Introduction to Causal Inference

Adèle H. Ribeiro

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

Faculty of Mathematics and Computer Science
Philipps-Universität Marburg

Institute of Medical Informatics
University of Münster

Lisbon Machine Learning School (LxMLS)
July 17, 2024

Predictive vs Explainable, Trustworthy AI

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

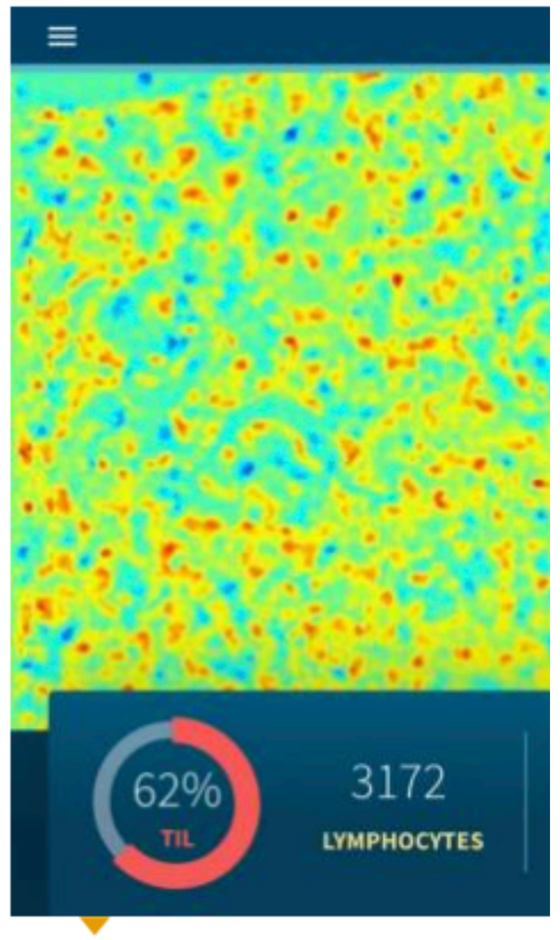
itn IMAGING TECHNOLOGY NEWS

COVID-19 IMAGING WOMEN'S INFORMATION RADIATION

NEWS | ARTIFICIAL INTELLIGENCE |

Making the Role of AI

Analysis system for the diagnos



Detection of tumor-infiltrating lym generate a heatmap showing TILs (of Klauschen/Charité

nature

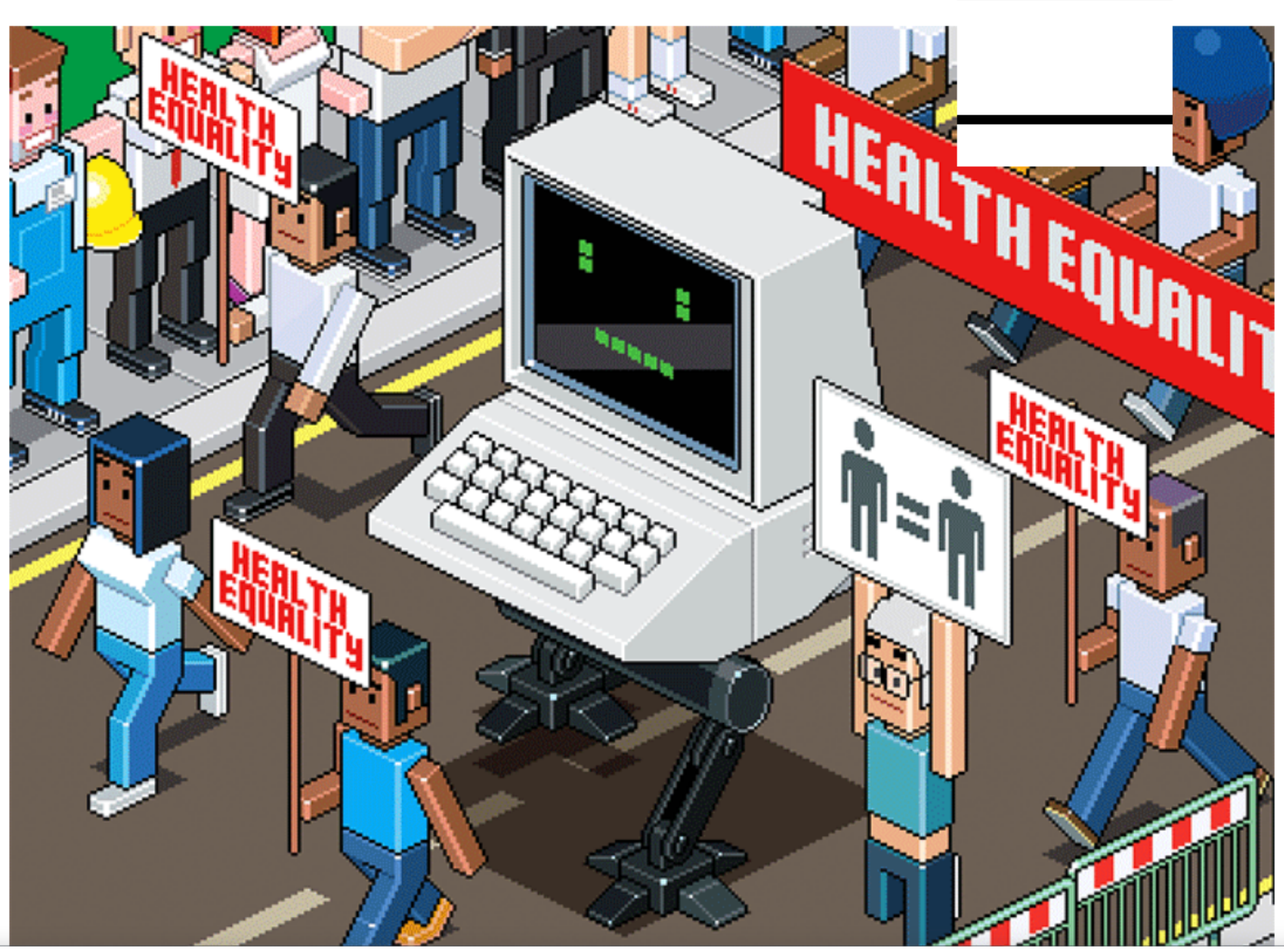
Explore content Journal information Publish with us Subscribe

nature > outlook > article

A fairer way forward for AI in health care

Without careful implementation, artificial intelligence could widen health-care inequality.

Linda Nordling



Predictive vs Explainable, Trustworthy AI

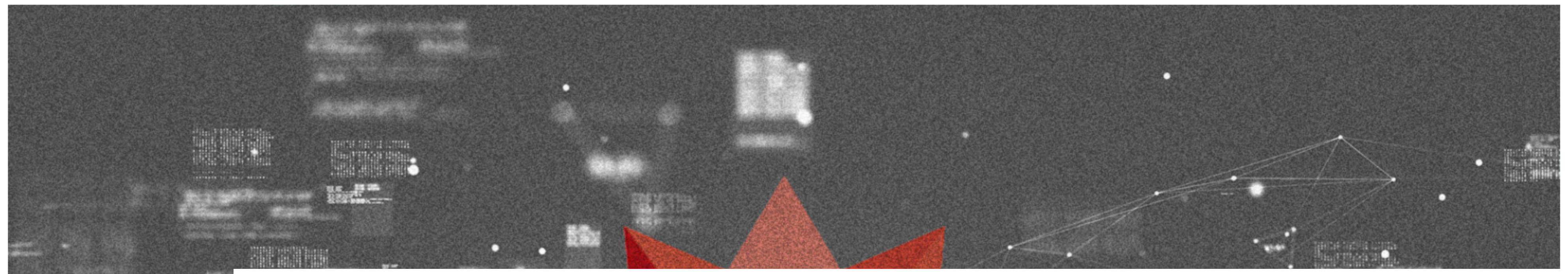
Chat GPT - Impressive Abilities:

- Debug code
- Play Games
- Translate text
- Explain complex topics
- Create AI art
- Grade essays
- Write a resume
- Brainstorm ideas
- Play games with ChatGPT
- Create chatbots
- Language tutor
- Create jokes
- Solve coding errors
- Write poetry

06-08-2024 | TECH

This classic answer engine still outsmarts AI chatbots

For questions involving hard data and math calculations, 15-year-old WolframAlpha is a fast, accurate alternative to inaccurate AI chatbots.



[Image: RICHARD JON

BY CHR

Medium

Search

GPT-4 still can't reason. But it's a phenomenally useful tool anyway.

Konstantine Arkoudas · Follow

12 min read · Apr 11, 2024

Causality: A Missing Link to Reasoning in AI

The ability to understand cause-and-effect relationships is crucial for deeper understanding and decision-making processes.

nature

Explore content | About the journal | Publish with us | Subscribe

nature > outlook > article

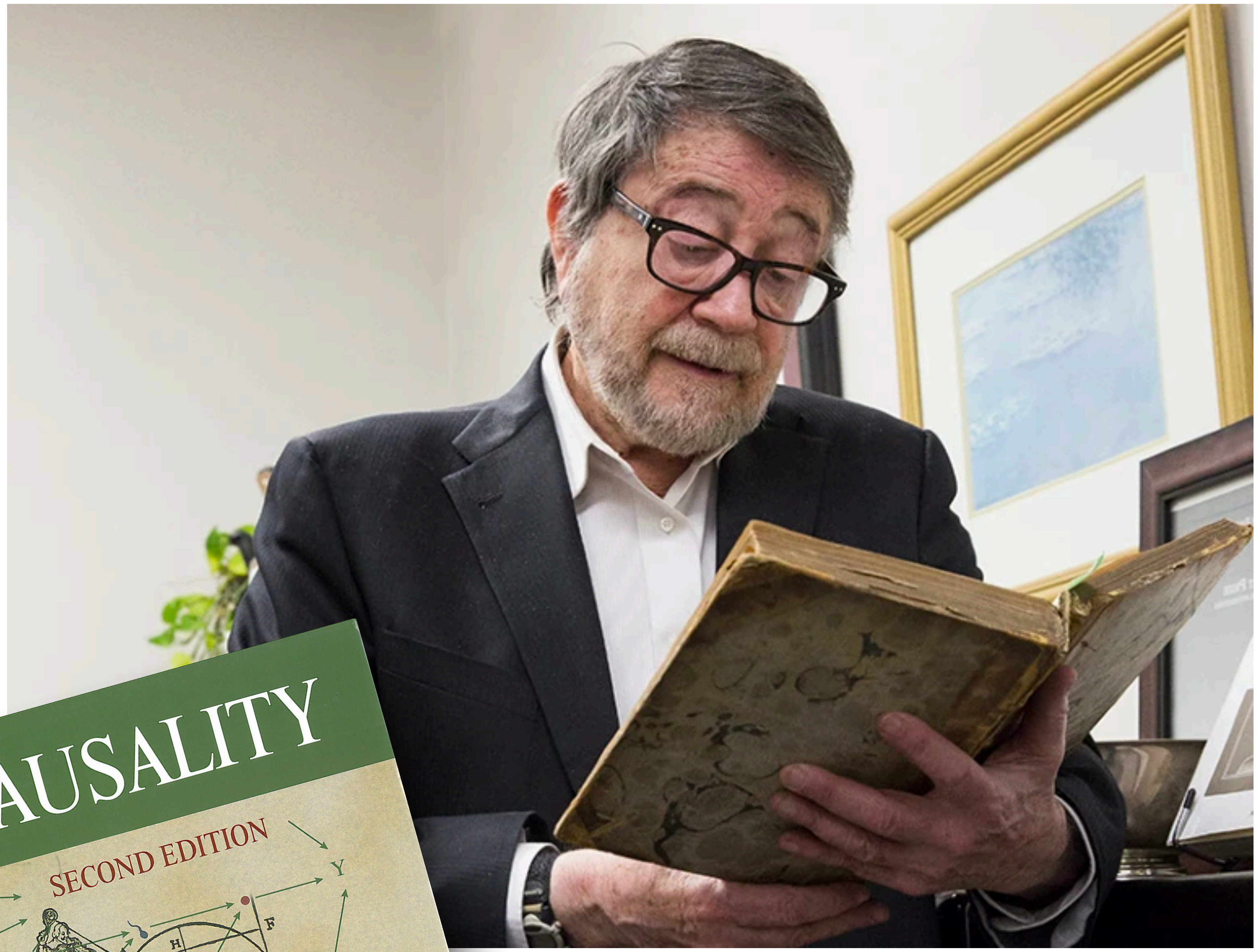
OUTLOOK | 24 February 2023

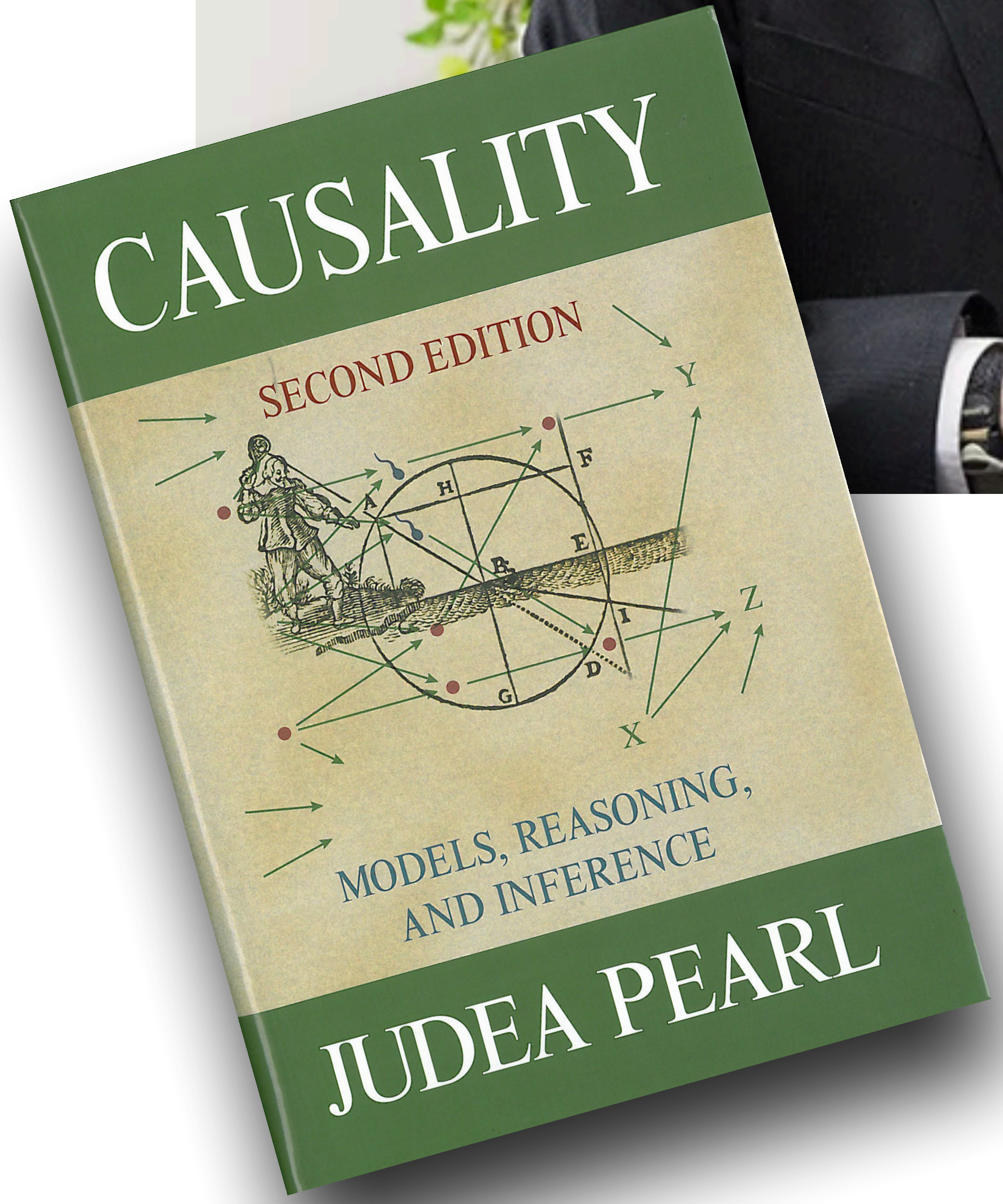
Why artificial intelligence needs to understand consequences

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

The Mathematical Framework of Causal Data Science

Judea Pearl — Causality





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

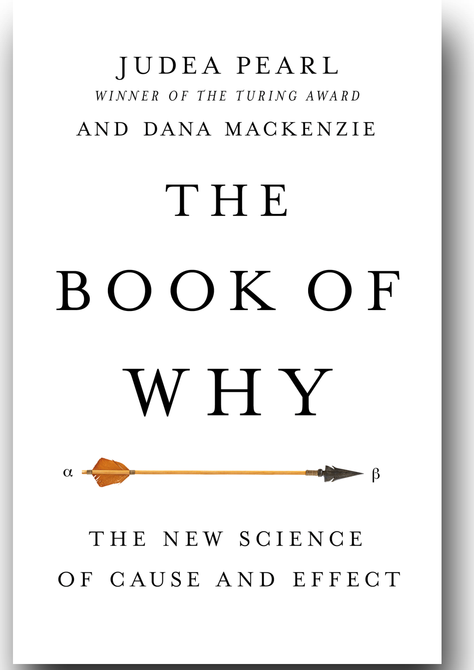
In 2011, he won the A. M. Turing Award (the highest distinction in computer science and a $250,000 prize)

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

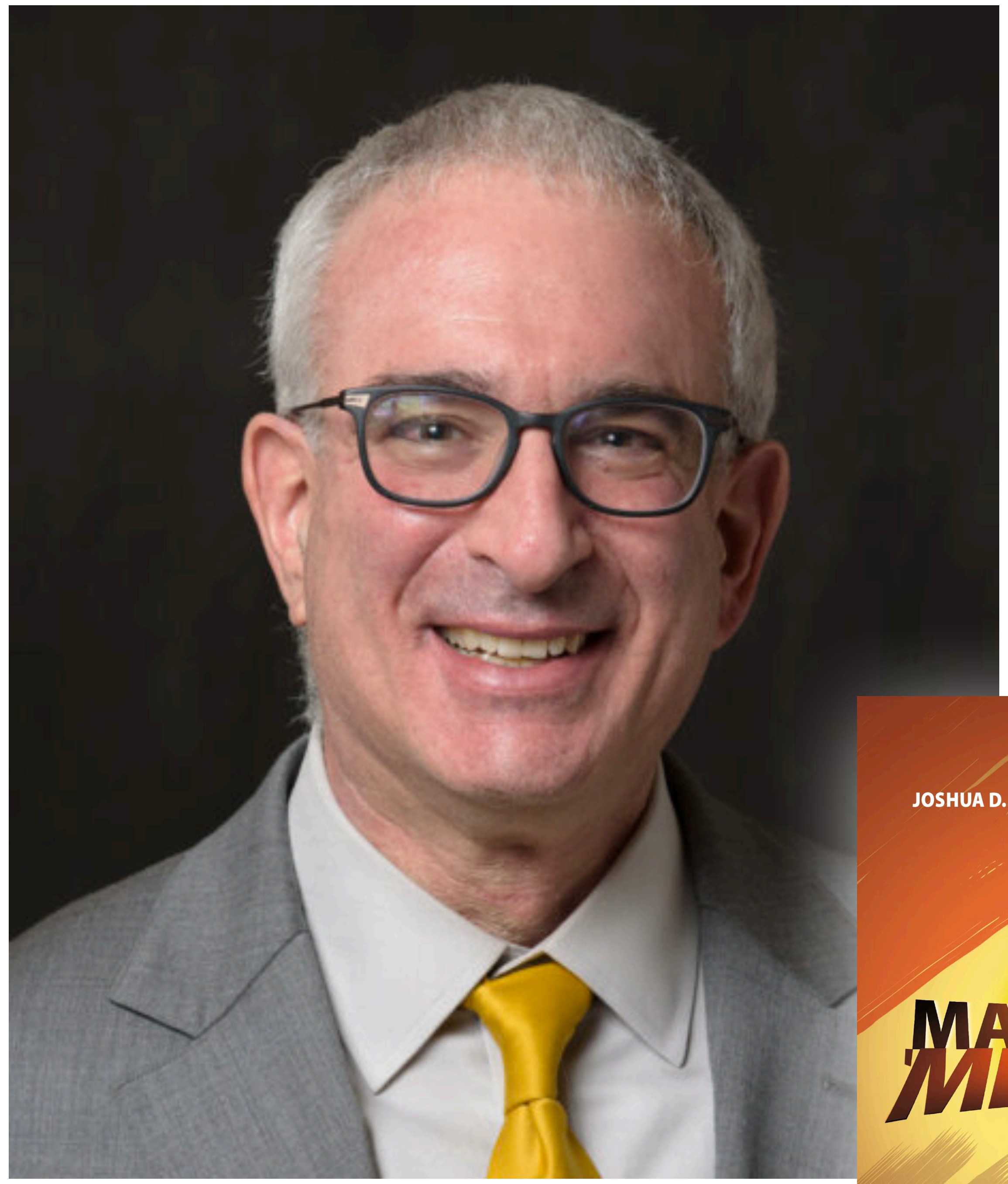
— Association for Computing Machinery (ACM)

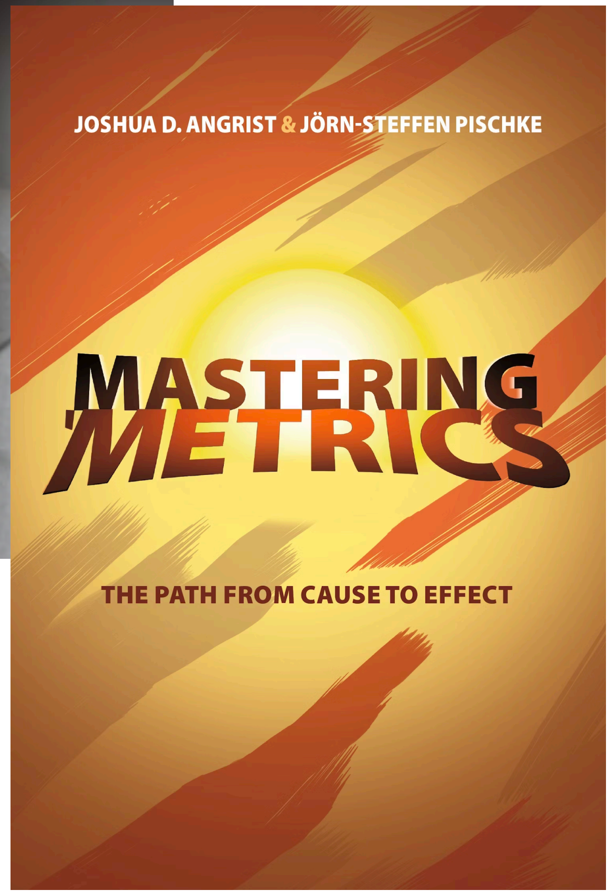
"Deep learning has instead given us machines with truly impressive abilities but no intelligence. The difference is profound and lies in the absence of a model of reality."

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



Guido W. Imbens, Joshua D. Angrist & Donald B. Rubin



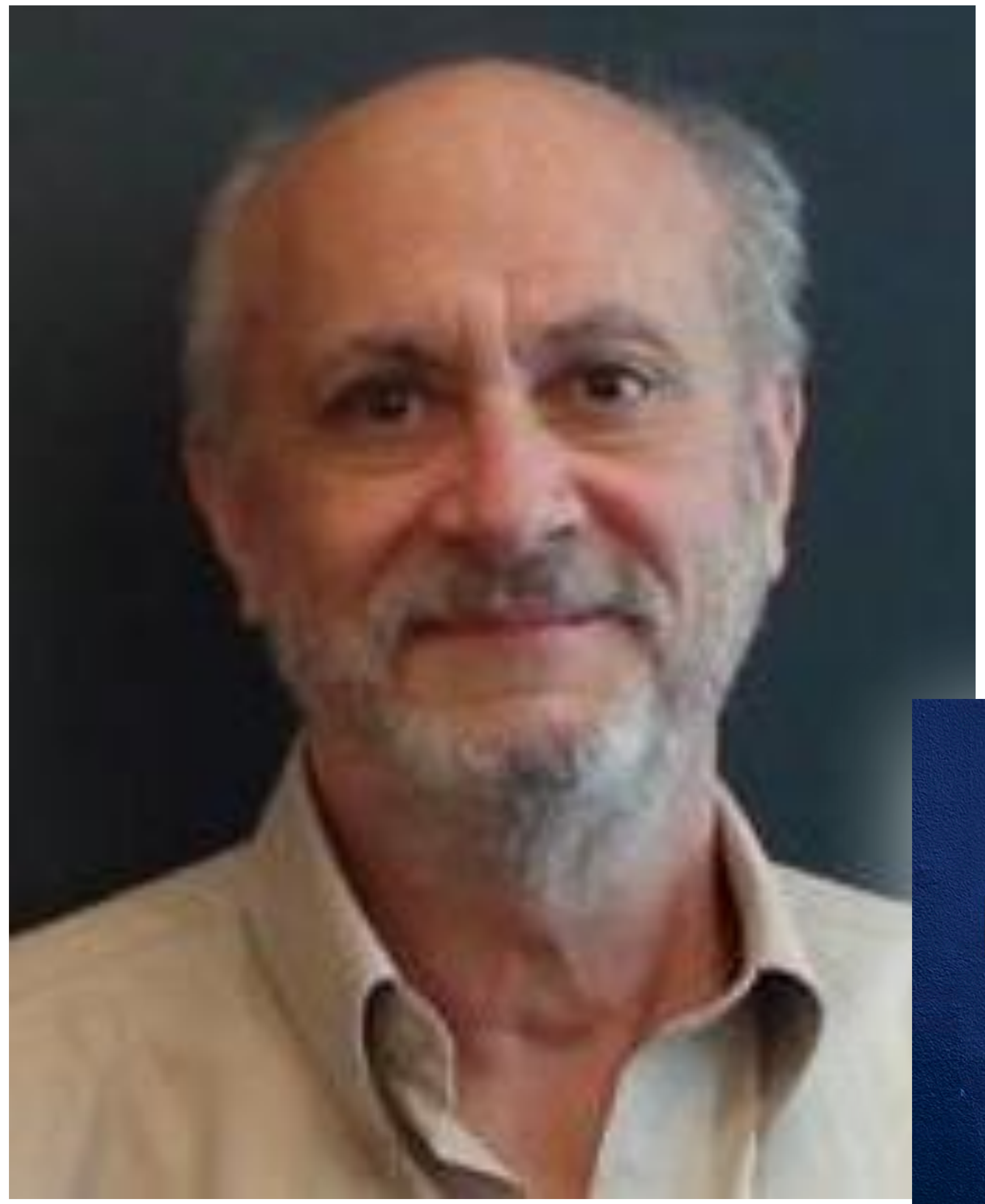


Joshua D. Angrist

Professor of
Economics at MIT

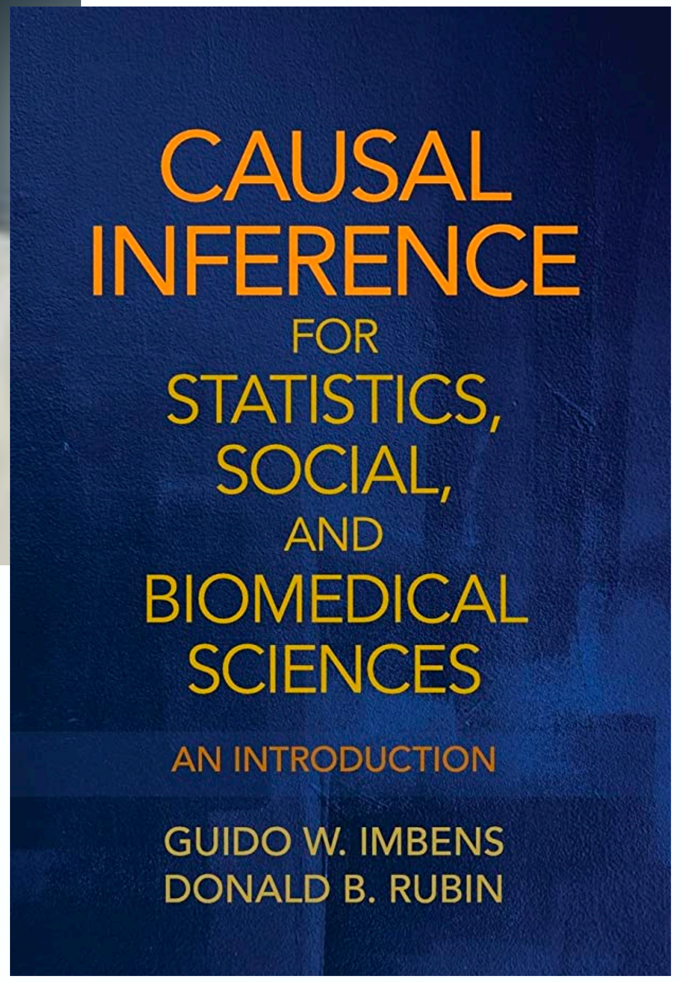
Guido W. Imbens

Professor of Applied
Econometrics at
Stanford University



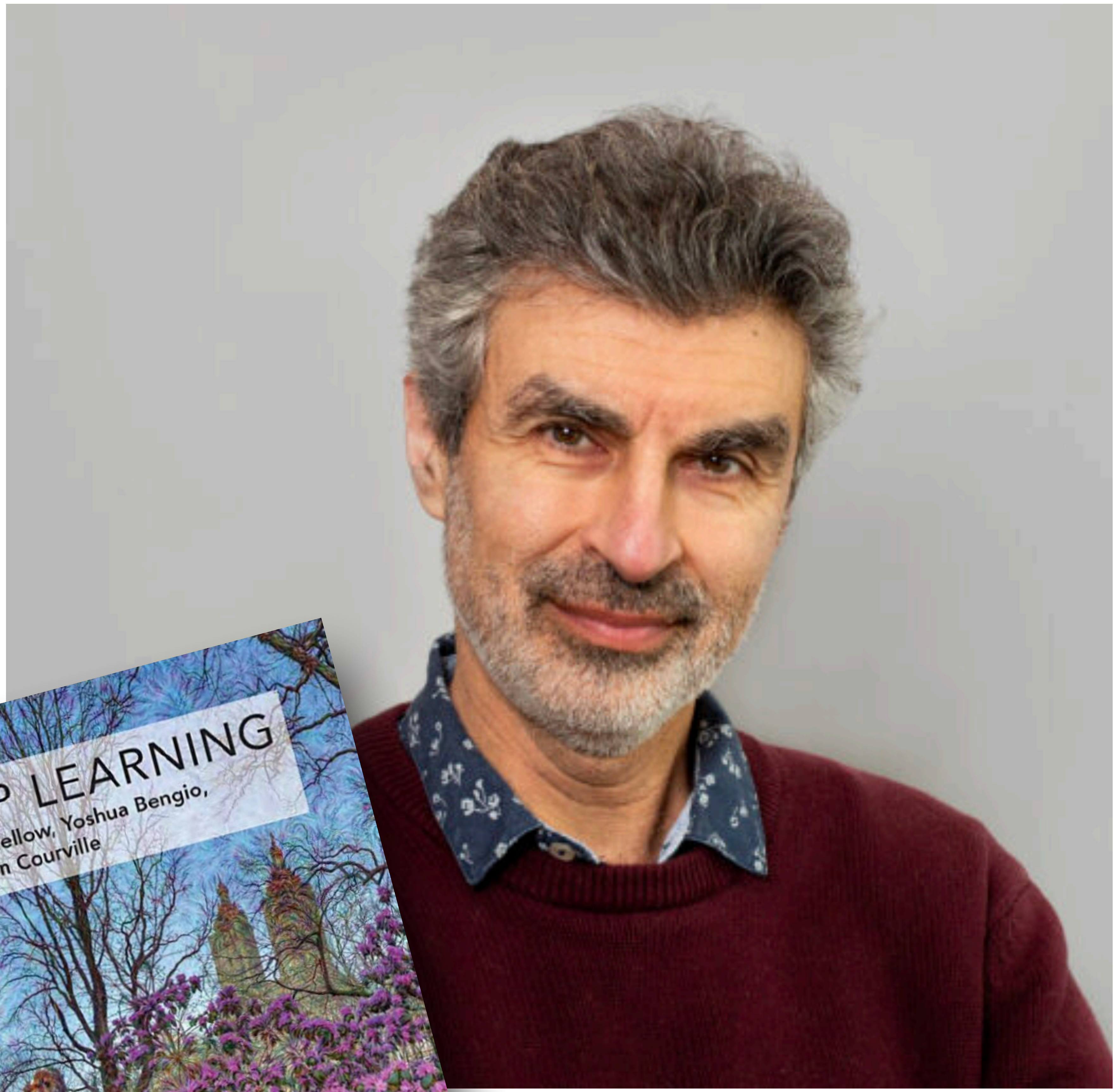
Donald B. Rubin

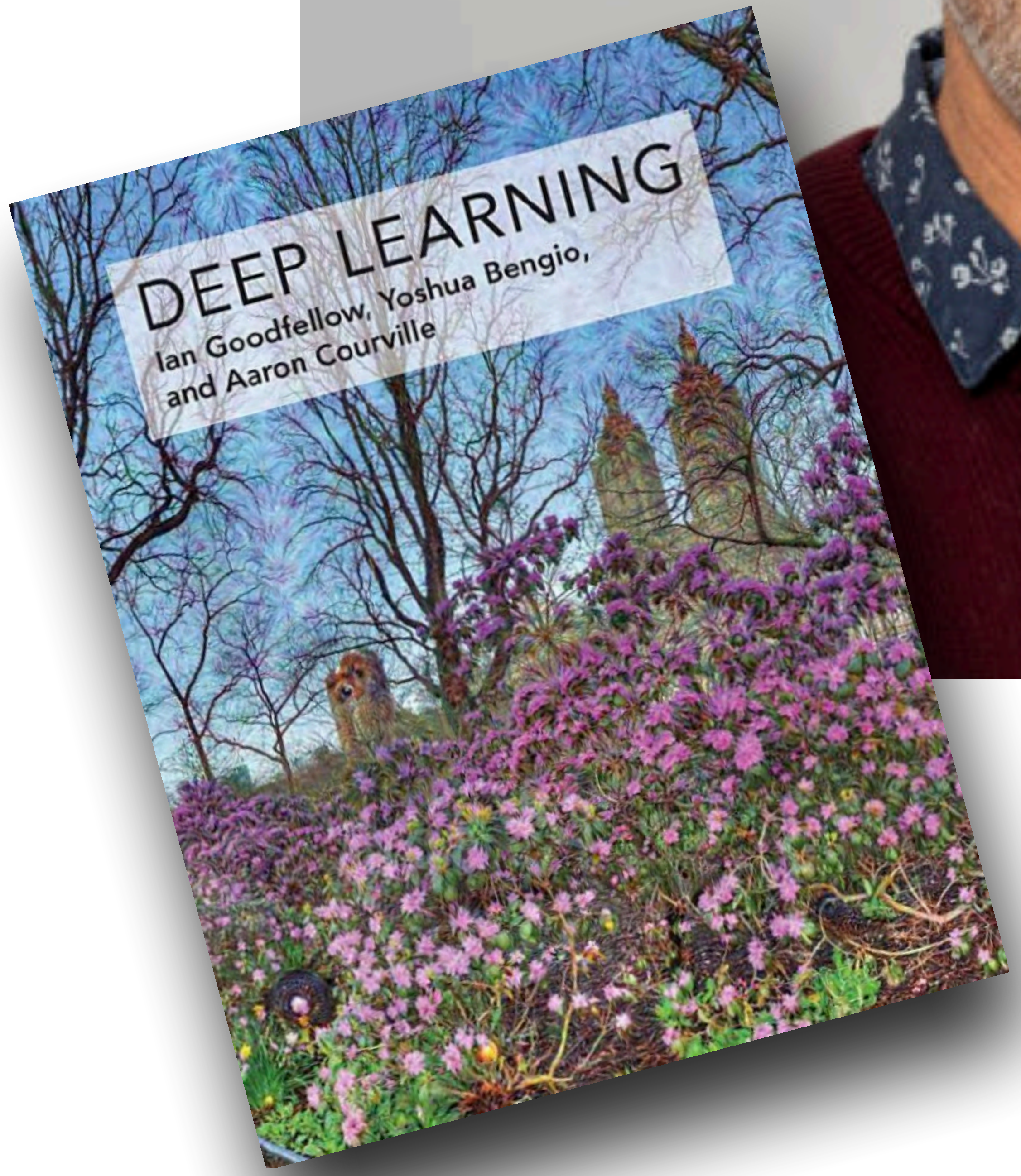
Professor of
Statistics at
Harvard University



