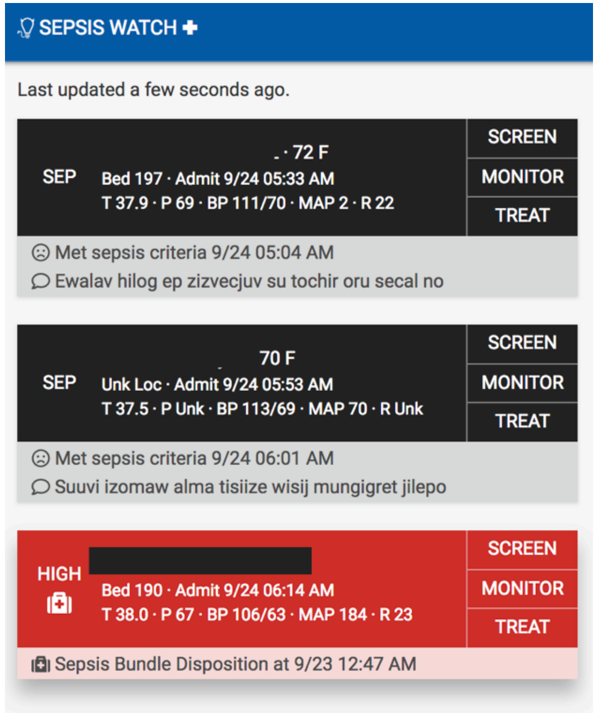
Beyond Bias Audits: Bringing Equity to the Machine Learning Pipeline

Irene Y. Chen
@irenetrampoline

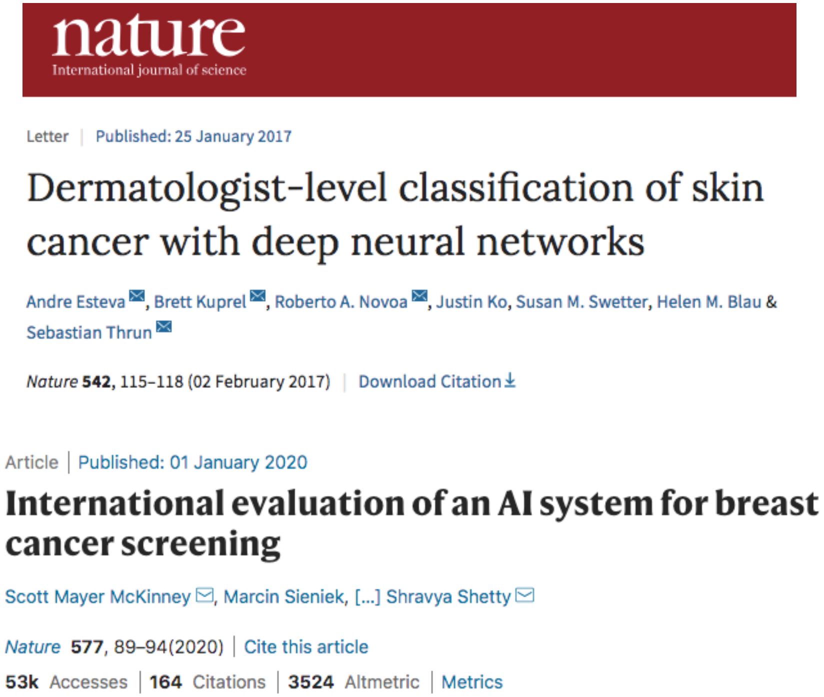
Joint work with David Sontag, Marzyeh Ghassemi, Fredrik D. Johansson, Rahul G. Krishnan, Sherri Rose, Emma Pierson, Shalmali Joshi, Kadija Ferryman, Bharti Khurana, Emily Alsentzer, Hyesun Park, Richard Thomas, Babina Gosangi, Rahul Gujrathi

MIT Clinical ML
www.clinicalml.org

Machine learning in healthcare settings show great promise



Combine multiple sources of clinical data



nature
International journal of science

Letter | Published: 25 January 2017

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

Nature **542**, 115–118 (02 February 2017) | Download Citation

Article | Published: 01 January 2020

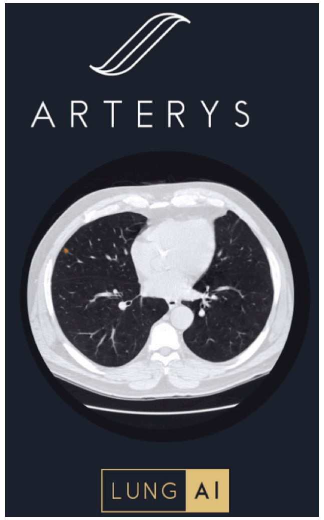
International evaluation of an AI system for breast cancer screening

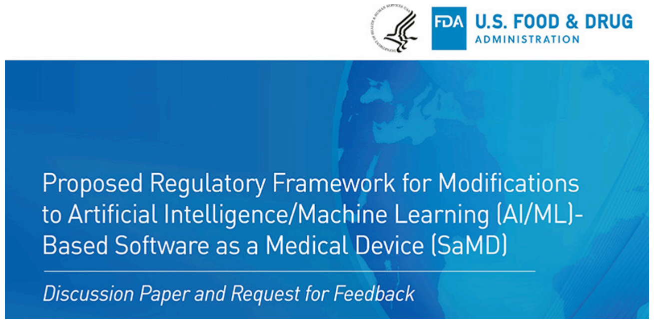
Scott Mayer McKinney, Marcin Sieniek, [...] Shravya Shetty

Nature **577**, 89–94(2020) | Cite this article

53k Accesses | **164** Citations | **3524** Altmetric | Metrics

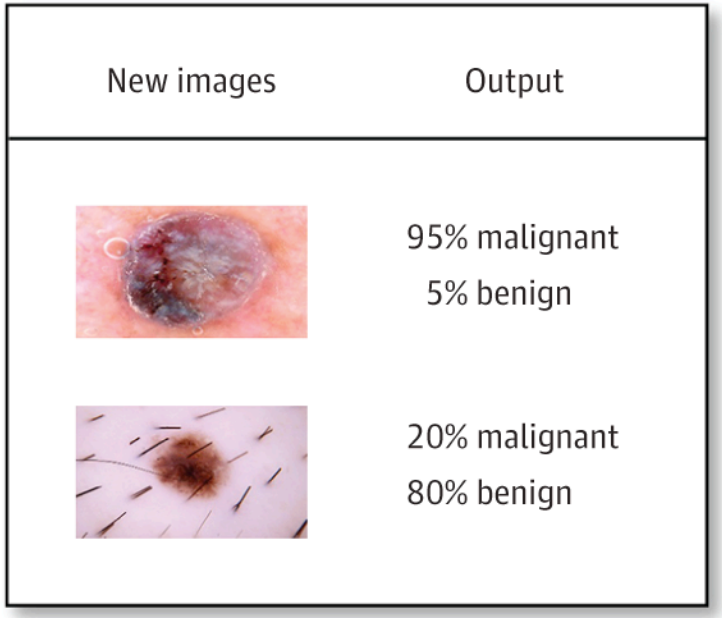
Meet/exceed human performance



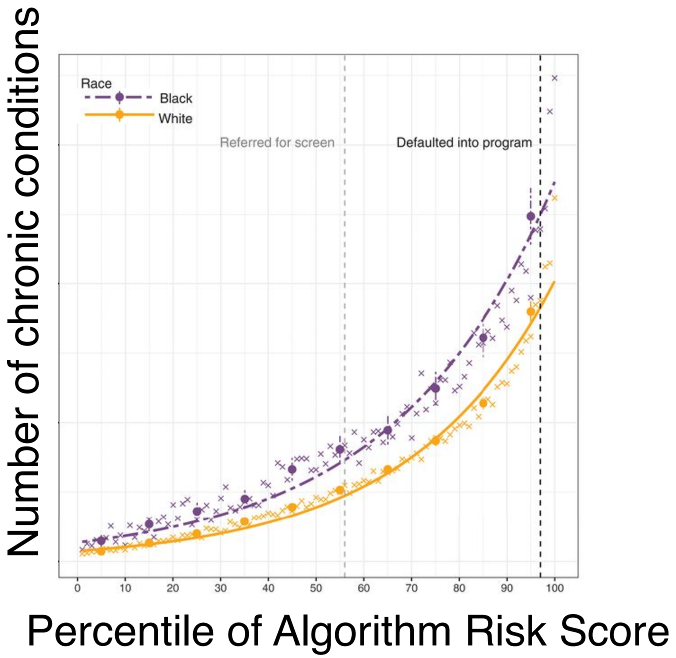


Receive regulatory approval

We are finding evidence of bias through audits



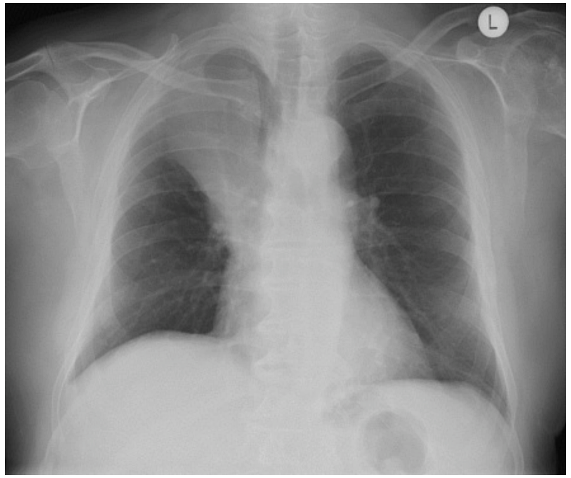
Dermatology algorithms are trained primarily on data from fair-skinned patients

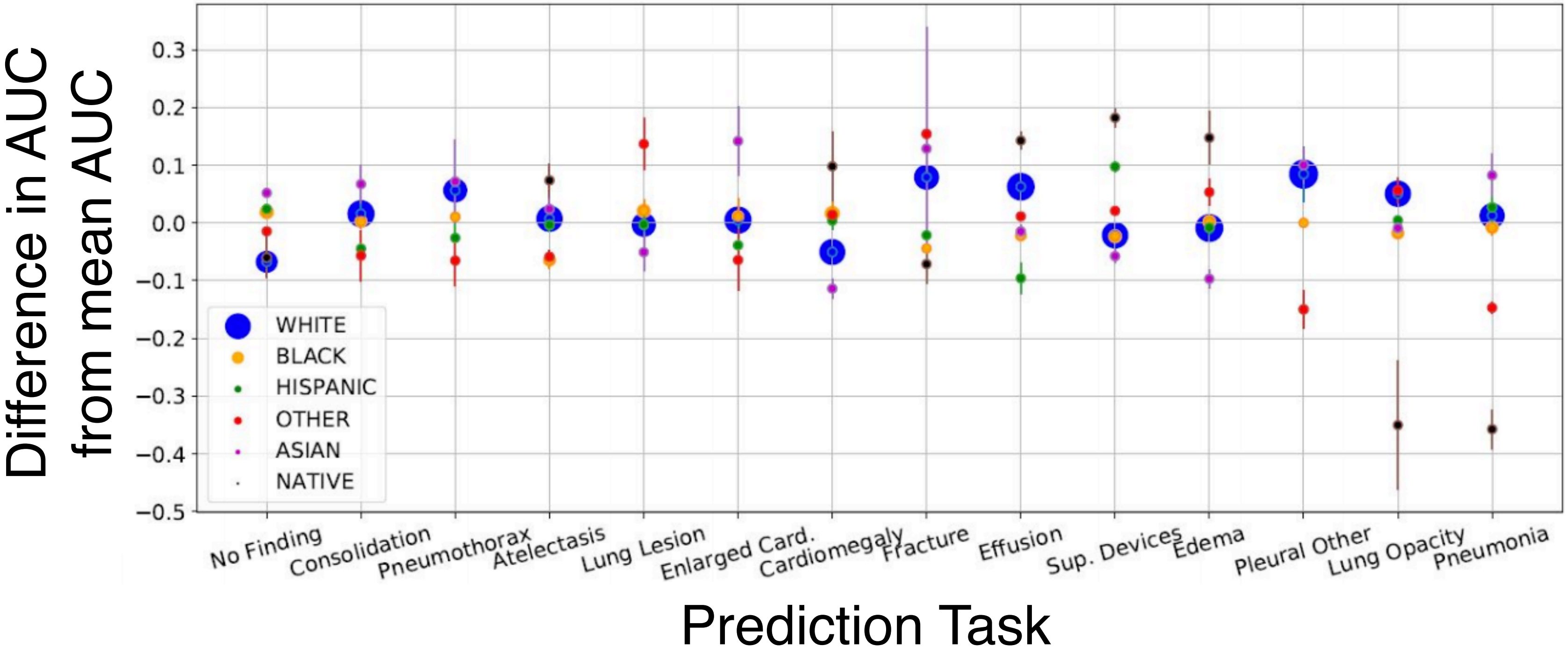


Care management algorithms show racial bias due to training on the “wrong” outcome

[1] Adamson and Smith, “Machine Learning and Health Care Disparities in Dermatology,” *JAMA Dermatology* 2018.
[2] Obermeyer et al, “Dissecting racial bias in algorithm used to manage the health of populations“, *Science* 2019.

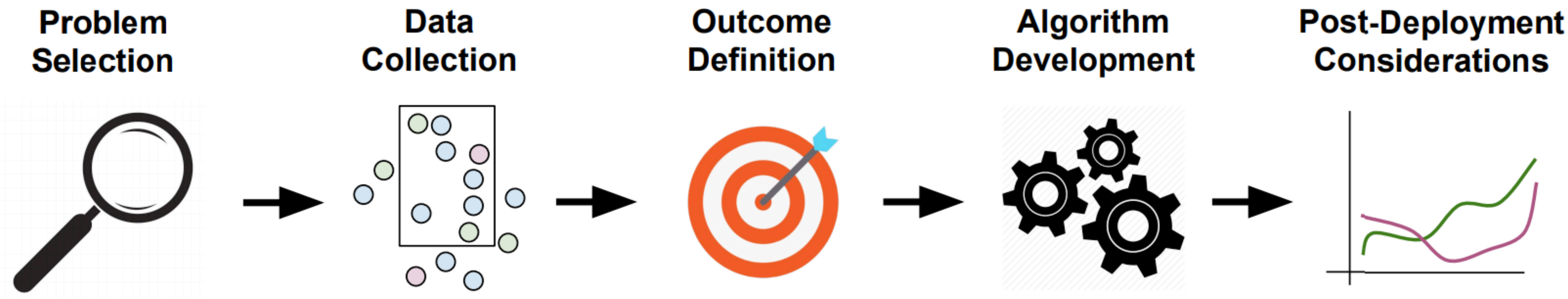
We are finding evidence of bias through audits





[1] Seyyed-Kalantari, Liu, McDermott, **Chen**, and Ghassemi. "CheXclusion: Fairness gaps in deep chest X-ray classifiers", PSB 2021.

