Applied Micro-Econometrics,Fall 2023

Lecture 3: Controls in DAGs

Zhaopeng Qu
Nanjing University Business School
9/27/2023

DAGs

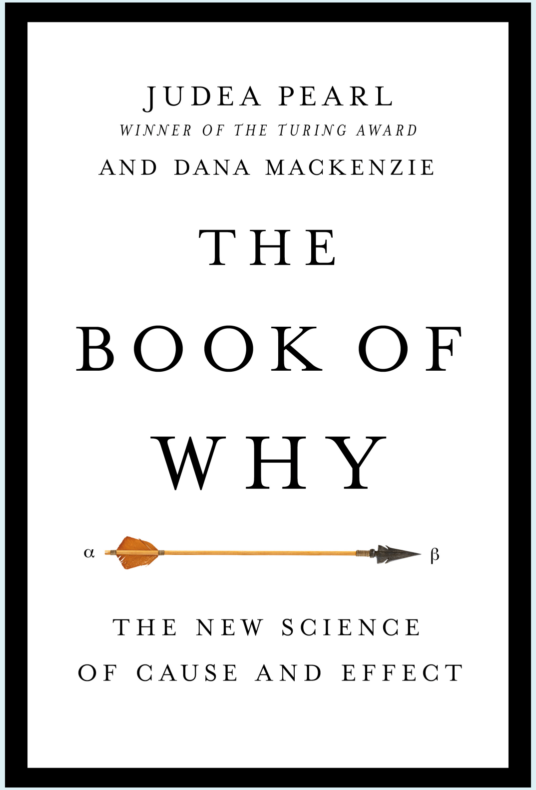
What's a DAG?

- DAG stands for **directed acyclic graph**, which graphically illustrates the causal relationships and non-causal associations within a network of random variables.

Judea Pearl(UCLA)

- Computer Scientist, **Turing Award** in 2011.

> "for fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning"

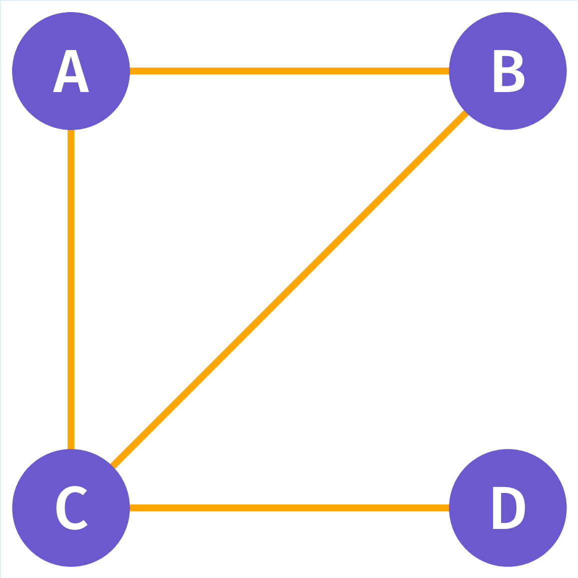


The Book of Why

Graphs

More formally

In graph theory, a graph is a collection of nodes connected by edges.

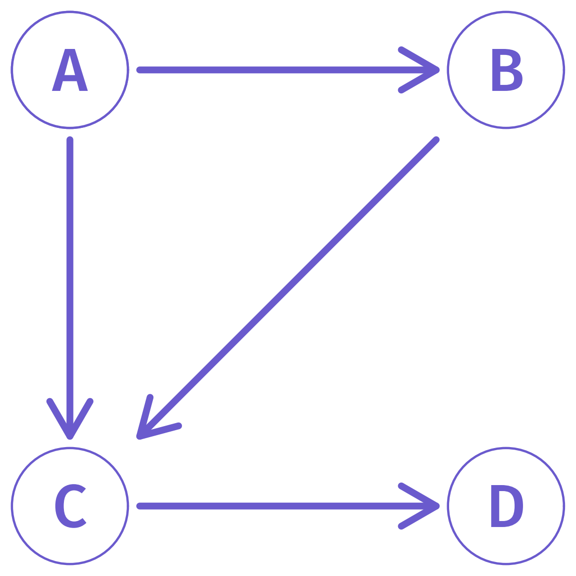


- Nodes(结点) connected by an edge(边或者连线) are called adjacent(邻近).
- Paths run along adjacent nodes, *e.g.*, A – B – C.
- The graph above is undirected, since the edges don't have direction.

Graphs

Directed

Directed graphs have edges with direction.

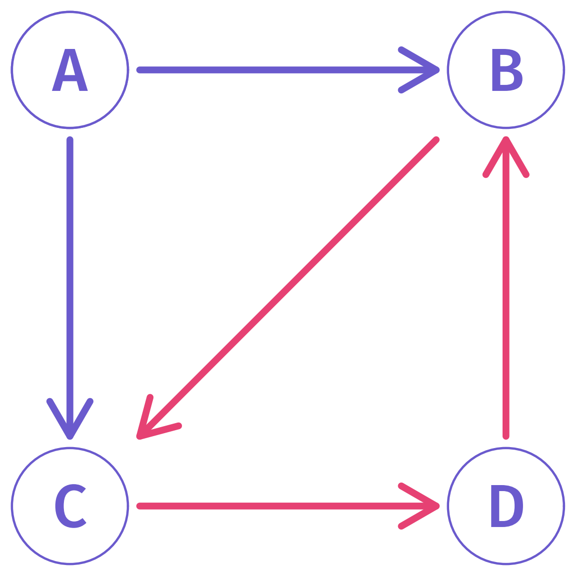


- Directed paths follow edges' directions, *e.g.*, $\mathrm{A} \to \mathrm{B} \to \mathrm{C}$.
- Nodes that precede a given node in a directed path are its ancestors.
- The opposite: descendants come after the node, *e.g.*, $\mathrm{D} = \mathrm{de}(\mathrm{C})$.

Graphs

Cycles

If a node is its own descendant (*e.g.*, $\mathrm{de}(\mathrm{D}) = \mathrm{D}$), your graph has a cycle.

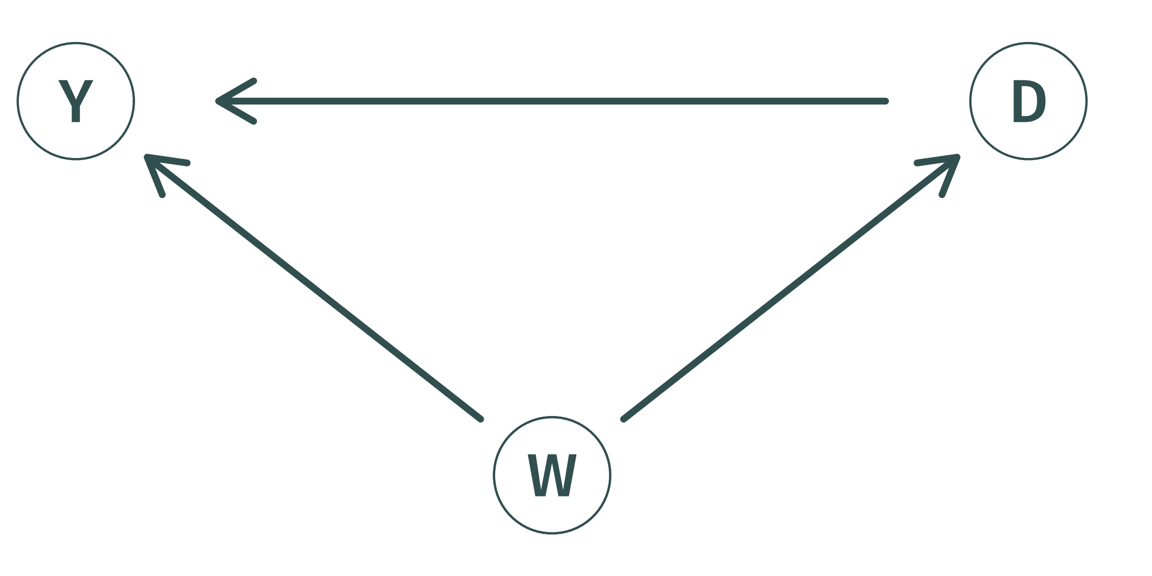


If your directed graph does not have any cycles, then you have a directed acyclic graph (DAG).

OVB in a DAG

OVB in a DAG

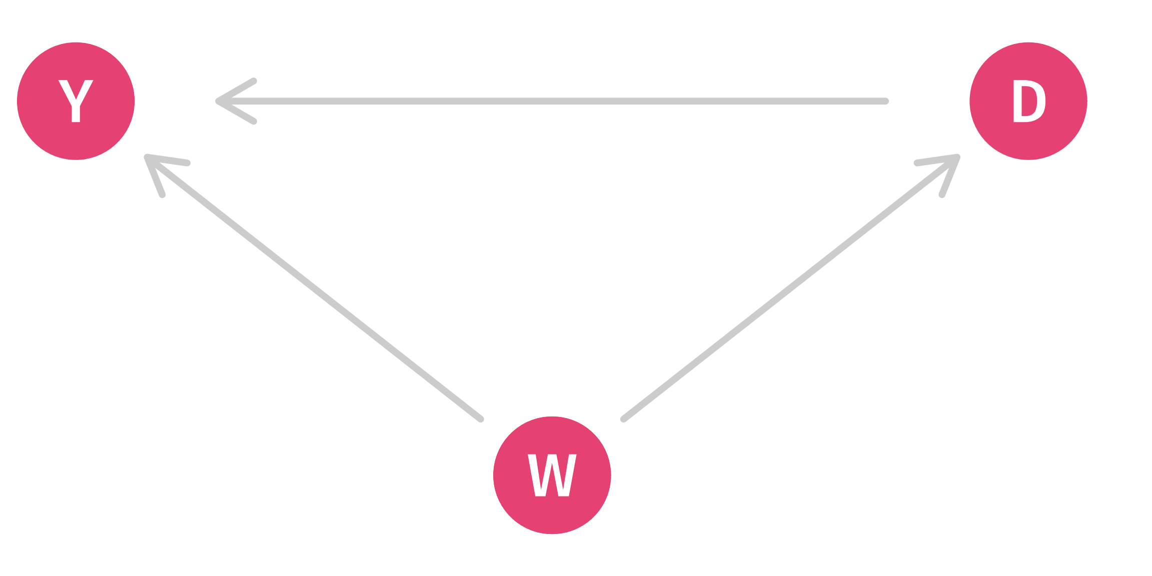
Example Omitted-variable bias in a DAG



A pretty standard DAG.

