From Prediction to Understanding: A Causal Perspective

Adèle H. Ribeiro

https://causxai.github.io/

adele.ribeiro@rwth-aachen.de

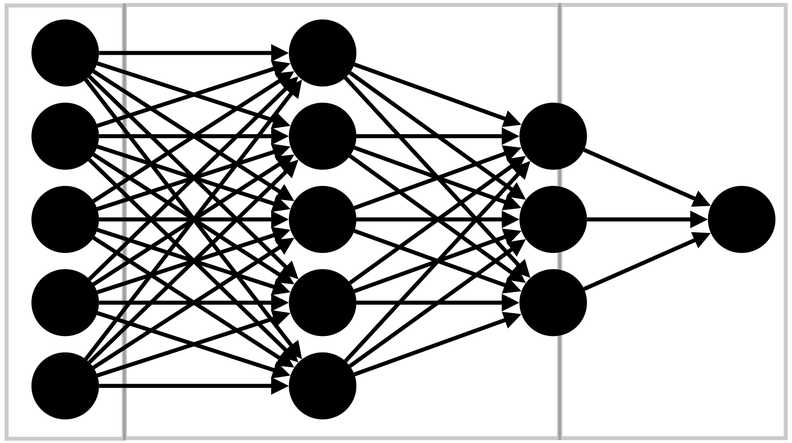
Causal and Explainable Artificial Intelligence in Life Sciences
RWTH Aachen University

Causal and Explainable
Artificial Intelligence
in Life Sciences

RWTH AACHEN UNIVERSITY

Artificial Intelligence & Machine Learning

- **Perception:** detecting and classifying patterns in medical imaging.
- **Language:** Translating documents, summarizing articles, flagging spam emails.
- **Prediction:** Forecasting demand, predicting disease progression
- **Integration:** Integrating multi-omics data for biological analysis
- **Generation:** Generating product images, transferring artistic style onto a photo.



Prediction

$P(Y = y \mid \mathbf{X})$

$E(Y \mid \mathbf{X})$

Medical Diagnosis

$P(Y = \text{disease} \mid X = \text{patient features})$

Image Classification

$P(Y = \text{dog} \mid X = \text{pixel array})$

Language Modeling

$P(Y = w_t \mid X = w_1, \ldots, w_{t-1})$

Speech Recognition

$P(Y = \text{text} \mid X = \text{audio})$

Weather Forecasting

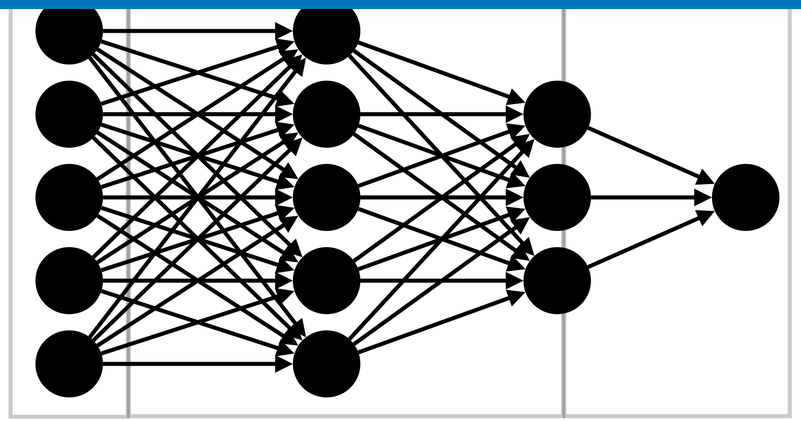
$P\left(Y = T_{t+h} \mid T_{t-1}, T_{t-2}, \ldots T_{t-p}\right)$

Artificial Intelligence & Machine Learning

- **Perception:** detecting and classifying patterns in medical imaging.
- **Language:** Translating documents, summarizing articles, flagging spam emails.
- **Prediction:** Forecasting demand, predicting disease progression
- **Integration:** Integrating multi-omics data for biological analysis

The Scaling Era: More Data, Larger Models, More Compute.

But is scale sufficient for scientific understanding?



$\mathbf{X}$: Input
Observed features

$\mathbf{Z}$: Learned features

Prediction

$P(Y = y \mid \mathbf{X})$

$E(Y \mid \mathbf{X})$

s
t features)

n

$P(Y = \text{dog} \mid X = \text{pixel array})$

Language Modeling

$P(Y = w_t \mid X = w_1, \ldots, w_{t-1})$

Speech Recognition

$P(Y = \text{text} \mid X = \text{audio})$

Weather Forecasting

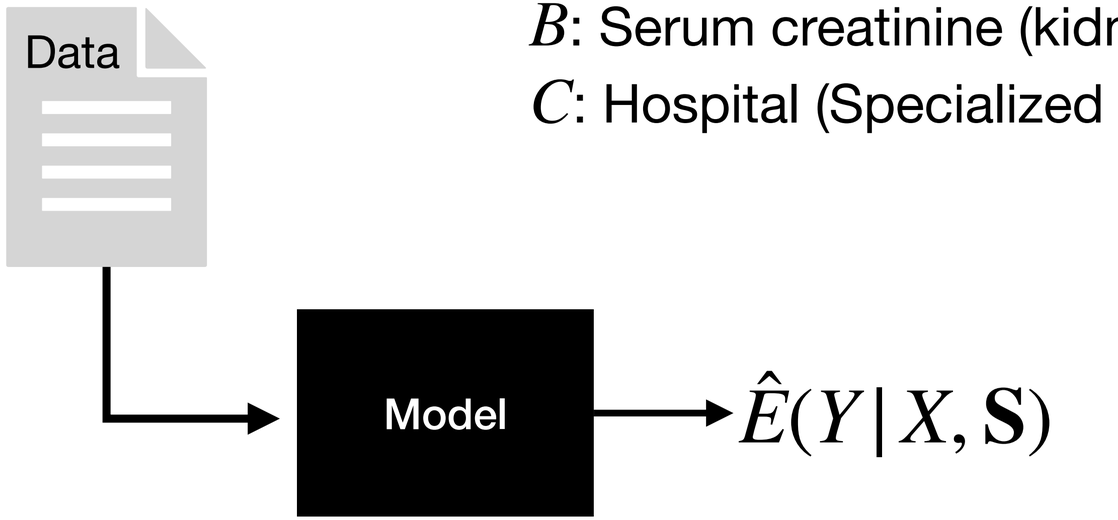
$P\left(Y = T_{t+h} \mid T_{t-1}, T_{t-2}, \ldots T_{t-p}\right)$

From Data to Predictive Models

Dose Response Prediction: How does the dose of an intravenous antihypertensive drug (X) affect a patient's systolic blood pressure (Y)?

$$\mathbf{V} = \{X, Y, A, B, C\}$$

Y: Systolic blood pressure difference after 24 hours (outcome)
X: IV drug infusion rate (treatment)
A: CYP450 pharmacogenomic variant (drug metabolism capacity)
B: Serum creatinine (kidney function marker)
C: Hospital (Specialized Cardiorespiratory Center vs. General)

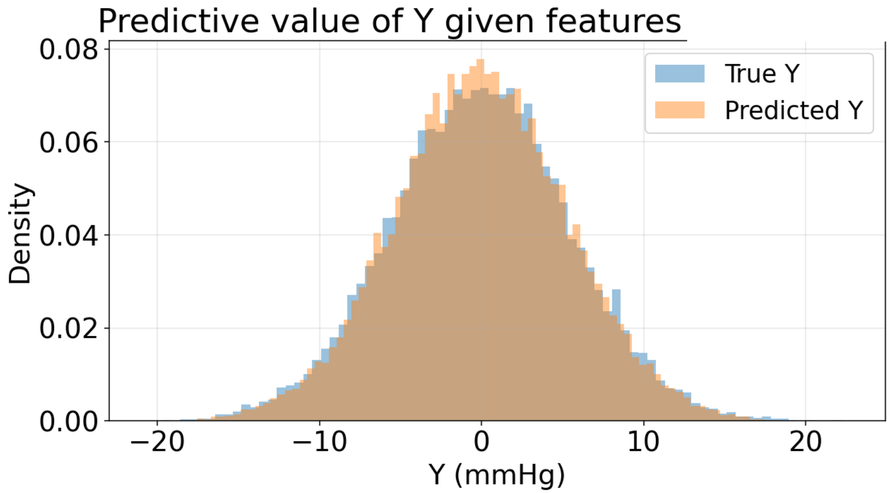


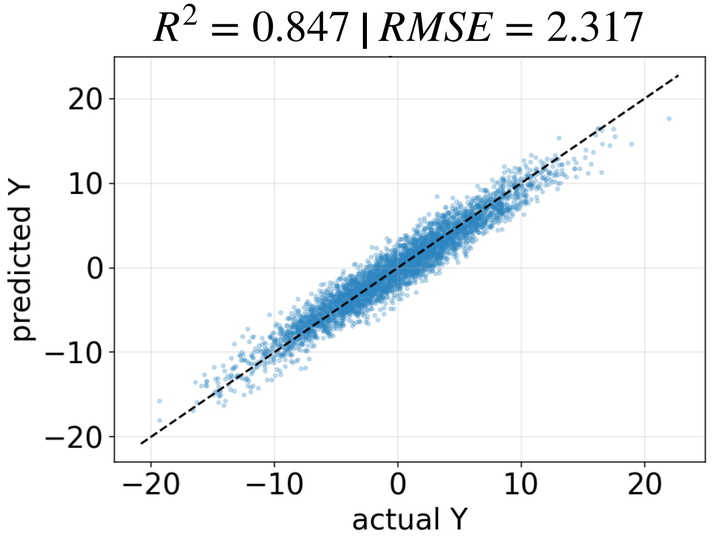
$$\mathbf{S} = \{A, B, C\}$$

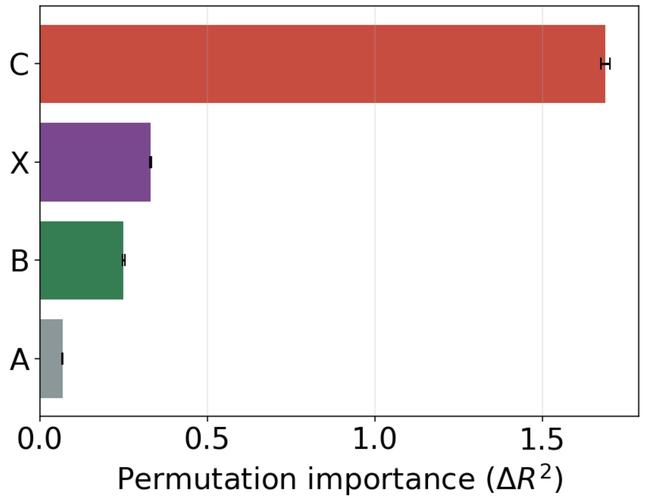
We can train a model using ML/AI.

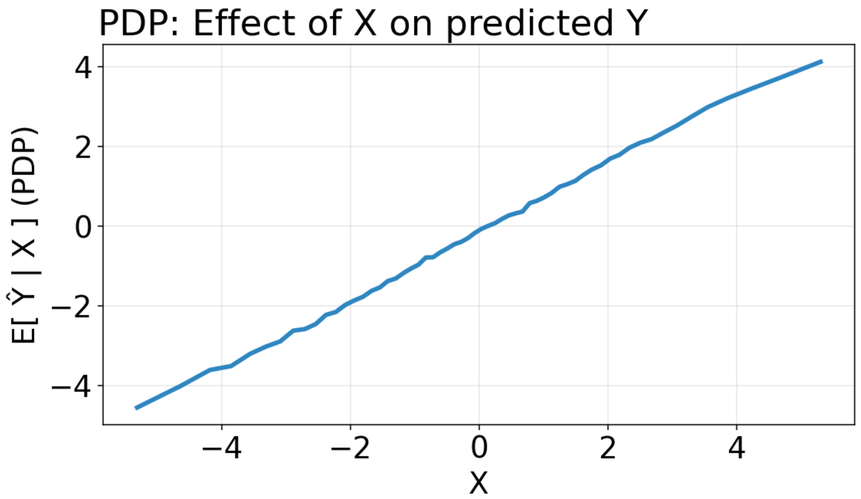
From Data to Predictive Models

Best of Random Forest, Gradient Boosting and Linear Regression









Y: Systolic blood pressure change

X: IV drug infusion rate (treatment)

A: CYP450 pharmacogenomic variant (drug metabolism capacity)

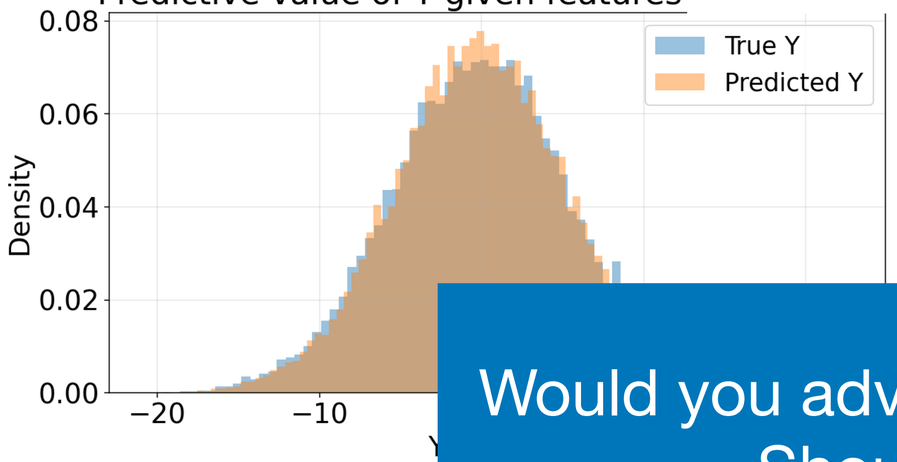
B: Serum creatinine (kidney function marker)

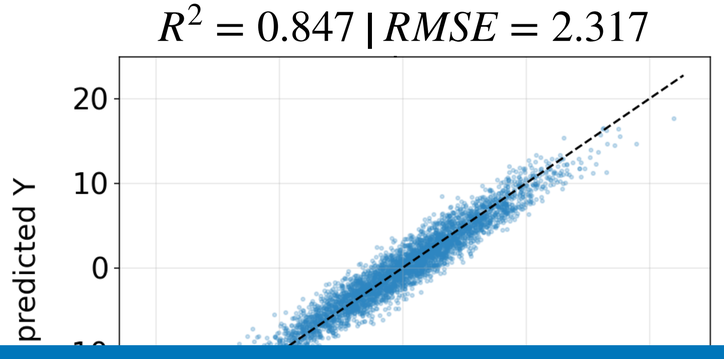
C: Hospital type

Higher drug doses (X) are associated with a greater elevation in SBP (Y), even after adjustment for $\{A, B, C\}$.

From Data to Predictive Models

Best of Random Forest, Gradient Boosting and Linear Regression





Y: Systolic blood pressure change

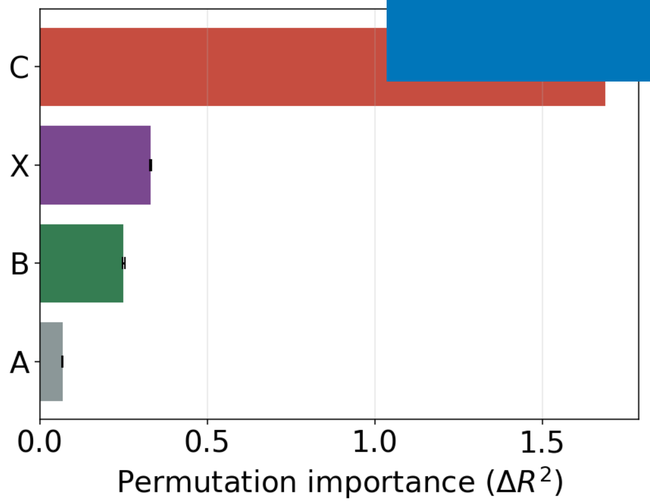
X: IV drug infusion rate (treatment)

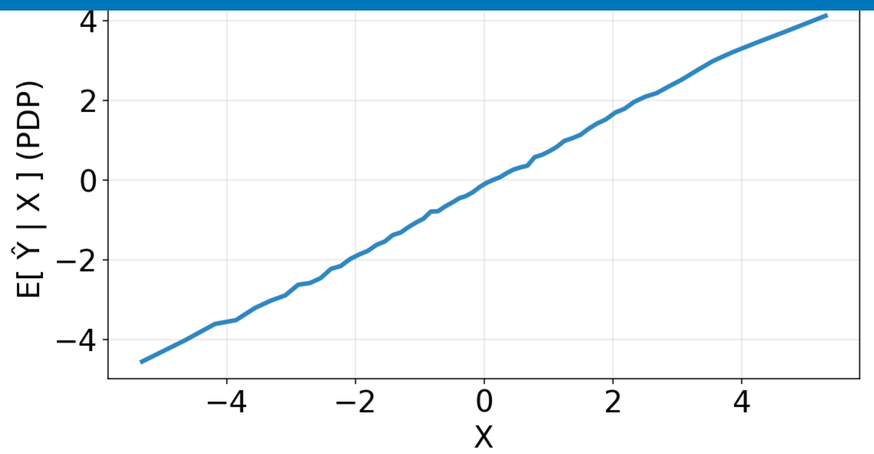
A: CYP450 pharmacogenomic variant (drug metabolism capacity)

B: Serum creatinine (kidney function marker)

C: Hospital type

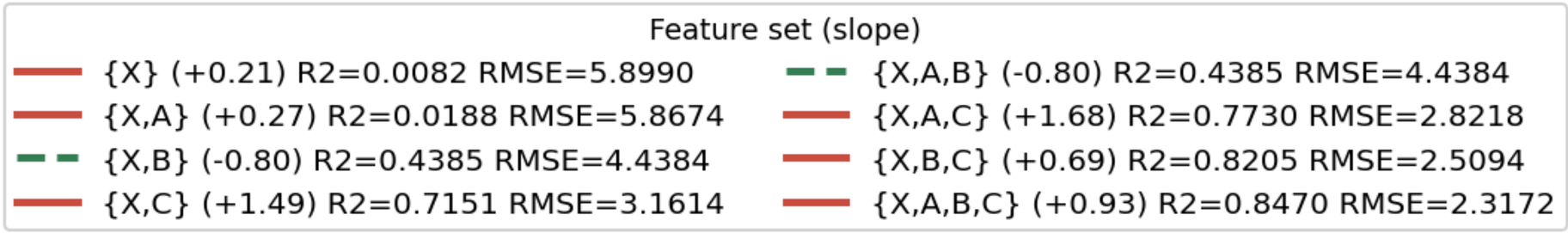
Would you advise clinicians to stop using this drug? Should clinical practice change?

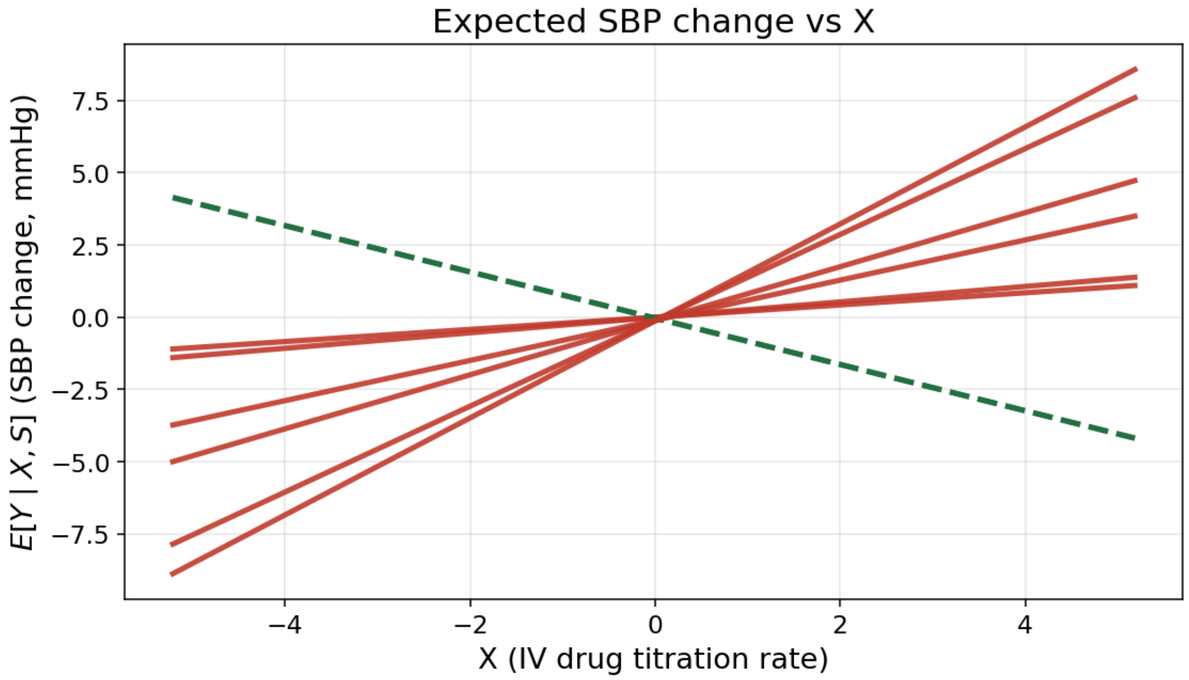


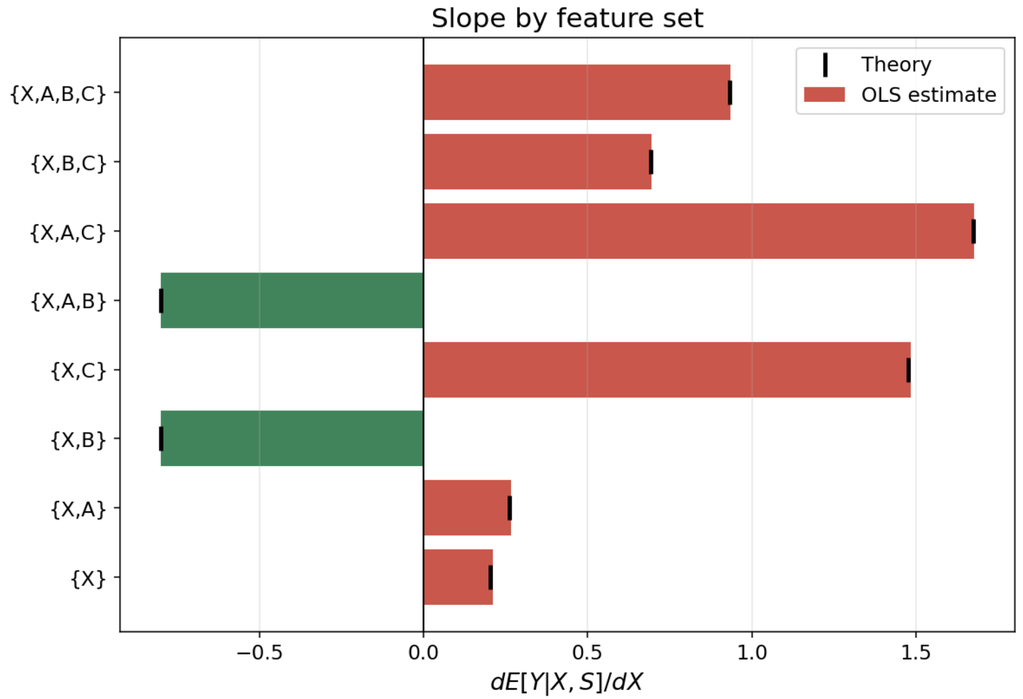


Higher drug doses (X) are associated with a greater elevation in SBP (Y), even after adjustment for $\{A, B, C\}$.

The Replication Crisis in Science







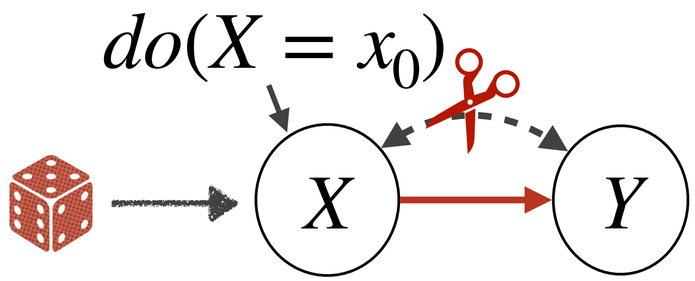
Why don't these analyses replicate? Should these estimates agree? Which one is the ***causal effect***? 🤷‍♀️

The Challenge of Causal Inference

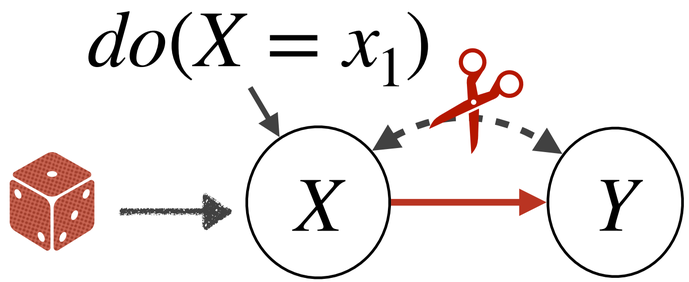
How can we extract ***reliable insights*** into underlying mechanisms?

