Security and Privacy of ML

Algorithmic Bias and Fairness (in ML)

Shang-Tse Chen

Department of Computer Science
& Information Engineering
National Taiwan University

Many slides adapted from MIT 6.S191: AI Bias and Fairness

Review: Fariness Formulation

X (features)

A (protected attribute)

Y (label)

X1	…	…	…	…	Race	Bail
0	…	0	1	…	1	Y
1	…	1	0	…	1	N
1	…	1	0	…	0	N
..	…	…	…	…	…	…

$$\mathbb{P}_a\{E\} = \mathbb{P}\{E \mid A = a\}.$$

Review: Demographic parity

Definition. Classifier C satisfies *demographic parity* if C is independent of A.

When C is binary $0/1$-variables, this means
$\mathbb{P}_a\{C = 1\} = \mathbb{P}_b\{C = 1\}$ for all groups a, b.

Approximate versions:

$$\frac{\mathbb{P}_a\{C = 1\}}{\mathbb{P}_b\{C = 1\}} \geq 1 - \epsilon$$

$$|\mathbb{P}_a\{C = 1\} - \mathbb{P}_b\{C = 1\}| \leq \epsilon$$

Review: Accuracy Parity

Definition. Classifier C satisfies *accuracy parity* if $\mathbb{P}_a\{C = Y\} = \mathbb{P}_b\{C = Y\}$ for all groups a, b.

- Pros:
  - Random guessing doesn't work
  - Allows perfect classifier
- Cons:
  - Error types matter!
  - Allows you to make up for rejecting qualified women by accepting unqualified men

Rewiew: True Positive Parity (TPP)
(or equal opportunity)

Assume C and Y are binary $0/1$-variables.

Definition. Classifier C satisfies *true positive parity* if $\mathbb{P}_a\{C = 1 \mid Y = 1\} = \mathbb{P}_b\{C = 1 \mid Y = 1\}$ for all groups a, b.

- When positive outcome (1) is desirable
- Equivalently, primary harm is due to false negatives
  - Deny bail when person will not recidivate

Review: False Positive Parity (FPP)

Assume C and Y are binary $0/1$-variables.

Definition. Classifier C satisfies *false positive parity* if $\mathbb{P}_a\{C = 1 \mid Y = 0\} = \mathbb{P}_b\{C = 1 \mid Y = 0\}$ for all groups a, b.

- TPP + FPP: Equalized Odds, or
  Positive Rate Parity

R satisfies equalized odds if
R is conditionally independent of A given Y.

Review: Predictive Value Parity

Assume C and Y are binary $0/1$-variables.

Definition. Classifier C satisfies

- *positive predictive value parity* if for all groups a, b:
  $\mathbb{P}_a\{Y = 1 \mid C = 1\} = \mathbb{P}_b\{Y = 1 \mid C = 1\}$
- *negative predictive value parity* if for all groups a, b:
  $\mathbb{P}_a\{Y = 1 \mid C = 0\} = \mathbb{P}_b\{Y = 1 \mid C = 0\}$
- *predictive value parity* if it satisfies both of the above.

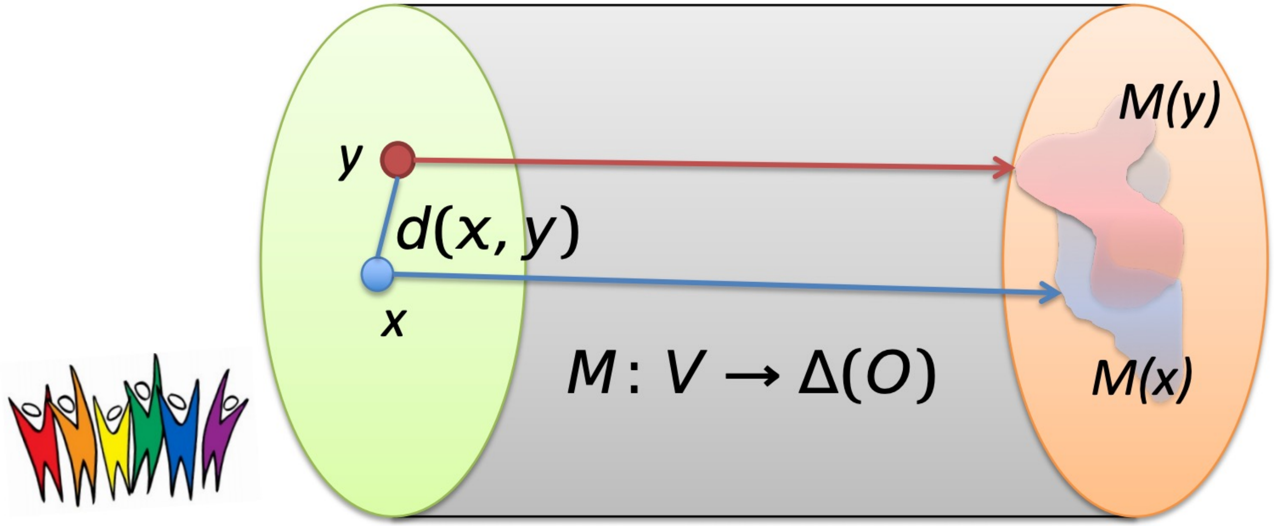
Equalized chance of success given acceptance

Review: Individual Fairness

Metric $d : V \times V \rightarrow \mathbb{R}$

Lipschitz condition $\|M(x) - M(y)\| \leq d(x, y)$

This talk: Statistical distance in [0,1]



V: Individuals O: outcomes

Today's Focus: Algorithmic Bias

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk – A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

Racial bias in a medical algorithm favors white patients over sicker black patients

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

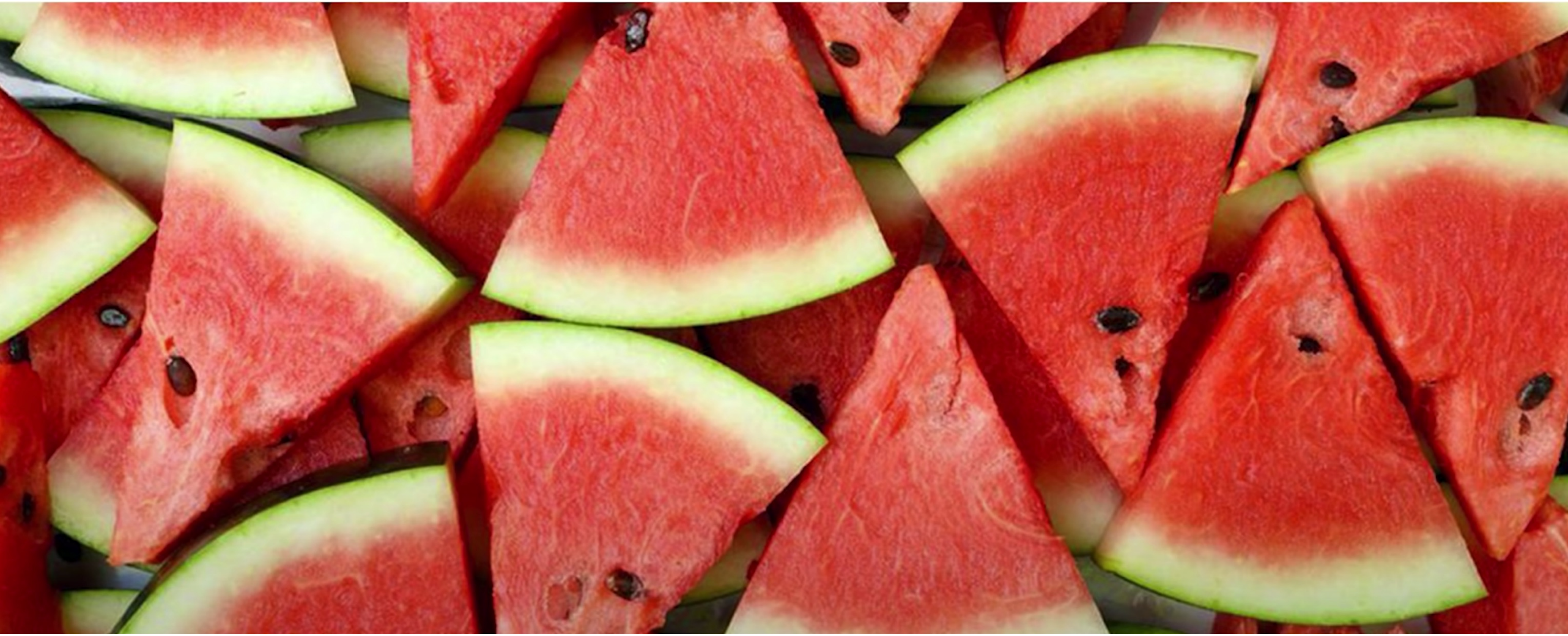
Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Artificial Intelligence has a gender bias problem – just ask Siri

The Best Algorithms Struggle to Recognize Black Faces Equally

US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

What is in This Image



What is in This Image

Watermelon

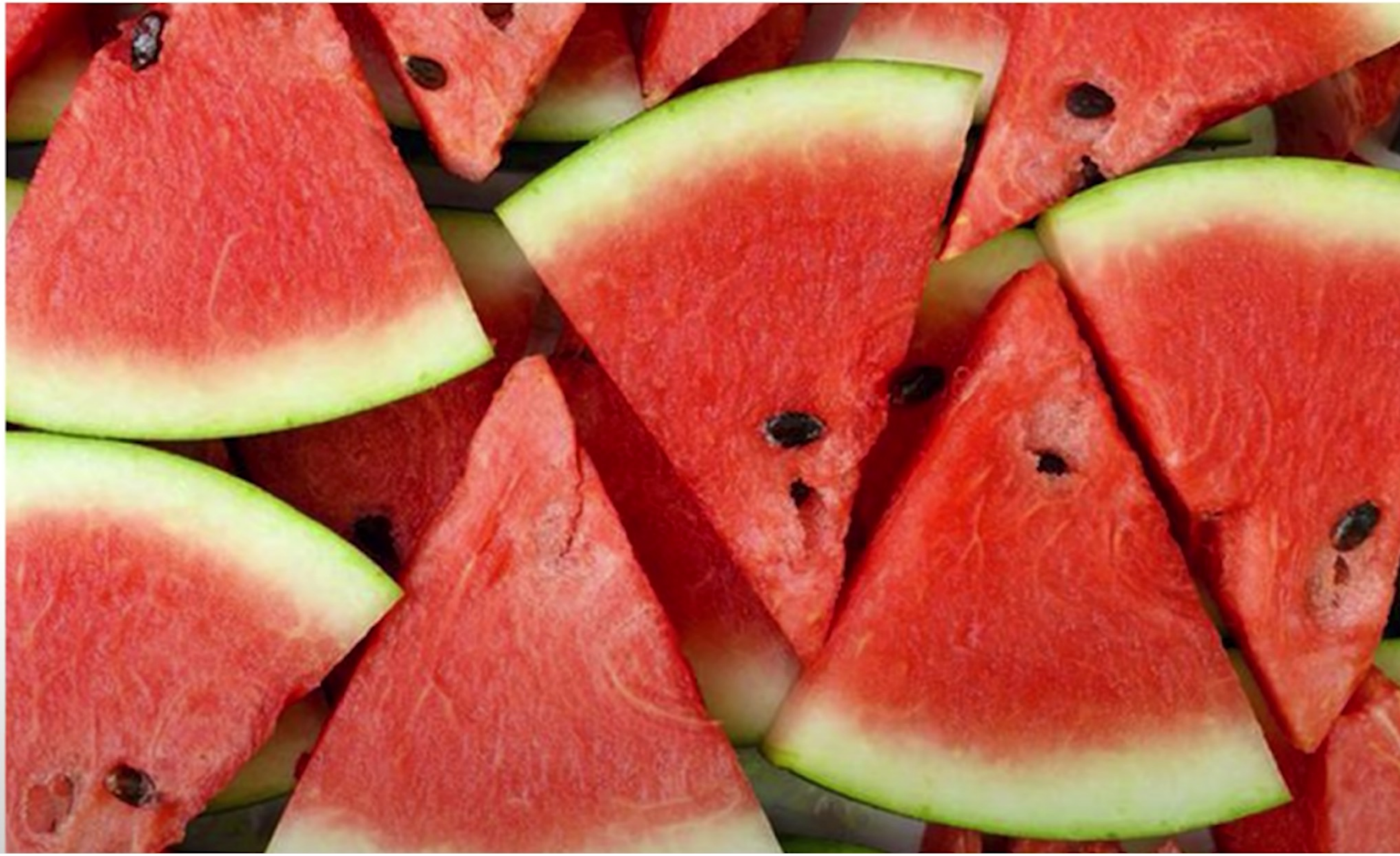
Watermelon slices

Watermelon with seeds

Juicy watermelon

Layers of watermelon

Watermelon slices next to each other



What is in This Image

Watermelon

Watermelon slices

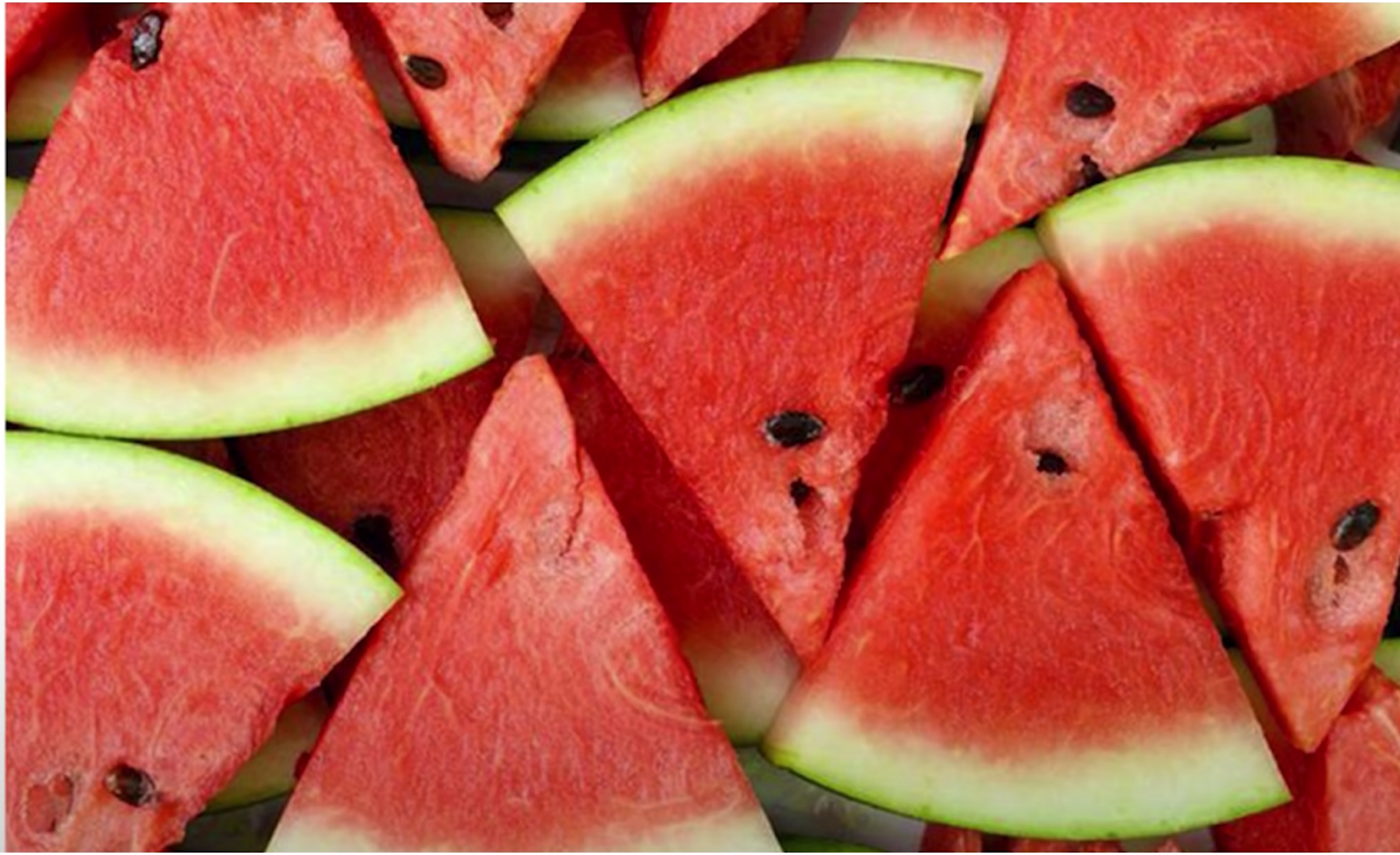
Watermelon with seeds

Juicy watermelon

Layers of watermelon

Watermelon slices next to each other

But what about
red watermelon?



What is in This Image

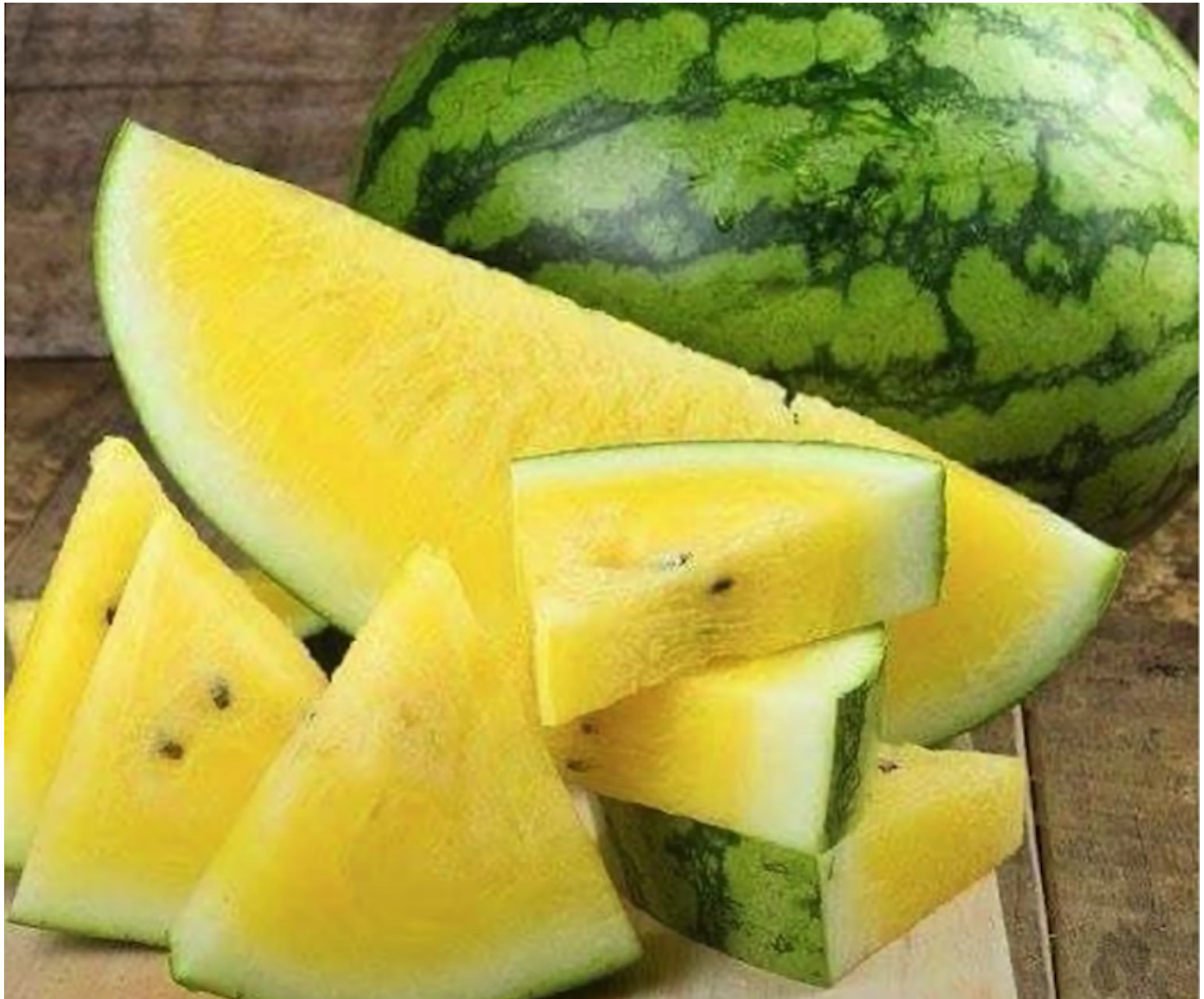
Yellow watermelon

Yellow watermelon slices

Yellow watermelon with seeds

Juicy yellow watermelon

...

