Fairness, Ethics, and Healthcare

IRENE CHEN (MIT)

CSC2541HS GUEST LECTURE

“Tuskegee Study of Untreated Syphilis in the Negro Male” (1932)

Ethics in healthcare is nothing new

- **Drug pricing:** The strange world of Canadian drug pricing (The Toronto Star, Jan 2019)
- **Opioid epidemic**: Massachusetts Attorney General Implicates Family Behind Purdue Pharma In Opioid Deaths (NPR, Jan 2019)
- **Retracted studies**: Harvard Calls for Retraction of Dozens of Studies by Noted Cardiac Researcher (NYT, Oct 2018)
- **Conflict of interest:** Sloan Kettering's Cozy Deal with Start-Up Ignites a New Uproar (NYT, Sept 2018)
- **Clinical trial populations**: Clinical Trials Still Don't Reflect the Diversity of America (NPR, Dec 2015)

What about algorithms?

Algorithms change the discussion

- What is reasonable safety for autonomous systems?
- Is the patient informed about risks and benefits?
- What about privacy and data collection?
- Who should regulate? Should these be for-profit black box algorithms?
- What about diversity? What populations are these tested on and then applied to?

Would you be okay with an algorithm for:

- Cardiovascular disease risk to **prescribe treatment**?
- Government disability severity to **allocate care**?
- Child endangerment risk to **decide in-home visits**?

Ann Intern Med. 2018 Jul 3;169(1):20-29. doi: 10.7326/M17-3011. Epub 2018 Jun 5.

Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk.

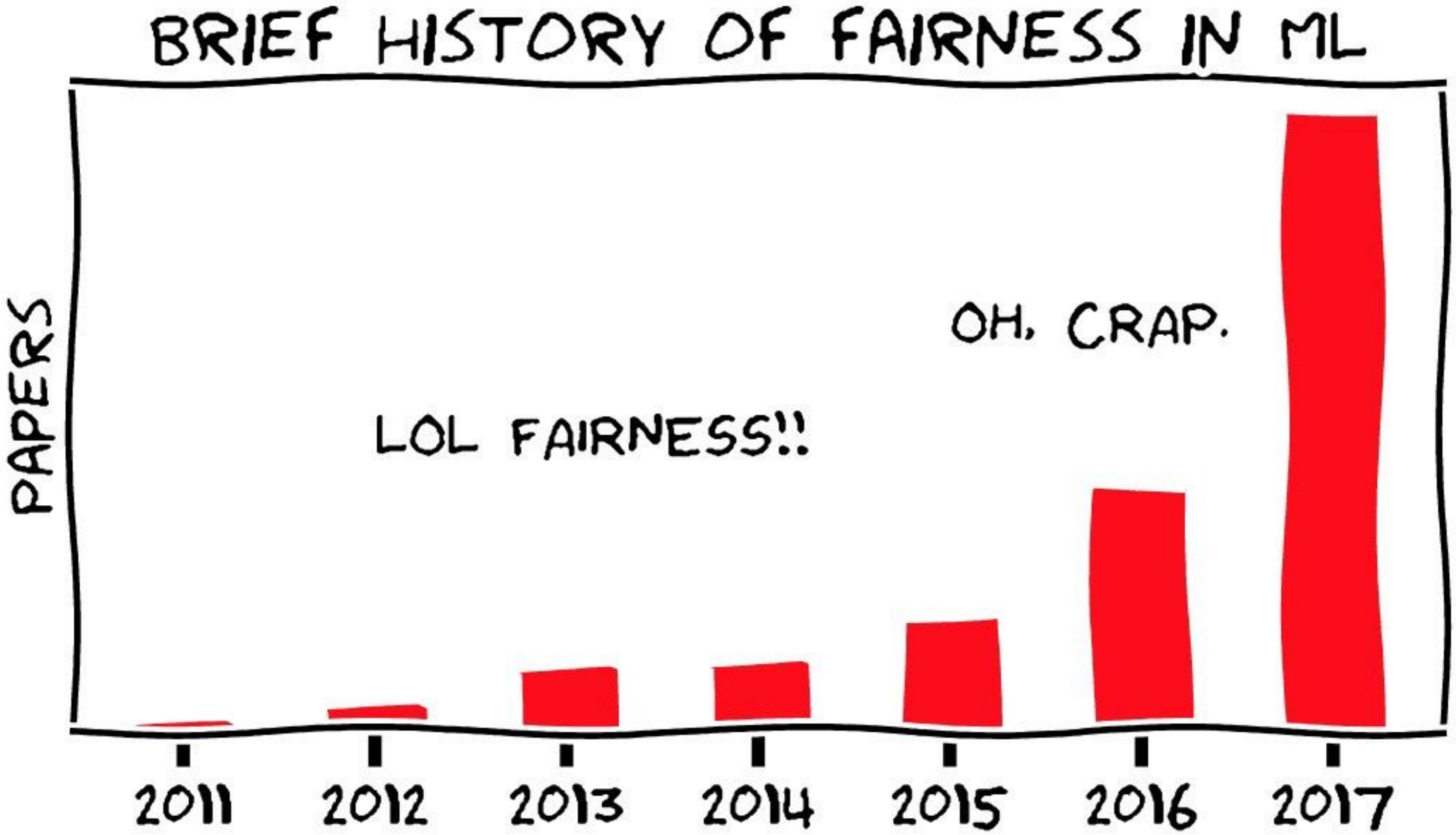
Yadlowsky S[1], Hayward RA[2], Sussman JB[2], McClelland RL[3], Min YI[4], Basu S[5].

SCIENCE

WHAT HAPPENS WHEN AN ALGORITHM CUTS YOUR HEALTH CARE

By Colin Lecher | @colinlecher | Mar 21, 2018, 9:00am EDT

Illustrations by William Joel; Photography by Amelia Holowaty Krales


BRIEF HISTORY OF FAIRNESS IN ML
PAPERS
OH, CRAP.
LOL FAIRNESS!!
2011
2012
2013
2014
2015
2016
2017

[Hardt, 2018]

Formalization of Fairness

- Fairness through blindness
- Demographic parity (or group fairness or statistical parity)
- Calibration (or predictive parity)
- Error rate balance (or equalized odds)
- Representation learning
- Causality and fairness
- … and many others! [Narayanan et al, 2018]

Discussion points

- What are relevant ***protected groups***?
- How do we define or measure ***unfairness***?
- What are areas of healthcare where we might be concerned about bias?

Fairness through Blindness

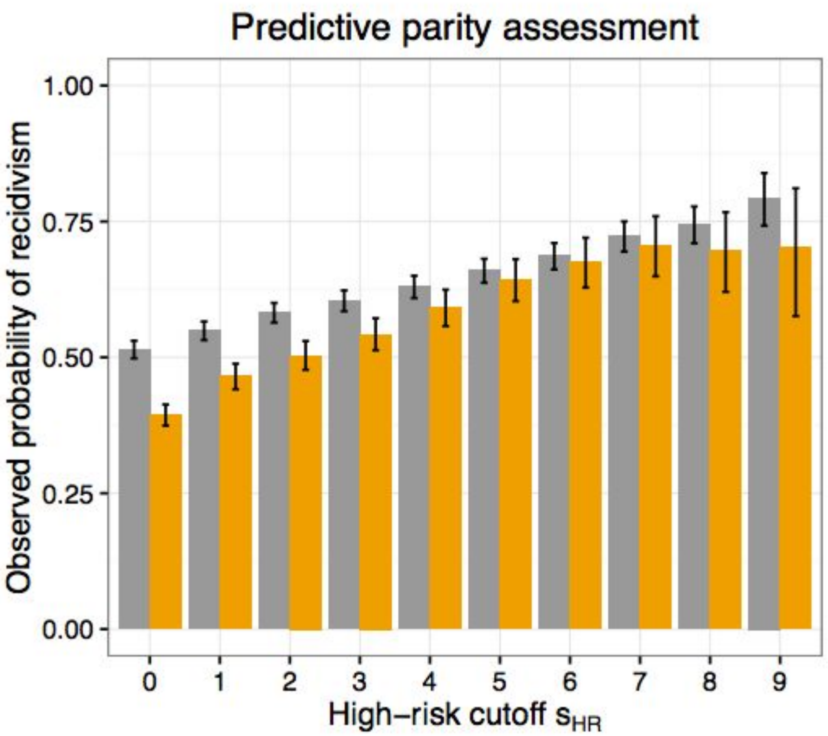
- **Plan**: Remove any sensitive group from data
- **Example**: Predict diabetes risk *Y* from clinical features *X* and race *A* using $P(\hat{Y} = Y | X)$ instead of $P(\hat{Y} = Y | X, A)$
- **Problems:**
  - *A* might have predictive value. What if *Y* = *A*?
  - Other features of *X* might be correlated with *A*

Demographic parity

- **Plan**: Require same fraction of $\hat{Y} = 1$ for each group *A*
- **Example**: Predict diabetes risk *Y* from clinical features *X* and race *A* such that $P(\hat{Y} = 1 | A = 1) = P(\hat{Y} = 1 | A = 0)$
- **Problems:**
  - What if true *Y* perfectly correlates with *A*?
  - Too strong: even perfect prediction $Y = \hat{Y}$ doesn't satisfy requirements
  - Too weak: doesn't control error rate, could be perfectly biased (wrong for all *A* = 1, correct for A = 0) and still have demographic parity

Calibration

- **Plan**: Same positive predictive value across groups
- **Example**: Predict diabetes risk Y from score S with threshold T from clinical features X and race A such that
$$P(Y = 1 | S > T, A = 0)$$
$$= P(Y = 1 | S > T, A = 1)$$
- **Problems:**
  - Might be in conflict with error rate balance



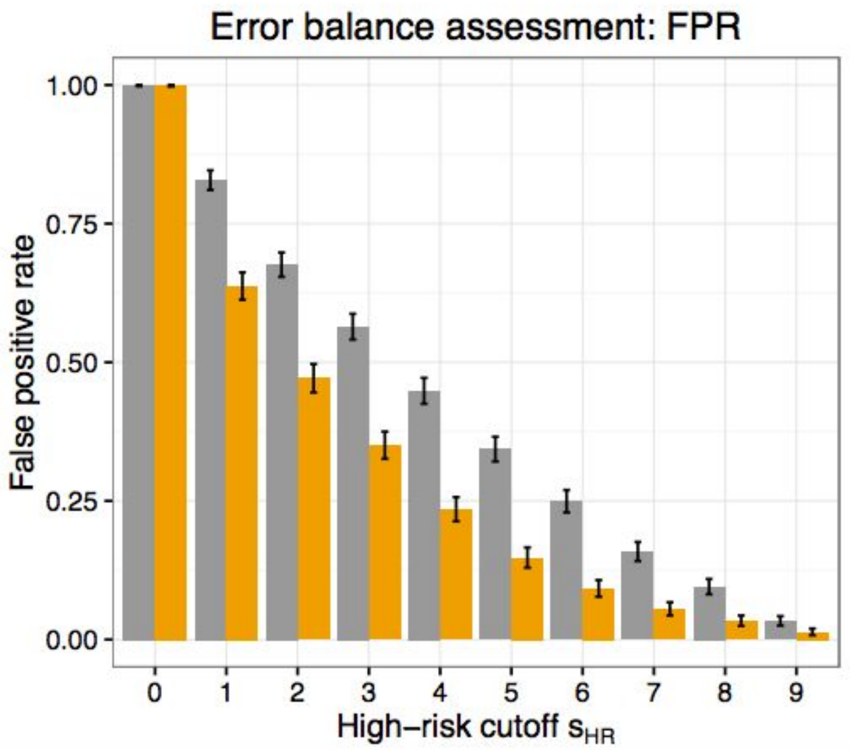
[Chouldechova, 2018]