: Randomized Experiments!

Randomization of the X's assignment



$$\mathbb{E}[Y \mid do(X = x_0)]$$



$$\mathbb{E}[Y \mid do(X = x_1)]$$

Average Causal Effect: $\mathbb{E}[Y \mid do(X = x_0)] - \mathbb{E}[Y \mid do(X = x_1)]$

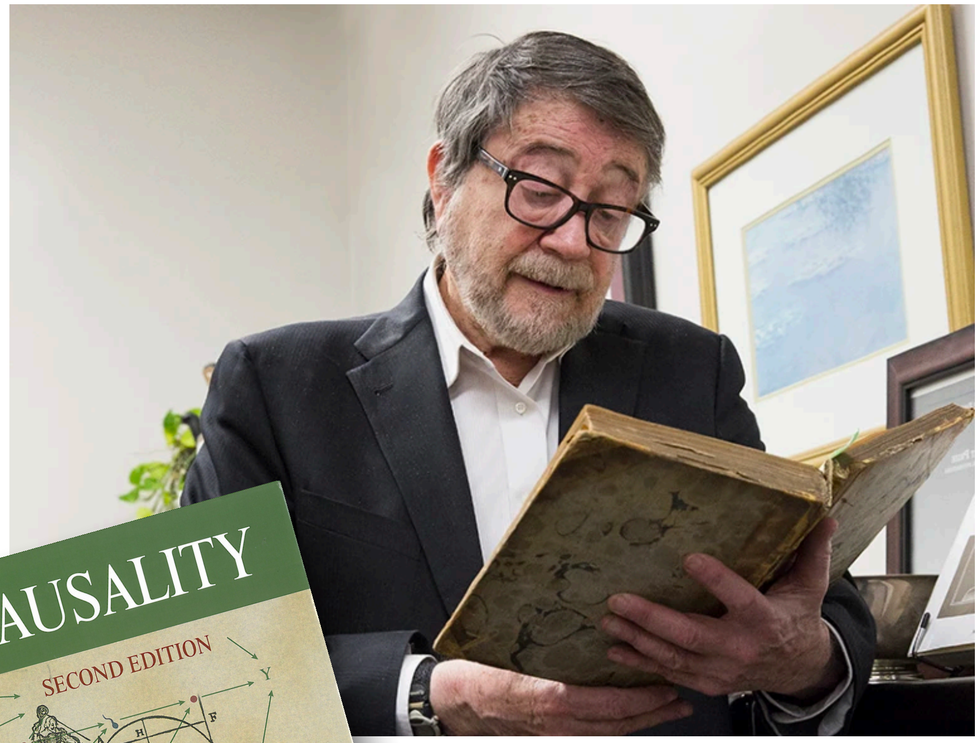
The Challenge of Causal Inference

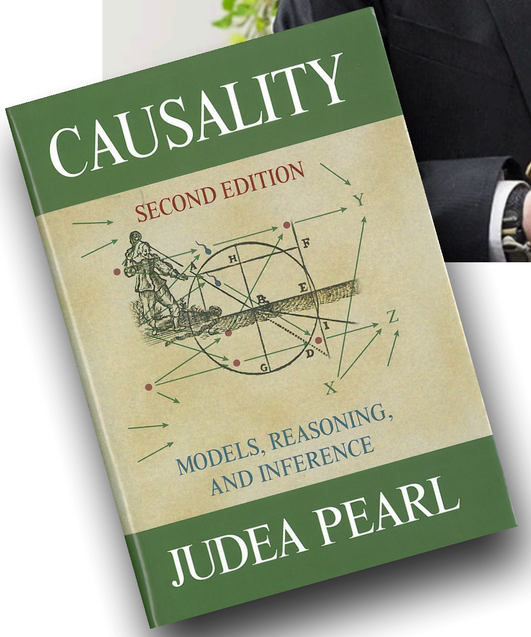
Randomized experiments are considered the gold standard for causal inference, but they have important limitations:

- May be costly, impractical, or ethically infeasible.
- Complete randomization and control are rarely possible in practice.
- Without careful targeting, many experiments are needed to gain a comprehensive understanding of the underlying system.

Can we instead leverage readily available ***observational*** data?

Causal Artificial Intelligence by Judea Pearl





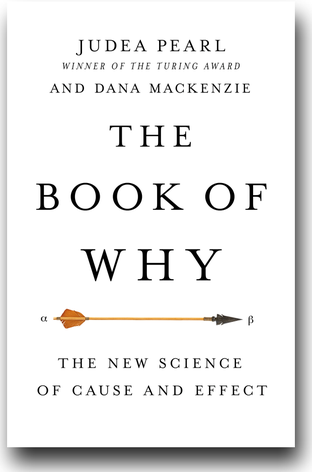
Director of the Cognitive Systems Laboratory at UCLA, received the 2011 A. M. Turing Award

“for fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning.”

— Association for Computing Machinery (ACM)

“Deep learning has instead given us machines with truly impressive abilities but no intelligence. The difference is profound and lies in the absence of a model of reality.”

— The Book of Why: The New Science of Cause and Effect



The Right Answer Depends on the Right Model

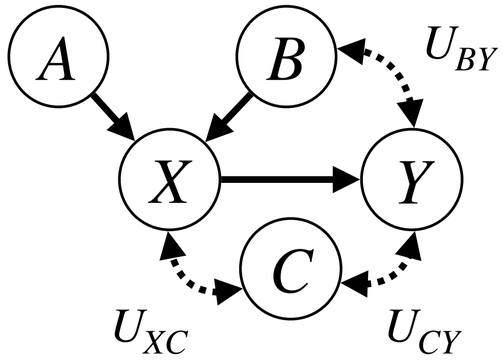
Data Generating Process

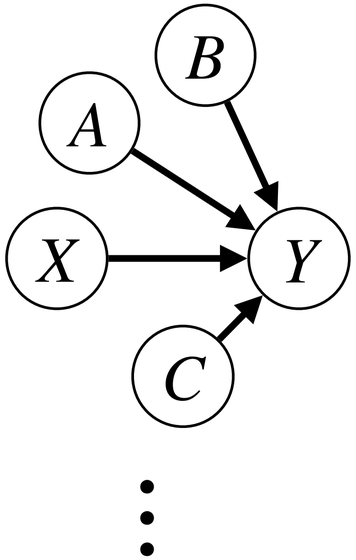
$$A = \varepsilon_A$$
$$X = 0.6\,A + 1.3\,U_{by} + 2.0\,U_{xc} + 0.4\,\varepsilon_X$$
$$B = 1.0\,U_{by} + 0.4\,\varepsilon_B$$
$$C = 1.5\,U_{xc} + 1.5\,U_{cy} + 0.4\,\varepsilon_C$$
$$Y = -0.8\,X + 5.0\,U_{by} - 4.0\,U_{cy} + 0.6\,\varepsilon_Y$$

Where $U_{xc}, U_{cy}, U_{by}, \varepsilon_X, \varepsilon_A, \varepsilon_B, \varepsilon_C, \varepsilon_Y \sim \mathcal{N}(0,1)$

Y: Systolic blood pressure change
X: IV drug infusion rate (treatment)
A: CYP450 pharmacogenomic variant
B: Serum creatinine (marker)
C: Hospital (Specialized vs. General)
U_{BY}: Chronic Kidney Disease
U_{CY}: Illness severity
U_{XC}: Socioeconomic/insurance tier

Causal Diagram





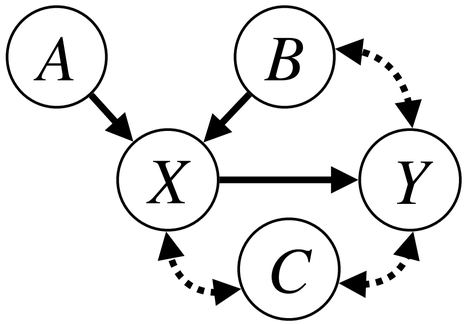
Many models are capable of approximating $P(Y \mid X, A, B, C)$, but they differ in their explanations!

Classical Causal Pipeline

1 **Query**

$$P(y \mid do(x))$$

2 **Causal Diagram**



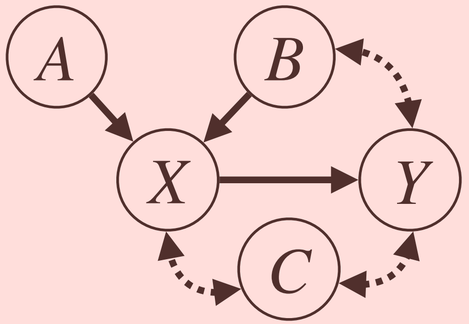
Classical Causal Pipeline

Symbolic system representation:
Requires structural causal knowledge for **every pair** of variables!

1 **Query**

$P(y \mid do(x))$

2 **Causal Diagram**



3 $P(x, y, a, b, c)$

Observational Distribution

Inference Engine

Solution

yes / no

$$P(y \mid do(x)) = \sum_b P(y \mid x, b)P(b)$$

$$\mathbb{E}(y \mid do(x)) = \sum_b \mathbb{E}(y \mid x, b)P(b)$$

Observational Data

$$\hat{\mathbb{E}}(y \mid do(x)) = \sum_b \hat{\mathbb{E}}((y \mid x, b)\hat{P}(b)$$

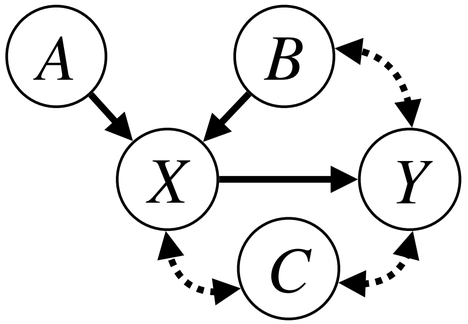
Interventional Distribution ← Available Distributions

Causal Modeling → Causal Effect Identification → Causal Effect Estimation

Pearl, J. 1995. Causal diagrams for empirical research. Biometrika, 82(4): 669–688.
Tian, J. and Pearl, J. (2002) A General Identification Condition for Causal Effects. In AAAI 2002, pp. 567–573, Menlo Park, CA.

The Causal Structure Explains Induced Biases

Causal Diagram



$$\hat{\mathbb{E}}(y \mid do(x)) = \sum_{\mathbf{z}} \hat{\mathbb{E}}((y \mid x, \mathbf{z})\hat{P}(\mathbf{z}) \qquad \text{with } \mathbf{Z} \in \{\{B\}, \{A, B\}\}$$

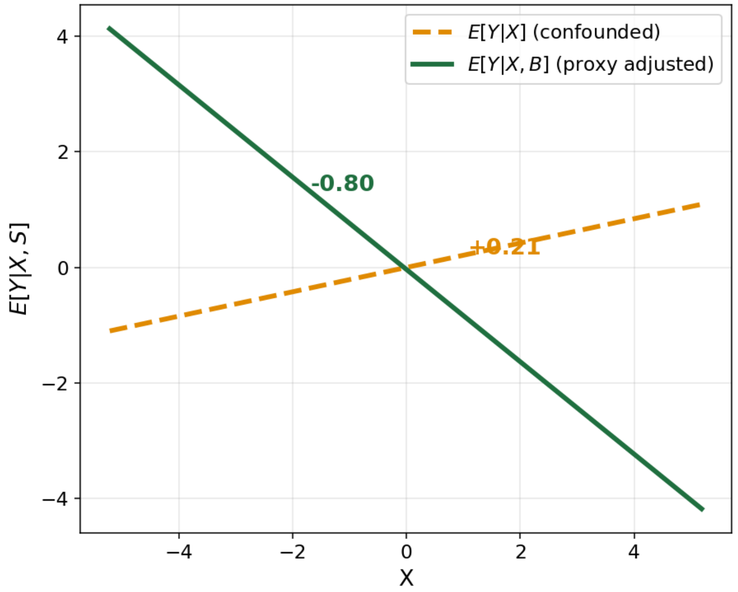
Adjusting for $\{B\}$or $\{A, B\}$ yields an
unbiased estimate of the underlying causal effect!

Increasing treatment dose by one unit is estimated to reduce SBP by about 0.77 mmHg on average.

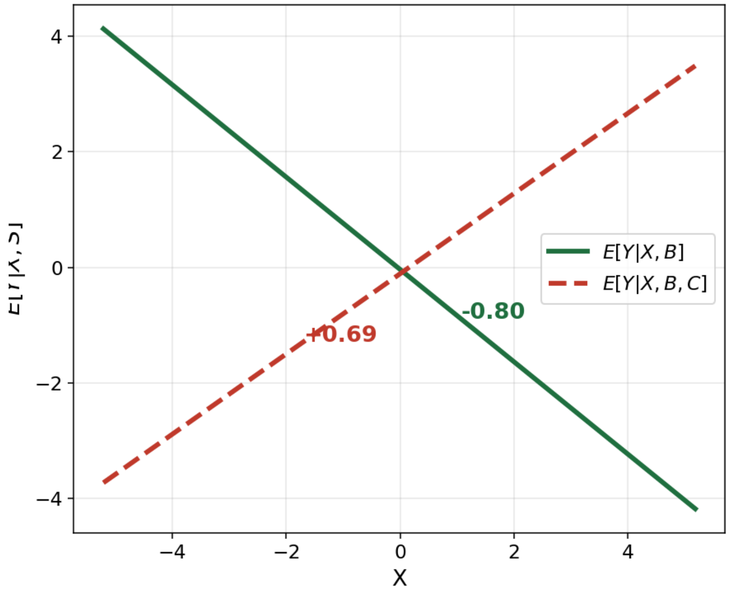
Simpson's paradox: Failing to adjust for a confounder can reverse an association.

Berkson's paradox:
Adjusting for a collider can create or reverse an association

Adjusting for B resolves Confounding



Adjusting for C induces Collider Bias



Classical Causal Pipeline

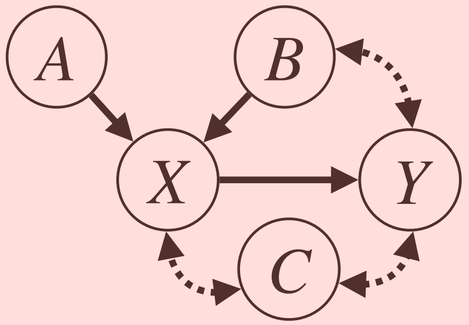
A fully specified causal diagram is rarely attainable in practice.

Can we relax such a requirement?

1 **Query**

$P(y \mid do(x))$

2 **Causal Diagram**



3

$P(x, y, a, b, c)$

Observational Distribution

Inference Engine

Solution

yes / no

$$P(y \mid do(x)) = \sum_b P(y \mid x, b)P(b)$$

$$\mathbb{E}(y \mid do(x)) = \sum_b \mathbb{E}(y \mid x, b)P(b) \rightarrow \hat{\mathbb{E}}(y \mid do(x)) = \sum_b \hat{\mathbb{E}}((y \mid x, b)\hat{P}(b)$$

Observational Data

Interventional Distribution ← Available Distributions

Causal Modeling → Causal Effect Identification → Causal Effect Estimation

Pearl, J. 1995. Causal diagrams for empirical research. Biometrika, 82(4): 669–688.

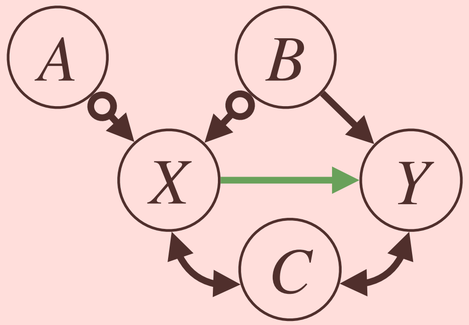
Tian, J. and Pearl, J. (2002) A General Identification Condition for Causal Effects. In AAAI 2002, pp. 567–573, Menlo Park, CA.

A Fully Data-Driven Causal Pipeline

1 **Query**

$P(y \mid do(x))$

2 **PAG**



3 $P(x, y, a, b, c)$

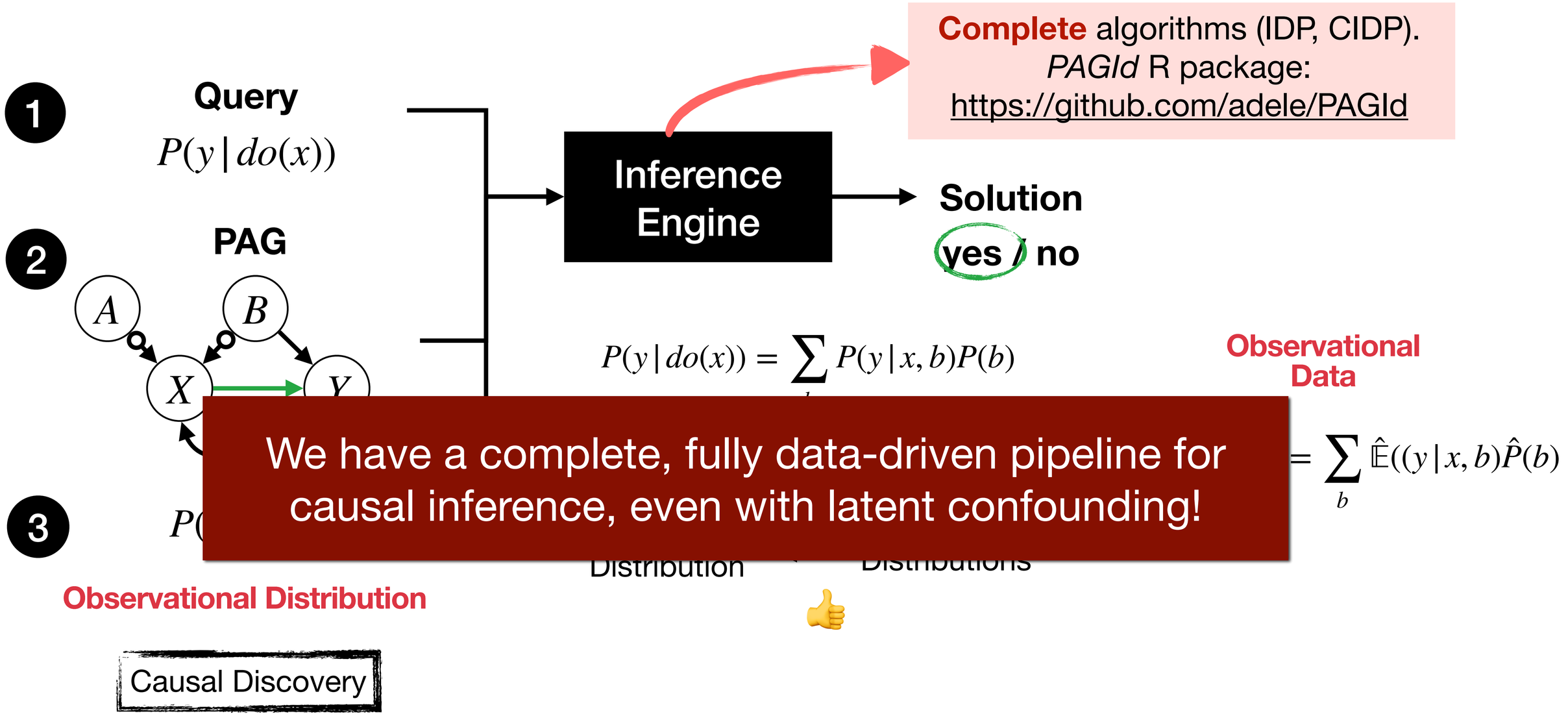
Observational Distribution

Causal Discovery

Can be constructed in a fully data-driven way!
(e.g. using the FCI algorithm)

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896. (Link)

A Fully Data-Driven Causal Pipeline



Jaber, A.*, **Ribeiro, A. H.***, Zhang, J., Bareinboim, E. (2022). Causal Identification under Markov equivalence: Calculus, Algorithm, and Completeness. *In Adv. Neural Inf Process Systems, 35, 3679-3690. (NeurIPS-22).* (Link)

Outline

Foundations

- Structural Causal Models
- Causal Bayesian Networks
- Pearl's Ladder of Causation

Foundations of Causal Discovery

- Bayesian Networks
- Conditional Independence & d-Separation
- Markov Equivalence Classes

Causal Discovery

- Constraint-Based Methods
- Conditional Independence Testing
- Fast Causal Inference (FCI)
- Recent Advances

Causal Effect Identification:

- Covariate Adjustment from Causal Graphs
- The ID Algorithm
- Effect Identification from Partial Ancestral Graphs (PAGs)

Structural Causal Model (SCM)

The true model behind the data

Full explainability for all layers of the causal hierarchy

Structural Causal Model (SCM)

Definition: A structural causal model $\mathscr{M}$ (or, data generating model) is a tuple $\langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$, where

- $\mathbf{V} = \{V_1, \ldots, V_n\}$: are endogenous variables
- $\mathbf{U} = \{U_1, \ldots, U_m\}$: are exogenous variables
- $\mathscr{F} = \{f_1, \ldots, f_n\}$: are functions determining $\mathbf{V}$, i.e., $v_i \leftarrow f_i(pa_i, u_i)$, where
  - $Pa_i \subseteq \mathbf{V}$ are endogenous causes (parents) of V_i
  - $U_i \subseteq \mathbf{U}$ are exogenous causes of V_i.
- $P(\mathbf{U})$ is the probability distribution over $\mathbf{U}$.

Assumption: $\mathscr{M}$ is recursive, i.e., there are no feedback (cyclic) mechanisms.

Causal (Intervention) Effects from SCMs

Pre-Interventional/ Observational SCM

$$\mathscr{M} = \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_{XY}, U_X, U_Y\} \\ \mathscr{F} = \begin{cases} X = f_X(U_X, U_{XY}) \\ Y = f_Y(X, U_Y, U_{XY}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

$do(X = x)$

Post-Interventional / Interventional SCM

$$\mathscr{M}_x = \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_{XY}, U_X, U_Y\} \\ \mathscr{F} = \begin{cases} X = x \\ Y = f_Y(x, U_Y, U_{XY}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Observational Distribution

$$P(\mathbf{V}) \doteq P_{\mathscr{M}}(\mathbf{V})$$

$\neq$

Interventional Distribution

$$P(\mathbf{V} \mid do(X = x)) \doteq P_{\mathscr{M}_x}(\mathbf{V})$$

Can we ***predict*** better the value of Y after ***observing*** that $X = x$?

$P(Y = y \mid X = x) \neq P(Y = y) \implies X$ is ***associated*** to Y

Can we ***predict*** better the value of Y after ***making an intervention*** $do(X = x)$?

$\exists x$ s.t. $P_{\mathscr{M}_x}(Y = y) \neq P(Y = y) \implies X$ is ***a cause*** of Y

SCM: Encoder of Functional Knowledge

The knowledge required to fully specify an SCM is usually *unavailable* in practice.

Randomized experiments are often costly, unfeasible, or unethical.

Is it possible to identify the effect of interventions from *observational* data without fully specifying the SCM (i.e., in a non-parametric fashion)?

Yes, with structural knowledge encoded as a causal diagram!

Causal Bayesian Network

A directed acyclic graph (augmented with latent variables) that encodes:

- All **conditional (in)dependencies** among the observed variables and
- the **causal structure** implied by an SCM

CBN: Encoder of Structural Causal Knowledge

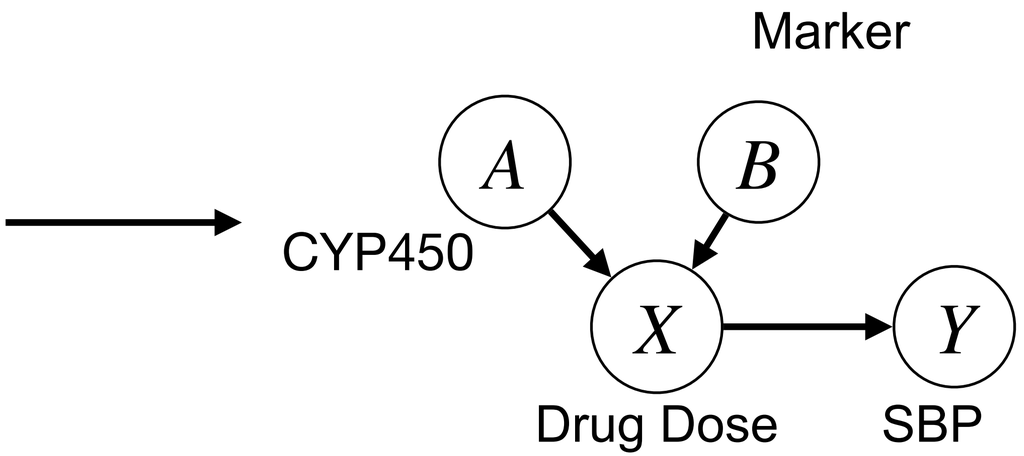
Structural Causal Model (SCM)

$$\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$$

$$\mathscr{M} = \begin{cases} \mathbf{V} = \{X, Y, A, B\} \\ \mathbf{U} = \{U_A, U_B, U_X, U_Y, U_{BY}\} \\ \mathscr{F} = \begin{cases} A \leftarrow f_A(U_A) \\ B \leftarrow f_B(U_B, U_{BY}) \\ X \leftarrow f_X(A, B, U_X) \\ Y \leftarrow f_X(X, U_Y, U_{BY}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Induced Causal Bayesian Network (CBN)

Causal Diagram



An SCM $\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$ induces a causal diagram such that, **for every** $V_i, V_j \in \mathbf{V}$:

$V_i \rightarrow V_j$, if V_i appears as argument of $f_j \in \mathscr{F}$.

