

Austin van Loon

Outline for lecture today

- The Morals of Prediction
- Defining and Measuring Fairness
- Sources of Bias
 - Insider Bias
 - Outsider Bias (in-class activity)
- Machine Learning and Inequality
- Week 2 review survey



The Morals of Prediction

Predictions (algorithmic and non-algorithmic), especially from those in power, can greatly impact people's lives

- Banks predict whether you will be able to pay off a loan
- Insurance companies predict your likelihood of filing a claim
- Police predict whether you have paraphernalia on you
- Online retailers predict the highest price you will buy something at
- Judges predict whether you are going to come back to court

The Morals of Prediction



Human Decision-makers

- We ask them to give an account of their decisions
- We develop constructs such as the “reasonable person” standard
- We define some bases of decision-making as off-limits (e.g., race, gender, age)

The Morals of Prediction



Human Decision-makers

- We ask them to give an account of their decisions
- We develop constructs such as the “reasonable person” standard
- We define some bases of decision-making as off-limits (e.g., race, gender, age)



Algorithmic Decision-makers

- “Black box” algorithms are inscrutable, even to those who develop them
- So much data is available in some cases, predictions of protected characteristics are trivial

The Morals of Prediction

The Kiviat (2017) reading described two moral frameworks that competed in debates in the U.S. about the use of credit scores in insurance pricing.

The Morals of Prediction

The Kiviat (2017) reading described two moral frameworks that competed in debates in the U.S. about the use of credit scores in insurance pricing.

1. **Moral deservingness** – If the algorithm would hypothetically penalize people for only unambiguously/reasonably bad behavior, it's okay
2. **Actuarial Fairness** – If a feature helps predict the outcome, it's okay to use

The Morals of Prediction

The Kiviat (2017) reading described two moral frameworks that competed in debates in the U.S. about the use of credit scores in insurance pricing.

1. **Moral deservingness** – If the algorithm would hypothetically penalize people for only unambiguously/reasonably bad behavior, it's okay
2. **Actuarial Fairness** – If a feature helps predict the outcome, it's okay to use

How else might we judge algorithms?

Defining and Measuring Fairness

Three common definitions of “fairness” in the ML literature with respect to a characteristic:

1

Anti-classification: not using the characteristic to train the algorithm

2

3

Defining and Measuring Fairness

Three common definitions of “fairness” in the ML literature with respect to a characteristic:

1

Anti-classification: not using the characteristic to train the algorithm

2

Classification parity: measures of predictive accuracy (e.g., accuracy, mean-squared error, etc.) are equal for all values of the characteristic

3

Defining and Measuring Fairness

Three common definitions of “fairness” in the ML literature with respect to a characteristic:

1

Anti-classification: not using the characteristic to train the algorithm

2

Classification parity: measures of predictive accuracy (e.g., accuracy, mean-squared error, etc.) are equal for all values of the characteristic

3

Calibration: given identical values of the algorithm’s estimates, the characteristic no longer correlates with the outcome

Defining and Measuring Fairness

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan

The setting

1. Two groups: $C = 0$ and $C = 1$
2. Some actual (binary for simplicity) outcome: Y
3. C does not *directly* affect Y , but C might be correlated with some X that is (i.e., $E(Y|X, C = 0) = E(Y|X, C = 1)$)
4. A predicted outcome (predicted probability) from a predictive model on the basis of X : \hat{Y}

Defining and Measuring Fairness

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan

The setting

1. Two groups: $C = 0$ and $C = 1$
2. Some actual (binary for simplicity) outcome: Y
3. C does not *directly* affect Y , but C might be correlated with some X that is (i.e., $E(Y|X, C = 0) = E(Y|X, C = 1)$)
4. A predicted outcome (predicted probability) from a predictive model on the basis of X : \hat{Y}

The “guarantees”

1. Within-group calibration: The algorithm should be relatively accurate (i.e., for both groups, $E(Y|\hat{Y}) \approx \hat{Y}$)
2. Positive class balance: Among those who were actually positive cases, (i.e., $Y = 1$), there should be no group difference in predicted probabilities (i.e., $(\hat{Y} | C = 0, Y = 1) \approx (\hat{Y} | C = 1, Y = 1)$)
3. Negative class balance: Among those who were actually negative cases, (i.e., $Y = 0$), there should be no group difference in predicted probabilities (i.e., $(\hat{Y} | C = 0, Y = 0) \approx (\hat{Y} | C = 1, Y = 0)$)
4. ~~Statistical parity: There should be no group differences in predicted probabilities (i.e., $(\hat{Y} | C = 0) \approx (\hat{Y} | C = 1)$)~~

Defining and Measuring Fairness

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan

The setting

1. Two groups: $C = 0$ and $C = 1$
2. Some actual (binary for simplicity) outcome: Y
3. C does not *directly* affect Y , but C might be correlated with some X that is (i.e., $E(Y|X, C = 0) = E(Y|X, C = 1)$)
4. A predicted outcome (predicted probability) from a predictive model on the basis of X : \hat{Y}

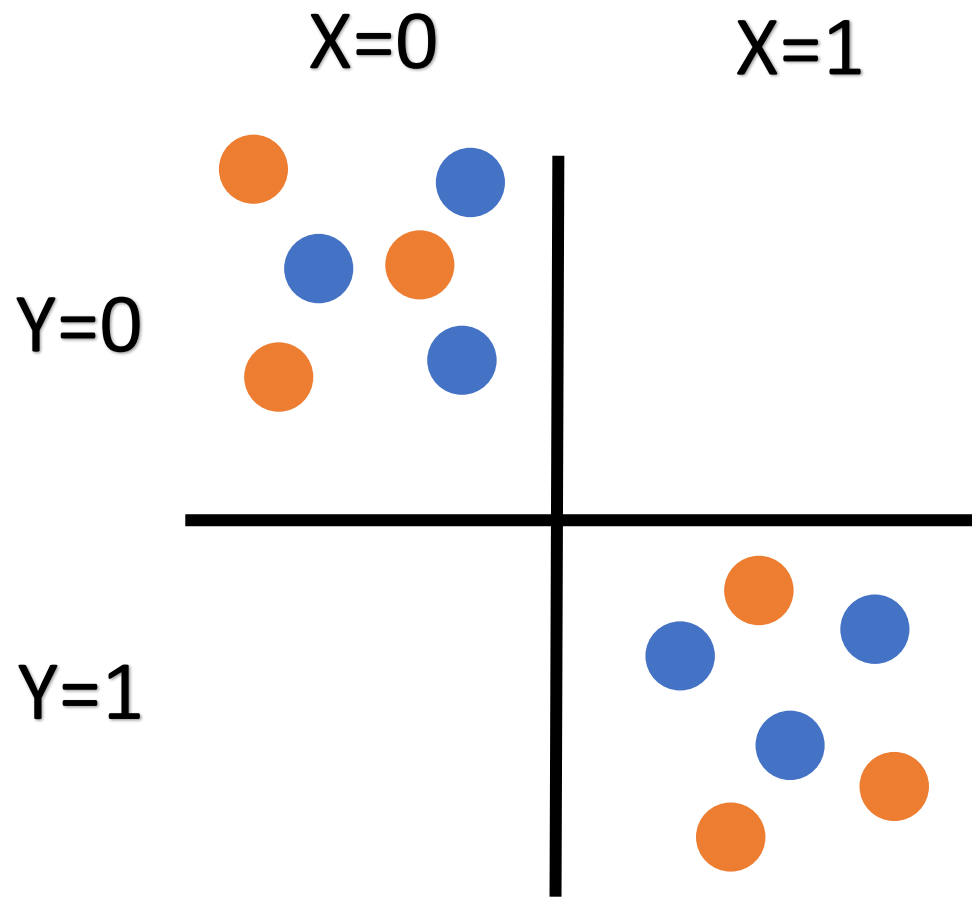
The “guarantees”

1. Within-group calibration: The algorithm should be relatively accurate (i.e., for both groups, $E(Y|\hat{Y}) \approx \hat{Y}$)
2. Positive class balance: Among those who were actually positive cases, (i.e., $Y = 1$), there should be no group difference in predicted probabilities (i.e., $(\hat{Y} | C = 0, Y = 1) \approx (\hat{Y} | C = 1, Y = 1)$)
3. Negative class balance: Among those who were actually negative cases, (i.e., $Y = 0$), there should be no group difference in predicted probabilities (i.e., $(\hat{Y} | C = 0, Y = 0) \approx (\hat{Y} | C = 1, Y = 0)$)
4. ~~Statistical parity: There should be no group differences in predicted probabilities (i.e., $(\hat{Y} | C = 0) \approx (\hat{Y} | C = 1)$)~~

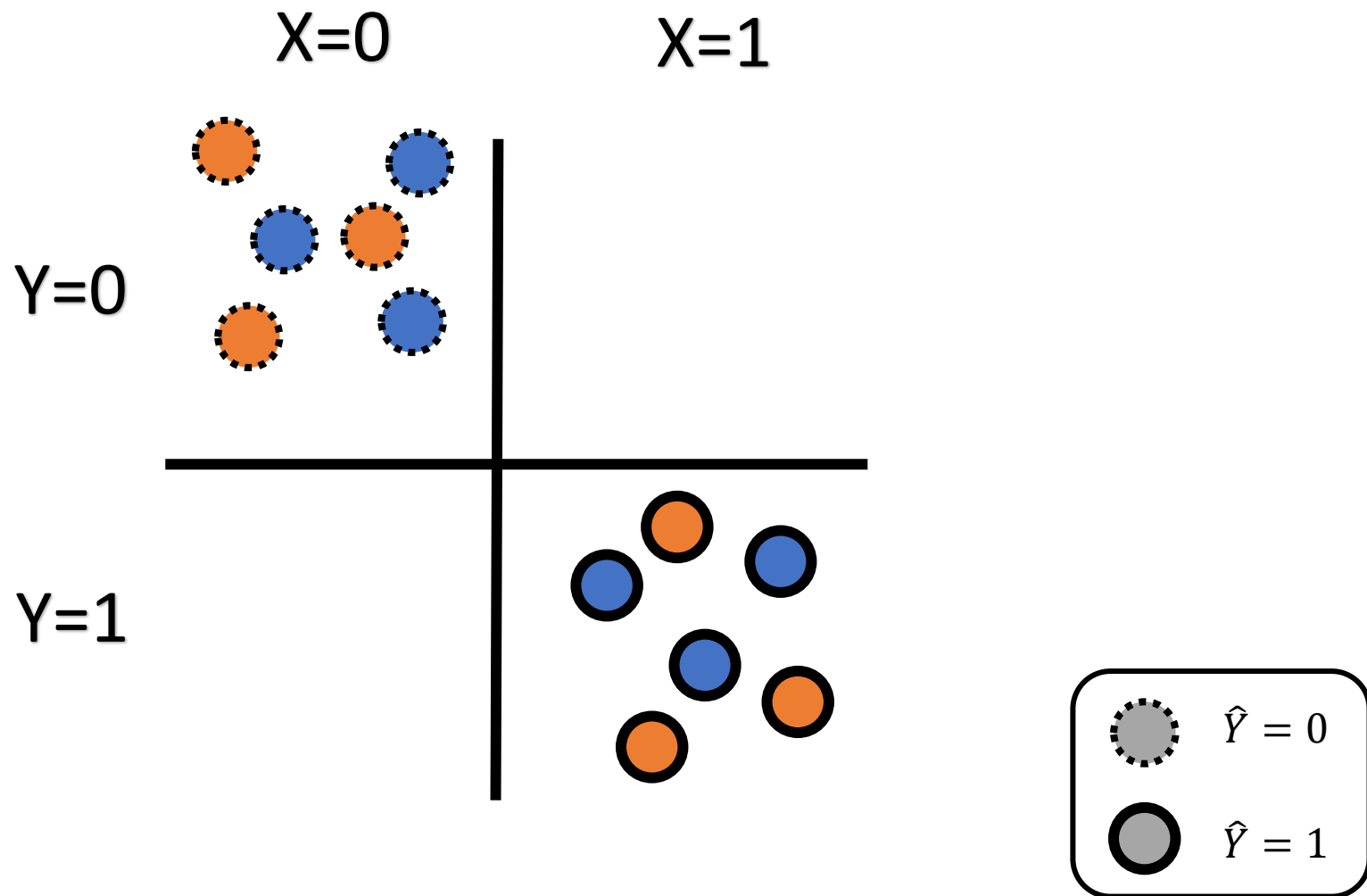
Guarantees (1) – (3) can only be achieved simultaneously when one of these is true:

1. Perfect Prediction: All positive cases get a predicted probability of one, all negative cases get a predicted probability of zero (i.e., $(\hat{Y} | Y = 0) = 0$; $(\hat{Y} | Y = 1) = 1$)
2. Equal Base Rates: Groups are the same on all observable characteristics associated with Y (i.e., $E(X|C = 0) = E(X|C = 1)$). If this were true, then the groups would also have the same average value of the outcome (i.e., $E(Y|C = 0) = E(Y|C = 1)$).