Error rate balance

- **Plan**: Same positive predictive value across groups
- **Example**: Predict diabetes risk *Y* from score *S* with threshold *T* from clinical features *X* and race *A* such that

$$P(S > T | Y = 0, A = 0)$$
$$= P(S > T | Y = 0, A = 1)$$

- **Problems:**
  - Might be in conflict with calibration



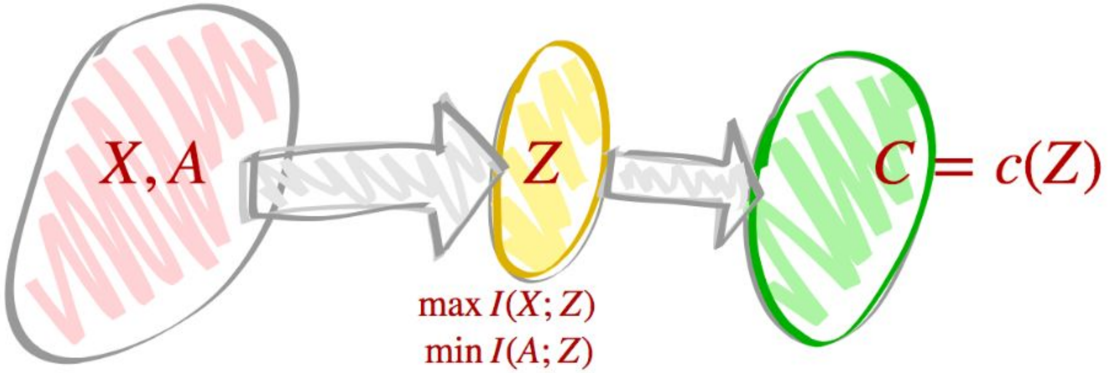
[Chouldechova, 2018]

Representation learning

- **Plan**: Learn latent representation to minimize group information
- **Example**: Predict diabetes risk Y from score S with threshold T from clinical features X and race A such that

  $$\max I(X;Z) \text{ and } min\, I(A;Z)$$

- **Problems:**
  - How to ensure you are not losing too much info and learning right representation?



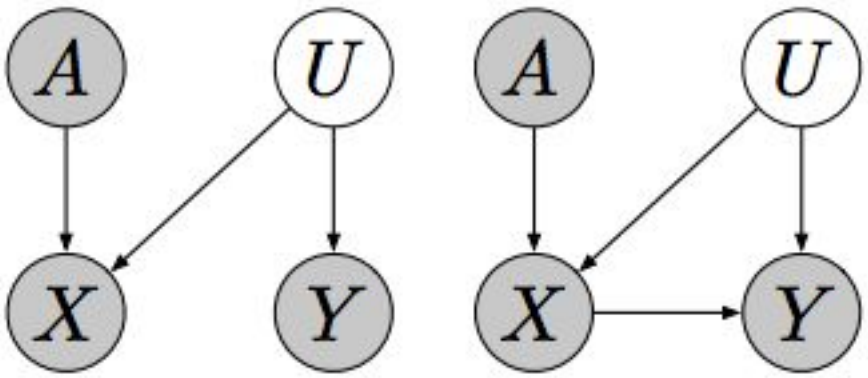
[Zemel et al, 2013]

Causal inference and fairness

- **Plan**: Group *A* should not be cause of prediction $\hat{Y}$
- **Example**: Predict diabetes risk *Y* from clinical features *X* and race *A* such that

$$P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a)$$

- **Problems:**
  - Creating a structural model encodes prior beliefs about world
  - Causal inference often requires ignorability assumptions

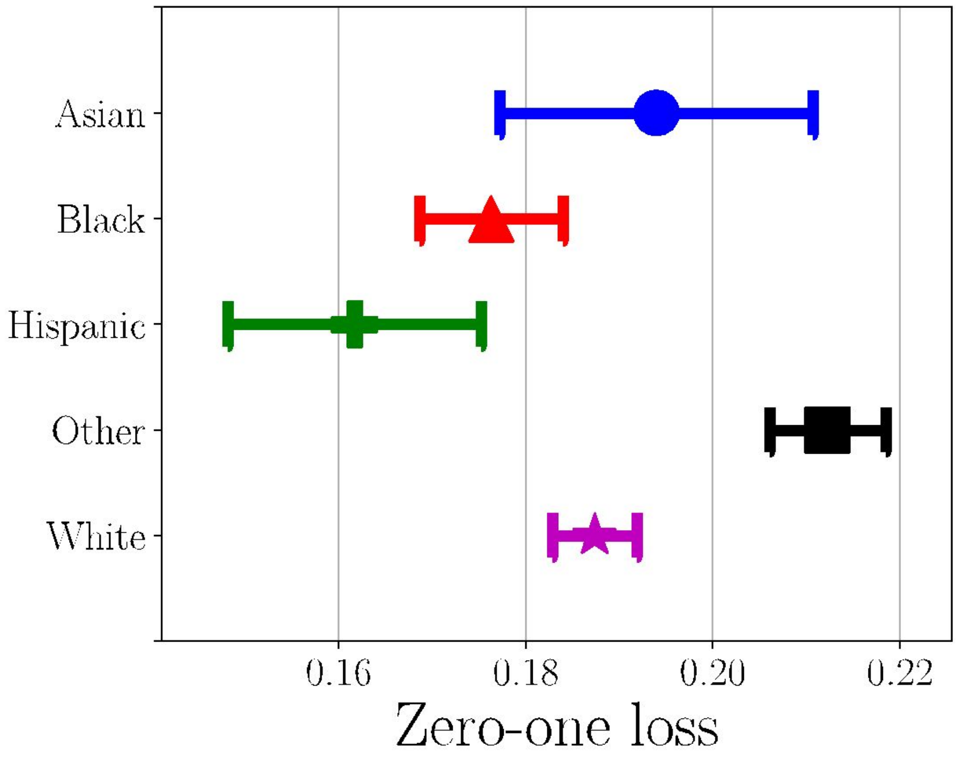


[Kusner et al, 2017]

What about the data?

Predicting hospital mortality from MIMIC

- Using clinical notes, can we predict hospital mortality from MIMIC data?
- We train a L1-regularized logistic regression.
- How do the accuracies differ by racial group?
- What might cause these discrepancies?

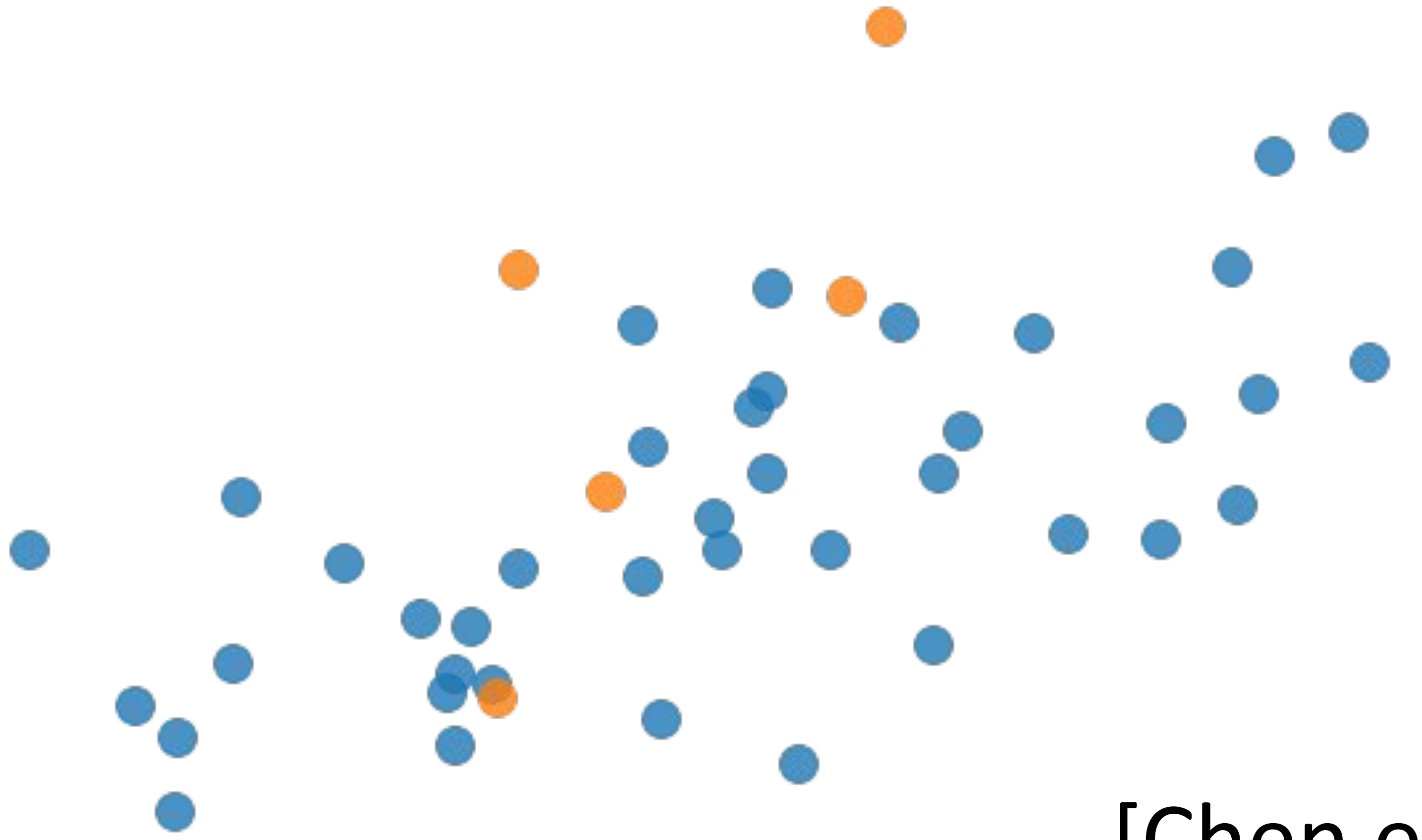


[Chen et al, 2018]

Why might my classifier be unfair?

