CSC 2541: Machine Learning for Healthcare

Lecture 5: Fairness, Ethics and Healthcare

Professor Marzyeh Ghassemi, PhD
University of Toronto, CS/Med
Vector Institute

Course Reminders!

- Submit the weekly reflection questions to MarkUs!
- Project proposals, Feb 6 at 5pm!
- Problem Set 2, Feb 14 at 11:59pm!

Logistics

- Course website:
  https://cs2541-ml4h2020.github.io

- Piazza:
  https://piazza.com/utoronto.ca/winter2020/csc2541

- Grading:
  - 20% Homework (2 problem sets)
  - 10% Weekly reflections on Markus (5 questions)
  - 10% Paper presentation done in-class (sign-up after the first lecture)
  - 60% course project (an eight-page write up)

Schedule

Jan 9, 2020, Lecture 1: Why is healthcare unique?

Jan 16, 2020, Lecture 2: Supervised Learning for Classification, Risk Scores and Survival

Jan 23, 2020, Lecture 3: Clinical Time Series Modelling

Jan 30, 2020, Lecture 4: Causal inference with Health Data --- Dr. Shalmali Joshi (Vector)

Problem Set 1 (Jan 31 at 11:59pm)

Feb 6, 2020, Lecture 5: Fairness, Ethics, and Healthcare

Project proposals (Feb 6 at 5pm)

Feb 13, 2020, Lecture 6: Deep Learning in Medical Imaging -- Dr. Joseph Paul Cohen (MILA)

Feb 20, 2020, Lecture 7: Clinical Reinforcement Learning

Feb 27, 2020, Lecture 8: Clinical NLP and Audio -- Dr. Tristan Naumann (MSR)

Problem Set 2 (Feb 27 at 11:59pm)

Mar 5, 2020, Lecture 9: Interpretability / Humans-In-The-Loop --- Dr. Rajesh Ranganath (NYU)

Mar 12, 2020, Lecture 10: Disease Progression Modelling/Transfer Learning -- Irene Chen (MIT)

Mar 19, 2020, Project Sessions/Lecture

Mar 26, 2020, Course Presentations

April 4, 2020, Course Presentations

Project Report (Apr 3 at 11:59pm)

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

Health Questions Beyond The Obvious

Across these use cases, a number of ethical, social, and political challenges are raised and the 10 most important are:

01 What effect will AI have on human relationships in health and care?

02 How is the use, storage and sharing of medical data impacted by AI?

03 What are the implications of issues around algorithmic transparency/explainability on health?

04 Will these technologies help eradicate or exacerbate existing health inequalities?

05 What is the difference between an algorithmic decision and a human decision? 6

06 What do patients and members of the public want from AI and related technologies?

07 How should these technologies be regulated?

08 Just because these technologies could enable access to new information, should we always use it?

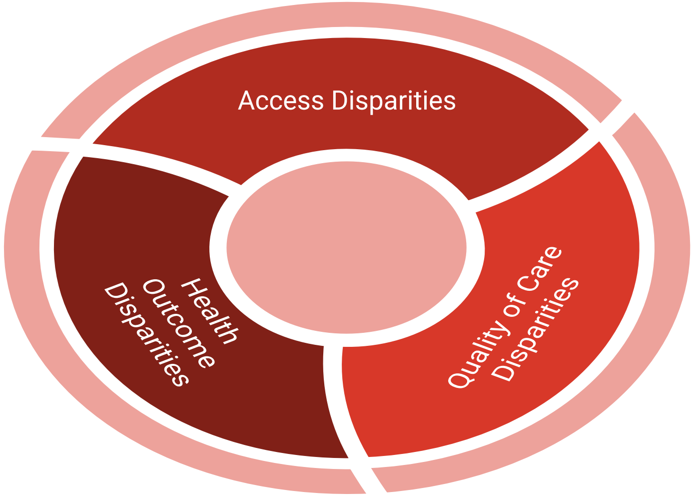
09 What makes algorithms, and the entities that create them, trustworthy?

10 What are the implications of collaboration between public and private sector organisations in the development of these tools?

[1] "Ethical, social, and political challenges of artificial intelligence in Health". *Wellcome Trust* April 2018. © Future Advocacy.

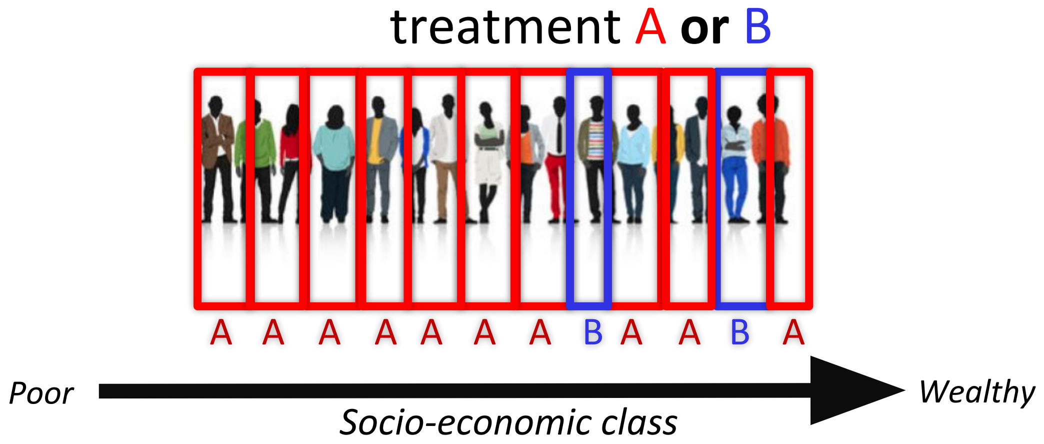
Inequality in Healthcare; A Categorization.

- Inequality of **access**
  - Mary gets to see a category of doctor that Ian doesn't.
- Inequality of **treatment**
  - Mary and Ian see the same category of doctor, but are given different treatments.
- Inequality of **outcome**
  - Given the same treatment, Mary recovers and Ian doesn't because of Ian's existing social determinants.



How Can We Improve Health Care For **All**?

- Patient populations have differences in treatment by race, sex, and socioeconomic status



- Are there differences in prediction accuracy by group?

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)
- **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.

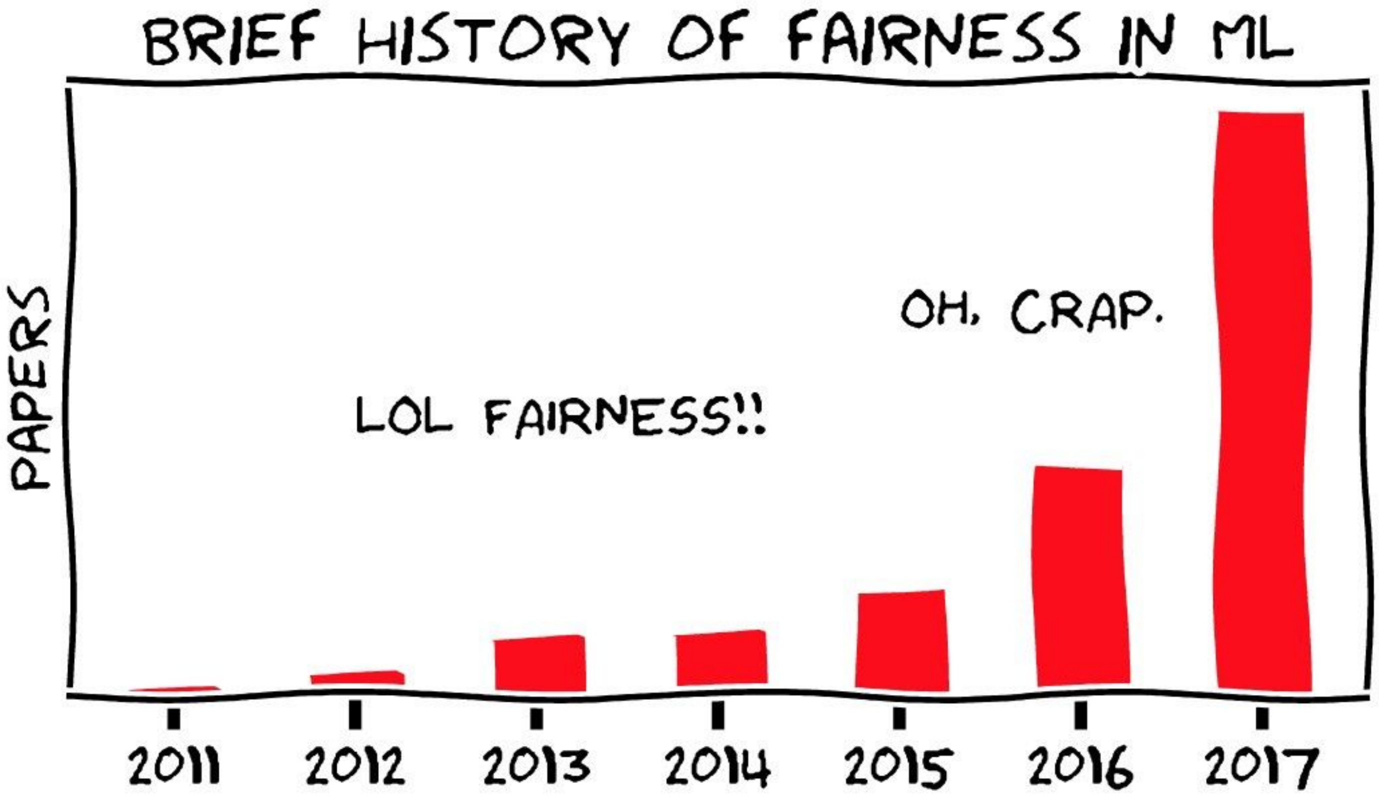
Yadlowsky S[1], Hayward RA[2], Sussman JB[2], McClelland RL[3], Min YI[4], Basu S[5].

