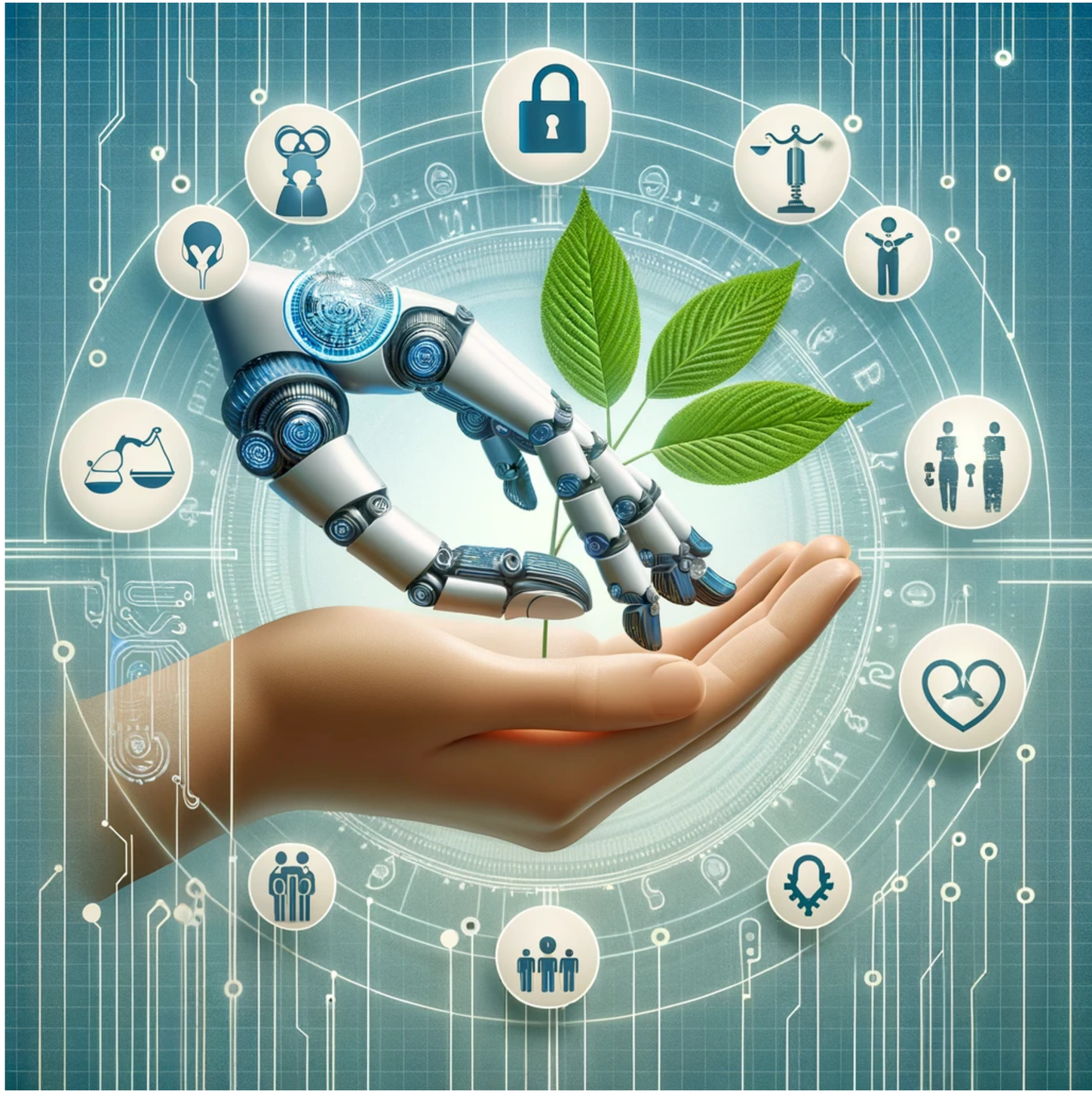
Responsible AI:

Seminar on Fairness, Safety, Privacy and more



https://nandofioretto.com

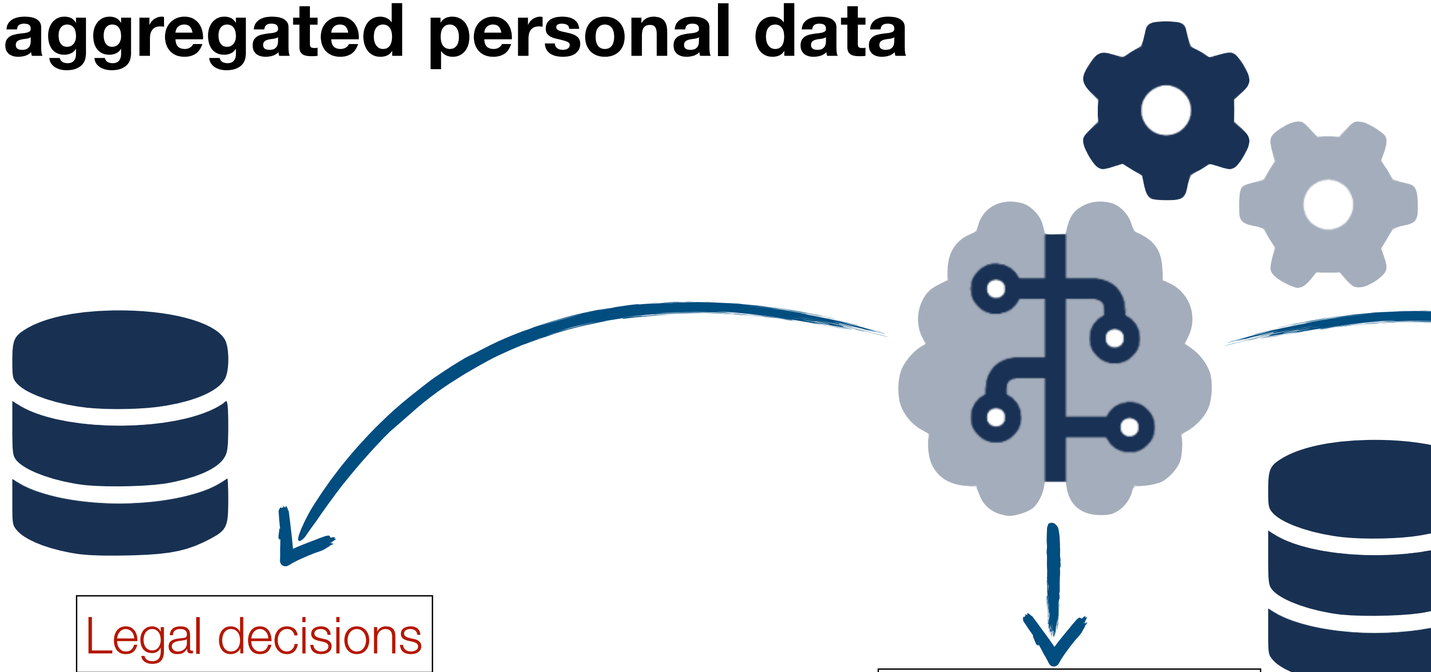
nandofioretto@gmail.com

@nandofioretto

Ferdinando Fioretto @UVA Spring 2024

AI-driven decision making

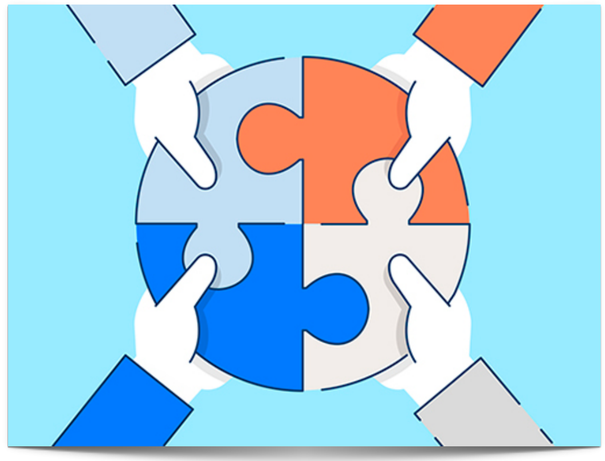
with aggregated personal data



Legal decisions

Loan predictions

Resource allocations



Key concerns

President Biden Signs Executive Order Advancing Racial Equity and Imposing Equity Principles on Government A.I.

February 16, 2023

Privacy

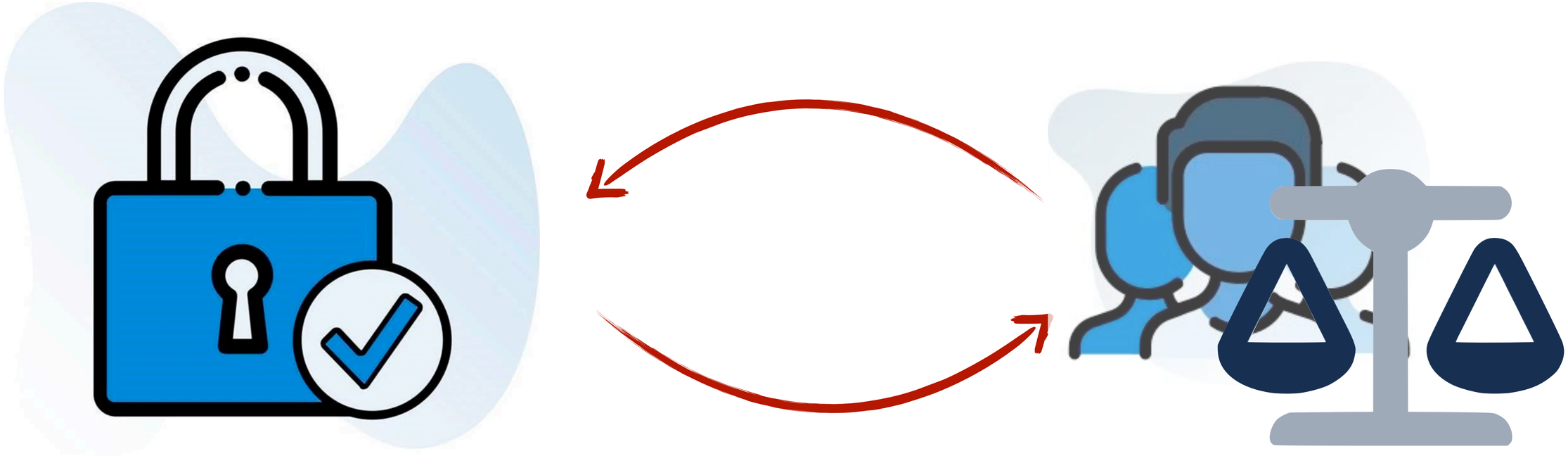
Fairness

Key concerns

Privacy

- Differential Privacy has become the paradigm of choice for protecting data privacy.
- Deployments are growing at a fast rate.

Key concerns



Privacy **Fairness**

Disproportionate impacts in decision making

Title 1 allotment

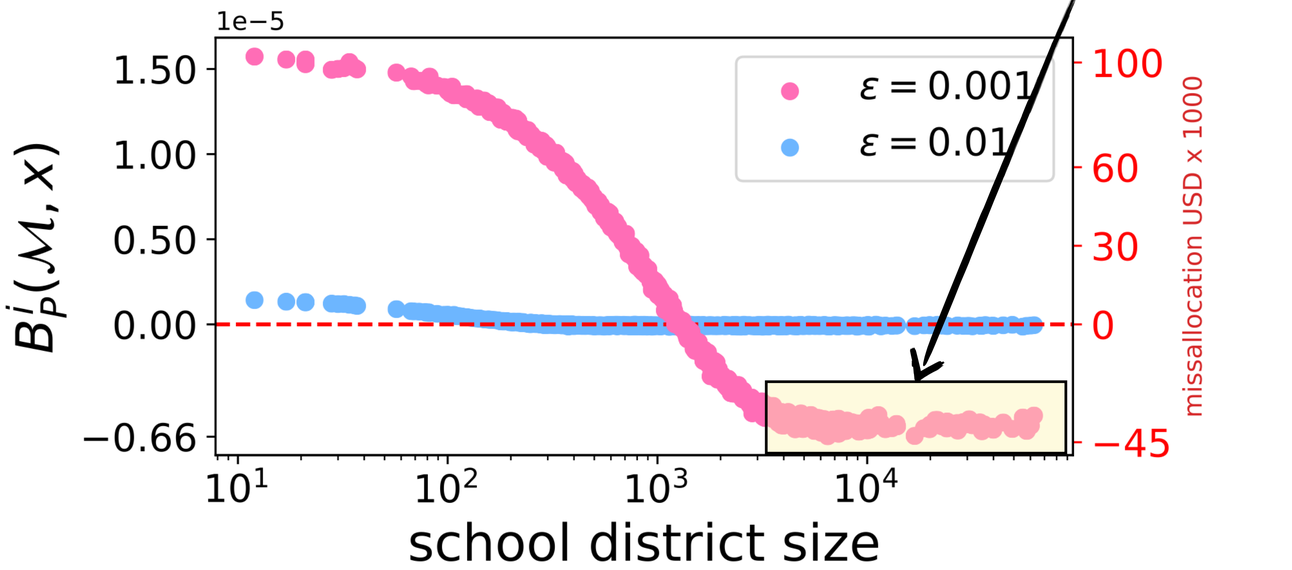
- Title 1 of the Elementary and Secondary Education Act is one of the largest U.S. program offering educational assistance to disadvantaged children.
- In the fiscal year 2021 alone, it distributed about $11.7 billion through several types of grants.
- **Allotment**:

count of children 5 to 17 in district i

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

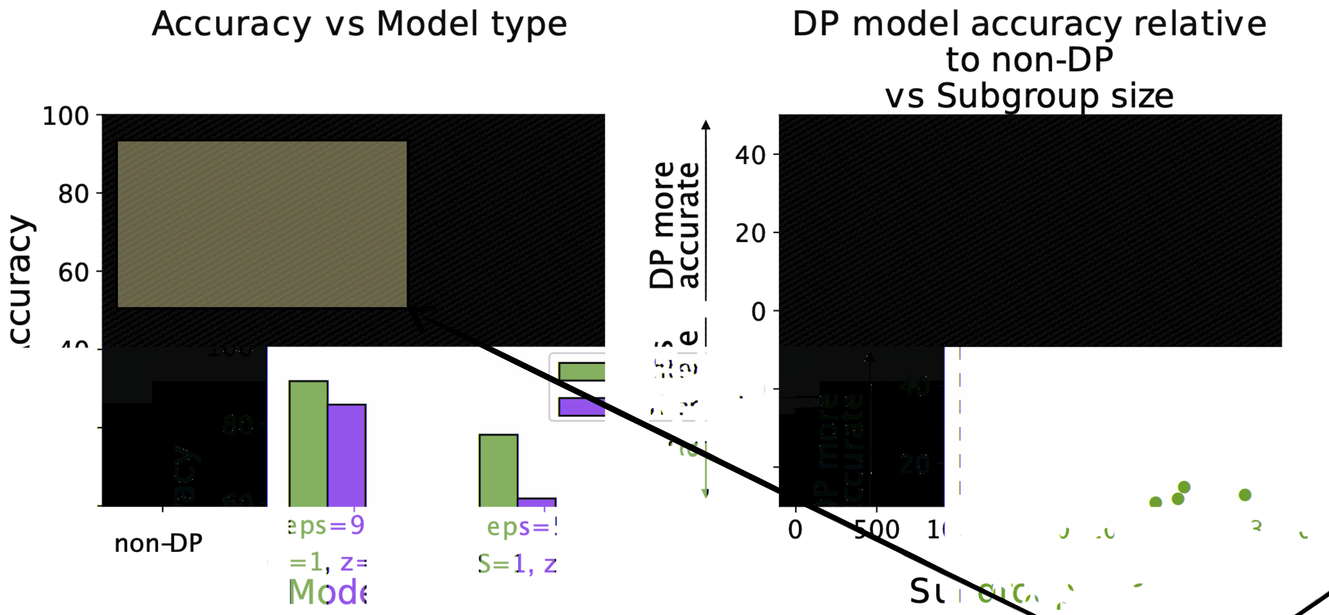
student expenditures in district i

Districts receiving up to 42K less than warranted

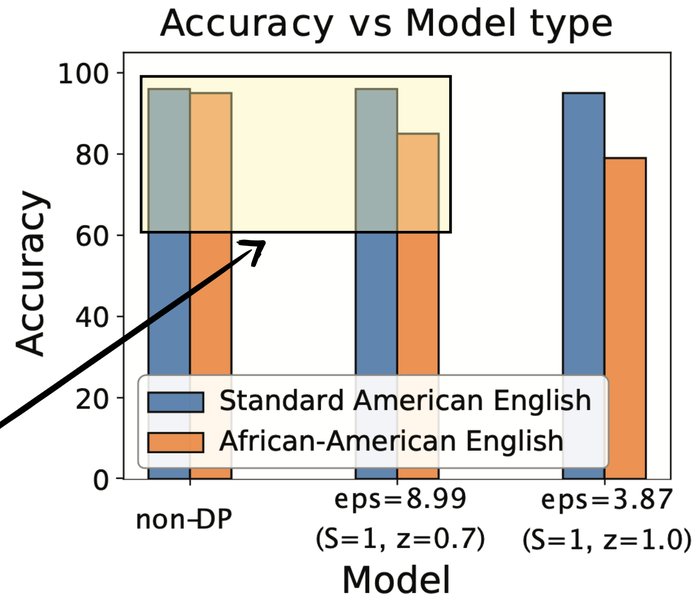


Disproportionate impacts in learning tasks

Gender and age classification on facial images



Sentiment analysis of tweets



As the privacy increases the accuracy disparity of the learning task increases

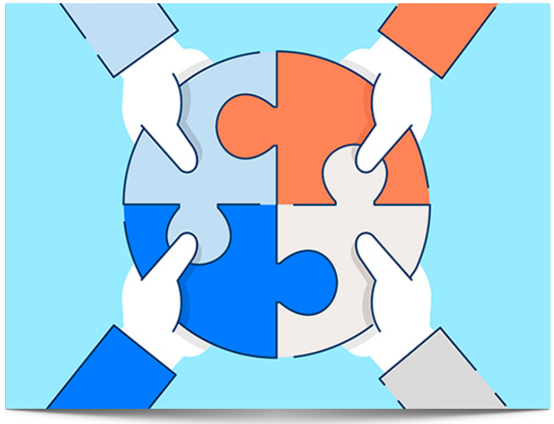
Bagdasaryan et al. 2019

Societal impact

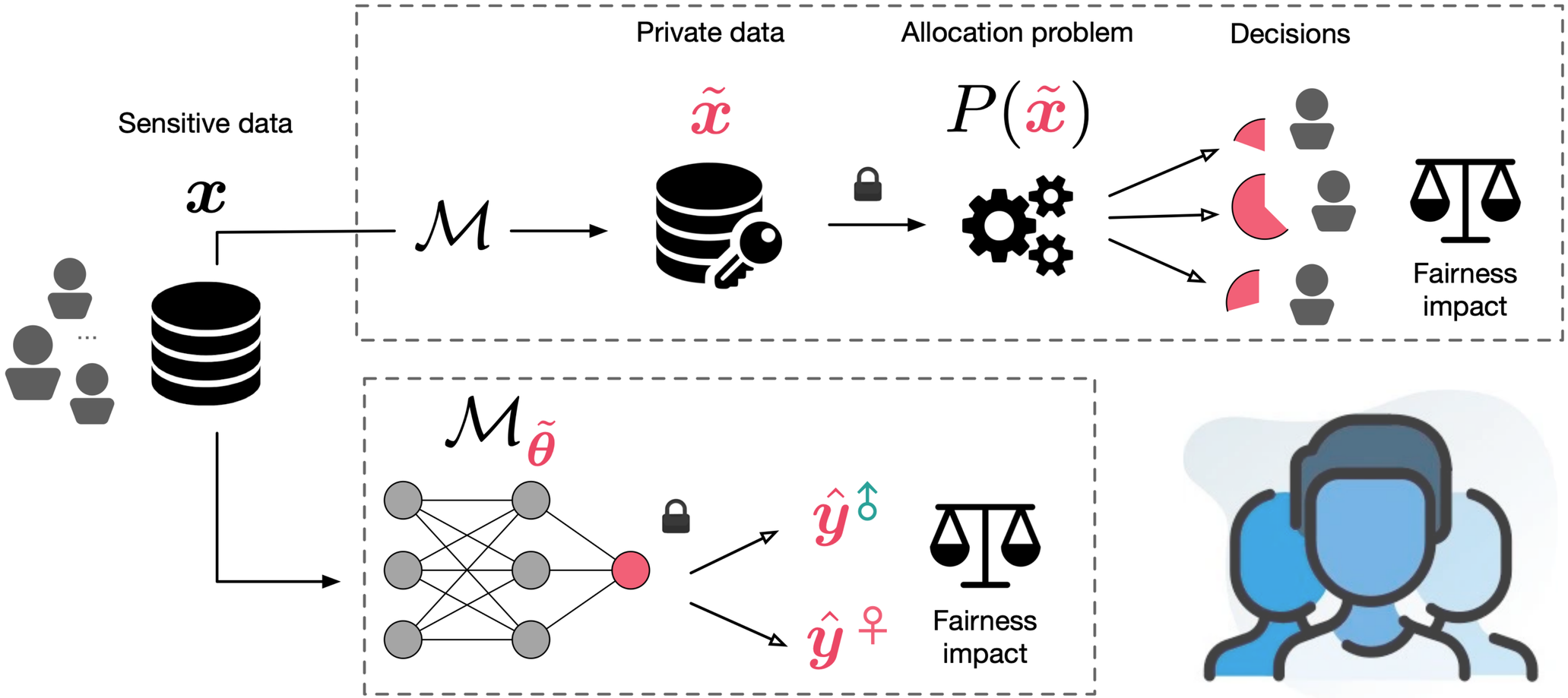
The resulting outcomes can have significant societal and economic impacts on the involved individuals:

- Classification errors may penalize some groups over others in important determinations including criminal assessment, hiring, or landing.
- Biased decisions can result in disparities regarding the allocation of critical funds, benefits, and therapeutics.

While these observations are becoming more apparent ***their causes are still largely understudied.***



Setting and talk outline

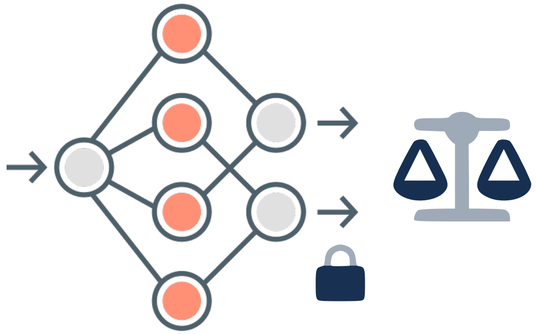


Agenda

Preliminaries

Fairness impacts of DP in decision making

Fairness impacts of DP in learning



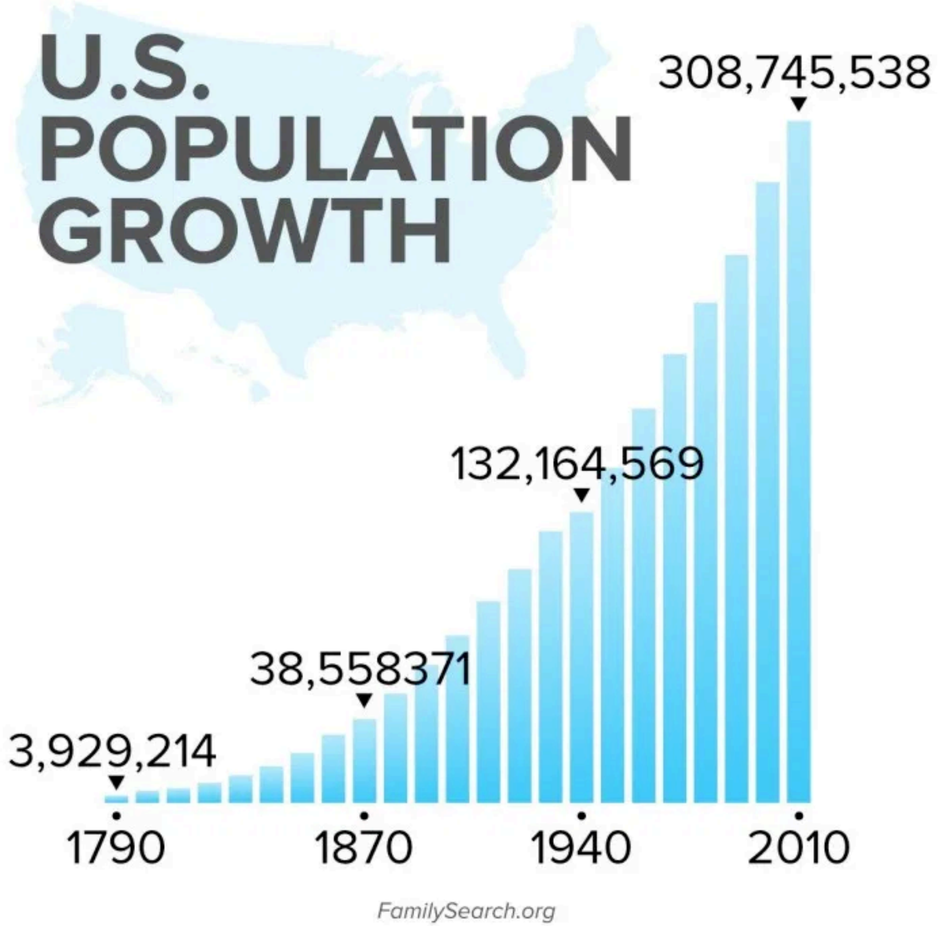
What's next?

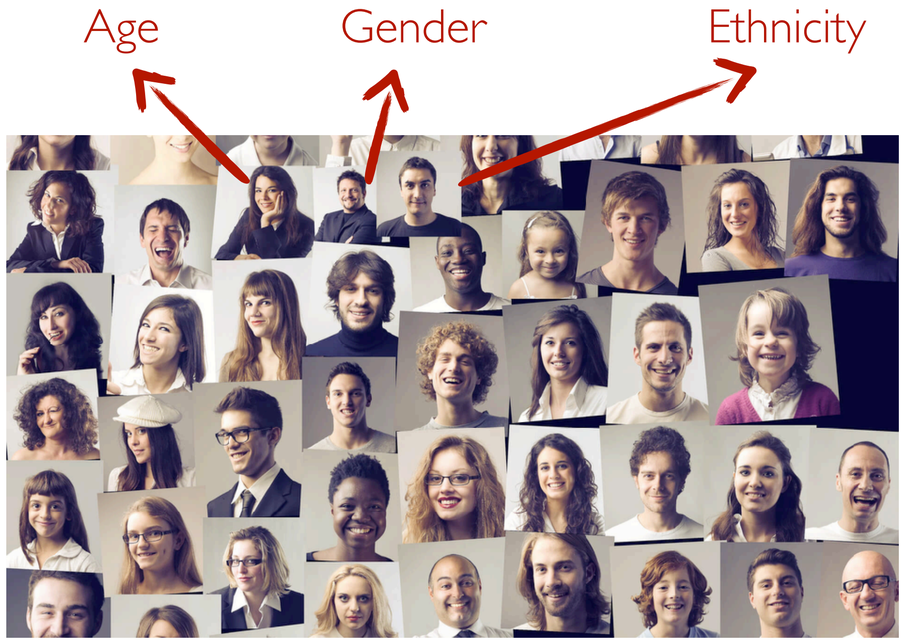
Counting and Stats Publishing

A Census data-release perspective

US Census data collection

Enumeration of the total population living the US





US Census data collection

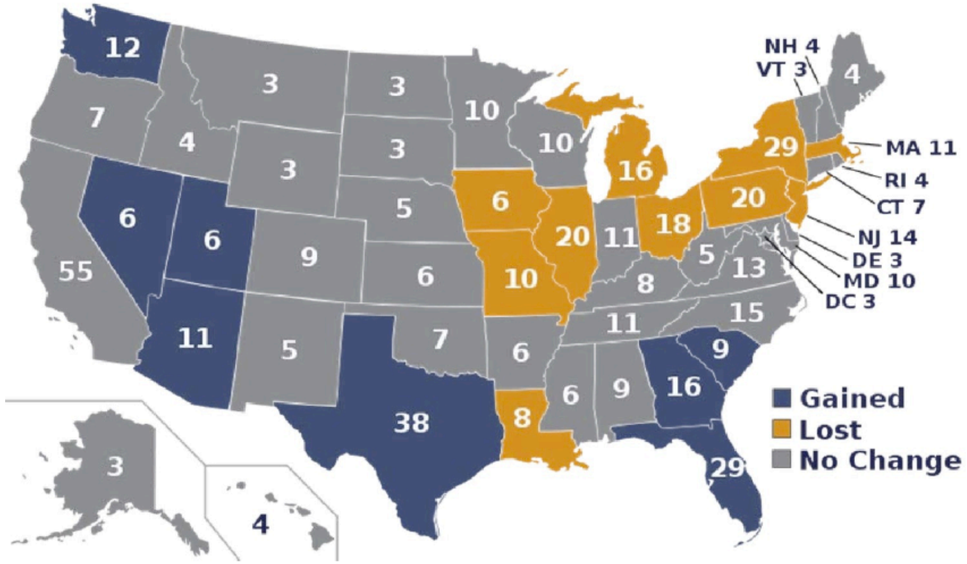
Accurate count is important

- 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

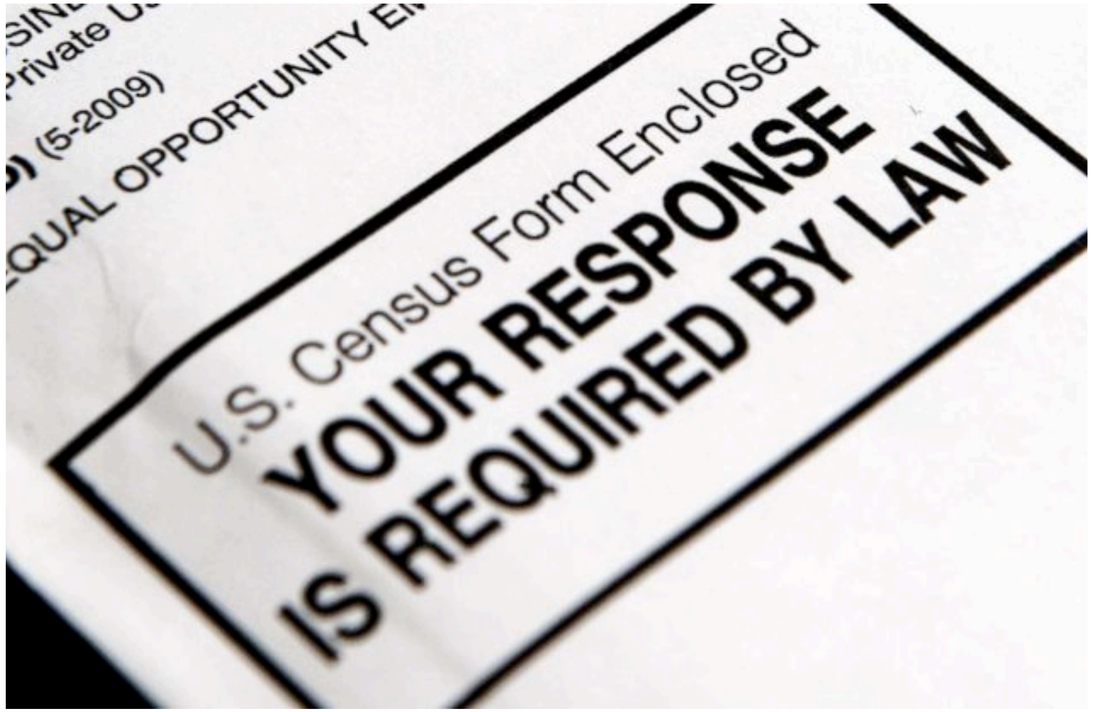


Determine the number of seats that states get in the US House of Representatives.

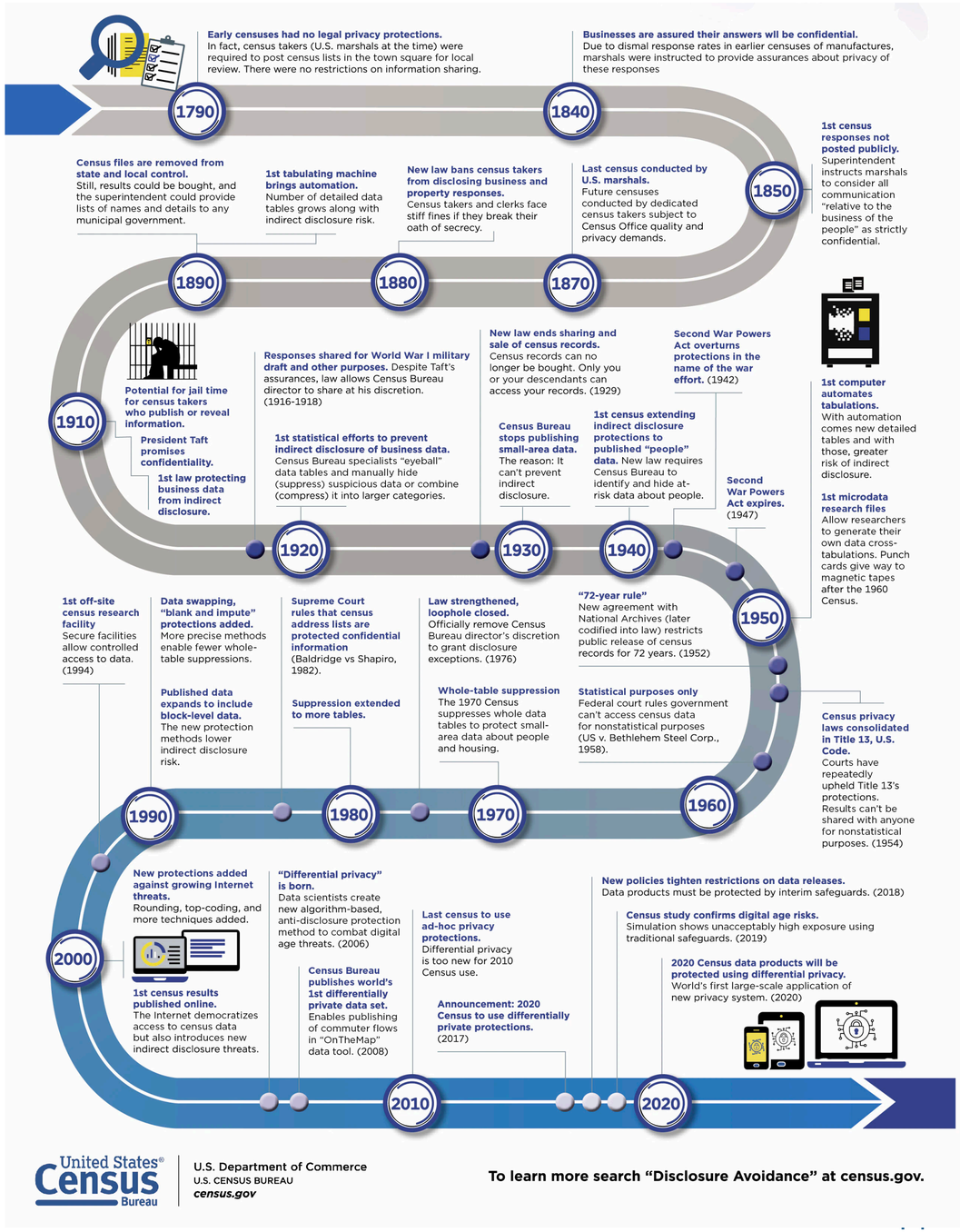
US Census data collection

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.