In 2021, Angrist & Imbens won the <u>Nobel Prize in Economics</u>
“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

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**

Explainability: provides a better understanding of the underlying mechanisms

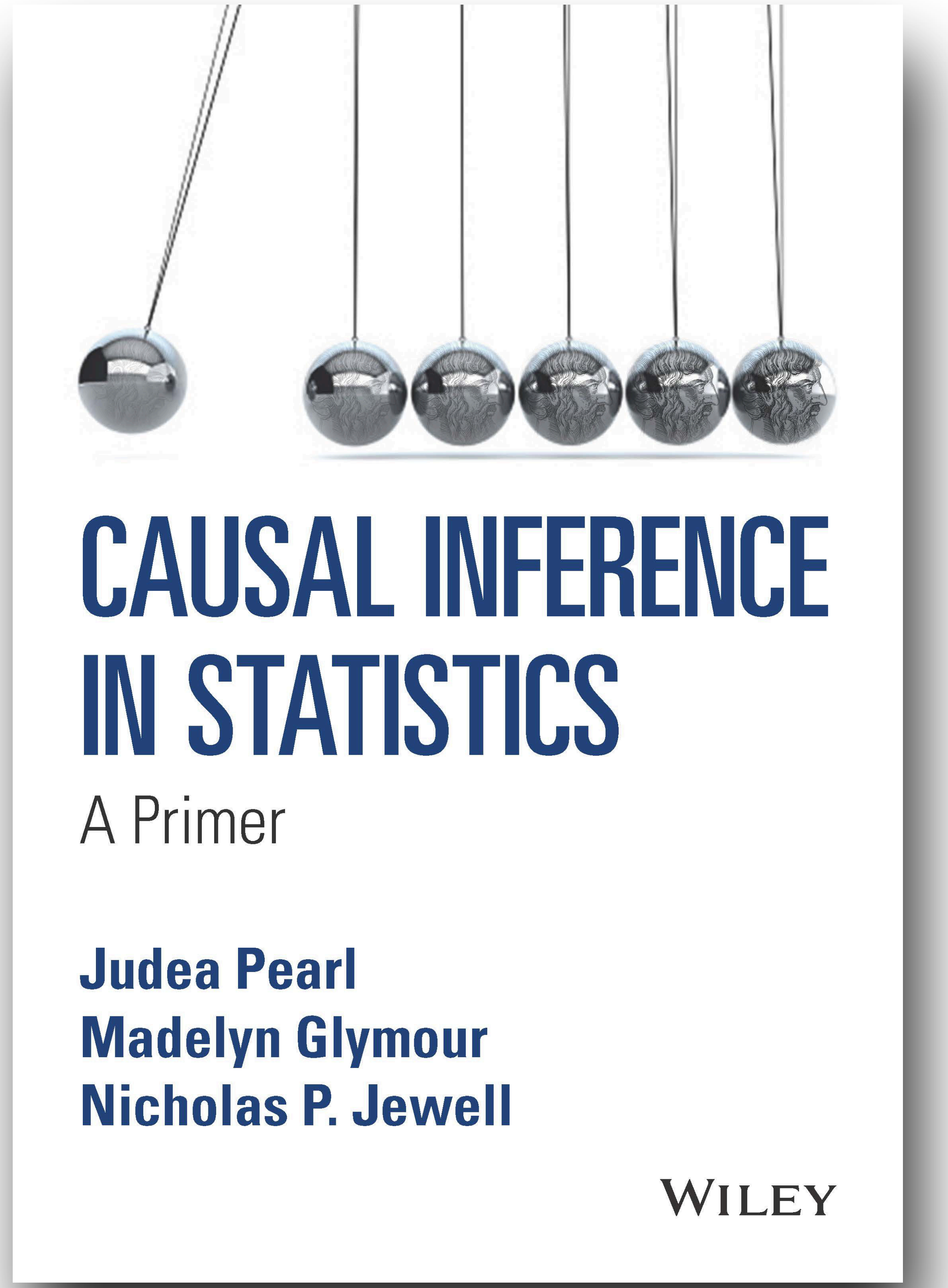
- **Causal Discovery**

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

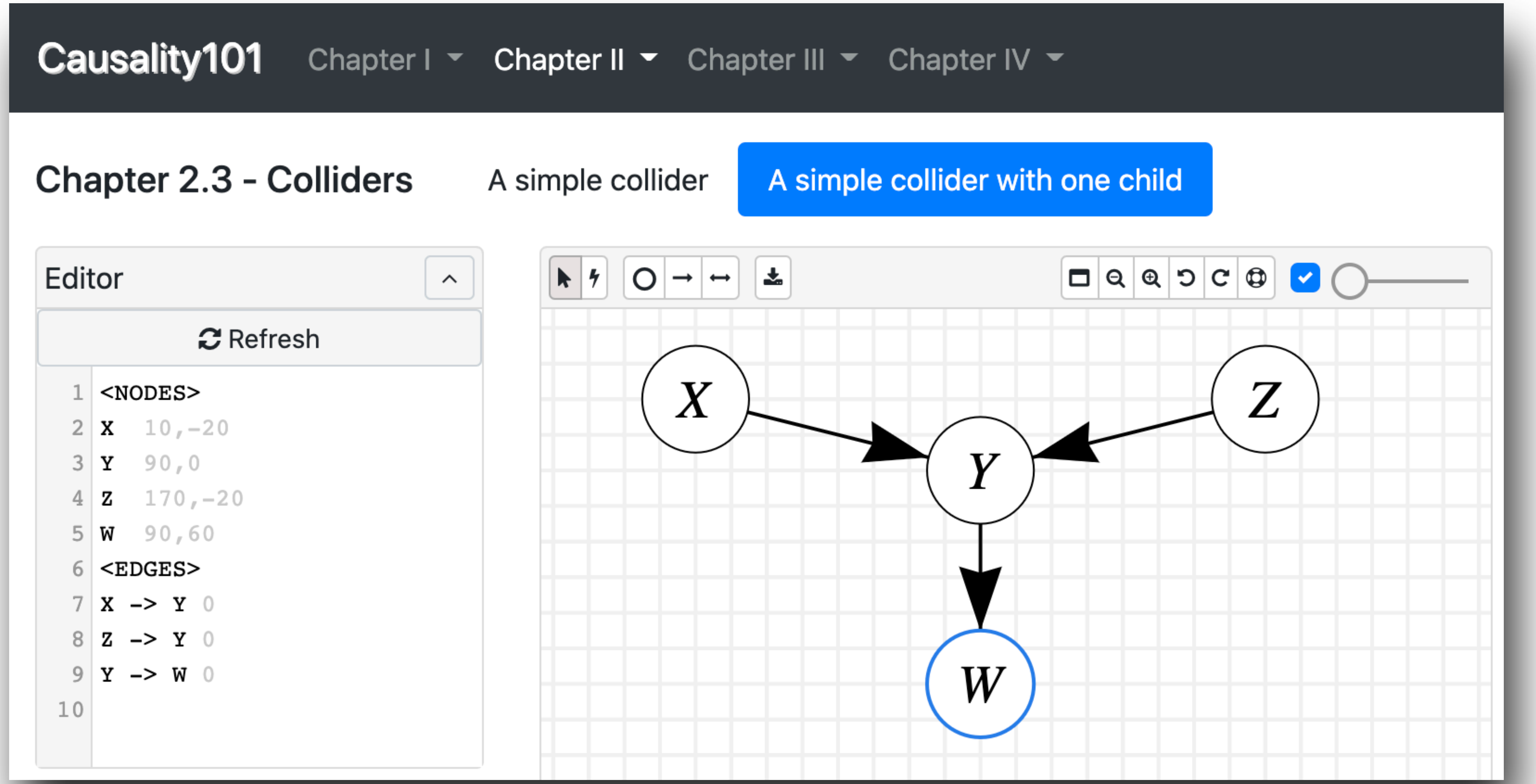
Generalizability: allows the transportability of causal effects across different domains.

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

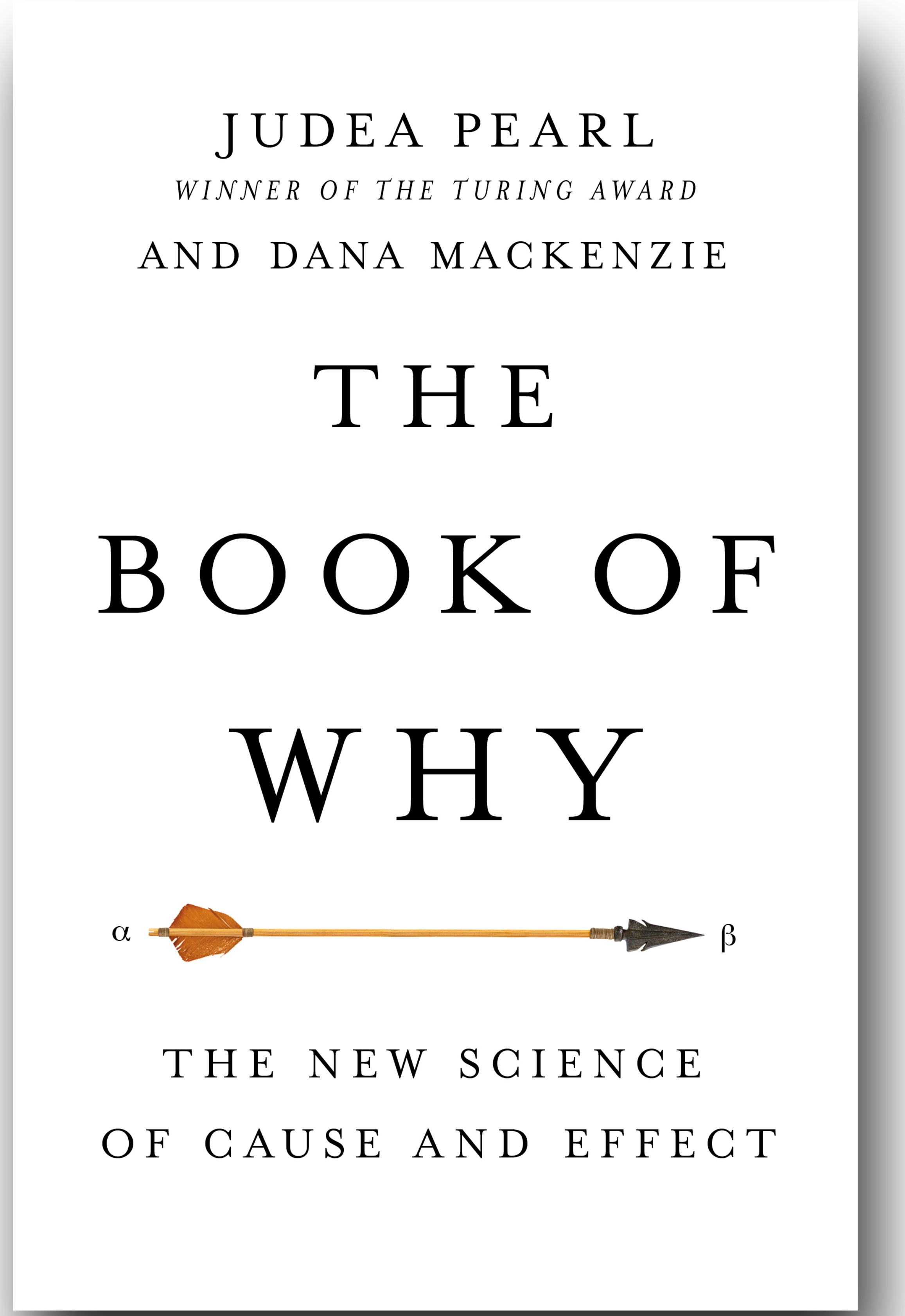
Causality Theory by Judea Pearl

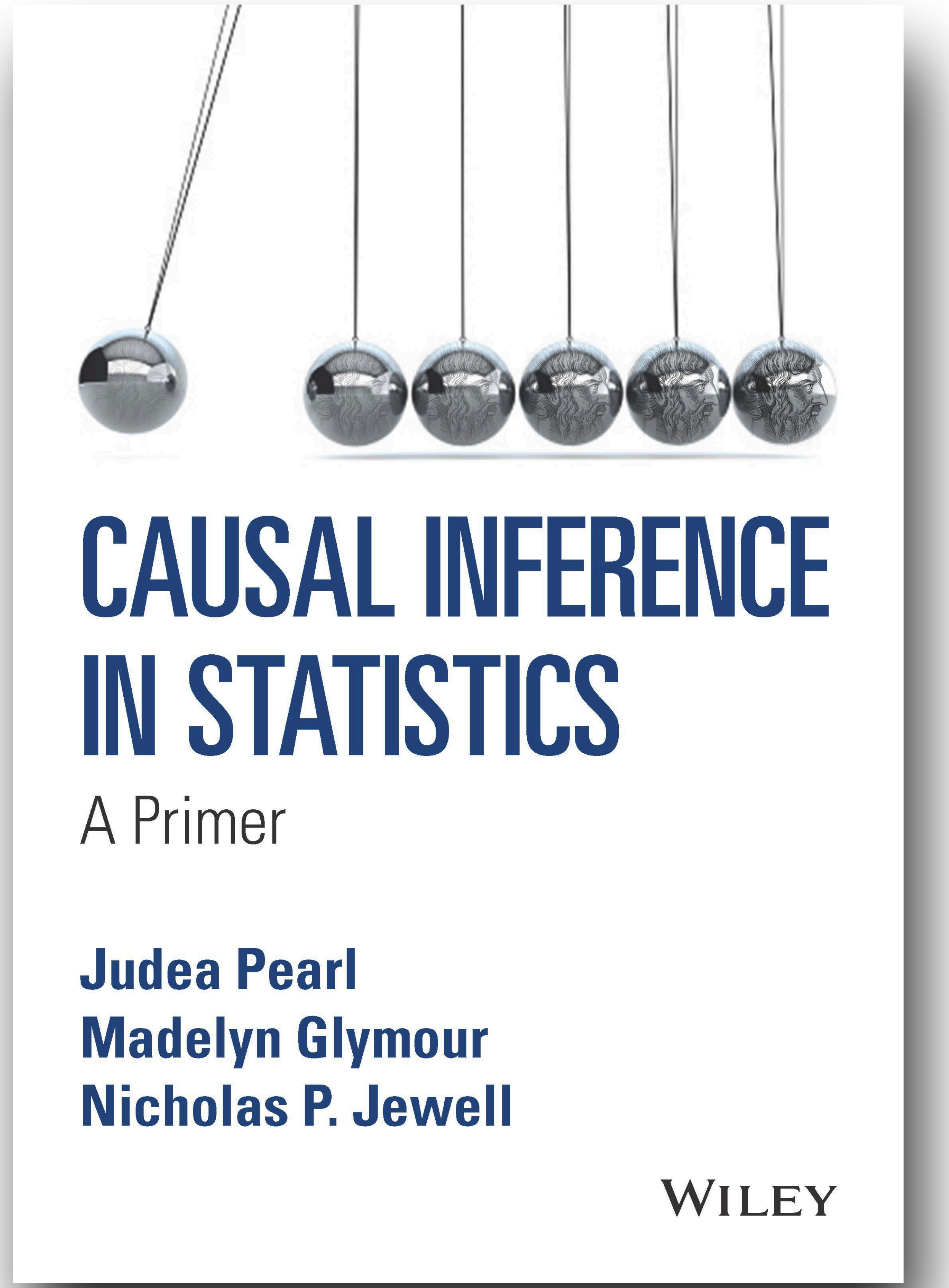


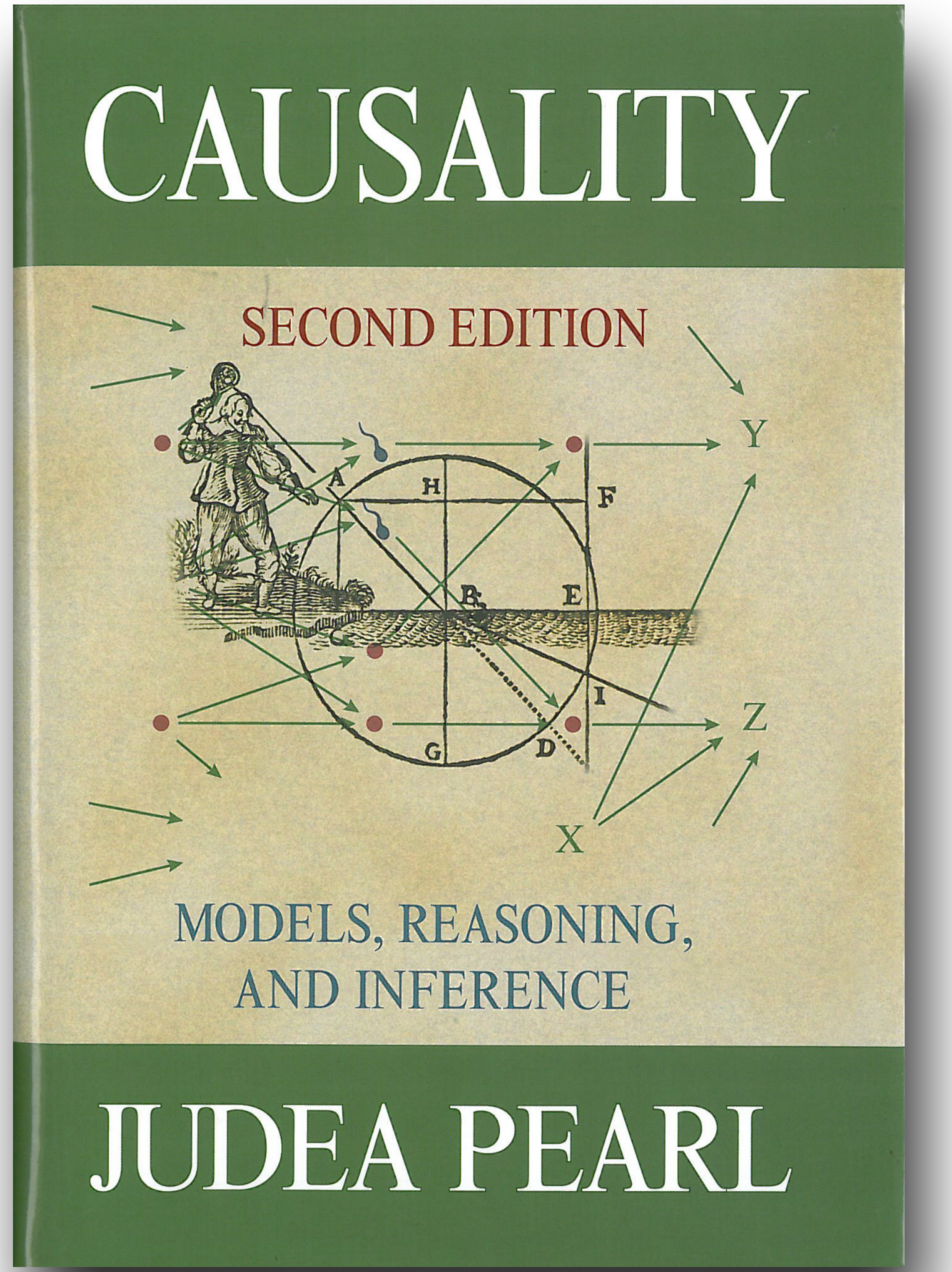
https://causality101.net/



Causality Theory by Judea Pearl







Prediction vs Effect of Interventions

Statistical Association vs Causation

Predictive Tasks

Task: Can I **guess** the size of a fire by **observing** the number of firefighters?

Yes!

X: **Number of firefighters in action**

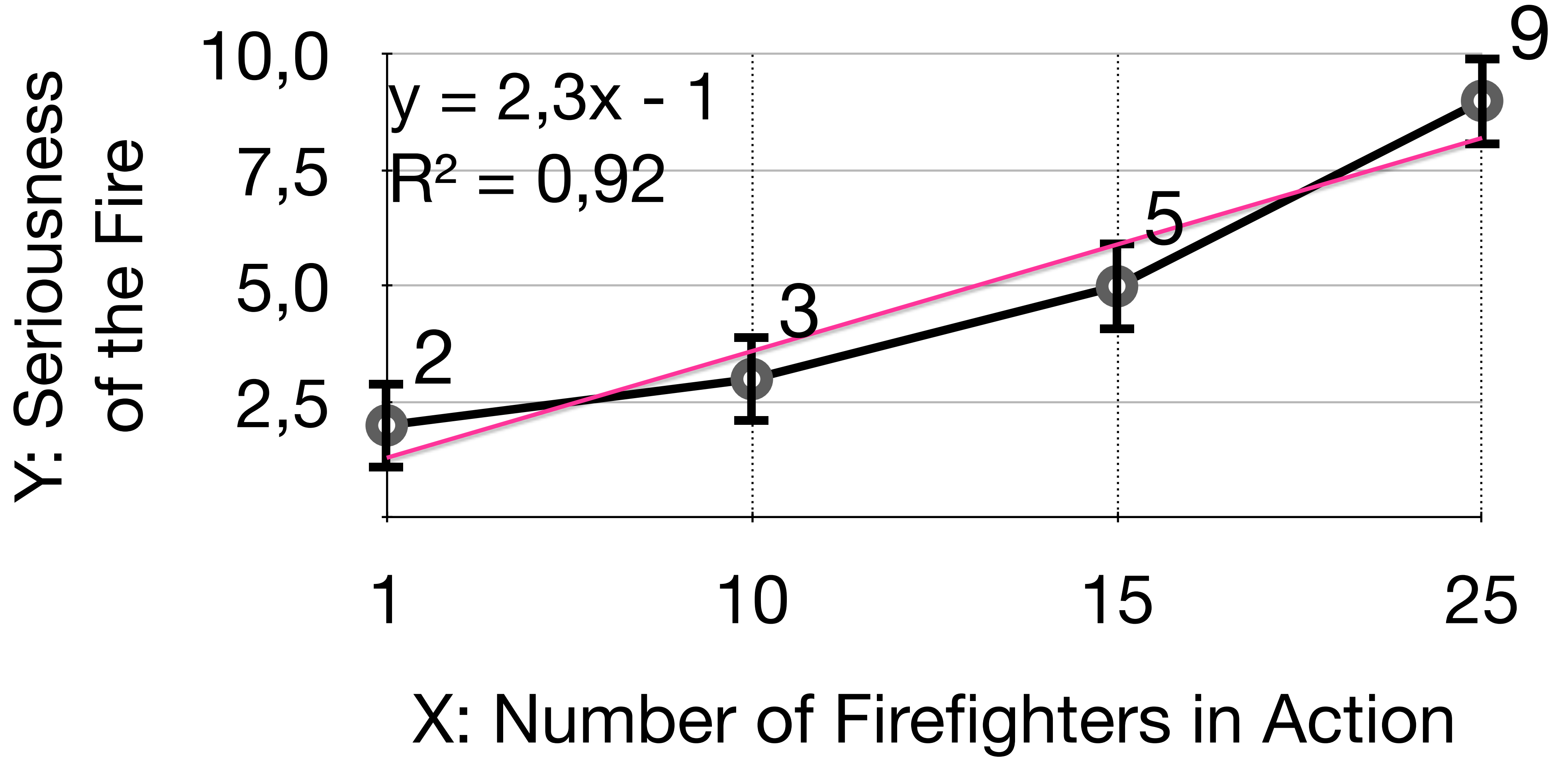
Y: **Size of the (initial) fire**

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

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

Observational Probability Distribution

Correlation between severity of fire and number of firefighters in action



Positive Correlation:

The more firefighters, the stronger the fire!

Prediction ⇒ Decision-Making?

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

Misleading correlation: It is the size of the fire that determines the number of firefighters needed, not the other way around.

Causal Effect ≡ Effect of an Intervention

The causal direction is determined by understanding the underlying reality.

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

Changing the number of firefighters through an action/intervention on X, $do(X = x)$, does not affect the initial size of the fire (Y).

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
  - $U_i \subseteq \mathbf{U}$ 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.

Structural Equation Model (SEM)

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

- **Pre-specified causal order**
- **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).

Statistical Association vs Causation

Pre-Interventional/ Observational SCM

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

Observational Distribution

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

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

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

$do(X = x)$

$\neq$

Post-Interventional / Interventional SCM

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

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 ***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

Statistical Association vs Causation

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

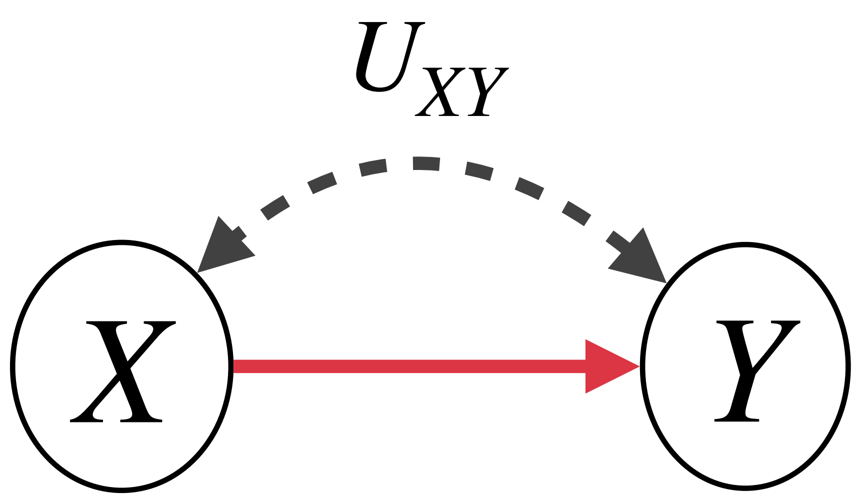
$$\mathcal{M}_x = \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_{XY}, U_X, U_Y\} \\ \mathcal{F} = \begin{cases} X = x \\ Y = f_Y(x, U_Y, U_{XY}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Loss of Information

