THE UNIVERSITY *of* EDINBURGH

Methods for Causal Inference
Lecture 3: Regression, graphs, conventions

Ava Khamseh

School of Informatics

2023-2024

Last time …

Language of probability: Variables, evens, samples space, probability law

Probability axioms, (conditional) total law of probability, independence, Bayes’ rule

Expected values, variance, correlation

Anscombe's Quartet

Group of 4 datasets with nearly identical simple descriptive statistical properties:

- Mean and sample variance of X
- Mean and sample variance of Y
- Correlation between X and Y
- Linear regression line (coefficient the same up to 2 or 3 decimal places)
- R^2 coefficient

A note on R^2: A measure for goodness-of-fit

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \ , \ y_i = f(x_i) \ , \ \bar{y} = \frac{1}{n} \sum_i y_i$$

If the fit y=f(x) is a perfect fit, the numerator is zero, $R^2 = 1$, and
$R^2 = 0$ implies the fit f(x) is no better than baseline average $\bar{y}$.
Negative values corresponds to models worse than the baseline average.

Anscombe's Quartet

Group of 4 datasets with nearly identical simple descriptive statistical properties:

- Mean and sample variance of X
- Mean and sample variance of Y
- Correlation between X and Y
- Linear regression line (coefficient the same up to 2 or 3 decimal places)
- R^2 coefficient

Yet, very different distributions, which can be observed by plotting the graphs

Same Pearson correlation, but,
different dependence structure
(X causes Y, but in different ways)

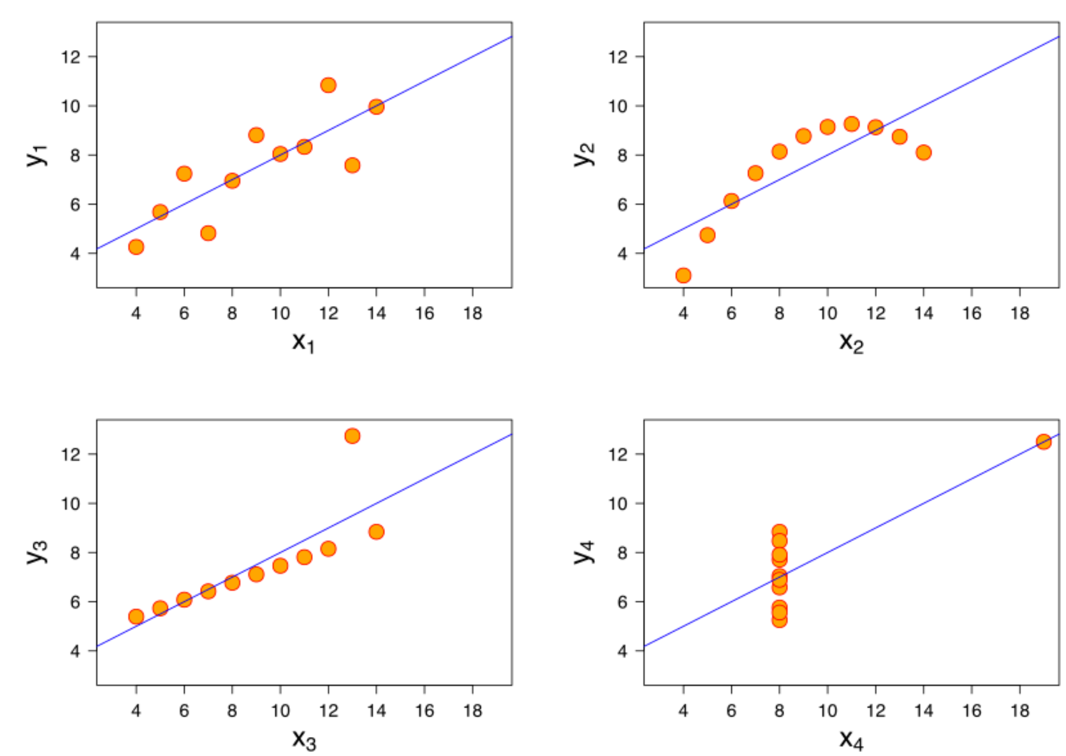


Figure from WikiPedia

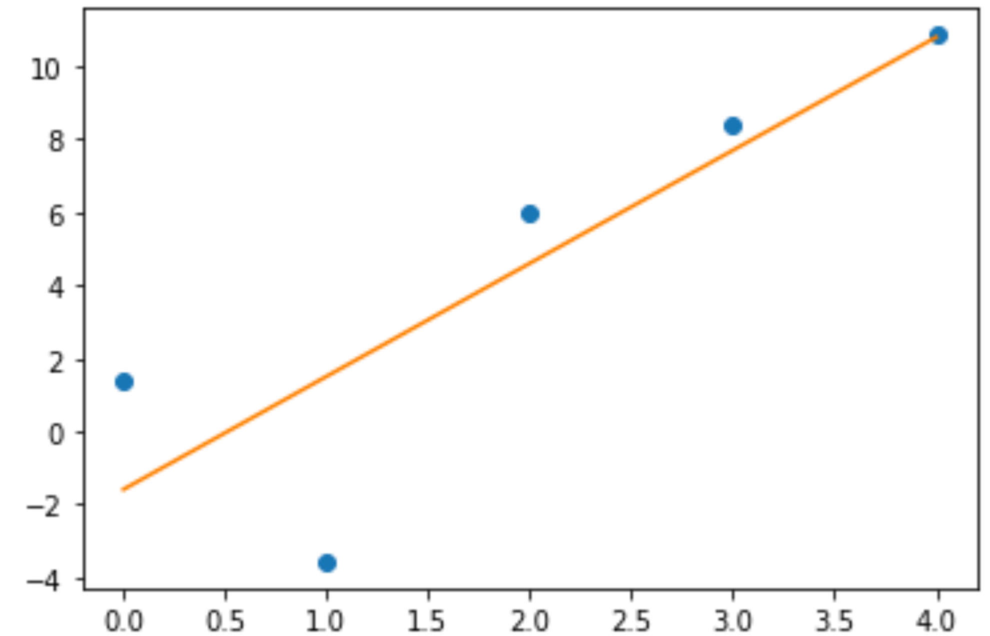
Regression

Suppose we wish to predict the value of an outcome Y, based on the value of some input X. The best prediction of Y based on X is given by $\mathbb{E}[Y|X=x]$ ('best': in terms of minimum loss function, on average, e.g. square loss)

Wish to estimate $\mathbb{E}[Y|X=x]$ from data -> **Regression**
Linear regression is <u>**a**</u> model that can be employed do this, but they are many other parametric (e.g. polynomial, GLMs) and non-parametric methods.

Let $f(x_i)$ be the value of the line $y = \alpha + \beta x$ at
The least squares regression line minimises:

$$\sum_i (y_i - f(x_i))^2 = \sum_i (y_i - \alpha - \beta x_i)^2$$



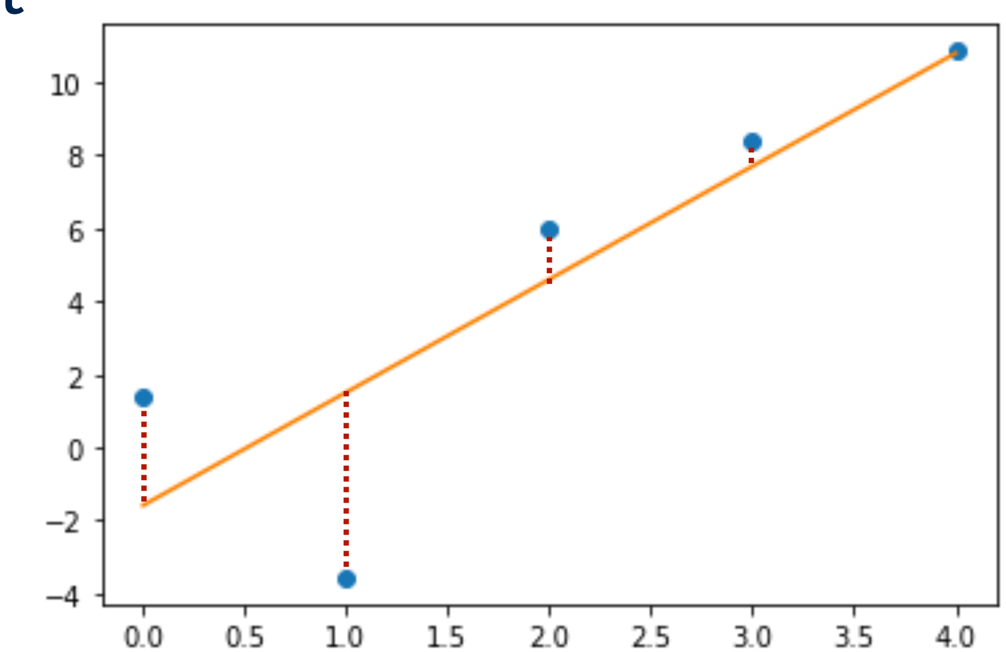
Regression

Suppose we wish to predict the value of an outcome Y, based on the value of some input X. The best prediction of Y based on X is given by $\mathbb{E}[Y|X=x]$ ('best': in terms of minimum loss function, on average, e.g. square loss)

Wish to estimate $\mathbb{E}[Y|X=x]$ from data -> **Regression**
Linear regression is a model that can be employed do this, but they are many other parametric (e.g. polynomial, GLMs) and non-parametric methods.

Let $f(x_i)$ be the value of the line $y = \alpha + \beta x$ at
The least squares regression line minimises:

$$\sum_i (y_i - f(x_i))^2 = \sum_i (y_i - \alpha - \beta x_i)^2$$



i.e. the sum of distances between the points and the line.

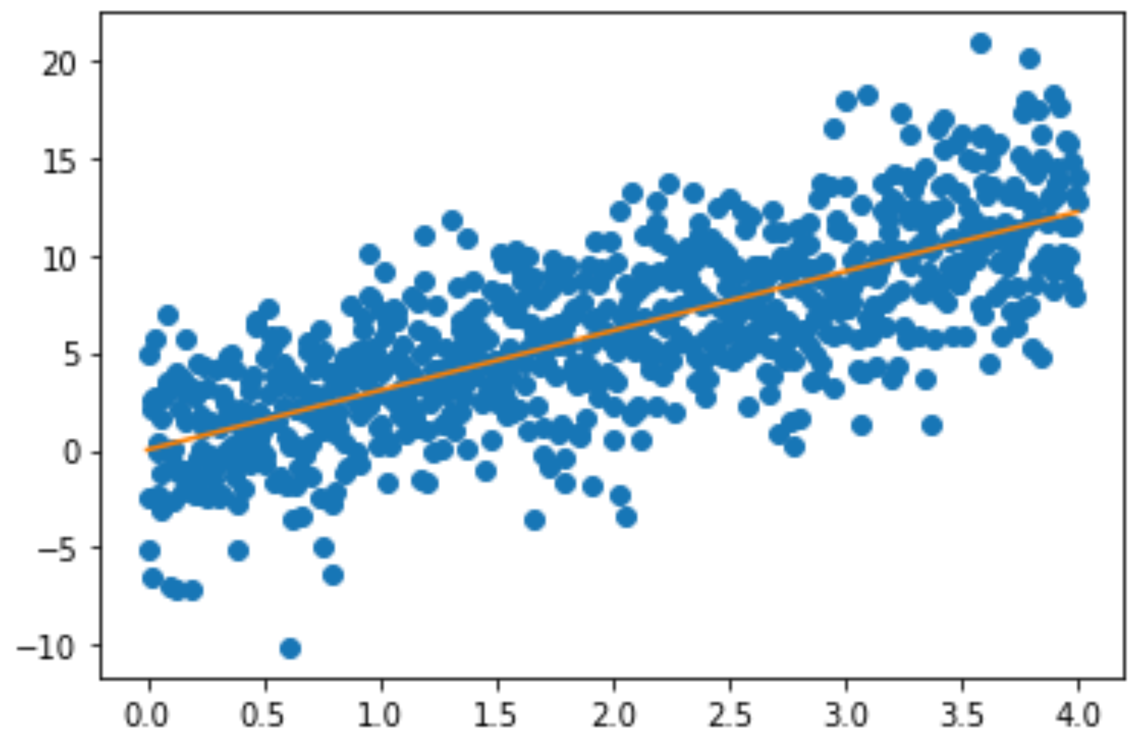
Regression

Suppose we wish to predict the value of an outcome Y, based on the value of some input X. The best prediction of Y based on X is given by $\mathbb{E}[Y|X=x]$ ('best': in terms of minimum loss function, on average, e.g. square loss)

Wish to estimate $\mathbb{E}[Y|X=x]$ from data -> **Regression**
Linear regression is a model that can be employed do this, but they are many other parametric (e.g. polynomial, GLMs) and non-parametric methods.

