

CSDS 452 Causality and Machine Learning

Lecture 14: Other Work about Causal Inference with Machine Learning

Instructor: Jing Ma

Fall 2024, CDS@CWRU

Outline

- Causal discovery (causal structural learning) through neural network
- Causal discovery with large language model

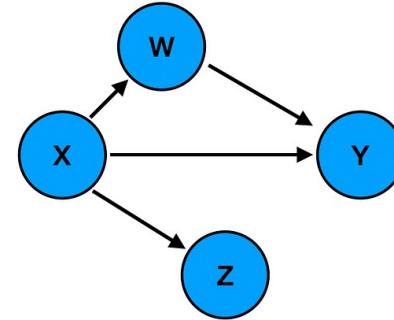
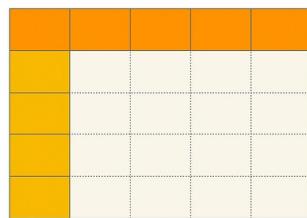
Causal Structural Learning

Want to know more?



Causal Structure Learning,
Heinze-Deml et. al., Annual
Reviews of Statistics, 2018

Goal: Learn a causal graph from data



Causal Structural Learning

Want to know more?



Causal Structure Learning,
Heinze-Deml et. al., Annual
Reviews of Statistics, 2018

Goal: Learn a causal graph from data

Problems:

- Space of causal graphs grows super-exponentially
- Finding the right DAG is hard
- Limited identifiability from purely observational data

Challenges:

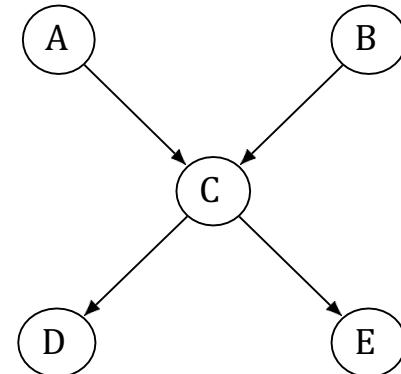
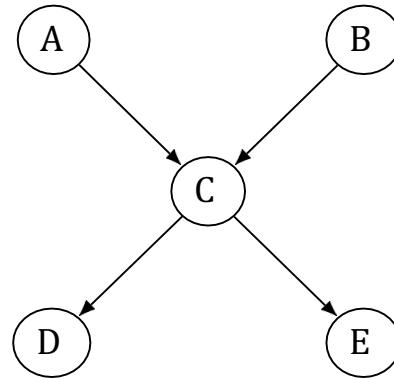
- Efficient Search in space of causal graphs

# of Nodes	# of DAGs	# of Graphs
1	1	1
2	3	2^2
3	25	2^6
4	543	2^{12}
5	29281	2^{20}
6	3781503	2^{30}
7	1138779265	2^{42}
8	783702329343	2^{56}
9	1213442454842881	2^{72}
10	4175098976430598143	2^{90}

Recap: The PC Algorithm

- Start with complete undirected graph
 - 3 steps
1. Identify the skeleton
 2. Identify immoralities and orient them
 3. Orient more edges incident on colliders
(based on the fact that we would have identified more immoralities in Step 2)

True causal graph



Common Assumptions

- **Faithfulness:** for any disjoint variable sets X, Y, Z

$$X \perp\!\!\!\perp_P Y | Z \quad \Rightarrow \quad X \perp\!\!\!\perp_G Y | Z$$

Data Causal graph

- **Causal Sufficiency:** there are no unobserved confounders of any of the variables in the graph
- **Acyclicity:** there are no cycles in the graph.

Causal Structural Learning

Want to know more?



Causal Structure Learning,
Heinze-Deml et. al., Annual
Reviews of Statistics, 2018

Causal Structure Learning

Constraint-based Approaches

- Use of conditional independence tests between variables
- Use such tests to iteratively delete edges and direct the remaining ones

Representative methods: Peter-Clark (PC),
Fast causal inference (FCI)

Score-based Approaches

- Aim to find the highest scoring graph according to a scoring metric:

$$G^* = \arg \max_{G \in \text{Causal Graphs}} S(D|G)$$

- Variations between methods:
 - Score Formulation
 - Search Procedure

Representative methods: Greedy Equivalent Search (GES), Fast GES

Score-based

- In general, score-based approaches seeks a candidate graph G^* (over the set of DAGs) that can maximize a scoring function S

$$G^* = \operatorname{argmax}_G S(D|G)$$

Score-based

- In general, score-based approaches seek a candidate graph G^* (over the set of DAGs) that can maximize a scoring function S

$$G^* = \operatorname{argmax}_G S(D|G)$$

- Limitation:
 - NP-hard, due to its nonconvex, combinatorial nature
 - The **acyclicity** constraint is a **combinatorial** constraint with the number of acyclic structures increasing **superexponentially** in the number of nodes

Structure Learning: Where Are We?

	MNs	BNs	Comments
constraint-based	✓	✓	need faithfulness
score-based, local search	✓	✓	combinatorial opt.

Challenges: The search space of DAGs is **combinatorial** and scales **superexponentially** with the number of nodes.

Structure Learning: Where Are We?

	MNs	BNs	Comments
constraint-based	✓	✓	need faithfulness
score-based, local search	✓	✓	combinatorial opt.
score-based, global search	✓ [†]	?*	continuous opt.

For undirected (Markov) graphs: some studies re-formulate structural learning as a **convex program** over **real, symmetric** matrices

=> **Efficient** algorithms for learning undirected graphs!

Structure Learning: Where Are We?

	MNs	BNs	Comments
constraint-based	✓	✓	need faithfulness
score-based, local search	✓	✓	combinatorial opt.
score-based, global search	✓ [†]	No TEARS	continuous opt.

*Challenges in Bayesian Networks:

- Directed graph → asymmetric matrix
- Acyclic graph → combinatorial constraint

DAGs with No TEARS

NO TEARS: Non-combinatoric Optimization via Trace
Exponential Augmented lagRangian Structure learning

DAGs with No TEARS

NO TEARS: Non-combinatorial Optimization via Trace Exponential Augmented lagRangian Structure learning

$$\max_G \text{score}(G)$$

$$s.t. G \in DAG$$

(combinatorial 😱)

Challenge: NP-hard, due to its nonconvex, combinatorial nature

DAGs with No TEARS

NO TEARS: Non-combinatorial Optimization via Trace
Exponential Augmented lagRangian Structure learning

$$\begin{aligned} & \max_G \text{score}(G) \\ & \text{s.t. } G \in \text{DAG} \end{aligned}$$

(combinatorial 😱)

Suppose we have d variables (nodes) in DAG $G = (V, E)$ on d nodes

Given n observations, we have
 n observational data samples for vector
 $X = (X_1, \dots, X_d)$

Challenge: NP-hard, due to its nonconvex, combinatorial nature

DAGs with No TEARS

NO TEARS: Non-combinatorial Optimization via Trace
Exponential Augmented lagRangian Structure learning

$$\begin{aligned} \max_G \text{score}(G) \\ \text{s.t. } G \in \text{DAG} \end{aligned} \qquad \iff \qquad$$

$$\begin{aligned} \max_W \text{score}(W) \\ \text{s.t. } G(W) \in \text{DAG} \end{aligned}$$

(combinatorial 😱)

We model X via a SCM with structural equations defined by a weighted adjacency matrix $W \in R^{d \times d}$
 $W = [w_1 | \dots | w_d]; \quad X_j = w_j^T X + noise_j$

Instead of discrete space, we will optimize over a continuous space of $d \times d$ real matrix