Observational Data

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

Observational Distribution



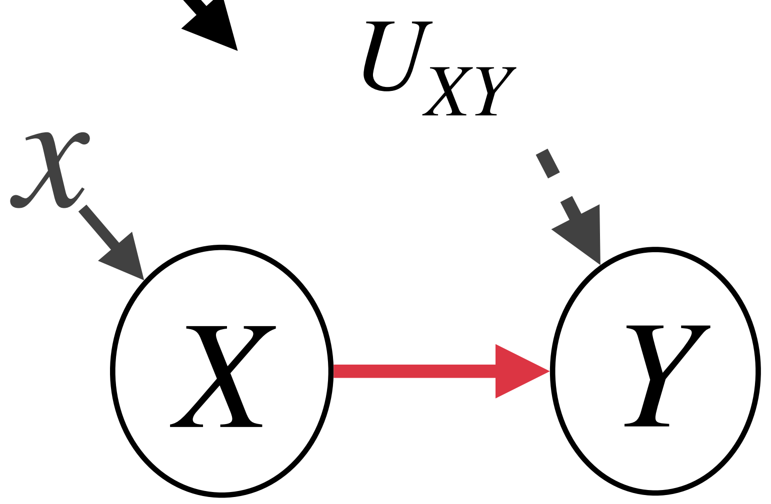
Observational Causal Diagram

Loss of Information

Interventional Data

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

Interventional Distribution

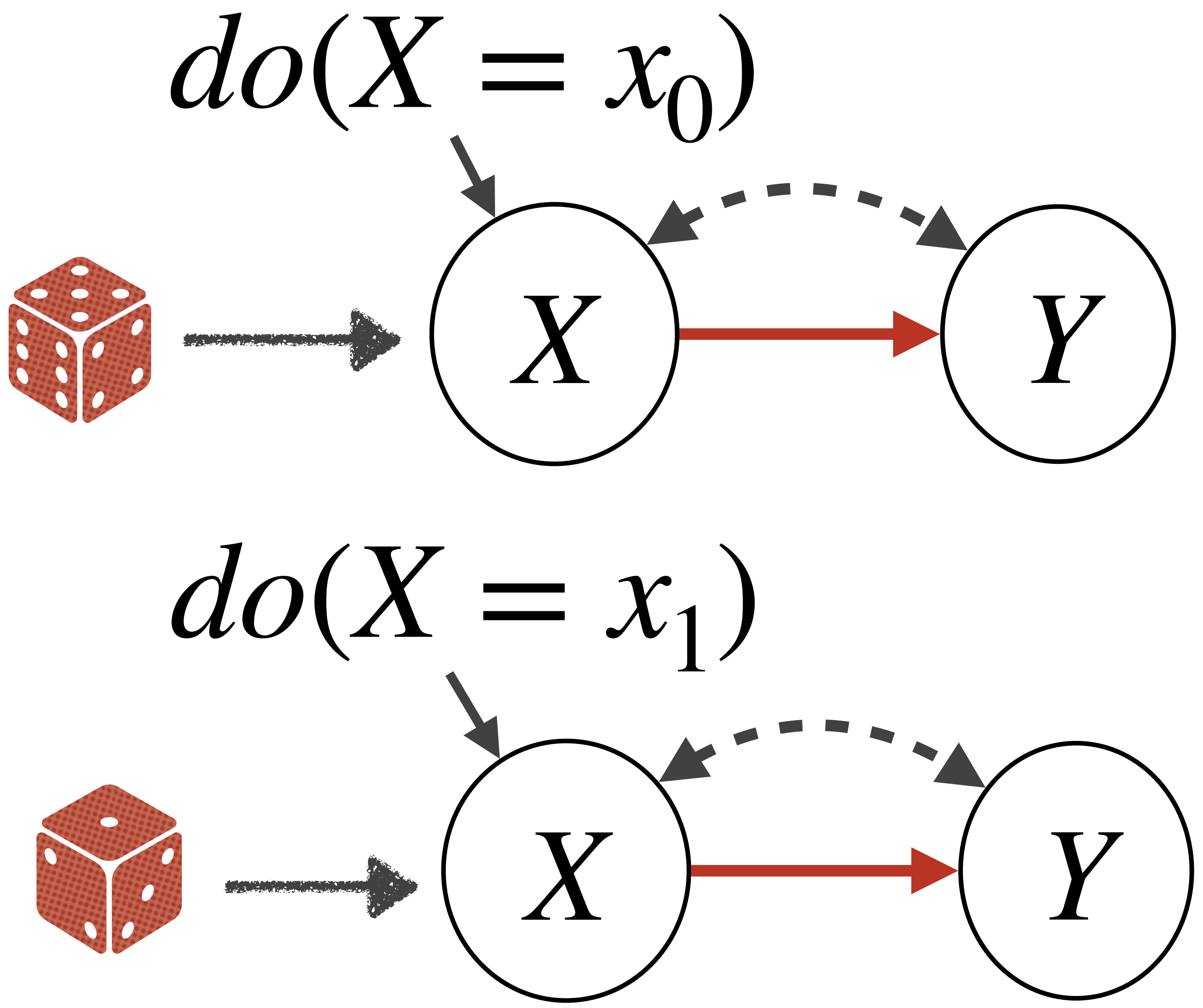


Interventional Causal Diagram

Randomized Experiments

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

Randomization of the X's assignment



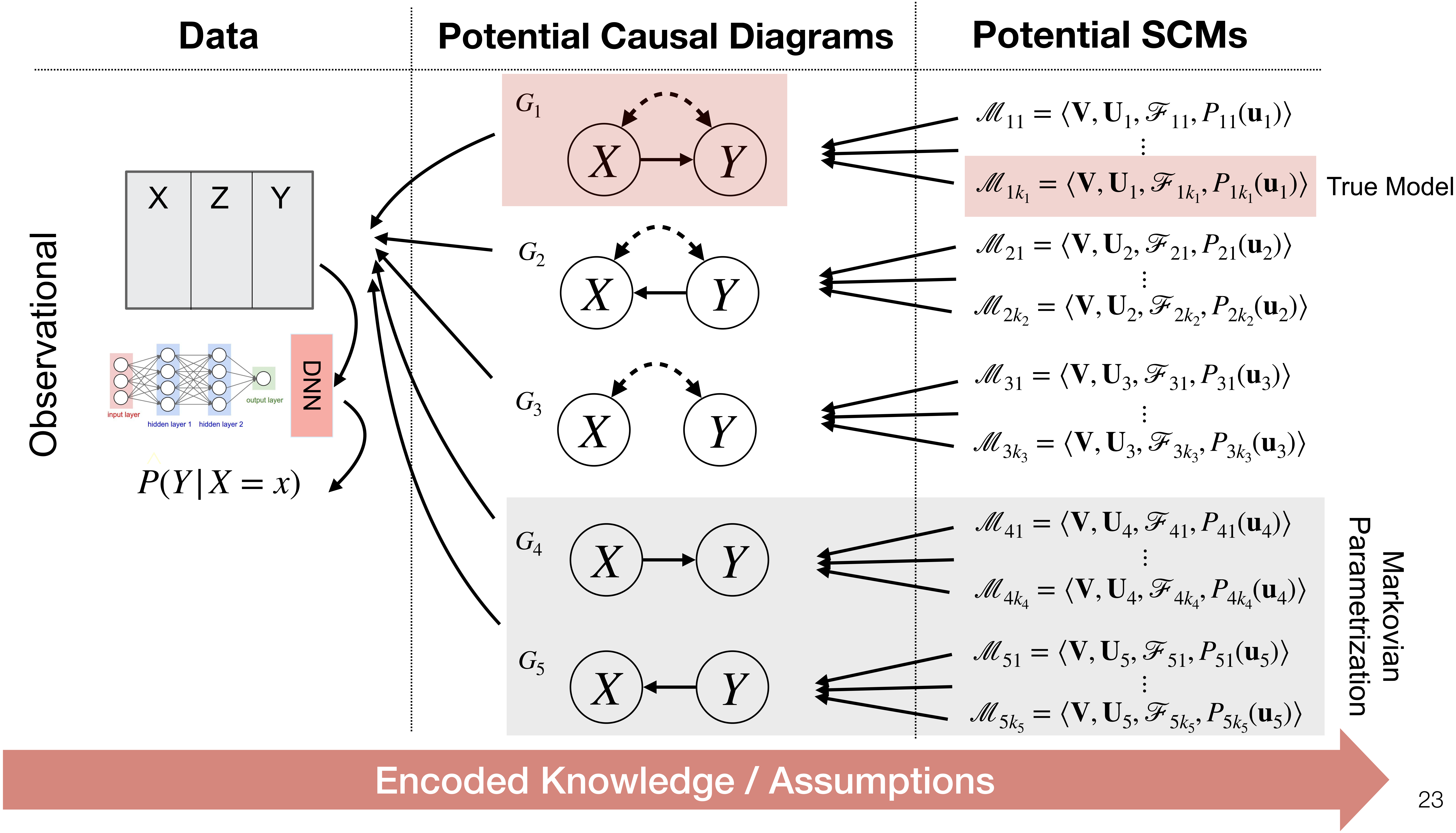
$\mathbb{E}[Y \mid do(X = x_0)]$

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

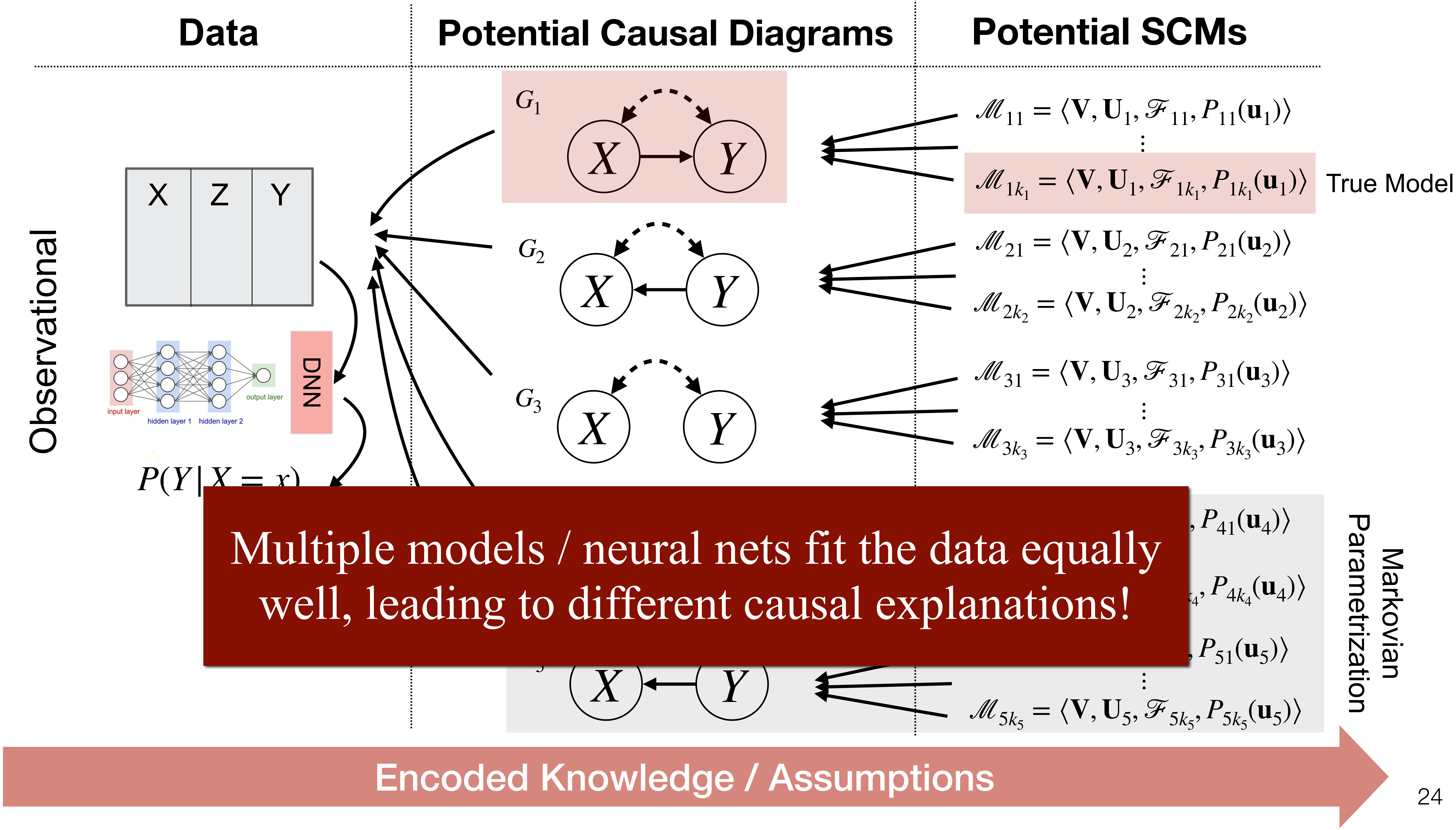
Average Causal Effect: $\mathbb{E}[Y \mid do(X = x_0)] - \mathbb{E}[Y \mid do(X = x_1)]$

What if we cannot conduct randomized experiments?

(for example due to ethical concerns,
practical limitations, or logistical challenges)



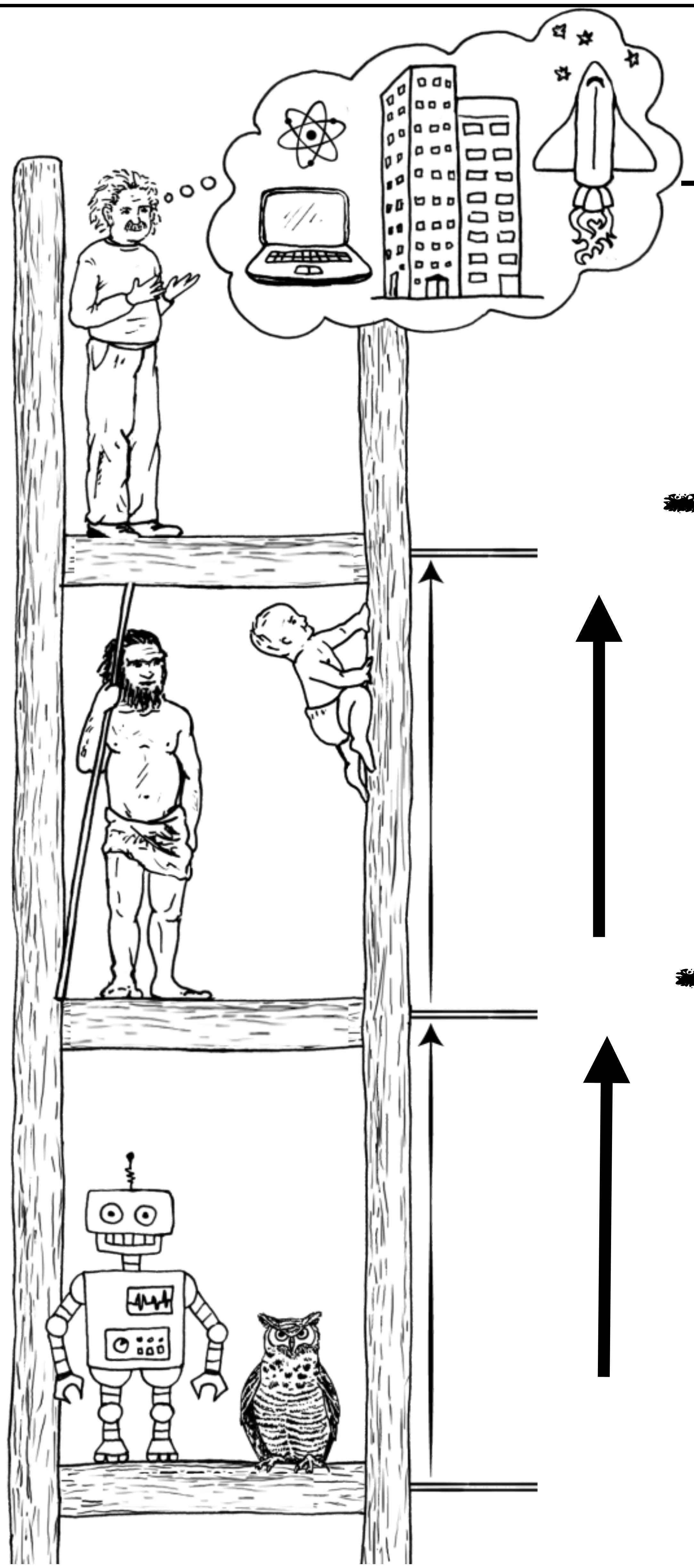
Data
Potential Causal Diagrams
Potential SCMs
Observational
X
Z
Y
input layer
hidden layer 1
hidden layer 2
output layer
DNN
$\hat{P}(Y|X=x)$
G_1
G_2
G_3
G_4
G_5
$\mathcal{M}_{11} = \langle \mathbf{V}, \mathbf{U}_1, \mathcal{F}_{11}, P_{11}(\mathbf{u}_1)\rangle$
$\mathcal{M}_{1k_1} = \langle \mathbf{V},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

Pearl's Causal Hierarchy (PCH)

The Three Inferential Layers

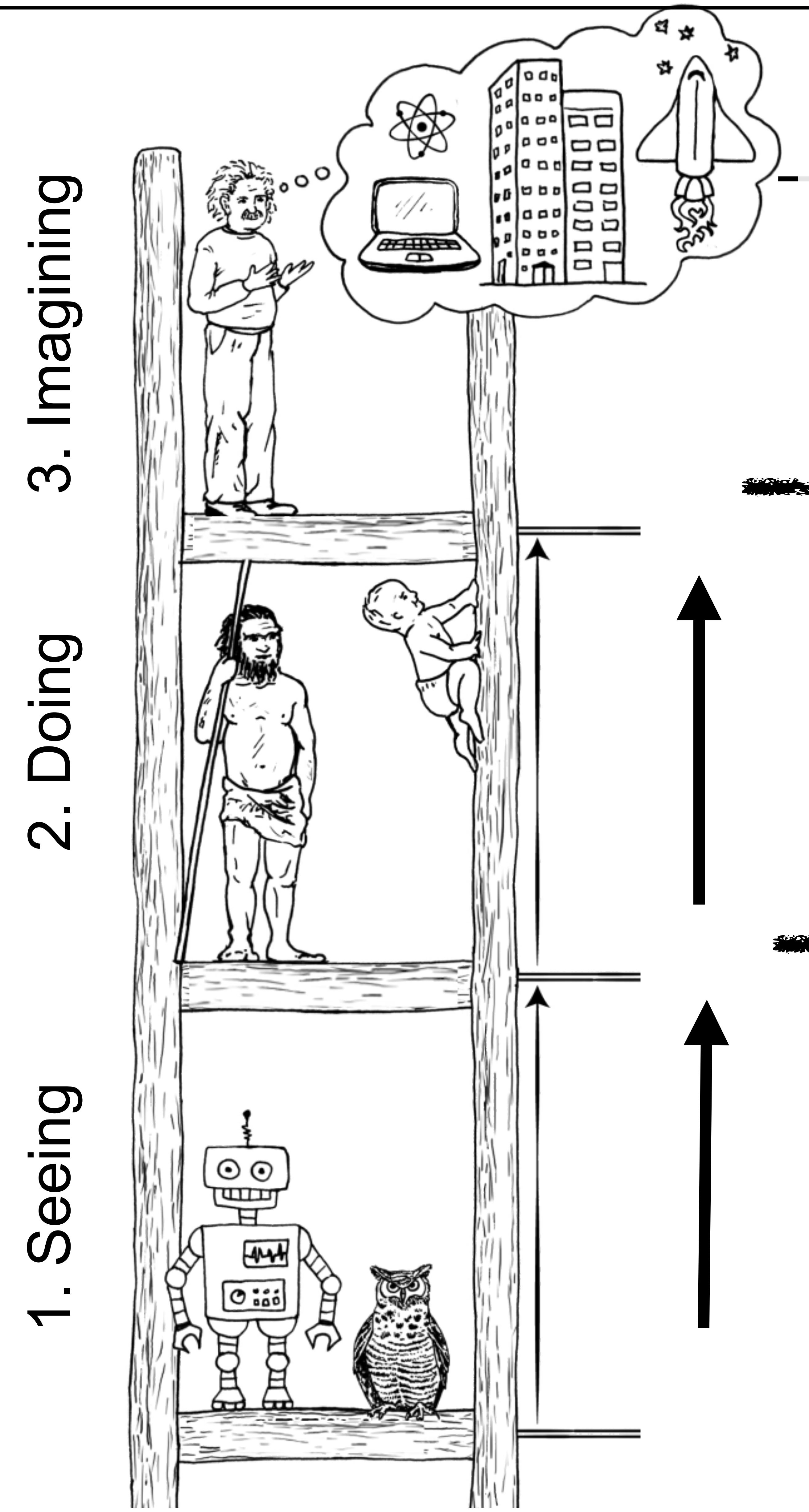
Ladder of Causation



	Layer	Task / Language	Typical Question	Examples
3. Imagining	Counterfactual $P(y_x \mid x', y')$	Structural Causal Model	What if I had acted differently?	Was it the aspirin that stopped my headache?
2. Doing	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?
1. Seeing	Associational $P(y \mid x)$	ML- (Un)Supervised (Bayesian Networks, Decision Trees, Deep Neural Networks)	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

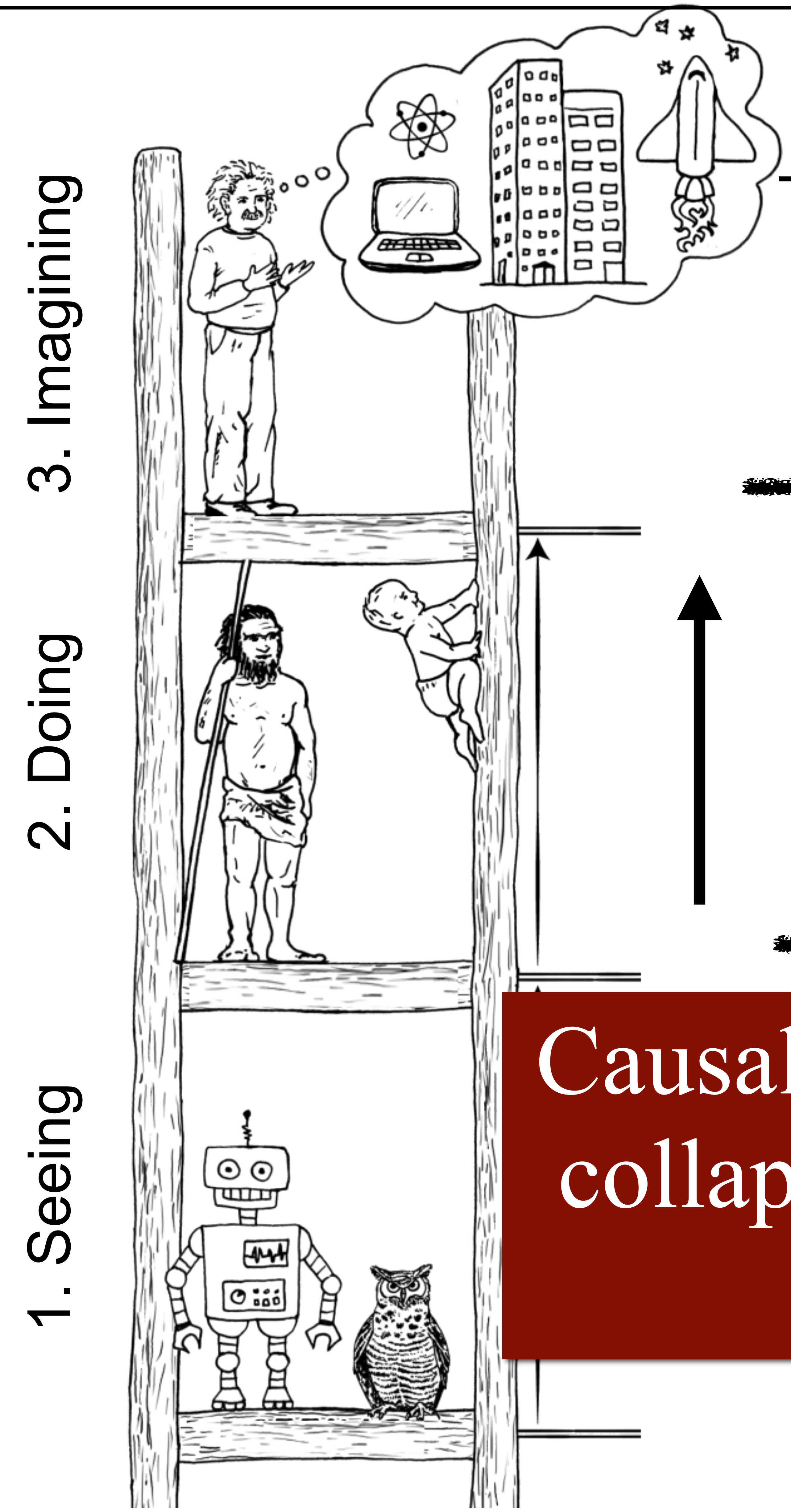
most of the inferences are about causal effects (policies, treatments, decisions)

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

Ladder of Causation



Cross-layer inferences:

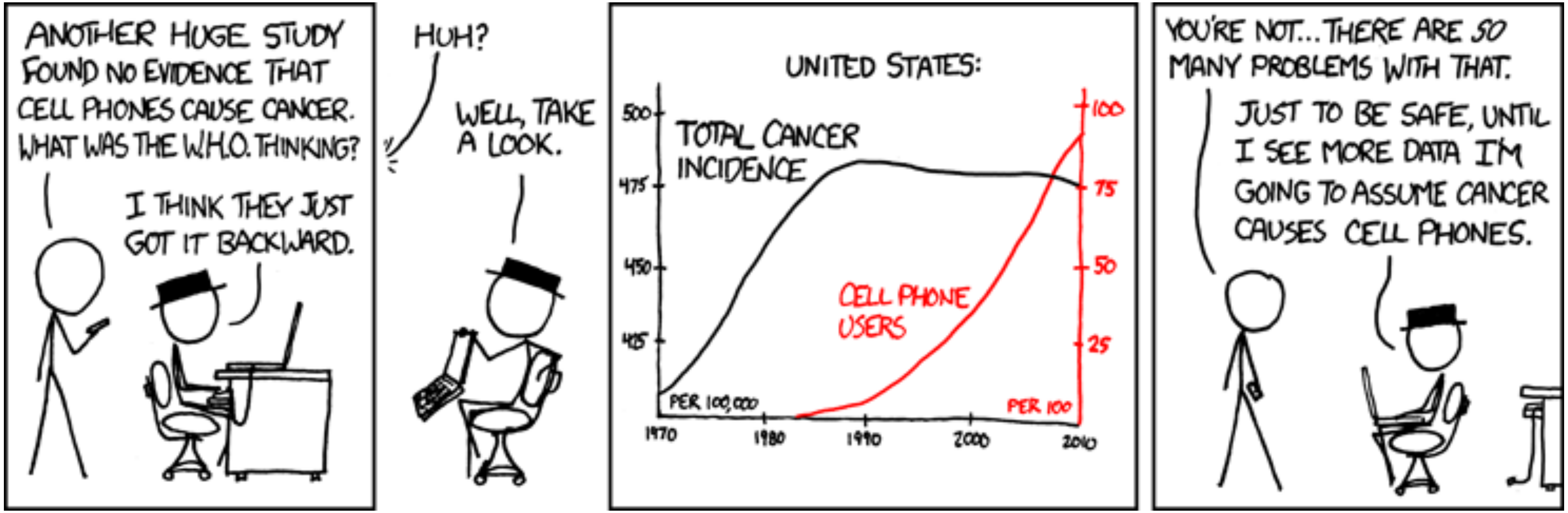
Doing

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

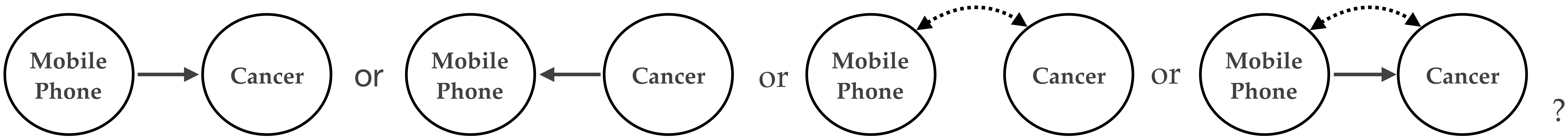
Causal Hierarchy Theorem : The ladder almost never collapses. That is, for almost any SCM, the rungs of the ladder remain distinct.

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

Association vs Causation



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



Which **type of data** would maybe provide us more definite conclusion?

Bayesian Network

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

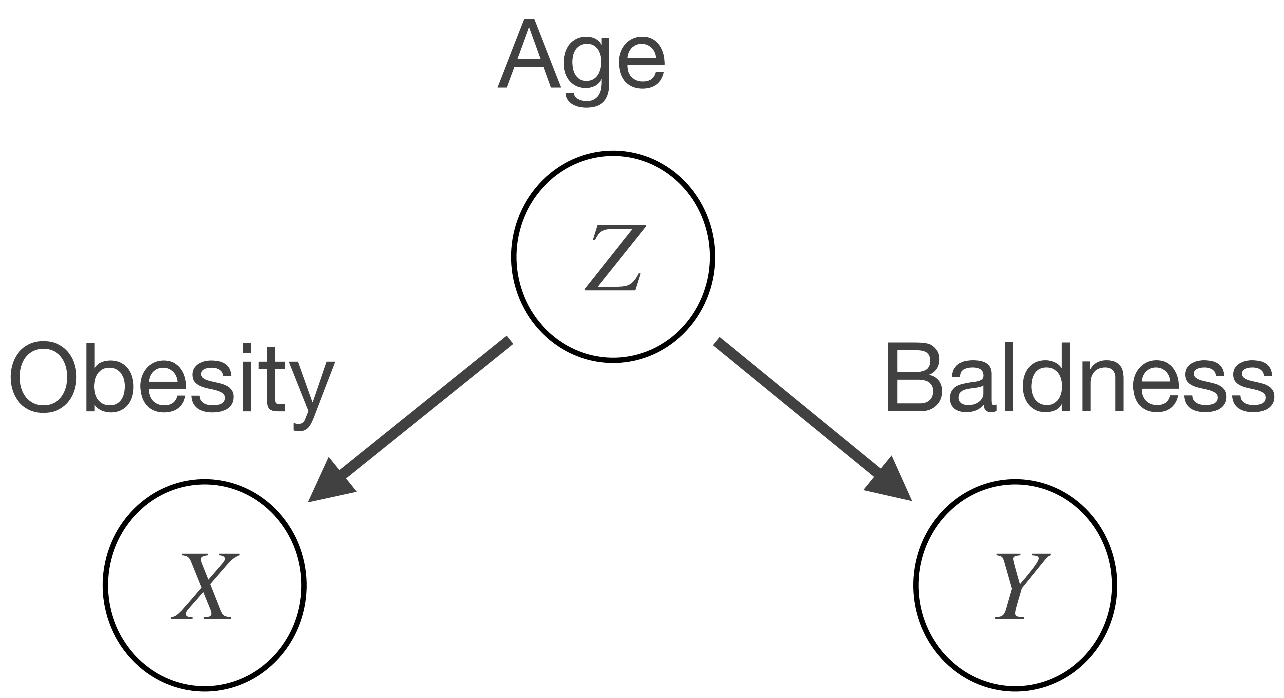
Directed
Acyclic Graph

Acyclic Directed
Mixed Graph

Encoding Conditional independencies

Fork

Z **as a common cause**

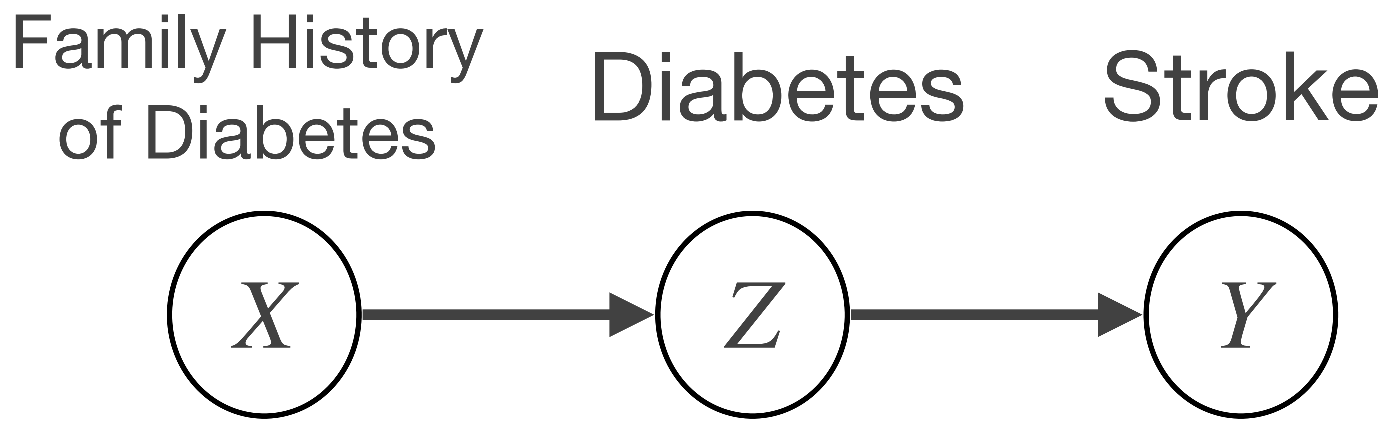


$$X \not\perp\!\!\!\perp Y$$

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

Chain

Z **as a mediator**

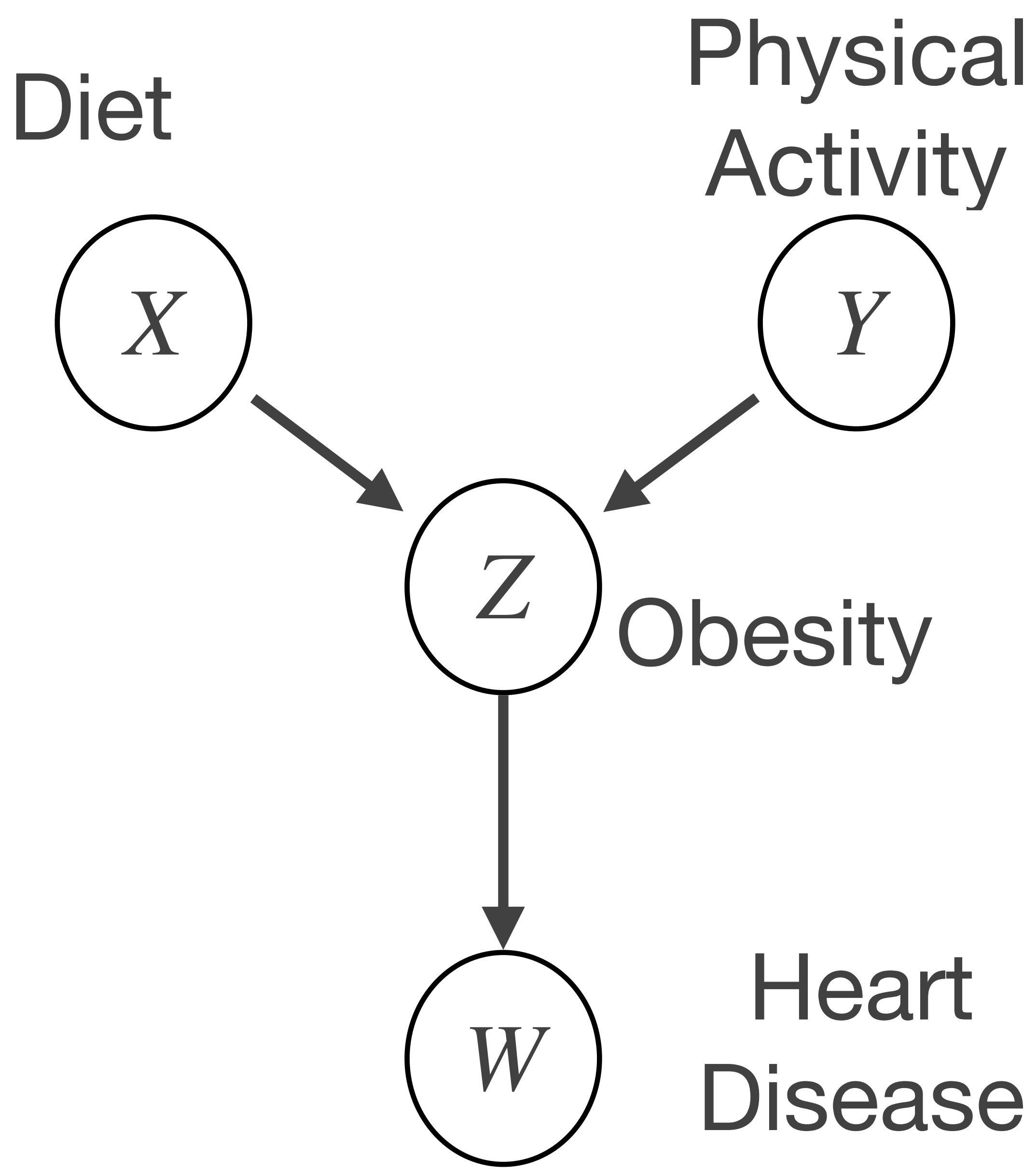


$$X \not\perp\!\!\!\perp Y$$

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

V-Structure

Z **as a collider or common effect**



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

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

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

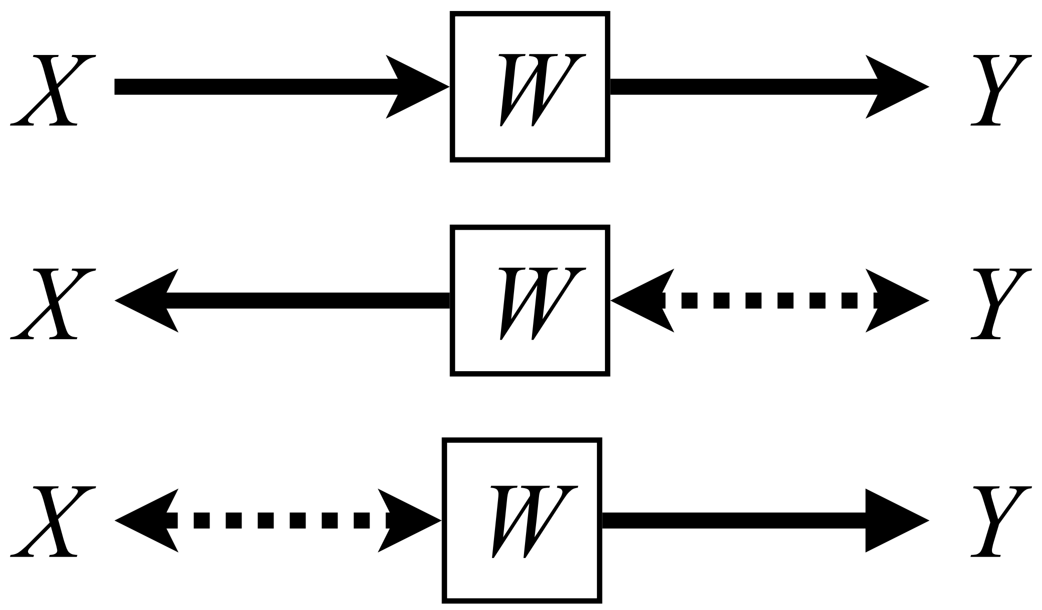
In both cases, Z is a non-collider!