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
LOL FAIRNESS!!
OH. CRAP.
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

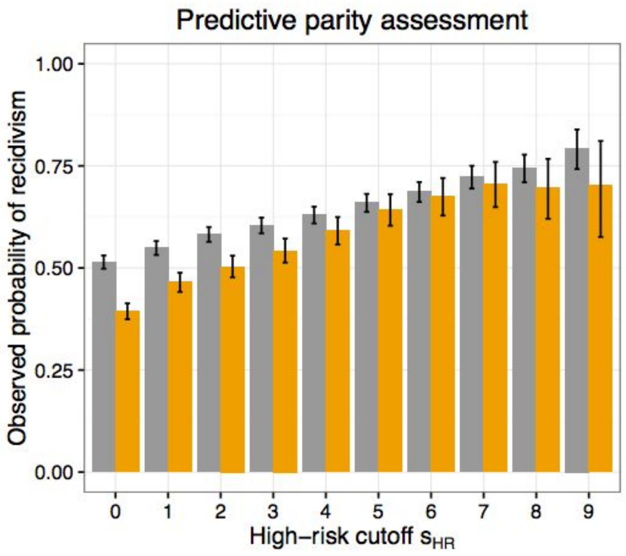
- **Plan**: Require same fraction of $\hat{Y} = 1$ for each group A
- **Example**: Predict diabetes risk Y from clinical features X and race A such that $P(\hat{Y} = 1 | A = 1) = P(\hat{Y} = 1 | A = 0)$
- **Problems:**
  - What if true Y perfectly correlates with A?
  - Too strong: even perfect prediction $Y = \hat{Y}$ 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



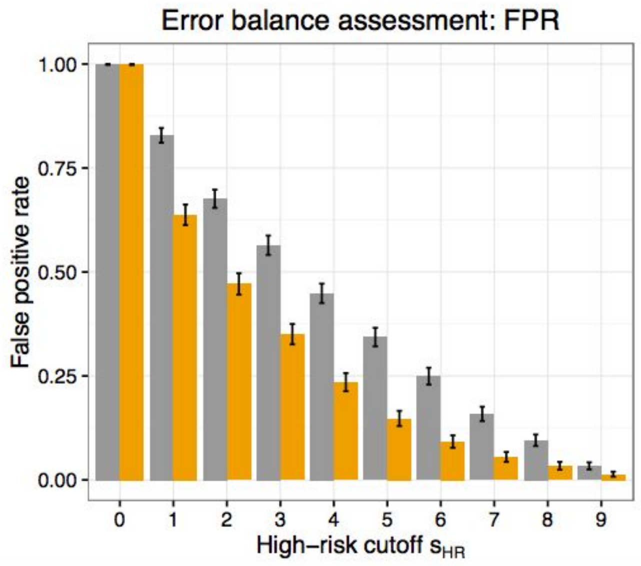
[Chouldechova, 2018]

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



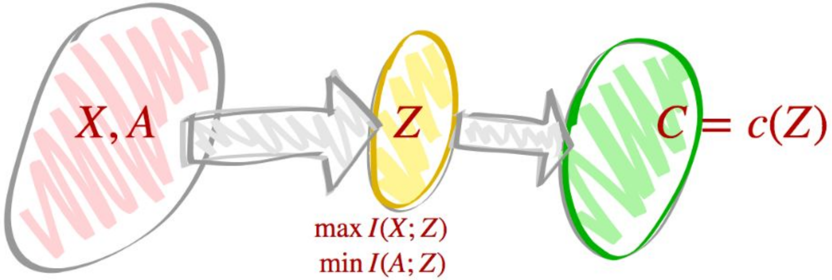
[Chouldechova, 2018]

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

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

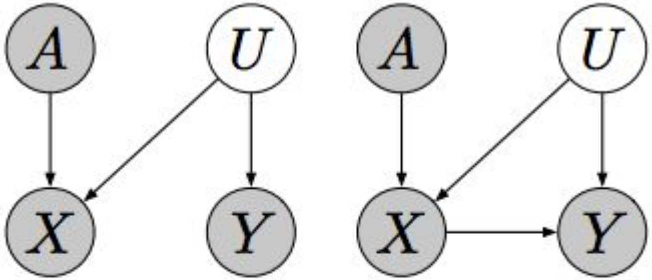
- **Problems:**
  - How to ensure you are not losing too much info and learning right representation?



[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



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

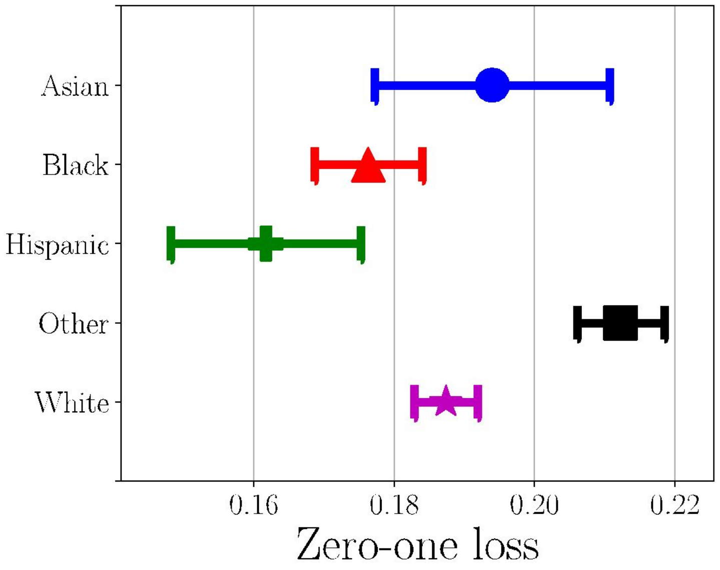
- **Problems:**
  - Creating a structural model encodes prior beliefs about world
  - Causal inference often requires ignorability assumptions

[Kusner et al, 2017]

What about the data?

Predicting hospital mortality in MIMIC

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

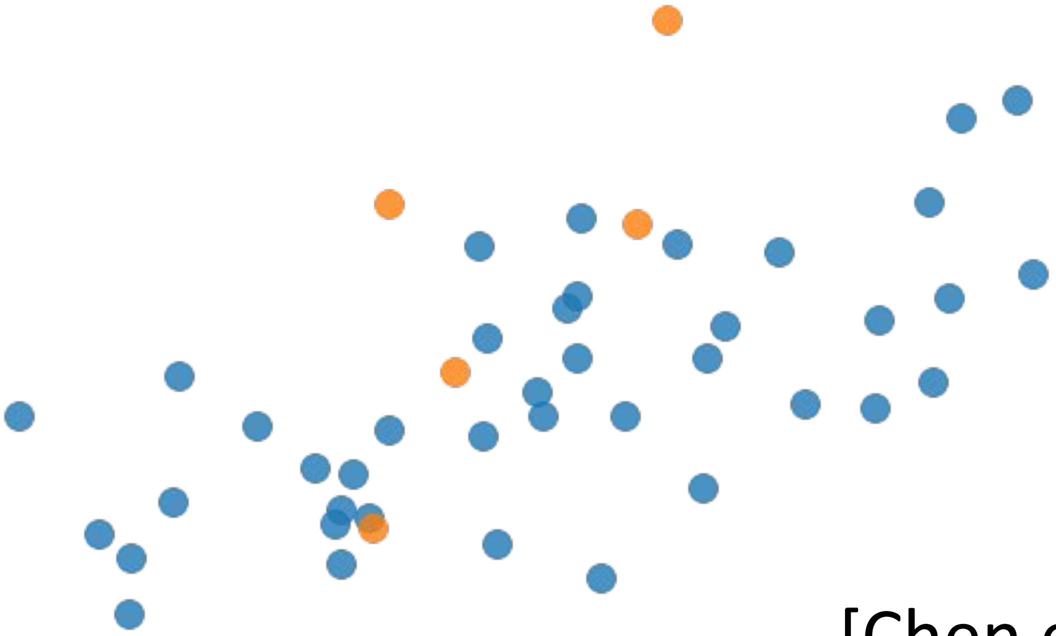


[Chen et al, 2018]

Why might my classifier be unfair?

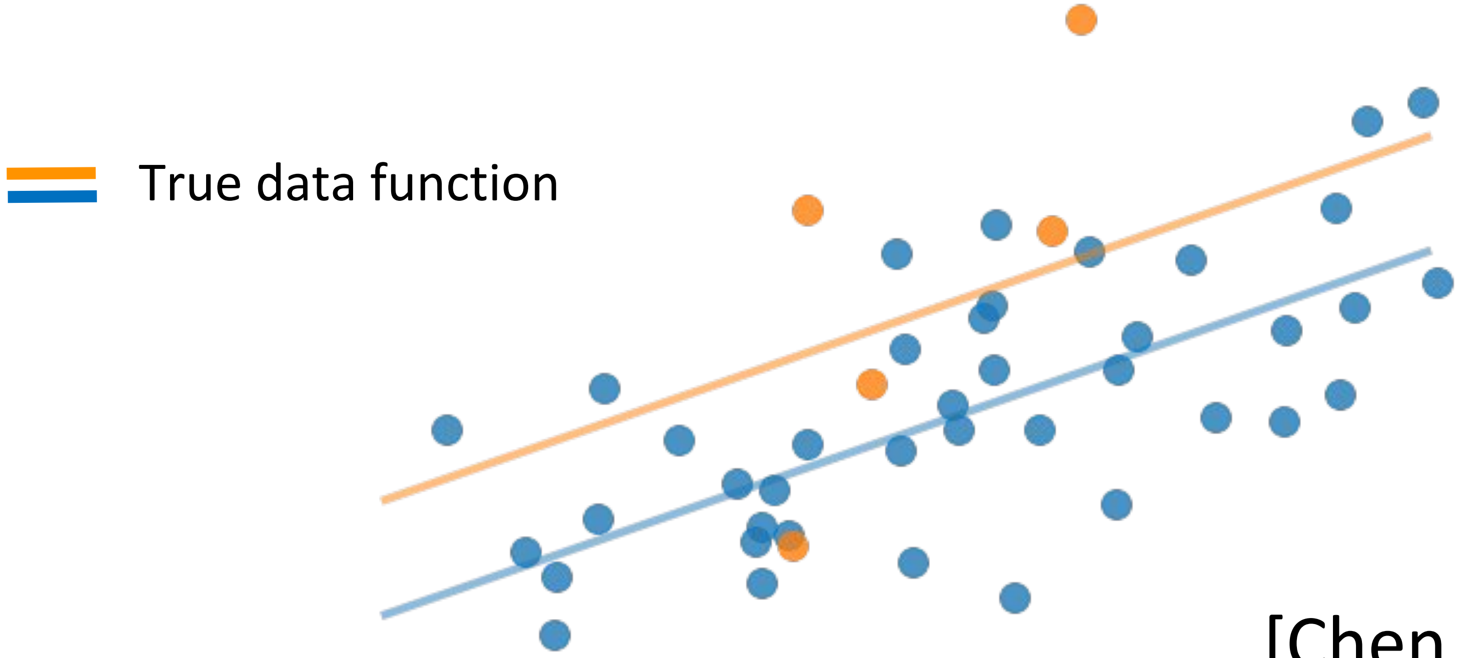
[Chen et al, 2018]

Why might my classifier be unfair?