OVB in a DAG

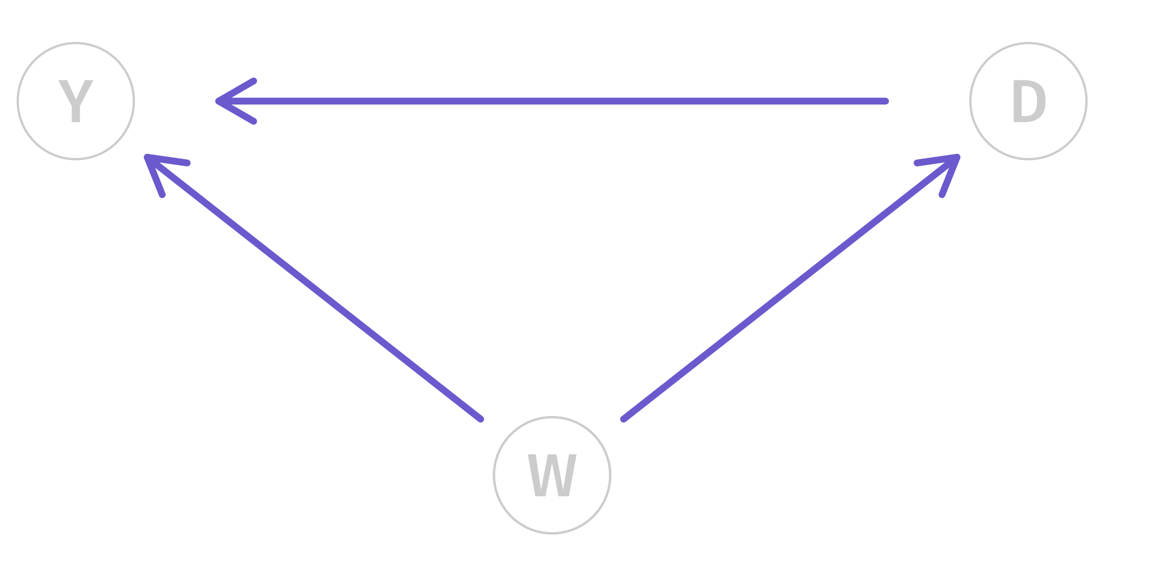
Example Omitted-variable bias in a DAG



Nodes are random variables.

OVB in a DAG

Example Omitted-variable bias in a DAG

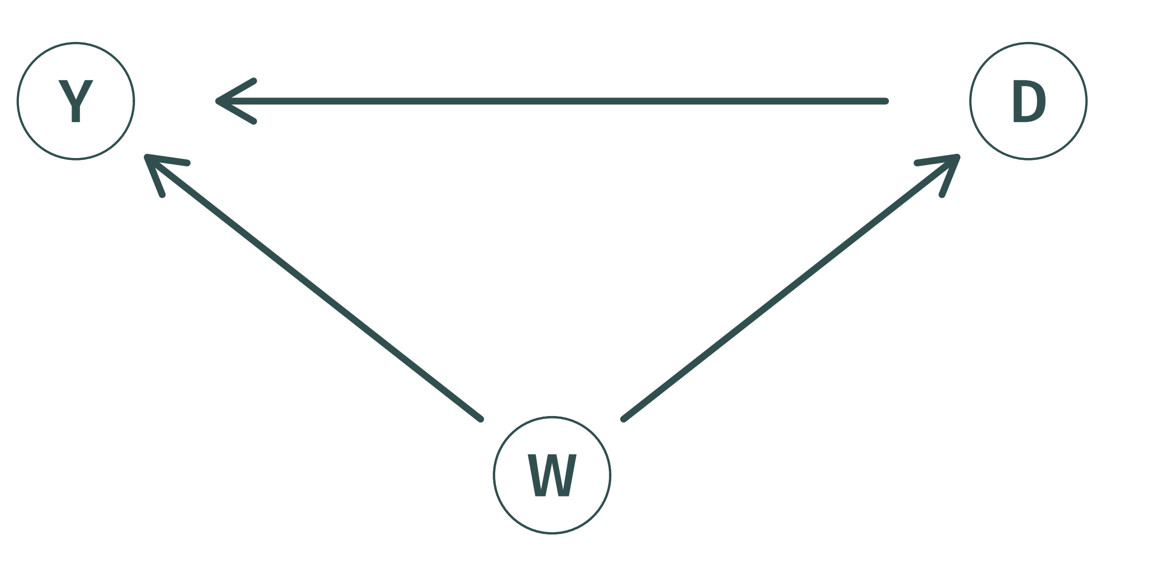


Edges depict causal links. Causality flows in the direction of the **arrows**.

- Connections matter!
- Direction matters (for causality).
- Non-connections also matter! (More on this topic soon.)

OVB in a DAG

Example Omitted-variable bias in a DAG



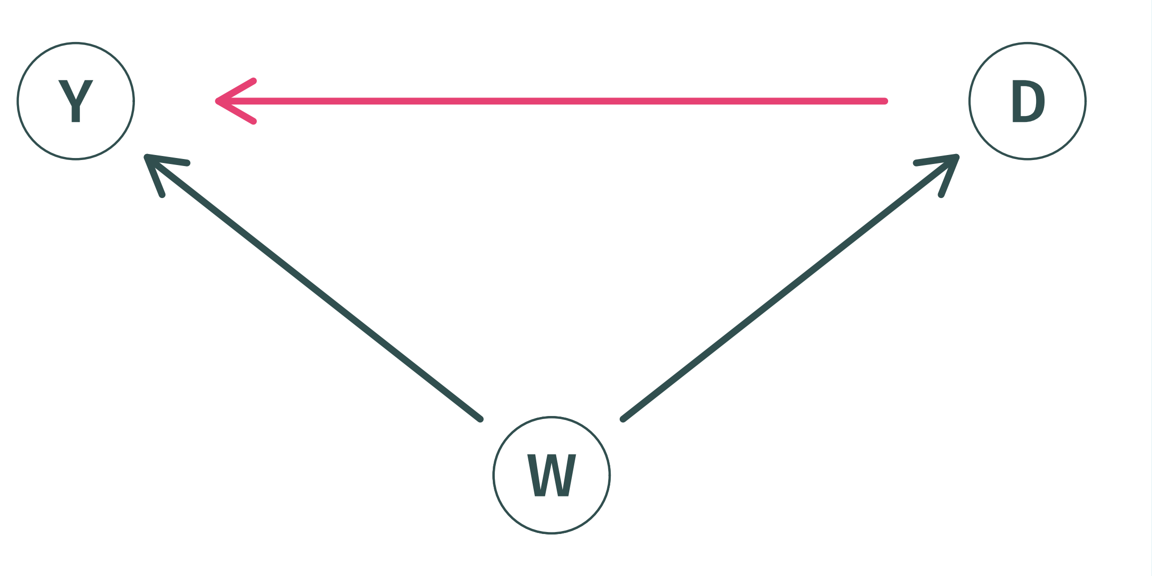
Here we can see that **Y** is affected by both **D** and **W**.

W also affects **D**.

Q How does this graph exhibit OVB?

OVB in a DAG

Example Omitted-variable bias in a DAG

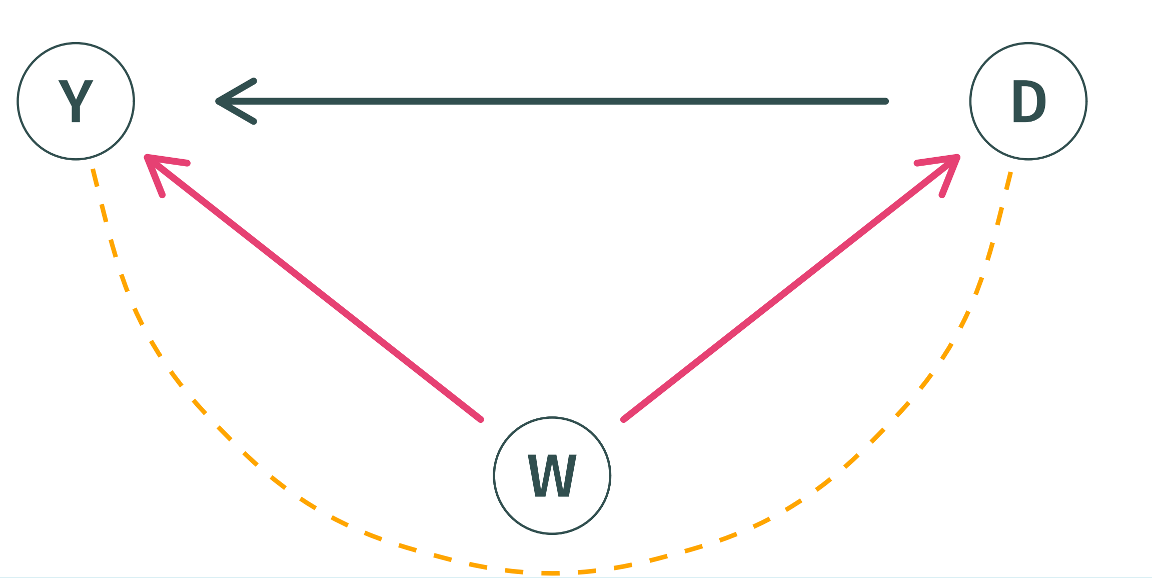


There are two pathways from D to Y.

