# Causal machine learning in practice

**Amit Sharma**

Principal Researcher

Microsoft Research

www.amitsharma.in
Twitter: **@amt_shrma**

# Two session series on Causal ML

- **Session 1:** Intro to causal machine learning
  - Estimating causal effect, explaining outcomes, and out-of-distribution generalization

- **Session 2:** Causal machine learning in practice
  - PyWhy/DoWhy and the promise of large language models

# **Session goals:** Causal ML in practice

- Understand **key challenges** of applying causal ML in the real world

- Learn about **open-source tools** for building causal models

- See **causal ML in action** through three real-world case studies
  - Customer rewards program in e-commerce platform
  - Debugging high latency in a microservice system architecture
  - Classifying images under spurious correlation

- **Looking ahead:** Explore how **large language models** may help with some of these challenges
  - How can LLMs help infer the correct causal graph?

# Outline

1. **The four key steps of causal inference**

   1. PyWhy software ecosystem for causal ML

   2. Open-source libraries: DoWhy, EconML, causal-learn, and more...

2. **Case-studies of applying causal ML using PyWhy/DoWhy**

   1. Effect estimation

   2. Root cause attribution

   3. Out-of-distribution prediction

3. **Looking ahead: Promise of large language models for causality**

# Section 1: The 4 Key Steps of Causal Inference

# Two Fundamental Challenges for Causal Inference

Multiple causal mechanisms and estimates can fit the same data distribution.

➡️ **1. Assumptions**

Estimation about different data distributions than the training distribution (no easy "cross-validation").

➡️ **2. Evaluation**

Real World:

Counterfactual World:

# The four key steps of causal inference to highlight the importance of assumptions

1. **Modeling assumptions:** Create a causal graph to encode assumptions.

2. **Identification:** Formulate what to estimate.
   1. E.g., use backdoor criterion.

3. **Estimation:** Compute the estimate.
   1. E.g., use matching or debiased ML.

4. **Refuting assumptions:** Validate the assumptions.

# To implement these 4 steps, we built DoWhy, an open-source library for causal inference

DoWhy makes assumptions front-and-center of any analysis

- Transparent declaration of assumptions

- Evaluation of those assumptions, to the extent possible

Most popular causal library on GitHub (> 2M downloads, 800+ forks)

- Taught in third-party tutorials and courses: O'Reilly, PyData, Northeastern, …

- Open-source community: 60 contributors

- Including major contributions: *Pandas dataframe extension*

Goal: An end-to-end platform for doing causal inference

# DoWhy provides a general API for the four steps of causal inference

1. **Modeling:** Create a causal graph to encode assumptions.

2. **Identification:** Formulate what to estimate.

3. **Estimation:** Compute the estimate.

4. **Refutation:** Validate the assumptions.

We'll discuss the four steps and show a code example using DoWhy.
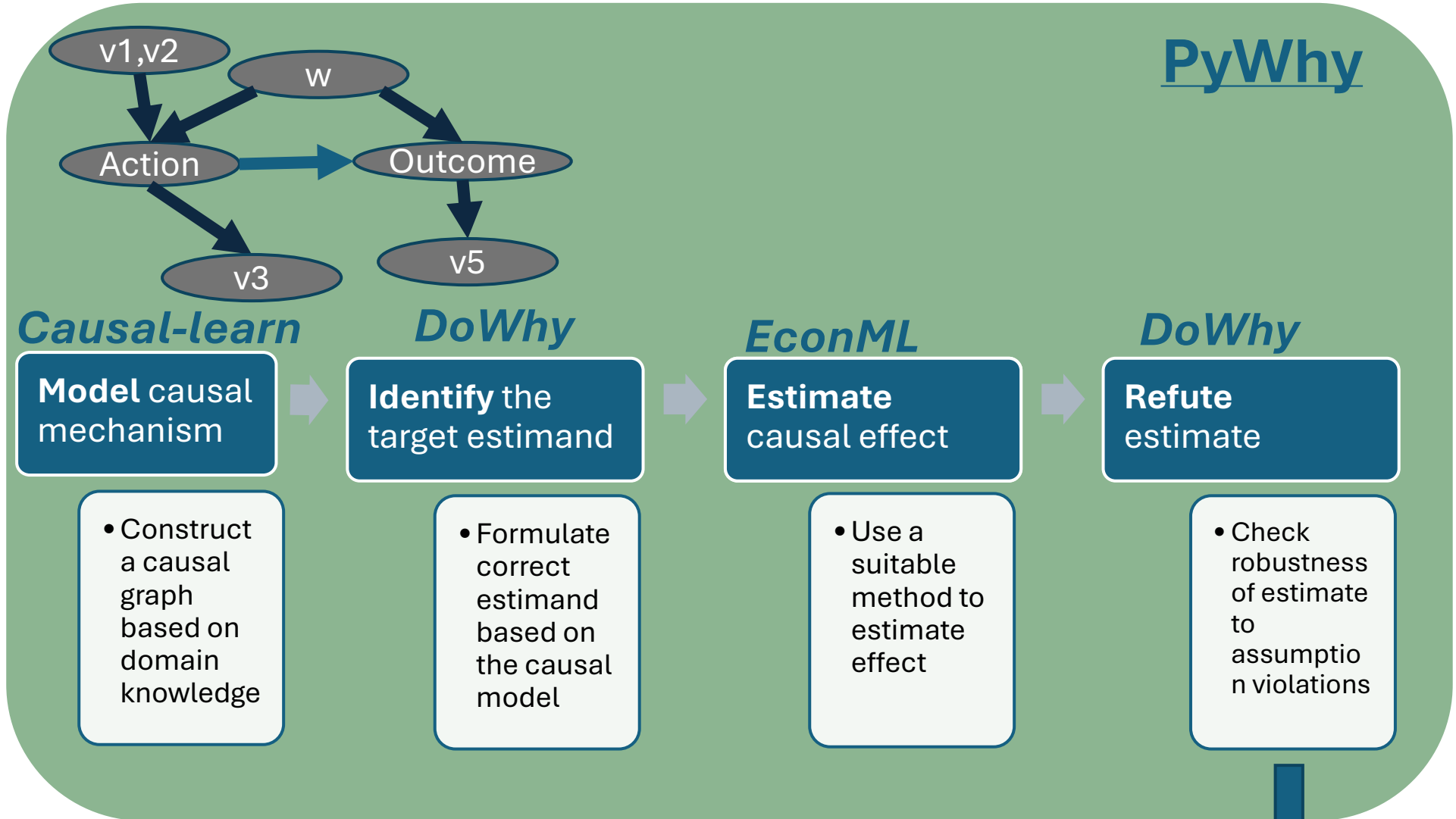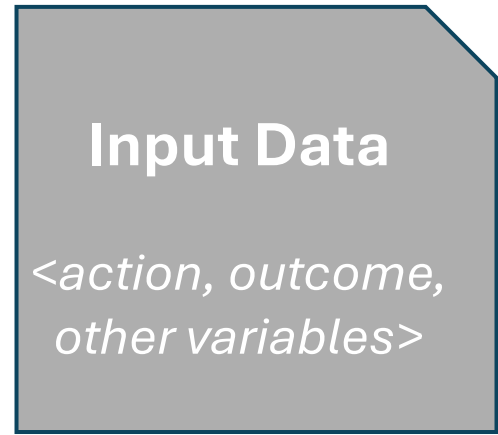
# An Open Source Ecosystem for Causal Machine Learning

# Section 2: Case studies using DoWhy/PyWhy

# **Case Study 1:** Estimating the effect of a customer loyalty rewards program

What is the impact of offering the customer loyalty program on total sales?

If the current members **had not signed up** for the program, how much less would they have spent?

