

Why and What If? Causal Inference for Everyone

Bruno Gonçalves

www.data4sci.com/newsletter

<https://github.com/DataForScience/PyData2020>



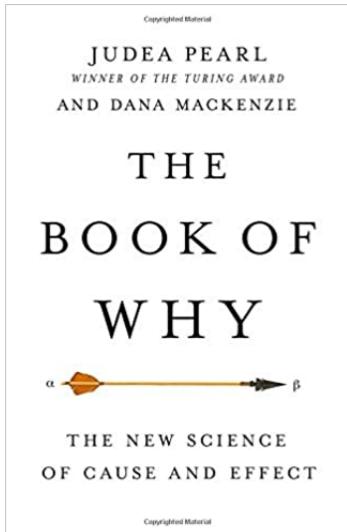
Who Am I?



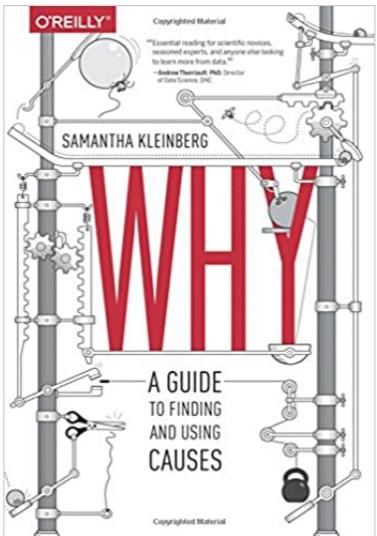
- **Name:** Bruno Gonçalves
- **Work:** Senior Data Scientist
- **Background:** Physics (PhD) and Computer Science (MS)
- **Experience:** 10+ years using large scale datasets to analyze individual human behavior.
- **Website:** www.data4sci.com
- **Twitter:** @data4sci
- **Email:** bgoncalves@data4sci.com

Bibliography

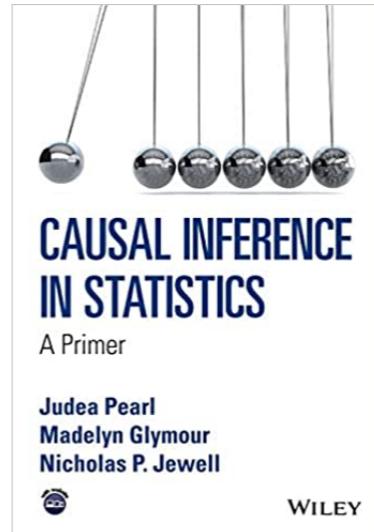
<https://github.com/DataForScience/PyData2020>



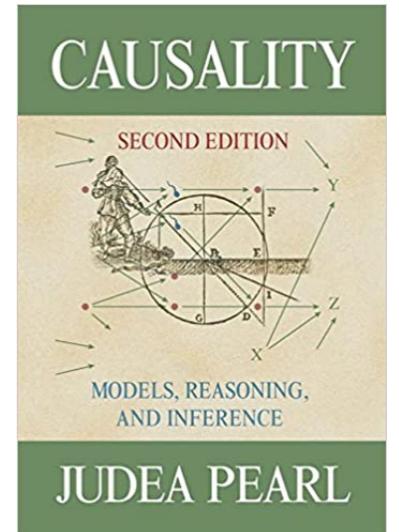
<https://amzn.to/34bOF8N>



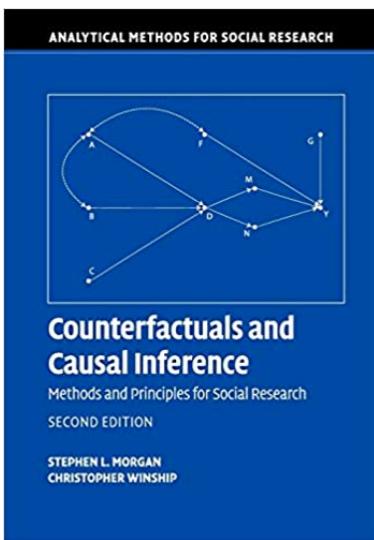
<https://amzn.to/323uK9n>



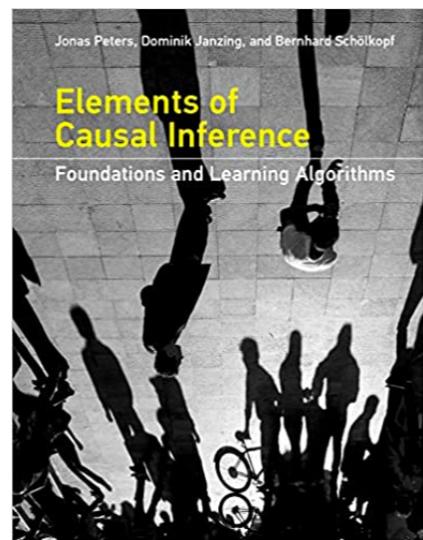
<https://amzn.to/3iS73aP>



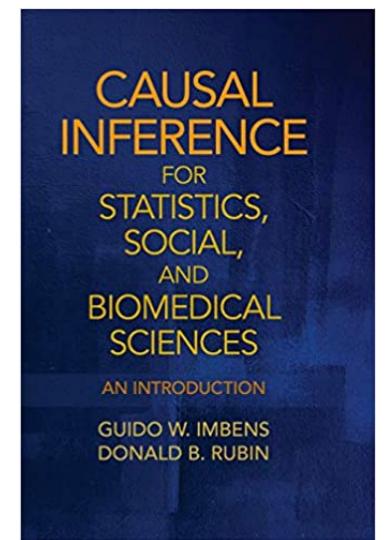
<https://amzn.to/3iW2l6m>



<https://amzn.to/34clQZV>



<https://amzn.to/34bvmMQ>



<https://amzn.to/34dlwJ5>



Table of Contents

1. Graphical Models
2. Interventions
3. Counterfactuals



1. Graphical Models

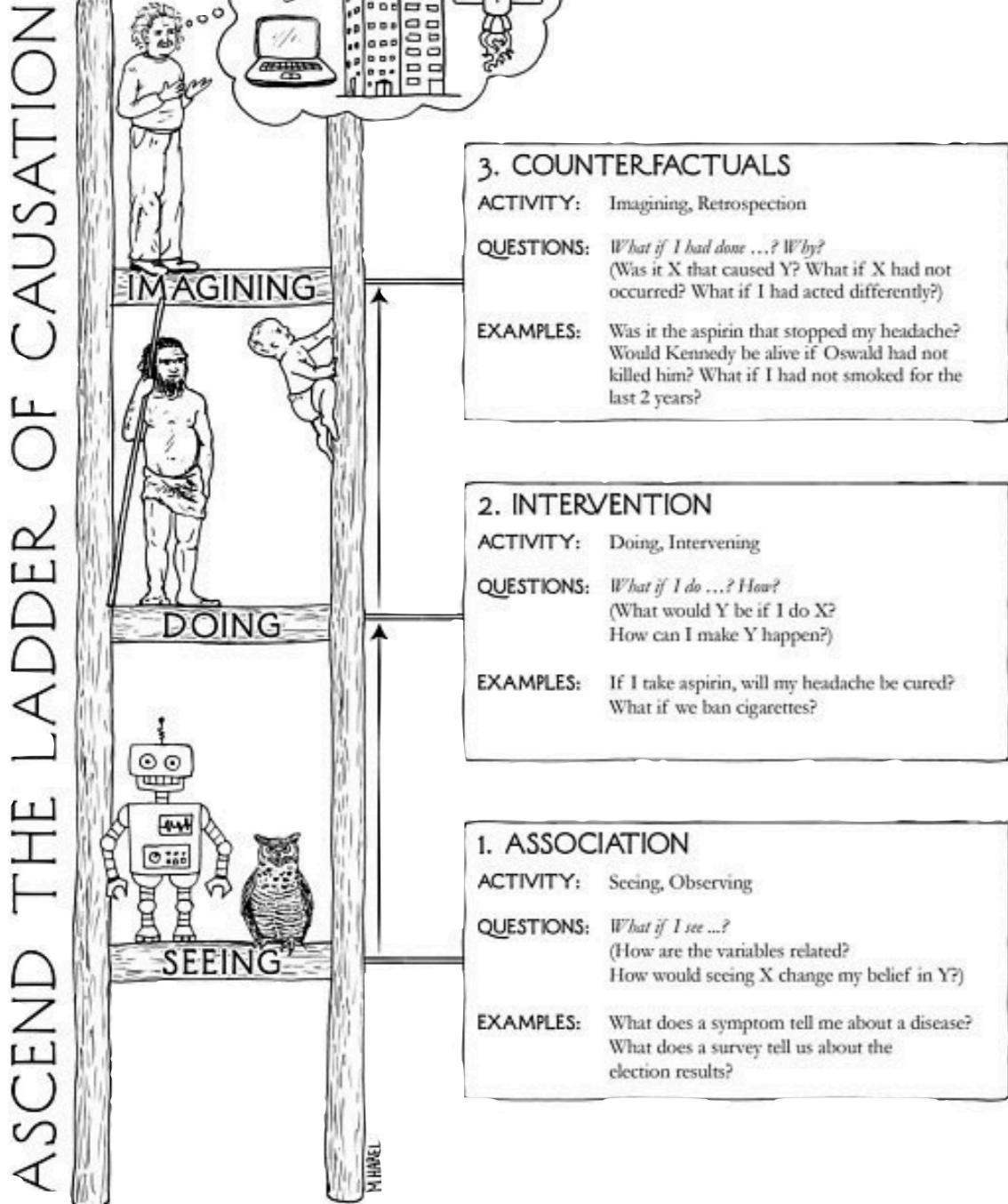
Causation

"More has been learned about **Causal Inference** in the last few decades than the sum total of everything that had been learned about it in all prior recorded history" (G. King)

- **What Causes** produced a given effect?
- **How** can we **Intervene** to produce a desired effect?
- **What would have** happened if the world was different?

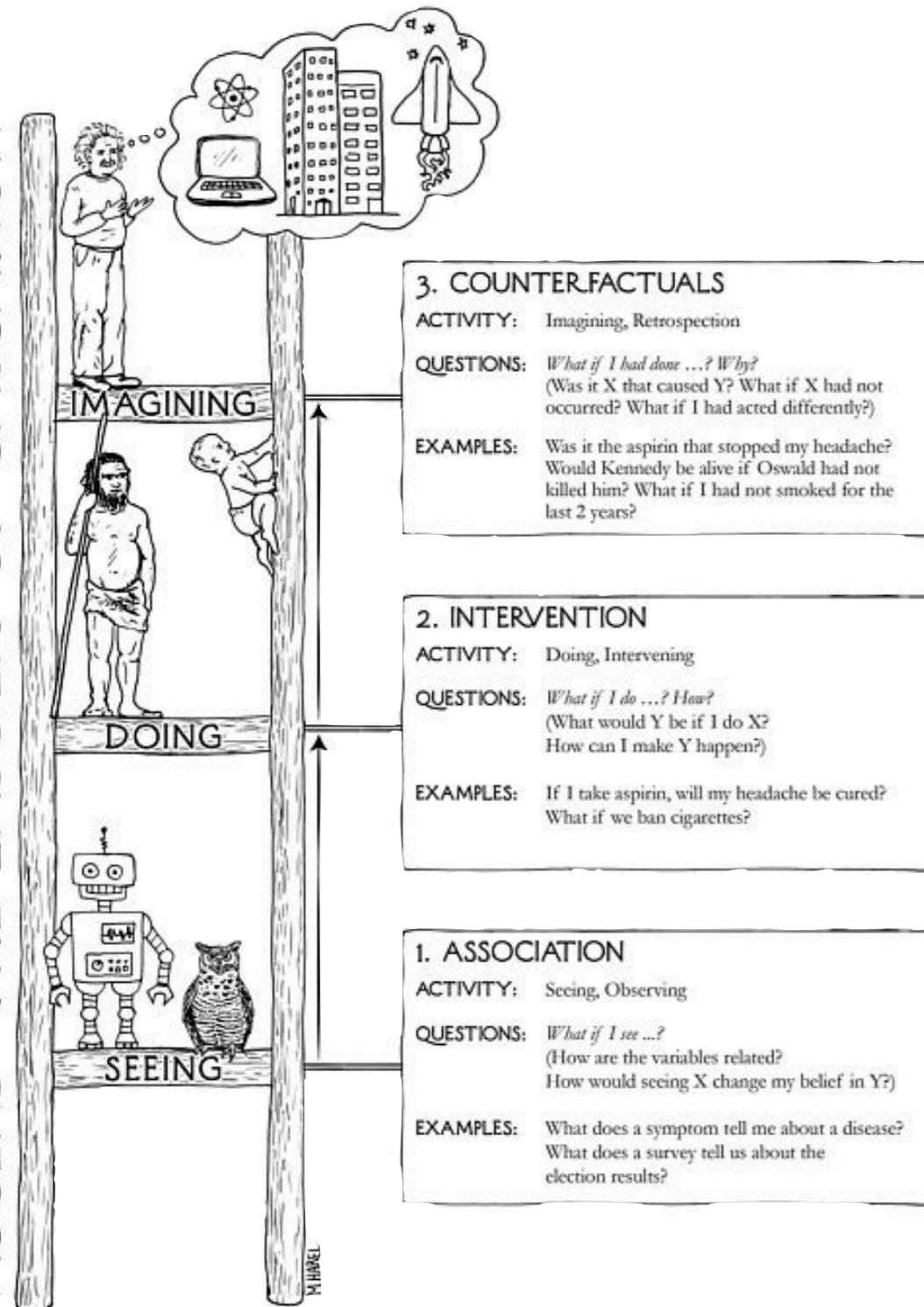
Ladder of Causality

<https://amzn.to/34bOF8N>



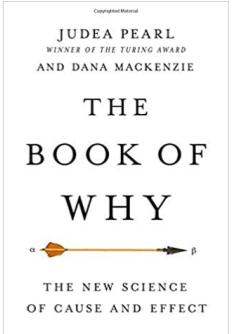
ASCEND THE LADDER OF CAUSATION

Ladder of Causality



The Three Layer Causal Hierarchy

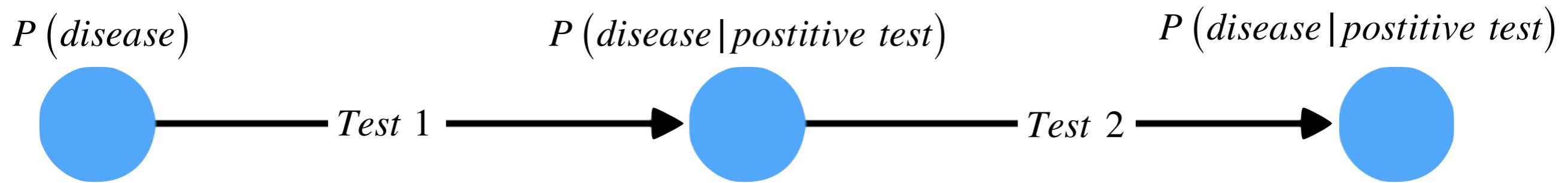
Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y x)$	Seeing	What is? How would seeing X change my belief in Y ?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y do(x), z)$	Doing Intervening	What if? What if I do X ?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x x', y')$	Imagining, Retrospection	Why? Was it X that caused Y ? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?



<https://amzn.to/34bOF8N>

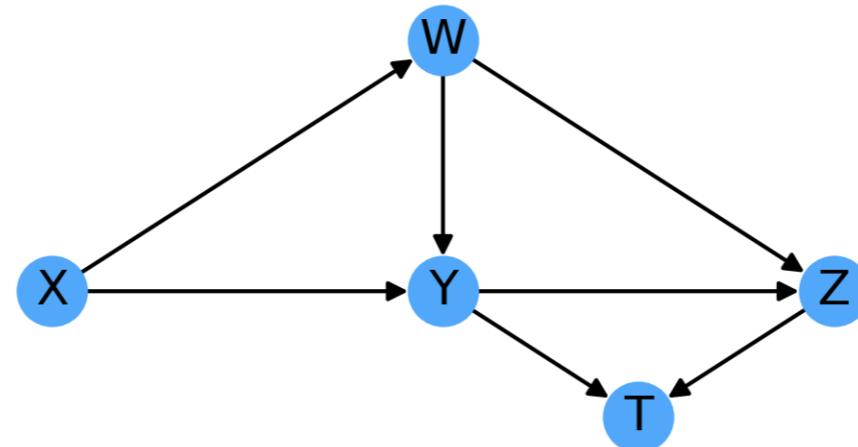
Graphical Models

- Our first Graphical Model:



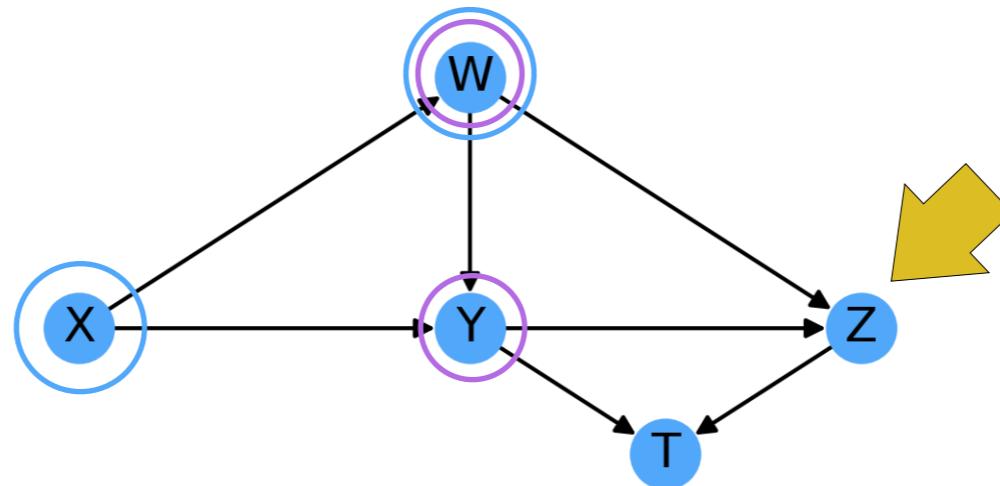
- Graphs are constituted by **Nodes** (subcomponents) interconnected by **edges** (connections, relationships, etc)
- In **Causal Models**:
 - **Nodes** represent variables of interest
 - **Edges** are **Directed** and imply a **causal relationship**
 - The **Graph** does not contain cycles: it's a **Directed Acyclical Graph**
- **Fundamental Assumption:** The value of a variable can **only depend** on the values of their **parents** that “point” to it (**incoming edges**) and can only influence the **children** they “point” to (**outgoing edges**)

Graphical Models



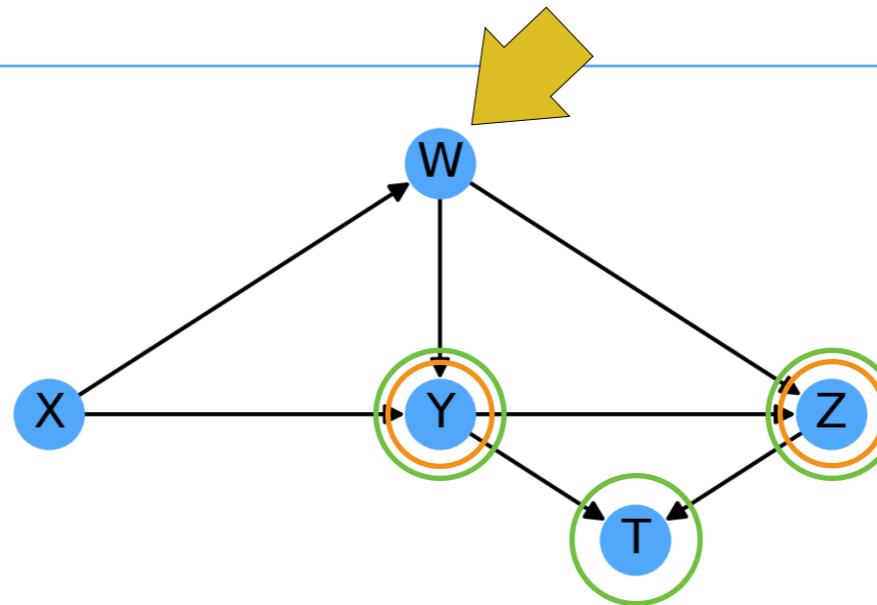
- 5 nodes: T, W, X, Y, Z
- 7 edges: [(X, W), (X, Y), (W, Y), (W, Z), (Y, Z), (Y, T), (Z, T)]