1. The path from D to Y $(D \to Y)$ is our casual relationship of interest.

OVB in a DAG

Example Omitted-variable bias in a DAG

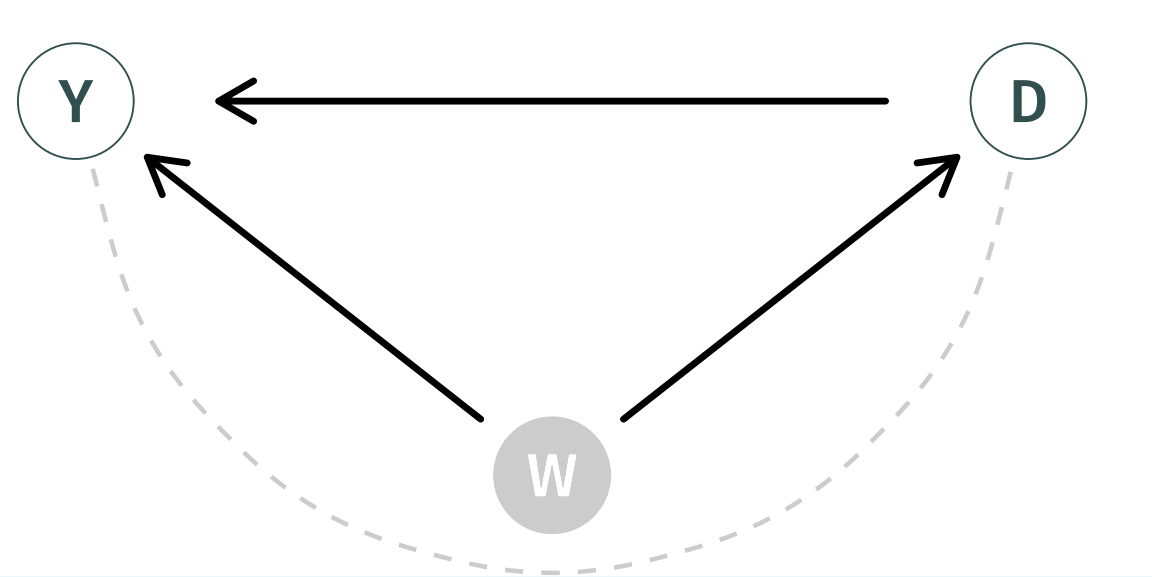


- There are two pathways from **D** to **Y**.

1. The path from **D** to **Y** ($D \to Y$) is our casual relationship of interest.
2. The path ($Y \leftarrow W \to D$) creates a non-causal association btn **D** and **Y**.

OVB in a DAG

Example Omitted-variable bias in a DAG



- There are two pathways from **D** to **Y**.

1. The path from **D** to **Y** $(D \rightarrow Y)$ is our casual relationship of interest.
2. The path $(Y \leftarrow W \rightarrow D)$ creates a non-causal association btn **D** and **Y**.

- To shut down this pathway creating a non-causal association, we must **condition on W**. Sound familiar?

DAGs in probability

The origin story

Many developments in *causal graphical models* came from work in probabilistic graphical models—especially Bayesian networks.

Recall what you know about joint probabilities:

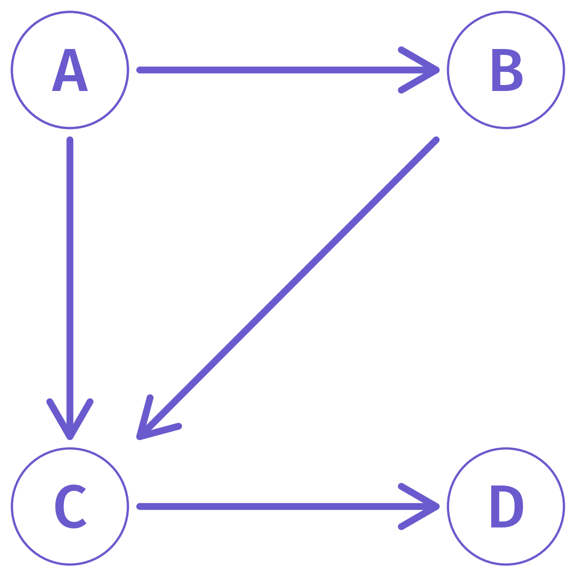
$$
\begin{aligned}
2 \qquad & P(x_1, x_2) = P(x_1)P(x_2|x_1) \\
3 \qquad & P(x_1, x_2, x_3) = P(x_1)P(x_2, x_3|x_1) = P(x_1)P(x_2|x_1)P(x_3|x_2, x_1) \\
\vdots \qquad & \\
n \qquad & P(x_1, x_2, \ldots, x_n) = P(x_1)\prod_{i=2}^{n} P(x_i|x_{i-1}, \ldots, x_1)
\end{aligned}
$$

This final product can include *a lot* of terms.
E.g., even when x_i are binary, $P(x_4|x_3, x_2, x_1)$ requires $2^3 = 8$ parameters.

DAGs in probability

Thinking locally

DAGs help us think through simplifying $P(x_k|x_{k-1}, x_{k-2}, \ldots, x_1)$.



Given a prob. dist. and a DAG, can we assume some independencies?

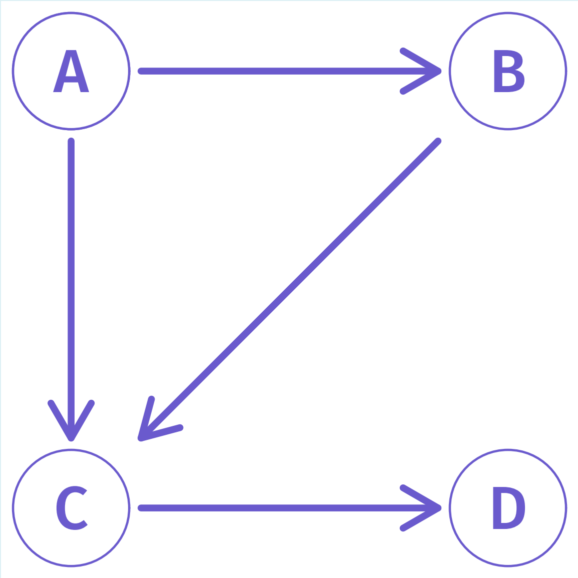
Given C, is it reasonable to assume D is independent of A and B?

Local Markov

This intuitive approach *is* the Local Markov Assumption

> Given its parents in the DAG, a node X is independent of all of its non-descendants.

Ex. Consider the DAG to the right:



With the Local Markov Assumption,
$P(\mathrm{D}|\mathrm{A}, \mathrm{B}, \mathrm{C})$ simplifies to $P(\mathrm{D}|\mathrm{C})$.

Conditional on its parent (C),
D is independent of A and B.

DAGs in probability

Local Markov and factorization

The Local Markov Assumption is equiv. to Bayesian Network Factorization

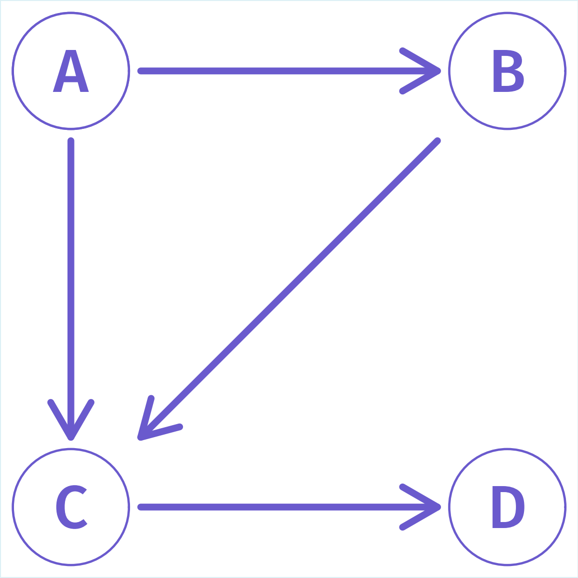
> For prob. dist. P and DAG G, P factorizes according to G if
>
> $$P(x_1, \ldots, x_n) = \prod_i P(x_i | \mathrm{pa}_i)$$
>
> where pa_i refers to x_i's parents in G.

Bayesian network factorization is also called *the chain rule for Bayesian networks* and *Markov compatibility*.

DAGs in probability

Factorize!

You can now (more easily) factorize the DAG/dist. below! (You're welcome.)



Factorization via B.N. chain rule

$$
\begin{aligned}
P(\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}) \\
&= \prod_i P(x_i | \mathrm{pa}_i) \\
&= P(\mathrm{A}) P(\mathrm{B}|\mathrm{A}) P(\mathrm{C}|\mathrm{A}, \mathrm{B}) P(\mathrm{D}|\mathrm{C})
\end{aligned}
$$

DAGs in probability

Independence

What have we learned so far? (Why should you care about this stuff?)

Local Markov and Bayesian Network Factorization tell us abount independencies within a probability distribution implied by the given DAG.

You're now able to say something about which variables are *independent*.

There's more: Great start, but there's more to life than independence.
We also want to say something about *dependence*.

DAGs in probability

Dependence

We need to strengthen our Local Markov assumption to be able to interpret adjacent nodes as dependent.

