Provably Debiasing Machine Learning Datasets

Mislav Balunović

ETH Zurich

SRILAB

ETH zürich

Why fairness and bias?

ML makes decisions that impact people:

- Should person get a loan?
- Is person likely to commit a crime?
- Should person get hired?

The European Commission is creating regulations with a goal that AI systems "do not create or reproduce bias“.

A.I. Could Worsen Health Disparities

In a health system riddled with inequity, we risk making dangerous biases automated and invisible.

The never-ending quest to predict crime using AI

The practice has a long history of skewing police toward communities of color. But that hasn't stopped researchers from building crime-predicting tools.

SCIENCEINSIDER | EUROPE

Europe plans to strictly regulate high-risk AI technology

How AI Is Deciding Who Gets Hired

Employee advocates say hiring software is making discrimination worse. But some applicants are hacking the system.

Major challenges

For wide adoption of fairness in machine learning we need to address the following challenges:

- How to **define** fairness?
- How to **enforce** fairness?
- How to **prove** fairness?

What does it mean to be fair?

Individual fairness

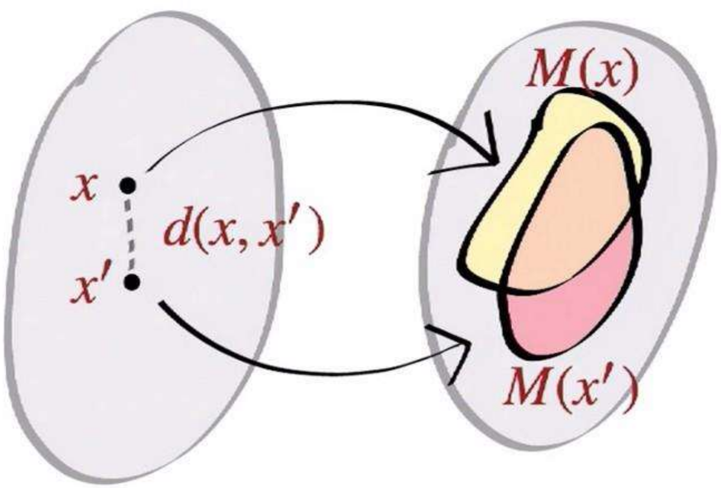


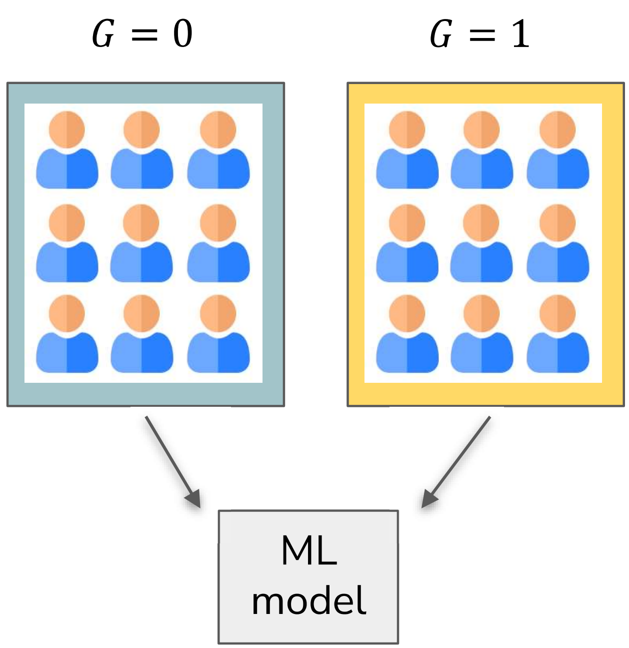
Image source: Moritz Hardt,
Fairness in Machine Learning, NIPS 2017

Requires that if two individuals $\boldsymbol{x}$ and $\boldsymbol{x}'$ are similar (according to some similarity notion), decisions of ML model $\mathrm{M}(\boldsymbol{x})$ and $\mathrm{M}(\boldsymbol{x}')$ should be similar for these two individuals.

Key challenge: finding a suitable distance similarity metric *d* (e.g., L_2 distance in feature space)

What does it mean to be fair?

