

Neural Networks bias recap

Prof. Luca Cagliero
Dipartimento di Automatica e Informatica
Politecnico di Torino



Lecture goal

- Recap fundamentals of Neural Networks bias
 - Bias definition
 - Selection bias
 - Sampling bias
 - Bias and fairness in language models

Disclaimer

- Neural networks and deep learning architectures are a prerequisite
 - This body of knowledge is intended to be a quick refresh of basic concepts that are particularly relevant to Deep NLP

Credits

- The following slides used as supporting materials are an extract of
 - Alexander Amini and Ava Soleimany
 - 6.S191 Introduction to Deep Learning. MIT Course
 - <https://introtodeeplearning.com>
 - © MIT license

Algorithmic Bias in the Headlines

Promoting Racism and
Systematic Bias in Machine
Learning Algorithms

Bias Bias in medical algorithm favors white
patients over black patients

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

Bias Bias in UK's Women's
Healthcare Race -> Challenge For

Gender bias in AI: building
fairer algorithms

Bias in AI: A problem recognized but
not yet resolved

What exactly does **bias** mean?

Millions of
bias in health-care algorithms

When It Comes to Gorillas, Google Photos Remains Blind

Study finds impact of gender bias in AI systems
and machine learning

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

Computer vision algorithm is moving
from the lab to the real world

Artificial intelligence has a gender bias
problem – just ask Siri

The Best Algorithms Struggle to Recognize Black Faces Equally

Algorithmic Bias in the Headlines

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

**Gender bias in AI: building
fairer algorithms**

**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.

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

**Bias in AI: A problem recognized but
still unresolved**

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

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

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.*

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.

Artificial Intelligence has a gender bias
problem – just ask Siri

Labeling, Prototyping, and Stereotyping

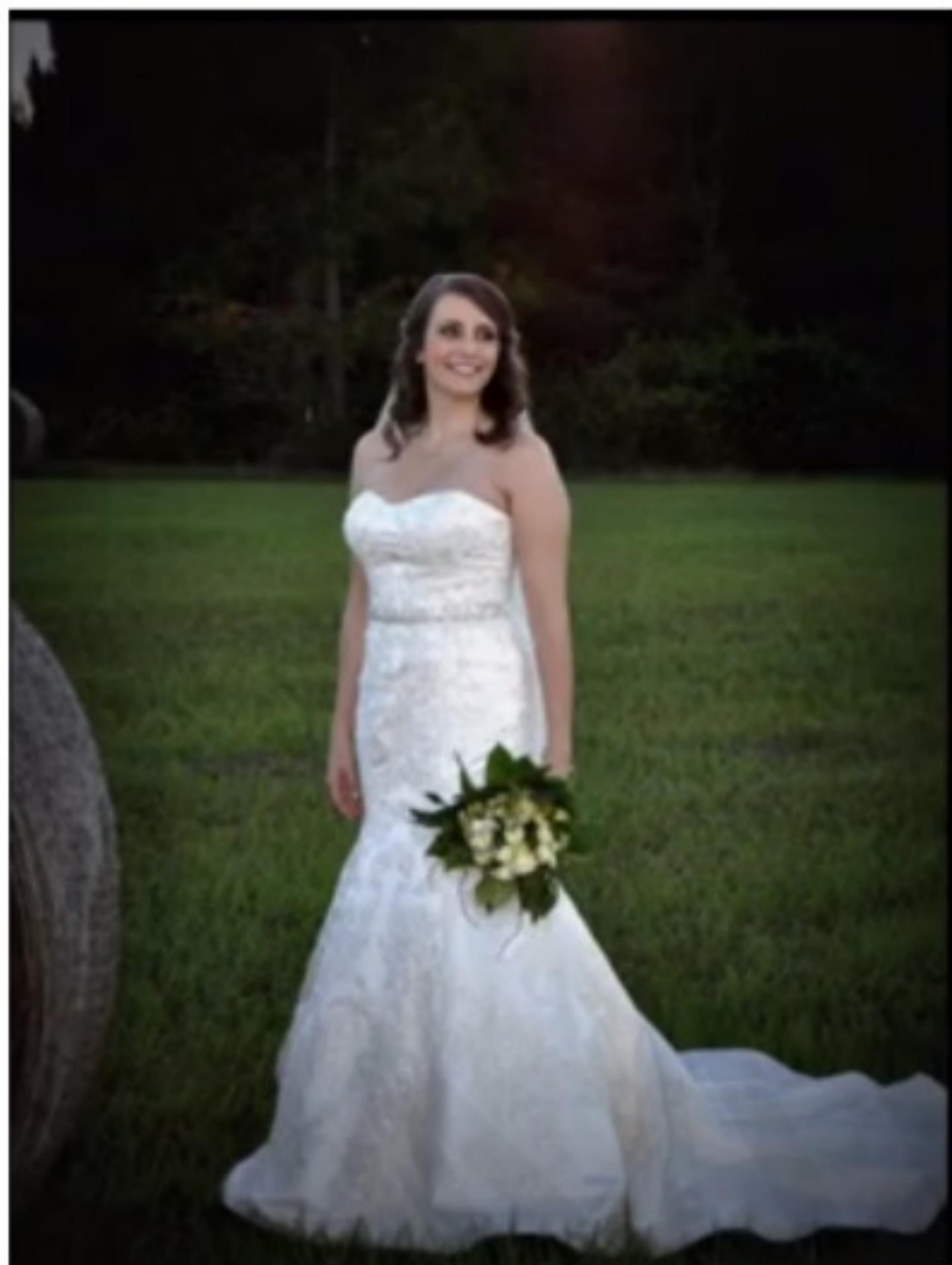
We **label** and **categorize** the world to reduce complex sensory inputs into **simplified** groups that are easier to work with.

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

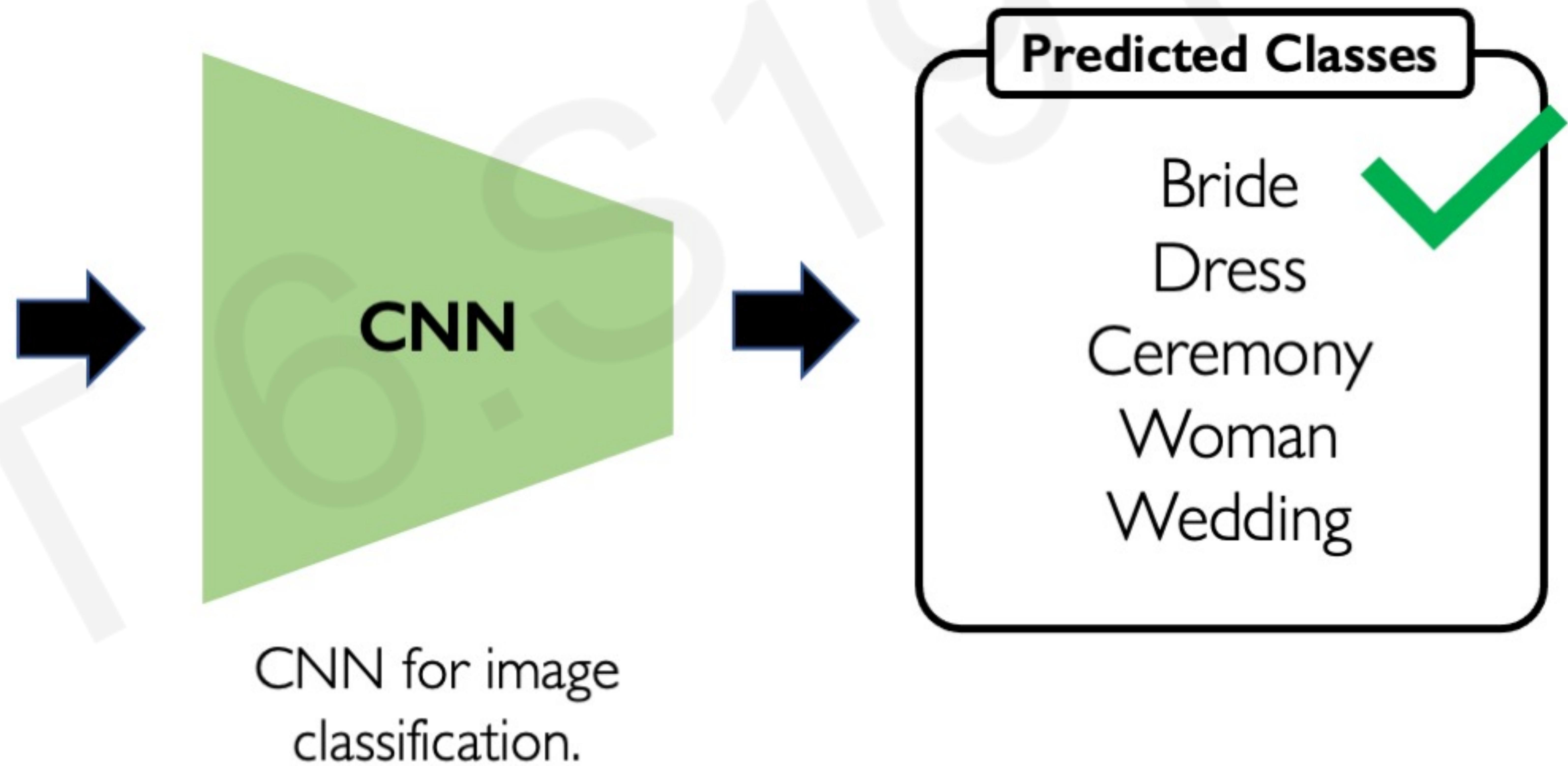
We tend to notice and talk about things that are **atypical**.