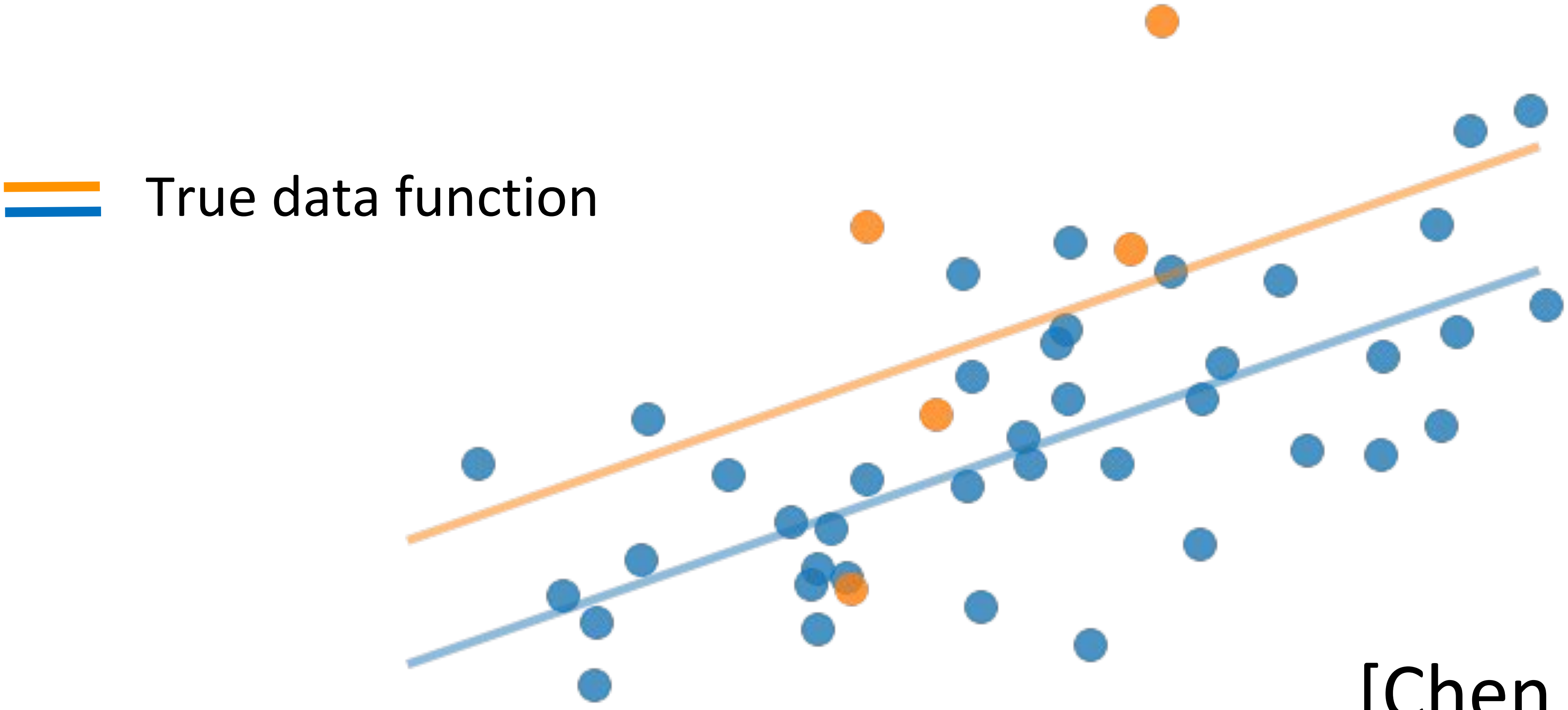
[Chen et al, 2018]

Why might my classifier be unfair?



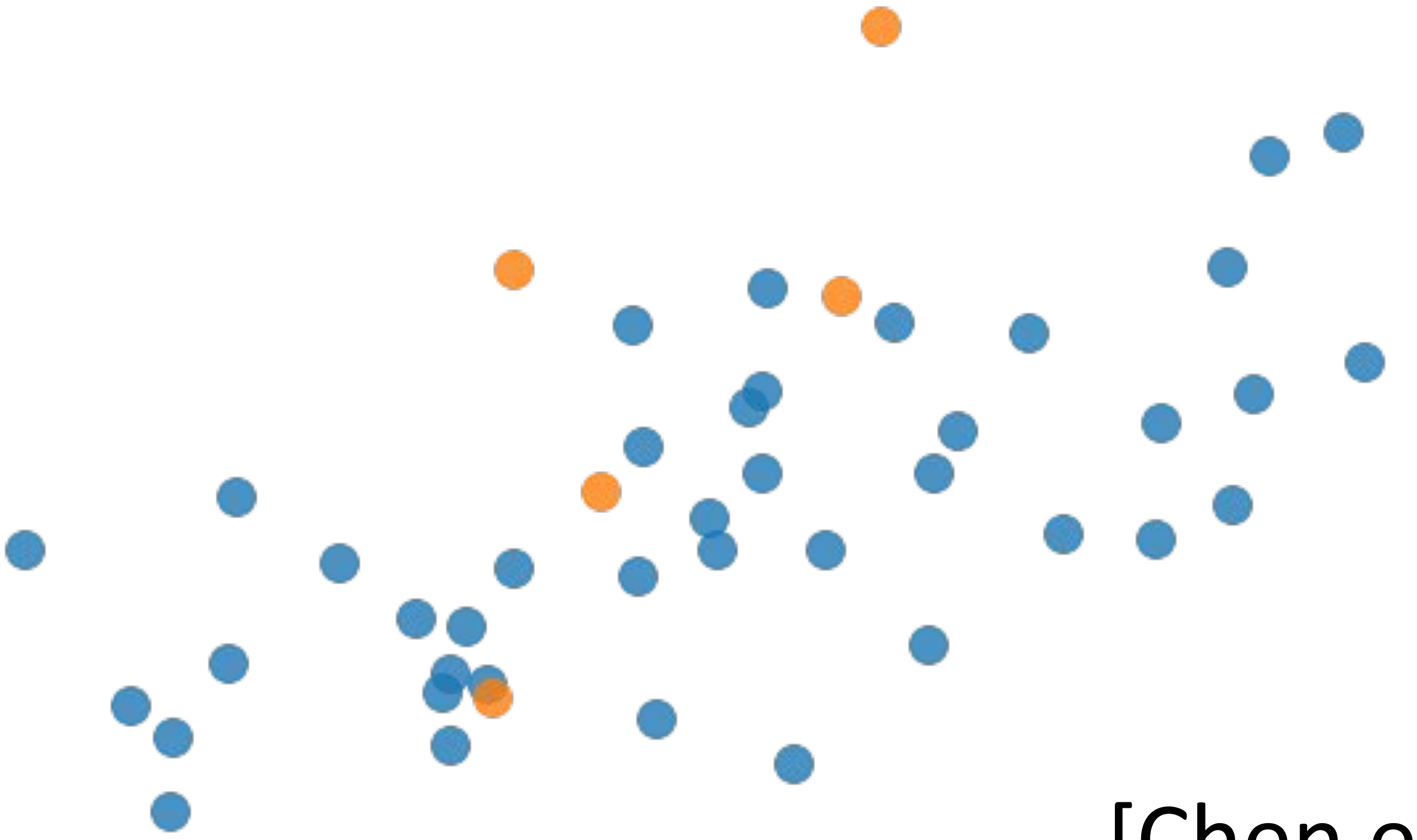
[Chen et al, 2018]

Why might my classifier be unfair?



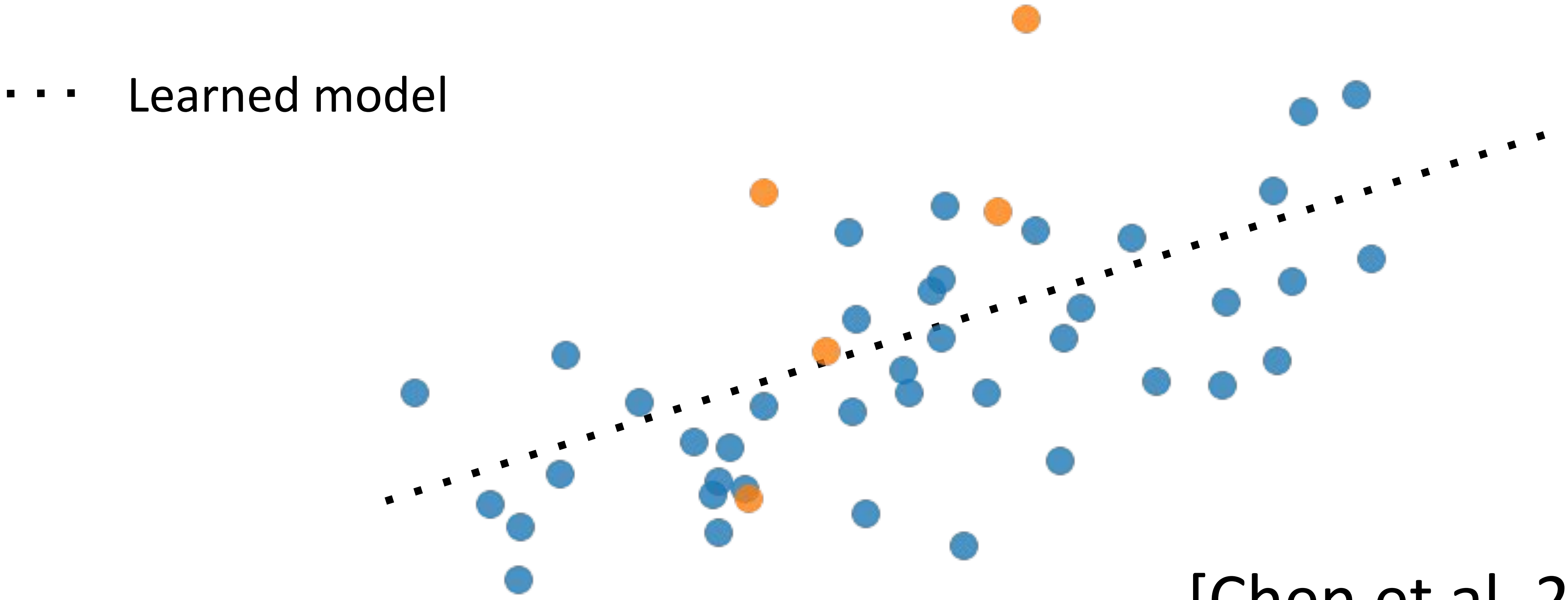
[Chen et al, 2018]

Why might my classifier be unfair?



