

Directed Acyclic Graphs

Manoel Horta Ribeiro
manoel@cs.princeton.edu

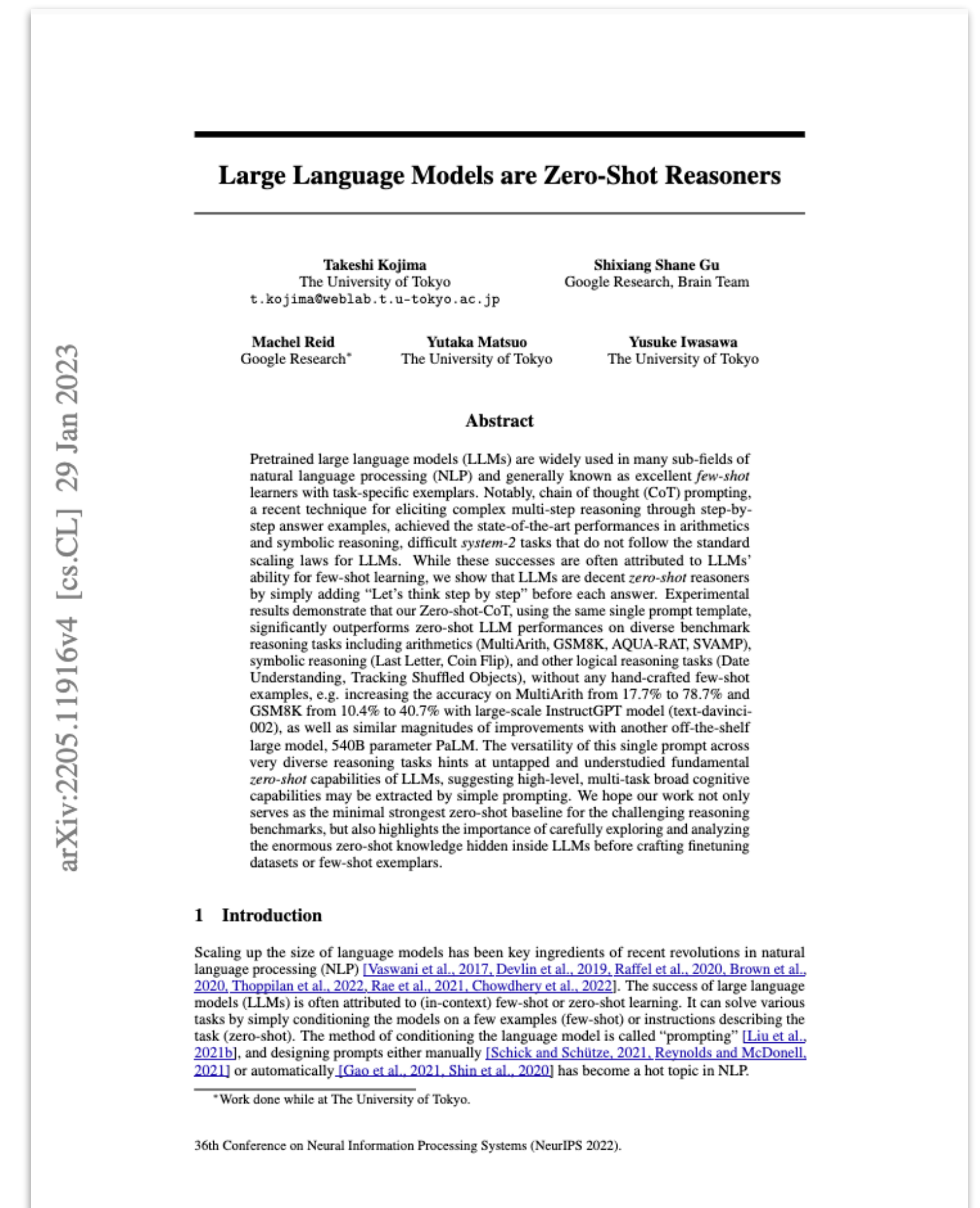
COS 424 / Fall 2026

Recap: Chain of Thought (CoT)

- In an influential paper, Kojima et al. (2022) proposed changing the way we prompt models.
- **Hypothesis:** Asking the model to reason improves performance.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there? A: The answer (arabic numerals) is

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there? **A: Let's think step by step.**



Recap: Experiments Rock

- We show we can estimate the ATE :

$$\hat{\tau} = \frac{1}{|\mathcal{D}_1|} \sum_{(i,r): T_{ir}=1} Y_{ir} - \frac{1}{|\mathcal{D}_0|} \sum_{(i,r): T_{ir}=0} Y_{ir}$$

- Exchangeability:

$$(Y_{ir}^0, Y_{ir}^1) \perp T_{ir}.$$

But at the same time, much can go wrong!

- Low power and multiple testing
 - Can't find an effect where there is one, or you find an effect when there's none.
- Non-compliance
 - Units do not actually receive the treatment they were assigned to!
- Attrition
 - Units assigned to the experiment are not observed in the outcomes.

And the observational case is worse!

- Suppose we wanted to assess the impact of adding “let’s think step by step” from *observational data*!
- People may be more likely to use CoT for harder problems.
- This means no exchangeability:

$$(Y_{ir}^0, Y_{ir}^1) \not\perp T_{ir}.$$

Recap: Average Treatment Effect

$$\begin{aligned}ATE &= E[Y_{ir}^1 - Y_{ir}^0] \\&= E[Y_{ir}^1] - E[Y_{ir}^0] && (1) \text{ Linearity of Expectations} \\&= \boxed{E[Y_{ir}^1 \mid T_{ir} = 1] - E[Y_{ir}^0 \mid T_{ir} = 0]} && (2) \text{ Exchangeability} \\&= E[Y_{ir} \mid T_{ir} = 1] - E[Y_{ir} \mid T_{ir} = 0] && (3) \text{ Consistency}\end{aligned}$$

Conditional Exchangeability

- Let Y and T be binary random variables specifying treatments and outcomes, and X be a confounder.
 - E.g., T is the prompt style, Y is whether the question is solved, and X is the question difficulty, again, “in the wild.”
- **Conditional Exchangeability** entails that:
$$Y^a \perp T \text{ for all } a \in \{0,1\} \mid X$$
 - I.e., within difficulty levels, treatments are if as random!

Average Treatment Effect Revisited

$$\begin{aligned}ATE &= E[Y_{ir}^1 - Y_{ir}^0] \\&= E[Y_{ir}^1] - E[Y_{ir}^0]\end{aligned}\quad (1) \text{ Linearity of Expectations}$$

$$= E_{X \sim P(X)}[E[Y_{ir}^1 | X]] - E_{X \sim P(X)}[E[Y_{ir}^0 | X]] \quad (2) \text{ Law of Iterated Expectations}$$

$$= E_{X \sim P(X)}[E[Y_{ir}^1 | X, T_{ir} = 1]] - E_{X \sim P(X)}[E[Y_{ir}^0 | X, T_{ir} = 0]] \quad (3) \text{ Conditional Exchangeability}$$

$$= E_{X \sim P(X)}[E[Y_{ir} | X, T_{ir} = 1]] - E_{X \sim P(X)}[E[Y_{ir} | X, T_{ir} = 0]] \quad (4) \text{ Consistency}$$

