Why and What If? Causal Inference for Everyone

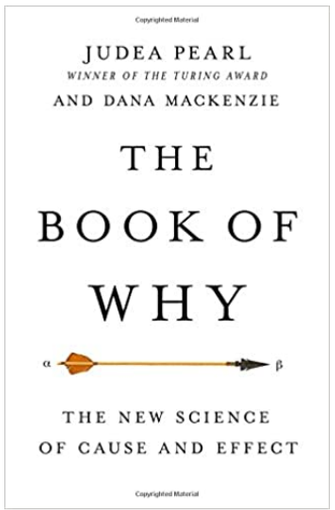
Bruno Gonçalves

www.data4sci.com/newsletter

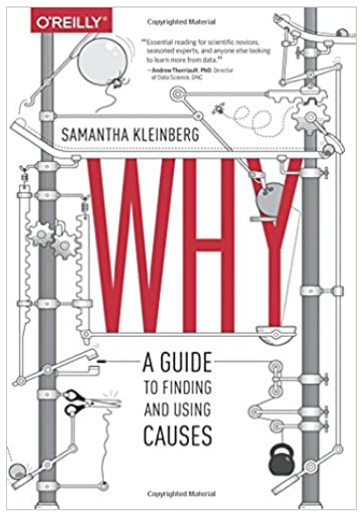
https://github.com/DataForScience/CausalInference

Bibliography

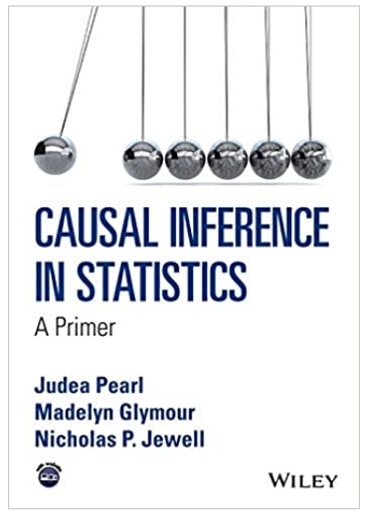
https://github.com/DataForScience/CausalInference



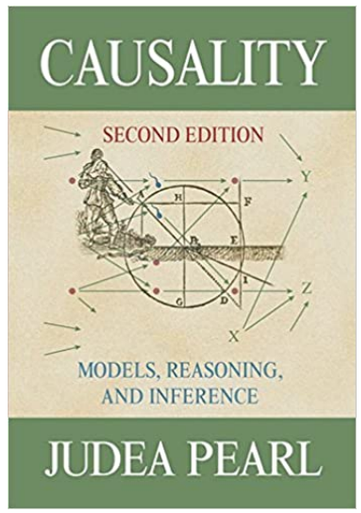
https://amzn.to/34bOF8N



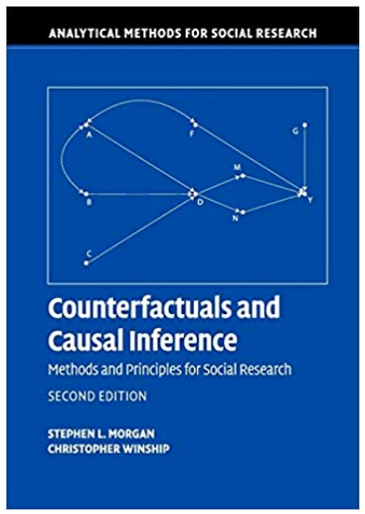
https://amzn.to/323uK9n



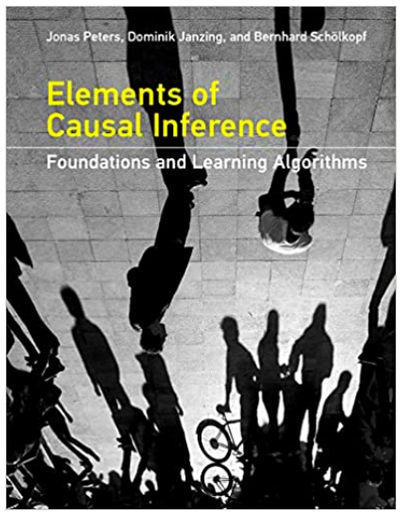
https://amzn.to/3iS73aP



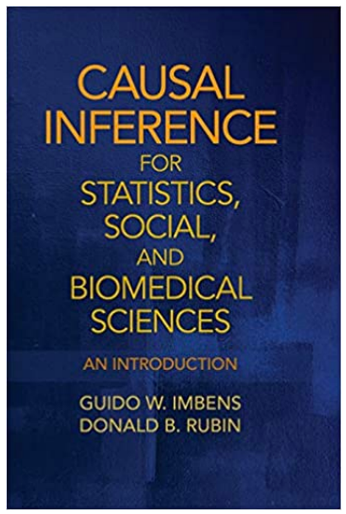
https://amzn.to/3iW2I6m



https://amzn.to/34clQZV



https://amzn.to/34bvmMQ

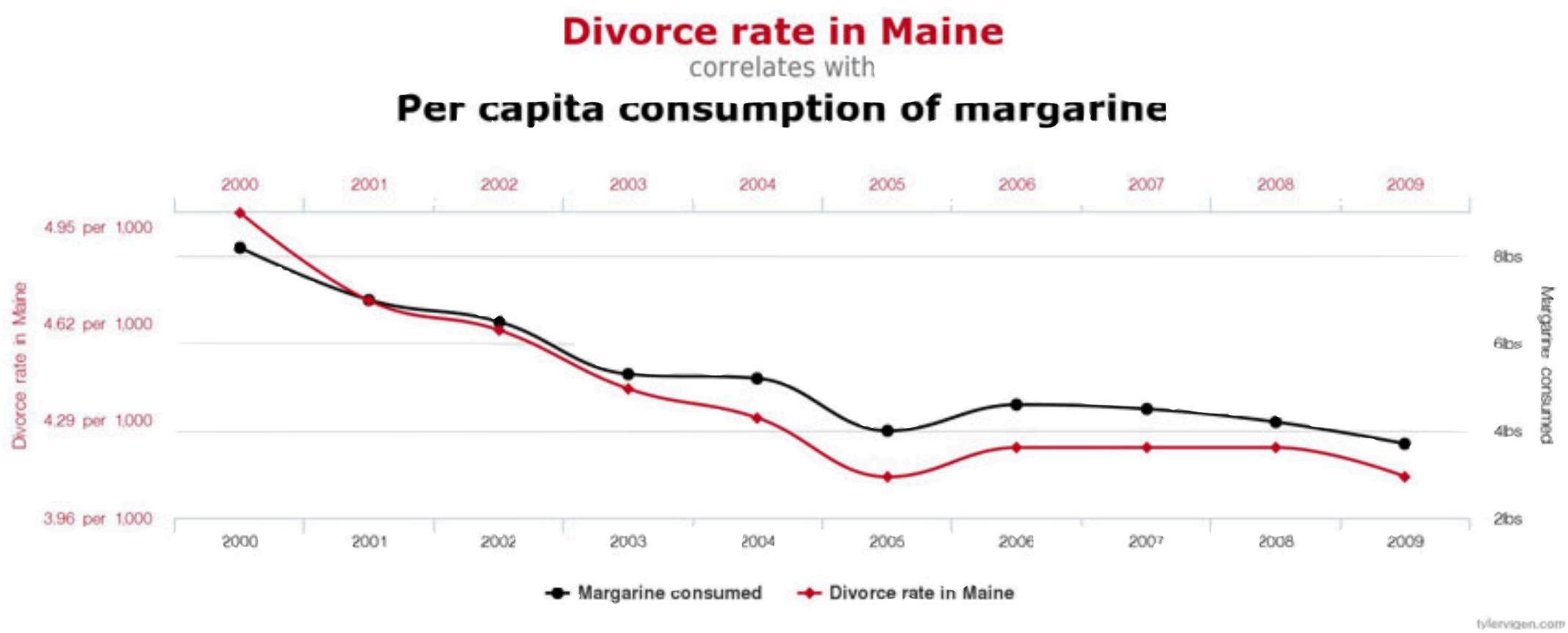


https://amzn.to/34dlwJ5

Table of Contents

1. Approaches to Causality

Correlations is not Causation



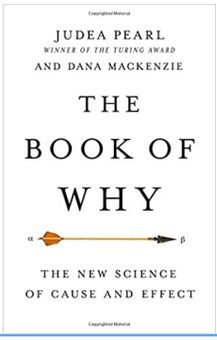
What is Causation?

Causation

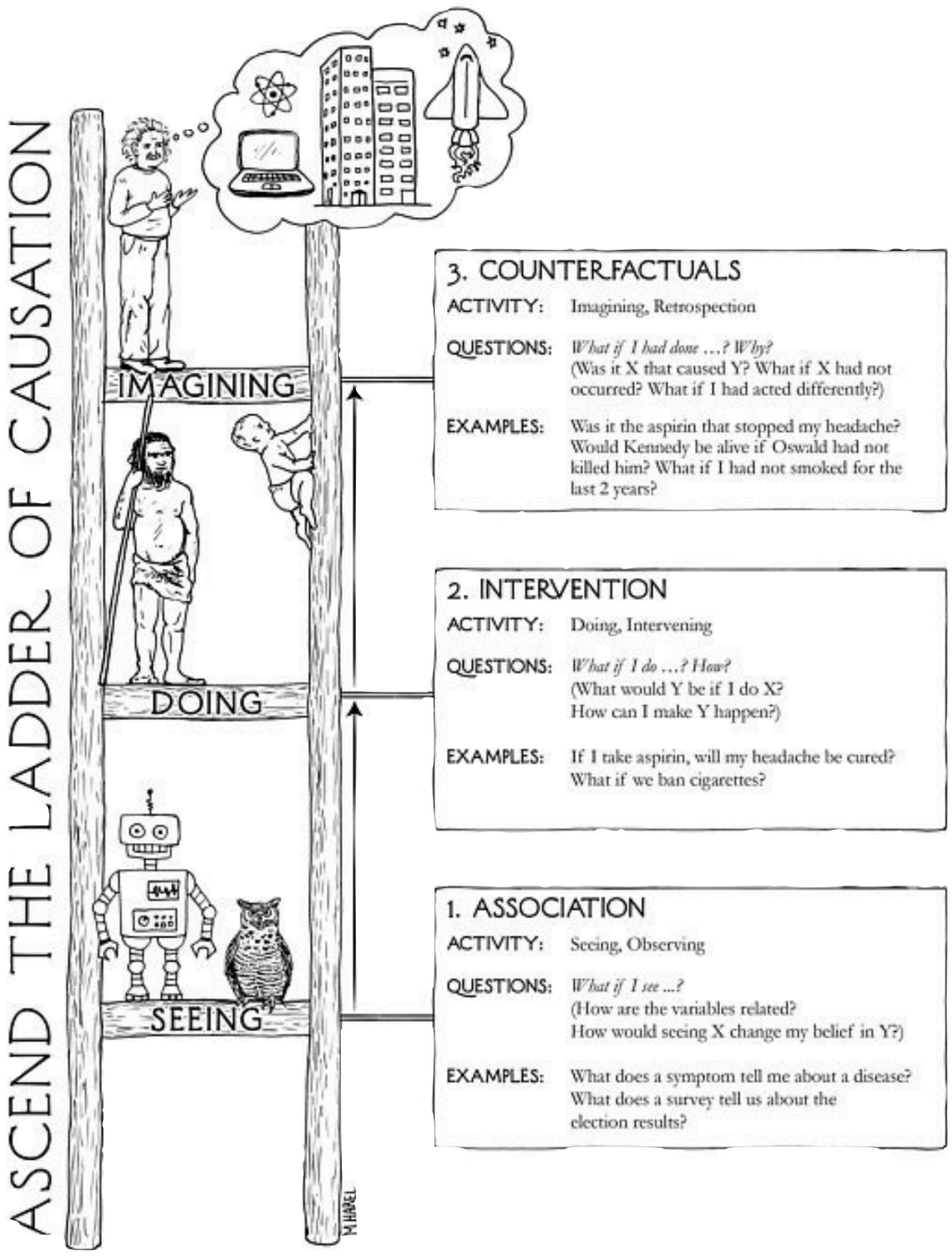
"More has been learned about Causal Inference in the last few decades than the sum total of everything that had been learned about it in all prior recorded history" (G. King)

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

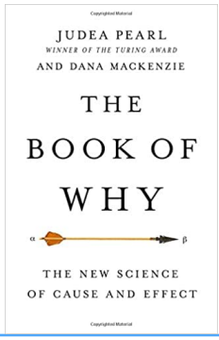
Ladder of Causality



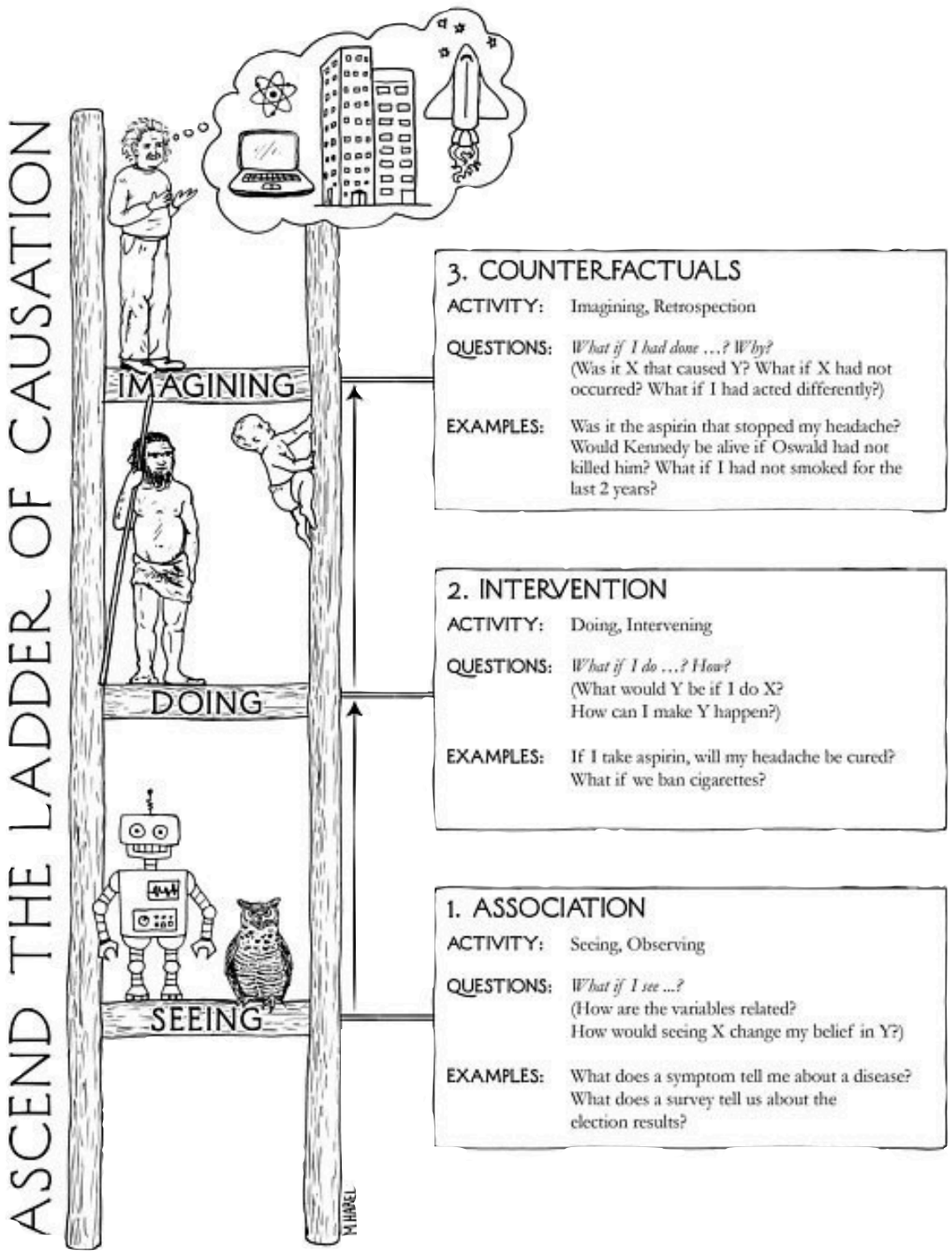
https://amzn.to/34bOF8N



Ladder of Causality



https://amzn.to/34bOF8N



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?

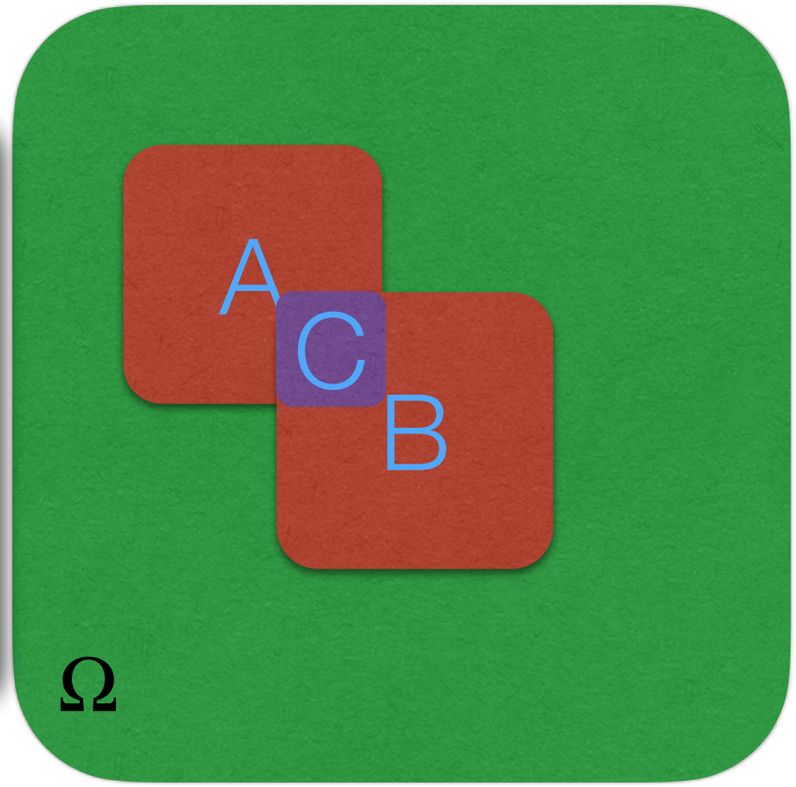
Probability Theory

Probability = Area

Prob(A) = Area A

Total Area = 1

Prob(A or B) = Area A + Area B - Area C



$$0 \leq P(A) \leq 1$$

$$P(\Omega) \equiv 1$$

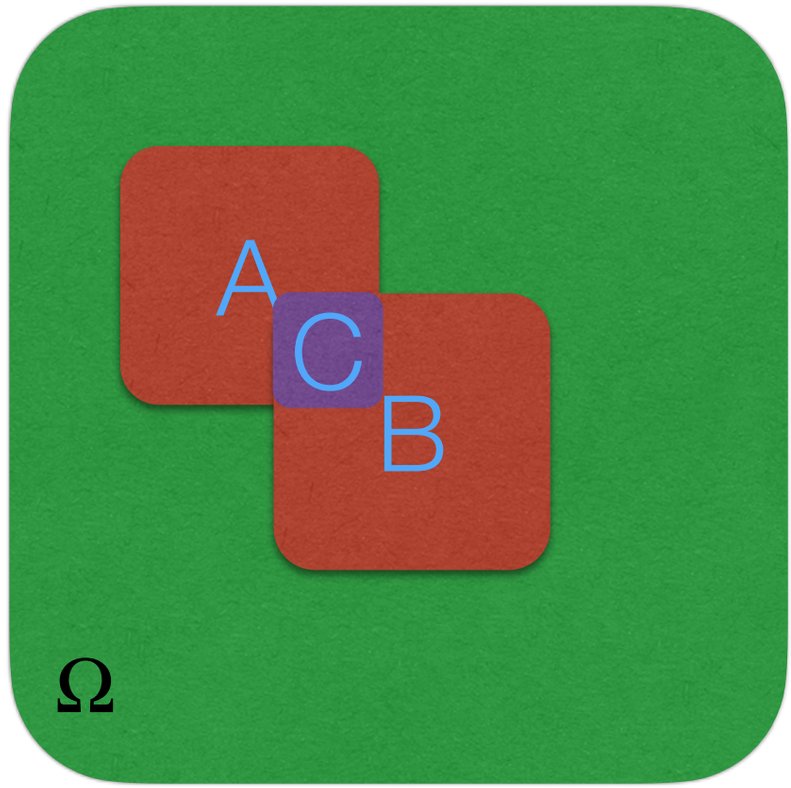
$$P(A \ or \ B) = P(A) + P(B) - P(C)$$

$$P(C) = P\left(A \ and \ B\right) = \text{overlap of A and B}$$

What's the probability that I'm in B given that I'm in A?

Conditional Probability

- What is the Probability of A given B?
- What is the probability that I'm in A if I know that I'm in B?
- Normalize the area of the overlap (C) by the area you're conditioning on (B).
- The conditional probability of A given B is defined as:



$$P\left(A\,|\,B\right) = \frac{P\left(C\right)}{P\left(B\right)}$$

- Which we can rewrite as:

$$P\left(A\,|\,B\right) = \frac{P\left(A,B\right)}{P\left(B\right)}$$

- As $P\left(C\right)$ is simply the overlap of A and B. Similarly for $P\left(B\,|\,A\right)$ we have:

$$P\left(B\,|\,A\right) = \frac{P\left(A,B\right)}{P\left(A\right)}$$

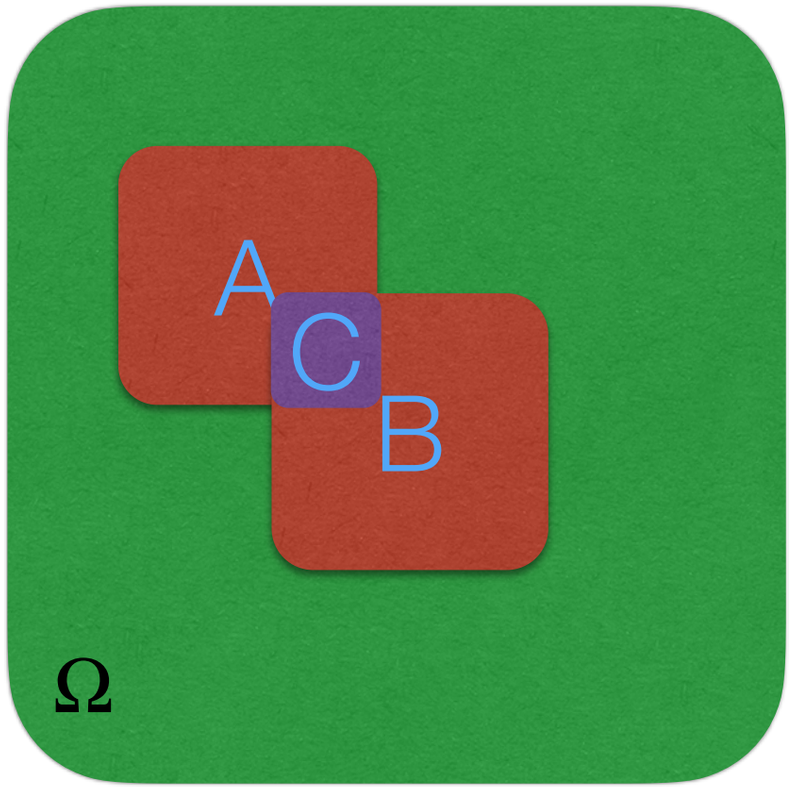
Bayes Theorem

- We can combine these two expressions to obtain:

$$P\left(A\,|\,B\right)P\left(B\right) = P\left(B\,|\,A\right)P\left(A\right)$$

- And finally:

$$P\left(A\,|\,B\right) = \frac{P\left(B\,|\,A\right)P\left(A\right)}{P\left(B\right)}$$



- This is known as Bayes' Theorem the basis of a whole new branch of statistics, Bayesian Statistics

- The process of conditioning reflects the inclusion of added information. You can Bayes' Theorem as a way of updating your belief about a given situation in the presence of new information.