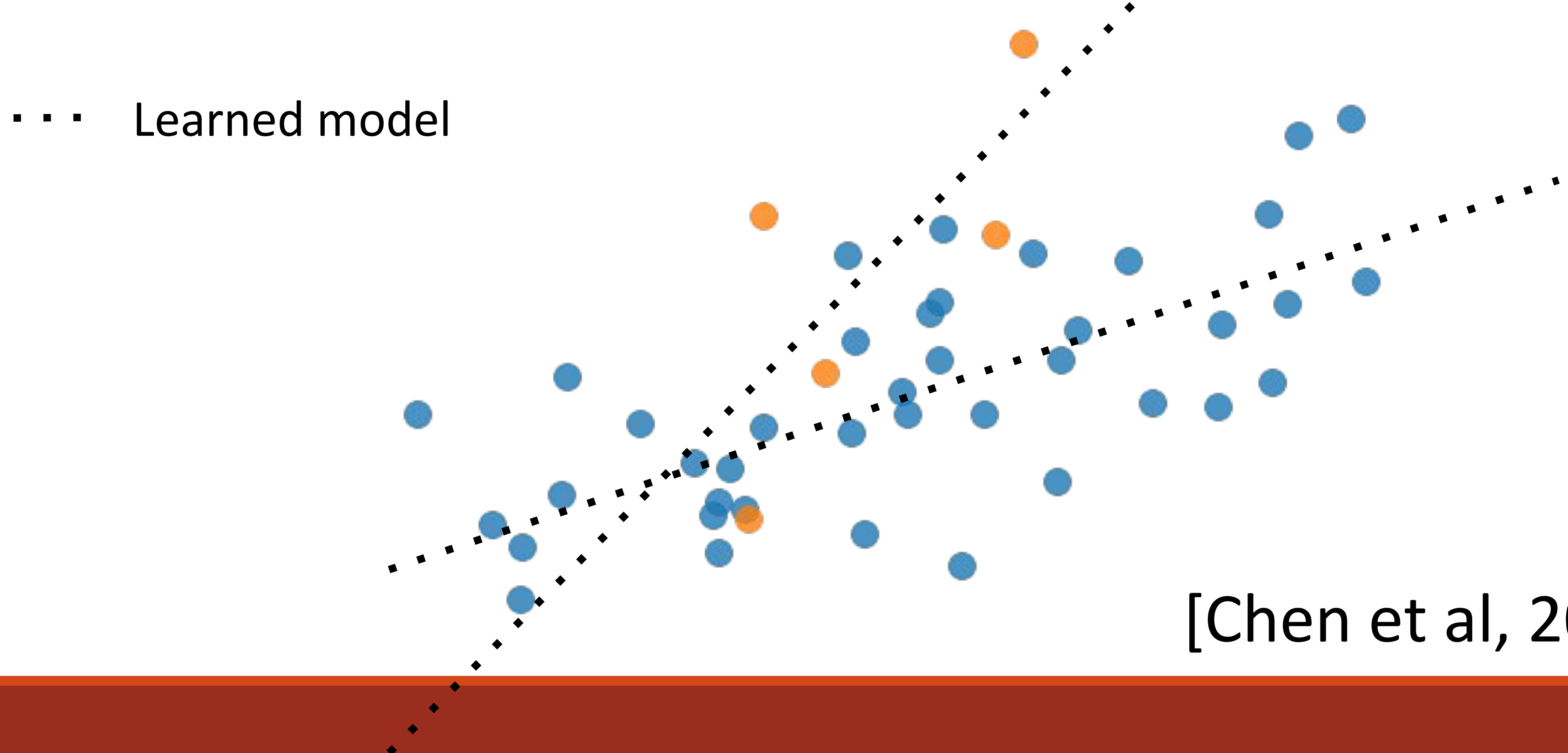
[Chen et al, 2018]

Why might my classifier be unfair?



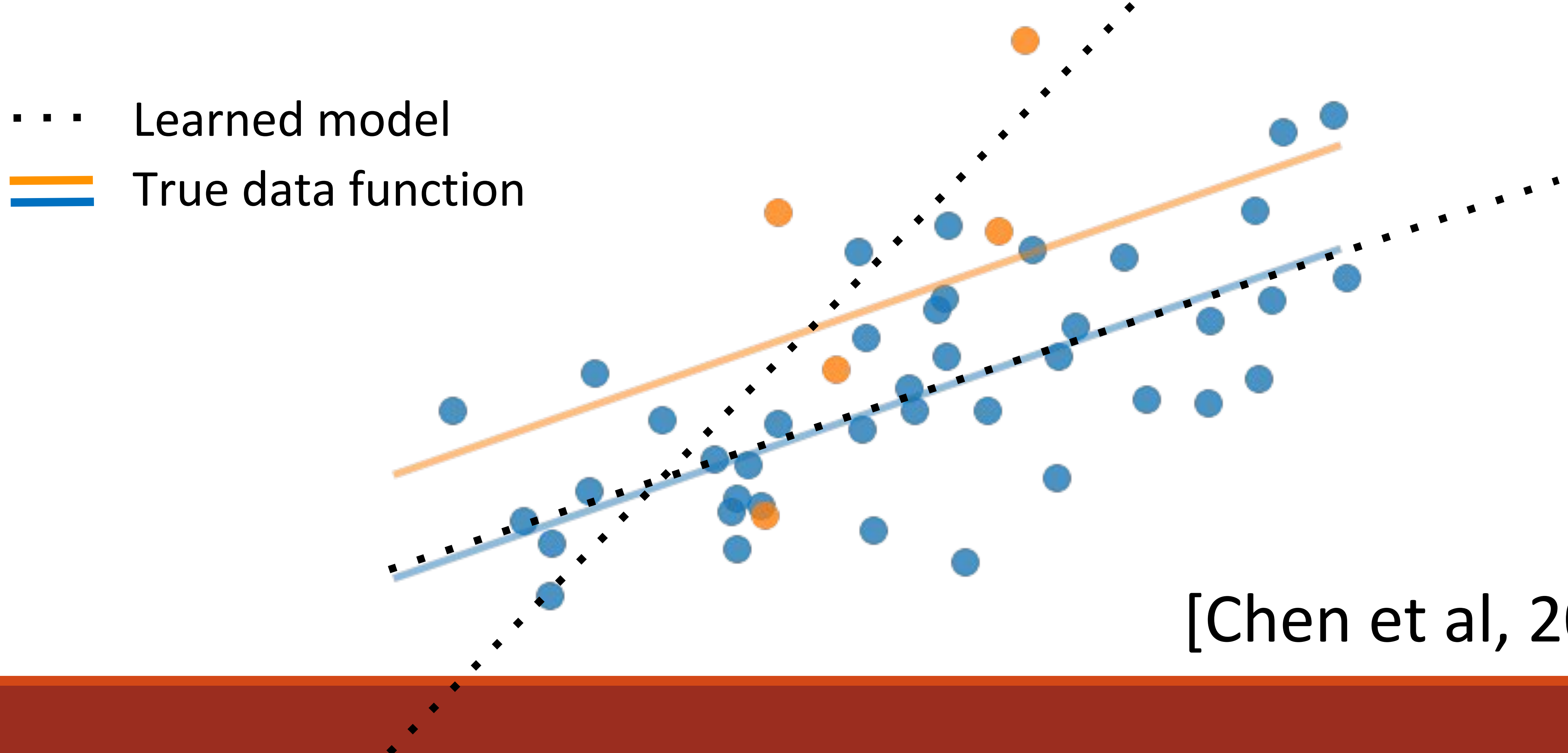
[Chen et al, 2018]

Why might my classifier be unfair?



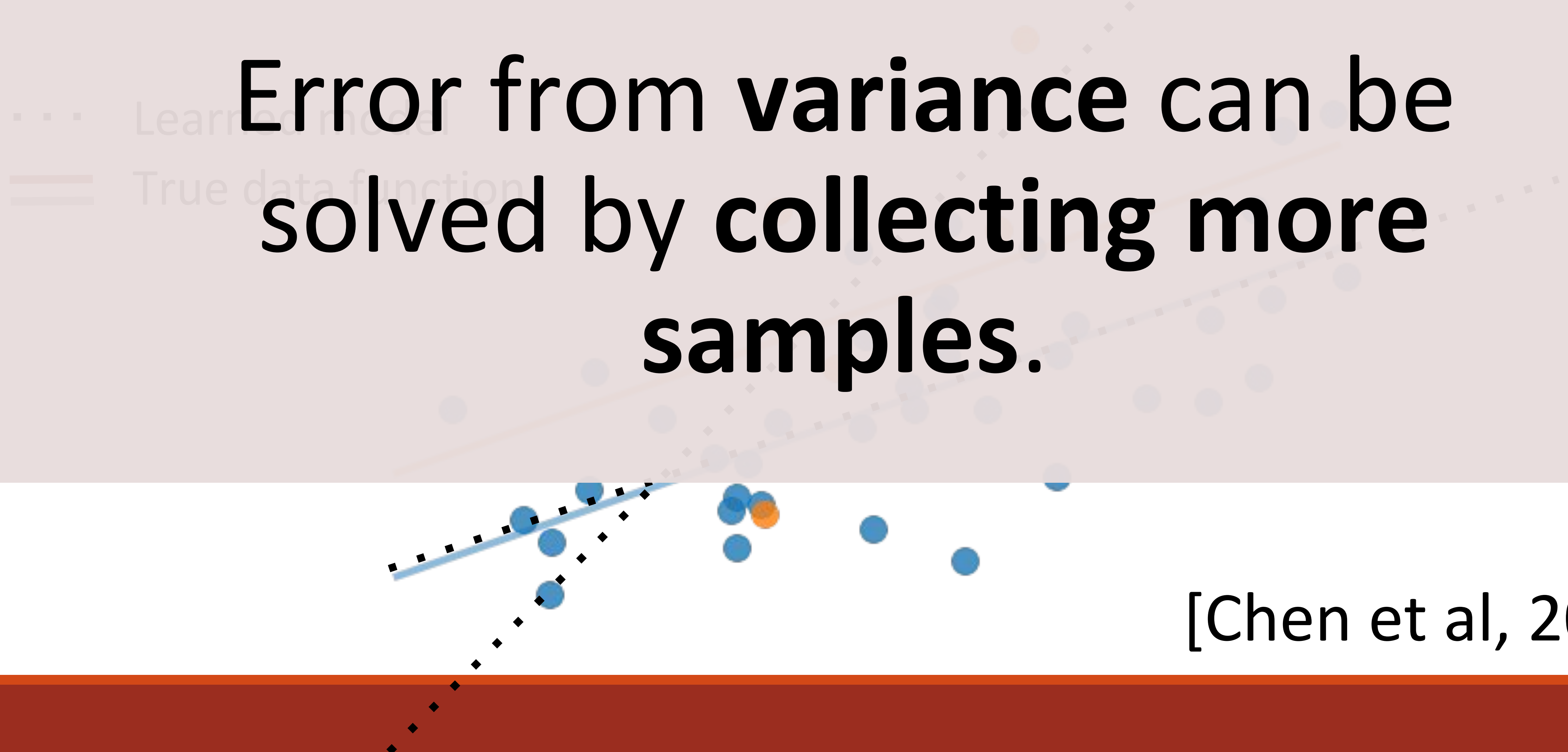
[Chen et al, 2018]

Why might my classifier be unfair?