DAGs with No TEARS

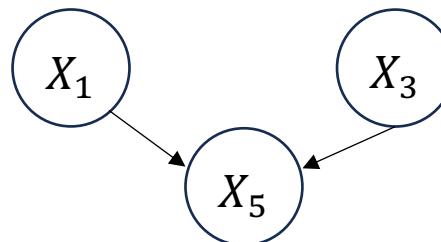
NO TEARS: Non-combinatorial Optimization via Trace
Exponential Augmented lagRangian Structure learning

$$\begin{aligned} \max_G \text{score}(G) \\ \text{s.t. } G \in \text{DAG} \end{aligned} \qquad \qquad \qquad \iff \qquad \qquad \qquad \begin{aligned} \max_W \text{score}(W) \\ \text{s.t. } G(W) \in \text{DAG} \end{aligned}$$

(combinatorial 😱)

We model X via a SCM with structural equations defined by a weighted adjacency matrix $W \in R^{d \times d}$
 $W = [w_1 | \dots | w_d]; \quad X_j = w_j^T X + noise_j$

$$w_5^T \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{bmatrix} + noise_j = X_5$$



Instead of discrete space, we will optimize over a continuous space of $d \times d$ real matrix

DAGs with No TEARS

NO TEARS: Non-combinatorial Optimization via Trace
Exponential Augmented lagRangian Structure learning

$$\begin{array}{ccc} \max_G score(G) & \iff & \max_W score(W) \\ s.t. G \in DAG & & s.t. G(W) \in DAG \end{array}$$

(combinatorial 😱)

Challenge: The acyclicity constraint is a combinatorial constraint with the number of acyclic structures increasing superexponentially in d

A new characterization of acyclicity

- Main idea: replace the combinatorial acyclicity constraint $G(W) \in DAG$ with a single smooth equality constraint $h(W) = 0$

A new characterization of acyclicity

- Main idea: replace the combinatorial acyclicity constraint $G(W) \in DAG$ with a single smooth equality constraint $h(W) = 0$
- Ideally, we want a function h :
 - $h(W) = 0$ if and only if W is acyclic (i.e., $G(W) \in DAG$)
 - The values of h quantify the “DAG-ness” of the graph
 - h is smooth;
 - h and its derivatives are easy to compute

DAGs with No TEARS

NO TEARS: Non-combinatorial Optimization via Trace
Exponential Augmented lagRangian Structure learning

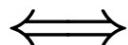
$$\begin{array}{ll} \max_G score(G) & \iff \\ s.t. \ G \in DAG & \max_W score(W) \\ & s.t. \ h(W) \leq 0 \text{ enforcing acyclicity} \end{array}$$

(combinatorial 😱)

(smooth 😎)

DAGs with No TEARS

$$\begin{aligned} \max_G \text{score}(G) \\ \text{s.t. } G \in \text{DAG} \end{aligned}$$



$$\begin{aligned} \max_W \text{score}(W) \\ \text{s.t. } h(W) \leq 0 \end{aligned}$$

(combinatorial 😱)

(smooth 😎)

Smooth Characterization of DAG

Such function exists: $h(W) = \text{tr}(e^{W \circ W}) - d$.

Moreover, simple gradient: $\nabla h(W) = (e^{W \circ W})^T \circ 2W$.

Summary

- A smooth characterization of DAG:

$$h(W) = \text{tr}(e^{W \circ W}) - d \leq 0 \iff G(W) \in \text{DAG}$$

- Use existing solvers for constrained optimization problem:

$$\begin{aligned} & \max_W \quad \text{score}(W) \\ & s.t. \quad h(W) \leq 0 \end{aligned}$$

- Bridge optimization and structure learning

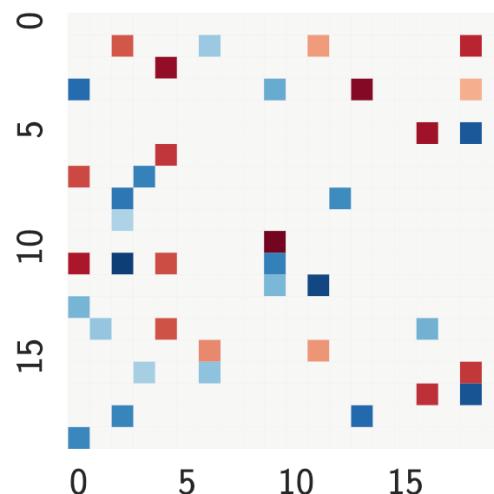
NO TEARS

- Available at: github.com/xunzheng/notears
 - 30 lines (function, gradient) + 20 lines (optimize) \approx 50 lines
 - Existing algorithms: \gg 1000 lines

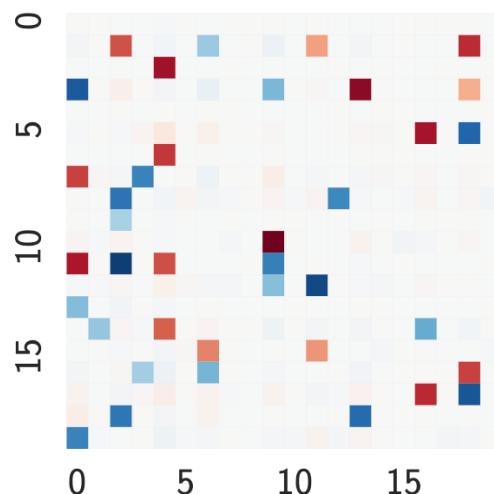
```
1 def notears_simple(X, max_iter=100, h_tol=1e-8, w_threshold=0.3):
2     n, d = X.shape
3     w_est, w_new = np.zeros(d * d), np.zeros(d * d)
4     rho, alpha, h, h_new = 1.0, 0.0, np.inf, np.inf
5     bnds = [(0, 0) if i == j else (None, None) for i in range(d) for j in range(d)]
6     for _ in range(max_iter):
7         while rho < 1e+20:
8             sol = sopt.minimize(_func, w_est, method='L-BFGS-B', jac=_grad, bounds=bnds)
9             w_new = sol.x
10            h_new = _h(w_new)
11            if h_new > 0.25 * h:
12                rho *= 10
13            else:
14                break
15            w_est, h = w_new, h_new
16            alpha += rho * h
17            if h <= h_tol:
18                break
19            w_est[np.abs(w_est) < w_threshold] = 0
20    return w_est.reshape([d, d])
```

Results: Recovering Erdos-Renyi Graph

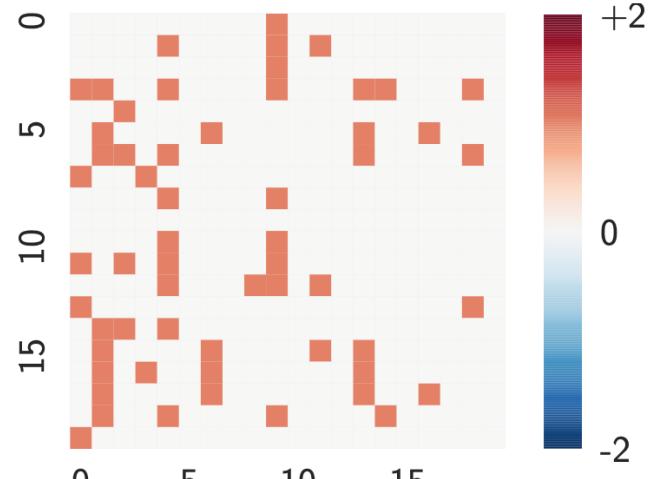
- ER graph is a type of graph models, where edges are randomly generated with certain ratio