Bayes Theorem - Terminology

Likelihood

Posterior

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Prior

Evidence

Bayes Theorem - Terminology

Likelihood
Prob Data Given the Hypothesis

Posterior
Prob Hypothesis
Given the Data

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Prior
Prob Hypothesis

Evidence
Prob Data

Medical Tests

Your doctor thinks you might have a rare disease that affects 1 person in 10,000. A test that is 99% accurate comes out positive. What's the probability of you having the disease?

Bayes Theorem:

$$P\left(disease|positive\ test\right) = \frac{P\left(positive\ test|disease\right)P\left(disease\right)}{P\left(positive\ test\right)}$$

Total Probability:

$$P\left(positive\ test\right) = P\left(positive\ test|disease\right)P\left(disease\right) + P\left(positive\ test|no\ disease\right)P\left(no\ disease\right)$$

Finally:

$$P\left(disease|positive\ test\right) = 0.0098$$

Medical Tests

Your doctor thinks you might have a rare disease that affects 1 person in 10,000. A test that is 99% accurate comes out positive. What's the probability of you having the disease?

Bayes Theorem:

$$P\left(disease \,|\, positive\ test\right) = \frac{P\left(positive\ test \,|\, disease\right) P\left(disease\right)}{P\left(positive\ test\right)}$$

Total Probability:

$$P\left(positive\ test\right) = P\left(positive\ test \,|\, disease\right) P\left(disease\right) + P\left(positive\ test \,|\, no\ disease\right) P\left(no\ disease\right)$$

Finally:

$$P\left(disease \,|\, positive\ test\right) = 0.0098$$

Base Rate Fallacy

Low Base Rate Value
+
Non-zero False Positive Rate

Medical Tests

Consider a population of 1,000,000 individuals. The numbers we should expect are:

	Disease	No Disease
Positive	99	9,999
Negative	1	989,901

$$P\left(disease \,|\, positive\ test\right) = \frac{TP}{TP + FP} = 0.0098$$

$$P\left(disease \,|\, negative\ test\right) = \frac{TP}{TN + FN} = 0.99999$$

Medical Tests

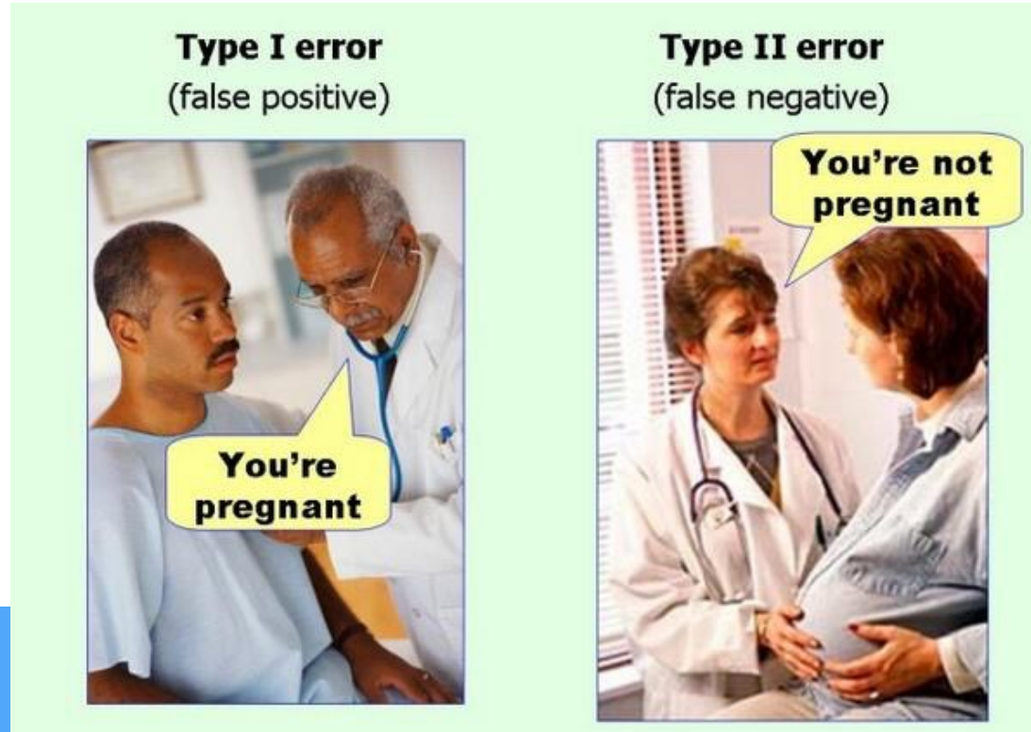
Consider a population of 1,000,000 individuals. The numbers we should expect are:

Marginals

	Disease	No Disease	
Positive	99	9,999	10,098
Negative	1	989,901	989,902
Marginals →	100	999,900	

$$P\left(disease \,|\, positive\ test\right) = \frac{TP}{TP + FP} = 0.0098$$

$$P\left(disease \,|\, negative\ test\right) = \frac{TP}{TN + FN} = 0.99999$$



A second Test

Bayes Theorem still looks the same:

$$P\left(disease|positive\ test\right) = \frac{P\left(positive\ test|disease\right)P\left(disease\right)}{P\left(positive\ test\right)}$$

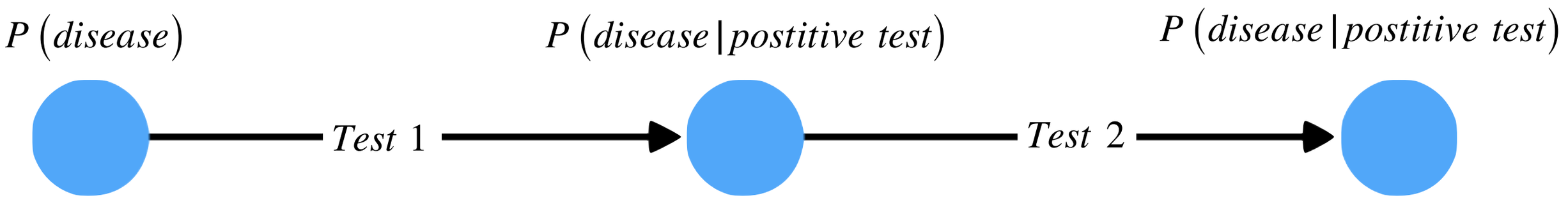
but now the probability that we have the disease has been updated:

$$P^{\dagger}\left(disease\right) = 0.0098$$

So this time we find:

$$P\left(disease|positive\ test\right) = 0.4949$$

Each test is providing new evidence, and Bayes theorem is simply telling us how to use it to update our beliefs.



Randomized Controlled Trial

https://en.wikipedia.org/wiki/Randomized_controlled_trial

- The classical question we are trying to answer is: Does my new treatment have an actual effect?
- Each patient either gets the medication or doesn't. No redos!
- How can we make sure the outcome doesn't depend on the individual patients?

Randomized Controlled Trial

https://en.wikipedia.org/wiki/Randomized_controlled_trial

- The classical question we are trying to answer is: Does my new treatment have an actual effect?
- Each patient either gets the medication or doesn't. No redos!
- How can we make sure the outcome doesn't depend on the individual patients?

Randomly Assign Patients

- Make sure that no specific individual characteristic determines the group (treatment/placebo) a specific patient is assigned to.
- Compare the outcomes

Hypothesis Testing

- Our hypothesis is that our intervention is effective
- The null-hypothesis is that there is no effect
- The main goal of Hypothesis Testing is to determine under what circumstances we can reject the null-hypothesis with a certain degree of certainty?
- In other words: How sure are we that we're not observing this difference just by chance (due to fluctuations as per the CLT)
- Select an appropriate test statistic to compare the two approaches

Hypothesis Testing

- In the case of binary outcomes, conversions follow a binomial distribution and the test statistic is the Z score:

$$Z = \frac{p_A - p_B}{\sqrt{SE_A^2 + SE_B^2}}$$

- where:

$$SE = \sqrt{\frac{p\left(1-p\right)}{N}}$$

- is the standard error for each instance.

- Under common assumptions, Z follows a Gaussian (normal) distribution centered at zero and with width one.

$$\mathcal{N}\left(0,1\right)$$

- Let's consider a practical example to clarify things

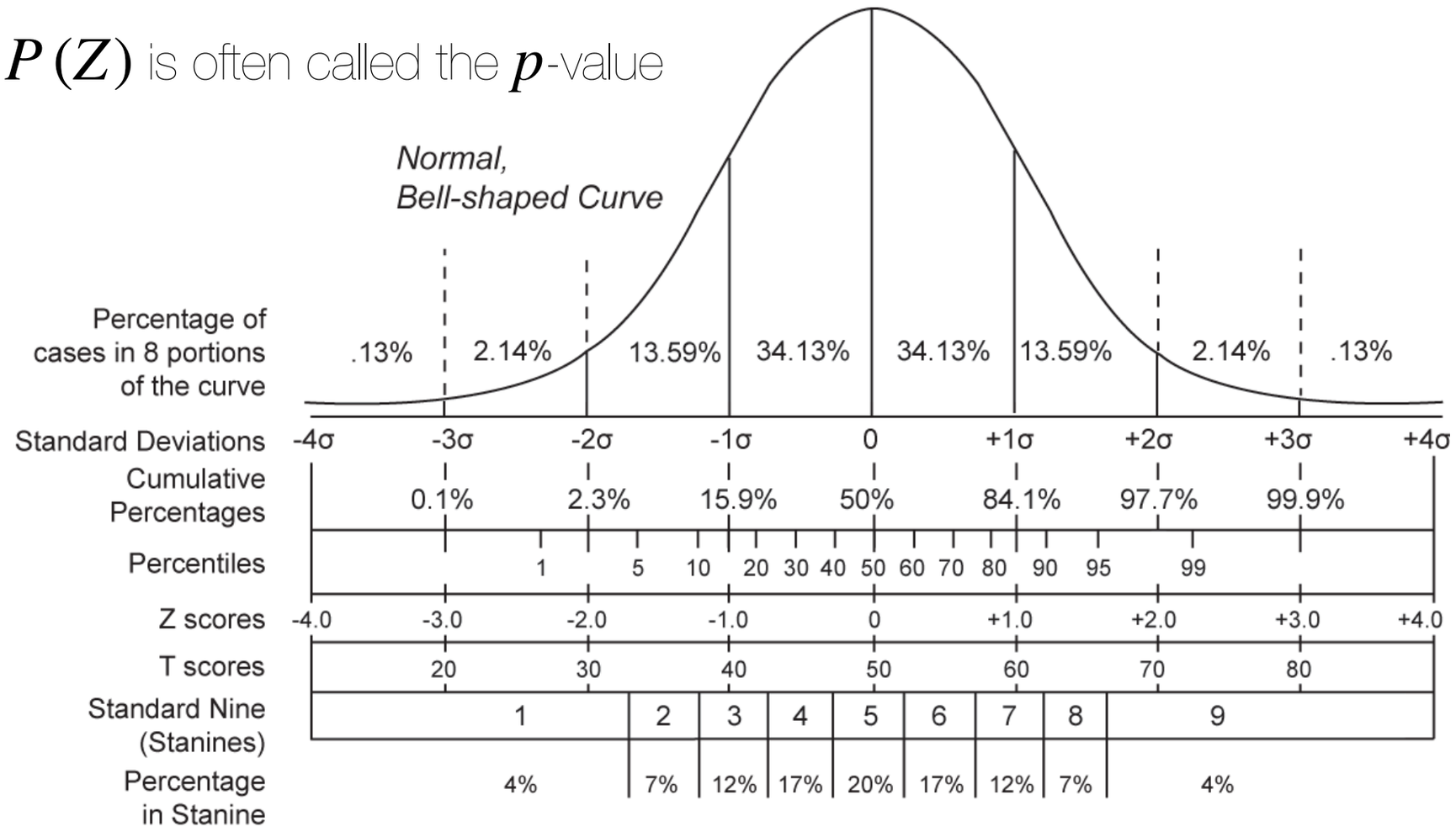
A/B Testing

- Which version of a headline results in more clicks?
- Divide users into two groups A and B and show each of them just one version
- Measure the click probability in each group, p_A and p_B
- The null hypothesis is that $p_A = p_B$. Can we reject it?

A/B Testing

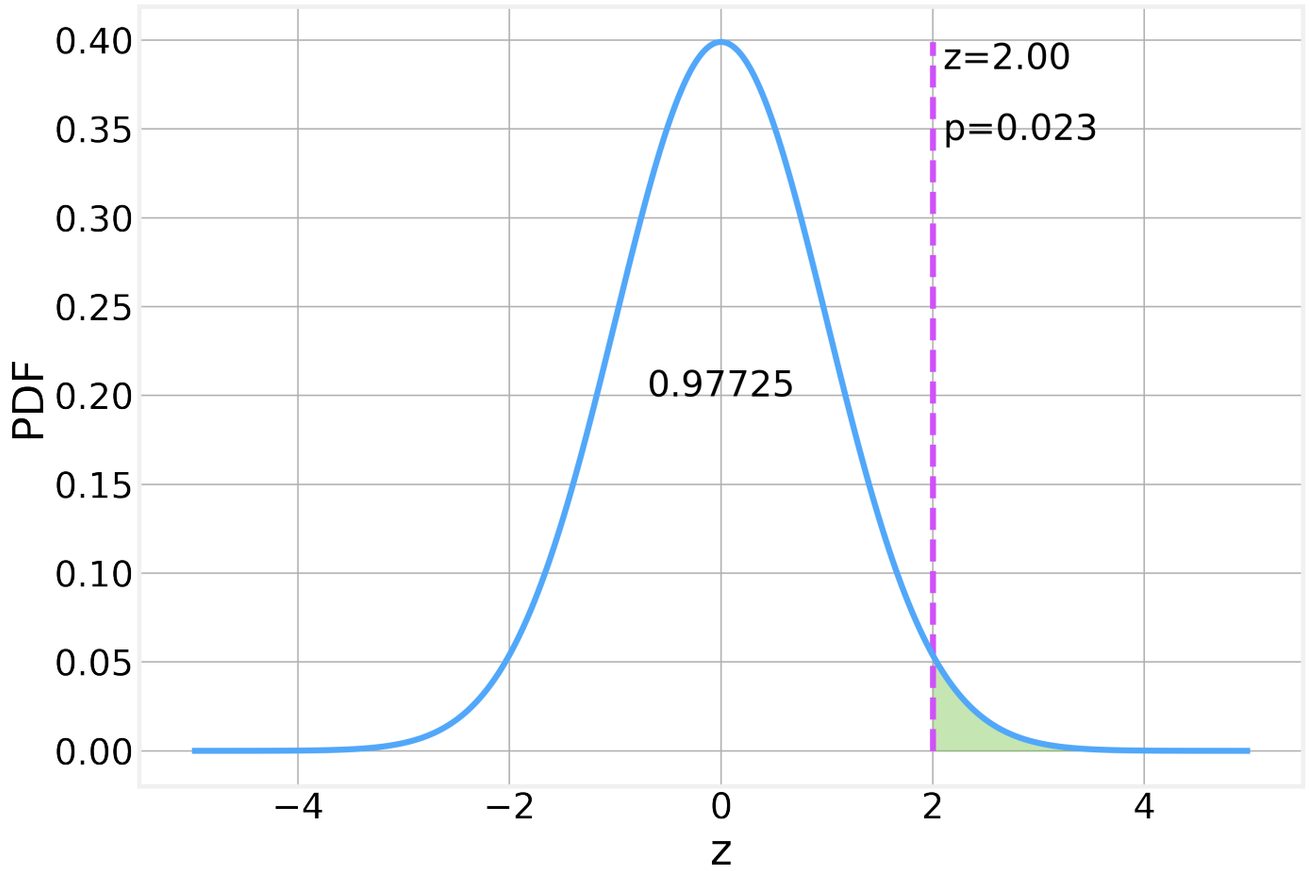
$$Z = \frac{p_A - p_B}{\sqrt{SE_A^2 + SE_B^2}}$$

- The value of $P(Z)$ effectively tells us how likely we are to observe this difference between p_A and p_B just due to sampling effects
- $P(Z)$ is often called the p-value



p-value

- Calculate the probability, p, of an event more extreme that the observation under the "null hypothesis"



- $p < 0.05$ Moderate
- $p < 0.01$ Strong
- $p < 0.001$ Very strong

evidence against the null-hypothesis

- The smaller the p-value the better.

Berkeley Discrimination Case

	Candidates	Acceptance Rate
Men	8442	0.44
Women	4321	0.35

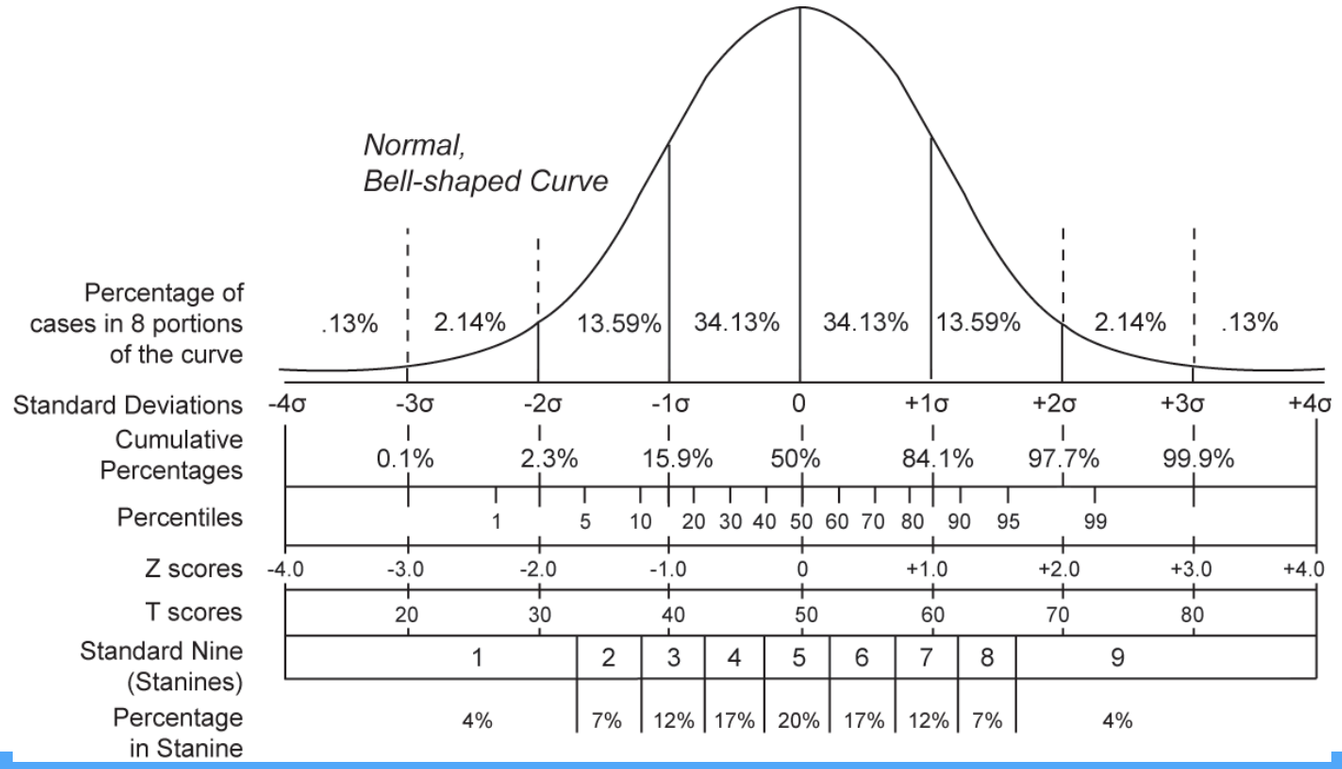
Were women being discriminated against when they applied to Berkley?