Graphical Models



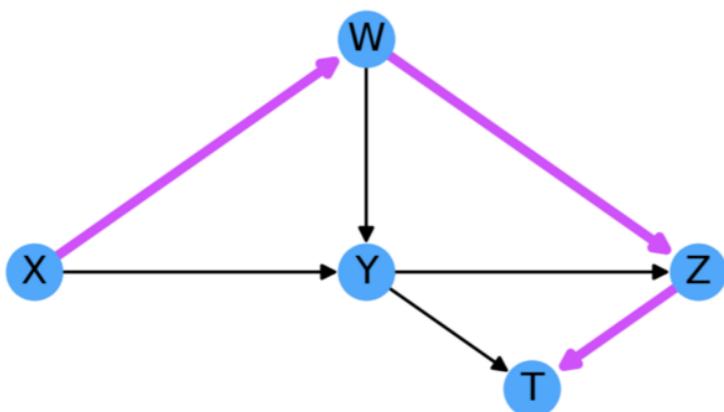
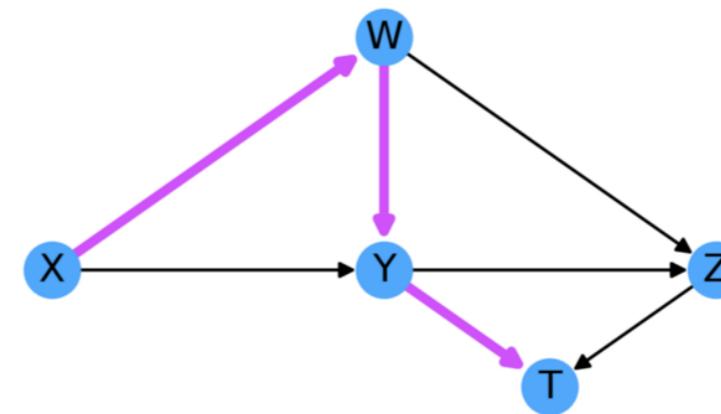
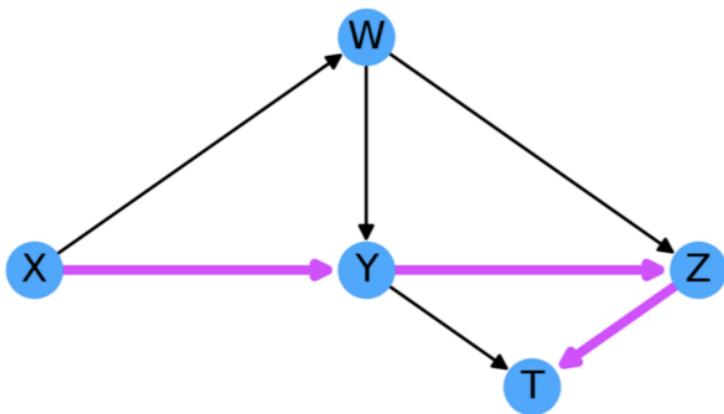
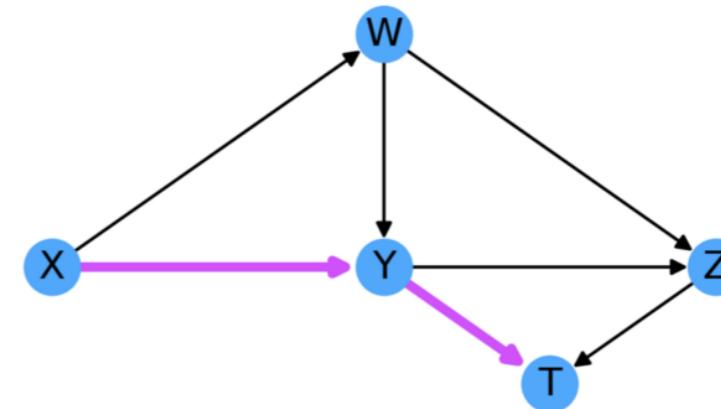
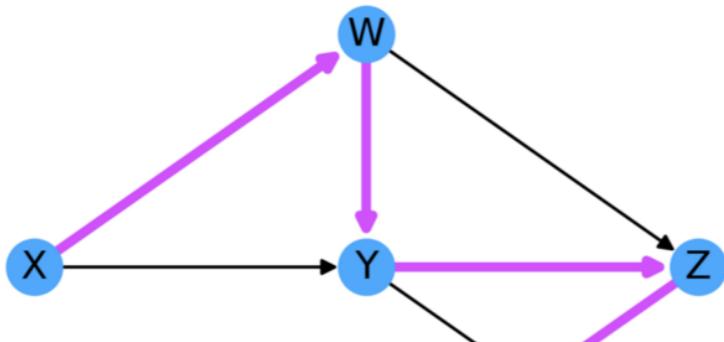
- 5 nodes: T, W, X, Y, Z
 - 7 edges: [(X, W), (X, Y), (W, Y), (W, Z), (Y, Z), (Y, T), (Z, T)]
 - Z has two parents, Y and W and three ancestors, Y, W and X.

Graphical Models



- 5 nodes: T, W, X, Y, Z
- 7 edges: [(X, W), (X, Y), (W, Y), (W, Z), (Y, Z), (Y, T), (Z, T)]
- Z has two **parents**, Y and W and three **ancestors**, Y, W and X.
- W has two **children**, Y and Z and three **descendants**, Y, Z, and T.

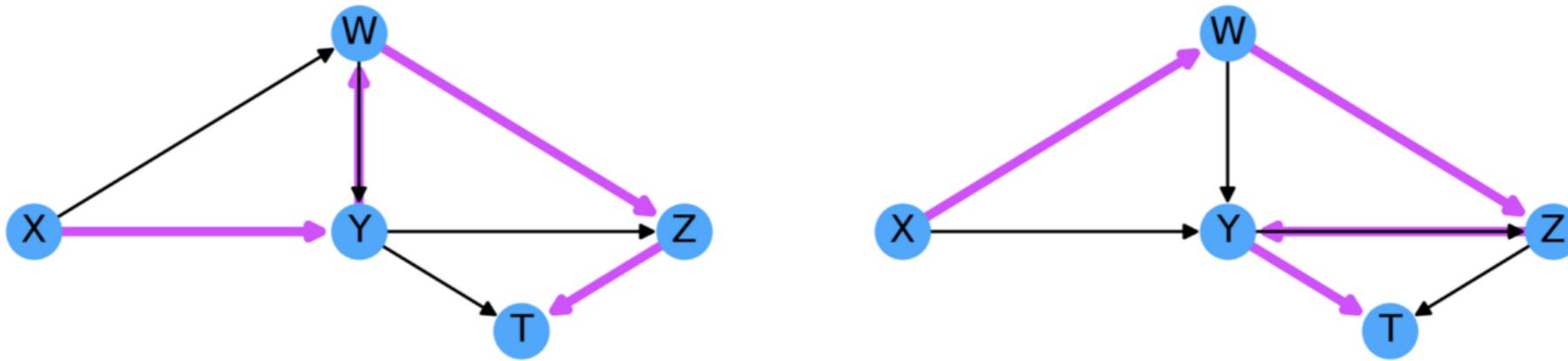
Graphical Models



Directed **paths** represent ways
in which one variable can
directly influence another

Graphical Models

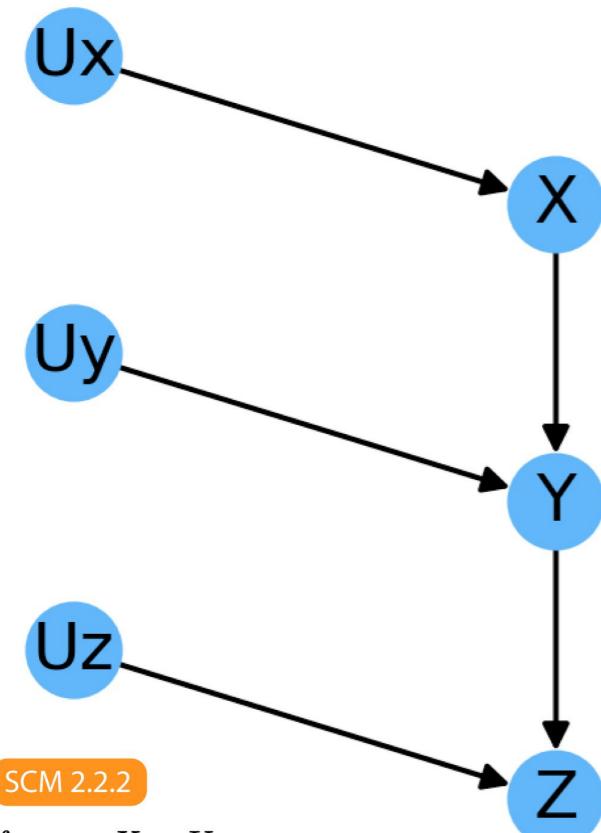
- Under certain circumstances, influence can also travel “**backwards**” along an edge:



Structural Causal Models

- A **Structural Causal Model (SCM)** is a DAGs together with a set of functions F fully specifying the form of the relationships
- The same DAG can describe a large number of SCMs
- A **Structural Causal Model (SCM)** consists of a set of **Endogenous (V)** and a set of **Exogenous (U)** variables connected through a **Directed Acyclical Graphs (DAG)** by a set of **Functions (F)** that determine the values of the variables in V based on the values of the variables in U .
- **Functions (F)** can be deterministic, stochastic or logical

Structural Causal Models



Exogenous variables (**U**) are often treated as error terms

$$U = \{U_X, U_Y, U_Z\}$$

$$V = \{X, Y, Z\}$$

$$F = \{f_X, f_Y, f_Z\}$$

SCM 2.2.1

$$f_X : X = U_X$$

$$f_Y : Y = \frac{X}{3} + U_Y$$

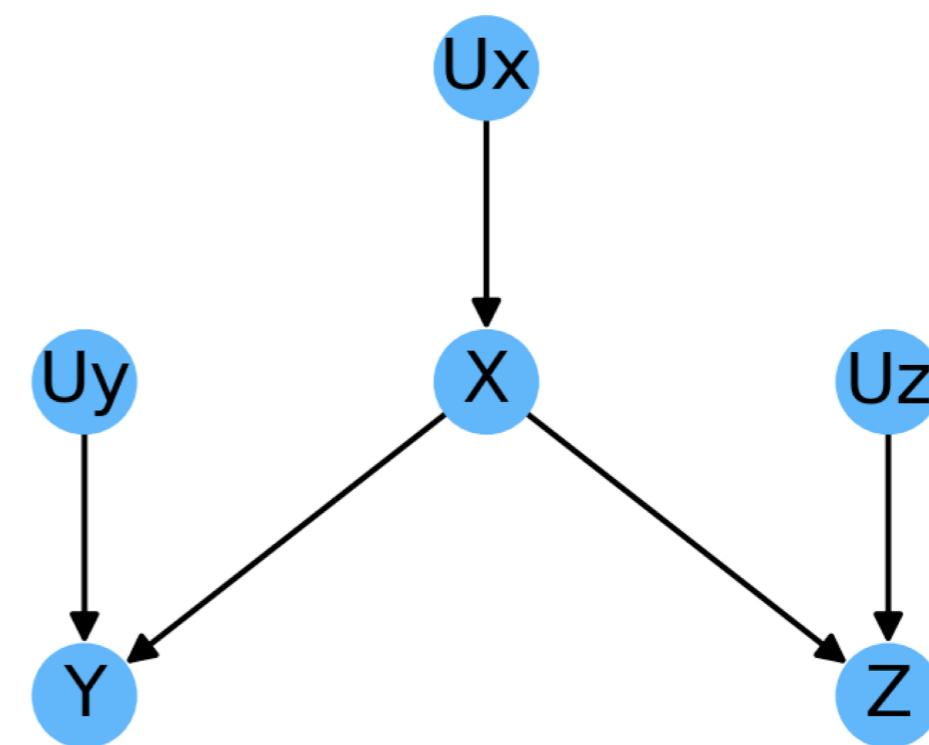
$$f_Z : Z = \frac{Y}{16} + U_Z$$

SCM 2.2.3

$$f_X : X = U_X$$

$$f_Y : Y = 84 - x + U_Y$$

$$f_Z : Z = \frac{100}{y} + U_Z$$



$$U = \{U_X, U_Y, U_Z\}$$

$$V = \{X, Y, Z\}$$

$$F = \{f_X, f_Y, f_Z\}$$

SCM 2.2.5

$$f_X : X = U_X$$

$$f_Y : Y = 4x + U_Y$$

$$f_Z : Z = \frac{x}{10} + U_Z$$

Structural Causal Models

- When given the full SCM specification, we can easily generate fake data to play with. We'll often do this to explore some ideas.
- **Rule of Product decomposition:** For any model whose graph is acyclic, the joint distribution of the variables in the model is given by the product of the conditional distributions $P(\text{child}|\text{parents})$ over all the "families" in the graph, or, mathematically:

$$P(X_1, X_2, \dots, X_n) = \prod_i P(X_i | \text{parents}_i)$$

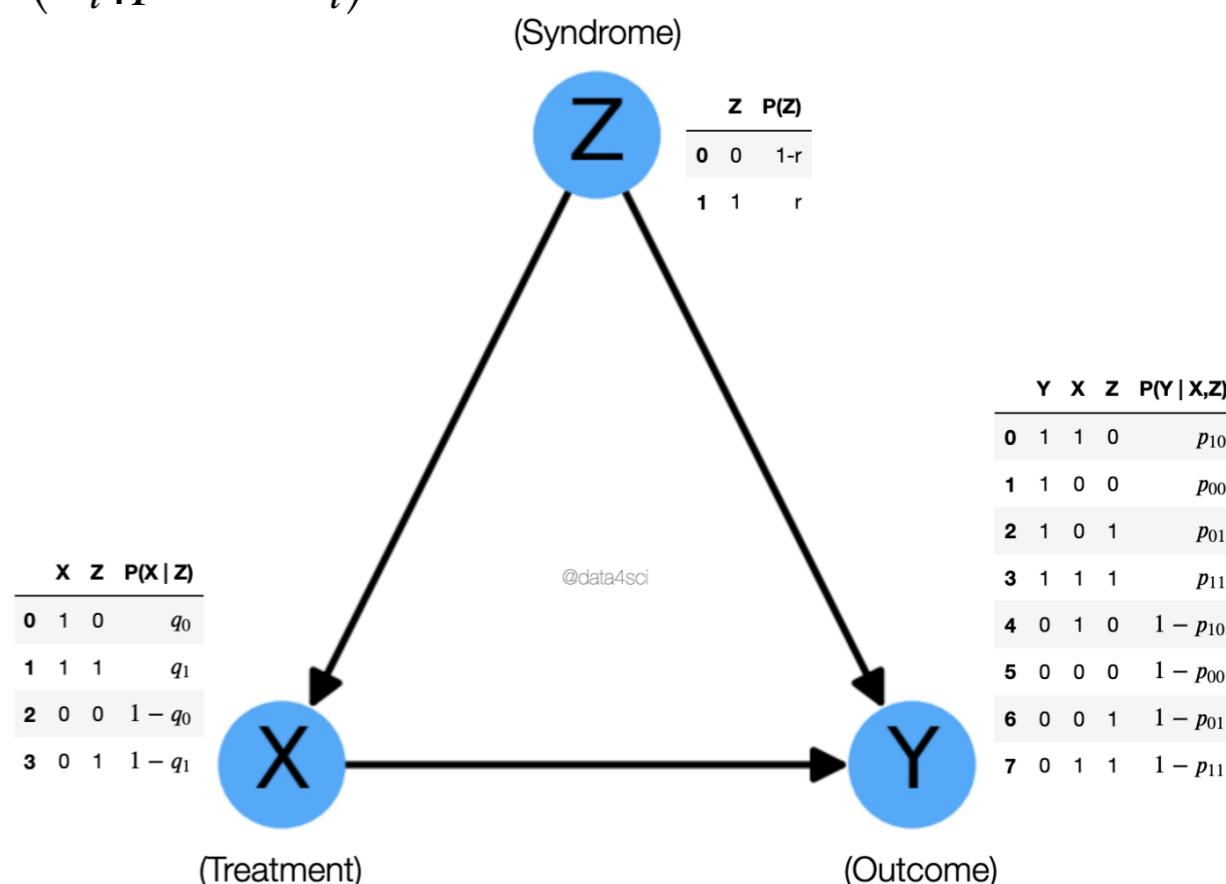
Structural Causal Models

- When given the full SCM specification, we can easily generate fake data to play with. We'll often do this to explore some ideas.
- **Rule of Product decomposition:** For any model whose graph is acyclic, the joint distribution of the variables in the model is given by the product of the conditional distributions $P(\text{child}|\text{parents})$ over all the "families" in the graph, or, mathematically:

$$P(X_1, X_2, \dots, X_n) = \prod_i P(X_i|\text{parents}_i)$$

- So, if given this SCM:

Probability tables fully specify the model.



Structural Causal Models

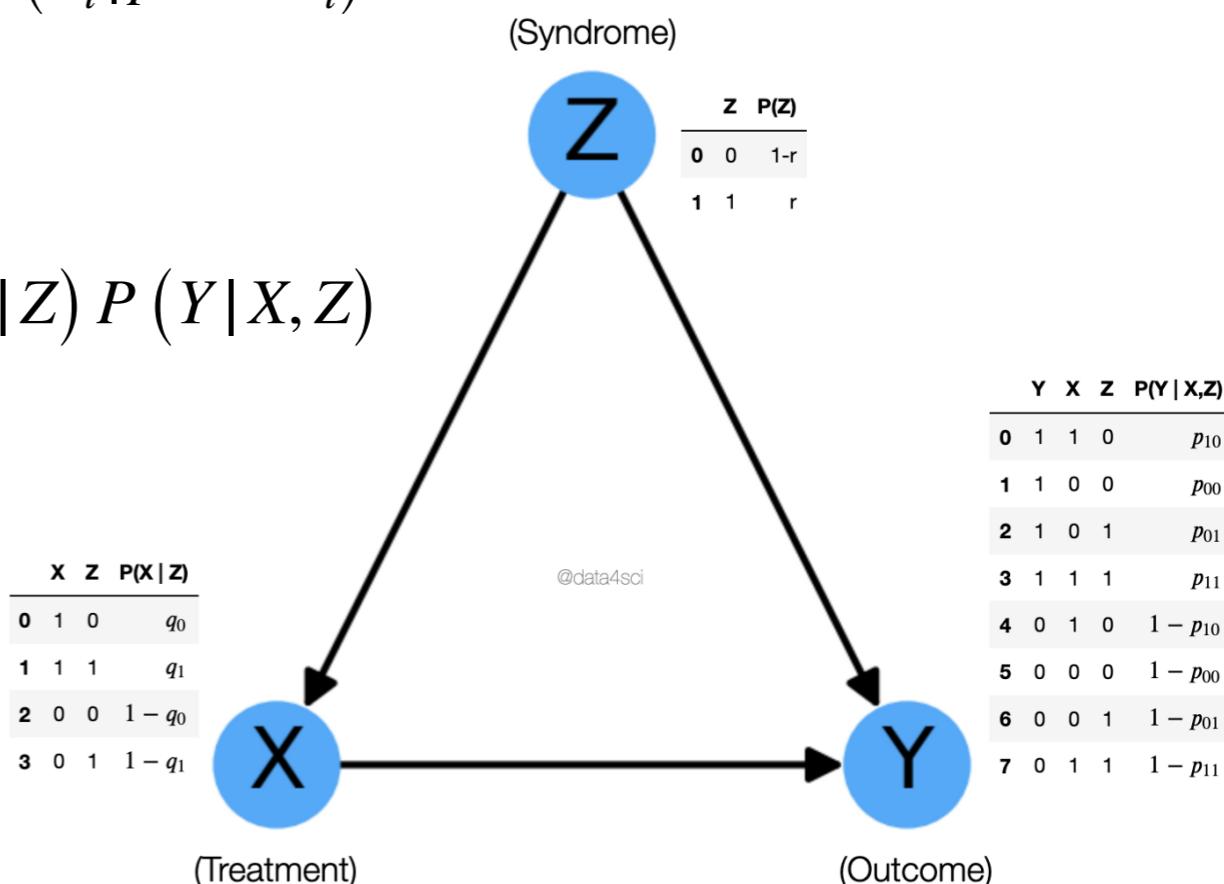
- When given the full SCM specification, we can easily generate fake data to play with. We'll often do this to explore some ideas.
- **Rule of Product decomposition:** For any model whose graph is acyclic, the joint distribution of the variables in the model is given by the product of the conditional distributions $P(\text{child}|\text{parents})$ over all the "families" in the graph, or, mathematically:

$$P(X_1, X_2, \dots, X_n) = \prod_i P(X_i|\text{parents}_i)$$

- So, if given this SCM, we can write:

$$P(X, Y, Z) = P(Z) P(X|Z) P(Y|X, Z)$$

Probability tables fully specify
the model.