Group fairness



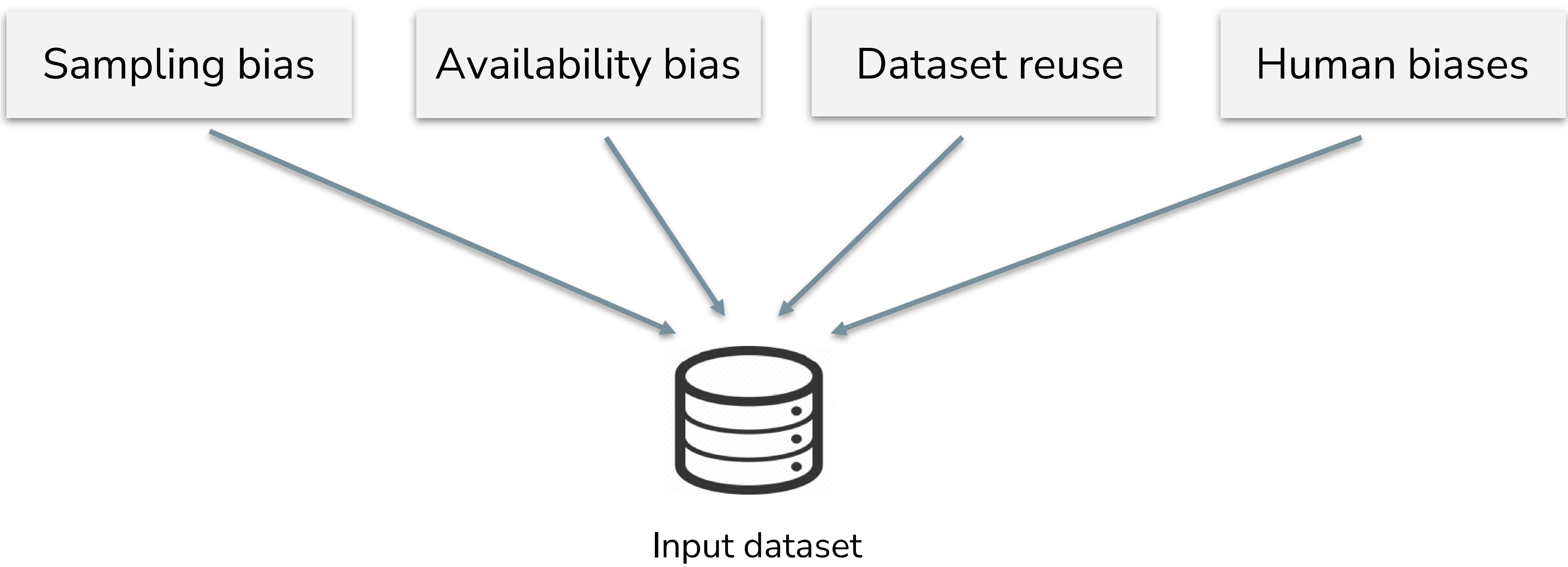
$$P(Y = 1 | G = 0) = P(Y = 1 | G = 1)$$

Requires the probability an ML model assigns a label to different groups is the same (e.g. groups can be different races).

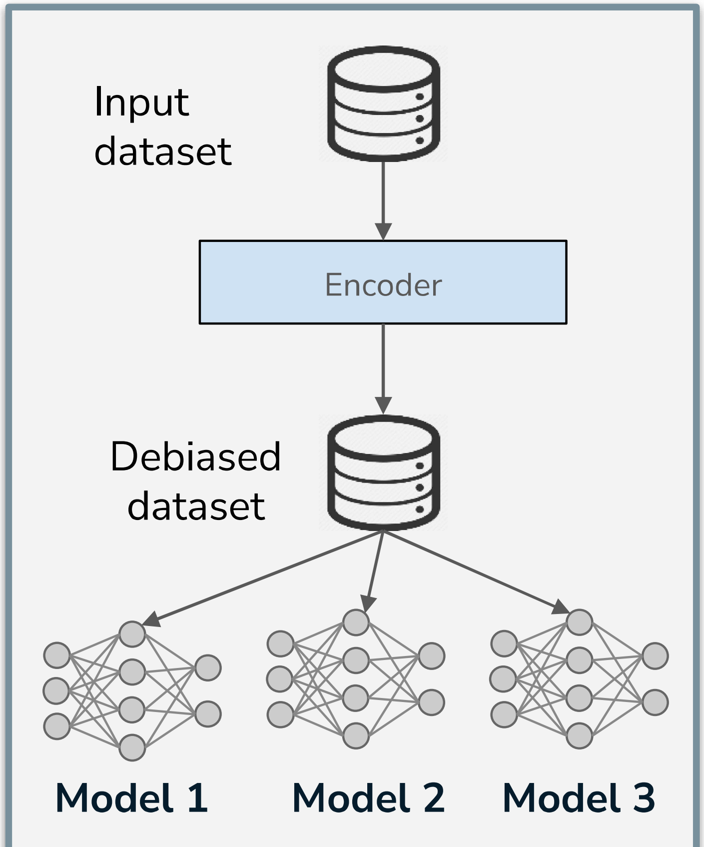
Variants of group fairness differ in the way groups are formed: demographic parity, equal opportunity, etc..

