Causality and its Role in Reasoning, Explainability, and Generalizability

Adèle H. Ribeiro

https://adele.github.io/ | adele.ribeiro@uni-marburg.de

Faculty of Mathematics and Computer Science
Philipps-Universität Marburg

Lisbon Machine Learning School (LxMLS)
July 20th, 2023

Recent Breakthroughs in AI

- We can learn models that makes **predictions** extremely well in high-dimensional settings.
- In particular, there are huge progresses in natural processing language, computer vision, and reinforcement learning.

Recent Breakthroughs in AI

The New York Times

Driverless Cars
/ to San Jose

Aside from the lidar range-finder unit on its r... Google's fleet of self-
icles, including this Lexus hybrid, look reasonably conventional. Google

NEWSLETTERS
Sign up to read our regular email newsletters

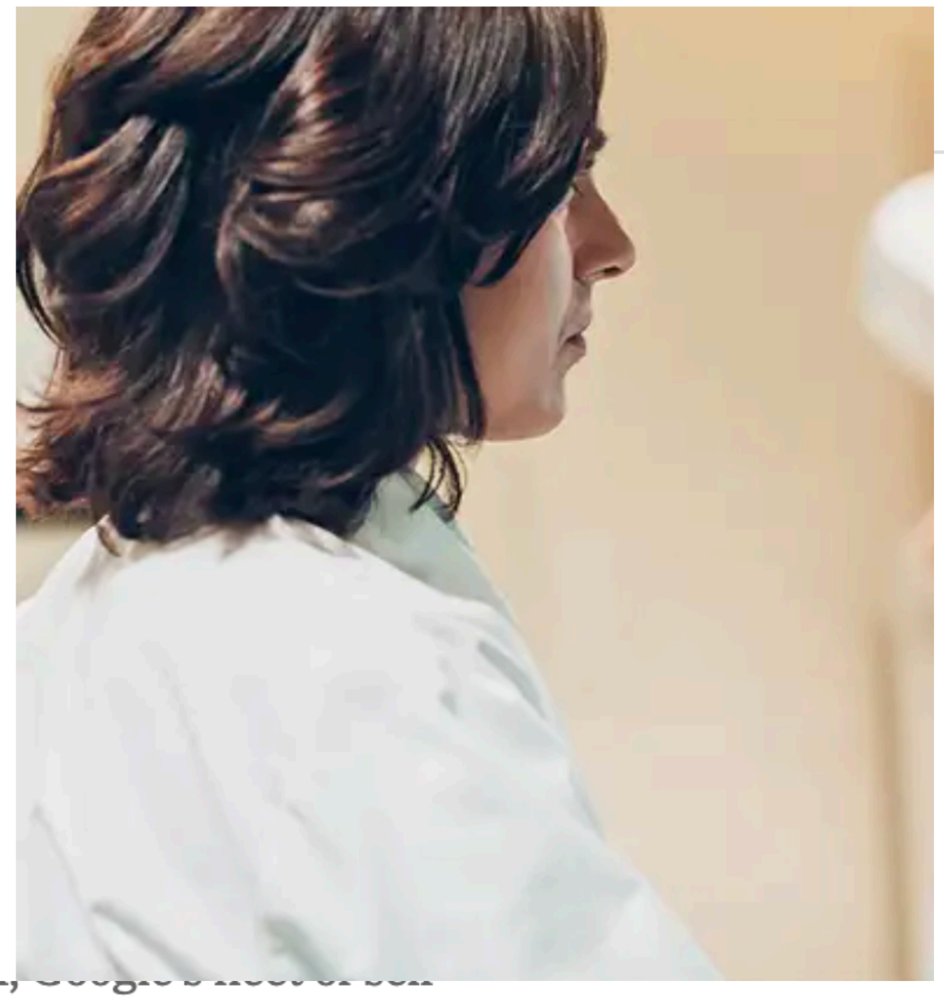
NewScientist

News Podcasts Video Technology Space Physics Health More Shop Courses Events

AI system is better than human doctors at predicting breast cancer

TECHNOLOGY 1 January 2020

By Jessica Hamzelou



The New York Times

Become an A.I. Expert | How Chatbots Work | Why Chatbots 'Hallucinate' | Testing Google Bard

THE SHIFT

GPT-4 Is Exciting and Scary

Today, the new language model from OpenAI may not seem all that dangerous. But the worst risks are the ones we cannot anticipate.

Current Challenges in AI

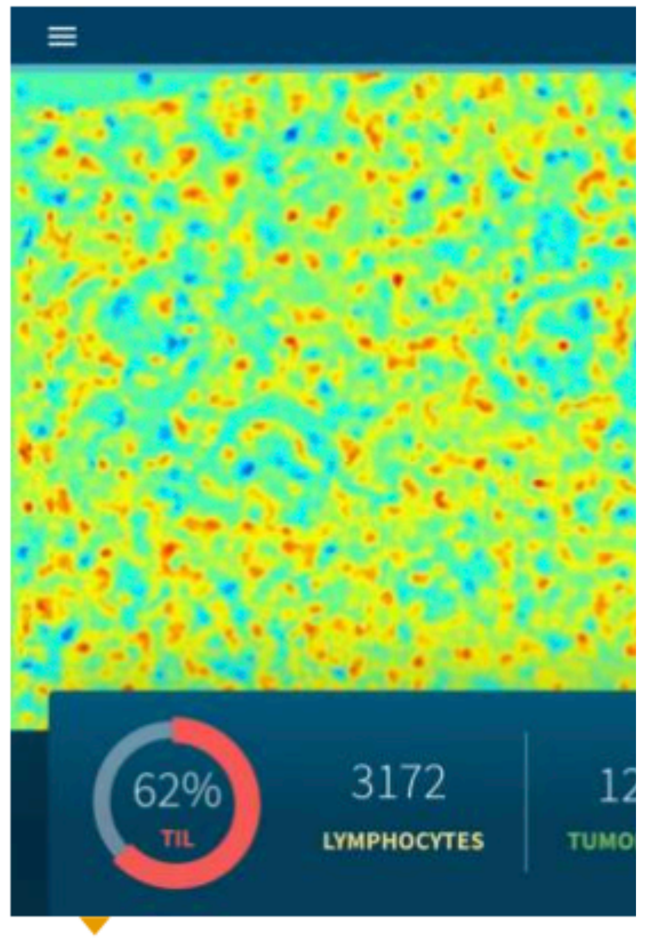
itn
IMAGING TECHNOLOGY NEWS

COVID-19 IMAGING WO HEA

NEWS | ARTIFICIAL INTELLIGENCE | MARC

Making the Role of AI in

Analysis system for the diagnosis of



Detection of tumor-infiltrating lymphocy generate a heatmap showing TILs (red) a of Klauschen/Charité

nature

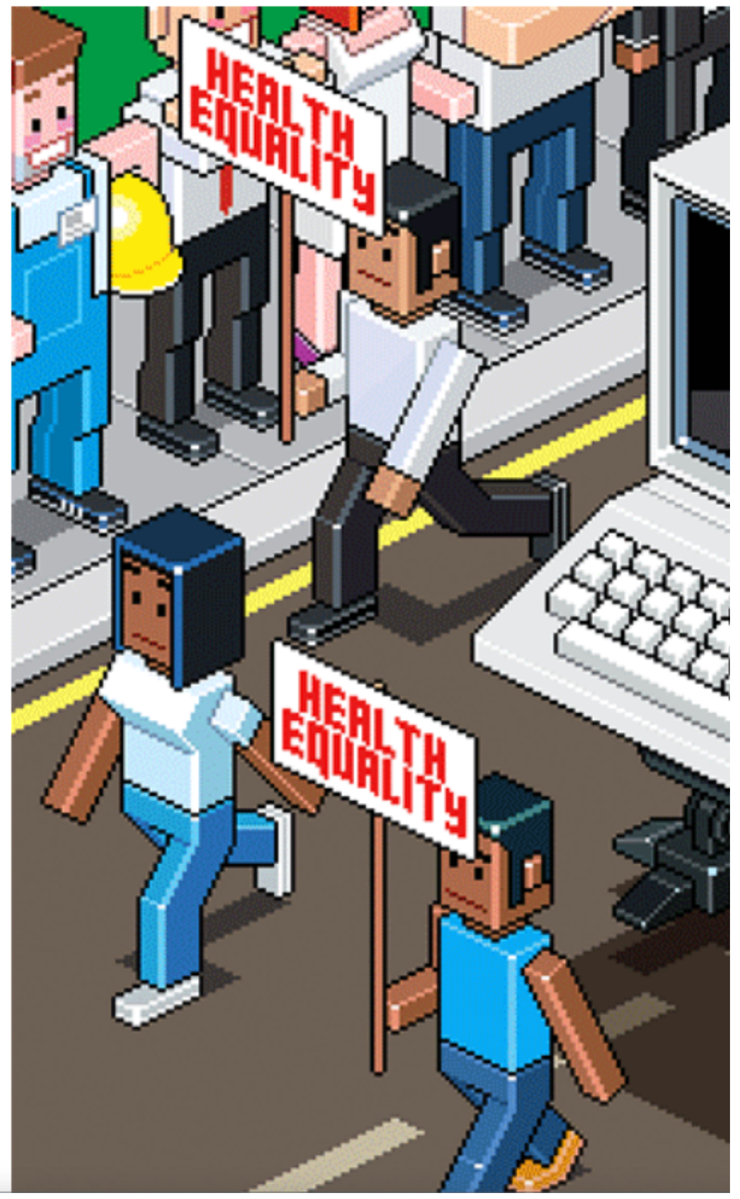
Explore content Journal information Publish with us Subscribe

nature > outlook > article

A fairer way forward for AI in health care

Without careful implemen inequality.

Linda Nordling



nature

Explore content About the journal Publish with us Subscribe

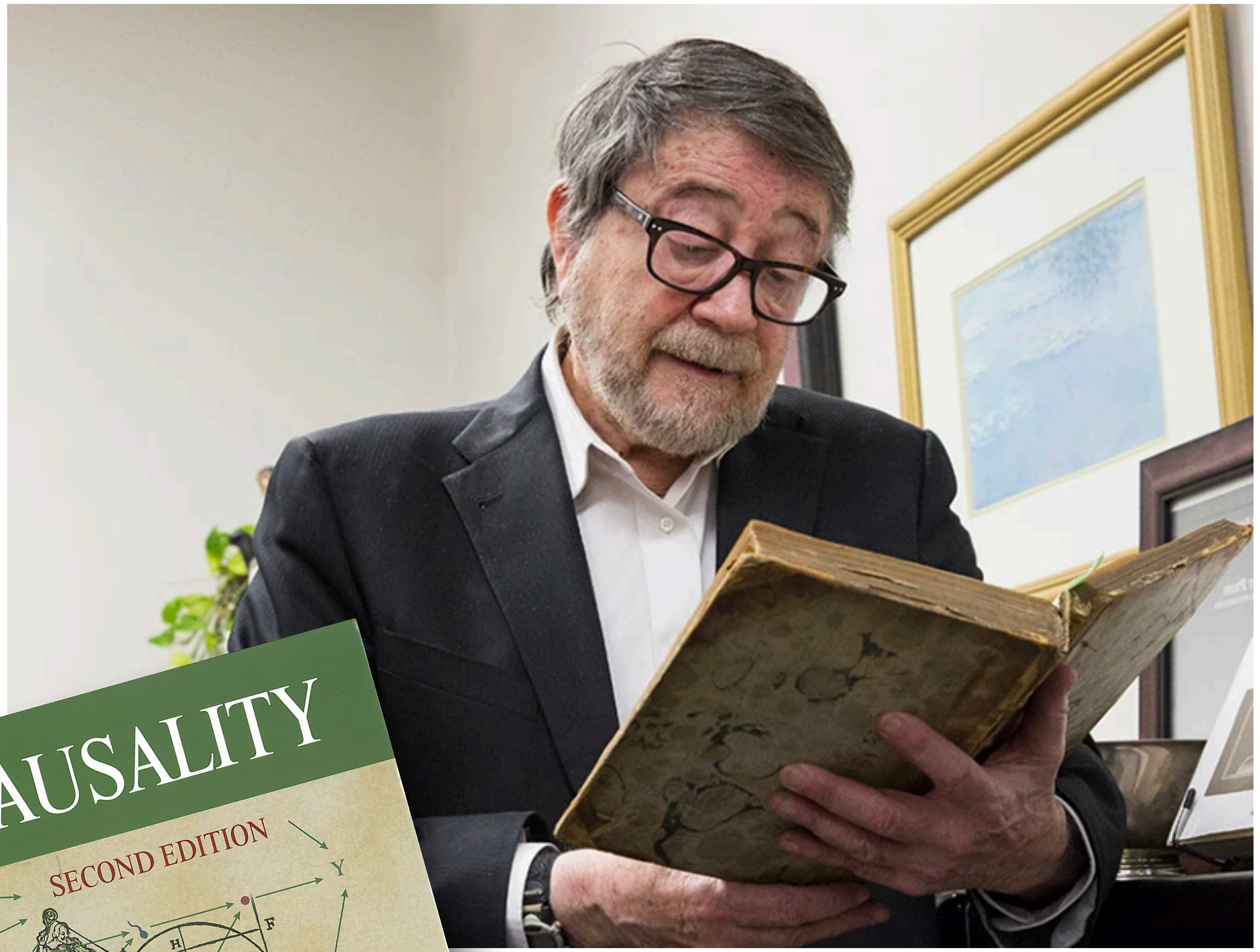
nature > outlook > article

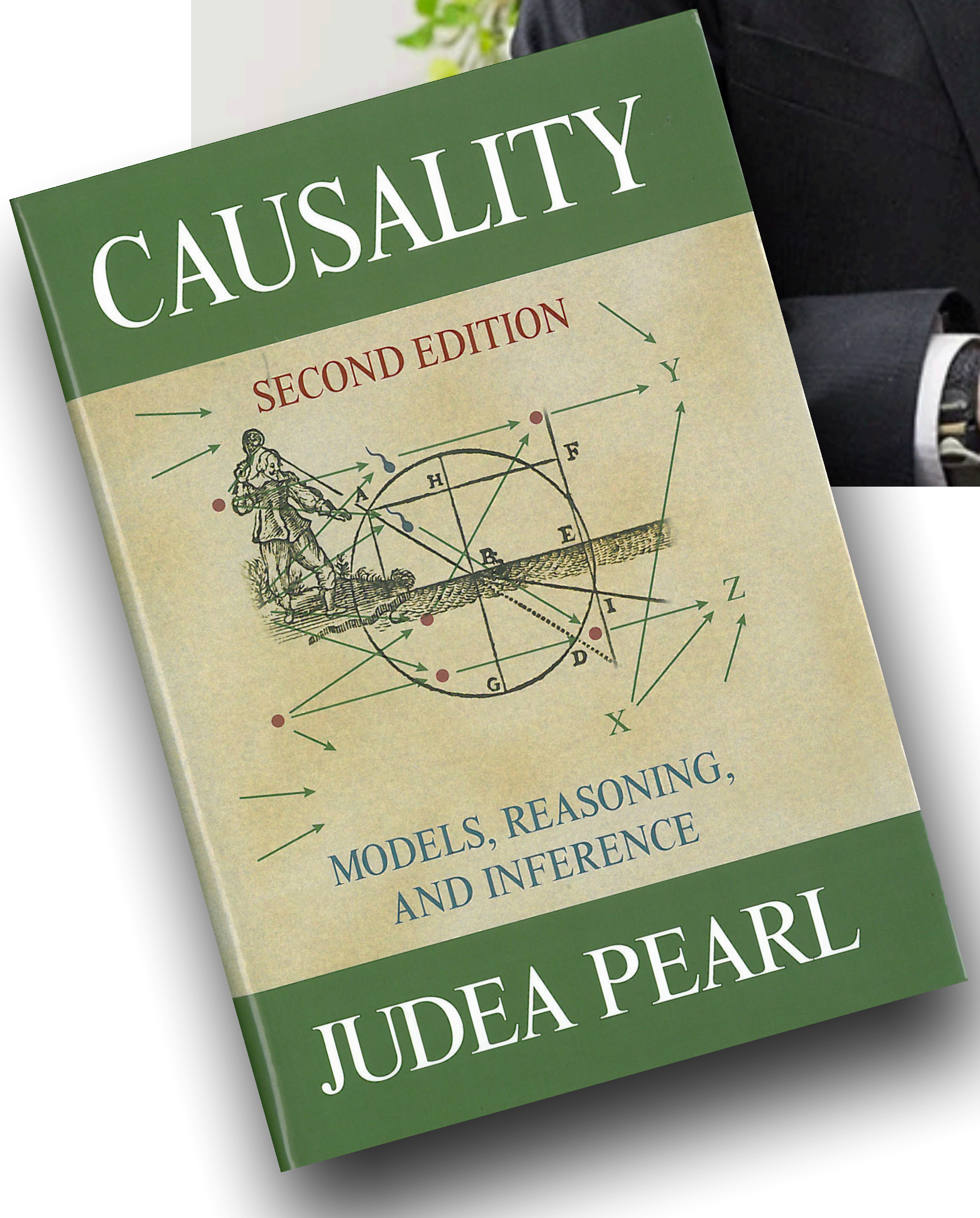
OUTLOOK | 24 February 2023

Why artificial intelligence needs to understand consequences

A machine with a grasp of cause and effect could learn more like a human, through imagination and regret.

Judea Pearl — Causality





Director of the Cognitive Systems Laboratory at the University of California, Los Angeles.

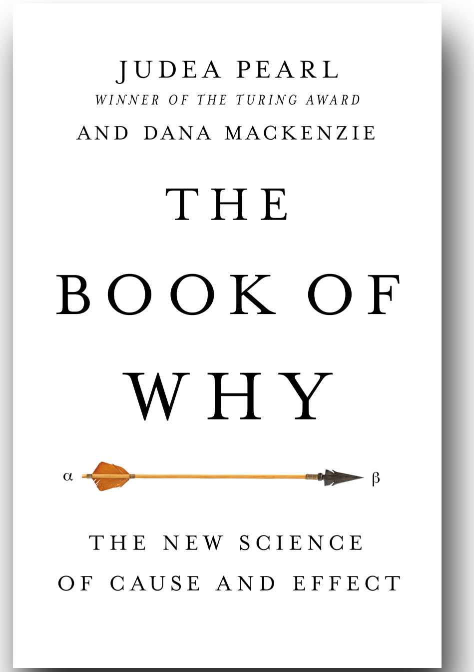
In 2011, he won the A. M. Turing Award (the highest distinction in computer science and a $250,000 prize)

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

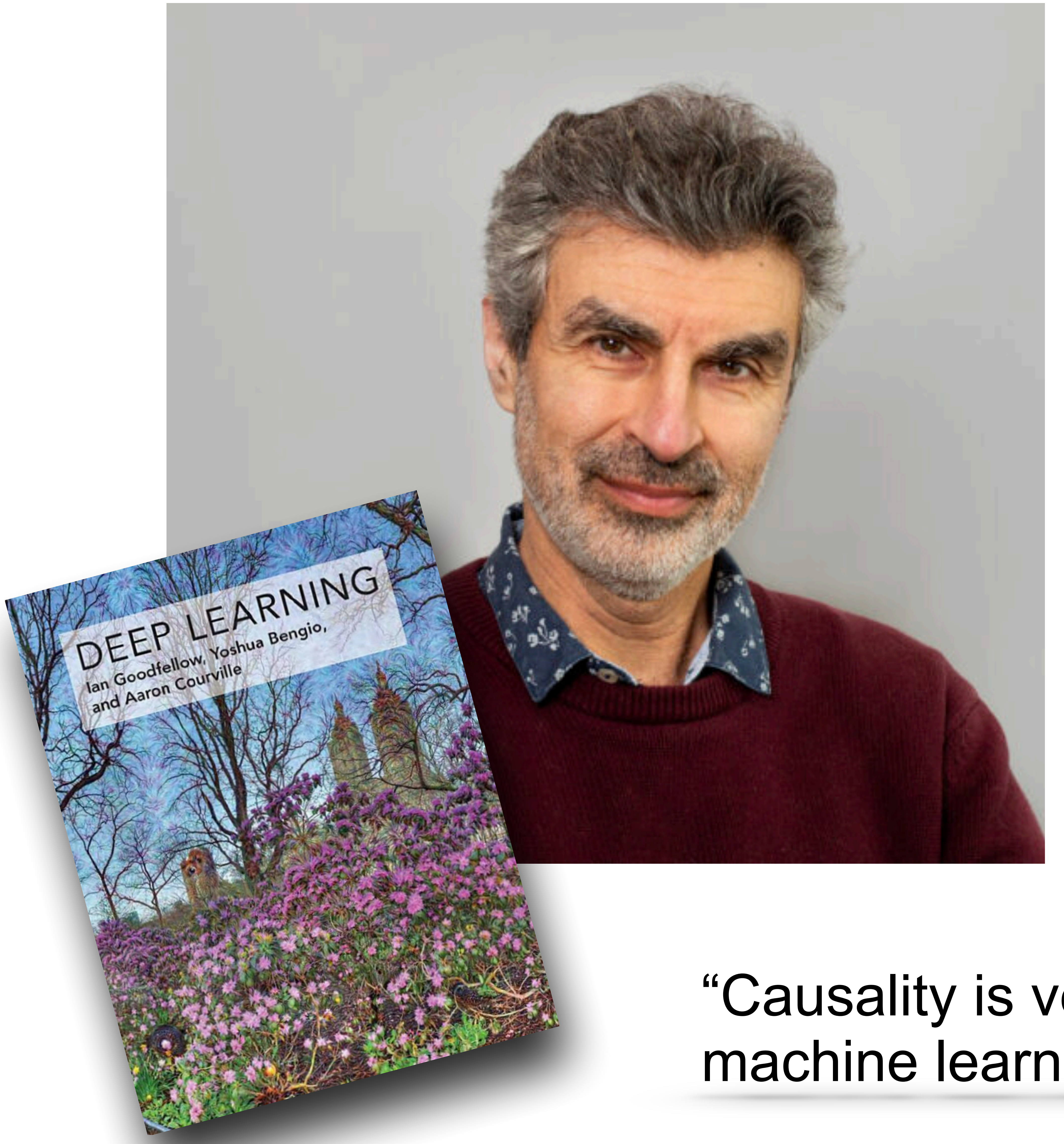
— Association for Computing Machinery (ACM)

“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



Yoshua Bengio — Deep Learning



Professor at the University of Montreal, and the Founder and Scientific Director of Mila – Quebec AI Institute

In 2018, he won the A. M. Turing Award, with Geoffrey Hinton, and Yann LeCun

“for conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing.”

— Association for Computing Machinery (ACM)

“Causality is very important for the next steps of progress of machine learning,” — interview with *IEEE Spectrum*, 2020.

Guido W. Imbens & Joshua D. Angrist

In 2021, they won the Nobel Prize in Economics (about $1 million)

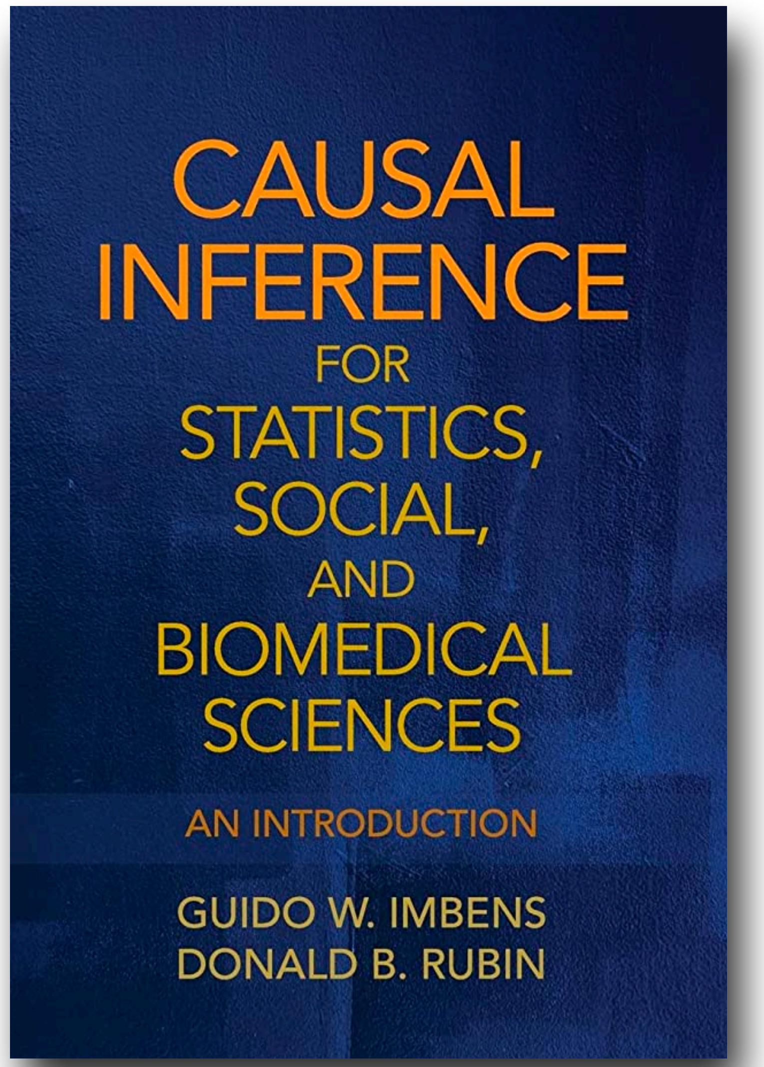
“for their methodological contributions to the analysis of causal relationships”

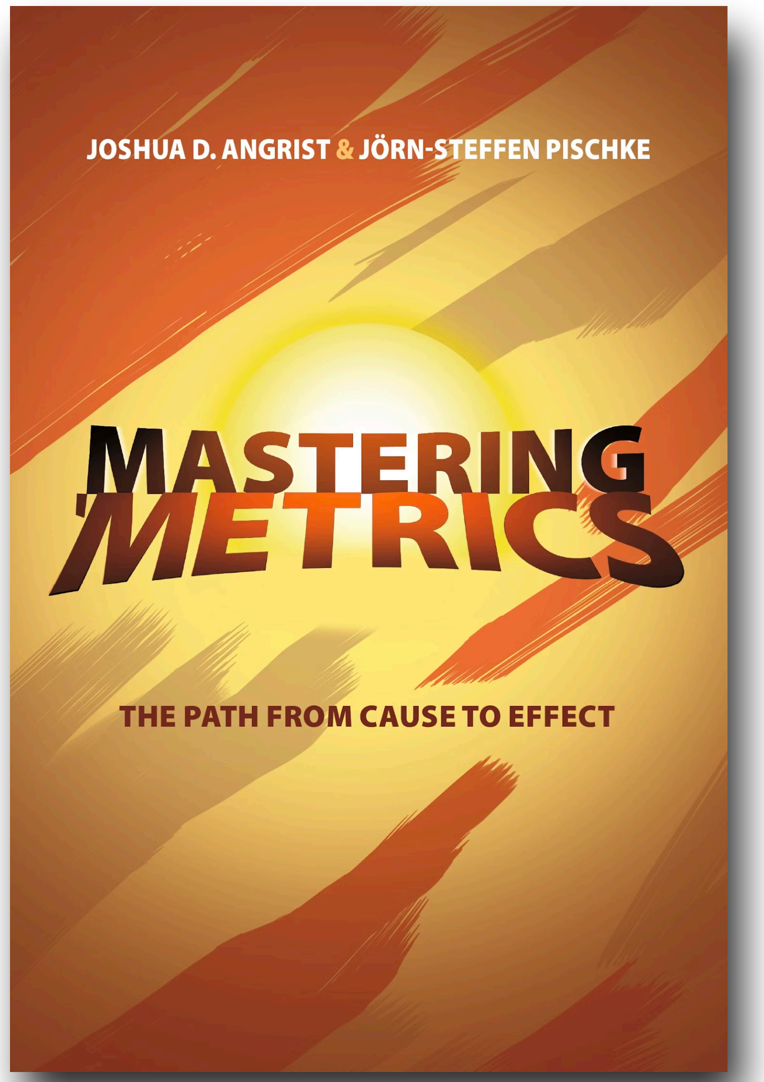
Guido W. Imbens

Professor of Applied Econometrics in Stanford University

Joshua D. Angrist

Professor of Economics at the Massachusetts Institute of Technology





Why causality is so important?

Causality is an essential component in the development of the new generation of Artificial Intelligence methods, allowing important capabilities such as

Explainability: provides a better understanding of the underlying mechanisms, e.g., learning directionality and confounding through causal structure learning.

Reasoning: can determine the effect of *unrealized* interventions rather than just predicting an outcome (i.e., can distinguish between association and causation).

Fairness: captures and disentangles any mechanisms of discrimination that may be present, including direct, indirect-mediated, and indirect-confounded.

Generalizability: allows the transportability of causal effects across different domains.

Data Fusion: provides language and theory to cohesively combine prior knowledge and data from multiple and heterogeneous studies.

Causal Data Science

Goal is to develop language, criteria, and algorithms for:

- **Data-Fusion:** cohesively combining heterogenous datasets,
- **Causal Inference:** inferring the effects of interventions, and
- **Decision-Making:** making robust and generalizable decisions.

PNAS

Causal inference and the data-fusion problem

Elias Bareinboim[a,b,1] **and Judea Pearl**[a]

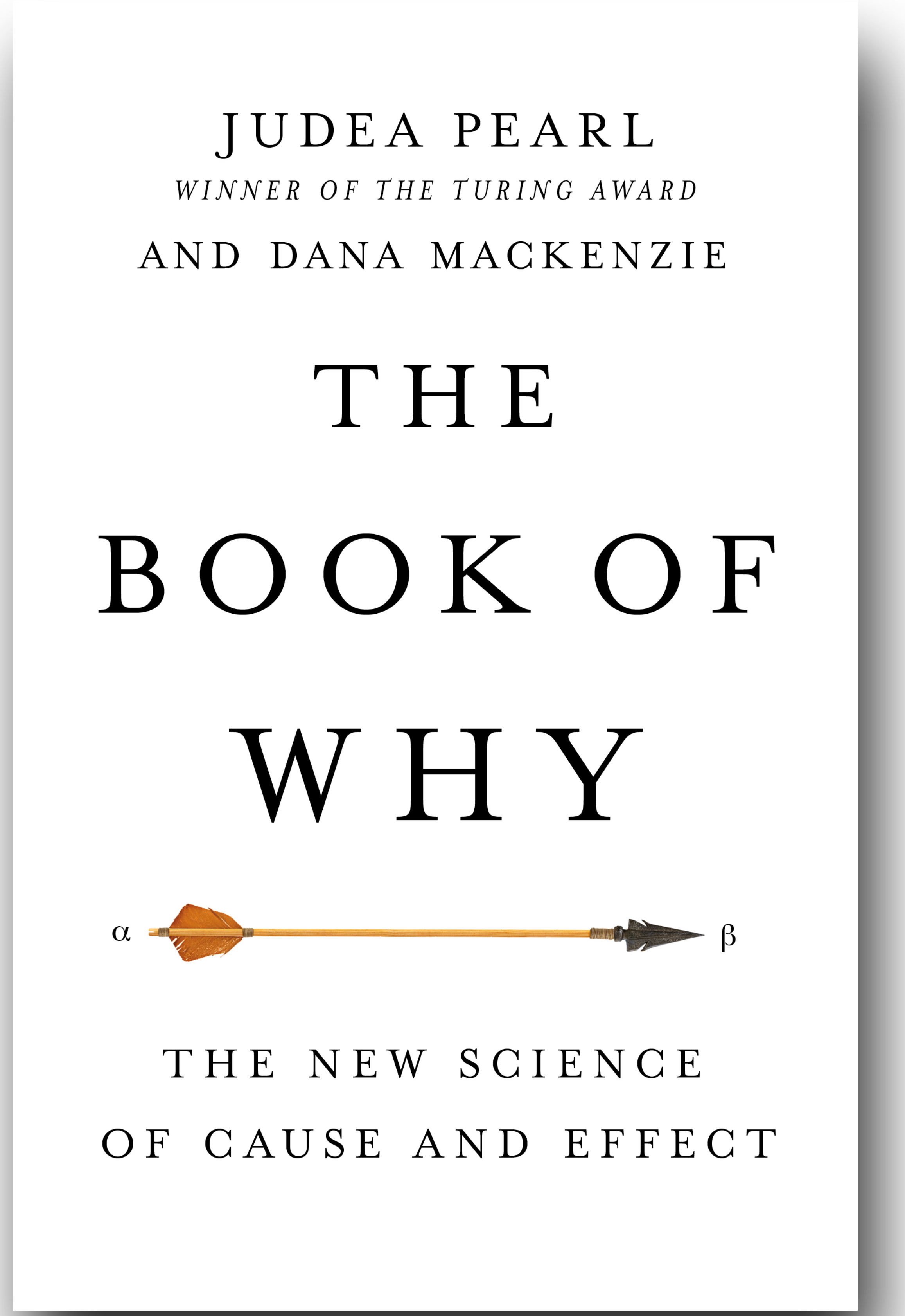
[a]Department of Computer Science, University of California, Los Angeles, CA 90095; and [b]Department of Computer Science, Purdue University, West Lafayette, IN 47907

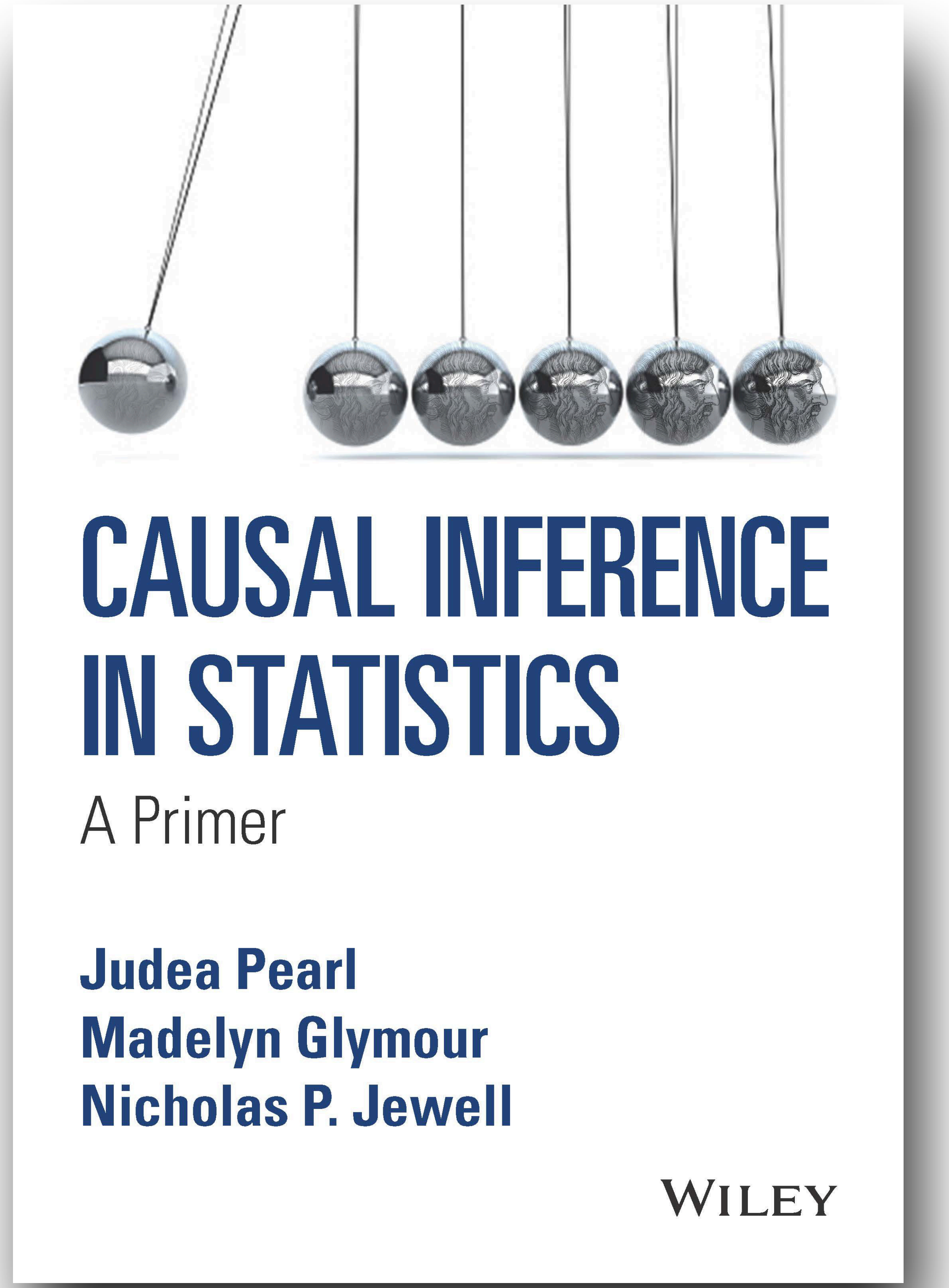
Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved March 15, 2016 (received for review June 29, 2015)