Assumptions:
1. **Linearity**: Y depends linearly on X
2. **Homoscedasticity**: variance of residual is the same for any value of X

Residual for every point: $y_i - f(x_i)$



Regression

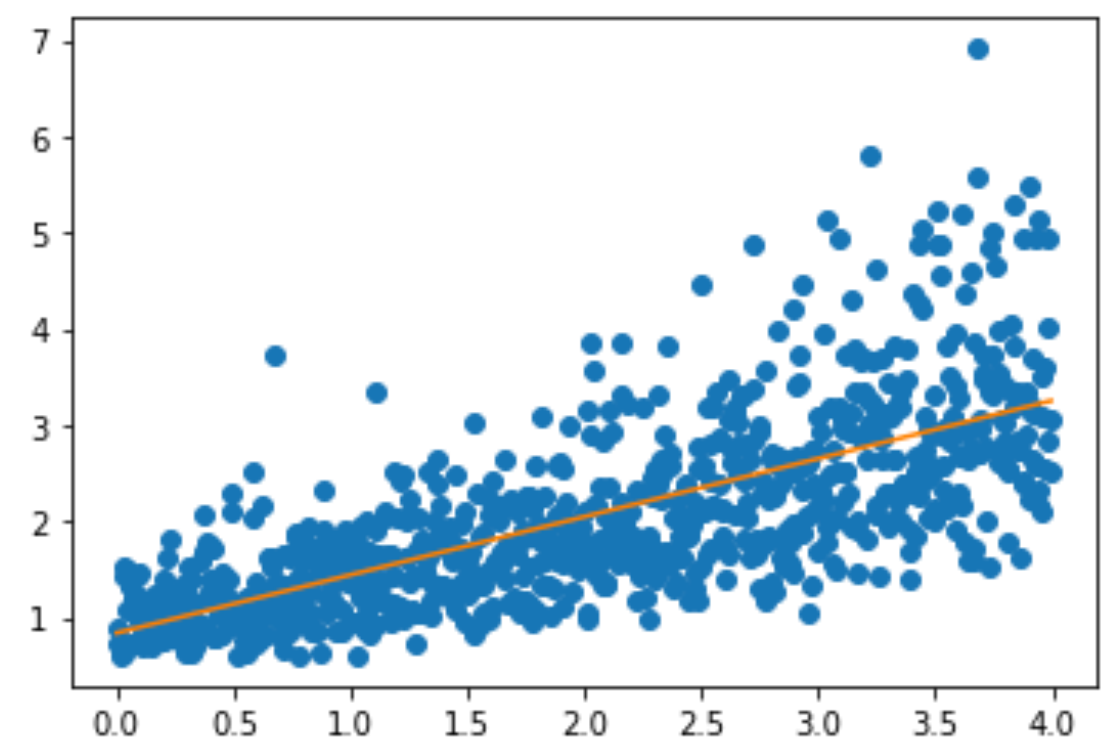
Suppose we wish to predict the value of an outcome Y, based on the value of some input X. The best prediction of Y based on X is given by $\mathbb{E}[Y|X=x]$ ('best': in terms of minimum loss function, on average, e.g. square loss)

Wish to estimate $\mathbb{E}[Y|X=x]$ from data -> **Regression**

Linear regression is <u>**a**</u> model that can be employed do this, but they are many other parametric (e.g. polynomial, GLMs) and non-parametric methods.

Assumptions:

1. **Linearity**: Y depends linearly on X
2. **Homoscedasticity**: variance of residual is the same for any value of X

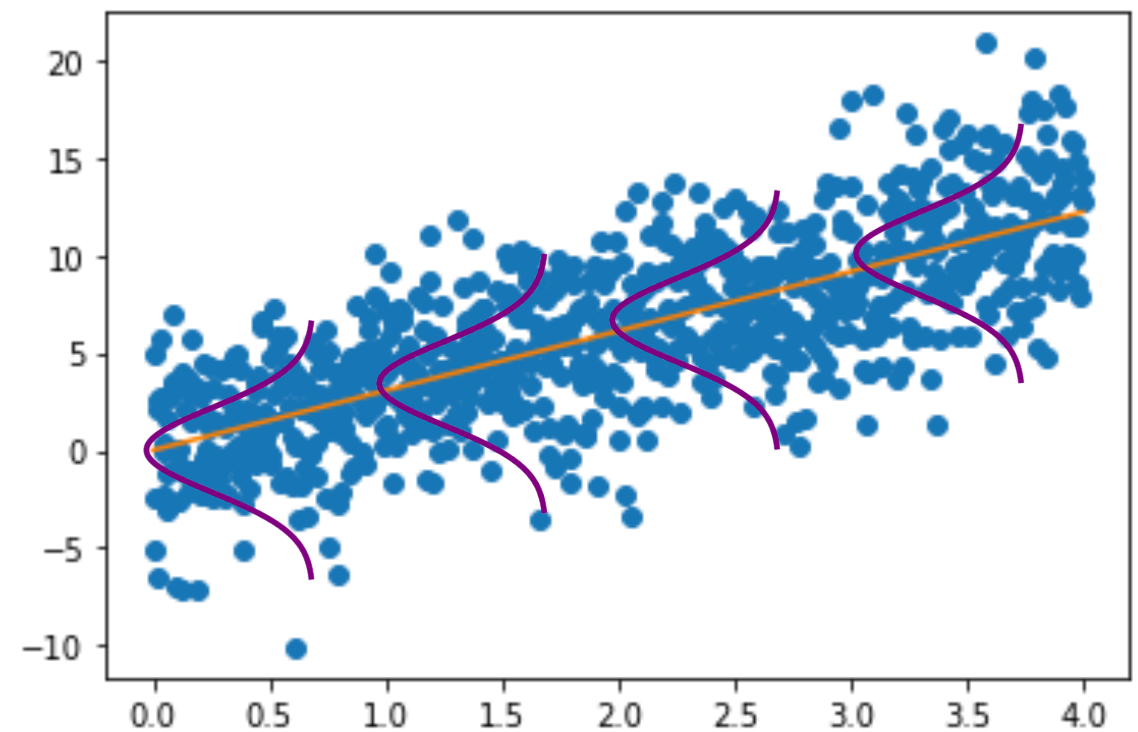


Regression

Suppose we wish to predict the value of an outcome Y, based on the value of some input X. The best prediction of Y based on X is given by $\mathbb{E}[Y|X=x]$ ('best': in terms of minimum loss function, on average, e.g. square loss)

Wish to estimate $\mathbb{E}[Y|X=x]$ from data -> **Regression**
Linear regression is a model that can be employed do this, but they are many other parametric (e.g. polynomial, GLMs) and non-parametric methods.

Assumptions:
1. **Linearity**: Y depends linearly on X
2. **Homoscedasticity**: variance of residual is the same for any value of X
3. **Independence** of observations
4. **Normality**: For any fixed value of X, Y is normally distributed



Regression

Suppose we wish to predict the value of an outcome Y, based on the value of some input X. The best prediction of Y based on X is given by $\mathbb{E}[Y|X=x]$ ('best': in terms of minimum loss function, on average, e.g. square loss)

Wish to estimate $\mathbb{E}[Y|X=x]$ from data -> **Regression**
Linear regression is <u>**a**</u> model that can be employed do this, but they are many other parametric (e.g. polynomial, GLMs) and non-parametric methods.

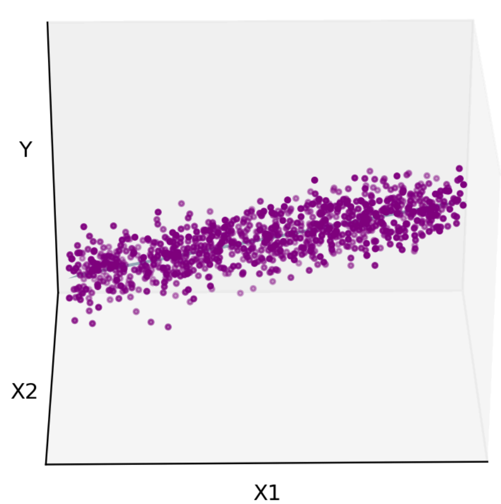
$$y = \alpha + \beta x \;\Rightarrow\; \beta = \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]}$$

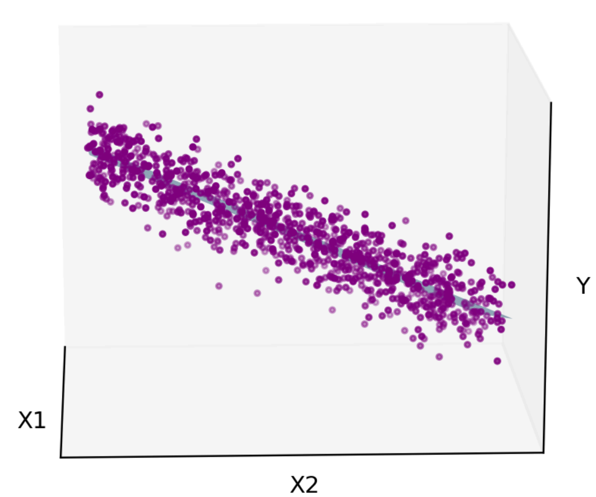
i.e. non-symmetric: Slope of Y on X is different from X on Y.
Positive correlation if $\beta > 0$, negative correlation if $\beta < 0$ (dependent)
No linear correlation if $\beta = 0$

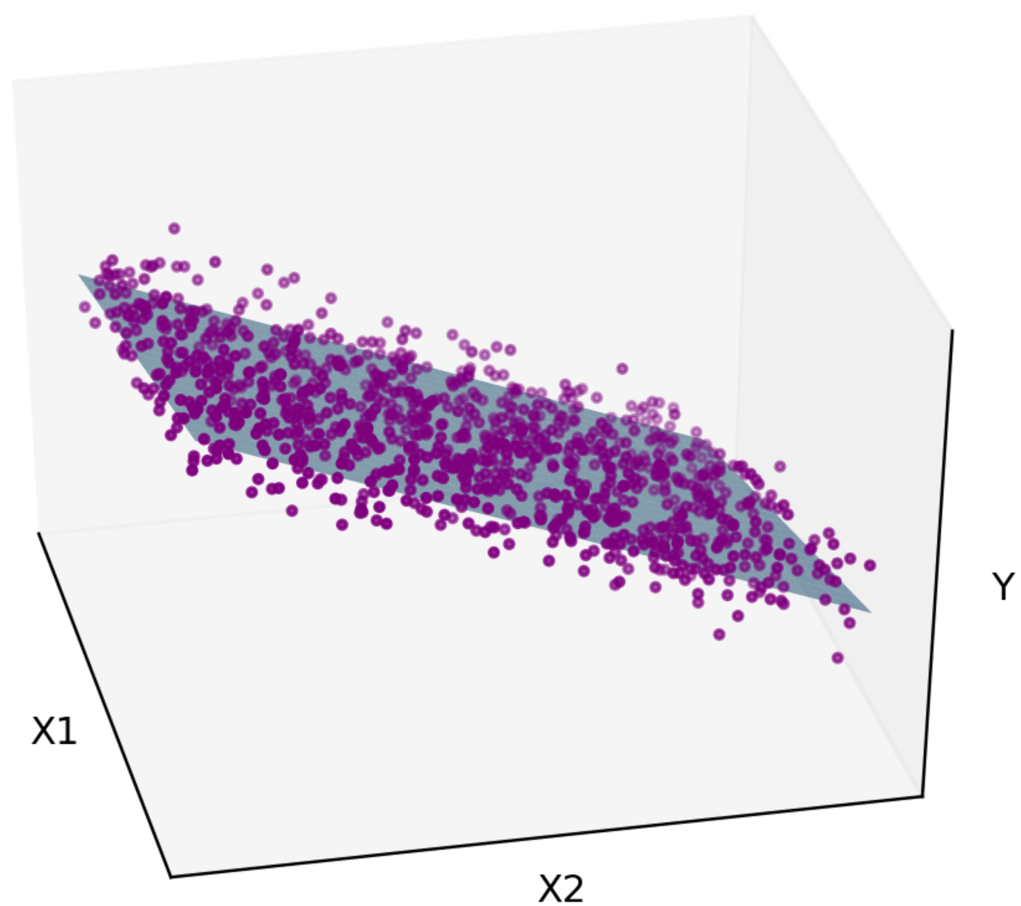
Multiple Regression

Regress Y on multiple variables, e.g., X_1 and X_2: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$ represents a plane in 3-dimensions.

In 2D: The regression lines with slopes β_1 and β_2 .



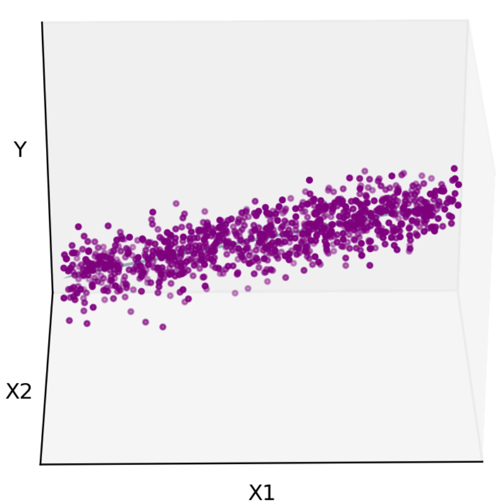


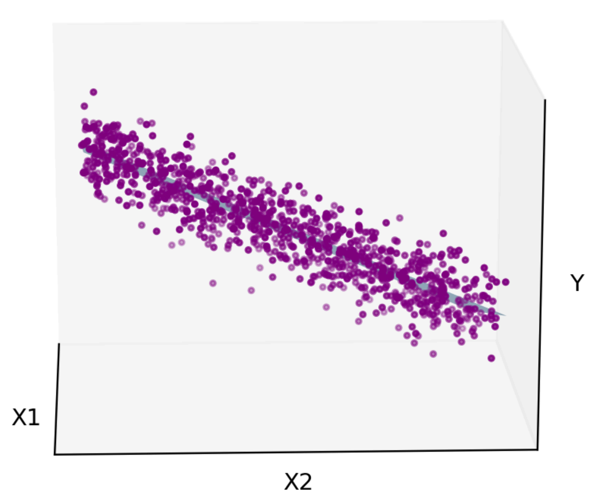


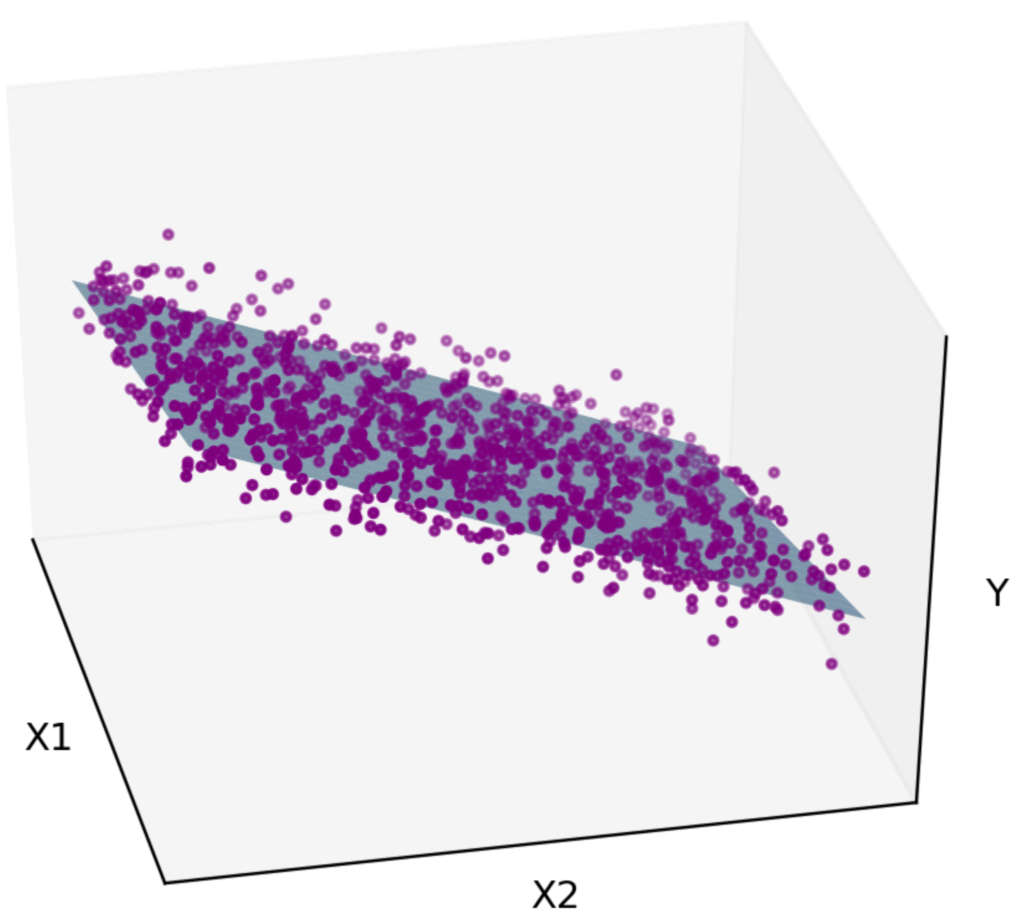
Multiple Regression

Regress Y on multiple variables, e.g., X_1 and X_2: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$ represents a plane in 3-dimensions.
In 2D: The regression lines with slopes β_1 and β_2 .