**ATT:** *Average treatment effect on the treated* (customers who signed up for the program)

| | user_id | signup_month | month | spend | treatment |
|---|---|---|---|---|---|
| 0 | 0 | 6 | 1 | 507 | True |
| 1 | 0 | 6 | 2 | 506 | True |
| 2 | 0 | 6 | 3 | 490 | True |
| 3 | 0 | 6 | 4 | 464 | True |
| 4 | 0 | 6 | 5 | 475 | True |
| .... | .... | .... | .... | .... | .... |
| 119995 | 9999 | 0 | 8 | 396 | False |
| 119996 | 9999 | 0 | 9 | 387 | False |
| 119997 | 9999 | 0 | 10 | 367 | False |
| 119998 | 9999 | 0 | 11 | 436 | False |

You can try out this example on Github:
github.com/microsoft/dowhy/blob/master/docs/source/example_notebooks/dowhy_example_effect_of_memberrewards_program.ipynb

# Step 1: Modeling. Create causal graph to encode assumptions.

```python
model = dowhy.CausalModel(data=df_i_signupmonth,
                          graph=causal_graph.replace("\n", " "),
                          treatment="treatment",
                          outcome="post_spends")
```
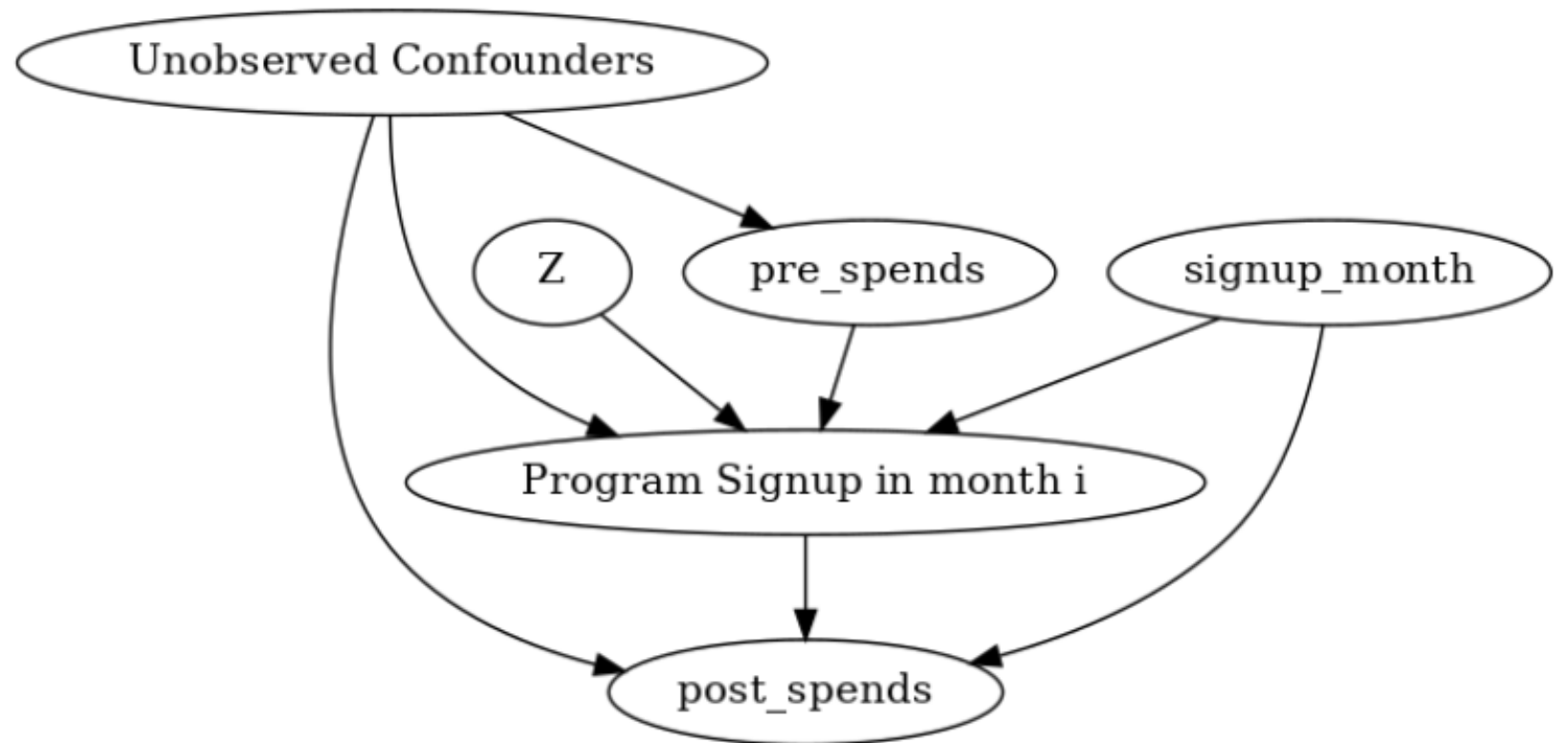
# Step 2: Identification. Formulate what to estimate

```python
identified_estimand = model.identify_effect(proceed_when_unidentifiable=True)
print(identified_estimand)
```

# Step 3: Estimation. Compute the estimate

```python
estimate = model.estimate_effect(identified_estimand,
                                  method_name="backdoor.propensity_score_matching",
                                  target_units="att")
print(estimate)
```

# Step 4: Refutation. Validate the assumptions

Obtained estimate depends on many (untestable) assumptions.

**Model:**

 Did we miss any unobserved variables in the assumed graph?

 Did we miss any edge between two variables in the assumed graph?

**Identify:**

 Did we make any parametric assumption for deriving the estimand?

**Estimate:**

 Is the assumed functional form sufficient for capturing the variation in data?

 Do the estimator assumptions lead to high variance?

# **Best practice:** Do refutation/robustness tests for as many assumptions as possible

**UNIT TESTS**

**Model:**

• Conditional Independence Test

**Identify:**

• D-separation Test

**Estimate:**

• Bootstrap Refuter

• Data Subset Refuter

**INTEGRATION TESTS**

**Test all steps at once.**

• Placebo Treatment Refuter

• Dummy Outcome Refuter

• Random Common Cause Refuter

• Sensitivity Analysis

• Simulated Outcome Refuter /Synth-validation [Schuler et al. 2017]

All these refutation methods are implemented in DoWhy.
**Caveat:** They can refute a given analysis, *but cannot prove its correctness.*

# Example 1: Conditional Independence Refuter

Through its edges, each causal graph implies certain conditional independence constraints on its nodes. *[d-separation, Pearl 2009]*

**Model refutation:** Check if the observed data satisfies the assumed model's independence constraints.

- Use an appropriate statistical test for independence *[Heinze-Demel et al. 2018]*.

- If not, the model is incorrect.
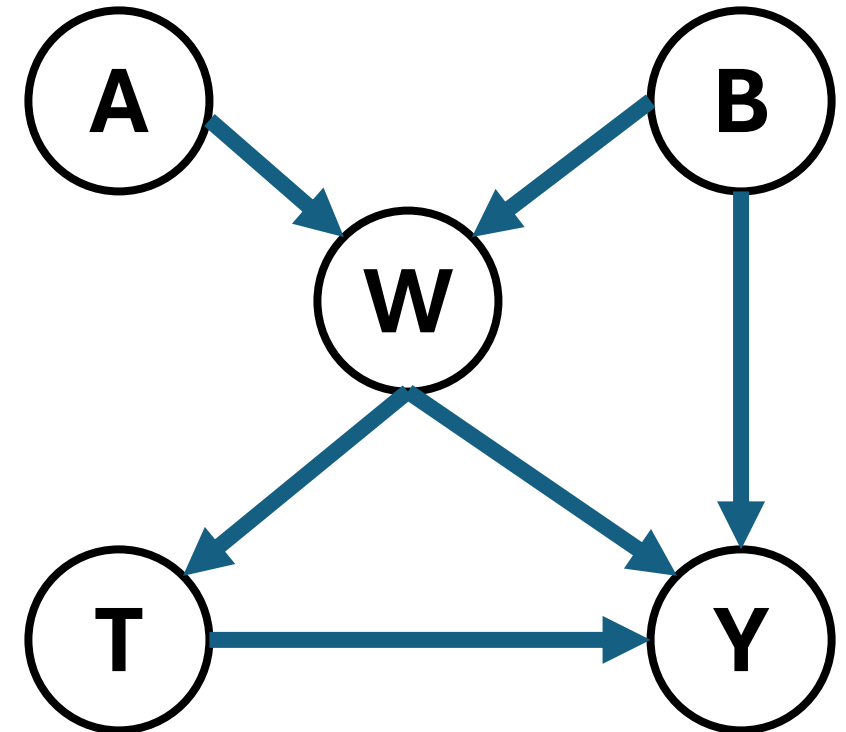


**Conditional Independencies:**

$$A \perp\!\!\!\perp B \qquad A \perp\!\!\!\perp T|W \qquad B \perp\!\!\!\perp T|W$$

# Example 2: Placebo Treatment ("A/A") Refuter

**Q:** *What if we can generate a dataset where the treatment does not cause the outcome?*

Then a correct causal inference method should return an estimate of zero.

**Placebo Treatment Refuter:**

Replace treatment variable T by a randomly generated variable (e.g., Gaussian).

