2025/07/24

Causal Machine Learning in Medical AI

DS with Alumni Seminar

연세대학교 통계데이터사이언스학과
DSL 12기 김민규

Contents

Introduction to Causality

Introduction to Generative Models

Theory of Causal Graph Discovery

Generative Models for Causal Graph Discovery

Generative Models based Causal Graph Discovery for Medical Domain

Additional Topics
- Generative Models for Causal Inference
- Generative Models for Causal Change Point Detection

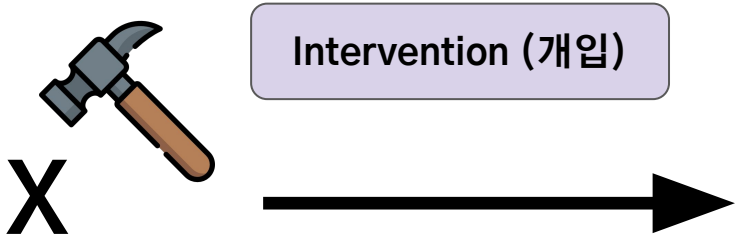
Presenter

- 이름: 김민규 [CV] [Personal Website] [LinkedIn]
- 소속: 연세대학교 통계데이터사이언스학과 통합 1학기
  - 2020.03 ~ 2025.02 연세대학교 상경대학 응용통계학과
  - 2025.03 ~ 현재 연세대학교 일반대학원 통계데이터사이언스학과 (MLAI https://mlai.yonsei.ac.kr/)
- 기타 경험
  - MLAI 학부 연구생 (2024.07 ~ 2025.02)
  - Data Science Lab 12기 학술부장 (2024.07 ~ 2025.05)
  - Yonsei Artificial Intelligence 14기 학회원 (2024.07 ~ 2025.06)

Introduction to Causality

The Goal of Treatment Effect Evaluation

우리는 유효한 '**Causal Effect**'를 추정하고 싶다!!

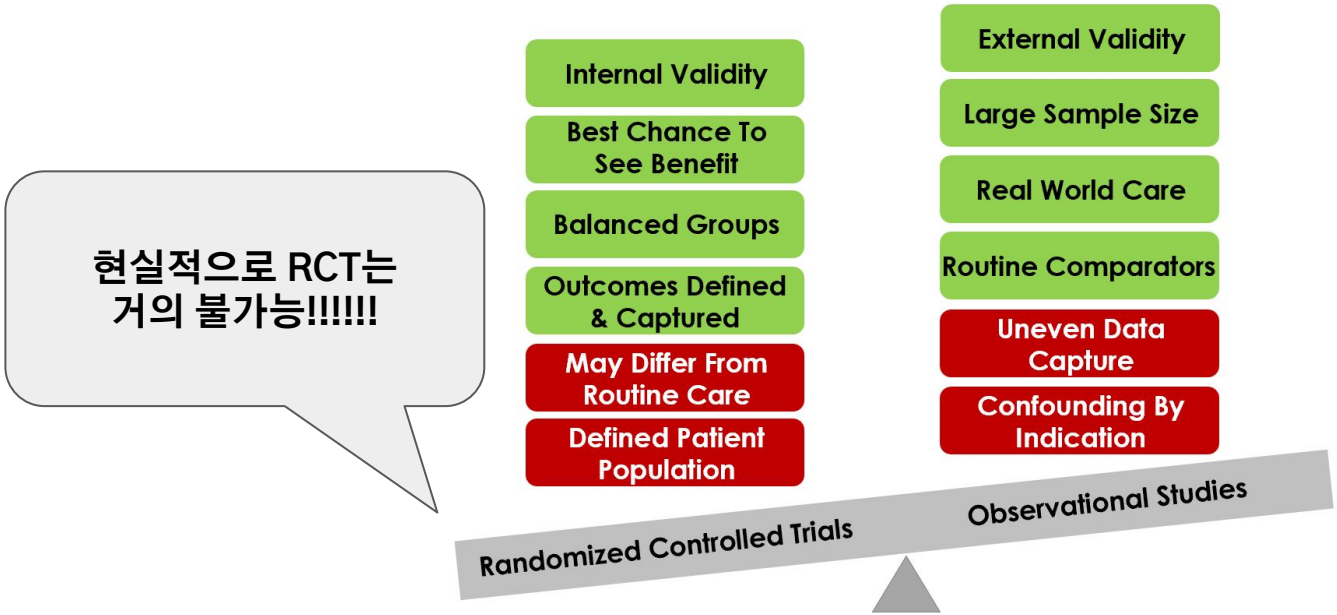


임의의 Treatment
(약 복용
환율 증감
출산 장려 정책 수행
새로운 수업방식)

다양한 Outcomes
(혈압 변화
경제성장률
출산율
수업 만족도)

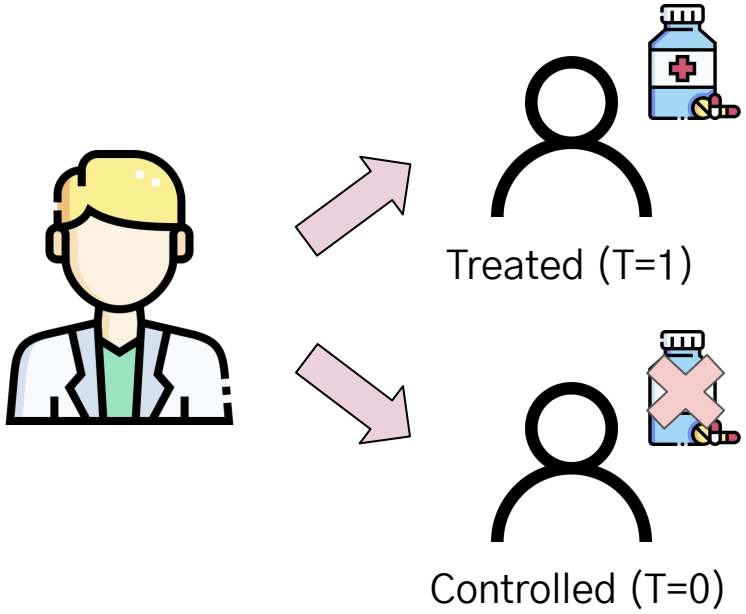
Randomized Controlled Trials vs. Observational Studies

RCT: 우리가 정말 관심있어 하는 treatment를 제외하고는, 모두 랜덤하게 배정된 상태
(예: 실제로 약을 복용한 사람들의 분포와 그렇지 않은 사람의 분포는 동일하다고 가정한다!)
⇒ Causal Effect를 추정할 수 있는 가장 완벽한 방법, 그러나…



Motivating Example

- 혈압약 복용 여부에 따른 **Average Treatment Effect (ATE)** 추정



Blood Pressure (Y | T=1)

수축기 혈압	132	128	140	…

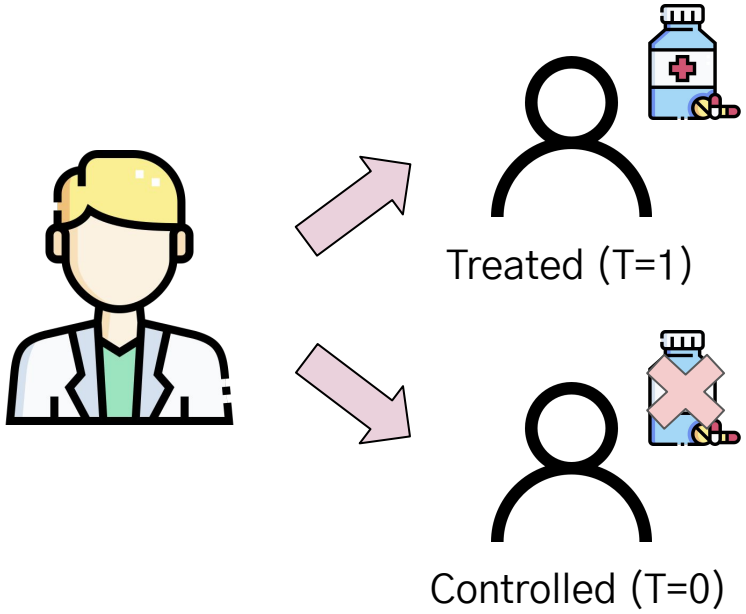
Blood Pressure (Y | T=0)

수축기 혈압	138	142	135	…

Q) Why ATE ≠ E(Y | T=1) − E(Y | T=0)?

Motivating Example

- 혈압약 복용 여부에 따른 **Average Treatment Effect (ATE)** 추정



Problem 1:
Inaccessibility to Counterfactual Result

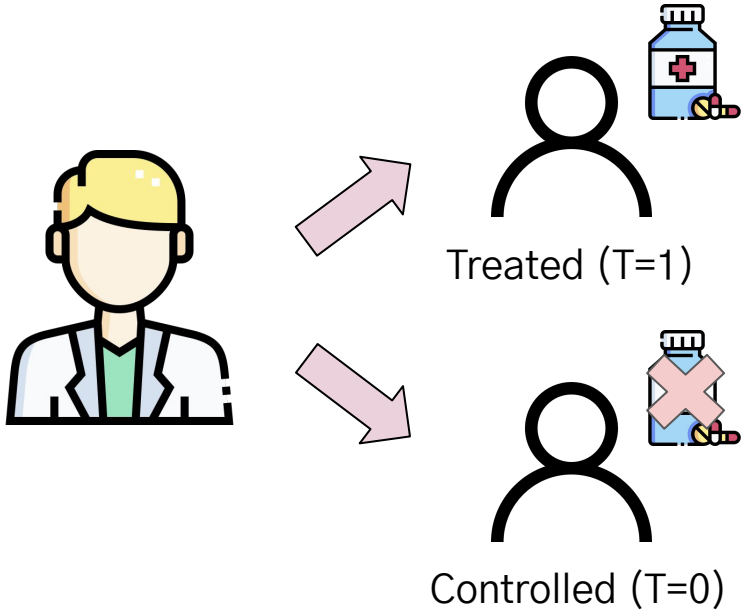
Blood Pressure (Y | T=1)	환자 1	환자 2	환자 3	…
수축기 혈압	132	?	?	…

Blood Pressure (Y | T=0)	환자 1	환자 2	환자 3	…
수축기 혈압	?	142	135	…

Motivating Example

- 혈압약 복용 여부에 따른 **Average Treatment Effect (ATE)** 추정

Problem 2:
Covariate Imbalance



Blood Pressure (Y | T=1)