ground truth



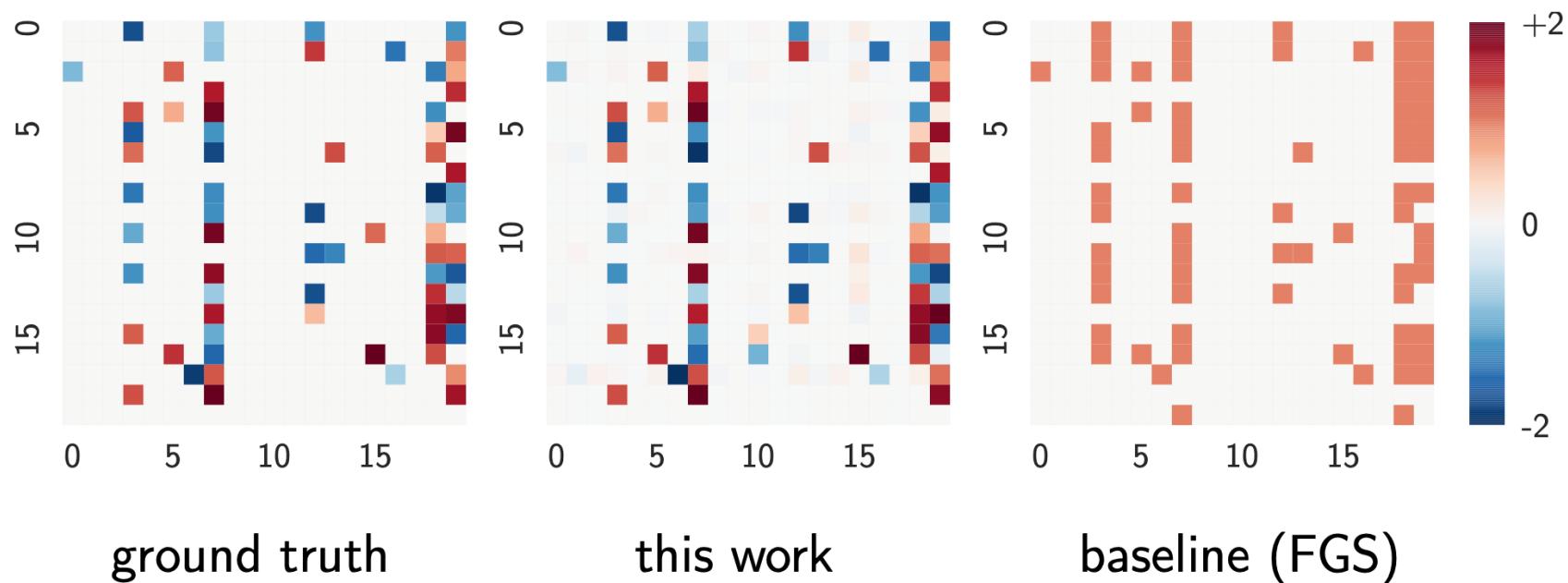
this work



baseline (FGS)

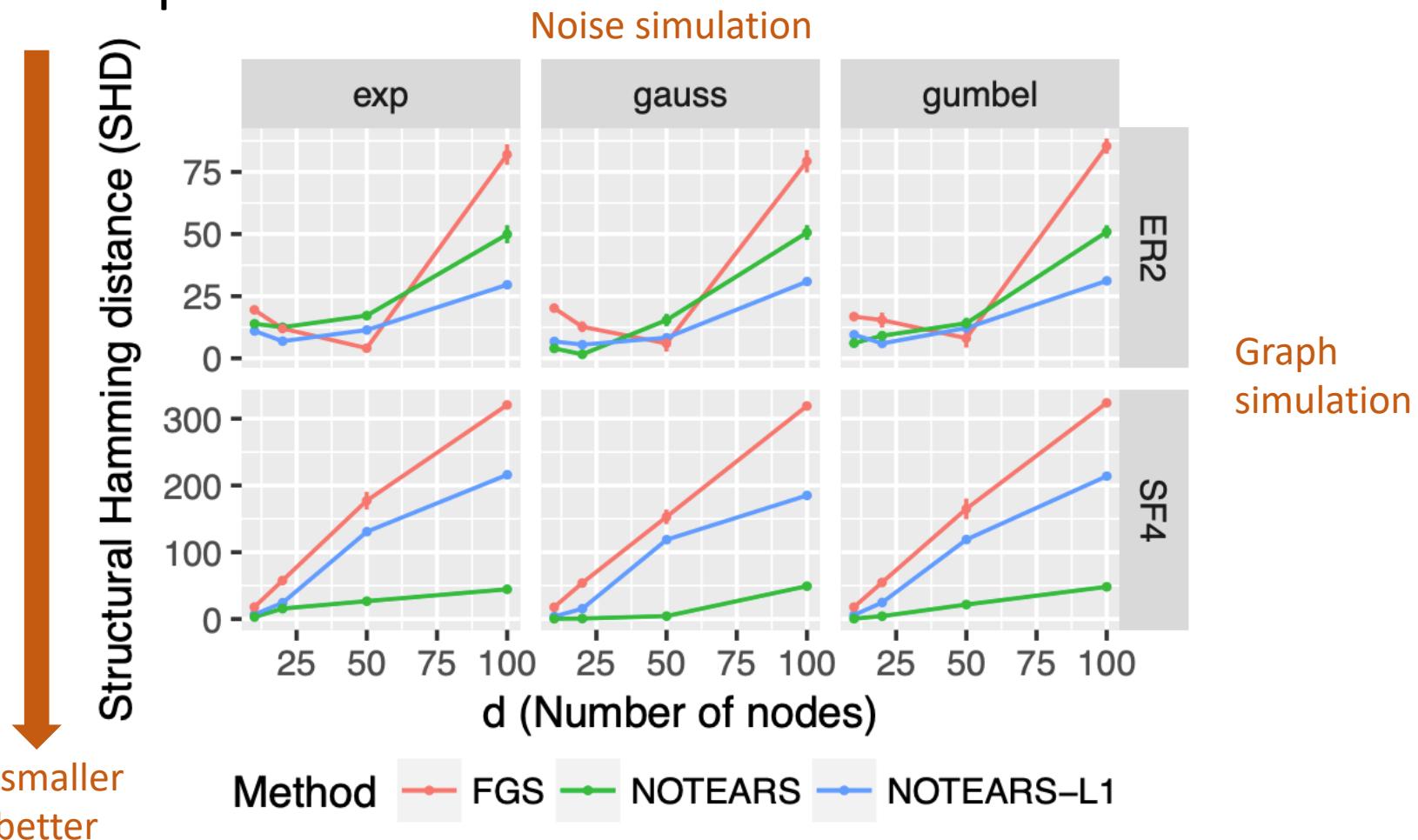
Results: Recovering Scale-free Graph

- Scale-free graph is a type of graph models, whose degree distribution follows a power law