CBN: Encoder of Structural Causal Knowledge

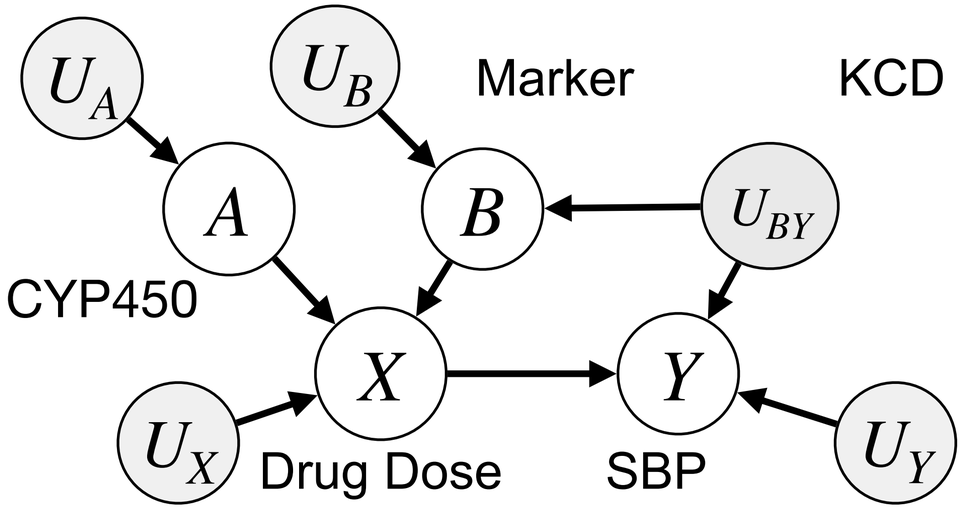
Structural Causal Model (SCM)

$$\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$$

$$\mathscr{M} = \begin{cases} \mathbf{V} = \{X, Y, A, B\} \\ \mathbf{U} = \{U_A, U_B, U_X, U_Y, U_{BY}\} \\ \mathscr{F} = \begin{cases} A \leftarrow f_A(U_A) \\ B \leftarrow f_B(U_B, U_{BY}) \\ X \leftarrow f_X(A, B, U_X) \\ Y \leftarrow f_X(X, U_Y, U_{BY}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Induced Causal Bayesian Network (CBN)

Causal Diagram



An SCM $\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$ induces a causal diagram such that, **for every** $V_i, V_j \in \mathbf{V}$:

$V_i \rightarrow V_j$, if V_i appears as argument of $f_j \in \mathscr{F}$.

CBN: Encoder of Structural Causal Knowledge

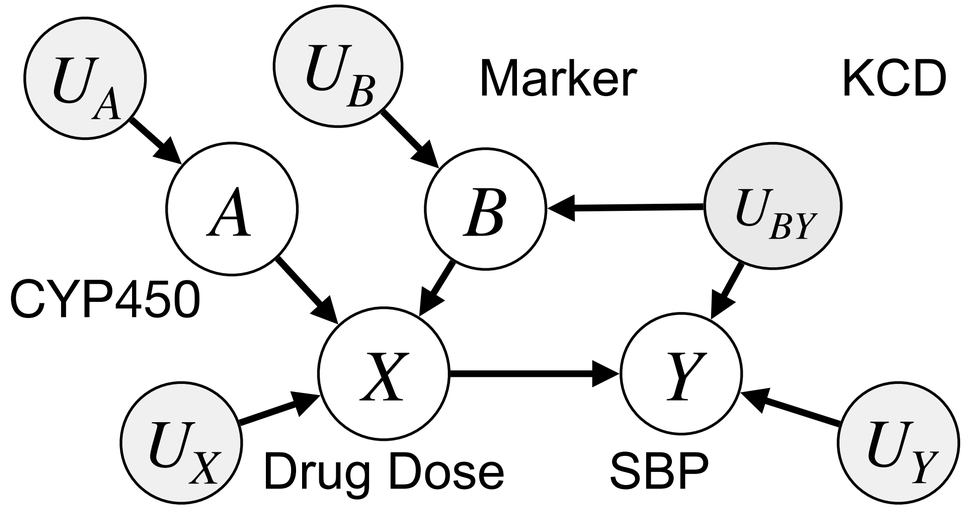
Structural Causal Model (SCM)

$$\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$$

$$\mathscr{M} = \begin{cases} \mathbf{V} = \{X, Y, A, B\} \\ \mathbf{U} = \{U_A, U_B, U_X, U_Y, U_{BY}\} \\ \mathscr{F} = \begin{cases} A \leftarrow f_A(U_A) \\ B \leftarrow f_B(U_B, U_{BY}) \\ X \leftarrow f_X(A, B, U_X) \\ Y \leftarrow f_X(X, U_Y, U_{BY}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Induced Causal Bayesian Network (CBN)

Causal Diagram



An SCM $\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$ induces a causal diagram such that, **for every** $V_i, V_j \in \mathbf{V}$:

$V_i \rightarrow V_j$, if V_i appears as argument of $f_j \in \mathscr{F}$.

$V_i \dashleftarrow\dashrightarrow V_j$ if the corresponding $U_i, U_j \in \mathbf{U}$ are correlated or f_i , f_j share some argument $U \in \mathbf{U}$.

CBN: Encoder of Structural Causal Knowledge

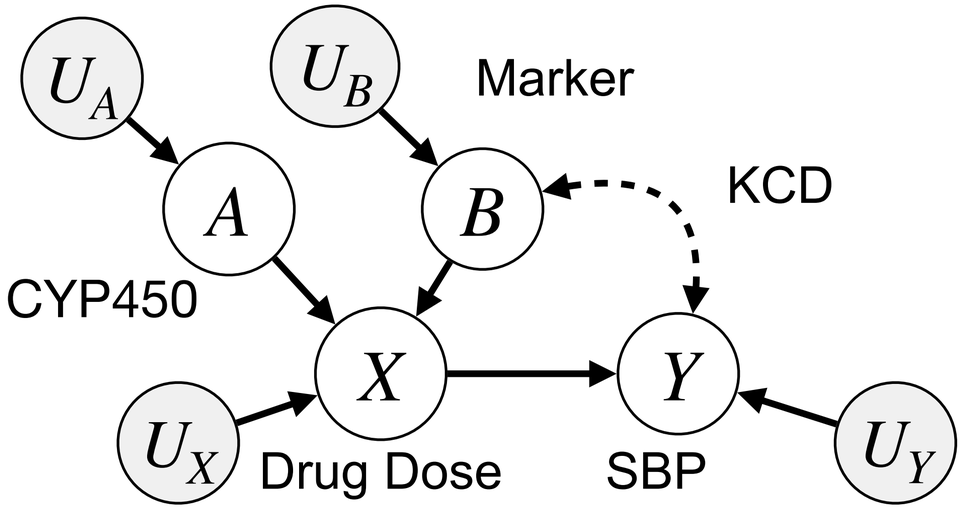
Structural Causal Model (SCM)

$$\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$$

$$\mathscr{M} = \begin{cases} \mathbf{V} = \{X, Y, A, B\} \\ \mathbf{U} = \{U_A, U_B, U_X, U_Y, U_{BY}\} \\ \mathscr{F} = \begin{cases} A \leftarrow f_A(U_A) \\ B \leftarrow f_B(U_B, U_{BY}) \\ X \leftarrow f_X(A, B, U_X) \\ Y \leftarrow f_X(X, U_Y, U_{BY}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Induced Causal Bayesian Network (CBN)

Causal Diagram



An SCM $\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$ induces a causal diagram such that, **for every** $V_i, V_j \in \mathbf{V}$:

$V_i \rightarrow V_j$, if V_i appears as argument of $f_j \in \mathscr{F}$.

$V_i \dashleftarrow\dashrightarrow V_j$ if the corresponding $U_i, U_j \in \mathbf{U}$ are correlated or f_i , f_j share some argument $U \in \mathbf{U}$.

CBN: Encoder of Structural Causal Knowledge

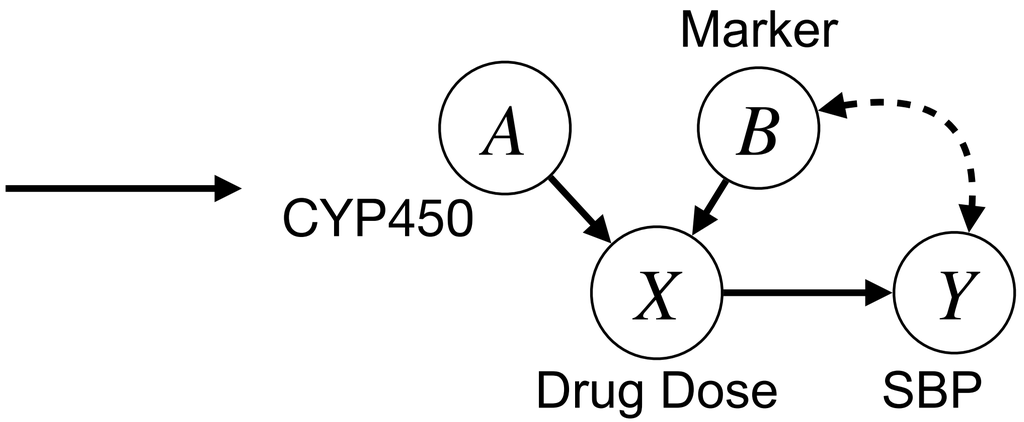
Structural Causal Model (SCM)

$$\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$$

$$\mathscr{M} = \begin{cases} \mathbf{V} = \{X, Y, A, B\} \\ \mathbf{U} = \{U_A, U_B, U_X, U_Y, U_{BY}\} \\ \mathscr{F} = \begin{cases} A \leftarrow f_A(U_A) \\ B \leftarrow f_B(U_B, U_{BY}) \\ X \leftarrow f_X(A, B, U_X) \\ Y \leftarrow f_X(X, U_Y, U_{BY}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Induced Causal Bayesian Network (CBN)

Causal Diagram



An SCM $\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$ induces a causal diagram such that, **for every** $V_i, V_j \in \mathbf{V}$:

$V_i \rightarrow V_j$, if V_i appears as argument of $f_j \in \mathscr{F}$.

$V_i \dashleftarrow\dashrightarrow V_j$ if the corresponding $U_i, U_j \in \mathbf{U}$ are correlated or f_i , f_j share some argument $U \in \mathbf{U}$.

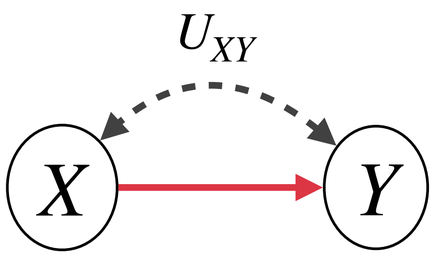
On the causal model idenfiability

What is induced by the SCM?

Observational SCM

$$\mathcal{M} = \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_{XY}, U_X, U_Y\} \\ \mathcal{F} = \begin{cases} X = f_X(U_X, U_{XY}) \\ Y = f_Y(X, U_Y, U_{XY}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Loss of Information



Observational Causal Diagram

$P_{\mathcal{M}}(\mathbf{V})$

Observational Distribution

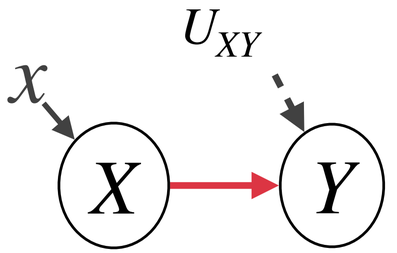
Observational Data

$do(X = x)$

Interventional SCM

$$\mathcal{M}_x = \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_{XY}, U_X, U_Y\} \\ \mathcal{F} = \begin{cases} X = x \\ Y = f_Y(x, U_Y, U_{XY}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Loss of Information



Interventional Causal Diagram

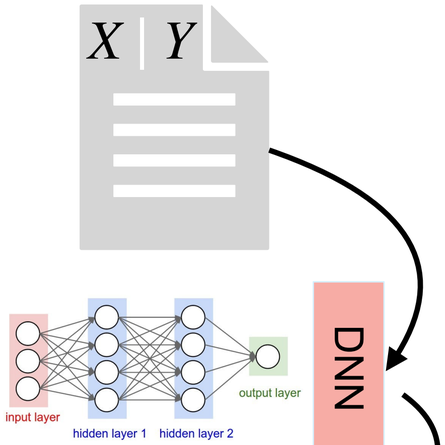
$P_{\mathcal{M}_x}(\mathbf{V}) \doteq P(\mathbf{V} \mid do(x))$

Interventional Distribution

Interventional Data

Correlation does not imply causation!

Observational



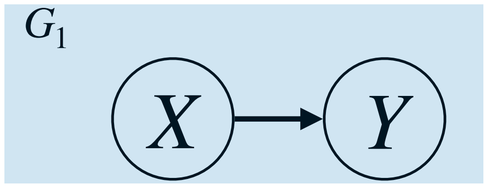
$P(Y \mid X = x)$

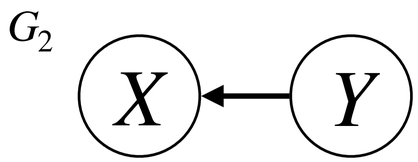
X: Treatment

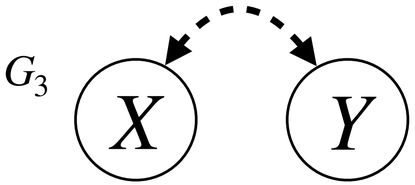
Y: Outcome

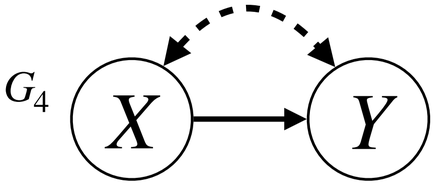
$P(Y \mid X = x) \neq P(Y)$

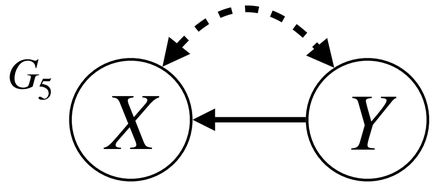
$\Rightarrow X$ improves prediction of Y











$\mathscr{M}_{11} = \langle \mathbf{V}, \mathbf{U}_1, \mathscr{F}_{11}, P_{11}(\mathbf{u}_1) \rangle$

$\mathscr{M}_{1k_1} = \langle \mathbf{V}, \mathbf{U}_1, \mathscr{F}_{1k_1}, P_{1k_1}(\mathbf{u}_1) \rangle$ Fitted Model

$\mathscr{M}_{21} = \langle \mathbf{V}, \mathbf{U}_2, \mathscr{F}_{21}, P_{21}(\mathbf{u}_2) \rangle$

$\mathscr{M}_{2k_2} = \langle \mathbf{V}, \mathbf{U}_2, \mathscr{F}_{2k_2}, P_{2k_2}(\mathbf{u}_2) \rangle$

$\mathscr{M}_{31} = \langle \mathbf{V}, \mathbf{U}_3, \mathscr{F}_{31}, P_{31}(\mathbf{u}_3) \rangle$

$\mathscr{M}_{3k_3} = \langle \mathbf{V}, \mathbf{U}_3, \mathscr{F}_{3k_3}, P_{3k_3}(\mathbf{u}_3) \rangle$

$\mathscr{M}_{41} = \langle \mathbf{V}, \mathbf{U}_4, \mathscr{F}_{41}, P_{41}(\mathbf{u}_4) \rangle$

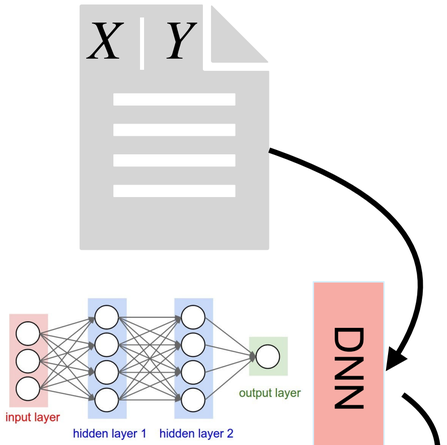
$\mathscr{M}_{4k_4} = \langle \mathbf{V}, \mathbf{U}_4, \mathscr{F}_{4k_4}, P_{4k_4}(\mathbf{u}_4) \rangle$

$\mathscr{M}_{51} = \langle \mathbf{V}, \mathbf{U}_5, \mathscr{F}_{51}, P_{51}(\mathbf{u}_5) \rangle$

$\mathscr{M}_{5k_5} = \langle \mathbf{V}, \mathbf{U}_5, \mathscr{F}_{5k_5}, P_{5k_5}(\mathbf{u}_5) \rangle$

Correlation does not imply causation!

Observational



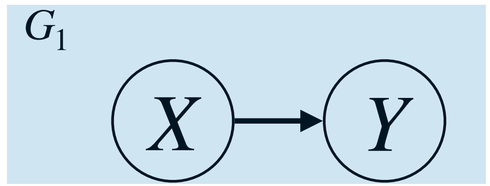
$P(Y \mid X = x)$

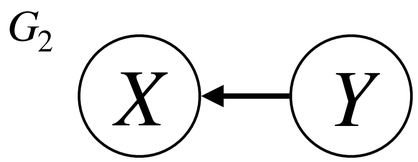
X: Treatment

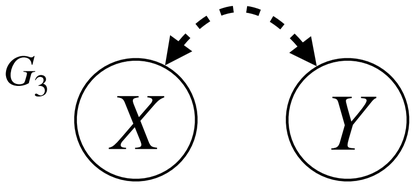
Y: Outcome

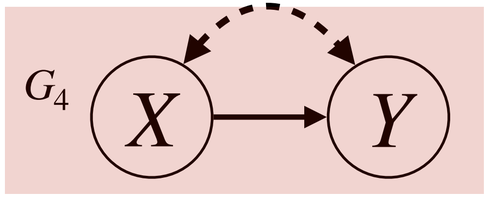
$P(Y \mid X = x) \neq P(Y)$

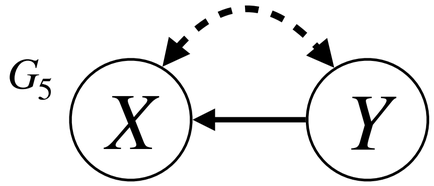
$\Rightarrow X$ improves prediction of Y











$\mathscr{M}_{11} = \langle \mathbf{V}, \mathbf{U}_1, \mathscr{F}_{11}, P_{11}(\mathbf{u}_1) \rangle$

$\mathscr{M}_{1k_1} = \langle \mathbf{V}, \mathbf{U}_1, \mathscr{F}_{1k_1}, P_{1k_1}(\mathbf{u}_1) \rangle$ Fitted Model

$\mathscr{M}_{21} = \langle \mathbf{V}, \mathbf{U}_2, \mathscr{F}_{21}, P_{21}(\mathbf{u}_2) \rangle$

$\mathscr{M}_{2k_2} = \langle \mathbf{V}, \mathbf{U}_2, \mathscr{F}_{2k_2}, P_{2k_2}(\mathbf{u}_2) \rangle$

$\mathscr{M}_{31} = \langle \mathbf{V}, \mathbf{U}_3, \mathscr{F}_{31}, P_{31}(\mathbf{u}_3) \rangle$

$\mathscr{M}_{3k_3} = \langle \mathbf{V}, \mathbf{U}_3, \mathscr{F}_{3k_3}, P_{3k_3}(\mathbf{u}_3) \rangle$

$\mathscr{M}_{41} = \langle \mathbf{V}, \mathbf{U}_4, \mathscr{F}_{41}, P_{41}(\mathbf{u}_4) \rangle$

$\mathscr{M}_{4k_4} = \langle \mathbf{V}, \mathbf{U}_4, \mathscr{F}_{4k_4}, P_{4k_4}(\mathbf{u}_4) \rangle$ True Model

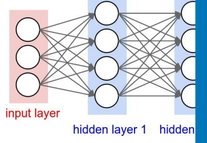
$\mathscr{M}_{51} = \langle \mathbf{V}, \mathbf{U}_5, \mathscr{F}_{51}, P_{51}(\mathbf{u}_5) \rangle$

$\mathscr{M}_{5k_5} = \langle \mathbf{V}, \mathbf{U}_5, \mathscr{F}_{5k_5}, P_{5k_5}(\mathbf{u}_5) \rangle$

Correlation does not imply causation!

Observational

$X \mid Y$



$P(Y \mid X$

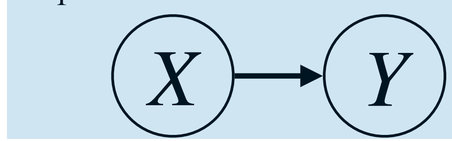
X: Treat

Y: Outcome

$P(Y \mid X = x) \neq P(Y)$

$\Rightarrow X$ improves prediction of Y

G_1



$\mathscr{M}_{11} = \langle \mathbf{V}, \mathbf{U}_1, \mathscr{F}_{11}, P_{11}(\mathbf{u}_1) \rangle$

$\vdots$

$\mathscr{M}_{1k_1} = \langle \mathbf{V}, \mathbf{U}_1, \mathscr{F}_{1k_1}, P_{1k_1}(\mathbf{u}_1) \rangle$ Fitted Model

$(\mathbf{u}_2)\rangle$

$_{k_2}(\mathbf{u}_2)\rangle$

$(\mathbf{u}_3)\rangle$

$_{k_3}(\mathbf{u}_3)\rangle$

$(\mathbf{u}_4)\rangle$

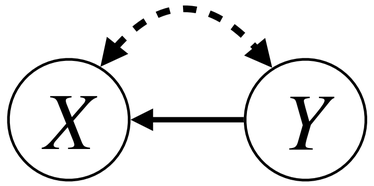
$_{4k_4}(\mathbf{u}_4)\rangle$ True Model

Model Explainability Tools:

- **Feature Importance:** SHAP, LIME, Permutation Importance
- **Visualization Techniques:** Partial Dependence Plots, Saliency Maps
- **Counterfactuals Explanations:** What-If Analysis
- **Model Probing:** Ablation Studies, Sensitivity Analysis

Model Explanations ≠ Reality Explanations

G_5



$\mathscr{M}_{51} = \langle \mathbf{V}, \mathbf{U}_5, \mathscr{F}_{51}, P_{51}(\mathbf{u}_5) \rangle$

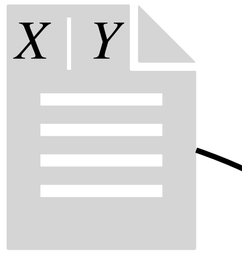
$\vdots$

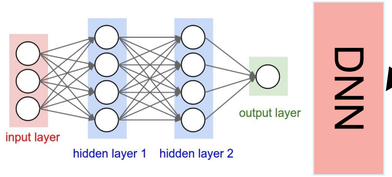
$\mathscr{M}_{5k_5} = \langle \mathbf{V}, \mathbf{U}_5, \mathscr{F}_{5k_5}, P_{5k_5}(\mathbf{u}_5) \rangle$

Correlation does not imply causation!