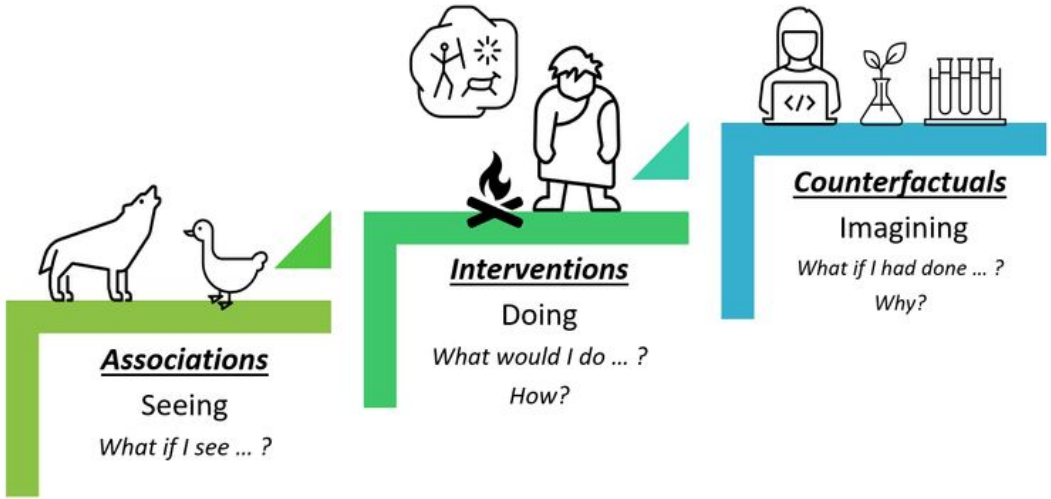
수축기 혈압	132	128	140	…
비만 (X)	O	O	O	…

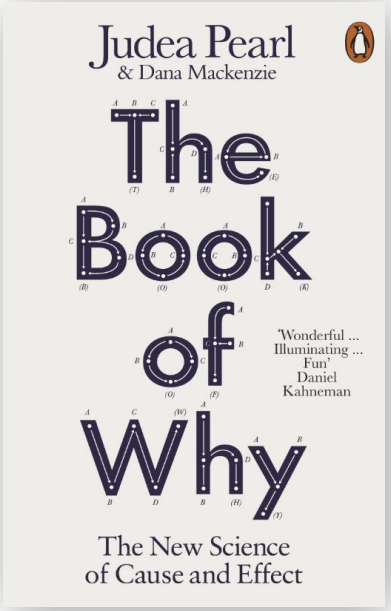
Blood Pressure (Y | T=0)

수축기 혈압	138	142	135	…
비만 (X)	X	O	X	…

Three Rungs of the *Ladder of Causation*

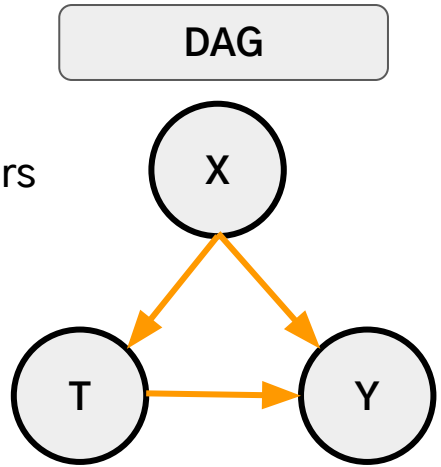
- Associational (Seeing) / Interventional (Doing) / **Counterfactual (Imagining)**





Directed Acyclic Graphs (DAG)의 역할

- 통계적 방법론은 상관관계를 파악할 수 있으나, 왜?라는 질문에 답변할 수 없다.
- 반면, Causal inference는 “What causes what?”에 대해 답한다!
- Directed Acyclic Graphs (DAGs) 이러한 인과 관계를 표현하는 주요한 도구!!
- Why use DAGs?
  1. 인과 추론을 위해 사용하는 가정을 시각화 가능
  2. 인과 관계와 상관 관계를 식별 가능: confounders / mediators / colliders
  3. 더 나은 분석 접근 가능



DAG Construction: From Manual to AI-Assisted Approach

DAG를 그리는 하나의 방법? LLM!

- Manual Approach (Current Practice)
  1. DAGs are typically drawn manually based on subject expertise.
  2. Tools like DAGitty or drawing software are commonly used.
  3. Limitations
     - 🌀 Time-consuming
     - 🌀 Requires deep prior knowledge
     - 🌀 Prone to missing latent variables
- Emerging Approach: Using LLMs to Draw DAGs

➡ **NEXT: Generative Models, Theory of DAG, Empirical Example of using LLM to draw DAG**

Introduction to Generative Models

Which image is real?

A B C

source: http://introtodeeplearning.com/

Which text is real?

Data Science Lab(DSL)은 데이터 과학에 관심 있는 학부생들이 주도적으로 참여하는 학술 동아리로, 정규 프로젝트, 탐색적 데이터 분석(EDA) 세션, 스터디와 세미나 등을 통해 실전 중심의 데이터 분석 및 머신러닝 역량을 키우는 것을 목표로 한다. 학회원들은 팀을 이루어 추천 시스템, 자연어처리, 컴퓨터비전, 강화학습 등 다양한 주제로 프로젝트를 수행하며, …

A

야. 내가 이미 이거 듣고 왔거든? 니네 진짜 다 쓸데없는 거 치워. 나와 빨리. 야. 야. 연고대. 나와. 나와 연고대. 아 진짜 너 어디야? 나와. 나와 이씨. 야. 서울대란 말이지. 관악산 자락 프라임 입지로써, 대학 본고사 도입 이후 전국 수석들 싹 다 빨아들인 학교가 서울대야. 내기할래? 내기할래? 야. 잘 들어봐. 니네 학교 예산, 여기 우리 본관 리모델링 비용이면 다 커버돼.

B

연세대학교는 신촌 한복판에 위치한 캠퍼스로, 학교보다 주변 상권이 더 유명하단 얘기도 종종 나옴. 학생들 분위기 전반적으로 자유로운 편이고, 인싸부터 조용한 연구실파까지 다양하게 공존함. 동아리, 학회, 공연, 축제 같은 문화 활동도 활발한 편이라 뭔가 할 거 찾으려면 끝도 없음. 공부는 학과 따라 강도가 갈리는데, 일부 과는 진짜 빡셈. 연대 특유의 느긋한 캠퍼스 감성과 동시에, 자기 일에 몰두하는 분위기 둘 다 공존하는 독특한 학교임.

C

All images and texts were generated by a generative model.

Introduction to Generative Models

Supervised Learning

- Data: (x,y)
- Goal: Learn function to map x -> y

Unsupervised Learning

- Data: (x)
- Goal: Learn the hidden or underlying structure of the data

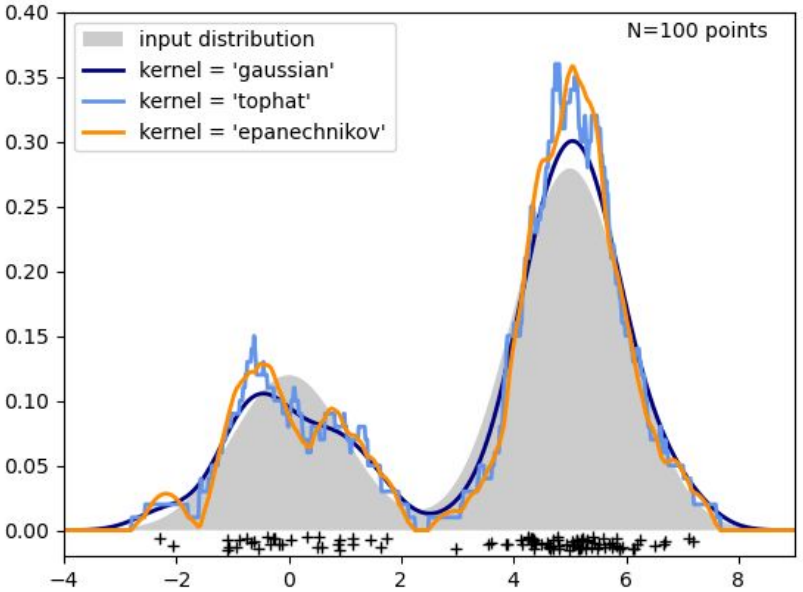
$P_{data}(x) \cong P_{\textbf{model}}(x)$
$P_{data}(x,y) \cong P_{\textbf{model}}(x,y)$

Generative Models

- Data: (x,y) or (x)
- Goal: Learn a model that represents the density with some observed samples
  - Density Estimation
  - Sample Generations

Generative Models

- Data: (x,y) or (x)
- Goal: Learn a model that represents the density with some observed samples
  - **Density Estimation**
  - Sample Generation



Generative Models

- Data: (x,y) or (x)
- Goal: Learn a model that represents the density with some observed samples
  - Density Estimation
  - **Sample Generation**

Training data ~ $P_{data}(x)$

Generated data ~ $P_{model}(x)$

OpenAI

Training of LLMs consists of multiple stages, but the most important pretraining phase is based on **next token prediction.**

A generative model is a type of artificial intelligence model that learns the distribution of data and can generate new data.

A _____
A generative _____
A generative model _____
…

Next token prediction?

ILYA "So let's suppose that you have consumed a **mystery novel,** and you are at the last page of the novel. And somewhere on the **last page**, there is a sentence where the detective is about to **announce the identity of whoever committed the crime**. And then there is this one word, which is the name of whoever did it. At that point, the system will make a guess of the next word. If the system is really, really good, it will have a good guess about that name; it might narrow it down to three choices or two choices. And if the neural network has paid really close attention (well, certainly that's how it works for people), if you pay really close attention in the mystery novel, and you think about it a lot, you can guess who did it at the end. So this suggests that **if a neural network could do a really good job of predicting the next word**, including this word, then **it would suggest that it's understood something very significant about the novel.**"

LLMs Are Also Generative Models

LLMs are generative models that learn the data distribution through next-token prediction and generate new text by sampling from the learned model distribution.

- Translation
  - **English: This sandwich is very tasty.**
  - **Spanish:** _____ _____ _____

- If you want to see a poem by Robert Frost about AI
  - **This is a poem written by Robert Frost about the perils of artificial intelligence.**
  - _____ _____ _____
  - _____ _____ _____
  - _____ _____ _____

Bold: Prompt

- Advantages of generative models
  - Can effectively handle various data types (images, text, tables, etc.)
  - Capable of obtaining better representations (embedding vectors) from given data, allowing for strong performance even on previously unseen data
  - **Can approximately estimate the population density from sample data**
  - **Can leverage the power of pretrained models (e.g., GPT)**

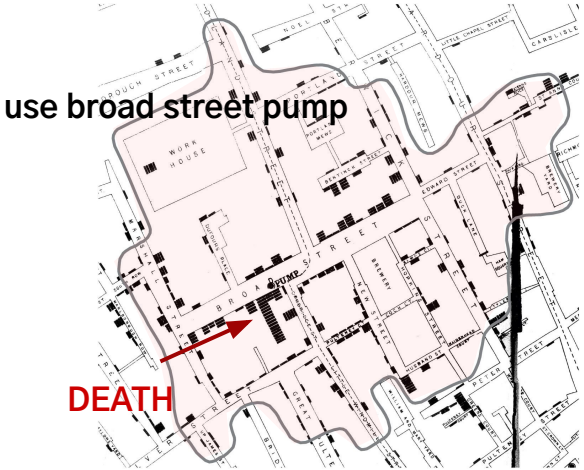
- When generative models are applied to causality
  - **Treatment effect estimation with hidden confounders**
  - **Leveraging the knowledge of LLMs for more accurate causal graph discovery**
  - Multimodal causality
  - Heterogeneous treatment effect (HTE) estimation
  - …

Theory of Causal Graph Discovery

Causal Graph

- **Definition:** a graphical structure that visually represents **causal relationships** between variables

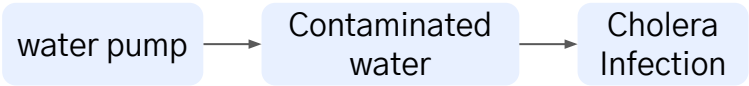
Q. Why Causal Graph is important? It supports effective decision making!



