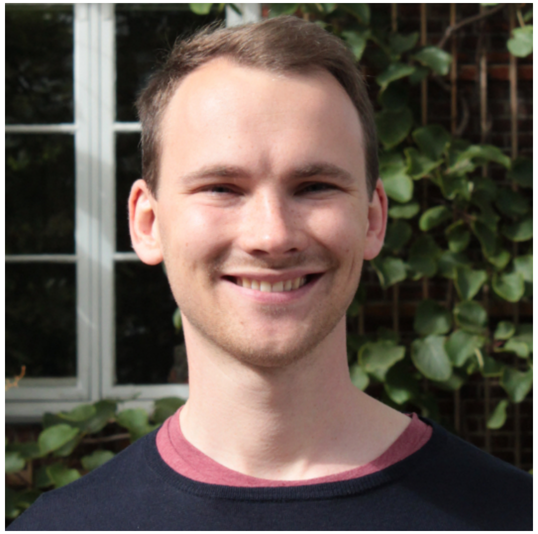
Why is My Classifier Discriminatory?

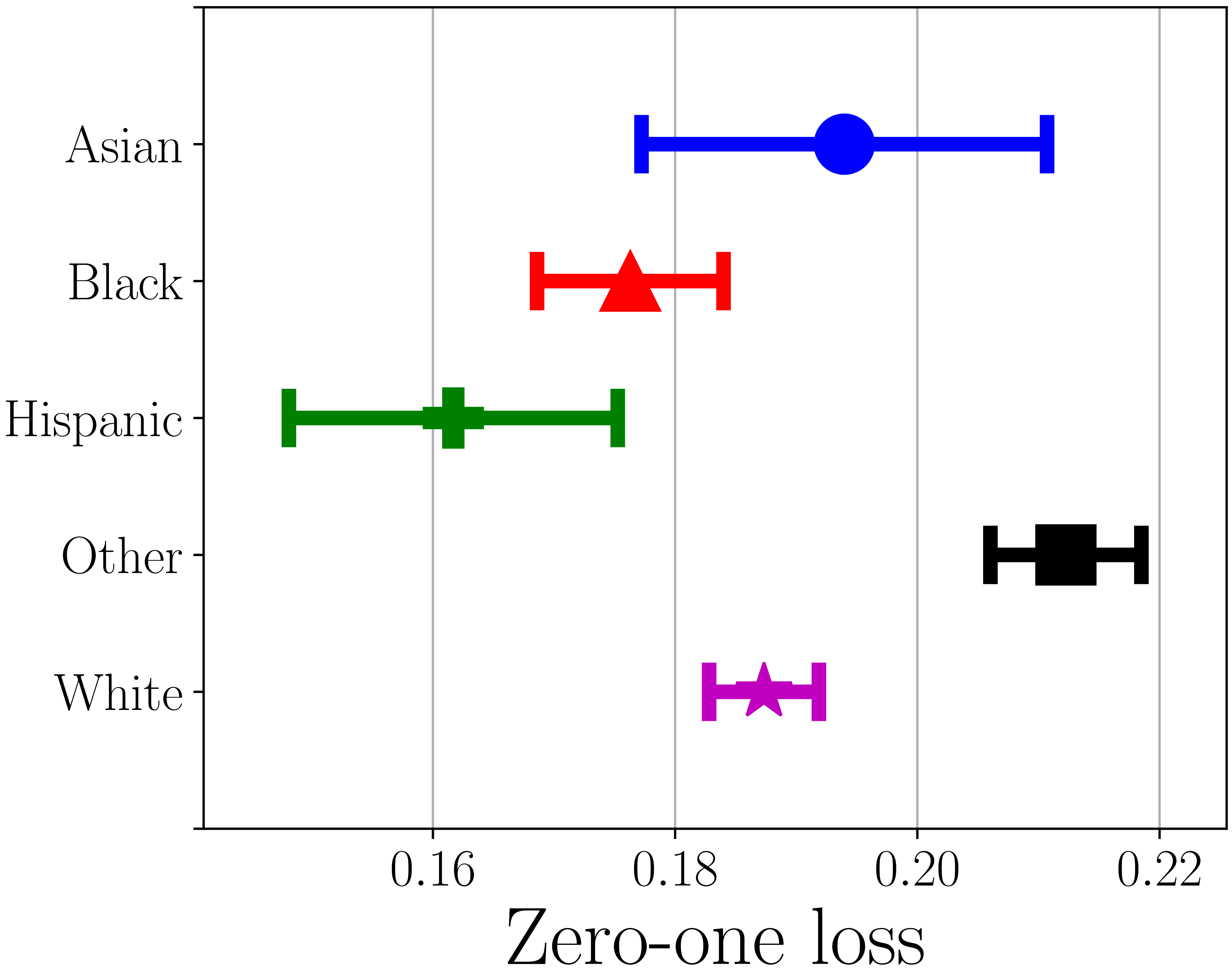

Irene Y. Chen, Fredrik D. Johansson, David Sontag

Massachusetts Institute of Technology (MIT)

NeurIPS 2018, Poster #120 Thurs 12/6 10:45am – 12:45pm @ 210 & 230

It is **surprisingly easy** to make a discriminatory algorithm.

Source: Shutterstock


Asian
Black
Hispanic
Other
White
0.16
0.18
0.20
0.22
Zero-one loss

In this paper

1. We want to find the **sources of unfairness** to guide resource allocation.

In this paper

1. We want to find the **sources of unfairness** to guide resource allocation.

2. We decompose unfairness into **bias, variance, and noise**.

In this paper

1. We want to find the **sources of unfairness** to guide resource allocation.

2. We decompose unfairness into **bias, variance, and noise**.

3. We demonstrate methods to guide **feature augmentation** and **training data collection** to fix unfairness.

Classification fairness: many factors

Model

- Loss function constraints
  - Kamairan et al, 2010; Zafar et al, 2017
- Representation learning
  - Zemel et al, 2013
- Regularization
  - Kamishima et al, 2007; Bechvod and Ligett, 2017
- Tradeoffs
  - Chouldechova, 2017; Kleinberg et al, 2016; Corbett-Davies et al, 2017

Classification fairness: many factors

Model

- Loss function constraints
  - Kamairan et al, 2010; Zafar et al, 2017
- Representation learning
  - Zemel et al, 2013
- Regularization
  - Kamishima et al, 2007; Bechvod and Ligett, 2017
- Tradeoffs
  - Chouldechova, 2017; Kleinberg et al, 2016; Corbett-Davies et al, 2017

Data

Classification fairness: many factors

Model

- Loss function constraints
  - Kamairan et al, 2010; Zafar et al, 2017
- Representation learning
  - Zemel et al, 2013
- Regularization
  - Kamishima et al, 2007; Bechvod and Ligett, 2017
- Tradeoffs
  - Chouldechova, 2017; Kleinberg et al, 2016; Corbett-Davies et al, 2017

Data

- Data processing
  - Haijan and Domingo-Ferrer, 2013; Feldman et al, 2015
- Cohort selection
- Sample size
- Number of features
- Group distribution

Classification fairness: many factors

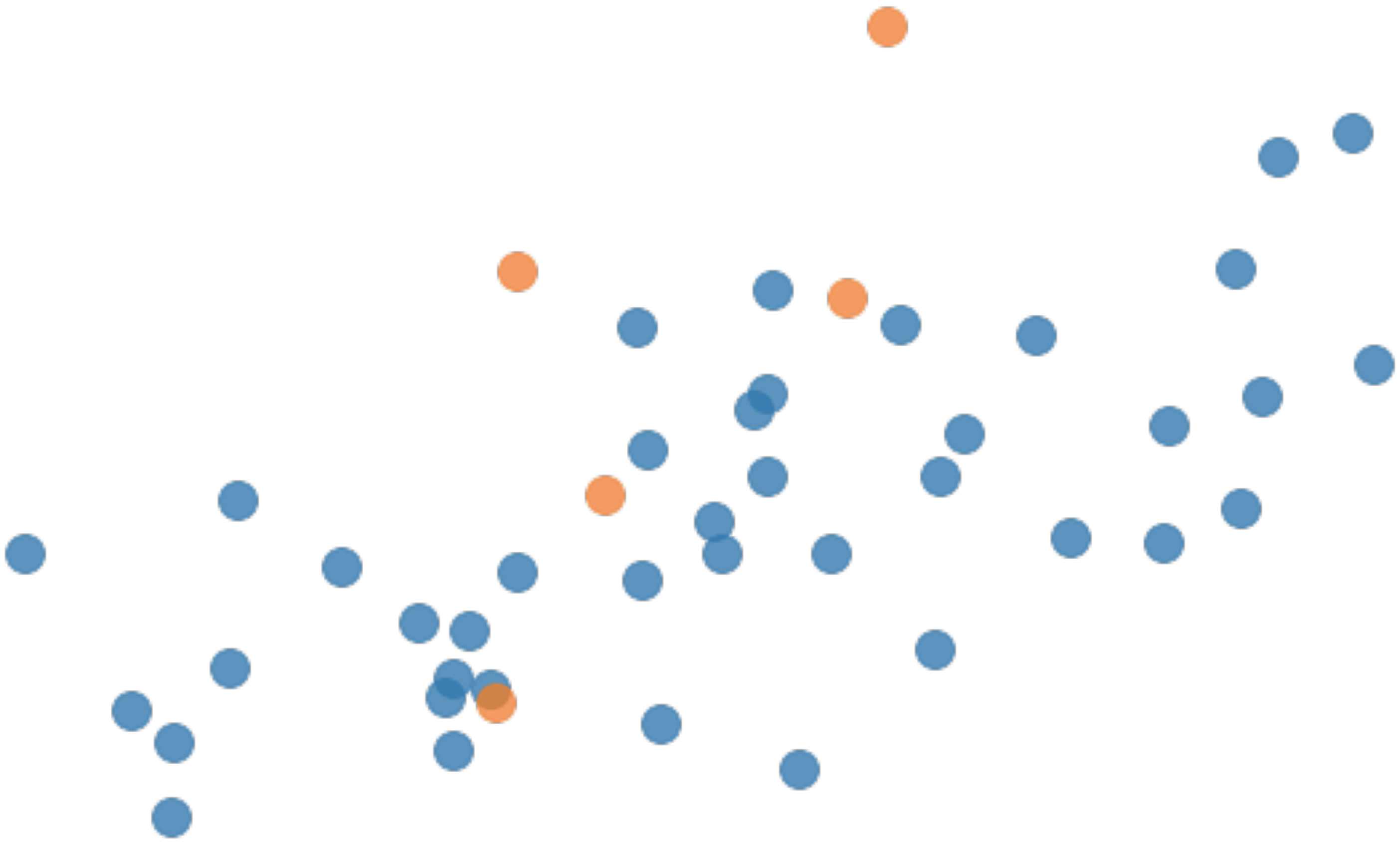
We should examine fairness algorithms in the context of the data and model.