Key challenge: How to define groups?

Sources of bias in datasets



Enforcing fairness

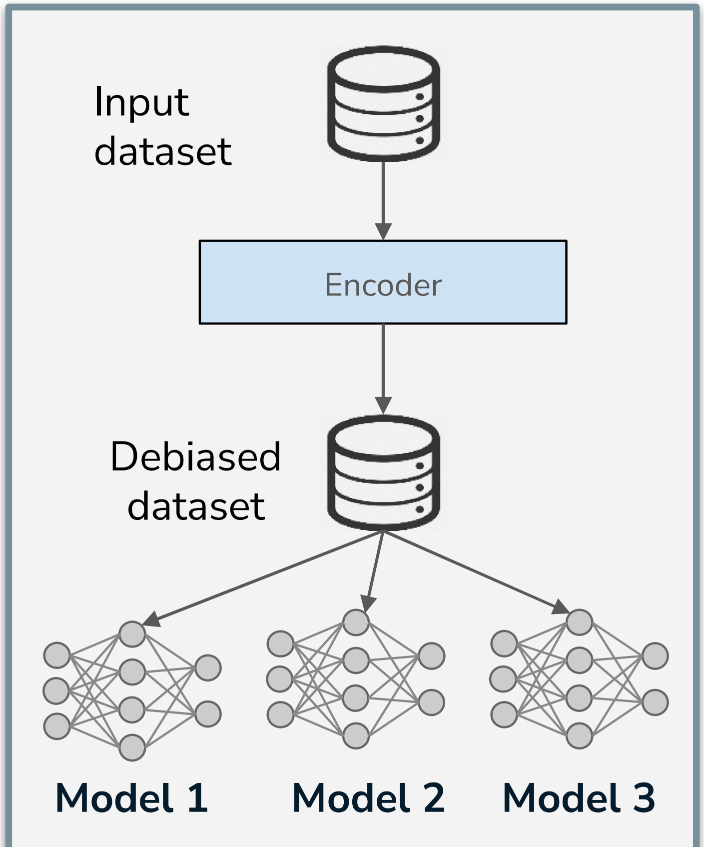


Pre-processing

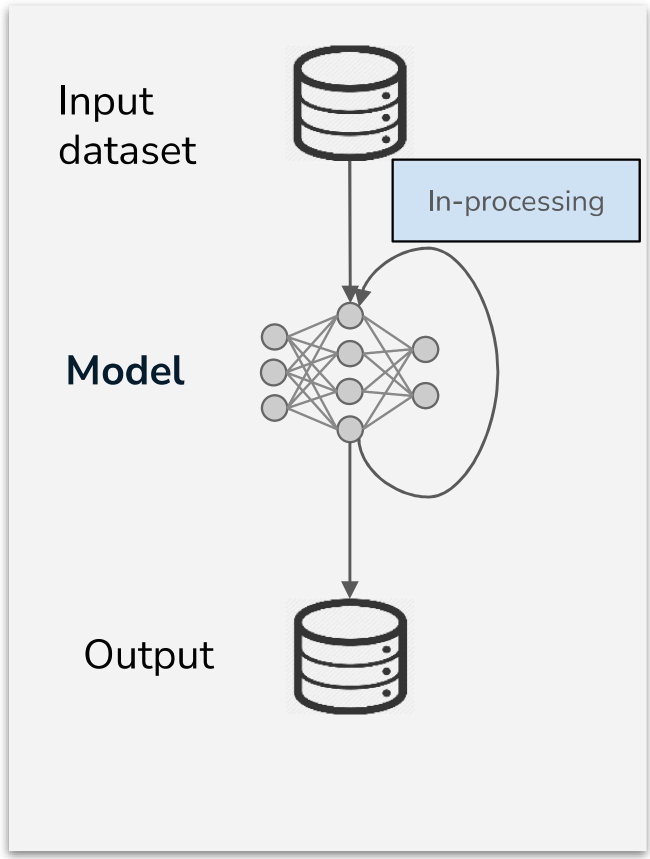
Pre-processing approach assumes there is an encoder f that transforms training dataset $x_1, x_2, \ldots, x_n$ into a new dataset $z_1, z_2, \ldots, z_n$ such that each training input x_i is transformed into a new representation $z_i = f(x_i)$.