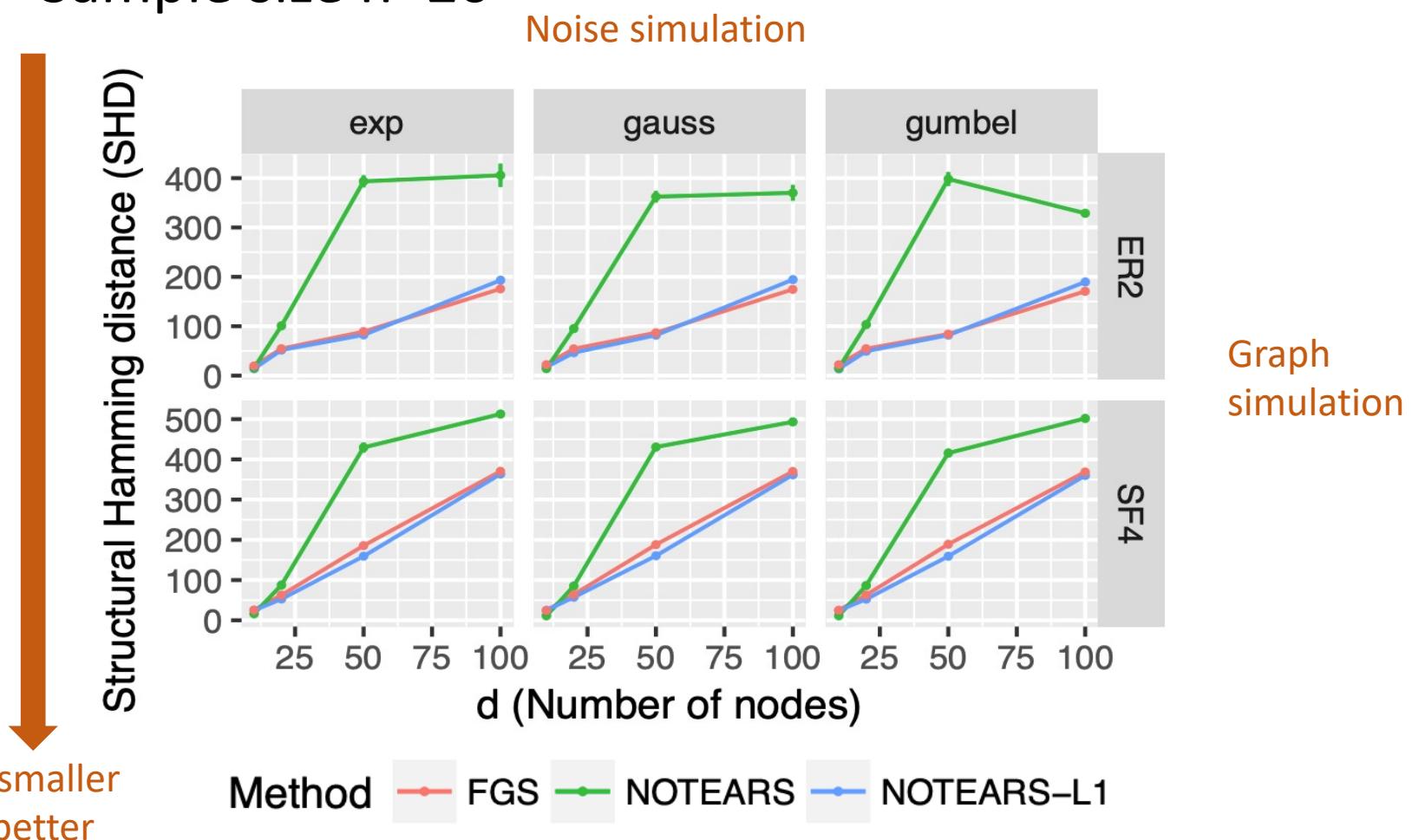
Under different numbers of nodes and samples

- Sample size n=1000



Under different numbers of nodes and samples

- Sample size n=20



More follow-up work

- Dag-gnn: Dag structure learning with graph neural networks, Yu et al., ICML 2018
- Scaling structural learning with no-bears to infer causal transcriptome networks, World Scientific 2019

More follow-up work

Method	Year	Data	Acycl.	Interv.	Output
CMS [152]	2014	low	-	no	Bi
NO TEARS [267]	2018	low	yes	no	DAG
CGNN [75]	2018	low	yes	no	DAG
Graphite [83]	2019	low/medium	no	no	UG
SAM [122]	2019	low/medium	yes	no	DAG
DAG-GNN [262]	2019	low	yes	no	DAG
GAE [177]	2019	low	yes	no	DAG
NO BEARS [142]	2019	low/medium/high	yes	no	DAG
Meta-Transfer [19]	2019	Bi	yes	yes	Bi
DEAR [214]	2020	high	yes	no	-
CAN [167]	2020	low/medium/high	yes	no	DAG
NO FEARS [251]	2020	low	yes	no	DAG
GOLEM [176]	2020	low	yes	no	DAG
ABIC [20]	2020	low	yes	no	ADMG/PAG
DYNOTEARNS [178]	2020	low	yes	no	SVAR
SDI [124]	2020	low	yes	yes	DAG
AEQ [64]	2020	Bi	-	no	direction
RL-BIC [272]	2020	low	yes	no	DAG
CRN [125]	2020	low	yes	yes	DAG
ACD [151]	2020	low	Granger	no	time-series DAG
V-CDN [145]	2020	high	Granger	no	time-series DAG
CASTLE (reg.) [138]	2020	low/medium	yes	no	DAG
GrandDAG [139]	2020	low	yes	no	DAG
MaskedNN [175]	2020	low	yes	no	DAG
CausalVAE [257]	2020	high	yes	yes	DAG
CAREFL [126]	2020	low	yes	no	DAG / Bi
Varando [244]	2020	low	yes	no	DAG
NO TEARS+ [268]	2020	low	yes	no	DAG
ICL [250]	2020	low	yes	no	DAG
LEAST [271]	2020	low/medium/high	yes	no	DAG

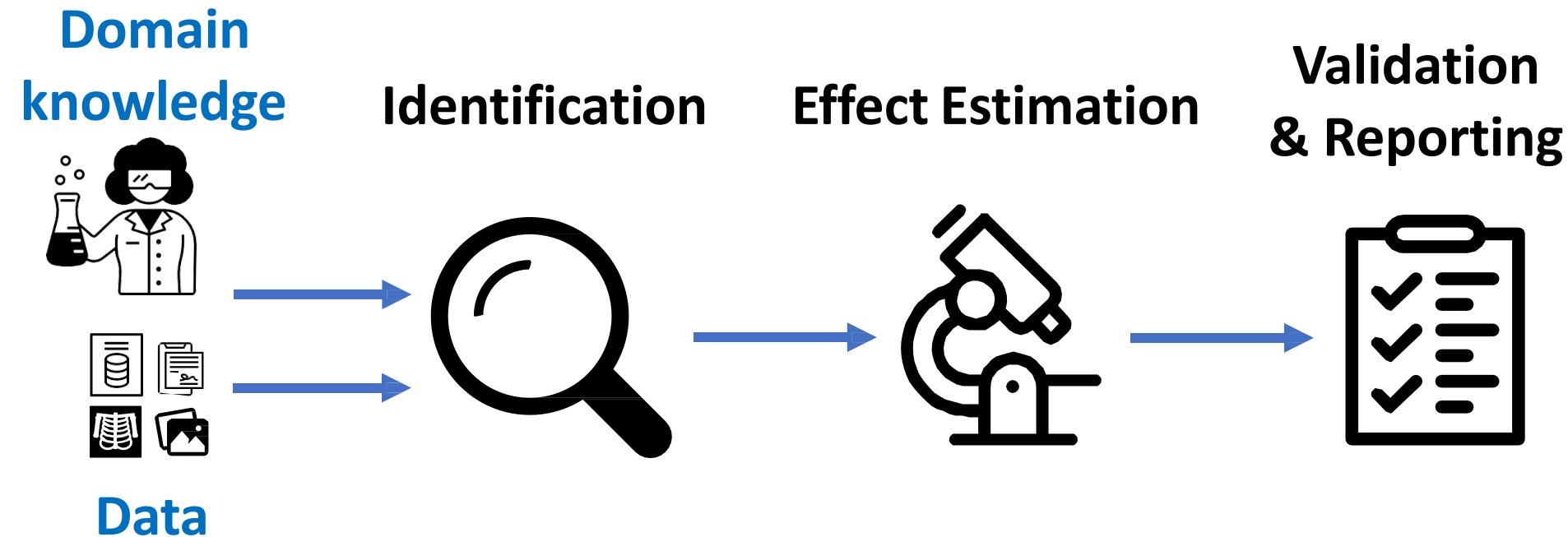
Data

- Low: <100 nodes
- Medium: >100
- High: >10,000

Outline

- Causal discovery (causal structural learning) through neural network
- Causal discovery with large language model