Model
- Loss function constraints
  - Kamairan et al, 2010; Zafar et al, 2017
- Representation learning
  - Zemel et al, 2013
- Regularization
  - Kamishima et al, 2007; Bechvod and Ligett, 2017
- Tradeoffs
  - Chouldechova, 2017; Kleinberg et al, 2016; Corbett-Davies et al, 2017

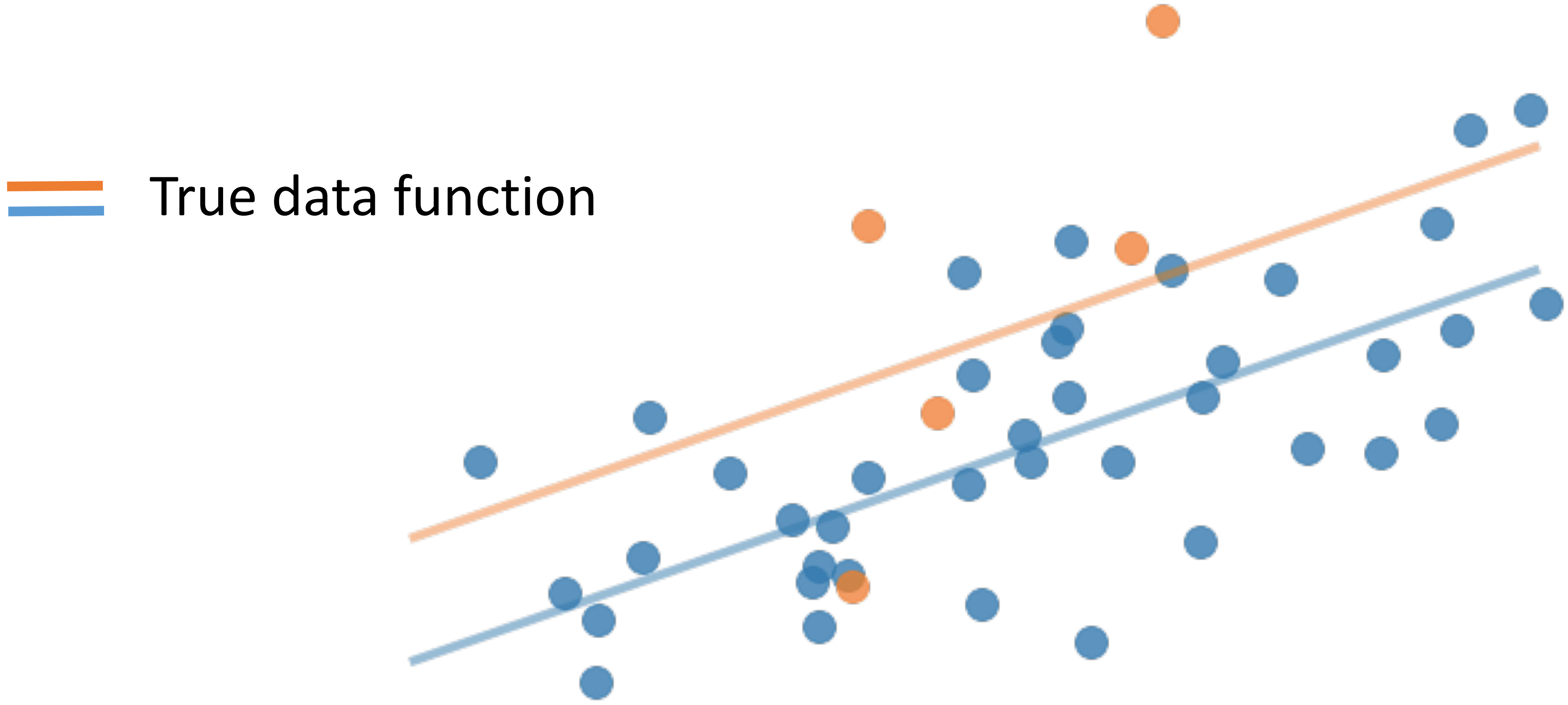
Data
- Data processing
  - Hajian and Domingo-Ferrer, 2013; Feldman et al, 2015
- Cohort selection
- Sample size
- Number of features
- Group distribution

Why might my classifier be unfair?

Why might my classifier be unfair?

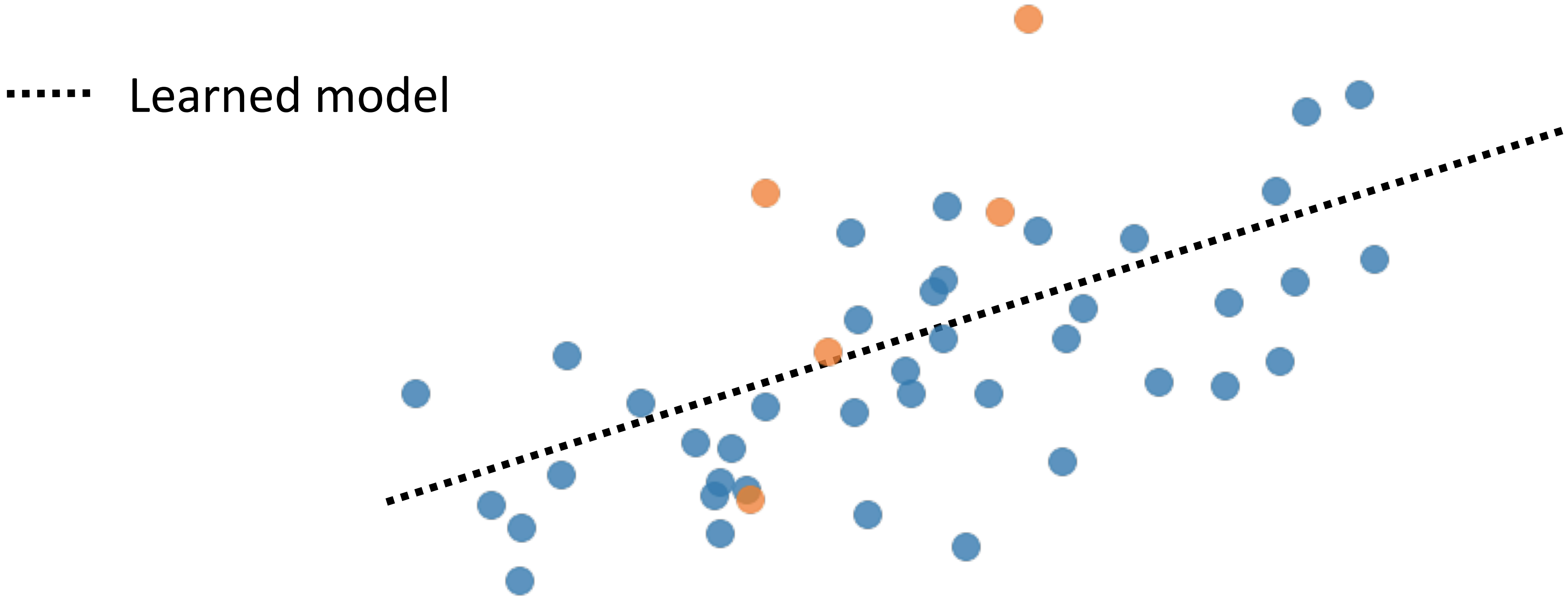


Why might my classifier be unfair?

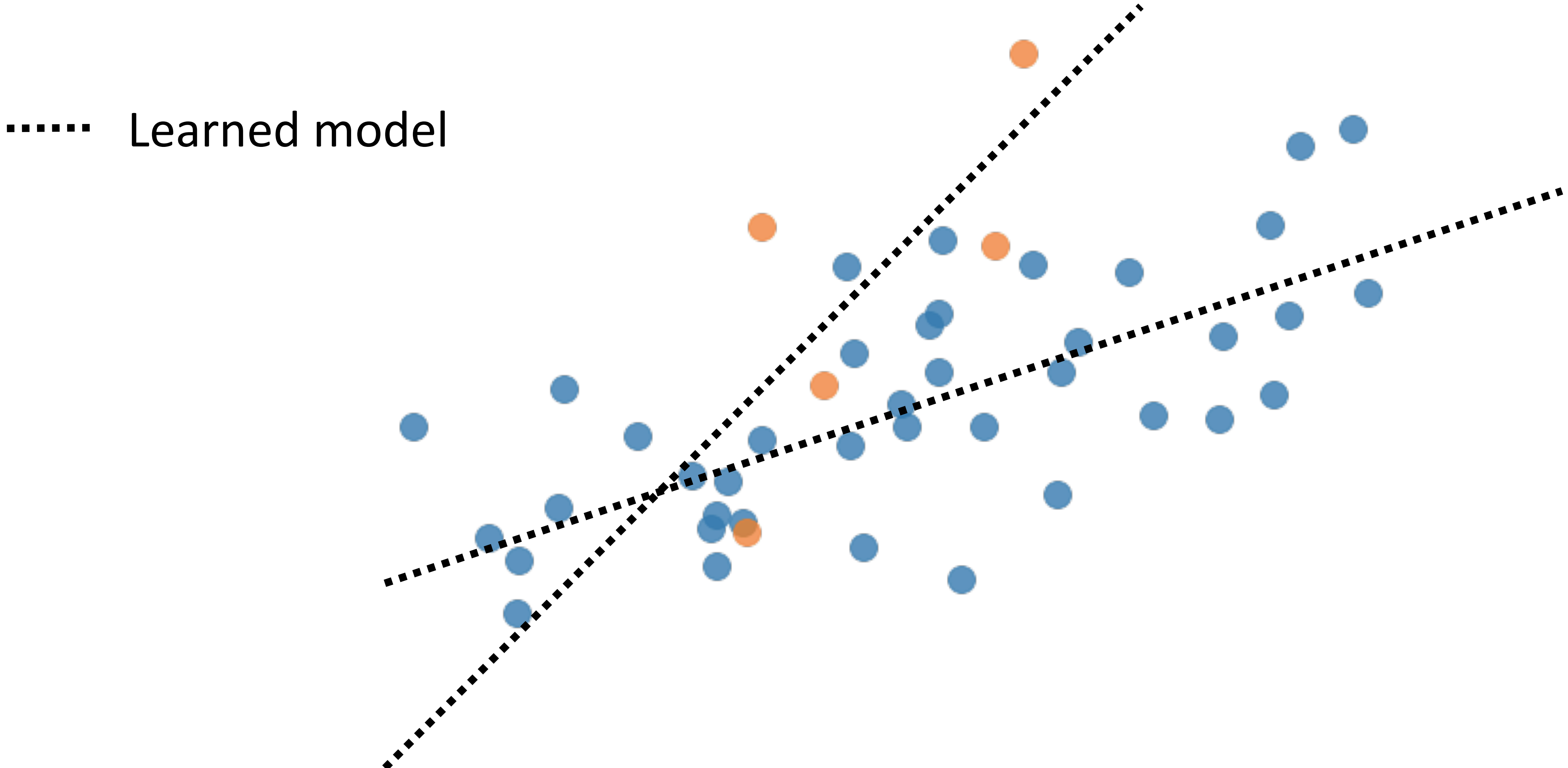


Why might my classifier be unfair?

