Machine Learning Approaches for Equitable Healthcare

Irene Y. Chen

PhD Student, Electrical Engineering and Computer Science

July 8, 2022

MIT Clinical ML
www.clinicalml.org

Why ML for healthcare?

Why ML for healthcare?

Electronic Medical Records

Genomics

Medical Imaging

Signals

Molecular Data

Wearable Data

Why ML for healthcare?

Electronic Medical Records

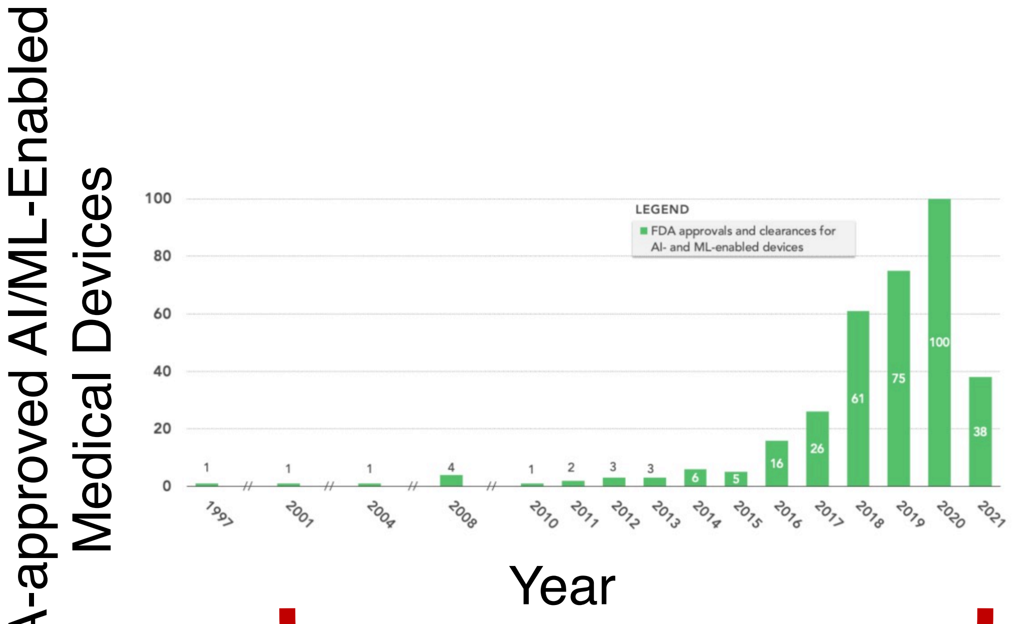
Genomics

Medical Imaging

Signals

Molecular Data

Wearable Data



343 total FDA approvals

38 in first half of 2021

FDA 2021, Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices

Why is machine learning for healthcare challenging?

Why is machine learning for healthcare challenging?

- Healthcare data are **limited and sparse**
  - Data for prediction models can have over 50% of values missing.[1]
  - Data sparsity can itself have patterns, e.g. time of lab test.[2]

- **Treatment variation** across hospitals and clinicians,[3] even for the same patient.[4]

- Healthcare **knowledge changes** all the time.
  - **13% of medical practice papers** are reversals**.**[5]

[1] Pantalone et al, Diabetes Medicine 2012; [2] Agniel et al, BMJ 2018; [3] Coburn et al, Breast Journal 2008; [4] Sporer et al, American Academy of Orthopaedic Surgeons 2006; [5] Prasad et al, JAMA Internal Medicine

Why is **equitable** healthcare challenging?

- Healthcare system has existing **health disparities**, for example maternal morbidity in Black women[1]

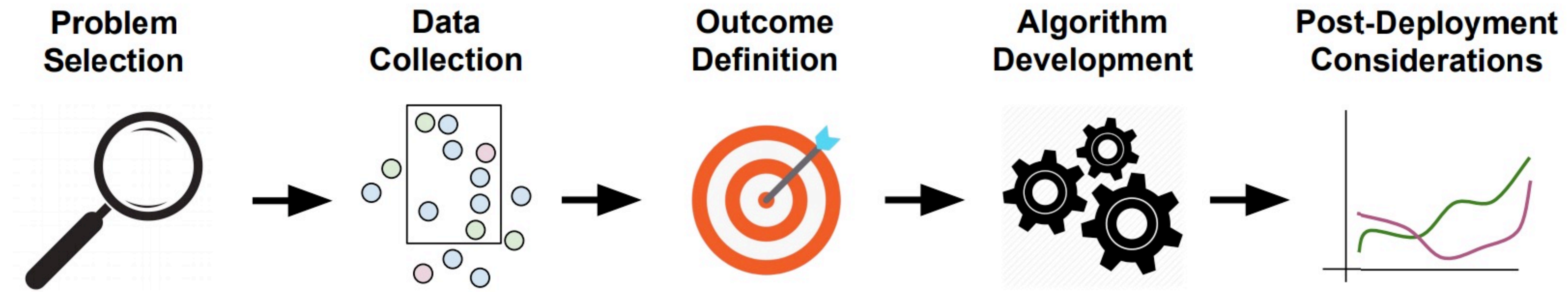
- **Uneven sample sizes** in data: 96% participants in GWAS datasets are of European descent[2]

- Subpopulations can face **differences in data distributions**, including differences in heart attack symptoms and care[3]

- Biased systems and biased datasets create **algorithmic bias**[4]

[1] NYC Government, Maternal Morbidity Report; [2] Need and Goldstein, Trends in Genetics 2009; [3] Goldberg et al, American Heart Journal 1998; [4] Obermeyer et al, *Science 2019.*

Machine Learning for Equitable Healthcare



Problem Selection

1. Early detection for intimate partner violence (PSB 2021)
2. Treating health disparities with AI (Nature Medicine 2020)

Data Collection

Collecting and researching insurance risk scores (ongoing)

Outcome Definition

Assessing different quality labels in intimate partner violence (ongoing)

Algorithm Development

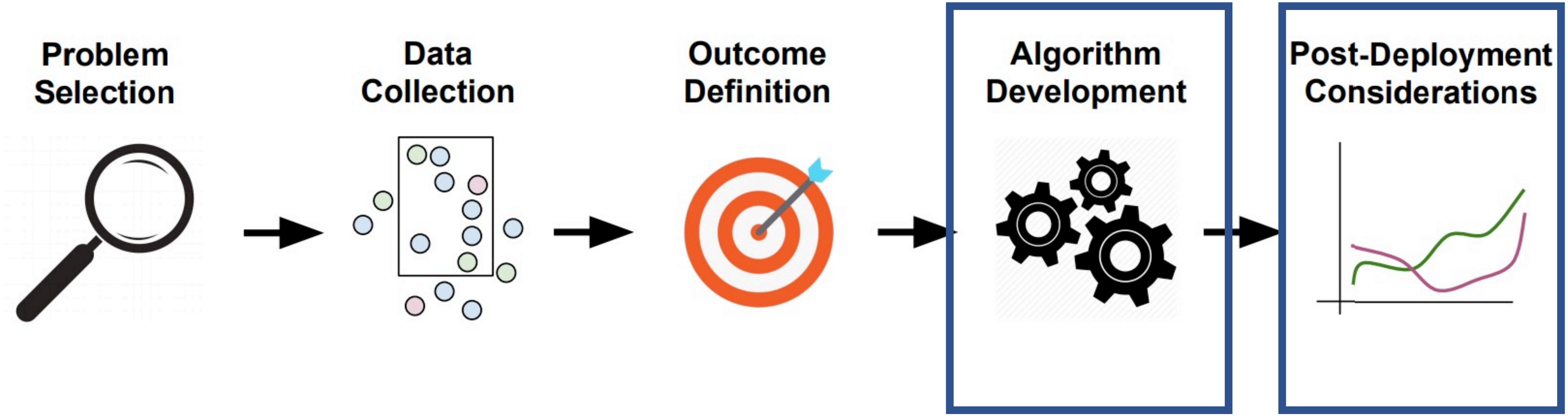
Correcting for patient access to care (AAAI 2022)

Post-Deployment Considerations

1. Bias auditing (AMA Journal of Ethics 2019, Nature Medicine 2021)
2. Mitigating algorithmic bias (NeurIPS 2018)

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

Machine Learning for Equitable Healthcare



Problem Selection

Data Collection

Outcome Definition

Algorithm Development

Post-Deployment Considerations

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

Today’s Talk

1. How can we **decompose** sources of **discrimination**? (NeurIPS 2018)

2. How can we **proactively build algorithms** that account for differences in access to care? (AAAI 2022)

How can we decompose sources of discrimination?

Chen, Johansson, Sontag, “Why is My Classifier Discriminatory?,” *NeurIPS 2018.*

Motivation: Risk Stratification for Clinical Interventions

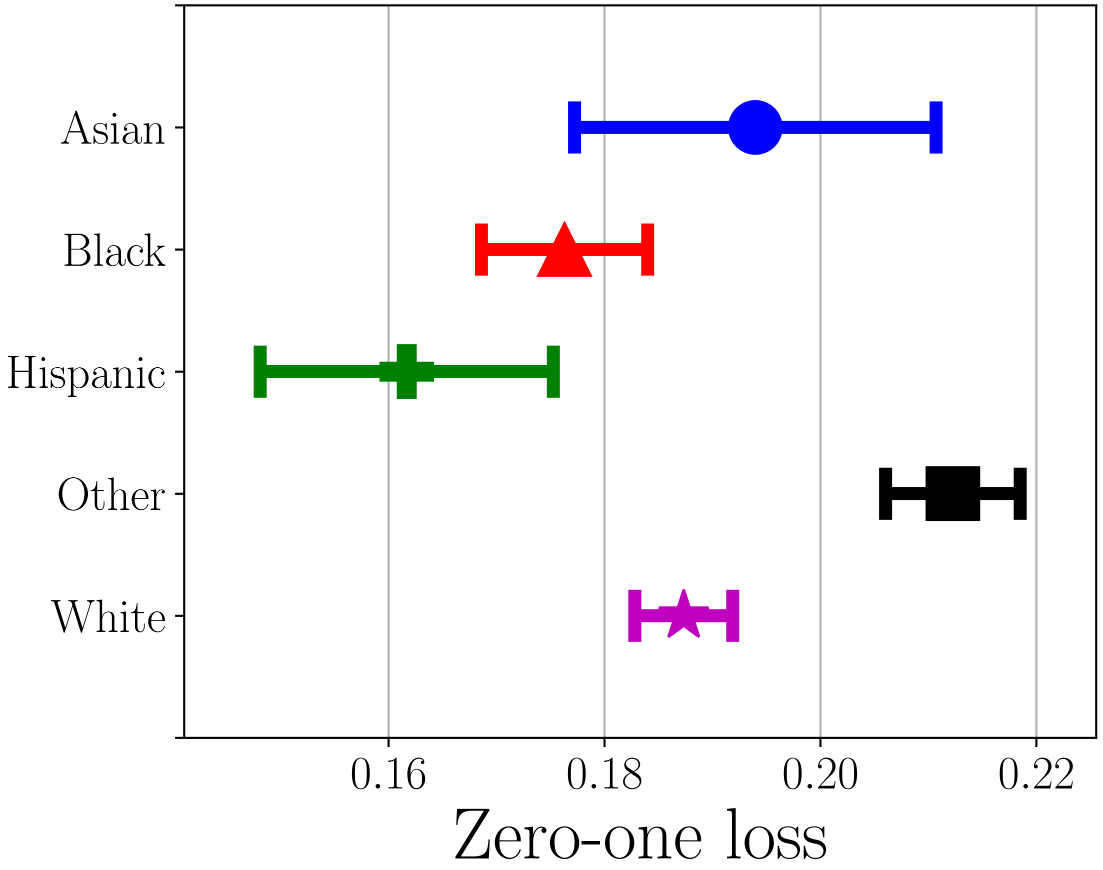
- Examples include APGAR score for **newborns**

- Risk stratification algorithms help clinicians choose **interventions** in real-time

- However, risk scores face new scrutiny as some are shown to generate **divergent risk estimates** for patients with identical risk profiles but different races.

Intensive Care Unit Mortality Prediction

Intensive Care Unit Mortality Prediction



How do we define fairness?

- We define fairness in the context of loss like false positive rate, false negative rate, etc.
- For outcome *Y* and prediction $\hat{Y}$ on data *D* , **zero-one loss** is:

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

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

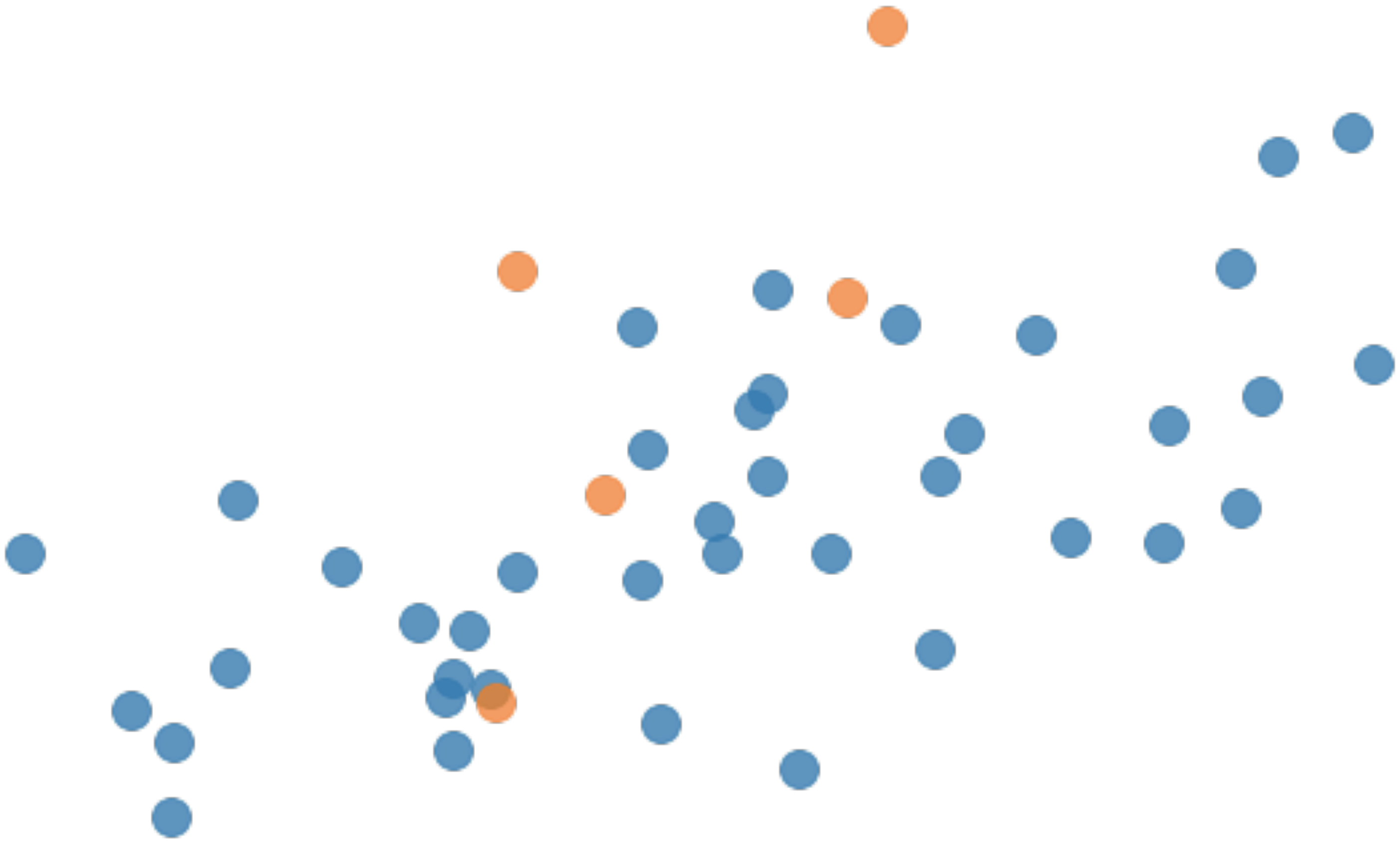
$$\bar{\Gamma}\left(\hat{Y}\right) := |\gamma_1 - \gamma_0|$$

Bias, variance, and noise

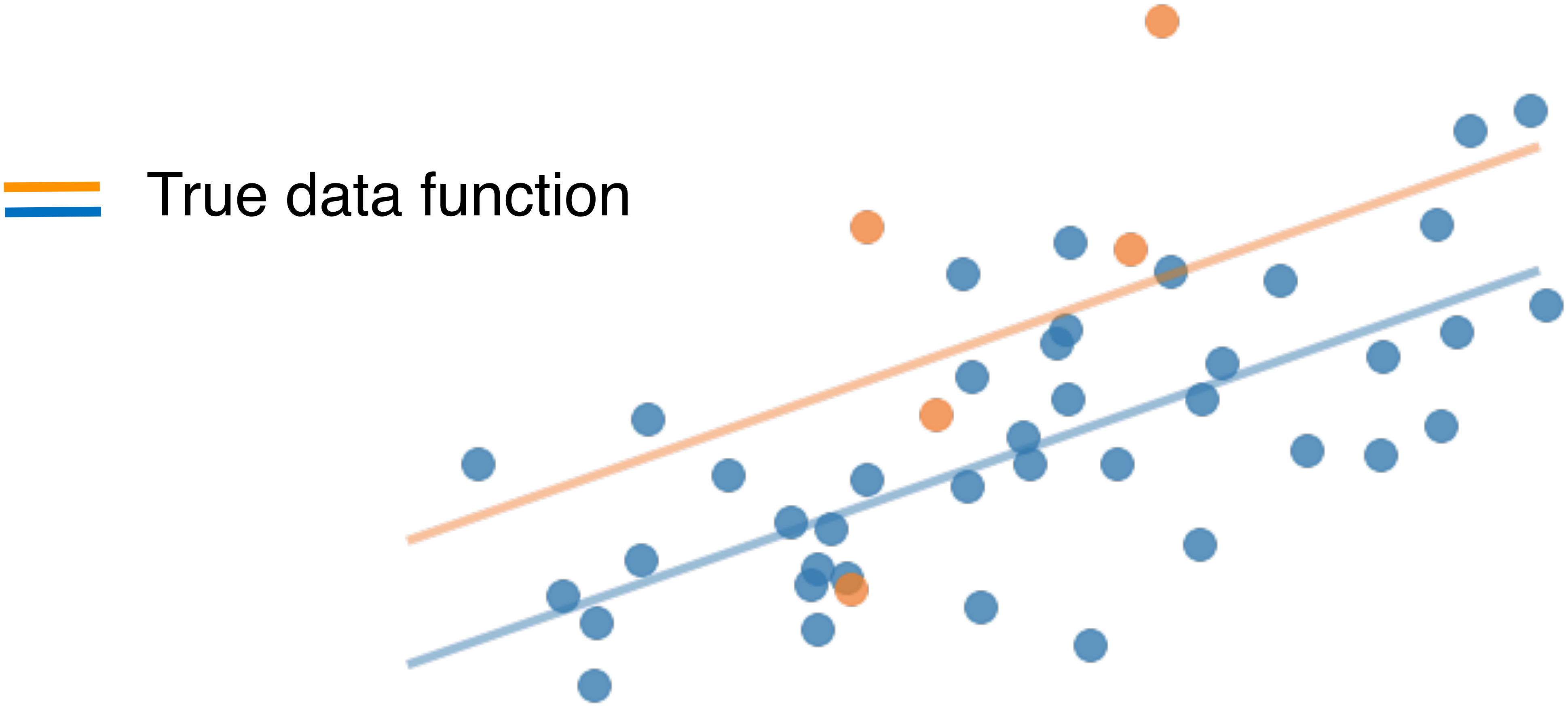
	Description	**How to fix**
Bias	How well model fits data	Change model class
Variance	How much sample size affects accuracy	Increase training data size
Noise	Error independent of model class and sample size	Increase number of features

Why might my classifier be unfair?