Causally Sufficient Models

Observational





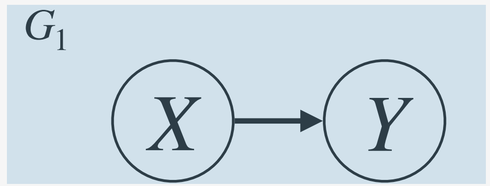
$P(Y|$

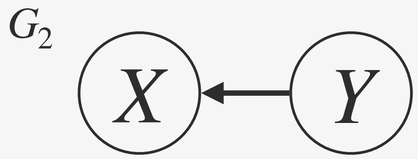
X: Tre

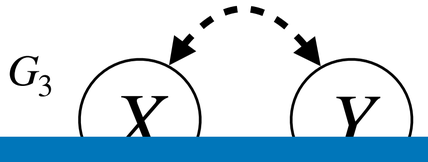
Y: Outcome

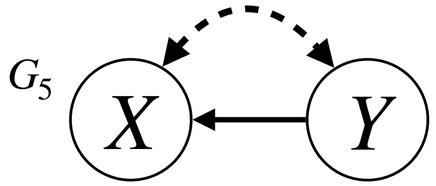
$$P(Y|X=x) \neq P(Y)$$

$\Rightarrow X$ improves prediction of Y









$\mathscr{M}_{11} = \langle \mathbf{V}, \mathbf{U}_1, \mathscr{F}_{11}, P_{11}(\mathbf{u}_1) \rangle$

$\mathscr{M}_{1k_1} = \langle \mathbf{V}, \mathbf{U}_1, \mathscr{F}_{1k_1}, P_{1k_1}(\mathbf{u}_1) \rangle$ Fitted Model

$\mathscr{M}_{21} = \langle \mathbf{V}, \mathbf{U}_2, \mathscr{F}_{21}, P_{21}(\mathbf{u}_2) \rangle$

$\mathscr{M}_{2k_2} = \langle \mathbf{V}, \mathbf{U}_2, \mathscr{F}_{2k_2}, P_{2k_2}(\mathbf{u}_2) \rangle$

$\mathscr{M}_{31} = \langle \mathbf{V}, \mathbf{U}_3, \mathscr{F}_{31}, P_{31}(\mathbf{u}_3) \rangle$

$P_{3k_3}(\mathbf{u}_3) \rangle$

$_{41}(\mathbf{u}_4) \rangle$

$P_{4k_4}(\mathbf{u}_4) \rangle$ True Model

$\mathscr{M}_{51} = \langle \mathbf{V}, \mathbf{U}_5, \mathscr{F}_{51}, P_{51}(\mathbf{u}_5) \rangle$

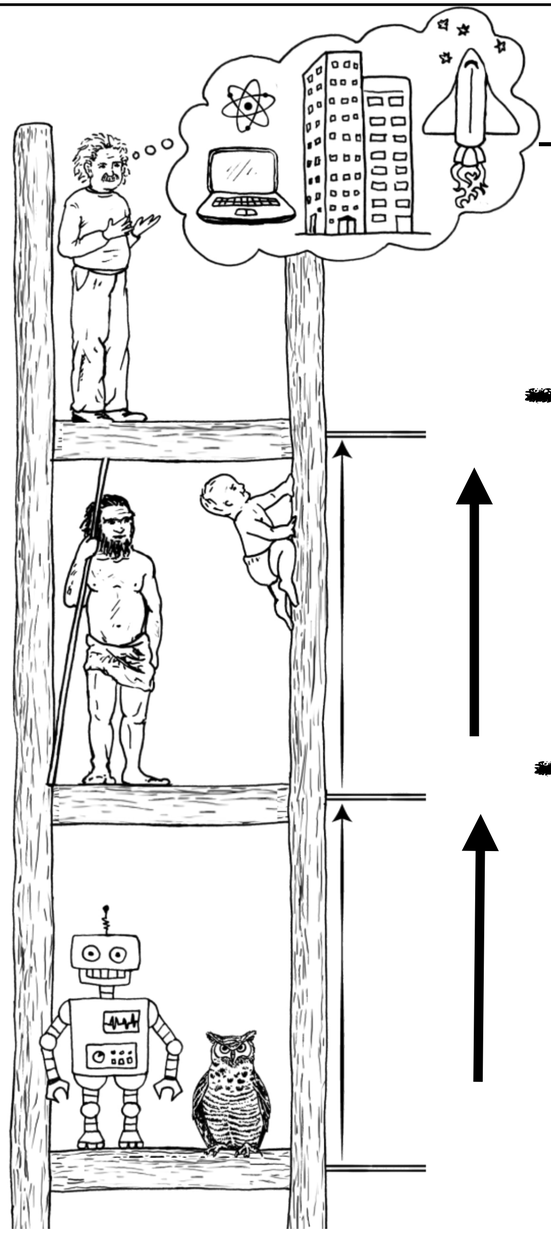
$\mathscr{M}_{5k_5} = \langle \mathbf{V}, \mathbf{U}_5, \mathscr{F}_{5k_5}, P_{5k_5}(\mathbf{u}_5) \rangle$

Non-identifiability even in the absence of lantent confounding and assuming known causal structure!

Pearl's Ladder of Causation

The Three Inferential Layers

Pearl's Ladder of Causation

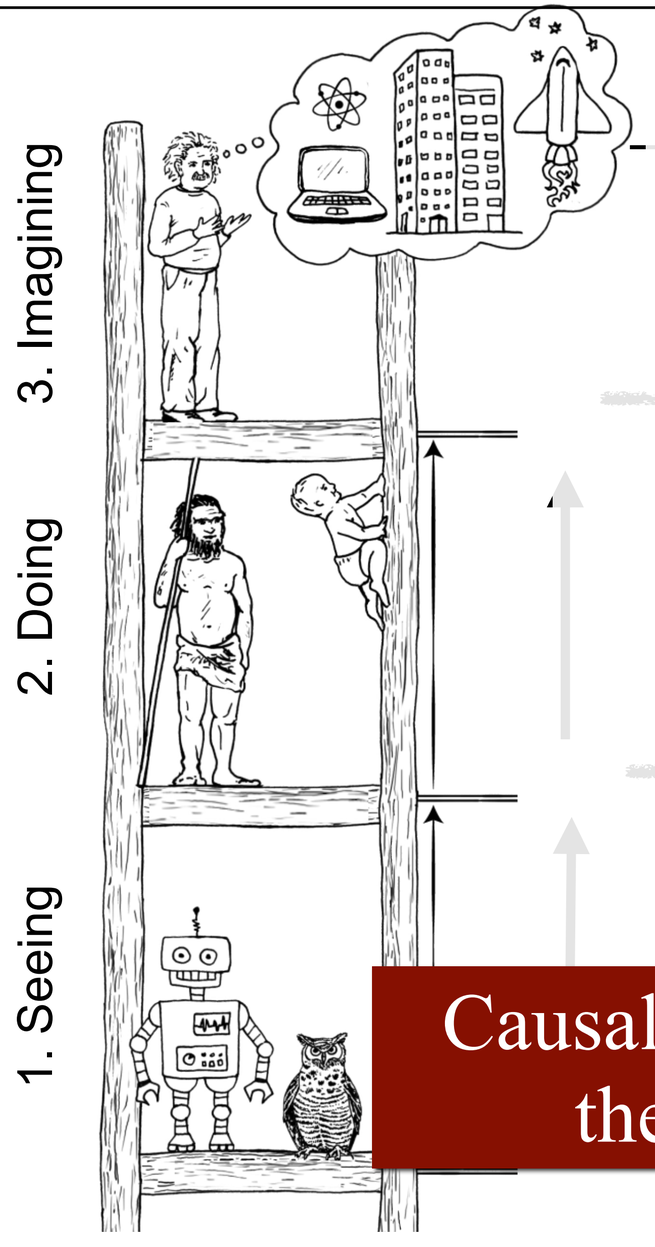


3. Imagining
2. Doing
1. Seeing

Layer	Task / Language	Typical Question	Examples
Counterfactual $P(y_x \mid x', y')$	Structural Causal Model / Deep SCMs	What would have happened to *it*, had *it* acted differently?	Was it the aspirin that stopped *my* headache?
Interventional $P(y \mid do(x), c)$	ML- Reinforcement Causal Effect Estimation	On average, what happens to Y if I intervene on X?	Will a headache be cured by an aspirin?
Associational $P(y \mid x)$	ML- (Un)Supervised (Decision trees, Deep neural net, LLMs/Transformers)	What does observing X tell me about Y	What does a symptom tell us about the disease?

* The Book of Why

Pearl's Ladder of Causation



Cross-layer inferences:

Doing — most of the inferences are about causal effects (policies, treatments, decisions)

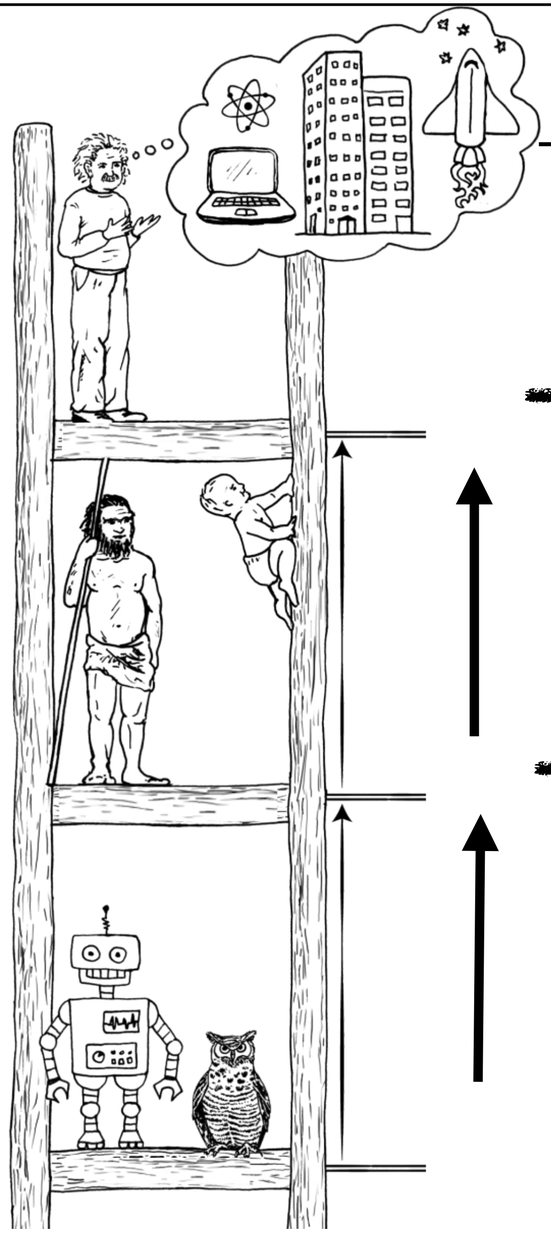
Seeing — most of the available data is observational, passively collected

Causal Hierarchy Theorem : For almost any SCM, the layers of the hierarchy remain distinct.

Book of Why & On Pearl's Hierarchy and the Foundations of Causal Inference, E. Bareinboim, J. Correa, D. Ibeling, T. Icard, in press. https://causalai.net/r60.pdf

Ladder of Causation / Intelligence

From Eric Xing's lecture



Philosophical Intelligence

3. Imagining

Physical Intelligence

2. Doing

Book Intelligence

1. Seeing

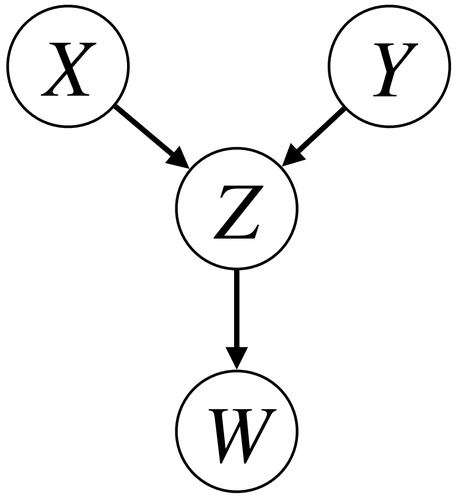
Layer	Task	What do you need?
Counterfactual $P(y_x \mid x', y')$	Digital twins, counterfactual fairness, root-cause analysis	Full abduction → action → prediction, **AND** identifiability of the mechanism (known DAG, provable monotonicity)*
Interventional $P(y \mid do(x), c)$	(Conditional) Population-level rollout / simulation	A causal graph and an *identifiable* causal effect
Associational $P(y \mid x)$	Prediction, pattern recognition, data imputation (MAR & MCAR)	Just the joint probability distribution, P(V)

* Nasr-Esfahany, A., & Kiciman, E. (2023). Counterfactual (non-) identifiability of learned structural causal models. *arXiv preprint arXiv:2301.09031*.

Bayesian Network

A DAG, possibly with latent confounders (ADMG),
representing the **joint distribution / conditional independences**
implied by an SCM

Graphical Kinship Notation



Directed Acyclic Graph (DAG)

X and Y are parents of Z, i.e., $X, Y \in Pa(Z)$

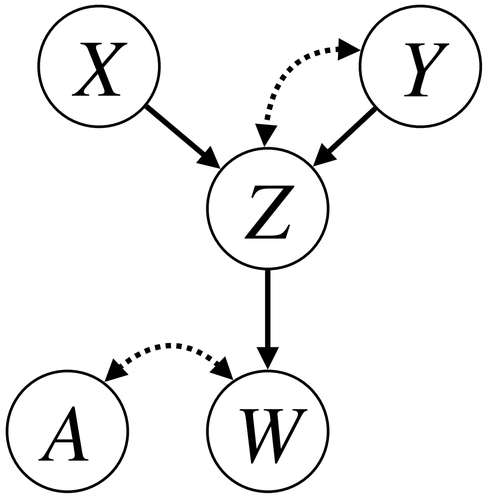
Z is a child of Y, i.e., $Z \in Ch(Y)$

W is a descendent of X, i.e., $W \in De(X)$

Y is ancestor of W, i.e., $Y \in An(W)$

Y is non-descendant of X, i.e., $Y \in NDesc(X)$

Graphical Kinship Notation



Acyclic Directed Mixed Graph (ADMG)

X and Y are parents of Z, i.e., $X, Y \in Pa(Z)$

Z is a child of Y, i.e., $Z \in Ch(Y)$

W is a descendent of X, i.e., $W \in De(X)$

Y is ancestor of W, i.e., $Y \in An(W)$

Y is non-descendant of X, i.e., $Y \in NDesc(X)$

A is spouse of W

Bayesian Networks & Markov Condition

A DAG G over $\mathbf{V}$ is a *Bayesian Network* for a joint probability distribution $P(\mathbf{V})$ if, for every $V_i \in \mathbf{V}$, it holds that $V_i \perp\!\!\!\perp NDesc_i \,|\, Pa_i$ and, therefore, $P(\mathbf{v})$ factorizes as follows:

P is satisfies the ***Markov Condition*** *w.r.t.* G

$$P(\mathbf{v}) = \prod_{V_i \in \mathbf{V}} P(v_i \,|\, v_{i-1}, \ldots, v_1)$$

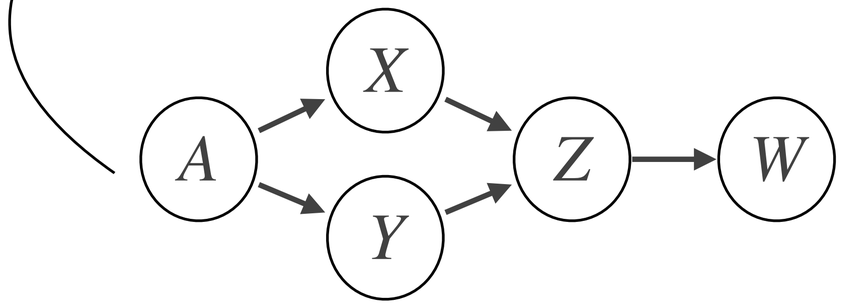
Chain Rule:

It holds for any topological order of G

$$= \prod_{V_i \in \mathbf{V}} P(v_i \,|\, pa_i)$$

$V_i \perp\!\!\!\perp NDesc_i \,|\, Pa_i, U_i$

Edges have no causal semantics!



$$P(\mathbf{v}) = P(w \,|\, z, x, y, a)\; P(z \,|\, x, y, a)\; P(x \,|\, y, a)\; P(y \,|\, a)\; P(a)$$
$$= P(w \,|\, z)\; P(z \,|\, x, y)\; P(x \,|\, a)\; P(y \,|\, a)\; P(a)$$

$W \perp\!\!\!\perp X, Y, A \,|\, Z$ $\quad$ $A \perp\!\!\!\perp Z \,|\, X, Y$ $\quad$ $Y \perp\!\!\!\perp X \,|\, A$

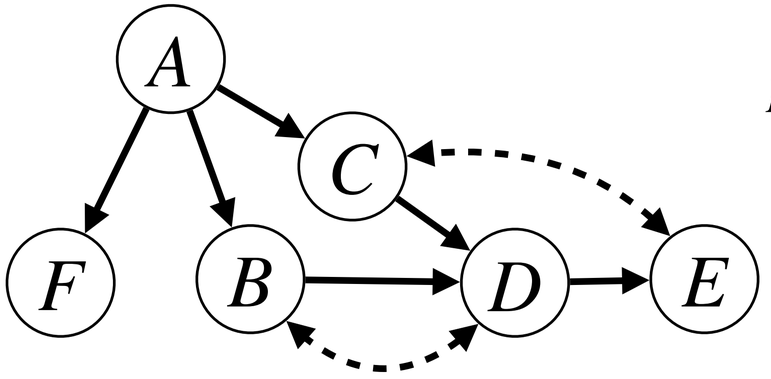
Bayesian Networks & Semi-Markov Condition

An ADMG G over $\mathbf{V}$ is a *Bayesian Network* for a joint probability distribution $P(\mathbf{V})$ if, for every $V_i \in \mathbf{V}$, it holds that $V_i \perp\!\!\!\perp NDesc_i \,|\, Pa_i^+$ and, therefore, $P(\mathbf{v})$ factorizes as follows:

P is satisfies the
Semi-*Markov Condition*
w.r.t. G

$$P(\mathbf{v}) = \prod_{V_i \in \mathbf{V}} P(v_i \,|\, pa_i^+).$$

The extended parents of V_i is defined as $Pa_i^+ = Pa^1(\{V \in \mathbf{C}(V_i) : V \leq V_i\}) \backslash \{V_i\}$, where $Pa^1(V) = Pa(V) \cup V$ and $\mathbf{C}(V_i)$ is a maximal path entirely made of bidirected edges.



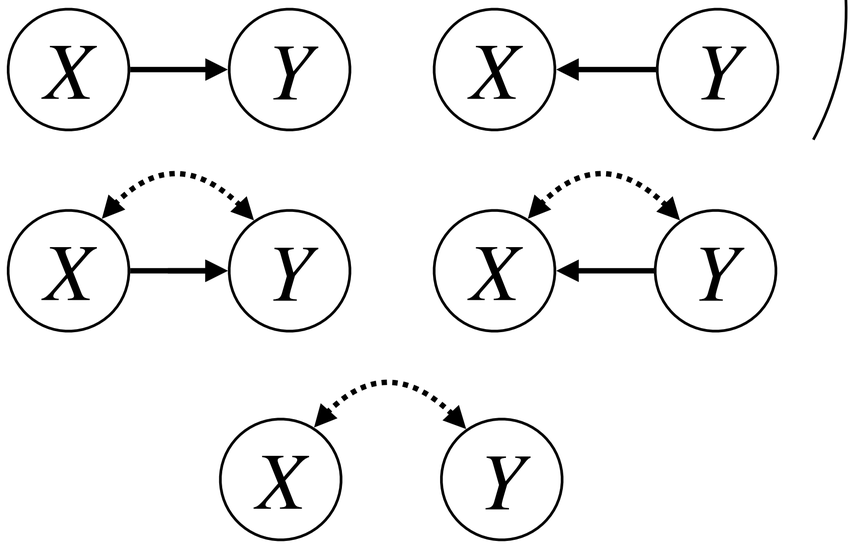
$$\begin{aligned} P(\mathbf{v}) &= P(e \,|\, d, c, b, a, f)\; P(d \,|\, c, b, a, f)\; P(c \,|\, b, a, f)\; P(b \,|\, a, f)\; P(f \,|\, a)\; P(a) \\ &= P(e \,|\, d, c, a)\; P(d \,|\, c, b, a)\; P(c \,|\, a)\; P(b \,|\, a)\; P(f \,|\, a)\; P(a) \end{aligned}$$

$E \perp\!\!\!\perp F, B \,|\, D, C, A$ $\quad D \perp\!\!\!\perp F \,|\, B, C, A$ $\quad C \perp\!\!\!\perp F, B \,|\, A$ $\quad B \perp\!\!\!\perp F \,|\, A$

Markov Equivalence Class

Definition (Markov Equivalence Class, MEC for short): A Markov Equivalence Class is a set of models that encode the same set of conditional independencies.

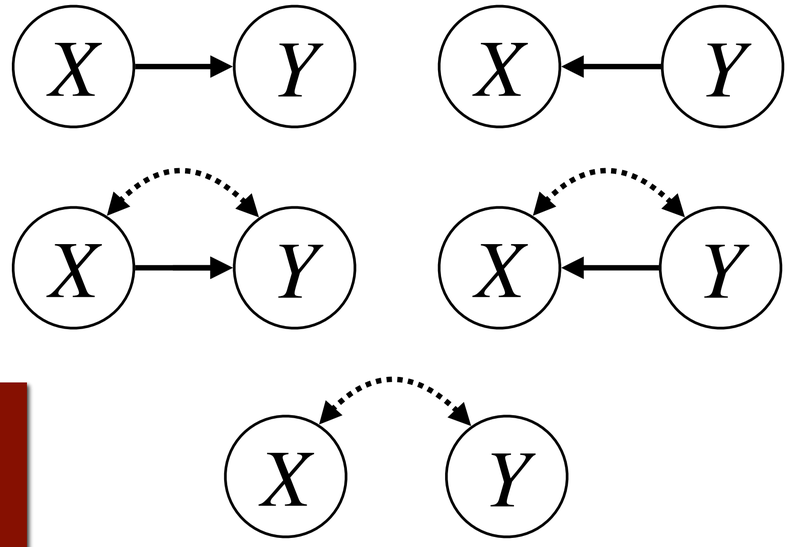
Markov Equivalent BNs

Distribution	Factorization	Bayesian Networks
$P(X, Y)$ with $P(Y \mid X) \neq P(Y)$ i.e., $X \not\perp\!\!\!\perp Y$	$P(x, y) = P(y \mid x)P(x)$ $P(x, y) = P(x \mid y)P(y)$	

Markov Equivalence Class

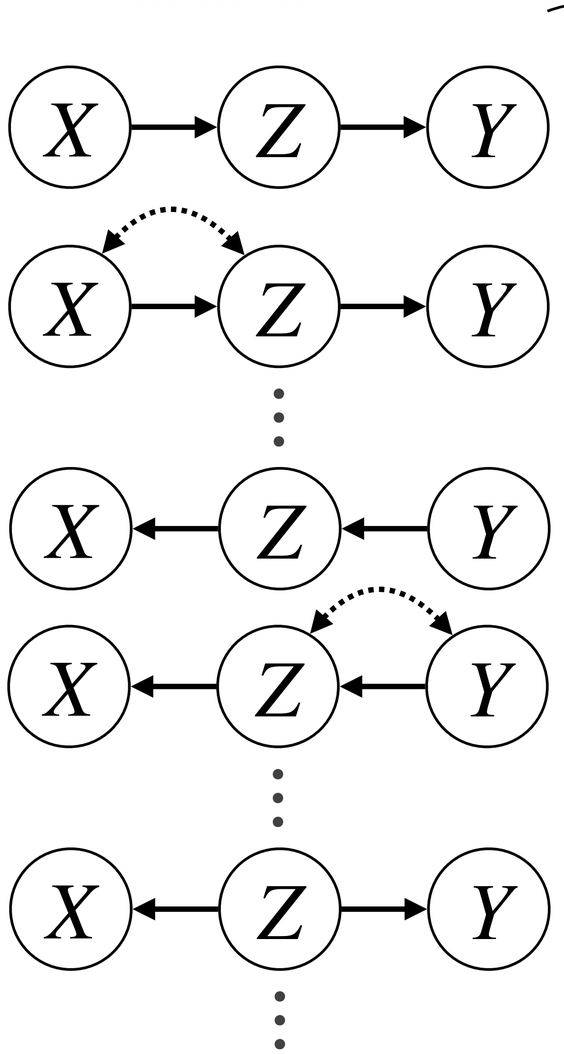
Definition (Markov Equivalence Class, MEC for short): A Markov Equivalence Class is a set of models that encode the same set of conditional independencies.