- Rerun the causal inference analysis.
- If the estimate is significantly away from zero, then analysis is incorrect.
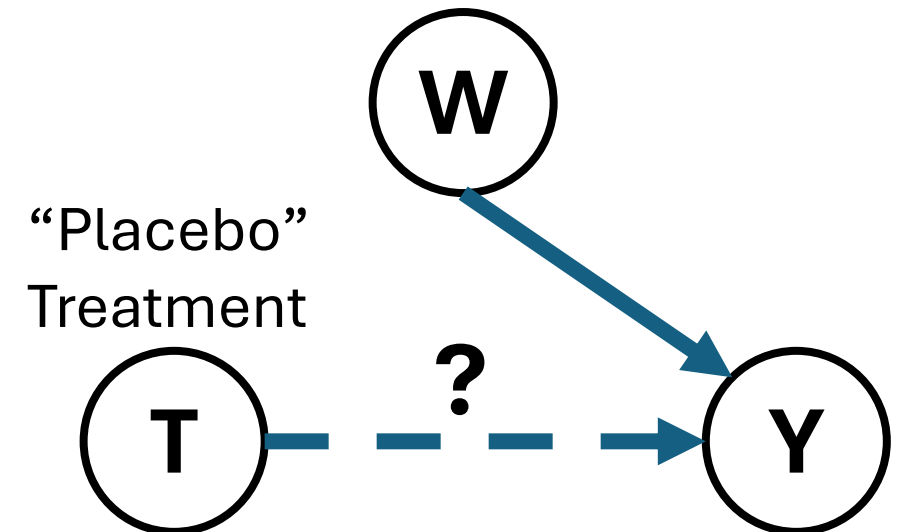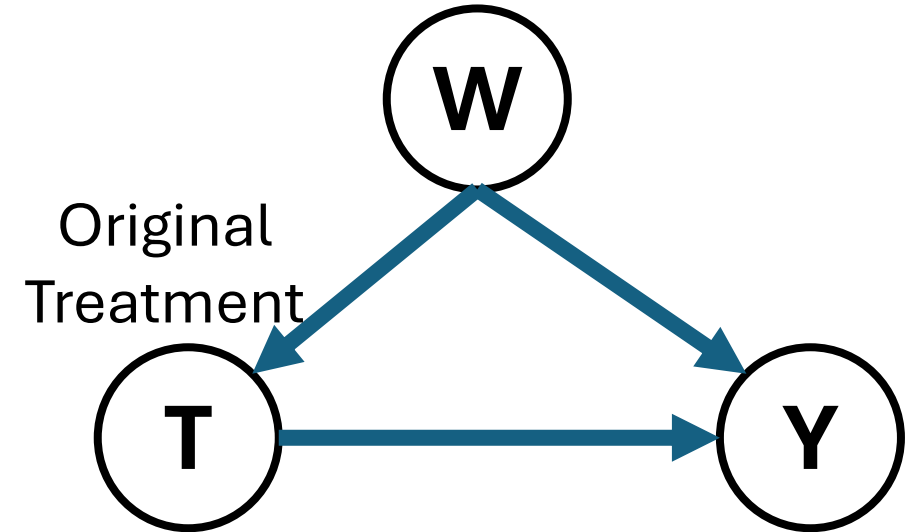
# Example 3: Add Unobserved Confounder to check sensitivity of an estimate

**Q:** *What if there was an unobserved confounder that was not included in the causal model?*

Check how sensitive the obtained estimate is after introducing a new confounder.

**Unobserved Confounder Refuter:**

- Simulate a confounder based on a given correlation $\rho$ with both treatment and outcome.
  - Maximum Correlation $\rho$ is based on the maximum correlation of any observed confounder.
- Re-run the analysis and check if the sign/direction of estimate flips.
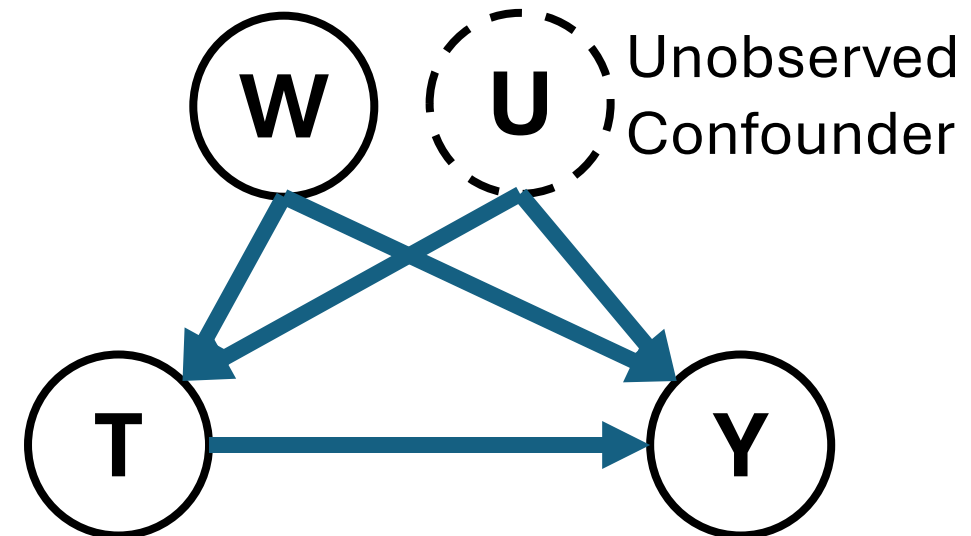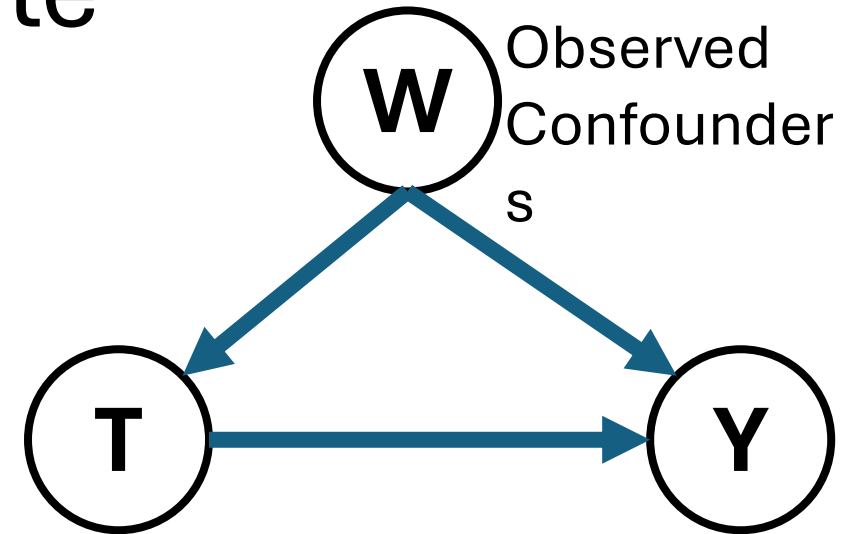
# **Step 4: Refutation.** Validate the assumptions
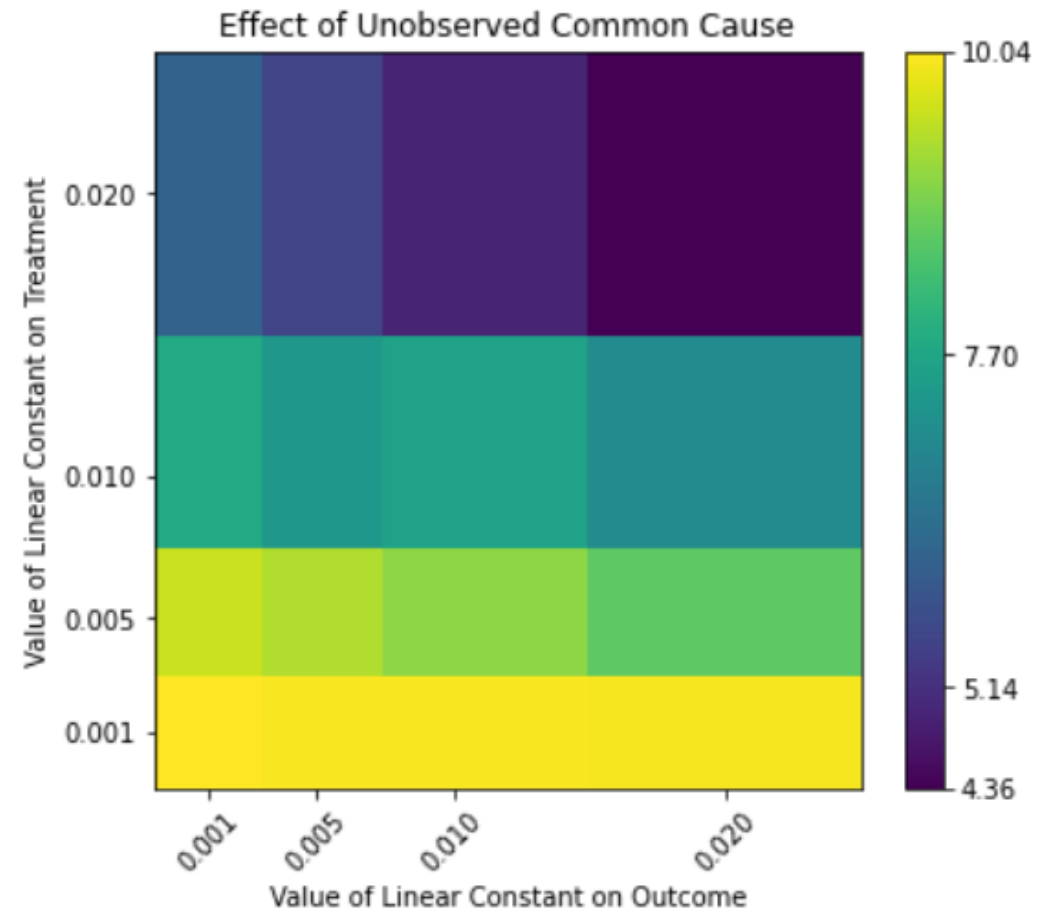
```
refutation = model.refute_estimate(identified_estimand, estimate, method_name="placebo_treatment_refuter",
                        placebo_type="permute", num_simulations=20)
print(refutation)
```

```
Refute: Use a Placebo Treatment
Estimated effect:100.03963044006804
New effect:0.6054947726720156
p value:0.24154316295878647
```



