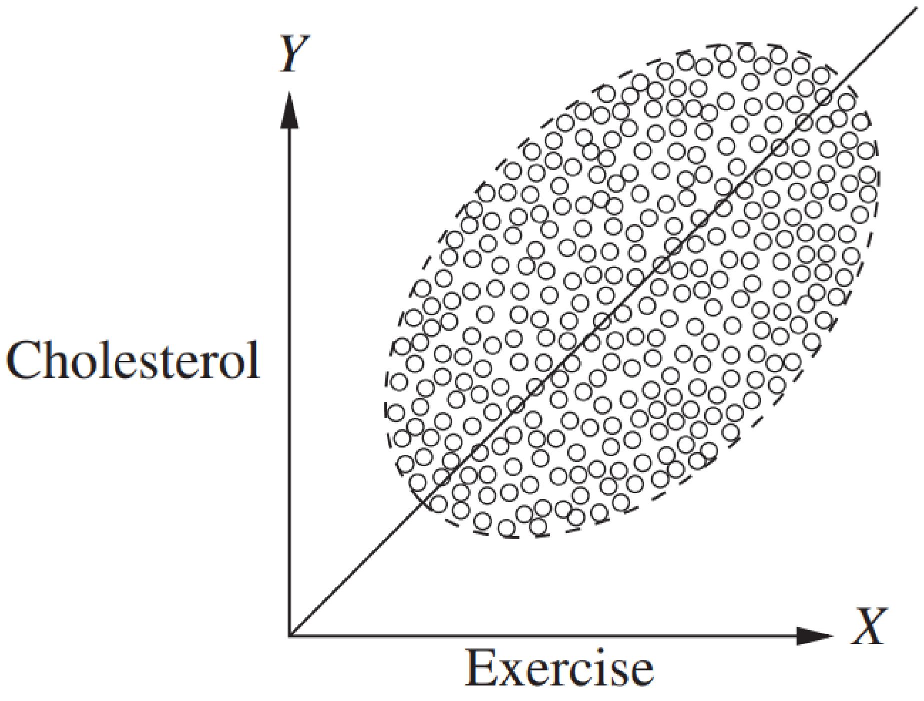
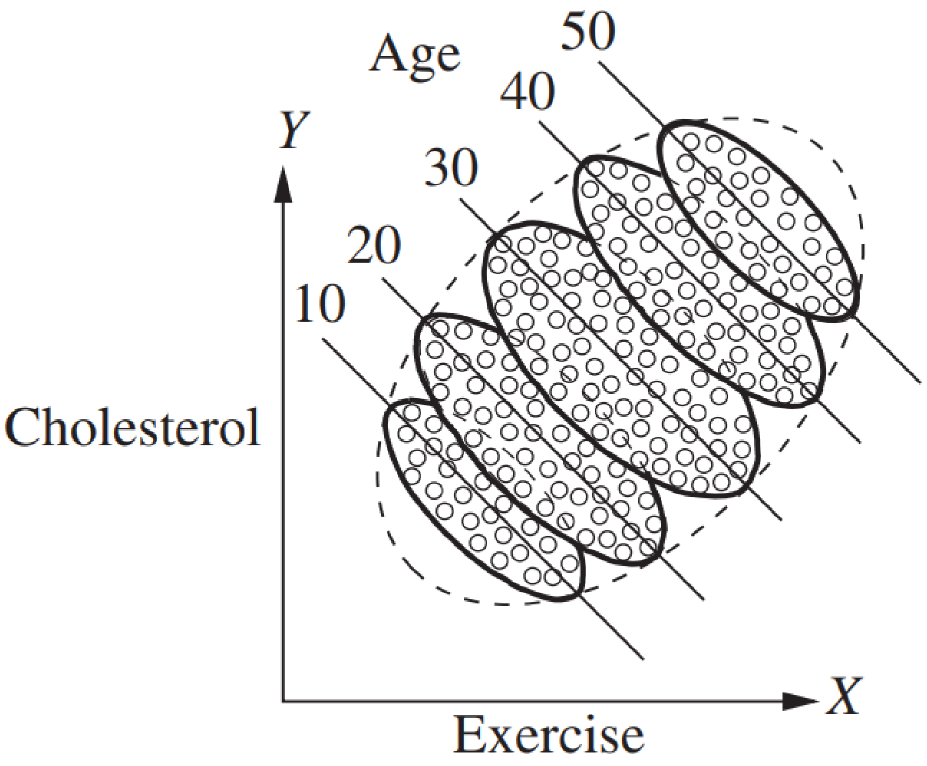
A Brief Introduction to Causal Discovery and Causal inference

Correlation vs. Causation





Chapter 1 (pp. 1-7 & 24-33) of J. Pearl, M. Glymour, and N.P. Jewell, Causal Inference in Statistics: A Primer, Wiley, 2016.

Correlation Is Not Causation

The gold rule of causal analysis: no causal claim can be established purely by a statistical method.

Statistical Implications of Causality

- Better to talk of (in)dependence rather than correlation.
- Most statisticians would agree that causality does tell us something about dependence.
- But dependence does tell us something about causality too.

(Conditional) Independence

- Two random variables X and Y are called independent if for each values of (X, Y) denoted by (x, y),
  - $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$
  - Denoted by $X \perp Y$ or $(X \perp Y)_D$
  - Otherwise they are dependent
- Two random variables X and Y are called conditionally independent given Z, if for each values of (X, Y, Z) denoted by (x, y, z),
  - $P(X = x, Y = y | Z = z) = P(X = x | Z = z) \cdot P(Y = y | Z = z)$
  - Denoted by $X \perp Y | Z$ or $(X \perp Y | Z)_D$
  - Otherwise they are conditionally dependent

Statistical Implications of Causality

- Reichenbach's *Common Cause Principle* (1956) links causality and (in)dependence.

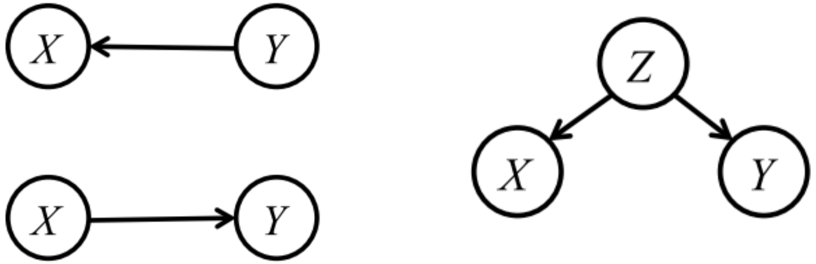
It seems that a dependence between events A and B indicates either that A causes B, or that B causes A, or that A and B have a common cause.

⬇

If A and B have a common cause C (only), then conditioning on C would make A and B independent. In this case, C is said to 'screen off' the dependence between A and B.

The Bridge: DAG

- Use the Directed Acyclic Graph (DAG) to represent the cause-effect relations
  - Nodes as variables
  - Edges as direct causal connections



- If a DAG represents the true causal relationship, then the DAG encodes all the conditional independence relations in the true distribution which can be read-off from the graph using d-separation.

Make Use of DAG for Causal Discovery and Causal Inference

DAG links causality with data

↓

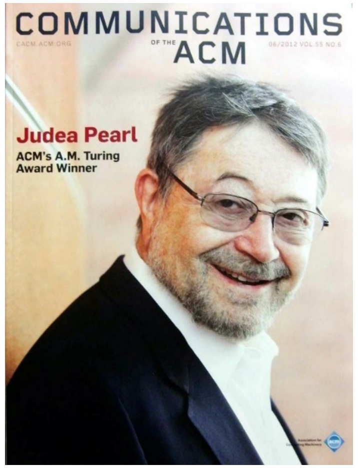
'Structural causal model', the more symbolic extension to the DAG

↙ ↘

Causal Discovery — Build DAG (causal graph) from observational data → Infer causal effects from given DAG (causal graph) — Causal Inference

Structural Equation/Causal Model

Structural Causal Model



- A causal model is triple $\mathcal{M} =< \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F} >$, where
  - $\boldsymbol{U}$ is a set of exogenous (hidden) variables whose values are determined by factors outside the model;
  - $\boldsymbol{V} = \{X_1, \cdots, X_i, \cdots\}$ is a set of endogenous (observed) variables whose values are determined by factors within the model;
  - $\boldsymbol{F} = \{f_1, \cdots, f_i, \cdots\}$ is a set of deterministic functions where each f_i is a mapping from $\boldsymbol{U} \times (\boldsymbol{V} \setminus X_i)$ to X_i. Symbolically, f_i can be written as

$$x_i = f_i(\boldsymbol{pa}_i, \boldsymbol{u}_i)$$

where $\boldsymbol{pa}_i$ is a realization of X_i's parents in $\boldsymbol{V}$, i.e., $\boldsymbol{Pa}_i \subseteq \boldsymbol{V}$, and $\boldsymbol{u}_i$ is a realization of X_i's parents in $\boldsymbol{U}$, i.e., $\boldsymbol{U}_i \subseteq \boldsymbol{U}$.

