# Introduction to Causal Inference

**Adèle H. Ribeiro**

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

Data Science in Biomedicine
Faculty of Mathematics and Computer Science
Philipps-Universität Marburg

**Tropical Probabilistic AI School**
**February 2nd, 2024**

# Current Challenges in AI

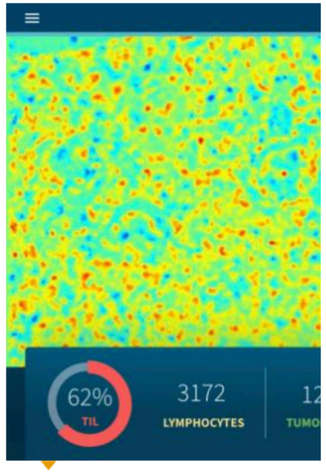
itn
IMAGING TECHNOLOGY NEWS

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

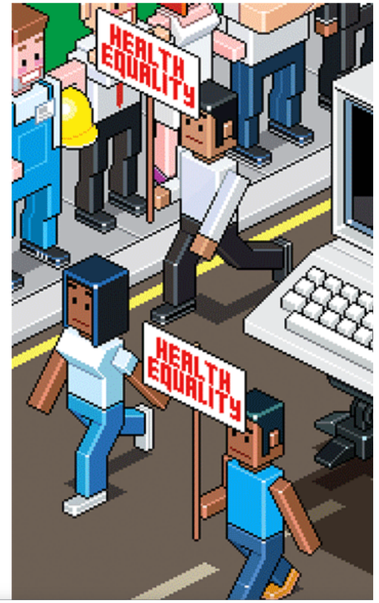
Explore content | Journal information | Publish with us | Subscribe

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

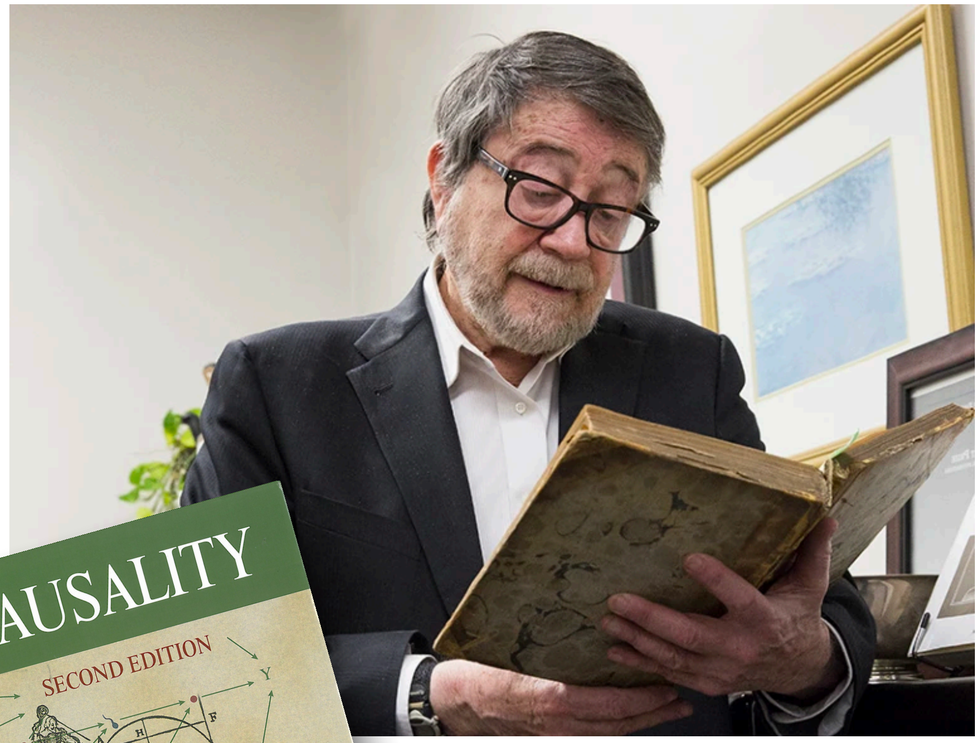
nature > outlook > article

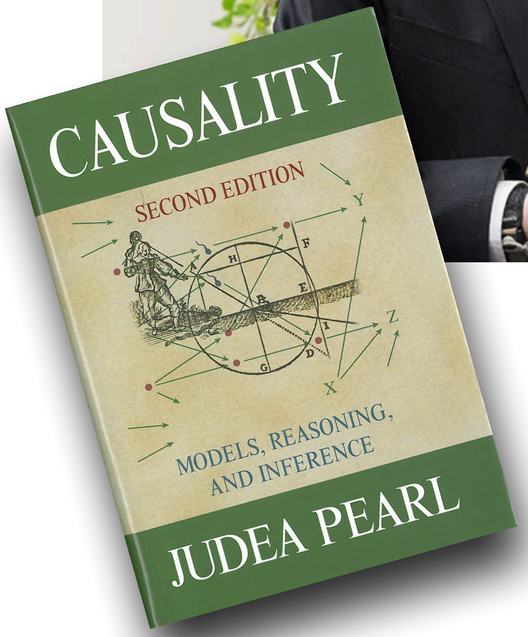
OUTLOOK | 24 February 2023

## 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.

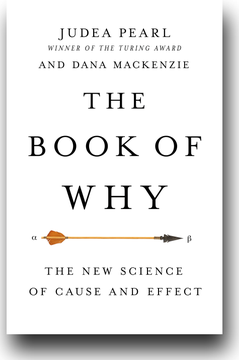
In 2011, he won the A. M. Turing Award (the highest distinction in computer science and a $250,000 prize)

"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



# Guido W. Imbens & Joshua D. Angrist

Guido W. Imbens

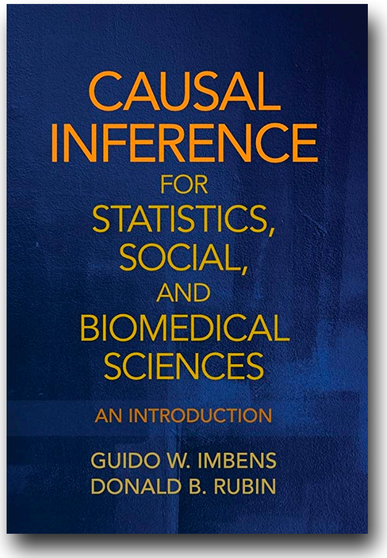
Professor of Applied Econometrics in Stanford University

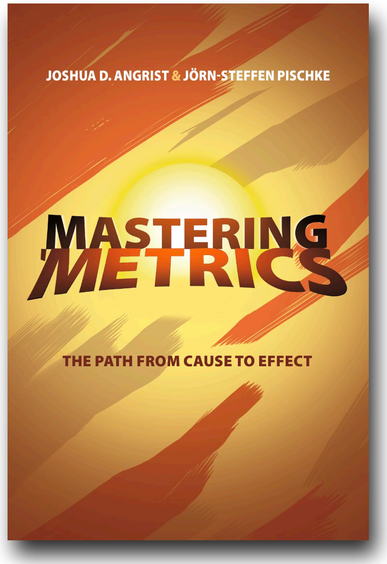
Joshua D. Angrist

Professor of Economics at the Massachusetts Institute of Technology

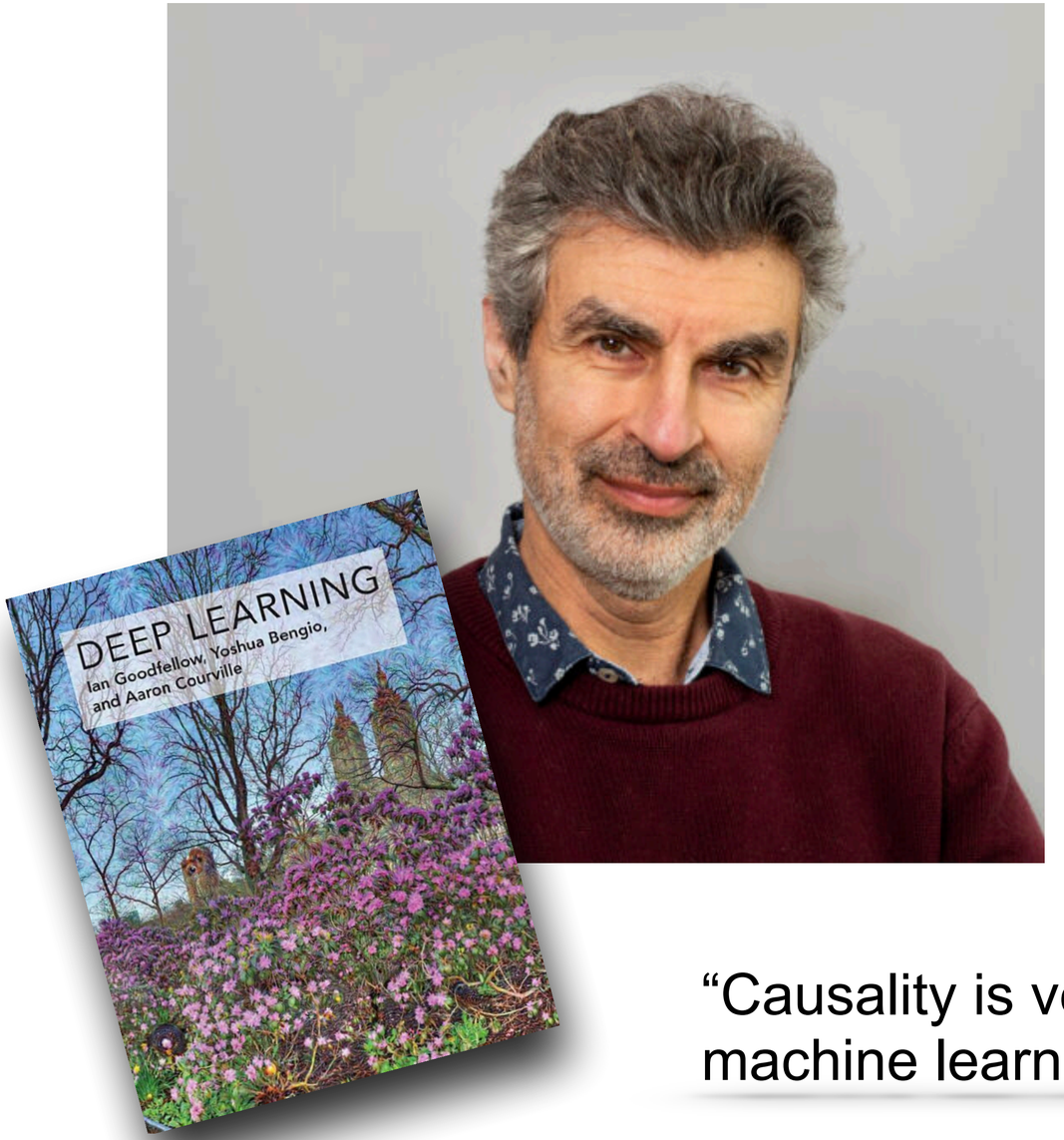
In 2021, they won the Nobel Prize in Economics (about $1 million)

"for their methodological contributions to the analysis of causal relationships"



# 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*.

# Why causality is so important?

Causality allows important capabilities such as

**Explainability:** provides a better understanding of the underlying mechanisms

- **Causal Discovery**

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

- **Causal Effect Identification and Estimation**

**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

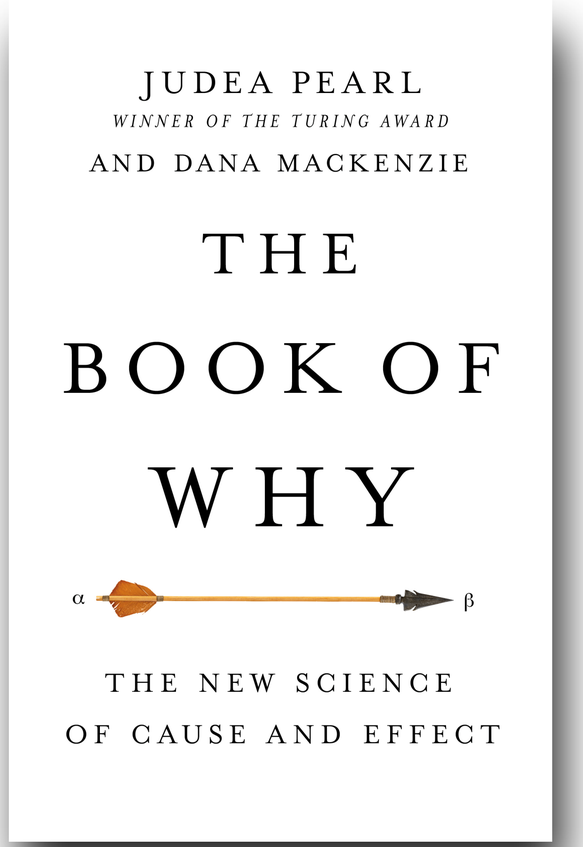
**Elias Bareinboim[a,b,1] and Judea Pearl[a]**

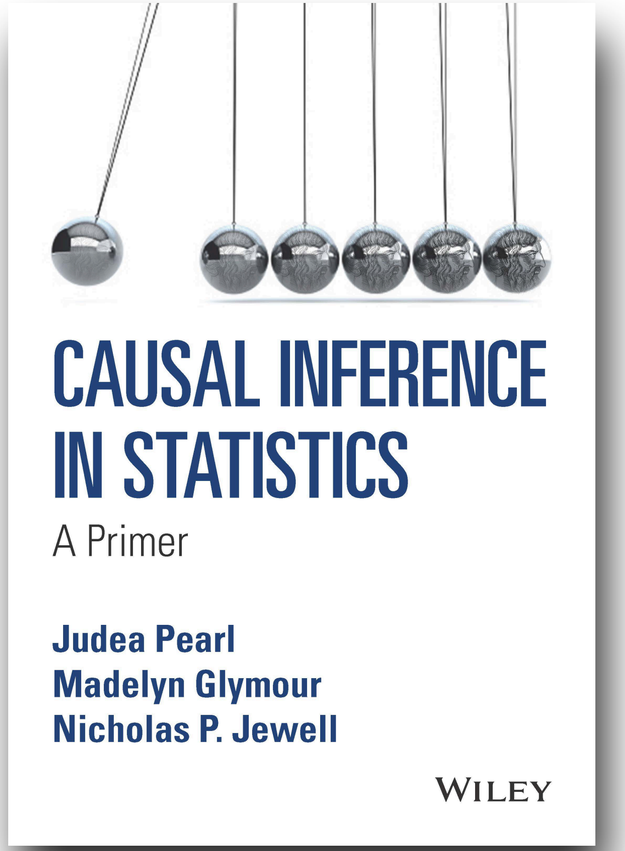
[a]Department of Computer Science, University of California, Los Angeles, CA 90095; and [b]Department of Computer Science, Purdue University, West Lafayette, IN 47907

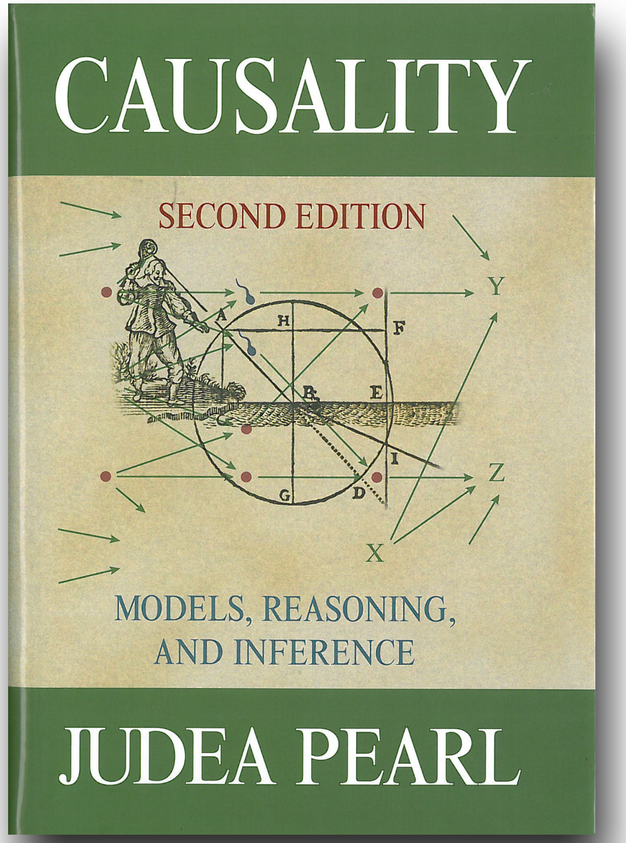
Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved March 15, 2016 (received for review June 29, 2015)

http://causalfusion.net

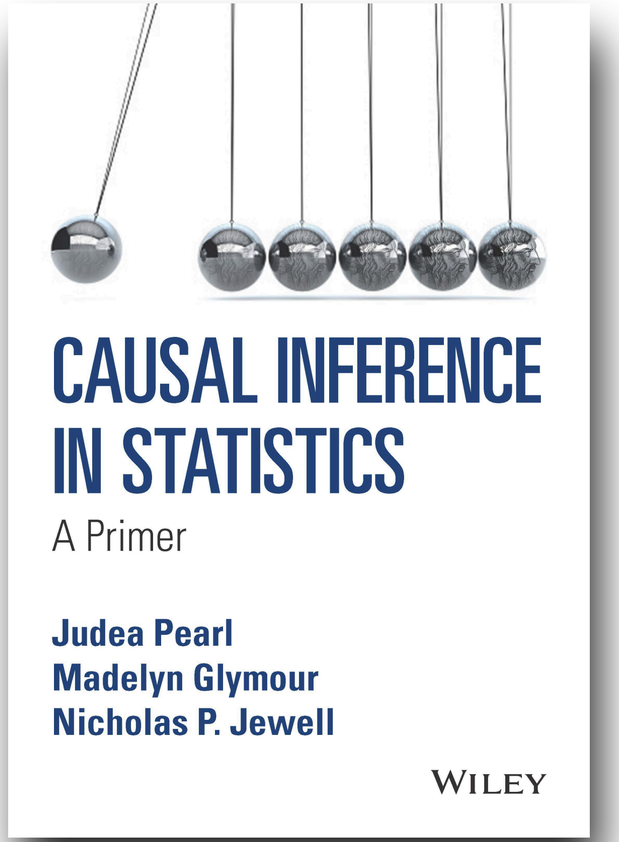
# Causality Theory by Judea Pearl



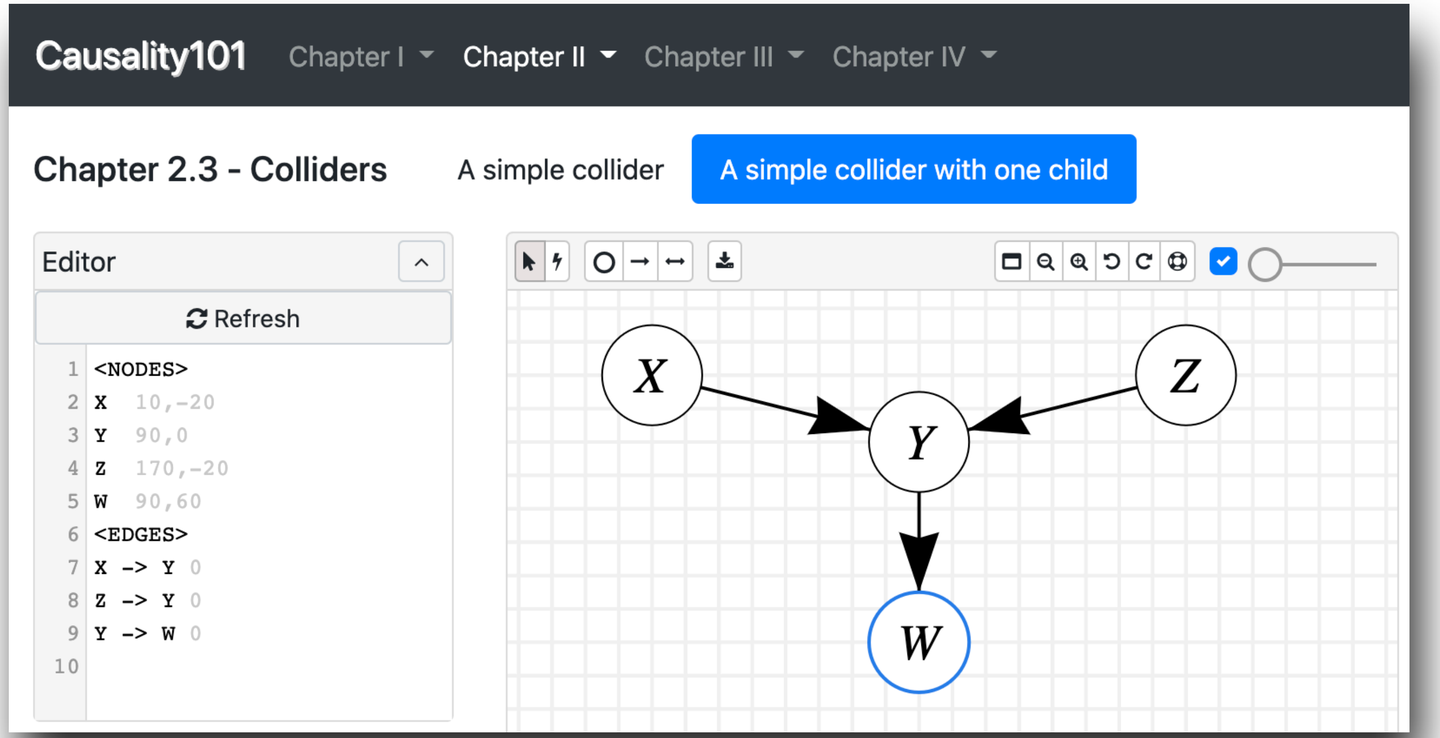



# Causality Theory by Judea Pearl




https://causality101.net/



# Prediction vs Effect of Interventions

## Statistical Association vs Causation

# Predictive Tasks

**Task:** Can I guess how severe is a fire by **observing** the number of firefighters?

**Yes!**

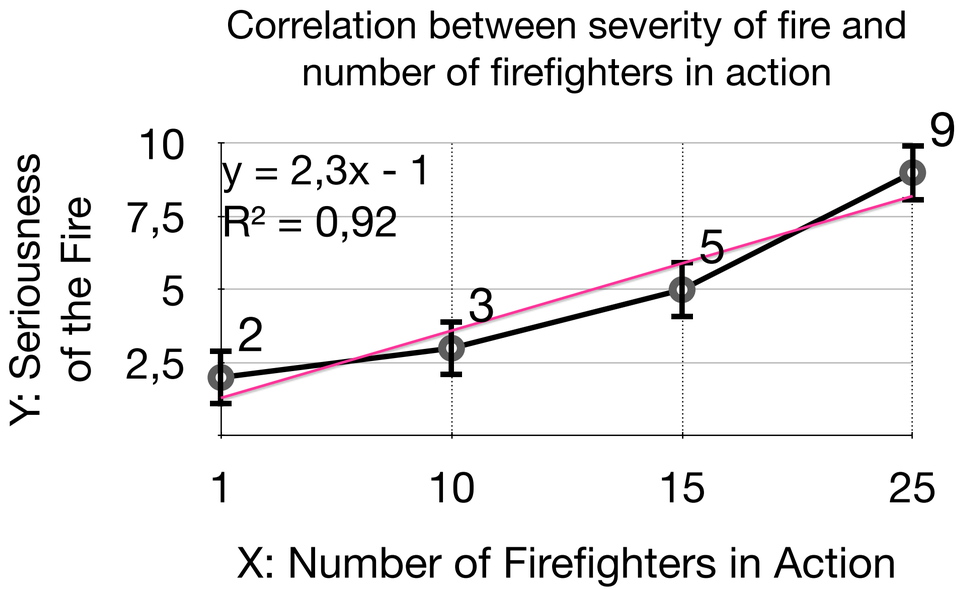
$X$: **Number of firefighters in action**

$Y$: **Severity of the (initial) fire**

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

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

**Observational Probability Distribution**




**Conclusion:** The severity of the fire increases with the number of firefighters.

# Prediction ⇒ Decision-Making?

**Conclusion:** The severity of the fire increases with the number of firefighters.

**The fewer firefighters, the weaker the fire.**

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

# Causal Effect $\neq$ Statistical Association

The **causal effect** of $X$ on $Y$ is a quantity that tells us how much $Y$ changes after an intervention $do(X = x)$, e.g., $E[Y \mid do(X = x)]$.