Effect of Unobserved Common Cause

# Case study 2: Finding the root cause of elevated latency in a microservice architecture

```
[2]: import pandas as pd

     normal_data = pd.read_csv("rca_microservice_architecture_latencies.csv")
     normal_data.head()
```

[2]:

| | Product DB | Customer DB | Order DB | Shipping Cost Service | Caching Service | Product Service | Auth Service | Order Service | API | www | Website |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.553608 | 0.057729 | 0.153977 | 0.120217 | 0.122195 | 0.391738 | 0.399664 | 0.710525 | 2.103962 | 2.580403 | 2.971071 |
| **1** | 0.053393 | 0.239560 | 0.297794 | 0.142854 | 0.275471 | 0.545372 | 0.646370 | 0.991620 | 2.932192 | 3.804571 | 3.895535 |
| **2** | 0.023860 | 0.300044 | 0.042169 | 0.125017 | 0.152685 | 0.574918 | 0.672228 | 0.964807 | 3.106218 | 4.076227 | 4.441924 |
| **3** | 0.118598 | 0.478097 | 0.042383 | 0.143969 | 0.222720 | 0.618129 | 0.638179 | 0.938366 | 3.217643 | 4.043560 | 4.334924 |
| **4** | 0.524901 | 0.078031 | 0.031694 | 0.231884 | 0.647452 | 1.081753 | 0.388506 | 0.711937 | 2.793605 | 3.215307 | 3.255062 |

# Why does the latency of website increase by 2 seconds?

```
[7]:  outlier_data = pd.read_csv("rca_microservice_architecture_anomaly.csv")
      outlier_data
```

[7]:

| | Product DB | Customer DB | Order DB | Shipping Cost Service | Caching Service | Product Service | Auth Service | Order Service | API | www | Website |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.493145 | 0.180896 | 0.192593 | 0.197001 | 2.130865 | 2.48584 | 0.533847 | 1.132151 | 4.85583 | 5.522179 | 5.572588 |

We are interested in the increased latency of `Website` which the customer directly experienced.

```
[8]:  outlier_data.iloc[0]['Website']-normal_data['Website'].mean()
```

[8]:  2.00895545064217

For this customer, `Website` was roughly 2 seconds slower than for other customers on average. Why?

# Step 1: Building the causal graph
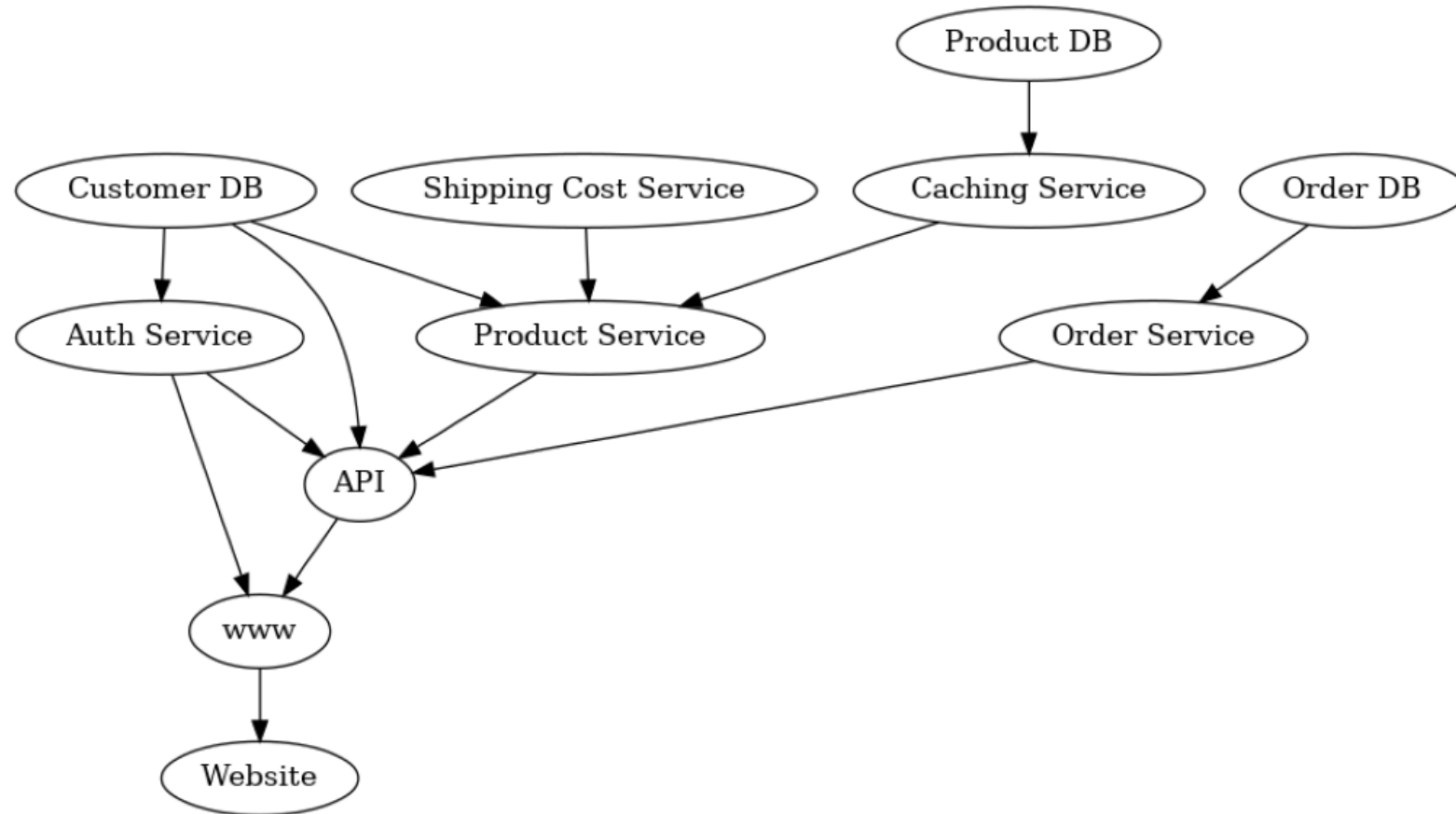
# Step 2: Identifying the estimate

In this case, we require a counterfactual estimate.