<Map made by John snow>

[source: https://johnsnow.matrix.msu.edu/book_images12.php/]

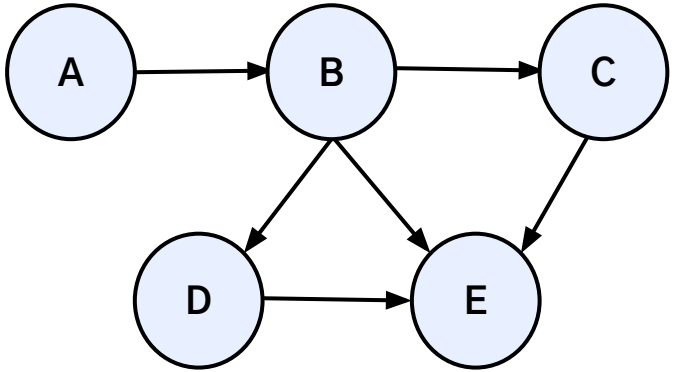
- When cholera broke out in London, patient addresses were mapped
- This revealed a concentration of cases around the **Broad Street pump**
- Discovery causal path



- Removing the pump handle ➡ cholera cases declined immediately

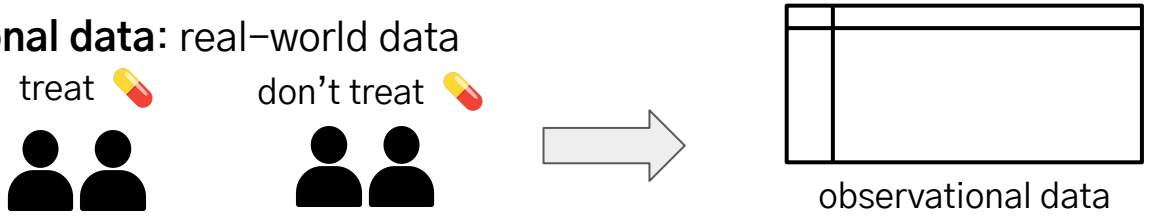
Directed Acyclic Graph (DAG)

- **Causal graph** usually represents as a **Directed Acyclic Graph (DAG)**
- **It consists of**
  - Set of Nodes (ex. A,B,C,D,E)
  - Set of Edges (ex. A→ B)
- Directed: each edge has a defined direction
- Acyclic: there are no loops (i.e., “cycles”) in the graph



Causal Discovery

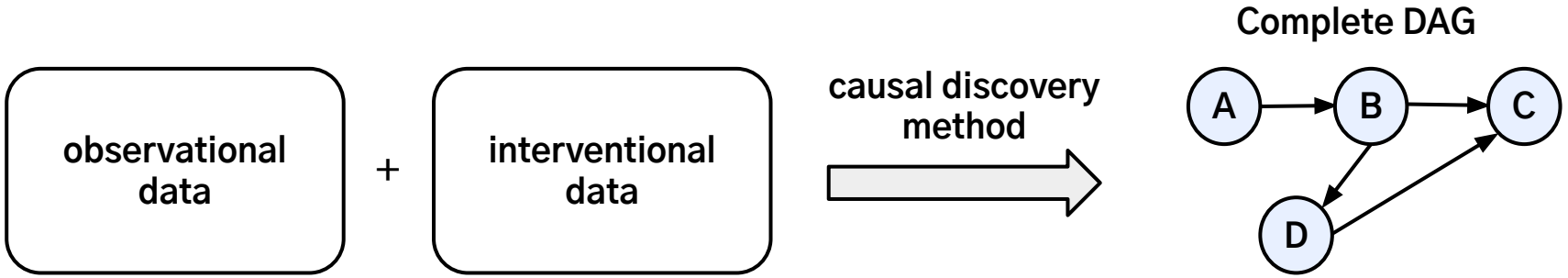
- The process of identifying causal relationships between variables from data
- It is generally known that a complete causal graph can be identified when both observational and interventional data are available
  - **observational data**: real-world data



  - **interventional data**: data collected through controlled experiments

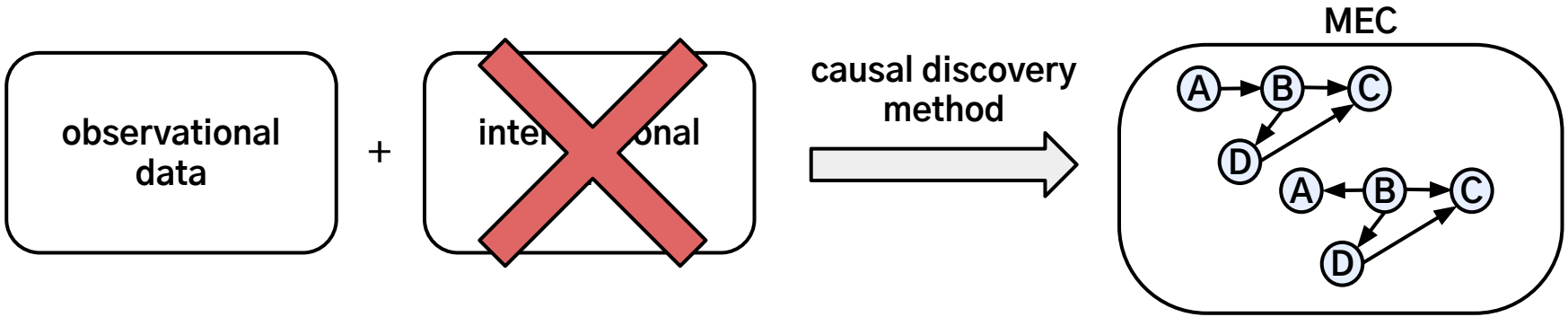
Causal Discovery

- The process of identifying causal relationships between variables from data
- It is generally known that a complete causal graph can be identified when both observational and interventional data are available



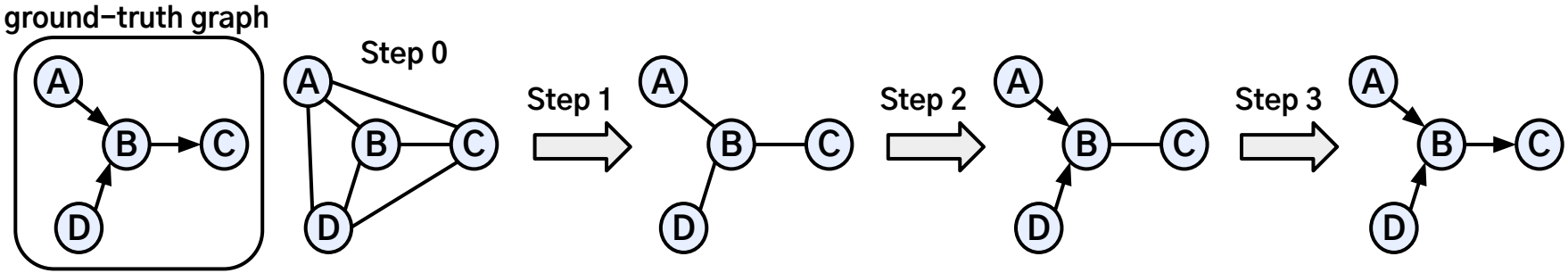
Causal Discovery

- The process of identifying causal relationships between variables from data
- It is generally known that a complete causal graph can be identified when both observational and interventional data are available
- Due to the high cost of interventions, observational data is often used, but it can only recover the causal structure up to the Markov Equivalence Class (MEC)
    - **MEC:** Set of DAGs that share the same conditional independence relationship



PC Algorithm

- A constraint-based method for causal discovery using conditional independence tests
- **Step**
  - Start with a complete undirected graph (all variables connected to each other)
  - Identify the skeleton (by unconditional / conditional independence)
  - Orient Colliders (X -> Z <- Y)
    (when X and Y are not adjacent and Z was not in the set that made X and Y independent))
  - Orient qualifying edges that are incident on colliders



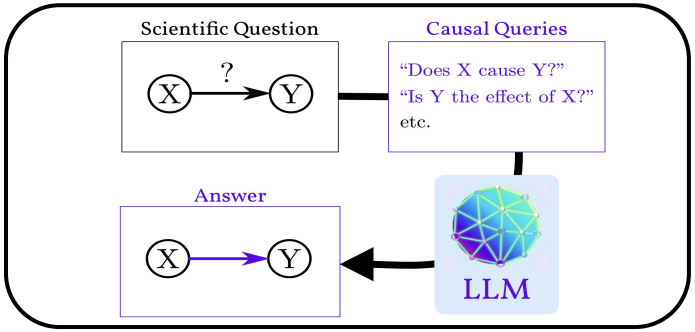
[1] Spirtes, Peter, Clark N. Glymour, and Richard Scheines. Causation, prediction, and search. MIT press, 2000.

Generative Models for Causal Graph Discovery

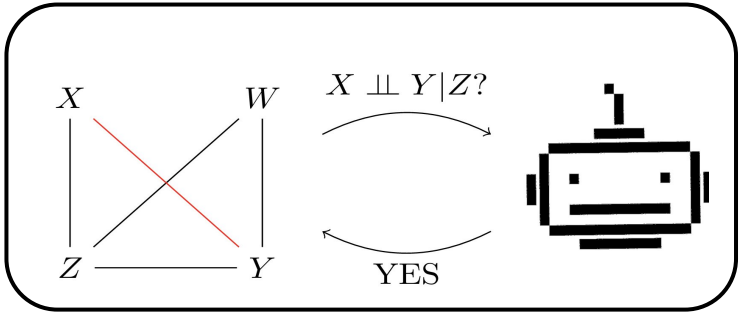
Causal Discovery with LLMs

- Large Language Models (LLMs) are transforming causal discovery by acting as scalable, generalized "meta-experts" that automate and accelerate expert-level reasoning
  - LLMs integrate information from various textual sources to assist in identifying causal patterns that may not be captured by conventional approaches alone

Examples of LLM causal discovery frameworks



Pairwise Approach



LLM Constraint based Approach

[2] Wan, Guangya, et al. "Bridging causal discovery and large language models: A comprehensive survey of integrative approaches and future directions." arXiv preprint arXiv:2402.11068 (2024).
[3] Zečević, Matej, et al. "Causal Parrots: Large Language Models May Talk Causality But Are Not Causal". Transactions on Machine Learning Research, 2023
[4] Cohrs, Kai-Hendrik, et al. "Large language models for constrained-based causal discovery." arXiv preprint arXiv:2406.07378 (2024).

Example) Does changing A cause a change in B?
Example) You are a helpful assistant for causal reasoning. Does changing A cause a change in B?