If a different number of firefighters were dispatched, $do(X = x)$, would this change the expected severity of the initial fire $Y$ ?

**Causal Effect $\equiv$ Effect of an Intervention**

This quantity is derived from the probability of $Y$ obtained after making the intervention $do(X = x)$, which is denoted by $P(Y \mid do(X = x))$ and is commonly referred as ***Interventional Distribution***.

# Causal Effect ≡ Effect of an Intervention

This question can also be answered if knowledge about the underlying reality is available!

$X$: **Number of firefighters in action**

$Y$: **(Initial) Severity of the fire**

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

**Underlying Structural Causal Model (SCM)**

$Y$ is not a function of $X$

In other words, $X$ **is not a cause of** $Y$

In this case, $\forall x, E[Y \mid do(X = x)] = E[Y]$.

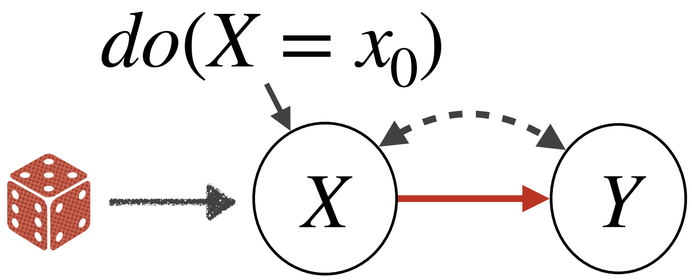
Changing the number of firefighters does not change the (initial) severity of the fire.

The action/intervention on $X$, $do(X = x)$ is independent of $Y$,
i.e., $P(Y = y \mid do(X = x)) = P(Y = y)$

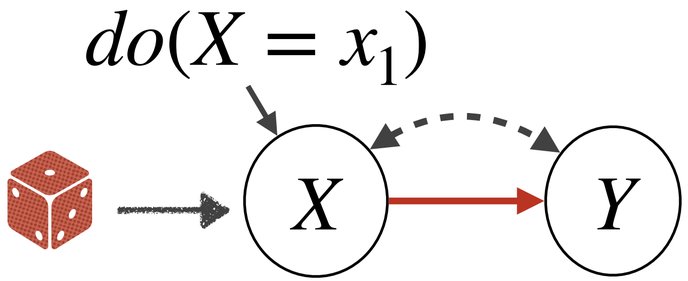
# Randomized Experiments

One way to access the interventional distribution $P(Y \mid do(X = x))$ is through a *perfectly realized* Randomized Experiments / Control Trials (e.g. RCT):

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

What is the causal effect of the number of firefighters $X$ on the severity of the initial fire $Y$ ?

# Pearl’s Causal Hierarchy (PCH)

## The Three Inferential Layers

# What is induced by the SCM?

**Observational SCM**

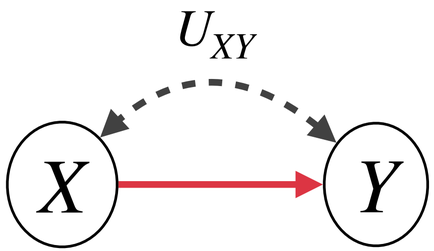
$$\mathcal{M} = \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_{XY}, U_X, U_Y\} \\ \mathcal{F} = \begin{cases} X = f_X(U_X, U_{XY}) \\ Y = f_Y(X, U_Y, U_{XY}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

$do(X = x)$

**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



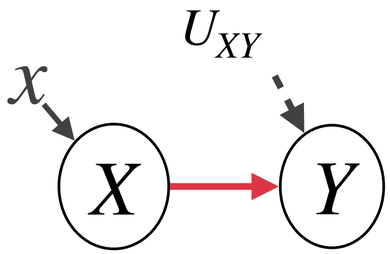

Observational Causal Diagram

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

Observational Distribution

Observational Data

Loss of Information

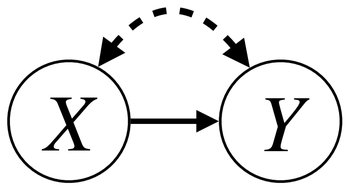
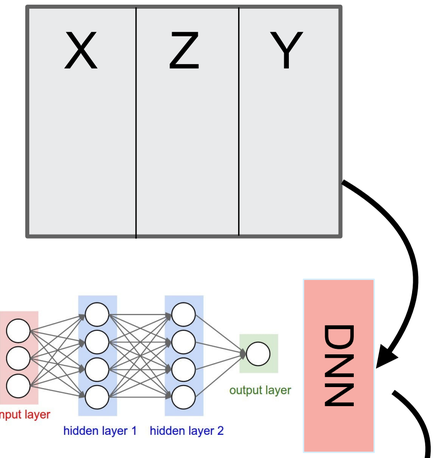
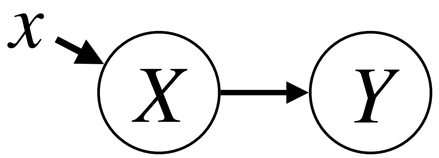



Interventional Causal Diagram

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

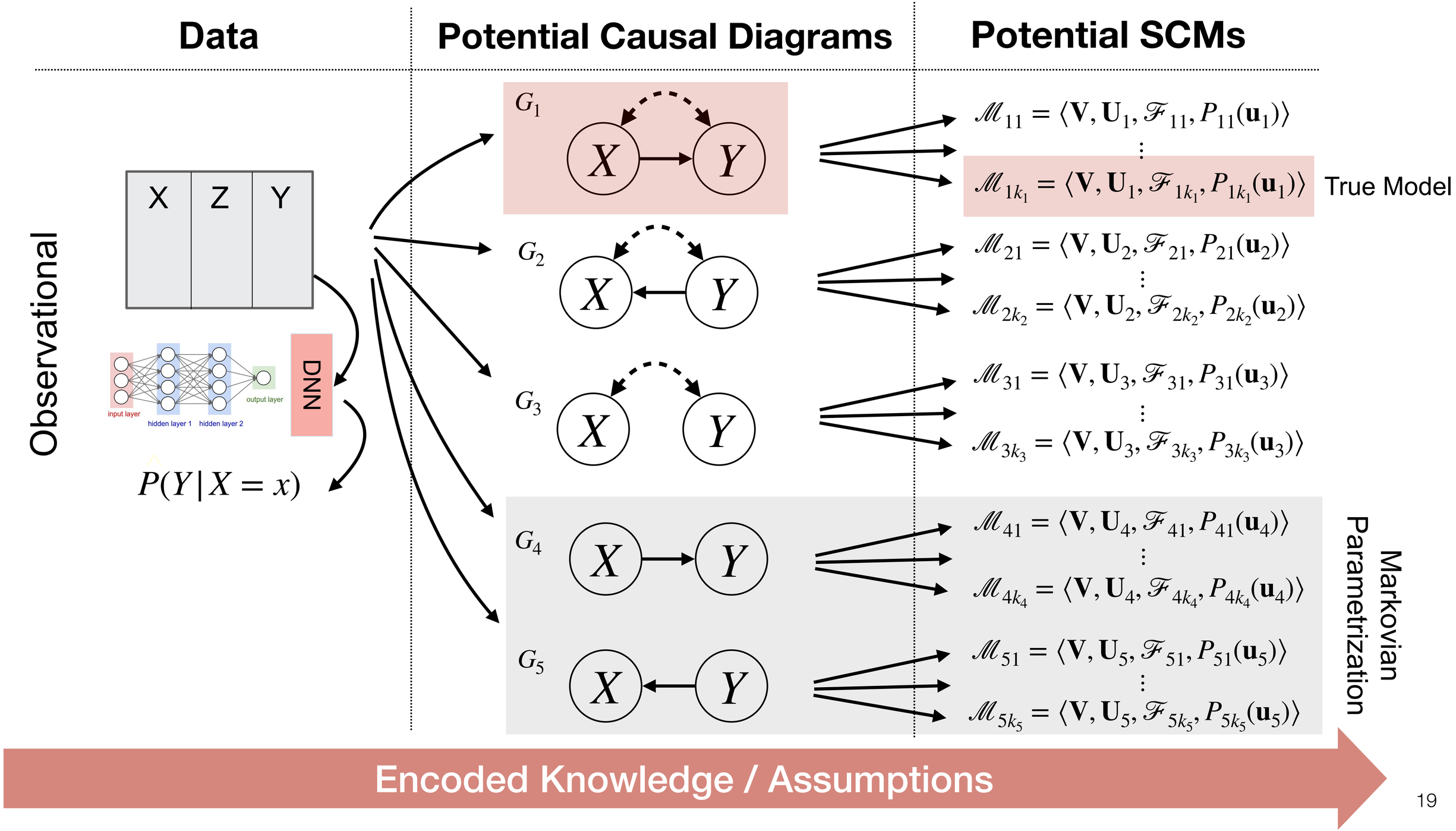
Interventional Distribution

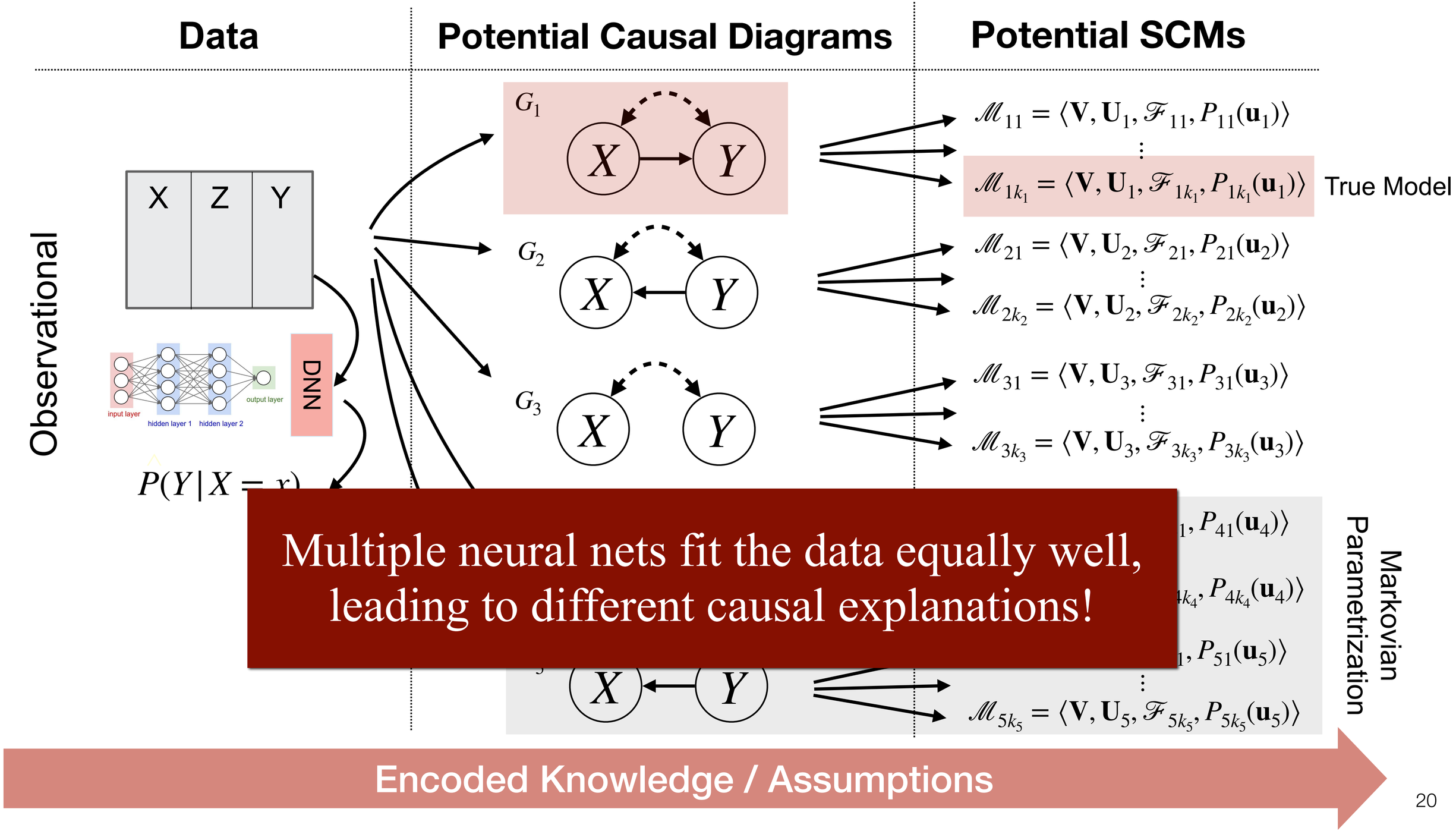
Interventional Data

| | Reality | Structural Knowledge | Data |
|---|---|---|---|
| Observational | Structural Causal Model (SCM) $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$ <br> $\mathcal{M} = \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_{XY}, U_X, U_Y\} \\ \mathcal{F} = \begin{cases} \boxed{X \leftarrow f_X(U_X, U_{XY})} \\ Y \leftarrow f_Y(X, U_Y, U_{XY}) \end{cases} \\ P(\mathbf{U}) \end{cases}$ | Causal Diagram <br> $G$ <br>   |   <br> $\hat{P}(Y \mid X = x)$ |
| Interventional | $\mathcal{M}_x = \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_{XY}, U_X, U_Y\} \\ \mathcal{F} = \begin{cases} \boxed{X \leftarrow x} \\ Y \leftarrow f_Y(x, U_Y, U_{XY}) \end{cases} \\ P(\mathbf{U}) \end{cases}$ | $G_{\overline{X}}$ <br>   | $\varnothing$ <br> $\hat{P}(Y \mid do(X = x)) = ?$ |

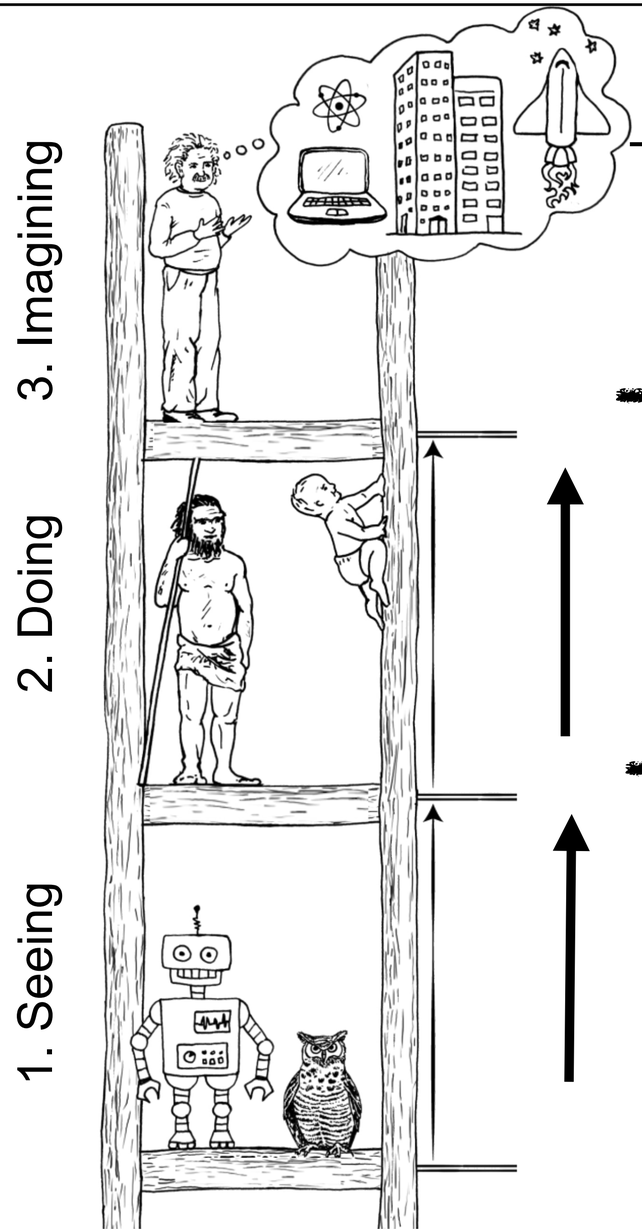
Seeing

Doing


Data
Potential Causal Diagrams
Potential SCMs
Observational
X
Z
Y
input layer
hidden layer 1
hidden layer 2
output layer
DNN
$P(Y|X=x)$
$G_1$
$G_2$
$G_3$
$G_4$
$G_5$
$\mathcal{M}_{11} = \langle \mathbf{V}, \mathbf{U}_1, \mathcal{F}_{11}, P_{11}(\mathbf{u}_1) \rangle$
$\mathcal{M}_{1k_1} = \langle \mathbf{V}, \mathbf{U}_1, \mathcal{F}_{1k_1}, P_{1k_1}(\mathbf{u}_1) \rangle$
True Model
$\mathcal{M}_{21} = \langle \mathbf{V}, \mathbf{U}_2, \mathcal{F}_{21}, P_{21}(\mathbf{u}_2) \rangle$
$\mathcal{M}_{2k_2} = \langle \mathbf{V}, \mathbf{U}_2, \mathcal{F}_{2k_2}, P_{2k_2}(\mathbf{u}_2) \rangle$
$\mathcal{M}_{31} = \langle \mathbf{V}, \mathbf{U}_3, \mathcal{F}_{31}, P_{31}(\mathbf{u}_3) \rangle$
$\mathcal{M}_{3k_3} = \langle \mathbf{V}, \mathbf{U}_3, \mathcal{F}_{3k_3}, P_{3k_3}(\mathbf{u}_3) \rangle$
$\mathcal{M}_{41} = \langle \mathbf{V}, \mathbf{U}_4, \mathcal{F}_{41}, P_{41}(\mathbf{u}_4) \rangle$
$\mathcal{M}_{4k_4} = \langle \mathbf{V}, \mathbf{U}_4, \mathcal{F}_{4k_4}, P_{4k_4}(\mathbf{u}_4) \rangle$
$\mathcal{M}_{51} = \langle \mathbf{V}, \mathbf{U}_5, \mathcal{F}_{51}, P_{51}(\mathbf{u}_5) \rangle$
$\mathcal{M}_{5k_5} = \langle \mathbf{V}, \mathbf{U}_5, \mathcal{F}_{5k_5}, P_{5k_5}(\mathbf{u}_5) \rangle$
Markovian Parametrization
Encoded Knowledge / Assumptions


Data
Potential Causal Diagrams
Potential SCMs
Observational
X
Z
Y
input layer
hidden layer 1
hidden layer 2
output layer
DNN
$P(Y|X=x)$
$G_1$
$G_2$
$G_3$
$X$
$Y$
$\mathscr{M}_{11} = \langle \mathbf{V}, \mathbf{U}_1, \mathscr{F}_{11}, P_{11}(\mathbf{u}_1) \rangle$
$\mathscr{M}_{1k_1} = \langle \mathbf{V}, \mathbf{U}_1, \mathscr{F}_{1k_1}, P_{1k_1}(\mathbf{u}_1) \rangle$
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

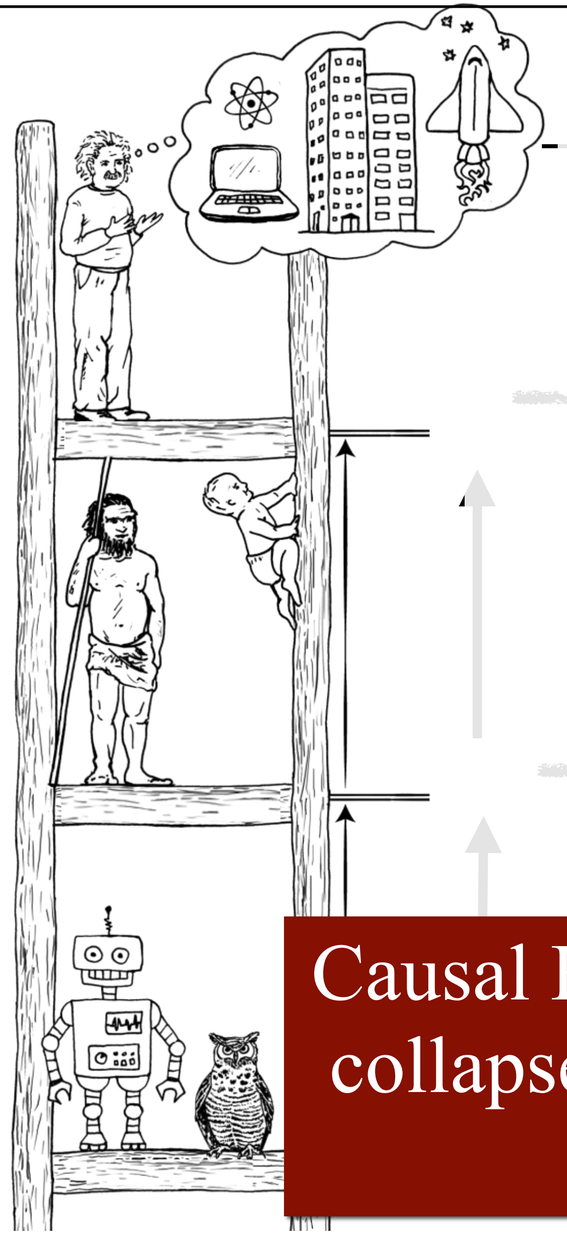
# Ladder of Causation




| 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 (MDPs, POMDPs, 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

Causal Hierarchy Theorem : The ladder almost never collapses. That is, for almost any SCM, the rungs of the ladder remain distinct.