Causal Graph

- Each causal model $\mathcal{M}$ is associated with a direct graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where
  - $\mathcal{V}$ is the set of nodes represent the variables $\boldsymbol{U} \cup \boldsymbol{V}$ in $\mathcal{M}$;
  - $\mathcal{E}$ is the set of edges determined by the structural equations in $\mathcal{M}$: for X_i, there is an edge pointing from each of its parents $\boldsymbol{Pa}_i \cup \boldsymbol{U}_i$ to it.
    - Each direct edge represents the potential direct causal relation.
    - Absence of direct edge represents zero direct causal relation.
- Assuming the acyclicity of causality, $\mathcal{G}$ is a directed acyclic graph (DAG).
- Standard terminology
  - parent, child, ancestor, descendent, path, direct path

A Causal Model and Its Graph

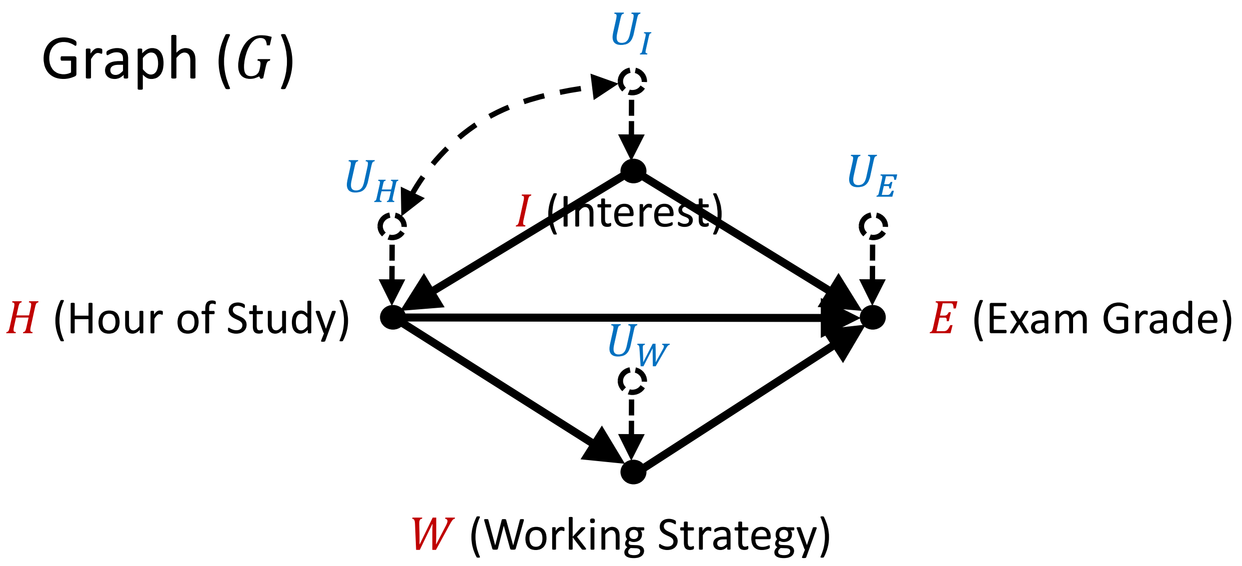
Observed Variables $\boldsymbol{V} = \{I, H, W, E\}$ Hidden Variables $\boldsymbol{U} = \{U_I, U_H, U_W, U_E\}$

Model (M)

$$i = f_I(u_I)$$
$$h = f_H(i, u_H)$$
$$w = f_W(h, u_W)$$
$$e = f_E(i, h, w, u_E)$$

Assume U_I and U_H are correlated.

Graph (G)



A Markovian Model and Its Graph

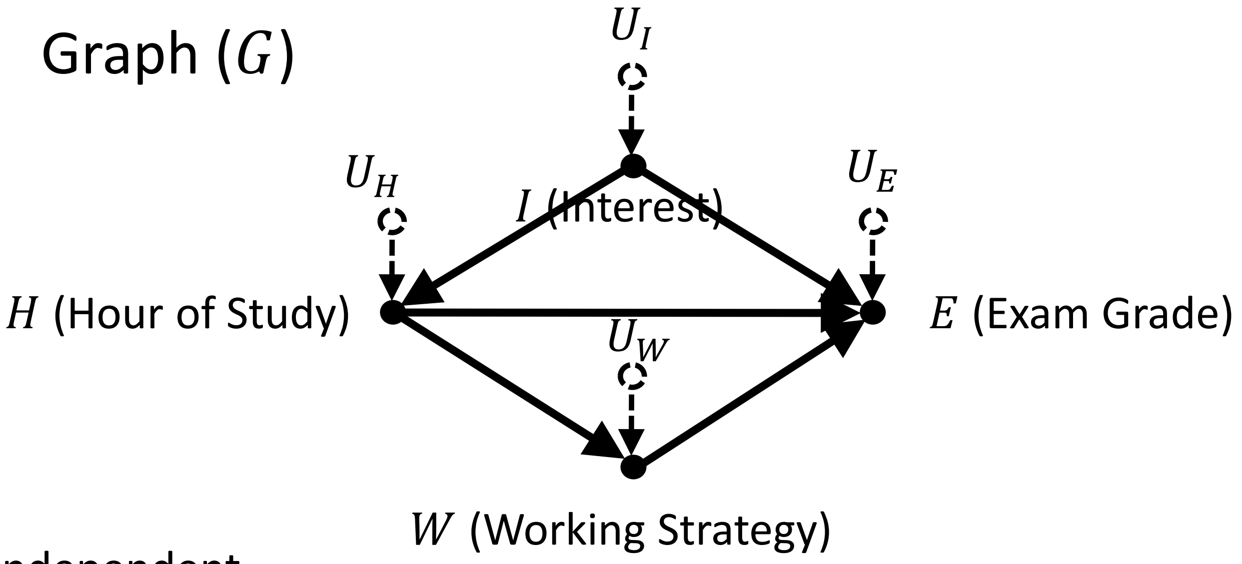
With causal sufficiency assumption

Model (M)

$$i = f_I(u_I)$$
$$h = f_H(i, u_H)$$
$$w = f_W(h, u_W)$$
$$e = f_E(i, h, w, u_E)$$

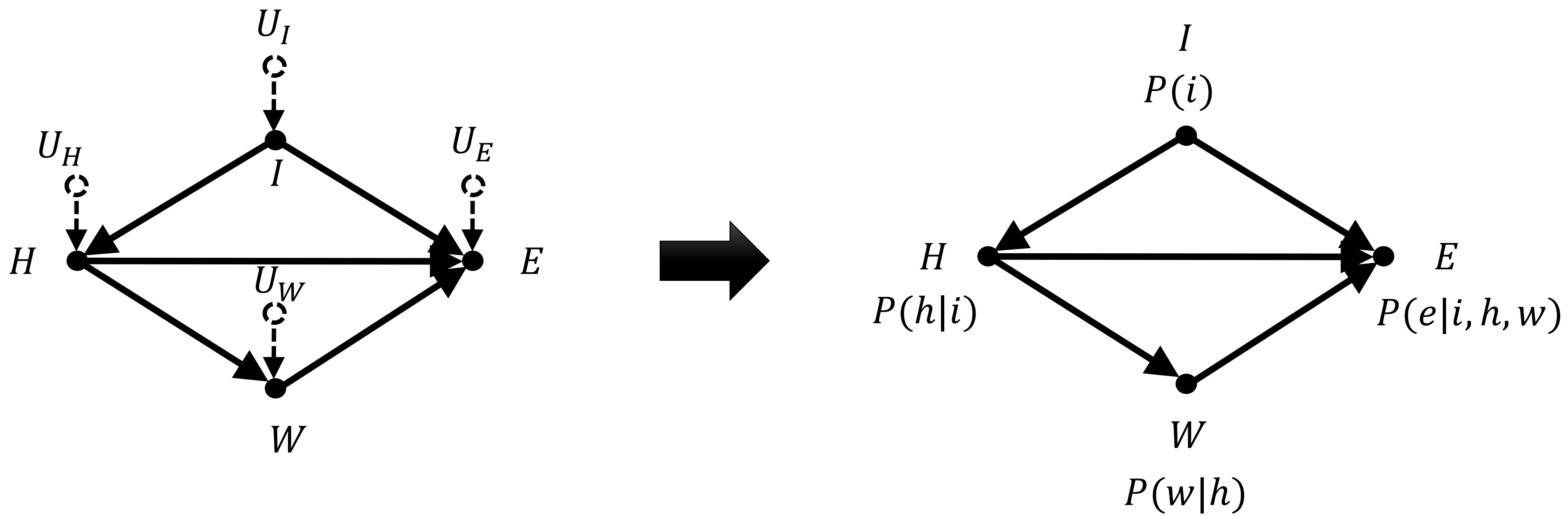
Assume U_I, U_H, U_W, U_E are mutually independent.

