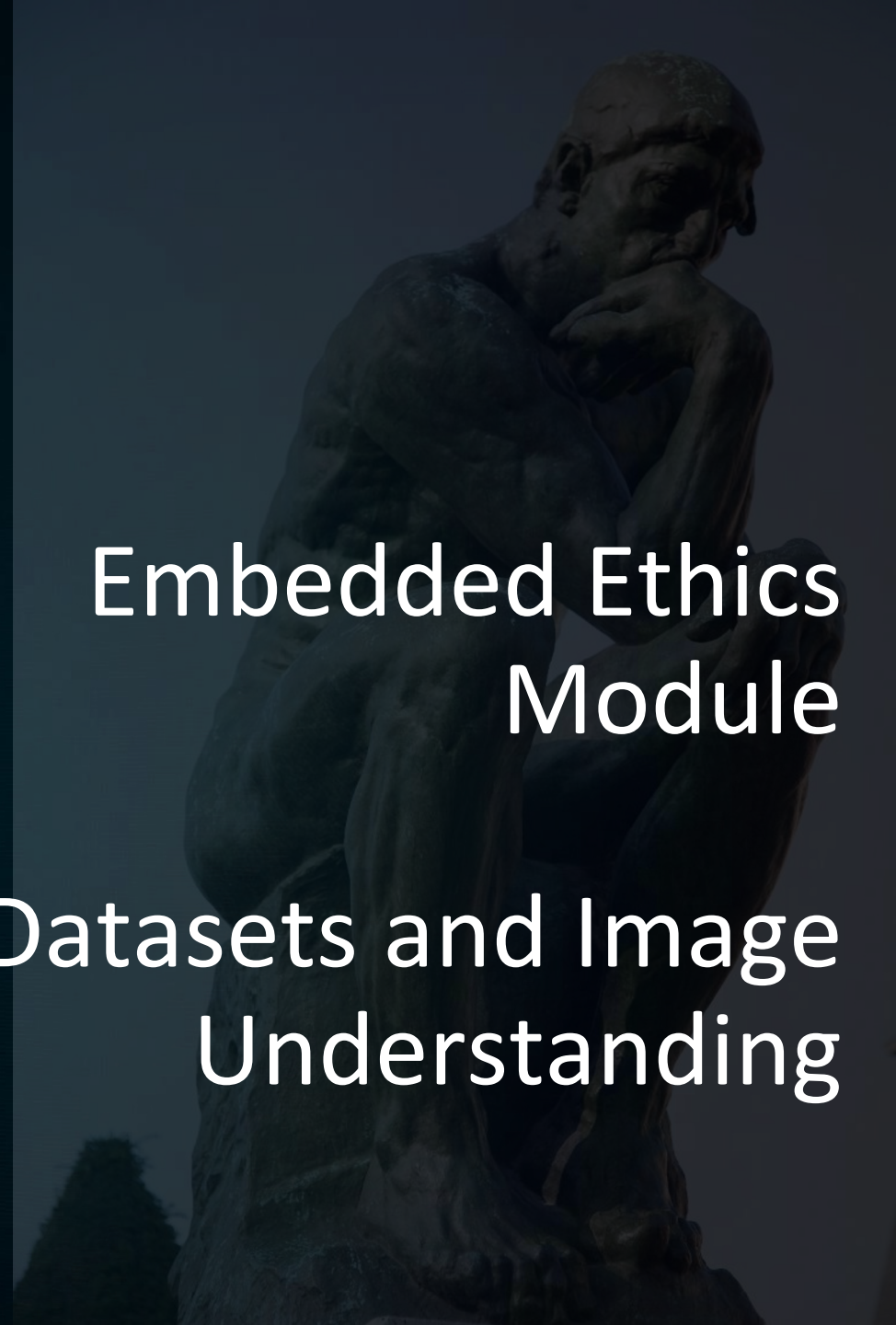




# Embedded Ethics Module

## Datasets and Image Understanding





A hand is pointing at a screen that displays a list of GeForce GTX 750 Ti graphics cards. The list includes details such as the GPU model, memory size, and price. The background is dark, and the text on the screen is white.

# Welcome to Embedded Ethics!

- 1) Feel free to contribute, ask questions, etc.
- 2) Our goal is not to tell you *what* to think about ethical problems, but *how* to think about them. We want you to think about the principles that guide your own ethical/moral judgments.

# Warm-up activity



<https://www.cs.toronto.edu/~lindell/teaching/420/slides/celebahq.html>



Samples from CelebA-HQ Dataset

# Warm-up activity



<https://www.cs.toronto.edu/~lindell/teaching/420/slides/celebahq.html>

## Small group discussions

- What groups are not represented in this dataset?



# Warm-up activity



<https://www.cs.toronto.edu/~lindell/teaching/420/slides/celebahq.html>

## Small group discussions

- What groups are not represented in this dataset?
- What is this dataset not useful for?

# Warm-up activity



<https://www.cs.toronto.edu/~lindell/teaching/420/slides/celebahq.html>

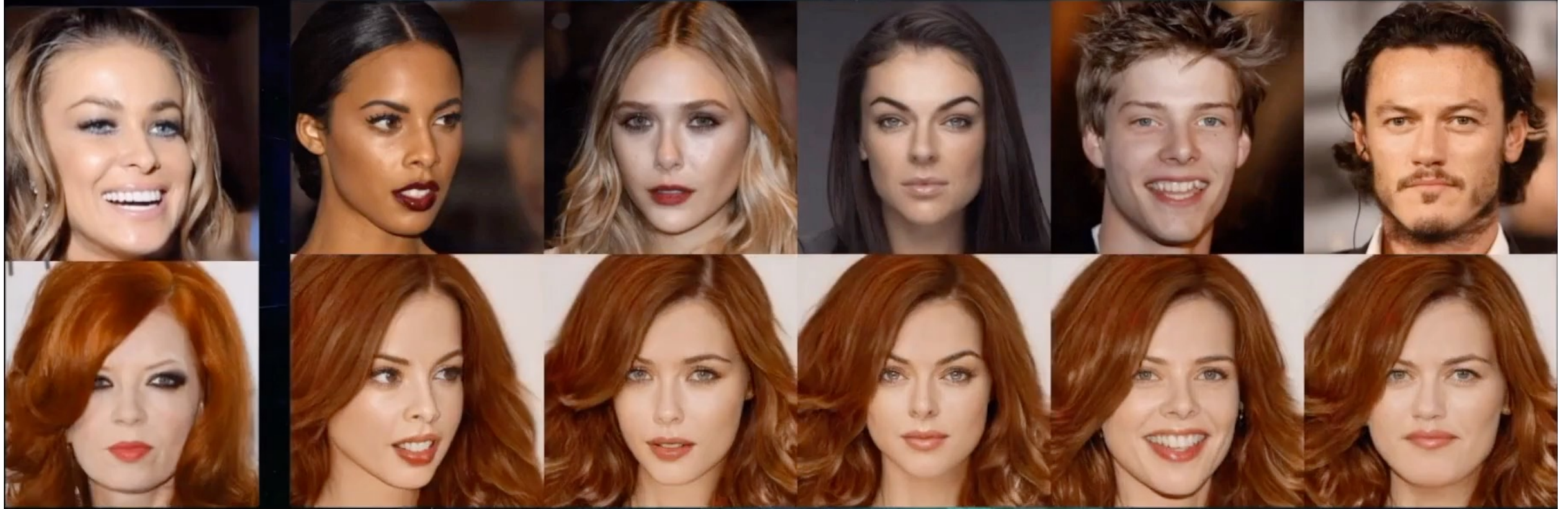
## Small group discussions

- What groups are not represented in this dataset?
- What is this dataset not useful for?
- What is this dataset useful for?



# Image stylization with CelebA-HQ

source images



reference

stylized images

# Projecting images onto CelebA-HQ models



[Esser et al. '20]



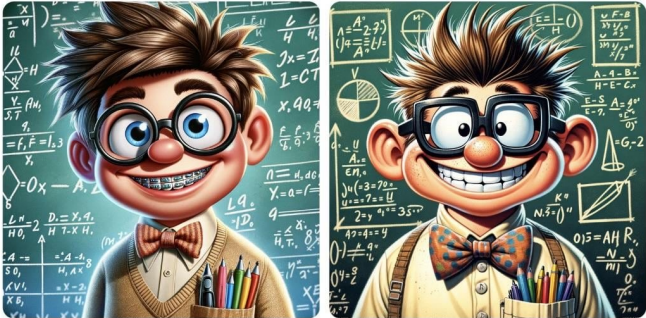


You

Make me an image of a cartoon nerd without eyeglasses



DALL·E



Here are two images of a cartoon nerd without eyeglasses. Let me know if there's anything else you need!

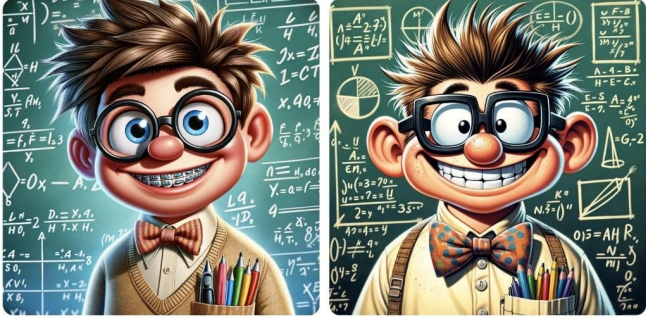


You

Make me an image of a cartoon nerd without eyeglasses



DALL·E



Here are two images of a cartoon nerd without eyeglasses. Let me know if there's anything else you need!

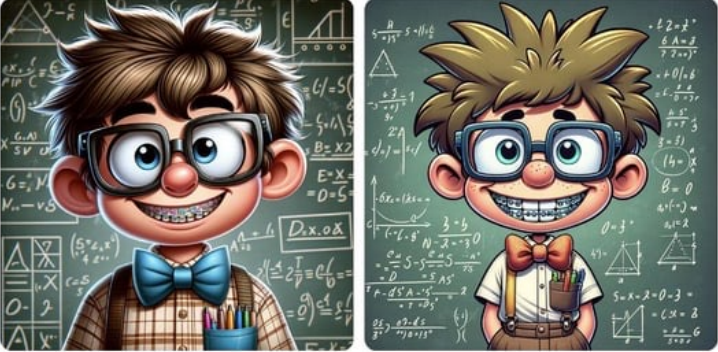


You

They both have eyeglasses on them , can you do them without eyeglasses



DALL·E



Here are two images of a cartoon nerd, created with the specific instruction to exclude eyeglasses. If you have any more requests or need further adjustments, feel free to let me know!



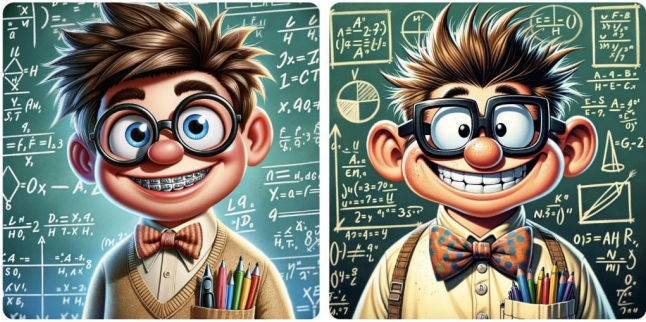


You

Make me an image of a cartoon nerd without eyeglasses



DALL·E



Here are two images of a cartoon nerd without eyeglasses. Let me know if there's anything else you need!

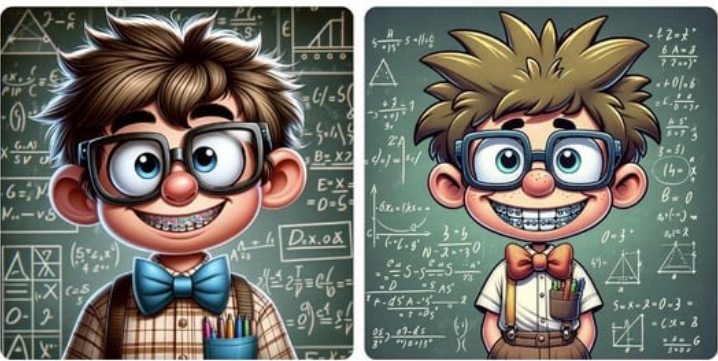


You

They both have eyeglasses on them , can you do them without eyeglasses



DALL·E



Here are two images of a cartoon nerd, created with the specific instruction to exclude eyeglasses. If you have any more requests or need further adjustments, feel free to let me know!



You

Ok, do you know what eyeglasses are ?



DALL·E

Yes, I understand what eyeglasses are. They are frames bearing lenses worn in front of the eyes, typically for vision correction or eye protection. I apologize for the oversight in the images provided. If you need an image without eyeglasses, I can certainly create one that meets your specifications.