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	←

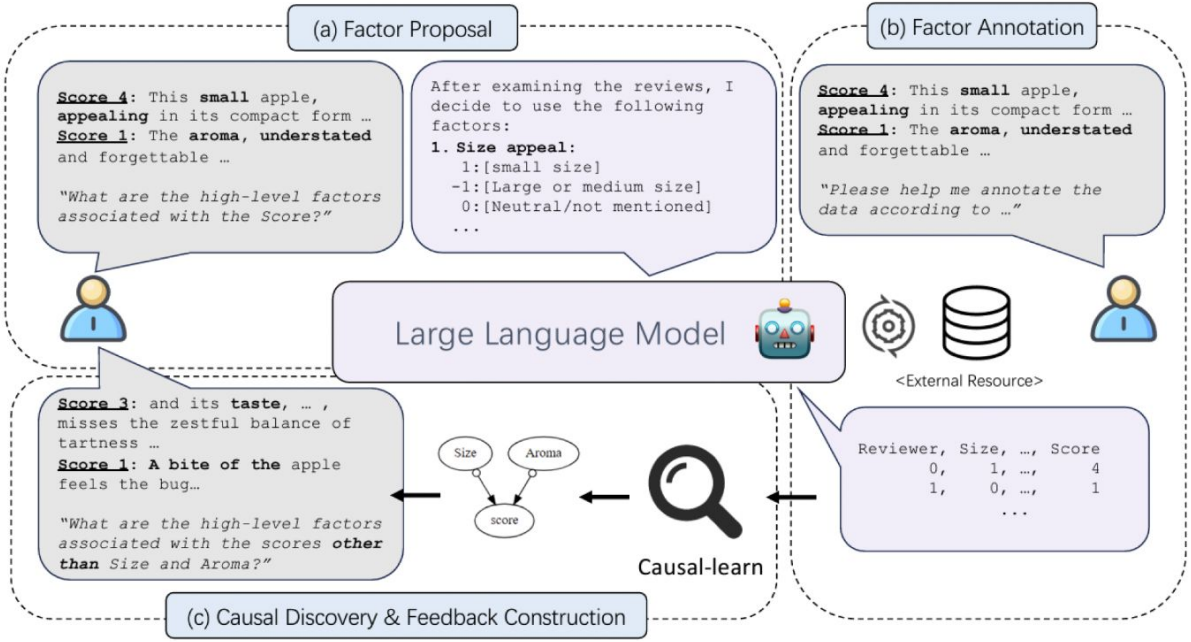
Model	Accuracy
ada	40.1
text-ada-001	50.0
babbage	50.0
text-babbage-001	50.9
curie	50.0
text-curie-001	50.0
davinci	38.4
text-davinci-001	50.0
text-davinci-002	51.7
text-davinci-003	55.1
gpt-3.5-turbo	71.1
gpt-3.5-turbo (neuropathic pain expert)	75.1
gp4-4	78.4
gpt-4 (neuropathic pain expert)	84.3
text-davinci-003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

[4] Cohrs, Kai-Hendrik, et al. "Large language models for constrained-based causal discovery." arXiv preprint arXiv:2406.07378 (2024).

Example: 25-1 DSL EDA Project with COAT

Generative Models for Causal Graph Discovery

COAT: Causal representatiOn AssistanT



Algorithm 1 The COAT Framework

Required: Dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$; LLM for factor proposal Ψ; Model for factor parsing Ψ_s; causal discovery algorithm $\mathcal{A}$; Feedback constructor $\mathcal{F}$; Maximal rounds T;
Random sampling $\widehat{\mathcal{D}}^1$;
Constructing $\boldsymbol{p}^1$;
while not converge and current round $t < T$ **do**
  $\mathcal{W}^t \leftarrow \Psi(\boldsymbol{p}^t, \widehat{\mathcal{D}}^t)$; //factor proposal
  $\boldsymbol{Z}^t \leftarrow \Psi_s(\mathcal{D}, \mathcal{W}^t, \boldsymbol{p}_p)$; //factor parsing
  $\mathcal{G}^t \leftarrow \mathcal{A}(\boldsymbol{Z}^{\leq t} \cup \{Y\})$; //causal discovery
  $(\widehat{\mathcal{D}}^{t+1}, \boldsymbol{p}^{t+1}) \leftarrow \mathcal{F}(\mathcal{G}^t, \mathcal{D}, \boldsymbol{p}^t)$; //feedback
end while
return $\mathcal{G}^T$

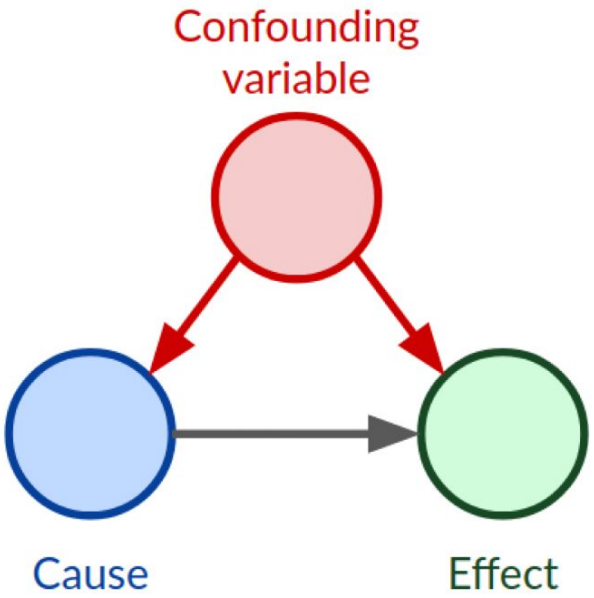
[5] Liu, Chen, et al. "Discovery of the Hidden World with Large Language Models" NeuIPS 2024 arXiv:2402.03941 (2024).

기울어진 아이스링크

: 2014 소치올림픽 여자 피겨 스케이팅
판정 논란 사례를 중심으로

25-1 EDA Sports Team
12기 신영군 김민규 13기 박수빈 이유주 조지성

Generative Models for Causal Graph Discovery



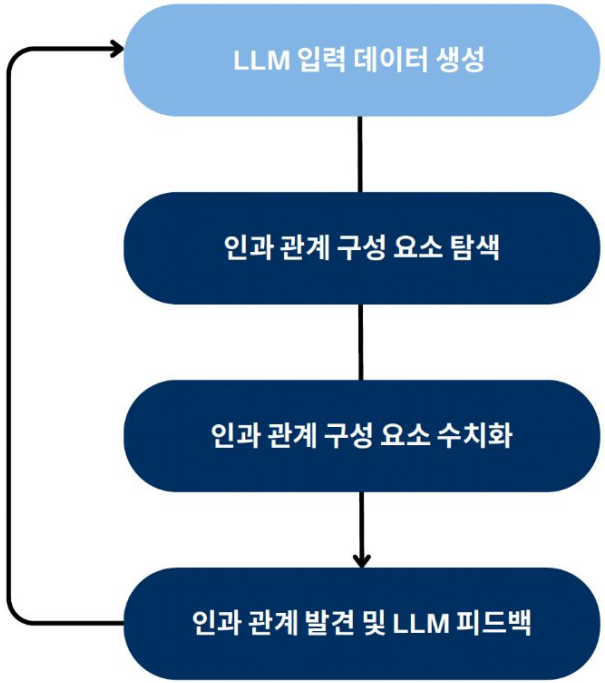
인과 관계 (Causal Relationship)
: 한 변수가 다른 변수에 직접적으로 영향을 주는 관계

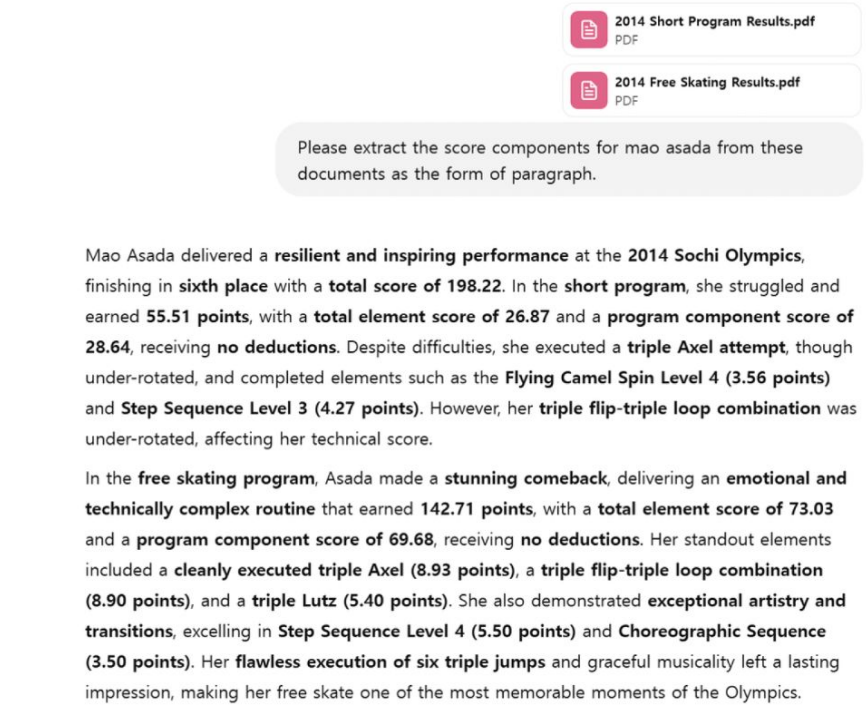
근데 문제는...
1) 영향을 주는 요소들을 일일이 모두 찾기는 어려움.
2) 인과 관계를 확인하기 위해 추상적인 요소들을
수치화하는 과정이 상당히 까다로움.

잠깐,
사전 학습된 언어 모델이 이 역할을 대신해준다면...?!

Example: 25-1 DSL EDA Project with COAT

Generative Models for Causal Graph Discovery



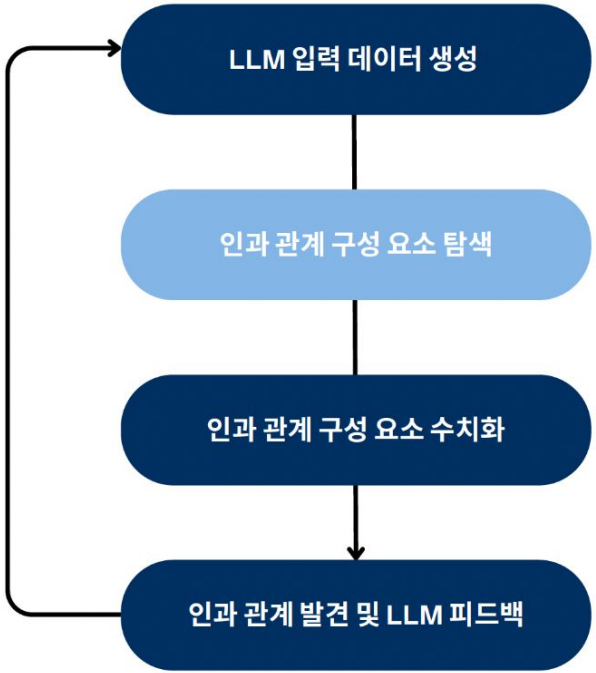


+ 댓글 Crawling

[5] Liu, Chen, et al. "Discovery of the Hidden World with Large Language Models" NeuIPS 2024 arXiv:2402.03941 (2024).

Example: 25-1 DSL EDA Project with COAT

Generative Models for Causal Graph Discovery



출력 결과 (여자 싱글)

- 기술 난이도 & 수행능력
- 구성 점수 (PCS)
- 일관성 & 깔끔함
- 점수에 대한 공정성
- 역사적, 문화적 영향력

여자 싱글 프로그램에서
공정성 요인과 관련된
인과관계 가능성이 제기됨!