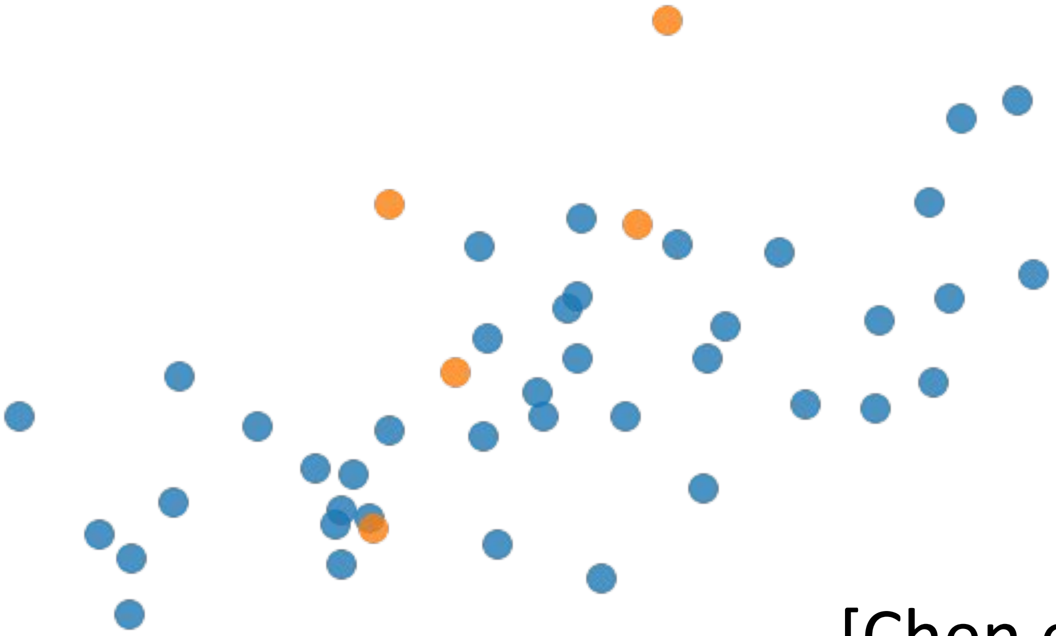
[Chen et al, 2018]

Why might my classifier be unfair?



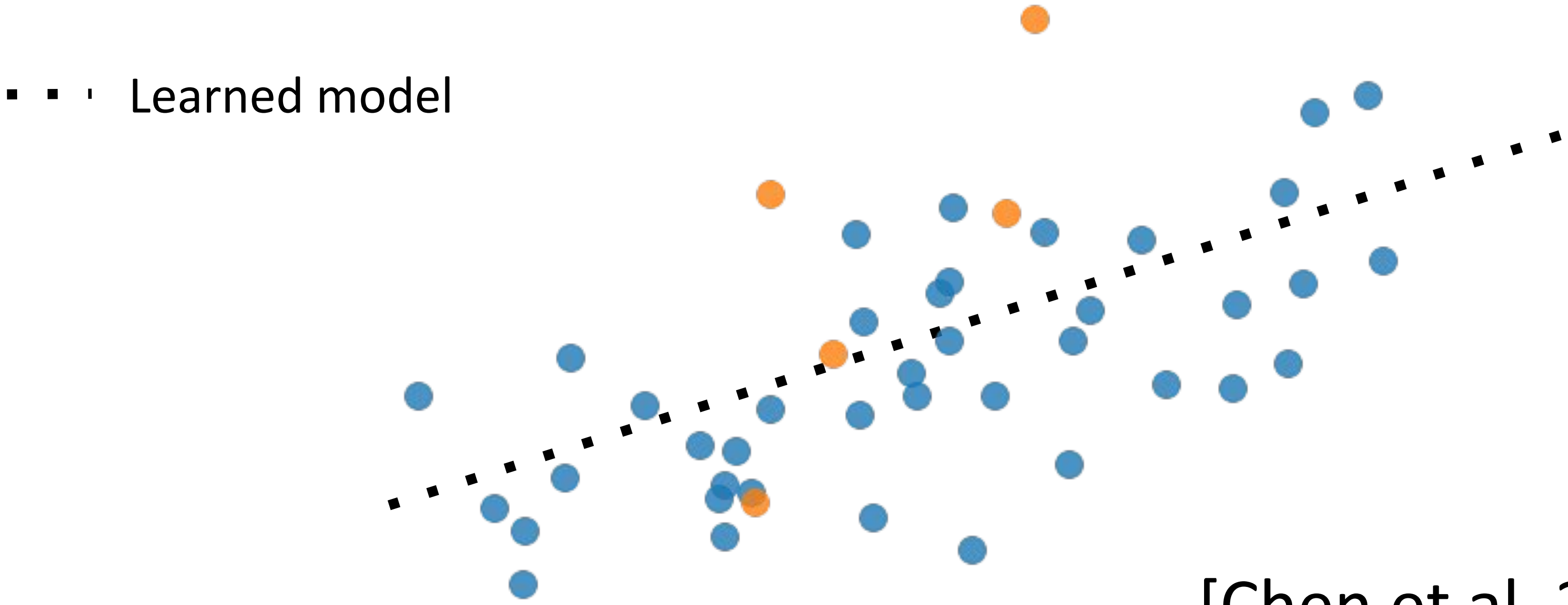
[Chen et al, 2018]

Why might my classifier be unfair?



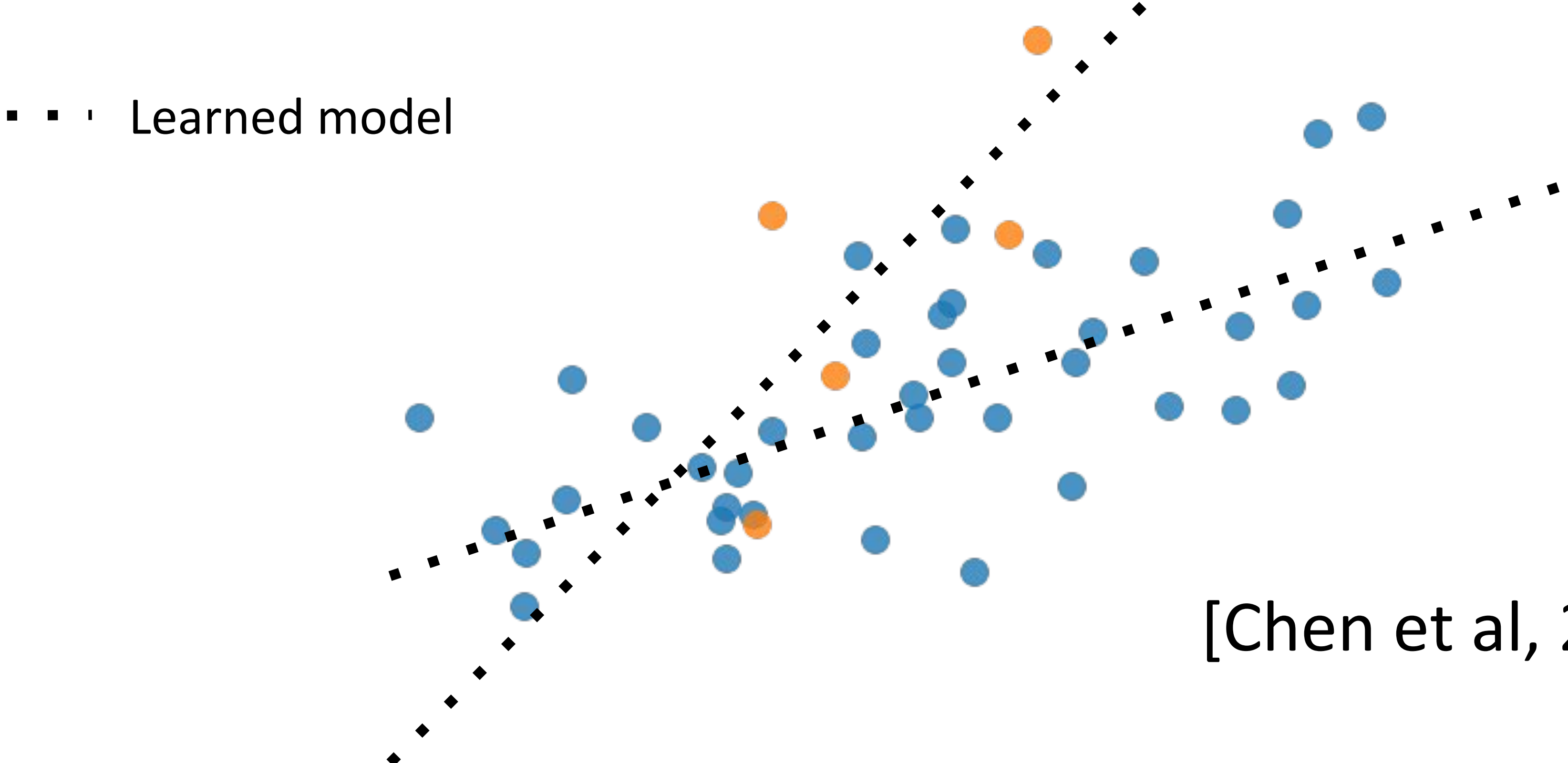
[Chen et al, 2018]

Why might my classifier be unfair?