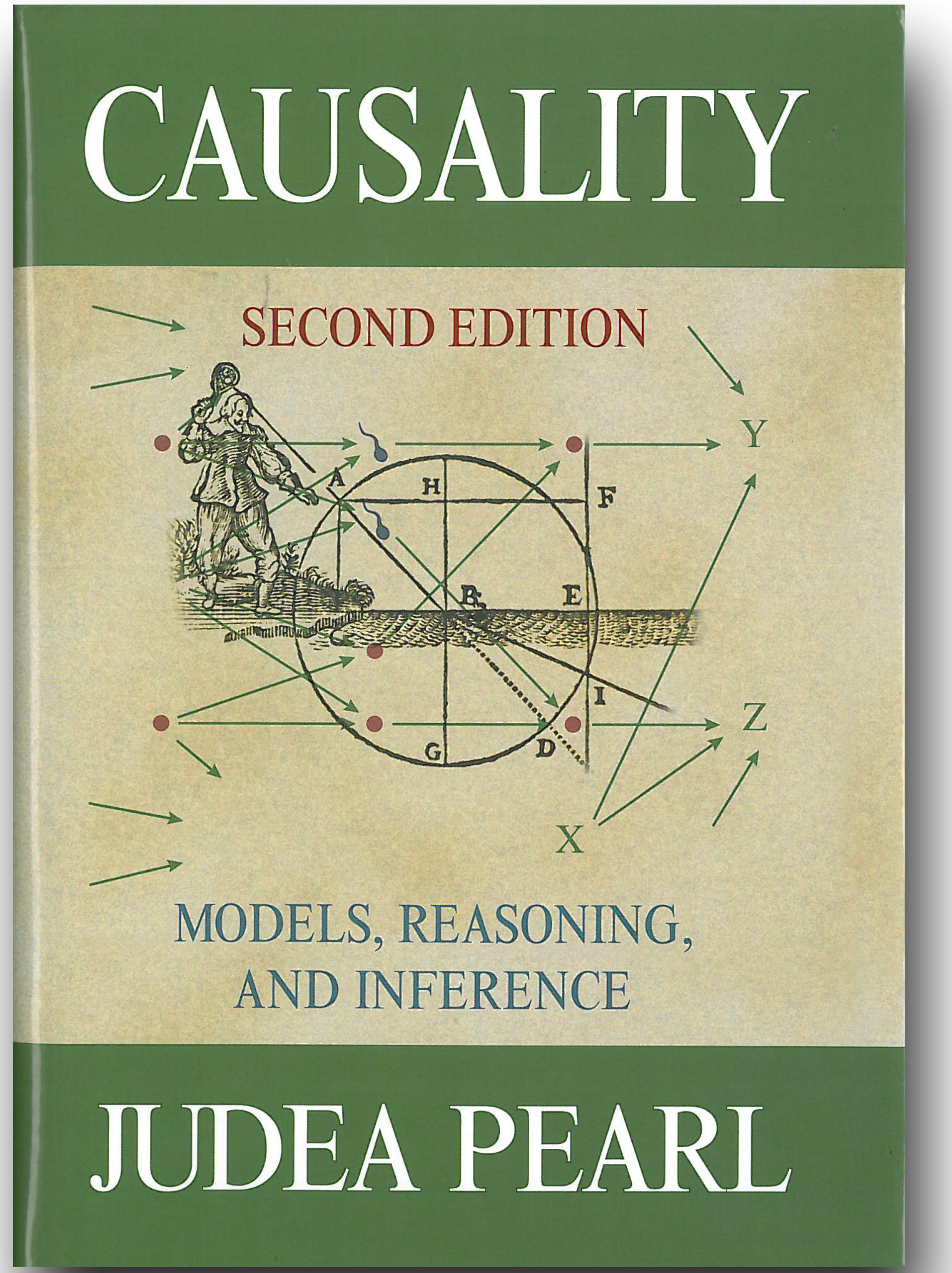
http://causalfusion.net

9

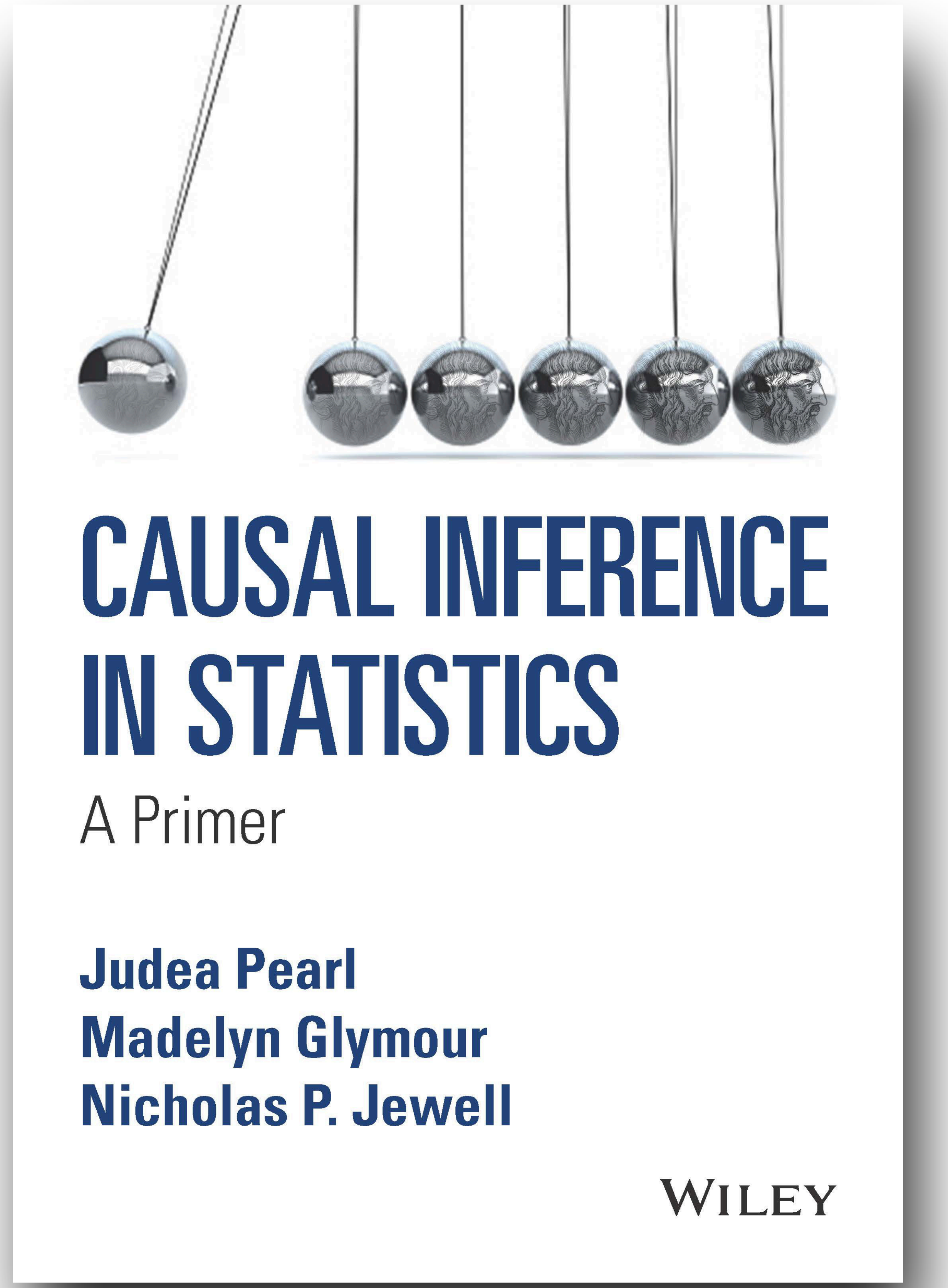
Causality Theory by Judea Pearl



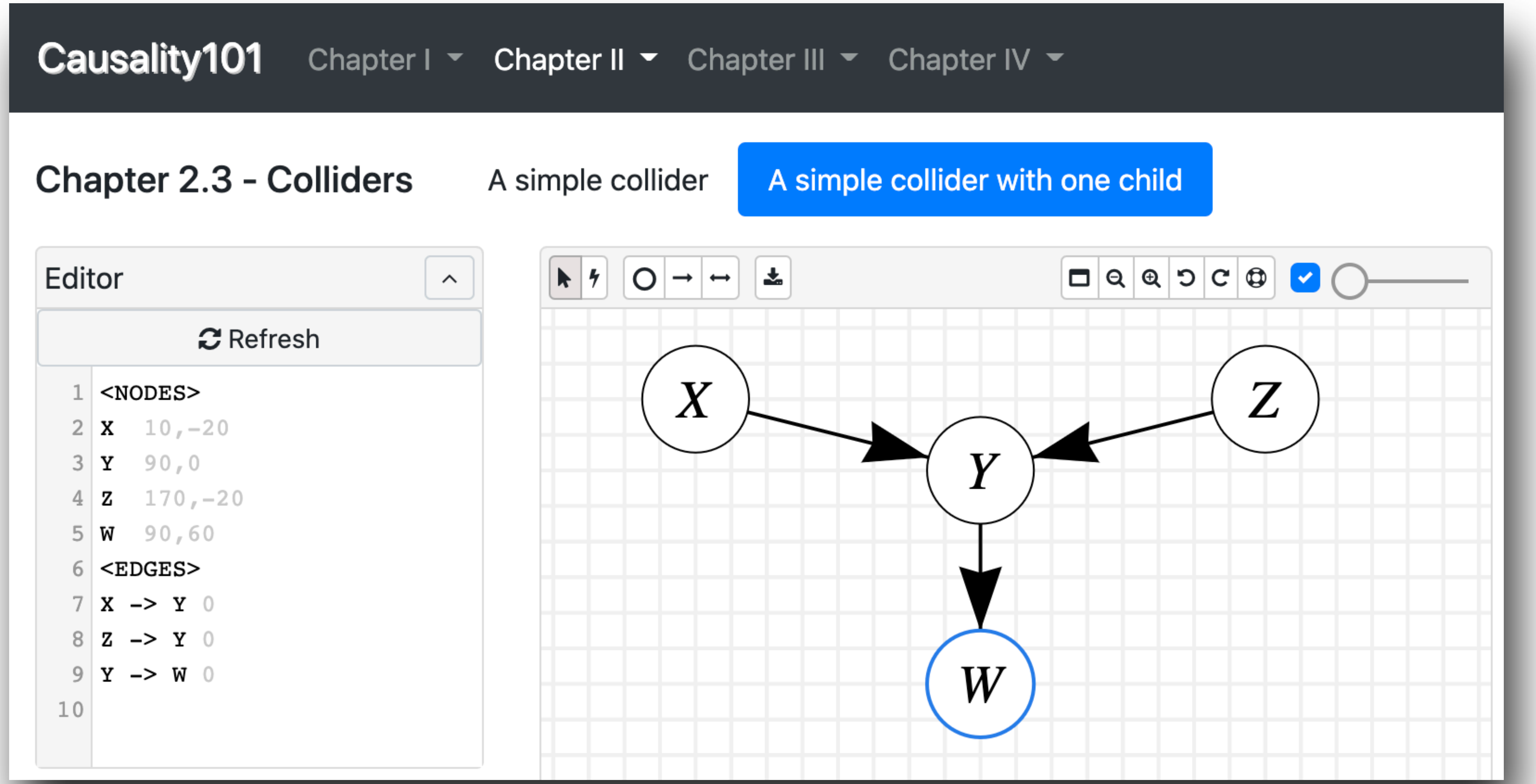




Causality Theory by Judea Pearl



https://causality101.net/



Prediction vs Reasoning

Statistical Association vs Causation

Predictive Tasks

Task: Can I guess how serious/big is the fire by the number of firefighters in action?

Yes!

X: **Number of firefighters in action**

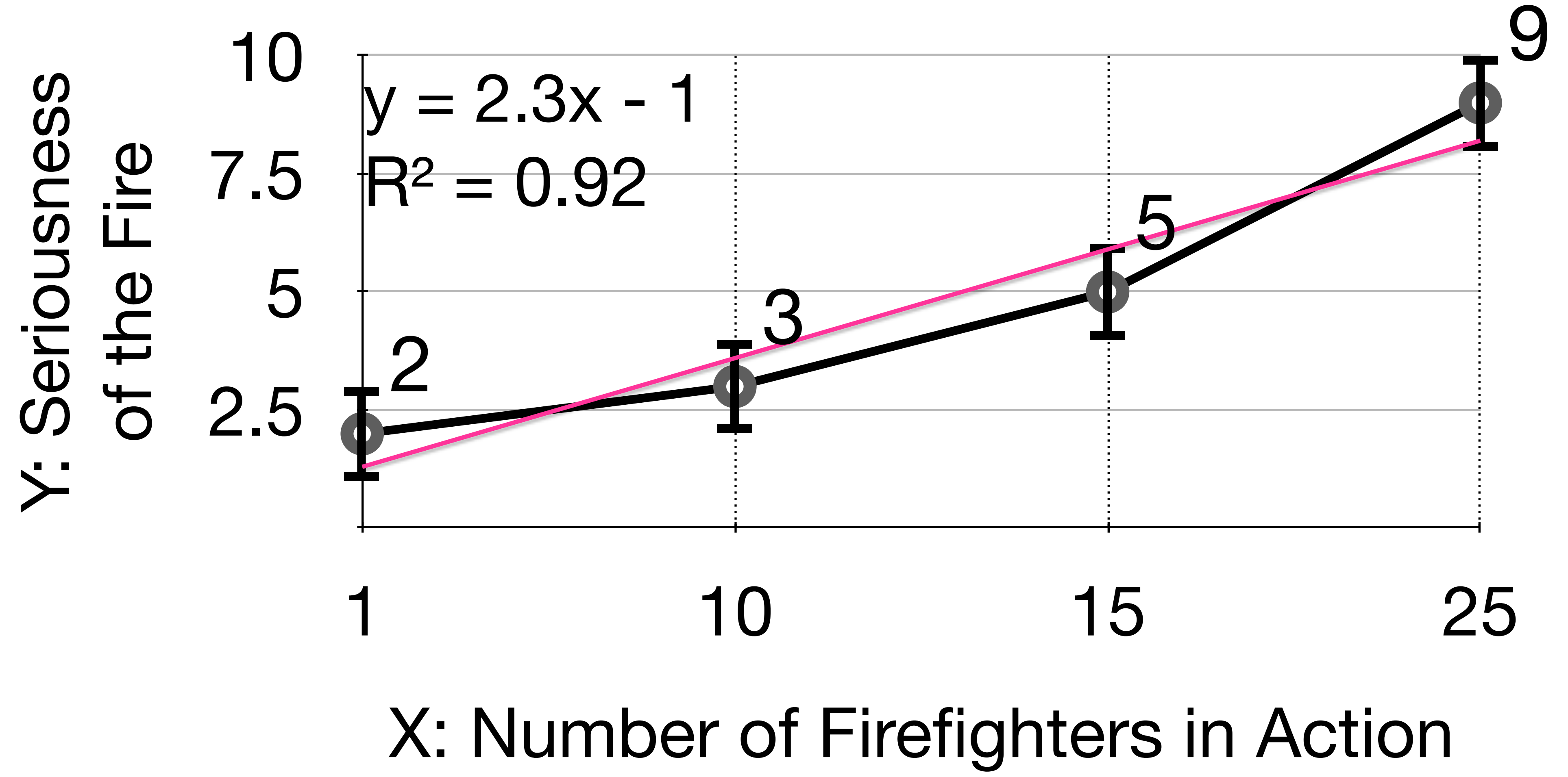
Y: **Seriousness of fire**

$\rho_{XY} \neq 0 \implies X$ **is a good predictor of** Y

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

Observational Probability Distribution

Correlation between seriousness of fire and number of firefighters in action



Conclusion: The seriousness of fire increases with the number of firefighters.

Prediction ⇒ Decision-Making / Reasoning?

Conclusion: The size of the fire increases with the number of firefighters.

In other words, the fewer the firefighters, the smaller the fire.

Should we decrease the number of firefighters to reduce the fire?

Effect of Interventions

X: Number of firefighters in action

Y: Seriousness of fire

$$\begin{cases} X = f_X(Y, U_X, U_{XY}) \\ Y = f_Y(U_Y, U_{XY}) \end{cases}$$

Underlying Model

Y is not a function of X

In other words, Y **is not caused by** X

Effect of Interventions

X: Number of firefighters in action

Y: Seriousness of fire

$$\begin{cases} \cancel{X = f_X(Y, U_X, U_{XY})} \quad X = x \\ Y = f_Y(U_Y, U_{XY}) \end{cases}$$

Underlying Model

Y is not a function of X

In other words, Y **is not caused by** X

Changing X won't change the value of Y

$$P(Y = y \mid do(X = x)) = P(Y = y)$$

Interventional Probability Distribution

The action/intervention on X, $do(X = x)$ is independent of Y

Conclusion: we cannot change the size of the fire by changing the number of firefighters.

Structural Causal Model (SCM)

EXPLAINABILITY AND THE DATA GENERATING MODEL

Structural Causal Model (SCM)

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

- $\mathbf{V} = \{V_1, \ldots, V_n\}$: are endogenous variables
- $\mathbf{U} = \{U_1, \ldots, U_m\}$: are exogenous variables
- $\mathcal{F} = \{f_1, \ldots, f_n\}$: 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 and $U_i \subseteq \mathbf{U}$ are exogenous causes of V_i.
- $P(\mathbf{U})$ is the probability distribution over $\mathbf{U}$.

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

Effect of Interventions in SCMs

Pre-Interventional/ Observational SCM

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

$do(X = x)$

Post-Interventional / Interventional SCM

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

Observational Distribution

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

$\neq$

Interventional Distribution

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

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

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

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

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

Structural Equation Model (SEM)

$$\mathcal{M} = \begin{cases} \mathbf{V} = \{X, Y, Z\} \\ \mathbf{U} = \{\epsilon_X, \epsilon_Y, \epsilon_Z\} \\ \mathcal{F} = \begin{cases} Z = \beta_{Z0} + \epsilon_Z \\ X = \beta_{X0} + \beta_{XZ} Z + \epsilon_X \\ Y = \beta_{Y0} + \beta_{YZ} Z + \beta_{YX} X + \epsilon_Y) \end{cases} \\ \mathbf{U} \sim \mathcal{N} \left(\mathbf{0}, \Sigma = \begin{bmatrix} \sigma_X & 0 & 0 \\ 0 & \sigma_Y & 0 \\ 0 & 0 & \sigma_Z \end{bmatrix} \right) \end{cases}$$

- **Linear functions**
- **Normal distribution**
- **Markovianity / Causal Sufficiency**: Error terms in $\mathbf{U}$ are independent of each other (diagonal covariance matrix).

Full specification of an SCM <u>requires</u> parametric and distributional assumptions.

Estimation of such models usually requires strong assumptions (e.g., Markovianity).

SCM: Encoder of Functional Knowledge

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

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!

Encoding Structural Causal Knowledge

Acyclic Directed Acyclic Graph (ADMG)
Causal Diagrams

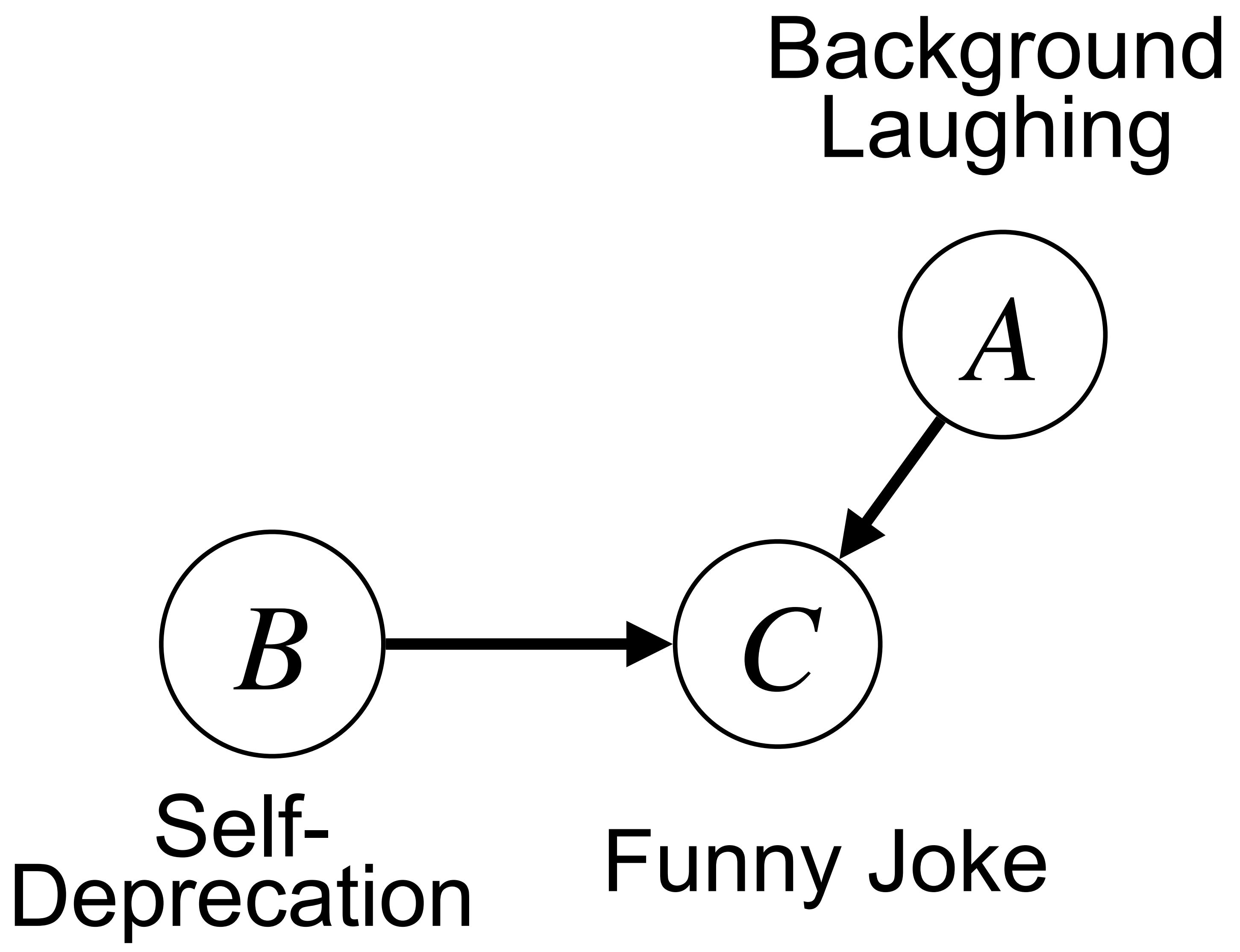
Causal Diagram: Encoder of Structural Knowledge

Structural Causal Model (SCM)

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

$$\mathcal{M} = \begin{cases} \mathbf{V} = \{A, B, C\} \\ \mathbf{U} = \{U_A, U_B, U_C, U_{AB}\} \\ \mathcal{F} = \begin{cases} A \leftarrow f_A(U_{AB}, U_A) \\ B \leftarrow f_B(U_{AB}, U_B) \\ C \leftarrow f_C(A, B, U_C) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Induced Causal Diagram



An SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$ 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 $f_j \in \mathcal{F}$.

Causal Diagram: Encoder of Structural Knowledge

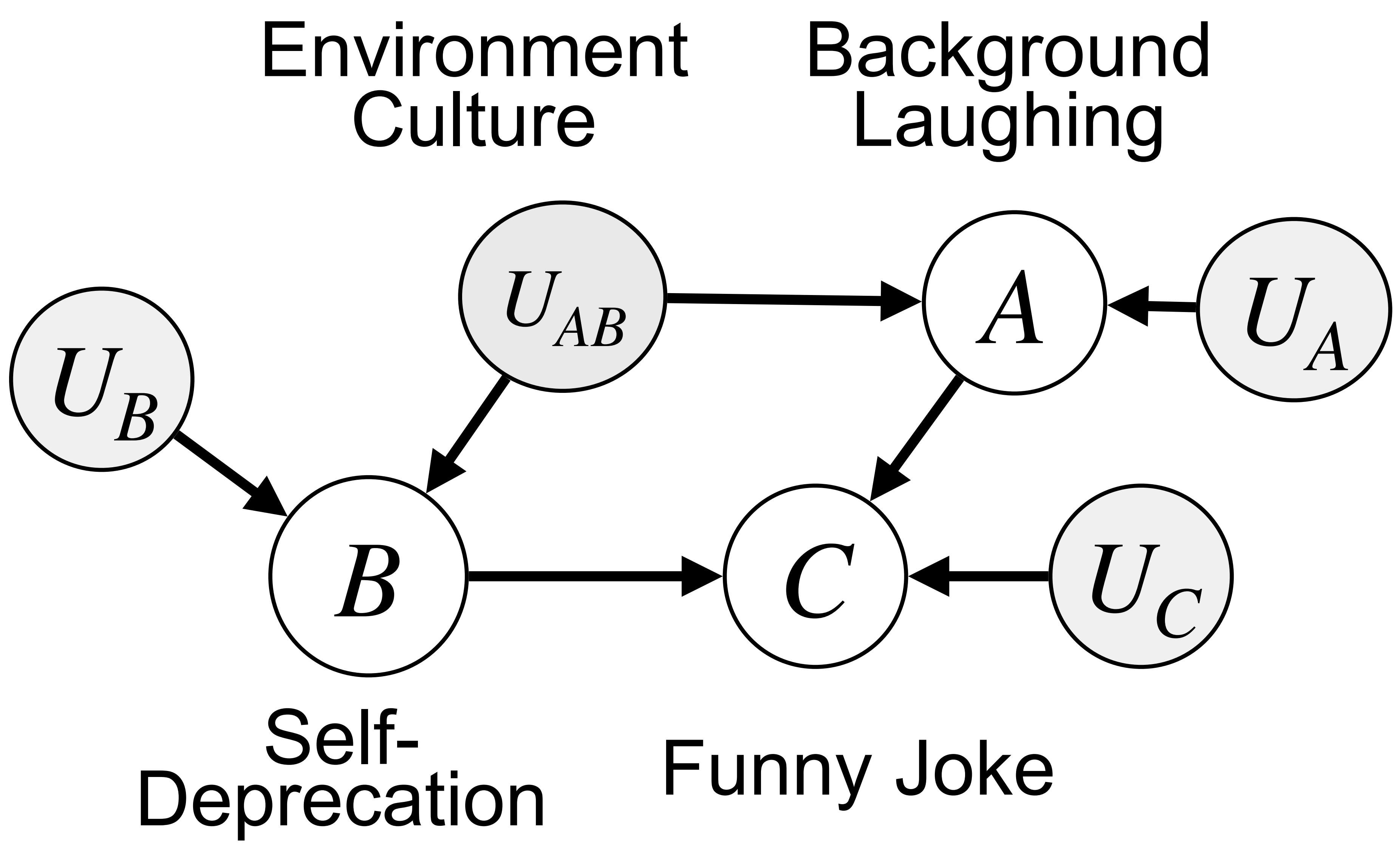
Structural Causal Model (SCM)

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

$$\mathcal{M} = \begin{cases} \mathbf{V} = \{A, B, C\} \\ \mathbf{U} = \{U_A, U_B, U_C, U_{AB}\} \\ \mathcal{F} = \begin{cases} A \leftarrow f_A(U_{AB}, U_A) \\ B \leftarrow f_B(U_{AB}, U_B) \\ C \leftarrow f_C(A, B, U_C) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

$\longrightarrow$

Induced Causal Diagram



An SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$ 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 $f_j \in \mathcal{F}$.

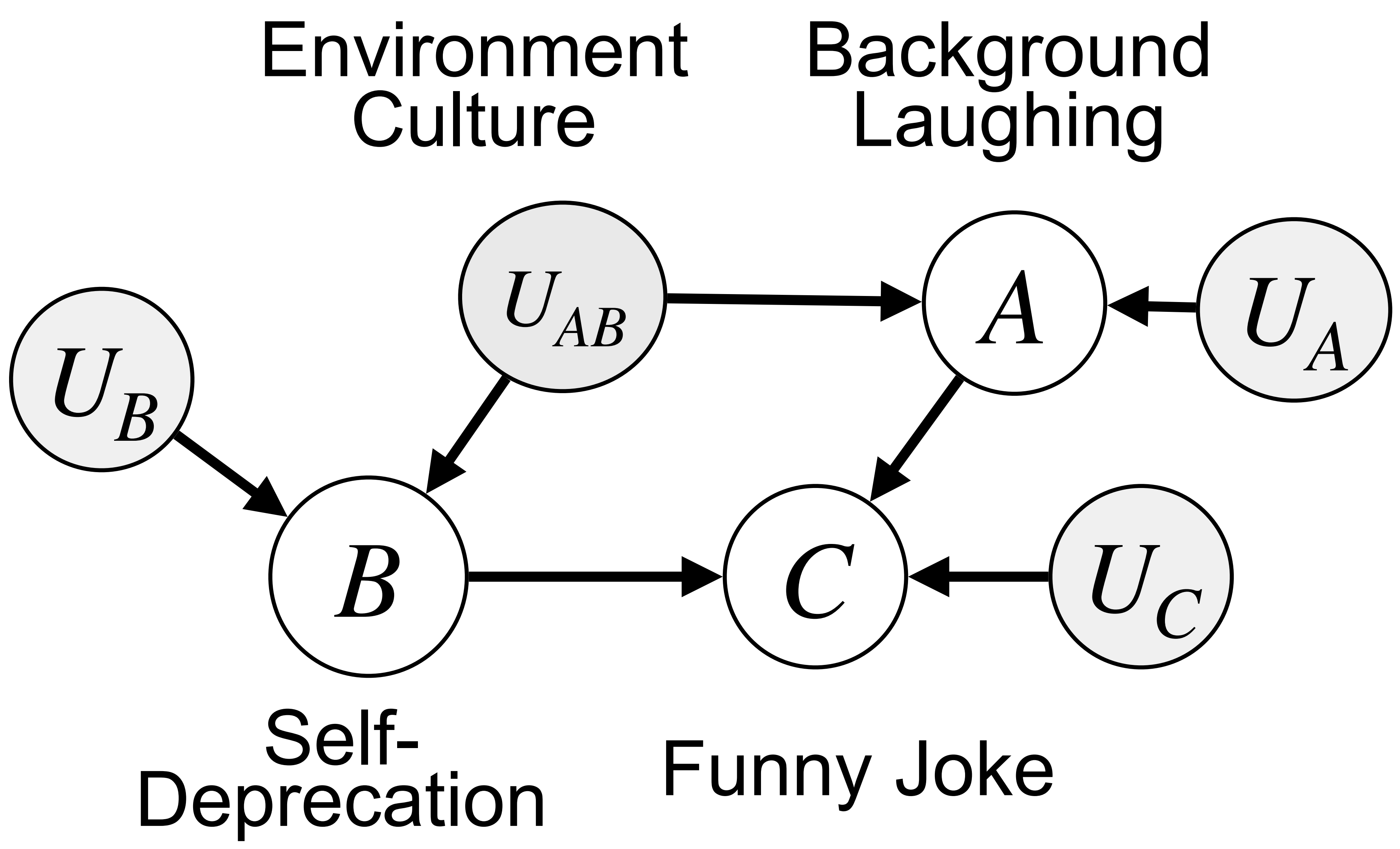
Causal Diagram: Encoder of Structural Knowledge

Structural Causal Model (SCM)

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

$$\mathcal{M} = \begin{cases} \mathbf{V} = \{A, B, C\} \\ \mathbf{U} = \{U_A, U_B, U_C, U_{AB}\} \\ \mathcal{F} = \begin{cases} A \leftarrow f_A(U_{AB}, U_A) \\ B \leftarrow f_B(U_{AB}, U_B) \\ C \leftarrow f_C(A, B, U_C) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Induced Causal Diagram



An SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$ 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 $f_j \in \mathcal{F}$.

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

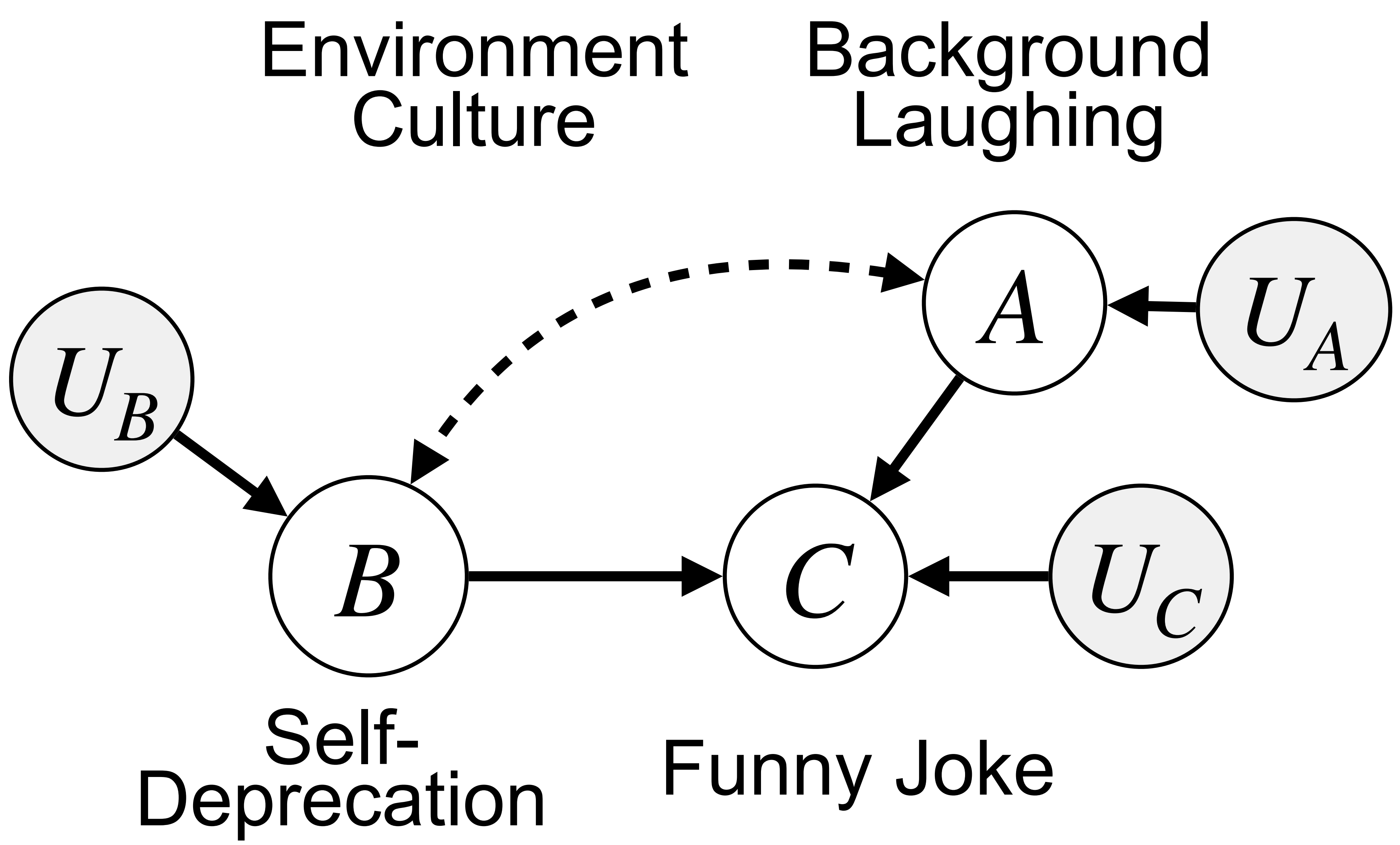
Causal Diagram: Encoder of Structural Knowledge

Structural Causal Model (SCM)

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

$$\mathscr{M} = \begin{cases} \mathbf{V} = \{A, B, C\} \\ \mathbf{U} = \{U_A, U_B, U_C, U_{AB}\} \\ \mathscr{F} = \begin{cases} A \leftarrow f_A(U_{AB}, U_A) \\ B \leftarrow f_B(U_{AB}, U_B) \\ C \leftarrow f_C(A, B, U_C) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Induced 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 $U \in \mathbf{U}$.