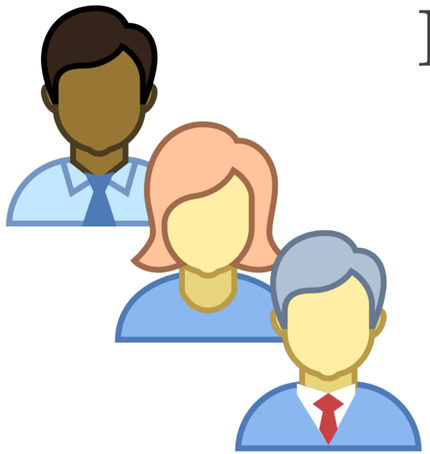
Ramachandran et al. 2012 McKenna et al. 2018

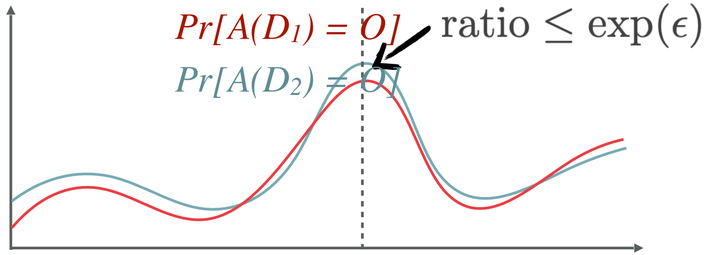
Differential Privacy

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:

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





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

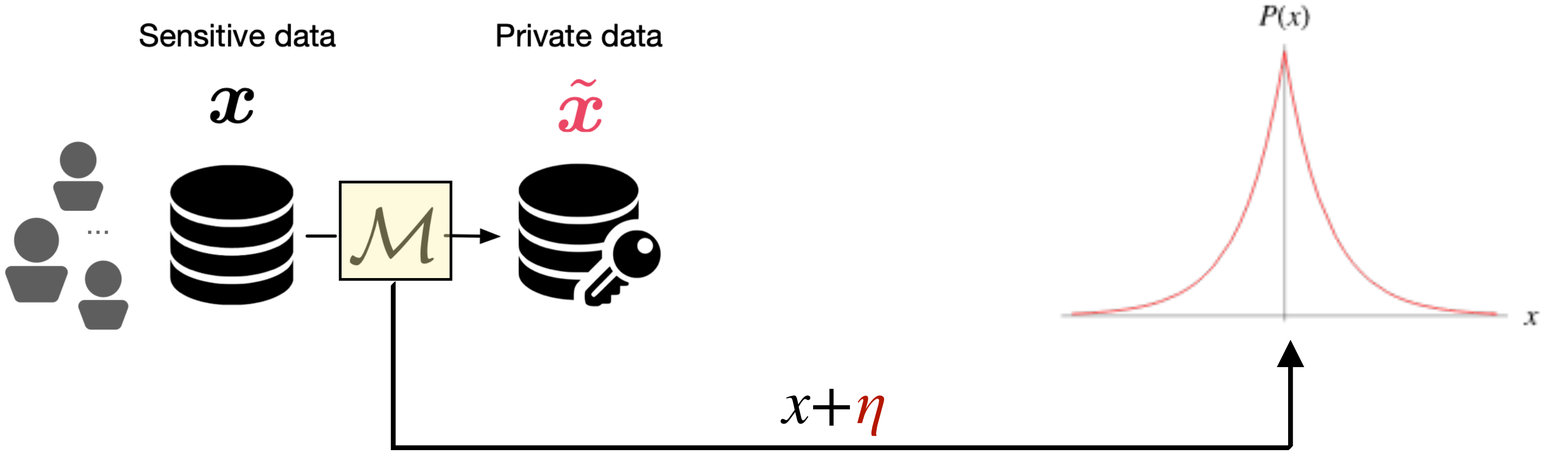
Dwork et al. 2006

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

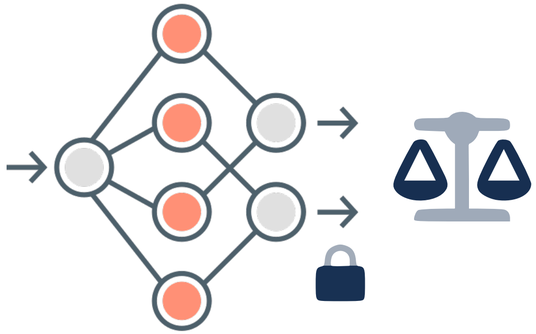


Agenda

Preliminaries

Fairness impacts of DP in decision making

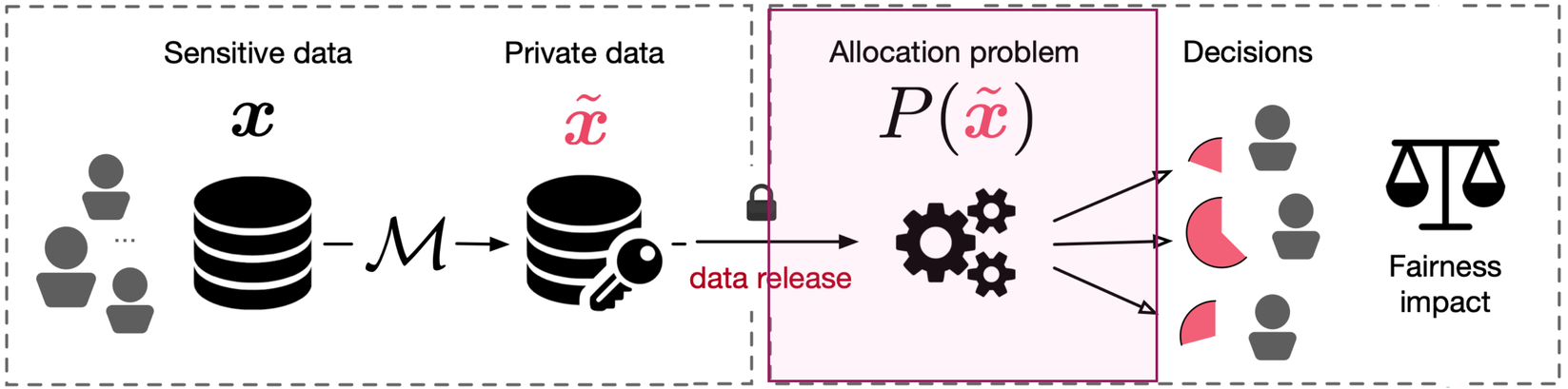
Fairness impacts of DP in learning



What's next?

Fairness in downstream decisions

Setting

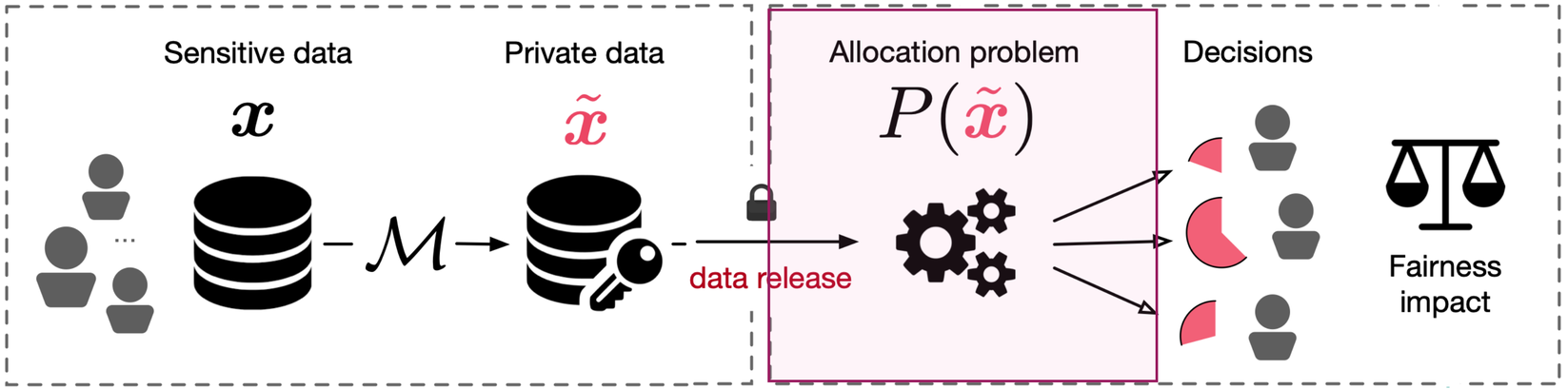


Analyze the unintended fairness impacts of a DP data-release mechanism to the outcome of a decision problem.

1. Allotment problems: which distributes a finite set of resources to some entity.
2. Decision rules: which determines whether an entity qualifies for some benefit.

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 (α-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 $i \in [n]$

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

Disproportionate impacts in downstream decisions

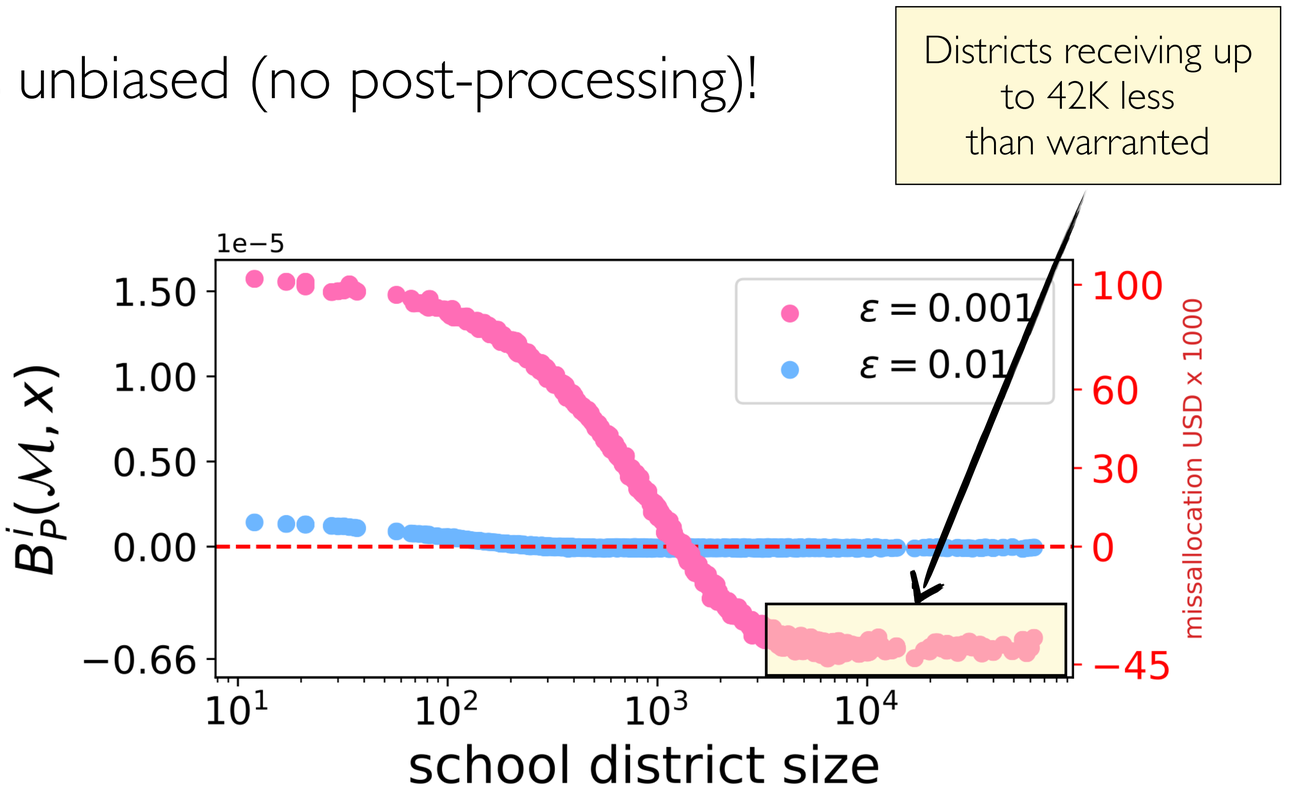
Title 1 allotment

- Even a simple allocation rule, applied on top of noisy data may produce biased decisions with significant fairness issues.
- This is true even if the DP mechanism is unbiased (no post-processing)!
- **Allotment**:

count of children 5 to 17 in district i

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

student expenditures in district i



Shape of the decision problem

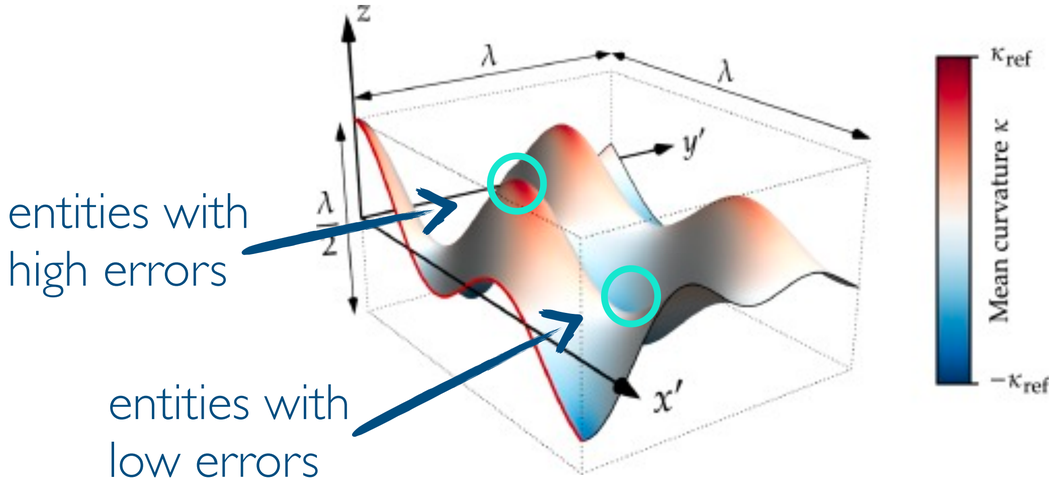
First key result

- **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)



- 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 $\boldsymbol{P_i}$ 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.

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:

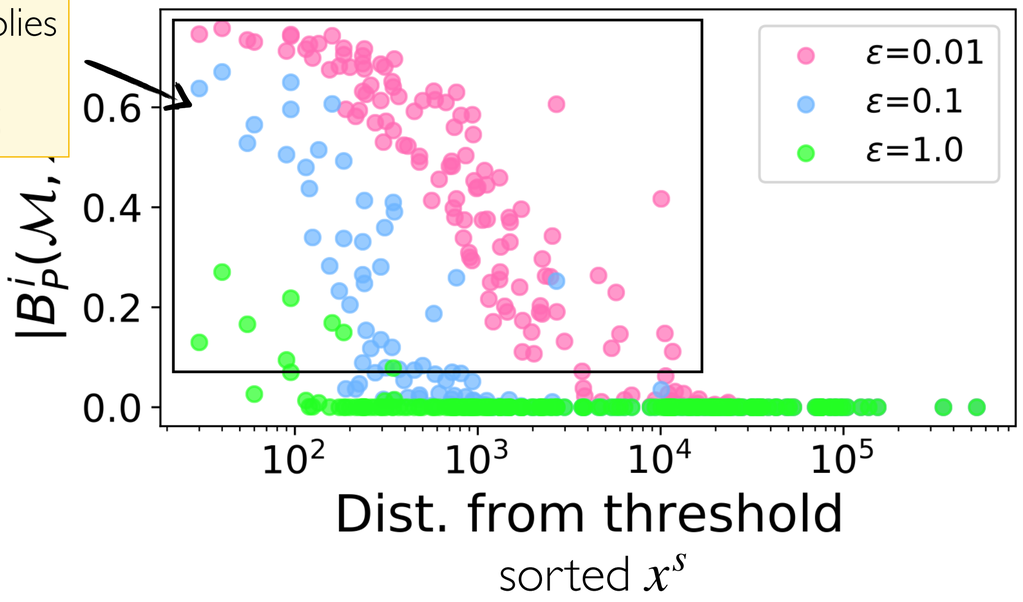
+ < 5th grade education

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

no. of ppl in i speaking minority language s

+ limited English proficiency

Misclassification implies potentially disenfranchising



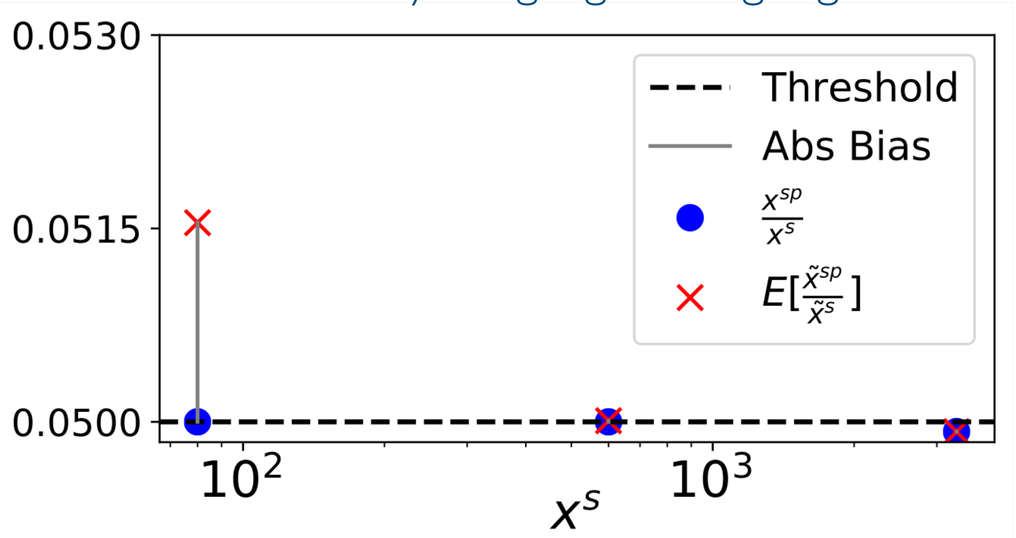
Fair Decision Rules

Ratio Functions

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

- *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

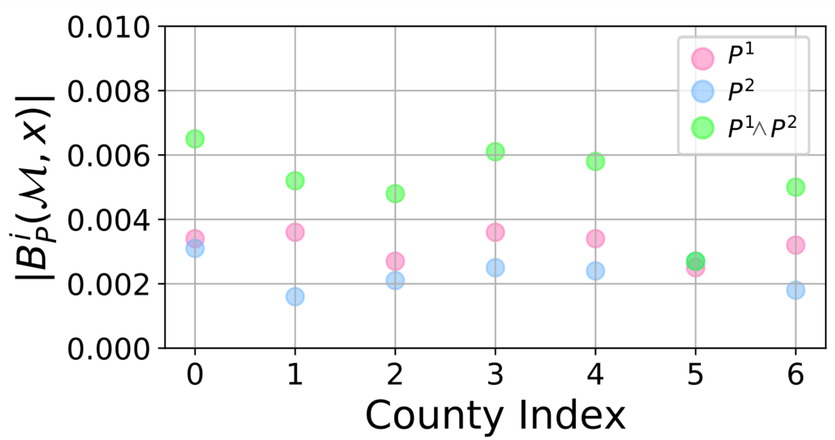
Second key result

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

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

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

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.