Markov Equivalent BNs

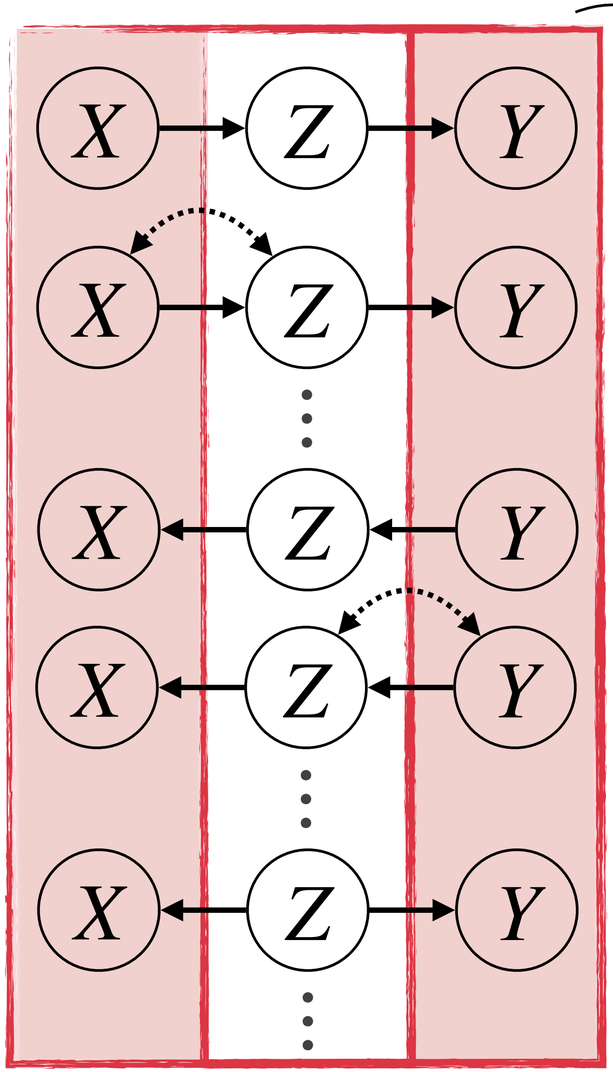
Distribution	Factorization	Bayesian Networks
$P(X, Y)$ with $P(Y \mid X) \neq P(Y)$ i.e., $X \not\perp\!\!\!\perp Y$	$P(x, y) = P(y \mid x)P(x)$ $P(x, y) = P(x \mid y)P(y)$	

All models imply no independence and no other invariance

Markov Equivalence Class

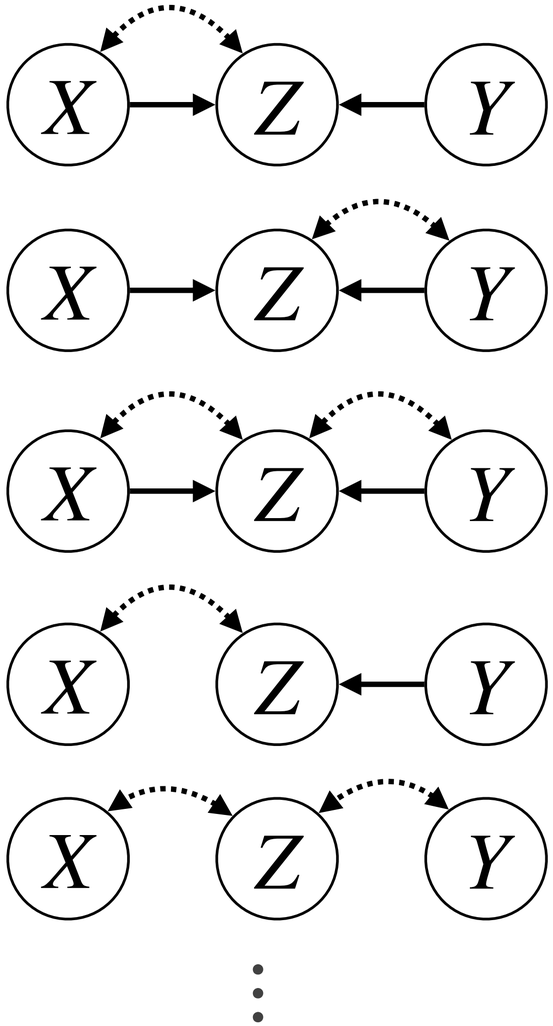
Distribution	Factorization	Bayesian Networks
$P(X, Y, Z)$ with $P(Y \mid X, Z) = P(Y \mid X)$ i.e., $X \perp\!\!\!\perp Y \mid Z$	$P(x, y, z) = P(y \mid x, z)P(z \mid x)P(x)$ $= P(y \mid z)P(z \mid x)P(x)$ $P(x, y, z) = P(x \mid y, z)P(y \mid z)P(z)$ $= P(x \mid z)P(z \mid y)P(y)$ $P(x, y, z) = P(y \mid x, z)P(x \mid z)P(z)$ $= P(y \mid z)P(x \mid z)P(z)$	

Markov Equivalence Class

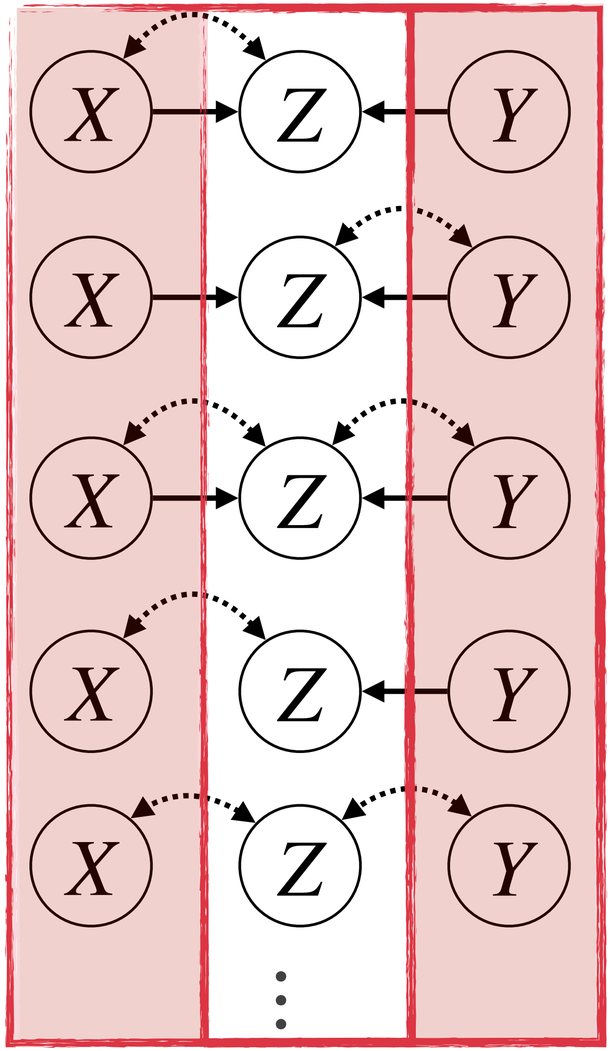
Distribution	Factorization	Bayesian Networks
$P(X, Y, Z)$ with $P(Y \mid X, Z) = P(Y \mid X)$ i.e., $X \perp\!\!\!\perp Y \mid Z$	$P(x,y,z) = P(y \mid x,z)P(z \mid x)P(x)$ $= P(y \mid z)P(z \mid x)P(x)$	
	$P(x,y,z) = P(x \mid y,z)P(y \mid z)P(z)$ $= P(x \mid z)P(z \mid y)P(y)$	
	$= P(y \mid z)P(x \mid z)P(z)$	

All models imply *only* $X \perp\!\!\!\perp Y \mid Z$ and
Z is always a *non-collider* in such models.

Markov Equivalence Class

Distribution	Factorization	Bayesian Networks
$P(X, Y, Z)$ with $P(Y \mid X) = P(Y)$ i.e., $X \perp\!\!\!\perp Y$	$P(x, y, z) = P(z \mid x, y)P(x \mid y)P(y)$ $= P(z \mid x, y)P(x)P(y)$	

Markov Equivalence Class

Distribution	Factorization	Bayesian Networks
$P(X, Y, Z)$ with $P(Y \mid X) = P(Y)$ i.e., $X \perp\!\!\!\perp Y$	$P(x, y, z) = P(z \mid x, y)P(x \mid y)P(y)$ $= P(z \mid x, y)P(x)P(y)$	

All models imply *only* $X \perp\!\!\!\perp Y$ and
Z is always a *collider* in such models,

Note: Z is never an ancestor of X or Y

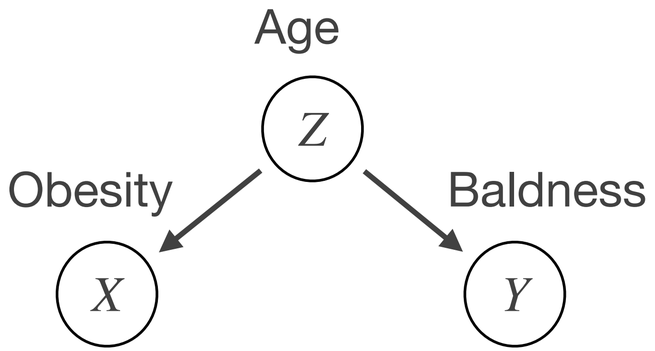
D-Separation

Graphical Tool for Identifying Conditional Independencies implied by Bayesian Networks

Implied Conditional independencies

Fork

Z as a common cause

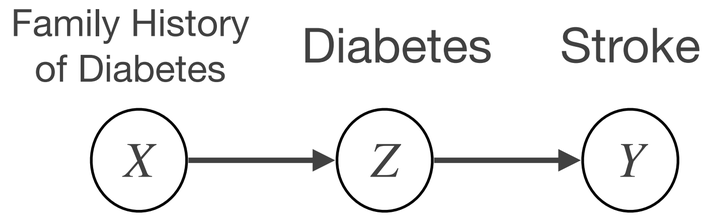


$$X \not\perp\!\!\!\perp Y$$

$$X \perp\!\!\!\perp Y \mid Z$$

Chain

Z as a mediator

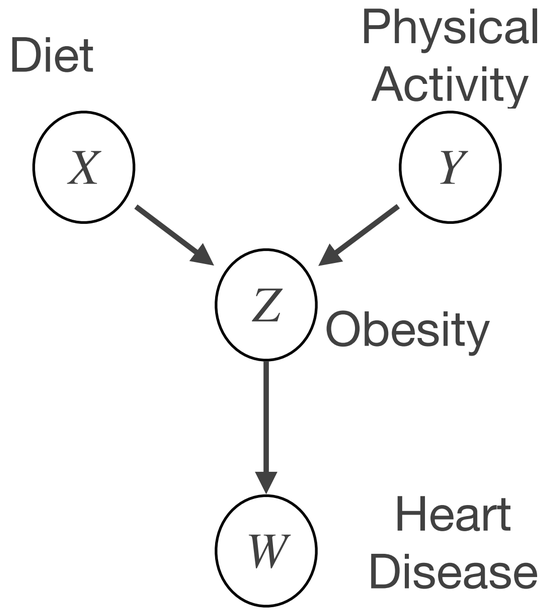


$$X \not\perp\!\!\!\perp Y$$

$$X \perp\!\!\!\perp Y \mid Z$$

V-Structure

Z as a collider or common effect



$$X \perp\!\!\!\perp Y$$

$$X \not\perp\!\!\!\perp Y \mid Z$$

$$X \not\perp\!\!\!\perp Y \mid W$$

Two Markov-equivalent models.

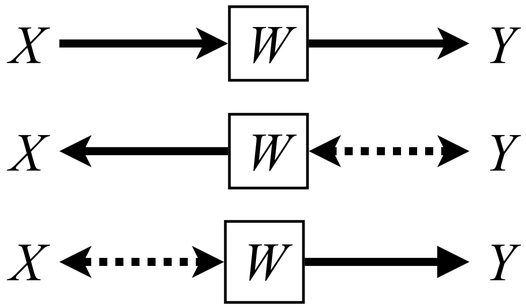
Note that in both cases Z is a non-collider!

Active and Inactive Triplets

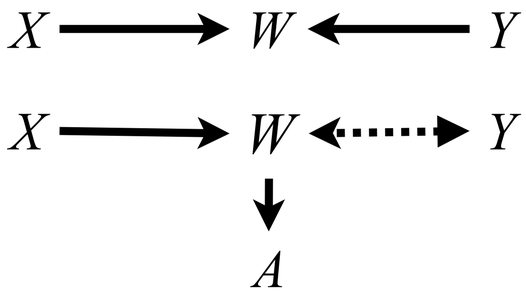
Definition (inactive): A triplet $\langle V_i, V_m, V_j \rangle$ is said to be ***inactive*** relative to a set $\mathbf{Z}$ if the middle node V_m:

1. Is a non-collider and is in $\mathbf{Z}$; or
2. Is a collider and neither it nor any of its descendants in $\mathbf{Z}$.

W is non-collider and $W \in \mathbf{Z}$



W is (descendant of) a collider and $W, A \notin \mathbf{Z}$

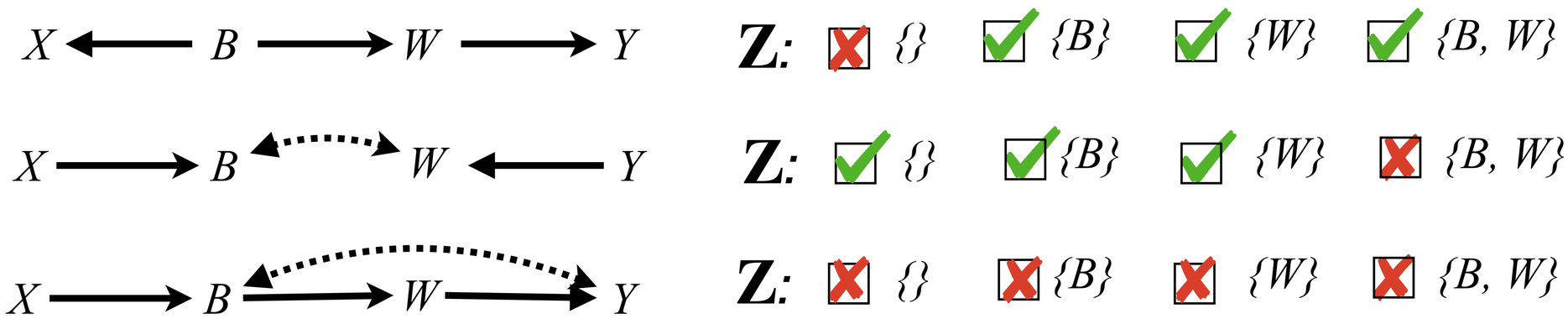


D-Separation

Definition (d-separation): A path p in an ADMG G is said to be ***d-separated*** (or blocked) by a set of variables $\mathbf{Z}$ if and only if p contains an inactive triplet in it.

A set $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ if and only if $\mathbf{Z}$ blocks every path between a node in $\mathbf{X}$ and a node in $\mathbf{Y}$. We denote that by $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_G$.

Does $\mathbf{Z}$ d-separate X and Y ?



$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_G \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P$$

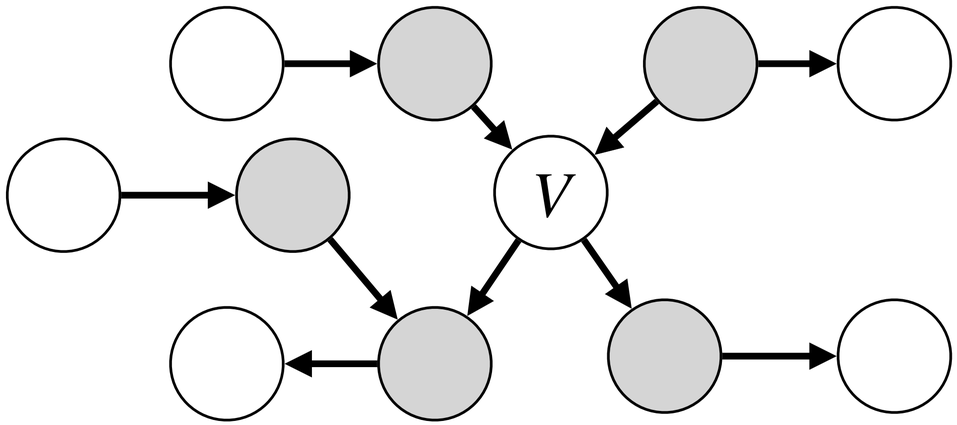
D-separations in G correspond to conditional independencies in P

The Markov Blanket: The Optimal Set for Prediction

For prediction, the Markov blanket is optimal:
adding variables outside the blanket cannot improve predictive performance

Markov Blanket (MB) of a *Markovian* BN over V: the union of parents, children, and parents of the children V.

$$\text{mb}_G(V) = Pa(V) \cup Ch(V) \cup Pa(Ch(V))$$



Markov Blanket of V

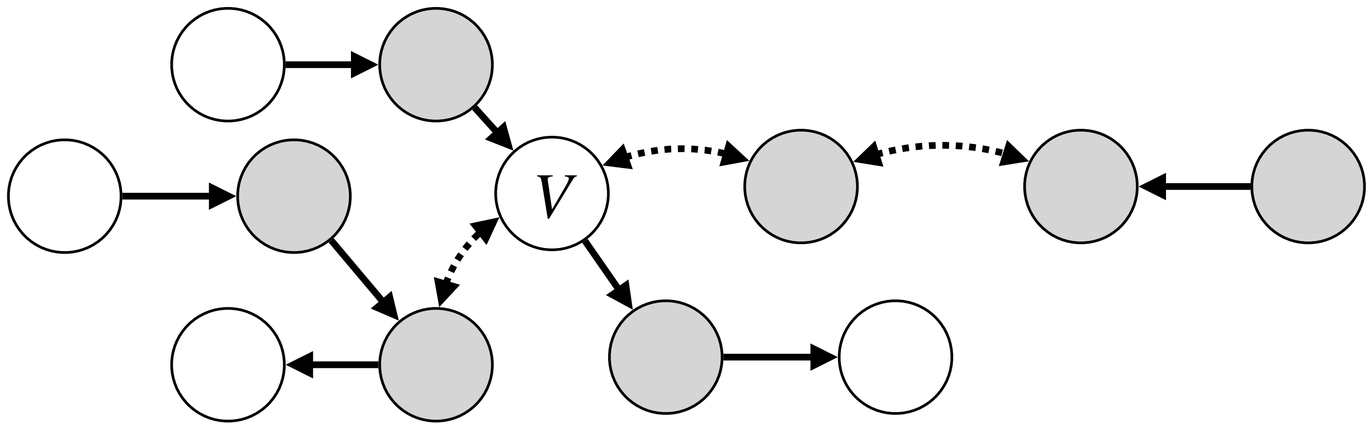
$$V \perp\!\!\!\perp \mathbf{V} \backslash \text{mb}_G(V) \mid \text{mb}_G(V)$$

The Markov Blanket: The Optimal Set for Prediction

Markov Blanket (MB) of a *Semi-Markovian* BN over V: is the district of V and the parents of the district of V (excluding V itself) i.e.:

$$\text{mb}_G(V) = \text{dis}_G(V) \cup \text{Pa}_G(\text{dis}_G(V)) \backslash \{V\}$$

District of V, $\text{dis}_G(V)$, is the set of variables connected with V through an edge or a bidirected path.



Markov Blanket of V

$$V \perp\!\!\!\perp \mathbf{V} \backslash \text{mb}_G(V) \mid \text{mb}_G(V)$$

Richardson, T. (2003). Markov Properties for Acyclic Directed Mixed Graphs. *Scandinavian Journal of Statistics*, *30*(1), 145–157

Can we learn the Markov Equivalence Class from observational data?

Learning the Markov Equivalence Class

Identifiability: In non-parametric settings (i.e., without making parametric or distributional assumptions) and solely from observational data, causal discovery algorithms can only learn a graphical representation of a *Markov equivalence class*!

Algorithms: Constraint-Based vs Score-Based

Systems: Causal Sufficient vs Causal Insufficient

Causal Sufficiency: assumption that all confounding variables have been observed — although strong, it has been widely employed to simplify causal discovery and inference.

Score-Based Causal Discovery Algorithms

Strategy: search for the most probable causal structure by assessing goodness-of-fit scores of different possible structures.

Common Scores: Bayesian Information Criterion (BIC) for Gaussian variables and the BDeu score for multinomial variables.

Under causal sufficiency:

- **GES:** Greedy Equivalence Search, by Chickering, 2003.
- **FGES:** Fast GES, by Ramsey et al., 2017 — extension of the GES that improves the runtime of the algorithm by using parallelization.

Score-Based Causal Discovery Algorithms

Accounting latent confounding:

Use BIC, assuming linear Gaussian models

- **GSMAG:** a greedy search algorithm for learning MAGs, by Triantafillou, S. and Tsamardinos, I., 2016.
- **MAGSL:** search based on **dynamic programming and branch and bound**, by Rantanen et al., 2021 — it is guaranteed to find a globally optimal MAG.
- **N-ADMG:** Neural ADMG Learning, by Ashman et al., 2023 — extends Diff-discovery to the setting where the true causal diagram is bow-free and corresponds to a non-linear SCM with additive noise.
- **AGFN:** Ancestral GFlowNets, by da Silva et al., 2026, — probabilistic causal discovery that integrates prior knowledge and actively elicits uncertain expert feedback.

Constraint-Based Causal Discovery Algorithms

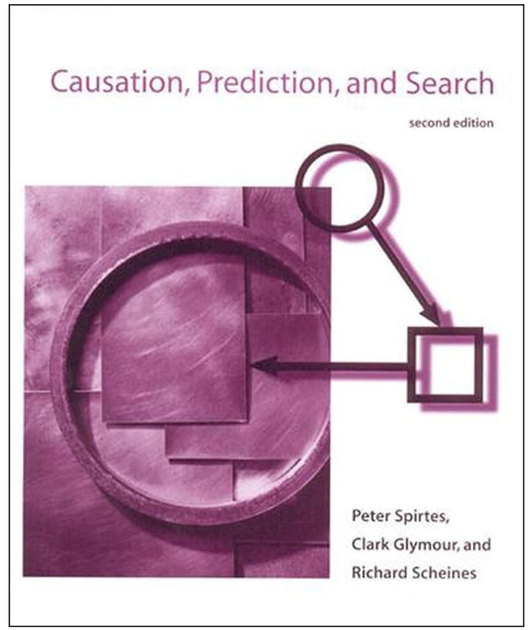
Strategy: construct a causal structure that aligns with all observed conditional independencies, identified using ***conditional independence tests***.

Under causal sufficiency:

IC: Inductive Causation, by Verma and Pearl, 1990.

PC: Peter-Clark, by Spirtes and Glymour, 1991.

They start with an adjacency (skeleton) phase, based on conditional independence tests, followed by an orientation phase.



Spirtes, P., Glymour, C., and Scheines, R. (2001).
Causation, Prediction, and Search, 2nd edn. Cambridge, MA: MIT Press.

Constraint-Based Causal Discovery Algorithms

Accounting for latent confounding:

- **FCI:** Fast Causal Inference, by Spirtes et al., 1995 . With the additional rules by Zhang, J. (2008), is a complete algorithm that handles both latent confounding and selection bias.
- **Multiple FCI variants:**
  - Anytime FCI (**AFCI**), by Spirtes P., 2001;
  - Conservative FCI (**CFCI)** and Really FCI (**RFCI**), by Colombo et al. 2012;
  - **FCI+**, by Claassen et al. 2013;
  - **SAT-Based FCI:** logical formalization of FCI, by Hyttinen et al., 2014.
  - **LPCMCI**: extension to time-series data, by Gerhardus, A., & Runge, J. (2020).
  - FCI for **cyclic graphs**, by Claassen, T. &; Mooij, J.M. (2023)
  - **dcFCI**, additional robustness in limited datasets, by Ribeiro & Heider, 2026.

