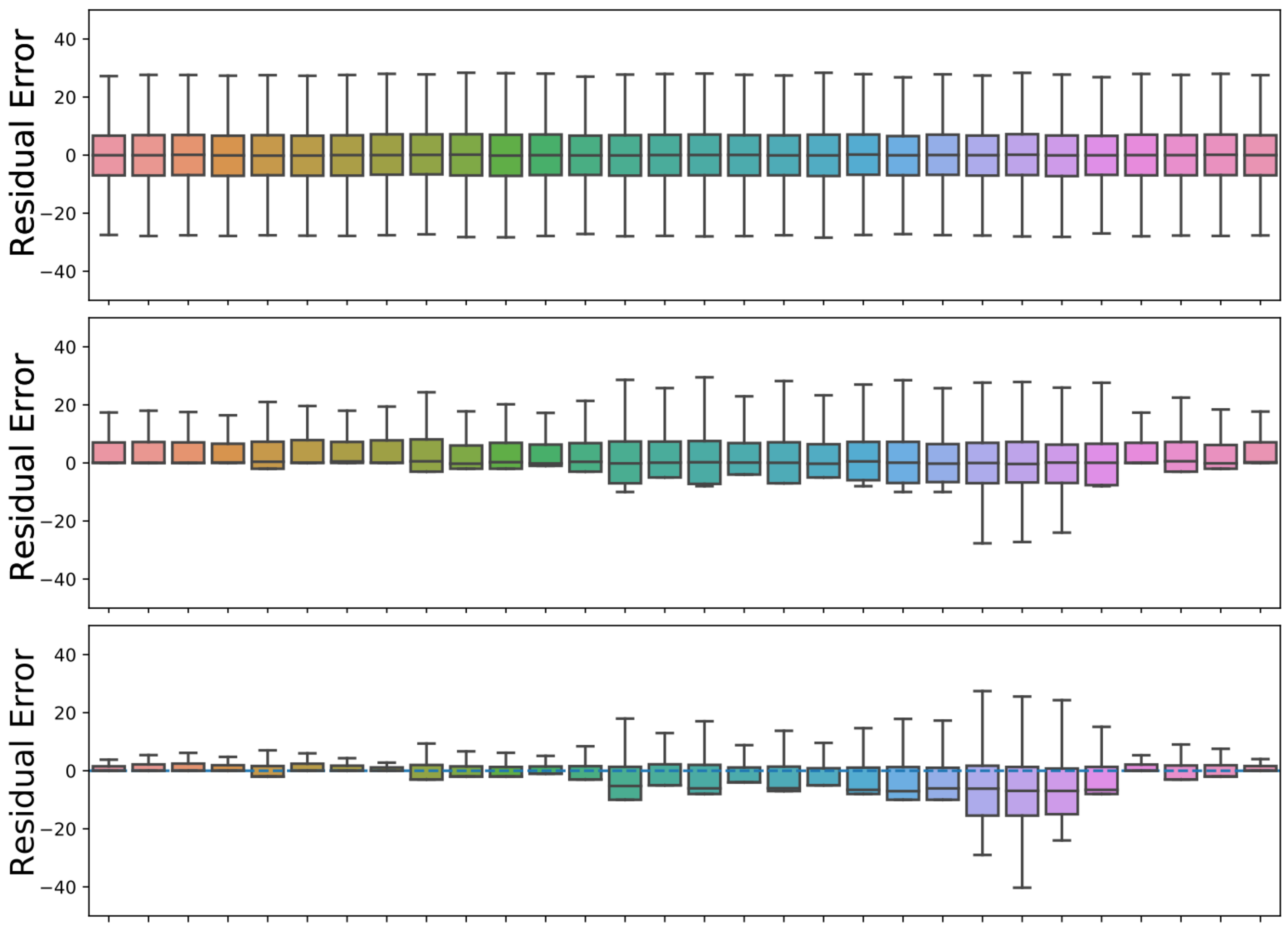
## Key result

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




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

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

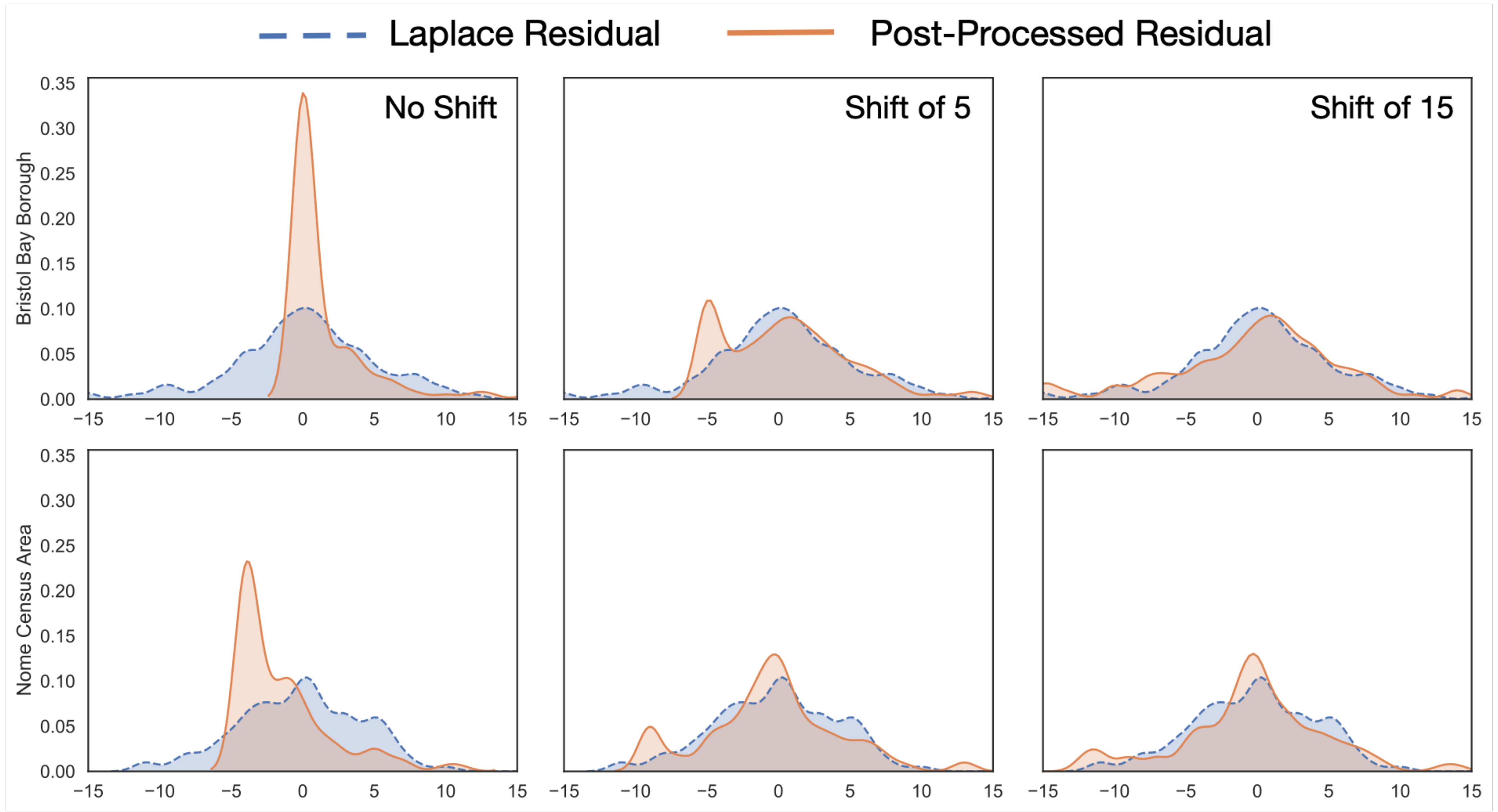


# Quantifying bias in post-processing

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

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

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

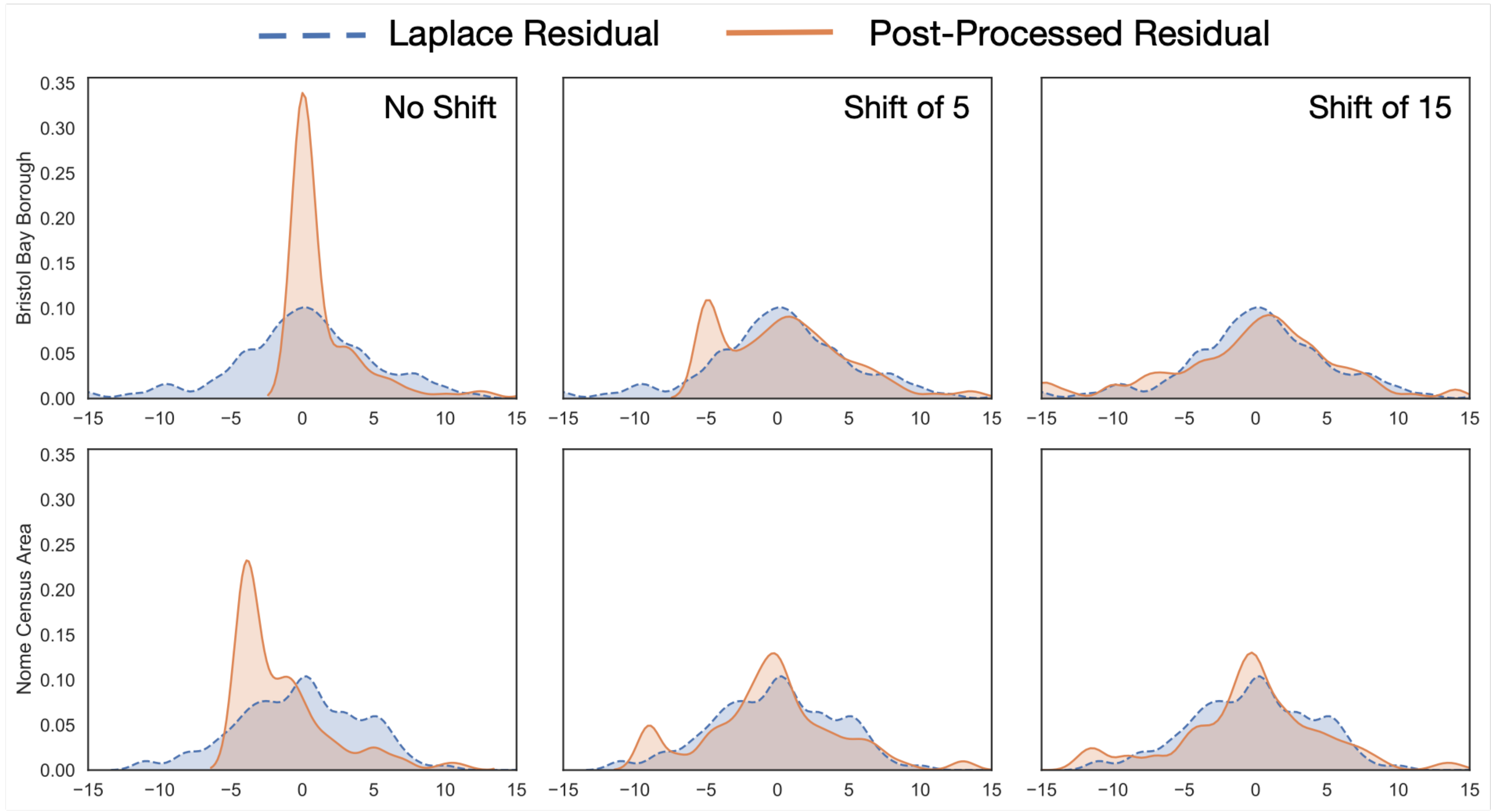



There is an $\ell_1$-ball of radius $r_m = \min_i x_i$ and centered in x which is a feasible subspace where there is no bias.