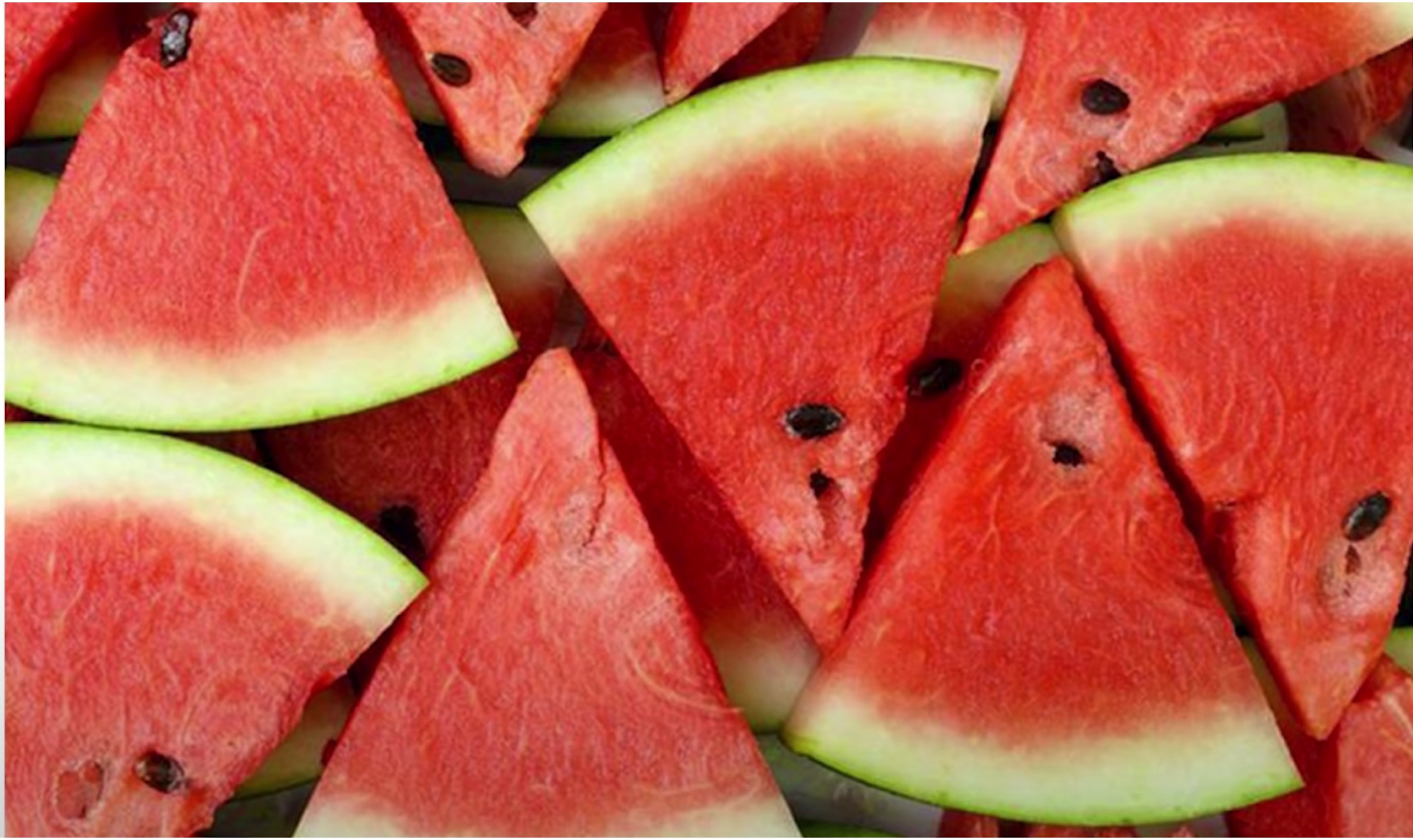


What is in This Image

But what about
red watermelon?

We tend not to think of
the contents of this image
as red watermelon.

Red is the *prototypical* color
for watermelon flesh.



Labeling, Prototyping, and Stereotyping

We **label** and **categorize** the world to reduce complex sensory inputs into **simplified** groups that are easier to work with.

Prototypes are “typical” representations of a concept or object.

We tend to notice and talk about things that are **atypical.**

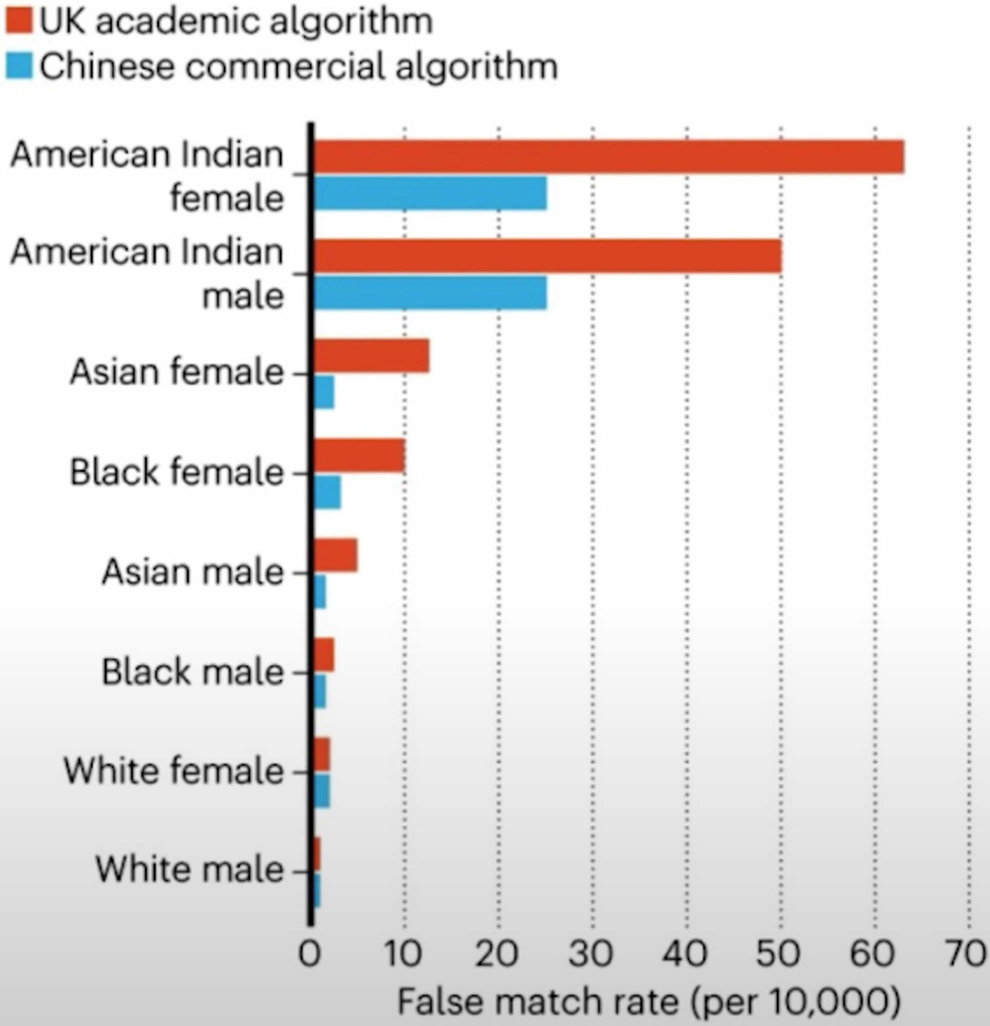
Biases and **stereotypes** arise when particular labels and features **confound decisions** – whether human or artificial.

Bias in Facial Detection

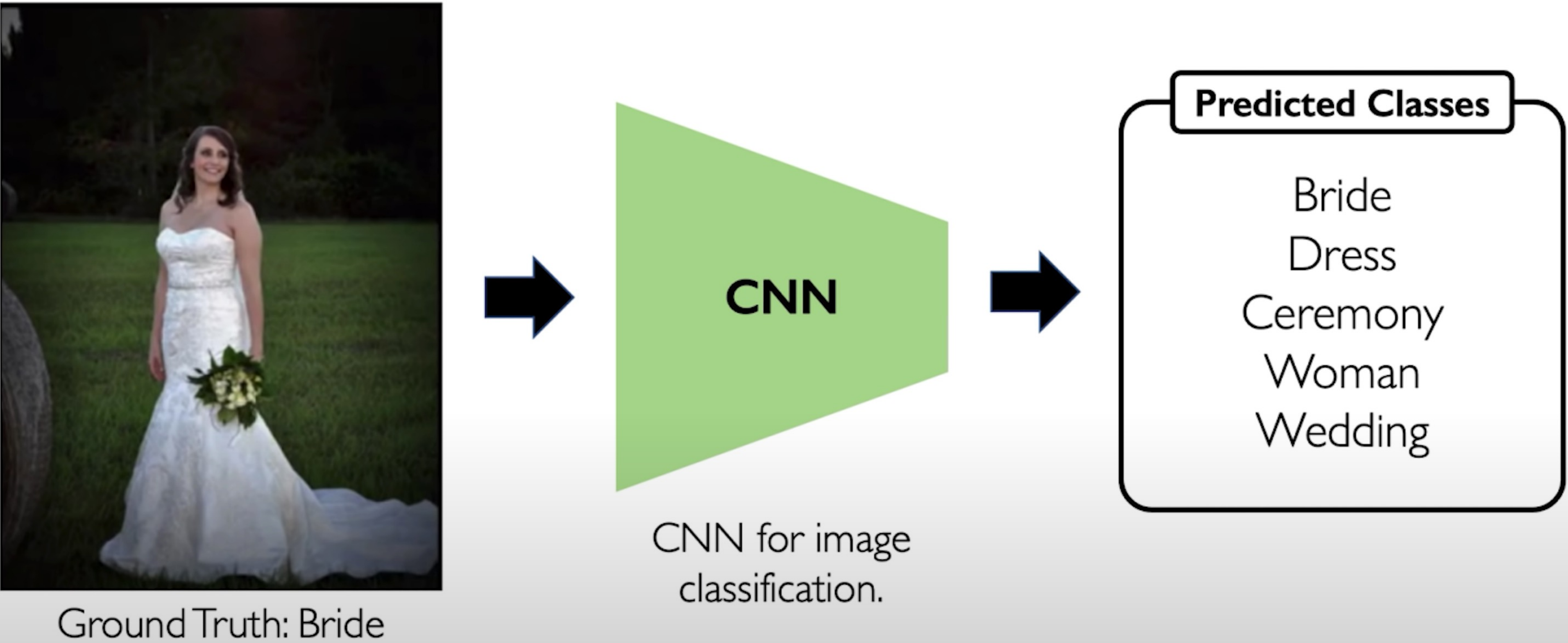
Independent Study I

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Independent Study II

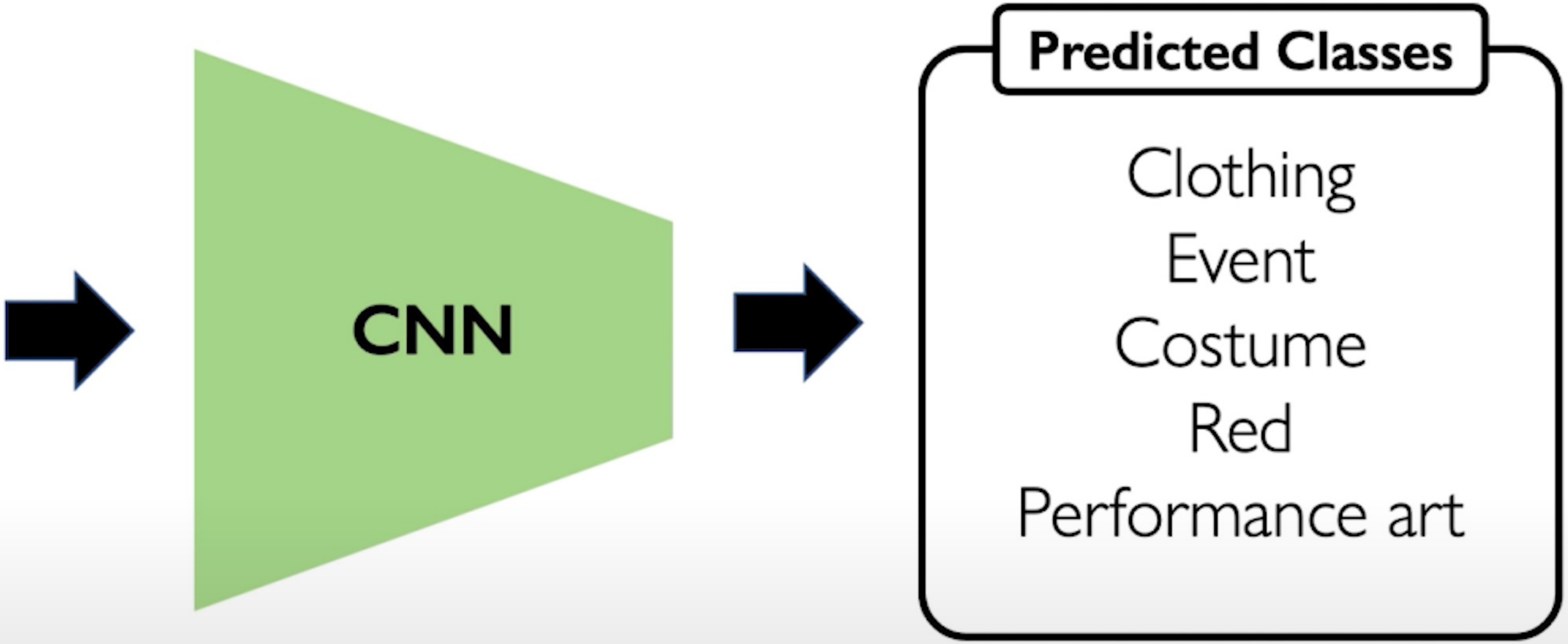


Bias in Image Classification



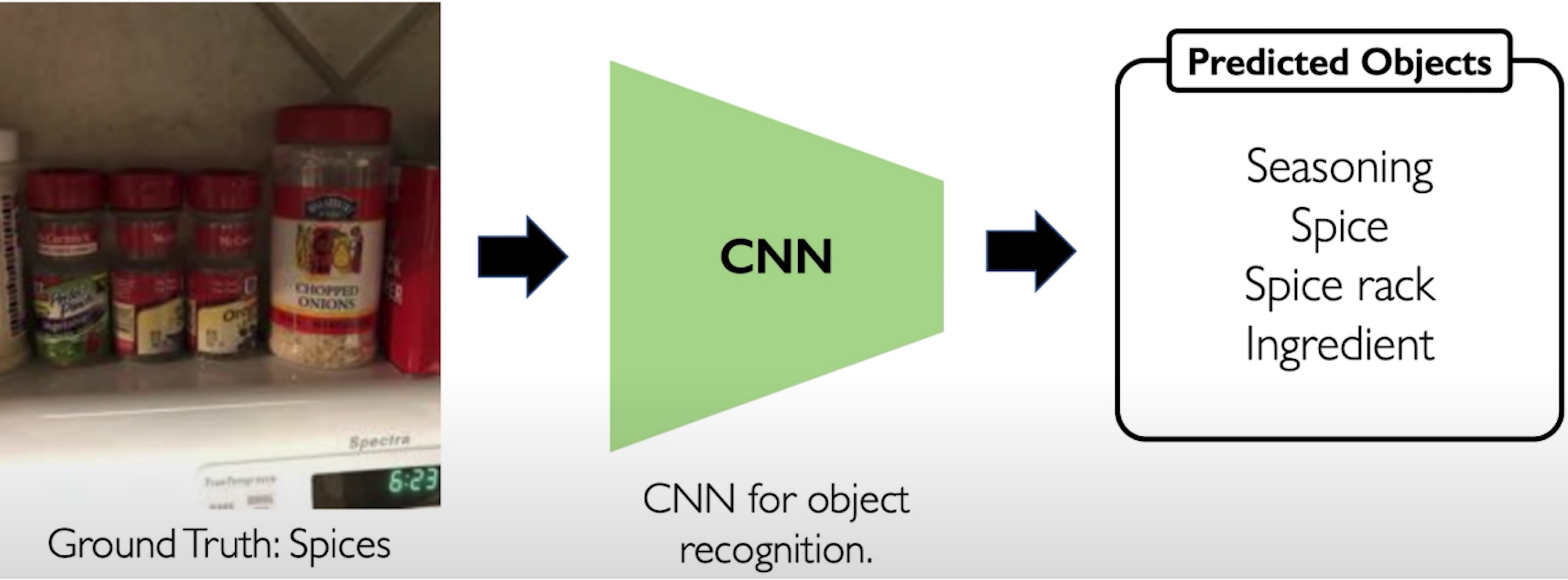
Bias in Image Classification

Ground Truth: Bride

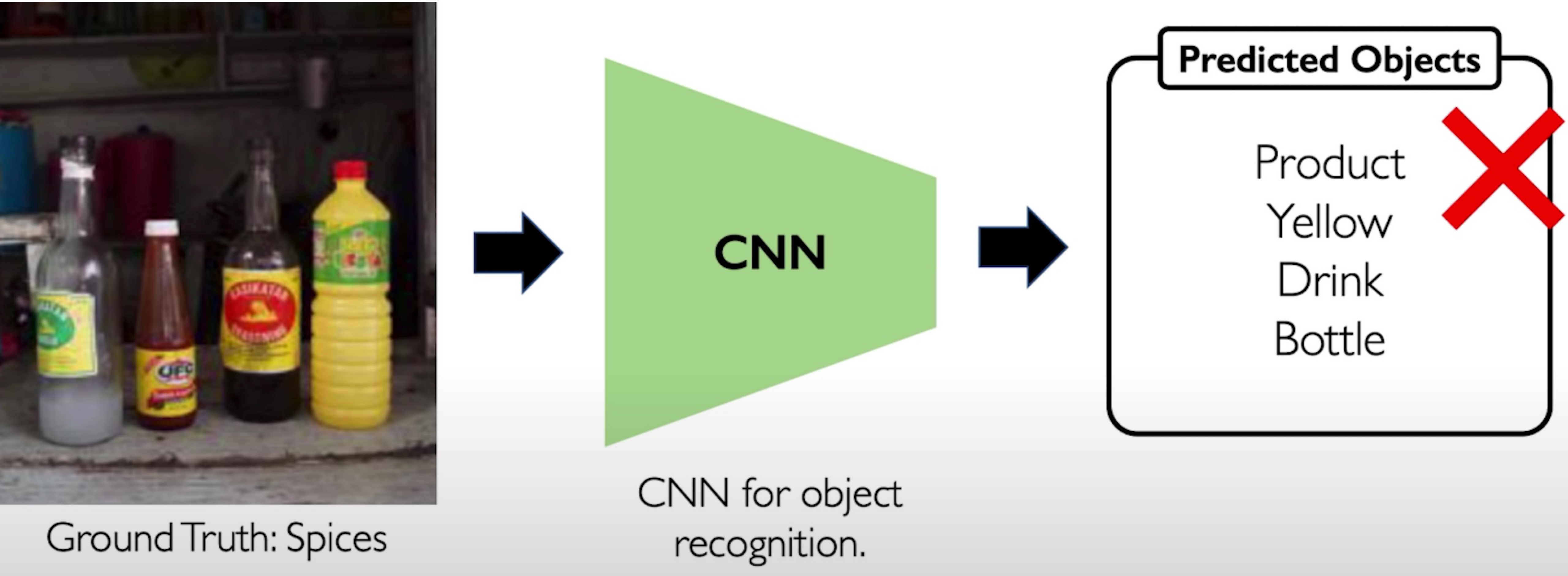


CNN for image classification.