Final Reported Factors:

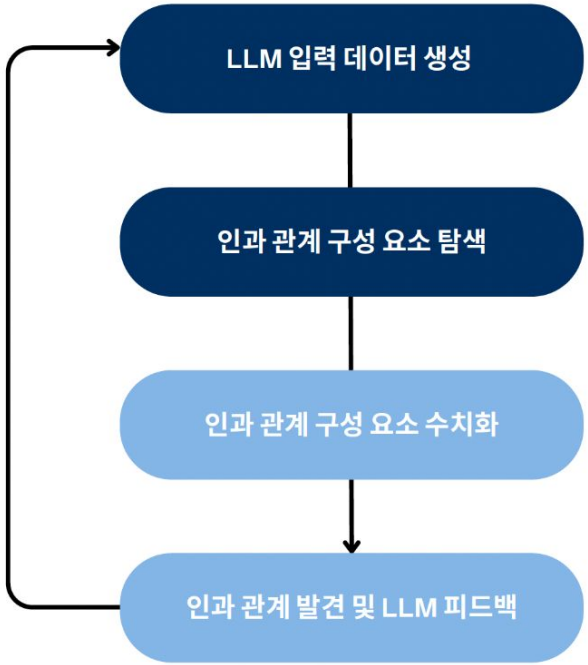
markdown 복사 편집

```
**Technical Difficulty & Execution**
- 1: High base value and positive GOE across elements.
- 0: Average difficulty or neutral GOE.
- -1: Major execution issues or weak base value.

**Program Components (PCS) - Artistry & Interpretation**
- 1: Exceptional artistry, performance, and interpretation.
- 0: Average or competent artistic performance.
- -1: Weak interpretation or lack of musical connection.

**Consistency & Cleanliness**
- 1: No falls, strong landings, seamless execution.
- 0: Minor mistakes but overall intact program.
- -1: Falls, stumbles, or disrupted rhythm.

**Perceived Scoring Fairness**
- 1: Widely accepted as fair.
- 0: Some debate but no major controversy.
- -1: Public or expert backlash, claims of bias.

**Historical & Cultural Impact**
- 1: Recognized as a legendary or influential performance.
- 0: Respected but not groundbreaking.
- -1: Seen as undeserving of historical recognition.
```

[5] Liu, Chen, et al. "Discovery of the Hidden World with Large Language Models" NeuIPS 2024 arXiv:2402.03941 (2024).

Example: 25-1 DSL EDA Project with COAT

Generative Models for Causal Graph Discovery

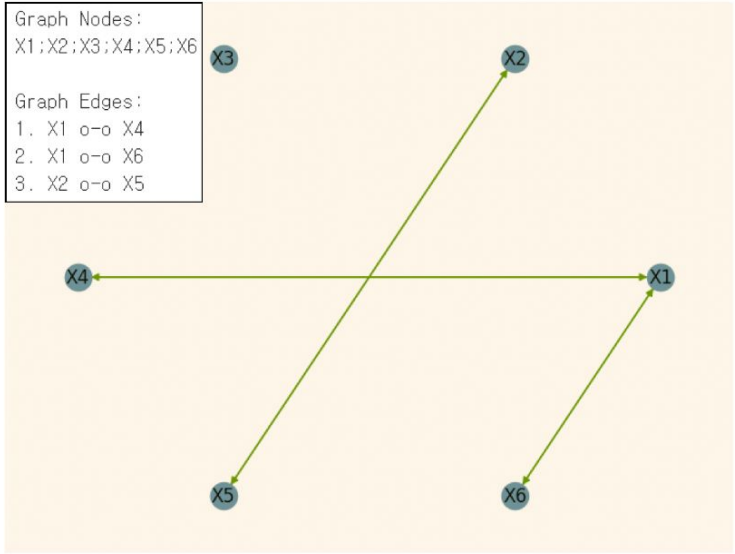


결과 시각화 (FCI)

X1: 기술 난이도 & 수행능력
X2: 구성 점수 (PCS)
X3: 일관성 & 깔끔함
X4: 점수에 대한 공정성
X5: 역사적, 문화적 영향력
X6 (Y): 대회 성적 (관심 변수)

여자 싱글 프로그램에서
공정성 요소가 기술 점수에
영향을 주어 최종적으로
결과와 인과관계를 맺고 있음!

Discovered using **PC algorithm**



[5] Liu, Chen, et al. "Discovery of the Hidden World with Large Language Models" NeuIPS 2024 arXiv:2402.03941 (2024).

Limits of LLM Approaches

Generative Models for Causal Graph Discovery

- **Sensitivity to Prompt Design**
  - Performance fluctuates with prompt phrasing (e.g., capitalization, triplet vs. pairwise formats)
  - Variable descriptions must align with LLMs' pre-trained knowledge, risking failure in unfamiliar domains

- **Hallucination Risks**
  - LLMs may generate plausible but incorrect causal links, especially in low-knowledge domains (e.g., novel medical variables)

- **Validation Dependency**
  - Without external validation, LLMs may generate unreliable causal links that compromise the trustworthiness of results

[6] Jiang, Haitao, et al. "Large language model for causal decision making." arXiv preprint arXiv:2312.17122 (2023).
[7] Kiciman, Emre, et al. "Causal reasoning and large language models: Opening a new frontier for causality." Transactions on Machine Learning Research (2023)

Jihee Kim, Minseol Jang, Miryoung Kim, Hyung Jin Han,
Kangjun Noh, Sumin Park, Kyungwoo Song, Hae Sun Suh *(ISPOR 2025)*

Generative Models based Causal Graph Discovery for Medical Domain

MAGIC (Multi-LLM Assisted Graph Inference and Correction)

Limitation

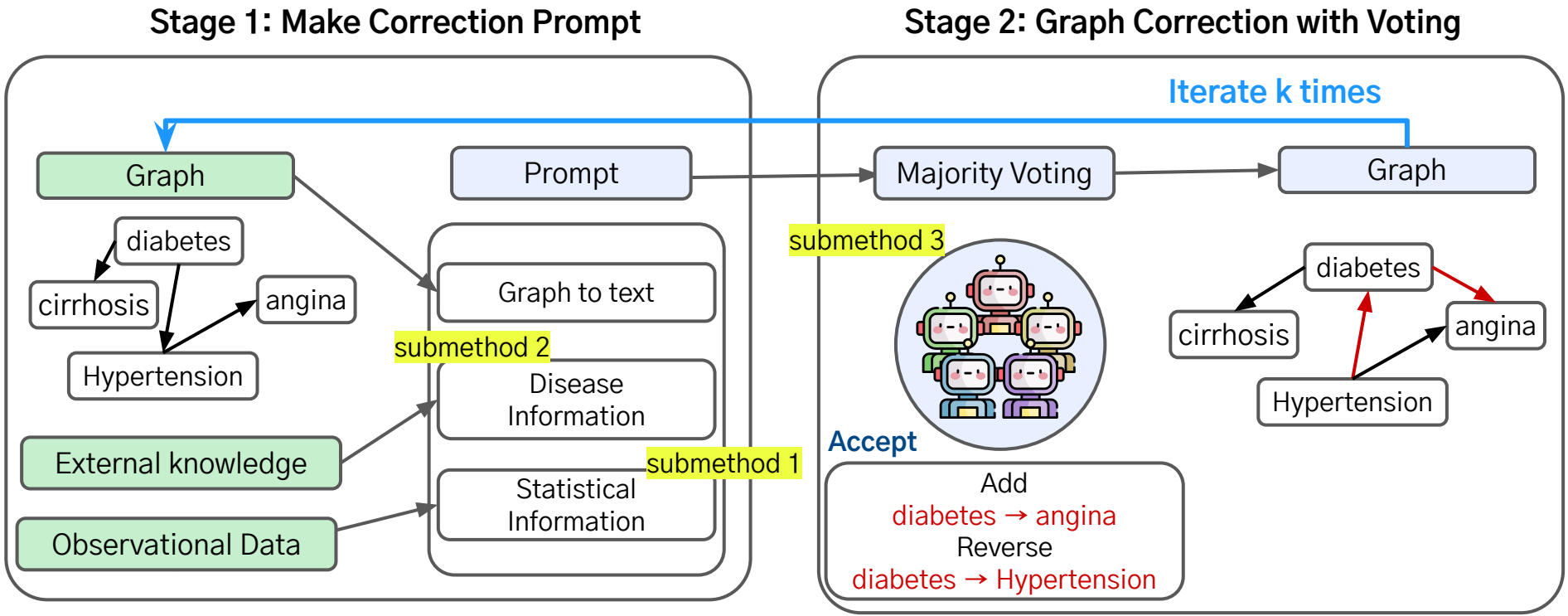
1. Lack of Domain-Specific Knowledge : Fails in unfamiliar domains or with novel variable descriptions
2. Inconsistent Outputs from Single LLM: Prone to hallucination and bias without validation

Our Model : MAGIC (Multi-LLM Assisted Graph Inference and Correction)

To enhance the reliability and validity of causal graphs generated by LLM + pairwise reasoning, we apply the following three correction methods

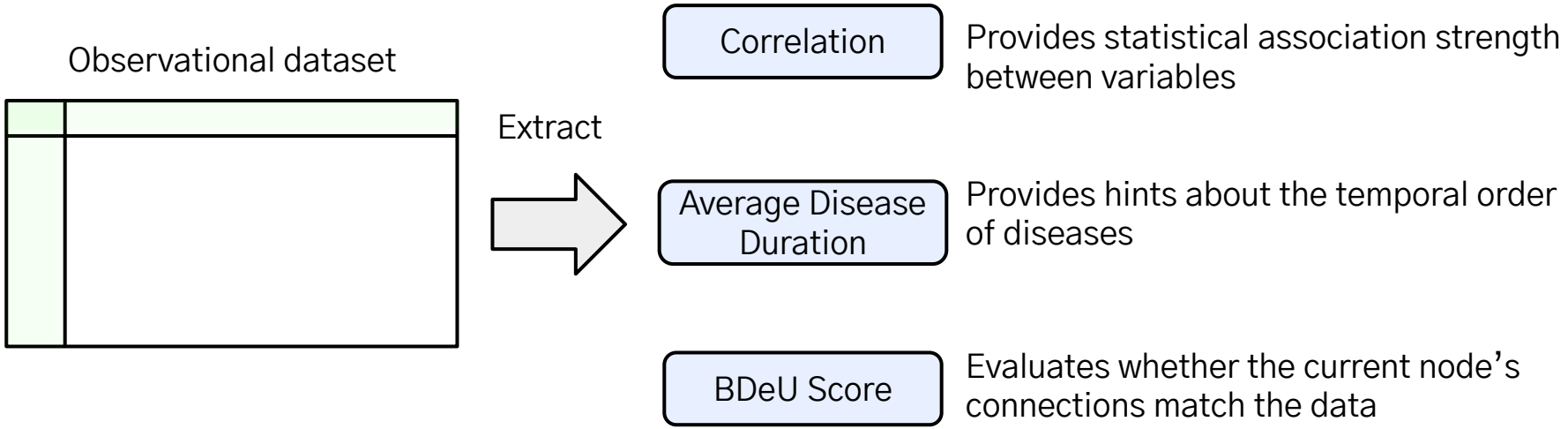
1. Statistical Feedback
2. External Knowledge Injection
3. Multi-LLM Voting

Overall Process of MAGIC



submethod 1: Statistical Information from Data

- Integrates objective statistical informations extracted from observational data into the prompt to enhance the data-driven causal inference ability of LLMs



submethod 1: Statistical Information from Data

- Integrates objective statistical informations extracted from observational data into the prompt to enhance the data-driven causal inference ability of LLMs

BDeU(Bayesian Dirichlet Equivalent Uniform) Score

- **Definition**: Evaluates how well the estimated DAG fits the data under a Bayesian framework

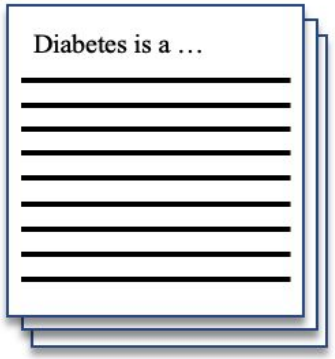
$$\log \mathrm{BDeu}(G \mid D) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \left[\log \Gamma(\alpha_{ij}) - \log \Gamma(\alpha_{ij} + N_{ij}) + \sum_{k=1}^{r_i} (\log \Gamma(\alpha_{ijk} + N_{ijk}) - \log \Gamma(\alpha_{ijk})) \right]$$

