38th Conference on Uncertainty in Artificial Intelligence
Eindhoven, Netherlands
August 1-5, 2022
uai2022

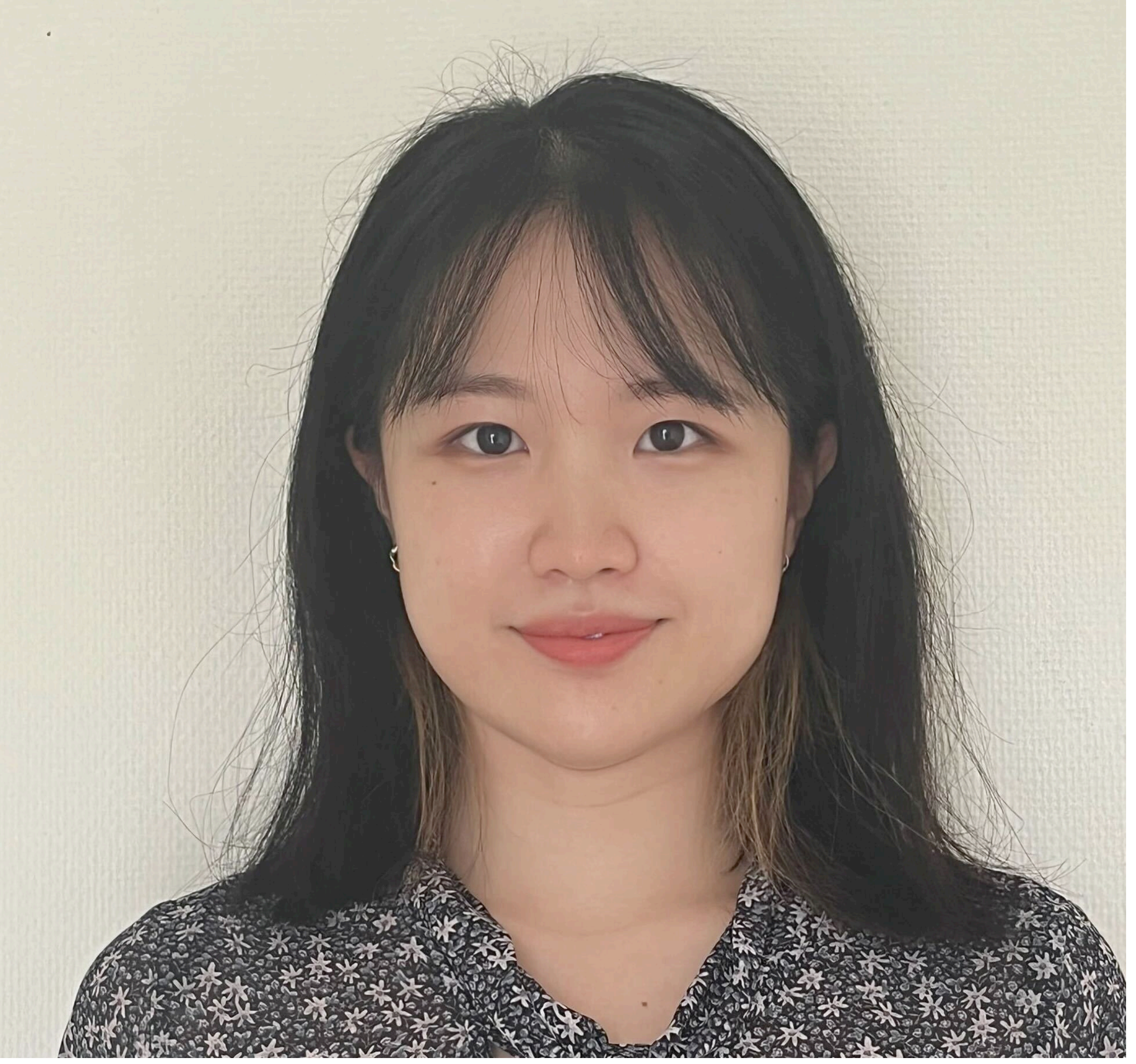
ETH zürich

ETH AI CENTER

How unfair is private learning ?

Amartya Sanyal, Yaxi Hu, Fanny Yang

Amartya



Yaxi

Fanny

Privacy and Fairness

Privacy and Fairness

Privacy and Fairness are both desirable properties in machine learning applications.

Privacy and Fairness

Privacy and Fairness are both desirable properties in machine learning applications.

Prior Work has mostly looked at the intersection:

Privacy and Fairness

Privacy and Fairness are both desirable properties in machine learning applications.

Prior Work has mostly looked at the intersection:

Privacy and Accuracy: Kasiviswanathan et al. 2008, Feldman and Xiao 2014, Alon et. al., 2022.

Privacy and Fairness

Privacy and Fairness are both desirable properties in machine learning applications.

Prior Work has mostly looked at the intersection:

Privacy and Accuracy: Kasiviswanathan et al. 2008, Feldman and Xiao 2014, Alon et. al., 2022.

Fairness and Accuracy: Sagawa et. al. 2019, Du et al. 2021, Goel et. al. 2021.

Privacy and Fairness

Privacy and Fairness are both desirable properties in machine learning applications.

Prior Work has mostly looked at the intersection:

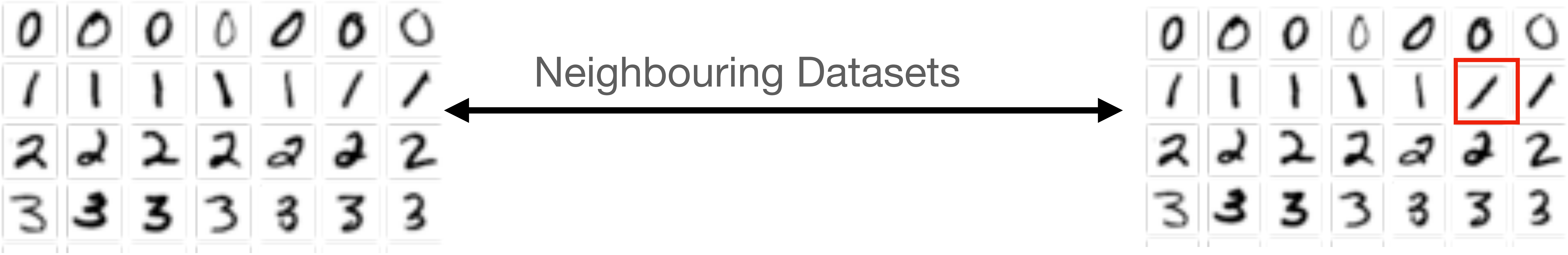
Privacy and Accuracy: Kasiviswanathan et al. 2008, Feldman and Xiao 2014, Alon et. al., 2022.

Fairness and Accuracy: Sagawa et. al. 2019, Du et al. 2021, Goel et. al. 2021.

THIS WORK: The interaction of Privacy and Fairness of nearly **accurate algorithms.**