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

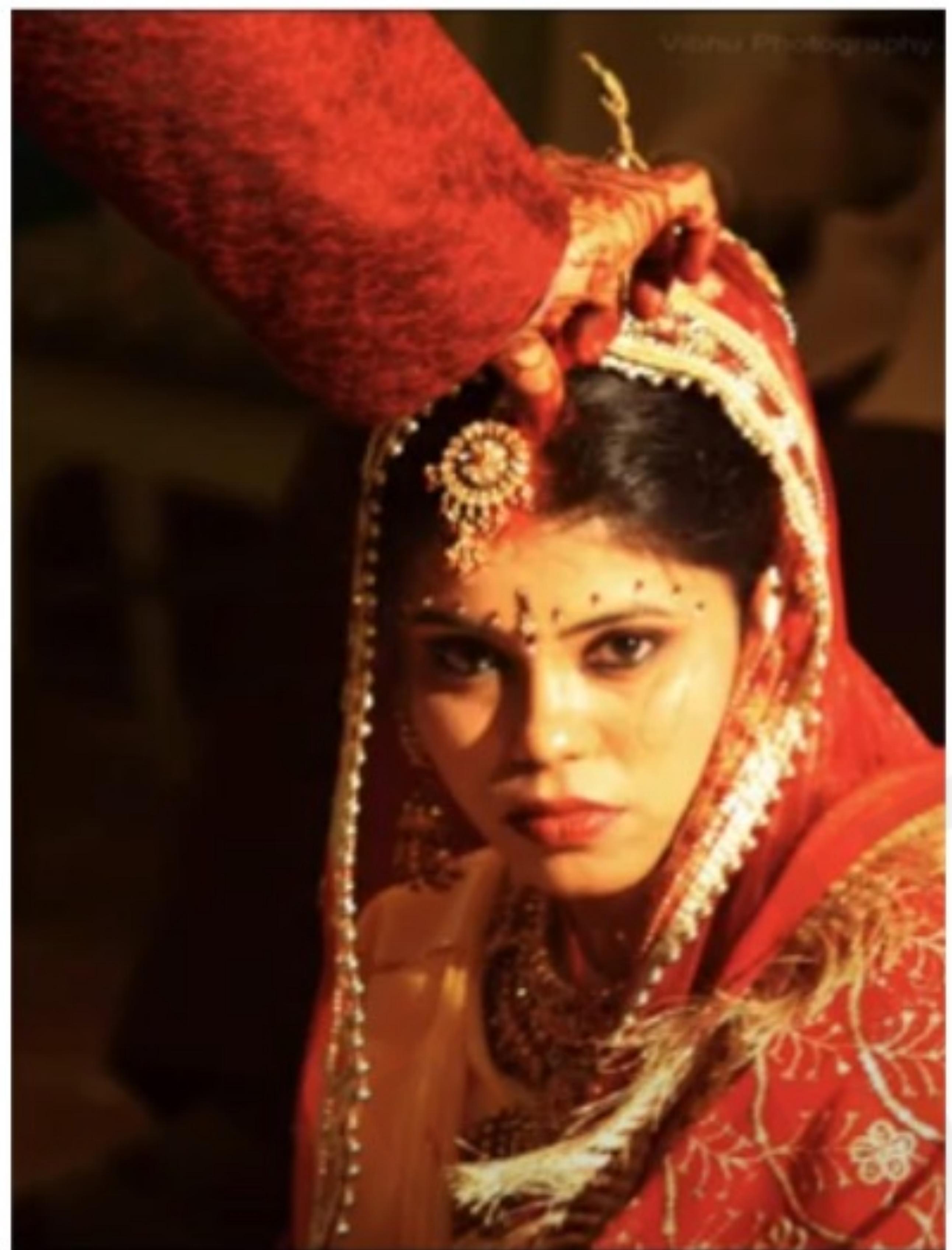
Bias in Image Classification



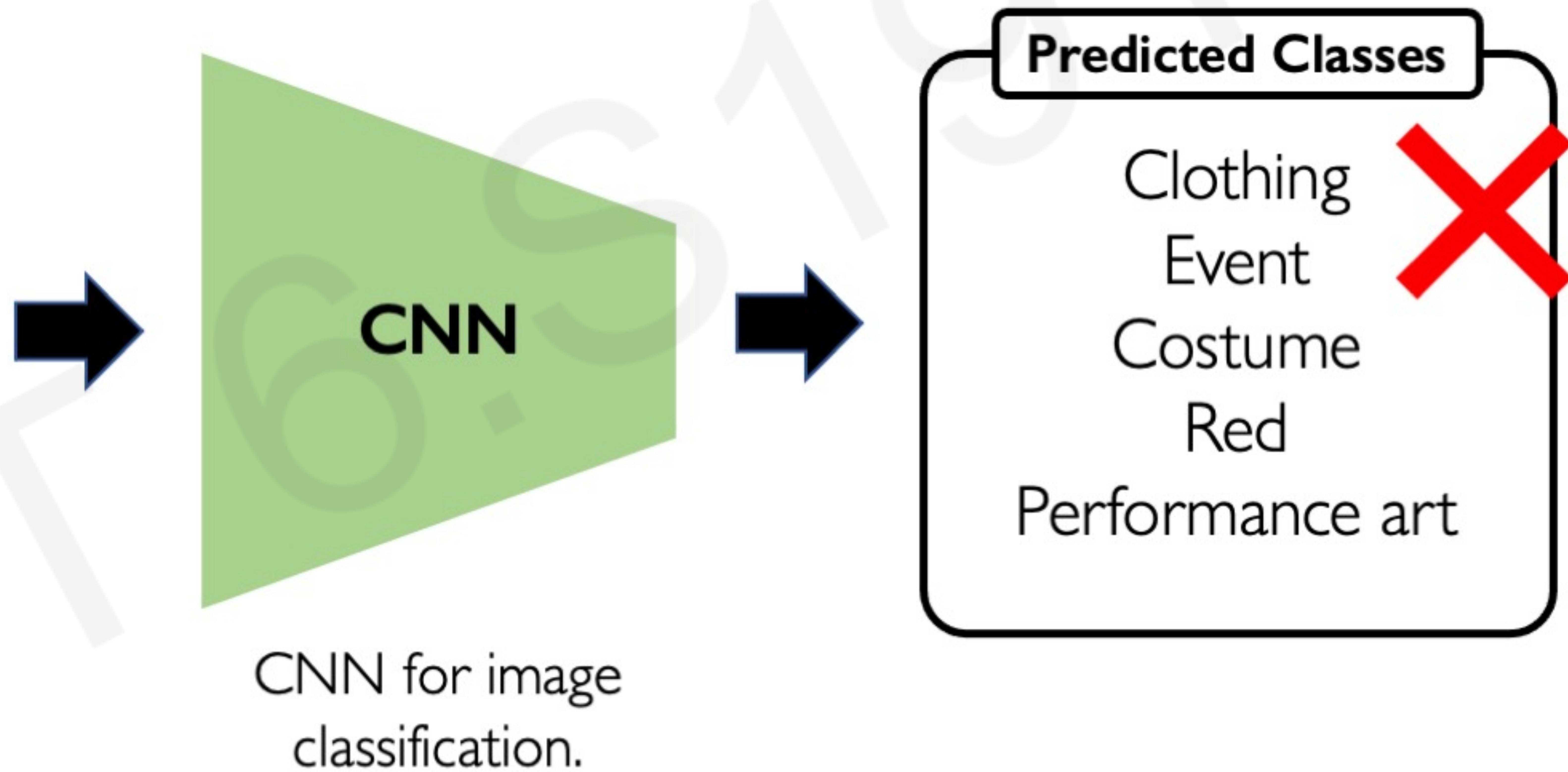
Ground Truth: Bride



Bias in Image Classification



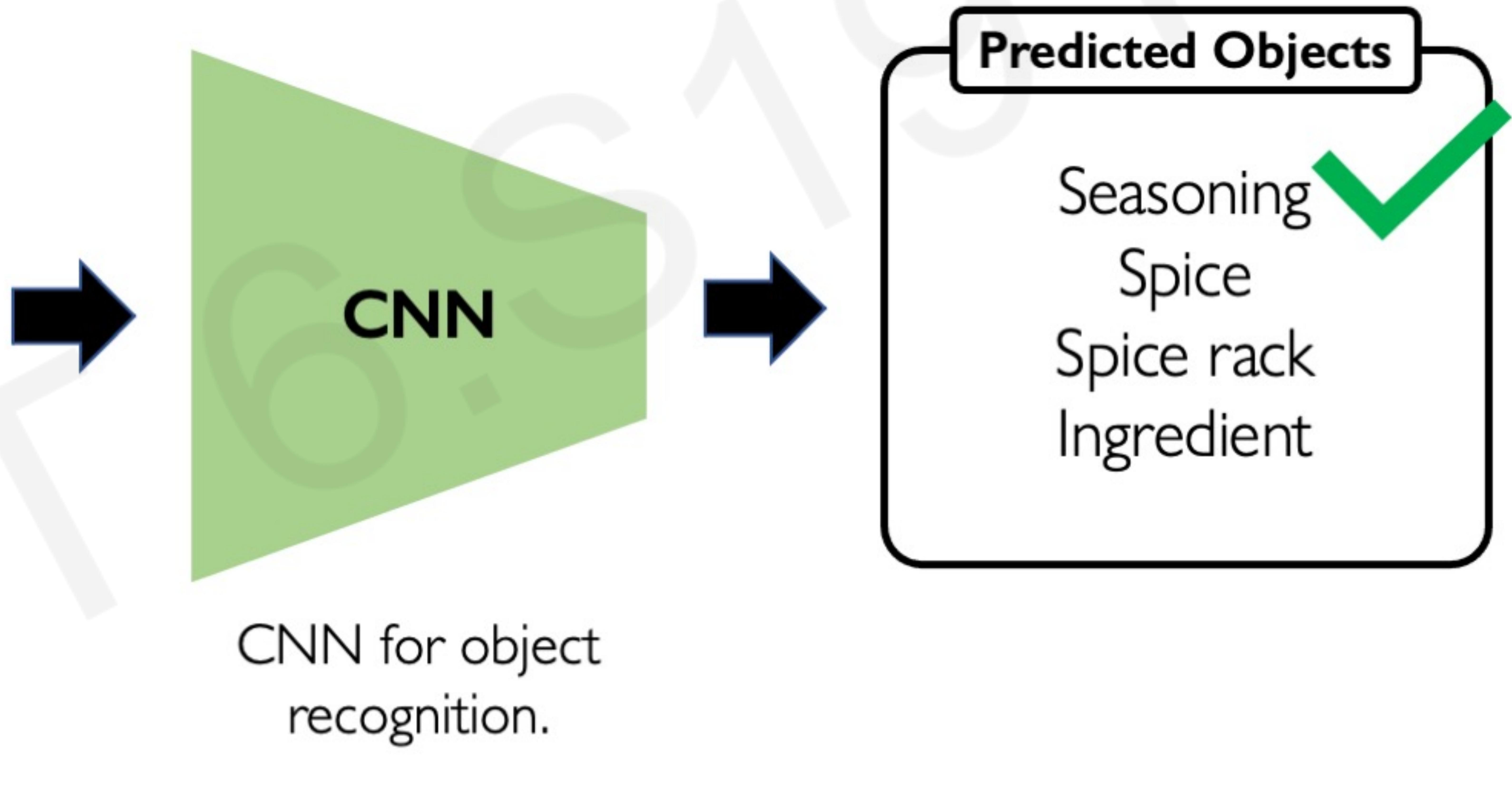
Ground Truth: Bride



Bias in Object Recognition



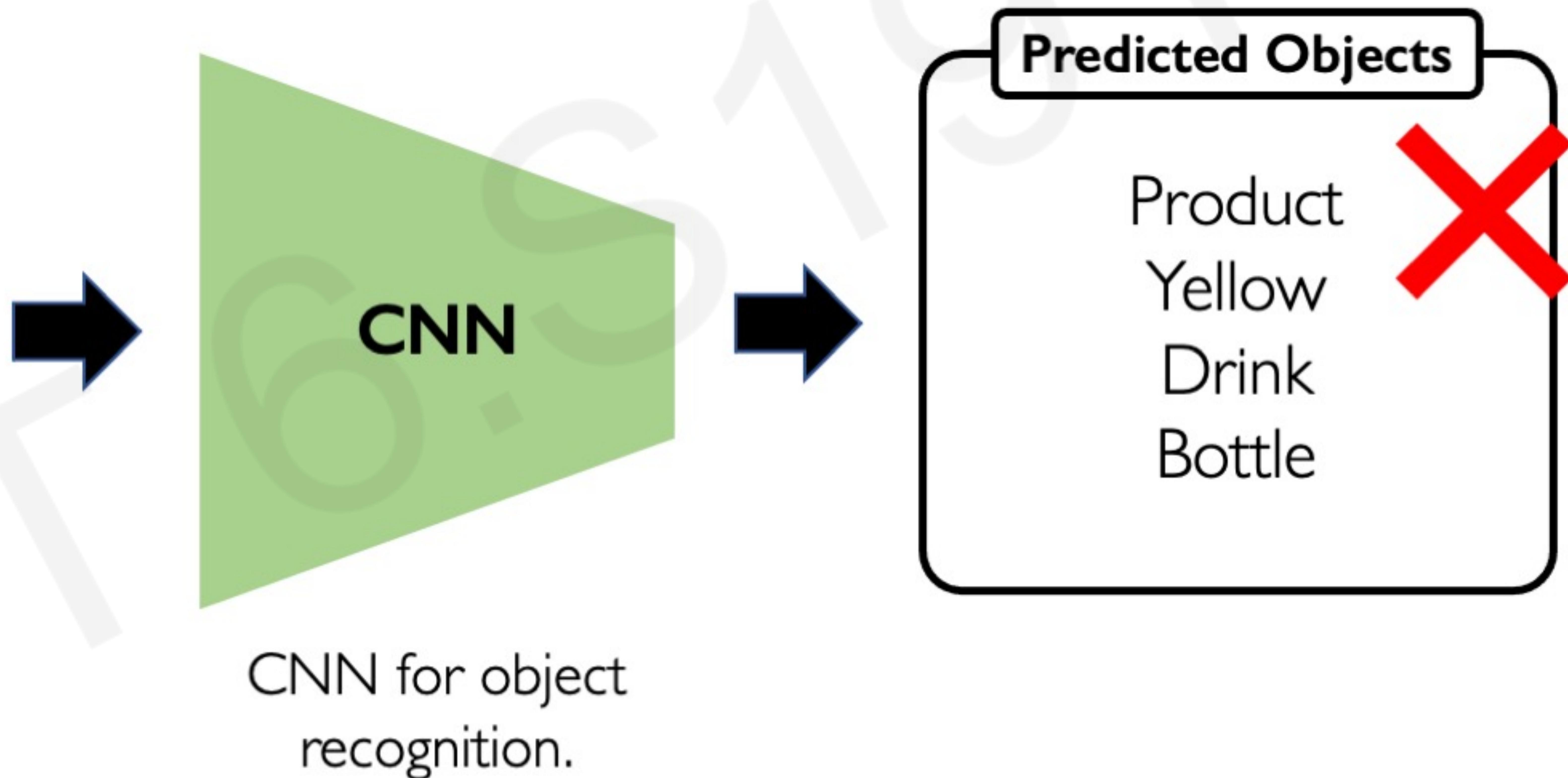
Ground Truth: Spices



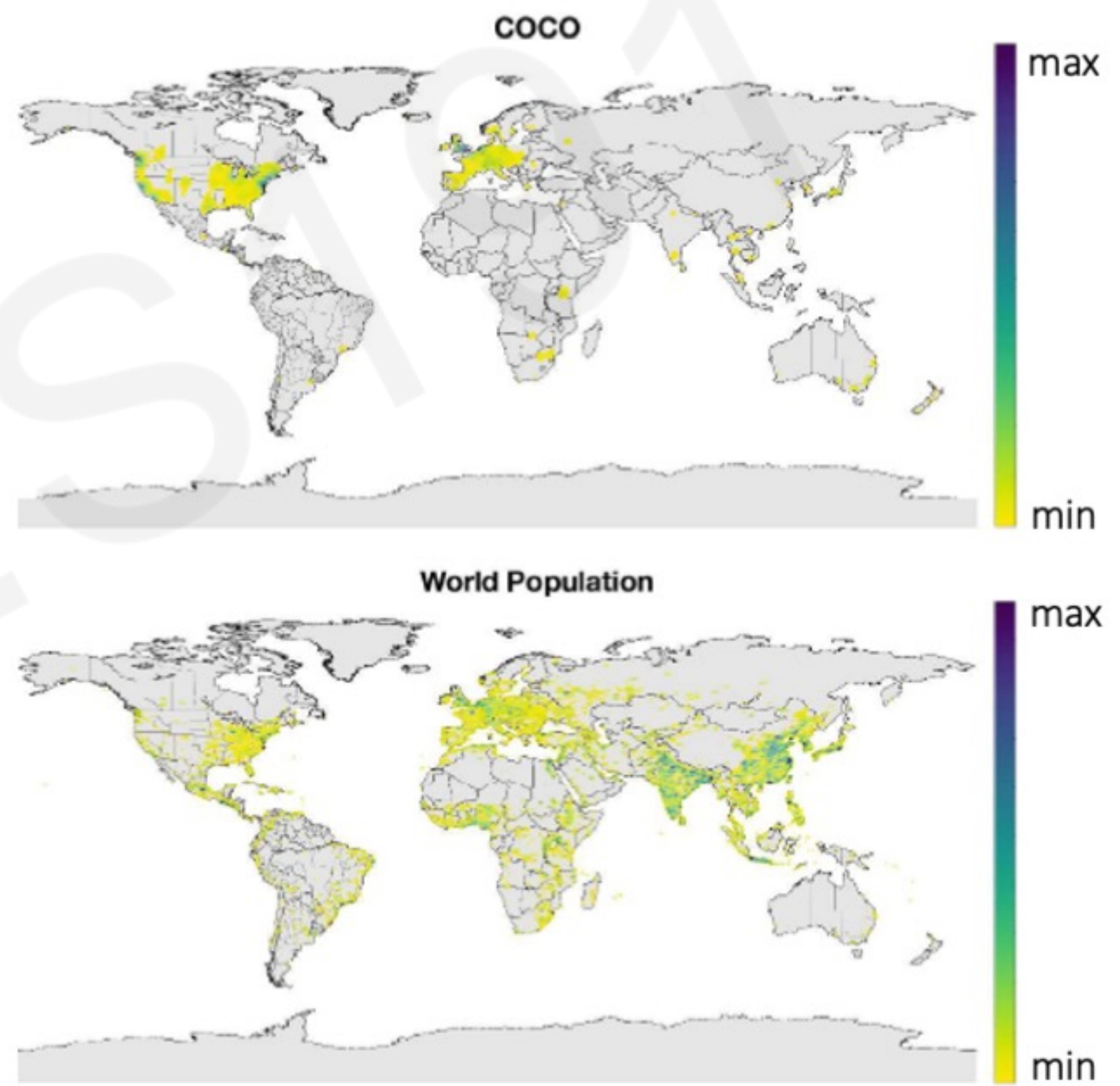
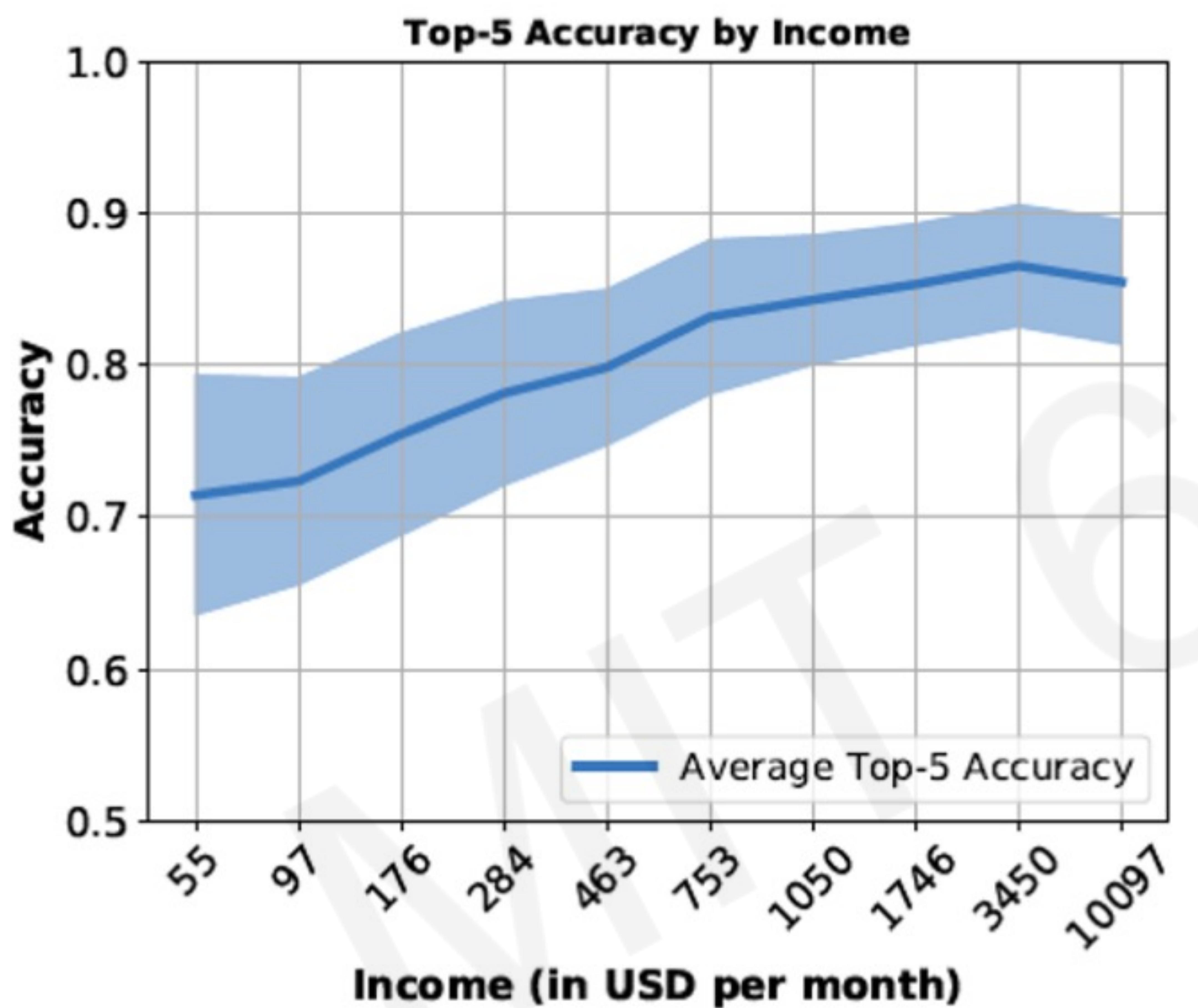
Bias in Object Recognition



Ground Truth: Spices



Bias Correlation with Income and Geography



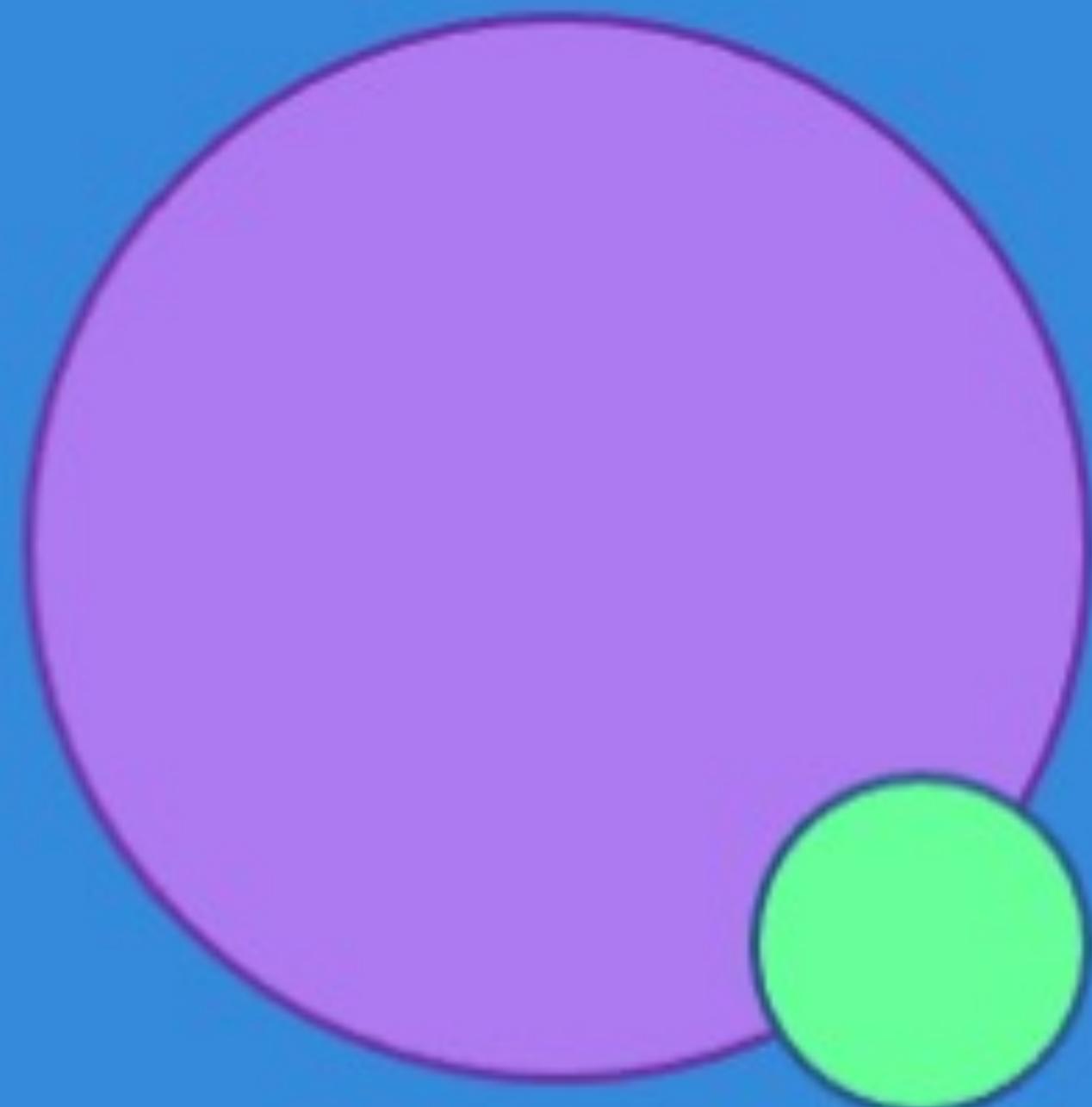
Bias at All Stages of the AI Life Cycle

- 1. Data:** imbalances with respect to class labels, features, input structure
- 2. Model:** lack of unified uncertainty, interpretability, and performance metrics
- 3. Training and deployment:** feedback loops that perpetuate biases
- 4. Evaluation:** done in bulk, lack of systematic analysis with respect to data subgroups
- 5. Interpretation:** human errors and biases distort meaning of results



Understanding and Mitigating Algorithmic Bias

Types and Sources of Bias



Strategies to Mitigate Bias



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 \neq Causation

Overgeneralization

"General" conclusions drawn from limited test data

Automation Bias

AI-generated decisions are favored over human-generation decisions

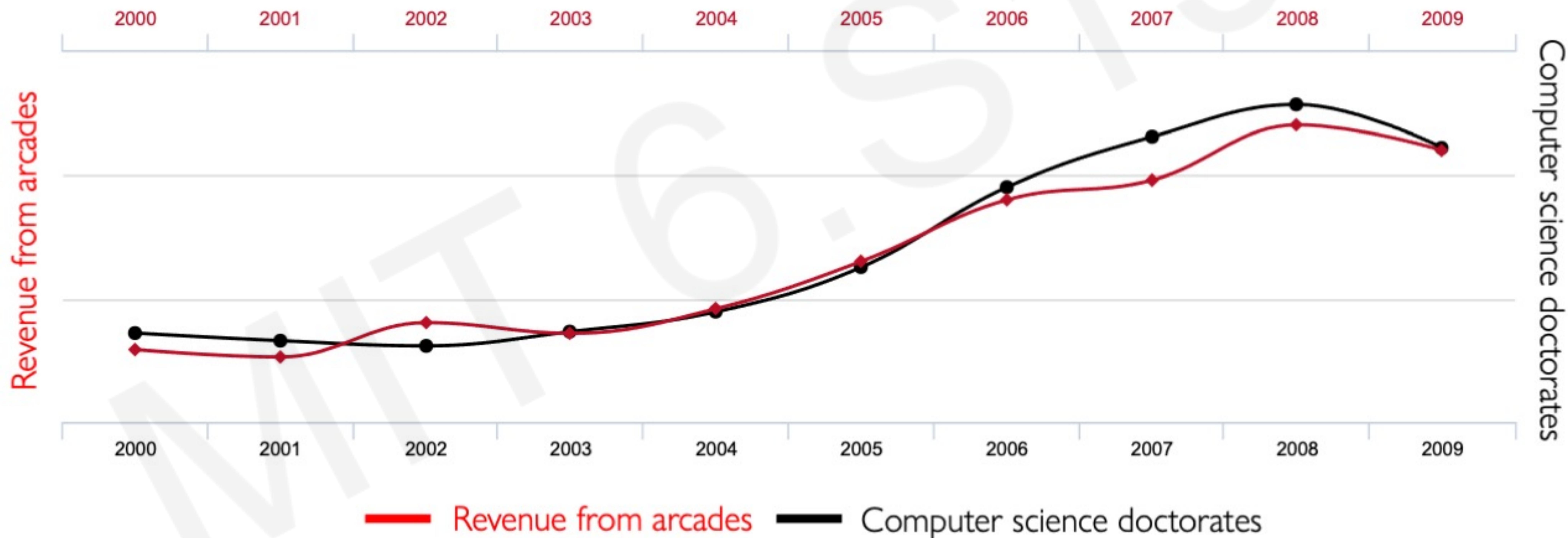
By no means an exhaustive list!

Bias from the Correlation Fallacy

Total revenue generated by arcades

correlates with

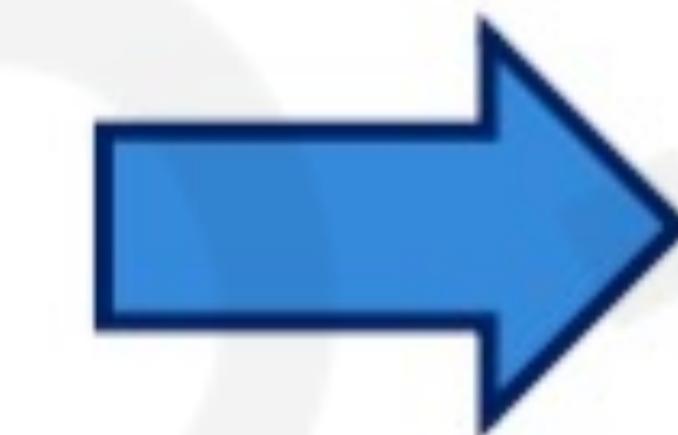
Computer science doctorates awarded in the US



Bias from Assuming Overgeneralization

Expectation:

Cups in my dataset



Reality:

Cups from many angles



Distribution shift can result in neural network bias.