Getting domain knowledge right is difficult

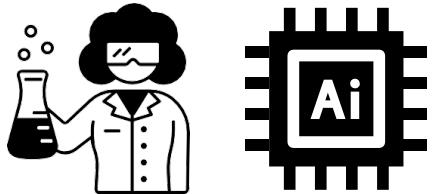


[\[2011.04216\] DoWhy: An End-to-End Library for Causal Inference \(arxiv.org\)](#)

[\[2305.06850\] A Causal Roadmap for Generating High-Quality Real-World Evidence \(arxiv.org\)](#)

LLM reduces burden on human domain expert

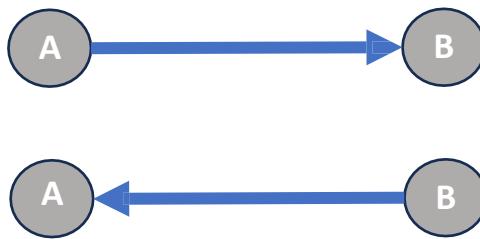
Domain
knowledge



Tasks

Pairwise relationships

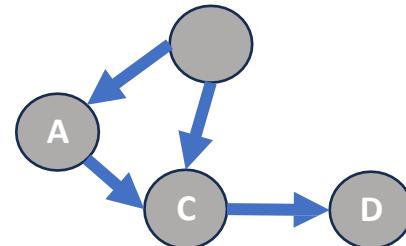
Given a pair of variables (A,B), decide whether A causes B or B causes A?



Full graph recovery

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

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



Benefit brought by LLM

- LLMs can infer causal structure by reasoning on **metadata** associated with the variables
 - E.g., the name of the variable and the problem context
 - Similar to how **domain experts** formalize their knowledge in a graph
- To differentiate from the existing covariance-based causal discovery, we call the LLMs capability as **knowledge-based** causal graph generation.

Applying LLMs to pairwise causal recovery

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.

Applying LLMs to pairwise causal recovery

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>.

Tübingen Benchmark for Pairwise relationships

- 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 discovery algorithms achieve 70-80% accuracy, Best is 83% [Mosaic, Wu & Fukumizu 2020].

Results: LLMs recall 97% correctly

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
ada	0.50	0.50
text-ada-001	0.49	0.50
babbage	0.51	0.50
text-babbage-001	0.50	0.50
curie	0.51	0.52
text-curie-001	0.50	0.50
davinci	0.48	0.47
text-davinci-001	0.50	0.50
text-davinci-002	0.79	0.79
text-davinci-003	0.82	0.83
LMPrior (Choi et al., 2022)	0.83	-
gpt-3.5-turbo	0.81	0.83
gpt-3.5-turbo (causal agent)	0.86	0.87
gpt-3.5-turbo (single prompt)	0.89	0.92
gpt-4 (single prompt)	0.96	0.97



Data-driven causal discovery



Knowledge-based causal recovery is competitive with *or much better than* data-driven approaches

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.	Model	Accuracy
Right L1 Radiculopathy	Right adductor tendonitis	→	ada	40.1
Pharyngeal discomfort	Right C3 Radiculopathy	←	text-ada-001	50.0
Right L5 Radiculopathy	Lumbago	→	babbage	50.0
Left PTA	Left L4 Radiculopathy	←	text-babbage-001	50.9
Left T3 Radiculopathy	Toracal dysfunction	→	curie	50.0
DLS L5-S1	Right S1 Radiculopathy	→	text-curie-001	50.0
Left C3 Radiculopathy	DLS C2-C3	←	davinci	38.4
Left C7 Radiculopathy	Left medial elbow problem	→	text-davinci-001	50.0
Right Ischias	Right L5 Radiculopathy	←	text-davinci-002	51.7
Right Morton trouble	Right L5 Radiculopathy	←	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

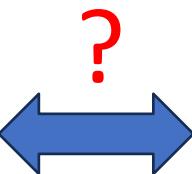
Table 3: Example cause-effect pairs from the Neuropathic pain diagnosis benchmark. ‘Dir.’ refers to the ground-truth causal direction between the variables.



LLM can make mistakes

- LLM can make such errors, not surprising, since it is not explicitly optimizing with prior knowledge or causal mechanisms.
- However, it is remarkable that such errors occur infrequently: The LLM made only 12 errors in the dataset of 108 pairs.

Even for mistakes, still, there is hope

Ozone  Radiation



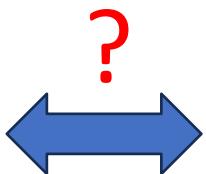
Correct answer: Radiation  Ozone

LLM initially gives a wrong answer

It is because the LLM interpreted the question as about ozone in the stratospheric layer, whereas the question was about ground-level in a city



Even for mistakes, still, there is hope

Ozone  Radiation



Correct answer: Radiation  Ozone

LLM initially gives a wrong answer

After adding this context, LLM generates correct answer



Plausible answer; benchmark says is wrong

Input Pair: L5 Radiculopathy \leftrightarrow Obesity ?

- 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 because obesity can put extra pressure on the spine and nerves, 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>

- Benchmark Answer:
- 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]

Construct Validity: Is Benchmark Memorized?



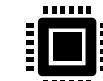
SYSTEM

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:

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

The diagram shows two probability terms stacked vertically: $P(Y|D)$ and $P(D)$. A red curly brace is positioned under the first term, and a blue curly brace is positioned under the second term. Below the red brace, the text "Likelihood that knowledge can be transformed to answer question Y" is written in red. Below the blue brace, the text "Likelihood that knowledge D is known by LLM" is written in blue.

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

Compared with NOTEARS

Algorithm	NHD	No. of Predicted Edges	Baseline NHD	Ratio
NOTEARS	0.22	10	0.32	0.69
DAG-GNN	0.37	20	0.44	0.83
gpt-3.5-turbo	0.21	21	0.38	0.55
gpt-4	0.14	25	0.48	0.28

Comparison on a novel Alzheimer dataset which cannot be used in training

Prompt strategies

How will you design a prompt for LLM for a causal question?

Basic prompt

- Give the question as the input string directly
- Different forms can be used: binary classification, multi-choice, open-question,...

Prompt:

You will be presented with a causal graph in the following form: A causes B, B causes D, B causes C, and B causes E.

Question: Does D serve as the parent node of E?

Answer (Yes or No ?):

Adversarial prompt

- Essential for comprehending the inherent risks and robustness of LLM
- Repeatedly ask the question, ignore or doubt previous answer

Adversarial-ignore

Prompt:

You will be presented with a causal graph in the following form:
A causes B, B causes D, B causes C, and B causes E.
Question: Does D serve as the parent node of E?
Answer (Yes or No ?): *model output 1*

Your former answer is: *model output 1*

Ignore the former answer and answer the question again.
You will be presented with a causal graph in the following form:
A causes B, B causes D, B causes C, and B causes E.
Question: Does D serve as the parent node of E?
Answer (Yes or No ?): *model output 2*

Adversarial-doubt

Prompt:

You will be presented with a causal graph in the following form:
A causes B, B causes D, B causes C, and B causes E.
Question: Does D serve as the parent node of E?
Answer (Yes or No ?): *model output 1*

Your former answer is: *model output 1*

Your former answer is wrong, please answer the question again.
You will be presented with a causal graph in the following form:
A causes B, B causes D, B causes C, and B causes E.
Question: Does D serve as the parent node of E?
Answer (Yes or No ?): *model output 2*

Chain-of-Thought (CoT)

- Decompose complex problems and perform intermediate reasoning steps to enhance their performance.

o-shot CoT

Prompt:

You will be presented with a causal graph in the following form: A causes B, B causes D, B causes C, and B causes E.

Question: Does D serve as the parent node of E? **Let's think step by step.**

Answer (Yes or No ?):

Manual CoT

Prompt:

Here are eight examples of causal attribution with chain of thought.