(*I.e.*, add it to our small set of assumptions.)

The Minimality Assumption†

> 1. Local Markov Given its parents in the DAG, a node X is independent of all of its non-descendants.
> 2. *(NEW)* Adjacent nodes in the DAG are dependent.

With the minimality assumption, we can learn both dependence and independence from connections (or non-connections) in a DAG.

† The name minimality refers to the minimal set of independencies for P and G—we cannot remove any more edges from the graph (while staying Markov compatible with G).

DAGs in probability

Causality

We need one last assumption move DAGs from *statistical* to *causal* models.

Strict Causal Edges Assumption

> Every parent is a direct cause of each of its children.

For Y, the set of *direct causes* is the set of variables to which Y responds.

This assumption actually strengthens the second part of Minimality:

> 2. Adjacent nodes in the DAG are dependent.

DAGs in probability

Assumptions

Thus, we only need two assumptions to turn DAGs into causal models:

1. Local Markov Given its parents in the DAG, a node X is independent of all of its non-descendants.
2. Strict Causal Edges Every parent is a direct cause of each of its children.

Not bad, right?

DAGs in probability

Flows

Brady Neal emphasizes the flow(s) of association and causation in DAGs, and I find it to be a super helpful way to think about these models.

Flow of association refers to whether two nodes are associated (statistically dependent) or not (statistically independent).

We will be interested in unconditional and conditional associations.

DAGs in probability

Building blocks

We will run through a few simple *building blocks* (DAGs) that make up more complex DAGs.

For each simple DAG, we want to ask a few questions:

1. Which nodes are unconditionally or conditionally **independent**?†
2. Which nodes are **dependent**?
3. What is the **intuition**?

† To prove A and B are conditionally independent, we can show $P(\mathrm{A}, \mathrm{B} | \mathrm{C})$ factorizes as $P(\mathrm{A} | \mathrm{C}) P(\mathrm{B} | \mathrm{C})$.

Building block 1: **Two unconnected nodes**

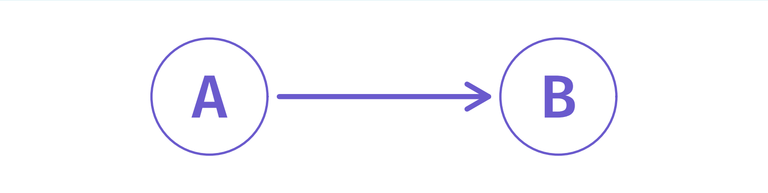
Intuition: A and B appear independent—no link between the nodes.

Proof: By Bayesian network factorization,

$$P(\mathrm{A}, \mathrm{B}) = P(\mathrm{A})P(\mathrm{B})$$

(since neither node has parents). ✓

Building block 2: **Two connected nodes**

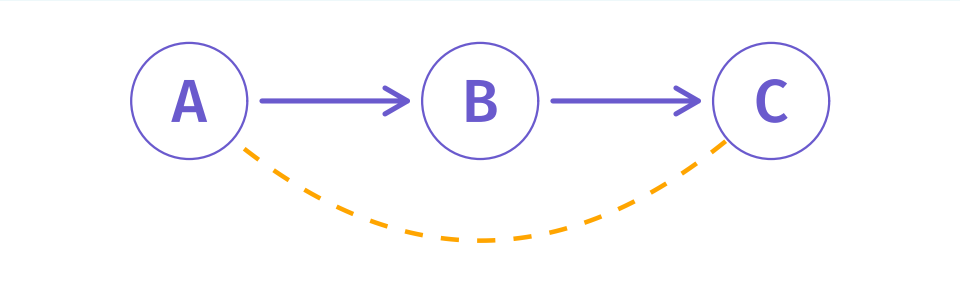


Intuition: A "is a cause of" B: there is clear (causal) dependence.†

Proof: By the Strict Causal Edges Assumption, every parent (here, A) is a direct cause of each of its children (B). ✓

† I'm not a huge fan of the "is a cause of" wording, but it appears to be (unfortunately) common in this literature. IMO, “A causes (or affects) B” would be clearer (and more grammatical), but no one asked me. One argument for "a cause of" (vs. "causes") is it emphasizes that events often have multiple causes.

Building block 3: Chains



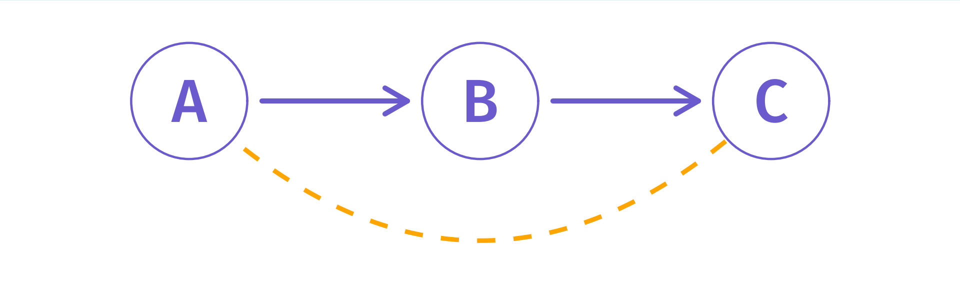
Intuition: We already showed two connected nodes are dependent:

- A and B are dependent.
- B and C are dependent.

The question is whether A and C are dependent:
Does association flow from A to C through B?

The answer *generally*† is "yes": changes in A typically cause changes in C.
† Section 2.2 of Pearl, Glymour, and Jewell provides a "pathological" example of "intransitive dependence". It's basically when A induces variation in B that is not relevant to C outcome.

Building block 3: **Chains**

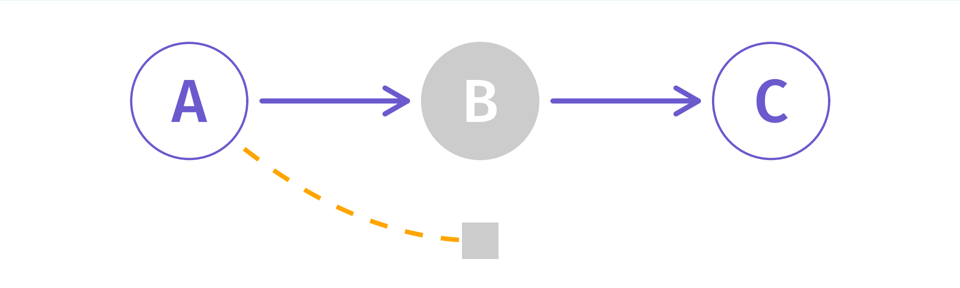


Proof: Here's the unsatisfying part.

Without more assumptions, we can't *prove* this association of A and C.

We'll think of this as a potential (even likely) association.

Building block 3: **Chains with conditions**



Q How does conditioning on B affect the association between A and C?

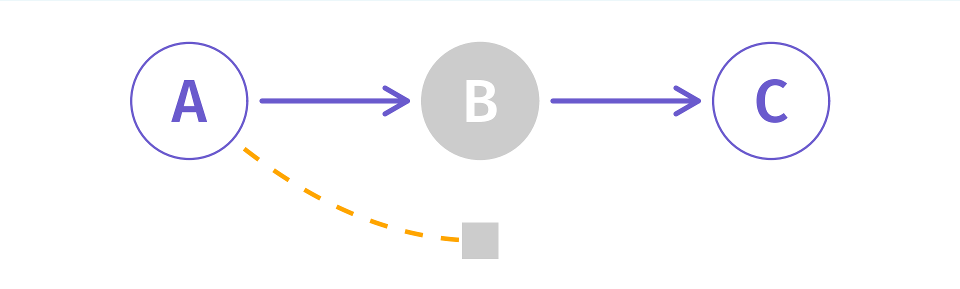
Intuition:

1. A affects C by changing B.
2. When we hold B constant, A cannot "reach" C.

We've blocked the path of association between A and C.

Conditioning blocks the flow of association **in chains**. ("Good" control!)

Building block 3: **Chains with conditions**



Proof: We want to show A and C are independent conditional on B, *i.e.*, $P(\mathrm{A},\mathrm{C}|\mathrm{B}) = P(\mathrm{A}|\mathrm{B})P(\mathrm{C}|\mathrm{B})$.

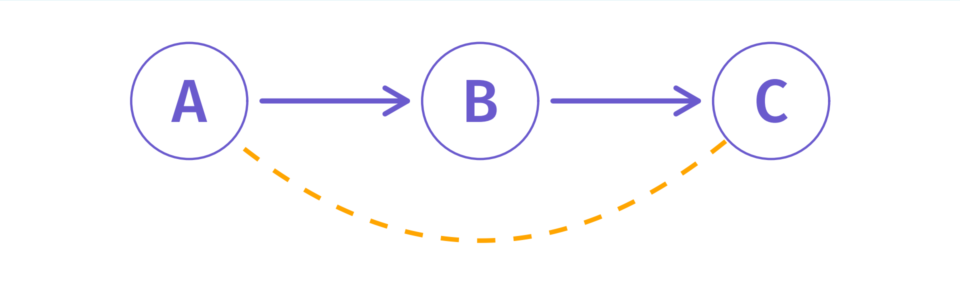
Start with BN factorization: $P(\mathrm{A},\mathrm{B},\mathrm{C}) = P(\mathrm{A})P(\mathrm{B}|\mathrm{A})P(\mathrm{C}|\mathrm{B})$.

Now apply Bayes' rule for the LHS of our goal: $P(\mathrm{A},\mathrm{C}|\mathrm{B}) = \frac{P(\mathrm{A},\mathrm{B},\mathrm{C})}{P(\mathrm{B})}$.