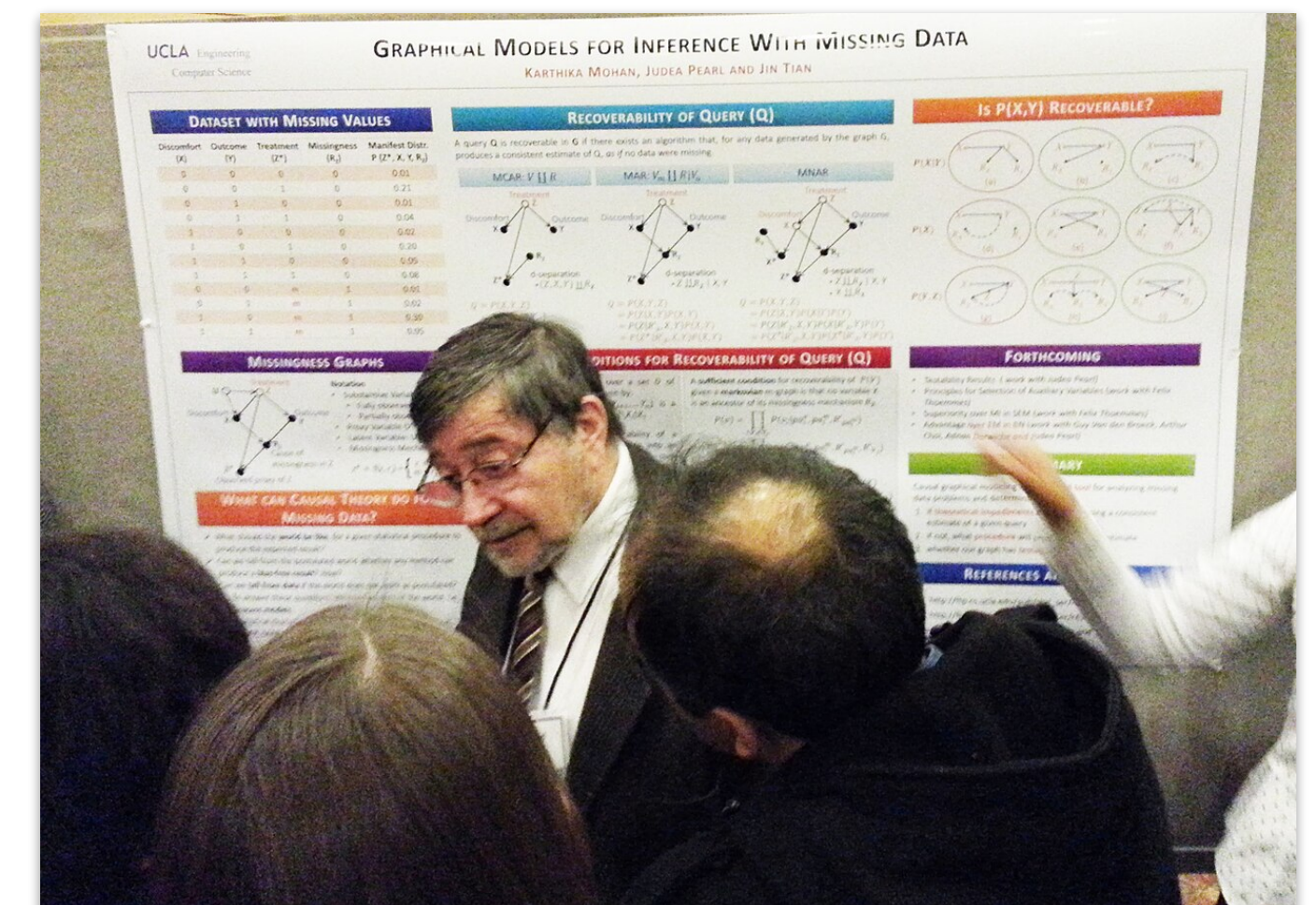
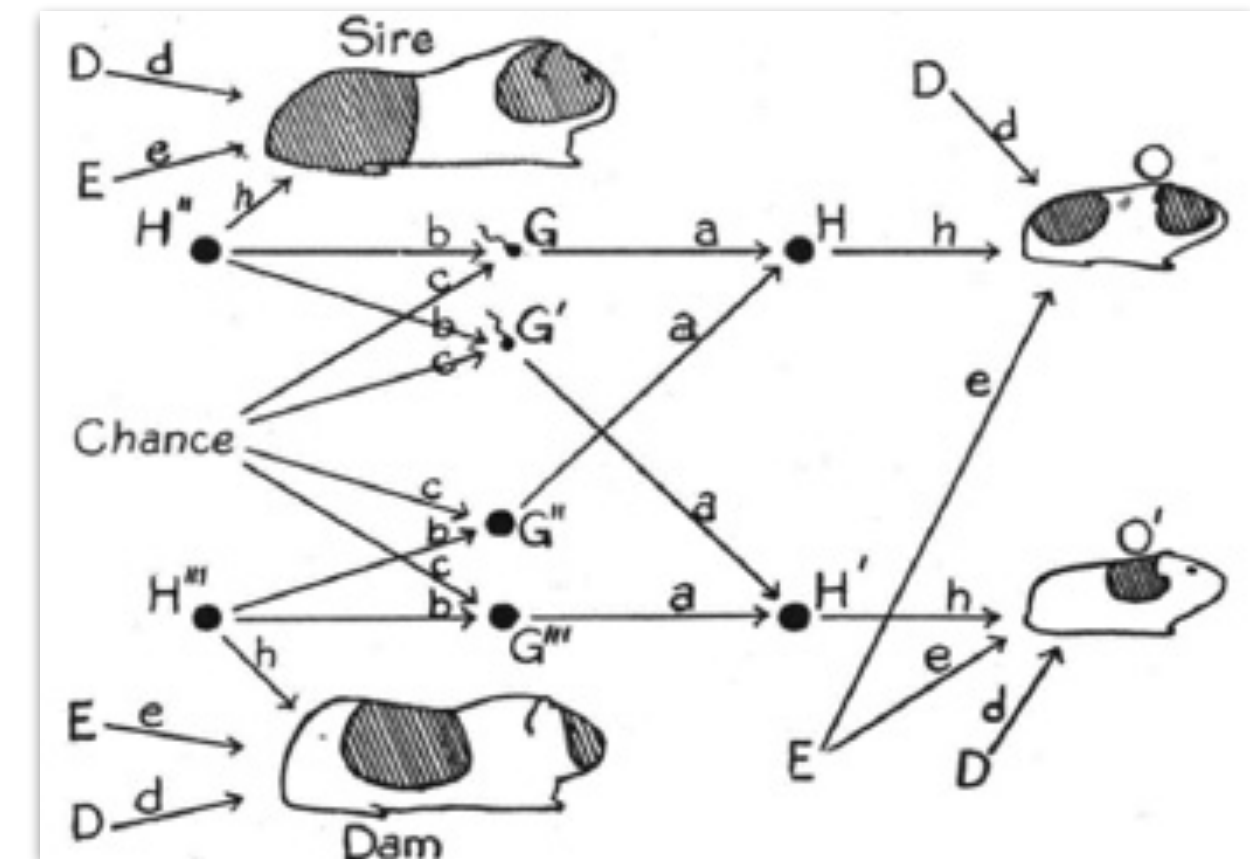
- Estimate the ATE at each level of the confounder.
- Average it according to the distribution of the population of interest ($X \sim P(x)$).

“The world is a complicated place”

- Naïve idea: controlling for everything!
 - Infants born to smokers were found to have a higher risk of mortality than infants born to non-smokers.
 - However, for infants with low birth weight, this relationship is reversed!
- If we “control” for birth weight, we would find that smoking decreases child mortality.
- Conditioning on birth weight mixes these causes and can spuriously suggest that smoking is protective.

DAGs to the rescue!

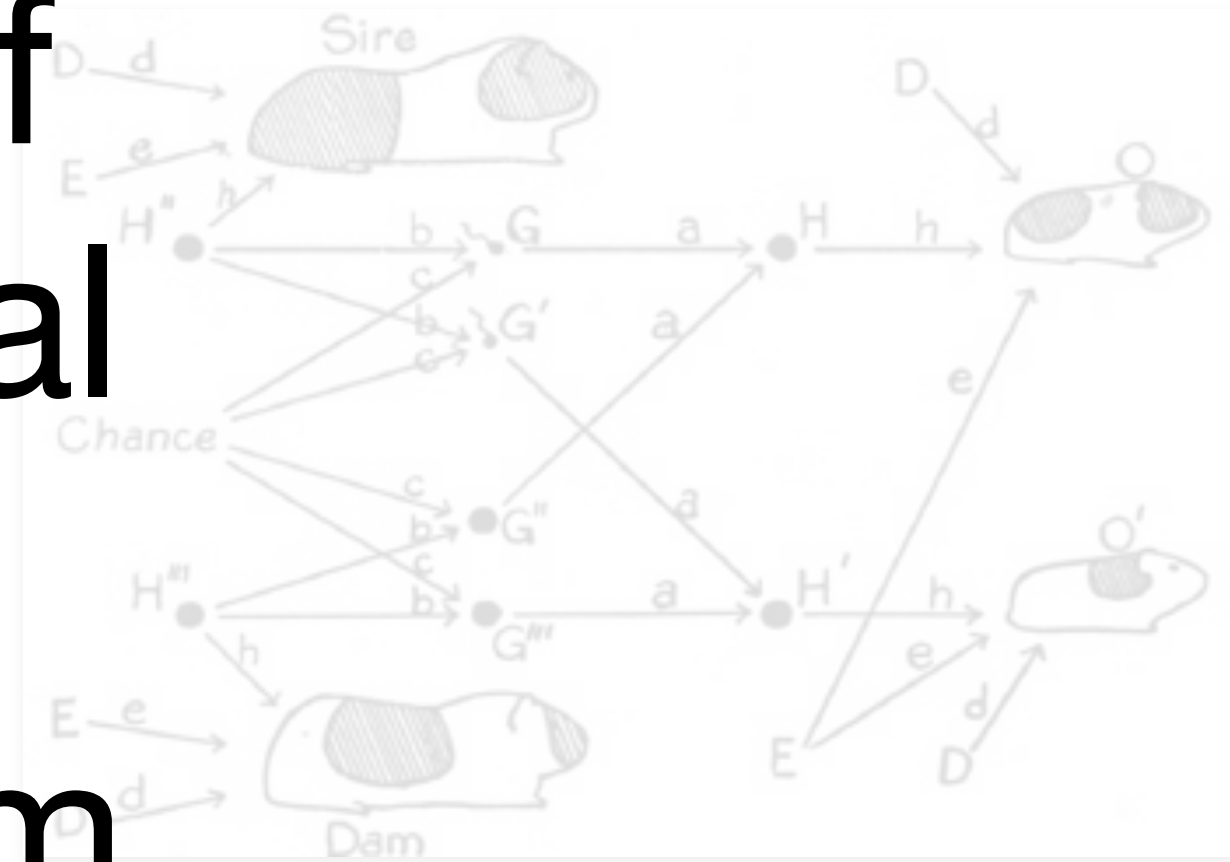
- Directed Acyclic Graphs are a great way to reason about how to identify causal effects!
- They date back to geneticist Sewall Wright (1921).
- Mostly developed alongside Judea Pearl's causal inference framework (structural causal model)



DAGs to the rescue!