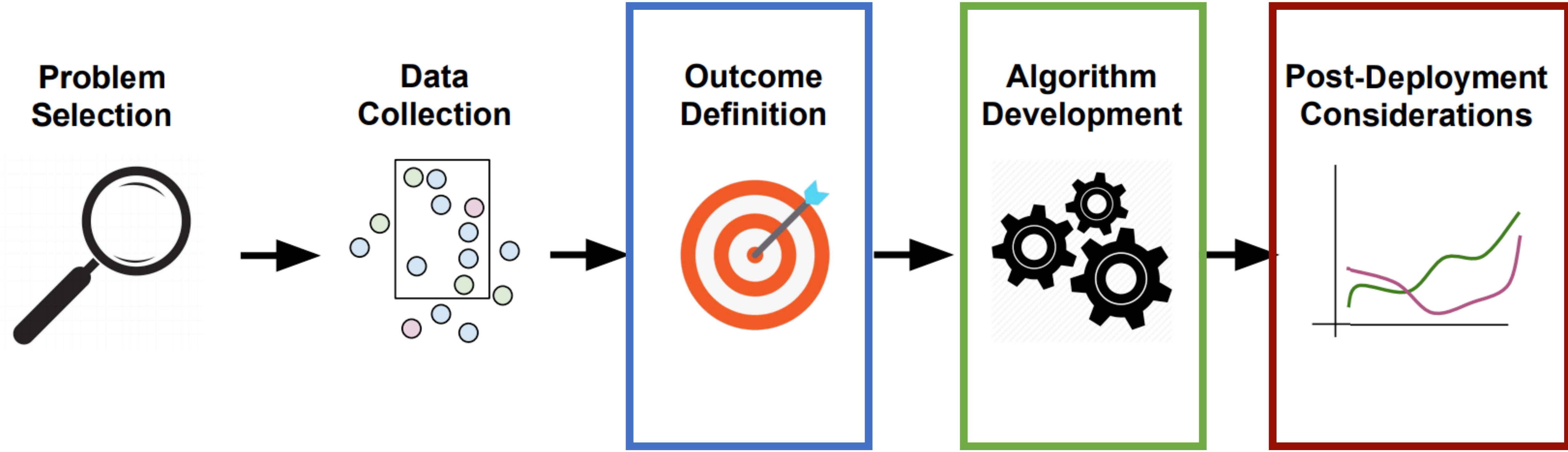
Ethical ML Pipeline



Bias Audits

Chen et al, “Ethical Machine Learning for Health Care,” *Annual Reviews for Biomedical Data Science 2021.*

Ethical ML Pipeline



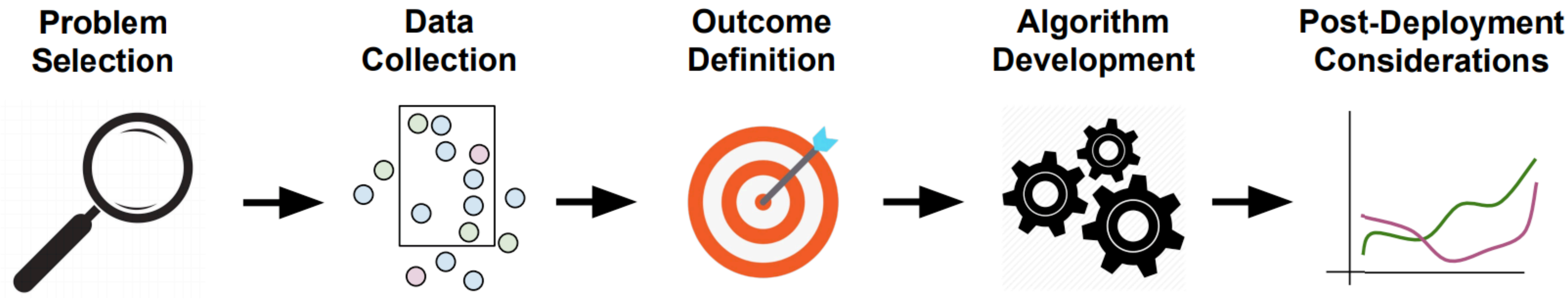
Chen et al, "Ethical Machine Learning for Health Care," *Annual Reviews for Biomedical Data Science 2021.*

We can create machine learning for equitable healthcare by:

1. Diagnosing sources of unfairness
2. Inferring access to care
3. Exploring appropriate labels for sensitive conditions

Diagnosing sources of unfairness in supervised learning

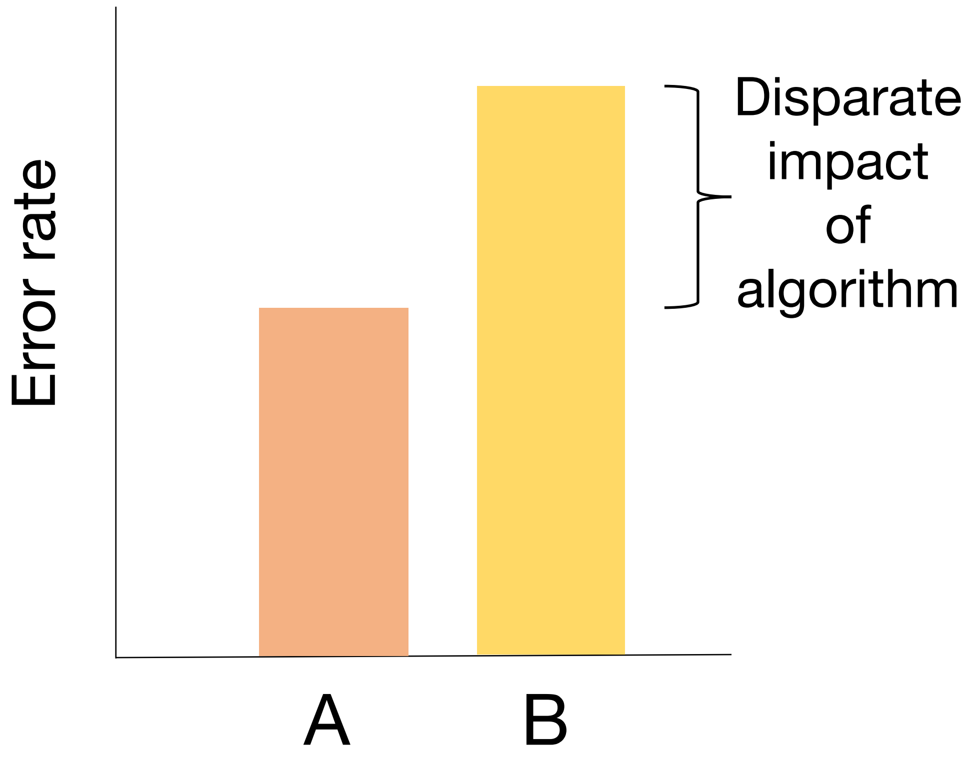
Ethical ML Pipeline



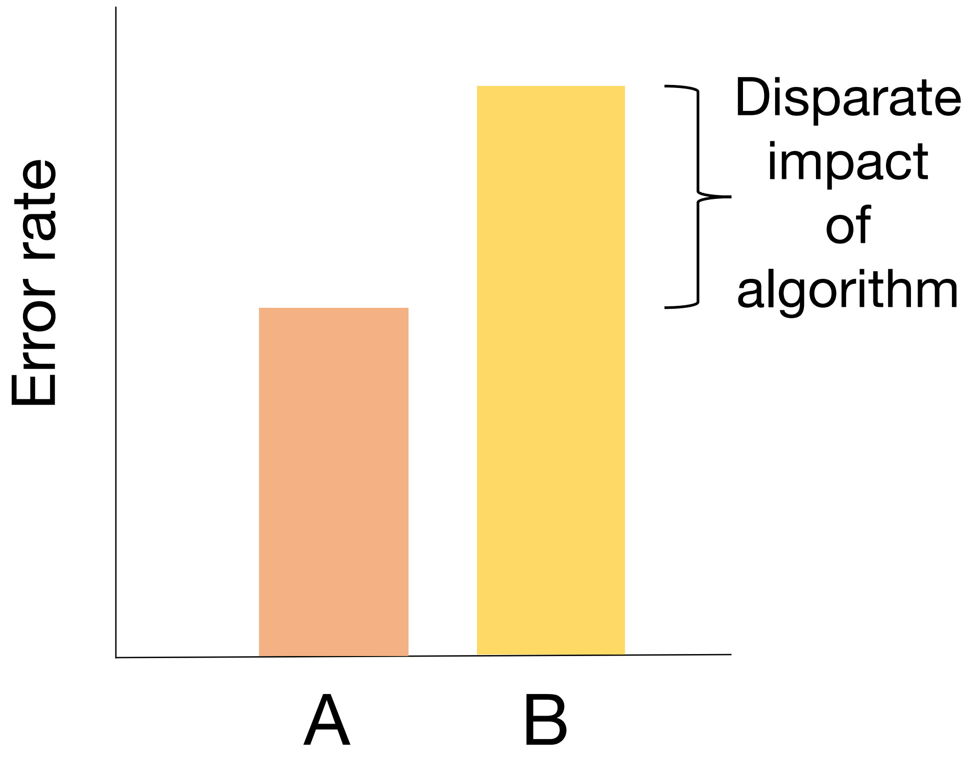
Diagnosing Sources of Unfairness in Supervised Learning

Chen et al, “Why is My Classifier Discriminatory?” NeurIPS 2018

Why might my algorithm be unfair?

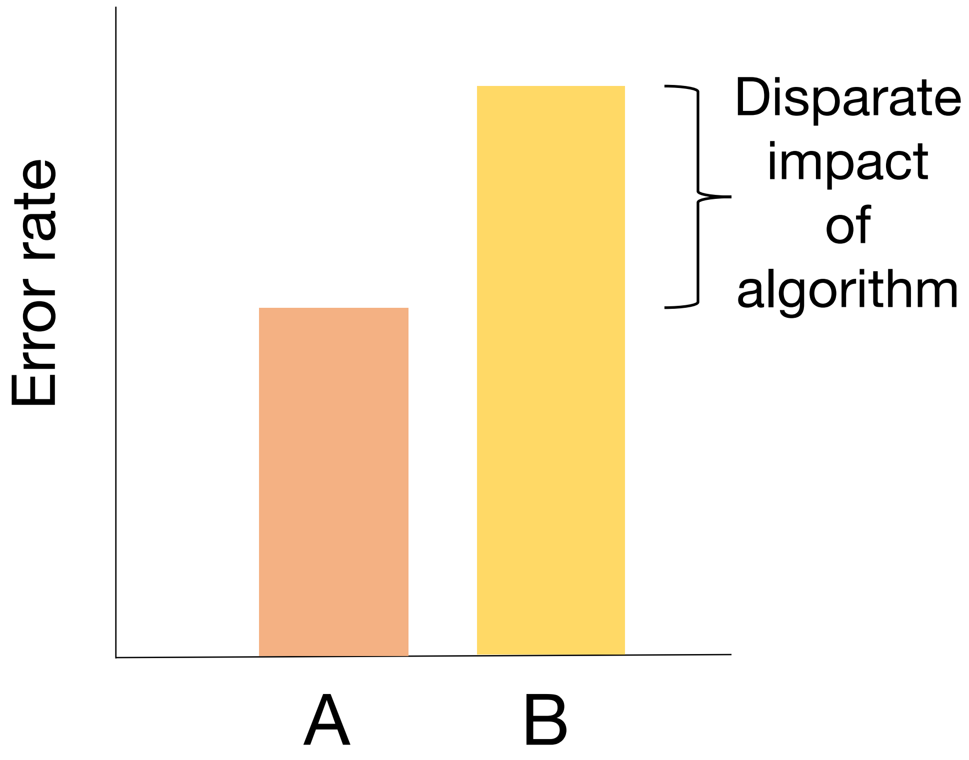


Why might my algorithm be unfair?



1. Group B is much smaller than Group A.
2. Group B has patterns in the data require more complex computational tools.
3. Measurements from Group B are less reliable.

Why might my algorithm be unfair?



1. Group B is much smaller than Group A. **VARIANCE**

2. Group B has patterns in the data require more complex computational tools. **BIAS**

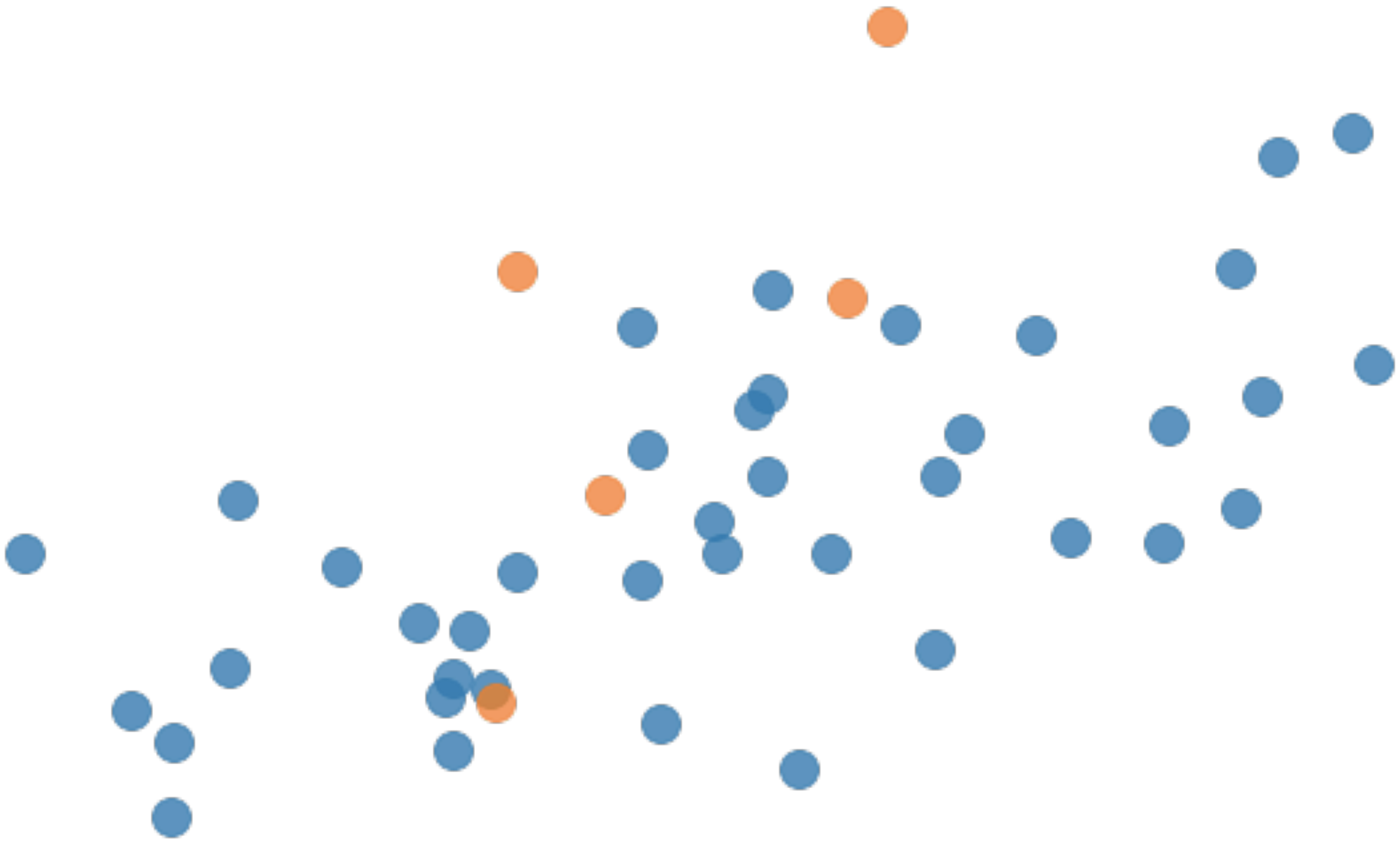
3. Measurements from Group B are less reliable. **NOISE**

Bias, variance, and noise

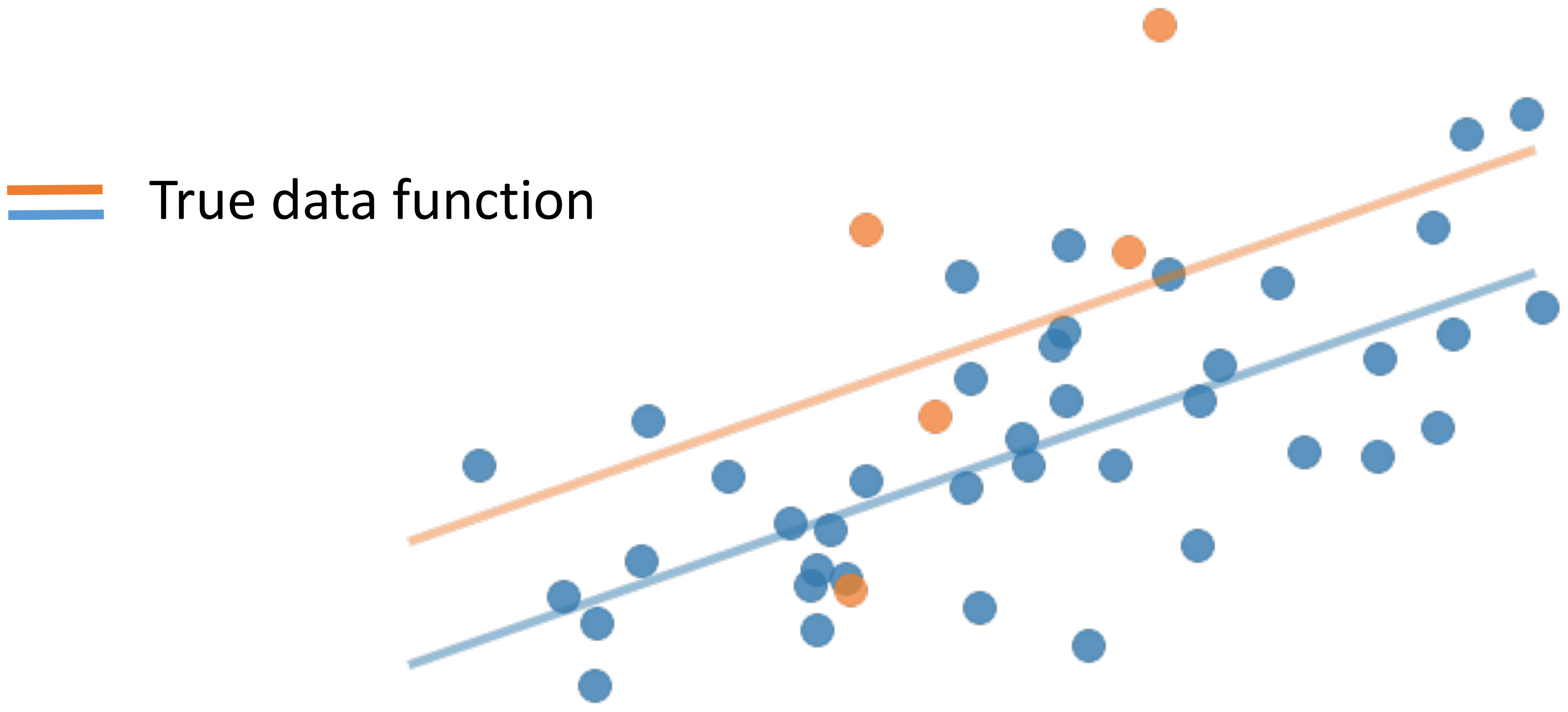
	Description	**How to fix**
Bias	How well model fits data	Change model class
Variance	How much sample size affects accuracy	Increase training data size
Noise	Error independent of model class and sample size	Increase number of features

Why might my algorithm be unfair?