And substitute our factorization into the Bayes' rule expression:

$$P(\mathrm{A},\mathrm{C}|\mathrm{B}) = \frac{P(\mathrm{A})P(\mathrm{B}|\mathrm{A})P(\mathrm{C}|\mathrm{B})}{P(\mathrm{B})} = P(\mathrm{A}|\mathrm{B})P(\mathrm{C}|\mathrm{B})$$ ✓ (Bayes rule again)

Building block 3: Chains

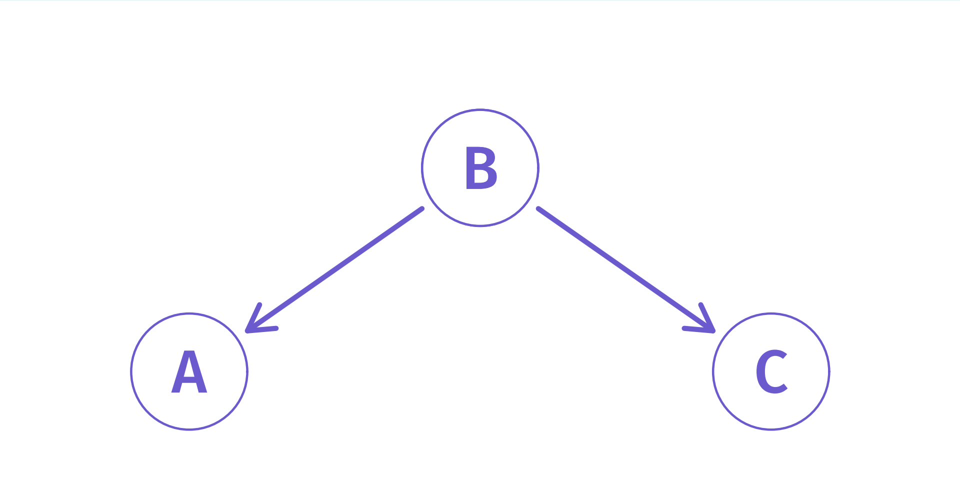


Note This association of A and C is not directional. (It is symmetric.)

On the other hand, causation **is** directional (and asymmetric).

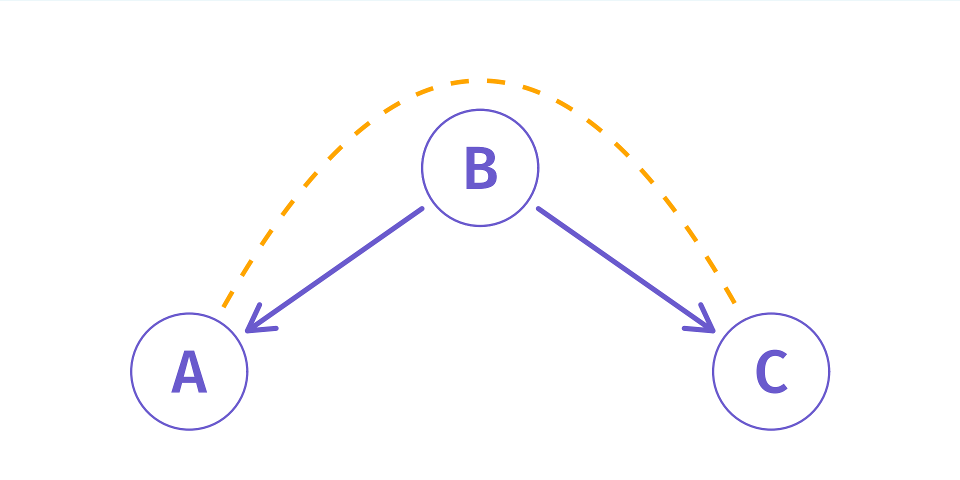
As you've been warned for years: Associations are not necessarily causal.

Building block 4: **Forks**



Forks are another very common structure in DAGs: $\mathrm{A} \leftarrow \mathrm{B} \rightarrow \mathrm{C}$.

Building block 4: **Forks**

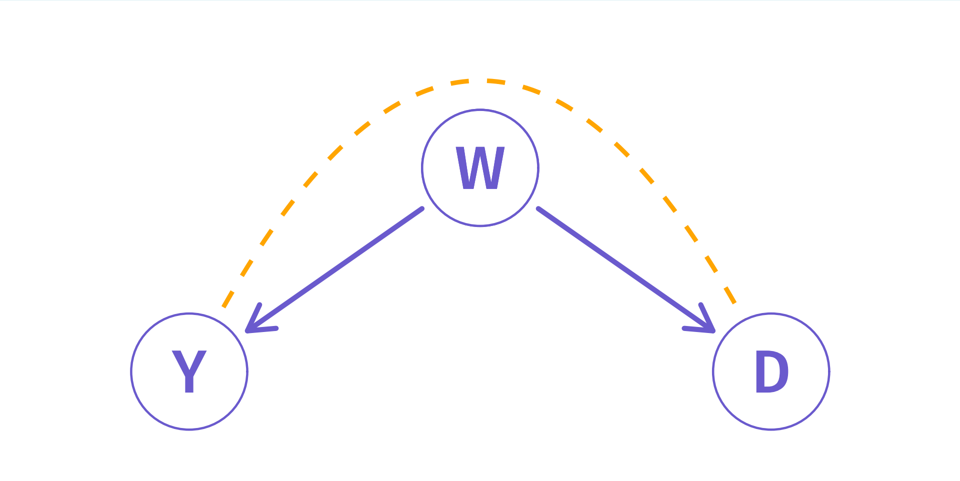


A and C are *usually* associated in forks. (As with chains.)

This chain of association follows the path $A \leftarrow B \rightarrow C$.

Intuition: B induces changes in A and B. An observer will see A change when C also changes—they are associated due to their common cause.

Building block 4: **Forks**

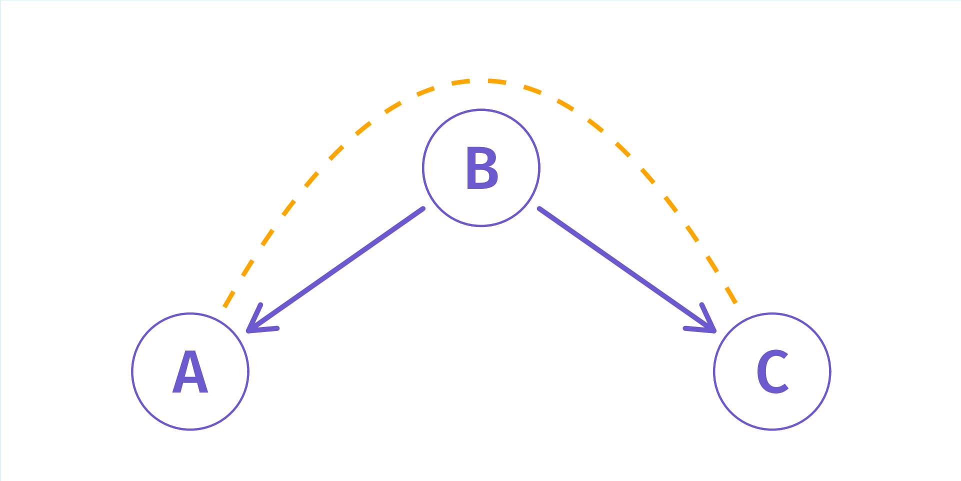


Another way to think about forks:

OVB when a treatment D does not affect the outcome Y.

Without controlling for W, Y and D are (usually) non-causally associated.

Building block 4: **Forks**

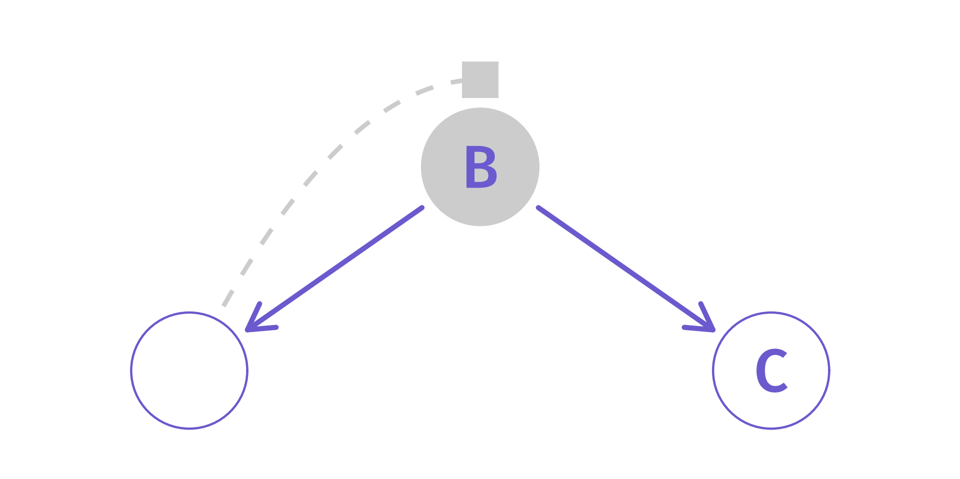


A and C are *usually* associated in forks. (As with chains.)

This chain of association follows the path $A \leftarrow B \rightarrow C$.

Proof: Same problem as chains: We can't show A and C are independent, so we assume they're likely (potentially?) dependent.