# Structural Causal Model (SCM)

## THE DATA GENERATING MODEL

# 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 $V_i$ and $U_i \subseteq \mathbf{U}$ 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.

# Structural Equation Model (SEM)

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

- **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).**

# 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*** $do(X = x)$?

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

# Effect of Interventions in SCMs

**Observational Distribution**

$$P(\mathbf{V}) \doteq P_{\mathscr{M}}(\mathbf{V}) = \sum_{\mathbf{u}} \prod_{V_i \in \mathbf{V}} P(v_i \mid pa_i, u_i) P(\mathbf{u})$$

**Factorization obtained by Chain Rule and conditional independencies implied by the SCM $\mathscr{M}$.**

$\neq$

$do(X = x)$

**Interventional Distribution**

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

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

**Truncated factorization implied by the SCM $\mathscr{M}_x$.**

# SCM: Encoder of <u>Functional</u> 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!

# Graphical Causal Model

**The DAG, possibly with latent confounders (ADMG), induced by an SCM**

Directed Acyclic Graphs

Acyclic Directed Mixed Graphs

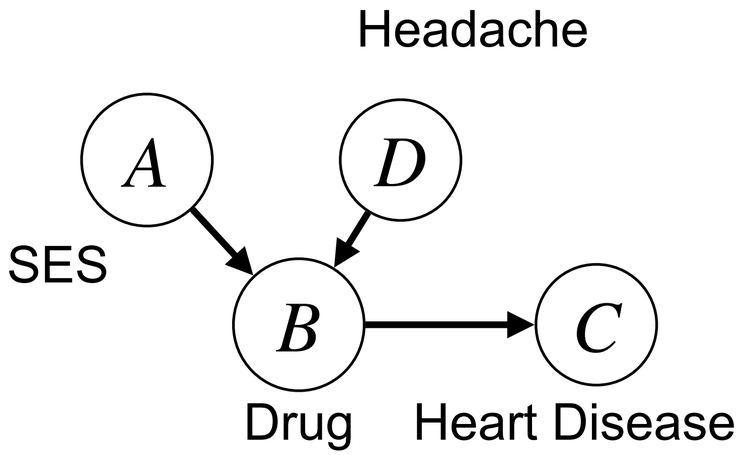
# Causal Diagram: Encoder of <u>Structural</u> Knowledge

Structural Causal Model (SCM)

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

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

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 $f_j \in \mathscr{F}$.

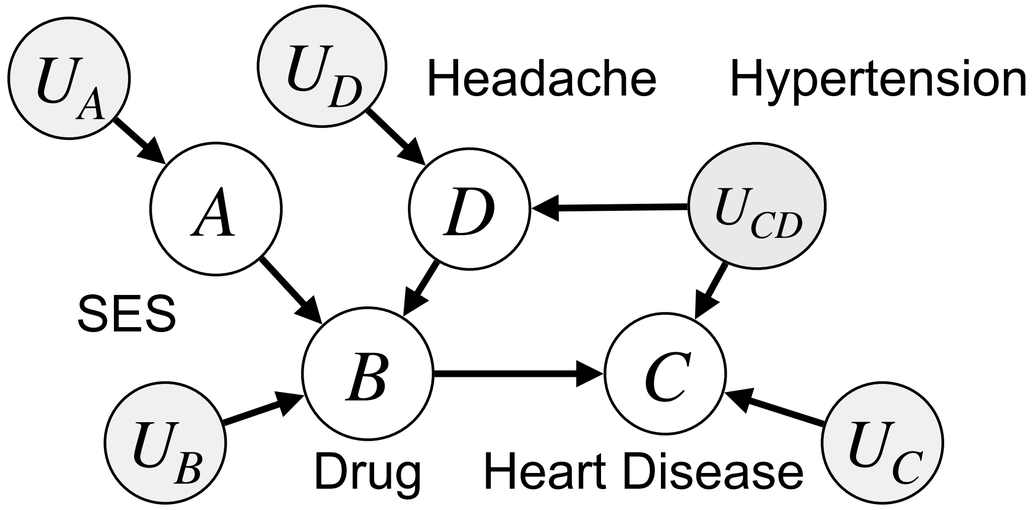
# Causal Diagram: Encoder of <u>Structural</u> Knowledge

Structural Causal Model (SCM)

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

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

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 $f_j \in \mathscr{F}$.

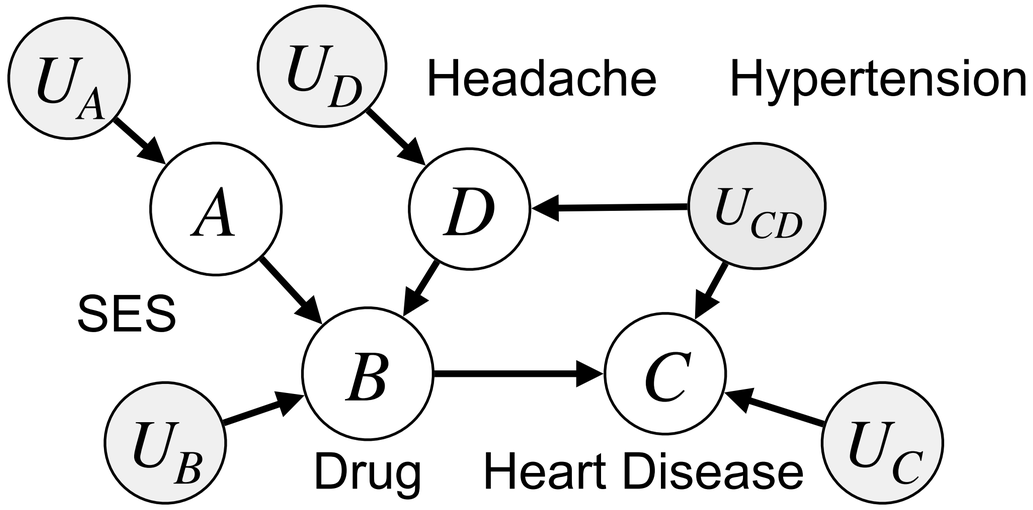
# Causal Diagram: Encoder of <u>Structural</u> Knowledge

Structural Causal Model (SCM)

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

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

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 $U \in \mathbf{U}$.

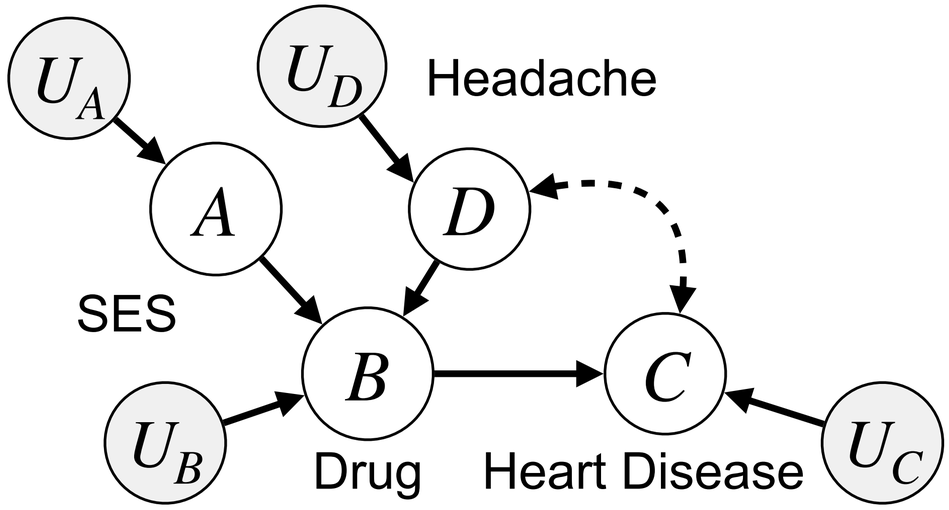
# Causal Diagram: Encoder of <u>Structural</u> Knowledge

Structural Causal Model (SCM)

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

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

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 $U \in \mathbf{U}$.

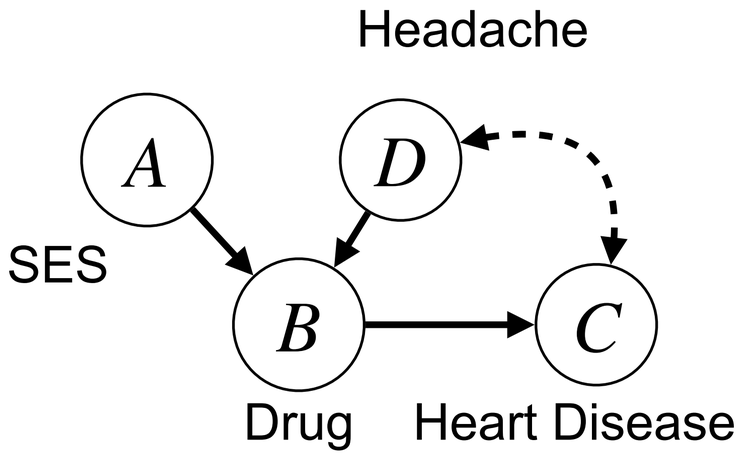
# Causal Diagram: Encoder of <u>Structural</u> Knowledge

Structural Causal Model (SCM)

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

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

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 $U \in \mathbf{U}$.

# D-Separation

Graphical Tool for DAGs and ADMGs

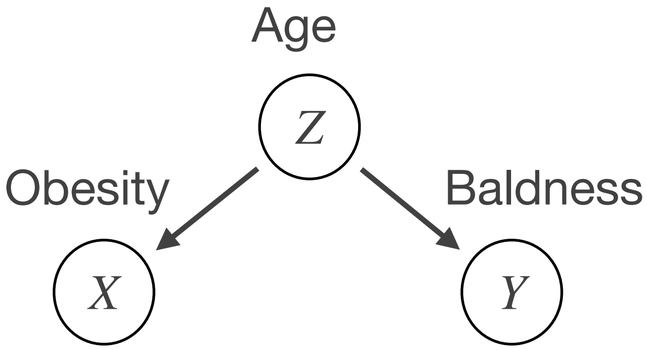
Directed Acyclic Graphs

Acyclic Directed Mixed Graphs

# Encoding Conditional Independencies

**Fork**

**$Z$ as a common cause**

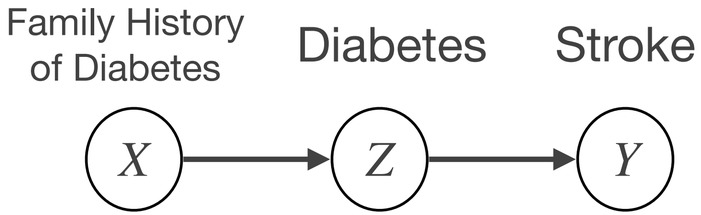



$$X \not\perp\!\!\!\perp Y$$

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

**Chain**

**$Z$ as a mediator**

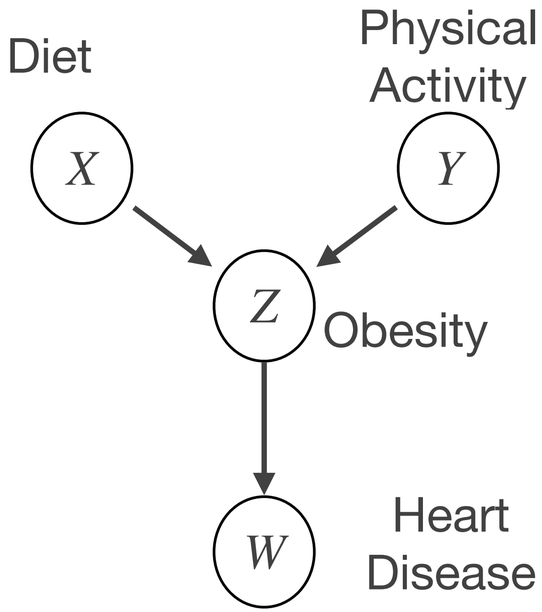



$$X \not\perp\!\!\!\perp Y$$

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

**V-Structure**

**$Z$ as a collider or common effect**




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

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

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

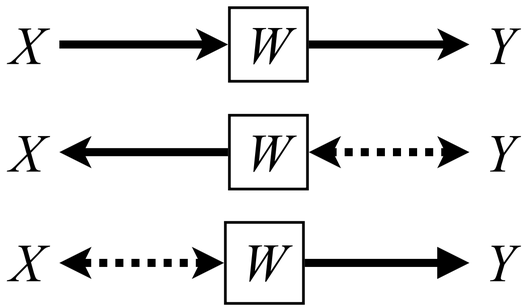
In both cases, $Z$ is a non-collider!

# Active and Inactive Triplets

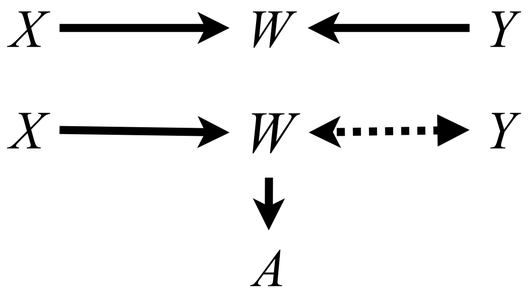
**Definition (inactive):** A triplet $\langle V_i, V_m, V_j \rangle$ is said to be ***inactive*** relative to a set $\mathbf{Z}$ if the middle node $V_m$:

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

An important tool that allows us to read the conditional (in)dependencies implied by an Acyclic Directed Mixed Graph (ADMG).

**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 $Y$ ?

$X \leftarrow B \rightarrow W \rightarrow Y$ $\quad$ $\mathbf{Z}$: ✗ $\{\}$ ✓ $\{B\}$ ✓ $\{W\}$ ✓ $\{B, W\}$

$X \rightarrow B \dashleftarrow\dashrightarrow W \leftarrow Y$ $\quad$ $\mathbf{Z}$: ✓ $\{\}$ ✓ $\{B\}$ ✓ $\{W\}$ ✗ $\{B, W\}$

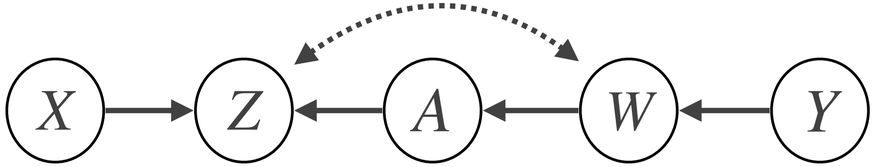
# D-Separation

An important tool that allows us to read the conditional (in)dependencies implied by an Acyclic Directed Mixed Graph (ADMG).

**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 $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_G$.

Which set d-separates $A$ and $Y$? $\{W\}$, i.e., $(A \perp\!\!\!\perp Y \mid W)_G$




What about $A$ and $X$? $\{\}$, i.e., $(A \perp\!\!\!\perp X)_G$

$Z$ and $Y$? None

**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 $P$

# Causal Effect Identification from Causal Diagrams

# 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,

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

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

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

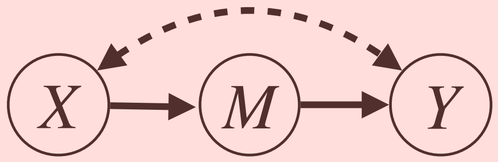
# Classical Causal Effect Identification

Structural knowledge available

1 **Query**

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

2 **Causal Contraints**




3 **Probability Distributions**

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

**Inference Engine**

**Solution**

**yes / no**

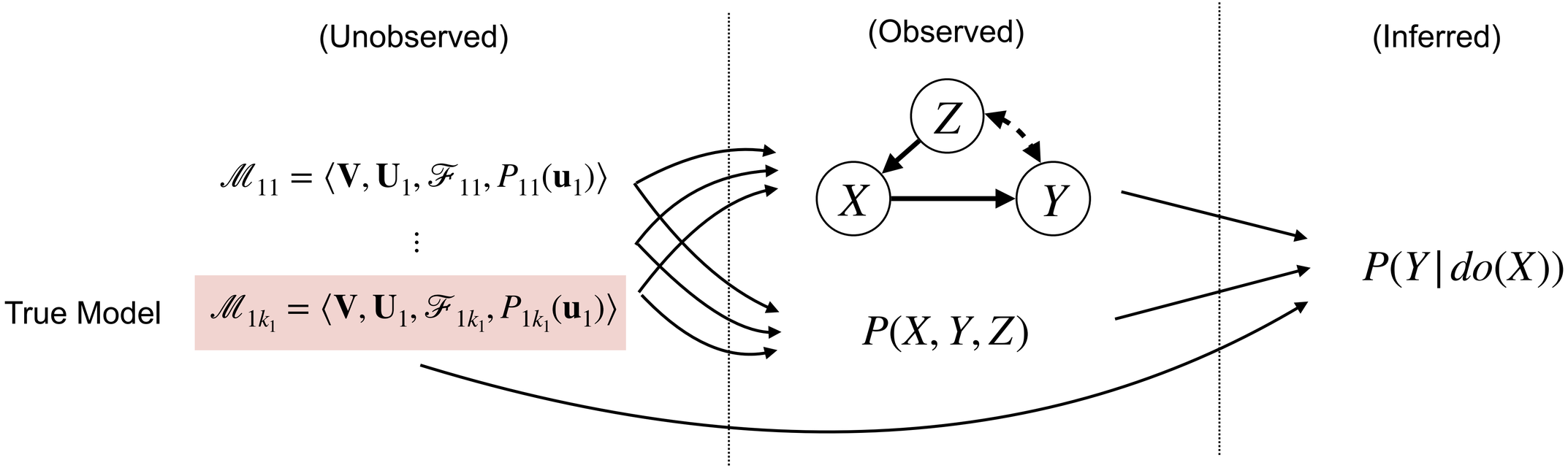
$$P(y \mid do(x)) = \sum_{m} P(m \mid x) \sum_{x'} P(y \mid m, x') P(x')$$

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.

# The Effect Identification Problem

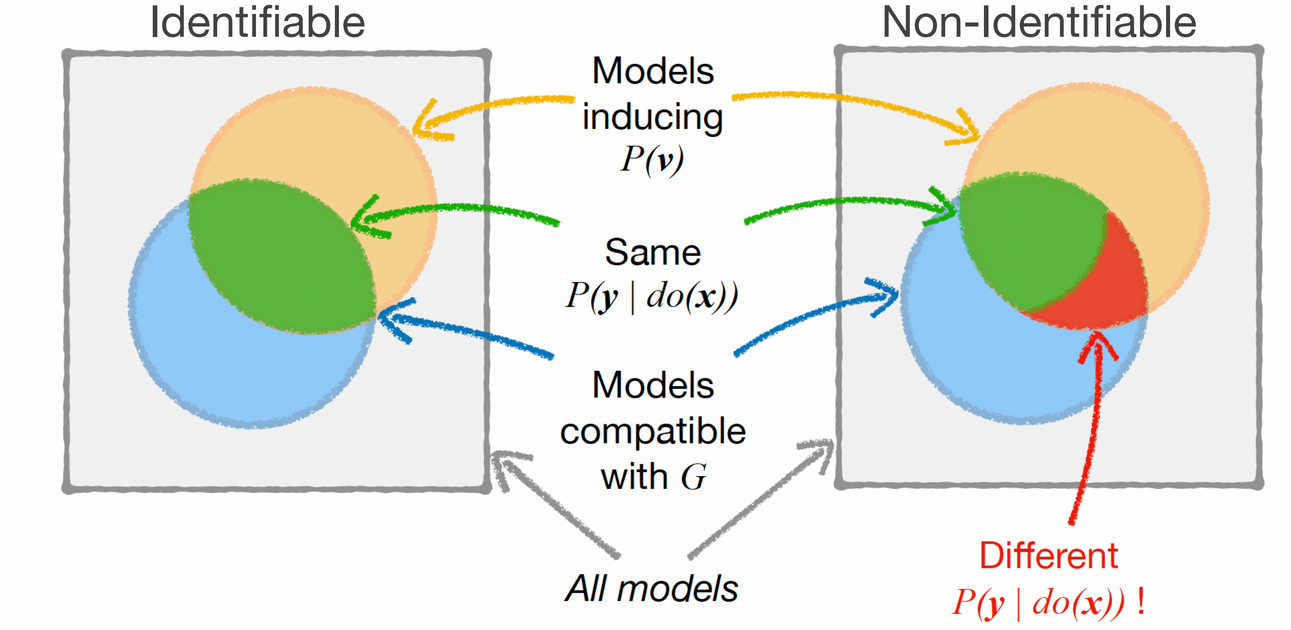
**Causal Effect Identifiability:** The causal effect of a (set of) treatment variable(s) $\mathbf{X}$ on a (set of) outcome variable(s) $\mathbf{Y}$ is said to be identifiable from a causal diagram $G$ and the probability of the observed variables $P(\mathbf{V})$ if the interventional distribution $P(\mathbf{Y} \mid do(\mathbf{X}))$ 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 $\mathscr{M}_1$ 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 $P(\mathbf{y} \mid do(\mathbf{x}))$.