Differential Privacy

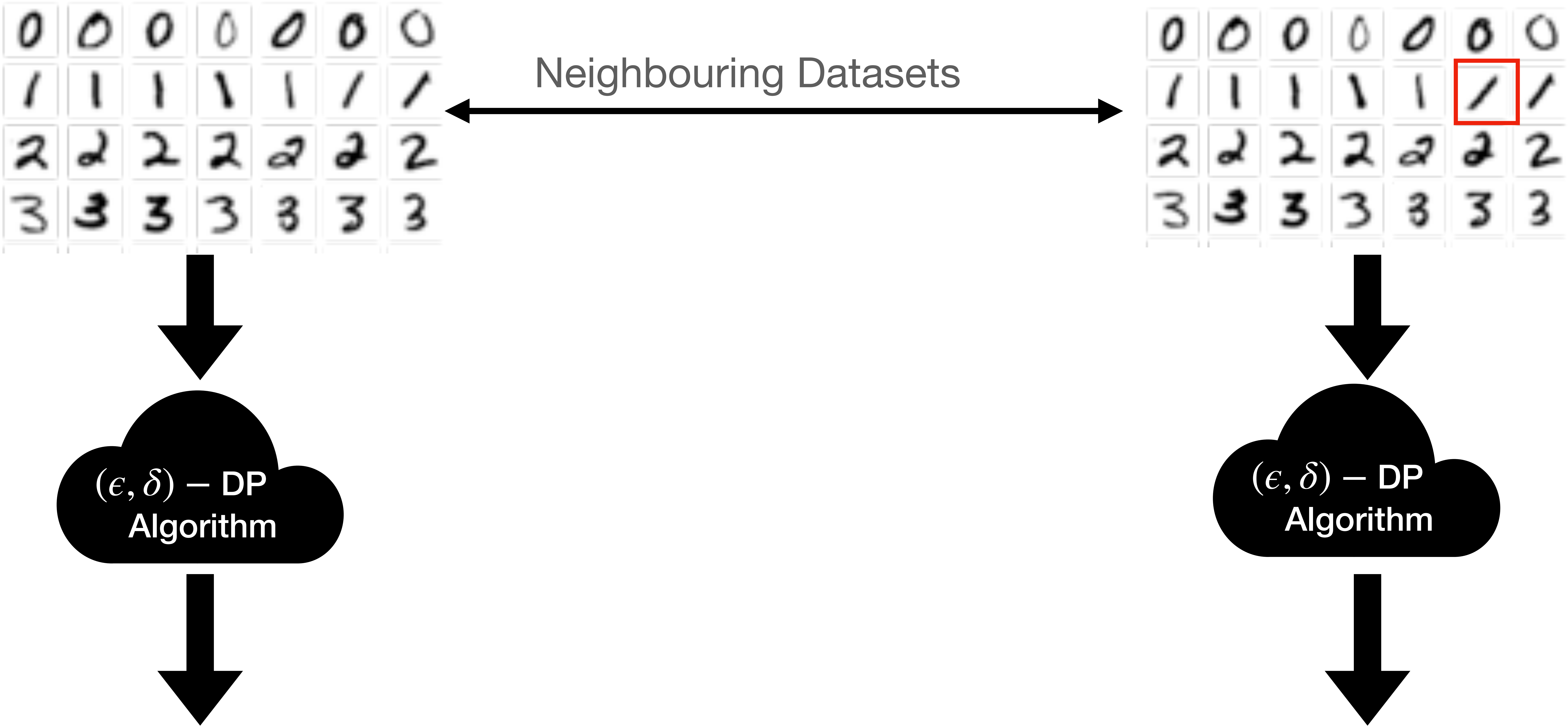
Differential Privacy



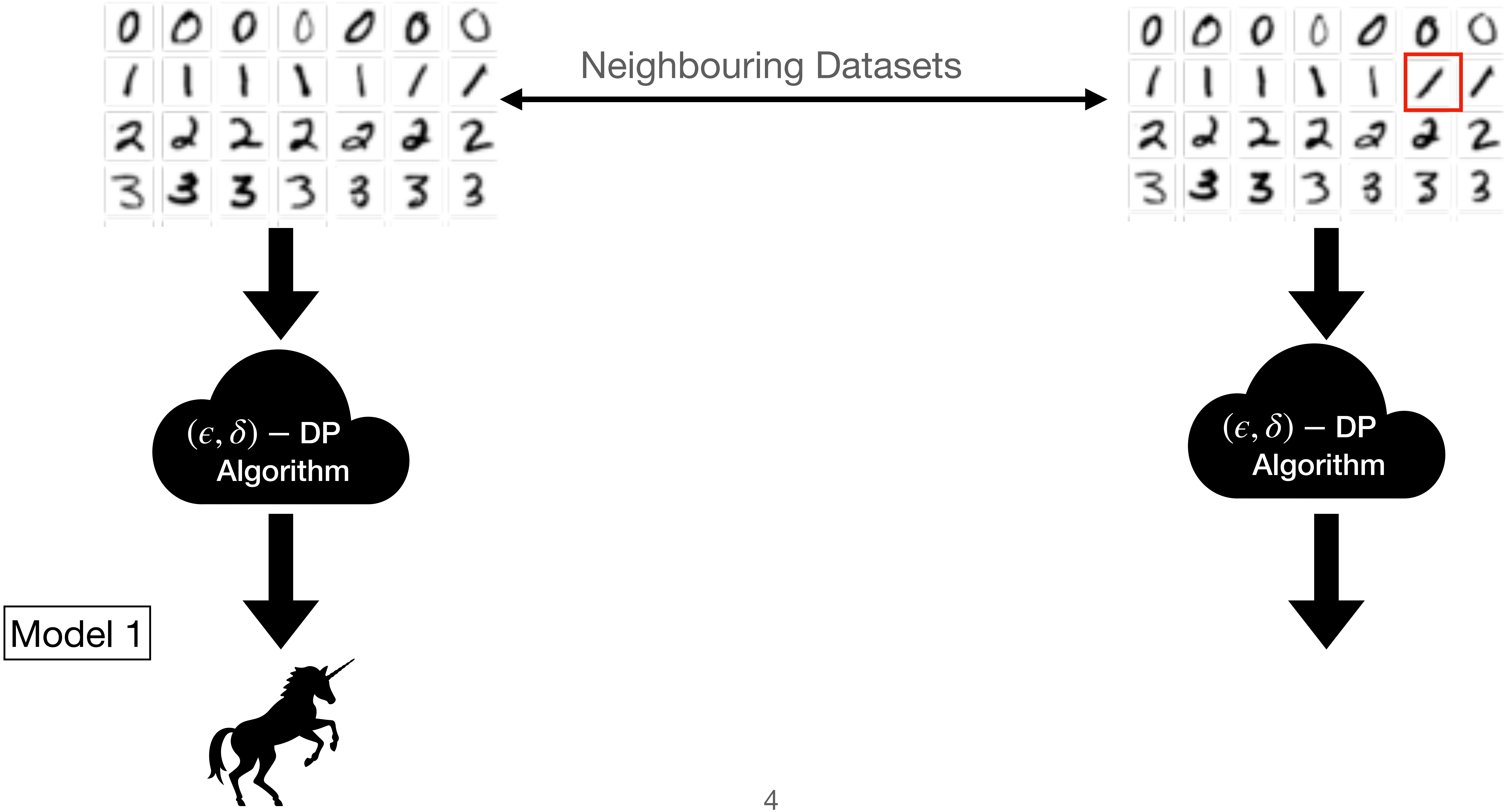
Differential Privacy

Differential Privacy

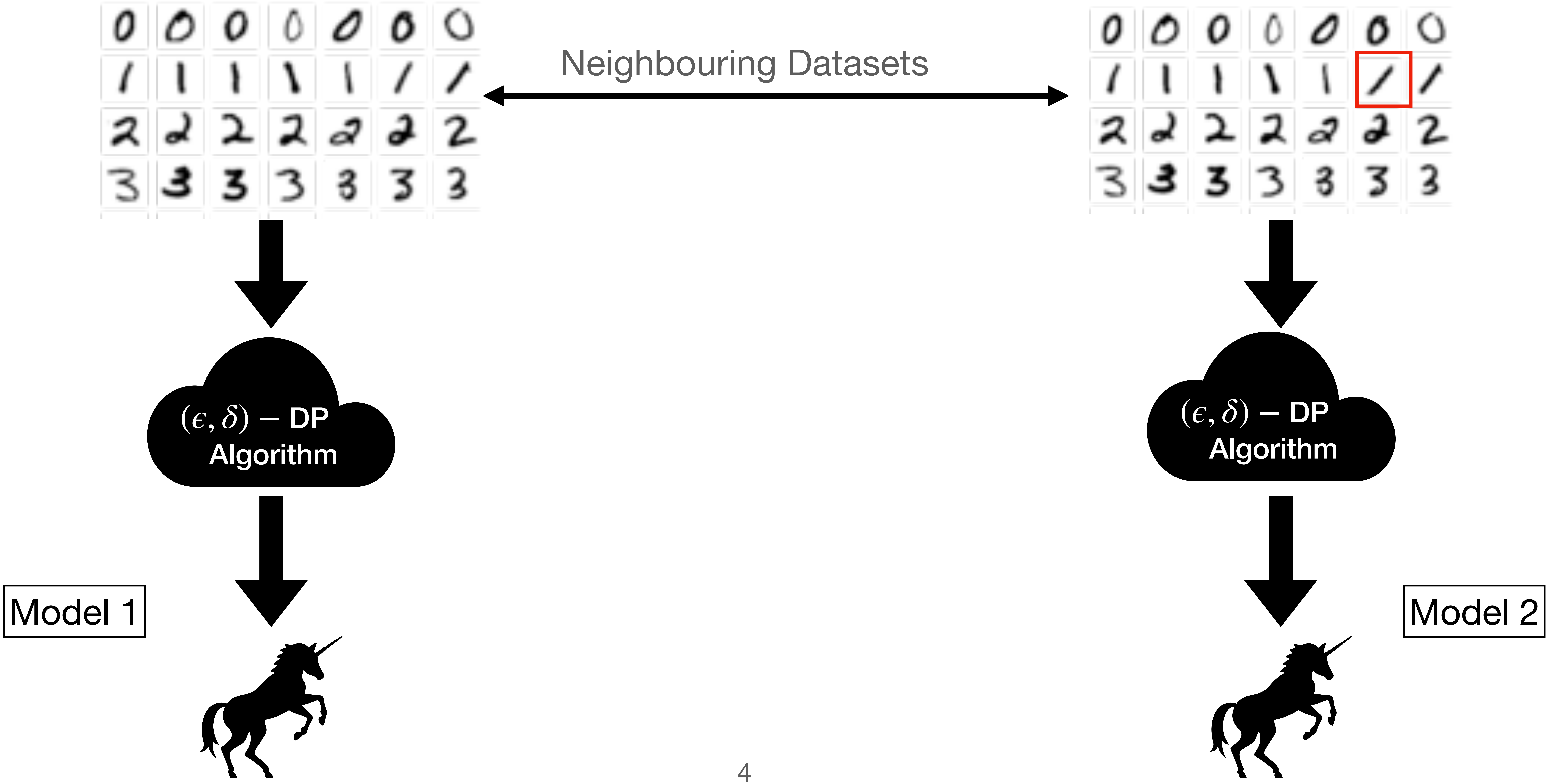
Differential Privacy



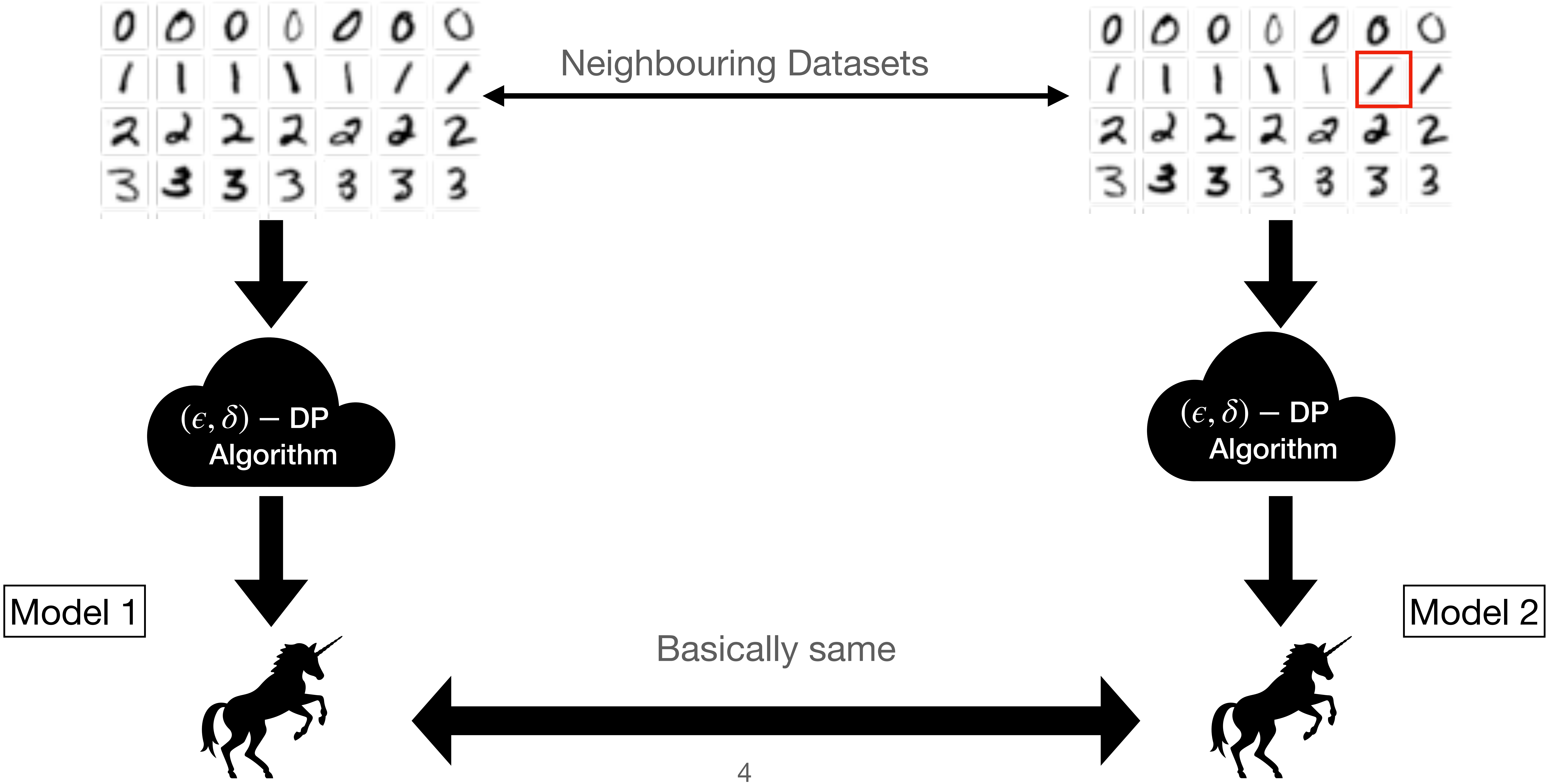
Differential Privacy



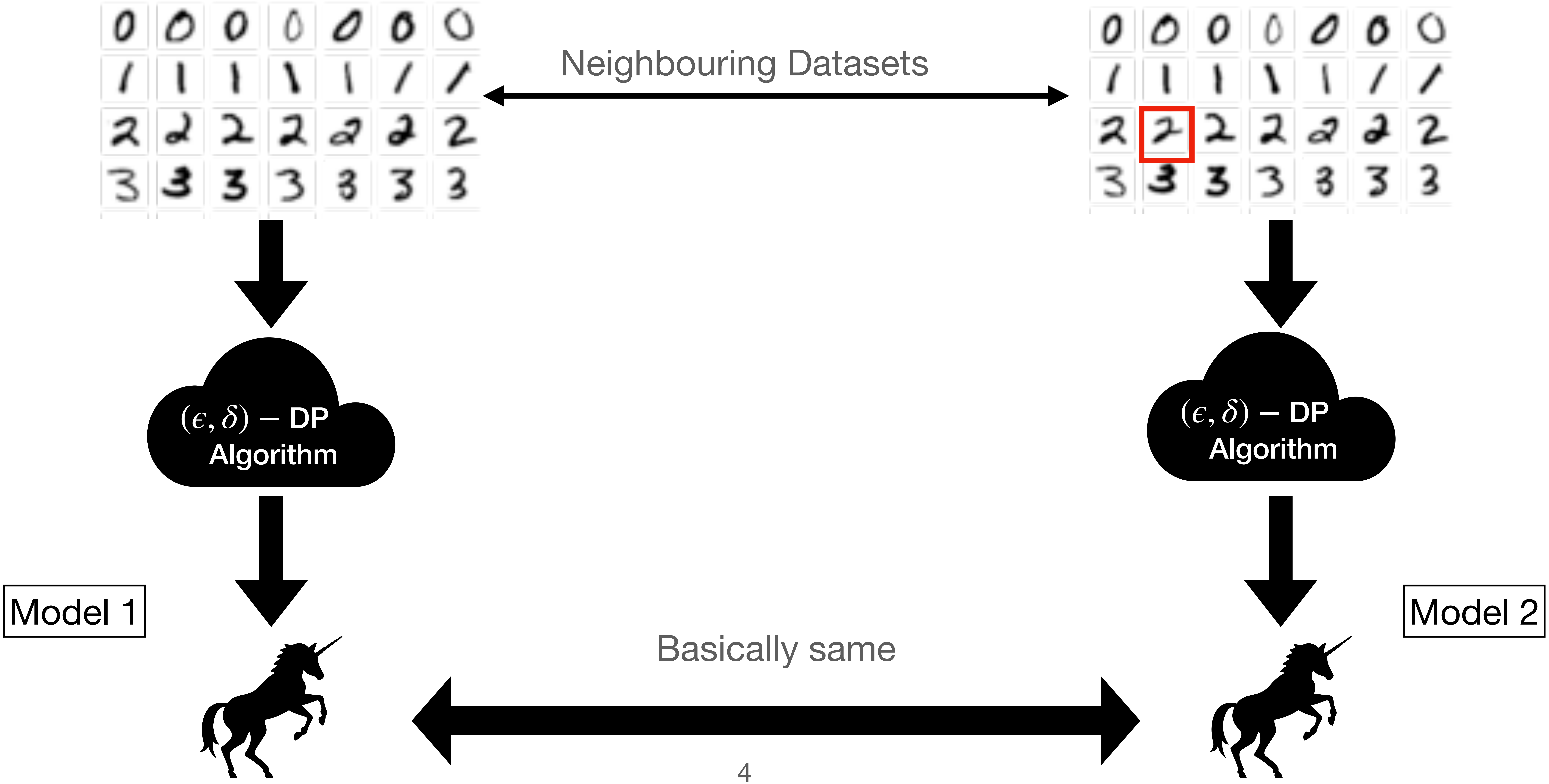
Differential Privacy



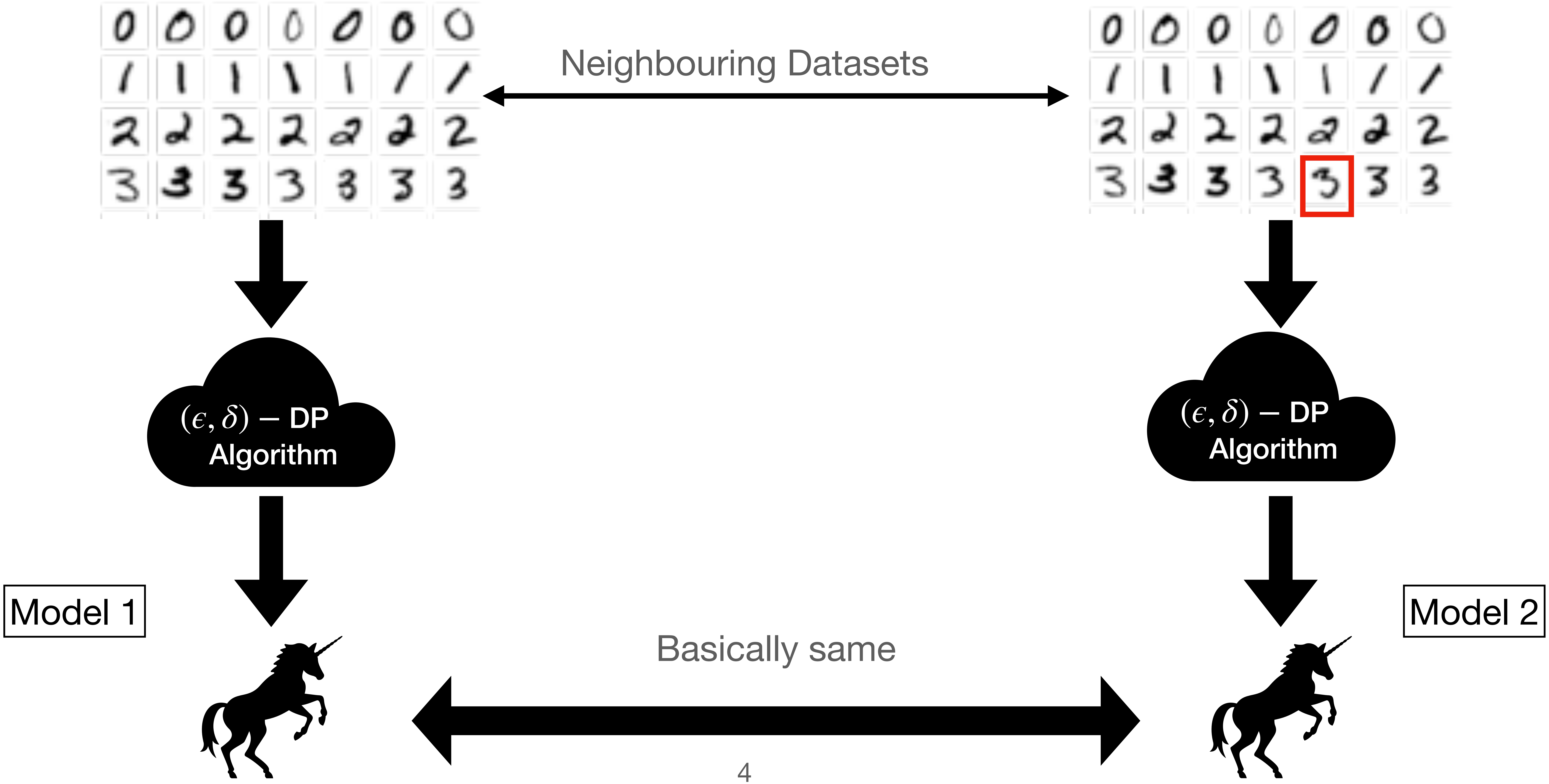
Differential Privacy



Differential Privacy



Differential Privacy



(Un) Fairness (Accuracy Discrepancy)

(Un) Fairness (Accuracy Discrepancy)

Genre

Thrillers

Superhero

B&W

Mimes

Silent

Puppet

Ostern

(Un) Fairness (Accuracy Discrepancy)

Genre	Thrillers	Superhero	B&W	Mimes	Silent	Puppet	Ostern
Proportion	40%	40%	4%	4%	4%	4%	4%

(Un) Fairness (Accuracy Discrepancy)

Minority subpopulations

Genre	Thrillers	Superhero	B&W	Mimes	Silent	Puppet	Ostern
Proportion	40%	40%	4%	4%	4%	4%	4%

(Un) Fairness (Accuracy Discrepancy)

	Majority subpopulations		Minority subpopulations				
Genre	Thrillers	Superhero	B&W	Mimes	Silent	Puppet	Ostern
Proportion	40%	40%	4%	4%	4%	4%	4%

(Un) Fairness (Accuracy Discrepancy)

ML Problem: Is the movie safe to watch for kids ?