- **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

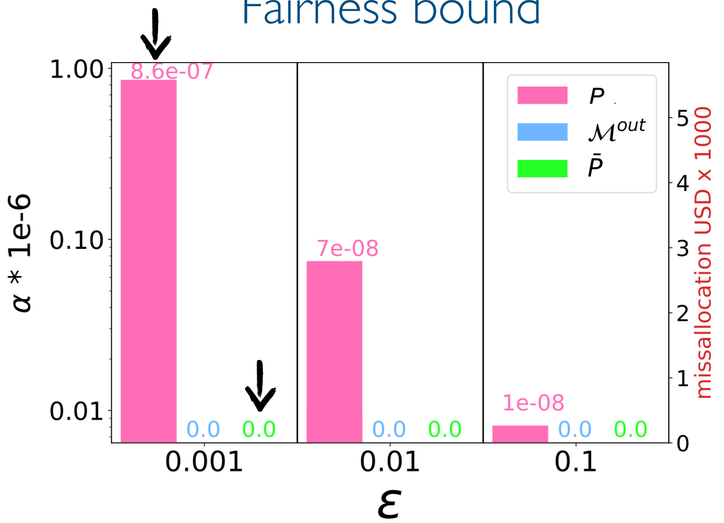
- Note that the observed issues are not data-driven, but problem-driven.
- **Corollary**: If P is a linear function, then mechanism M is fair w.r.t. P.
- Linearizing the allotment problem — General idea: Given a problem P_i derive a linear approximation $\tilde{P}_i$ of P_i

Redundant data release

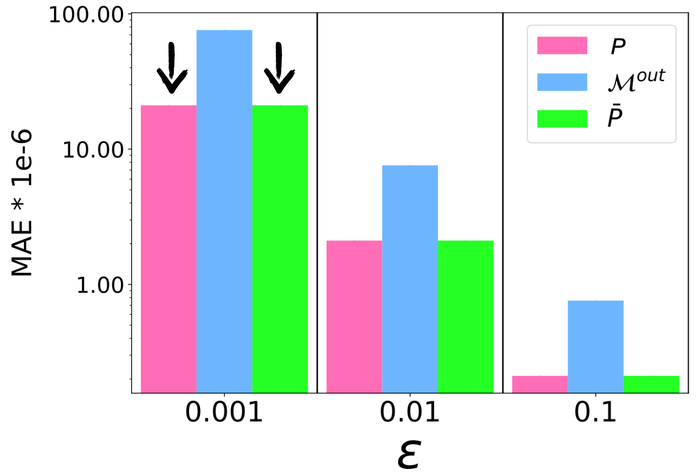
$$P_i^F(\boldsymbol{x}) \stackrel{\text{def}}{=} \left(\frac{x_i \cdot a_i}{\sum_{i \in [n]} x_i \cdot a_i} \right)$$

Release its (noisy) version as a constant

Fairness bound



Mean Abs Error



Mitigation solution

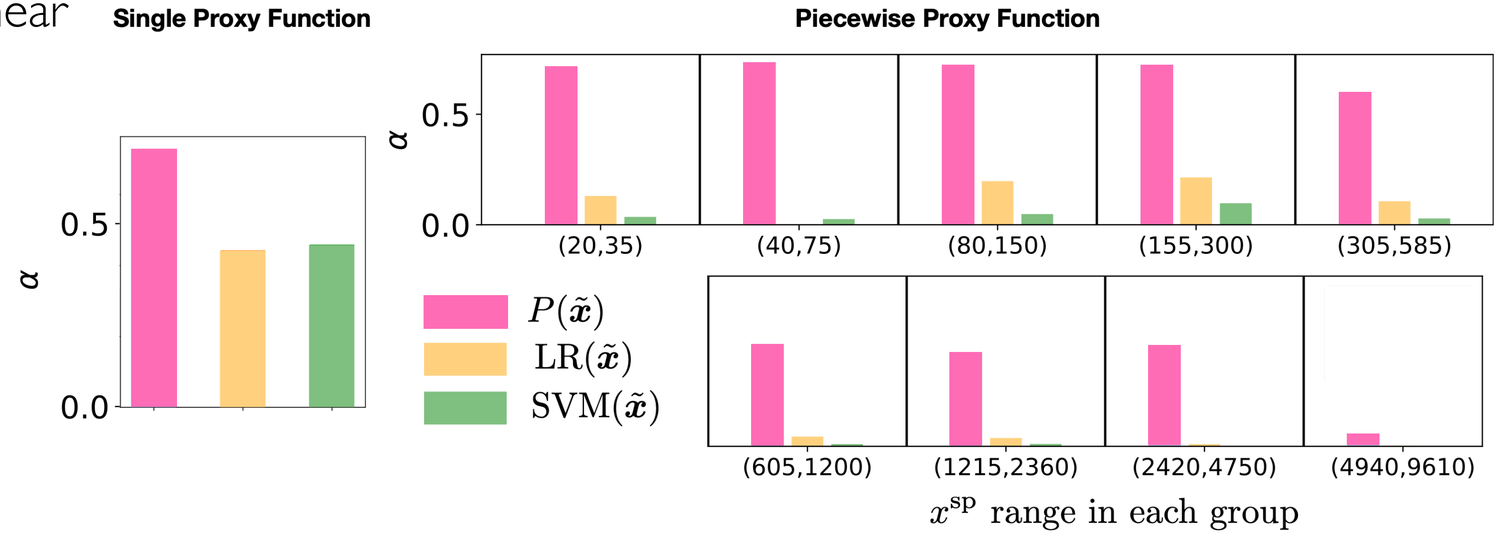
Fair Decision Rules

- Much more difficult scenario. But we could resort to the linear approximation trick again.

1. Partition dataset into groups x^{sp}
2. Train subgroups using features x using a linear classifier
3. Use the parameters of the proxy linear model LR(x) or SVM(x) to make a decision i.e., to approximate P_i^M

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

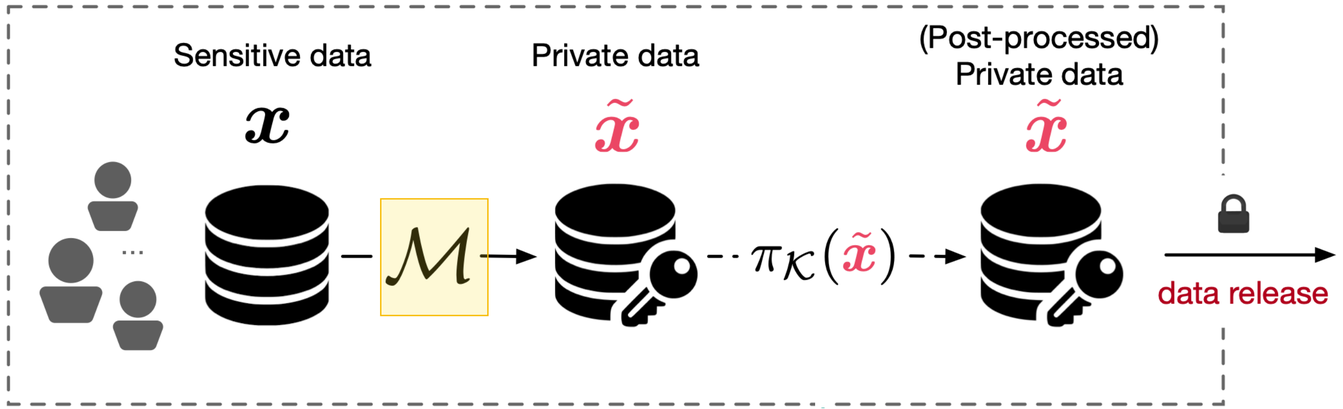
Result summary: Fairness violation decreases substantially, within each subgroup.



DP Post-processing

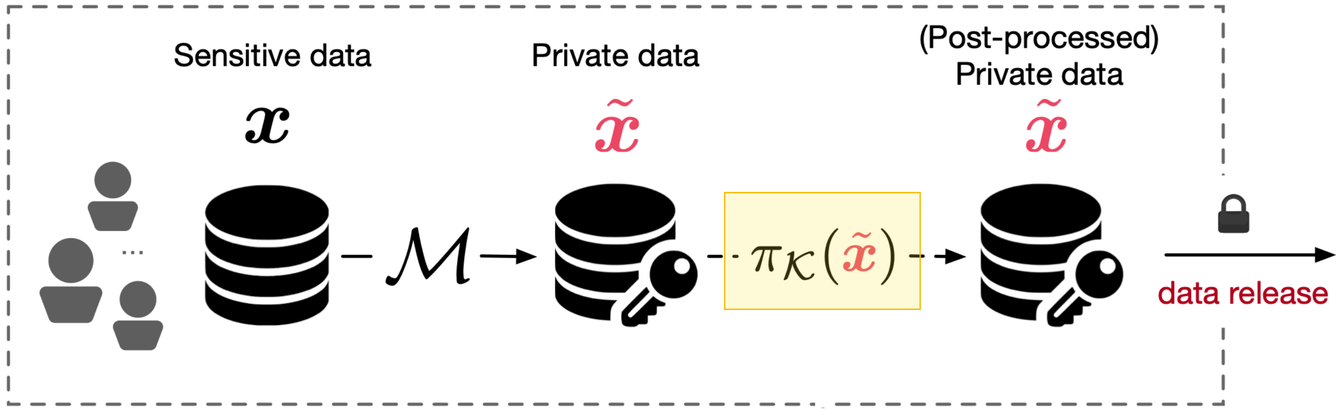
Fairness impact

DP data release with post-processing



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

DP data release with post-processing



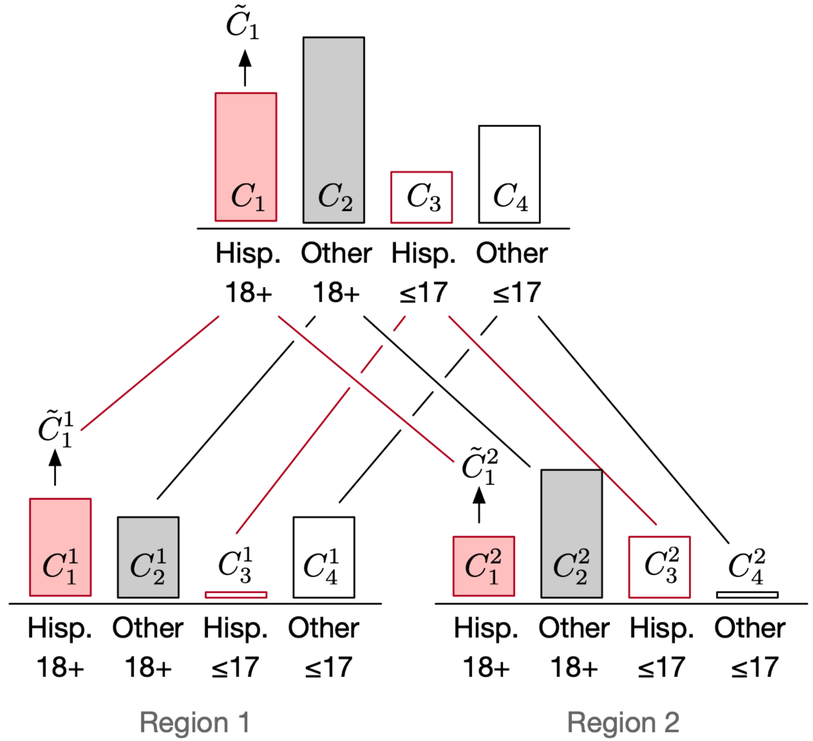
1. Apply noise with appropriate parameter $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \text{Noise}$

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

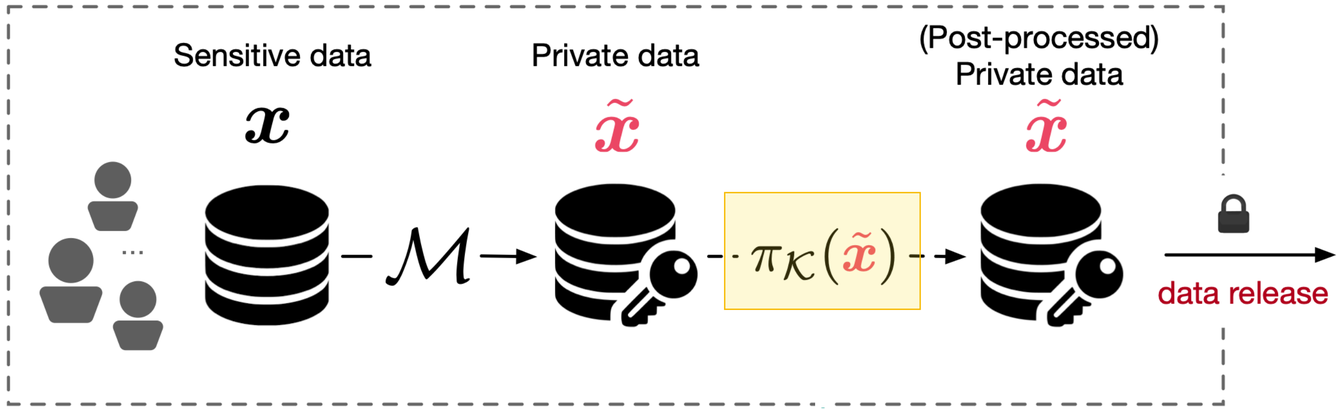
$$\pi_{\mathcal{K}}(\tilde{\boldsymbol{x}}) : \underset{\boldsymbol{v} \in \mathcal{K}}{\operatorname{argmin}} \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2$$

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



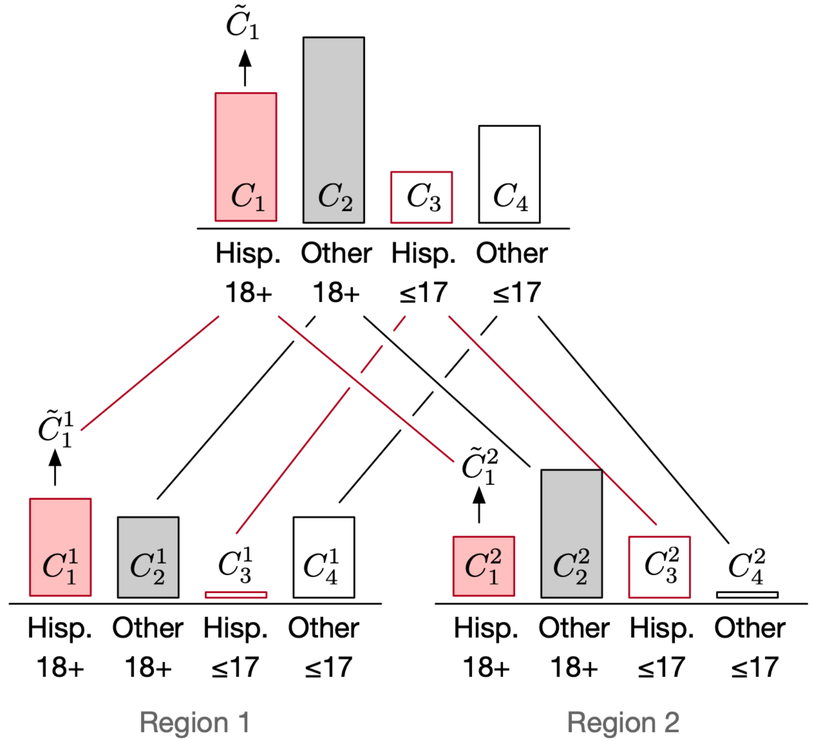
1. Apply noise with appropriate parameter $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \text{Noise}$

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

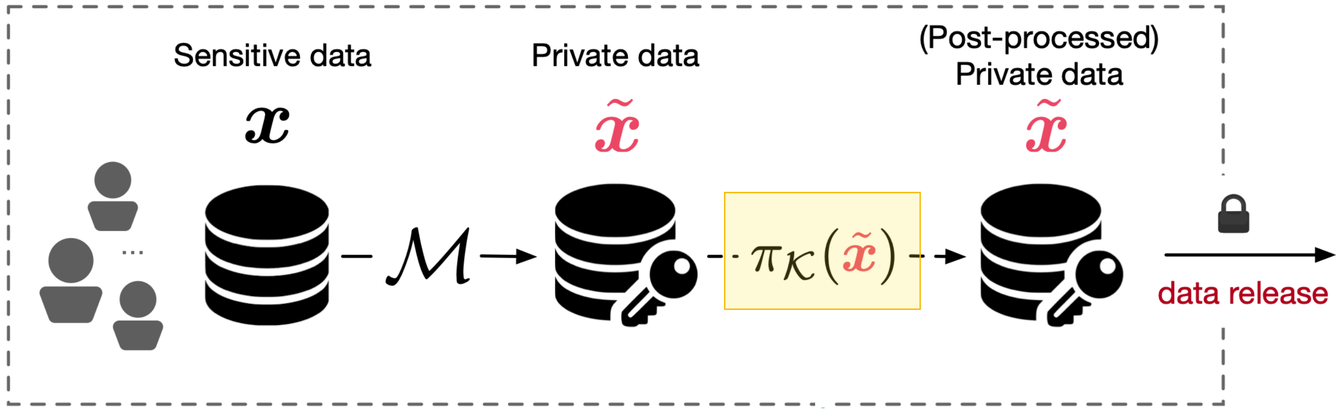
$$\pi_{\mathcal{K}}(\tilde{\boldsymbol{x}}) : \underset{\boldsymbol{v} \in \mathcal{K}}{\operatorname{argmin}} \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2$$

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



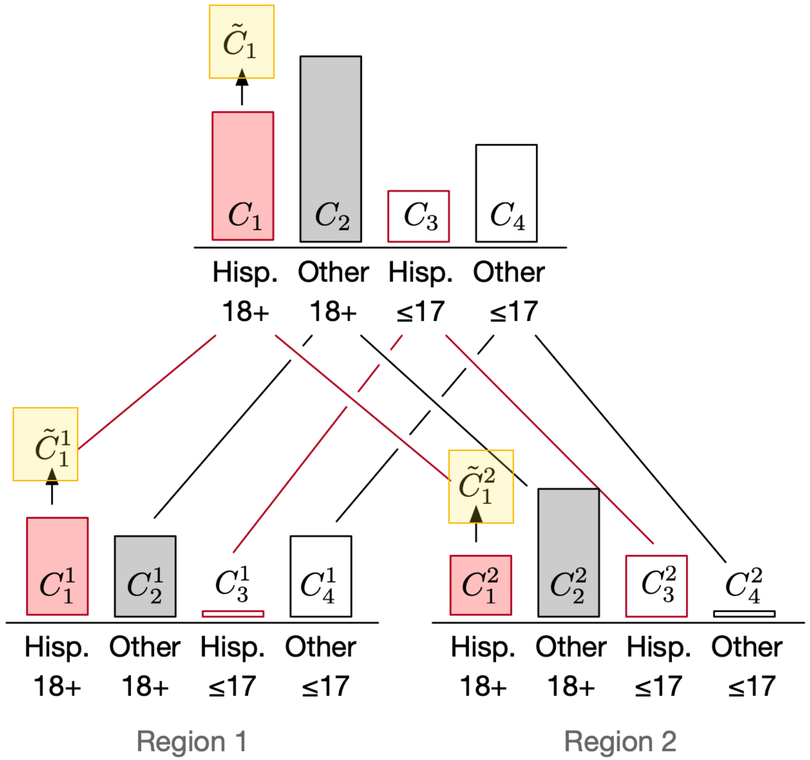
1. Apply noise with appropriate parameter $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \text{Noise}$

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

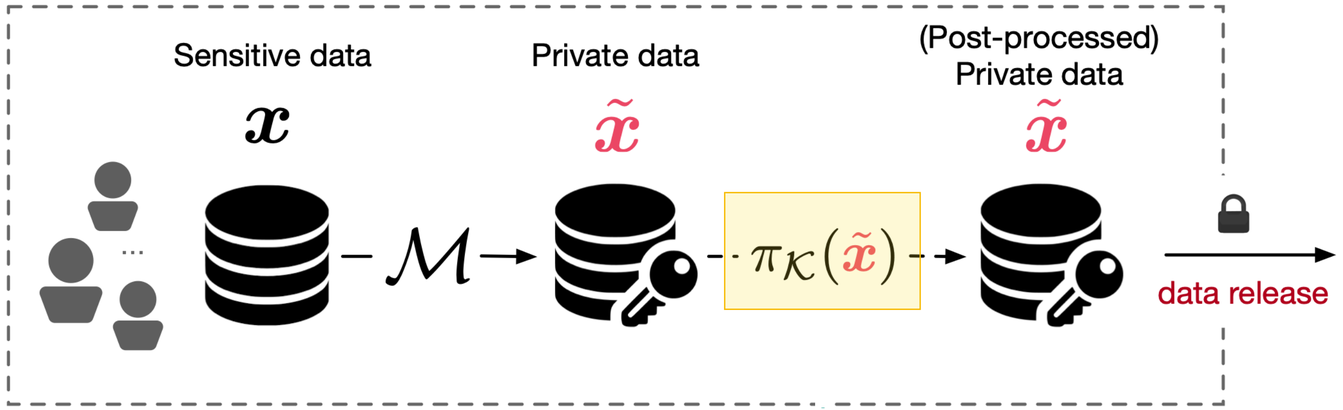
$$\pi_{\mathcal{K}}(\tilde{\boldsymbol{x}}) : \underset{\boldsymbol{v} \in \mathcal{K}}{\operatorname{argmin}} \|\boldsymbol{v} - \tilde{\boldsymbol{x}}\|_2$$