- G: DAG structure.
- D: Dataset.
- n: Number of variables.
- q_i: Number of parent configurations for variable i.
- r_i: Number of possible values for variable i.
- N_{ij}: Number of data points with parent configuration j.
- α_{ijk}: Hyperparameter (Dirichlet prior count).

submethod 2: External clinical knowledge

- Incorporates summarized clinical knowledge from Wikipedia into prompts to enhance LLM's causal reasoning with factual medical context

first sections of Wikipedia page



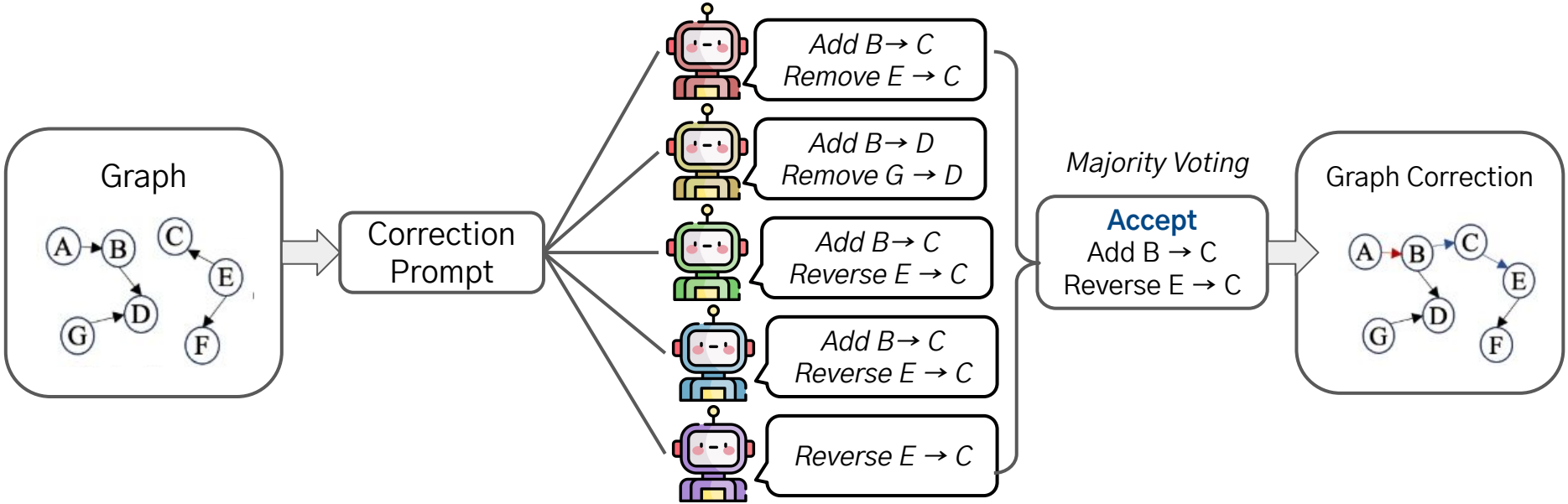
Summarize into a single paragraph

In prompt

Below is the medical information:
Diabetes mellitus is a chronic metabolic disorder characterized by prolonged high blood sugar levels due to insufficient insulin production or cellular resistance to insulin. ⋯
Hypertension, ⋯
Dyslipidemia, ⋯

submethod 3: Multi-LLM Voting

- A collective intelligence approach that aggregates independent judgments from multiple LLMs to offset individual hallucinations and biases



Dataset Overview

- **Our Goal**: accurately discover causal relationships between chronic diseases by constructing a reliable causal graph

- **Dataset Construction**
  - In collaboration with Kyung Hee University of Pharmacy, we built a dataset on 13 chronic diseases.
  - Source: KNHANES (Korean National Health and Nutrition Examination Survey)
  - Size: 36,107 participants
  - Variables: 13 chronic diseases
    - Disease variables are categorized in 10-year intervals based on disease duration

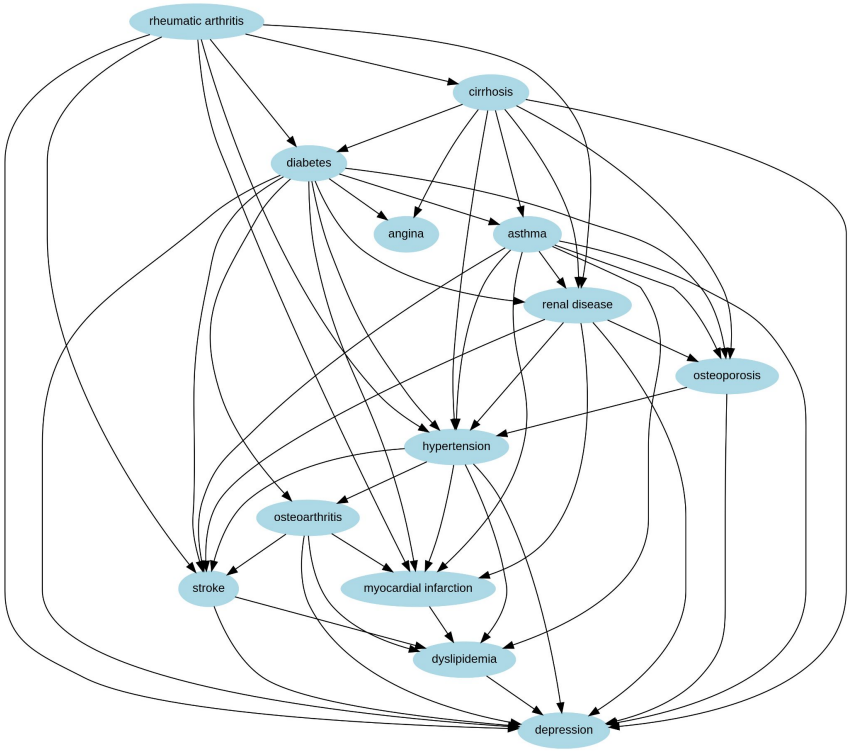
Reference DAG Construction

- No absolute ground truth graph exists in the medical field for causal evaluation

  ➔ A reference DAG is needed to benchmark model performance

- Reviewed 138 peer-reviewed medical publications
- Collaborated with medical experts to refine and finalize the graph

Reference DAG



- **Baselines**
  - Statistical Methods
    - PC (Peter-Clark): Conditional Independence based
    - GES (Greedy Equivalence Search): Score based Search
    - LiNGAM: Functional-Based Methods
  - LLM Based Methods (Use gpt-4o)
    - Pairwise: Constructs the graph by asking causal questions for each variable pair
    - BFS (Breadth-First Search): Starts from a key variable and expands causal links step by step
- **MAGIC**
  - We use 5 LLM models (gpt 4o, claude 3.7 sonnet, Gemini-2-flash, deepseek r1, Llama 3.3 70B)

[8] Chickering, David Maxwell. "Optimal structure identification with greedy search." Journal of machine learning research 3.Nov (2002): 507-554.
[9] Shimizu, Shohei, et al. "DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model." Journal of Machine Learning Research-JMLR 12.Apr (2011): 1225-1248.
[10] Jiralerspong, Thomas, et al. "Efficient causal graph discovery using large language models." arXiv preprint arXiv:2402.01207 (2024).

1. Skeleton Accuracy

- Evaluates only the presence of edges, ignoring direction

2. Orientation Accuracy

- Evaluates the correctness of edge directions

3. SHD (Structural Hamming Distance)

- Measures how many edge insertions, deletions, or direction reversals are needed to match the ground truth graph
- SHD = Number of edges to add & delete (in the skeleton) + Number of edge direction differences

Performance Comparison

Generative Models based Causal Graph Discovery for Medical Domain

Model	Skeleton Precision (%)	Skeleton Recall (%)	Skeleton F1 Score (%)	Orientation Precision (%)	Orientation Recall (%)	Orientation F1 Score (%)	SHD
PC	87.50	42	56.76	39.13	18	24.66	44
GES	82.61	38	52.05	27.78	10	14.71	49
LiNGAM	90.48	38	53.52	57.14	24	33.80	40
LLM_Pairwise	**96.30**	52	67.53	**77.78**	42	54.55	30
LLM_BFS	86.67	26	40	46.67	14	21.54	45
MAGIC (Ours)	94.12	**64**	**76.19**	73.53	**50**	**59.52**	**27**

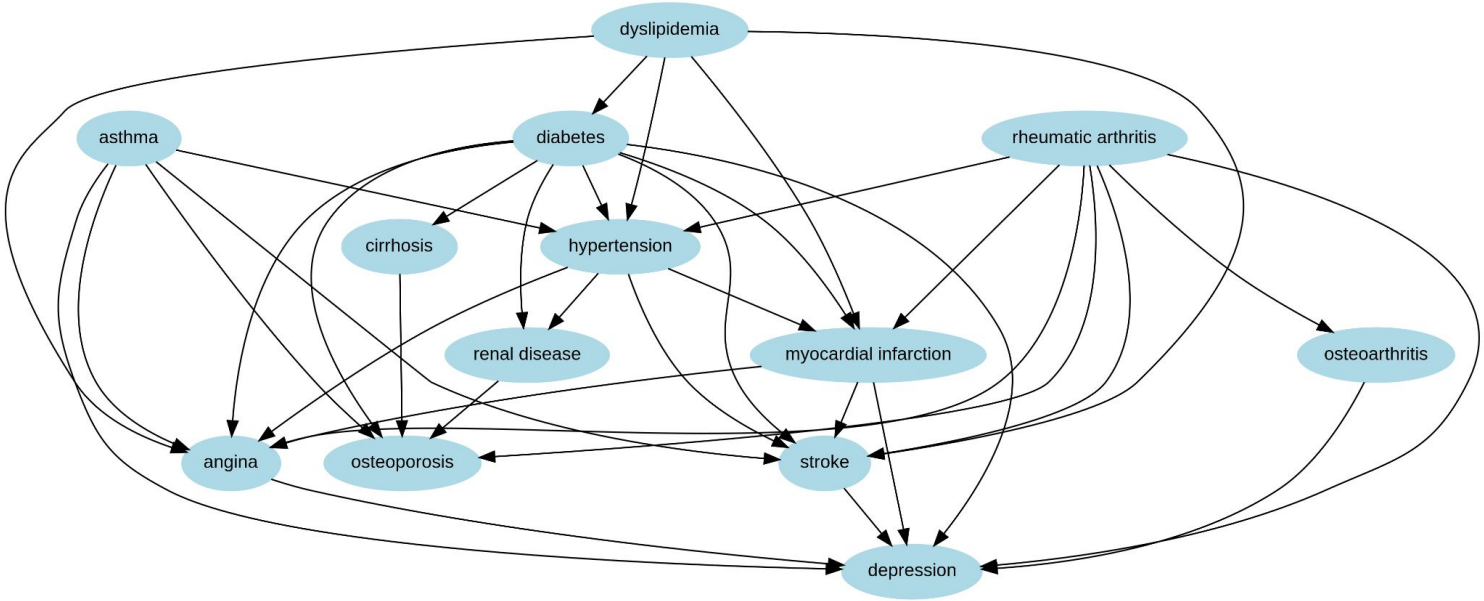
- MAGIC achieved the best overall performance based on F1-score and SHD.
- While skeleton and orientation precision were second-best, MAGIC had the lowest SHD (27) and highest F1-scores.