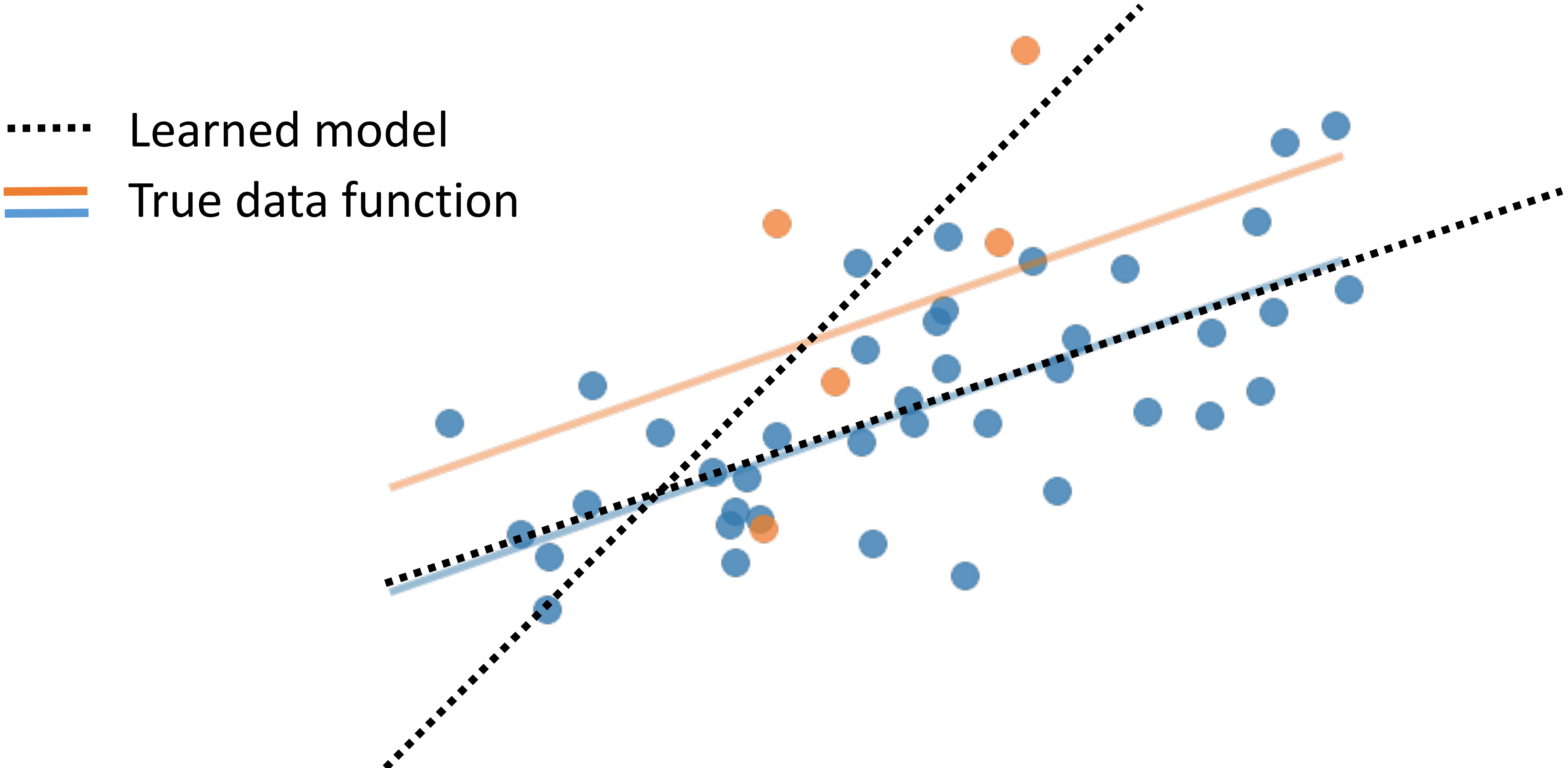
Why might my classifier be unfair?



Why might my classifier be unfair?



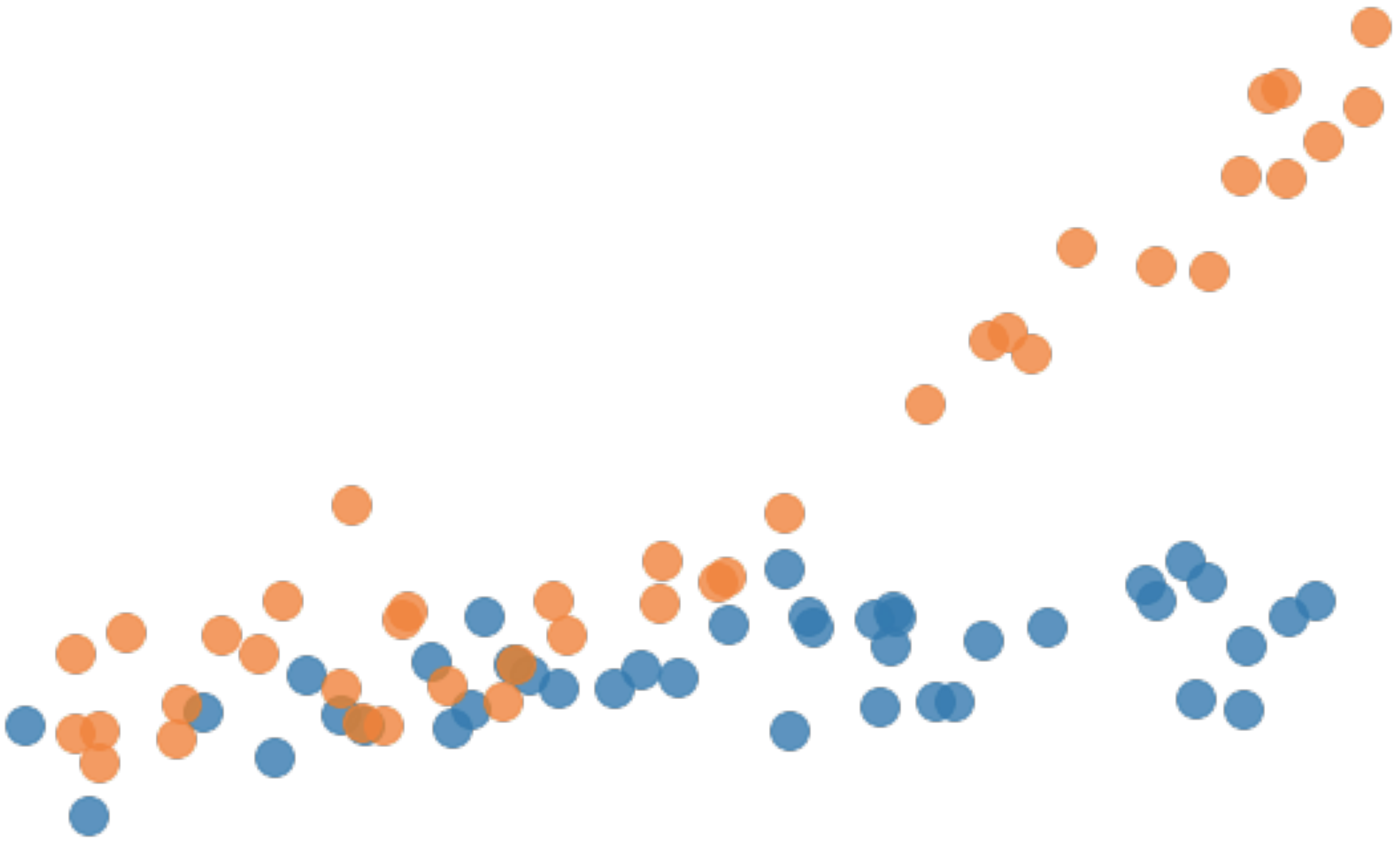
Why might my classifier be unfair?