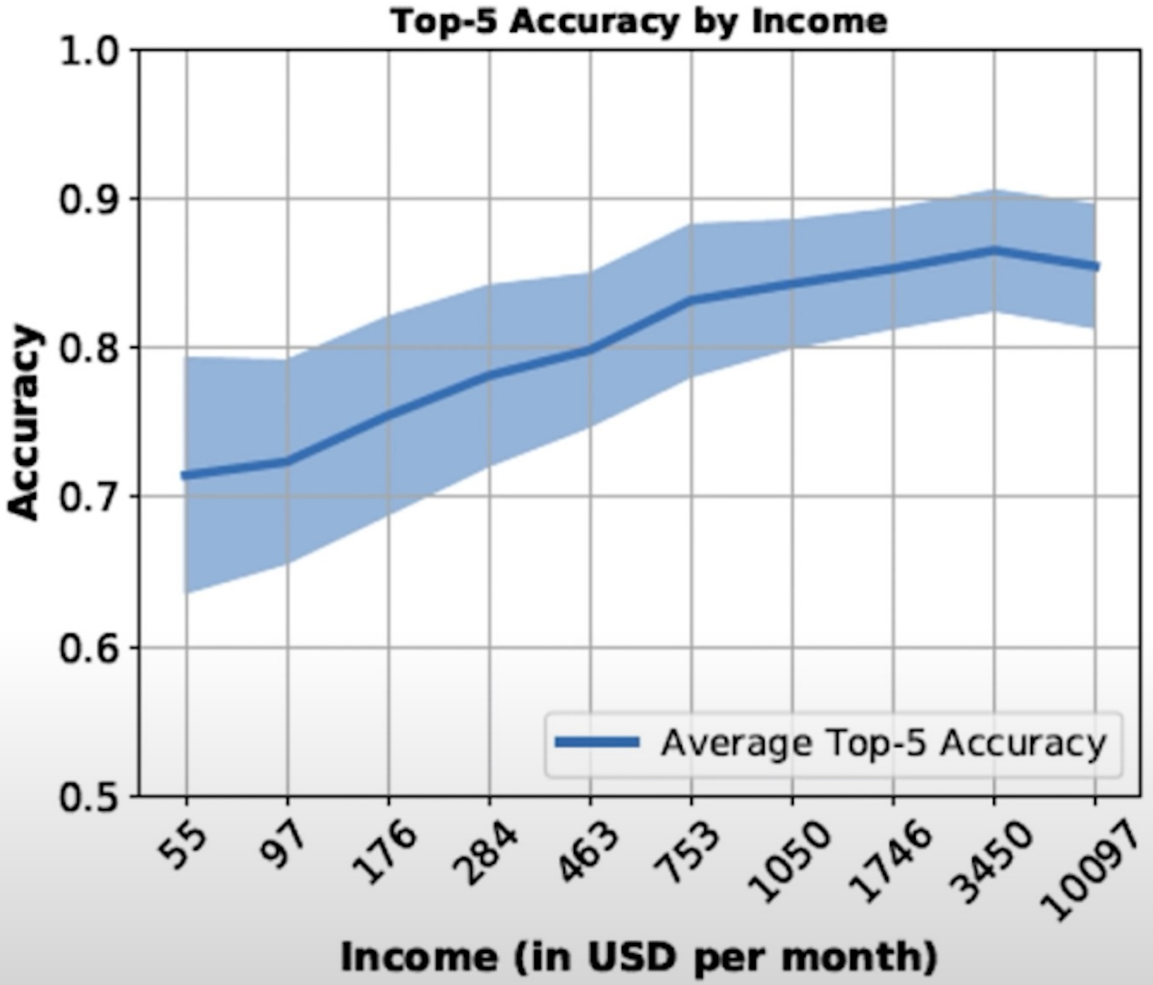
Bias in Image Classification

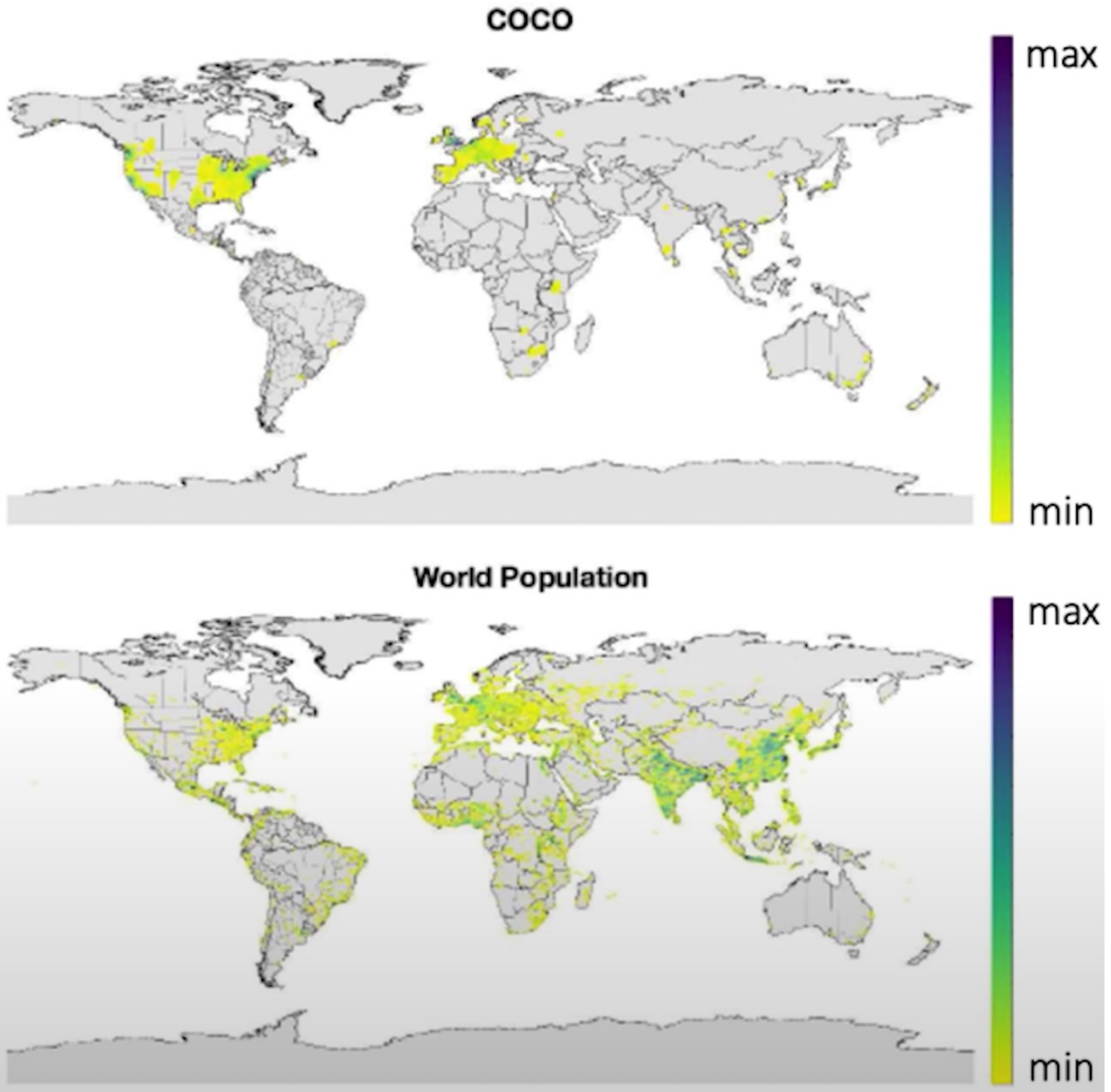


Bias in Image Classification



Bias Correlation with Income and Geography





Bias at All Stages of AI Life Cycle

- Data
- Model
- Training and Deployment
- Evaluation
- Interpretation

Taxonomy of Common Biases

Data-Driven

Selection Bias

Data selection does not reflect randomization
Ex: class imbalance

Reporting Bias

What is shared does not reflect real likelihood
Ex: news coverage

Sampling Bias

Particular data instances are more frequently sampled
Ex: hair, skin tone

Interpretation-Driven

Correlation Fallacy

Correlation != Causation

Overgeneralization

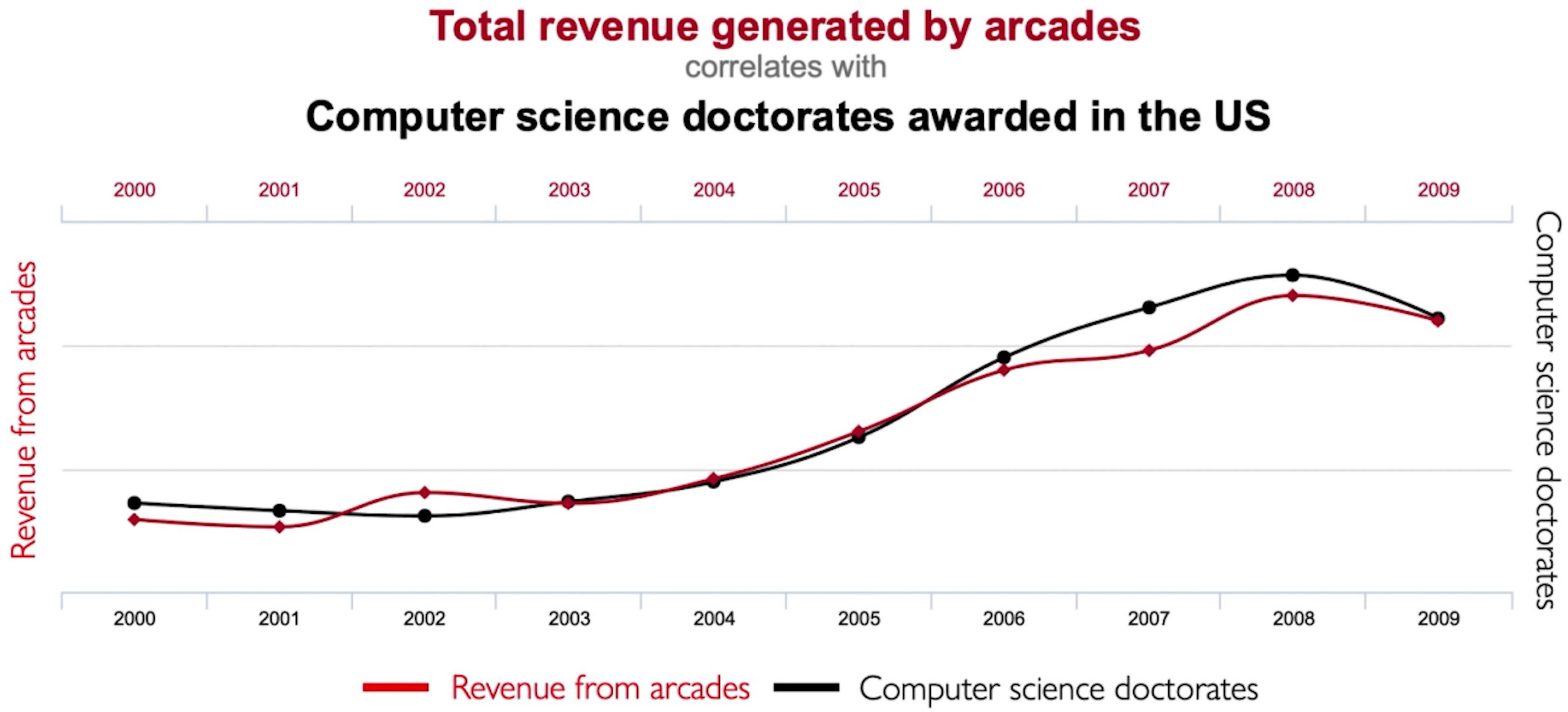
"General" conclusions drawn from limited test data

Automation Bias

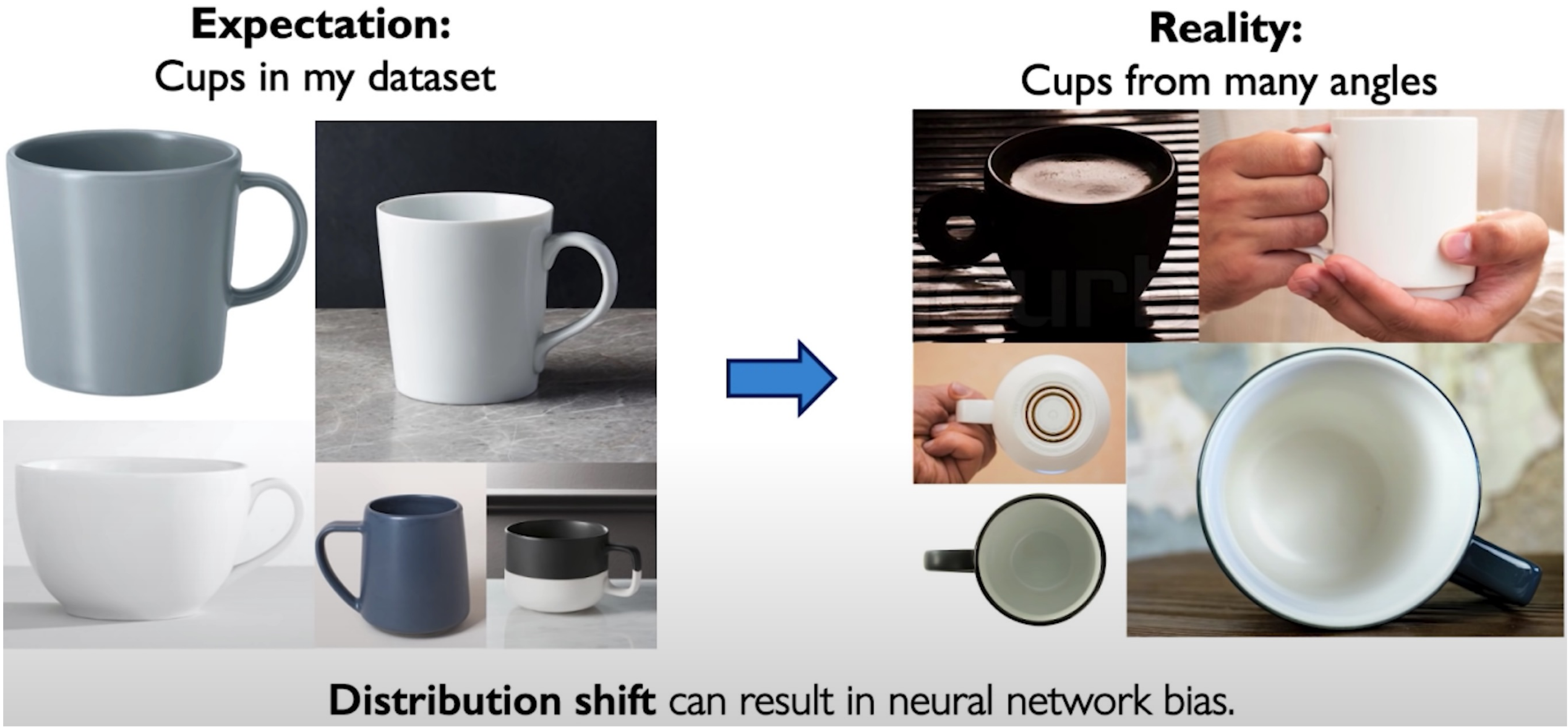
AI-generated decisions are favored over human-generation decisions

By no means an exhaustive list!