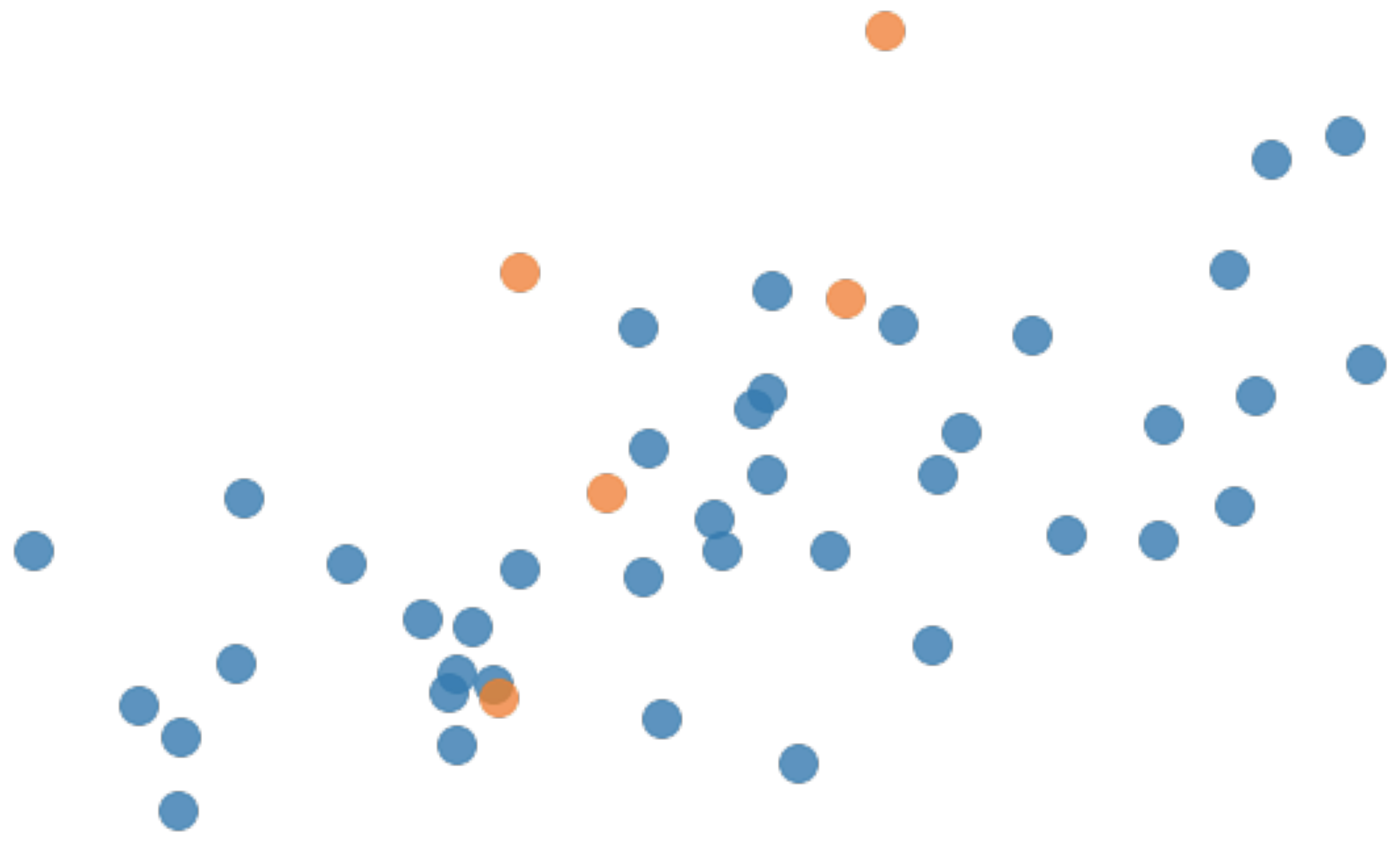
Why might my algorithm be unfair?



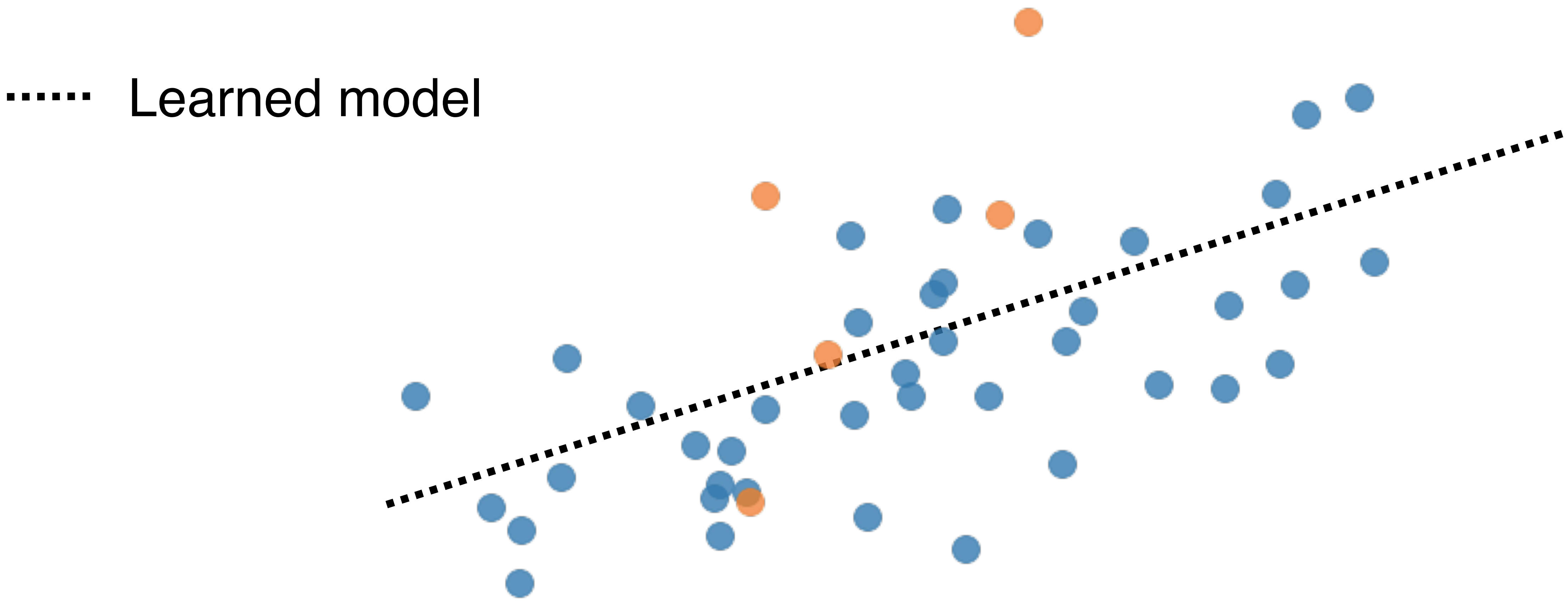
Why might my algorithm be unfair?



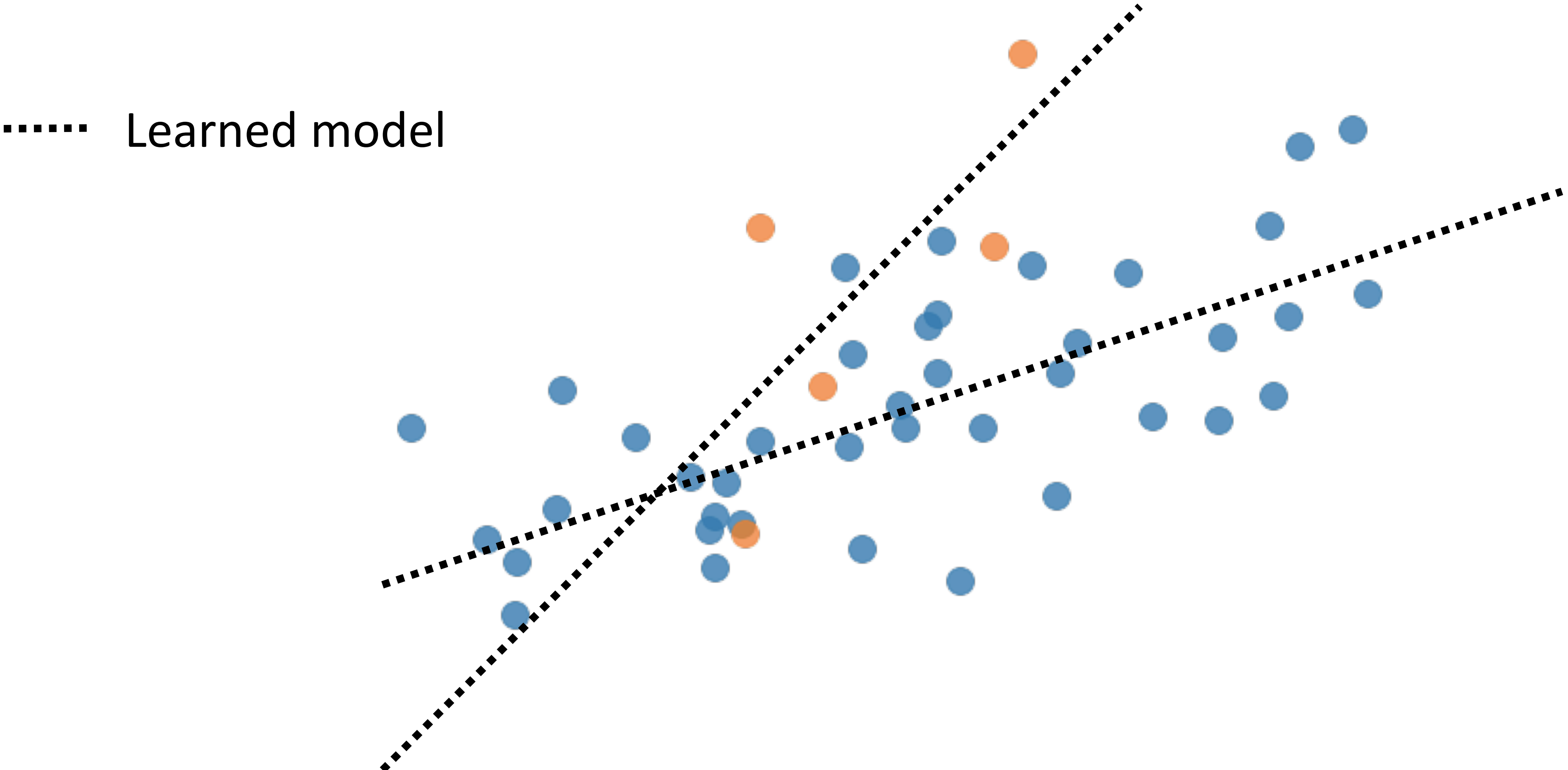
Why might my algorithm be unfair?



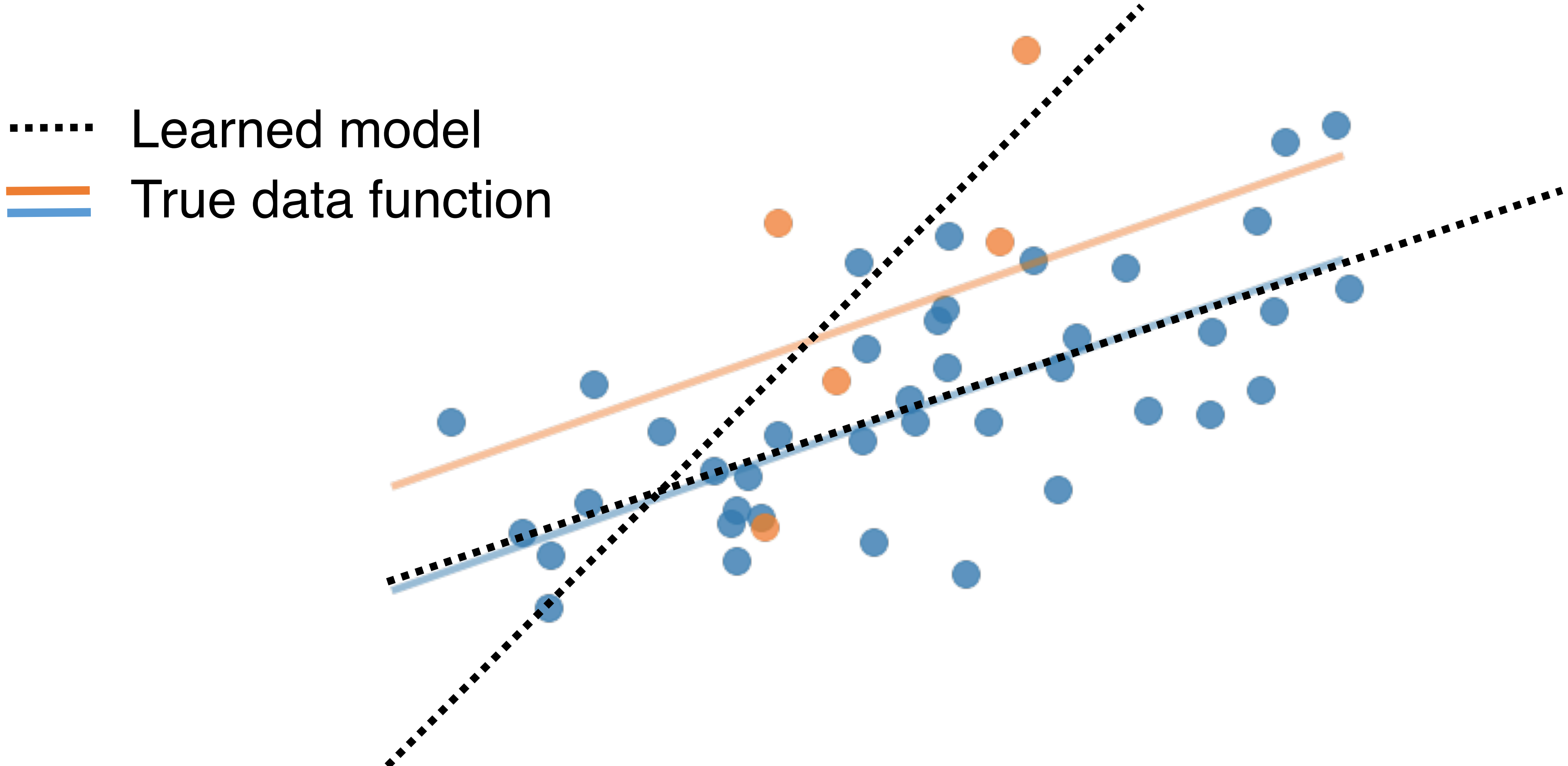
Why might my algorithm be unfair?