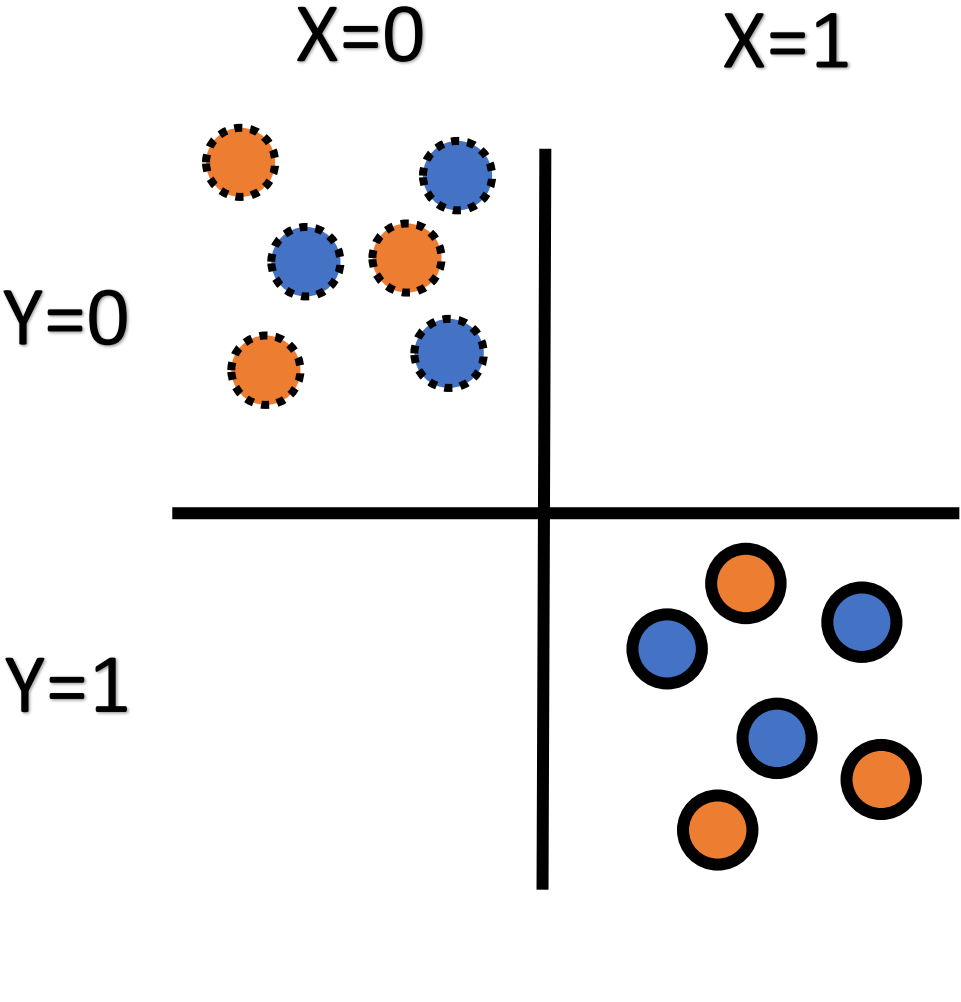
Defining and Measuring Fairness



Defining and Measuring Fairness

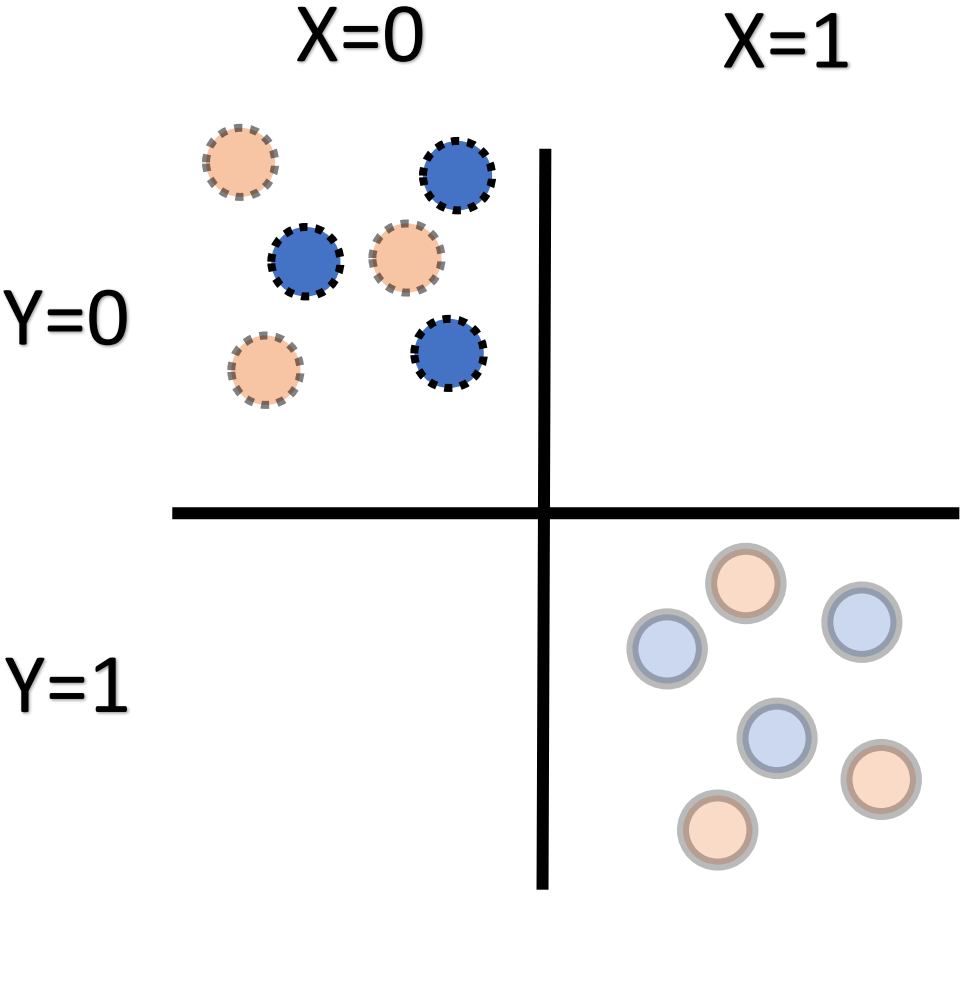


Defining and Measuring Fairness

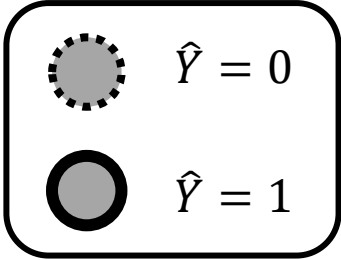


Group	$(\hat{Y} \mid Y = 0)$	$(\hat{Y} \mid Y = 1)$
Blue		
Orange		

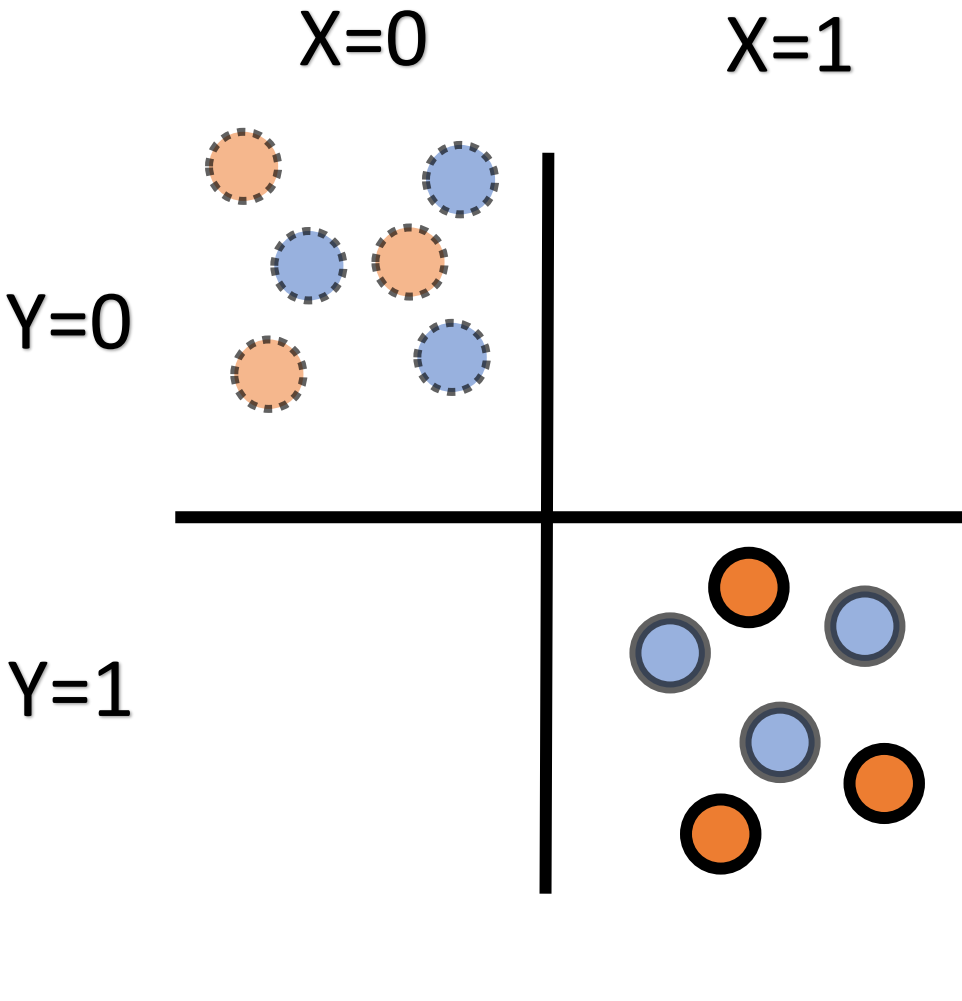
Defining and Measuring Fairness



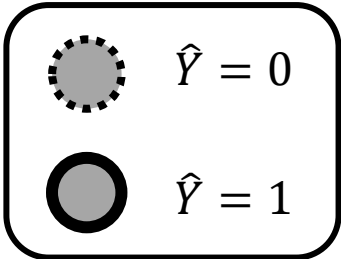
Group	$(\hat{Y} Y = 0)$	$(\hat{Y} Y = 1)$
Blue	?	
Orange		



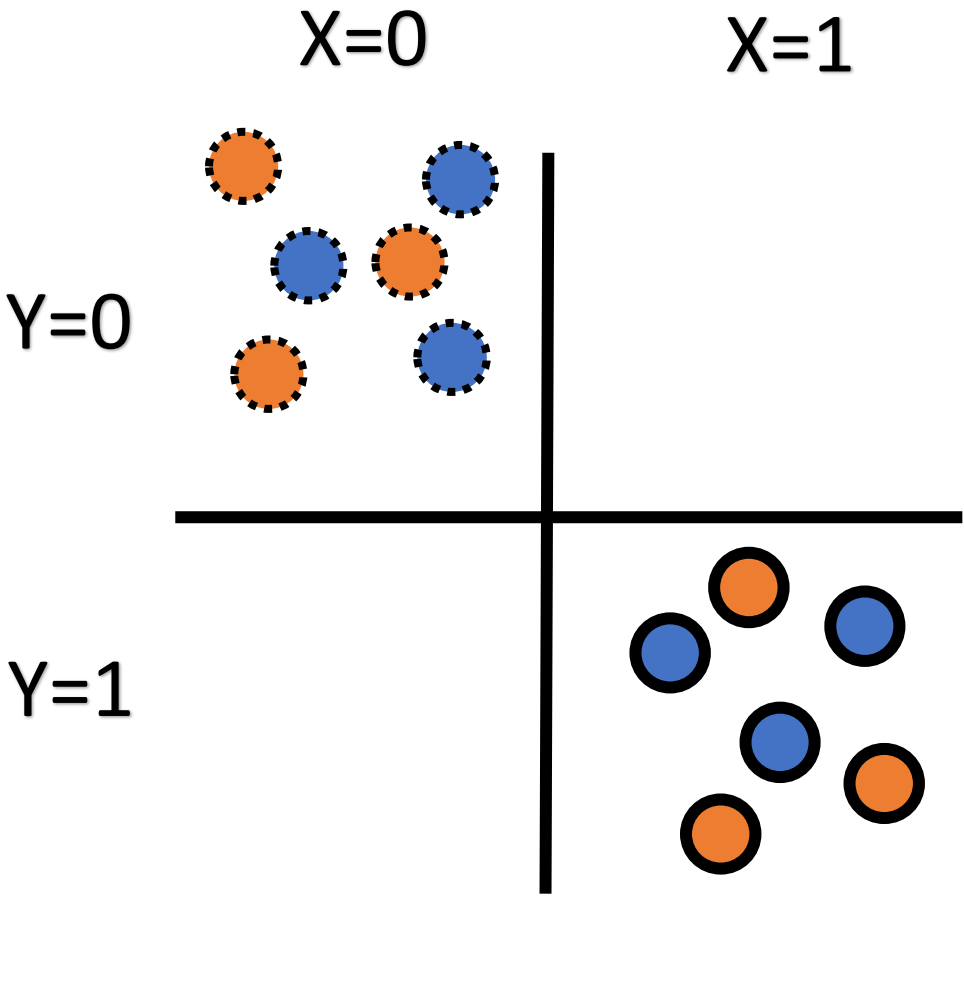
Defining and Measuring Fairness



Group	$(\hat{Y} Y = 0)$	$(\hat{Y} Y = 1)$
Blue	0	
Orange		?