# 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:

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

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

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

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

**Causal Effect of $\mathbf{X}$ on $\mathbf{Y}$:**

$$P(\mathbf{y} \,|\, do(\mathbf{x})) = \sum_{\mathbf{V} \setminus (\mathbf{Y} \cup \mathbf{X})} \prod_{V_i \in \mathbf{V} \setminus \mathbf{X}} P(v_i \,|\, pa_i) \Big|_{\mathbf{X}=\mathbf{x}}$$

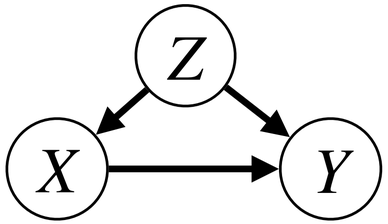
- In Markovian Models, the joint interventional distribution (and hence any causal effect) is always identifiable.
- This factorization is a.k.a “manipulation theorem” (Spirtes et al. 1993) or G-computation (Robins 1986, p. 1423).

# Example: Identifiable Effect

**Causal Effect of $\mathbf{X}$ on $\mathbf{Y}$:**

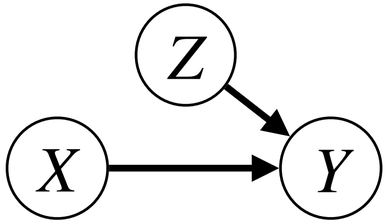
$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{V} \setminus (\mathbf{Y} \cup \mathbf{X})} \prod_{V_i \in \mathbf{V} \setminus \mathbf{X}} P_{\mathbf{x}}(v_i \mid pa_i) \Big|_{\mathbf{X}=\mathbf{x}}$$

$G$




$do(X = x)$

$G_{\overline{X}}$




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

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

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

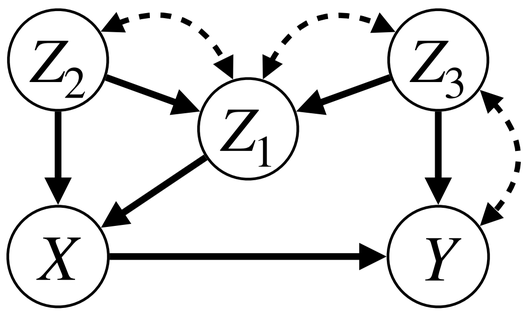
# Identification Via Adjustment over Parents

Let $G$ be a causal graph with all parents observed.

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. Try at home!




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

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

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

# Identification in Semi-Markovian Models

**Truncated Factorization — Semi-Markovian:** Let $G$ be the causal diagram for the collection $\mathbf{P}_*$ of all interventional distributions $P(\mathbf{V} \mid do(\mathbf{x}))$, for any $\mathbf{X} \subseteq \mathbf{V}$. The interventional distribution $P(\mathbf{V} \mid do(\mathbf{x}))$ factorizes as followings:

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

For Semi-Markovian Models, more advanced tools are necessary for evaluating $P_{\mathbf{x}}(\mathbf{v})$ as an expression that depends only on $P(\mathbf{V})$, i.e., a do-free and U-free expression.

# 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:

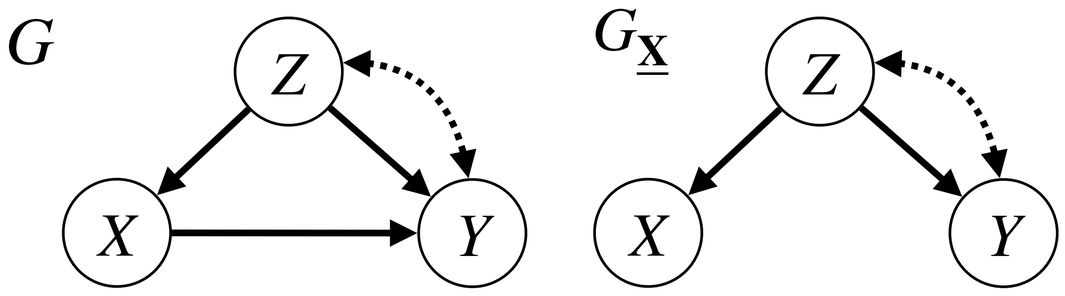
1. $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ in the graph $G_{\underline{\mathbf{X}}}$, i.e., the graph resulting from cutting the arrows out of $\mathbf{X}$
2. no node in $\mathbf{Z}$ is a descendant of a variable $X \in \mathbf{X}$ in $G$ (all variables in $\mathbf{Z}$ are pre-treatment)

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:

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

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

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




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:

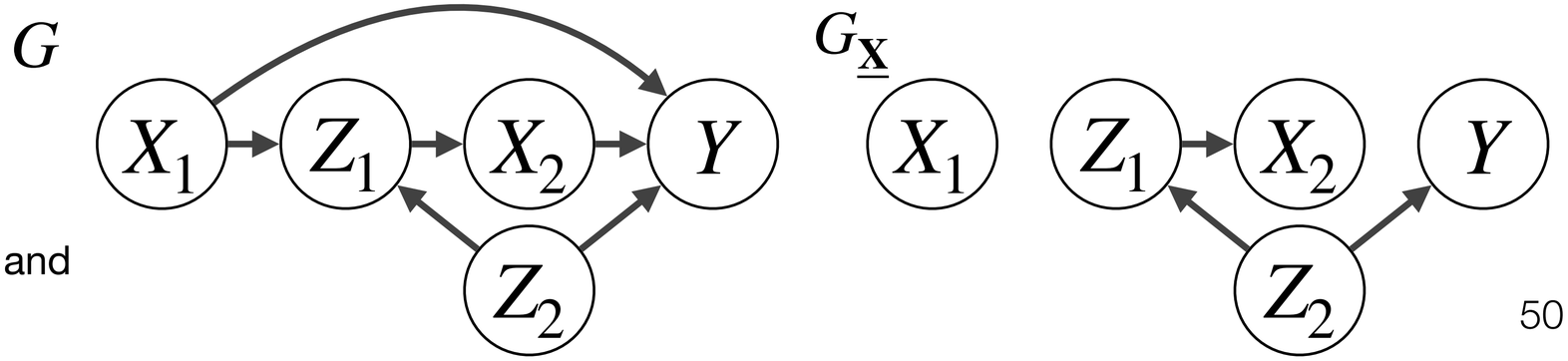
1. $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ in the graph $G_{\underline{\mathbf{X}}}$, i.e., the graph resulting from cutting the arrows out of $\mathbf{X}$
2. no node in $\mathbf{Z}$ is a descendant of a variable $X \in \mathbf{X}$ in $G$ (all variables in $\mathbf{Z}$ are pre-treatment)

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:

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

$\mathbf{X} = \{X_1, X_2\}$
$\mathbf{Y} = \{Y\}$
$\mathbf{Z} = \{Z_2\}$

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

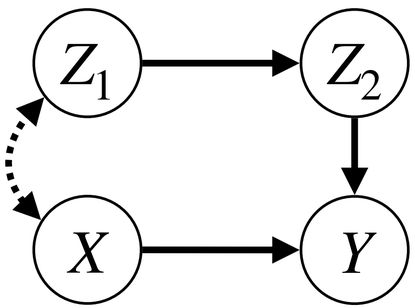


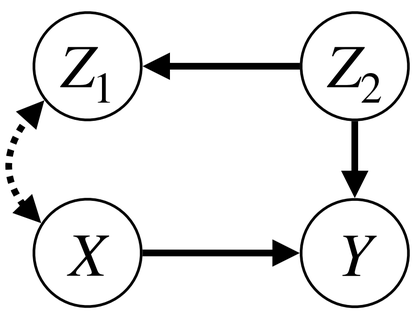

Judea Pearl. Comment: Graphical models, causality and intervention. Stat. Sci., 8:266–269, 1993.

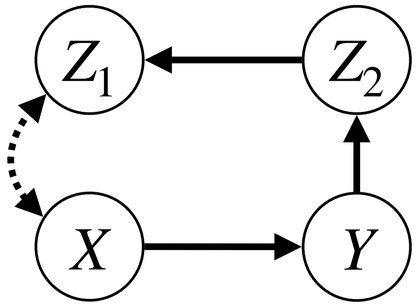
# Admissible Sets for BD Adjustment

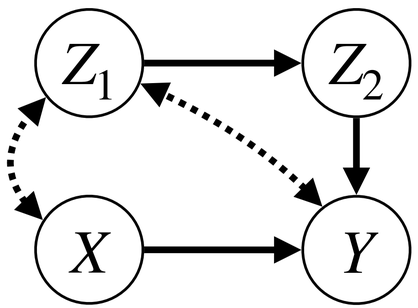
$\mathbf{Z}$ satisfies the ***backdoor criterion*** for or $(\mathbf{X}, \mathbf{Y})$ in the causal graph $G$ if:

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)



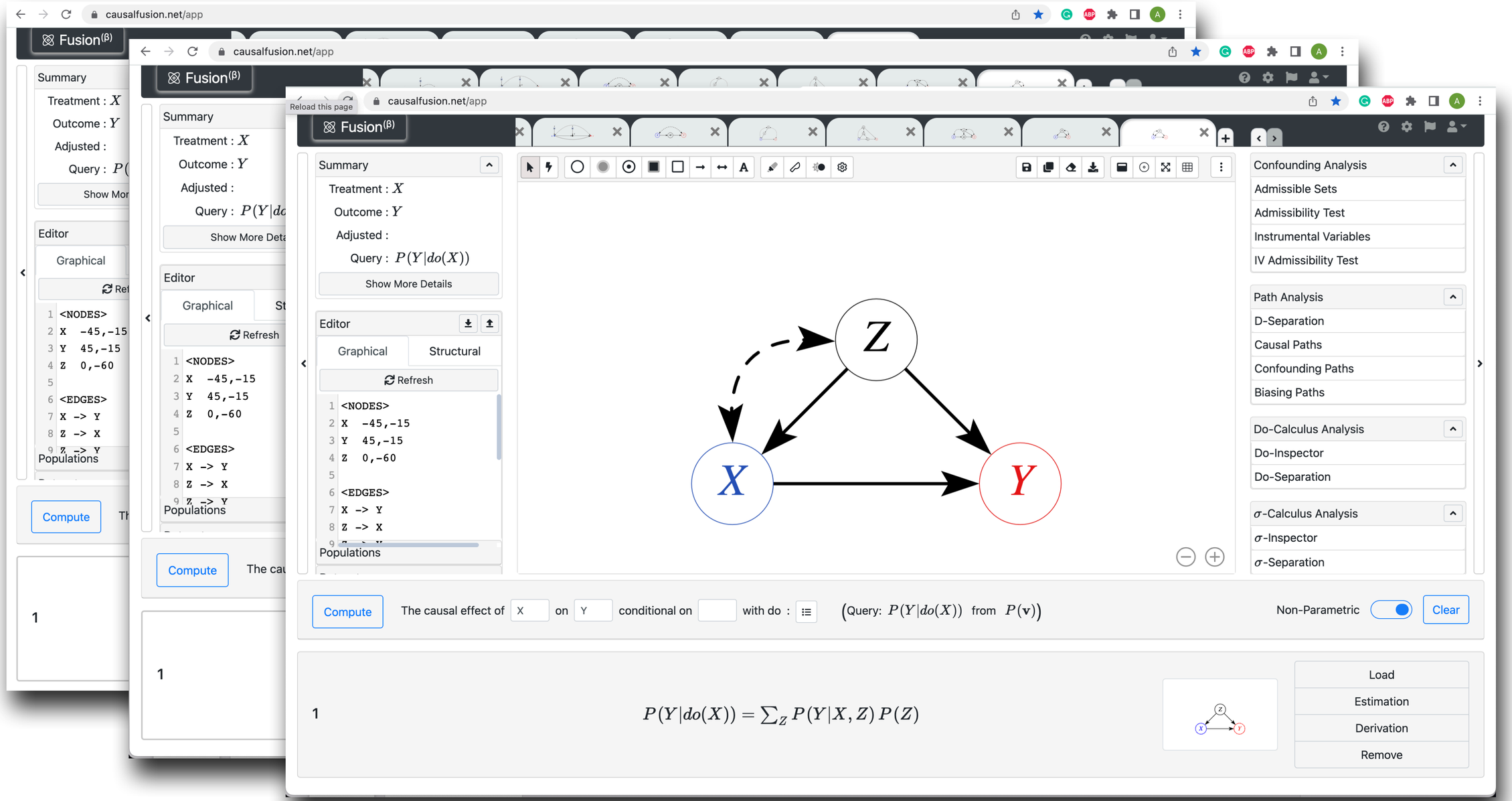



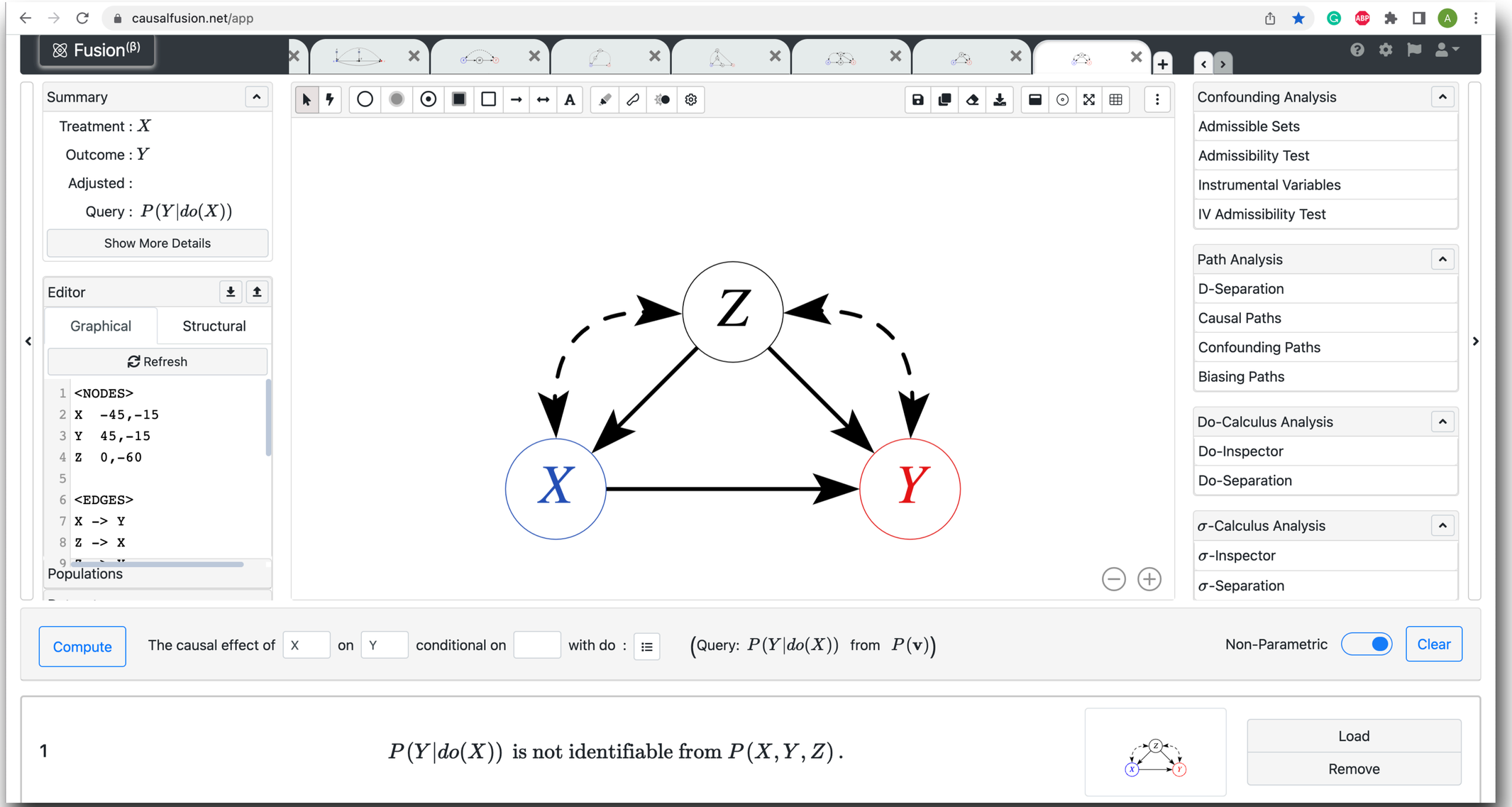




| | | | |
|---|---|---|---|
| Minimal BD Adjustment Sets | $\{Z_1\}$, $\{Z_2\}$, | $\{\}$, | $\{\}$ | There is no BD Adjustment Set! |
| | $\{Z_1, Z_2\}$ | $\{Z_2\}$, $\{Z_1, Z_2\}$ | | |
| | $P(y \mid do(x)) = \sum_{z_1} P\left(y \mid x, z_1\right) P\left(z_1\right)$ | $P(y \mid do(x)) = P\left(y \mid x\right)$ | $P(y \mid do(x)) = P\left(y \mid x\right)$ | $P(y \mid do(x))$ is non-identifiable |

# http://causalfusion.net



http://causalfusion.net



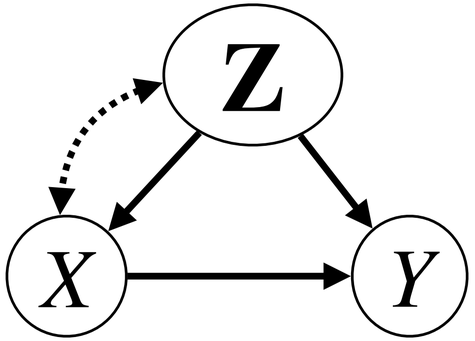
# Counterfactual Interpretation of Backdoor

Theorem 4.3.1, Pearl's Primer Book

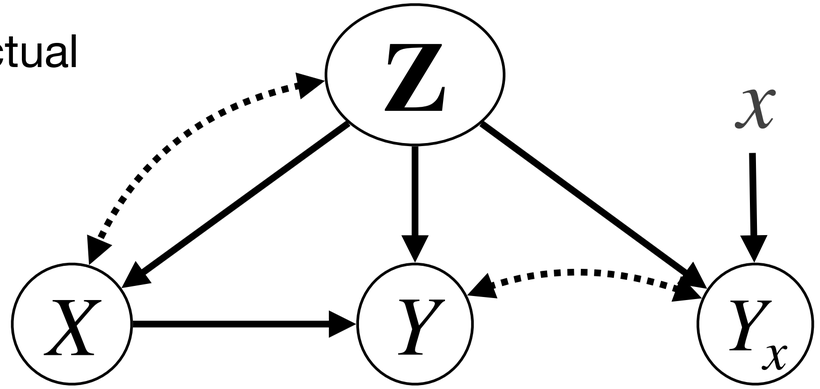
**Theorem:** If a set $\mathbf{Z}$ satisfies the *backdoor criterion* w.r.t. the ordered pair $(X, Y)$, then, for all $x$, it holds that $Y_x \perp\!\!\!\perp X \mid \mathbf{Z}$.

Although the satisfiability of $\mathbf{Z}$ to the *backdoor criterion* can be tested given a causal diagram or a PAG, the condition $Y_x \perp\!\!\!\perp X \mid \mathbf{Z}$ is sometimes framed as an assumption, referred to as **(conditional) ignorability**, **exchangeability** or **unconfoundedness**.

Observational Graph

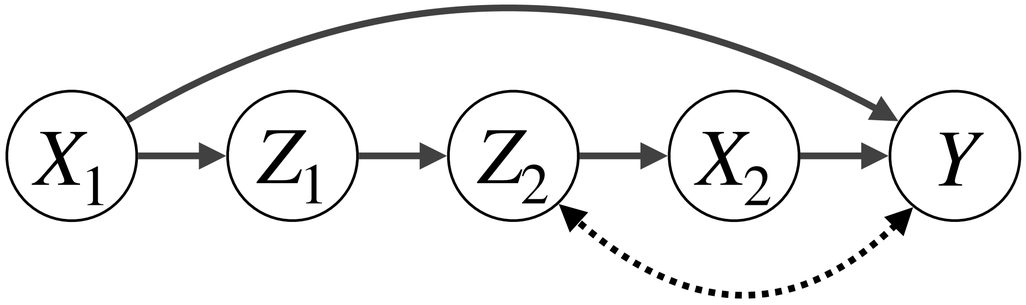



Counterfactual Graph




$Y_x \perp\!\!\!\perp X \mid Z$

# Is BD complete for covariate adjustment?




There are open backdoor paths between $(\{X_1, X_2\}, \{Y\})$, e.g.. $\langle X_2, Z_2, Y\rangle$ and $\langle X_2, Z_2, Z_1, X_1, Y\rangle$.

However, no set is admissible for ***backdoor adjustment*** for $(\{X_1, X_2\}, \{Y\})$, as both $Z_1$ and $Z_2$ are descendants of $\{X_1, X_2\}$ in $G$.

However, we have $P(y \mid do(x_1, x_2)) = \sum_z P\left(y \mid x_1, x_2, z_1, z_2\right) P\left(z_1, z_2\right)$

Verify on causalfusion.net

Backdoor Criterion is **sound (sufficient)** but not **complete (necessary)** for covariate adjustment.

There are descendants of $\mathbf{X}$ that may be used (and sometimes needed) for adjustment.

**Can we have a sound and complete graphical criterion for covariate adjustment?**

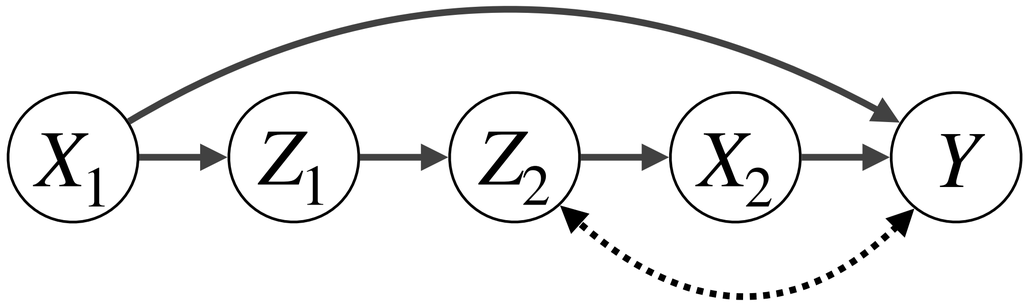
# The Adjustment Criterion

**Sound and Complete Graphical Criterion for Covariate Adjustment:**

- Shpitser. et al (2010) - Shpitser, I., VanderWeele, T., & Robins, J. M. On the validity of covariate adjustment for estimating causal effects. UAI'10: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence - *(Link)*

# Proper Paths

**Definition (Proper Path):** A path $p$ between a node in $\mathbf{X}$ and a node in $\mathbf{Y}$ is said to be *proper* if only its first node is in $\mathbf{X}$.




$$\mathbf{X} = \{X_1, X_2\}$$
$$\mathbf{Y} = \{Y\}$$

Causal paths between $\mathbf{X}$ and $\mathbf{Y}$:

$\langle X_1, Y \rangle$
$\langle X_2, Y \rangle$ } Proper causal paths

$\langle X_1, Z_1, Z_2, X_2, Y \rangle$ Non-proper causal path

Non-causal paths between $\mathbf{X}$ and $\mathbf{Y}$:

$\langle X_2, Z_2, Y \rangle$ } Proper non-causal path

$\langle X_2, Z_2, Z_1, X_1, Y \rangle$ Non-proper non-causal path

# Adjustment Criterion for Causal Diagrams

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 proper non-causal path between $X$ and $Y$, and
2. no node in $\mathbf{Z}$ is a descendant of a variable $W \notin \mathbf{X}$ which lies on a proper causal path from $\mathbf{X}$ to $\mathbf{Y}$

$W \in Desc(W)$

$\mathbf{Z}$ cannot contain a variable in a proper causal path or a descendant of such variable.