You will be presented with a causal graph in the following form: A causes D, A causes B, C causes E, and D causes E.

Question: Does D serve as the parent node of E?

Answer (Yes or No ?): D causes E, so D is the parent node of E. Thus the answer is Yes.

You will be presented with a causal graph in the following form: A causes B, A causes I, A causes C, B causes H, B causes F, B causes K, B causes I, B causes J, C causes J, C causes D, D causes F, D causes E, D causes G, D causes H, F causes J, G causes I, H causes K, and I causes J.

Question: Does J serve as the parent node of K?

Answer (Yes or No ?):

]} 8x

In-Context Learning

- A model learns new tasks through a set of examples “learn from analogy”
- In ICL, a prompt contains a causal task description and several input-output pairs, showing how the task inputs can be answered.

0/1/3-shot IcL

Prompt:

Determine whether or not a variable can serve as the parent of another variable in a given causal graph.

You will be presented with a causal graph in the following form: A causes D, A causes B, C causes E, and D causes E.

Question: Does D serve as the parent node of E?

Answer (Yes or No ?): Yes

0/1/3x

You will be presented with a causal graph in the following form: A causes B, A causes I, A causes C, B causes H, B causes F, B causes K, B causes I, B causes J, C causes J, C causes D, D causes F, D causes E, D causes G, D causes H, F causes J, G causes I, H causes K, and I causes J.

Question: Does J serve as the parent node of K?

Answer (Yes or No ?):

Explicit Function

- Recent studies have elucidated that language models may have emotional awareness analogous to humans
- Use encouraging and positive language within prompts

Prompt:

You are a helpful assistant for causal attribution (parent node).

You will be presented with a causal graph in the following form: A causes B, B causes D, B causes C, and B causes E.

Question: Does D serve as the parent node of E?

Answer (Yes or No ?):

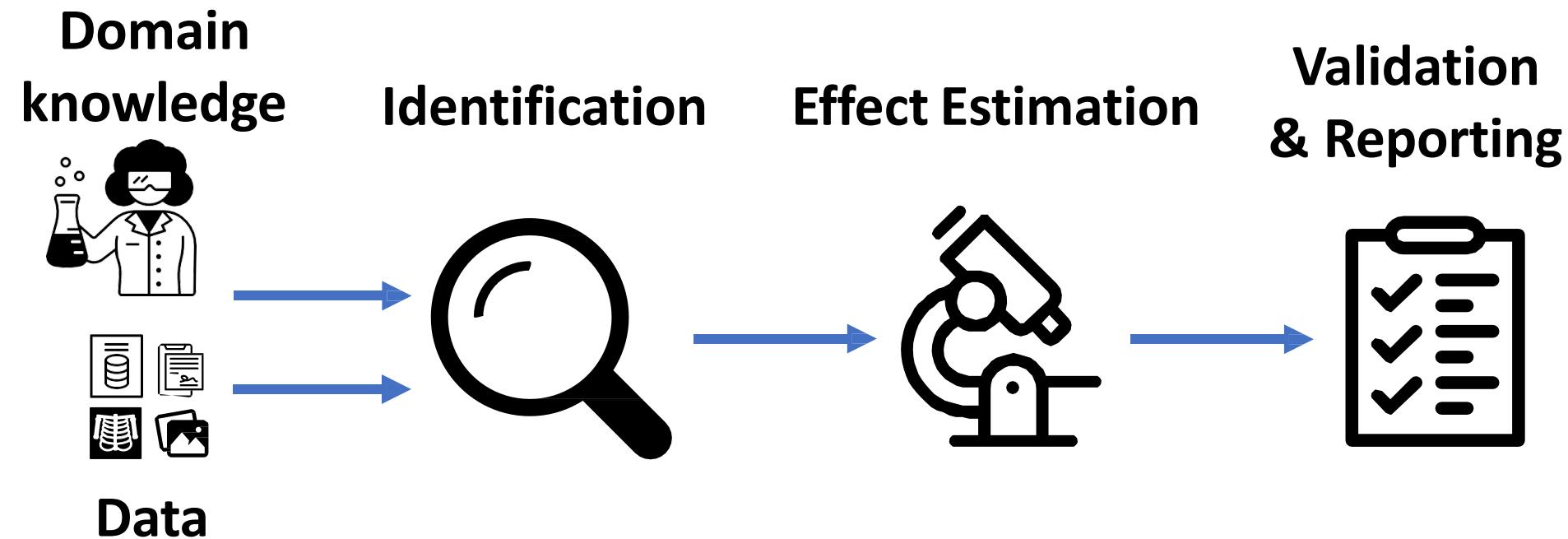
Comparison of performance

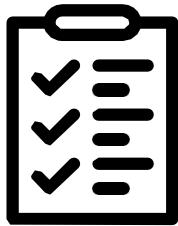
Model	CEPairs		E-CARE		COPA		CALM-CA Binary	Neuro Pain Choice
	Binary	Choice	Binary	Choice	Binary	Binary		
ada	0.50	0.48	0.49	0.49	0.49	0.57	0.40	
text-ada-001	0.49	0.49	0.33	0.50	0.35	0.48	0.50	
Llama2 (7B)	-	0.53	0.50	0.41	0.35	0.32	-	
Llama2 (13B)	-	0.52	0.50	0.44	0.36	0.42	-	
Llama2 (70B)	-	0.52	0.44	0.50	0.45	0.49	-	
babbage	0.51	0.49	0.36	0.49	0.40	0.58	0.50	
text-babbage-001	0.50	0.50	0.50	0.49	0.50	0.56	0.51	
curie	0.51	0.50	0.50	0.50	0.50	0.58	0.50	
text-curie-001	0.50	0.50	0.50	0.51	0.50	0.58	0.50	
davinci	0.48	0.50	0.49	0.50	0.51	0.58	0.38	
text-davinci-001	0.50	0.50	0.50	0.50	0.50	0.52	0.50	
text-davinci-002	0.79	0.66	0.64	0.80	0.67	0.69	0.52	
text-davinci-003	0.82	0.77	0.66	0.90	0.77	0.80	0.55	
GPT-3.5-Turbo	0.81	0.80	0.66	0.92	0.66	0.72	0.71	
GPT-4	-	0.74	0.68	0.90	0.80	0.93	0.78	
GPT-4 (0-shot ICL)	-	0.83	0.71	0.97	0.78	0.90	-	
GPT-4 (1-shot ICL)	-	0.81	0.70	0.93	0.76	0.90	-	
GPT-4 (3-shot ICL)	-	0.71	0.70	0.80	0.81	0.91	-	
GPT-4 (0-shot CoT)	-	0.77	0.68	0.91	0.79	0.92	-	
GPT-4 (Manual CoT)	-	0.79	0.73	0.97	0.82	0.95	-	
GPT-4 (EF)	-	0.83	0.71	0.98	0.80	0.92	0.84	

Takeaways

- LLMs enable knowledge-based causal discovery or recovery
 - Strong performance for pairwise causal relationships
 - Across multiple datasets in varied domains incl. medicine and climate science
 - Full graph recovery poses additional challenges (e.g., direct/indirect causes)
- GPT-3.5, GPT-4 have memorized Tübingen benchmark
 - Valid measurement of LLM *transformation of knowledge into a causal answer*
 - Does not give likelihood that arbitrary relationship has been memorized

How else might LLMs help?





Validation LLMs suggest negative controls

USER: In a causal observational study of vaccine efficacy, we are matching people who have voluntarily gotten a vaccine with individuals who have not yet done so. We are matching on attributes intended to capture all confounding effects. To help validate our analysis, what are some examples of negative controls where we might expect to see zero treatment effect?

GPT first misunderstands and gives examples of factors that not caused by the treatment:

- hair color, blood type, ...