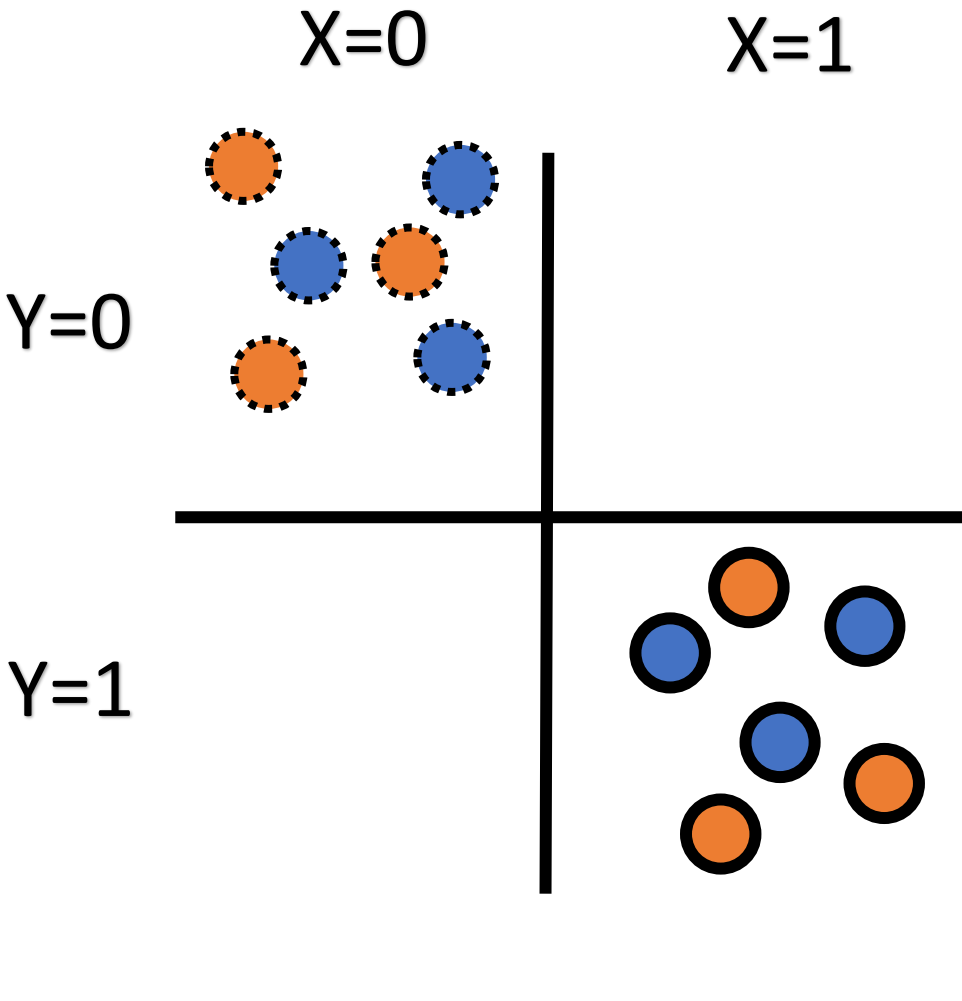


Defining and Measuring Fairness

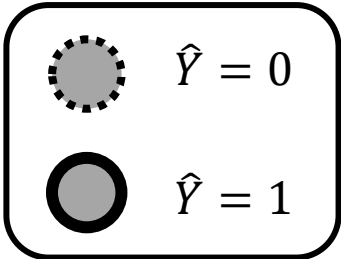


Group	$(\hat{Y} Y = 0)$	$(\hat{Y} Y = 1)$
Blue	0	?
Orange	?	1

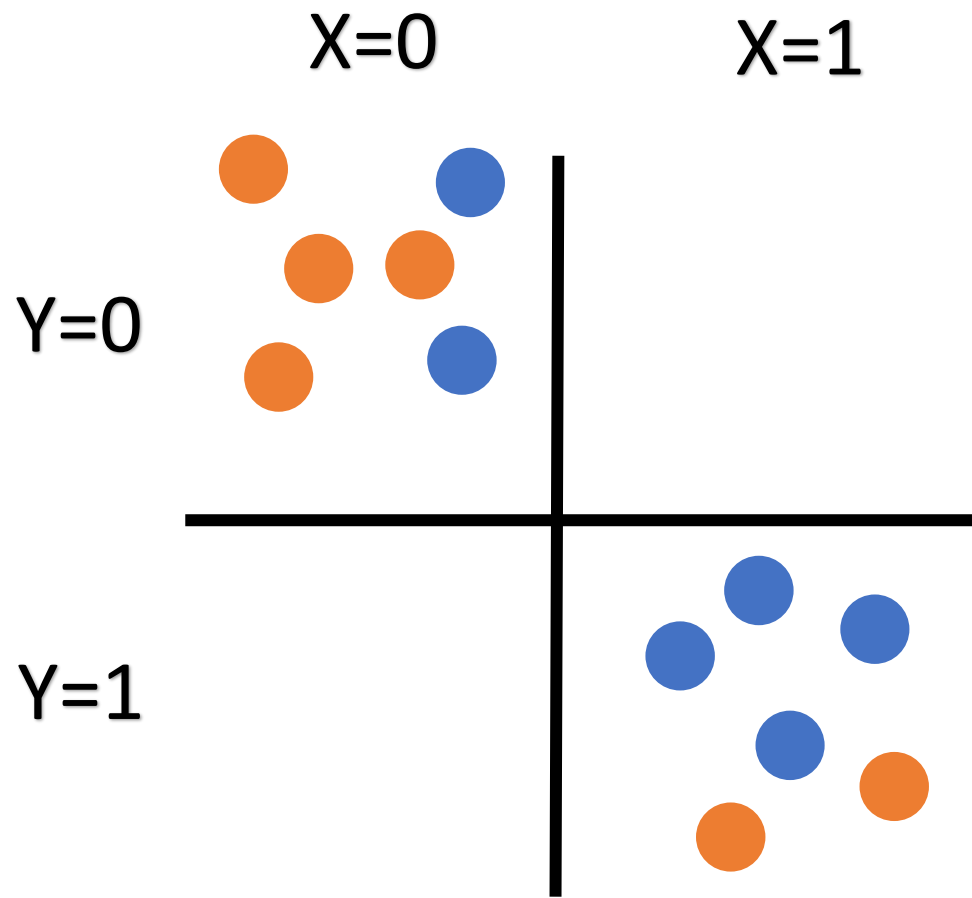
Defining and Measuring Fairness



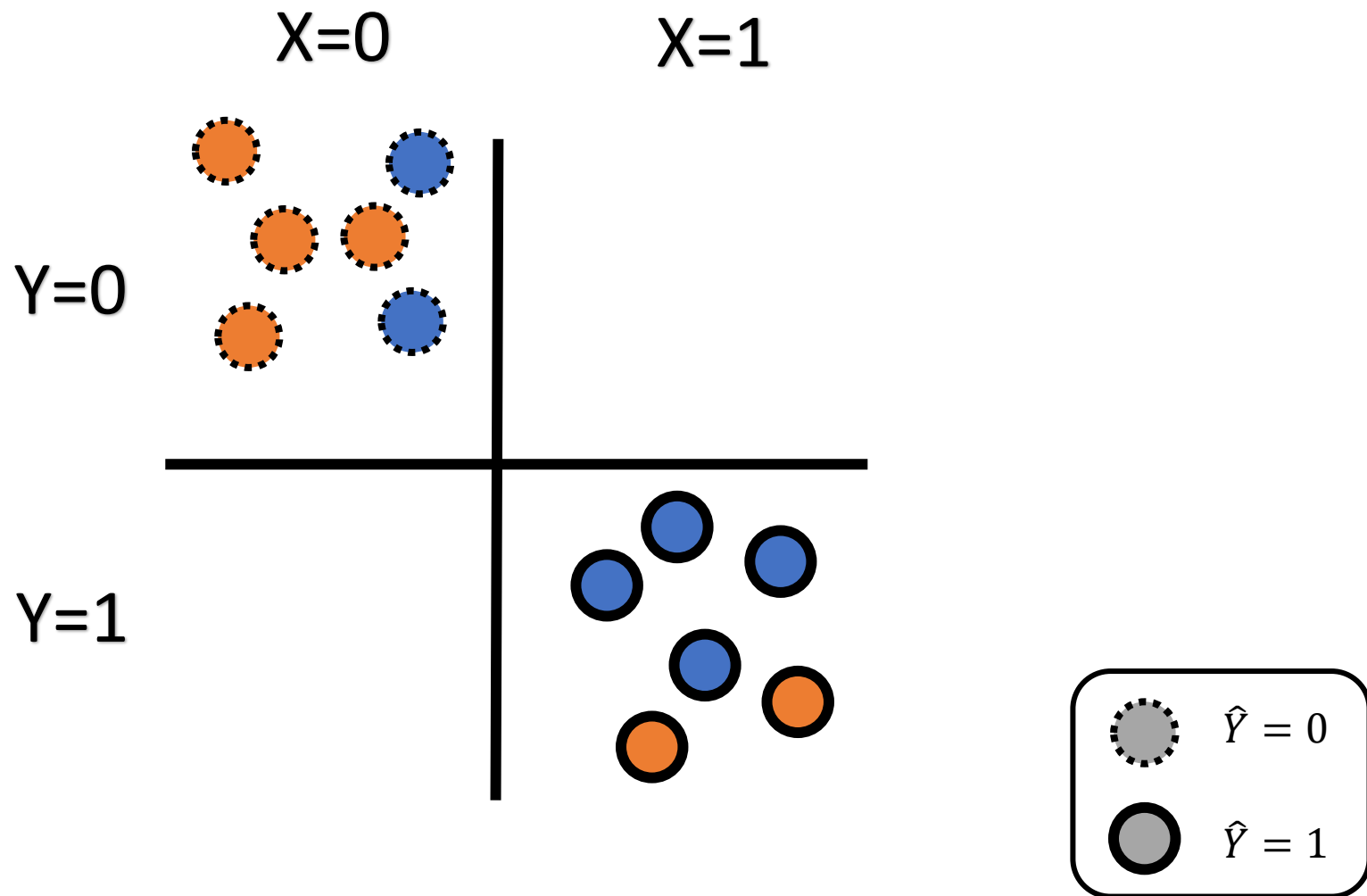
Group	$(\hat{Y} \mid Y = 0)$	$(\hat{Y} \mid Y = 1)$
Blue	0	1
Orange	0	1



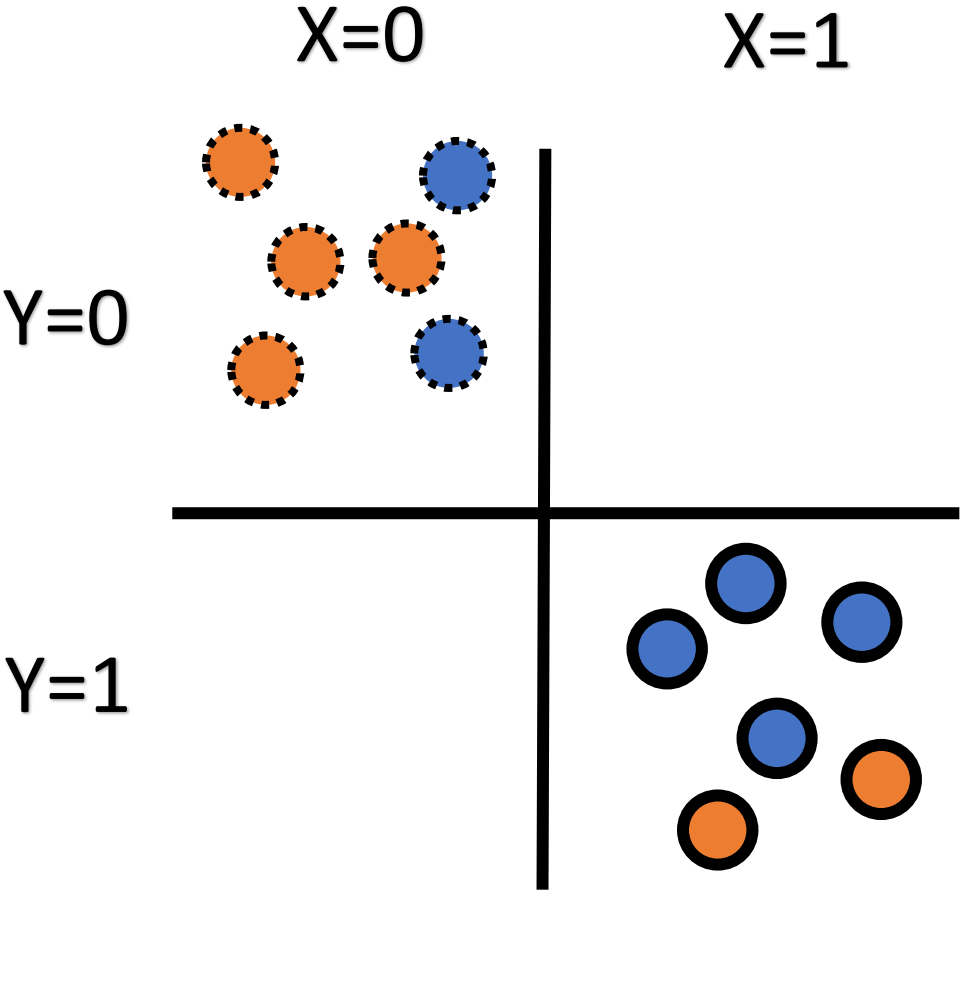
Defining and Measuring Fairness



Defining and Measuring Fairness

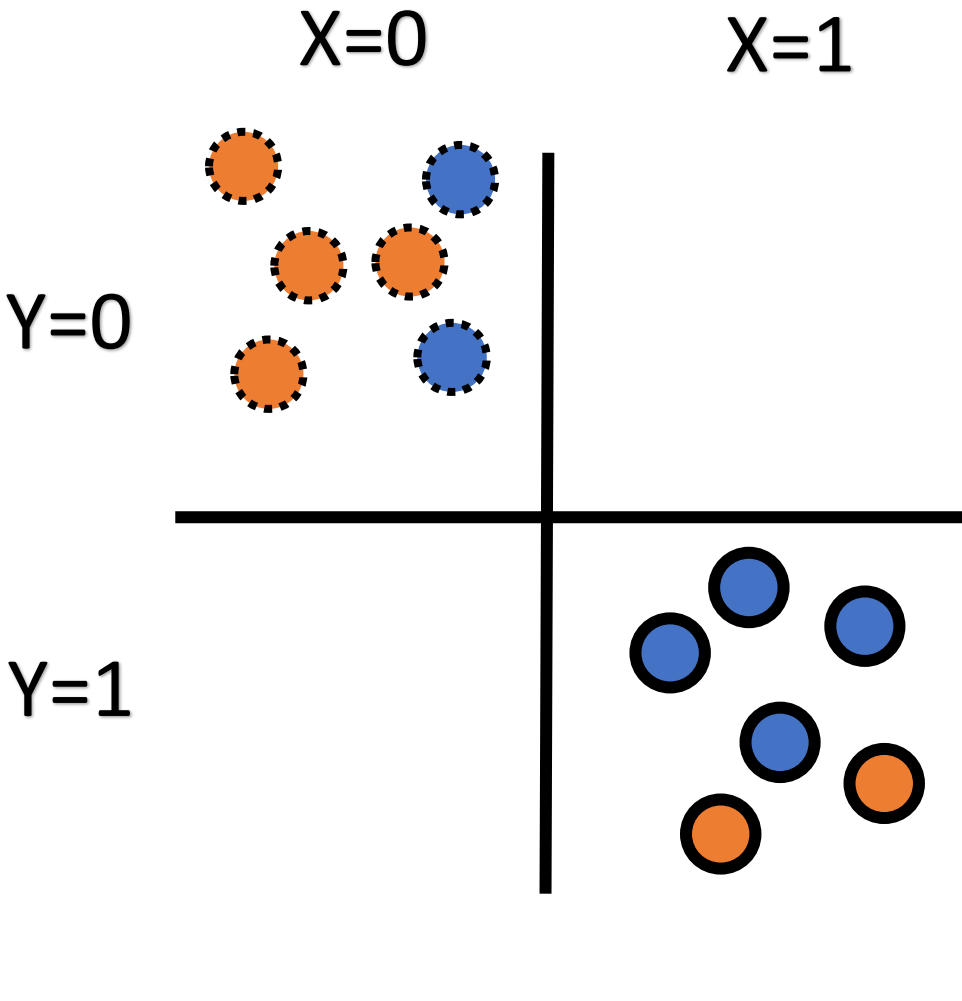


Defining and Measuring Fairness

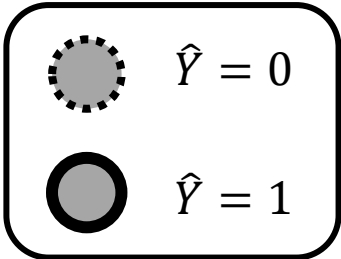


Group	$(\hat{Y} \mid Y = 0)$	$(\hat{Y} \mid Y = 1)$
Blue	?	?
Orange	?	?