Graph (G)



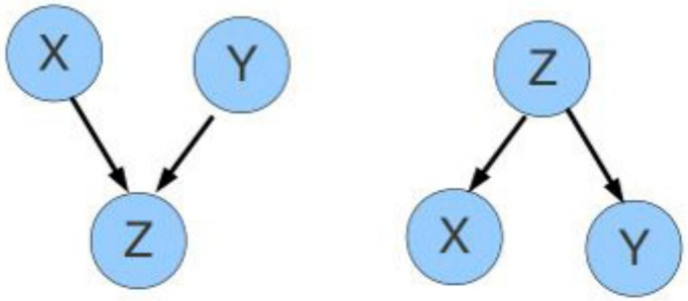
Causal Graph of Markovian Model

Each node is associated with an observable conditional probability table (CPT) $P(x_i|\boldsymbol{pa}_i)$



Causal Discovery

d-Separation



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

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

- Definition of d-separation

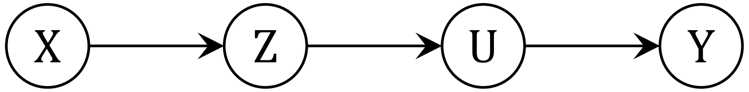
> - A path q is said to be blocked by conditioning on a set $\boldsymbol{Z}$ if
>   - q contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node m is in $\boldsymbol{Z}$, or
>   - q contains a collider $i \rightarrow m \leftarrow j$ such that the middle node m is not in $\boldsymbol{Z}$ and such that no descendant of m is in $\boldsymbol{Z}$.
> - $\boldsymbol{Z}$ is said to d-separate X and Y if $\boldsymbol{Z}$ blocks every path from X to Y, denoted by $(X \perp Y|Z)_G$

- If the DAG represents the true causal relationship

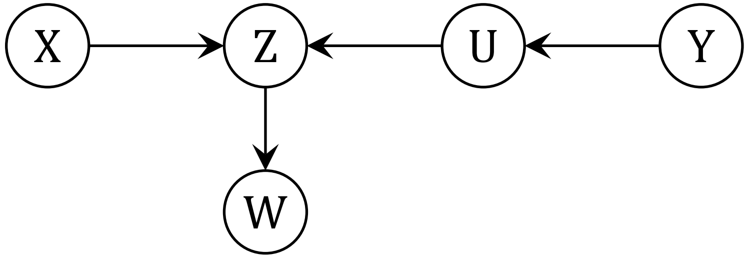
$$(X \perp Y|Z)_G \Leftrightarrow (X \perp Y|Z)_D$$

d-Separation

- Example (blocking of paths)



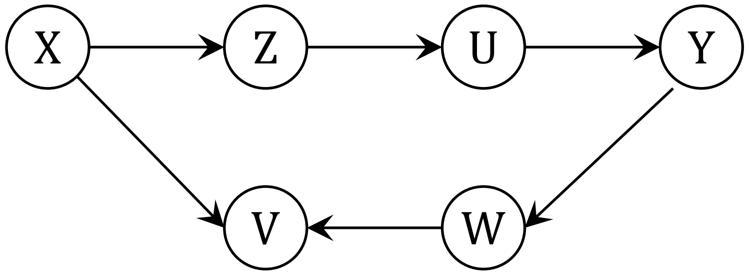
  - Path from X to Y is blocked by conditioning on $\{U\}$ or $\{Z\}$ or both $\{U, Z\}$
- Example (unblocking of paths)



  - Path from X to Y is blocked by $\emptyset$ or $\{U\}$
  - Unblocked by conditioning on $\{Z\}$ or $\{W\}$ or both $\{Z, W\}$

d-Separation

- Examples (*d*-separation)



- We have following *d*-separation relations
  - $(X \perp Y|Z)_G, (X \perp Y|U)_G, (X \perp Y|ZU)_G$
  - $(X \perp Y|ZW)_G, (X \perp Y|UW)_G, (X \perp Y|ZUW)_G$
  - $(X \perp Y|VZUW)_G$
- However we do NOT have
  - $(X \perp Y|VZU)_G$

PC Algorithm (Peter Spirtes & Clark Glymour)

- Faithfulness assumption
- Causal sufficiency (no hidden common cause) assumption
- The BEST we can do without further assumptions (or knowledge).
- Usually CANNOT identity the unique causal graph (up to the Markov equivalent class)

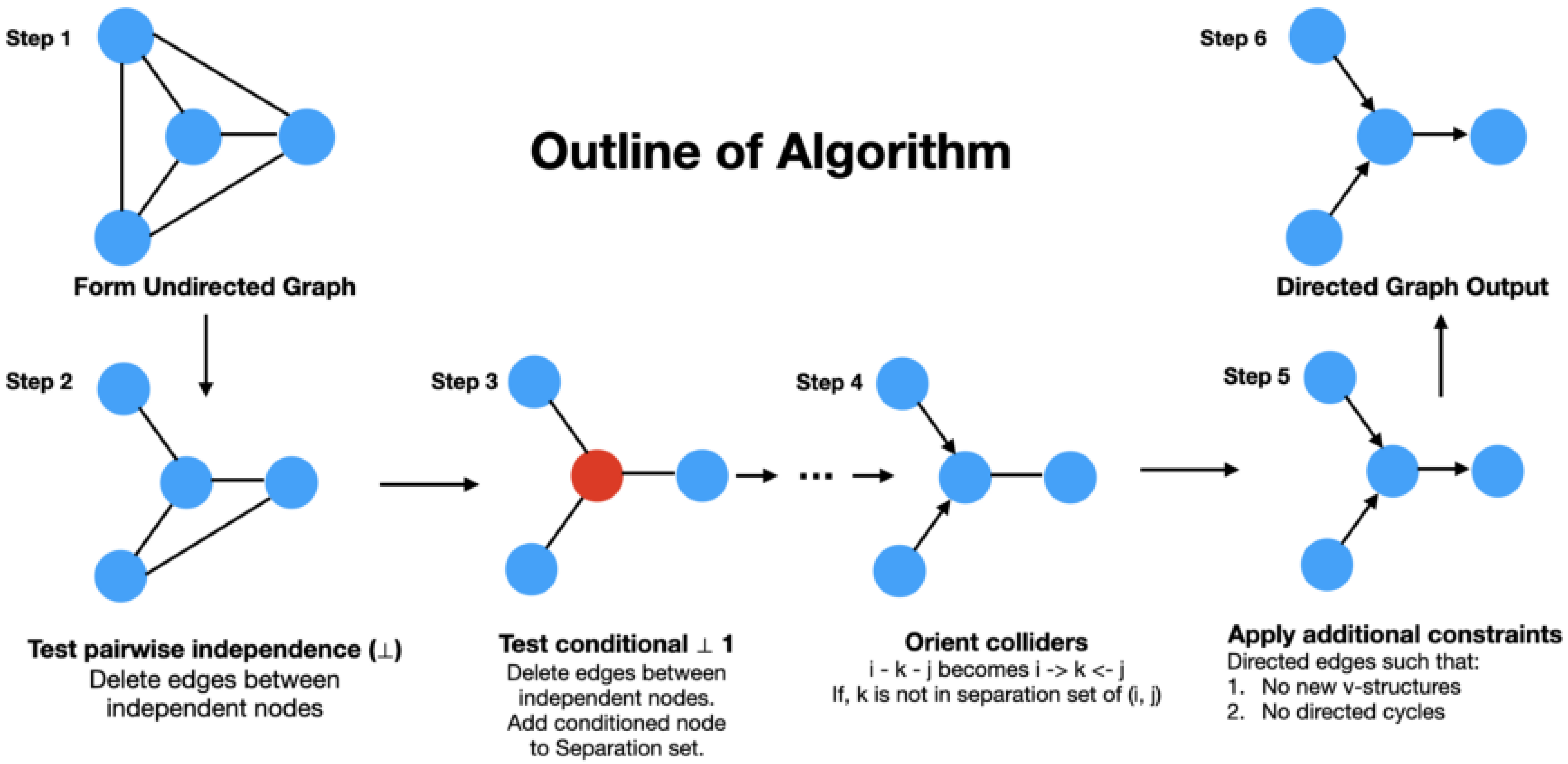
$$\left.\begin{array}{l} X \rightarrow Z \rightarrow Y \\ X \leftarrow Z \rightarrow Y \\ X \leftarrow Z \leftarrow Y \end{array}\right\}$$