Bias from the Correlation Fallacy



Bias from Assuming Generalization



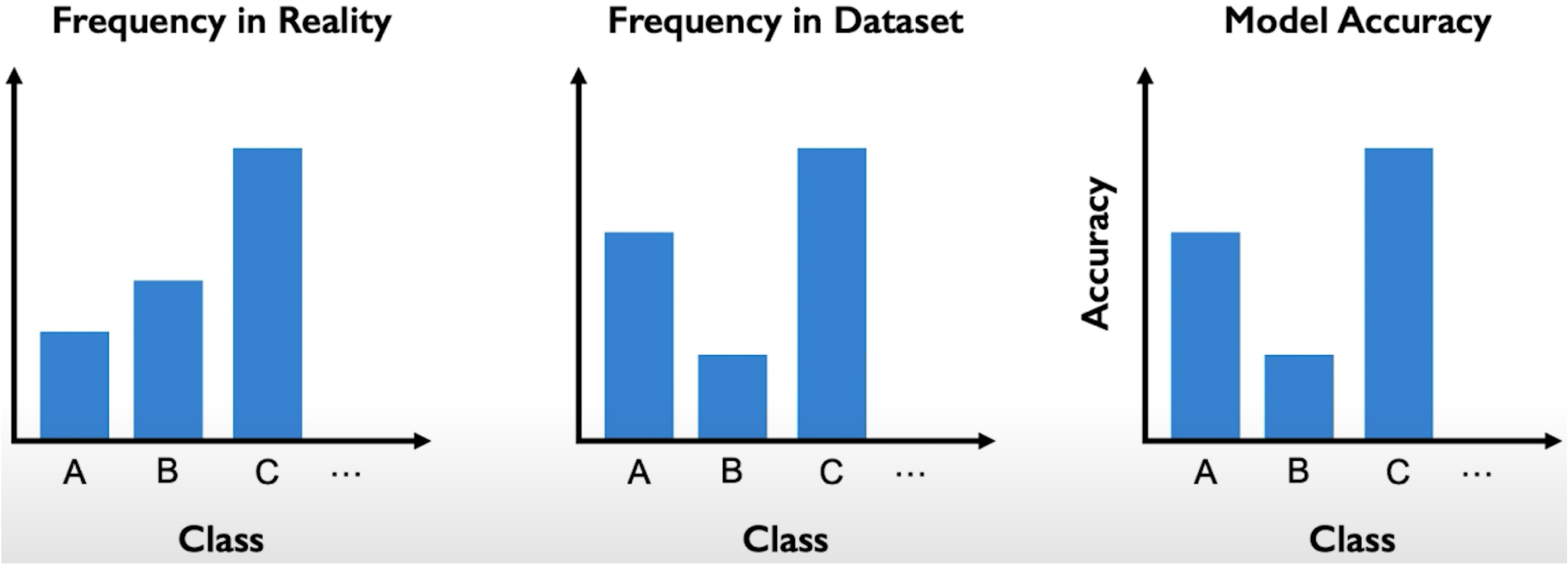
Datasets with Distribution Shift

	Train			Test	
Satellite Image (x)					
Year / Region (d)	2002 / Americas	2009 / Africa	2012 / Europe	2016 / Americas	2017 / Africa
Building / Land Type (y)	shopping mall	multi-unit residential	road bridge	recreational facility	educational institution

Task: Building / land classification

Distribution shift: Time / geographic region

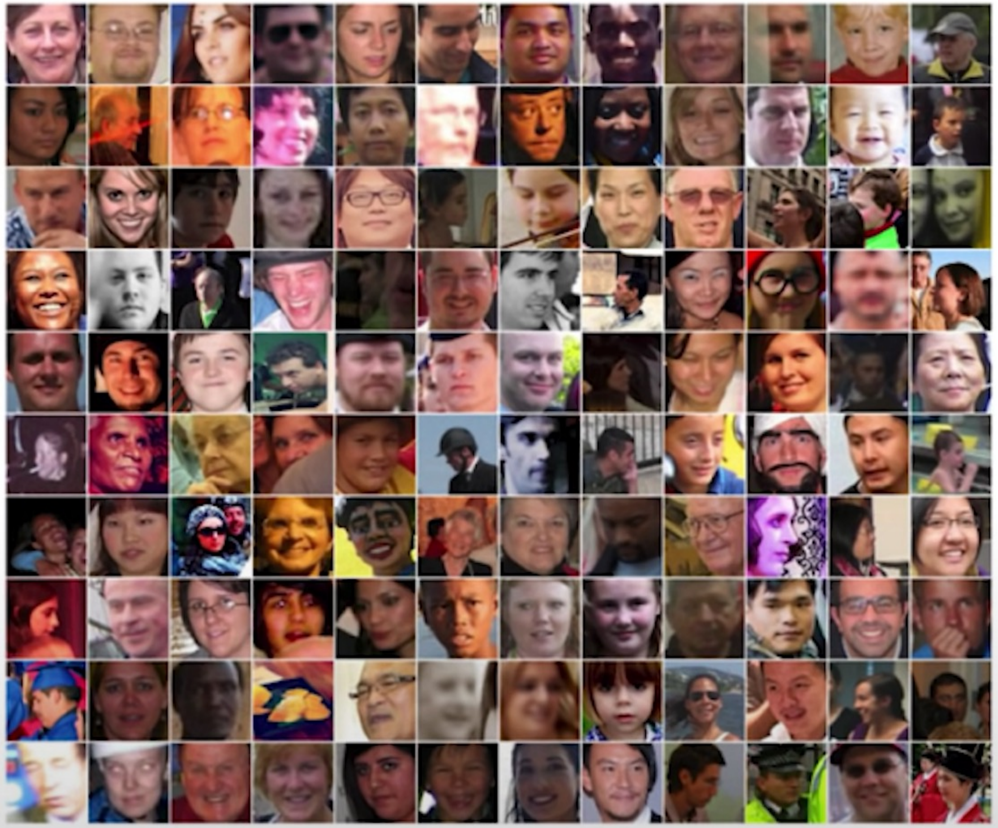
Bias due to Class Imbalance



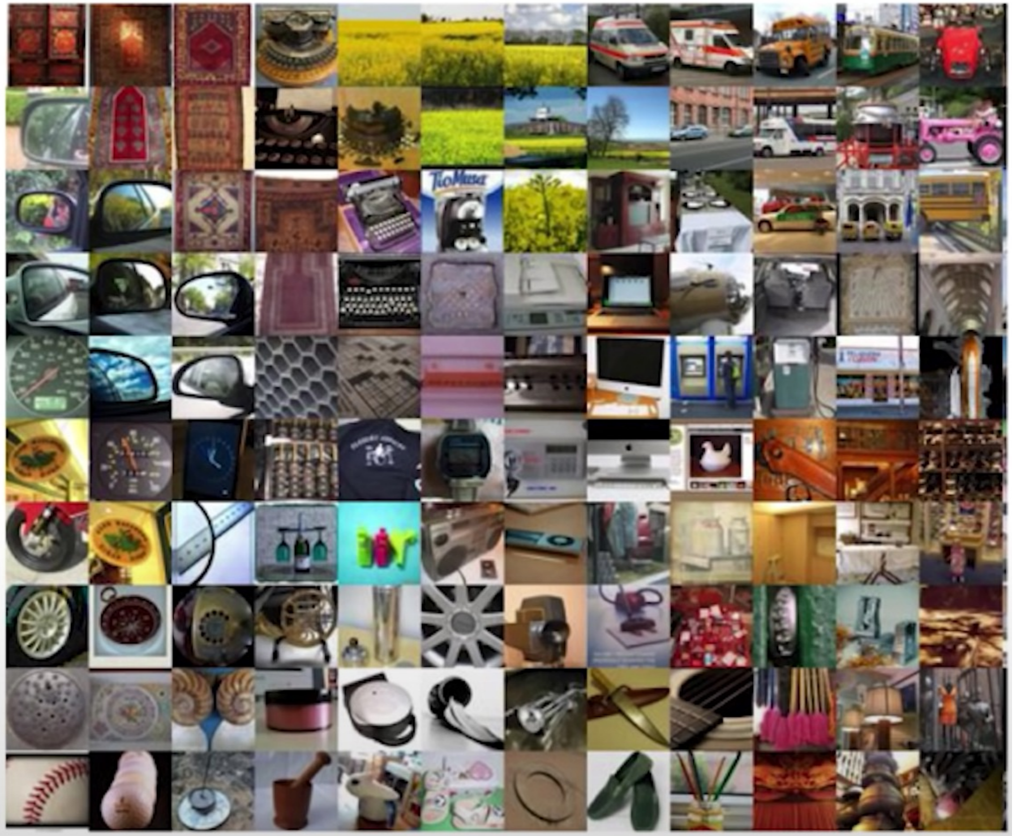
Bias in Features

Consider training a facial detection system on images of faces and images of non-faces:

Faces

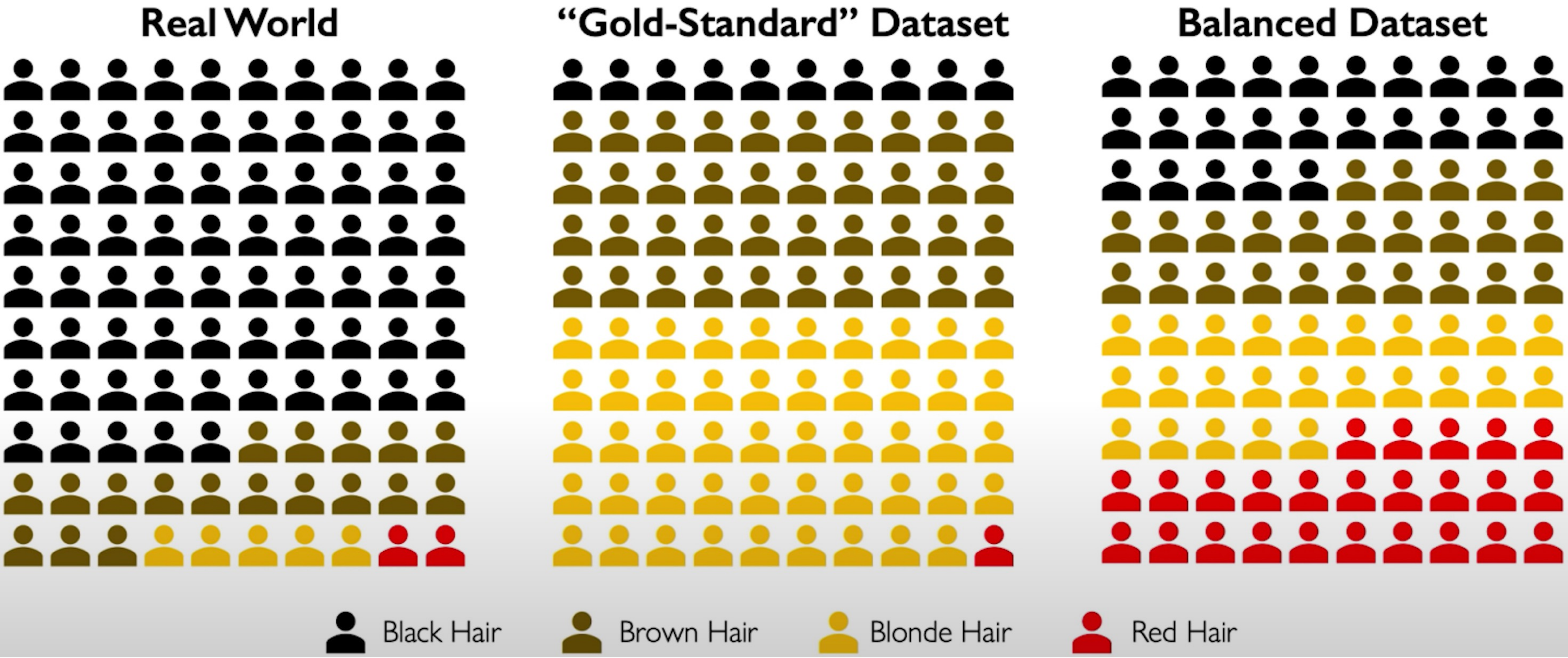


Non-Faces

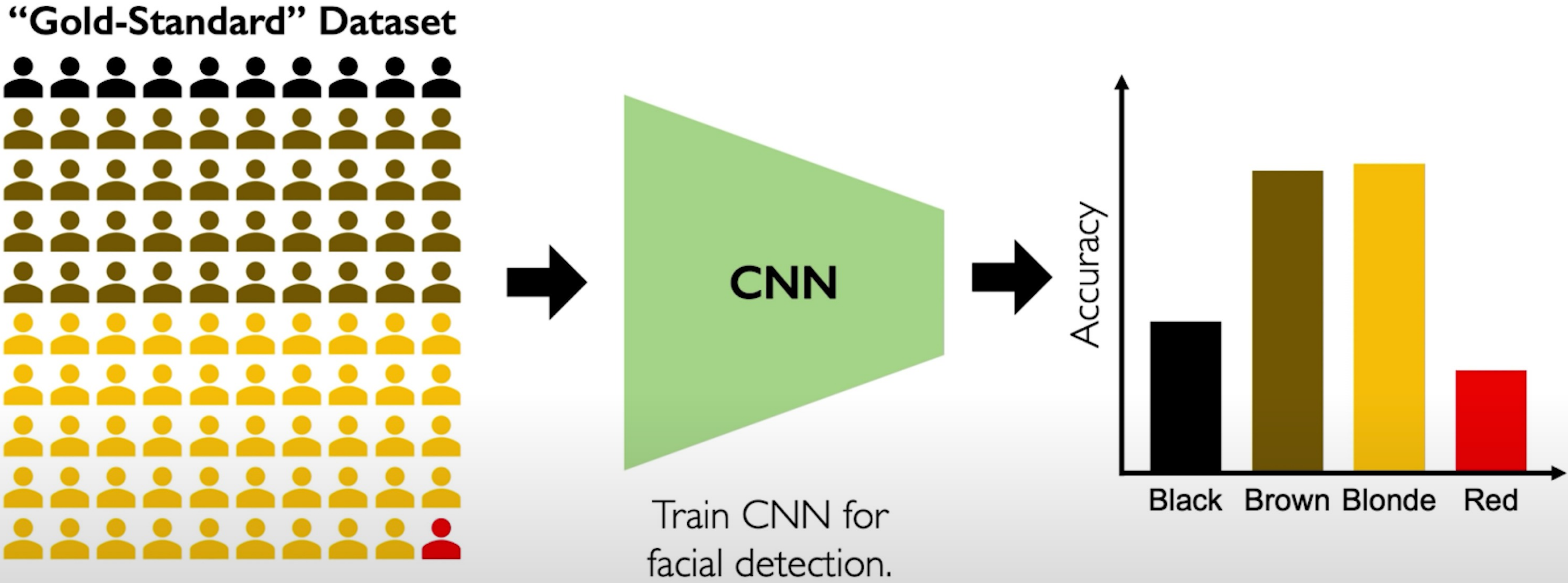


Potential biases hidden **within each class** can be even more dangerous.

Case Study: Bias in Facial Detection



Case Study: Bias in Facial Detection

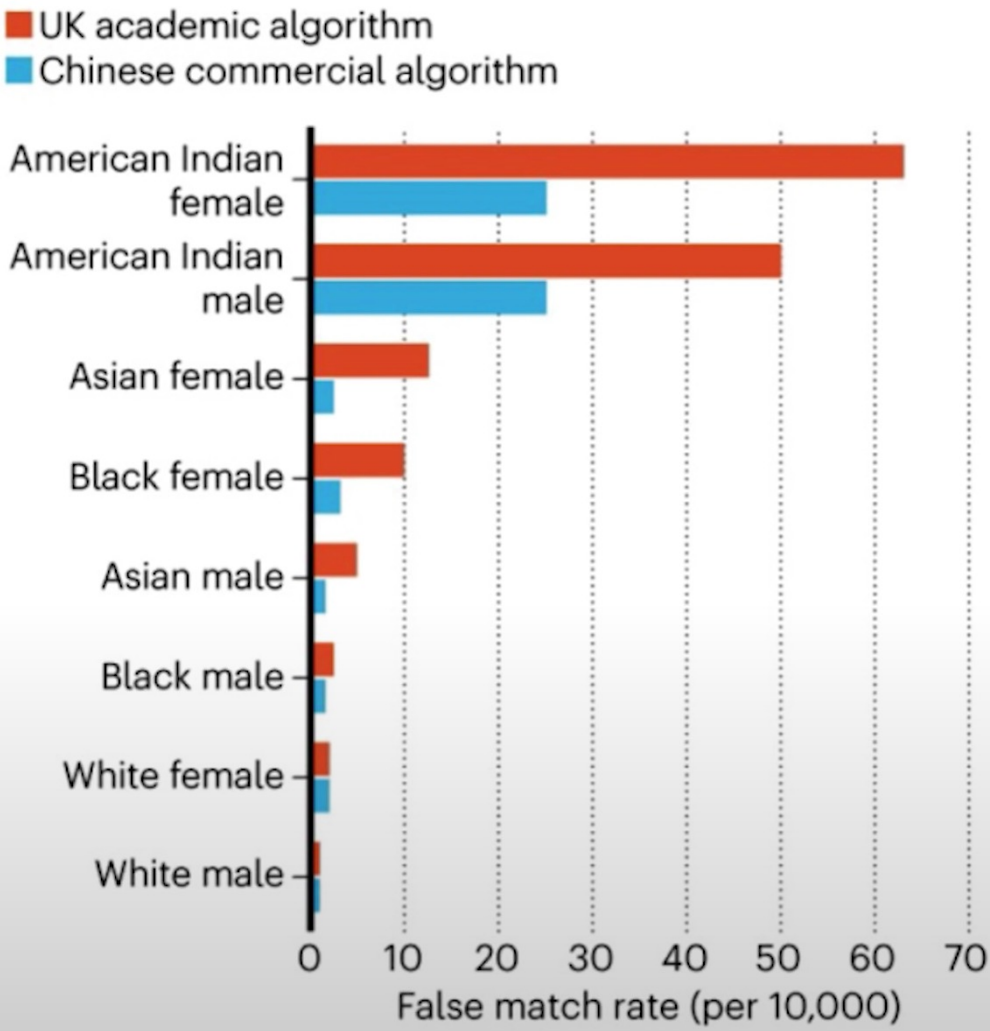


Case Study: Bias in Facial Detection

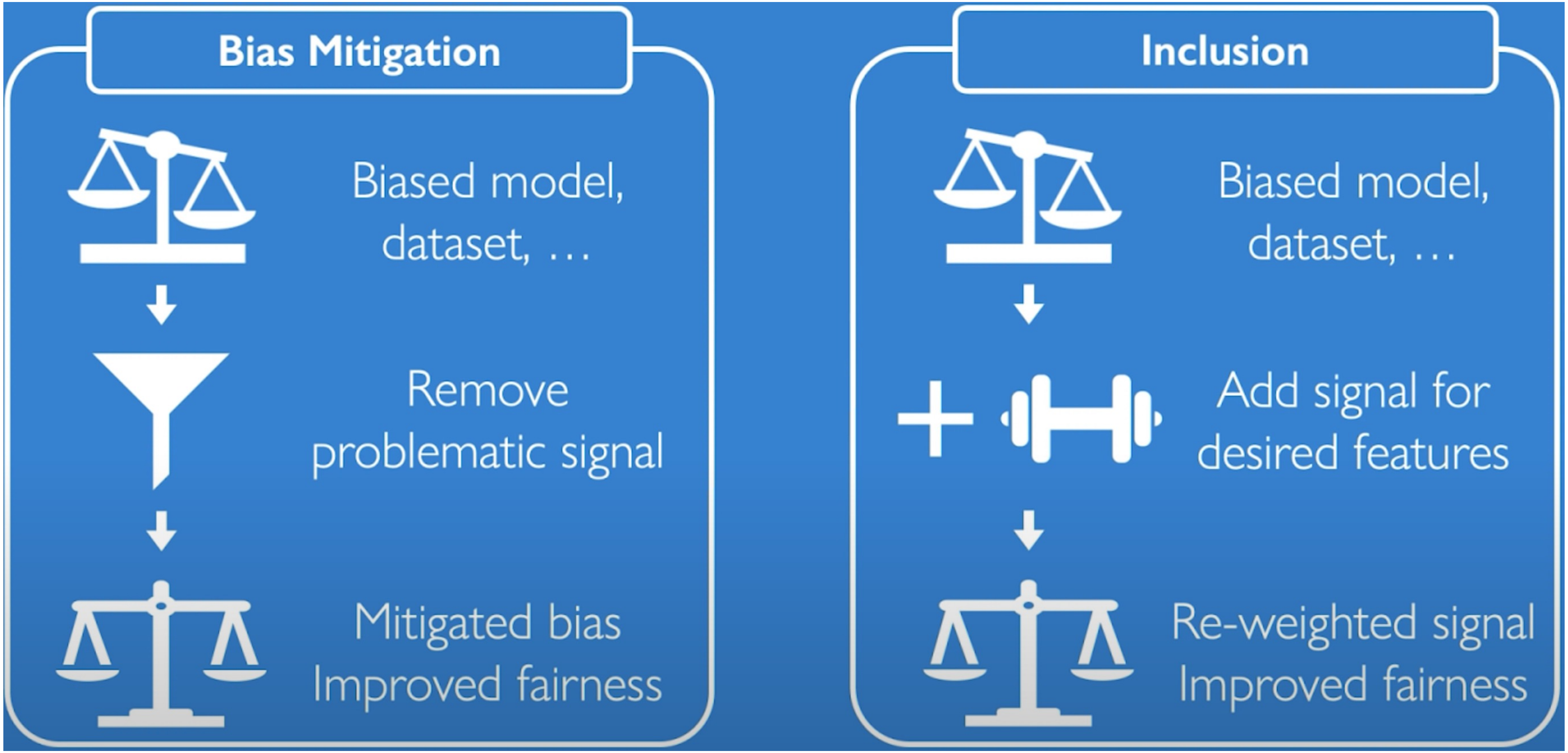
Independent Study I

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Independent Study II



Learning Techniques to Improve Fairness



Bias & Fairness in Supervised Learning

A classifier's output decision should be the **same across sensitive characteristics**, given what the correct decision should be.