Constraint-Based Causal Discovery Algorithms

$$\mathscr{M}_1 = \begin{cases} \mathbf{V} = \{X, Y, Z\} \\ \mathbf{U} = \{U_x, U_Y, U_Z\} \\ \mathscr{F} = \begin{cases} X \leftarrow f_X(U_X) \\ Z \leftarrow f_Z(X, Y, U_Z) \\ Y \leftarrow f_Y(U_Y) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

$$\mathscr{M}_{N-1} = \begin{cases} \mathbf{V} = \{X, Y, Z\} \\ \mathbf{U} = \{U_{XZ}, U_{YZ}, U_X, U_Y, U_Z\} \\ \mathscr{F} = \begin{cases} X \leftarrow f_X(U_{XZ}, U_X) \\ Z \leftarrow f_Z(Y, U_{XZ}, U_Z) \\ Y \leftarrow f_Y(U_Y) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

$$\mathscr{M}_N = \begin{cases} \mathbf{V} = \{X, Y, Z\} \\ \mathbf{U} = \{U_{XZ}, U_{YZ}, U_X, U_Y, U_Z\} \\ \mathscr{F} = \begin{cases} X \leftarrow f_X(U_{XZ}, U_X) \\ Z \leftarrow f_Z(U_{XZ}, U_{YZ}, U_Z) \\ Y \leftarrow f_Y(U_{YZ}, U_Y) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Data

Conditional (in)dependencies

$P(\mathbf{v})$

$X \perp\!\!\!\perp Y$

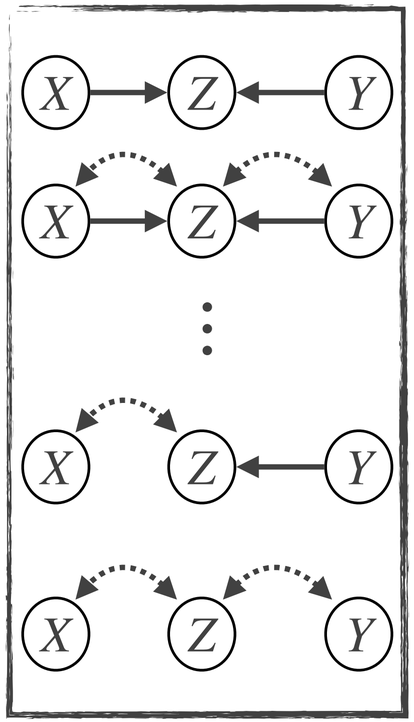
$X \not\perp\!\!\!\perp Z$

$Z \not\perp\!\!\!\perp Y$

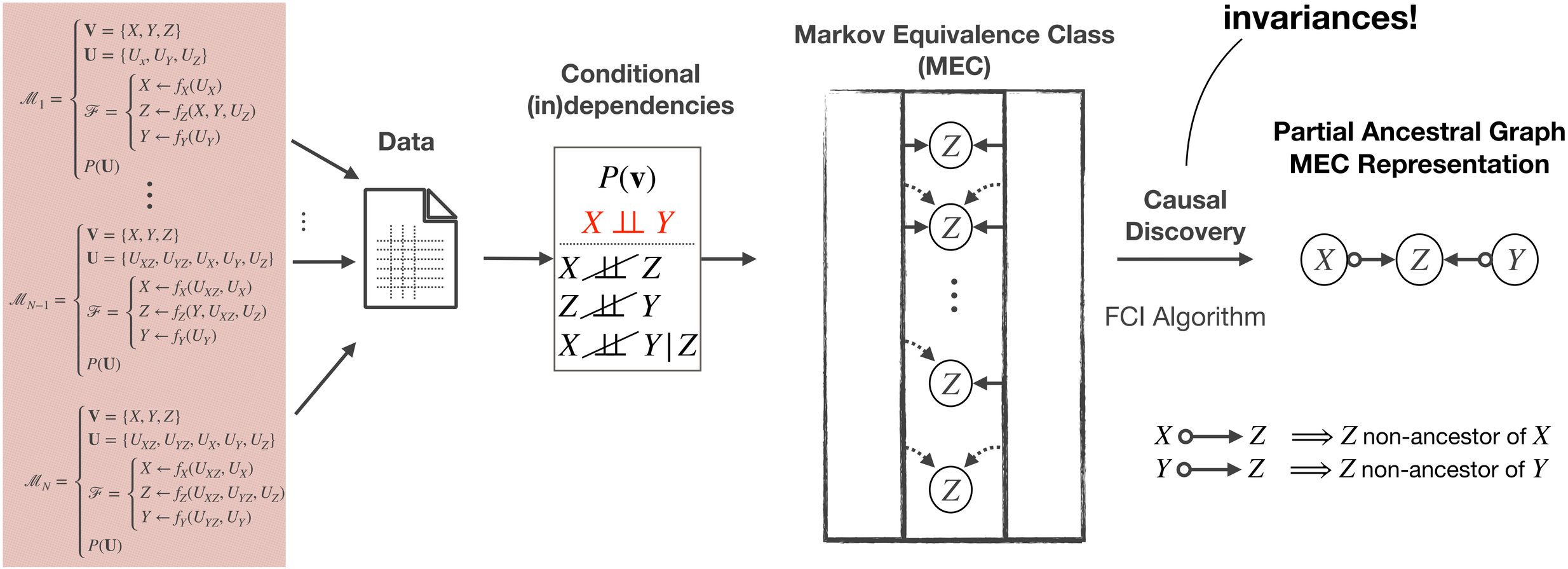
$X \not\perp\!\!\!\perp Y \mid Z$

$$P(x, y, z) = P(z \mid x, y)P(x \mid y)P(y)$$
$$= P(z \mid x, y)P(x)P(y)$$

Markov Equivalence Class (MEC)



Constraint-Based Causal Discovery Algorithms

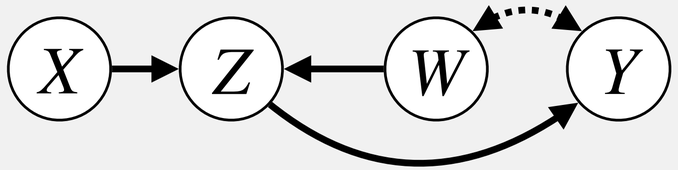


Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896. Link

FCI Algorithm - Pipeline

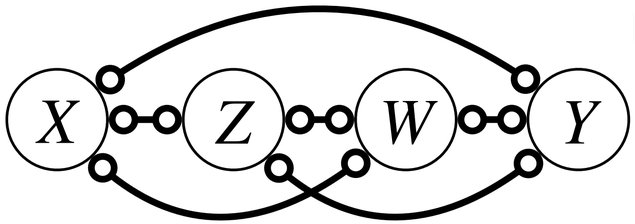
Unknown Reality

True causal diagram



$X \perp\!\!\!\perp W$
$X \perp\!\!\!\perp Y \mid Z, W$

Implied by the ADMG using d-separation

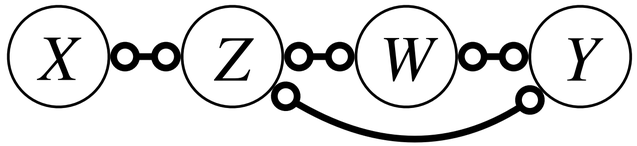


Complete Graph

Conditional Independence Tests

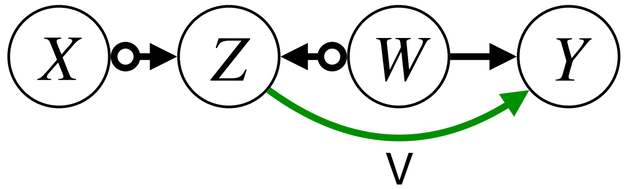
$X \perp\!\!\!\perp W$
$X \perp\!\!\!\perp Y \mid Z, W$

By **faithfulness**, are correctly observed in the data



Skeleton

FCI Rules
$(R0) + (R1 \text{ to } R10)$



V

Partial Ancestral Graph (PAG)

Implied by the PAG using m-separation

$X \perp\!\!\!\perp W$
$X \perp\!\!\!\perp Y \mid Z, W$

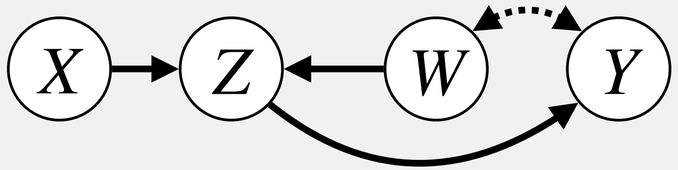
$A \circ\!\!\rightarrow B \implies$ B non-ancestor of A
$A \rightarrow B \implies$ A ancestror of B
$A \leftrightarrow B \implies$ spurious association
$A - B \implies$ selection bias

Z is not an ancestor of X or W.
Z and W are ancestors of Y.
Z is not confounded with Y.

Applying FCI's Rules

Unknown Reality

True causal diagram



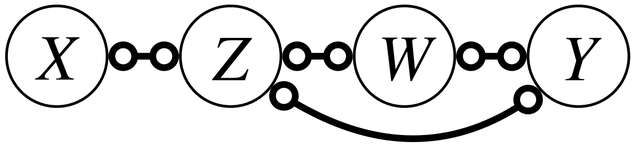
$X \perp\!\!\!\perp W$

$X \perp\!\!\!\perp Y \mid Z, W$

Implied by the ADMG using d-separation

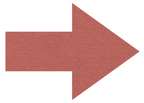
R0: If $\langle V_1, V_2, V_3 \rangle$ is unshielded and $V_2 \notin Sepset(V_1, V_3)$, then

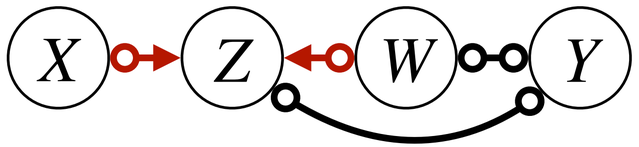
That is the only way for the path between V_1 and V_3 to be blocked when not conditioning on V2



Skeleton

Applying R0:





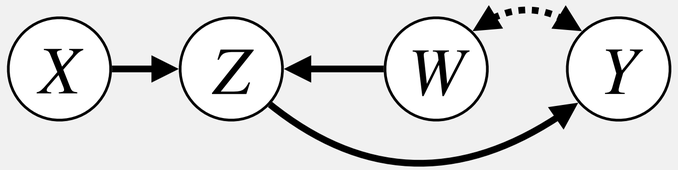
$X \perp\!\!\!\perp W$, where $Z \notin Sepset(X, W)$

$X \perp\!\!\!\perp Y \mid Z, W$, where $Z \in Sepset(X, Y)$

Applying FCI's Rules

Unknown Reality

True causal diagram

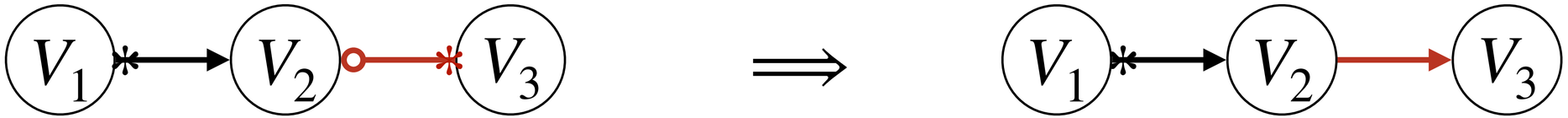


$X \perp\!\!\!\perp W$

$X \perp\!\!\!\perp Y \mid Z, W$

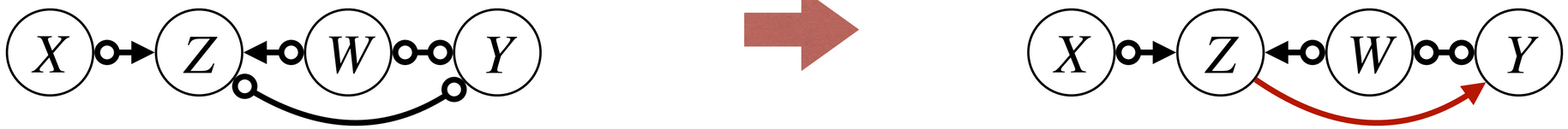
Implied by the ADMG using d-separation

R1:



where V_1 and V_3 are not adjacent

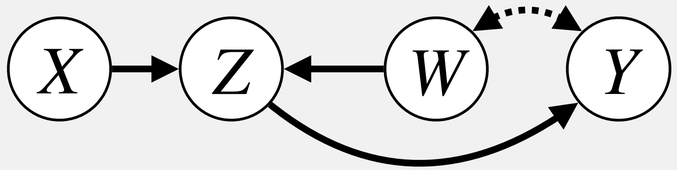
FCI's R1:



Applying FCI's Rules

Unknown Reality

True causal diagram

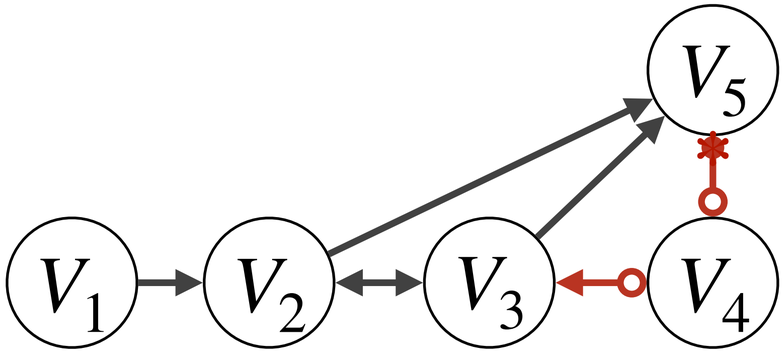


$X \perp\!\!\!\perp W$

$X \perp\!\!\!\perp Y \mid Z, W$

Implied by the ADMG using d-separation

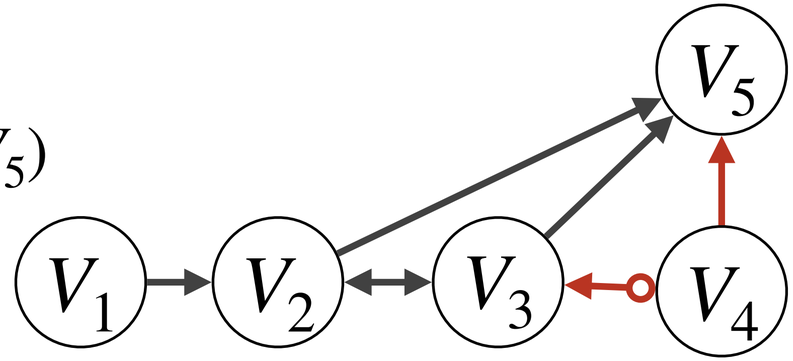
R4:



$\langle V_1, V_2, V_3, V_4, V_5 \rangle$ is a discriminating path for V_4

$V_4 \in Sepset(V_1, V_5)$

$\Longrightarrow$

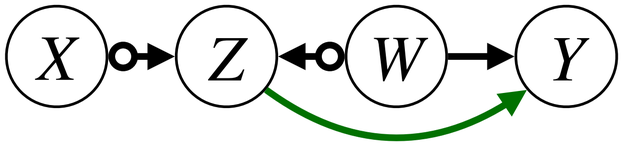


FCI's R4:

$X \perp\!\!\!\perp Y \mid Z, W$

$W \in Sepset(X, Y)$

PAG: Representation of the Markov Equivalence Class



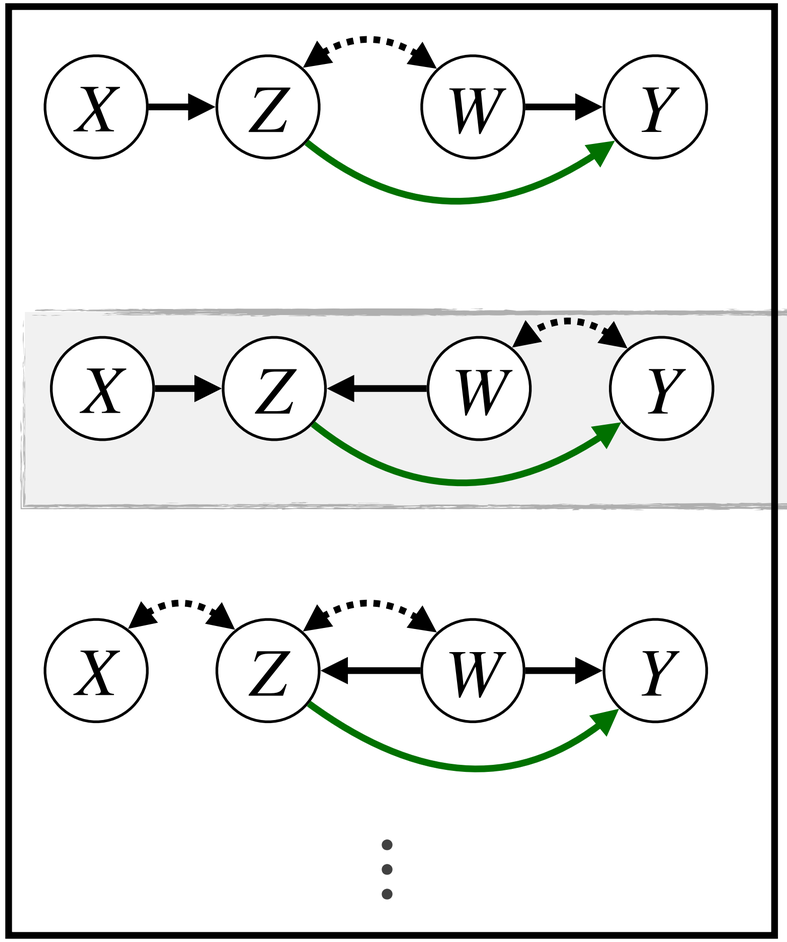
V

Partial Ancestral Graph (PAG)

Z is not an ancestor of X or W.

Z and W are ancestors of Y.

Z is not confounded with Y.



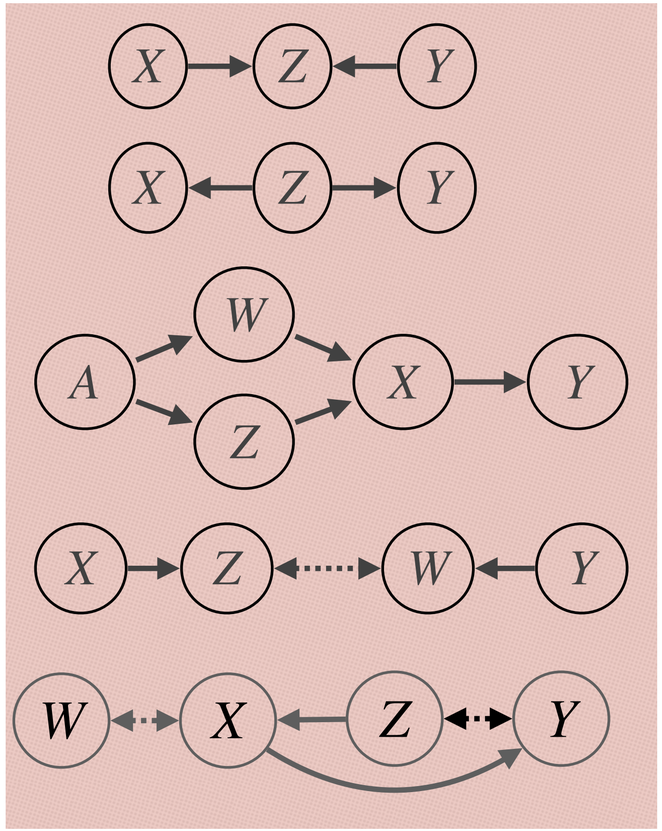
True (unknown) causal diagram

$X \perp\!\!\!\perp W$

$X \perp\!\!\!\perp Y | Z, W$

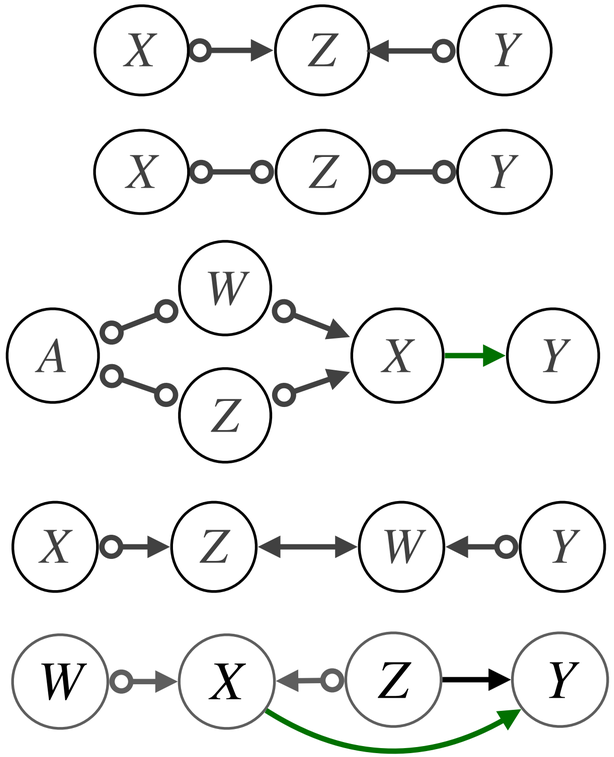
Fast Causal Inference (FCI) Algorithm

Underlying Causal Diagram



Data → FCI → E.C.

Partial Ancestral Graph



Available Implementations of the FCI

R Packages:

- pcalg R package:
  - https://cran.r-project.org/web/packages/pcalg/
  - https://github.com/cran/pcalg/
- RPy-Tetrad (Wrapper in R): https://github.com/cmu-phil/py-tetrad/tree/main/pytetrad/R

Python Packages:

- Do-discover in PyWhy: https://github.com/py-why/dodiscover
- Causal-Learn: https://causal-learn.readthedocs.io/en/latest/index.html
- Py-Tetrad (Wrapper in Python): https://github.com/bd2kccd/py-causal

Conditional Independence Tests

Assuming **independent observations:**

Linear models with **Gaussian errors**: partial correlation test

- **Fisher, R.A.** (1921). *On the Probable Error of a Coefficient of Correlation Deduced from a Small Sample.*
- R package: https://cran.r-project.org/web/packages/pcalg/

Linear models with **heterogeneous data types**- likelihood ratio tests based on GLM

- **Tsagris, M., Borboudakis, G., Lagani, V. *et al.*** (2018) Constraint-based causal discovery with mixed data. *Int J Data Sci Anal* **6**, 19–30. (Link)
- R package: https://cran.r-project.org/web/packages/MXM/

Linear / **non-linear** models with **heterogeneous data types —** kernel-based non-parametric test

- **Zhang, K., Peters, J., Janzing, D., & Schölkopf, B.** (2012). *Kernel-based conditional independence test and application in causal discovery.* In: Uncertainty in artificial intelligence. AUAI Press; 2011. p.804–13
- R package: https://cran.r-project.org/web/packages/CondIndTests

Fisher's Partial Correlation Test

Linear models with **Gaussian errors**: partial correlation test

- **Fisher, R.A.** (1921). *On the Probable Error of a Coefficient of Correlation Deduced from a Small Sample.*
- R package: https://cran.r-project.org/web/packages/pcalg/

Testing $X \perp\!\!\!\perp Y \,|\, \mathbf{S}$**:**

For multivariate Gaussian variables, the **partial correlation test** is equivalent to a conditional independence test.

1) Obtain the **partial correlation coefficient** via univariate **linear regression**:

Give i.i.d. data $D = \{x_i, y_i, s_{ki} : k = 1, \dots |\mathbf{S}|\}$

$$x_i = \alpha^X + \sum_{k=1}^{|\mathbf{S}|} \beta_k^X s_{ki} + \varepsilon_i^X, \text{ where } \varepsilon_i^X \sim \mathcal{N}(0, \sigma^2_{X|\mathbf{S}})$$

$$y_i = \alpha^Y + \sum_{k=1}^{|\mathbf{S}|} \beta_k^Y s_{ki} + \varepsilon_i^Y, \text{ where } \varepsilon_i^Y \sim \mathcal{N}(0, \sigma^2_{Y|\mathbf{S}})$$

Then, $\rho_{XY|\mathbf{S}} = \text{corr}\left(\boldsymbol{\varepsilon}^X, \boldsymbol{\varepsilon}^Y\right)$

The partial correlation coefficient is the Pearson correlation between the residuals!

Fisher's Partial Correlation Test

2) Test the **significance** of the partial correlation coefficient:

Hypothesis test: $H_0 : \rho_{XY|\mathbf{S}} = 0$ vs $H_1 : \rho_{XY|\mathbf{S}} \neq 0$

Fisher's z-transformed partial correlation coefficient:

$$z(\rho_{XY|\mathbf{S}}) = \frac{1}{2} \ln \left(\frac{1 + \rho_{XY|\mathbf{S}}}{1 - \rho_{XY|\mathbf{S}}} \right).$$

Under H_0, the test statistic $T = \sqrt{N - |\mathbf{S}| - 3} \; |z(\rho_{XY|\mathbf{S}})|$ asymptotically follows a standard Gaussian distribution (zero mean and unit variance).

We reject H_0 if p-value < α, where p-value for the two-tailed test is given by

$$P(T > |t| \,|\, H_0) = 2 \times (1 - cdf(|t|),$$

t is the observed (estimated) value of T, and

cdf is the cumulative distribution function of T under H_0.

How can we accommodate heterogeneous data types?

Linear models with **heterogeneous data types**- likelihood ratio tests based on **GLM**

- **Tsagris, M., Borboudakis, G., Lagani, V. *et al.*** (2018) Constraint-based causal discovery with mixed data. *Int J Data Sci Anal* **6**, 19–30. (Link)
- R package: https://cran.r-project.org/web/packages/MXM/

Likelihood Ratio Test (LRT) for testing $X \perp\!\!\!\perp Y \,|\, \mathbf{S}$:

1) First, we test $H_{0_1} : P(Y \,|\, X, \mathbf{S}) = P(Y \,|\, \mathbf{S})$

1.a) Fit the null and alternative models for Y using **generalized linear models (GLMs)**:

Give i.i.d. data $D = \{x_i, y_i, s_{ki} : k = 1, \ldots |\mathbf{S}| \}$

Null model $\mathscr{M}_0$: $g(E[Y_i \,|\, \mathbf{S}_i]) = \eta_i^{(0)} = \alpha + \sum_{k=1}^{|\mathbf{S}|} \beta_{S_k} s_{ki}$

Alternative model $\mathscr{M}_1$: $g(E[Y_i \,|\, X_i, \mathbf{S}_i]) = \eta_i^{(1)} \alpha + \beta_X X_i + \sum_{k=1}^{|\mathbf{S}|} \beta_{S_k} s_{ki}$

where $g(\cdot)$ is the link function appropriated to the data type of Y (e.g., identity for Gaussian, logit for binary, log for counts).