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

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

$\mathbf{Z} = \{\}$ ✗

$\mathbf{Z} = \{Z_2\}$ ✗

$\mathbf{Z} = \{Z_1, Z_2\}$ ✓

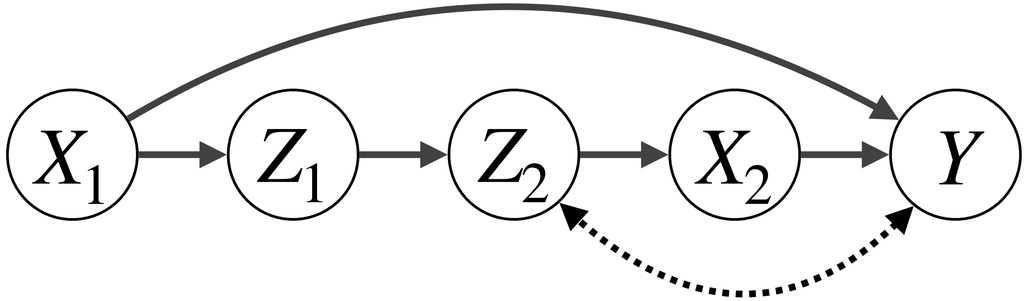
Proper non-causal paths between $\mathbf{X}$ and $\mathbf{Y}$:

$\langle X_2, Z_2, Y \rangle$

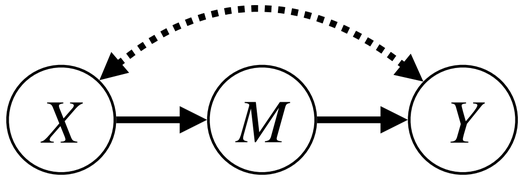
Proper causal paths between $\mathbf{X}$ and $\mathbf{Y}$:

$\langle X_1, Y \rangle$

$\langle X_2, Y \rangle$

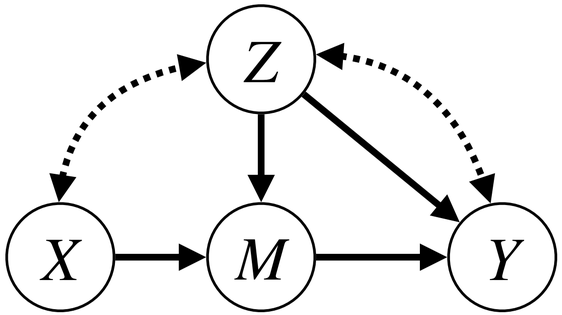


# Many Scenarios Beyond Adjustment!



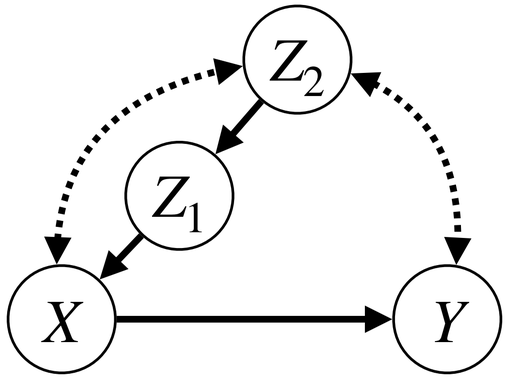

Front-Door

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



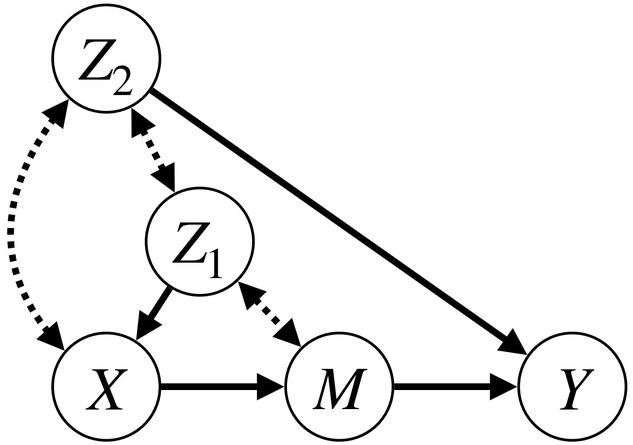

Conditional Front-Door

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




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

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

And many others….

# Tools for Causal Identification

1. Markovian Models (No Unobserved Confounders)
   i. Truncated Factorization / G-computation or G-formula
2. Adjustment over Parents (No Unobserved Parents)
3. Non-Markovian Models (Under the Presence of Unobserved Confounders)
   i. Graphical criteria (Backdoor Adjustment, Generalized Adjustment, Front-door Adjustment)
   ii. Do-Calculus (a.k.a Causal Calculus)
   iii. Identify Algorithm (a.k.a. ID algorithm)

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.

# Questions?

# Coding Exercises

## Causality Tutorial:

- **Code:** https://github.com/adele/Causality-Tutorial/blob/main/main.Rmd
- **HTML Output:** https://github.com/adele/Causality-Tutorial/blob/main/main.html

## Check Part I:

1. Causal Modeling
2. Causal Effect Identification from Causal Diagrams

What if domain knowledge does not allow you construct a causal diagram?

# Markov Equivalence Class

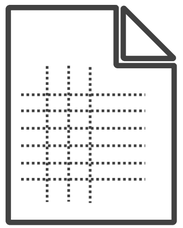
$$\mathcal{M}_1 = \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_x, U_Y\} \\ \mathcal{F} = \begin{cases} f_X(U_X) \\ f_Y(X, U_Y) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

$$\vdots$$

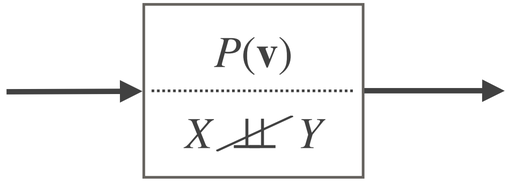
$$\mathcal{M}_{N-1} = \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_x, U_Y, U_{X,Y}\} \\ \mathcal{F} = \begin{cases} f_X(Y, U_X, U_{X,Y}) \\ f_Y(U_Y, U_{X,Y}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

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

Data



Conditional
(in)dependencies

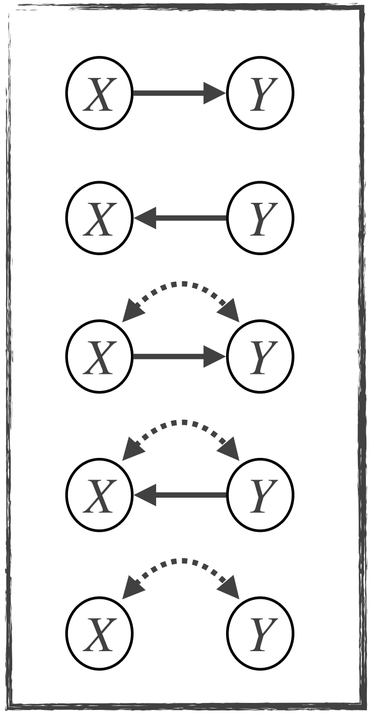



$$P(x, y) = \sum_{u_x, u_y} P(x \mid y) P(y) P(u_x, u_y)$$

$$P(x, y) = \sum_{u_x, u_y} P(y \mid x) P(x) P(u_x, u_y)$$

**Markov Equivalence Class**

(class of models implying the same set of conditional independencies)




**Correlation does not imply causation!**

# Super-Exponential Growth

The space of DAGs grows super-exponentially with the number *n* of variables, as shown by the following recurrence relation (Robinson, 1973):

$$|DAG(n)| = \sum_{i=1}^{n} \binom{n}{1} 2^{i(n-i)} |DAG(n-1)|$$

| $n$ | $\lvert DAG(n) \rvert$ |
|---|---|
| 2 | 3 |
| 3 | 27 |
| 4 | 729 |
| 5 | 59,049 |
| 6 | $1.4349 \times 10^{7}$ |
| 7 | $1.0460 \times 10^{10}$ |
| 8 | $2.2877 \times 10^{13}$ |

Inference trough enumeration
is not a good idea!

# Super-Exponential Growth

The space of ADMGs also grows super-exponentially with the number *n* of variables, and it is much bigger than the space of DAGs:

$$|ADMG(n)| = |DAG(n)| \times 2^{n(n-1)/2}$$

$$|ADMG(n)| \gg |DAG(n)|$$

| $n$ | $\lvert DAG(n) \rvert$ | $\lvert ADMG(n) \rvert$ |
|---|---|---|
| 2 | 3 | 6 |
| 3 | 27 | 216 |
| 4 | 729 | 46,656 |
| 5 | 59,049 | $6.0457 \times 10^{7}$ |
| 6 | $1.4349 \times 10^{7}$ | $4.7019 \times 10^{11}$ |
| 7 | $1.0460 \times 10^{10}$ | $2.1936 \times 10^{16}$ |
| 8 | $2.2877 \times 10^{13}$ | $6.1410 \times 10^{21}$ |

# Ancestral Graphs

**Markov Equivalence Class (MEC):** Two models over $\mathbf{V}$, $\mathscr{M}_1$ and $\mathscr{M}_2$, belong to the same MEC if for every disjoint sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subset \mathbf{V}$, $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathscr{M}_1} \Leftrightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathscr{M}_2}$

Given any ADMG $G$ over $\mathbf{V} \cup \mathbf{U} \cup \mathbf{S}$, there exists a *maximal ancestral graph* (MAG) $\mathscr{M}_G$ over $\mathbf{V}$ alone such that for any three disjoint sets of variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$,

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}, \mathbf{S})_G \Leftrightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}, \mathbf{S})_{\mathscr{M}_G}$$

ADMG $G_1$

ADMG $G_2$

$\vdots$

ADMG $G_K$

$\Rightarrow$ $\mathscr{M}_{G_i}$

$\mathscr{M}_G$ encodes all probabilistic constraints of multiple ADGM $G$ without considering the $\mathbf{U}$ variables!

# Space of MAGs and Posets

The space of MAGs is the same as the space of partially ordered sets ("posets").

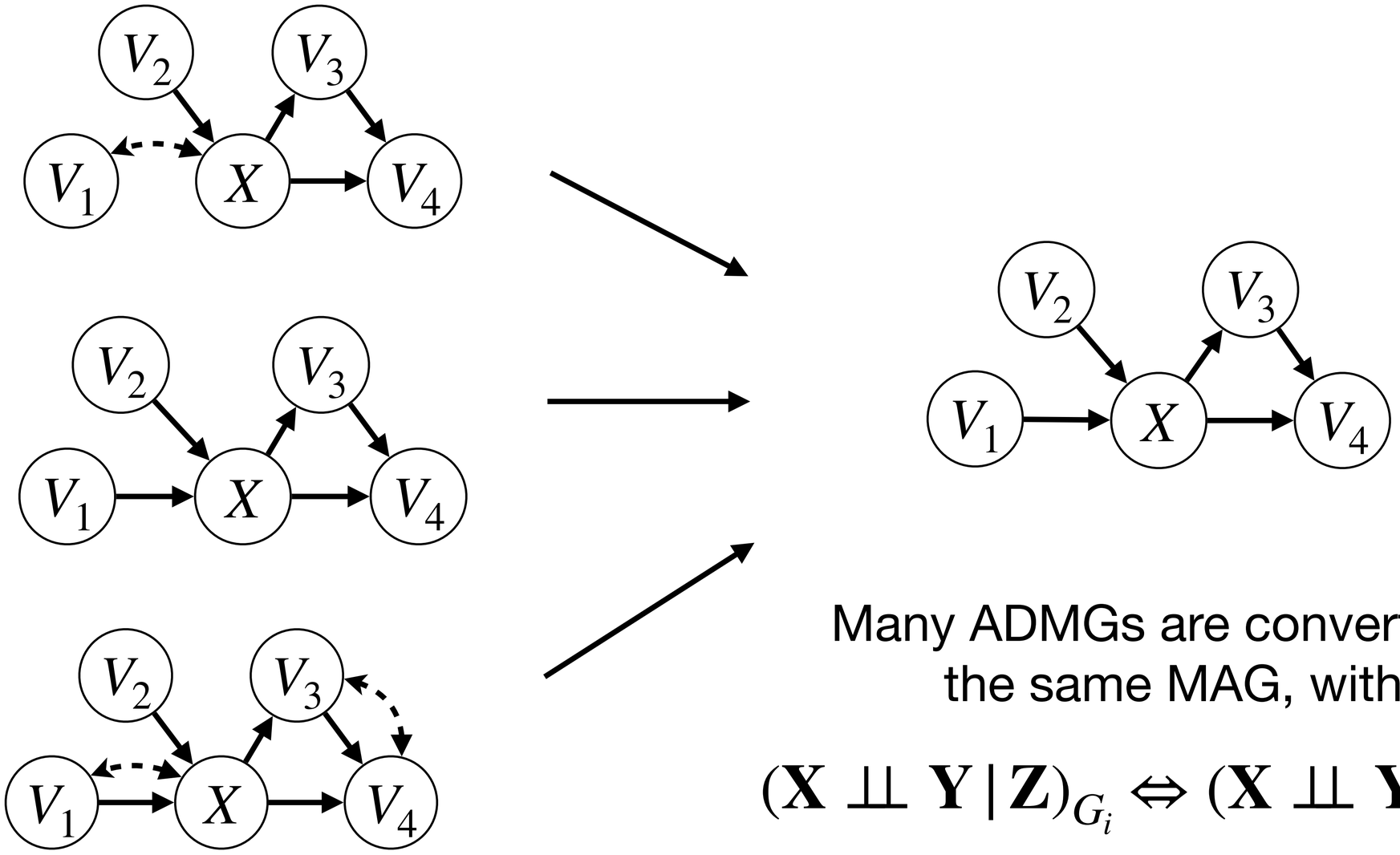
The recurrence can be found here: https://oeis.org/A001035

Growth is still super exponential, but slower than for DAGs and ADMGs.

| $n$ | $\lvert DAG(n) \rvert$ | $\lvert ADMG(n) \rvert$ | $\lvert MAG(n) \rvert$ |
|---|---|---|---|
| 2 | 3 | 6 | 3 |
| 3 | 27 | 216 | 19 |
| 4 | 729 | 46,656 | 219 |
| 5 | 59,049 | $6.0457 \times 10^7$ | 4231 |
| 6 | $1.4349 \times 10^7$ | $4.7019 \times 10^{11}$ | 130,023 |
| 7 | $1.0460 \times 10^{10}$ | $2.1936 \times 10^{16}$ | 6,129,859 |
| 8 | $2.2877 \times 10^{13}$ | $6.1410 \times 10^{21}$ | $4.3172 \times 10^8$ |

D. J. Kleitman and B. L. Rothschild. Asymptotic enumeration of partial orders on a finite set. Transactions of the American Mathematical. Society, 205:205–220, 1975.

# MAGs: Equivalence Class of ADMGs

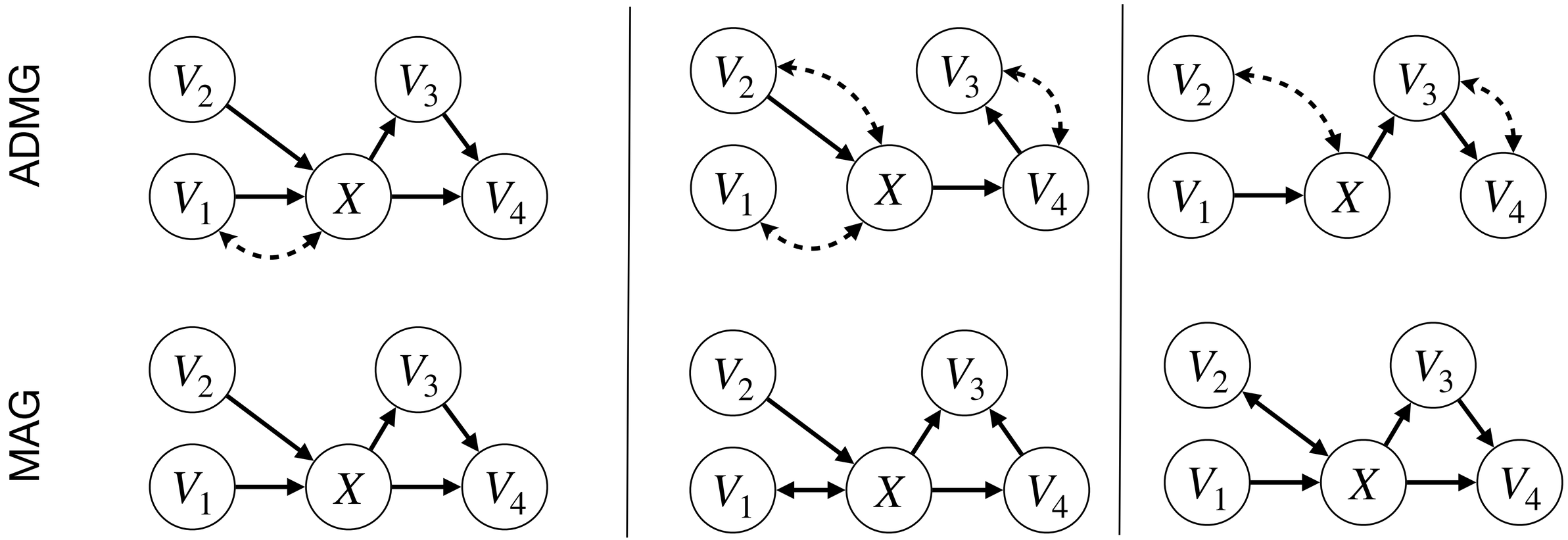



$$V_1 \perp\!\!\!\perp V_2$$
$$V_1 \perp\!\!\!\perp V_3 \mid X$$
$$V_2 \perp\!\!\!\perp V_3 \mid X$$
$$V_2 \perp\!\!\!\perp V_4 \mid X$$
$$V_1 \perp\!\!\!\perp V_4 \mid X, V_3$$
$$\vdots$$

Many ADMGs are converted to the same MAG, with:

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{G_i} \Leftrightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{M}_{G_i}}$$

# Markov Equivalent MAGs




Different MAGs may entail the same set of conditional independencies.

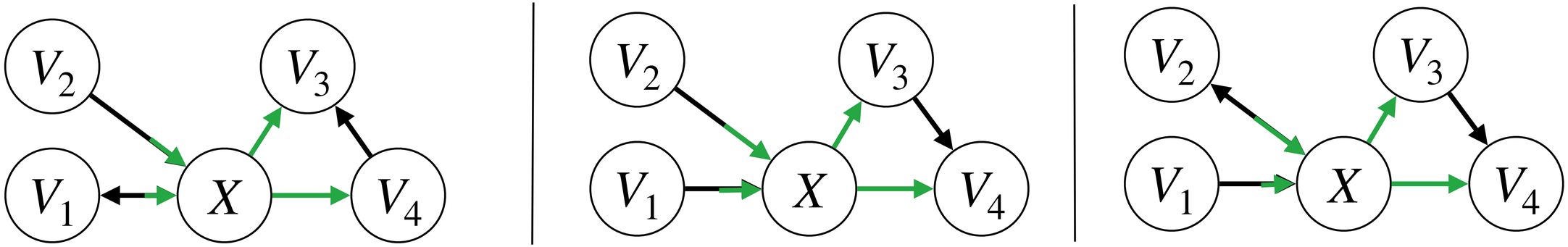
# PAGs: Markov Equivalence Class of MAGs

A **partial ancestral graph (PAG)** for $[\mathscr{G}]$ is a graph $\mathscr{P}$ with possibly three kinds of marks (and hence six kinds of edges: $-$, $\rightarrow$, $\leftrightarrow$, $\circ\!\!-$, $\circ\!\!-\!\!\circ$, $\circ\!\!\rightarrow$ ), such that

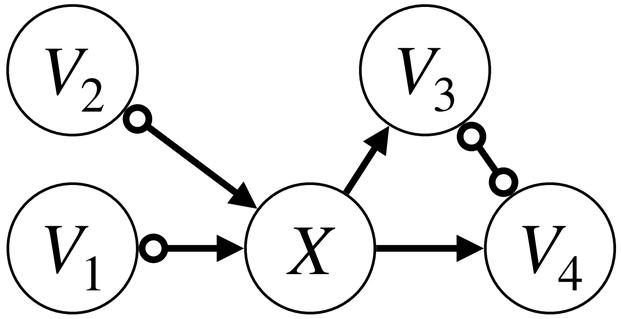
(1) $\mathscr{P}$ has the same adjacencies as any member of $\mathscr{G}$ does; and

(2) every non-circle mark in $\mathscr{P}$ is an invariant mark in $[\mathscr{G}]$.

MAG




PAG



# Learning the Markov Equivalence Class

## Causal Discovery:

Many models are statistically indistinguishable without additional parametric / distributional assumptions.

In non-parametric settings, causal discovery algorithms can only learn a graphical representation of its *Markov equivalence class* (MEC)!