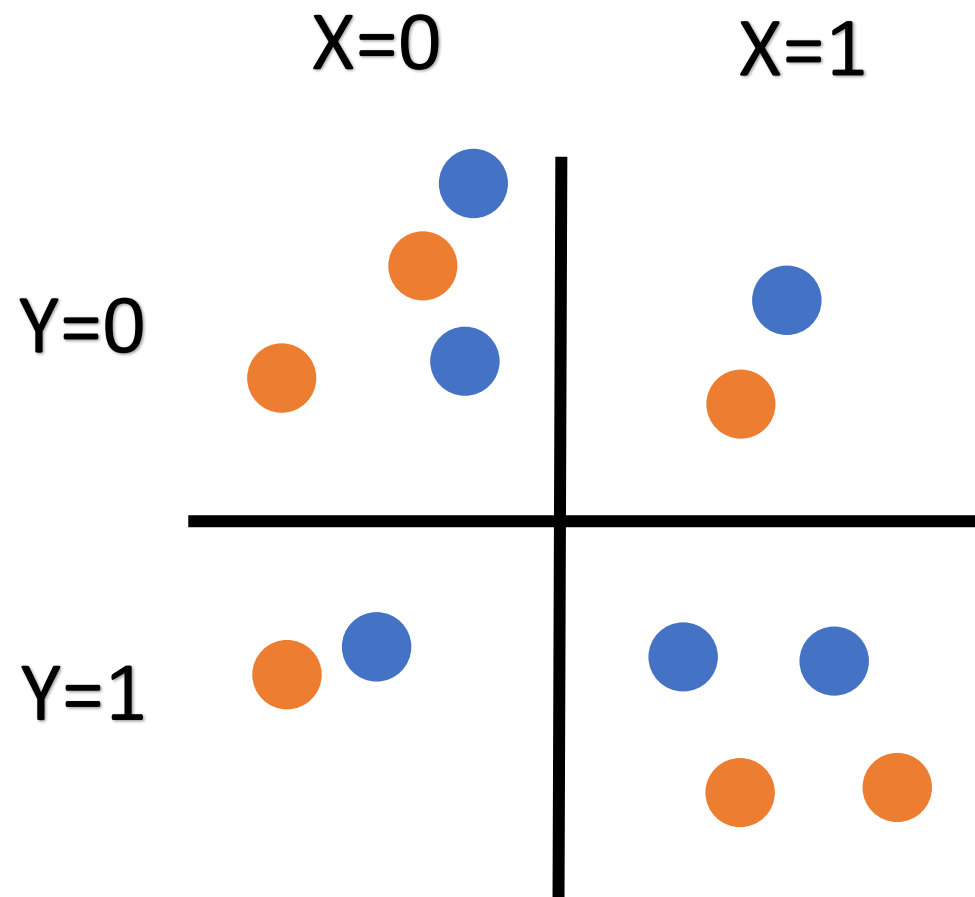
Defining and Measuring Fairness



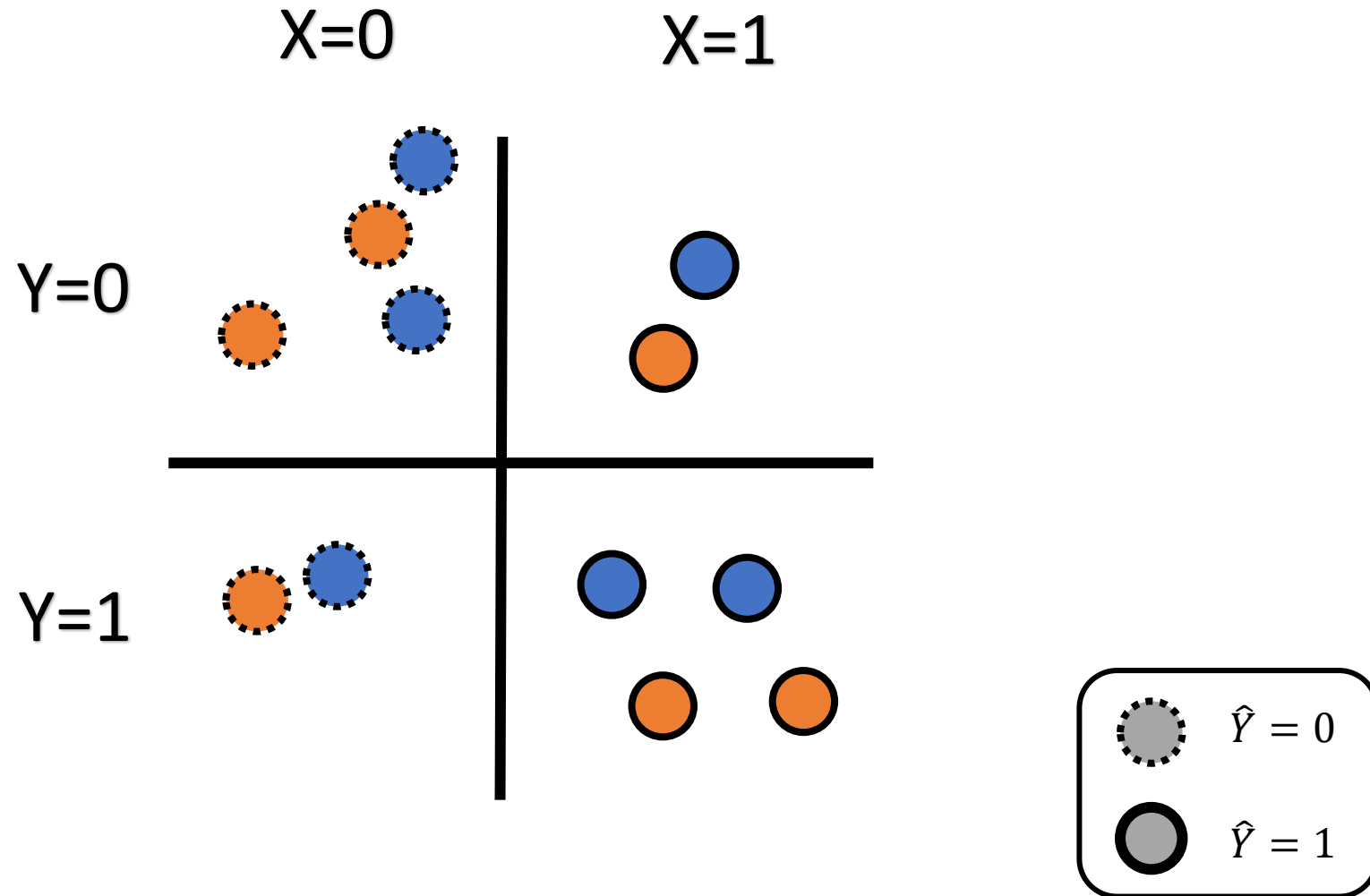
Group	$(\hat{Y} \mid Y = 0)$	$(\hat{Y} \mid Y = 1)$
Blue	0	1
Orange	0	1



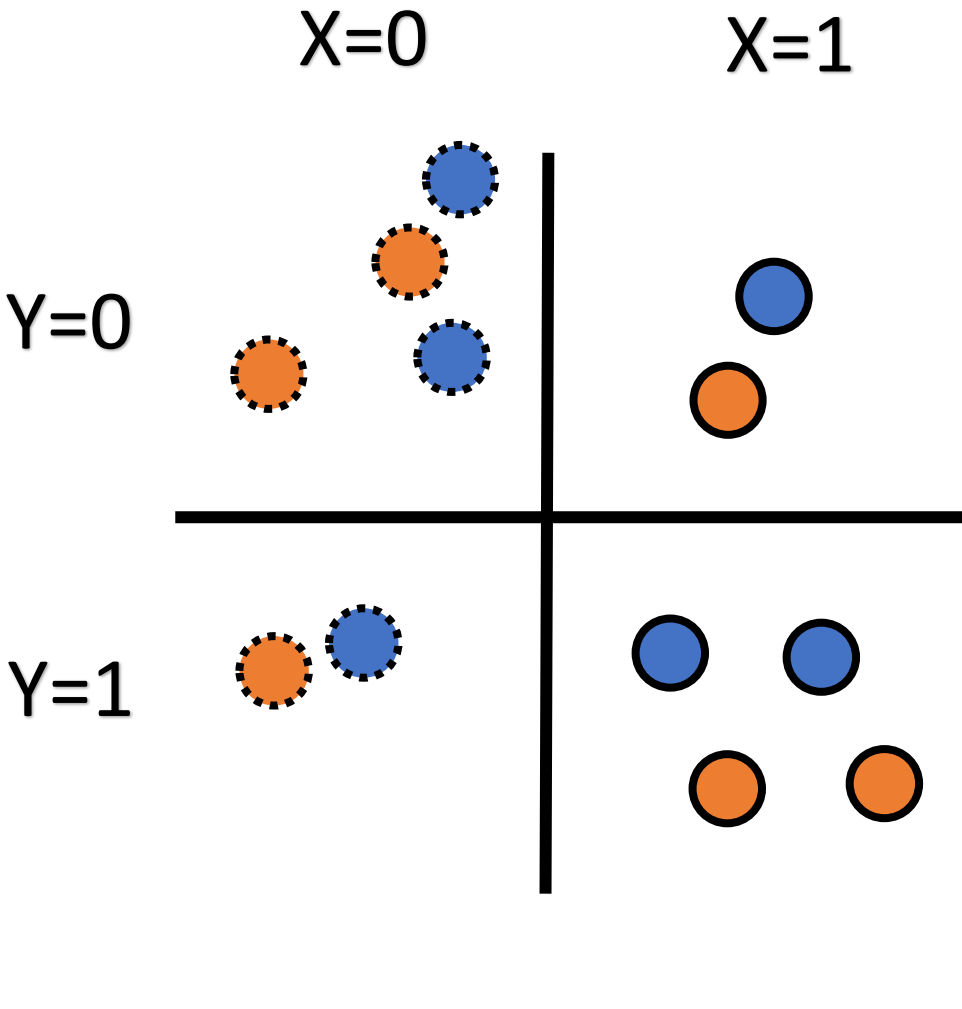
Defining and Measuring Fairness




Defining and Measuring Fairness




Defining and Measuring Fairness



Group	$(\hat{Y} Y = 0)$	$(\hat{Y} Y = 1)$
Blue	?	?
Orange	?	?