Berkeley Discrimination Case

	Candidates	Acceptance Rate	SE
Men	8442	0.44	5.4×10^{-3}
Women	4321	0.35	7.2×10^{-3}

Were women being discriminated against when they applied to Berkley?

$$Z = \frac{p_A - p_B}{\sqrt{SE_A^2 + SE_B^2}}$$



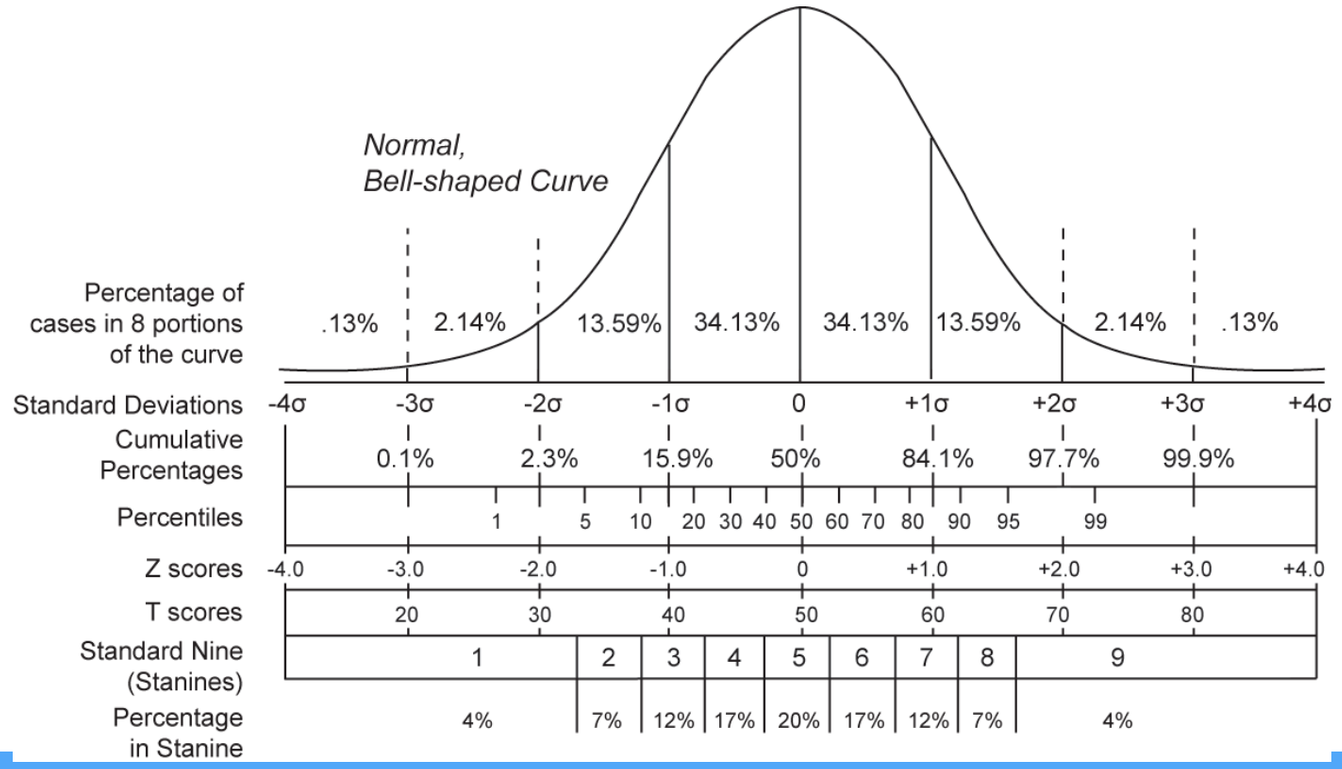
Berkeley Discrimination Case

	Candidates	Acceptance Rate	SE
Men	8442	0.44	5.4×10^{-3}
Women	4321	0.35	7.2×10^{-3}

Were women being discriminated against when they applied to Berkley?

$$Z = \frac{p_A - p_B}{\sqrt{SE_A^2 + SE_B^2}}$$

$$p \approx 10^{-23}$$



Berkeley Discrimination Case

Science 187, 398 (1975)

	Candidates	Acceptance Rate	SE
Men	8442	0.44	5.4×10^{-3}
Women	4321	0.35	7.2×10^{-3}

	Men		Women	
	Candidates	**Acceptance**	**Candidates**	**Acceptance**
A	825	0.62	108	0.82
B	560	0.63	25	0.68
C	325	0.37	594	0.34
D	417	0.33	375	0.35
E	191	0.28	393	0.24
F	272	0.06	341	0.07
Total	2590	0.46	1835	0.30

Berkeley Discrimination Case

Science 187, 398 (1975)

	Candidates	Acceptance Rate	SE
Men	8442	0.44	5.4×10^{-3}
Women	4321	0.35	7.2×10^{-3}

	Men		Women	
	Candidates	**Acceptance**	**Candidates**	**Acceptance**
A	825	0.62	108	**0.82**
B	560	0.63	25	**0.68**
C	325	0.37	594	0.34
D	417	0.33	375	0.35
E	191	**0.28**	393	0.24
F	272	**0.06**	341	0.07
Total	2590	0.46	1835	0.30

Simpson's Paradox

Science 187, 398 (1975)

	Candidates	Acceptance Rate	SE
Men	8442	0.44	5.4×10^{-3}
Women	4321	0.35	7.2×10^{-3}

	Men		Women	
	Candidates	Acceptance	Candidates	Acceptance
A	825	0.62	108	**0.82**
B	560	0.63	25	**0.68**
C	325	0.37	594	0.34
D	417	0.33	375	0.35
E	191	**0.28**	393	0.24
F	272	**0.06**	341	0.07
Total	2590	0.46	1835	0.30

Simpson's Paradox

https://en.wikipedia.org/wiki/Simpson%27s_paradox

"aggregated data can appear to reverse important trends in the numbers being combined"

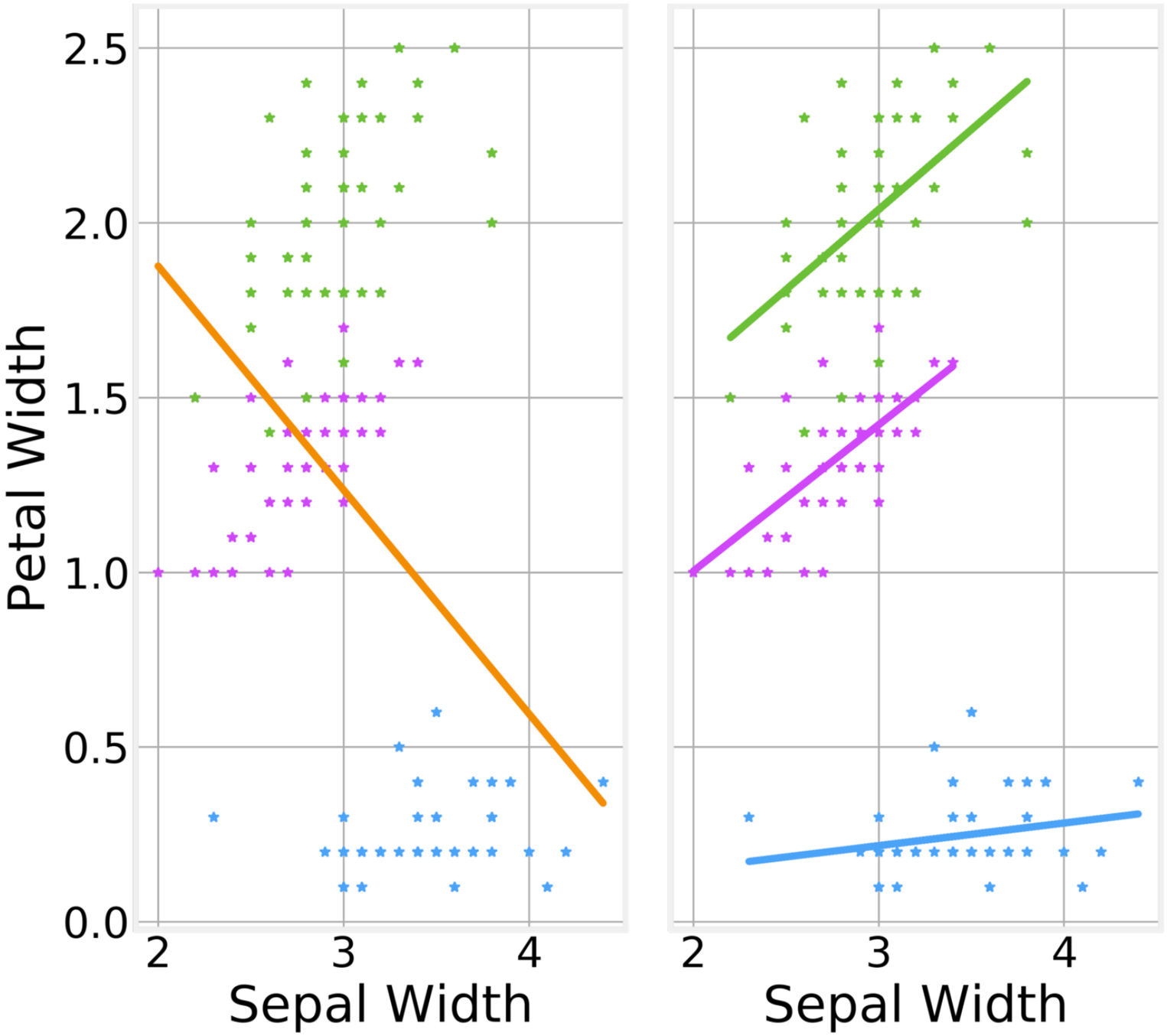
WSJ, Dec 2, 2009

- Simpson's Paradox is likely to appear whenever you have confounding factors.

- In the Berkeley Case, we had two factors:
  - Men and Women prefer different departments
  - Departments have widely varying acceptance rates
- Most women applied to departments with low acceptance rates, while most men applied to departments with high acceptance rates

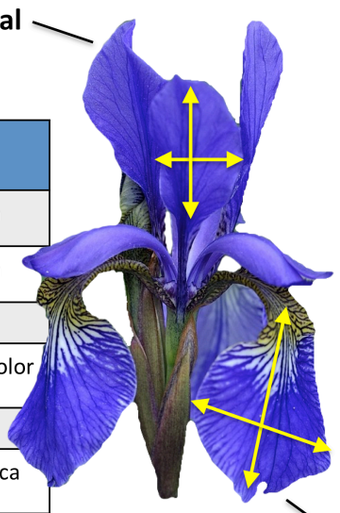
Simpson's Paradox

https://en.wikipedia.org/wiki/Simpson%27s_paradox



Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

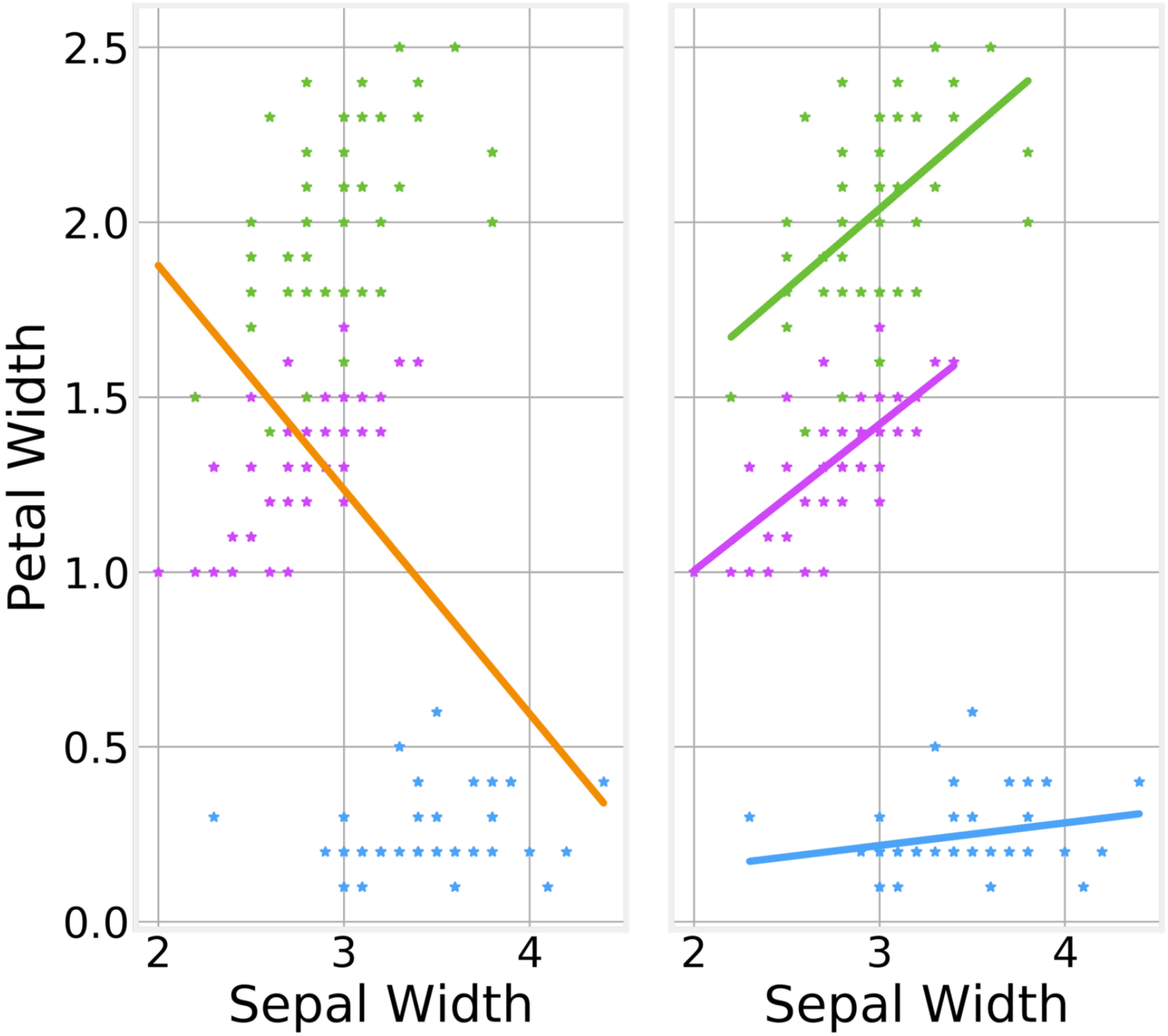


Class labels
(targets)

Features
(attributes, measurements, dimensions)

Simpson's Paradox

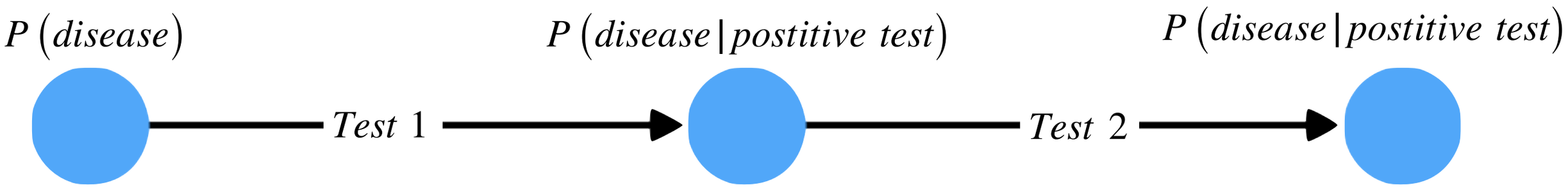
https://en.wikipedia.org/wiki/Simpson%27s_paradox



Understanding the data generating process (the history behind the data) is paramount to properly understand causal relationships

Graphical Models

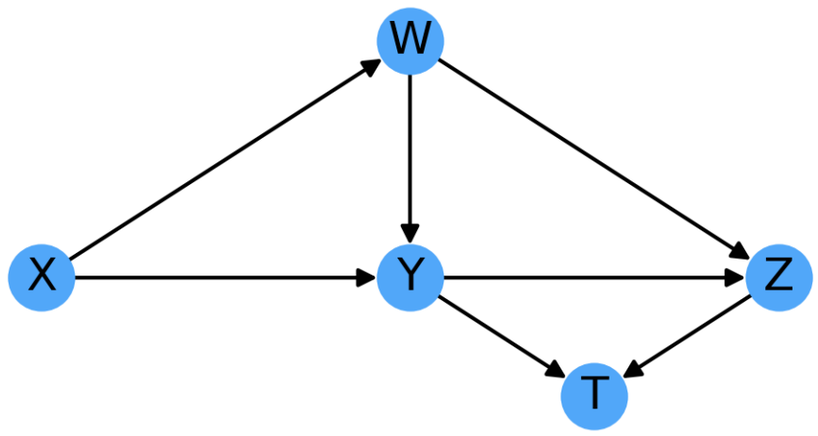
- We've already encountered our first Graphical Model:



- Graphs are constituted by Nodes (subcomponents) interconnected by edges (connections, relationships, etc)

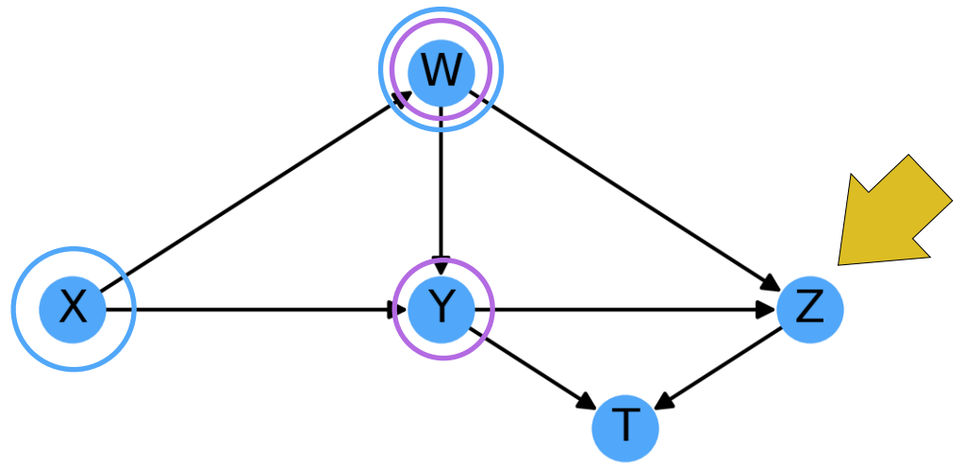
- In Causal Models:
  - Nodes represent variables of interest
  - Edges are Directed and imply a causal relationship
  - The Graph does not contain cycles: it's a Directed Acyclical Graph
- Fundamental Assumption: The value of a variable can only depend on the values of their parents that "point" to it (incoming edges) and can only influence the children they "point" to (outgoing edges)

Graphical Models



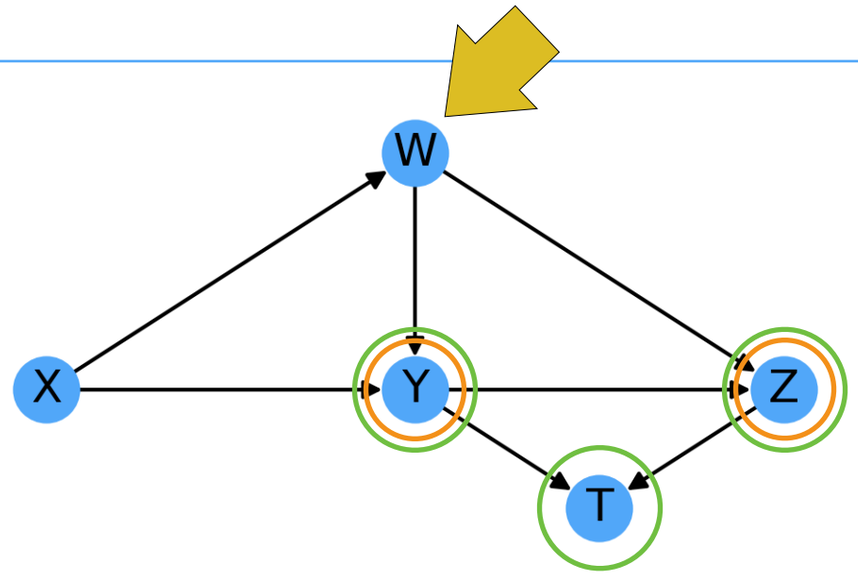
- 5 nodes: T, W, X, Y, Z
- 7 edges: [(X, W), (X, Y), (W, Y), (W, Z), (Y, Z), (Y, T), (Z, T)]

Graphical Models



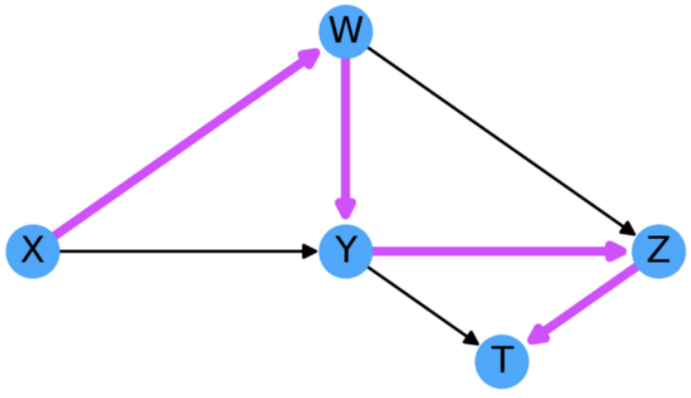
- 5 nodes: T, W, X, Y, Z
- 7 edges: [(X, W), (X, Y), (W, Y), (W, Z), (Y, Z), (Y, T), (Z, T)]
- Z has two parents, Y and W and three ancestors, Y, W and X.

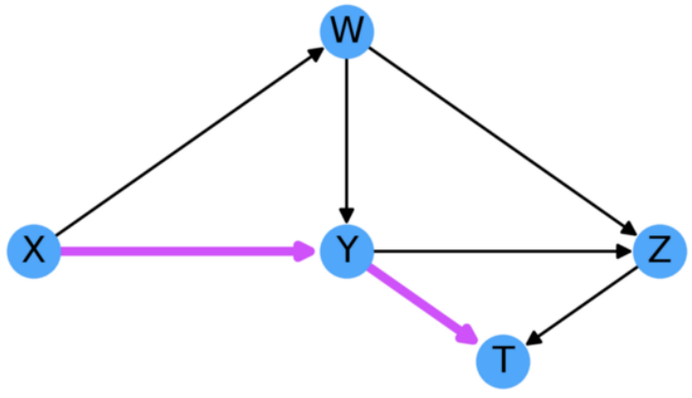
Graphical Models

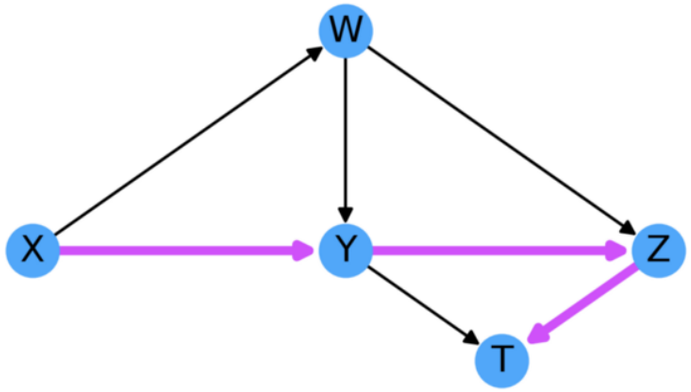


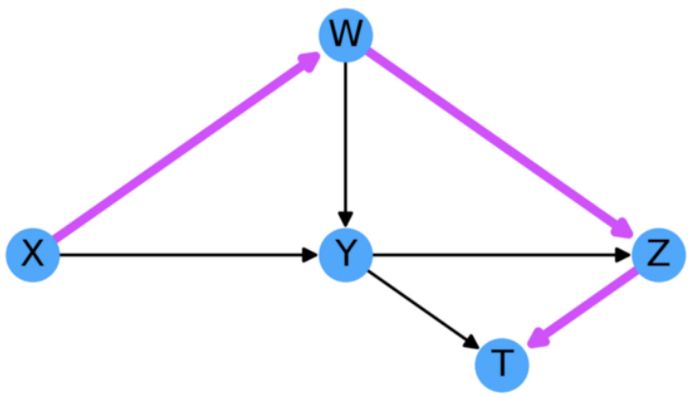
- 5 nodes: T, W, X, Y, Z
- 7 edges: [(X, W), (X, Y), (W, Y), (W, Z), (Y, Z), (Y, T), (Z, T)]
- Z has two parents, Y and W and three ancestors, Y, W and X.
- W has two children, Y and Z and three descendants, Y, Z, and T.

Graphical Models





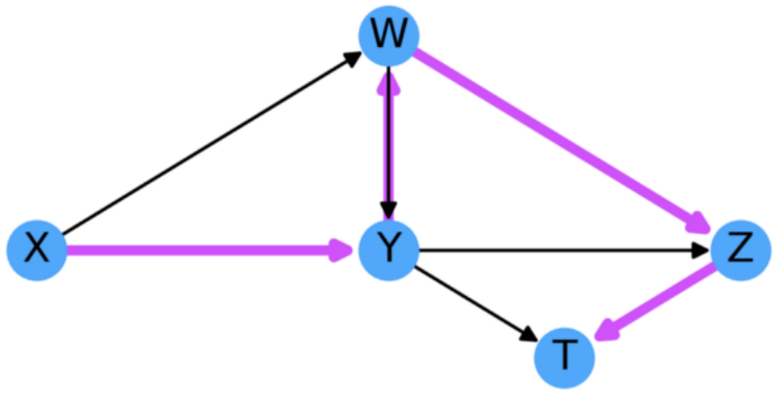


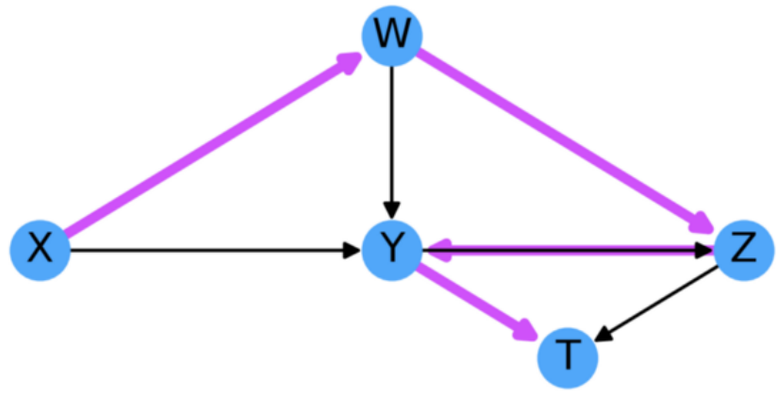


Directed paths represent ways in which one variable can directly influence another

Graphical Models

- Under certain circumstances, influence can also travel "backwards" along an edge:





Code - Approaches

https://github.com/DataForScience/CausalInference

2. Properties of Graphical Models

Structural Causal Models

- We've already seen we can use Directed Acyclical Graphs (DAG) to describe causal relationships

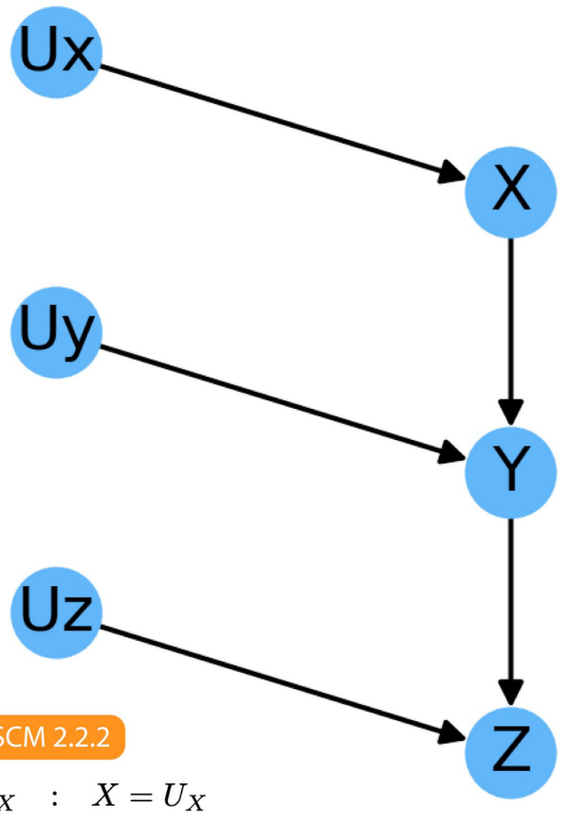
- A Structural Causal Model (SCM) is a DAGs together with a set of functions $\boldsymbol{F}$ fully specifying the form of the relationships

- 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 and set of Functions (F) that determine the values of the the variables in V based on the values of the variables in U.