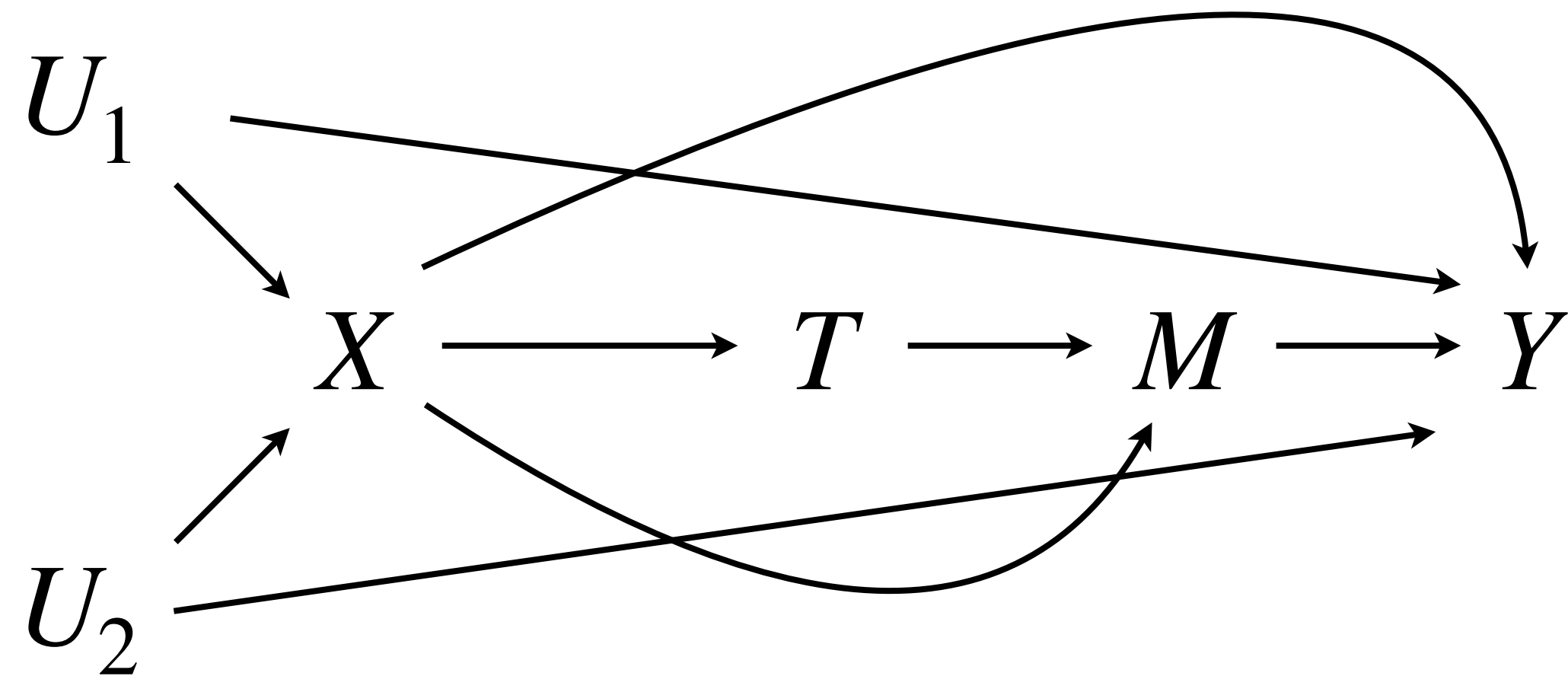
The possibility of separating causal and non-causal associations from data!

- Directed Acyclic Graphs are a great way to reason about how to identify causal effects!
- They date back to genetics: Sewall Wright (1921):
- Mostly developed alongside Judea Pearl's causal inference framework (structural causal model)