Structural Causal Models

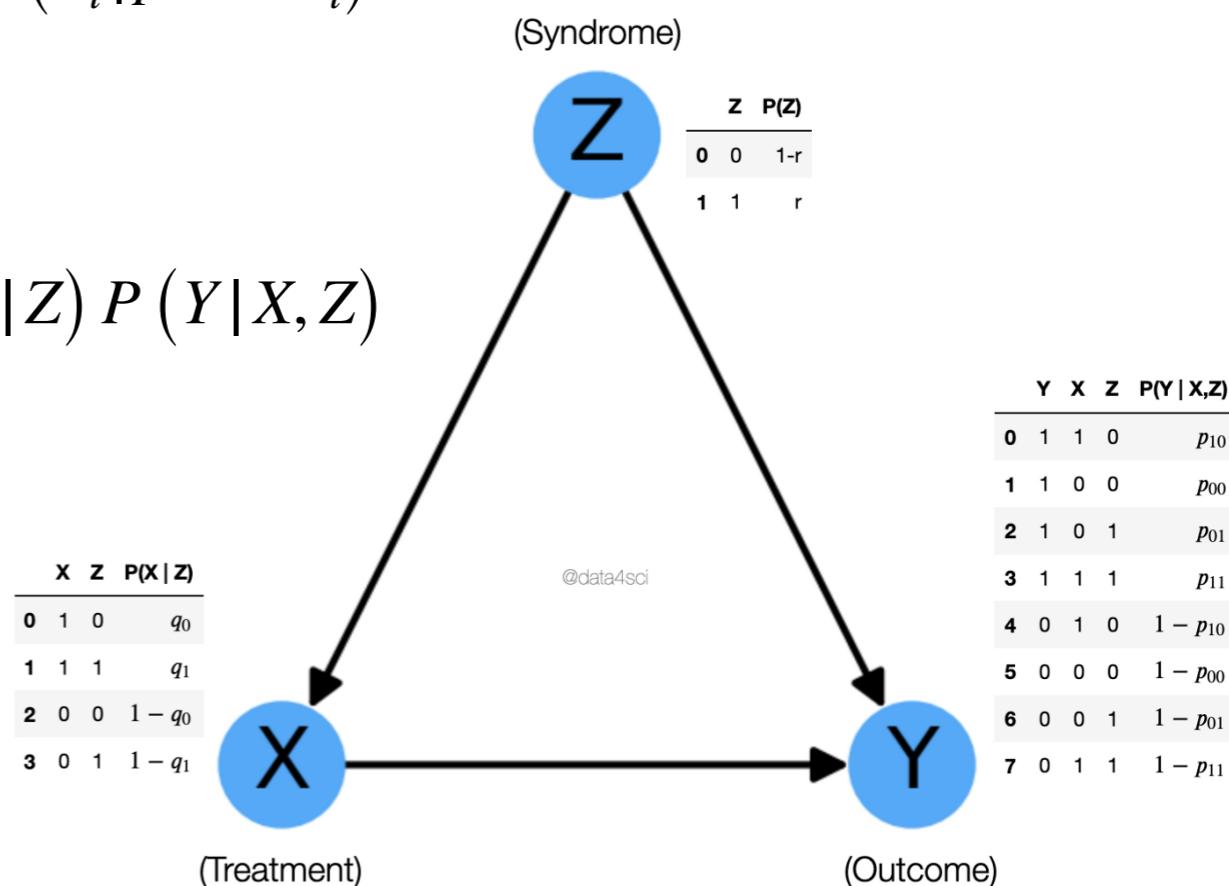
- When given the full SCM specification, we can easily generate fake data to play with. We'll often do this to explore some ideas.
- **Rule of Product decomposition:** For any model whose graph is acyclic, the joint distribution of the variables in the model is given by the product of the conditional distributions $P(\text{child}|\text{parents})$ over all the "families" in the graph, or, mathematically:

$$P(X_1, X_2, \dots, X_n) = \prod_i P(X_i|\text{parents}_i)$$

- So, if given this SCM, we can write:

$$P(X, Y, Z) = P(Z) P(X|Z) P(Y|X, Z)$$

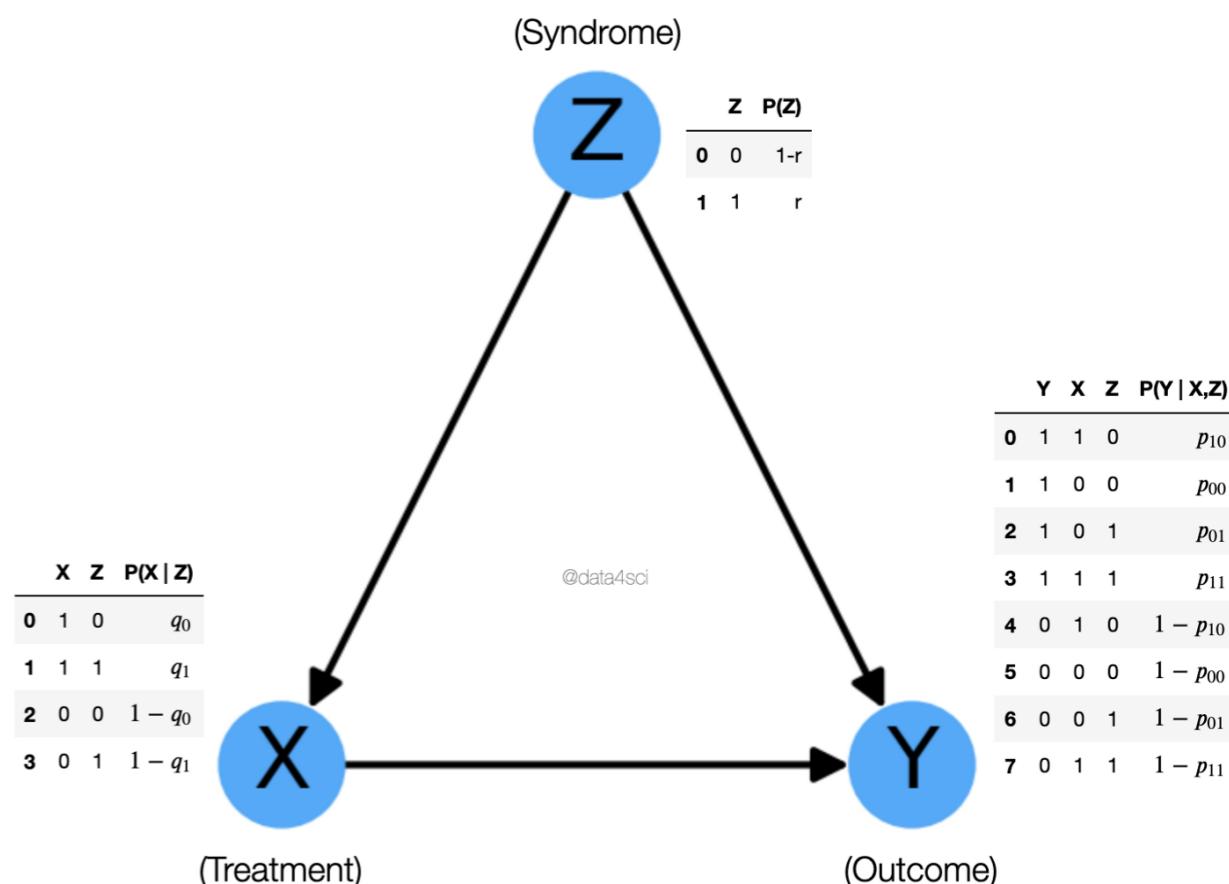
- And this joint probability distribution is sufficient to answer any question we might have about the behavior of the system



Structural Causal Models

- When analyzing the behavior of different graphical models we will make use of several rules.
The most fundamental of one is:

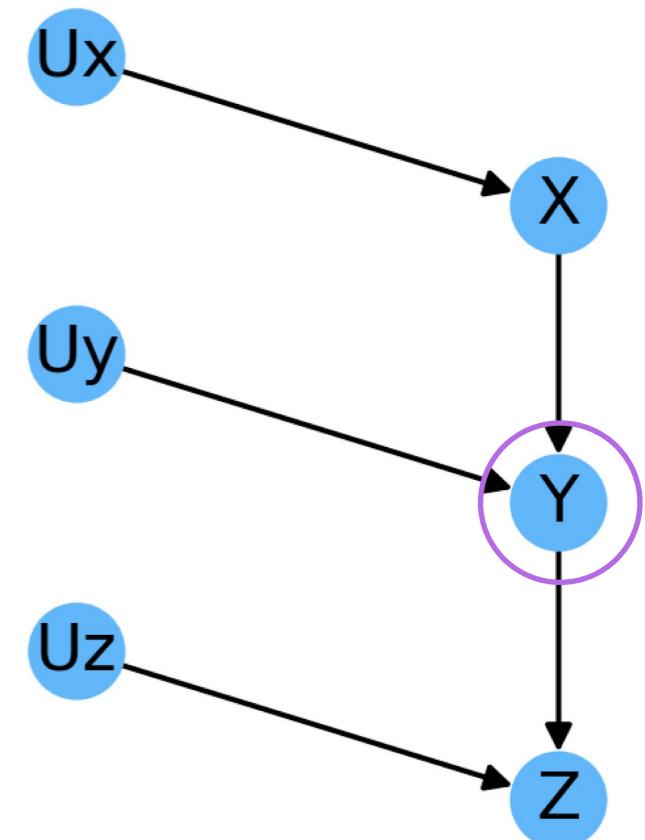
Rule 0 (Edge dependency) — Any two variables with a directed edge between them are dependent



Chains

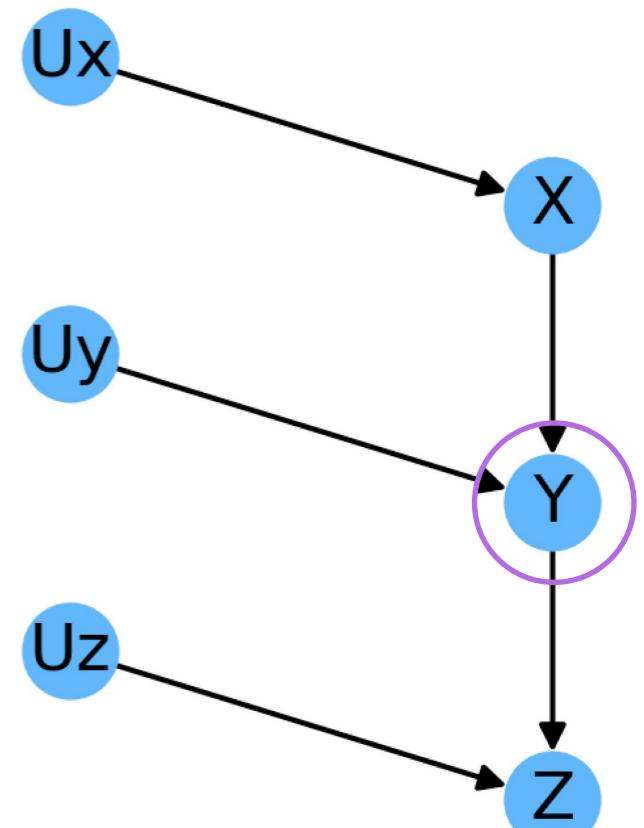
- From this Graph, we can immediately infer:

- Z and Y are **dependent** – $P(Z|Y) \neq P(Z)$
- Y and X are **dependent** – $P(Y|Z) \neq P(Y)$
- Z and X are likely **dependent** – $P(Z|X) \neq P(Z)$



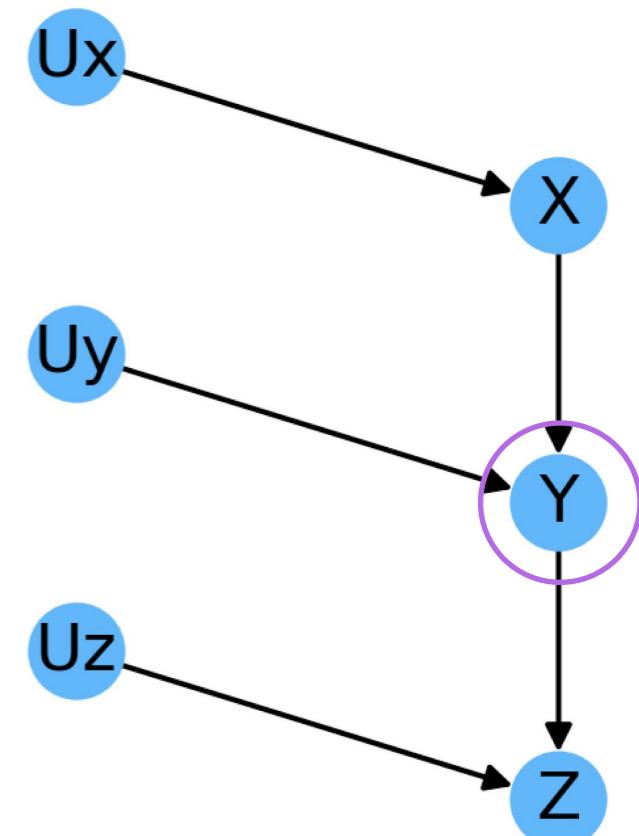
Chains

- From this Graph, we can immediately infer:
 - Z and Y are **dependent** – $P(Z|Y) \neq P(Z)$
 - Y and X are **dependent** – $P(Y|Z) \neq P(Y)$
 - Z and X are likely **dependent** – $P(Z|X) \neq P(Z)$
- We can also easily see how by setting the value of Y , we remove any influence that X can have on Z (as that can only happen through changes in Y). In other words:
 - Z and X are **independent** conditional on Y – $P(Z|X, Y) = P(Z|Y)$



Chains

- From this Graph, we can immediately infer:
 - Z and Y are **dependent** – $P(Z|Y) \neq P(Z)$
 - Y and X are **dependent** – $P(Y|Z) \neq P(Y)$
 - Z and X are likely **dependent** – $P(Z|X) \neq P(Z)$
- We can also easily see how by setting the value of Y , we remove any influence that X can have on Z (as that can only happen through changes in Y). In other words:
 - Z and X are **independent** conditional on Y – $P(Z|X, Y) = P(Z|Y)$

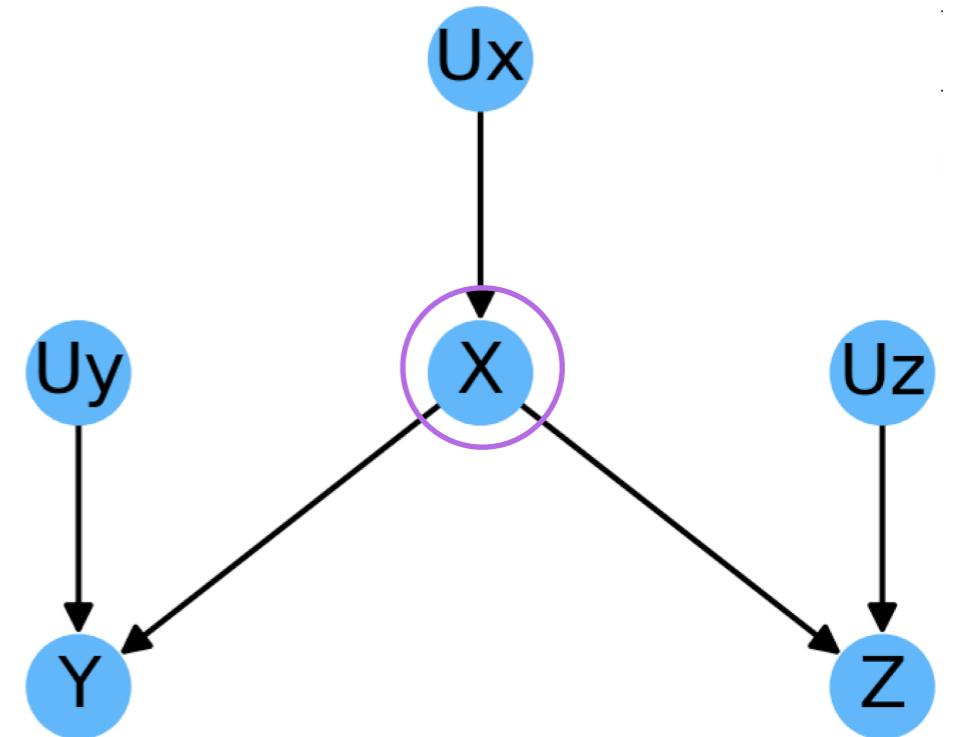


Rule 1 (Conditional Independence on Chains) – Two variables, X and Z , are conditionally independent given Y , if there is only one unidirectional path between X and Z and Y is any set of variables that intercepts that path.

Forks

- From this Graph we can infer:

- X and Y are **dependent** – $P(X|Y) \neq P(X)$
- X and Z are **dependent** – $P(X|Z) \neq P(X)$



Forks

- From this Graph we can infer:

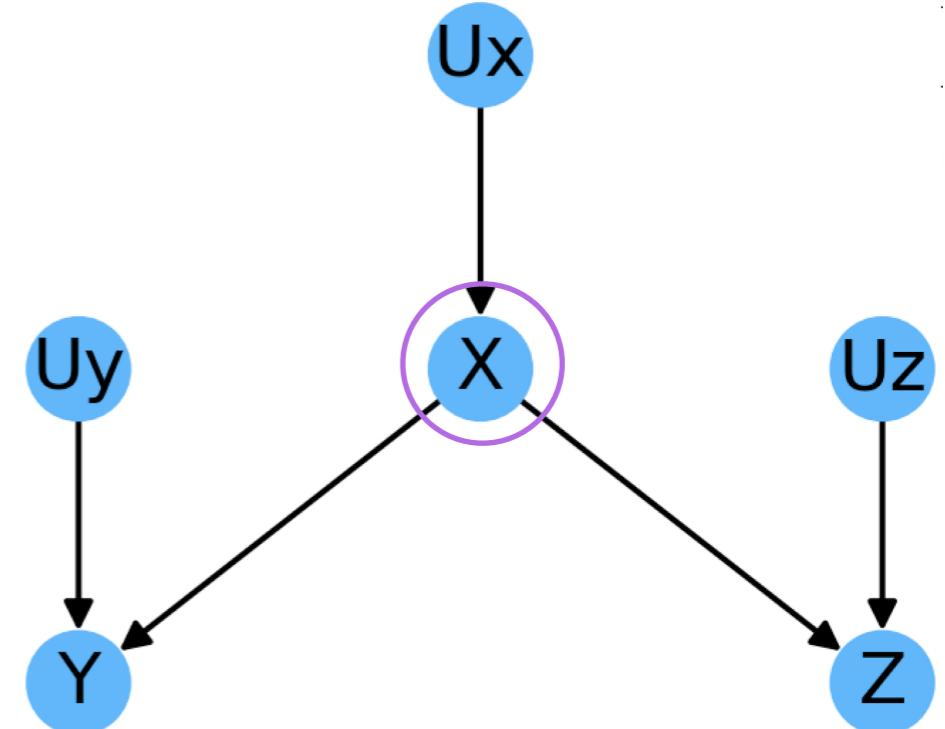
- X and Y are **dependent** – $P(X|Y) \neq P(X)$

- X and Z are **dependent** – $P(X|Z) \neq P(X)$

- Furthermore, we can see:

- Z and Y are likely **dependent** –
 $P(Z|Y) \neq P(Z)$ as their values are both directly determined by the value of X

- Y and Z are **independent**, conditional on X – $P(Y|Z, X) = P(Y|X)$ since from a given value of X , Y and Z are free to vary according to $P(Y|X)$ and $P(Z|X)$, respectively



Forks

- From this Graph we can infer:

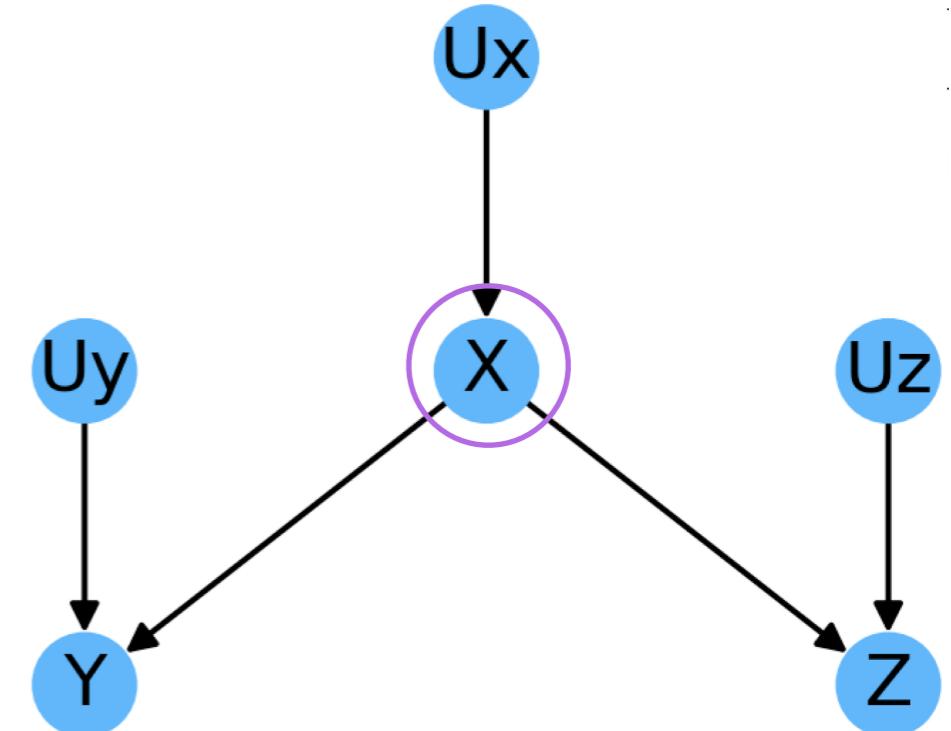
- X and Y are **dependent** — $P(X|Y) \neq P(X)$

- X and Z are **dependent** — $P(X|Z) \neq P(X)$

- Furthermore, we can see:

- Z and Y are likely **dependent** —
 $P(Z|Y) \neq P(Z)$ as their values are both directly determined by the value of X

- Y and Z are **independent**, conditional on X — $P(Y|Z, X) = P(Y|X)$ since from a given value of X , Y and Z are free to vary according to $P(Y|X)$ and $P(Z|X)$, respectively



Rule 2 (Conditional Independence in Forks) — If a variable X is a common cause of variables Y and Z , and there is only one path between Y and Z , then Y and Z are **independent** conditional on X .