A classifier, $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is **biased** if its decision changes after being exposed to additional sensitive feature inputs. It is fair with respect to variables $\boldsymbol{z}$ if:

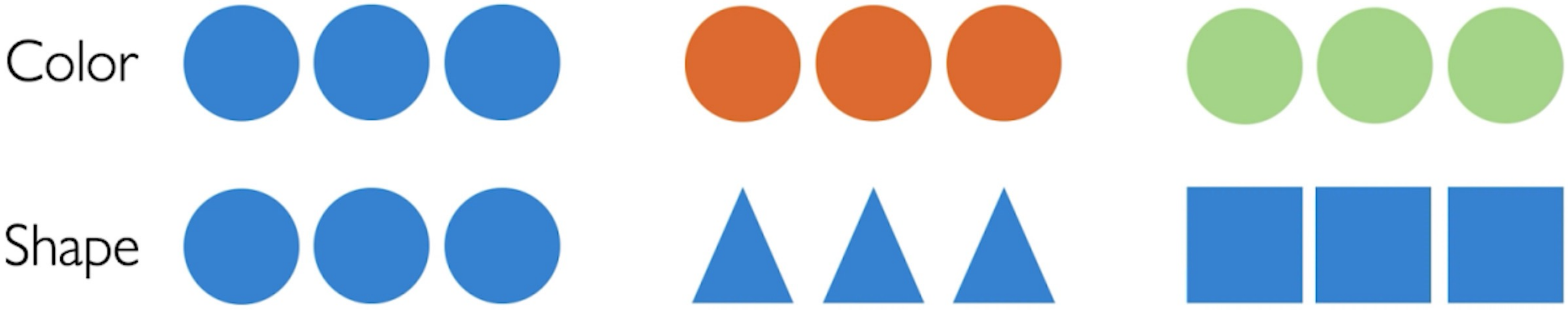
$$f_\theta(x) = f_\theta(x, z)$$

For example, for a single binary variable $\boldsymbol{z}$, fairness means:

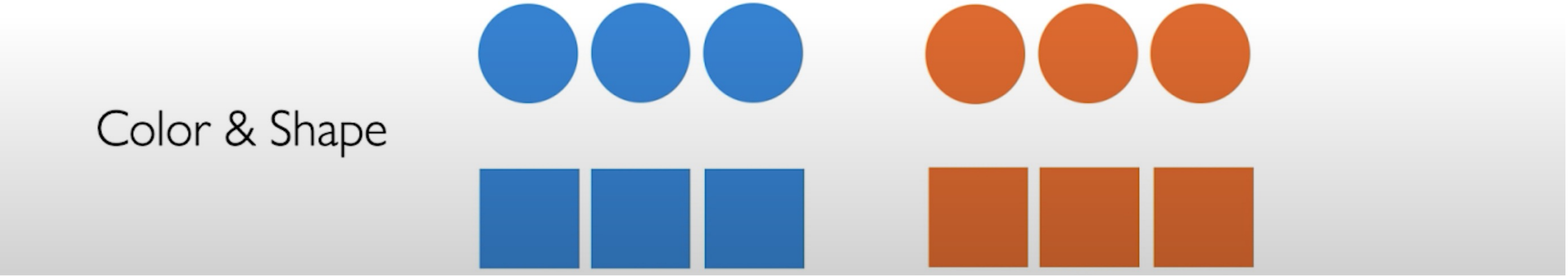
$$P[\hat{\boldsymbol{y}} = 1 | z = 0, y = 1] = P[\hat{\boldsymbol{y}} = 1 | z = 1, y = 1]$$

Evaluating Bias and Fairness

Disaggregated evaluation: evaluate performance with respect to different subgroups



Intersectional evaluation: evaluate performance with respect to subgroup intersections

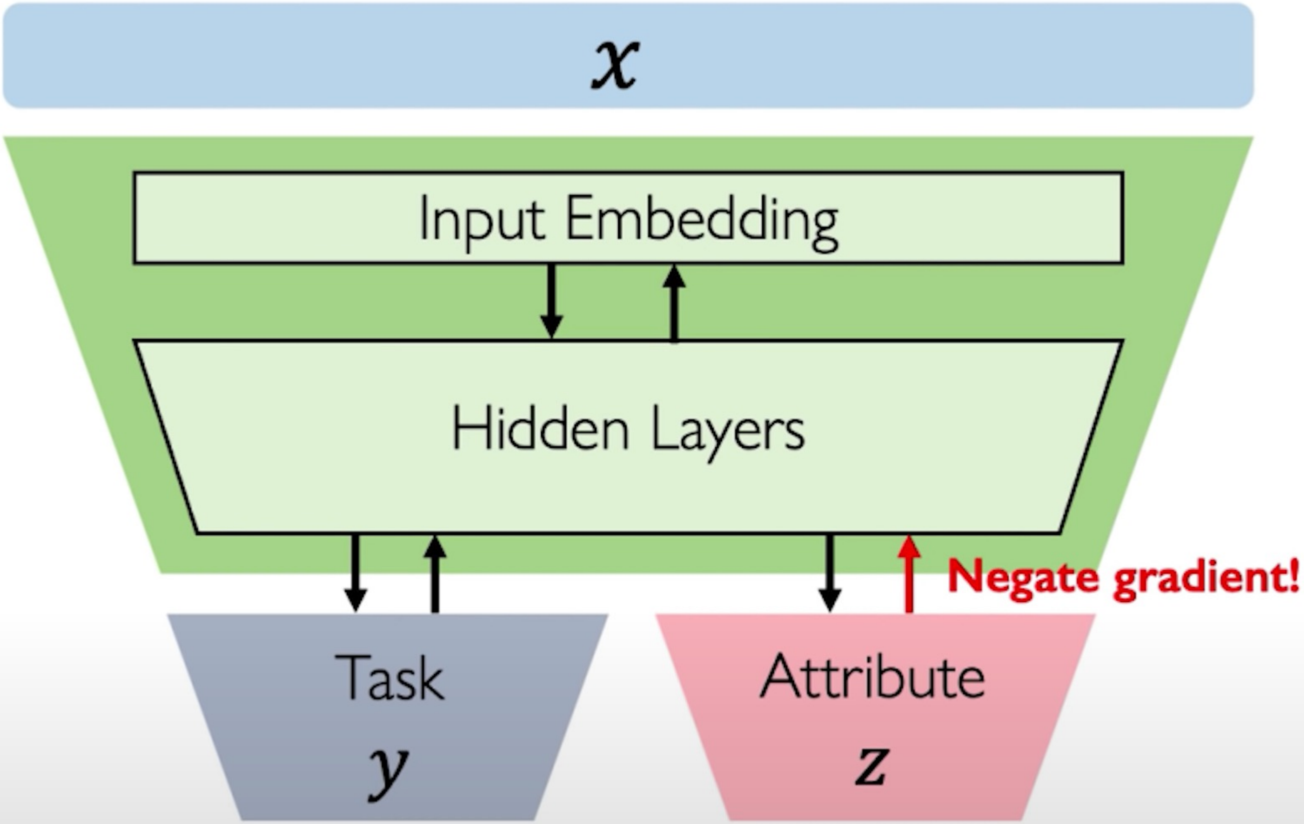


Adversarial Multi-Task Learning to Mitigate Bias

Setup: specify attribute z for which we seek to mitigate bias. Jointly predict output y and z.

Two discriminator output heads:
1. Target / class label y
2. Sensitive attribute z

Train adversarially:
1. Predict sensitive attribute z
2. Negate gradient for z head
3. "Remove" effect of z on task decision



Jointly predict output label y and sensitive attribute z to remove from decision

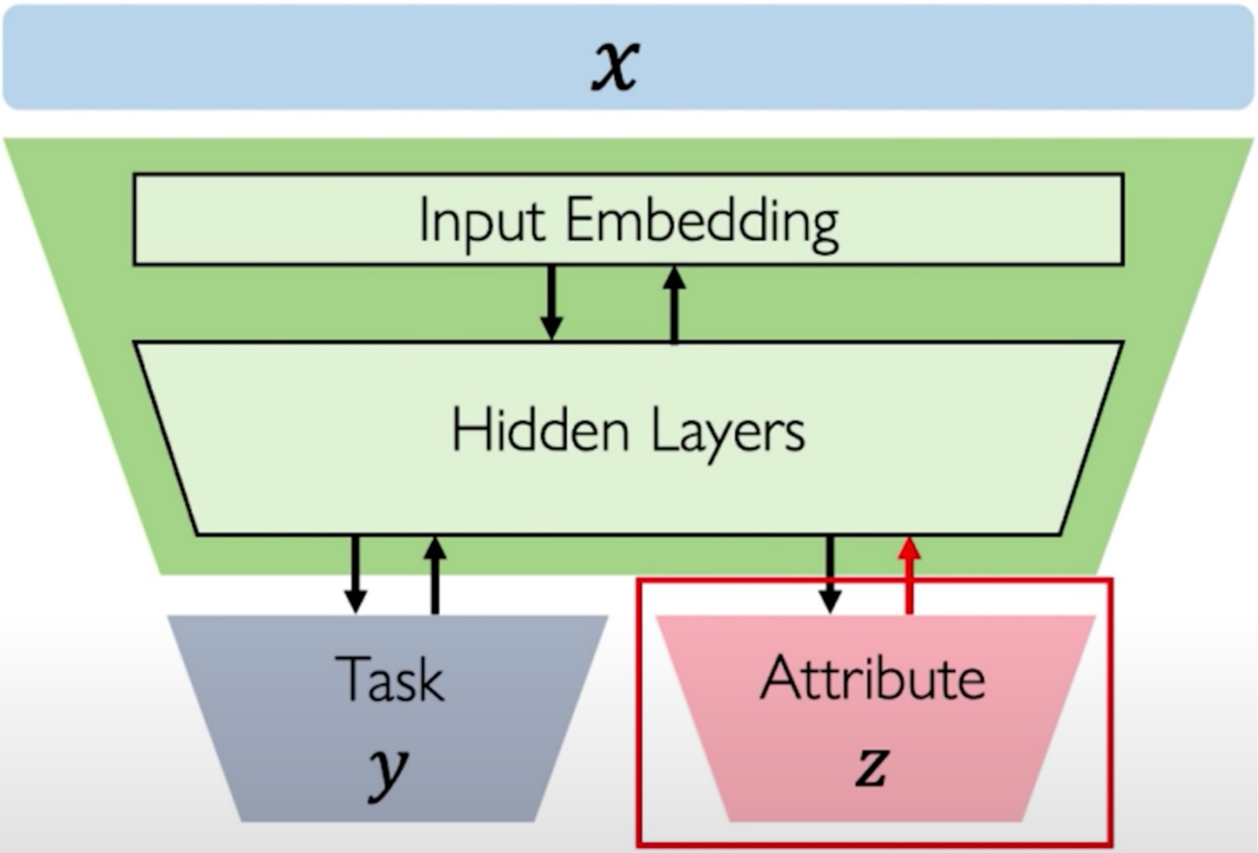
[Zhang et al. AAAI/AIES 2018]

Application to Language Modeling

Task: language model to complete analogies
He is to **she**, as **doctor** is to ?

biased		debiased	
neighbor	similarity	neighbor	similarity
nurse	1.0121	nurse	0.7056
nanny	0.9035	obstetrician	0.6861
fiancée	0.8700	pediatrician	0.6447
maid	0.8674	dentist	0.6367
fiancé	0.8617	surgeon	0.6303
mother	0.8612	physician	0.6254
fiance	0.8611	cardiologist	0.6088
dentist	0.8569	pharmacist	0.6081
woman	0.8564	hospital	0.5969

Sensitive attribute: Gender



Jointly predict output label y and sensitive attribute z to remove from decision

[Zhang et al. AAAI/AIES 2018]

Adaptive Resampling for Automatic Debiasing

Generative models can uncover the **underlying latent variables** in a dataset.

VS

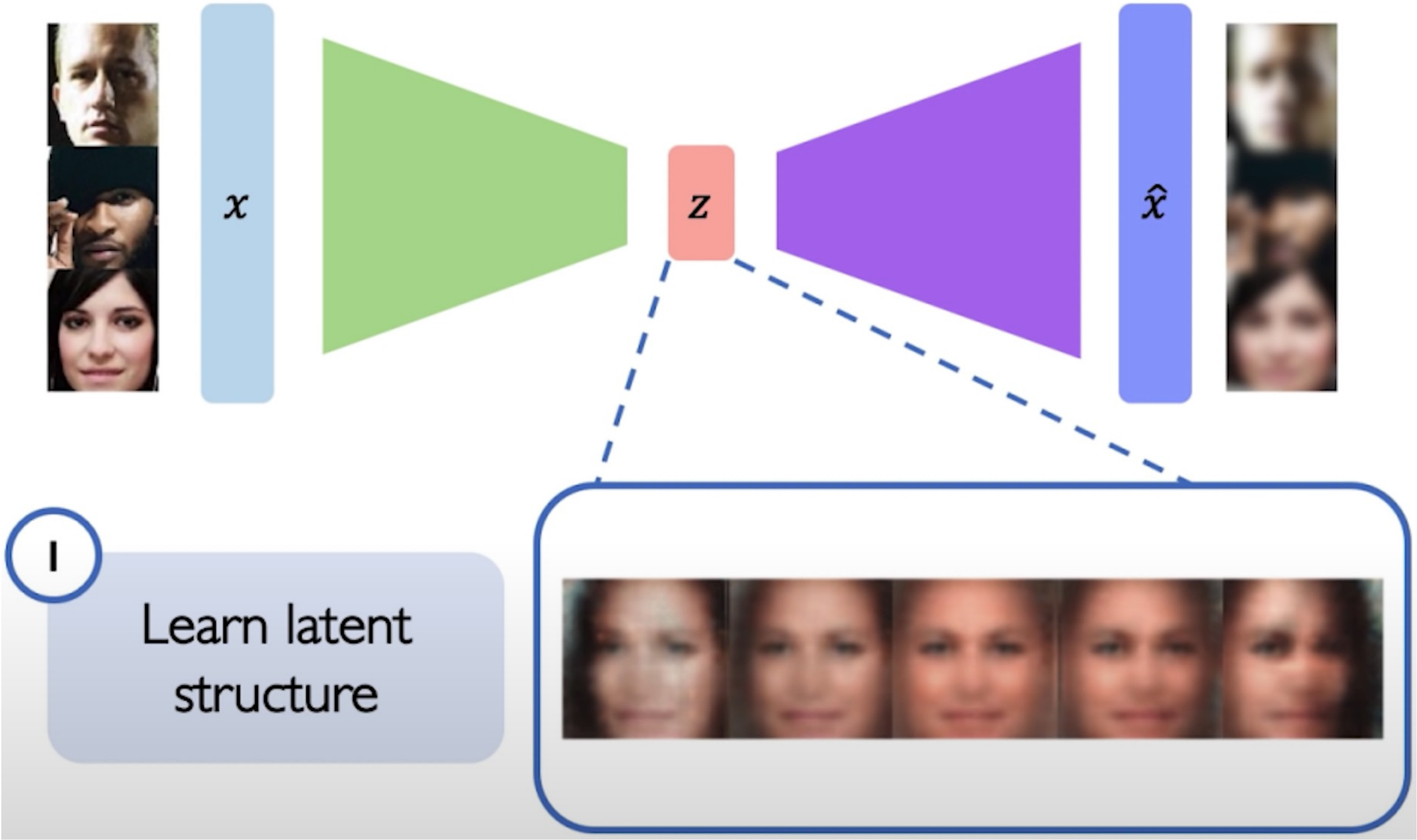
Homogeneous skin color, pose

Diverse skin color, pose, illumination

Can we use latent distributions to identify unwanted biases?