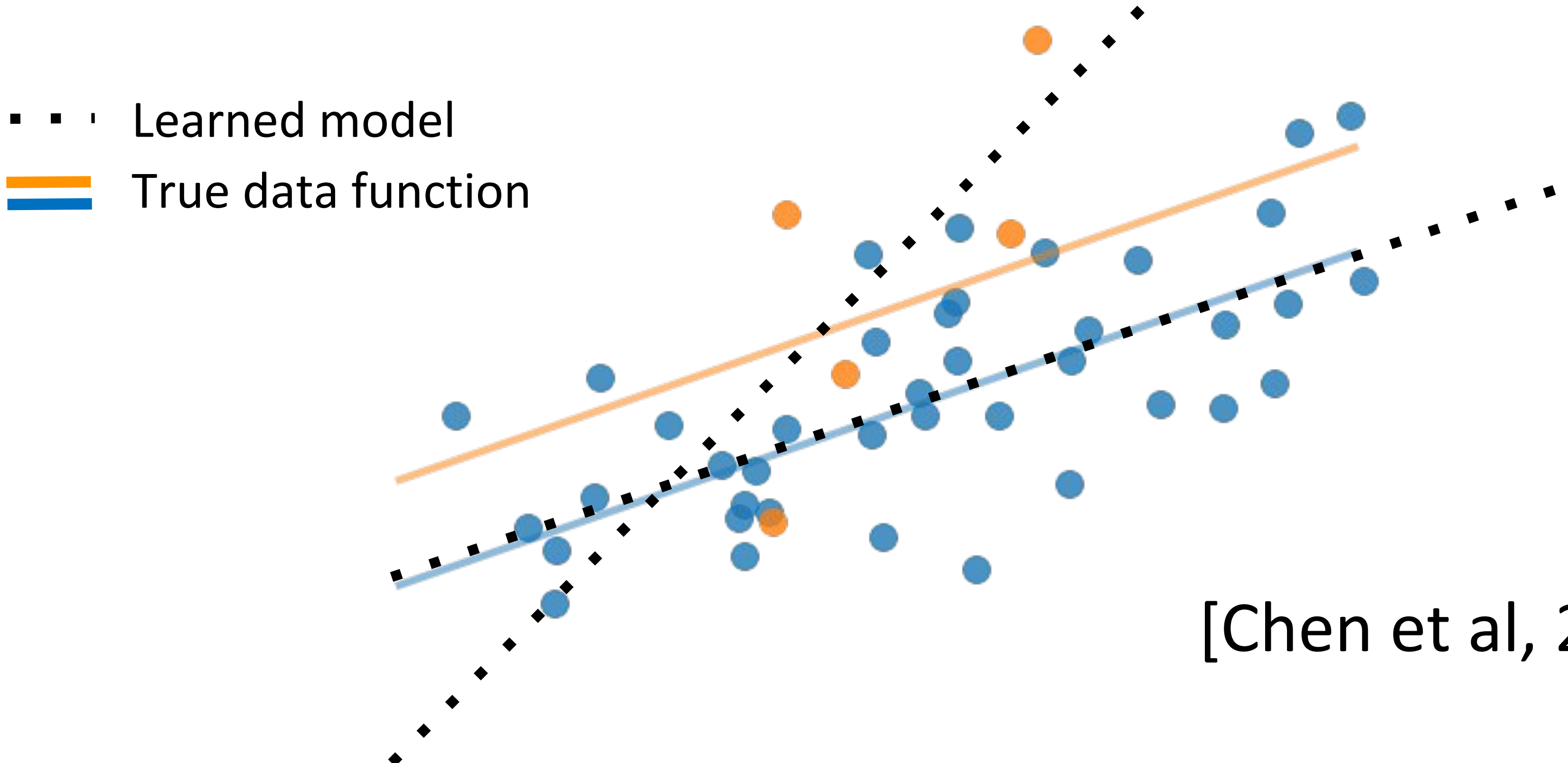
[Chen et al, 2018]

Why might my classifier be unfair?



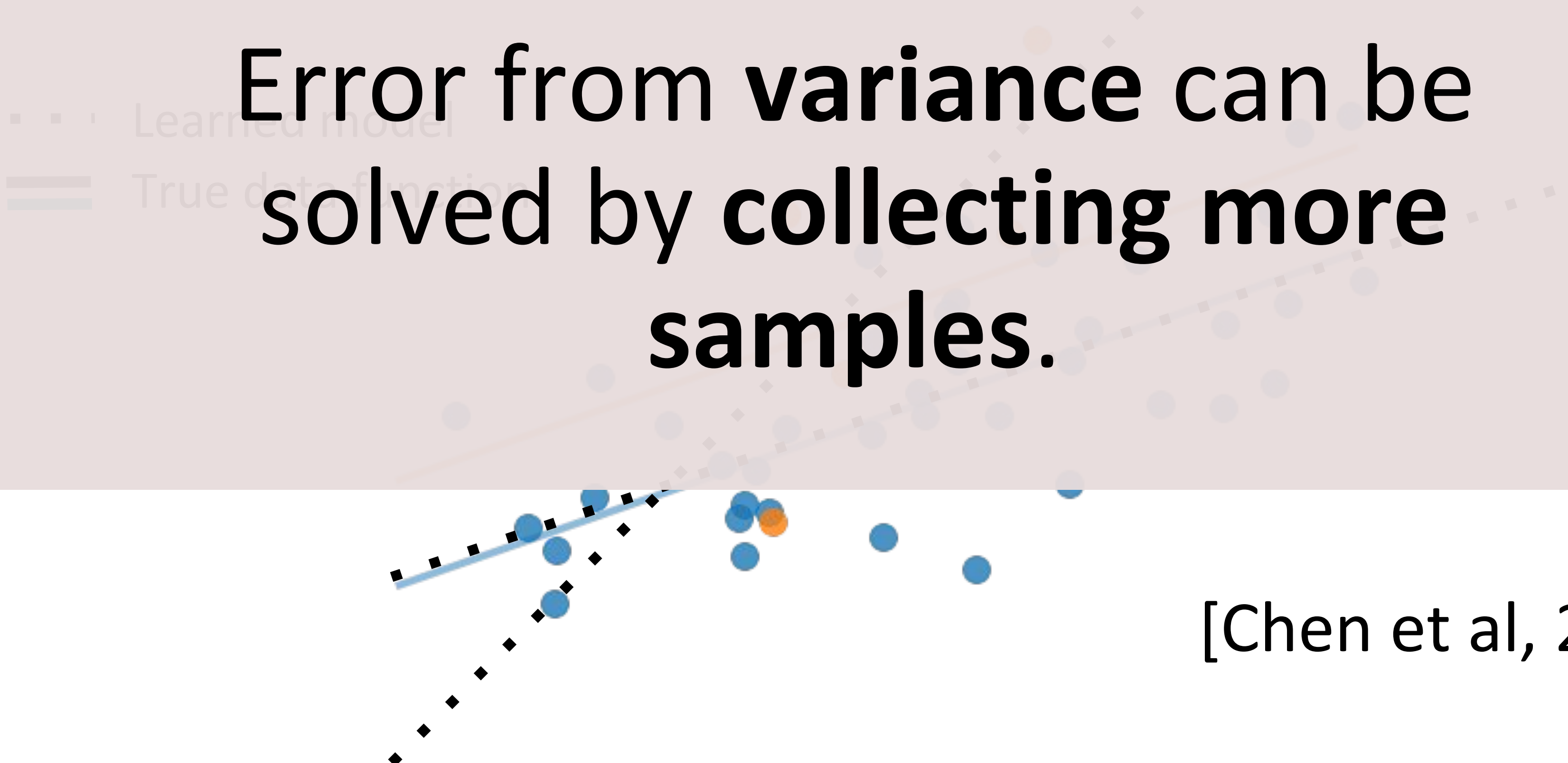
[Chen et al, 2018]

Why might my classifier be unfair?



[Chen et al, 2018]

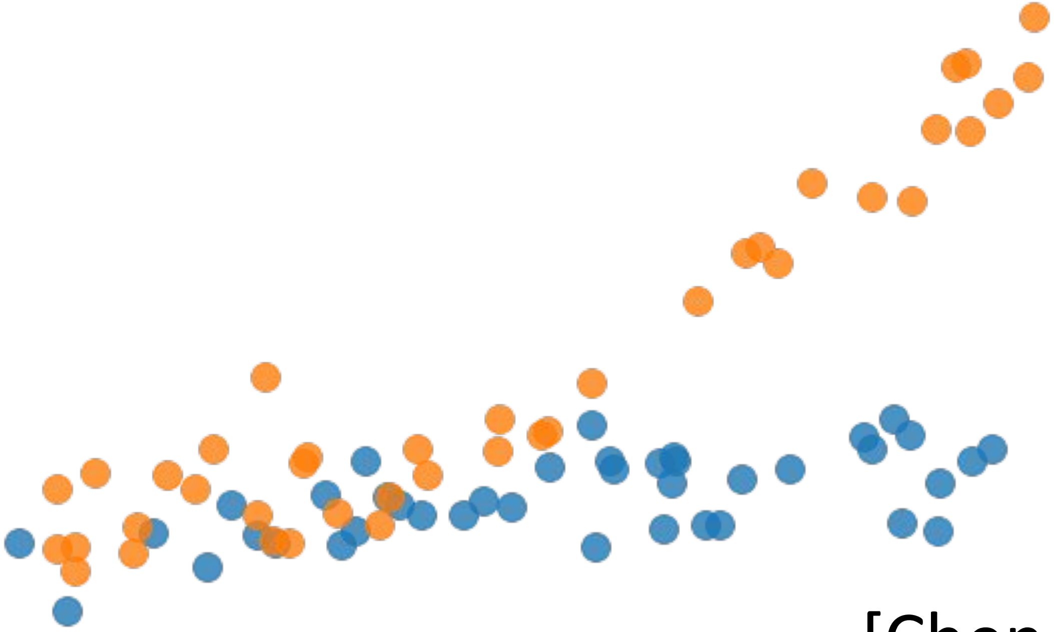
Why might my classifier be unfair?