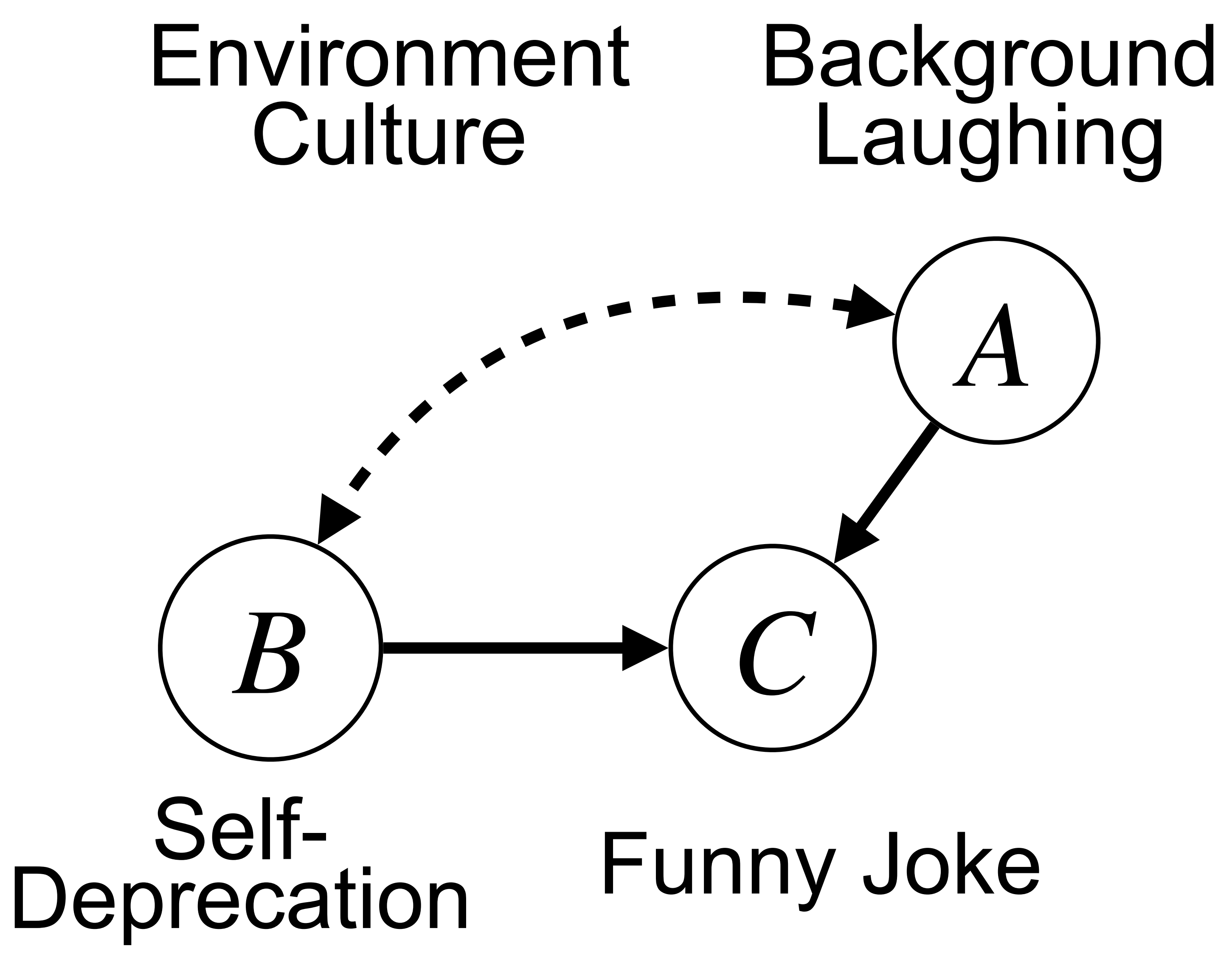
Causal Diagram: Encoder of <u>Structural</u> Knowledge

Structural Causal Model (SCM)

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

$$\mathcal{M} = \begin{cases} \mathbf{V} = \{A, B, C\} \\ \mathbf{U} = \{U_A, U_B, U_C, U_{AB}\} \\ \mathcal{F} = \begin{cases} A \leftarrow f_A(U_{AB}, U_A) \\ B \leftarrow f_B(U_{AB}, U_B) \\ C \leftarrow f_C(A, B, U_C) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Induced Causal Diagram



An SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$ 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 $f_j \in \mathcal{F}$.

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

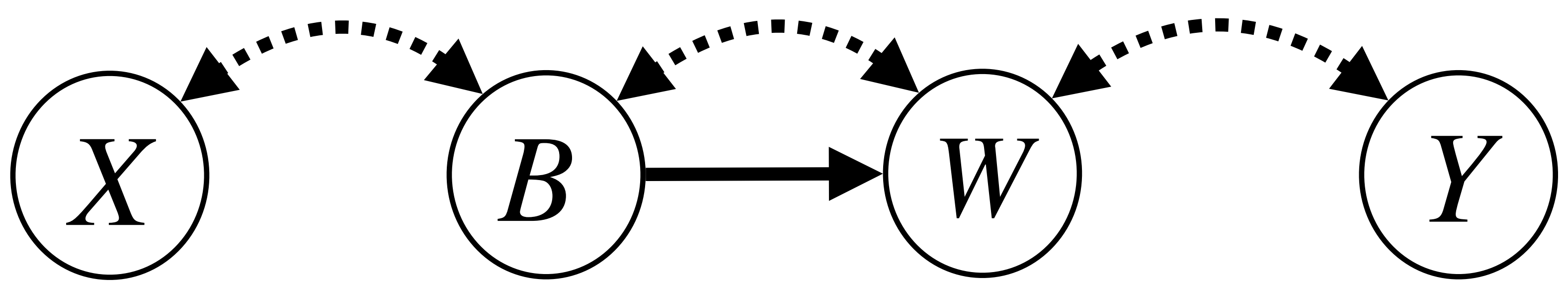
D-Separation and Implied Conditional Independencies

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 $\mathbf{Z}$.

Definition (d-separation): A path p in a causal diagram 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 $\mathbf{Y}$.



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

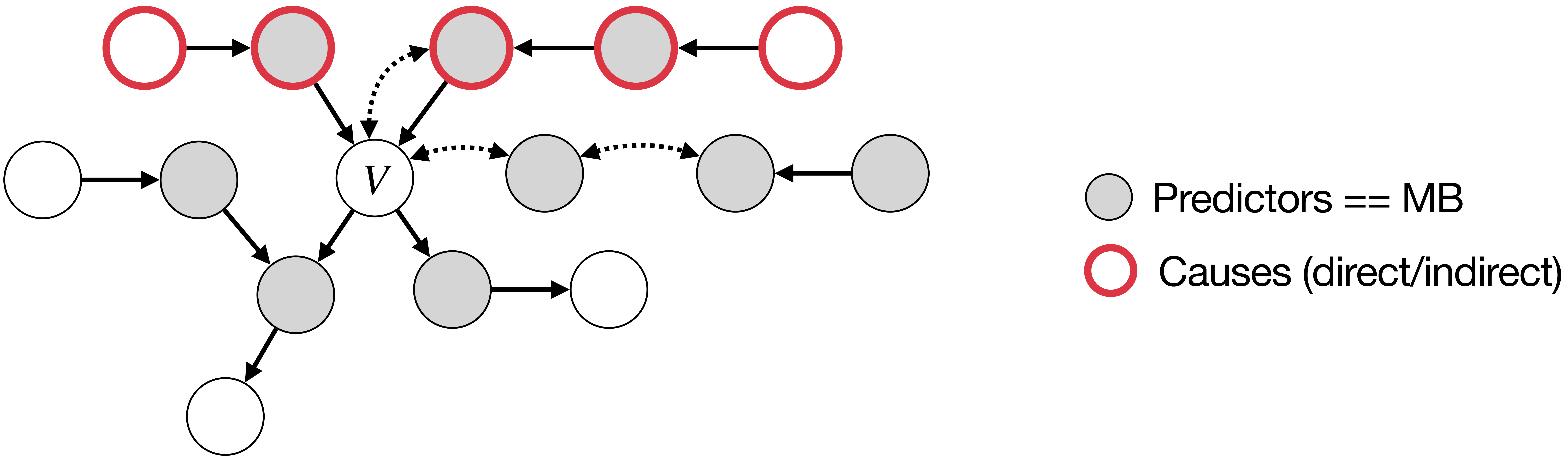
$\mathbf{Z}$: ✓ $\{\}$ ✓ $\{B\}$ ✗ $\{W\}$ ✗ $\{B, W\}$

We have that $(X \perp\!\!\!\perp Y)_G$, $(X \perp\!\!\!\perp Y \mid B)_G$, and $(X \not\perp\!\!\!\perp Y \mid W)_G$, but $(X \not\perp\!\!\!\perp Y \mid B, W)_G$

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

D-separations in G imply conditional independencies in P

Graphically Explaining Causes and Predictors

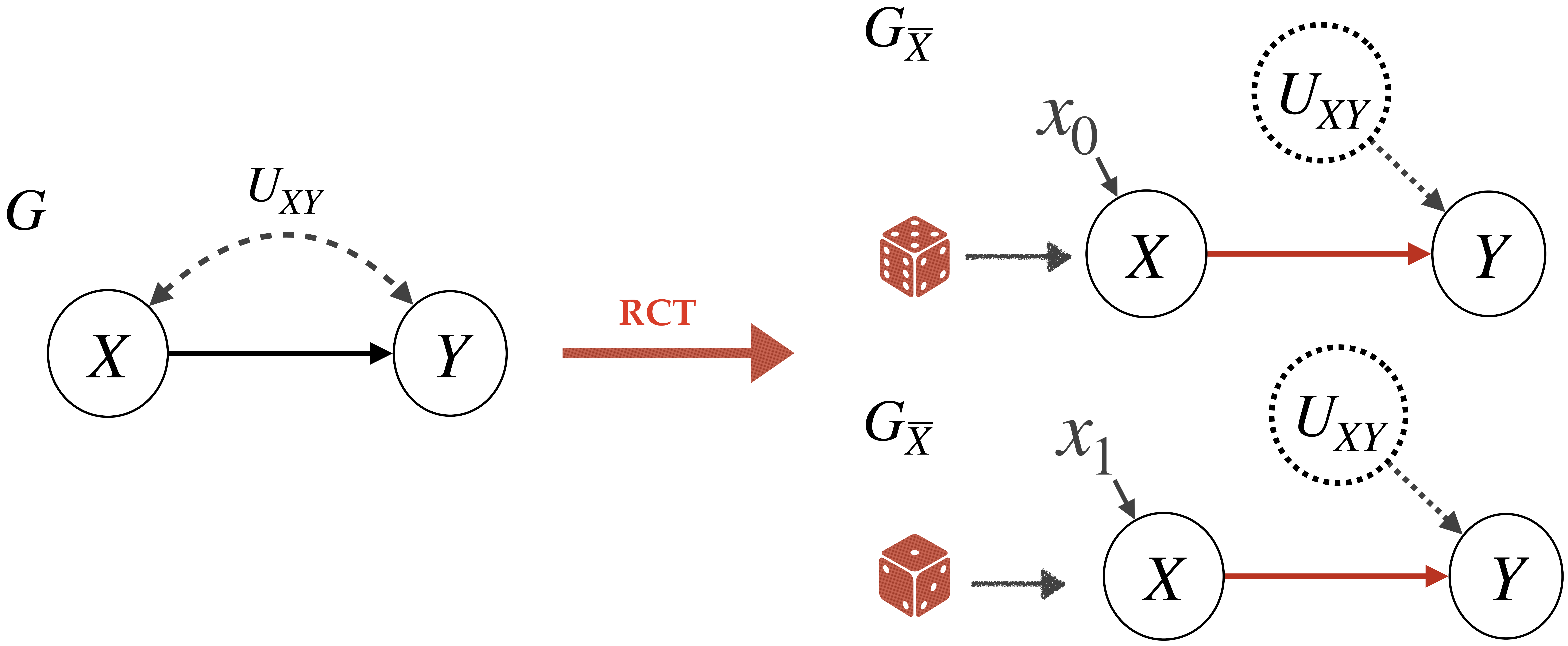


Markov Blanket (MB) of V: the bidirected connected component (district) of V (excluding V itself) and the parents of the district of V, i.e.:

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

Randomized Experiments

Randomized Experiments / Control Trials (e.g. RCT) allow the identification of causal effects by leveraging randomization of the treatment assignment.



Pearl’s Inferential Hierarchy

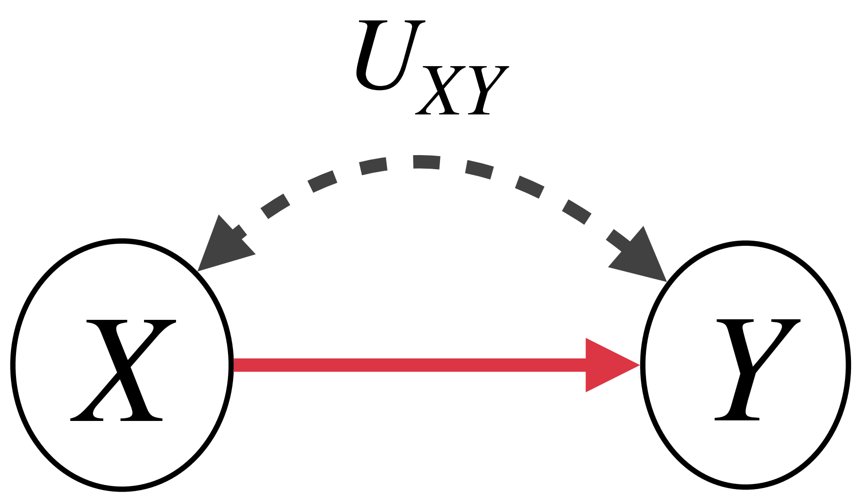
Associational vs Interventional vs Counterfactual

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

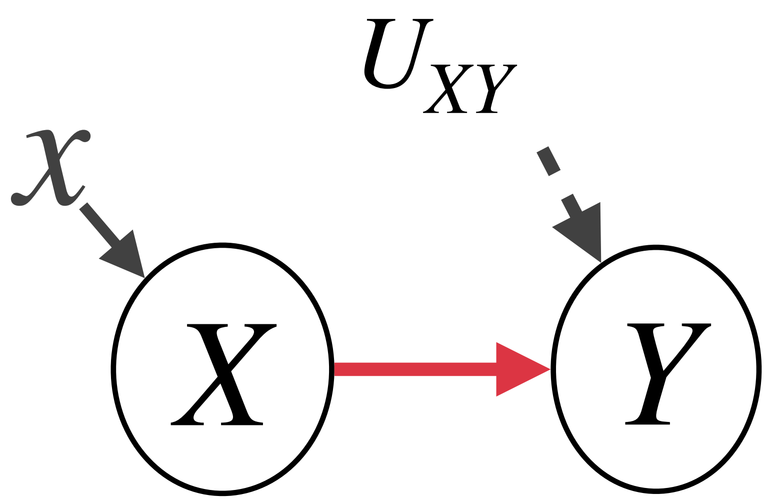
Observational Data

$do(X = x)$

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

Reality	Structural Knowledge	Data

Observational

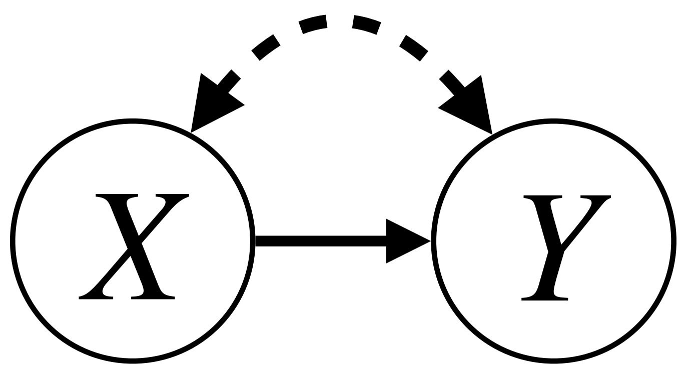
Structural Causal Model (SCM)

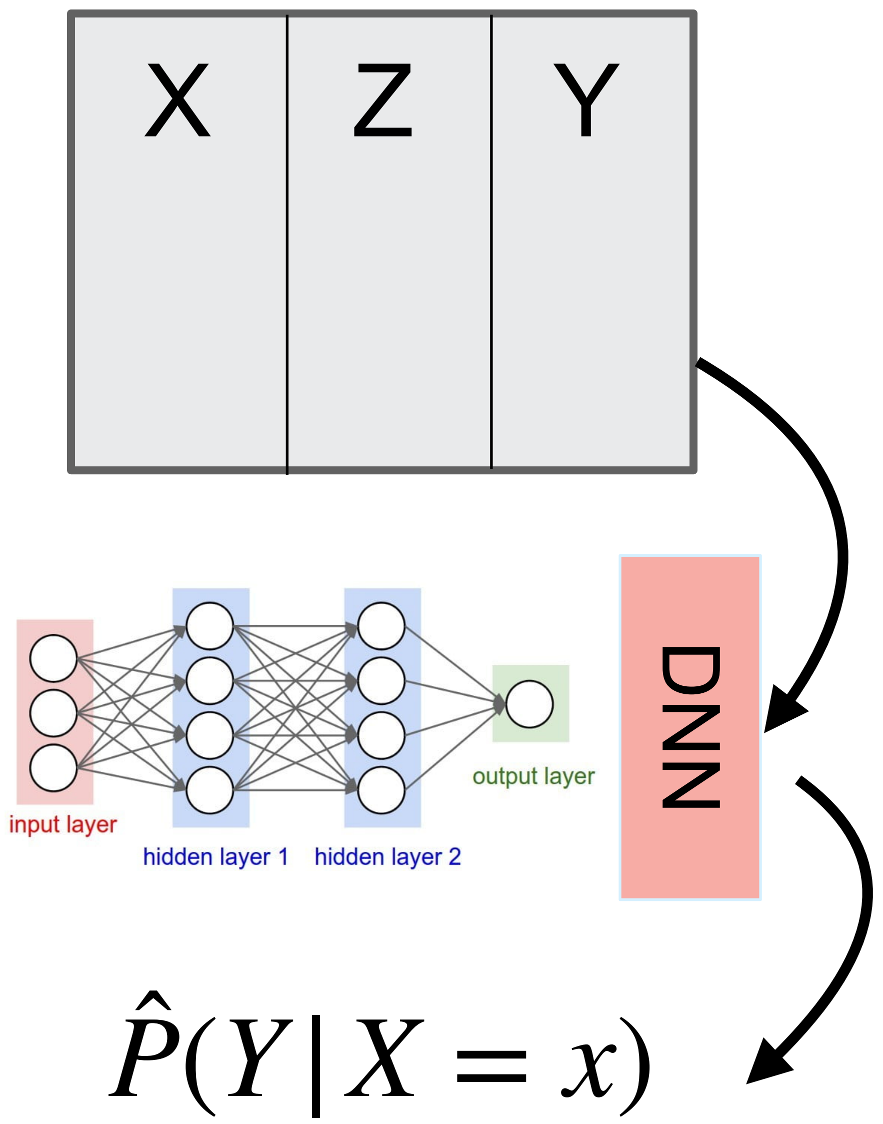
$\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$

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

Causal Diagram

G





$\hat{P}(Y \mid X = x)$

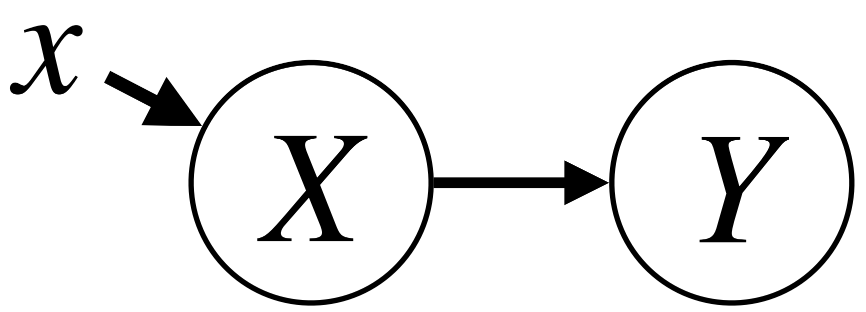
Seeing

Doing

Interventional

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

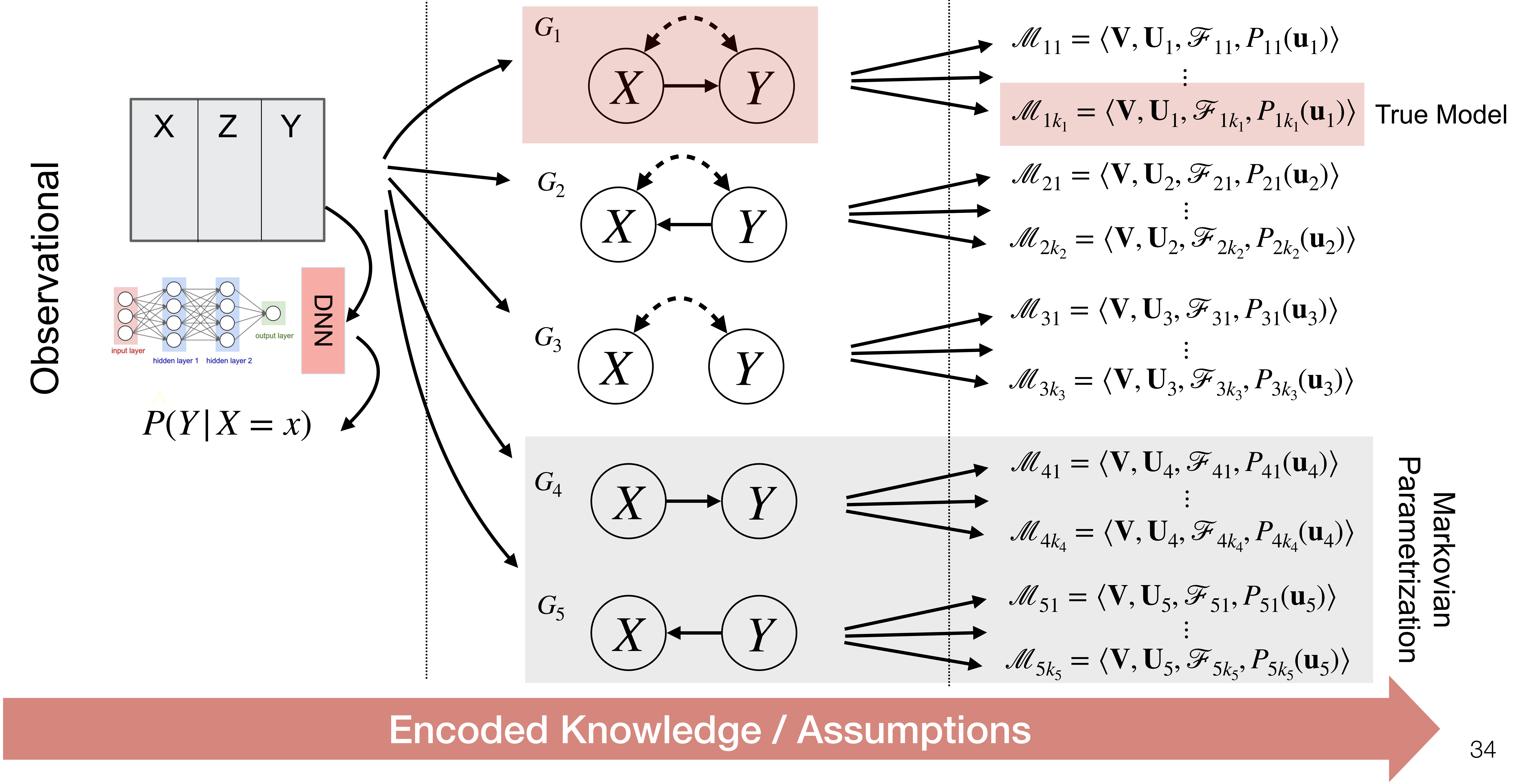
$G_{\overline{X}}$

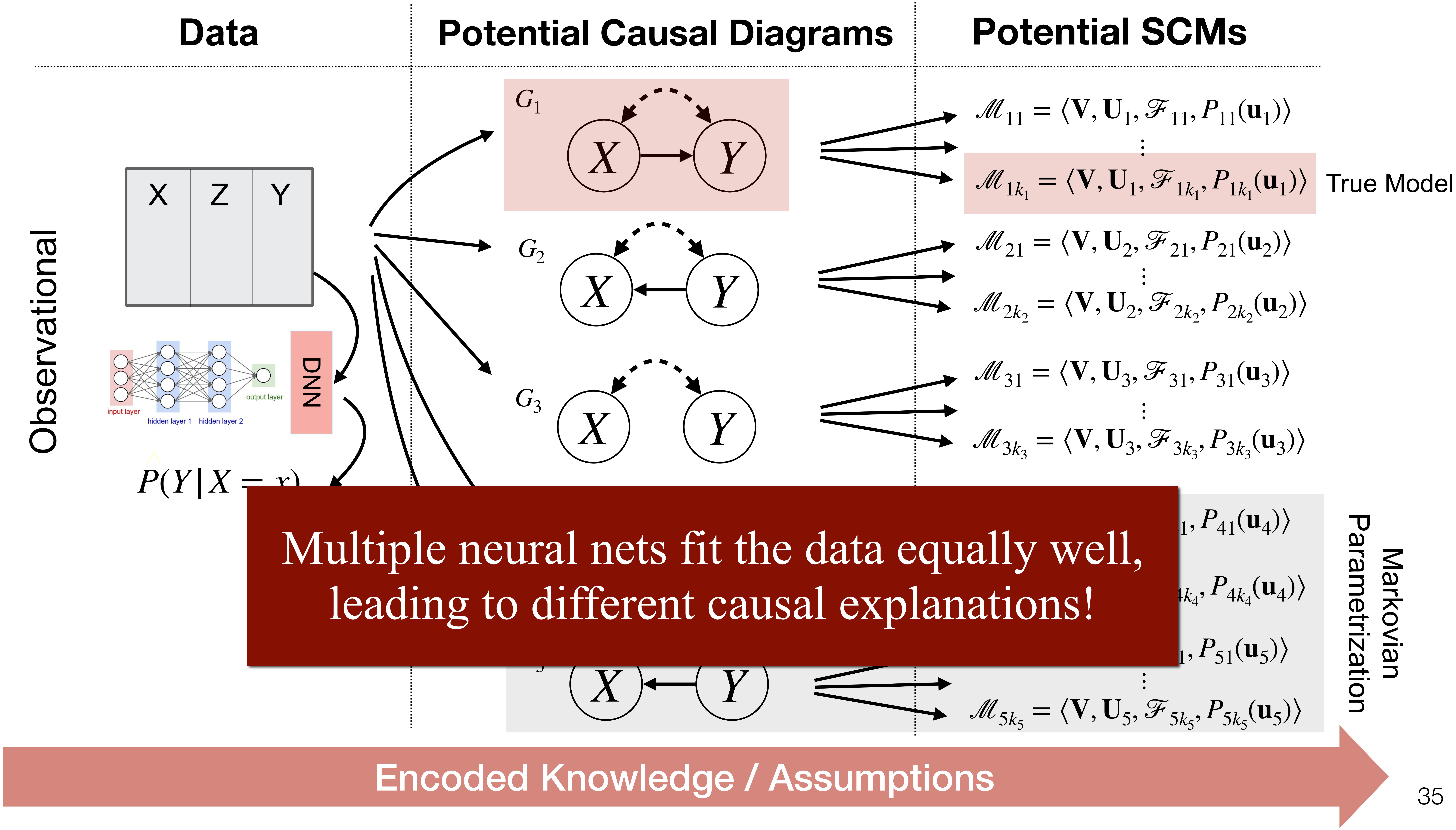


$\varnothing$

$\hat{P}(Y \mid do(X = x)) = ?$

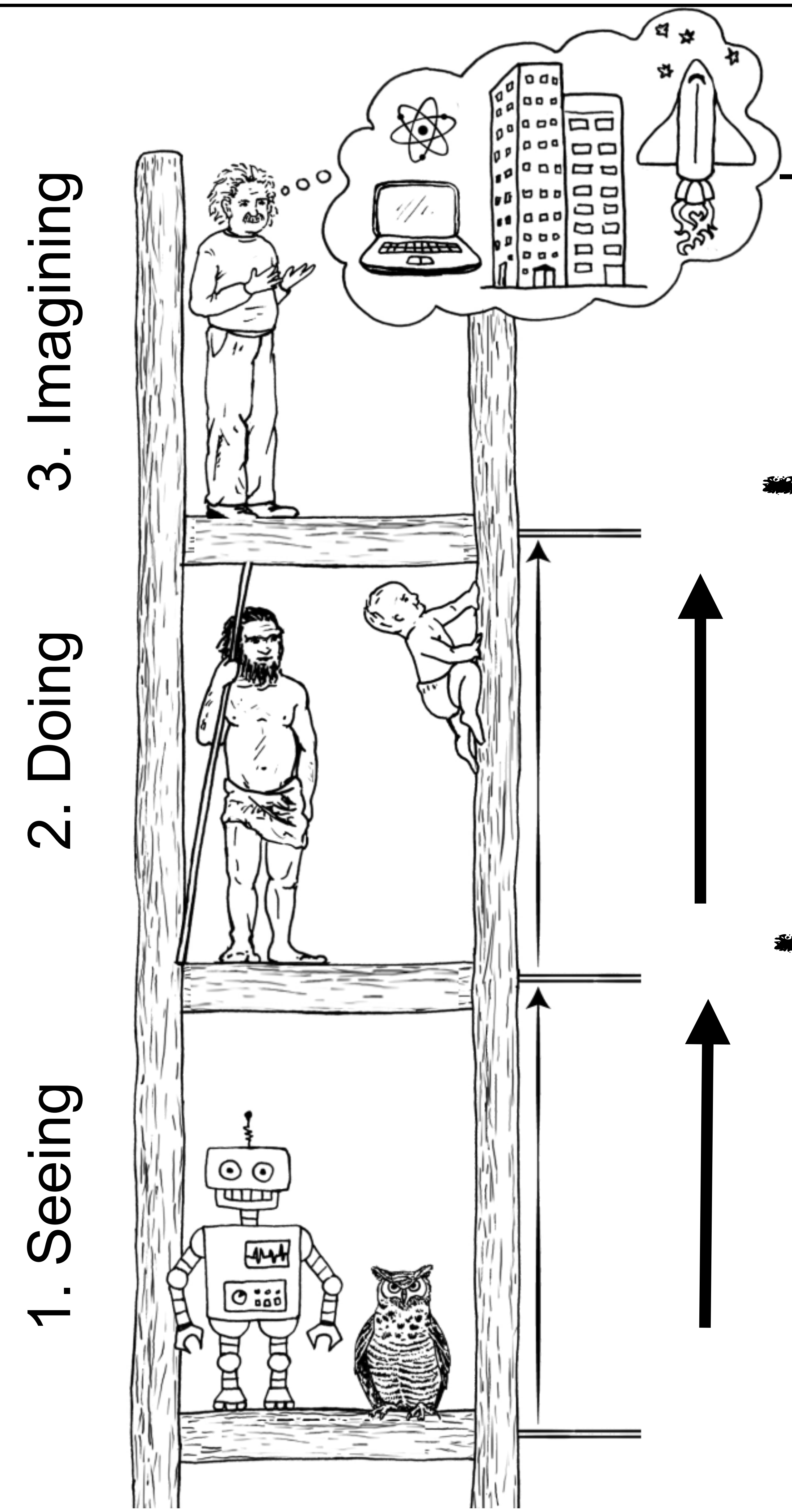
Data | Potential Causal Diagrams | Potential SCMs





Data
Potential Causal Diagrams
Potential SCMs
X
Z
Y
Observational
input layer
hidden layer 1
hidden layer 2
output layer
DNN
$\hat{P}(Y|X=x)$
G_1
G_2
G_3
$\mathscr{M}_{11} = \langle \mathbf{V}, \mathbf{U}_1, \mathscr{F}_{11}, P_{11}(\mathbf{u}_1) \rangle$
$\mathscr{M}_{1k_1} = \langle \mathbf{V}, \mathbf{U}_1, \mathscr{F}_{1k_1}, P_{1k_1}(\mathbf{u}_1) \rangle$
True Model
$\mathscr{M}_{21} = \langle \mathbf{V}, \mathbf{U}_2, \mathscr{F}_{21}, P_{21}(\mathbf{u}_2) \rangle$
$\mathscr{M}_{2k_2} = \langle \mathbf{V}, \mathbf{U}_2, \mathscr{F}_{2k_2}, P_{2k_2}(\mathbf{u}_2) \rangle$
$\mathscr{M}_{31} = \langle \mathbf{V}, \mathbf{U}_3, \mathscr{F}_{31}, P_{31}(\mathbf{u}_3) \rangle$
$\mathscr{M}_{3k_3} = \langle \mathbf{V}, \mathbf{U}_3, \mathscr{F}_{3k_3}, P_{3k_3}(\mathbf{u}_3) \rangle$
$\mathscr{M}_{5k_5} = \langle \mathbf{V}, \mathbf{U}_5, \mathscr{F}_{5k_5}, P_{5k_5}(\mathbf{u}_5) \rangle$
Markovian Parametrization
Multiple neural nets fit the data equally well, leading to different causal explanations!
Encoded Knowledge / Assumptions

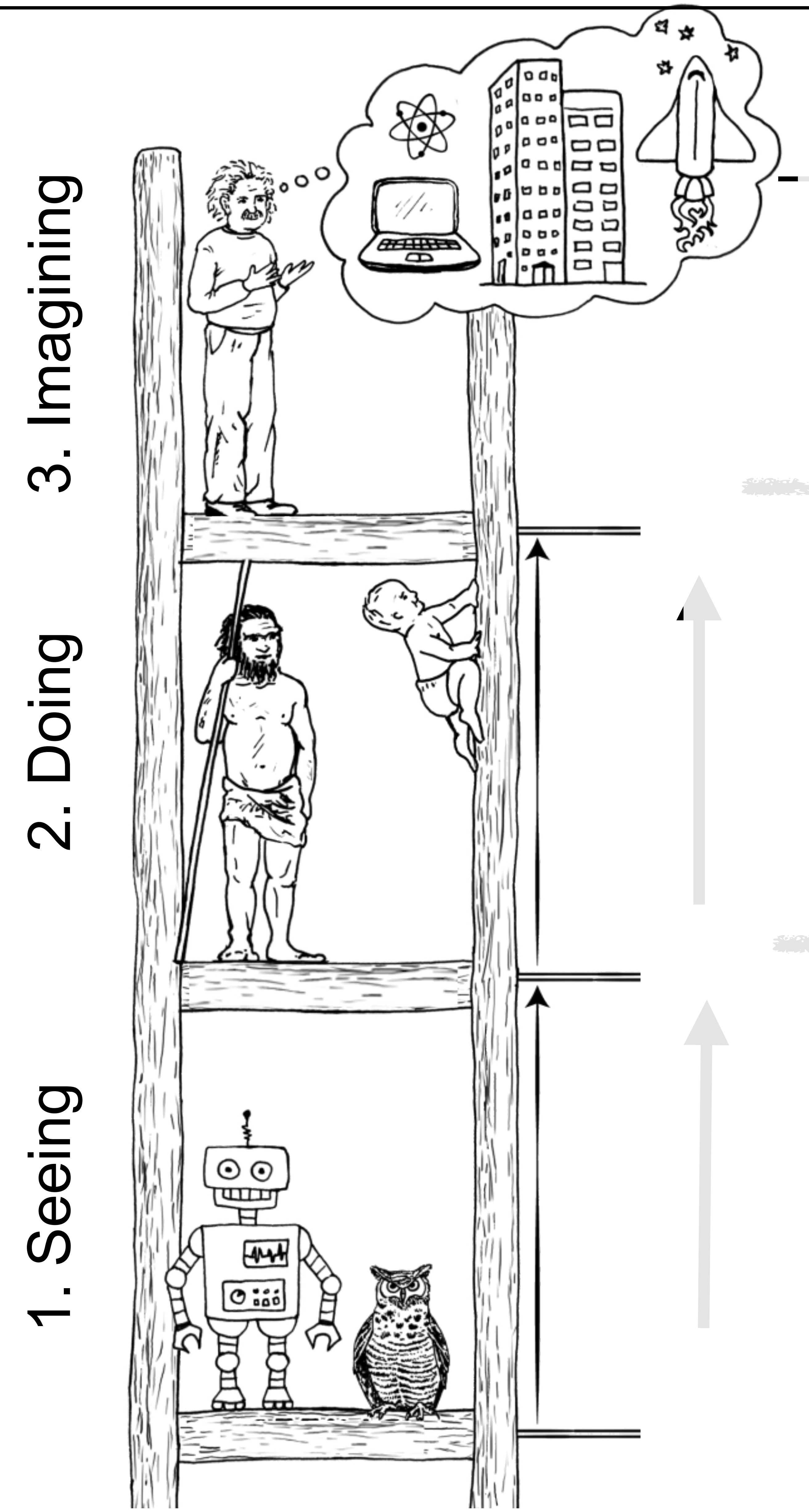
Ladder of Causation



Layer	Task / Language	Typical Question	Examples
Counterfactual $P(y_x \mid x', y')$	Structural Causal Model	What if I had acted differently?	Was it the aspirin that stopped my headache?
Interventional $P(y \mid do(x), c)$	ML- Reinforcement (Causal Bayes Net)	What if I do X? What would Y be if I intervene on X?	Will my headache be cured if I take aspirin?
Associational $P(y \mid x)$	ML- (Un)Supervised (Decision trees, Deep nets, …)	What if I see? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?