Active and Inactive Triplets

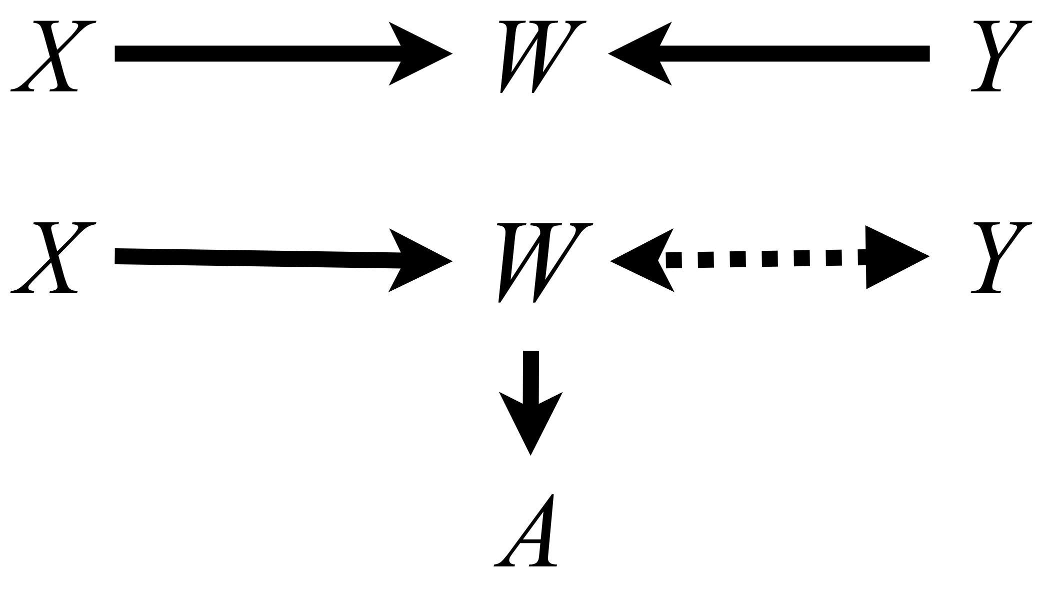
Definition (inactive): A triplet $\langle V_i, V_m, V_j \rangle$ is said to be ***inactive*** relative to a set $\mathbf{Z}$ if the middle node V_m:

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 $W, A \notin \mathbf{Z}$

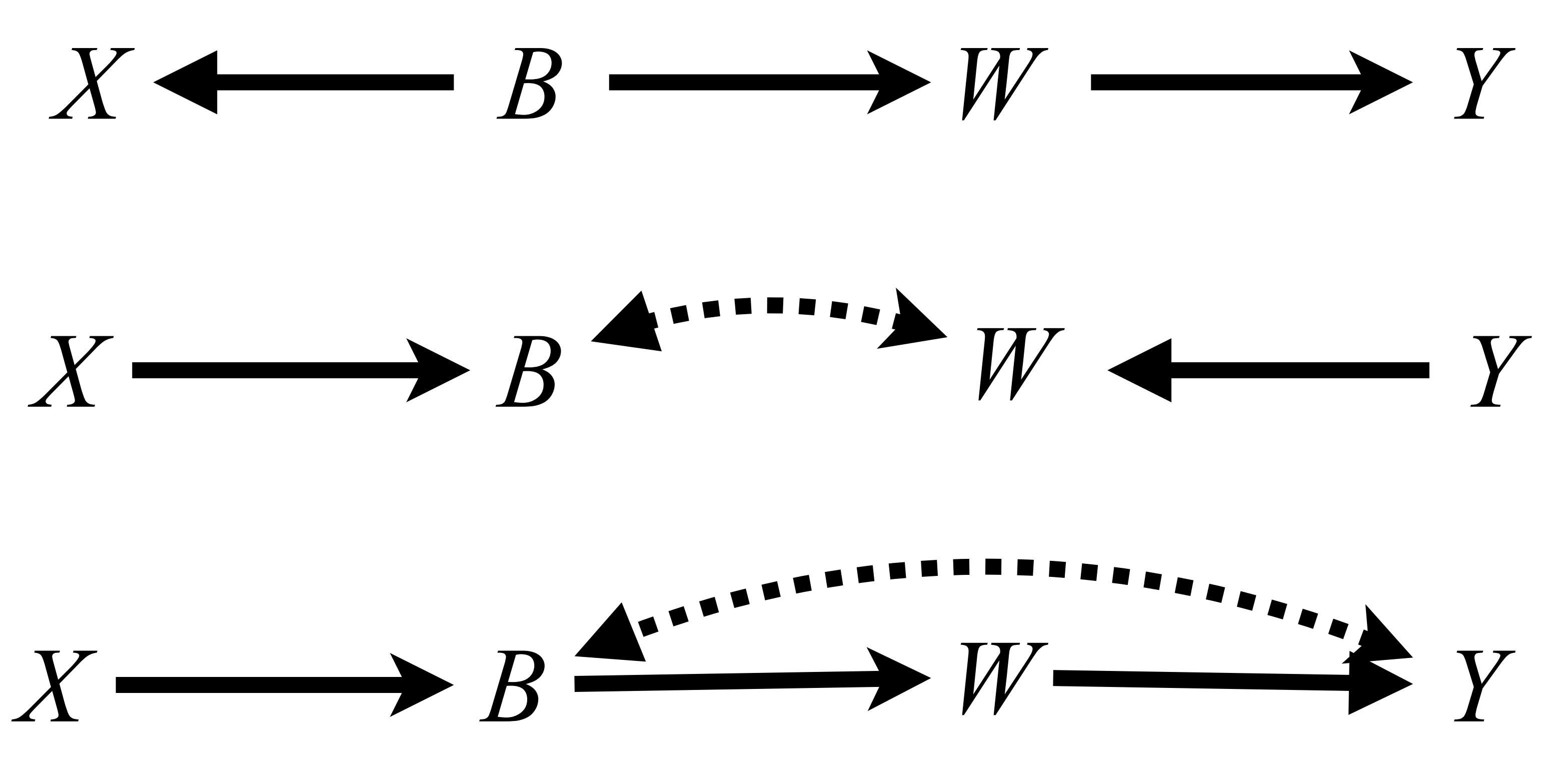


D-Separation

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

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



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

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

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

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

D-separations in G imply conditional independencies in P

BN - Encoder of Conditional Independences

Bayesian Networks (BN) are **Minimal Independence Maps**:

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

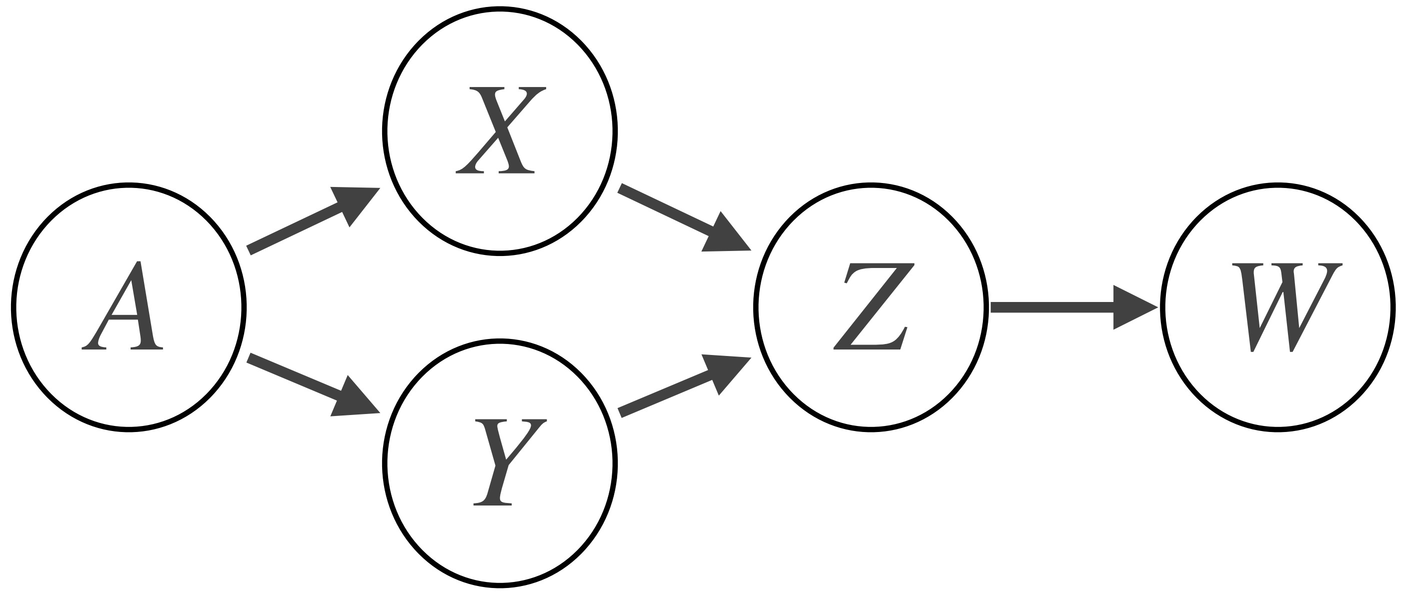
No edges of G can be removed without ceasing such a property.

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 $\mathscr{M}$.

Edges have no causal semantics!



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

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

BN - Encoder of Conditional Independences

Bayesian Networks (BN) are **Minimal Independence Maps**:

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

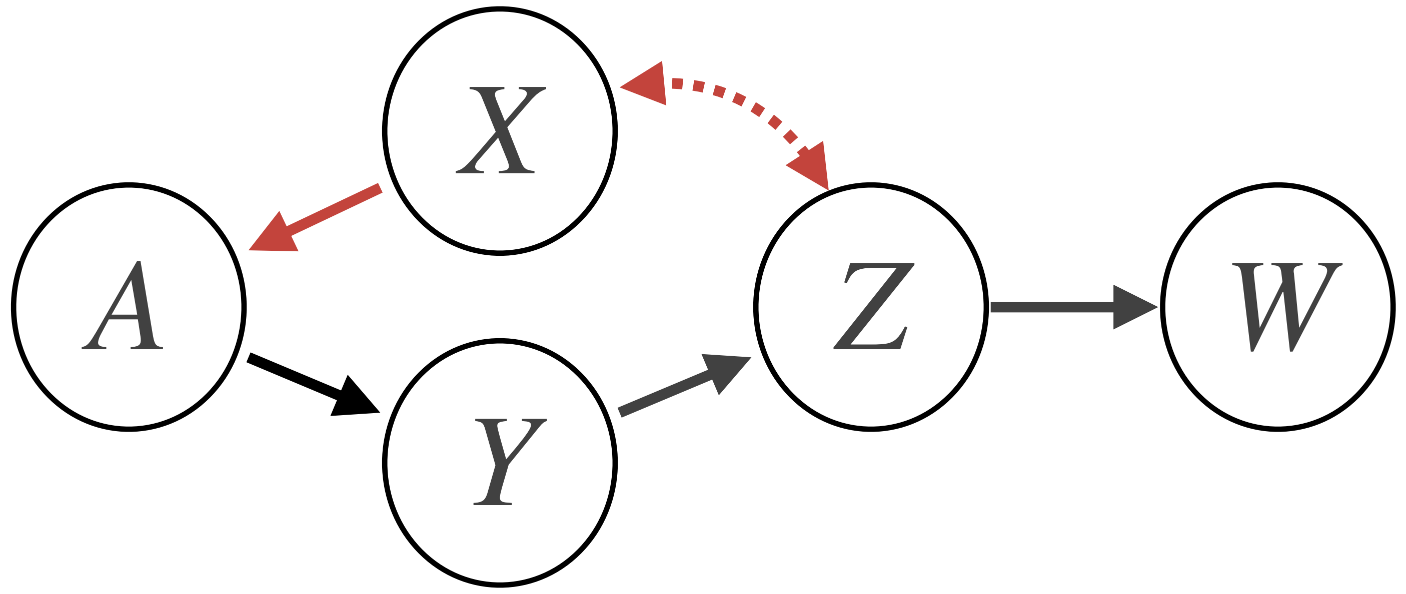
No edges of G can be removed without ceasing such a property.

Observational Distribution

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

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

Edges have no causal semantics!



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

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

Markov Equivalence Class

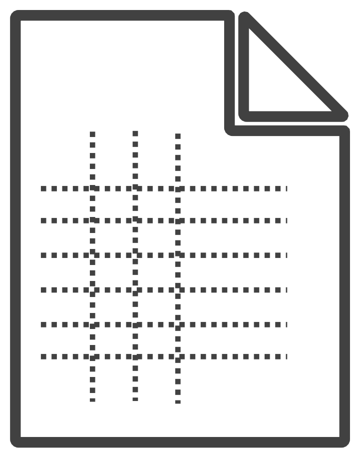
$$\mathcal{M}_1 = \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_x, U_Y\} \\ \mathcal{F} = \begin{cases} f_X(U_X) \\ f_Y(X, U_Y) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

$$\vdots$$

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

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

Data



Conditional (in)dependencies

$$P(\mathbf{v})$$

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

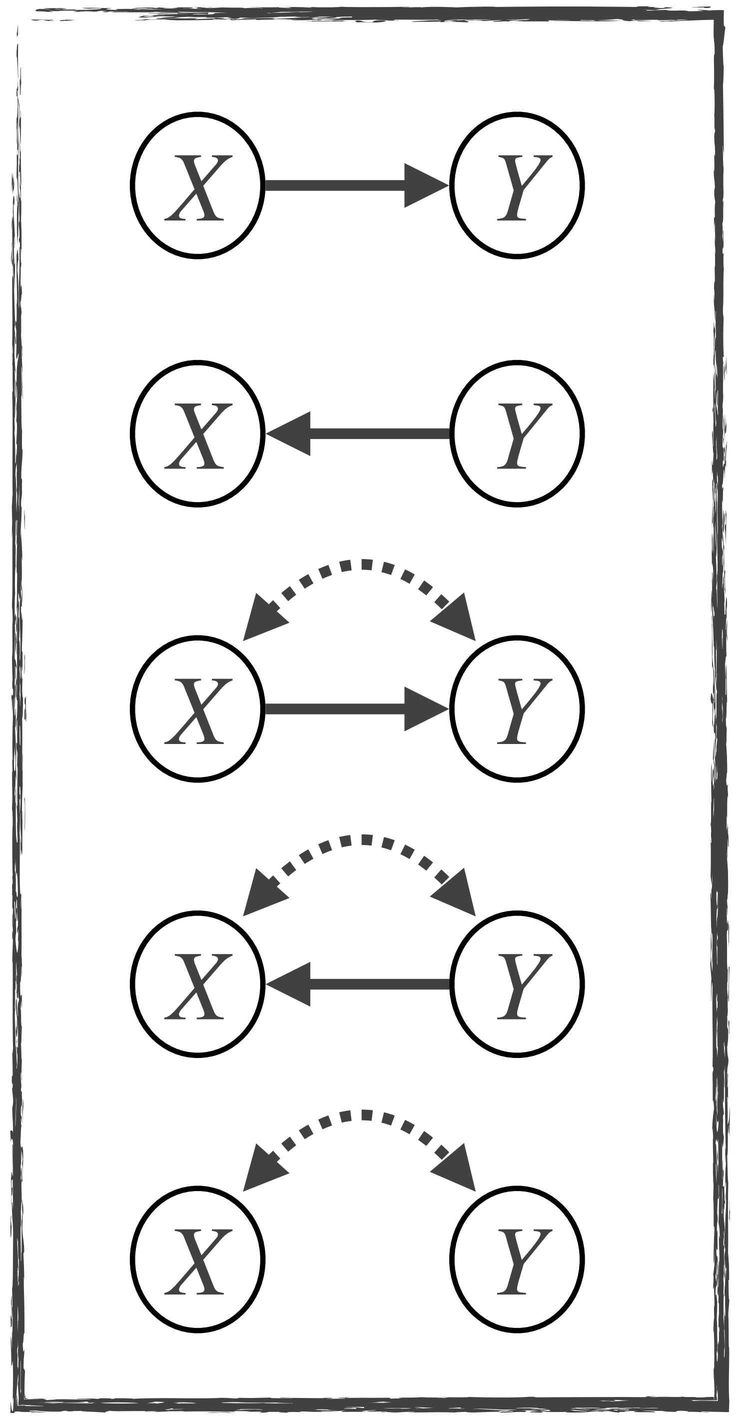
$$P(x, y) = \sum_{u_x, u_y} P(x \mid y) P(y) P(u_x, u_y)$$

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

$$\vdots$$

Markov Equivalence Class

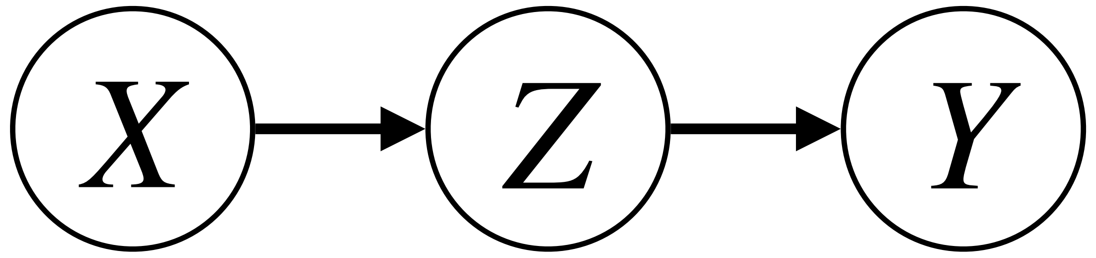
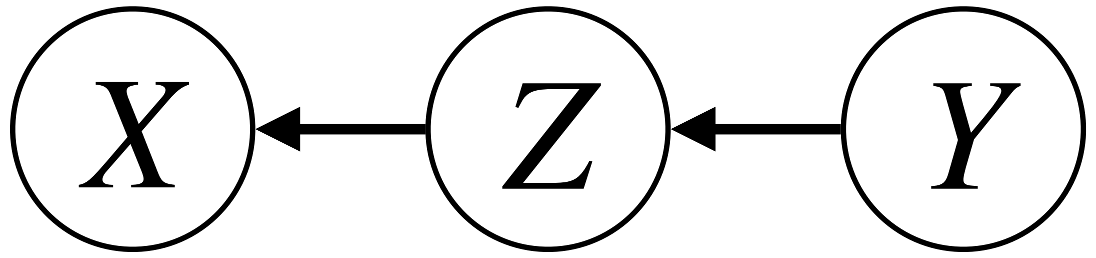
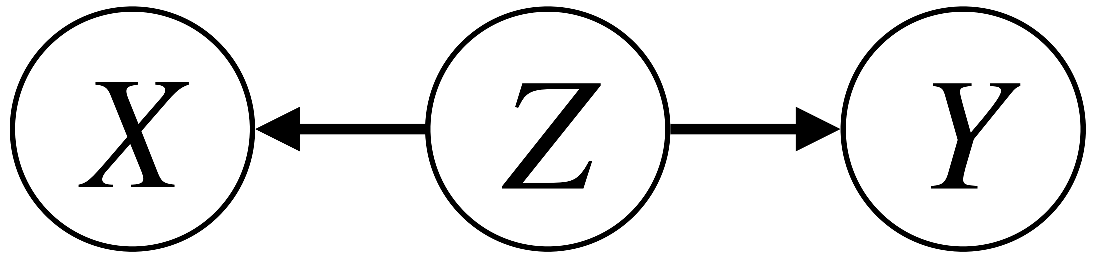
(class of models implying the same set of conditional independencies)



Correlation does not imply causation!

Equivalent Bayesian Networks

Observational Data

Distribution	**Factorization**	**Bayesian Networks**
	$$P(x,y,z) = \color{red}{P(y\,\vert\,x,z)}\color{black}P(z\,\vert\,x)P(x) = \color{red}{P(y\,\vert\,z)}\color{black}P(z\,\vert\,x)P(x)$$	
$P(X,Y,Z)$ with $P(Y\,\vert\,X,Z) = P(Y\,\vert\,X)$ i.e., $X \perp\!\!\!\perp Y\,\vert\,Z$	$$P(x,y,z) = \color{red}{P(x\,\vert\,y,z)}\color{black}P(y\,\vert\,z)P(z) = \color{red}{P(x\,\vert\,z)}\color{black}P(z\,\vert\,y)P(y)$$	 Markov Equivalent
	$$P(x,y,z) = \color{red}{P(y\,\vert\,x,z)}\color{black}P(x\,\vert\,z)P(z) = \color{red}{P(y\,\vert\,z)}\color{black}P(x\,\vert\,z)P(z)$$	

Two models are considered ***Markov equivalent*** if they imply the same conditional independencies.

Equivalent Bayesian Networks

Distribution	Factorization	Bayesian Networks

$$P(X, Y, Z)$$

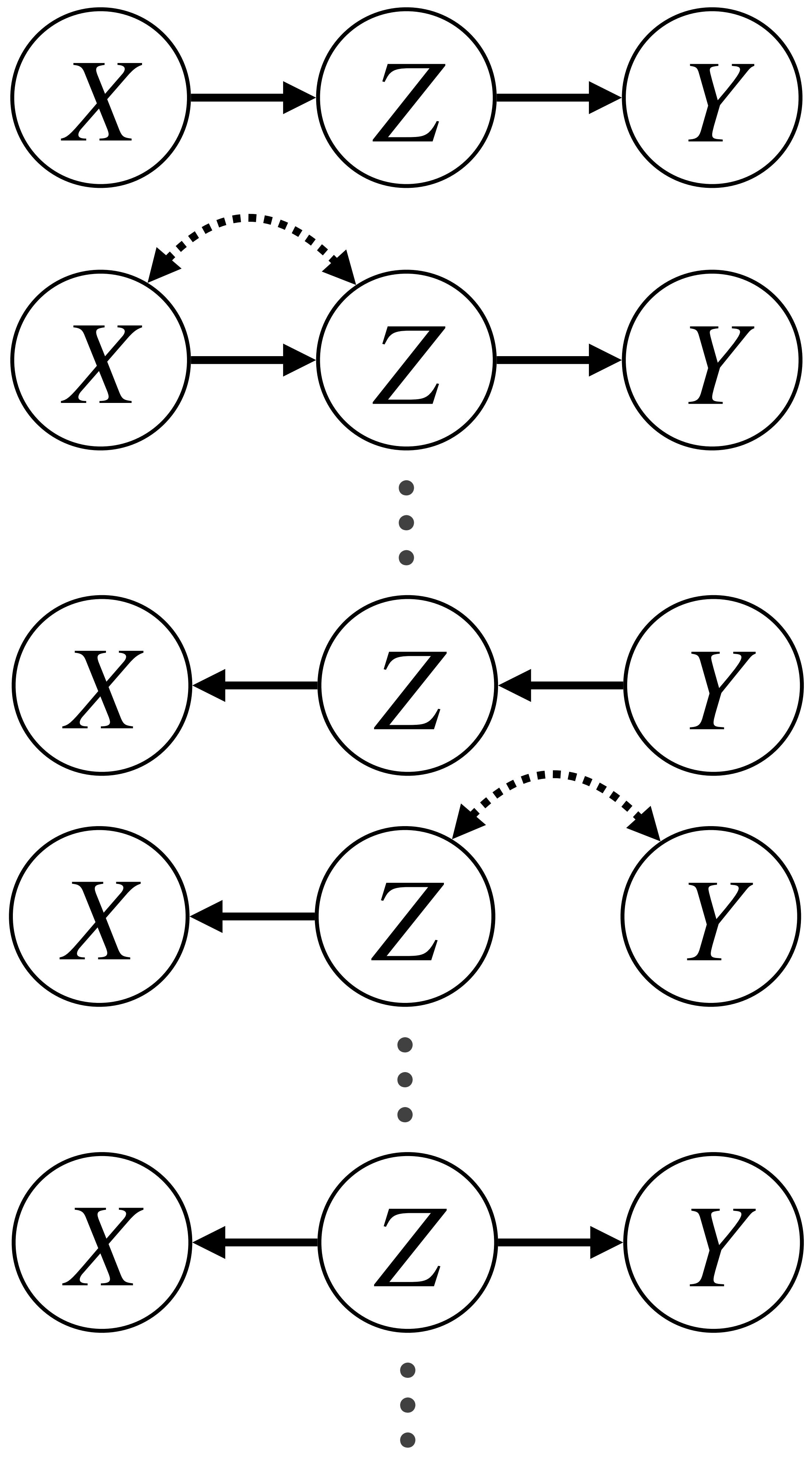
with $P(Y \mid X, Z) = P(Y \mid X)$

i.e., $X \perp\!\!\!\perp Y \mid Z$

$$P(x, y, z) = \color{red}{P(y \mid x, z)}P(z \mid x)P(x)$$
$$= \color{red}{P(y \mid z)}P(z \mid x)P(x)$$

$$P(x, y, z) = \color{red}{P(x \mid y, z)}P(y \mid z)P(z)$$
$$= \color{red}{P(x \mid z)}P(z \mid y)P(y)$$

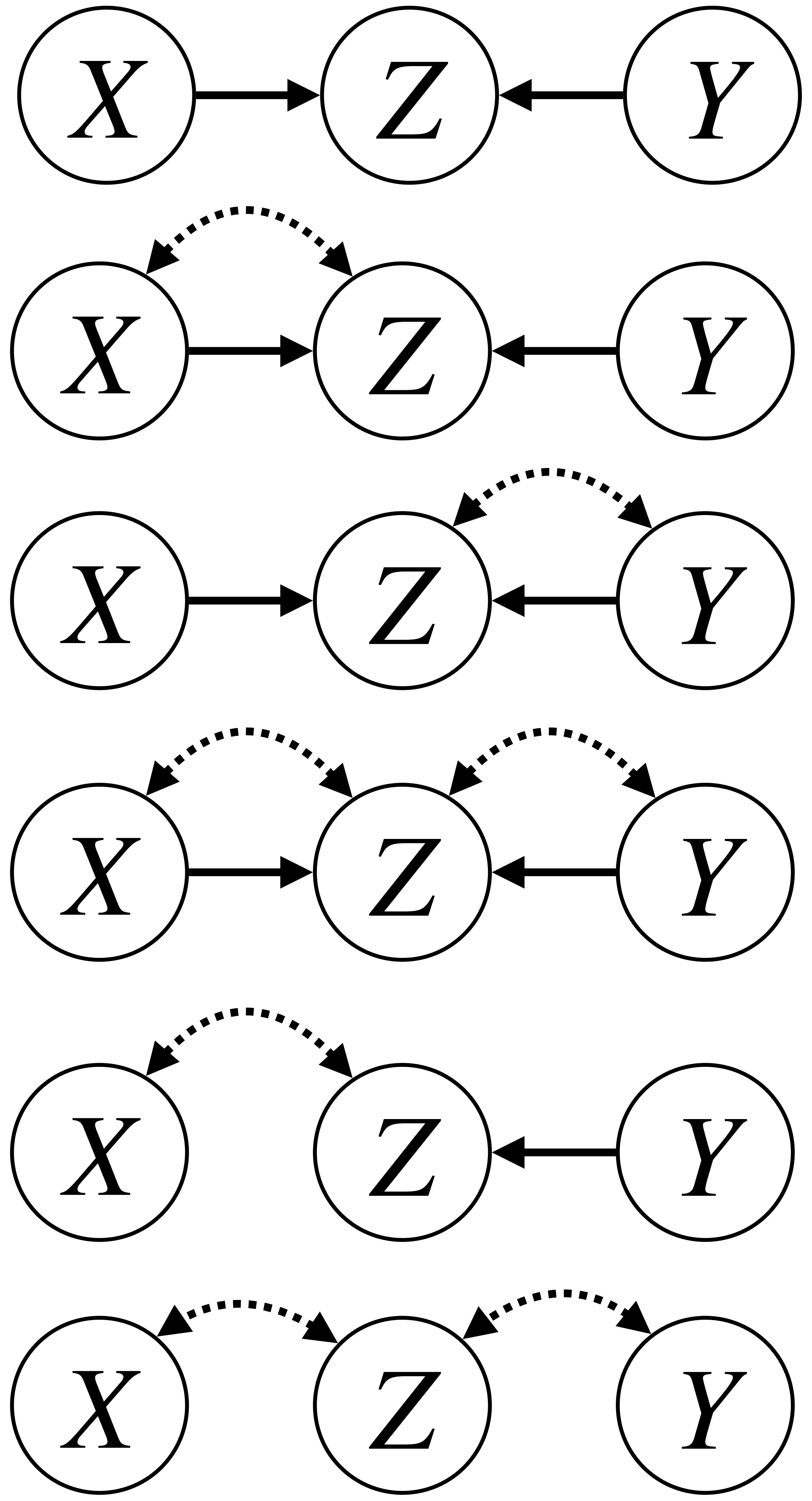
$$P(x, y, z) = \color{red}{P(y \mid x, z)}P(x \mid z)P(z)$$
$$= \color{red}{P(y \mid z)}P(x \mid z)P(z)$$



Markov Equivalent

Invariance:
Z is **never** a collider
(either ancestor of X and Y).

Equivalent Bayesian Networks

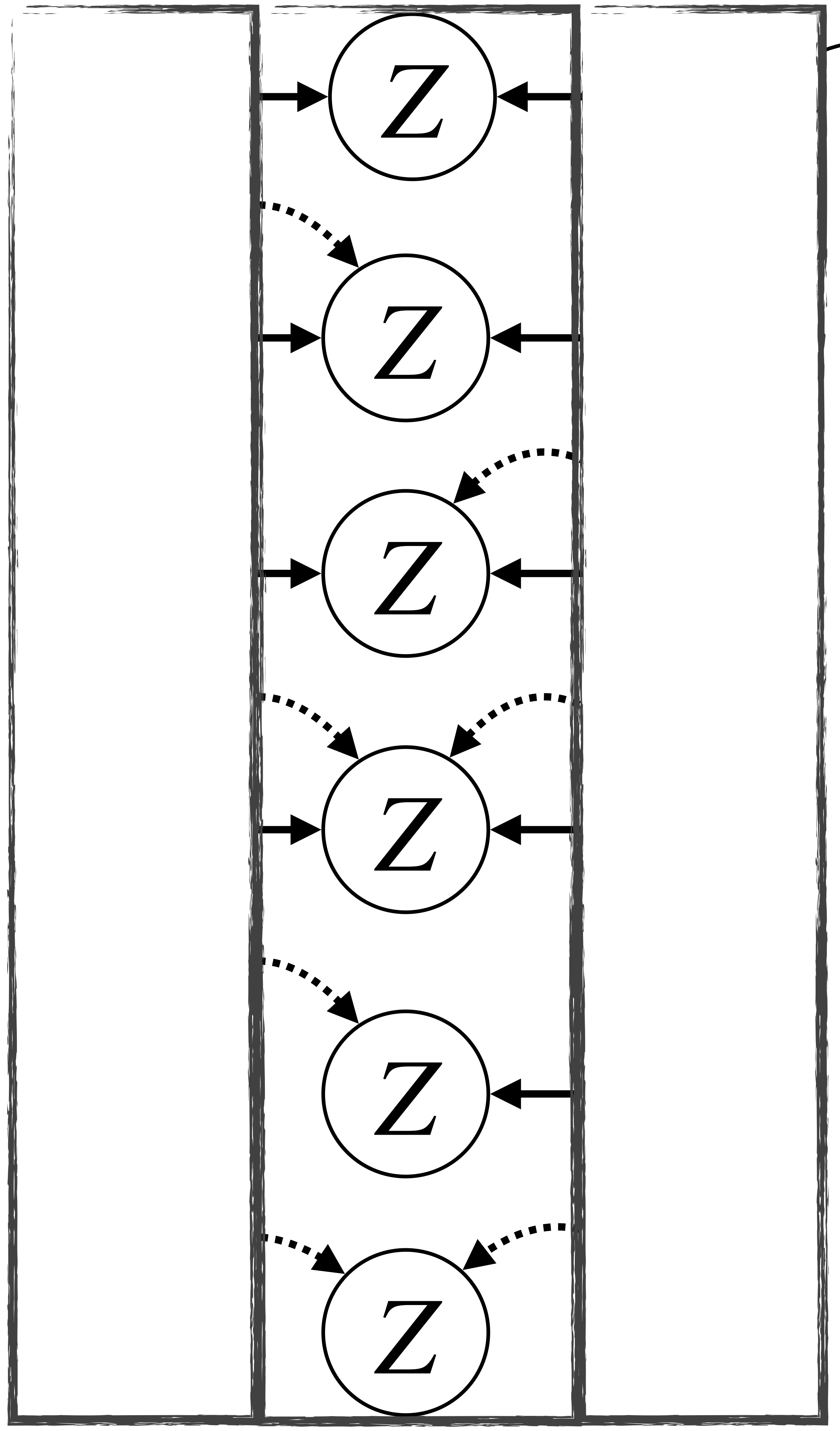
Distribution	Factorization	Bayesian Networks
$P(X, Y, Z)$ with $P(Y \mid X) = P(Y)$ i.e., $X \perp\!\!\!\perp Y$	$P(x, y, z) = P(z \mid x, y)\color{red}{P(x \mid y)}\color{black}P(y)$ $= P(z \mid x, y)\color{red}{P(x)}\color{black}P(y)$	

Identifiable if no latent confounder

Markov Equivalent

Equivalent Bayesian Networks

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



Identifiable if no latent confounder

Markov Equivalent

Invariance:
Z is **always** a collider
(non-ancestor of X and Y).

Causal Bayesian Network

A DAG, possibly with latent confounders (ADMG),
representing the **causal and confounding relationships**
implied by an SCM

CBN: Encoder of Structural Causal Knowledge

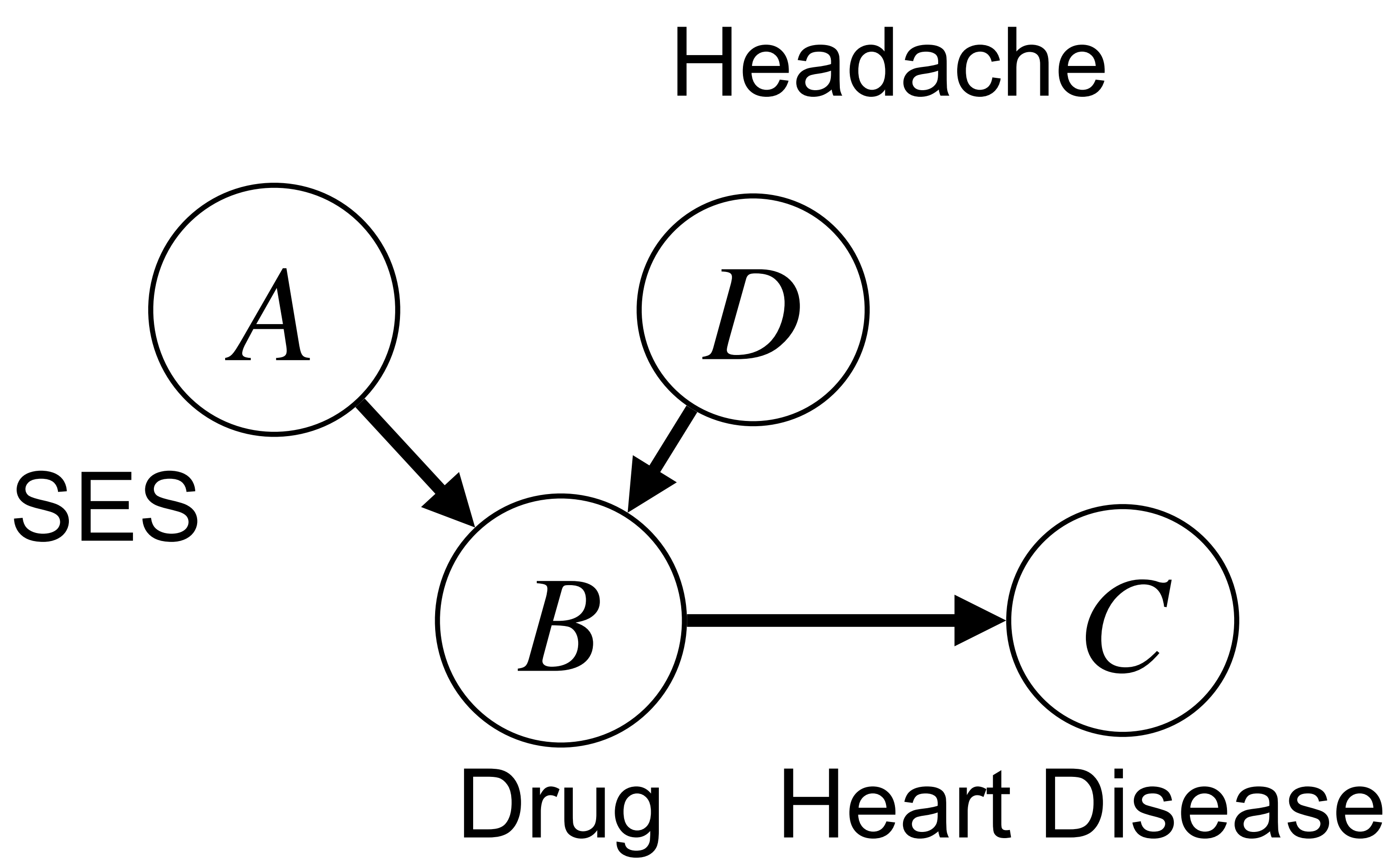
Structural Causal Model (SCM)

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

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

Induced Causal Bayesian Network (CBN)

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 $f_j \in \mathcal{F}$.