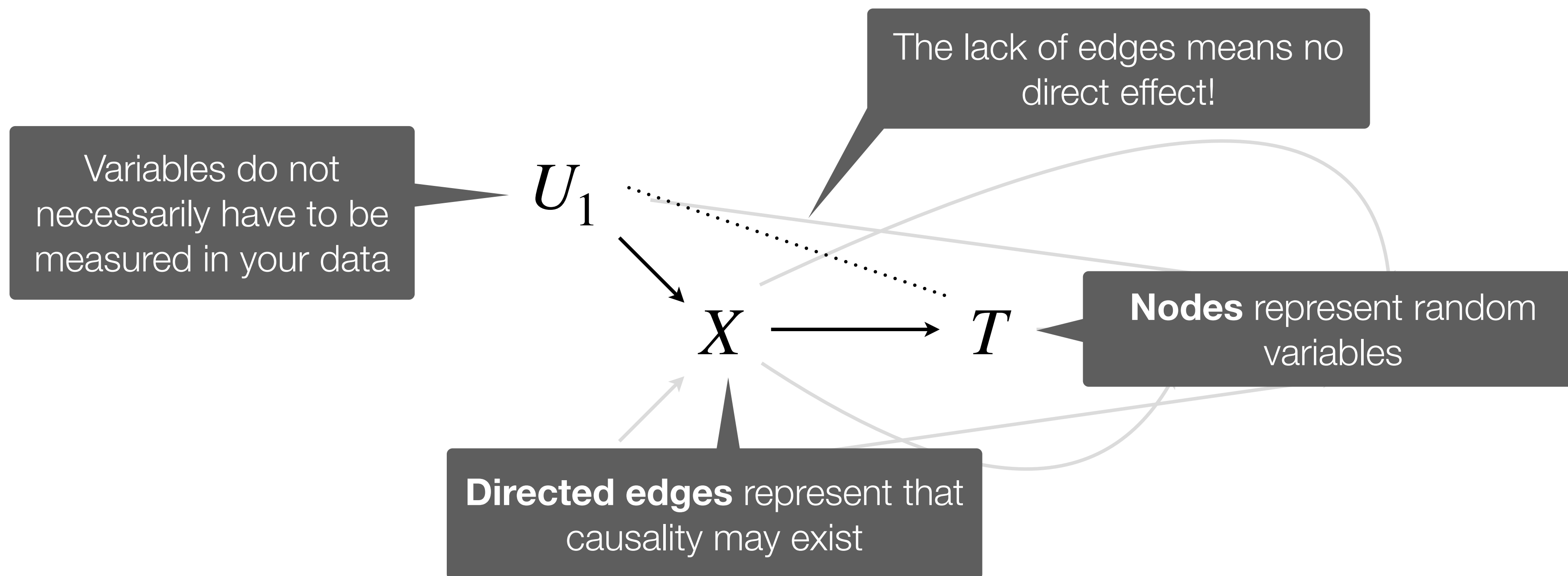
DAG Basics

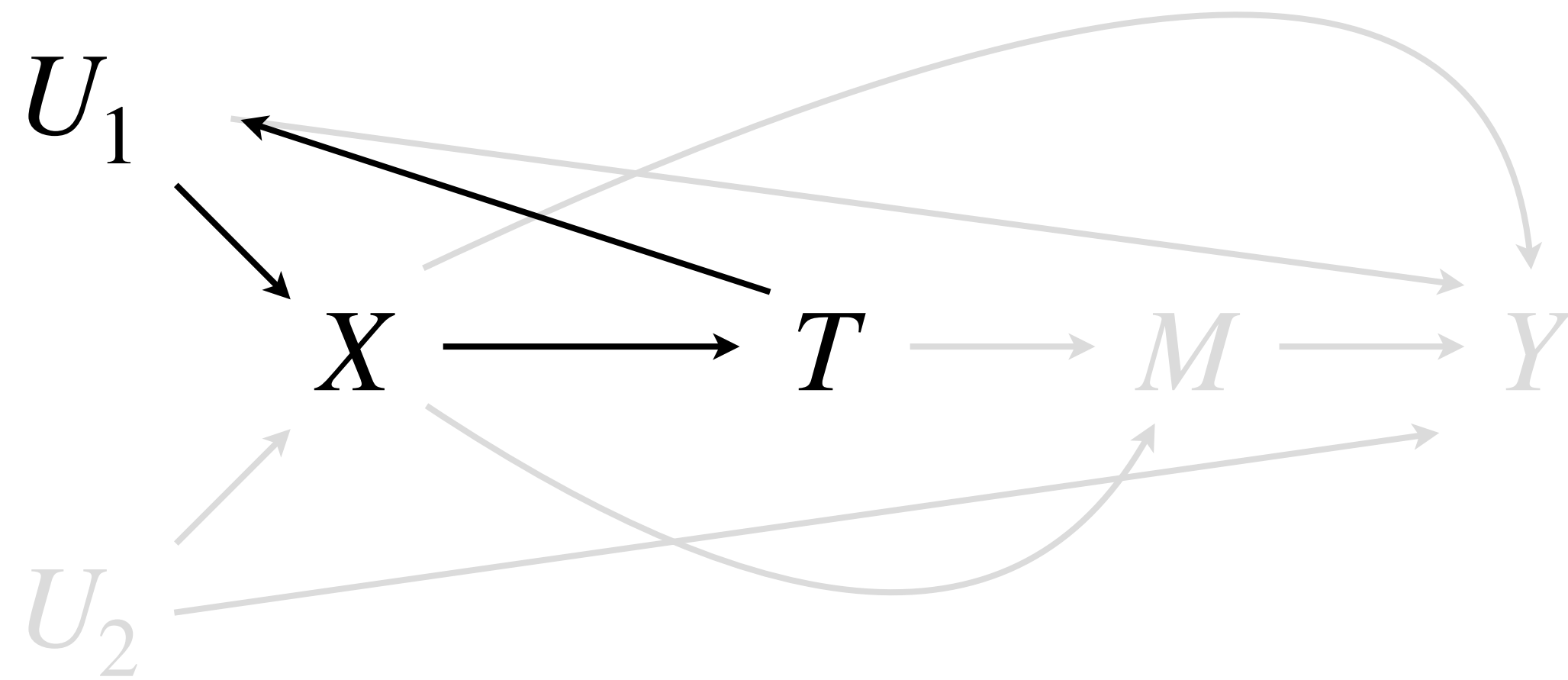




This is a **Directed Acyclic Graph**







There are no **cycles**! This is not a DAG!



X, U_1, Y is also a path

U_1

X

T

M

Y

U_2

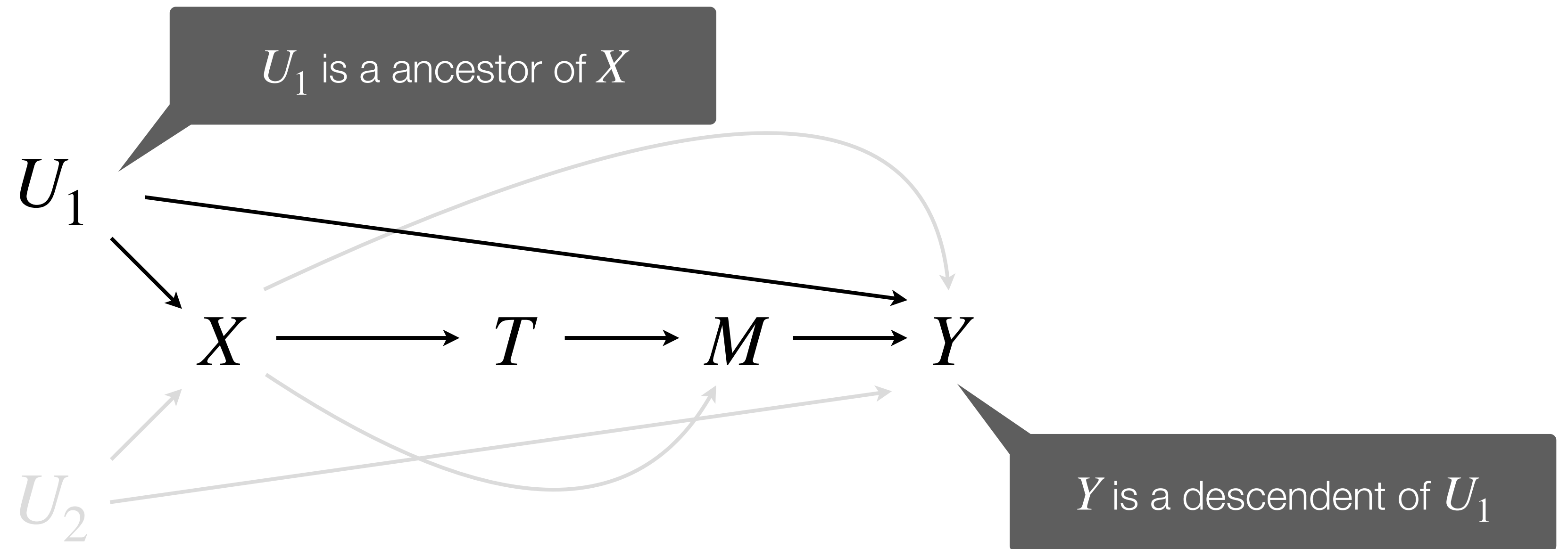
X, T, M, Y is a path

A **path** is any sequence of connected edges ignoring arrow direction





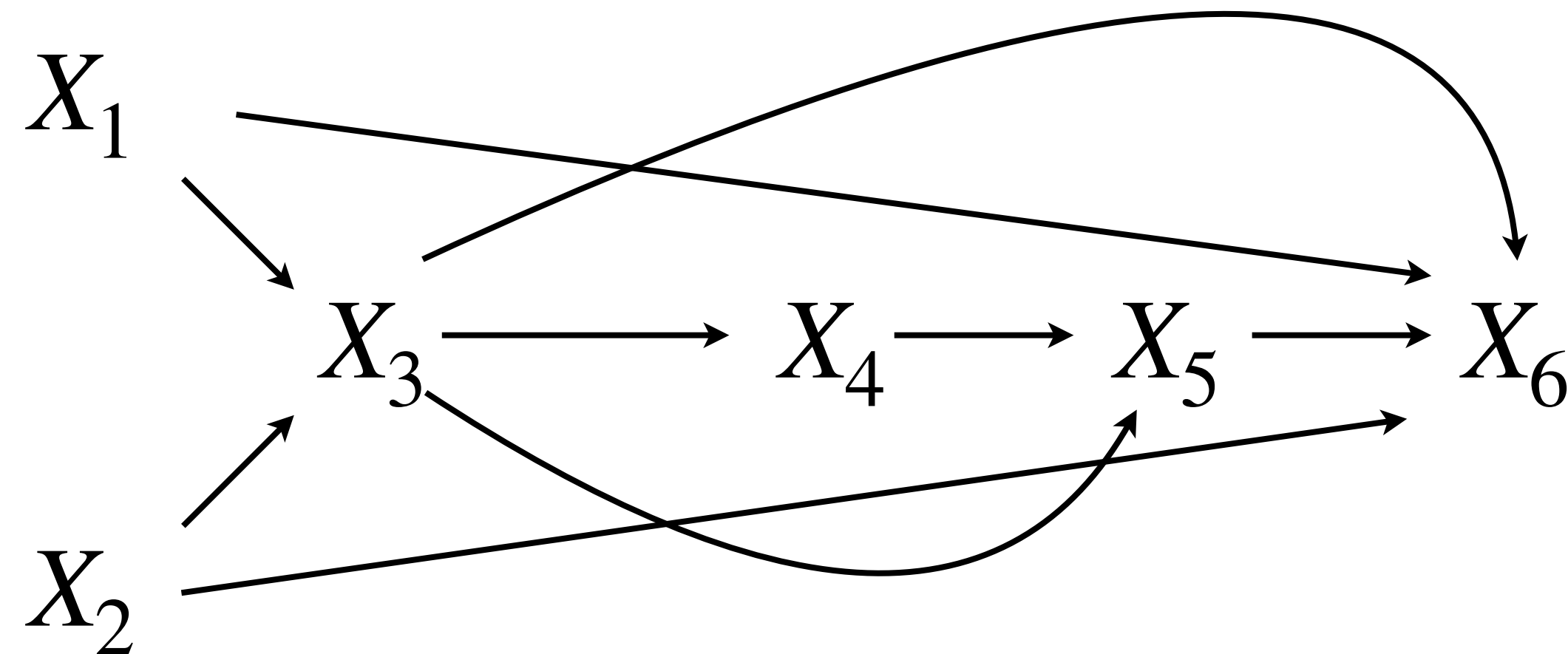
Nodes can be **descendants**
or **ancestors** of other nodes



Modeling the joint distribution

Chain Rule:

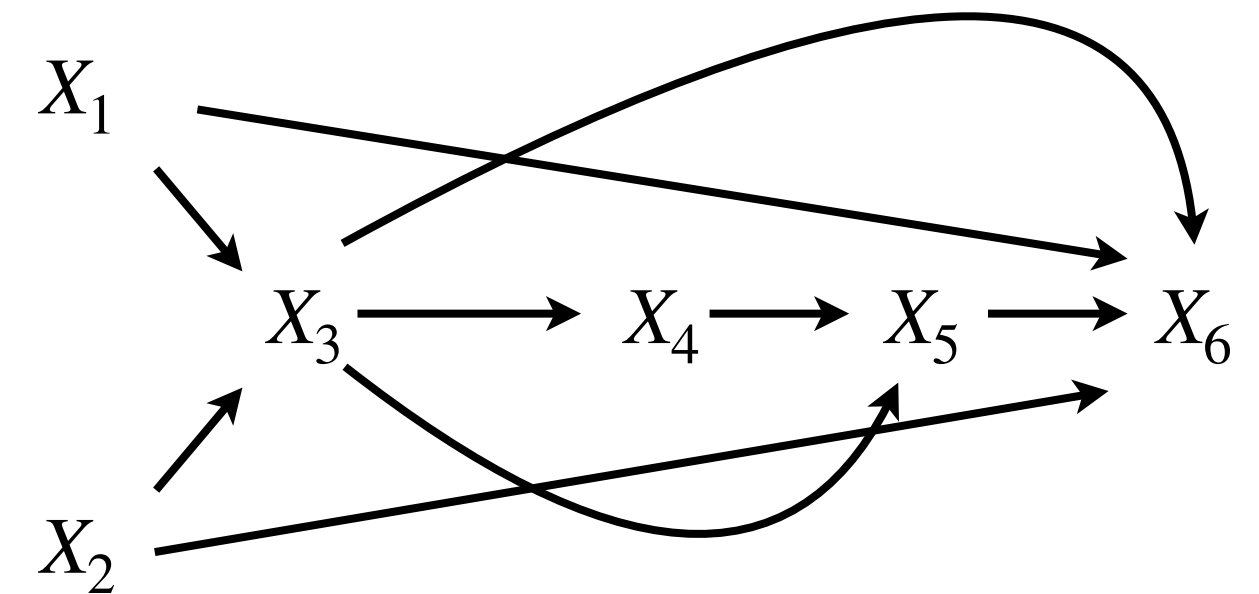
$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2 | X_1)P(X_3 | X_2, X_1)P(X_4 | X_3, X_2, X_1) \dots$$



Local Markov Assumption

A node X is independent of all of its non-descendants given its parents.

$$P(X_1, \dots, X_n) = \prod_i P(X_i \mid \text{Parents of } X_i)$$



$$P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2, X_1)P(X_4 \mid X_3, X_2, X_1)P(X_5 \mid X_4, X_3, X_2, X_1)P(X_6 \mid X_5, X_4, X_3, X_2, X_1)$$

$$P(X_1)P(X_2)P(X_3 \mid X_2, X_1)P(X_4 \mid X_3, X_2, X_1)P(X_5 \mid X_4, X_3, X_2, X_1)P(X_6 \mid X_5, X_4, X_3, X_2, X_1)$$

$$P(X_1)P(X_2)P(X_3 \mid X_2, X_1)P(X_4 \mid X_3)P(X_5 \mid X_4, X_3, X_2, X_1)P(X_6 \mid X_5, X_4, X_3, X_2, X_1)$$

$$P(X_1)P(X_2)P(X_3 \mid X_2, X_1)P(X_4 \mid X_3)P(X_5 \mid X_4, X_3)P(X_6 \mid X_5, X_4, X_3, X_2, X_1)$$

$$P(X_1)P(X_2)P(X_3 \mid X_2, X_1)P(X_4 \mid X_3)P(X_5 \mid X_4, X_3)P(X_6 \mid X_5, X_2, X_1)$$

DAG Building Blocks

Two nodes:

Unconnected nodes

X_1 X_2

$$P(X_1, X_2) = P(X_1)P(X_2)$$

Connected nodes

$X_1 \longrightarrow X_2$

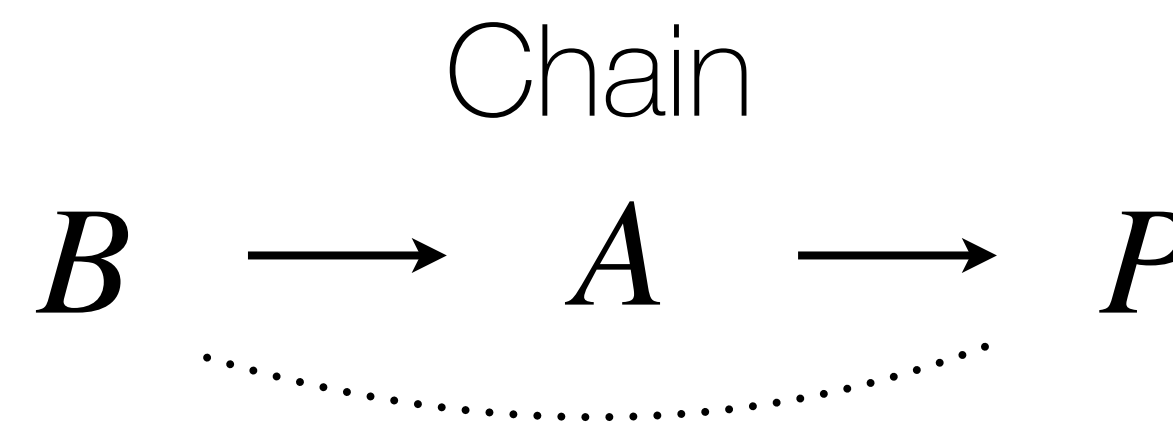


$$P(X_1, X_2) = P(X_2 | X_1)P(X_1)$$

Statistical dependence

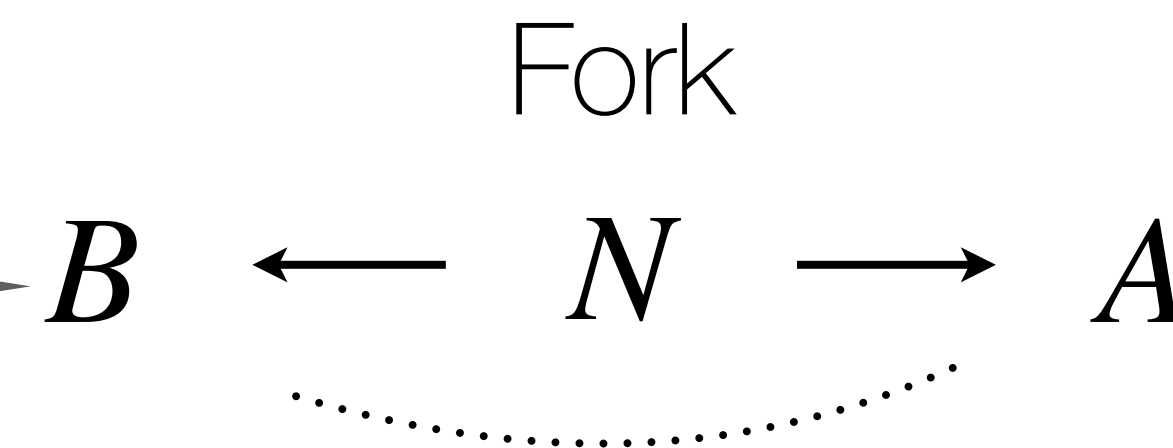
DAG Building Blocks

Three nodes:

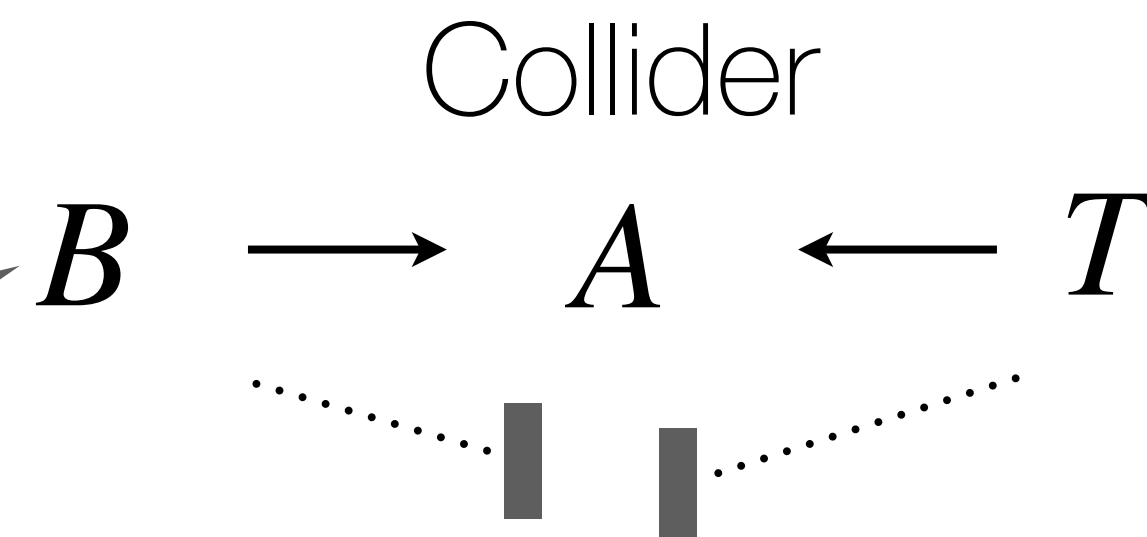


B : Car gets broken into
 A : Car alarm goes off
 P : Police show up

B : Car gets broken into
 N : Parking location
 A : Car alarm goes off



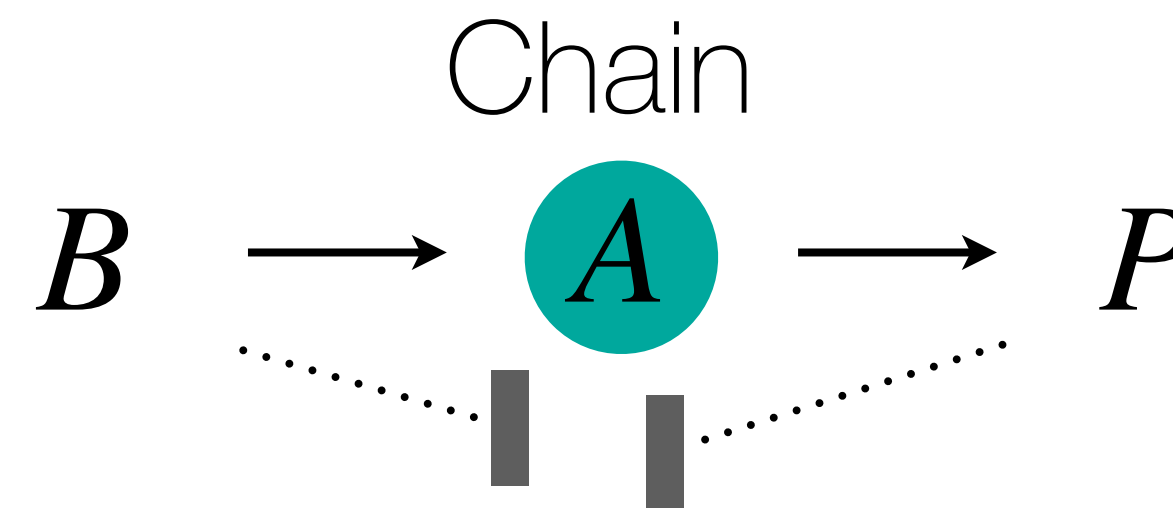
$$P(B, T) = P(B)P(T)$$



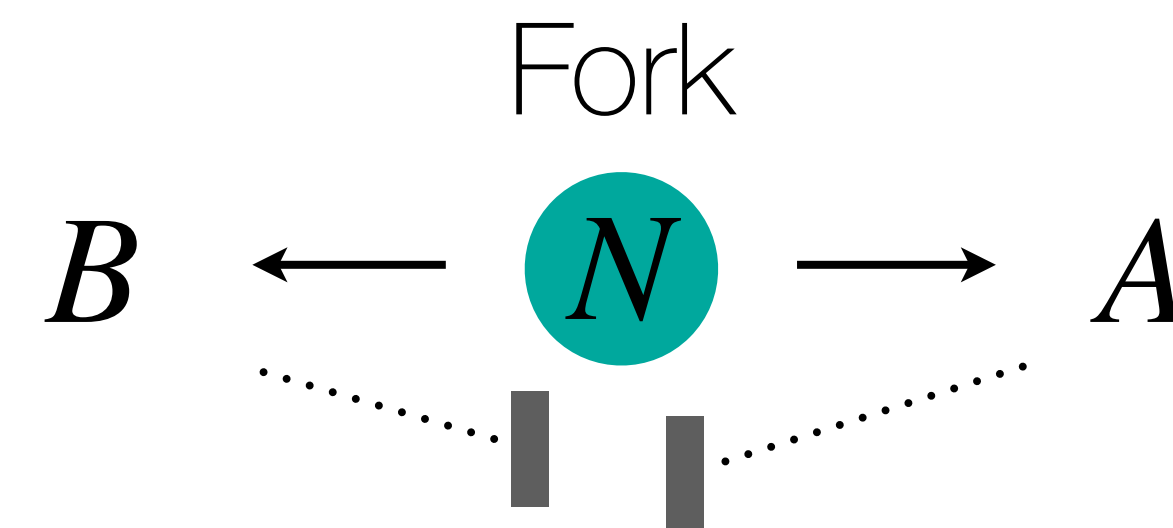
B : Car gets broken into
 A : Car alarm goes off
 T : Thunderstorm

Conditioning

Three nodes:

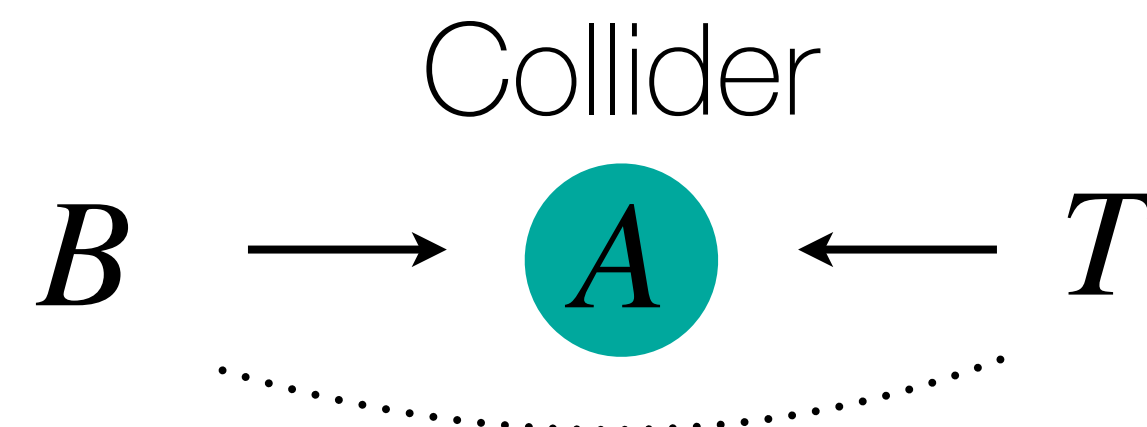


$$\begin{aligned} P(P, B | A) &= P(B | A, P)P(P | A) \\ &= P(B | A)P(P | A) \end{aligned}$$

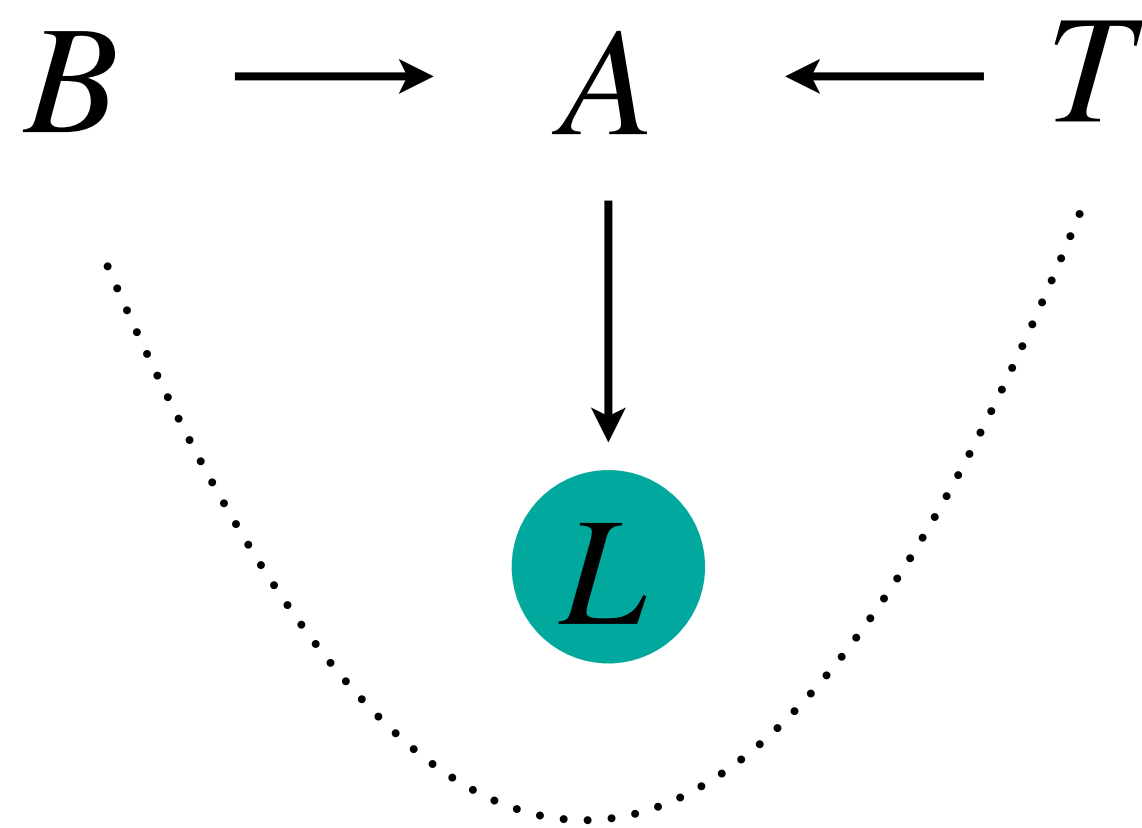


$$\begin{aligned} P(A, B | N) &= P(B | N)P(A | B, N) \\ &= P(B | N)P(A | N) \end{aligned}$$

Given that the alarm goes off (A)...
... if there is a thunderstorm (B)...
... then, the car theft (T) decreases!