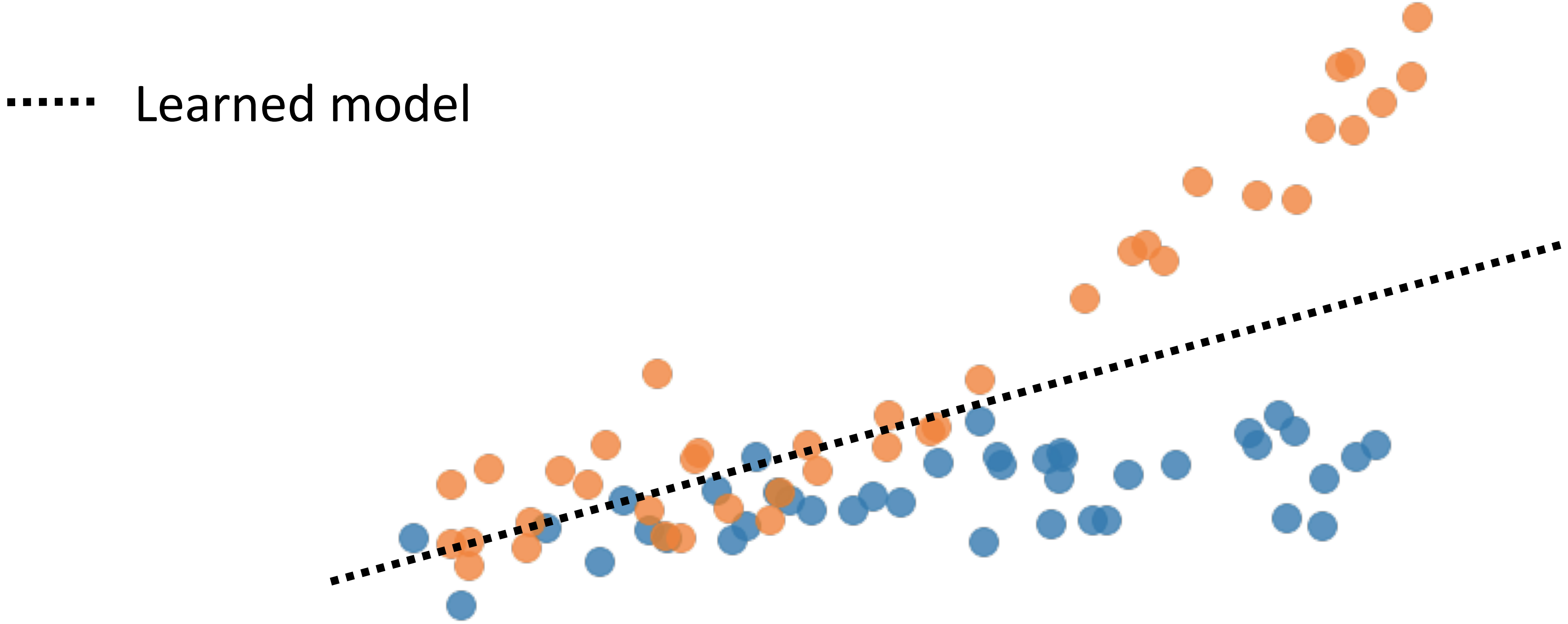
Why might my classifier be unfair?

Error from **variance** can be solved by **collecting more samples.**

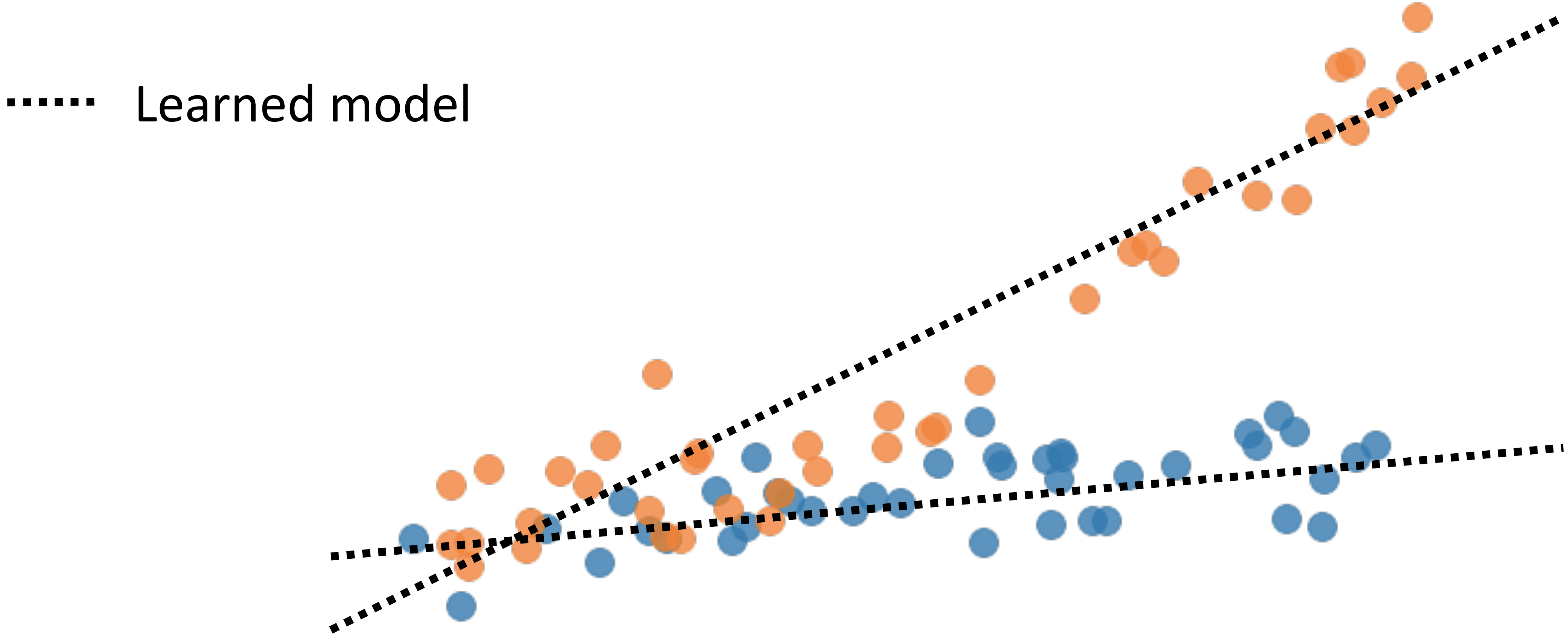
Why might my classifier be unfair?



Why might my classifier be unfair?



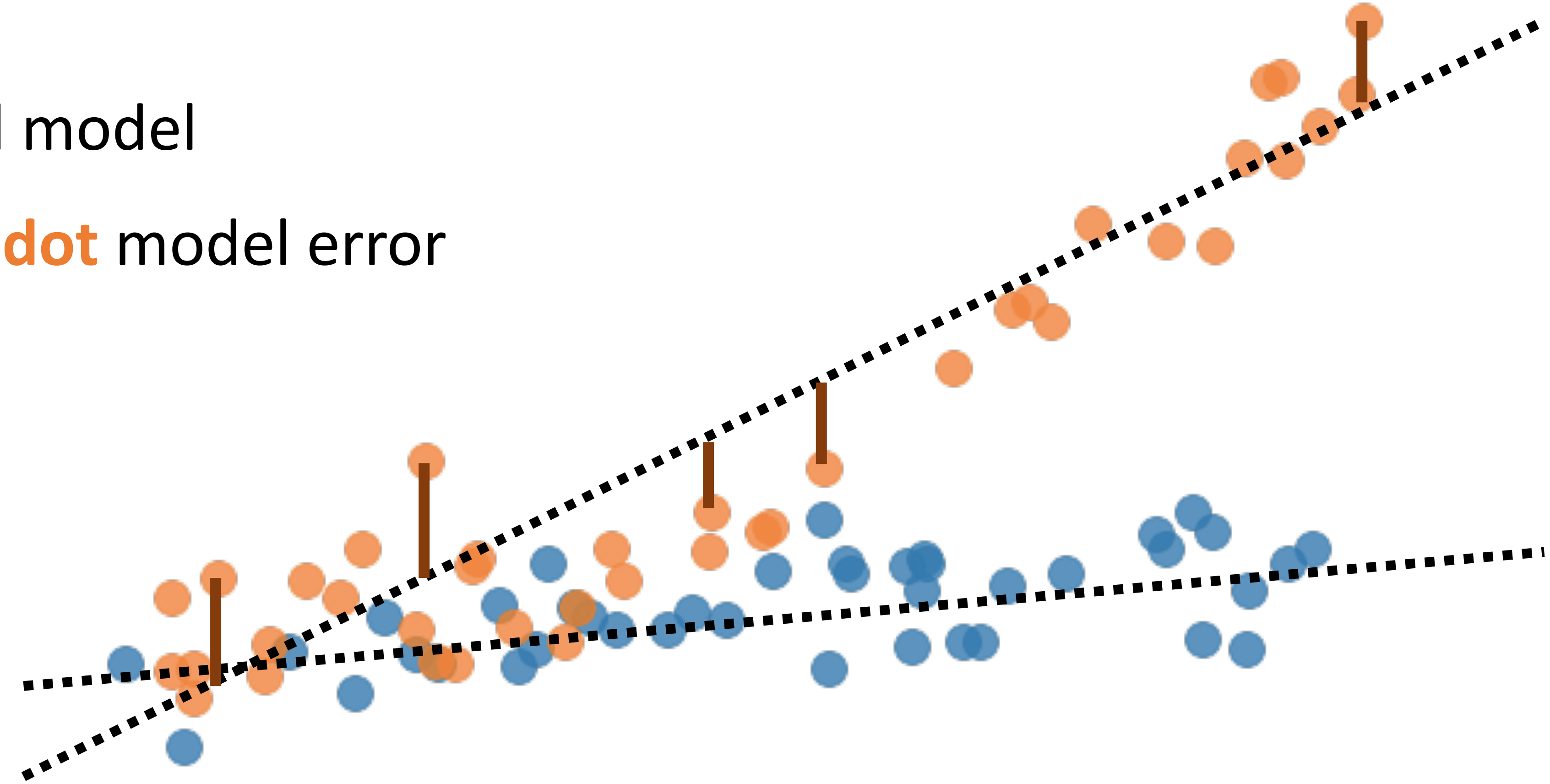
Why might my classifier be unfair?