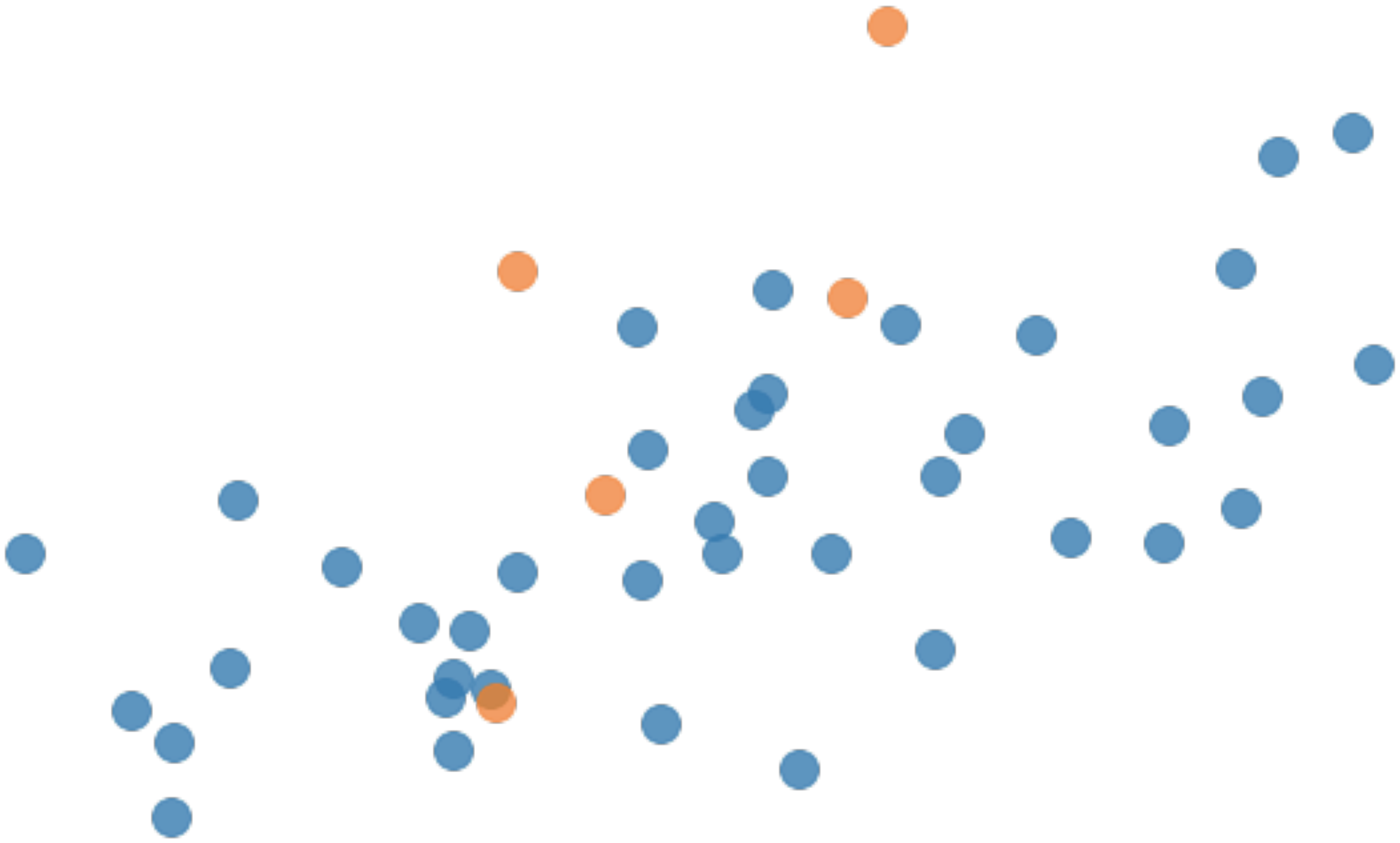
Why might my classifier be unfair?



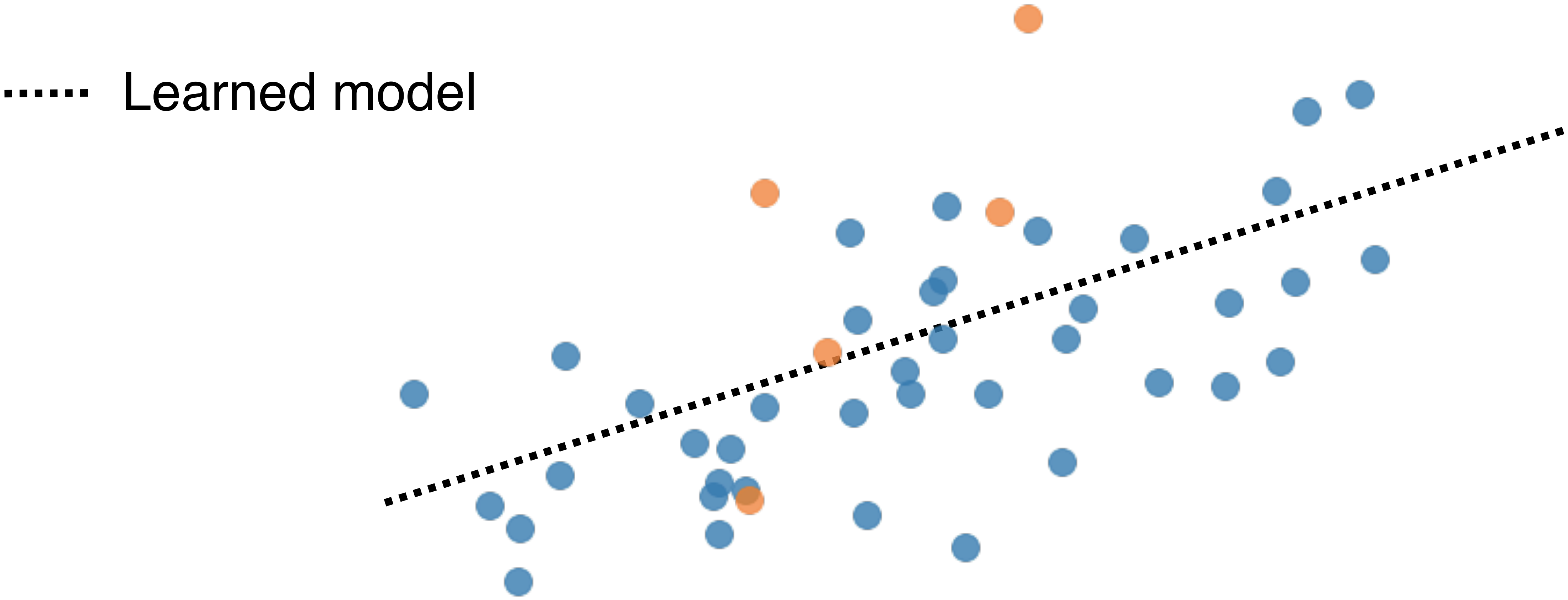
Why might my classifier be unfair?



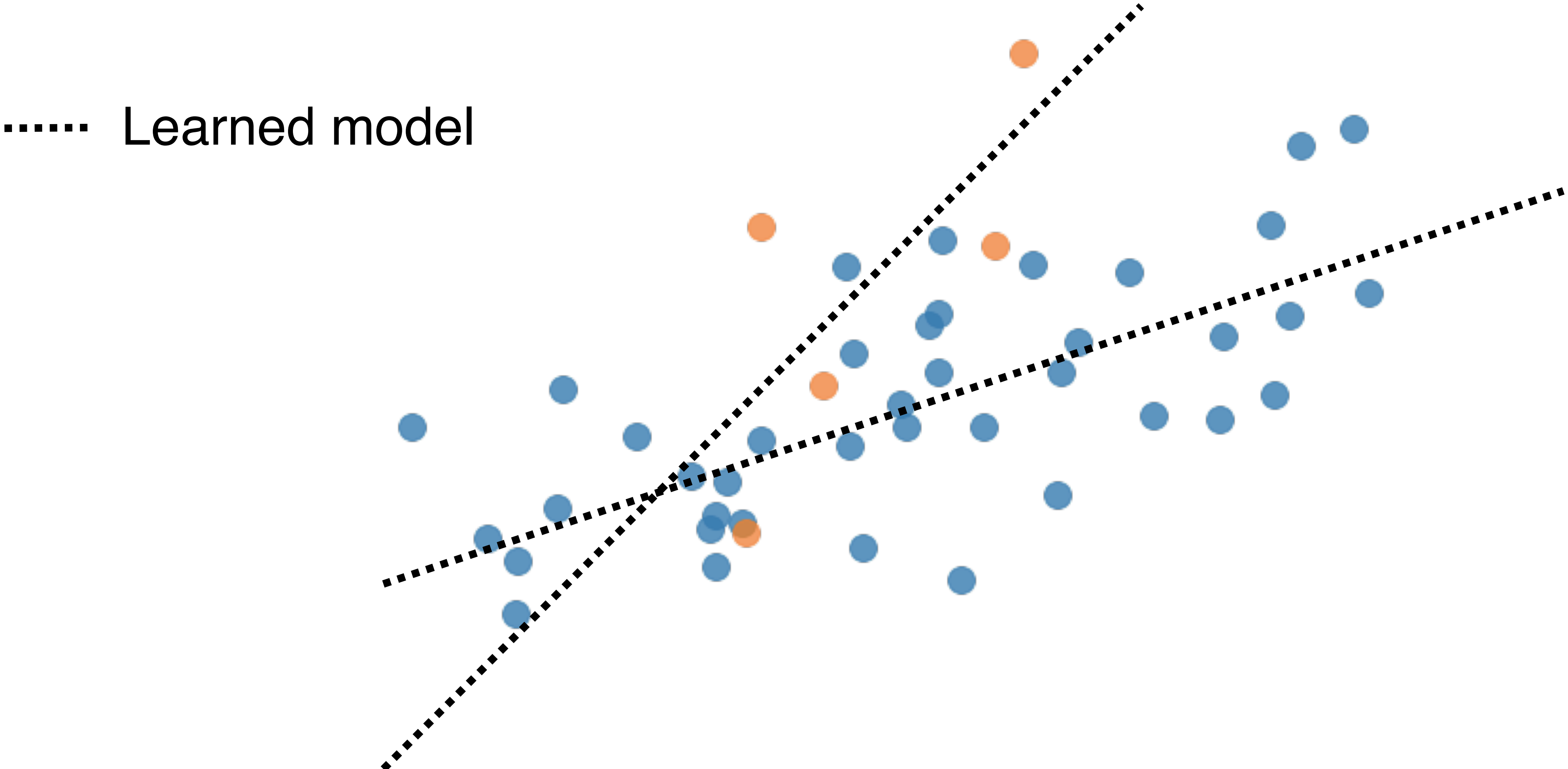
Why might my classifier be unfair?



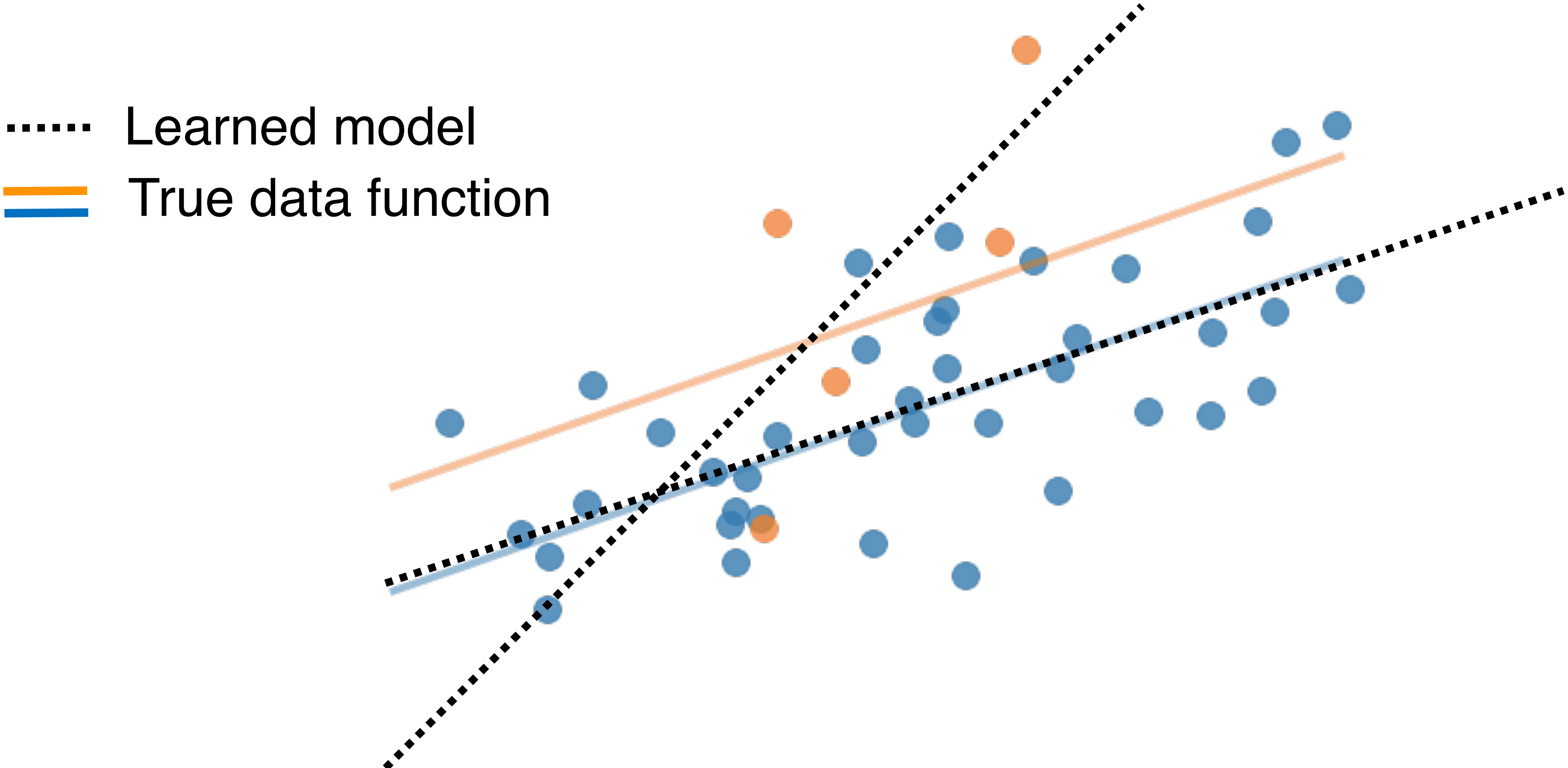
Why might my classifier be unfair?



Why might my classifier be unfair?



Why might my classifier be unfair?

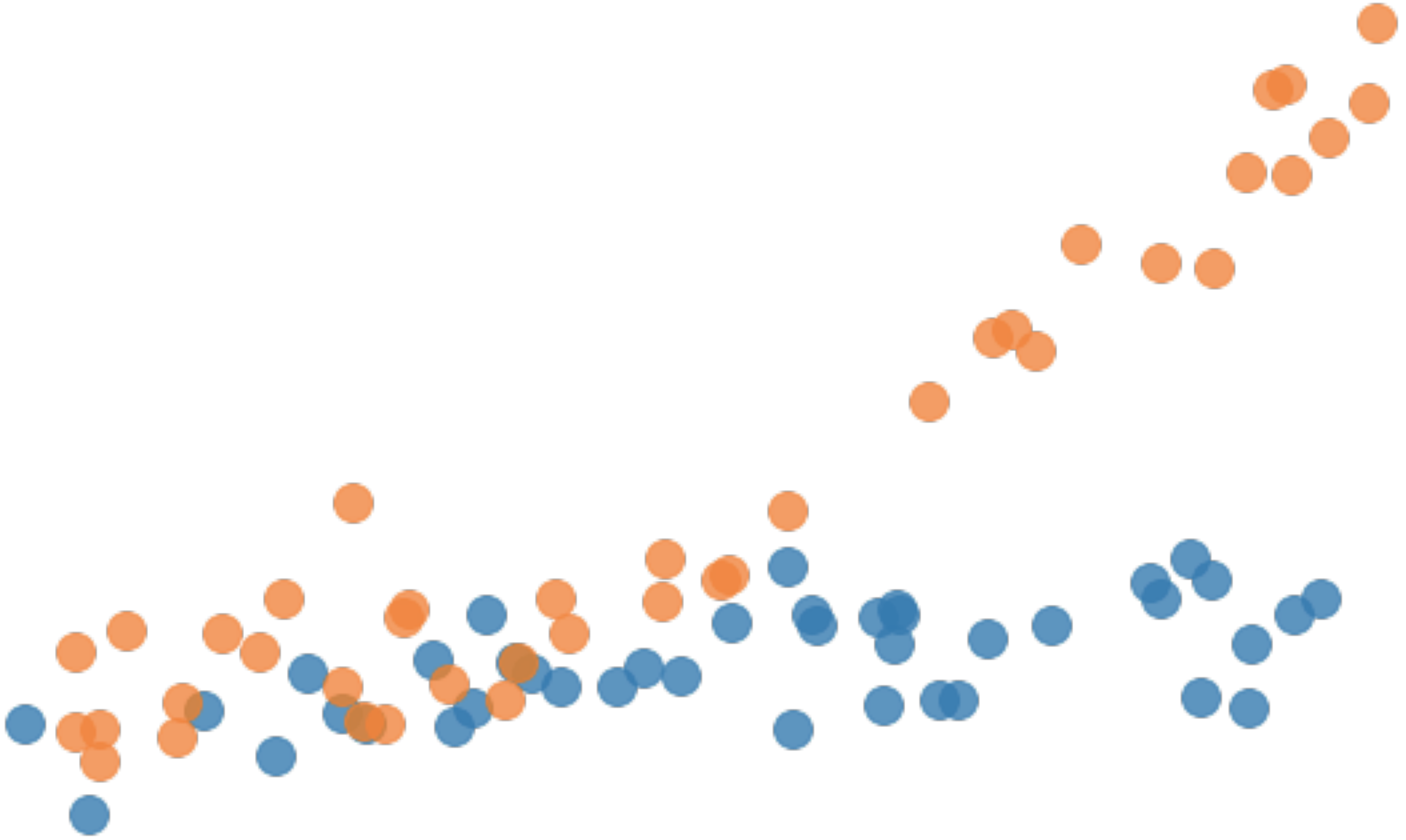


Why might my classifier be unfair?

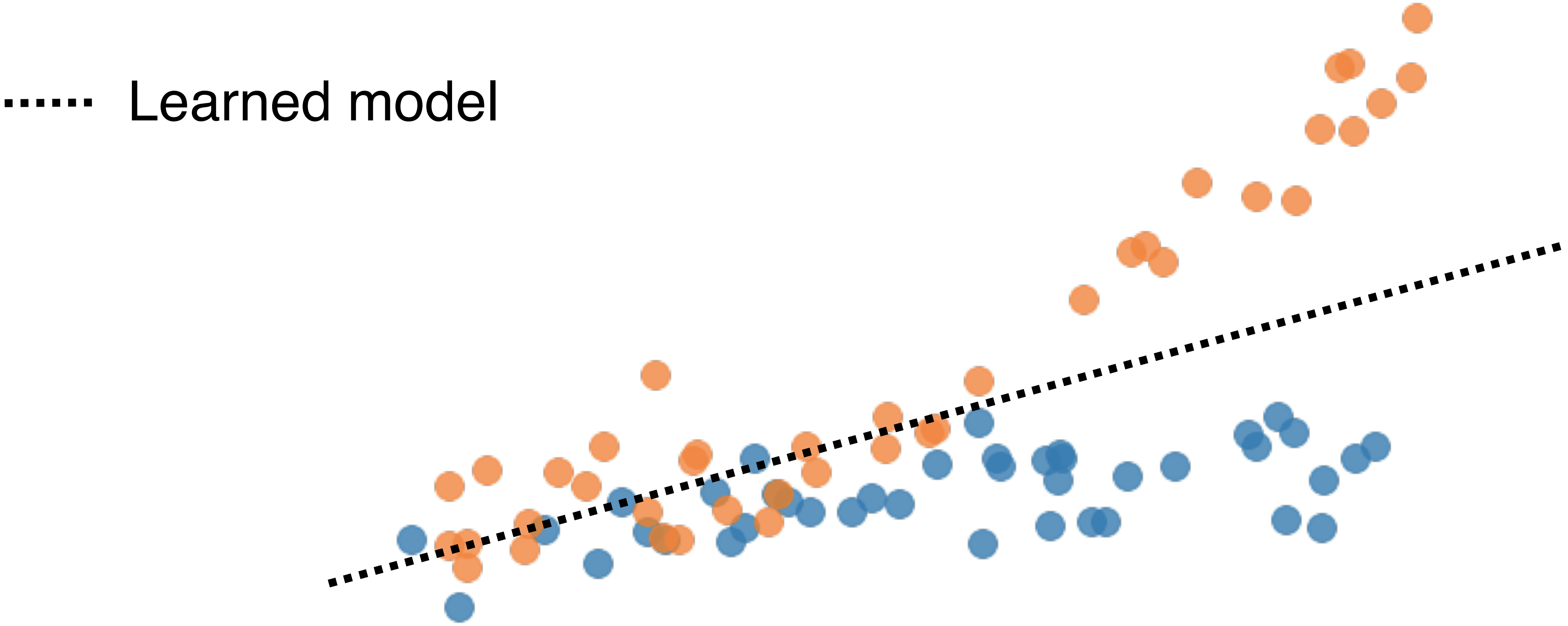
Error from **variance** can be solved by **collecting more samples**.