**Fast Causal Inference (FCI):** Sound and complete causal discovery algorithm, even in the presence of unobserved confounders and 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

# Learning Structural Invariances

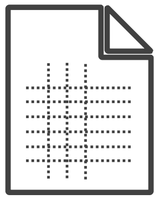
$$\mathcal{M}_1 = \begin{cases} \mathbf{V} = \{X, Y, Z\} \\ \mathbf{U} = \{U_x, U_Y, U_Z\} \\ \mathcal{F} = \begin{cases} X \leftarrow f_X(U_X) \\ Z \leftarrow f_Z(X, Y, U_Z) \\ Y \leftarrow f_Y(U_Y) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

$$\vdots$$

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

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

**Data**



**Conditional (in)dependencies**

$P(\mathbf{v})$

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

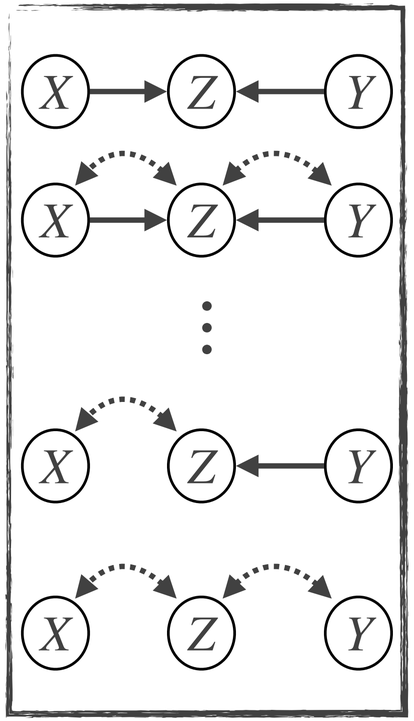
$X \not\perp\!\!\!\perp Z$

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

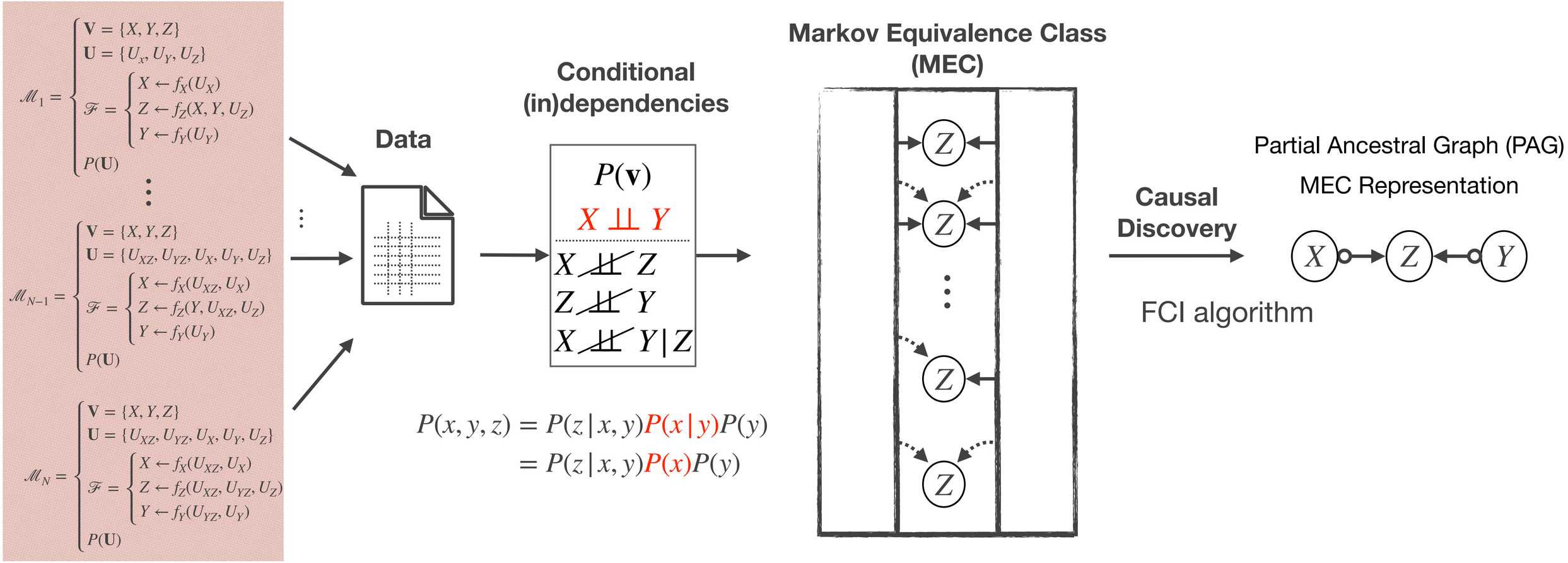
$X \not\perp\!\!\!\perp Y \mid Z$

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

**Markov Equivalence Class (MEC)**

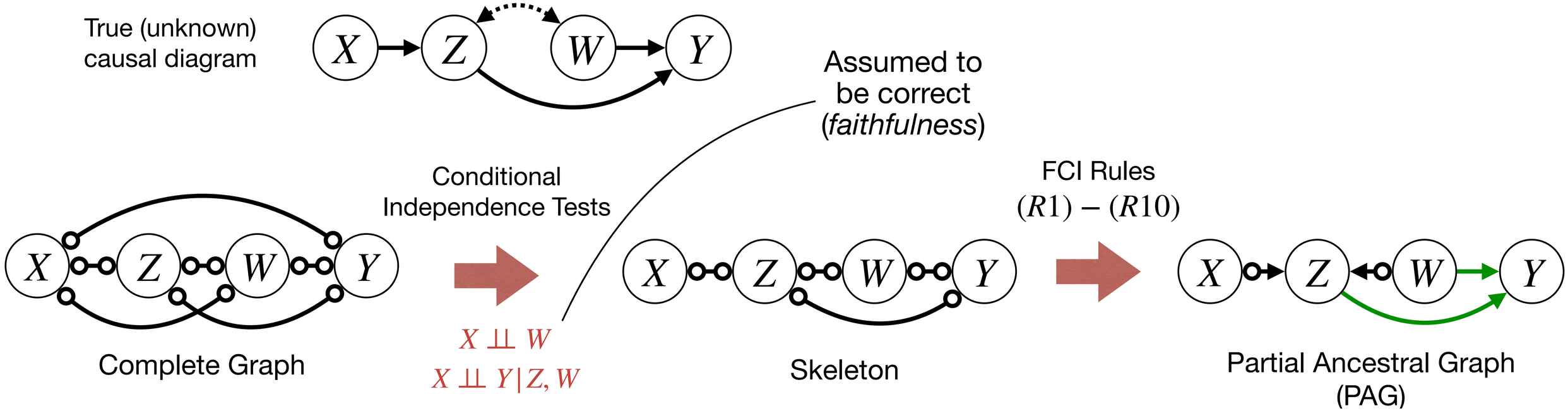


# Learning Structural Invariances




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

# Fast Causal Inference (FCI) Algorithm




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

**Tail** $\Longrightarrow$ **ancestrally**

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

$A \longleftrightarrow B$ — spurious association

$A \longmapsto B$ — selection bias

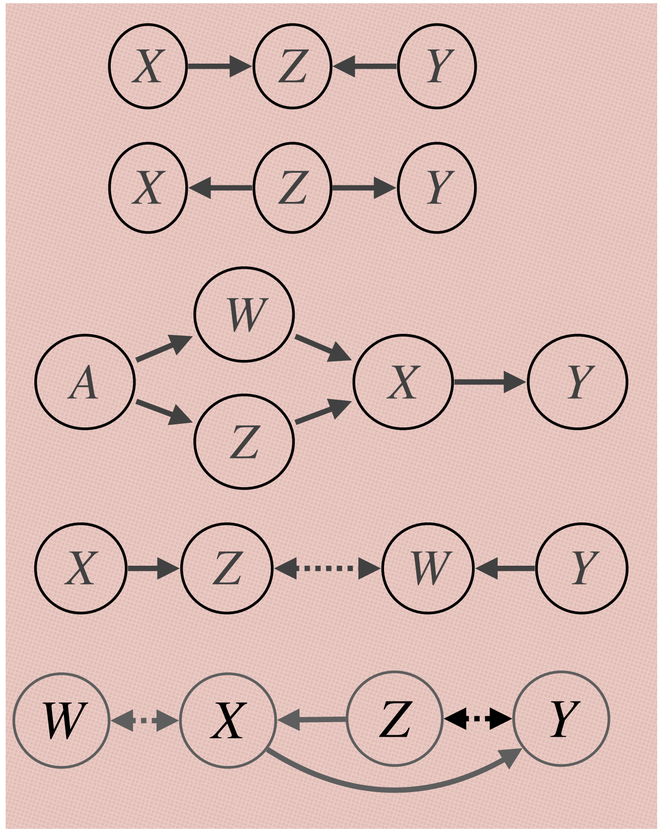
**$Z$ is not an ancestor of $X$ or $W$.**

**$Z$ and $W$ are ancestors (and definite causes) of $Y$.**

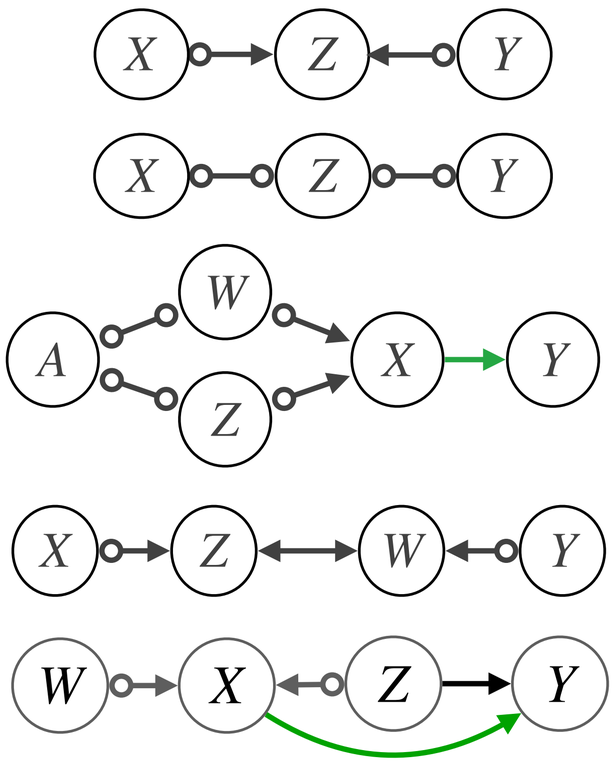
Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896. Link

# Fast Causal Inference (FCI) Algorithm

**Underlying Causal Diagram**



**Partial Ancestral Graph**



# 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

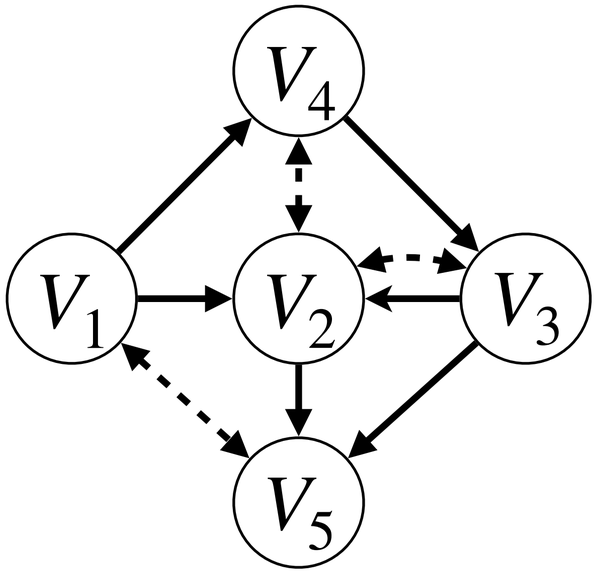
- Tsagris, M., Borboudakis, G., Lagani, V. *et al.* (2018) Constraint-based causal discovery with mixed data. *Int J Data Sci Anal* **6**, 19–30. (Link)
- R package: https://cran.r-project.org/web/packages/MXM/

Gaussian errors and correlated observations (family data) :

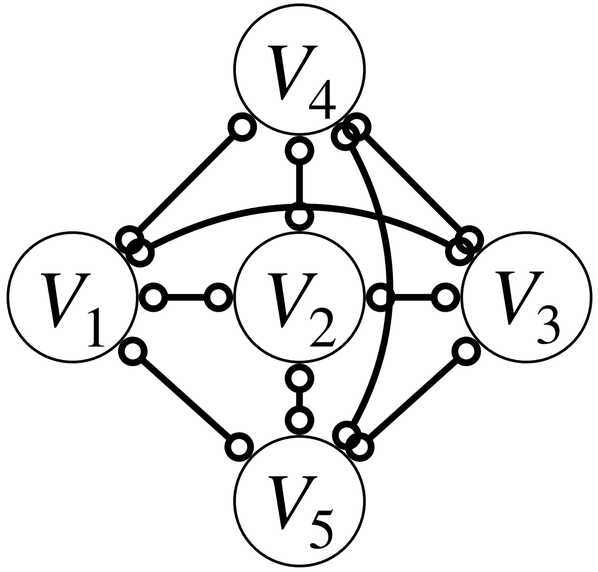
**Ribeiro A.H.,** Soler J.M.P. (2020). *Learning Genetic and environmental graphical models from family data*, Statistics in Medicine.
R package: https://github.com/adele/FamilyBasedPGMs

# FCI - Skeleton

Form a complete graph on the set of variables, in which there is a circle-circle edge between every pair of variables;

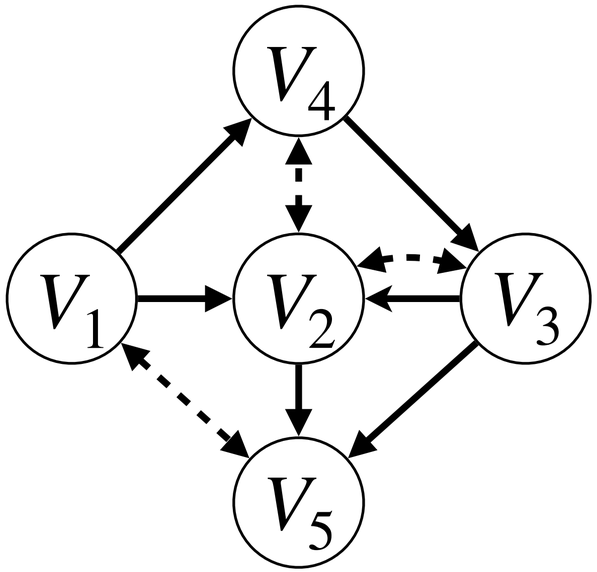



True, unknown ADMG

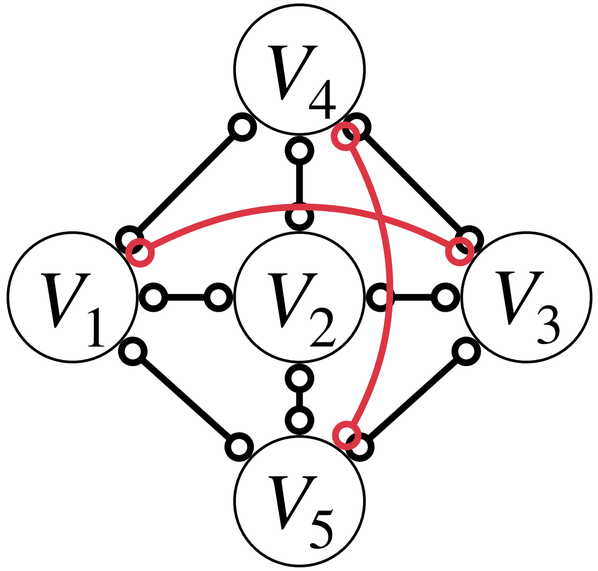


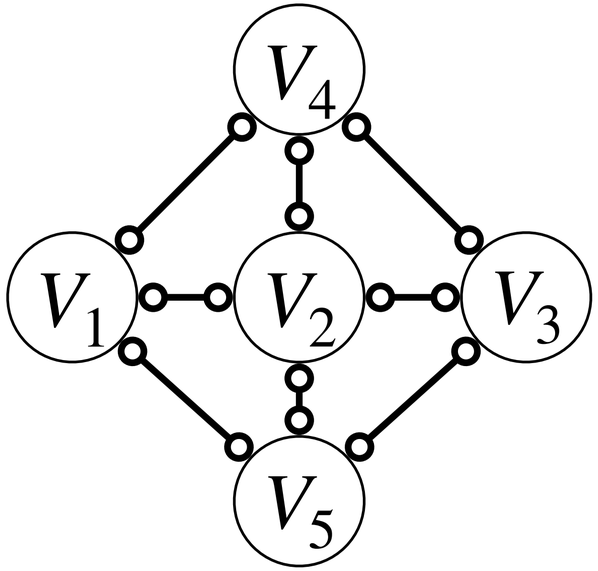
# FCI - Skeleton

For every pair of variables $V_1$ and $V_2$, if exists a set $\mathbf{S}_{1,2}$ such that $V_1 \perp\!\!\!\perp V_2 \,|\, \mathbf{S}_{1,2}$, then remove the edge between $V_1$ and $V_2$ and add $\mathbf{S}_{1,2}$ in *Sepset*$(V_1, V_2)$.




True, unknown ADMG



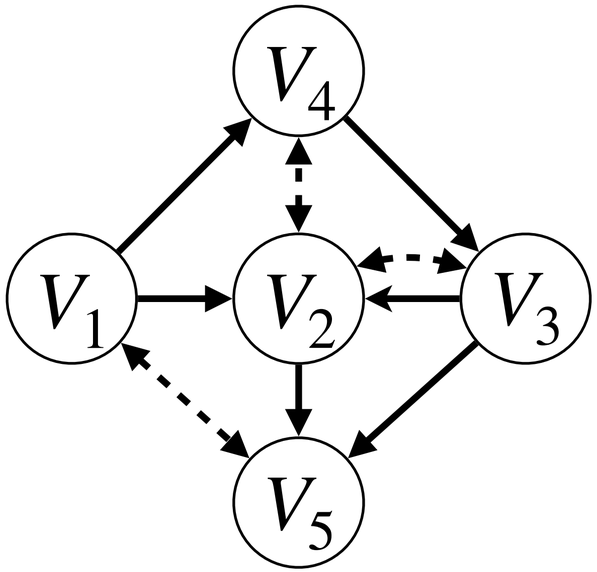




$V_1 \perp\!\!\!\perp V_3 \,|\, V_4$ and $V_4 \perp\!\!\!\perp V_5 \,|\, V_1, V_2, V_3$

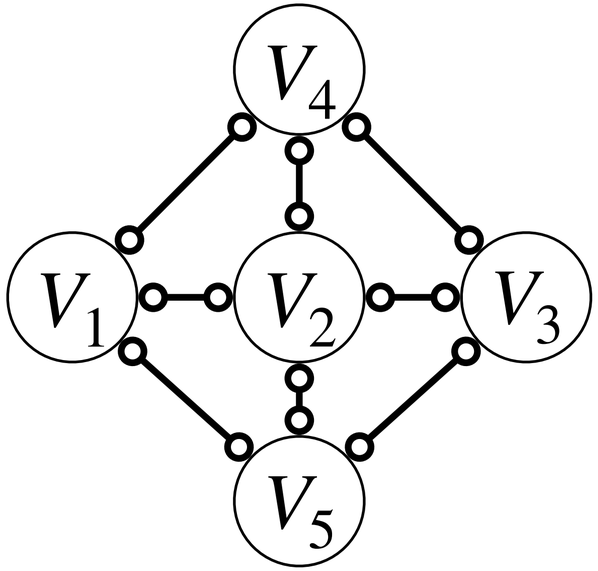
# FCI - Orienting the Colliders

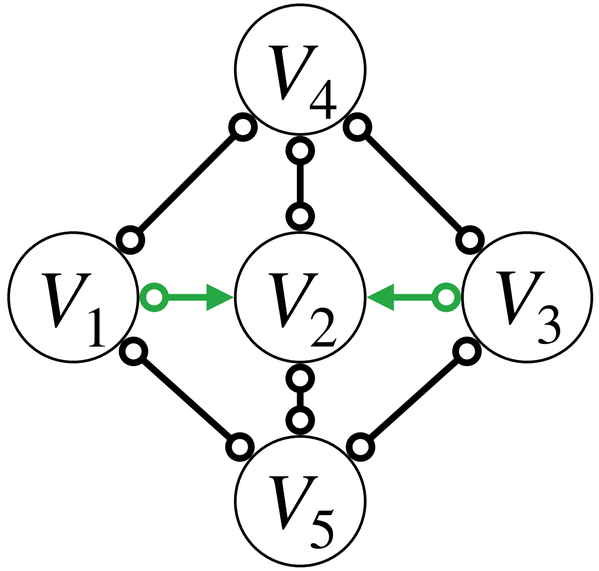
**R0:** If $\langle V_1, V_2, V_3 \rangle$ is unshielded and $V_2 \notin Sepset(V_1, V_3)$, then

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




True, unknown ADMG



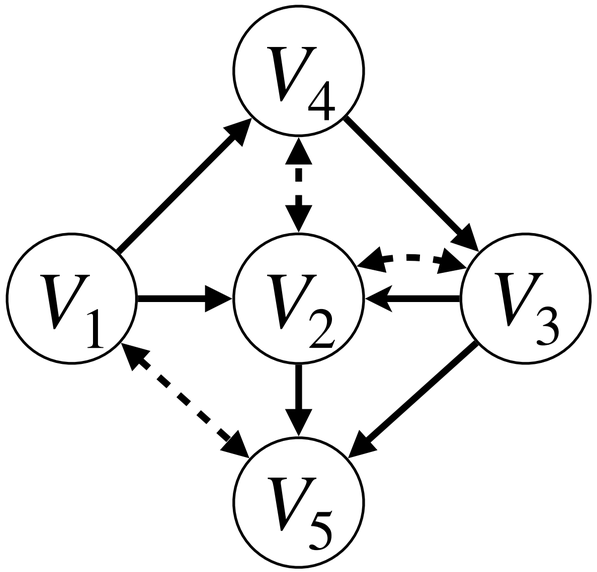




$V_1 \perp\!\!\!\perp V_3 \mid V_4$ and $V_1 \not\perp\!\!\!\perp V_3 \mid V_4, V_2$

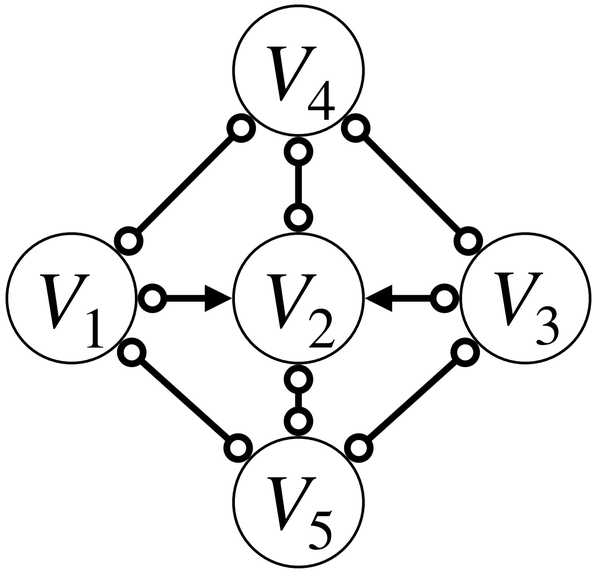
# FCI - Orienting the Colliders

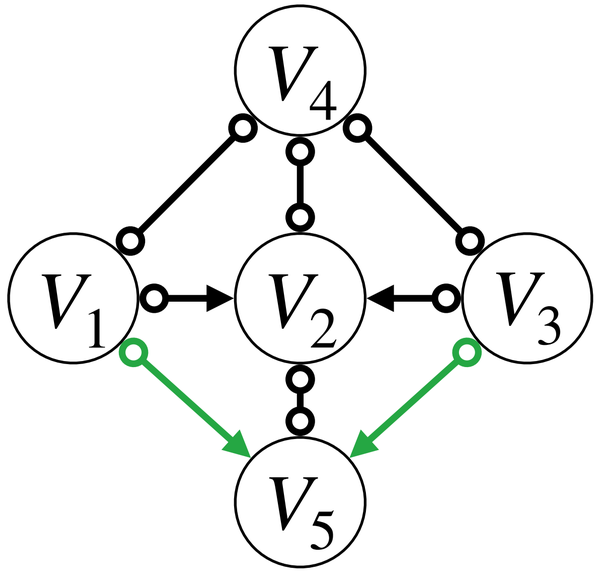
**R0:** If $\langle V_1, V_2, V_3 \rangle$ is unshielded and $V_2 \notin Sepset(V_1, V_3)$, then

We apply R0 until no more collider can be oriented!