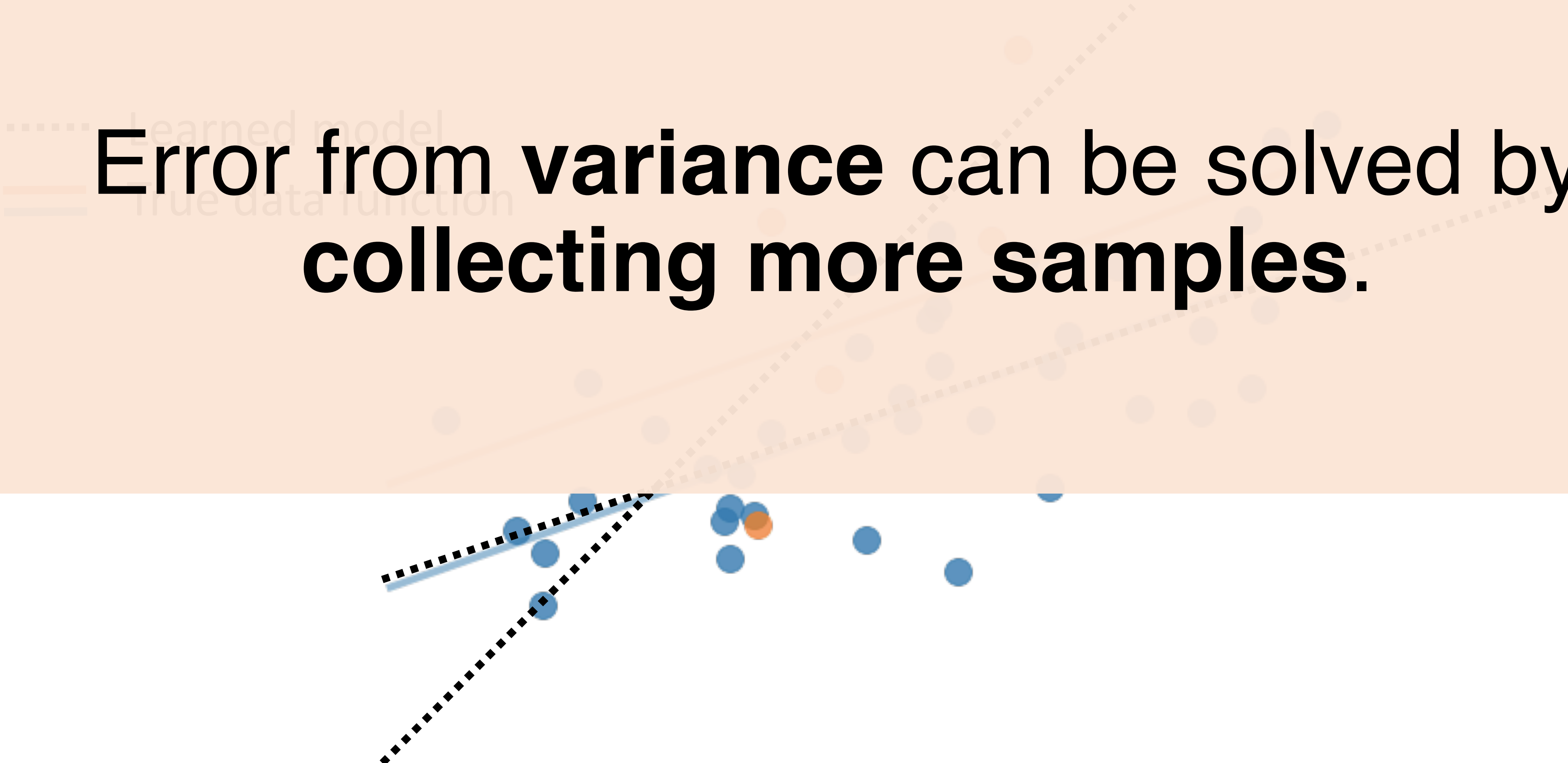
Why might my algorithm be unfair?



Why might my algorithm be unfair?

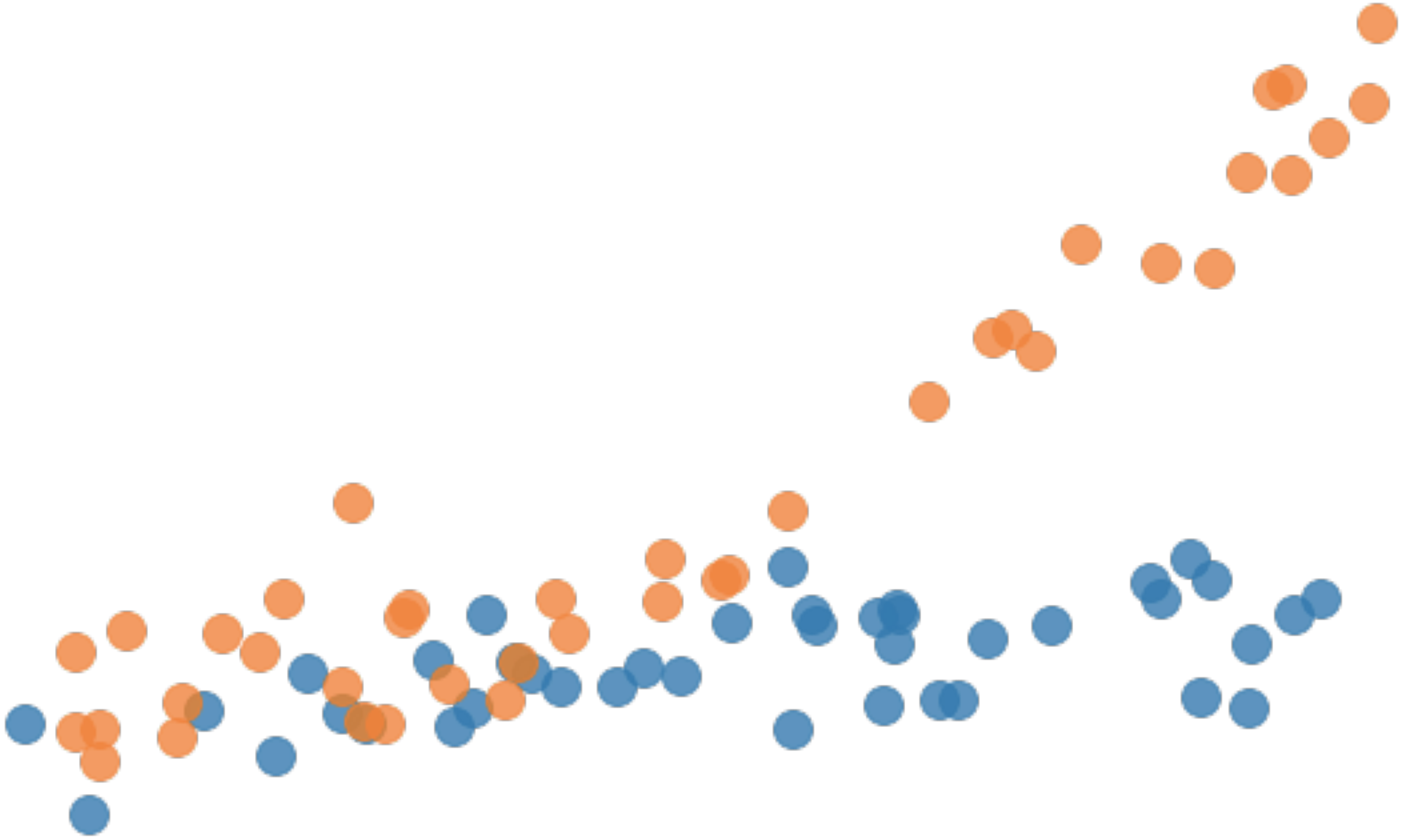


Why might my classifier be unfair?

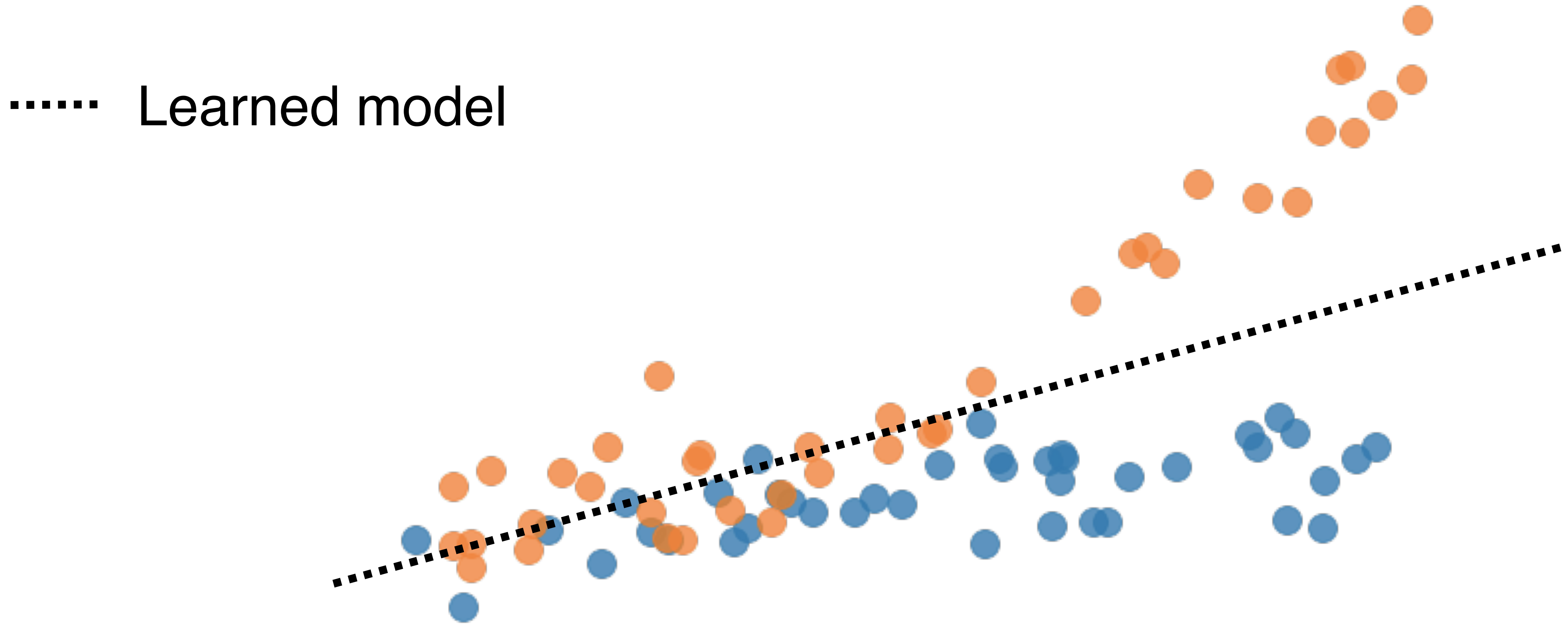


Error from **variance** can be solved by **collecting more samples**.

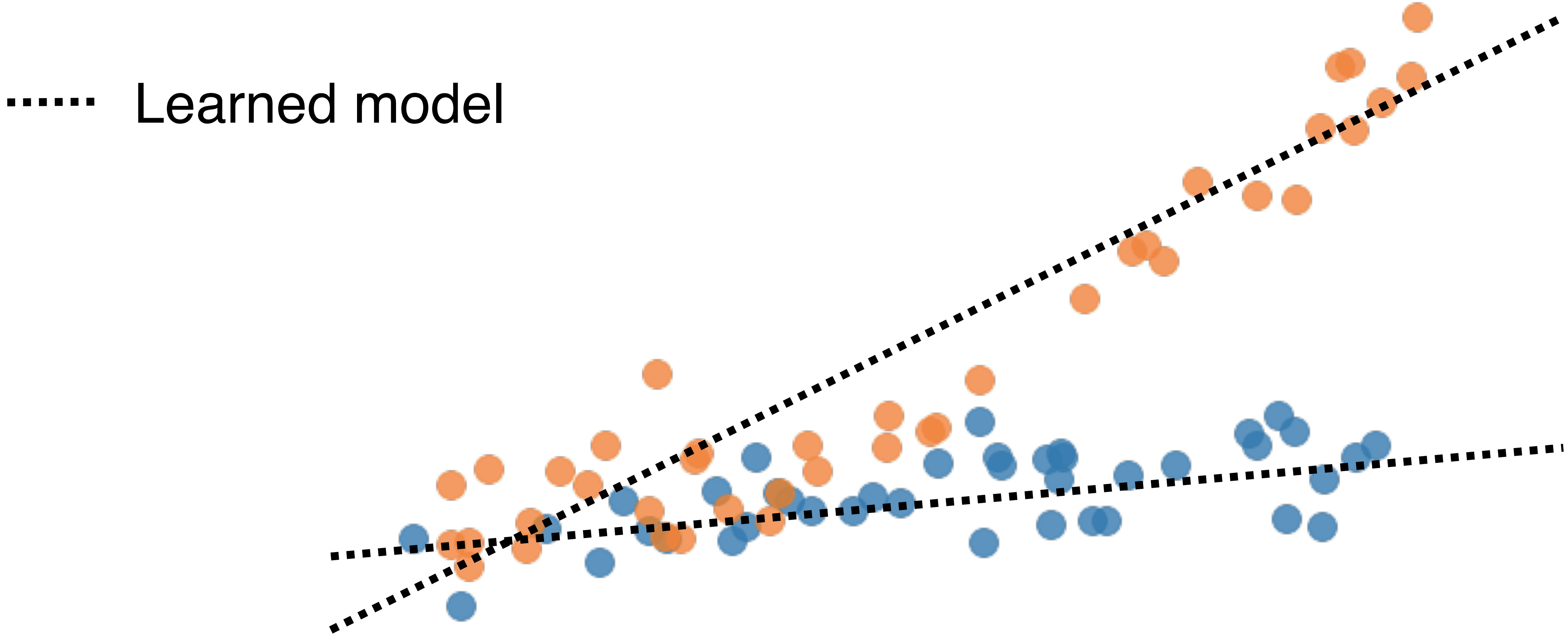
Why might my algorithm be unfair?



Why might my algorithm be unfair?