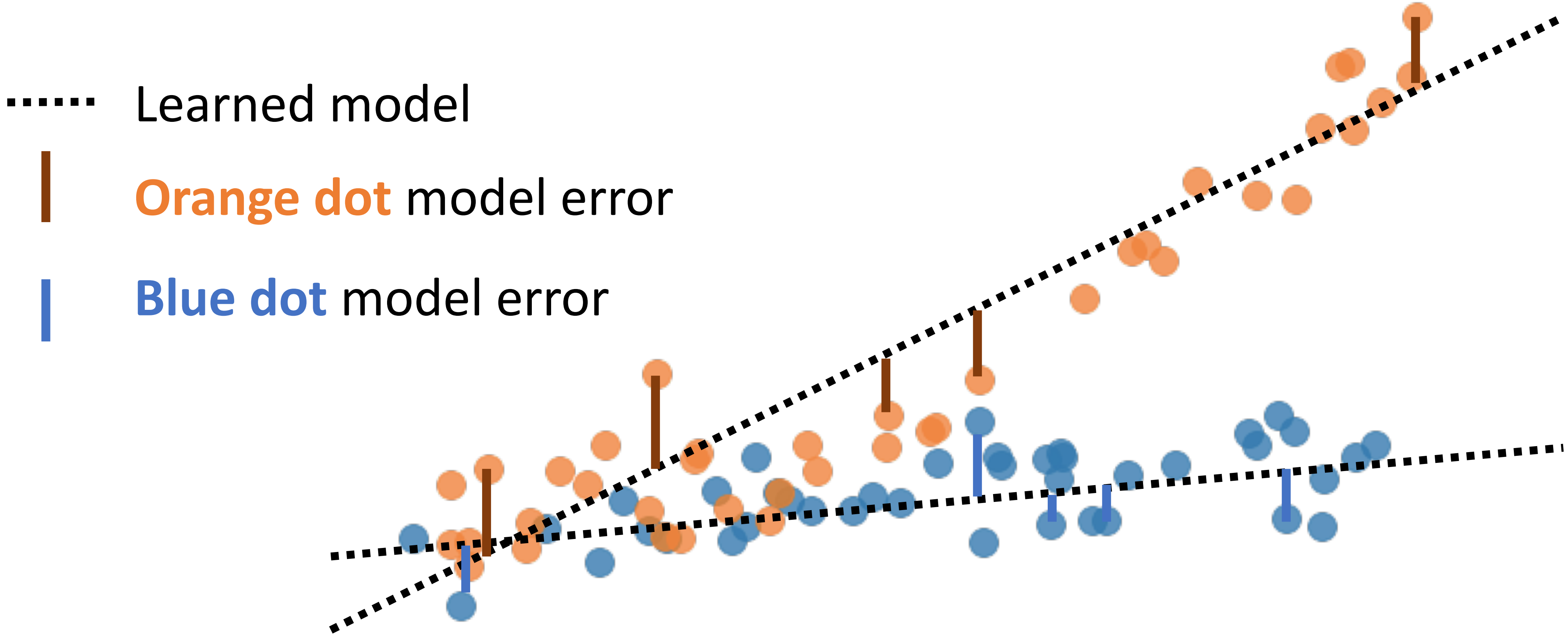
Why might my classifier be unfair?

Learned model

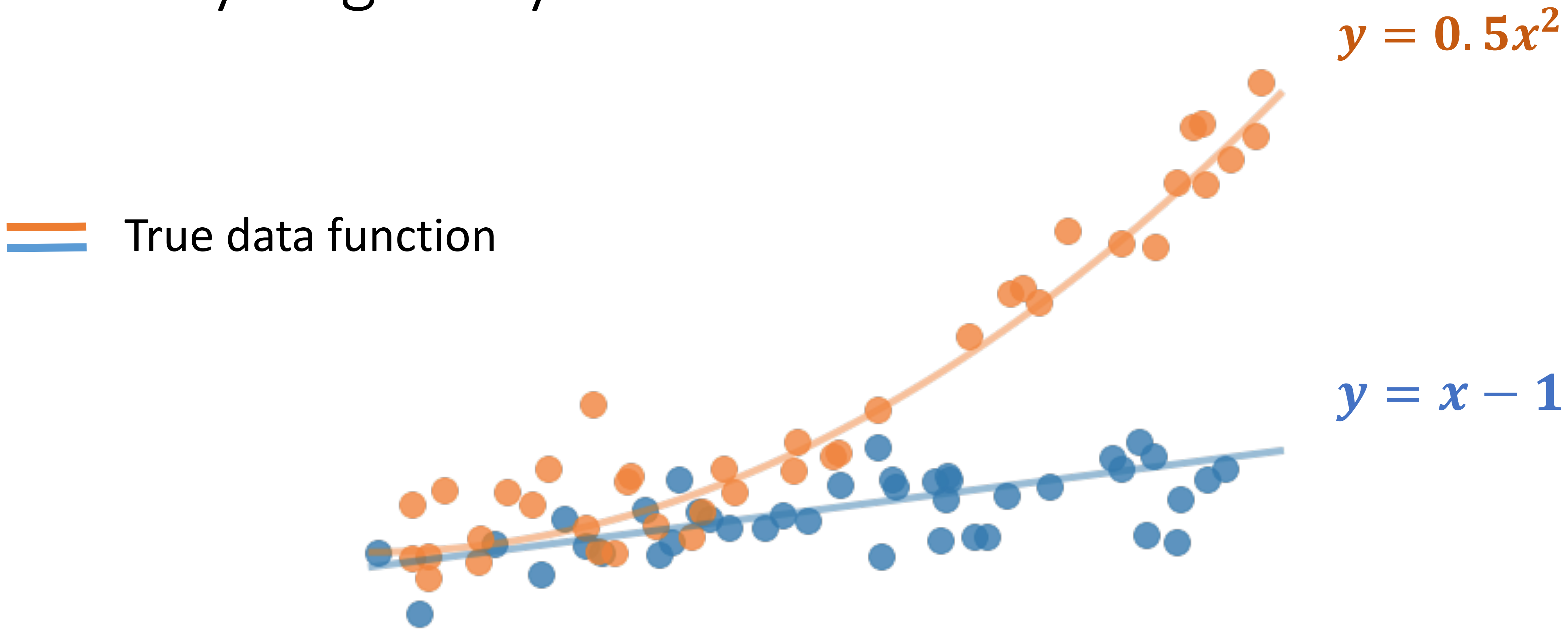
Orange dot model error



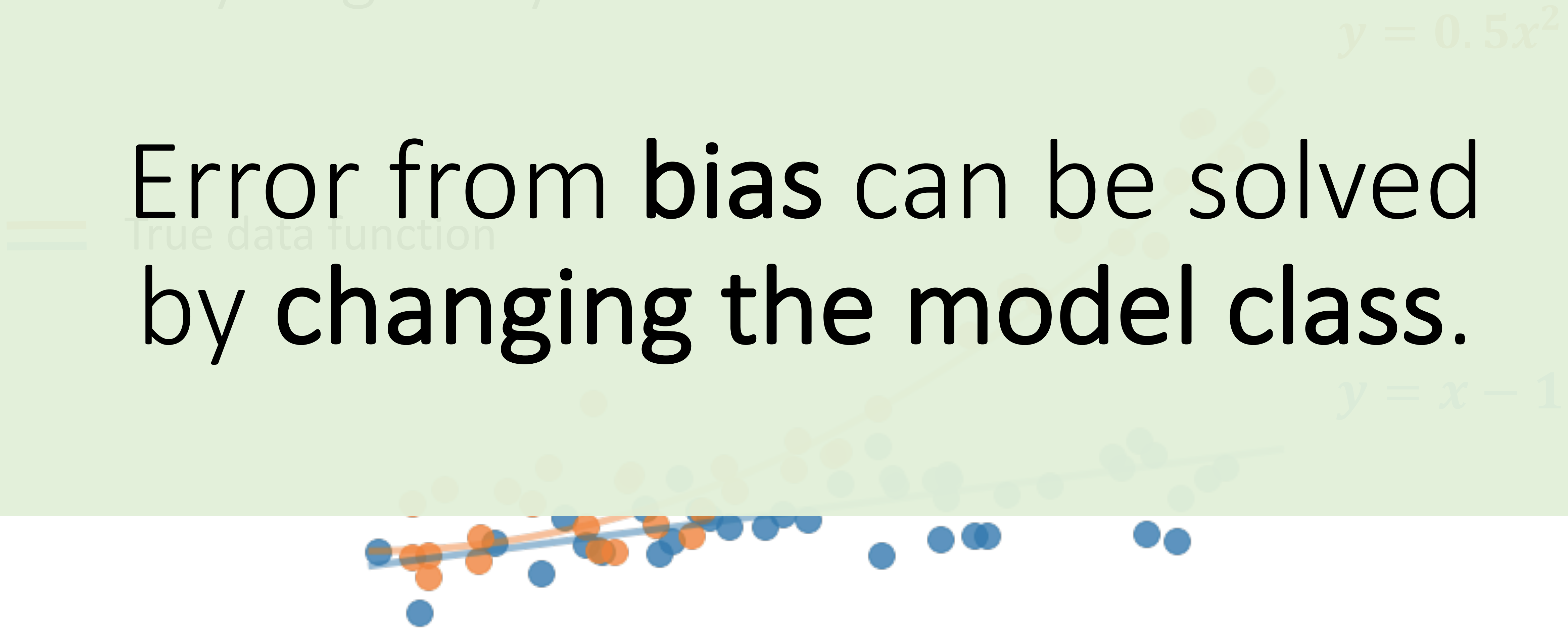
Why might my classifier be unfair?



Why might my classifier be unfair?