X_1 is positively correlated with Y, irrespective of X_2, since $X_1 \perp\!\!\!\perp X_2$







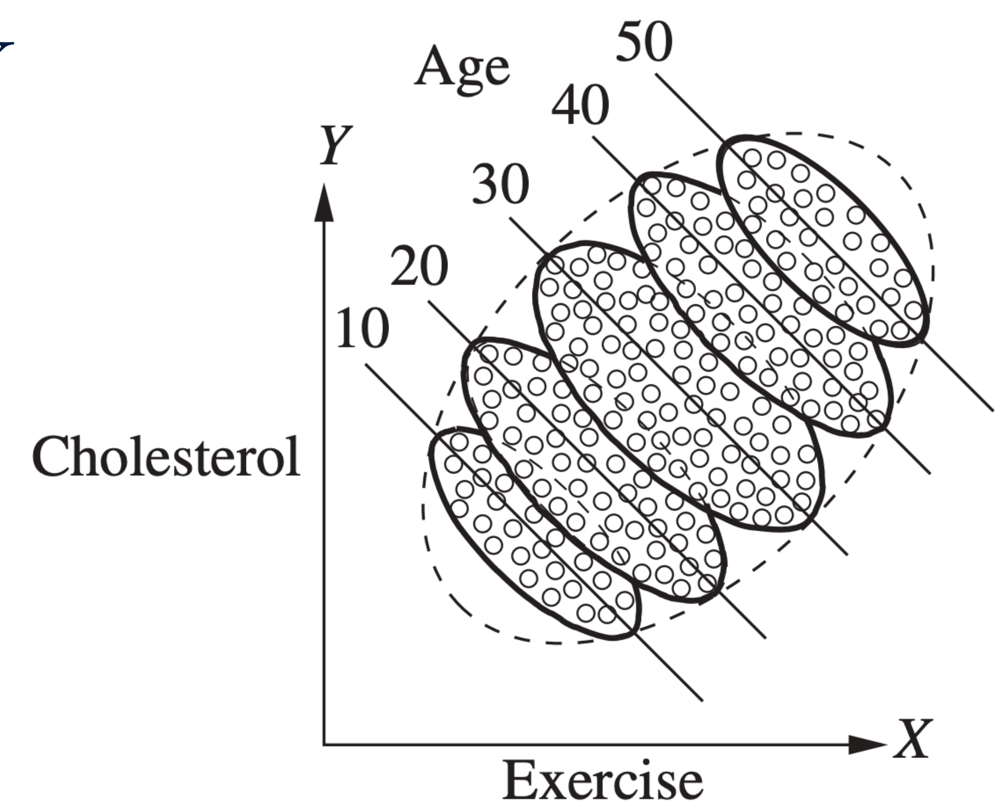
Multiple Regression

Regress Y on multiple variables, e.g., X_1 and X_2: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$ represents a plane in 3-dimensions.
In 2D: The regression lines with slopes β_1 and β_2 .

X_1 is positively correlated with Y, irrespective of X_2, since $X_1 \perp\!\!\!\perp X_2$

But when $X_1 \not\perp\!\!\!\perp X_2$ it is possible for
X_1 to be positively correlated with Y overall,
but for fixed X_2 be negatively correlated with Y

Example: Simpson's paradox



Improving estimate via ensemble learning [non-examinable]

- Do we need the additivity assumption?
- In fact, ignoring covariate-treatment interaction can be a source of bias
- Data driven approach:

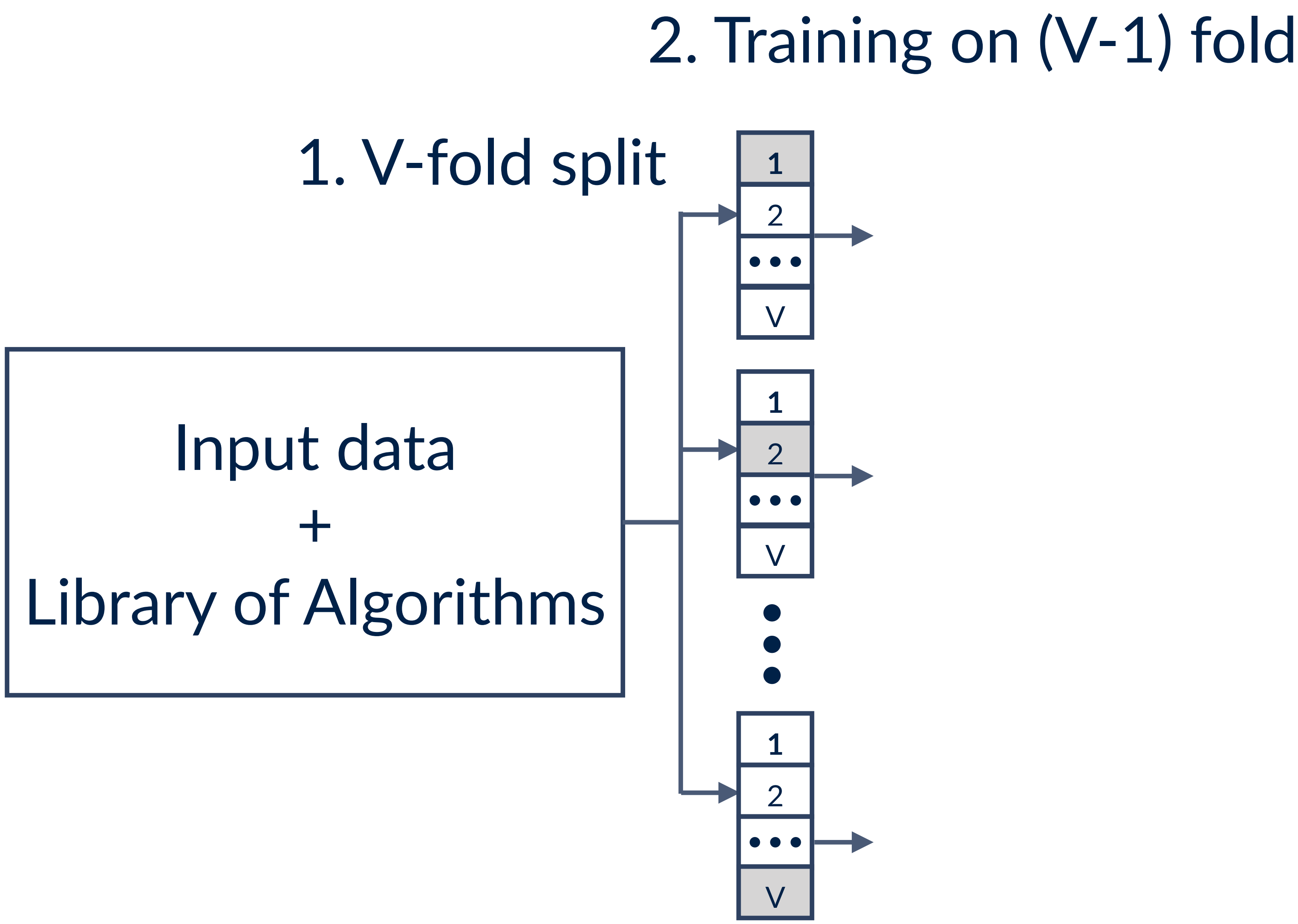
$$\mathbb{E}_0\left(Y|T,X\right) = \beta_0 + \beta_X X + \beta_T T + \gamma XT$$

$$\mathbb{E}_0\left(Y|T,X\right) = \beta_0 + \beta_X X + \beta_T T + \gamma XT + \beta'_X X^2$$

$$\mathbb{E}_0\left(Y|T,X\right) = \beta_0 + \beta_X X + \beta_T T + \gamma XT + \beta'_X X^2 + \gamma' X^2 T$$

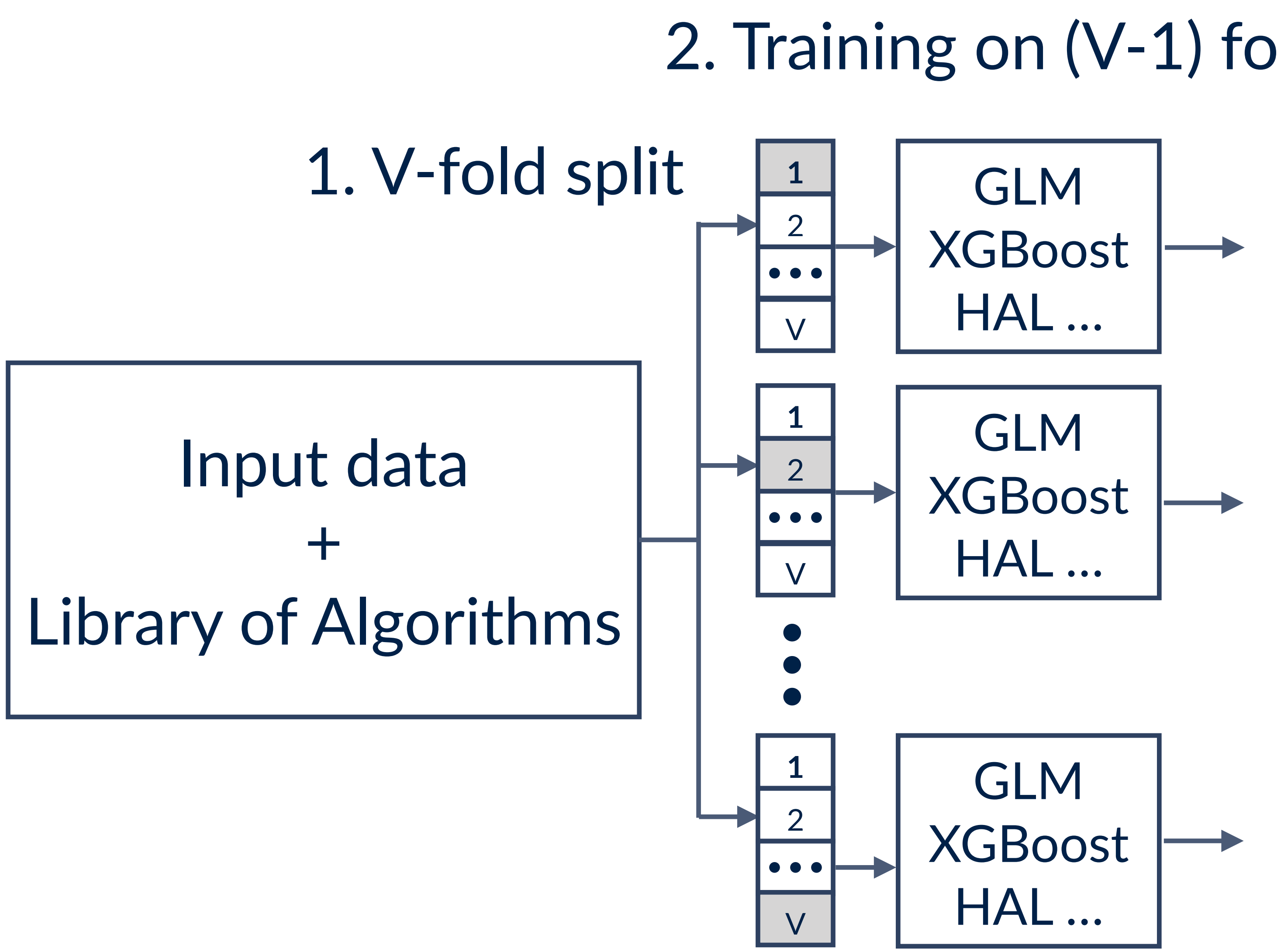
- V-fold cross-validation using an ensemble learning, e.g. super-learner
- Appropriate **choice of loss function**, e.g., L1 for conditional median, L2 for conditional mean, log loss for binary outcome, …

Continuous Super Learner [non-examinable]