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

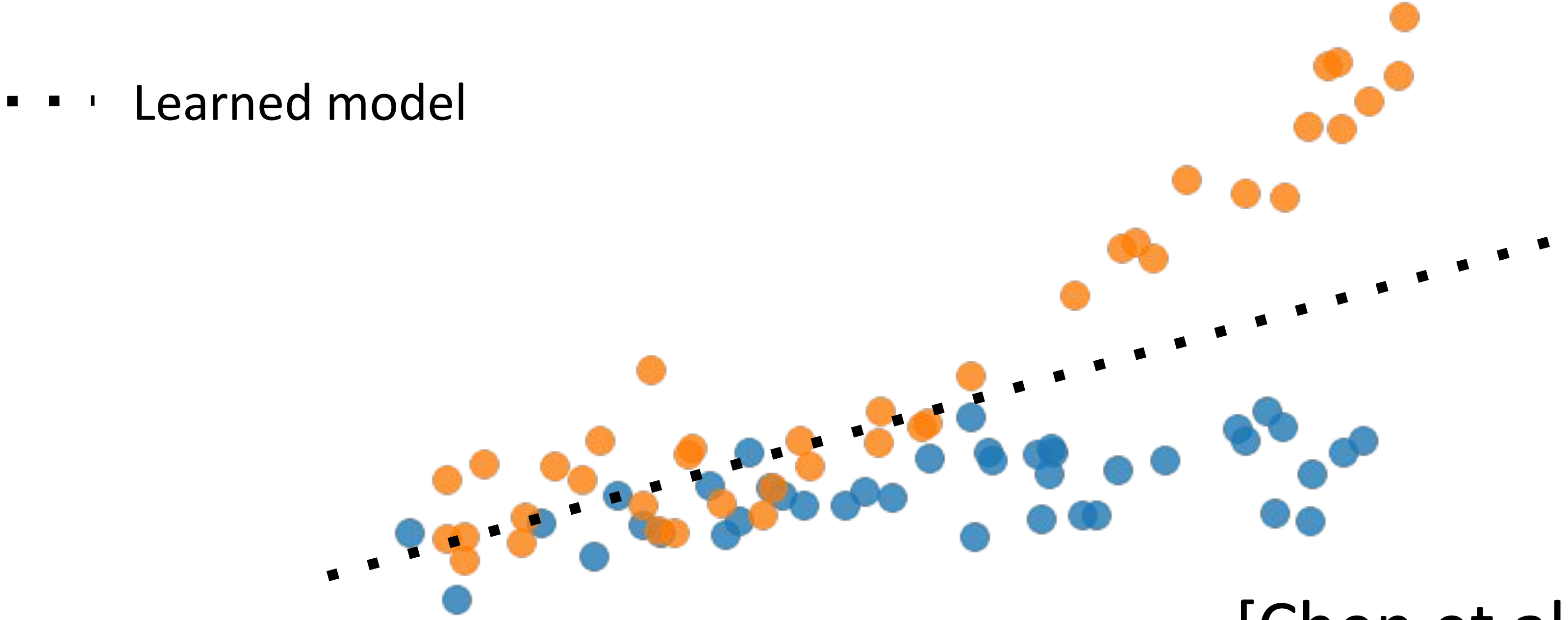
[Chen et al, 2018]

Why might my classifier be unfair?



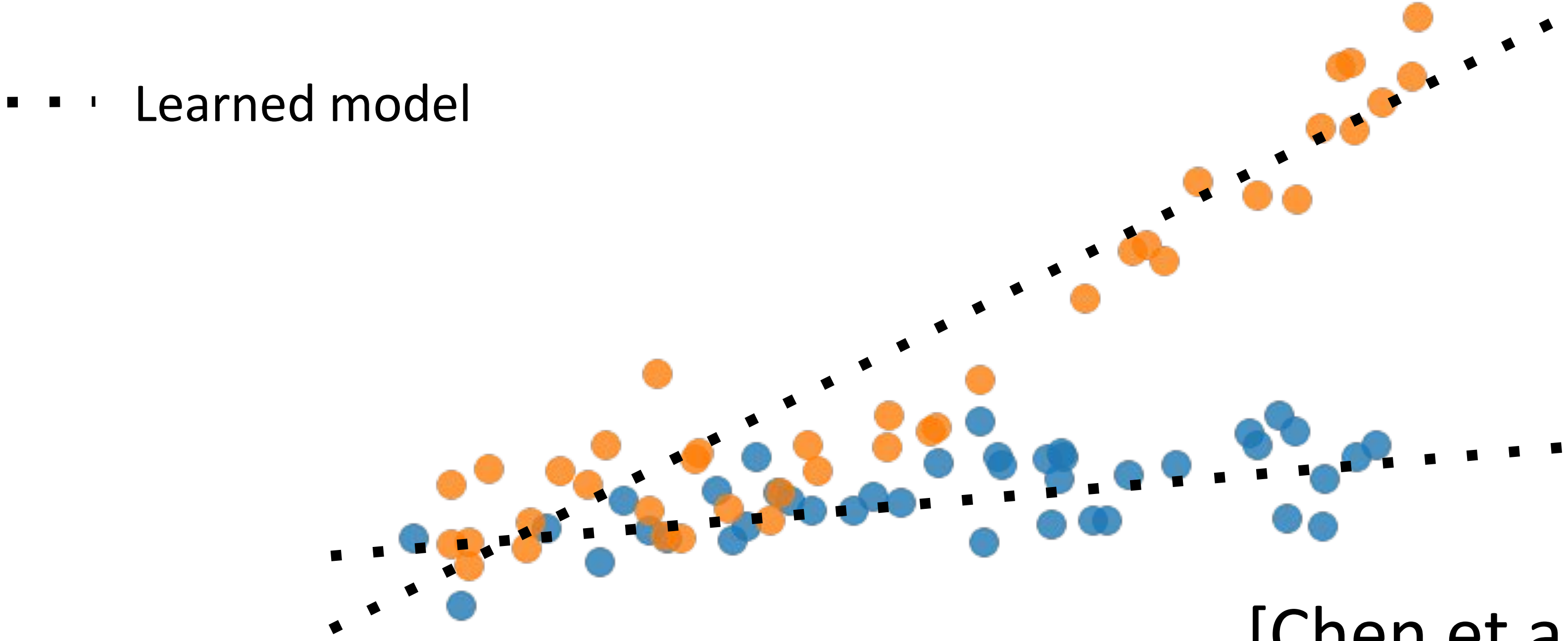
[Chen et al, 2018]

Why might my classifier be unfair?



[Chen et al, 2018]

Why might my classifier be unfair?

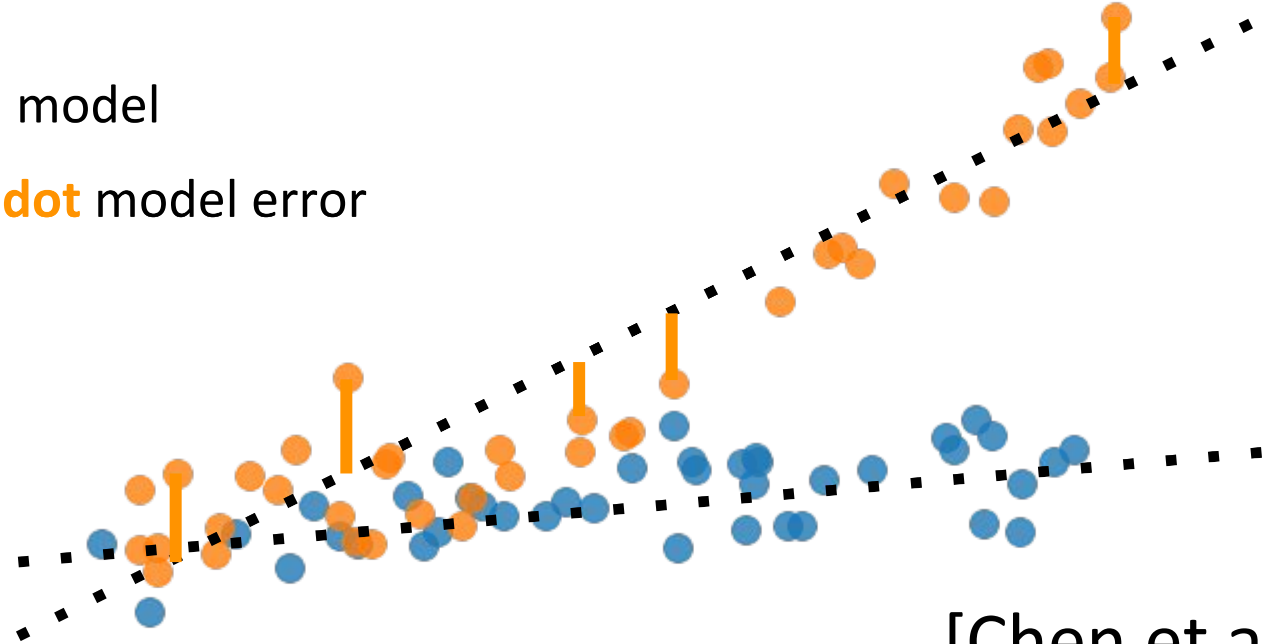


[Chen et al, 2018]

Why might my classifier be unfair?