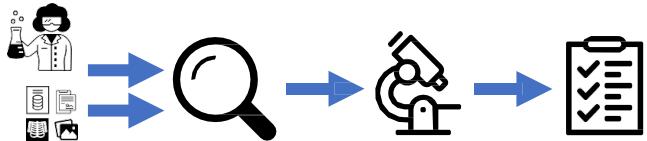
USER: In what subpopulations would we expect to see zero treatment effect on disease prevention? (for the disease being vaccinated against)

GPT then gives reasonable answers:

- Individuals with pre-existing immunity
- Individuals with specific immune deficiencies
- Nonresponders
- Individuals vaccinated post-infection

If we add that we have longitudinal infection data GPT identifies time-bound negative controls

- Pre-vaccination period
- Short time window post-vaccination



Adding to the Open Source Ecosystem for Causality

PyWhy-LLM

Python library for using LLMs in causal analysis process

Integrates with PyWhy libraries
(DoWhy, EconML, ...)

Work-in-progress

<https://pywhy.org/>

<https://github.com/py-why/pywhy-llm/>

Latent confounders

```
In [9]:  
variables = ["ice cream sales", "temperature", "cavities"]  
latents = modeler.suggest_confounders(variables, treatment="ice cream sales", outcome = "shark attacks")  
  
print(latents)
```

system You are a helpful assistant for causal reasoning.

What latent confounding factors might influence the relationship between `ice cream sales` and `shark attacks`?

user We have already considered the following factors `['ice cream sales', 'temperature', 'cavities']`. Please do not repeat them.

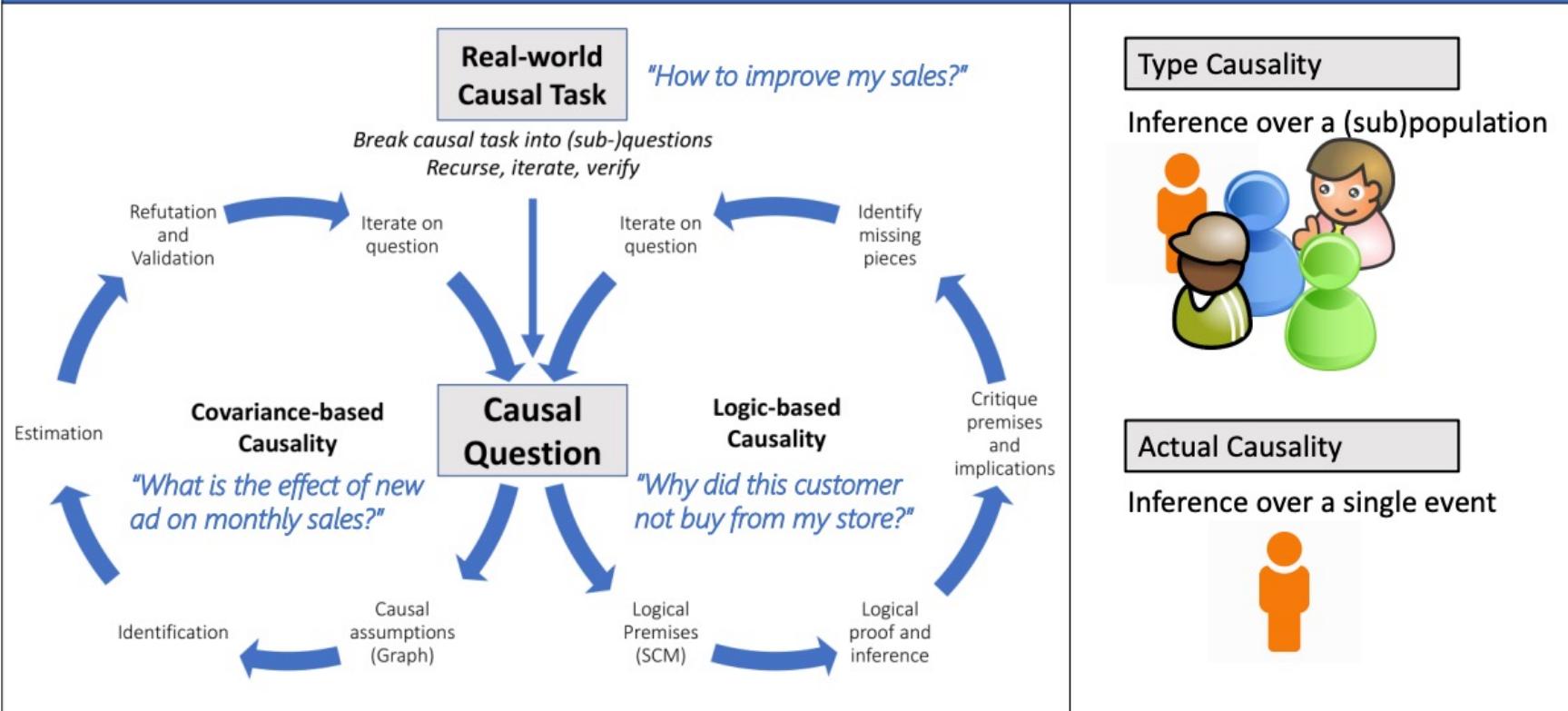
List the confounding factors between `ice cream sales` and `shark attacks` enclosing the name of each factor in `<conf>` `</conf>` tags.

assistant `<conf>Beach Attendance</conf>, <conf>Season of the Year</conf>, <conf>Water Temperature</conf>, <conf>Public Holidays</conf>, <conf>Availability of Ice Cream Vendors</conf>, <conf>Shark Population</conf>, <conf>Swimming Conditions</conf>, <conf>Tourist Season</conf>.`

```
[ 'Beach Attendance', 'Season of the Year', 'Water Temperature', 'Public Holidays', 'Availability of Ice Cream Vendors', 'Shark Population', 'Swimming Conditions', 'Tourist Season' ]
```

Many Kinds of Causality

Different tasks: Graph Discovery, Effect inference, Attribution, Prediction



Event/Actual Causality and Causal Judgments

Type Causality

Inference over a (sub)population



- Bob has lung cancer and smokes.
Did Bob's smoking cause his cancer?
- A customer saw a newspaper ad and bought toys.
What would have happened if they hadn't seen the ad?
- A doctor washes their hands before surgery.
What would have happened if the Dr hadn't washed their hands?

Actual Causality

Inference over a single event



Causal context is hard to formalize

- **Causal frame:** Candidate causal factors relevant to causal question
- **Necessary causality:** Did cause **need to happen** for outcome to occur?
- **Sufficient causality:** Is cause **alone enough** for outcome to occur?
- **Normality:** Do events line up statistical/social/... norms?
- **Other human factors:** bias towards action, intention, epistemic, ...

Necessary and sufficient causes

- **Necessary causality:**
If an event C does not occur, then the outcome event, E , will not occur.
- **Sufficient causality:**
If an event C occurs then the outcome event, E , will occur.
- **Robust sufficient causality:**
... even if other contributing factors did not occur.