Key advantage: we can reuse the debiased dataset for several different tasks!

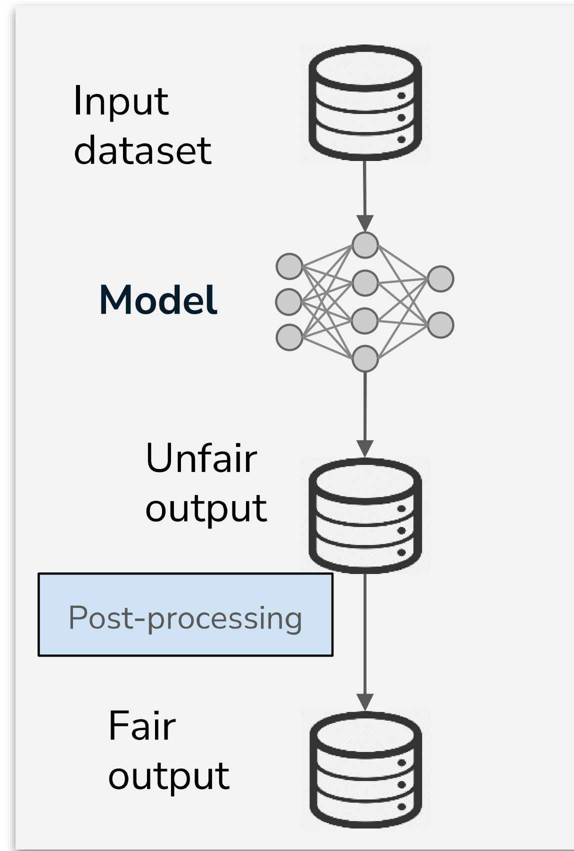
Enforcing fairness



Pre-processing



In-processing



Post-processing

Fairness: Application Domains

Tabular data

Age	Salary	Loan
37	85K	True
26	60K	False
52	100K	True

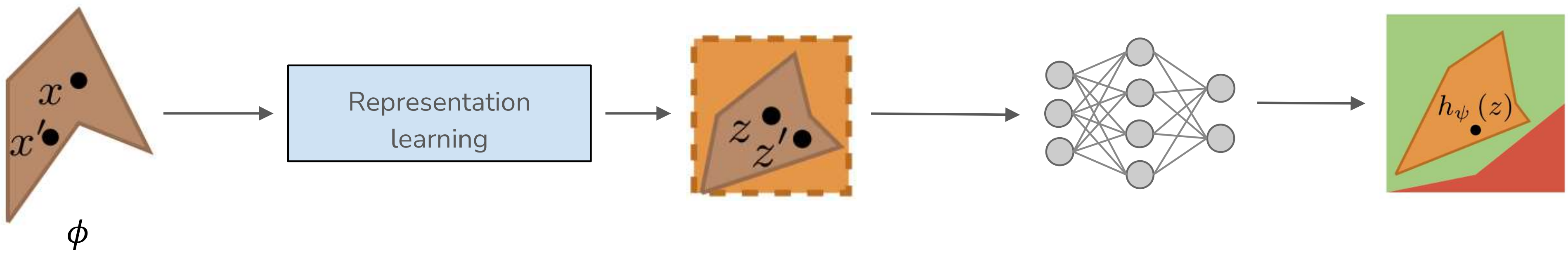
Images

NLP

The first is a training problem. A.I. must learn to diagnose disease on large data sets, and if that data doesn't include enough patients from a particular background, it won't be as reliable for them. Evidence from other fields suggests this isn't just a theoretical concern. A recent study found that some facial recognition programs incorrectly classify less than 1 percent of light-skinned men but more than one-third of dark-skinned women. What happens when we rely on such algorithms to diagnose melanoma on light versus dark skin?