So, we assume all variables are observed, and fit an SCM.

```python
from scipy.stats import halfnorm

causal_model = gcm.StructuralCausalModel(causal_graph)

for node in causal_graph.nodes:
    if len(list(causal_graph.predecessors(node))) > 0:
        causal_model.set_causal_mechanism(node, gcm.AdditiveNoiseModel(gcm.ml.create_linear_regressor()))
    else:
        causal_model.set_causal_mechanism(node, gcm.ScipyDistribution(halfnorm))
```
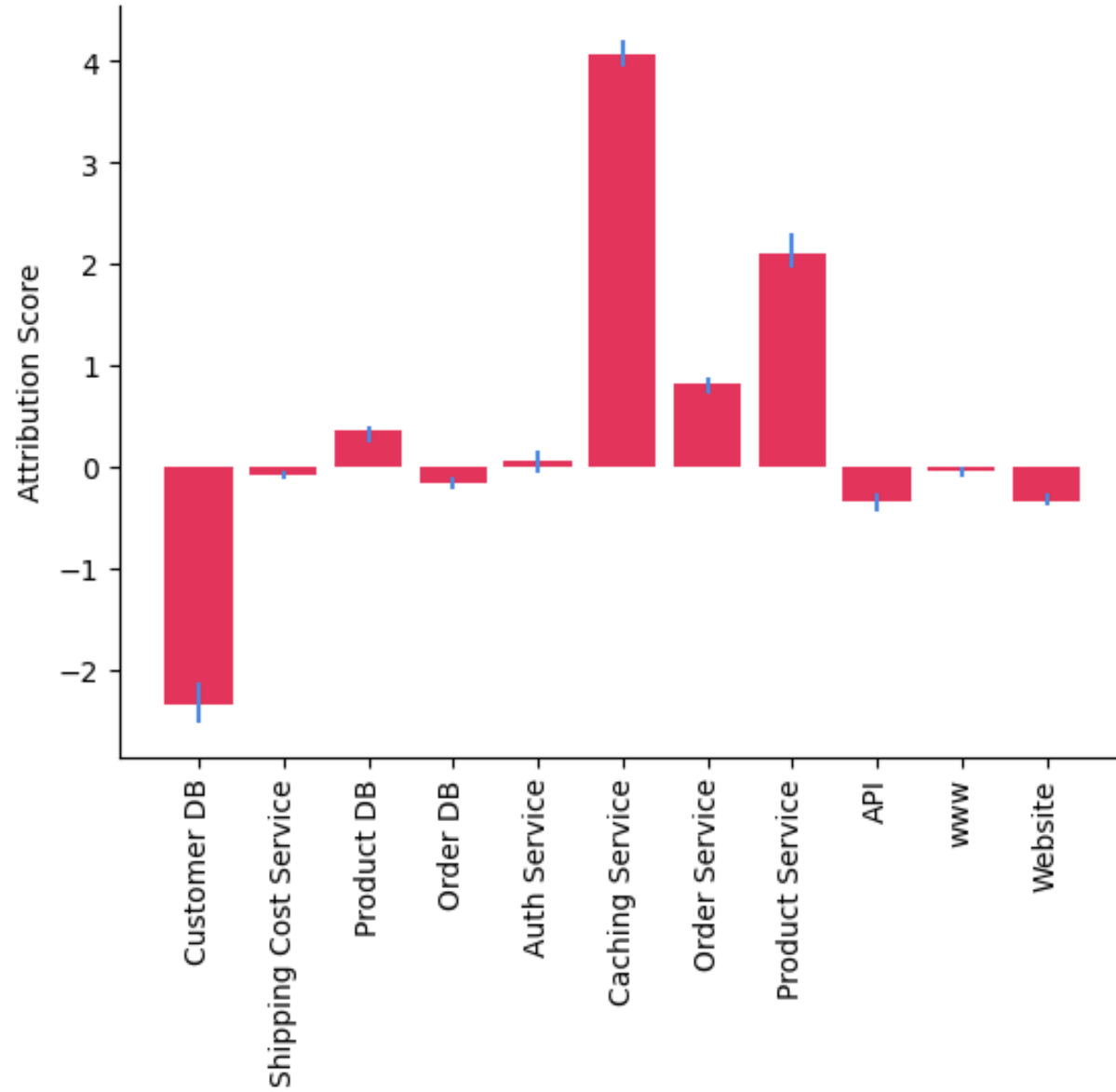
# Step 3: Estimating the counterfactual + attribution

```
[9]:  gcm.config.disable_progress_bars() # to disable print statements when computing Shapley values

      median_attribs, uncertainty_attribs = gcm.confidence_intervals(
          gcm.fit_and_compute(gcm.attribute_anomalies,
                              causal_model,
                              normal_data,
                              target_node='Website',
                              anomaly_samples=outlier_data),
          num_bootstrap_resamples=10)
```

By default, a quantile-based anomaly score is used that estimates the negative log-probability of a sample being normal. This is, the higher the probabilty of an outlier, the larger the score. The library offers different kinds of outlier scoring functions, such as the z-score, where the mean is the expected value based on the causal model.

```
[10]: bar_plot(median_attribs, uncertainty_attribs, 'Attribution Score')
```

# Case study 3: Classifying images under spurious correlation



train domain 1 ($E_1$)        train domain 2 ($E_2$)        test domain ($E_3$)

Domains in `MNISTCausalIndAttribute`

**Task:** detect the digit in each image.

# Step 1: Model the causal graph
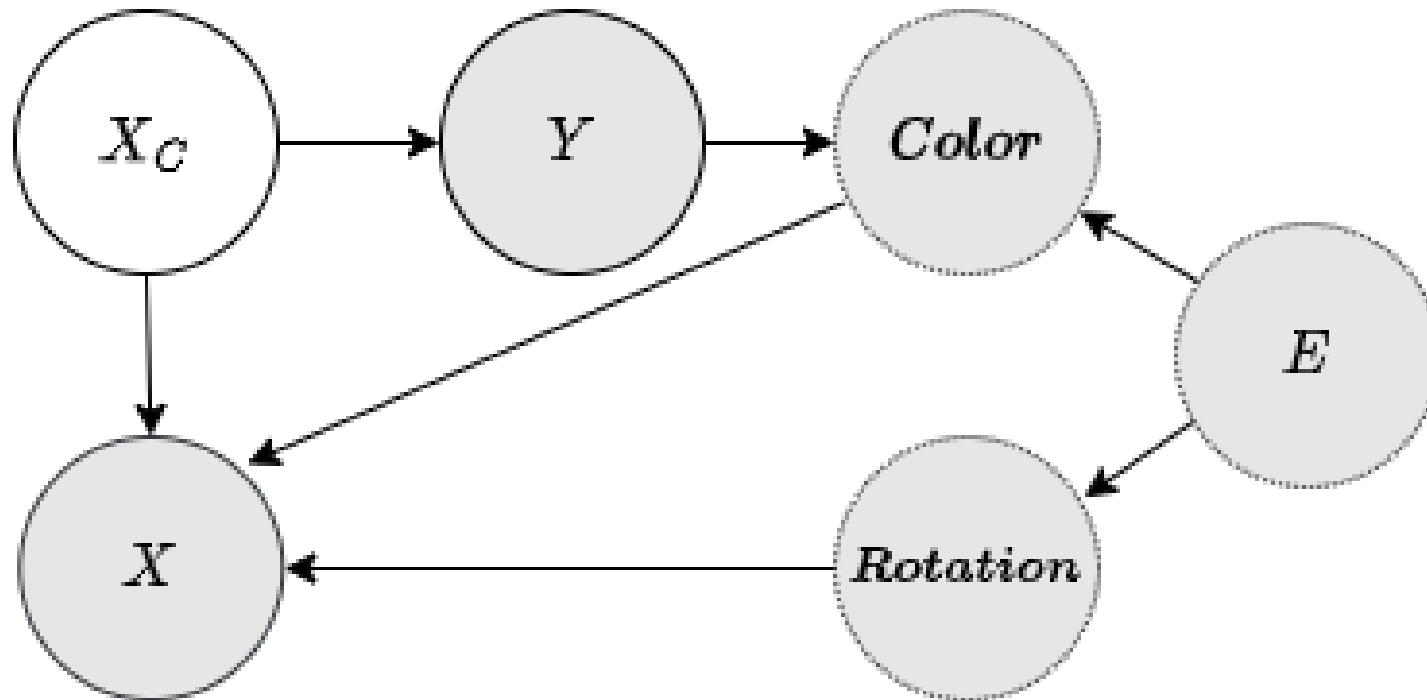
# Step 2: Identify the causal predictor

MNISTCausalIndAttribute: Multi-attribute *Causal+Independent* shift

```
[19]:  from dowhy.causal_prediction.datasets.mnist import MNISTCausalIndAttribute

       data_dir = 'data'
       dataset = MNISTCausalIndAttribute(data_dir)
```

```
[20]:  # `attr_types` should be ordered consistent with the attribute order in dataset class
       algorithm = CACM(model, lr=1e-3, gamma=1e-2, attr_types=['causal', 'ind'], lambda_causal=100., lambda_ind=10
```
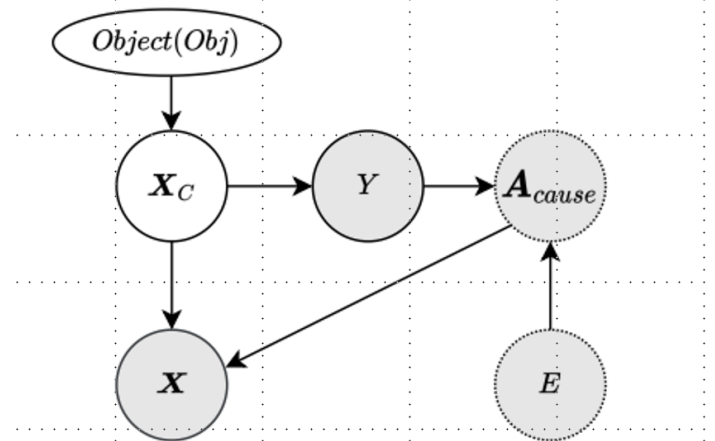


*Causal*

$$X_c \perp\!\!\!\perp A_{cause} \mid Y, E \checkmark$$
$$X_c \perp\!\!\!\perp A_{cause} \mid E \times$$

# Step 3: Fit the causal predictor

```
[15]: trainer = pl.Trainer(devices=1, max_epochs=5)

      trainer.fit(algorithm, loaders['train_loaders'], loaders['val_loaders'])
```

```
GPU available: False, used: False
TPU available: False, using: 0 TPU cores
IPU available: False, using: 0 IPUs
HPU available: False, using: 0 HPUs

  | Name  | Type       | Params
------------------------------------
0 | model | Sequential | 306 K
------------------------------------
306 K       Trainable params
0           Non-trainable params
306 K       Total params
1.226       Total estimated model params size (MB)
```

Testing DataLoader 0: 100% |████████████████████████| 157/157 [00:00<00:00, 158.14it/s]

```
     Test metric           DataLoader 0
─────────────────────────────────────────
      test_acc          0.6485999822616577
      test_loss         0.6817014217376709
─────────────────────────────────────────
```

# Section 3: The promise of large language models (LLMs) for causality

To what extent can we learn the causal graph?