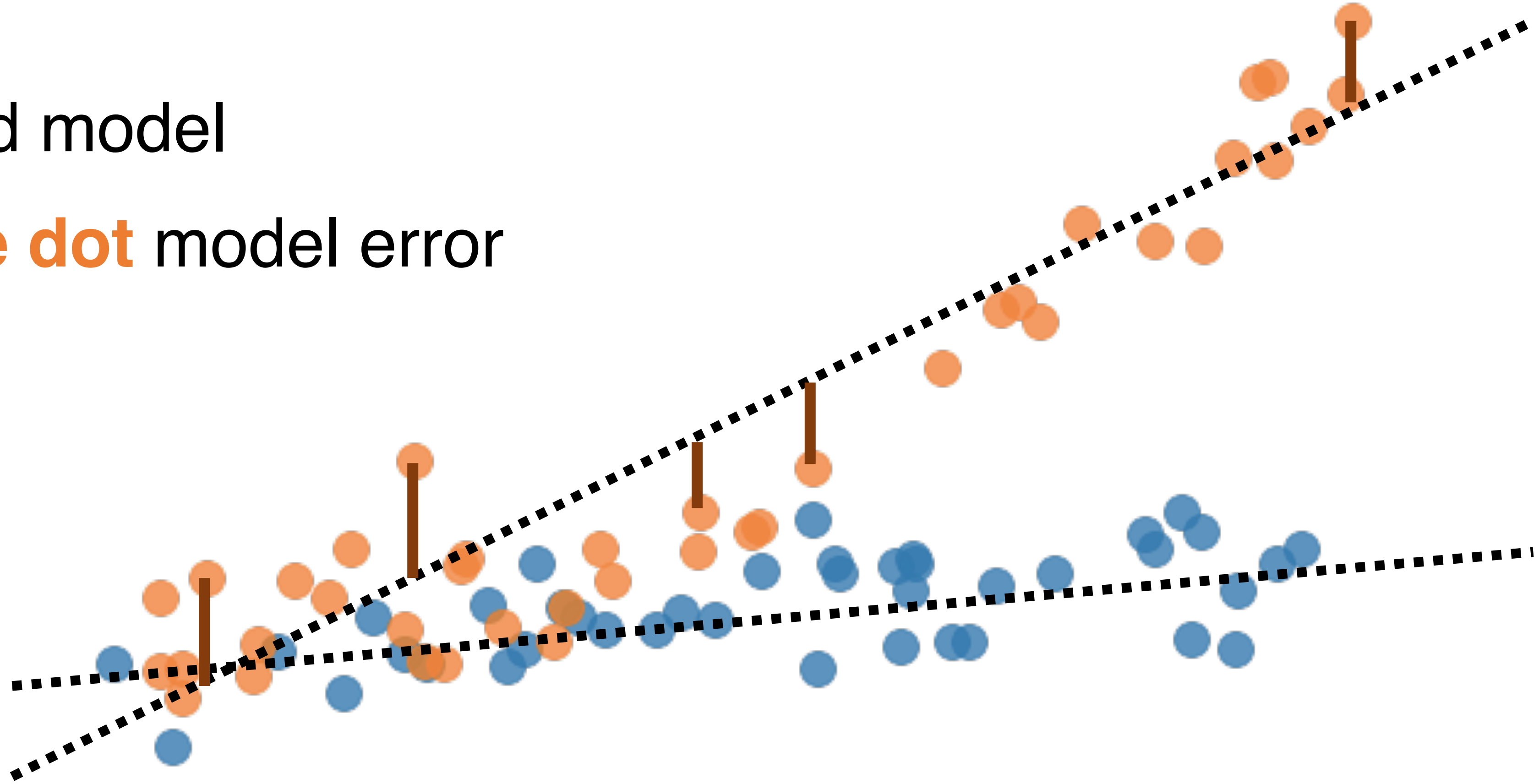
Why might my algorithm be unfair?



Why might my algorithm be unfair?

Learned model

Orange dot model error

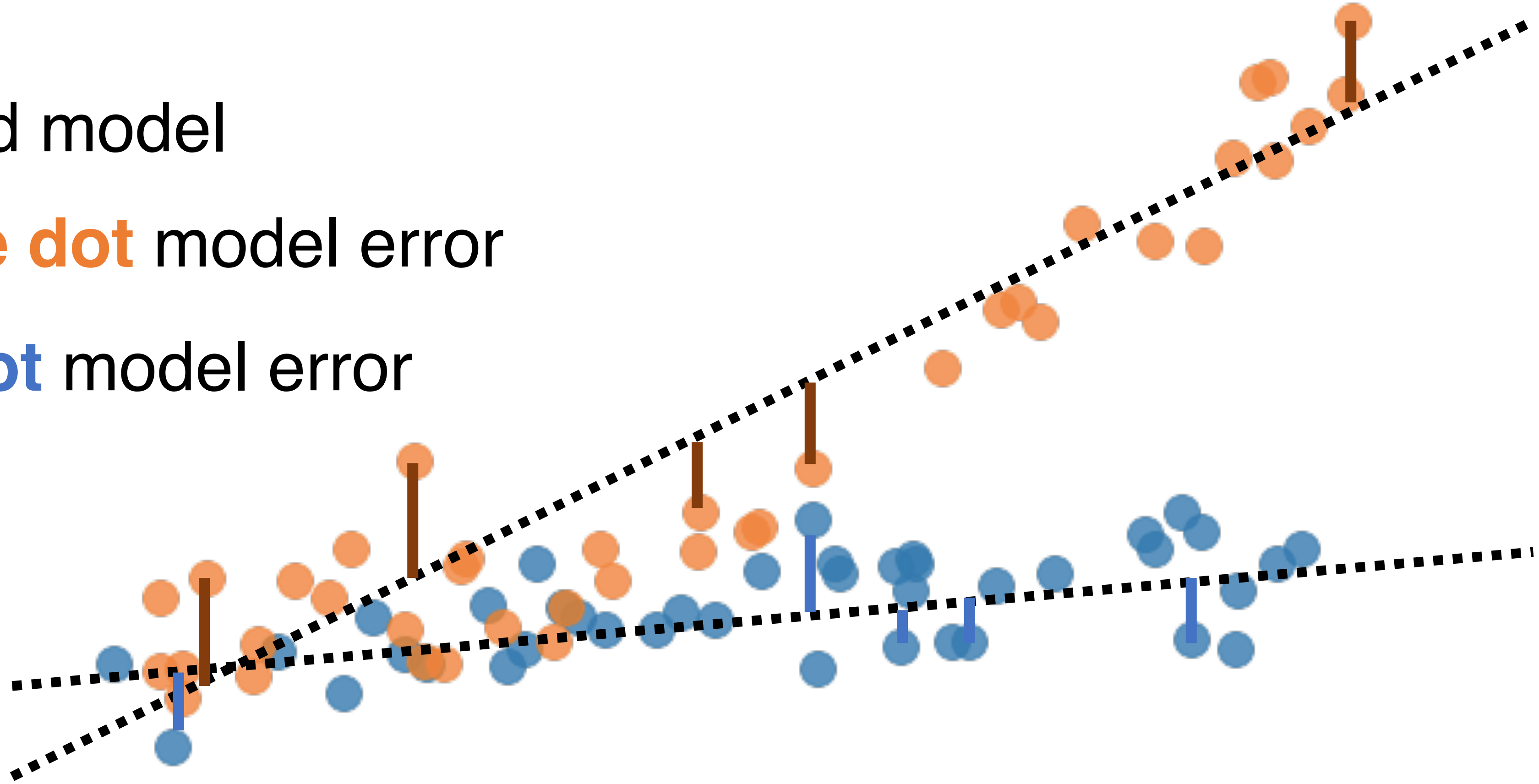


Why might my algorithm be unfair?

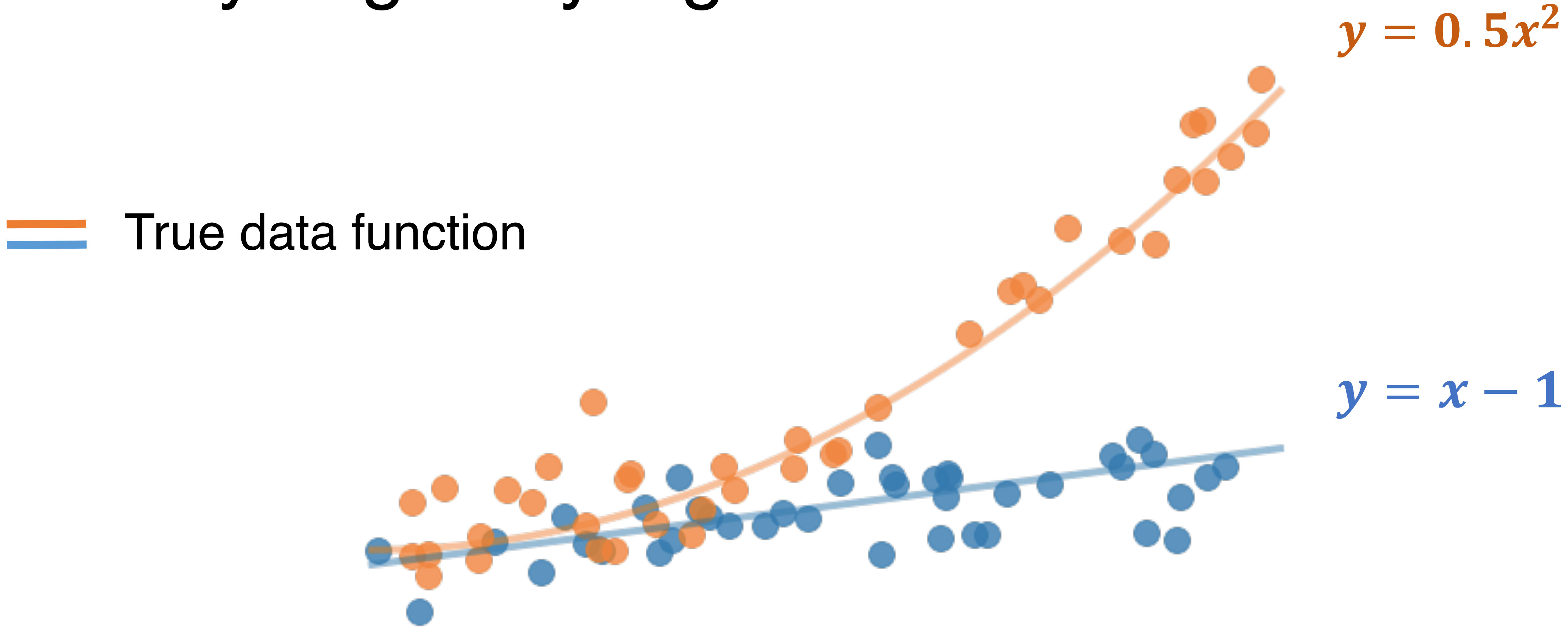
Learned model

Orange dot model error

Blue dot model error



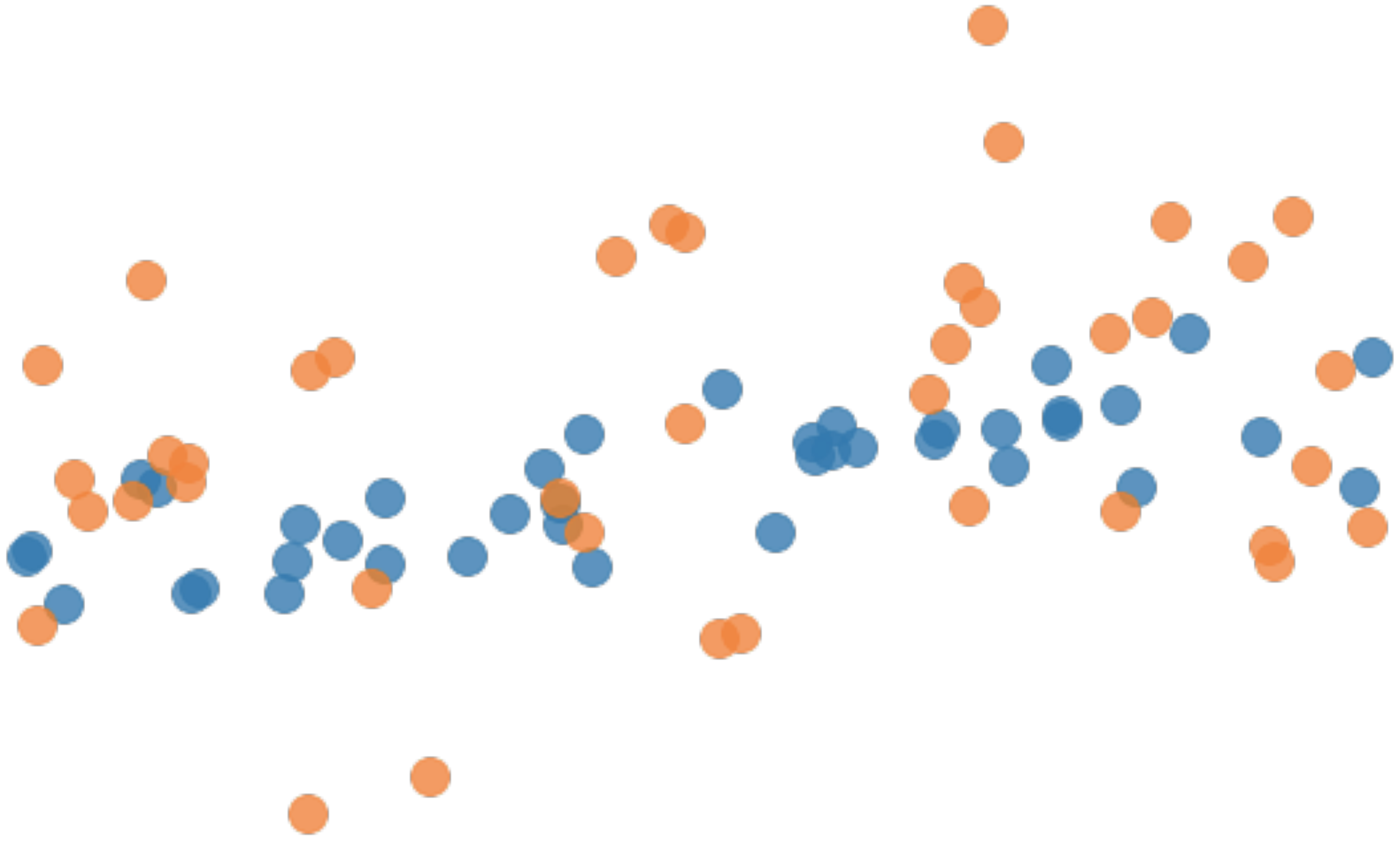
Why might my algorithm be unfair?



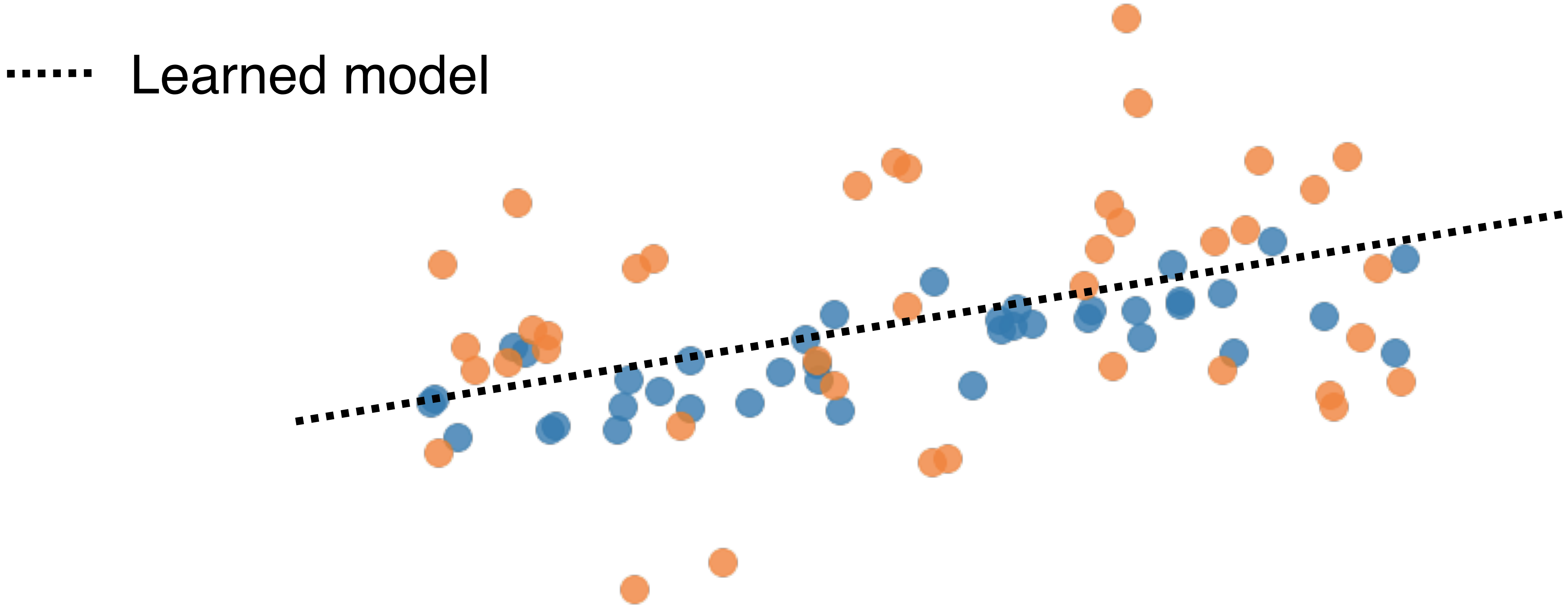
Why might my classifier be unfair?