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



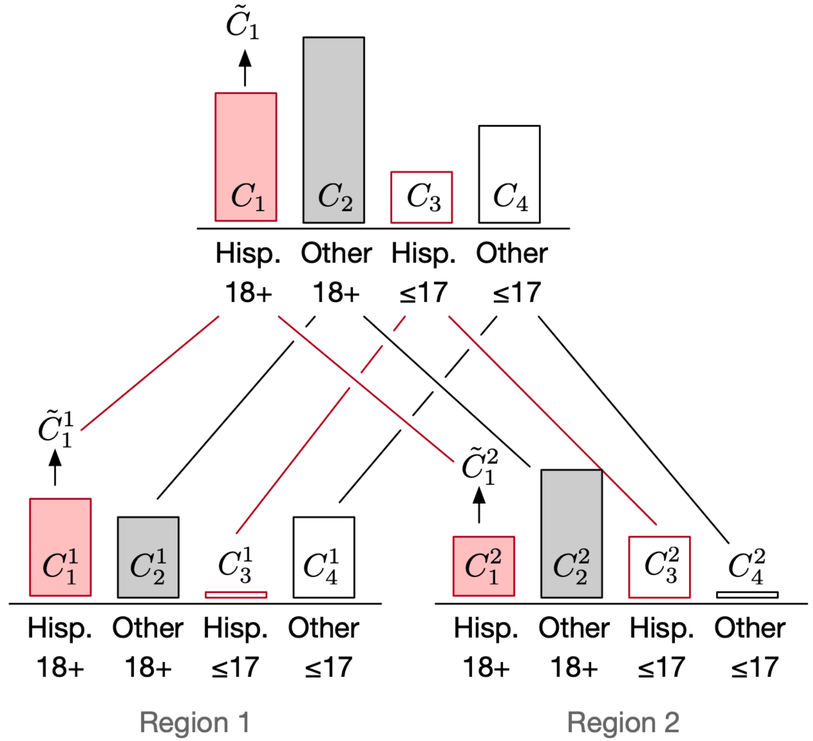
1. Apply noise with appropriate parameter $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \text{Noise}$

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

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

with feasible region defined as

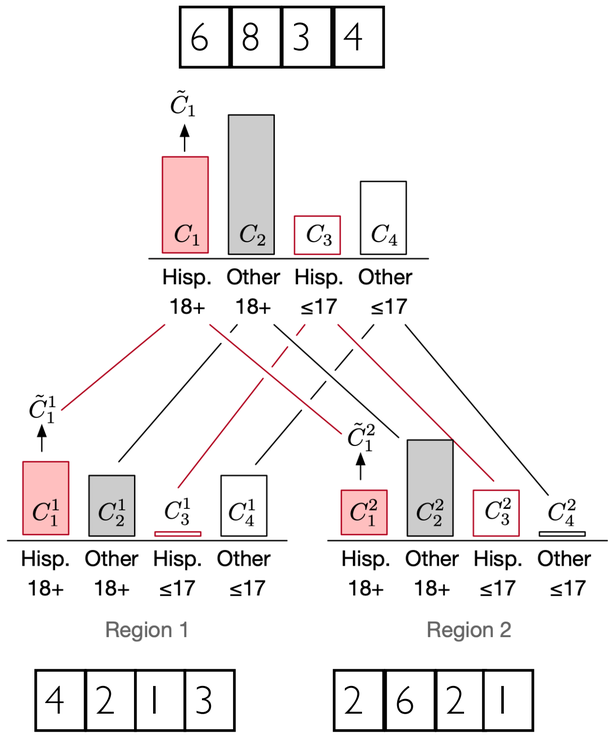
$$\mathcal{K} = \left\{ \boldsymbol{v} \mid \sum_{i=1}^{n} v_i = C, \boldsymbol{v} \geq 0 \right\}$$



Satisfies DP due to post-processing immunity

DP post-processing

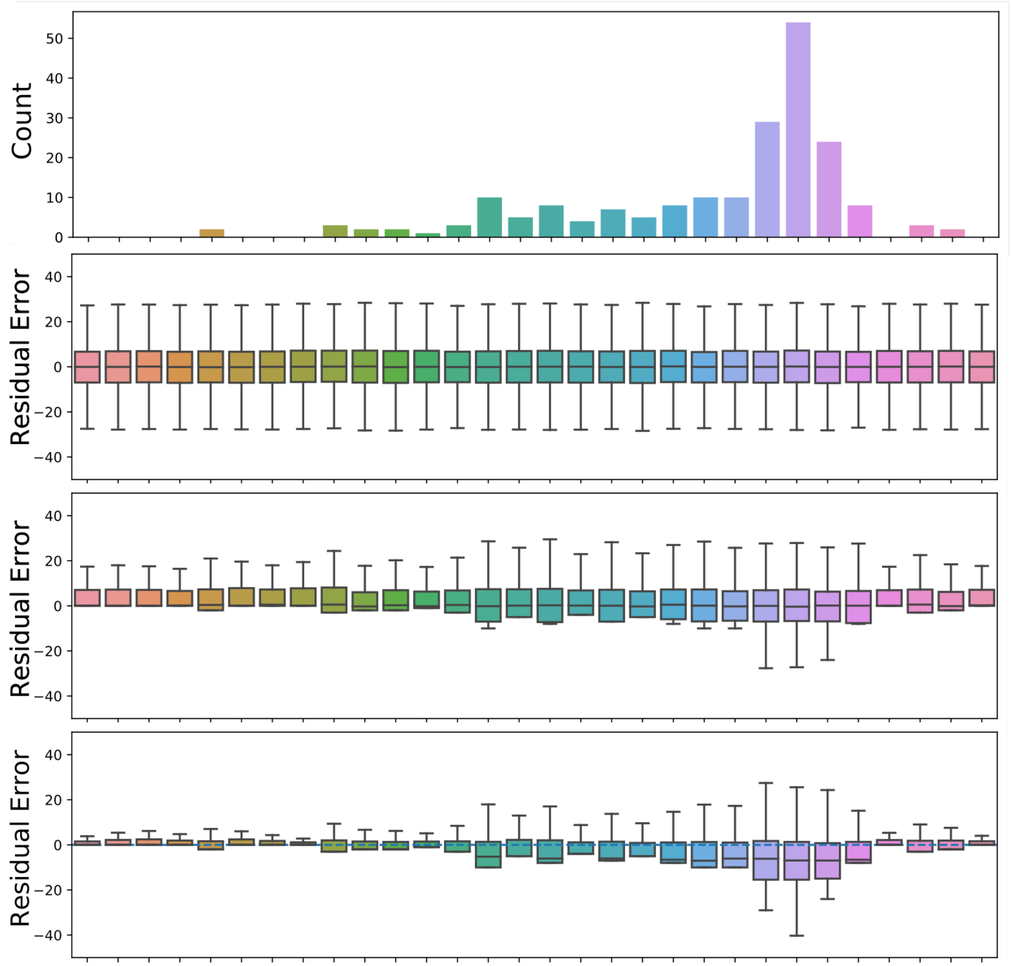
Error and bias



Laplace mechanism

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

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



DP post-processing

Error and bias

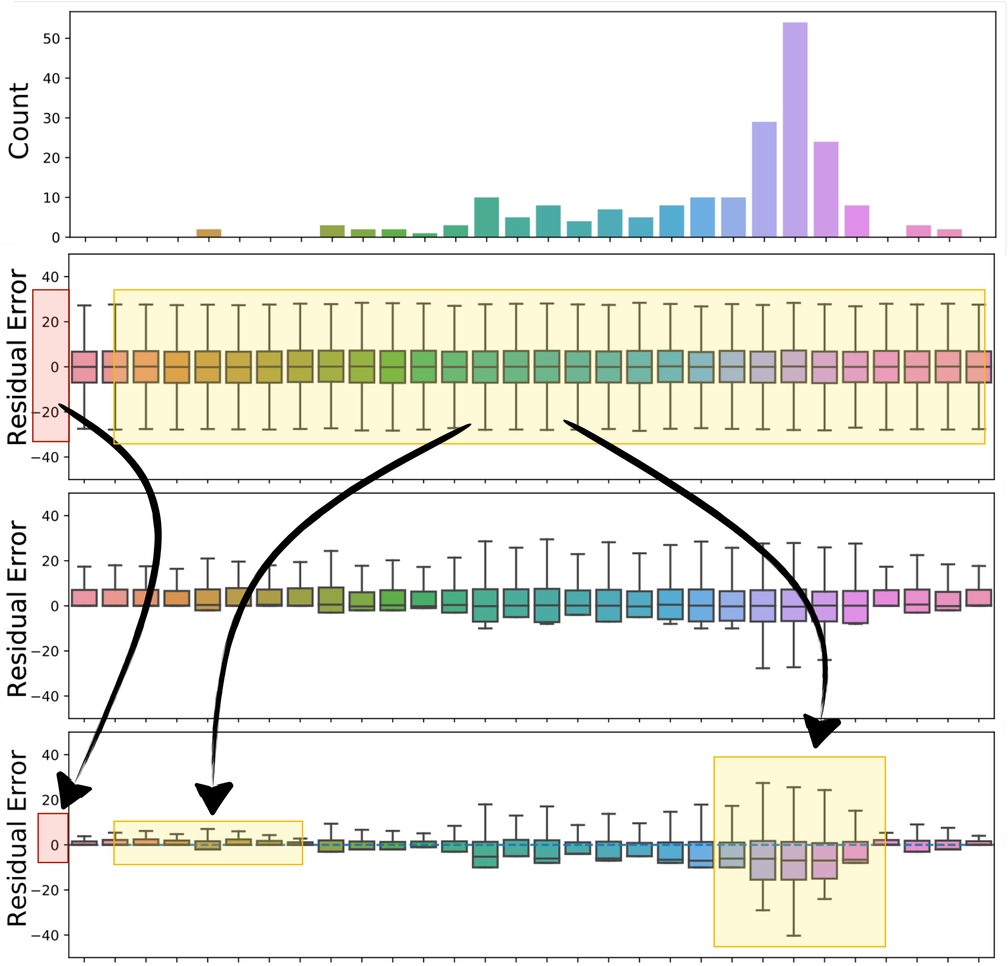
Observe that post-processing reduces the errors.

However, it increases unfairness!

Laplace mechanism

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

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

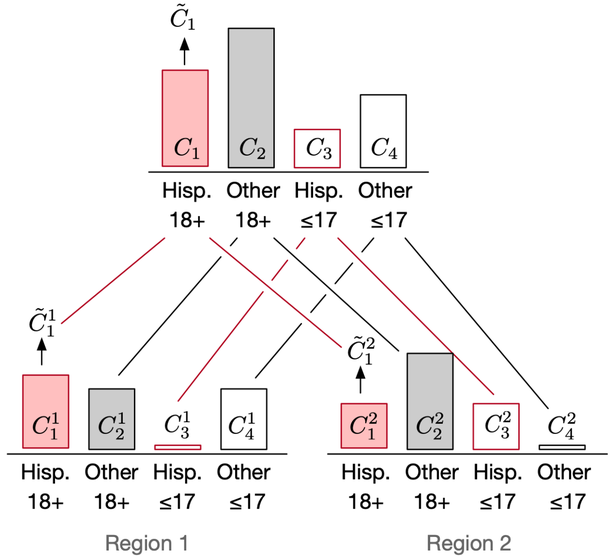


Zhu et al. AAAI:2021
Zhu et al. IJCAI:2022

Bias of post-processing

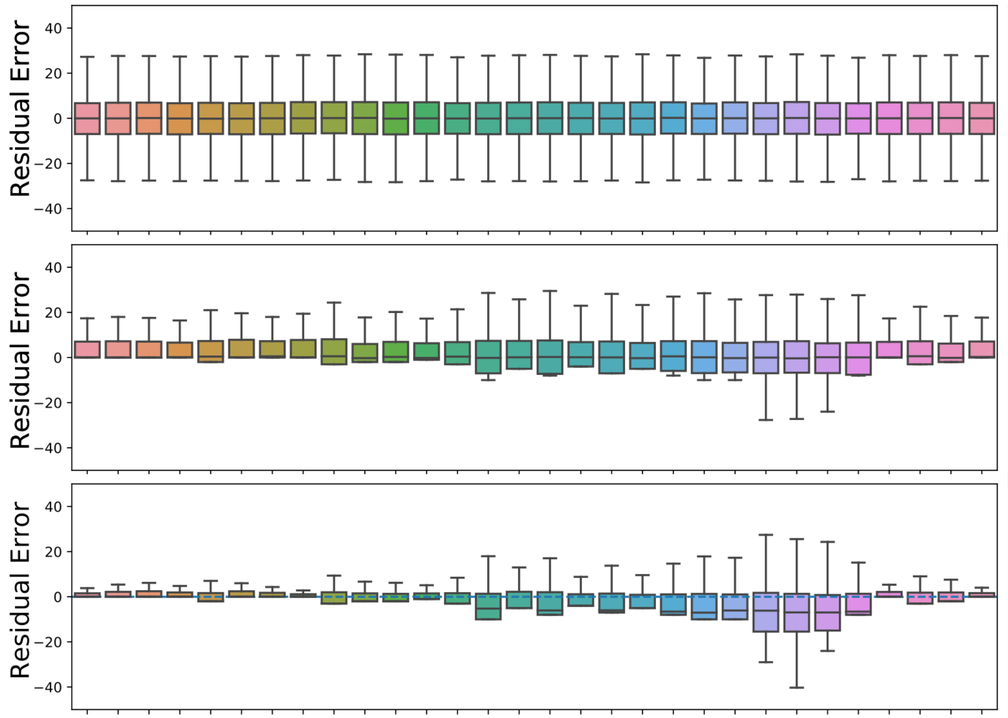
Key result

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



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

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



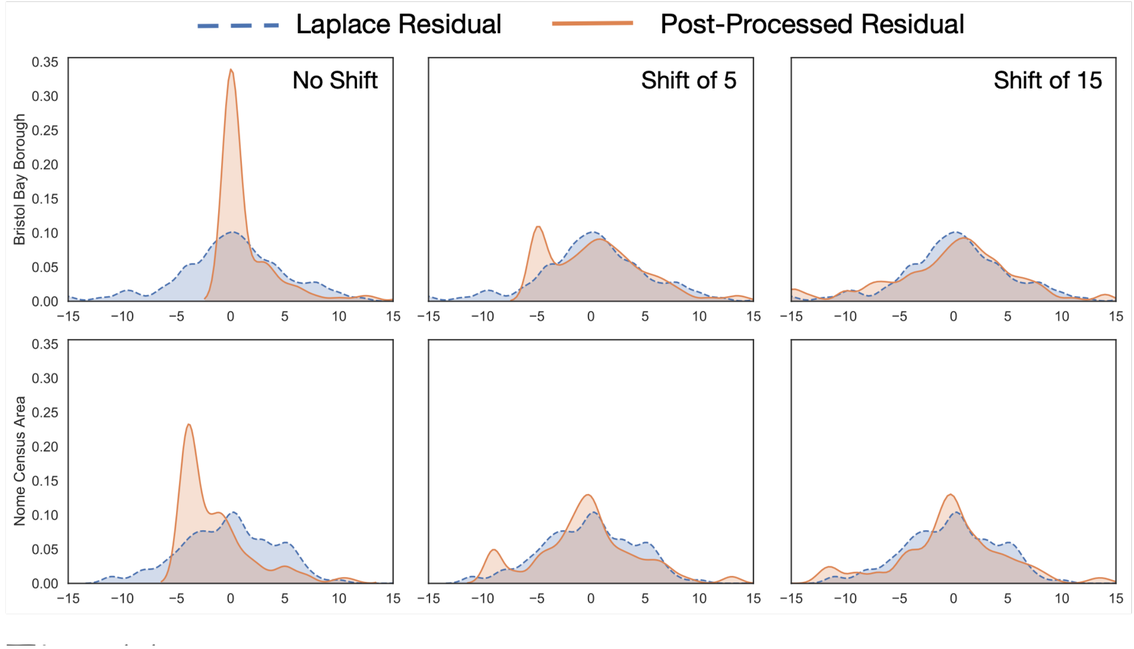
 Zhu et al. AAAI:2021 Zhu et al. IJCAI:2022

Quantifying bias in post-processing

Theorem: *Suppose that the noisy data $\tilde{\boldsymbol{x}}$ 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*

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

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



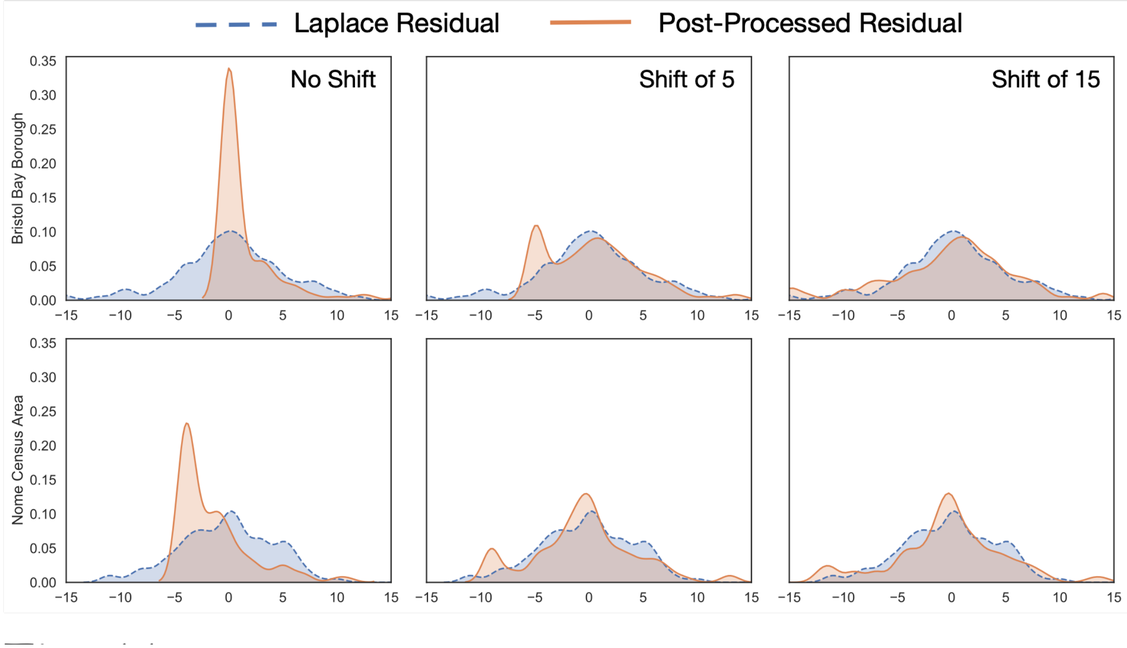
There is an ℓ_1-ball of radius $r_m = \min_i x_i$ and centered in x which is a feasible subspace where there is no bias.