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

# 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!

# DP post-processing

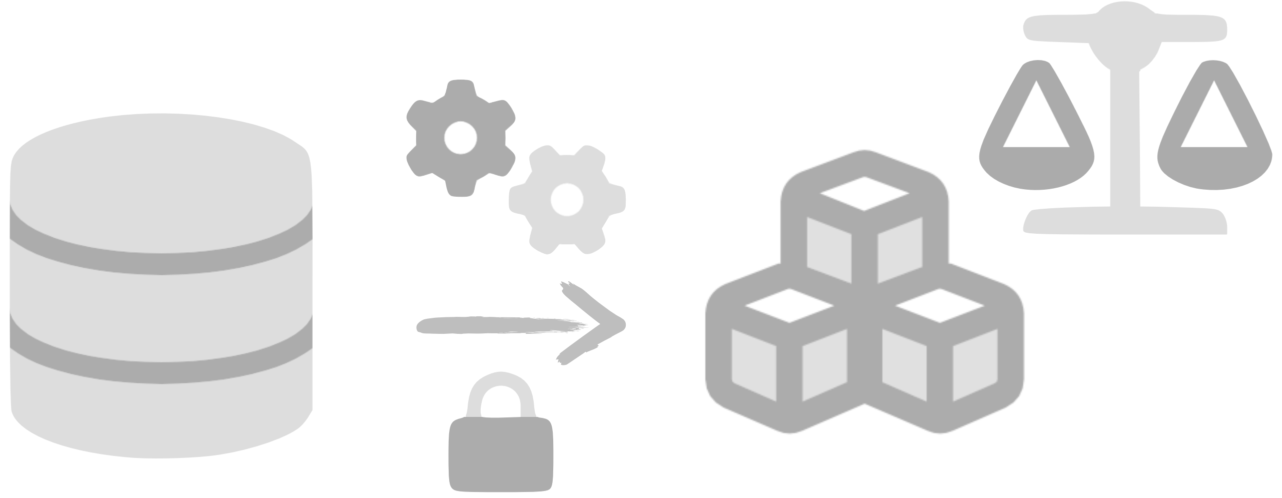
## Important conclusion

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

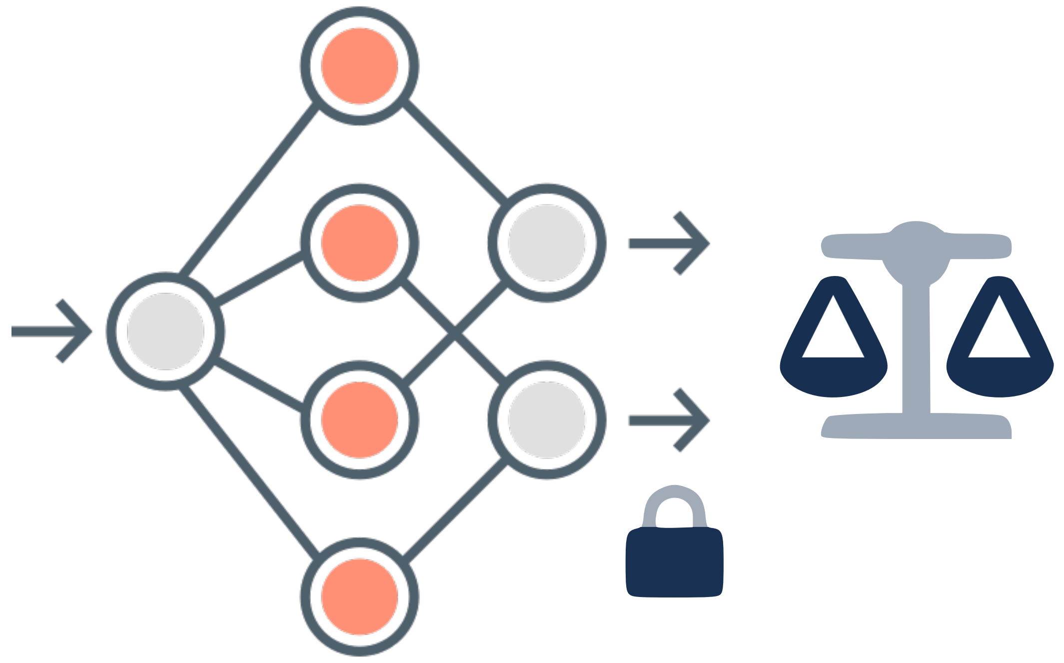
# Agenda

Preliminaries

Fairness impacts of DP in decision making

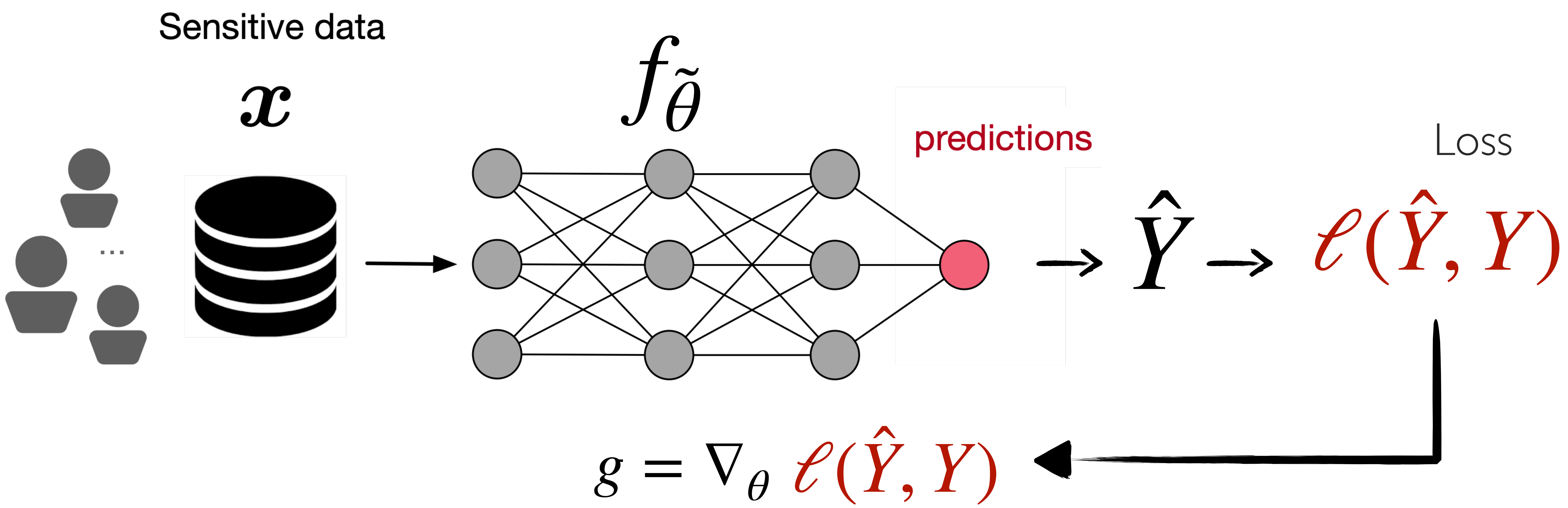


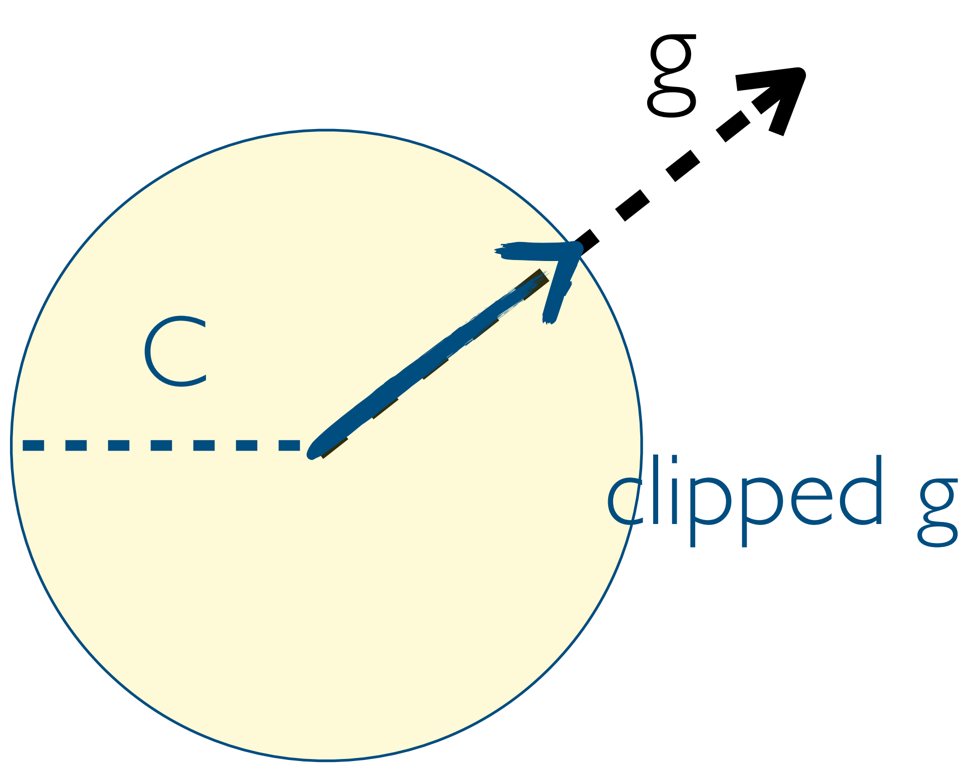
Fairness impacts of DP in learning



What's next?

# DP Stochastic Gradient Descent




1. Clip g in $\ell_2$ ball of size C
2. Add noise from $\mathcal{N}\left(0, IC^2\sigma^2\right)$ to the aggregated gradients in a mini-batch
3. Update parameters $\theta$

# Fairness issues in DP-SGD

**Theorem:** Consider an ERM problem with twice differentiable loss w.r.t. the model parameters. The expected loss of a group a at iteration t+1 is:

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1}; D_a)\right] = \underbrace{\mathcal{L}(\boldsymbol{\theta}_t; D_a) - \eta\left\langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D \right\rangle + \frac{\eta^2}{2}\mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B\right]}_{\textit{non-private term}}$$

$$+ \underbrace{\eta\left(\left\langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D \right\rangle - \left\langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_D \right\rangle\right) + \frac{\eta^2}{2}\left(\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] - \mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B\right]\right)}_{\textit{private term due to clipping}} \qquad (R_a^{\text{clip}})$$

$$+ \underbrace{\frac{\eta^2}{2}\operatorname{Tr}(\boldsymbol{H}_\ell^a) C^2 \sigma^2}_{\textit{private term due to noise}} \qquad (R_a^{\text{noise}})$$

$$+ \; O(\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|^3),$$

*where the expectation is taken over the randomness of the private noise and the mini-batch selection, and the terms $\boldsymbol{g}_Z$ and $\bar{\boldsymbol{g}}_Z$ denote, respectively, the average non-private and private gradients over subset $Z$ of $D$ at iteration $t$ (the iteration number is dropped for ease of notation).*

# Shameless plug

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

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

Watch video

Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence
Survey Track. Pages 5470-5477. https://doi.org/10.24963/ijcai.2022/766

PDF BibTeX

# Shameless plug 2

**New open-access book on DP in the era of AI**

Differential Privacy in Artificial Intelligence

From Theory to Practice

**Ferdinando Fioretto, Pascal Van Hentenryck, James Anderson, Kallista Bonawitz, Konstantinos Chatzikokolakis, Giovanni Cherubin, Graham Cormode, Rachel Cummings, Damien Desfontaines, Liyue Fan, Marco Gaboardi, Marzyeh Ghassemi, Bryant Gipson, Anna Goldenberg, Michael Hay, Peter Kairouz, Steven H. Low, Ashwin Machanavajjhala, Brendan McMahan, Catuscia Palamidessi, Nicolas Papernot, David Pujol, Reza Shokri, Jeremy Seeman, Thomas Steinke, Vinith M. Suriyakumar, Yurii Sushko, Yuchao Tao, Christine Task, Andreas Terzis, Abhradeep Thakurta, Salil Vadhan, Jiayuan Ye, Juba Ziani, Fengyu Zhou**

the essence of knowledge

## Chapter 1 already on ArXiv

# Why clipping causes unfairness?

## Gradient norms and excessive risk

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1}; D_a)\right] = \underbrace{\mathcal{L}(\boldsymbol{\theta}_t; D_a) - \eta \langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D \rangle + \frac{\eta^2}{2} \mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B\right]}_{\text{non-private term}}$$

$$+ \underbrace{\eta \left( \langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D \rangle - \langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_D \rangle \right) + \frac{\eta^2}{2} \left( \mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] - \mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B\right] \right)}_{\text{private term due to clipping}} \qquad (R_a^{\text{clip}})$$

$$+ \underbrace{\frac{\eta^2}{2} \mathrm{Tr}(\boldsymbol{H}_\ell^a) C^2 \sigma^2}_{\text{private term due to noise}} \qquad (R_a^{\text{noise}})$$

$$+ O(\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|^3),$$

# Why clipping causes unfairness?

## Gradient norms and excessive risk

$$+\underbrace{\eta\left(\langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D\rangle - \langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_D\rangle\right) + \frac{\eta^2}{2}\left(\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] - \mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B\right]\right)}_{\textit{private term due to clipping}} \qquad (R_a^{\text{clip}})$$

# Why clipping causes unfairness?

## Gradient norms and excessive risk

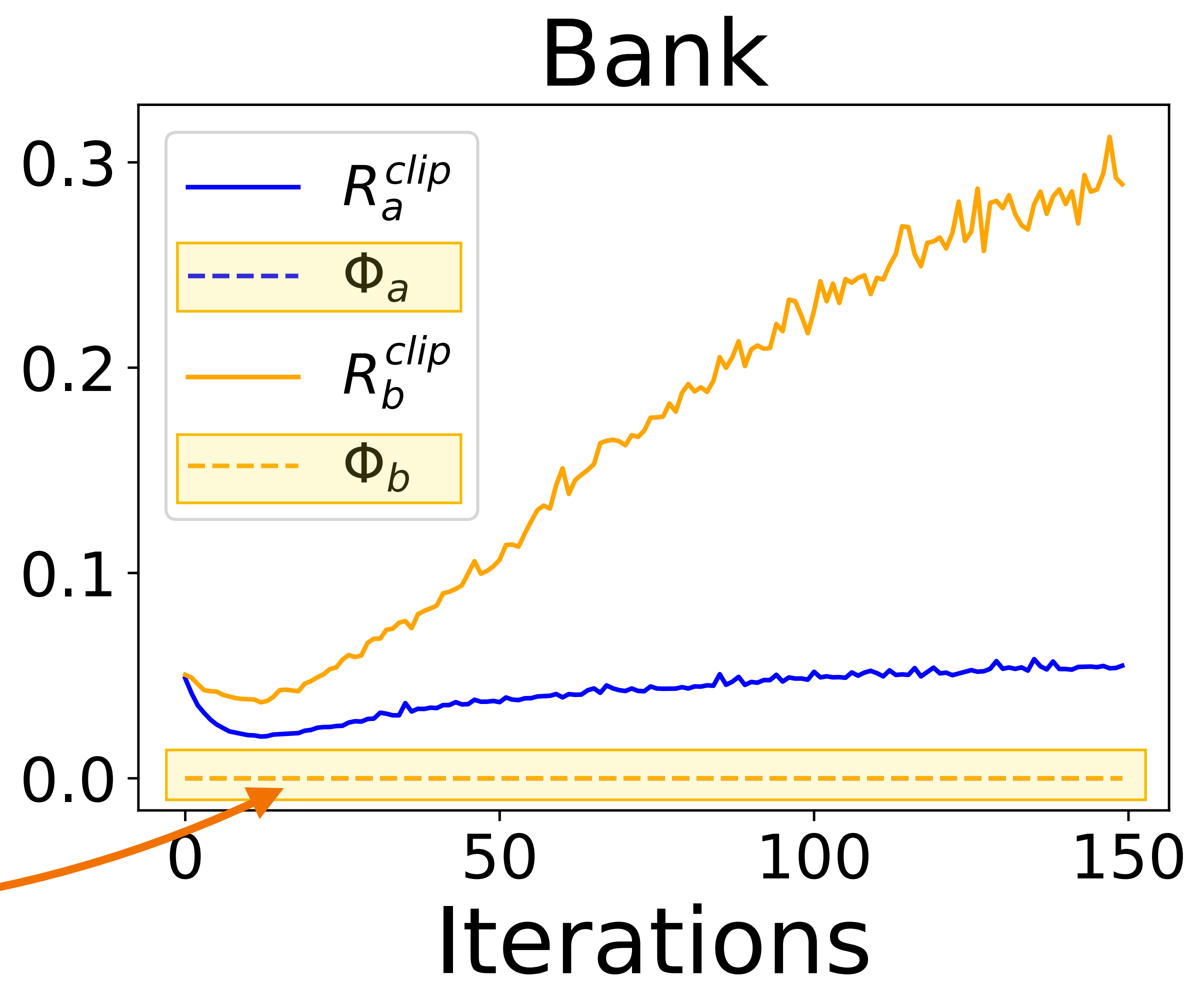
$$+\underbrace{\eta\left(\langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D\rangle - \langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_D\rangle\right) + \frac{\eta^2}{2}\left(\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] - \mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B\right]\right)}_{\textit{private term due to clipping}} \qquad (R_a^{\text{clip}})$$

# Why clipping causes unfairness?

## Gradient norms and excessive risk

Negligible impact

$$+\underbrace{\eta\left(\langle \boxed{\boldsymbol{g}_{D_a}}, \boldsymbol{g}_D\rangle - \langle \boxed{\boldsymbol{g}_{D_a}}, \bar{\boldsymbol{g}}_D\rangle\right) + \boxed{\frac{\eta^2}{2}\left(\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] - \mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B\right]\right)}}_{\textit{private term due to clipping}} \qquad (R_a^{clip})$$

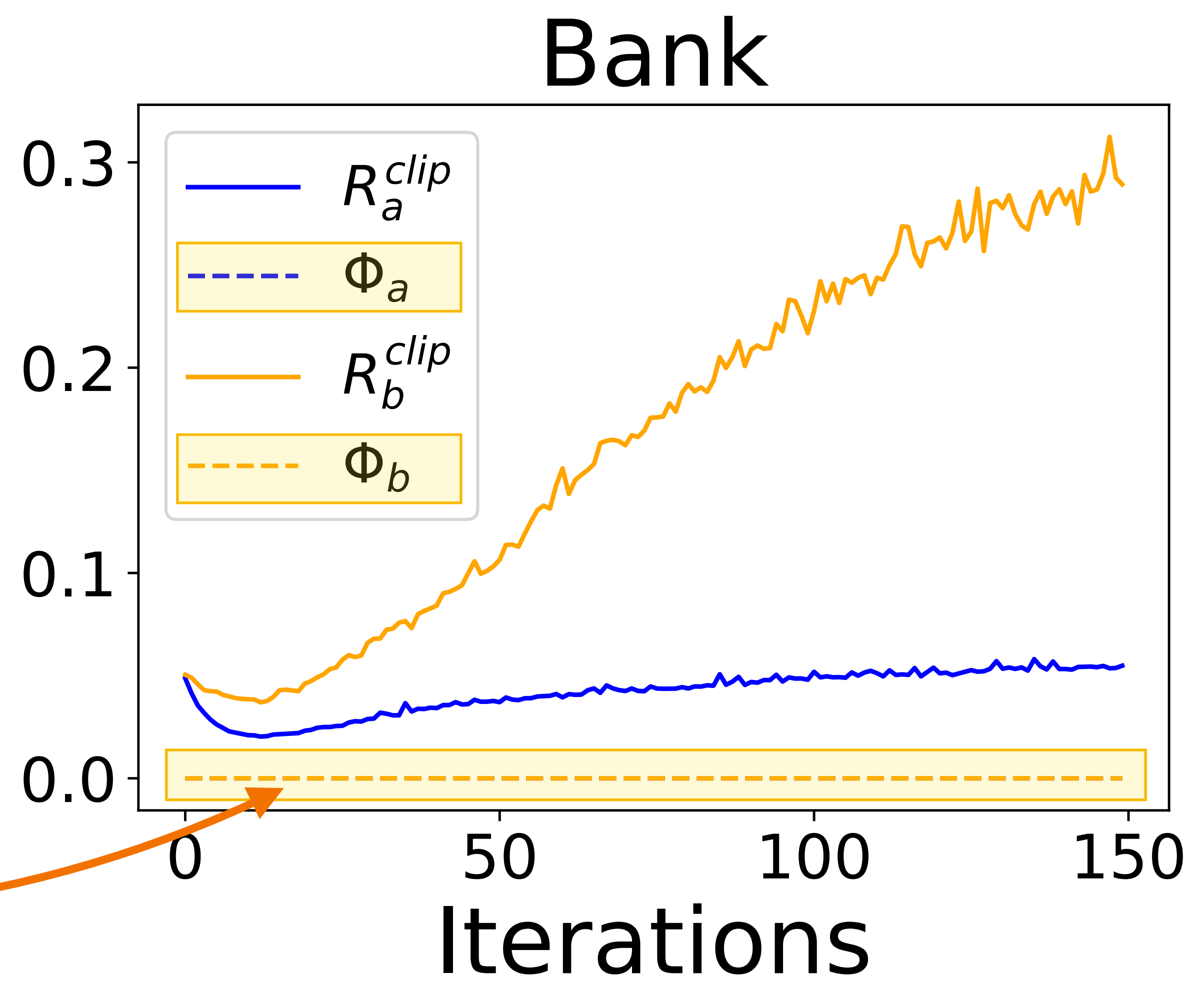


# Why clipping causes unfairness?

## Gradient norms and excessive risk

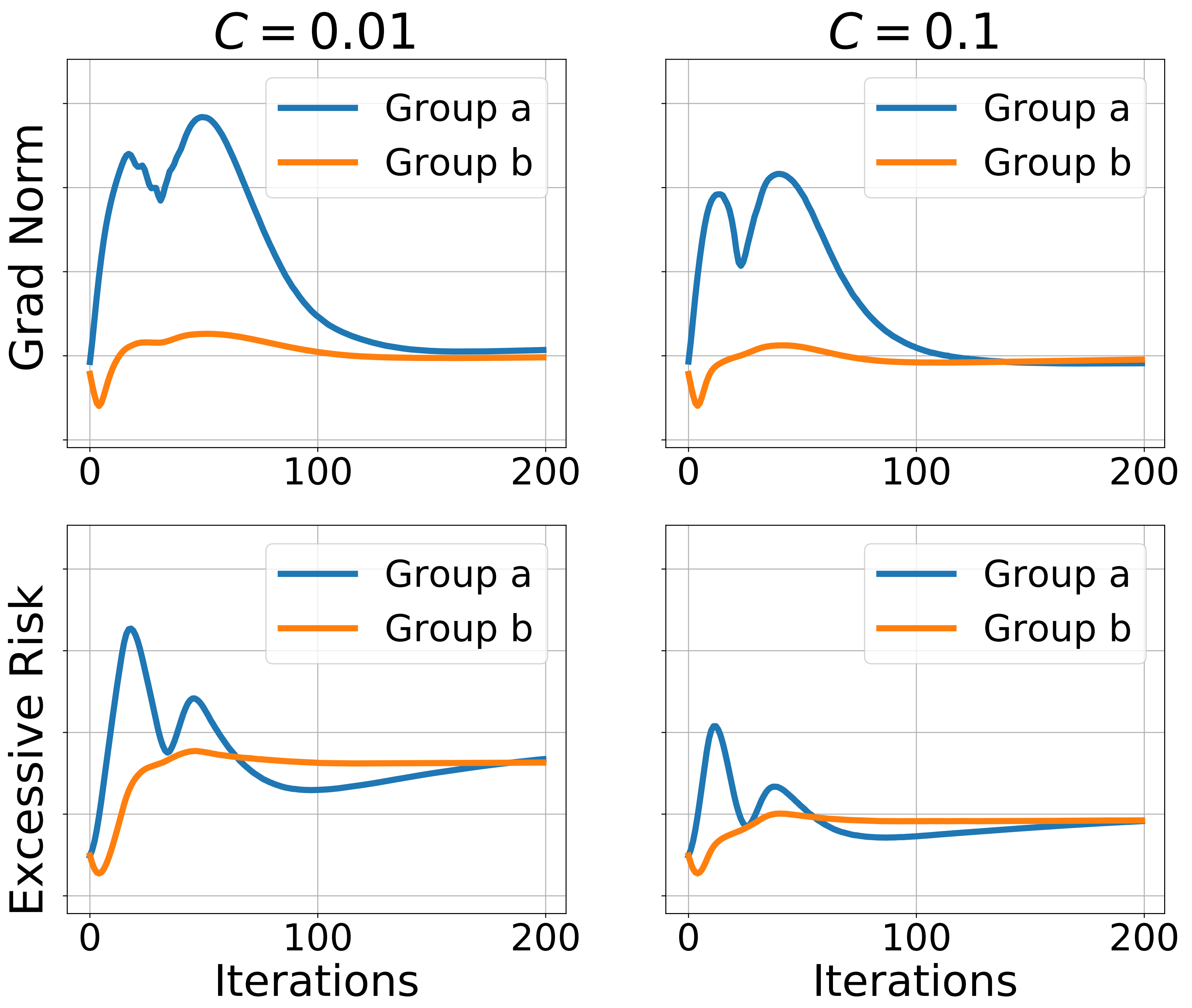
Negligible impact

$$\underbrace{\eta\left(\langle \boxed{\boldsymbol{g}_{D_a}}, \boldsymbol{g}_D\rangle - \langle \boxed{\boldsymbol{g}_{D_a}}, \bar{\boldsymbol{g}}_D\rangle\right) + \boxed{\frac{\eta^2}{2}\left(\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] - \mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B\right]\right)}}_{\textit{private term due to clipping}}$$