$$X \rightarrow Z \leftarrow Y$$

PC Algorithm: The Sketch

1. Construct the skeleton
    1. Start with a fully connected undirected graph
    2. Remove all edges $X - Y$ with $X \perp Y$
    3. Remove all edges $X - Y$ for which there is a neighbor $Z \neq Y, X$ with $X \perp Y | Z$
    4. Remove all edges $X - Y$ for which there are two neighbors $Z_1, Z_2 \neq Y, X$ with $X \perp Y | Z_1, Z_2$
    5. …
2. Orient the arrows by finding v-structures $X \rightarrow Z \leftarrow Y$

Example of PC Algorithm



Open-source Software

- http://www.phil.cmu.edu/tetrad/
- Implement a large set of Constraint-Based and Score-Based causal discovery algorithms.

- https://github.com/py-why/causal-learn
- A python package for causal discovery that implements both classical and state-of-the-art causal discovery algorithms, which is a Python translation and extension of Tetrad.

We Can Do Better Than PC Algorithm

- Given X, Y, can we distinguish $X \rightarrow Y$ and $X \leftarrow Y$?

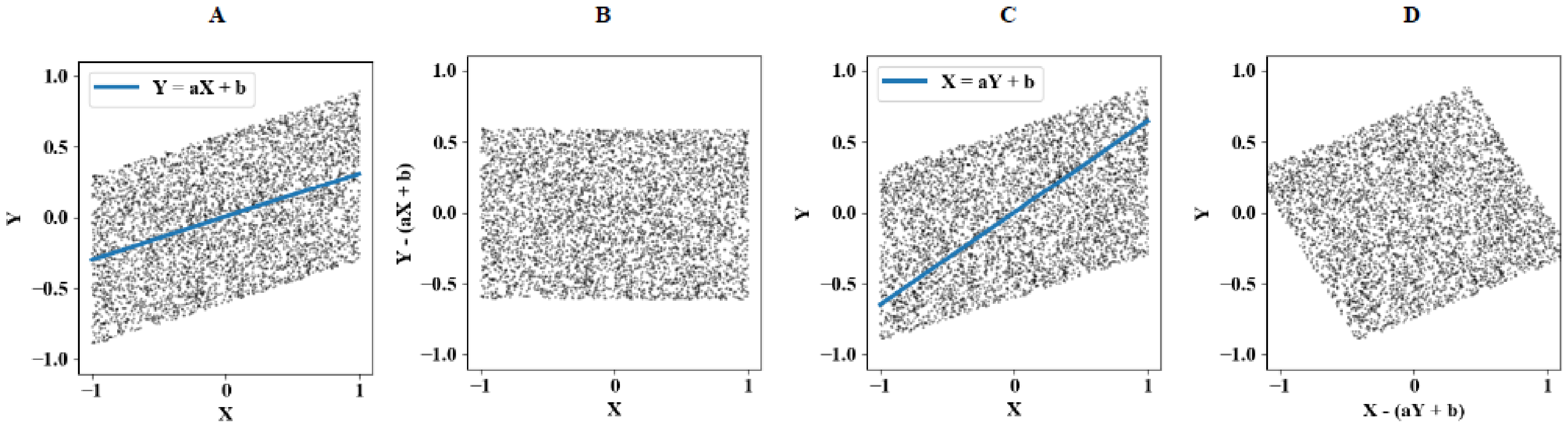
- If some additional assumptions are made about the functional and/or parametric forms of the underlying true data-generating structure, then one can exploit asymmetries in order to identify the direction of a structural relationship.

Additive Noise

- Given the linear structural equations

$X = U_X$ and $Y = X + U_Y$ such that $U_Y \perp U_X$

- If U_X or U_Y is non-Gaussian
- Then the causal direction $X \rightarrow Y$ is identifiable



Independent Causal Mechanisms

- The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.
- Suppose $X \rightarrow Y$, then $P(x)$ and $P(y|x)$ should be independent.
- In other words, semi-supervised learning, i.e., unsupervised learning on X should not improve supervised learning $X \mapsto Y$.
- Will be different if decompose the distribution to $P(y)$ and $P(x|y)$.

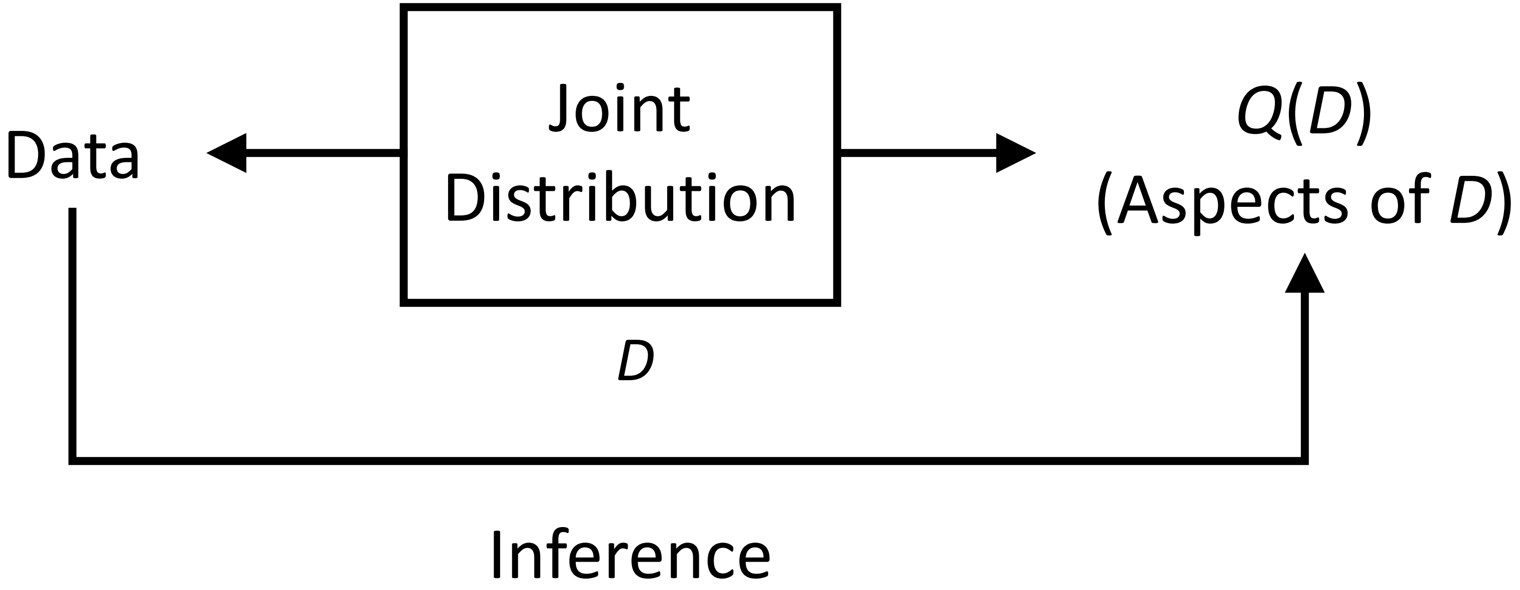
Causal Inference

The BIG Idea(s)

1. Every causal inference task must rely on judgmental, extra-data assumptions (or experiments).
2. We have ways of encoding those assumptions mathematically and test their implications.
3. We have a mathematical machinery to take those assumptions, combine them with data and derive answers to questions of interest.
4. We have a way of doing (2) and (3) in a language that permits us to judge the scientific plausibility of our assumptions and to derive their ramifications swiftly and transparently.
5. Items (2)-(4) make causal inference manageable, fun, and profitable.