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

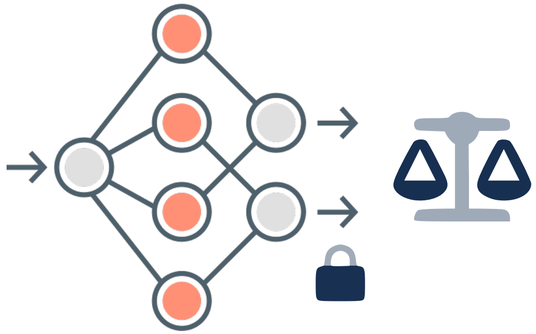
Although post-processing reduces errors, its application to policy determinations should take into account fairness issues.

Agenda

Preliminaries

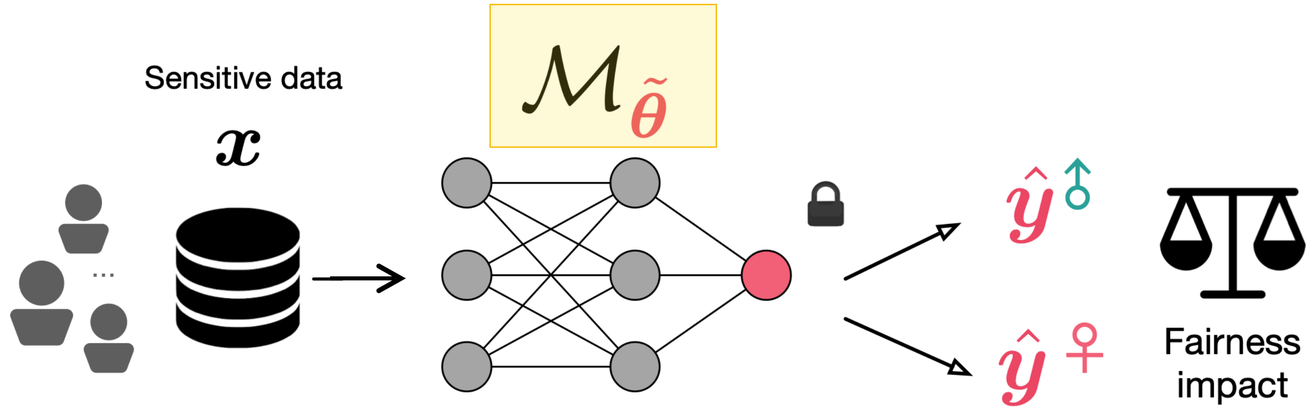
Fairness impacts of DP in decision making

Fairness impacts of DP in learning



What's next?

Fairness in DP learning tasks



- Given a dataset consisting of data points $(X_i,\ A_i,\ Y_i)$ the goal is to learn a classifier f_θ that guarantees privacy of the individual data points and the learning task minimizes

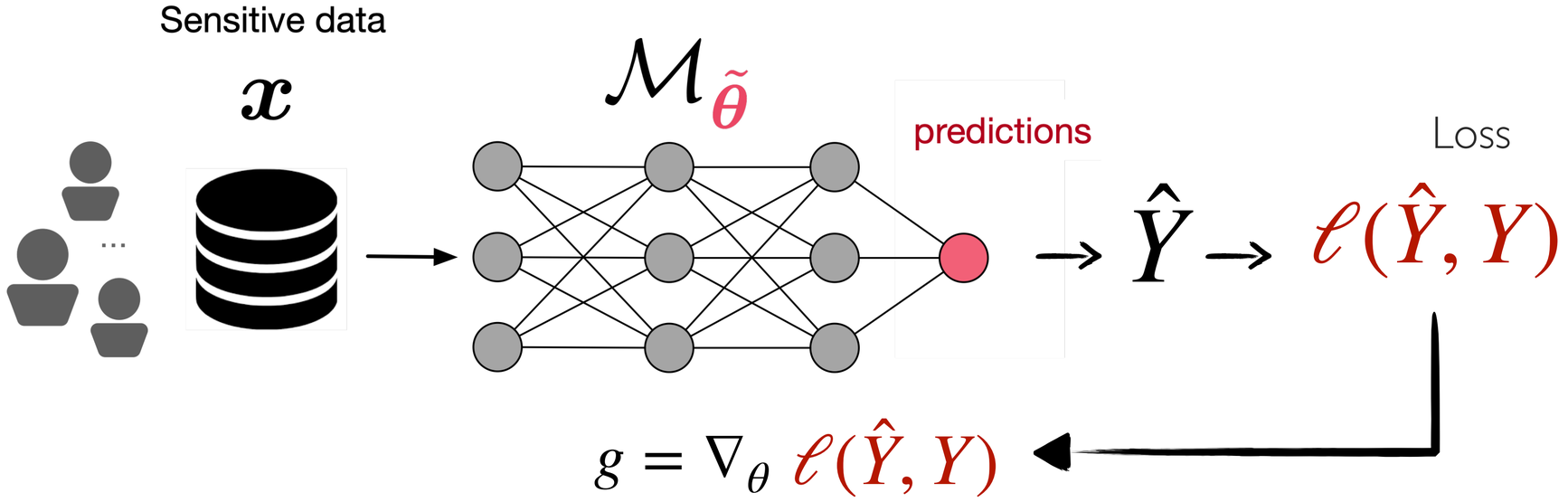
$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; D) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(X_i), Y_i)$$

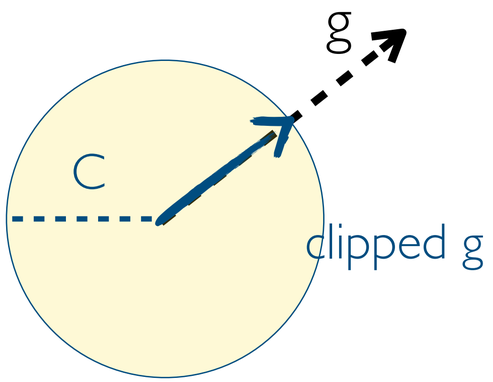
- Fairness focuses on the notion of excessive risk: $R(\boldsymbol{\theta}, D) = \mathbb{E}_{\tilde{\boldsymbol{\theta}}} \left[\mathcal{L}(\tilde{\boldsymbol{\theta}}; D) \right] - \mathcal{L}(\boldsymbol{\theta}^*; D)$, and is measured with respect to the excessive risk gap

$$\xi_a = |R_a(\boldsymbol{\theta}) - R(\boldsymbol{\theta})|$$

group-level population level

DP Stochastic Gradient Descent





1. Clip g in ℓ_2 ball of size C
2. Add noise from $\mathcal{N}\left(0, IC^2\sigma^2\right)$
3. Average gradients g's in a mini-batch
4. Update parameters θ

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:

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

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

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

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

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).

Why clipping causes unfairness?

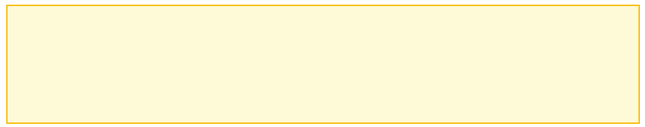
Gradient norms and excessive risk

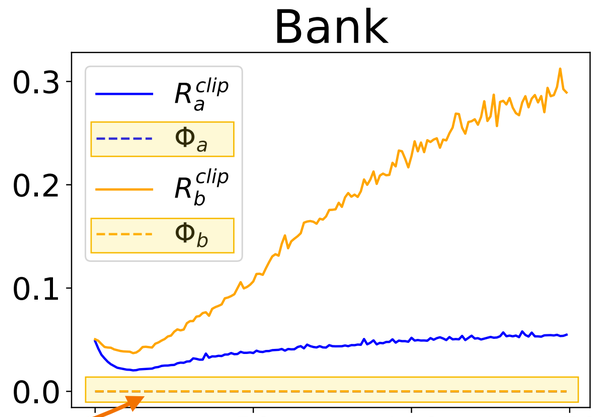
$$+\underbrace{\eta\left(\langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D\rangle - \langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_D\rangle\right) + \frac{\eta^2}{2}\left(\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] - \mathbb{E}\left[\boldsymbol{g}_B^T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

Why clipping causes unfairness?

Gradient norms and excessive risk

Negligible impact





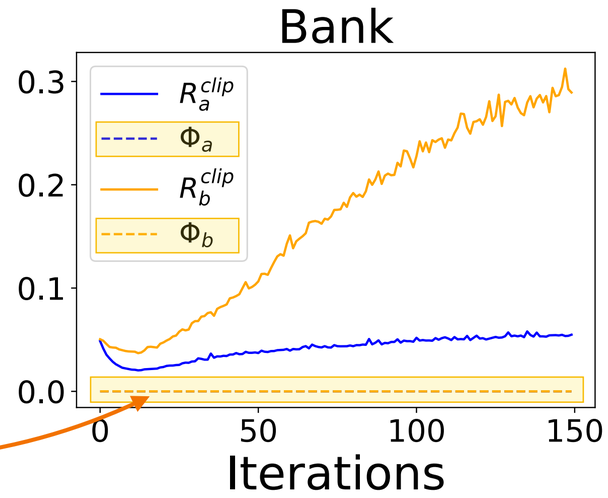
$$+\underbrace{\eta\left(\langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D\rangle - \langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_D\rangle\right) + \frac{\eta^2}{2}\left(\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] - \mathbb{E}\left[\boldsymbol{g}_B^T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

Why clipping causes unfairness?

Gradient norms and excessive risk

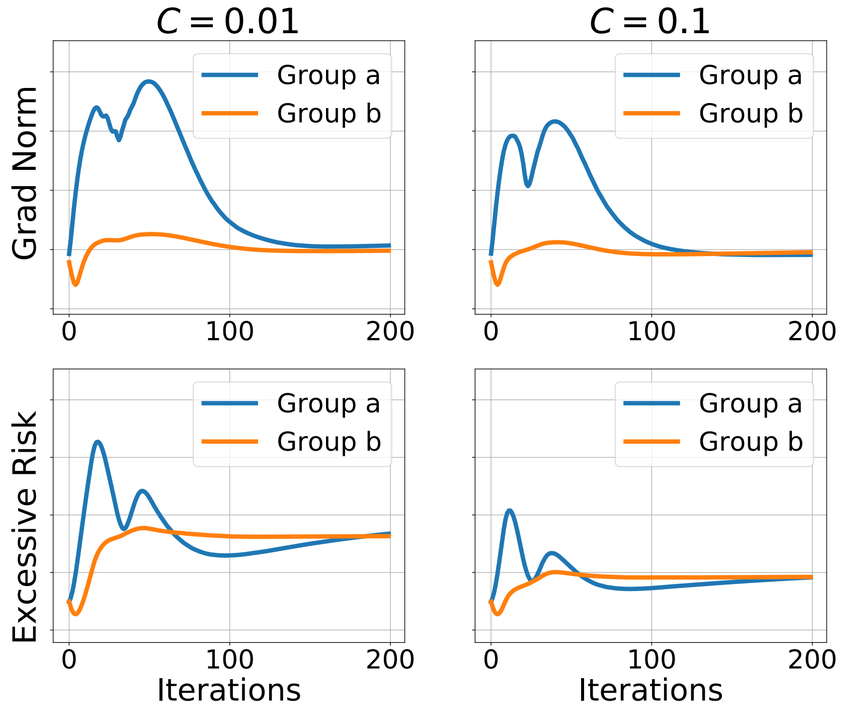
Negligible impact

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



Crucial Proxy to Unfairness (due to clipping)

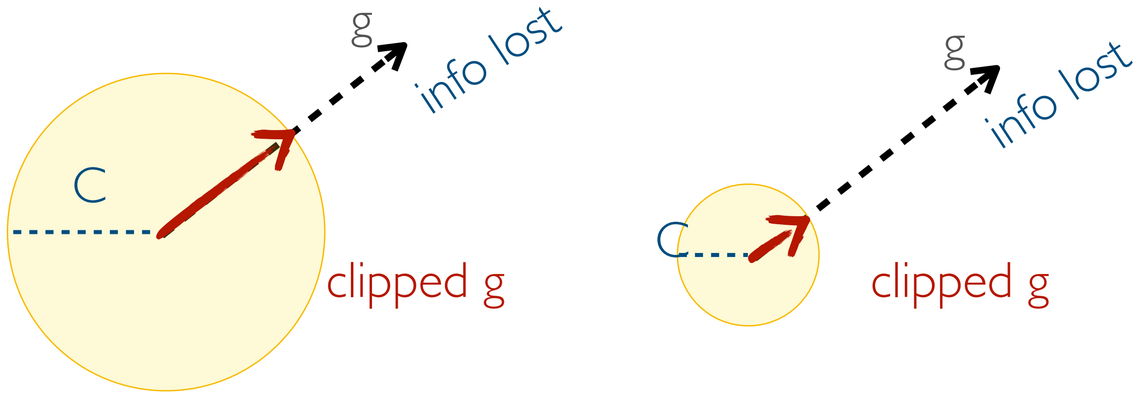
Theorem (informal): Gradient flow affects the excessive risk (unfairness) of the individuals and groups.



Why clipping causes unfairness?

Gradient norms and excessive risk

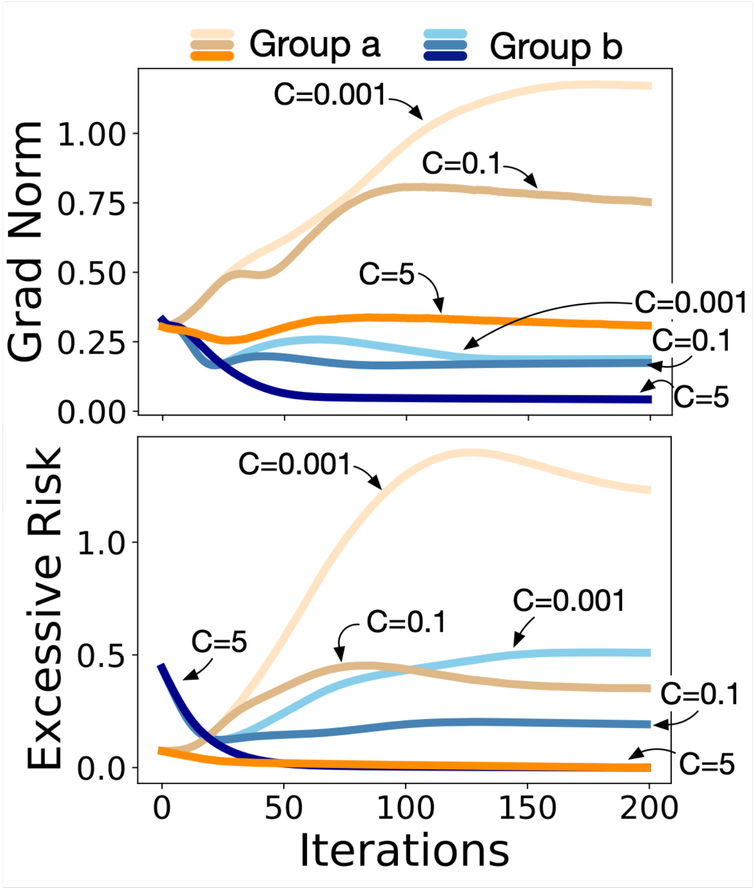
- When clipping, the smaller C, the higher is the information loss of the average gradients that are backpropagated.



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:*

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

Impact of gradient clipping (Bank dataset)



Why noise causes unfairness in DP-SGD?

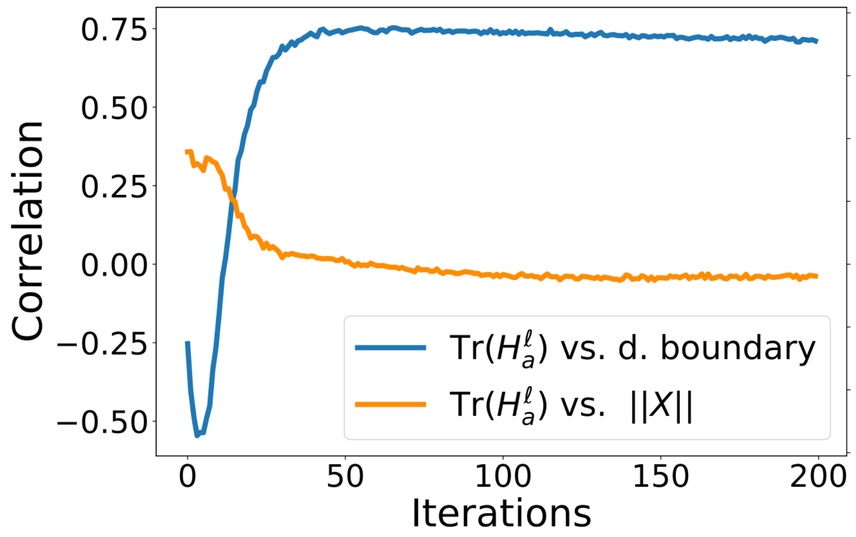
$$+ \underbrace{\frac{\eta^2}{2}\,\mathrm{Tr}(\boldsymbol{H}_\ell^a)C^2\sigma^2}_{\textit{private term due to noise}} \qquad (R_a^{noise})$$

Why noise causes unfairness in DP-SGD?