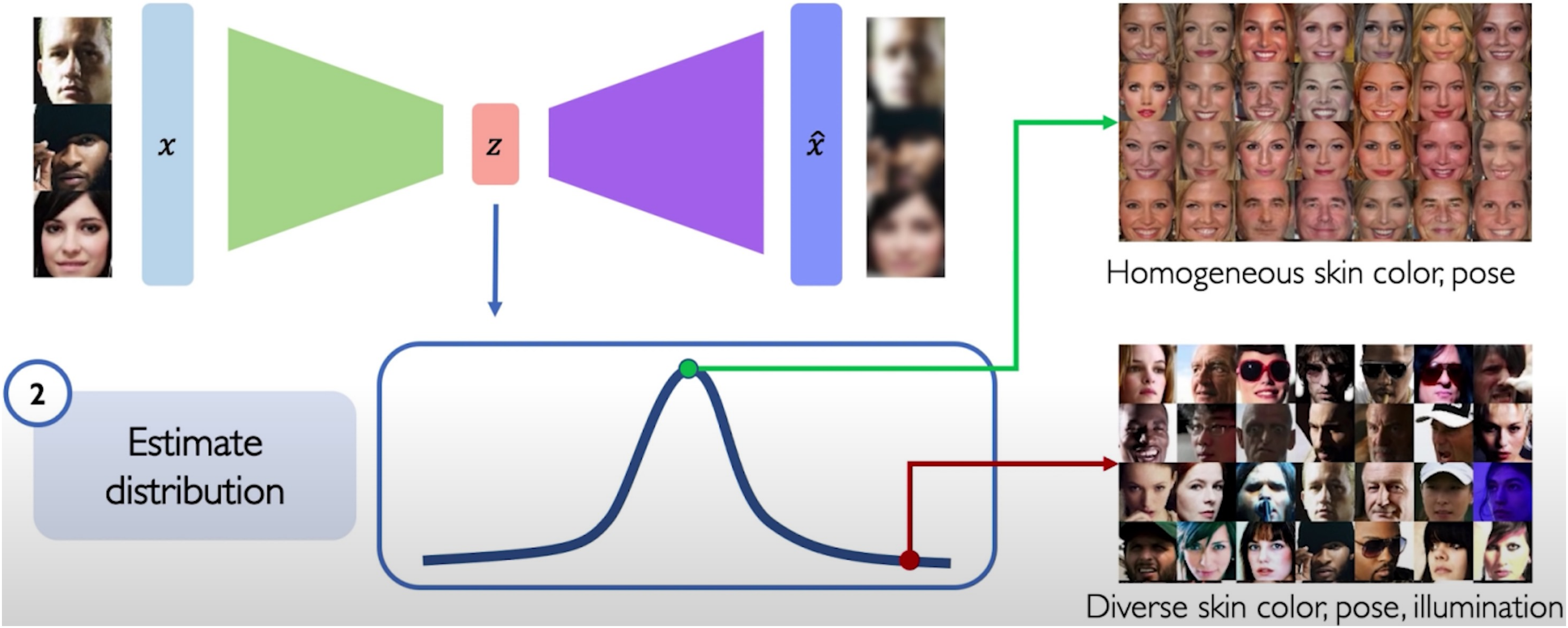
[Amini et al. AAAI/AIES 2019]

Mitigating Bias through Learned Latent Structure



[Amini et al. AAAI/AIES 2019]

Mitigating Bias through Learned Latent Structure



[Zhang et al. AAAI/AIES 2018]

Using Latent Variables for Automatic Debiasing

Approximate the distribution of the latent space with a joint histogram over the latent variables:

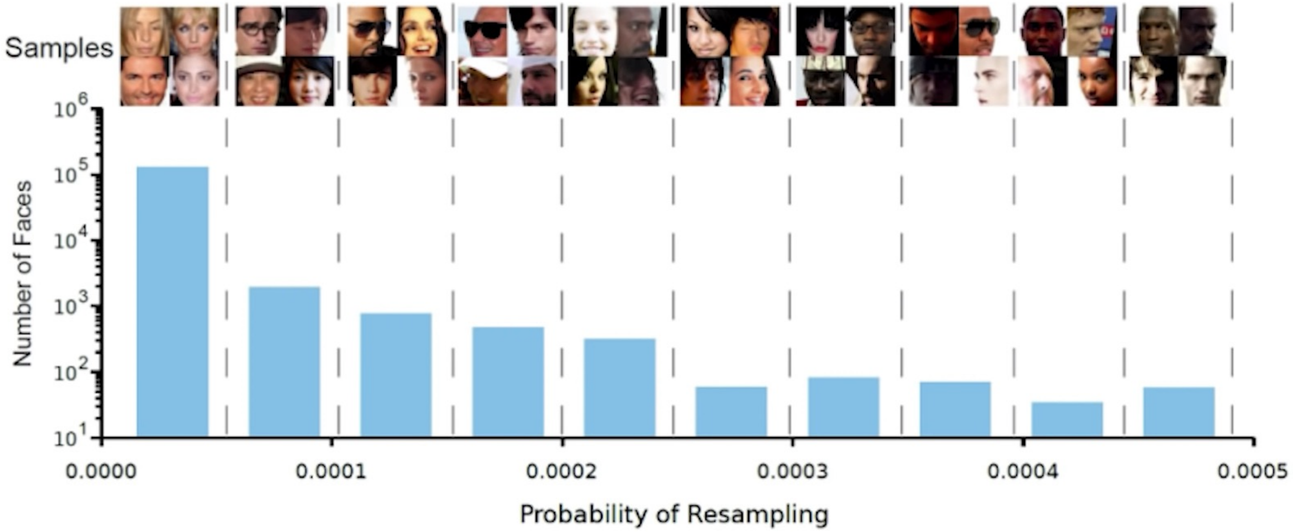
$$\underbrace{\hat{Q}(\boldsymbol{z}|X)}_{\text{Estimated joint distribution}} \propto \underbrace{\prod_i}_{\text{Independence to approximate}} \underbrace{\hat{Q}_i(\boldsymbol{z}_i|X)}_{\text{Histogram for each latent variable } z_i}$$

Define **adjusted probability** for sampling a particular datapoint $\boldsymbol{x}$ during training:

$$\underbrace{W(\boldsymbol{z}(\boldsymbol{x})|X)}_{\text{Probability of selecting datapoint}} \propto \prod_i \frac{1}{\underbrace{\hat{Q}_i(\boldsymbol{z}_i(\boldsymbol{x})|X)}_{\text{Histogram for each latent variable } z_i} + \underbrace{\alpha}_{\text{Debiasing parameter}}}$$

[Zhang et al. AAAI/AIES 2018]

Adaptive Adjustment of Resampling Probability



Top 10 faces with Lowest Resampling Probability

Top 10 faces with Highest Resampling Probability

Random Batch Sampling During Standard Face Detection Training

Homogenous skin color, pose

Mean Sample Prob: 7.57 x 10^{-6}

Batch Sampling During Training with Learned Debiaising

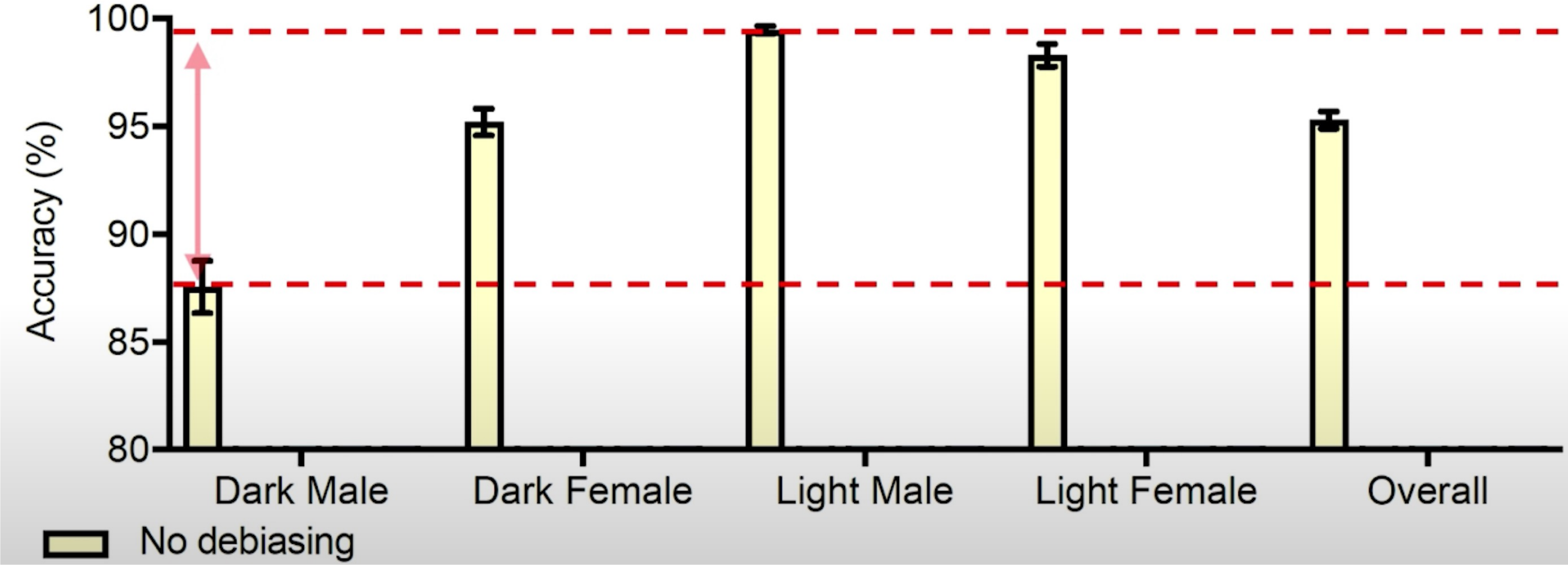
Diverse skin color, pose, illumination

Mean Sample Prob: 1.03 x 10^{-4}

Adaptive resampling based on automatically **learned features →** no need to specify attributes to debias against!

Evaluation: Decreased Categorical Bias

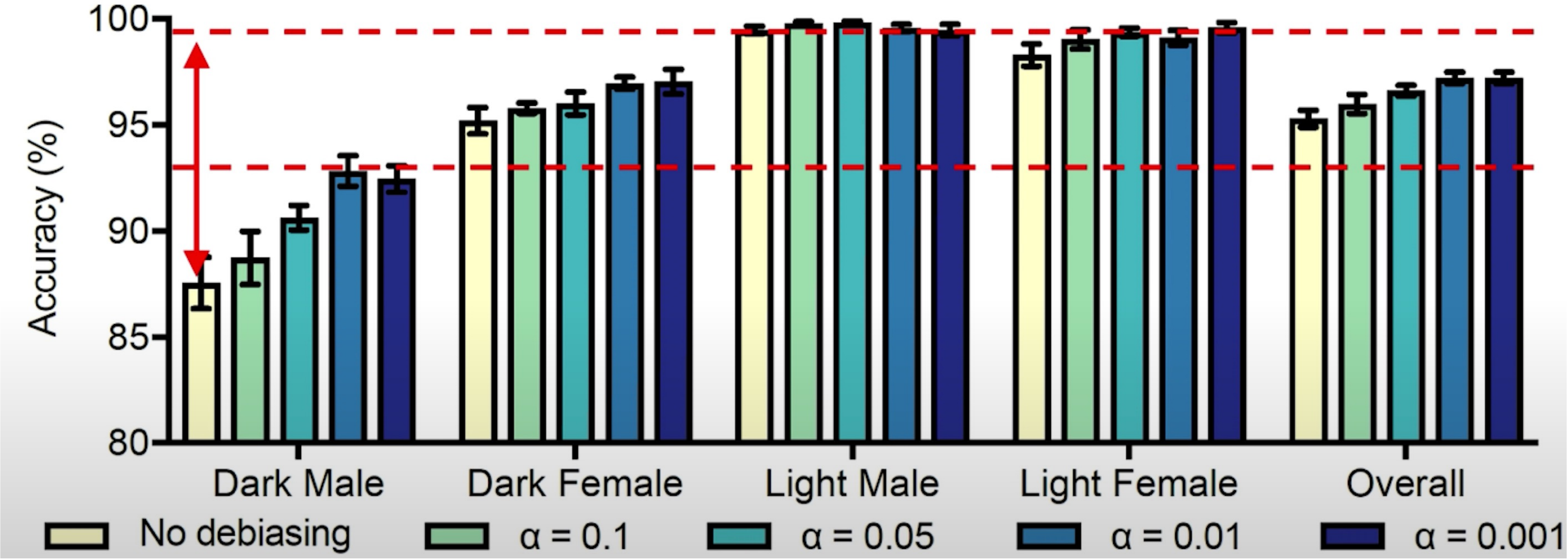
Disaggregated and intersectional evaluation: evaluate performance across subgroups and combinations of subgroups



[Zhang et al. AAAI/AIES 2018]

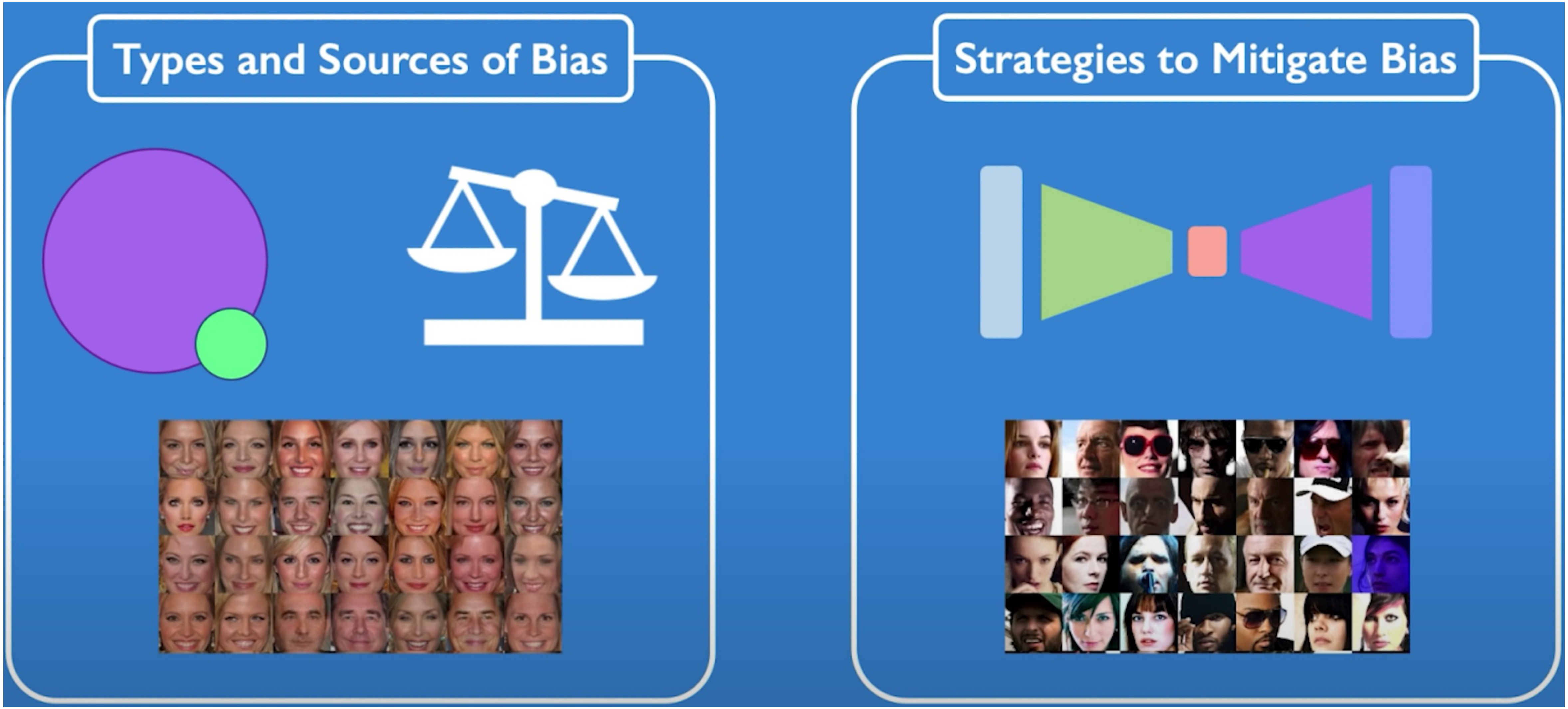
Evaluation: Decreased Categorical Bias

Disaggregated and intersectional evaluation: evaluate performance across subgroups and combinations of subgroups



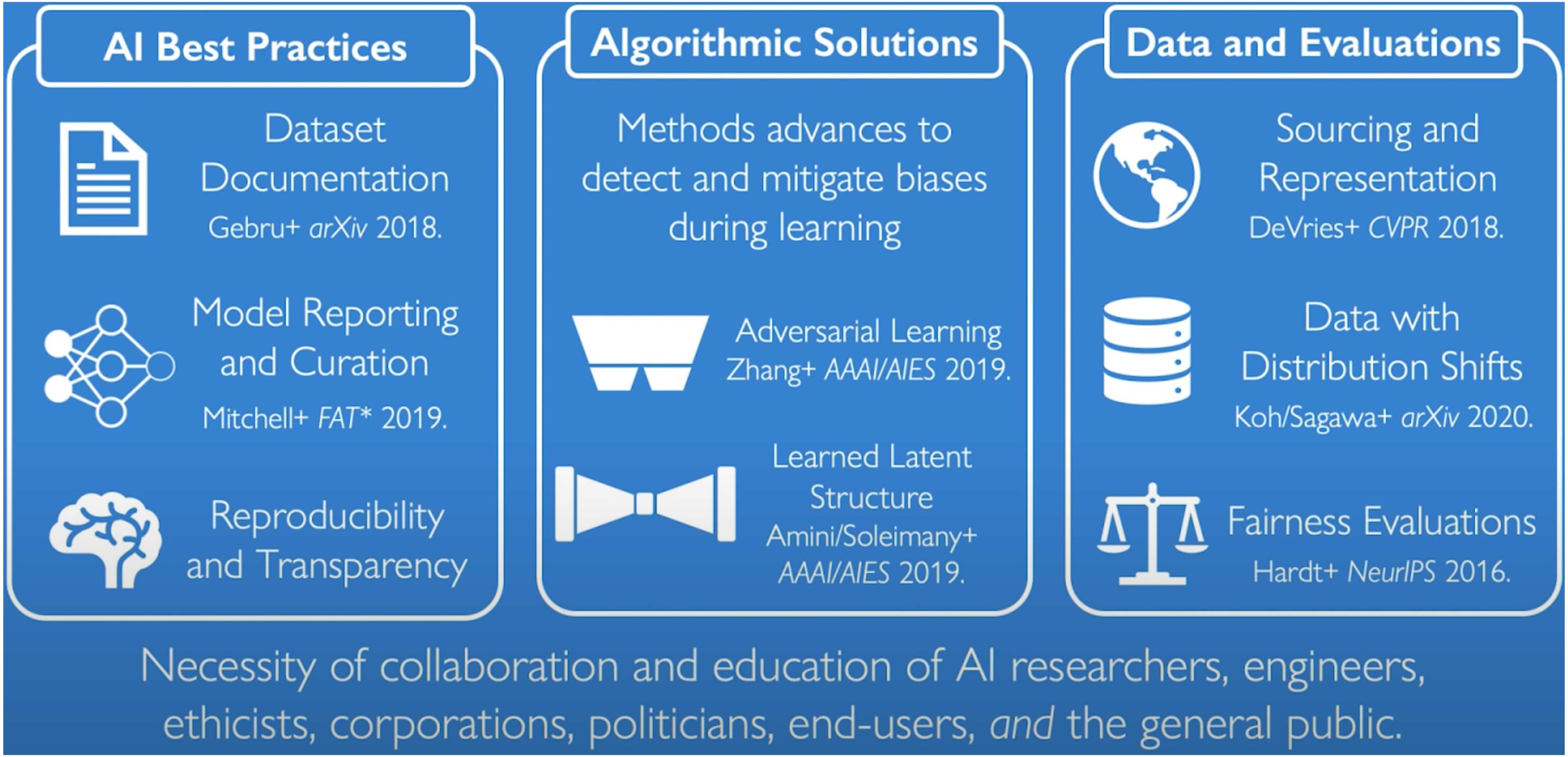
[Zhang et al. AAAI/AIES 2018]

Understanding and Mitigating Algorithmic Bias



[Zhang et al. AAAI/AIES 2018]

AI Fairness: Summary and Future Consideration



[Zhang et al. AAAI/AIES 2018]