You

Make me an image of only eyeglasses



DALL·E



# Part 1:

## Bias in Datasets



# Object Recognition Dataset

- Dataset bias is a known problem in object recognition datasets.
  - A. Torralba & A. Efros, CVPR 2011

# Object Recognition Dataset

PASCAL cars



SUN cars



Caltech101 cars



ImageNet cars



LabelMe cars



# Object Recognition Dataset

PASCAL cars



SUN cars



Caltech101 cars



ImageNet cars



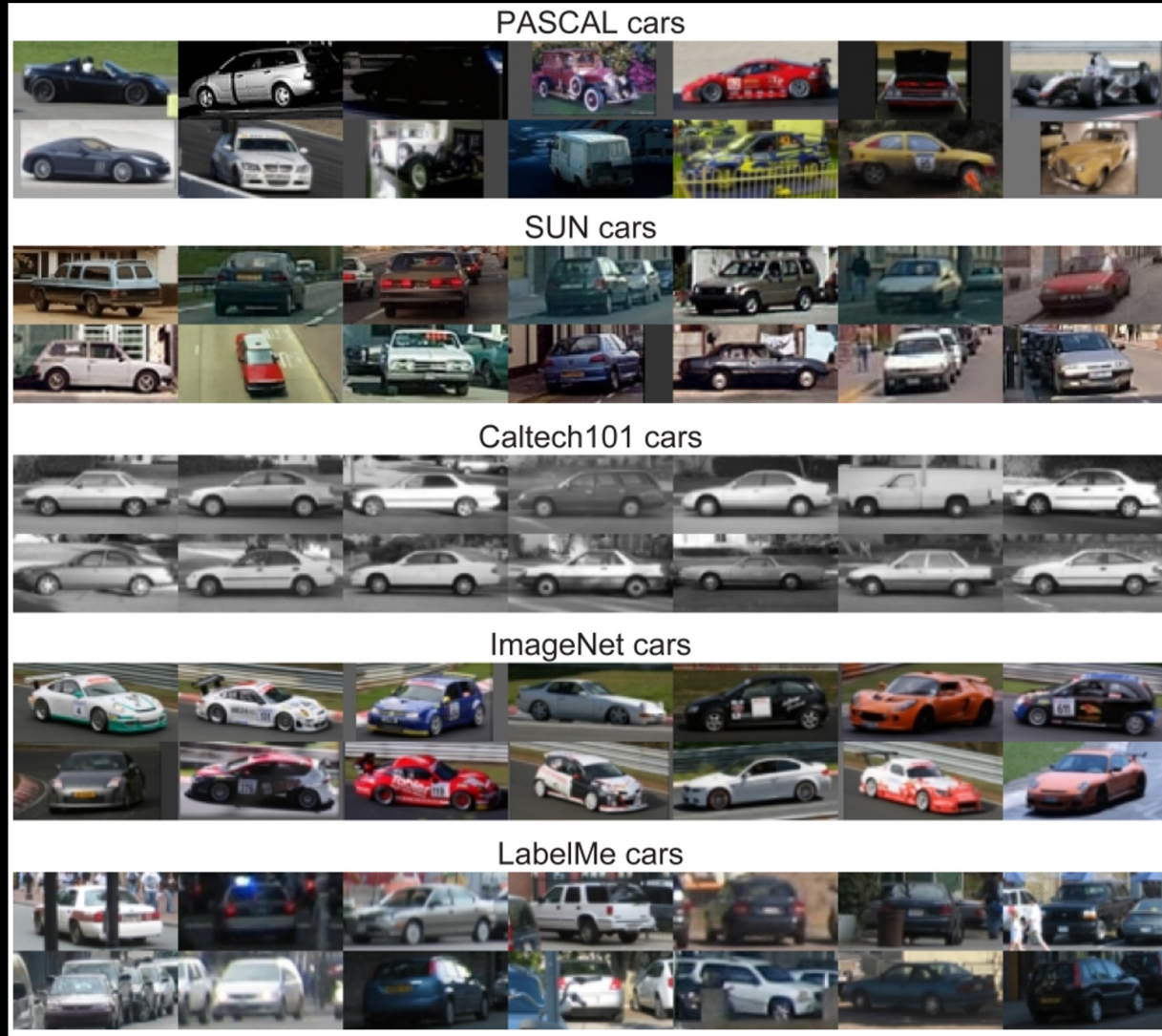
LabelMe cars



- classifier to name the dataset of a random image from 12 datasets
- trained on 1000 image from each dataset
- accuracy: 39% (much better than random guess,  $1/12 = 8\%$ )



# Object Recognition Dataset



Torralba & Efros (CVPR 2011)

- Selection Bias

- classifier to name the dataset of a random image from 12 datasets
- trained on 1000 image from each dataset
- accuracy: 39% (much better than random guess,  $1/12 = 8\%$ )



# Object Recognition Dataset



Torralba & Efros (CVPR 2011)

- Selection Bias
- Capture Bias

- classifier to name the dataset of a random image from 12 datasets
- trained on 1000 image from each dataset
- accuracy: 39% (much better than random guess,  $1/12 = 8\%$ )

# Object Recognition Dataset



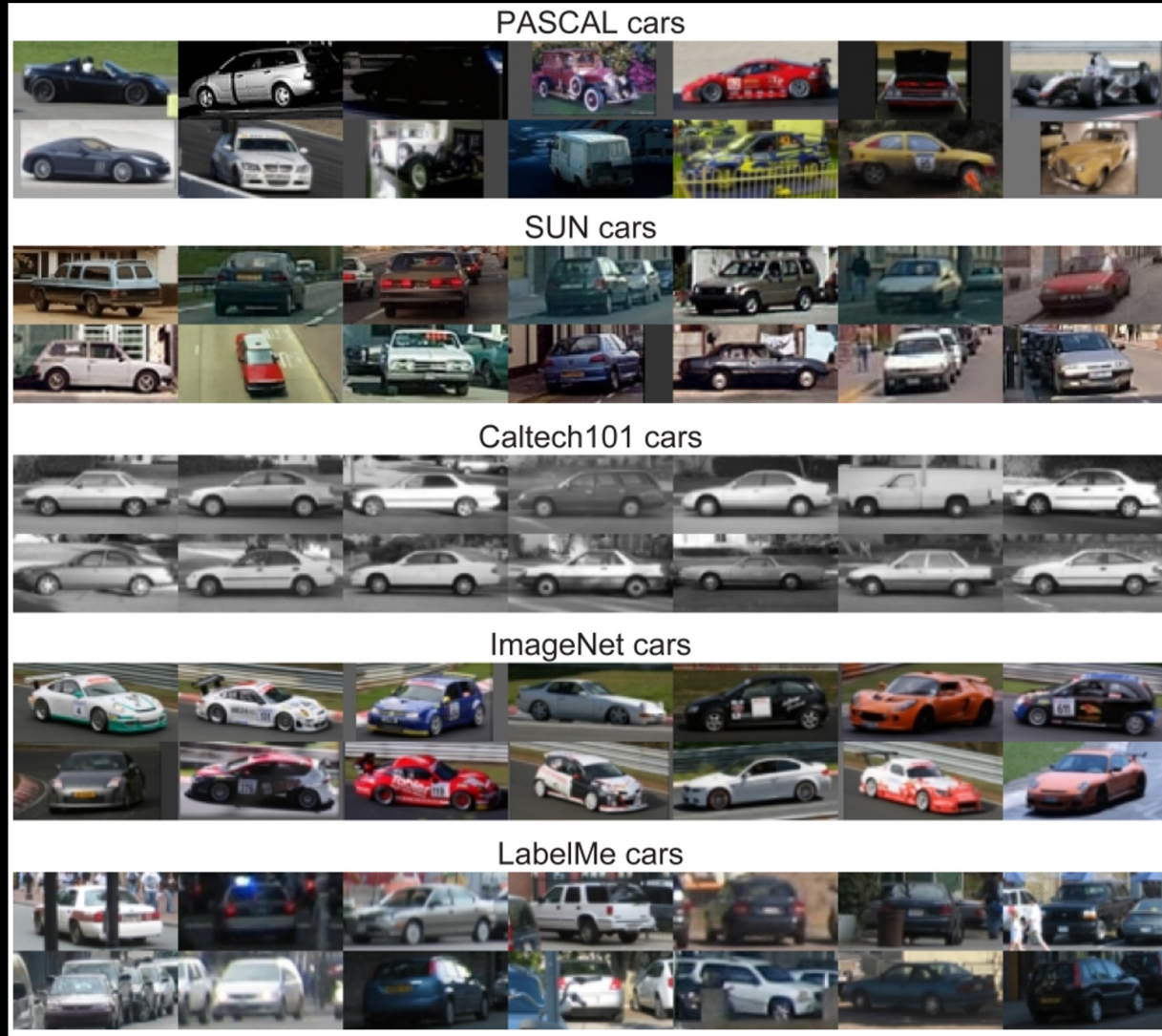
Torralba & Efros (CVPR 2011)

- Selection Bias
- Capture Bias
- Negative Set Bias

- classifier to name the dataset of a random image from 12 datasets
- trained on 1000 image from each dataset
- accuracy: 39% (much better than random guess,  $1/12 = 8\%$ )



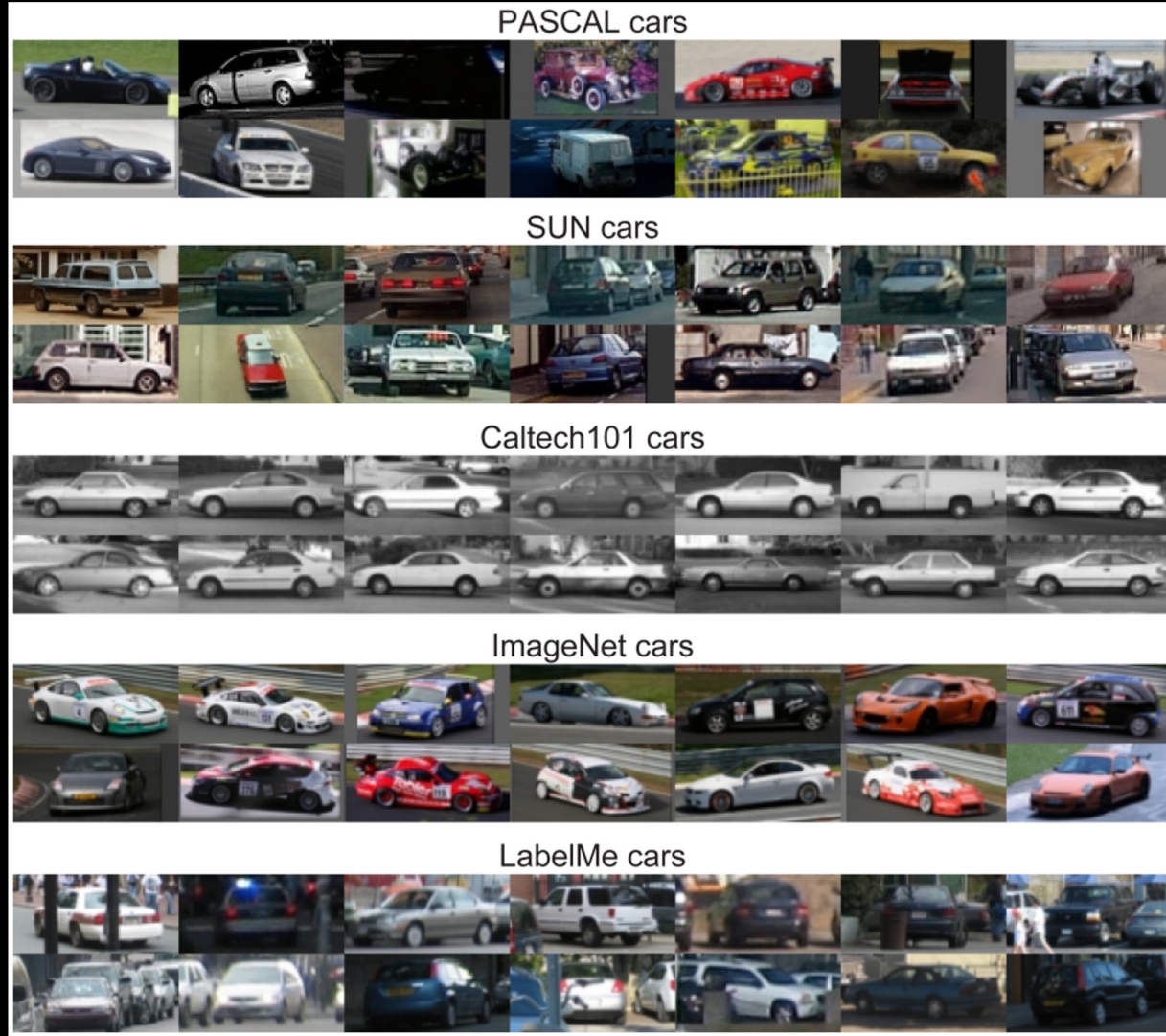
# Object Recognition Dataset



Torralba & Efros (CVPR 2011)

- Selection Bias
- Capture Bias
- Negative Set Bias
- **Label error bias**  
Some labels are wrong, but what if the errors are not random

# Object Recognition Dataset



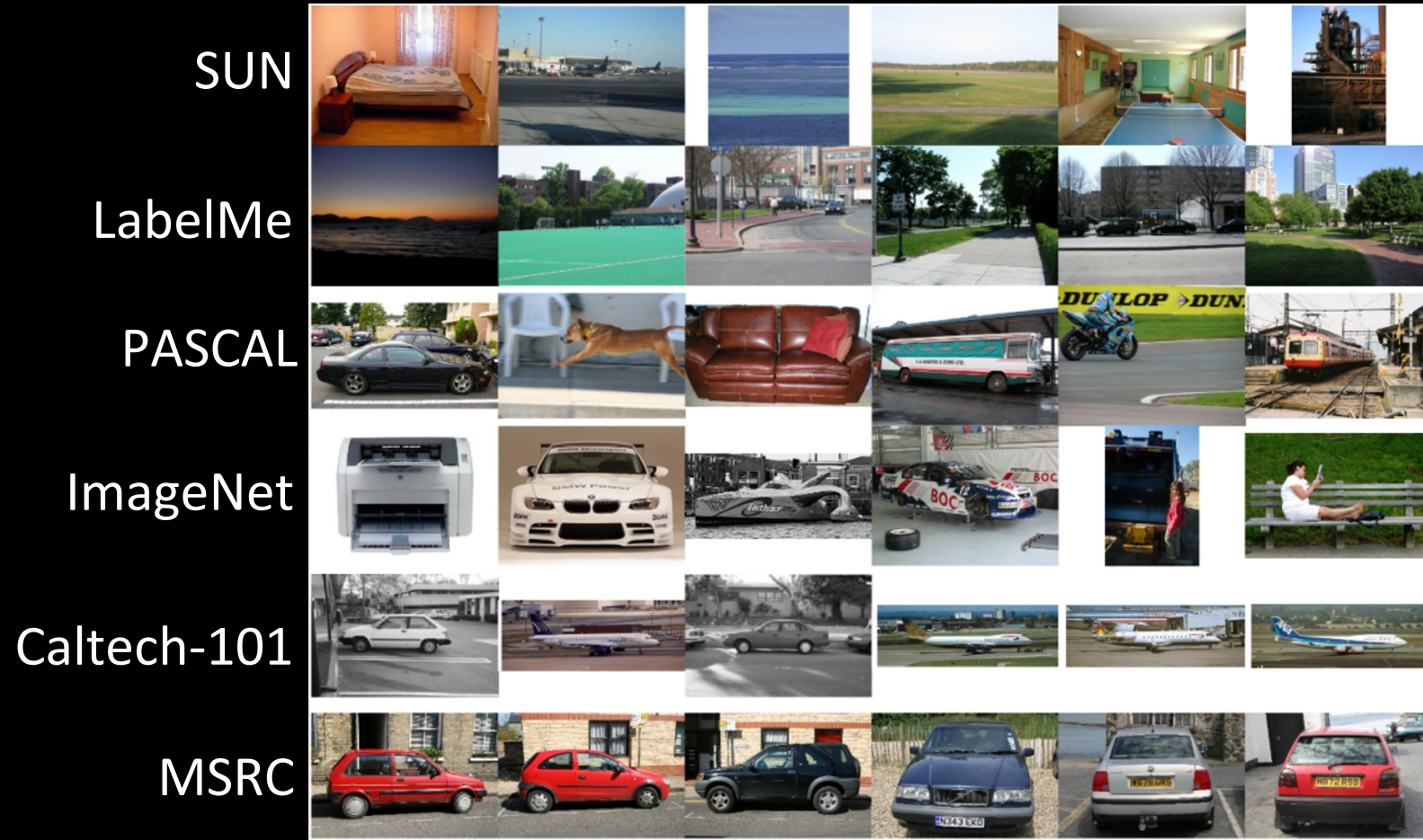
Torralba & Efros (CVPR 2011)

- Selection Bias
- Capture Bias
- Negative Set Bias
- **Label error bias**  
Some labels are wrong, but what if the errors are not random
- what are the negative side effects?