From Statistics to Causal Modeling

- Traditional statistical inference paradigm:



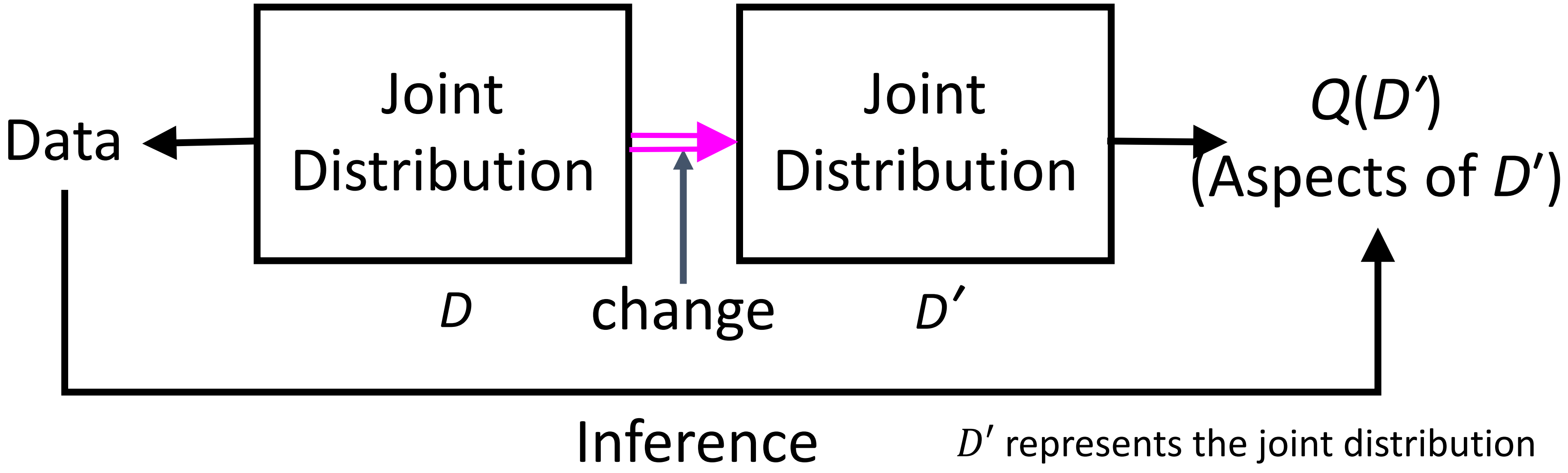
- What is the chance of getting Grade A for the students who study 1 hour each day?

Estimate $Q(D) = P_D(E = \text{‘}A\text{’} \mid H = 1)$

E (Exam Grade)
H (Hour of Study)
I (Interest)
W (Working Strategy)

From Statistics to Causal Modeling

- What is the chance of getting Grade A if a new policy requires all students to study 2 hours each day?
  - The question cannot be solved by statistics.

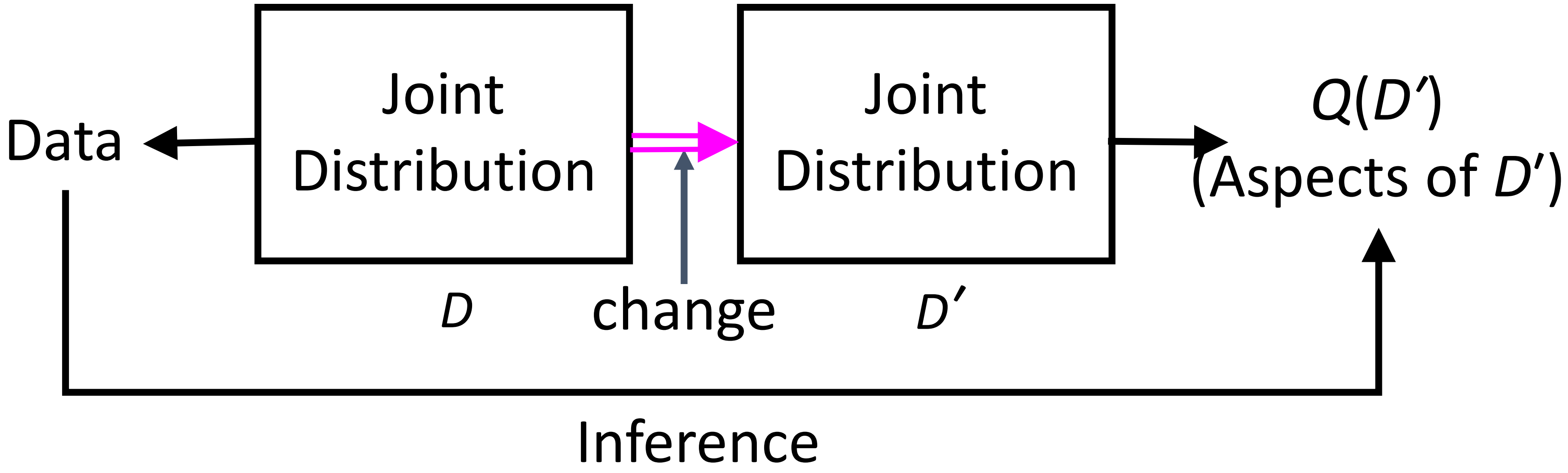


Estimate $Q(D') = P_{D'}(E = \text{‘}A\text{’})$

D' represents the joint distribution after adopting the new policy.

From Statistics to Causal Modeling

- What is the chance of getting Grade A if a new policy requires all students to study 2 hours each day?
  - The question cannot be solved by statistics.

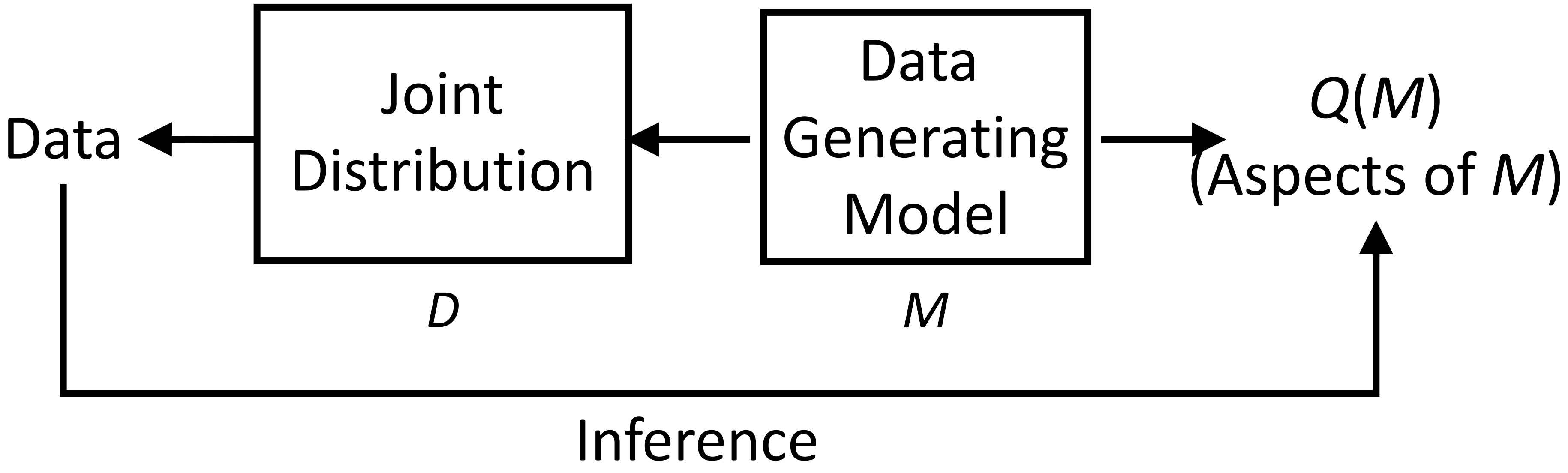


$$P_{D'}(\,E = \text{‘}A\text{’}\,) \;\neq\; P_D\,(\,E = \text{‘}A\text{’}\,|\,H = 2\,)$$

The probability of getting Grade A of the students who study 2 hours each day at the first place.