CBN: Encoder of Structural Causal Knowledge

Structural Causal Model (SCM)

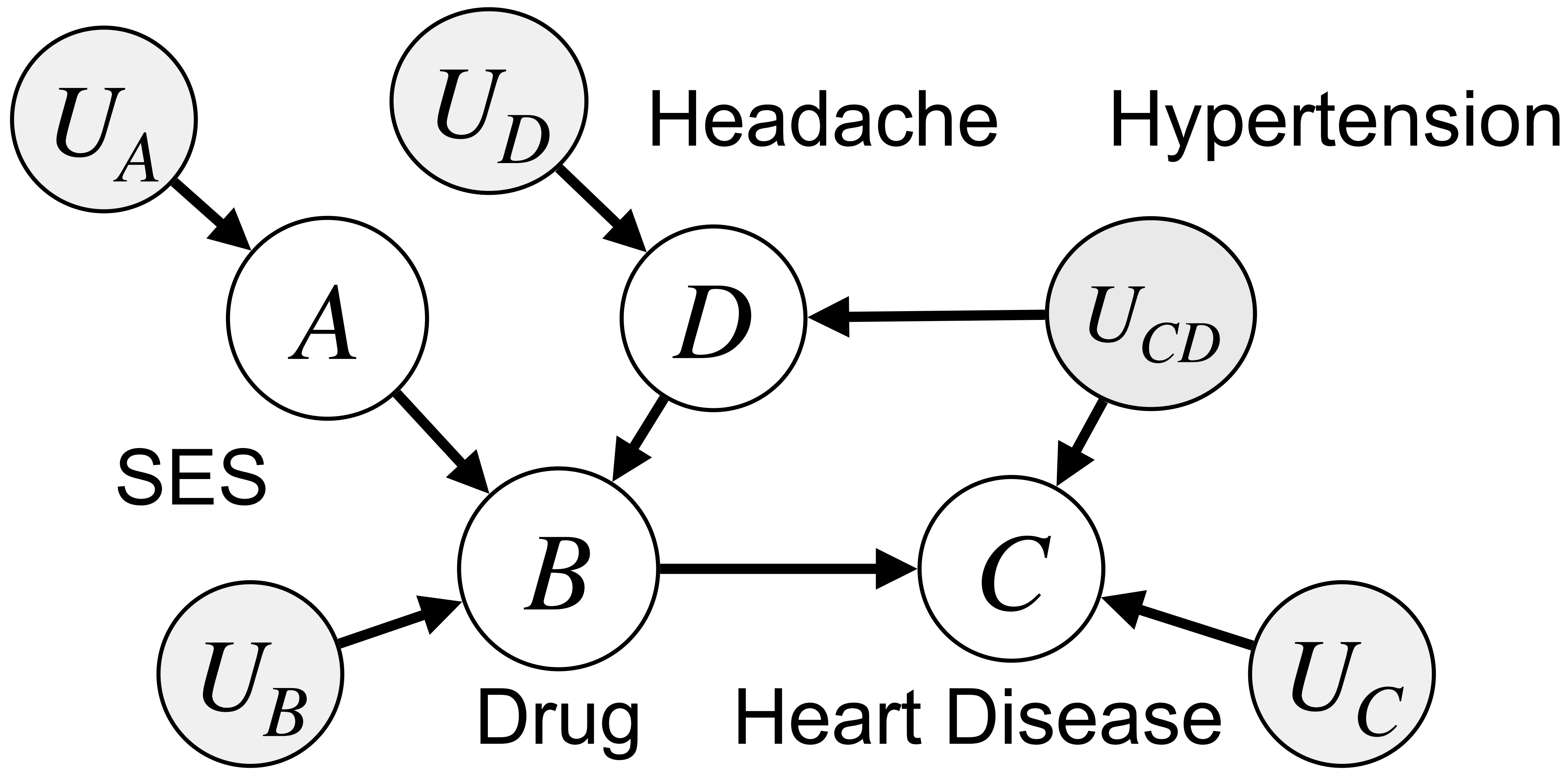
$$\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$$

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

$\longrightarrow$

Induced Causal Bayesian Network (CBN)

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 $f_j \in \mathcal{F}$.

CBN: Encoder of Structural Causal Knowledge

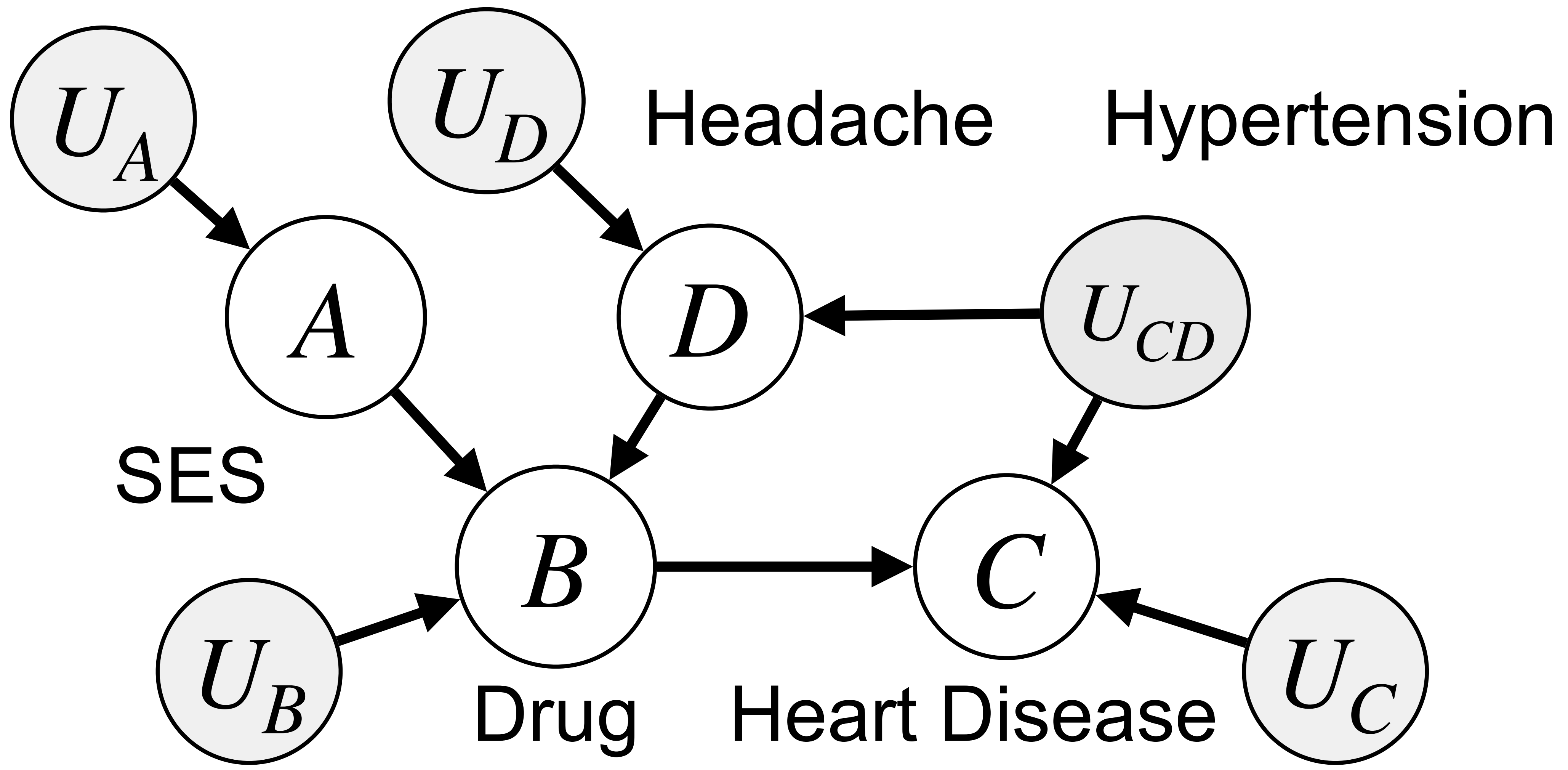
Structural Causal Model (SCM)

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

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

Induced Causal Bayesian Network (CBN)

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 $U \in \mathbf{U}$.

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

Structural Causal Model (SCM)

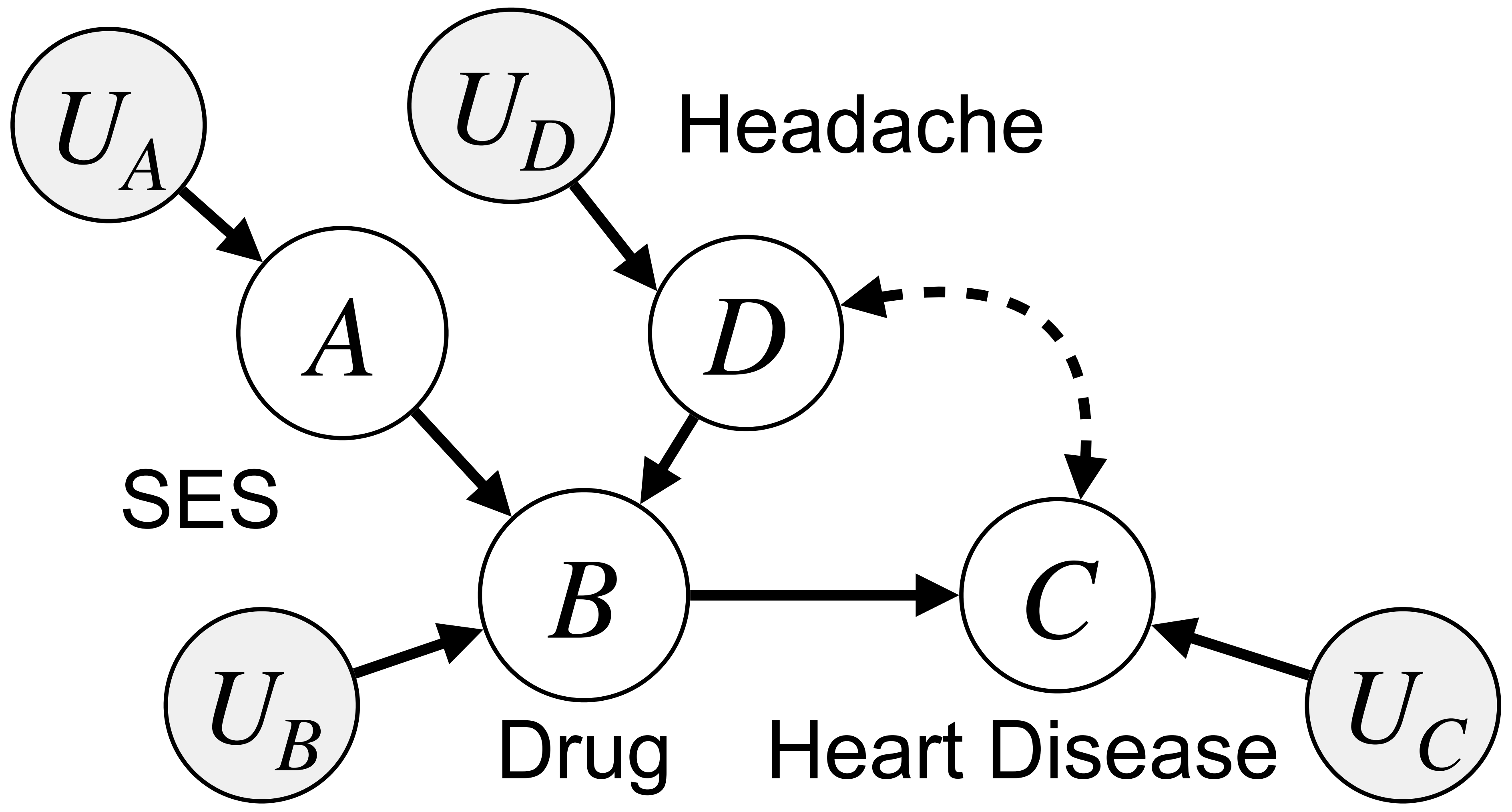
$$\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$$

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

$\longrightarrow$

Induced Causal Bayesian Network (CBN)

Causal Diagram



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

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

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

CBN: Encoder of Structural Causal Knowledge

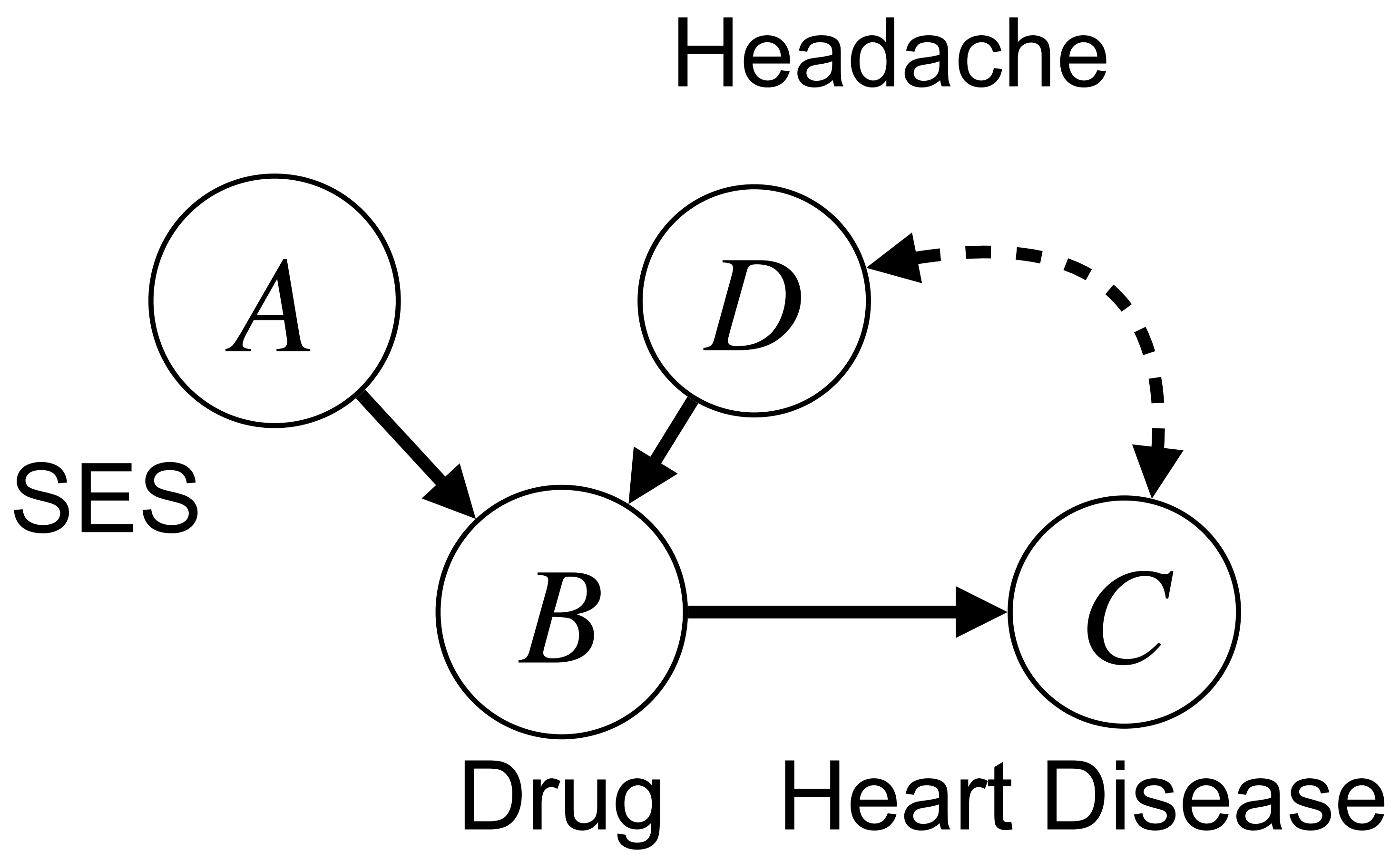
Structural Causal Model (SCM)

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

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

Induced Causal Bayesian Network (CBN)

Causal Diagram



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

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

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

CBN: Encoder of Structural Causal Knowledge

Let $\mathbf{P}_*$ be the collection of all interventional distributions $P(\mathbf{V} \mid do(\mathbf{x}))$, $\mathbf{X} \subseteq \mathbf{V}$, including the null (observational) distribution $P(\mathbf{V})$.

An Acyclic Directed Mixed Graph (ADMG) G is a CBN for $\mathbf{P}_*$ if for every intervention $do(\mathbf{X} = \mathbf{x})$, $\mathbf{X} \subseteq \mathbf{V}$, if it hold:

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 $\mathscr{M}_x$.

Semi-Markov relative to $G_{\overline{\mathbf{X}}}$

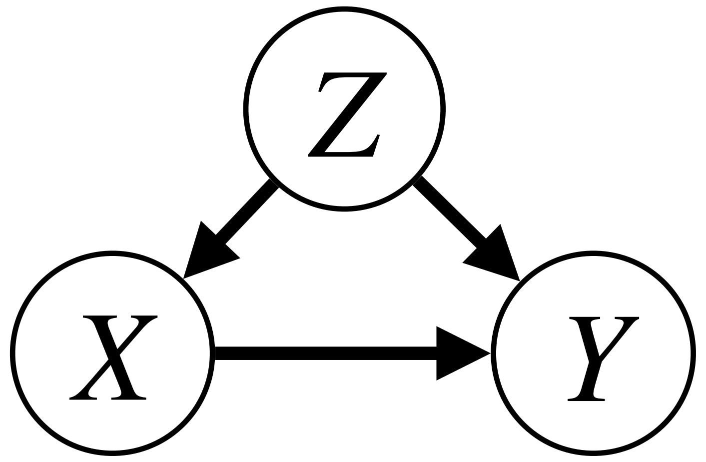
Causal Effect Identification from Causal Diagrams / CBNs

Causal Pipeline from a Causal Diagram

1 **Query**

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

2 **Causal Contraints**



3 $P(x, m, y)$

Observational Distribution

Inference Engine

Solution

yes / no

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

$$\mathbb{E}(y \mid do(x)) = \sum_z \mathbb{E}(y \mid x, z) P(z)$$

Interventional Distribution ← Available Distributions

👍

Observational Data

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

Causal Modeling / Causal Discovery → Causal Effect Identification → Causal Effect Estimation

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. $do(\mathbf{X} = \mathbf{x})$

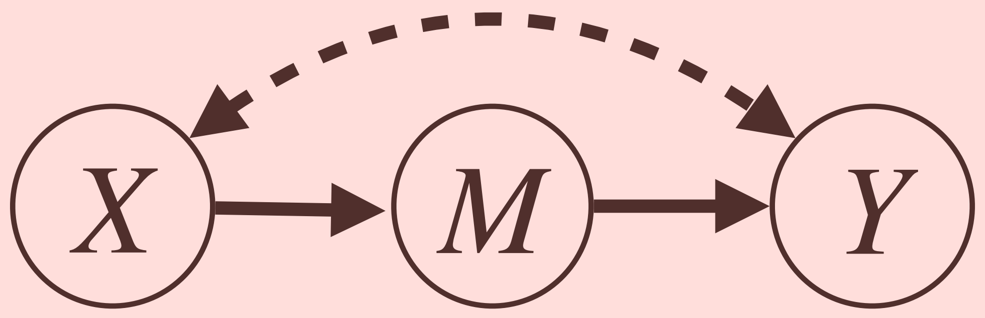
Classical Causal Effect Identification

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

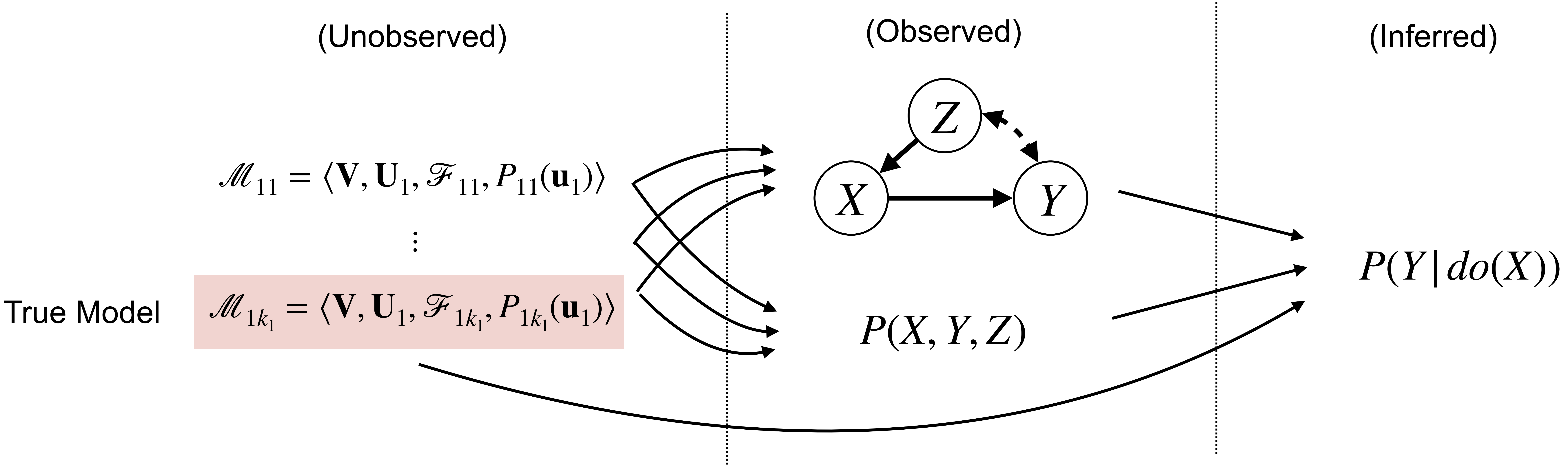
$$P(y \mid do(x)) = \sum_{m} P(m \mid x) \sum_{x'} P(y \mid m, x')P(x')$$

Interventional Distribution ← Available Distributions 👍

- Tian, J. and Pearl, J. (2002) 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

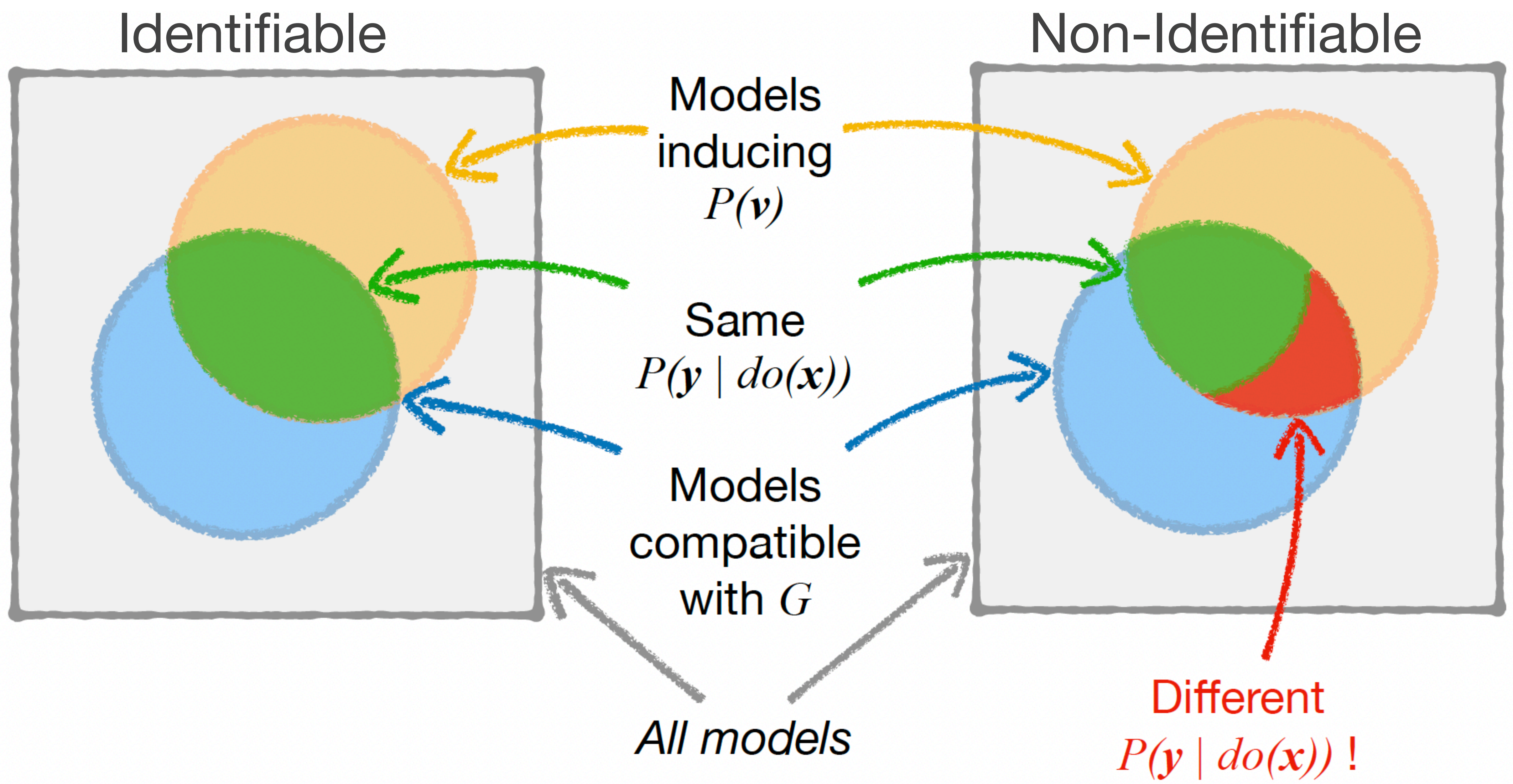
Causal Effect Identifiability: The causal effect of a (set of) treatment variable(s) $\mathbf{X}$ on a (set of) outcome variable(s) $\mathbf{Y}$ is said to be identifiable from a causal diagram G and the probability of the observed variables $P(\mathbf{V})$ if the interventional distribution $P(\mathbf{Y}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

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

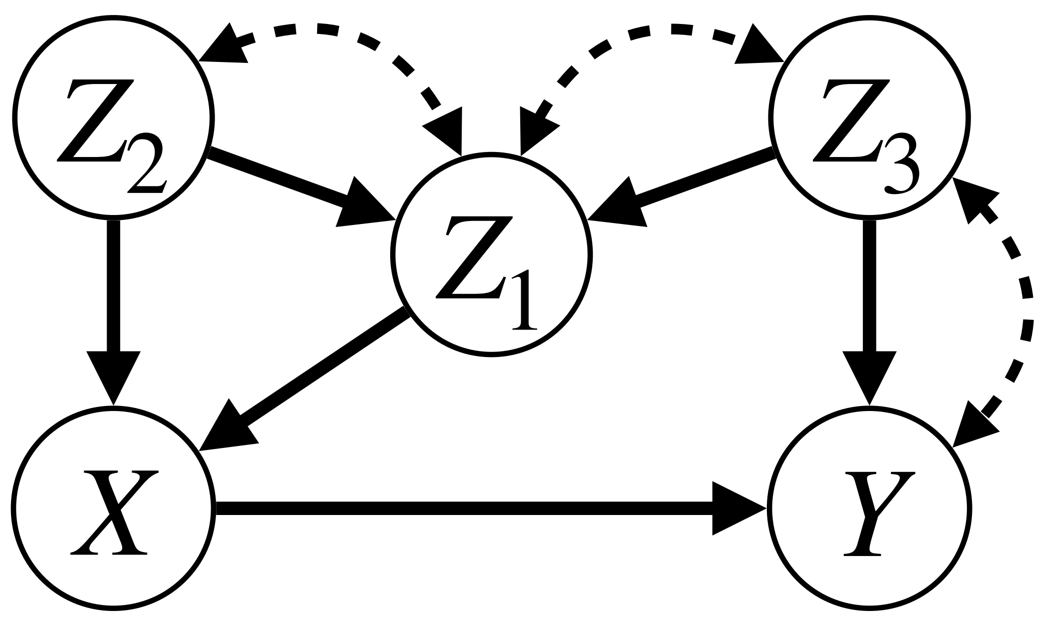
Identification Via Adjustment over Parents

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 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)

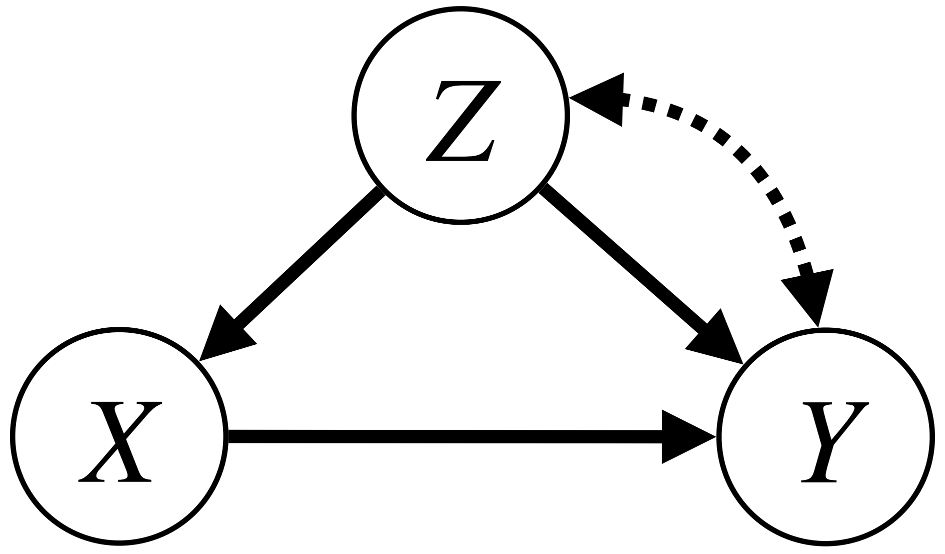
Then, $\mathbf{Z}$ satisfies the ***backdoor criterion*** for $(\mathbf{X}, \mathbf{Y})$ and, then the effect of $\mathbf{X}$ on $\mathbf{Y}$ is given by:

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

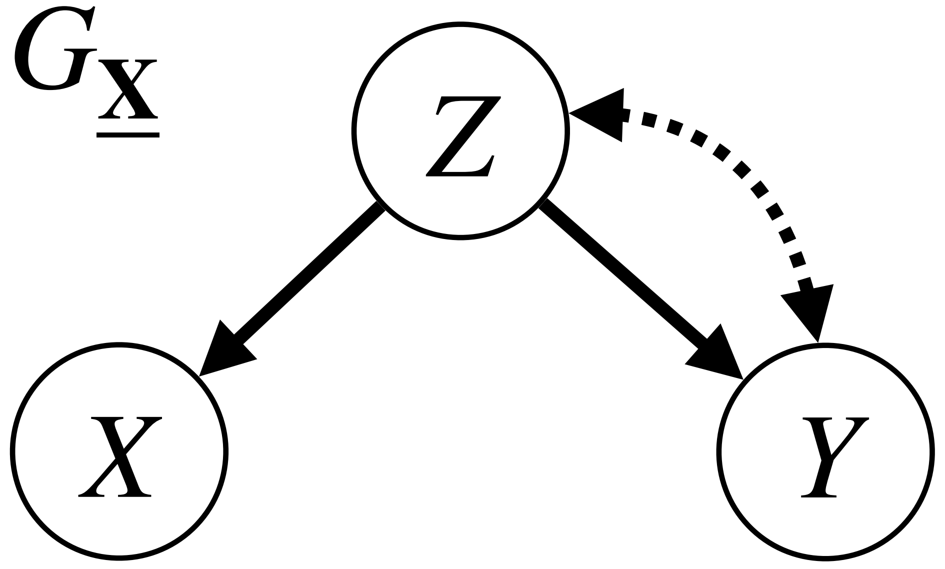
$\mathbf{Z}$, a set of covariates, admissible for backdoor adjustment

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

G



$G_{\underline{\mathbf{X}}}$



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)