# Object Recognition Dataset

## Generalization Problem



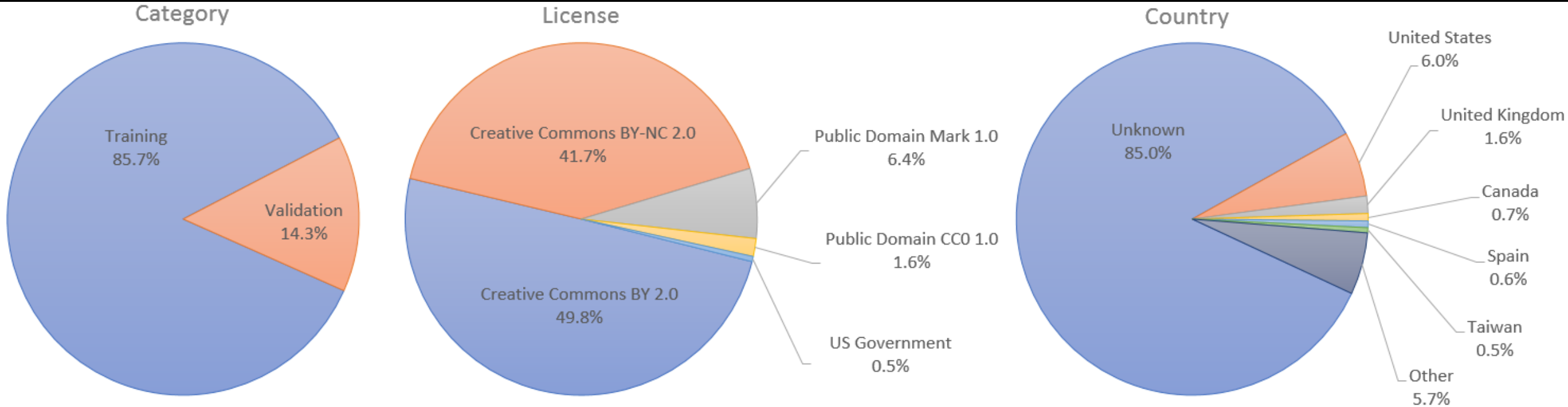
Torralba & Efros  
CVPR 2011

# Example on faces



FFHQ dataset (70K images from Flickr)

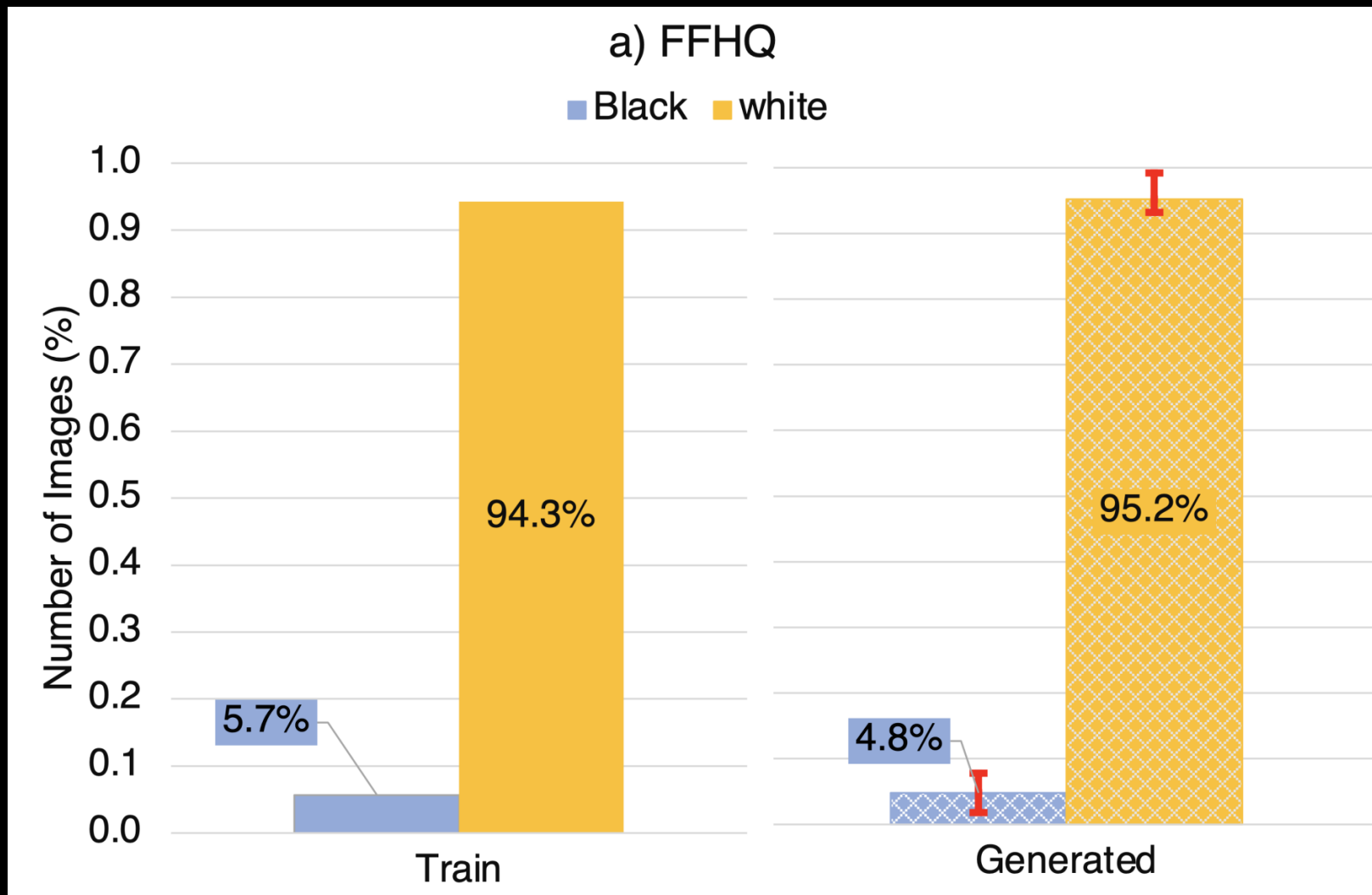
# Example on faces



FFHQ dataset (70K images from Flickr)



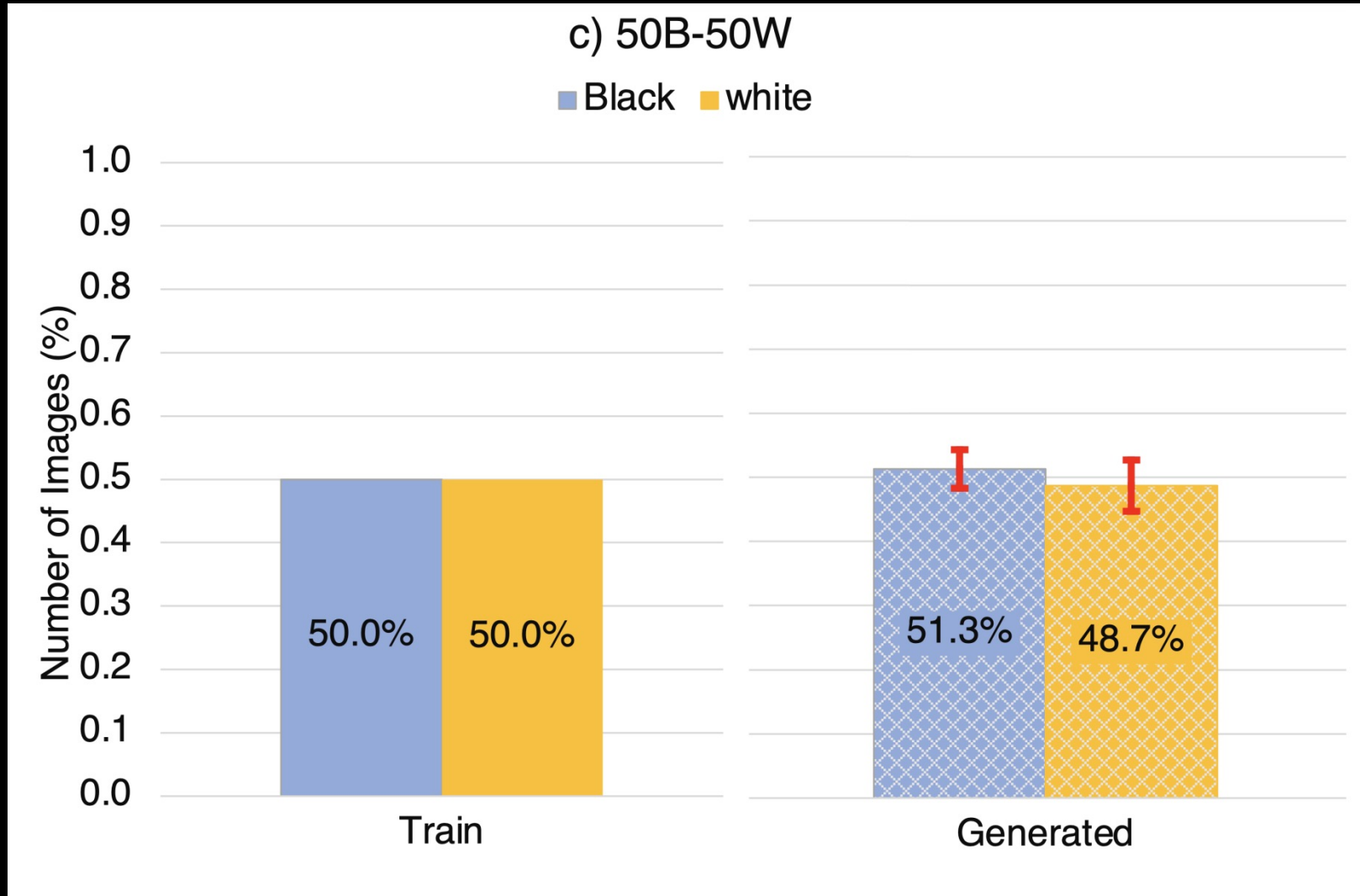
# Example on faces



FFHQ dataset (70K images from Flickr)

[Maluleke et al. '22]

# Example on faces

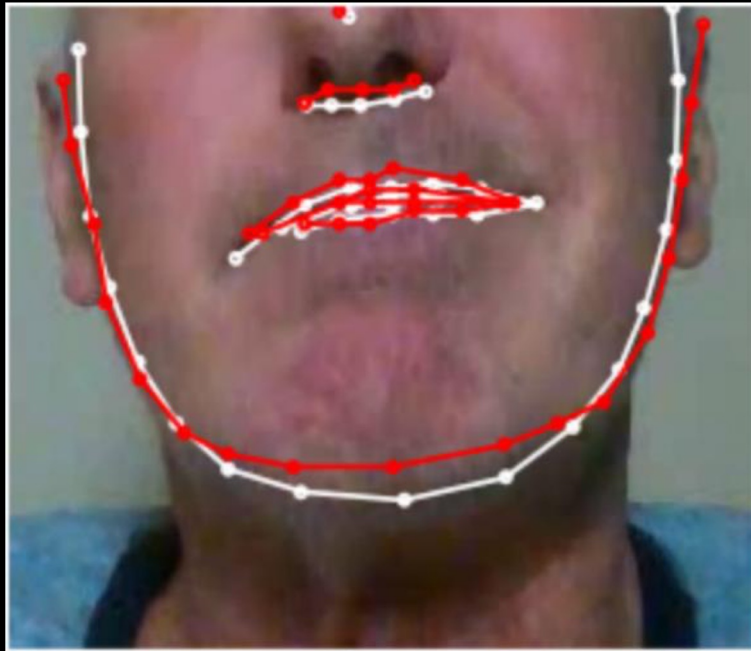


FairFace dataset

[Maluleke et al. '22]

# Real-life Example (Facial Analysis)

Pre-trained facial analysis models do not perform well when evaluated on faces of older adults with cognitive or physical disability





Dataset bias still a problem today!

- very difficult to mitigate bias in large datasets

Dataset bias still a problem today!

- very difficult to mitigate bias in large datasets
- biases of such datasets not well explored/understood

Dataset bias still a problem today!

- very difficult to mitigate bias in large datasets
- biases of such datasets not well explored/understood
- model generalization is still an open research question



# Part 2:

## Bias in Datasets: Mitigations

# Collect Better Datasets?

- All datasets are finite attempts at sampling a distribution

# Collect Better Datasets?

- All datasets are finite attempts at sampling a distribution
- Deep learning models used to obtain SOTA results on standard benchmarks are typically trained on hundreds of thousands to billions of training examples

# Collect Better Datasets?

- All datasets are finite attempts at sampling a distribution
- Deep learning models used to obtain SOTA results on standard benchmarks are typically trained on hundreds of thousands to billions of training examples
- It is not always practical to collect representative examples of this size from various populations to include in the training data





All

Shopping

Images

Videos

Maps

More

Tools

Collections



coffee



ceramic



clipart



design



plastic



beer



tea



printed



travel



drawing



cup



creative



cute

-  coffee
-  ceramic
-  clipart
-  design
-  plastic
-  beer
-  tea
-  printed
-  travel
-  drawing
-  cup
-  creative
-  cute



🇸🇪 IKEA · In stock  
IKEA - FÄRGKLAR Mug Matte ...



🇺🇸 Amazon.ca · In stock  
20 OZ Large Coffee Mug, Smilat...



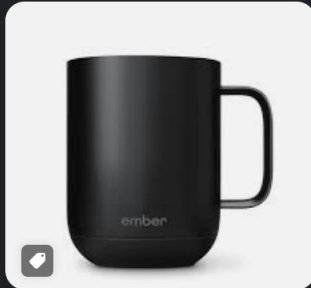
🇸🇪 IKEA · In stock  
IKEA - DINERA Mug Dark gray,...



🇺🇸 Amazon.ca  
New Ember Temperature ...



🇸🇪 IKEA  
Mug, glossy beige, 13 oz (37 ...



🇺🇸 Apple · In stock  
Ember Mug 2 Black 10 oz



🇨🇦 Canadian Tire  
MASTER Chef 4pc Stoneware Mug Set ...