- Functions (F) can be deterministic, stochastic or logical

Structural Causal Models



$$U = \{U_X, U_Y, U_Z\}$$
$$V = \{X, Y, Z\}$$
$$F = \{f_X, f_Y, f_Z\}$$

SCM 2.2.1

$$f_X : X = U_X$$
$$f_Y : Y = \frac{X}{3} + U_Y$$
$$f_Z : Z = \frac{Y}{16} + U_Z$$

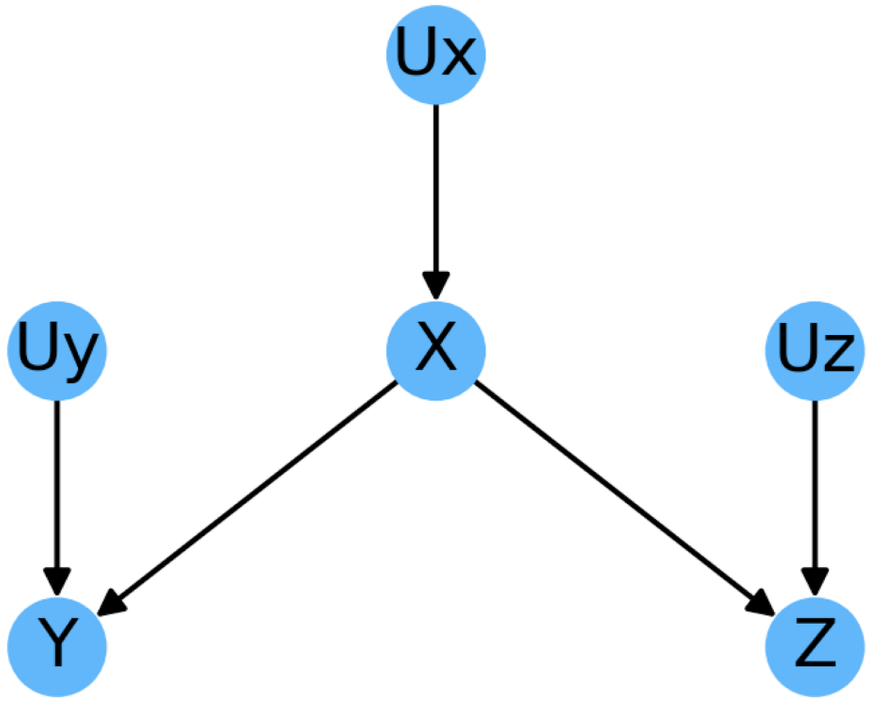
SCM 2.2.3

$$f_X : X = U_X$$
$$f_Y : Y = 84 - x + U_Y$$
$$f_Z : Z = \frac{100}{y} + U_Z$$

SCM 2.2.2

$$f_X : X = U_X$$
$$f_Y : Y = \begin{cases} \text{Closed IF (X = Up AND } U_Y = 0) \text{ OR (X = Down AND } U_Y = 1) \\ \text{Open otherwise} \end{cases}$$
$$f_Z : Z = \begin{cases} \text{On IF (Y = Closed AND } U_Z = 0) \text{ OR (Y = Open AND } U_Z = 1) \\ \text{Off otherwise} \end{cases}$$

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 2.2.5

$$f_X : X = U_X$$
$$f_Y : Y = 4x + U_Y$$
$$f_Z : Z = \frac{x}{10} + U_Z$$

SCM 2.2.6

$$f_X : X = U_X$$
$$f_Y : Y = \begin{cases} \text{On IF (X = Up AND } U_Y = 0) \text{ OR (X = Down AND } U_Y = 1) \\ \text{Off otherwise} \end{cases}$$
$$f_Z : Z = \begin{cases} \text{On IF (X = Up AND } U_Z = 0) \text{ OR (X = Down AND } U_Z = 1) \\ \text{Off otherwise} \end{cases}$$

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 $P\left(child|parents\right)$ over all the "families" in the graph, or, mathematically:

$$P\left(X_1, X_2, \cdots X_n\right) = \prod_i P\left(X_i|parents_i\right)$$

Structural Causal Models

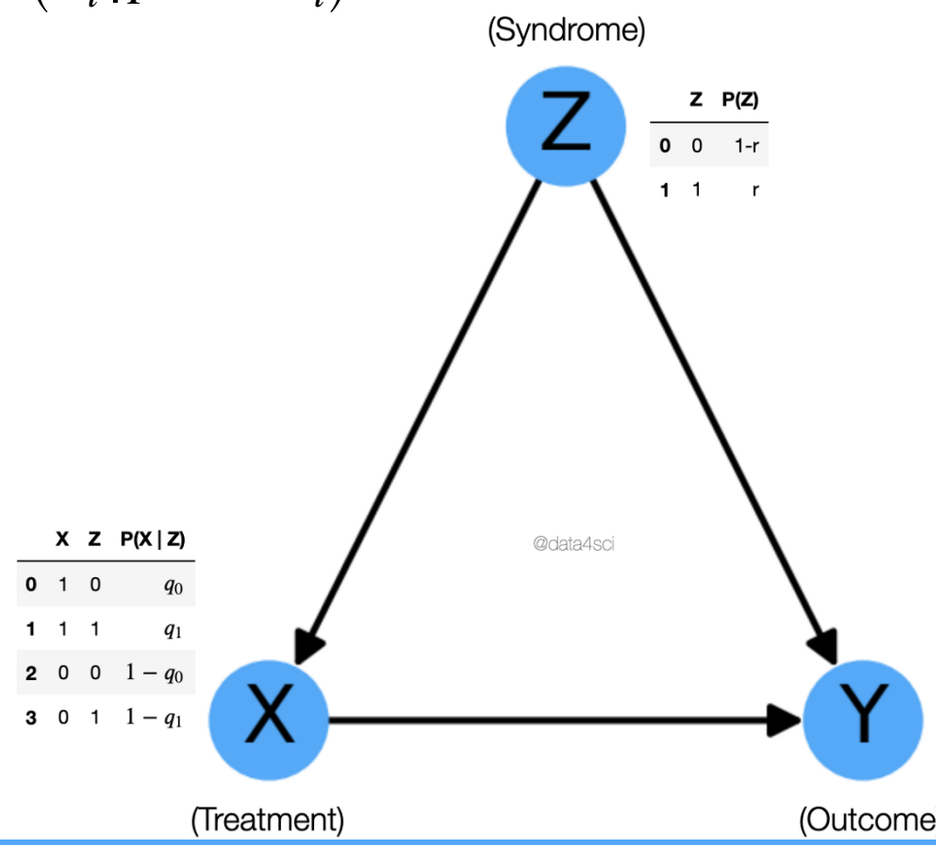
- When given the full SCM specification, we can easily generate fake data to play with. We'll often do this to explore some ideas.

- 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 $P\left(child|parents\right)$ over all the "families" in the graph, or, mathematically:

$$P\left(X_1, X_2, \cdots X_n\right) = \prod_i P\left(X_i|parents_i\right)$$

- So, if given this SCM:

Probability tables fully specify the model.



Z	P(Z)
0	1-r
1	r

X	Z	P(X | Z)
0	0	q_0
1	1	q_1
0	0	$1-q_0$
0	1	$1-q_1$

Y	X	Z	P(Y | X,Z)
1	1	0	p_{10}
1	0	0	p_{00}
1	0	1	p_{01}
1	1	1	p_{11}
0	1	0	$1-p_{10}$
0	0	0	$1-p_{00}$
0	0	1	$1-p_{01}$
0	1	1	$1-p_{11}$

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.

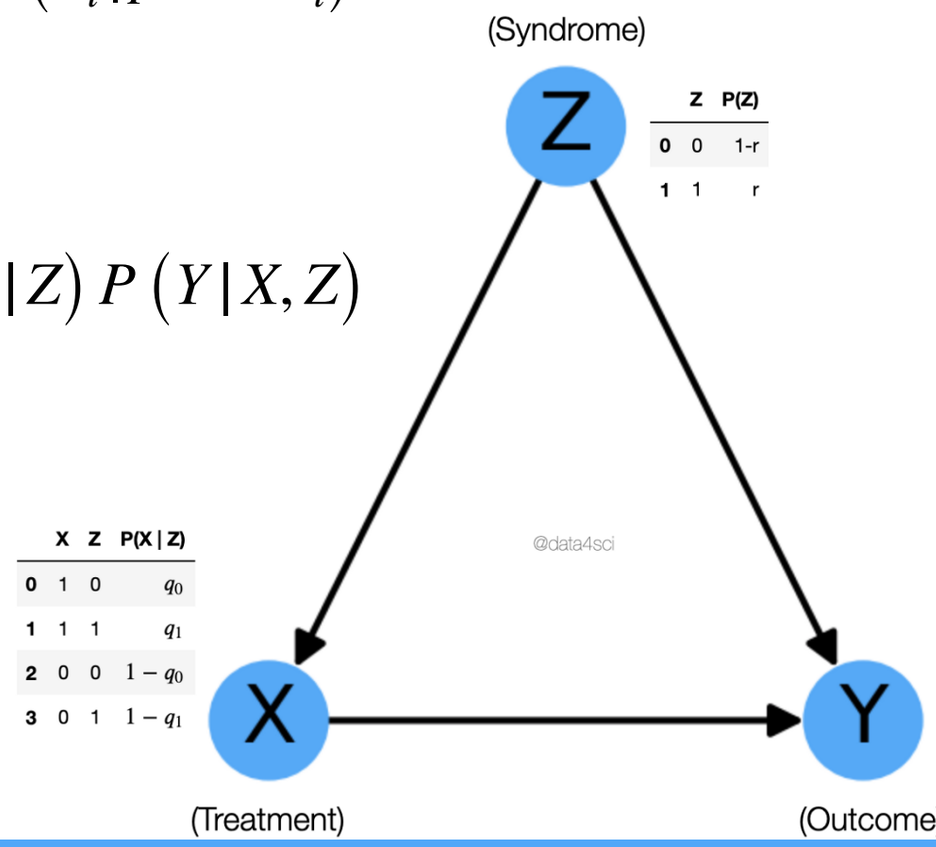
- Rule of Product decomposition: For any model whose graph is acyclic, the joint distribution of the variables in the model is given by the product of the conditional distributions $P\left(child|parents\right)$ over all the "families" in the graph, or, mathematically:

$$P\left(X_1, X_2, \cdots X_n\right) = \prod_i P\left(X_i|parents_i\right)$$

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

$$P\left(X, Y, Z\right) = P\left(Z\right) P\left(X|Z\right) P\left(Y|X, Z\right)$$

Probability tables fully specify the model.



	Z	P(Z)
0	0	1-r
1	1	r

	X	Z	P(X | Z)
0	1	0	q_0
1	1	1	q_1
2	0	0	$1-q_0$
3	0	1	$1-q_1$

	Y	X	Z	P(Y | X,Z)
0	1	1	0	p_{10}
1	1	0	0	p_{00}
2	1	0	1	p_{01}
3	1	1	1	p_{11}
4	0	1	0	$1-p_{10}$
5	0	0	0	$1-p_{00}$
6	0	0	1	$1-p_{01}$
7	0	1	1	$1-p_{11}$

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.

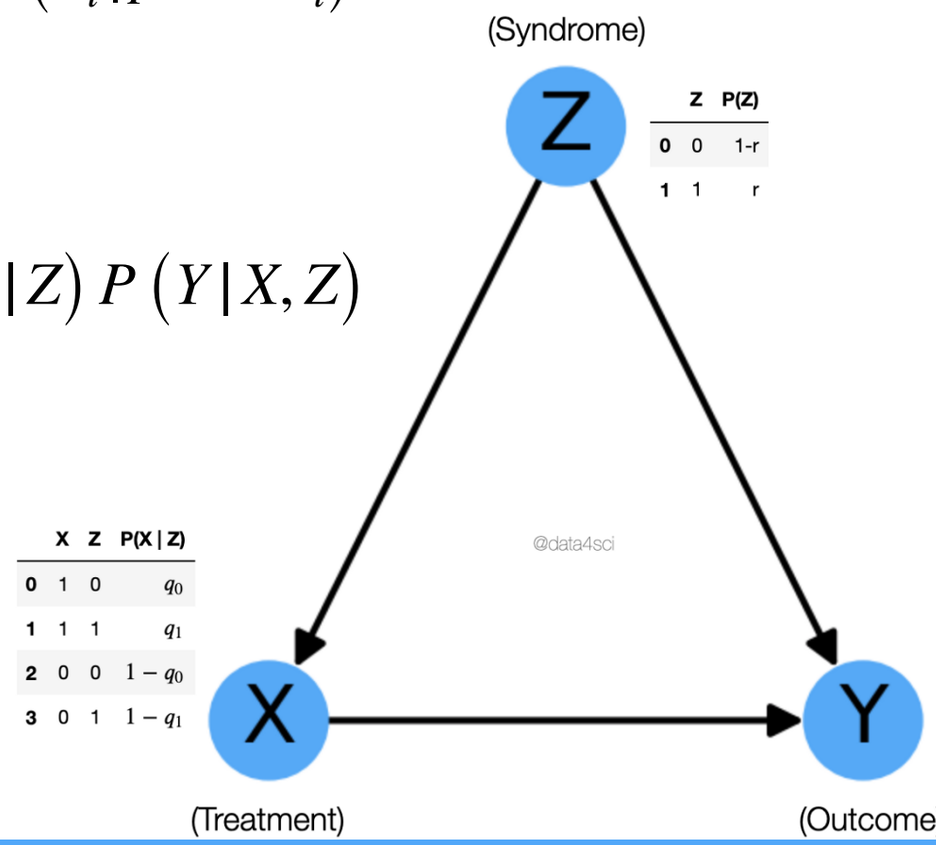
- Rule of Product decomposition: For any model whose graph is acyclic, the joint distribution of the variables in the model is given by the product of the conditional distributions $P\left(child|parents\right)$ over all the "families" in the graph, or, mathematically:

$$P\left(X_1, X_2, \cdots X_n\right) = \prod_i P\left(X_i|parents_i\right)$$

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

$$P\left(X, Y, Z\right) = P\left(Z\right) P\left(X|Z\right) P\left(Y|X, Z\right)$$

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



Z	P(Z)
0	1-r
1	r

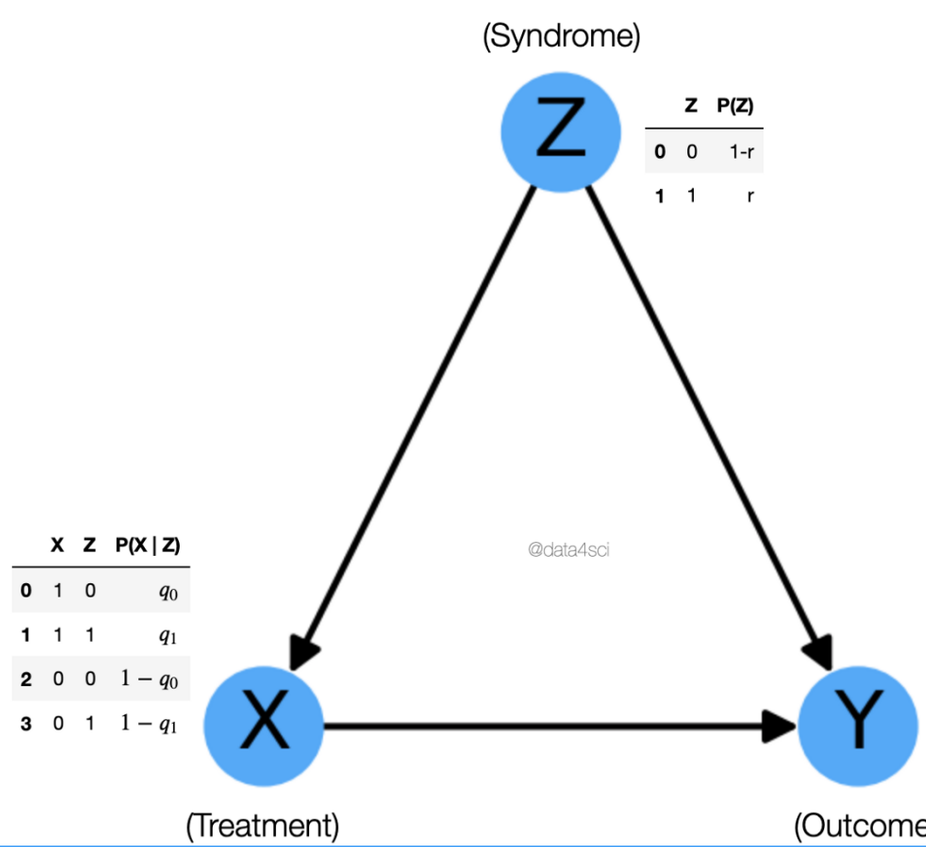
X	Z	P(X | Z)
0	0	q_0
1	1	q_1
0	0	$1-q_0$
1	0	$1-q_1$

Y	X	Z	P(Y | X,Z)
1	1	0	p_{10}
1	0	0	p_{00}
1	0	1	p_{01}
1	1	1	p_{11}
0	1	0	$1-p_{10}$
0	0	0	$1-p_{00}$
0	0	1	$1-p_{01}$
0	1	1	$1-p_{11}$

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



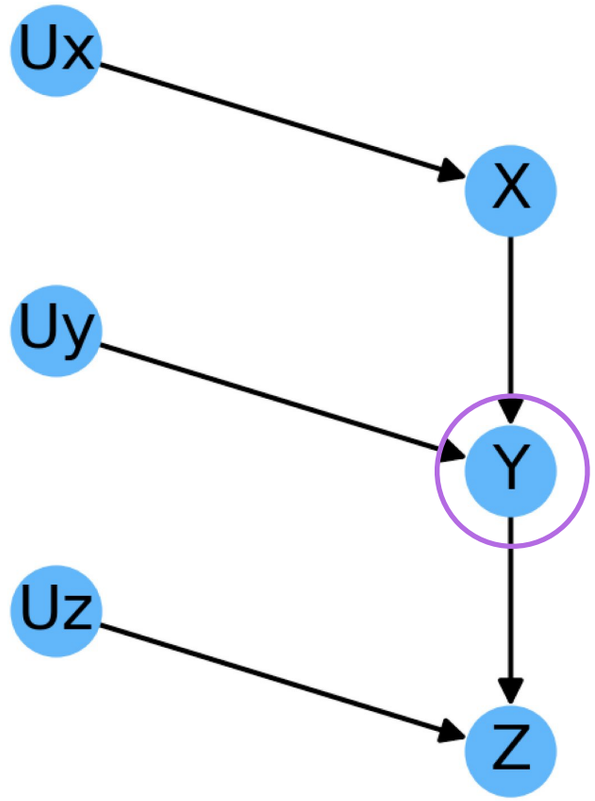
Z	P(Z)
0	1-r
1	r

X	Z	P(X | Z)
1	0	q_0
1	1	q_1
0	0	$1-q_0$
0	1	$1-q_1$

Y	X	Z	P(Y | X,Z)
1	1	0	p_{10}
1	0	0	p_{00}
1	0	1	p_{01}
1	1	1	p_{11}
0	1	0	$1-p_{10}$
0	0	0	$1-p_{00}$
0	0	1	$1-p_{01}$
0	1	1	$1-p_{11}$

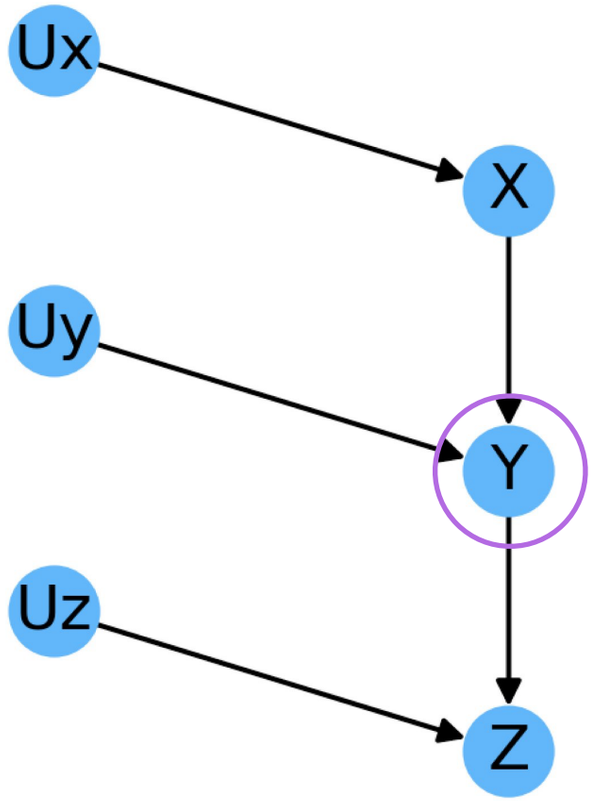
Chains

- From this Graph, we can immediately infer:
  - Z and Y are dependent — $P\left(Z|Y\right) \neq P\left(Z\right)$
  - Y and X are dependent — $P\left(Y|Z\right) \neq P\left(Y\right)$
  - Z and X are likely dependent — $P\left(Z|X\right) \neq P\left(Z\right)$