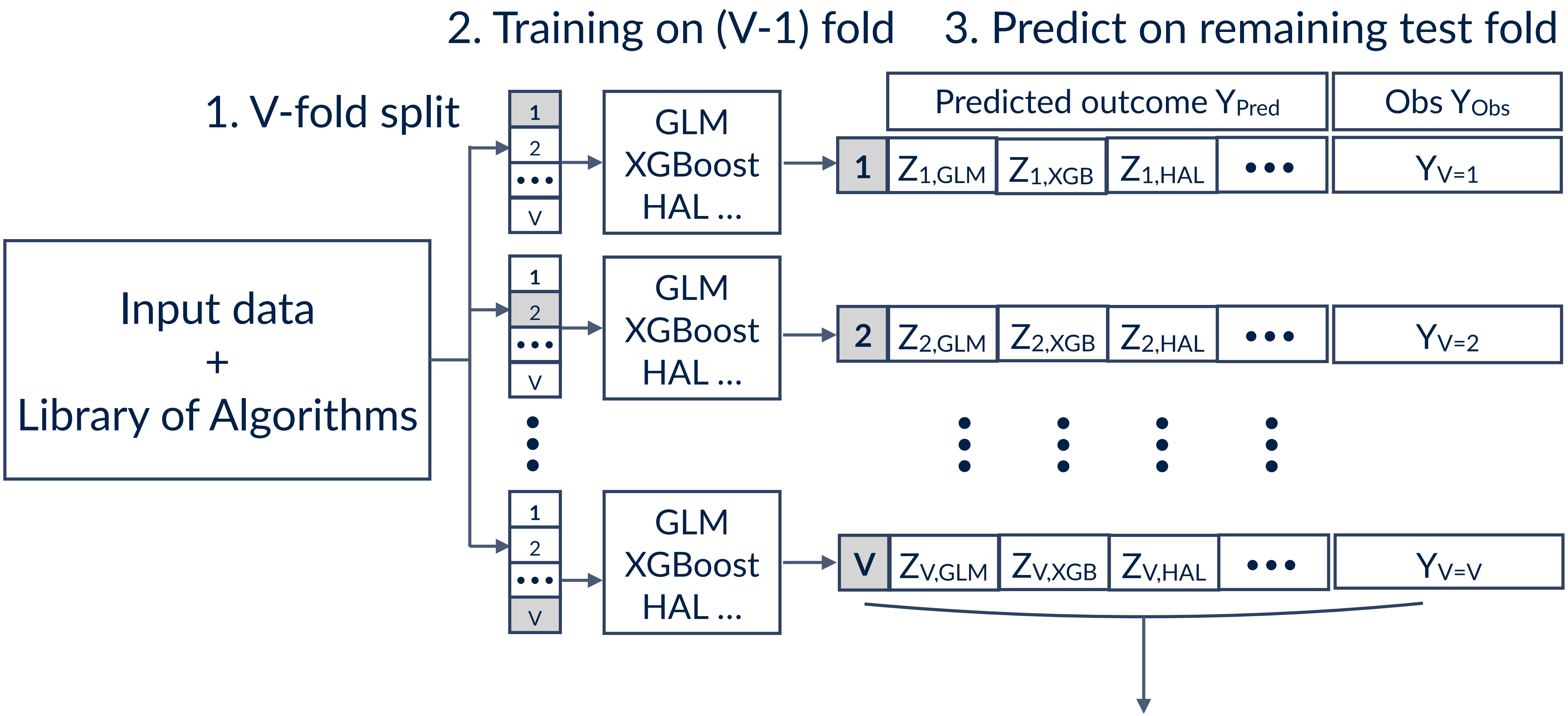
Eric Polley, Mark van der Laan, Sherri Rose 2011

Continuous Super Learner [non-examinable]



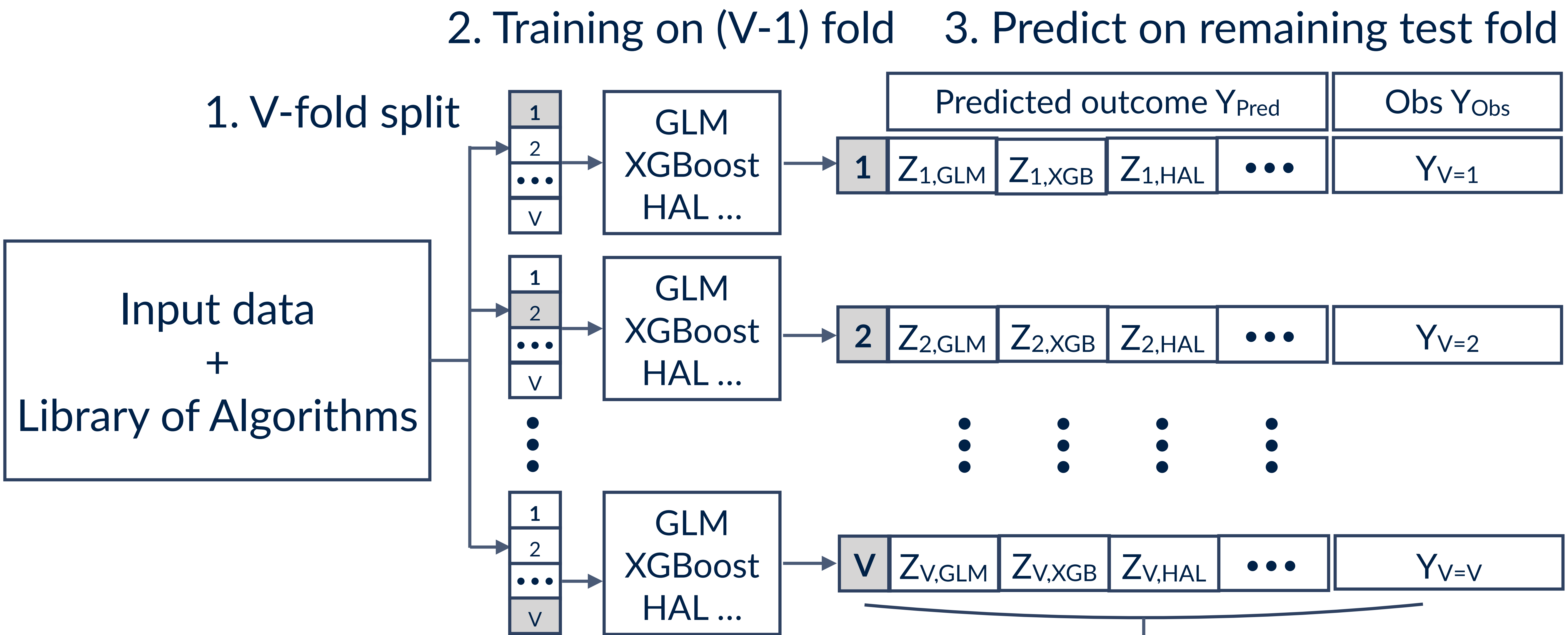
Eric Polley, Mark van der Laan, Sherri Rose 2011

Continuous Super Learner [non-examinable]



Eric Polley, Mark van der Laan, Sherri Rose 2011

Continuous Super Learner [non-examinable]



4. Fit the weight α for each algorithm

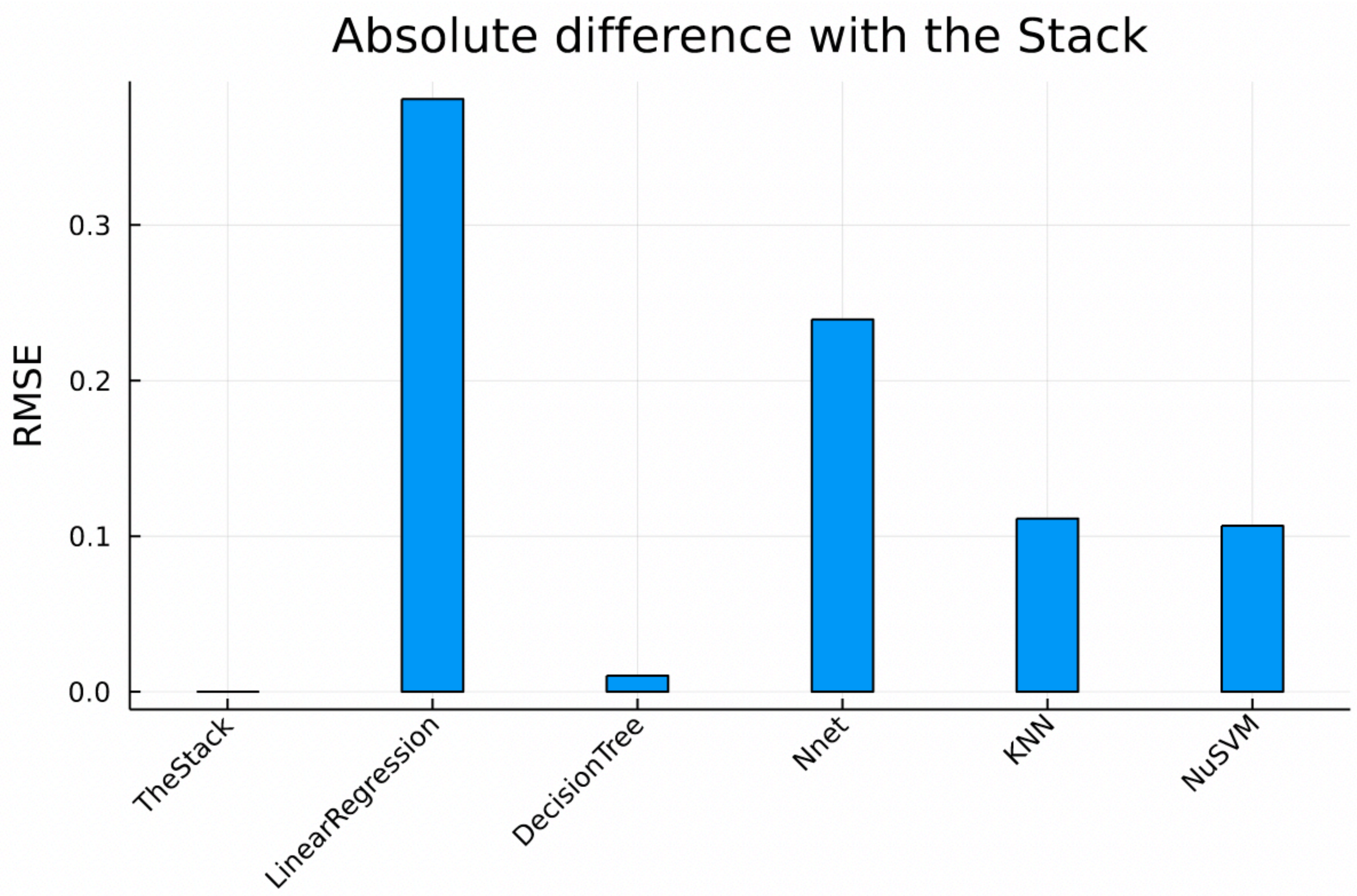
$$\mathbb{E}[Y_{\mathrm{Obs}}|Y_{\mathrm{Pred}}] = \alpha_1 Y_{\mathrm{GLM}} + \alpha_2 Y_{\mathrm{HAL}} + \alpha_3 Y_{\mathrm{XGB}} + \cdots$$

5. Train each algorithm on entire dataset combined with fitted weights

+ verify goodness-of-fit

Eric Polley, Mark van der Laan, Sherri Rose 2011

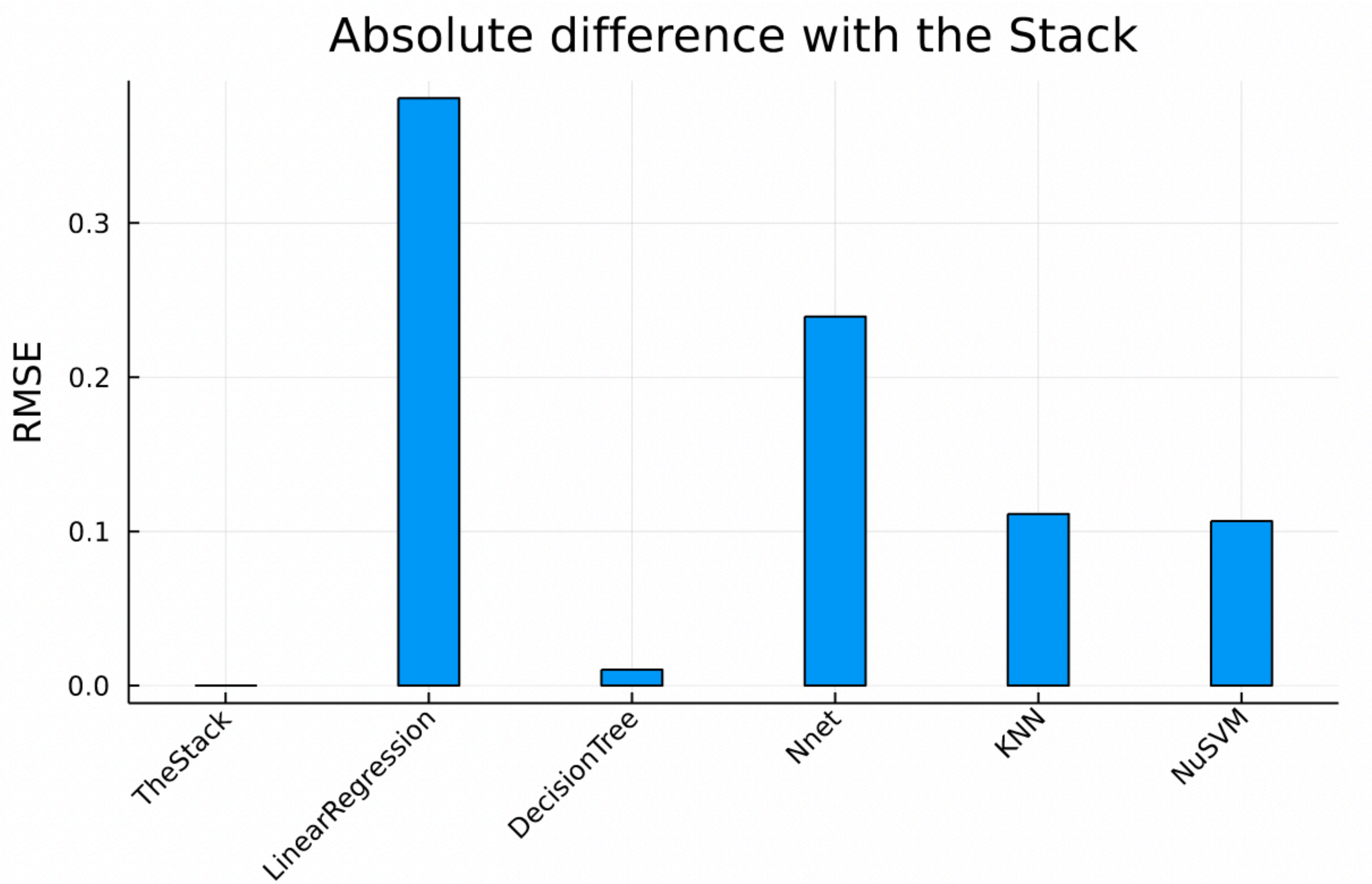
Discrete Super Learner [non-examinable]



Smaller mean squared error = better performance

Simulation by Olivier Labayle Pabet

Discrete Super Learner [non-examinable]



Theorem (Van der Laan, Polley, Hubbard; 2007)

Asymptotically, the stack always wins

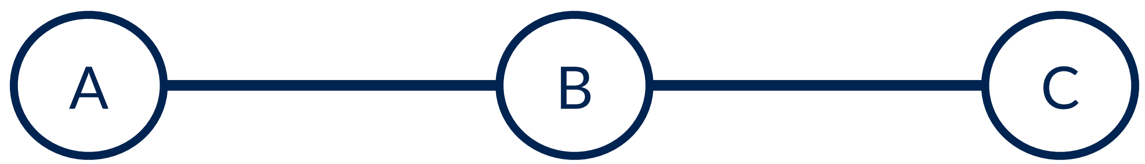
Simulation by Olivier Labayle Pabet

Basics of Graphs

Simpson's paradox: concrete example of why data alone is not enough!