	Majority subpopulations		Minority subpopulations				
Genre	Thrillers	Superhero	B&W	Mimes	Silent	Puppet	Ostern
Proportion	40%	40%	4%	4%	4%	4%	4%

(Un) Fairness (Accuracy Discrepancy)

ML Problem: Is the movie safe to watch for kids ?

	Majority subpopulations		Minority subpopulations				
Genre	Thrillers	Superhero	B&W	Mimes	Silent	Puppet	Ostern
Proportion	40%	40%	4%	4%	4%	4%	4%
Error	5%	5%	65%	75%	80%	80%	50%

(Un) Fairness (Accuracy Discrepancy)

ML Problem: Is the movie safe to watch for kids ?

	Majority subpopulations		Minority subpopulations				
Genre	Thrillers	Superhero	B&W	Mimes	Silent	Puppet	Ostern
Proportion	40%	40%	4%	4%	4%	4%	4%
Error	5%	5%	65%	75%	80%	80%	50%
	Majority Error = 5%		Minority Error = 70%				

(Un) Fairness (Accuracy Discrepancy)

ML Problem: Is the movie safe to watch for kids ?

	Majority subpopulations		Minority subpopulations				
Genre	Thrillers	Superhero	B&W	Mimes	Silent	Puppet	Ostern
Proportion	40%	40%	4%	4%	4%	4%	4%
Error	5%	5%	65%	75%	80%	80%	50%
	Majority Error = 5%		Minority Error = 70%				

Total Error = 18%

(Un) Fairness (Accuracy Discrepancy)

ML Problem: Is the movie safe to watch for kids ?

	Majority subpopulations		Minority subpopulations				
Genre	Thrillers	Superhero	B&W	Mimes	Silent	Puppet	Ostern
Proportion	40%	40%	4%	4%	4%	4%	4%
Error	5%	5%	65%	75%	80%	80%	50%
	Majority Error = 5%		Minority Error = 70%				

Total Error = 18%

Accuracy Discrepancy = Minority Error - Total Error

(Un) Fairness (Accuracy Discrepancy)

ML Problem: Is the movie safe to watch for kids ?

	Majority subpopulations		Minority subpopulations				
Genre	Thrillers	Superhero	B&W	Mimes	Silent	Puppet	Ostern
Proportion	40%	40%	4%	4%	4%	4%	4%
Error	5%	5%	65%	75%	80%	80%	50%
	Majority Error = 5%		Minority Error = 70%				

Total Error = 18%

Accuracy Discrepancy = 70 - 18 = 52%

Example dataset

CelebA

Example dataset

CelebA

Example dataset

CelebA

40 binary attributes with each image

Eyeglasses

Bangs

Pointy Noise

Example dataset

CelebA

40 binary attributes -> **2^{40} subpopulations.**

40 binary attributes with each image

Eyeglasses

Bangs

Pointy Noise

Example dataset

CelebA

40 binary attributes -> **2^{40} subpopulations.**

- **Subpopulation 1**: Eyeglasses, bangs, …, pointy nose.

40 binary attributes with each image

Eyeglasses

Bangs

Pointy Noise

Example dataset

CelebA

<u>40 binary</u> <u>attributes -> **2^{40} subpopulations.**</u>

- **Subpopulation 1**: Eyeglasses, bangs, …, pointy nose.
- **Subpopulation 2:** No eyeglasses, bangs,……,pointy noise.

<u>40 binary attributes with each image</u>

Eyeglasses

Bangs

Pointy Noise

Example dataset

CelebA

40 binary attributes with each image

40 binary attributes -> **2^{40} subpopulations.**

- **Subpopulation 1**: Eyeglasses, bangs, …, pointy nose.
- **Subpopulation 2:** No eyeglasses, bangs,……,pointy noise.
- …
- …
- **Subpopulation 2^{40}:** No eyeglasses, no bangs,…, no pointy nose.

Example dataset

CelebA

40 binary attributes -> **2^{40} subpopulations.**

- **Subpopulation 1**: Eyeglasses, bangs, …, pointy nose.
- **Subpopulation 2:** No eyeglasses, bangs,……,pointy noise.
- …
- …
- **Subpopulation 2^{40}:** No eyeglasses, no bangs,…, no pointy nose.

40 binary attributes with each image

Eyeglasses

Bangs

Pointy Noise

Example dataset

CelebA

<u>40 binary attributes with each image</u>

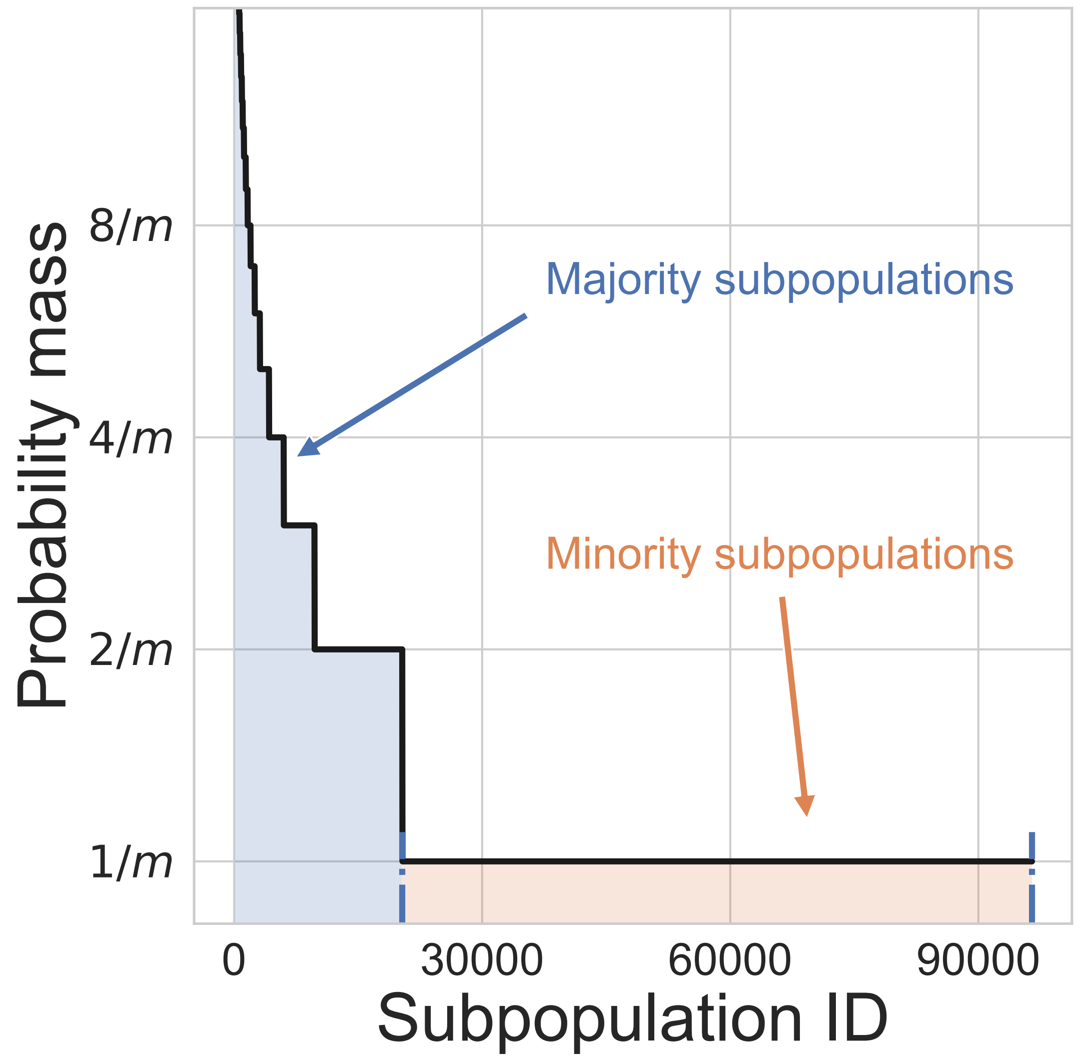
Eyeglasses

Bangs

Pointy Noise

<u>**40 binary** attributes -> **2^{40} subpopulations.**</u>

- **Subpopulation 1**: Eyeglasses, bangs, …, pointy nose.
- **Subpopulation 2:** No eyeglasses, bangs,……,pointy noise.
- …
- …
- **Subpopulation 2^{40}:** No eyeglasses, no bangs,…, no pointy nose.