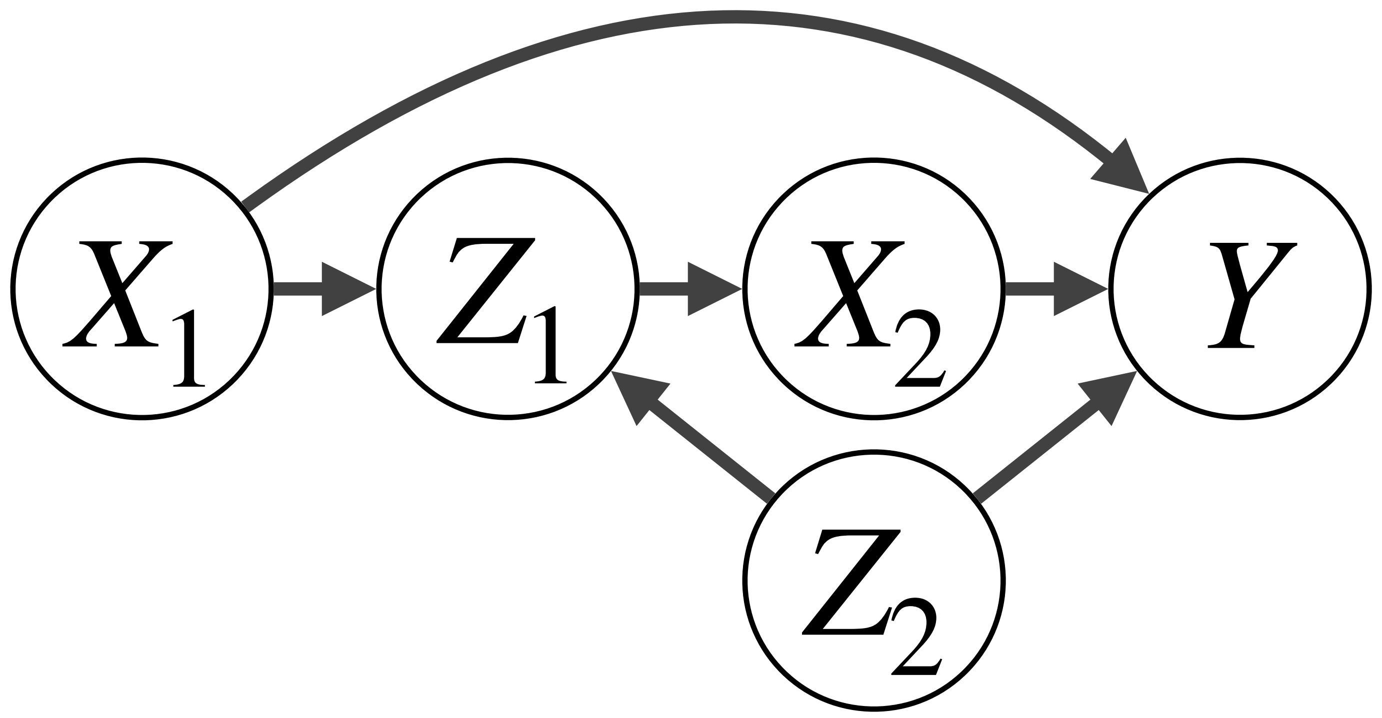
Then, $\mathbf{Z}$ satisfies the ***backdoor criterion*** for $(\mathbf{X}, \mathbf{Y})$ and, then the effect of $\mathbf{X}$ on $\mathbf{Y}$ is given by:

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

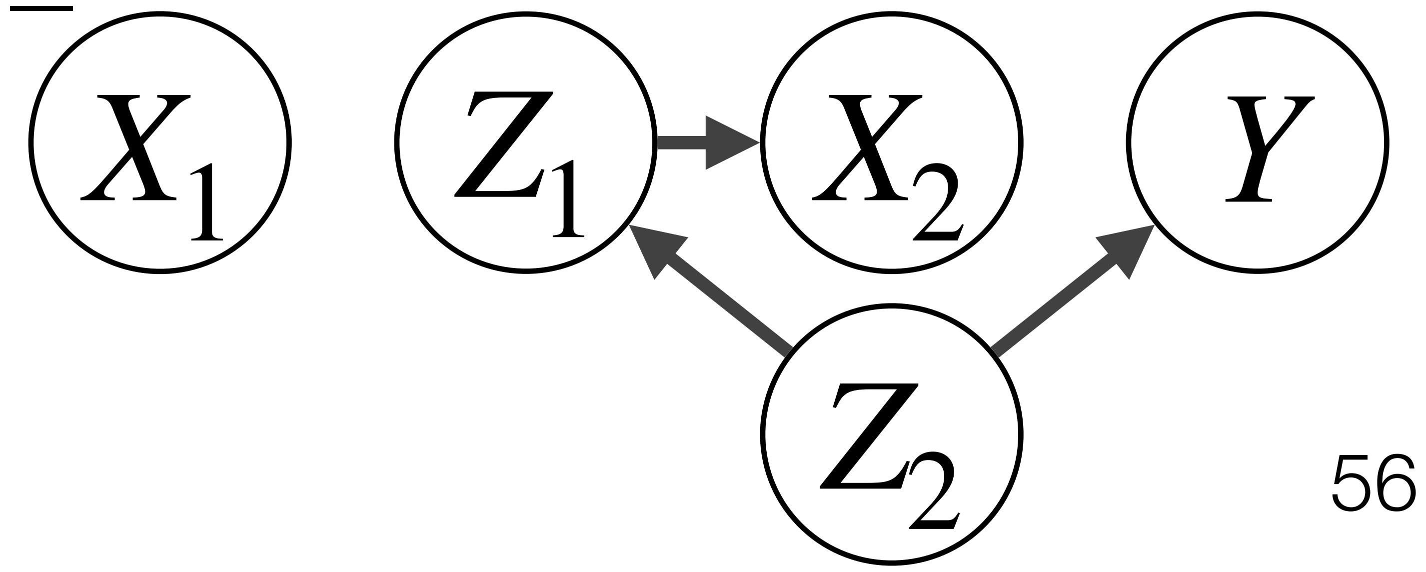
$\mathbf{X} = \{X_1, X_2\}$
$\mathbf{Y} = \{Y\}$
$\mathbf{Z} = \{Z_2\}$

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

G



$G_{\underline{\mathbf{X}}}$

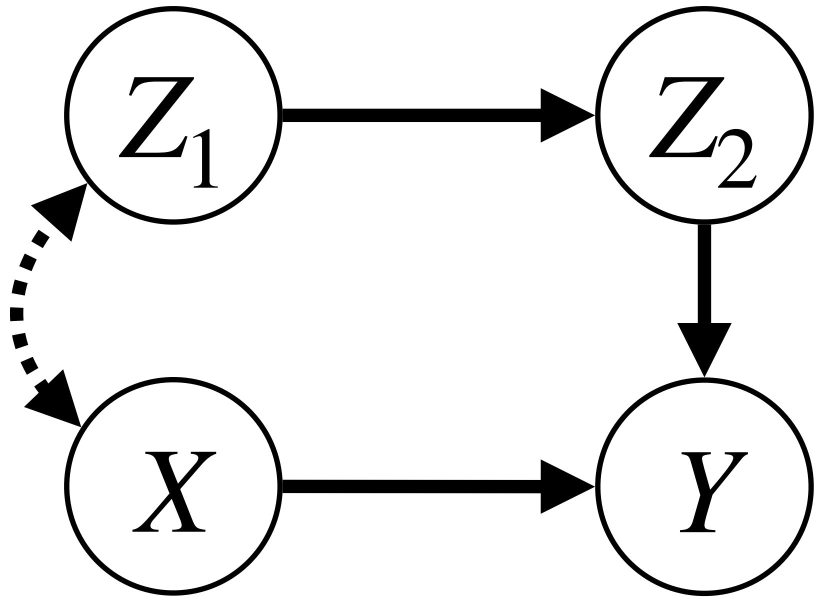


Judea Pearl. Comment: Graphical models, causality and intervention. Stat. Sci., 8:266–269, 1993.

Admissible Sets for BD Adjustment

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

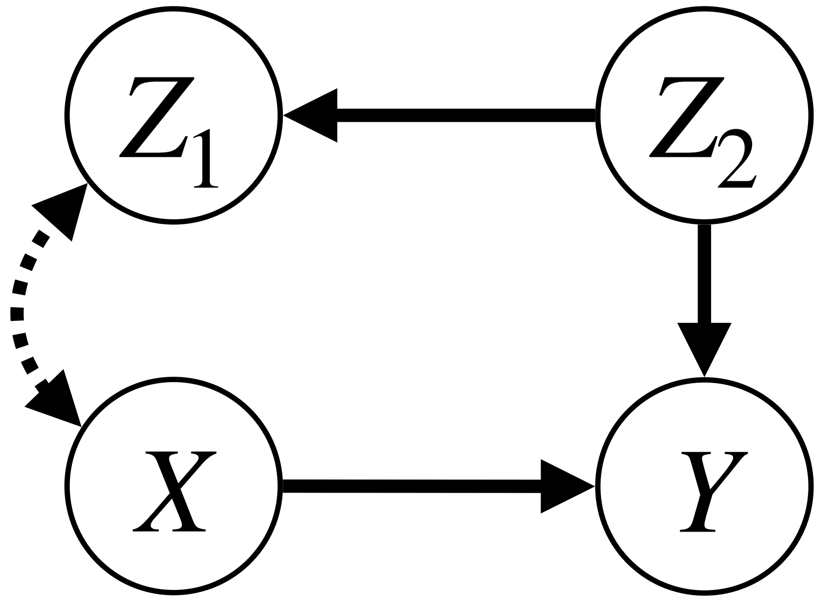
1. $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ in the graph $G_{\underline{\mathbf{X}}}$, i.e., the graph resulting from cutting the arrows out of $\mathbf{X}$
2. no node in $\mathbf{Z}$ is a descendant of a variable $X \in \mathbf{X}$ in G (all variables in $\mathbf{Z}$ are pre-treatment)



Minimal BD Adjustment Sets: $\{Z_1\}$, $\{Z_2\}$,

$\{Z_1, Z_2\}$

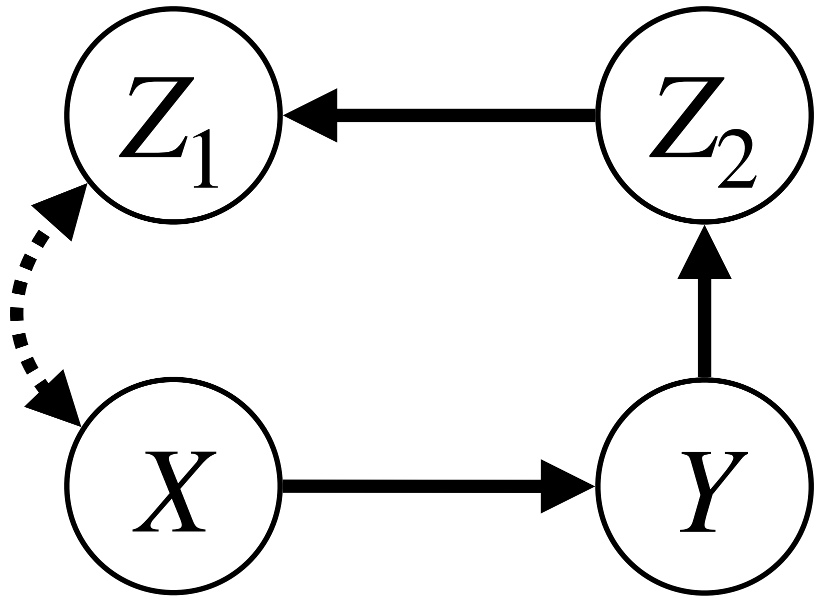
$$P(y \mid do(x)) = \sum_{z_1} P\left(y \mid x, z_1\right) P\left(z_1\right)$$



$\{\}$,

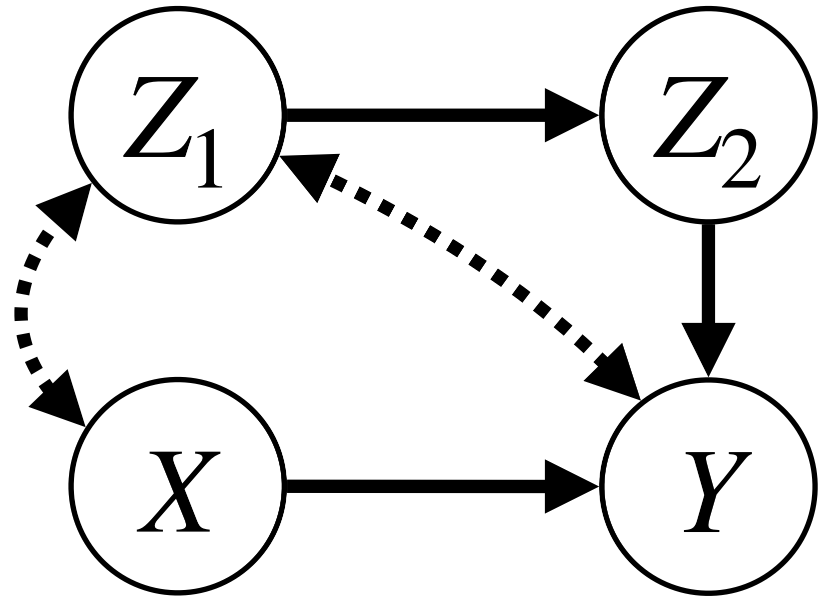
$\{Z_2\}$,
$\{Z_1, Z_2\}$

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



$\{\}$

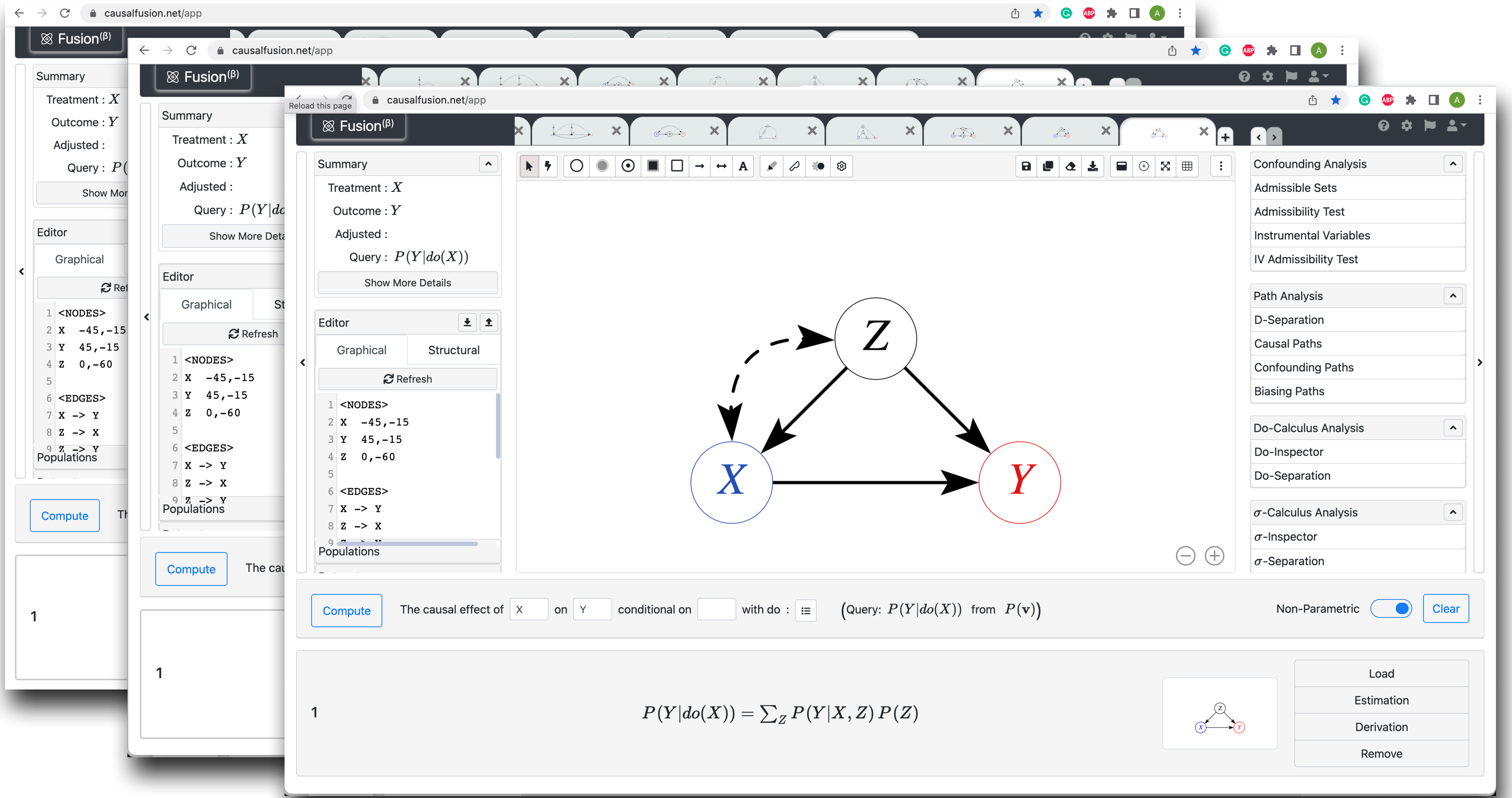
$$P(y \mid do(x)) = P\left(y \mid x\right)$$



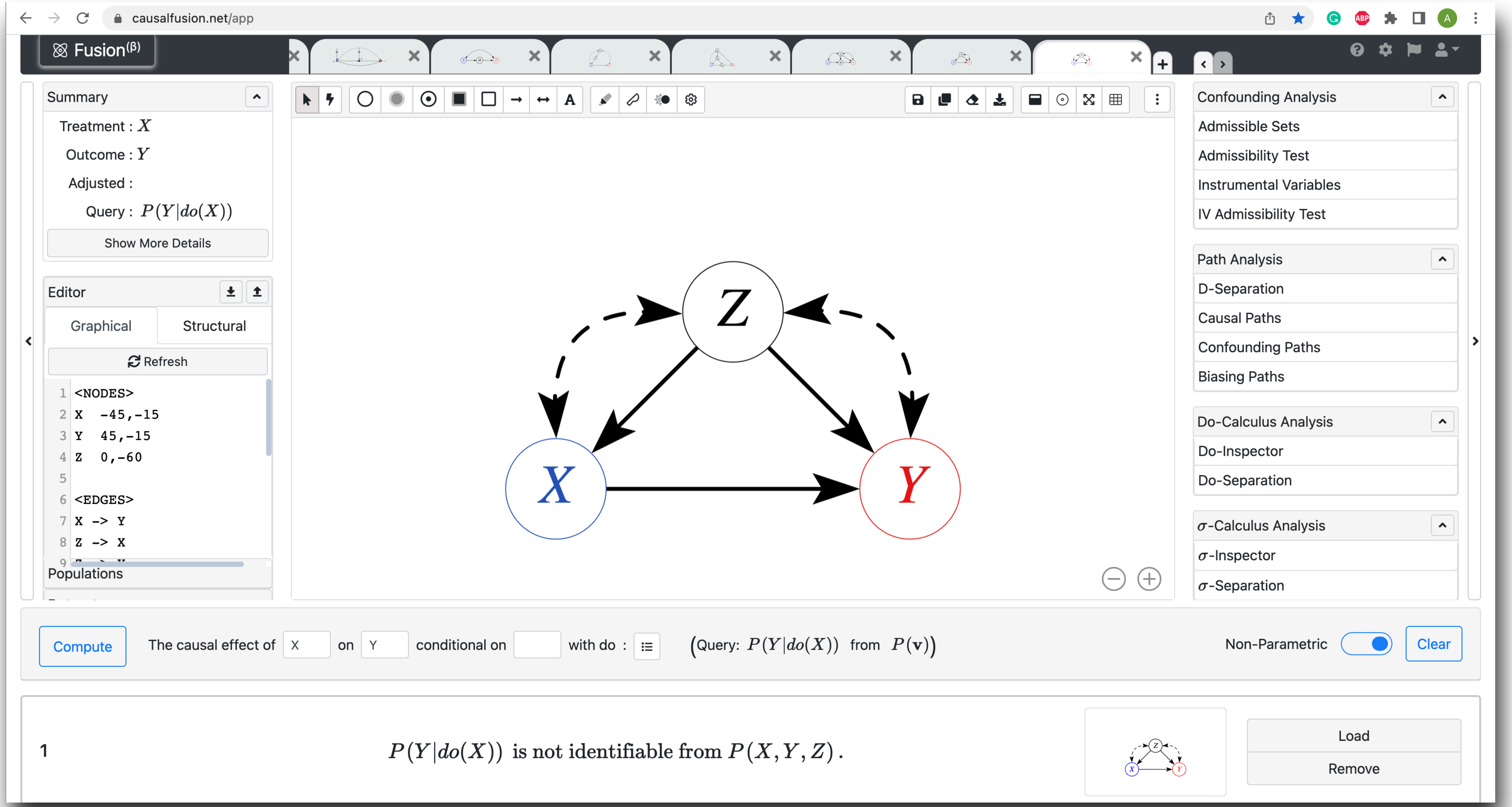
There is no BD Adjustment Set!

$P(y \mid do(x))$ is non-identifiable

http://causalfusion.net



http://causalfusion.net



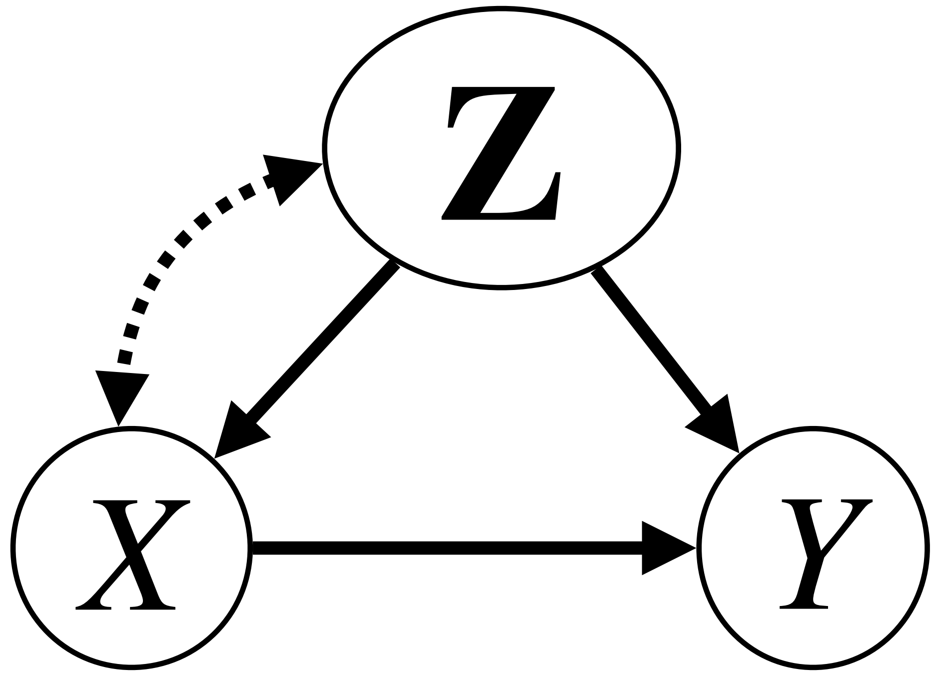
Counterfactual Interpretation of Backdoor

Theorem 4.3.1, Pearl's Primer Book

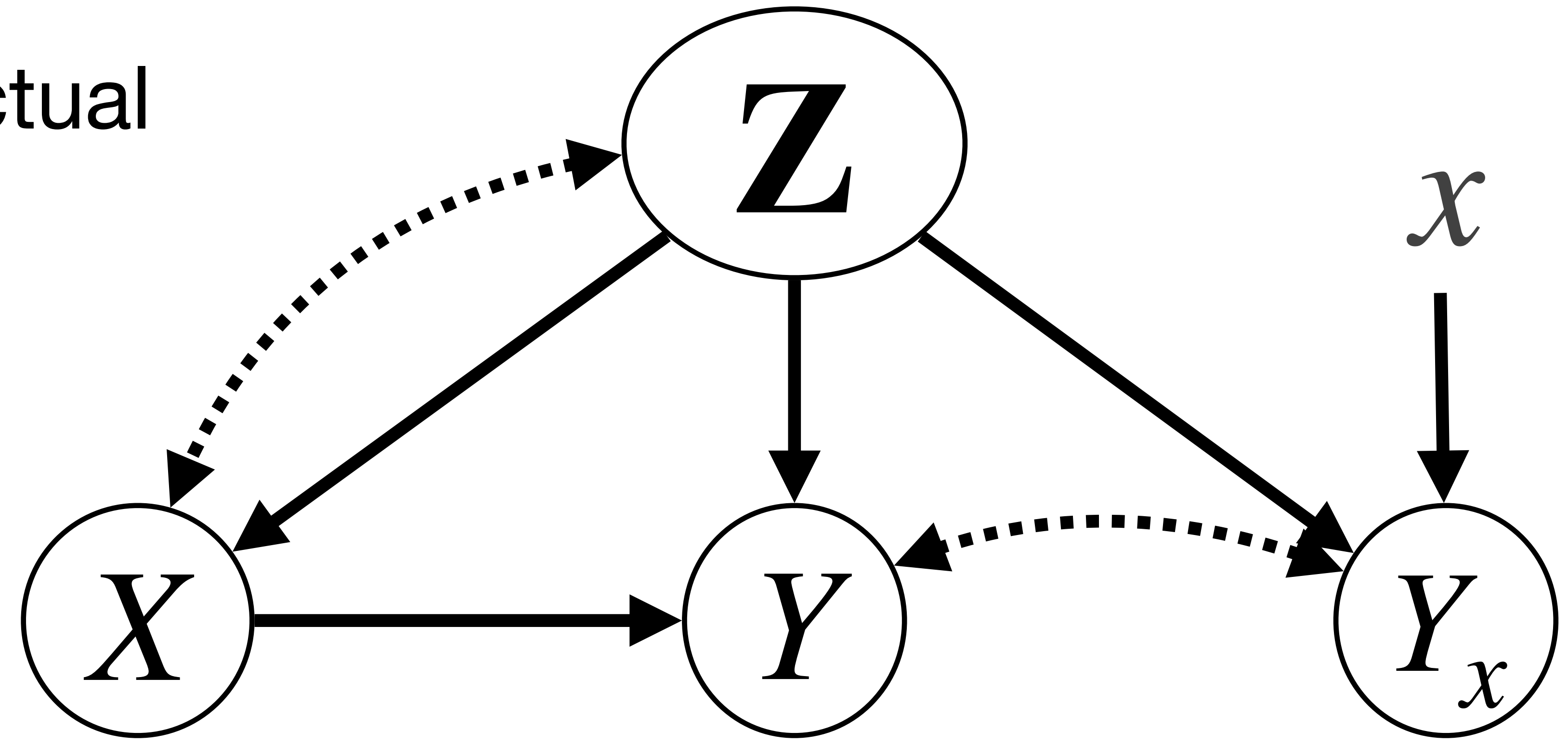
Theorem: If a set $\mathbf{Z}$ satisfies the *backdoor criterion* w.r.t. the ordered pair (X, Y), then, for all x, it holds that $Y_x \perp\!\!\!\perp X \mid \mathbf{Z}$.

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

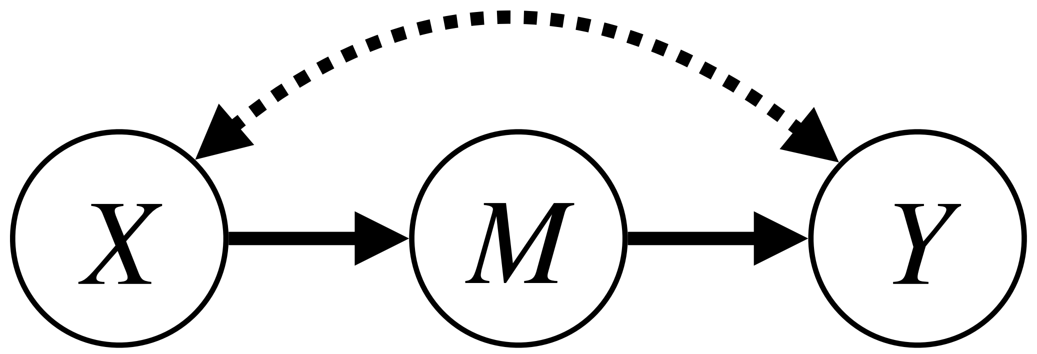


Counterfactual Graph



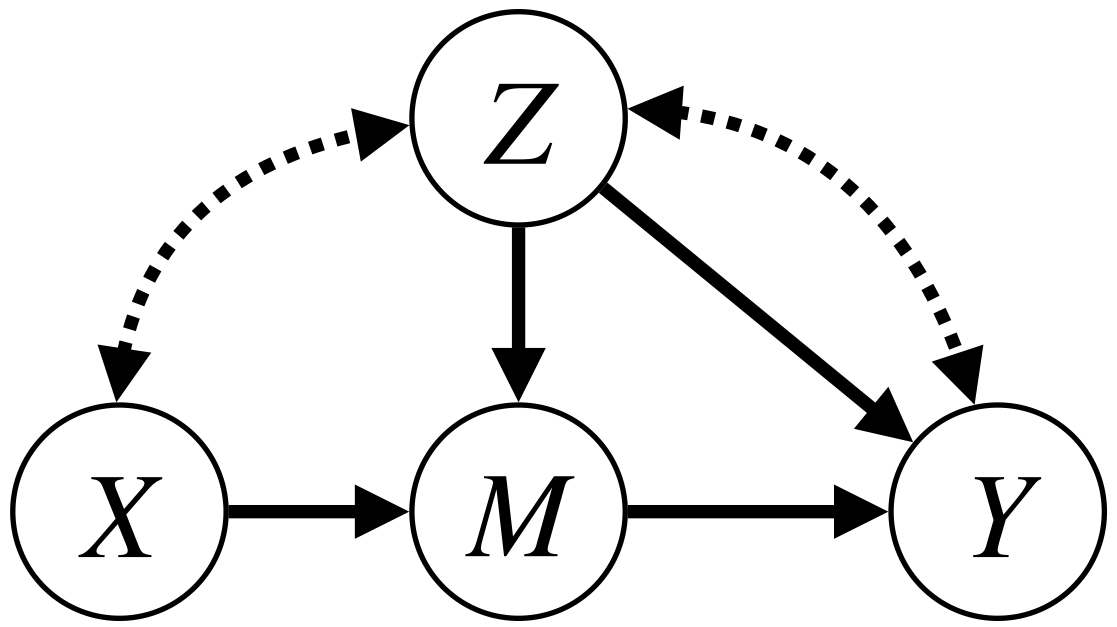
$$Y_x \perp\!\!\!\perp X \mid Z$$

Many Scenarios Beyond Adjustment!

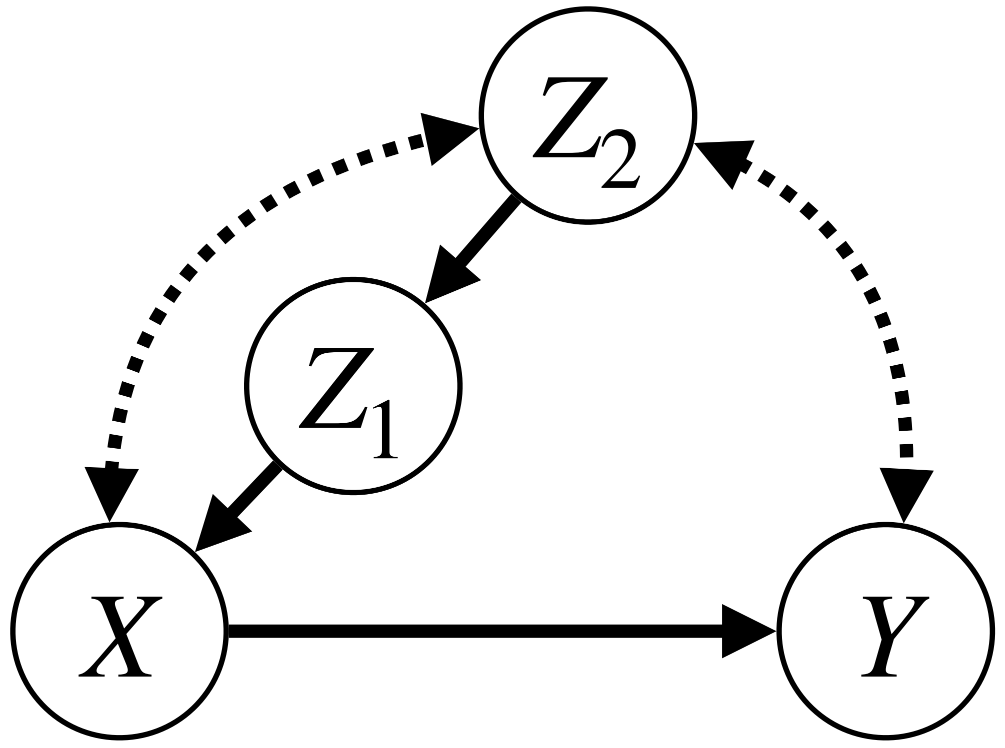


$$P(y\,|\,do(x)) = \sum_m P(m\,|\,x) \sum_{x'} P(y\,|\,m, x')P(x')$$

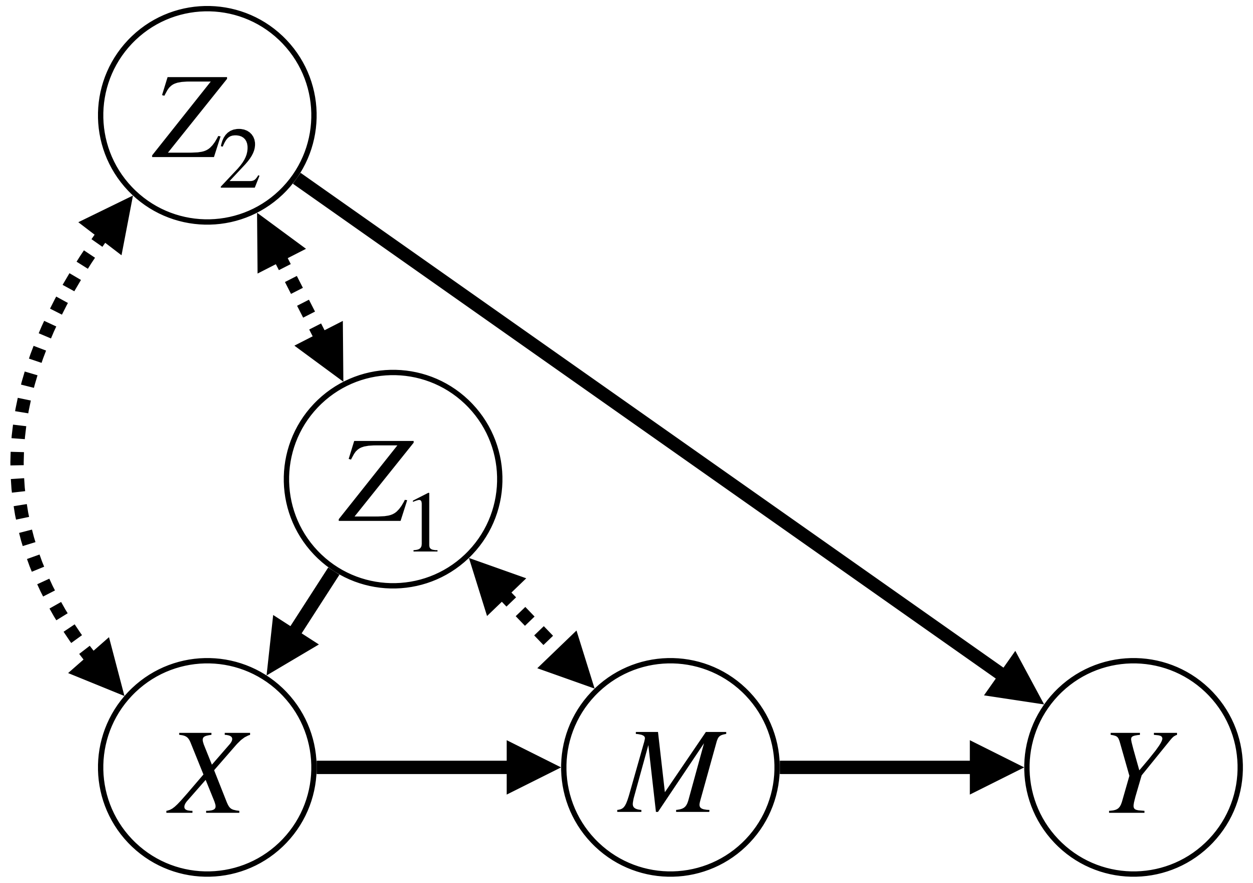
Front-Door



Conditional Front-Door



Napkin



Unnamed

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

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

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

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.

