Directed Acyclic Graphs

Manoel Horta Ribeiro
manoel@cs.princeton.edu

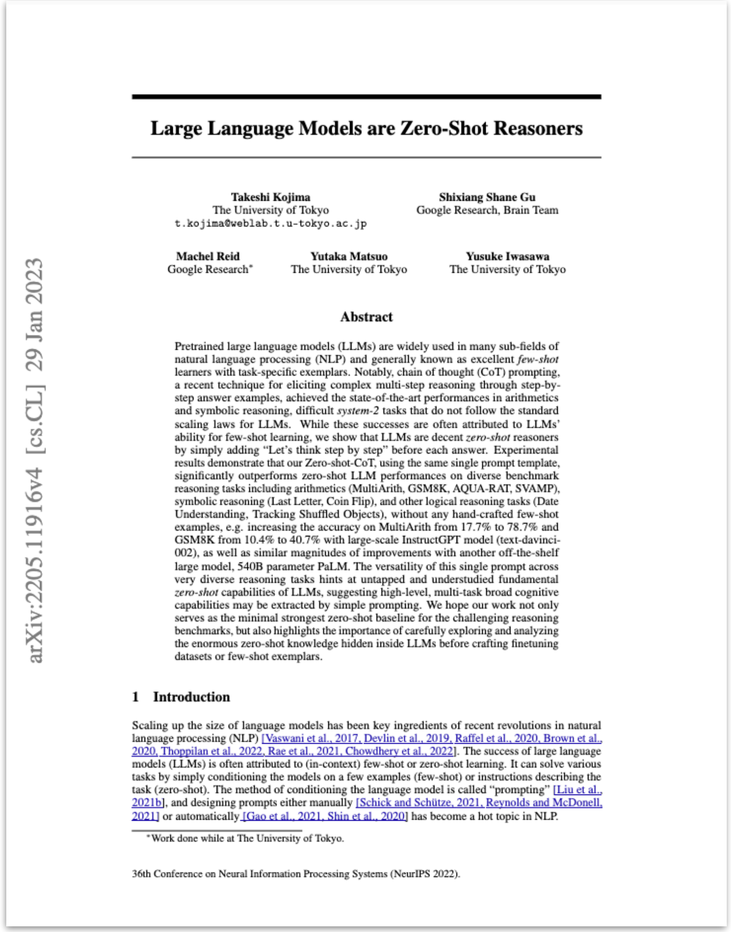
humans & machines lab

COS 598D / Spring 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.**



arXiv:2205.11916v4 [cs.CL] 29 Jan 2023

Large Language Models are Zero-Shot Reasoners

Takeshi Kojima
The University of Tokyo
t.kojima@weblab.t.u-tokyo.ac.jp

Shixiang Shane Gu
Google Research, Brain Team

Machel Reid
Google Research*

Yutaka Matsuo
The University of Tokyo

Yusuke Iwasawa
The University of Tokyo

Abstract

Pretrained large language models (LLMs) are widely used in many sub-fields of natural language processing (NLP) and generally known as excellent *few-shot* learners with task-specific exemplars. Notably, chain of thought (CoT) prompting, a recent technique for eliciting complex multi-step reasoning through step-by-step answer examples, achieved the state-of-the-art performances in arithmetics and symbolic reasoning, difficult *system-2* tasks that do not follow the standard scaling laws for LLMs. While these successes are often attributed to LLMs' ability for few-shot learning, we show that LLMs are decent *zero-shot* reasoners by simply adding "Let's think step by step" before each answer. Experimental results demonstrate that our Zero-shot-CoT, using the same single prompt template, significantly outperforms zero-shot LLM performances on diverse benchmark reasoning tasks including arithmetics (MultiArith, GSM8K, AQUA-RAT, SVAMP), symbolic reasoning (Last Letter, Coin Flip), and other logical reasoning tasks (Date Understanding, Tracking Shuffled Objects), without any hand-crafted few-shot examples, e.g. increasing the accuracy on MultiArith from 17.7% to 78.7% and GSM8K from 10.4% to 40.7% with large-scale InstructGPT model (text-davinci-002), as well as similar magnitudes of improvements with another off-the-shelf large model, 540B parameter PaLM. The versatility of this single prompt across very diverse reasoning tasks hints at untapped and understudied fundamental *zero-shot* capabilities of LLMs, suggesting high-level, multi-task broad cognitive capabilities may be extracted by simple prompting. We hope our work not only serves as the minimal strongest zero-shot baseline for the challenging reasoning benchmarks, but also highlights the importance of carefully exploring and analyzing the enormous zero-shot knowledge hidden inside LLMs before crafting finetuning datasets or few-shot exemplars.

1 Introduction

Scaling up the size of language models has been key ingredients of recent revolutions in natural language processing (NLP) [Vaswani et al., 2017, Devlin et al., 2019, Raffel et al., 2020, Brown et al., 2020, Thoppilan et al., 2022, Rae et al., 2021, Chowdhery et al., 2022]. The success of large language models (LLMs) is often attributed to (in-context) few-shot or zero-shot learning. It can solve various tasks by simply conditioning the models on a few examples (few-shot) or instructions describing the task (zero-shot). The method of conditioning the language model is called "prompting" [Liu et al., 2021b], and designing prompts either manually [Schick and Schütze, 2021, Reynolds and McDonell, 2021] or automatically [Gao et al., 2021, Shin et al., 2020] has become a hot topic in NLP.

*Work done while at The University of Tokyo.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Recap: Experiments Rock

- We show we can estimate the ATE:

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

- Exchangeability:

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

But at the same time, much can go wrong!

- Low power and multiple testing
  - Do not have to do with the bias of the estimate. do not have to do with the bias of the estimate.
- 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

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

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:

$$Y^a \perp T \text{ for all } | X \quad a \in \{0,1\}$$

  - I.e., within difficulty levels, treatments are if as random!