Example dataset

CelebA

40 binary attributes with each image

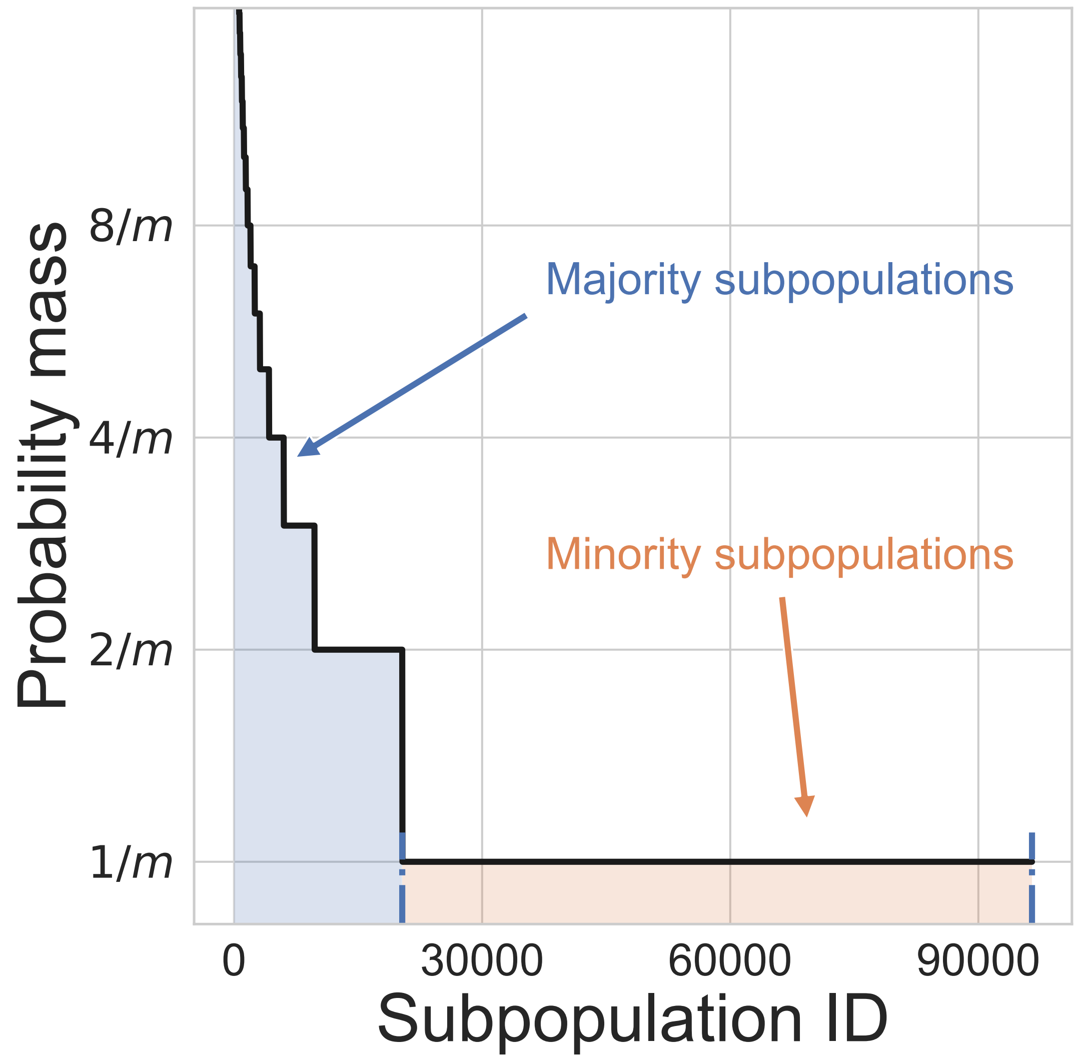
Eyeglasses

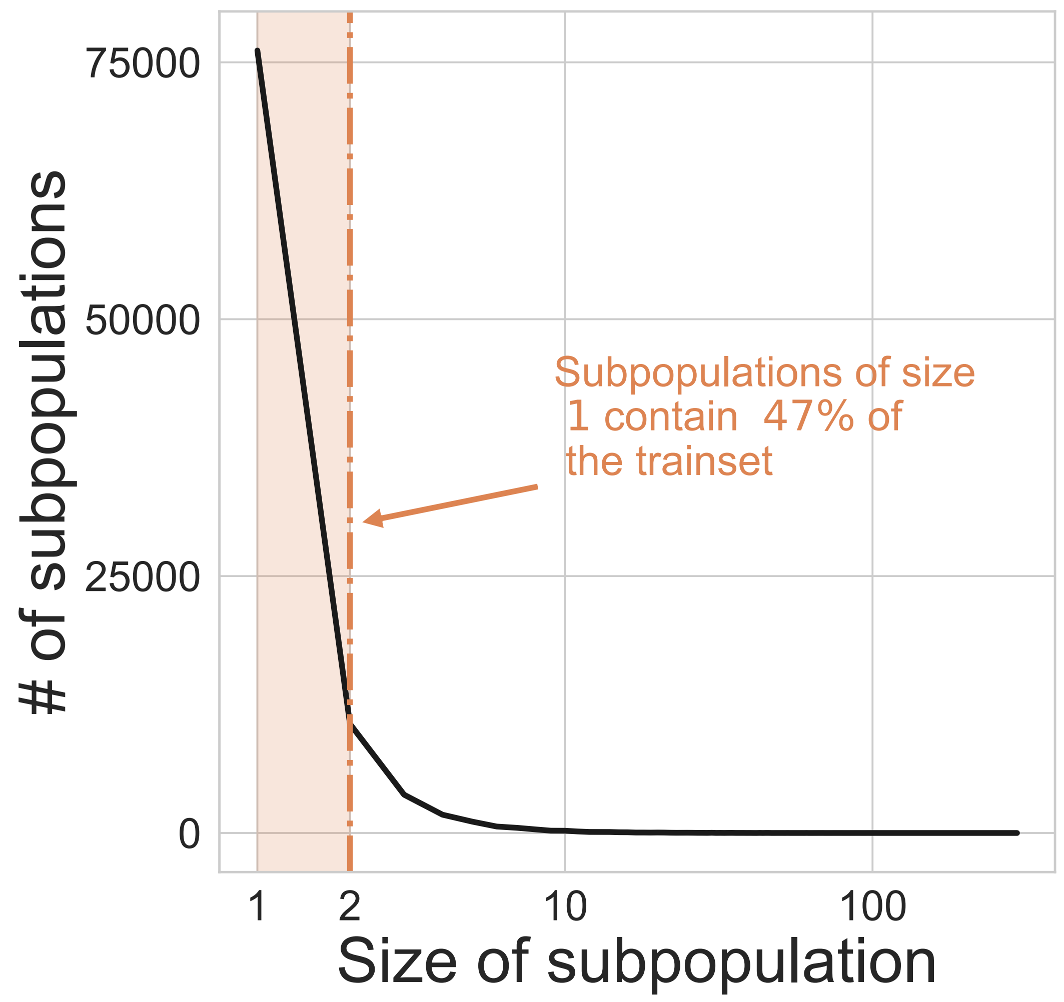
Bangs

Pointy Noise

40 binary attributes -> 2^{40} subpopulations.

- **Subpopulation 1**: Eyeglasses, bangs, …, pointy nose.
- **Subpopulation 2:** No eyeglasses, bangs,……,pointy noise.
- …
- …
- **Subpopulation 2^{40}:** No eyeglasses, no bangs,…, no pointy nose.