Interesting Papers at ICLR 2024

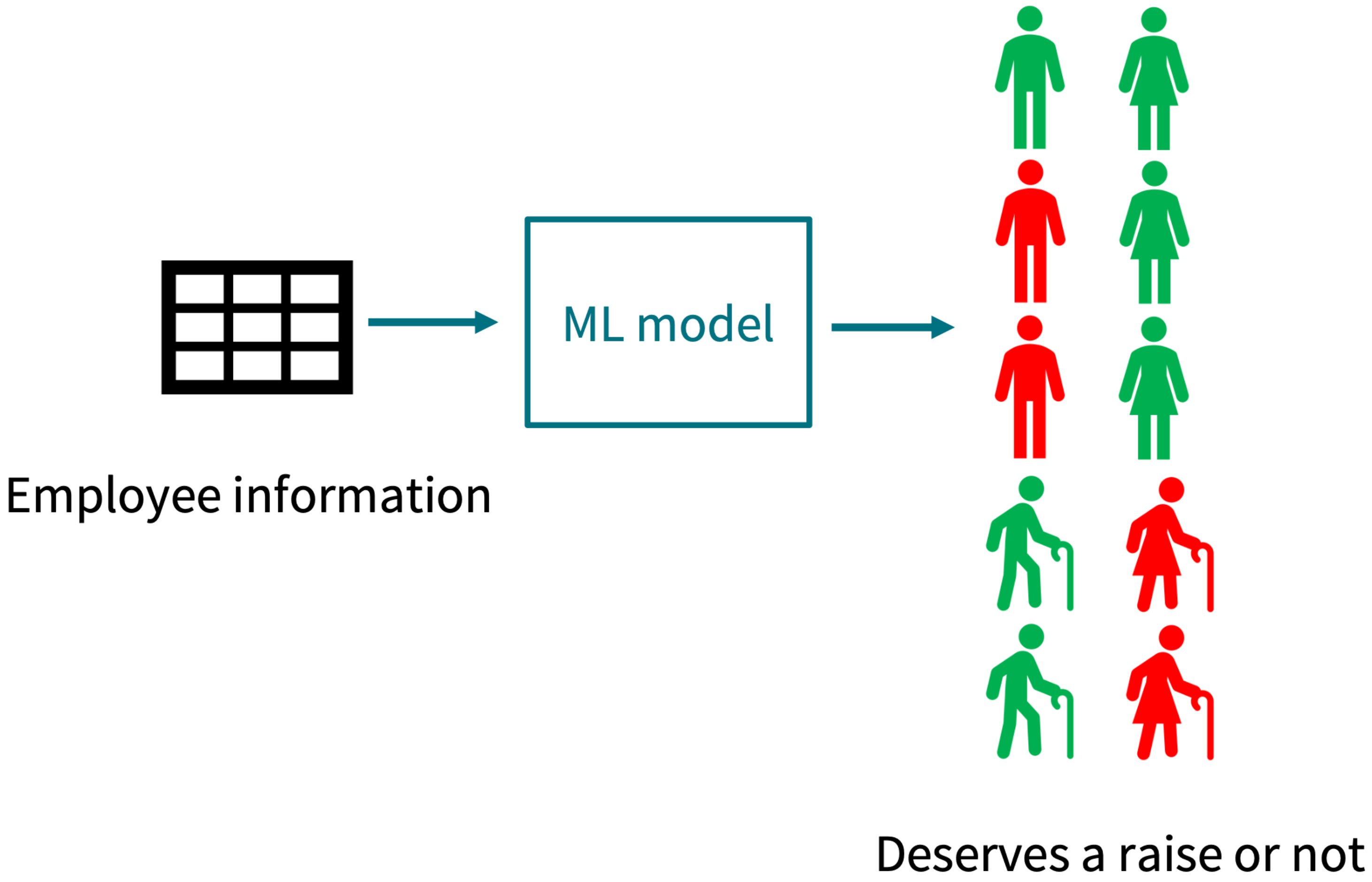
ICLR 2024 Test of Time Award

- Winner: **Auto-Encoding Variational Bayes**
- Runner Up: **Intriguing properties of neural networks**

ON THE FAIRNESS ROAD: ROBUST OPTIMIZATION FOR ADVERSARIAL DEBIASING

Vincent Grari[*,1,2,4], **Thibault Laugel**[*,1,2,4], **Tatsunori Hashimoto**[2], **Sylvain Lamprier**[3], **Marcin Detyniecki**[1,4,5]
[1] AXA Group Operations
[2] Stanford University
[3] LERIA, Université d'Angers, France
[4] TRAIL, Sorbonne Université, Paris, France
[5] Polish Academy of Science, IBS PAN, Warsaw, Poland
{grari,laugel}@stanford.edu
code: https://github.com/axa-rev-research/ROAD-fairness/

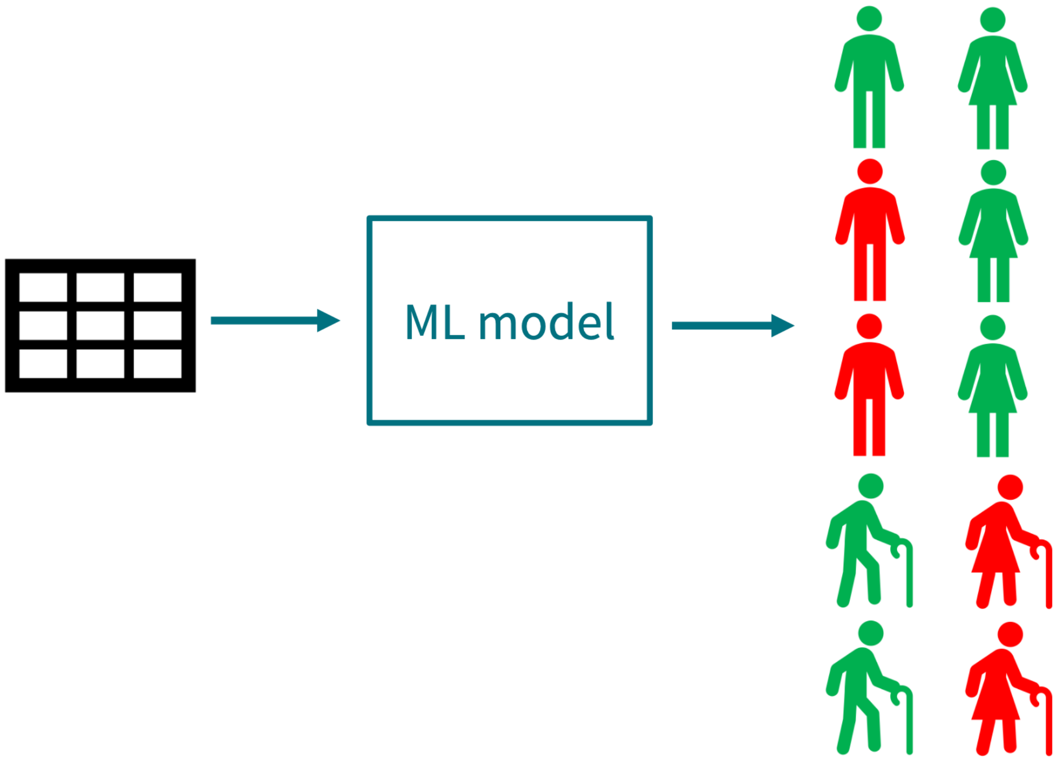
Group Fairness



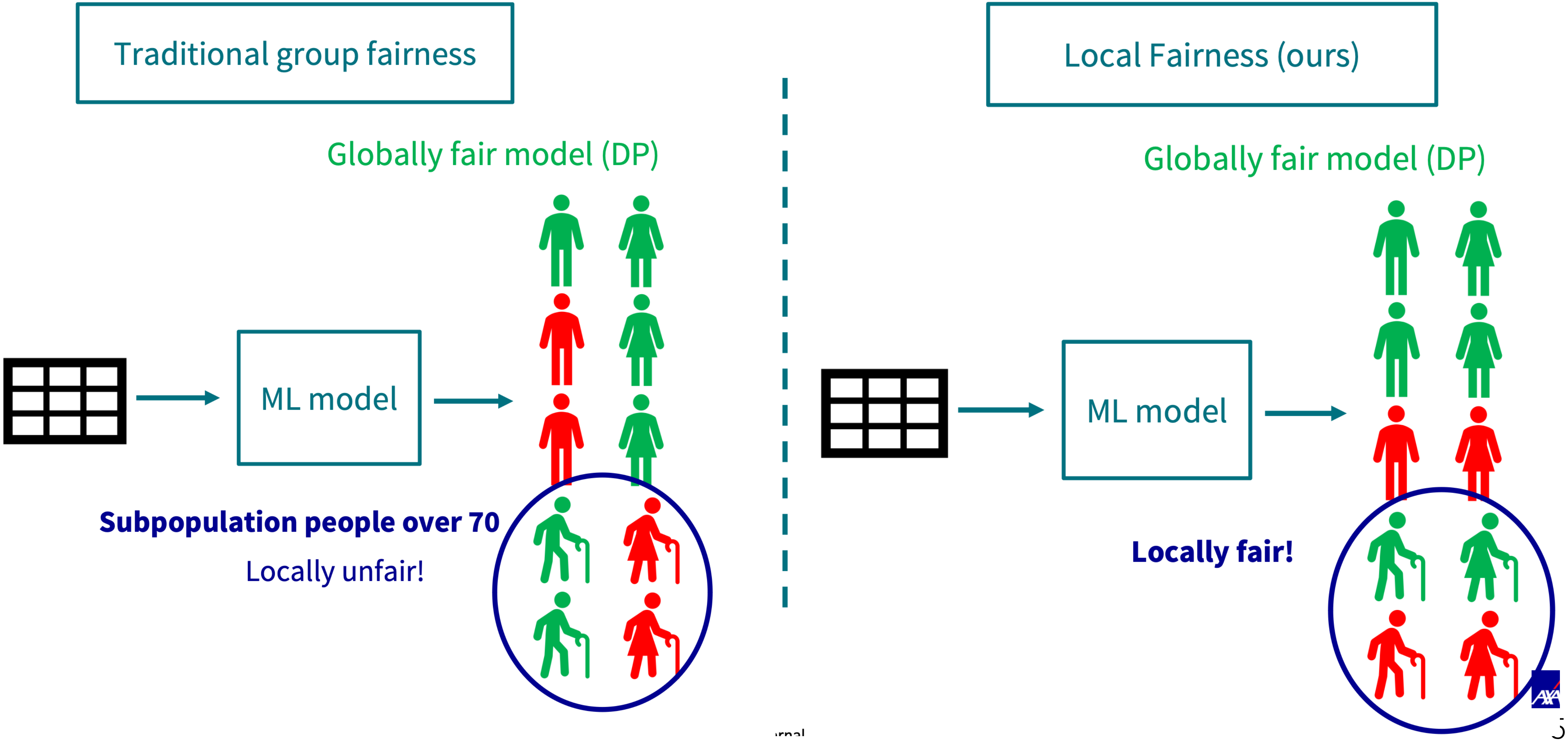
Group Fairness

Traditional group fairness

Globally fair model (DP): $\mathbb{P}(\hat{Y} = 1 | S = 1) = \mathbb{P}(\hat{Y} = 1 | S = 0)$

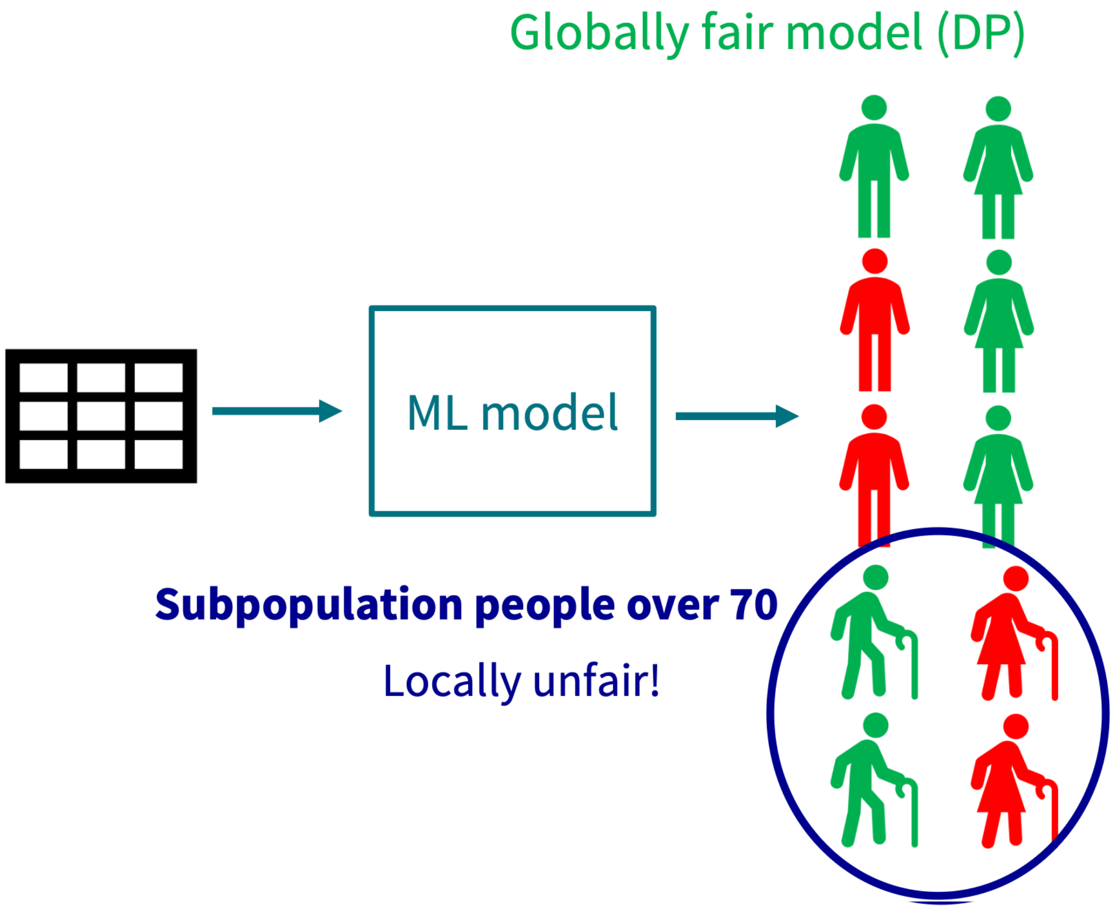


The Local (Un)fairness Problem

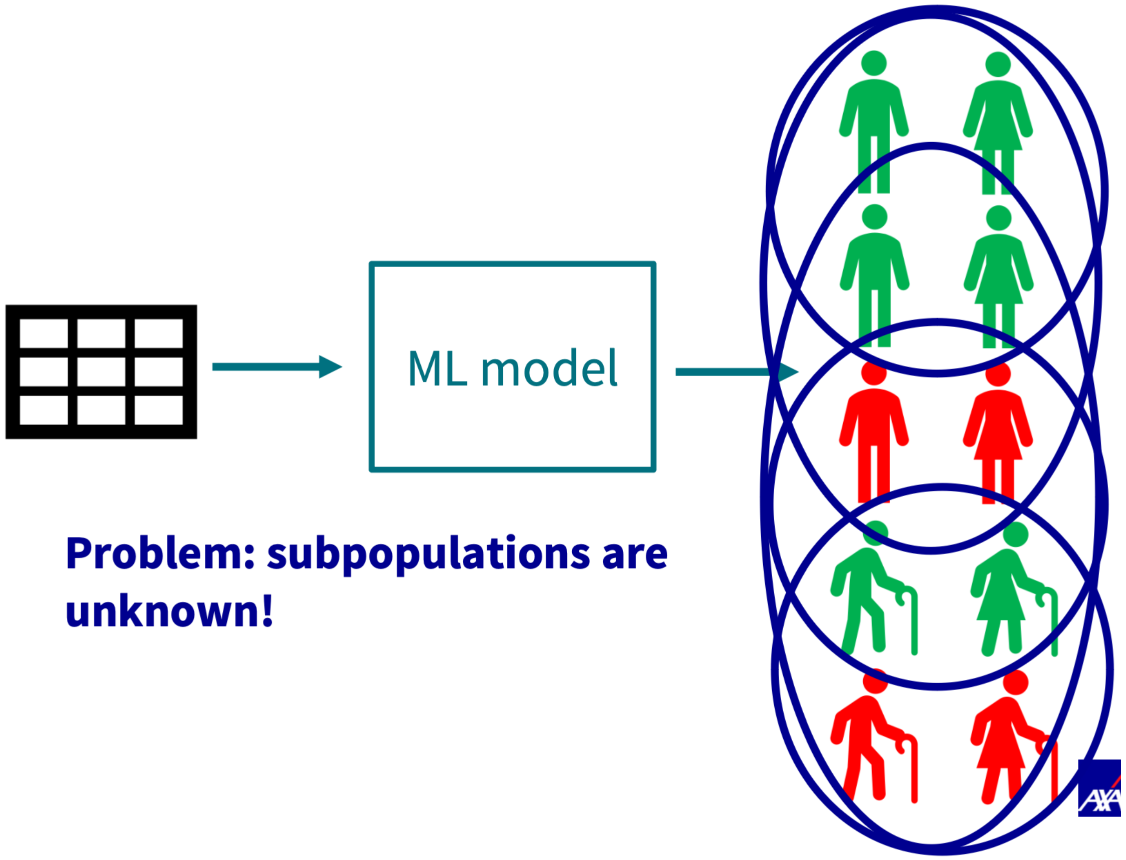


The Local (Un)fairness Problem

Traditional group fairness



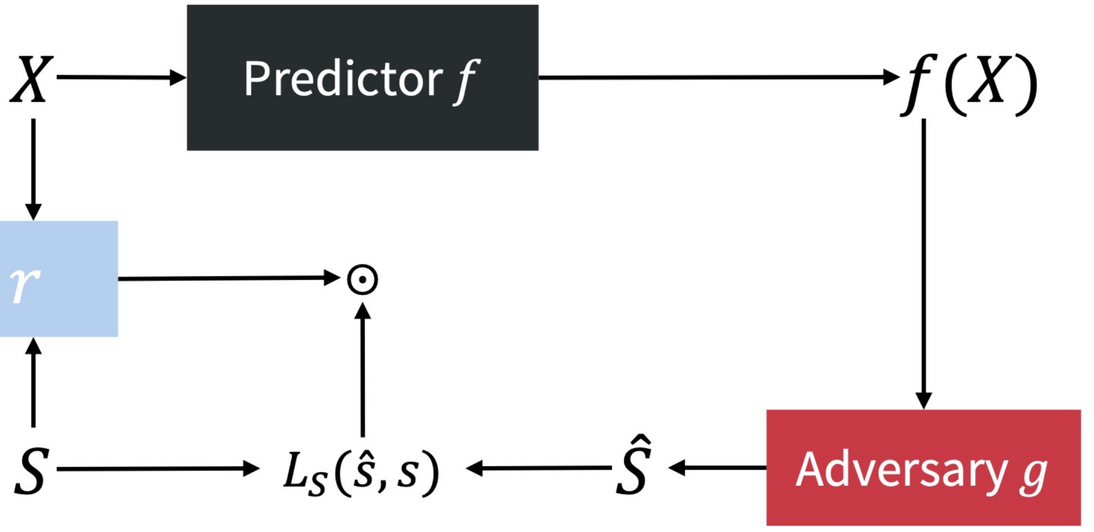
Local Fairness (ours)