From Statistics to Causal Modeling

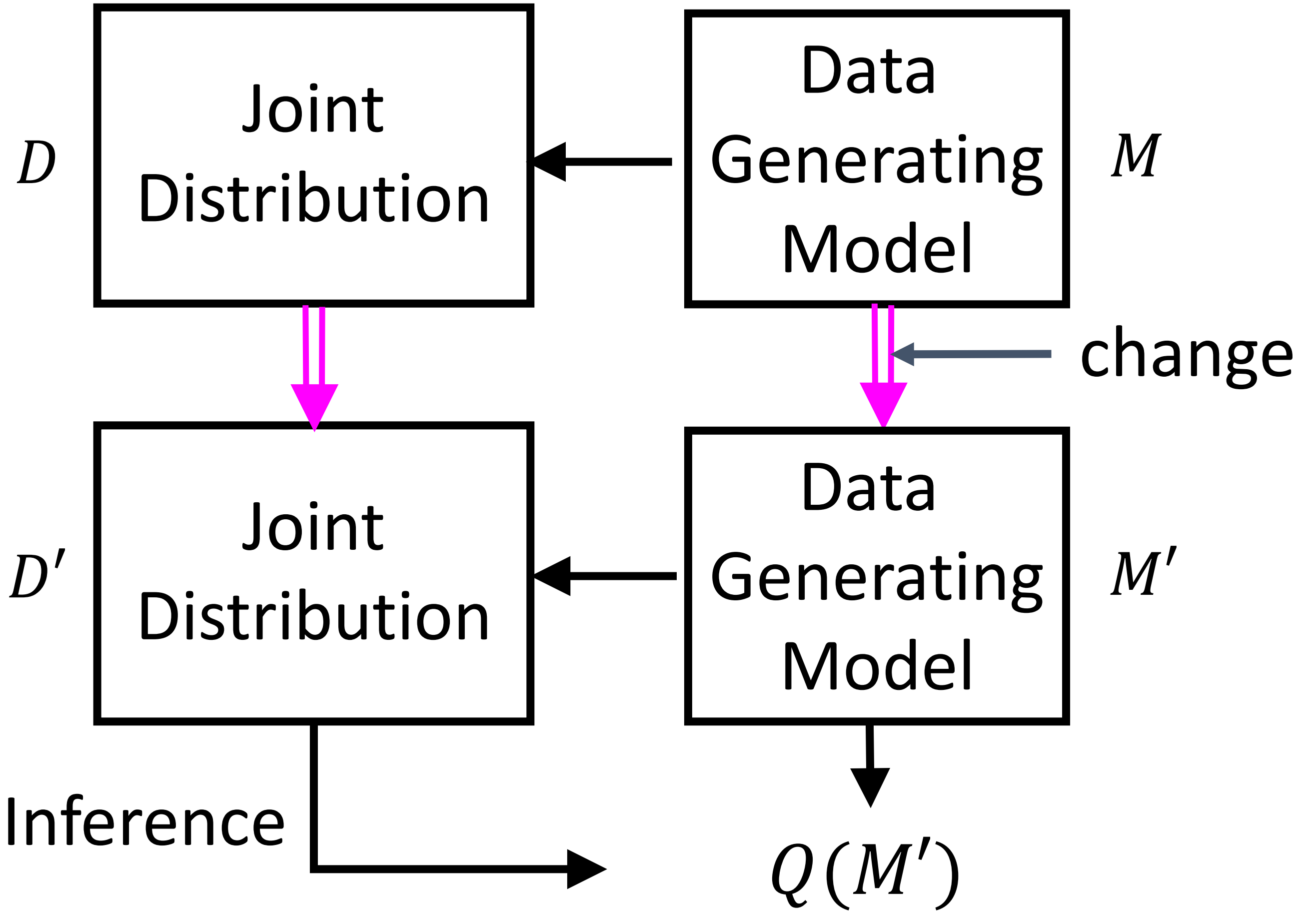
- Causal inference



M – Data generation model that encodes the causal assumptions/knowledge.
D – model of data, M – model of reality

From Statistics to Causal Modeling

- Causal inference



WHAT KIND OF QUESTIONS SHOULD THE CAUSAL MODEL ANSWER

THE CAUSAL HIERARCHY

- Observational Questions:
- "What if we see A"

(What is?) $P(y \mid A)$

- Action Questions:
- "What if we do A?"

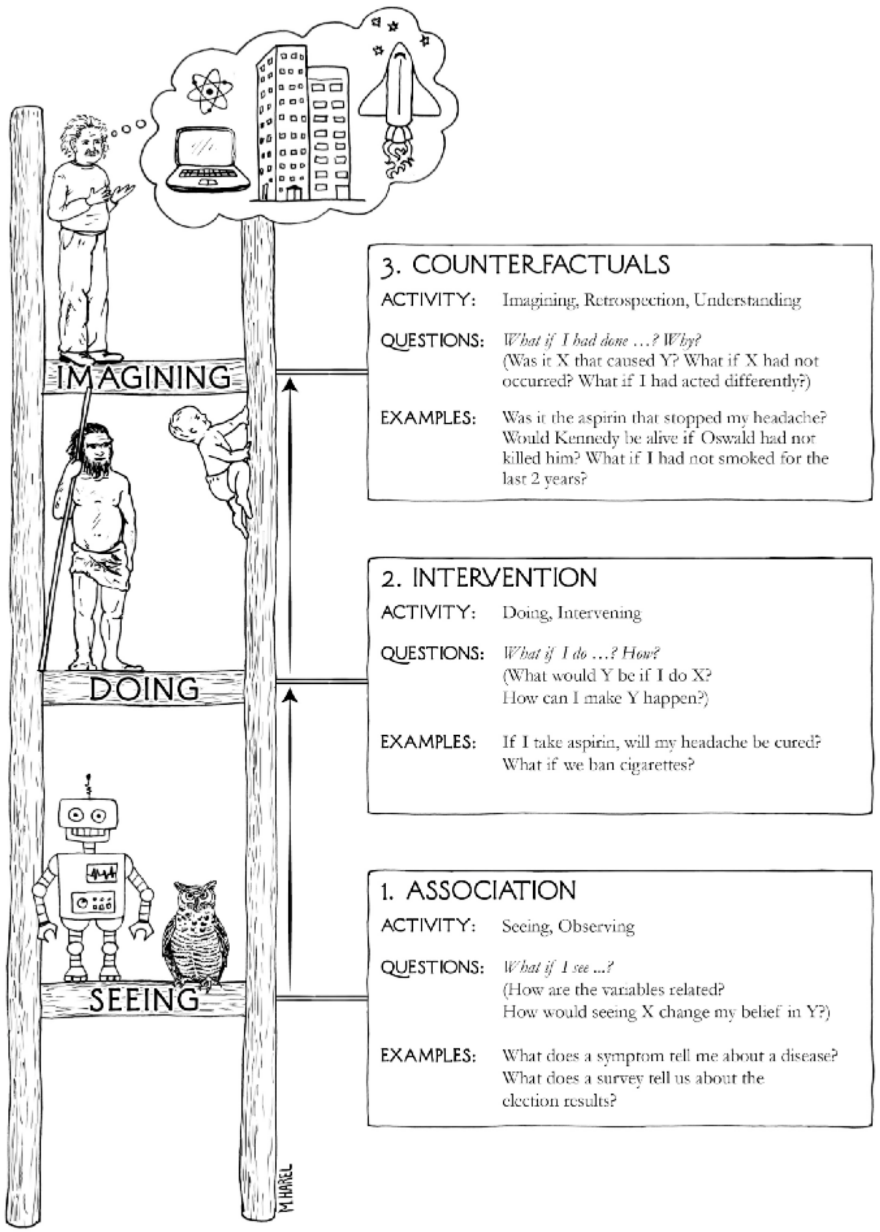
(What if?) $P(y \mid do(A))$

- Counterfactuals Questions:
- "What if we did things differently?"