Need to represent causal knowledge as part of a graph **Graph theory**

Graph: A collection of **nodes** (vertices) and **edges**.



Adjacent nodes: If there is an edge connecting them: A and B, B and C

Complete graph: There exist an edge between every pair of nodes (not above)

Path: sequences of nodes beginning with node X and ending with X', e.g.,

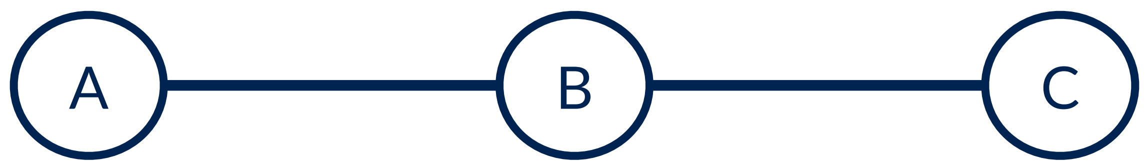
There is a path from A to C because A is connected to B and B is connected to C.

Basics of Graphs

Simpson's paradox: concrete example of why data alone is not enough!

Need to represent causal knowledge as part of a graph → **Graph theory**

Graph: A collection of **nodes** (vertices) and **edges**.



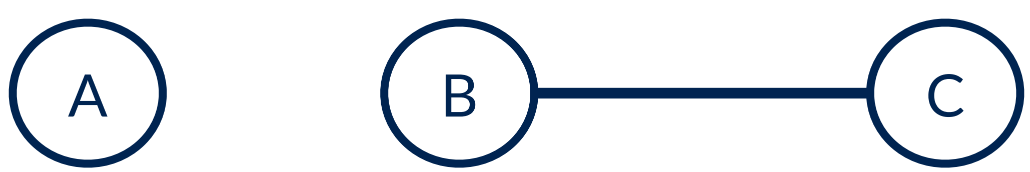
Adjacent nodes: If there is an edge connecting them: A and B, B and C

Complete graph: There exist an edge between every pair of nodes (not above)

Path: sequences of nodes beginning with node X and ending with X', e.g.,

There is a path from A to C because A is connected to B and B is connected to C.

i.e., not this:



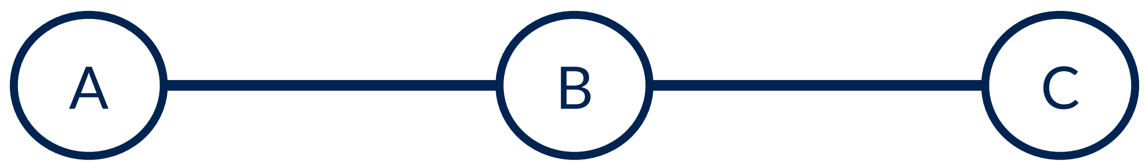
Basics of Graphs

Simpson's paradox: concrete example of why data alone is not enough!

Need to represent causal knowledge as part of a graph → **Graph theory**

Graph: A collection of **nodes** (vertices) and **edges**.

Undirected



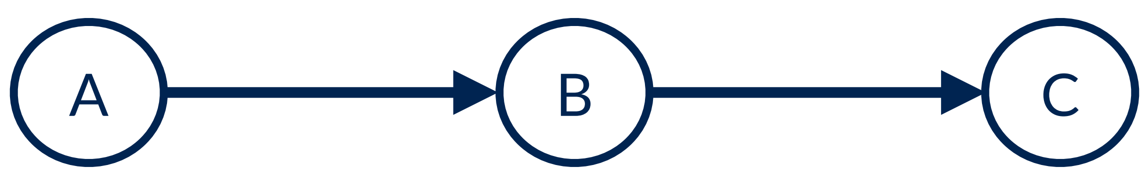
Adjacent nodes: If there is an edge connecting them: A and B, B and C

Complete graph: There exist an edge between every pair of nodes (not above)

Path: sequences of nodes beginning with node X and ending with X', e.g.,

Directed/Undirected: If the edges have in/out arrows

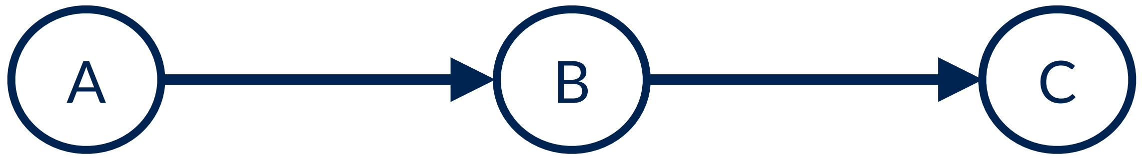
Directed



The node that a directed edge starts from: parent

The node a directed edge goes into: child of the node the edge comes from

Directed Graphs



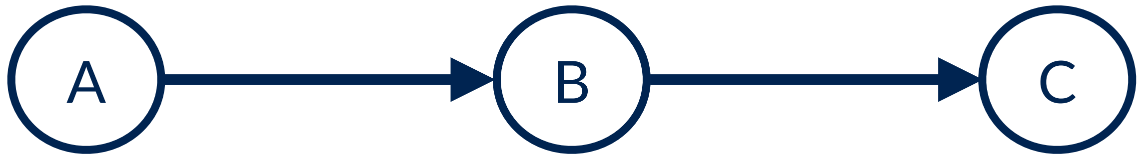
The node that a directed edge starts from: **parent**

The node a directed edge goes into: **child** of the node the edge comes from

E.g., A is the parent of B, B is the parent of C.

B is a child of A and C is a child of B

Directed Graphs

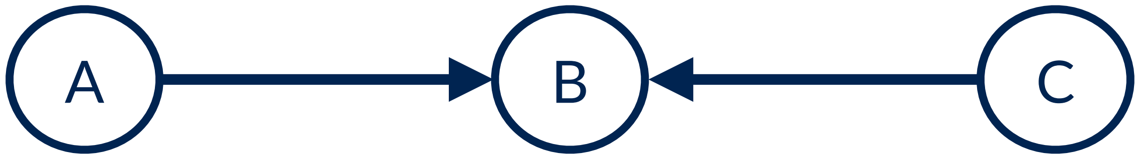


The node that a directed edge starts from: **parent**

The node a directed edge goes into: **child** of the node the edge comes from

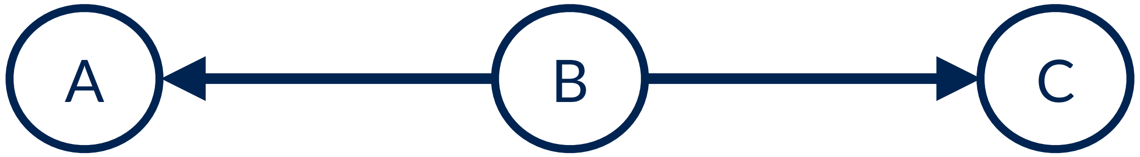
Directed Path: If the path can be traced along the arrows, i.e., A to B to C above.

Not:

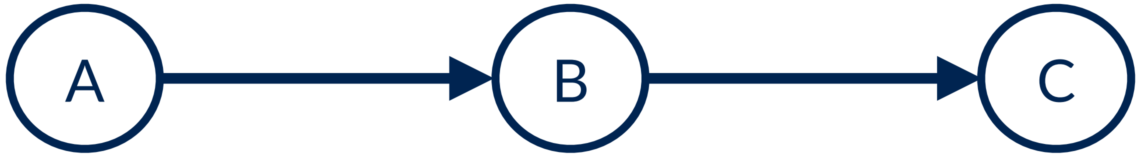


and

Not:



Directed Graphs



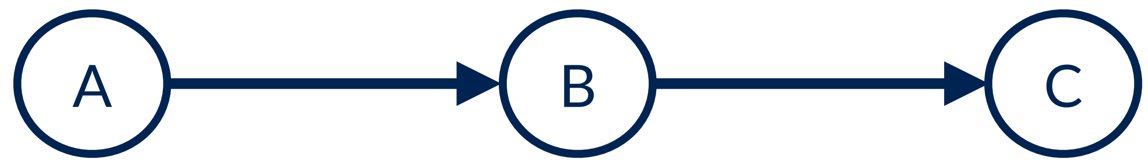
The node that a directed edge starts from: **parent**

The node a directed edge goes into: **child** of the node the edge comes from

Directed Path: If the path can be traced along the arrows, i.e., A to B to C above.

Two nodes connected by a direct path, first node (A) is the **ancestor** of every node in the path (B and C) and every node on the path is a **descendant** of it.

Directed Graphs



The node that a directed edge starts from: **parent**

The node a directed edge goes into: **child** of the node the edge comes from

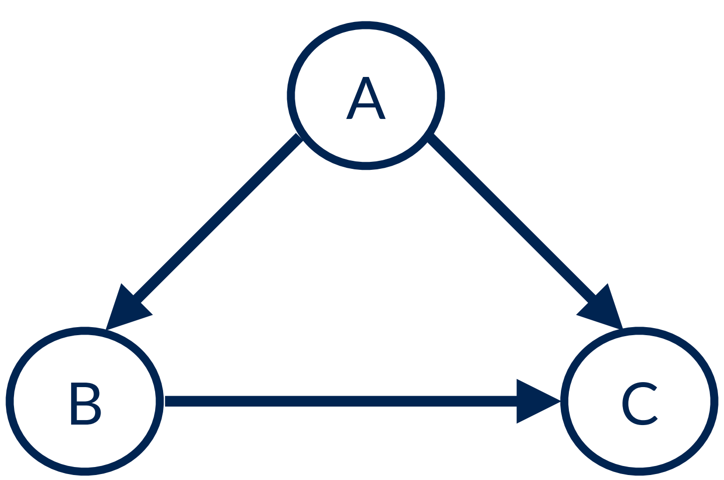
Directed Path: If the path can be traced along the arrows, i.e., A to B to C above.

Two nodes connected by a direct path, first node (A) is the **ancestor** of every node in the path (B and C) and every node on the path is a **descendant** of it.

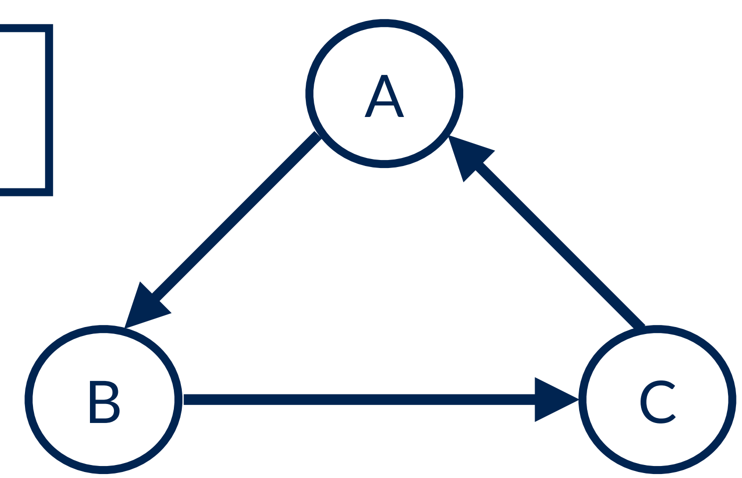
Cyclic: When a directed path exists from a node to itself (**complicates things!!**)

A direct graph with no cycles is **acyclic**.

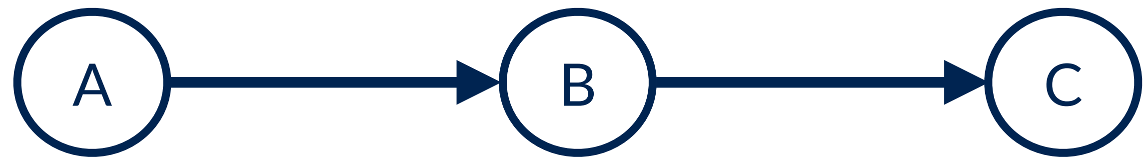
Acyclic



Cyclic



Directed Graphs



The node that a directed edge starts from: **parent**

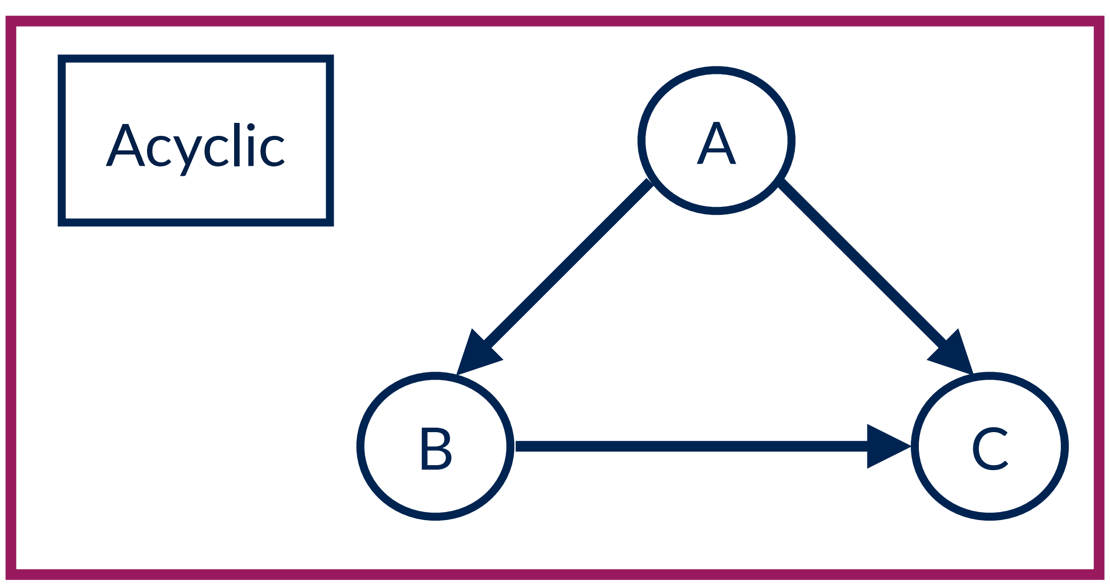
The node a directed edge goes into: **child** of the node the edge comes from

Directed Path: If the path can be traced along the arrows, i.e., A to B to C above.