[Chen et al, 2018]

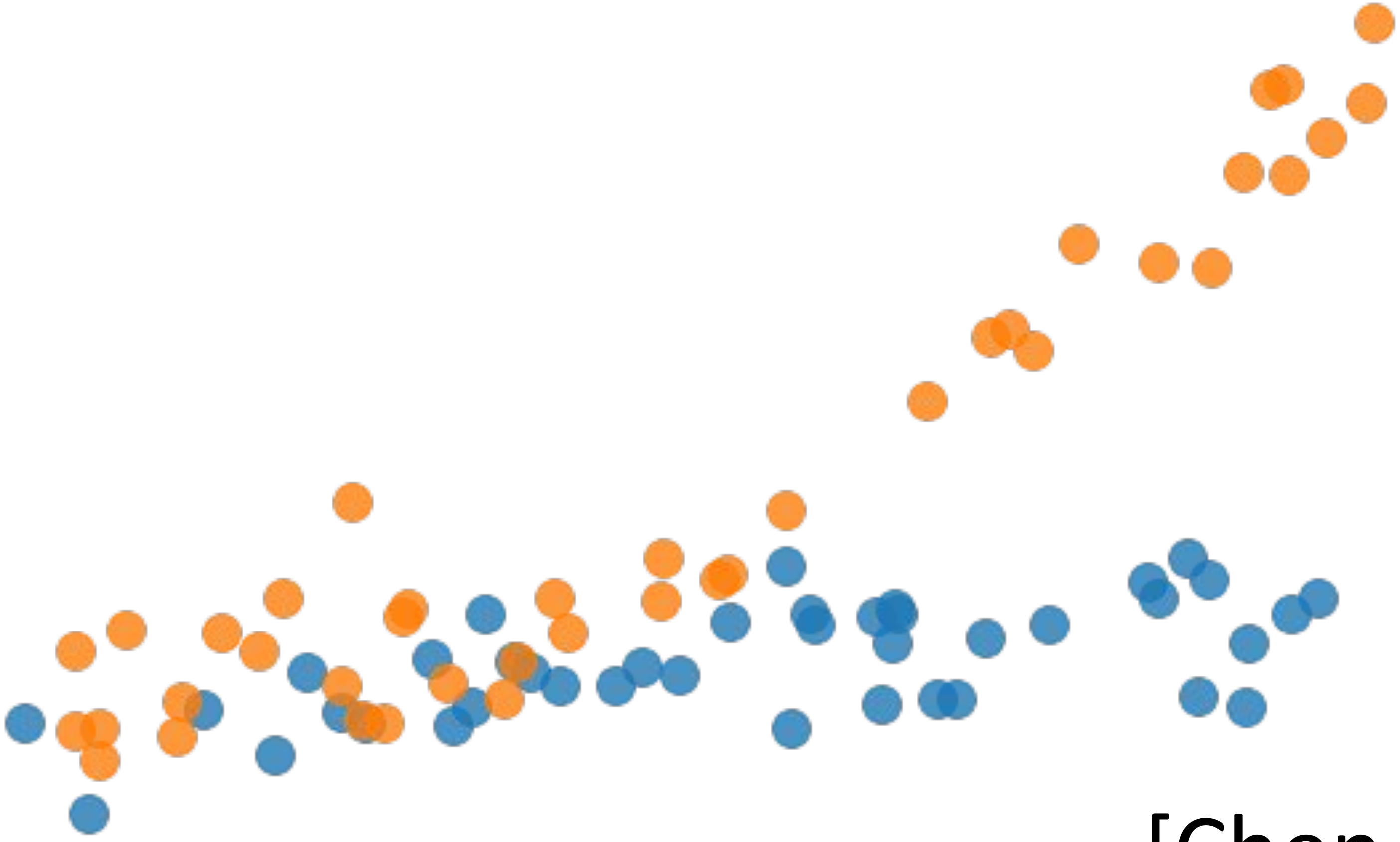
Why might my classifier be unfair?



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

[Chen et al, 2018]

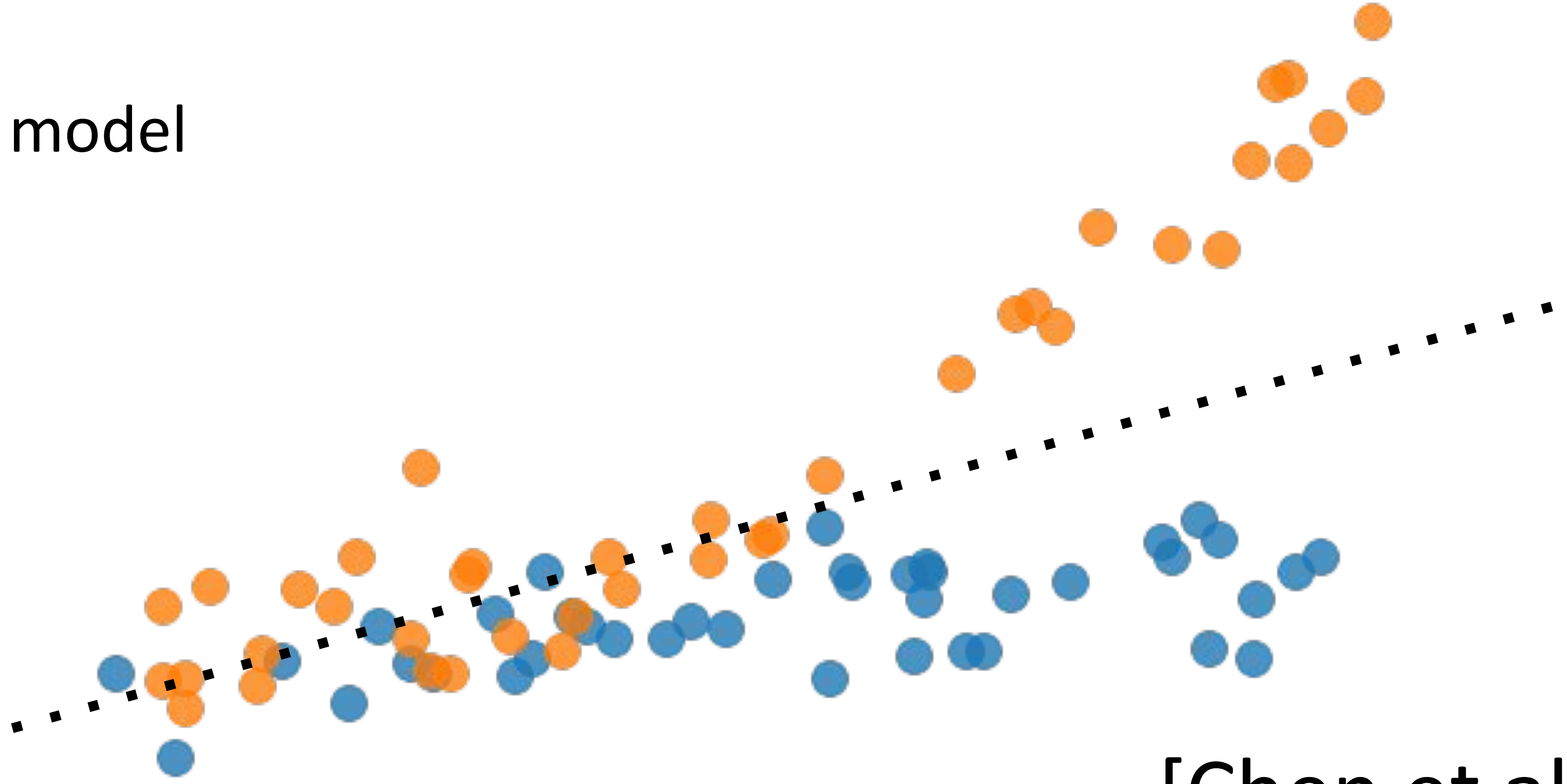
Why might my classifier be unfair?



[Chen et al, 2018]

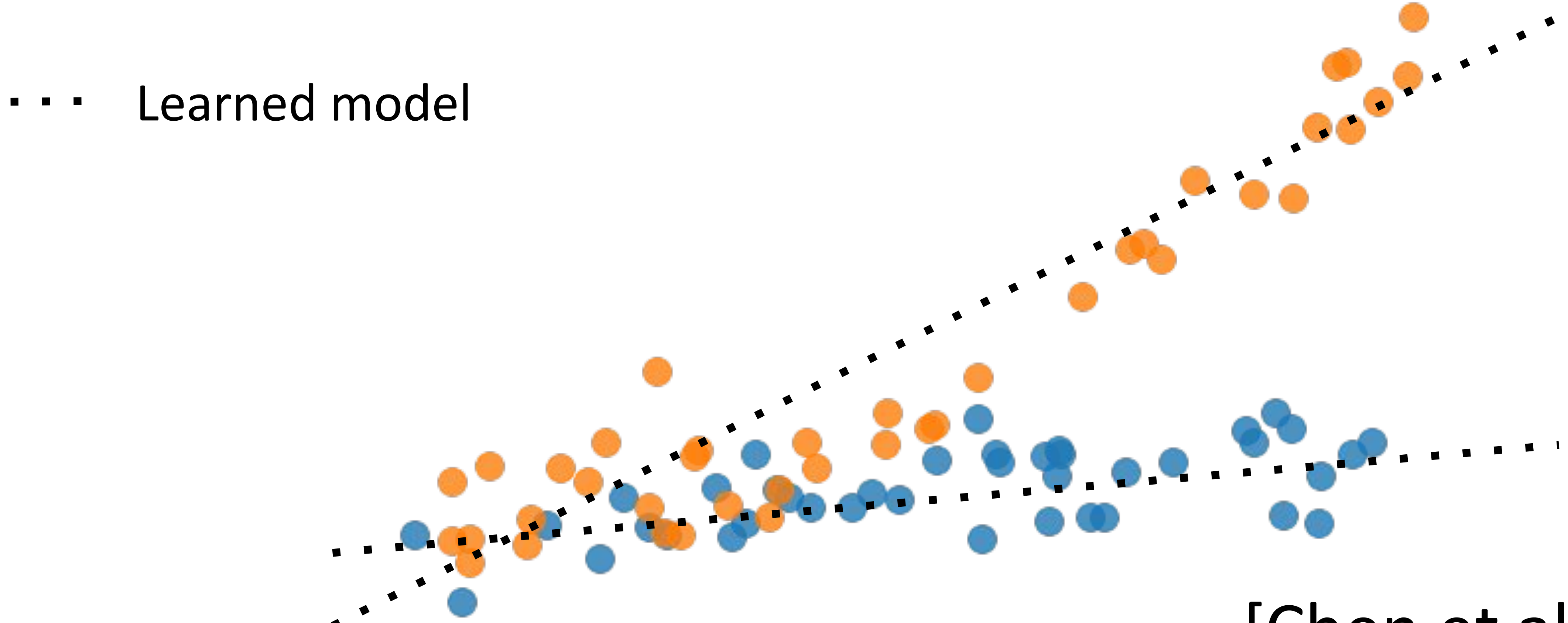
Why might my classifier be unfair?

··· Learned model



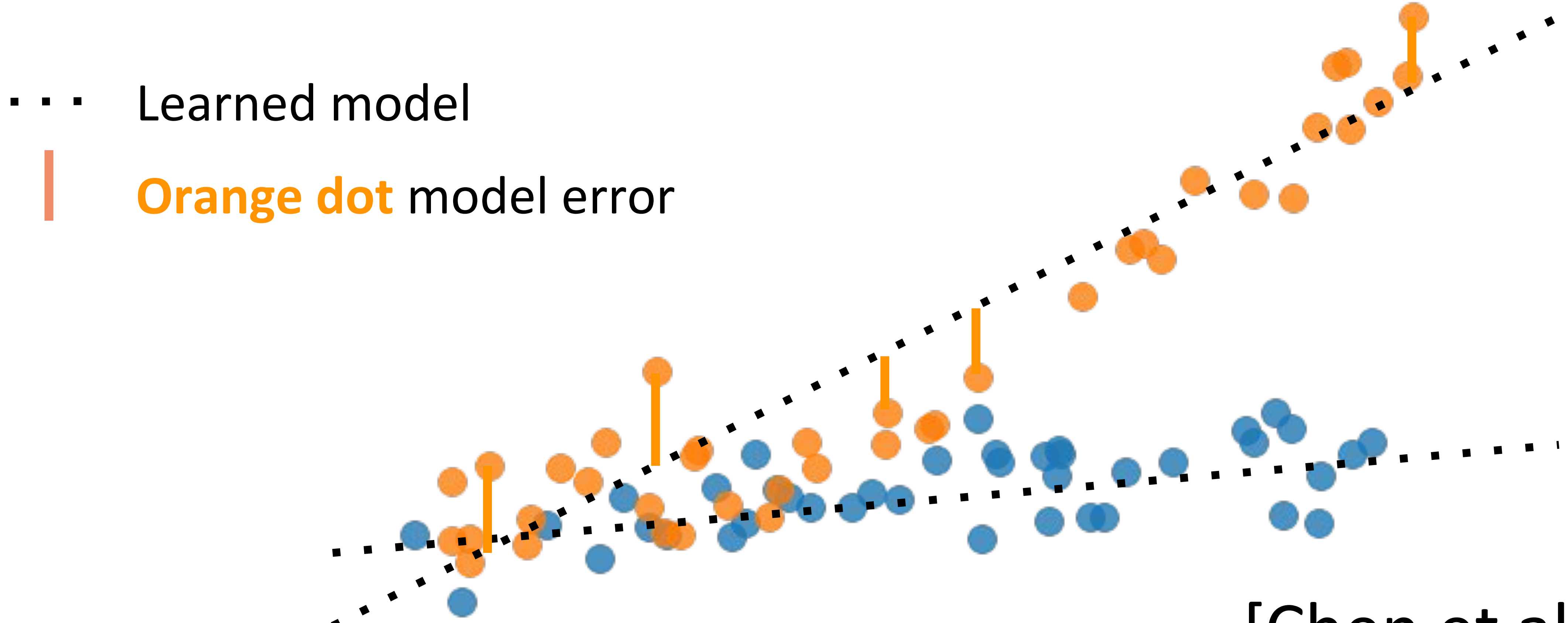
[Chen et al, 2018]

Why might my classifier be unfair?