# Domain knowledge is critical to correct causal analysis
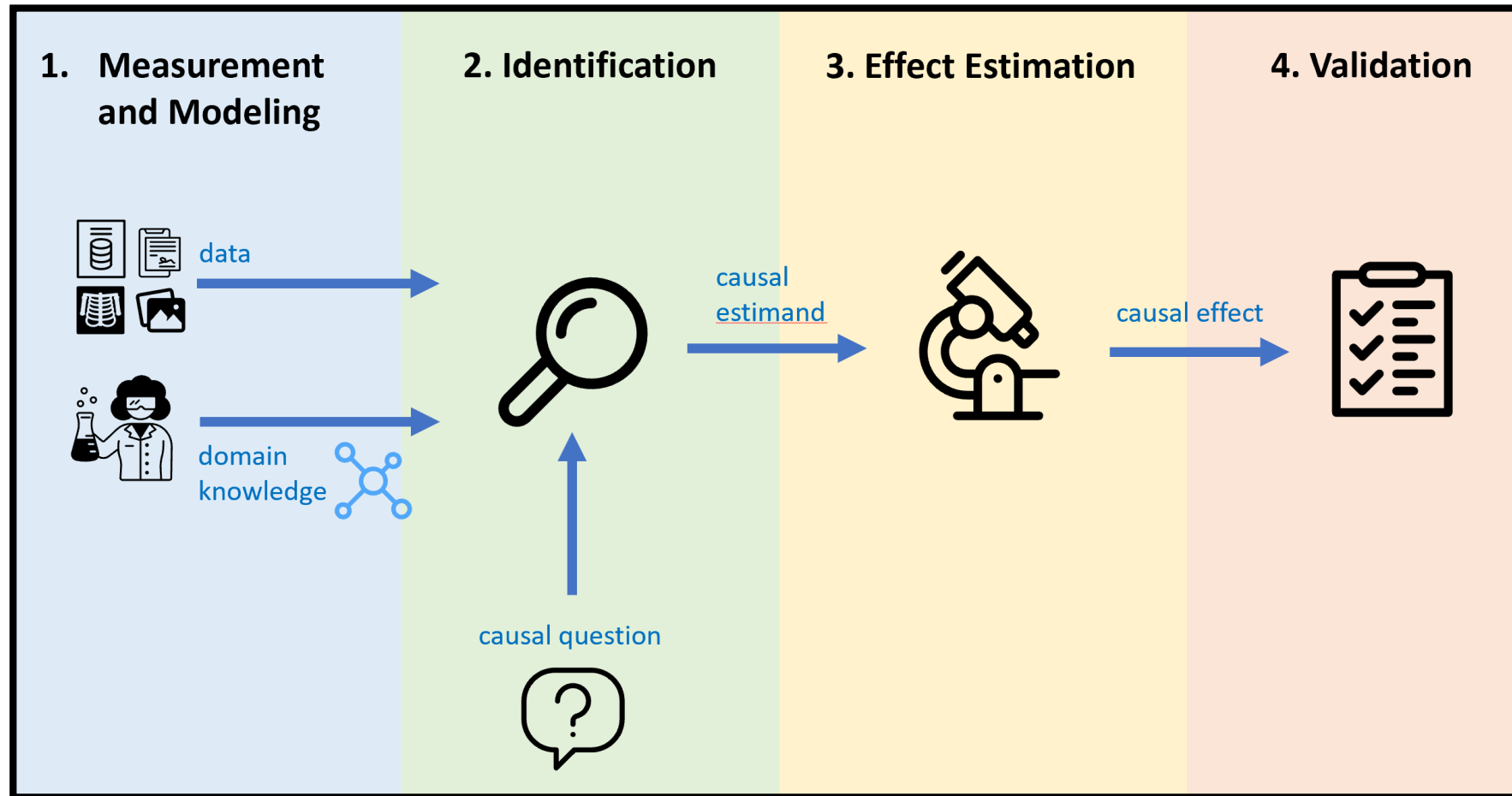


**Sobering results using SoTA graph discovery algorithms on real-world data,** due to the difficulty of learning causal relationships from data alone, [Tu et al. 2019, Huang et al. 2021, Kaiser & Sipos 2022]

# **Causal discovery:** Two common tasks

## **Pairwise discovery**

Given a pair of variables (A,B), decide whether A causes B or B causes A?
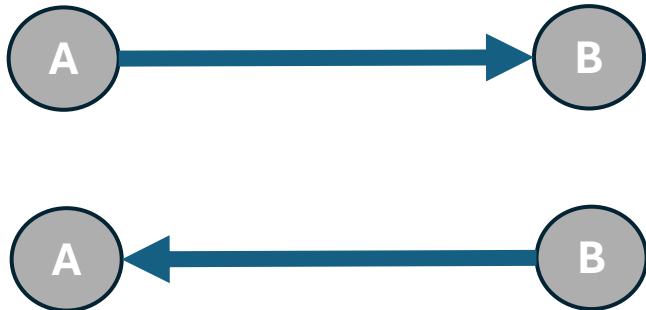
## **Full graph discovery**

Given a set of variables infer a directed acyclic graph over them.

- Infer which pairs of variables form an edge, and their direction.
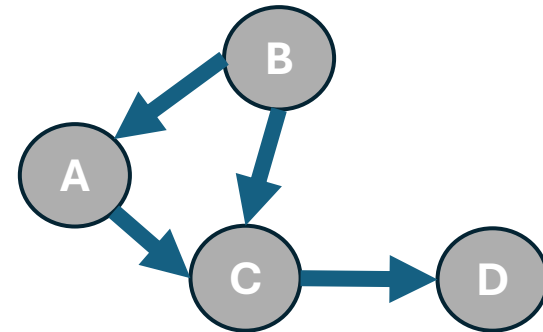
# Pairwise discovery
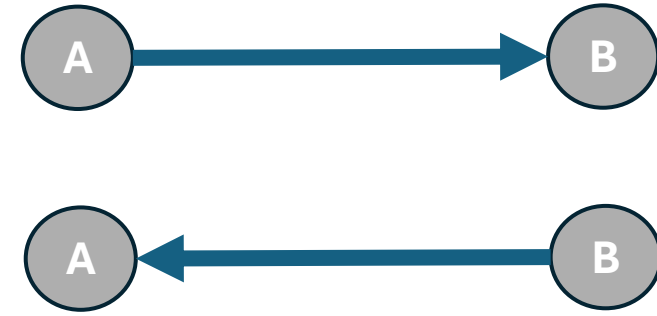


- Observed distribution: P(A,B)

- Graph 1 factorization: P(B|A) P(A)

- Graph 2 factorization: P(A|B) P(B)