Learned model
True data function

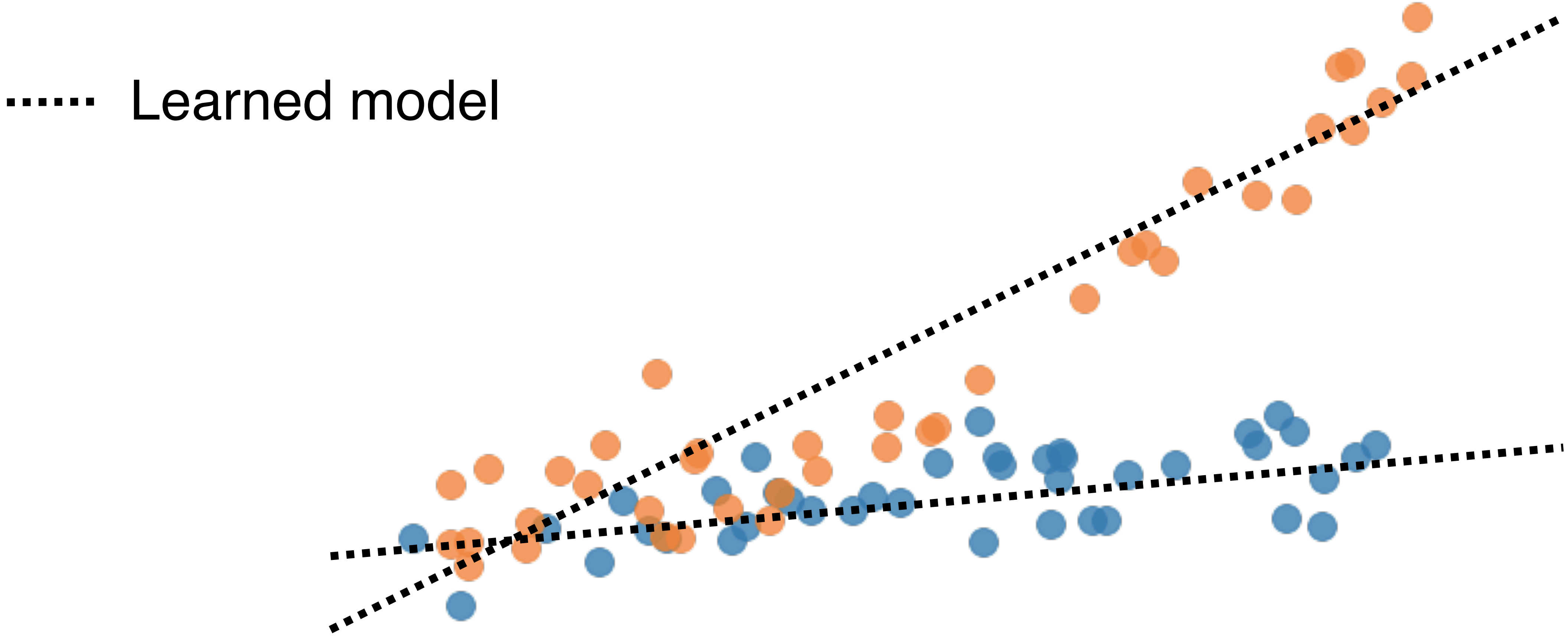
Why might my classifier be unfair?



Why might my classifier be unfair?



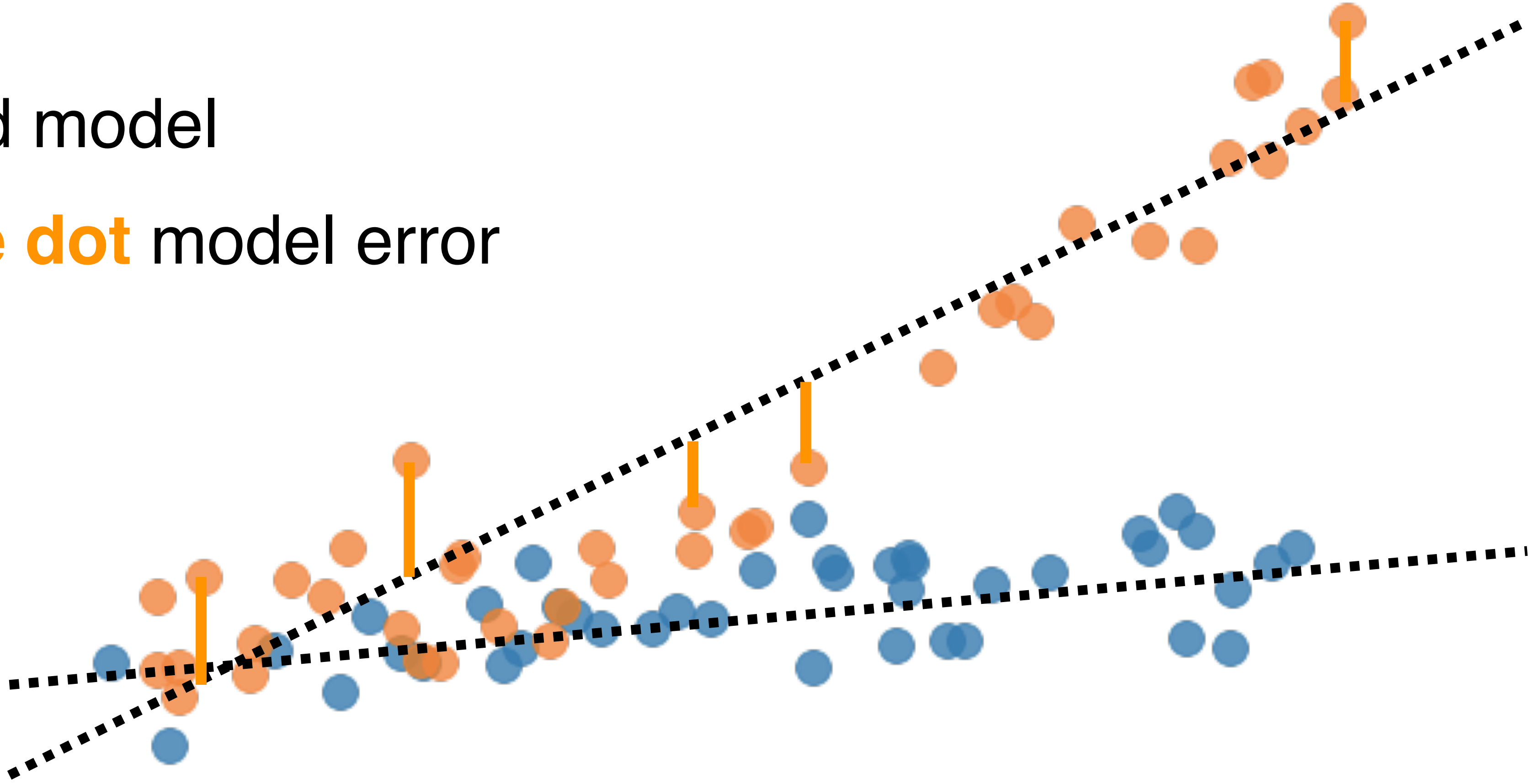
Why might my classifier be unfair?



Why might my classifier be unfair?

⋯⋯ Learned model

| **Orange dot** model error

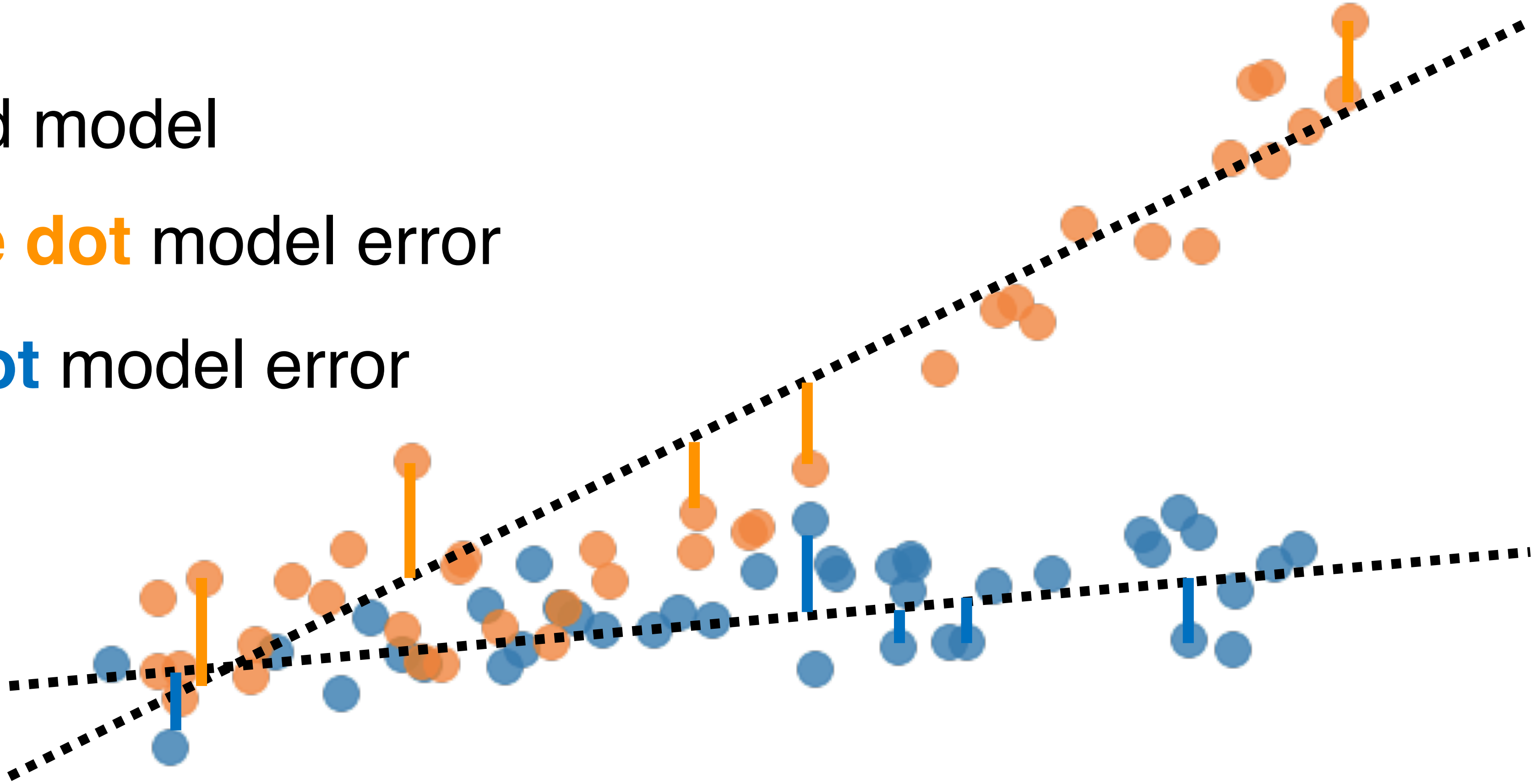


Why might my classifier be unfair?

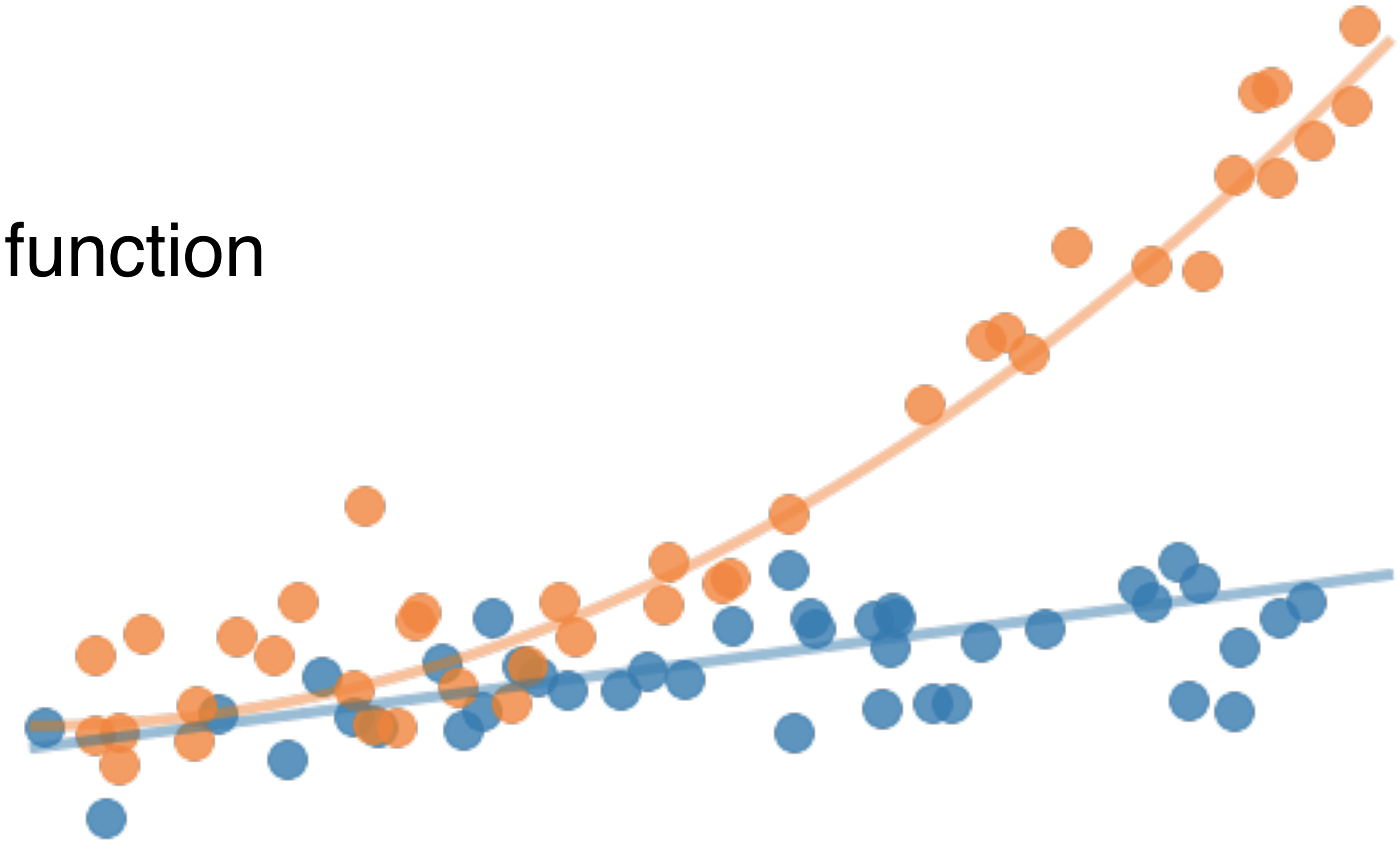
······ Learned model

| **Orange dot** model error

| **Blue dot** model error



Why might my classifier be unfair?

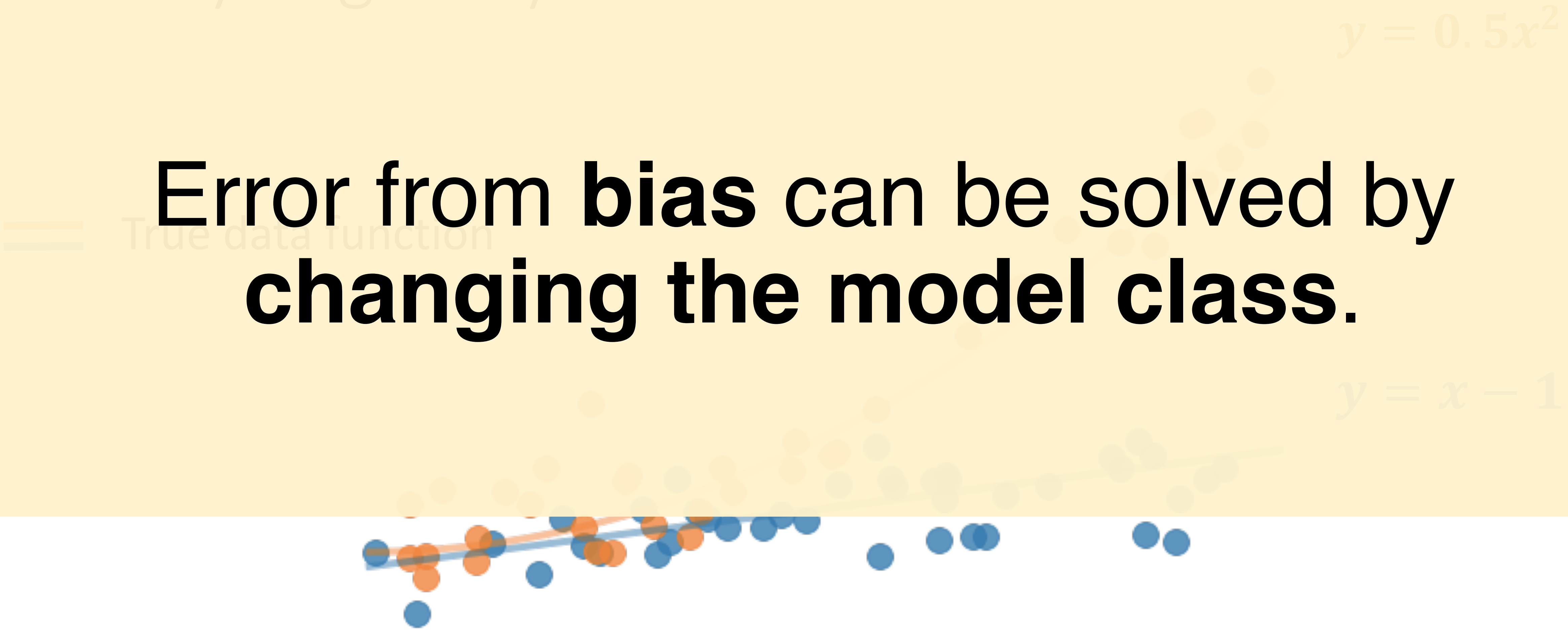


$y = 0.5x^2$

True data function

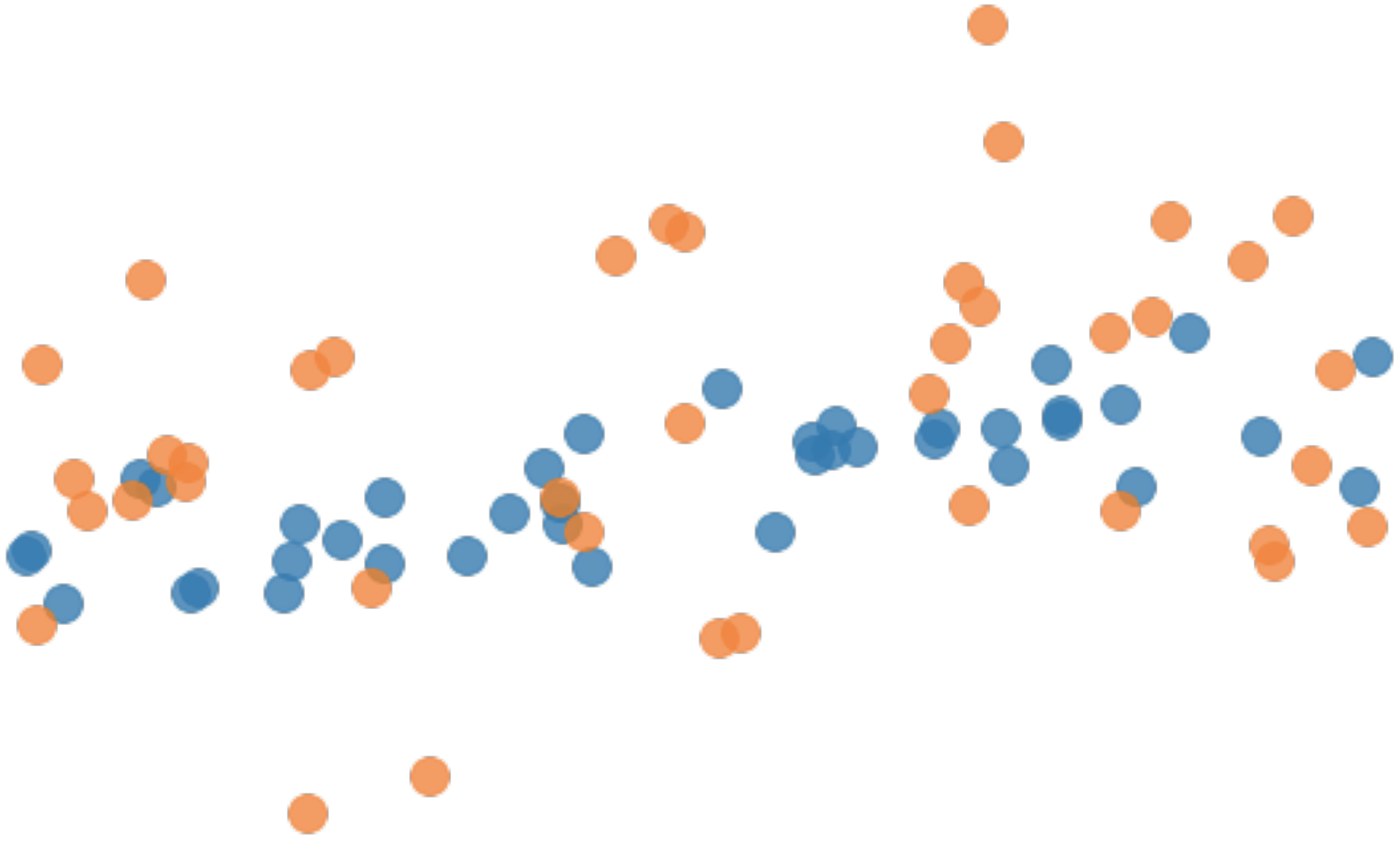
$y = x - 1$

Why might my classifier be unfair?