Average Treatment Effect Revisited

$$
\begin{aligned}
ATE &= E\left[Y_{ir}^1 - Y_{ir}^0\right] \\
&= E\left[Y_{ir}^1\right] - E\left[Y_{ir}^0\right] && (1)\ \text{Linearity of Expectations} \\
&= E_{X\sim P(X)}\Big[\, E\left[Y_{ir}^1 \mid X\right]\Big] \;-\; E_{X\sim P(X)}\Big[\, E\left[Y_{ir}^0 \mid X\right]\Big] && (2)\ \text{Law of Iterated Expectations} \\
&= E_{X\sim P(X)}\Big[\, E\left[Y_{ir}^1 \mid X, T_{ir}=1\right]\Big] \;-\; E_{X\sim P(X)}\Big[\, E\left[Y_{ir}^0 \mid X, T_{ir}=0\right]\Big] && (3)\ \text{Conditional Exchangeability} \\
&= E_{X\sim P(X)}\Big[\, E\left[Y_{ir} \mid X, T_{ir}=1\right]\Big] \;-\; E_{X\sim P(X)}\Big[\, E\left[Y_{ir} \mid X, T_{ir}=0\right]\Big] && (4)\ \text{Consistency}
\end{aligned}
$$

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

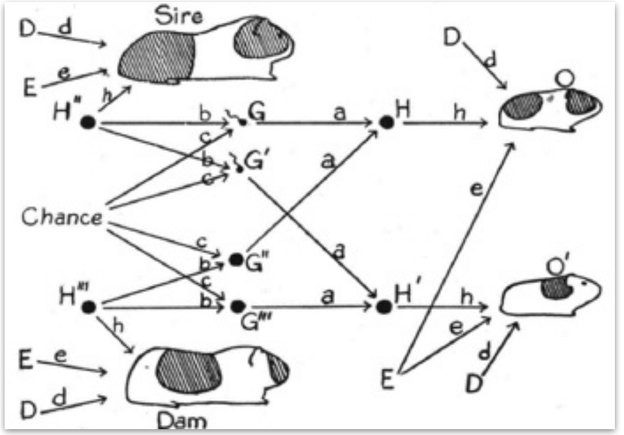
“The world is a complicated place”

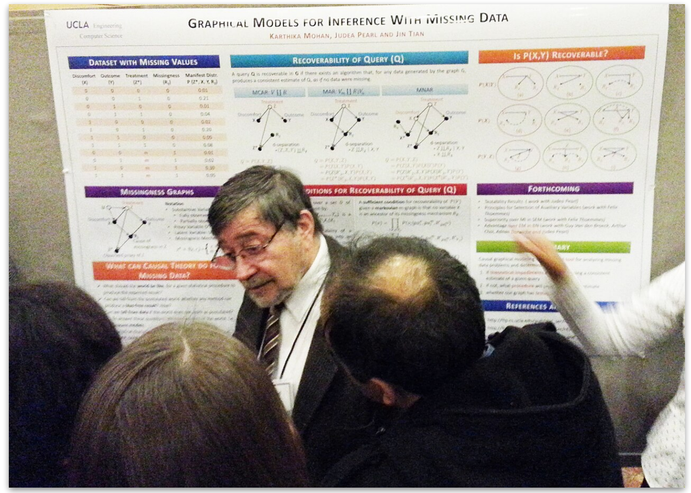
- Naïve idea: controlling for everything!
  - Infants born to smokers were found to have a higher risk of mortality than infants born to non-smokers.
  - However, for infants with low birth weight, this relationship is reversed!

- 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)



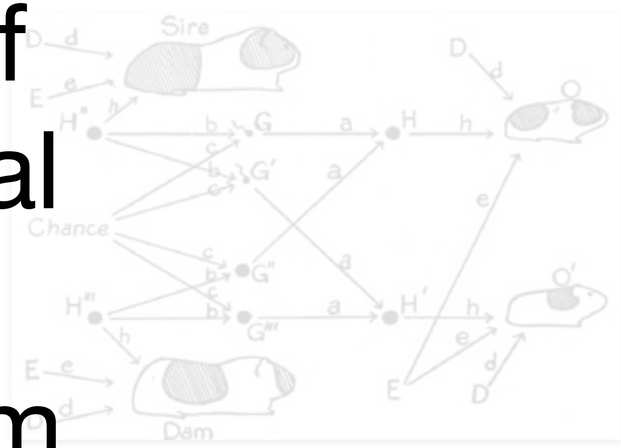


DAGs to the rescue!

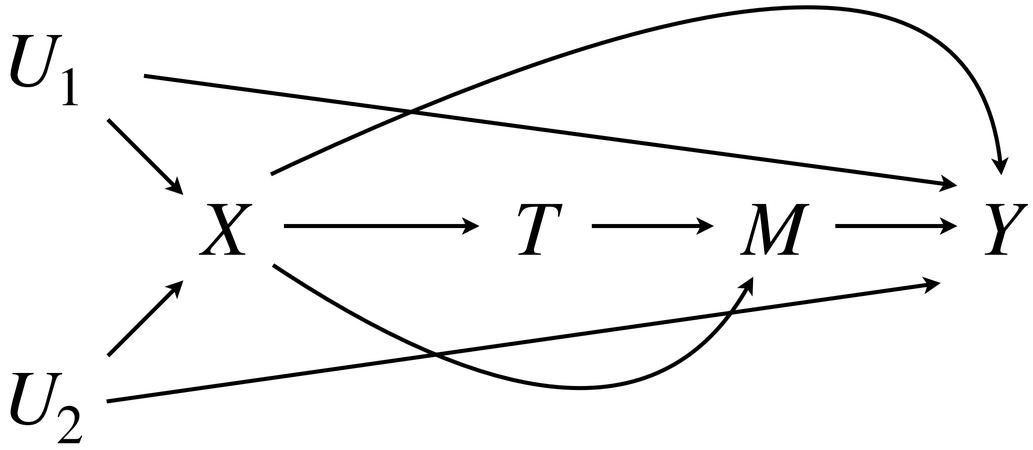
- Directed Acyclic Graphs are a great way to reason about causal effects!
- They date back to at least Sewall Wright (1921)
- Mostly developed alongside Judea Pearl's causal inference framework (structural causal model)