Conditioning on descendants



If I know my wife messaged me (M)...
... as she often does when the alarm goes off (A)...
... if there is a thunderstorm (B)...
... then, the car theft (T) decreases!

That's wild!



With these building blocks, we
can analyze the flow of
association in any DAG!

Some more definitions

A path between nodes X Y is **blocked** by a potentially empty conditioning set $\mathbf{Z}$ if along the path there is...

- A chain $\dots \rightarrow W \rightarrow \dots$ where $W \in \mathbf{Z}$
- A fork $\dots \leftarrow W \rightarrow \dots$ where $W \in \mathbf{Z}$
- A collider $\dots \rightarrow W \leftarrow \dots$ where $W \notin \mathbf{Z}$ and none of the descendants are in $\mathbf{Z}$

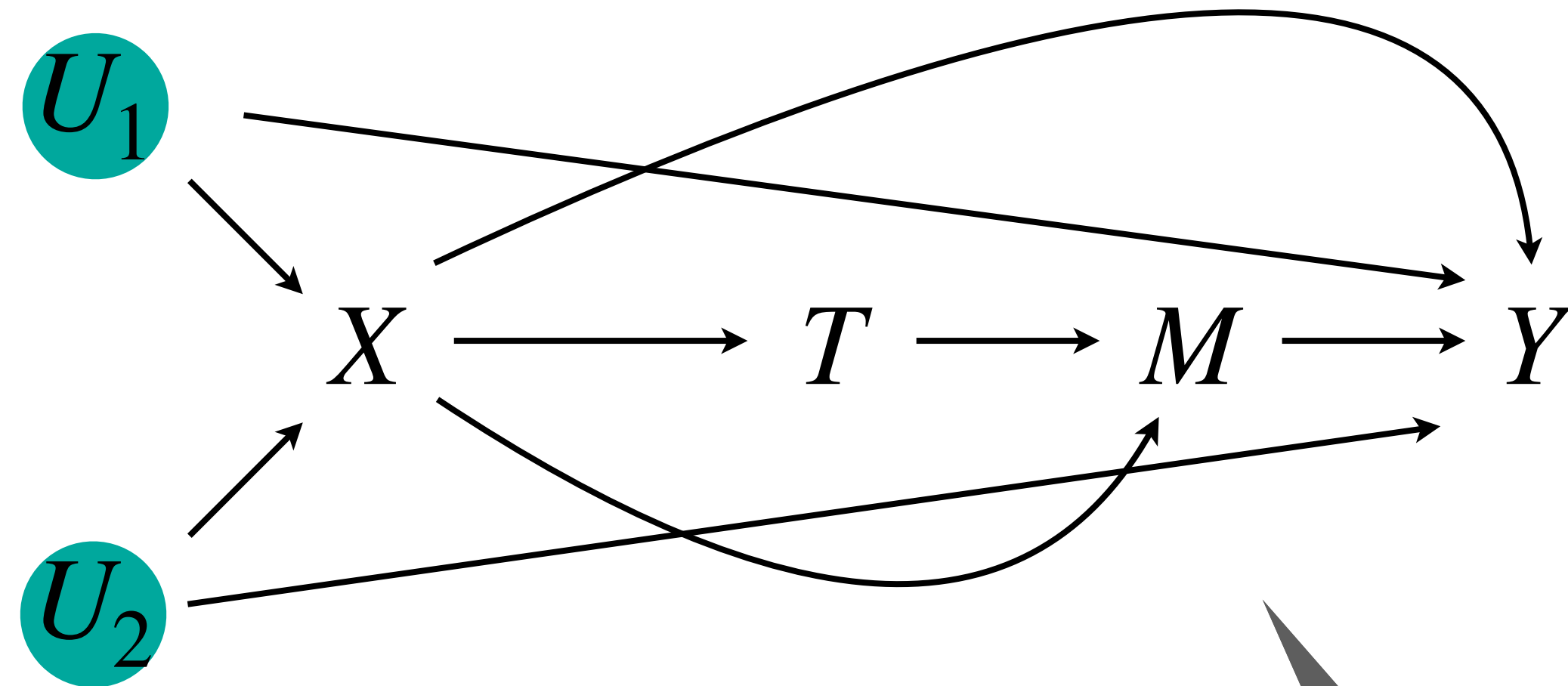
Association is not causation!

Backdoor criterion

A graph-based criterion that allows researchers to identify a set of variables $\mathbf{Z}$ that, when controlled for, allow them to estimate the causal effect of a variable X on another variable Y

1. No node in $\mathbf{Z}$ must be a descendant of X .
2. $\mathbf{Z}$ blocks every path from X to Y that starts with an arrow into X .

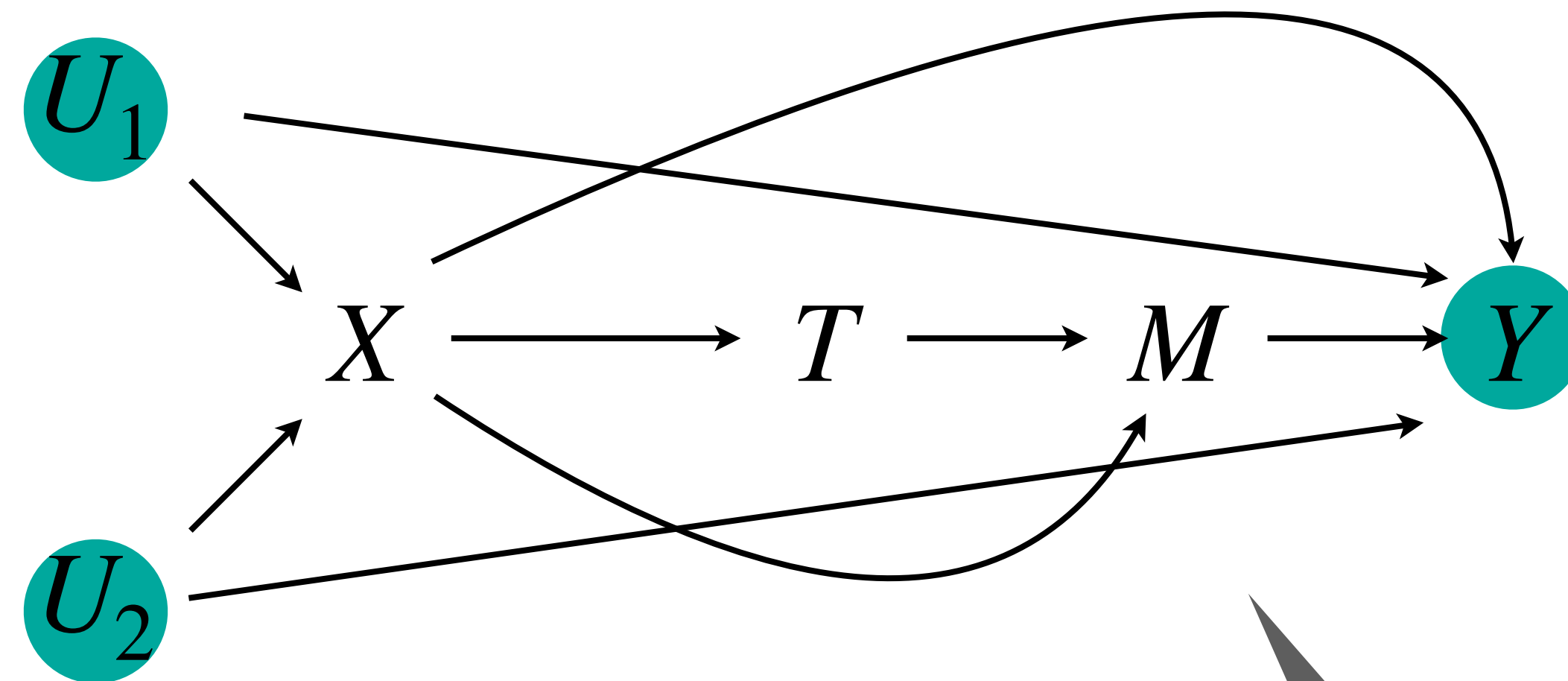
Backdoor criterion exercises



Yes!