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

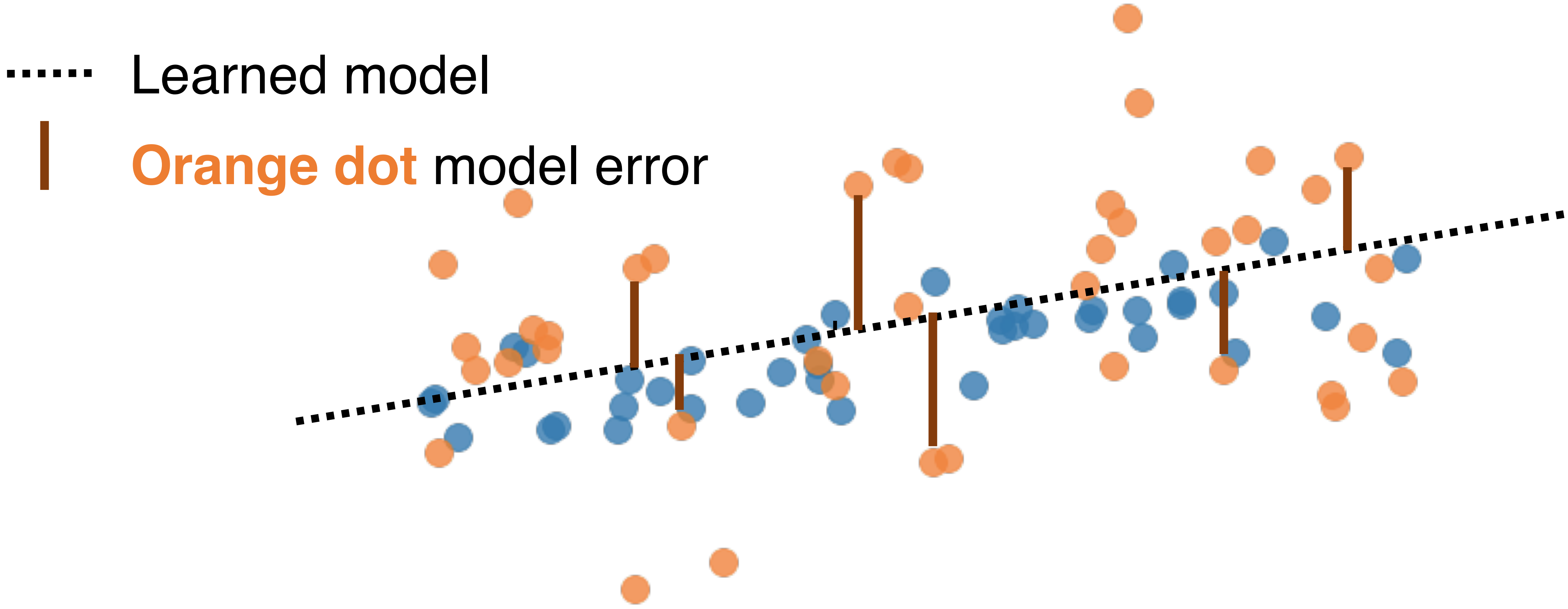
Why might my algorithm be unfair?



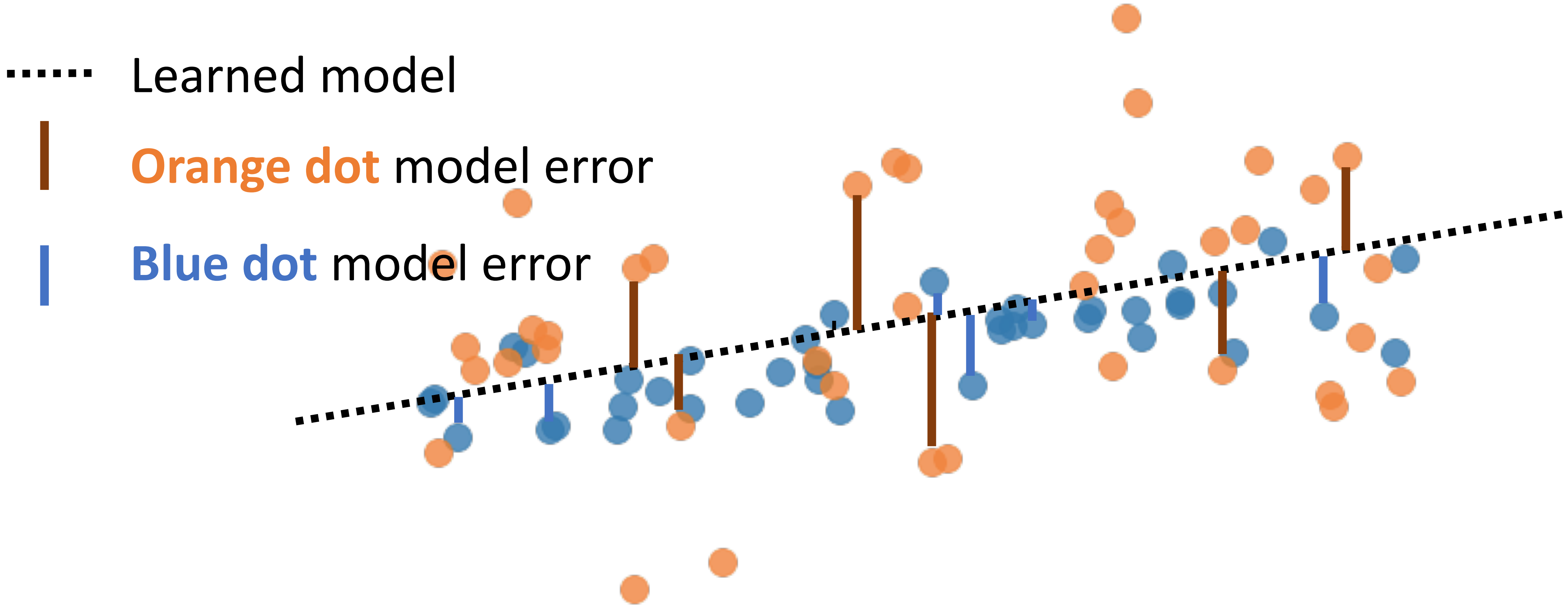
Why might my algorithm be unfair?



Why might my algorithm be unfair?

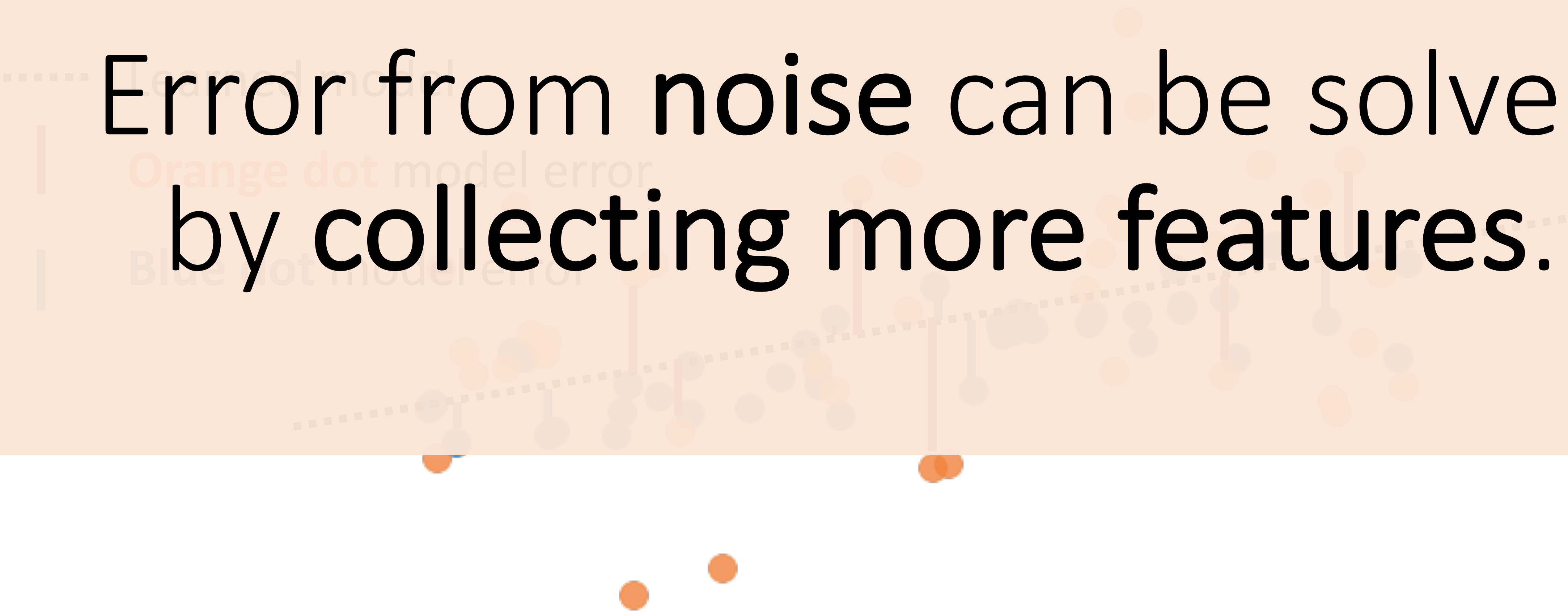


Why might my algorithm be unfair?



Why might my classifier be unfair?

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



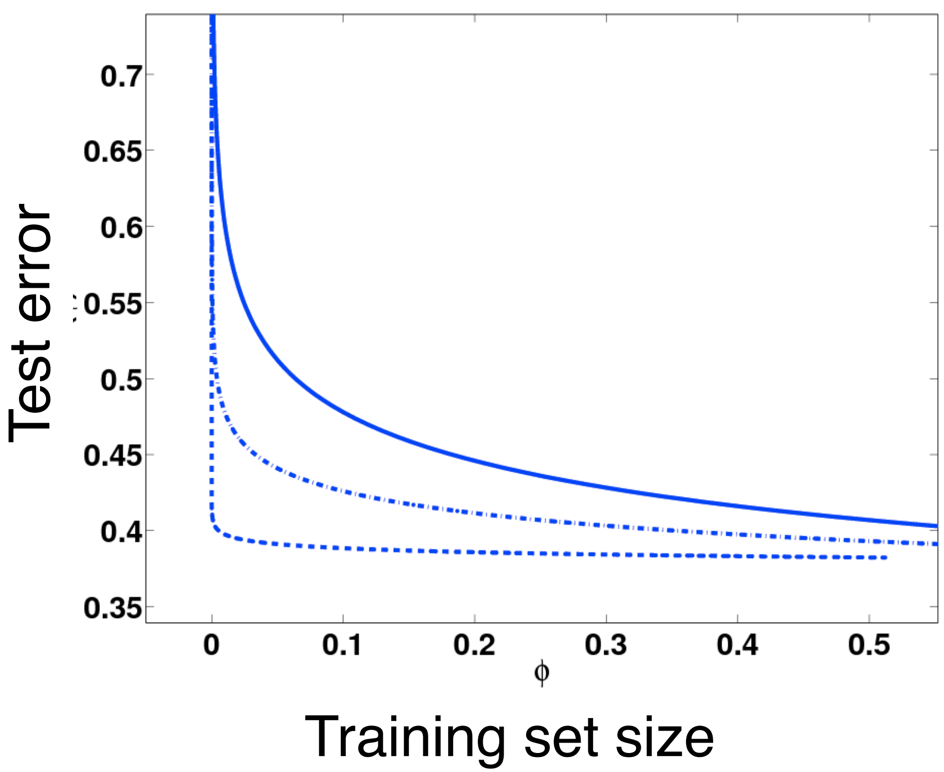
Bias, variance, and noise

	Description	How to detect	How to fix
Bias	How well model fits data	Experiment with model complexity	Change model class
Variance	How much sample size affects accuracy	Fit inverse power low from subsampling	Increase training data size
Noise	Error independent of model class and sample size	Estimate Bayes error with distance metrics	Increase number of features

Detect Variance: Change training set size

- Plotting model performance versus training data size is known as a **Type II learning curve** [Domhan et al, 2015]
- Empirically we can fit Type II learning curves with **inverse-power laws**.

$$\bar{\gamma}_a(\hat{Y}, n_a) = \alpha_a n_a^{-\beta_a} + \delta_a$$

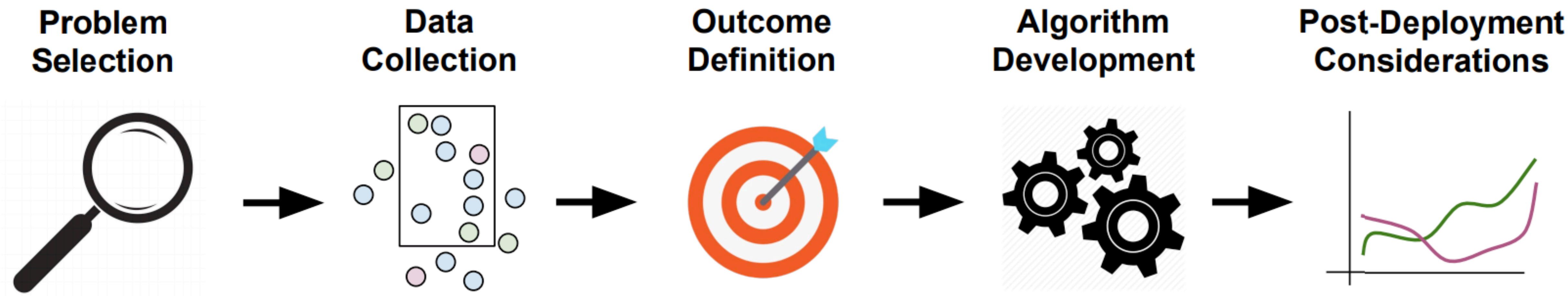


Bias, variance, and noise

	Description	How to detect	How to fix
Bias	How well model fits data	Experiment with model complexity	Change model class
Variance	How much sample size affects accuracy	Fit inverse power low from subsampling	Increase training data size
Noise	Error independent of model class and sample size	Estimate Bayes error with distance metrics	Increase number of features

Clustering Left-Censored Multivariate Time-Series for Disease Phenotyping

Ethical ML Pipeline



Clustering Left-Censored Multivariate Time-Series for Disease Phenotyping

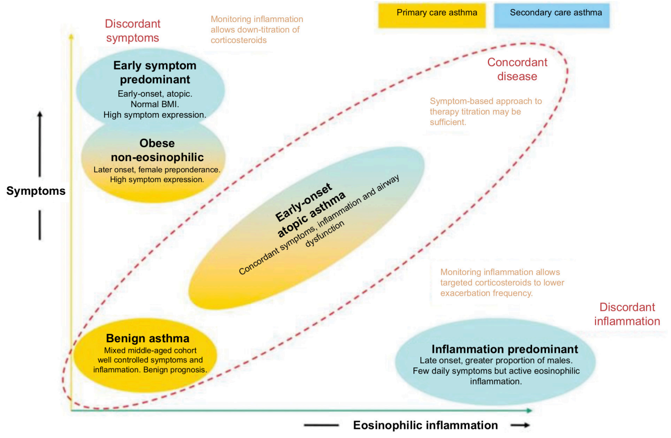
Chen, Krishnan, Sontag; Under Review, arxiv.org/abs/2102.07005

Systemic health disparities cause "noise"

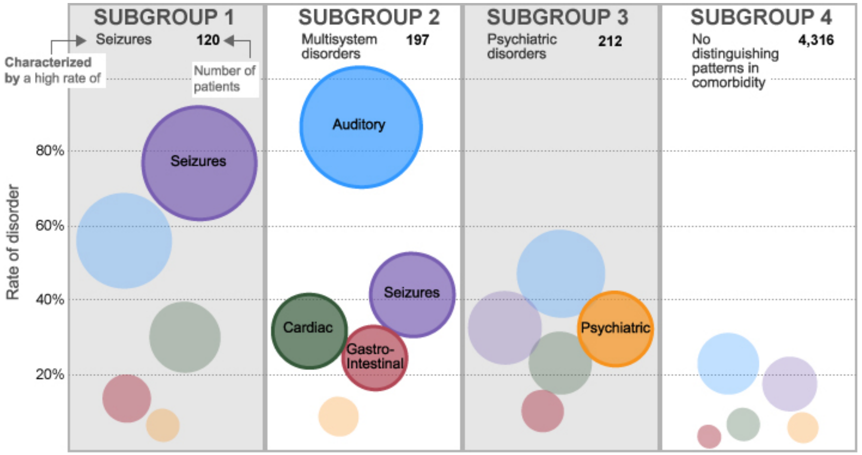
- **Disparities in access to care**
  - Rural hospitals closing, insurance coverage, trust in healthcare system, medical adherence
- **Disparities in treatment**
  - Different treatments for same conditions, same treatments for different physiological systems
- **Disparities in outcomes**
  - Life expectancy by socioeconomic status, maternal morbidity/mortality by race