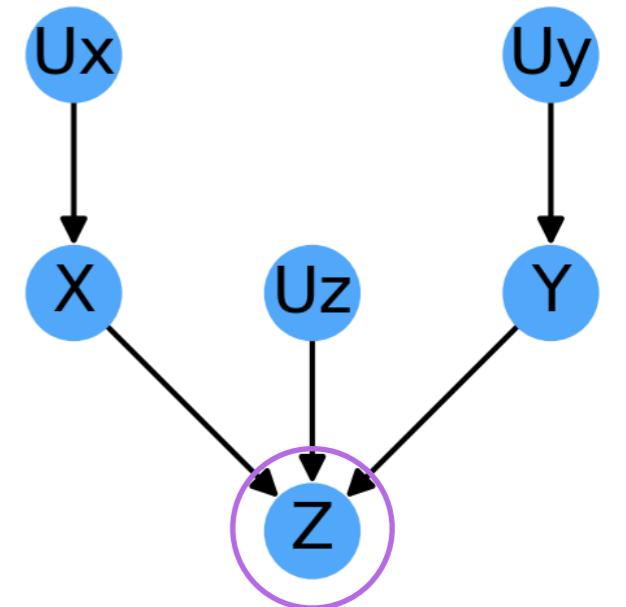
Colliders

- From this Graph, we infer:

- X and Z are **dependent** – $P(X|Z) \neq P(X)$
- Y and Z are **dependent** – $P(Y|Z) \neq P(Y)$ as the value of Z is determined by both X and Y and so it provides information about their values (**Information traveling “backwards”**)
- X and Y are **independent** – $P(X|Y) = P(X)$ as they share no common ancestors

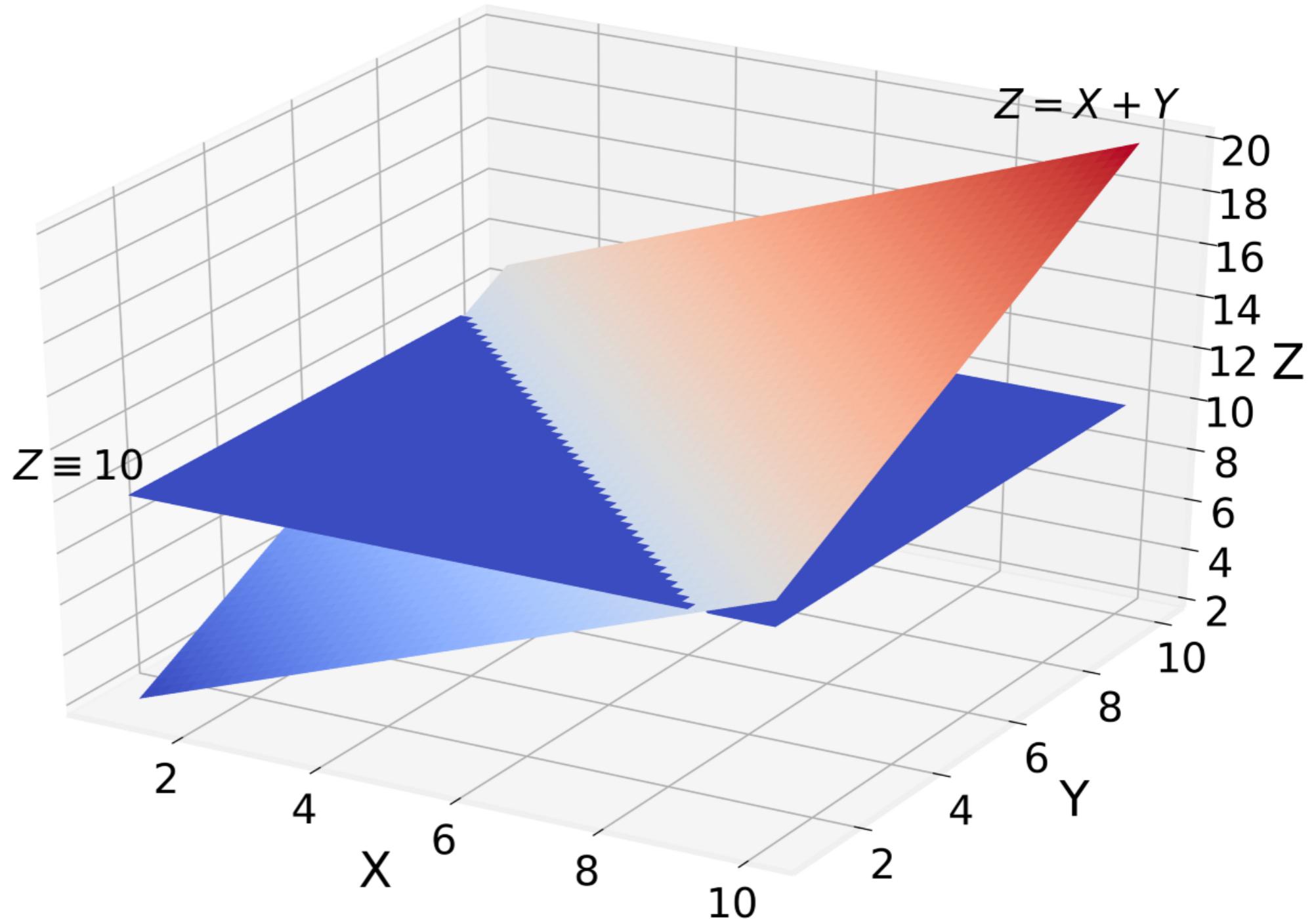
- Furthermore:

- X and Y are **dependent** conditional on Z – $P(X|Y,Z) \neq P(X|Z)$ as only specific values of X and Y can be combined to produce the value of Z
- Let's consider a simple example: $Z = X + Y$



Colliders

Fixing the value of Z defines the intersecting plane to the $X+Y$ surface, limiting the possible values of X and Y



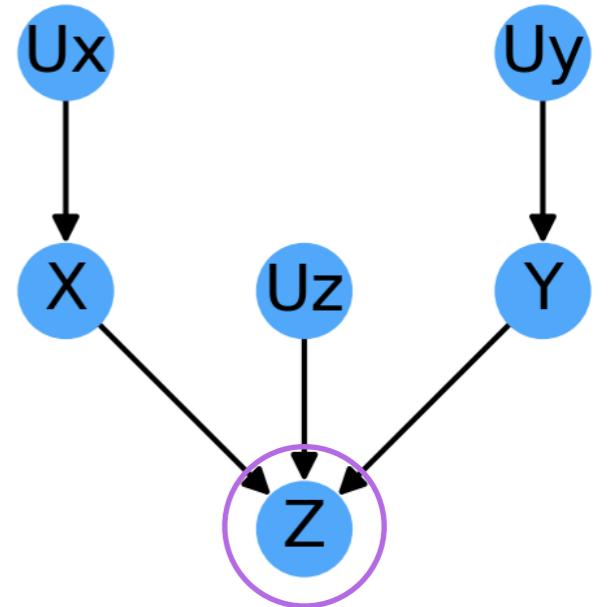
Colliders

- From this Graph, we infer:

- X and Z are **dependent** – $P(X|Z) \neq P(X)$
- Y and Z are **dependent** – $P(Y|Z) \neq P(Y)$ as the value of Z is determined by both X and Y and so it provides information about their values (**Information traveling “backwards”**)
- X and Y are **independent** – $P(X|Y) = P(X)$ as they share no common ancestors

- Furthermore:

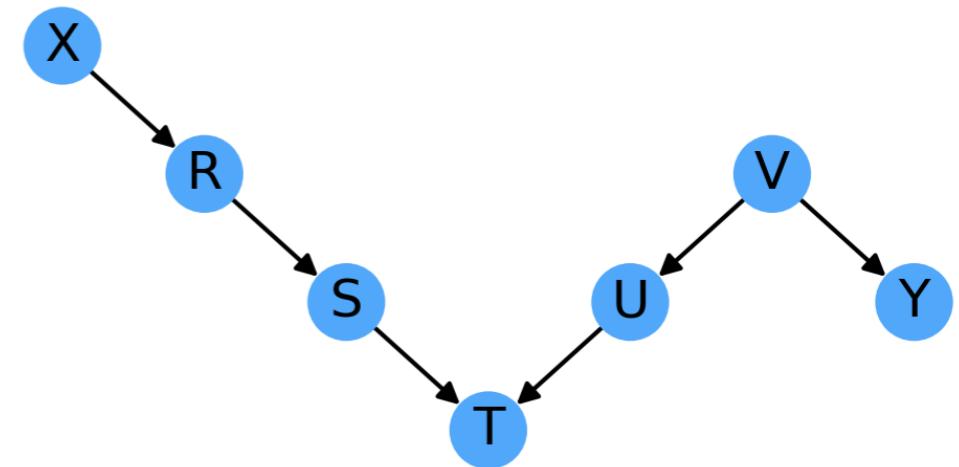
- X and Y are **dependent** conditional on Z – $P(X|Y,Z) \neq P(X|Z)$ as only specific values of X and Y can be combined to produce the value of Z



Rule 3 (Conditional Independence in Colliders): If a variable Z is the collision node between two variables X and Y , and there is only one path between X and Y , then X and Y are unconditionally **independent** but are **dependent** conditional on Z and any descendants of Z .

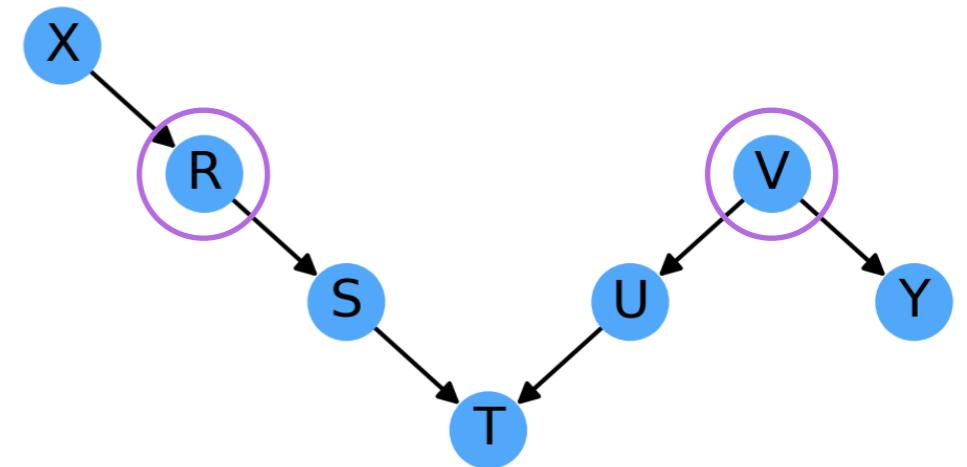
d-separation

- Using the rules introduced above we can reason about this Graph.



d-separation

- Using the rules introduced above we can reason about this Graph.
- If we condition on the values of **R** and **V** we can apply them to determine which variables should be **independent**:

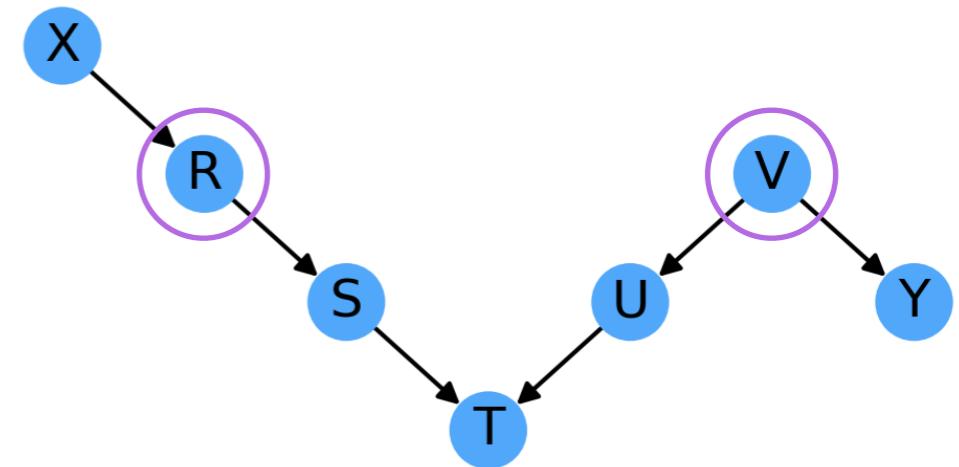


- **Rule 0** - Exclude all adjacent pairs of variables: (**S**, **T**), (**V**, **T**) are **dependent** conditional on **R** and **V**
- **Rule 1** - (**X**, **S**) and (**X**, **T**) are **independent** conditional on **R**. Also, **R** and **V** block the only path from **X** to **Y**, making (**X**, **Y**), (**X**, **U**) and (**S**, **Y**) are also **independent**
- **Rule 2** - (**U**, **Y**) and (**T**, **Y**) are **independent** conditional on **V**
- **Rule 3** - (**S**, **U**) are **independent**

d-separation

- Using the rules introduced above we can reason about this Graph.
- If we condition on the values of **R** and **V** we can apply them to determine which variables should be **independent**:

- **Rule 0** - Exclude all adjacent pairs of variables: (**S**, **T**), (**V**, **T**) are **dependent** conditional on **R** and **V**
- **Rule 1** - (**X**, **S**) and (**X**, **T**) are **independent** conditional on **R**. Also, **R** and **V** block the only path from **X** to **Y**, making (**X**, **Y**), (**X**, **U**) and (**S**, **Y**) are also **independent**
- **Rule 2** - (**U**, **Y**) and (**T**, **Y**) are **independent** conditional on **V**
- **Rule 3** - (**S**, **U**) are **independent**



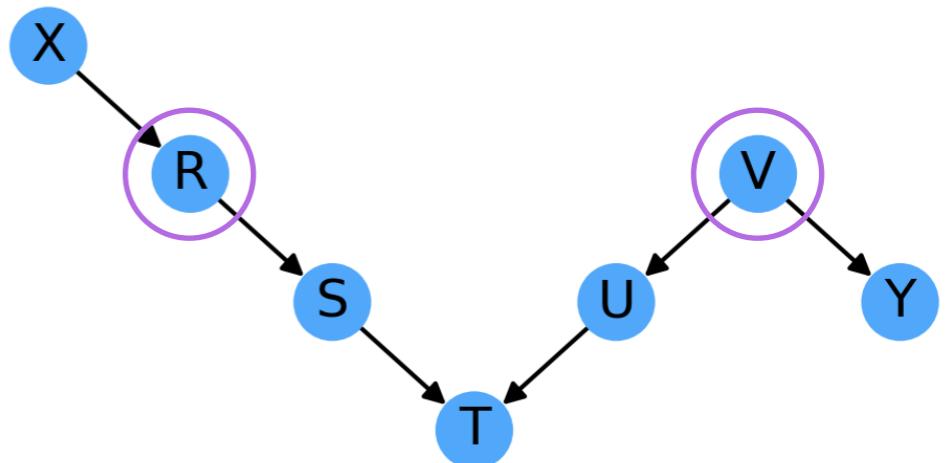
	source	target	Result
1	X	S	Rule 1, Independent
2	X	T	Rule 1, Independent
3	X	U	Blocked, Independent
5	X	Y	Blocked, Independent
11	S	T	Rule 0, Dependent
12	S	U	Rule 3, Independent
14	S	Y	Blocked, Independent
15	T	U	Rule 0, Dependent
17	T	Y	Rule 2, Independent
19	U	Y	Rule 2, Independent

d-separation

- We can verify all these relationships by simulating the model and performing a linear regression fit between the variables of interest.
- When we fit:

$$X \sim 1 + Y + Z$$

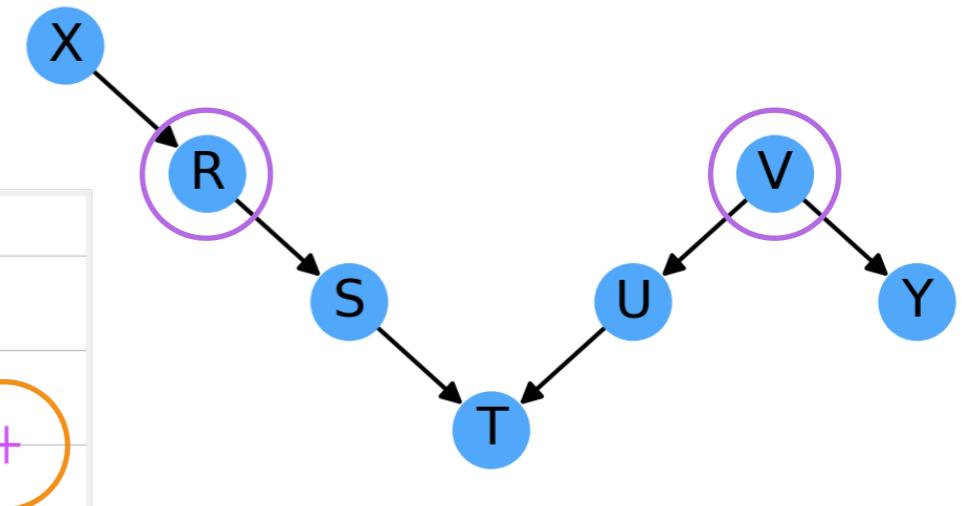
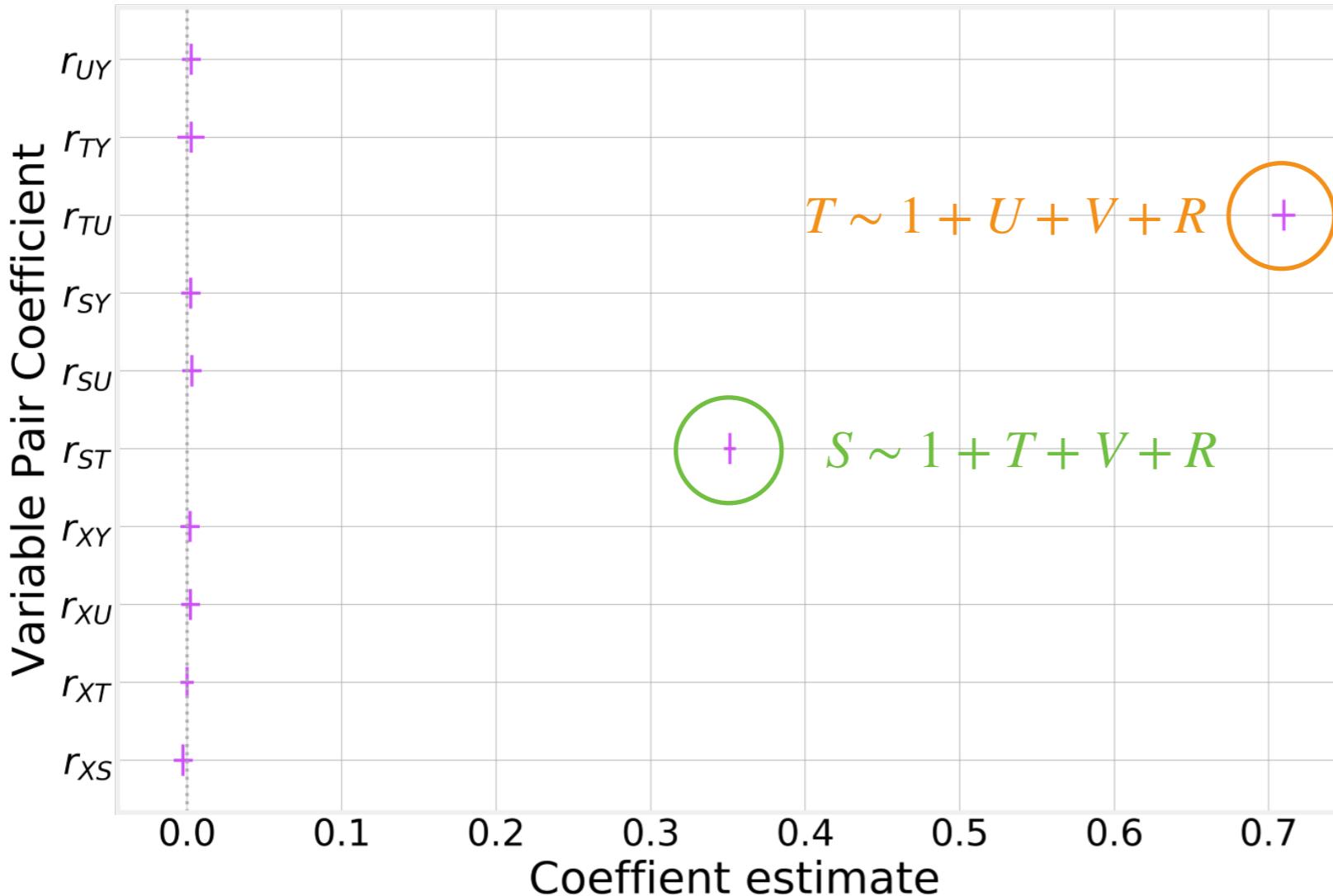
- If we find the r_{XY} coefficient to be zero, we can say that X is independent on Y conditional on Z , otherwise we say they are dependent.



	source	target	Result
1	X	S	Rule 1, Independent
2	X	T	Rule 1, Independent
3	X	U	Blocked, Independent
5	X	Y	Blocked, Independent
11	S	T	Rule 0, Dependent
12	S	U	Rule 3, Independent
14	S	Y	Blocked, Independent
15	T	U	Rule 0, Dependent
17	T	Y	Rule 2, Independent
19	U	Y	Rule 2, Independent

d-separation

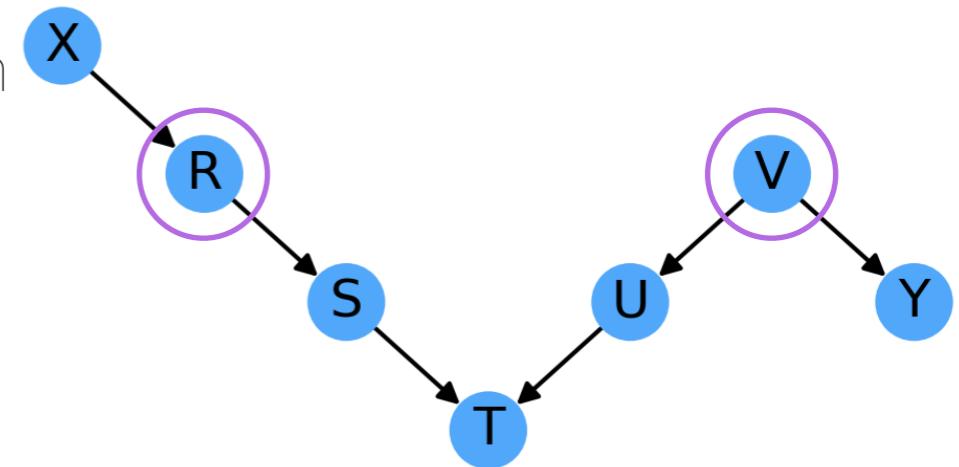
- Performing all the fits, we obtain:



	source	target	Result
1	X	S	Rule 1, Independent
2	X	T	Rule 1, Independent
3	X	U	Blocked, Independent
5	X	Y	Blocked, Independent
11	S	T	Rule 0, Dependent
12	S	U	Rule 3, Independent
14	S	Y	Blocked, Independent
15	T	U	Rule 0, Dependent
17	T	Y	Rule 2, Independent
19	U	Y	Rule 2, Independent

d-separation

- The concept of d-separations allows us to easily reason about the independency of any pair of variables in a simple way



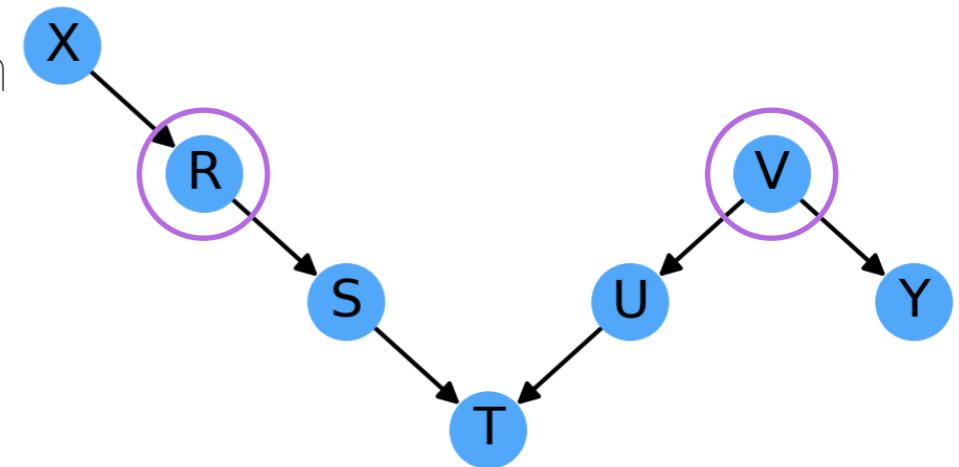
Definition (d-separation) — A path p is blocked by a set of nodes Z , if and only if:

- p contains a **chain** of nodes $A \rightarrow B \rightarrow C$, or a **fork** $A \leftarrow B \rightarrow C$ such that the middle node B is in Z , or
- p is a **collider** $A \rightarrow B \leftarrow C$ such that the collision node B is not in Z , and no descendent of B is in Z .