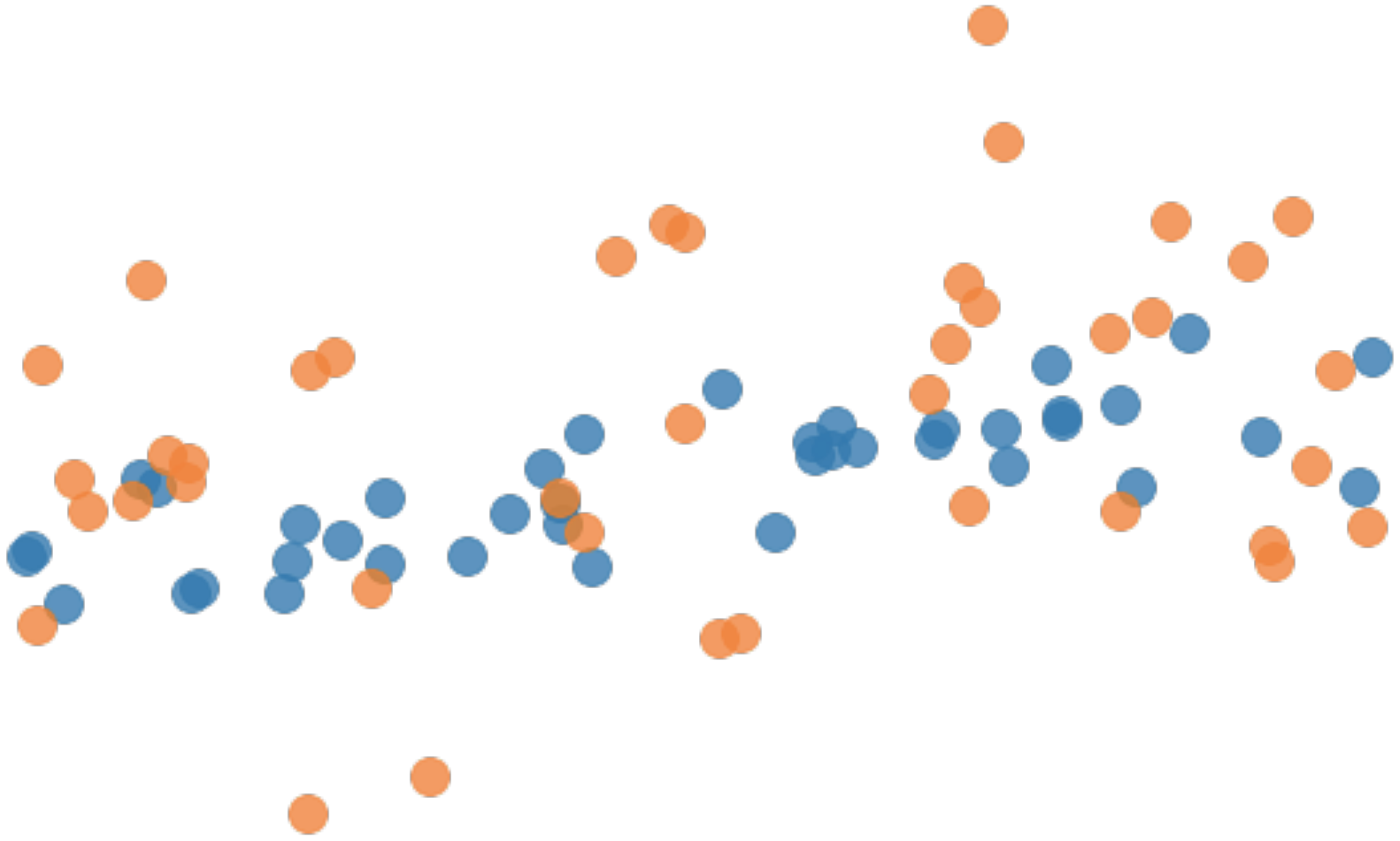


Why might my classifier be unfair?

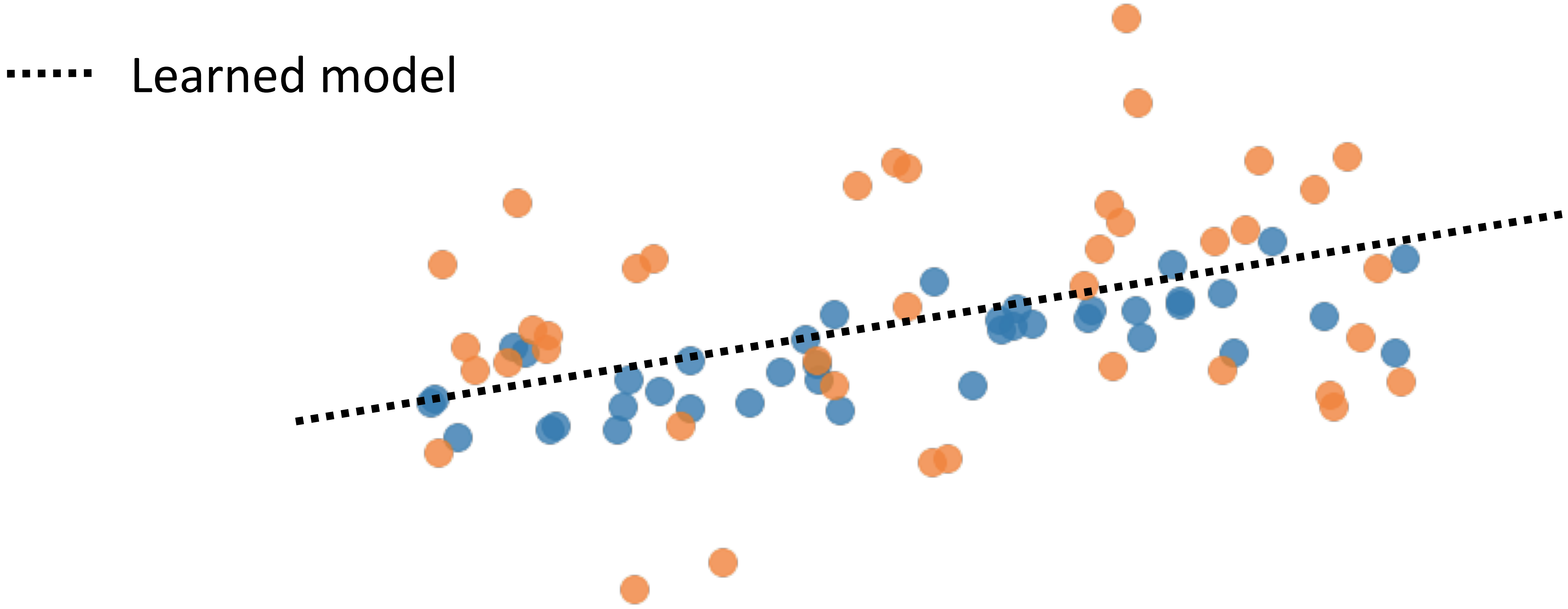


Error from **bias** can be solved by **changing the model class.**

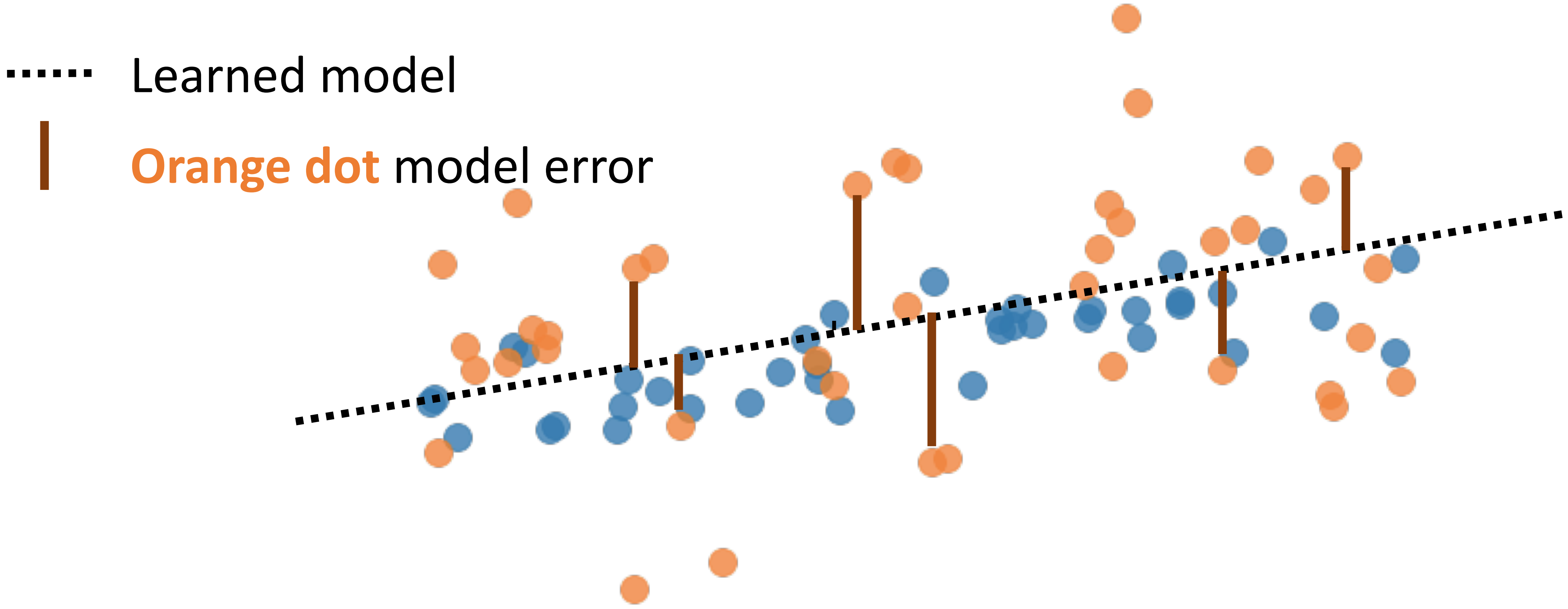
Why might my classifier be unfair?



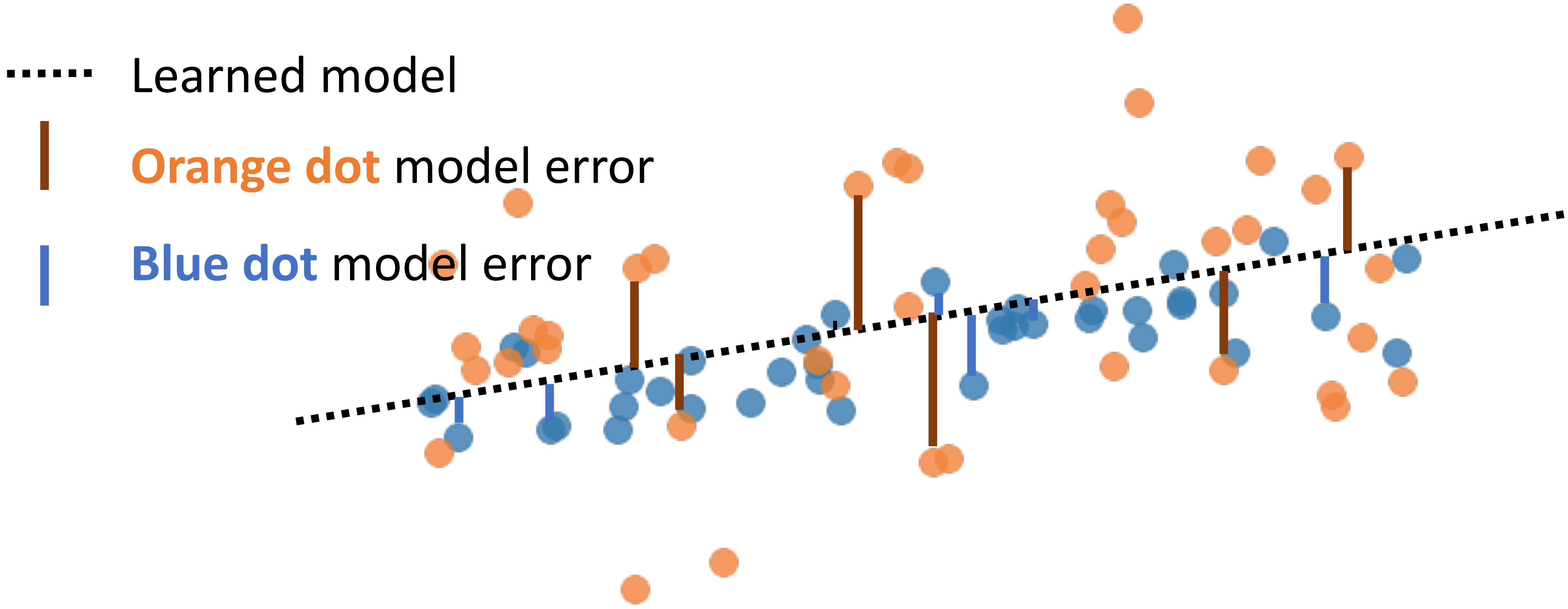
Why might my classifier be unfair?