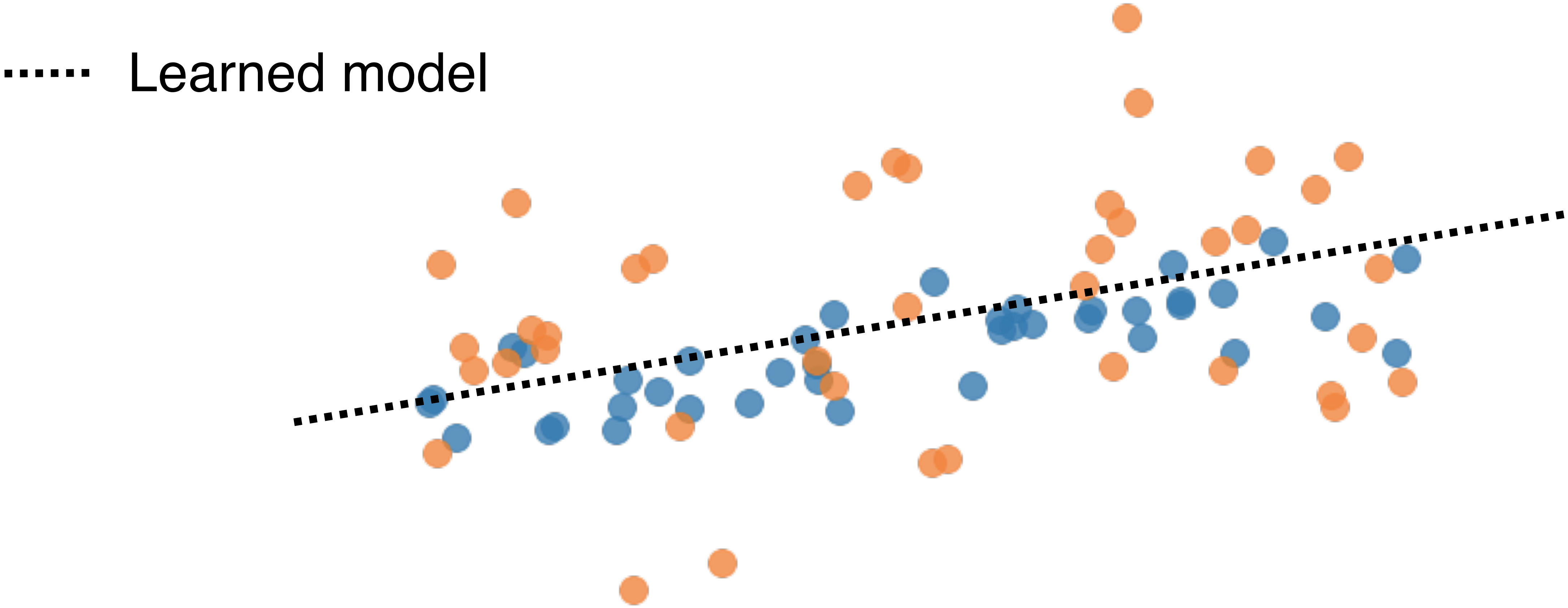


Error from **bias** can be solved by **changing the model class**.

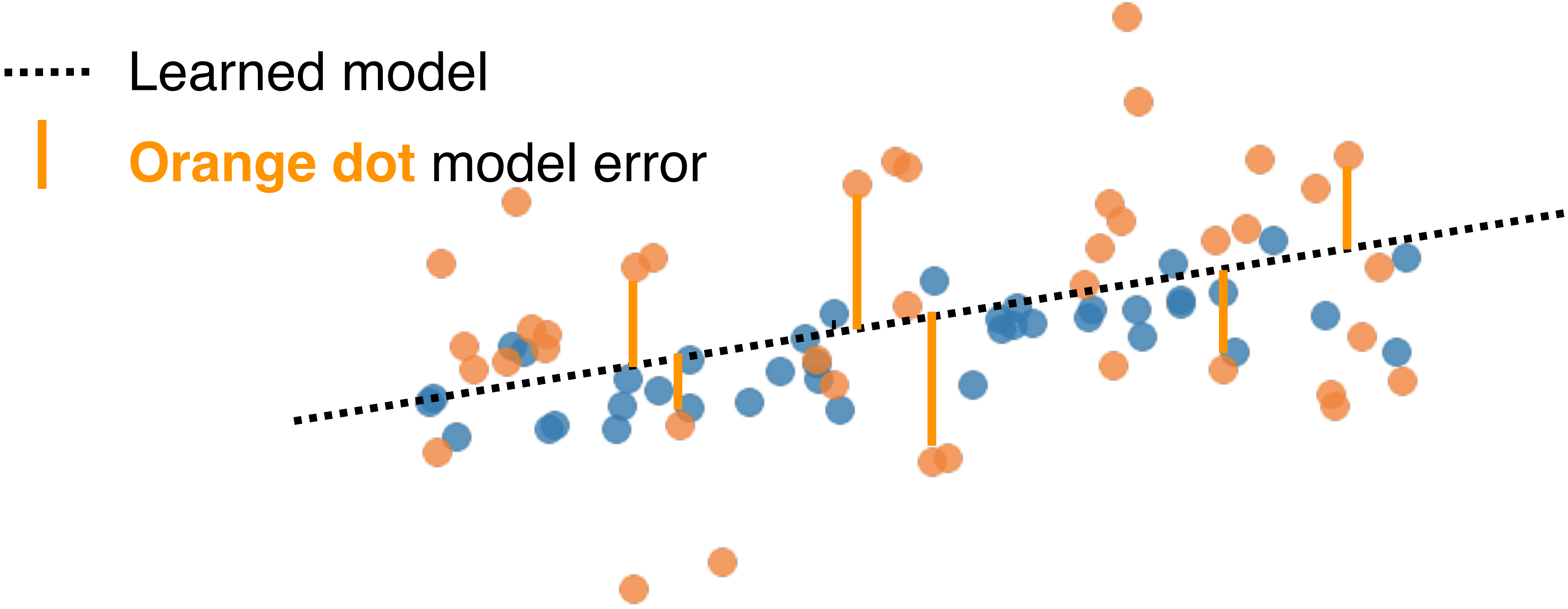
Why might my classifier be unfair?



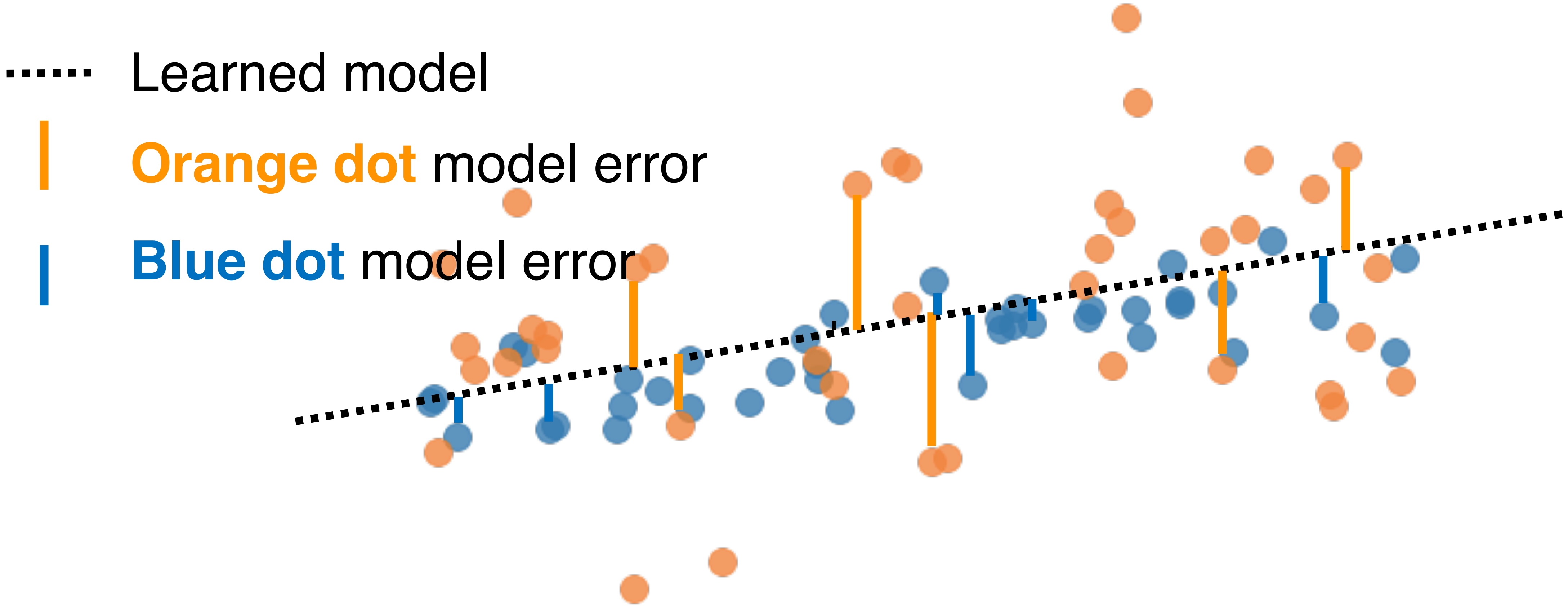
Why might my classifier be unfair?



Why might my classifier be unfair?



Why might my classifier be unfair?



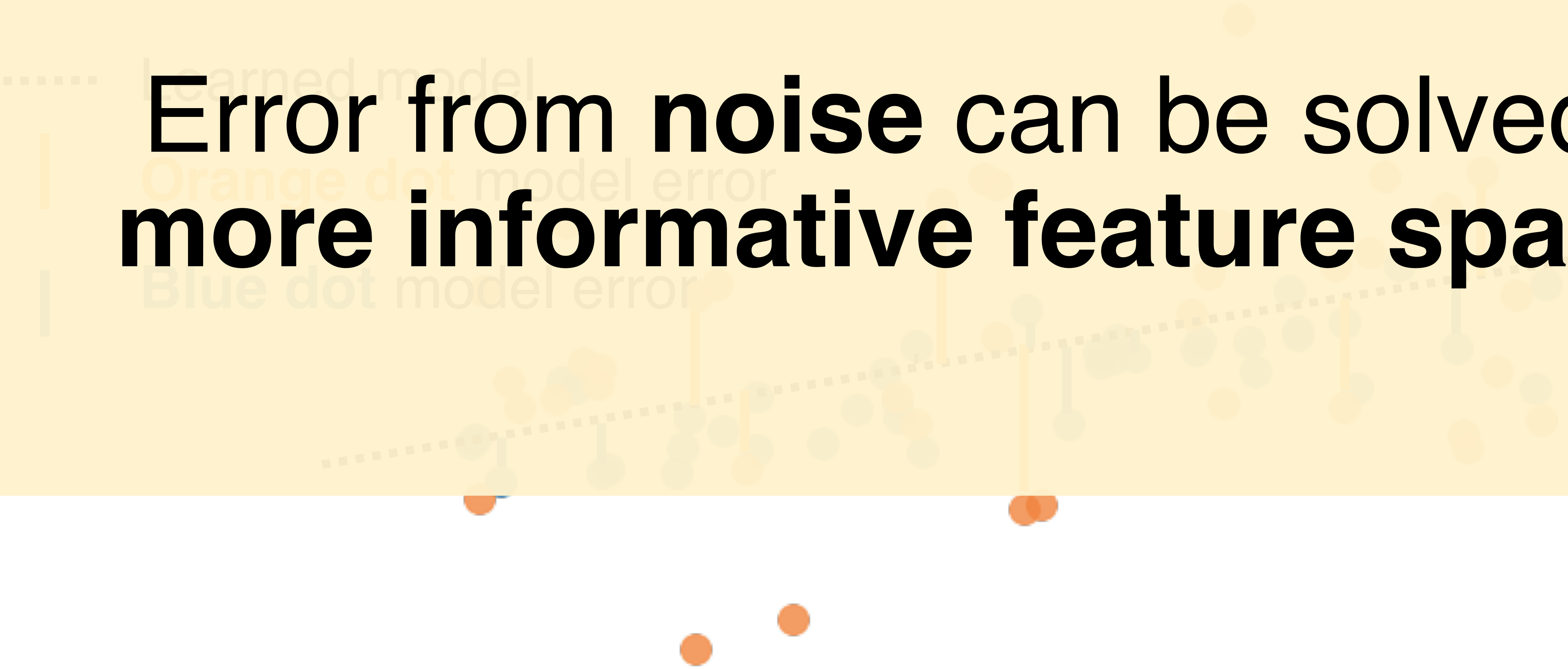
······ Learned model

| **Orange dot** model error

| **Blue dot** model error

Why might my classifier be unfair?

Error from **noise** can be solved by **more informative feature spaces**.



Contribution: Sources of unfairness

“unfairness”

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

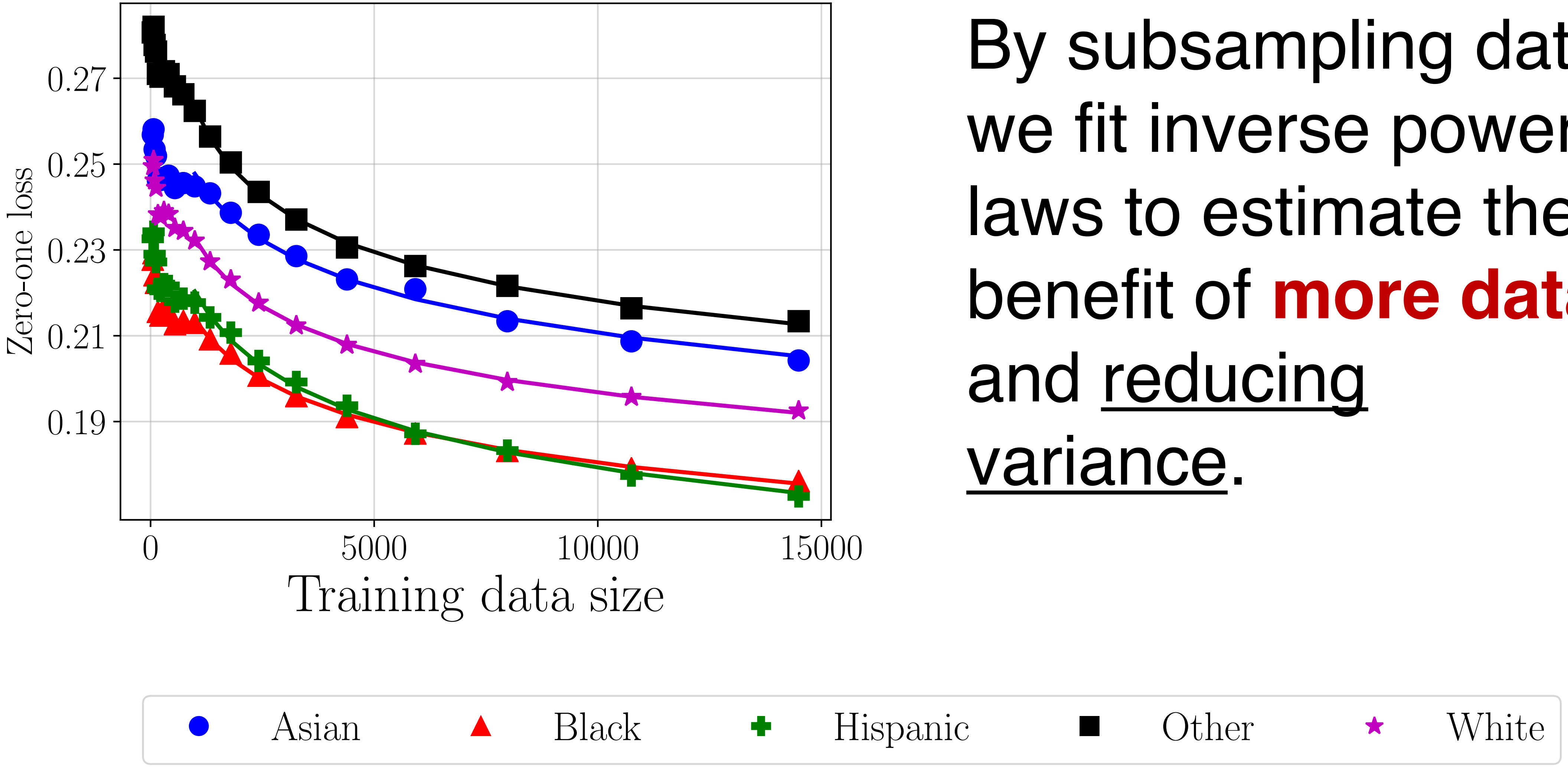
difference in bias **difference in variances** **difference in noise**

How can we realistically estimate $\bar{B}_a$, $\bar{V}_a$, and $\bar{N}_a$?