(Why?) $P(y_{A'} \mid e)$

- Options:
- "With what probability?"

Ladder of Causality



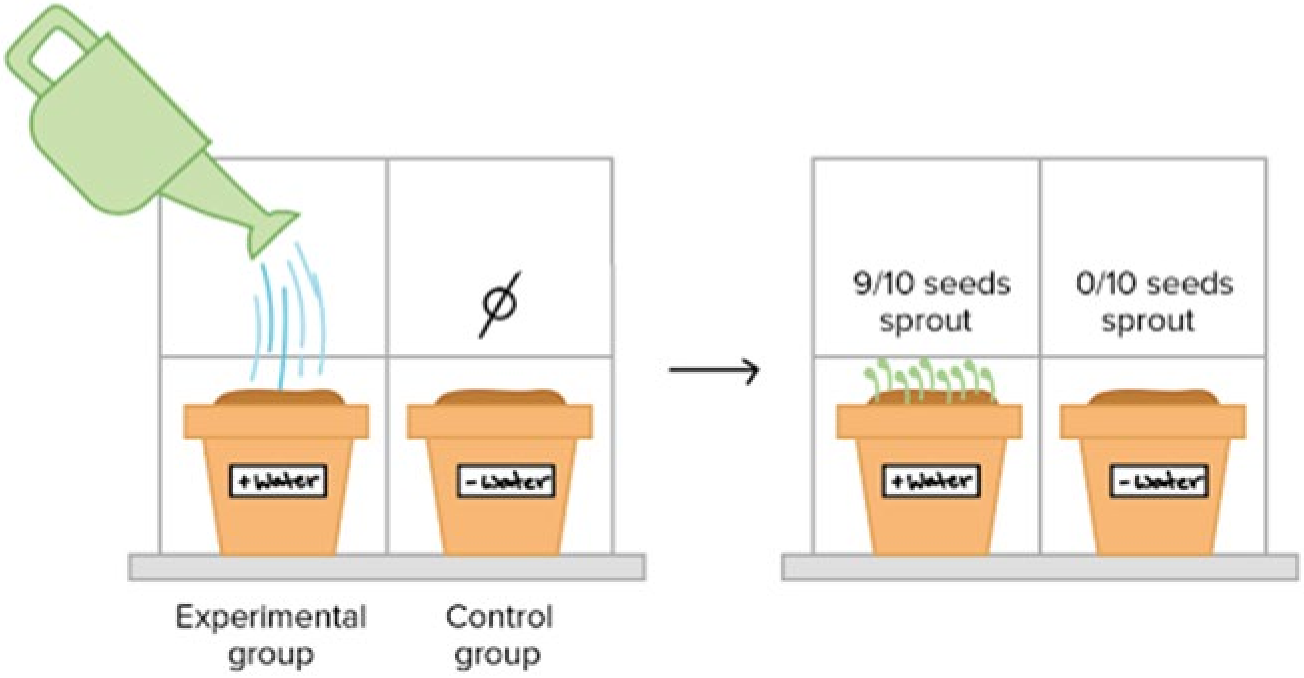
Judea, Pearl, and Mackenzie Dana. "The Book of Why: The New Science of Cause and Effect." *Basic Books*. 2018.

Causal Inference

- Question: What is the chance of getting grade A if we change the study hour to 2?
  - The above probability does not equal to $P(E = \text{'}A\text{'} | H = 2)$, i.e., the conditional probability of getting grade A given study hour equals to 2.

Intervention

- Physical intervention



Intervention and *do*-Operation

- The basic operation of manipulating a causal model.
  - Simulate the physical intervention.
  - Forces some observed variables $\boldsymbol{X} \in \boldsymbol{V}$ to take certain constants $\boldsymbol{x}$.

- Mathematically formulated as $do(\boldsymbol{X} = \boldsymbol{x})$ or simply $do(\boldsymbol{x})$.

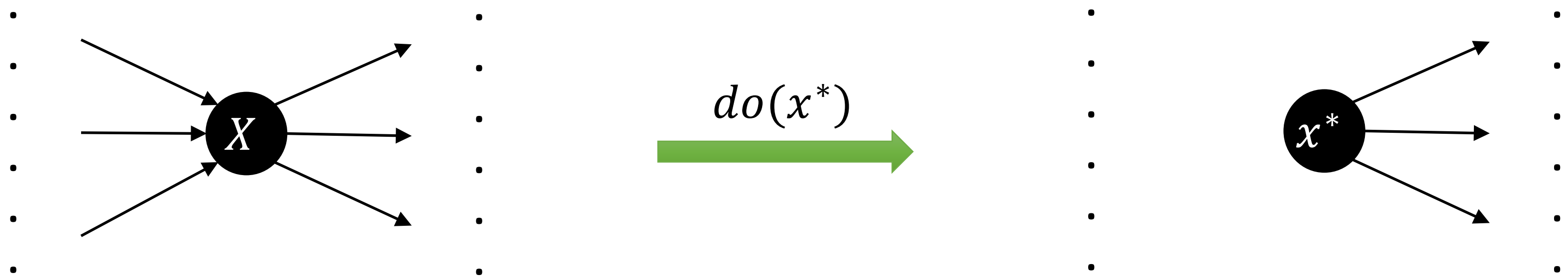
- The effect of intervention on all other observed variables $\boldsymbol{Y} = \boldsymbol{V} \backslash \boldsymbol{X}$ is represented by the post-intervention distribution of $\boldsymbol{Y}$.
  - Denoted by $P(\boldsymbol{Y} = \boldsymbol{y} | do(\boldsymbol{X} = \boldsymbol{x}))$ or simply $P(\boldsymbol{y} | do(\boldsymbol{x}))$;

Intervention and *do*-Operation

- In causal model $\mathcal{M}$, intervention $do(x^*)$ is defined as the substitution of structural equation $x = f_X(\boldsymbol{pa}_X, \boldsymbol{u}_X)$ with value x^*. The causal model after performing $do(x^*)$ is denoted by $\mathcal{M}_{x^*}$.

$$\mathcal{M}: \quad x = f_X(\boldsymbol{pa}_X, \boldsymbol{u}_X) \quad \xrightarrow{do(x^*)} \quad \mathcal{M}_{x^*}: \quad x = x^*$$

- From the point of view of the causal graph, performing $do(x^*)$ is equivalent to setting the node X to value x^* and removing all the incoming edges in X.



Intervention in Markovian Model

- In the Markovian model, the post-intervention distribution $P(\boldsymbol{y}|do(\boldsymbol{x}))$ can be calculated from the CPTs, known as the truncated factorization:

$$P(\boldsymbol{y}|do(\boldsymbol{x})) = \prod_{Y \in \boldsymbol{Y}} P(y|\boldsymbol{Pa}_Y)\delta_{\boldsymbol{X} \leftarrow \boldsymbol{x}}$$

  - where $\delta_{\boldsymbol{X} \leftarrow \boldsymbol{x}}$ means assigning attributes in $\boldsymbol{X}$ involved in the term ahead with the corresponding values in $\boldsymbol{x}$.

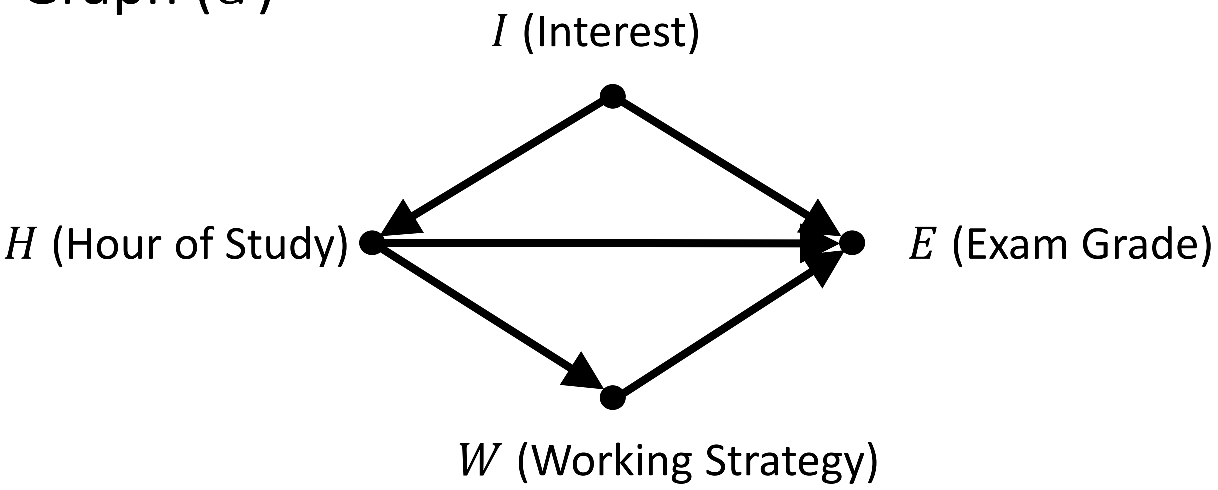
- Specifically, for a single attribute Y given an intervention on a single attribute X,

$$P(y|do(x)) = \sum_{\substack{\boldsymbol{V} \backslash \{X,Y\} \\ Y=y}} \prod_{V \in \boldsymbol{V} \backslash \{X\}} P(v|\boldsymbol{Pa}_V)\delta_{X \leftarrow x}$$

Intervention Example

- What is the probability of getting grade A if we change the study hour to 2?