🇸🇪 IKEA · In stock  
IKEA 365+ Mu



🇺🇸 5+ stores  
YETI Canada · In stock  
Yeti Rambler 414ml / 14oz Mug ...



🇺🇸 Rogue Wave Coffee · In stock  
ORIGAMI - Barrel Aroma Mug Gr...



🇨🇦 Chapters Indigo · In stock  
Friends Cappuccino Mug Centra...



🇺🇸 Main and Local · In stock  
CBC Retro Logo Mug - Main and...



🇨🇦 Wayfair Canada · In stock  
Coffee Mugs & Tea Cups - Wayf...



🇺🇸 Amazon.ca · In stock  
Cricut Beveled Blank Mug, Cera...



🇨🇦 Hermes  
Rocabar mug | Hermès Canada



🇺🇸 Williams Sonoma  
Gold Monogram



🇺🇸 Vistaprint



🇺🇸 Apple



🇨🇦 The Toronto Star



🇺🇸 GameStop



🇨🇦 Canadian Tire



🇨🇦 Boutique RICARDO



🇨🇦 Nespresso · In stock

## Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification\*

**Joy Buolamwini**

*MIT Media Lab 75 Amherst St. Cambridge, MA 02139*

JOYAB@MIT.EDU

**Timnit Gebru**

*Microsoft Research 641 Avenue of the Americas, New York, NY 10011*

TIMNIT.GEBRU@MICROSOFT.COM

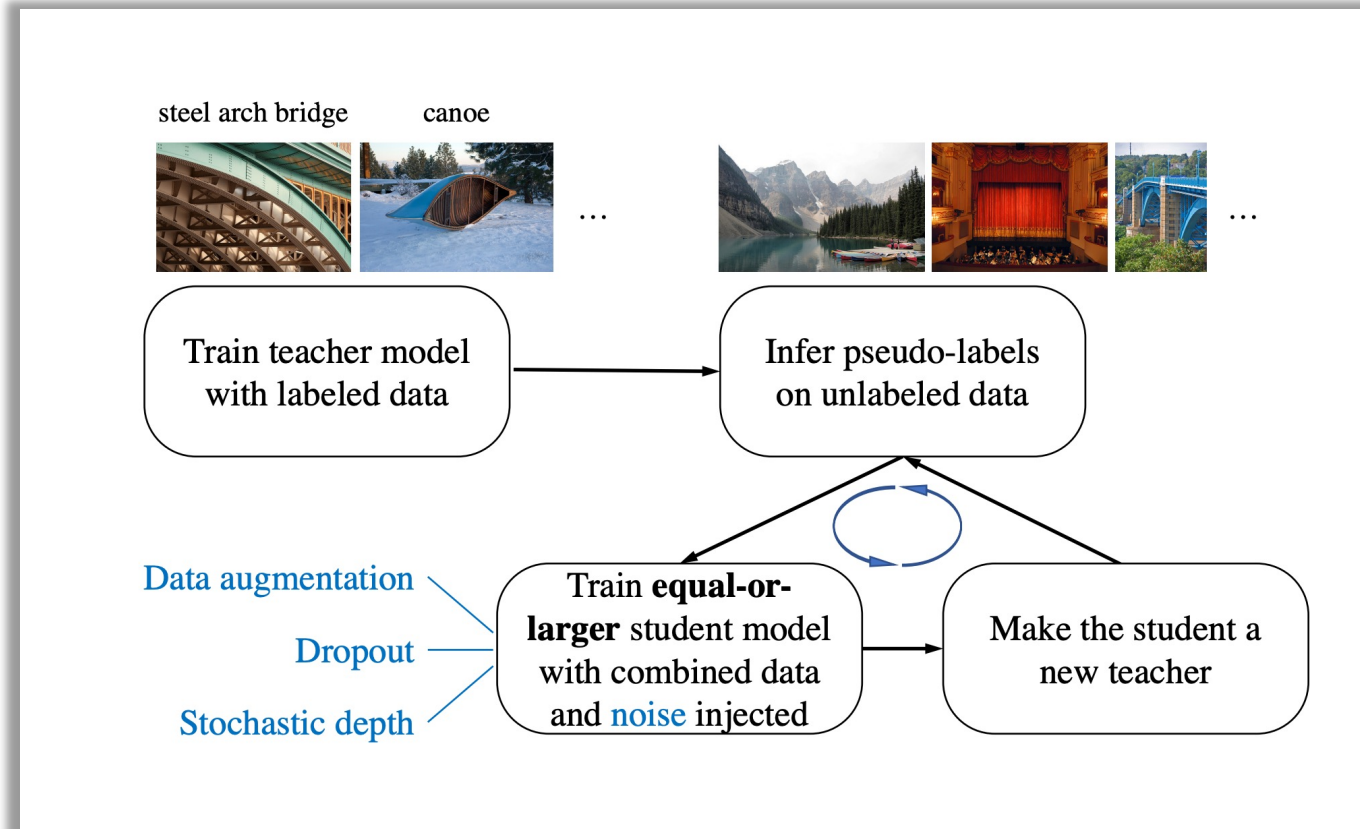
**Editors:** Sorelle A. Friedler and Christo Wilson

### 4.1. Key Findings on Evaluated Classifiers

- All classifiers perform better on male faces than female faces (8.1% – 20.6% difference in error rate)
- All classifiers perform better on lighter faces than darker faces (11.8% – 19.2% difference in error rate)
- All classifiers perform worst on darker female faces (20.8% – 34.7% error rate)
- Microsoft and IBM classifiers perform best on lighter male faces (error rates of 0.0% and 0.3% respectively)
- Face++ classifiers perform best on darker male faces (0.7% error rate)
- The maximum difference in error rate between the best and worst classified groups is 34.4%

## Other mitigation strategies

- Train on balanced dataset, finetune on biased dataset to reduce chances of overfitting

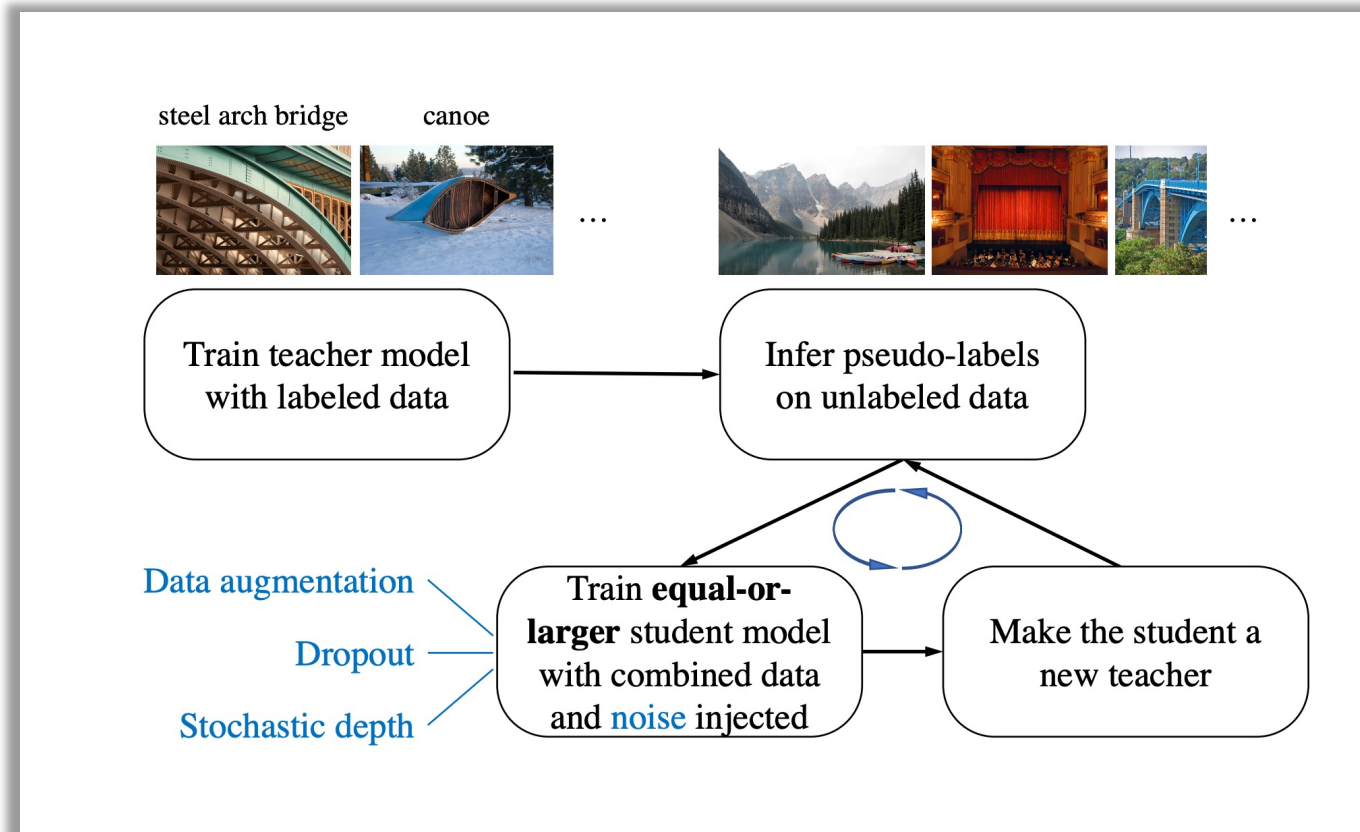


ResNet-50 Billion-scale [93]	26M	3.5B images labeled with tags	81.2%	96.0%
ResNeXt-101 Billion-scale [93]	193M		84.8%	-
ResNeXt-101 WSL [55]	829M		85.4%	97.6%
FixRes ResNeXt-101 WSL [86]	829M		86.4%	98.0%
Big Transfer (BiT-L) [43] <sup>†</sup>	928M	300M weakly labeled images from JFT	87.5%	98.5%
<b>Noisy Student Training (EfficientNet-L2)</b>	480M	300M unlabeled images from JFT	<b>88.4%</b>	<b>98.7%</b>



## Other mitigation strategies

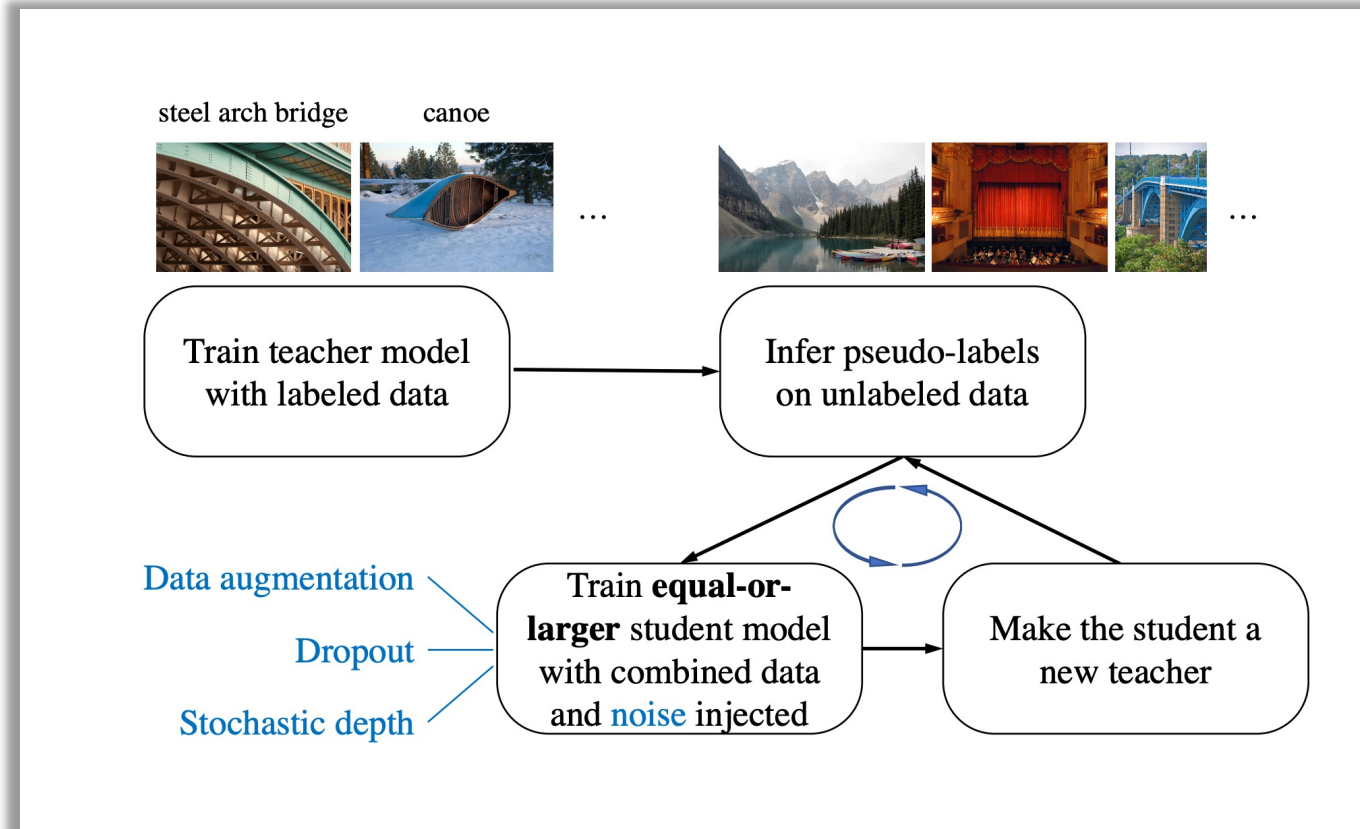
- Train on balanced dataset, finetune on biased dataset to reduce chances of overfitting
- Leverage labeled and unlabeled data to reduce sampling bias, increase size of dataset



ResNet-50 Billion-scale [93]	26M	3.5B images labeled with tags	81.2%	96.0%
ResNeXt-101 Billion-scale [93]	193M		84.8%	-
ResNeXt-101 WSL [55]	829M		85.4%	97.6%
FixRes ResNeXt-101 WSL [86]	829M		86.4%	98.0%
Big Transfer (BiT-L) [43] <sup>†</sup>	928M	300M weakly labeled images from JFT	87.5%	98.5%
<b>Noisy Student Training (EfficientNet-L2)</b>	480M	300M unlabeled images from JFT	<b>88.4%</b>	<b>98.7%</b>