Medicine has long struggled to include enough women and minorities in research, despite knowing they have different risk

Enforcing individual fairness: LCIFR (Ruoss et al., NeurIPS'20)

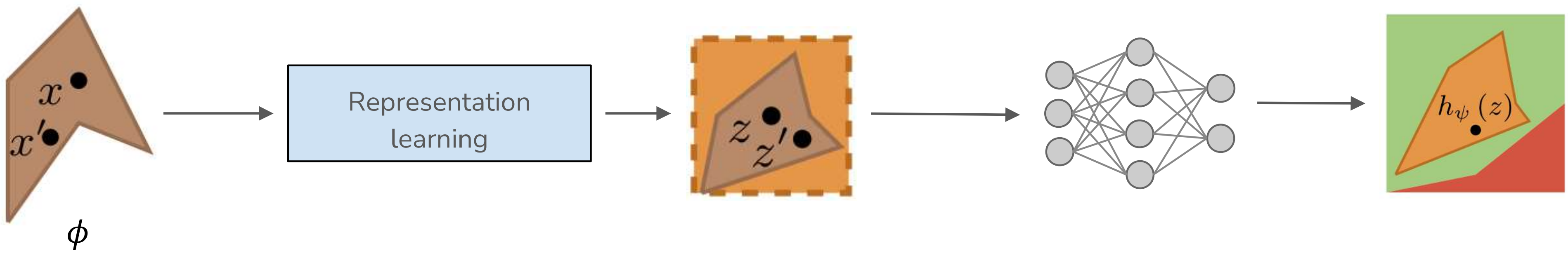


Example of an individual fairness formula ϕ:

Persons $\boldsymbol{x}$ and $\boldsymbol{x}'$ are similar if and only if:

- They differ in age by at most 10
- They have same or different race
- All of their other attributes are the same.

Enforcing individual fairness: LCIFR (Ruoss et al., NeurIPS'20)

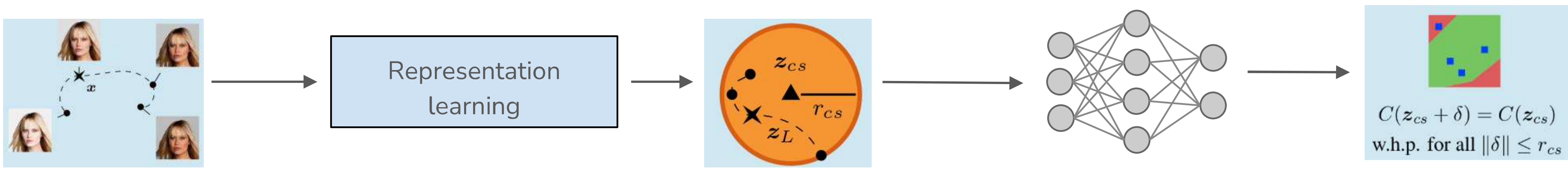


1. Given data point $\boldsymbol{x}$, compute new data representation $\boldsymbol{z}$ which provably guarantees that all data points $\boldsymbol{x}'$ similar to $\boldsymbol{x}$ will get mapped to the neighborhood of $\boldsymbol{z}$:

$$\phi(x, x') \Rightarrow \|z - z'\|_2 < \delta$$

2. Given data representation $\boldsymbol{z}$, train a classifier that is robust to ε-perturbations in the latent space

Enforcing individual fairness: LASSI (Peychev et al., ECCV'22)