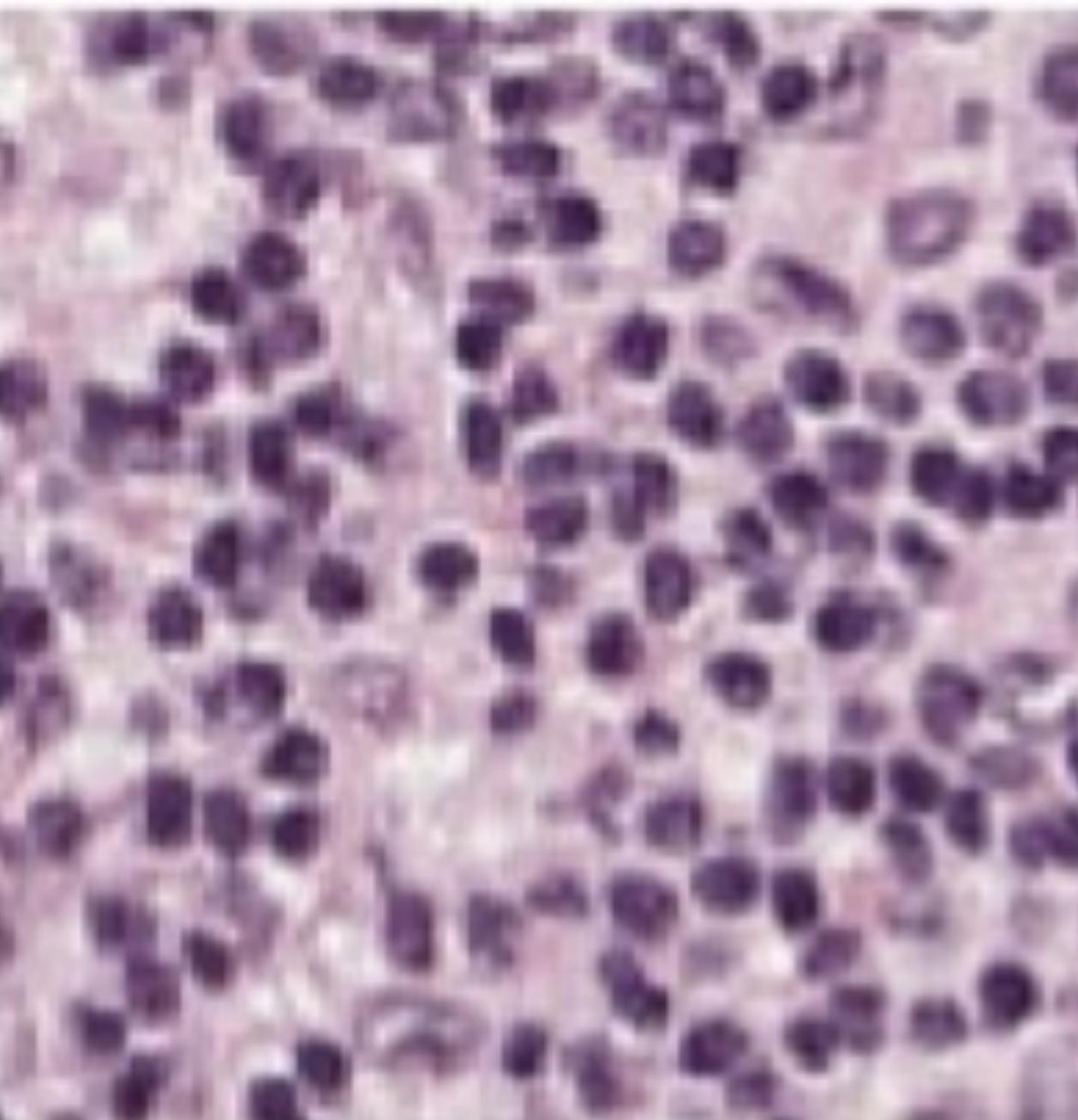
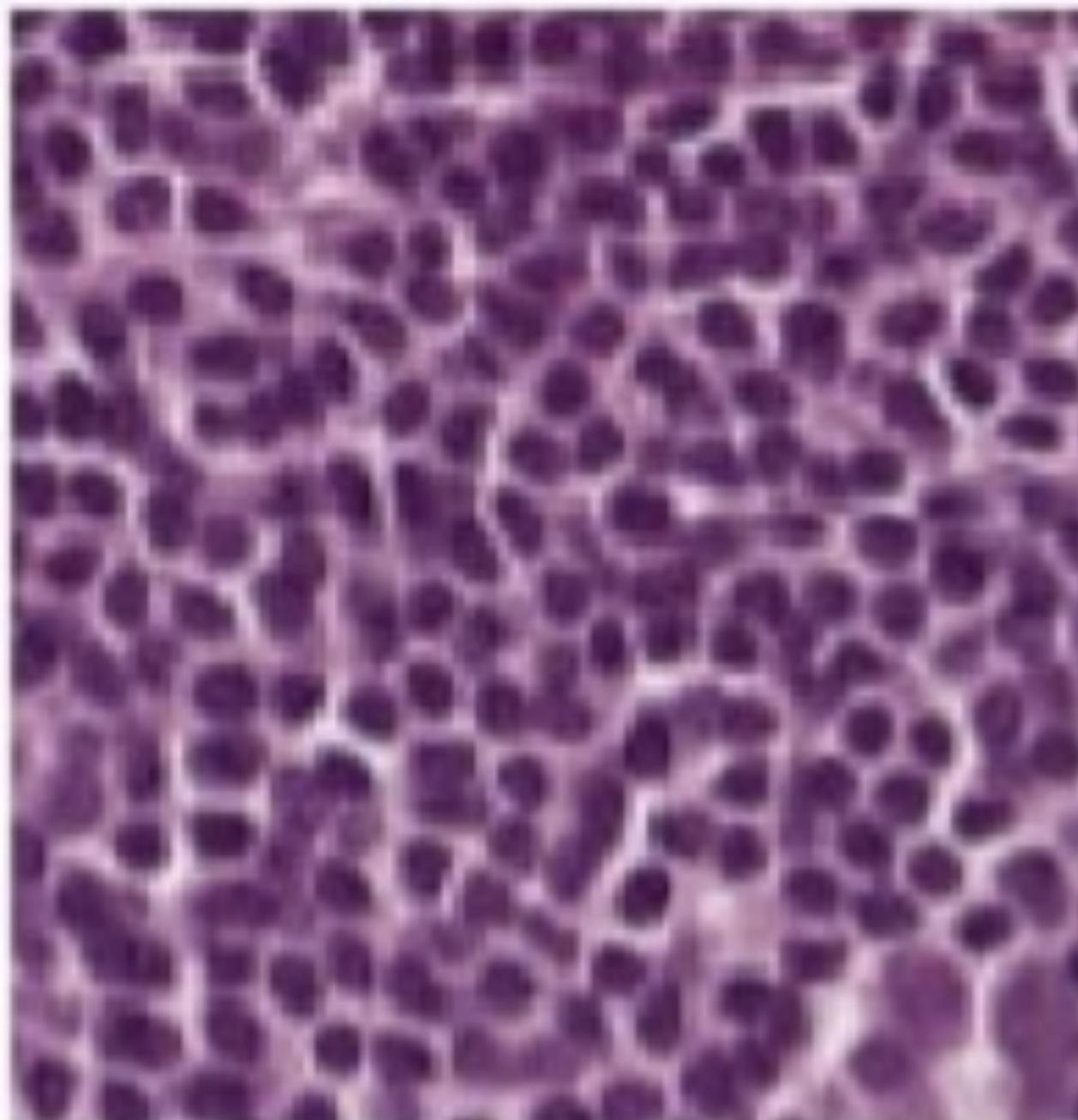
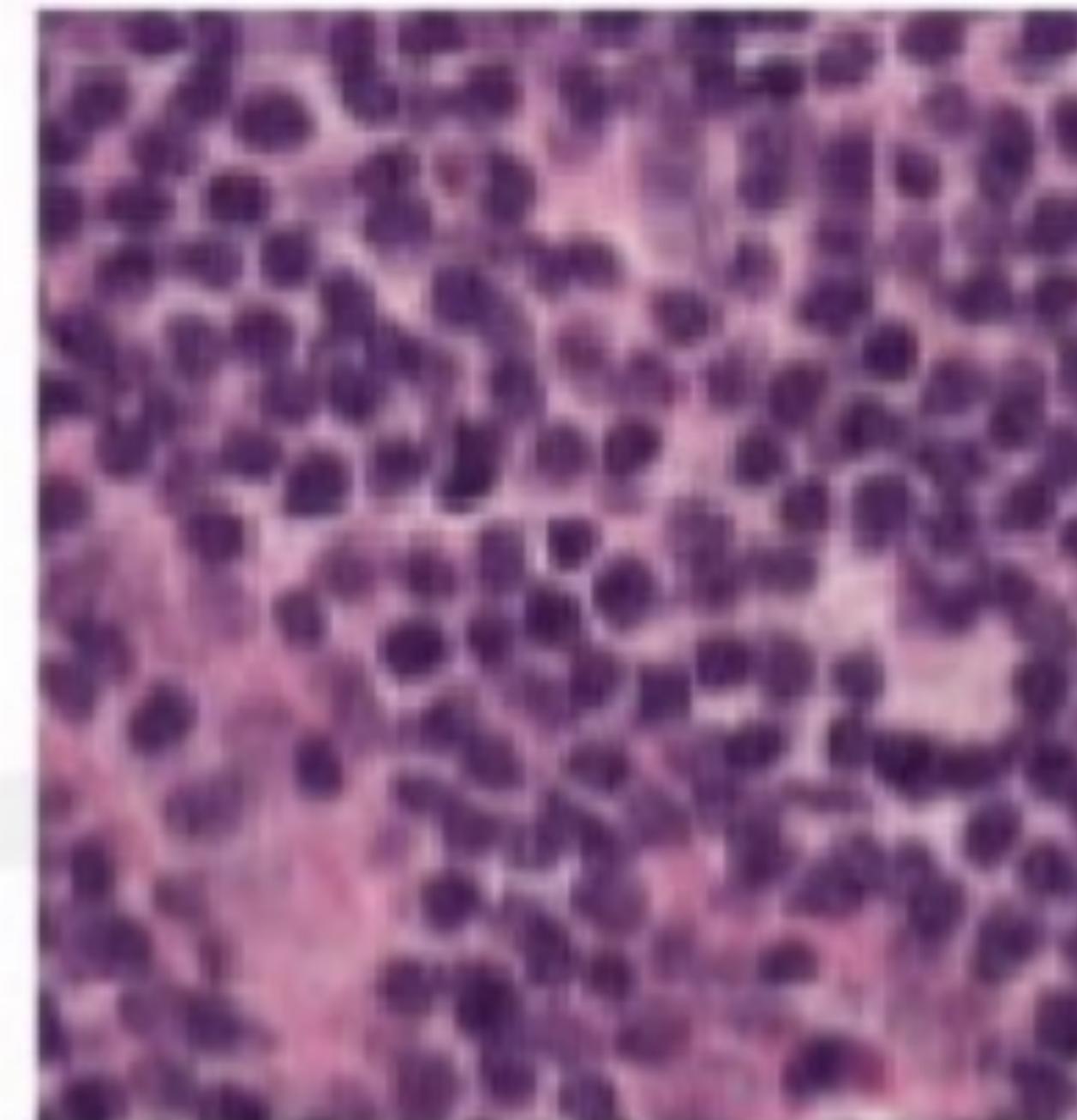
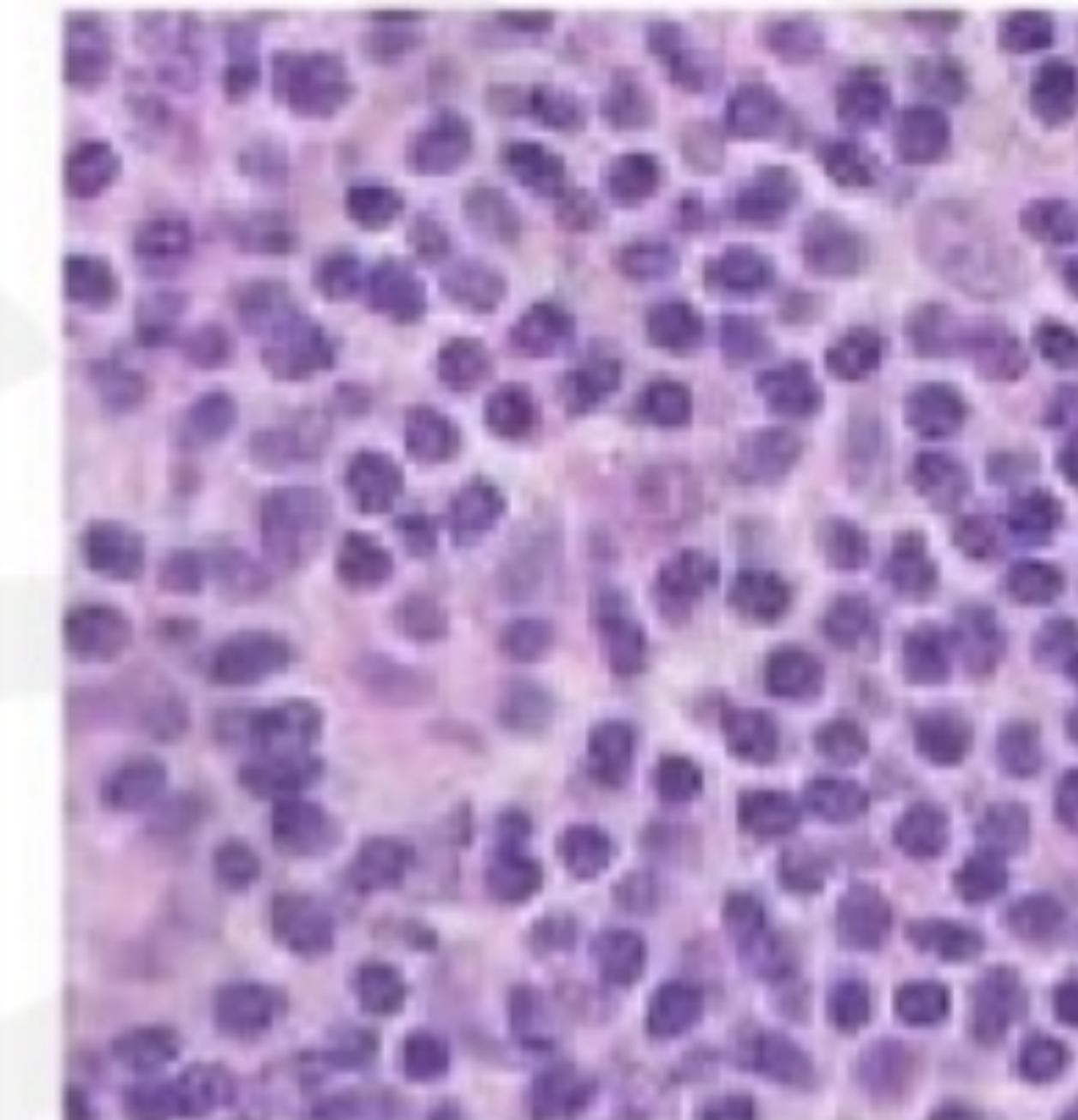
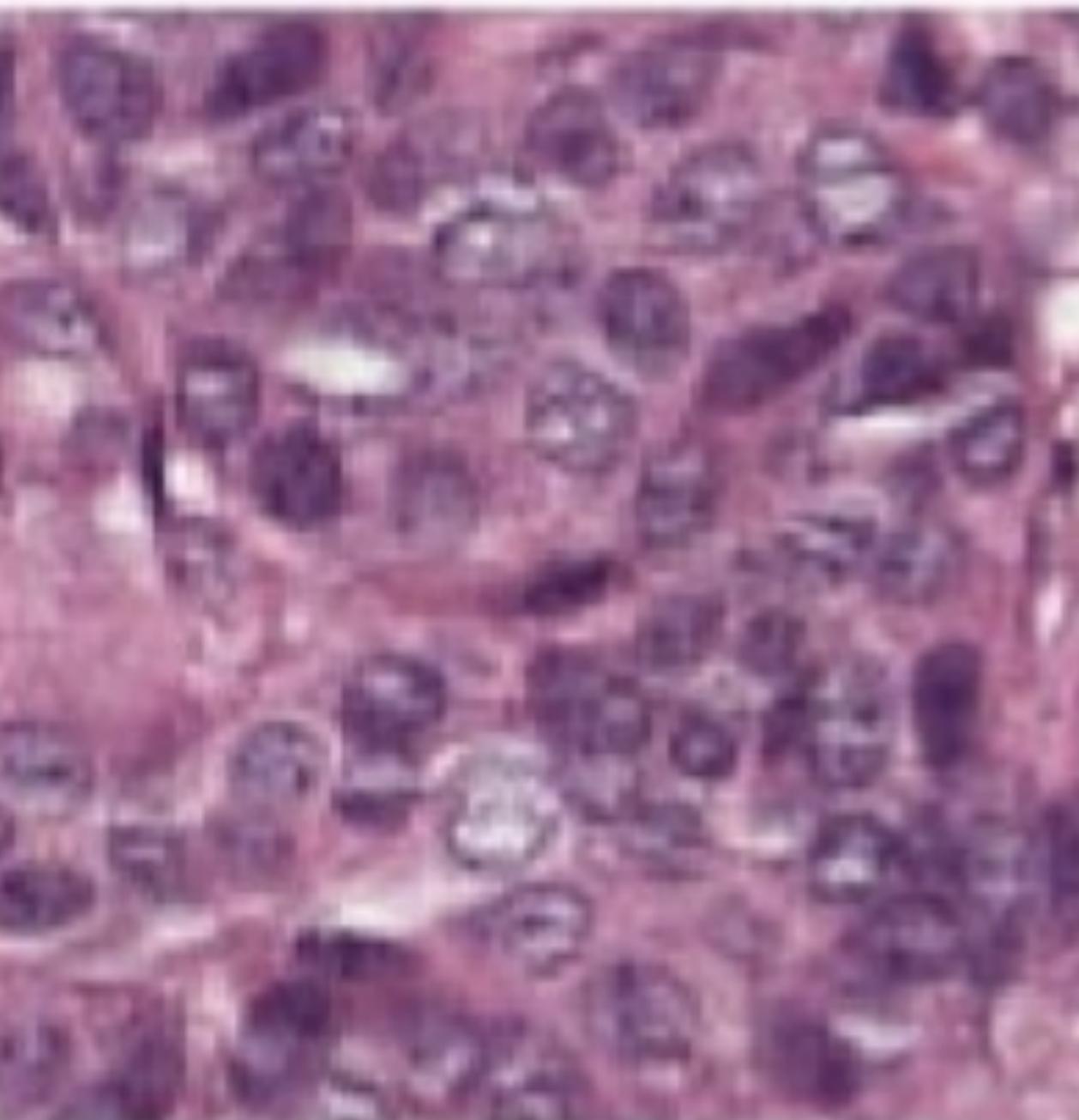
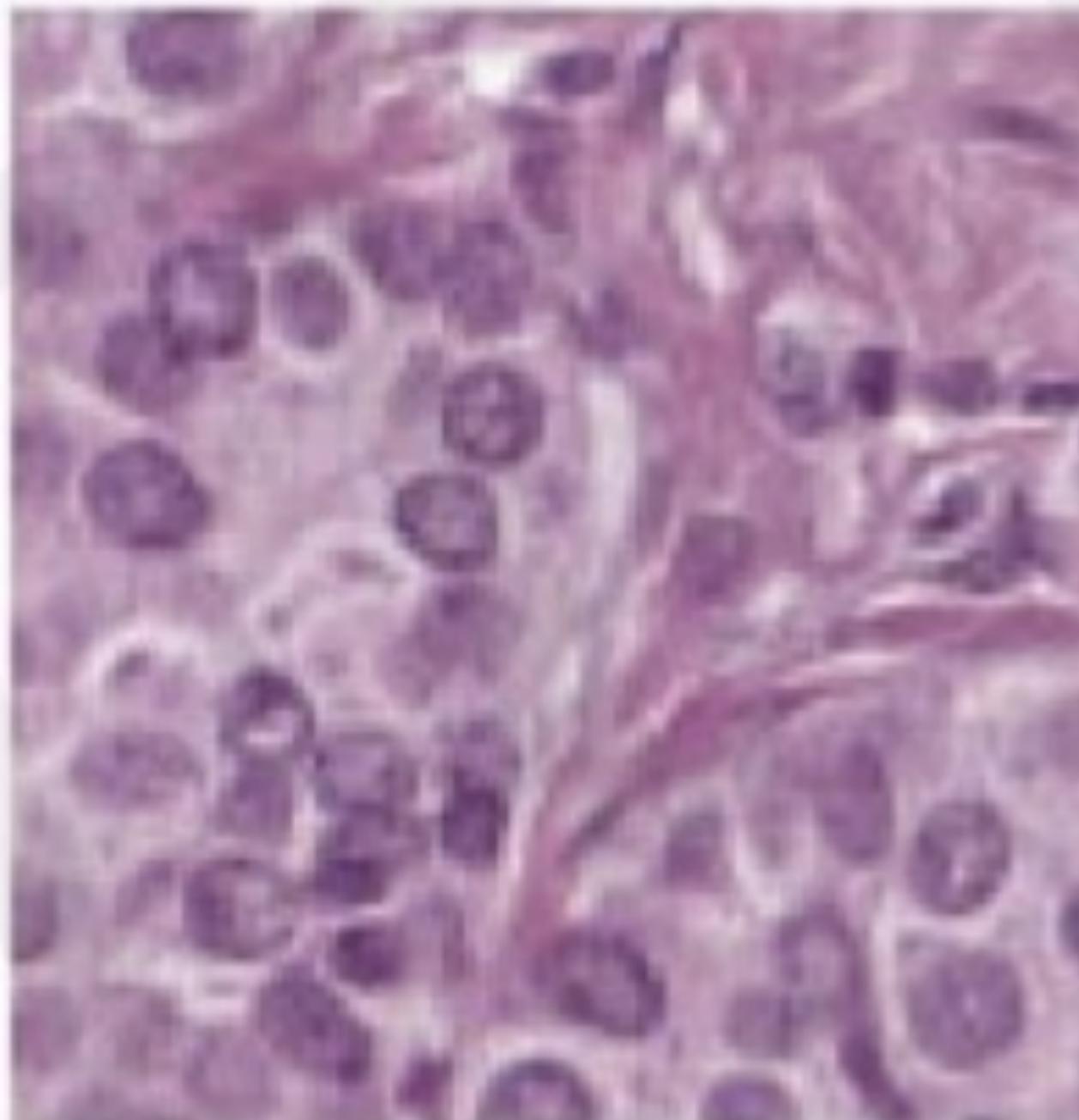
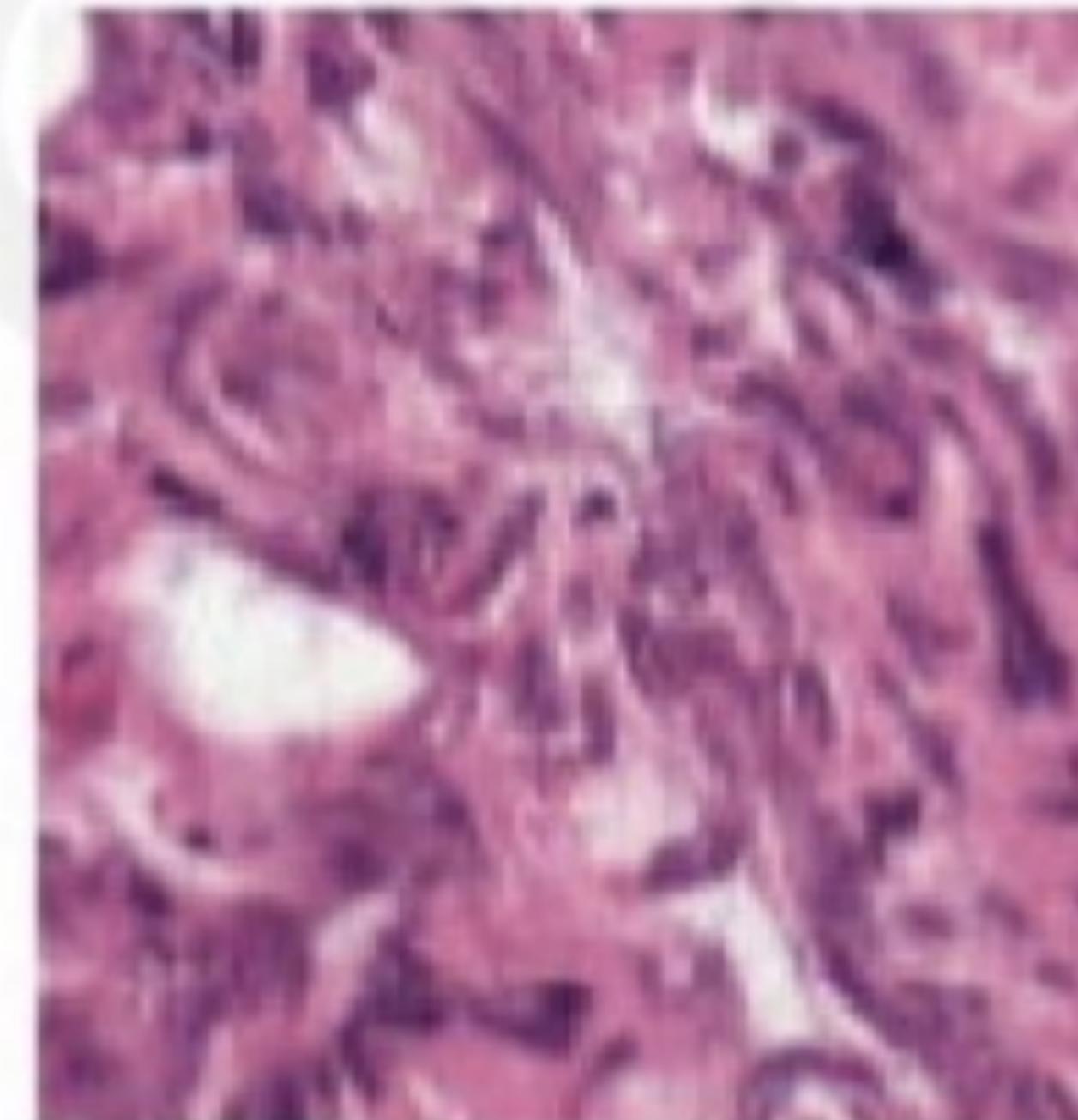
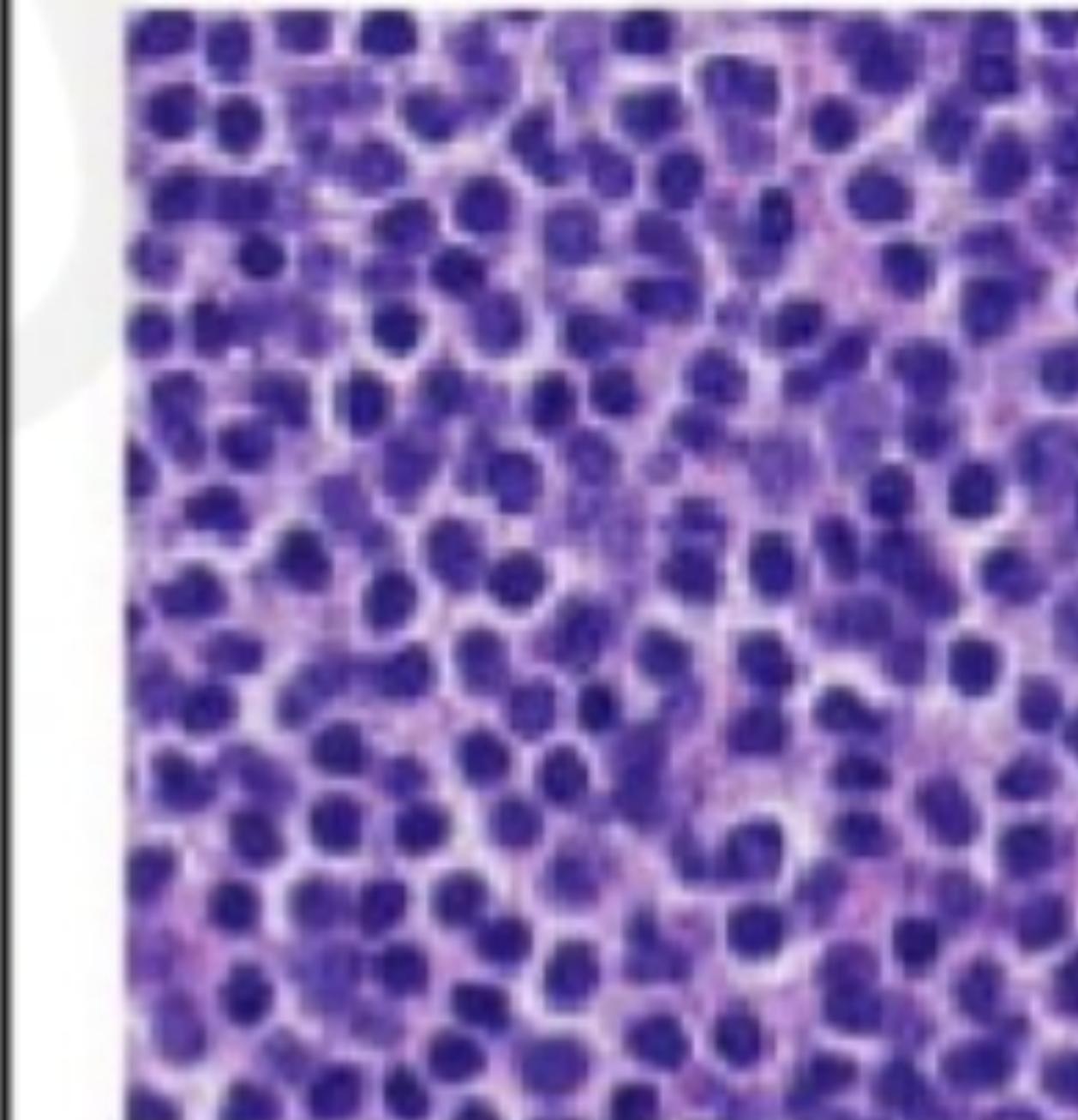
Datasets with Distribution Shifts

	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

Datasets with Distribution Shifts

	Train		Val (OOD)	Test (OOD)
$y = \text{Normal}$	$d = \text{Hospital 1}$ 	$d = \text{Hospital 2}$ 	$d = \text{Hospital 3}$ 	$d = \text{Hospital 4}$ 
	$y = \text{Tumor}$ 	$d = \text{Hospital 2}$ 	$d = \text{Hospital 3}$ 	$d = \text{Hospital 5}$ 

Task: Disease classification from histopathology images

Distribution shift: Hospital source

Taxonomy of Common Biases

Data-Driven

Selection Bias

Data selection does not reflect randomization
Ex: class imbalance

Sampling Bias

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

Interpretation-Driven

Correlation Fallacy

Correlation \neq Causation

Overgeneralization

"General" conclusions drawn from limited test data

Automation Bias

AI-generated decisions are favored over human-generation decisions

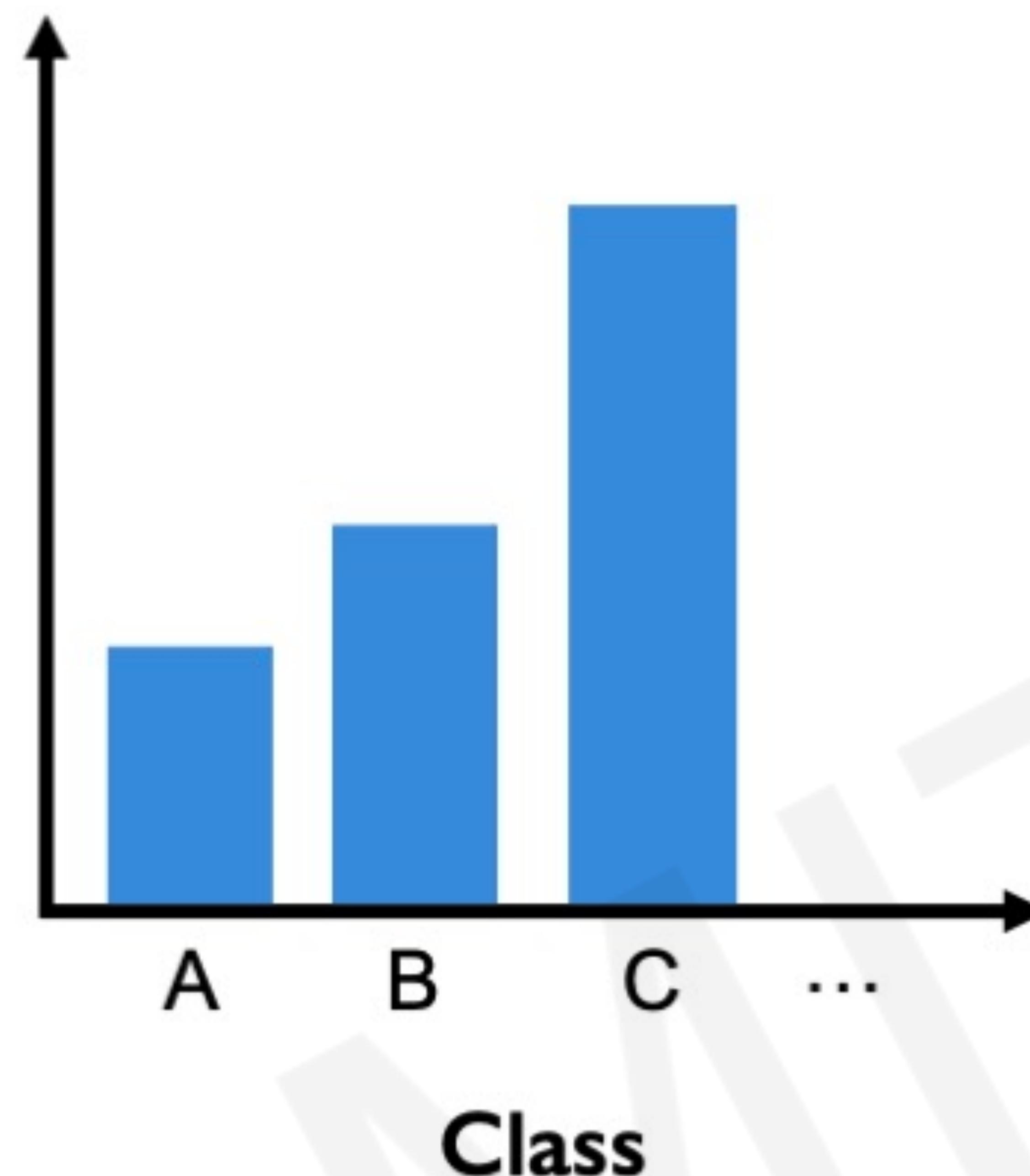
By no means an exhaustive list!