identify

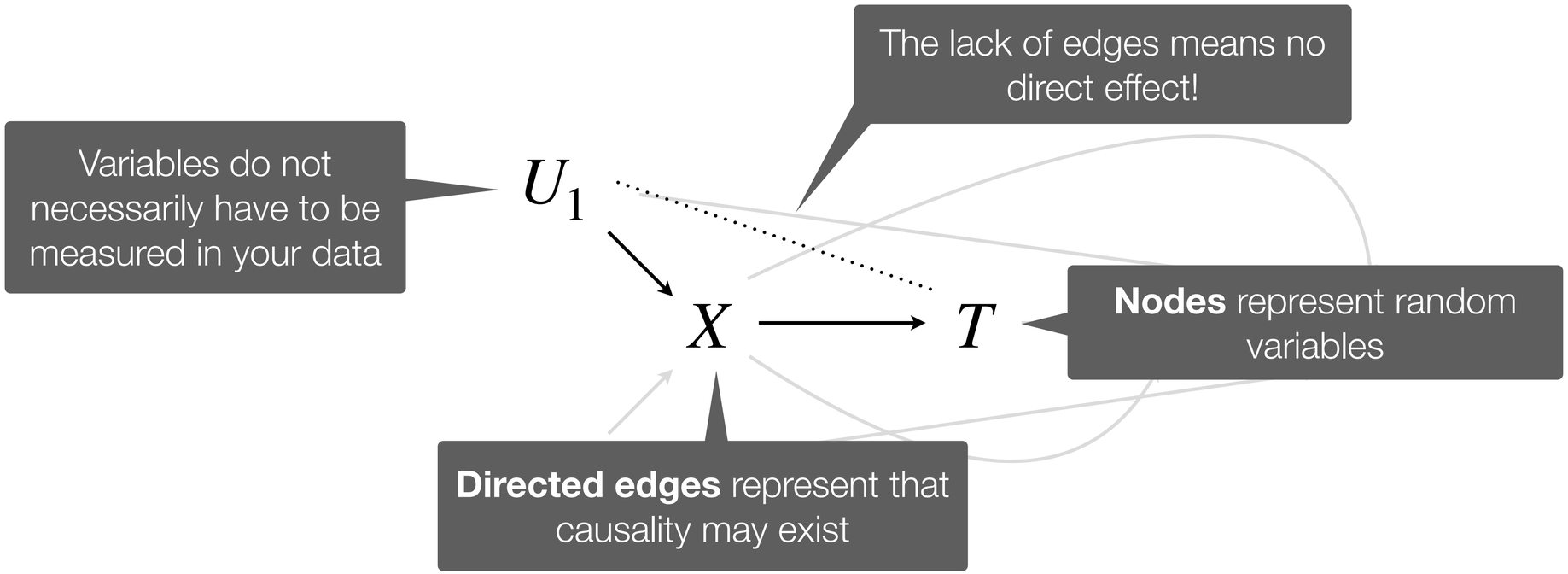
The possibility of separating causal and non-causal associations from data!