Two nodes connected by a direct path, first node (A) is the **ancestor** of every node in the path (B and C) and every node on the path is a **descendant** of it.

Cyclic: When a directed path exists from a node to itself (**complicates things!!**)

A direct graph with no cycles is **acyclic**.



Directed Acyclic Graphs (DAGs)

A Brief Introduction to Structural Casual Models (SCMs)

Causality: Need to formally state our assumptions about the causal model,
the relevant features of the data, the role they play, how they relate to each other.

A Brief Introduction to Structural Casual Models (SCMs)

Causality: Need to formally state our assumptions about the causal model, the relevant features of the data, the role they play, how they relate to each other.

SCM: Consists of 2 set**s** of variables U and V, and a set of functions f. f assigns each variable in V a value based on other variables in U and V.

A Brief Introduction to Structural Casual Models (SCMs)

Causality: Need to formally state our assumptions about the causal model, the relevant features of the data, the role they play, how they relate to each other.

SCM: Consists of 2 sets of variables U and V, and a set of functions f. f assigns each variable in V a value based on other variables in U and V.

“A variable X is a **direct cause** of variable Y if X appears in the function that assigns Y’s value.

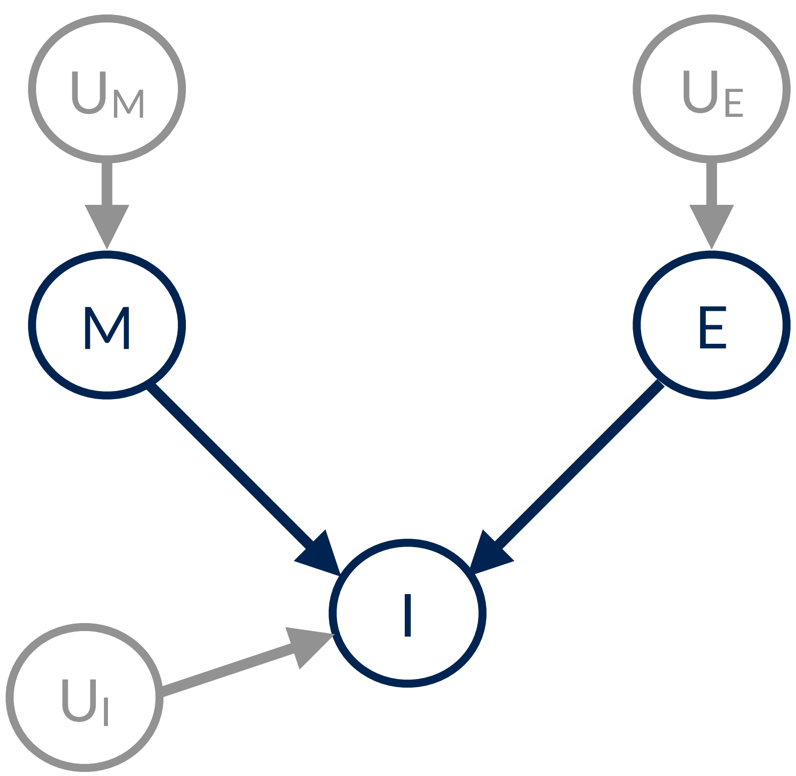
X is a cause of Y if it is a direct cause of Y or of any cause of Y.”

U: exogenous variables ‘external to the model’, e.g. noise or we simply do not explain how they are caused. Not descendants of any other variables. Roots.
V: endogenous variable which is a descendant of at least one exogenous variable

A Brief Introduction to Structural Casual Models (SCMs)

$$V = \{M, E, I\}$$
$$U = \{U_M, U_E, U_I\}$$

$$f_M : M = U_M$$
$$f_E : E = U_E$$
$$f_I : I = 2M + 3E + U_I$$



M: Exam Marks

E: Experience with coding

I: Internship funding

For causality need both the SCM and the graph

Product Decomposition Rule

Graphical models: Express joint distributions very efficiently

The joint distributions of the variables given by the product of conditional probability distributions:

$$P(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} P(x_i | pa_i)$$

where pa_i denote the parents of X_i .
(Discussed in later lectures in more detail). Example:

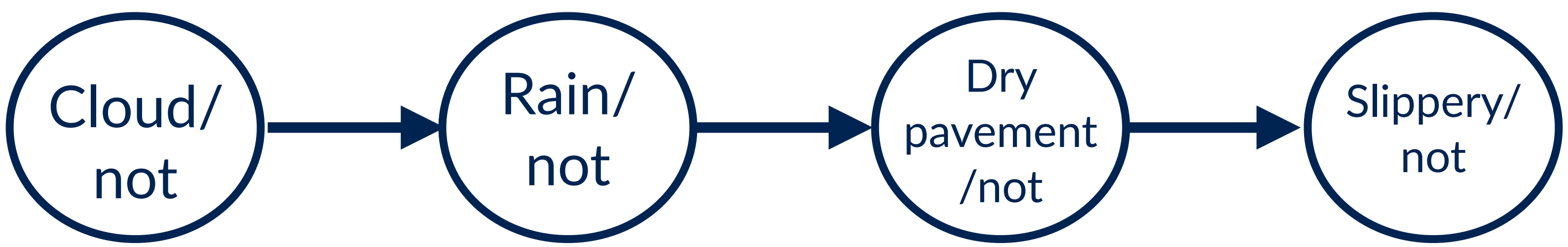
$$X \longrightarrow Y \longrightarrow Z$$

$$P(X = x, Y = y, Z = z) = P(X = x)P(Y = y|X = x)P(Z = z|Y = y)$$

Graph assumptions: High-dim estimation $\longrightarrow$ Few lower-dim probabilities
Graph simplifies the estimation problem and implies more precise estimators
(can draw the graph without necessarily needing the functional form)

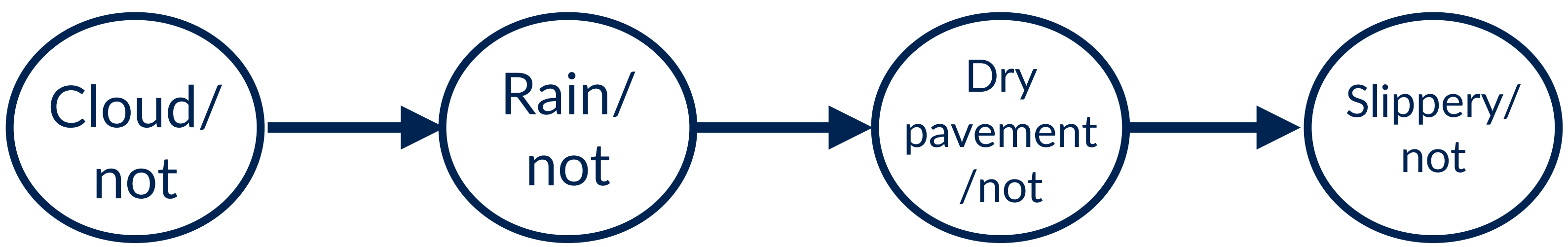
Product Decomposition Rule

p(clouds, no-rain, dry-pavement, slippery pavement) = ?



Product Decomposition Rule

p(clouds, no-rain, dry-pavement, slippery pavement) = ’5% or 10% or 15%?’

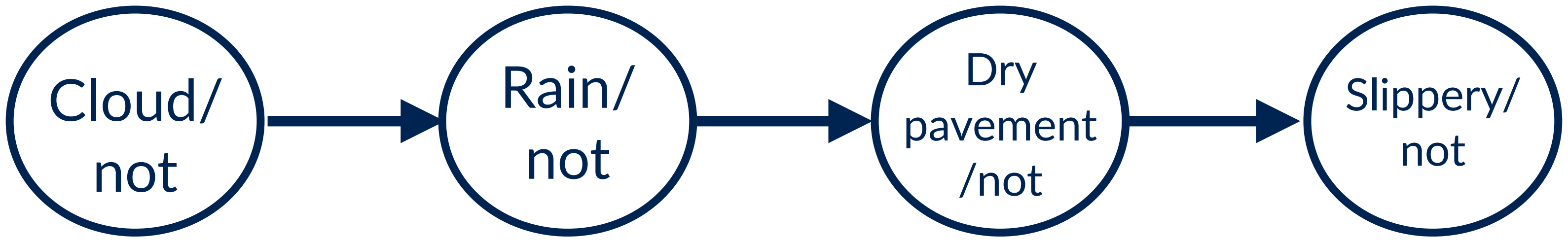


Product Decomposition Rule

p(clouds, no-rain, dry-pavement, slippery pavement) = ’5% or 10% or 15%?’

p(clouds)p(no rain | clouds)p(dry pavement | no rain) x
p(slippery pavement | dry pavement) ~

0.6 x 0.7 x 0.9 x 0.05 ~ 0.02

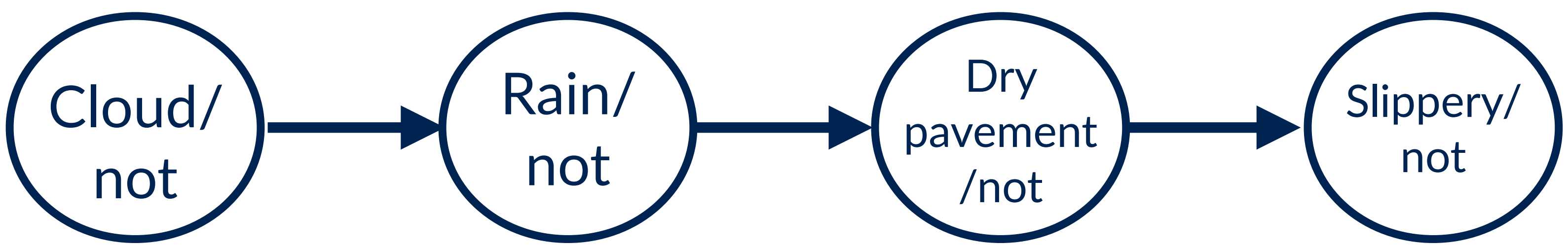


Product Decomposition Rule

p(clouds, no-rain, dry-pavement, slippery pavement) = ’5% or 10% or 15%?’

p(clouds)p(no rain | clouds)p(dry pavement | no rain) x
p(slippery pavement | dry pavement) ~

0.6 x 0.7 x 0.9 x 0.05 ~ 0.02

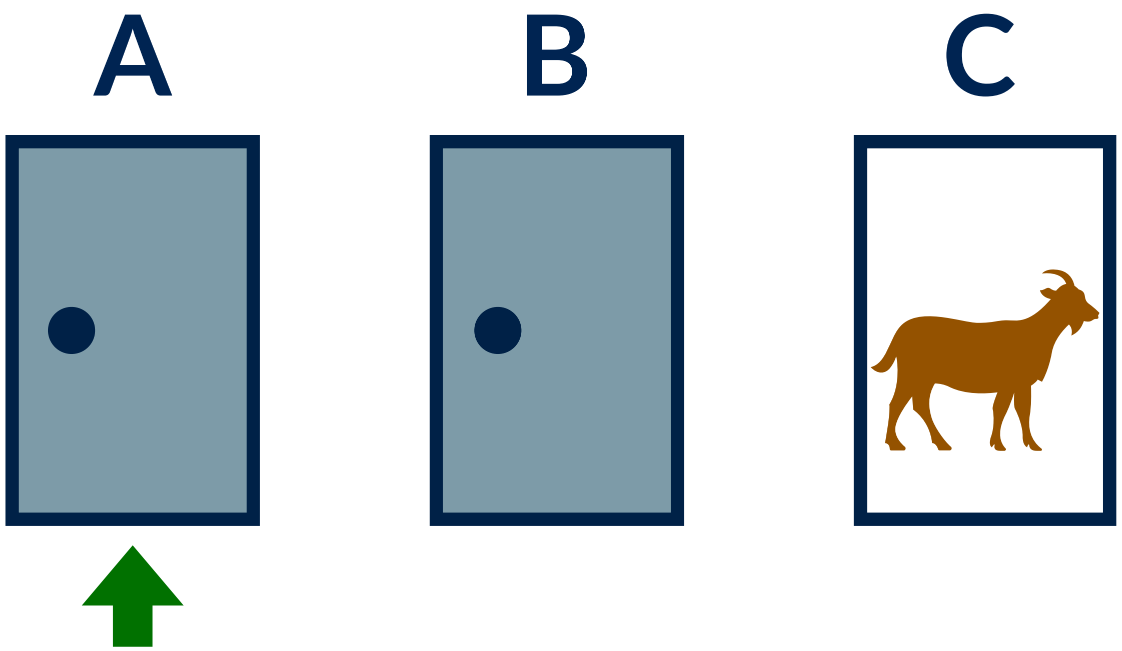


Combinations: $2^4 - 1 = 15$
Suppose we have 45 data points of these 4 observations
Approx, 45/15 = 3 observations per outcome, some may get 2 or 1 or empty.

Need far more data to estimate the joint distribution as compared to each of the conditional distributions.