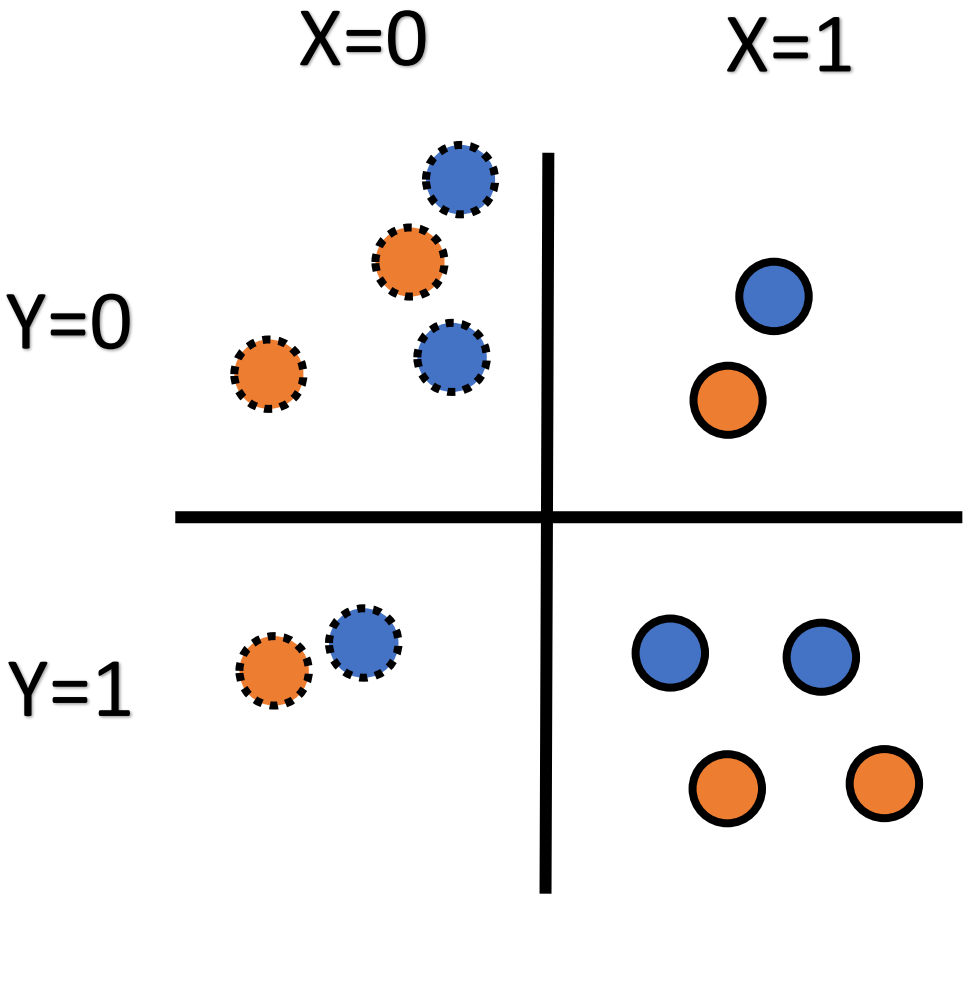


$\hat{Y} = 0$



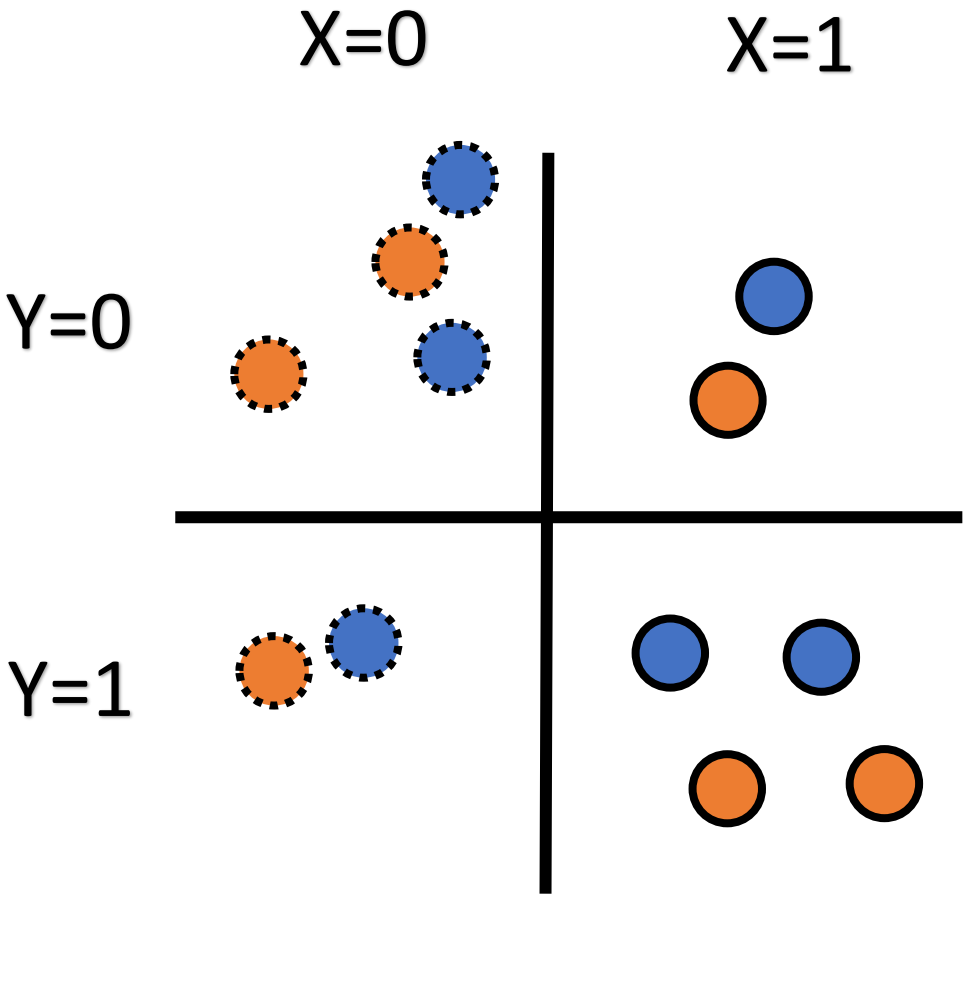
$\hat{Y} = 1$

Defining and Measuring Fairness

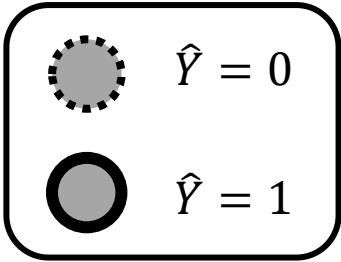


Group	$(\hat{Y} Y = 0)$	$(\hat{Y} Y = 1)$
Blue	1/3	2/3
Orange	1/3	2/3

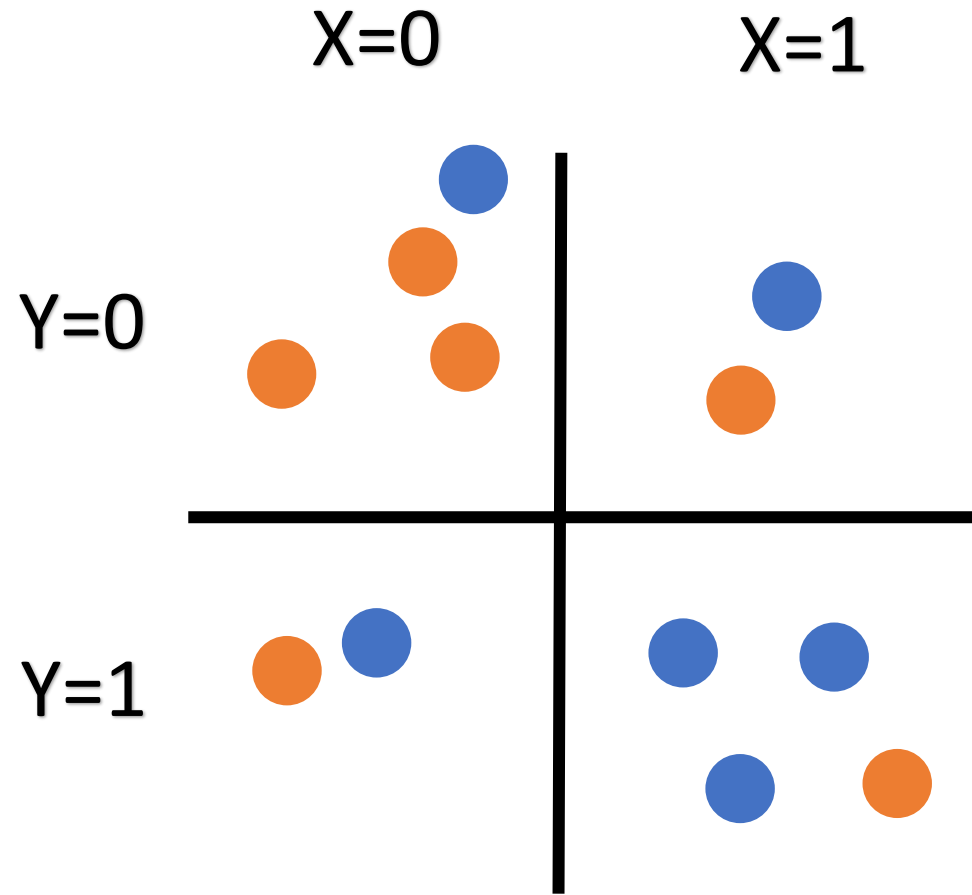
Defining and Measuring Fairness



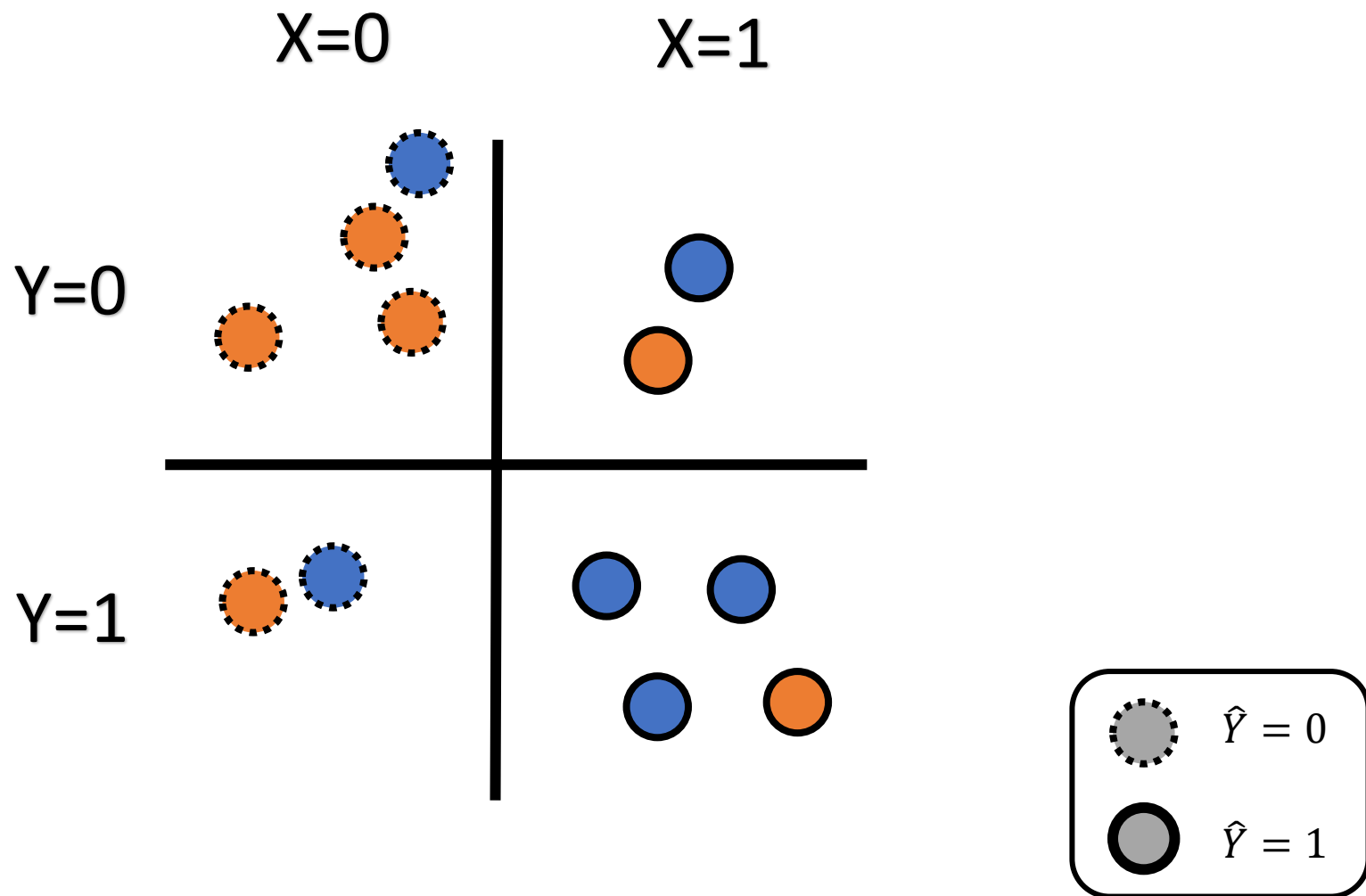
Group	$(\hat{Y} Y = 0)$	$(\hat{Y} Y = 1)$
Blue	1/3	2/3
Orange	1/3	2/3



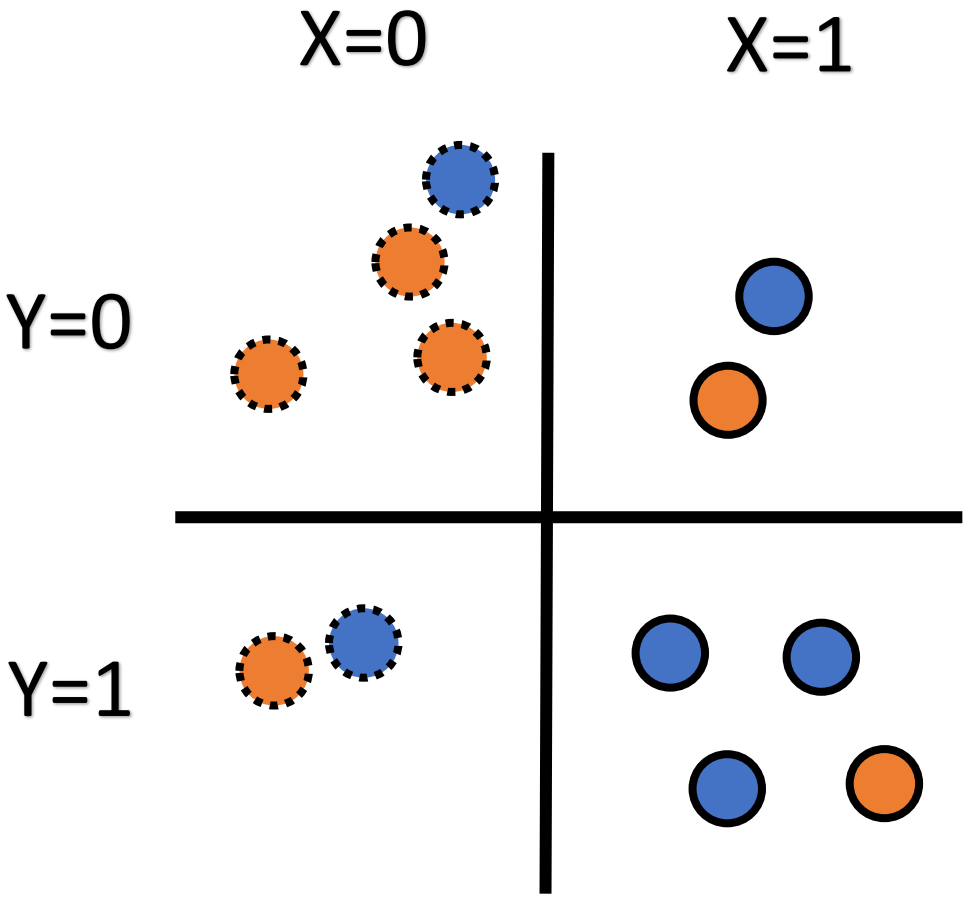
Defining and Measuring Fairness



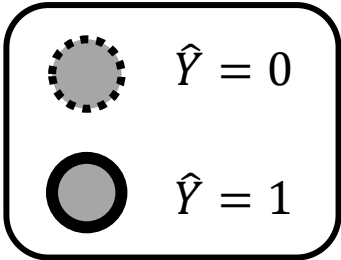
Defining and Measuring Fairness



Defining and Measuring Fairness



Group	$(\hat{Y} \mid Y = 0)$	$(\hat{Y} \mid Y = 1)$
Blue	1/2	3/4
Orange	1/4	1/2



Sources of Bias

Insider Bias

- A small group of people
- Those with intimate access to the algorithm's learning process

Outsider Bias

- Society structures the relationships between variables
- Some variables are difficult to measure

Sources of Bias

Insider Bias

- A small group of people
- Those with intimate access to the algorithm's learning process

Outsider Bias

- Society structures the relationships between variables
- Some variables are difficult to measure

Insider Bias

Who decided what to measure and what the outcome would be?

- Search for/inclusion of features
- Selecting an outcome

Insider Bias

The background of the slide features a grayscale image of a person's profile, wearing glasses and working on a laptop. A large, solid blue arrow points from the left side of the slide towards the right, passing behind the text boxes.

Who decided what to measure and what the outcome would be?

- Search for/inclusion of features
- Selecting an outcome

How did you measure your outcome and features?

- Human raters have biases
- Meaning of behavior varies

Insider Bias

Who decided what to measure and what the outcome would be?

- Search for/inclusion of features
- Selecting an outcome

What is represented in your dataset?

- Frequent class drives model predictions
- Distinctiveness can become a feature

How did you measure your outcome and features?

- Human raters have biases
- Meaning of behavior varies



Insider Bias

Who decided what to measure and what the outcome would be?

- Search for/inclusion of features
- Selecting an outcome

What is represented in your dataset?

- Frequent class drives model predictions
- Distinctiveness can become a feature

How did you measure your outcome and features?

- Human raters have biases
- Meaning of behavior varies

Which intuitions drove the training process?

- “Sanity checks” and heuristics galore
- When is “good enough”?

Sources of Bias

Insider Bias

- A small group of people
- Those with intimate access to the algorithm's learning process

Outsider Bias

- **Society structures the relationships between variables**
- **Some variables are difficult to measure**

In-class activity!

- Get into groups of two or three
- **You must not know anyone in your group from somewhere other than class!**
- Find the in-class activity the following link:
<https://tinyurl.com/InClassML>
- We'll go through the first “scenario” together

Things to do for each scenario...

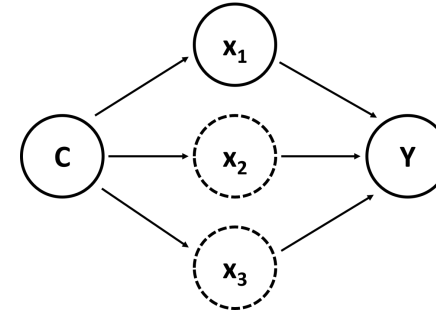
Before running the code...

- Make sure you understand the diagram
 - Does the group variable have a causal pathway to the outcome?
 - Is the algorithm using the group variable to make predictions?
- Imagine a set of real-world variables that might make some sense for this diagram (but keep S4 and S5 the same)
 - C might be age, gender, race/ethnicity, political party identity, native/immigrant status, etc.
 - Y might be something positive or negative (it's a continuous outcome in the simulation, but if it's helpful you can also think of it as probabilities of a binary outcome; all the mechanics are the same)
 - Causal pathways can be straight-forward (e.g., people who are older than 60 [C] start to lose their ability to learn new things because of decreased neuroplasticity) or complex (e.g., women on average have less STEM knowledge because of discrimination within STEM fields and gender norms)

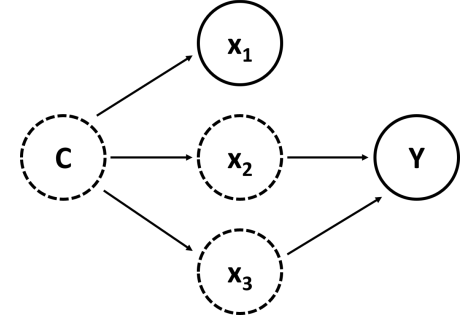
After running the code...

- How accurately did the algorithm predict the outcome?
- Is there a group difference in predicted values?
- What was the algorithm's weight on C (if it used it)?
- How does the correlation of the group variable with the algorithm's predictions compare to its correlation with the actual outcome?
- Given how you defined your variables and pathways before you ran the code, does this outcome seem fair to your group (it's okay to disagree)?