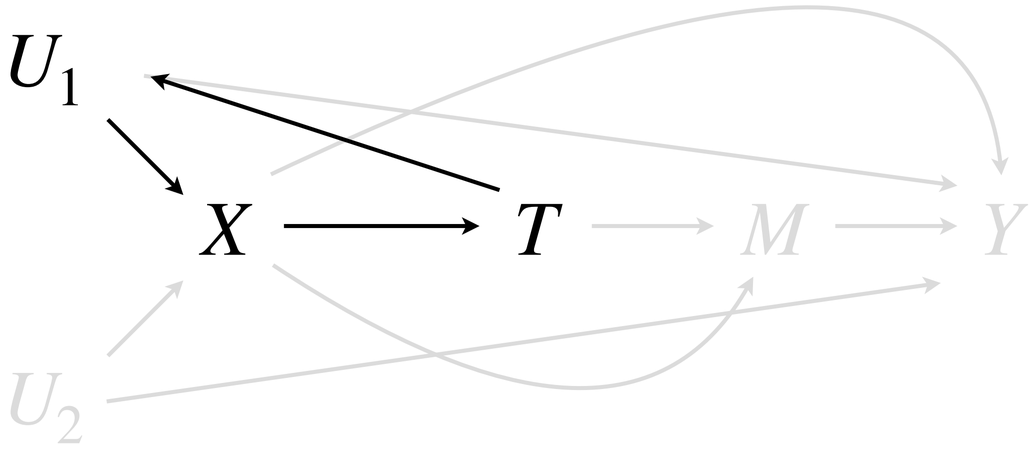


DAG Basics

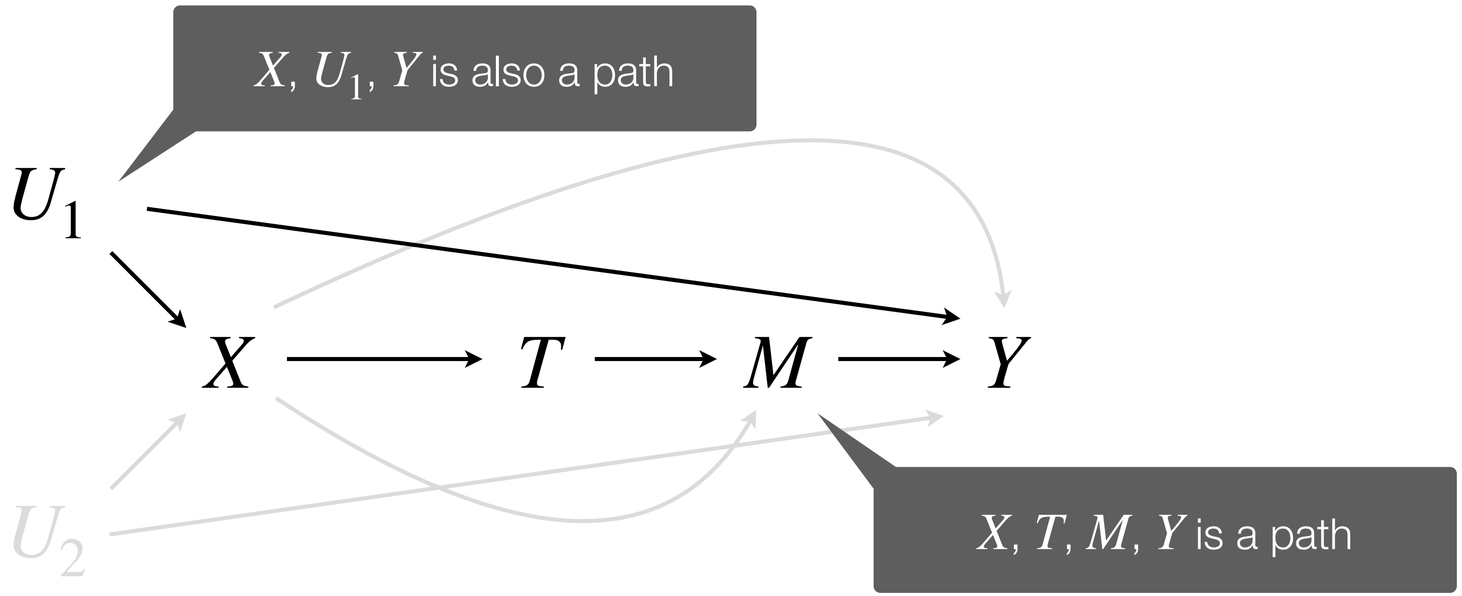


This is a **Directed Acyclic Graph**

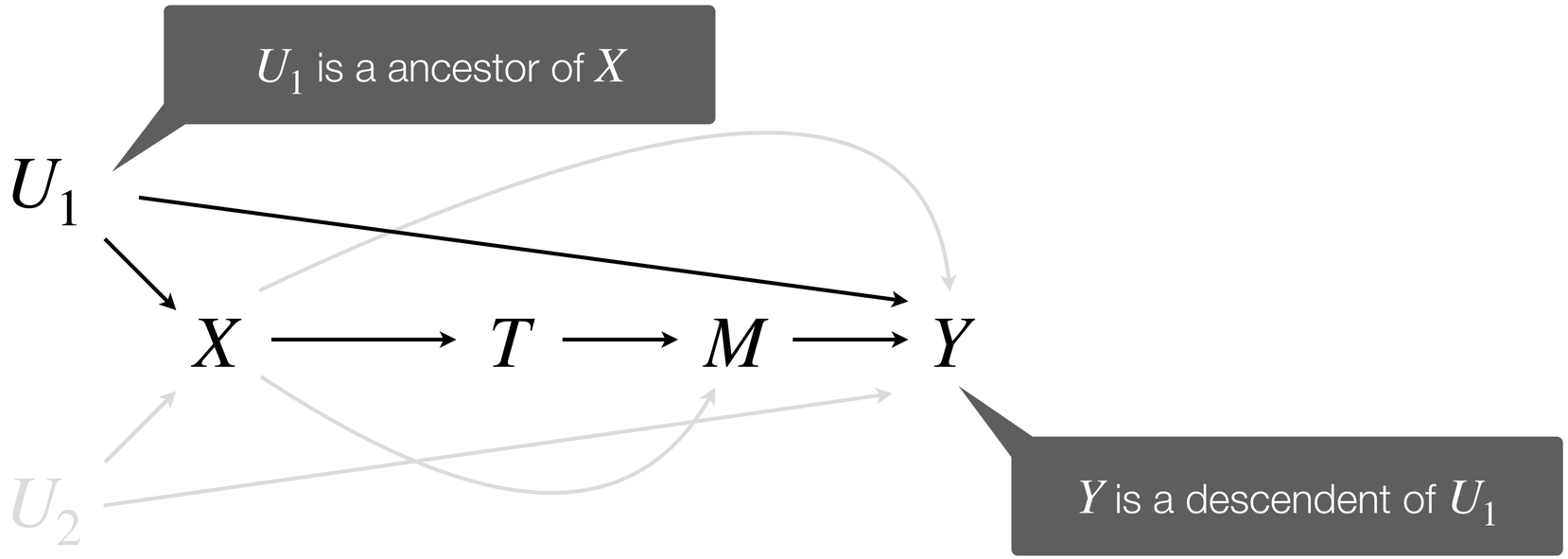

The lack of edges means no direct effect!
Variables do not necessarily have to be measured in your data
U_1
X
T
Nodes represent random variables
Directed edges represent that causality may exist



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



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

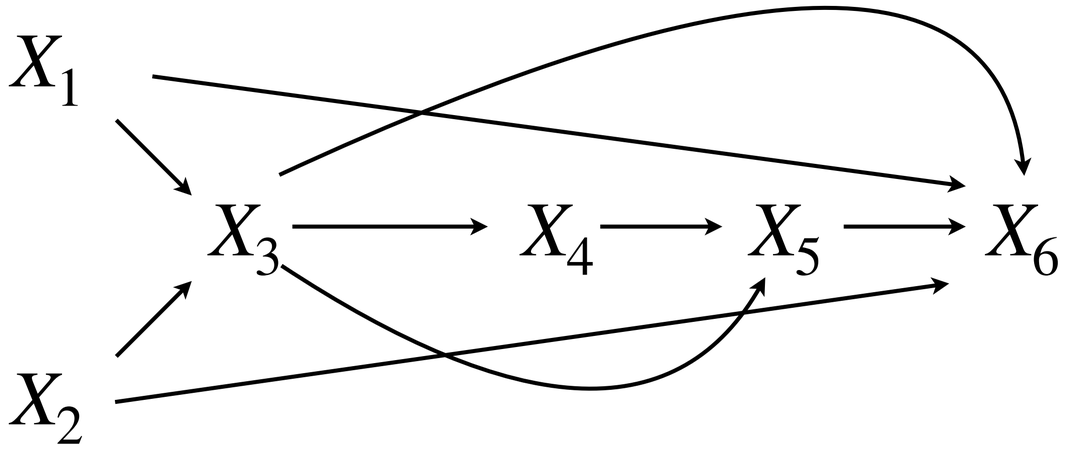


Nodes can be **descendants** or **ancestors** of other nodes

Modeling the joint distribution

Chain Rule:

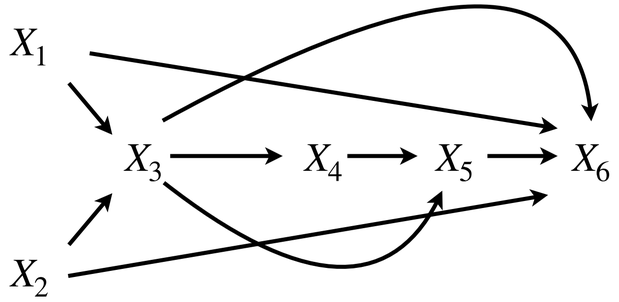
$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2, X_1)P(X_4 \mid X_3, X_2, X_1)\ldots$$



Local Markov Assumption

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

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



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

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

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

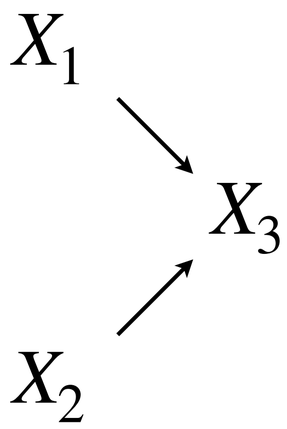
$$P(X_1)P(X_2)P(X_3 \mid X_2, X_1)P(X_4 \mid X_3)P(X_5 \mid X_4, X_3,)P(X_6 \mid X_5, X_4, X_3, X_2, X_1)$$

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

“Good’ol Bayesian Networks”

	$P(X_1 = a)$
$a = 0$	30 %
$a = 1$	70 %

	$P(X_2 = a)$
$a = 0$	60 %
$a = 1$	40 %



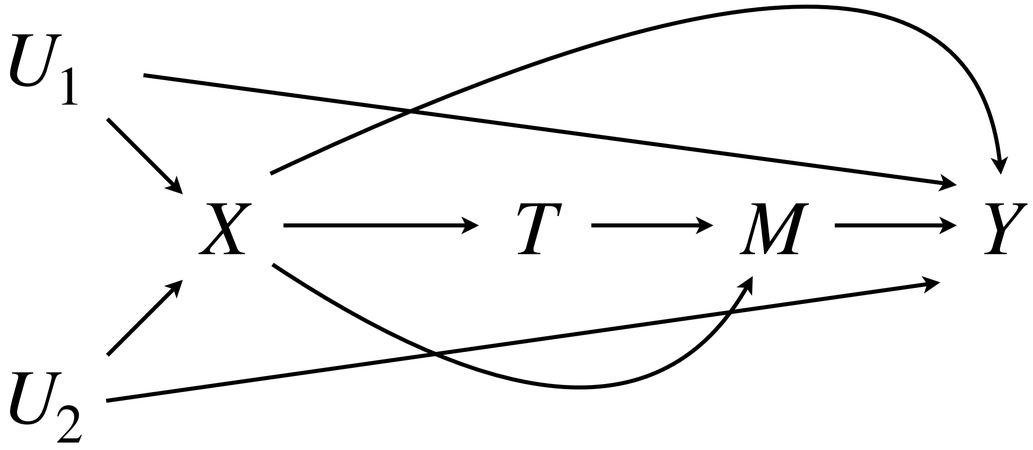
$P(X_3 = a \mid \ldots)$	$a = 0$	$a = 1$
$X_1 = 0, \quad X_2 = 0$	15 %	85 %
$X_1 = 0, \quad X_2 = 1$	25 %	75 %
$X_1 = 1, \quad X_2 = 0$	20 %	80 %
$X_1 = 1, \quad X_2 = 1$	35 %	65 %

The table grows exponentially with the number of parents! Can be very tricky to estimate with data!

Causal Edges Assumption

Every parent is a direct **cause** of all their children!

X "causes" Y if Y can change in response to changes in X, all else held equal



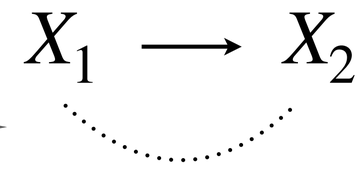
DAG Building Blocks

Two nodes:

Unconnected nodes

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

Connected nodes

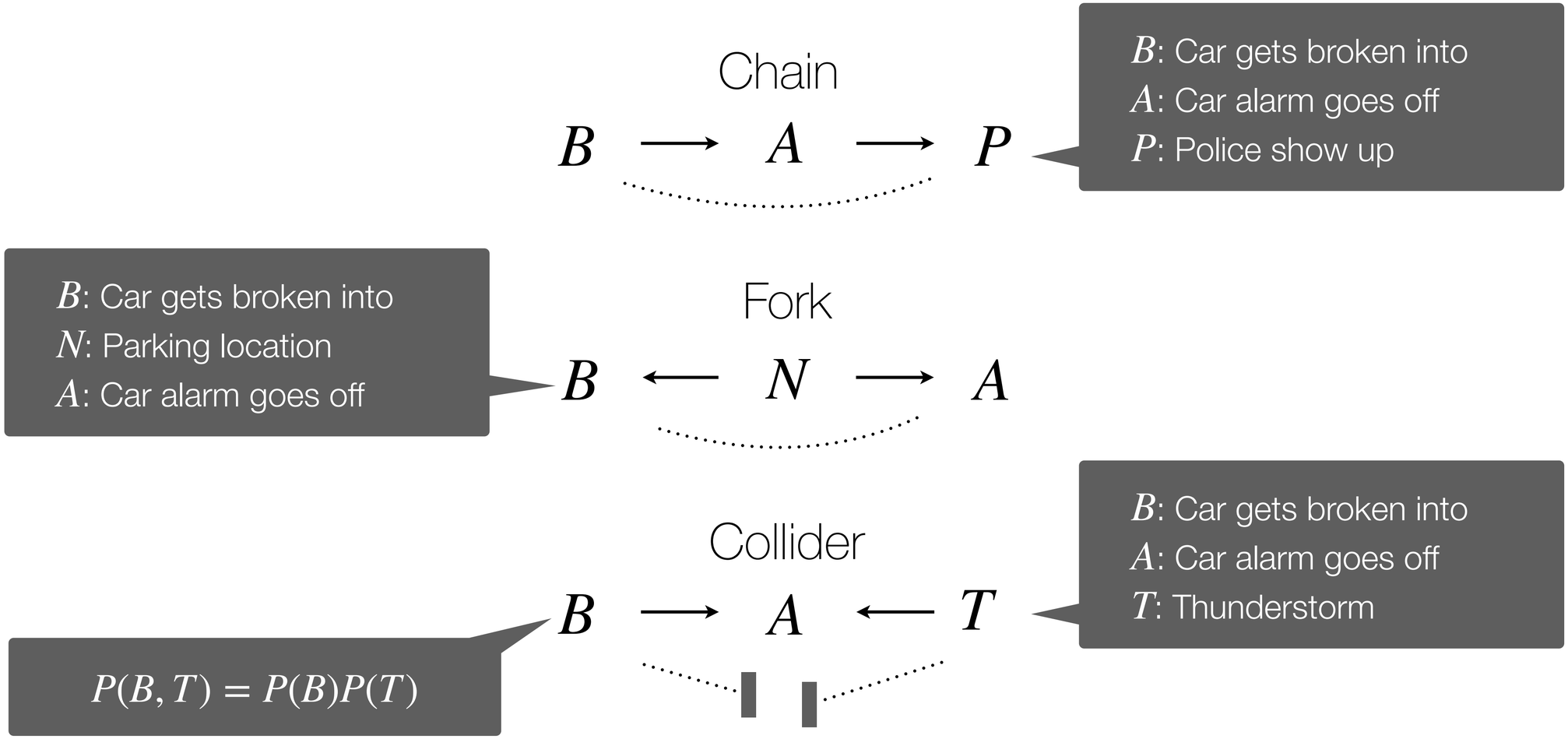


$$P(X_1, X_2) = P(X_2 \mid X_1)P(X1)$$

Statistical dependence

DAG Building Blocks

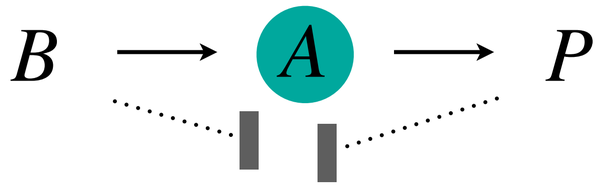
Three nodes:



Conditioning

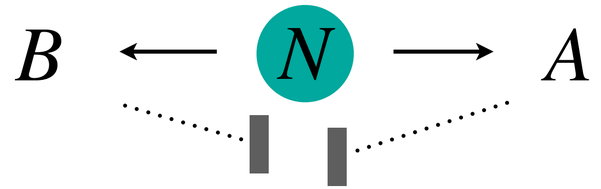
Three nodes:

Chain



$$P(P, B \mid A) = P(B \mid A, P)P(P \mid A)$$
$$= P(B \mid A)P(P \mid A)$$

Fork

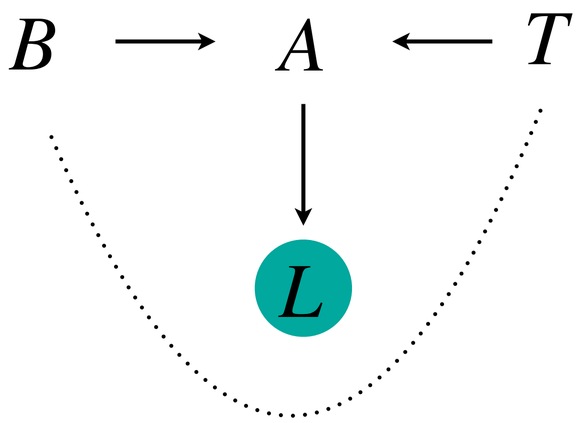


$$P(A, B \mid N) = P(B \mid N)P(A \mid B, N)$$
$$= P(B \mid N)P(A \mid N)$$

Collider

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…

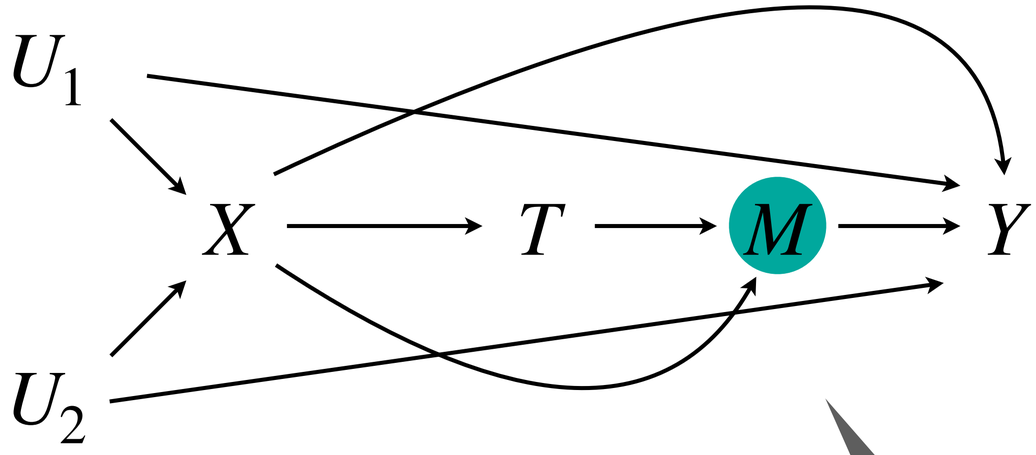
- A chain $\ldots \rightarrow W \rightarrow \ldots$ where $W \in \mathbf{Z}$
- A fork $\ldots \leftarrow W \rightarrow \ldots$ where $W \in \mathbf{Z}$
- A collider $\ldots \rightarrow W \leftarrow \ldots$ where $W \notin \mathbf{Z}$ and none of the descendants are in $\mathbf{Z}$

d-separation

Two sets of nodes $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by a set of nodes $\mathbf{Z}$ if all paths between any node in $\mathbf{X}$ and any node in $\mathbf{Y}$ are blocked.