$$X, Y, Z = \{U_1, U_2\}$$

Backdoor criterion exercises



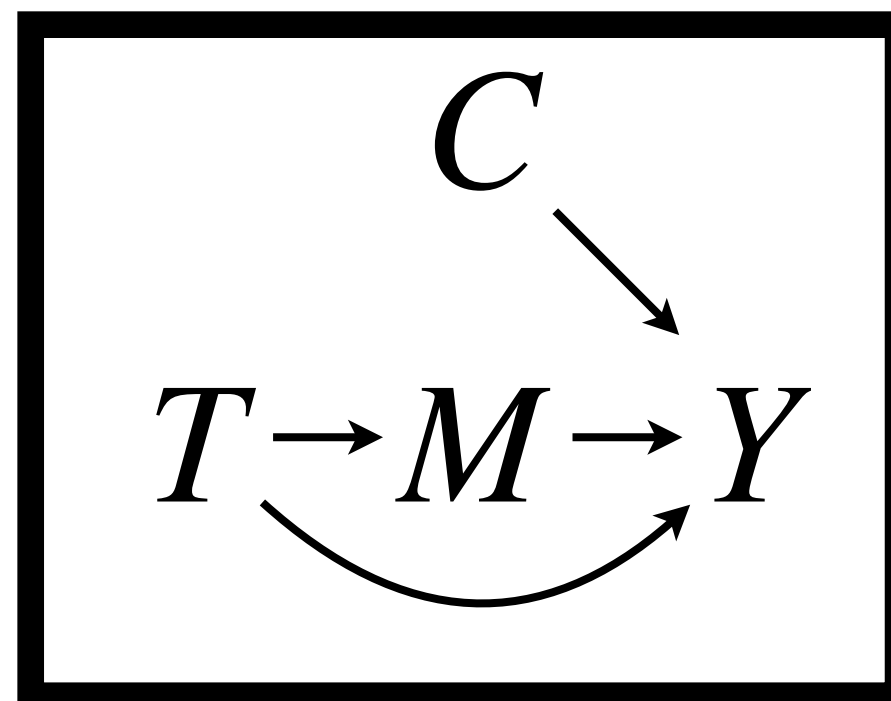
Wrong! Y is a descendant of X

$$X, M, Z = \{U_1, U_2, Y\}$$

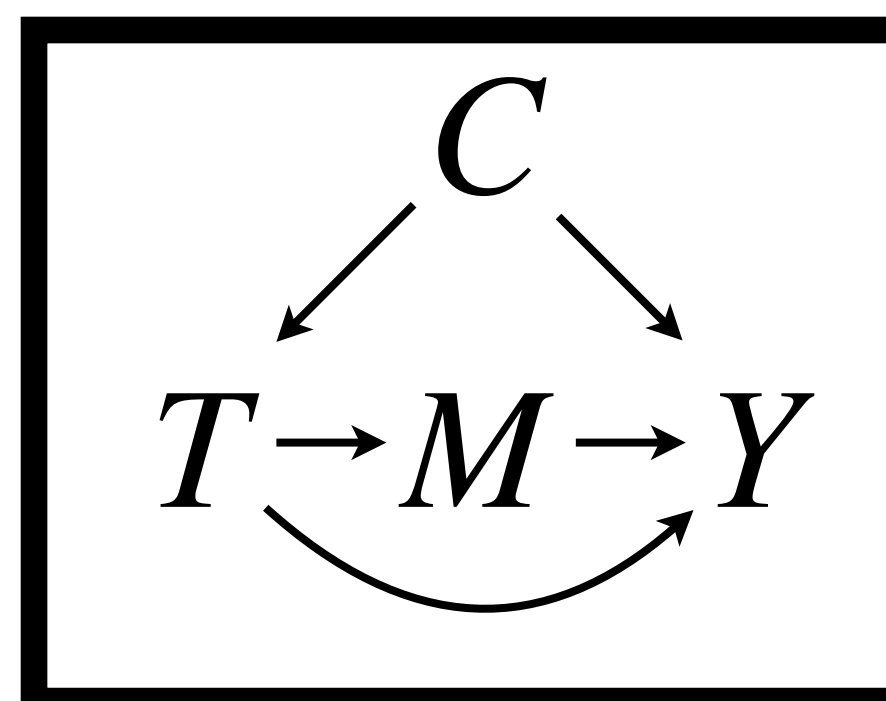
Back to CoT

- Y : whether the model solves the problem
- C : the difficulty of the question.
- T : indicate the prompt (CoT/standard)
- M : whether the model used intermediate tokens

Experiment



Observational



$T \rightarrow Y$ direct causal effect

$T \rightarrow M \rightarrow Y$ indirect causal effect

How do I use DAGs?

- We spent so much time on the minutia of DAGs... But what are DAGs really useful for?
- **Opinionated take:** To help you reason about identification in the real world!
- Let's discuss how to sharpen your causal reasoning with DAGs!

What causal effect are you interested in?

- We start with a causal question:
 - Does smoking cause cancer?
 - Does CoT prompting improve the correctness of LLMs' responses in math questions?
- Think about whether you care about direct effects? Or are mediators fine?
 - Does smoking cause tar in the lungs, which causes cancer?
 - Does CoT cause the addition of reasoning tokens that improve correctness of LLMs' responses in math questions?

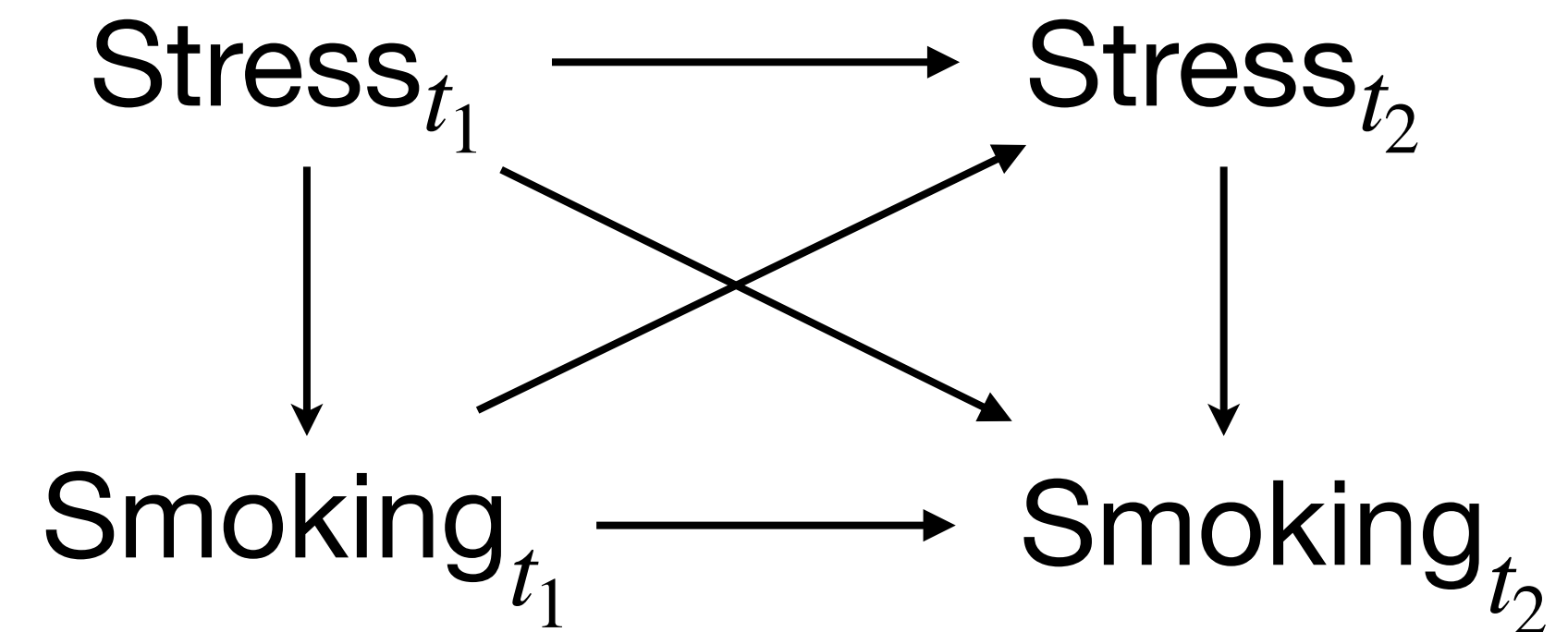
Think about causes of Y and X

- What causes the “outcome” and the “treatment”
 - Will CoT prompts be used in harder questions?
 - Will harder questions be more likely to
- Start adding arrows with potential mechanisms you identify... If some things are unclear, draw undirected arrows denoting uncertainty.
- If two variables are obviously related but you don't want to commit to the mechanism, add “unknown” variables U .
 - There are definitely a bunch of lifestyle and cultural factors that cause both smoking and cancer types (e.g., drinking)

Make time kill cycles!

- Sometimes you will be likely to draw cycles due to time!
Split them up!

Stress $\rightleftarrows$ Smoking



Remember: it is a model!

- Your DAG is a *model*, it should capture what matters most for your specific problem!
- Include variables that seem relevant!
- **Ask:** What assumptions does this DAG imply? What would falsify them?
- **Ask:** Can I estimate the causal effect?
- Draw many DAGs, iterate through this process!



Give us some feedback!



[https://tinyurl.com/
reasoningwithdata](https://tinyurl.com/reasoningwithdata)