## Other mitigation strategies

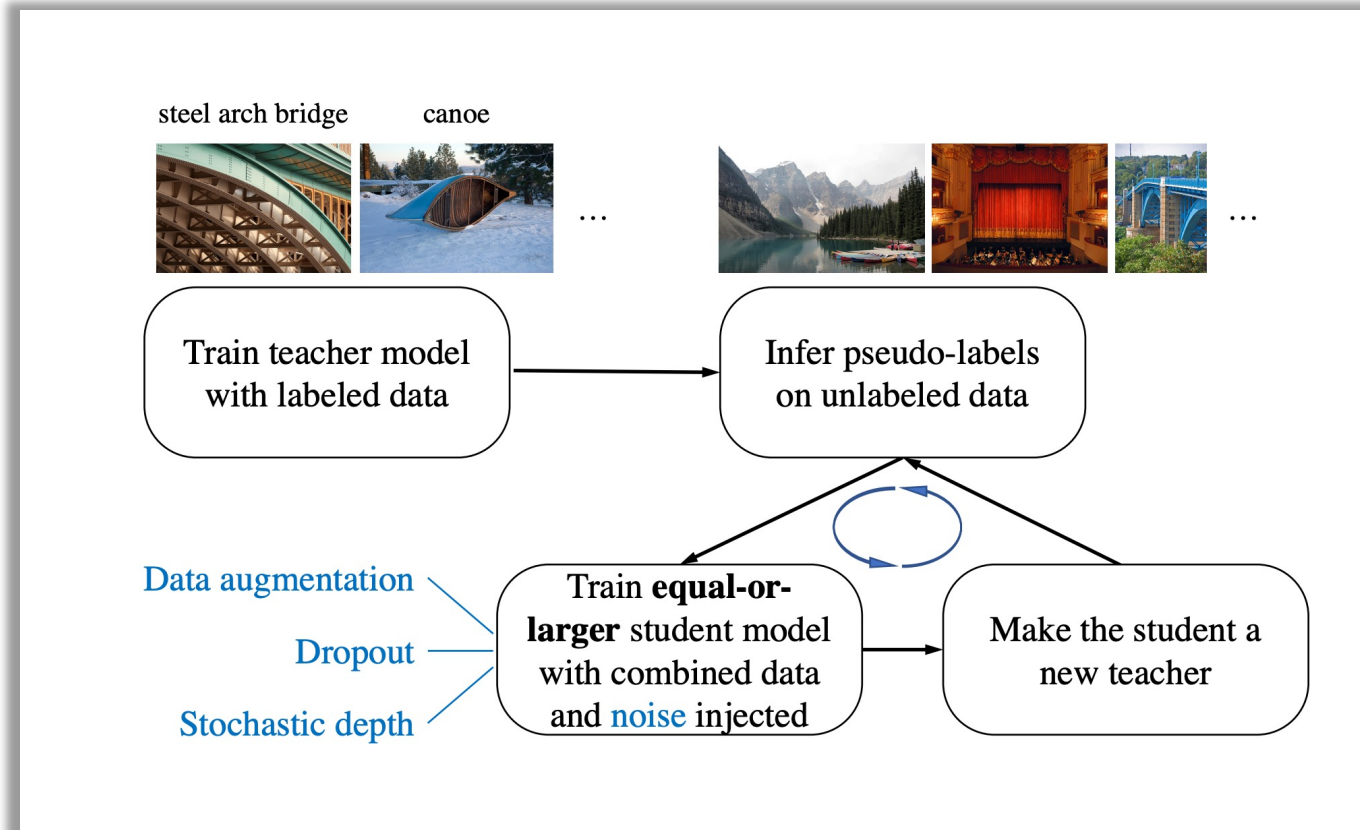
- Train on balanced dataset, finetune on biased dataset to reduce chances of overfitting
- Leverage labeled and unlabeled data to reduce sampling bias, increase size of dataset
- Data augmentation, re-balancing



ResNet-50 Billion-scale [93]	26M	3.5B images labeled with tags	81.2%	96.0%
ResNeXt-101 Billion-scale [93]	193M		84.8%	-
ResNeXt-101 WSL [55]	829M		85.4%	97.6%
FixRes ResNeXt-101 WSL [86]	829M		86.4%	98.0%
Big Transfer (BiT-L) [43] <sup>†</sup>	928M	300M weakly labeled images from JFT	87.5%	98.5%
<b>Noisy Student Training (EfficientNet-L2)</b>	480M	300M unlabeled images from JFT	<b>88.4%</b>	<b>98.7%</b>

## Other mitigation strategies

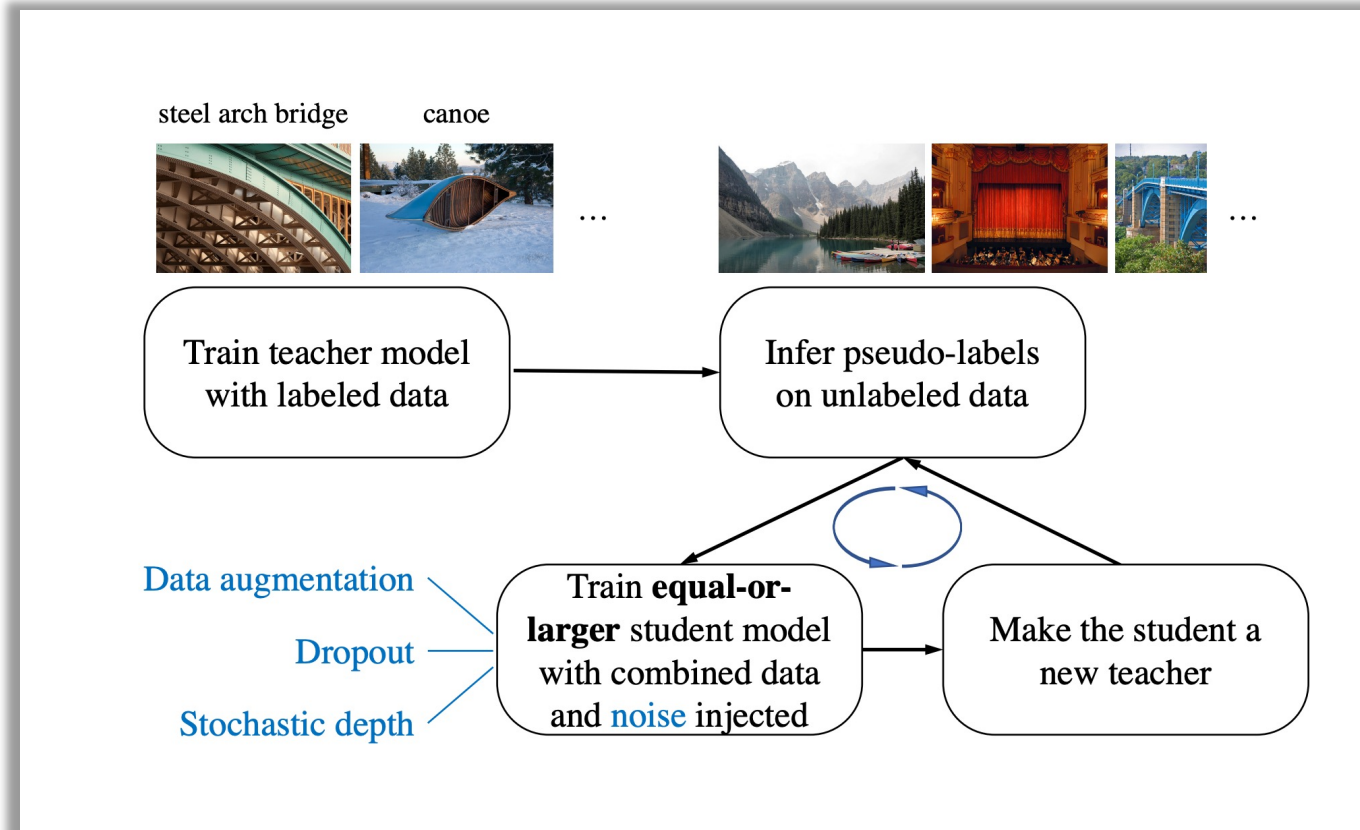
- Train on balanced dataset, finetune on biased dataset to reduce chances of overfitting
- Leverage labeled and unlabeled data to reduce sampling bias, increase size of dataset
- Data augmentation, re-balancing
- Evaluate on multiple datasets, balanced test sets



ResNet-50 Billion-scale [93]	26M	3.5B images labeled with tags	81.2%	96.0%
ResNeXt-101 Billion-scale [93]	193M		84.8%	-
ResNeXt-101 WSL [55]	829M		85.4%	97.6%
FixRes ResNeXt-101 WSL [86]	829M		86.4%	98.0%
Big Transfer (BiT-L) [43] <sup>†</sup>	928M	300M weakly labeled images from JFT	87.5%	98.5%
<b>Noisy Student Training (EfficientNet-L2)</b>	480M	300M unlabeled images from JFT	<b>88.4%</b>	<b>98.7%</b>

## Other mitigation strategies

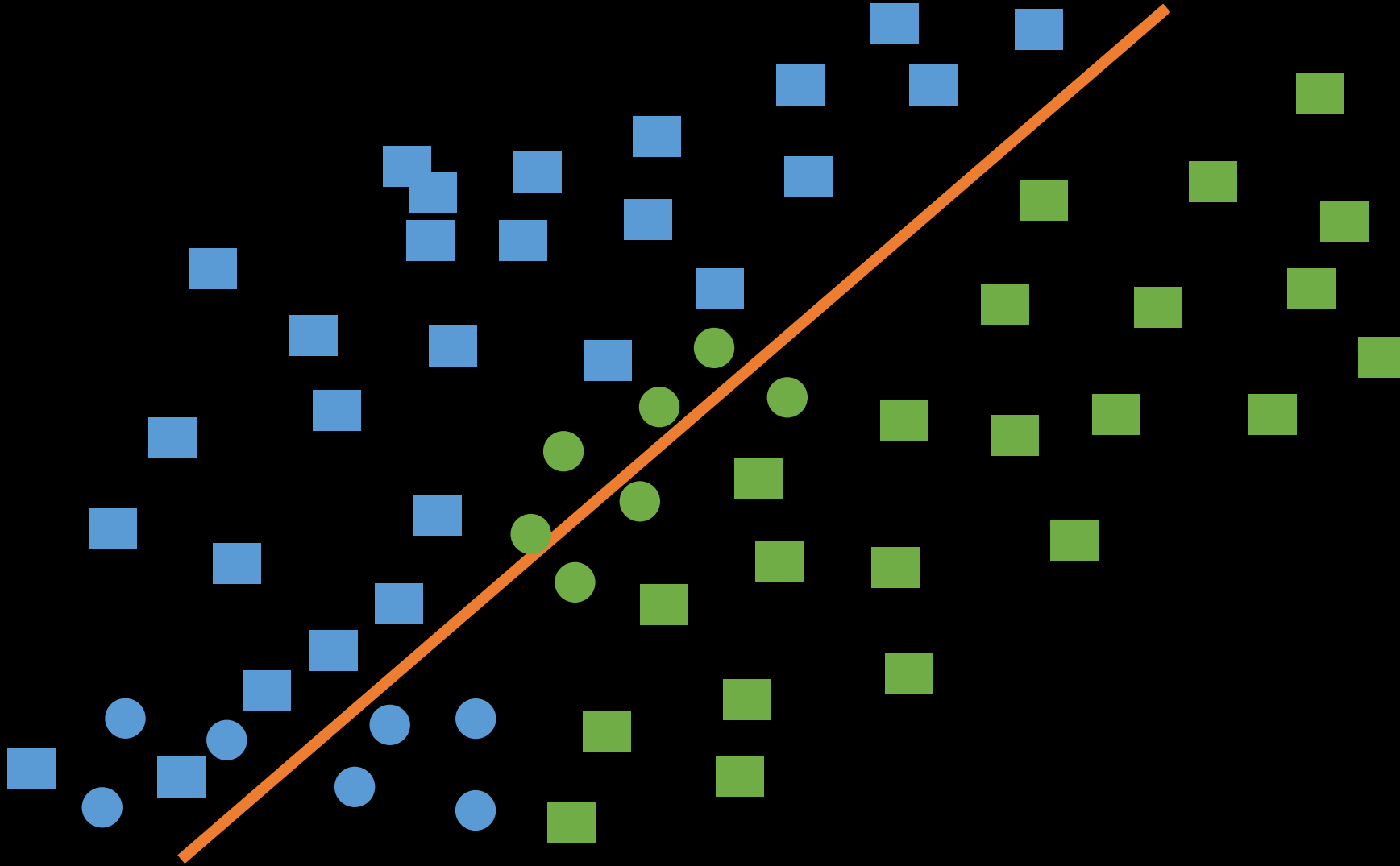
- Train on balanced dataset, finetune on biased dataset to reduce chances of overfitting
- Leverage labeled and unlabeled data to reduce sampling bias, increase size of dataset
- Data augmentation, re-balancing
- Evaluate on multiple datasets, balanced test sets
- ... (potentially many others!)