Why might my classifier be unfair?

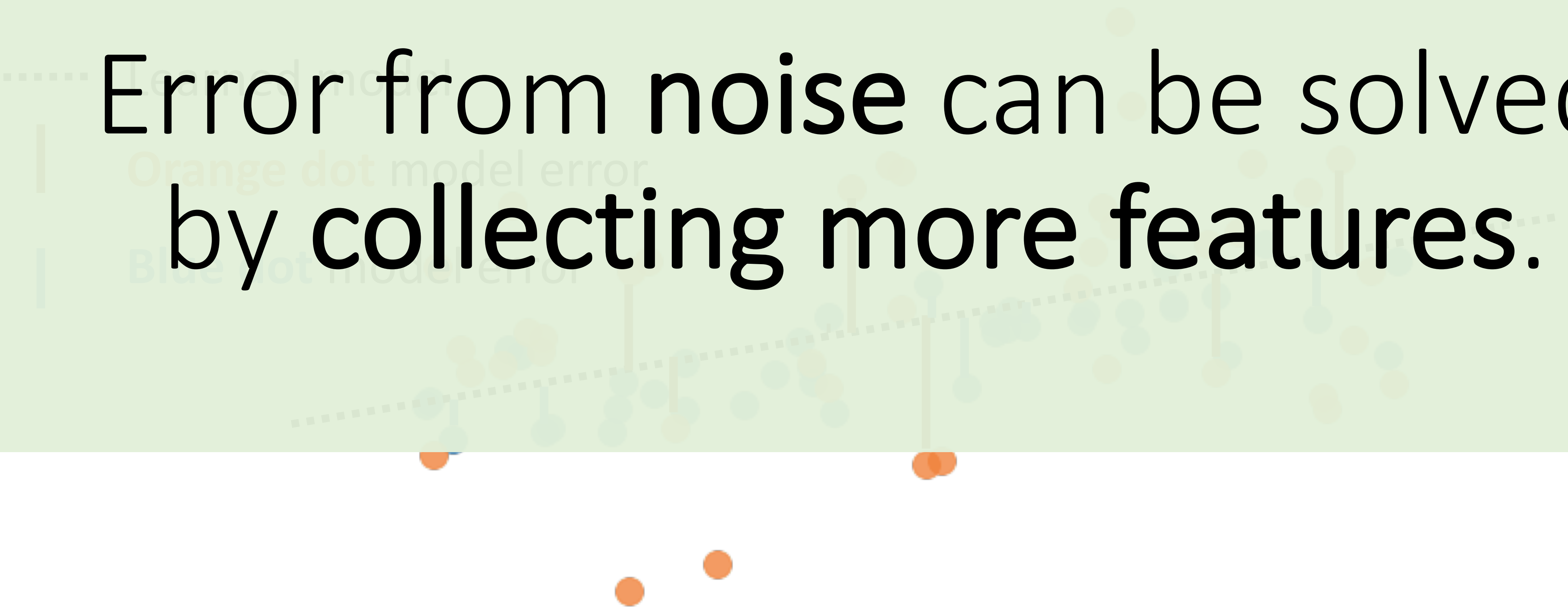


Why might my classifier be unfair?



Why might my classifier be unfair?

Error from **noise** can be solved by **collecting more features**.



How do we define fairness?

How do we define fairness?

We define fairness in the **context of loss** like false positive rate, false negative rate, etc.

For example, zero-one loss for data D and prediction $\hat{Y}$:

$$\gamma_a\left(\hat{Y}, Y, D\right) := P_D\left(\hat{Y} \neq Y \;\middle|\; A = a\right)$$

How do we define fairness?

We define fairness in the **context of loss** like false positive rate, false negative rate, etc.

For example, zero-one loss for data D and prediction $\hat{Y}$:

$$\gamma_a\left(\hat{Y}, Y, D\right) := P_D\left(\hat{Y} \neq Y \;\middle|\; A = a\right)$$

We can then formalize **unfairness as group differences.**

$$\bar{\Gamma}\left(\hat{Y}\right) := |\,\gamma_1 - \gamma_0|$$

We rely on accurate Y labels and focus on algorithmic error.

Why might my classifier be unfair?

Theorem 1: For error over group *a* given predictor $\hat{Y}$:

$$\bar{\gamma}_a(\hat{Y}) = \bar{B}_a(\hat{Y}) + \bar{V}_a(\hat{Y}) + \bar{N}_a$$

Note that $\overline{N_a}$ indicates the expectation of N_a over *X* and data *D*.

Why might my classifier be unfair?

Theorem 1: For error over group *a* given predictor $\hat{Y}$:

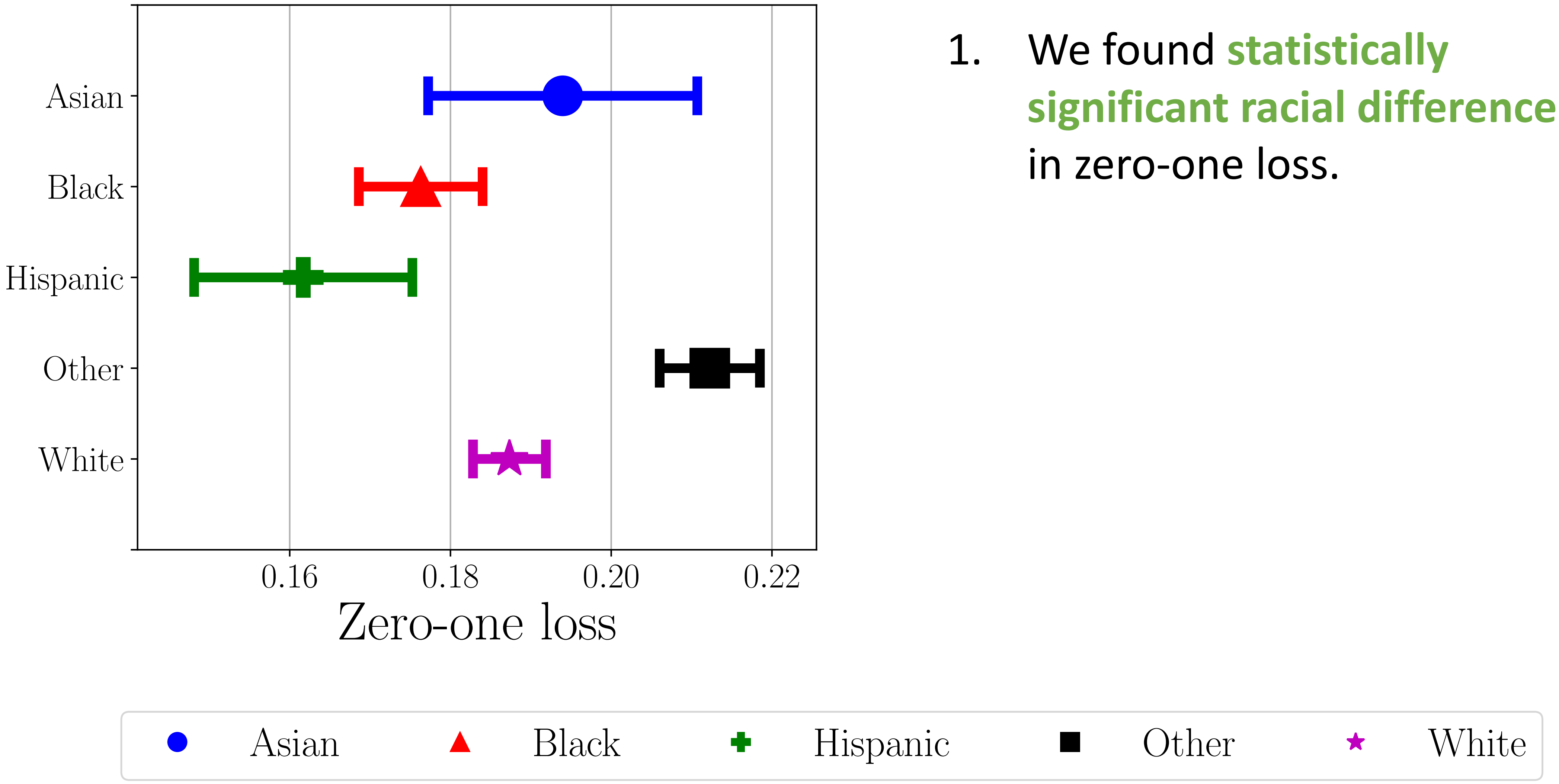
$$\bar{\gamma}_a(\hat{Y}) = \bar{B}_a(\hat{Y}) + \bar{V}_a(\hat{Y}) + \bar{N}_a$$

Note that $\overline{N_a}$ indicates the expectation of N_a over *X* and data *D*.

Accordingly, the expected discrimination level $\bar{\Gamma} := |\overline{\gamma_1} - \bar{\gamma}_0|$ can be decomposed into differences in bias, differences in variance, and differences in noise.