Contribution: Estimation Techniques

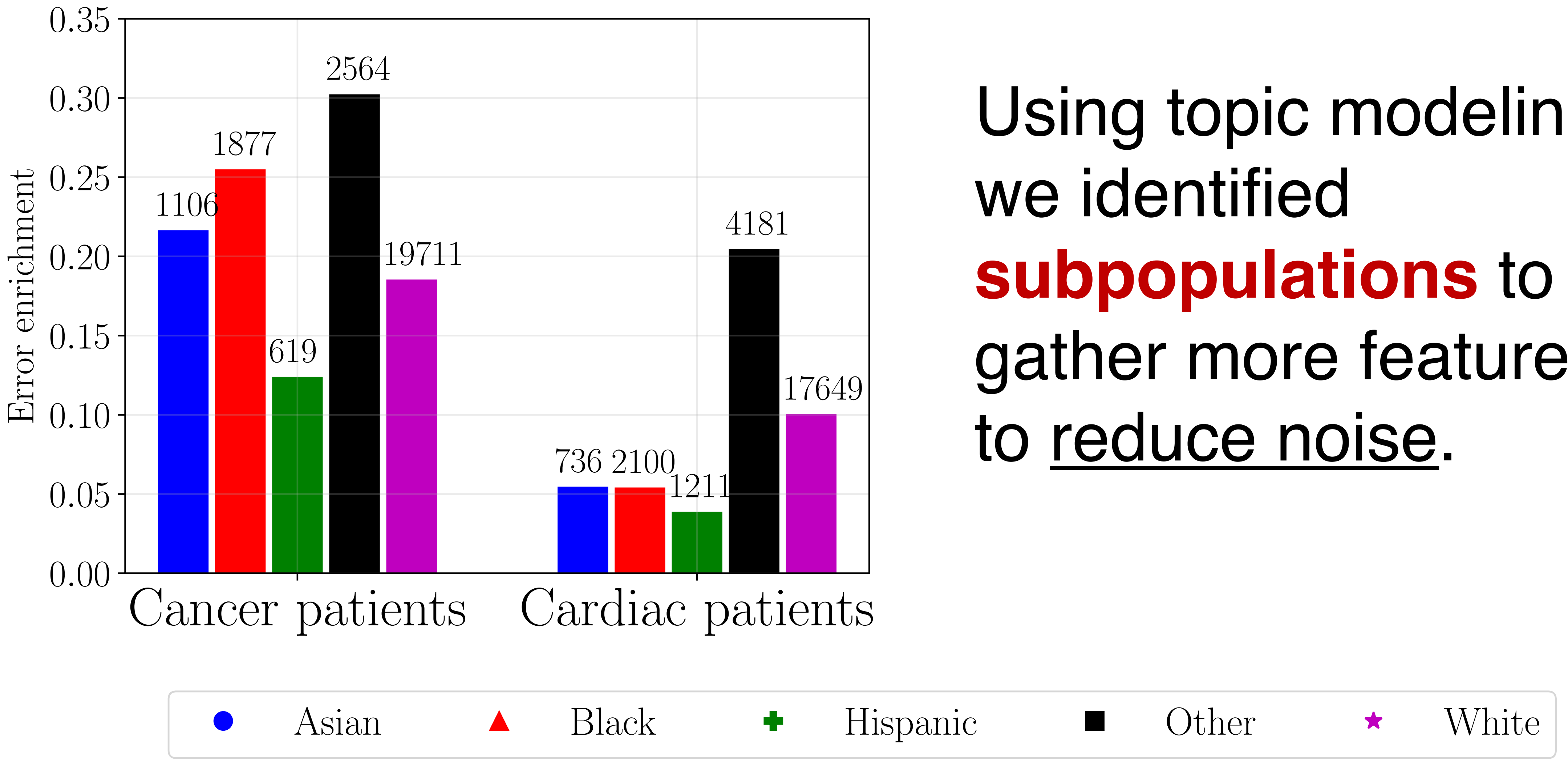
	Description	**How to estimate**	**How to fix**
Bias	How well model fits data	Experiment with model complexity	Change model class
Variance	How much sample size affects accuracy	Fit inverse power law from subsampling	Increase training data size
Noise	Error independent of model class and sample size	Estimate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

Mortality prediction from MIMIC-III clinical notes



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



Using topic modeling, we identified **subpopulations** to gather more features to reduce noise.

Collaboration: Independence Blue Cross

- Partnership with Independence Blue Cross, a health insurer based in Philadelphia

- Working to audit the **case management algorithms** and relevant subcomponents, including likelihood of hospitalization and high-risk pregnancy

How can we audit and address algorithmic bias?

1. **Decompose** sources of discrimination into statistical bias, variance, noise

2. Propose **practical actions** for detecting these components and mitigating discrimination

3. Techniques useful for other **high-stakes settings** including finance data, education data, or climate data

Machine Learning for Equitable Healthcare

1. Equity Audits for Machine Learning

Chen, Johansson, Sontag. (NeurIPS 2018)

Chen, Szolovits, Ghassemi. (AMA Journal of Ethics 2019)

Seyyed-Kalantari, Liu, McDermott, Chen, Ghassemi. (Nature Medicine 2021)

Chen, Agrawal, Horng, Sontag. (PSB 2020)

2. Machine Learning for Equity

Chen, Krishnan, Sontag. (AAAI 2022)

Chen, Joshi, Ghassemi. (Nature Medicine 2020)

Chen, Alsentzer, Park, Thomas, Gosangi, Gujrathi, Khurana. (PSB 2021)

Chen, Pierson, Rose, Joshi, Ferryman, Ghassemi. (Annual Reviews for Biomedical Data Science 2021)

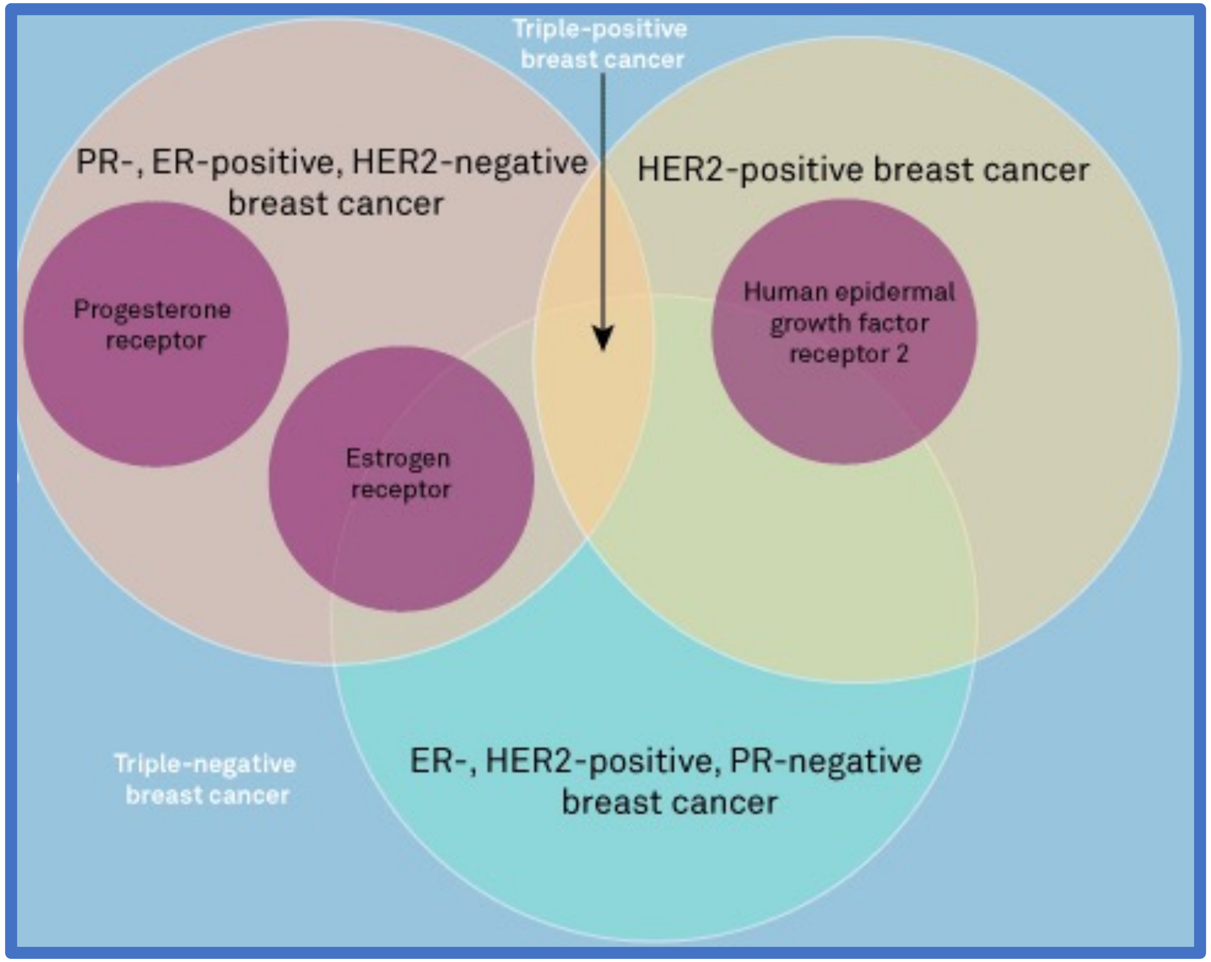
How can we build algorithms that account for differences in access to care?

Chen, Krishnan, Sontag, “Clustering Interval-Censored Time-Series for Disease Phenotyping,” *AAAI 2022.*

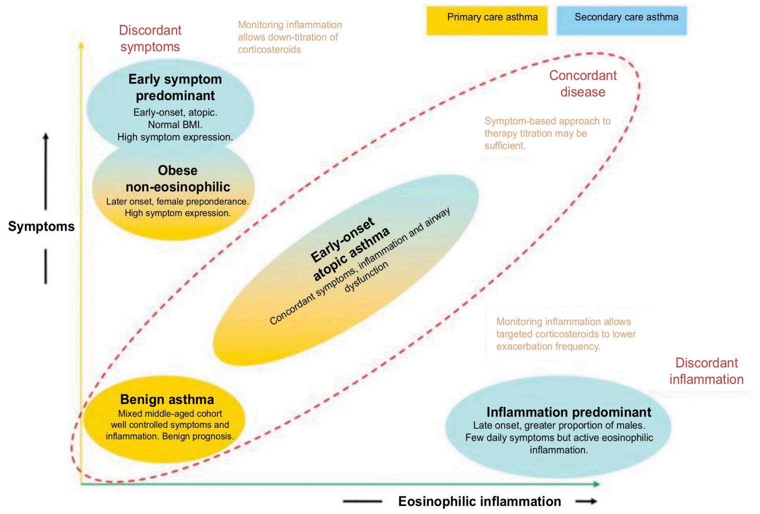
Systemic Health Disparities

- **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

Motivation: Disease Subtyping

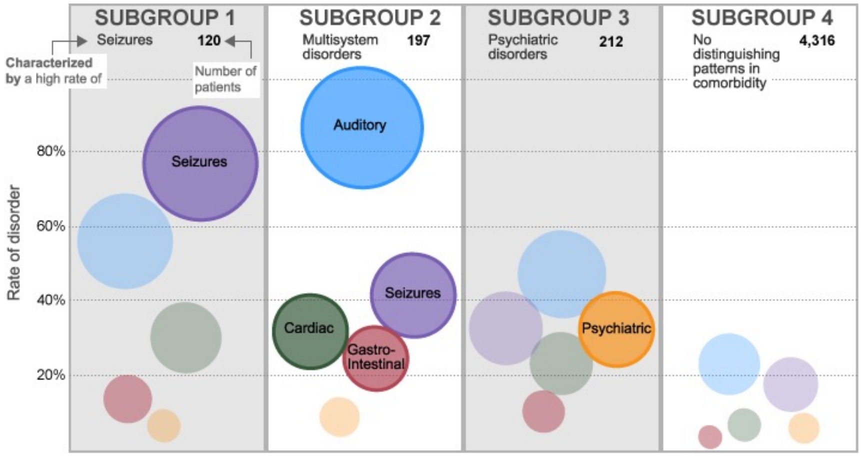

Triple-positive breast cancer
PR-, ER-positive, HER2-negative breast cancer
HER2-positive breast cancer
Progesterone receptor
Human epidermal growth factor receptor 2
Estrogen receptor
Triple-negative breast cancer
ER-, HER2-positive, PR-negative breast cancer

Many diseases are biologically heterogeneous despite a common diagnosis

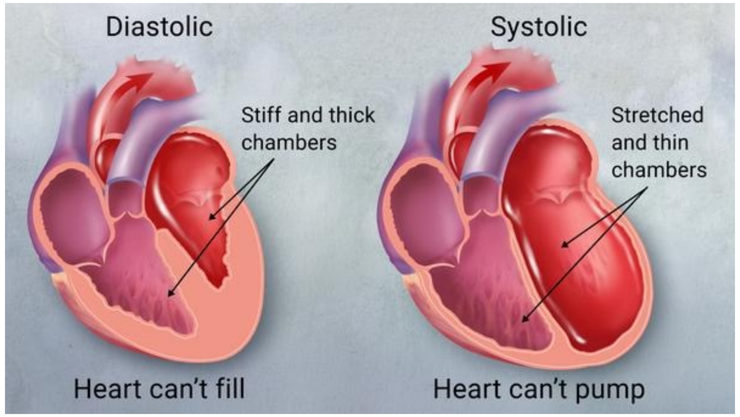


Primary care asthma
Secondary care asthma
Discordant symptoms
Monitoring inflammation allows down-titration of corticosteroids
Early symptom predominant
Early-onset, atopic. Normal BMI. High symptom expression.
Obese non-eosinophilic
Later onset, female preponderance. High symptom expression.
Symptoms
Concordant disease
Symptom-based approach to therapy titration may be sufficient.
Early-onset atopic asthma
Concordant symptoms, inflammation and airway dysfunction
Monitoring inflammation allows targeted corticosteroids to lower exacerbation frequency.
Discordant inflammation
Benign asthma
Mixed middle-aged cohort well controlled symptoms and inflammation. Benign prognosis.
Inflammation predominant
Late onset, greater proportion of males. Few daily symptoms but active eosinophilic inflammation.
Eosinophilic inflammation

Asthma



Autism



Heart Failure

Nissen et al, *Journal of Asthma and Allergy* 2018; Kohane et al, *PLoS One*, 2012; Mayo Clinic

Our goal is to find disease subtypes

- Subtypes are **“similar” patients**

- Subtypes are useful tools to design patient treatments or **expand understanding** of human health

- We want to account for **systemic health disparities**

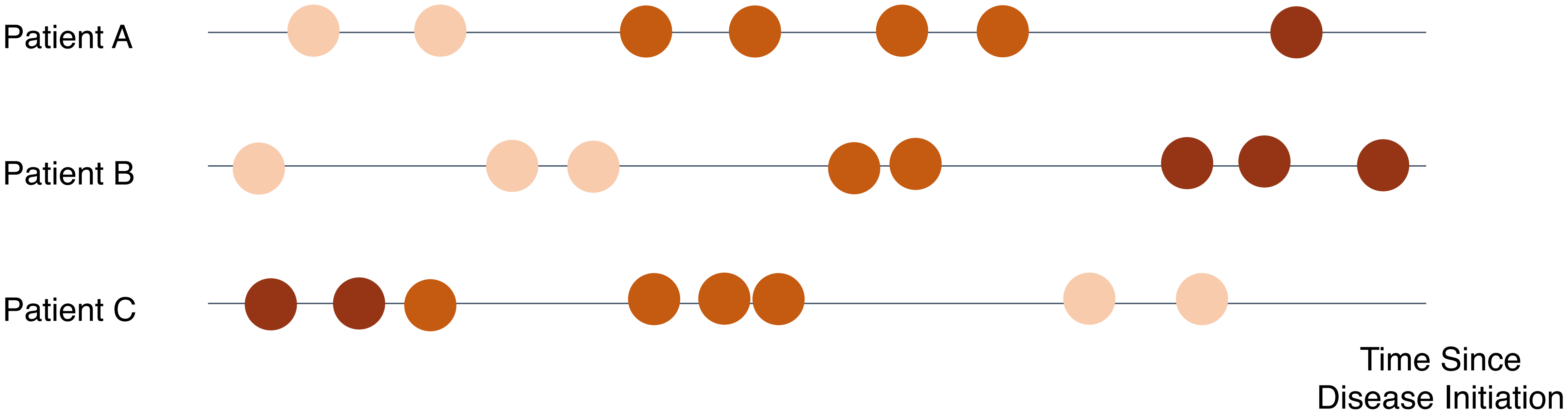
Idealized health data

Biomarker Severity

● = Mild

● = Moderate

● = Severe



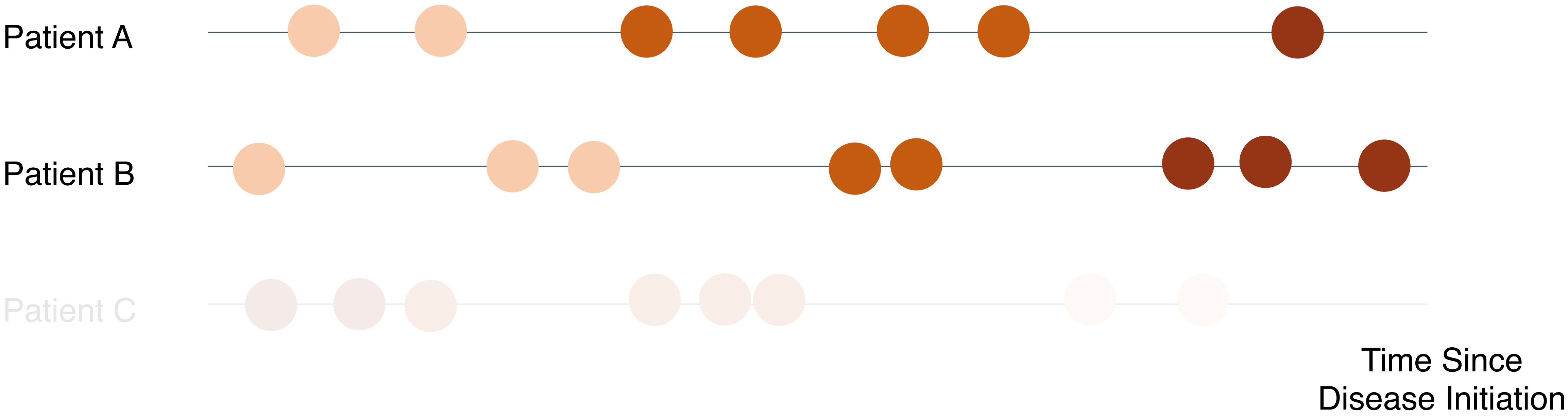
Idealized health data

Biomarker Severity

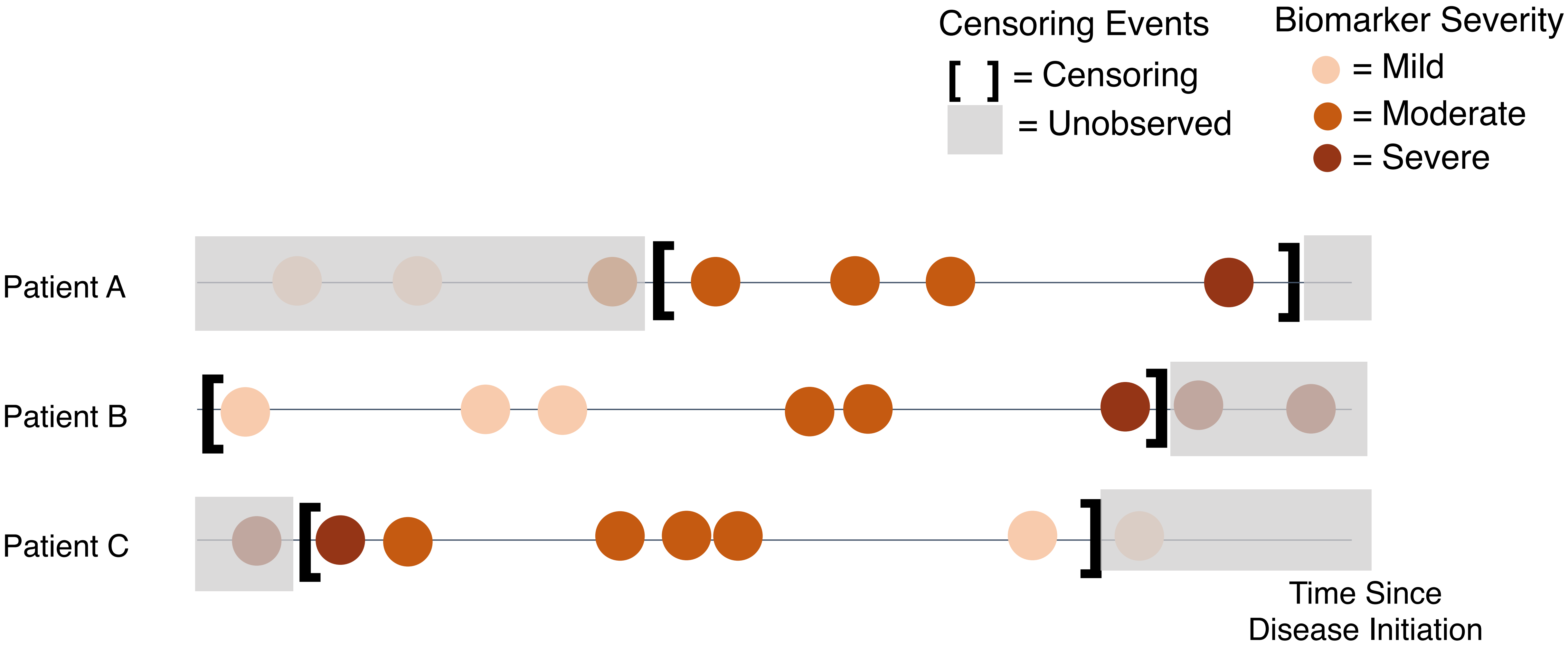
● = Mild

● = Moderate

● = Severe



A and B have very similar patient profiles! They should be assigned to the same cluster.