Biases due to Class Imbalance

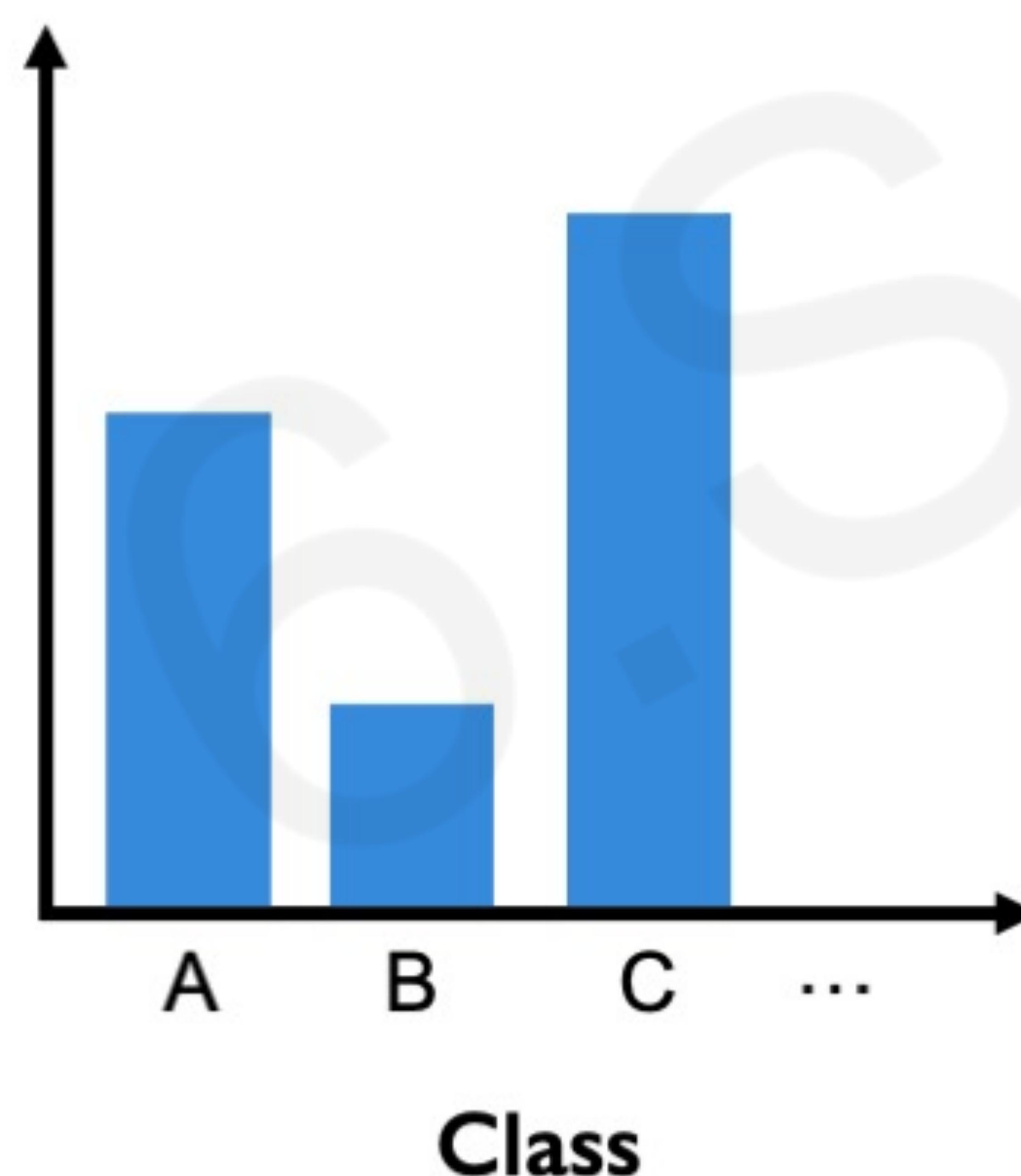


Biases due to Class Imbalance

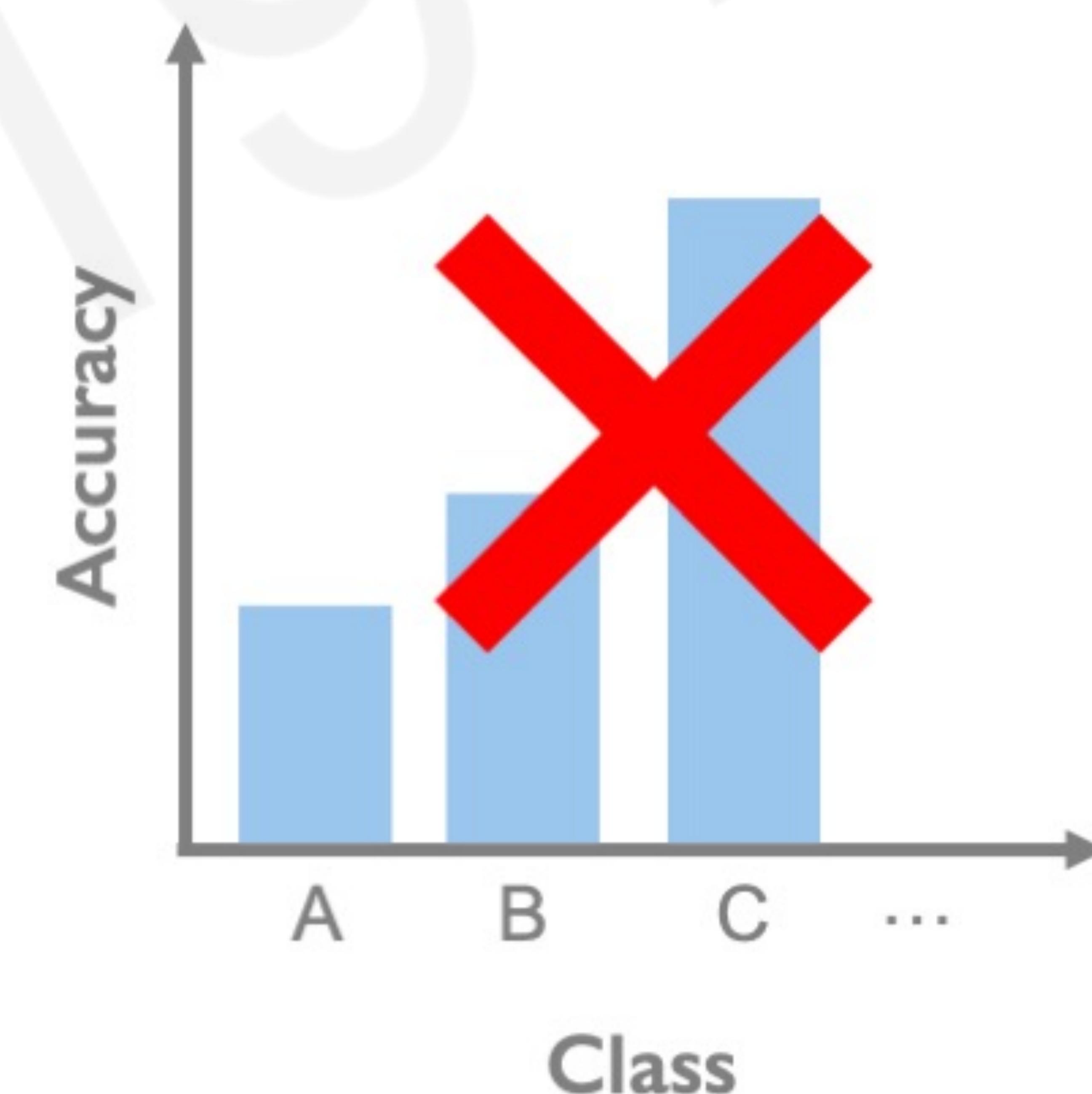
Frequency in Reality



Frequency in Dataset

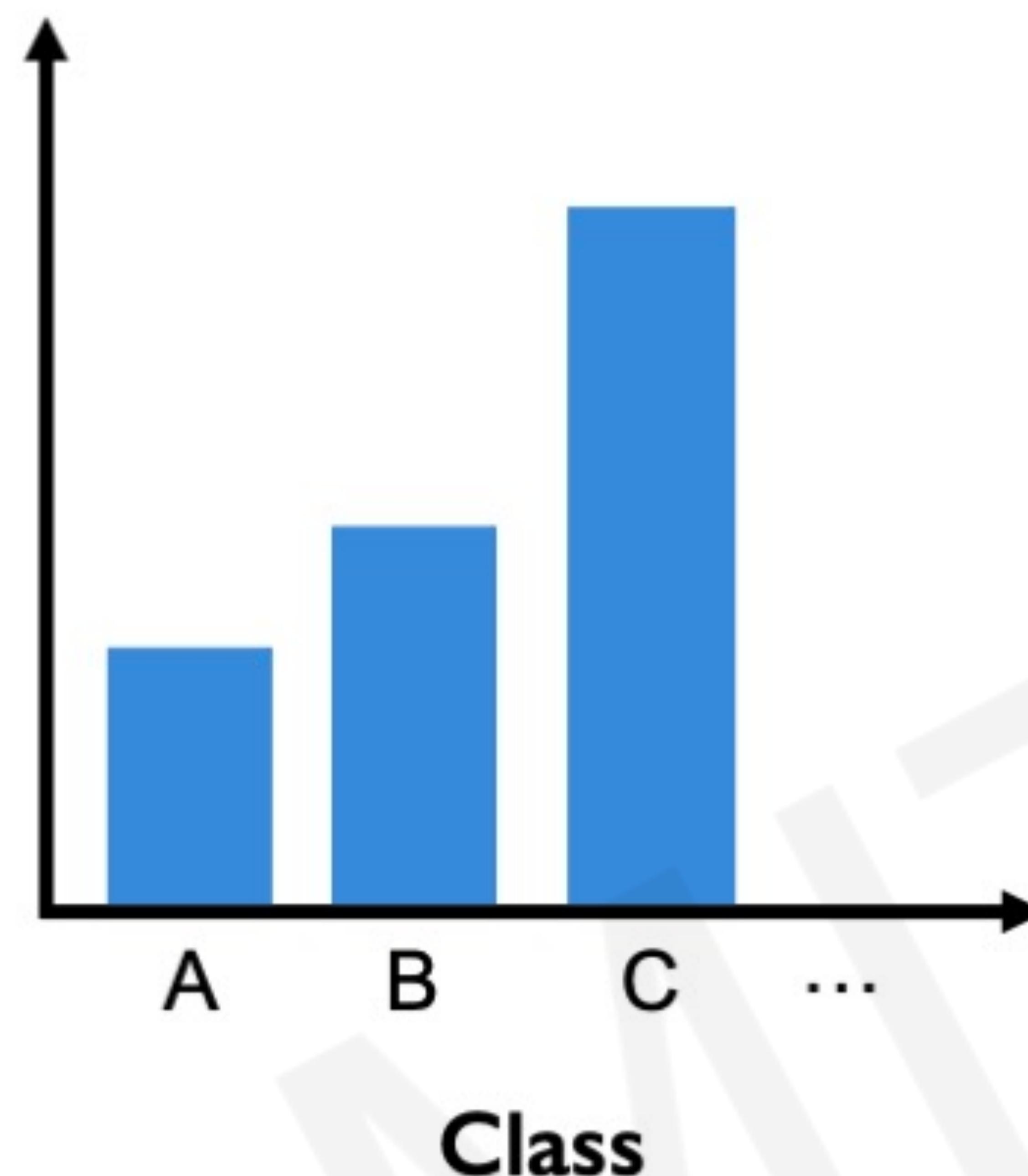


Model Accuracy

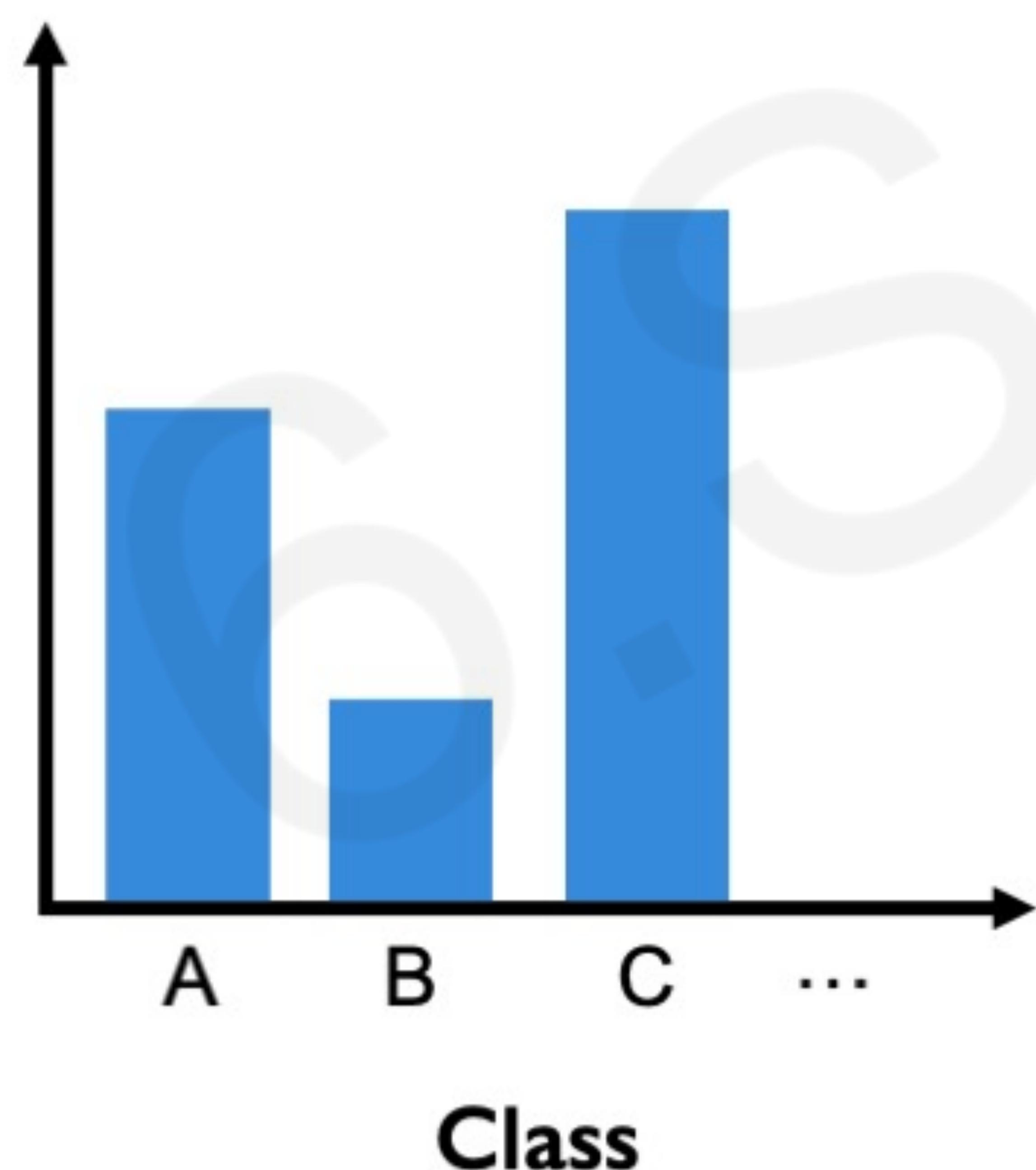


Biases due to Class Imbalance

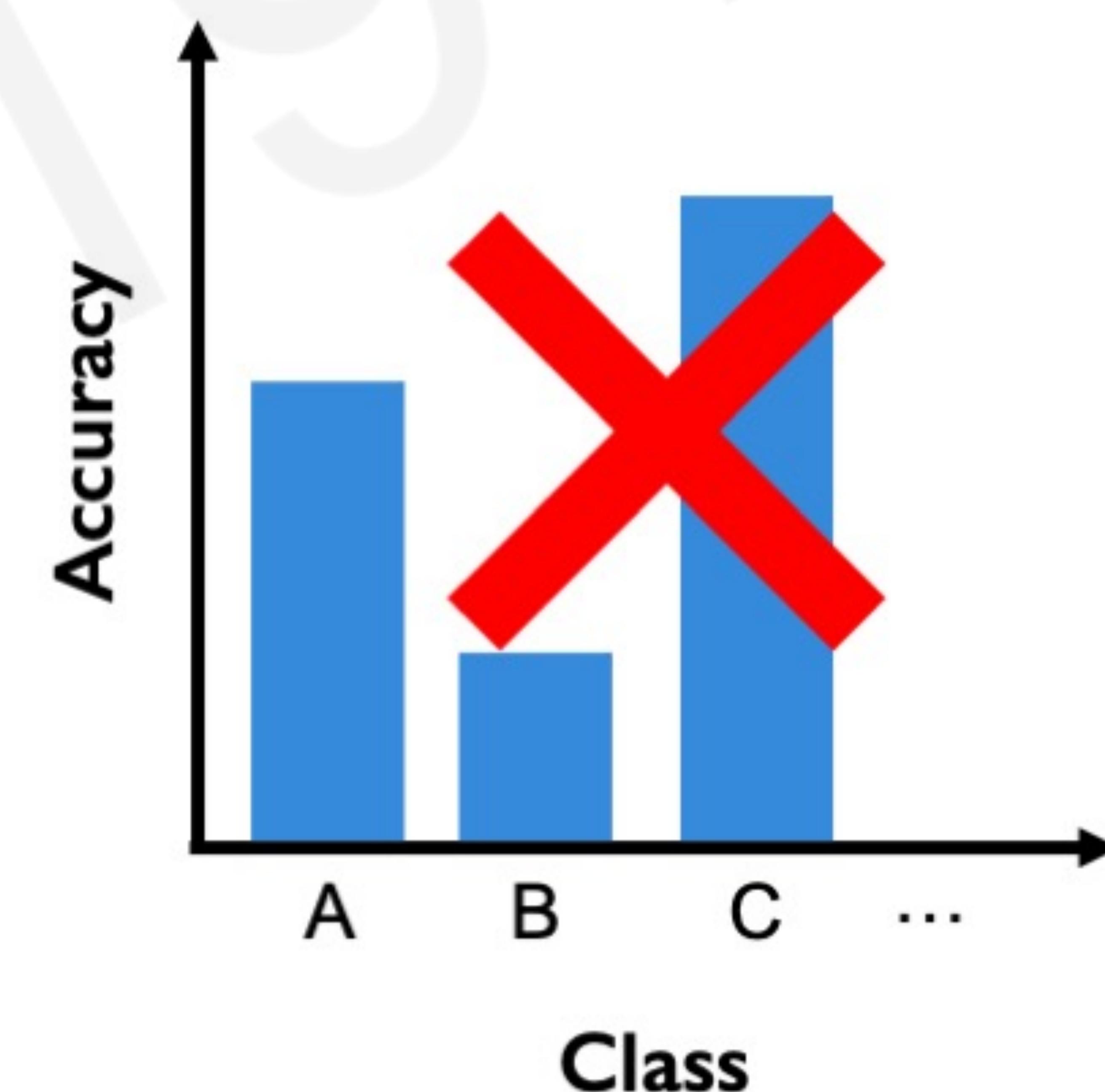
Frequency in Reality



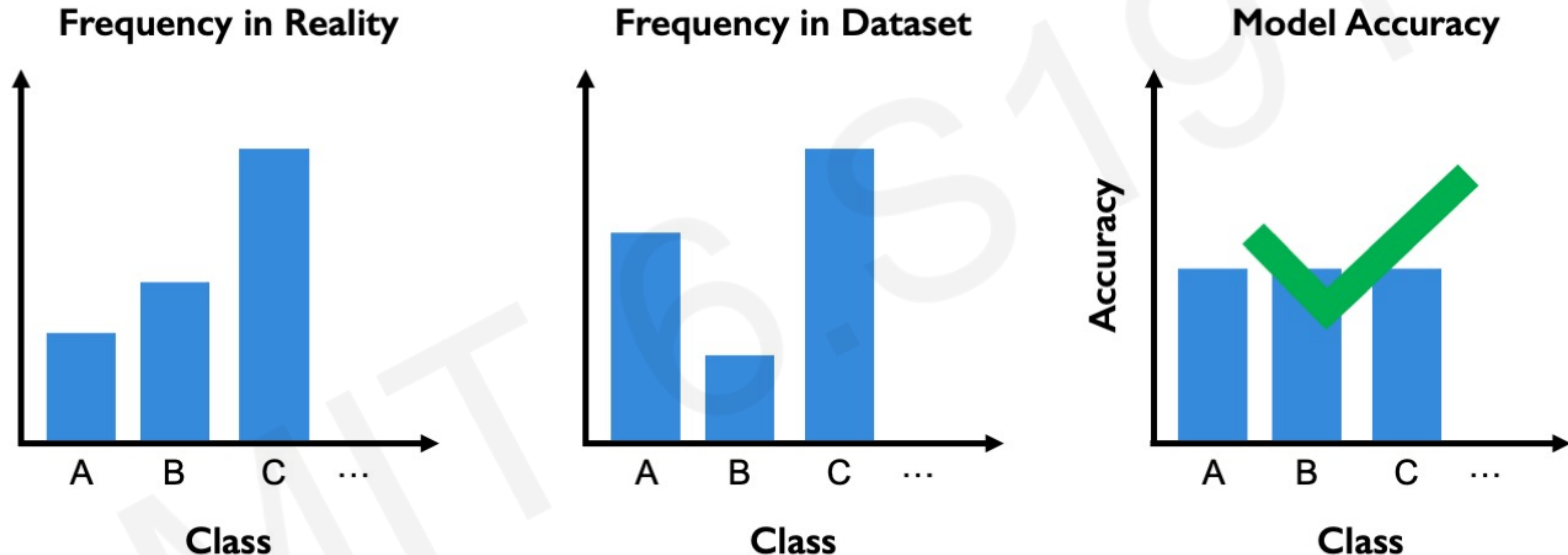
Frequency in Dataset



Model Accuracy



Biases due to Class Imbalance

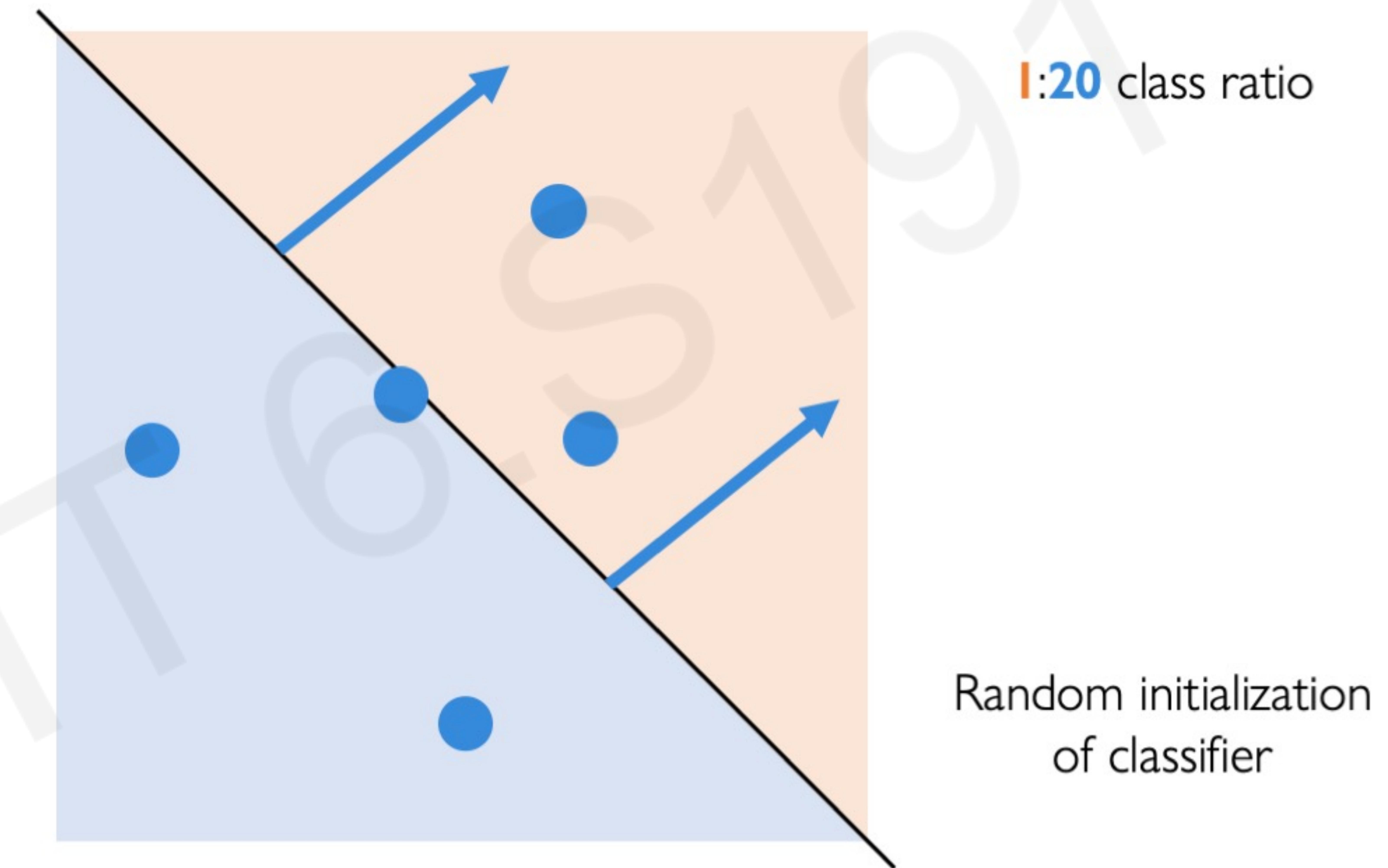


Goal: fair performance for all classes.
Why is class imbalance problematic?