Sufficiency is harder, because we have to determine causal frame

Evaluation Vignettes

Vignette Type	Input Context	Event	Actor	Nec.	Suff.
Overdetermination	Alice (AF) and Bob (BF) each fire a bullet at a window, simultaneously striking the window, shattering it (WS).	window shattering	Alice	No	Yes
Switch	Alice pushes Bob. Therefore, Bob is hit by a truck. Bob dies. Otherwise, Bob would have been hit by a bus, which would have killed him as well.	Bob's death	Alice	No	Yes
Late preemption	Alice (AF) and Bob (BF) each fire a bullet at a window. Alice's bullet hits the window first (AH). The window shatters (WS). Bob's bullet arrives second and does not hit the window (BH).	window shattering	Alice	No	Yes
Vignette Type	Input Context	Event	Actor	Nec.	Suff.
Overdetermination	There is a fire in the chemistry lab. A can of water would douse the fire. Agents X and Y both spray a can of water each, dousing the fire.	fire doused	being Agent X	No	Yes
Switch	Reagent X is added to a mixture, which leads to an explosion and kills Sam. Otherwise, Reagent Y in Sam's pocket would have infected him and killed him as well.	Sam's death	Reagent X	No	Yes
Late preemption	Any of Reagent X or Reagent Y can be added to a mixture to convert it into a crystal. Reagent X is added first and the mixture turns to crystal. Reagent Y is added later and but does not mix since the crystal is already formed.	crystal formation	Reagent X	No	Yes

Necessary and Sufficient - Results

Vignette Type	Necessary	Sufficient
<i>gpt-3.5-turbo</i>		
Overdetermination	✓, ✓	X, ✓
Switch	X, X	✓, X
Late preemption	X	X
Early preemption	X, ✓, X	X, X, ✓
Double preemption	✓	✓
Bogus preemption	✓	X,
Short circuit	X	X
Miscellaneous	✓, ✓, ✓, X	✓, ✓, X, ✓
Total Accuracy	46.6%	46.6%
<i>gpt-4</i>		
Overdetermination	✓, ✓	✓, ✓
Switch	✓, ✓	✓, ✓
Late preemption	✓	✓
Early preemption	✓, ✓, ✓	✓,
Double preemption	✓	X
Bogus preemption	✓	✓
Short circuit	X	X
Miscellaneous	✓, ✓, ✓	✓, ✓, ✓, ✓
Total Accuracy	86.6%	86.6%

Vignette Type	Necessary	Sufficient
<i>gpt-3.5-turbo</i>		
Overdetermination	✓, ✓	X, ✓
Switch	X, ✓	✓, X
Late preemption	X	✓
Early preemption	✓, X	X, X
Double preemption	✓	✓
Bogus preemption	✓	X
Short circuit	X	X
Miscellaneous	✓, ✓, ✓, X	✓, X, X, ✓
Total Accuracy	64.2%	42.8%
<i>gpt-4</i>		
Overdetermination	✓, ✓	✓, ✓
Switch	✓, ✓	X, ✓
Late preemption	✓	✓
Early preemption	✓, ✓	X
Double preemption	✓	✓
Bogus preemption	✓	✓
Short circuit	✓	✓
Miscellaneous	✓, ✓, ✓	✓, ✓, X, ✓
Total Accuracy	92.8%	78.5%

GPT-3.5

GPT-4

Takeaways on text-based reasoning

GPT-4 understands scenarios, identifies necessity and sufficiency

- Not possible before

See paper for additional experiments

- Counterfactual reasoning: GPT-4: 92.44% accuracy
 - Only 6% below human baseline
- Normality: TL;DR: 70% accuracy with GPT-4

Wrapping Up

What's new with causality now

LLMs provide...

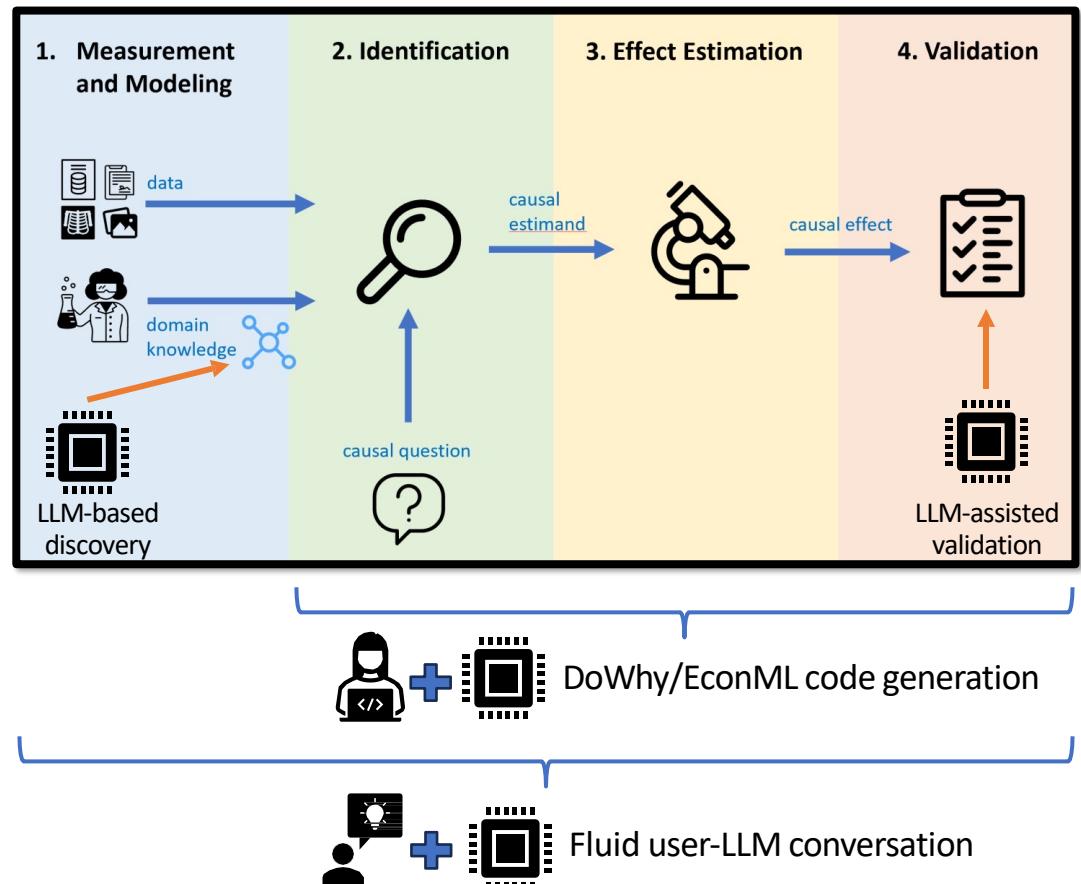
- Domain knowledge only available via human experts before
 - Provide when explicitly asked
 - Also implicitly, e.g., in background knowledge for actual causality vignettes.
- New capability to extract key primitives of text-based reasoning
 - (necessity, sufficiency, normality, etc.)
 - Possibility of system to analyze actual causality for practical scenarios

What's not changing with causality

- Need for rigorous, well-documented, and verifiable analyses
 - Especially for high-risk and high-value tasks
 - Must ensure correctness for decision making

Implications for Practitioners

- Augmenting human expertise with LLMs
 - Assisting in graph creation, validation, and robustness checks
 - Case study: LLM-assisted identification of negative controls
- LLMs can enable end-to-end causal tools
 - Case study: asking LLM to generate DoWhy and EconML analysis code
- LLM as a fluid conversational interface



Conclusion: A New Frontier for Causality

- Human domain knowledge critical for causal analysis
- LLMs mimic this capability
 - Building causal assumptions and arguments, counterfactual inference, and systematization of necessity, sufficiency, ...
- Implications for practice:
Reduce burden on human domain expert
- New research questions:
Combining data-driven and text-based analysis?

<https://pywhy.org/>

<https://arxiv.org/abs/2305.00050>

References & Reading Materials

- Zheng X, Aragam B, Ravikumar P K, et al. Dags with no tears: Continuous optimization for structure learning[J]. Advances in neural information processing systems, 2018, 31.
 - <https://proceedings.neurips.cc/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf>
- Kıcıman, Emre, et al. "Causal reasoning and large language models: Opening a new frontier for causality." arXiv preprint arXiv:2305.00050 (2023)
- Ma, Jing. "Causal Inference with Large Language Model: A Survey." arXiv preprint arXiv:2409.09822 (2024).

Thank you!
Questions?