Censoring Events

[ ] = Censoring

▨ = Unobserved

Biomarker Severity

● = Mild

● = Moderate

● = Severe

Patient A

Patient B

Patient C

Time Since Disease Initiation

Data is collected in a **censored interval** for each patient

How can we learn disease subtyping?

- Option 1: Manually re-align the subtypes
  - Clinician time is expensive
  - Time-consuming for large datasets

- Option 2: Ignore alignment in learning subtypes
  - Subtypes may learn interval censoring instead of biologically interesting findings

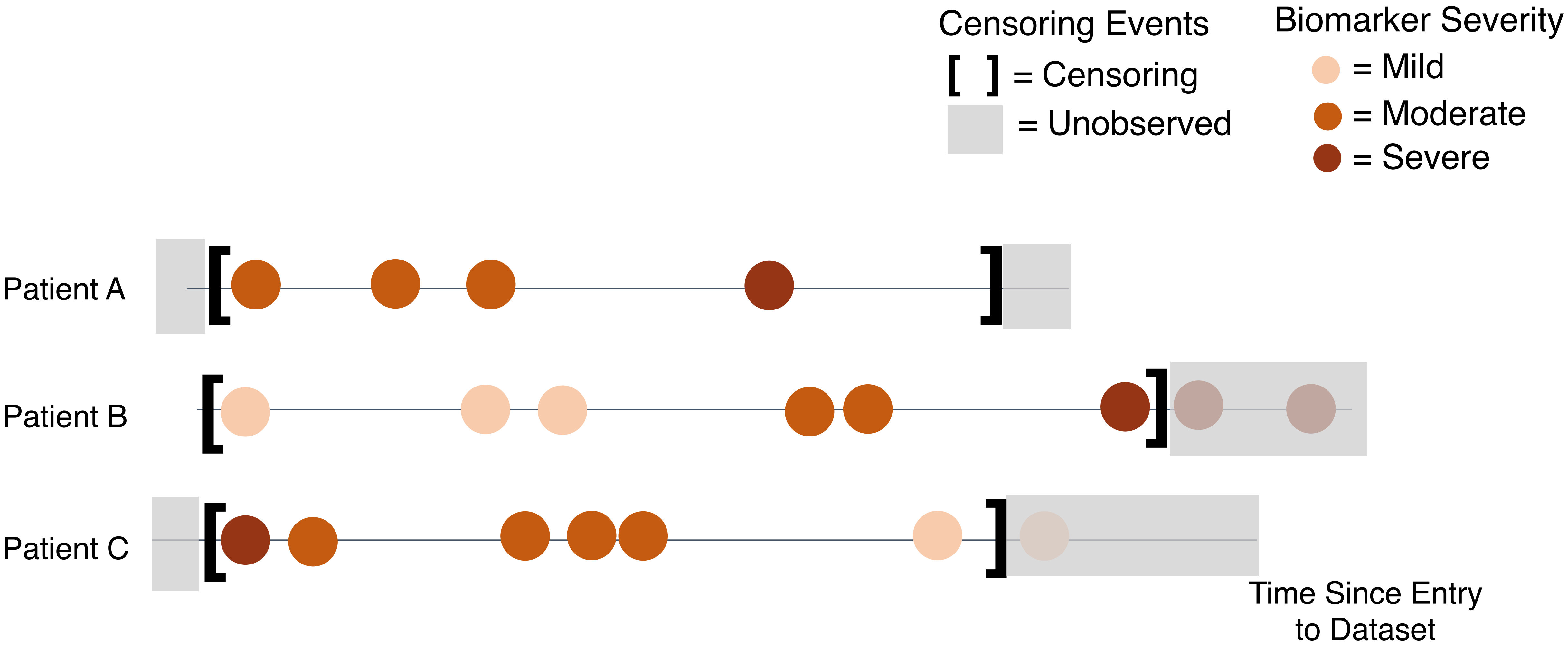
- Option 3: Incorporate alignment into a statistical model used for clustering
  - Explicitly *disentangle* between subtype identity and alignment

How can we learn disease subtyping?

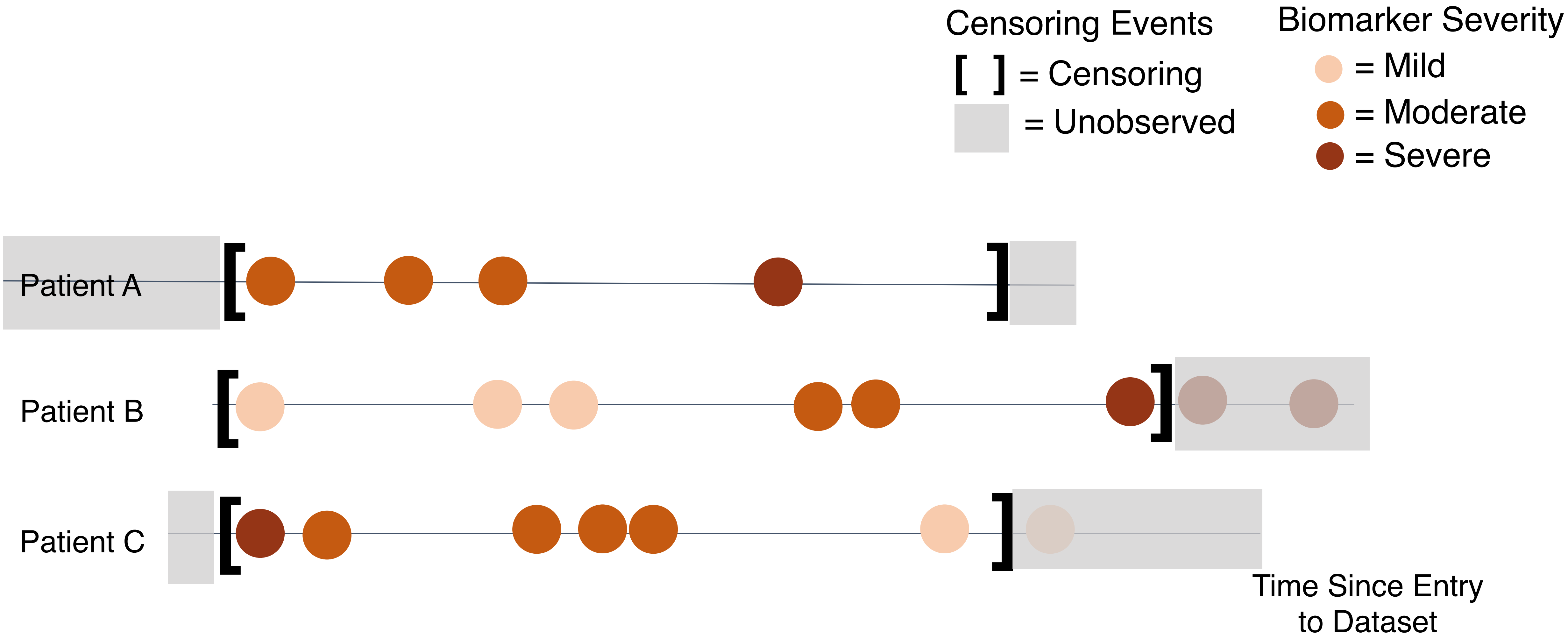
- Option 1: Manually re-align the subtypes
  - Clinician time is expensive
  - Time-consuming for large datasets

- **Option 2: Ignore alignment in learning subtypes**
  - **Subtypes may learn interval censoring instead of biologically interesting findings**

- Option 3: Incorporate alignment into a statistical model used for clustering
  - Explicitly *disentangle* between subtype identity and alignment



Option 2: Assume time-series start at the same stage of disease progression.



Option 2: We may inadvertently cluster based on **disease stage** instead of biologically interesting clusters.

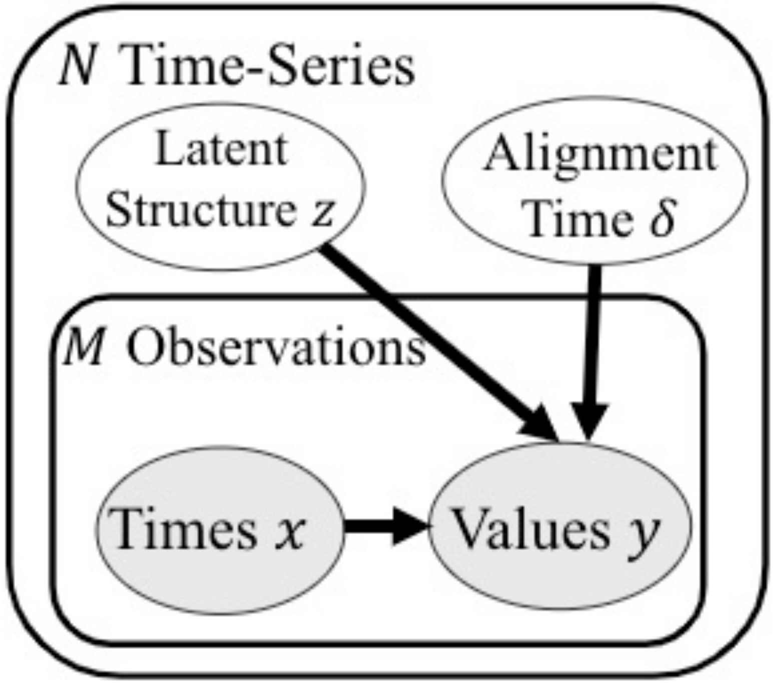
How can we learn disease subtyping?

- Option 1: Manually re-align the subtypes
  - Clinician time is expensive
  - Time-consuming for large datasets

- Option 2: Ignore alignment in learning subtypes
  - Subtypes may learn interval censoring instead of biologically interesting findings

- **Option 3: Incorporate alignment into a statistical model used for clustering**
  - **Explicitly *disentangle* between subtype identity and alignment**

SubLign is a deep generative model that jointly learns patient subtype and alignment



Variational inference to approximate likelihood

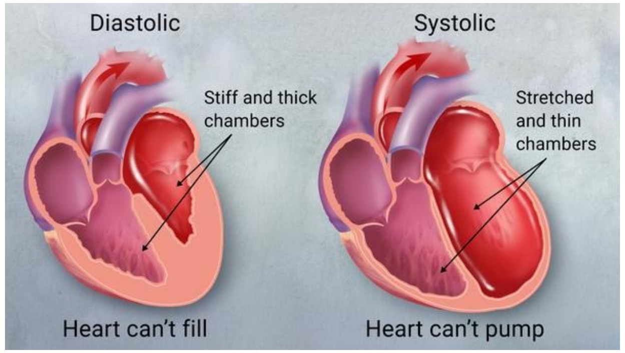
$$P_{\theta_1} = P_{\theta_2} \quad \Rightarrow \quad \theta_1 = \theta_2$$

$$\text{for all } \theta_1, \theta_2 \in \Theta.$$

Identifiability results show sufficient conditions

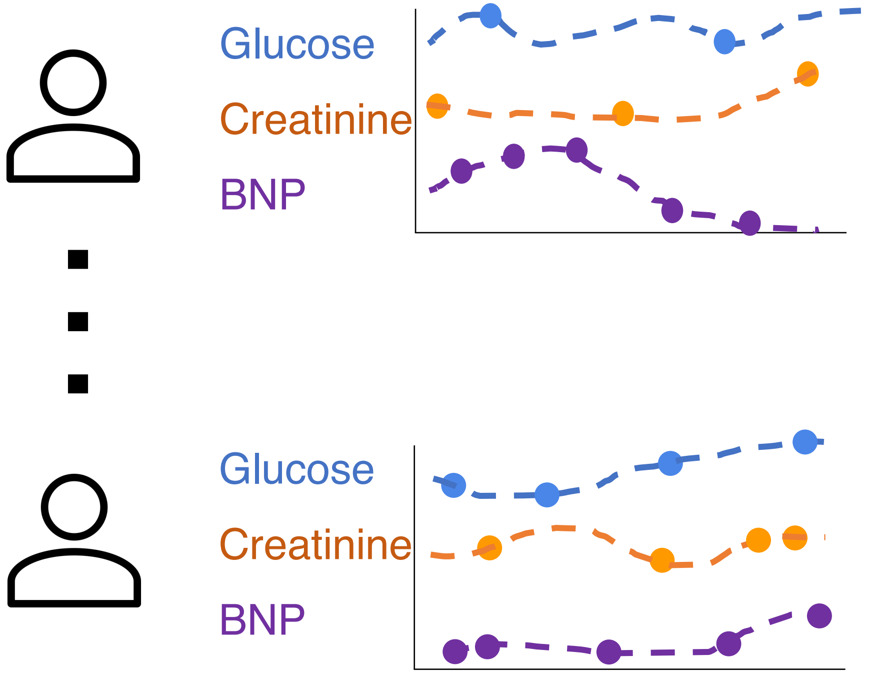
PARKINSON'S PROGRESSION MARKERS INITIATIVE

Play a Part in Parkinson's Research



Experiment results recover known clinical findings

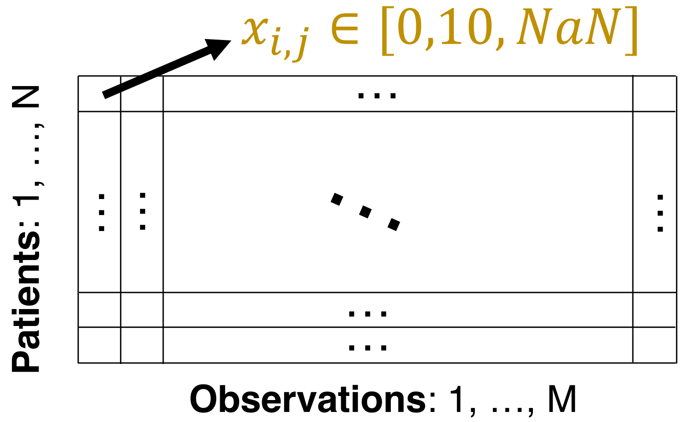
How can we model the clinical data?

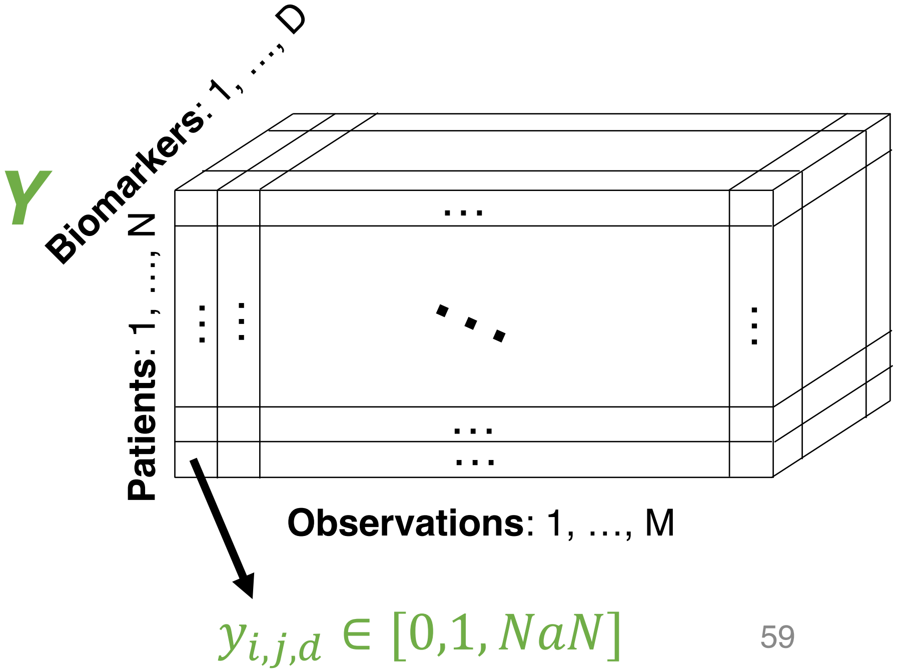


Irregularly Sampled
Multivariate Time-Series

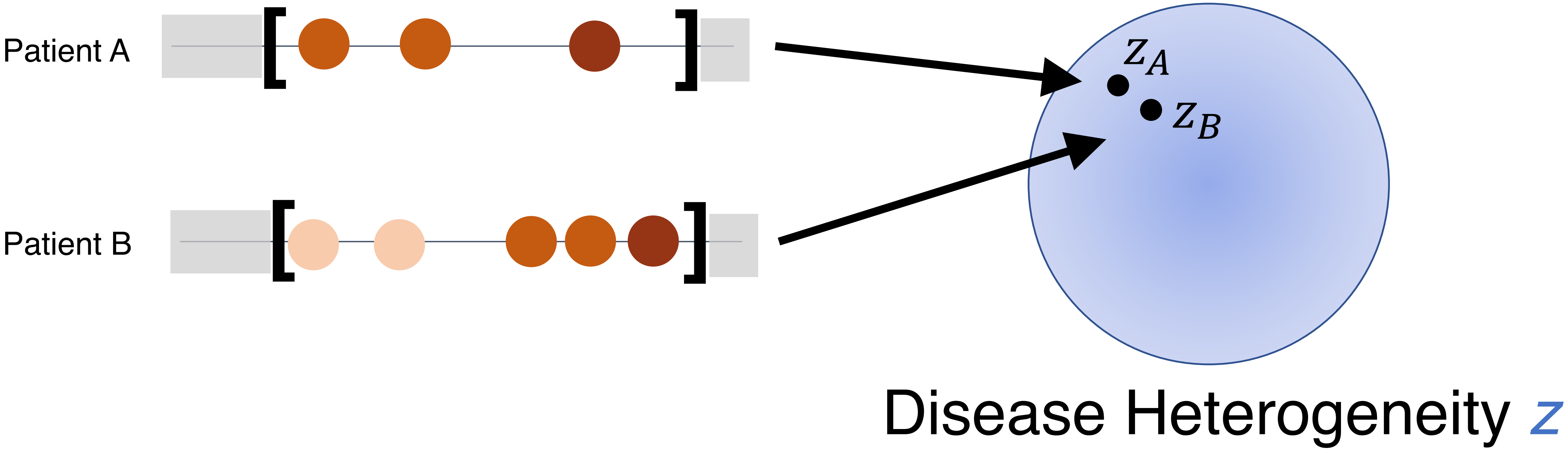
Observed
Times ***X***

Biomarkers ***Y***



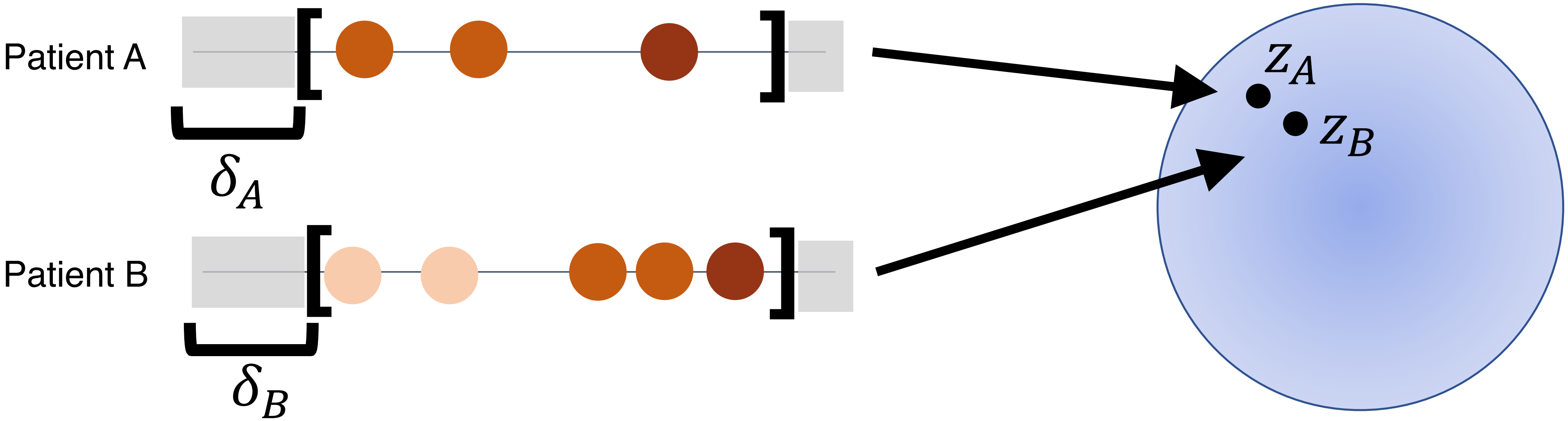


SubLign: Subtype and Alignment



Similar patients are close together in **latent representation** space.
Subtypes can be found by clustering the continuous space.