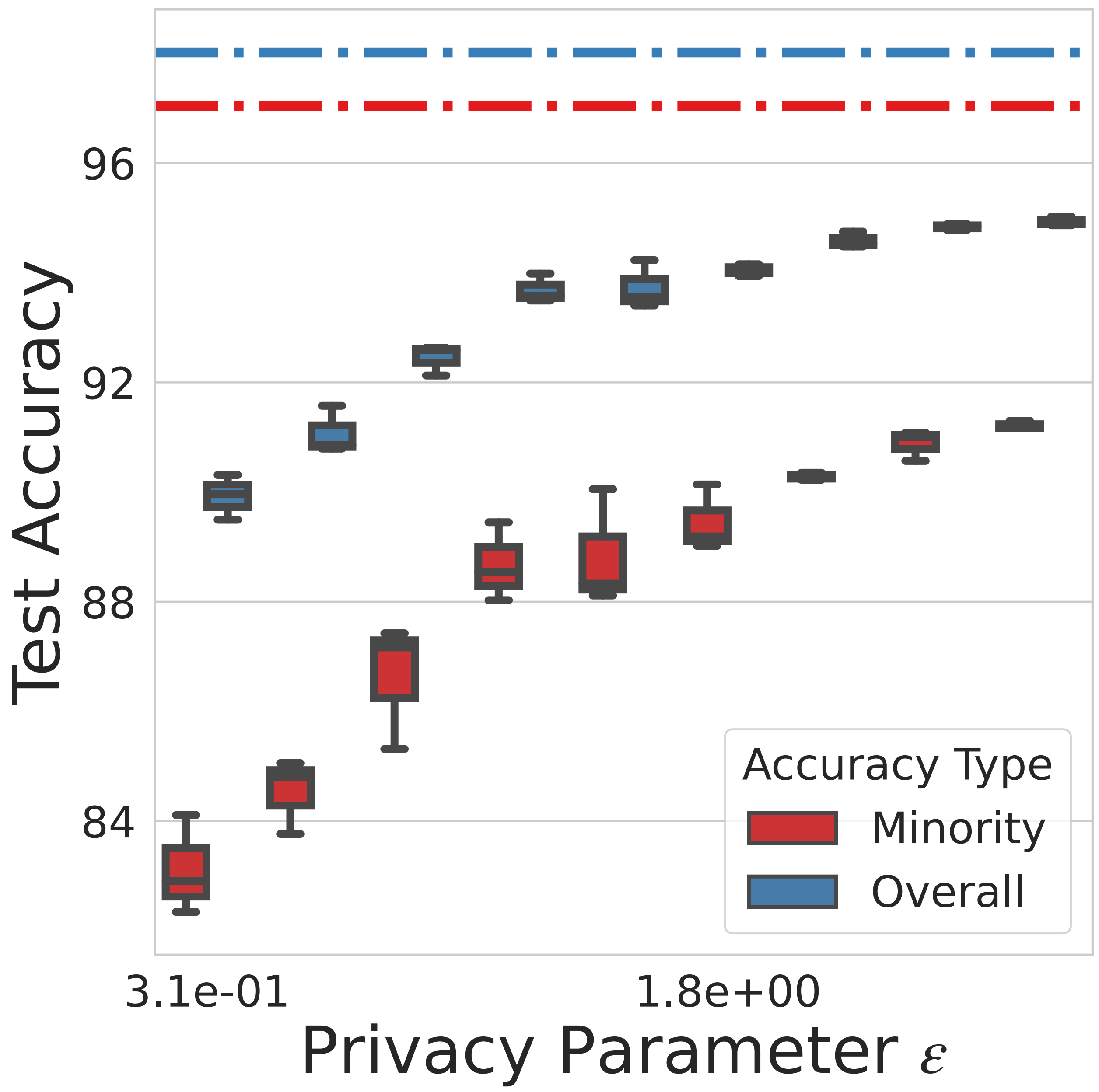


Privacy vs Fairness

CelebA

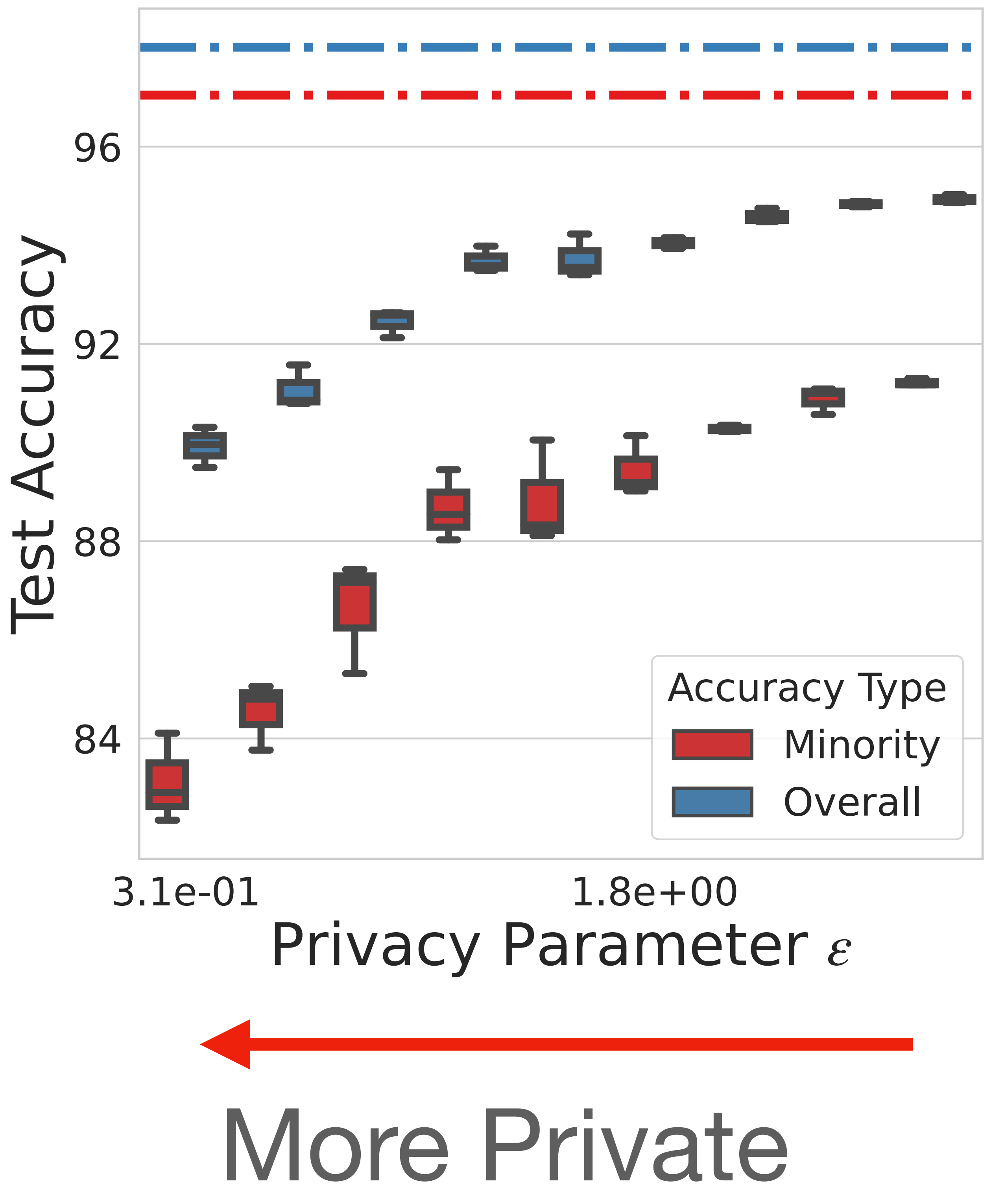
Privacy vs Fairness

CelebA



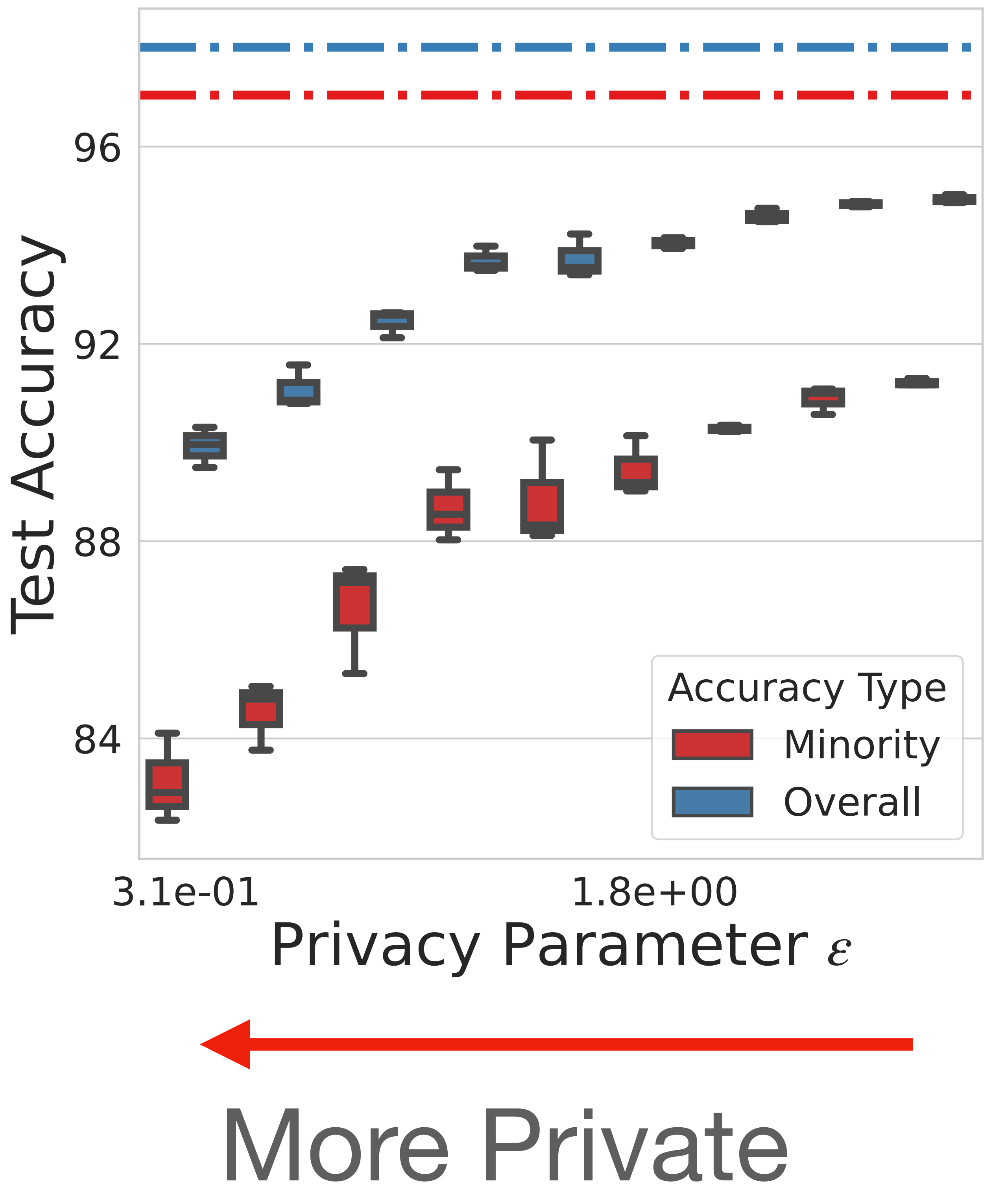
Privacy vs Fairness

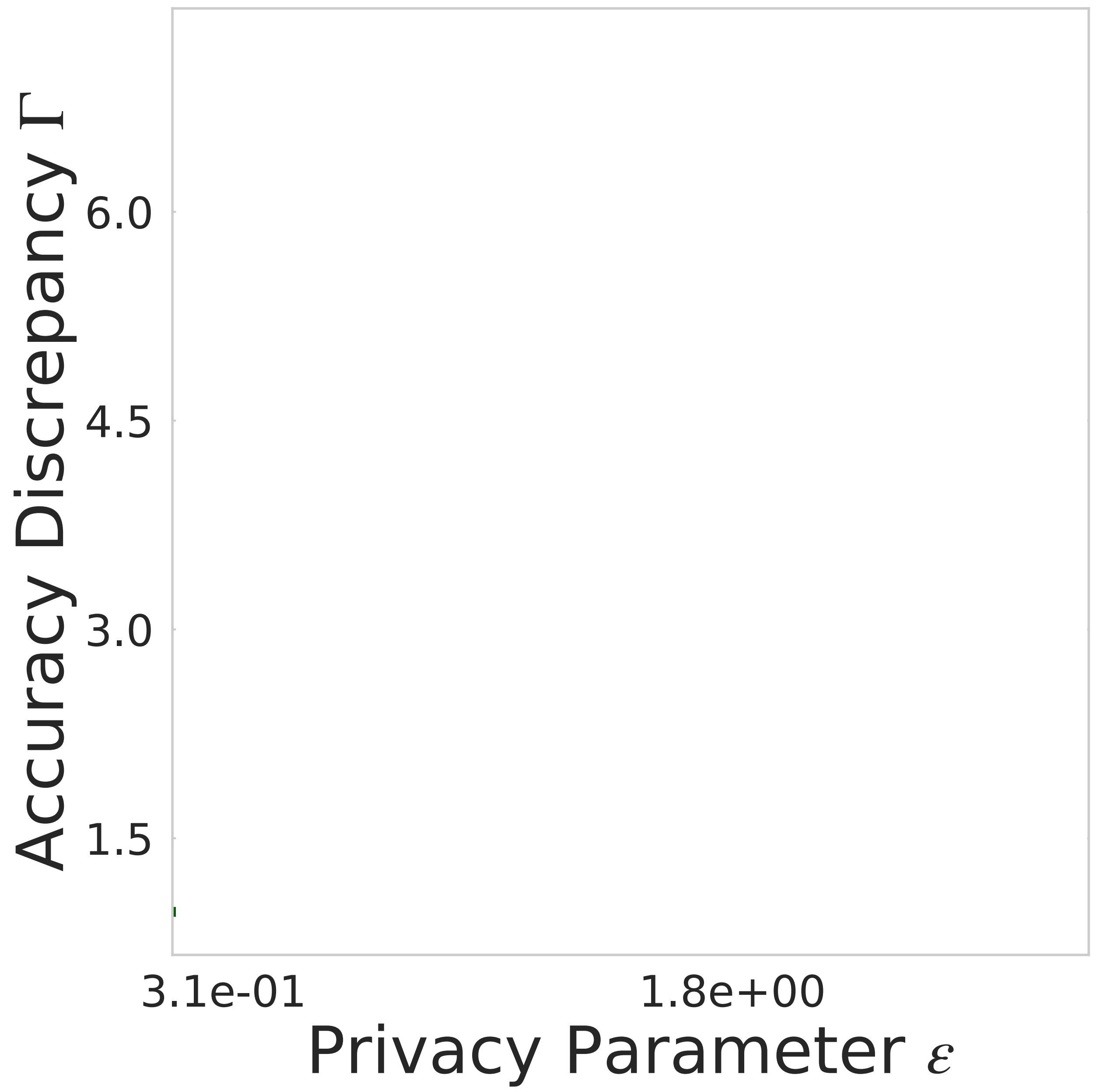
CelebA



Privacy vs Fairness

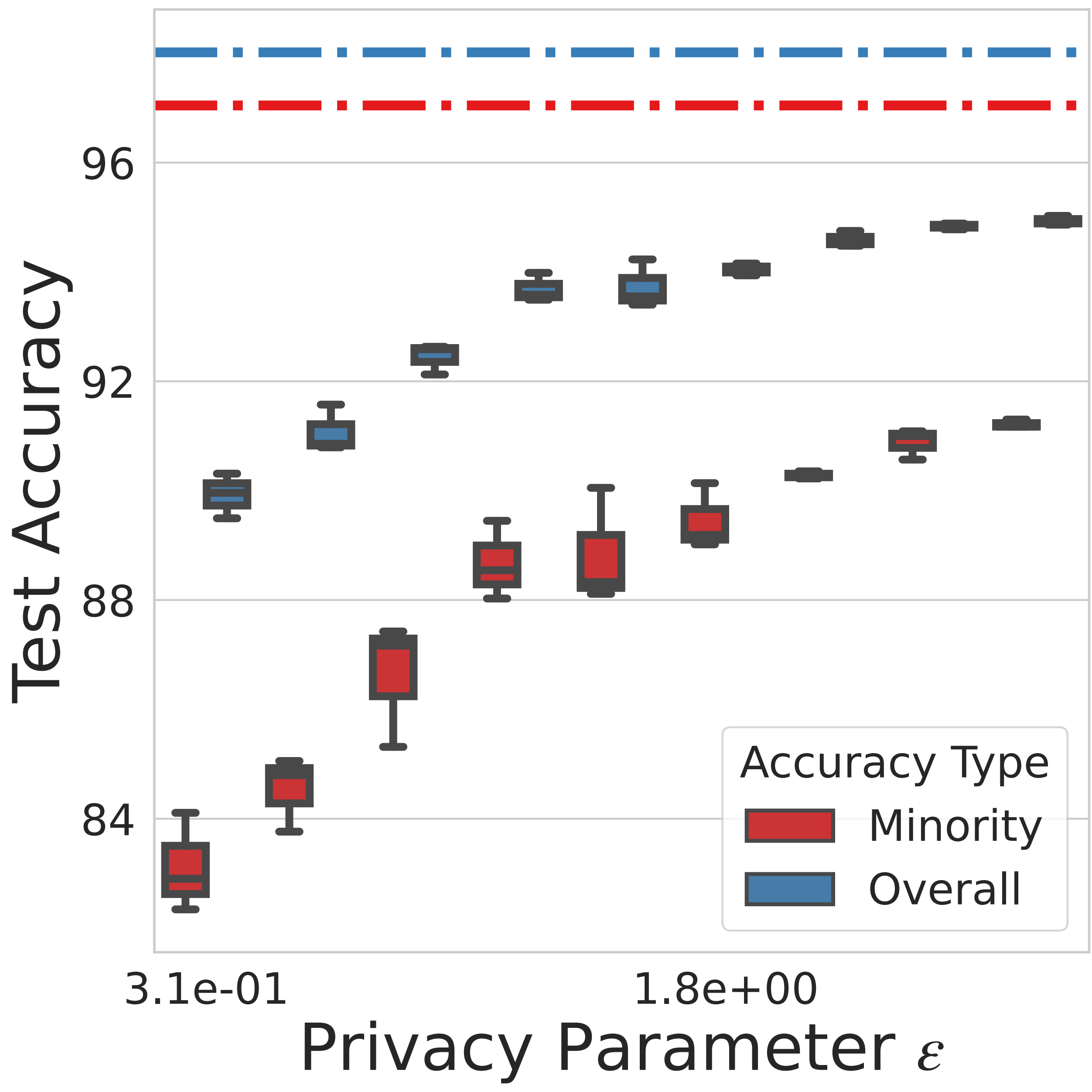
CelebA