Chains

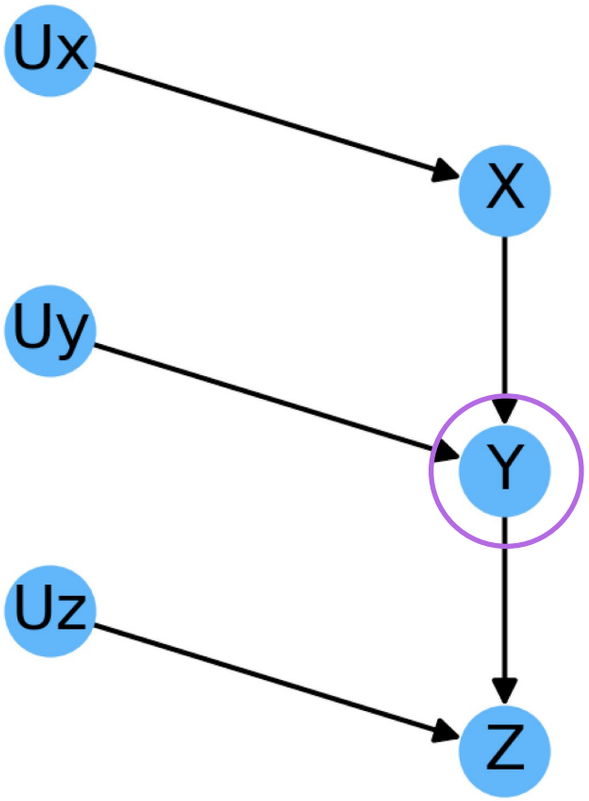
- From this Graph, we can immediately infer:
  - Z and Y are dependent — $P\left(Z|Y\right) \neq P\left(Z\right)$
  - Y and X are dependent — $P\left(Y|Z\right) \neq P\left(Y\right)$
  - Z and X are likely dependent — $P\left(Z|X\right) \neq P\left(Z\right)$



- 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— $P\left(Z|X,Y\right) = P\left(Z|Y\right)$

Chains

- From this Graph, we can immediately infer:
  - Z and Y are dependent — $P\left(Z|Y\right) \neq P\left(Z\right)$
  - Y and X are dependent — $P\left(Y|Z\right) \neq P\left(Y\right)$
  - Z and X are likely dependent — $P\left(Z|X\right) \neq P\left(Z\right)$

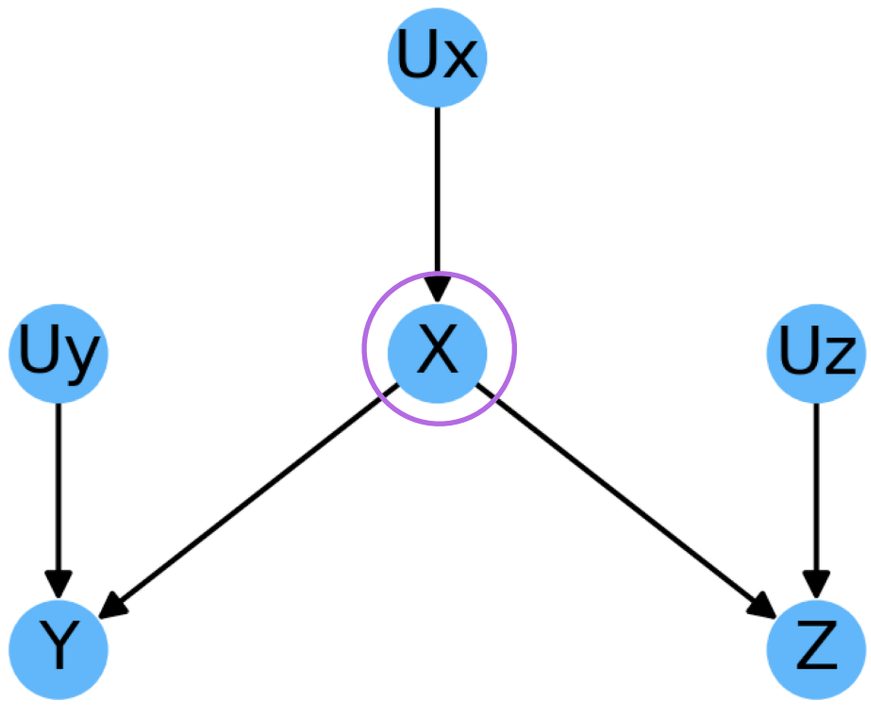


- We can also easily see how by setting the value of Y, we remove any influence that X can have on Z (as that can only happen through changes in Y. In other words:
  - Z and X are independent conditional on Y— $P\left(Z|X,Y\right) = P\left(Z|Y\right)$

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.

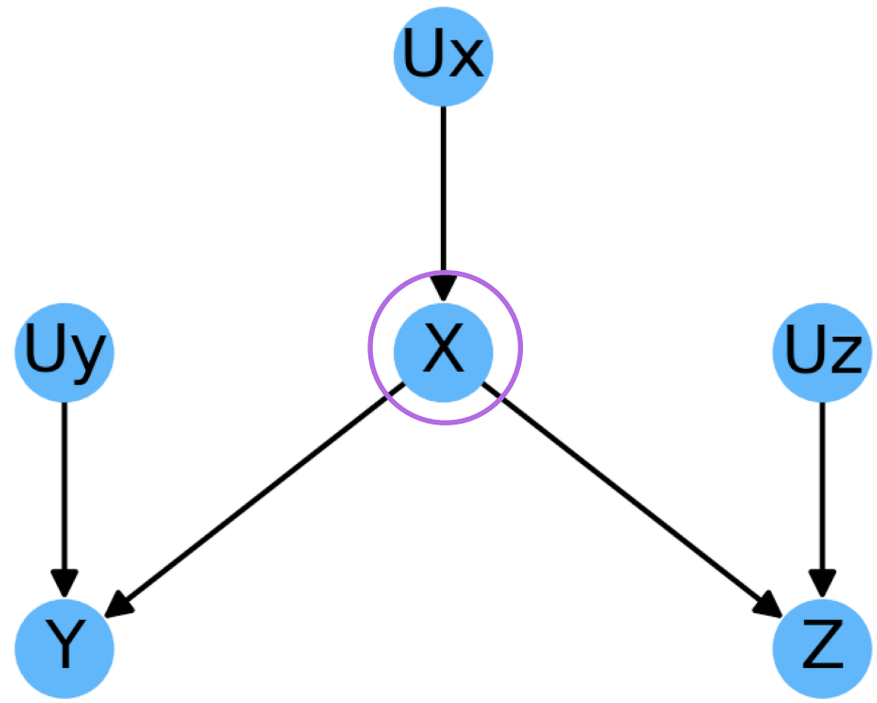
Forks

- From this Graph we can infer:
  - X and Y are dependent — $P\left(X|Y\right) \neq P\left(X\right)$
  - X and Z are dependent — $P\left(X|Z\right) \neq P\left(X\right)$



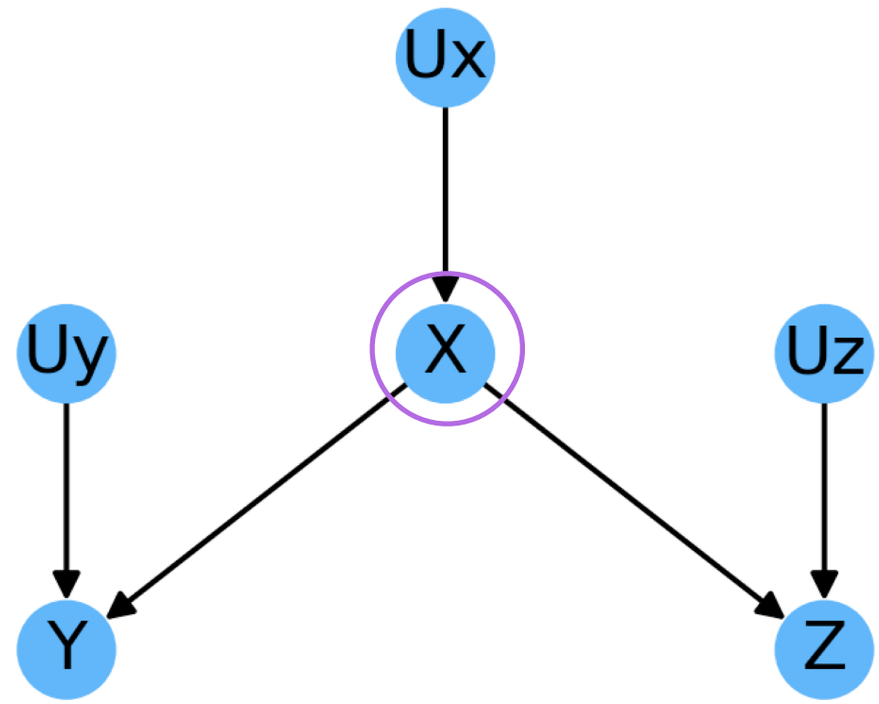
Forks

- From this Graph we can infer:
  - X and Y are dependent — $P\left(X|Y\right) \neq P\left(X\right)$
  - X and Z are dependent — $P\left(X|Z\right) \neq P\left(X\right)$
- Furthermore, we can see:
  - Z and Y are likely dependent — $P\left(Z|Y\right) \neq P\left(Z\right)$ as their values are both directly determined by the value of X
  - Y and Z are independent, conditional on X — $P\left(Y|Z,X\right) = P\left(Y|X\right)$ since from a given value of X, Y and Z are free to vary according to $P\left(Y|X\right)$ and $P\left(Z|X\right)$, respectively



Forks

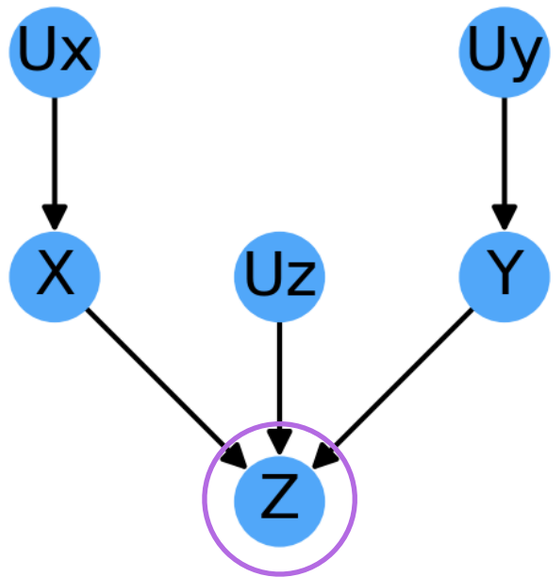
- From this Graph we can infer:
  - X and Y are dependent — $P\left(X|Y\right) \neq P\left(X\right)$
  - X and Z are dependent — $P\left(X|Z\right) \neq P\left(X\right)$
- Furthermore, we can see:
  - Z and Y are likely dependent — $P\left(Z|Y\right) \neq P\left(Z\right)$ as their values are both directly determined by the value of X
  - Y and Z are independent, conditional on X — $P\left(Y|Z,X\right) = P\left(Y|X\right)$ since from a given value of X, Y and Z are free to vary according to $P\left(Y|X\right)$ and $P\left(Z|X\right)$, 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

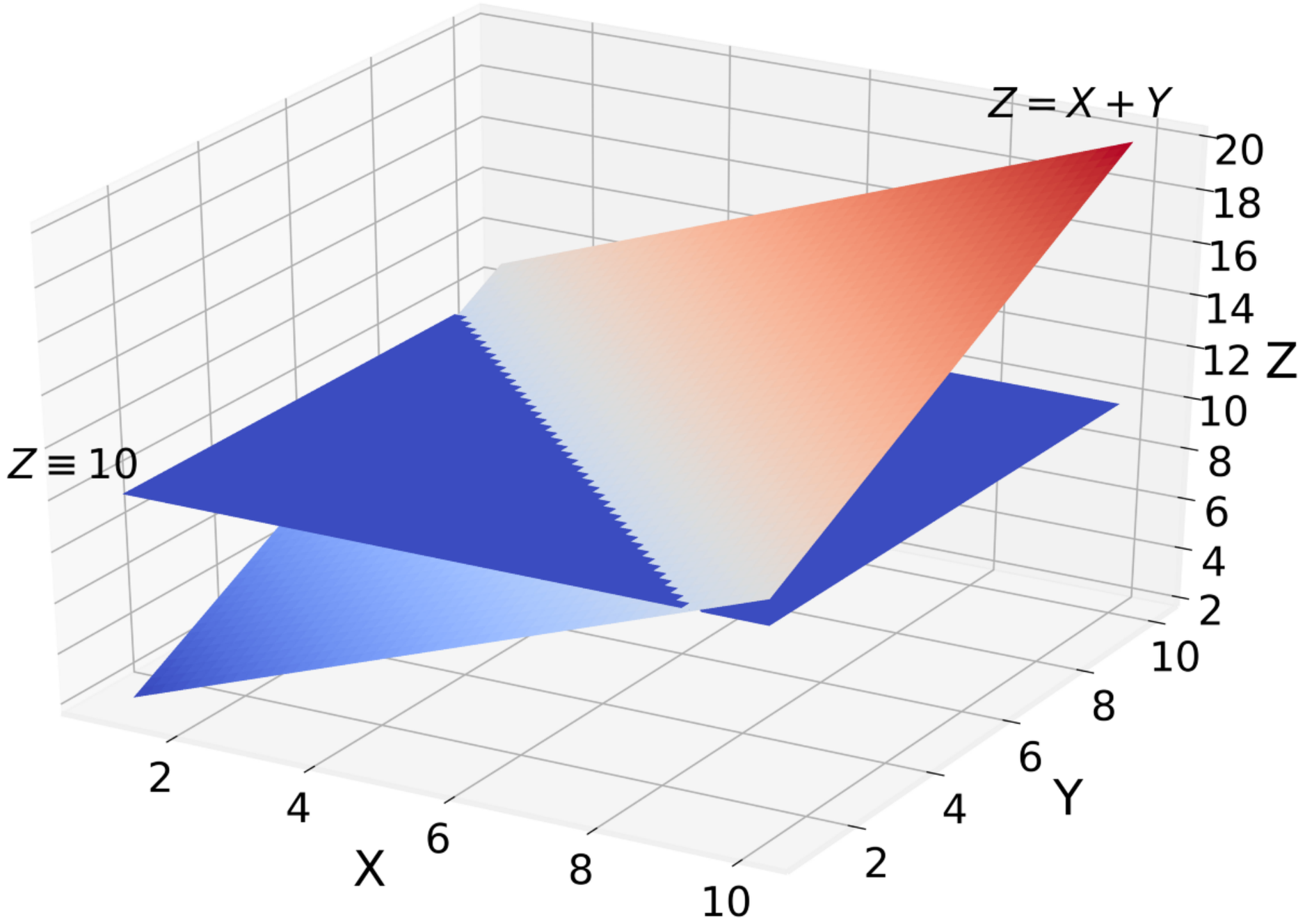
- From this Graph, we infer:
  - X and Z are dependent — $P\left(X|Z\right) \neq P\left(X\right)$
  - Y and Z are dependent — $P\left(Y|Z\right) \neq P\left(Y\right)$ as the value of Z is determined by both X and Y and so it provides information about their values (Information traveling "backwards")



  - X and Y are independent — $P\left(X|Y\right) = P\left(X\right)$ as they share no common ancestors
- Furthermore:
  - X and Y are dependent conditional on Z — $P\left(X|Y,Z\right) \neq P\left(X|Z\right)$ as only specific values of X and Y can be combined to produce the value of Z
- Let's consider a simple example: $Z = X + Y$

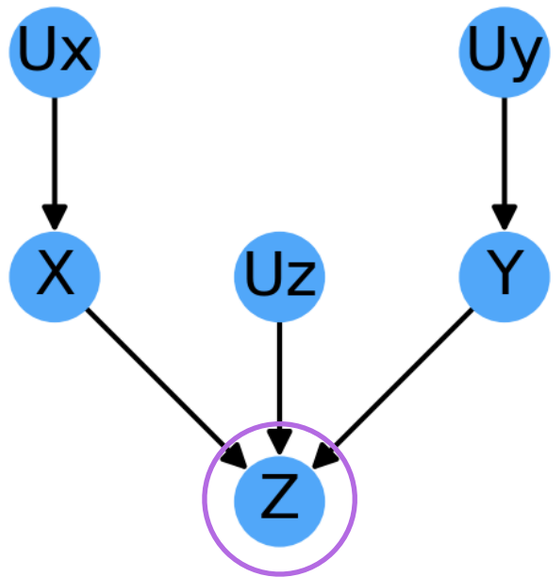
Colliders

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 — $P\left(X|Z\right) \neq P\left(X\right)$
  - Y and Z are dependent — $P\left(Y|Z\right) \neq P\left(Y\right)$ as the value of Z is determined by both X and Y and so it provides information about their values (Information traveling "backwards")

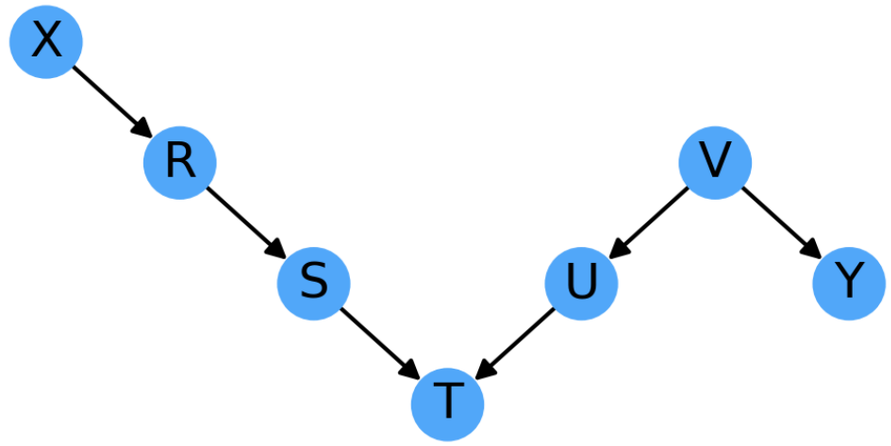


  - X and Y are independent — $P\left(X|Y\right) = P\left(X\right)$ as they share no common ancestors
- Furthermore:
  - X and Y are dependent conditional on Z — $P\left(X|Y,Z\right) \neq P\left(X|Z\right)$ as only specific values of X and Y can be combined to produce the value of Z

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.

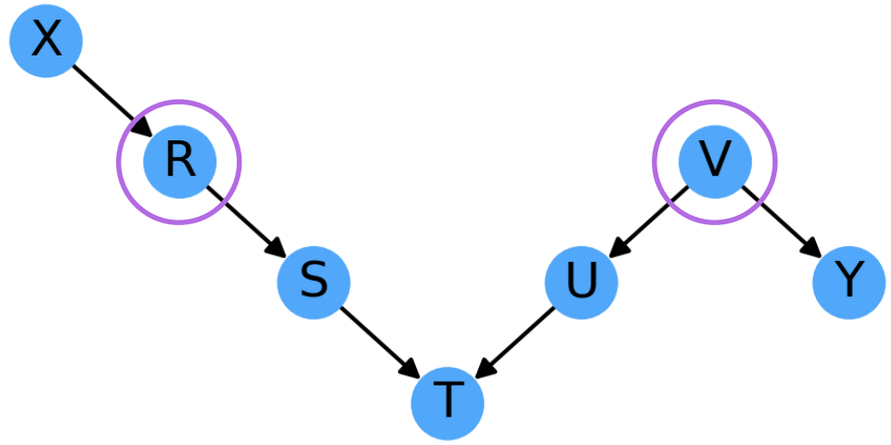
d-separation

- 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.

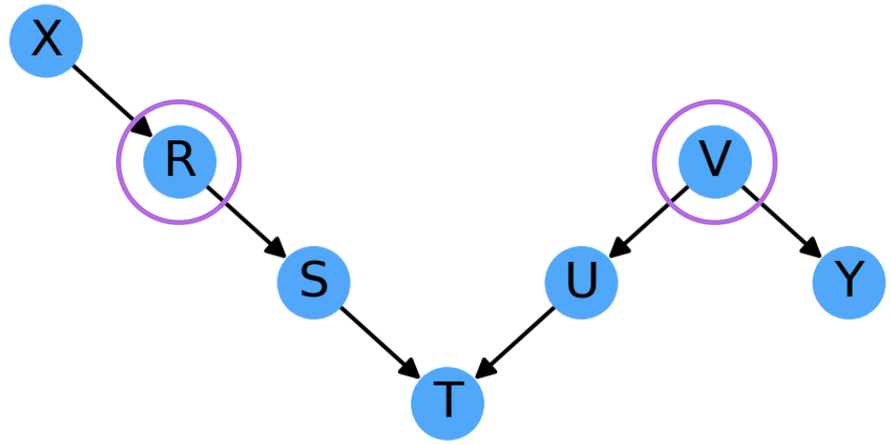
- If we condition on the values of R and V we can apply them to determine which variables should be independent:

  - 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

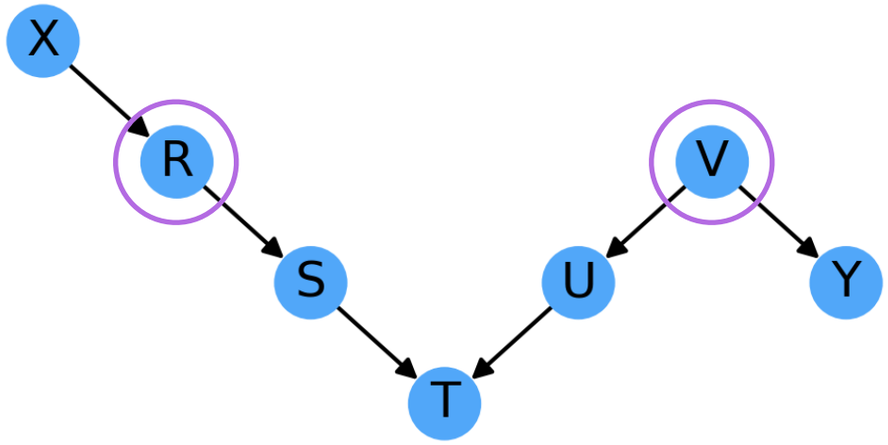
d-separation

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

- When we fit:

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

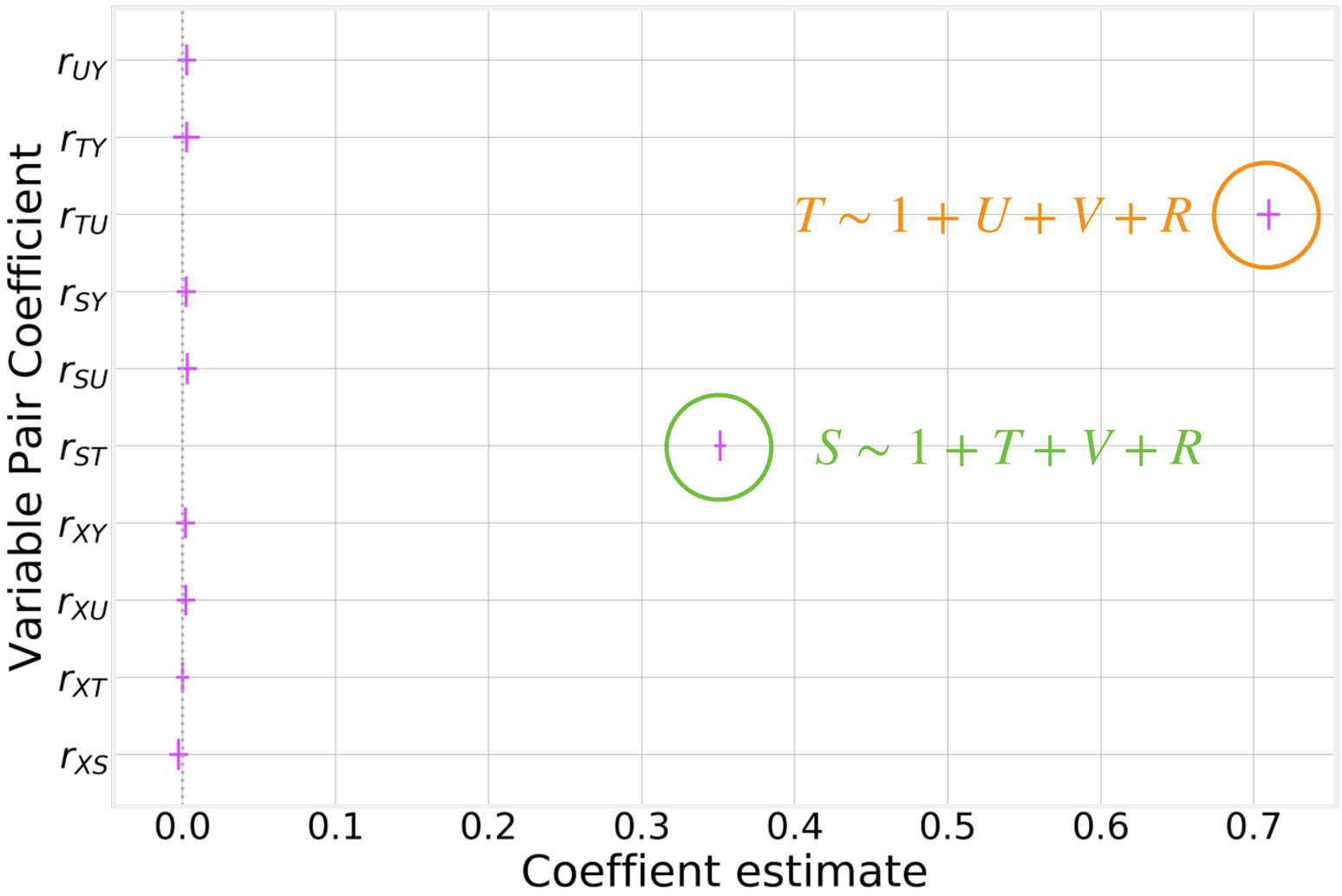
- If we find the r_{XY} coefficient to be zero, we can say that X is independent on Y conditional on Z, otherwise we say they are dependent.