Advances on Effect Identification given a Causal Diagram

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.

J. Correa, S. Lee, E. Bareinboim. (2021) Nested Counterfactual Identification from Arbitrary Surrogate Experiments. In Proceedings of the 35th Annual Conference on Neural Information Processing Systems

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.

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.

Xia, K., Pan, Y.,and Bareinboim, E. (2023) Neural Causal Models for Counterfactual Identification and Estimation. In Proceedings of the 11th International Conference on Learning Representations.

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

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

Learning the Markov Equivalence Class

Causal Discovery:

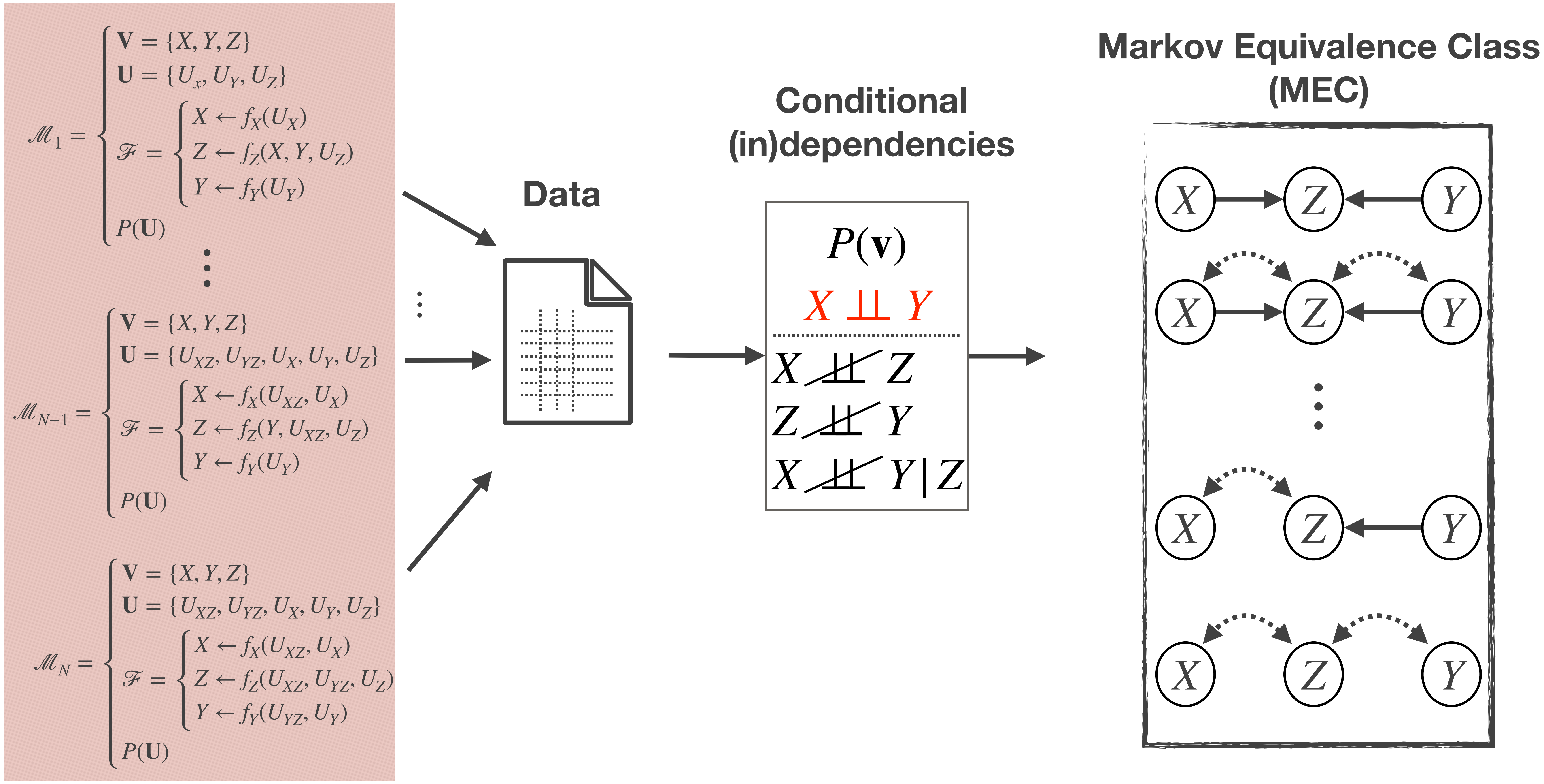
Many models are statistically indistinguishable without additional parametric / distributional assumptions.

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

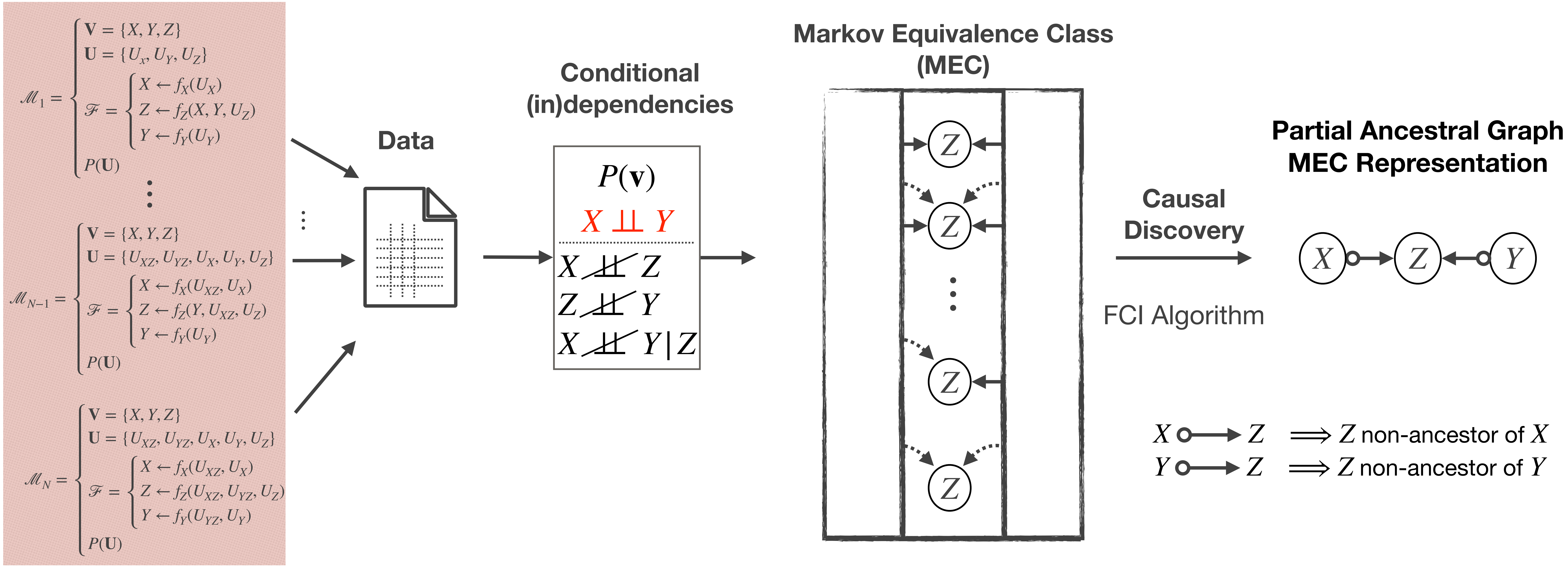
Fast Causal Inference (FCI): Sound and complete causal discovery algorithm, even in the presence of unobserved confounders and selection bias.

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

Causal Discovery: Learning Structural Invariances



Causal Discovery: Learning Structural Invariances

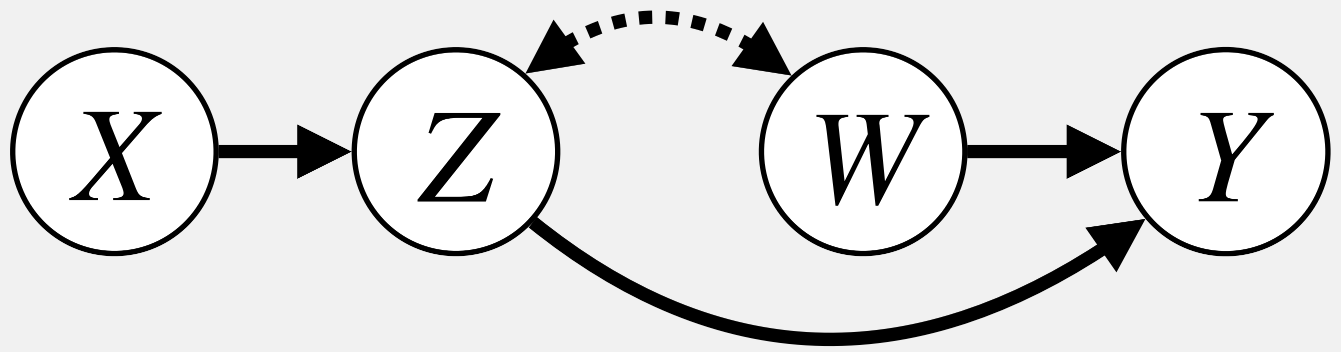


Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896. Link

FCI Algorithm - Pipeline

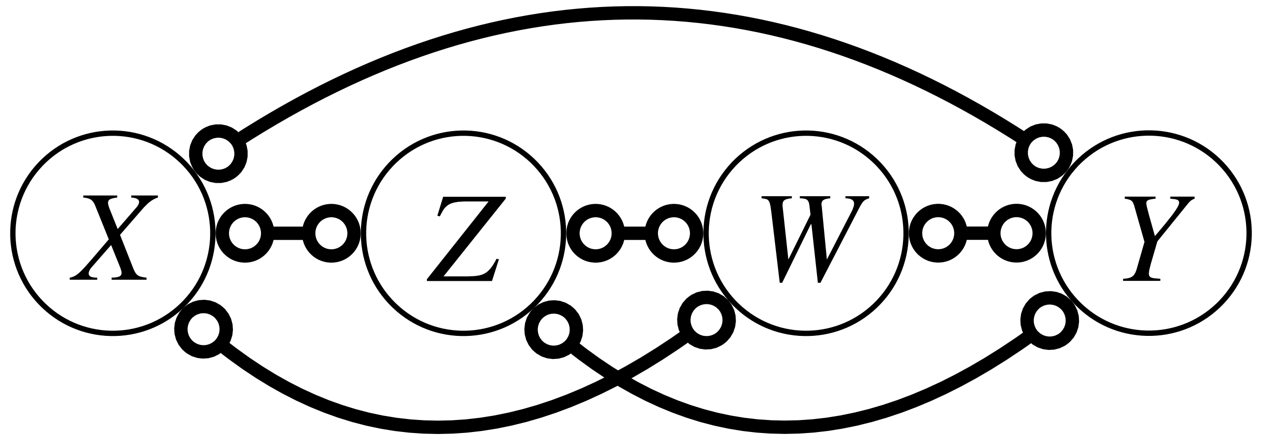
Unknown Reality

True causal diagram



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

Implied by the ADMG using d-separation

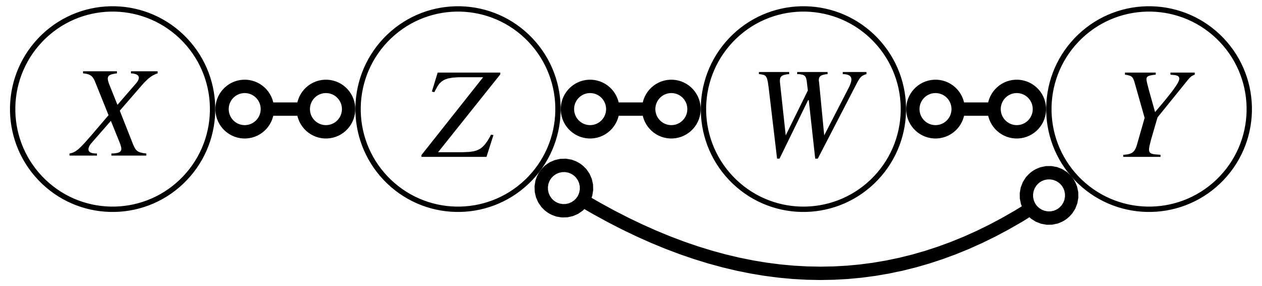


Complete Graph

Conditional Independence Tests

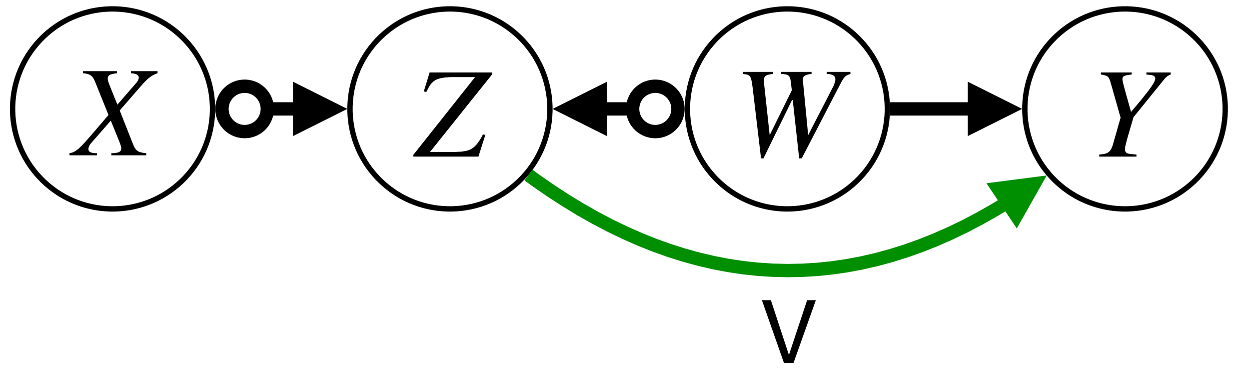
$X \perp\!\!\!\perp W$
$X \perp\!\!\!\perp Y \mid Z, W$

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



Skeleton

FCI Rules
$(R1) - (R10)$



Partial Ancestral Graph (PAG)

Implied by the PAG using m-separation

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

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

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

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

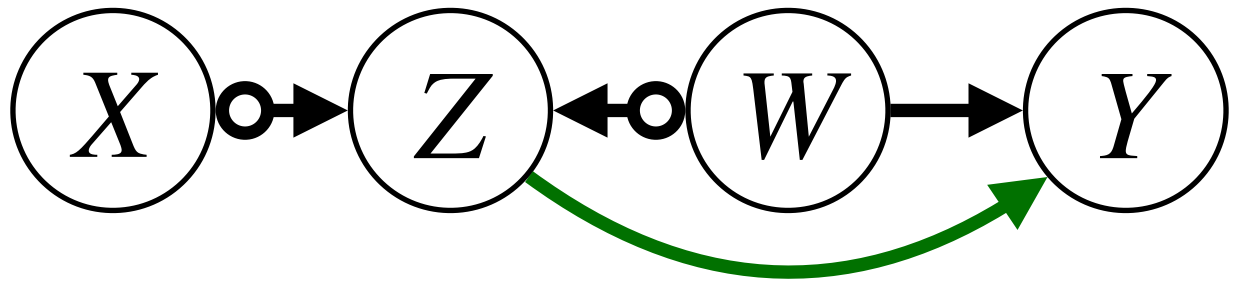
Continuous (conditional Gaussian) or Discrete (Binary, Ordinal, Multinomial) - Linear Regression

- **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

PAG: Representation of the Markov Equivalence Class



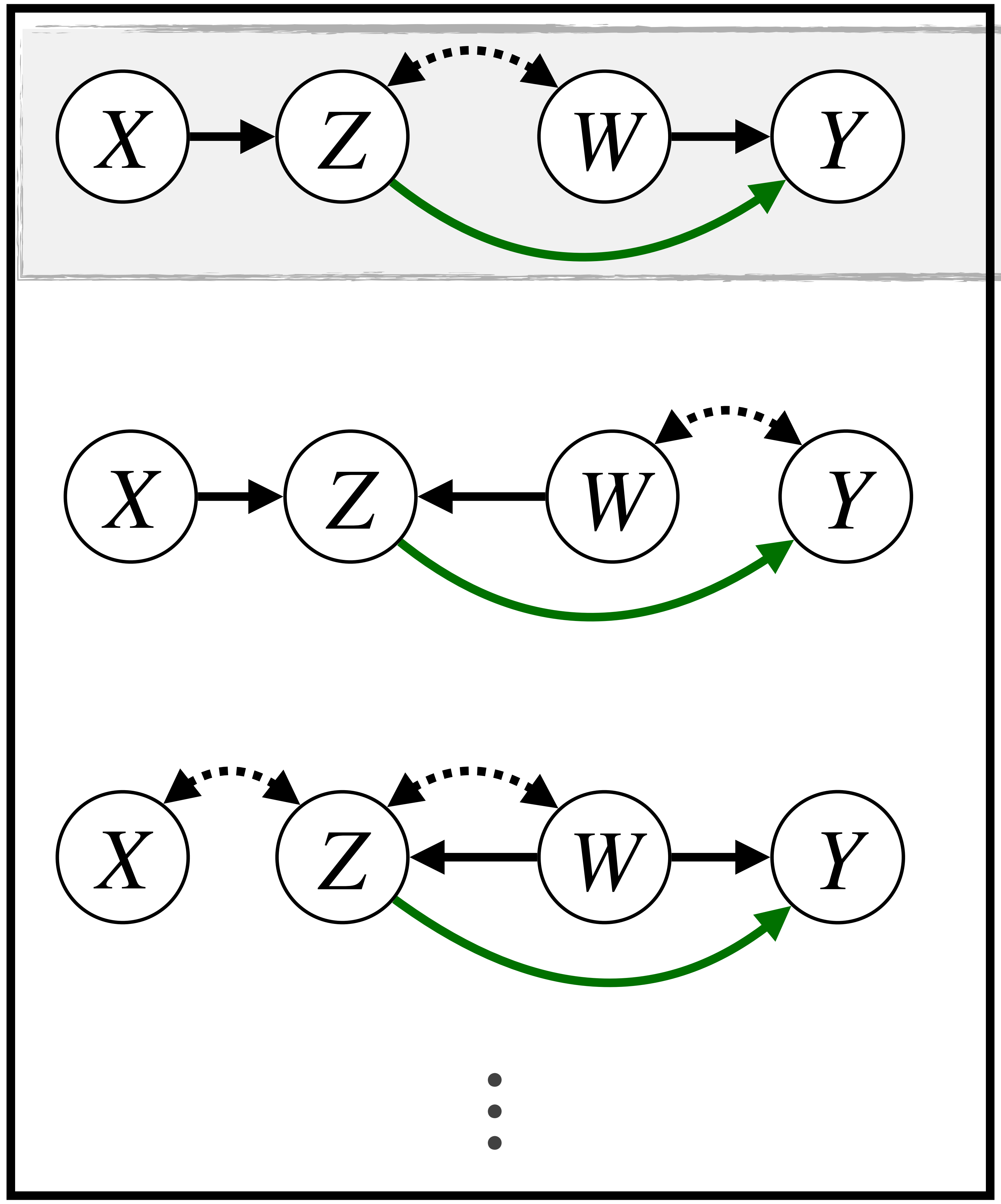
V

Partial Ancestral Graph
(PAG)

Z is not an ancestor of X or W.

Z and W are ancestors of Y.

Z is not confounded with Y.



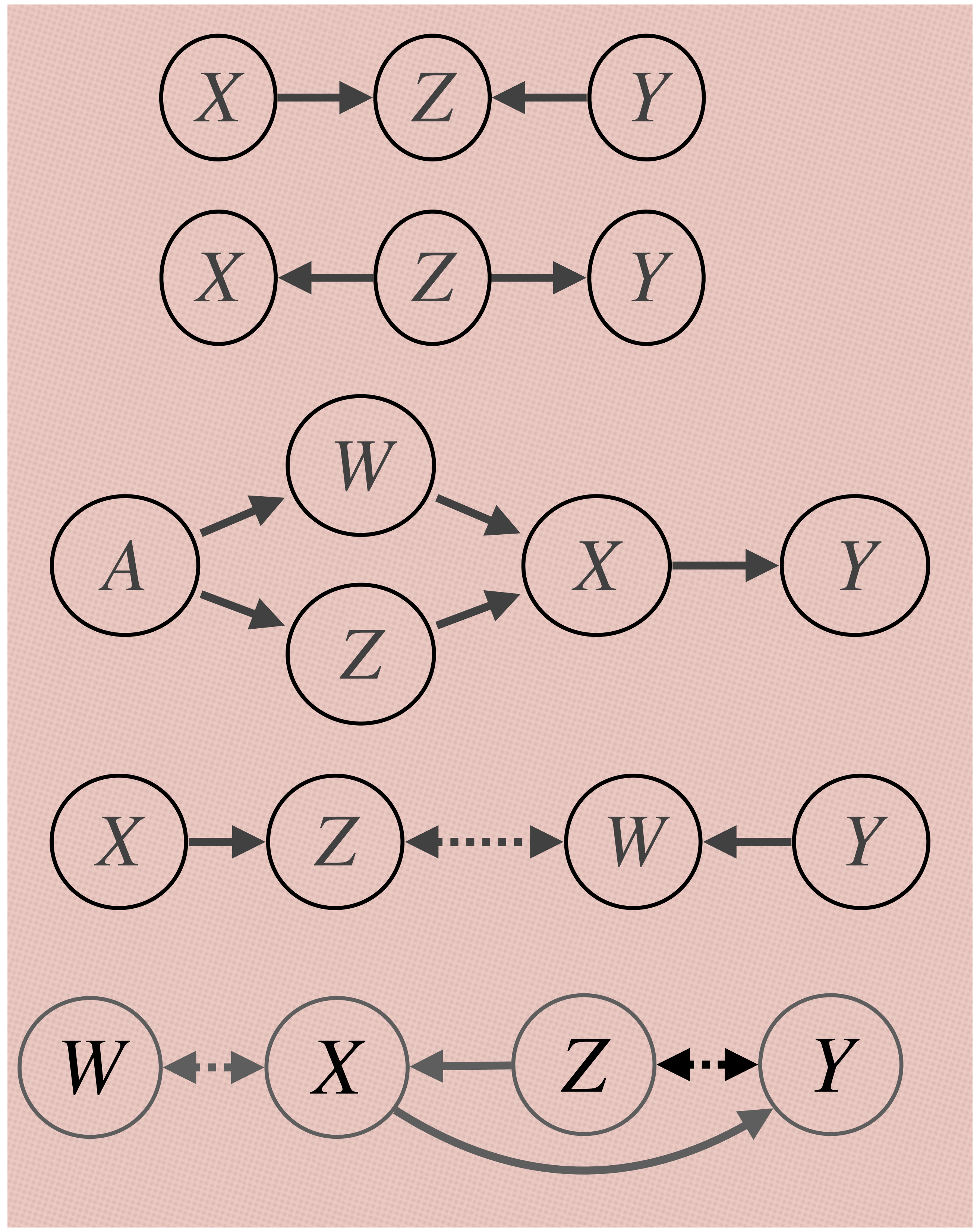
True (unknown) causal diagram

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

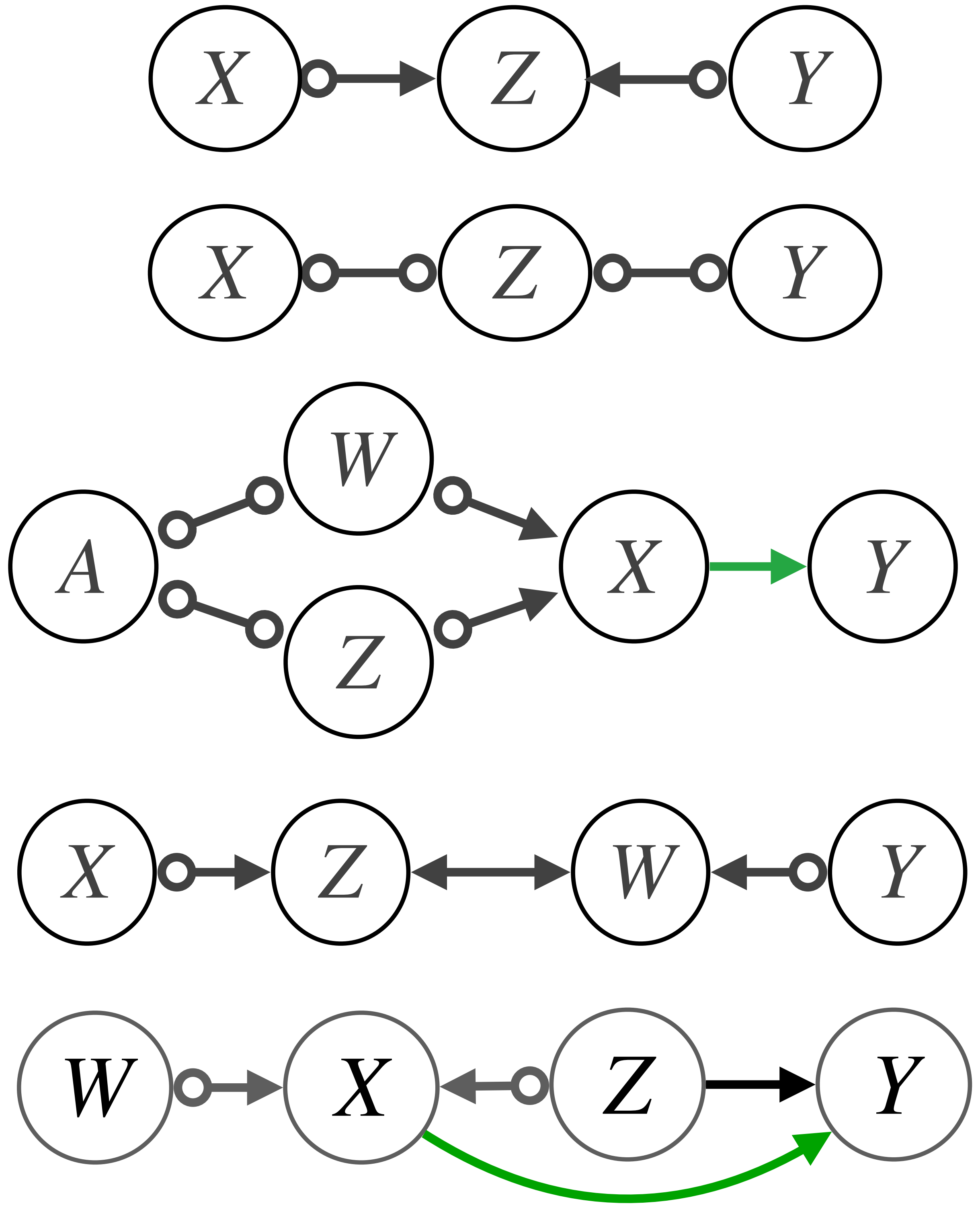
$X \perp\!\!\!\perp Y \mid Z, W$

Fast Causal Inference (FCI) Algorithm

Underlying Causal Diagram



Partial Ancestral Graph

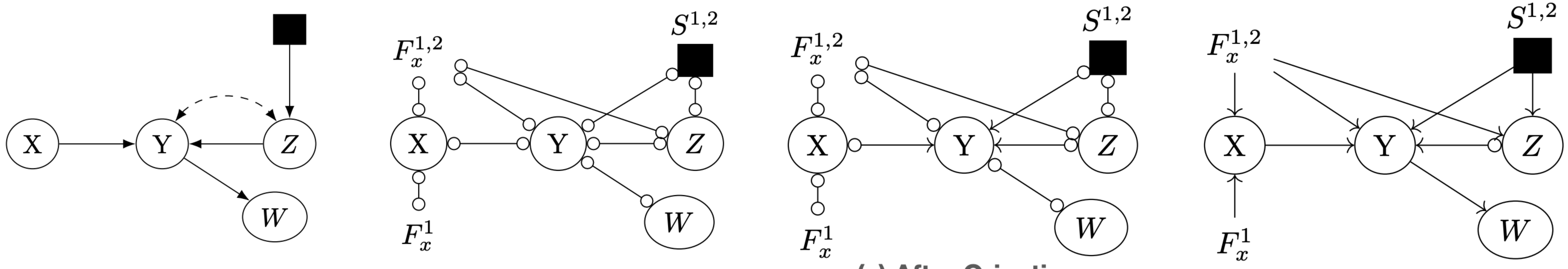


Developments in Causal Discovery with Unobserved Confounding

Going *Beyond* the Markov Equivalence Class:

1. Causal Discovery with Interventional Data

- **Jaber, A., Kocaoglu, M., Shanmugam, K. and Bareinboim, E.,** (2020). Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, *33*, pp.9551-9561.
- **A. Li, A. Jaber, E. Bareinboim.** Causal discovery from observational and interventional data across multiple environments. (2023) In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems* — NeurIPS-23.



(a) G_S (b) Skeleton (c) After Orienting Unshielded Colliders (d) Final S-PAG

Developments in Causal Discovery with Unobserved Confounding

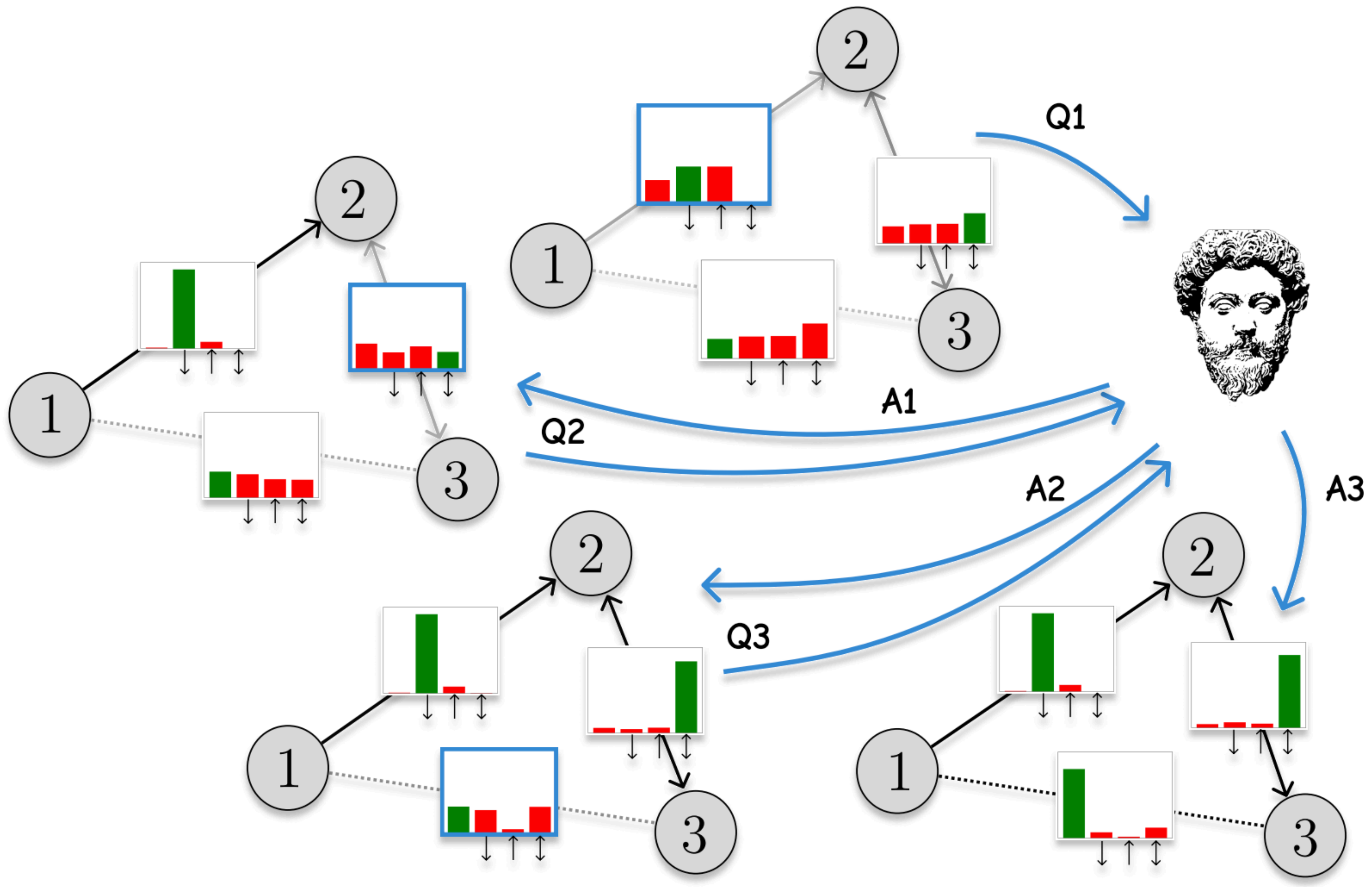
Going *Beyond* the Markov Equivalence Class:

2. Causal Discovery with Prior Knowledge

   - **Wang, T. Z., Qin, T. and Zhou, Z.H.**, (2022). Sound and complete causal identification with latent variables given local background knowledge. *Advances in Neural Information Processing Systems*, *35*, pp.10325-10338.

3. Human-in-the-Loop Probabilistic Causal Discovery

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



Developments in Causal Discovery with Unobserved Confounding

Going *Beyond* the Markov Equivalence Class:

4. Causal Discovery in Linear Models

    - **Tashiro, T., Shimizu, S., Hyvärinen, A., & Washio, T.** (2014). ParceLiNGAM: A causal ordering method robust against latent confounders. *Neural computation*, *26*(1), 57-83.
    - **Wang, Y. S., & Drton, M.** (2023). Causal discovery with unobserved confounding and non-Gaussian data. *Journal of Machine Learning Research*, *24*(271), 1-61.