If Z blocks every path between two nodes X and Y , then X and Y are **d-separated**, conditional on Z , and thus are **independent** conditional on Z

d-separation

- The concept of d-separations allows us to easily reason about the independency of any pair of variables in a simple way:



Definition (d-separation) — A path p is blocked by a set of nodes Z , if and only if:

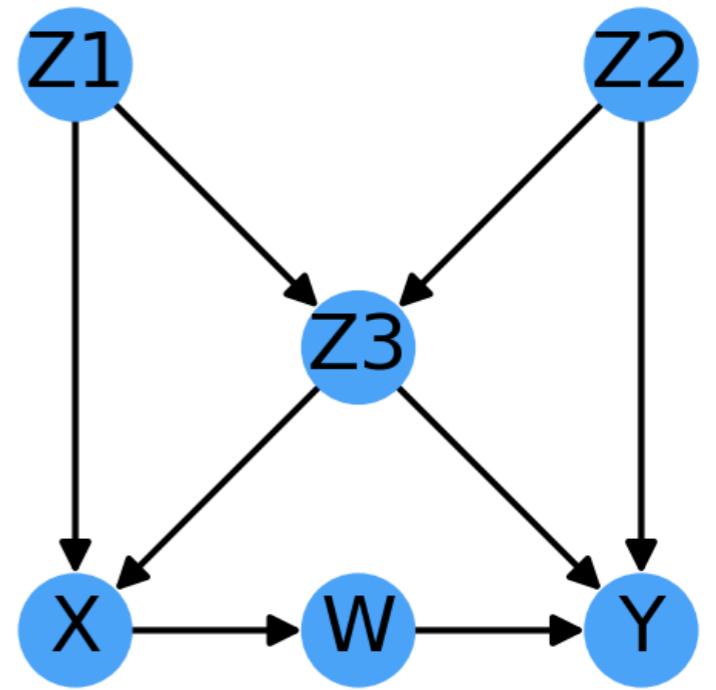
- p contains a **chain** of nodes $A \rightarrow B \rightarrow C$, or a **fork** $A \leftarrow B \rightarrow C$ such that the middle node B is in Z , or
- p is a **collider** $A \rightarrow B \leftarrow C$ such that the collision node B is not in Z , and no descendent of B is in Z .

If Z blocks every path between two nodes X and Y , then X and Y are **d-separated**, conditional on Z , and thus are **independent** conditional on Z

- In other words: Any two nodes X , and Y that are **d-separated** conditional on Z are necessarily **independent** conditional on Z .

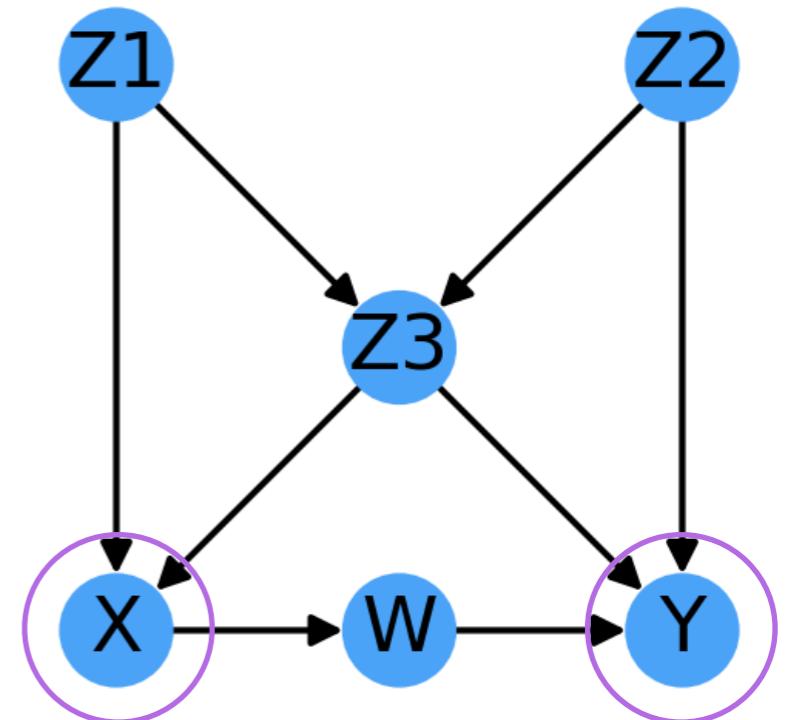
d-separation

- Let's consider another example.
- We have 3 Forks (Z_1 , Z_2 , and Z_3) and 3 Colliders (Z_3 , X , and Y)



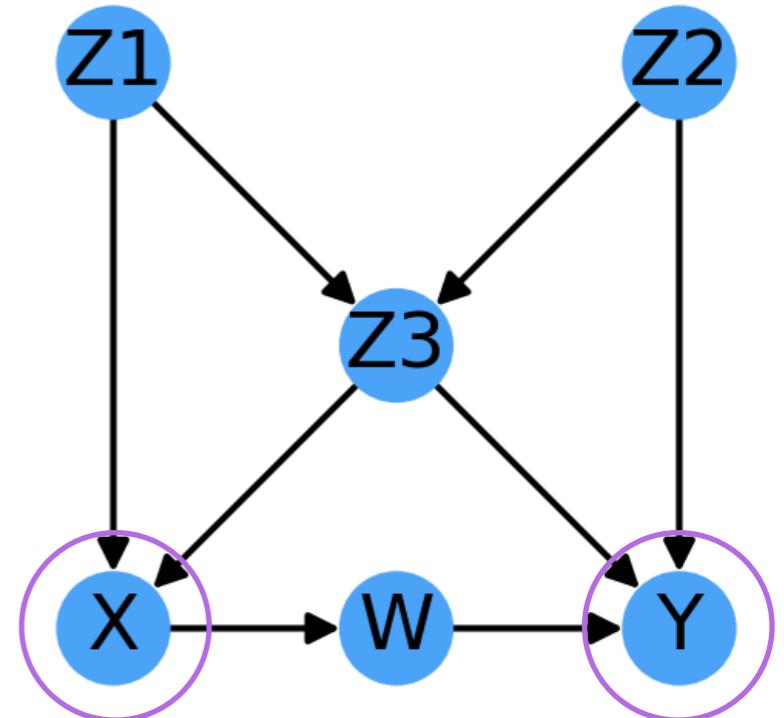
d-separation

- Let's consider another example.
- We have 3 Forks (Z_1 , Z_2 , and Z_3) and 3 Colliders (Z_3 , X , and Y)
- We see that X and Y are d-connected through W , making them unconditionally dependent.
- How can we make them independent (d-separated)?



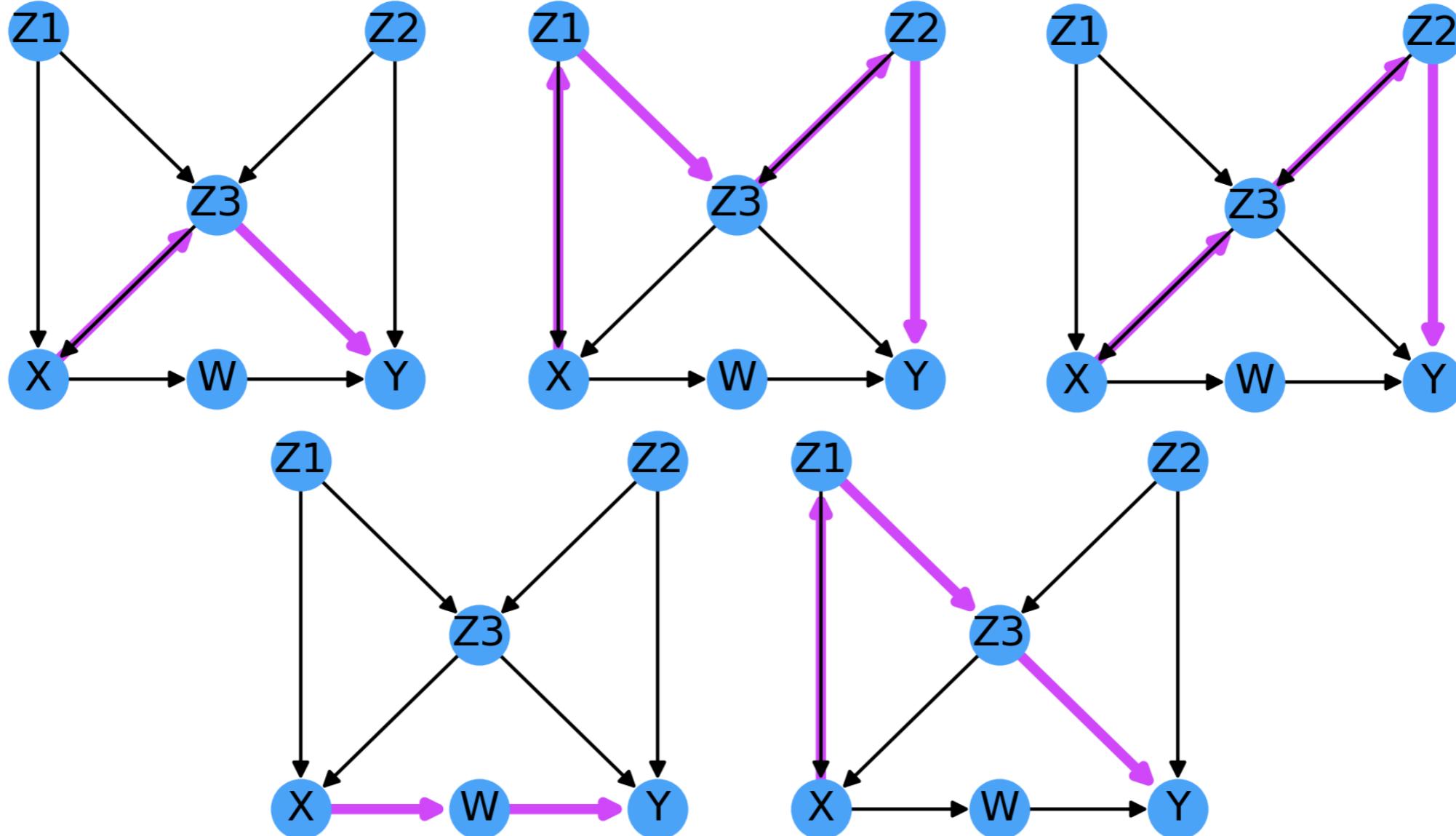
d-separation

- Let's consider another example.
- We have 3 Forks (Z_1 , Z_2 , and Z_3) and 3 Colliders (Z_3 , X , and Y)
- We see that X and Y are d-connected through W , making them unconditionally dependent.
- How can we make them independent (d-separated)?
- We must find the minimum number of conditioning variables that blocks ALL paths between X and Y



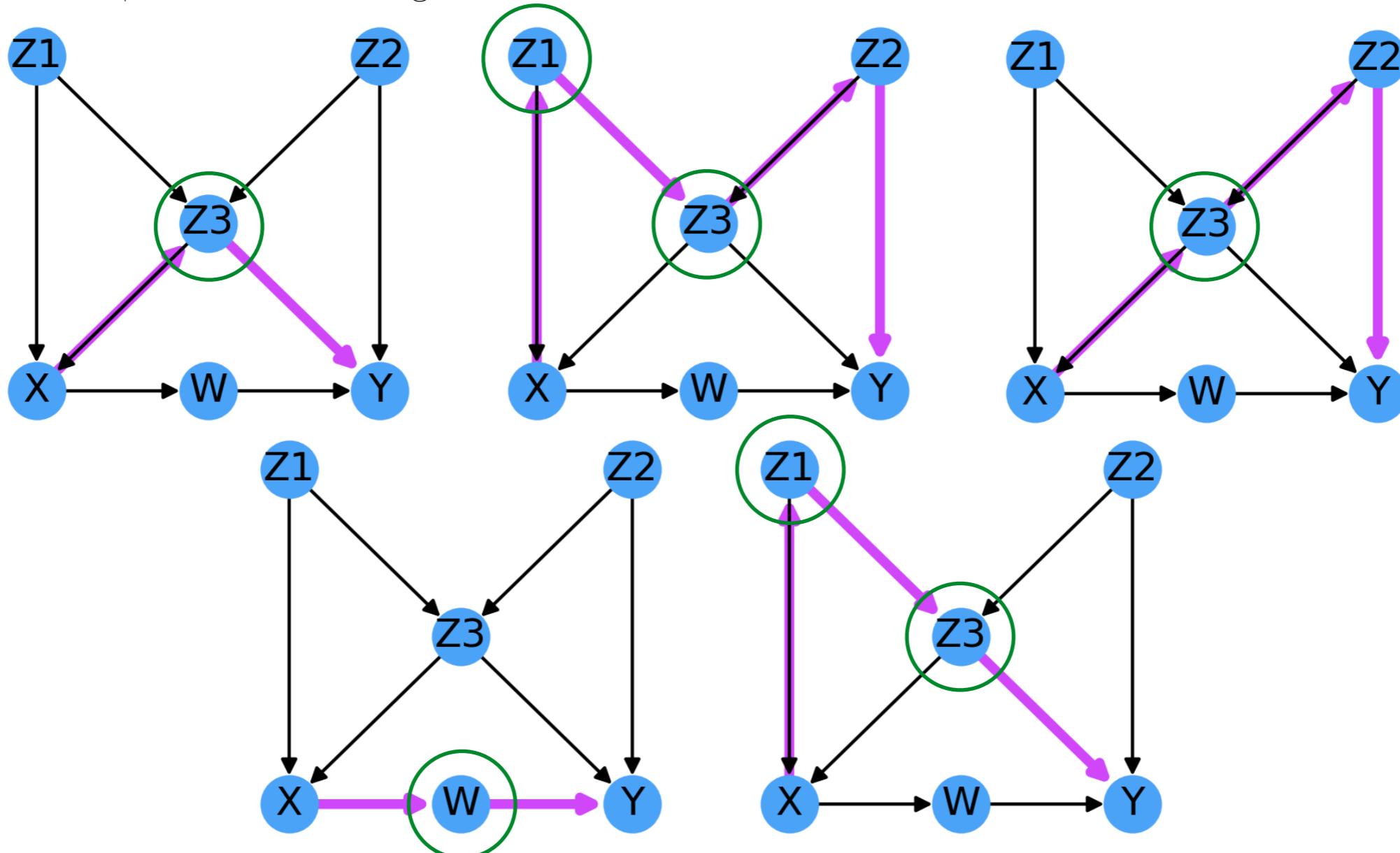
d-separation

- There are 5 paths connecting X and Y



d-separation

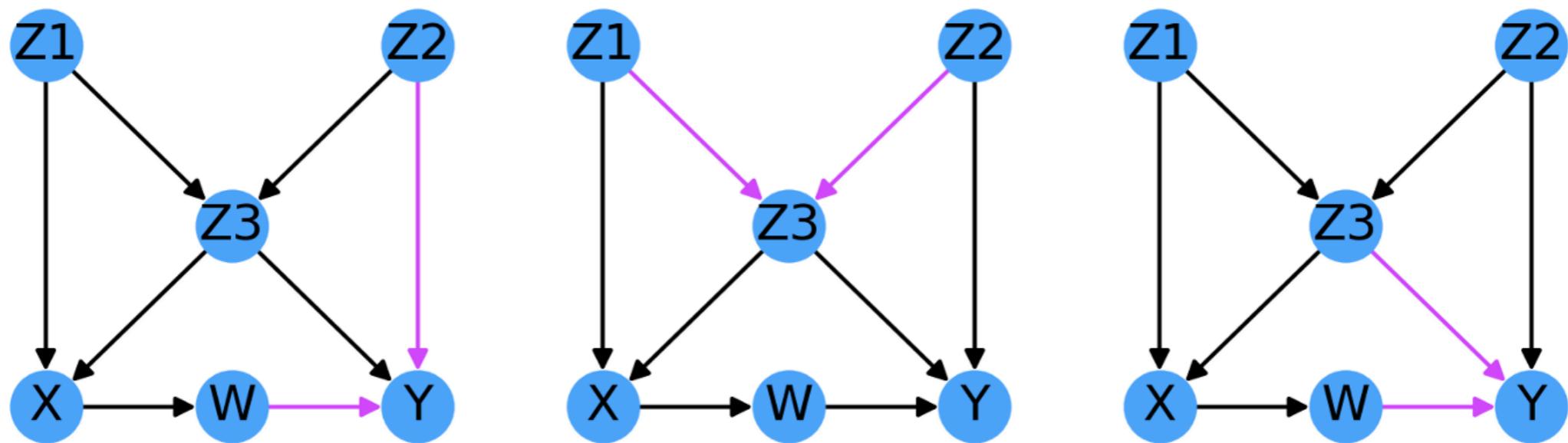
- There are 5 paths connecting X and Y



- We can block them all by conditioning on W , Z_1 and Z_3

v-structures

- v-structures are a Graph motif related to Colliders
- v-structures are defined as converging arrows whose tails are not connected by an arrow.
- In the previous DAG we have 3 v-structures:



- An individual Collider can give rise to multiple v-structures

Model Testing and Causal Search

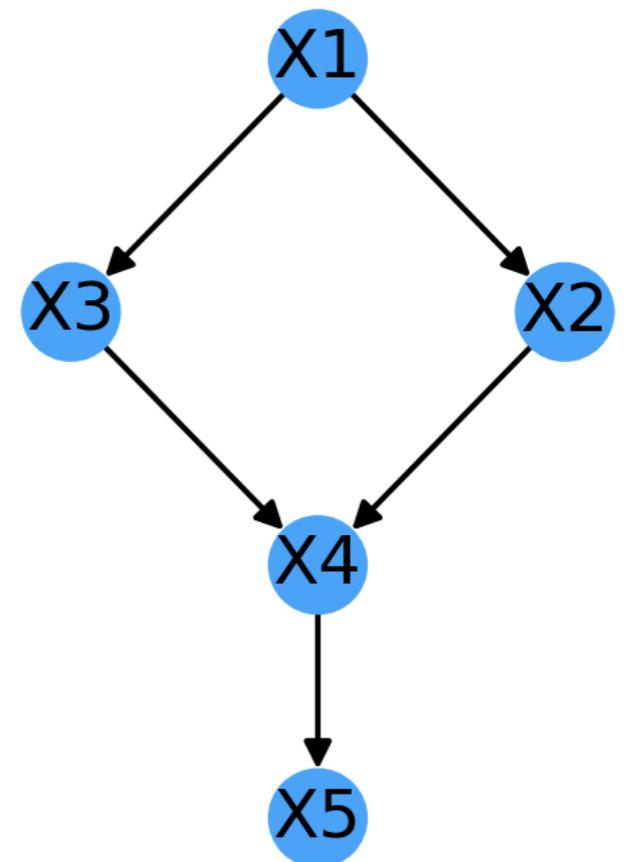
- **d-separation** allows us to quickly identify testable implications that can be tested directly on an empirical data set.
- A model is validated by the data if **every d-separation** condition is empirically verified by the data.