Causal Inference in Statistics, Pearl (2016)

SCM for the Monty Hall Problem



X = Door chosen by player

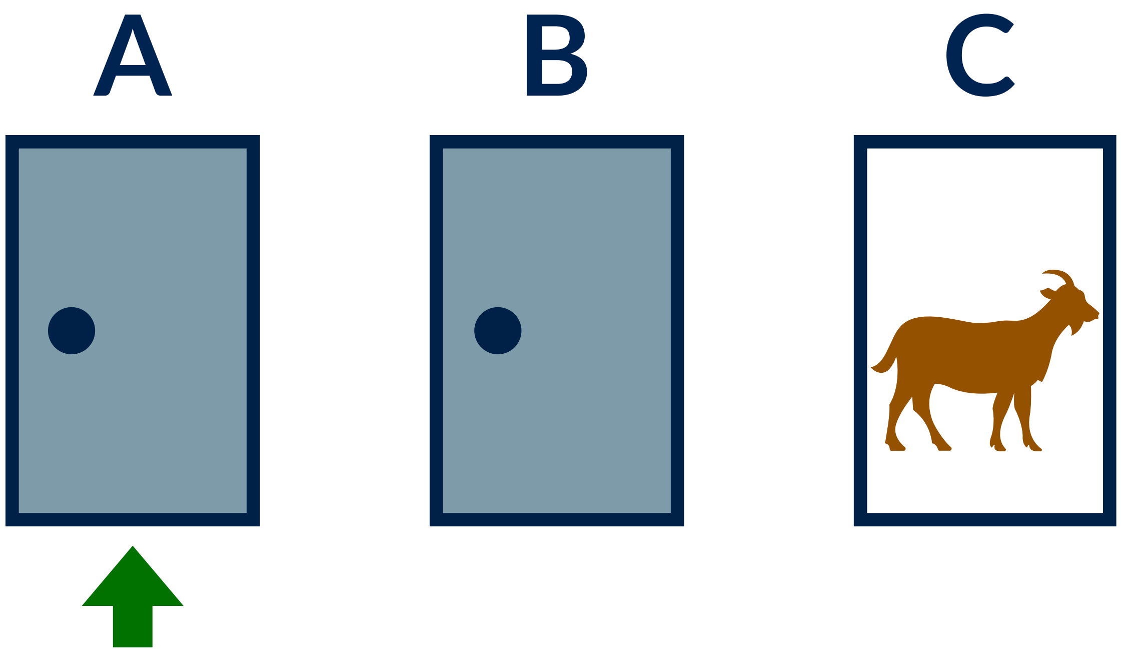
Y = Door hiding the car

Z = Door opened by host

The player can choose any door with p = 1/3

The car can be behind any door with p = 1/3

SCM for the Monty Hall Problem



X = Door chosen by player

Y = Door hiding the car

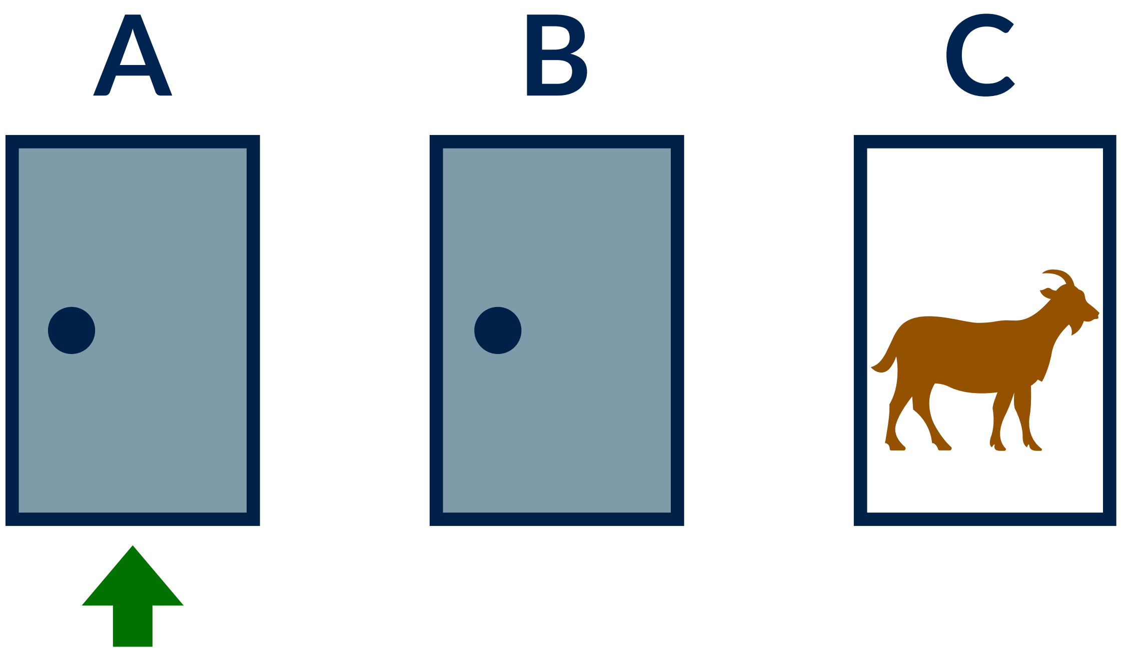
Z = Door opened by host

Z needs to use 2 pieces of information:

(1) not be the door chosen by player

(2) not be the door that hides the car

SCM for the Monty Hall Problem



X = Door chosen by player

Y = Door hiding the car

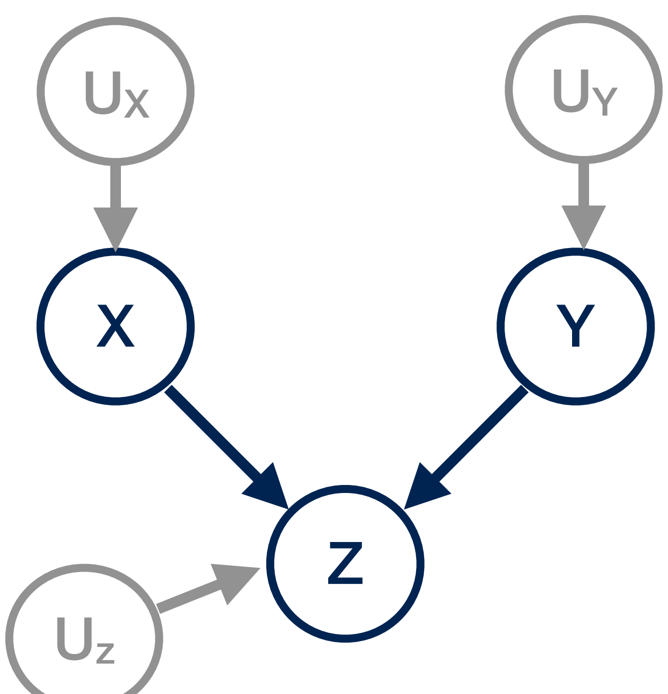
Z = Door opened by host

Z needs to use 2 pieces of information:

(1) not be the door chosen by player

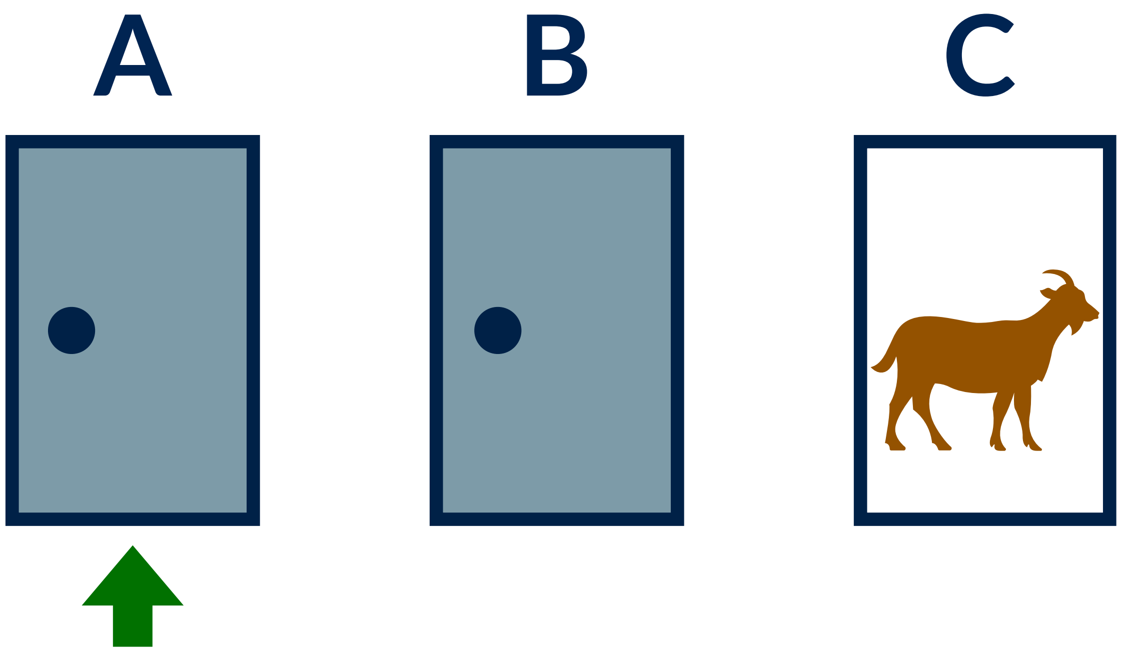
(2) not be the door that hides the car

$$V = \{X, Y, Z\}$$
$$U = \{U_X, U_Y, U_Z\}$$
$$F = \{f\}$$



$$X = U_X$$
$$Y = U_Y$$
$$Z = f(X, Y) + U_Z$$

SCM for the Monty Hall Problem

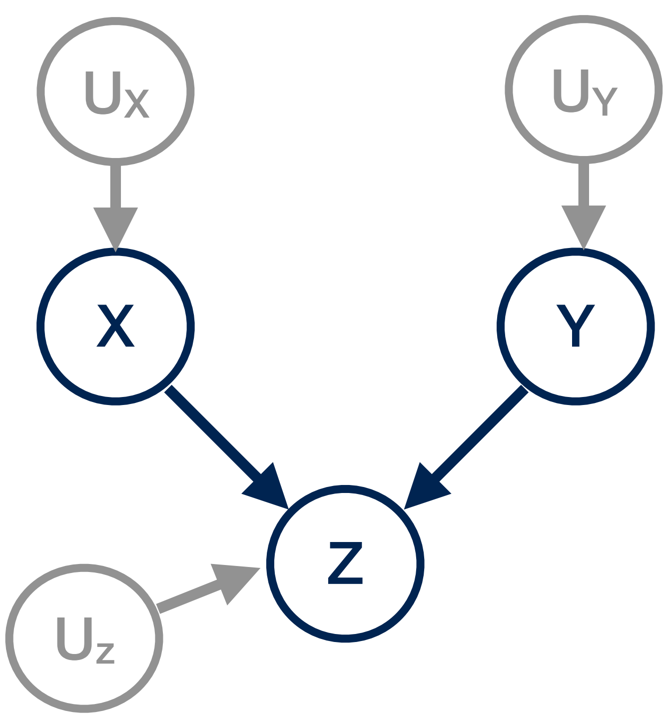


X = Door chosen by player

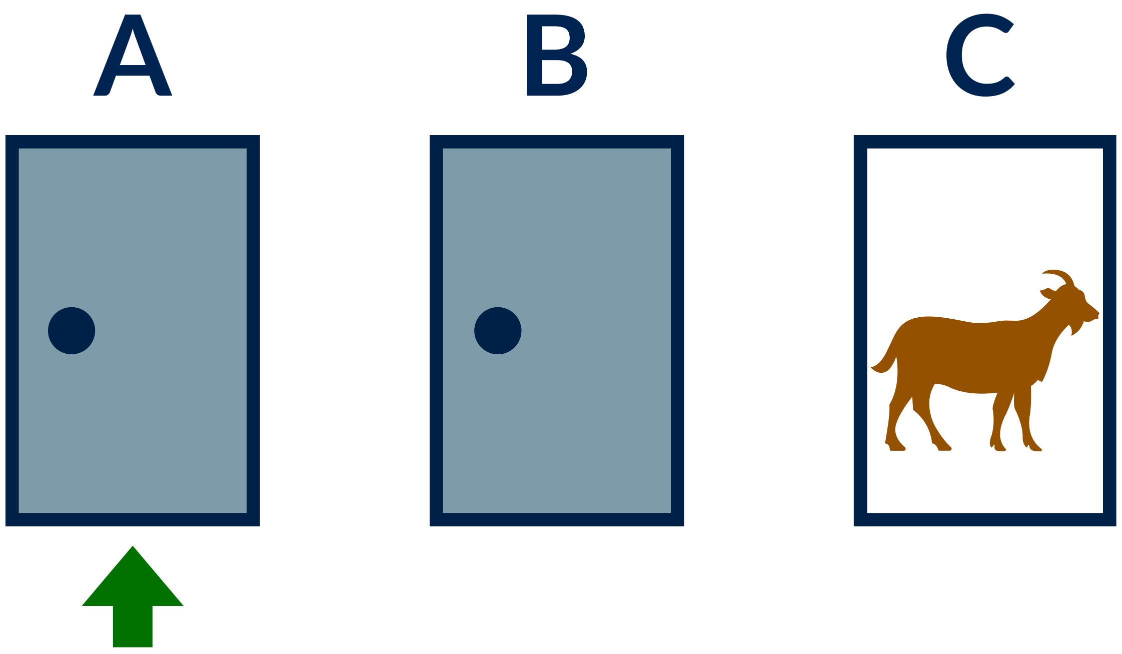
Y = Door hiding the car

Z = Door opened by host

The joint probability: $P(X,Y,Z) = P(Z|X,Y)P(Y)P(X)$



SCM for the Monty Hall Problem

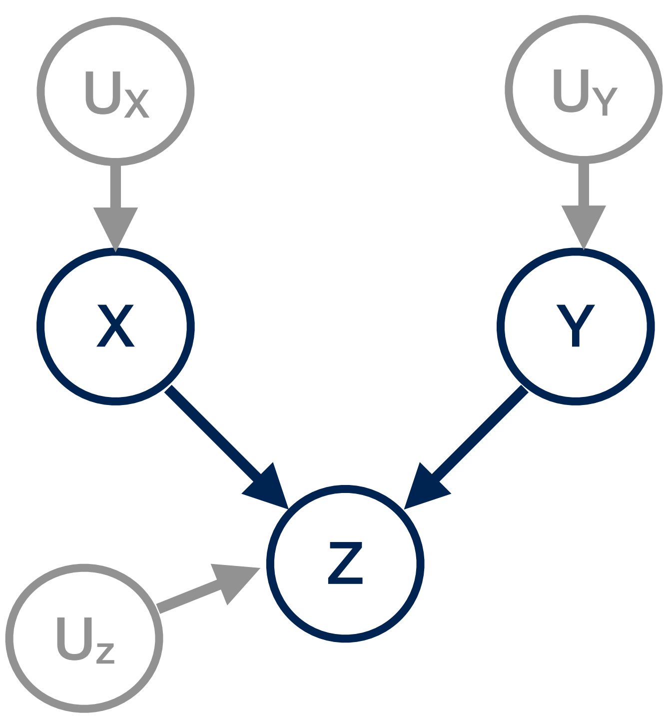


X = Door chosen by player

Y = Door hiding the car

Z = Door opened by host

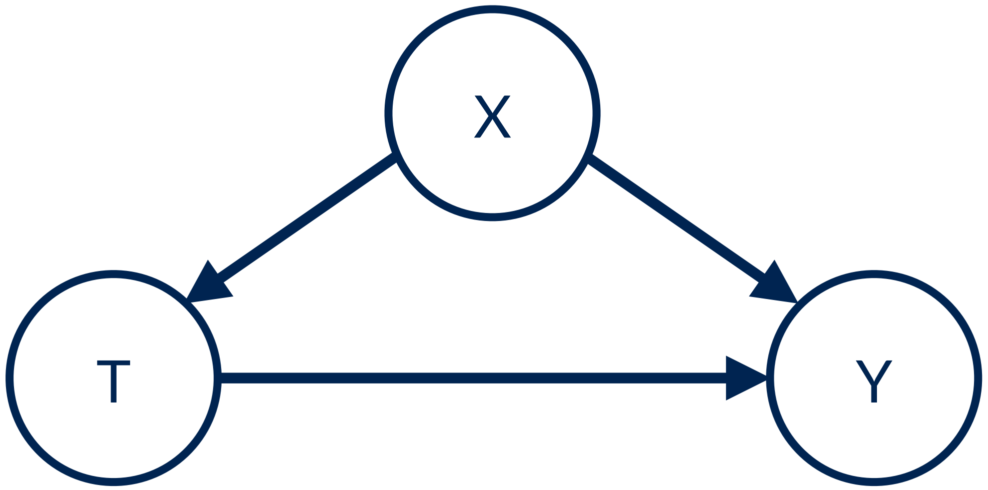
The joint probability: $P(X,Y,Z) = P(Z|X,Y)\overset{1/3}{P(Y)}\overset{1/3}{P(X)}$



$$P(Z|X,Y) = \begin{cases} 0.5 \text{ for } x = y \neq z \\ 1 \text{ for } x \neq y \neq z \\ 0 \text{ for } z = x \text{ or } z = y \end{cases}$$

Conventions

- Variable to be manipulated: **treatment (T)**, e.g. medication
- Variable we observe as response: **outcome (Y)**, e.g. success/failure of medication
- Other observable variables that can affect treatment and outcome causally and we wish to correct for: **confounders (X)**, e.g. age, sex, socio-economic status, ...
- Unobservable confounder **(U)**

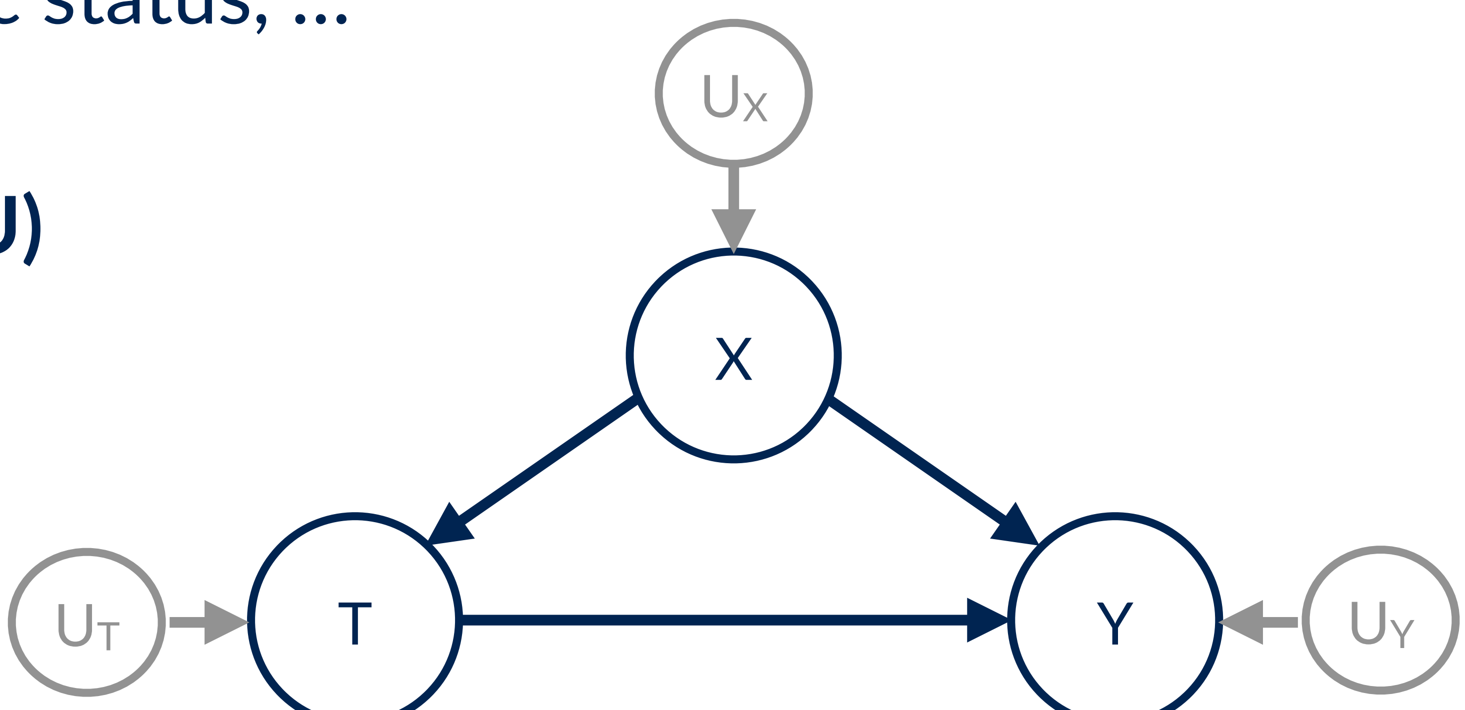


Conventions

- Variable to be manipulated: **treatment (T)**, e.g. medication
- Variable we observe as response: **outcome (Y)**, e.g. success/failure of medication
- Other observable variables that can affect treatment and outcome causally and we wish to correct for: **confounders (X)**, e.g. age, sex, socio-economic status, …
- Unobservable confounder **(U)**

For simplicity drop U_i's from graphs **if**:

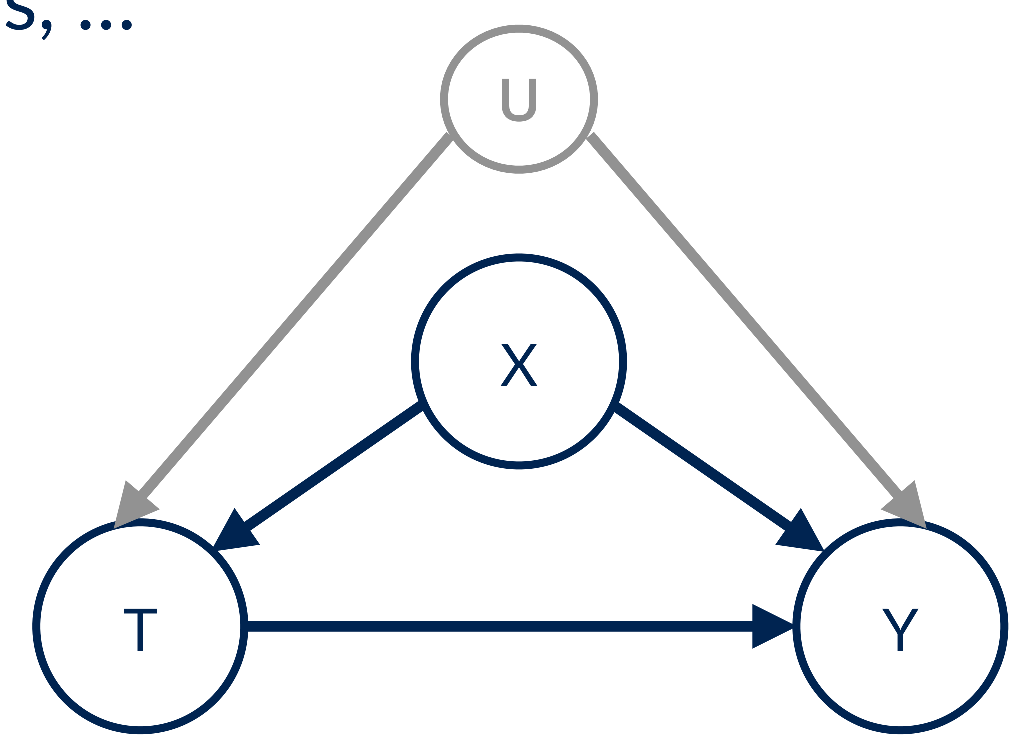
$$U_T \perp\!\!\!\perp U_X \perp\!\!\!\perp U_Y$$



Conventions

- Variable to be manipulated: **treatment (T)**, e.g. medication
- Variable we observe as response: **outcome (Y)**, e.g. success/failure of medication
- Other observable variables that can affect treatment and outcome causally and we wish to correct for: **confounders (X)**, e.g. age, sex, socio-economic status, …
- Unobservable confounder **(U)**

A different story when Us are dependent or a confounder: See IV



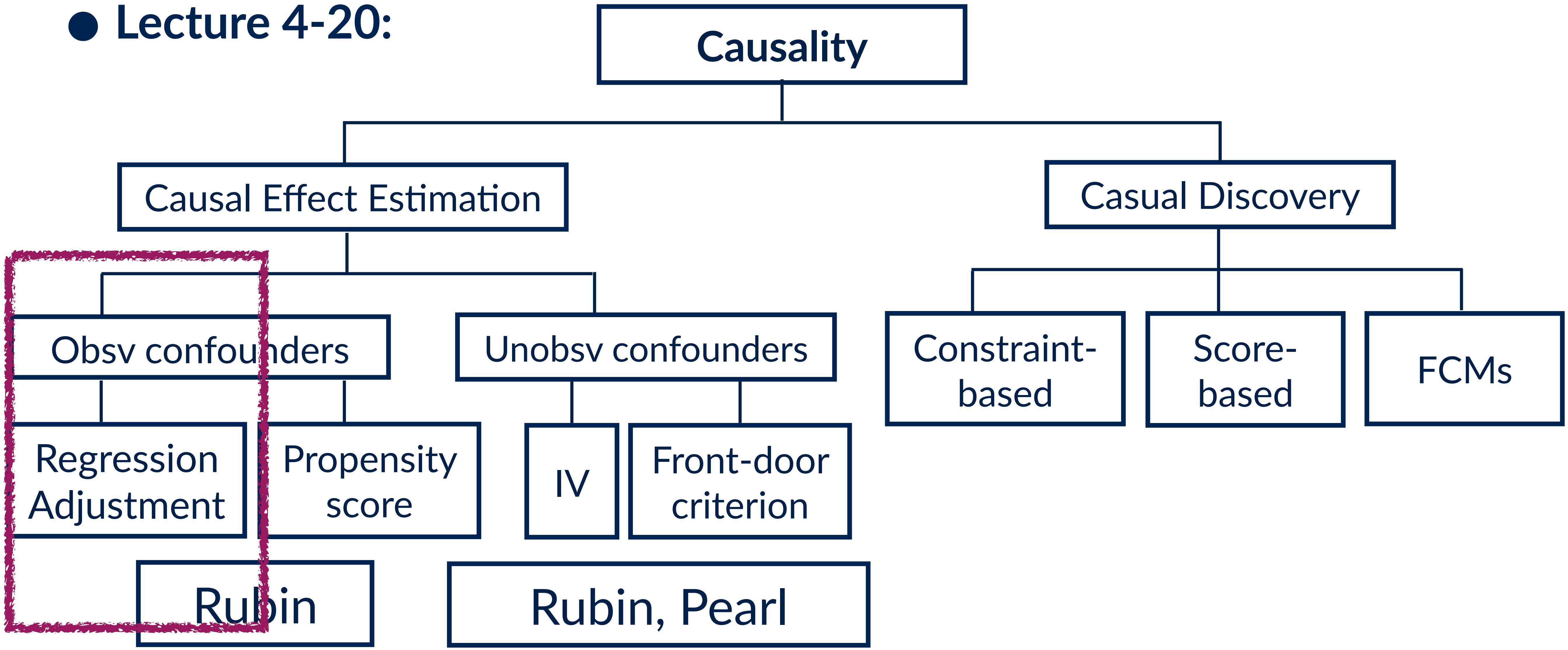
Causal Identification vs Estimation

Causal Identification problem: Is it possible to express a causal quantity in terms of the probability distribution of the observed data, and if so, how?

Estimation problem: How to estimate the functional relationship between treatment T and outcome Y, given other variables X in the system.
For example: $\mathbb{E}[Y|T,X] = f(T,X)$

Overview of the course

- **Lecture 1**: Introduction & Motivation, why do we care about causality? Why deriving causality from observational data is non-trivial.
- **Lecture 2:** Recap of probability theory, variables, events, conditional probabilities, independence, law of total probability, Bayes' rule
- **Lecture 3**: Recap of regression, multiple regression, graphs, SCM
- **Lecture 4-20:**



Methods for Causal Inference
Lecture 3: Basics of probability

Ava Khamseh

School of Informatics
2023-2024