* 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



Cross-layer inferences:

Doing

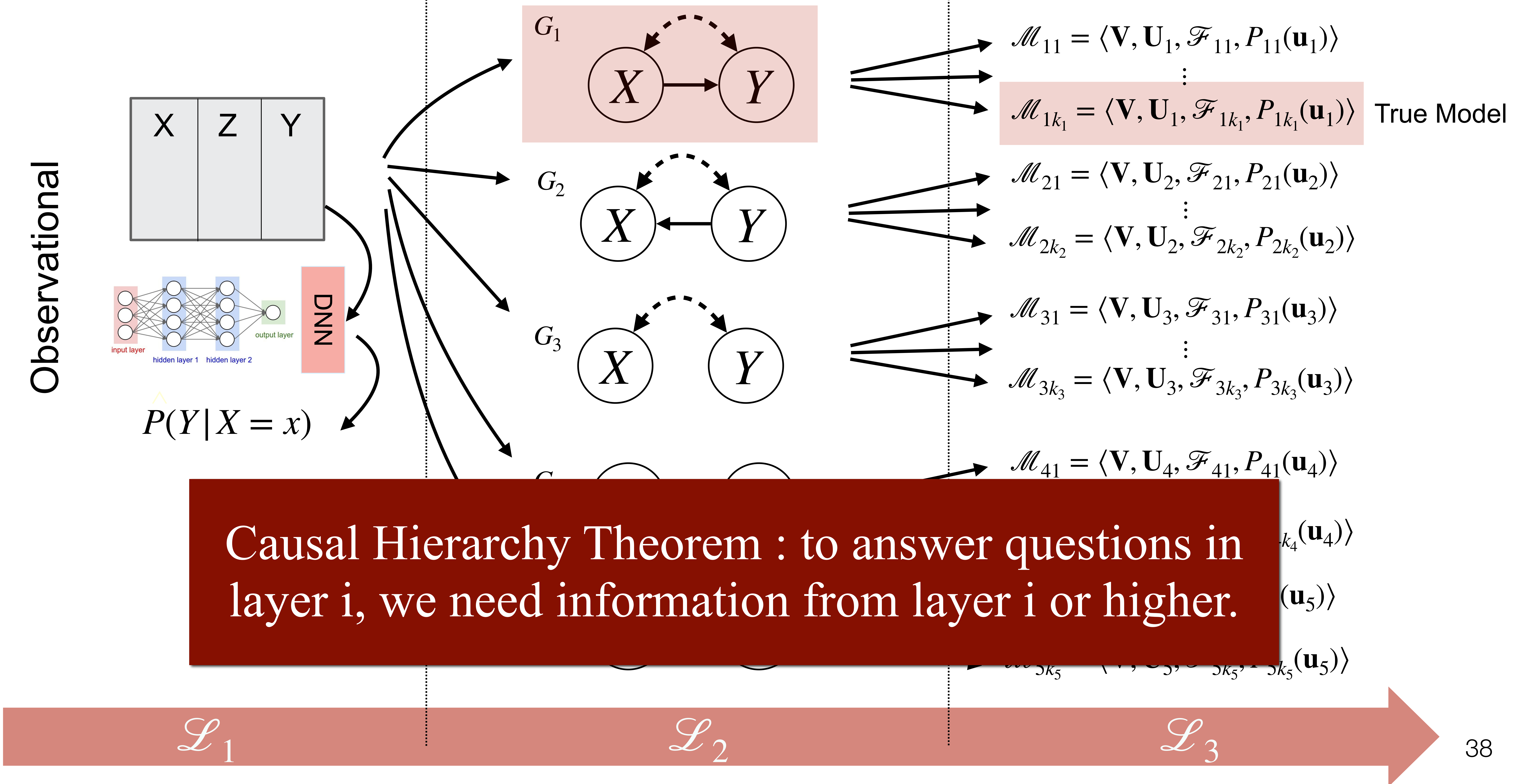
most of the inferences are about causal effects (policies, treatments, decisions)

Seeing

most of the available data is observational, passively collected

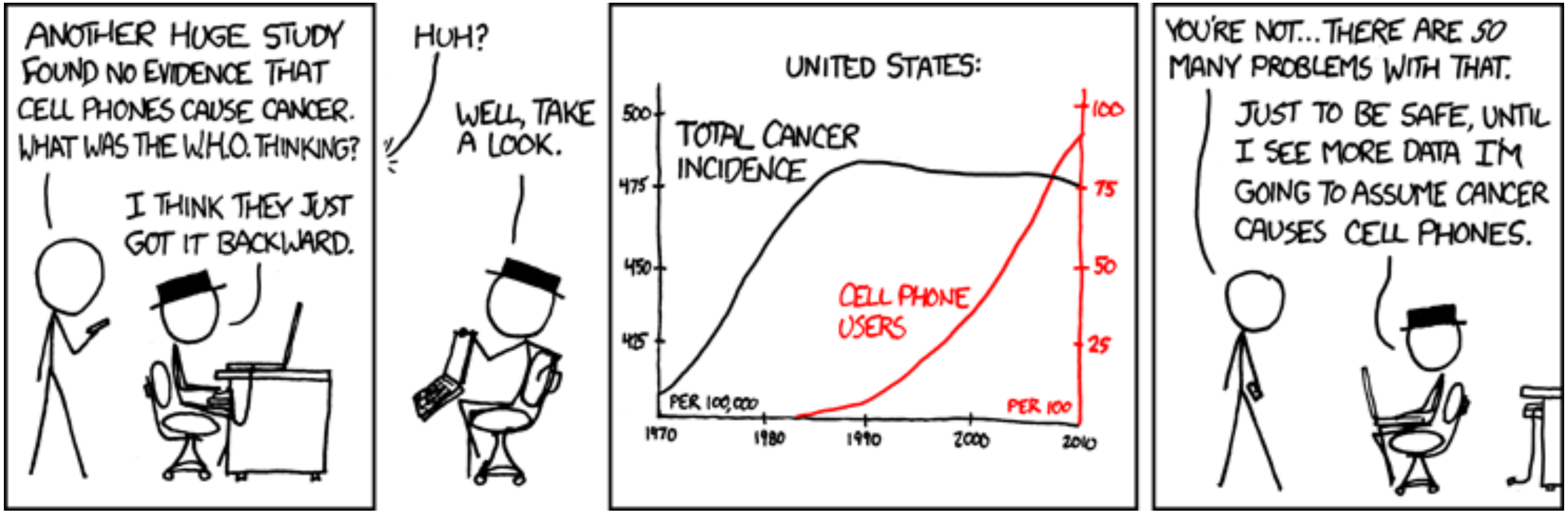
* Book of Why & On Pearl's Hierarchy and the Foundations of Causal Inference, E. Bareinboim, J. Correa, D. Ibeling, T. Icard, in press. https://causalai.net/r60.pdf

Data	Potential Causal Diagrams	Potential SCMs

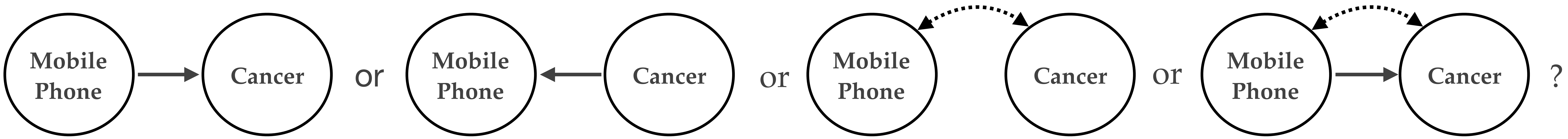


Causal Hierarchy Theorem : to answer questions in layer i, we need information from layer i or higher.

Association vs Causation



Will we be able to decide the true relationship just by “seeing” more data?



Causal Effect Identification

Graphical Criteria, Do-Calculus, and ID-Algorithm

Causal Effect

The **causal effect** of a (set of) treatment variable(s) $\mathbf{X}$ on a (set of) outcome variable(s) $\mathbf{Y}$ is a quantity derived from $P(\mathbf{Y} \mid do(\mathbf{X}))$ that tells us how much $\mathbf{Y}$ changes due to an intervention $do(\mathbf{X} = \mathbf{x})$.

Examples:

- *Average Treatment Effect* (ATE) for discrete treatments:

$$\mathbb{E}[\mathbf{Y} \mid do(\mathbf{X} = \mathbf{x}')] - \mathbb{E}[\mathbf{Y} \mid do(\mathbf{X} = \mathbf{x})],$$

where $\mathbb{E}[\mathbf{Y} \mid do(\mathbf{X} = \mathbf{x})] = \sum_{\mathbf{y} \in \Omega_{\mathbf{Y}}} \mathbf{y} P(\mathbf{y} \mid do(\mathbf{x}))$

defined for two treatment levels $\mathbf{x}'$ and $\mathbf{x}$ of $\mathbf{X}$.

- *Average Treatment Effect* (ATE) for continuous treatments,

$$\frac{\partial \mathbb{E}[Y_i \mid do(X_j = x_j)]}{\partial x_j}, \text{ for all } Y_i \in \mathbf{Y}, \text{ and } X_j \in \mathbf{X}.$$

The derivative shows the rate of change of $\mathbf{Y}$ w.r.t. $do(\mathbf{X} = \mathbf{x})$

Jacobian of $\mathbb{E}[\mathbf{Y} \mid do(\mathbf{X} = \mathbf{x})]$, where

$$\mathbb{E}[\mathbf{Y} \mid do(\mathbf{X} = \mathbf{x})] = \int_{\Omega_{\mathbf{y}}} \mathbf{y} P(\mathbf{y} \mid do(\mathbf{x})) d\mathbf{y},$$

and $\Omega_{\mathbf{Y}}$ is the space of all possible values that $\mathbf{Y}$ might take on

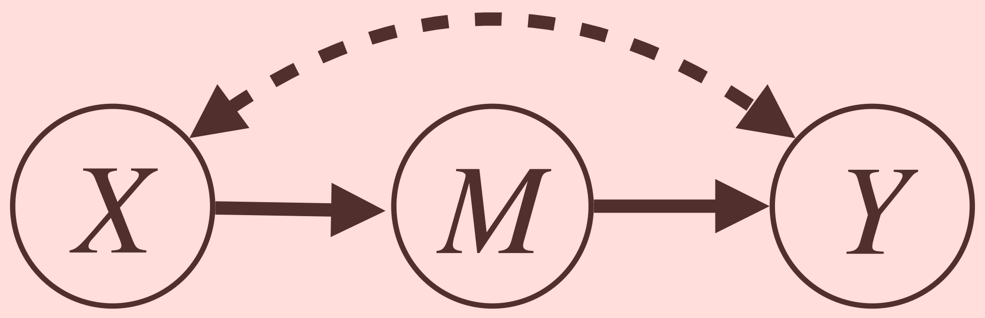
Classical Causal Effect Identification

Structural knowledge available

1 **Query**

$P(y\,|\,do(x))$

2 **Causal Contraints**



3 **Probability Distributions**

$P(x, m, y)$ **Observational Distribution**

Inference Engine

Solution

yes / no

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

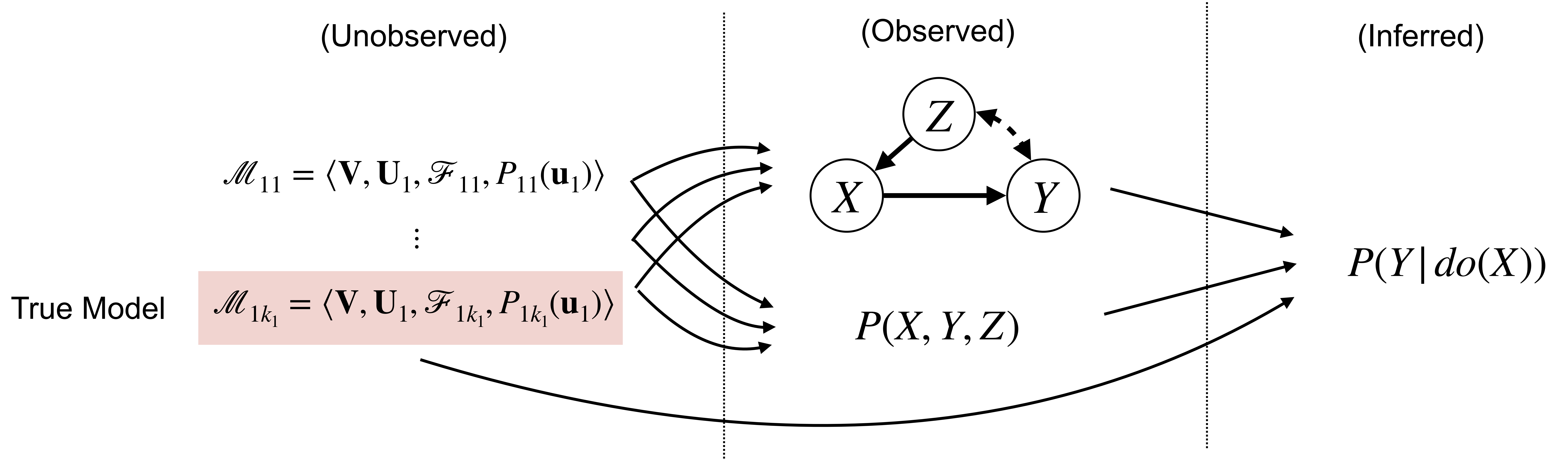
Interventional Distribution ← Available Distributions

👍

- 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.

The Effect Identification Problem

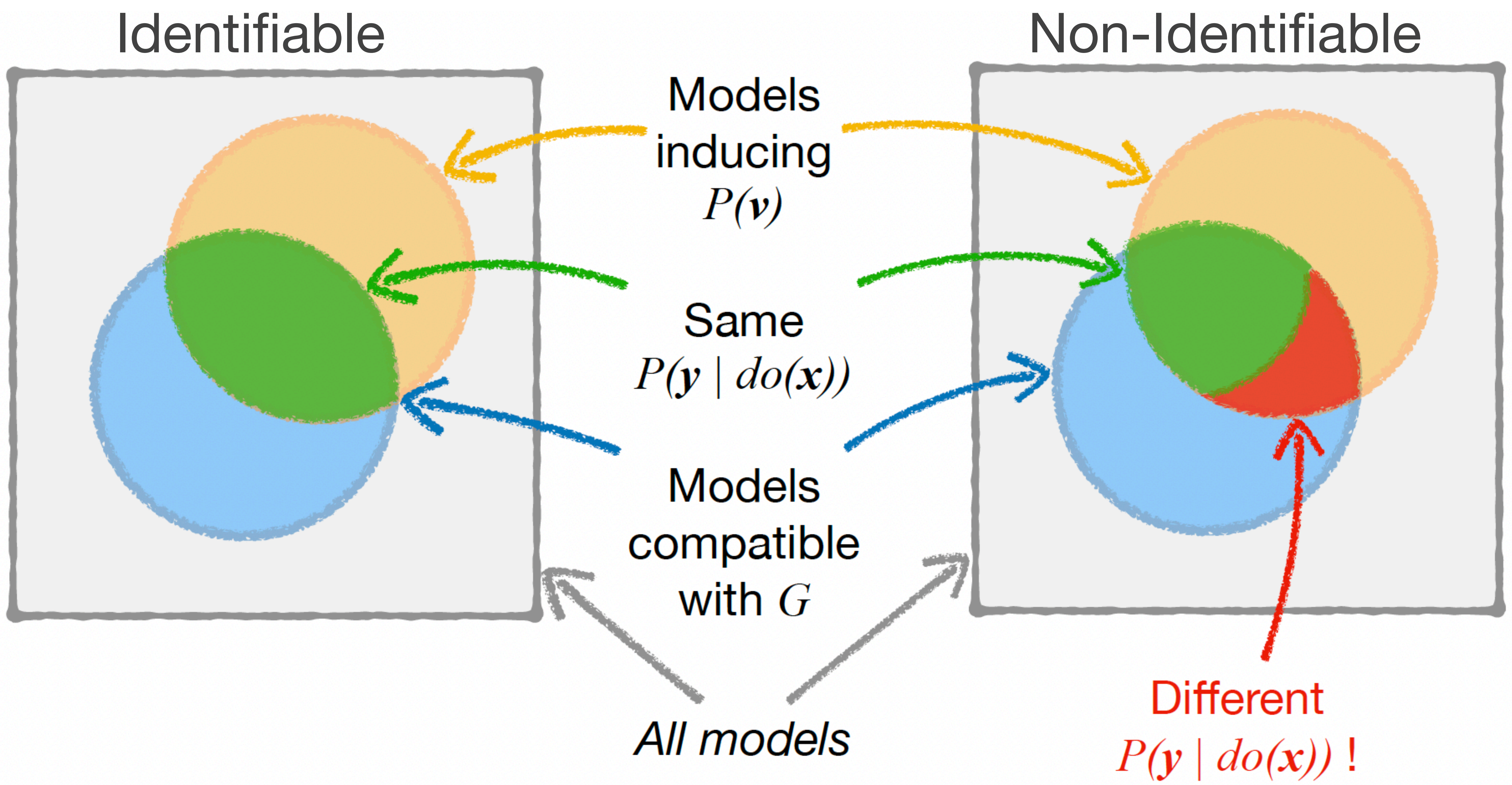
Causal Effect Identifiability: The causal effect of a (set of) treatment variable(s) $\mathbf{X}$ on a (set of) outcome variable(s) $\mathbf{Y}$ is said to be identifiable from a causal diagram G and the probability of the observed variables $P(\mathbf{V})$ if the interventional distribution $P(\mathbf{Y} \mid do(\mathbf{X}))$ is *uniquely computable,* i.e., if for every pair of SCMs $\mathscr{M}_1$ and $\mathscr{M}_2$ that induce G and $P^{\mathscr{M}_1}(\mathbf{V}) = P^{\mathscr{M}_2}(\mathbf{V}) = P(\mathbf{V}) > 0$, $P^{\mathscr{M}_1}(\mathbf{Y} \mid do(\mathbf{X})) = P^{\mathscr{M}_2}(\mathbf{Y} \mid do(\mathbf{X})) = P(\mathbf{Y} \mid do(\mathbf{X}))$.



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 $P(\mathbf{y} \mid do(\mathbf{x}))$.

The Effect Identification Problem

Causal Effect Identifiability: The causal effect of a (set of) treatment variable(s) $\mathbf{X}$ on a (set of) outcome variable(s) $\mathbf{Y}$ is said to be identifiable from a causal diagram G and the probability of the observed variables $P(\mathbf{V})$ if the interventional distribution $P(\mathbf{Y} \mid do(\mathbf{X}))$ is *uniquely computable,* i.e., if for every pair of SCMs $\mathcal{M}_1$ and $\mathcal{M}_2$ that induce G and $P^{\mathcal{M}_1}(\mathbf{V}) = P^{\mathcal{M}_2}(\mathbf{V}) = P(\mathbf{V}) > 0$, $P^{\mathcal{M}_1}(\mathbf{Y} \mid do(\mathbf{X})) = P^{\mathcal{M}_2}(\mathbf{Y} \mid do(\mathbf{X})) = P(\mathbf{Y} \mid do(\mathbf{X}))$.



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

Tools for Causal Identification

1. Truncated Factorization / G-computation formula

Markovian Models

2. Graphical criteria
   1. Parent adjustment
   2. Backdoor Adjustment
   3. Front-door Adjustment

A few interesting (albeit still constrained) scenarios

3. Do-Calculus (a.k.a Causal Calculus)
4. Identify Algorithm (a.k.a. ID algorithm)

General Semi-Markovian Scenarios

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press, New York. http://dx.doi.org/10.1017/CBO9780511803161

Jin Tian. Studies in causal reasoning and learning. PhD thesis, University of California, Los Angeles, 2002.

Identification in Markovian Models

Truncated Factorization — Markovian: Let G be a causal diagram for the collection $\mathbf{P}_*$ of all interventional distributions $P_{\mathbf{x}}(\mathbf{V})$, for any $\mathbf{X} \subseteq \mathbf{V}$. It follows that $P_{\mathbf{x}}(\mathbf{V})$ factorizes as:

$$P_{\mathbf{x}}(\mathbf{v}) \doteq P(\mathbf{v} \mid do(\mathbf{x})) = \prod_{V_i \in \mathbf{V} \setminus \mathbf{X}} P_{\mathbf{x}}(v_i \mid pa_i) \Big|_{\mathbf{X}=\mathbf{x}}$$

Follows from $P_{\mathbf{x}}(\mathbf{v}) \doteq P(\mathbf{v} \mid do(\mathbf{x}))$ being *Markov* relative to $G_{\overline{\mathbf{X}}}$

$$= \prod_{V_i \in \mathbf{V} \setminus \mathbf{X}} P(v_i \mid pa_i) \Big|_{\mathbf{X}=\mathbf{x}}$$

Markovian SCMs have the modularity property, i.e., $P_{\mathbf{x}}(v_i \mid pa_i) = P(v_i \mid pa_i)$

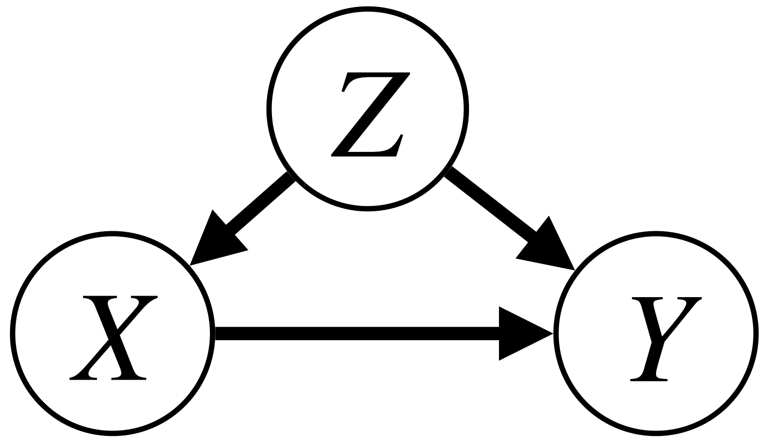
Causal Effect of $\mathbf{X}$ on $\mathbf{Y}$: $$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{V} \setminus (\mathbf{Y} \cup \mathbf{X})} \prod_{V_i \in \mathbf{V} \setminus \mathbf{X}} P(v_i \mid pa_i) \Big|_{\mathbf{X}=\mathbf{x}}$$

- In Markovian Models, the joint interventional distribution (and hence any causal effect) is always identifiable.
- This factorization is a.k.a "manipulation theorem" (Spirtes et al. 1993) or G-computation (Robins 1986, p. 1423).

Example: Identifiable Effect

Causal Effect of $\mathbf{X}$ on $\mathbf{Y}$: $$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{V} \setminus (\mathbf{Y} \cup \mathbf{X})} \prod_{V_i \in \mathbf{V} \setminus \mathbf{X}} P_{\mathbf{x}}(v_i \mid pa_i) \Big|_{\mathbf{X}=\mathbf{x}}$$

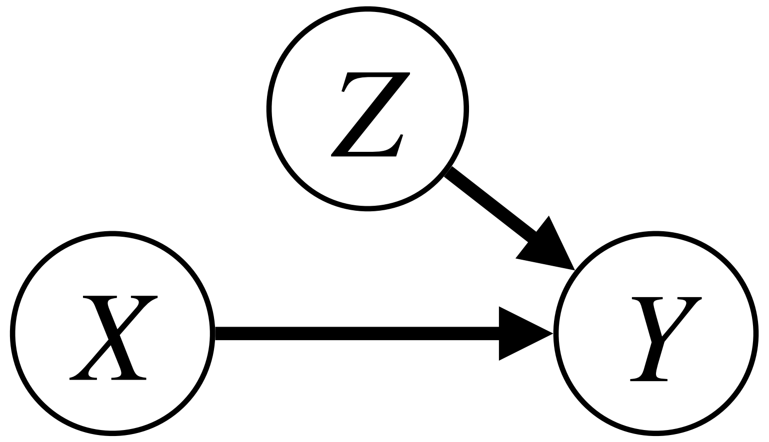
G



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

$do(X = x)$

$G_{\overline{X}}$



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

$$\Longrightarrow P(y \mid do(x)) = \sum_z P(z)P(y \mid x, z)$$

Identification in Semi-Markovian Models

Adjustment over parents:

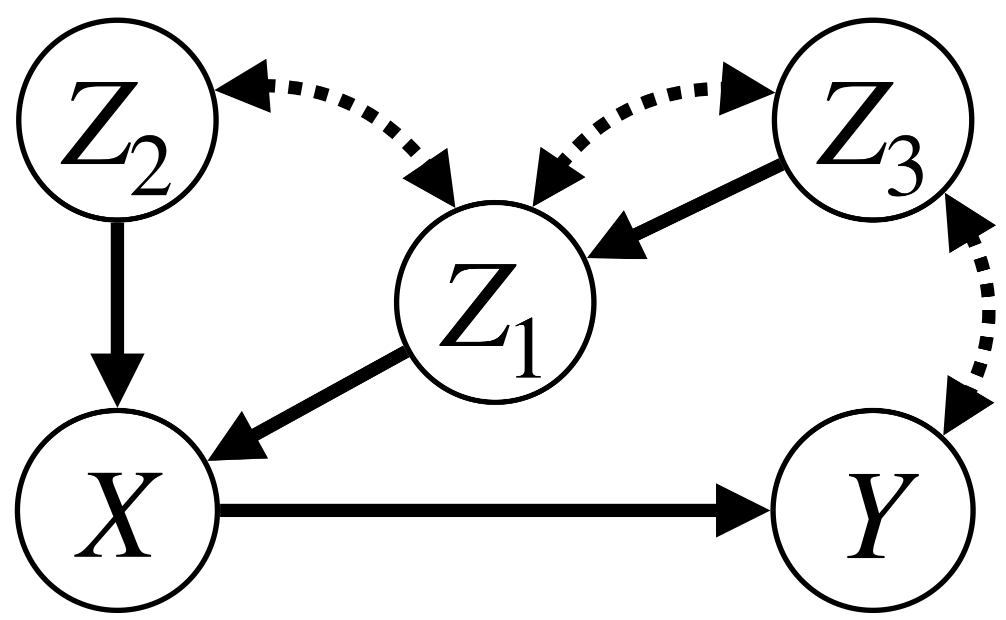
Let G be a causal graph with **no unmeasured parents**.

Then, the effect of $\mathbf{X}$ on $\mathbf{Y}$ is given by:

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

Proof follows from the truncated factorization for Markovian models!

$$\mathbf{X} = \{X\}$$
$$\mathbf{Y} = \{Y\}$$
$$\mathbf{Pa_X} = \{Z_1, Z_2\}$$



$$Pa_x = \{Z_1, Z_2\}$$

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

Identification in Semi-Markovian Models

Adjustment over parents:

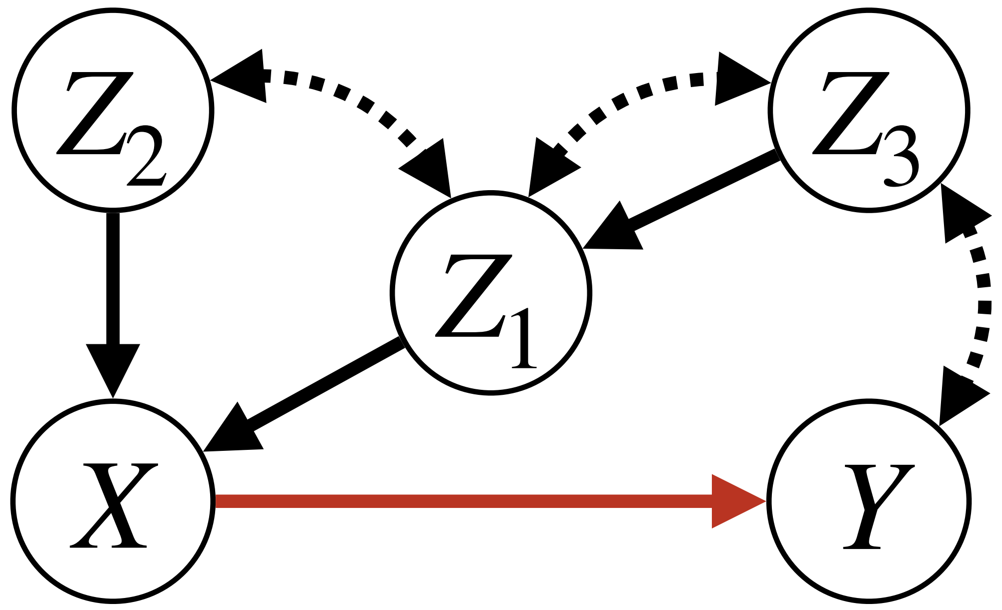
Let G be a causal graph with **no unmeasured parents**.

Then, the effect of $\mathbf{X}$ on $\mathbf{Y}$ is given by:

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

Proof follows from the truncated factorization for Markovian models!

$$\mathbf{X} = \{X\}$$
$$\mathbf{Y} = \{Y\}$$
$$\mathbf{Pa_X} = \{Z_1, Z_2\}$$



$$Pa_x = \{Z_1, Z_2\}$$

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

After conditioning on the parents, the association between X and Y is only due to the direct path.

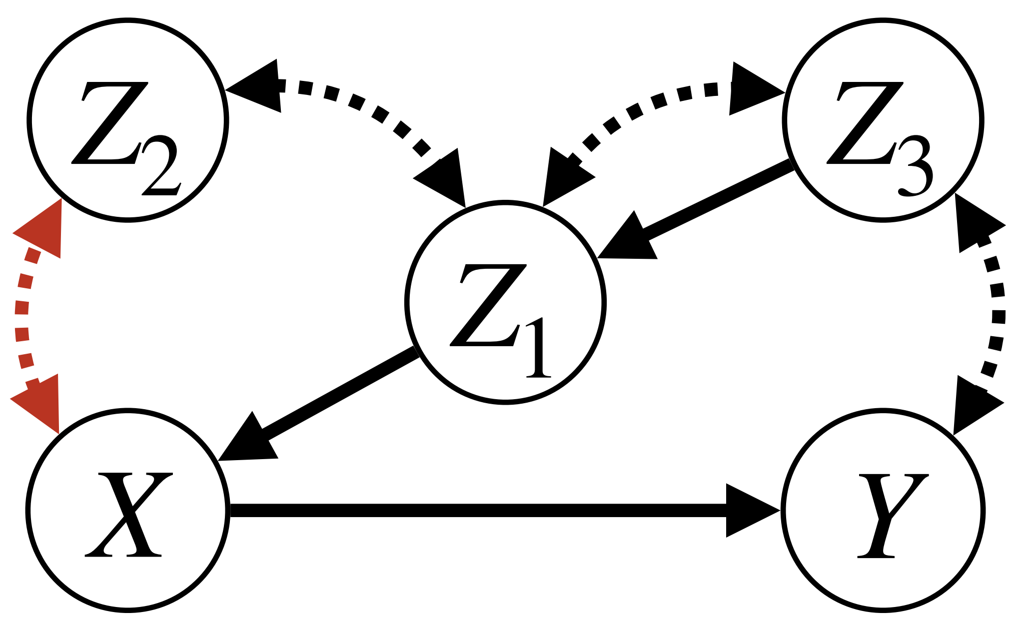
Identification in Semi-Markovian Models

Adjustment over parents:

Let G be a causal graph with **no unmeasured parents**.

Then, the effect of $\mathbf{X}$ on $\mathbf{Y}$ is given by:

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



$$Pa_x = \{Z_2\}$$
$$U_x = \{U_{X,Z2}\}$$

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

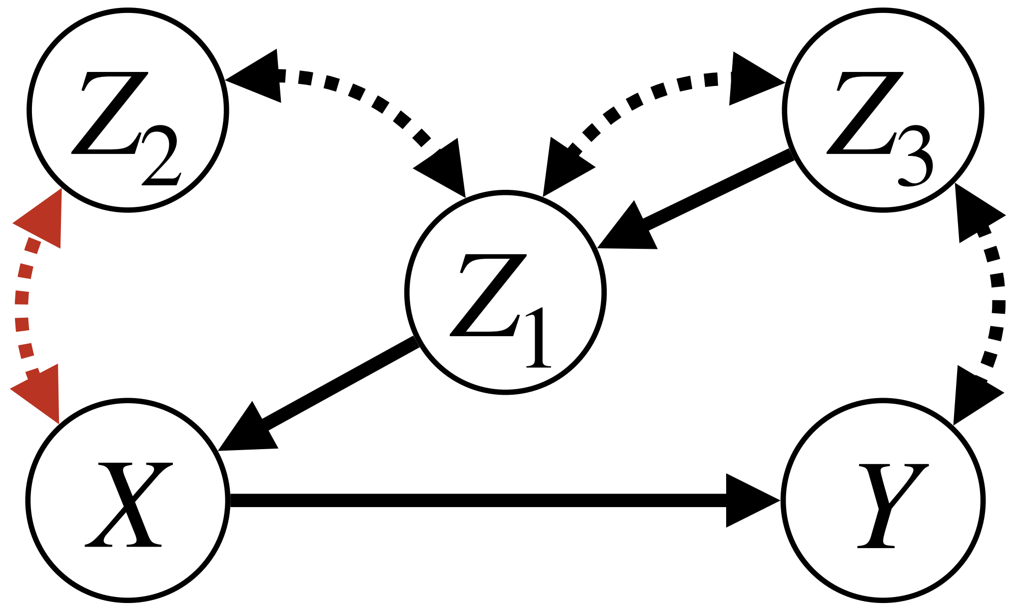
Identification in Semi-Markovian Models

Adjustment over parents:

Let G be a causal graph with **no unmeasured parents**.

Then, the effect of $\mathbf{X}$ on $\mathbf{Y}$ is given by:

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



$$Pa_x = \{Z_2\}$$
$$U_x = \{U_{X,Z2}\}$$

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

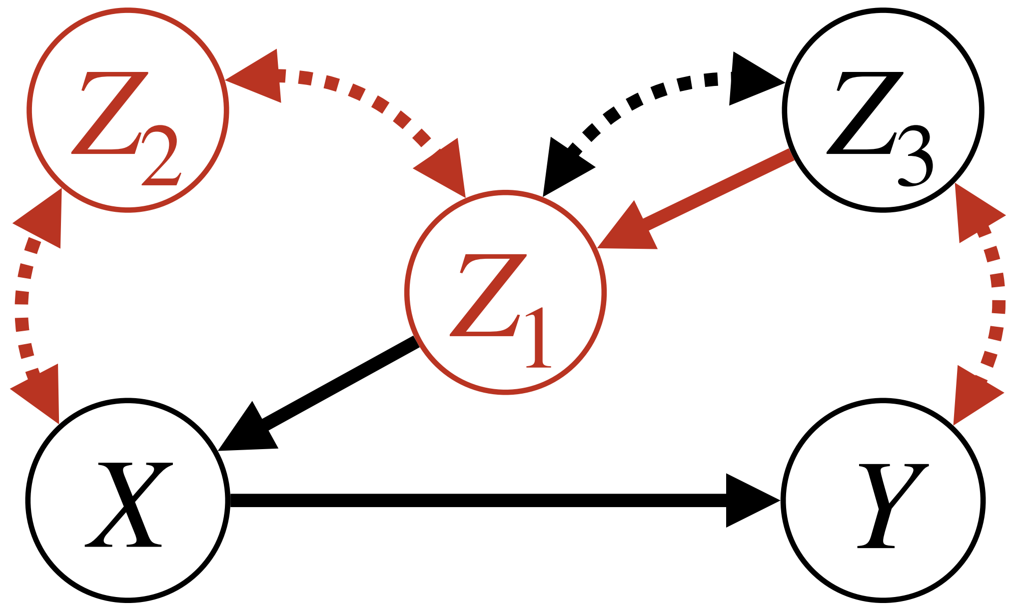
Identification in Semi-Markovian Models

Adjustment over parents:

Let G be a causal graph with **no unmeasured parents**.

Then, the effect of $\mathbf{X}$ on $\mathbf{Y}$ is given by:

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



$$Pa_x = \{Z_2\}$$
$$U_x = \{U_{X,Z2}\}$$

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

After conditioning on the $\{Z_1, Z_2\}$, the association between X and Y is also due to a spurious / confounding path.

Backdoor Adjustment

Also known as ***confounding paths***, or ***backdoor paths***.

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:

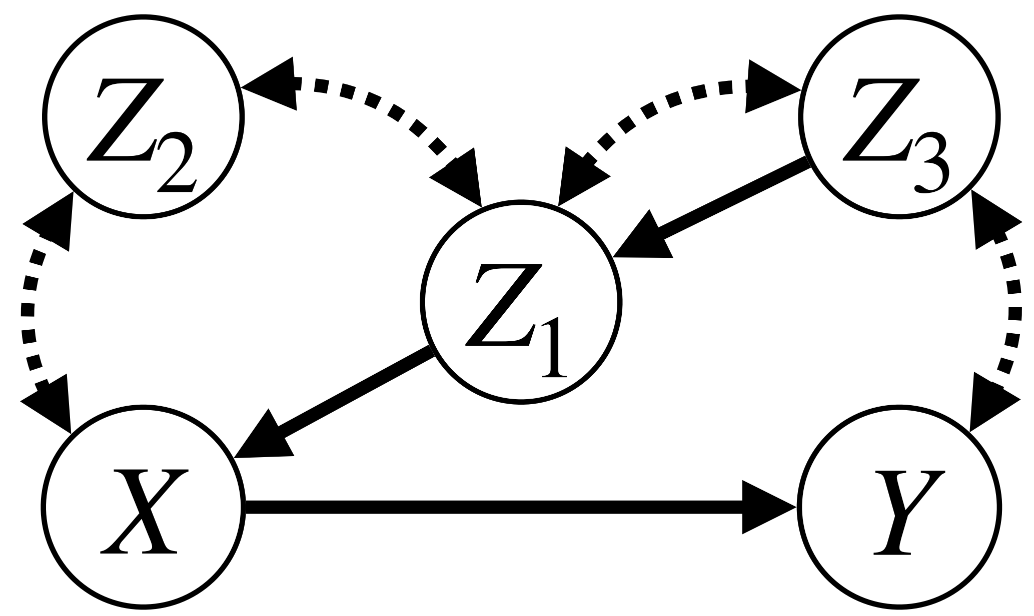
1. for every $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$, $\mathbf{Z}$ blocks every path between X and Y that has an arrow into X, and
2. no node in $\mathbf{Z}$ is a descendant of a variable $X \in \mathbf{X}$ (all variables in $\mathbf{Z}$ are pre-treatment)

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:

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

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

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



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

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

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

Backdoor Adjustment

Also known as ***confounding paths***, or ***backdoor paths***.