How can we accommodate heterogeneous data types?

1.b) Compare the models using a likelihood-ratio test (LRT):

Let:

- $\ell_0 = \log \mathcal{L}(\theta_{\mathcal{M}_0}) = \sum_{i=1}^{n} \log f(y_i \mid S_i^\top \beta_S)$, be the maximized log-likelihood under $\mathcal{M}_0$
- $\ell_1 = \log \mathcal{L}(\theta_{\mathcal{M}_1}) = \sum_{i=1}^{n} \log f(y_i \mid S_i^\top \beta_S + X_i^\top \beta_X)$ be the maximized log-likelihood under $\mathcal{M}_1$

Then the LRT statistic $\Lambda = -2\left(\ell_0 - \ell_1\right)$ measures how much better the alternative model $\mathcal{M}_1$ fits the data compared to the null model $\mathcal{M}_0$.

Under $H_{0_1} : \beta_x = 0$ and regularity conditions:

$$\Lambda \xrightarrow{d} \chi^2_k, \quad \text{with } k = \dim(\beta_X).$$

This enables computation of p-value p_1 corresponding to testing $H_{0_1} : P(Y \mid X, \mathbf{Z}) = P(Y \mid \mathbf{Z})$

How can we accommodate heterogeneous data types?

2) Repeat steps 1.a) and 1.b), switching the roles of Y and X, to test $H_{0_2} : P(X \mid Y, \mathbf{S}) = P(X \mid \mathbf{S})$

This enables computation of p-value p_2.

3) Tests are not necessarily symmetric, so a **final p-value** can be computed as following:

$p = \min(p1, p2)$ Favors dependencies

$p = \max(p1, p2)$ Favors independencies

$p = \min\{2 \times \min(p1, p2), \max(p1, p2)\}$ Heuristic shown to have a good control of FP and FN

Priority: Continuous > Nominal > Ordinal

Tsagris, M., Borboudakis, G., Lagani, V. *et al.* (2018) Constraint-based causal discovery with mixed data. *Int J Data Sci Anal* **6**, 19–30. (Link)

LR-Based CI Test for Heterogenous Data Types

Implementations in R:

MXM R Package — functions `ci.mm2` and `ci.fast2`

- Handles **binary, nominal** and **ordinal** variables

Lagani V, Athineou G, Farcomeni A, Tsagris M, Tsamardinos I (2017). “Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets.” *Journal of Statistical Software*, 80(7). doi:10.18637/jss.v080.i07

CRAN: Check latest archived version: https://cran.r-project.org/src/contrib/Archive/MXM/

GitHub: https://github.com/mensxmachina/MXM-R-Package

FCI.Utils R Package — function `mixedCITest`

- Handles **binary, nominal, ordinal, proportions, and (zero inflated) count** variables

GitHub: https://github.com/adele/FCI.Utils

Equivalence of Residual, Likelihood, and Coefficient-Based Tests

Under **joint Gaussianity**, partial correlation, regression-based tests (e.g., t-tests), and likelihood ratio tests all provide **equivalent** ways to test for conditional independence:

Zero partial correlation between X and Y given $\mathbf{S}$

(zero correlation between the residuals of X and Y after regressing out $\mathbf{S}$)

$\Leftrightarrow$

No improvement in model fit when adding X to a regression of Y on $\mathbf{S}$

(as tested by an LRT or F-test)

$\Leftrightarrow$

No statistical significance when testing whether $\beta_X = 0$ (e.g., via a t-test)

in the linear model $Y = \beta_X X + \mathbf{S}^T \beta_{\mathbf{S}} + \varepsilon$

Conditional Independence Decisions in Finite Samples

Given a significance level α (e.g., 0.05),

$$p \leq \alpha \implies X \not\perp\!\!\!\perp Y \,|\, \mathbf{S}$$
$$p > \alpha \implies X \perp\!\!\!\perp Y \,|\, \mathbf{S}$$

Even when conditional independence tests are **correctly specified,** errors may arise due to:

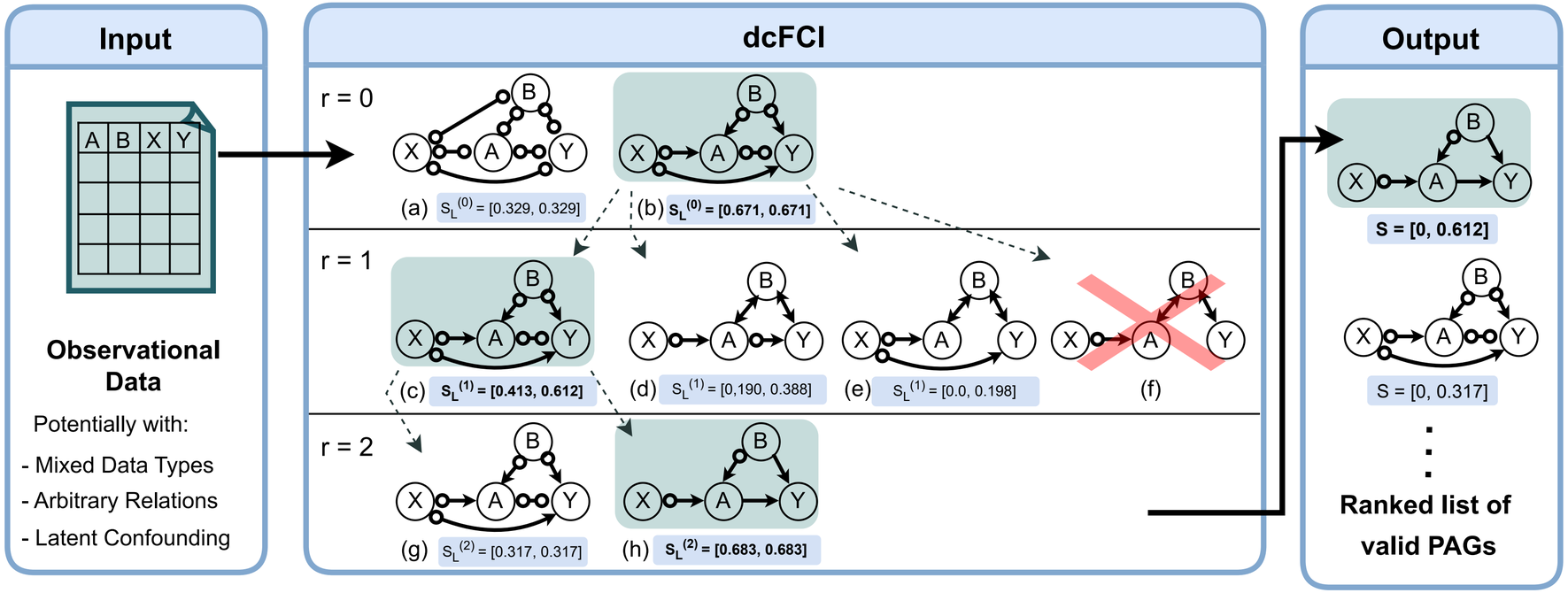
- Limited sample size
- Weak effect sizes
- Multicollinearity or measurement noise

Consequences include:

- Failure to reject H_0 — **false independencies**
- Incorrect rejection of H_0 - **false dependencies**

Uncertainty Modeling with data-compatible FCI (dcFCI)

- Introduced the ***first non-parametric Bayesian score*** to assess PAG fitness to data and quantify uncertainty in the learning process.
- Developed a ***hybrid causal discovery algorithm*****, dcFCI**, which generates a ranked list of data-compatible PAGs based on their scores.

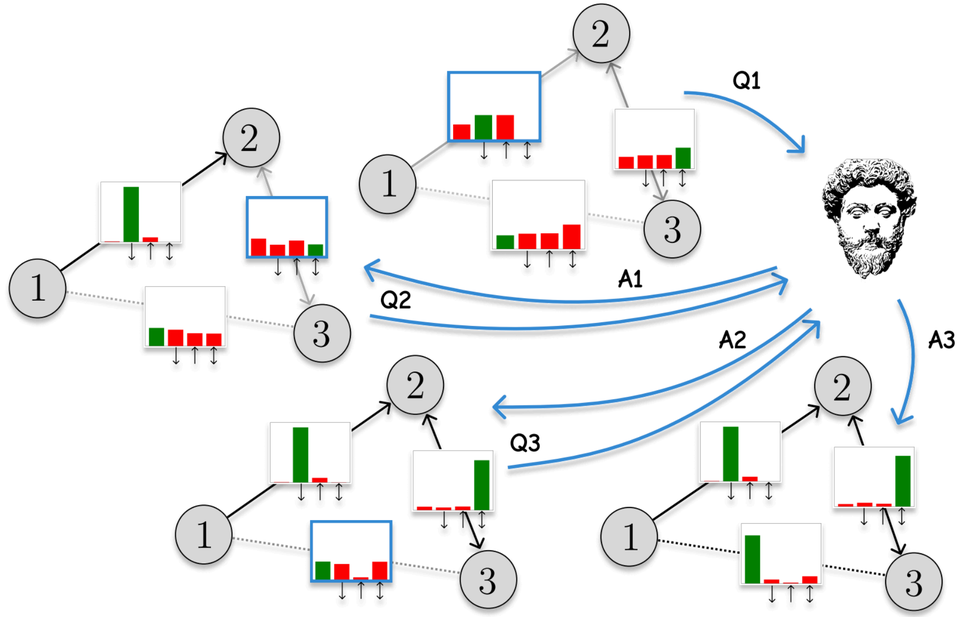


Ribeiro, A. H. and Heider, D. (2026). dcFCI: Robust causal discovery under latent confounding, unfaithfulness, and mixed data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (TPAMI), 1-12 (Link).

Probabilistic Causal Discovery: Active Learning and EITL

Expert-in-the-Loop Probabilistic Causal Discovery

- We introduce **Ancestral Generative Flow Networks (AFGN)**.
- First learns the distribution of ancestral graphs from data.
- Integrates an ***optimal*** strategy for ***eliciting feedback*** on causal relations, allowing knowledge to be incorporated from both ***experimental evidence*** and ***domain expertise***.
- Experts may provide ***uncertain*** or ***probabilistic knowledge***.

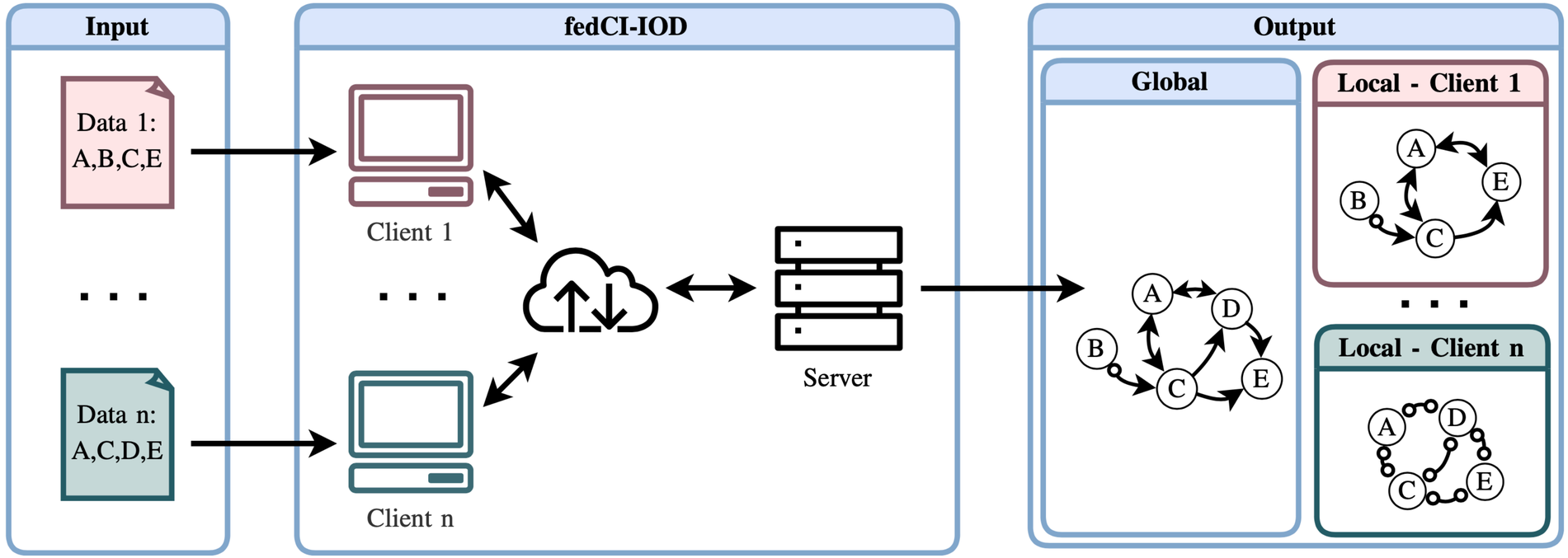


Learned AGFN is progressively refined with feedbacks from an expert.

da Silva, T., Bazaluk, B., Silva, E., Góis, A., Lahlou, S., Heider, D., Kaski, S. and Mesquita, D., **Ribeiro, A. H.** (2026). Expert-Aided Causal Discovery of Ancestral Graphs. *Information Sciences*, 123816. (Link)

Federated Causal Discovery from Multiple Datasets

- Federated conditional independence testing: preserving ***data privacy*** by sharing only model parameters.
- First federated causal discovery framework that accounts for ***latent confounding*** and ***selection bias*** across ***heterogeneous*** datasets with ***non-identical variable sets***, ***mixed data types***.



Available as a prebuilt docker image on docker hub: maximilianhahn/fedci-iod-app

Hahn M., Zajak, A., Heider, D., **Ribeiro A. H.** (2026). Federated Causal Discovery Across Heterogeneous Datasets under Latent Confounding —arXiv:2603.05149 — *Under Review.*

Causal Effect Identification Given a Causal Diagram

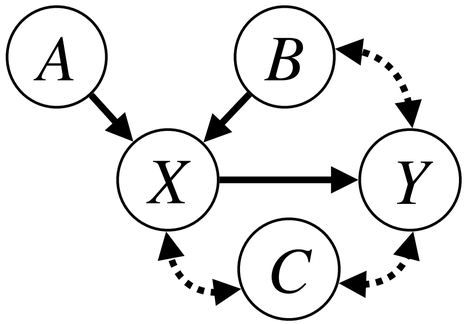
Graphical Criteria, Do-Calculus, and ID-Algorithm

Classical Causal Pipeline

1 **Query**

$P(y \mid do(x))$

2 **Causal Diagram**



3 $P(x, y, a, b, c)$

Observational Distribution

Inference Engine

Solution

yes / no

$$P(y \mid do(x)) = \sum_b P(y \mid x, b)P(b)$$

$$\mathbb{E}(y \mid do(x)) = \sum_b \mathbb{E}(y \mid x, b)P(b)$$

Observational Data

$$\hat{\mathbb{E}}(y \mid do(x)) = \sum_b \hat{\mathbb{E}}((y \mid x, b)\hat{P}(b)$$

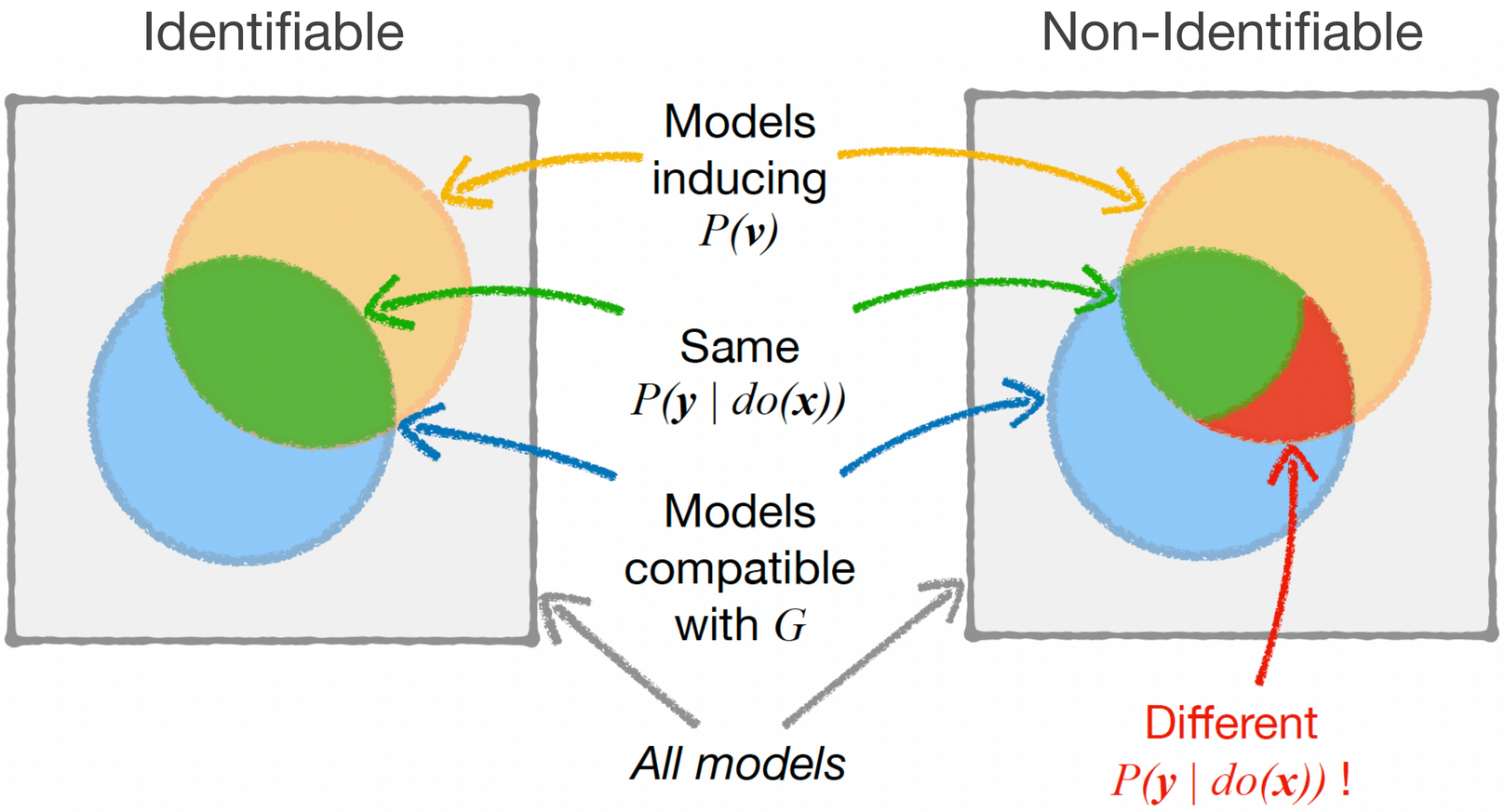
Interventional Distribution ← Available Distributions

Causal Modeling → Causal Effect Identification → Causal Effect Estimation

Pearl, J. 1995. Causal diagrams for empirical research. Biometrika, 82(4): 669–688.

Tian, J. and Pearl, J. (2002) A General Identification Condition for Causal Effects. In AAAI 2002, pp. 567–573, Menlo Park, CA.

The Effect Identification Problem



In words, causal effect identifiability means that, no matter the form of true SCM, for all models $\mathscr{M}$ agreeing with $\langle G, P(\mathbf{V}) \rangle$, they also agree in $P(\mathbf{y} \mid do(\mathbf{x}))$.

Identification via Backdoor Criterion

In $G_{\underline{\mathbf{X}}}$, all non-backdoor paths are severed

Let $\mathbf{X}$ be a set of treatment variables and $\mathbf{Y}$ a set of outcome variables in the causal graph G.

If there exists a set $\mathbf{Z}$ such that:

1. $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ in the graph $G_{\underline{\mathbf{X}}}$, i.e., the graph resulting from cutting the arrows out of $\mathbf{X}$
2. no node in $\mathbf{Z}$ is a descendant of a variable $X \in \mathbf{X}$ in G (all variables in $\mathbf{Z}$ are pre-treatment)

Then, $\mathbf{Z}$ satisfies the ***backdoor criterion*** for $(\mathbf{X}, \mathbf{Y})$ and, then the effect of $\mathbf{X}$ on $\mathbf{Y}$ is given by:

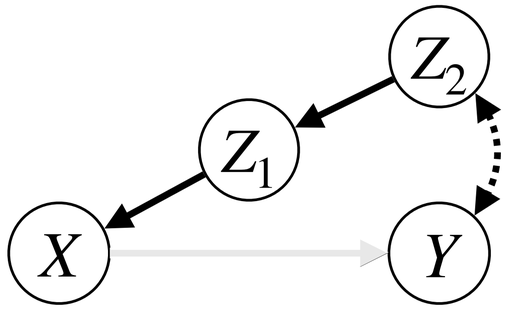
$$P(\mathbf{y} \,|\, do(\mathbf{x})) = \sum_{\mathbf{z}} P\left(\mathbf{y} \,|\, \mathbf{x}, \mathbf{z}\right) P\left(\mathbf{z}\right)$$

$\mathbf{Z}$, a set of covariates, admissible for backdoor adjustment

$\mathbf{Z} = \{Z_1\}$

$\mathbf{Z} = \{Z_1, Z_2\}$

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



Judea Pearl. Comment: Graphical models, causality and intervention. Stat. Sci., 8:266–269, 1993.

Identification via Backdoor Criterion

In $G_{\underline{\mathbf{X}}}$, all non-backdoor paths are severed

Let $\mathbf{X}$ be a set of treatment variables and $\mathbf{Y}$ a set of outcome variables in the causal graph G.

If there exists a set $\mathbf{Z}$ such that:

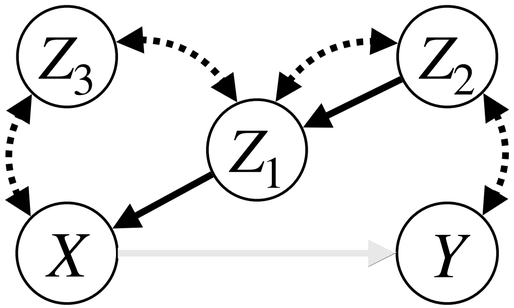
1. $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ in the graph $G_{\underline{\mathbf{X}}}$, i.e., the graph resulting from cutting the arrows out of $\mathbf{X}$
2. no node in $\mathbf{Z}$ is a descendant of a variable $X \in \mathbf{X}$ in G (all variables in $\mathbf{Z}$ are pre-treatment)

Then, $\mathbf{Z}$ satisfies the ***backdoor criterion*** for $(\mathbf{X}, \mathbf{Y})$ and, then the effect of $\mathbf{X}$ on $\mathbf{Y}$ is given by:

$$P(\mathbf{y} \,|\, do(\mathbf{x})) = \sum_{\mathbf{z}} P\left(\mathbf{y} \,|\, \mathbf{x}, \mathbf{z}\right) P\left(\mathbf{z}\right)$$

$\mathbf{Z}$, a set of covariates, admissible for backdoor adjustment

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



$\mathbf{Z} = \{Z_1\}$

$\mathbf{Z} = \{Z_1, Z_2\}$

Judea Pearl. Comment: Graphical models, causality and intervention. Stat. Sci., 8:266–269, 1993.

Identification via Backdoor Criterion

In $G_{\underline{\mathbf{X}}}$, all non-backdoor paths are severed

Let $\mathbf{X}$ be a set of treatment variables and $\mathbf{Y}$ a set of outcome variables in the causal graph G.

If there exists a set $\mathbf{Z}$ such that:

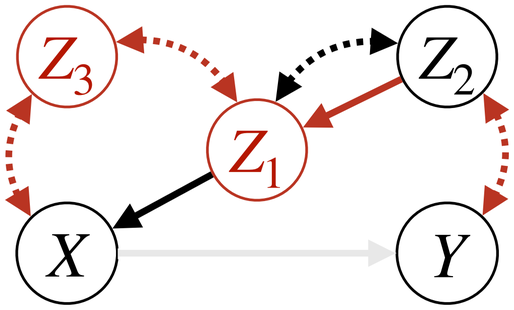
1. $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ in the graph $G_{\underline{\mathbf{X}}}$, i.e., the graph resulting from cutting the arrows out of $\mathbf{X}$
2. no node in $\mathbf{Z}$ is a descendant of a variable $X \in \mathbf{X}$ in G (all variables in $\mathbf{Z}$ are pre-treatment)

Then, $\mathbf{Z}$ satisfies the ***backdoor criterion*** for $(\mathbf{X}, \mathbf{Y})$ and, then the effect of $\mathbf{X}$ on $\mathbf{Y}$ is given by:

$$P(\mathbf{y} \,|\, do(\mathbf{x})) = \sum_{\mathbf{z}} P\left(\mathbf{y} \,|\, \mathbf{x}, \mathbf{z}\right) P\left(\mathbf{z}\right)$$

$\mathbf{Z}$, a set of covariates, admissible for backdoor adjustment

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



$\mathbf{Z} = \{Z_1\}$

$\mathbf{Z} = \{Z_1, Z_3\}$

$\mathbf{Z} = \{Z_1, Z_3\}$ ✘

Judea Pearl. Comment: Graphical models, causality and intervention. Stat. Sci., 8:266–269, 1993.