Chen, Joshi, Ghassemi; *Nature Medicine* 2020

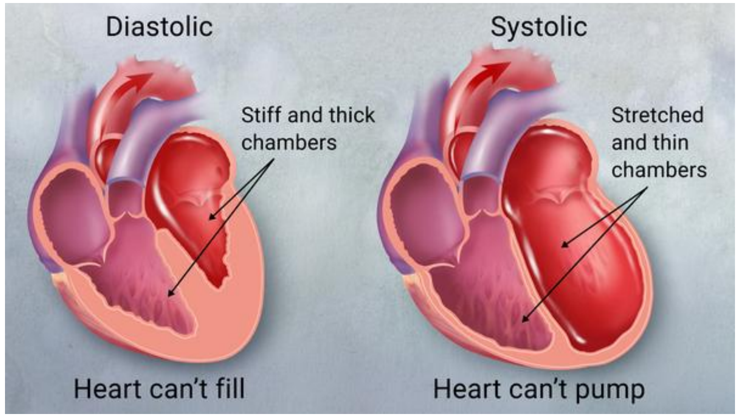
Case study: Many diseases are biologically heterogeneous despite a common diagnosis



Asthma



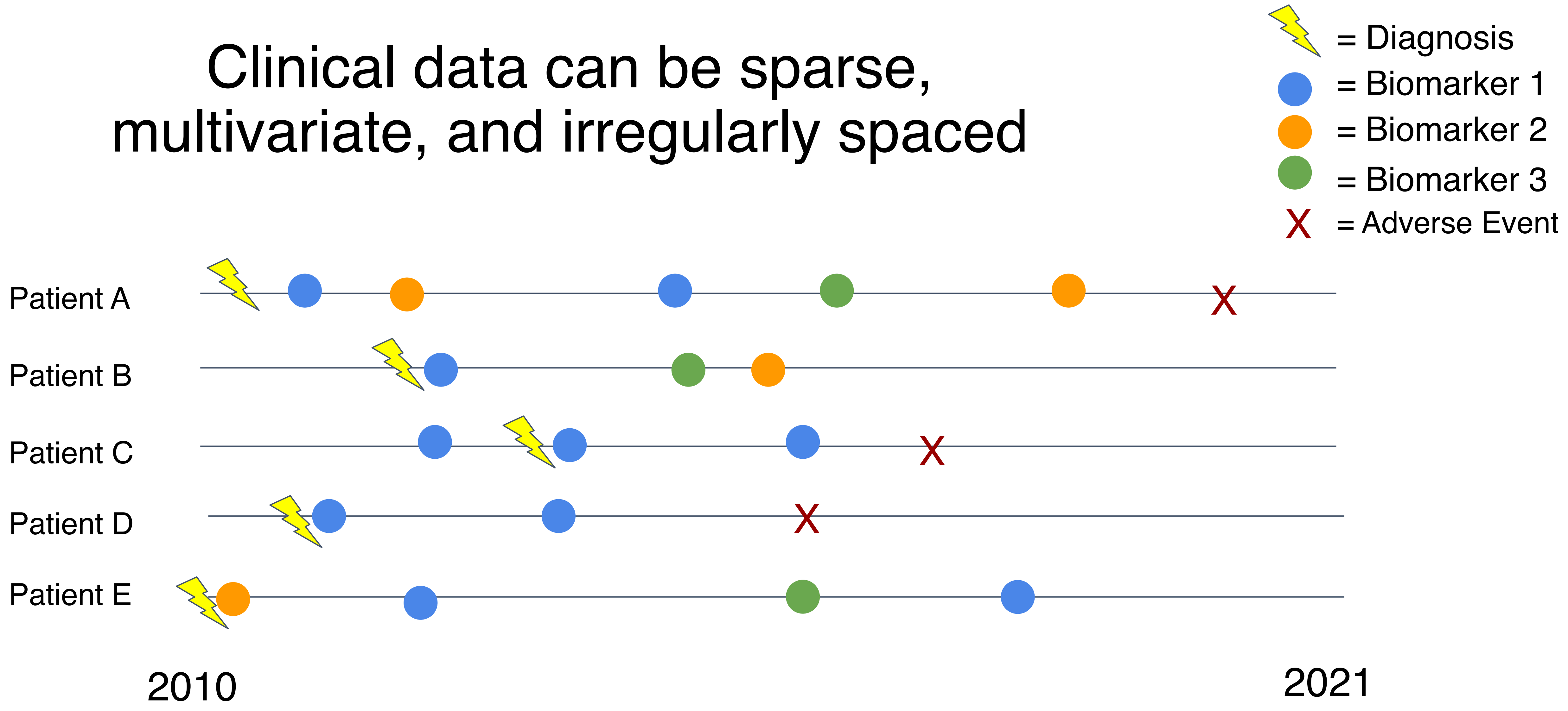
Autism



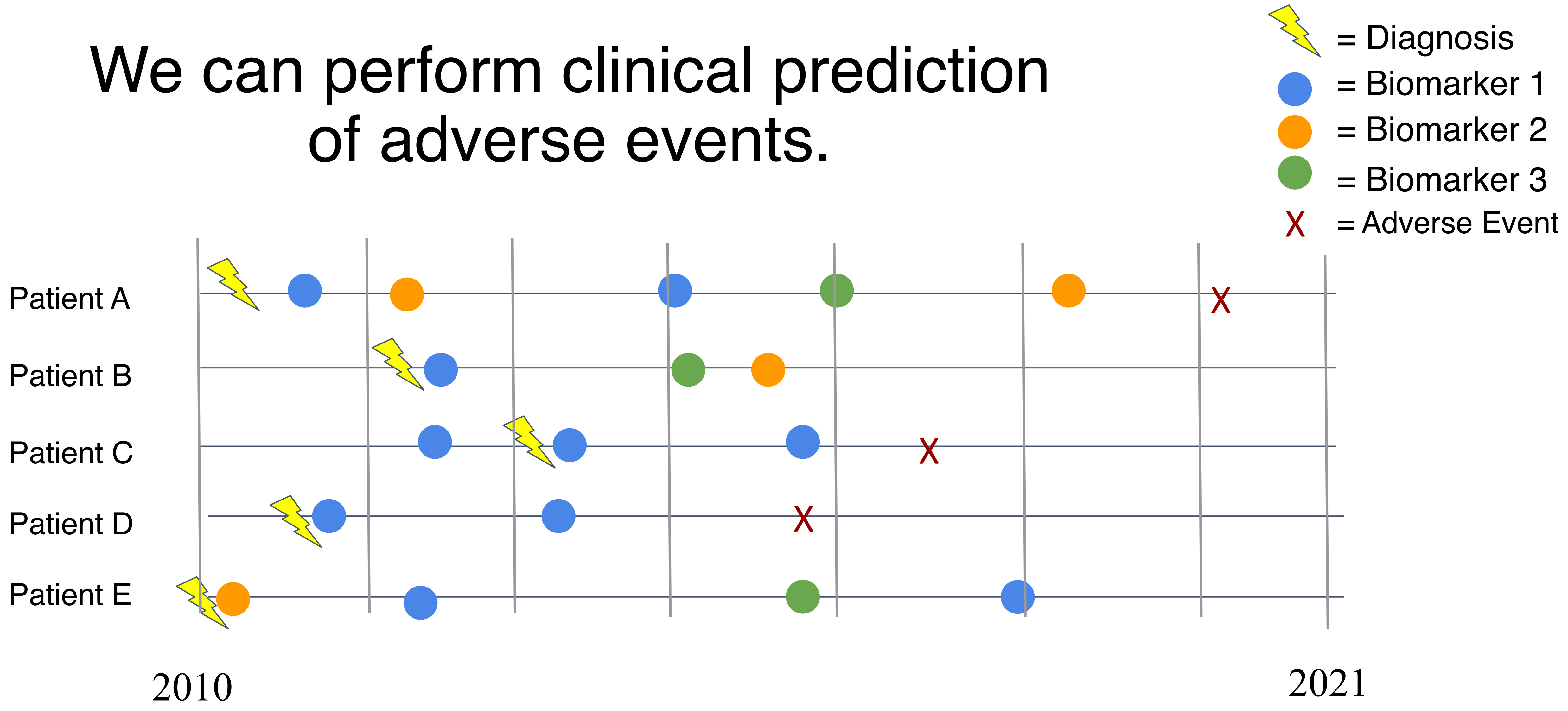
Heart Failure

[1] Nissen et al, *Journal of Asthma and Allergy* 2018.
[2] Kohane et al, *PLoS One*, 2012.
[3] Mayo Clinic

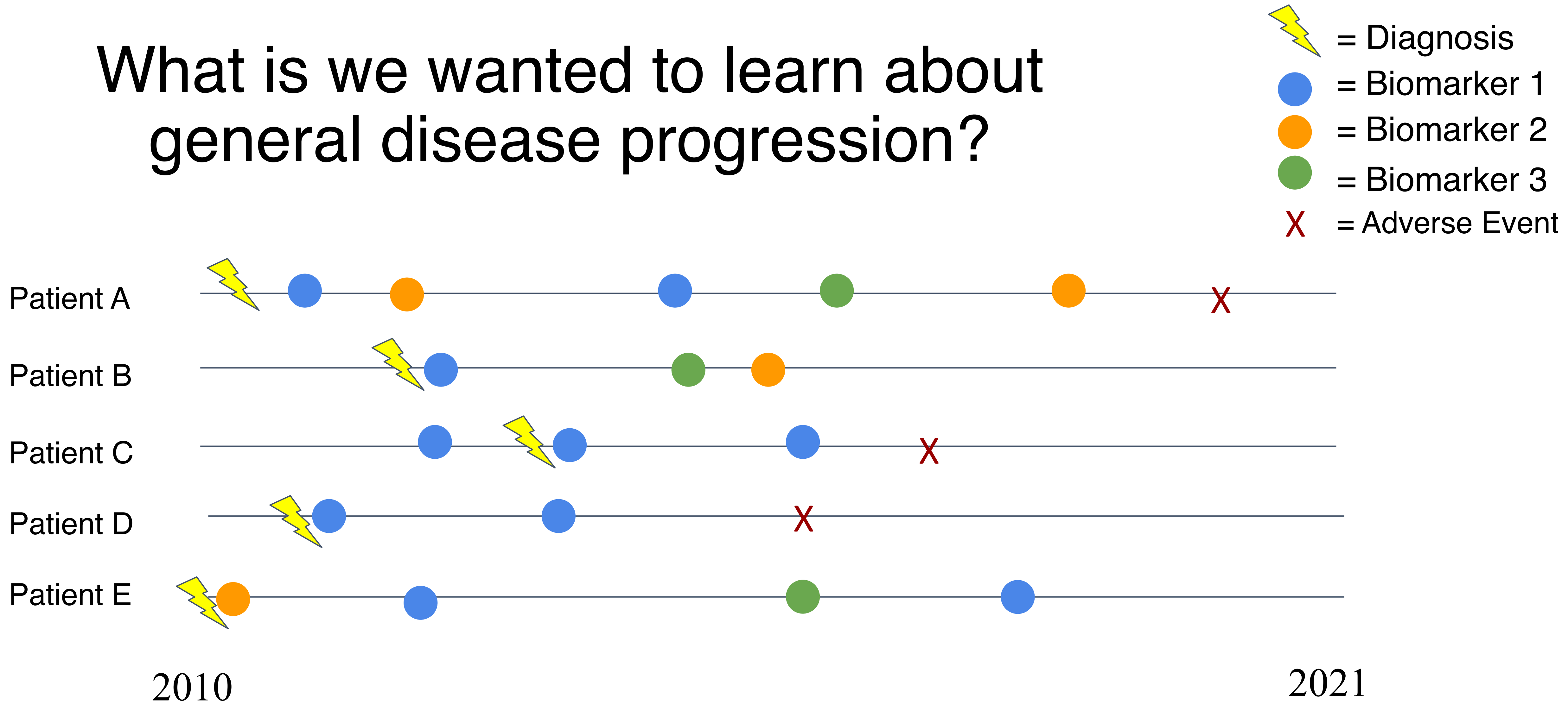
Clinical data can be sparse, multivariate, and irregularly spaced



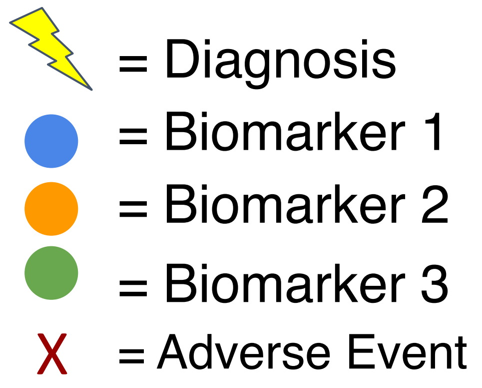
We can perform clinical prediction of adverse events.

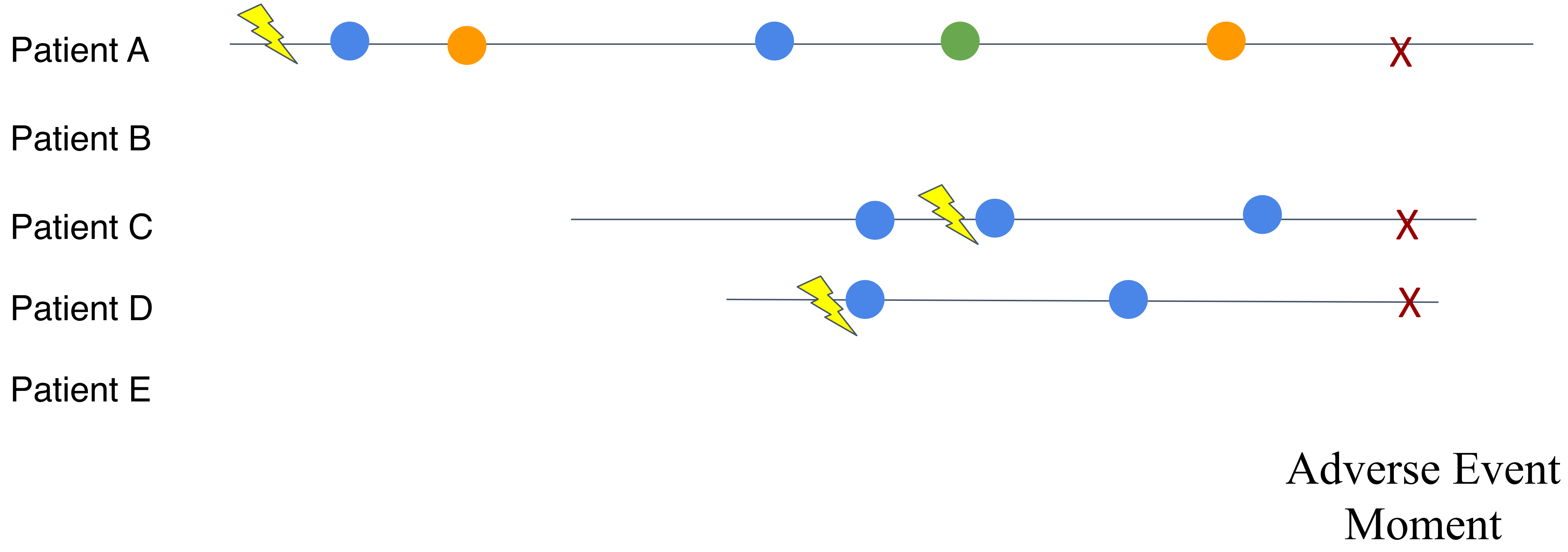


What is we wanted to learn about general disease progression?