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:

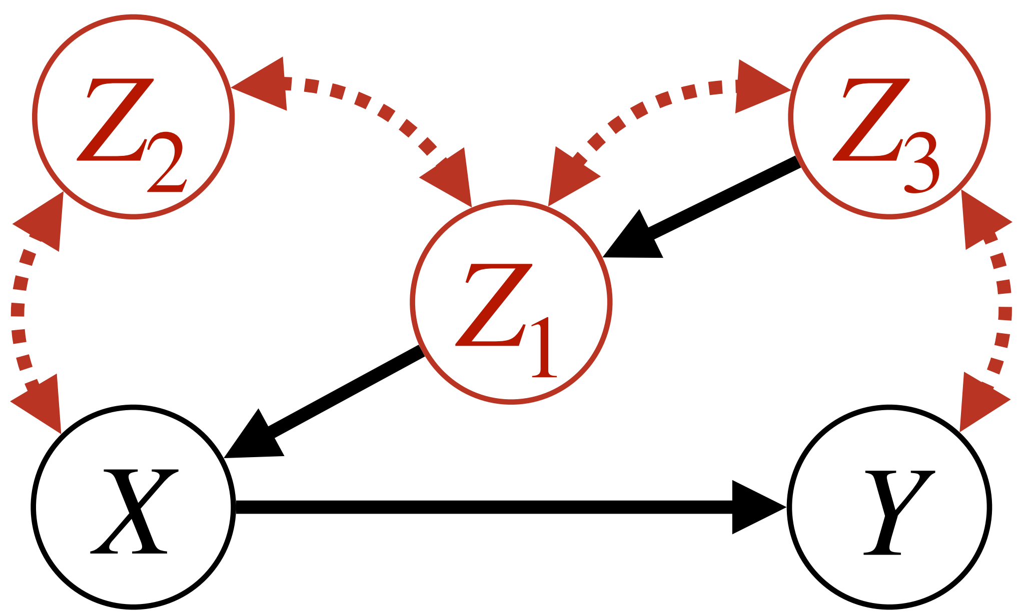
1. for every $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$, $\mathbf{Z}$ blocks every path between X and Y that has an arrow into X, and
2. no node in $\mathbf{Z}$ is a descendant of a variable $X \in \mathbf{X}$ (all variables in $\mathbf{Z}$ are pre-treatment)

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.

Backdoor Adjustment

Also known as ***confounding paths***, or ***backdoor paths***.

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:

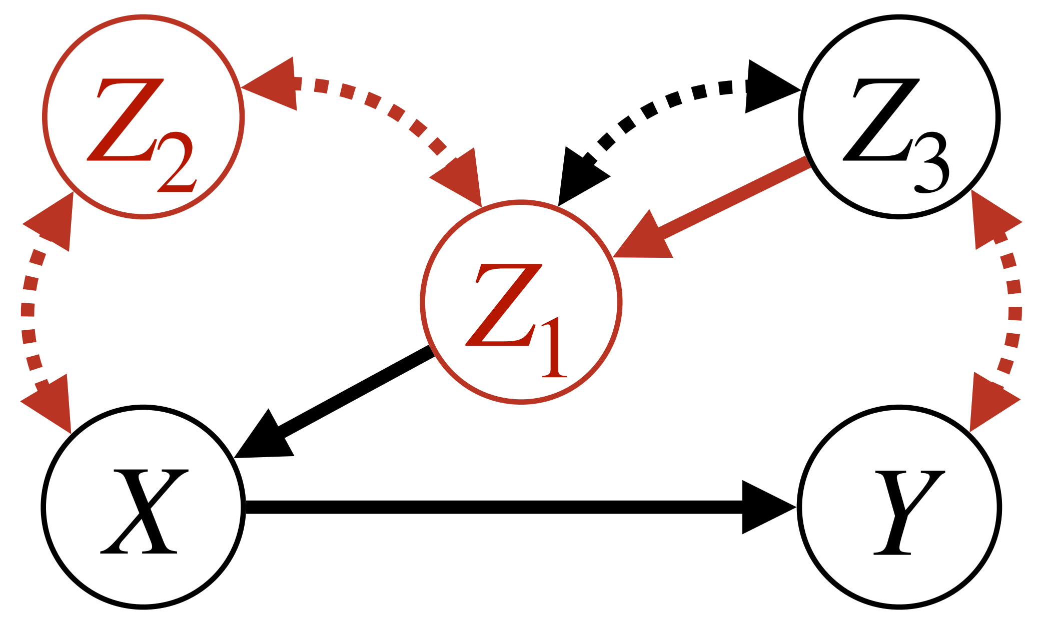
1. for every $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$, $\mathbf{Z}$ blocks every path between X and Y that has an arrow into X, and
2. no node in $\mathbf{Z}$ is a descendant of a variable $X \in \mathbf{X}$ (all variables in $\mathbf{Z}$ are pre-treatment)

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:

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

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

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



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

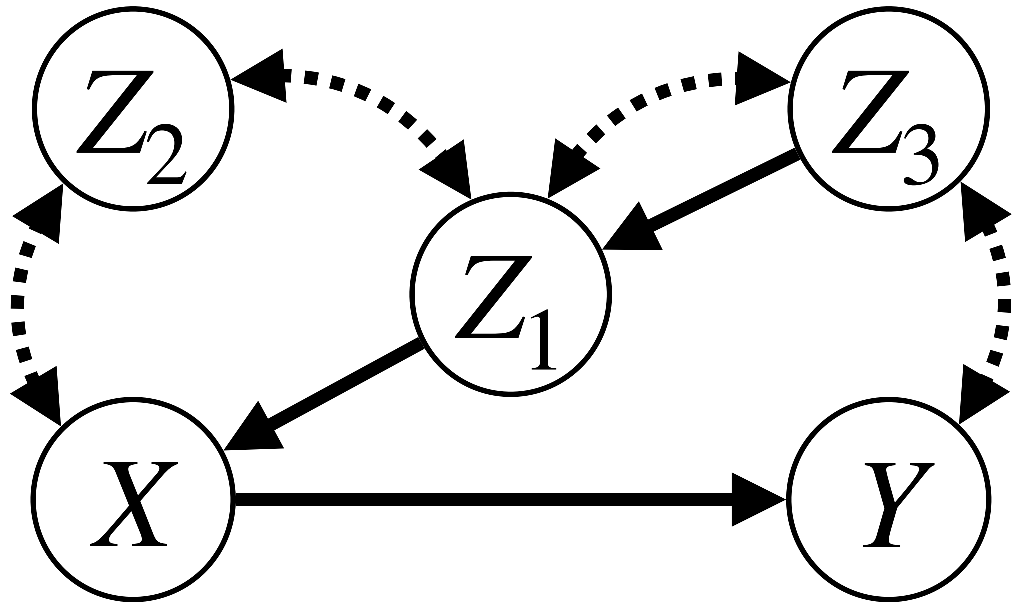
$\mathbf{Z} = \{Z_1, Z_3\}$

$\mathbf{Z} = \{Z_1, Z_2\}$ ✘

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

Estimation via Propensity Scores

Theorem: If the set $\mathbf{Z}$ *satisfies the parent / backdoor criterion* w.r.t. the ordered pair (X, Y) in the causal graph G, then the causal effect of X on Y is identifiable (**uniquely computable**) and given by:



Backdoor Adjustment ≡ Conditional Ignorability:
$Y_x \perp\!\!\!\perp X \mid \mathbf{Z}$

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

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

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

propensity score
neural nets

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

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

Only if $\mathbf{Z}$ satisfies the BD criterion, Inverse Probability Weighting/ Propensity Score can be used to estimate $P(y \mid do(x))$.

56

What if backdoor adjustment does not work?

Identification via Front-Door Adjustment

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

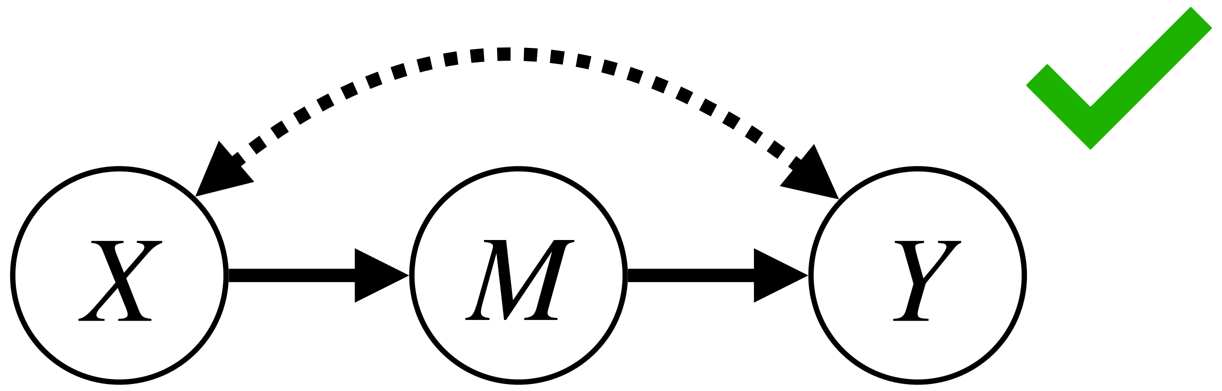
If there exists a set $\mathbf{M}$ such that:

1. $\mathbf{M}$ intercepts all directed paths from any vertex $X \in \mathbf{X}$ to any vertex $Y \in \mathbf{Y}$;
2. There is no unblocked back-door path from any vertex $X \in \mathbf{X}$ to vertex $M \in \mathbf{M}$; and
3. All back-door paths from any vertex $M \in \mathbf{M}$ to any vertex $Y \in \mathbf{Y}$ are blocked by $\mathbf{X}$.

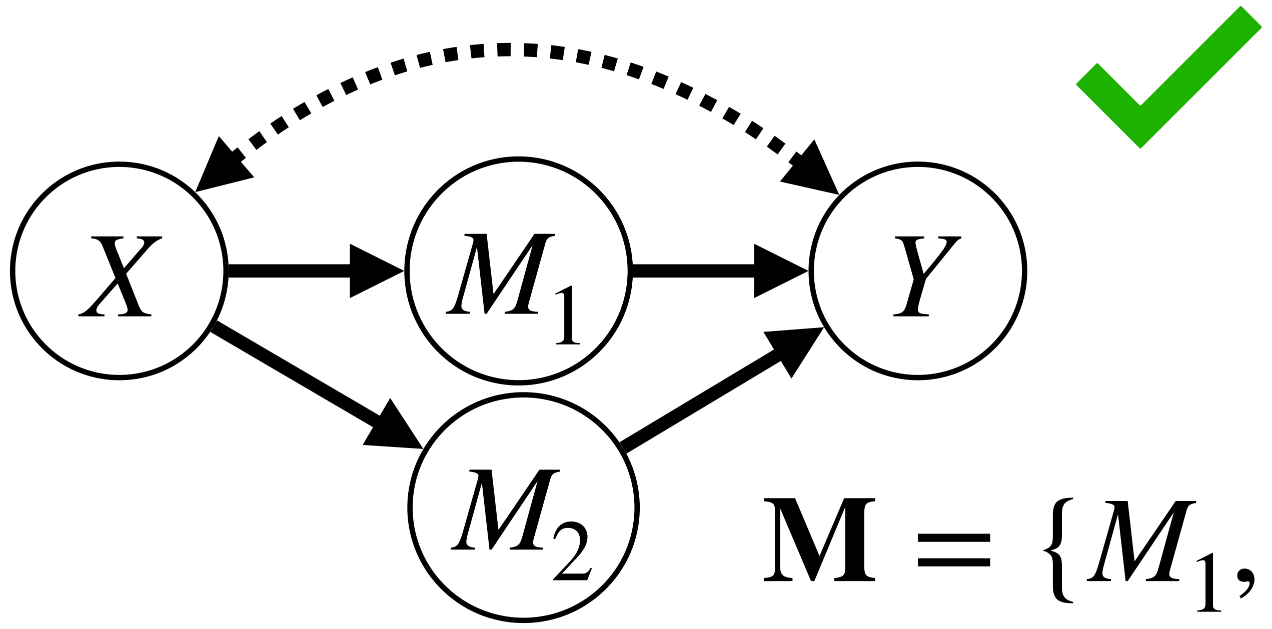
Then, $\mathbf{M}$ satisfies the ***front-door criterion*** and, then the effect of $\mathbf{X}$ on $\mathbf{Y}$ is given by:

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

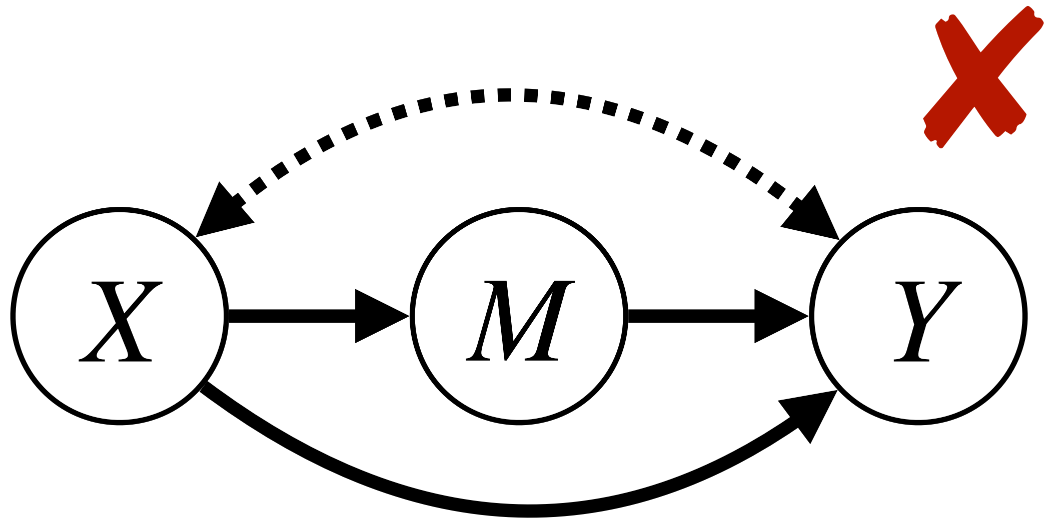
$\mathbf{X} = \{X\}$
$\mathbf{Y} = \{Y\}$

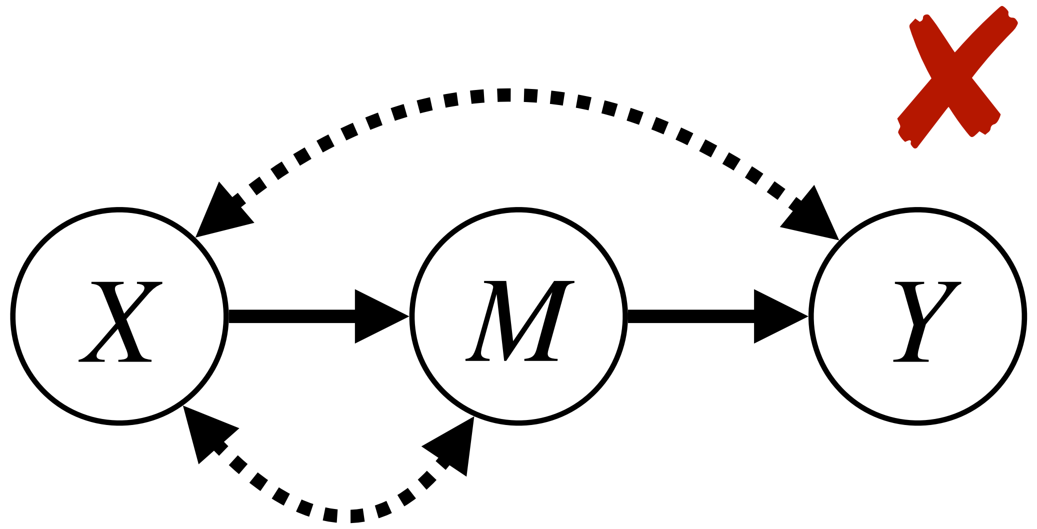


$\mathbf{M} = \{M\}$

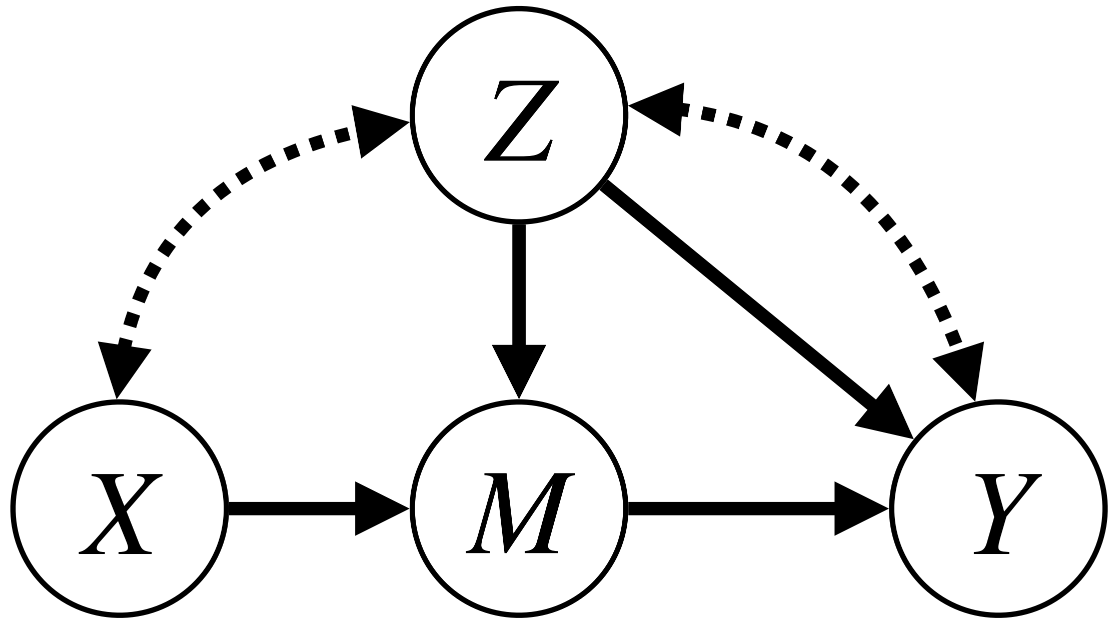


$\mathbf{M} = \{M_1, M_2\}$

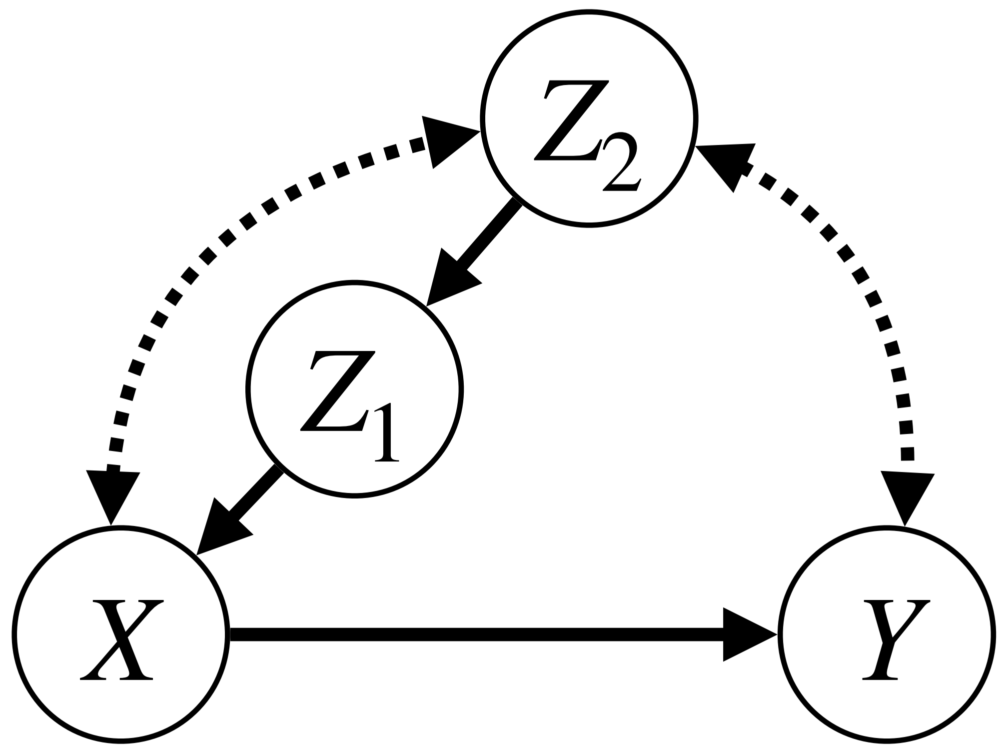




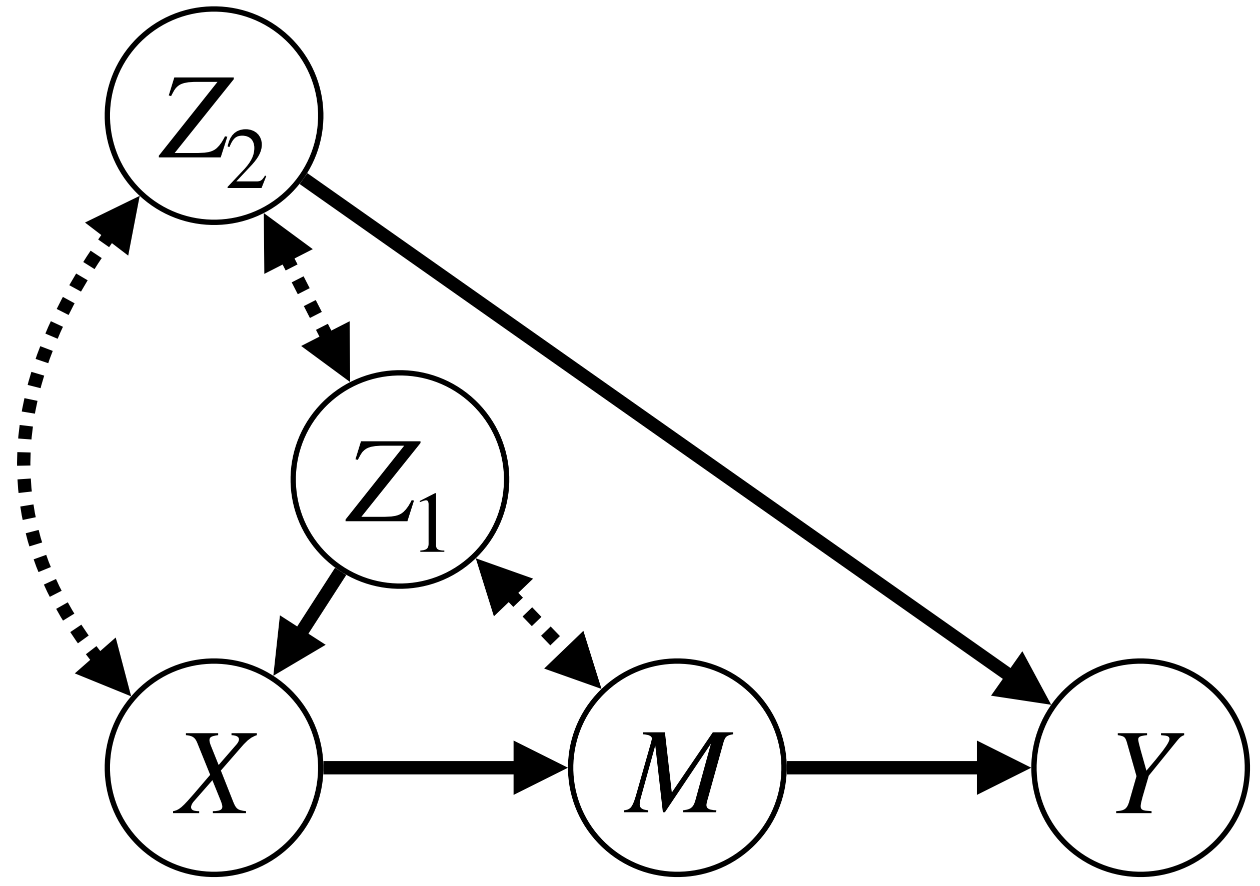
Many scenarios beyond back-door and front-door!



Conditional Front-Door



Napkin



Unnamed

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

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

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

And many others….

Do-Calculus (a.k.a. Causal Calculus)

Pearl, 1995

Graphical conditions implying invariances between observational ($\mathscr{L}_1$) and interventional ($\mathscr{L}_2$) distributions

Theorem: Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}$ be any disjoint subjects of variables.

Rule 1 (Insertion/Deletion of Observations)

$$P(\mathbf{y} \mid do(\mathbf{w}), \mathbf{x}, \mathbf{z}) = P(\mathbf{y} \mid do(\mathbf{w}), \mathbf{z}), \text{ if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}, \mathbf{W})_{G_{\overline{\mathbf{W}}}}$$

Rule 2 (Exchange of Actions and Observations)

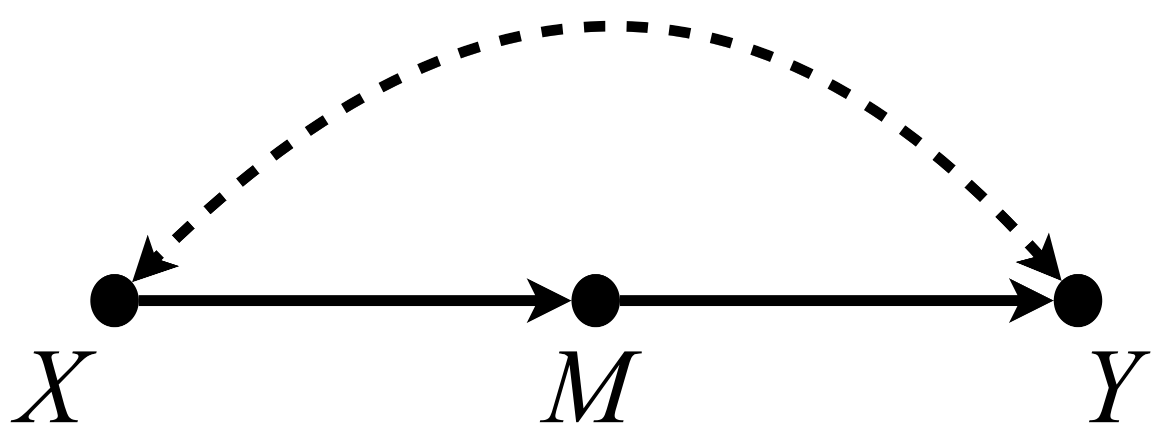
$$P(\mathbf{y} \mid do(\mathbf{w}), do(\mathbf{x}), \mathbf{z}) = P(\mathbf{y} \mid do(\mathbf{w}), \mathbf{x}, \mathbf{z}), \text{ if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}, \mathbf{W})_{G_{\overline{\mathbf{W}}\underline{\mathbf{X}}}}$$

Rule 3 (Insertion/Deletion of Actions)

$$P(\mathbf{y} \mid do(\mathbf{w}), do(\mathbf{x}), \mathbf{z}) = P(\mathbf{y} \mid do(\mathbf{w}), \mathbf{z}), \text{ if } (\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}, \mathbf{W})_{G_{\overline{\mathbf{W}}, \overline{\mathbf{X}(\mathbf{Z})}}}$$

$G_{\overline{\mathbf{W}}\underline{\mathbf{X}}}$: graph G after removing the incoming arrows into $\mathbf{W}$ and the outgoing arrows from $\mathbf{X}$;
$\mathbf{X}(\mathbf{Z})$: set of $\mathbf{X}$-nodes that are not ancestors of any $\mathbf{Z}$-node in $G_{\overline{\mathbf{W}}}$.

Identification in Non-Markovian Models



$$
\begin{aligned}
P(y \mid do(x)) &= \sum_m P(y \mid do(x), m) P(m \mid do(x)) && \text{Probability Axioms} \\
&= \sum_m P(y \mid do(x), do(m)) P(m \mid do(x)) && \text{Rule 2} \\
&= \sum_m P(y \mid do(x), do(m)) P(m \mid x) && \text{Rule 2} \\
&= \sum_m P(y \mid do(m)) P(m \mid x) && \text{Rule 3} \\
&= \sum_{x'} \sum_m P(y \mid do(m), x') P(x' \mid do(m)) P(m \mid x) && \text{Probability Axioms} \\
&= \sum_{x'} \sum_m P(y \mid m, x') P(x' \mid do(m)) P(m \mid x) && \text{Rule 2} \\
&= \sum_{x'} \sum_m P(y \mid m, x') P(x' \mid m) P(m \mid x) && \text{Rule 3}
\end{aligned}
$$

The Identify (ID) Algorithm

Algorithm 1 $\mathbf{ID}(\mathbf{x}, \mathbf{y})$ given Causal Diagram $\mathcal{G}$

Input: two disjoint sets $\mathbf{X}, \mathbf{Y} \subset \mathbf{V}$
Output: Expression for $P_{\mathbf{x}}(\mathbf{y})$ or `FAIL`

```
Let D = An(Y)_{G_{V\X}}
Let the c-components of G_D be D_i, i = 1, ..., k
P_x(y) = Σ_{d\y} ∏_i IDENTIFY(D_i, V, P)

function IDENTIFY(C, T, Q = Q[T])
    if C = T then return Q[T]
    /* Let S^B denote the c-component of {B} in G_T */
    if ∃B ∈ T \ C such that S^B ∩ Ch(B) = ∅ then
        Compute Q[T \ {B}] from Q;          ▷ Lemma 1
        return IDENTIFY(C, T \ {B}, Q[T \ {B}])
    else
        throw FAIL
```

Lemma 1. *Given a causal diagram $\mathcal{D}$ over $\mathbf{V}$, $X \in \mathbf{T} \subseteq \mathbf{V}$, and $P_{\mathbf{v} \setminus \mathbf{t}}$, i.e., an expression for $Q[\mathbf{T}]$. If X is not in the same c-component with a child in $\mathcal{D}_{\mathbf{T}}$, then $Q[\mathbf{T} \setminus \{X\}]$ is identifiable and given by*

$$Q[\mathbf{T} \setminus \{X\}] = \frac{P_{\mathbf{v} \setminus \mathbf{t}}}{Q[S^X]} \times \sum_x Q[S^X] \tag{2}$$

where S^X is the c-component of X in $\mathcal{D}_{\mathbf{T}}$ and $Q[S^X]$ is computable from $P_{\mathbf{v} \setminus \mathbf{t}}$ by [Tian, 2002, Lemma 11].

Jin Tian. Studies in causal reasoning and learning. PhD thesis, University of California, Los Angeles, 2002. Link

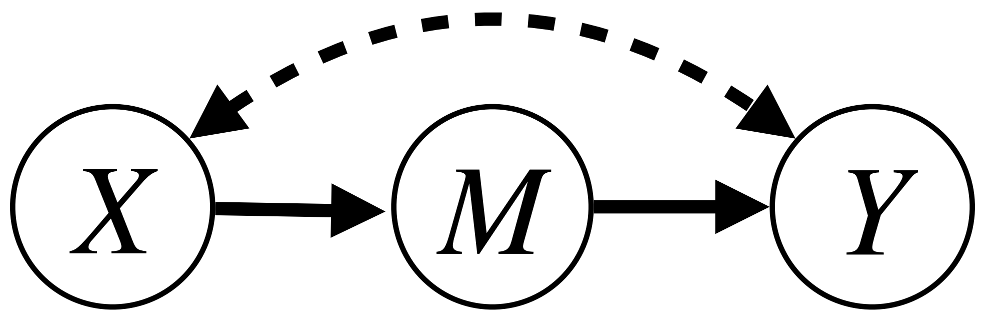
Causal Effect Identification

ID-Algorithm and many recent generalizations.

1 **Query**

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

2 **Causal Contraints**



3 **Probability Distributions**

$P(x, m, y)$ **Observational Distribution**

Inference Engine

Solution

yes / no

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

Interventional Distribution ← Available Distributions

- 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.

More on Causal Effect Identification

Identification from observational and experimental data:

Lee, S., Correa, J., and Bareinboim, E. (**2019**). General identifiability with arbitrary surrogate experiments. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, volume 35, Tel Aviv, Israel. AUAI Press. Link

Identification of stochastic/soft (and possibly imperfect) interventions:

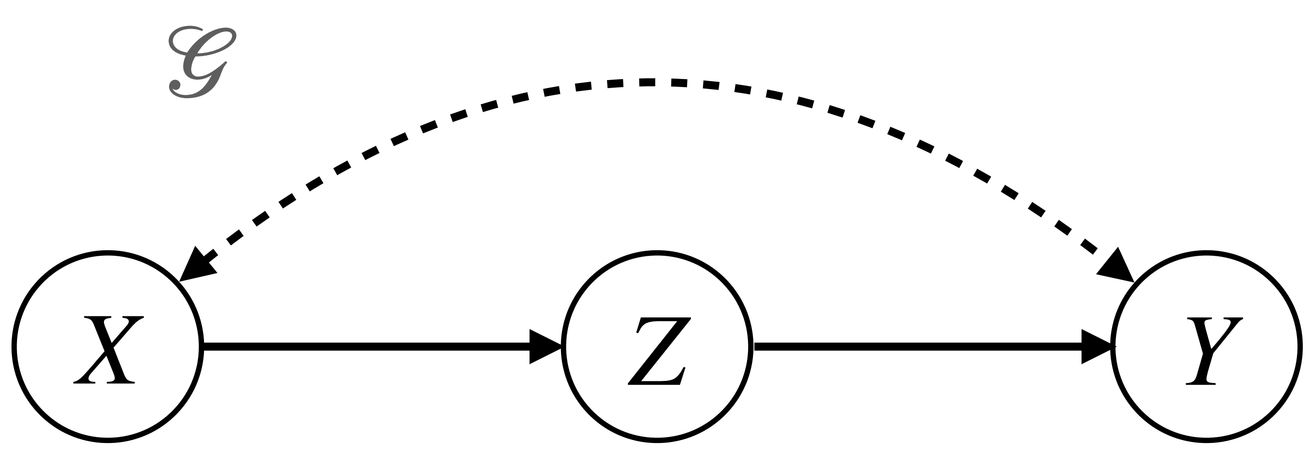
Correa, J. and Bareinboim, E. (2020). A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY. AAAI Press. Link

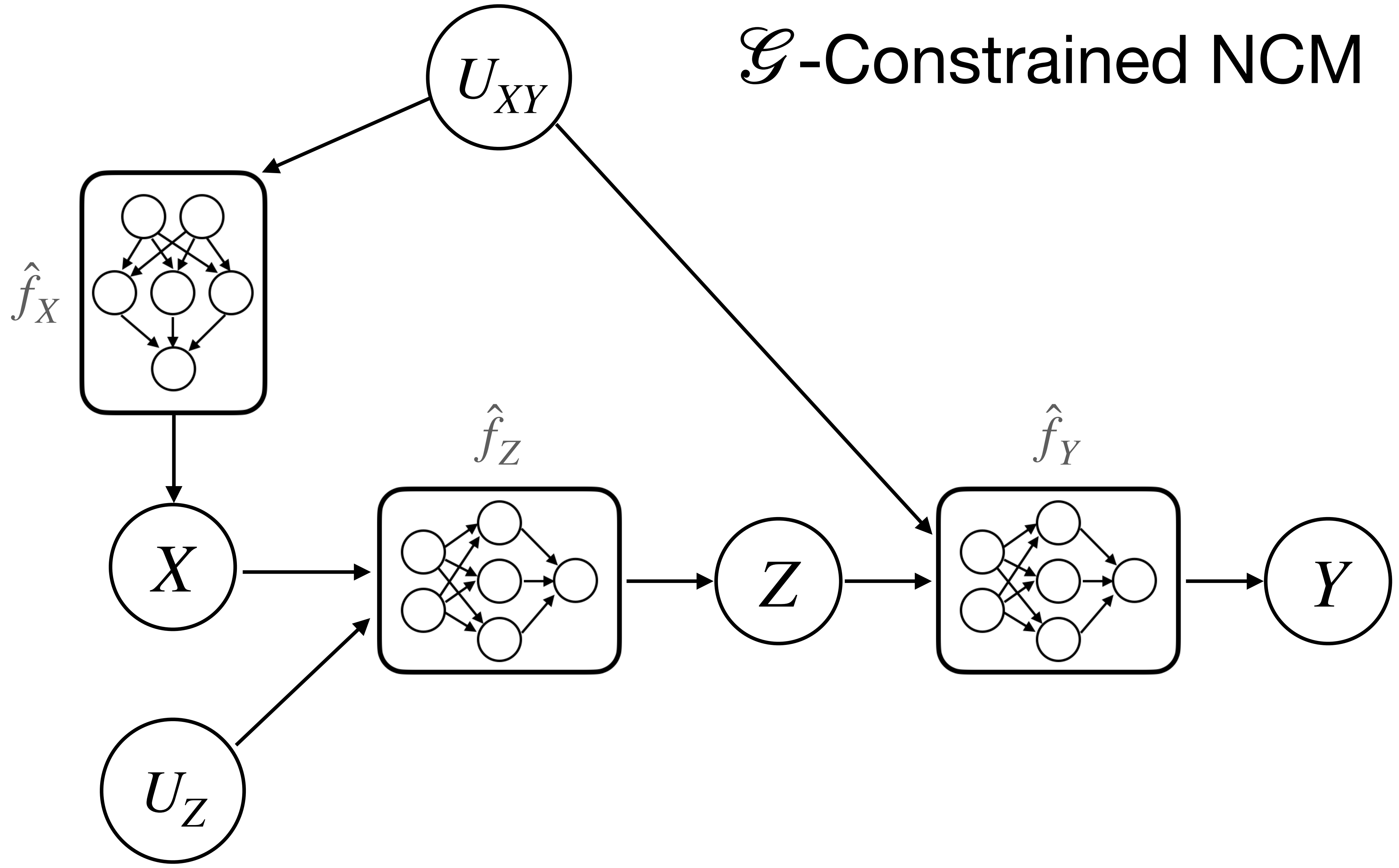
Identification and Estimation via Deep Neural Networks:

Xia, K., Lee, K.-Z., Bengio, Y., and Bareinboim, E. (2021). The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34. Link

Identification and Estimation via Deep Neural Networks

True Model:
SCM $\mathscr{M}^*$



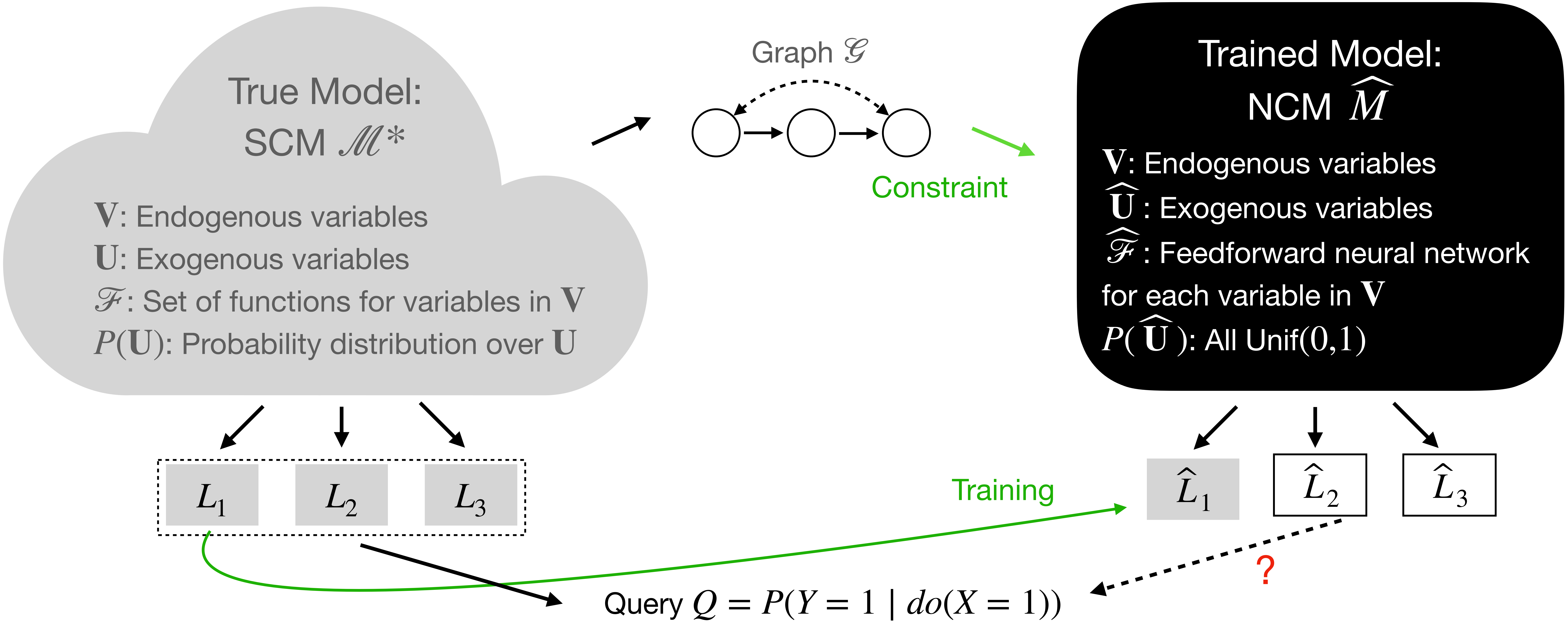


Trained Model:
$\mathcal{G}$-NCM $\widehat{M}$

$\mathbf{V}$: Endogenous variables
$\widehat{\mathbf{U}}$: Create one for every bidirected clique
$\widehat{\mathscr{F}}$: Feedforward neural network for each variable in $\mathbf{V}$ with parents from the graph
$P(\widehat{\mathbf{U}})$: All Unif$(0,1)$

Inductive bias based on the causal diagram: the enforced constraints empower the NCM with the ability to solve causal inference tasks.