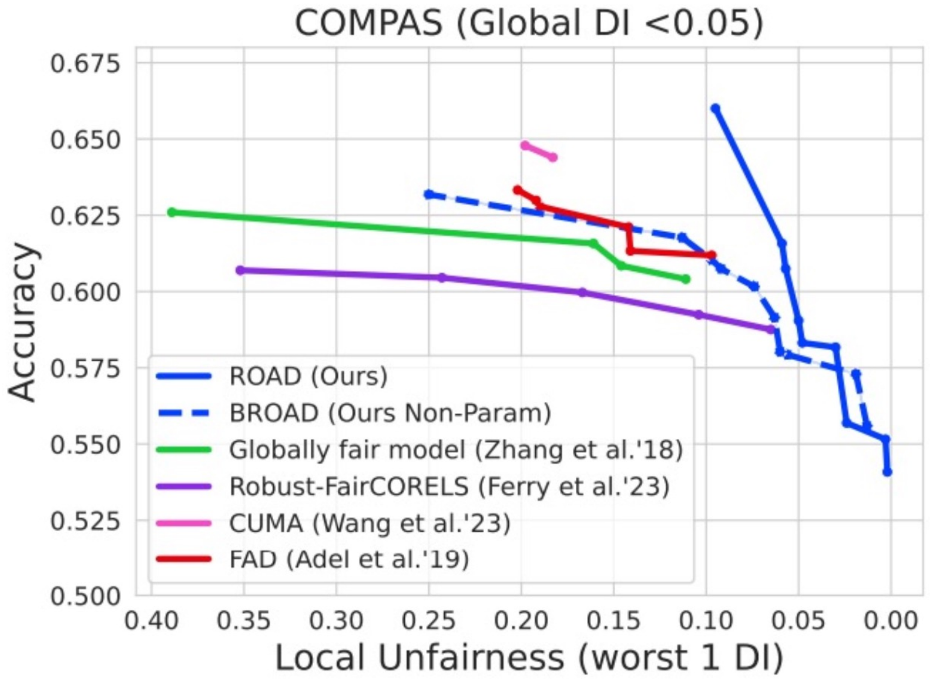
Internal

Distributionally Robust Optimization (DRO) for Fairness

$$L_Y(f(x), y) - \lambda r(x,s) L_s(\hat{s}, s) + KL\ constraint$$



Classical adversarial approach for fairness



Results: more fair locally for the same levels of group fairness and accuracy

Distributionally Robust Optimization (DRO) for Fairness

Traditional group fairness

$$\min_{w_f} \mathbb{E}_p[L_Y\left(f_{w_f}(x), y\right)]$$
$$s.t.\, DI_{(x,s)\sim p}\left(f_{w_f}(x), s\right) < \epsilon$$

Local Fairness (ours)

$$\min_{w_f} \mathbb{E}_p[L_Y\left(f_{w_f}(x), y\right)]$$
$$s.t. \max_{q\in Q} DI_{(x,s)\sim q}\left(f_{w_f}(x), s\right) < \epsilon$$

Q : set of "plausible" distributions
~set of subpopulations

In practice: KL divergence-ball around p

THE DEVIL IS IN THE NEURONS: INTERPRETING AND MITIGATING SOCIAL BIASES IN PRE-TRAINED LANGUAGE MODELS

Yan Liu[♦] **Yu Liu**[♣] **Xiaokang Chen**[♥] **Pin-Yu Chen**[★] **Daoguang Zan**[♣]
Min-Yen Kan[▶] **Tsung-Yi Ho**[♦]
[♦]**Chinese University of Hong Kong** [♥]**Peking University**
[▶]**National University of Singapore** [♣]**Microsoft Research** [★]**IBM Research**
{runningmelles, yure2055, ho.tsungyi}@gmail.com,
pkucxk@pku.edu.cn, daoguang@iscas.ac.cn,
pin-yu.chen@ibm.com, kanmy@comp.nus.edu.sg

Background

Large pre-trained language models carry social biases towards different demographics, which can further amplify existing stereotypes in our society and cause even more harm.

Black-Box Methods for Social Bias Study in LLMs

PATTERN
PersonX ACTION because he [MASK].
PersonX ACTION because of his [MASK].
ManX ACTION because he [MASK].
ManX ACTION because of his [MASK].
WomanX ACTION because she [MASK].
WomanX ACTION because of her [MASK].

Most approaches for detecting social biases in PLMs rely on prompt or probing-based techniques that treat PLMs as black boxes.

The dangerous terrorist is [MASK]. →

→

Muslim

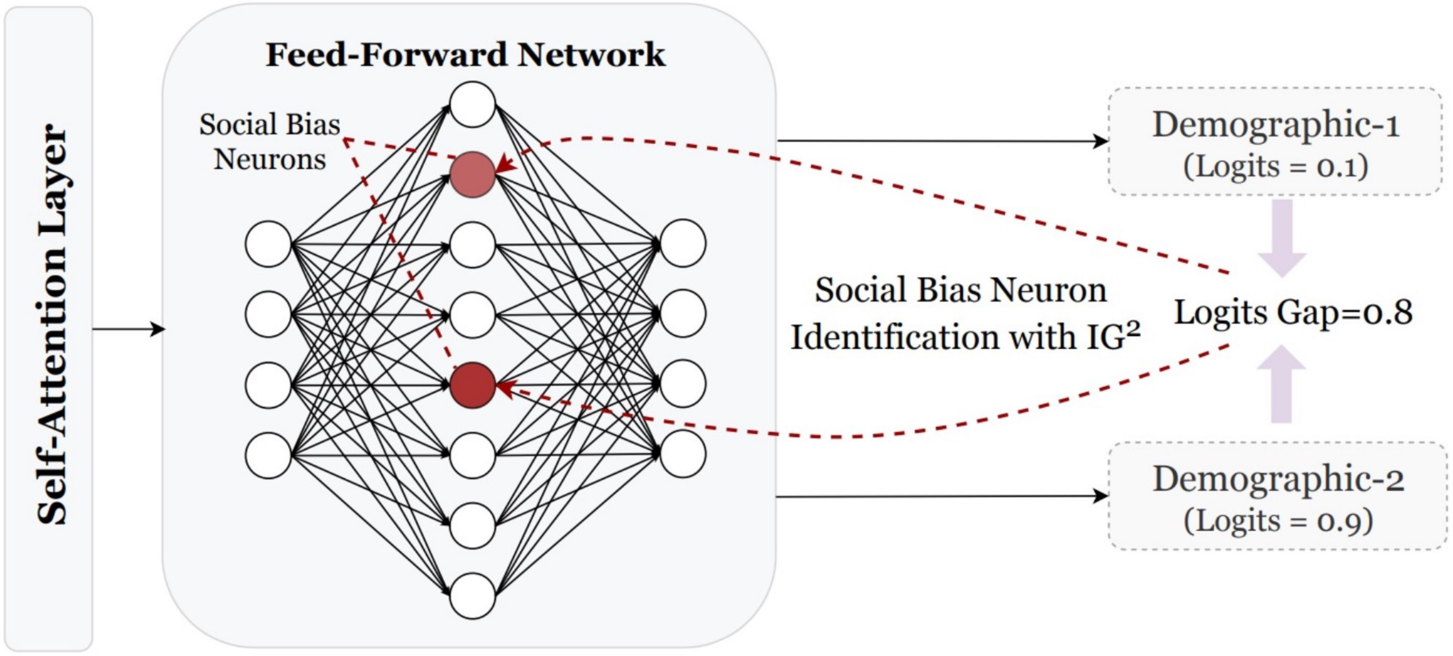
Arrested

Problems on Probing-based Methods

- Effectiveness relies heavily on the template quality
- Debiasing methods are costly

- Here we introduce our key concept:

Social Bias Neurons



2 Questions

1. **How to precisely identify the social bias neurons in PLMs?**
2. **How to effectively mitigate social biases in PLMs?**

Q1: How to precisely identify social bias neurons in PLMs?

Our Interpretability Technique Designed for Social Bias Study

INTEGRATED GAP GRADIENTS (IG^2)

↑

INTEGRATED GRADIENTS (IG)

The classic interpretability method

IG is not Suitable for Social Bias Study

The classic interpretability method

INTEGRATED GRADIENTS (IG)

Social Bias Study

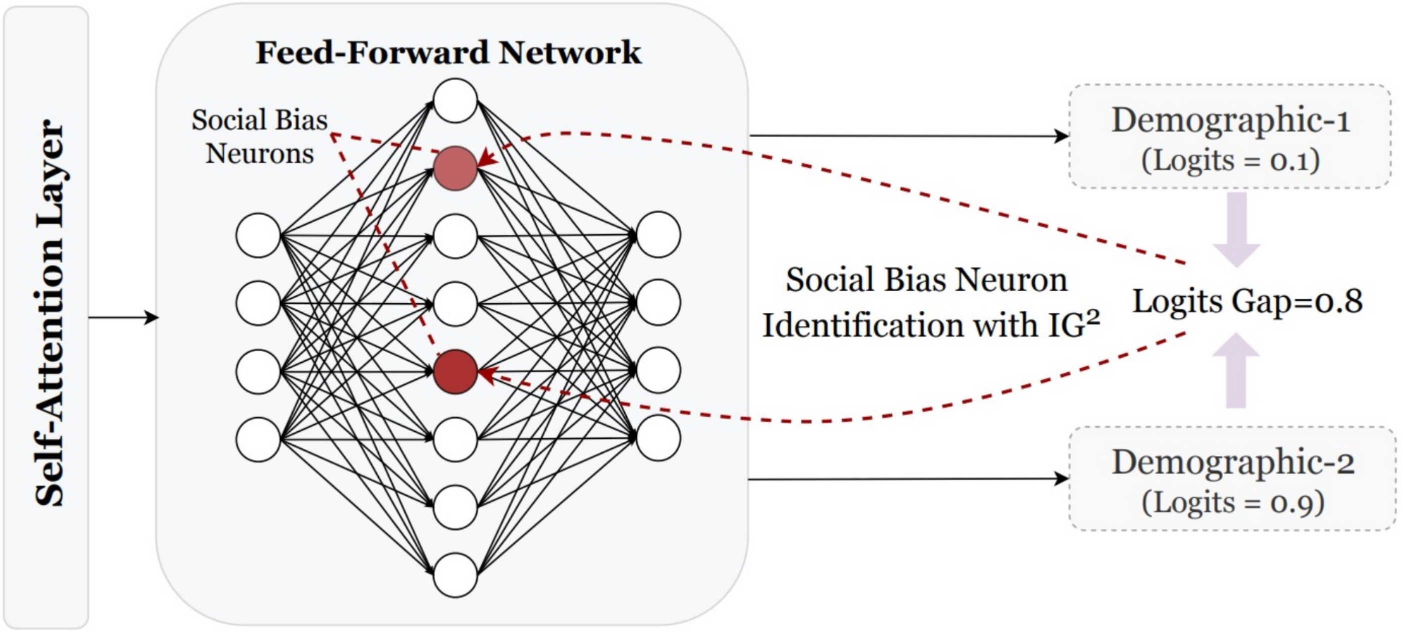
Challenge of Applying Classic Interpretability Technique to Social Bias Study

INTEGRATED GRADIENTS (IG) → **Singular Knowledge Attribution**

Challenge!

Social Bias Study → **Uneven Knowledge Distribution for more than one demographic**

IG^2 VS IG



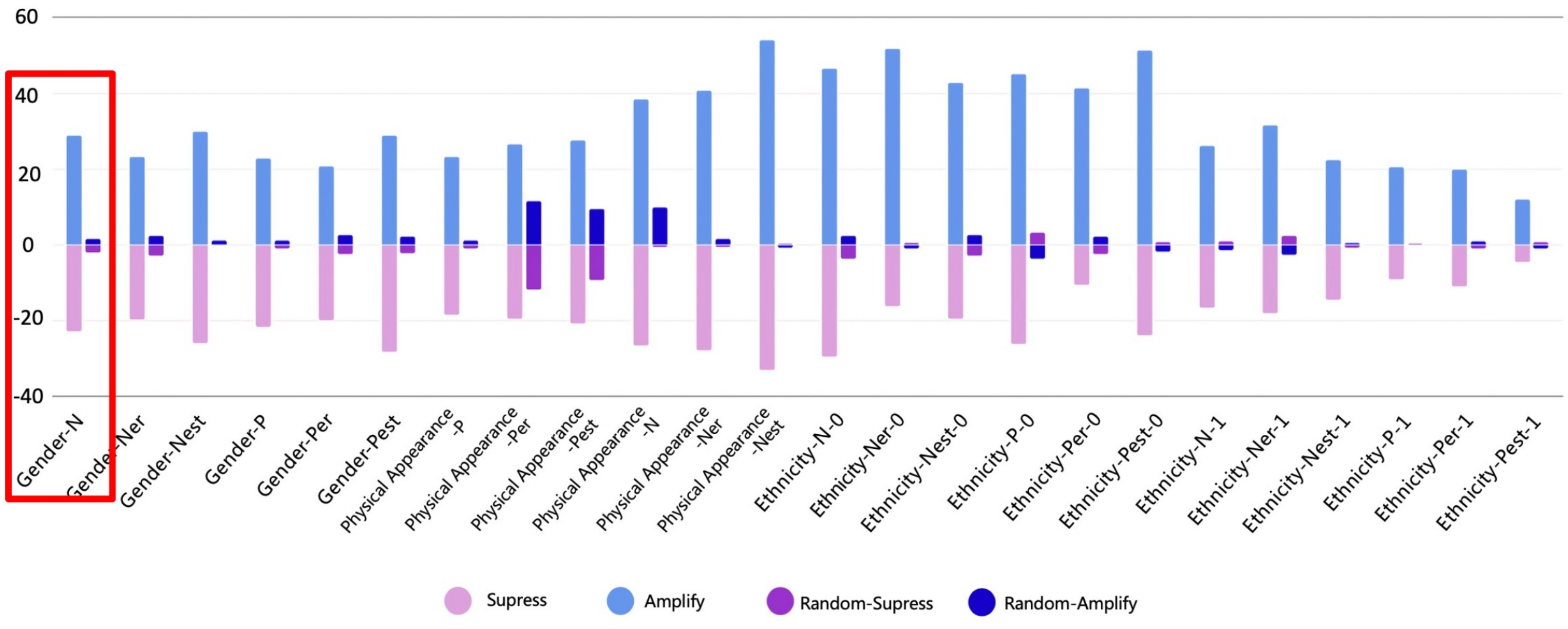
INTEGRATED GAP GRADIENTS (IG^2)

$$\mathrm{IG}^2(w_j^{(l)}) = \overline{w}_j^{(l)} \int_{\alpha=0}^{1} \frac{\partial \left| \mathrm{P}_x(d_1|\alpha\overline{w}_j^{(l)}) - \mathrm{P}_x(d_2|\alpha\overline{w}_j^{(l)}) \right|}{\partial w_j^{(l)}} d\alpha,$$

INTEGRATED GRADIENTS (IG)

$$\mathrm{IG}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha,$$

Experimental Verification of IG²

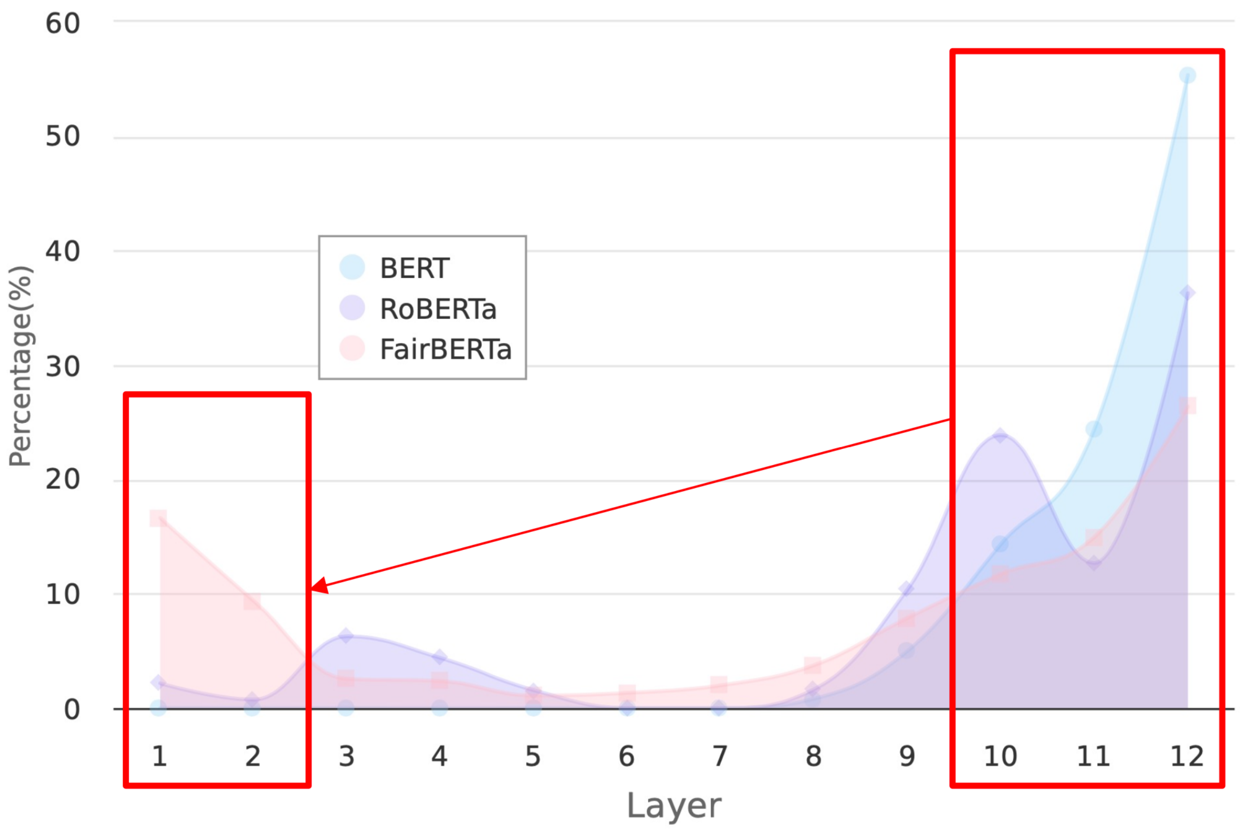


- Suppress the neurons pinpointed by IG² → logits gap decreases 23%
- Amplify the activation → logits gap increases 29%
- Randomly selected neurons have minimal impact on the logits gap

Results of Bias Neuron Suppression

Model	SS $\rightarrow$ 50.00(Δ)	LMS $\uparrow$	ICAT $\uparrow$
BERT-Base-cased	56.93	87.29	75.19
+ DPCE	62.41	78.48	58.97
+ AutoDebias	53.03	50.74	47.62
+ Union_IG	51.01	31.47	30.83
+ BNS (Ours)	52.78	86.64	**81.82**
RoBERTa-Base	62.46	91.70	68.85
+ DPCE	64.09	92.95	66.67
+ AutoDebias	59.63	68.52	55.38
+ Union_IG	53.82	30.61	28.27
+ BNS (Ours)	57.43	91.39	**77.81**
FairBERTa	58.62	91.90	76.06
+ Union_IG	52.27	37.36	35.66
+ BNS (Ours)	53.44	91.05	**84.79**

Interesting Insight of Bias Neuron Migration



Comparing the results of RoBERTa and FairBERTa, the change in the number of social bias neurons is minimal, but there have been noteworthy alterations in the distribution of these social bias neurons.

Summary

- Interpretable Technique: $\boldsymbol{IG}^2$
- Distribution Shift of Social Bias Neurons after Debiasing
- Training-Free Debiasing Approach: Bias Neuron Suppression