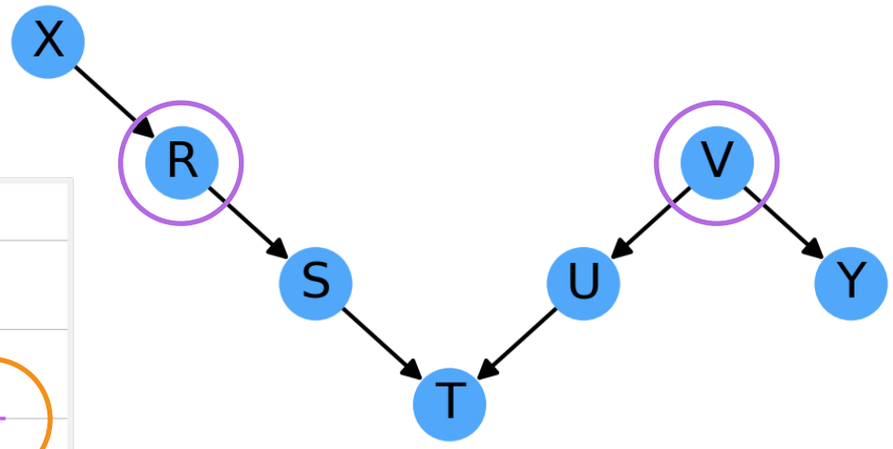


	source	target	Result
1	X	S	Rule 1, Independent
2	X	T	Rule 1, Independent
3	X	U	Blocked, Independent
5	X	Y	Blocked, Independent
11	S	T	Rule 0, Dependent
12	S	U	Rule 3, Independent
14	S	Y	Blocked, Independent
15	T	U	Rule 0, Dependent
17	T	Y	Rule 2, Independent
19	U	Y	Rule 2, Independent

d-separation

- Performing all the fits, we obtain:

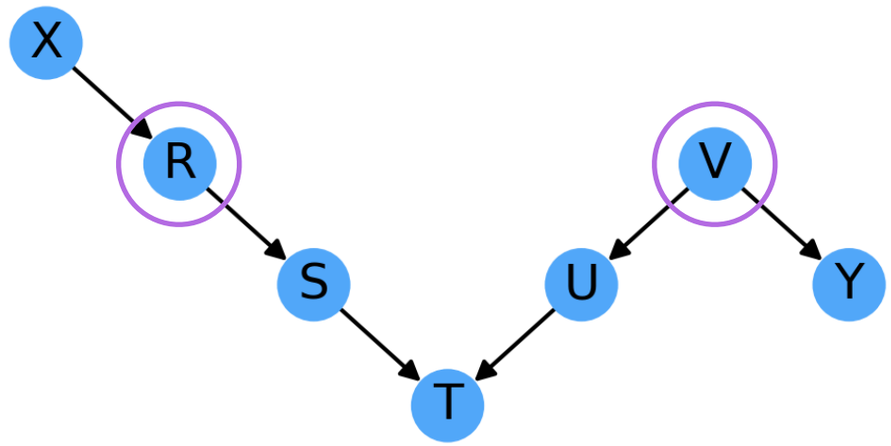




	source	target	Result
1	X	S	Rule 1, Independent
2	X	T	Rule 1, Independent
3	X	U	Blocked, Independent
5	X	Y	Blocked, Independent
11	S	T	Rule 0, Dependent
12	S	U	Rule 3, Independent
14	S	Y	Blocked, Independent
15	T	U	Rule 0, Dependent
17	T	Y	Rule 2, Independent
19	U	Y	Rule 2, Independent

d-separation

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

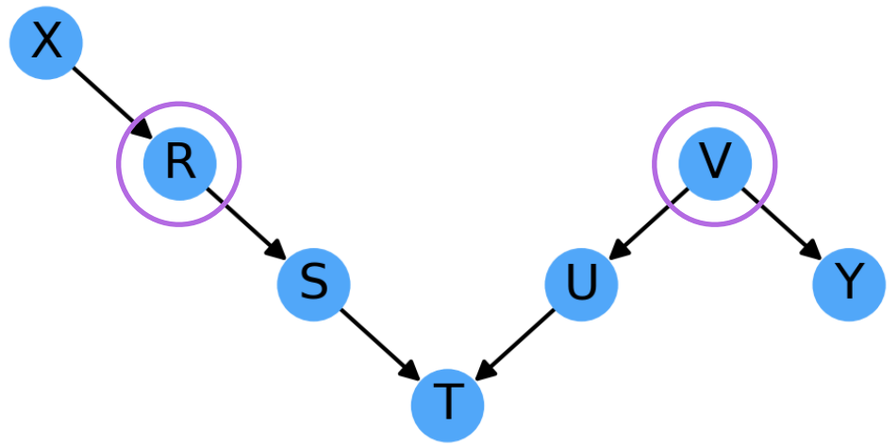


Definition (d-separation) — A path p is blocked by a set of nodes Z, if and only if:
1. p contains a chain of nodes A → B → C, or a fork A ← B → C such that the middle node B is in Z, or
2. p is a collider A → B ← C such that the collision node B is is not in Z, and no descendent of B is in Z.

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:



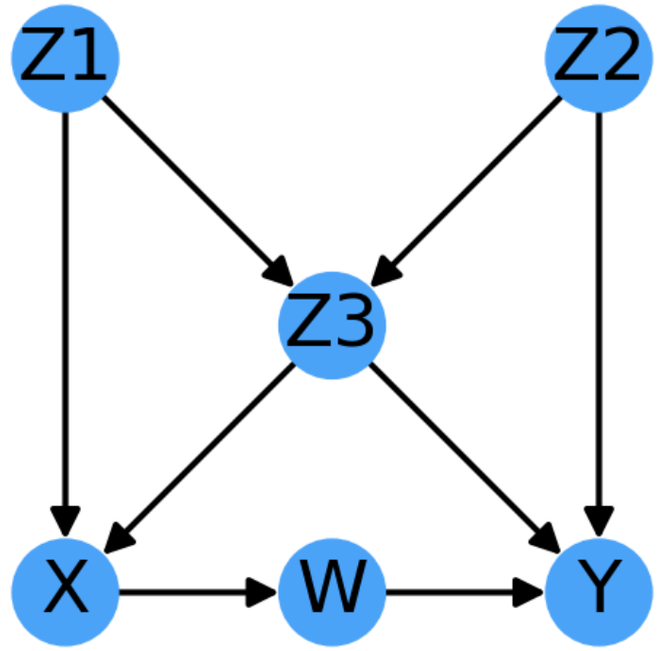
Definition (d-separation) — A path p is blocked by a set of nodes Z, if and only if:
1. p contains a chain of nodes A → B → C, or a fork A ← B → C such that the middle node B is in Z, or
2. p is a collider A → B ← C such that the collision node B is is not in Z, and no descendent of B is in Z.

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.

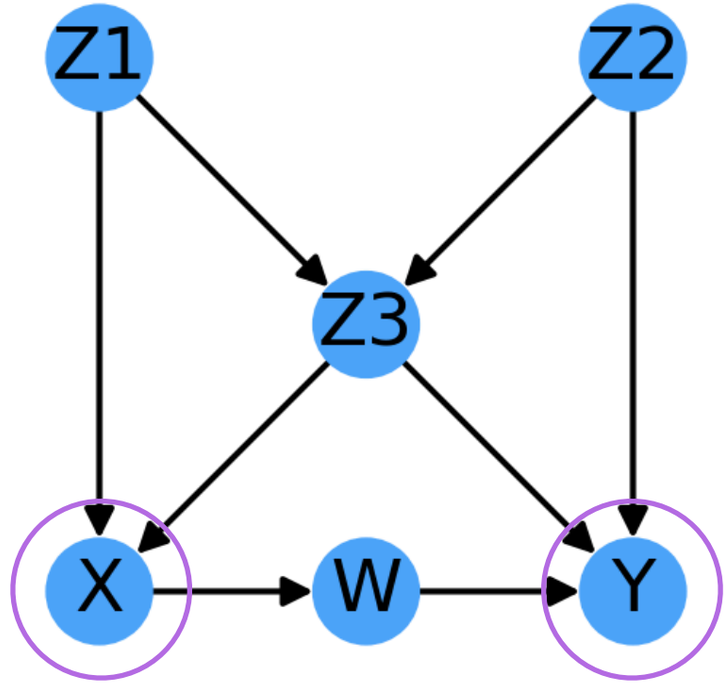
d-separation

- Let's consider another example.
- We have 3 Forks (Z1, Z2, and Z3) and 3 Colliders (Z3, X, and Y)



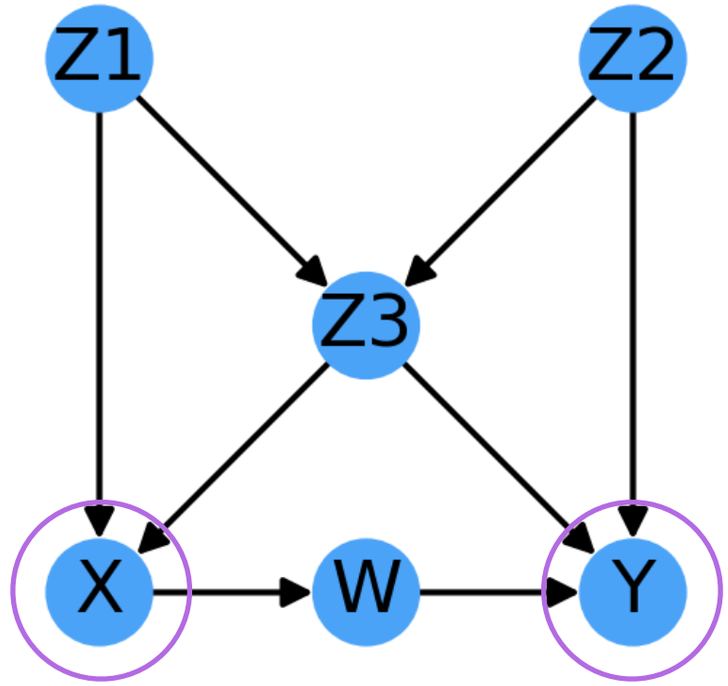
d-separation

- Let's consider another example.
- We have 3 Forks (Z1, Z2, and Z3) and 3 Colliders (Z3, 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)?



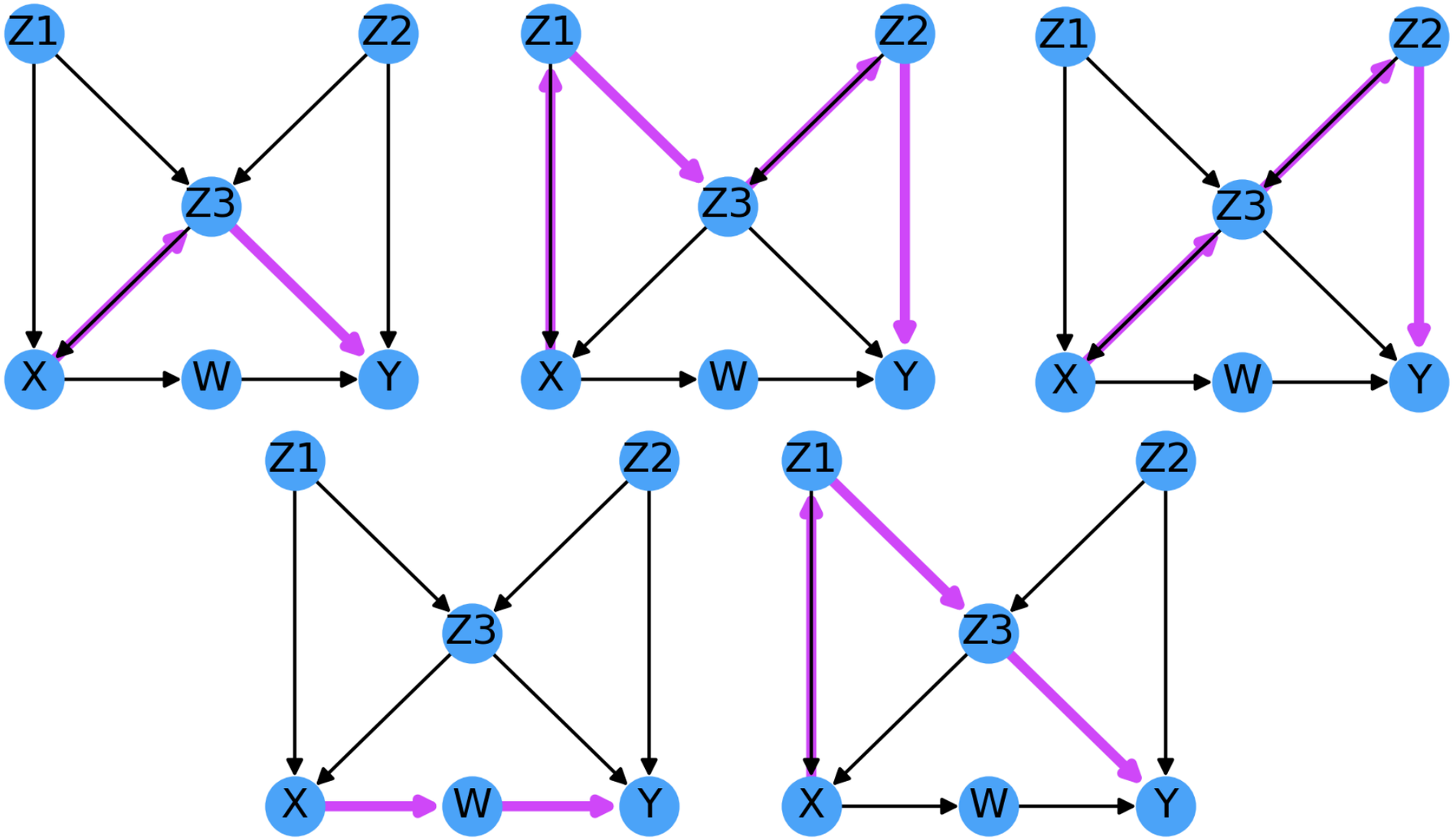
d-separation

- Let's consider another example.
- We have 3 Forks (Z1, Z2, and Z3) and 3 Colliders (Z3, 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



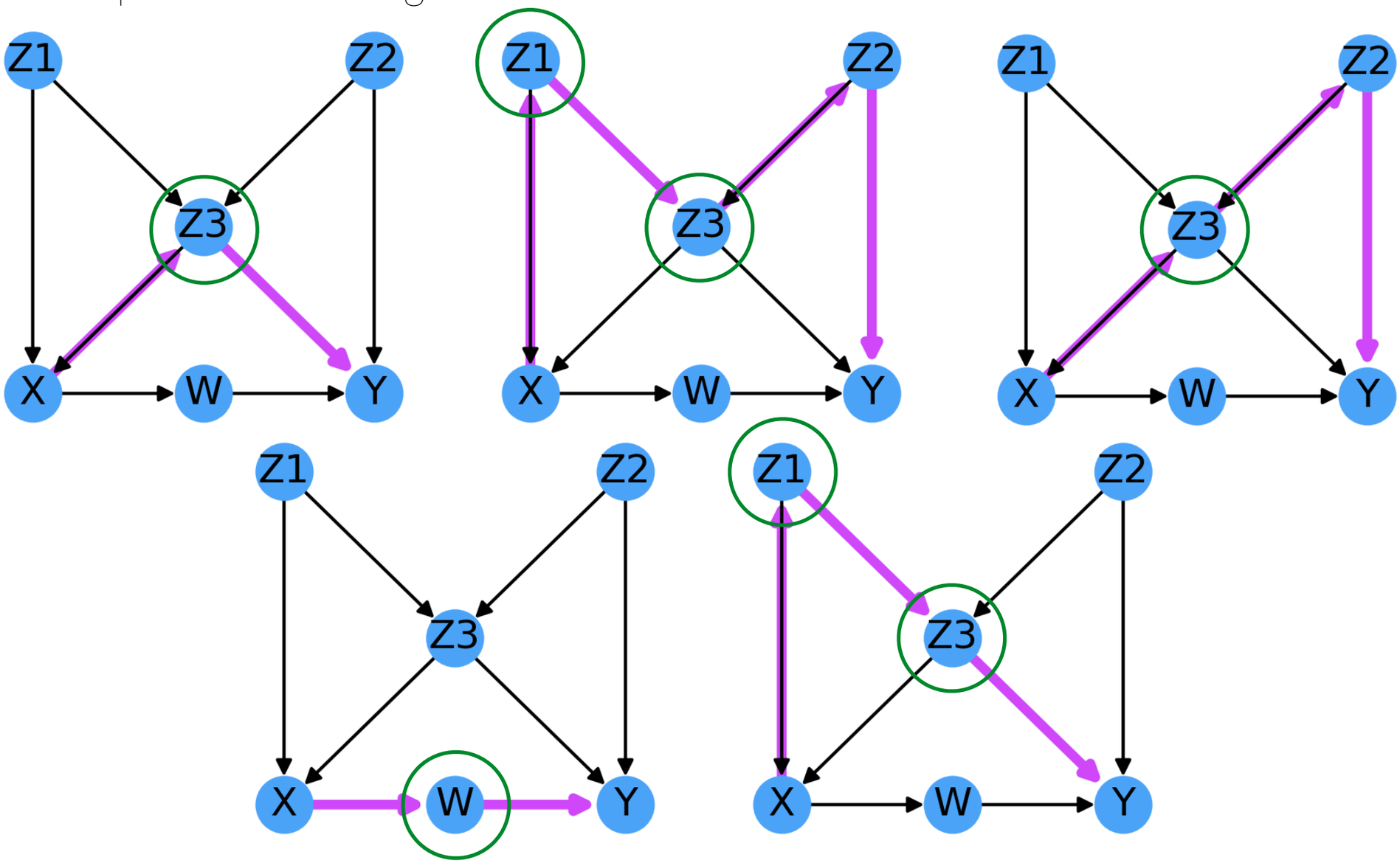
d-separation

- There are 5 paths connecting X and Y



d-separation

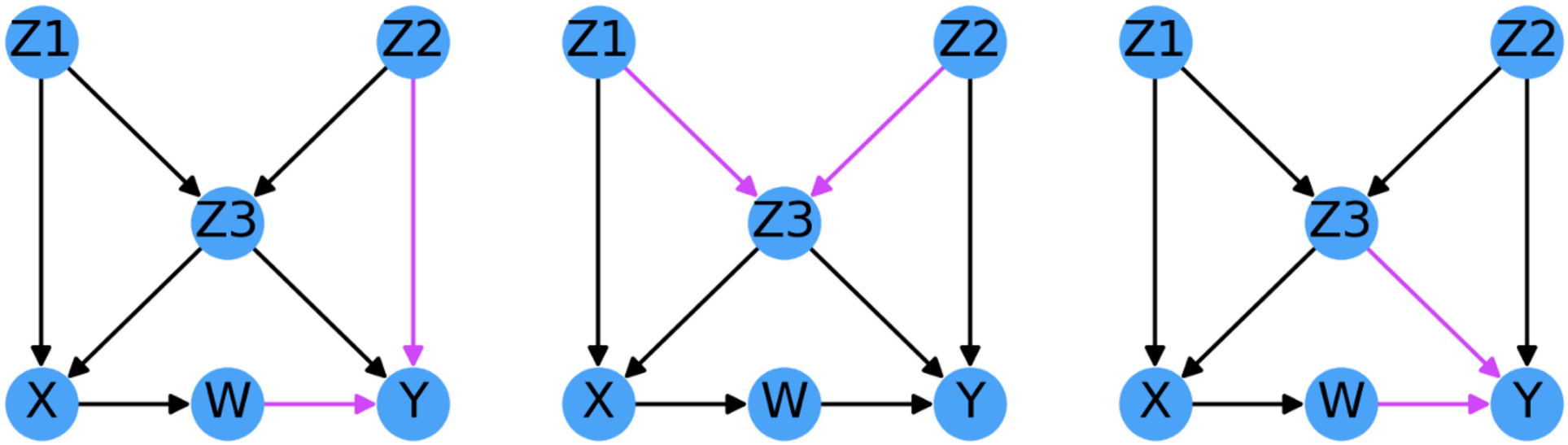
- There are 5 paths connecting X and Y



- We can block them all by conditioning on W, Z1 and Z3

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.

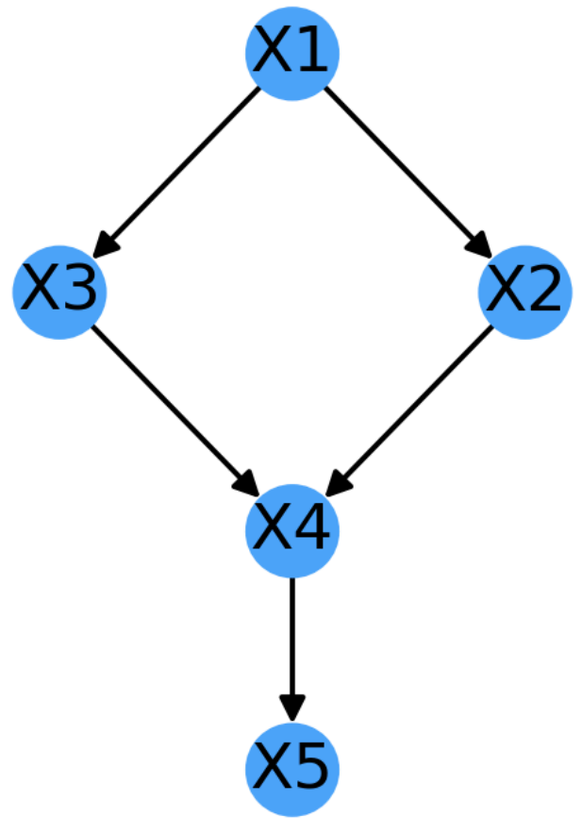
- A model is validated by the data if every d-separation condition is empirically verified by the data.