Expressiveness of NCMs



Thm: For any SCM $\mathscr{M}^*$, there exists an NCM $\widehat{M}$ such that $\widehat{M}$ matches $\mathscr{M}^*$ on all three PCH layers!

This does not imply that the estimated NCM $\widehat{M}$ matches the true SCM $\mathscr{M}^*$!

Solution: A Neural Algorithm for Identification

Algorithm 1: Identifying/estimating queries with NCMs.

Input : causal query $Q = P(\mathbf{y} \mid do(\mathbf{x}))$, L_1 data $P(\mathbf{v})$, and causal diagram $\mathcal{G}$

Output : $P^{\mathcal{M}^*}(\mathbf{y} \mid do(\mathbf{x}))$ if identifiable, `FAIL` otherwise.

$\widehat{M} \leftarrow$ `NCM(`$\mathbf{V}, \mathcal{G}$`)` `// from Def. 7`

$\boldsymbol{\theta}^*_{\min} \leftarrow \arg\min_{\boldsymbol{\theta}} P^{\widehat{M}(\boldsymbol{\theta})}(\mathbf{y} \mid do(\mathbf{x}))$ s.t. $L_1(\widehat{M}(\boldsymbol{\theta})) = P(\mathbf{v})$

$\boldsymbol{\theta}^*_{\max} \leftarrow \arg\max_{\boldsymbol{\theta}} P^{\widehat{M}(\boldsymbol{\theta})}(\mathbf{y} \mid do(\mathbf{x}))$ s.t. $L_1(\widehat{M}(\boldsymbol{\theta})) = P(\mathbf{v})$

if $P^{\widehat{M}(\boldsymbol{\theta}^*_{\min})}(\mathbf{y} \mid do(\mathbf{x})) \neq P^{\widehat{M}(\boldsymbol{\theta}^*_{\max})}(\mathbf{y} \mid do(\mathbf{x}))$ **then**

    return `FAIL`

else

    return $P^{\widehat{M}(\boldsymbol{\theta}^*_{\min})}(\mathbf{y} \mid do(\mathbf{x}))$ `// choose min or max arbitrarily`

Maximize and minimize the induced causal query Q while maintaining L_1-consistency (can be done with likelihood estimation).

Thm. : Q is identifiable if and only if they match!

Corol: If Q is identifiable, then we can compute it by performing the mutilation procedure on $\widehat{M}$!

The approach is equivalent to established symbolic approaches (Thm. 4), and in identifiable cases, the result is an NCM that can serve as a proxy model for estimating the query (Corol. 2).

Can we relax some causal assumptions?

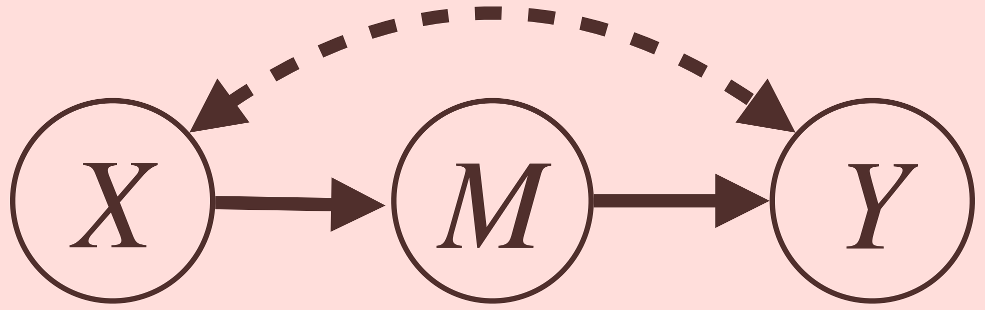
Causal Effect Identification

Usually hard to be fully specified.

1 **Query**

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

2 **Causal Contraints**



3 **Probability Distributions**

$P(x, m, y)$ **Observational Distribution**

Inference Engine

Solution

yes / no

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

Interventional Distribution ← Available Distributions

👍

- 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.

Is a Causal Diagram Still Too Much?

Causal diagrams are powerful tools that allow for inferences based on weaker knowledge (structural invariances) than the encoded in the true, underlying SCM.

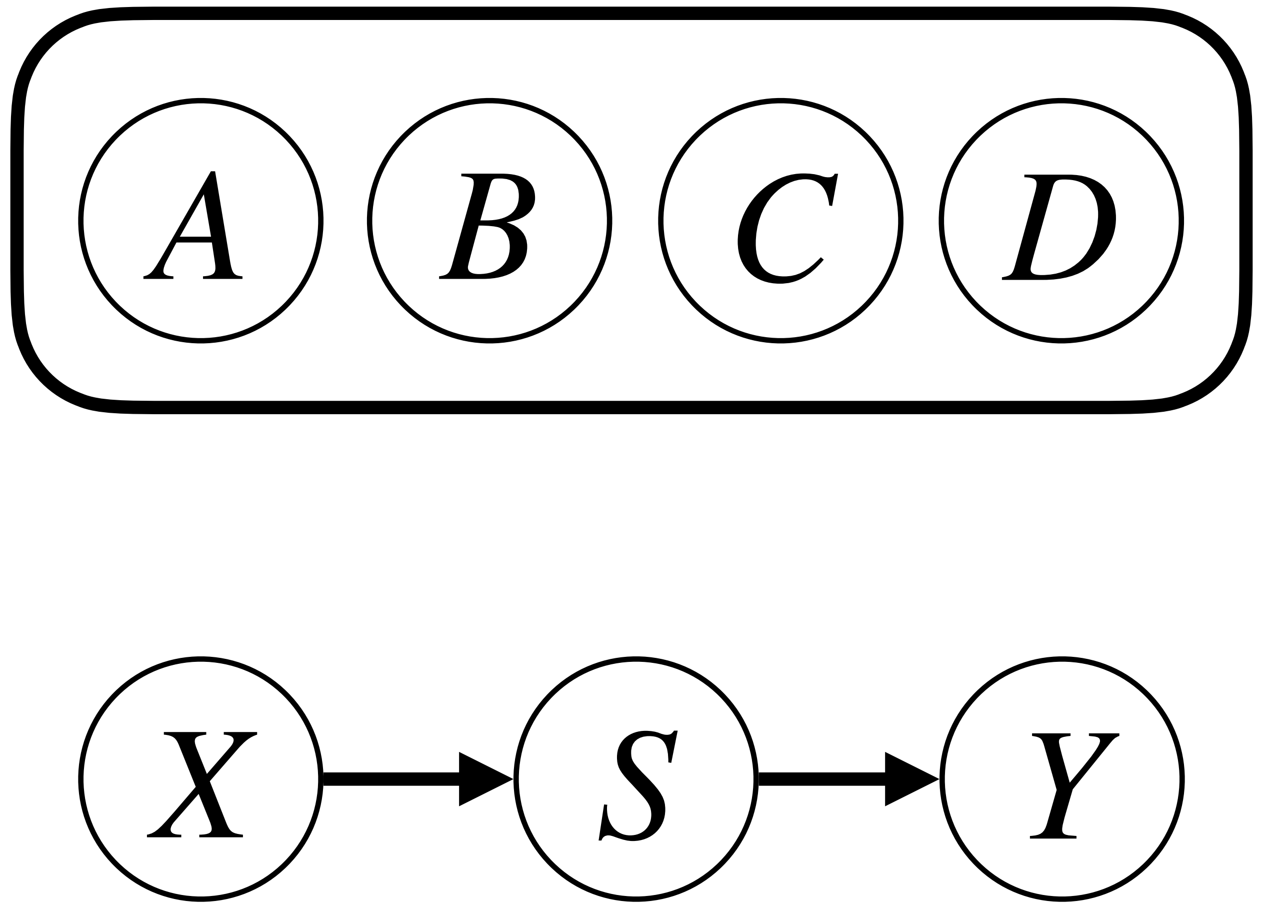
Still, structural knowledge for **every pair of variables** may not be available in many real-world, complex, high-dimensional systems.

Question:

Is it possible to relax the assumption of having a fully specified causal diagram and still be able to identify a causal effect?

Partially Understood Systems

A) Age
(B) Blood pressure
(C) Comorbidities
(D) Medication history
(X) Lisinopril
(S) Sleep Quality
(Y) Stroke

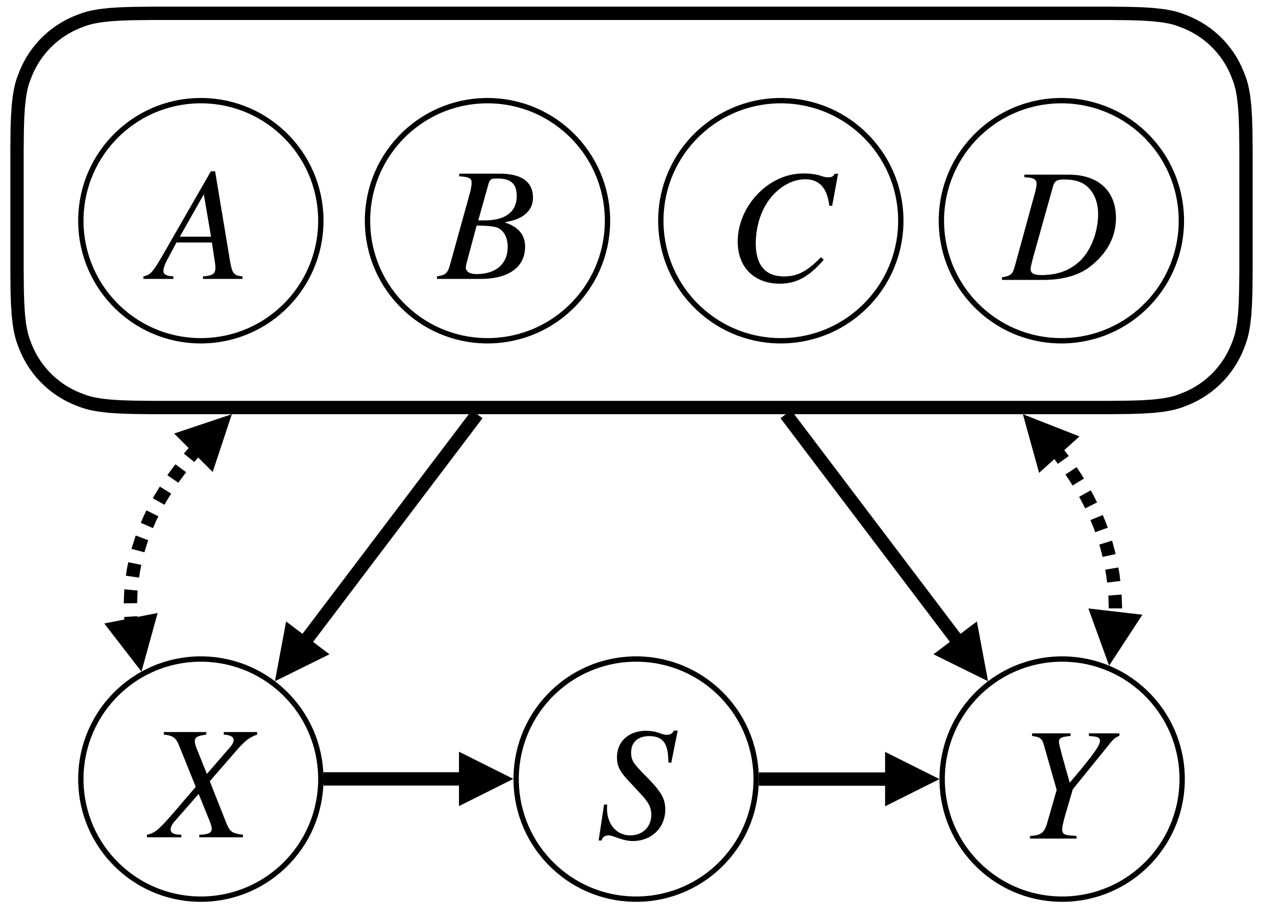


A causal diagram cannot be specified given the existing knowledge!

How can we identify $P(y \mid do(x))$ in this case?

Cluster DAGs (C-DAGs)

A) Age
(B) Blood pressure
(C) Comorbidities
(D) Medication history
(X) Lisinopril
(S) Sleep Quality
(Y) Stroke



$\{\{X\}, \{S\}, \{Y\}, \{A, B, C, D\}\}$

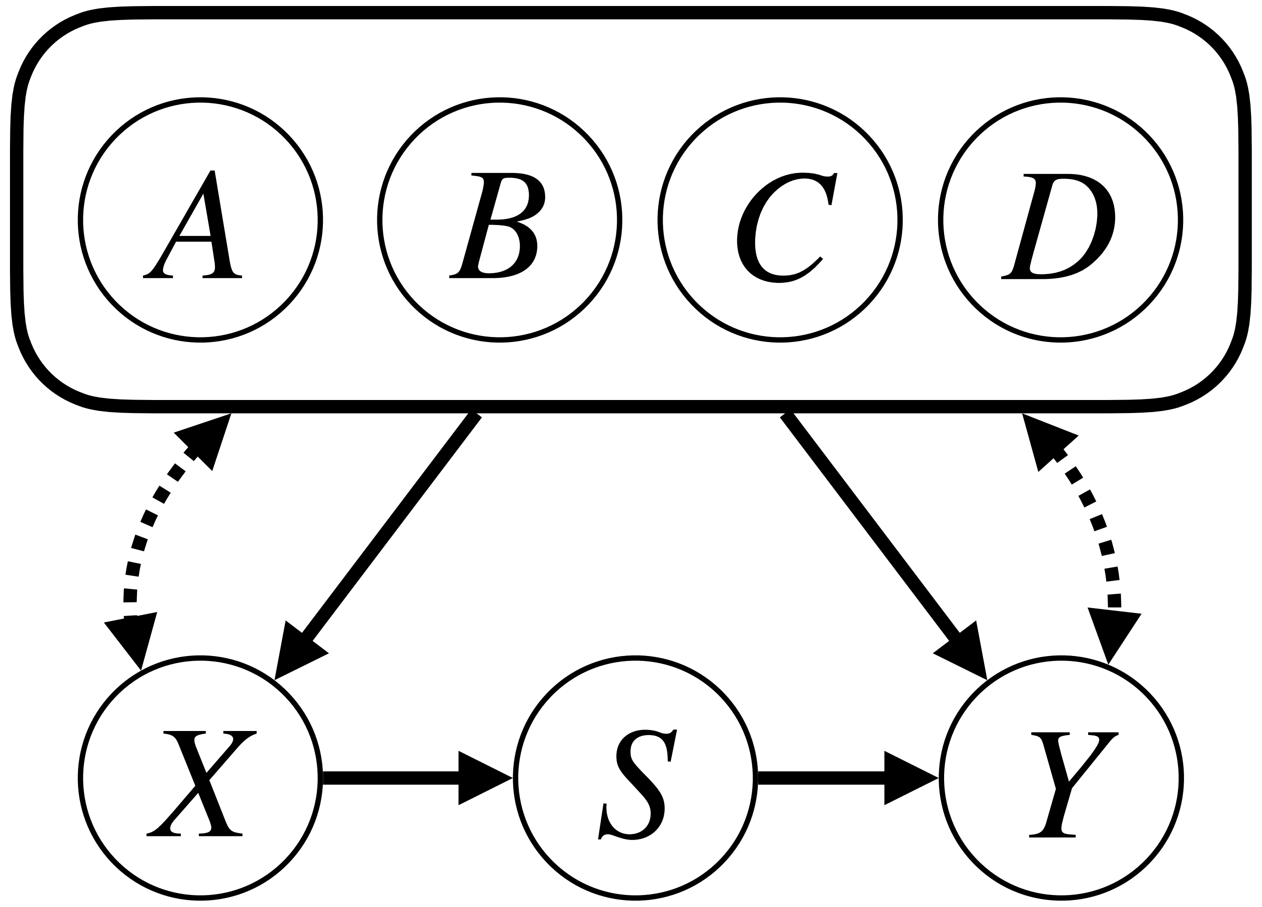
A *cluster DAG* $G_{\mathbf{C}}$ over a given partition $\mathbf{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_k\}$ of $\mathbf{V}$ is compatible with a causal diagram G over $\mathbf{V}$ if **for every $\mathbf{C}_i, \mathbf{C}_j \in \mathbf{C}$**:

- $\mathbf{C}_i \rightarrow \mathbf{C}_j$ if $\exists V_i \in \mathbf{C}_i$ and $V_j \in \mathbf{C}_j$ such that $V_i \rightarrow V_j$
- $\mathbf{C}_i \leftarrow\!\!\dashrightarrow \mathbf{C}_j$ if $\exists V_i \in \mathbf{C}_i$ and $V_j \in \mathbf{C}_j$ such that $V_i \leftarrow\!\!\dashrightarrow V_j$

and $G_{\mathbf{C}}$ contains no cycles.

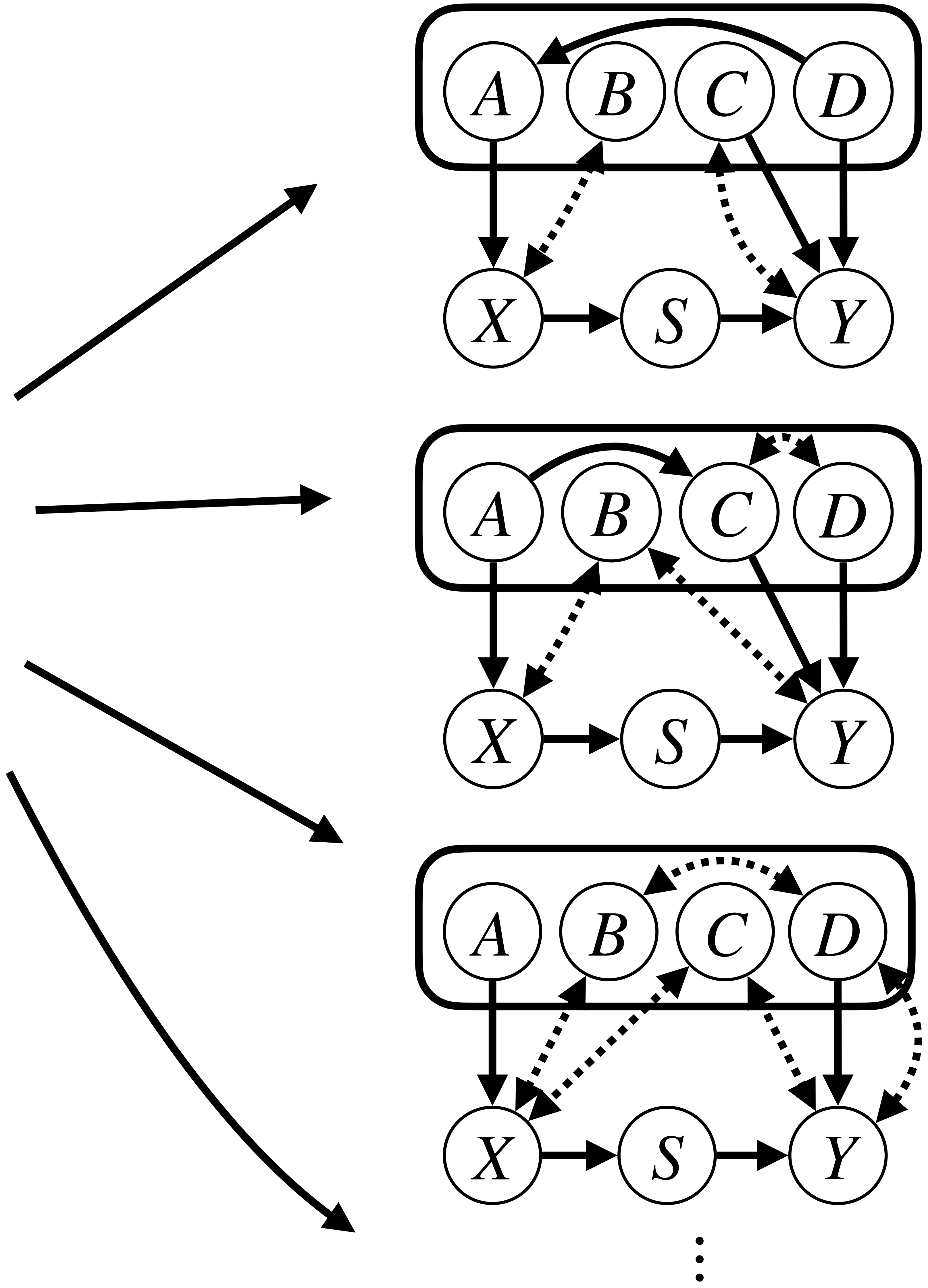
Partially Understood Systems

Many causal diagrams are compatible with the current knowledge!

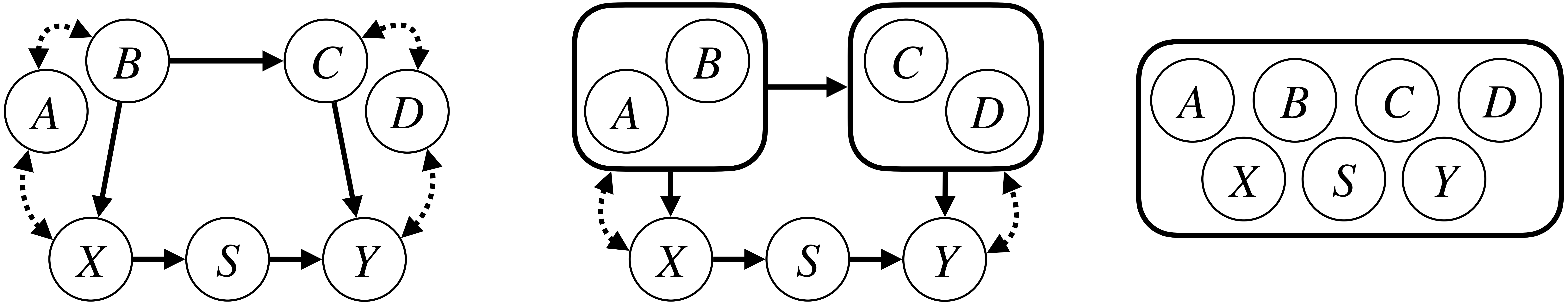


Can be seen as an *equivalence class* of causal diagrams, where any relationships are allowed among the variables within each cluster.

Can we infer causal effects without deciding on any one particular causal diagram?



C-DAG: Flexible Encoder of Model Assumptions

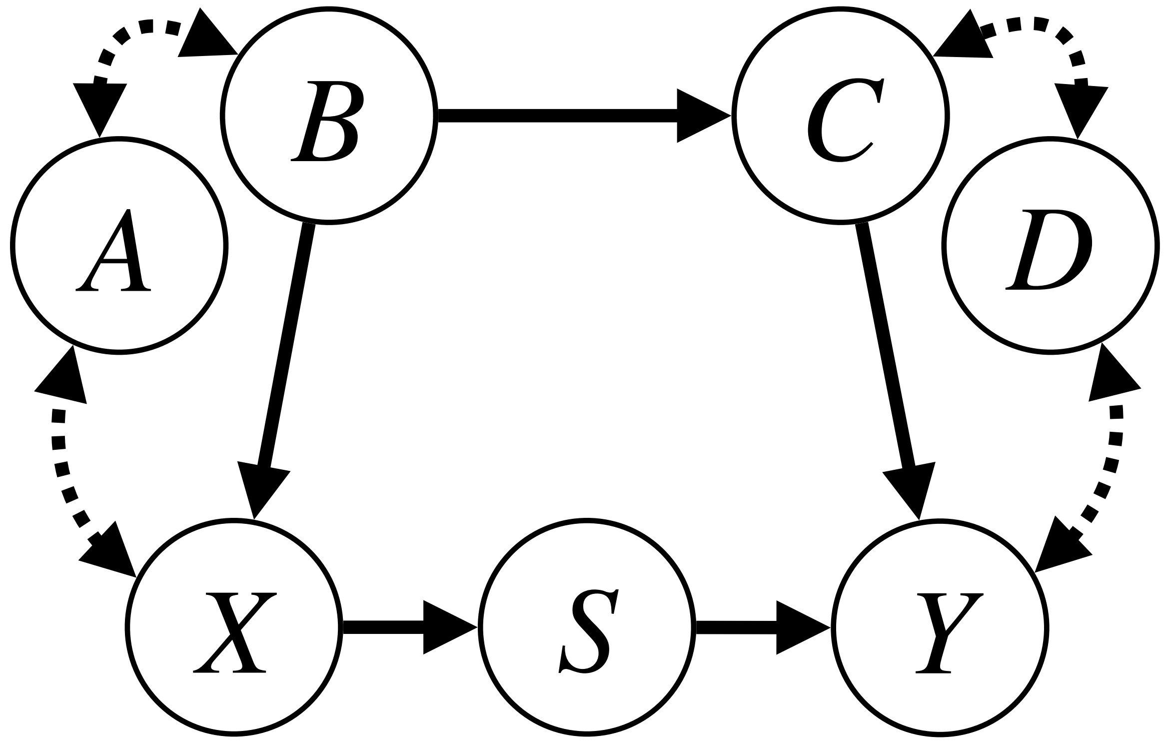


N clusters of size one
(full knowledge - DAG)

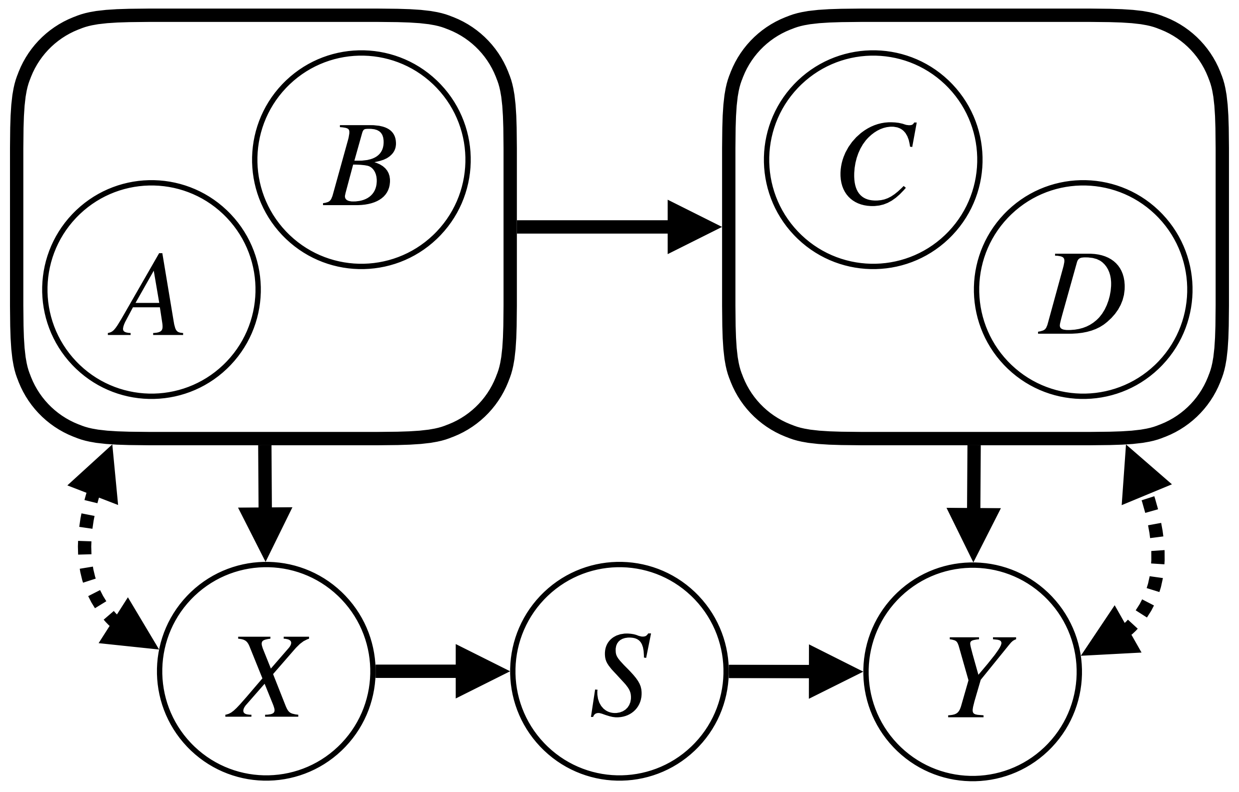
…
(partial knowledge - C-DAG)

One cluster of size N
(no knowledge)

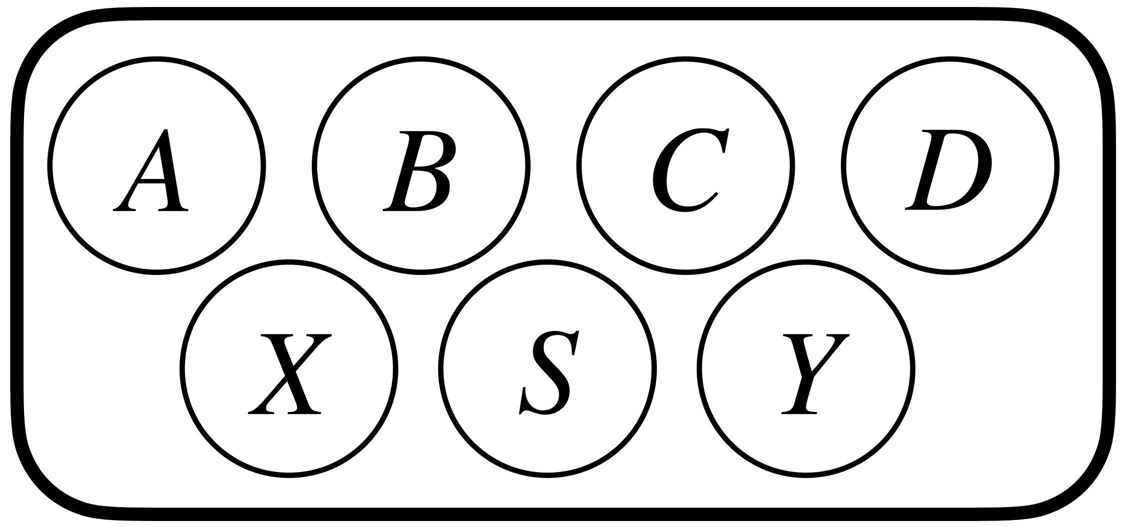
C-DAG: Flexible Encoder of Model Assumptions



N clusters of size one
(full knowledge - DAG)



…
(partial knowledge - C-DAG)



One cluster of size N
(no knowledge)

Clusters are manually created by domain experts:

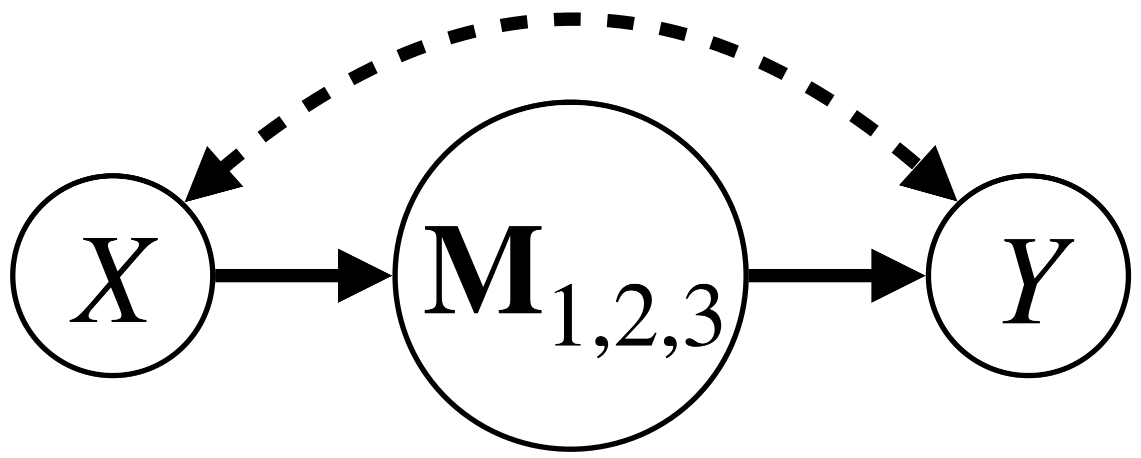
- due to lack of knowledge, consensus, or interest on the internal causal structure;
- to communicate relationships among semantically meaningful entities.