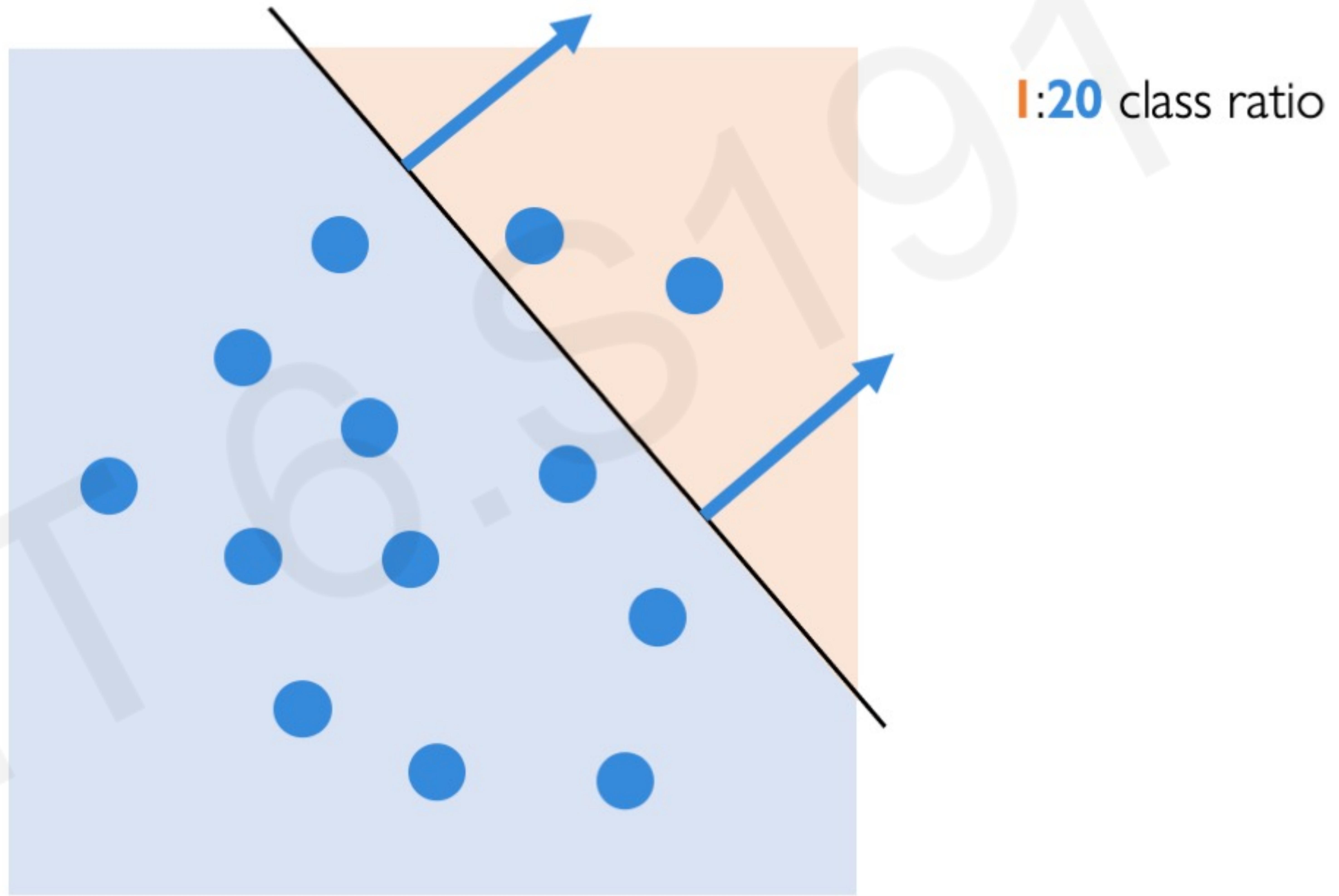
Learning in the Face of Class Imbalance

Incremental updates are made to the classifier during learning



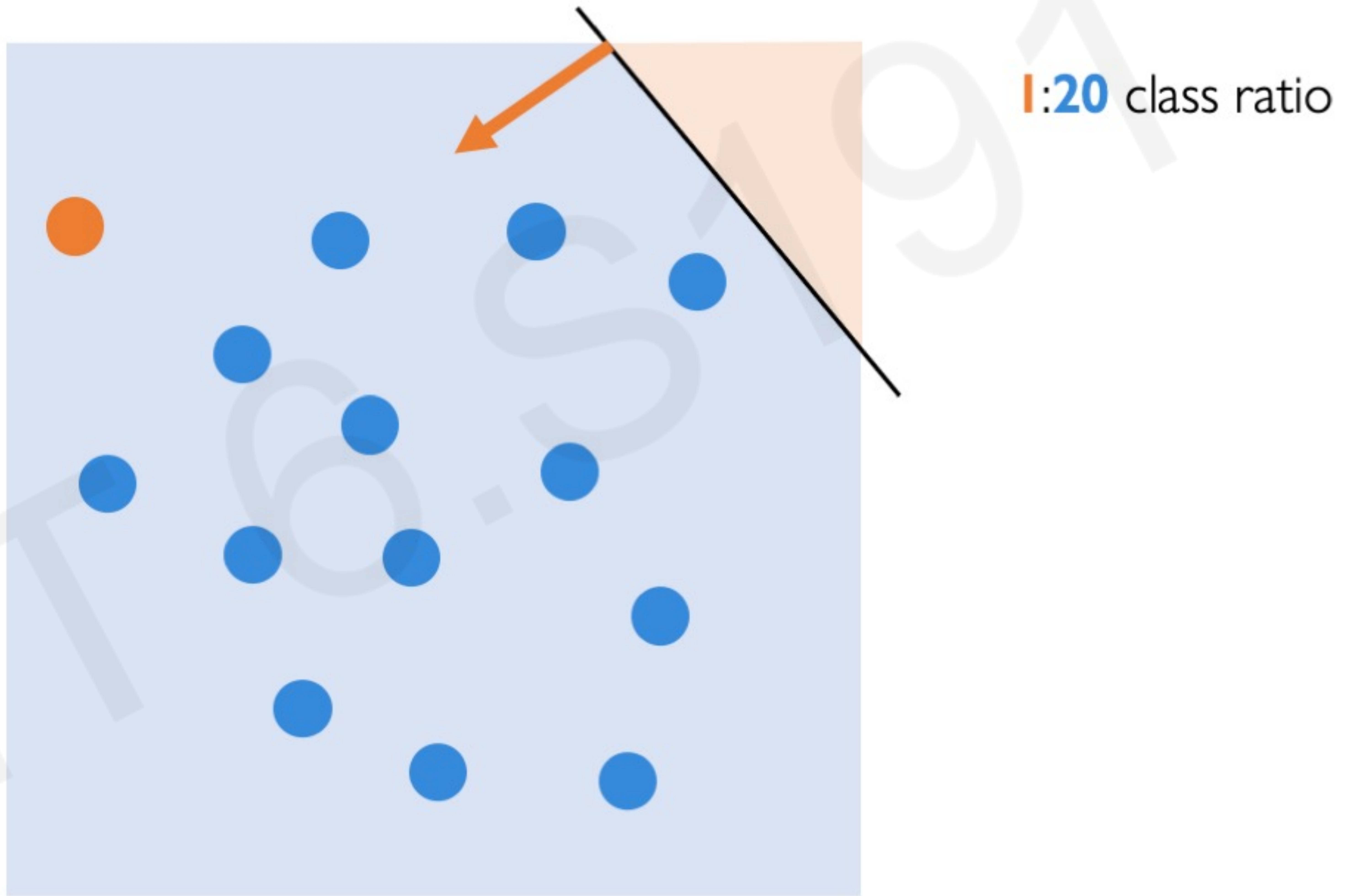
Learning in the Face of Class Imbalance

Incremental updates are made to the classifier during learning

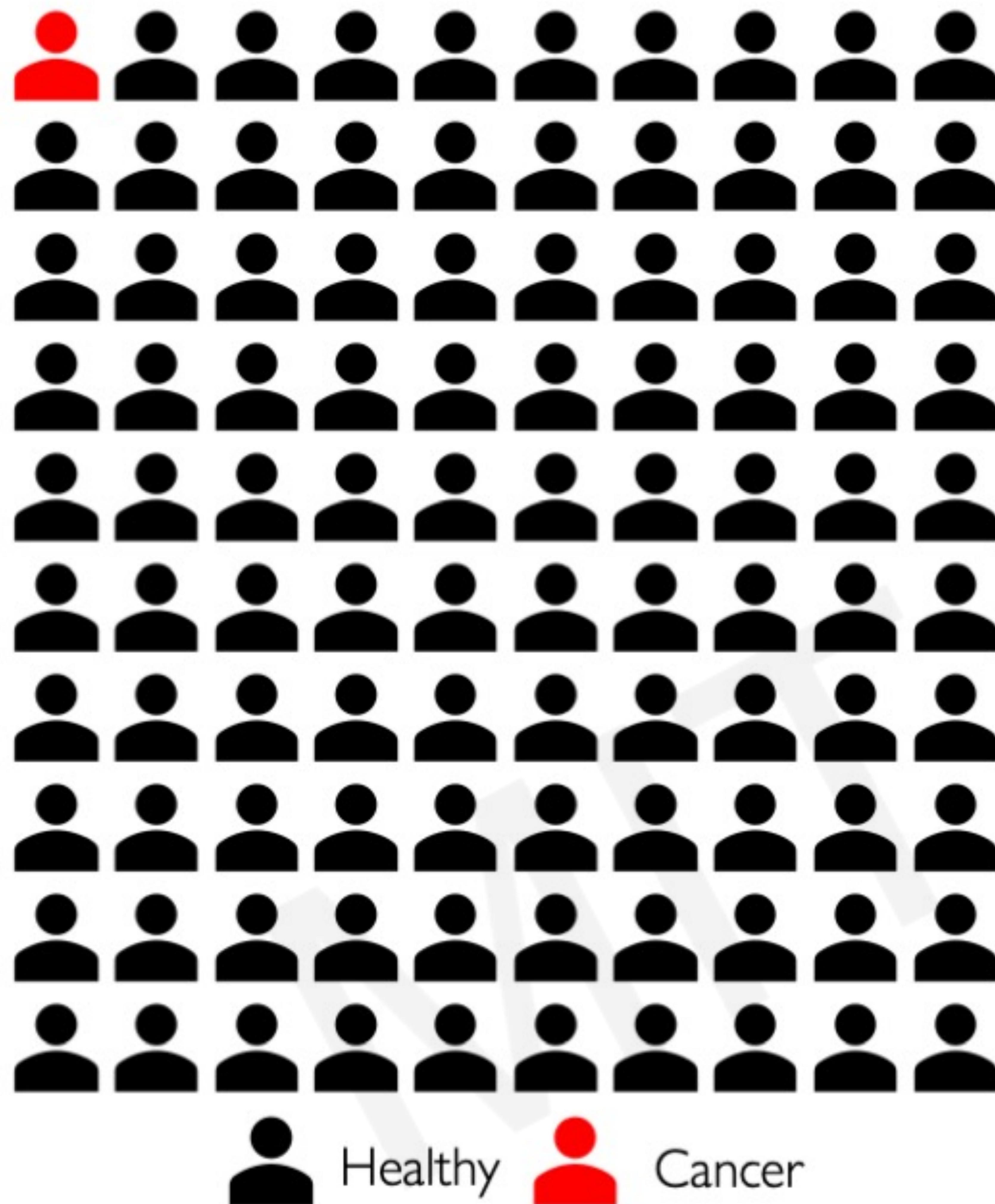


Learning in the Face of Class Imbalance

Incremental updates are made to the classifier during learning



Case Study: The Danger of Class Imbalance

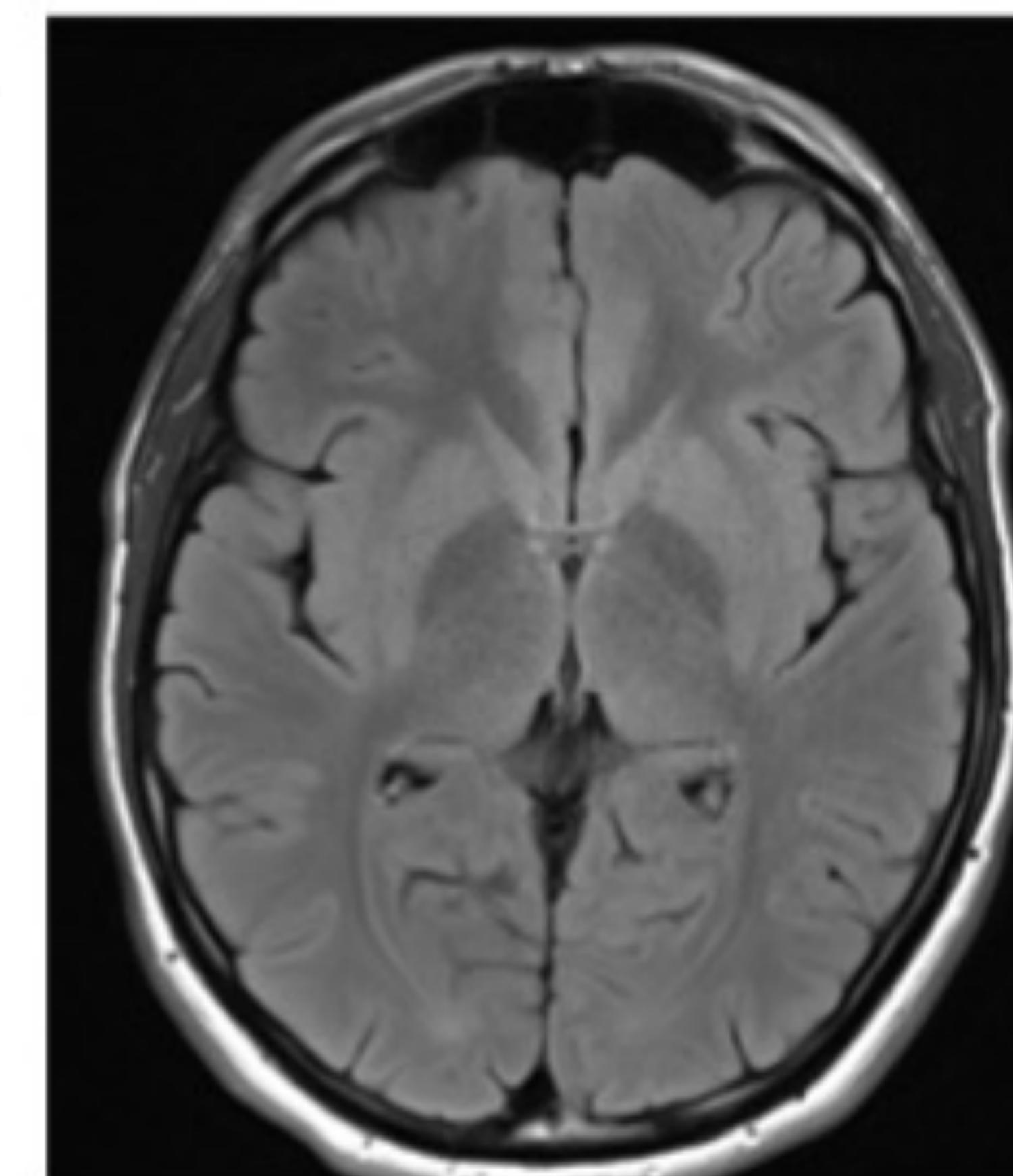


Case Study: Cancer Detection from Medical Images

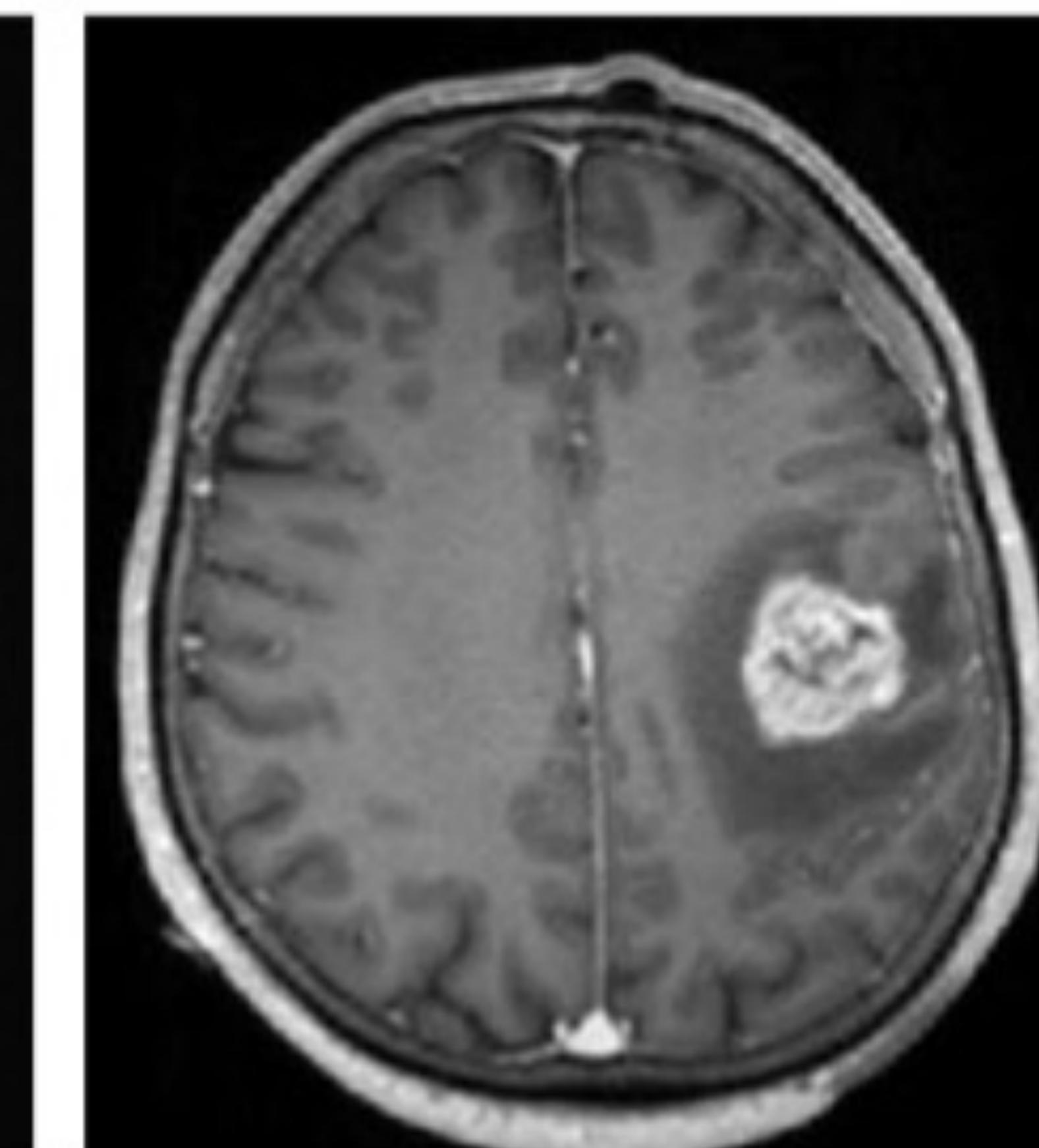
- Glioblastoma (GBM): most aggressive and deadliest brain tumor
- GBM incidence in USA: **3.19 per 100,000** individuals!
- Task: train CNN to detect GBM from MRI scans of the brain



What if class incidence in dataset reflected real-world incidence?



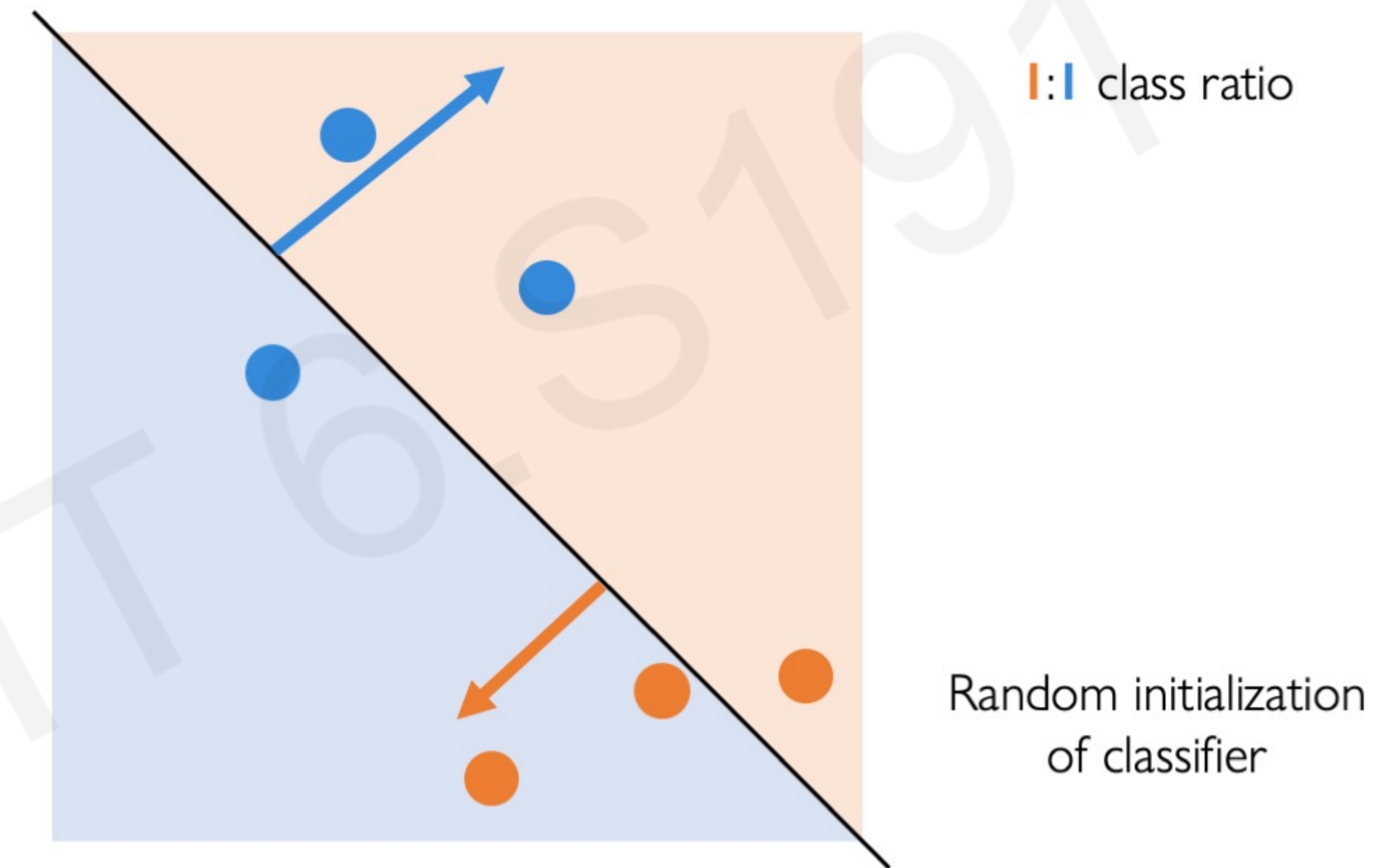
99.997%



0.003%

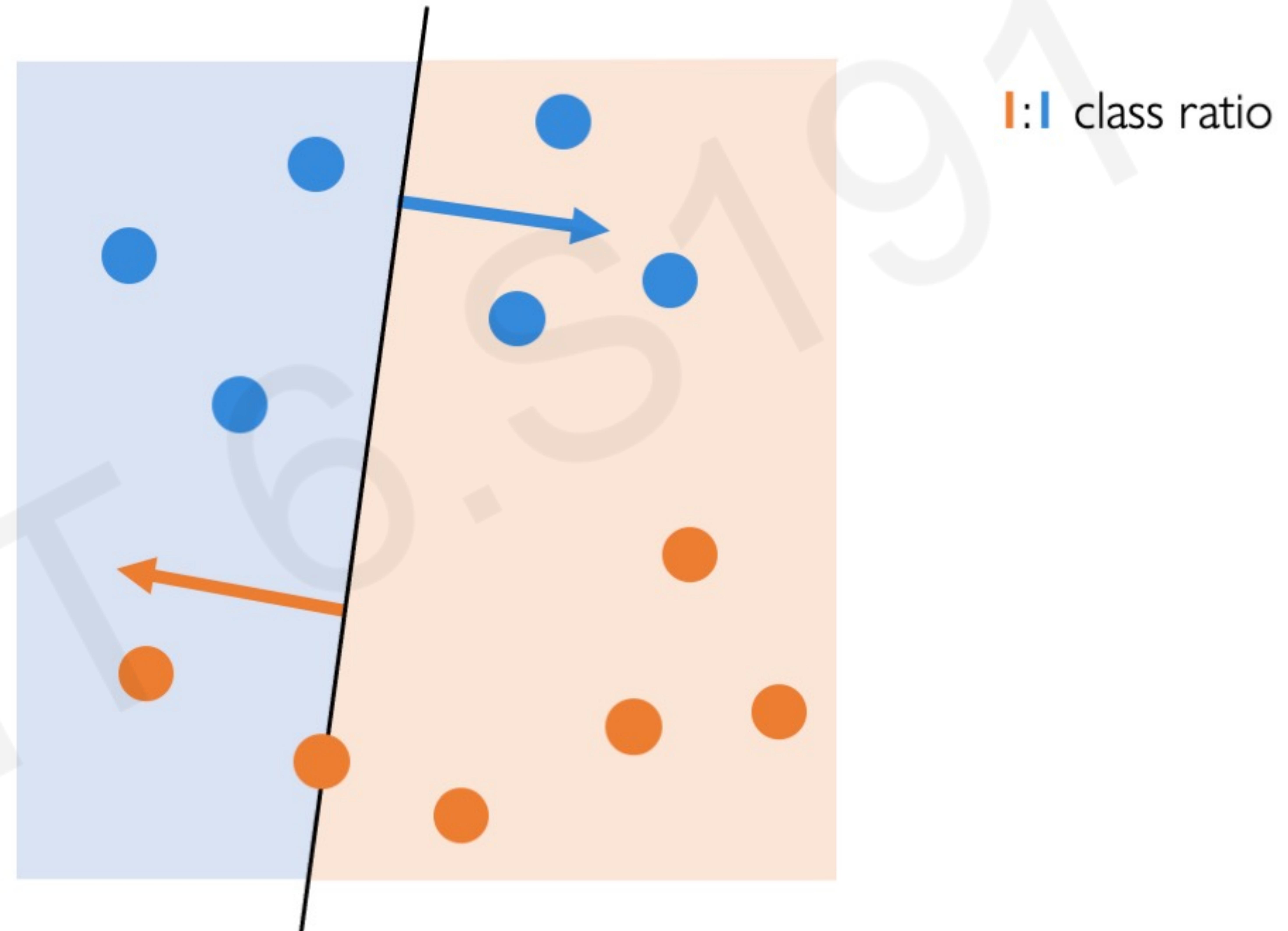
Learning from Class Balanced Data: Batch Selection

Incremental updates are made to the classifier during learning

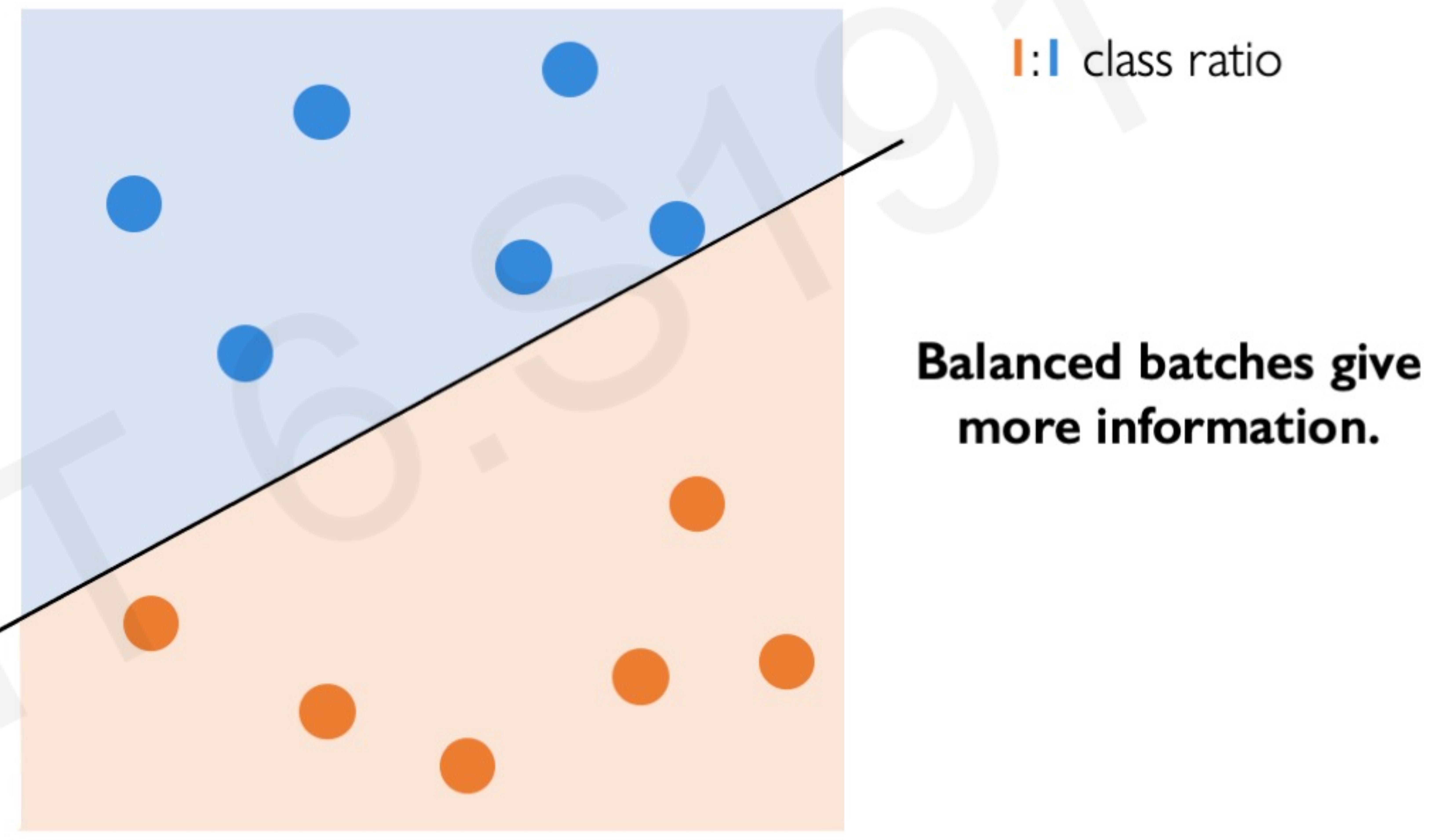


Learning from Class Balanced Data: Batch Selection

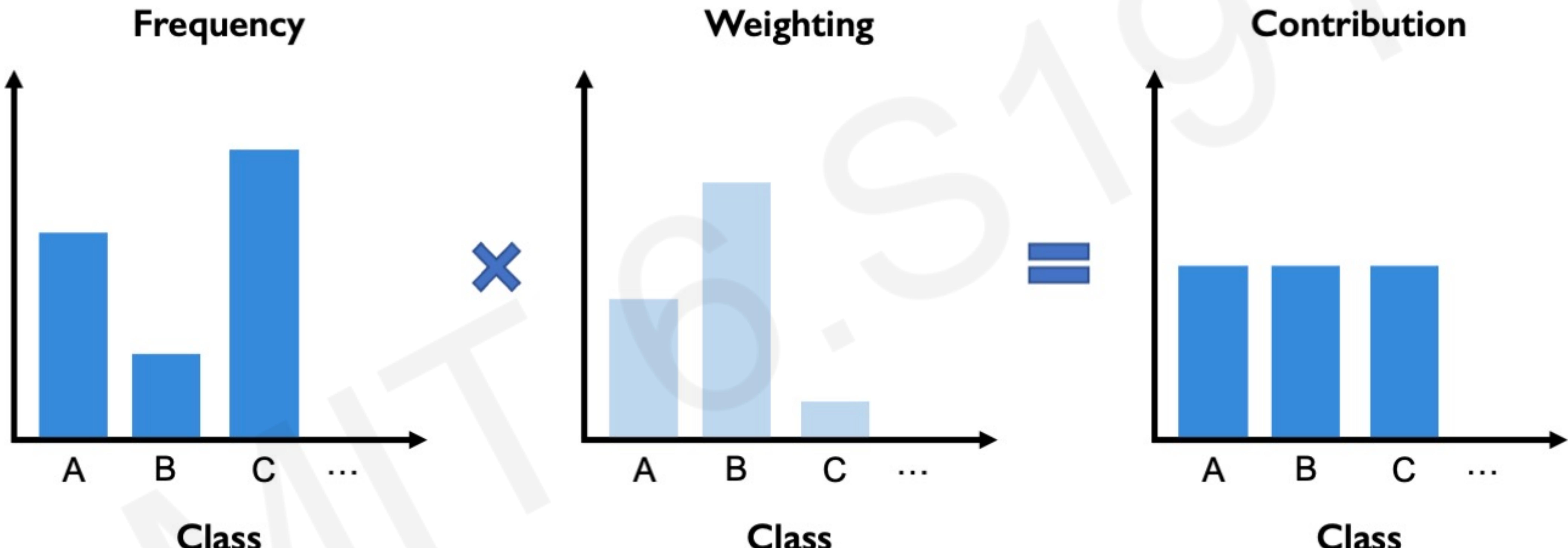
Incremental updates are made to the classifier during learning