Learned model

Orange dot model error



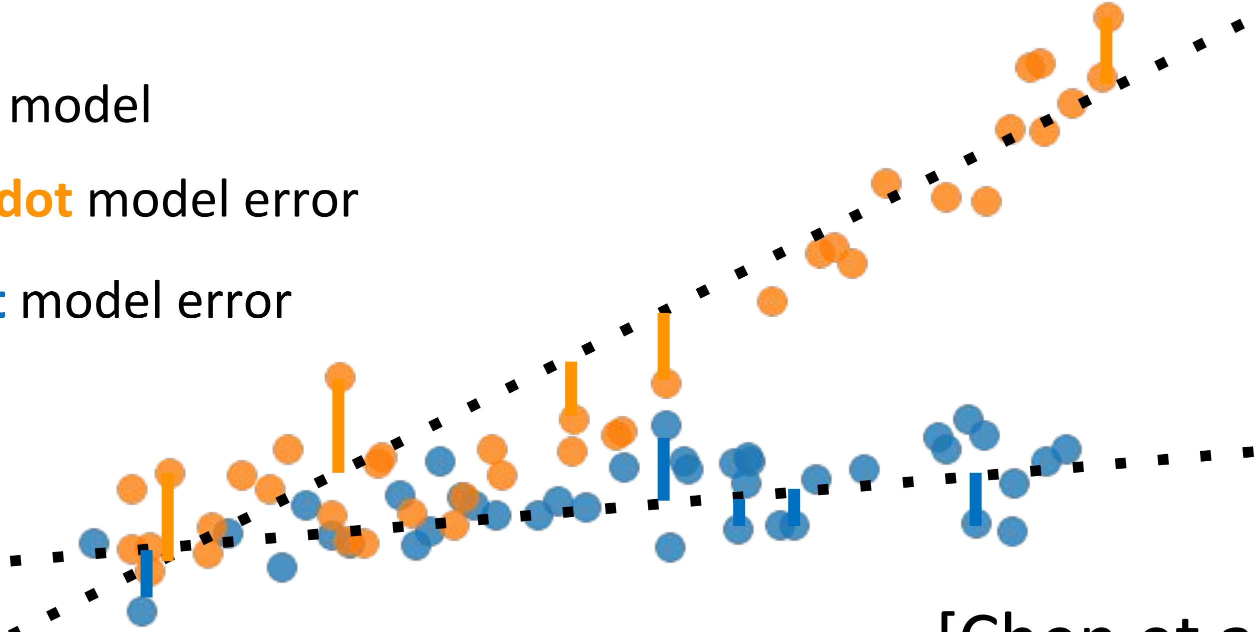
[Chen et al, 2018]

Why might my classifier be unfair?

Learned model

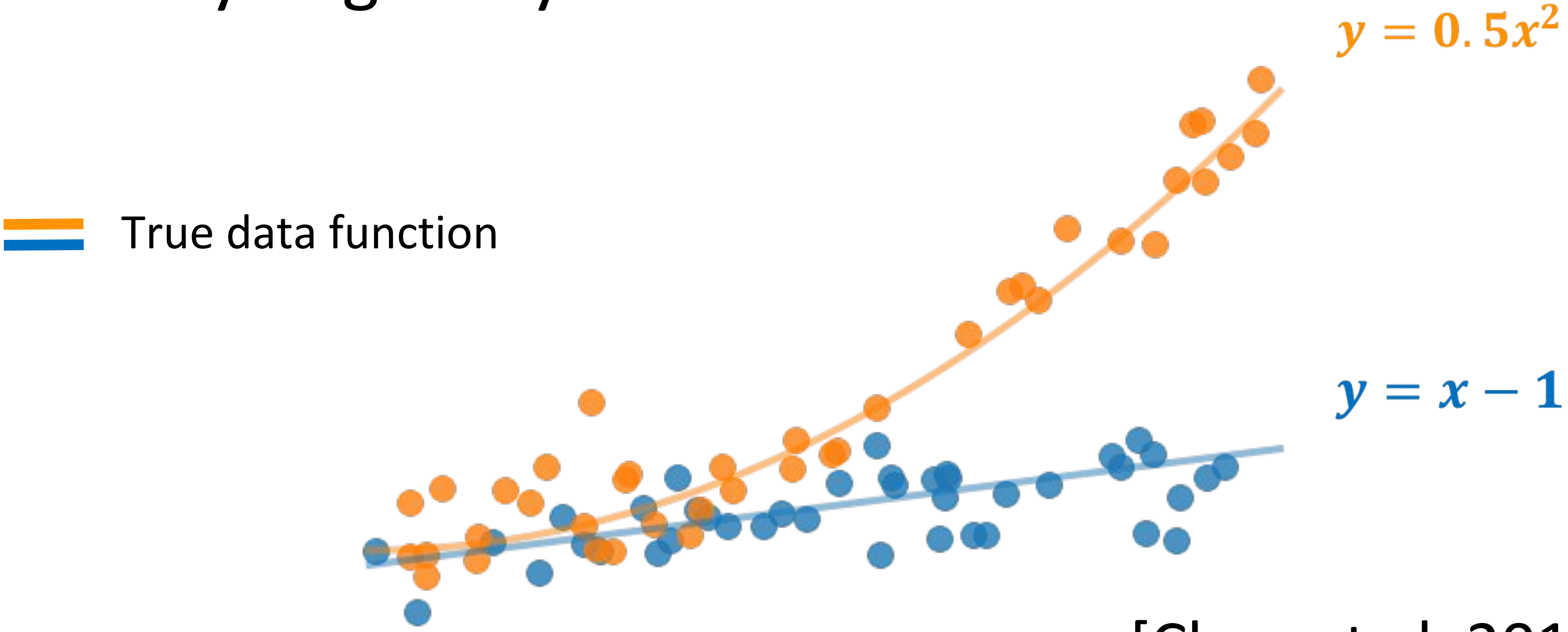
Orange dot model error

Blue dot model error



[Chen et al, 2018]

Why might my classifier be unfair?



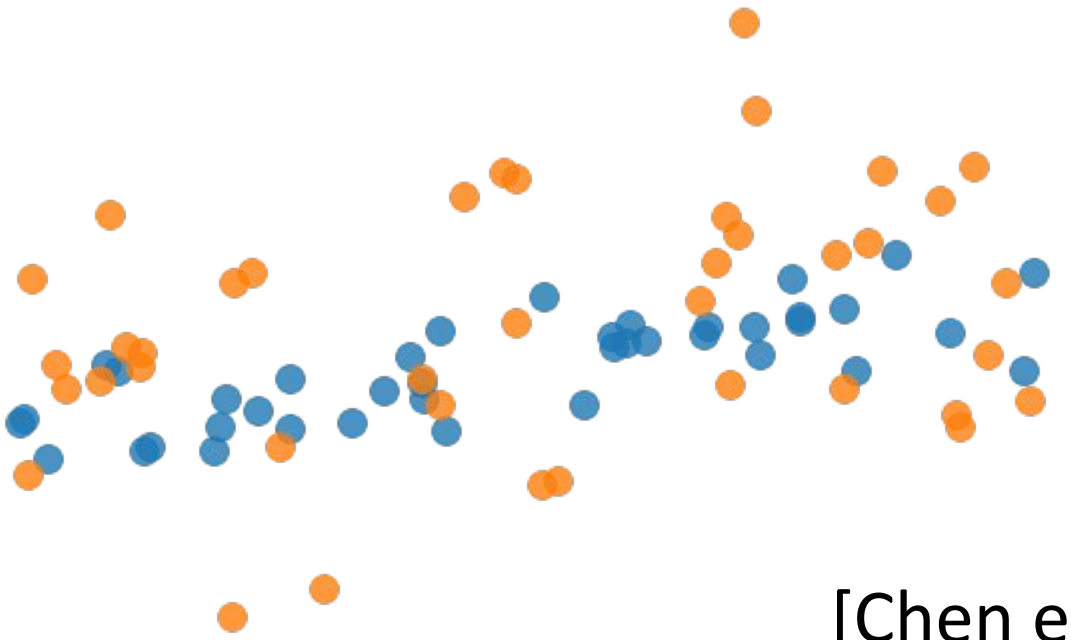
[Chen et al, 2018]

Why might my classifier be unfair?

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

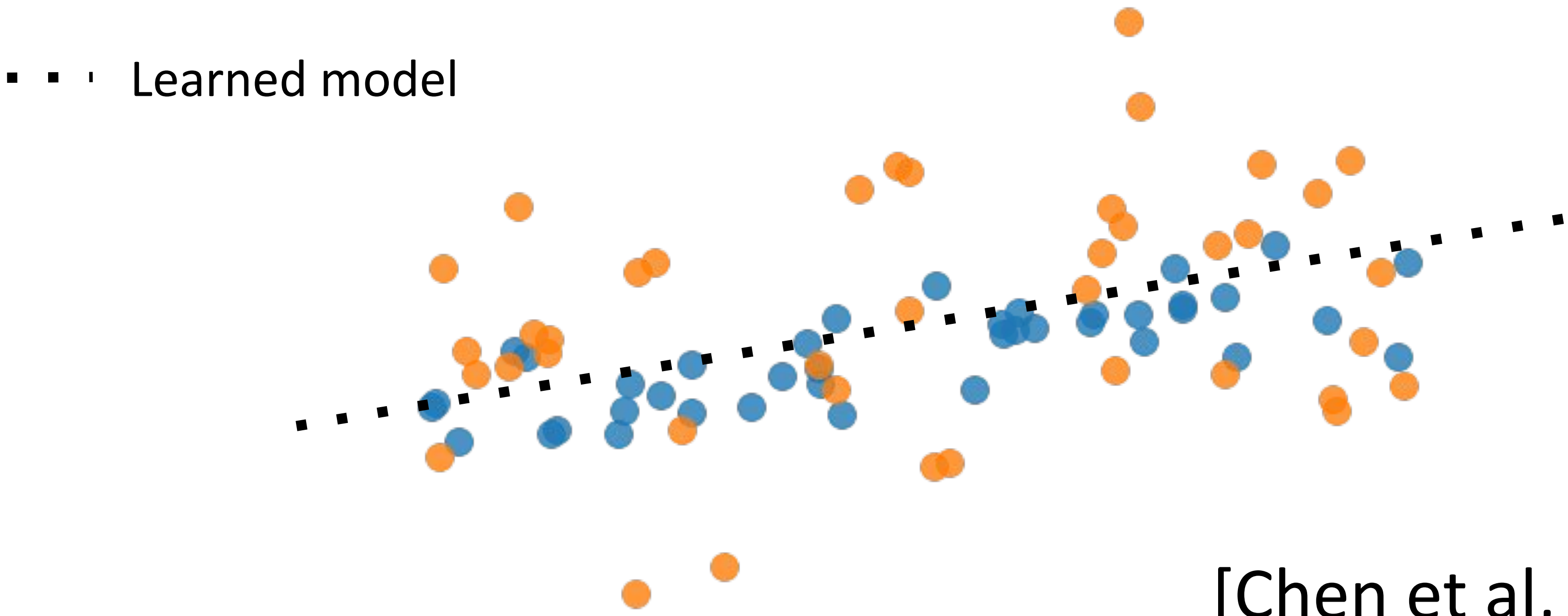
[Chen et al, 2018]

Why might my classifier be unfair?