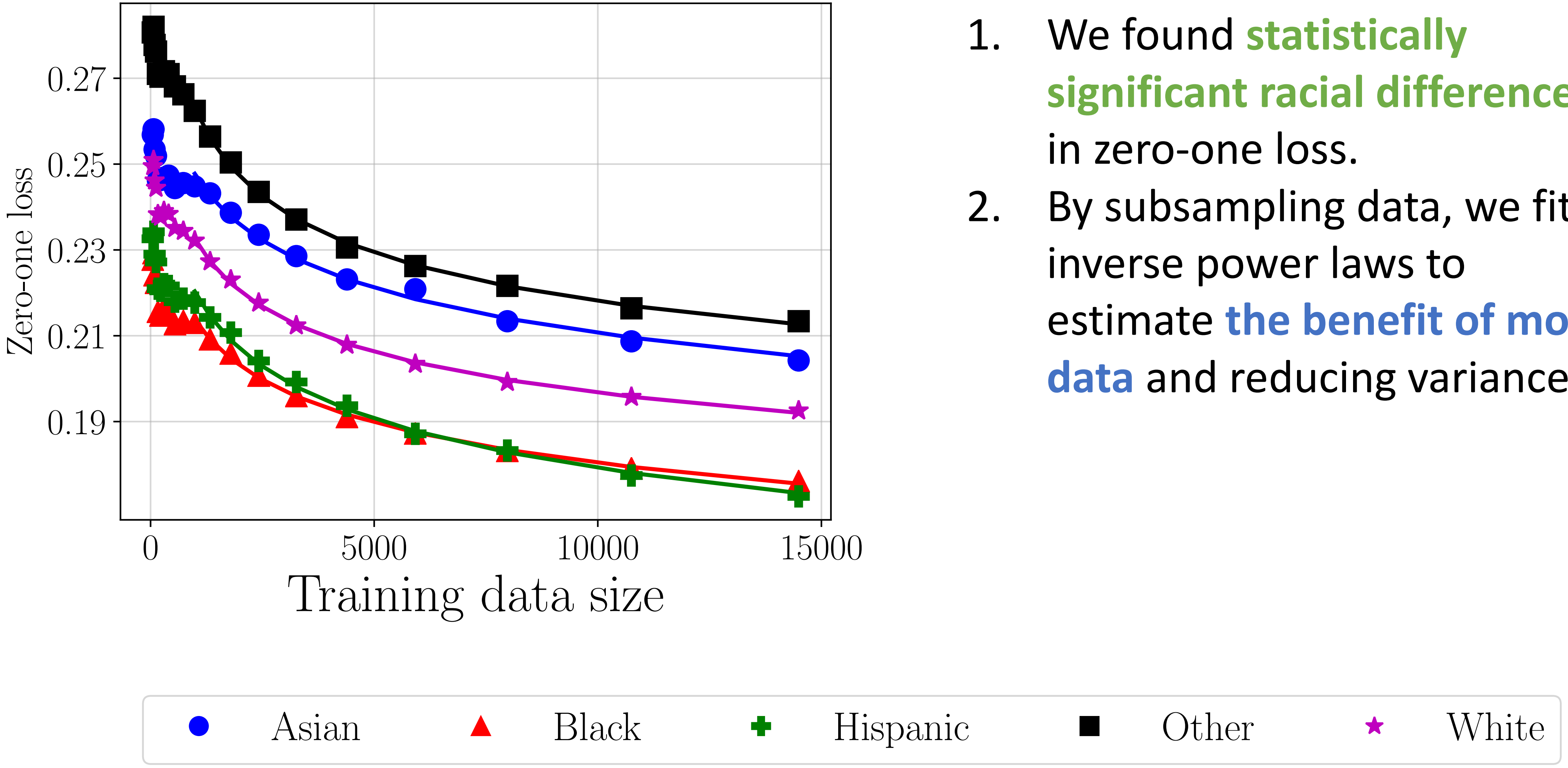
$$\bar{\Gamma} = |(\bar{B}_1 - \bar{B}_0) + (\bar{V}_1 - \bar{V}_0) + (\bar{N}_1 - \bar{N}_0)|$$

Mortality prediction from MIMIC-III clinical notes



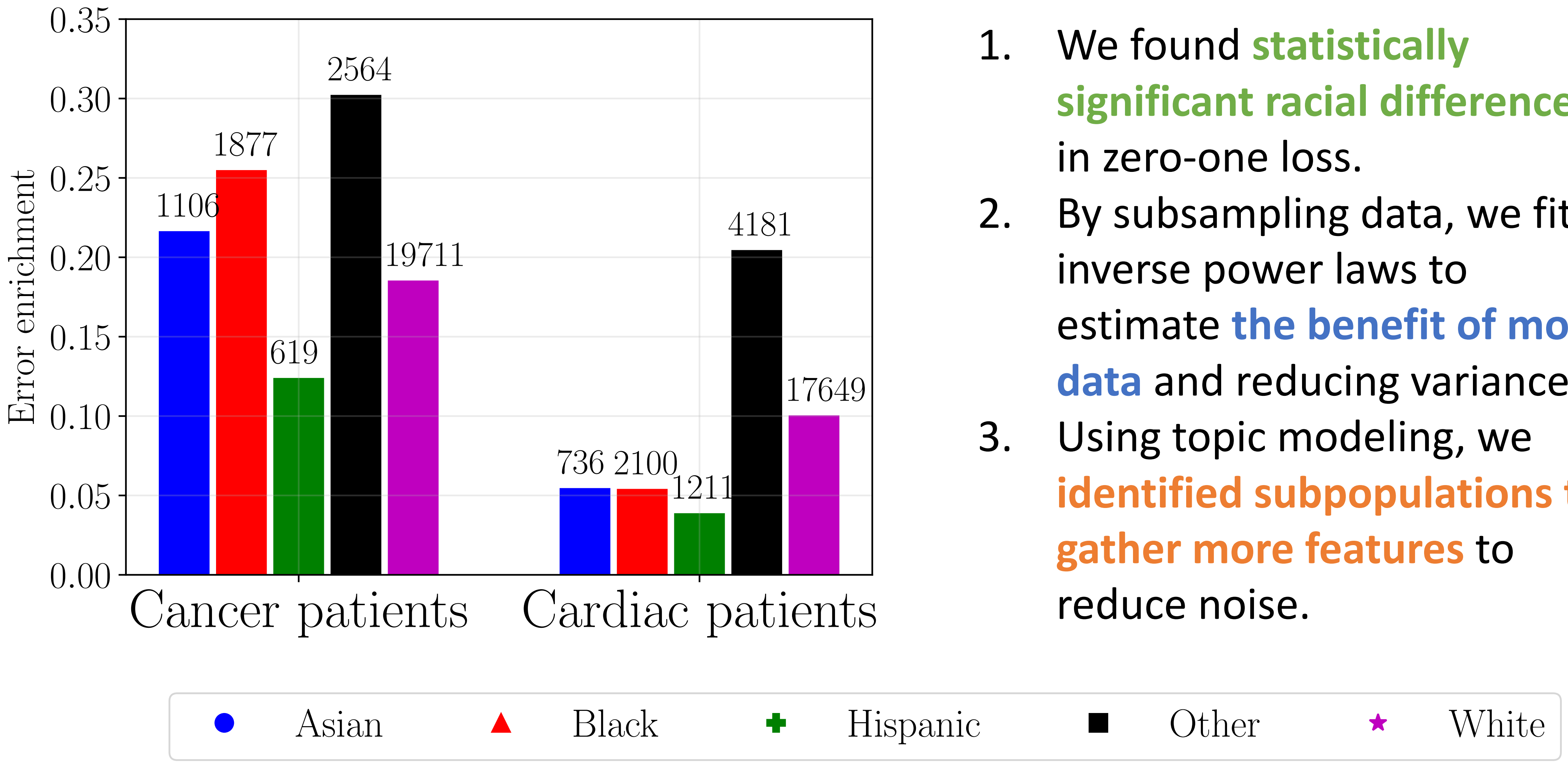
1. We found **statistically significant racial differences** in zero-one loss.

Mortality prediction from MIMIC-III clinical notes



1. We found **statistically significant racial differences** in zero-one loss.
2. By subsampling data, we fit inverse power laws to estimate **the benefit of more data** and reducing variance.

Mortality prediction from MIMIC-III clinical notes



1. We found **statistically significant racial differences** in zero-one loss.
2. By subsampling data, we fit inverse power laws to estimate **the benefit of more data** and reducing variance.
3. Using topic modeling, we **identified subpopulations to gather more features** to reduce noise.

Where do we go from here?

1. For **accurate and fair models** deployed in real world applications, both the data and model should be considered.

Where do we go from here?

1. For **accurate and fair models** deployed in real world applications, both the data and model should be considered.

2. Using **easily implemented fairness checks**, we hope others will check their algorithms for bias, variance, and noise--which will guide further efforts to reduce unfairness.

Read our paper:
https://arxiv.org/abs/1805.12002

Come find us at NeurIPS:

- **Spotlight talk:** Thurs 12/6 10:20am – 10:25am @ 220 CD.
- **Poster #120:** Thurs 12/6 10:45am – 12:45pm @ 210 & 230.

Why Is My Classifier Discriminatory?

Irene Chen
MIT
iychen@mit.edu

Fredrik D. Johansson
MIT
fredrikj@mit.edu

David Sontag
MIT
dsontag@csail.mit.edu

Abstract

Recent attempts to achieve fairness in predictive models focus on the balance between fairness and accuracy. In sensitive applications such as healthcare or criminal justice, this trade-off is often undesirable as any increase in prediction error could have devastating consequences. In this work, we argue that the fairness of predictions should be evaluated in context of the data, and that unfairness induced by inadequate samples sizes or unmeasured predictive variables should be addressed through data collection, rather than by constraining the model. We decompose cost-based metrics of discrimination into bias, variance, and noise, and propose actions aimed at estimating and reducing each term. Finally, we perform case-studies on prediction of income, mortality, and review ratings, confirming the value of this analysis. We find that data collection is often a means to reduce discrimination without sacrificing accuracy.

1 Introduction

As machine learning algorithms increasingly affect decision making in society, many have raised concerns about the fairness and biases of these algorithms, especially in applications to healthcare or criminal justice, where human lives are at stake (Angwin et al., 2016; Barocas & Selbst, 2016). It is often hoped that the use of automatic decision support systems trained on observational data will remove human bias and improve accuracy. However, factors such as data quality and model choice may encode unintentional discrimination, resulting in systematic disparate impact.