SubLign: Subtype and Alignment



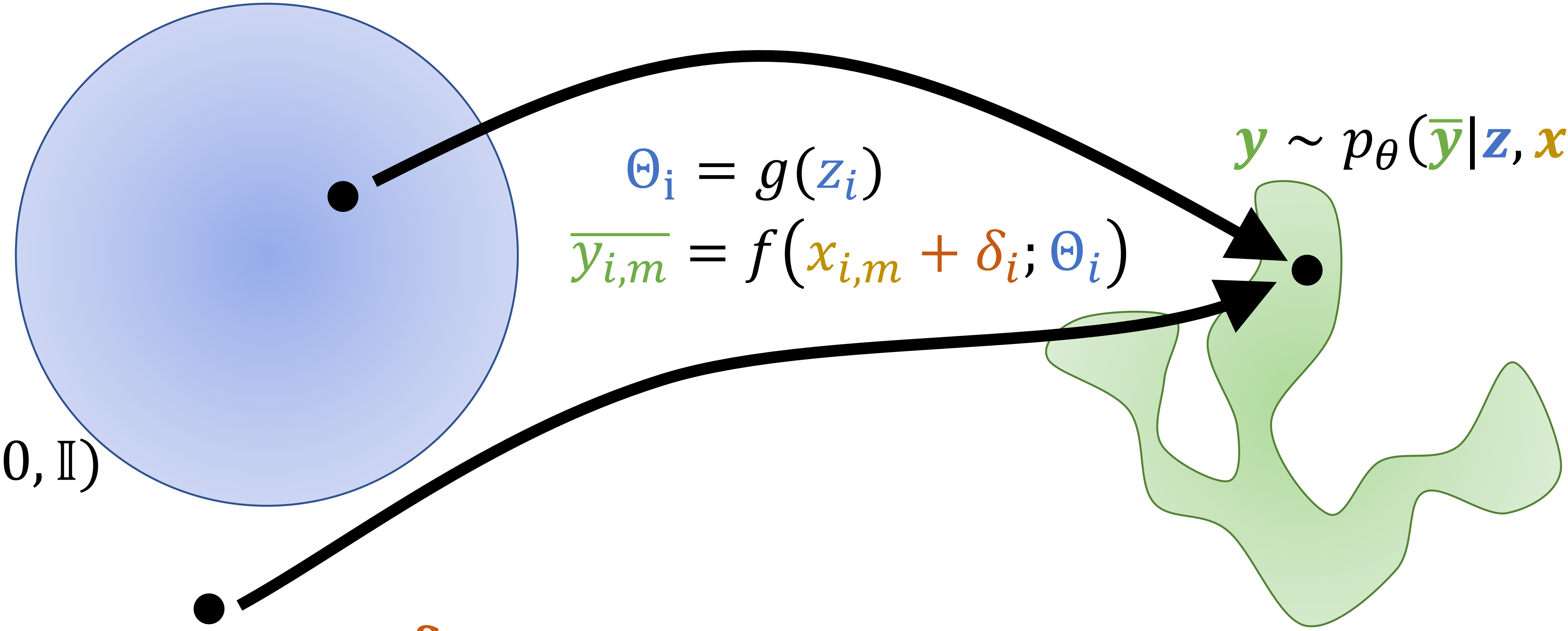
Alignment Value δ

We want to learn heterogeneity that corrects for a **latent alignment** value

SubLign Data Generation

Observed Times X

Biomarkers Y



Disease Heterogeneity $\boldsymbol{z}$

$$\Theta_{\mathrm{i}} = g(z_i)$$

$$\overline{y_{i,m}} = f\left(x_{i,m} + \delta_i ; \Theta_i\right)$$

$$\boldsymbol{y} \sim p_\theta(\overline{\boldsymbol{y}} | \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\delta})$$

$$\boldsymbol{z} \sim N(0, \mathbb{I})$$

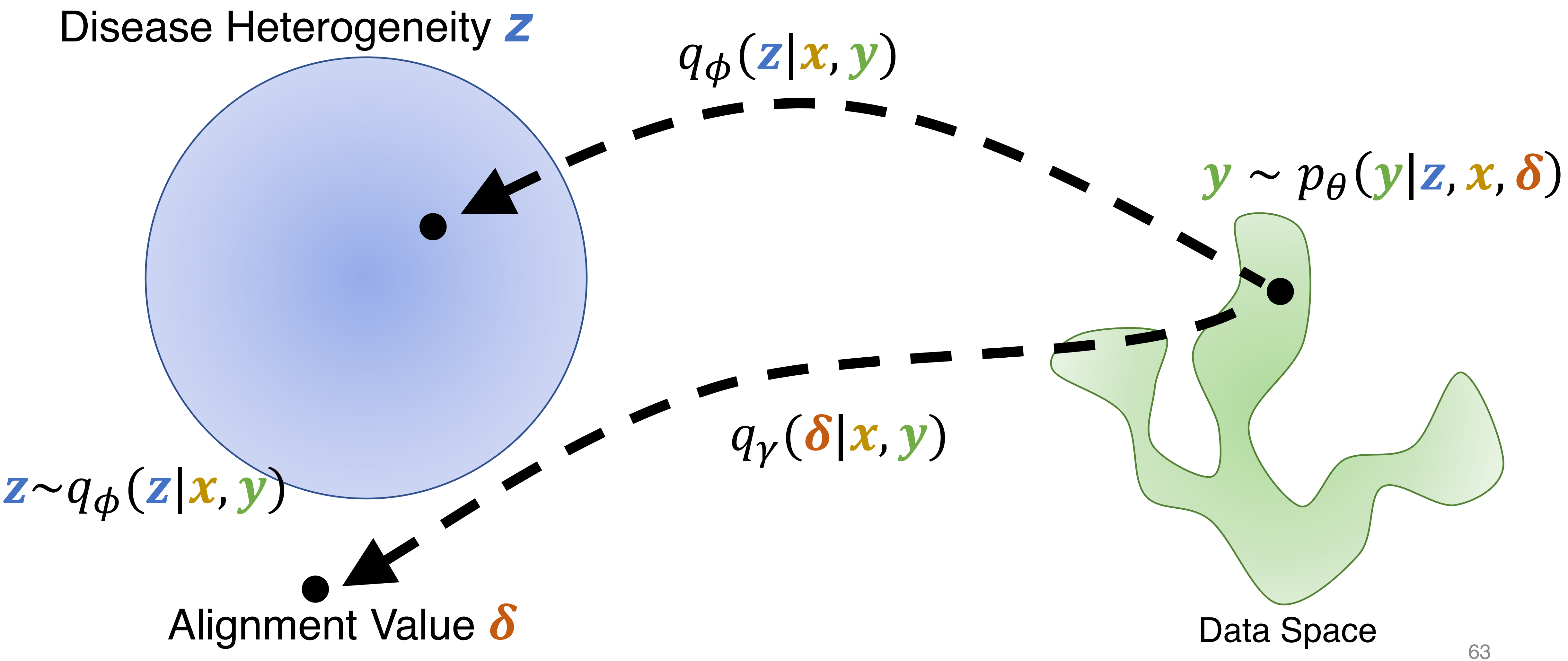
Alignment Value $\boldsymbol{\delta}$

Data Space

SubLign Representation Inference

Observed Times X

Biomarkers Y



Disease Heterogeneity $\boldsymbol{z}$

$q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})$

$\boldsymbol{y} \sim p_\theta(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{x},\boldsymbol{\delta})$

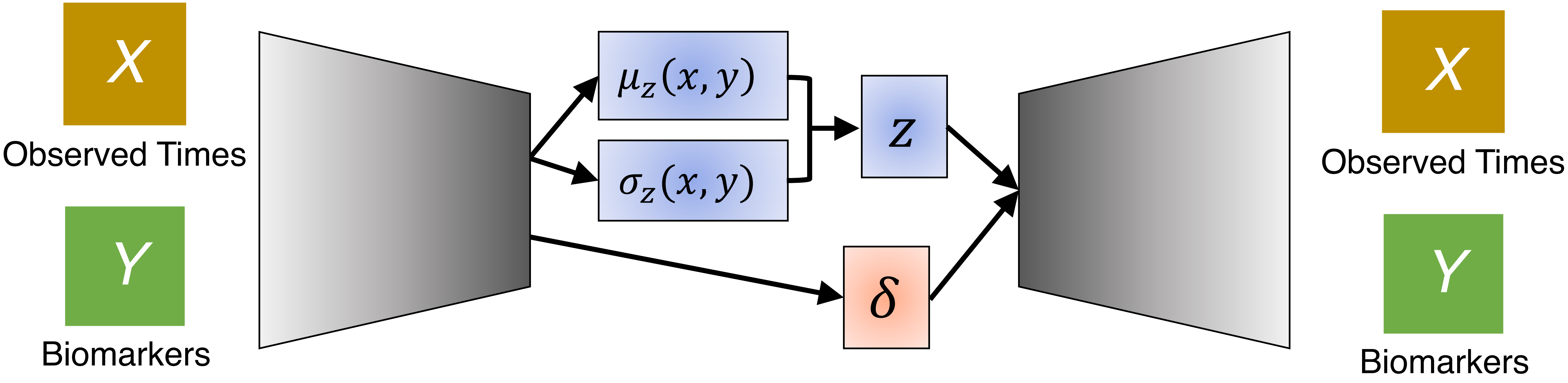
$q_\gamma(\boldsymbol{\delta}|\boldsymbol{x},\boldsymbol{y})$

$\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})$

Alignment Value $\boldsymbol{\delta}$

Data Space

SubLign Model Architecture



Recurrent Neural Network Encoder

$$q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})$$

$$q_\gamma(\boldsymbol{\delta}|\boldsymbol{x},\boldsymbol{y})$$

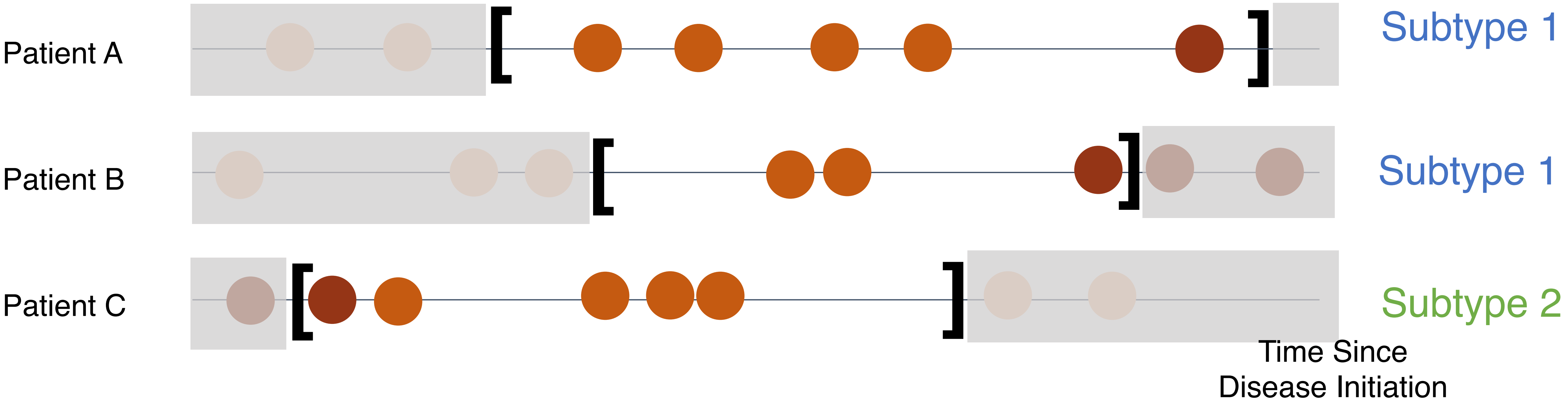
Neural Network Decoder

$$\Theta_\mathrm{i} = g(z_i)$$

$$\overline{y_{i,m}} = f\left(x_{i,m} + \delta_i; \Theta_i\right)$$

$$\boldsymbol{y} \sim p_\theta(\overline{\boldsymbol{y}}|\boldsymbol{z},\boldsymbol{x},\boldsymbol{\delta})$$

Identifiability: When can we recover the correct subtypes?



A, B, and C look so similar that it might be impossible to discover the correct subtypes.

Censoring Events

[ ] = Censoring

▨ = Unobserved

Biomarker Severity

● = Mild

● = Moderate

● = Severe

Identifiability: When can we recover the correct subtypes?

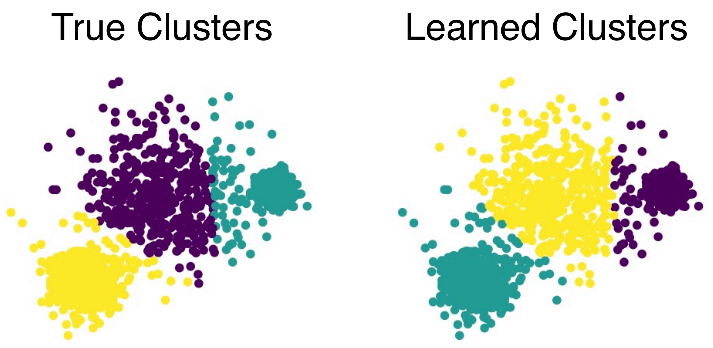
- Theoretical question: *Are there situations where we can reliably disentangle subtype from alignment time?*

Identifiability: When can we recover the correct subtypes?

- Theoretical question: *Are there situations where we can reliably disentangle subtype from alignment time?*

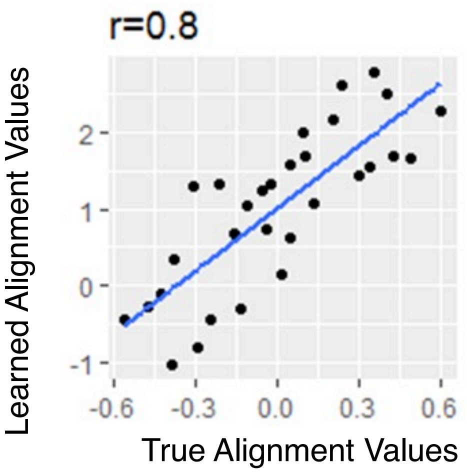
- Yes! We can prove identifiability under a noiseless, parameterized version of SubLign

How do we evaluate SubLign?



1. Clustering
   - **Adjusted Rand index** (ARI): quantitative measure of label concordance
   - We lack ground truth in baseline data, so we use baseline data (not included in SubLign) to validate known clinical findings

2. Alignment
   - **Swaps metric**: How many swaps to get values in correct order, as a percent?
   - **Pearson correlation coefficient**: How correlated are the aligned values and the true values?



How well does SubLign recover cluster and alignment values on synthetic data?

	Cluster performance	Alignment performance	Alignment performance
Model	ARI ↑	Swaps ↓	Pearson ↑

How well does SubLign recover cluster and alignment values on synthetic data?

	Cluster performance	Alignment performance	Alignment performance
Model	ARI ↑	Swaps ↓	Pearson ↑
SubLign	**0.94 ± 0.02**	**0.09 ± 0.00**	**0.85 ± 0.04**
SubNoLign	0.81 ± 0.21	–	–

SubLign outperforms deep generative model **without alignment**

How well does SubLign recover cluster and alignment values on synthetic data?

	Cluster performance	Alignment performance	Alignment performance
Model	ARI ↑	Swaps ↓	Pearson ↑
SubLign	**0.94 ± 0.02**	**0.09 ± 0.00**	**0.85 ± 0.04**
KMeans+Loss	0.67 ± 0.04	0.21 ± 0.03	0.49 ± 0.01

SubLign outperforms **greedy** approach of clustering then aligning

How well does SubLign recover cluster and alignment values on synthetic data?

	Cluster performance	Alignment performance	Alignment performance
Model	ARI ↑	Swaps ↓	Pearson ↑
SubLign	**0.94 ± 0.02**	**0.09 ± 0.00**	**0.85 ± 0.04**
SuStaIn	0.66 ± 0.02	0.16 ± 0.00	0.30 ± 0.02
PAGA	0.32 ± 0.05	0.52 ± 0.07	0.04 ± 0.20

SubLign outperforms algorithms assuming **cross-sectional** data and **linear** data

How well does SubLign recover cluster and alignment values on synthetic data?

	Cluster performance	Alignment performance	Alignment performance
MODEL	ARI ↑	SWAPS ↓	PEARSON ↑
SubLign	**0.94 ± 0.02**	**0.09 ± 0.00**	**0.85 ± 0.04**
BayLong	0.19 ± 0.18	0.48 ± 0.00	0.01 ± 0.02

SubLign outperforms algorithm with Bayesian model assumptions

How well does SubLign recover cluster and alignment values on synthetic data?

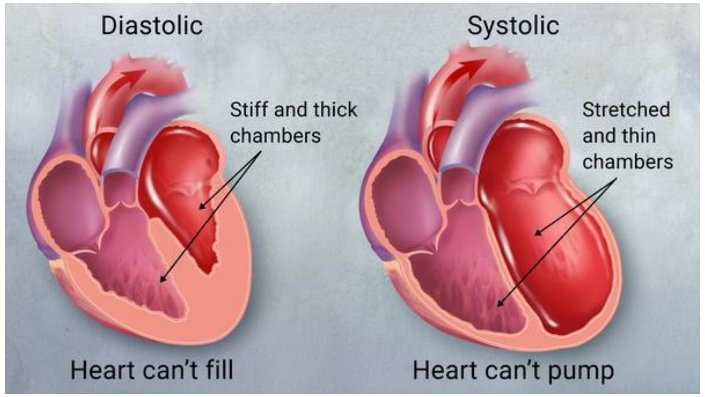
SubLign outperforms baselines! →

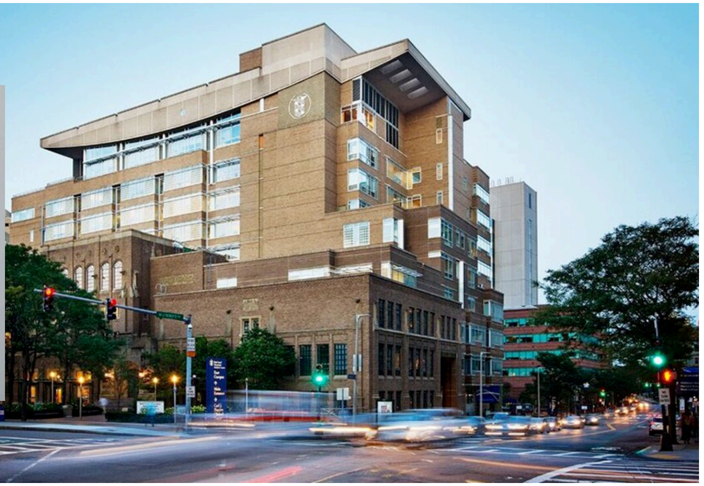
	Cluster performance	Alignment performance	Alignment performance
MODEL	ARI ↑	SWAPS ↓	PEARSON ↑
SubLign	**0.94 ± 0.02**	**0.09 ± 0.00**	**0.85 ± 0.04**
SubNoLign	0.81 ± 0.21	–	–
KMeans+Loss	0.67 ± 0.04	0.21 ± 0.03	0.49 ± 0.01
SuStaIn	0.66 ± 0.02	0.16 ± 0.00	0.30 ± 0.02
BayLong	0.19 ± 0.18	0.48 ± 0.00	0.01 ± 0.02
PAGA	0.32 ± 0.05	0.52 ± 0.07	0.04 ± 0.20

(Including 4 other baselines)

How well does SubLign recover cluster and alignment values on **clinical** data?