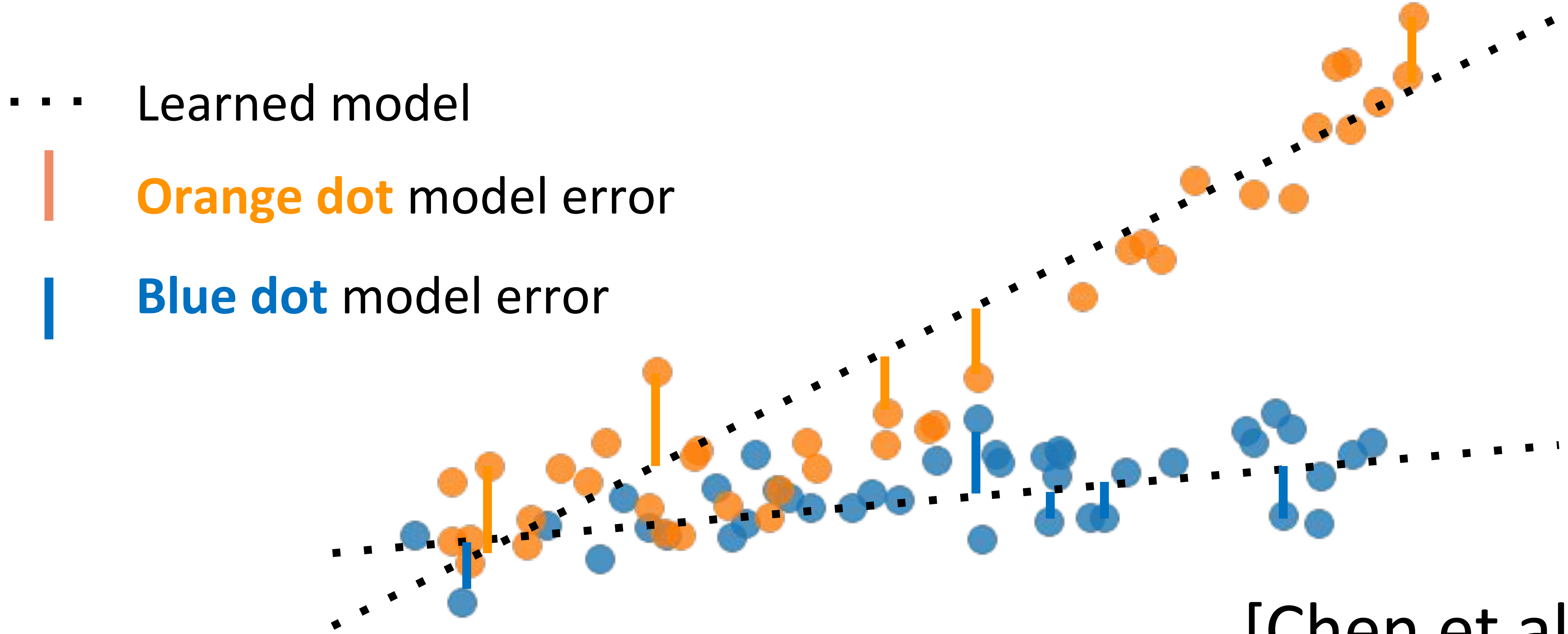
[Chen et al, 2018]

Why might my classifier be unfair?



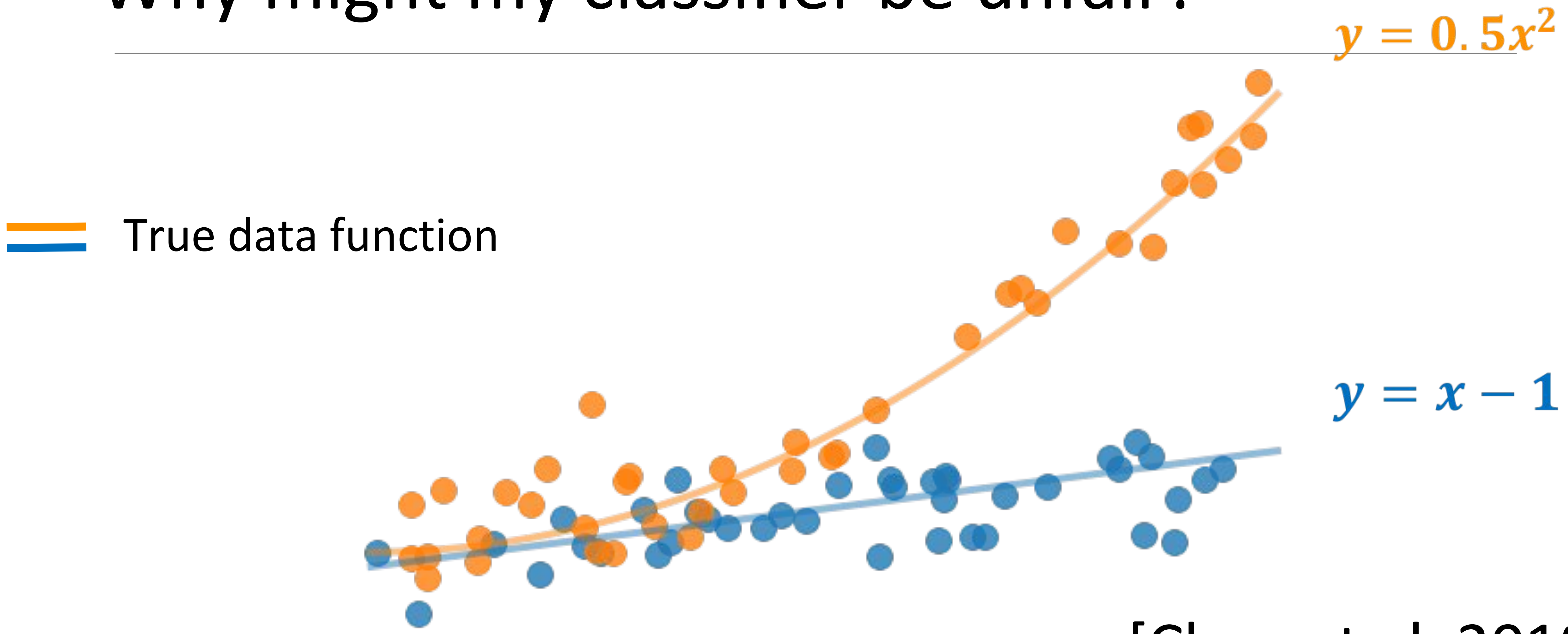
[Chen et al, 2018]

Why might my classifier be unfair?



[Chen et al, 2018]

Why might my classifier be unfair?



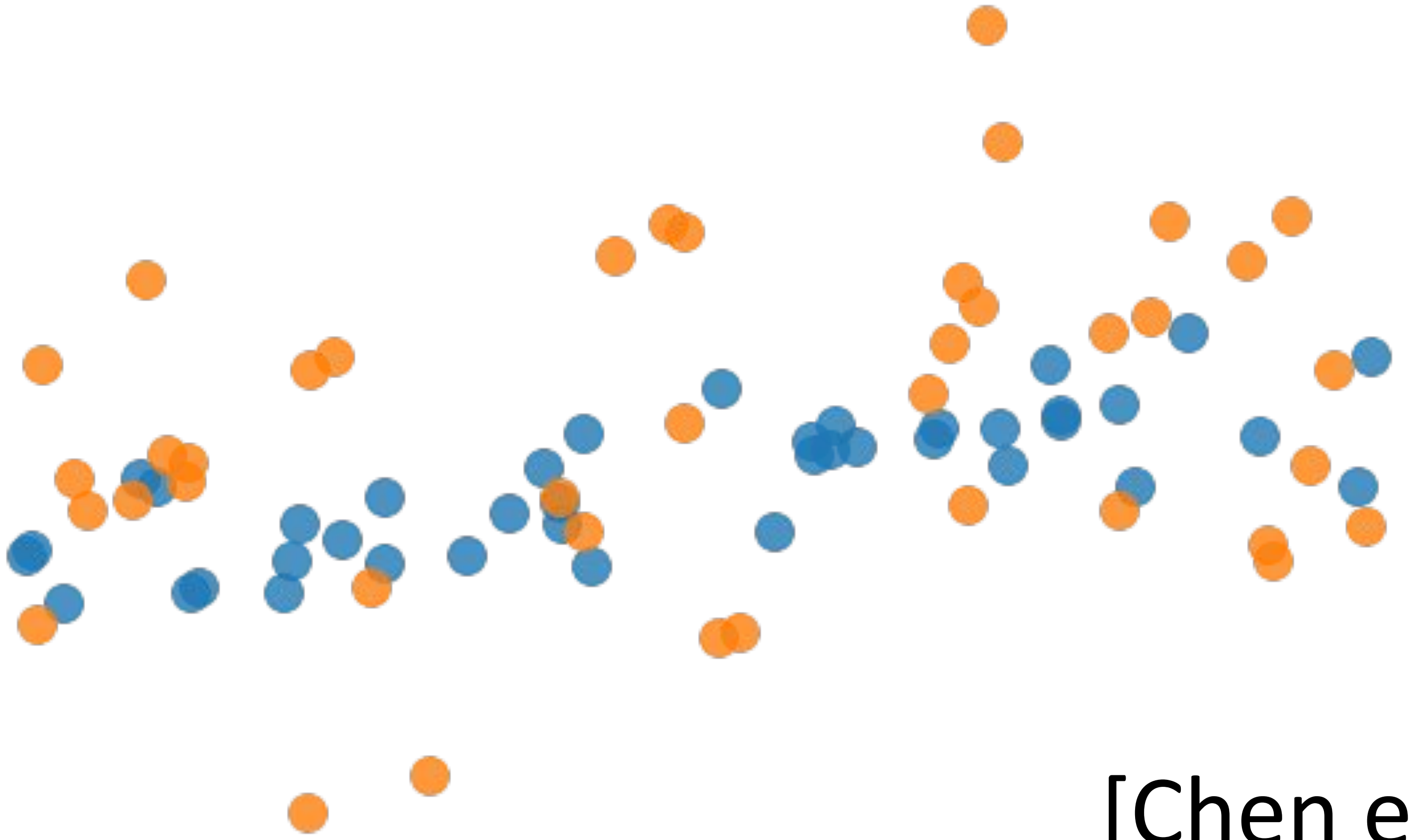
[Chen et al, 2018]

Why might my classifier be unfair?