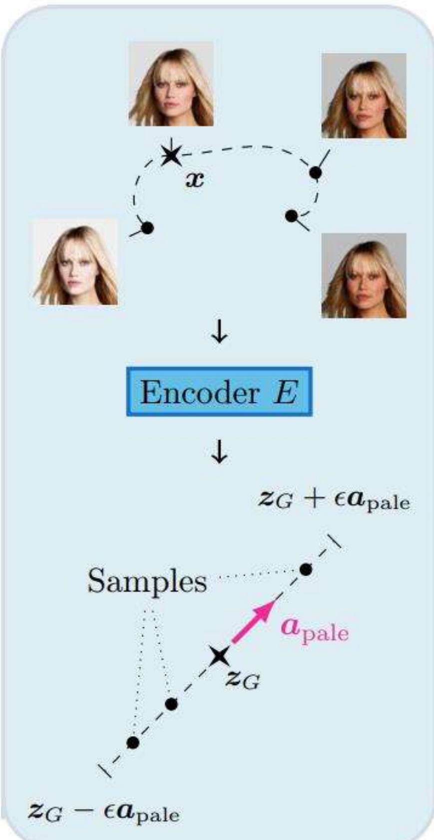


1. Use generative model to capture the set of images similar to $\boldsymbol{x}$

2. Use smoothing to guarantee that representations of similar individuals get mapped to similar representations *with high probability*

3. Use smoothing to guarantee that similar representations get classified the same *with high probability*

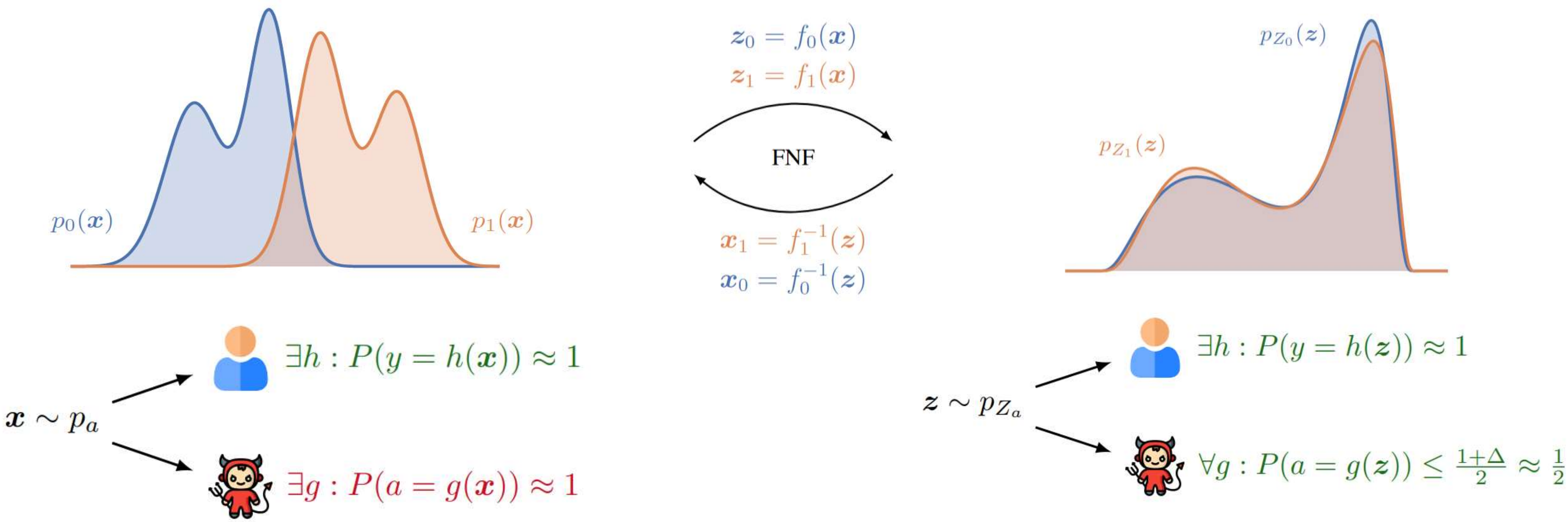
Individual similarity using generative model



Set of images similar to $\boldsymbol{x}$ lies on a curve that cannot be easily captured in the image space