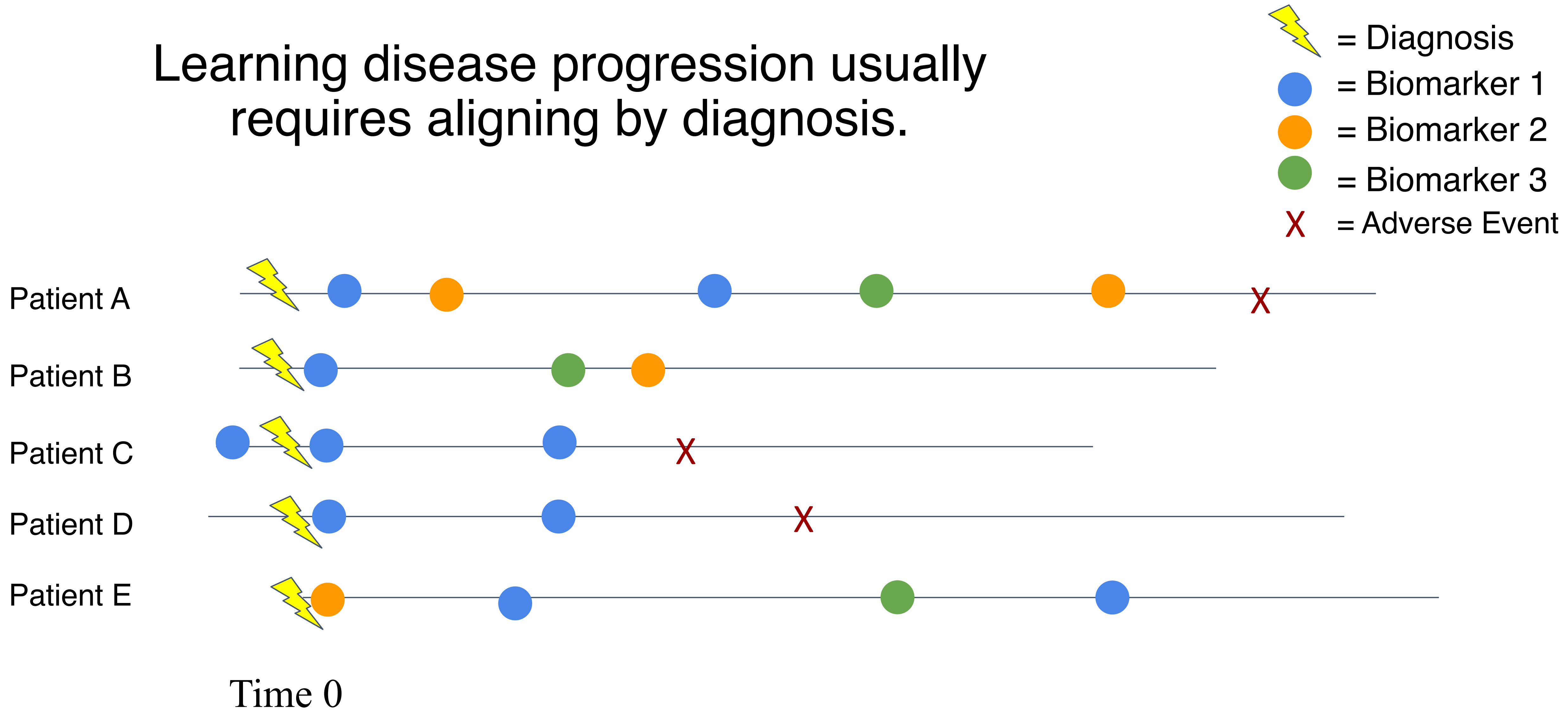


We could align by adverse event, but this limits our dataset.





Learning disease progression usually requires aligning by diagnosis.



Left-censoring hides data before diagnosis

Access to health insurance

Feb 24, 2020, 02:13pm EST | 11,106 views

1 In 4 Rural Hospitals Are At Risk Of Closure And The Problem Is Getting Worse

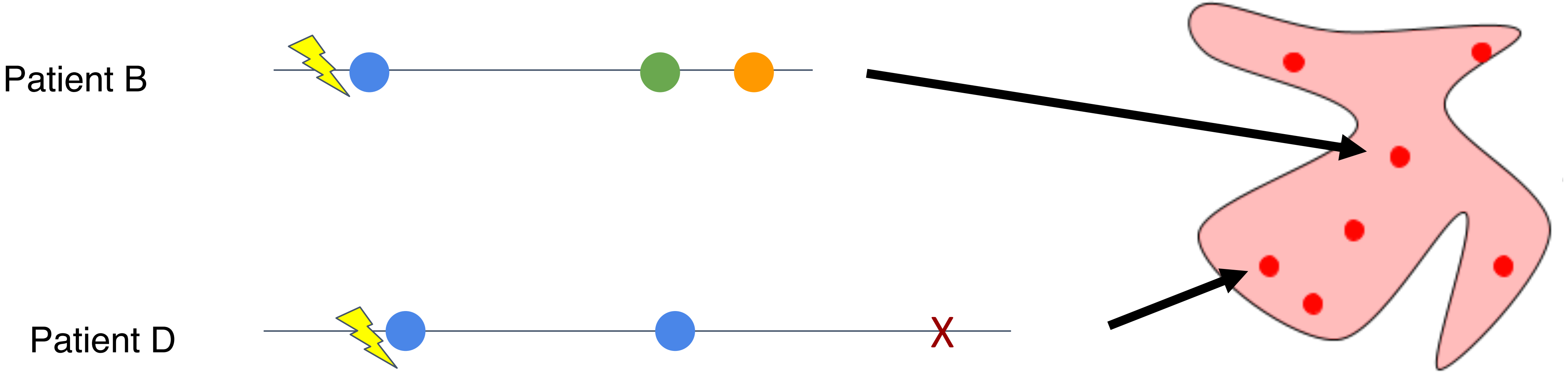
Clary Estes Former Contributor

Healthcare

Geographic proximity to hospitals

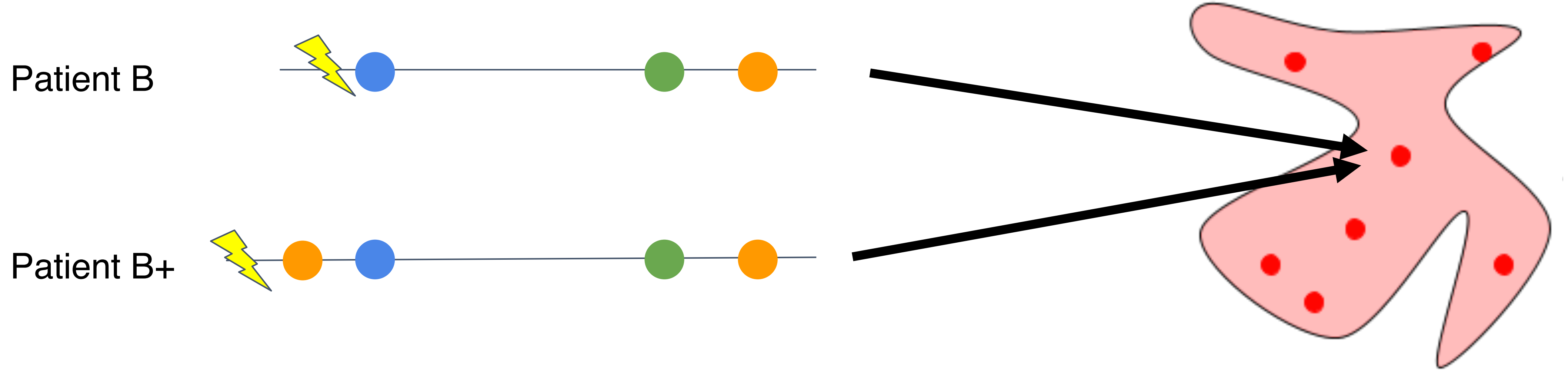
Medical mistrust

A deep generative model maps patients to a low-dimensional latent space



Patients close together are more similar.

A deep generative model maps patients to a low-dimensional latent space



Similar patients with different left-censorship should still be close together.

SubLign: Can we recover clinical subtypes?

	HFpEF		HFrEF
Feature	A (674)	B (444)	C (416)
Age	75.985	74.736	69.438
Female	0.712	0.234	0.435
Anemia	0.230	0.167	0.142
Atherosclerosis	0.285	0.349	0.401
Atrial Fibrillation	0.445	0.550	0.430
Chronic KD	0.277	0.349	0.341
Diastolic HF	0.504	0.363	0.067
Obese	0.568	0.653	0.462
Old MI	0.123	0.142	0.245
Pulmonary HD	0.295	0.225	0.190
Systolic HF	0.093	0.270	0.534

Recovers known heart failure subtypes and suggests other heterogeneity

	Parkinson's patients	Control patients
Model	A (321)	B (298)
Healthy Control	0.551	0.064
Biological Dad With PD	0.028	0.068
Full Sibling With PD	0.010	0.058
UPSIT Part 1	7.558	5.493
UPSIT Part 2	7.648	5.695
UPSIT Part 3	6.988	5.238
UPSIT Part 4	7.539	5.624
UPSIT Total	29.732	22.050

Recovers known features of Parkinson's patients

Intimate Partner Violence and Injury Prediction from Radiology Reports

Ethical ML Pipeline

Problem Selection → **Data Collection** → **Outcome Definition** → **Algorithm Development** → **Post-Deployment Considerations**

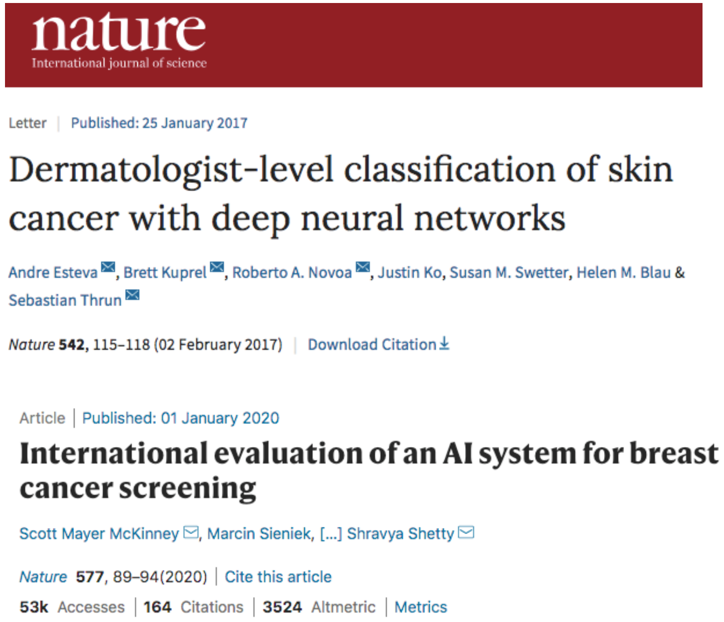
Intimate Partner Violence and Injury Prediction

Chen et al, "Intimate Partner Violence and Injury Prediction from Radiology Reports" PSB 2021.

How can we detect IPV victims early?

Half of all women killed globally are killed by intimate partners or family.[1]

IPV victims reporter higher rates of clinical visits.[2]



nature
International journal of science

Letter | Published: 25 January 2017

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

Nature 542, 115–118 (02 February 2017) | Download Citation

Article | Published: 01 January 2020

International evaluation of an AI system for breast cancer screening

Scott Mayer McKinney, Marcin Sieniek, [...] Shravya Shetty

Nature 577, 89–94(2020) | Cite this article

53k Accesses | 164 Citations | 3524 Altmetric | Metrics

Algorithms can screen patients with performance that exceeds humans.

1. U. N. O. on Drugs and Crime, Global Study on Homicide: Gender-related Killing of Women and Girls (UNODC, United Nations Office on Drugs and Crime, 2018).
2. C. Wisner, T. Gilmer, L. Saltzman and T. Zink, Intimate partner violence against women, Journal of family practice 48, 439 (1999).

How do we get accurate IPV labels?

- Biggest barrier to early intervention is **underreporting** by the patient because of shame, economic dependency, or lack of trust in healthcare providers
- IPV victims use healthcare services like the **emergency department** or **imaging studies** at higher rates than other patients
- We examine 1,479 victims and control patients at Brigham and Women's Hospital (BWH) in Boston

What kind of labels could we use?

1. **ICD codes**: Based on clinical staff assessment

2. **Patient self-report**: Based on patient enrollment in violence prevention program

3. **Radiologist labeling**: Based on injuries in radiology reports

1) Self-report labels

- **Inclusion Criteria**
  - <u>IPV victims</u>: Identified as entering a violence prevention program at BWH, for IPV, with at least one radiology study at BWH
  - <u>Control cohort</u>: Age- and sex-matched patients in the BWH patient population with at least one radiology study at BWH
- **Features**
  - Radiology report text, extracted from template
- **Label**
  - Was this person a **self-report to the BWH violence prevention program**?

Passageway – Domestic Abuse Intervention and Prevention

CCHHE's Passageway program works to improve the health, wellbeing, and safety of those experiencing abuse from an intimate partner. We offer the following support services to hospital and health center patients, employees, and community members:

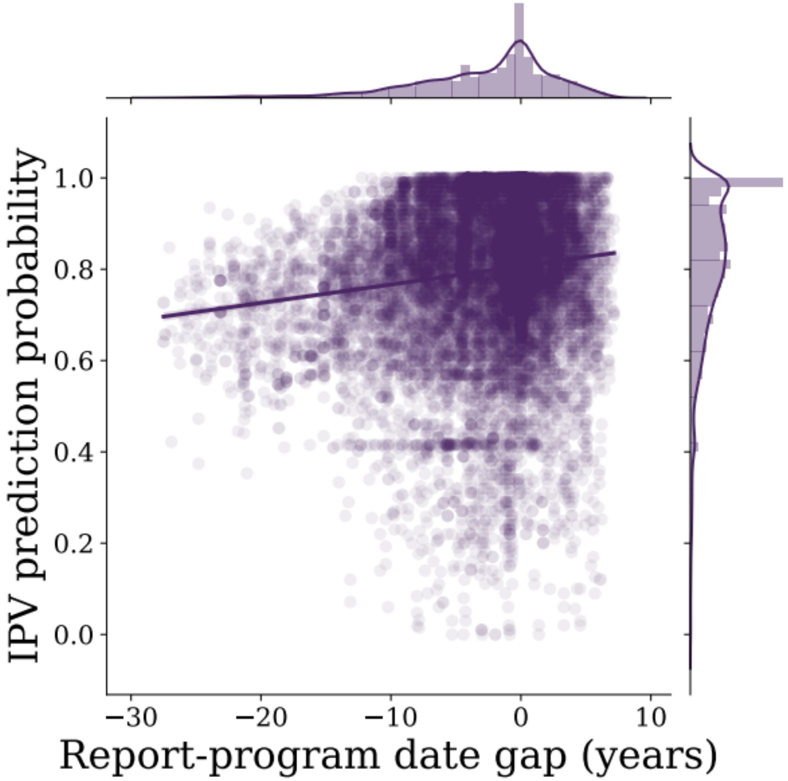
- Free and confidential advocacy services*
- Safety planning
- Individual counseling and support
- A safe place to talk
- Information about the health effects of domestic violence
- Support groups
- Medical advocacy
- Legal and court advocacy
- Referrals to community resources (health care, housing, shelter, lawyers, and others)

2) Radiology injury label

- **Inclusion Criteria**
  - Data from BWH
- **Features**
  - Radiology report text, extracted from template
  - Each report text treated as separate
- **Label**
  - **Fellowship-trained emergency radiologists** provided injury labels

How do predictions differ on the two label sets?

- Models performance for both labels are **comparable**
  - Self-report label: 0.84 ± 0.03
  - Radiologist label: 0.87 ± 0.01
- We **can use self-report labels**, which are much less time intensive than radiologist labels.
- We can detect IPV a **median of 3.08 years** before program entry (sensitivity 64%, specificity 95%)



Ethical ML Pipeline

Problem Selection → **Data Collection** → **Outcome Definition** → **Algorithm Development** → **Post-Deployment Considerations**

Intimate Partner Violence and Injury Prediction

Clustering Left-Censored Multivariate Time-Series for Disease Phenotyping

Diagnosing Sources of Unfairness in Supervised Learning