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

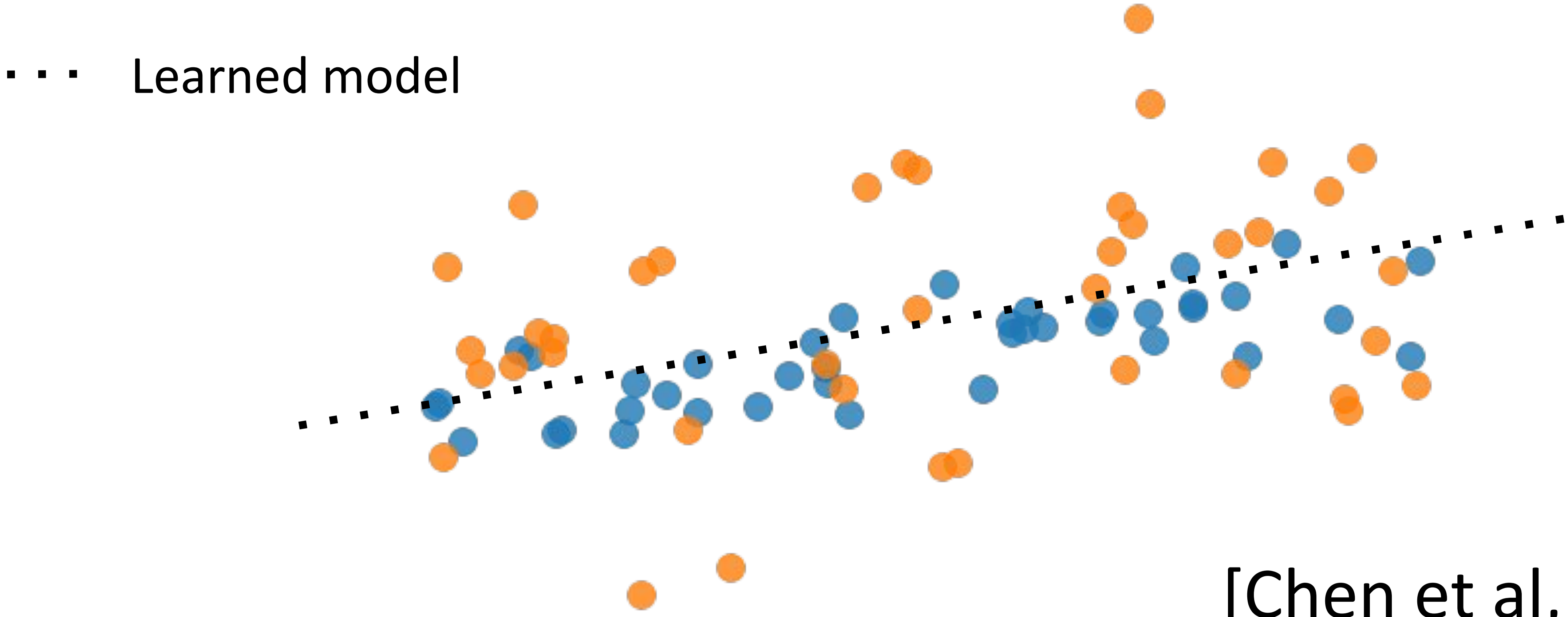
[Chen et al, 2018]

Why might my classifier be unfair?



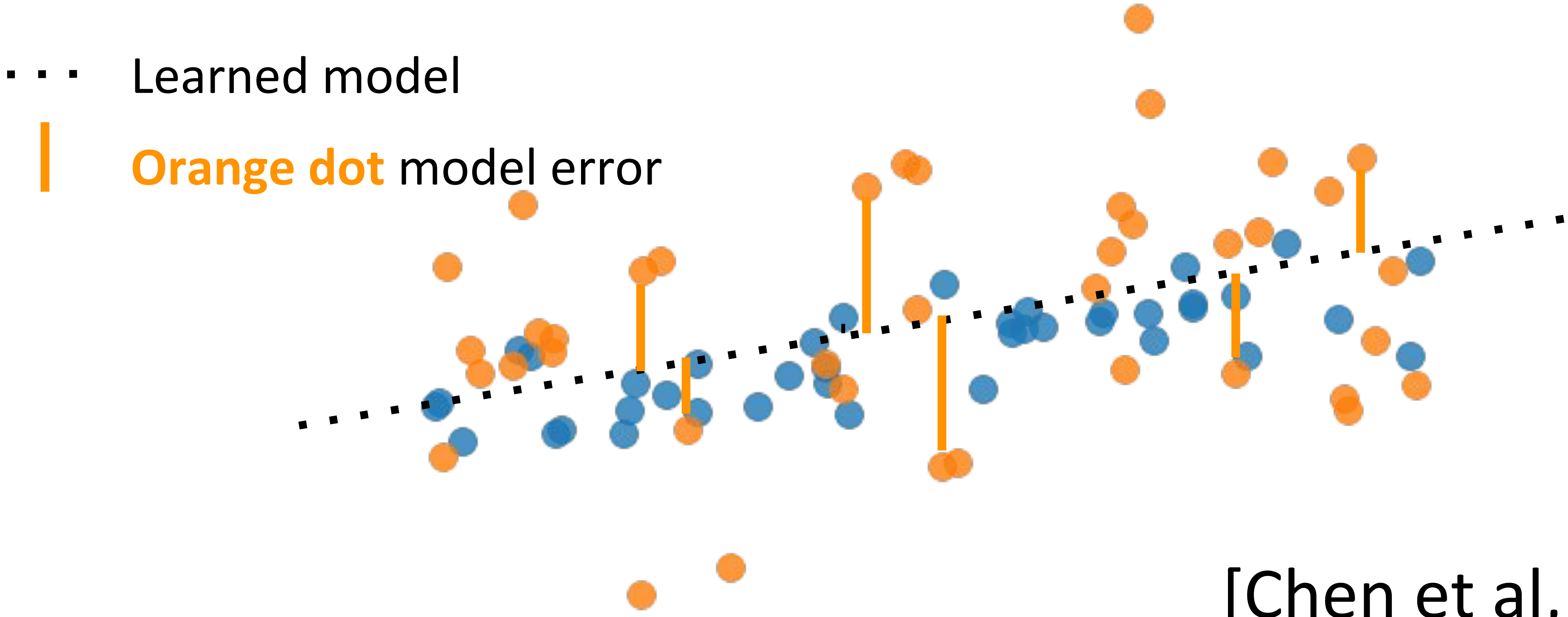
[Chen et al, 2018]

Why might my classifier be unfair?



[Chen et al, 2018]

Why might my classifier be unfair?

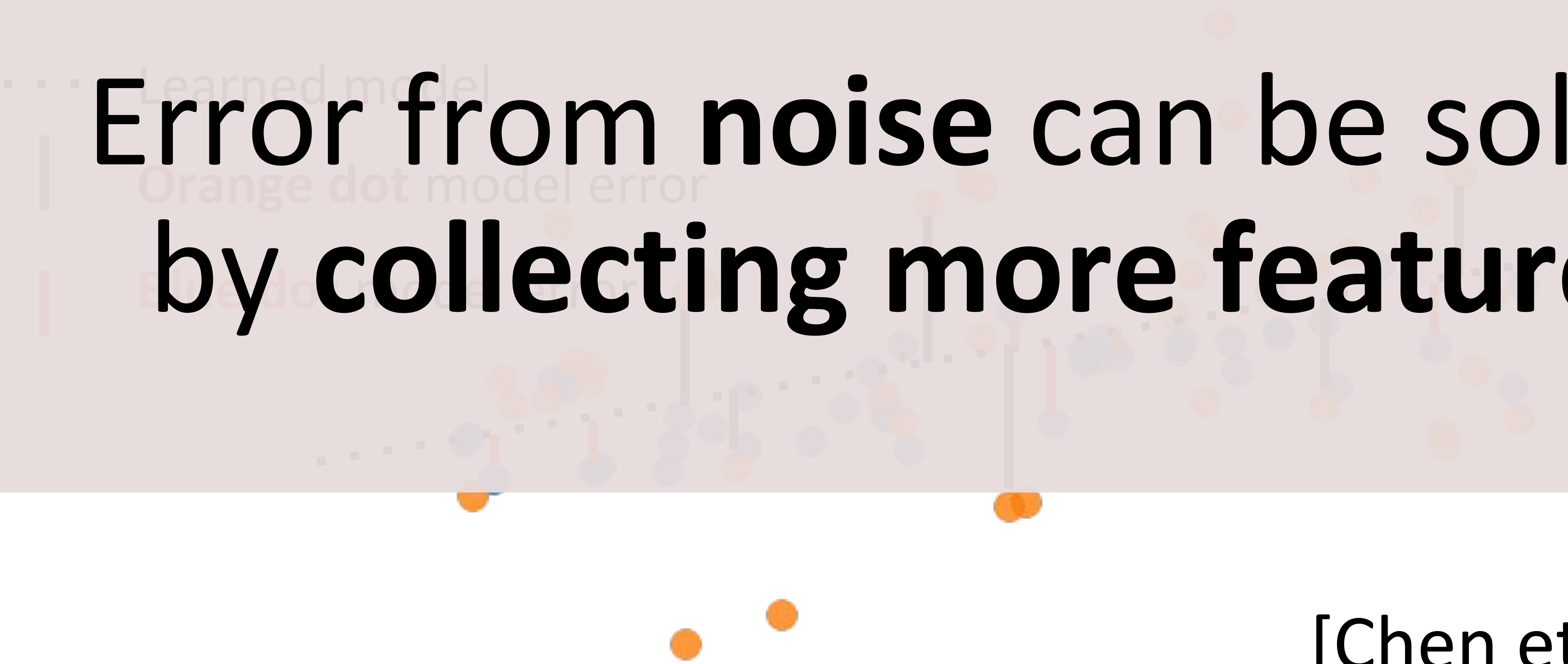


[Chen et al, 2018]

Why might my classifier be unfair?

[Chen et al, 2018]

Why might my classifier be unfair?



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

[Chen et al, 2018]

Bias, variance, noise

We can decompose how a predictor $\hat{Y}$ performs based on protected group a, features x, and data D through Bayes optimal predictor y^* , majority predictor $\tilde{y}$

- Bias $B_a(\hat{Y}, x, a) = L(y^*(x,a), \tilde{y}(x,a))$
- Variance $V_a(\hat{Y}, x, a) = E_D[L(\tilde{y}(x,a), \hat{y}_D(x,a)]$
- Noise $N(x,a) = E_Y[L(y^*(x,a)) \mid X, A]$

[Domingos, 2000]

What about 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 $\hat{Y}$:

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

[Chen et al, 2018]

What about 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 $\hat{Y}$:

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

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

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

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

[Chen et al, 2018]

Bias, variance, noise for fairness

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*.