Crucial Proxy to Unfairness (due to clipping)

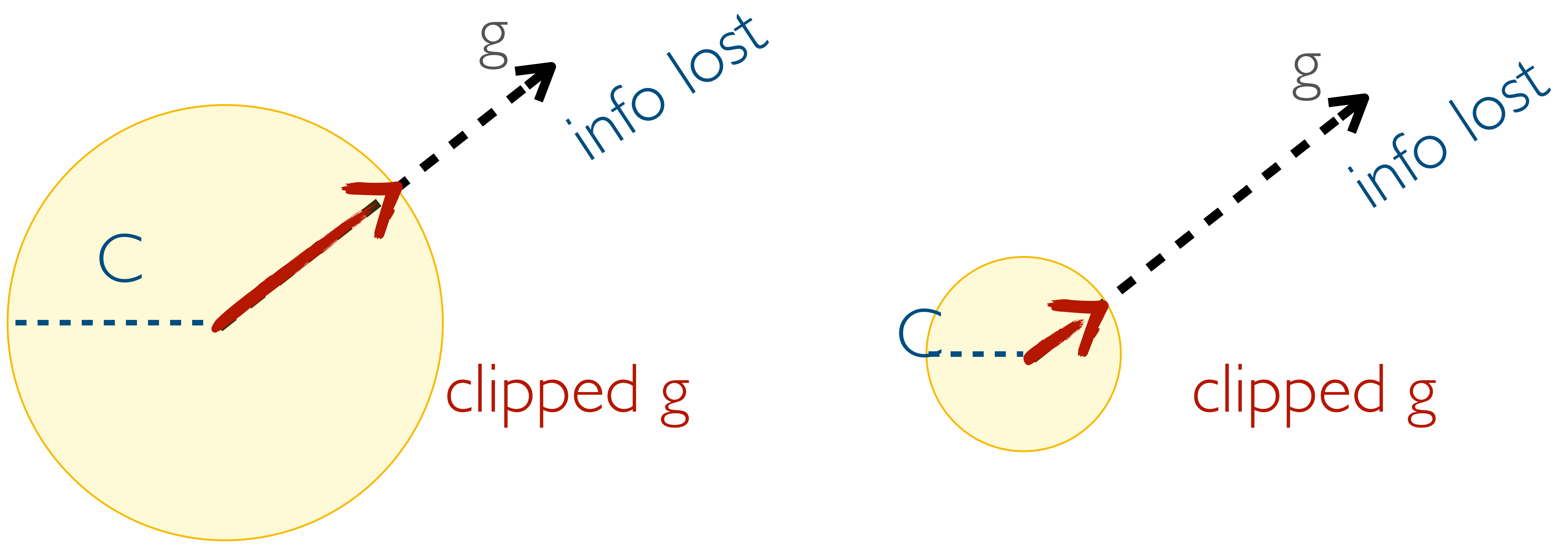
| |
|---|
| Theorem (informal): Gradient flow affects the excessive risk (unfairness) of the individuals and groups. |



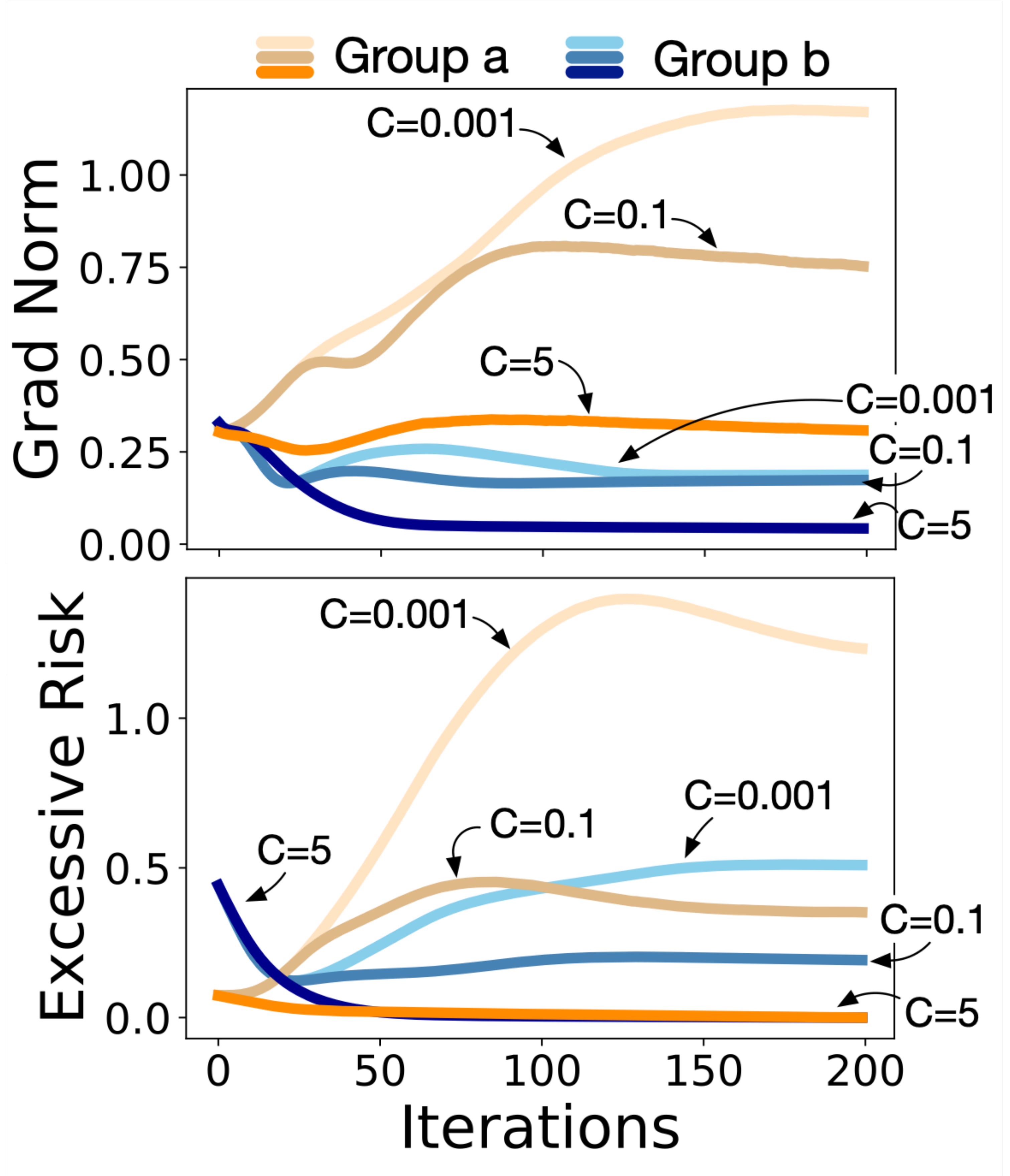
# Why clipping causes unfairness?

## Gradient norms and excessive risk

- When clipping, the smaller C, the higher is the information loss of the average gradients that are backpropagated.



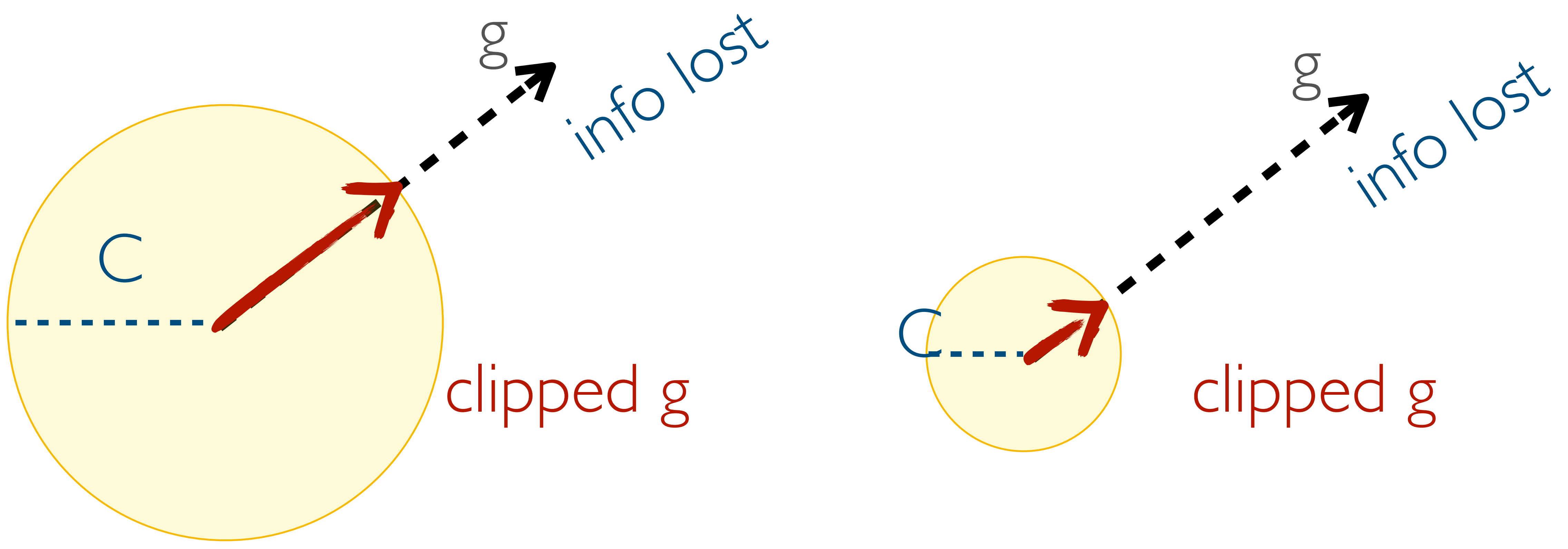

Impact of gradient clipping
(Bank dataset)



# Why clipping causes unfairness?

## Gradient norms and excessive risk

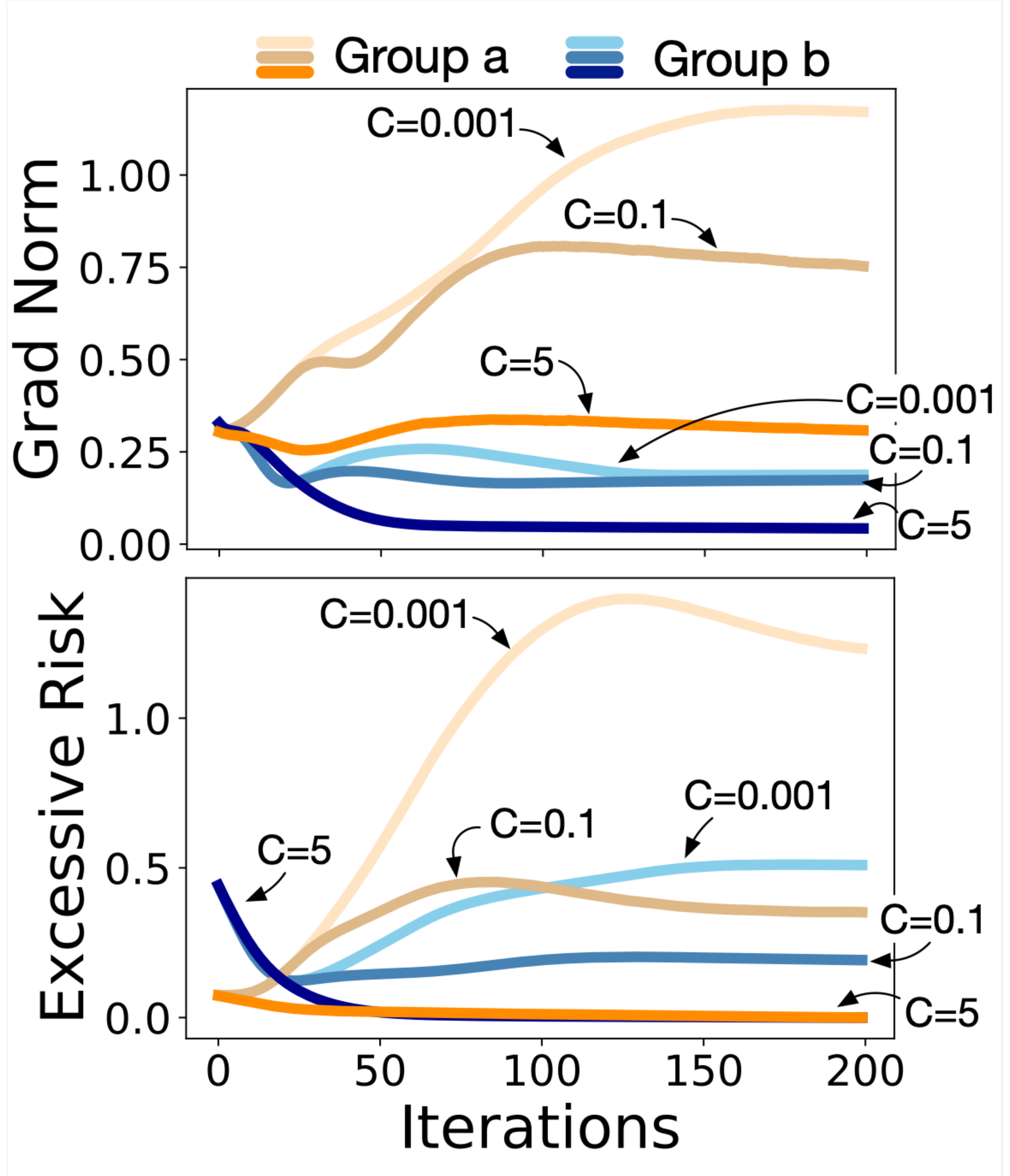
- When clipping, the smaller C, the higher is the information loss of the average gradients that are backpropagated.




**Theorem:** *Let $p_z = |D_z|/|D|$ be the fraction of training samples in group $z \in \mathcal{A}$. For groups $a, b \in \mathcal{A}$, $R_a^{clip} > R_b^{clip}$ whenever:*

$$\|g_{D_a}\|\left(p_a - \frac{p_a^2}{2}\right) \geq \frac{5}{2}C + \|g_{D_b}\|\left(1 + p_b + \frac{p_b^2}{2}\right).$$

Impact of gradient clipping
(Bank dataset)



# Why noise causes unfairness in DP-SGD?

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1}; D_a)\right] = \underbrace{\mathcal{L}(\boldsymbol{\theta}_t; D_a) - \eta \left\langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D \right\rangle + \frac{\eta^2}{2}\mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B\right]}_{\textit{non-private term}}$$

$$+ \underbrace{\eta \left( \left\langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D \right\rangle - \left\langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_D \right\rangle \right) + \frac{\eta^2}{2}\left( \mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] - \mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B\right] \right)}_{\textit{private term due to clipping}} \qquad (R_a^{\text{clip}})$$

$$+ \underbrace{\frac{\eta^2}{2} \operatorname{Tr}(\boldsymbol{H}_\ell^a) C^2 \sigma^2}_{\textit{private term due to noise}} \qquad (R_a^{\text{noise}})$$

$$+ \; O(\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|^3),$$

# Why noise causes unfairness in DP-SGD?

$$+ \underbrace{\frac{\eta^2}{2} \mathrm{Tr}(\boldsymbol{H}_\ell^a) C^2 \sigma^2}_{\text{private term due to noise}} \qquad \boxed{(R_a^{\text{noise}})}$$