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

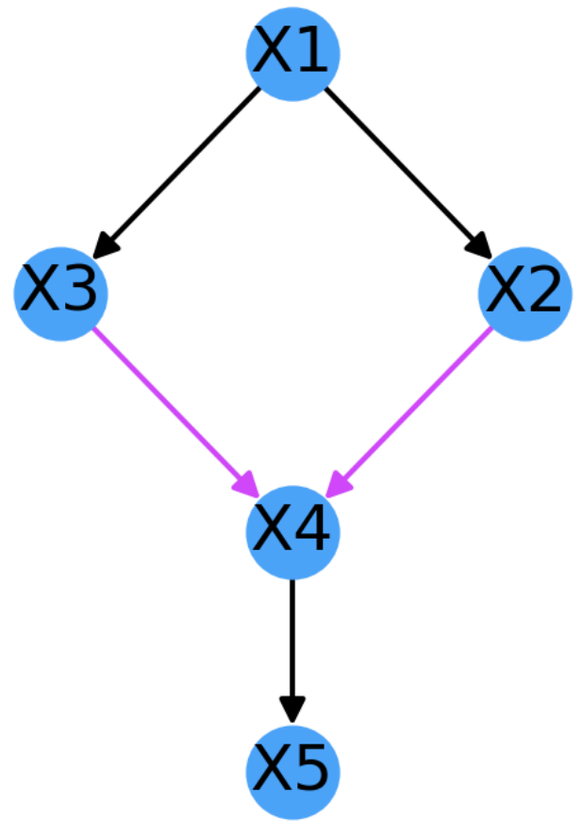
- Let us consider this DAG:



Model Testing and Causal Search

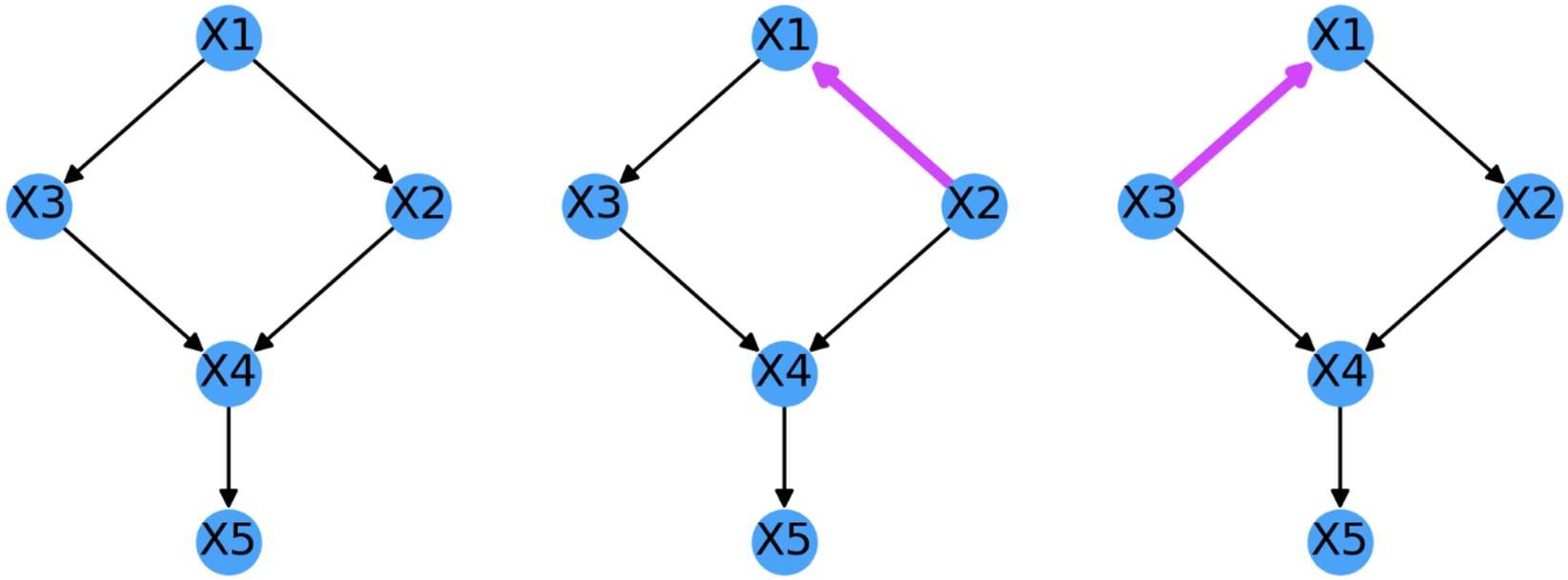
- Let us consider this DAG:

- 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

3. Interventions

Interventions

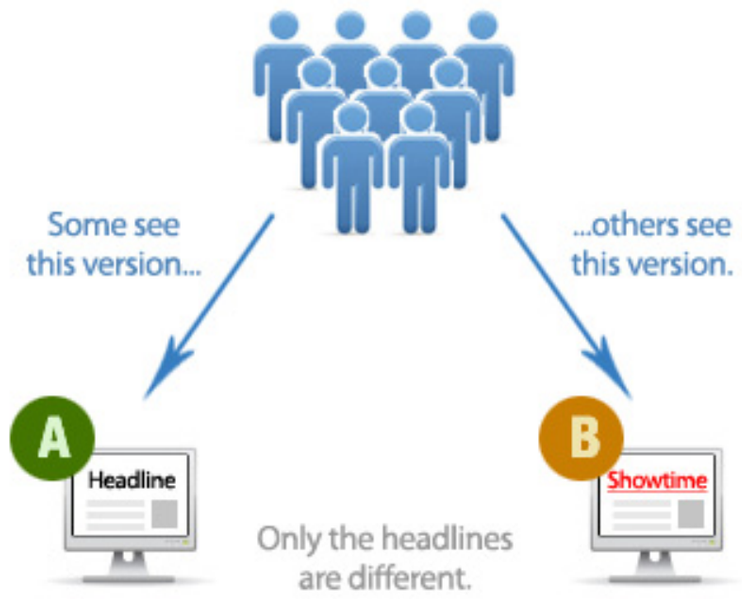
- In many cases we are interested in predicting the effects of interventions:
    - What happens if we take medication X?
    - What happens if we change the color of bottom Y?
    - What happens if we increase the price by Z?

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

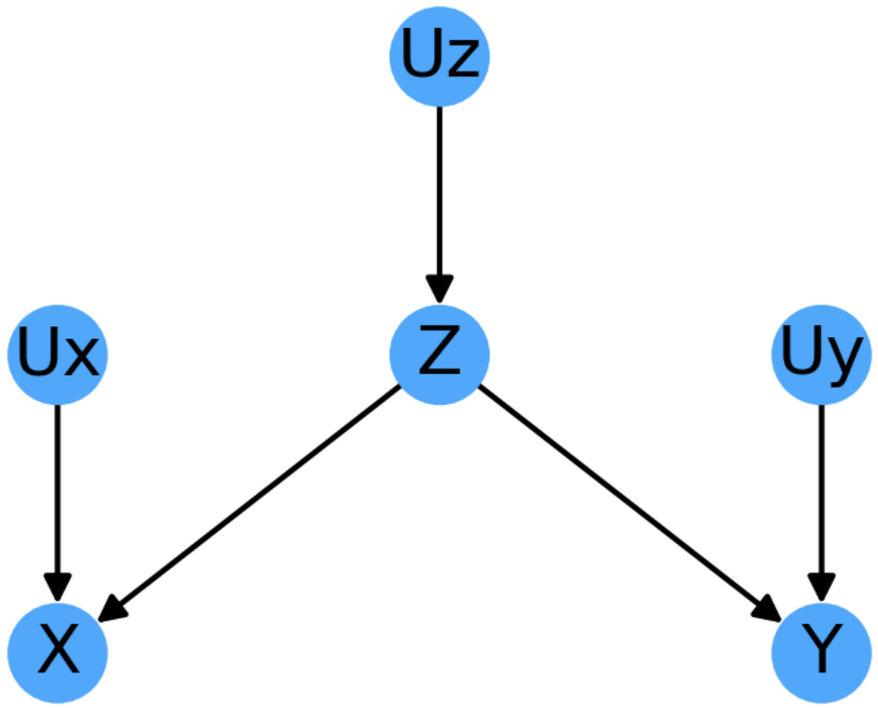
- Unfortunately, it's not always possible to perform the randomized experiment



Interventions

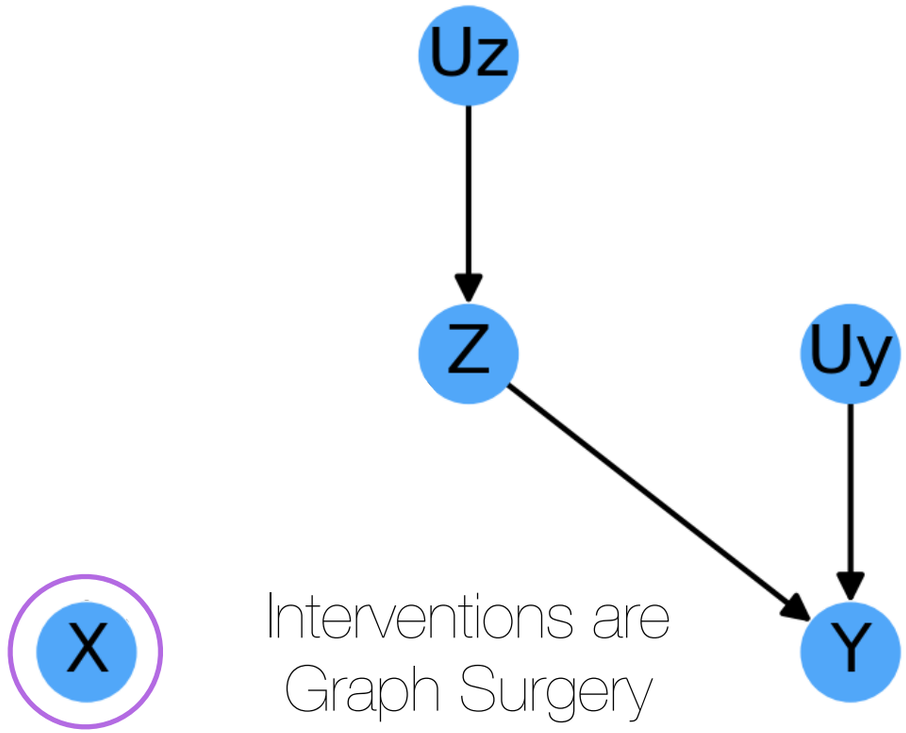
- We can measure the effects of an intervention using purely observational data.

- 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.

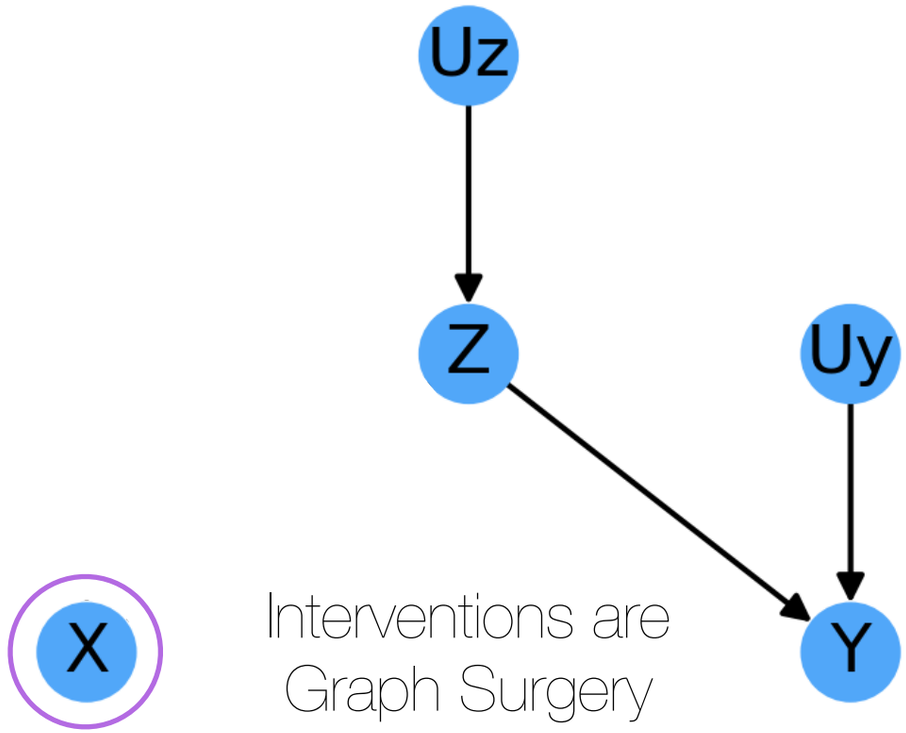
- Intervening requires that we fix the value of a given variable. This is equivalent to removing all incoming edges

- In this DAG, if we were to intervene on X, we would obtain

- Mathematically, we represent the intervention using the $\mathbf{do}\,()$ operator:

$$P_m\left(Y \,|\, \mathrm{do}\,(X = x)\right)$$

- 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 $P\left(Y|X\right)$



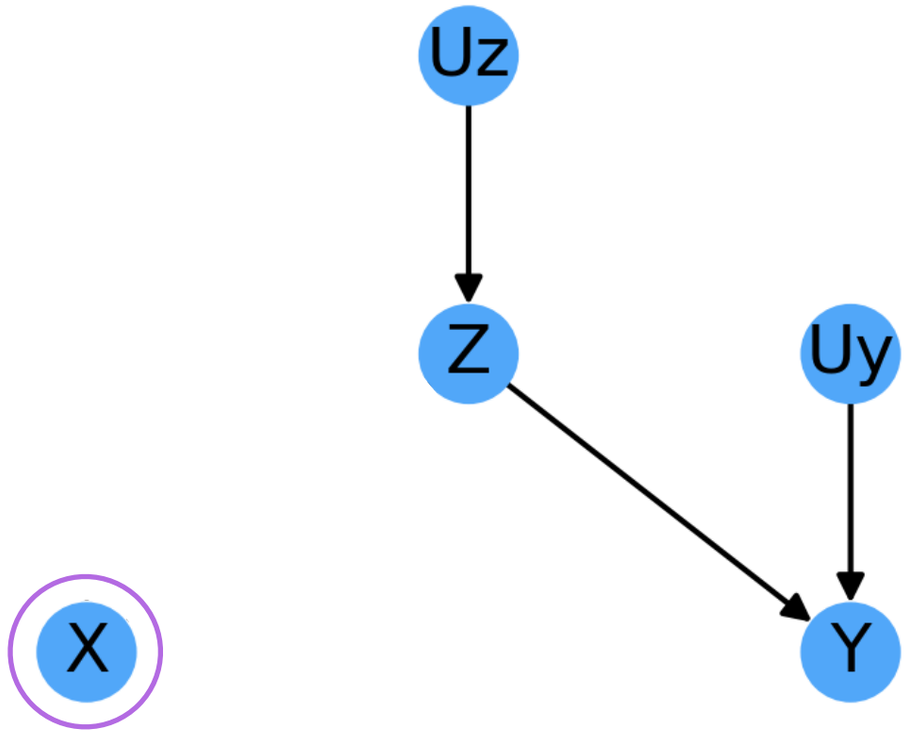
Average Causal Effect

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

$$P_m\left(Y \,|\, \mathrm{do}\,(X=1)\right) - P_m\left(Y \,|\, \mathrm{do}\,(X=0)\right)$$

- 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

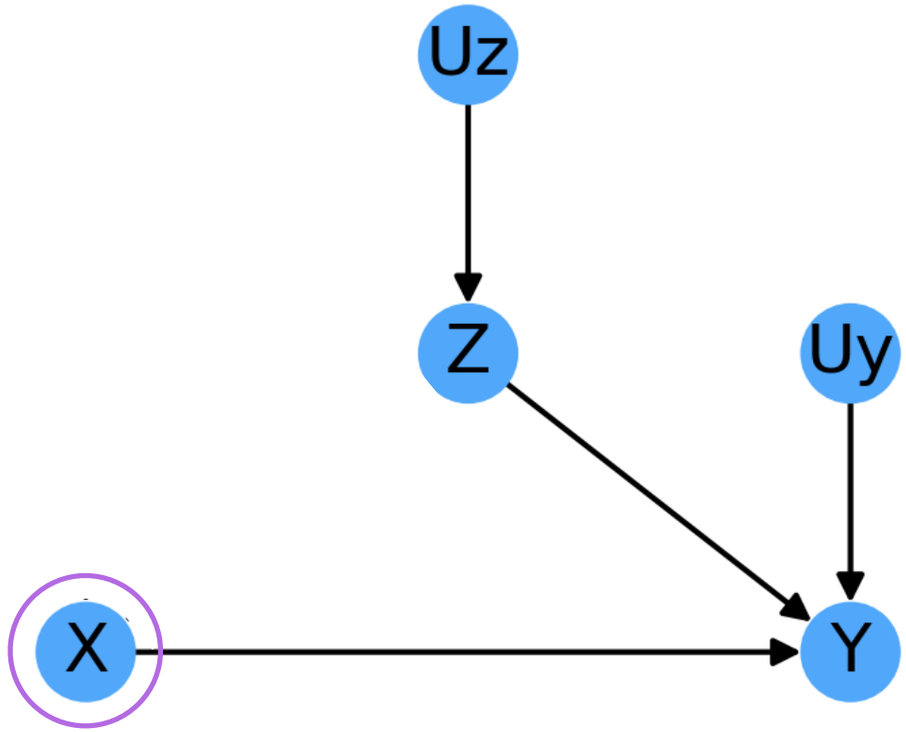
- We quantify the effect of the intervention by calculating the difference between intervening or not:

$$P_m\left(Y \,|\, \mathrm{do}\,(X=1)\right) - P_m\left(Y \,|\, \mathrm{do}\,(X=0)\right)$$

- 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



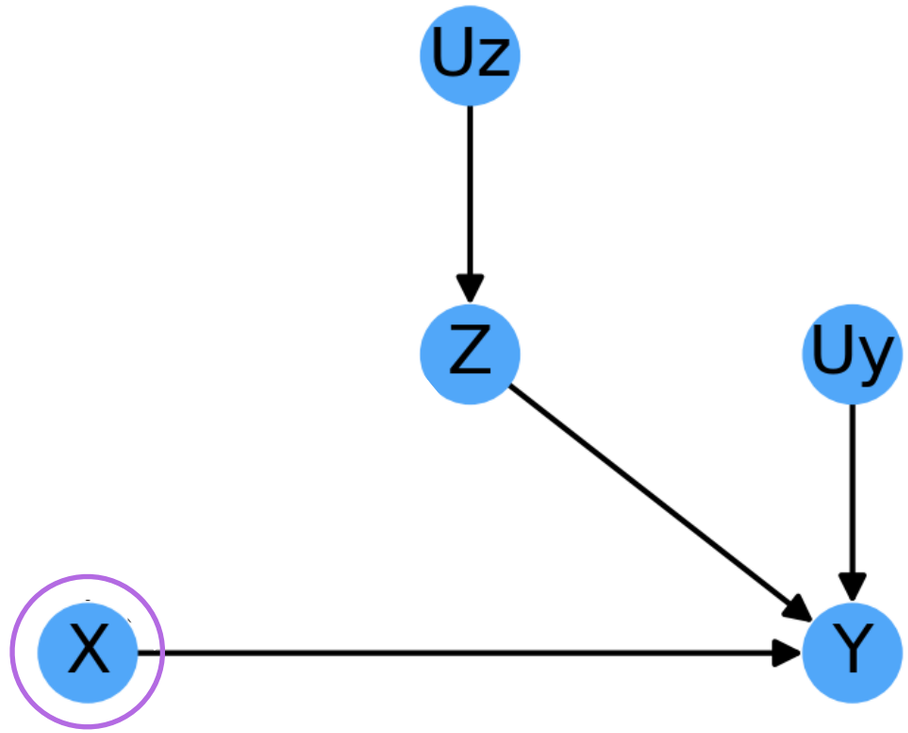
Average Causal Effect

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

$$P_m\left(Y|\mathrm{do}\,(X=1)\right) - P_m\left(Y|\mathrm{do}\,(X=0)\right)$$

- 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:

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

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

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

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

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:

$$P\left(Y|\,\mathrm{do}\,(X)\right) = \sum_{PA} P\left(Y|X,PA\right) P\left(PA\right)$$

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 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:

$$P\left(Y|\,\mathrm{do}\,(X)\right) = \sum_{Z} P\left(Y|X,Z\right) P\left(Z\right)$$

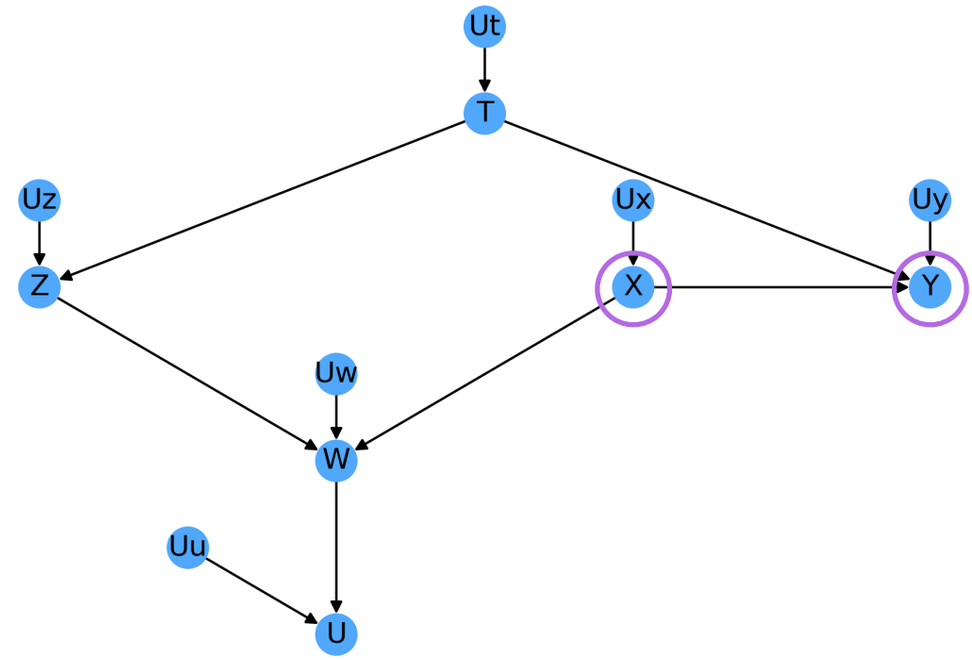
and is always satisfied by PA.