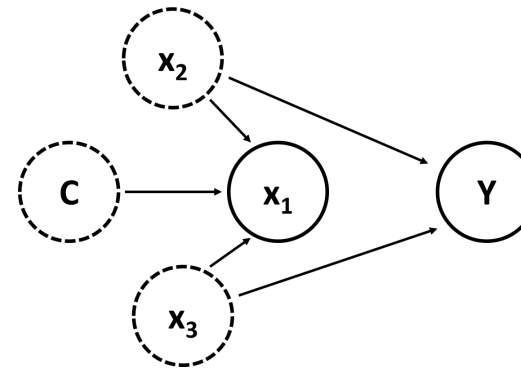
Scenario 2



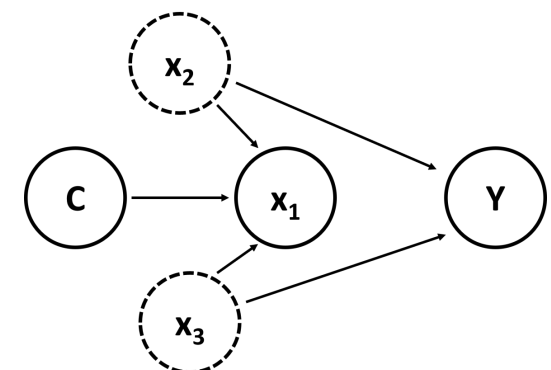
Scenario 3



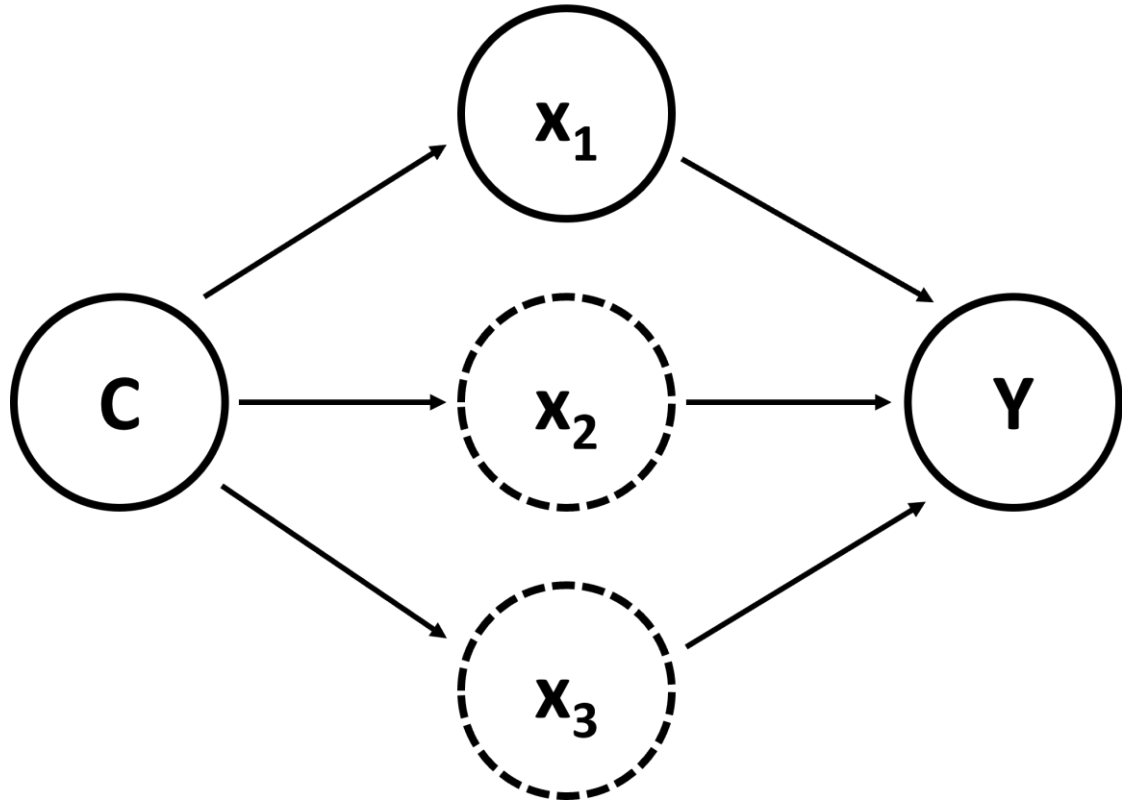
Scenario 4



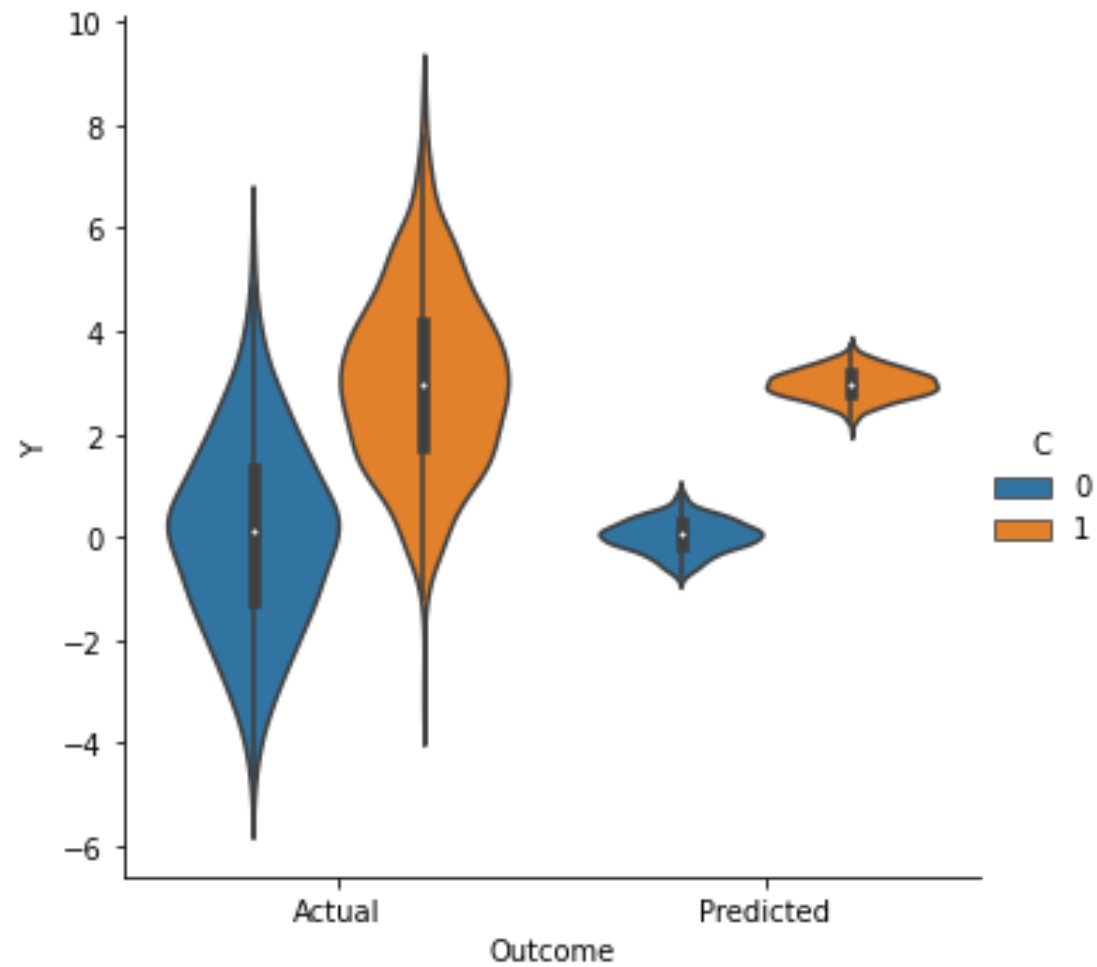
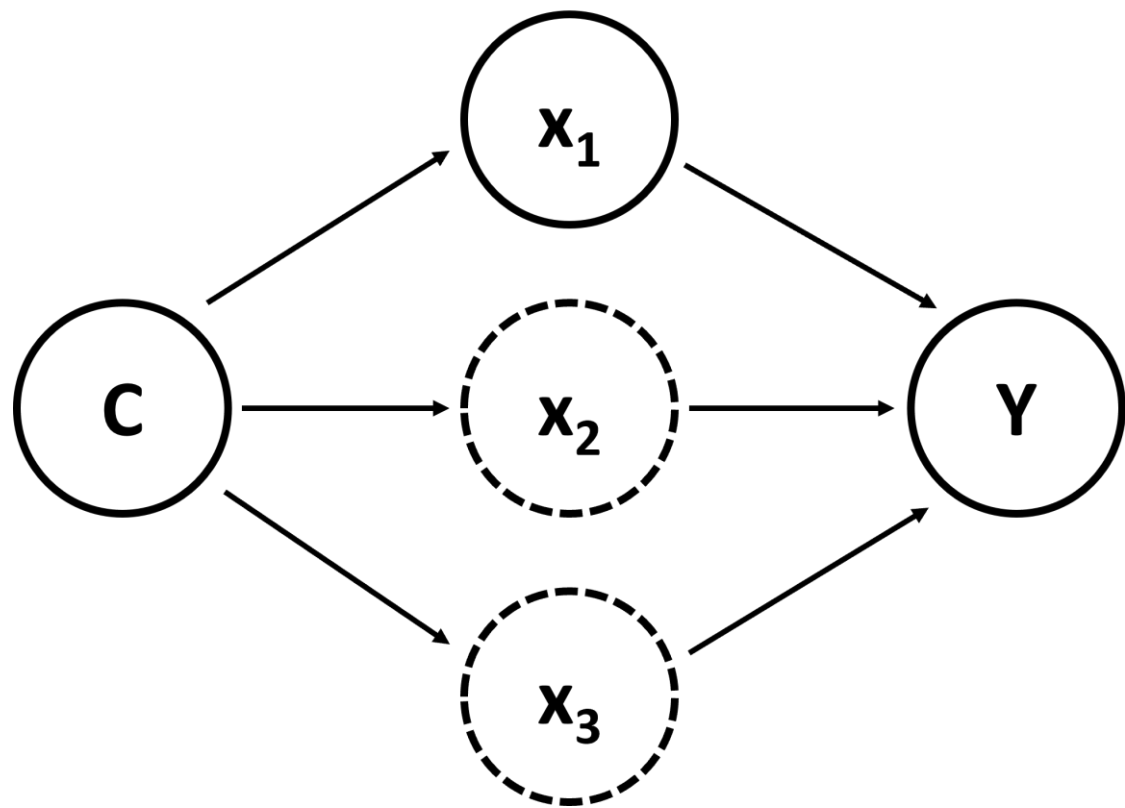
Scenario 5



Scenario 2

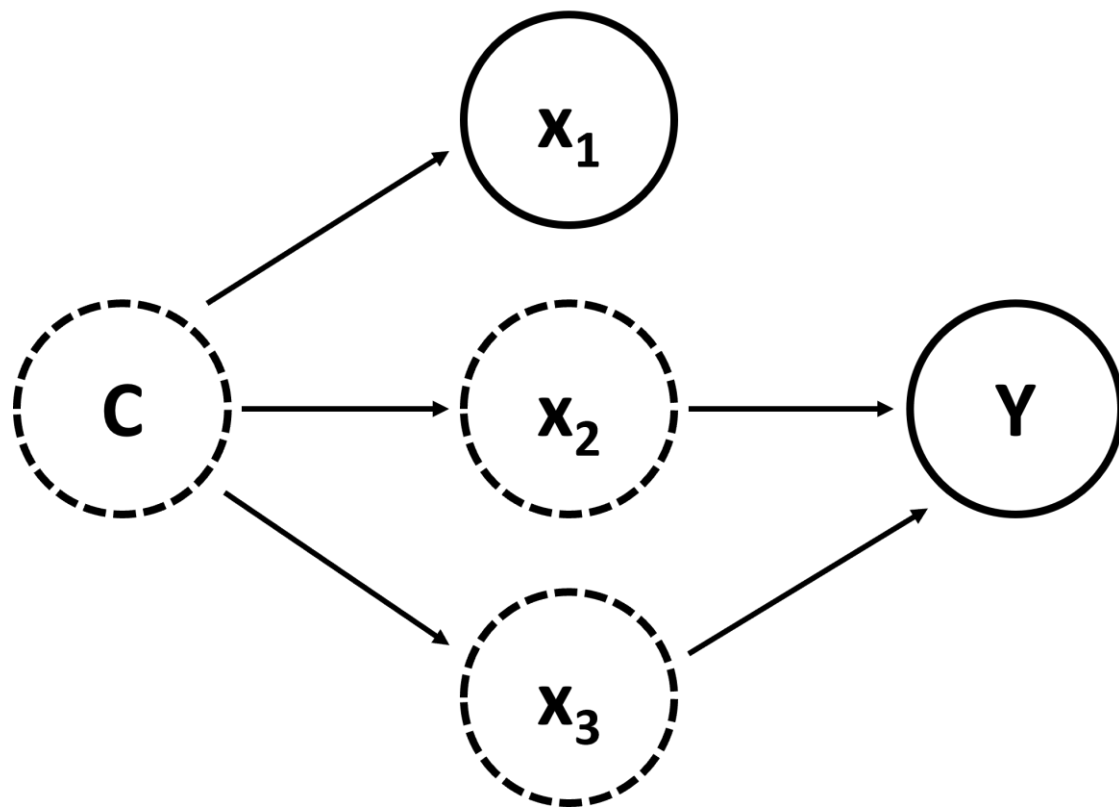


Scenario 2

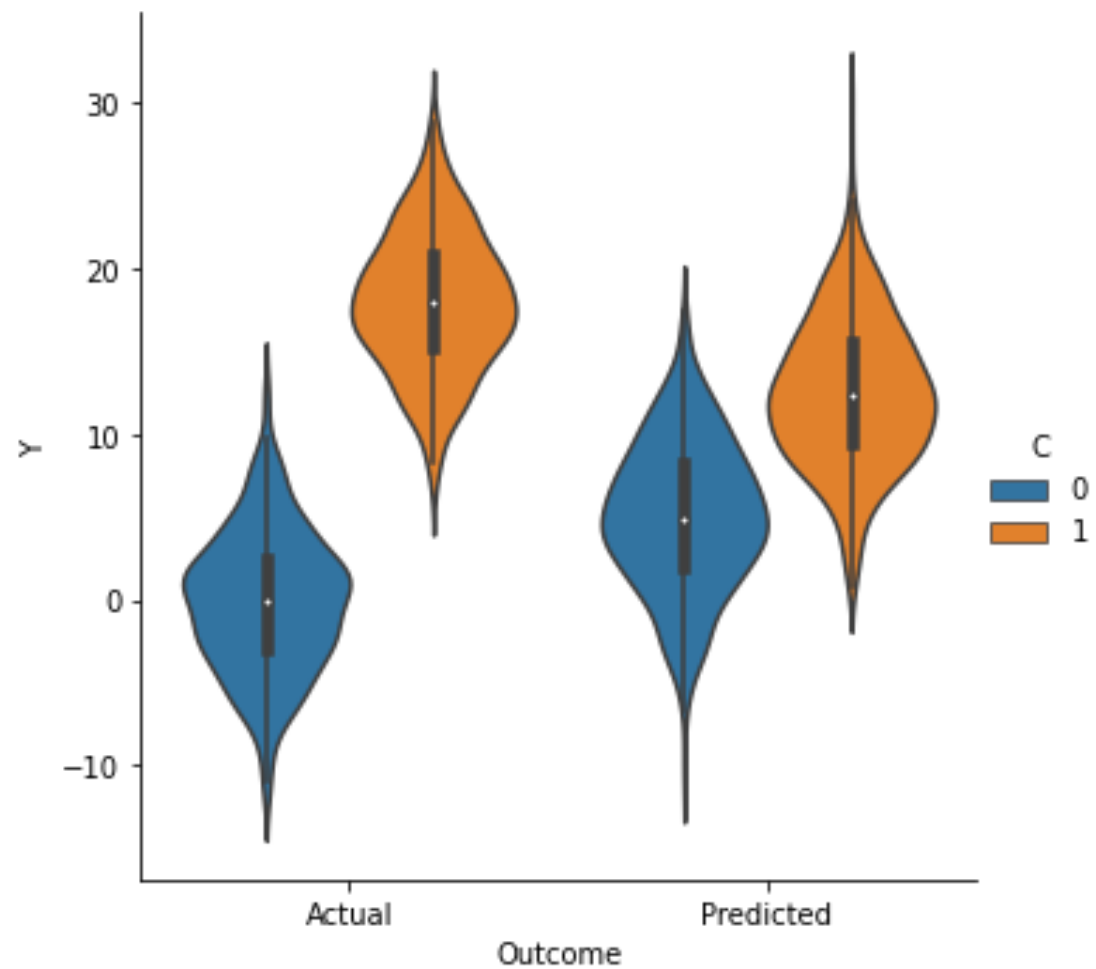
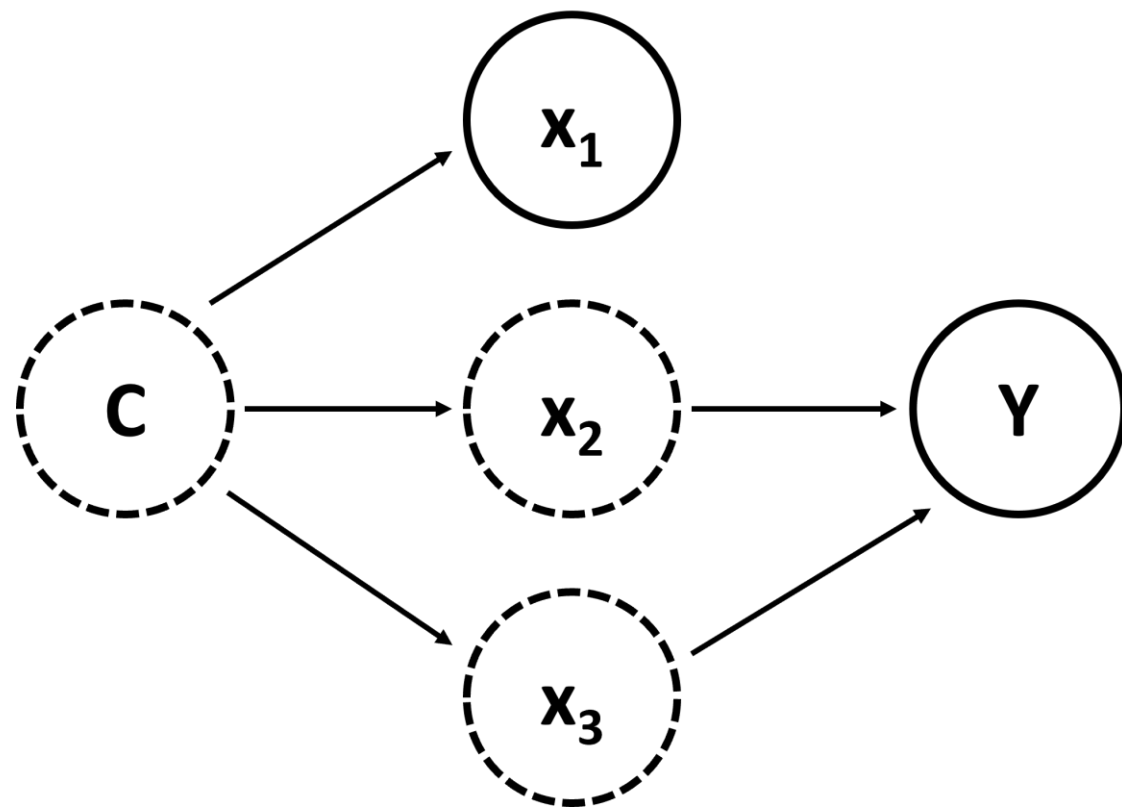