True, unknown ADMG




$V_1 \perp\!\!\!\perp V_3 \mid V_4$ and $V_1 \not\perp\!\!\!\perp V_3 \mid V_4, V_5$

# Rules R1 to R10

Next, we apply Rules 1 to 10, in any order, whenever any of them applies.

**Theorem by Ali et al, 2005 [Arrowhead Completeness]:**

- R1 to R4, together with R0, are complete for arrowhead (non-ancestrality) orientation.

**Theorem by Zhang J., 2008 [Tail Completeness]:**

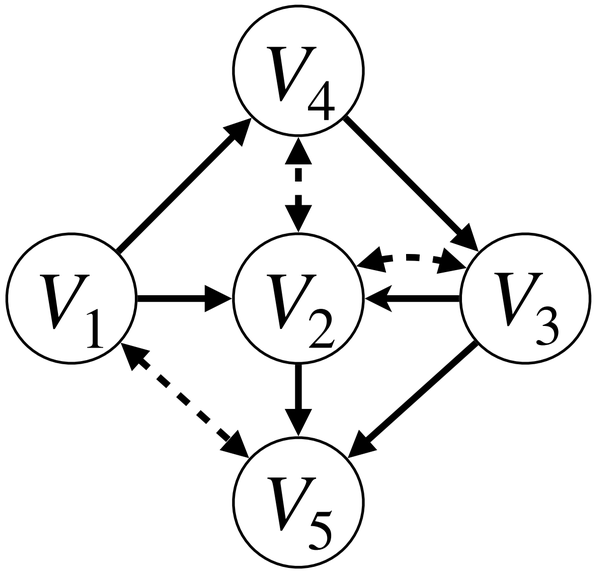
- R5 to R10, together with R0 to R4, are complete for arrowhead (non-ancestrality) and tail (ancestrality) orientation.
- If no selection variables are present, then rules R5 to R7 are not necessary.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

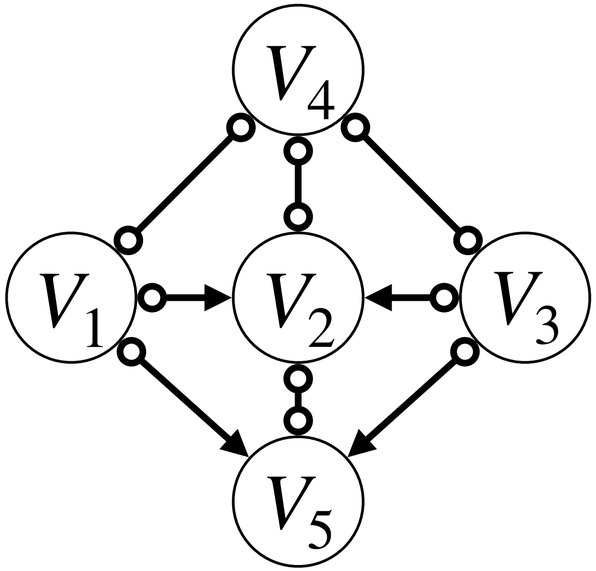
R.A. Ali, T. Richardson, P. Spirtes, J. Zhang, Towards characterizing Markov equivalence classes for directed acyclic graphs with latent variables. In Proceedings of the 21th Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2005, pp. 10–17. (Link)

Zhang, J., 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, *172*(16-17), pp.1873-1896.