Both have the **same likelihood given observed data.**

**Impossible** to determine direction non-parametrically.

 **Recent work:** assumes non-linear models or non-gaussian noise (e.g. LinGAM)

# Full graph discovery
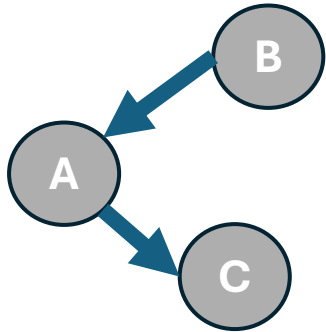


- **Chain**
- B indep C|A
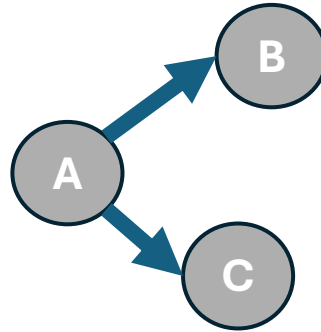
**Fork**
B indep C|A
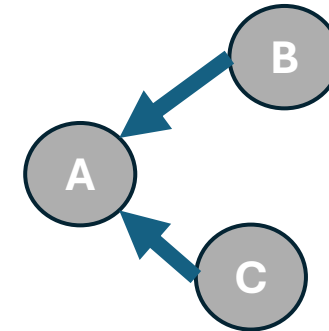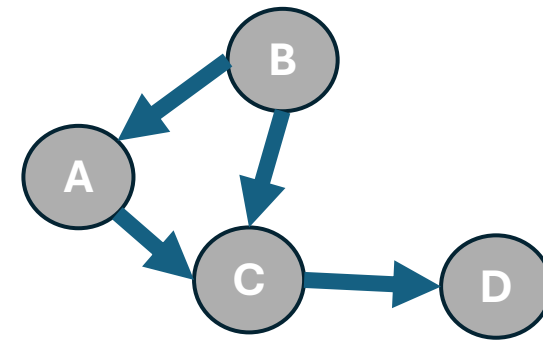
**Collider**
B indep C

For some edges, can determine the direction whenever there is a collider (PC algorithm)

**Recent work:** Formulates it as a continuous optimization problem and simply learns a graph with acyclic constraint (e.g., NOTEARS)

Now let's look at how LLMs may help

# **Pairwise discovery:** Tübingen Benchmark

- 104 variable pairs spanning range of fields [Mooij et al. 2016]

| Variable A | Variable B | Domain |
| --- | --- | --- |
| Age of Abalone | Shell weight | Zoology |
| Cement | Compressive strength of concrete | Engineering |
| Alcohol | Mean corpuscular volume | Biology |
| Organic carbon in soil | Clay content in soil | Pedology |
| PPFD (Photosynthetic Photon Flux Density) | Net Ecosystem productivity | Physics |
| Drinking water access | Infant mortality | Epidemiology |
| Ozone concentration | Radiation | Atmospheric Science |
| Contrast of tilted Gabor patches | Accuracy of detection by participants | Cognitive Science |
| Time for 1/6 rotation of a Stirling engine | Heat bath temperature | Engineering |
| Time for passing first segment of a ball track | Time for passing second segment | Basic Physics |

- **Challenging task:** Most existing algorithms achieve 70-80% accuracy, Best is 83% [Mosaic, Wu & Fukumizu 2020].

# Applying LLMs to pairwise causal discovery

**Method:** For each pair, input below prompts to an LLM and record the output.

## Two prompts per pair

**Template**:
- Does changing {A} cause a change in {B}? Please answer in a single word: yes or no.
- Does changing {B} cause a change in {A}? Please answer in a single word: yes or no.

**Examples:**
- Does changing the altitude cause a change in temperature? Please answer in a single word: yes or no.
- Does changing the temperature cause a change in altitude? Please answer in a single word: yes or no.

## Single prompt

**Template:**
- Which cause-and-effect relationship is more likely?
A. changing {A} causes a change in {B}.
B. changing {B} causes a change in {A}.

Let's work this out in a step by step way to be sure that we have the right answer. Then provide your final answer within the tags <Answer>A/B</Answer>.

**Example:**
- Which cause-and-effect relationship is more likely?
A. changing the altitude causes a change in temperature.
B. changing the temperature causes a change in altitude.

Let's work this out in a step by step way to be sure that we have the right answer. Then provide your final answer within the tags <Answer>A/B</Answer>.

SYSTEM: You are a helpful assistant for causal reasoning.

USER: Which cause-and-effect relationship is more likely?

A. changing the age of an abalone causes a change in its length.

B. changing the length of an abalone causes a change in its age.

Let's work this out in a step by step way to be sure that we have the right answer. Then provide your final answer within the tags <Answer>A/B</Answer>.

GPT-3.5-TURBO: 1. Cause: Changing the age of an abalone

Effect: a change in its length.

2. Cause: Changing the length of an abalone

Effect: a change in its age.

Based on the biological knowledge about abalones, it is more likely that changing the age of an abalone causes a change in its length. This is because the length of an abalone is known to increase as it ages, and not the other way around. Therefore, the correct answer is <Answer>A</Answer>.

(a) Correct answer

SYSTEM: You are a helpful assistant for causal reasoning.

USER: Which cause-and-effect relationship is more likely?

A. changing the age of an abalone causes a change in its diameter.

B. changing the diameter of an abalone causes a change in its age.

Let's work this out in a step by step way to be sure that we have the right answer. Then provide your final answer within the tags <Answer>A/B</Answer>.

GPT-3.5-TURBO: Step 1: Determine which variable is the cause and which is the effect.

In Option A, "age" is the cause and "diameter" is the effect.

In Option B, "diameter" is the cause and "age" is the effect.

Step 2: Consider which variable is more likely to cause a change in the other.

Abalone growth is primarily determined by environmental factors such as water temperature, food availability, and water quality. It is unlikely that changing the age of an abalone would cause its diameter to change significantly. On the other hand, changes in the environment could potentially impact the size of the abalone, and therefore, its diameter could change. Therefore, Option B is more likely: changing the diameter of an abalone causes a change in its age.

<Answer>B</Answer>

(b) Incorrect answer

Figure 3: **Probing reasoning of LLMs for deciding causal direction.** Two example outputs from a LLM (gpt-3.5-turbo). The first gives the impression of the LLM performing causal reasoning to arrive at the correct answer—that age may cause the length of abalone to increase. The second example asks a question that requires arguably the same causal knowledge but the LLM proceeds with an incoherent argument and provides an incorrect answer.

# **Results:** GPT-4 obtains 96% accuracy, 13 points higher than previous SoTA

Evaluate all available models from OpenAI

**Two prompts template:**

- With increased model size (and perhaps RLHF), begin to see gains.
  - Smaller models (ada, babbage, curie) are similar to random chance.

- With a single system prompt change, obtain **5 points gain**!
  - System prompt: "You are a helpful assistant for causal reasoning"

**Single prompt template:**
- Obtain further gains in accuracy
- GPT-4 obtains 97%
  - Surprising how few LLM errors are!

| Model | Acc. | Wt. Acc. |
|---|---|---|
| Slope (Marx & Vreeken, 2017) | 0.75 | 0.83 |
| bQCD (Tagasovska et al., 2020) | 0.68 | 0.75 |
| PNL-MLP (Zhang & Hyvarinen, 2012) | 0.75 | 0.73 |
| Mosaic (Wu & Fukumizu, 2020) | 83.3 | 81.5 |

**Covariance-based**

**Knowledge-based**

# Still LLMs need specific instructions

**Input Pair:** Ozone concentration, Radiation

LLM: 1. Ozone is a gas in Earth's atmosphere that absorbs most of the Sun's ultraviolet (UV) radiation.
2. Chlorofluorocarbons (CFCs), pollutants that were once widely used in refrigerants and aerosol sprays, have been found to destroy ozone molecules in the atmosphere.
3. The destruction of the ozone layer in the upper atmosphere causes more UV radiation to reach the surface of Earth.
4. Increased UV radiation can lead to various negative effects such as skin cancer, cataracts, and damage to crops.
From these facts, we can determine that a decrease in ozone concentration would lead to an increase in radiation. Therefore, Option A is more likely:
<Answer>A</Answer>

USER: Wait, this question is not about stratospheric ozone concentration. Rather, I'm asking about the ground-level ozone concentration measured in a city and the corresponding ground-level radiation due to sunlight.

GPT-3.5-TURBO:: I see, thank you for clarifying. In that case, the answer is likely option B: changing radiation causes a change in ground-level ozone concentration. This is because ground-level ozone is formed through a chemical reaction between nitrogen oxides (NOx) and volatile organic compounds (VOCs) in the presence of sunlight. This reaction is known as photochemical smog and occurs more frequently on hot, sunny days when there is more radiation. Therefore, an increase in radiation would likely lead to an increase in ground-level ozone concentration.