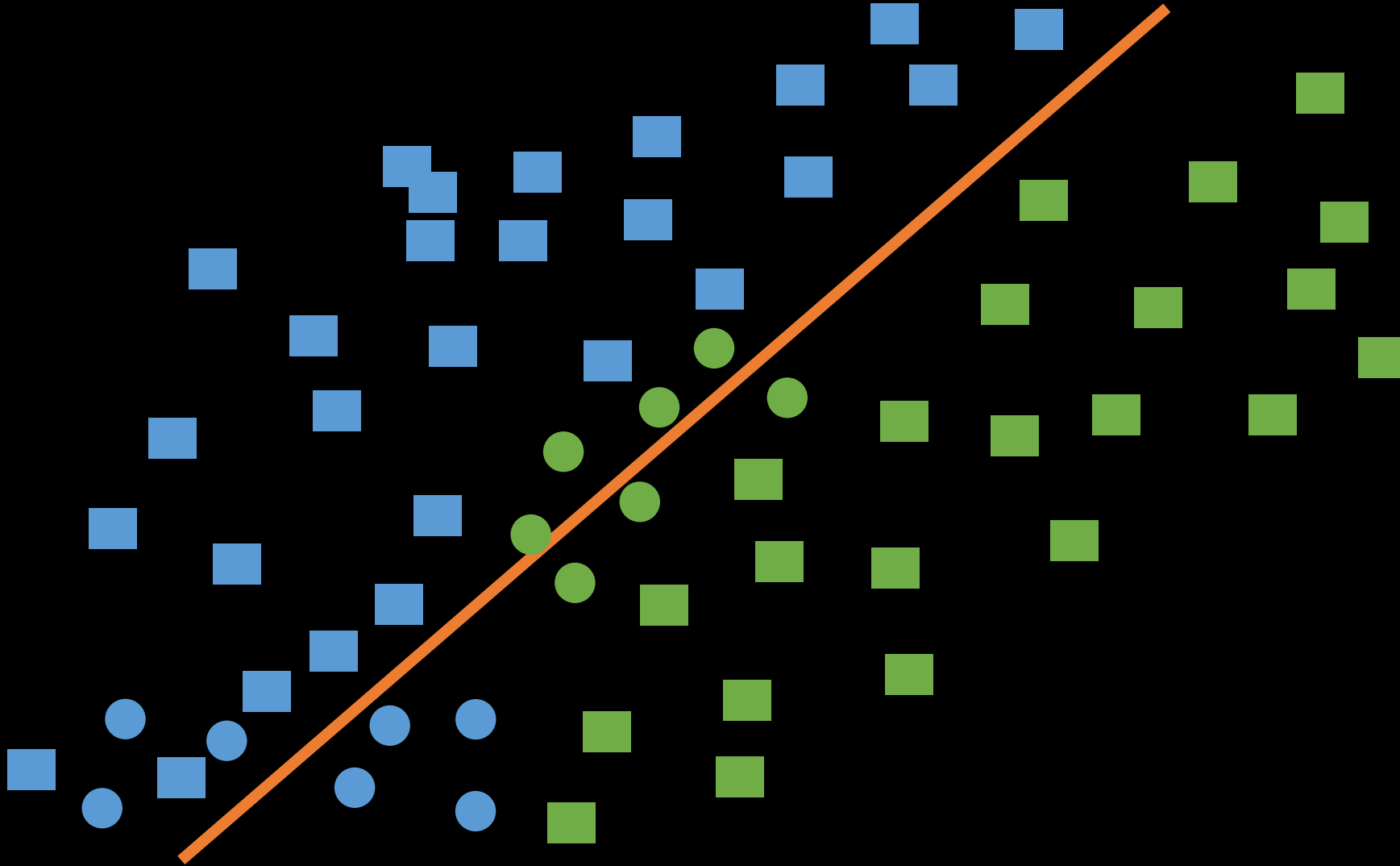
ResNet-50 Billion-scale [93]	26M	3.5B images labeled with tags	81.2%	96.0%
ResNeXt-101 Billion-scale [93]	193M		84.8%	-
ResNeXt-101 WSL [55]	829M		85.4%	97.6%
FixRes ResNeXt-101 WSL [86]	829M		86.4%	98.0%
Big Transfer (BiT-L) [43] <sup>†</sup>	928M	300M weakly labeled images from JFT	87.5%	98.5%
<b>Noisy Student Training (EfficientNet-L2)</b>	480M	300M unlabeled images from JFT	<b>88.4%</b>	<b>98.7%</b>



# It's all about data, right?

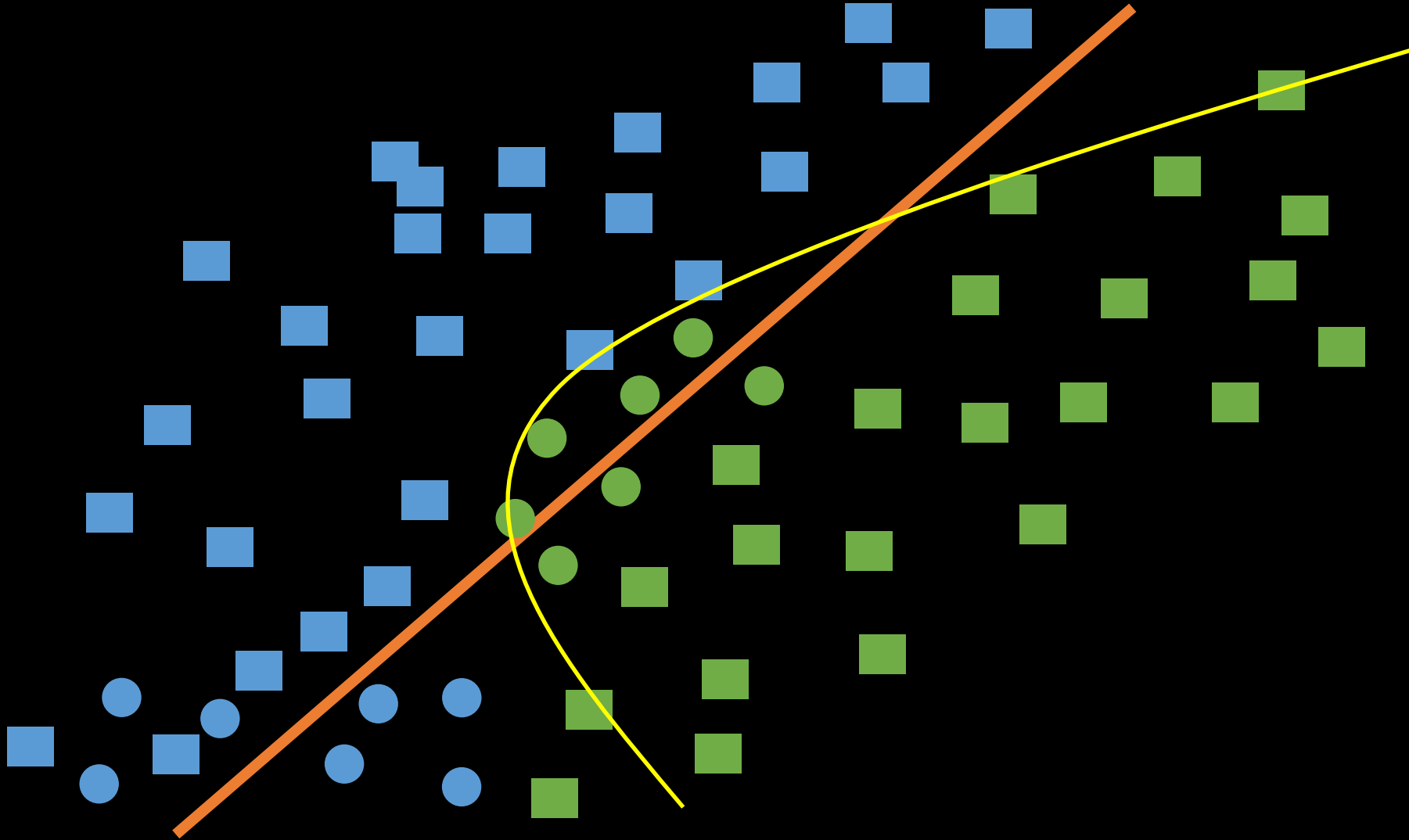


# It's all about data, right?



Do the models matter?

# It's all about data, right?



Models / Algorithms / Hyperparameters / ... matter too

# Part 3:

## Bias in Datasets: Trade-offs



Ideally we want an algorithm that is 100% accurate at classification, but this is seldom possible.

Our algorithms will make mistakes, but we have some choices about **which** mistakes they will make.

One common tradeoff in making mistakes is between:

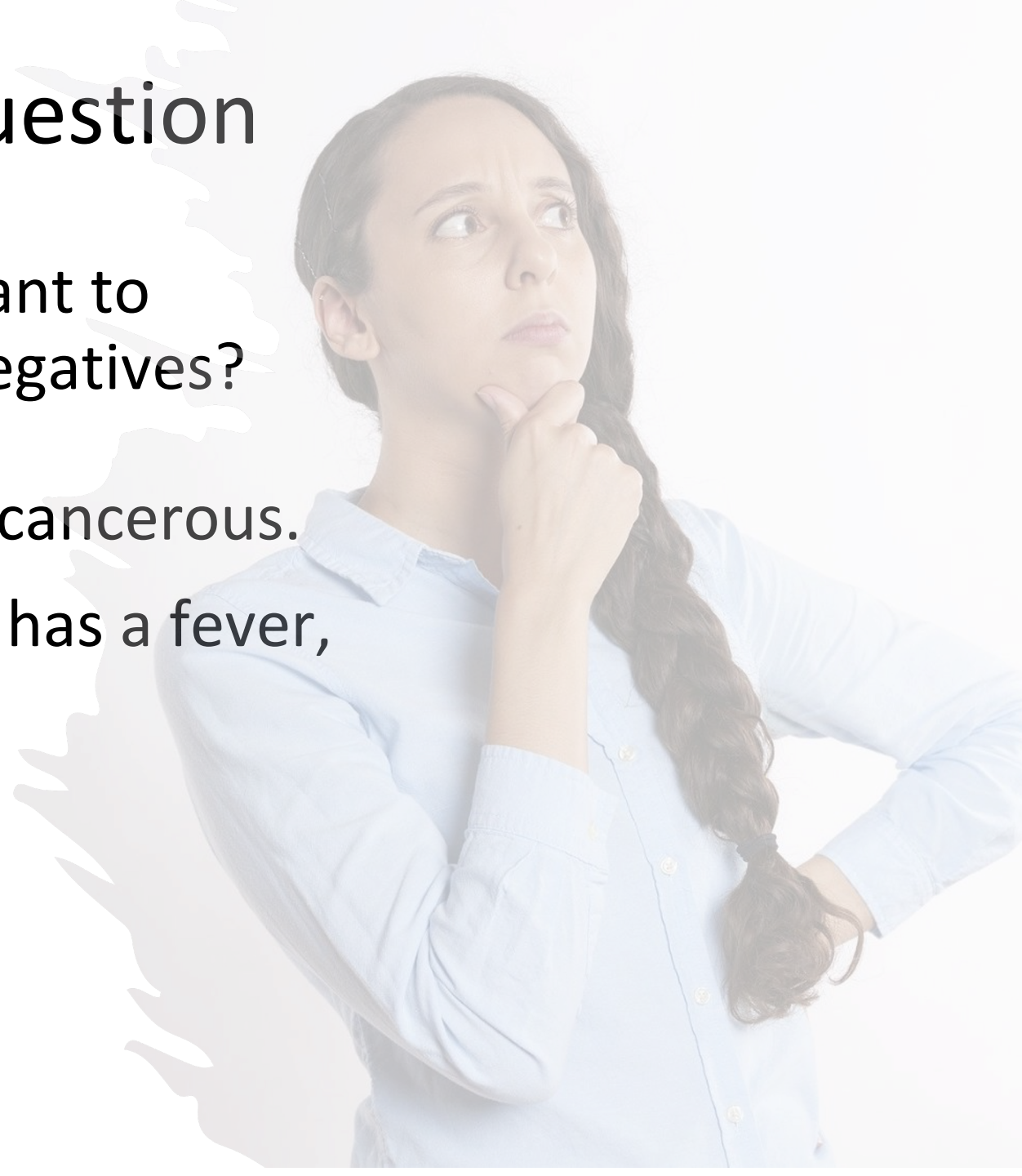
False positives: when the algorithm identifies someone as **having** X, but they **do not** have X

False negatives: when the algorithm identifies someone as **not having** X, but they **do** have X

# Discussion Question

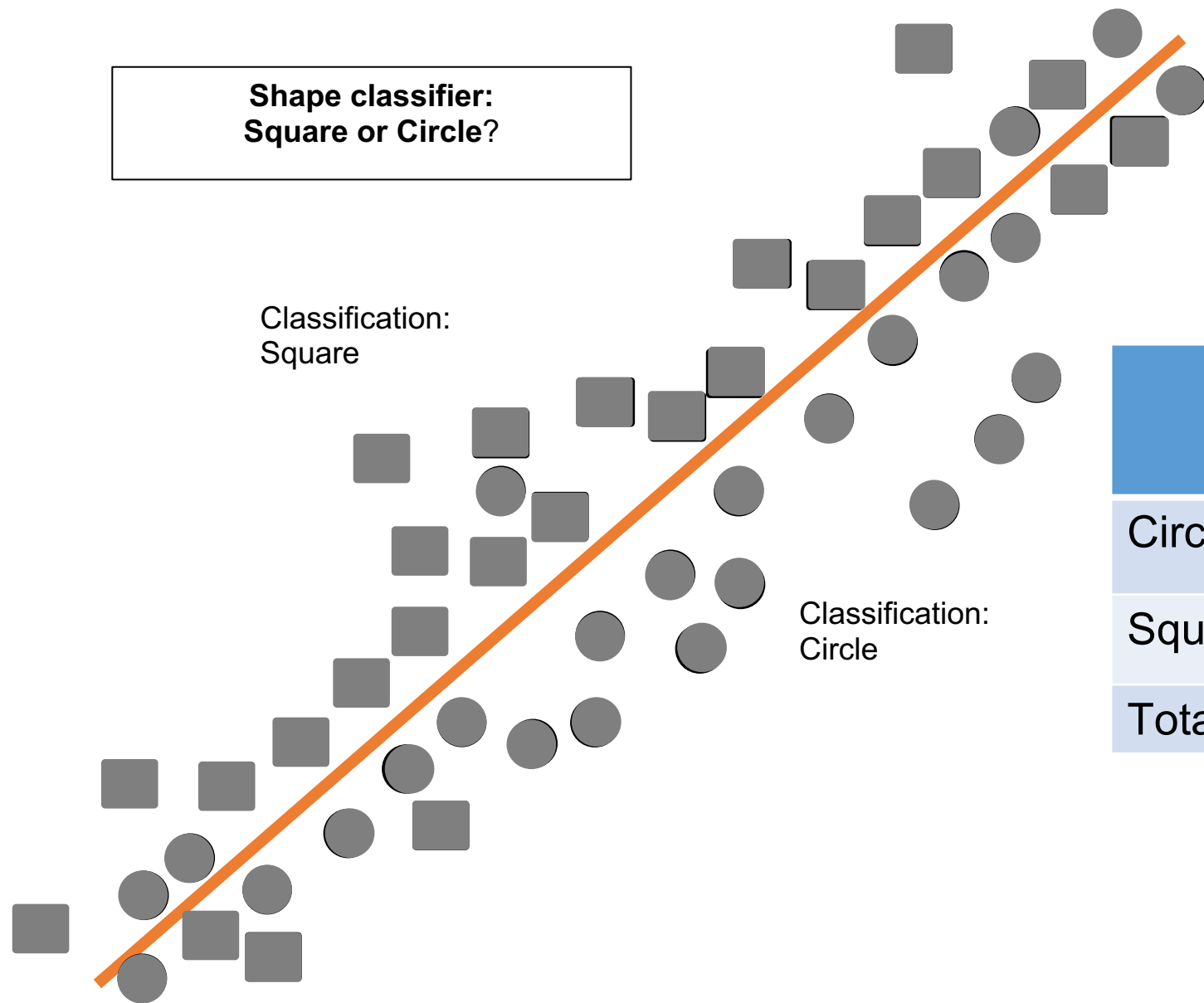
In the following, is it more important to minimize false positives or false negatives?

- 1) Determining whether a mole is cancerous.
- 2) Determining whether someone has a fever, for COVID screening at an airport.