Identification of Causal Effects from C-DAGs

1 **Query**

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

2 **C-DAG**



3 **Data**

$P(x, m_1, m_2, m_3, y)$

Inference Engine

Solution

yes / no

$$P(y \mid do(x)) = \sum_{\mathbf{m}_{123}} P(\mathbf{m}_{123} \mid x) \sum_{x'} P(y \mid \mathbf{m}_{123}, x')P(x')$$

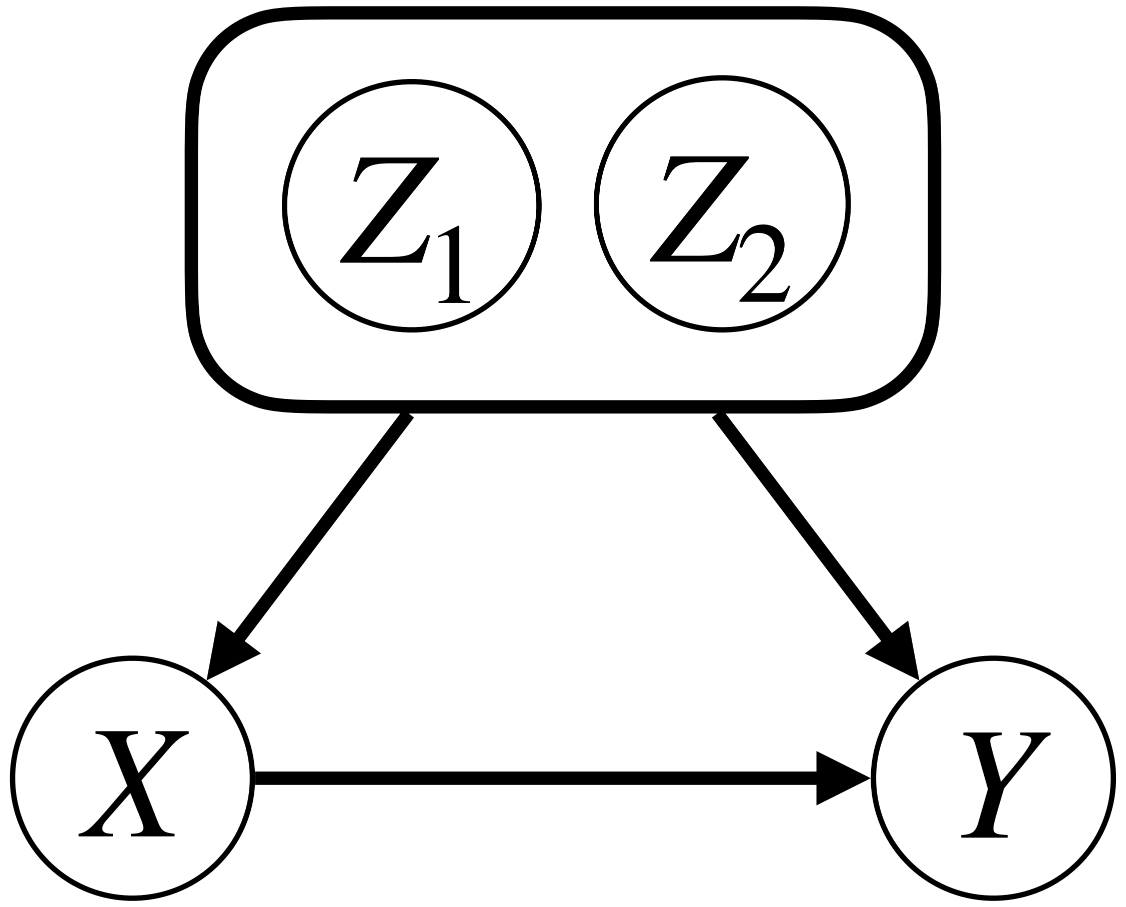
Inferred (Interventional) Distribution ← Available (Observational) Distribution

👍

Anand, T. V., **Ribeiro A. H.,** Tian, J., & Bareinboim, E. (2023). Causal Effect Identification in Cluster DAGs. In Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence.

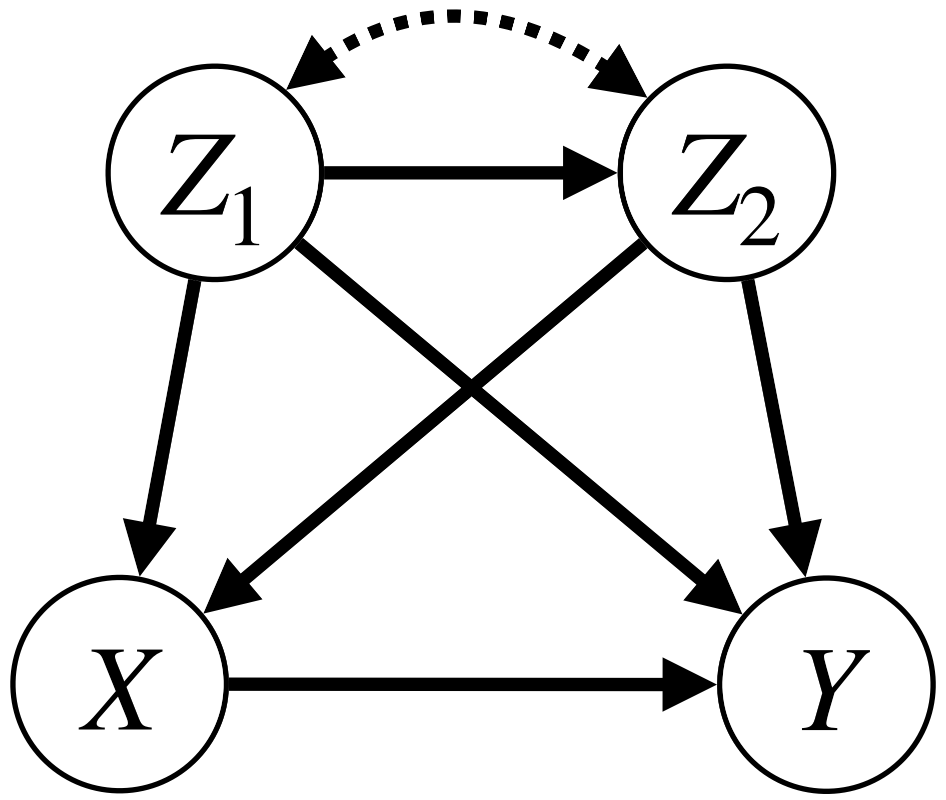
Effect Identifiabiliy given a C-DAG

$G_{\mathbf{C}}$



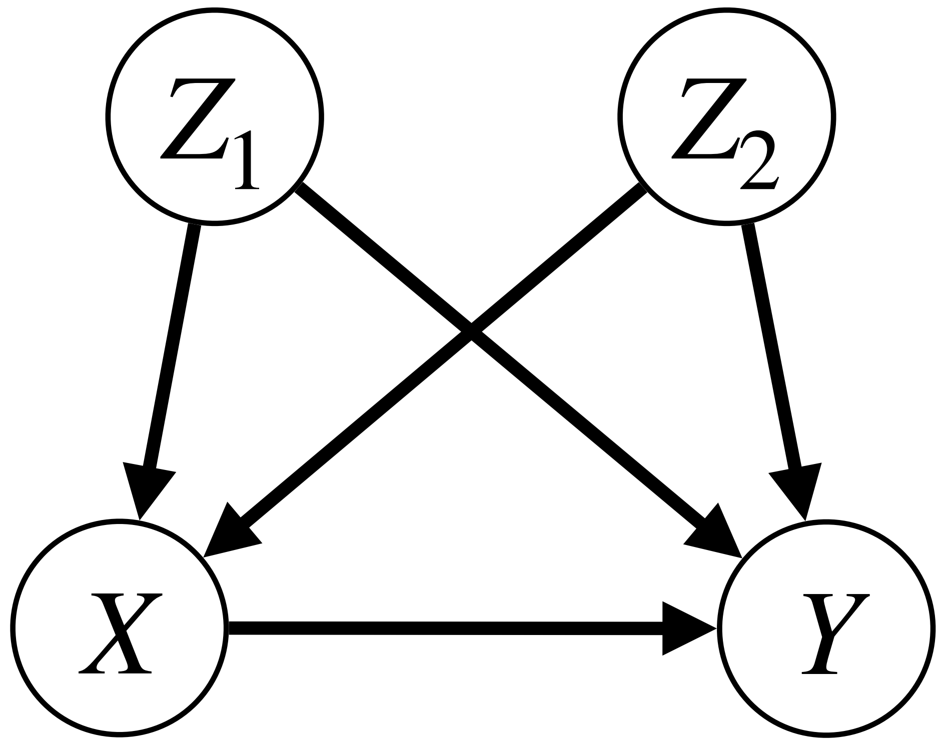
$$P(y \mid do(x)) = \sum_{\mathbf{z}} P\left(y \mid x, \mathbf{z}\right) P\left(\mathbf{z}\right)$$

G_1



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

G_2

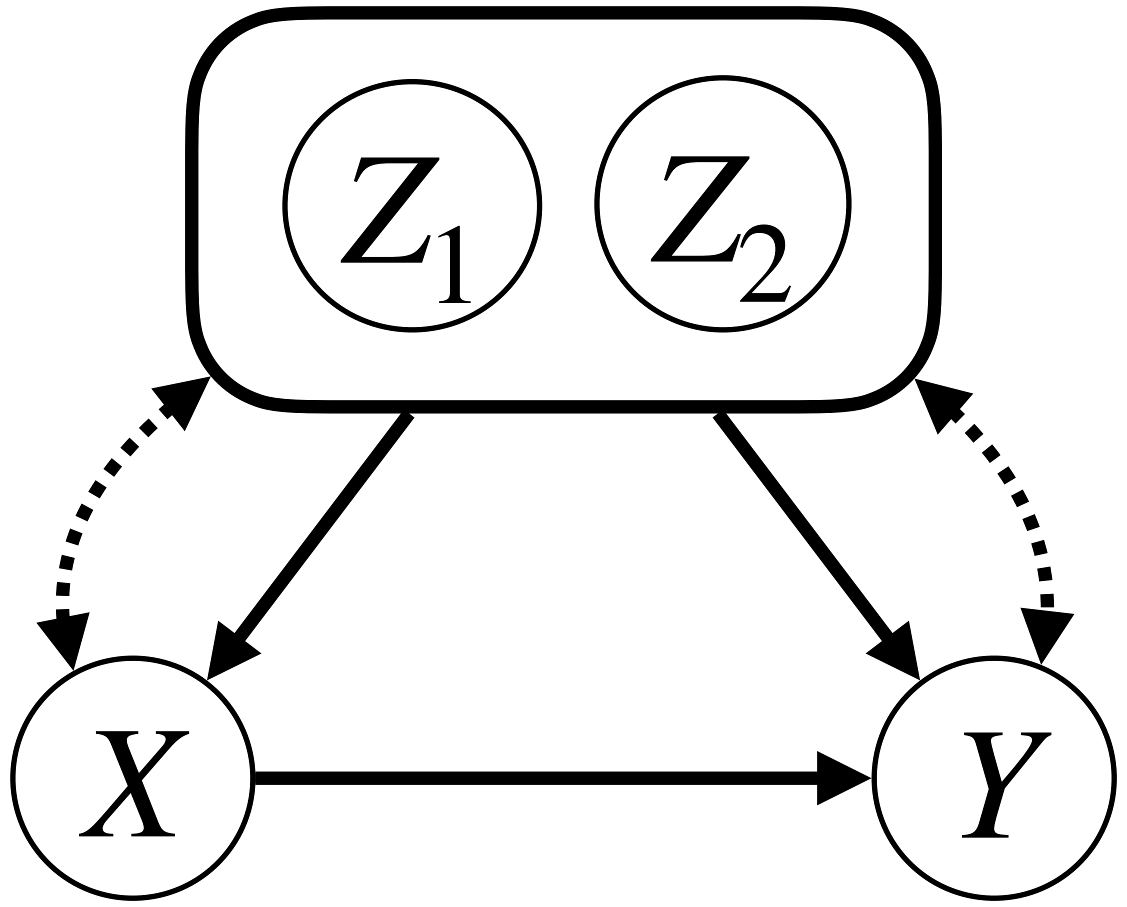


$$P(y \mid do(x)) = \sum_{z_1, z_2} P(y \mid x, z_1, z_2) P(z_1, z_2)$$

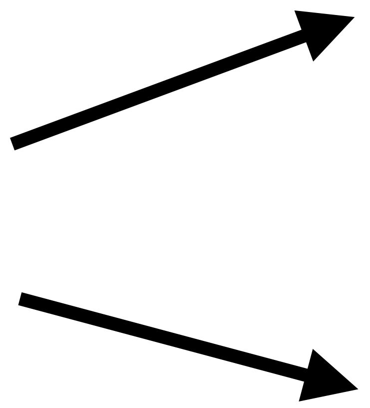
An effect identifiable in a C-DAG $G_{\mathbf{C}}$ is identifiable in all compatible causal diagrams G using the same identification formula!

Effect Non-Identifiabiliy given a C-DAG

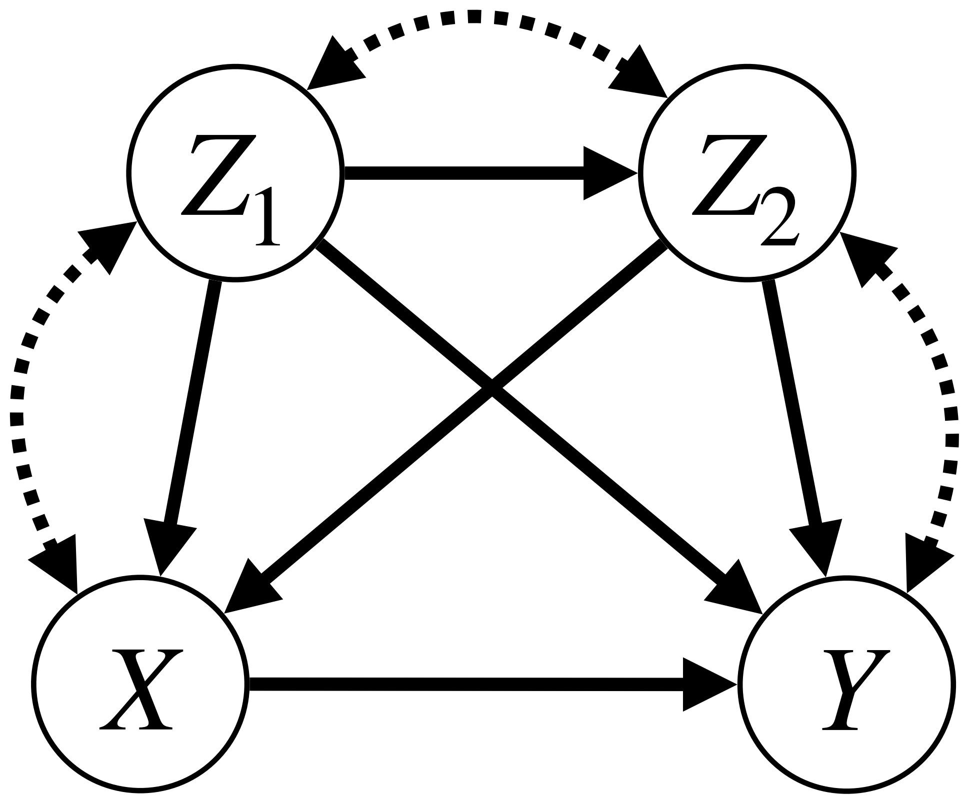
$G_{\mathbf{C}}$



$P(y \mid do(x))$ is not identifiable

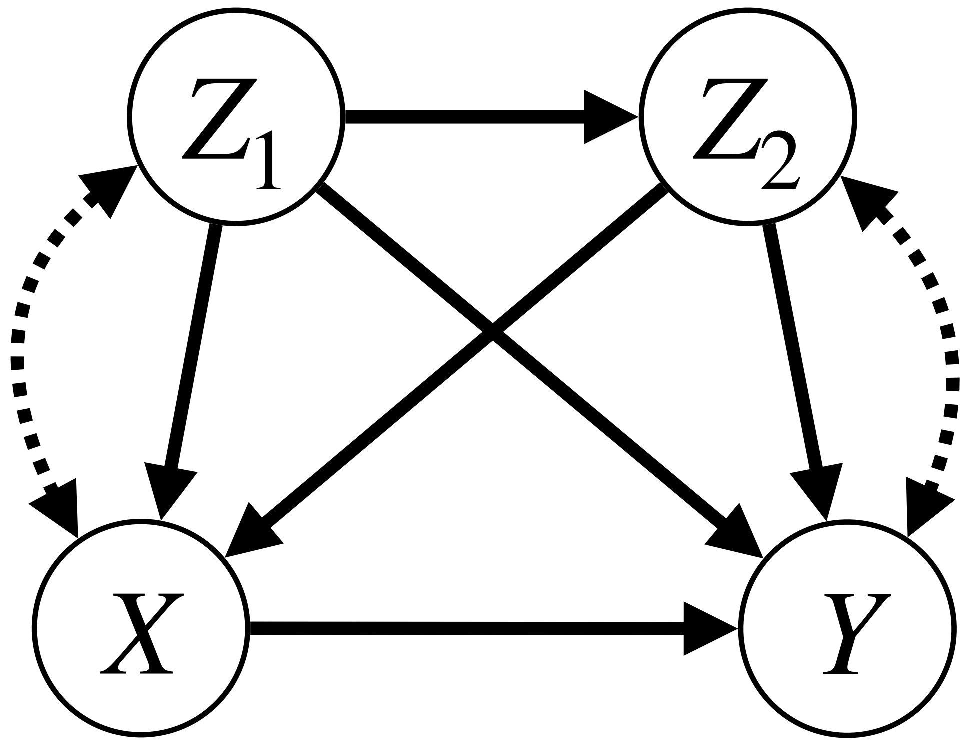


G_1



$P(y \mid do(x))$ is not identifiable

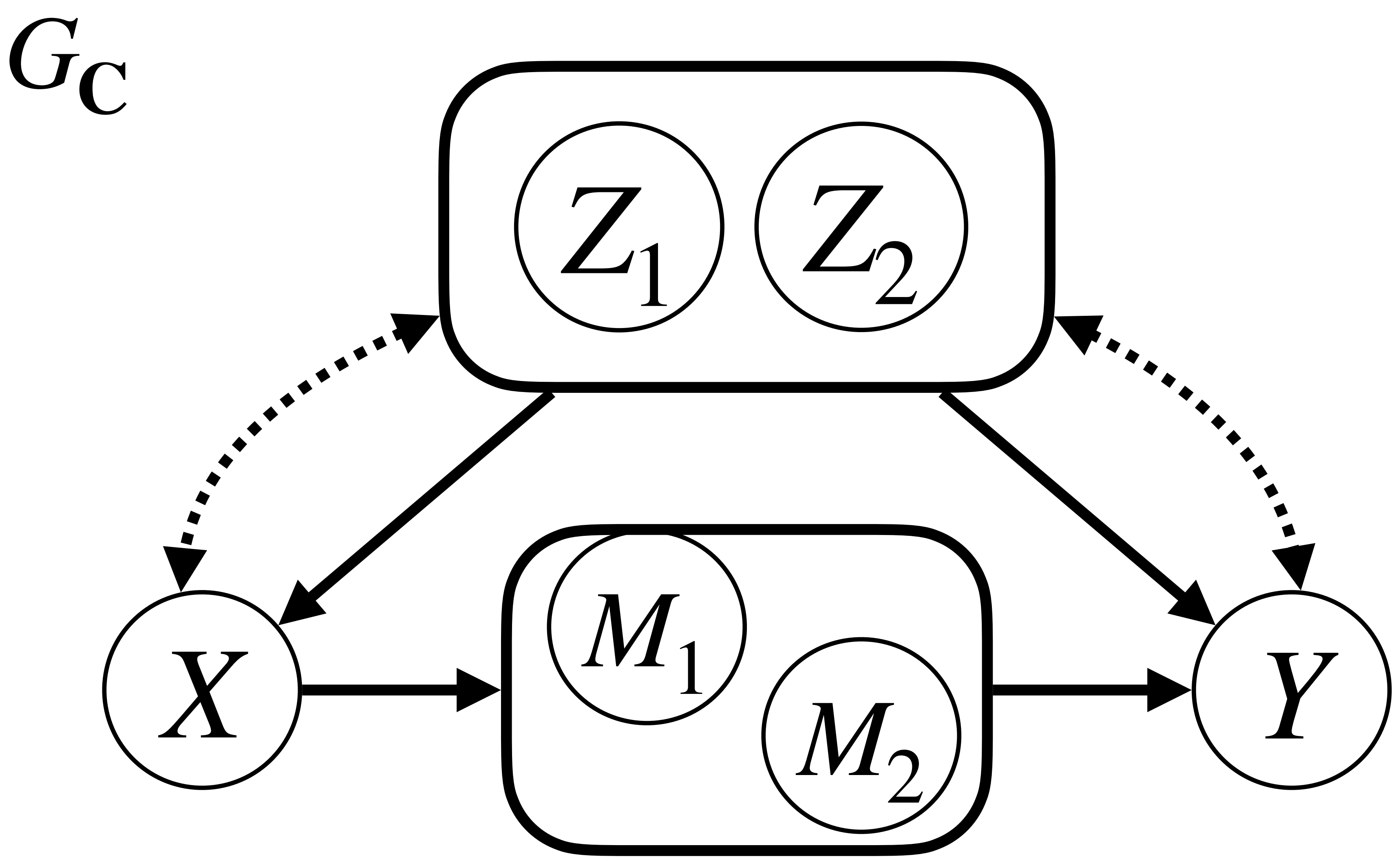
G_2



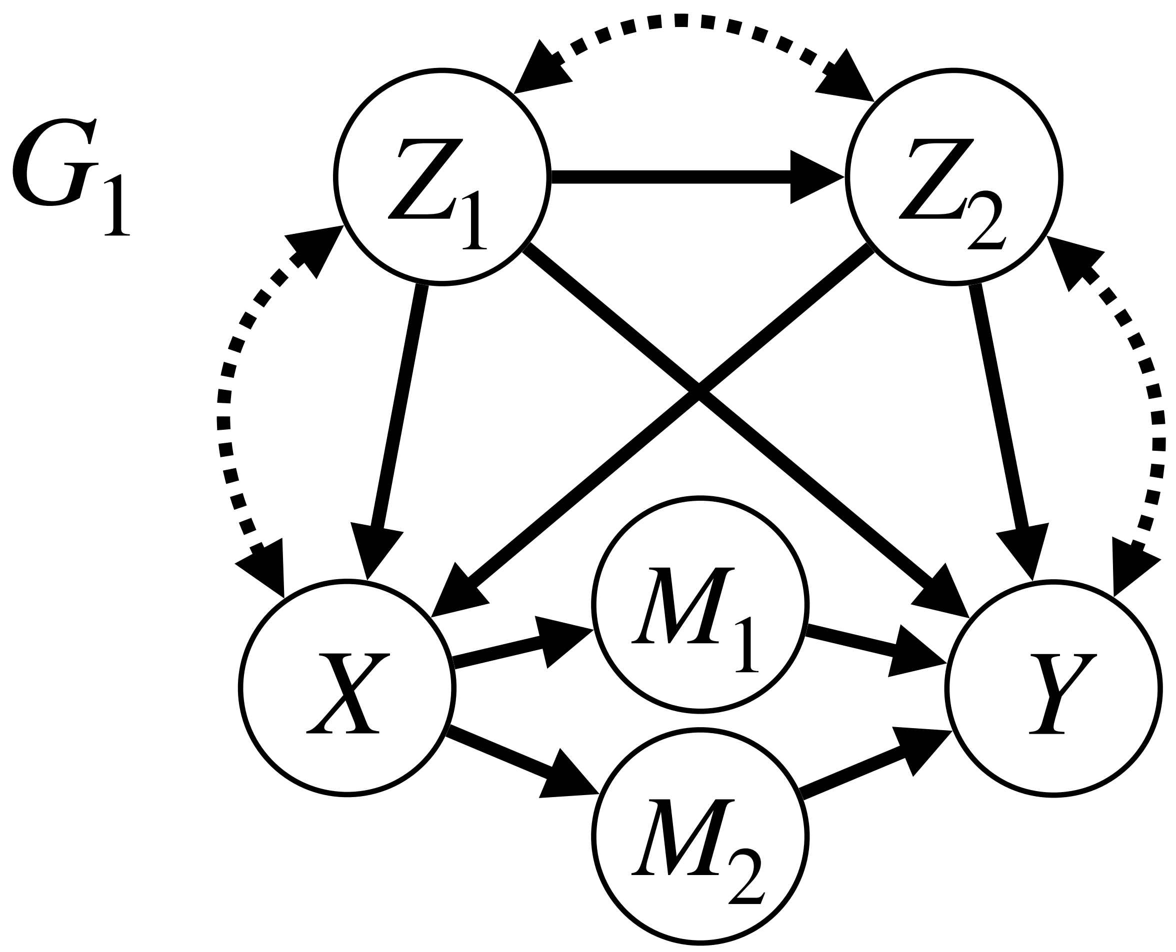
$$P(y \mid do(x)) = \sum_{z_1, z_2} P(y \mid x, z_1, z_2) P(z_1, z_2)$$

An effect is not identifiable in a C-DAG $G_{\mathbf{C}}$ if there exists at least one compatible causal diagrams G in which the effect is not identifiable.

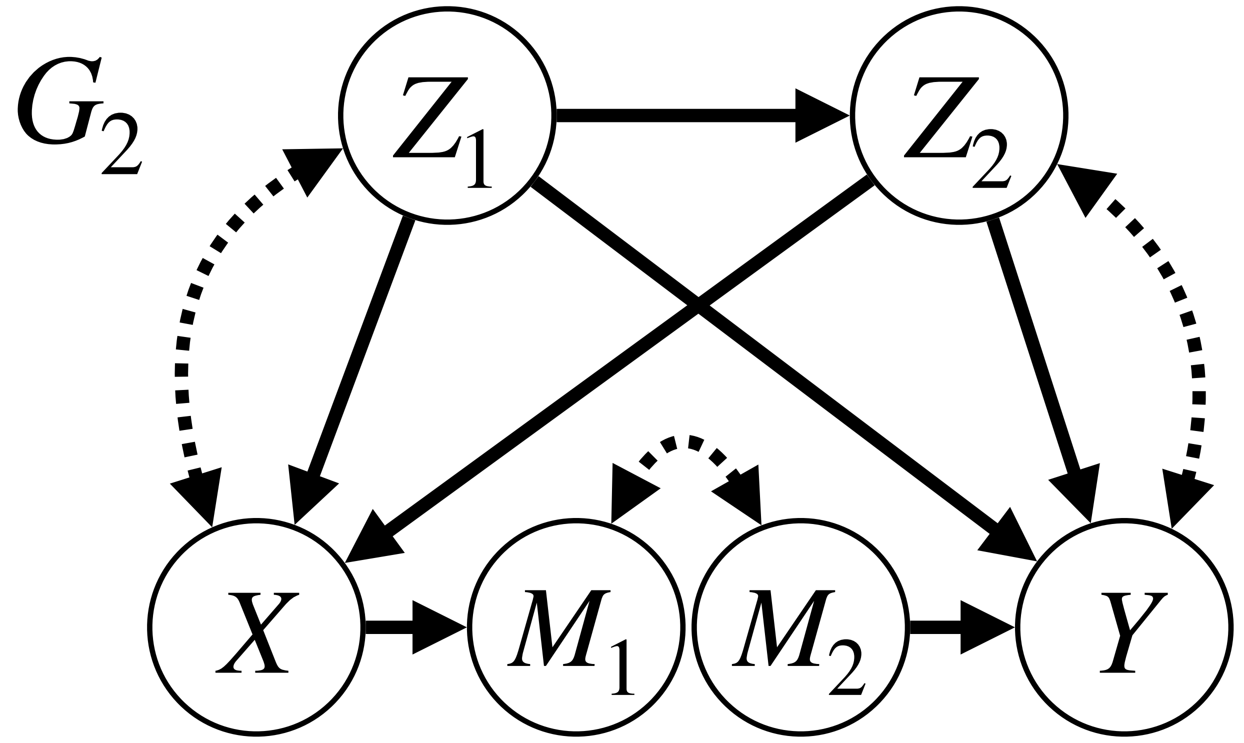
Beyond Backdoor Adjustment



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



$$P(y \mid do(x)) = \sum_{m_1,m_2} P(m_1, m_2 \mid x) \sum_{x'} P(y \mid m_1, m_2, x')P(x')$$



$$P(y \mid do(x)) = \sum_{m_1,m_2} P(m_1, m_2 \mid x) \sum_{x'} P(y \mid m_1, m_2, x')P(x')$$

Again, an effect identifiable in a C-DAG $G_{\mathbf{C}}$ is identifiable in all compatible causal diagrams G using the same identification formula!

What if no knowledge is available?

Can we learn a causal diagram $\mathscr{G}$ from observational data?

Causal Discovery:

In non-parametric settings, we can't learn the true causal diagram, but algorithms such as the Fast Causal Inference (FCI) can learn a graphical representation of its *Markov equivalence class*!

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

Causal Discovery

Fast Causal Inference (FCI)

A constraint-based causal discovery algorithms that
accounts for unobserved confounders

Causal Discovery

Goal: Learn a graphical representation of the Markov Equivalence Class from observational data.

Assumptions: the observed distribution is the marginal of a distribution P that satisfies the following conditions for the true causal diagram G (an **ADMG**):

1) **I-Map / Semi-Markov Condition:** for any disjoint subsets $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$:

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z})_G \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z})_P.$$

G is an ***I-Map of*** P

P is ***semi-Markov relative*** to G.

2) **Faithfulness Condition:** for any disjoint subsets $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$:

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z})_P \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z})_G.$$

P is ***faithful*** to G

Note: Estimation of the marginal distribution from limited data requires and **additional assumption**:

3) An adequate *conditional independence test* is available.

Fast Causal Inference (FCI) Algorithm

FCI: Learn a PAG $\mathscr{P}$ representing the **Markov Equivalence Class (MEC)** from P, i.e.:

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z})_{\mathscr{P}} \Leftrightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z})_{P;G}$$

Evaluated through m-separation

Every non-circle edge mark represents an invariance in the MEC in terms of ancestral and non-ancestral relationships

Arrowhead $\Longrightarrow$ **non-ancestrality**

Tail $\Longrightarrow$ **ancestrally**

Circle $\Longrightarrow$ **non-invariance**

$A \longrightarrow B \quad \Longrightarrow$ ancestrally

$A \circ\!\!\longrightarrow B \quad \Longrightarrow$ non-ancestrality

$A \longleftrightarrow B \quad \Longrightarrow$ spurious association

$A \vdash\!\!\!-\!\!\!-\!\!\!\dashv B \quad \Longrightarrow$ selection bias

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

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

Conditional Independence Tests

Gaussian errors and independent observations: 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/

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

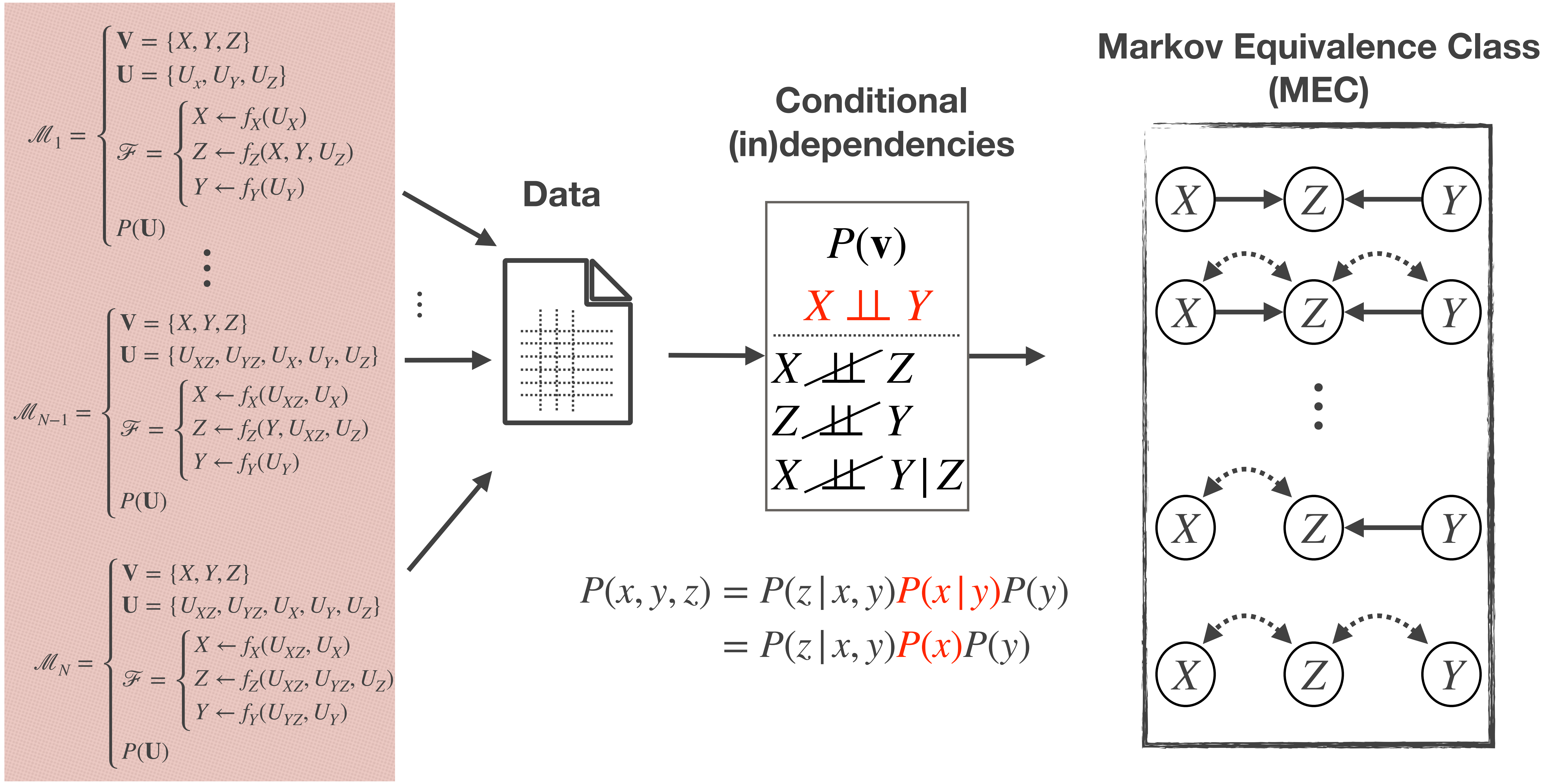
Continuous (conditional Gaussian) or Discrete (Binary, Ordinal, Multinomial) - Linear Regression

- 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/

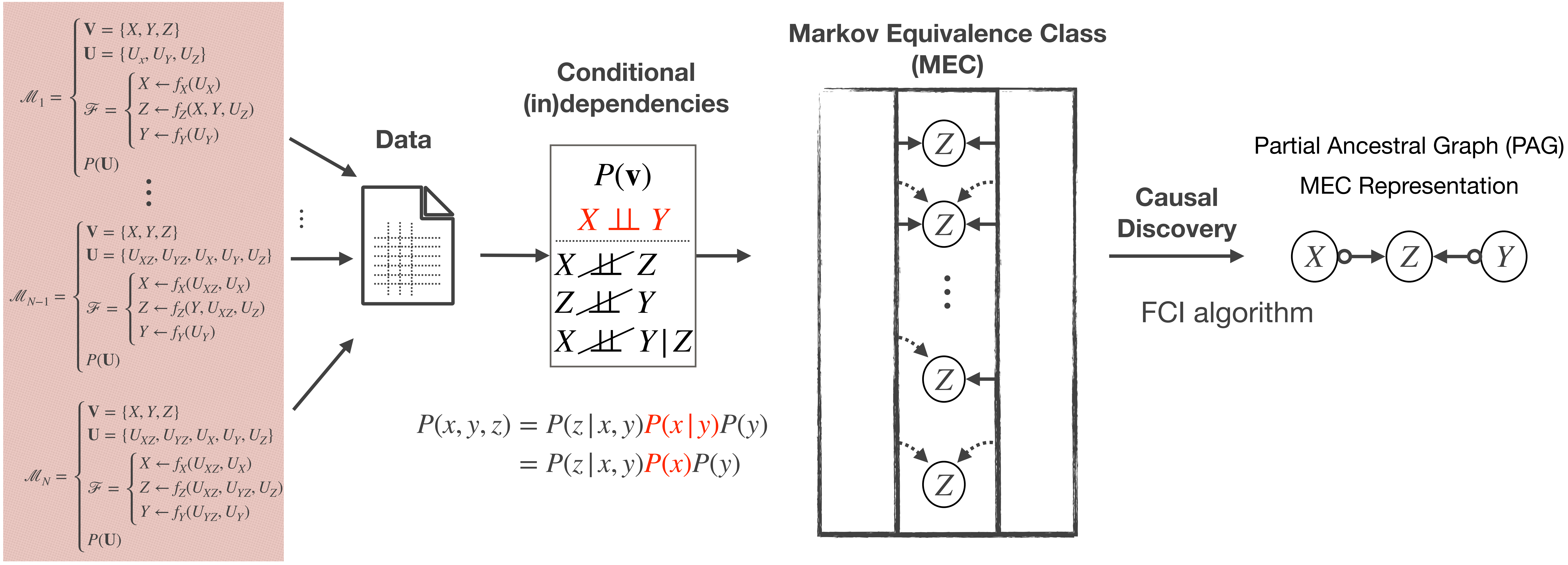
Gaussian errors and correlated observations (family data) :

Ribeiro A.H.**,** Soler J.M.P. (2020). *Learning Genetic and environmental graphical models from family data*, Statistics in Medicine.
R package: https://github.com/adele/FamilyBasedPGMs