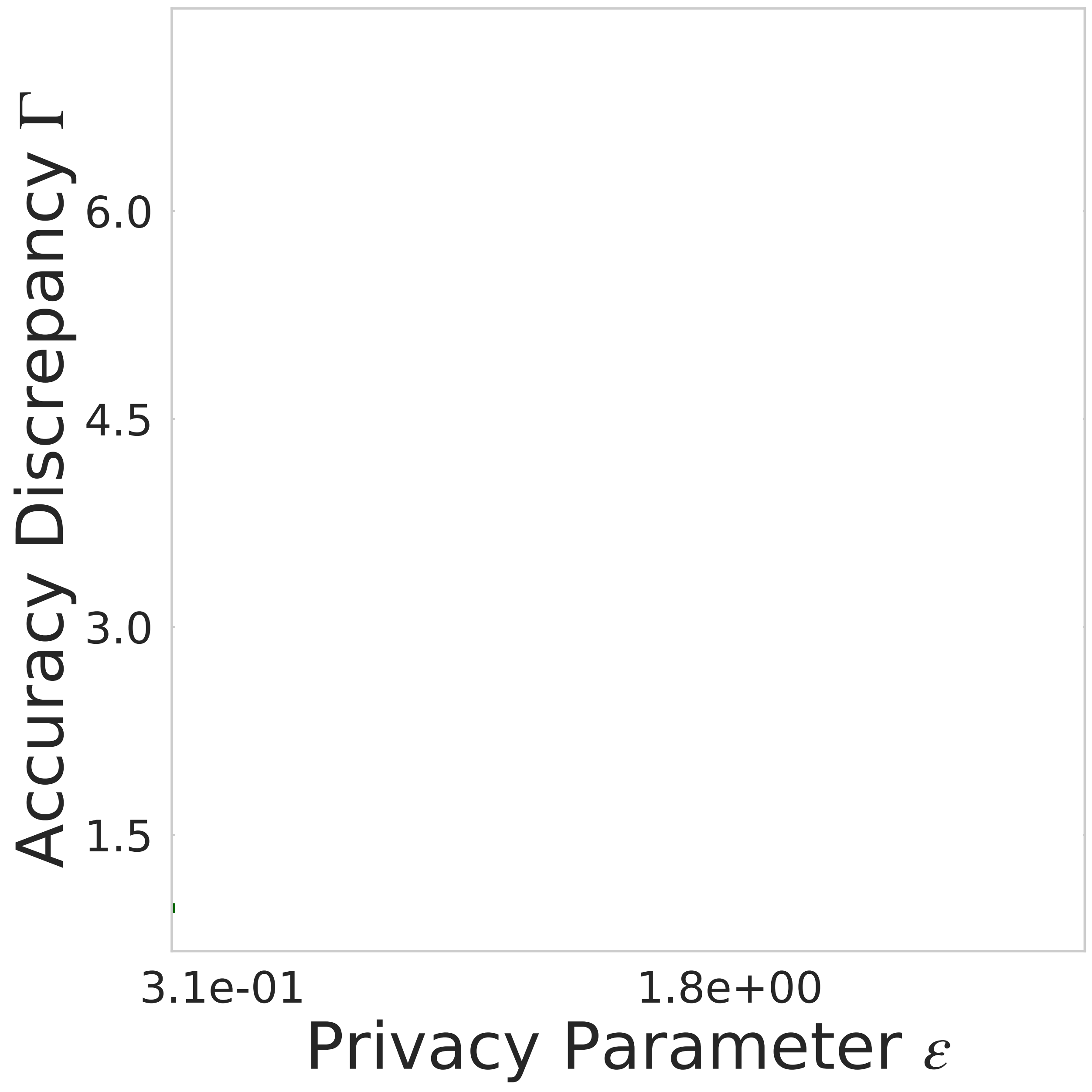


Privacy vs Fairness

CelebA



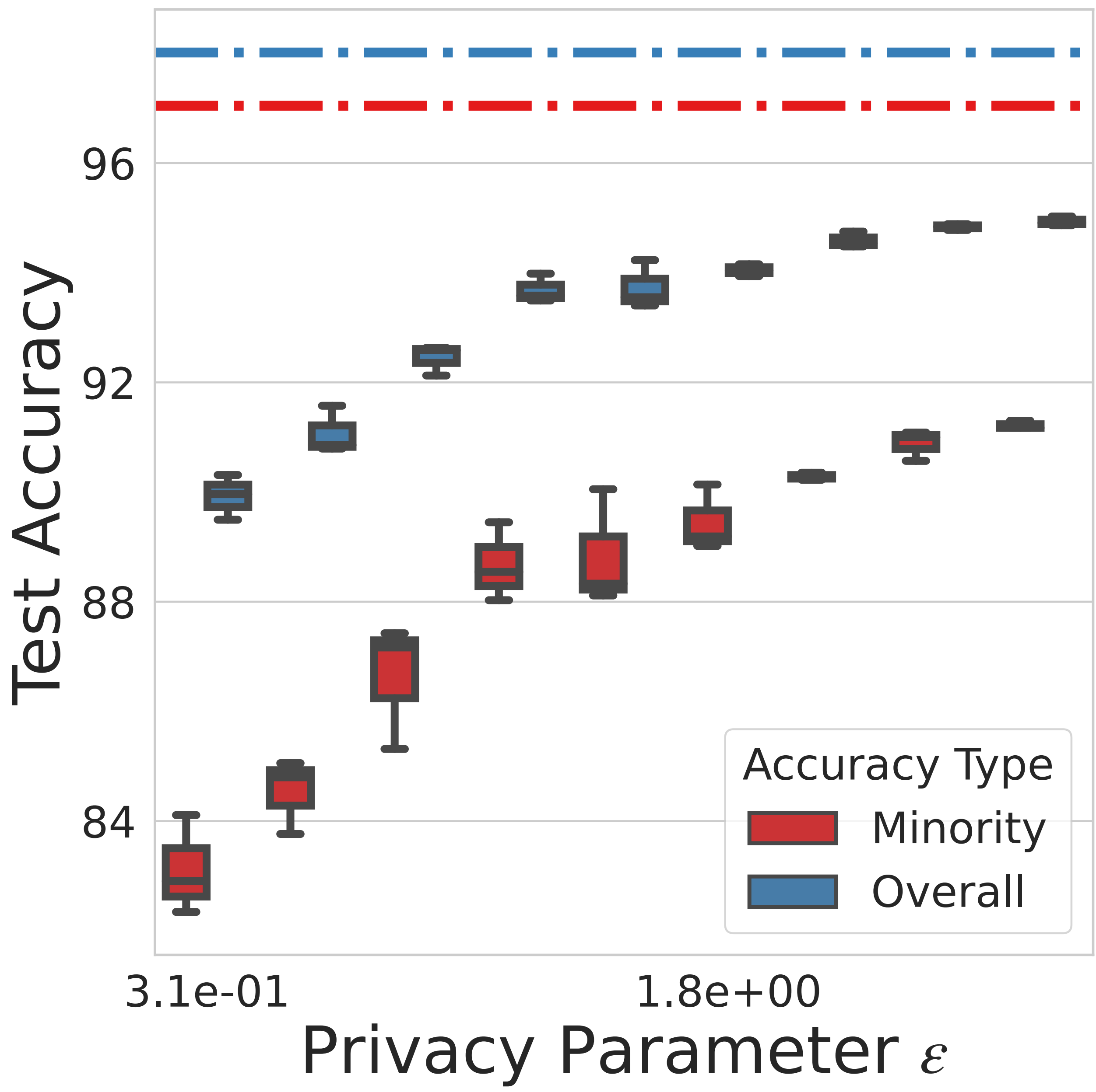
More Private



More Private

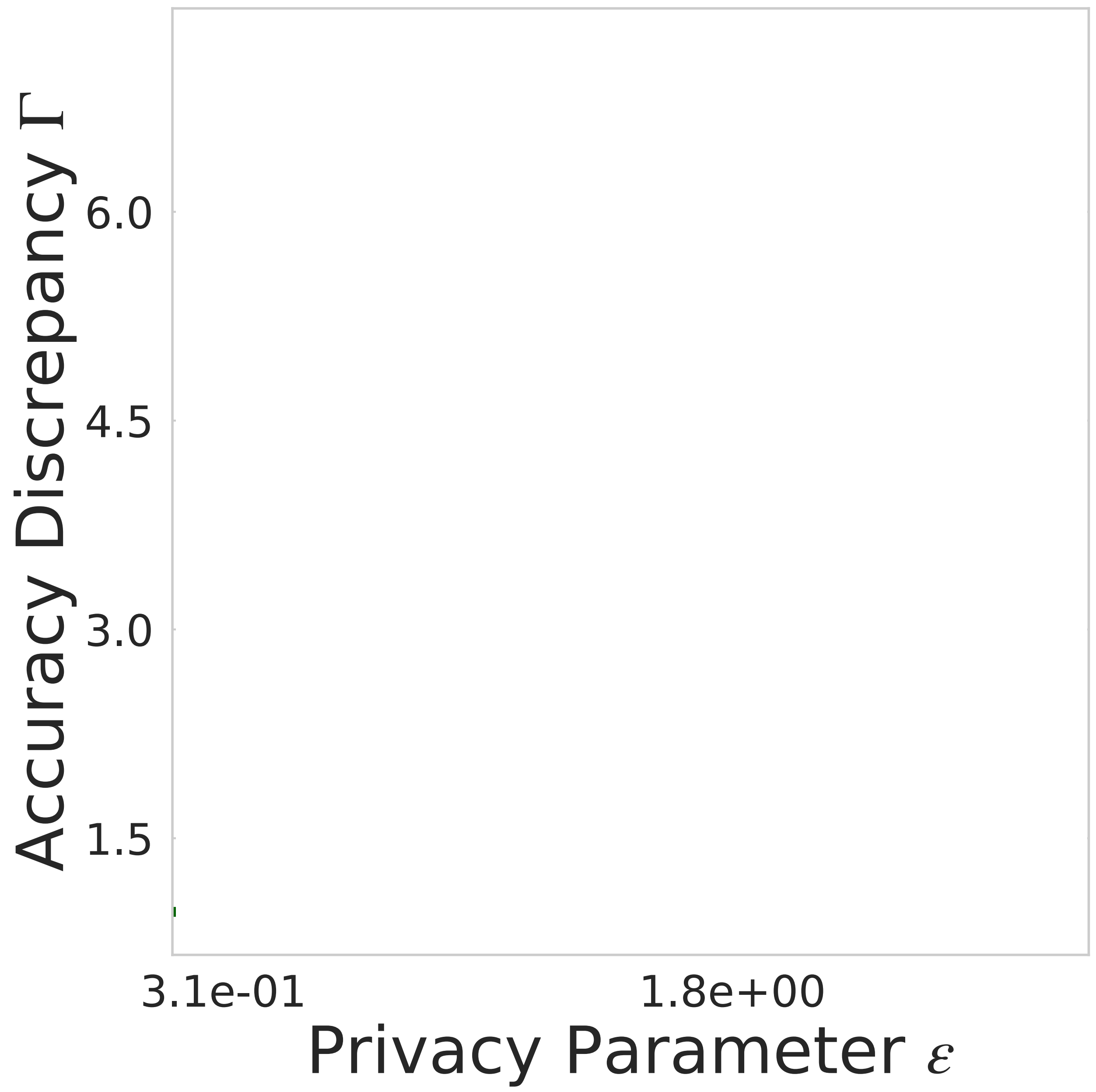
Privacy vs Fairness

CelebA



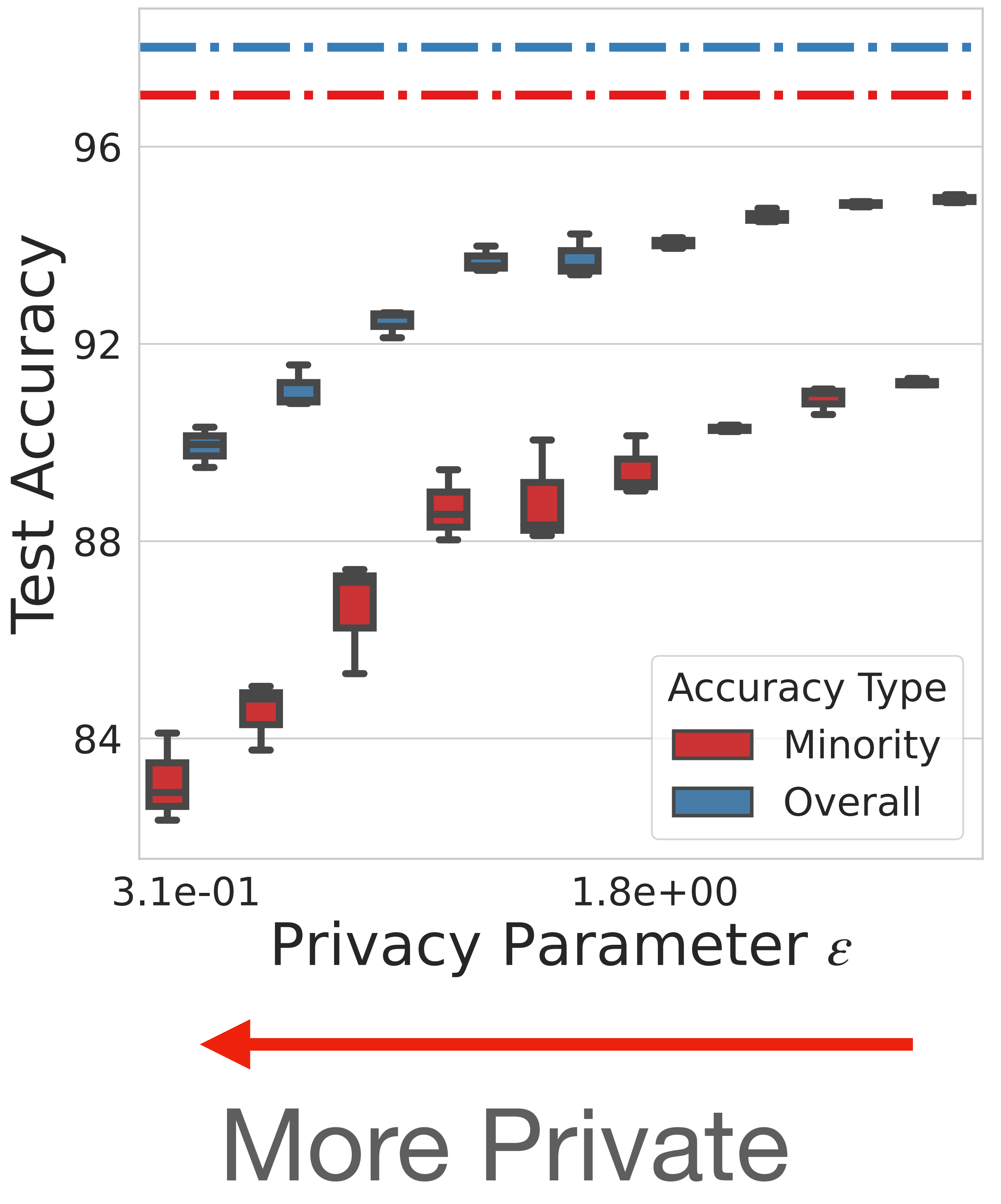
More Private

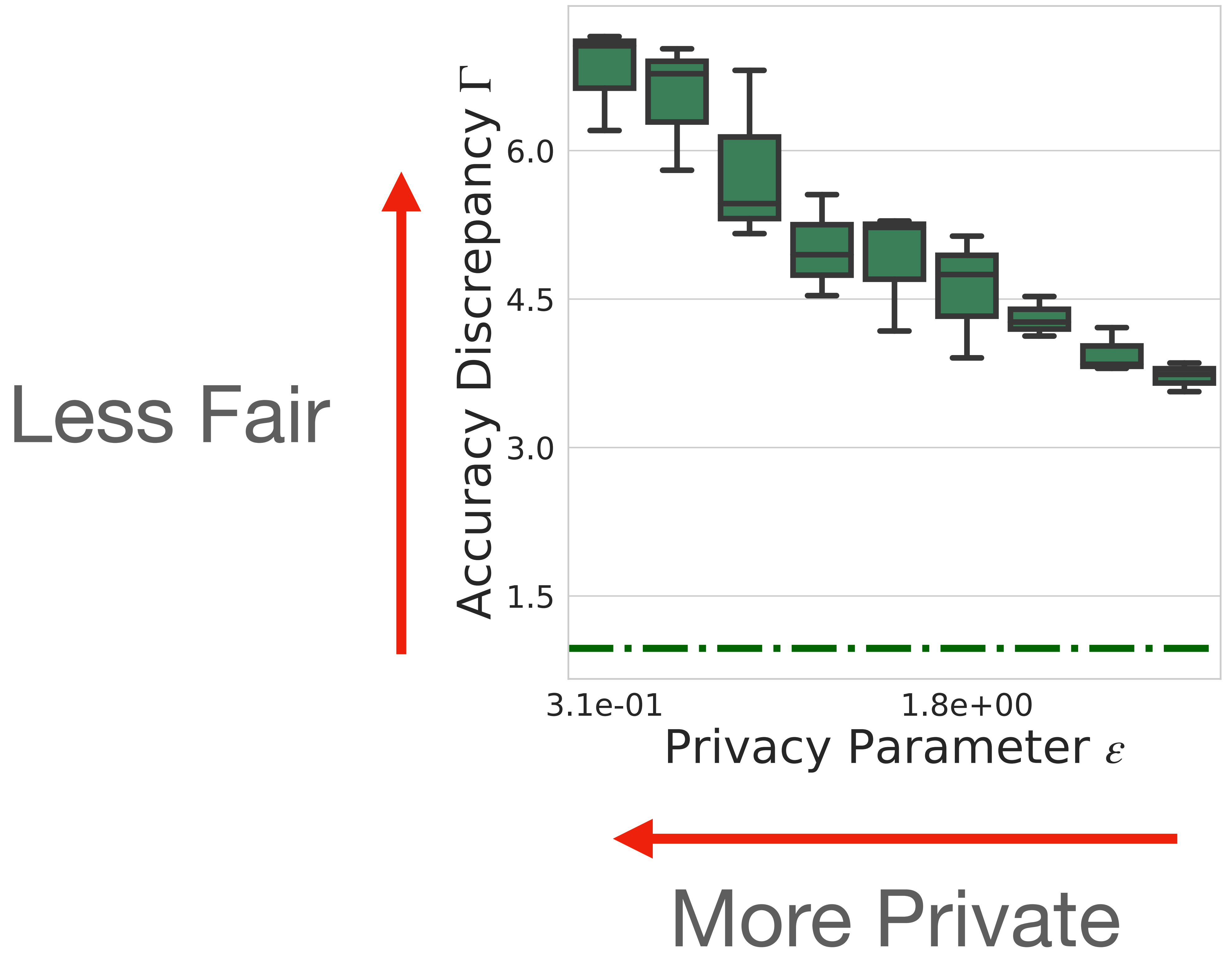
Less Fair



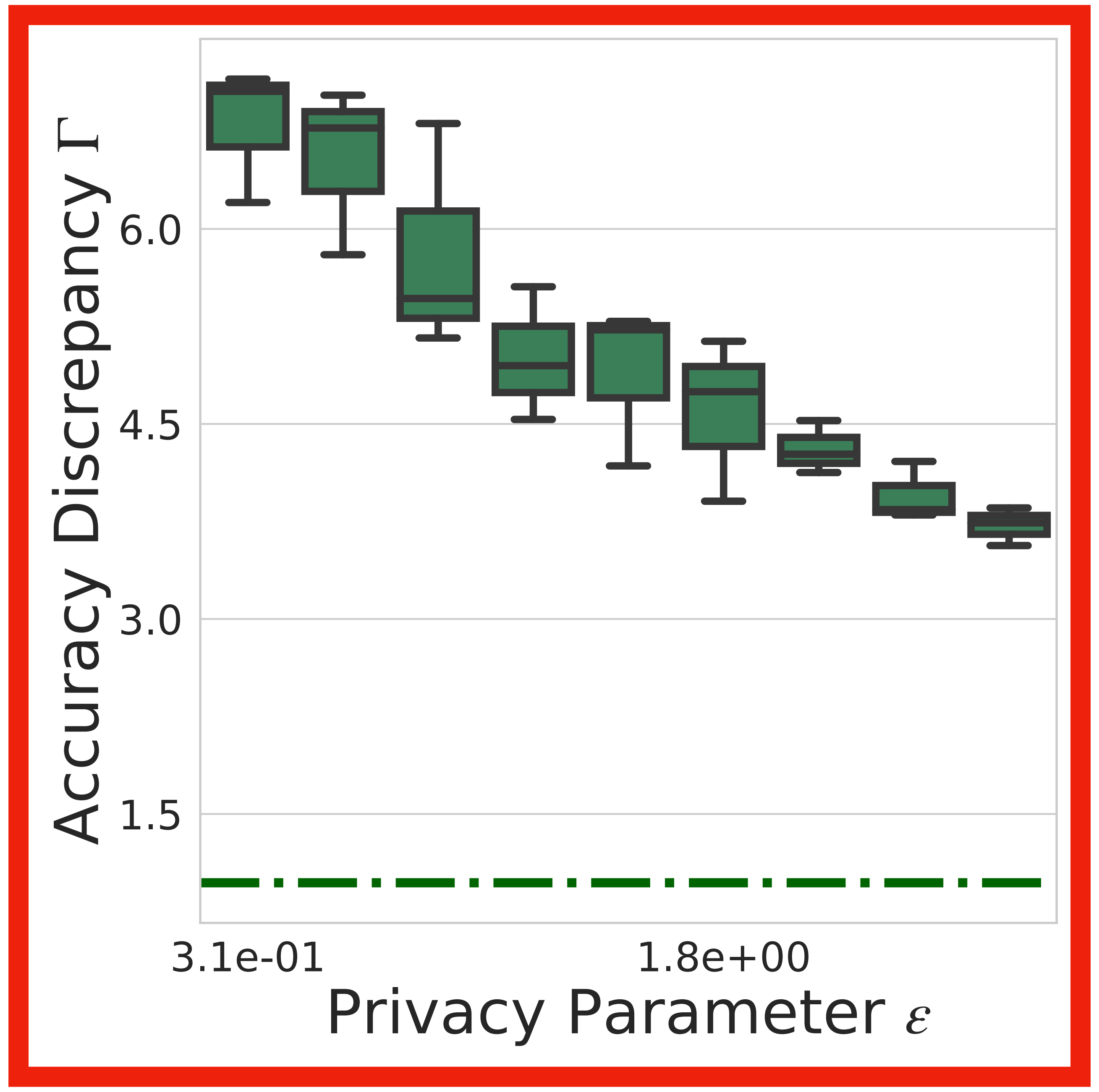