Performance Gain per Iteration in MAGIC

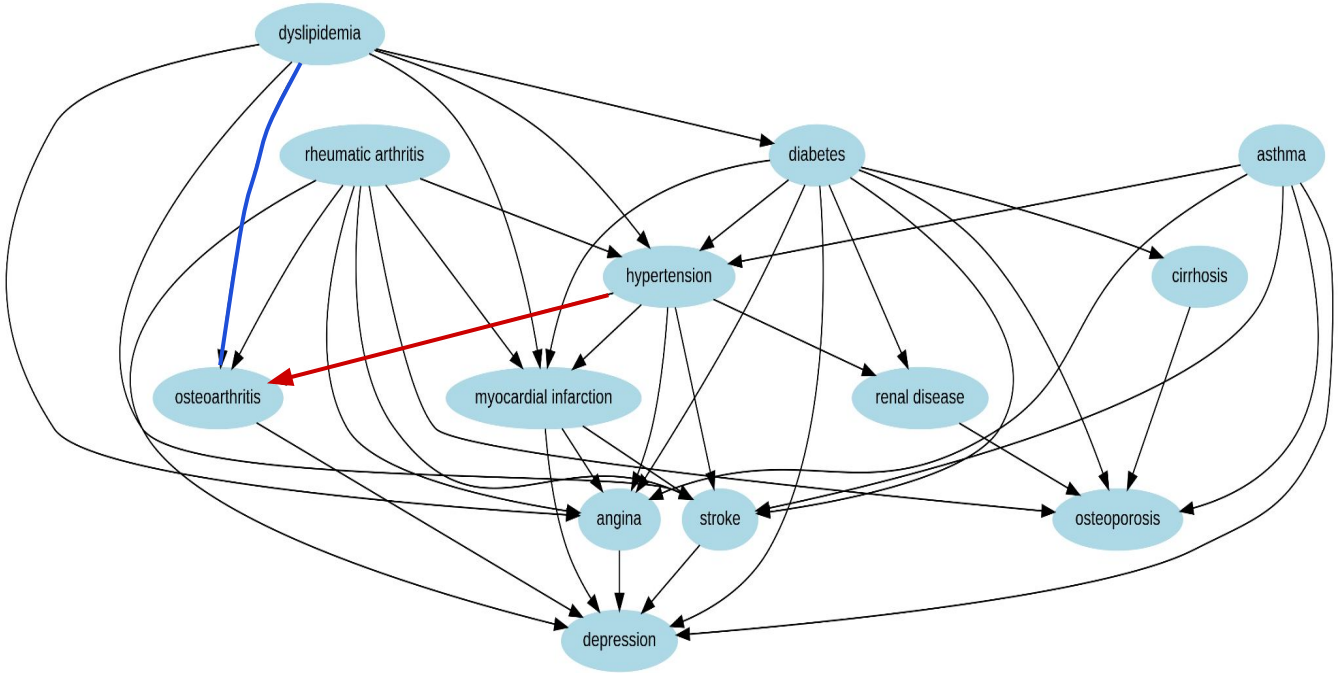
Iteration	Skeleton Precision (%)	Skeleton Recall (%)	Skeleton F1 Score (%)	Orientation Precision (%)	Orientation Recall (%)	Orientation F1 Score (%)	SHD
1	**96.55**	56	70.89	**75.86**	44	55.70	29
2	93.94	62	74.70	72.73	48	57.83	28
3	94.12	**64**	**76.19**	73.53	**50**	**59.52**	**27**
4	94.12	**64**	**76.19**	73.53	**50**	**59.52**	**27**
5	94.12	**64**	**76.19**	73.53	**50**	**59.52**	**27**

- MAGIC showed progressive improvements in recall and F1−score over iterations.
- Performance converged at the third iteration, indicating stable final results.

Initial graph (Output of LLM Pairwise Approach)



Iteration 1

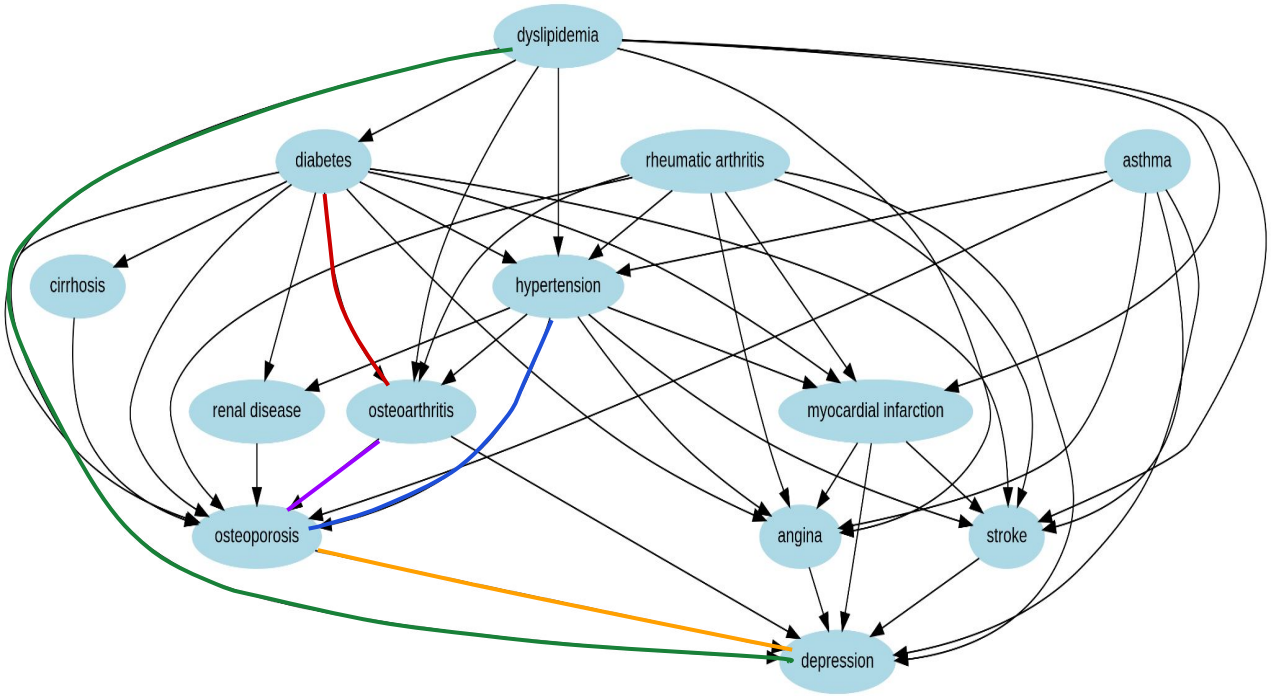


Accepted

hypertension → osteoarthritis (add)

dyslipidemia → osteoarthritis (add)

Iteration 2

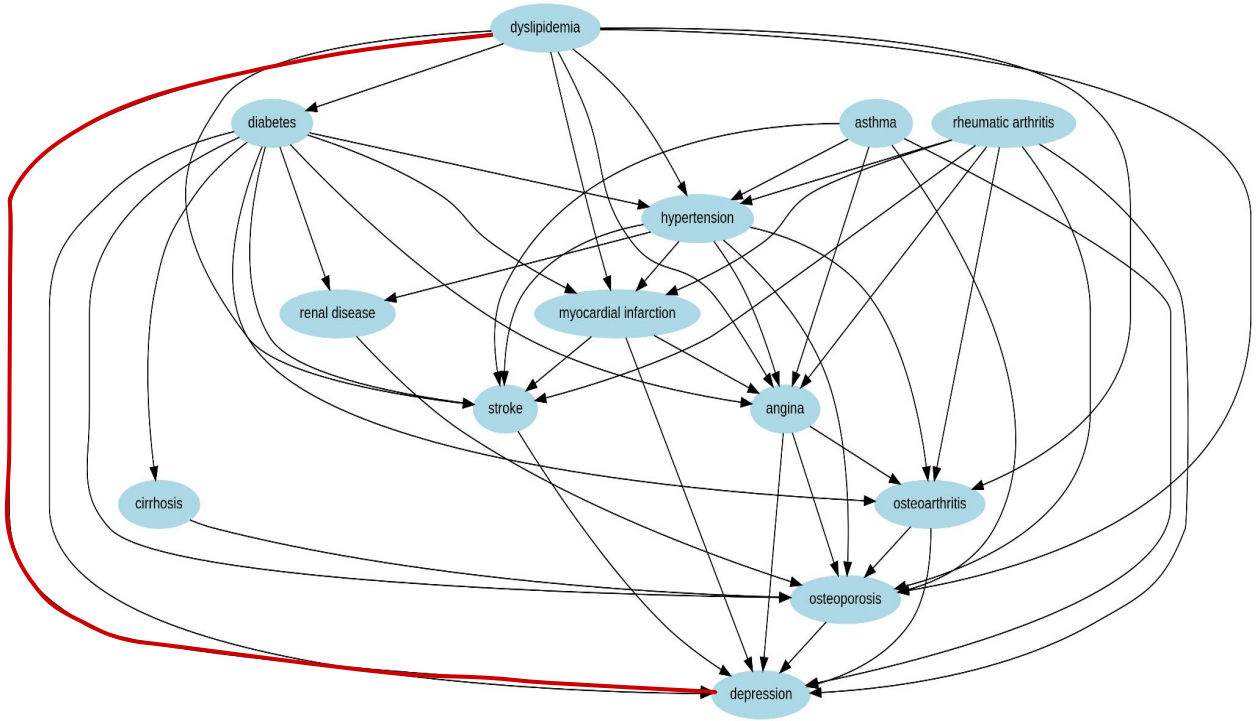


Accepted

diabetes → osteoarthritis (3/5) (add)
hypertension → osteoporosis (3/5) (add)
dyslipidemia → osteoporosis (4/5) (add)
osteoarthritis → osteoporosis (4/5) (add)
osteoporosis → depression (3/5) (add)

MAGIC: Edge Change during Correction Process – Iteration 3 (Convergence)

Iteration 3



Accepted

dyslipidemia → depression (3/5) (add)

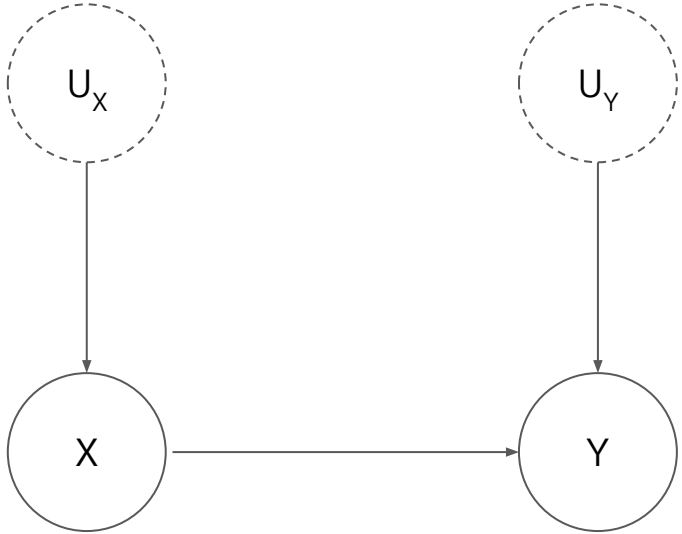
Additional Topics
– Causal Change Point Detection