Model Testing and Causal Search

- **d-separation** allows us to quickly identify testable implications that can be tested directly on an empirical data set.
- A model is validated by the data if **every d-separation** condition is empirically verified by the data.
- However, different models that result in the same set of **d-separation** conditions are indistinguishable and said to be part of the same **equivalence class**.
- Two causal model DAGs are said to belong to the same **equivalence class** if:
 - They contain the same **skeleton** (set of **undirected** edges)
 - Exactly the same set of **v-structures**
 - We introduce no cycles in the resulting graph.
- It's easy to see that there is no edge we can change in the previous graph that would not break at least one of these conditions

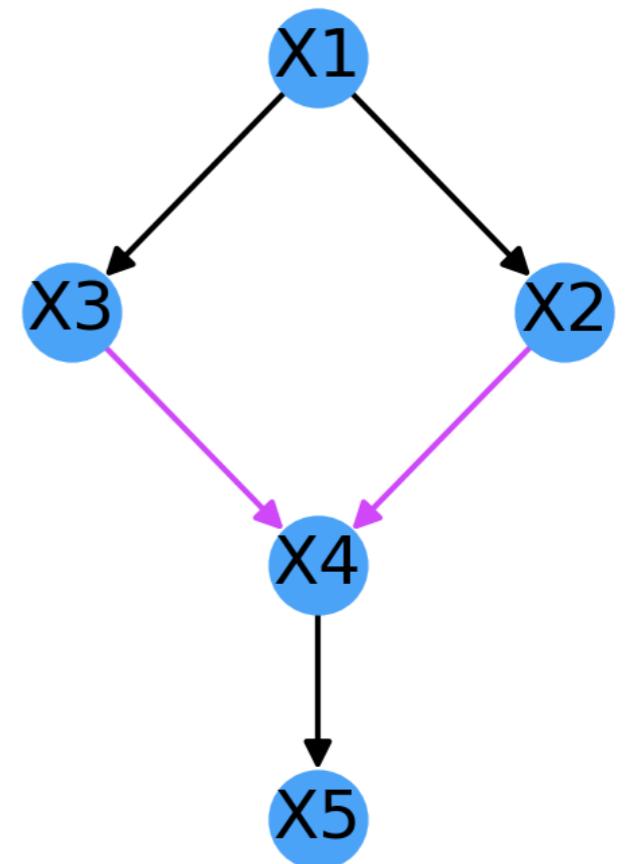
Model Testing and Causal Search

- Let us consider this DAG:



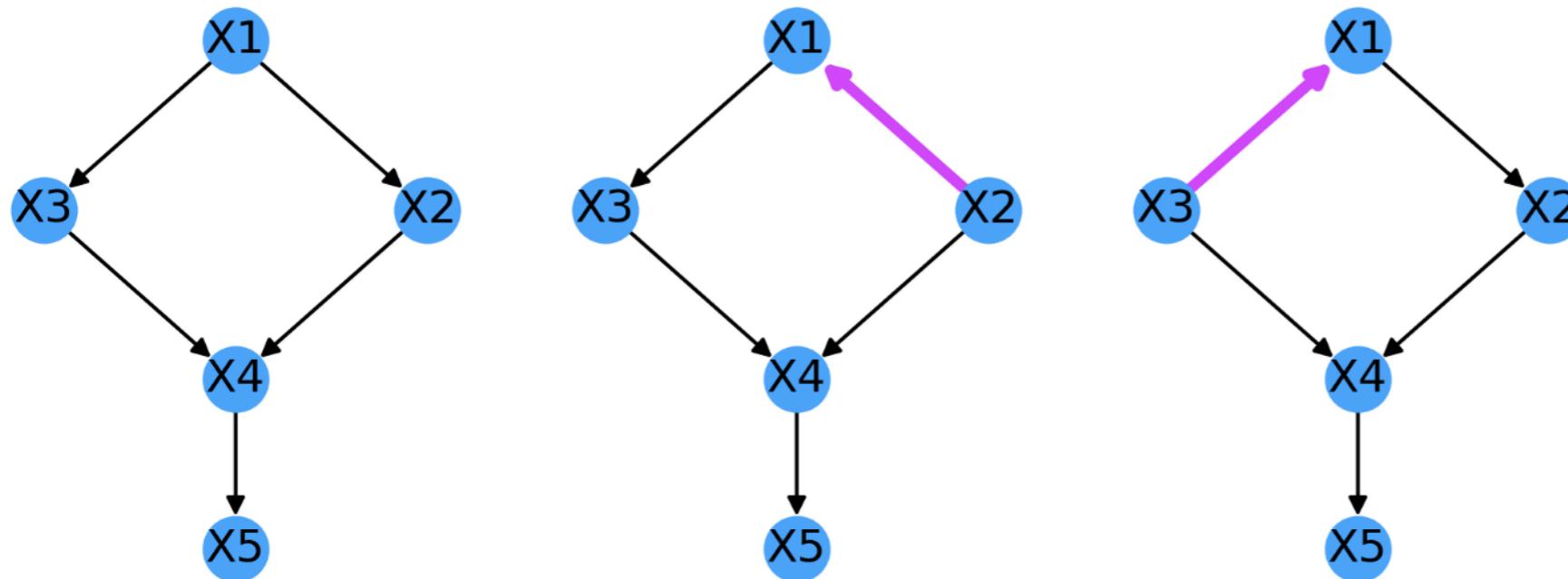
Model Testing and Causal Search

- Let us consider this DAG:
- Here we have exactly one **v-structure**:



Model Testing and Causal Search

- Let us consider this DAG:
- Here we have exactly one **v-structure**:
- There are two edges we can change without breaking the rules, so graph has an equivalence class of 3 graphs:



- Which implies that we have no way of determining the direction of the (X1, X2) and (X1, X3) edges



Code - Graphical Models
<https://github.com/DataForScience/CausalInference>



2. Interventions

Interventions

- In many cases we are interested in predicting the effects of interventions:
 - What happens if we take medication **X**?
 - What happens if we change the color of bottom **Y**?
 - What happens if we increase the price by **Z**?

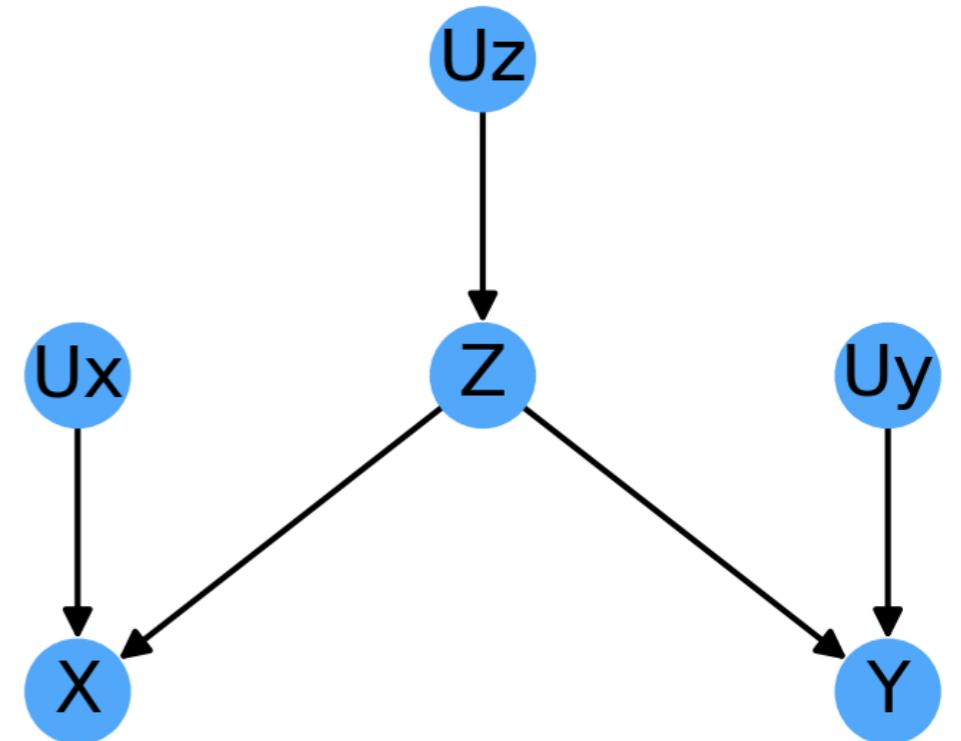
Interventions

- In many cases we are interested in predicting the effects of interventions:
 - What happens if we take medication **X**?
 - What happens if we change the color of bottom **Y**?
 - What happens if we increase the price by **Z**?
- The gold standard to study the effect of interventions is the Randomized Control Trial, like the A/B test we considered before
 - Randomly divide the population into two groups
 - Apply the intervention to one and no the other
 - Analyze the differences
- Unfortunately, it's not always possible to perform the randomized experiment



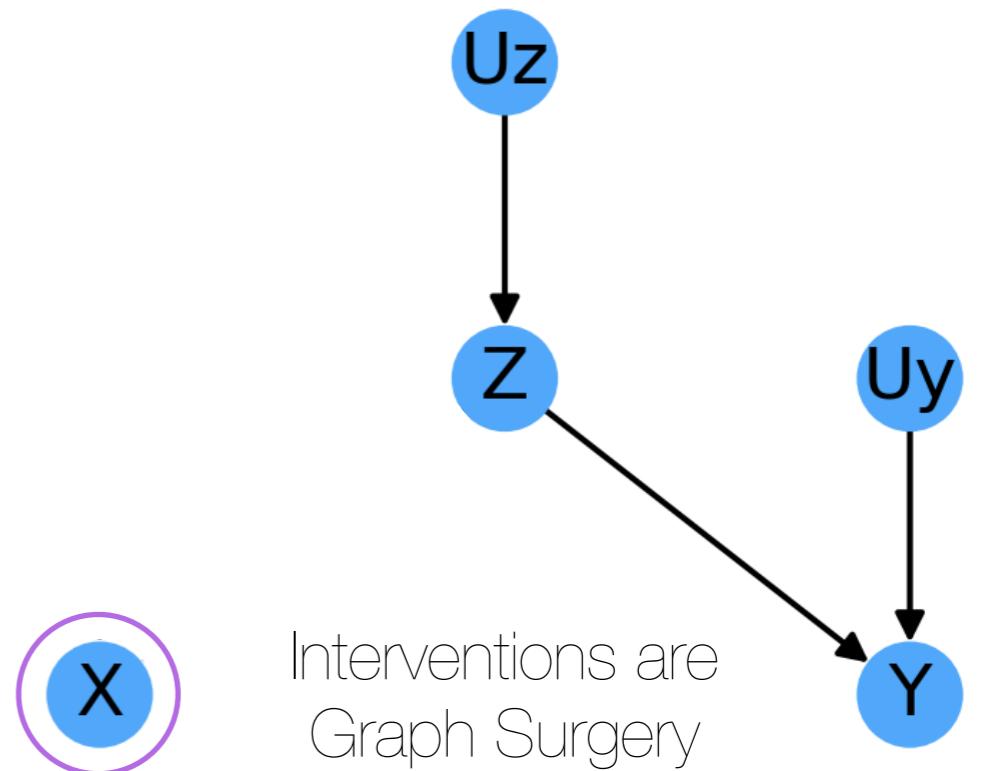
Interventions

- We can measure the effects of an intervention using purely observational data.
- Intervening requires that we **fix the value** of a given variable. This is equivalent to **removing all incoming edges**



Interventions

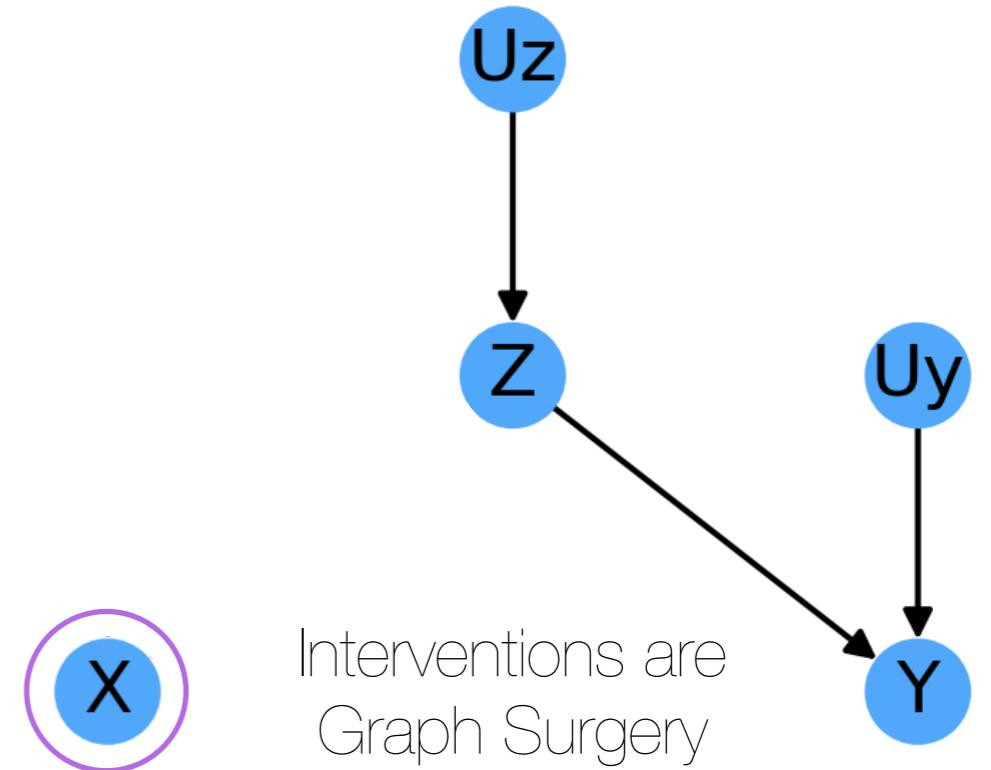
- We can measure the effects of an intervention using purely observational data.
- Intervening requires that we **fix the value** of a given variable. This is equivalent to **removing all incoming edges**
- In this DAG, if we were to intervene on **X**, we would obtain



Interventions

- We can measure the effects of an intervention using purely observational data.
- Intervening requires that we **fix the value** of a given variable. This is equivalent to **removing all incoming edges**
- In this DAG, if we were to intervene on **X**, we would obtain
- Mathematically, we represent the intervention using the **do ()** operator:

$$P_m(Y | \text{do}(X = x))$$



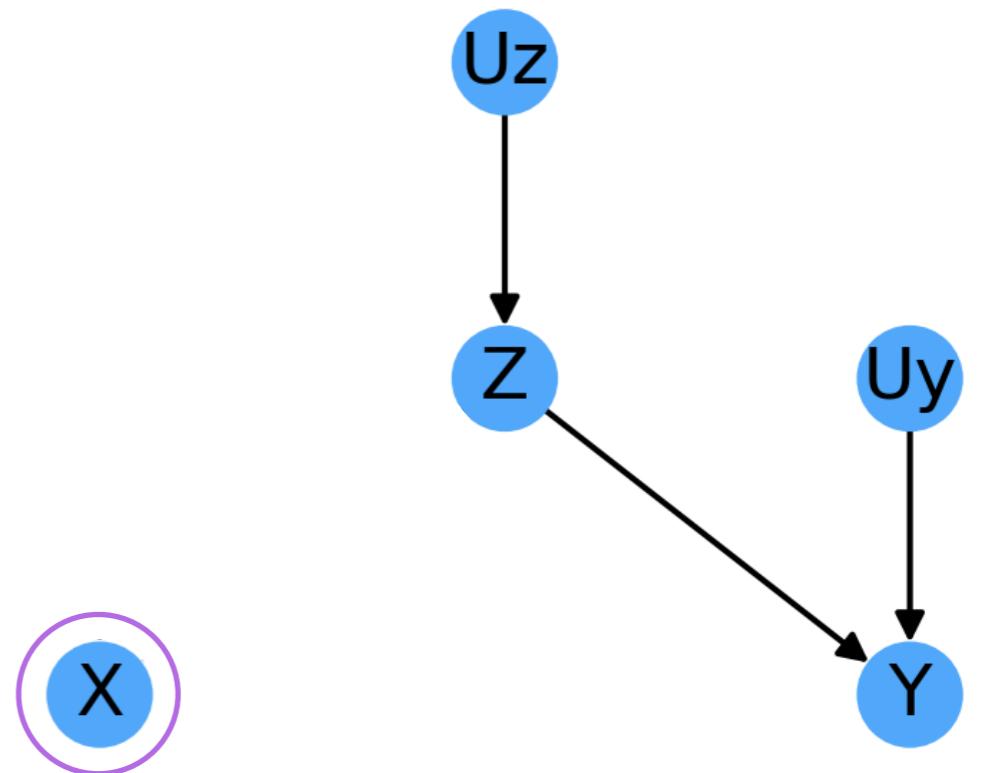
- Where the subscripted Probability **P_m** in highlights the fact that this is distribution observed in the modified Graph and in general is different than $P(Y|X)$

Average Causal Effect

- We quantify the effect of the intervention by calculating the difference between intervening or not:

$$P_m(Y | \text{do}(X = 1)) - P_m(Y | \text{do}(X = 0))$$

- This is similar in spirit to the $p_A - p_B$ term in the A/B analysis
- The detailed values of the modified Probability distributions can be calculated using the new Graph and our understanding of Causal Graphs

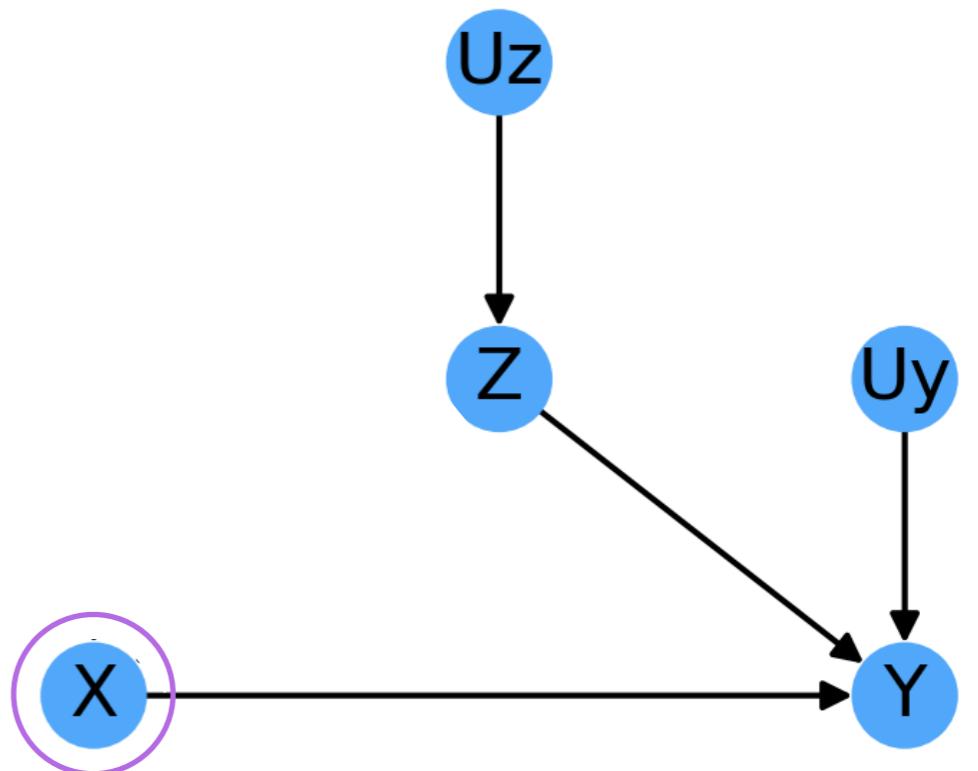


Average Causal Effect

- We quantify the effect of the intervention by calculating the difference between intervening or not:

$$P_m(Y | \text{do}(X = 1)) - P_m(Y | \text{do}(X = 0))$$

- This is similar in spirit to the $p_A - p_B$ term in the A/B analysis
- The detailed values of the modified Probability distributions can be calculated using the new Graph and our understanding of Causal Graphs
- Let us consider a slightly more **sophisticated** example



Average Causal Effect

- We quantify the effect of the intervention by calculating the difference between intervening or not:

$$P_m(Y|do(X=1)) - P_m(Y|do(X=0))$$

- This is similar in spirit to the $p_A - p_B$ term in the A/B analysis

- The detailed values of the modified Probability distributions can be calculated using the new Graph and our understanding of Causal Graphs

- Let us consider a slightly more **sophisticated** example

- The important thing to note is that there are some distributions that don't change between the two versions:

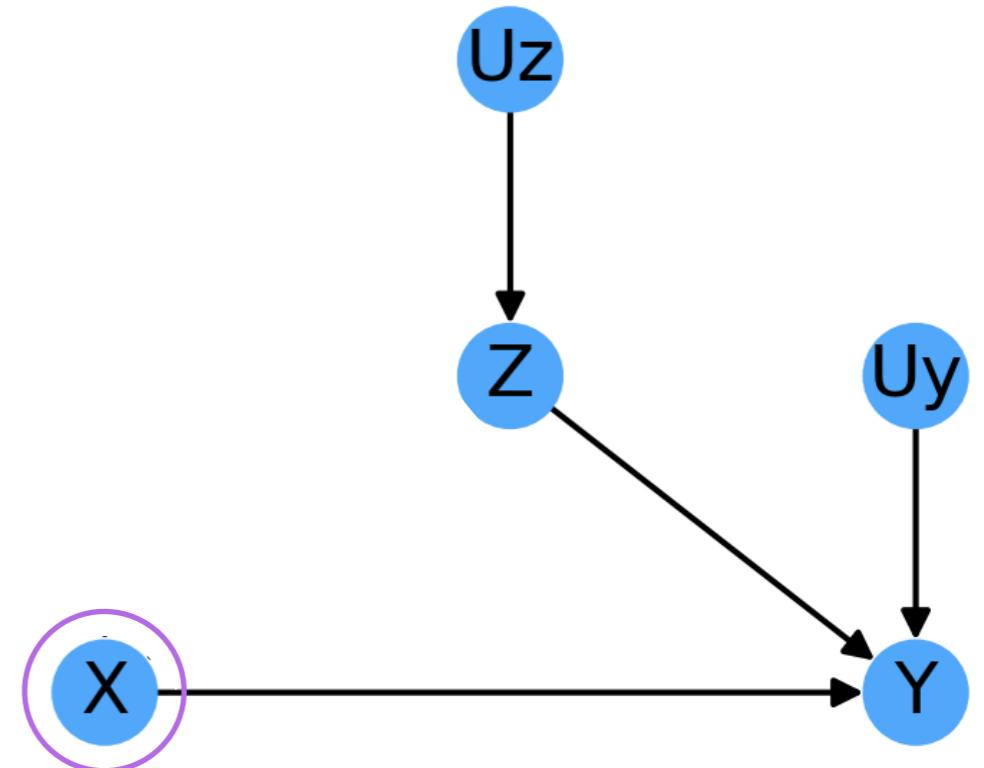
$$P_m(Z) \equiv P(Z)$$

$$P_m(Y|X,Z) \equiv P(Y|X,Z)$$

As the dependence of Y on X and Z aren't affected by our intervention. And

$$P_m(Z|X) \equiv P_m(Z) \equiv P(Z)$$

As X and Z are now d-separated



Causal Effect Rule

Causal Effect Rule - Given a Graph in which a set of variables PA are signaled as the parents of X , the causal effect of X on Y is given by:

$$P(Y|\text{do}(X)) = \sum_{\text{PA}} P(Y|X, \text{PA}) P(\text{PA})$$

where the sum over PA ranges over all the combinations of values allowed

- The sum over the values of the Parents of X while counter-intuitive is simple to understand:
 - When we fix the value of X , we break off the connection between X and PA so the variables in PA are able to vary freely
 - Even though we severed the edges between X and PA , the variables in PA can still be connected to the other variables in the DAG and are able to influence them

Back-Door Criterion

- The Back-Door Criterion allows us to determine under what conditions we are able to compute the effects of interventions using just observational data

Back-Door Criterion - A set of variables Z satisfies the backdoor criterion relative to (X, Y) if no node in Z is a descendant of X , and Z blocks every path between X and Y that contains an arrow into X

- In other words, the Back-Door Criterion simply means that we are controlling for all the variables necessary to prevent any information from leaking from Y back to X
- If a set of variables Z satisfy the backdoor criterion, then:

$$P(Y | \text{do}(X)) = \sum_Z P(Y | X, Z) P(Z)$$

and is always satisfied by **PA**.