# Why noise causes unfairness in DP-SGD?

## Distance to the decision boundary and excess risk

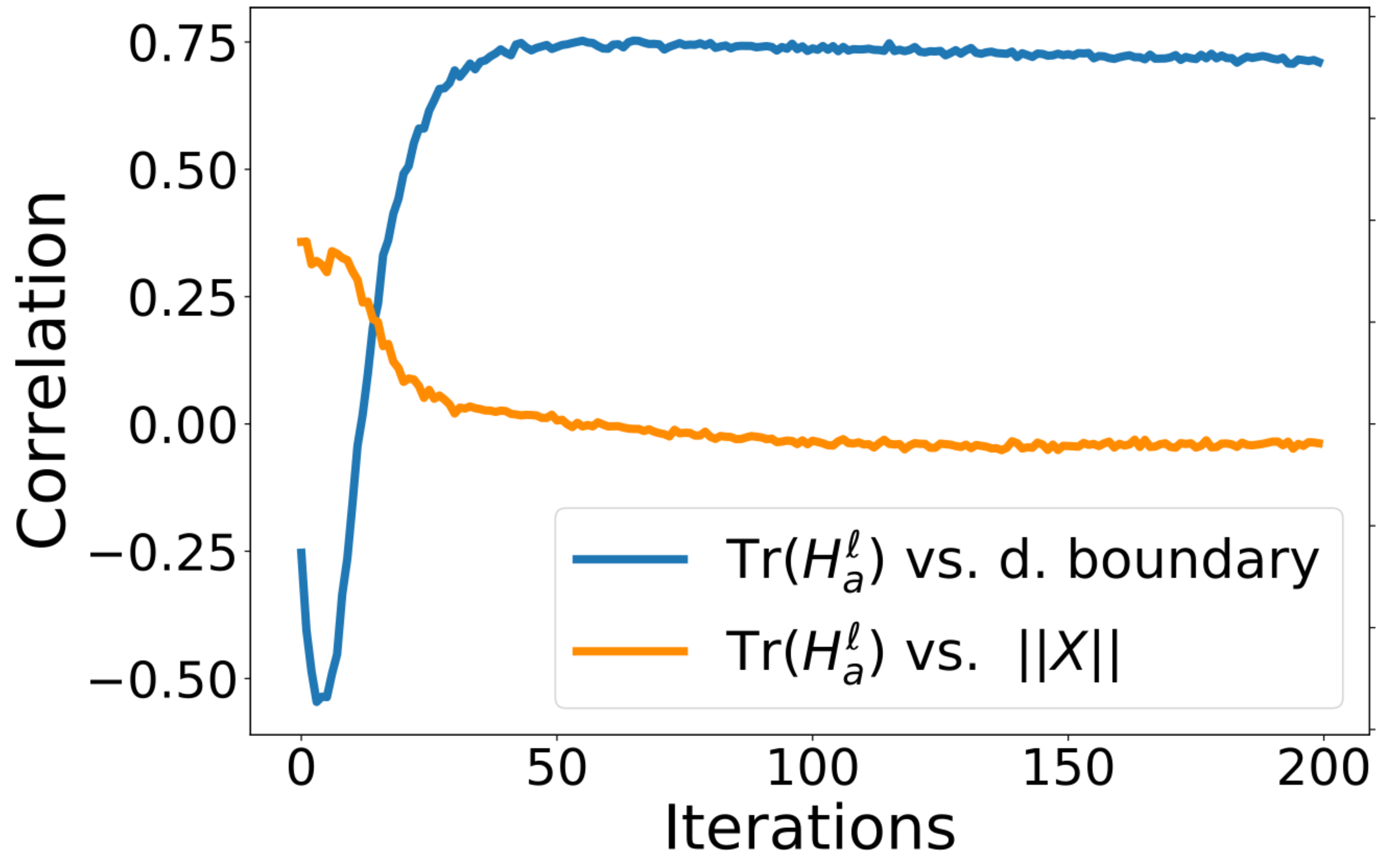
$$+ \underbrace{\frac{\eta^2}{2}\,\mathrm{Tr}(\boldsymbol{H}_\ell^a)C^2\sigma^2}_{\textit{private term due to noise}}$$

# Why noise causes unfairness in DP-SGD?

## Distance to the decision boundary and excess risk

$$+ \quad \underbrace{\frac{\eta^2}{2}\mathrm{Tr}(\boxed{\boldsymbol{H}_\ell^a})C^2\sigma^2}_{\textit{private term due to noise}}$$

Correlation between Hessian trace and closeness to the decision boundary and input norms



# Why noise causes unfairness in DP-SGD?

## Distance to the decision boundary and excess risk

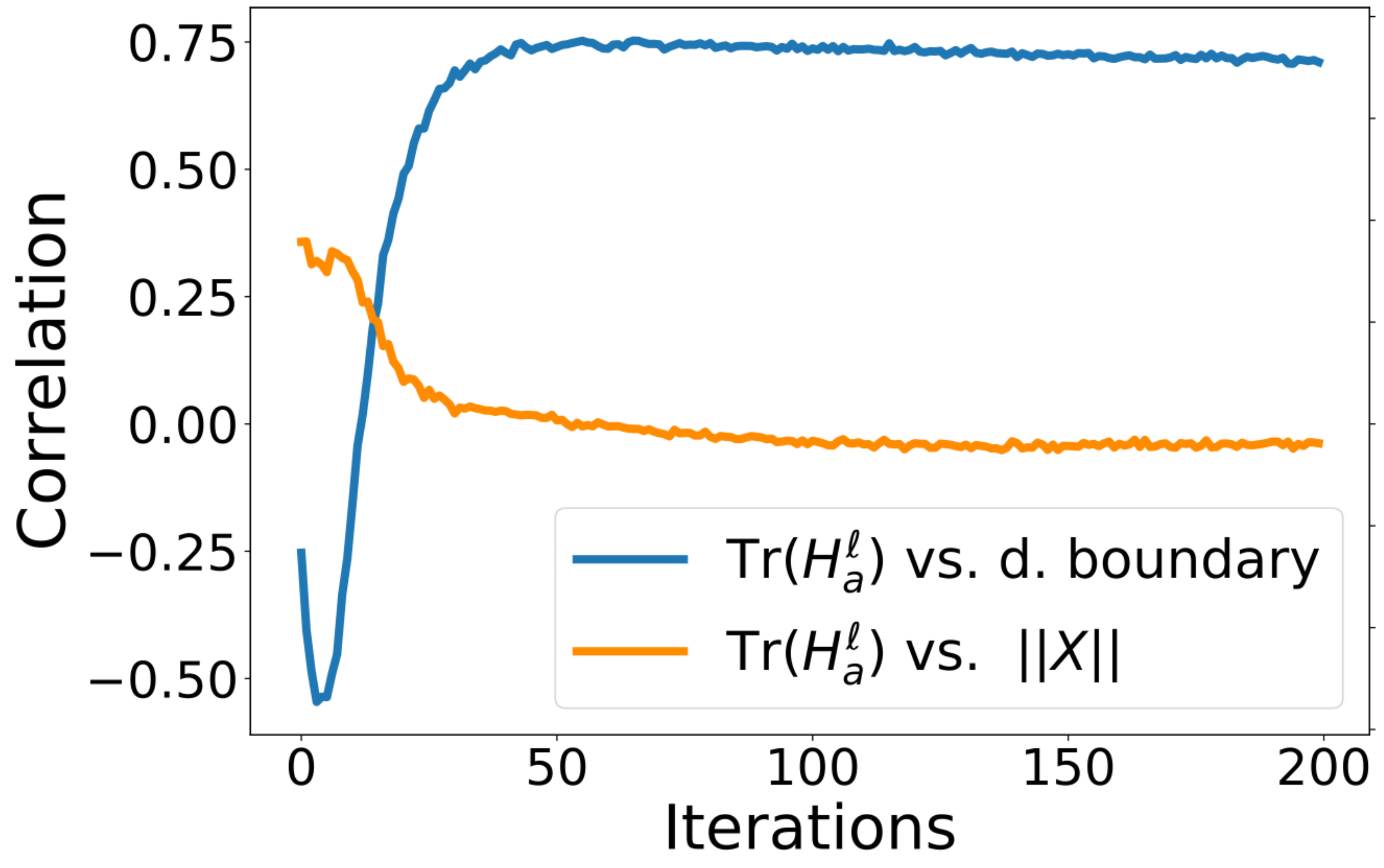
$$\underbrace{\frac{\eta^2}{2}\,\mathrm{Tr}(\boxed{\boldsymbol{H}_\ell^a})C^2\sigma^2}_{\textit{private term due to noise}}$$

Crucial Proxy to Unfairness (due to noise)

**Theorem (informal):** Individuals whose outputs are close to the decision boundary will have higher Hessian traces (high local curvatures of the loss).

Intuitively, the model decisions for samples which are close to the decision boundary are less robust to the presence of noise w.r.t. samples which are farther away from the boundary.

Correlation between Hessian trace and closeness to the decision boundary and input norms

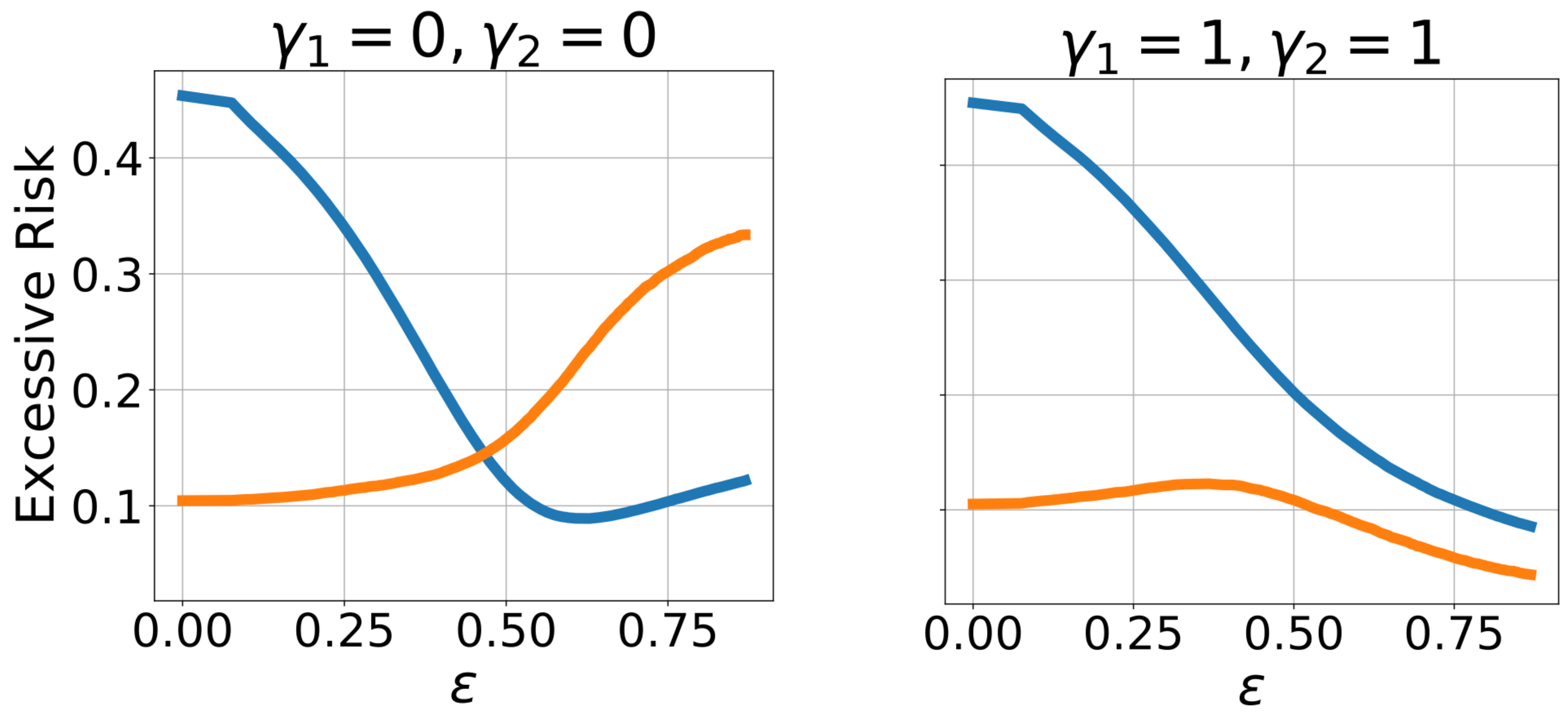


# Mitigating solutions

Modify training so to equalize the factors affecting the excessive risk due to clipping and to noise addition

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; D) + \sum_{a \in \mathcal{A}} \left( \gamma_1 \left| \langle \boldsymbol{g}_{D_a} - \boldsymbol{g}_D, \boldsymbol{g}_{D_a} - \bar{\boldsymbol{g}}_D \rangle \right| + \gamma_2 \left| \mathrm{Tr}(\boldsymbol{H}_\ell^a) - \mathrm{Tr}(\boldsymbol{H}_\ell) \right| \right),$$