Distance to the decision boundary and excess risk

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

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

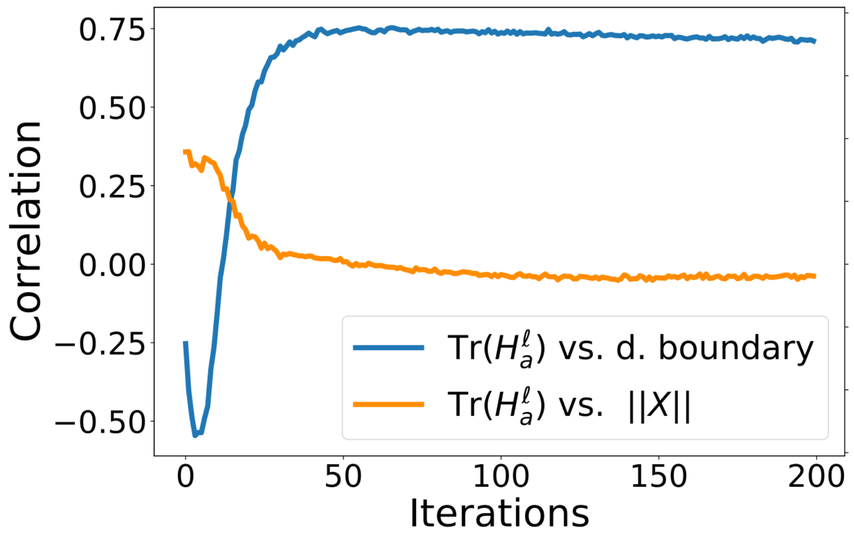
$$\underbrace{\frac{\eta^2}{2}\mathrm{Tr}(\boldsymbol{H}_\ell^a)C^2\sigma^2}_{\textit{private term due to noise}}$$

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



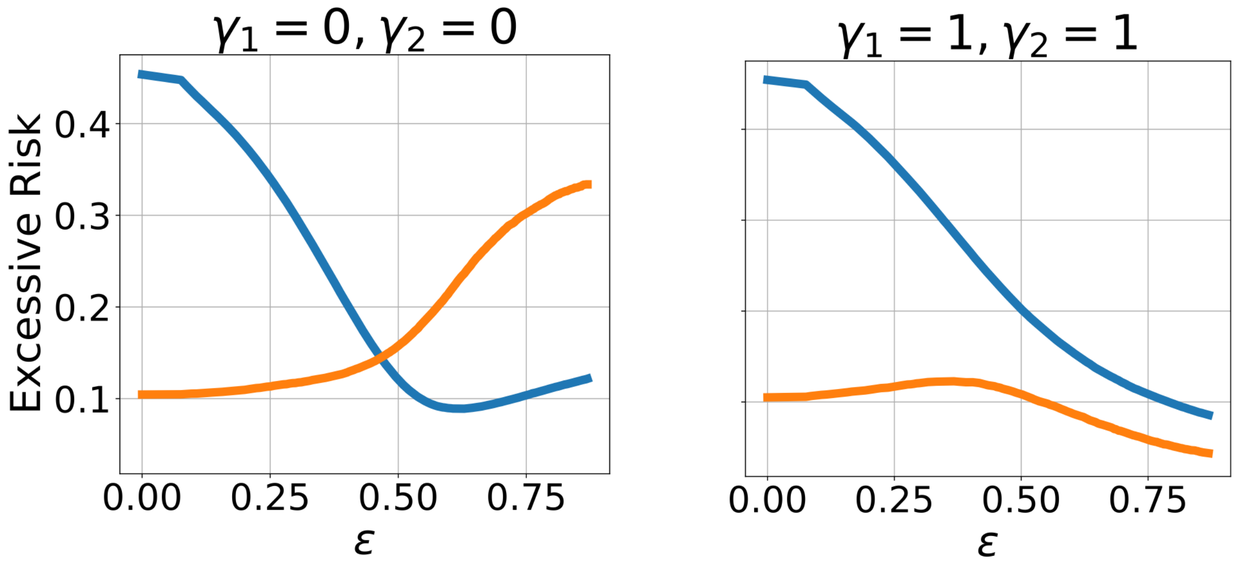
Mitigating solutions

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

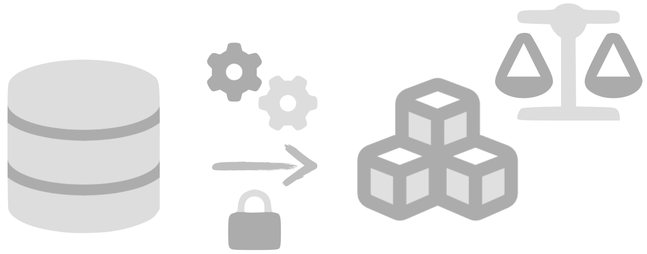
Majority group



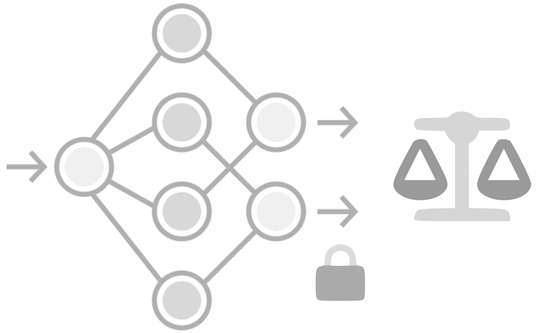
Agenda

Preliminaries

Fairness impacts of DP in decision making



Fairness impacts of DP in learning



What's next?

Privacy and Equity of decision making

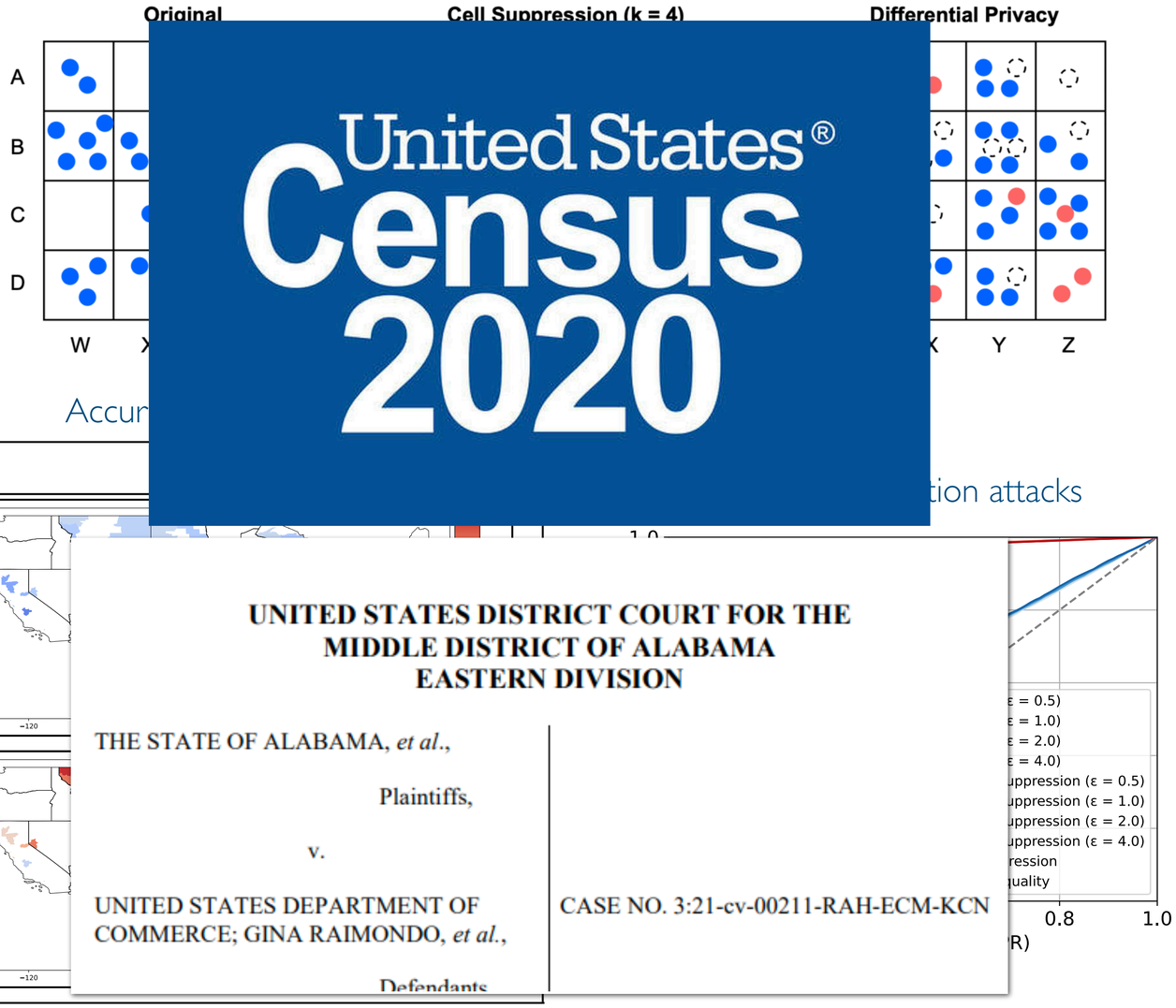
Need

Analyze the cost of traditional disclosure avoidance techniques on public policy decisions.

Challenges

How to compare traditional disclosure avoidance techniques with modern techniques (DP).

- Privacy
- Bias and Variance
- Fairness



Unfairness in Hardware Network Pruning

Pruning has a disparate impact on model accuracy

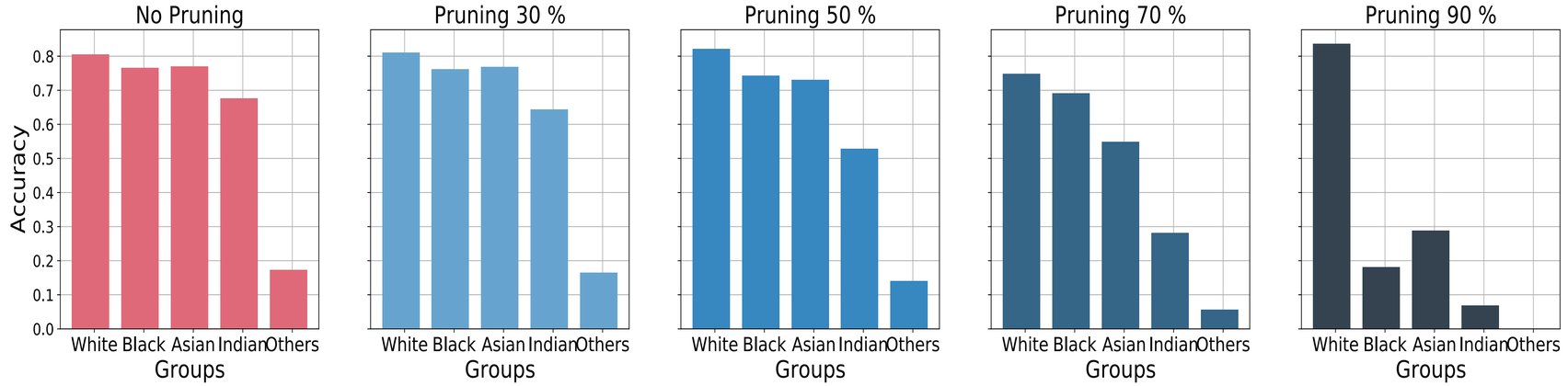


Figure 1: Accuracy of each demographic group in the UTK-Face dataset using Resnet18 [18], at the increasing of the pruning rate.

Unfairness in Hardware Tooling

On The Fairness Impacts of Hardware Selection in Machine Learning

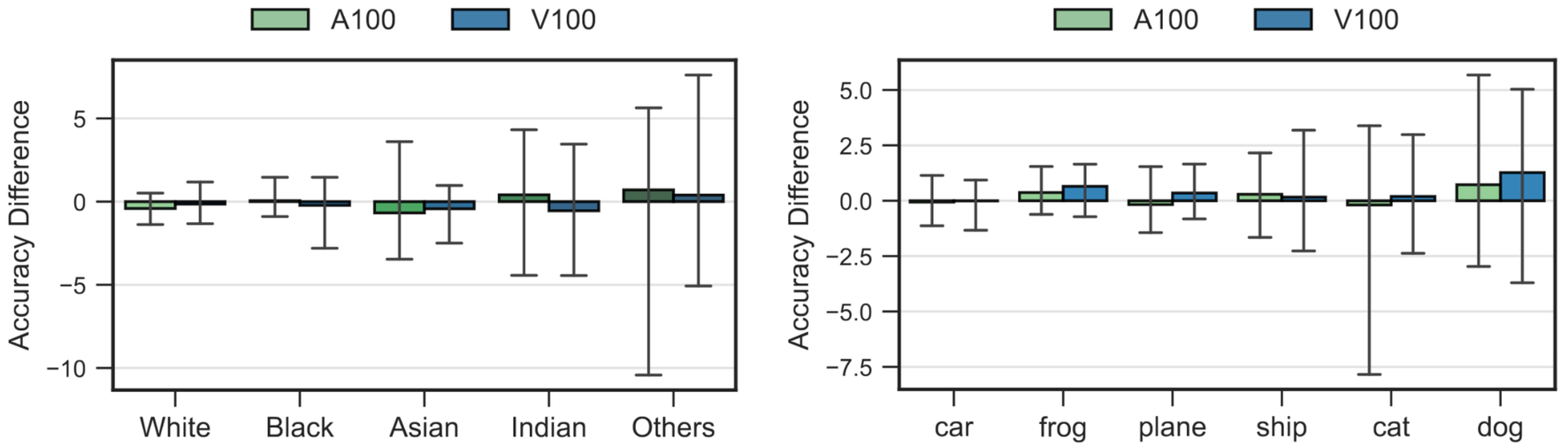


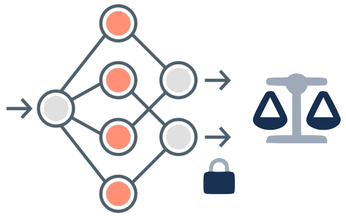
Figure 1: A training model (ResNet34) with the same parameters (random seeds, epochs, batch-size) on different hardware can have vastly different performance results, especially for minority groups (dark colors). The reference hardware is T4. **Left**: UTK-Face, **Right**: CIFAR-10.

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.

Discussion

- In ML, there is often a trade-off between accuracy and fairness. Can you think of examples where this trade-off is evident? How should it be managed?
- **Trade-offs**: When and when not DP may lead to fairness issues in AI models? How can these trade-offs be balanced or mitigated?
- **Impact of DP on Data Representation**: Can DP affect representation of minority groups in datasets?
- **Role of Differential Privacy in Bias Mitigation:** How can differential privacy be used as a tool to mitigate bias in AI models?
- How does the concept of fairness intersect with other performance metrics of a model? Is it possible to achieve high fairness without sacrificing other metrics?

- **Policy considerations:**
- Discuss the ethical implications of not addressing fairness and privacy in AI systems. Who is most at risk?
- What role should policy and regulation play in ensuring fairness and privacy in AI?

Responsible AI:
Seminar on Fairness, Safety, Privacy and more

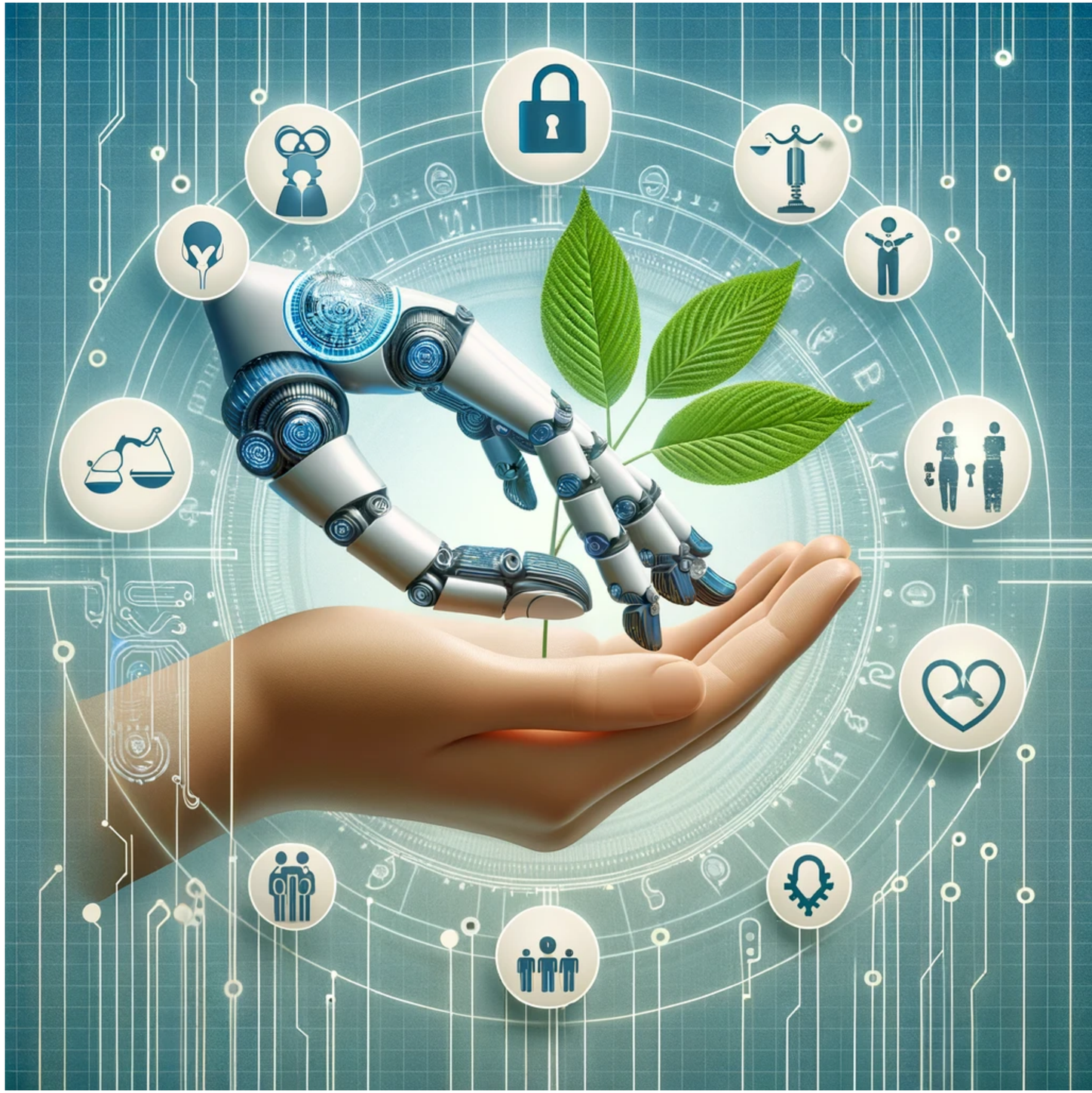
Thank you!

https://nandofioretto.com

nandofioretto@gmail.com

@nandofioretto

Ferdinando Fioretto @UVA Spring 2024



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.*

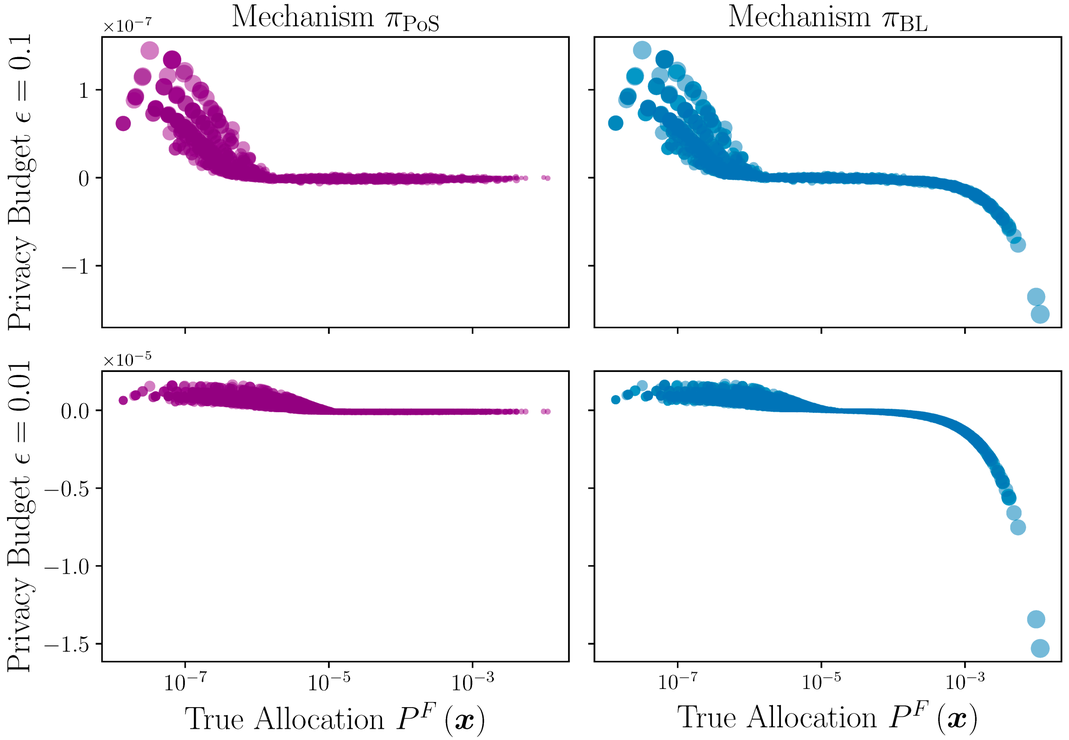
$$\pi_{\mathrm{PoS}}\left(\tilde{\boldsymbol{x}}\right) := \underset{\boldsymbol{v}\in\Delta^n}{\arg\min} \left\|\boldsymbol{v} - P^F\left(\tilde{\boldsymbol{x}}\right)\right\|_2 \qquad (P_{\mathrm{PoS}})$$