More Private

Privacy vs Fairness
CelebA



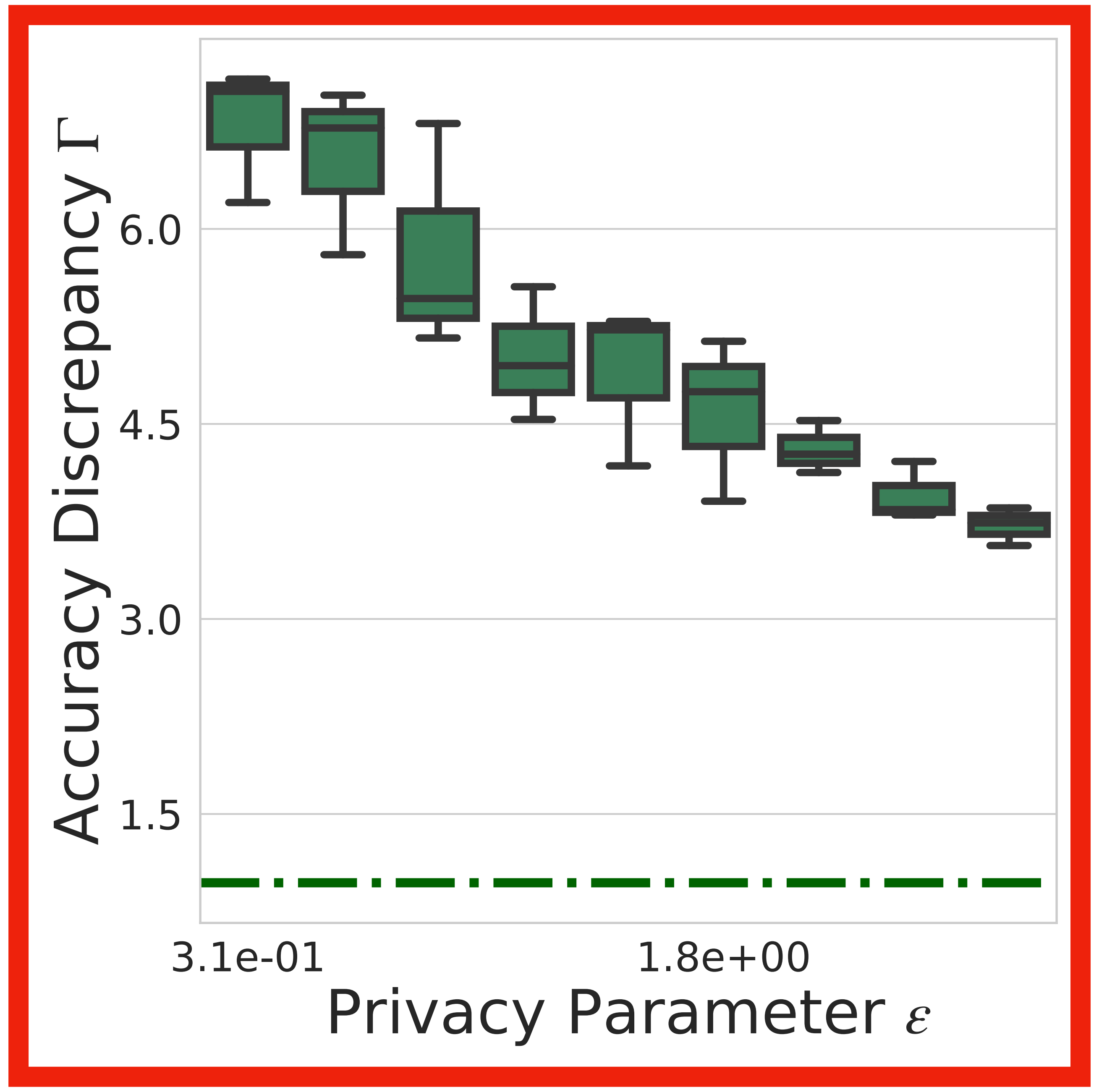


Is this trend systematic?



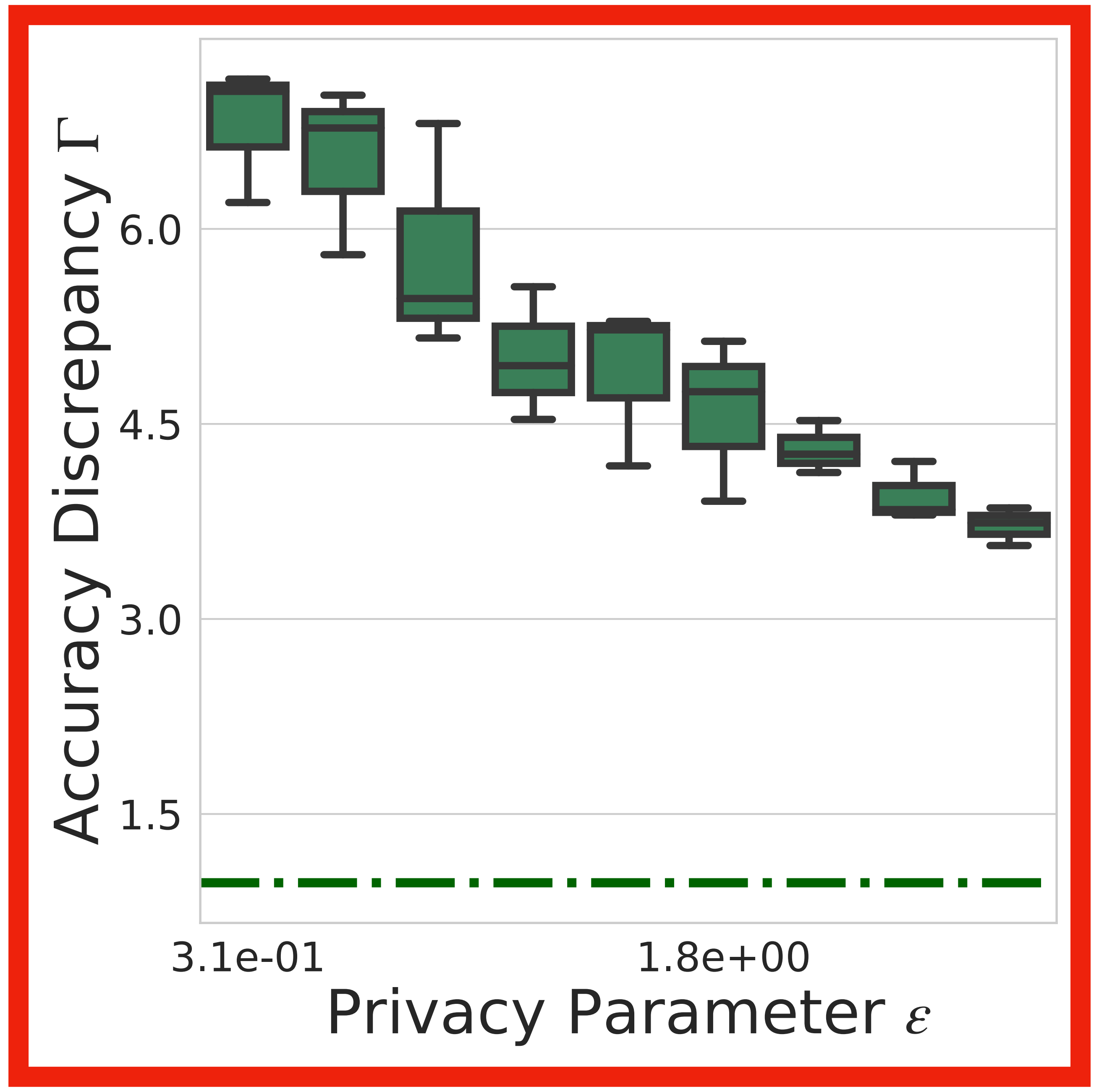
Is this trend systematic?

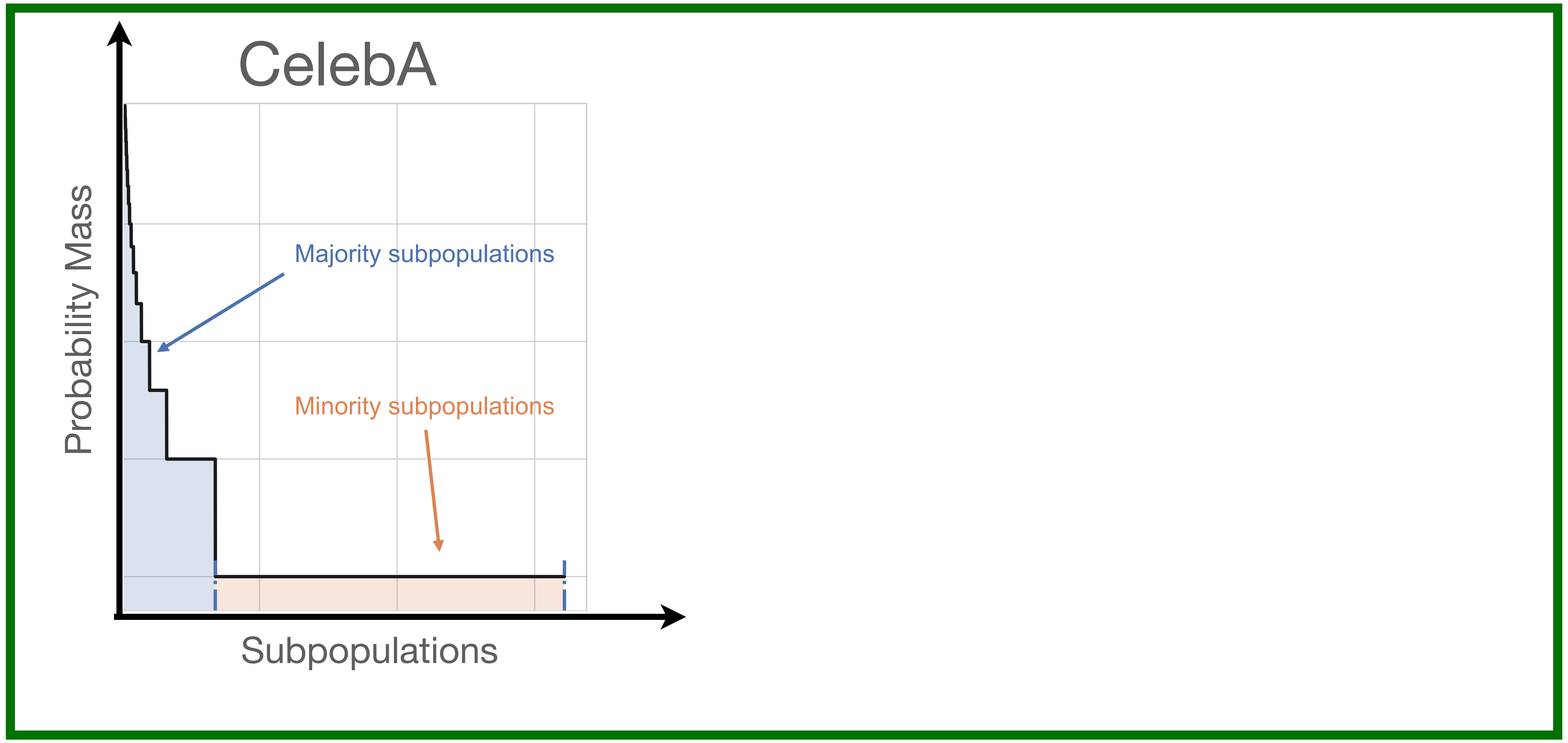
Main Contribution:
We prove this trend in a model-agnostic setting



Is this trend systematic?