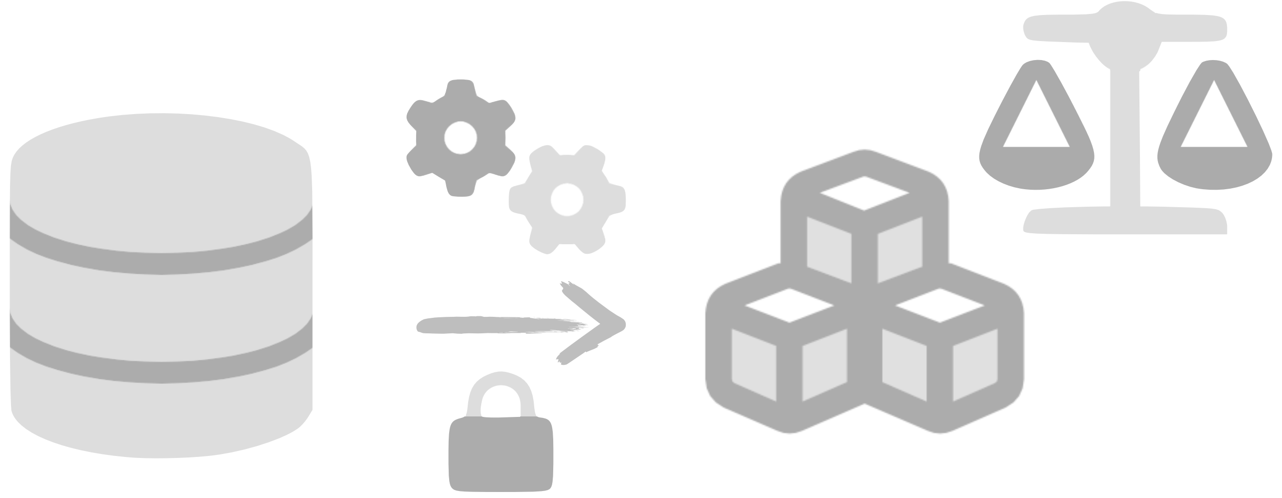
Minority group
Majority group



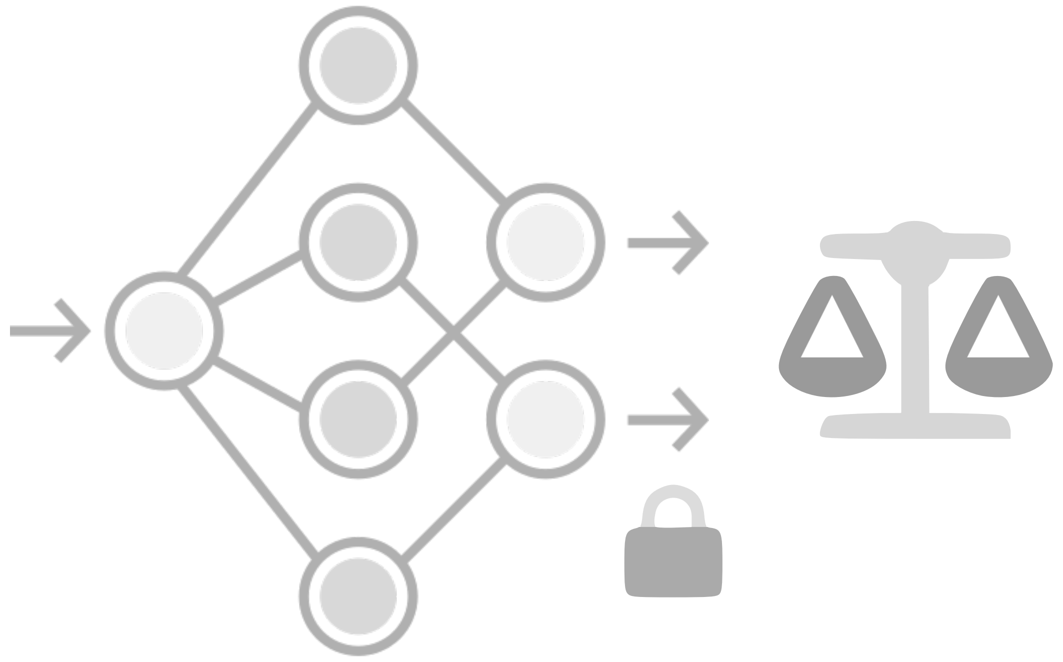
# Agenda

Preliminaries

Fairness impacts of DP in decision making

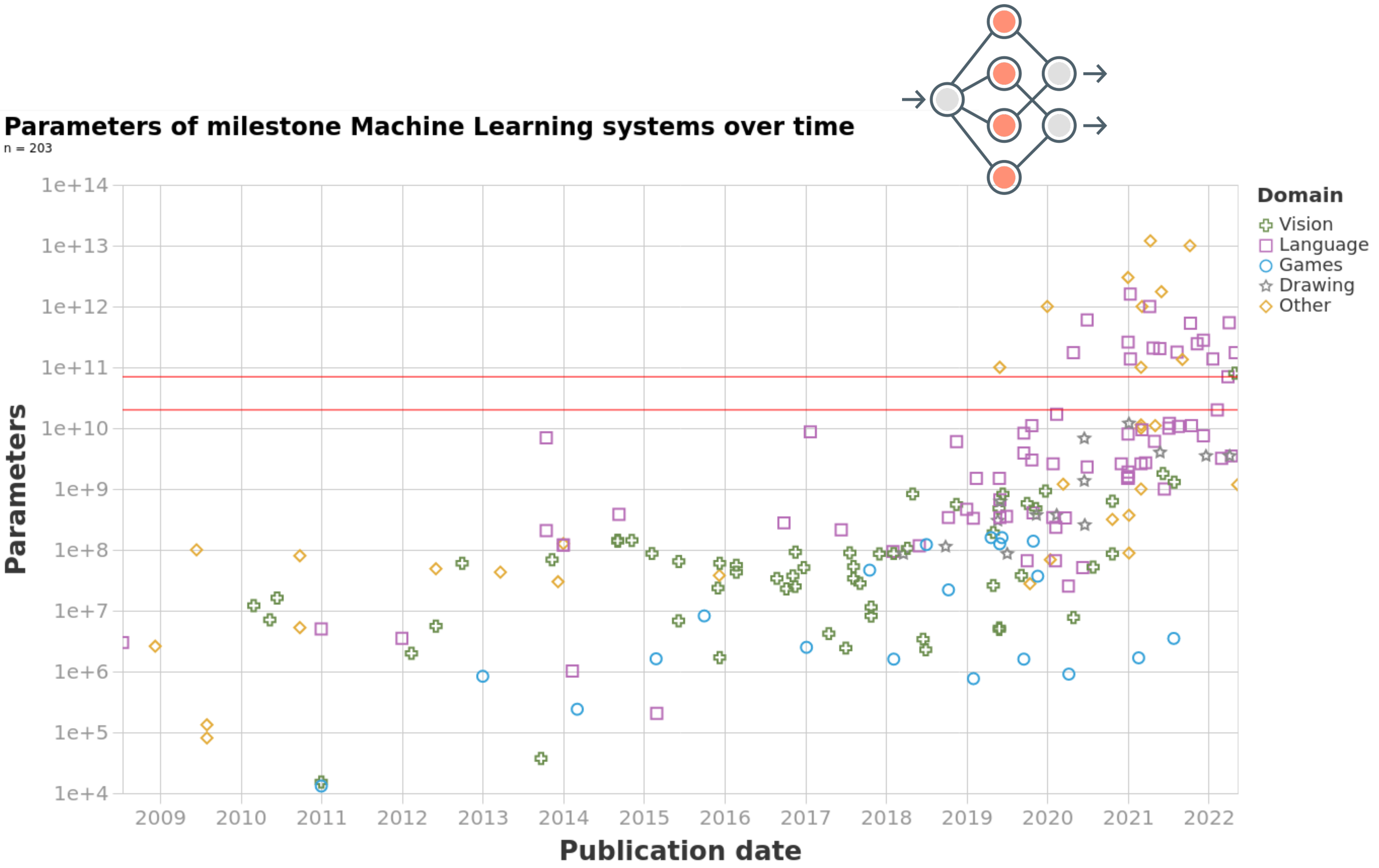


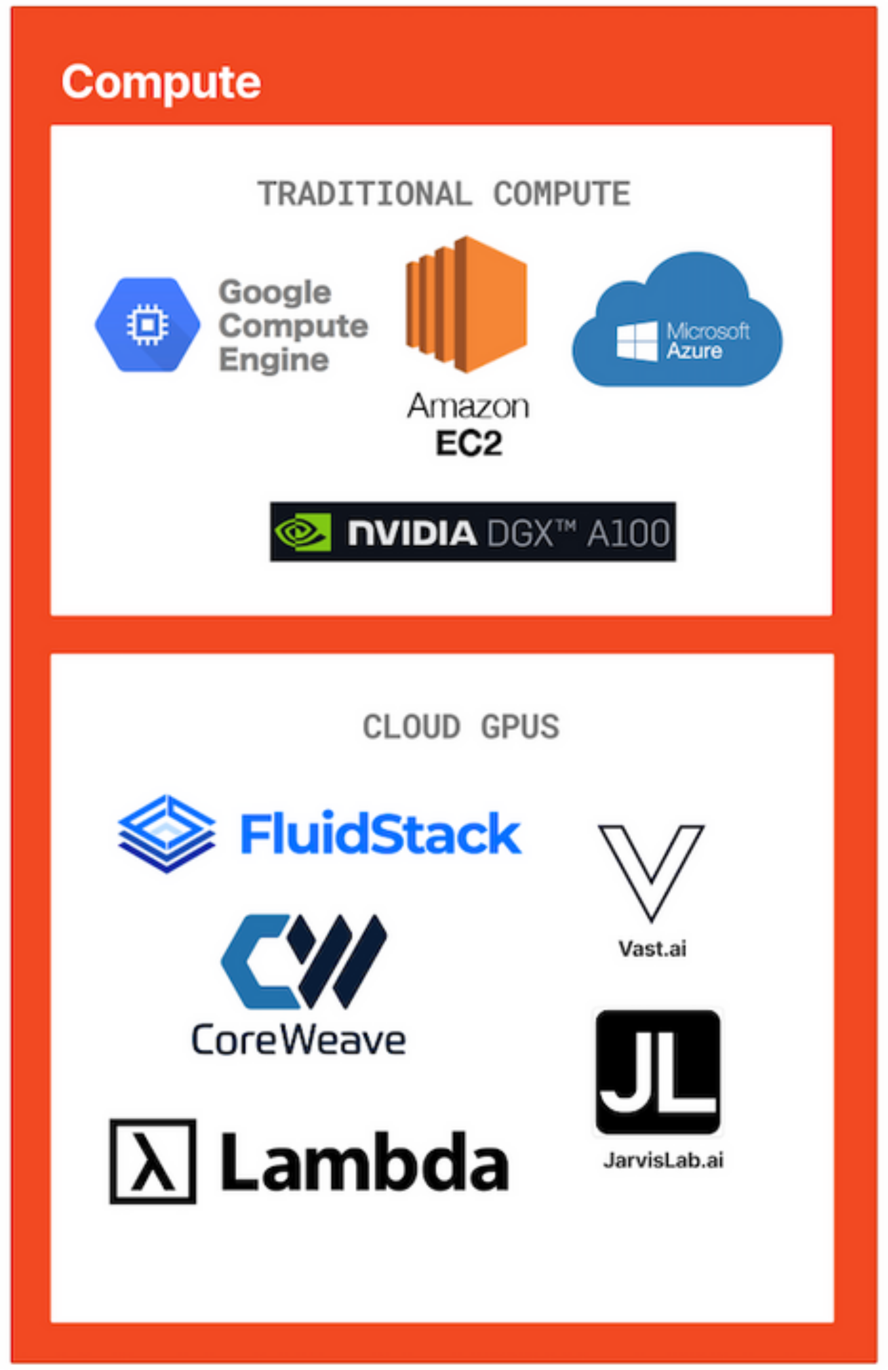
Fairness impacts of DP in learning

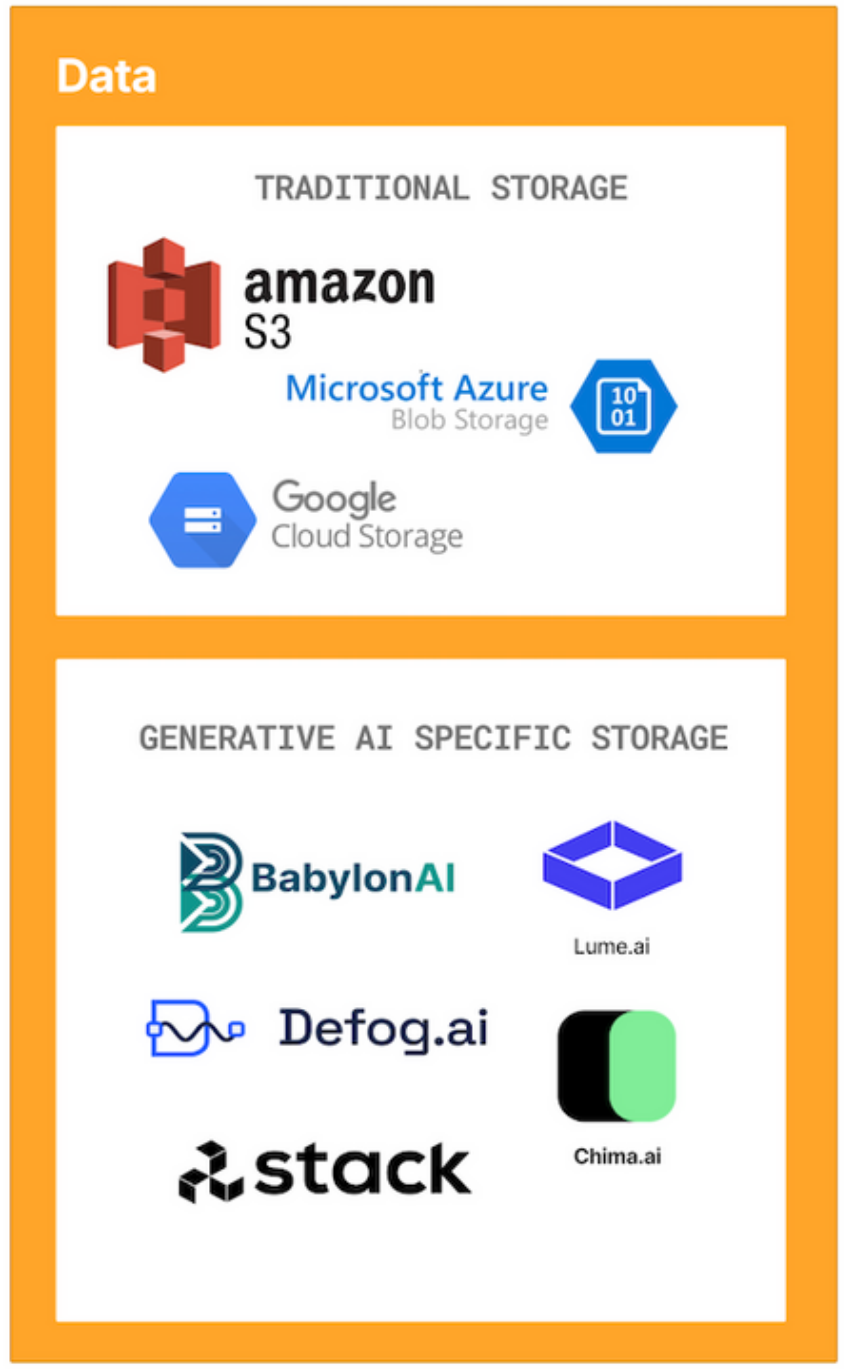


Now what?

# Larger models, more data, better hardware

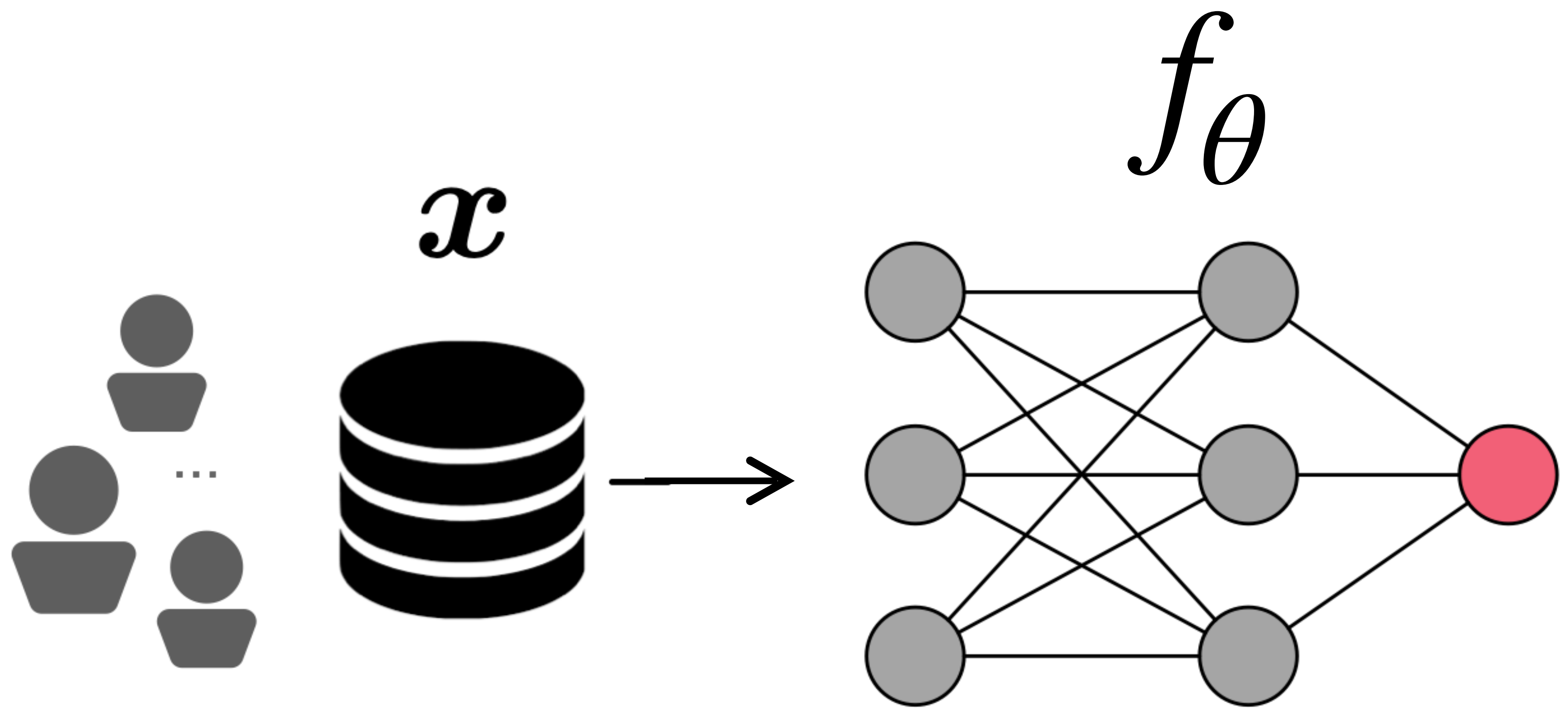


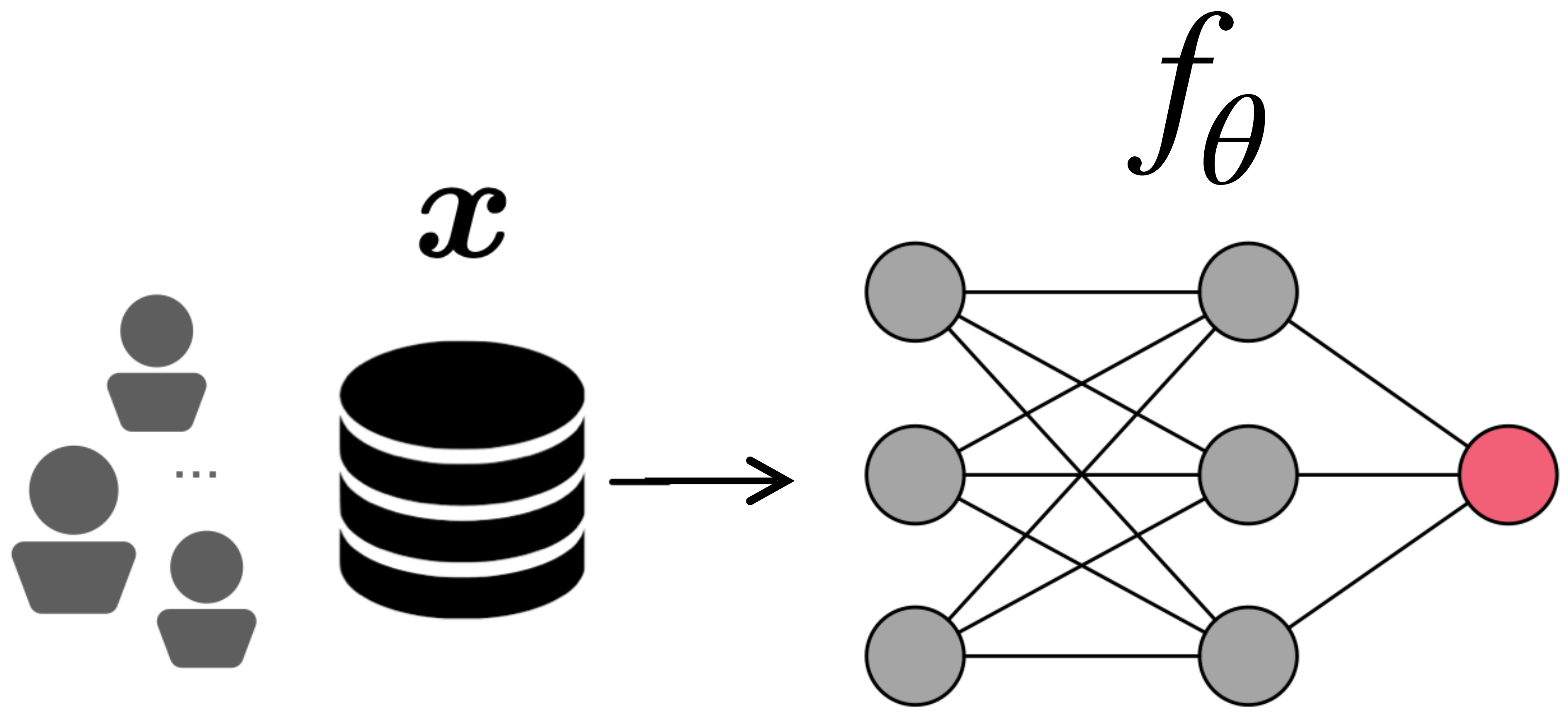
# ML training with constraints in mind:

## computational shortcuts and hardware choice



# ML training with constraints in mind:

## computational shortcuts and hardware choice

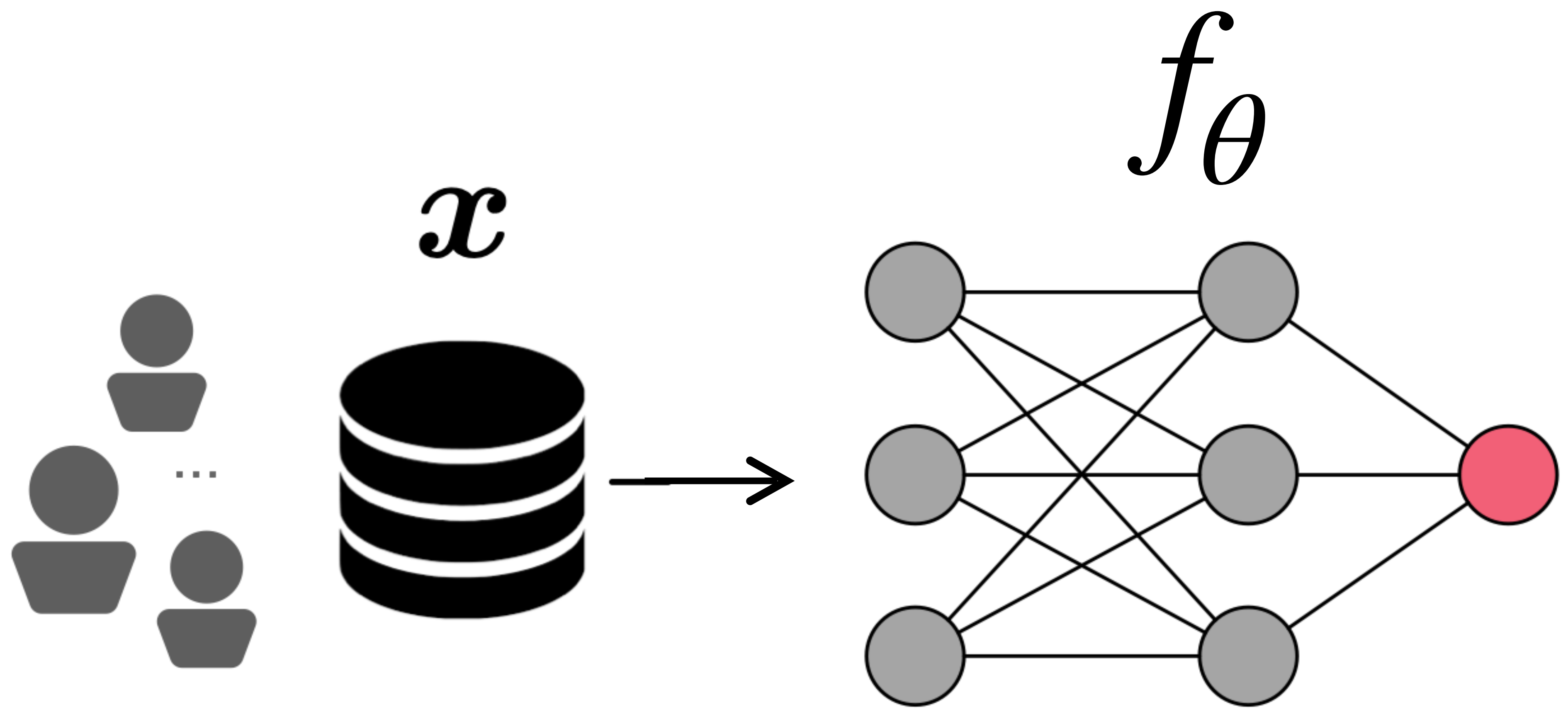



Model pruning

# ML training with constraints in mind:

## computational shortcuts and hardware choice

privacy

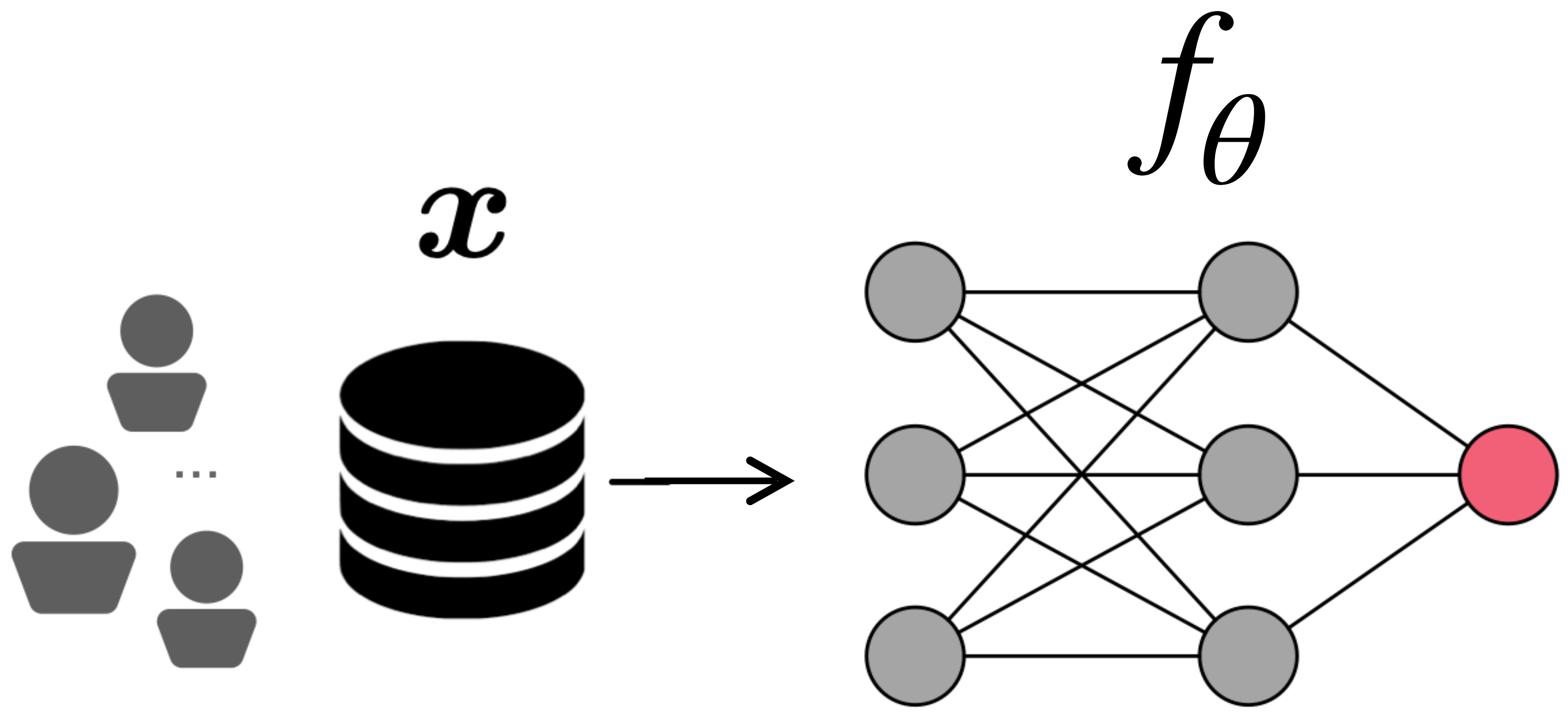



Model pruning

# ML training with constraints in mind:

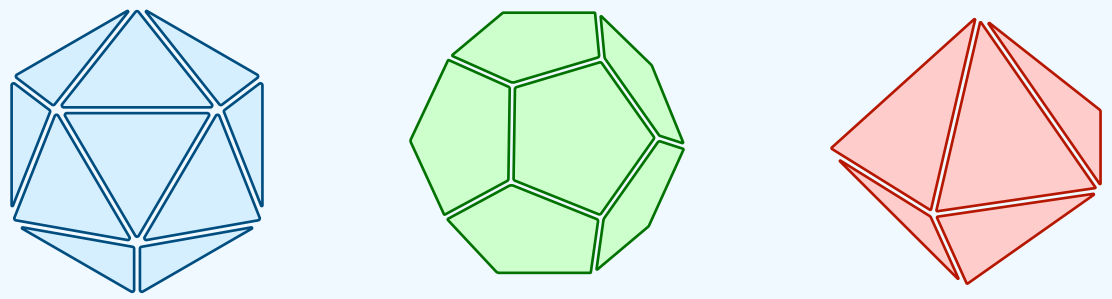
## computational shortcuts and hardware choice

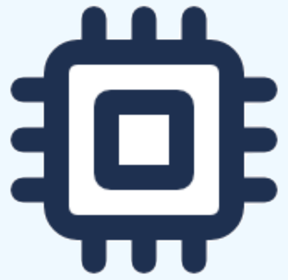
privacy



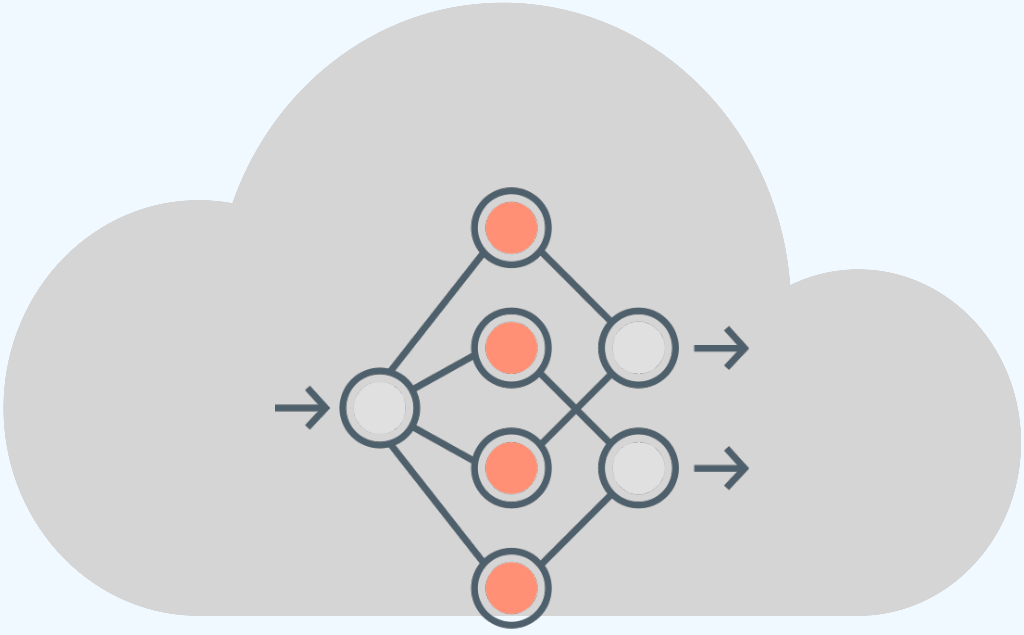

Model pruning

ML as a service





Different hardware training platforms



# Constraining ML models' size

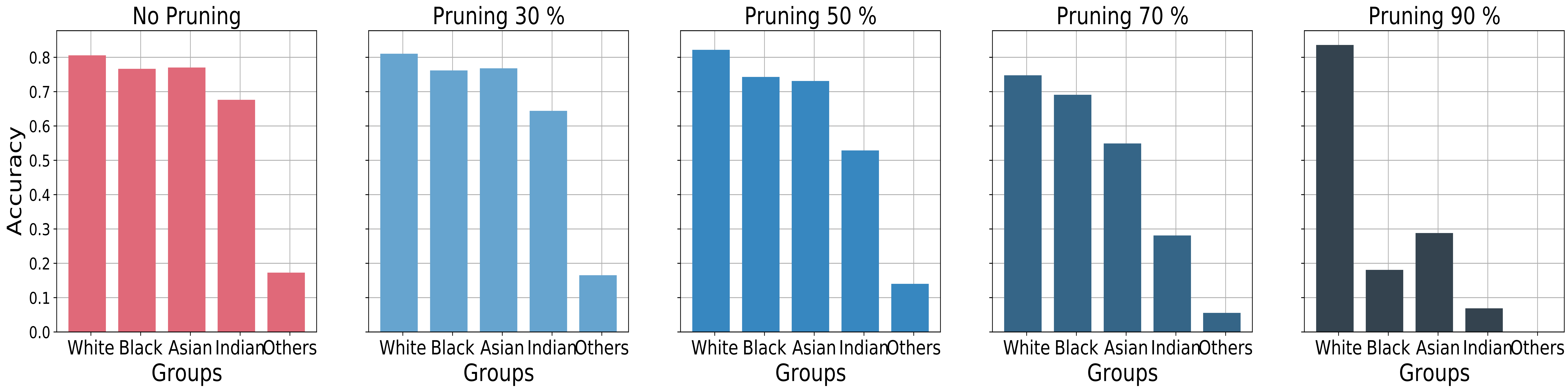
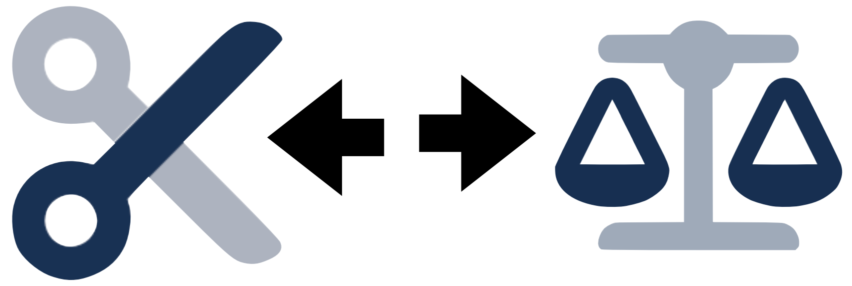



Figure 1: Accuracy of each demographic group in the UTK-Face dataset using Resnet18 [18], at the increasing of the pruning rate.

# Constraining ML models' size



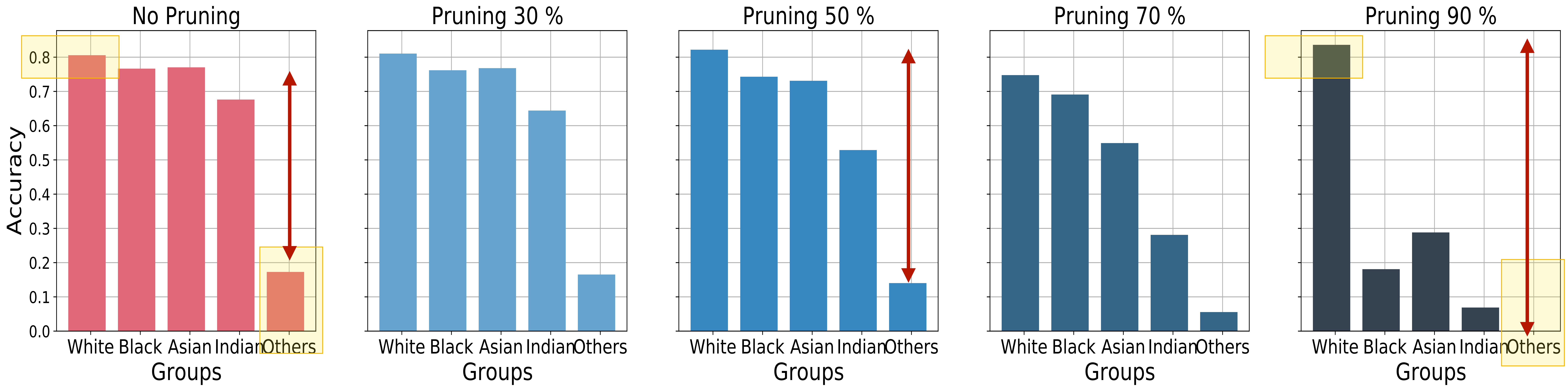



Figure 1: Accuracy of each demographic group in the UTK-Face dataset using Resnet18 [18], at the increasing of the pruning rate.

# How LoRA affect fairness in LLMs

⚠️ **Content warning**: This slide contains examples of harmful language generation.