If two sets of variables $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by a set $\mathbf{Z}$ in $\mathcal{G}$, then in every probability distribution that factorizes according to $\mathcal{G}$, we have $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$

d-separation exercises



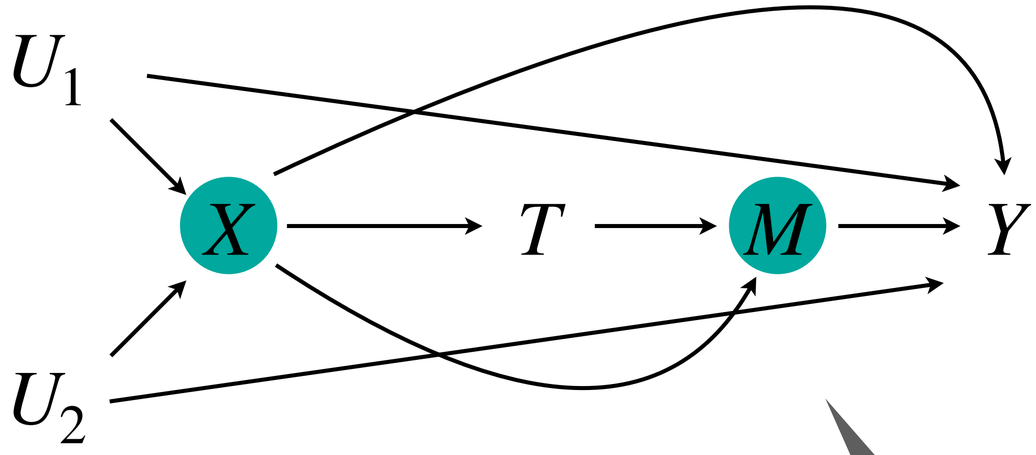
No! $T \leftarrow X \rightarrow Y$

$\mathbf{X} = \{T\}$

$\mathbf{Y} = \{Y\}$

$\mathbf{Z} = \{M\}$

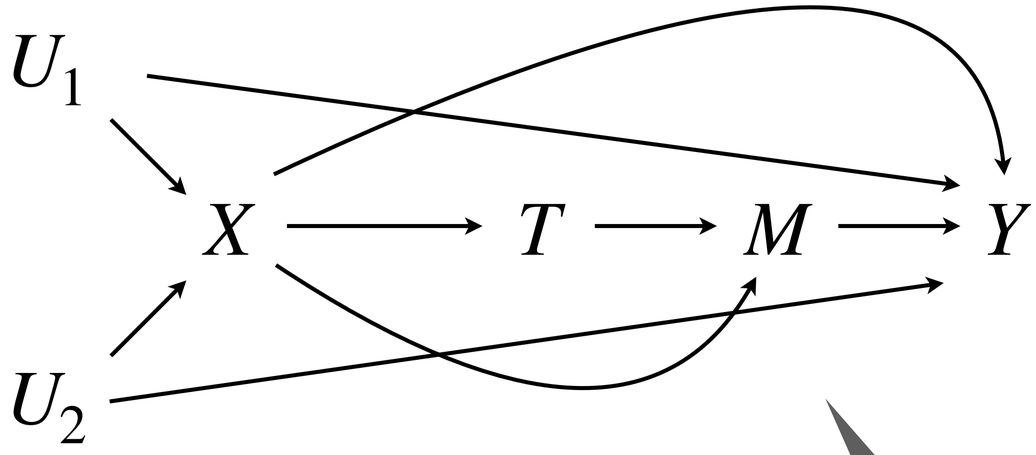
d-separation exercises



Yes!

$$\mathbf{X} = \{T\}$$
$$\mathbf{Y} = \{Y\}$$
$$\mathbf{Z} = \{M, X\}$$

d-separation exercises



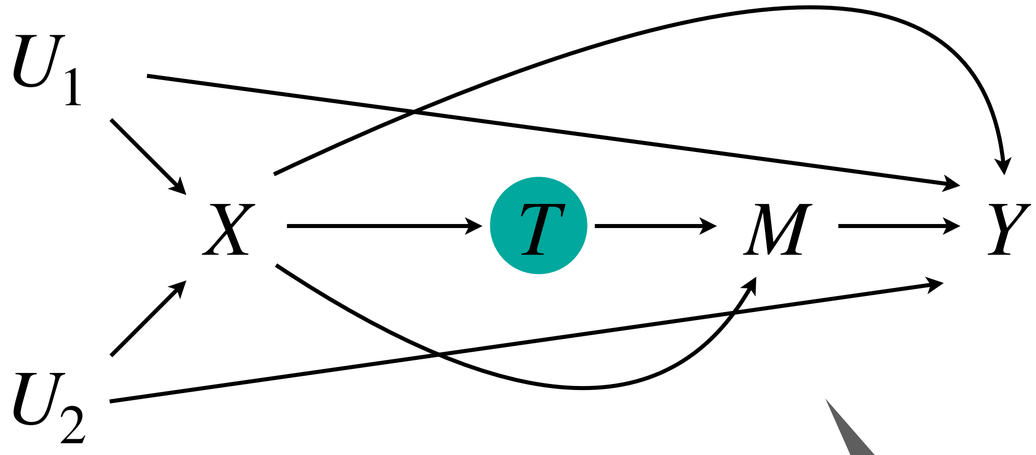
Yes!