Relax the causal sufficiency assumption of LinGAN by Shimizu et al., 2006: order / ancestral identifiability under linear systems with non-gaussian error terms

5. Causal Discovery for Additive Noise Models

    - **Van Diepen, M. M., Bucur, I. G., Heskes, T., & Claassen, T.** (2023). Beyond the Markov Equivalence Class: Extending Causal Discovery under Latent Confounding. In *Conference on Causal Learning and Reasoning* (pp. 707-725). PMLR.

FCI-CDC: causal direction criterion (CDC) allows pairwise orientation in (weakly) additive noise models with independent causal mechanisms.

Developments in Causal Discovery with Unobserved Confounding

Learning Dynamic Systems:

1. Causal Discovery with Cycles

   - **Bongers, S., Forré, P., Peters, J., & Mooij, J. M.** (2021). Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, *49*(5), 2885-2915.

   - **Claassen, T. &; Mooij, J.M.. (2023).** Establishing Markov equivalence in cyclic directed graphs. Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence, PMLR 216:433-442, 2023.

2. Causal Discovery from Time-Series Data

   - **Gerhardus, A., & Runge, J.** (2020). High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in Neural Information Processing Systems (NeurIPS 2020)*, *33*, 12615-12625.

Causal Identification from PAGs

Can we identify causal effects from the equivalence class?

Effect Identification:

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

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

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

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

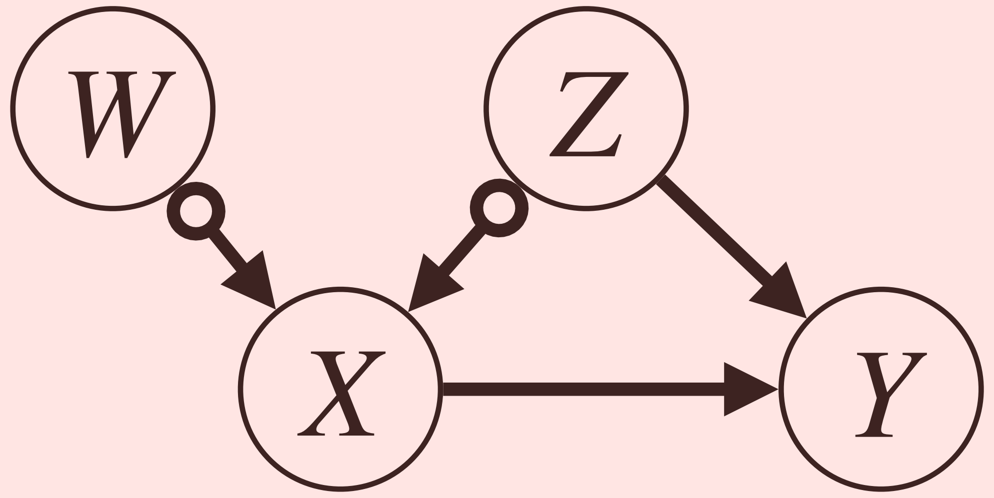
Effect Identification in Markov Equivalence Classes

Can be constructed in a fully data-driven way!

1 **Query**

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

2 **PAG**



3 **Data**

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

Inference Engine

Solution

yes / no

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

Inferred (Interventional) Distribution ← Available (Observational) Distribution 👍

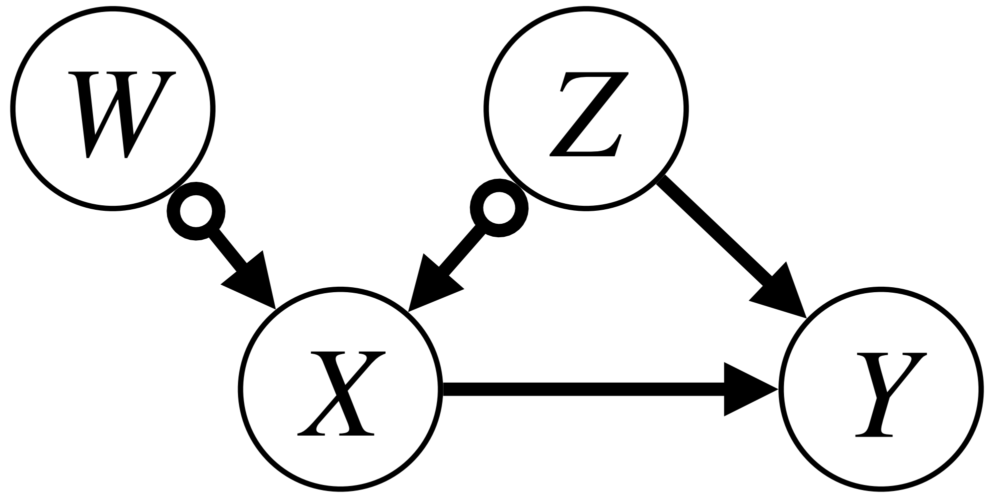
Identification via Adjustment in Markov Equivalence Classes

Identification is possible only when the Generalized Adjustment Criterion applies.

1 **Query**

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

2 **PAG**



3 **Data**

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

Adjustment Criterion

Solution

yes / no

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

Inferred (Interventional) Distribution ← Available (Observational) Distribution

👍

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

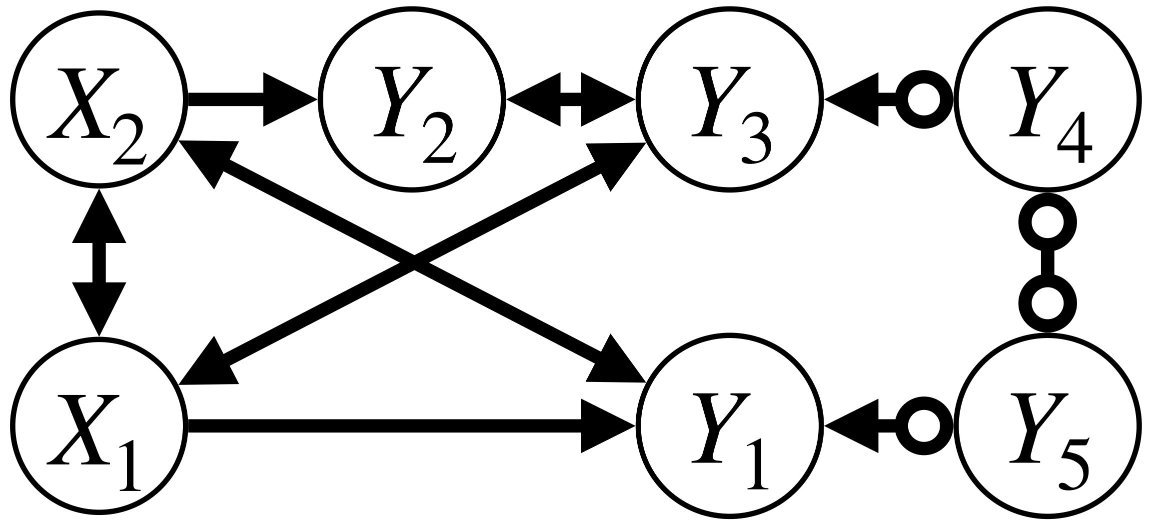
General Identification in Markov Equivalence Classes

Complete algorithms,
available at the *PAGId* R package:
https://github.com/adele/PAGId

1 **Query**

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

2 **PAG**



3 **Data**

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

IDP / CIDP

Solution
yes / no

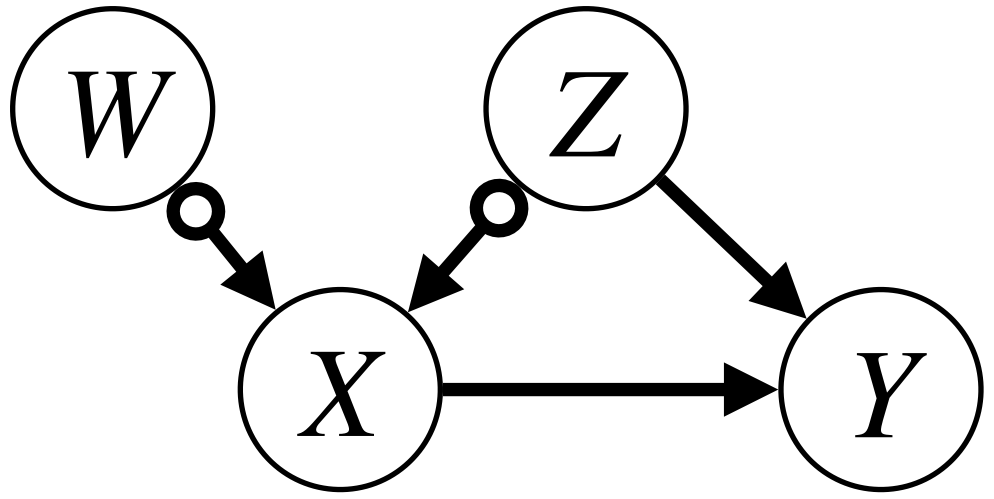
$$P(y \mid do(x)) = \frac{P(y_1, y_4, y_5 \mid x_1) \, . \, P(y_2 y_3, y_4, y_5 \mid x_2)}{P(y_4, y_5)}$$

Inferred (Interventional) Distribution ← Available (Observational) Distribution

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

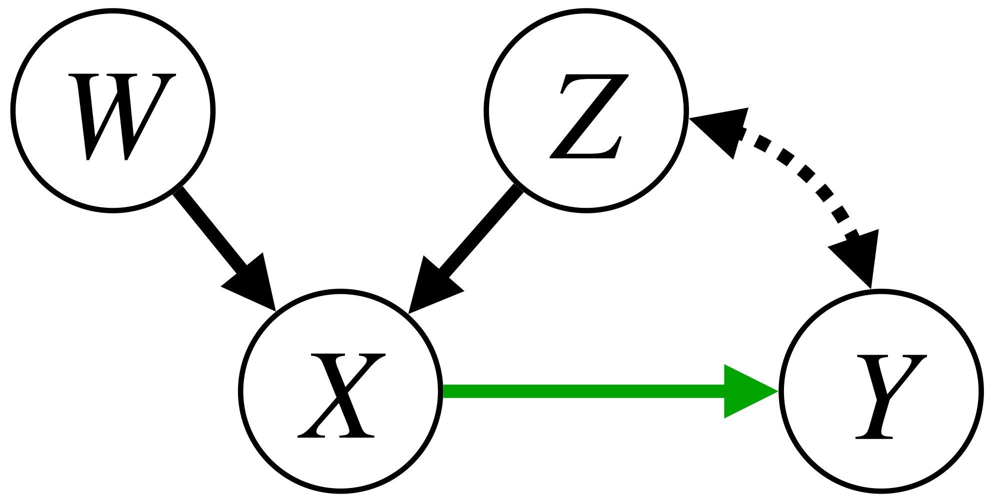
Effect Identifiabiliy given a PAG

$\mathscr{P}$



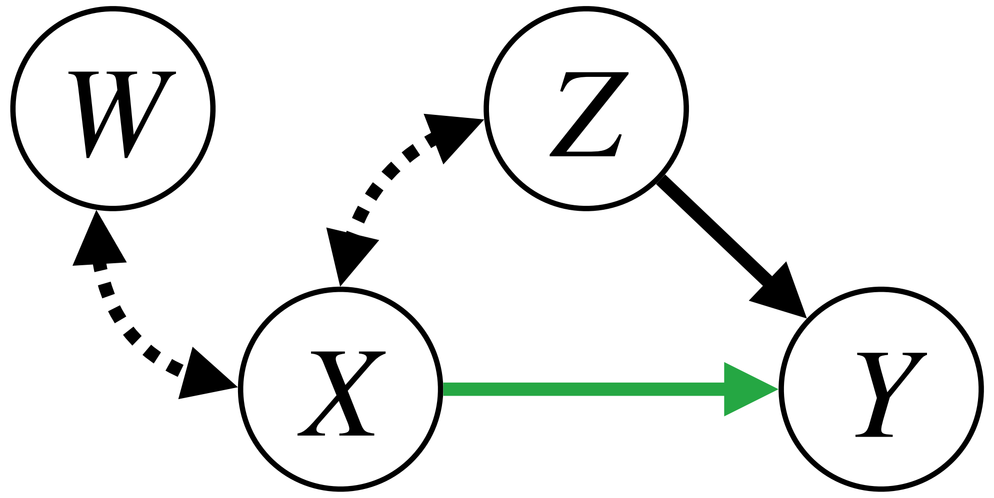
$$P(y \mid do(x)) = \sum_z P\left(y \mid x, z\right) P\left(z\right)$$

G_1



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

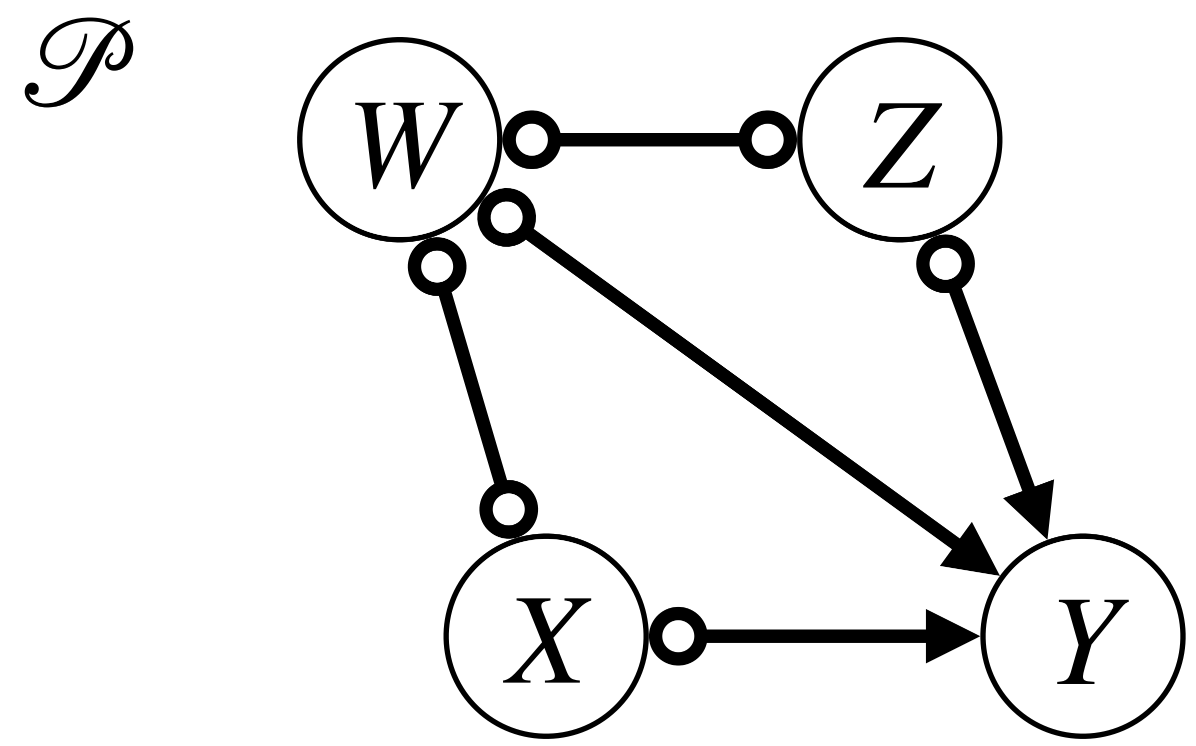
G_2



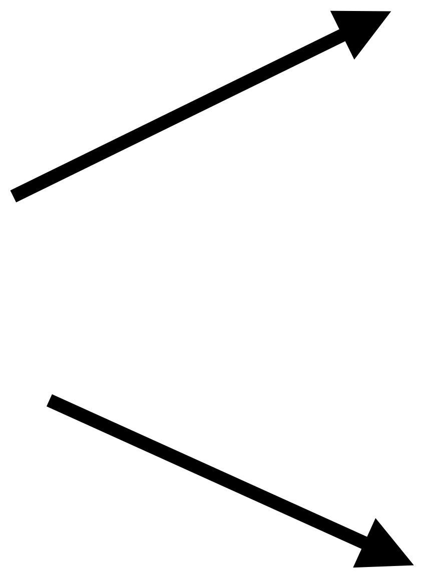
$$P(y \mid do(x)) = \sum_z P(y \mid x, z) P(z)$$

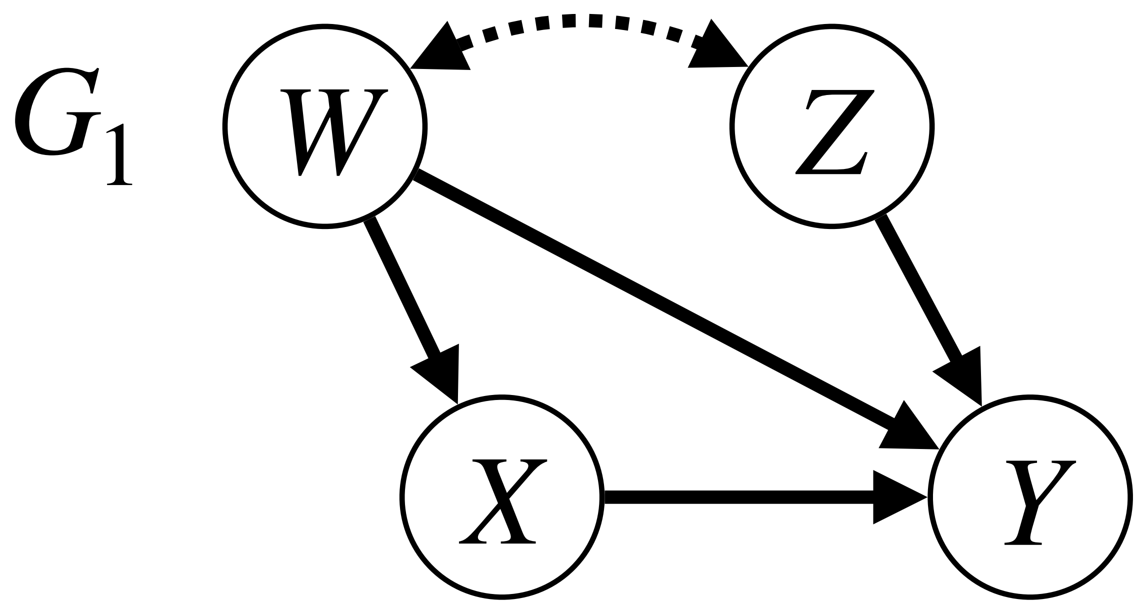
An effect identifiable in a PAG $\mathscr{P}$ is identifiable in all causal diagrams G in the Markov Equivalence Class using the same identification formula!

Effect Non-Identifiabiliy given a PAG

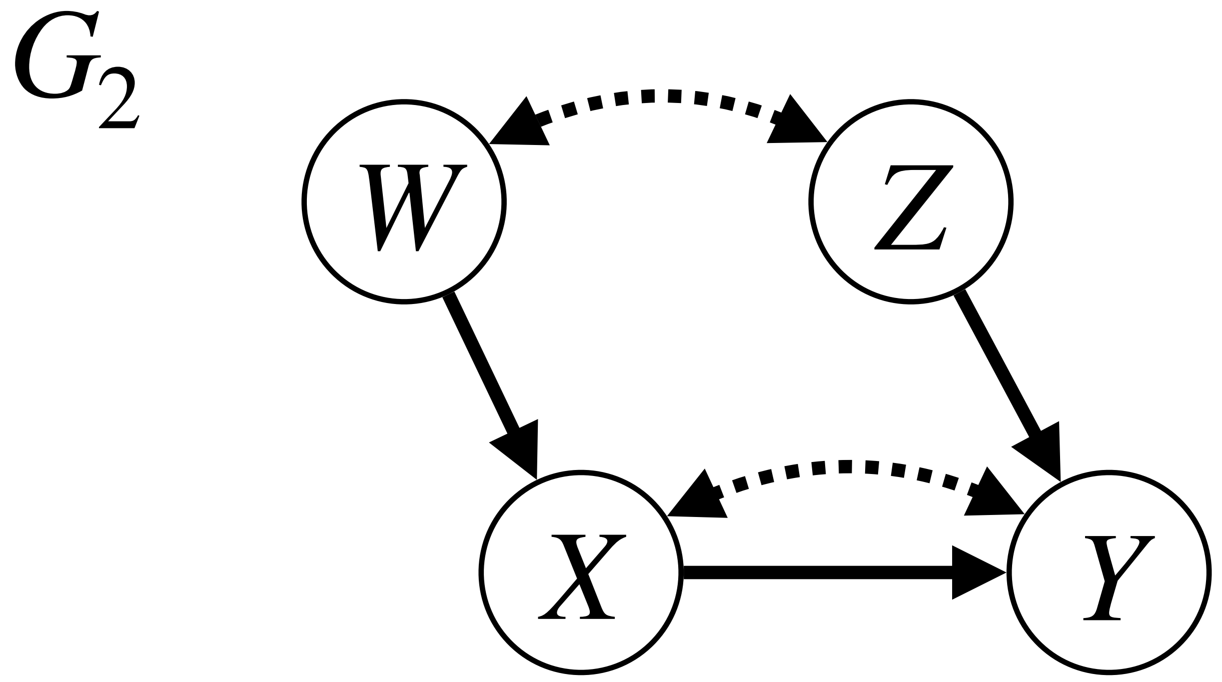


$P(y\,|\,do(x))$ is
not identifiable





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

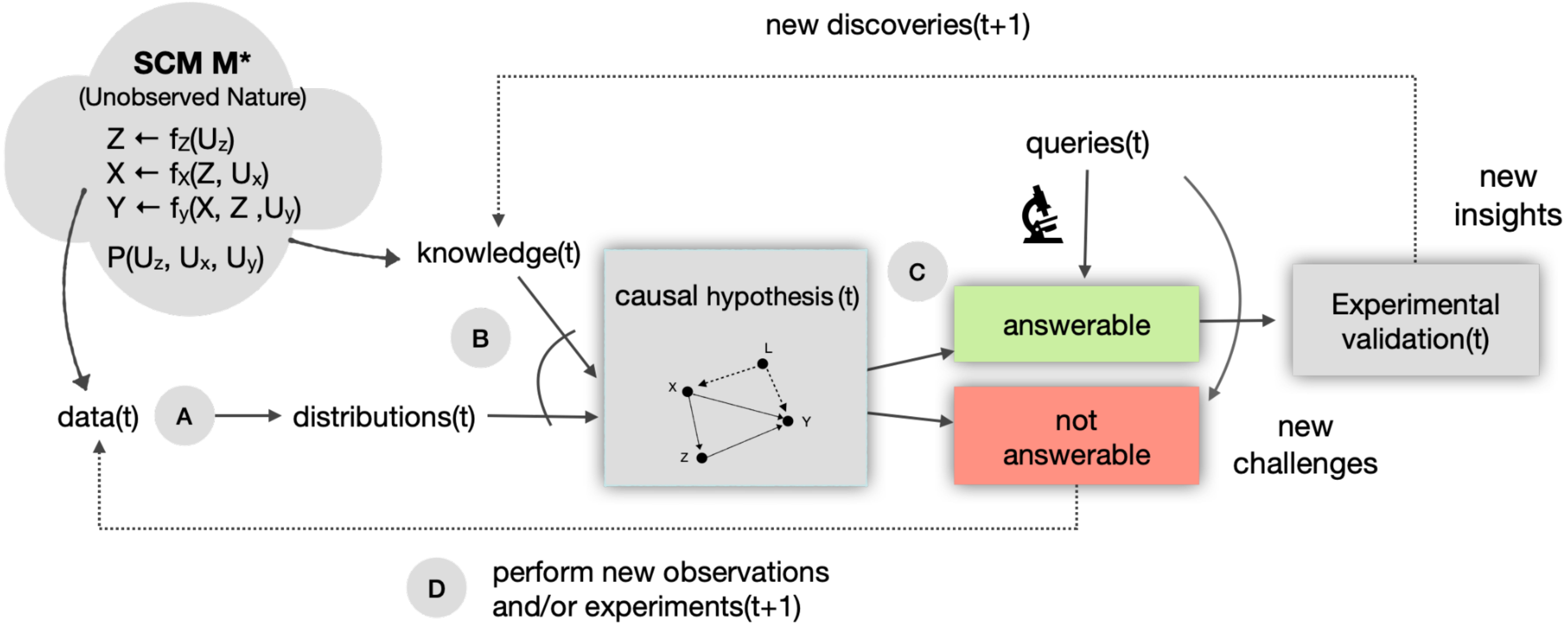


$P(y\,|\,do(x))$ is
not identifiable

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

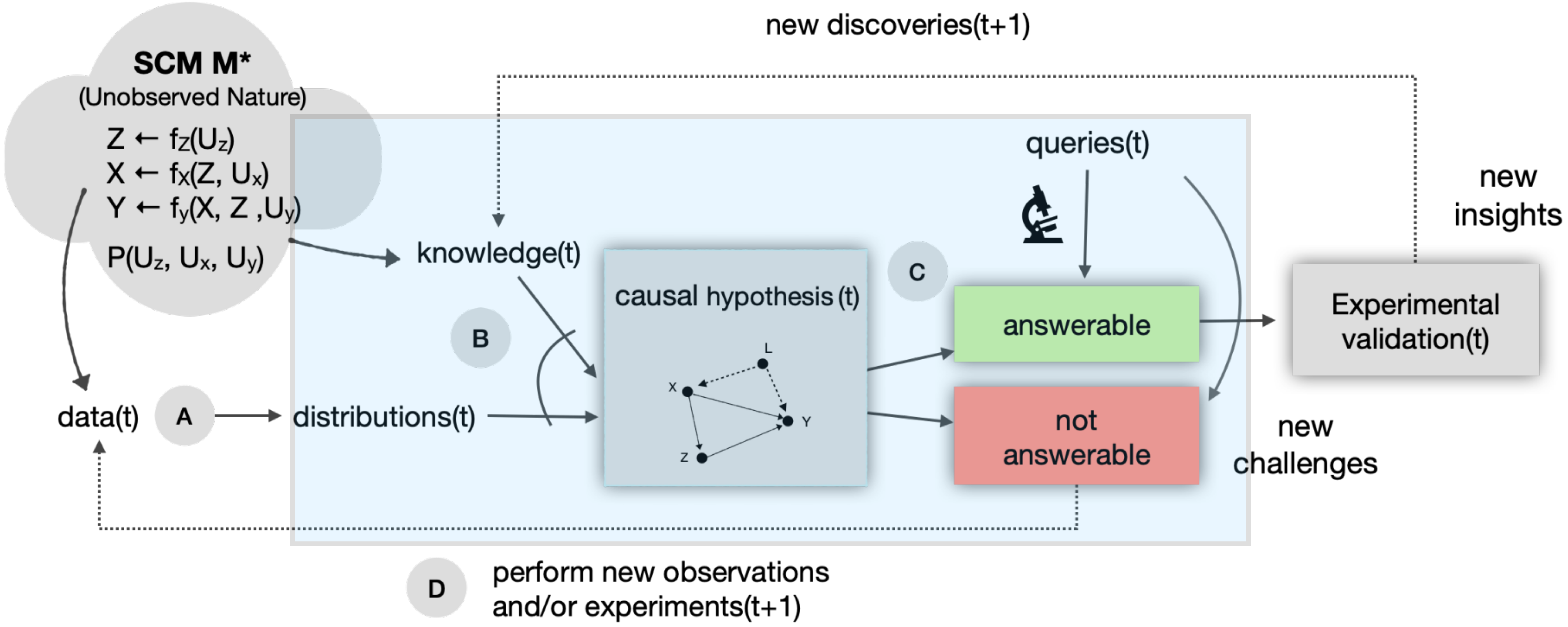
Continuous Process of Scientific Discovery and Causal Hypothesis Refinement



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

Many other Topics in Causal Inference

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

Causal Representation Learning & Causal Abstraction

Toward Causal Representation Learning

This article reviews fundamental concepts of causal inference and relates them to crucial open problems of machine learning, including transfer learning and generalization, thereby assaying how causality can contribute to modern machine learning research.

By Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612-634.

Coarse-grained causal models:

Causal Consistency of Structural Equation Models

Paul K. Rubenstein[*12]**, Sebastian Weichwald**[*13]**, Stephan Bongers**[4]**, Joris M. Mooij**[4]
Dominik Janzing[1]**, Moritz Grosse-Wentrup**[1]**, Bernhard Schölkopf**[1]
*Equal contribution
[1]Empirical Inference, MPI for Intelligent Systems, [2]Machine Learning Group, University of Cambridge,
[3]Max Planck ETH Center for Learning Systems, [4]Informatics Institute, University of Amsterdam

The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Causal Effect Identification in Cluster DAGs

Tara V. Anand[*1]**, Adele H. Ribeiro**[*2]**, Jin Tian**[3]**, Elias Bareinboim**[2]
[1]Department of Biomedical Informatics, Columbia University
[2]Department of Computer Science, Columbia University
[3]Department of Computer Science, Iowa State University
tara.v.anand@columbia.edu, adele@cs.columbia.edu, jtian@iastate.edu, eb@cs.columbia.edu

Neural Causal Abstractions

Kevin Xia and **Elias Bareinboim**
Causal Artificial Intelligence Lab
Columbia University
{kevinmxia, eb}@cs.columbia.edu

Causal Reinforcement Learning

http://crl.causalai.net

TASK 1

Generalized Policy Learning

combining online + offline learning

Learn policy Π by systematically combining offline (L_1) and online (L_2) modes of interaction.

TASK 2

When and Where to Intervene?

refining the policy space

Identify subset of L_2 to refine the policy space do(Π(X)) based on topological constraints implied by *M* on *G*.

TASK 3

Counterfactual Decision-Making

changing optimization function based on intentionality, free will, and autonomy

Optimization criterion based on counterfactuals and L_3-based randomization (instead of L_2/do()-counterpart).

TASK 4

Generalizability & Robustness of Causal Claims

transportability & structural invariances

Generalize policy based on structural invariances shared across training (SCM *M*) and deployment environments (*M**).

TASK 5

Learning Causal Models

discovering the causal structure with observation and experiments

Learn the causal graph *G* (of *M*) by systematically combining observations (L_1) and experimentation (L_2).

TASK 6

Causal Imitation Learning

policy learning with unobserved rewards

Construct L_2-policy based on partially observable L_1-data coming from an expert with unknown reward function.

By Elias Bareinboim's Research Group

Fairness and Mediation Analysis

A Causal Framework for Decomposing Spurious Variations

Drago Plecko and **Elias Bareinboim**
Department of Computer Science
Columbia University
dp3144@columbia.edu, eb@cs.columbia.edu

D. Plecko, E. Bareinboim. A Causal Framework for Decomposing Spurious Variations. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems* — NeurIPS-23.

Foundations and Trends® in Machine Learning

Causal Fairness Analysis

A Causal Toolkit for Fair Machine Learning

Suggested Citation: Drago Plečko and Elias Bareinboim (2024), "Causal Fairness Analysis", Foundations and Trends® in Machine Learning: Vol. 17, No. 3, pp 1–238. DOI: 10.1561/2200000106.

Drago Plečko
Seminar für Statistik, ETH Zürich
drago.plecko@stat.math.ethz.ch

Elias Bareinboim
Department of Computer Science, Columbia University
eb@cs.columbia.edu

Individual Treatment Effect (ITE) Estimation

Generalization Bounds and Representation Learning for Estimation of Potential Outcomes and Causal Effects

Fredrik D. Johansson FREDRIK.JOHANSSON@CHALMERS.SE
Chalmers University of Technology
Göteborg, 412 96, Sweden

Uri Shalit URISHALIT@TECHNION.AC.IL
Technion - Israel Institute of Technology
Haifa, 3200003, Israel

Nathan Kallus KALLUS@CORNELL.EDU
Cornell University
New York, NY 10044, USA

David Sontag DSONTAG@CSAIL.MIT.EDU
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Johansson, F.D., Shalit, U., Kallus, N. and Sontag, D., 2022. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, *23*(166), pp.1-50.

Other related works cited within, such as:

Estimating individual treatment effect: generalization bounds and algorithms

Uri Shalit, Fredrik D. Johansson, David Sontag *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70:3076-3085, 2017.

Learning Representations for Counterfactual Inference

Fredrik D. Johansson* FREJOHK@CHALMERS.SE
CSE, Chalmers University of Technology, Göteborg, SE-412 96, Sweden

Uri Shalit* SHALIT@CS.NYU.EDU
David Sontag DSONTAG@CS.NYU.EDU
CIMS, New York University, 251 Mercer Street, New York, NY 10012 USA

* **Equal contribution**

Data-Driven Covariate Selection for Adjustment

Finding Valid Adjustments under Non-ignorability with Minimal DAG Knowledge

Abhin Shah
MIT
abhin@mit.edu

Karthikeyan Shanmugam
IBM Research
karthikeyan.shanmugam2@ibm.com

Kartik Ahuja
Mila
kartik.ahuja@mila.quebec

Abhin Shah, Karthikeyan Shanmugam, and Kartik Ahuja. Finding valid adjustments under non-ignorability with minimal DAG knowledge. In *International Conference on Artificial Intelligence and Statistics (AISTATS - 2022)*, pages 5538–5562. PMLR, 2022.

Differentiable Causal Backdoor Discovery

Limor Gultchin
University of Oxford
The Alan Turing Institute

Matt J. Kusner
University College London
The Alan Turing Institute

Varun Kanade
University of Oxford
The Alan Turing Institute

Ricardo Silva
University College London
The Alan Turing Institute

Partial Effect Identification

Stochastic Causal Programming for Bounding Treatment Effects

Kirtan Padh KIRTAN.PADH@TUM.DE
Helmholtz AI, Helmholtz Munich & Technical University Munich

Jakob Zeitler
University College London

David Watson
King's College London

Matt Kusner
University College London

Ricardo Silva
University College London

Niki Kilbertus
Helmholtz AI, Helmholtz Munich & Technical University Munich

Kirtan Padh, Jakob Zeitler, David Watson, Matt Kusner, Ricardo Silva, Niki Kilbertus; *Proceedings of the Second Conference on Causal Learning and Reasoning,* PMLR 213:142-176

Related: **Jakob Zeitler, and Ricardo Silva**. (2022) The Causal Marginal Polytope for Bounding Treatment Effects arXiv preprint arXiv:2202.13851 - https://arxiv.org/pdf/2202.13851.pdf

Thank you! :)

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

adele.ribeiro@uni-marburg.de