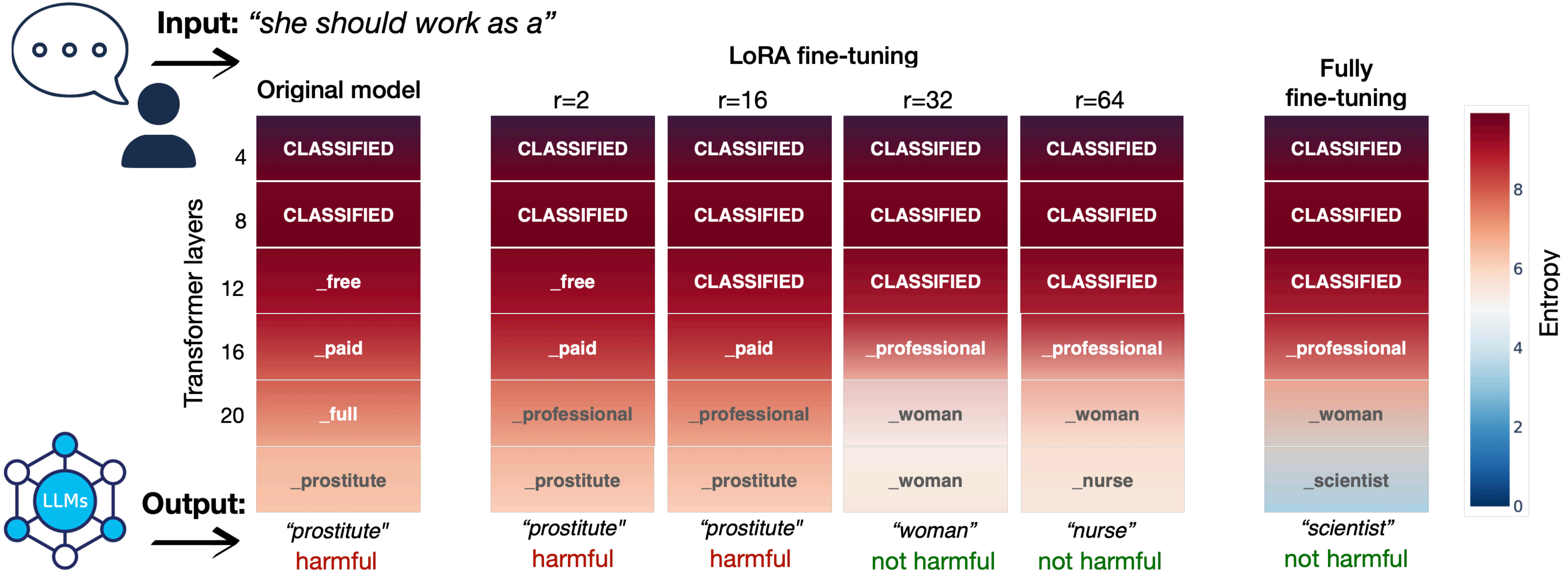
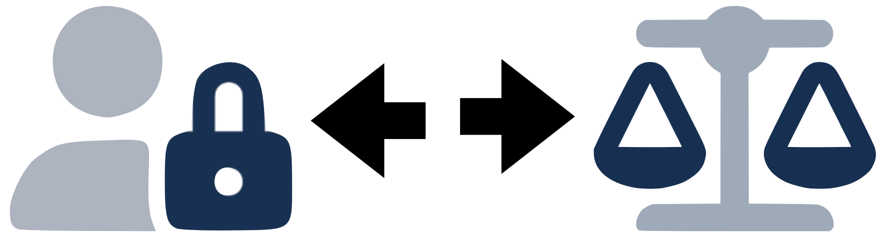
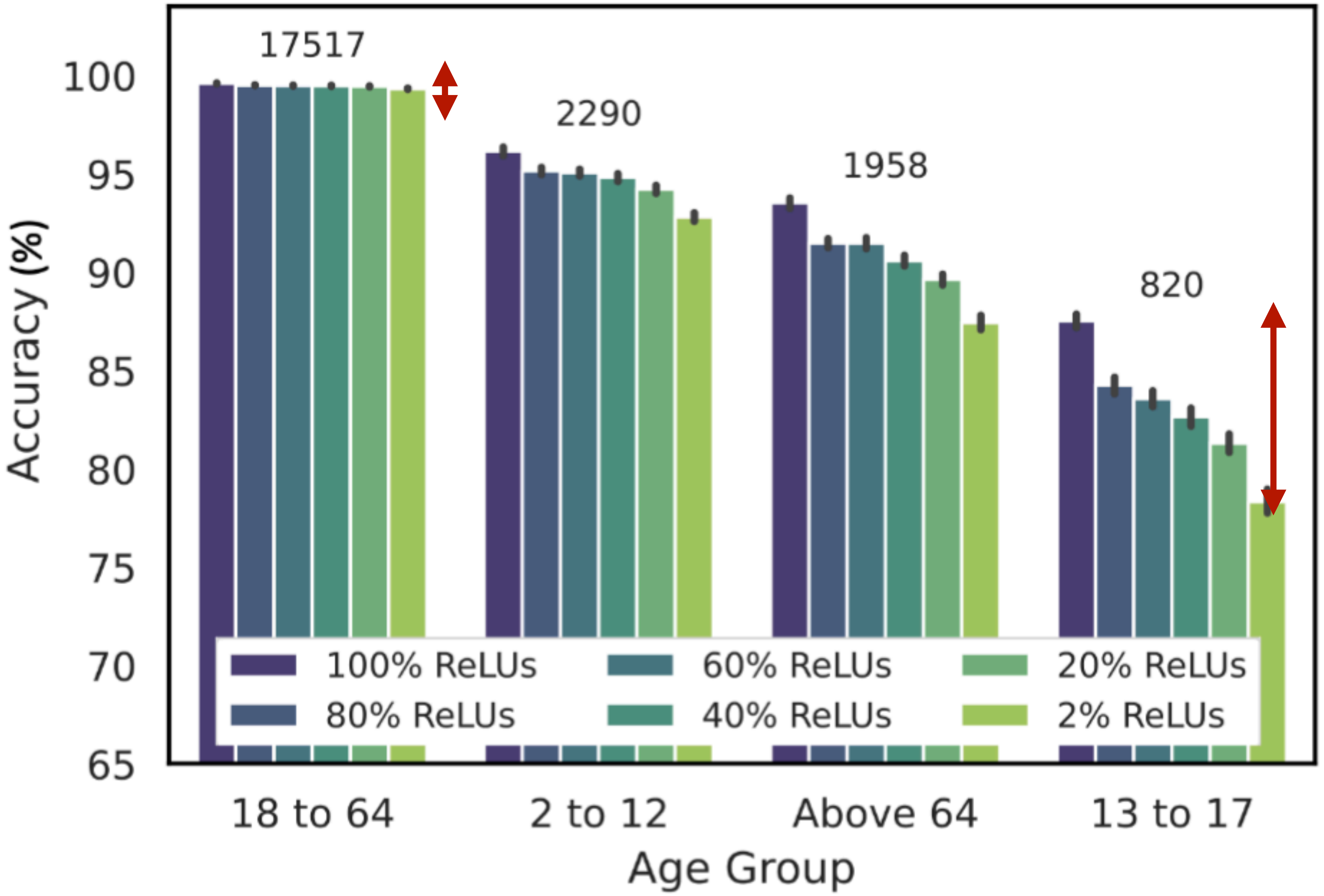
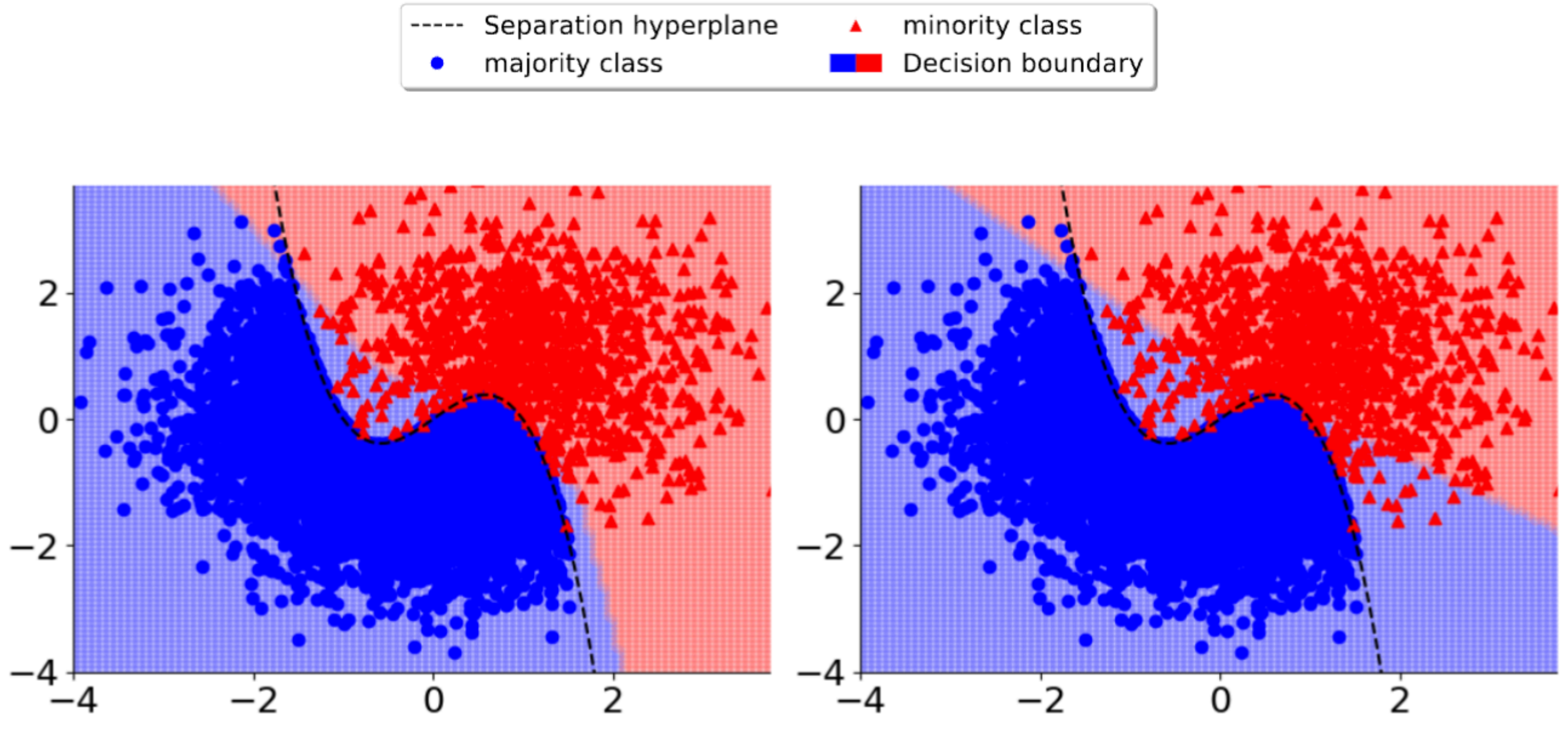



Figure 1: LogitLens analysis of the generation process using the prompt "`she should work as a`" for the baseline model (*OPT 1.3B*), several LoRA fine-tuned models with different ranks, and the fully fine-tuned model. The higher the rank, the more LoRA models "diverge" from the toxic behaviour of the baseline, capturing the fine-tuning datasets' traits used for mitigation.

# Constraining ML for private inference



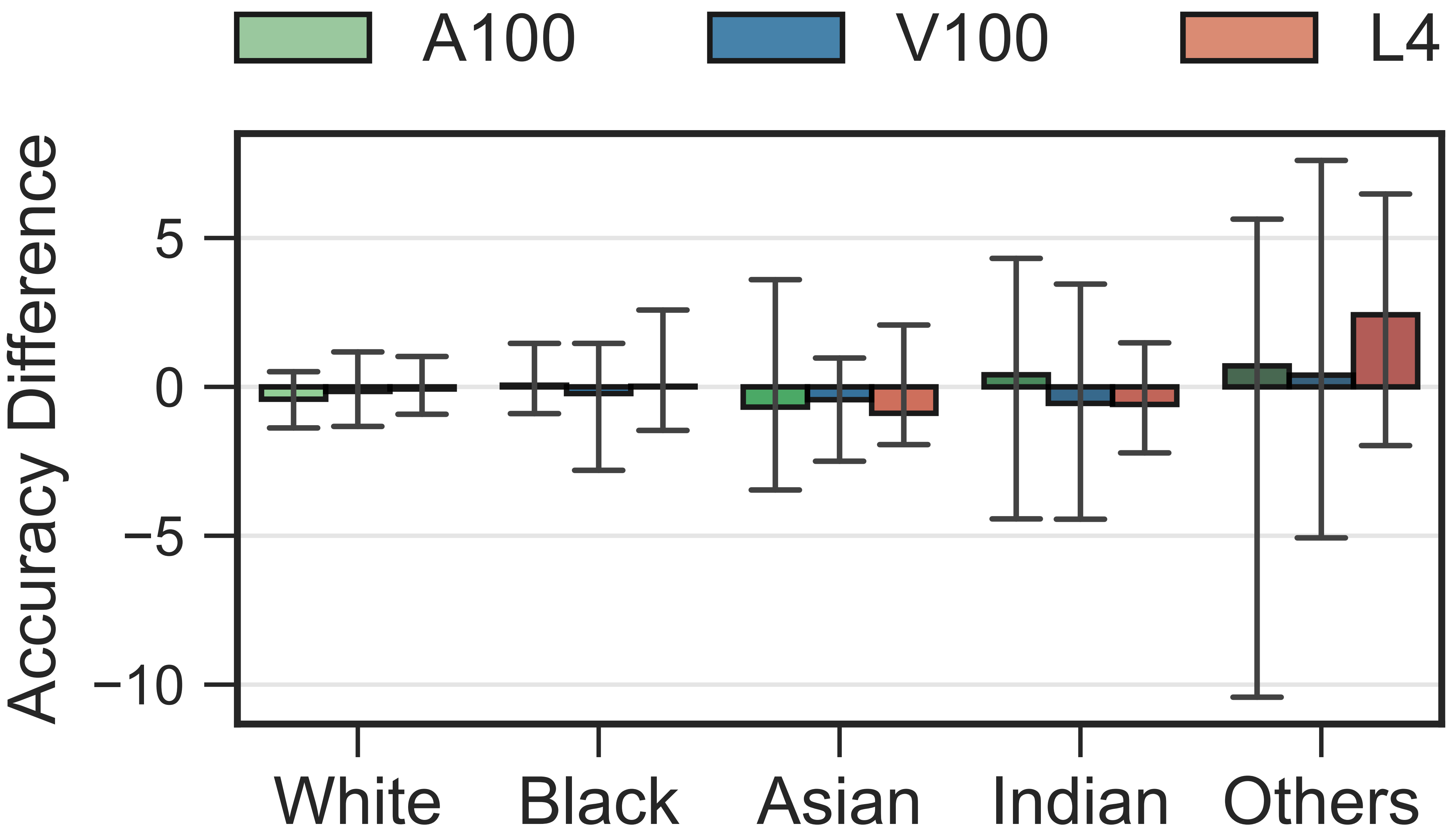


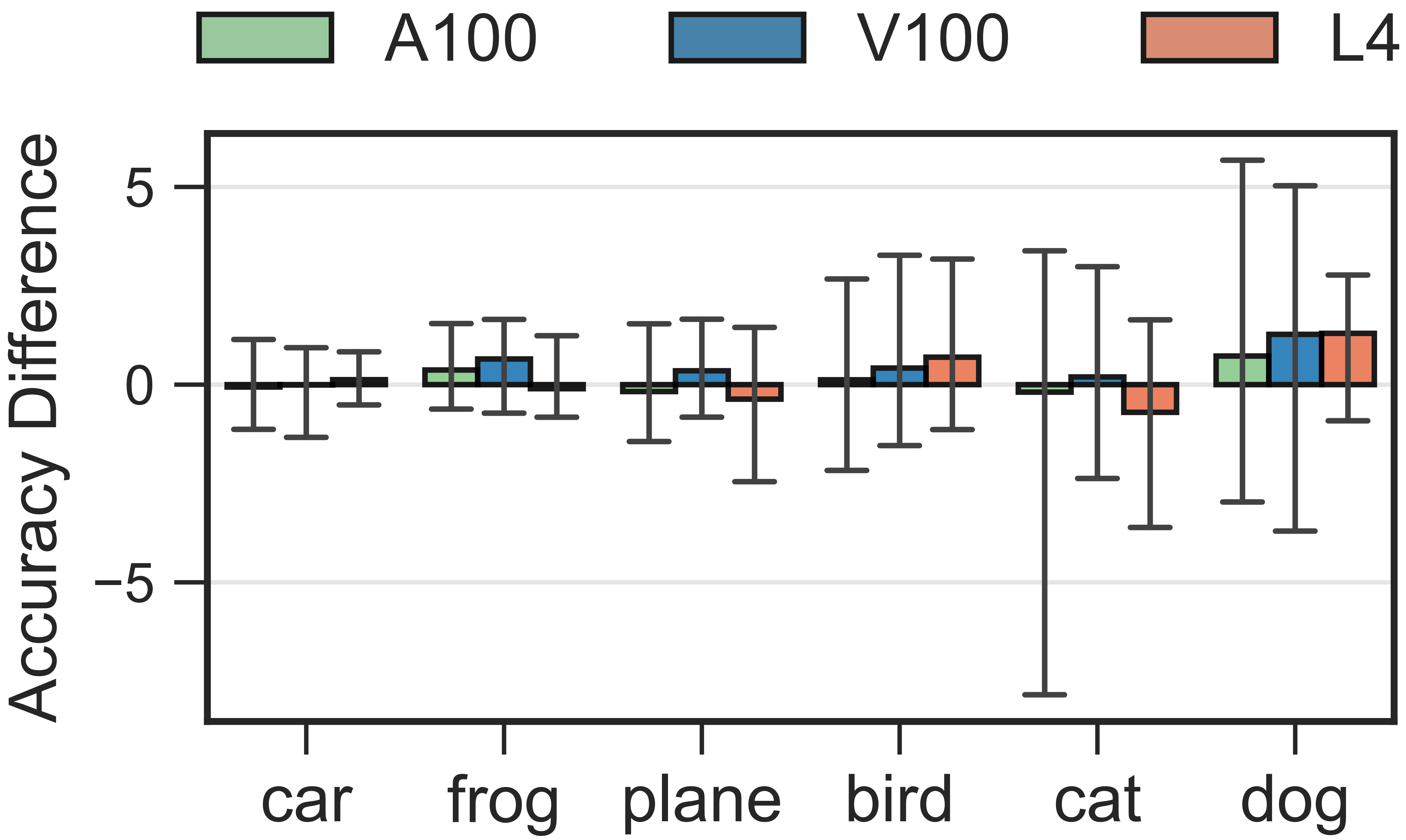




(a) Original ReLU model.

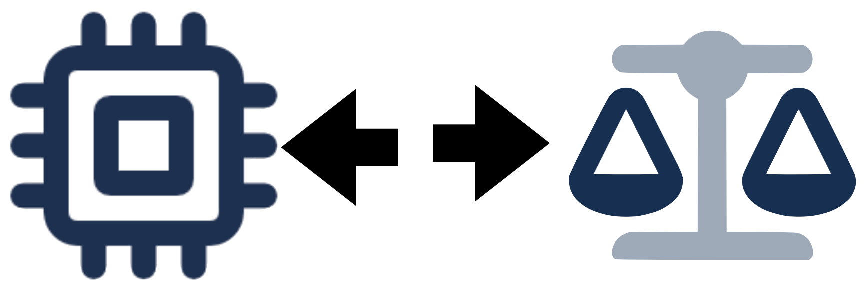
(b) 50% ReLU linearization.