# Going back to the example…



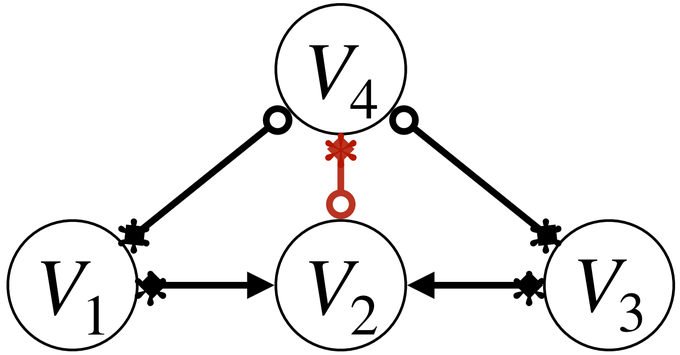

True, unknown ADMG




After Skeleton + R0

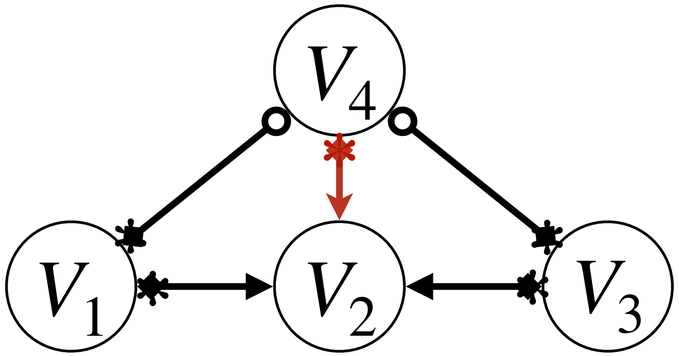
# Applying Mark Inference Rules

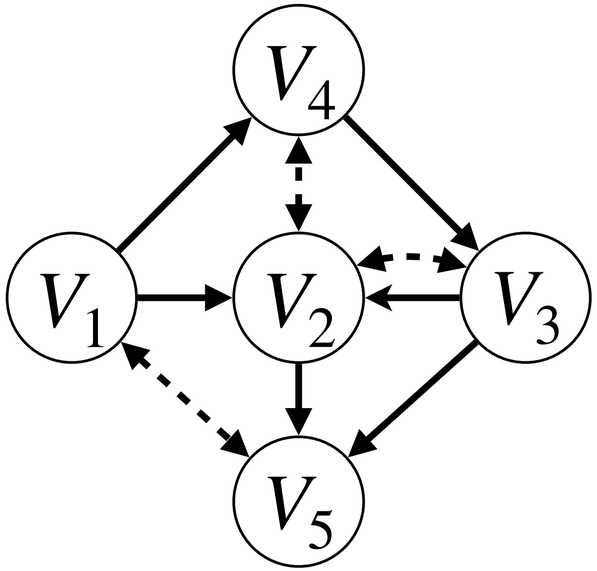
**R3:**



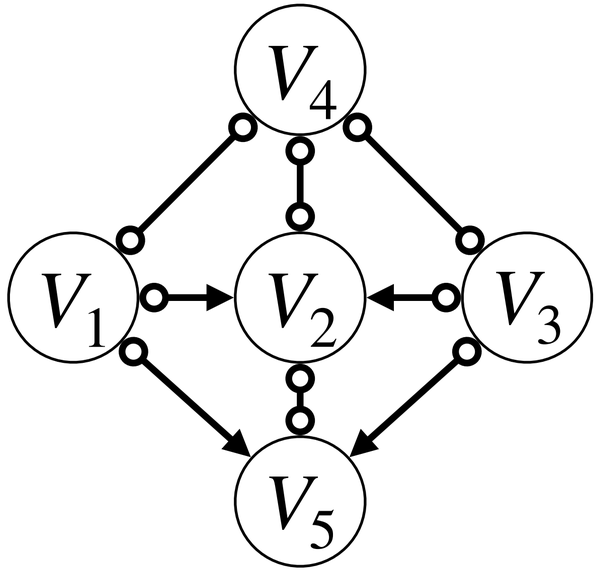

where $V_1$ and $V_3$ are not adjacent

$\Longrightarrow$

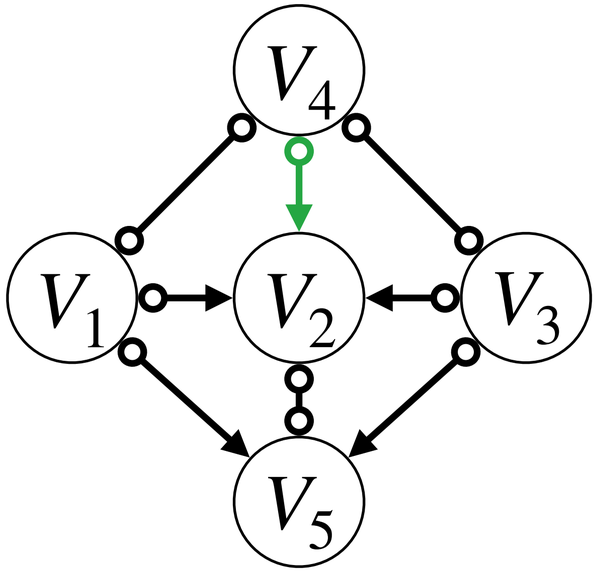



True, unknown ADMG



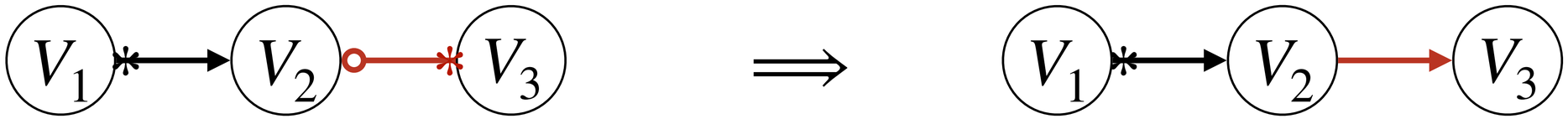

After Skeleton + R0



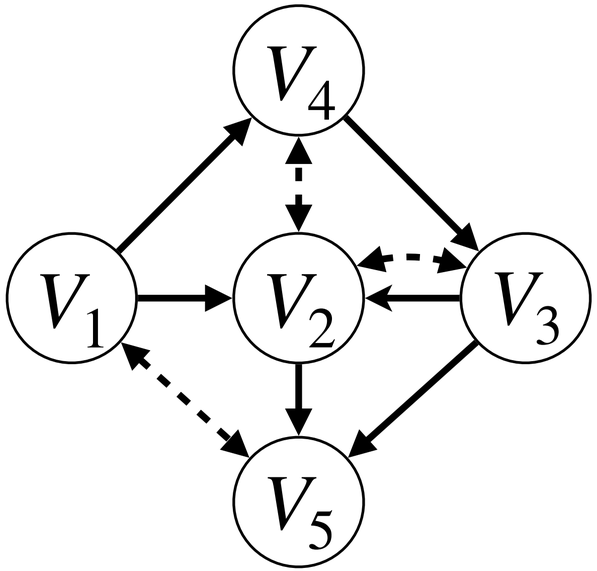

Applying R3

# Applying Mark Inference Rules

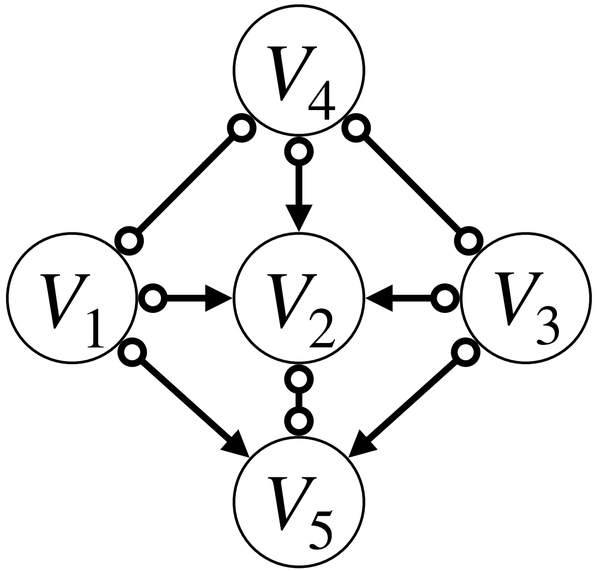
**R1:**



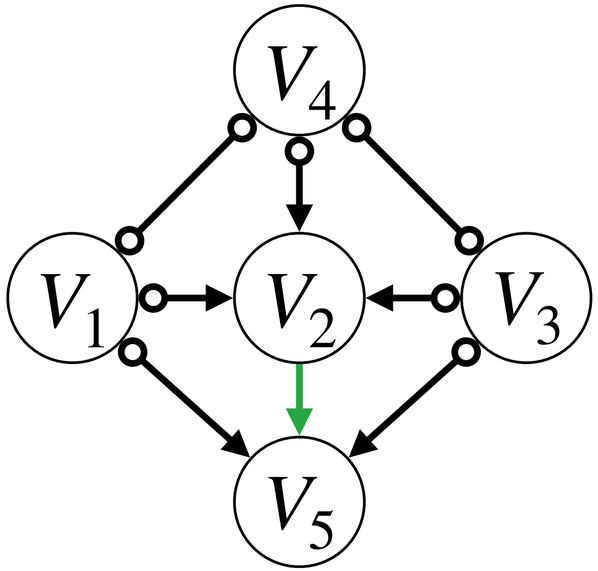

where $V_1$ and $V_3$ are not adjacent




True, unknown ADMG



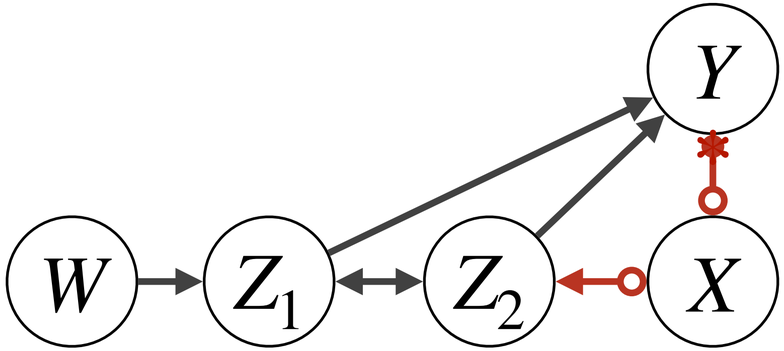

After Skel + R0 + R3




Applying R1

# Applying Mark Inference Rules

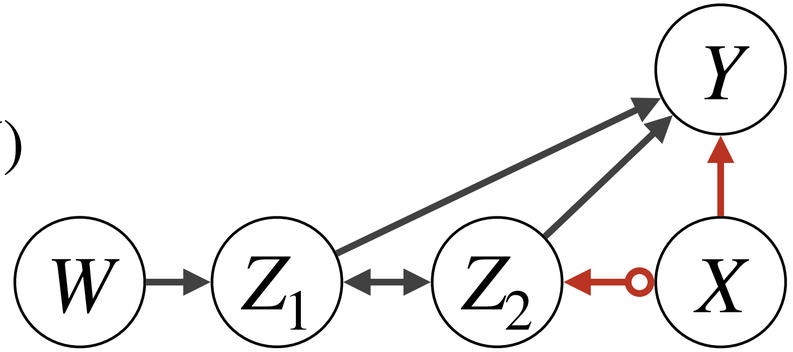
**R4:**

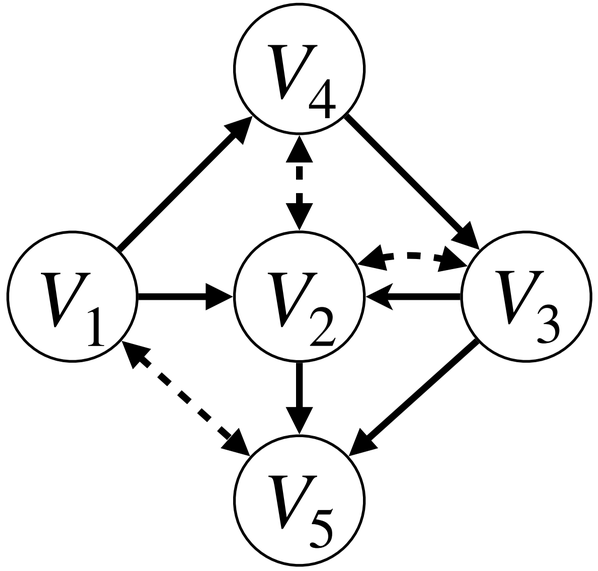



$\langle W, Z_1, Z_2, X, Y \rangle$ is a discriminating path for $X$

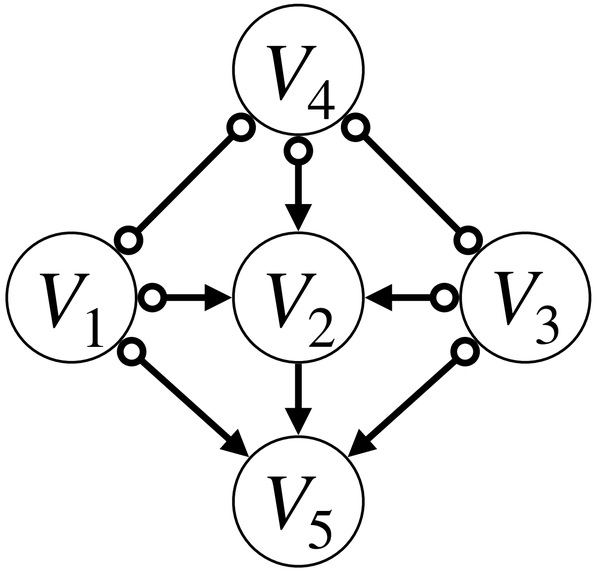
$X \in Sepset(W, Y)$

$\Longrightarrow$



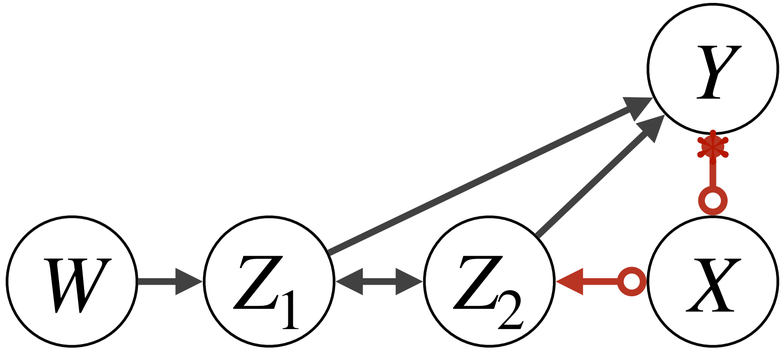

True, unknown ADMG




After Skel + R0 + R3 + R1

# Applying Mark Inference Rules

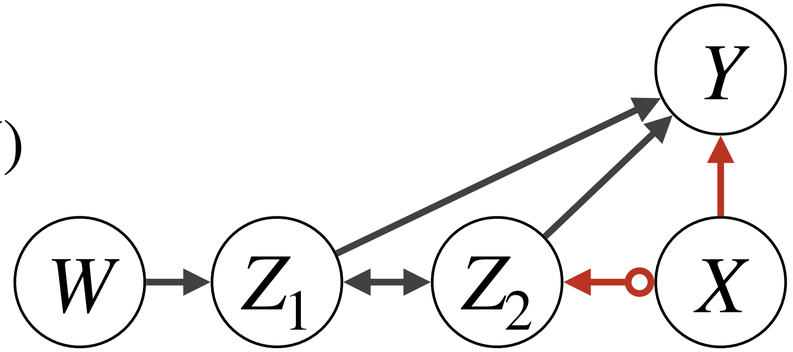
**R4:**

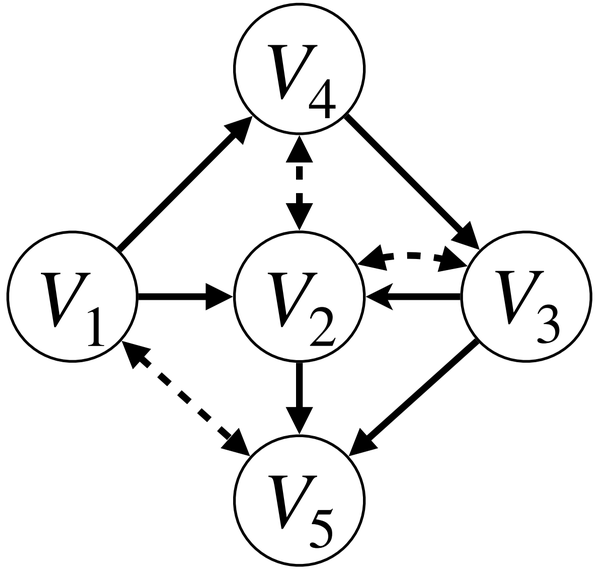



$\langle W, Z_1, Z_2, X, Y \rangle$ is a discriminating path for $X$

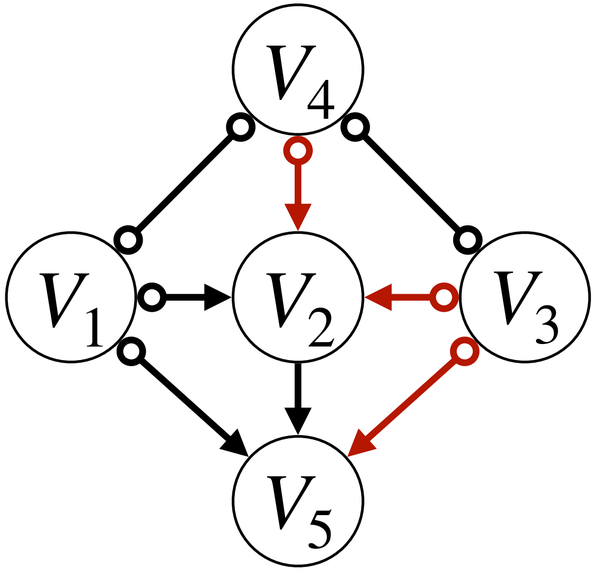
$X \in Sepset(W, Y)$

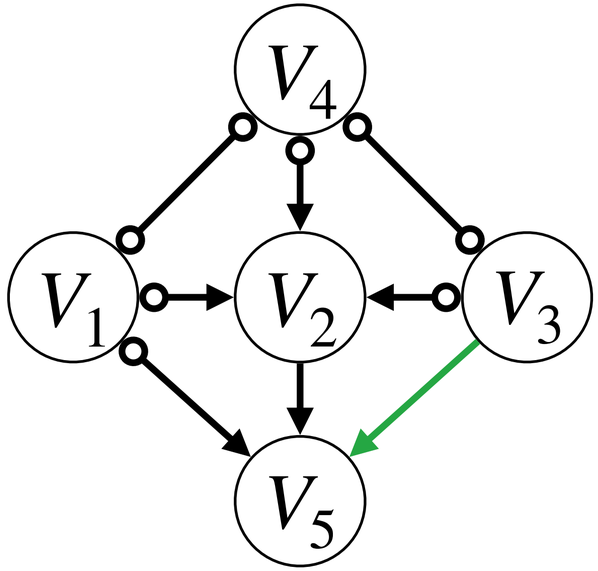
$\Longrightarrow$




True, unknown ADMG



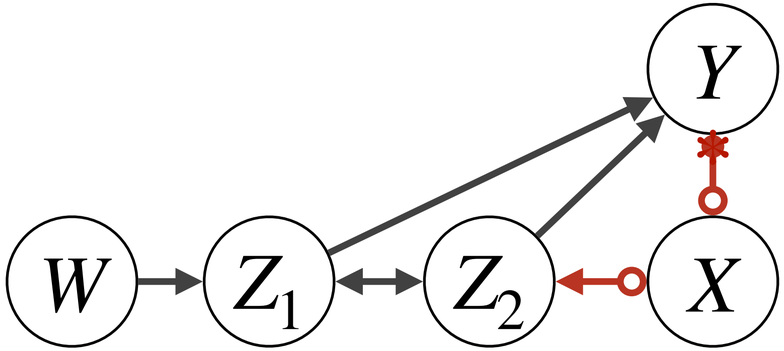




$\langle V_4, V_2, V_3, V_5 \rangle$ is a discriminating path for $V_3$ and

$V_3 \in Sepset(V_4, V_5) \;-\; V_4 \perp\!\!\!\perp V_5 \,|\, V_1, V_2, V_3$

# Applying Mark Inference Rules

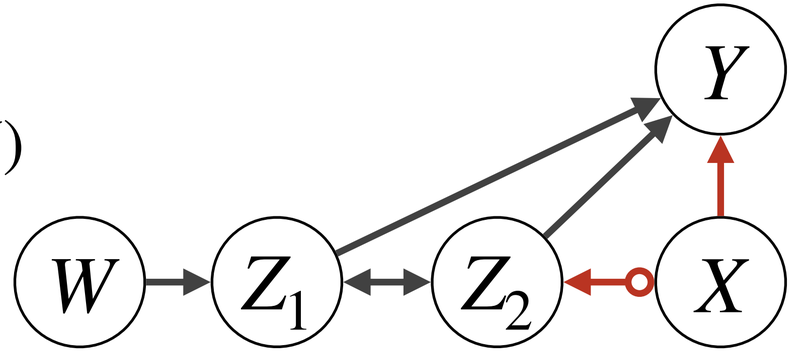
**R4:**

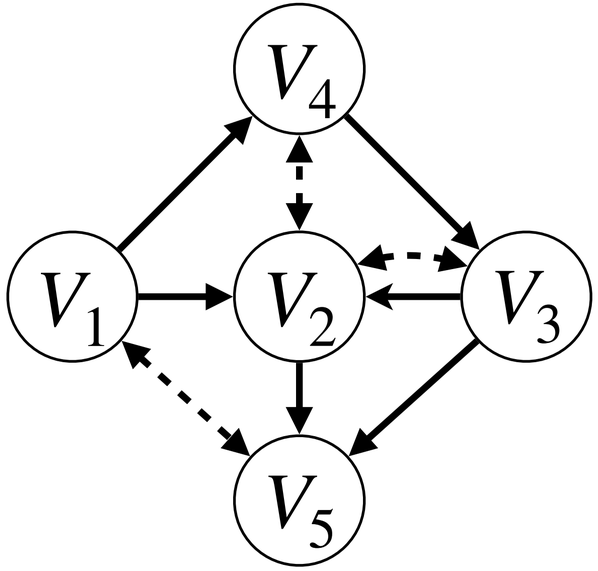



$\langle W, Z_1, Z_2, X, Y \rangle$ is a discriminating path for $X$

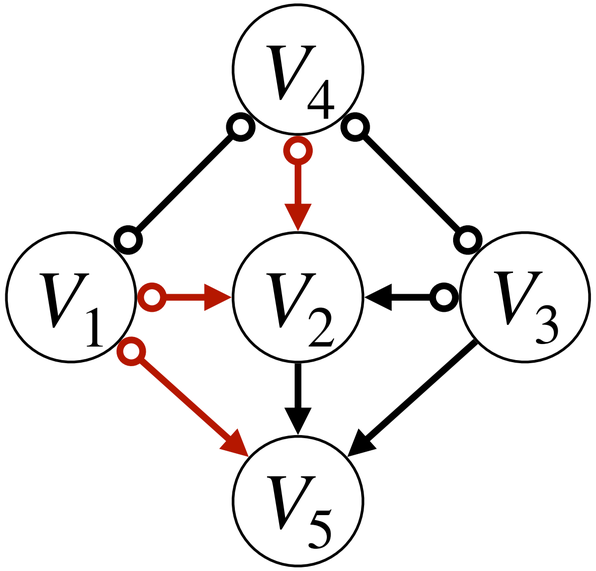
$X \in Sepset(W, Y)$

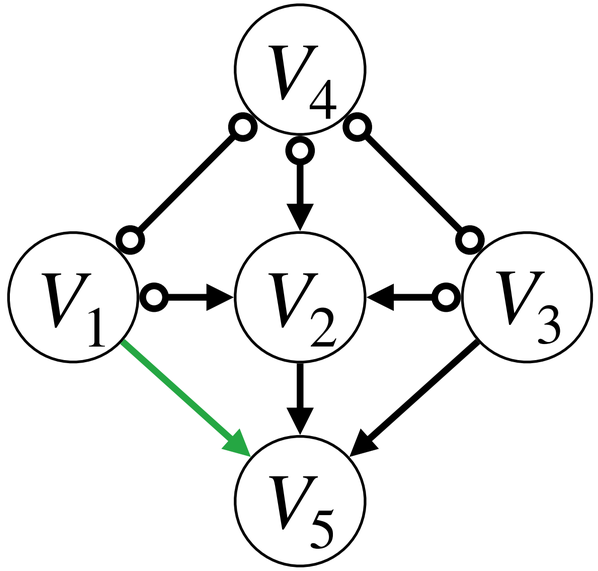
$\Longrightarrow$



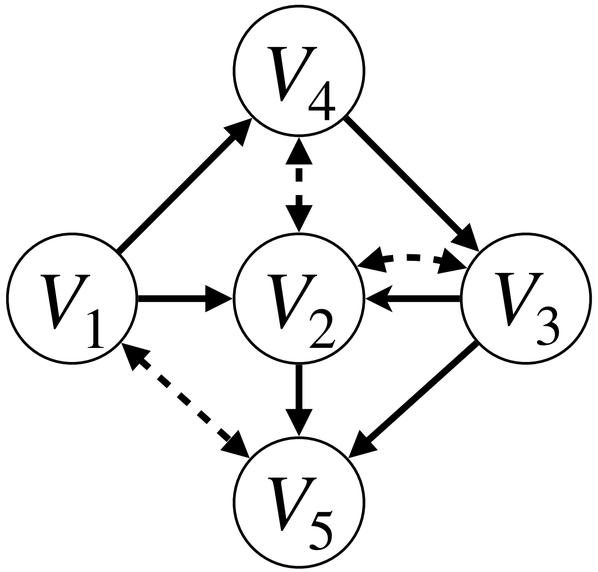

True, unknown ADMG




$\langle V_4, V_2, V_1, V_5 \rangle$ is a discriminating path for $V_1$ and $V_1 \in Sepset(V_4, V_5)$ $-$ $V_4 \perp\!\!\!\perp V_5 \mid V_1, V_2, V_3$

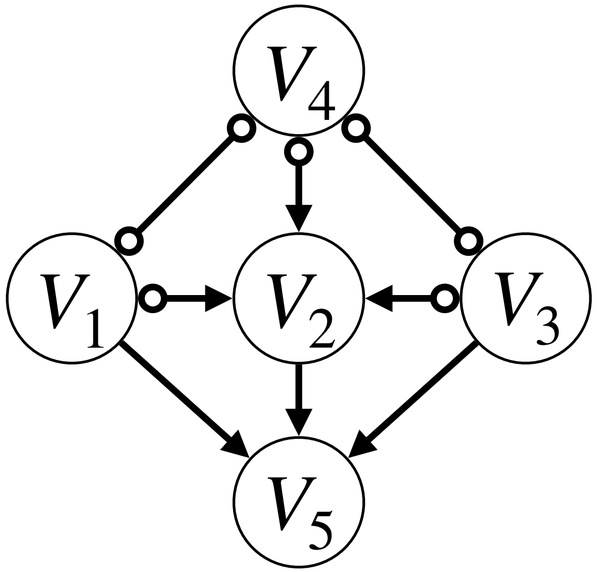
# Final PAG




True, unknown ADMG

$$V_1 \perp\!\!\!\perp V_3 \mid V_4$$

$$V_4 \perp\!\!\!\perp V_5 \mid V_1, V_2, V_3$$




Final PAG

After Skel + R0 + R3 + R1 + R4 + R4

$$V_1 \perp\!\!\!\perp V_3 \mid V_4$$

$$V_4 \perp\!\!\!\perp V_5 \mid V_1, V_2, V_3$$

# More on Causal Discovery

1. Interesting generalizations of the FCI algorithms:
   i. Cyclic Systems
   ii. Multiple Datasets
   iii. Observational + (Soft) Interventional Data
   iv. Time-Series Data
   v. Local Background Knowledge
   vi. Logical Characterization (Conflict Resolution and Background Knowledge)
   vii. Robustness to Violations of Faithfulness

# Some Pointers on Causal Discovery

**Causal discovery from observational and experimental data:**

- Jaber, A., Kocaoglu, M., Shanmugam, K., Bareinboim, E. , 2020. Causal Discovery from Soft Interventions with Unknown Targets: Characterization & Learning. In Advances in Neural Information Processing Systems.
- Li, A., Jaber, A. and Bareinboim, E., 2023, Causal discovery from observational and interventional data across multiple environments. In *Thirty-seventh Conference on Neural Information Processing Systems*.

**Causal discovery from time-series data:**

- Runge, J., Gerhardus, A., Varando, G., Eyring, V. and Camps-Valls, G., 2023. Causal inference for time series. *Nature Reviews Earth & Environment*, *4*(7), pp.487-505.

Vancouver

**Causal discovery from cyclic systems:**

- Mooij, J.M. and Claassen, T., 2020, August. Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. In *Conference on Uncertainty in Artificial Intelligence* (pp. 1159-1168). PMLR.

**Probabilistic Causal discovery with Human in the Loop:**

- da Silva, T., Silva, E., Ribeiro, A., Góis, A., Heider, D., Kaski, S. and Mesquita, D., 2023. Human-in-the-Loop Causal Discovery under Latent Confounding using Ancestral GFlowNets. *arXiv preprint arXiv:2309.12032*.

# 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)

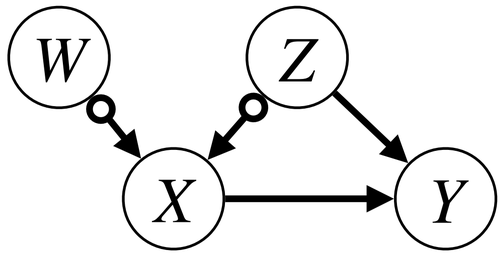
# Adjustment in Markov Equivalence Classes

Identification via Generalized Adjustment Criterion is available in the *pcalg* and *dagitty* R packages.

**1** **Query**

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

**2** **PAG**




**3** **Data**

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

**Adjustment Criterion**

**Solution**

**yes / no**

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

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

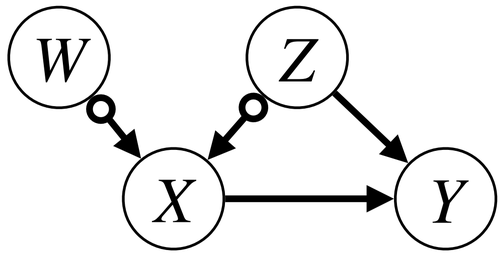
# General Identification in Markov Equivalence Classes

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**

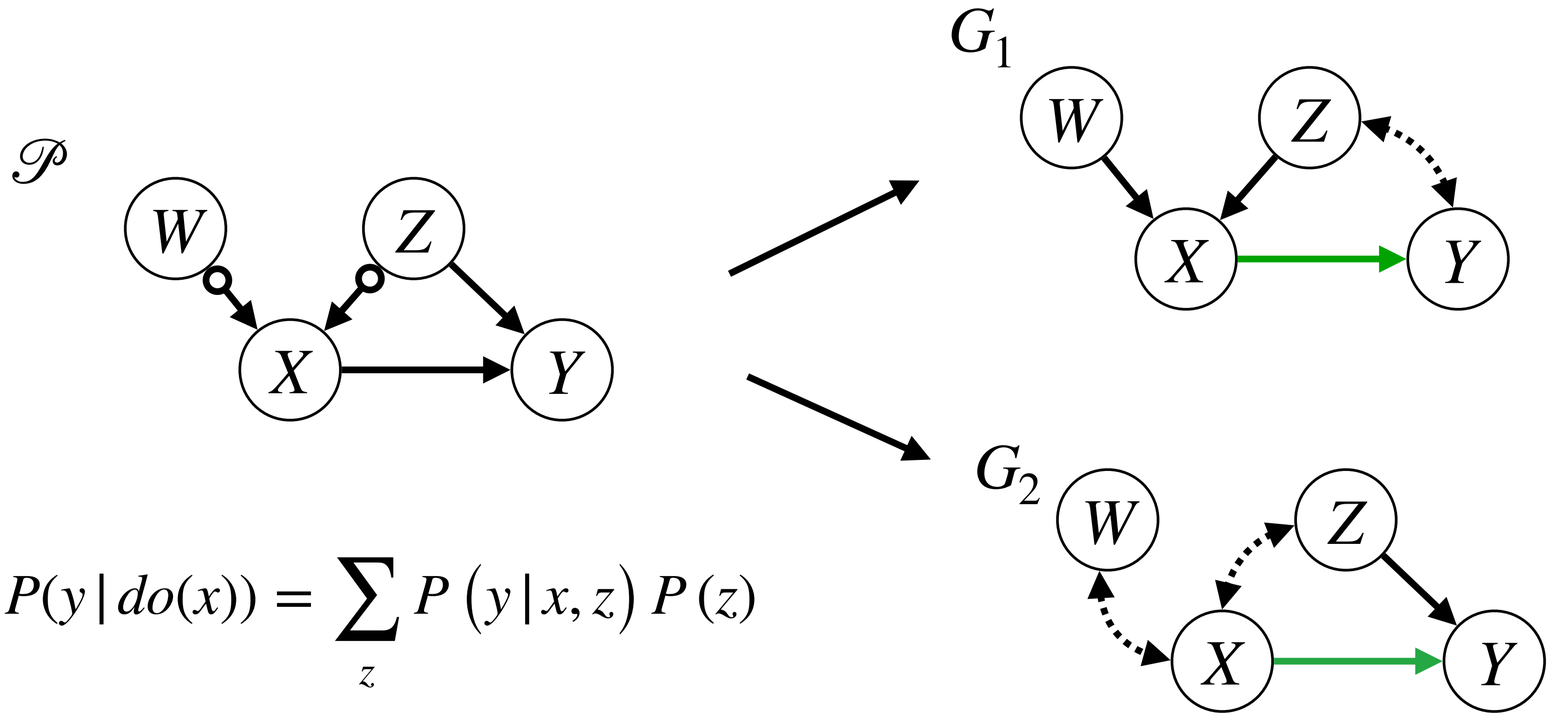
**yes / no**

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

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



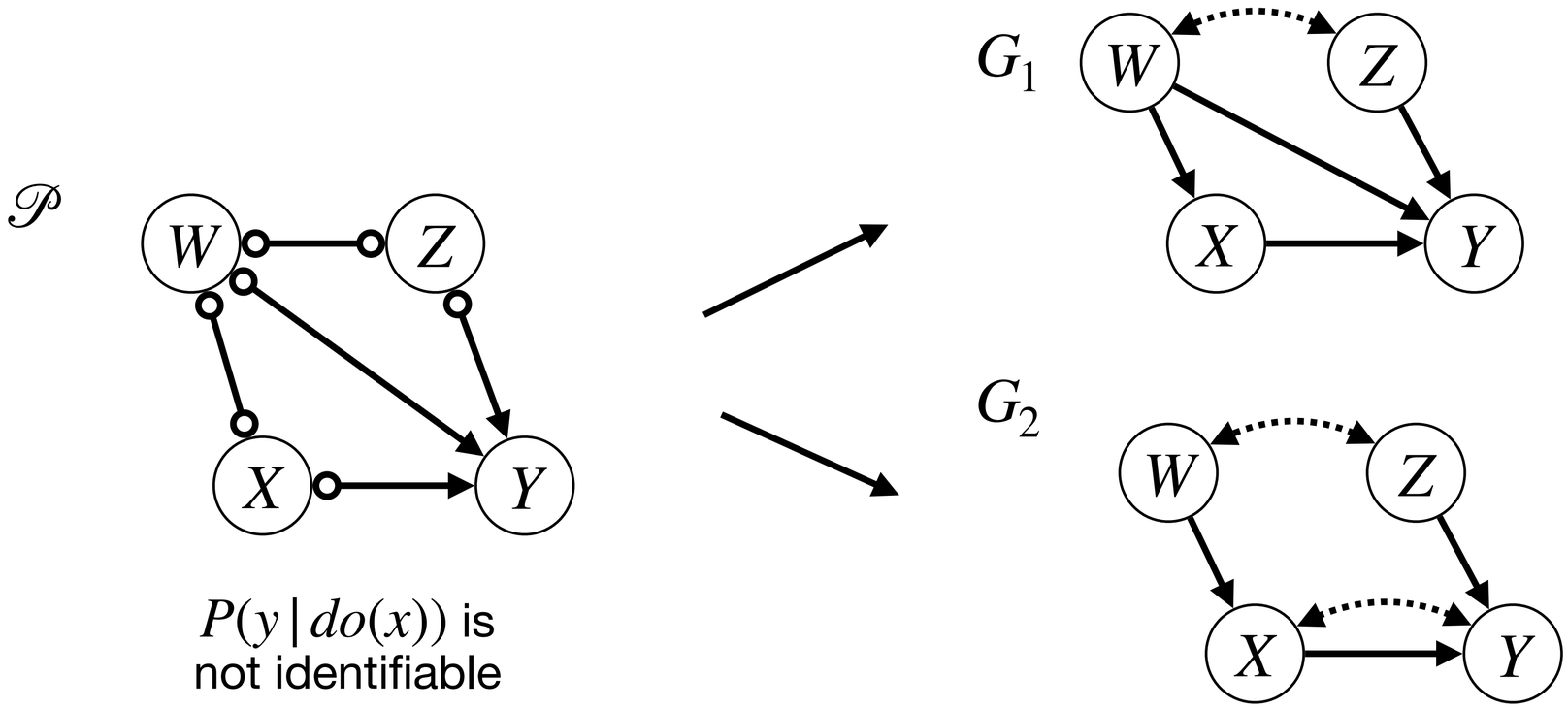

$$P(y \mid do(x)) = \sum_z P\left(y \mid x, z\right) P\left(z\right)$$

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

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

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

# More on Causal Identification

1. Identification and Estimation in Linear Systems.
2. Complete non-identifiability criteria.
3. Generalized Effect Identification:
   - For when surrogate experimental distributions such as $P(\mathbf{y} \mid do(\mathbf{w}))$ are available.
4. Identification of effects from soft-interventions:
   - Hard / $do$-interventions vs soft / $\sigma$-interventions.
5. Partial Identification:
   - For when the effect is not point-identifiable, but an interval for it can be derived.
6. Effect Transportability:
   - From when the target is a different population/domain.

# Other topics

7. Identifiability and Estimation from Deep Neural Networks
8. Counterfactual Identification:
    i. Axioms and Identification Algorithms of $\mathscr{L}_3$ quantities, such as $P(\mathbf{y}_{\mathbf{x}} \mid \mathbf{x}', \mathbf{z})$.
    ii. Mediation Analysis / Fairness
9. More on estimation of the adjustment formula:
    i. Variety of frameworks leveraging machine learning
    ii. Doubly robustness
    iii. Balancing
10. Optimal Causal Experimental Design
11. Causal Reinforcement Learning

# Some Pointers on Causal Identification

**Identification from observational and experimenta 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

Xia, K. , Pan, Y., Bareinboim E.. Neural Causal Models for Counterfactual Identification and Estimation. I*n Proceedings of the 11th International Conference on Learning Representations*. Link

# Coding Exercises

## Causality Tutorial:

- **Code:** https://github.com/adele/Causality-Tutorial/blob/main/main.Rmd
- **HTML Output:** https://github.com/adele/Causality-Tutorial/blob/main/main.html

## Check Part II:

1. Causal Discovery
2. Causal Effect Identification from PAGs

# Thank you! :)

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

adele.ribeiro@uni-marburg.de