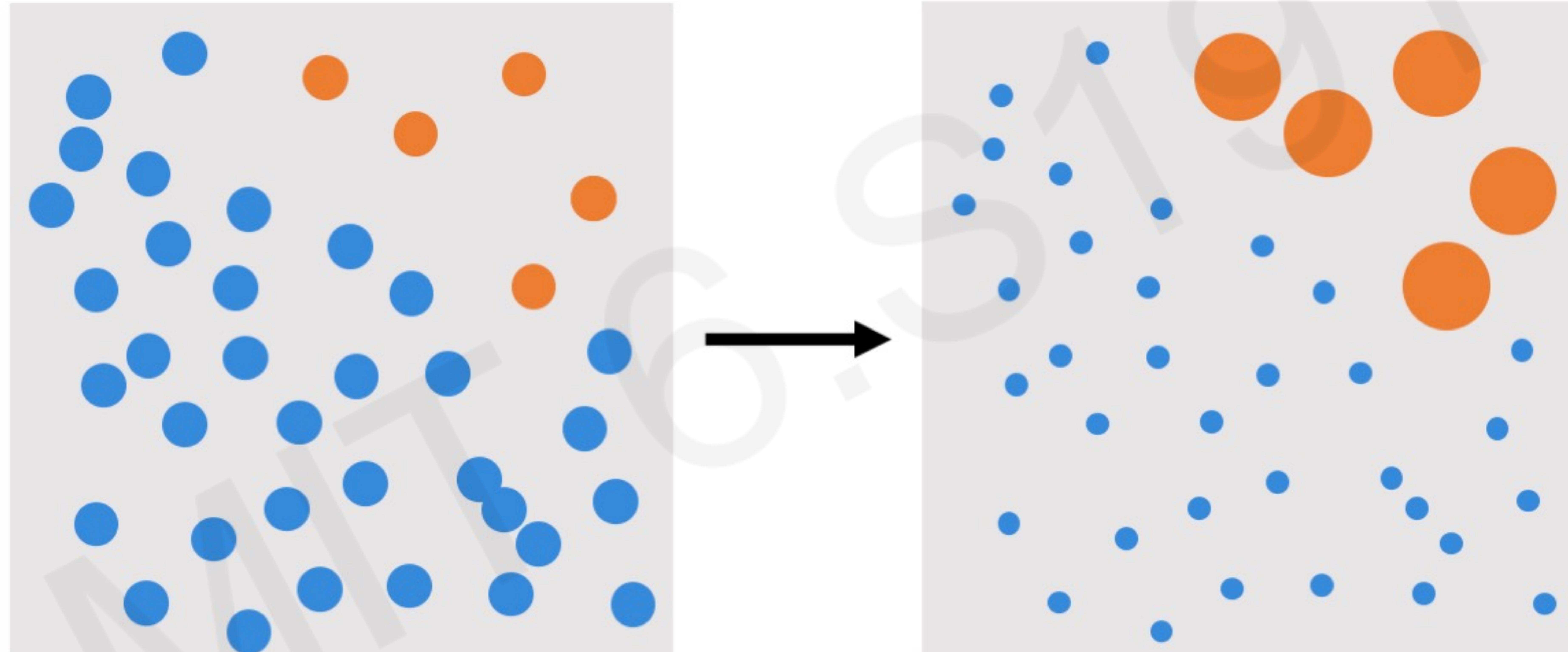
Learning from Class Balanced Data: Batch Selection



Learning from Class Balanced Data: Example Weighting



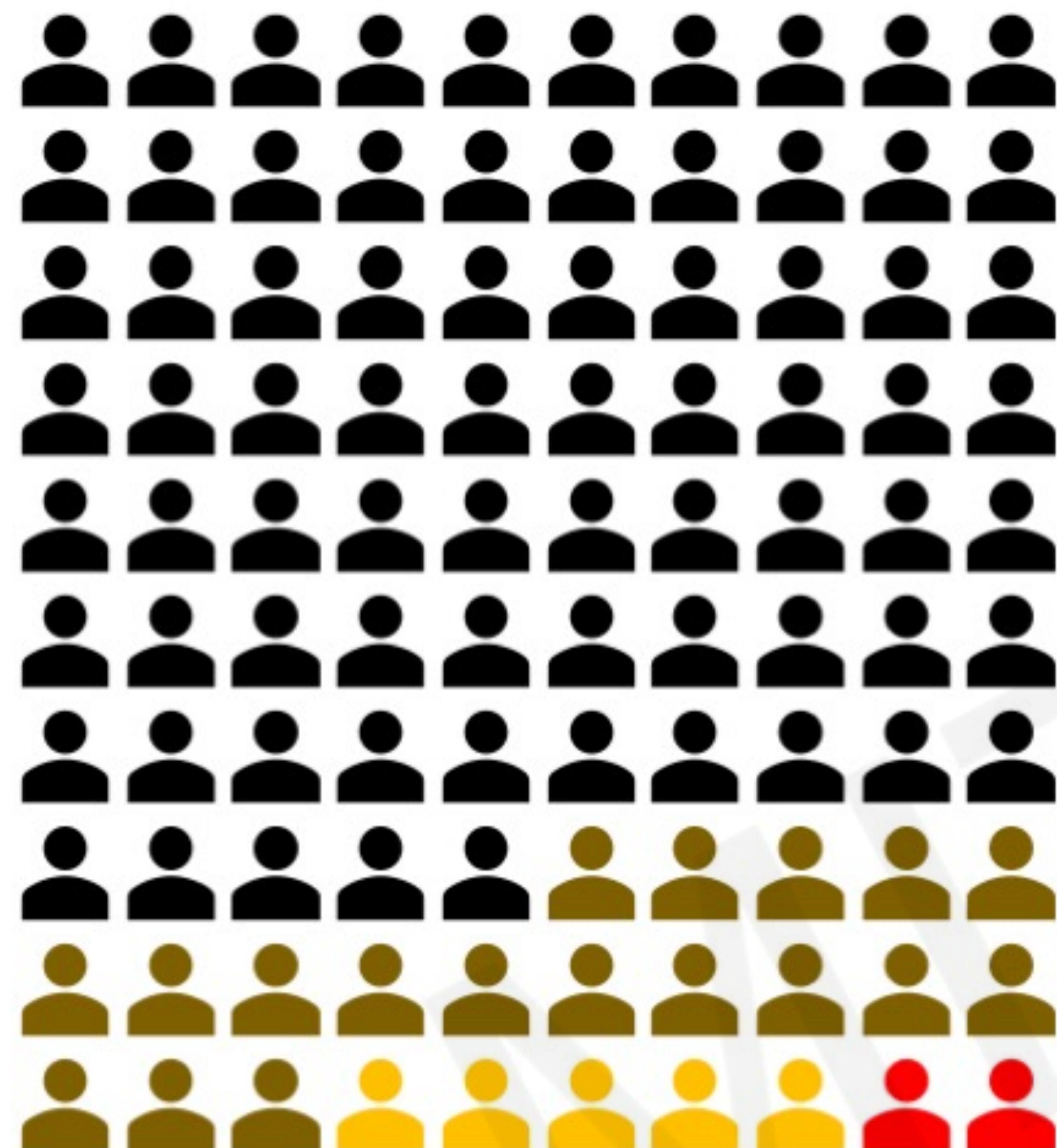
Learning from Class Balanced Data: Example Weighting



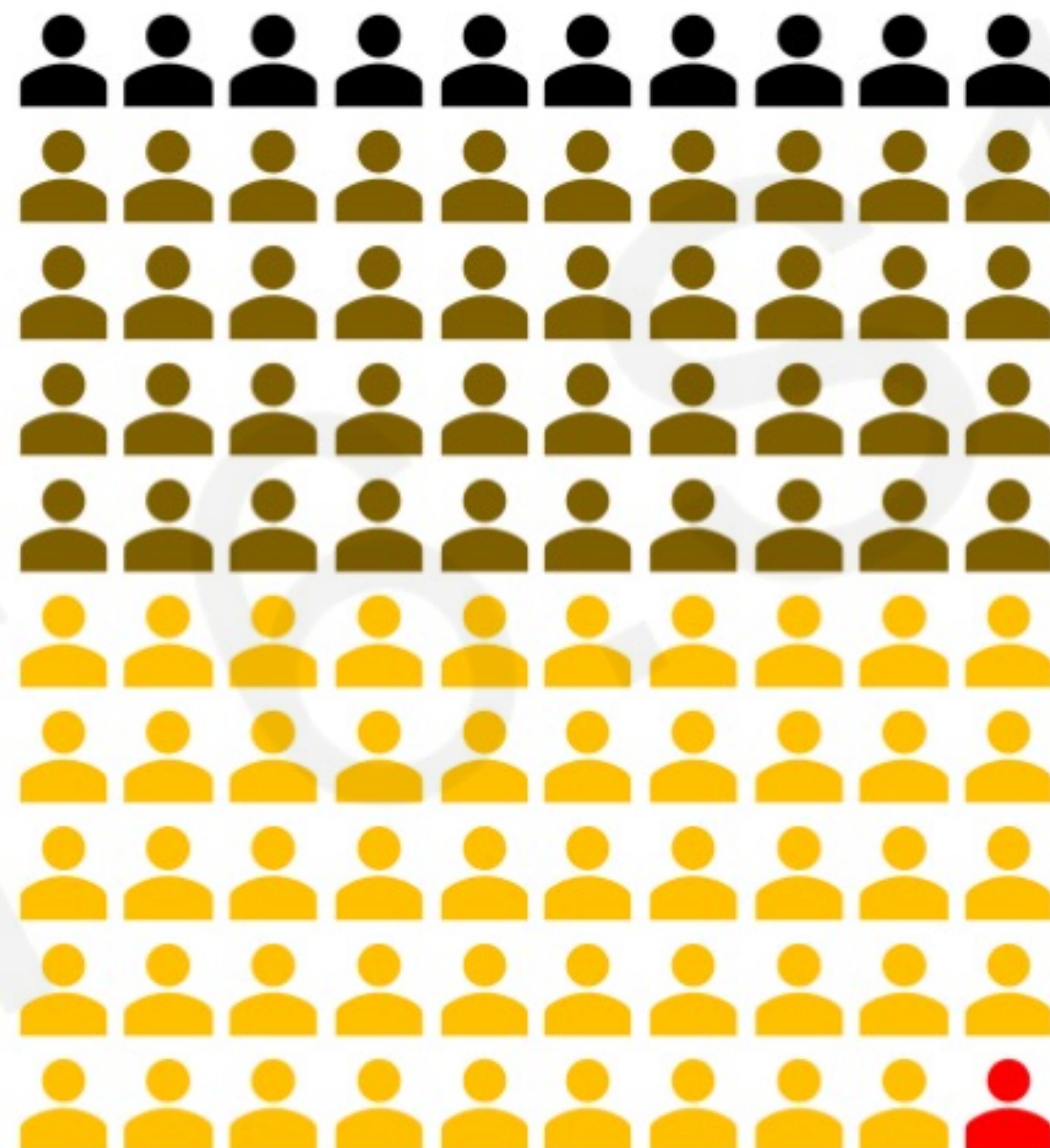
Size == probability of selection during training

Case Study: Hidden Bias in Facial Detection

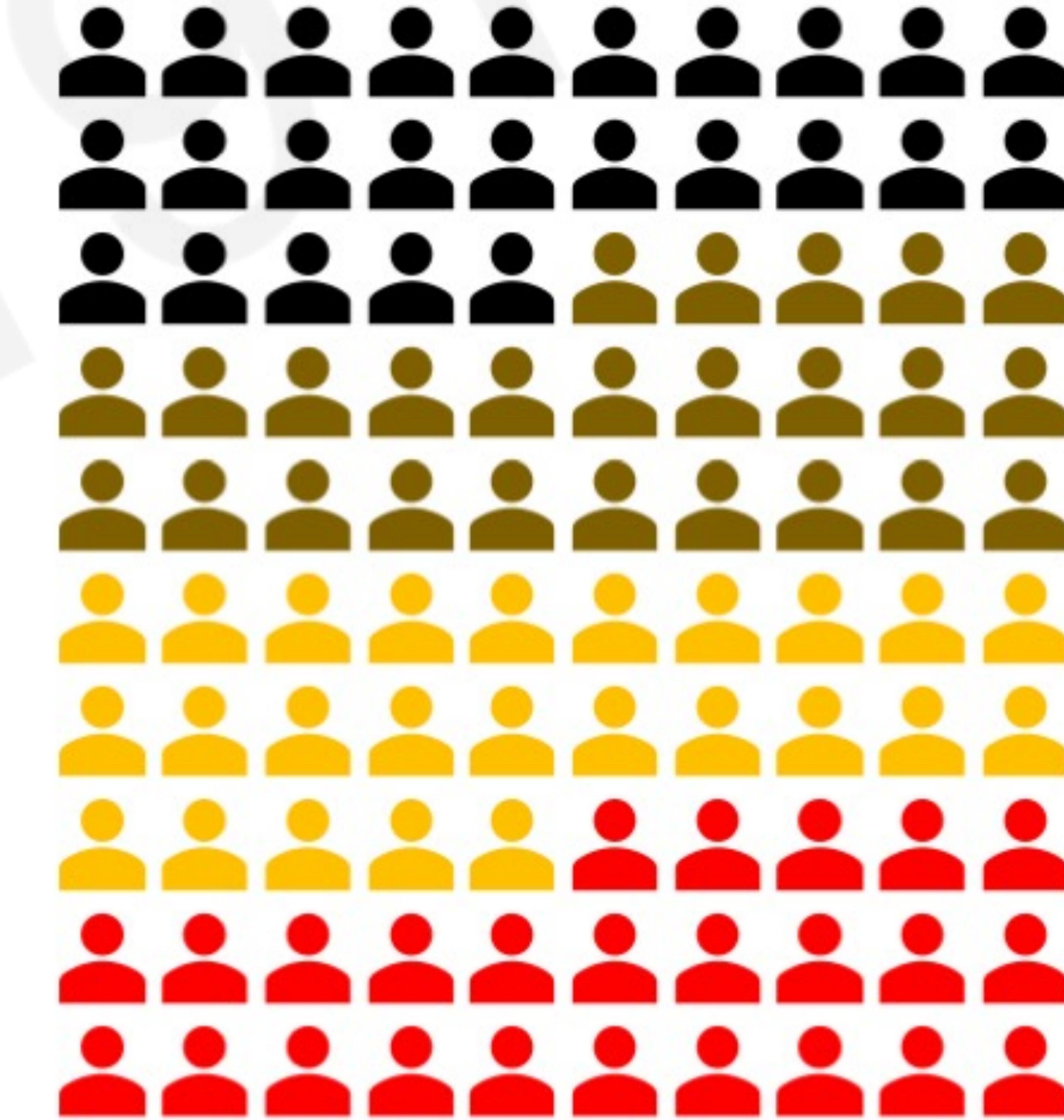
Real World



“Gold-Standard” Dataset



Balanced Dataset



Black Hair



Brown Hair



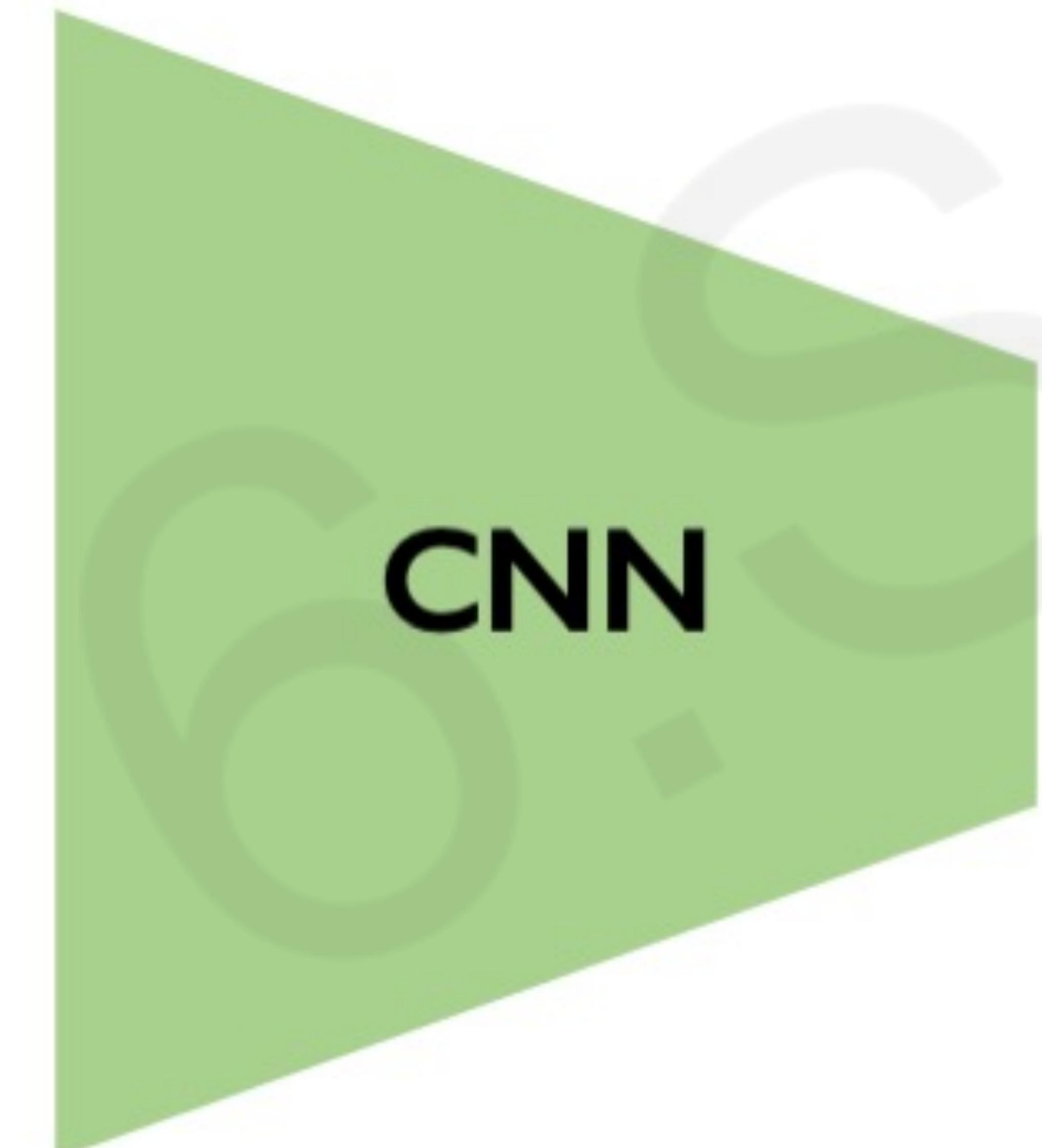
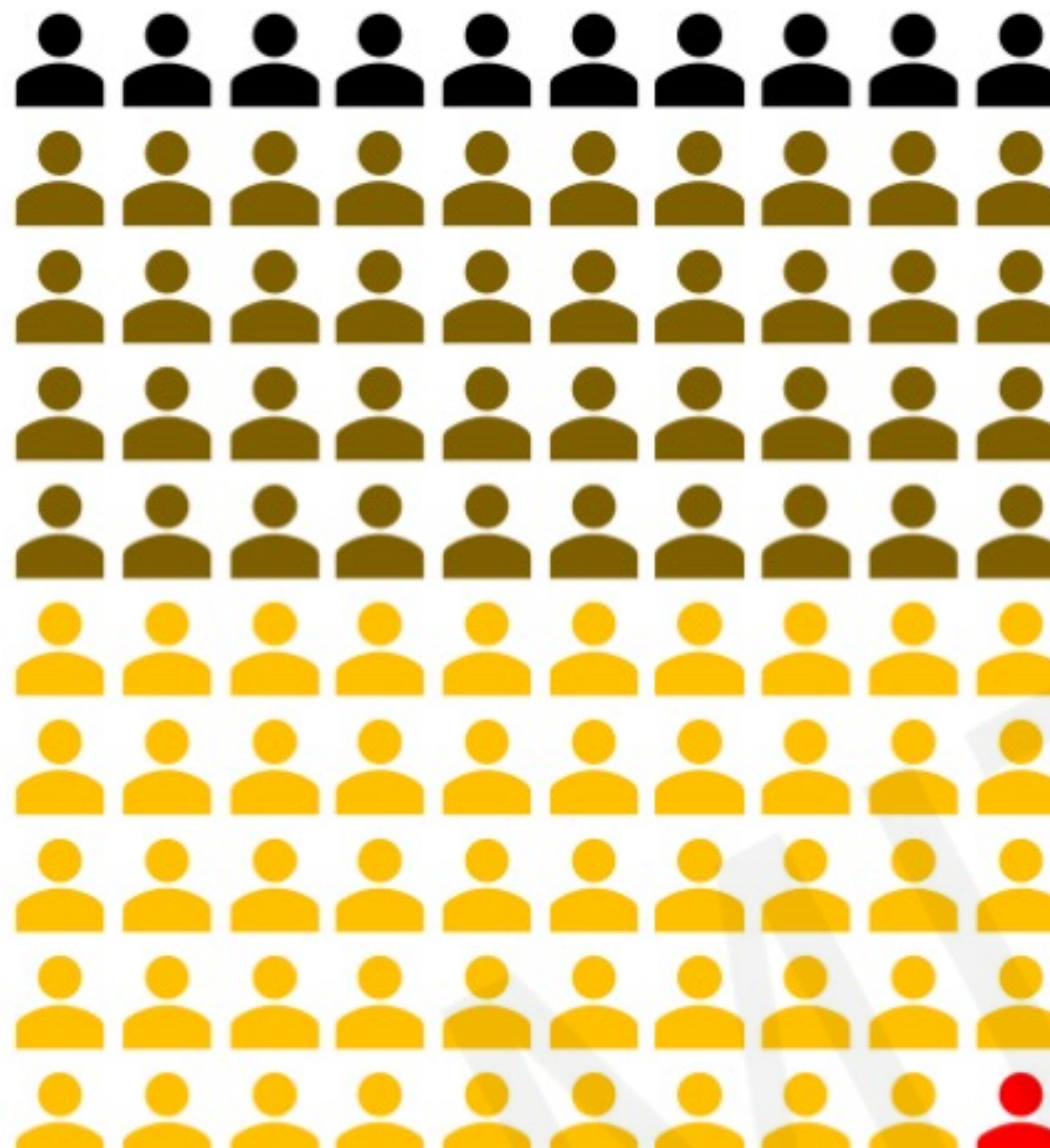
Blonde Hair