[Chen et al, 2018]

Bias, variance, noise for fairness

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

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

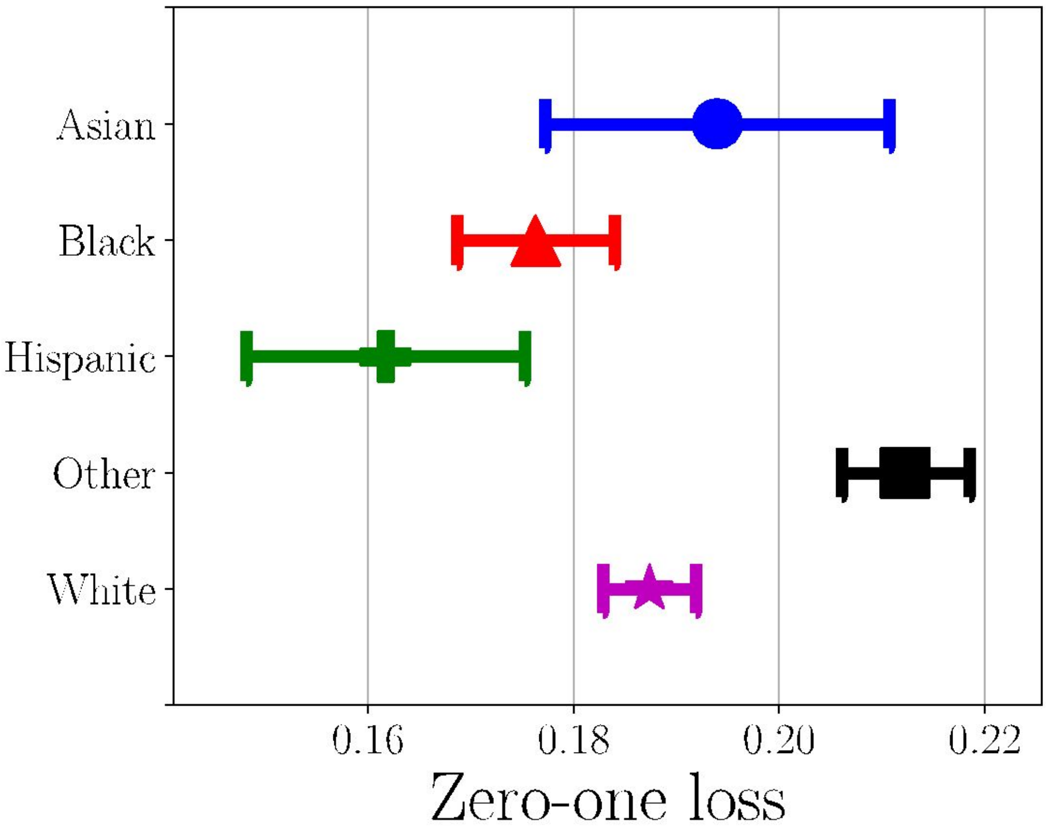
Note that $\overline{N_a}$ indicates the expectation of N_a over *X* and data *D*.

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.

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

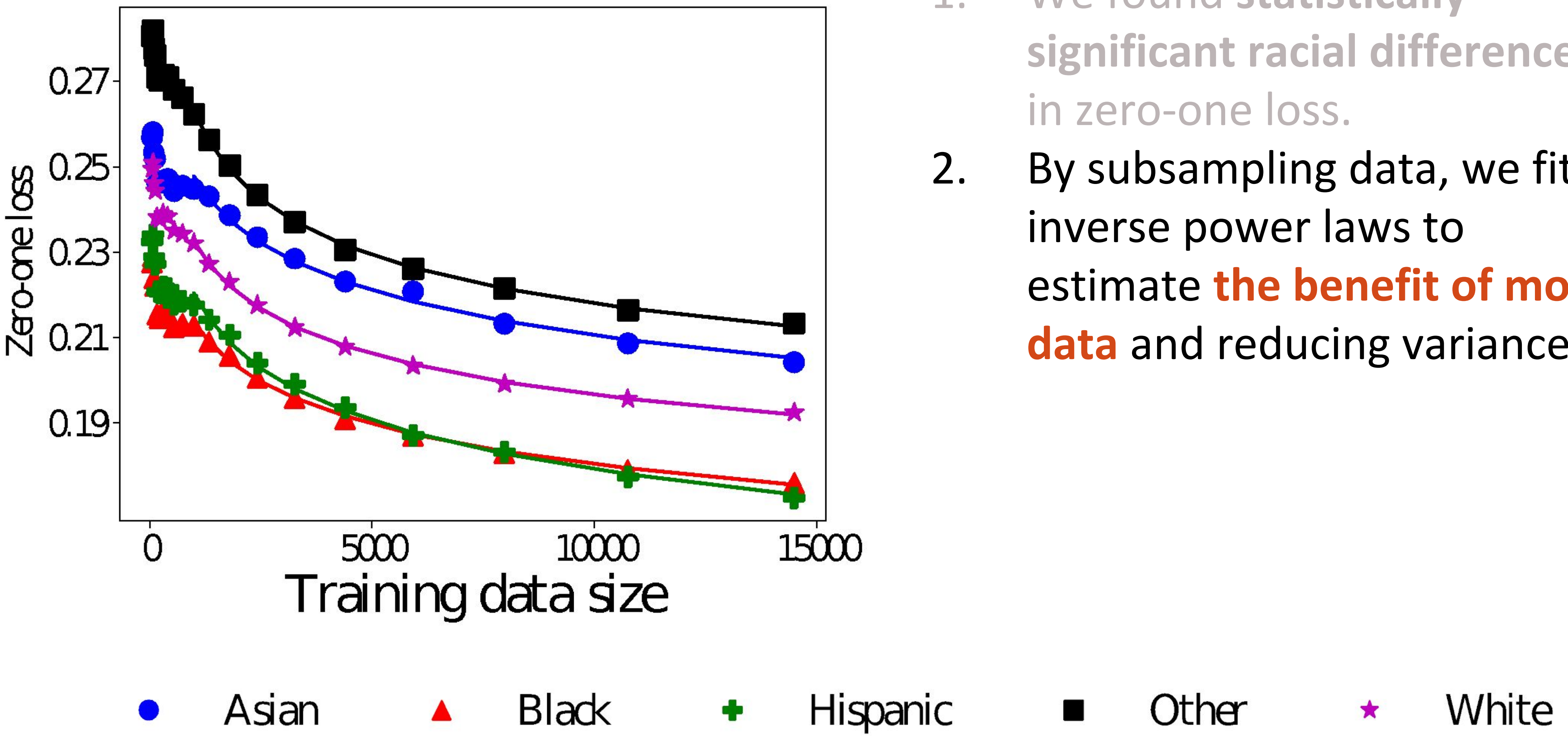
[Chen et al, 2018]

Mortality prediction from MIMIC-III clinical notes



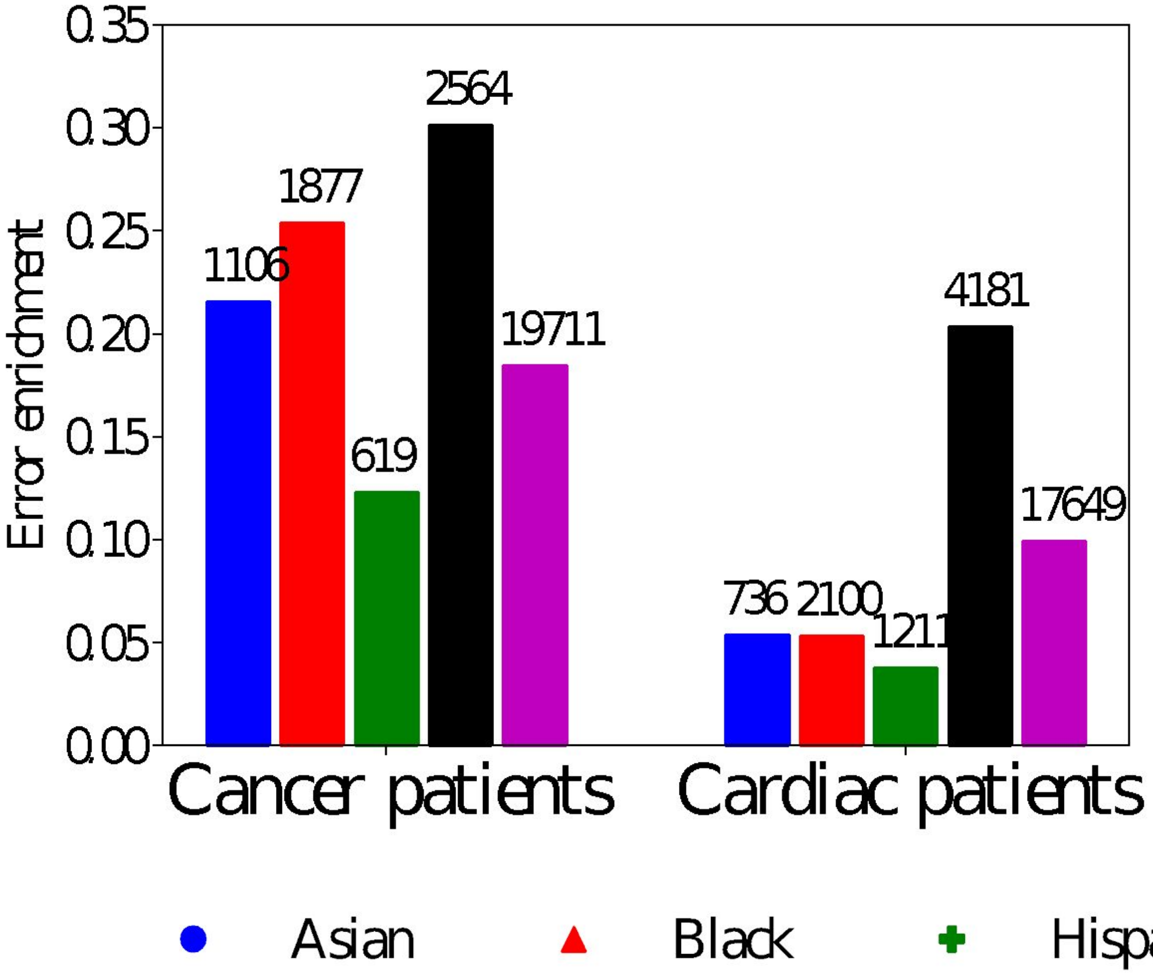
1. We found **statistically significant racial differences** in zero-one loss.

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.

Other Fairness in Healthcare

- **Dermatology:** “AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind” (The Atlantic, Aug 2018)
- **Clinical trials population:** “Clinical Trials Still Don’t Reflect the Diversity of America” (NPR, Dec 2015)
- **End of life care**: “Modeling Mistrust in End-of-Life Care” (MLHC 2018)
- **Alzheimer’s detection from speech:** “Technology analyzes speech to detect Alzheimer’s” (YouAreUNLTD, May 2018)
- **Cardiovascular Disease**: **”**Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk” (Annals of Internal Medicine, July 2018)

What’s next?

- How should we define fairness? How should it differ for healthcare, criminal justice, or other fields?
- What does it mean to study fairness or un-fairness?
- How can we “certify” fairness? If smaller components are all fair, does that mean the composite is fair?
- What does auditing a model entail? How might a model’s intended use and training data differ?
- What are protected groups? What about intersectionality?
- What about downstream effects over time? How can humans help or hurt?