Building block 4: **Blocked forks**



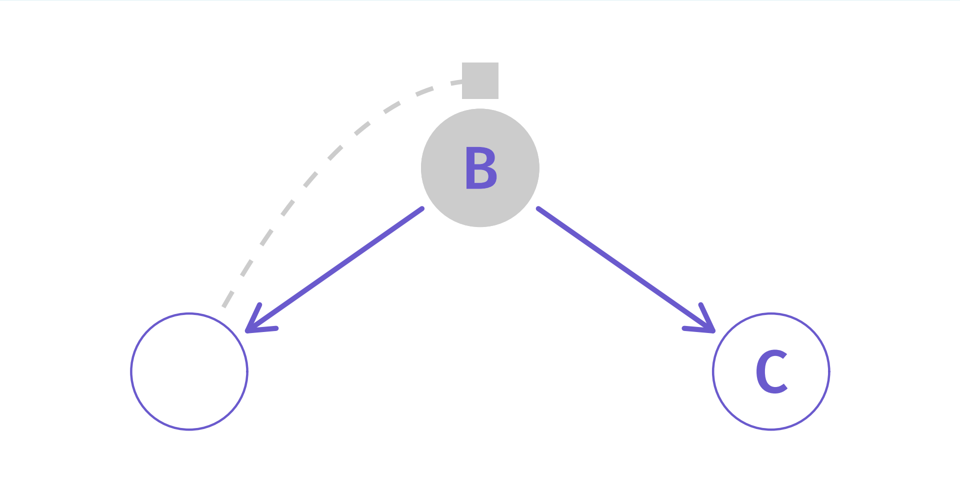
Conditioning on B makes A and C independent. (As with chains.)

Intuition: A and C are only associated due to their common cause B.

When we shutdown (hold constant) this common cause (B),
there is way for A and C to associate.

Also: Think about Local Markov. Or think about OVB.

Building block 4: **Blocked forks**



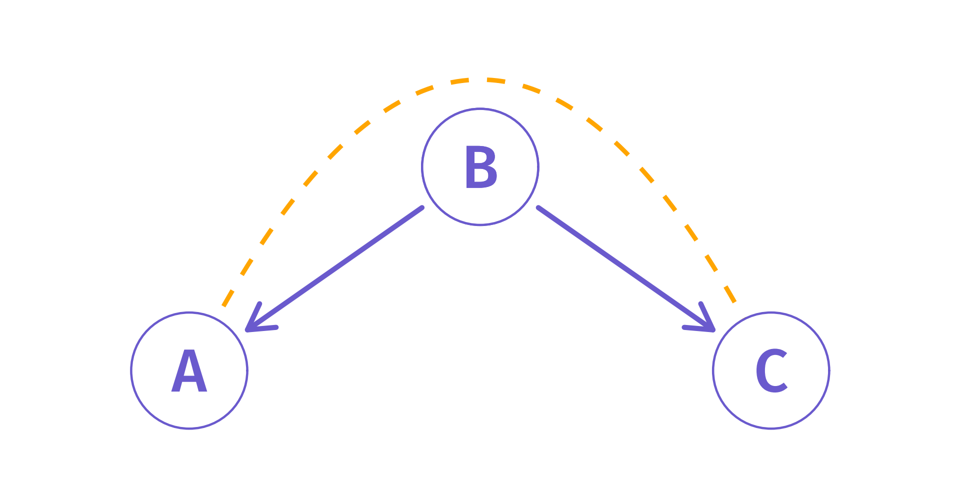
Proof: We want to show $P(\mathrm{A},\mathrm{C}|\mathrm{B}) = P(\mathrm{A}|\mathrm{B})P(\mathrm{C}|\mathrm{B})$.

Step 1: Bayesian net. factorization: $P(\mathrm{A},\mathrm{B},\mathrm{C}) = P(\mathrm{B})P(\mathrm{A}|\mathrm{B})P(\mathrm{C}|\mathrm{B})$

Step 2: Bayes' rule: $P(\mathrm{A},\mathrm{C}|\mathrm{B}) = \frac{P(\mathrm{A},\mathrm{B},\mathrm{C})}{P(\mathrm{B})}$

Step 3: Combine 2 & 1: $P(\mathrm{A},\mathrm{C}|\mathrm{B}) = \frac{P(\mathrm{A},\mathrm{B},\mathrm{C})}{P(\mathrm{B})} = P(\mathrm{A}|\mathrm{B})P(\mathrm{C}|\mathrm{B})$ ✓

Building block 4: **Forks**

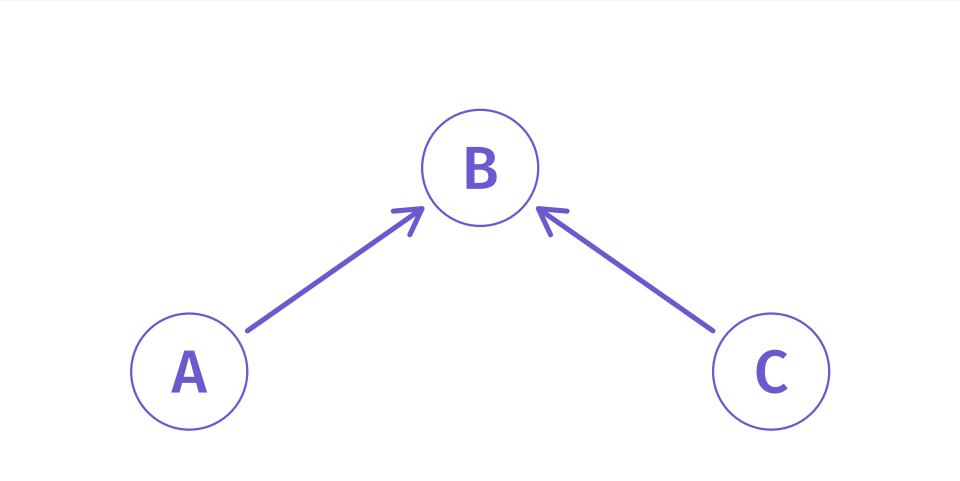


Two more items to emphasize:

1. **Association** need not follow paths' directions, *e.g.*, $\text{A} \leftarrow \text{B} \rightarrow \text{C}$.

2. **Causation** follows directed paths.

Building block 5: **Immoralities**



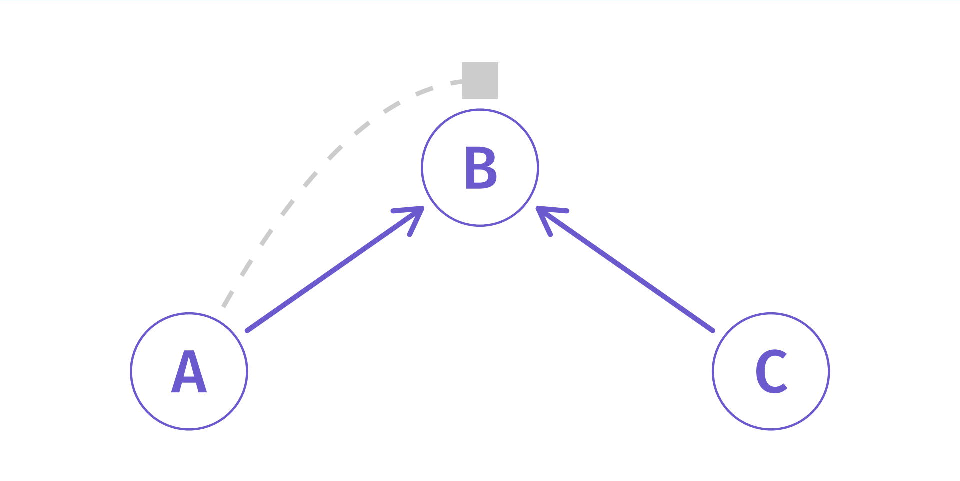
An immorality occurs when two nodes share a child without being otherwise connected.† $A \rightarrow B \leftarrow C$

The child (here: B) at the center of this immorality is called a collider.

Notice: An immorality is a fork with reversed directions of the edges.

† I'm not making this up.

Building block 5: **Immoralities**



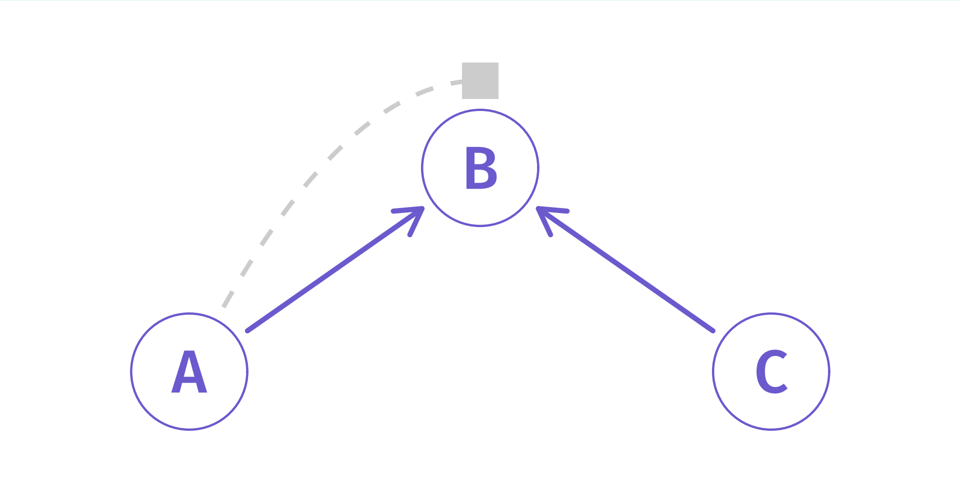
Q Are A and C independent?

A Yes. $A \perp\!\!\!\perp C$.

Intuition: Causal effects flow from A and C and stop there.

- Neither A nor C is a descendant of the other.
- A and C do not share any common causes.

Building block 5: **Immoralities**



Proof: Start with *marginalizing* dist. of A and C. Then BNF.