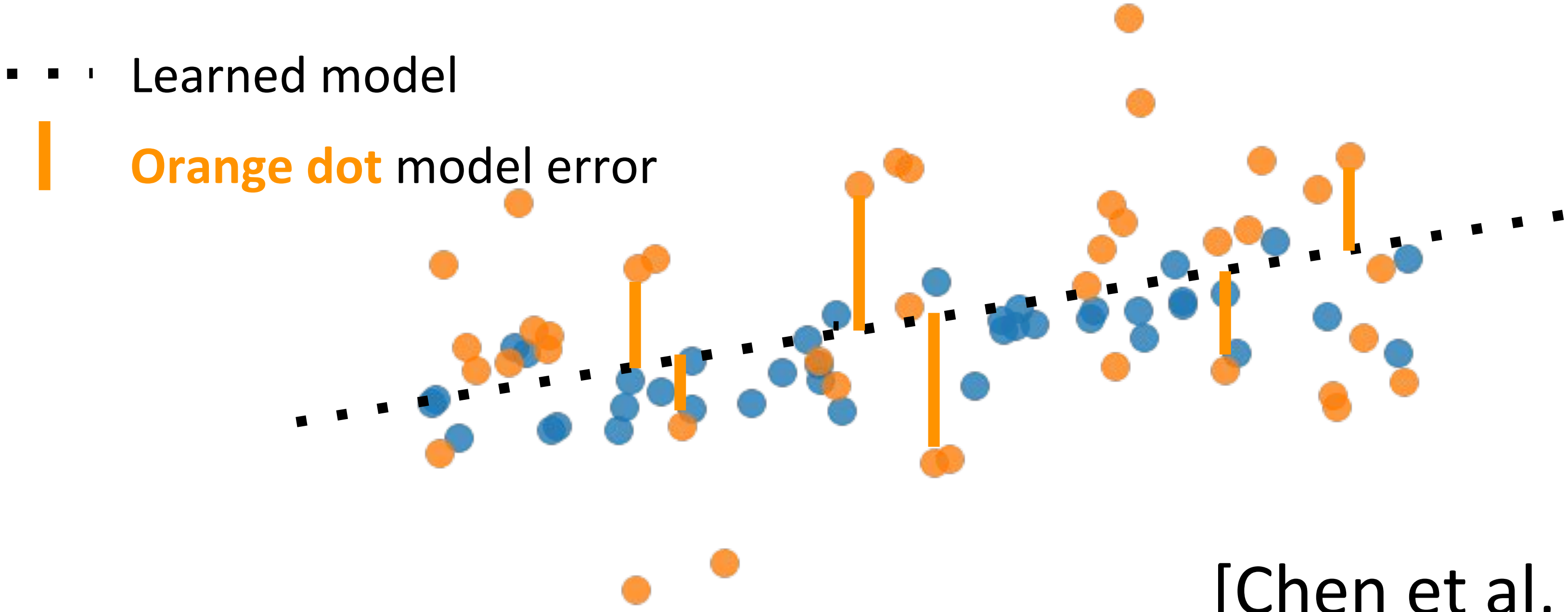
[Chen et al, 2018]

Why might my classifier be unfair?



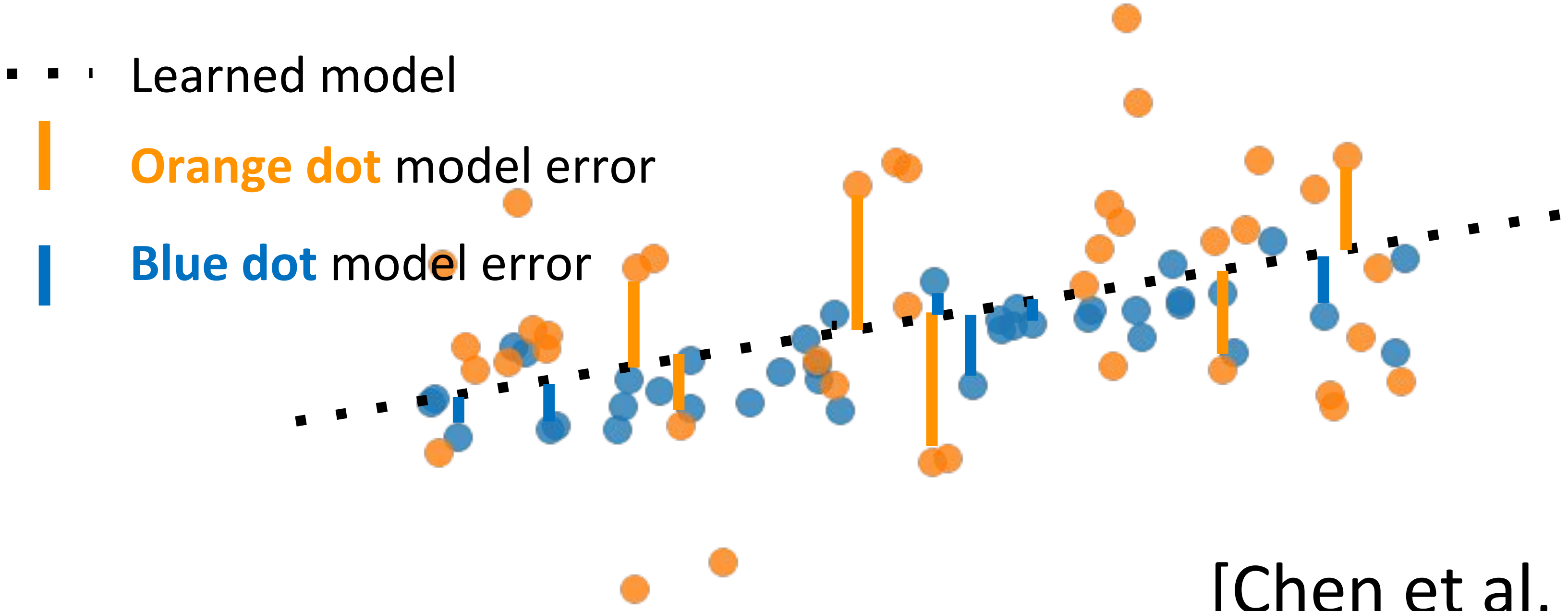
[Chen et al, 2018]

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 $\hat{Y}$:

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

[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 $\hat{Y}$:

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

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

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

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 $\hat{Y}$:

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

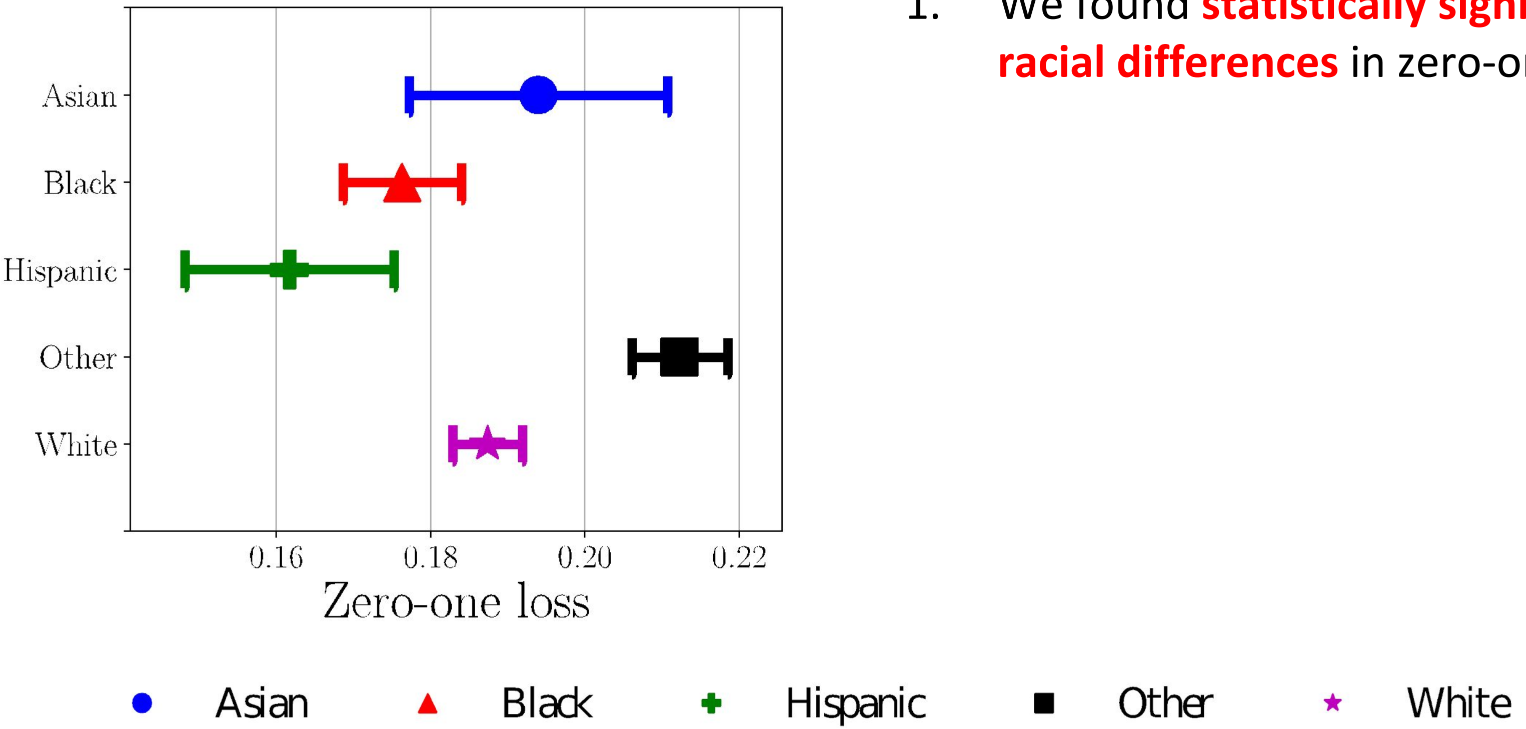
Note that $\overline{N_a}$ indicates the expectation of N_a over *X* and data *D*.