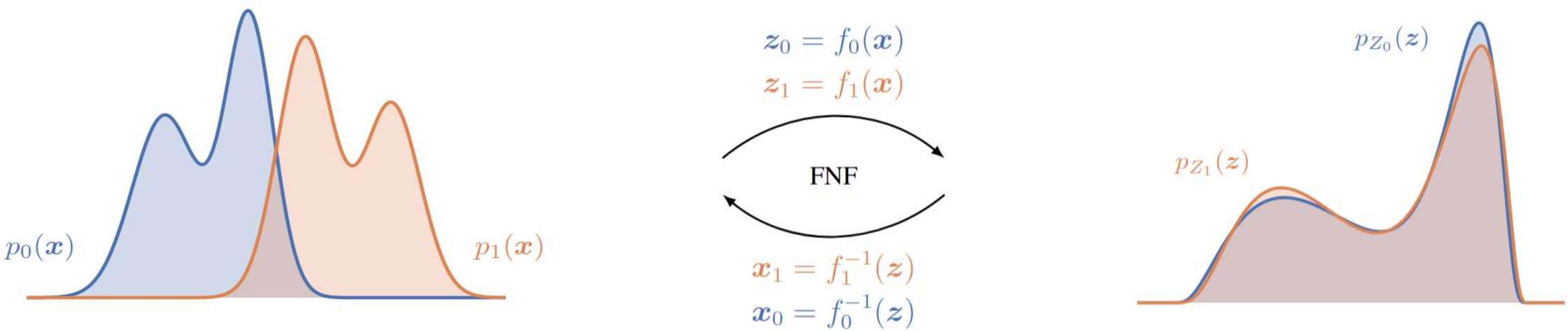
Instead we can capture these images using a line segment in the latent space of a generative model

Enforcing group fairness: FNF (Balunović et al., ICLR'22)



Key idea: Compute representations such that data points $\boldsymbol{x}$ from group 0 get mapped to a new data representation $\boldsymbol{z}$ which provably cannot be distinguished from data points $\boldsymbol{x}'$ from group 1, meaning that $\boldsymbol{p}_{Z_0} \approx \boldsymbol{p}_{Z_1}$

Enforcing group fairness: FNF (Balunović et al., ICLR'22)



We use **bijective** encoder architecture (normalizing flows) which enables us to transform input to output distribution, ultimately allowing for training the encoder to map two groups to similar distributions.

Enforcing group fairness: FNF (Balunović et al., ICLR'22)

$$z \sim p_{Z_a}$$

$$\exists h : P(y = h(z)) \approx 1$$

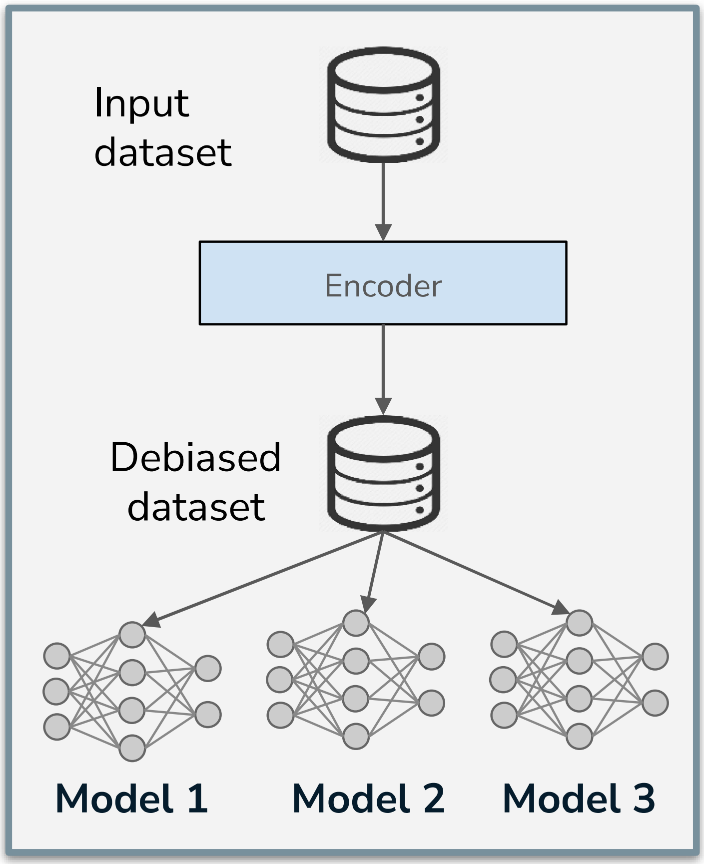
$$\forall g : P(a = g(z)) \leq \frac{1+\Delta}{2} \approx \frac{1}{2}$$

What do we prove?

Assuming we know the probability distribution over inputs x, we can estimate statistical distance Δ over latent representations z.

This allows us to bound maximum accuracy of the adversary (with high confidence).

Conclusion



Pre-processing