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

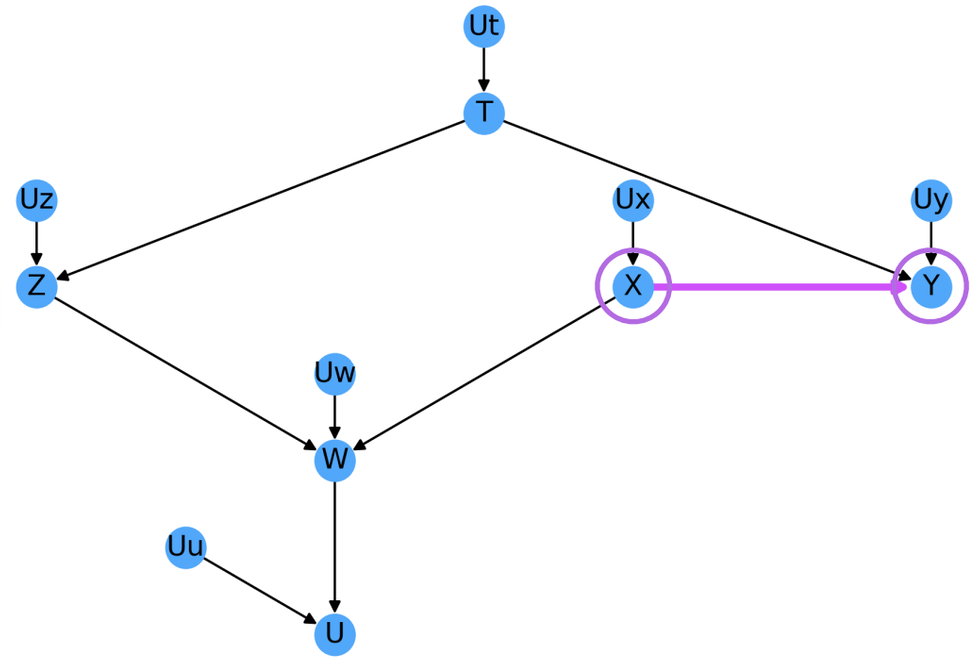


- Let's consider an example. We want to quantify:

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

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:

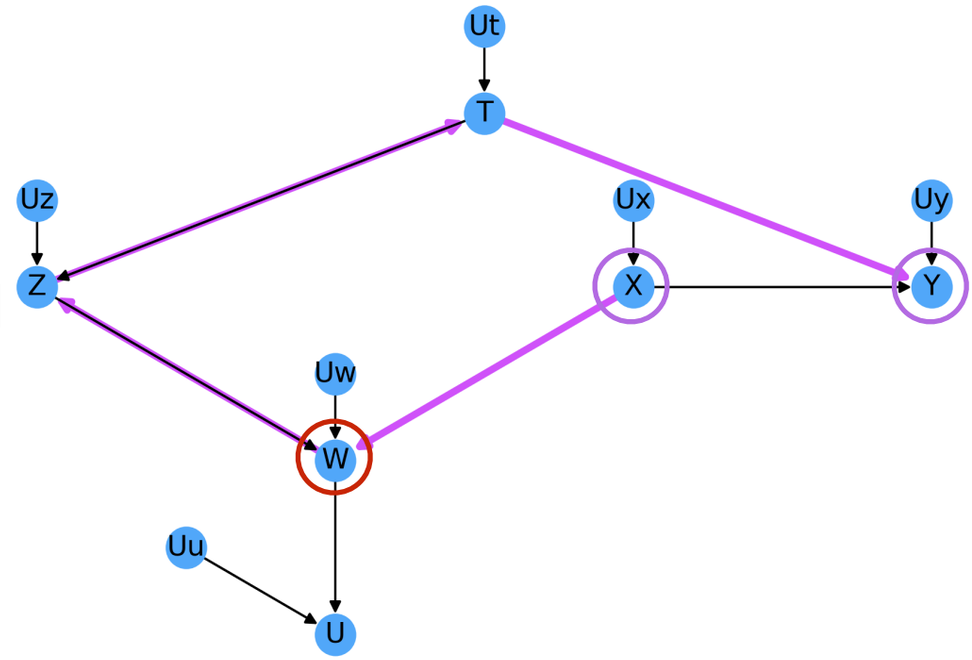
$$P\left(Y|\,\mathrm{do}\,(X)\right)$$

- The backdoor path is blocked by the collider at W (d-separation), so

$$P\left(Y|\,\mathrm{do}\,(X)\right) \equiv P\left(Y|X\right)$$

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:

$$P\left(Y|\,\mathrm{do}\,(X)\right)$$

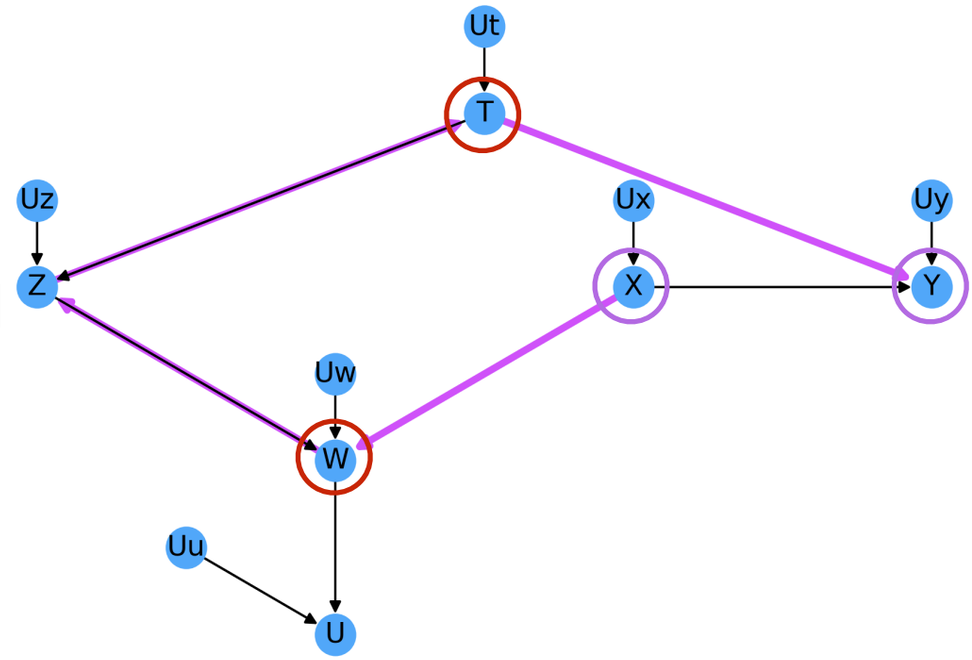
- The backdoor path is blocked by the collider at W (d-separation), so

$$P\left(Y|\,\mathrm{do}\,(X)\right) \equiv P\left(Y|X\right)$$

- 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:

$$P\left(Y|\,\mathrm{do}\,(X)\right)$$

- The backdoor path is blocked by the collider at W (d-separation), so

$$P\left(Y|\,\mathrm{do}\,(X)\right) \equiv P\left(Y|X\right)$$

- 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 $P\left(Y|\,\mathrm{do}\,(X),W\right)$ will be biased. In this case, we would compute:

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

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

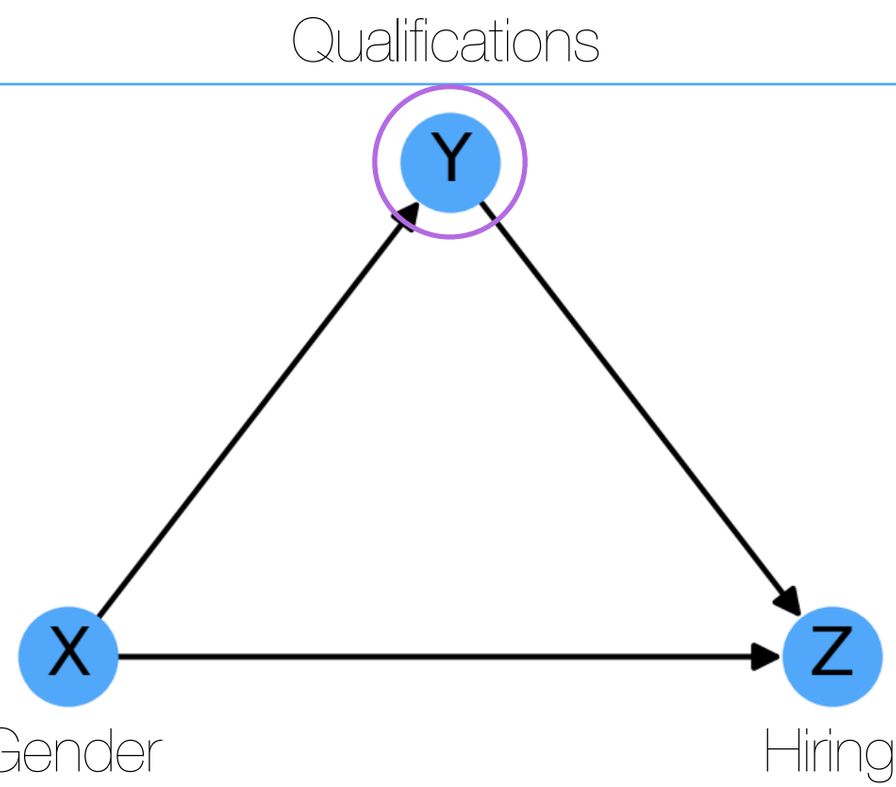
Front-Door Adjustment - If Z satisfies the front door criterion relative to (X, Y) and $P(X, Z) > 0$ then the causal effect of X on Y is:

$$P\left(Y|\,\mathrm{do}\,(X)\right) = \sum_{Z} P\left(Z|X\right) \sum_{X'} P\left(Y|X', Z\right) P\left(X'\right)$$

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

$$P\left(Z|X,Y\right)$$

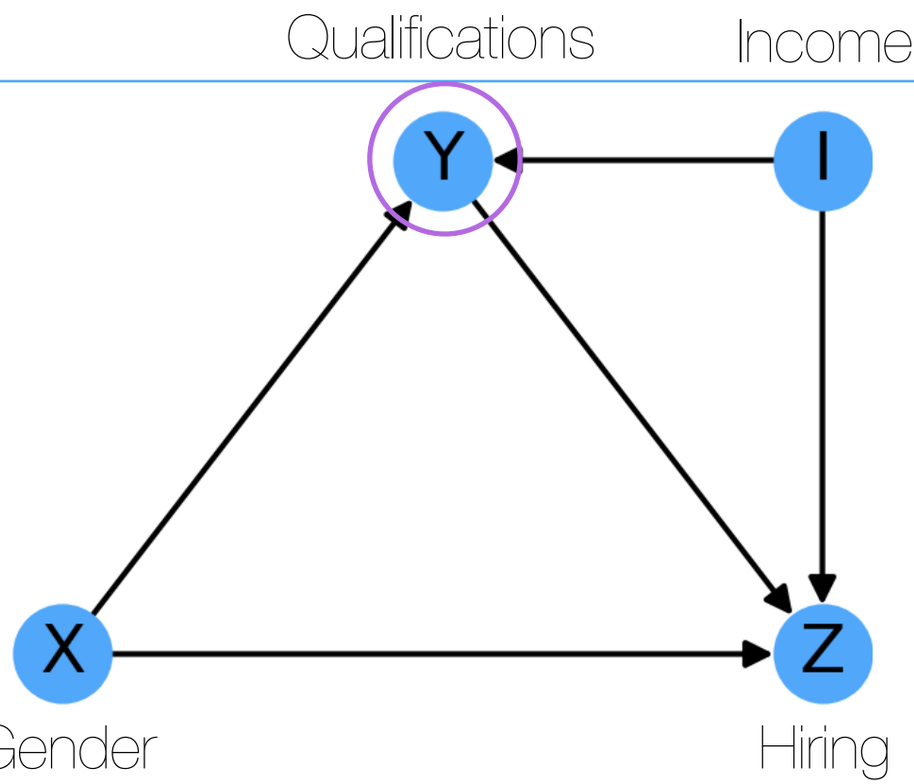


Mediation

- 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

$$P\left(Z|X,Y\right)$$

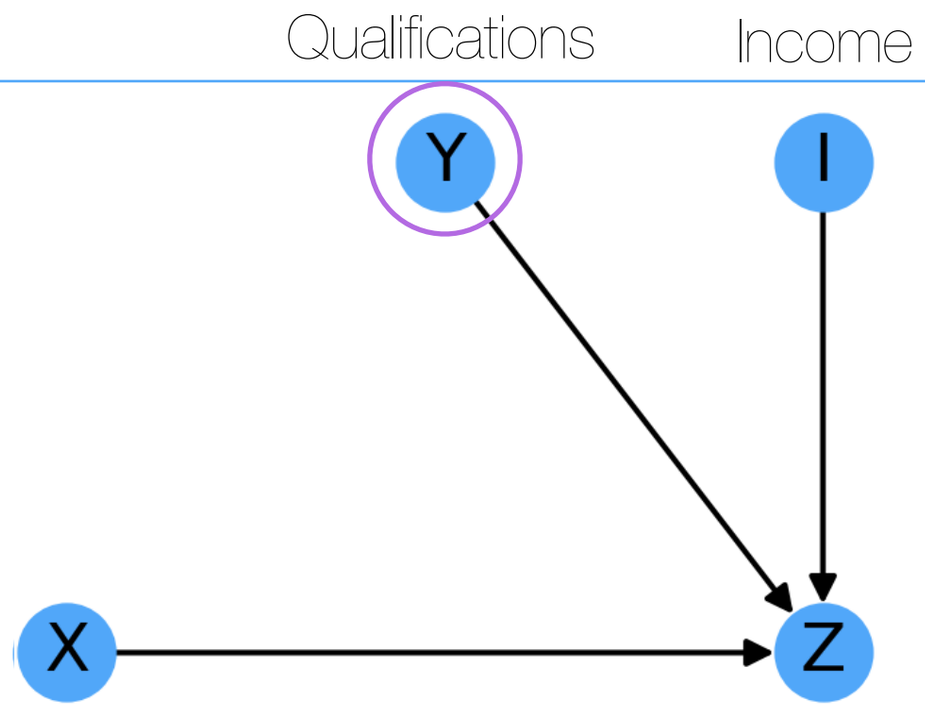
- But what if there is a cofounder of Y?
- Now if we condition on Y we are conditioning on a Collider and unblocking the path through I.



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

$$P\left(Z|X,Y\right)$$



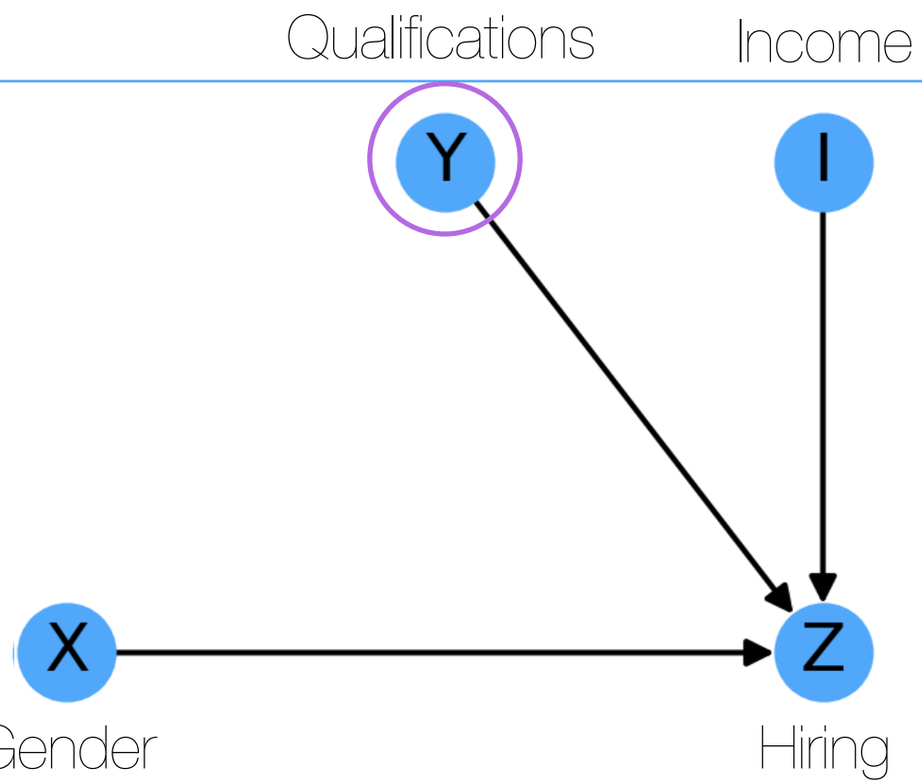
- But what if there is a cofounder of Y?
- Now if we condition on Y we are conditioning on a Collider and unblocking the path through I.
- One way around this difficulty is to intervene on Y, which removes the edges from X and I
- Now we can measure the Controlled Direct Effect (CDE):

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

- 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 backdown 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

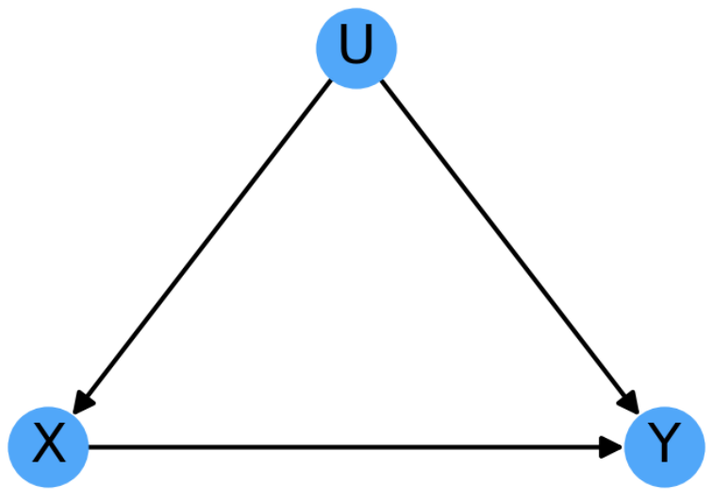
https://github.com/DataForScience/CausalInference

4. 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 out 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:



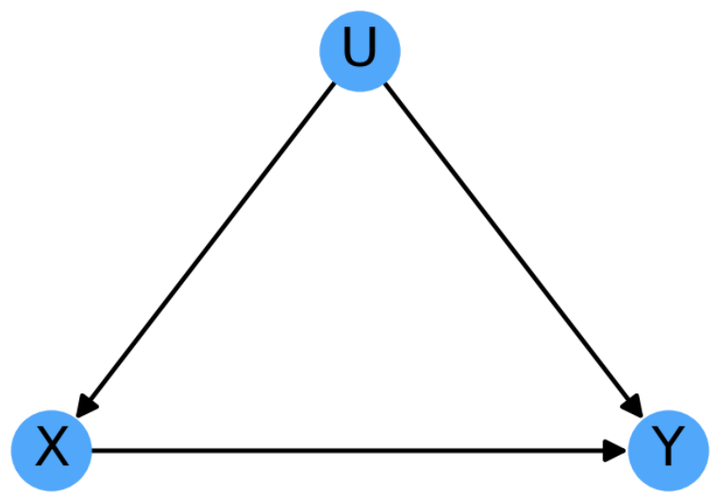
$$X = aU \qquad Y = bX + U$$

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 solving, we obtain

$$X_y(u) = au$$

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

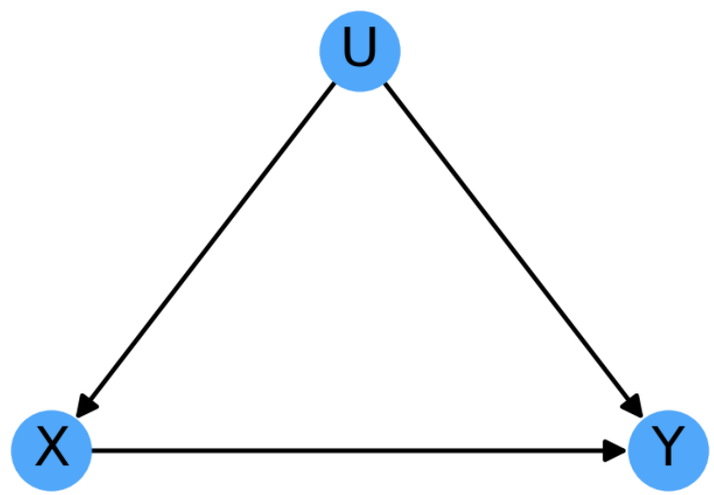
$$X = aU \qquad Y = bX + U$$

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 solving, we obtain

$$X_y(u) = au$$

$X = aU$ $\quad$ $Y = bX + U$

- As might have bene 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 conterfactual is defined as

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

Steps to Compute Counterfactuals

- The procedure requires 3 steps:

1. Abduction - Use evidence $\boldsymbol{E = e}$ to determine the value of $\boldsymbol{U}$

2. Action - Modify the model $\boldsymbol{M}$, by setting $\boldsymbol{X = x}$ to obtain $\boldsymbol{M_x}$

3. Prediction - Use $\boldsymbol{M_x}$ and $\boldsymbol{U}$ to compute $\boldsymbol{Y_x(u)}$

Steps to Compute Counterfactuals

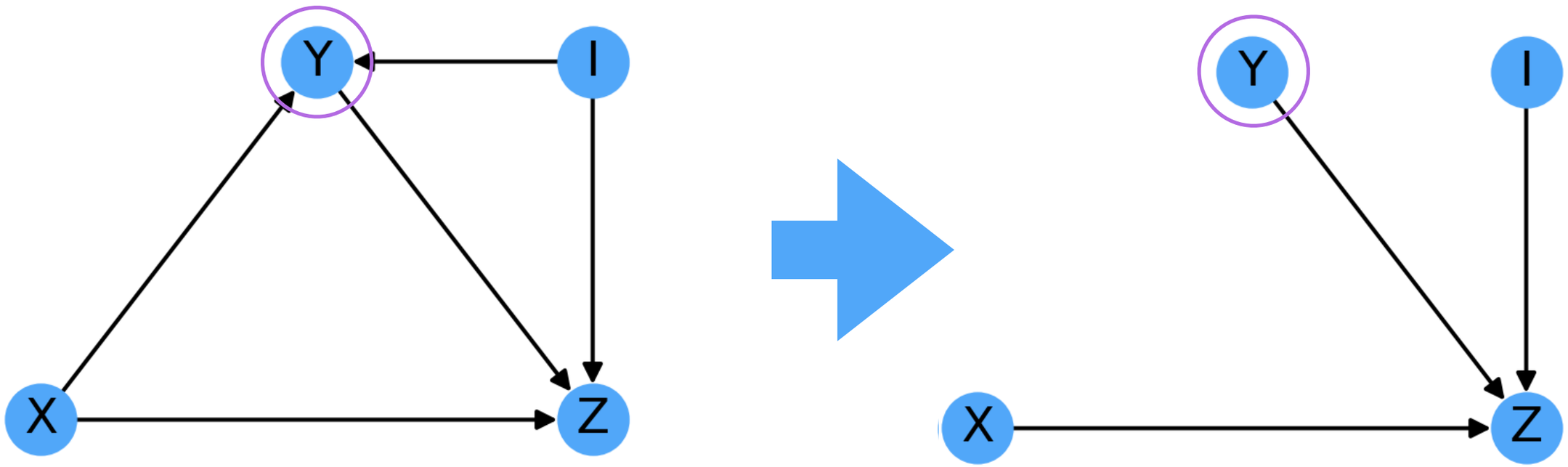
- The procedure requires 3 steps:

  1. Abduction - Use evidence $\boldsymbol{E = e}$ to determine the value of $\boldsymbol{U}$

  2. Action - Modify the model $\boldsymbol{M}$, by setting $\boldsymbol{X = x}$ to obtain $\boldsymbol{M_x}$

  3. Prediction - Use $\boldsymbol{M_x}$ and $\boldsymbol{U}$ to compute $\boldsymbol{Y_x(u)}$

- Step 2 is similar to an intervention



Machine Learning

- Supervised Learning
  - Predict output given input
  - Training set of known inputs and outputs is provided
- Unsupervised Learning
  - Autonomously learn an good representation of the dataset
  - Find clusters in input
- Reinforcement Learning
  - Learn sequence of actions to maximize payoff
  - Discount factor for delayed rewards

Machine Learning

- Machine Learning seeks to infer properties of the distribution underlying the data set
- Causal Inference assumes a strong underlying structure and explores their consequences
- In Machine Learning we often are limited by the available data, while still trying to extrapolate to unseen instances
- In Causal Inference we learn rules to explore the effects of interventions, counterfactuals, etc

Events

www.data4sci.com/newsletter

http://paypal.me/data4sci

Data Visualization with matplotlib and seaborn for Everyone

Nov 12, 2020 - 5am-9am (PST)

Deep Learning for Everyone

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

Time Series for Everyone

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

Advanced Time Series for Everyone

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

Natural Language Processing (NLP) from Scratch

http://bit.ly/LiveLessonNLP - On Demand

END