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.

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

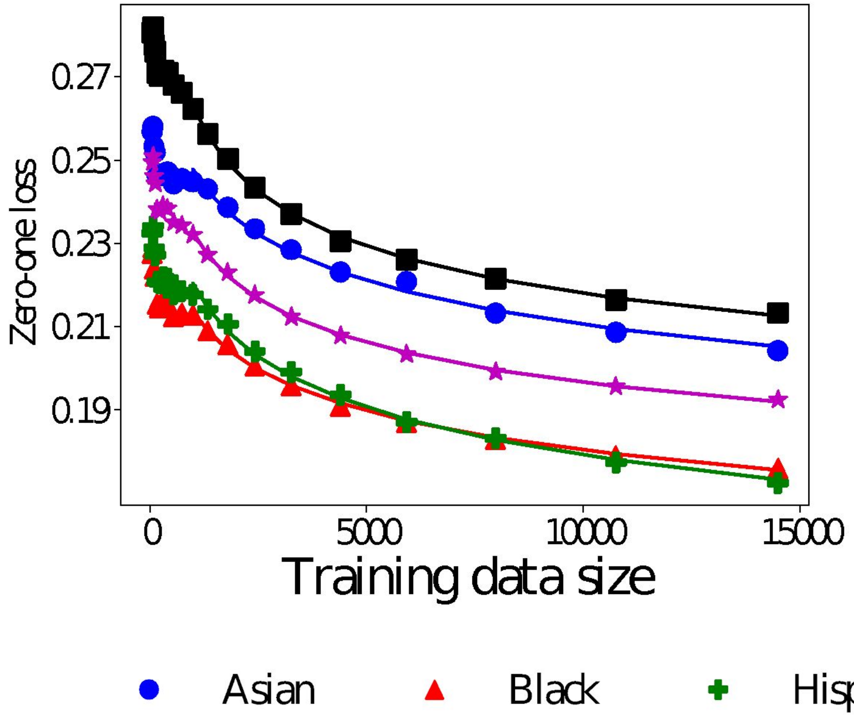
[Chen et al, 2018]

Mortality prediction in MIMIC-III clinical notes



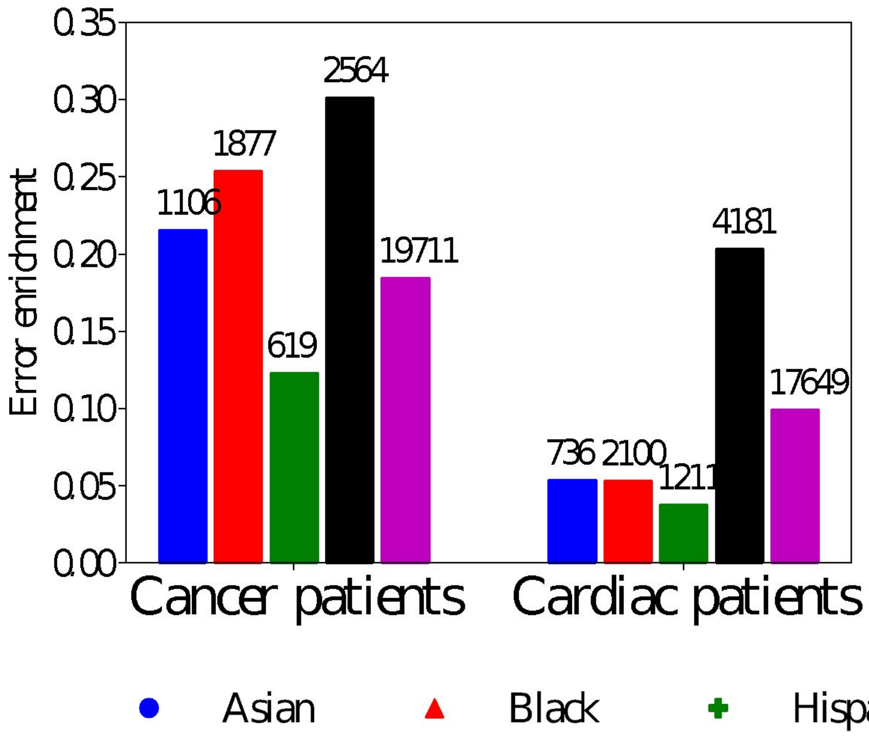
1. We found **statistically significant racial differences** in zero-one loss.

Mortality prediction in 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 in 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 in healthcare, criminal justice, or other fields?
- What does it mean to study fairness or un-fairness?
- How can we “certify” fairness?
- 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?

Sidebar - Ethics in Helping Human Decision Making

- Ultimately, the goal is **improved care**.

- Example: Software designed to improve OB decision making during labour did **not improve clinical outcomes**.

> "Use of computerised interpretation of cardiotocographs in women who have continuous electronic fetal monitoring in labour does not improve clinical outcomes for mothers or babies."

- Human decisions about routine practice will need to be **justified with or without ML**.

Brocklehurst P, Field D, Greene K, Juszczak E, Keith R, Kenyon S, Linsell L, Mabey C, Newburn M, Plachcinski R, Quigley M. Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial. The Lancet. 2017 Apr 29;389(10080):1719-29.

We Can Get People To Trust Explanations

- Trust is a **process** rather than a status, and that systems should be designed as to allow for maintenance of that expectation rather than reaching a state.

- In robotics, there has been work demonstrating that, humans tend to **overtrust robotic systems** in scenarios where

    - 1) a person accepts risk because that person believes the **robot can perform a function that it cannot** or

    - 2) the person accepts too much risk because the expectation is that the **system will mitigate the risk**.

[1] http://www.jeffreymbradshaw.net/publications/50.%20Trust%20in%20Automation.pdf
[2] https://dl.acm.org/doi/10.1145/3278721.3278786

Explainable AI In Health Is A Bad Idea

- Recent work on the interplay between ML/**human decisions** found

  ‘no significant improvement in the degree to which people follow the predictions of a "clear" model with few features compared to the other experimental conditions’.

- Worse, models with more "transparency" **hampered people’s ability** to detect when a model **makes serious mistakes**.

- Models that are more "transparent" can make people **feel like the choice is good**, and therefore don't do a more aggressive audit.

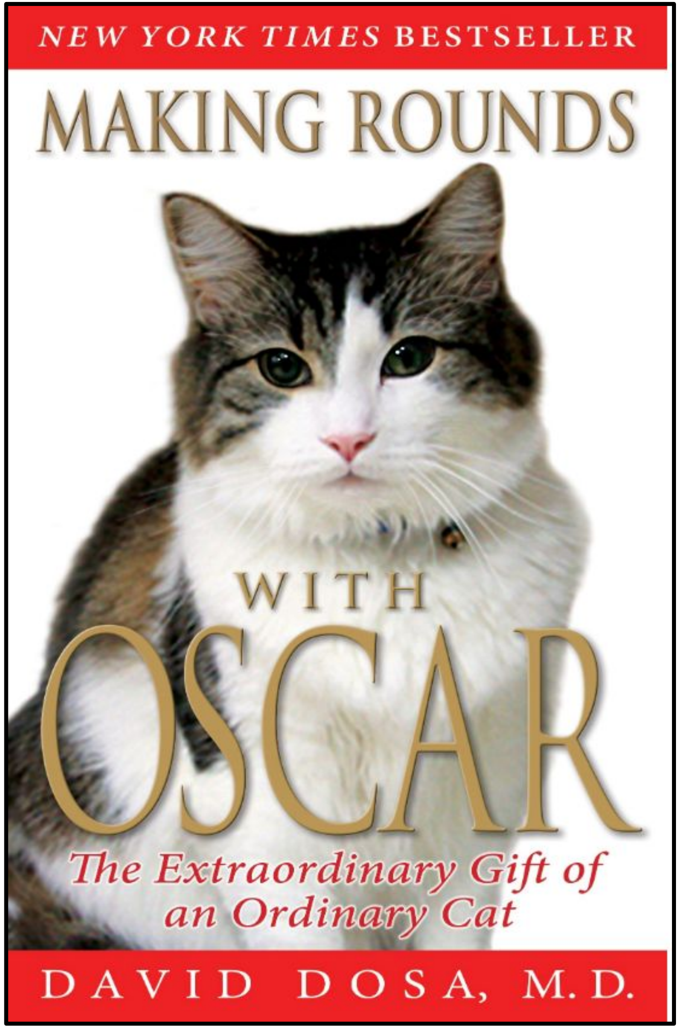
Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Vaughan JW, Wallach H. Manipulating and measuring model interpretability. 2018. (https://arxiv.org/pdf/1802.07810.pdf)

If The Cat Can Do It

- Oscar the cat, who appeared able to,

  "predict the impending death of terminally ill patients"

  by choosing to nap next to people a few hours before they die.



https://www.ncbi.nlm.nih.gov/pubmed/17652647