Identification via Backdoor Criterion

In $G_{\underline{\mathbf{X}}}$, all non-backdoor paths are severed

Let $\mathbf{X}$ be a set of treatment variables and $\mathbf{Y}$ a set of outcome variables in the causal graph G.

If there exists a set $\mathbf{Z}$ such that:

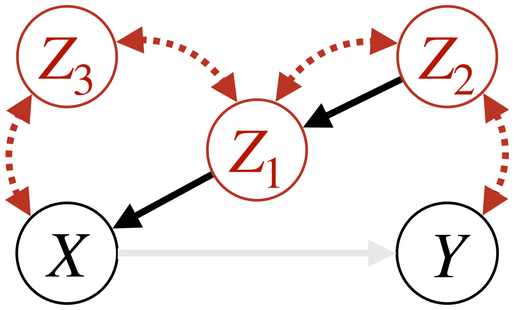
1. $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ in the graph $G_{\underline{\mathbf{X}}}$, i.e., the graph resulting from cutting the arrows out of $\mathbf{X}$
2. no node in $\mathbf{Z}$ is a descendant of a variable $X \in \mathbf{X}$ in G (all variables in $\mathbf{Z}$ are pre-treatment)

Then, $\mathbf{Z}$ satisfies the ***backdoor criterion*** for $(\mathbf{X}, \mathbf{Y})$ and, then the effect of $\mathbf{X}$ on $\mathbf{Y}$ is given by:

$$P(\mathbf{y} \,|\, do(\mathbf{x})) = \sum_{\mathbf{z}} P\left(\mathbf{y} \,|\, \mathbf{x}, \mathbf{z}\right) P\left(\mathbf{z}\right)$$

$\mathbf{Z}$, a set of covariates, admissible for backdoor adjustment

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



$\mathbf{Z} = \{Z_1\}$

$\mathbf{Z} = \{Z_1, Z_3\}$

$\mathbf{Z} = \{Z_1, Z_2, Z_3\}$ ✘

Judea Pearl. Comment: Graphical models, causality and intervention. Stat. Sci., 8:266–269, 1993.

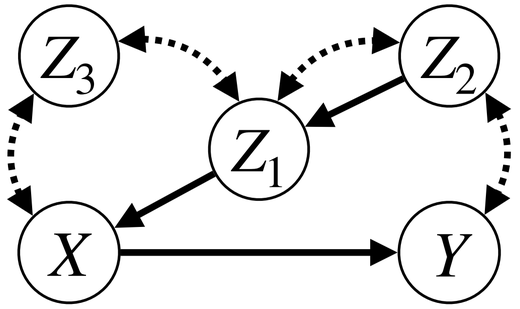
Estimation via Propensity Scores

Consider the case in which the causal effect of X on Y is identifiable through adjustment over a set of variables $\mathbf{Z}$, i.e.,

$$P(y \mid do(x)) = \sum_{\mathbf{z}} P(y \mid x, \mathbf{z}) P(\mathbf{z})$$

$$= \sum_{\mathbf{z}} \frac{P(y \mid x, \mathbf{z}) P(x \mid \mathbf{z}) P(\mathbf{z})}{P(x \mid \mathbf{z})}$$

$$= \sum_{\mathbf{z}} \frac{P(y, x, \mathbf{z})}{P(x \mid \mathbf{z})}$$



$\mathbf{Z} = \{Z_1\}$

$\mathbf{Z} = \{Z_1, Z_2\}$

Only if $\mathbf{Z}$ is admissible for adjustment, Propensity Score can be used to estimate $P(y \mid do(x))$.

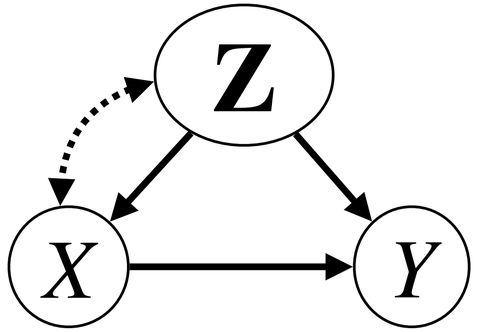
For X is binary/categorial: logistic/multinomial regression or ML-based classification

For X continuous: ML-based regression techniques.

The interventional joint distribution can be easily derived by reweighing the observational joint distribution with the inverse of the propensity score!

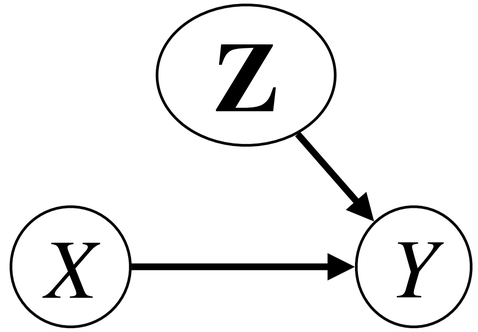
Inverse Probability Weighting (IPW)

After reweighing the observational samples, we obtain *pseudo* interventional samples:



$P(X|\mathbf{Z}) \neq P(X)$

Reweighing samples with $\frac{1}{P(X|\mathbf{Z})}$



$P(X|\mathbf{Z}) = P(X)$

Original Sample

		$P(X\|Z)$	$\frac{1}{P(X\|Z)}$
X=0 (Control Group)		1/4	4
		2/3	1.5
X = 1 (Treated Group)		3/4	1.33
		1/3	3

Imbalanced

Pseudo interventional Sample

X=0 (Control Group)	
	
X = 1 (Treated Group)	
	

Balanced

Inverse Probability Weighting (IPW)

This gives us the following estimator of $E(Y \mid do(x))$, from a sample $\{x_i, y_i, \mathbf{z_i}\}_{i=1}^{N}$:

$$\hat{E}(Y \mid do(x)) = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i \mathbf{1}_{\{x_i = x\}}}{\hat{P}(x_i \mid \mathbf{z}_i)}$$

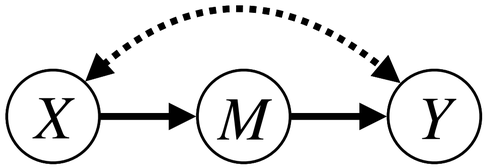
The mean of all values y_i, inversely weighted according to the propensity score.

The Average Treatment Effect (ATE) of a binary treatment can be estimated as:

$$\hat{E}(Y \mid do(X = 1)) - \hat{E}(Y \mid do(X = 0))$$

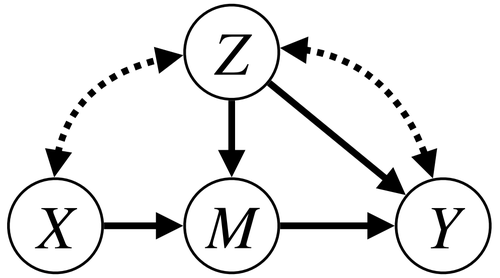
$$= \frac{1}{N} \sum_{i=1}^{N} \left(\frac{y_i \mathbf{1}_{\{x_i = 1\}}}{\hat{P}(X = 1 \mid \mathbf{z}_i)} - \frac{y_i \mathbf{1}_{\{x_i = 0\}}}{\hat{P}(X = 0 \mid \mathbf{z}_i)} \right)$$

Many Scenarios Beyond Backdoor …



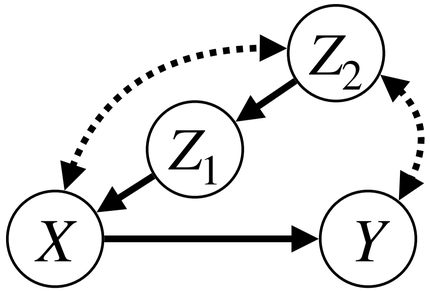
Front-Door

$$P(y \mid do(x)) = \sum_{m} P(m \mid x) \sum_{x'} P(y \mid m, x')P(x')$$



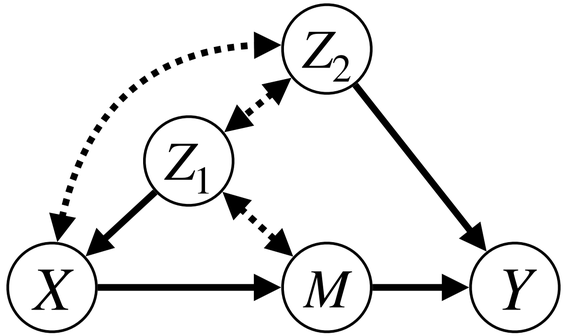
Conditional Front-Door

$$P(y \mid do(x)) = \sum_{m,z} P(m \mid x, z) \sum_{x'} P(y \mid m, x', z)P(x', z)$$



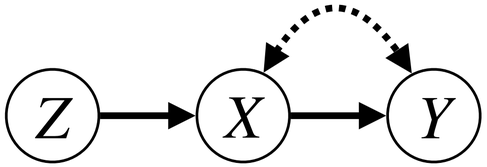
Napkin

$$P(y \mid do(x)) = \frac{\sum_{z_2} P(x, y \mid z_1, z_2)P(z_2)}{\sum_{z_2} P(x \mid z_1, z_2)P(z_2)}$$



Unnamed

$$P(y \mid do(x)) = \sum_{z_2,z_3} P(y \mid x, z_1, z_2, z_3)P(z_2) \sum_{z_1} P(z_3 \mid x, z_1)P(z_1)$$

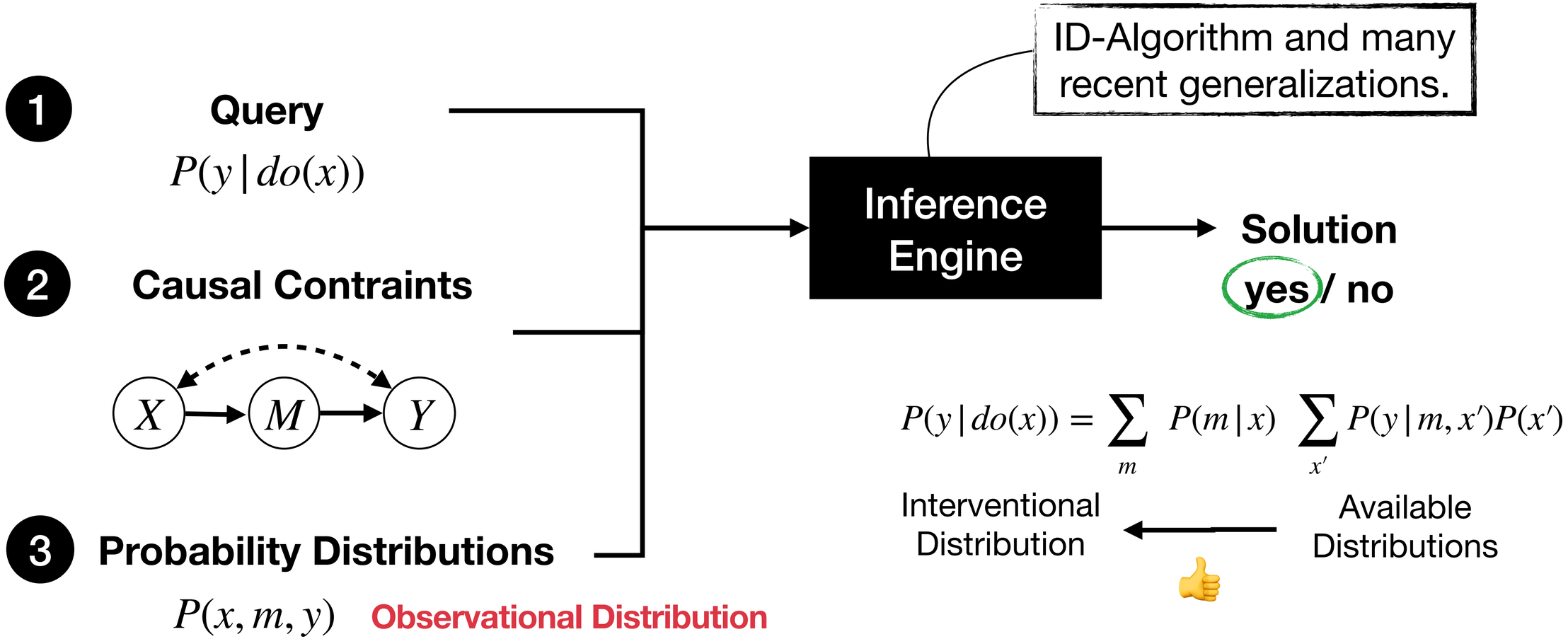


Linear IV Graph

$$\frac{d}{dx}\mathbb{E}[Y \mid \text{do}(X = x)] = \beta = \frac{Cov(Z, Y)}{Cov(Z, X)}$$

And many others….

Causal Effect Identification (ID) Algorithm



- Tian, J. and Pearl, J. A General Identification Condition for Causal Effects. In Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002), pp. 567–573, Menlo Park, CA, 2002. AAAI Press/MIT Press.

Causal Identification from PAGs

Can we identify causal effects from the equivalence class?

Effect Identification:

For Covariate Adjustment, we can use the Generalized Adjustment Criterion.

Recently, we proposed complete calculus and algorithms for the identification of marginal and conditional causal effect in PAGs!

Perkovic, E., Textor, J. C., Kalisch, M., & Maathuis, M. H. (2018). Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. Journal of Machine Learning Research 18 (2018) 1-62

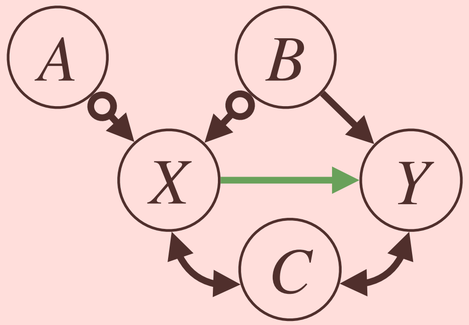
Jaber A., **Ribeiro A. H.,** Zhang, J., Bareinboim, E. (2022) Causal Identification under Markov Equivalence - Calculus, Algorithm, and Completeness. In Proceedings of the 36th Annual Conference on Neural Information Processing Systems, NeurIPS. (Link)

Effect Identification in Markov Equivalence Classes

1 **Query**

$P(y \mid do(x))$

2 **PAG**



3 $P(x, y, a, b, c)$

Observational Distribution

Causal Discovery

Can be constructed in a fully data-driven way!
(e.g. using the FCI algorithm)

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896. (Link)

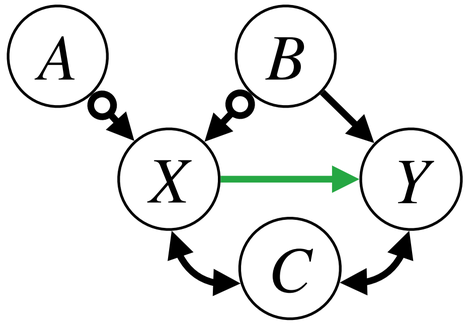
Effect Identification in Markov Equivalence Classes

Generalized Adjustment Criterion enables effect identification via adjustment in PAGs.

1 **Query**

$P(y \mid do(x))$

2 **PAG**



3 $P(x, y, a, b, c)$

Observational Distribution

Causal Discovery

Inference Engine

Solution

yes / no

$$P(y \mid do(x)) = \sum_b P(y \mid x, b)P(b)$$

$$\mathbb{E}(y \mid do(x)) = \sum_b \mathbb{E}(y \mid x, b)P(b)$$

Interventional Distribution ← Available Distributions

👍

Perkovic, E., Textor, J. C., Kalisch, M., & Maathuis, M. H. (2018). Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. Journal of Machine Learning Research 18 (2018) 1-62

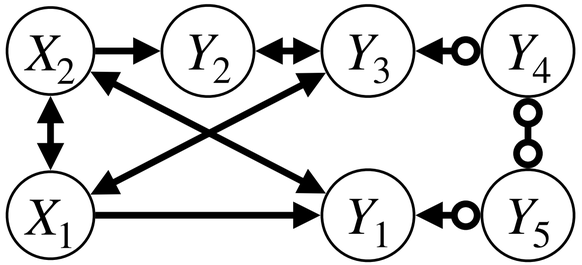
General Identification in Markov Equivalence Classes

Complete algorithms (IDP, CIDP).
PAGId R package:
https://github.com/adele/PAGId

1 **Query**

$P(y \mid do(x))$

2 **PAG**



3 **Data**

$P(x, y, z, w)$ **Observational Distribution**

IDP / CIDP

Solution

yes / no

$$P(y \mid do(x)) = \frac{P(y_1, y_4, y_5 \mid x_1) \,.\, P(y_2 y_3, y_4, y_5 \mid x_2)}{P(y_4, y_5)}$$

Inferred (Interventional) Distribution ← Available (Observational) Distribution

Jaber, A.*, **Ribeiro, A. H.***, Zhang, J., Bareinboim, E. (2022). Causal Identification under Markov equivalence: Calculus, Algorithm, and Completeness. *In Adv. Neural Inf Process Systems, 35, 3679-3690. (NeurIPS-22).* (Link)

Gut Microbiota's Causal Role in Major Depressive Disorder

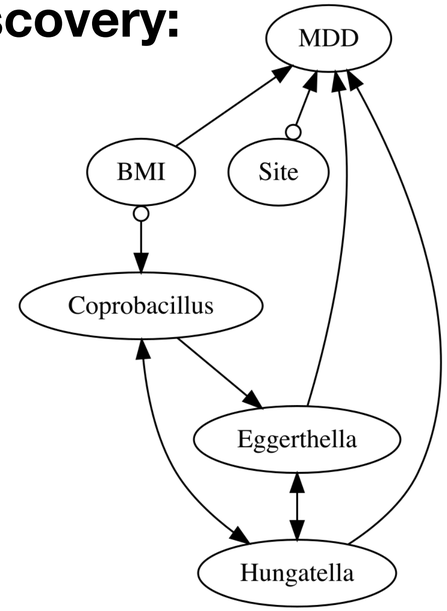
DFG FOR2107 dataset, including microbiome and clinical data from 1,269 patients.

Differential Abundance Analysis:

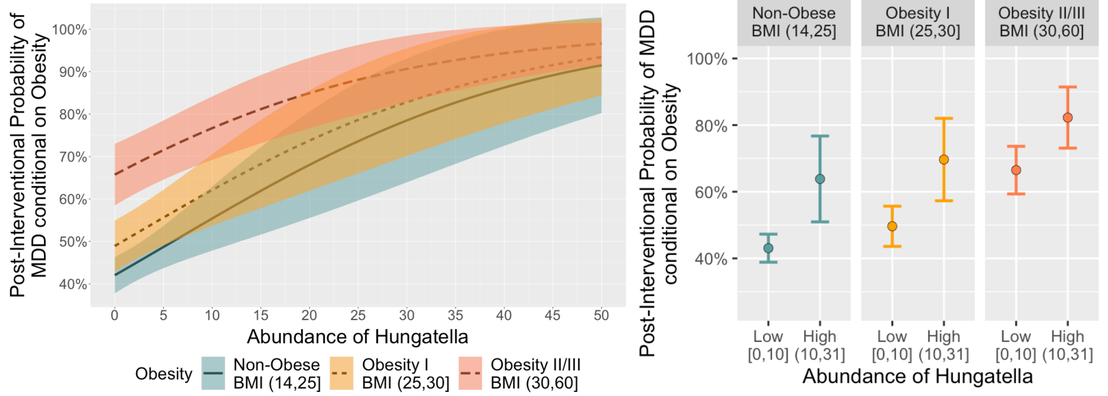
Genus	FDR-corr. p-values	
	LinDA	ZicoSeq
Hungatella	0.0002	0.0071
Eggerthella	0.0063	0.0071
Coprobacillus	0.0070	0.0071
Lachnospiraceae FCS020 group	0.0063	0.011

Causal Discovery:

FCI with a robustness-enhancing strategy.



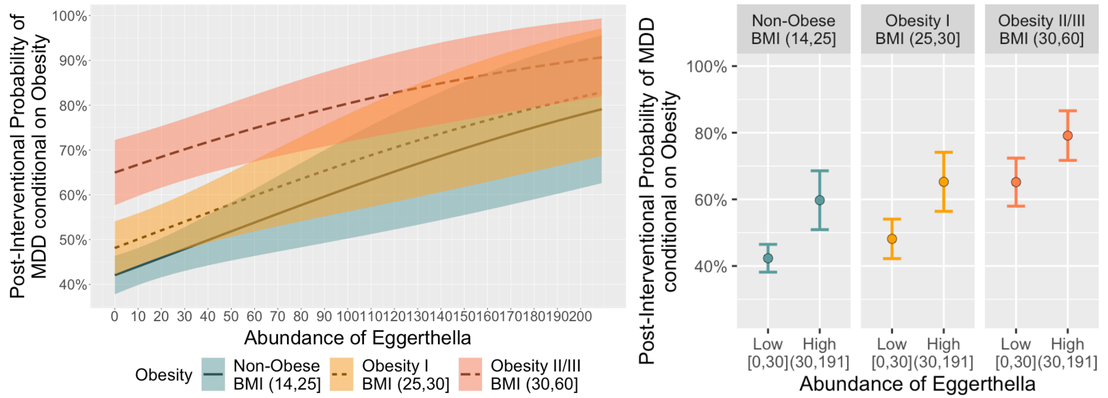
Obesity-specific causal effect of *Hungatella* on MDD



Obesity	Effect	95% C.I.	
No	0.2078	0.07230	0.3433
I	0.2002	0.06261	0.3379
II/II	0.1577	0.04138	0.2741

ACE: 0.1984
95% CI: [0.06607, 0.3307]

Obesity-specific causal effect of *Eggerthella* on MDD



Obesity	Effect	95% C.I.	
No	0.1742	0.07659	0.2718
I	0.1712	0.06460	0.2778
II/II	0.1397	0.03607	0.2433

ACE: 0.1683
95% CI: [0.06812, 0.2684]

Fehse L.*, **Ribeiro, A.H.***, Winter, N. R., …, Dominik Heider, Tim Hahn. (2024) From Gut to Mind: Dissecting the Causal Contribution of Gut-Microbiota to Major Depressive Disorder in Humans. MedRxiv Preprint, DOI: 10.1101/2024.12.05.24318549 * Equal contribution

Current Focus: Causality in Complex Systems

Application Side:

1. ***Malaria Research —*** *Identify and understand the factors influencing individual malaria risk in Brazil's Amazon (complex relational and dependency structure).*

2. ***Causal Analysis from Single-Cell Multi-Omics Data —*** *Inferring gene regulatory mechanisms from high-dimensional, multi-modal single-cell measurements.*

3. ***High-level modeling of Causal Functional Brain Networks —*** *Causal structure learning from real neuroimaging and simulated neuronal data, while identifying sample units (causal representation learning), and mapping low-levelto high-level dynamics (causal abstraction).*

4. ***Epidemiological, Longitudinal Studies —*** *Causal inference for cancer, cardiometabolic disease, and other chronic conditions tracked over time, typically in a privacy preserving setting.*

Current Focus: Causality in Complex Systems

Methodological Side:

1. ***Robust Inference Integrating Experimental Data:*** Causal discovery and inference from observational and experimental data under uncertainty, heterogeneity, and multi-modality.

2. ***Heterogeneity and Multi-Modality:*** Conditional independence tests for non-linear, multimodal data with complex dependency structures (e.g., temporal, spatial, genetic).

3. **Cyclic & Dynamic Modeling:** Causal discovery and inference for systems with feedback loops and time-varying structure, beyond standard acyclic assumptions.

4. ***Scalability and Goal-Oriented Learning:*** Adaptive discovery of larger causal graphs, expanding variable sets based on user-defined goals or score thresholds.

Many Other Topics in Causal Inference

1. Causal Representation Learning & Causal Abstraction
2. Causal Reinforcement Learning
3. Data-Driven Covariate Selection for Adjustment and Transportability
4. Partial Effect Identification
5. Causal Design of Experiments
6. Individual Treatment Effect (ITE) Estimation
7. Fairness & Mediation Analysis
8. Many more…

I am happy to discuss more if you are interested! :)