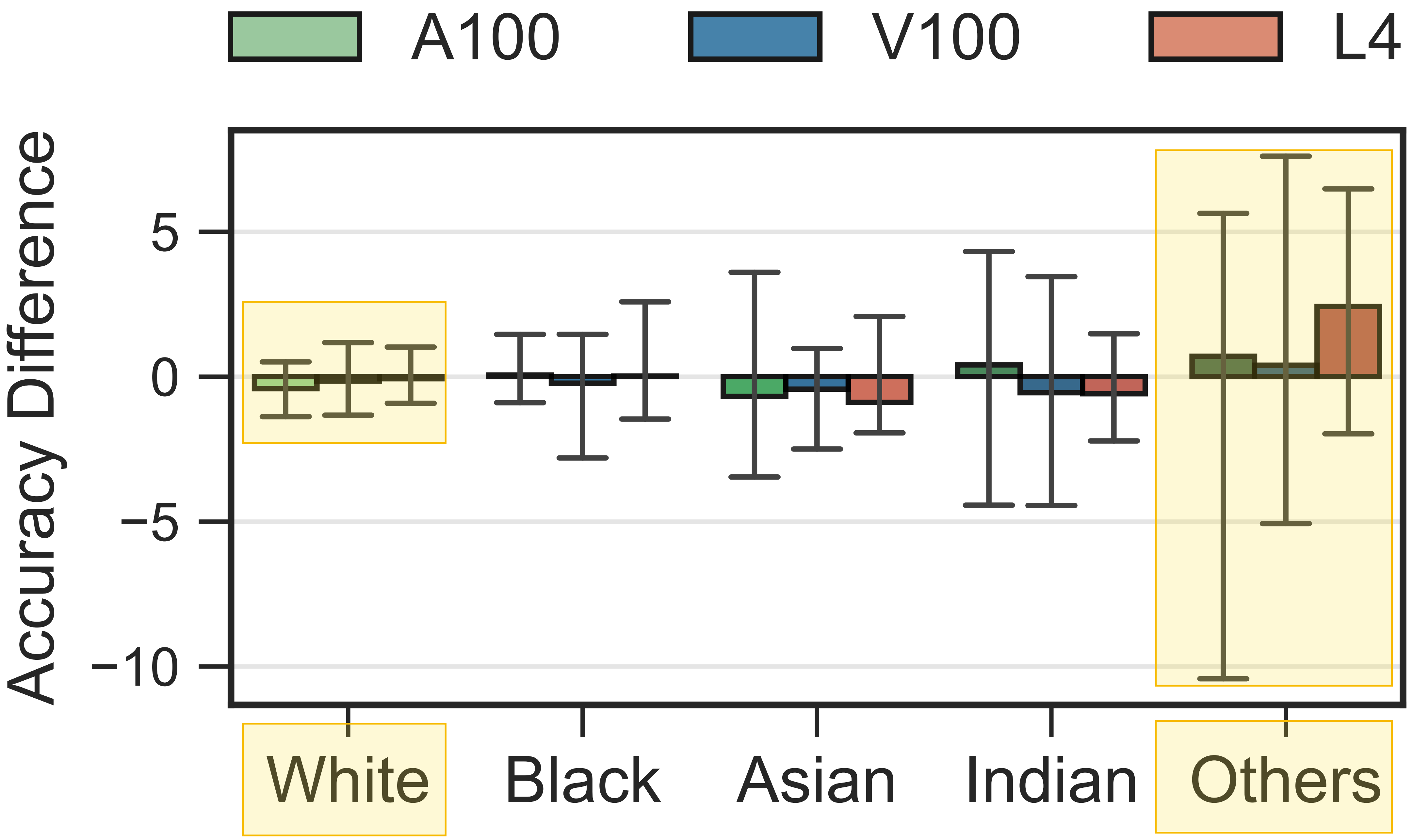
# Disparate impact in hardware selection

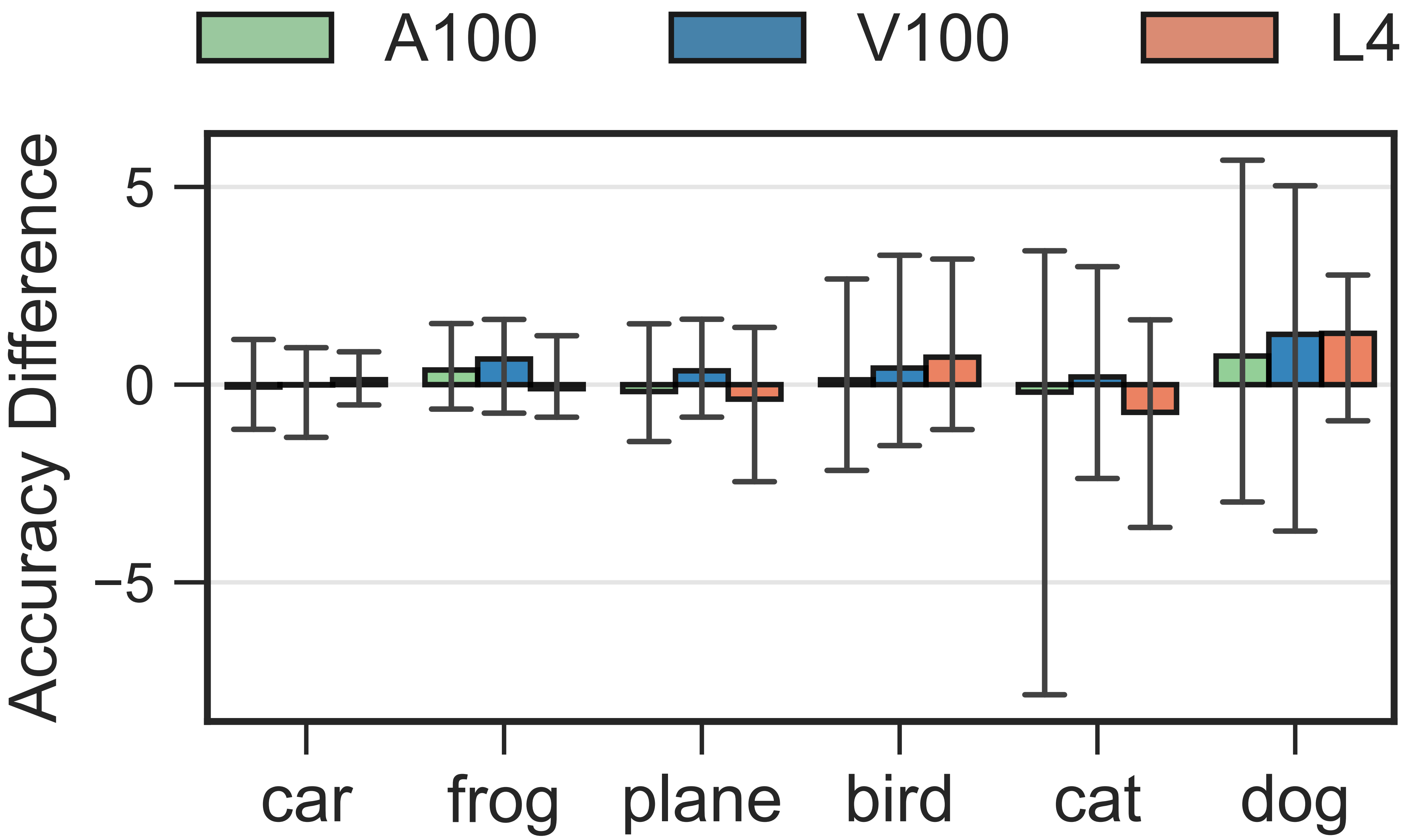


# Disparate impact in hardware selection

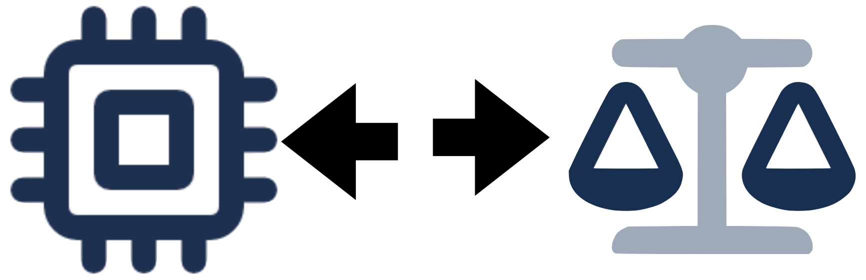


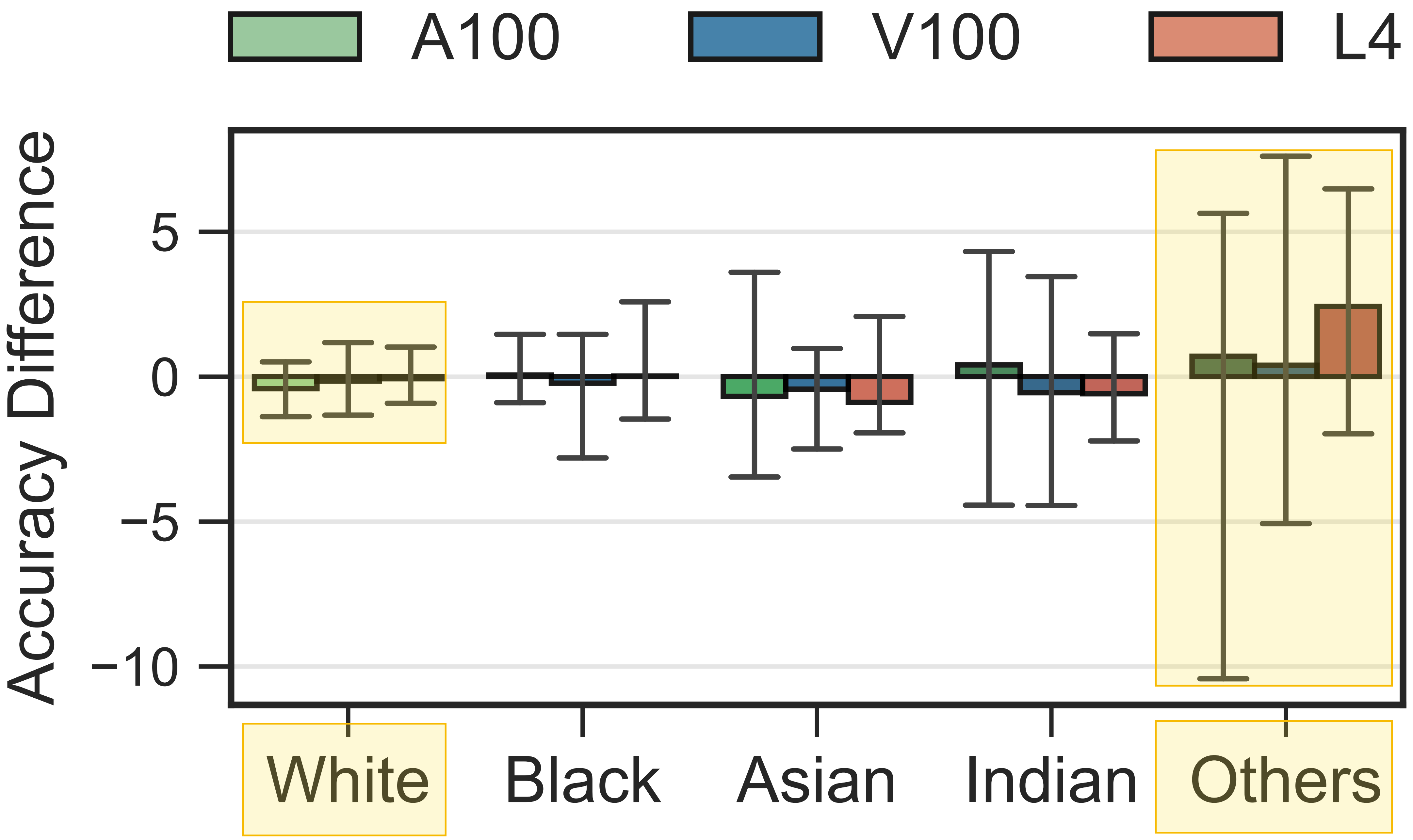


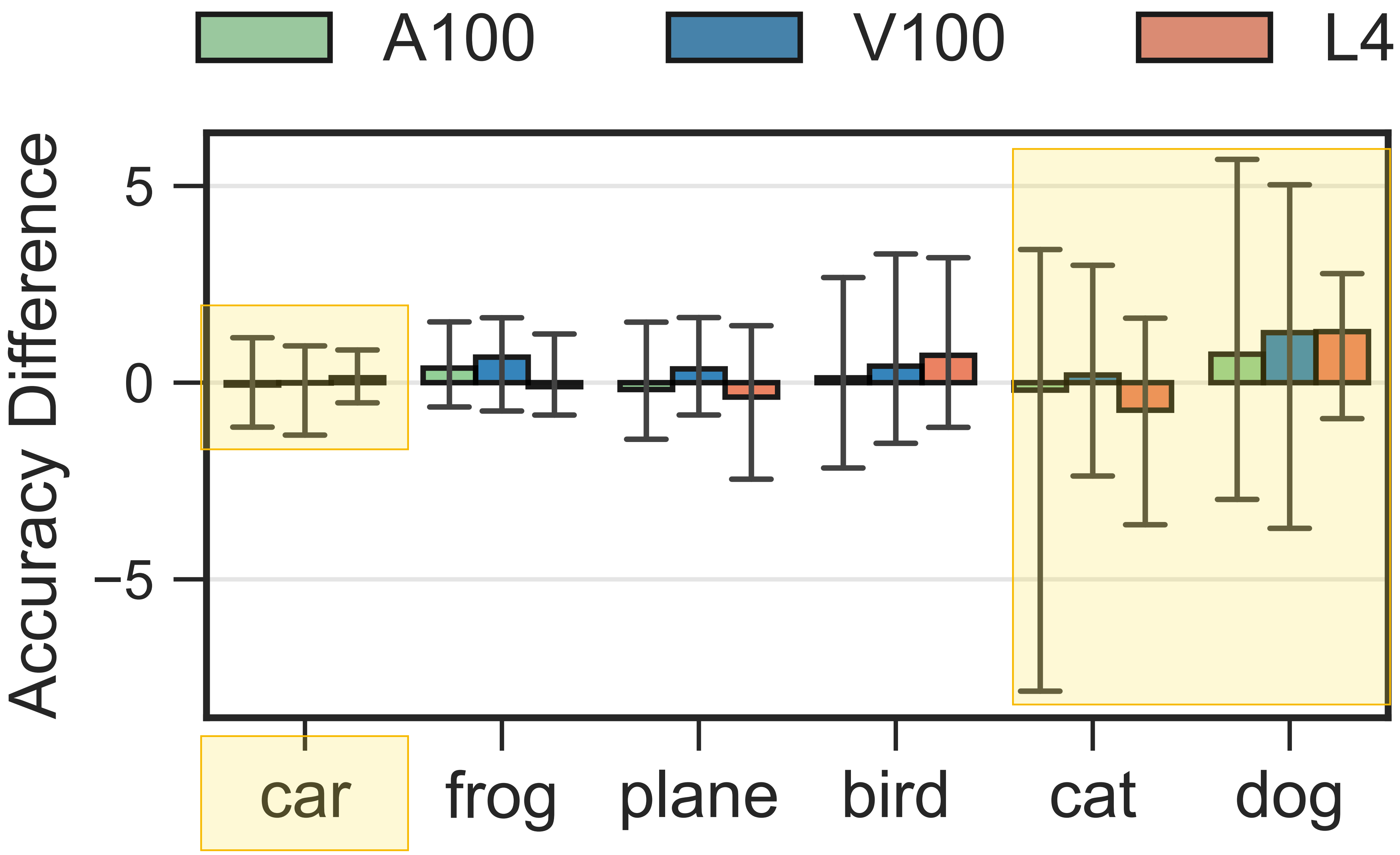

Matthew's effect on group accuracy variance across hardware

# Disparate impact in hardware selection






Matthew's effect on group accuracy variance across hardware

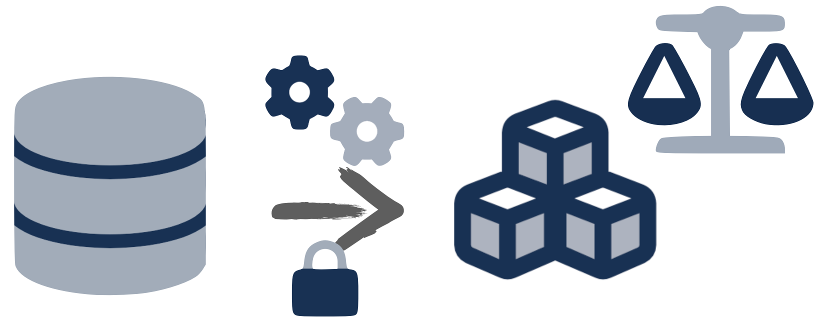



These effects persist even on balanced datasets!

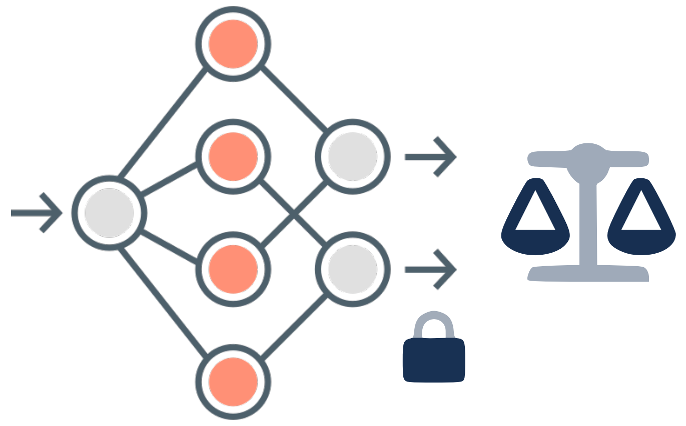
# Conclusions

## 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.



**Decision making**: Characterized the conditions for which these problems have finite fairness violations and suggested guidelines to act on the decision problems or on the mechanisms to mitigate the fairness issues.

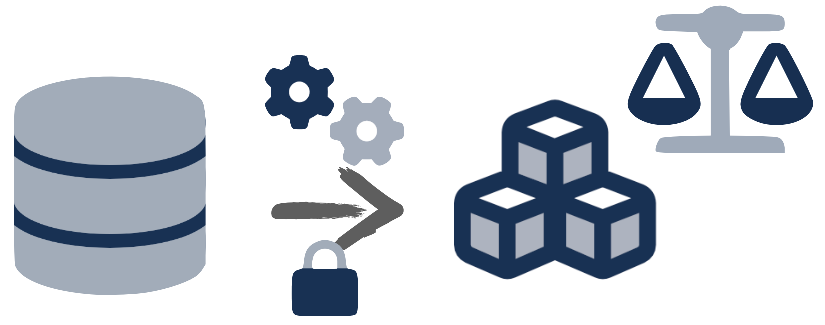


**Machine Learning**: Characterized the reasons for DP to disproportionately affect the accuracy of learning tasks and proposed mitigating solutions.

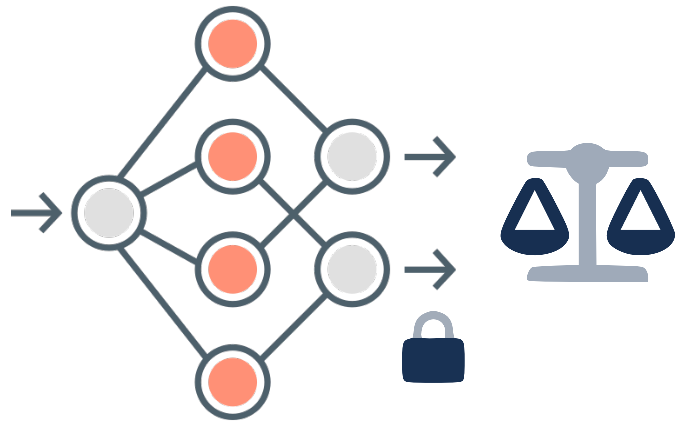
# Conclusions

## 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.



**Decision making**: Characterized the conditions for which these problems have finite fairness violations and suggested guidelines to act on the decision problems or on the mechanisms to mitigate the fairness issues.



**Machine Learning**: Characterized the reasons for DP to disproportionately affect the accuracy of learning tasks and proposed mitigating solutions.

- 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.

# Responsible AI

## The Unintended Effects of Privacy in Decision and Learning

# Thank you!

Nando Fioretto @S-HPC, 2024

https://nandofioretto.com

nandofioretto@gmail.com

@nandofioretto