SCM Definition

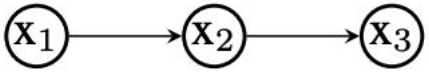
A tuple of the following sets:

- A set of endogenous variables
- A set of exogenous variables
- A set of functions, one to generate each endogenous variable as a function of the other variables

- $X = f_X(U_x)$
- $Y = f_Y(X, U_Y)$



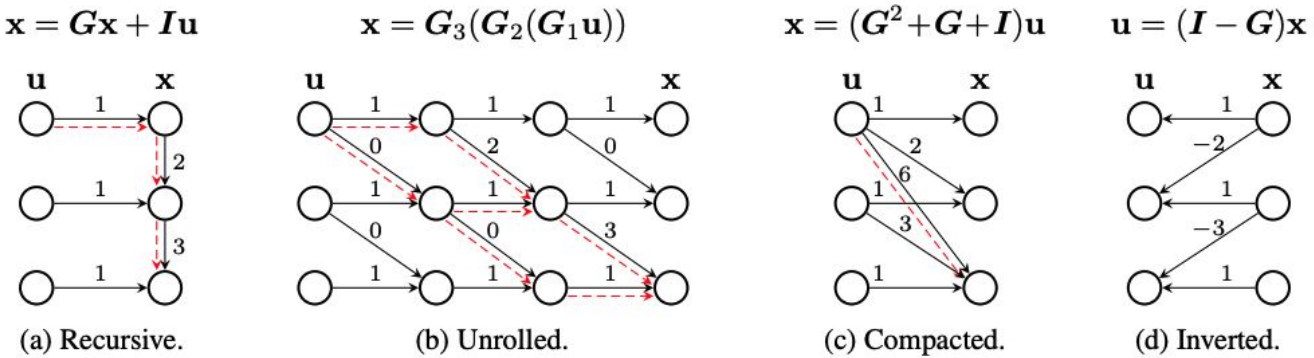
TMI (triangular monotonic increasing)



$$\pi = (1 \quad 2 \quad 3)$$

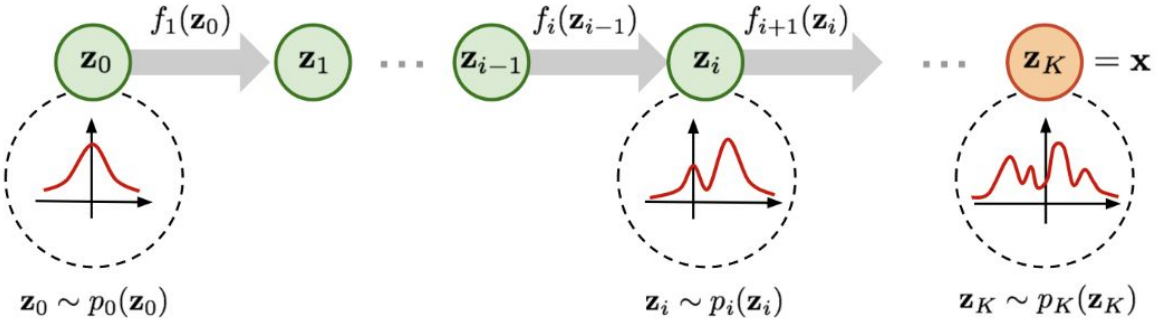
$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

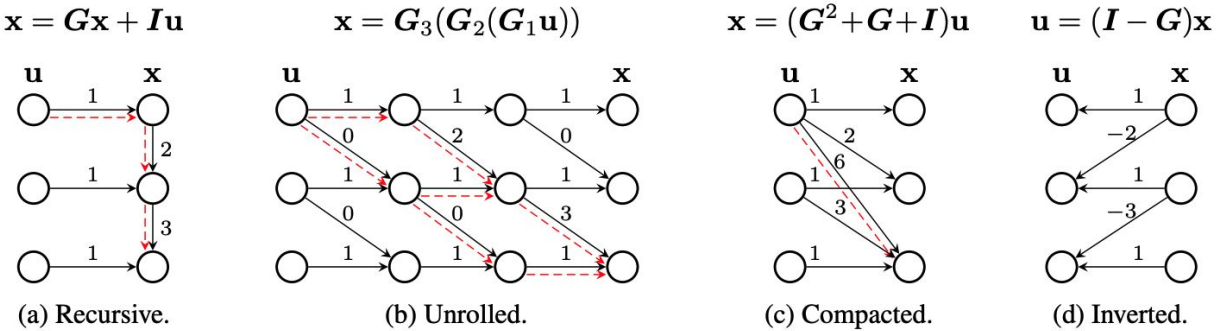
Figure 2: Causal graph, and its causal ordering π and adjacency matrix $\boldsymbol{A}$.



[11] Javaloy, Adrián, Pablo Sánchez-Martín, and Isabel Valera. "Causal normalizing flows: from theory to practice." Advances in Neural Information Processing Systems 36 (2023): 58833-58864.

Additional Topics – Causal Change Point Detection





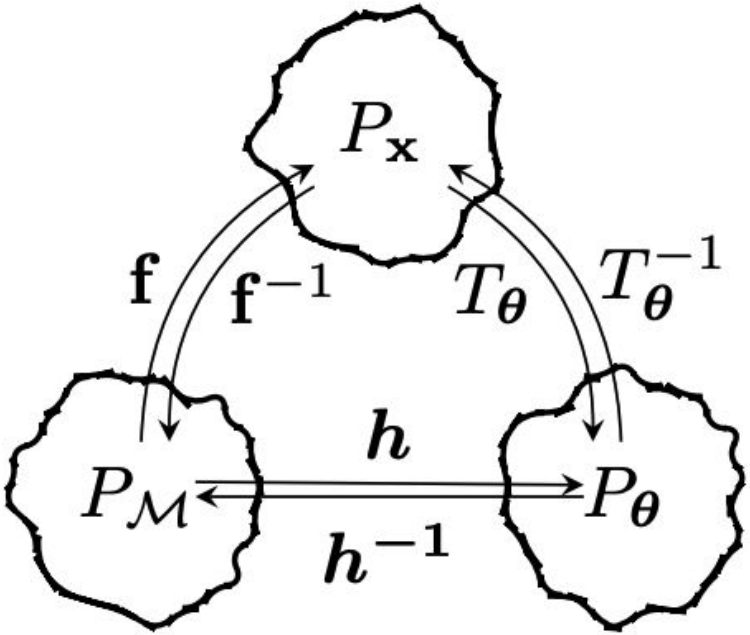
When using Normalizing Flows:

1) **It is possible to infer the SCM (data generating process) using only a causal graph and observational data.**
2) Interventions can be performed, enabling the estimation of causal effects.

[11] Javaloy, Adrián, Pablo Sánchez-Martín, and Isabel Valera. "Causal normalizing flows: from theory to practice." Advances in Neural Information Processing Systems 36 (2023): 58833-58864.

Additional Topics – Causal Change Point Detection

F: set of all TMI (triangular monotonic increasing) maps
P_u: set of all fully-factorized distributions p(u)



Theorem 1 (Identifiability). If two elements of the family $\mathcal{F} \times \mathcal{P}_{\mathbf{u}}$ (as defined above) produce the same observational distribution, then the two data-generating processes differ by an invertible, component-wise transformation of the variables $\mathbf{u}$.

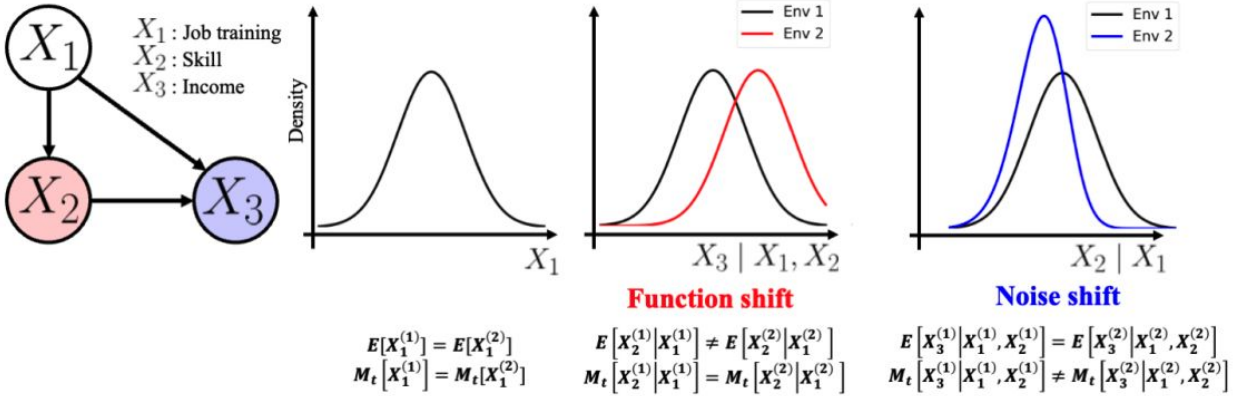
Theorem 1 on identifiability, which states that if two causal Normalizing Flow (NF) models within the family $F \times P_u$ produce the same observational distribution P_x, their learned transformations (T_θ) from observed data *x* to exogenous variables *u* differ from the true data-generating process only by an invertible, component-wise transformation *h* of these latent variables.

This implies that a Causal NF (T_θ) can recover the true exogenous variables up to this invertible function *h*, ensuring the functional dependencies learned by the NF are causally consistent with the underlying structural causal model.

[11] Javaloy, Adrián, Pablo Sánchez-Martín, and Isabel Valera. "Causal normalizing flows: from theory to practice." Advances in Neural Information Processing Systems 36 (2023): 58833-58864.

Research conducted in our lab (by Gyeongdeok Seo, Jaeyoon Shim, **Mingyu Kim**, Hoyoon Byun, Yonghan Jung, Kyungwoo Song)

Additional Topics – Causal Change Point Detection



$$E[X_1^{(1)}] = E[X_1^{(2)}]$$
$$M_t\left[X_1^{(1)}\right] = M_t[X_1^{(2)}]$$

$$E\left[X_2^{(1)} \middle| X_1^{(1)}\right] \neq E\left[X_2^{(2)} \middle| X_1^{(2)}\right]$$
$$M_t\left[X_2^{(1)} \middle| X_1^{(1)}\right] = M_t\left[X_2^{(2)} \middle| X_1^{(2)}\right]$$

$$E\left[X_3^{(1)} \middle| X_1^{(1)}, X_2^{(1)}\right] = E\left[X_3^{(2)} \middle| X_1^{(2)}, X_2^{(2)}\right]$$
$$M_t\left[X_3^{(1)} \middle| X_1^{(1)}, X_2^{(1)}\right] \neq M_t\left[X_3^{(2)} \middle| X_1^{(2)}, X_2^{(2)}\right]$$

When using Normalizing Flows:

1) It is possible to determine whether the **SCM has changed over time.**
2) If a change is detected, **it is possible to identify which node has changed**, whether the exogenous variables have changed, **or whether the functional mechanisms (f) have changed**.

[12] Dissecting Causal Mechanism Shifts via FANS: Function And Noise Separation using Flows

Conclusion

- Generative models aim to estimate the population density from samples.
- In causal inference, especially in real-world settings, accurately estimating hidden confounders is crucial.
- Generative-model-based causal inference methods are effective at accounting for hidden confounders and estimating treatment effects.
- Such generative models, including large language models (LLMs), can also be widely applied to causal graph discovery.

※ Most slides are adapted from
Innovations in Regulatory Science: Real-World Evidence and Causal AI, June 19 2025.