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

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

which closely approximate the optimal post-processing mechanism

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



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**

Properties of the training data

Fairness impact

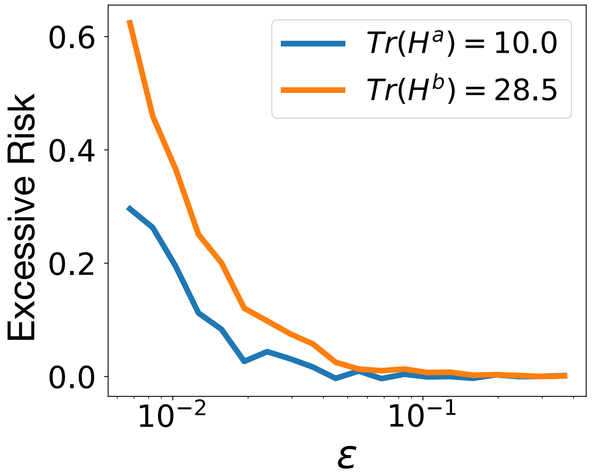
Warm up: output perturbation

Input norms

- Adds Gaussian noise $\mathcal{N}\left(0, \Delta_\ell^2\, \sigma^2\right)$ to the optimal model parameters θ^*
- **Thm:** For ℓ twice differentiable convex functions, the excessive risk gap is approximated as

$$\xi_a \approx \frac{1}{2}\Delta_\ell^2\sigma^2 \left|\mathrm{Tr}(\boldsymbol{H}_\ell^a) - \mathrm{Tr}(\boldsymbol{H}_\ell)\right|,$$

privacy parameters — local curvature of the losses



- Groups with larger Hessian traces may have larger ER compared to those with smaller Hessian traces
- **Corollary**: Consider the ERM for a linear model $f_\theta(X) = \theta^\top X$ with ℓ_2 loss. Then output perturbation cannot guarantee pure fairness
- The Hessian of the ℓ_2 loss for a group depends solely on the input norms of the elements in D_a

$$\mathrm{Tr}(\mathbf{H}_\ell^a) = \mathbb{E}_{X\sim D_a}\mathrm{Tr}(XX^\top) = \mathbb{E}_{X\sim D_a}\|X\|^2$$

Fioretto al. NeurIPS:2021

Properties of the training data

Input norms

- Adds Gaussian noise $\mathcal{N}\left(0, \Delta_\ell^2\, \sigma^2\right)$ to the optimal model parameters θ^*
- **Thm:** For ℓ twice differentiable convex functions, the excessive risk gap is approximated as

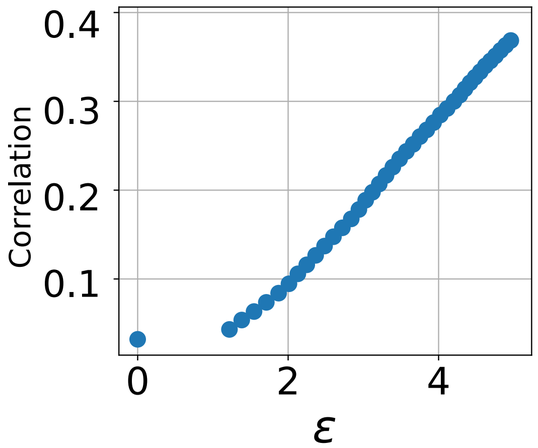
$$\xi_a \approx \frac{1}{2}\Delta_\ell^2\sigma^2 \left|\mathrm{Tr}(\boldsymbol{H}_\ell^a) - \mathrm{Tr}(\boldsymbol{H}_\ell)\right|,$$

privacy parameters — local curvature of the losses

- Groups with large input norms (often observed at the tail of the data distribution) may lead to large Hessian loss values.

- **Corollary**: For groups *a* and *b*, if their average group norms $\mathbb{E}_{X_A \sim D_a}\|X_a\| = \mathbb{E}_{X_b \sim D_b}\|X_b\|$ have identical values, then output perturbation with ℓ_2 loss function achieves pure fairness.

correlation between input norms and excessive risk

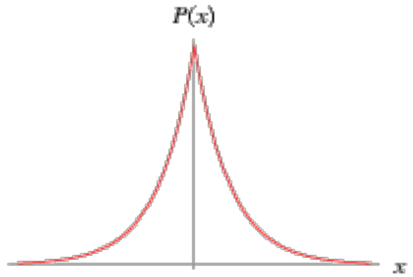


generated using a large non-linear (deep net) model!

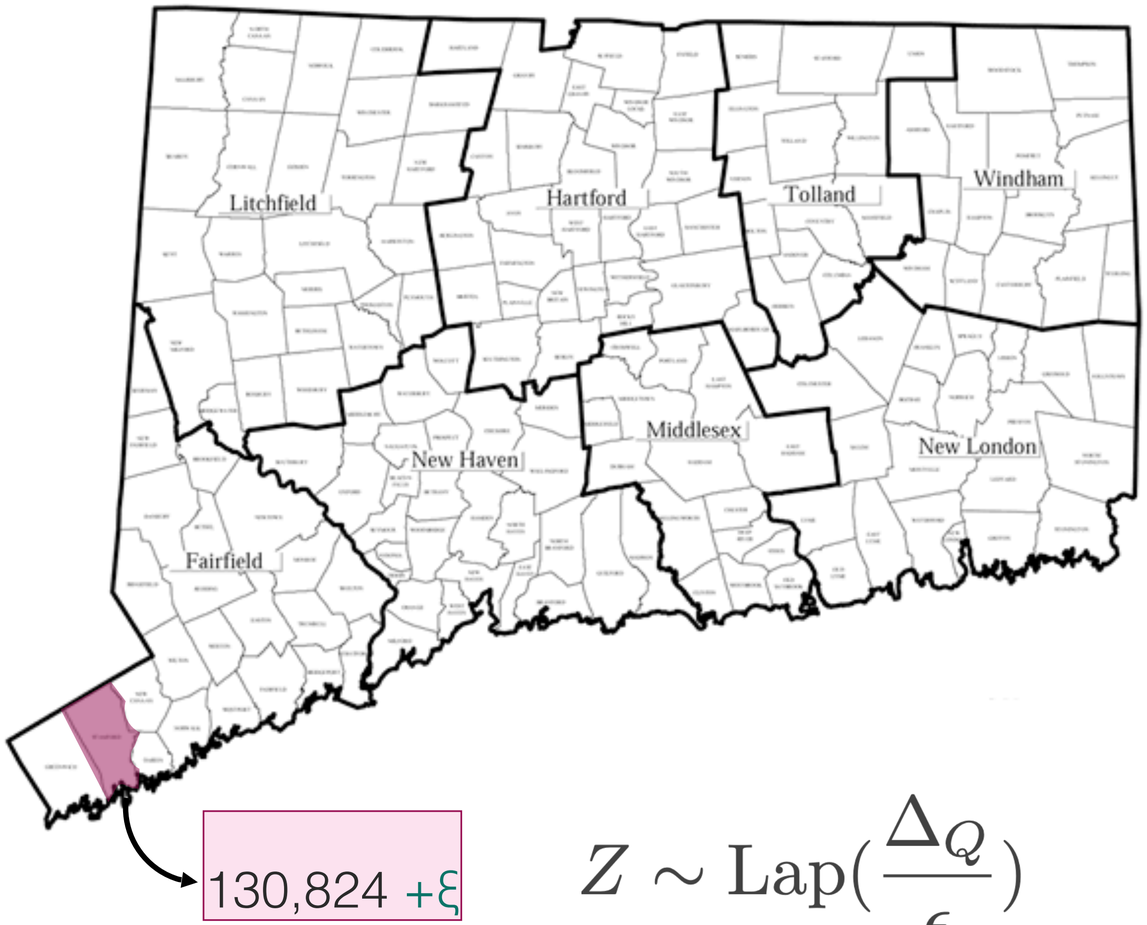
The Census Data Release Problem

- GOAL: Release socio-demographic feature of the population grouped by:

1. Census blocks
2. Counties
3. States
4. National level



PDF $f(x \mid \mu = 0, b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$

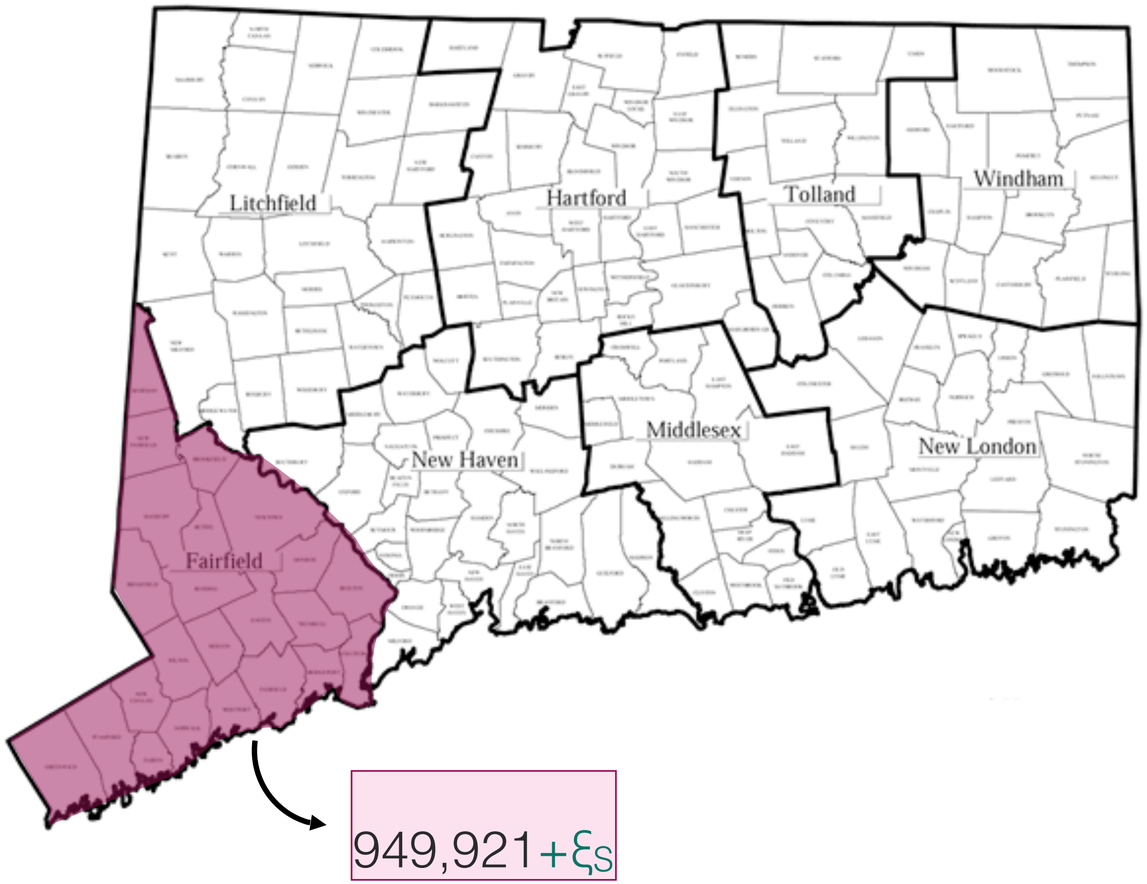


$$Z \sim \text{Lap}\left(\frac{\Delta_Q}{\epsilon}\right)$$

The Census Data Release Problem

- GOAL: Release socio-demographic feature of the population grouped by:

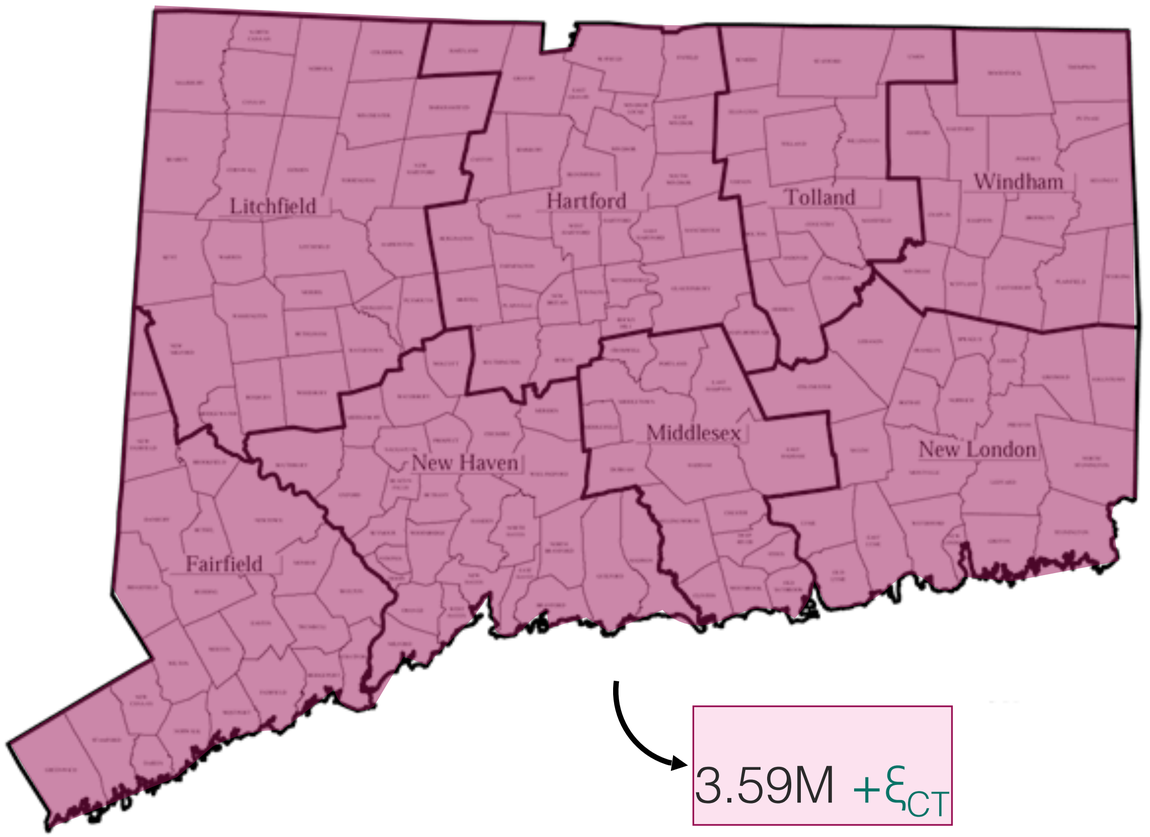
1. Census blocks
2. Counties
3. States
4. National level



The Census Data Release Problem

- GOAL: Release socio-demographic feature of the population grouped by:

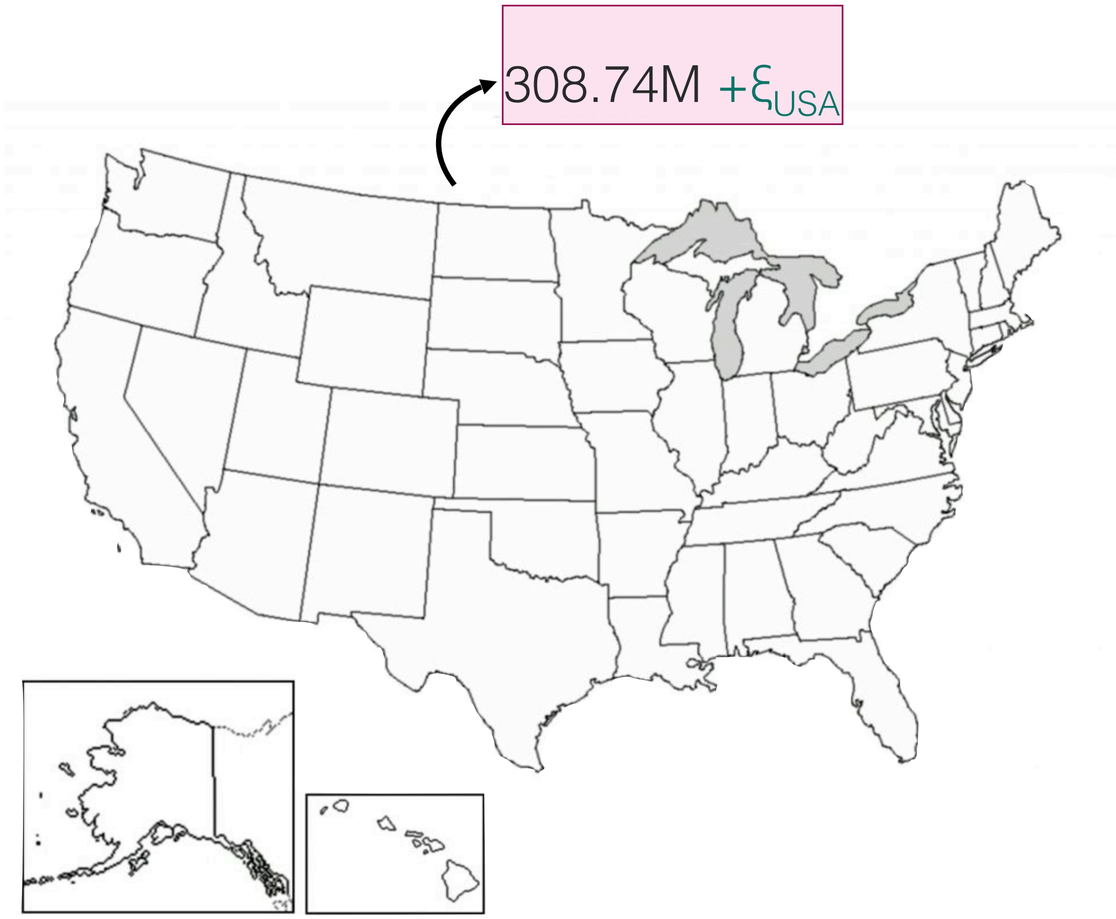
1. Census blocks
2. Counties
3. States
4. National level



The Census Data Release Problem

- GOAL: Release socio-demographic feature of the population grouped by:

1. Census blocks
2. Counties
3. States
4. National level



The consistency issue

- **Requirements:**
  1. **Privacy**
  2. Hierarchical **Consistency**
  3. **Validity**: The private values are non-negative
- Noise is applied independently to each estimate
- The noisy quantities at a "level" (e.g., state) are inconsistent with the sum of the noisy quantities at the "children levels" (e.g., counties of that state)