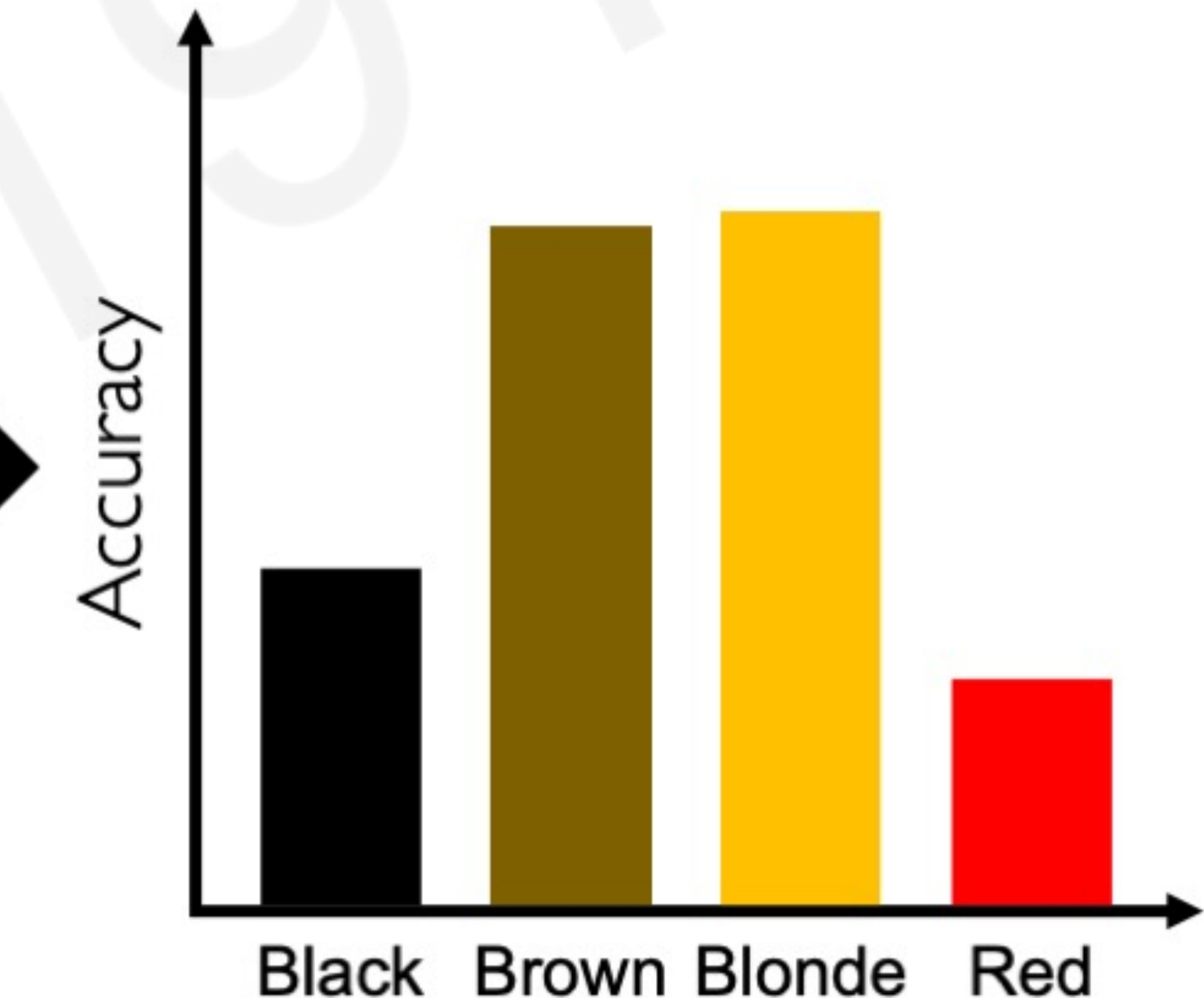
Red Hair

Case Study: Hidden Bias in Facial Detection

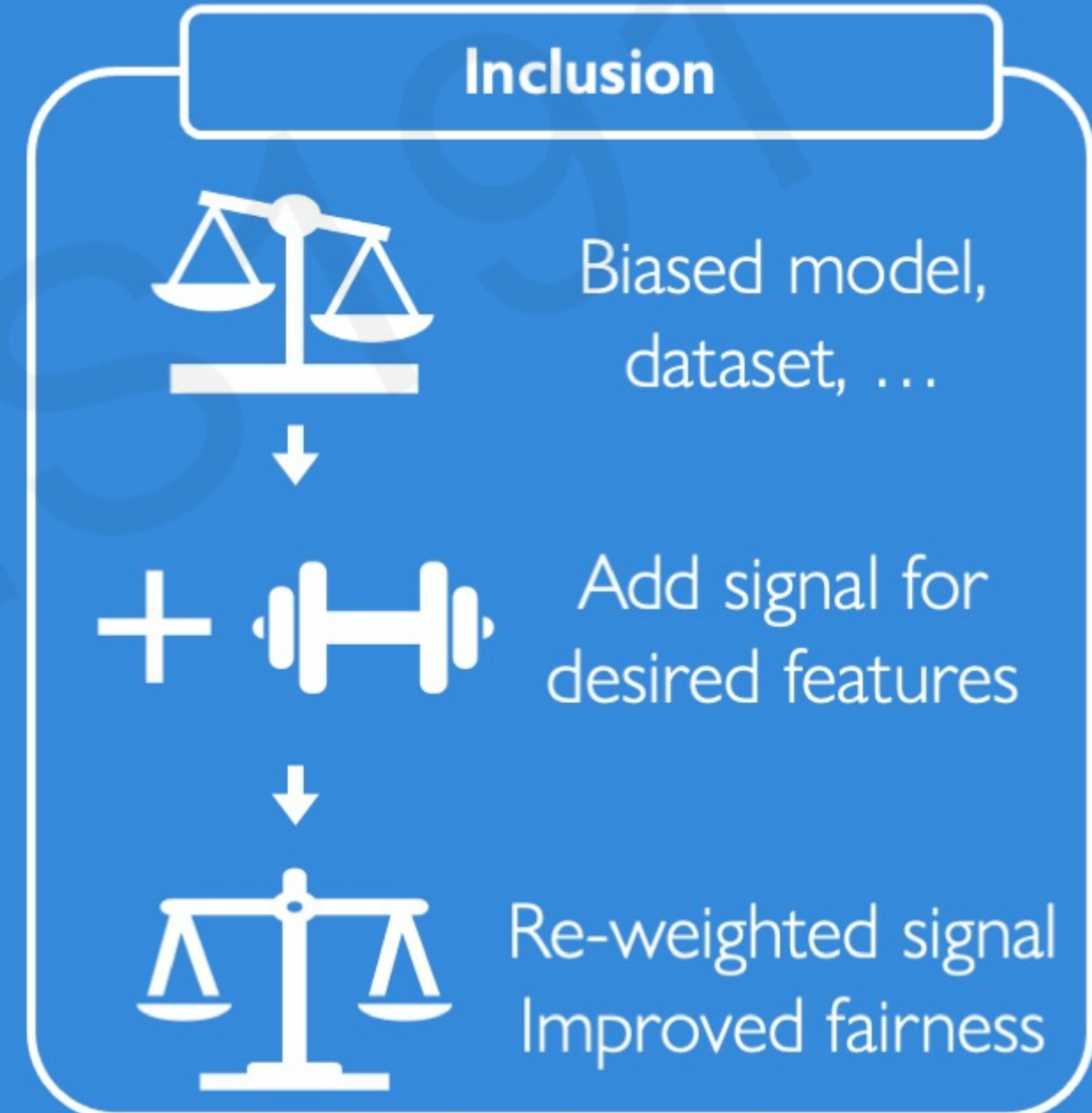
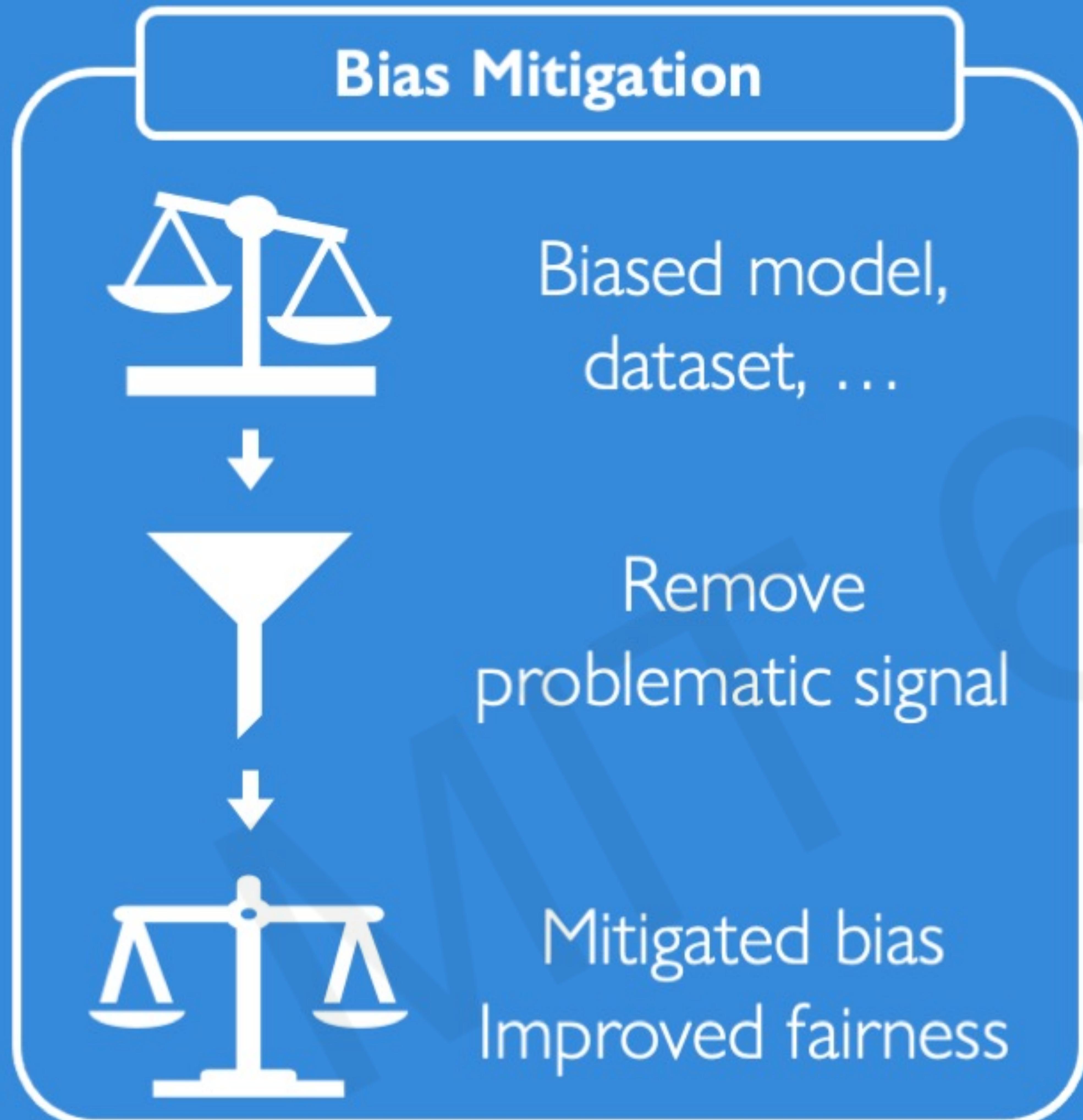
“Gold-Standard” Dataset



Train CNN for
facial detection.



Learning Techniques to Improve Fairness



Bias and Fairness in Supervised Classification

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

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

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

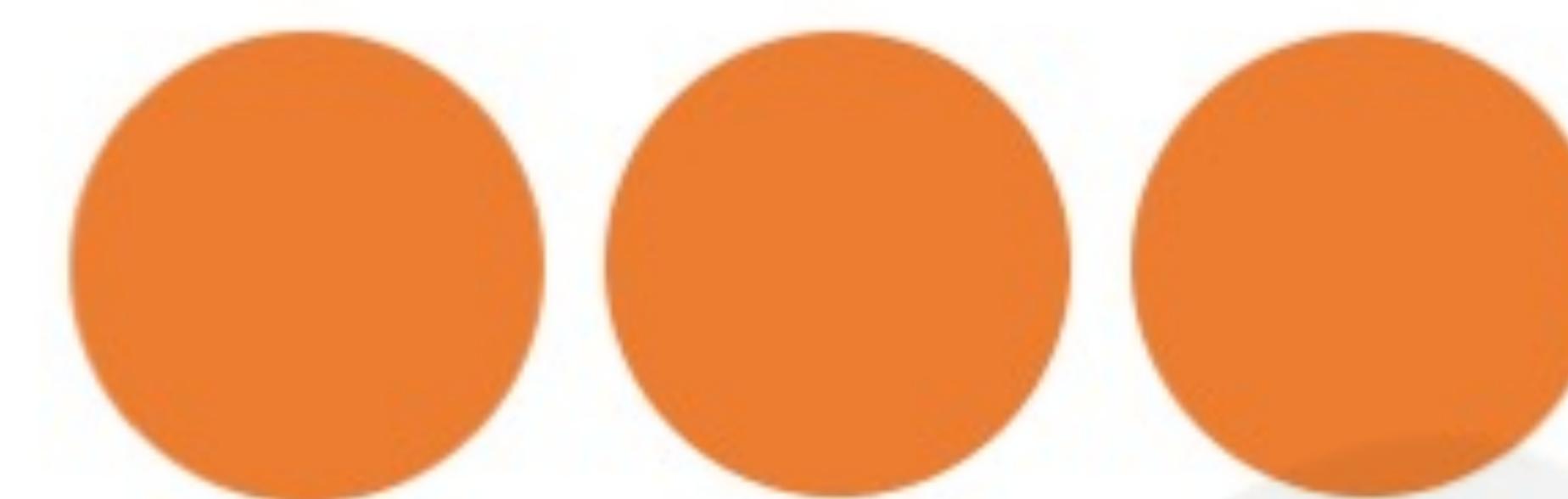
For example, for a single binary variable z , fairness means:

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

Evaluating Bias and Fairness

Disaggregated evaluation: evaluate performance with respect to different subgroups

Color

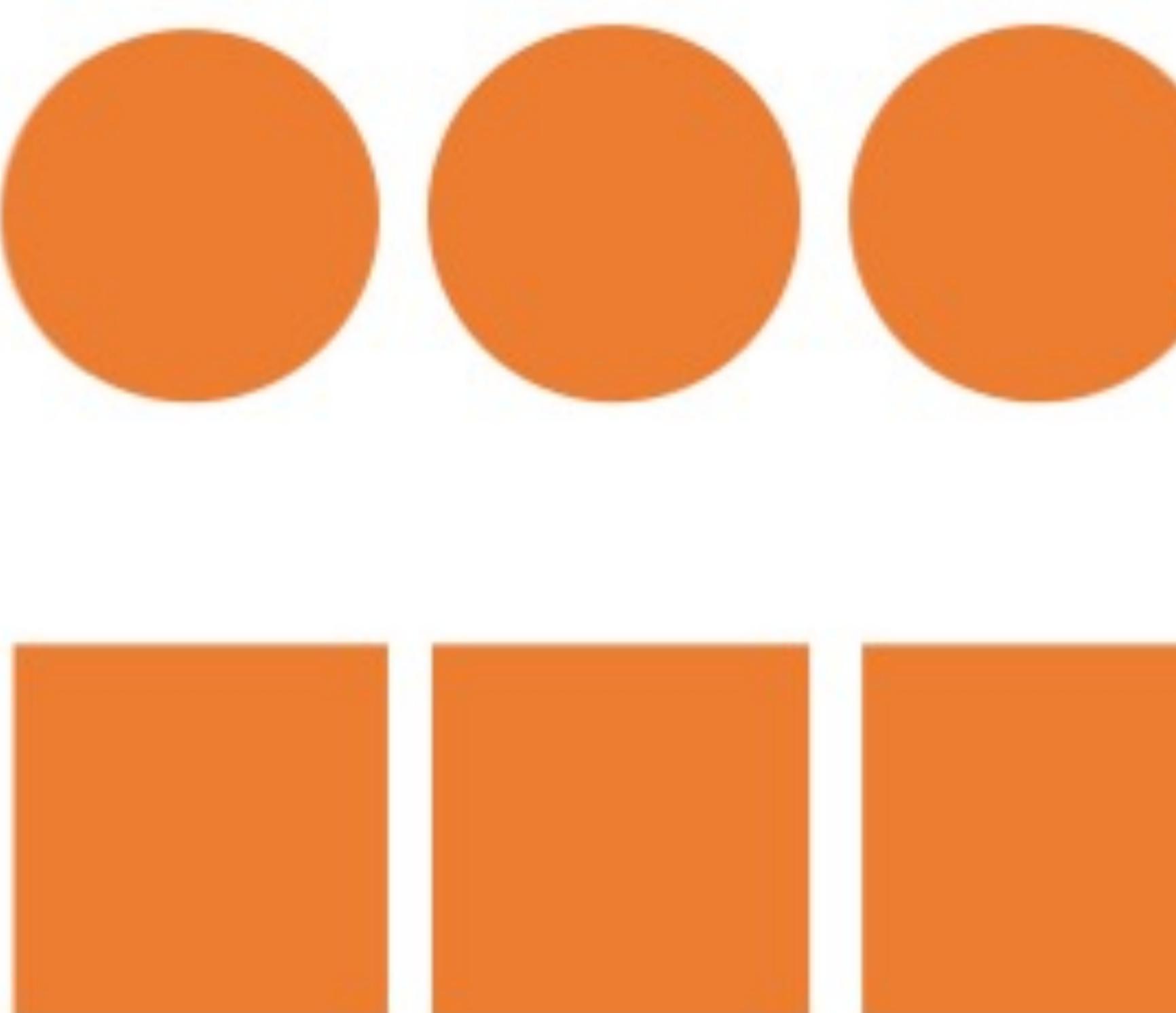
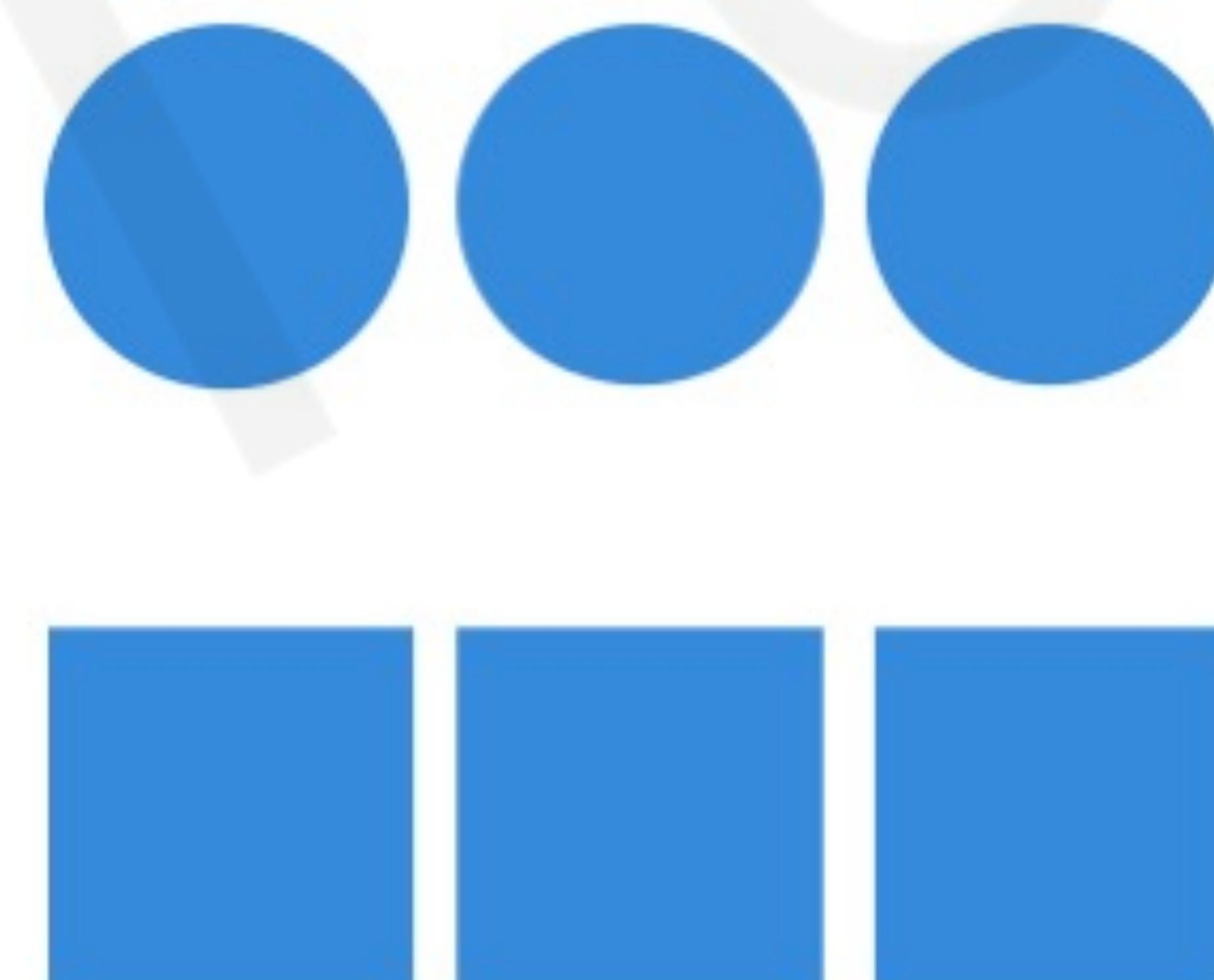


Shape



Intersectional evaluation: evaluate performance with respect to subgroup intersections

Color & Shape



Adversarial Multi-Task Learning to Mitigate Bias

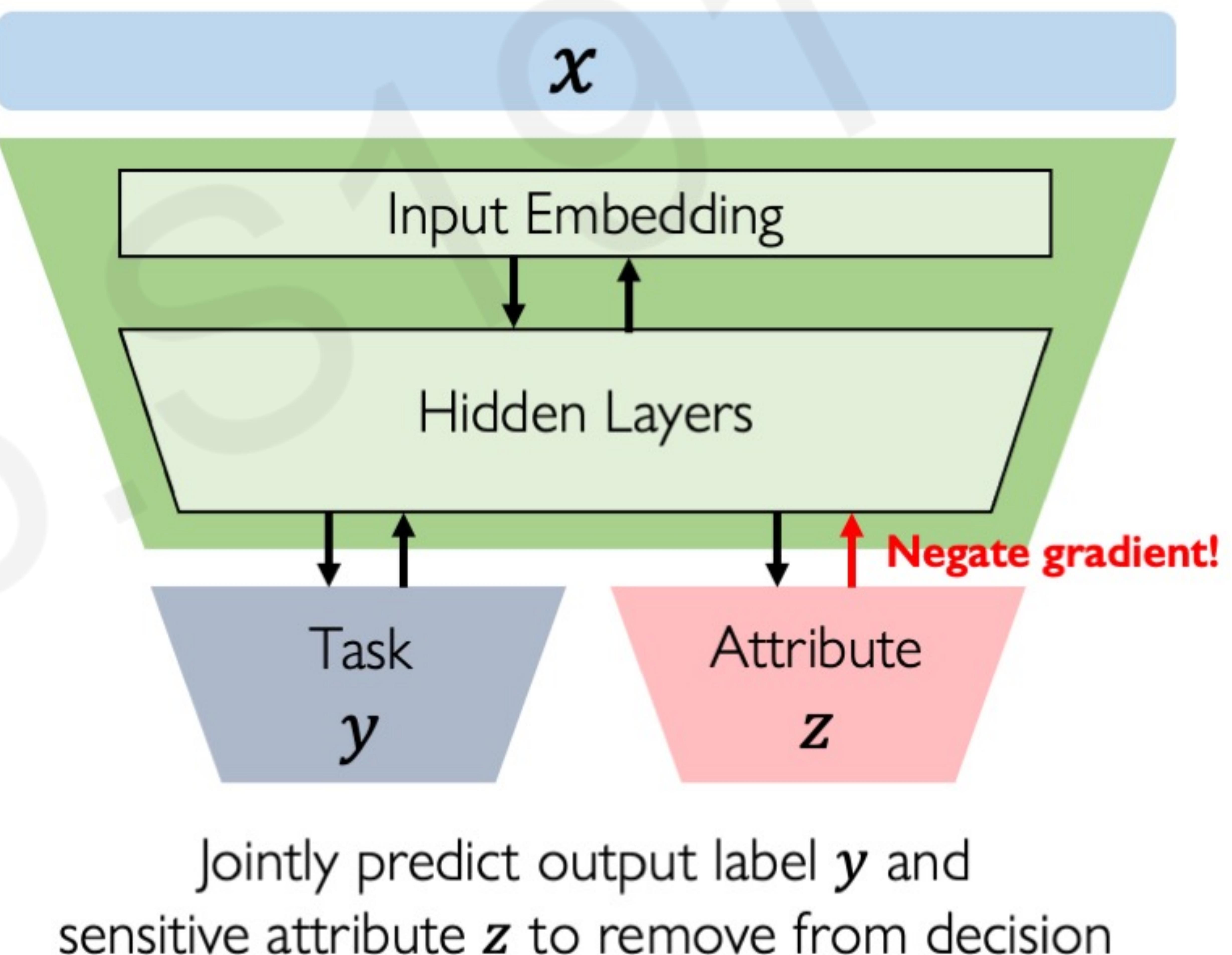
Setup: specify attribute z for which we seek to mitigate bias. Jointly predict output y and z .