<Answer>B</Answer>

# Similar results on a neuropathic pain dataset

221 nodes & 475 edges about neuropathic pain diagnosis [Tu et al. 2019]. Use all edges as pairs.

| Variable A | Variable B | Dir. |
|---|---|---|
| Right L1 Radiculopathy | Right adductor tendonitis | → |
| Pharyngeal discomfort | Right C3 Radiculopahty | ← |
| Right L5 Radiculopathy | Lumbago | → |
| Left PTA | Left L4 Radiculopahty | ← |
| Left T3 Radiculopahty | Toracal dysfunction | → |
| DLS L5-S1 | Right S1 Radiculopathy | → |
| Left C3 Radiculopathy | DLS C2-C3 | ← |
| Left C7 Radiculopathy | Left medial elbow problem | → |
| Right Ischias | Right L5 Radiculopathy | ← |
| Right Morton trouble | Right L5 Radiculopathy | ← |

Table 3: Example cause-effect pairs from the Neuropathic pain diagnosis benchmark. 'Dir.' refers to the ground-truth causal direction between the variables.

| Model | Accuracy |
|---|---|
| ada | 40.1 |
| text-ada-001 | 50.0 |
| babbage | 50.0 |
| text-babbage-001 | 50.9 |
| curie | 50.0 |
| text-curie-001 | 50.0 |
| davinci | 38.4 |
| text-davinci-001 | 50.0 |
| text-davinci-002 | 51.7 |
| text-davinci-003 | 55.1 |
| gpt-3.5-turbo | 71.1 |
| gpt-3.5-turbo (neuropathic pain expert) | 75.1 |
| gp4-4 | 78.4 |
| gpt-4 (neuropathic pain expert) | 84.3 |
| text-davinci-003 (single prompt) | 86.0 |
| gpt-3.5-turbo (single prompt) | 85.5 |
| gpt-4 (single prompt) | **96.2** |

# Example: "Incorrect", but plausible LLM answer

**Input Pair:** L5 Radiculopathy, Obesity

**Neuropathic Pain dataset:** L5 Radiculopathy causes Obesity.

**LLM (gpt-3.5-turbo):** Obesity causes L5 Radiculopathy.

**Medical Literature:** Obesity may be a risk factor for radiculopathy [Atchison & Vincent, 2012; Tomic et al.,2009]

LLM: Step 1: Understand the terms
- Neuropathic pain: pain caused by damage or disease affecting the nervous system
- L5 Radiculopathy: a condition where the nerve root that exits the spine at the L5 level is compressed or irritated, causing pain, numbness, or weakness in the leg
- Obesity: a medical condition characterized by excessive body fat that increases the risk of health problems
Step 2: Analyze the cause-and-effect relationship
- Option A: Left L5 Radiculopathy causes Left Obesity. This relationship does not make sense because L5 radiculopathy is a nerve condition that affects the leg, while obesity is a whole-body condition related to excessive body fat. It is unlikely that a leg nerve condition would cause obesity.
- Option B: Left Obesity causes Left L5 Radiculopathy. This relationship is more plausible <u>because obesity can put extra pressure on the spine and nerves</u>, leading to nerve compression or irritation. It is possible that obesity could cause L5 radiculopathy.
Step 3: Choose the more likely cause-and-effect relationship
- Based on the analysis, Option B is more likely: Left Obesity causes Left L5 Radiculopathy.
<Answer>B</Answer>

# **Full graph discovery:** Neuropathic pain dataset

- Use a 100 pair subsample [Tu et al. 2023]
  - 50 correct edges, 50 non-existent edges
  - Need to infer existence + direction of an edge

---

**Single prompt**

---

**Template:**
- Which cause-and-effect relationship is more likely? Consider only direct causal mechanism and ignore any effect due to common causes.
A. {A} causes {B}.
B. {B} causes {A}.
C. No causal relationship exists.

Let's work this out in a step by step way to be sure that we have the right answer. Then provide your final answer within the tags <Answer>A/B</Answer>.
**Example:**
- Which cause-and-effect relationship is more likely? Consider only direct causal mechanism and ignore any effect due to common causes.
A. Left L4 Radiculopathy causes Right Iliac Crest pain.
B. Right Iliac Crest pain causes Left L4 Radiculopathy
C. No causal relationship exists.

Let's work this out in a step by step way to be sure that we have the right answer. Then provide your final answer within the tags <Answer>A/B</Answer>.
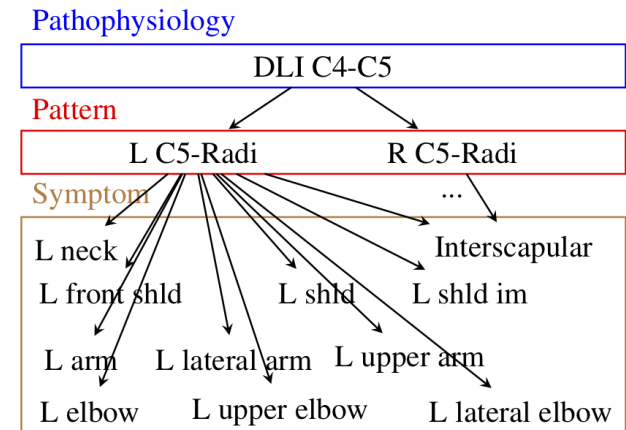
---

Ground-truth graph

# **Prompt matters!** With an improved prompt, we obtain 3x better accuracy in graph discovery

| Model | Precision | Recall | F1 |
|---|---|---|---|
| Random | 0.25 | 0.5 | 0.33 |
| chatGPT (Tu et al., 2023) | 1 | 0.12 | 0.21 |

- Simple use of ChatGPT resulted in **0.21 F1** on retrieving edges correctly.

- With our single prompt, gpt-3.5-turbo (ChatGPT) obtains **0.68 F1.**

# On an atmospheric science dataset, LLMs obtain higher accuracy than recent deep learning algorithms

**Arctic sea ice dataset:** 12 variable, 48 edges graph on the drivers of sea ice thickness in the Arctic region.

**Variables:** Total cloud water path, sea level pressure, geopotential height, meridional and zonal wind at 10m, net shortwave and longwave flux at surface, etc.

| Algorithm | NHD | No. of predicted edges | Baseline NHD | Ratio |
|---|---|---|---|---|
| TCDF | 0.33 | 9 | 0.39 | 0.84 |
| NOTEARS (Static) | 0.33 | 15 | 0.44 | 0.75 |
| NOTEARS (Temporal) | 0.35 | 7 | 0.38 | 0.92 |
| DAG-GNN (Static) | 0.32 | 23 | 0.49 | 0.65 |
| DAG-GNN (Temporal) | 0.34 | 16 | 0.44 | 0.77 |
| gpt-3.5-turbo | 0.33 | 62 | 0.76 | 0.43 |
| gpt-4 | **0.22** | 46 | 0.65 | **0.34** |

Table 7: Normalized hamming distance (NHD) for different causal discovery algorithms. Since NHD depends on the number of predicted edges, we compare the ratio of NHD and baseline NHD across algorithms. A lower NHD ratio is better. LLM-based discovery (gpt-3.5-turbo) obtains comparable NHD and the lowest NHD ratio compared to recent covariance-based discovery algorithms.
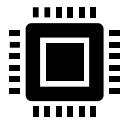
# Construct Validity: Is Benchmark Memorized?

**SYSTEM**

You are an AI assistant that has read many sources of text from the internet. I am looking at text from the dataset, published by as . Here is the README for the dataset: I am going to list some sample rows of data, and I want you to complete each row as best as possible. I am testing your memory.

**USER**

pair0005 Age

**GPT**

pair0005 Age Length Abalone ->

|          | Cells  | Rows   |
|----------|--------|--------|
| **GPT-3.5** | 58.9%  | 19.8%  |
| **GPT-4**   | 61%    | 25%    |

**Yes, Tübingen dataset clearly in the training dataset.**

# Construct Validity: What are we Measuring?

Let us model knowledge-based discovery as:

$$P(Y|D)\ P(D)$$

**Likelihood that knowledge can be transformed to answer question Y**    **Likelihood that knowledge D is known by LLM**

- With memorized benchmark data, we are *not* measuring $P(D)$
- We *are* measuring: how LLM can process and transform $D$ into the necessary causal relationship $Y$

# Takeaways from the causal discovery section

- **LLMs enable knowledge-based causal discovery**
  - Competitive performance in determining pairwise causal relationships
  - Across multiple datasets in varied domains incl. medicine and climate science
- **Full graph discovery poses additional challenges**
  - E.g., distinguishing between direct and indirect causes
  - LLMs provide non-trivial utility for inferring full causal graphs
- **GPT-3.5, GPT-4 have memorized Tübingen benchmark**
  - Our results are valid measurements of LLM ability to transform knowledge into a causal answer
  - Not valid for estimating the likelihood that arbitrary relationship has been memorized

# Conclusion

- Causal ML is important whenever we have decision-making or attribution tasks, or want generalizability of predictive model beyond the training distribution.

- Causal graph is the most important assumption.
  - "No causes in, no causes out" – Judea Pearl

- Open-source tools help make this assumption explicit and transparently share analysis with others.

- But obtaining a causal graph is still the hardest problem.
  - Large language models can help the domain expert in building it.

# Causal ML extras: What we **did not cover**

- Causal identification using **instrumental variables**
- **Propensity-based estimation** methods for effect estimation
- **Refutation and sensitivity analyses** for causal ML
- Connections to **reinforcement learning**

Refer to book draft, Causal Reasoning: Fundamentals and Machine Learning Applications (**causalinference.gitlab.io/book**)

# Thank you!

Amit Sharma