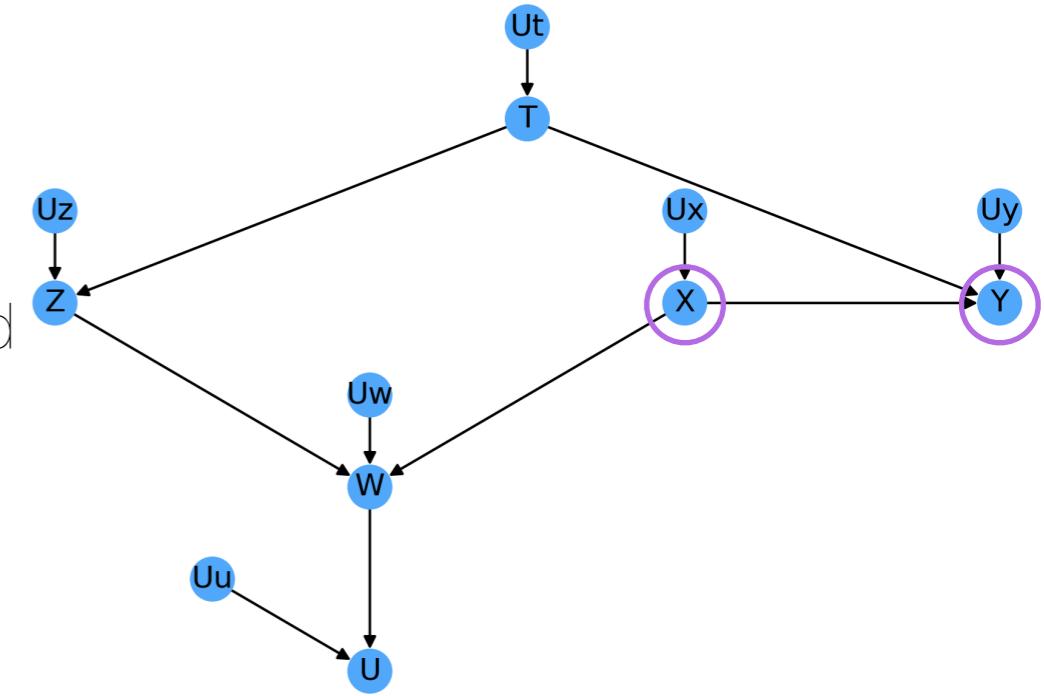
Back-Door Criterion

- Intuition:
 - Block all spurious paths between X and Y
 - Leave all directed paths between X and Y unperturbed
 - Create no new spurious paths

Back-Door Criterion

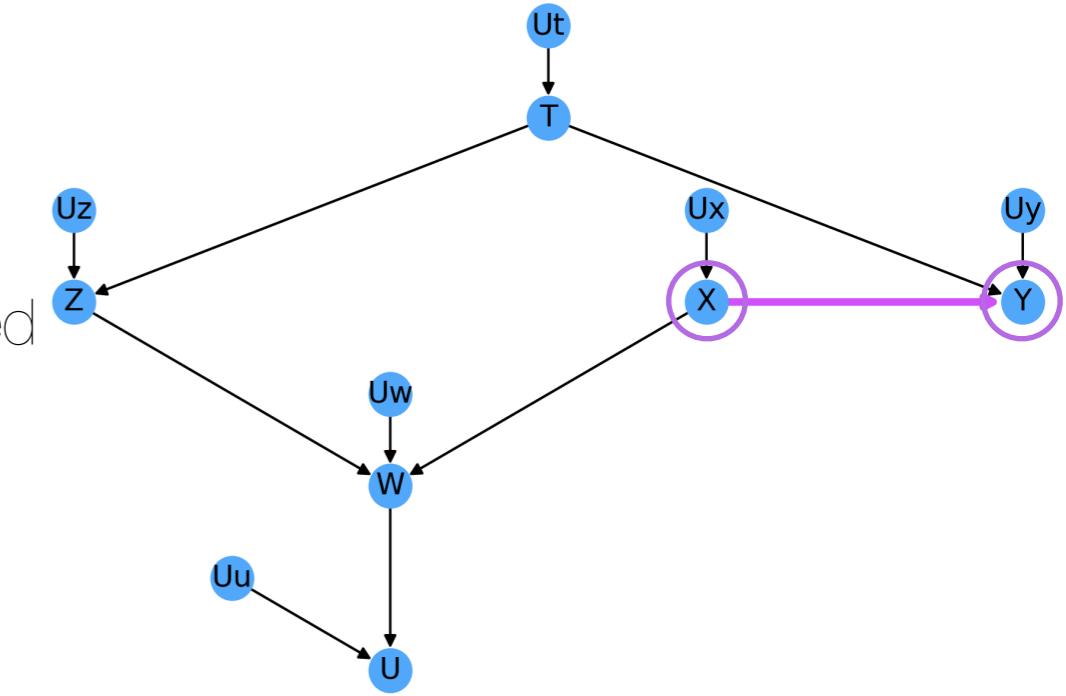
- Intuition:
 - Block all spurious paths between X and Y
 - Leave all directed paths between X and Y unperturbed
 - Create no new spurious paths
- Let's consider an example. We want to quantify:

$$P(Y | \text{do}(X))$$



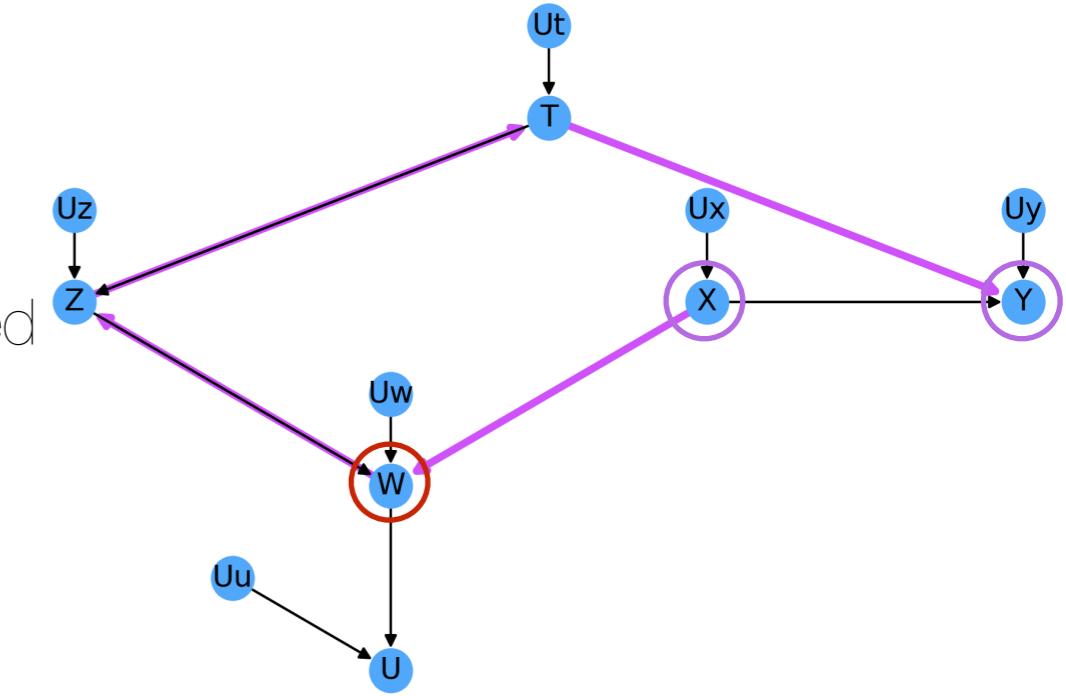
Back-Door Criterion

- Intuition:
 - Block all spurious paths between X and Y
 - Leave all directed paths between X and Y unperturbed
 - Create no new spurious paths
- Let's consider an example. We want to quantify:
$$P(Y | \text{do}(X))$$
- The backdoor path is blocked by the collider at W (**d-separation**), so
$$P(Y | \text{do}(X)) \equiv P(Y | X)$$



Back-Door Criterion

- Intuition:
 - Block all spurious paths between X and Y
 - Leave all directed paths between X and Y unperturbed
 - Create no new spurious paths
- Let's consider an example. We want to quantify:
$$P(Y | \text{do}(X))$$
- The backdoor path is blocked by the collider at W (**d-separation**), so
$$P(Y | \text{do}(X)) \equiv P(Y | X)$$
- Now let's imagine that we are conditioning on W , opening the back-door path.



Back-Door Criterion

- Intuition:
 - Block all spurious paths between X and Y
 - Leave all directed paths between X and Y unperturbed
 - Create no new spurious paths
- Let's consider an example. We want to quantify:
$$P(Y | \text{do}(X))$$
- The backdoor path is blocked by the collider at W (**d-separation**), so
$$P(Y | \text{do}(X)) \equiv P(Y | X)$$
- Now let's imagine that we are conditioning on W , opening the back-door path.
- In this case we have to condition on another variable, like T in order to block it, otherwise our measurement of $P(Y | \text{do}(X), W)$ will be biased. In this case, we would compute:

$$P(Y | \text{do}(X), W) = \sum_T P(Y | X, W, T) P(T | W)$$

Front-Door Criterion

- The Front-Door Criterion is a complementary approach to the Back-Door Criterion to identify sets of covariates that allow us to calculate the results of interventions using observational data.

Front-Door Criterion - A set of variables Z satisfies the front door criterion relative to (X, Y) if:

- Z intercepts all directed paths from X , to Y
- There is no unblocked path from X to Z
- All back-door paths from Z to Y are blocked by X

Front-Door Adjustment - If Z satisfies the front door criterion relative to (X, Y) and $P(X, Z) > 0$ then the causal effect of X on Y is:

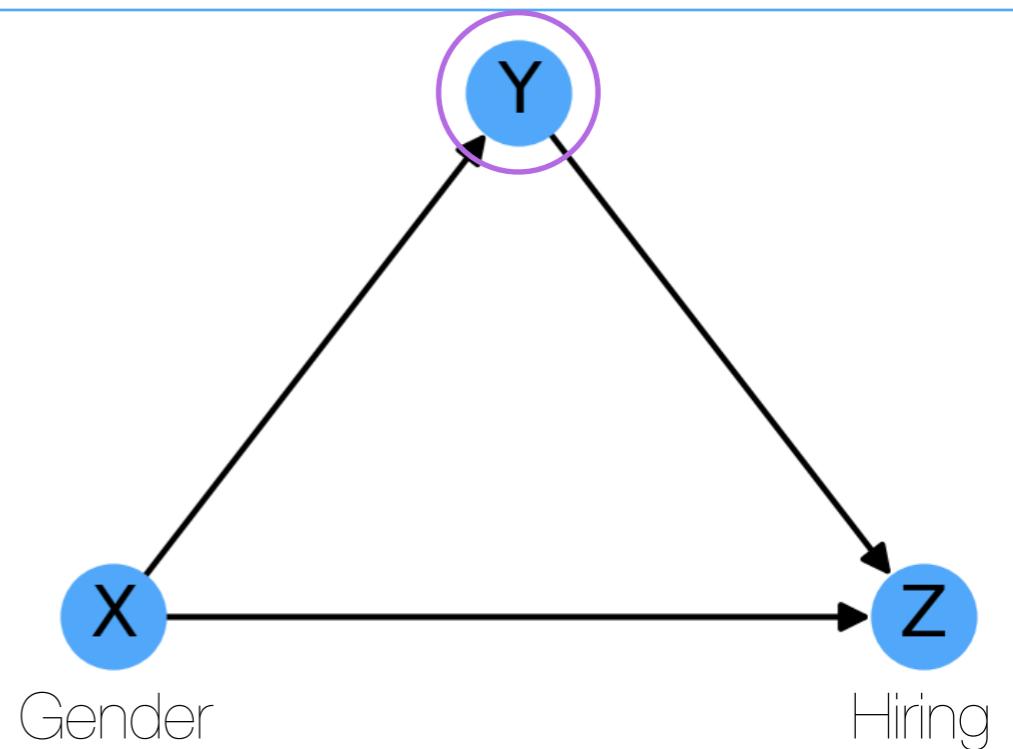
$$P(Y | \text{do}(X)) = \sum_Z P(Z|X) \sum_{X'} P(Y|X', Z) P(X')$$

Mediation

- Often a variable affects another both directly and indirectly through a set of **mediating variables**
- **X** is influencing **Z** both directly and through **Y**
- We can quantify the direct influence of **X** on **Z** by conditioning on **Y**

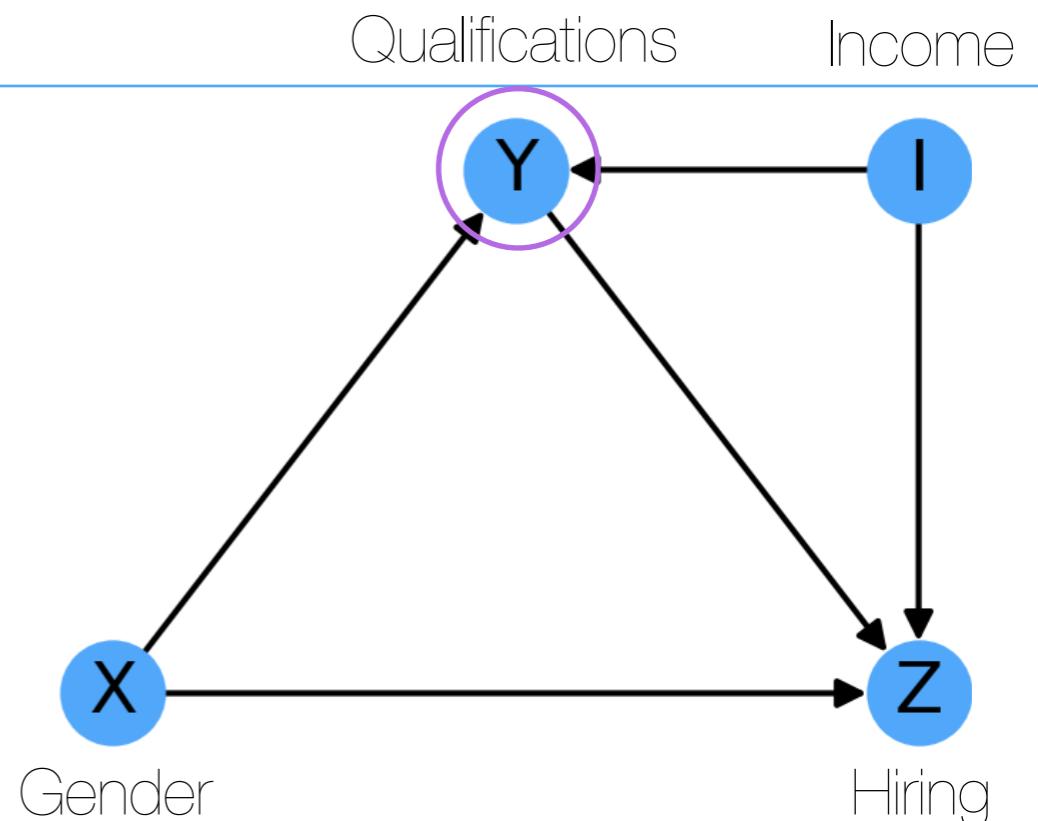
$$P(Z|X, Y)$$

Qualifications



Mediation

- Often a variable affects another both directly and indirectly through a set of **mediating variables**
- **X** is influencing **Z** both directly and through **Y**
- We can quantify the direct influence of **X** on **Z** by conditioning on **Y**
$$P(Z|X, Y)$$
- But what if there is a cofounder of **Y**?
- Now if we condition on **Y** we are conditioning on a **Collider** and unblocking the path through **I**.



Mediation

- Often a variable affects another both directly and indirectly through a set of **mediating variables**

- X is influencing Z both directly and through Y

- We can quantify the direct influence of X on Z by conditioning on Y

$$P(Z|X, Y)$$

- But what if there is a cofounder of Y ?

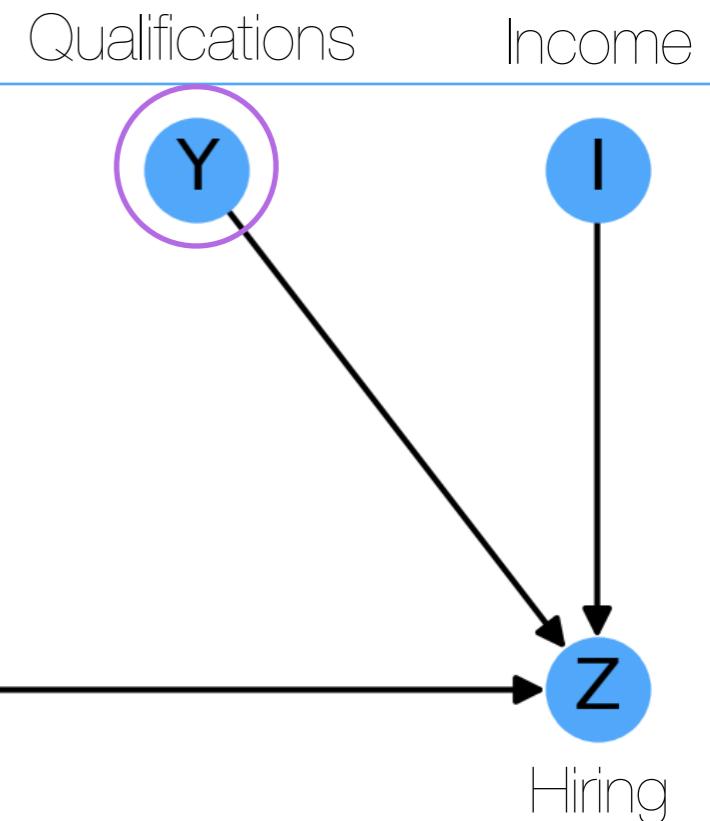
- Now if we condition on Y we are conditioning on a **Collider** and unblocking the path through I .