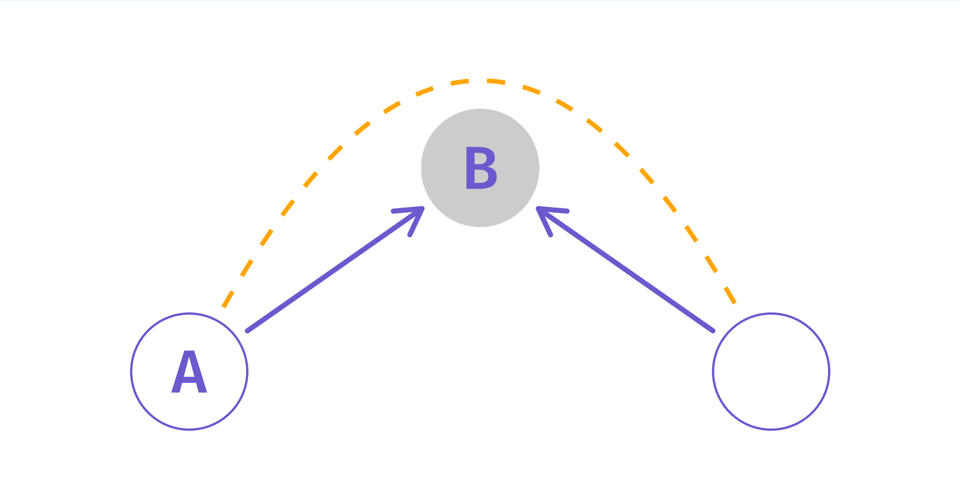
$$P(A, C) = \sum_B P(A, B, C)$$

$$P(A, C) = \sum_B P(A)P(C)P(B|A, C)$$

$$P(A, C) = P(A)P(C) \left(\sum_B P(B|A, C) = 1\right)$$

$$P(A, C) = P(A)P(C)$$ ✓ ($A \perp\!\!\!\perp C$ without conditioning)

Building block 5: **Immoralities with conditions**

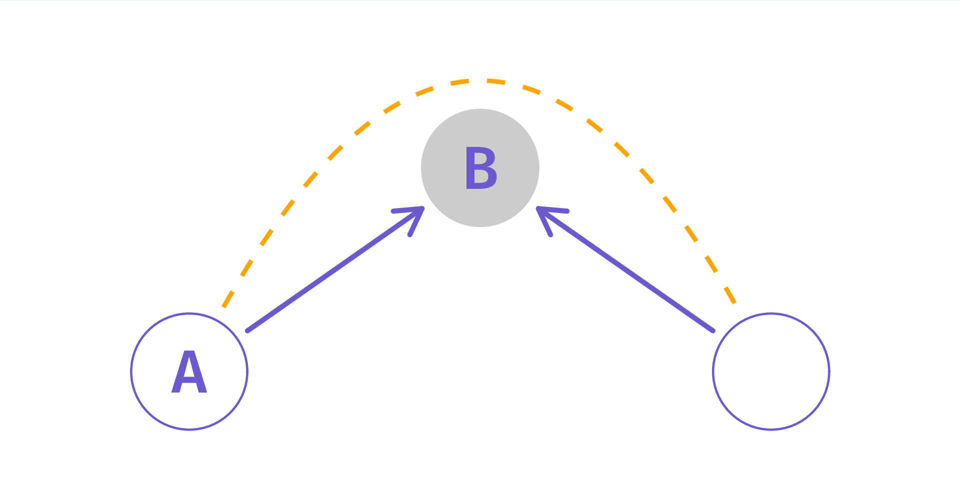


Q What happens when we condition on B?
A We unblock (or open) the previously blocked (closed) path.

While A and C are independent, they are conditionally dependent.

Important: When you condition on a collider, you open up the path.

Building block 5: **Immoralities with conditions**



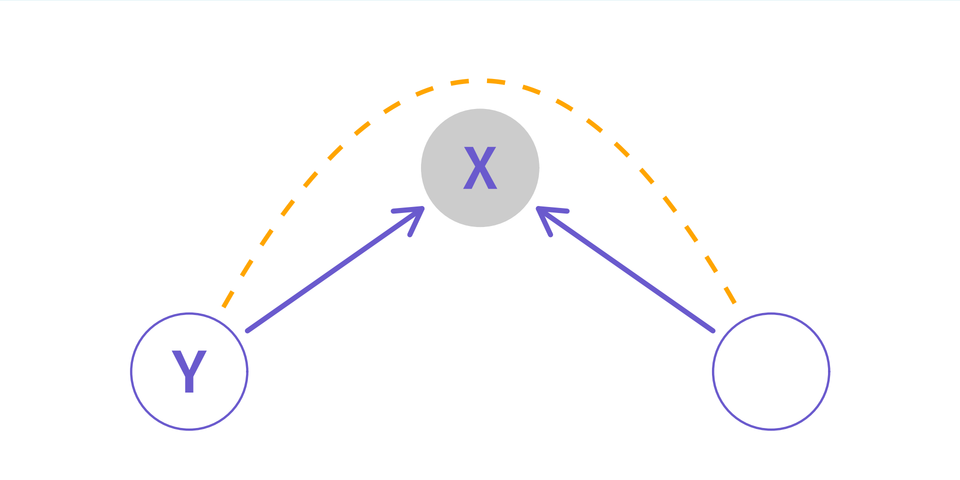
Intuition: B is a combination of A and C.

Conditioning on a value of B jointly constrains A and C—they can no longer move independently.

Example: Let A take on $\{0, 1\}$ and C take on $\{0, 1\}$ (independently).

Conditional on $B = 1$, A and C are perfectly negatively correlated.

Building block 5: **Immoralities with conditions**



In *MHE* vocabulary: The collider X is a *bad control*.

X is affected by both your treatment D and outcome Y.

The result: A spurious relationship between Y and D
Remember: they're actually (unconditionally) independent.

This spurious relationship is often called collider bias.

Blocked paths

Let's formally define a blocked path (blocking is important).

A path between X and Y is blocked by conditioning on a set of variables Z (possibly empty) if either of the following statements is true:

1. On the path, there is a **chain** $(\cdots \rightarrow \mathrm{W} \rightarrow \ldots)$ or a **fork** $(\cdots \leftarrow \mathrm{W} \rightarrow \ldots)$, and we condition on W $(\mathrm{W} \in \mathrm{Z})$.

2. On the path, there is a **collider** $(\cdots \rightarrow \mathrm{W} \leftarrow \ldots)$, and we *do not* condition on W $(\mathrm{W} \notin \mathrm{Z})$ or any of its **descendants** $\left(\mathrm{de}(\mathrm{W}) \not\subseteq \mathrm{Z}\right)$.

Association flows along unblocked paths.

d-separation and d-connected(-ness)

Finally, we'll define whether nodes are separated or connected in DAGs.

Separation: Nodes X and Y are d-separated by a set of nodes Z if all paths between X and Y are blocked by Z.

Notation for d-separation: $\mathrm{X} \perp\!\!\!\perp_G \mathrm{Y} | \mathrm{Z}$

Connection: If there is at least one path between X and Y that is unblocked, then X and Y are d-connected.

d-separation and causality

d-separation tells us that two nodes are not associated.

To measure the causal effect of X on Y:
We must eliminate non-causal association.

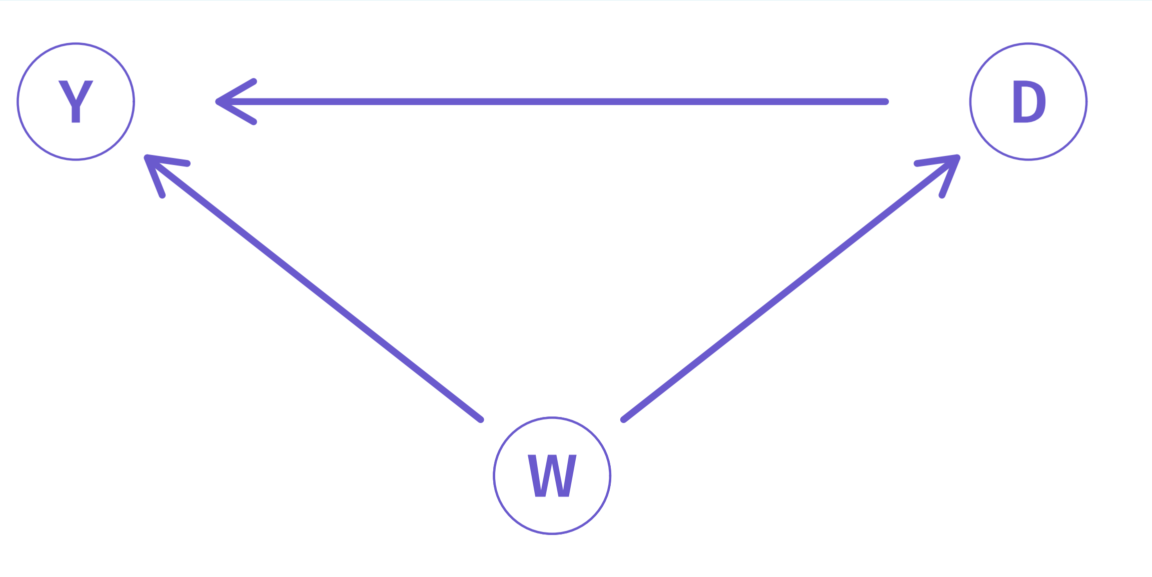
Putting these ideas together, here is our criterion to isolate causal effects:

> If we remove all edges flowing **out** of X (its causal effects),
> then X and Y should be d-separated.

This criterion ensures that we've closed the backdoor paths that generate non-causal associations between X and Y.