$\mathbf{X} = \{U_1\}$

$\mathbf{Y} = \{U_2\}$

$\mathbf{Z} = \{\}$

d-separation exercises



No! $U_1 \leftarrow X \rightarrow U_2$

$\mathbf{X} = \{U_1\}$

$\mathbf{Y} = \{U_2\}$

$\mathbf{Z} = \{T\}$

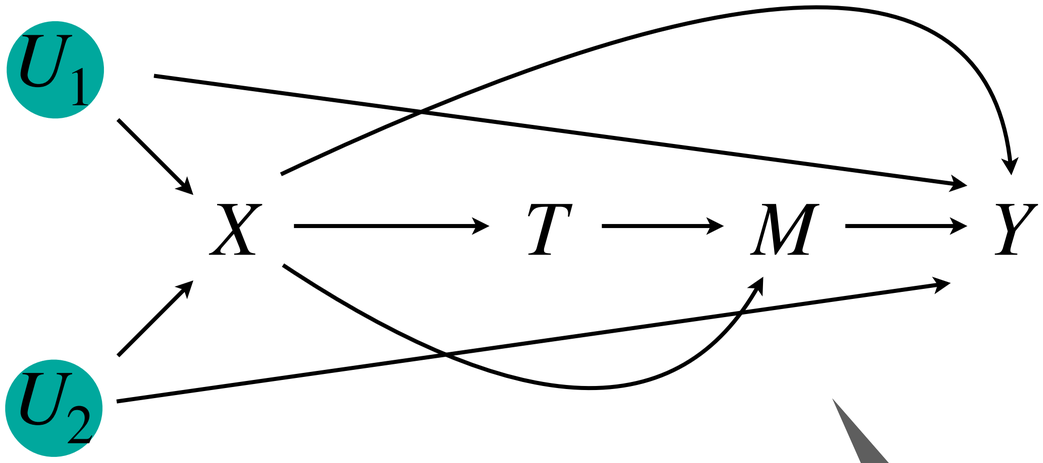
Association is not causation!

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.

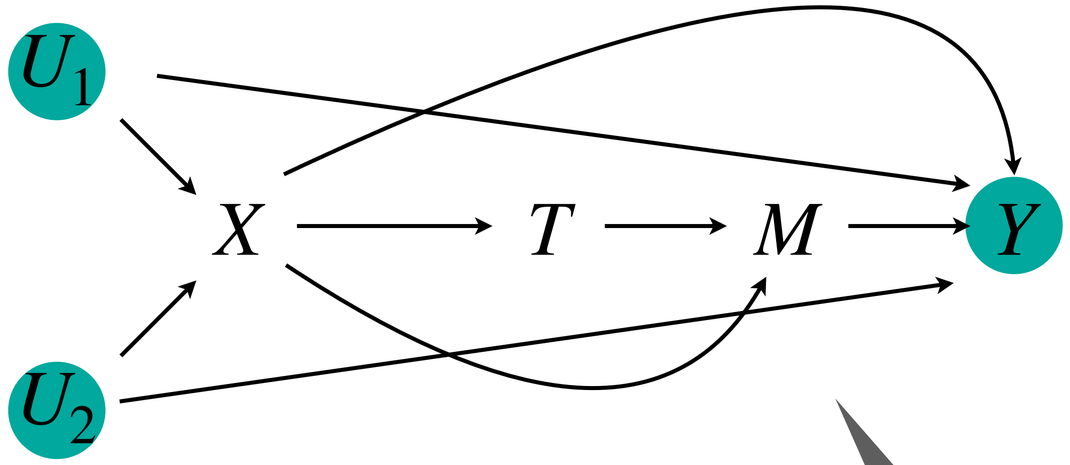
Backdoor criterion exercises



Yes!

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

Backdoor criterion exercises



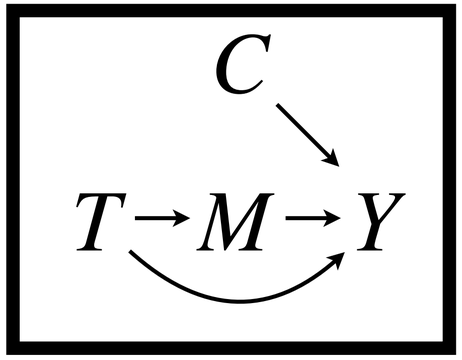
$X, M, \mathbf{Z} = \{U_1, U_2, Y\}$

Wrong! Y is a descendant of X

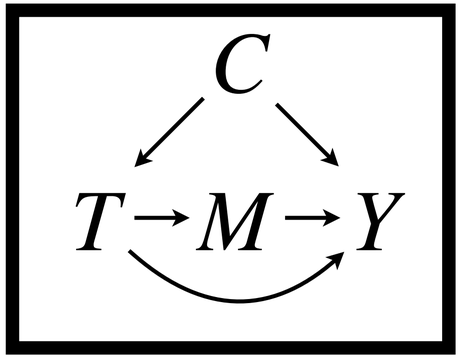
Back to CoT

- 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?

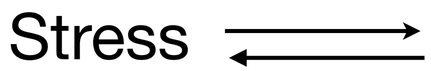
- We start with a causal question:
  - Does smoking cause cancer?
  - Does CoT prompting improve the correctness of LLMs' responses in math questions?
- Think about whether you care about direct effects? Or are mediators fine?
  - Does smoking cause tar in the lungs, which causes cancer?
  - Does CoT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

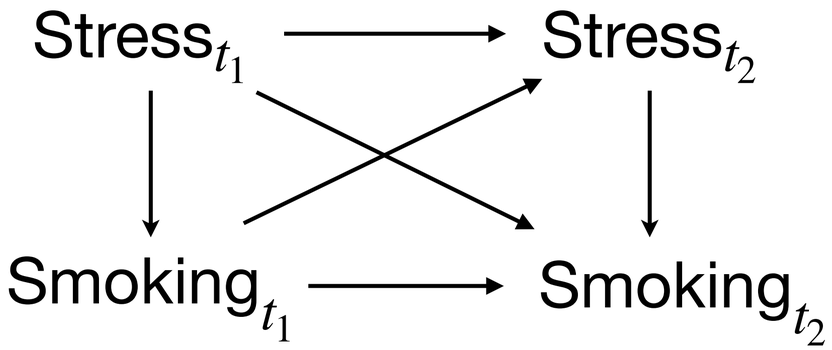
Think about causes of Y and X

- 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!





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!