- One way around this difficulty is to **intervene** on Y , which removes the edges from X and I

- Now we can measure the **Controlled Direct Effect (CDE)**:

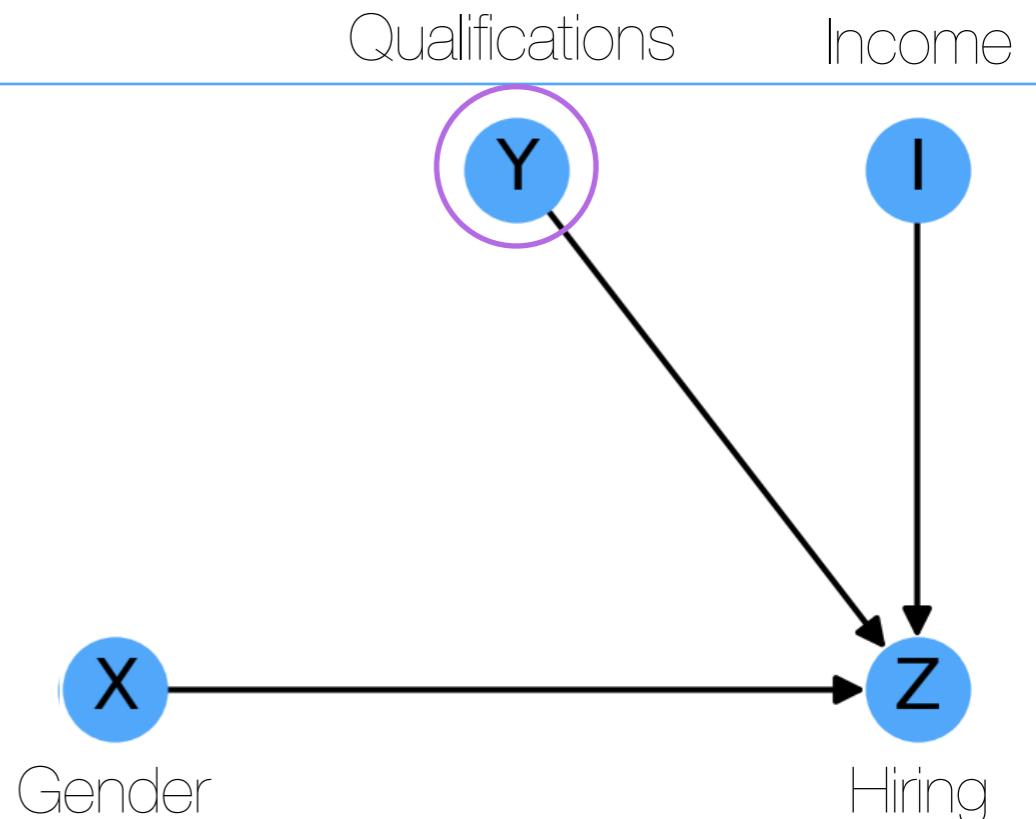
$$CDE = P(Y \mid \text{do}(X = 1) \text{ do}(Z)) - P(Y \mid \text{do}(X = 0) \text{ do}(Z))$$

- The Mediation equivalent of the **Average Causal Effect**



Mediation

- The CDE of **X** on **Y**, mediated by **Z** requires:
 - There is a set **S1** of variables that block all backdoor paths from **Z** to **Y**
 - There is a set **S2** of variables that block all backdoor paths from **X** to **Y** after removing all incoming edges from **Z**
- When both conditions are satisfied, we can express the CDE as a combination of purely observational quantities (without any **do()** operators)





Code - Interventions

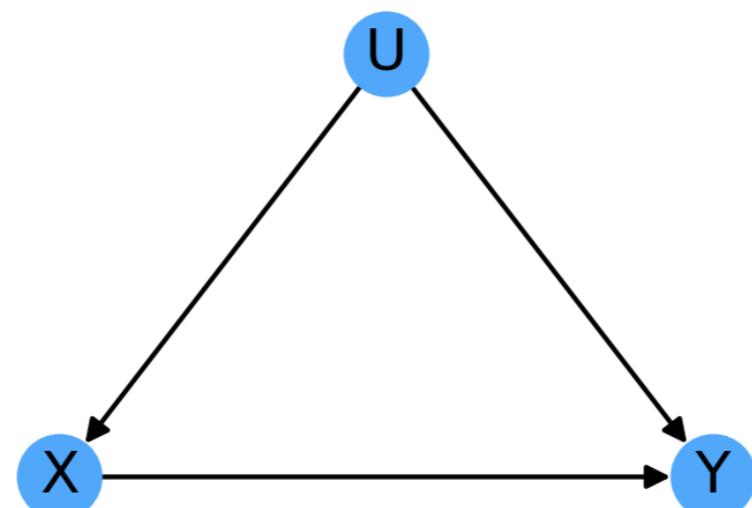
<https://github.com/DataForScience/CausalInference>



3. Counterfactuals

Counterfactuals

- **Counterfactual** - An "if" statement where the "if" part is untrue or unrealized
- Represent "What if" questions about things that we didn't do
 - What if I had taken a right turn instead of a left
- For our analysis they require a fully specified SCM, including both DAG and structural equations
- We use $Y_x(u) = y$ to represent Y would have value y , had X had the value x in situation $U=u$.
- Consider a simple SCM:



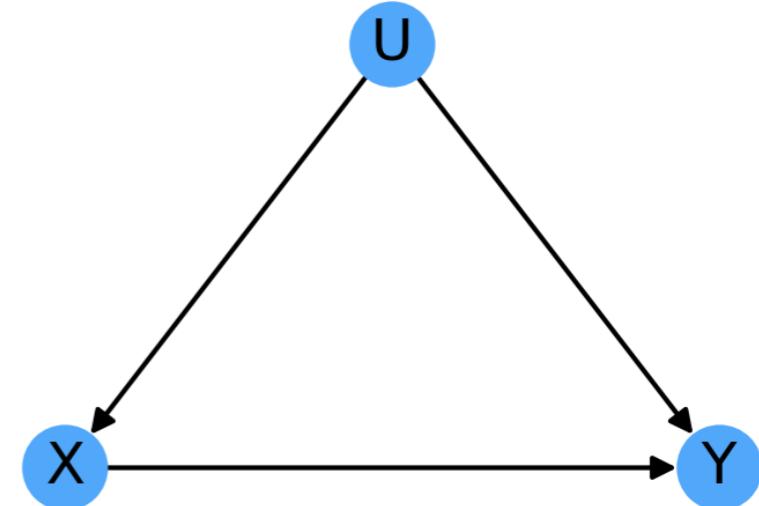
$$X = aU$$

$$Y = bX + U$$

Counterfactuals

- To compute $Y_x(u)$ we start by plugging in the value $X = x$ to obtain: $Y = bx + U$
- Finally, plugging in $U = u$, we obtain:

$$Y_x(u) = bx + u$$



- We can also compute $X_y(u)$ by plugging in $Y = y$ and $X = aU$ solving, we obtain

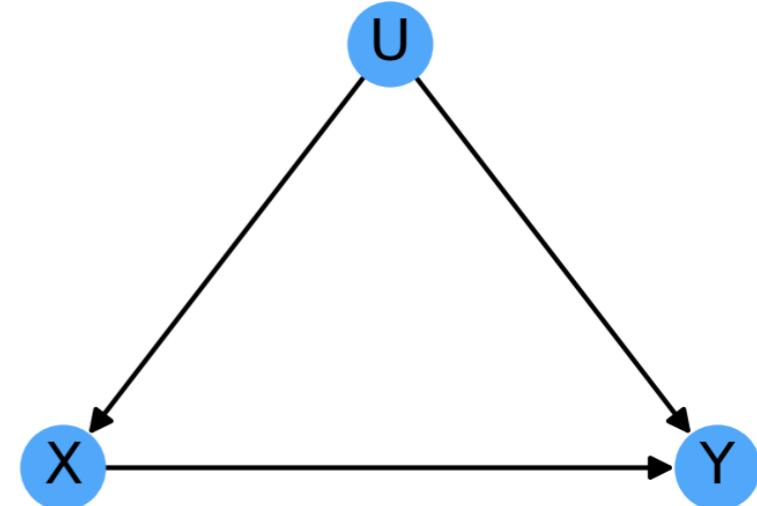
$$X_y(u) = au$$

- As might have been expected from the fact that X does not depend on Y explicitly

Counterfactuals

- To compute $Y_x(u)$ we start by plugging in the value $X = x$ to obtain: $Y = bx + U$
- Finally, plugging in $U = u$, we obtain:

$$Y_x(u) = bx + u$$



- We can also compute $X_y(u)$ by plugging in $Y = y$ and $X = aU$ solving, we obtain

$$X_y(u) = au$$

- As might have been expected from the fact that X does not depend on Y explicitly

Counterfactual - Consider a pair of variables X, Y . Let M_x represent the modified version of the SCM M , where the equation for X is replaced by the value x . A counterfactual is defined as

$$Y_x(u) = Y_{M_x}(u)$$

Steps to Compute Counterfactuals

- The procedure requires 3 steps:
 1. **Abduction** - Use evidence $\mathbf{E} = \mathbf{e}$ to determine the value of \mathbf{U}
 2. **Action** - Modify the model \mathbf{M} , by setting $\mathbf{X} = \mathbf{x}$ to obtain \mathbf{M}_x
 3. **Prediction** - Use \mathbf{M}_x and \mathbf{U} to compute $\mathbf{Y}_x(\mathbf{u})$

Steps to Compute Counterfactuals

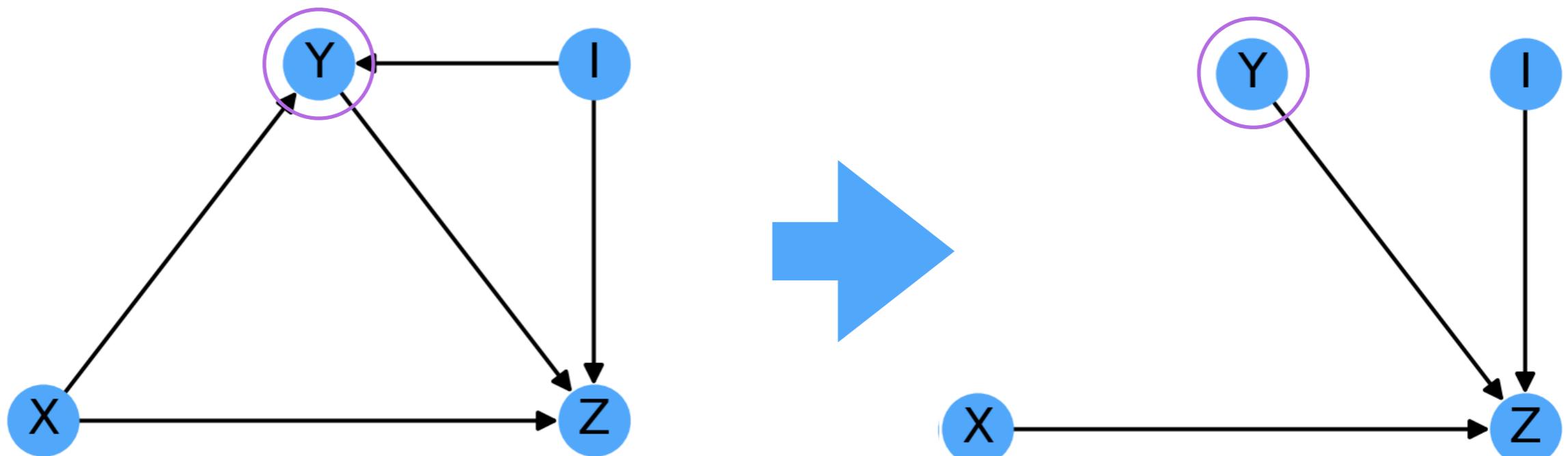
- The procedure requires 3 steps:

1. **Abduction** - Use evidence $E = e$ to determine the value of U

2. **Action** - Modify the model M , by setting $X = x$ to obtain M_x

3. **Prediction** - Use M_x and U to compute $Y_x(u)$

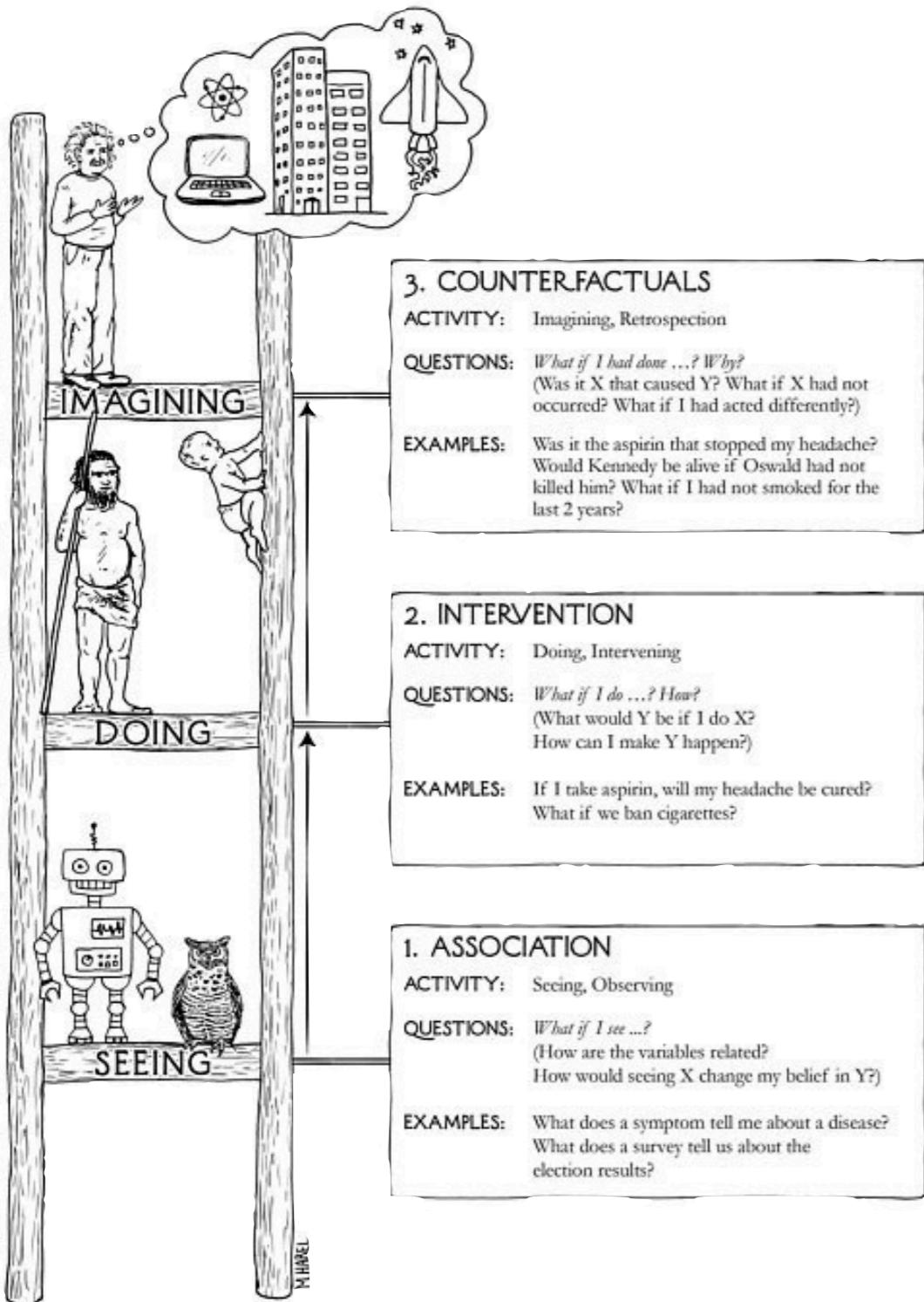
- Step 2 is similar to an intervention



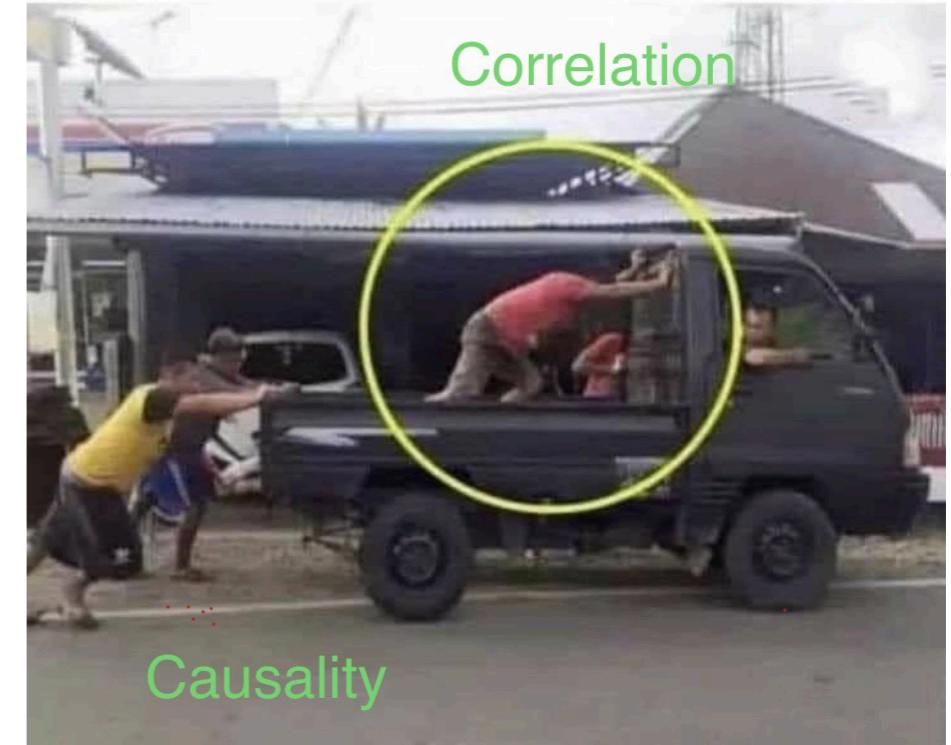
Causation Vs Correlation - Revisited



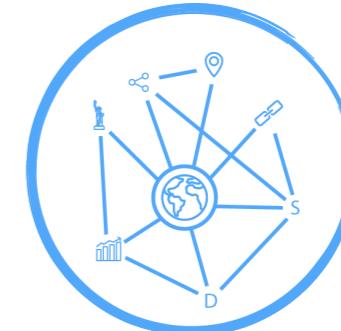
ASCEND THE LADDER OF CAUSATION



- You now have a basic understanding of the theoretical framework needed to properly disentangle pure correlations from causations and the empirically determine causal effects using empirical data!



Events



www.data4sci.com/newsletter



<http://paypal.me/data4sci>



Deep Learning for Everyone

<https://bit.ly/2UnPXHC>

Nov 19, 2020 - 5am-9am (PST)

Time Series for Everyone

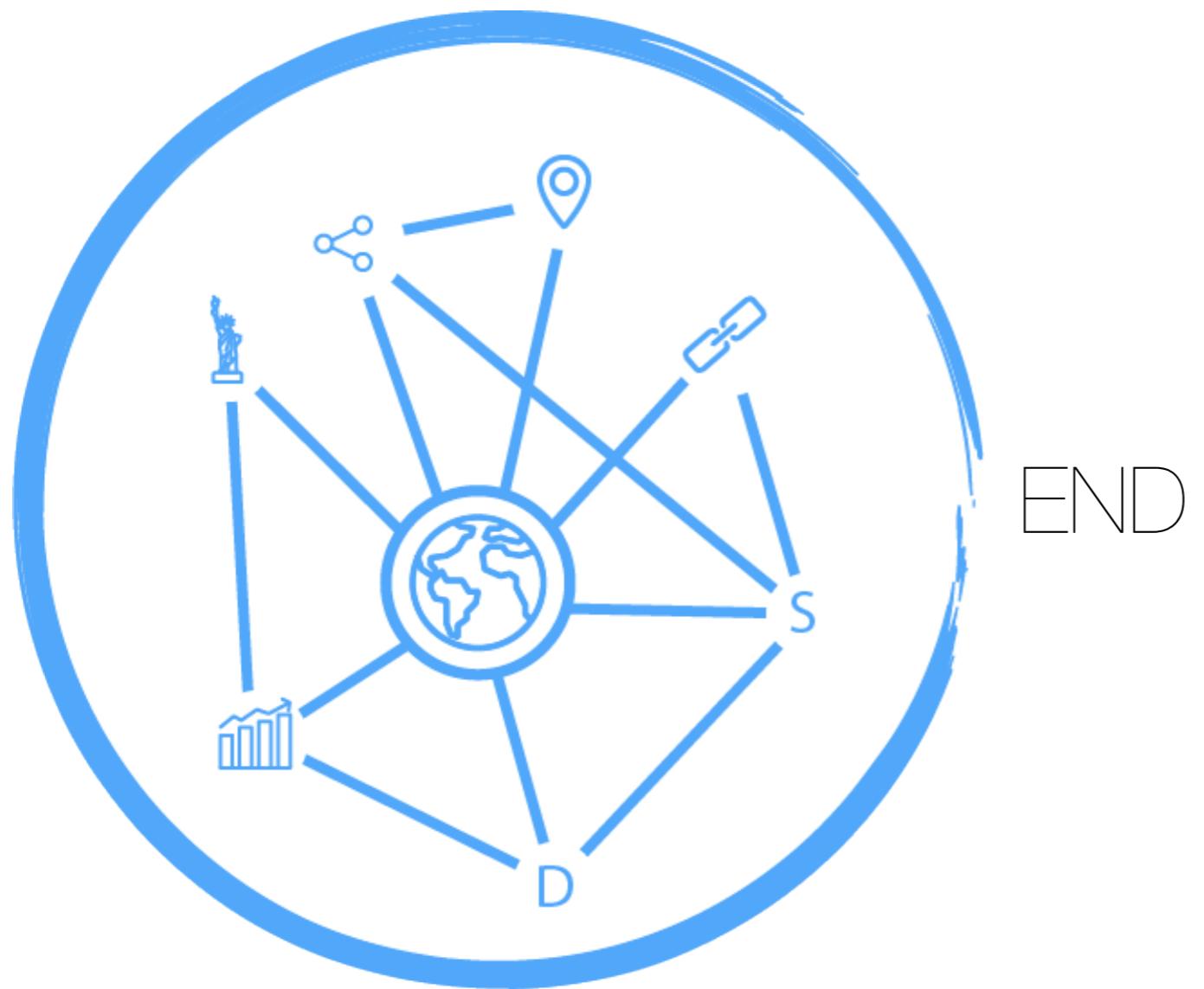
<https://bit.ly/3IsMMdN>

Dec 04, 2020 - 5am-9am (PST)

Advanced Time Series for Everyone

<https://bit.ly/32CDfJA>

Dec 11, 2020 - 5am-9am (PST)



END