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



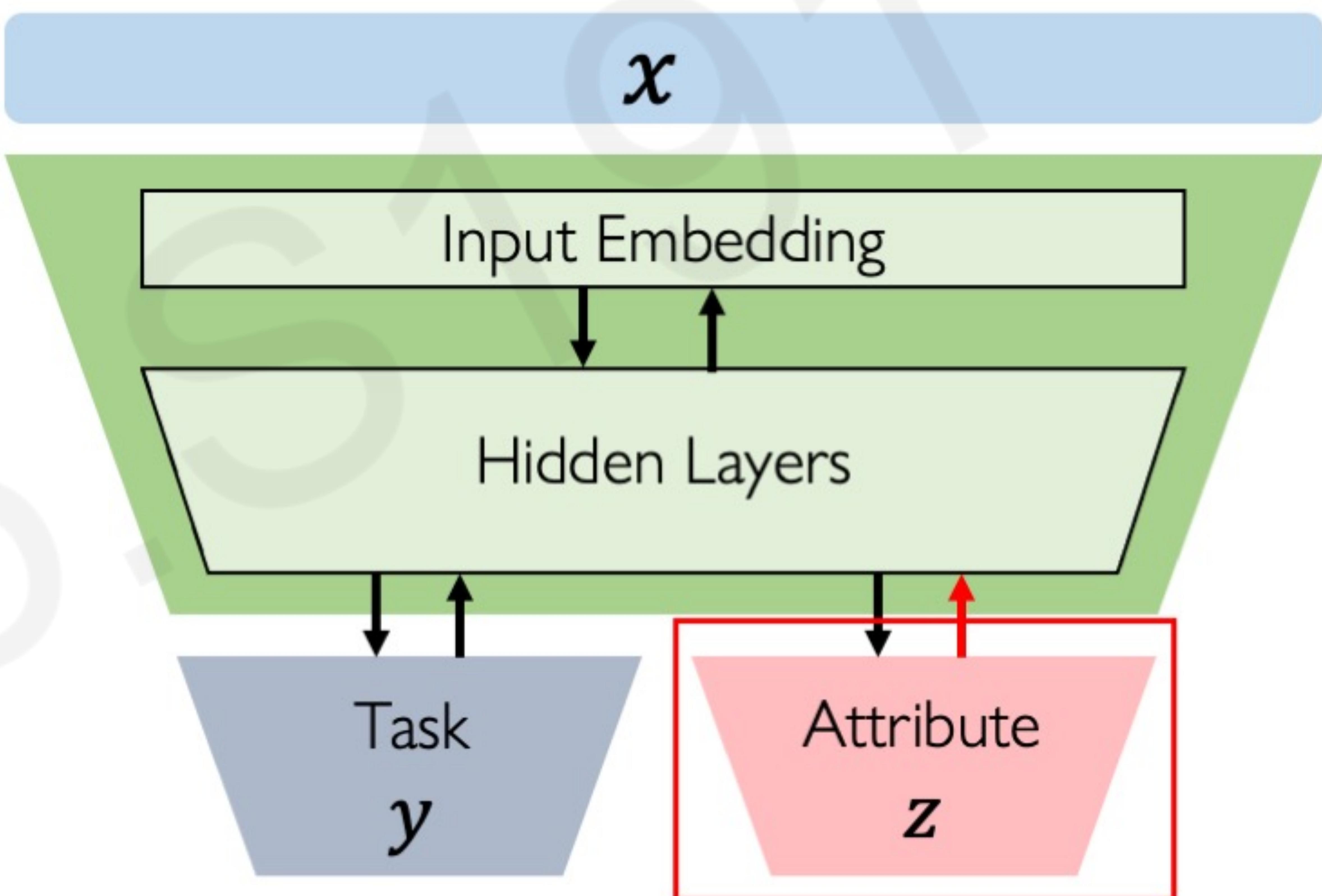
Application to Language Modeling

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

Adaptive Resampling for Automated Debiasing

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



Homogeneous skin color, pose

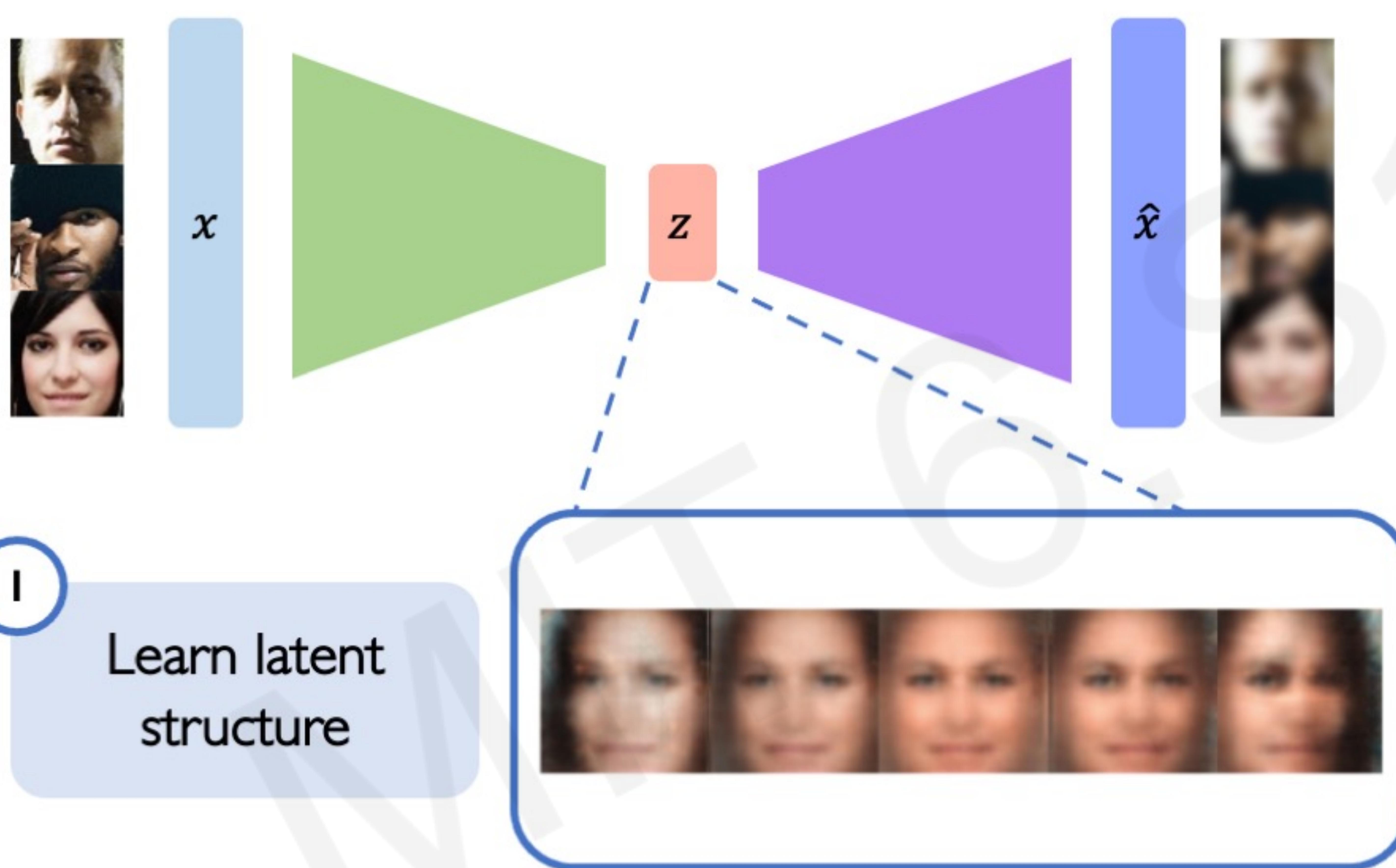
VS



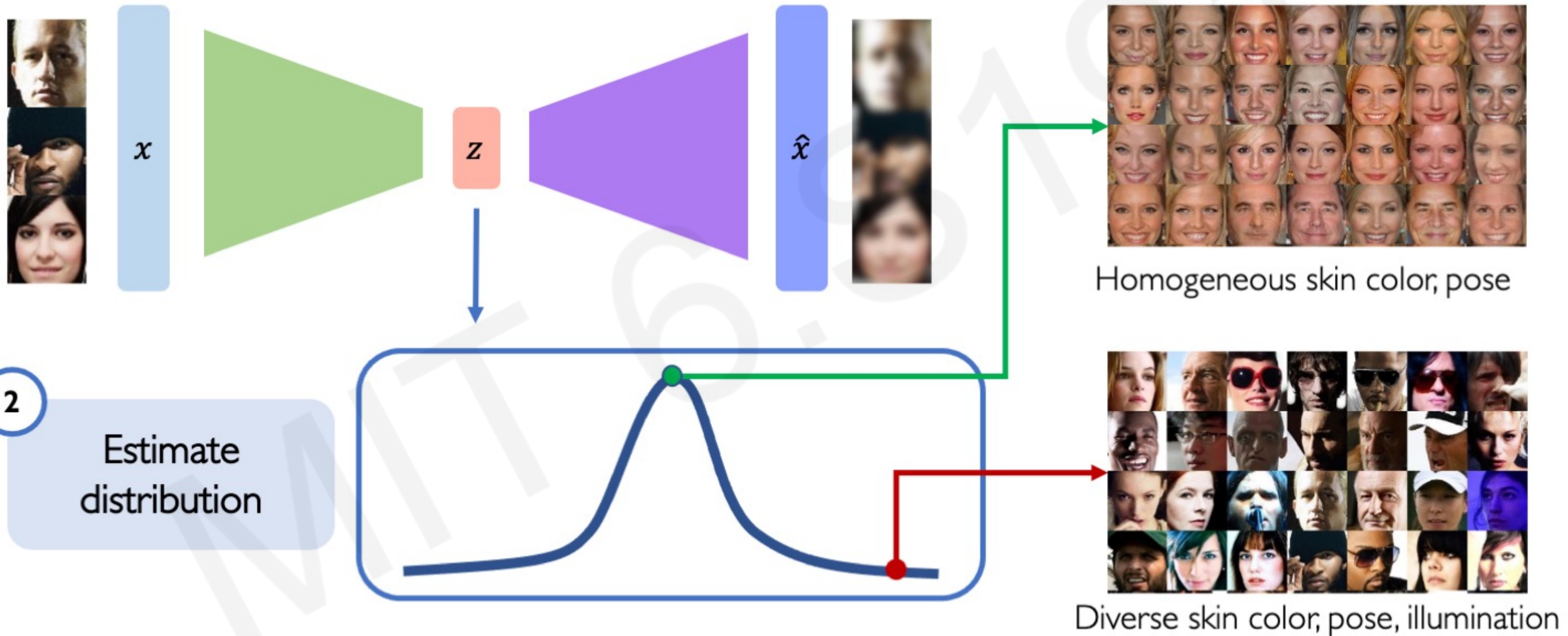
Diverse skin color, pose, illumination

Can we use latent distributions to identify unwanted biases?

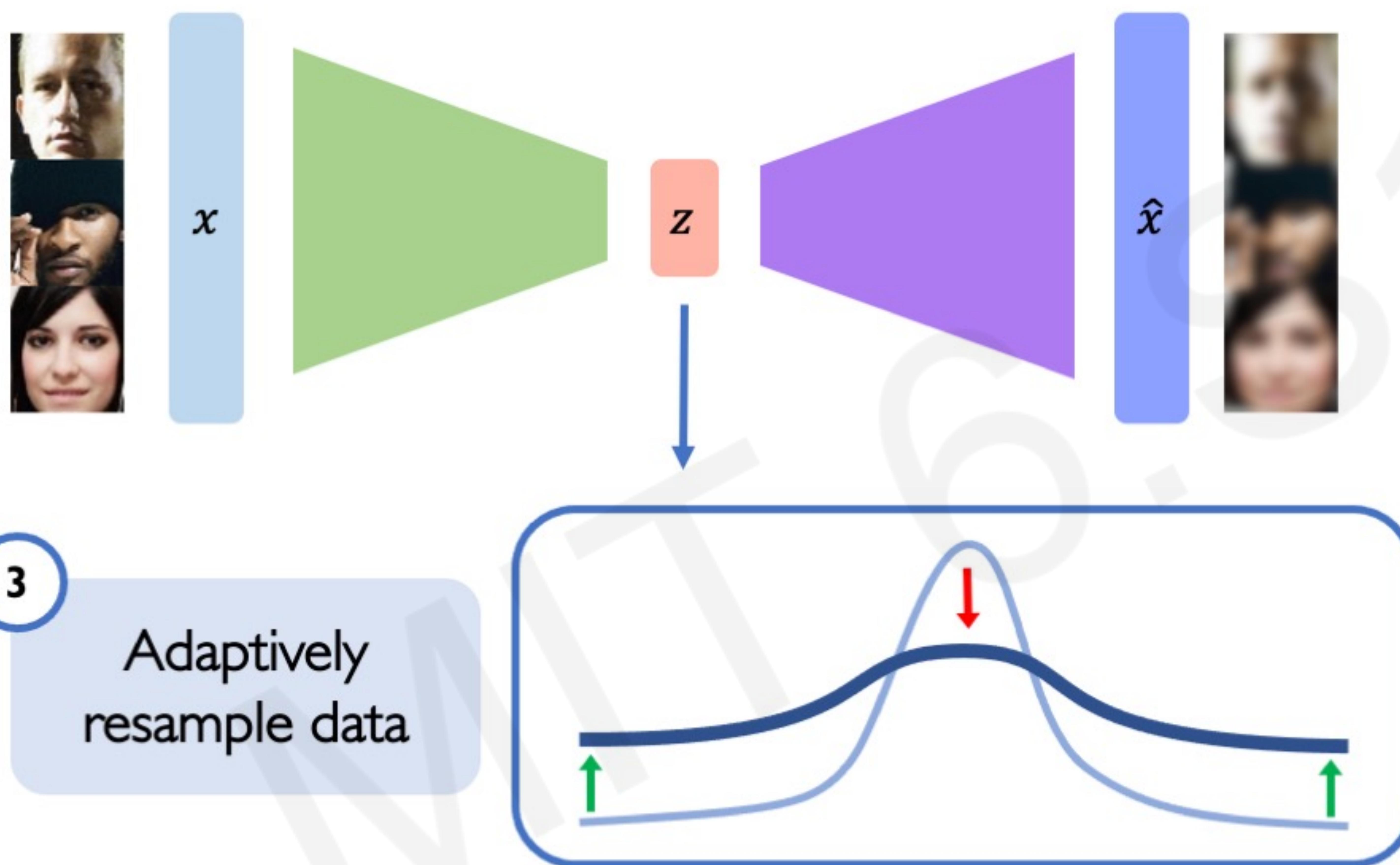
Mitigating Bias through Learned Latent Structure



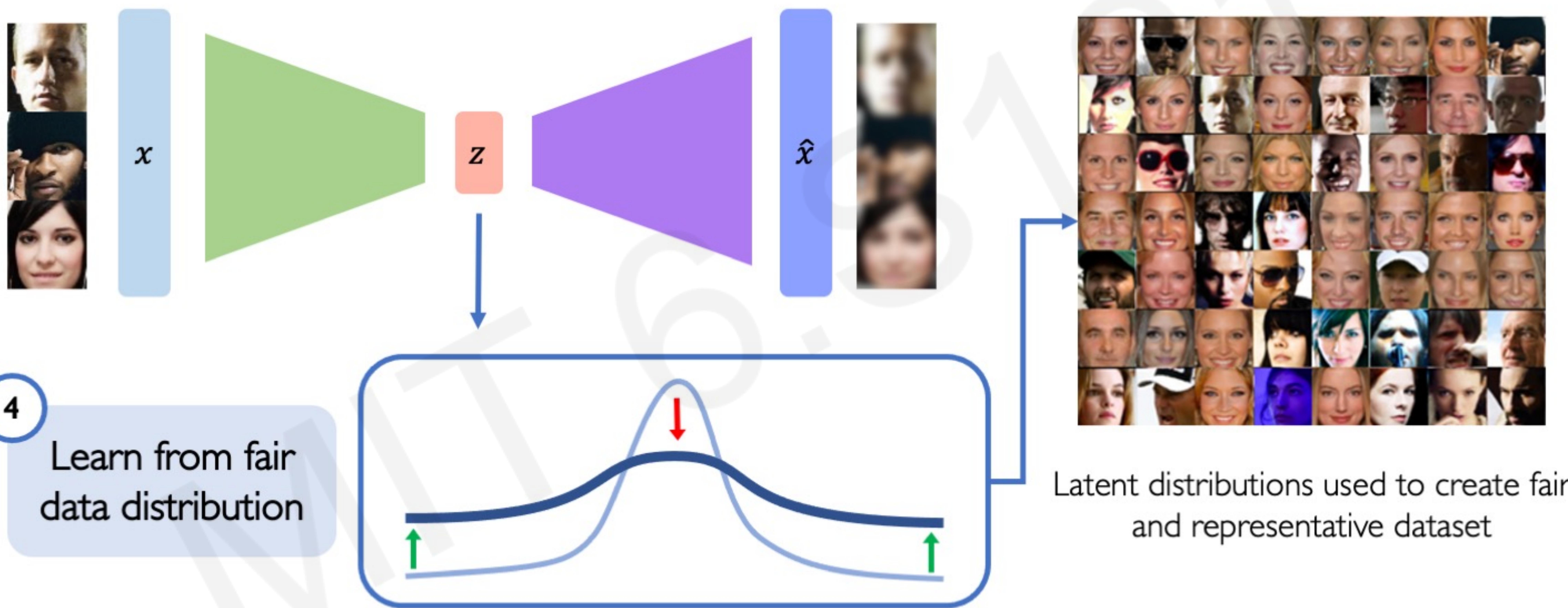
Mitigating Bias through Learned Latent Structure



Mitigating Bias through Learned Latent Structure



Mitigating Bias through Learned Latent Structure



Using Latent Variables for Automated Debiasing

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

$$\hat{Q}(z|X) \propto \prod_i \hat{Q}_i(z_i|X)$$

Estimated joint distribution *i* **Histogram for each latent variable z_i approximate**

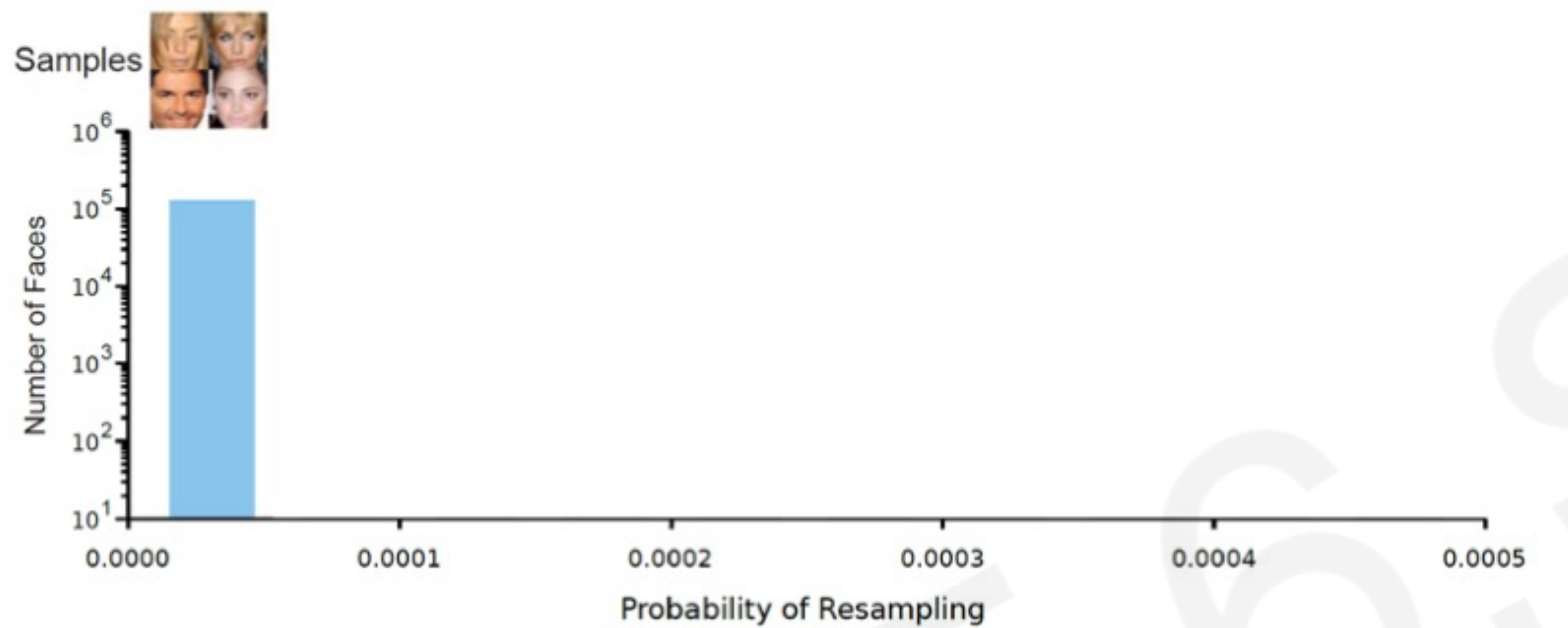
Independence to approximate

Define **adjusted probability** for sampling a particular datapoint x during training:

$$W(z(x)|X) \propto \prod_i \frac{1}{\hat{Q}_i(z_i(x)|X) + \alpha}$$

Probability of selecting datapoint **Histogram for each latent variable z_i** **Debiasing parameter**

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: 7.57×10^{-6}

Batch Sampling During Training with Learned Debiasing

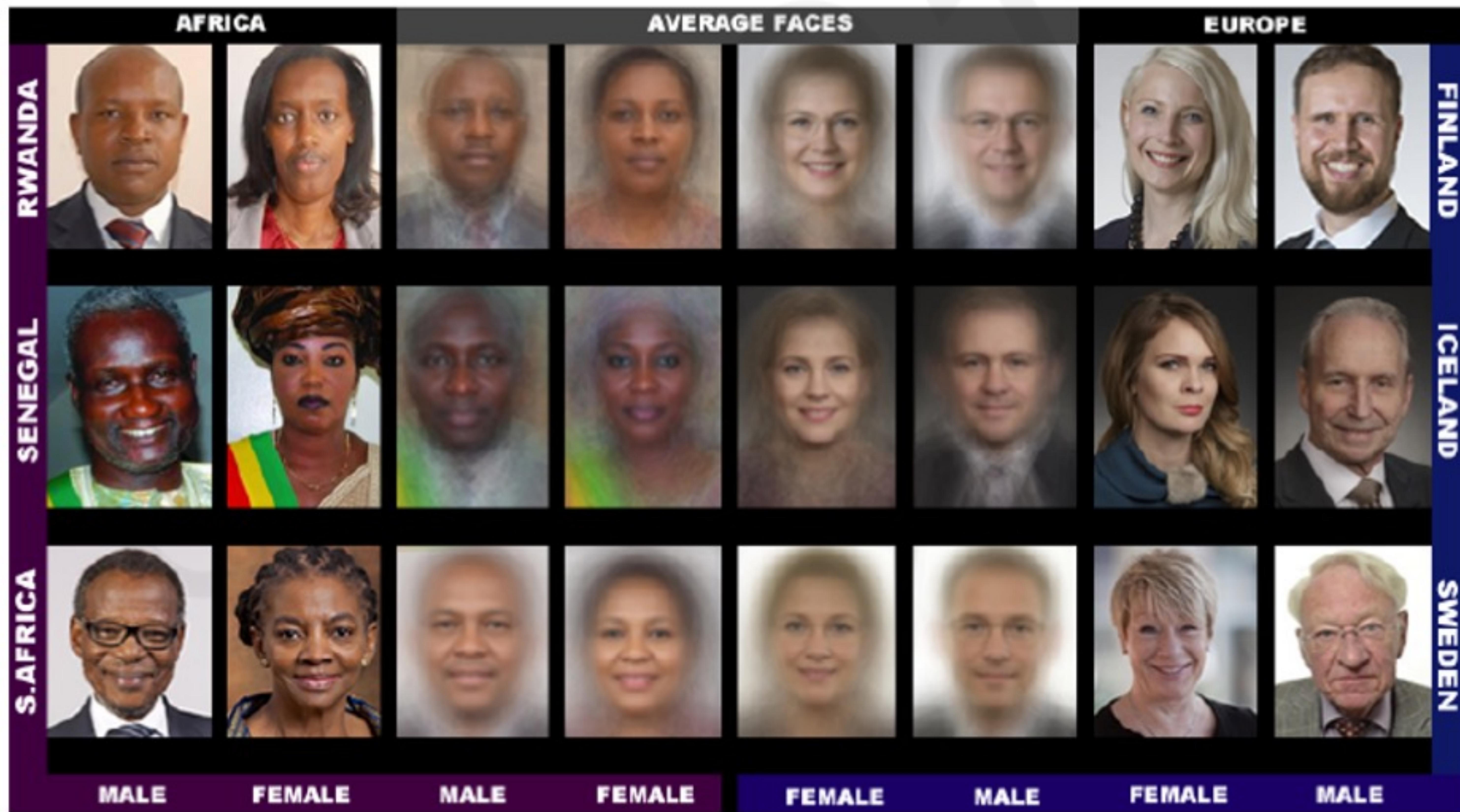


Diverse skin color, pose, illumination
Mean Sample Prob: 1.03×10^{-4}

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

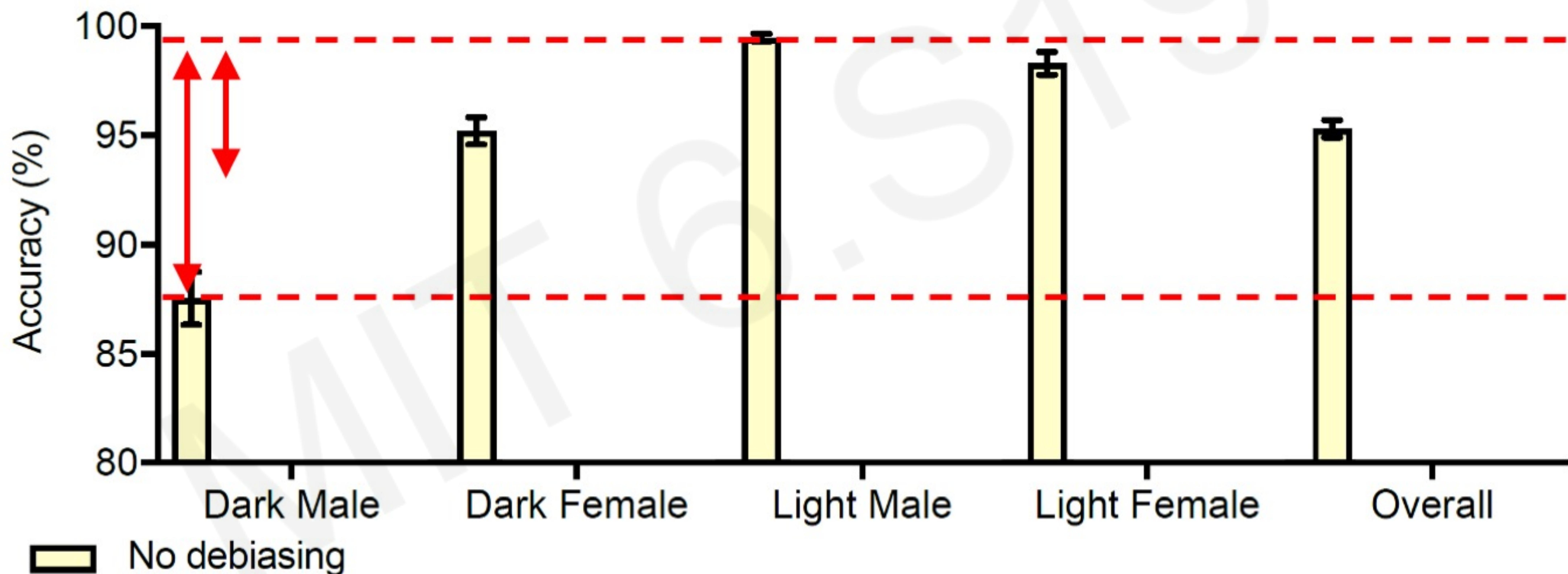
Balanced Dataset for Evaluation

- Pilot Parliaments Benchmark (PPB) dataset
- Evaluation of facial detection algorithms
- Skin tone, male/female binary



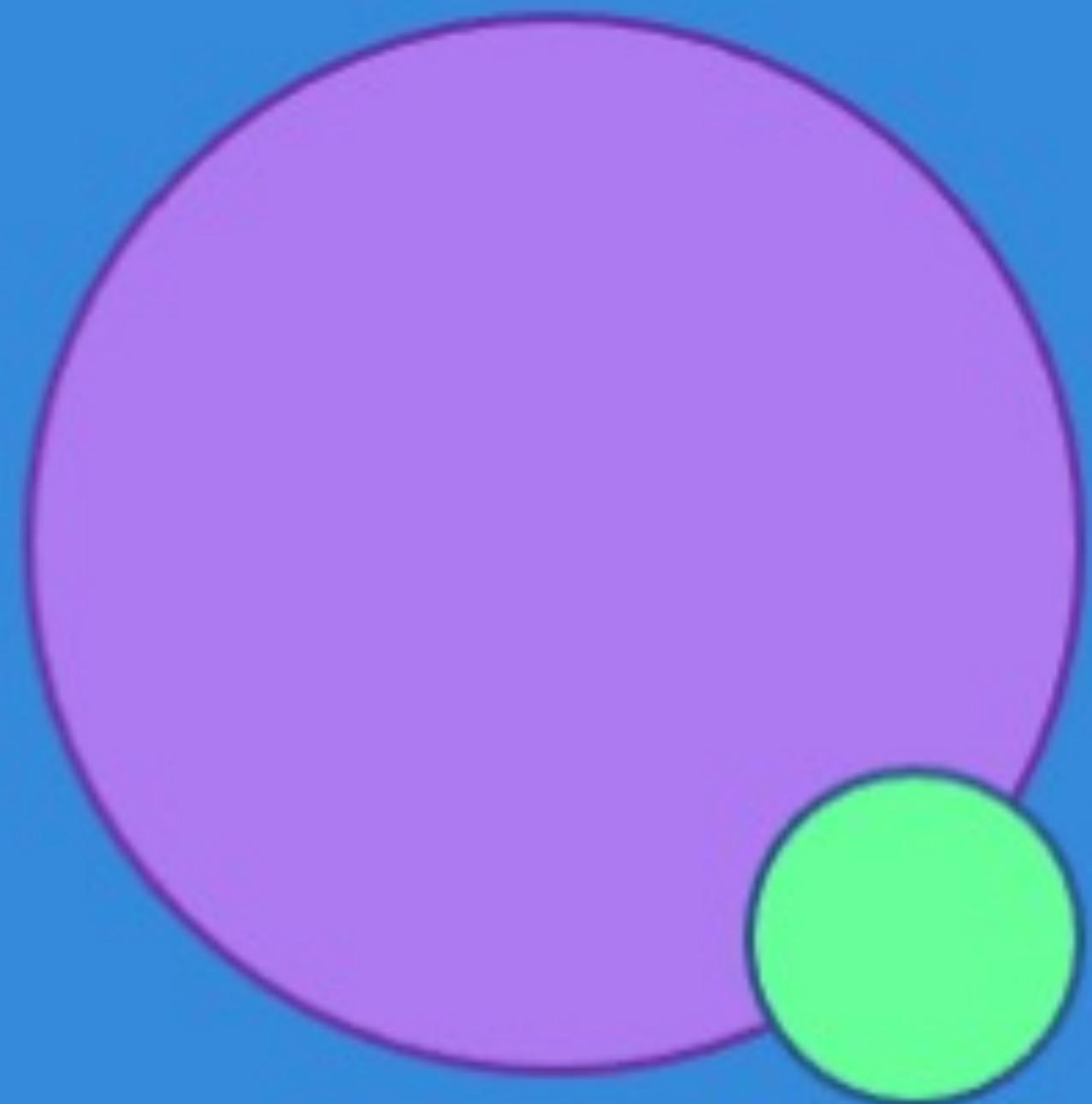
Evaluation: Decreased Categorical Bias

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



Understanding and Mitigating Algorithmic Bias

Types and Sources of Bias



Strategies to Mitigate Bias



AI Fairness: Summary and Future Considerations

AI Best Practices



Dataset
Documentation
Gebru+ *arXiv* 2018.



Model Reporting
and Curation
Mitchell+ *FAT** 2019.



Reproducibility
and Transparency

Algorithmic Solutions

Methods advances to
detect and mitigate biases
during learning



Adversarial Learning
Zhang+ *AAAI/AIES* 2019.



Learned Latent
Structure
Amini/Soleimany+
AAAI/AIES 2019.

Data and Evaluations



Sourcing and
Representation
DeVries+ *CVPR* 2018.



Data with
Distribution Shifts
Koh/Sagawa+ *arXiv* 2020.



Fairness Evaluations
Hardt+ *NeurIPS* 2016.

Necessity of collaboration and education of AI researchers, engineers, ethicists, corporations, politicians, end-users, and the general public.