Examples

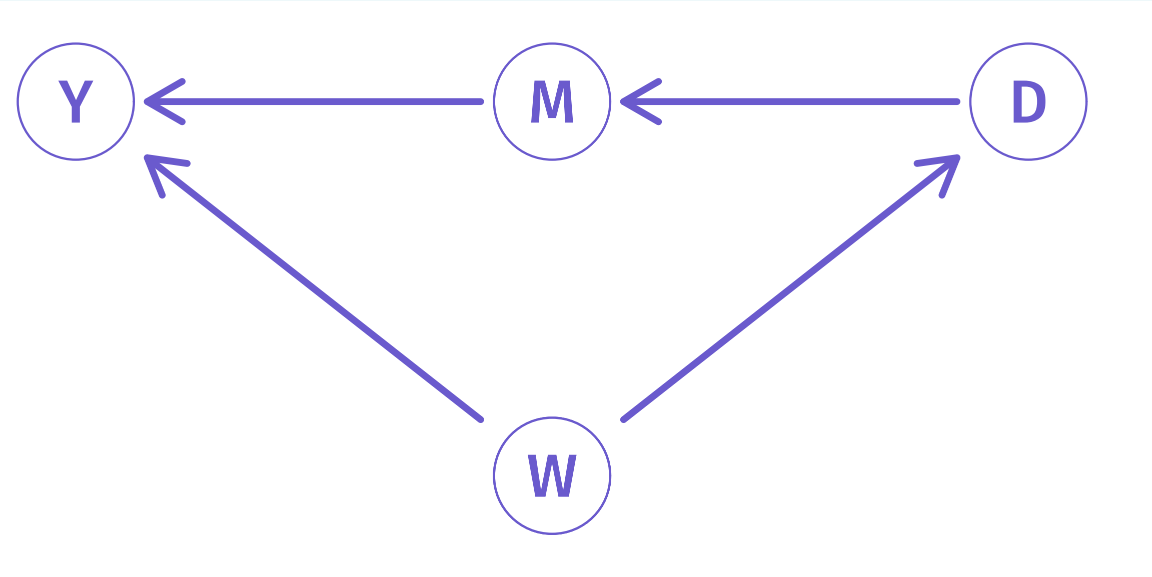
Example 1: **OVB**



Q OVB using DAG fundamentals: When can we isolate causal effects?

Example 2: **Mediation**

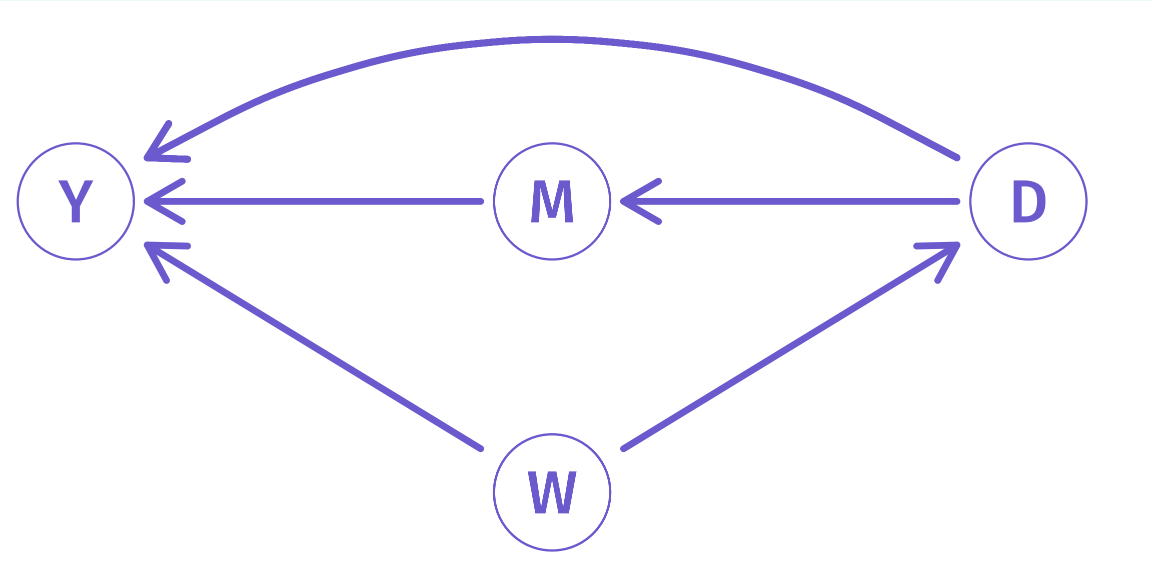
Here M is a mediator: it mediates the effect of D on Y.



Q1 What do we need to condition on to get the effect of D on Y?

Q2 What happens if we condition on W and M?

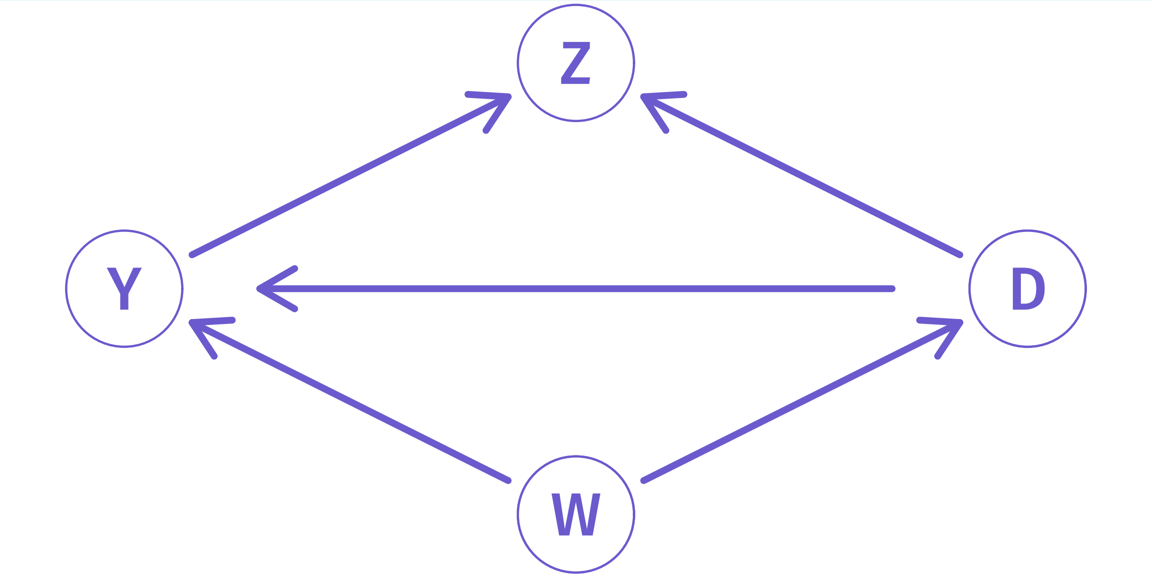
Example 3: **Partial mediation**



Q1 What do we need to condition on to get the effect of D on Y?

Q2 What happens if we condition on W and M?

Example 4: **Non-mediator descendants**

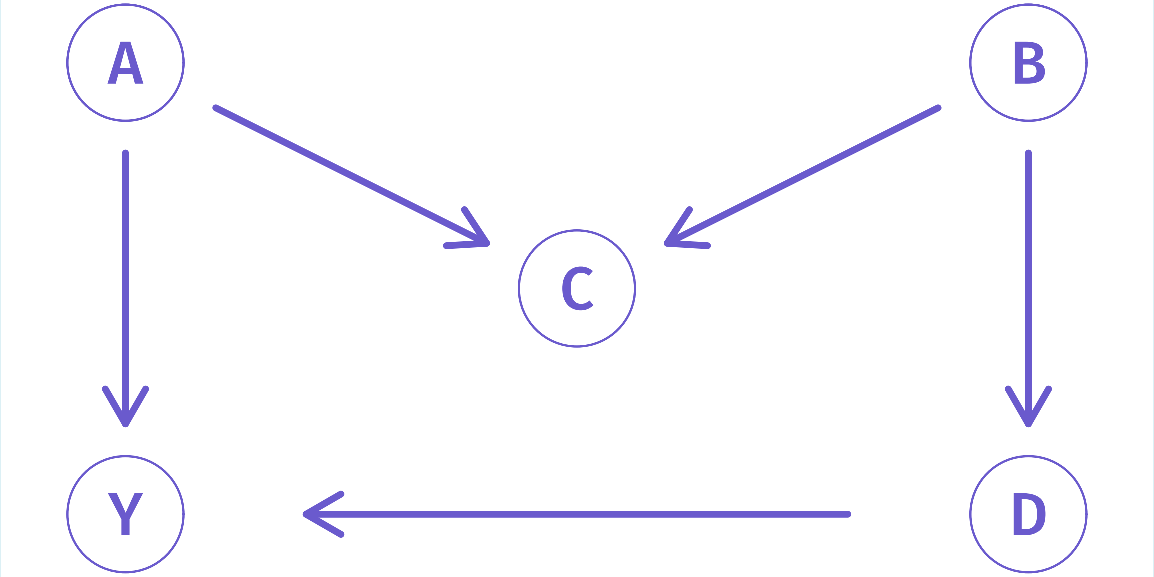


Q1 What do we need to condition on to get the effect of D on Y?

Q2 What happens if we condition on W and/or Z?

Example 5: **M-Bias**

Notice that C here is *not* a result of treatment (could be "pre-treatment").



Q1 What do we need to condition on to get the effect of D on Y?

Q2 What happens if we condition on C?

Q3 What happens if we condition on C along with B and/or C?

One more note:

DAGs are often drawn without "noise variables" (disturbances).

But they still exist—they're just "outside of the model."

DAGs

Limitations

So what can't DAGs do (well)?

- Simultaneity: Defined causality as unidirectional and prohibited cycles.
- Dynamics: You can sort of allow a variable to affect itself... $Y_{t=1} \rightarrow Y_{t=2}$.
- Uncertainty: DAGs are most useful when you can correctly draw them.