Weight on C : 2.9

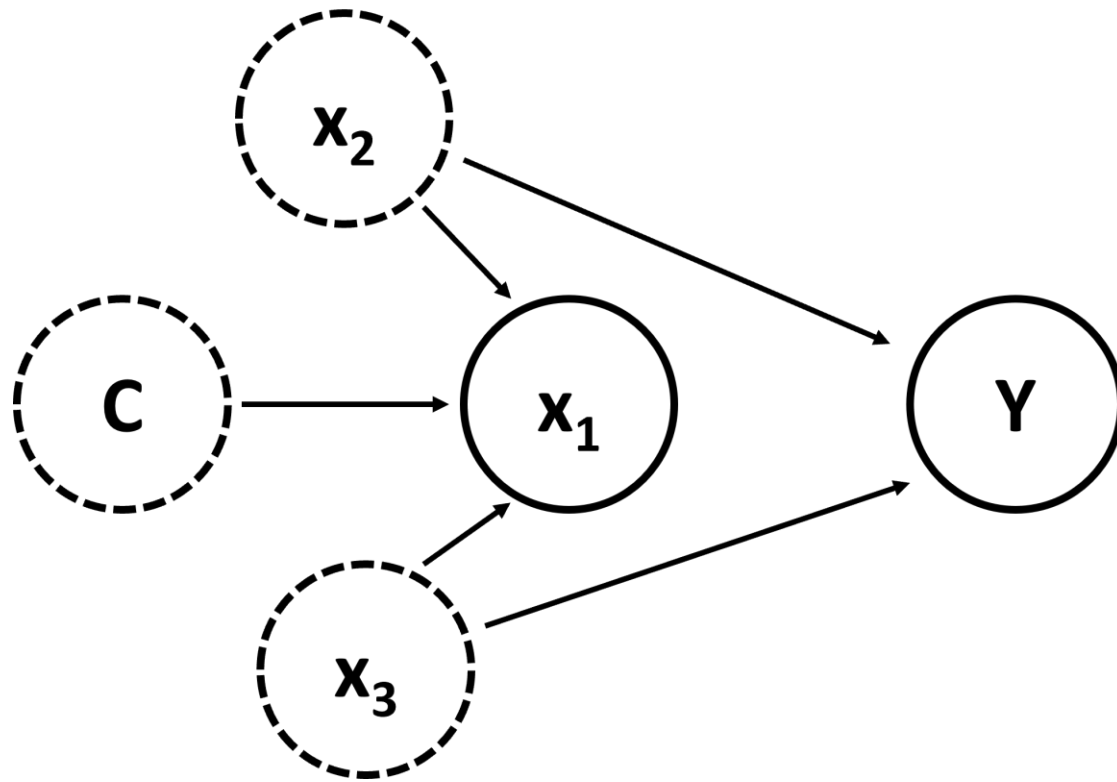
Scenario 3



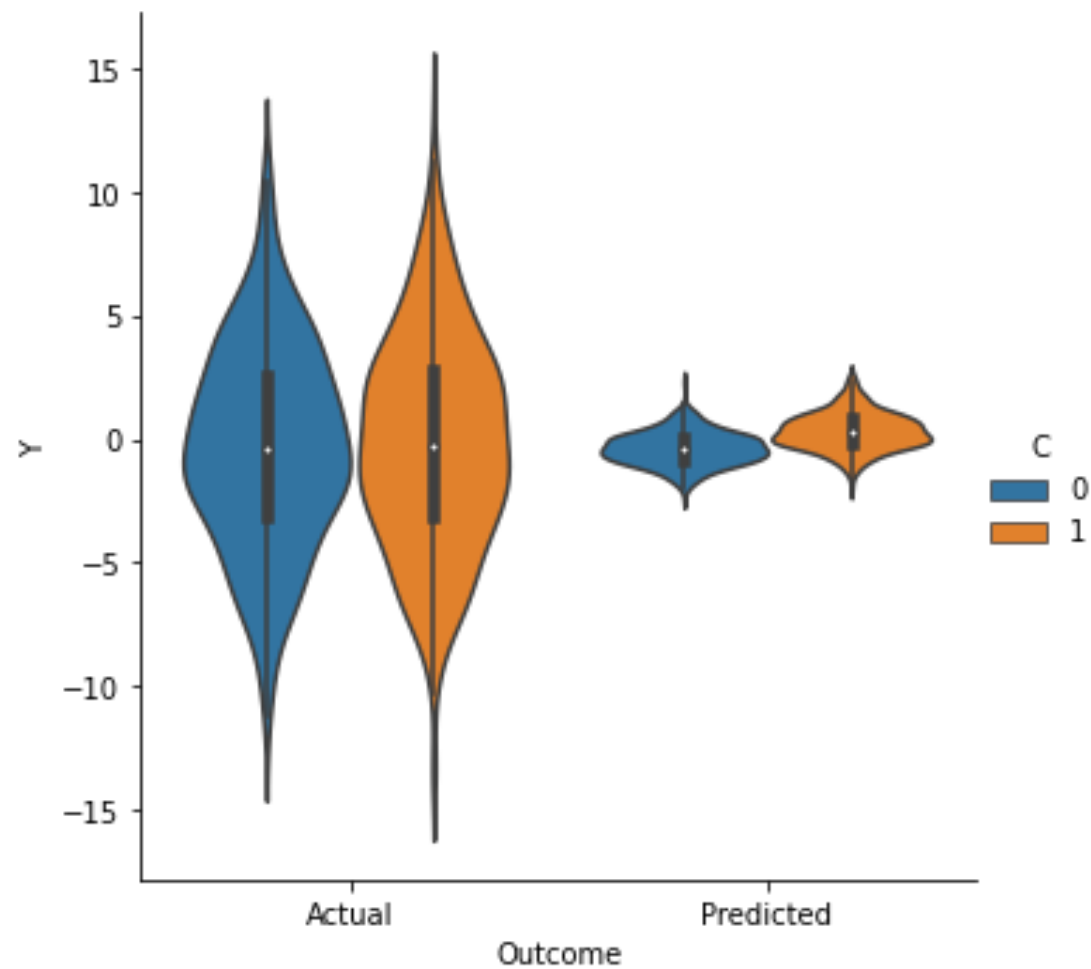
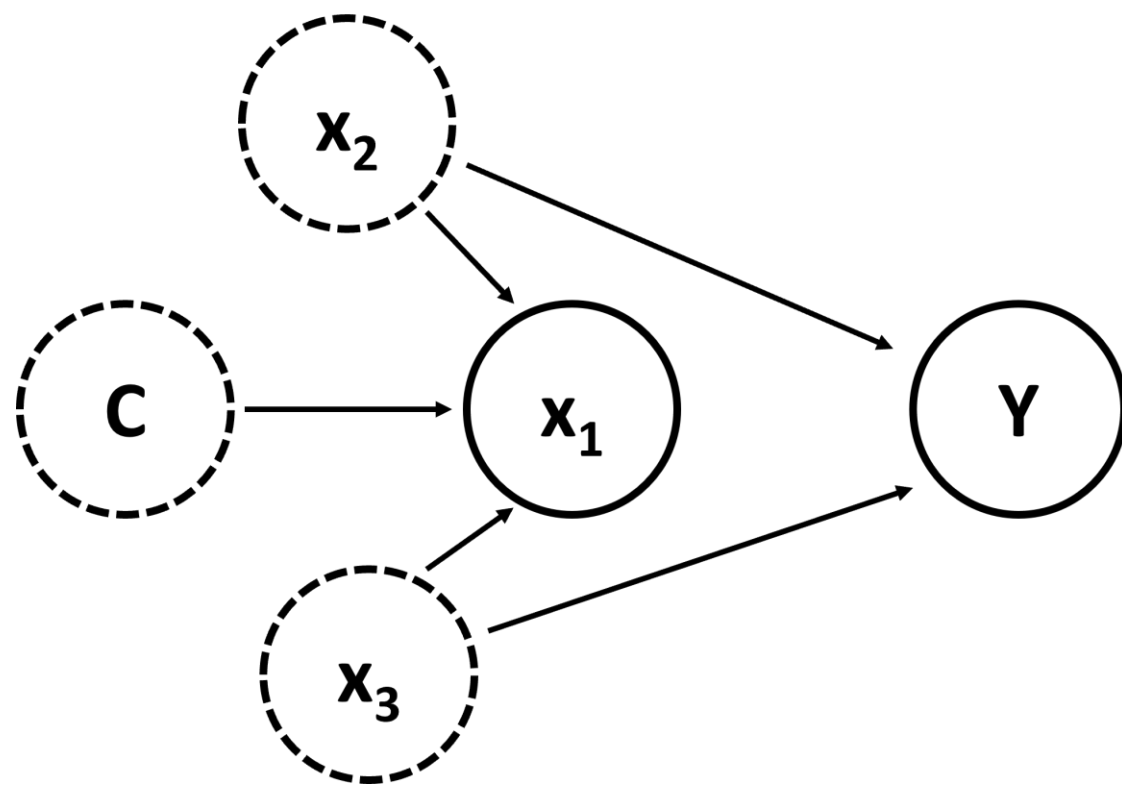
Scenario 3