Learning Structural Invariances

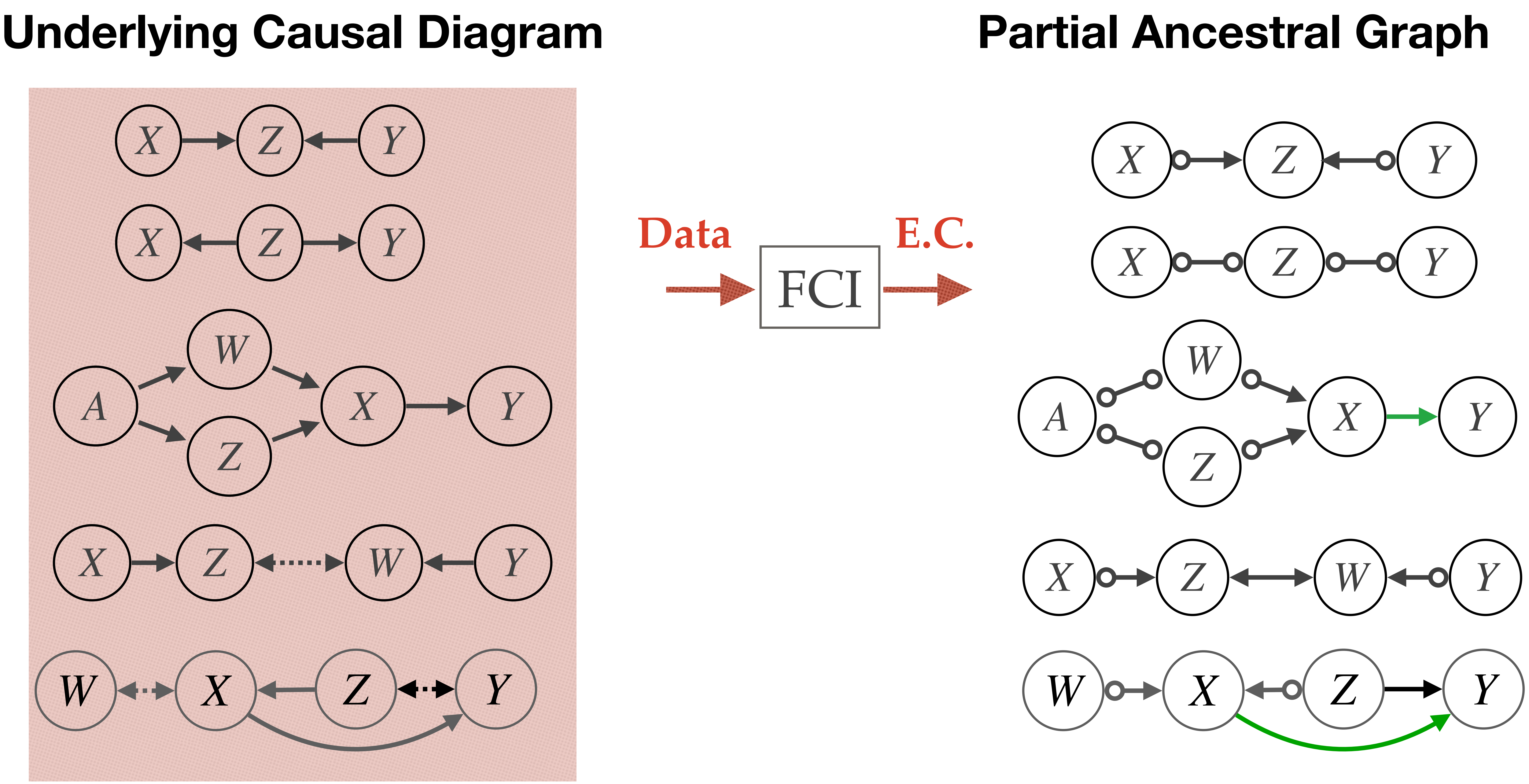


Learning Structural Invariances



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

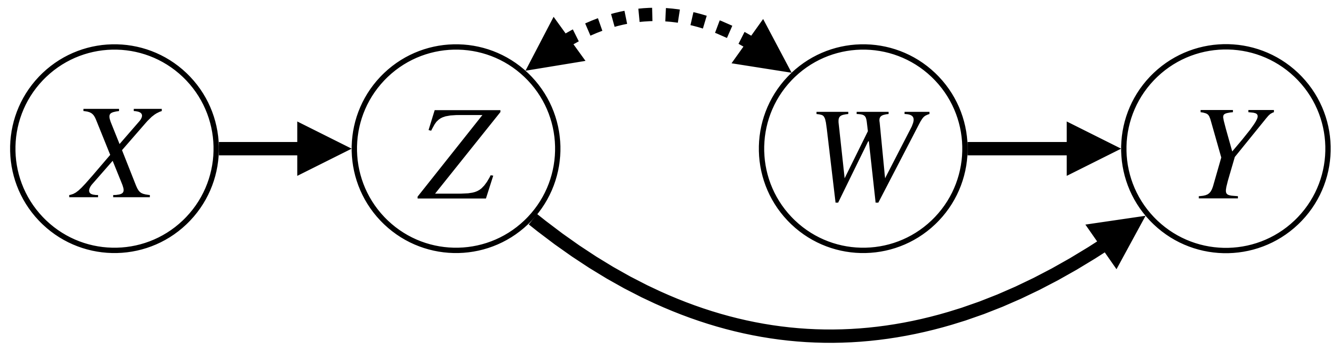
Other examples



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]

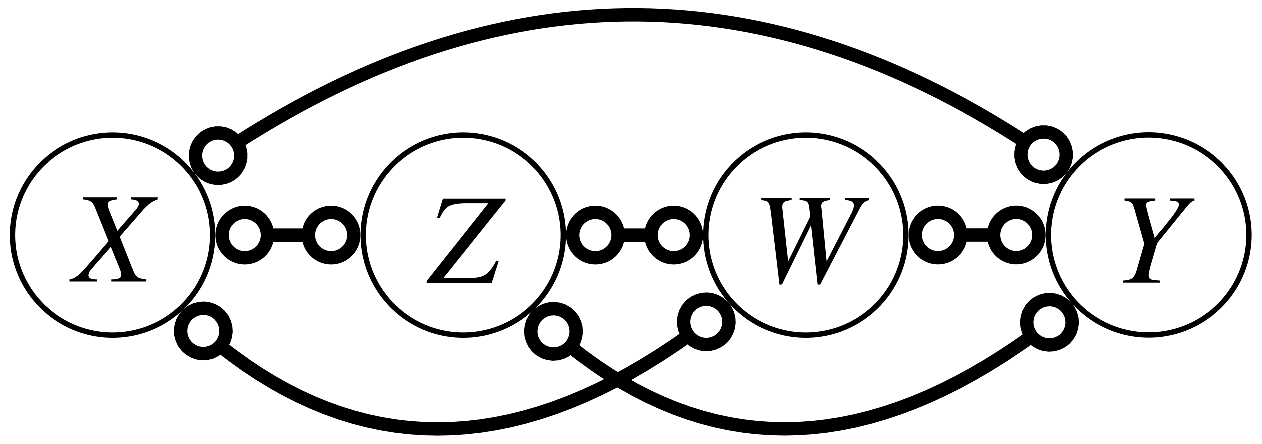
Fast Causal Inference (FCI) Algorithm

True (unknown) causal diagram



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

Implied by the ADMG using d-separation

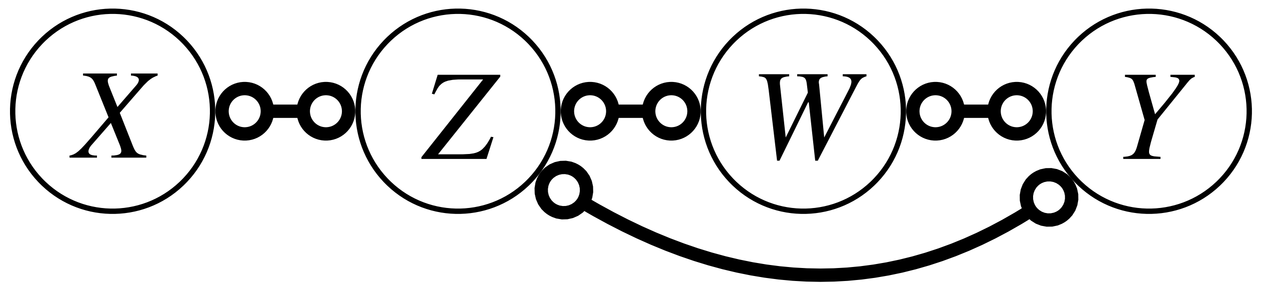


Complete Graph

Conditional Independence Tests

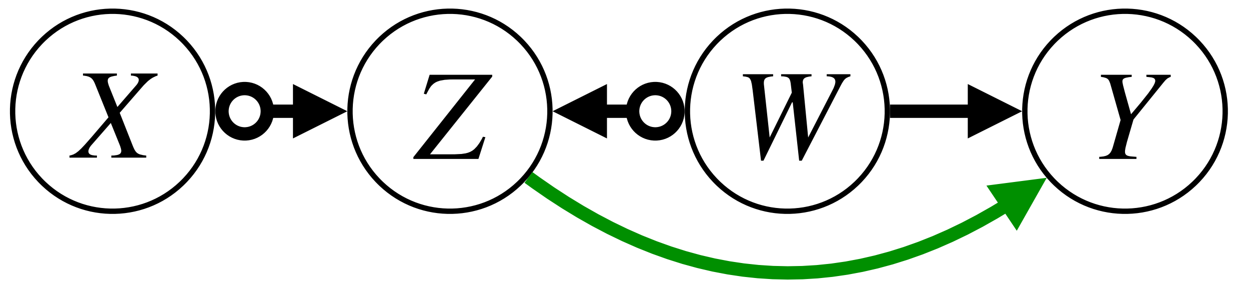
$X \perp\!\!\!\perp W$
$X \perp\!\!\!\perp Y \mid Z, W$

observed in the data



Skeleton

FCI Rules
$(R1) - (R10)$



Partial Ancestral Graph (PAG)

Implied by the PAG using m-separation

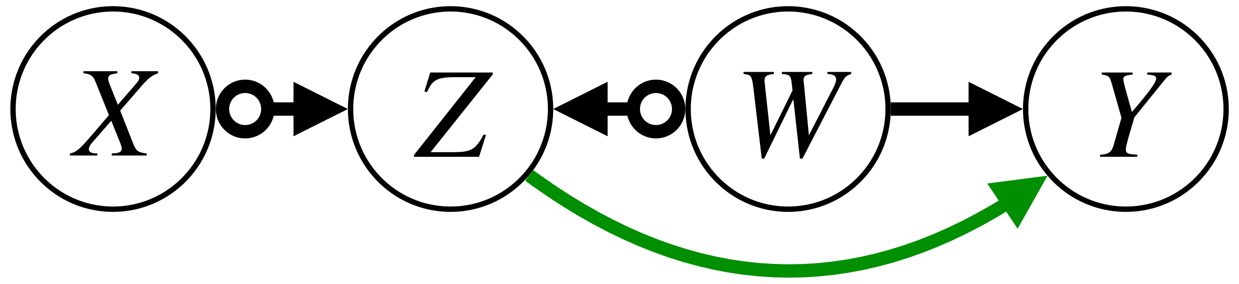
$X \perp\!\!\!\perp W$
$X \perp\!\!\!\perp Y \mid Z, W$

Z is not an ancestor of X or W.

Z and W are ancestors of Y.

Z is not confounded with Y.

PAG represents the Markov Equivalence Class

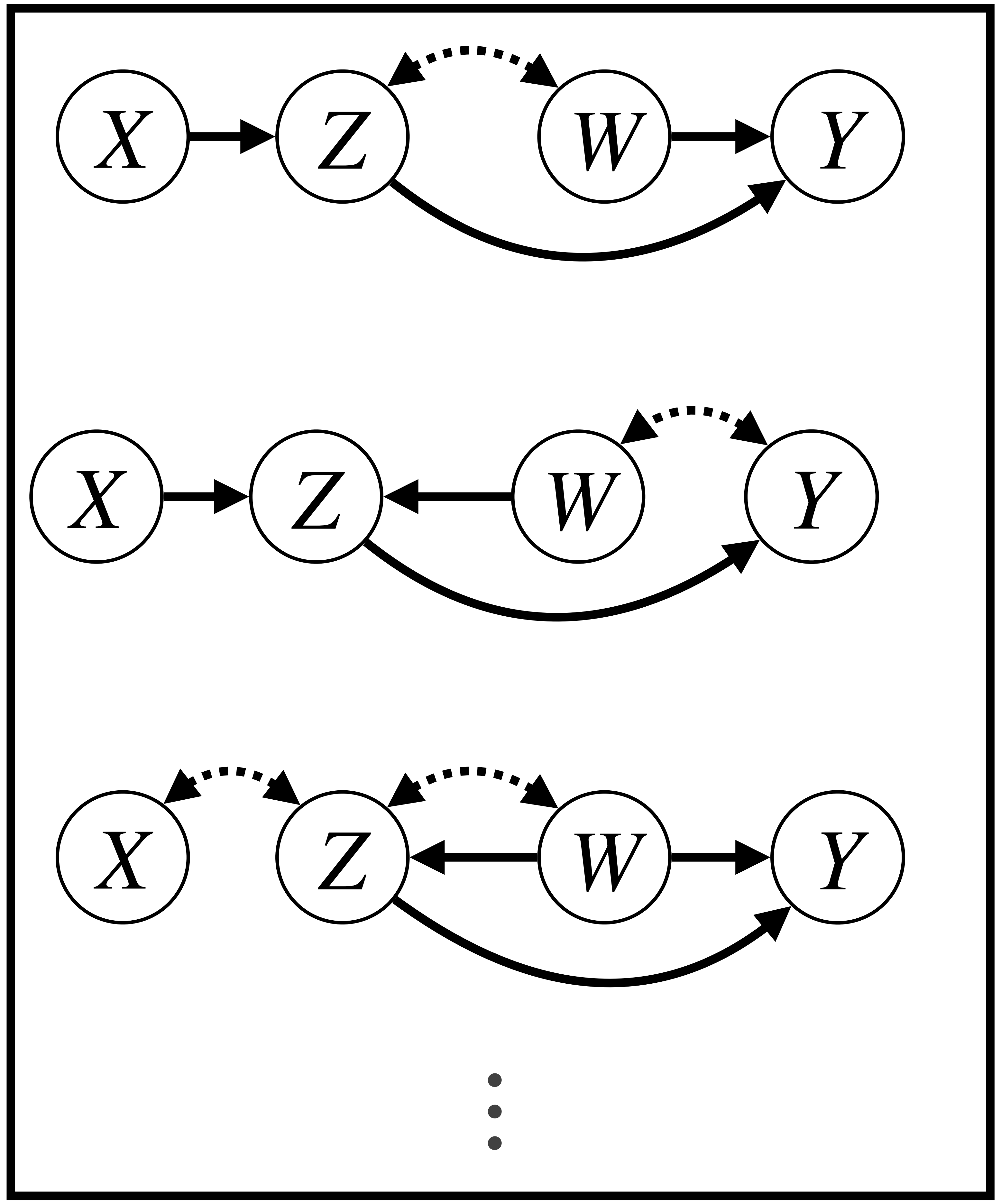


Partial Ancestral Graph
(PAG)

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

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

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

More on Causal Discovery

Causal discovery from observational and experimental data:

- Gonçalo Rui Alves Faria, Andre Martins, Mario A. T. Figueiredo. Differentiable Causal Discovery Under Latent Interventions. Proceedings of the First Conference on Causal Learning and Reasoning, PMLR 177:253-274, 2022.

- Kocaoglu, M., Jaber, A., Shanmugam, K., Bareinboim, E. Characterization and Learning of Causal Graphs with Latent Variables from Soft Interventions. In Proceedings of the 33rd Annual Conference on Neural Information Processing Systems. 2019.

- Jaber, A., Kocaoglu, M., Shanmugam, K., Bareinboim, E. Causal Discovery from Soft Interventions with Unknown Targets: Characterization & Learning. In Advances in Neural Information Processing Systems 2020.

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)

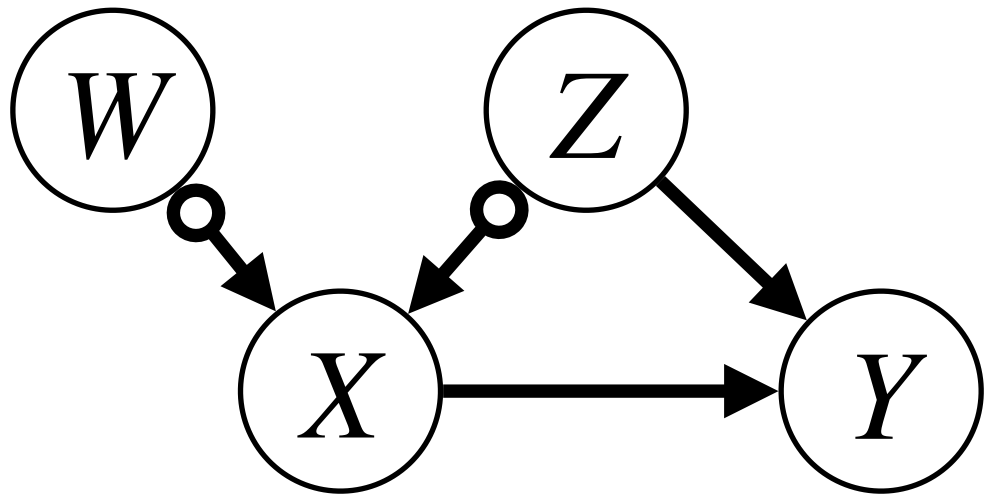
General Identification in Markov Equivalence Classes

The CIDP and IDP algorithms are available at the *PAGId* R package: https://github.com/adele/PAGId

1 **Query**

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

2 **PAG**



3 **Data**

$P(x, m, y)$ **Observational Distribution**

IDP / CIDP

Solution

yes / no

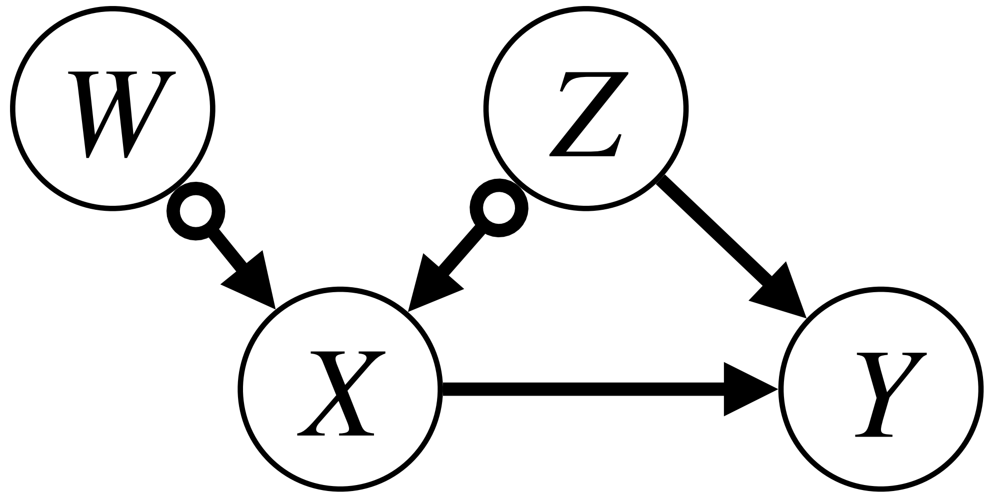
$$P(y \mid do(x)) = \sum_z P(y \mid x, z)P(z)$$

Interventional Distribution ← Available Distributions 👍

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

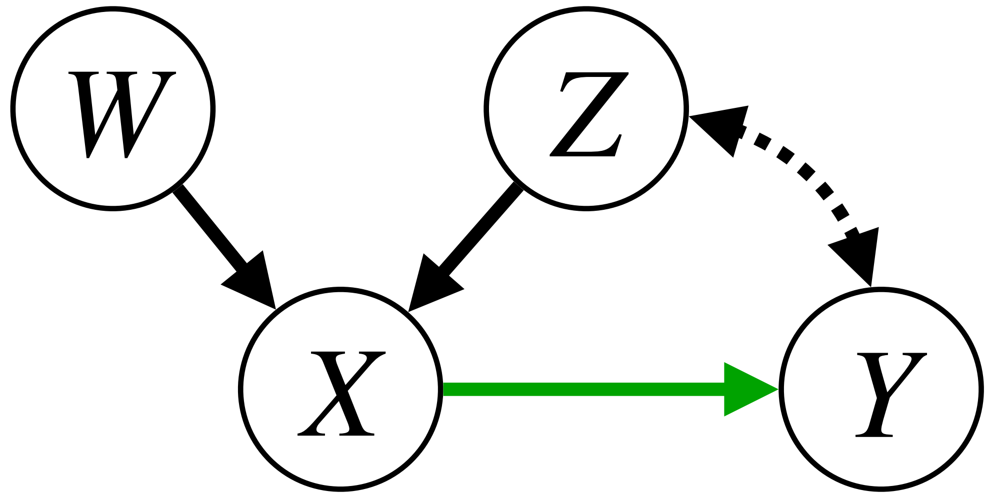
Effect Identifiabiliy given a PAG

$\mathscr{P}$



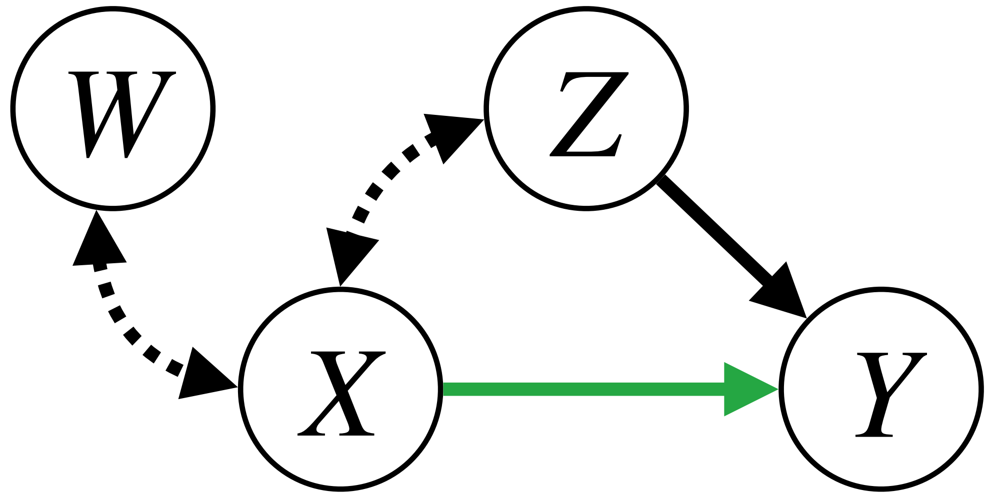
$$P(y \mid do(x)) = \sum_z P\left(y \mid x, z\right) P\left(z\right)$$

G_1



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

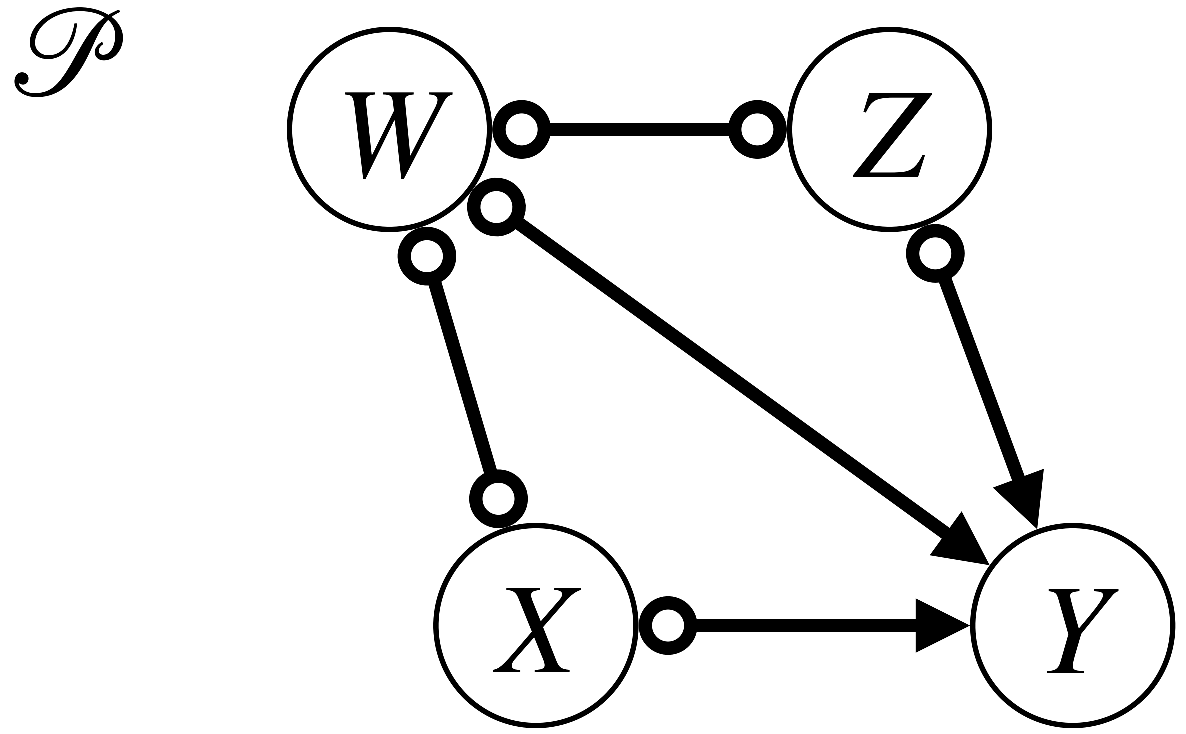
G_2



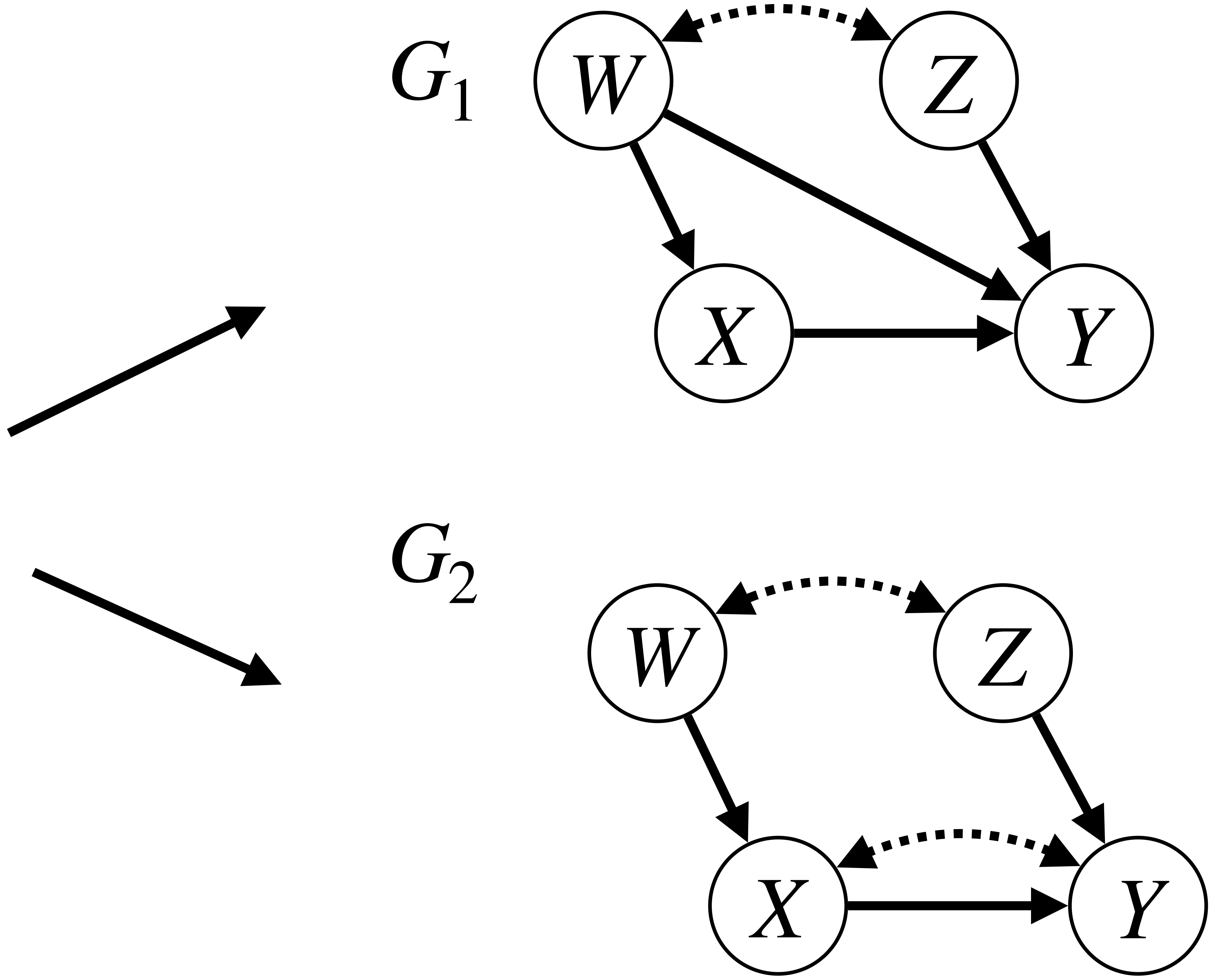
$$P(y \mid do(x)) = \sum_z P(y \mid x, z) P(z)$$

An effect identifiable in a PAG $\mathscr{P}$ is identifiable in all causal diagrams G in the Markov Equivalence Class using the same identification formula!

Effect Non-Identifiabiliy given a PAG



$P(y \mid do(x))$ is
not identifiable



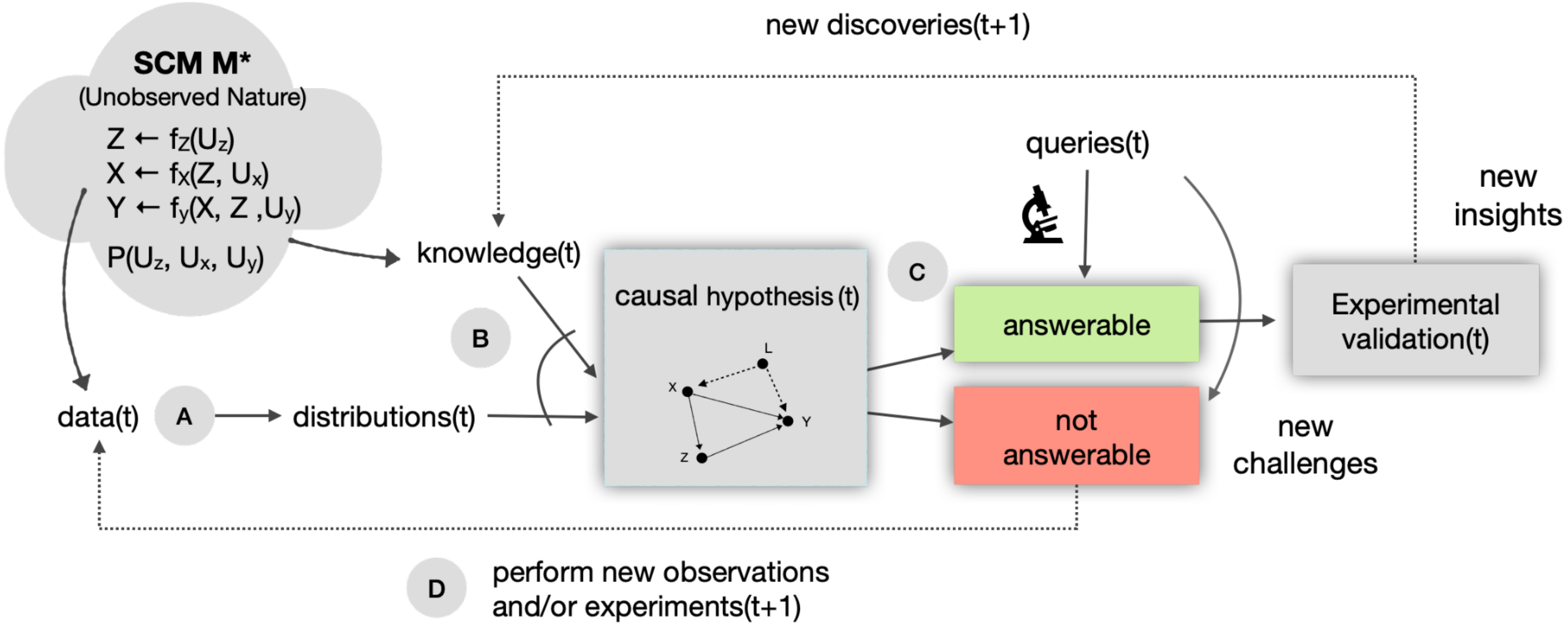
$$P(y \mid do(x)) = \sum_z P(y \mid x, z) P(z)$$

$P(y \mid do(x))$ is
not identifiable

An effect not identifiable in a PAG $\mathscr{P}$ is not identifiable in at least one causal diagrams G in the Markov Equivalence Class

Causal Inference Workflow

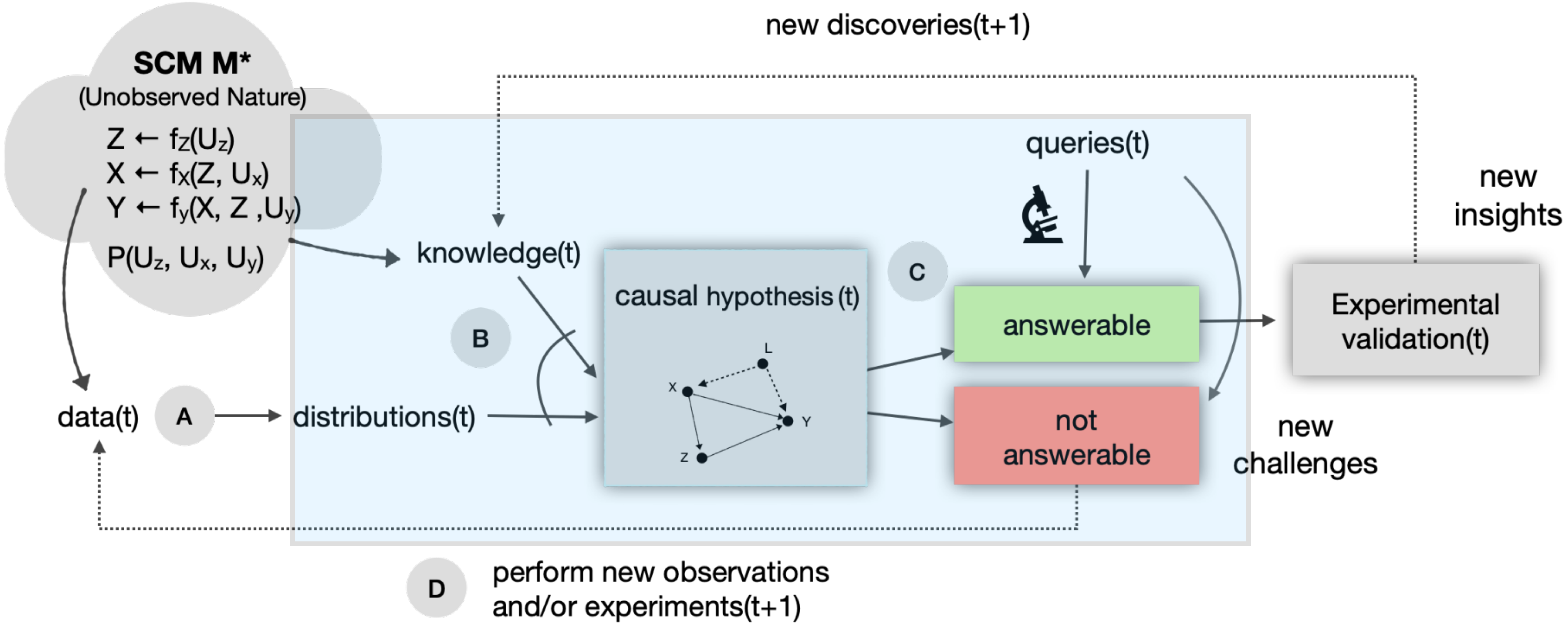
Continuous Process of Scientific Discovery and Causal Hypothesis Refinement



A Statistical Learning B Causal Learning C Causal Inference D Causal Exp. Design

Causal Inference Workflow

Continuous Process of Scientific Discovery and Causal Hypothesis Refinement



A Statistical Learning | B Causal Learning | C Causal Inference | D Causal Exp. Design

There is much more to learn…

1. Generalized Effect Identification:
   - For when multiple observational and experimental datasets are available, possible under partial observability.
2. Partial Identification:
   - For when the effect is not point-identifiable, but an interval for it can be derived.
3. Effect Transportability:
   - For when the target is a different population/domain.
4. Counterfactual Identification:
   - Identification of $\mathscr{L}_3$ quantities, such as $P(\mathbf{y}_{\mathbf{x}} \,|\, \mathbf{x}')$ and $P(\mathbf{y}_{\mathbf{x}} \,|\, \mathbf{x}', \mathbf{z})$.
5. Fairness Evaluation:
   - To identify path-specify effects related to protective variables.
6. Effect Estimation beyond backdoor scenarios:
   - Via doubly robust machine learning and different plug-in density estimators.

Conclusions

Causal inference can help overcome critical challenges in Artificial Intelligence, including robustness, generalizability, explainability, and fairness.

Causal Data Science: principled way of combining data and substantive knowledge about the phenomenon under investigation to generate causal explanations and better decision-making.

Recent developments for causal inference when knowledge is largely unavailable and coarse are expected to help the practice of causal data analysis and meet the growing demand in the Empirical Sciences for sound causal explanations and more robust and generalizable decision-making.

Thank you! :)

Feel free to reach out to me if you have any questions:

adele.ribeiro@uni-marburg.de