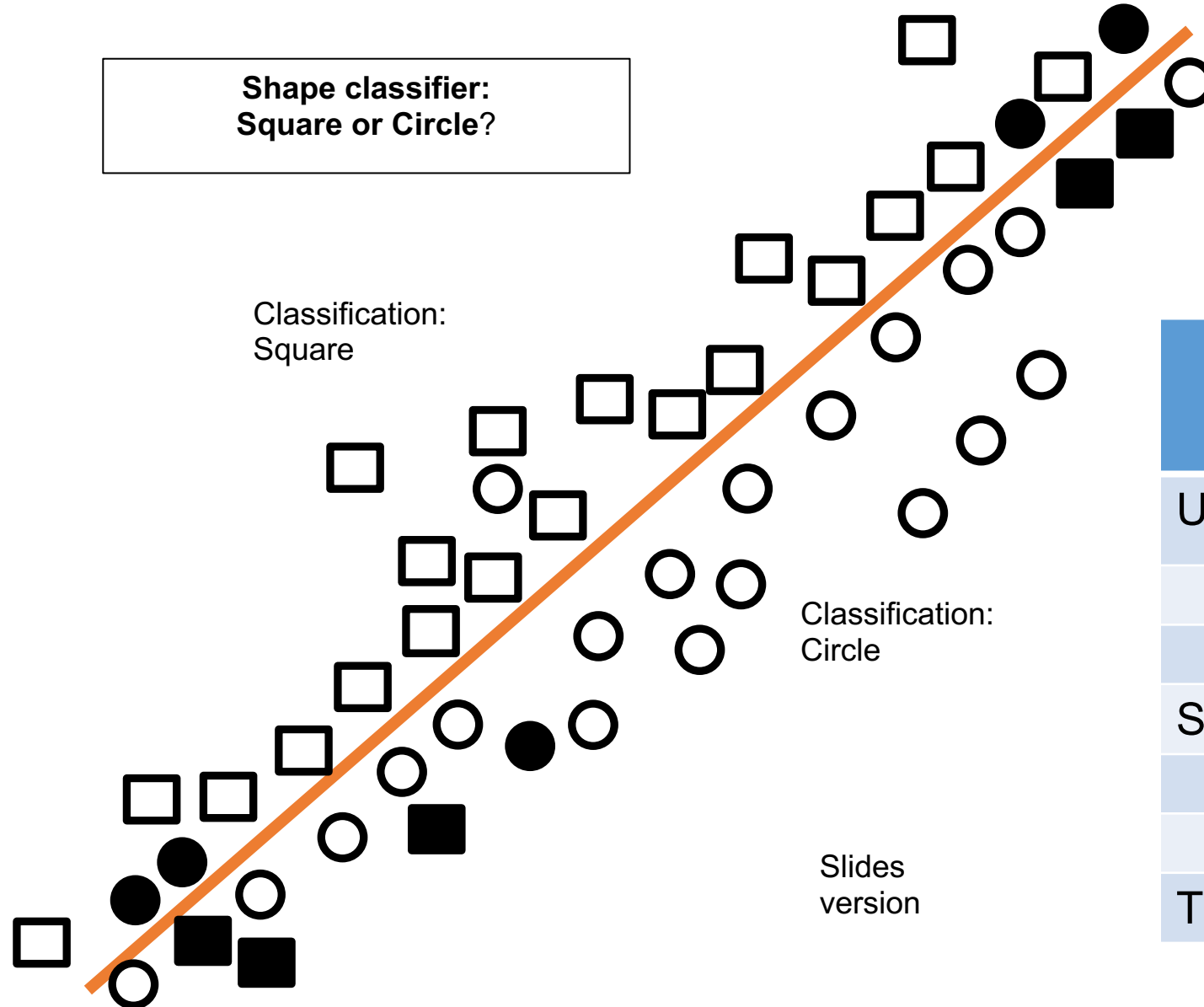


# Worksheet: Part 1

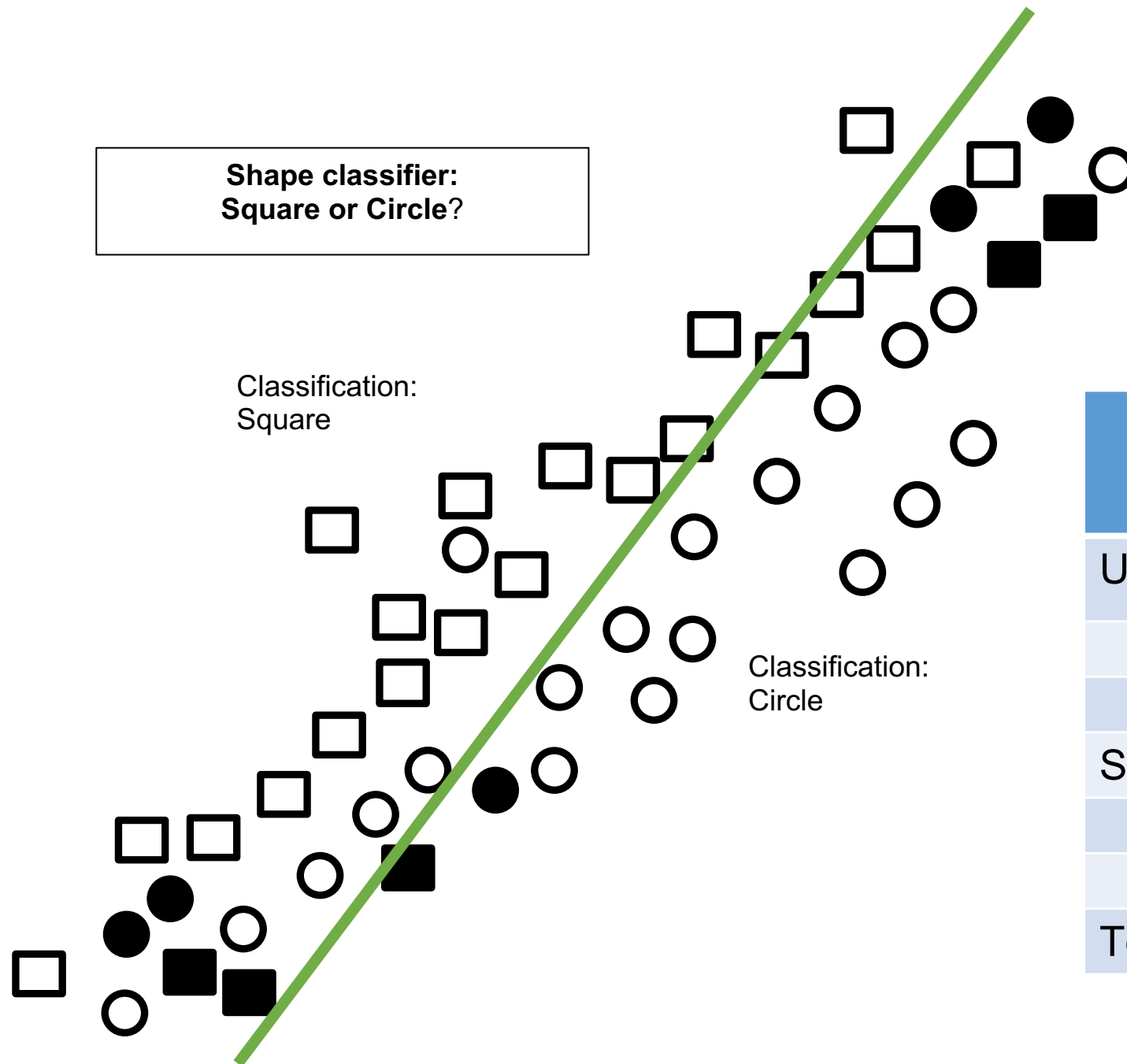




	Correctly classified
Circles	20/25
Squares	20/25
Total	40/50



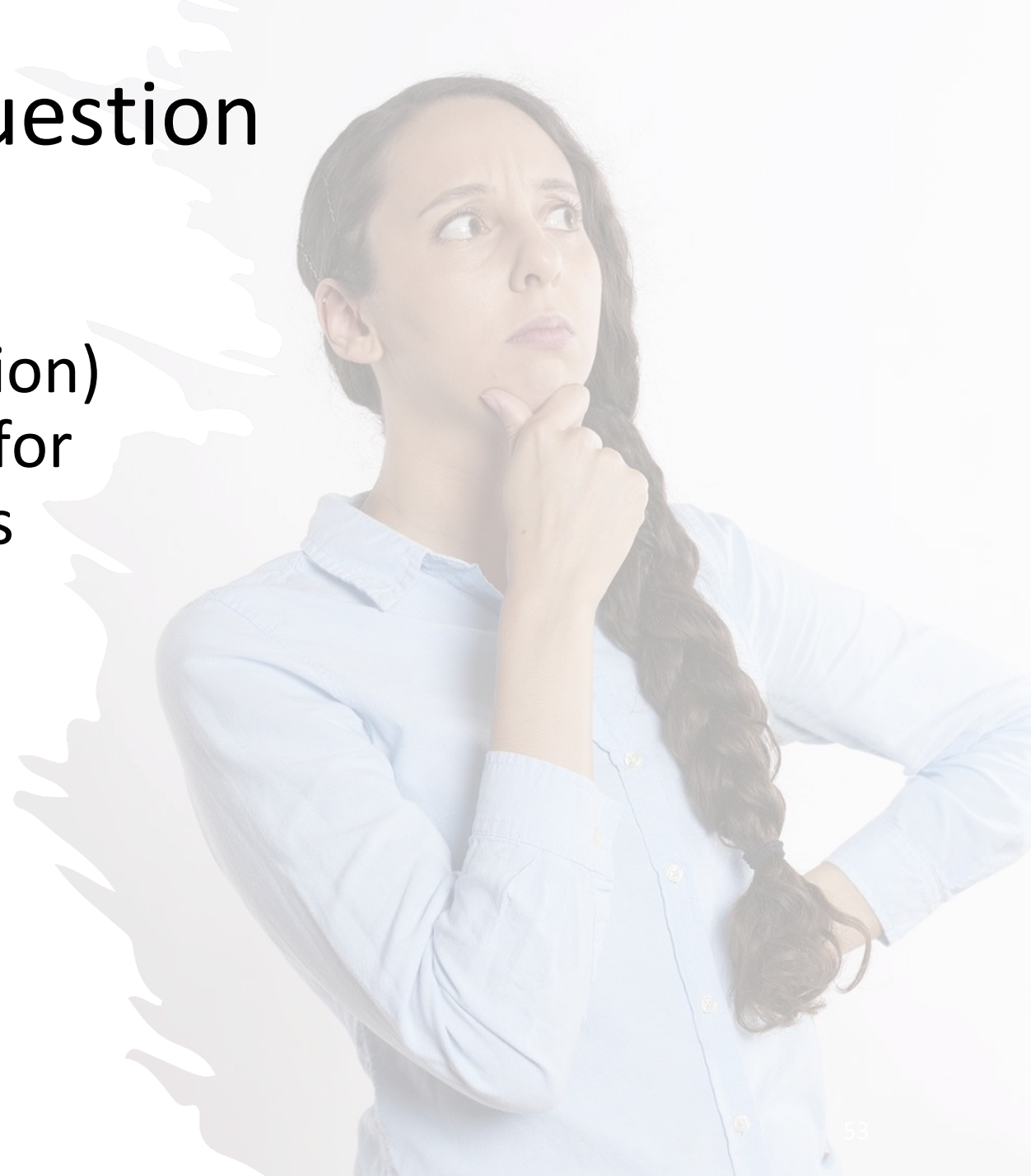
	Correctly classified
Unshaded	39/40
<i>Squares</i>	20/20
<i>Circles</i>	19/20
Shaded	1/10
<i>Squares</i>	0/5
<i>Circles</i>	1/5
Total	41/50



	Correctly classified
Unshaded	29/40
<i>Squares</i>	16/20
<i>Circles</i>	13/20
Shaded	5/10
<i>Squares</i>	2/5
<i>Circles</i>	3/5
Total	35/50

# Discussion Question

What are some machine learning techniques (beyond linear regression) that would improve performance for both shaded and unshaded shapes without compromising overall accuracy?



# Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

*by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica*

*May 23, 2016*

## Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

From COMPAS study: False Negatives / False Positives sometimes differ across subgroups even when total accuracy is the same



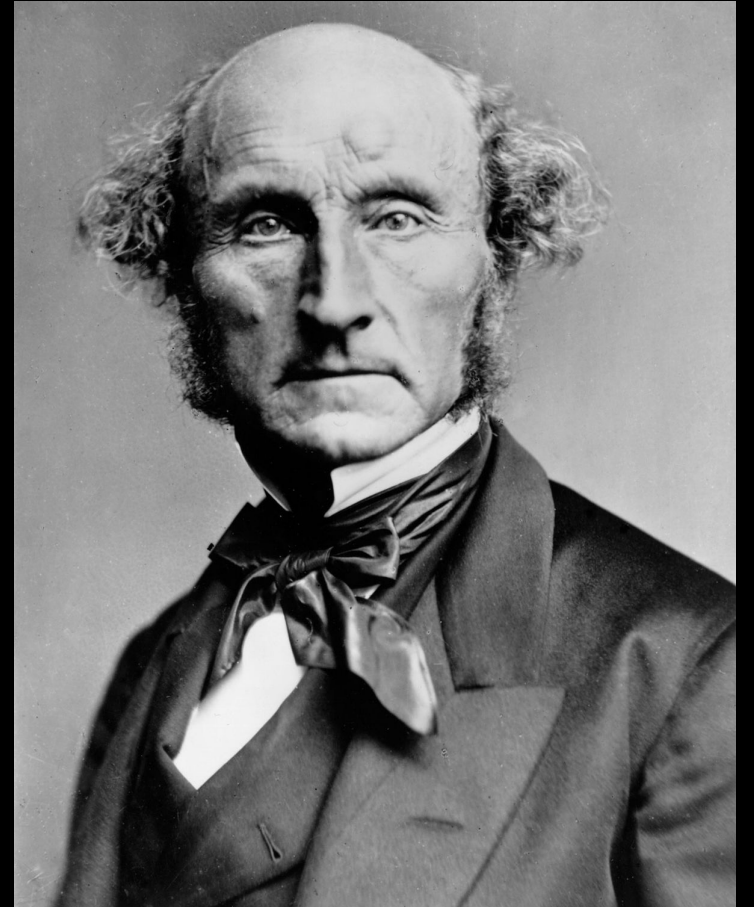
# Worksheet: Part 2

**Trade-offs:** Should we prioritize overall accuracy or subgroup accuracy?

# The (Simple) Utilitarian Answer: Overall Accuracy

Philosopher John Stuart Mill:

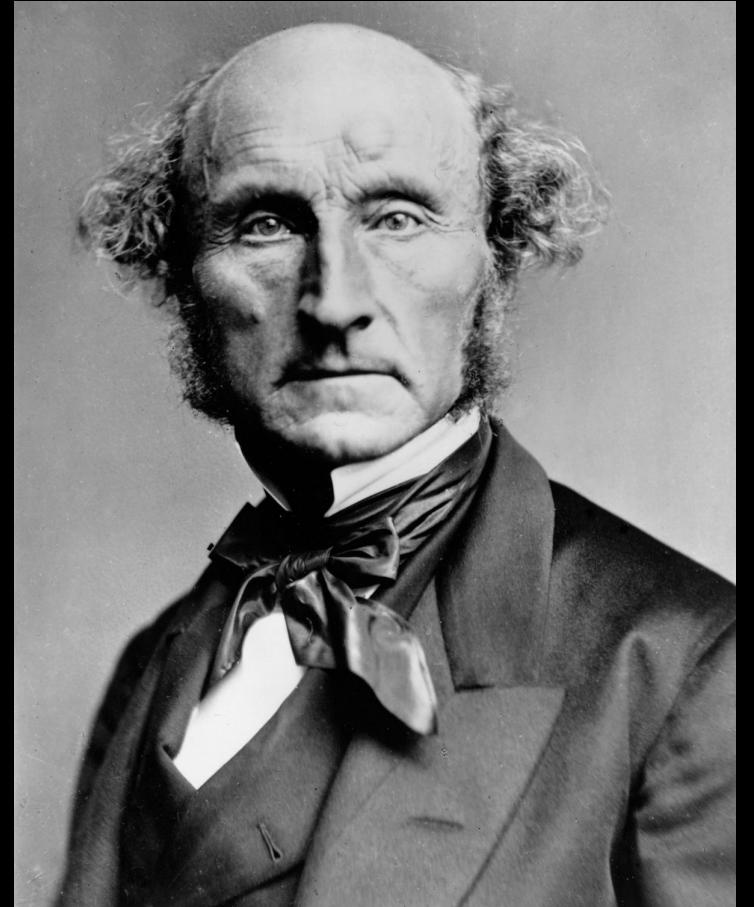
Utilitarianism says that action is morally right when it maximizes **good consequences** (e.g. happiness) and minimizes **bad consequences** (e.g. pain)



## The (Simple) Utilitarian Answer: Overall Accuracy

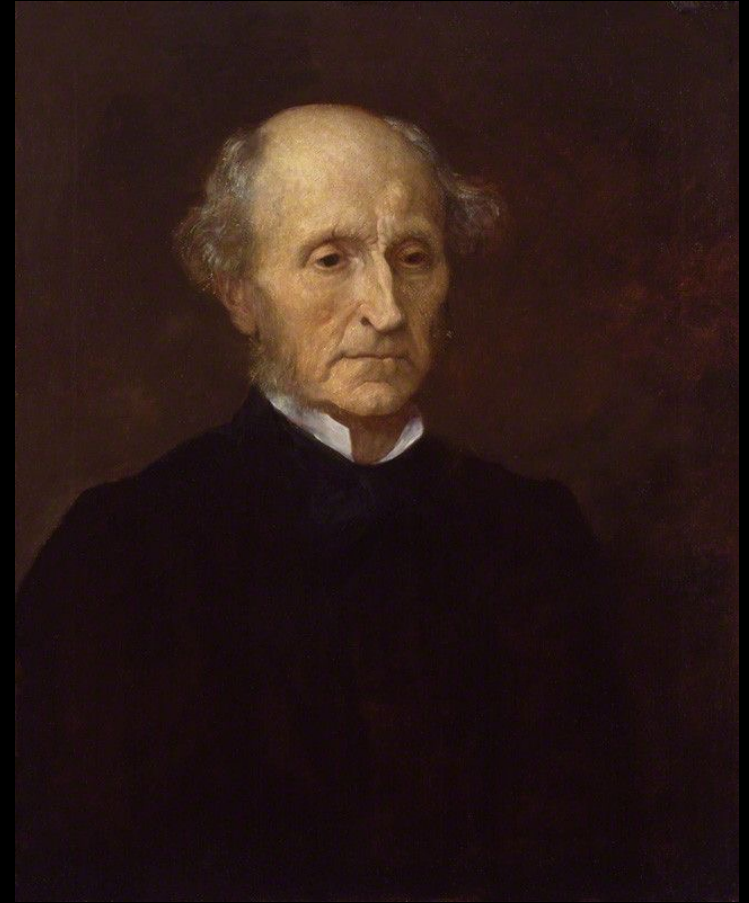
Utilitarianism: you should minimize the total number of errors.

Every time you prioritize subgroup accuracy over overall accuracy, there are more errors (e.g. more misdiagnoses of cancerous moles -> more people getting cancer)



## The (Complex) Utilitarian Answer: Sometimes Both

However, a utilitarian could also argue people from some subgroups are **more vulnerable** to harms from misdiagnoses.





# Discussion Question

Why might elderly or racialized individuals be more vulnerable to harms from misdiagnoses?



## Other Possible Reasons to Prioritize Subgroup Accuracy:

1. Have some subgroups suffered more harms already (so it would be **unfair** to give them more harms)?

**Forehead thermometers may be less accurate at detecting fevers in Black people, study finds**



By Jacqueline Howard, CNN

🕒 Updated 6:48 PM ET, Wed September 7, 2022

# Other Possible Reasons to Prioritize Subgroup Accuracy:

2. Is the algorithm **discriminatory** towards subgroups?



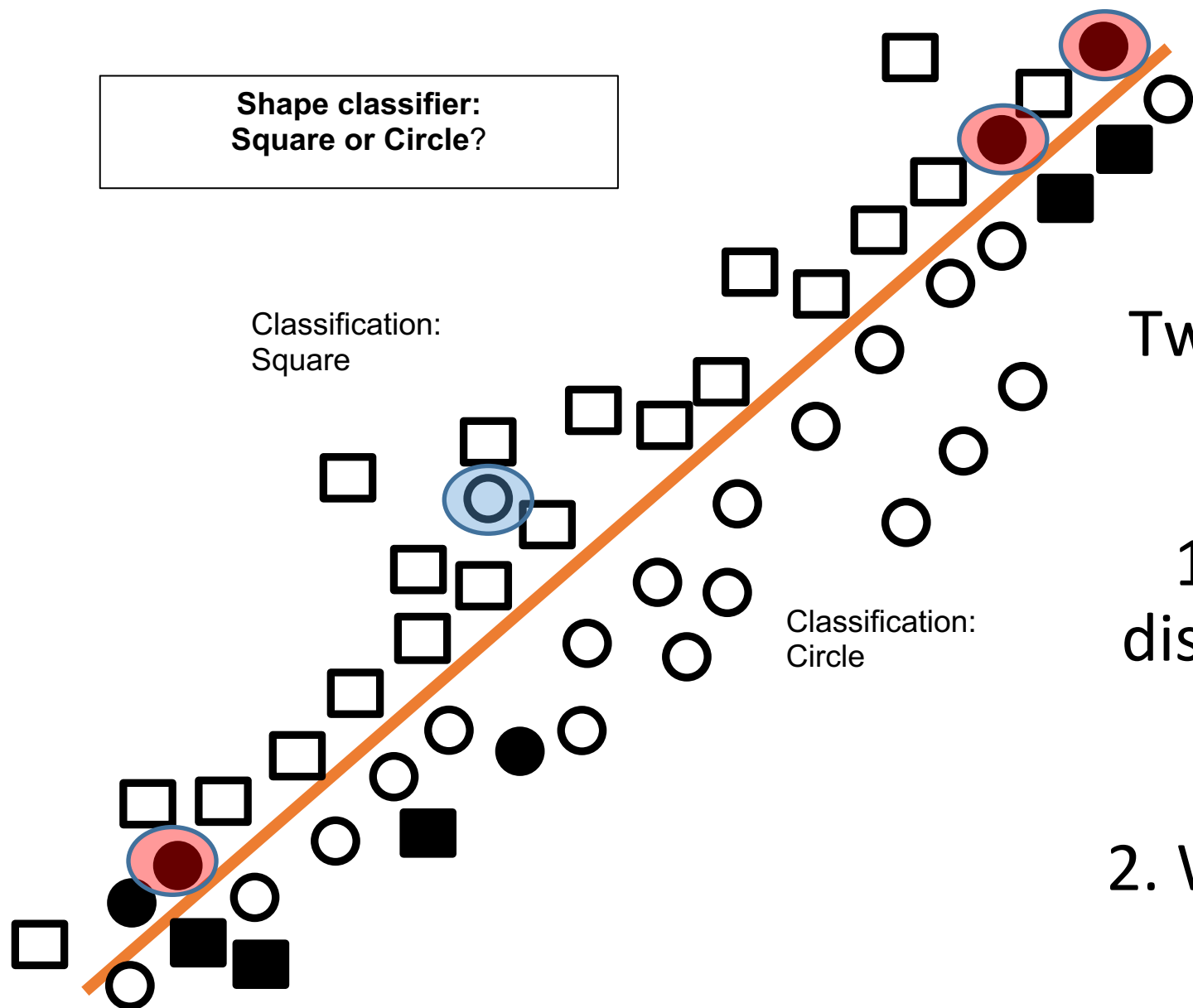
## Other Possible Reasons to Prioritize Subgroup Accuracy:

2. Is the algorithm **discriminatory** towards subgroups?

Some straightforward cases:

- The negative effect on the subgroup is **intentional**.
- The subgroup error is due to a subgroup being **under-represented** in a dataset.

But many would say that an algorithm can be discriminatory in other cases too.



## Two Challenges for Theories of Discrimination:

1. How do we distinguish discrimination from ordinary misclassification?
2. What makes discrimination worse than ordinary misclassification?





“Discrimination is a distinction which, **whether intentional or not but based on** grounds relating to **personal characteristics of the individual or group**, has an effect which **imposes disadvantages not imposed upon others** or which withholds or limits access to advantages available to other members of society.”

Supreme Court of Canada, *Andrews v. Law Society of British Columbia* (1989)

Some of the standard personal characteristics involved in discrimination:

- Gender identity
- Race
- Sexuality
- Disability
- Age

(In Canadian Law, these are called “enumerated grounds”.)





To understand **when** an algorithm is discriminatory, we need to think about **why** discrimination is particularly wrong.

Philosophers actually have big disagreements about this! But here are some of their theories...





**Irrationality:** Discrimination is wrong because it involves treating someone arbitrarily or “without rational justification”. (Cotter 2006, 10)



**Immutability:** Discrimination is wrong because it treats someone differently on the basis of an immutable trait, which is a trait that is “so central to a person’s identity that it would be abhorrent for government to penalize a person for refusing to change them, regardless of how easy that change might be physically” (Clarke 2015: 1)





**Demeaning:** Discrimination is wrong because it involves demeaning another person, or expressing that that person has lower moral status because they belong to some group (Hellman 2018: 102)

# **How airport scanners discriminate against passengers of color**

Full-body scanners often have trouble reading thick hair and certain head coverings — contributing to racist profiling.

# Discussion Question

Do any of these considerations {irrationality, immutability, demeaningness} apply when someone is misdiagnosed because they are elderly or racialized?



# Part 4:

## The Law of Bias and Discrimination





As of October 2022, as part of Bill C-27, the Parliament of Canada is considering legislation that will regulate AI: the *Artificial Intelligence and Data Act*:

Anyone who processes data or makes it available for use must take steps to ensure it is **anonymous**.

Anyone who is responsible for a “high impact system” must take steps to identify, assess and mitigate the risks of **harm** or **biased output**.



**Harm** means

- (a) physical or psychological harm to an individual;
- (b) damage to an individual's property; or
- (c) economic loss to an individual.

**biased output** means content that is generated, or a decision, recommendation or prediction that is made, by an artificial intelligence system and that adversely differentiates, directly or indirectly and without justification, in relation to an individual on one or more of the prohibited grounds of discrimination set out in section 3 of the Canadian Human Rights Act, or on a combination of such prohibited grounds.





Every person who commits an offence under section 38 or 39

(a) is liable, on conviction on indictment,

(i) to a fine of not more than the greater of **\$25,000,000** and **5% of the person's gross global revenues** in its financial year before the one in which the person is sentenced, in the case of a person who is not an individual, and

(ii) to a fine in the discretion of the court or to a term of **imprisonment of up to five years less a day**, or to both, in the case of an individual;

# Part 5: Final Summary

## Final Summary

- There are many sources of bias in image processing and image understanding algorithms: in the **datasets** and in the **algorithms** themselves
- By being aware of the potential for bias, we can take some concrete steps to **mitigate** it.
- Sometimes there may be **trade-offs** between overall accuracy and subgroup activity, and we need to think carefully in these cases about what to prioritize.

# Acknowledgements

This module was created as part of an Embedded Ethics Education Initiative (E3I), a joint project between the Department of Computer Science<sup>1</sup> and the Schwartz Reisman Institute for Technology and Society<sup>2</sup>, University of Toronto.

## **Instructional Team:**

Philosophy: Steven Coyne

Computer Science: Babak Taati, David Lindell

## **Faculty Advisors:**

Diane Horton<sup>1</sup>, David Liu<sup>1</sup>, and Sheila McIlraith<sup>1,2</sup>

**Department of Computer Science**  
**Schwartz Reisman Institute for Technology and Society**  
**University of Toronto**



Computer Science  
UNIVERSITY OF TORONTO



UNIVERSITY OF  
TORONTO





# References

- Altman, Andrew, "Discrimination", *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), Edward N. Zalta (ed.), URL = [<https://plato.stanford.edu/archives/win2020/entries/discrimination/>](https://plato.stanford.edu/archives/win2020/entries/discrimination/).
- Clarke, Jessica A. 2015. "Against Immutability," *Yale Law Journal* 125(1): 2-102.
- Cotter, Anne-Marie Mooney, 2006. *Race Matters: An International Legal Analysis of Race Discrimination*. London: Routledge
- Government of Canada, "C-27", URL=< <https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading>>
- Hellman, Deborah. 2018. "Discrimination and Social Meaning," in *The Routledge Handbook of the Ethics of Discrimination* (ed. Kasper Lippert-Rasmussen). London: Routledge
- ProPublica. 2016. "Machine Bias: Risk Assessments in Criminal Sentencing," URL=<<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>>
- The Verge. 2021. "Black Teen barred from skating rink by inaccurate facial recognition," URL=<<https://www.theverge.com/2021/7/15/22578801/black-teen-skating-rink-inaccurate-facial-recognition>>.