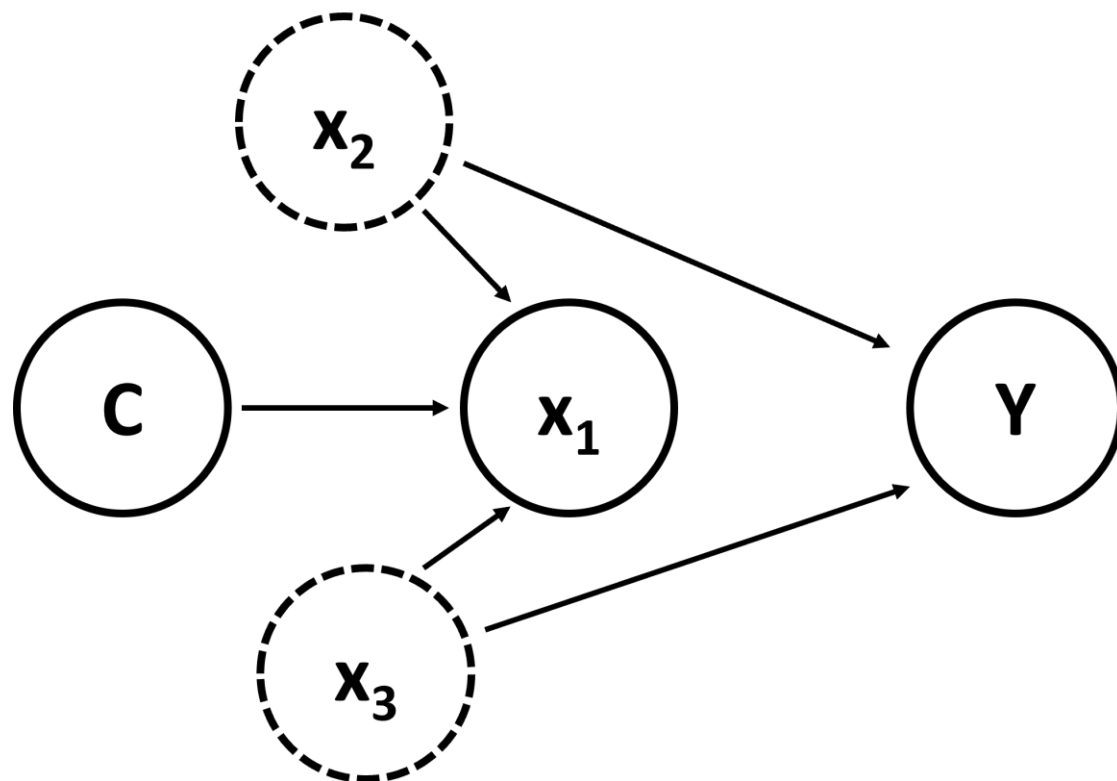
Scenario 4



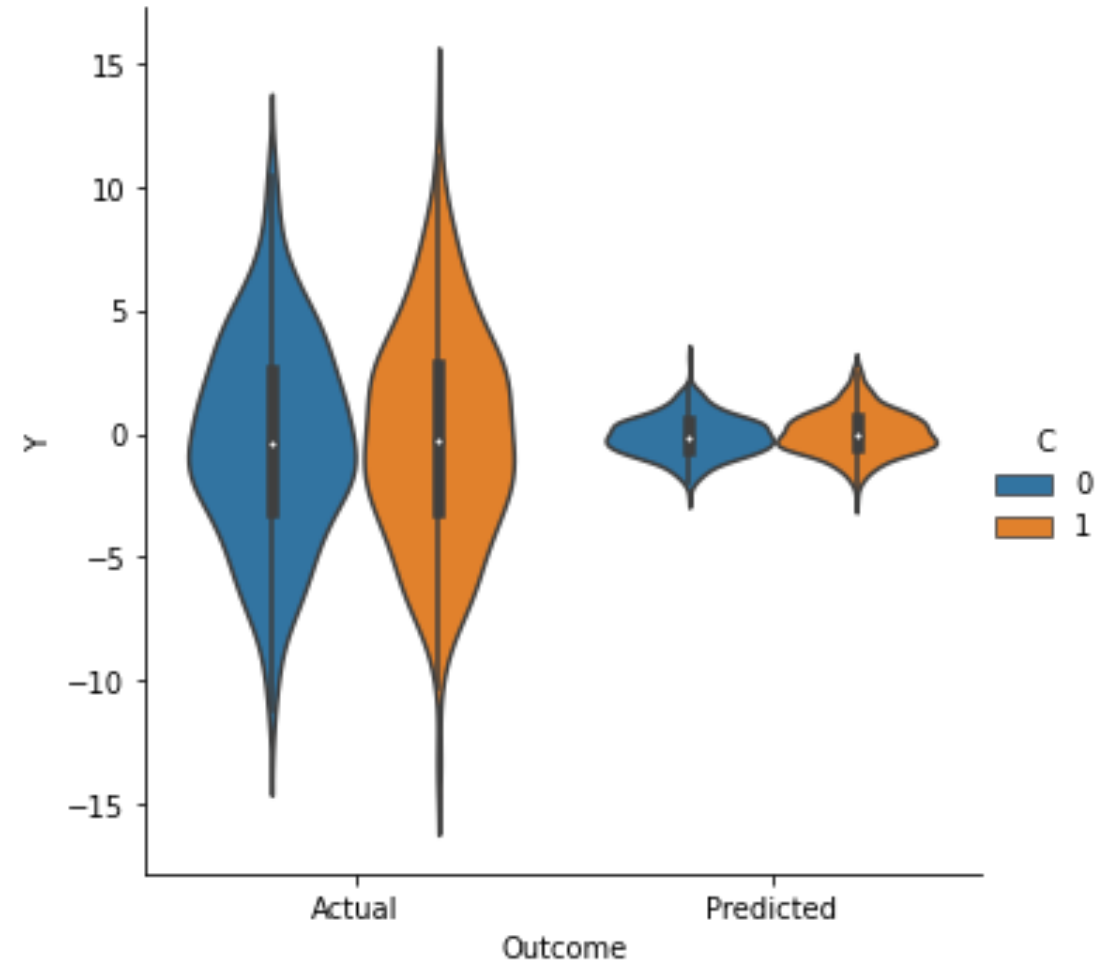
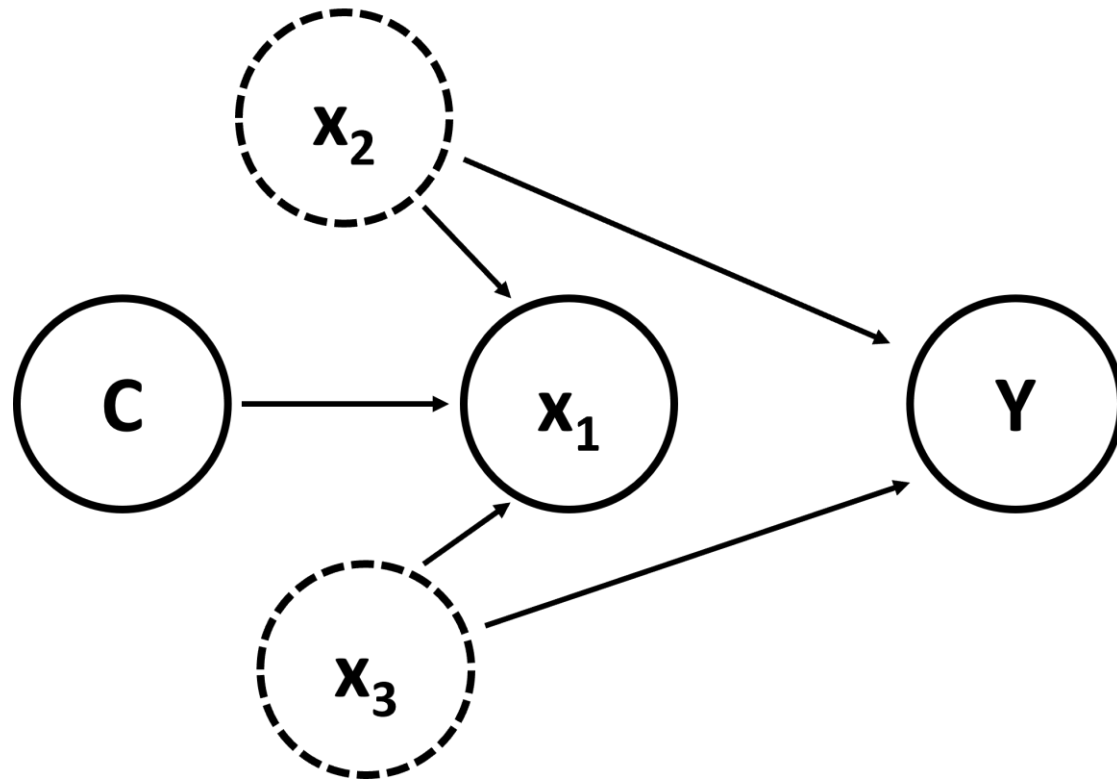
Scenario 4



Scenario 5



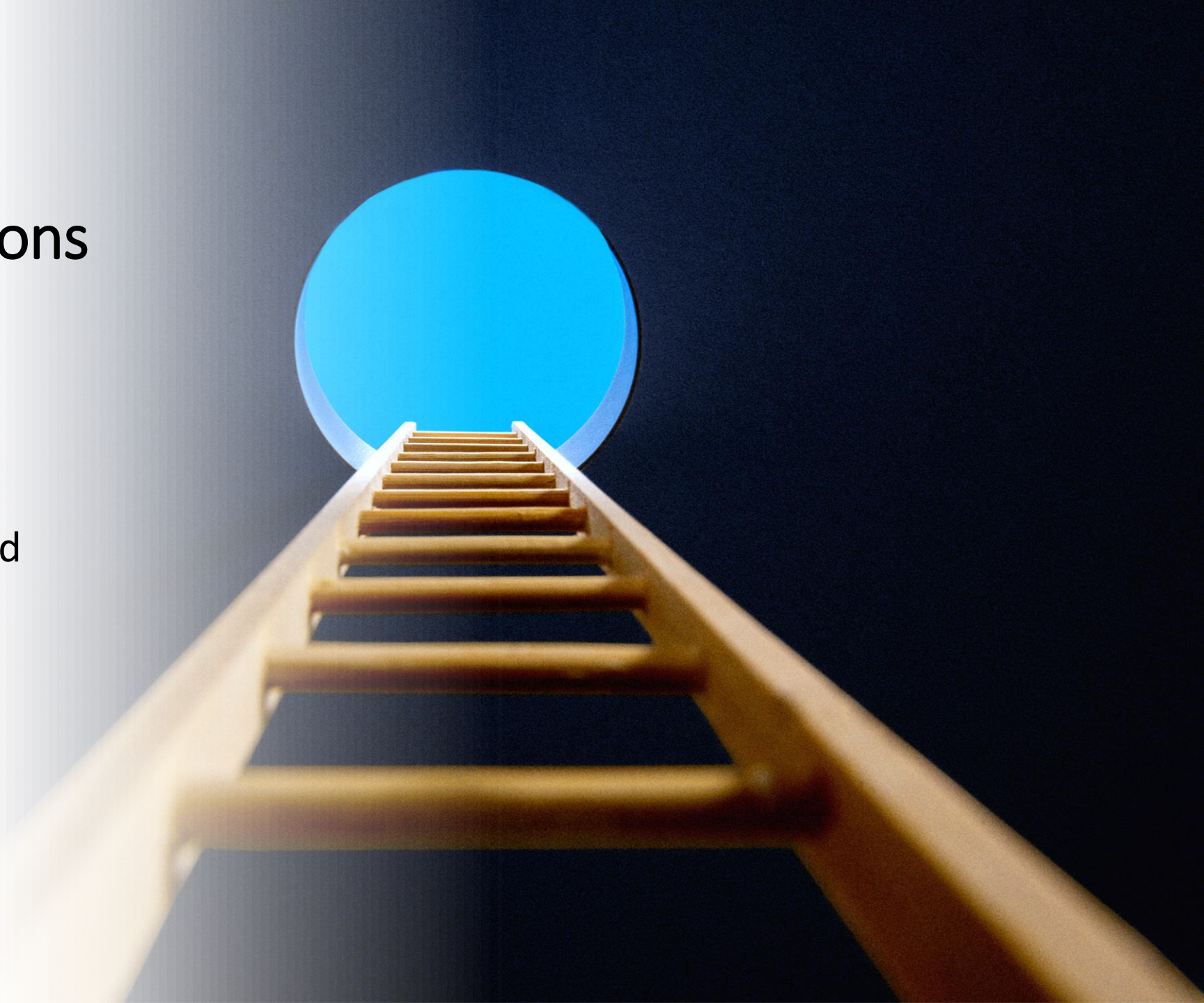
Scenario 5



Weight on C : -0.6

Inequality and Algorithmic Predictions

- Algorithms are often built to further one's interests over (and often at the expense of) others'
- Even after an algorithm is built, how/when it is used and when it is deemed no longer useful is left up to someone (typically someone in power)
- Predictions, especially algorithmic ones, follow us through our lives and feed into future, cementing our place in society



Inequality and Algorithmic Predictions

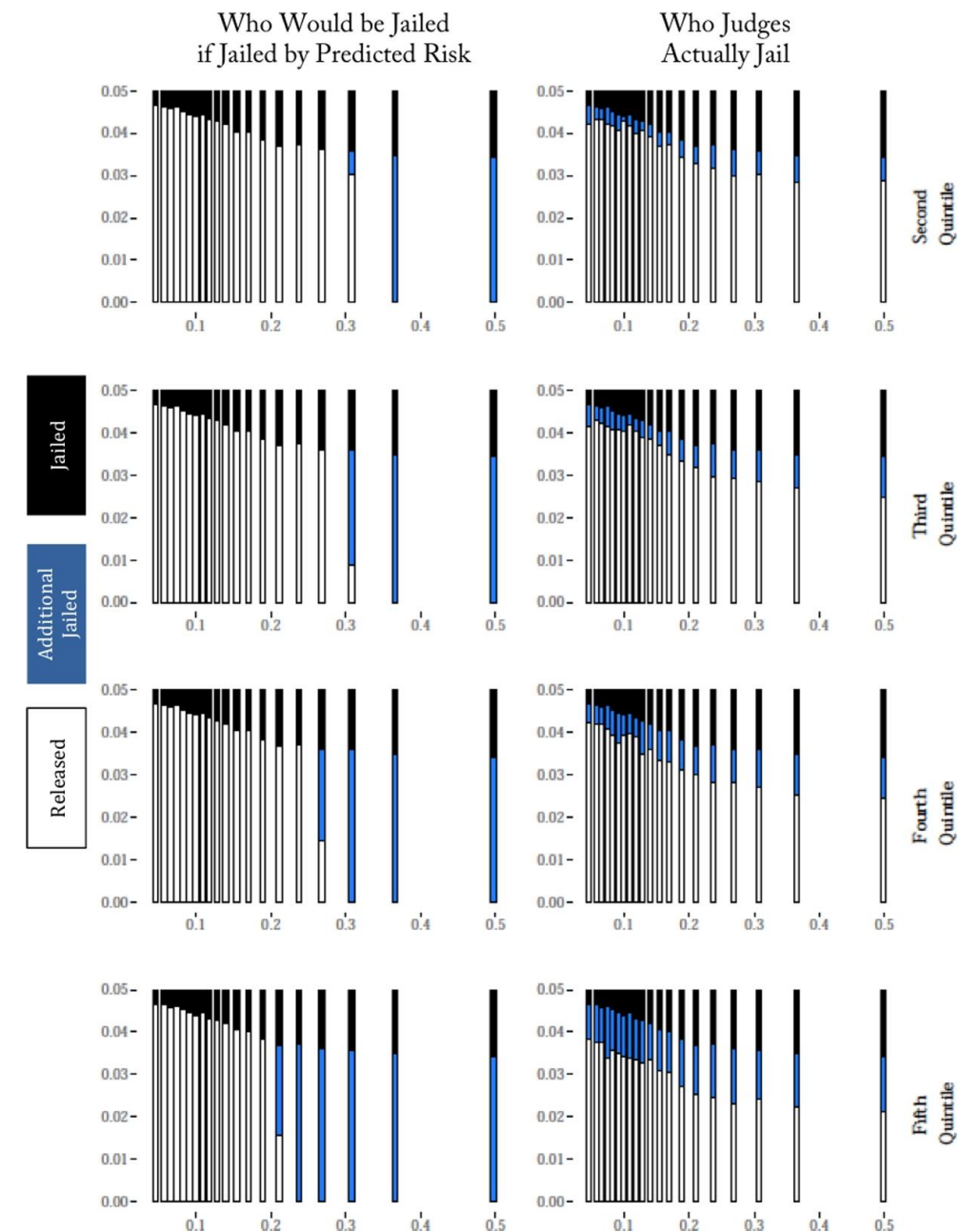
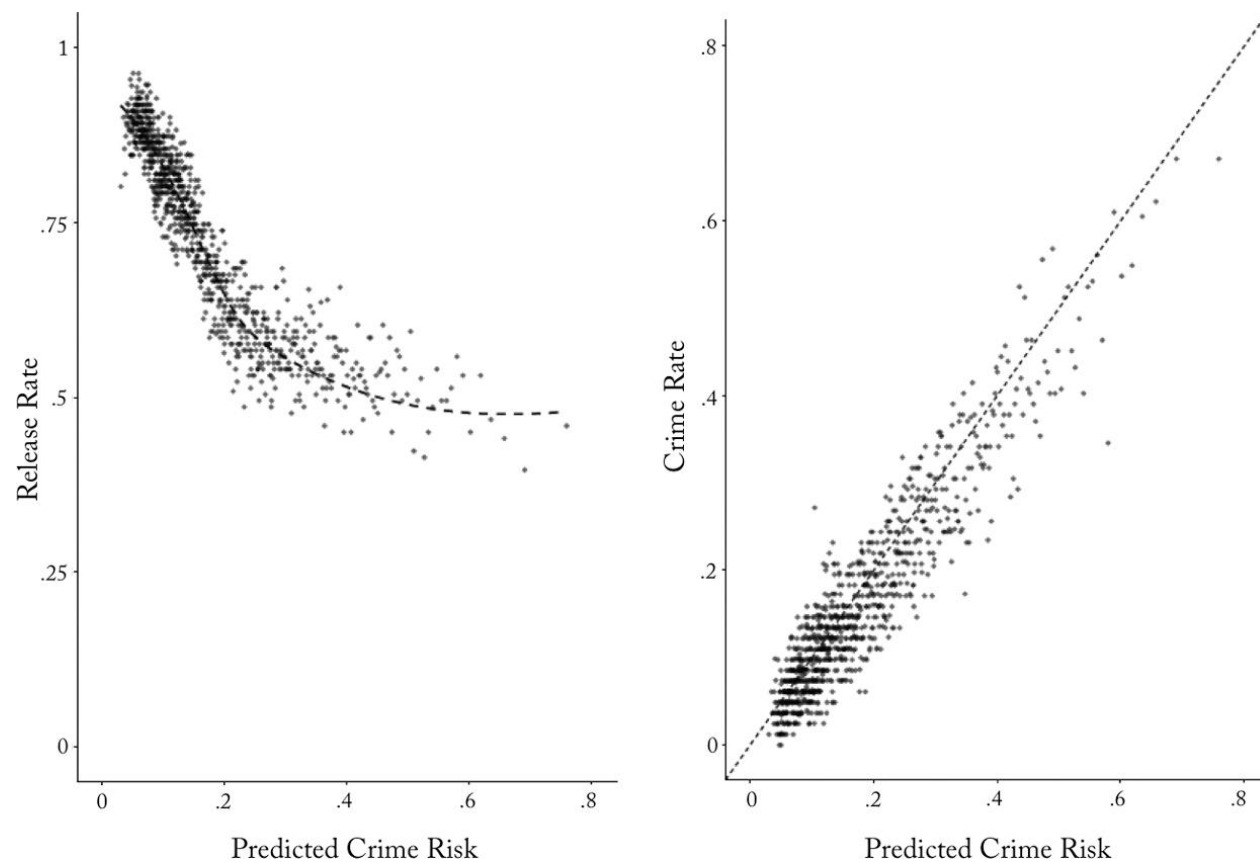
Human Decisions and Machine Predictions*

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan

The Quarterly Journal of Economics, Volume 133, Issue 1, February 2018, Pages 237–293,

<https://doi.org/10.1093/qje/qjx032>

Published: 26 August 2017



Inequality and Algorithmic Predictions

Simple rules to guide expert classifications

Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel✉, Daniel G. Goldstein

First published: 27 May 2020 | <https://doi.org/10.1111/rssa.12576> | Citations: 8

Table 3. Simple rule for estimating the risk of flight, where a defendant’s risk is obtained by summing the appropriate scores for age and prior history of FTA

<i>Feature</i>	<i>Score</i>	<i>Feature</i>	<i>Score</i>
$18 \leq \text{age} < 21$	3	No prior FTAs	0
$21 \leq \text{age} < 31$	2	1 prior FTA	2
$31 \leq \text{age} < 51$	1	2 or more prior FTAs	3
$51 \leq \text{age}$	0		

Please fill out the survey at :
tinyurl.com/week2pa