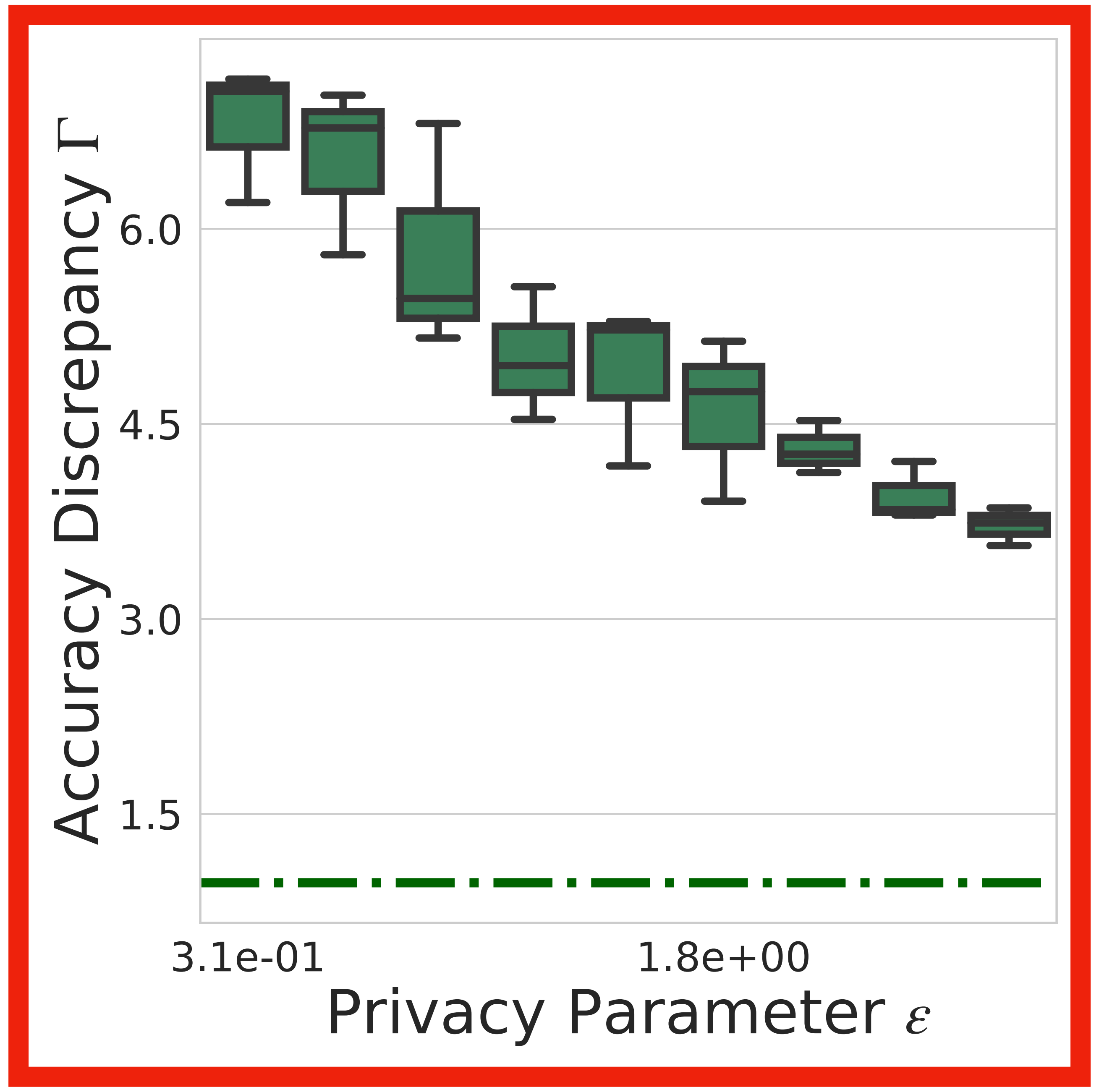
Main Contribution:
We prove this trend in a model-agnostic setting for long-tailed distribution.

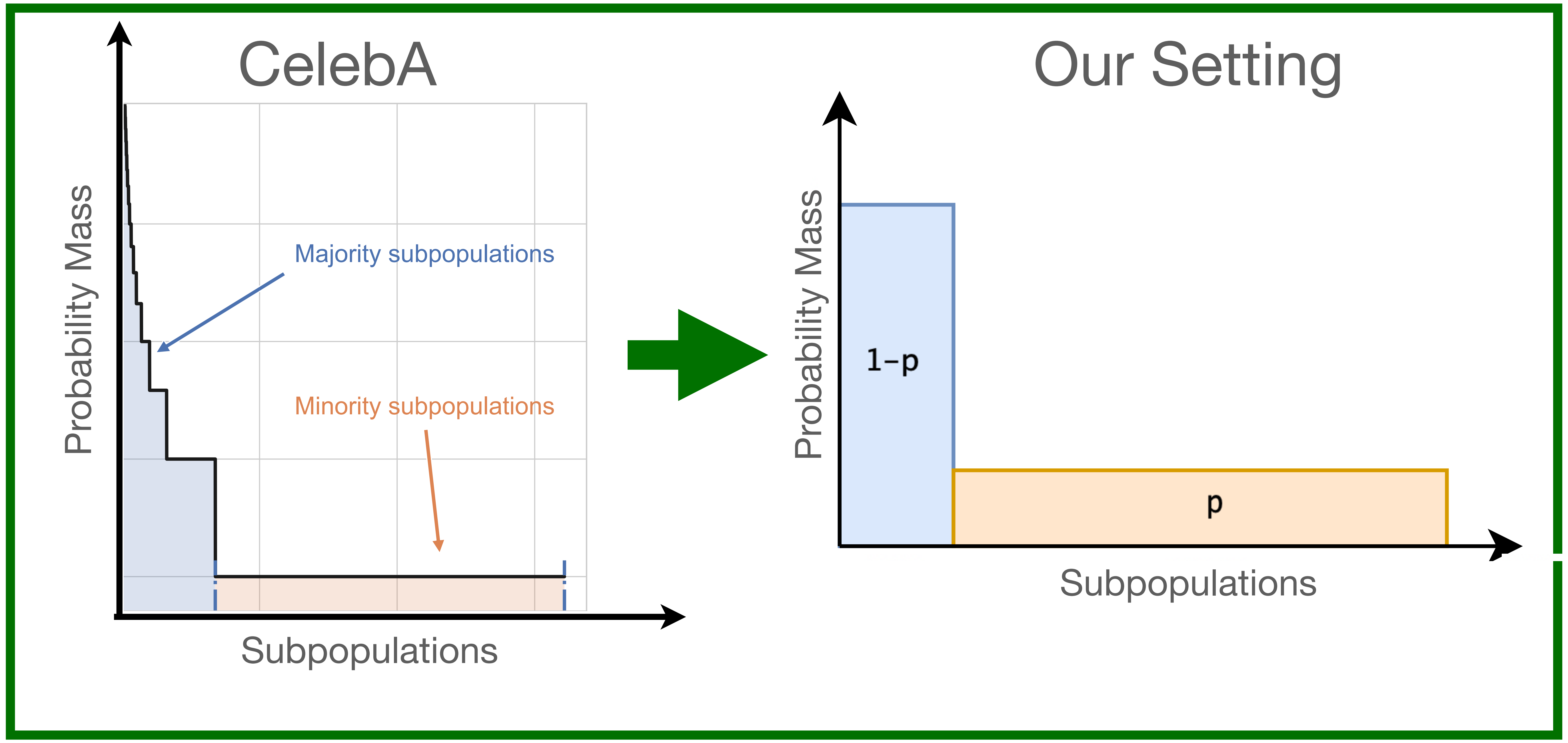




Is this trend systematic?

Main Contribution:
We prove this trend in a model-agnostic setting for long-tailed distribution.





Definitions of error and fairness

Definitions of error and fairness

- Error

Definitions of error and fairness

- Error $\operatorname{err}(A, \Pi, F) =$

Definitions of error and fairness

- Error $\mathrm{err}\left(A, \Pi, F\right) =$

Learning Algorithm

Definitions of error and fairness

- Error

$$\text{err}\left(\boxed{A}, \boxed{\Pi}, F\right) =$$

Learning Algorithm

Data Distribution

Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- Error

$$\mathrm{err}\left(A, \Pi, F\right) =$$

Learning Algorithm

Data Distribution

Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- Error

$$\mathrm{err}\left(A, \Pi, F\right) = \mathbb{P}\left[h(x) \neq f(x)\right]$$

Learning Algorithm

Probability is over $S \sim \Pi^m, f \sim F, h \sim A(S_f)$, and $x \sim \Pi_{p,N}$

Data Distribution

Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- Error

$$\mathrm{err}\,(A, \Pi, F) = \mathbb{P}\left[h(x) \neq f(x)\right]$$

Learning Algorithm

Data Distribution

Probability is over $\boxed{S \sim \Pi^m}$, $f \sim F, h \sim A(S_f)$, and $x \sim \Pi_{p,N}$

Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- Error

$$\mathrm{err}\,(A, \Pi, F) = \mathbb{P}\left[h(x) \neq f(x)\right]$$

Learning Algorithm

Probability is over $S \sim \Pi^m, \boxed{f \sim F}, h \sim A(S_f)$, and $x \sim \Pi_{p,N}$

Data Distribution

Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- Error $\quad \mathrm{err}\,(A, \Pi, F) = \mathbb{P}\left[h(x) \neq f(x)\right]$

Learning Algorithm

Probability is over $S \sim \Pi^m, f \sim F, \boxed{h \sim A(S_f),}$ and $x \sim \Pi_{p,N}$

Data Distribution

Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- Error

$$\mathrm{err}\left(A, \Pi, F\right) = \mathbb{P}\left[h(x) \neq f(x)\right]$$

Learning Algorithm

Probability is over $S \sim \Pi^m, f \sim F, h \sim A(S_f)$, and $\boxed{x \sim \Pi_{p,N}}$

Data Distribution

Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- Error

$$\mathrm{err}\left(A, \Pi, F\right) = \mathbb{P}\left[h(x) \neq f(x)\right]$$

Learning Algorithm

Probability is over $S \sim \Pi^m, f \sim F, h \sim A(S_f)$, and $x \sim \Pi_{p,N}$

Data Distribution

Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- Error

$$\mathrm{err}(A, \Pi, F) = \mathbb{P}\left[h(x) \neq f(x)\right]$$

Learning Algorithm

Probability is over $S \sim \Pi^m, f \sim F, h \sim A(S_f)$, and $x \sim \Pi_{p,N}$

Data Distribution

- Accuracy Discrepancy

Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- Error

$$\mathrm{err}\left(\boxed{A}, \boxed{\Pi}, \boxed{F}\right) = \boxed{\mathbb{P}}\left[h(x) \neq f(x)\right]$$

Learning Algorithm

Probability is over $S \sim \Pi^m, f \sim F, h \sim A(S_f)$, and $x \sim \Pi_{p,N}$

Data Distribution

- Accuracy Discrepancy

$$\Gamma\left(A, \Pi, F\right) = \mathrm{err}_{\mathrm{Minority}}\left(A, \Pi, F\right) - \mathrm{err}\left(A, \Pi, F\right)$$

Definitions of error and fairness

Prior distribution over labelling functions $\subseteq Y^X$

- Error

$$\mathrm{err}\left(A, \Pi, F\right) = \mathbb{P}\left[h(x) \neq f(x)\right]$$

Learning Algorithm

Data Distribution

Probability is over $S \sim \Pi^m, f \sim F, h \sim A(S_f)$, and $x \sim \Pi_{p,N}$

- Accuracy Discrepancy

Marginalised over minority subpopulations

$$\Gamma\left(A, \Pi, F\right) = \boxed{\mathrm{err}_{\mathrm{Minority}}\left(A, \Pi, F\right)} - \mathrm{err}\left(A, \Pi, F\right)$$

Privacy at the cost of fairness

Privacy at the cost of fairness

Consider any (ϵ, δ)**-DP algorithm** that obtains low error on a long-tailed distribution.

Privacy at the cost of fairness

Consider any (ϵ, δ)**-DP algorithm** that obtains low error on a long-tailed distribution.

(Informal Theorem A) We prove an asymptotic lower bound for accuracy discrepancy which

- (Privacy) Increases with privacy parameter ϵ.

Privacy at the cost of fairness

Consider any **(ϵ, δ)-DP algorithm** that obtains low error on a long-tailed distribution.

N: # Minority subpopulations
m: # Training points

(Informal Theorem A) We prove an asymptotic lower bound for accuracy discrepancy which

- (Privacy) Increases with privacy parameter ϵ.

Privacy at the cost of fairness

Consider any (ϵ, δ)-**DP algorithm** that obtains low error on a long-tailed distribution.

(Minority Subpopulations) Let $\frac{N}{m} \to c$ as $N, m \to \infty$.

N: # Minority subpopulations
m: # Training points

(Informal Theorem A) We prove an asymptotic lower bound for accuracy discrepancy which

- (Privacy) Increases with privacy parameter ϵ.

Privacy at the cost of fairness

Consider any (ϵ, δ)**-DP algorithm** that obtains low error on a long-tailed distribution.

(Minority Subpopulations) Let $\frac{N}{m} \to c$ as $N, m \to \infty$.

N: # Minority subpopulations
m: # Training points

(Informal Theorem A) We prove an asymptotic lower bound for accuracy discrepancy which

- (Privacy) Increases with privacy parameter ϵ.
- (Long-tailed) Increases with (relative) # of minority subpopulations c.

Privacy at the cost of fairness

Consider any (ϵ, δ)**-DP algorithm** that obtains low error on a long-tailed distribution.

(Minority Subpopulations) Let $\frac{N}{m} \to c$ as $N, m \to \infty$.

N: # Minority subpopulations
m: # Training points
F: Label prior

(Informal Theorem A) We prove an asymptotic lower bound for accuracy discrepancy which

- (Privacy) Increases with privacy parameter ϵ.
- (Long-tailed) Increases with (relative) # of minority subpopulations c.

Privacy at the cost of fairness

Consider any (ϵ, δ)**-DP algorithm** that obtains low error on a long-tailed distribution.

(Minority Subpopulations) Let $\frac{N}{m} \to c$ as $N, m \to \infty$.

(Label prior Entropy) Define $\|F\|_\infty = \max_{x,y} \mathbb{P}_{f \sim F}\left[f(x) = y\right]$

N: # Minority subpopulations
m: # Training points
F: Label prior

(Informal Theorem A) We prove an asymptotic lower bound for accuracy discrepancy which

- (Privacy) Increases with privacy parameter ϵ.
- (Long-tailed) Increases with (relative) # of minority subpopulations c.

Privacy at the cost of fairness

Consider any (ϵ, δ)**-DP algorithm** that obtains low error on a long-tailed distribution.

(Minority Subpopulations) Let $\frac{\mathrm{N}}{m} \to c$ as $N, m \to \infty$.

(Label prior Entropy) Define $\|F\|_\infty = \max_{x,y} \mathbb{P}_{f\sim F}\left[f(x) = y\right]$

N: # Minority subpopulations

m: # Training points

F: Label prior

(Informal Theorem A) We prove an asymptotic lower bound for accuracy discrepancy which

- (Privacy) Increases with privacy parameter ϵ.
- (Long-tailed) Increases with (relative) # of minority subpopulations c.
- (Label prior) Increases with entropy of the label prior.

Thank you