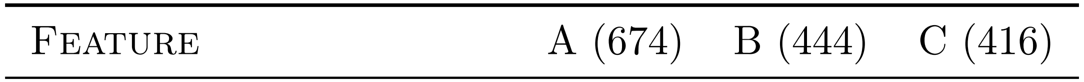
- **Observational** data from Beth Israel Deaconess Medical Center (Boston)
- 1,534 **heart failure** patients suffering from heart failure
- 12 features over time based on on echocardiograms
- **Validate** subtypes based on demographic and diagnosis data





How well does SubLign recover cluster and alignment values on **clinical** data?

Clusters learned by SubLign are reasonably sized



Feature	A (674)	B (444)	C (416)

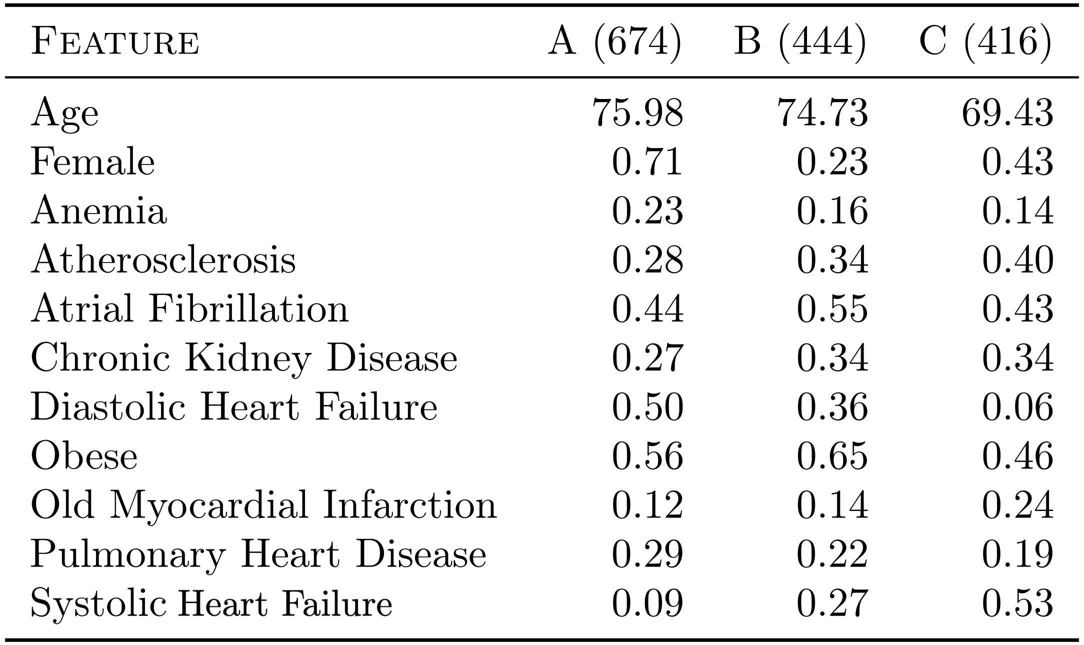
How well does SubLign recover cluster and alignment values on **clinical** data?

11 features (of 24) are **statistically significant** based on an ANOVA test with p<0.05 with a Benjamini-Hochberg correction

Feature	A (674)	B (444)	C (416)
Age			
Female			
Anemia			
Atherosclerosis			
Atrial Fibrillation			
Chronic Kidney Disease			
Diastolic Heart Failure			
Obese			
Old Myocardial Infarction			
Pulmonary Heart Disease			
Systolic Heart Failure			

How well does SubLign recover cluster and alignment values on **clinical** data?

We report cluster means for each feature

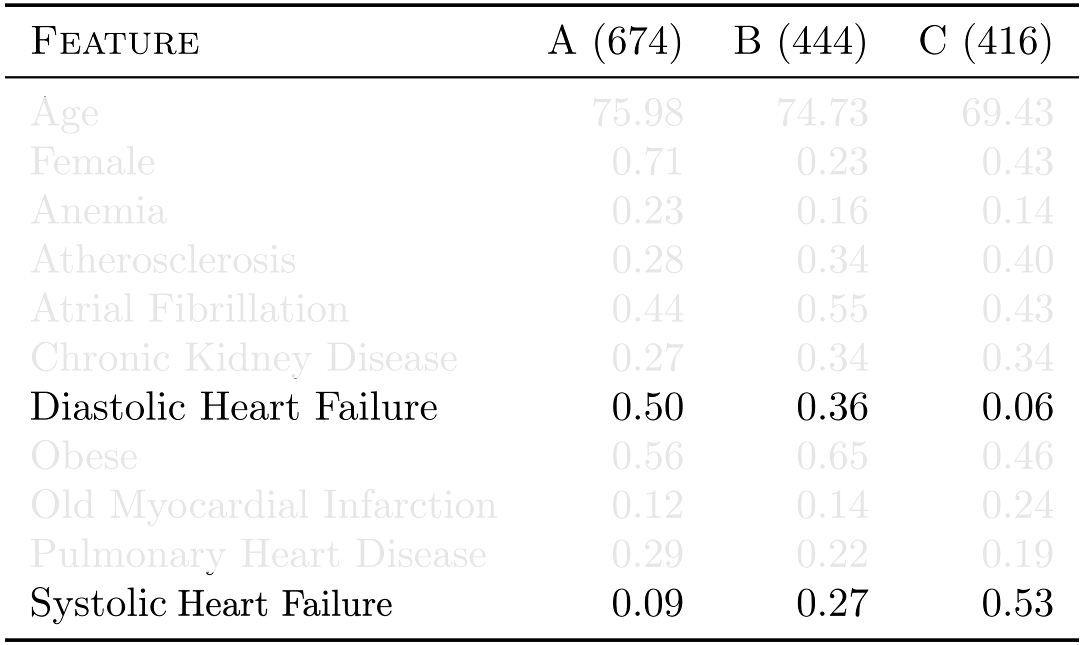


Feature	A (674)	B (444)	C (416)
Age	75.98	74.73	69.43
Female	0.71	0.23	0.43
Anemia	0.23	0.16	0.14
Atherosclerosis	0.28	0.34	0.40
Atrial Fibrillation	0.44	0.55	0.43
Chronic Kidney Disease	0.27	0.34	0.34
Diastolic Heart Failure	0.50	0.36	0.06
Obese	0.56	0.65	0.46
Old Myocardial Infarction	0.12	0.14	0.24
Pulmonary Heart Disease	0.29	0.22	0.19
Systolic Heart Failure	0.09	0.27	0.53

How well does SubLign recover cluster and alignment values on **clinical** data?

Diastolic (A) and systolic (C) heart failure are known subtypes.

B has patient from both diastolic and systolic heart failure.

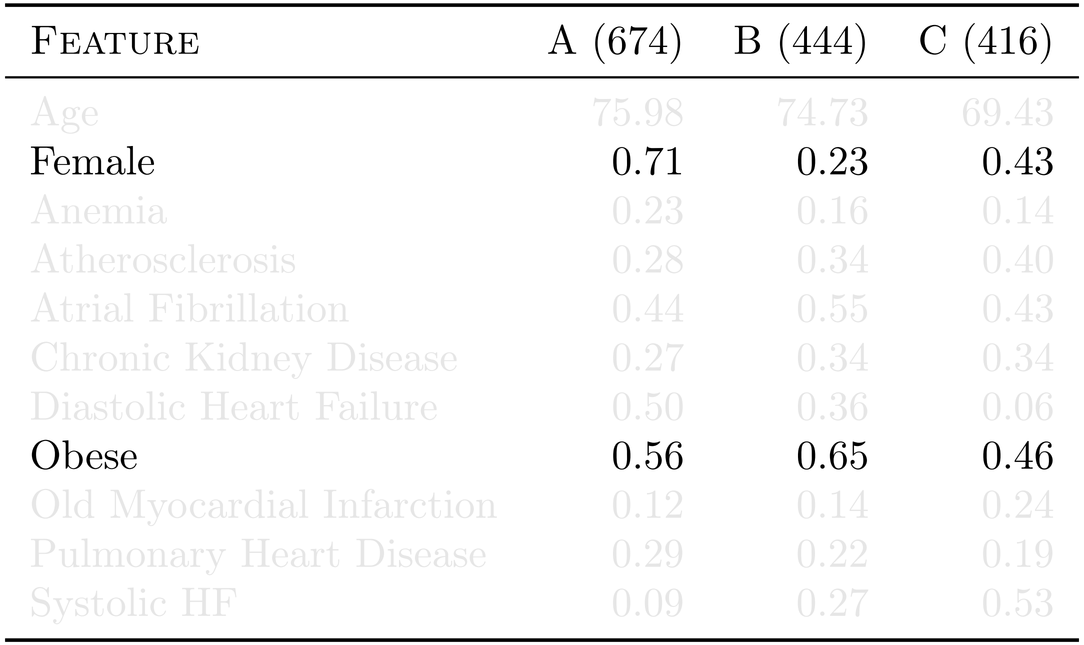


FEATURE	A (674)	B (444)	C (416)
Age	75.98	74.73	69.43
Female	0.71	0.23	0.43
Anemia	0.23	0.16	0.14
Atherosclerosis	0.28	0.34	0.40
Atrial Fibrillation	0.44	0.55	0.43
Chronic Kidney Disease	0.27	0.34	0.34
Diastolic Heart Failure	0.50	0.36	0.06
Obese	0.56	0.65	0.46
Old Myocardial Infarction	0.12	0.14	0.24
Pulmonary Heart Disease	0.29	0.22	0.19
Systolic Heart Failure	0.09	0.27	0.53

Shah et al, Phenomapping for novel classification of heart failure with preserved ejection fraction, Circulation 2014.

How well does SubLign recover cluster and alignment values on **clinical** data?

Clinical literature suggests that **women**[1] and **obese**[2] patients may manifest heart failure differently

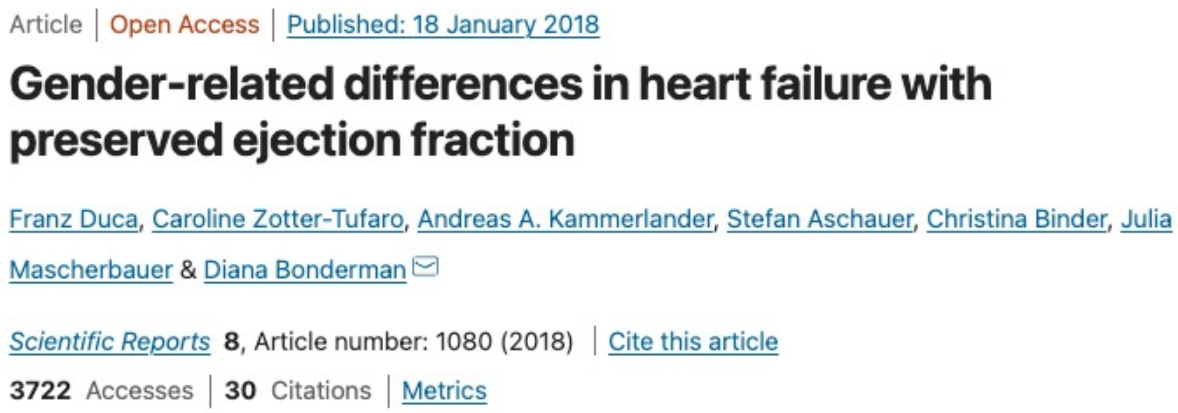


FEATURE	A (674)	B (444)	C (416)
Age	75.98	74.73	69.43
Female	0.71	0.23	0.43
Anemia	0.23	0.16	0.14
Atherosclerosis	0.28	0.34	0.40
Atrial Fibrillation	0.44	0.55	0.43
Chronic Kidney Disease	0.27	0.34	0.34
Diastolic Heart Failure	0.50	0.36	0.06
Obese	0.56	0.65	0.46
Old Myocardial Infarction	0.12	0.14	0.24
Pulmonary Heart Disease	0.29	0.22	0.19
Systolic HF	0.09	0.27	0.53

[1] Duca et al, Scientific Reports 2018. [2] Tadic and Cuspidi, Heart Failure Reviews 2019.

How well does SubLign recover cluster and alignment values on **clinical** data?

Clinical literature suggests that **women**[1] and **obese**[2] patients may manifest heart failure differently



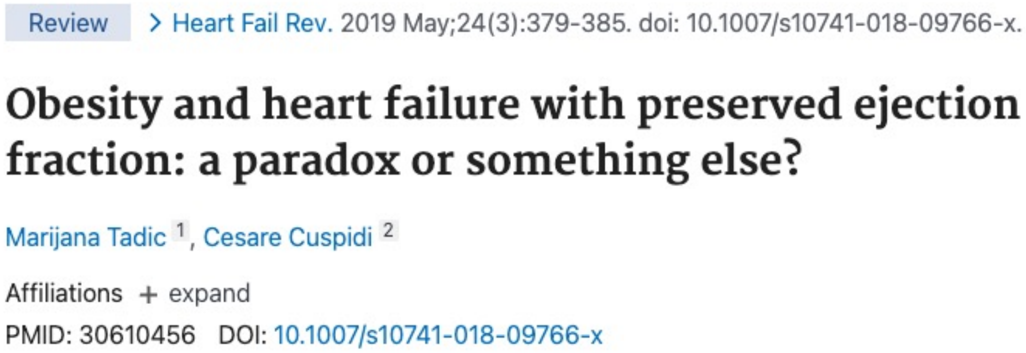
Article | Open Access | Published: 18 January 2018

Gender-related differences in heart failure with preserved ejection fraction

Franz Duca, Caroline Zotter-Tufaro, Andreas A. Kammerlander, Stefan Aschauer, Christina Binder, Julia Mascherbauer & Diana Bonderman

Scientific Reports **8**, Article number: 1080 (2018) | Cite this article

3722 Accesses | **30** Citations | Metrics



Review > Heart Fail Rev. 2019 May;24(3):379-385. doi: 10.1007/s10741-018-09766-x.

Obesity and heart failure with preserved ejection fraction: a paradox or something else?

Marijana Tadic [1], Cesare Cuspidi [2]

Affiliations + expand

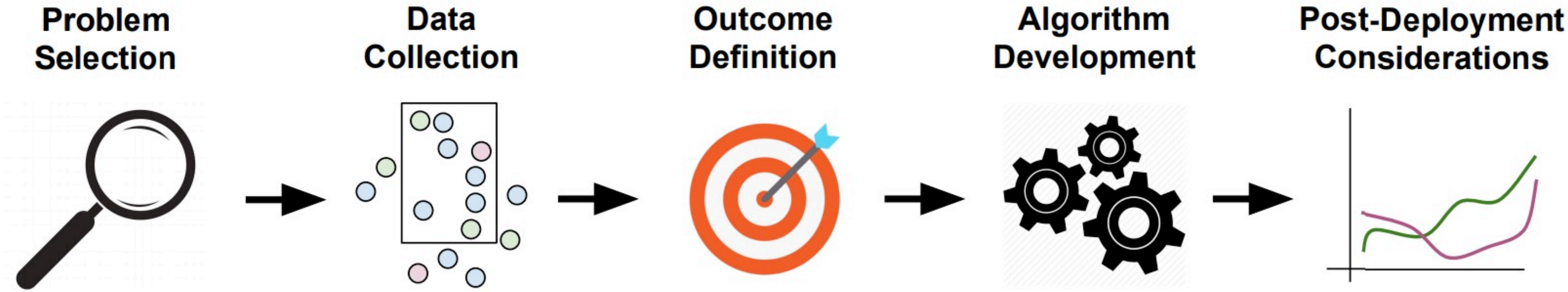
PMID: 30610456 DOI: 10.1007/s10741-018-09766-x

[1] Duca et al, Scientific Reports 2018. [2] Tadic and Cuspidi, Heart Failure Reviews 2019.

How can we accommodate differences in access to care?

1. Model access to care as a **latent variable**
2. Design **deep generative model** to infer disease subtyping and alignment
3. Prove conditions under which disease subtyping is **identifiable**
4. Algorithm improves over baselines in synthetic setting and validates **known subtypes** on real-world data

Machine Learning for Equitable Healthcare



Problem Selection

1. Early detection for intimate partner violence (PSB 2021)
2. Treating health disparities with AI (Nature Medicine 2020)

Data Collection

Collecting and researching insurance risk scores (ongoing)

Outcome Definition

Assessing different quality labels in intimate partner violence (ongoing)

Algorithm Development

Correcting for patient access to care (AAAI 2022)

Post-Deployment Considerations

1. Bias auditing (AMA Journal of Ethics 2019, Nature Medicine 2021)
2. Mitigating algorithmic bias (NeurIPS 2018)

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

Acknowledgements

David Sontag

Pete Szolovits

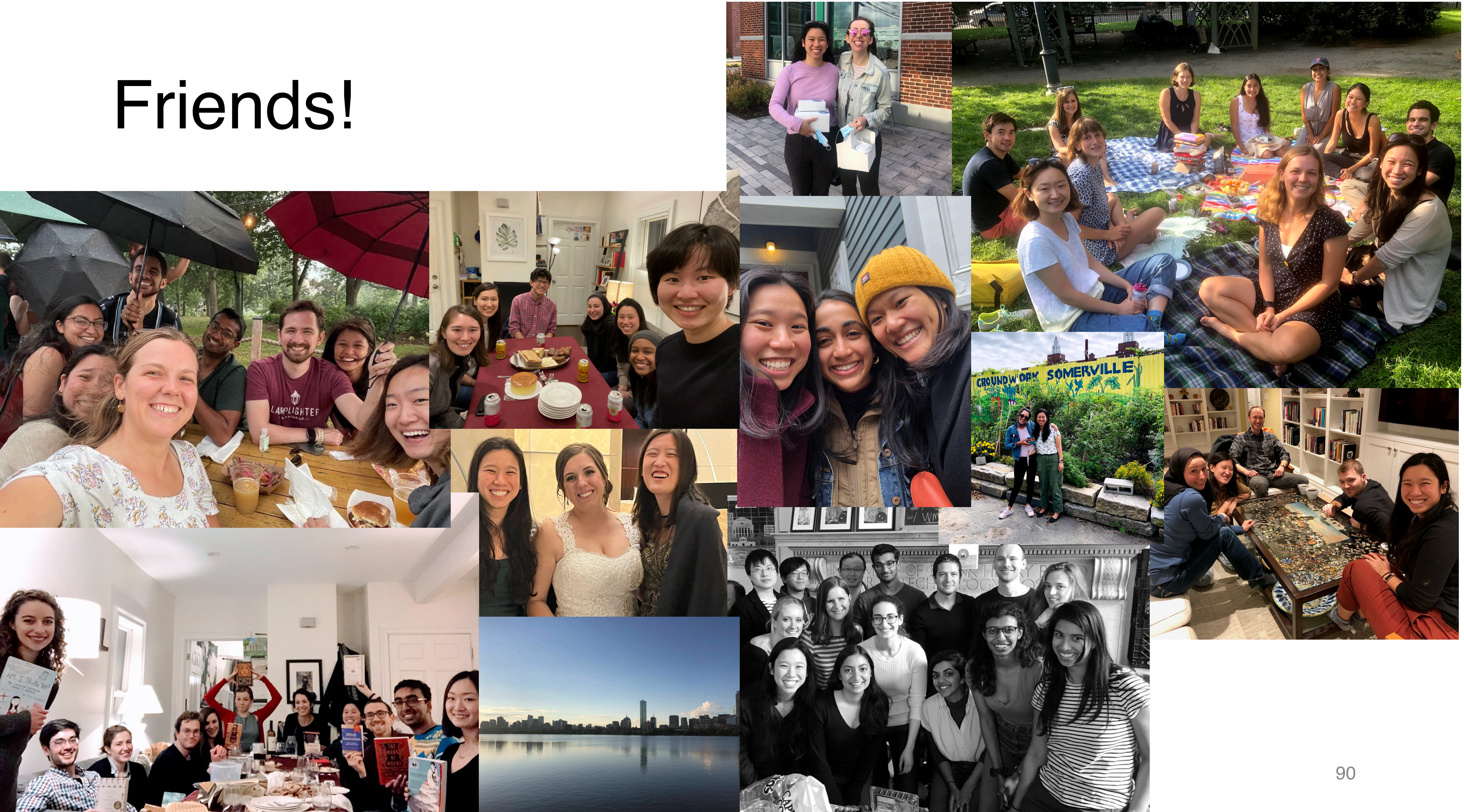
Marzyeh Ghassemi

Co-authors and Mentors

Clinical Machine Learning Group

Clinical Machine Learning Group

Friends!



Family