Graph (G)



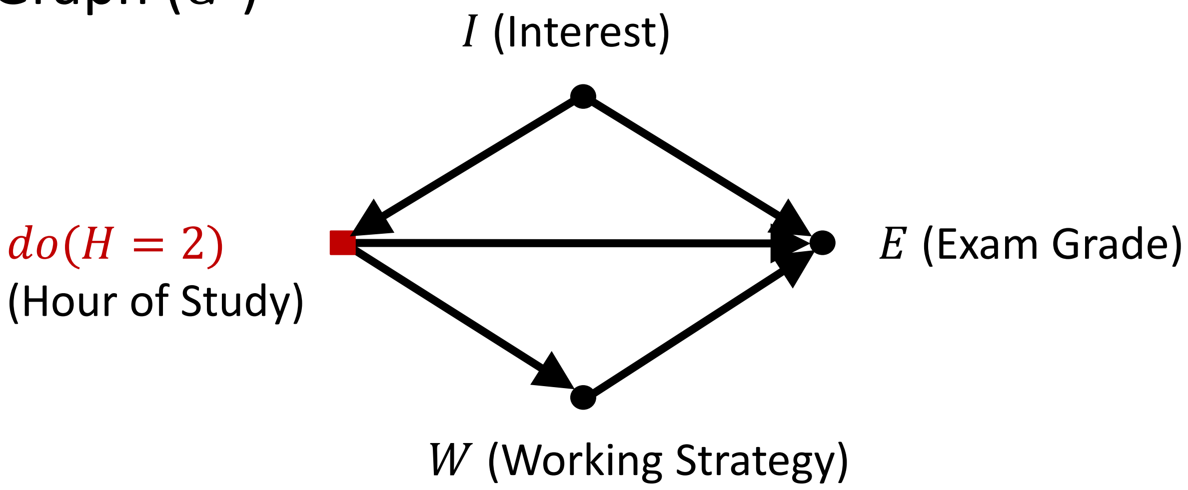
Model (M)

$$i = f_I(u_I)$$
$$h = f_H(i, u_H)$$
$$w = f_W(h, u_W)$$
$$e = f_E(i, h, w, u_E)$$

Intervention Example

- What is the probability of getting grade A if we change the study hour to 2, i.e., $do(H=2)$?

Graph (G')



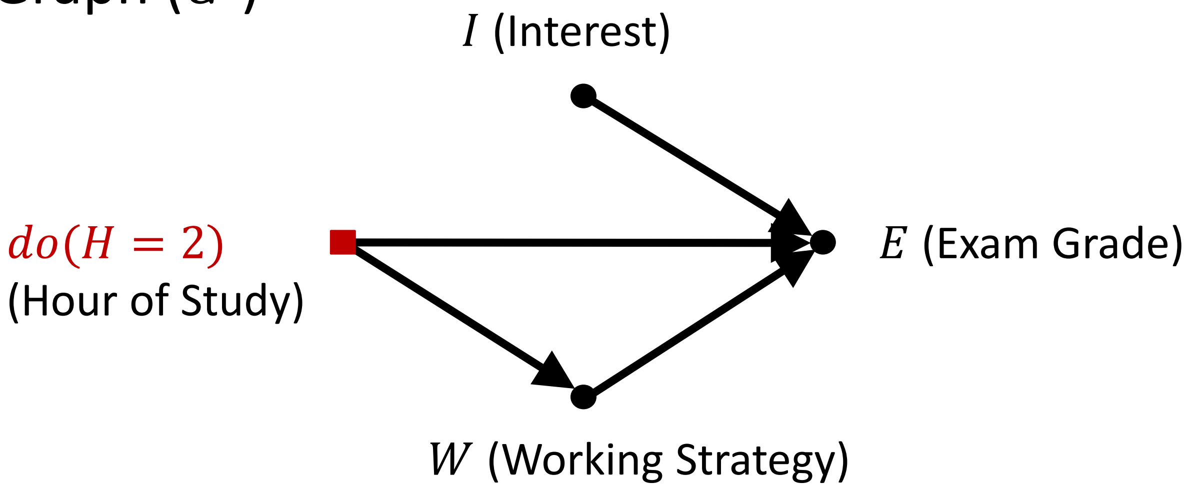
Model (M')

$$i = f_I(u_I)$$
$$h = 2$$
$$w = f_W(h, u_W)$$
$$e = f_E(i, h, w, u_E)$$

- Find $P(E = \text{'}A\text{'} \mid do(H=2))$

Intervention Example

Graph (G')



Model (M')

$i = f_I(u_I)$
$h = 2$
$w = f_W(h, u_W)$
$e = f_E(i, h, k, u_E)$

$$P(y|do(x)) = \sum_{\substack{\boldsymbol{V}\setminus\{X,Y\} \\ Y=y}} \prod_{V \in \boldsymbol{V}\setminus\{X\}} P(v|\boldsymbol{Pa}_V)\delta_{X \leftarrow x}$$

$$P(E = \text{‘}A\text{’}|do(H = 2)) = \sum_{I,W} P(i)P(w|H = 2)P(E = \text{‘}A\text{’}|i, H = 2, w)$$

Applications of CI in ML

- Fair machine learning
- Reinforcement learning
- Transfer learning and multi-task learning
- Robust machine learning

Multi-task Learning

- Predict a target Y from some features X.
- Consider D training tasks where each task k has a different distribution $\mathbb{P}^k$ for generating data, i.e., $(X^k, Y^k) \sim \mathbb{P}^k$, $k \in \{1, \ldots, D\}$.
- To improve performance in some tasks (aka., test tasks).

Invariant Models based on Causal Methodology

- Assume there exists an invariant subset S^*, i.e.,

$$Y^k | X^k_{S^*} = Y^{k'} | X^{k'}_{S^*} \quad \forall k, k' \in \{1, \ldots, D\}$$

- Missing data approach to combine invariance and task-specific information.
  - Assume that features other than X_{S^*} are missing.
  - Let $Z_i = (X_{S^*,i}, X_{N,i}, Y)$ be a pooled sample of the available data from all the tasks in which $X_{N,i}$ is considered missing if i is drawn from a training task.
  - EM algorithm is used to maximize log-likelihood

$$\ell(\Sigma) = \text{const} - \frac{1}{2} \sum_{i=1}^{n} \det(\Sigma_i) - \frac{1}{2} \mathbf{Z}^T_{obs,i} \Sigma_i^{-1} \mathbf{Z}_{obs,i},$$

Rojas-Carulla, M., Schölkopf, B., Turner, R., & Peters, J. (2018). Invariant models for causal transfer learning. The Journal of Machine Learning Research, 19(1), 1309-1342.

Relation to Causality

- Suppose that there is an SCM over variables (X, Y).
- Suppose that the different tasks $\mathbb{P}^1, \dots, \mathbb{P}^D$ are post-interventional distributions of an underlying SCM with graph structure G.
- Suppose that the target variable Y has not been intervened on.

- Then: The set $S^* \coloneqq \boldsymbol{Pa}_Y$ is an invariant set.

Rojas-Carulla, M., Schölkopf, B., Turner, R., & Peters, J. (2018). Invariant models for causal transfer learning. The Journal of Machine Learning Research, 19(1), 1309-1342.

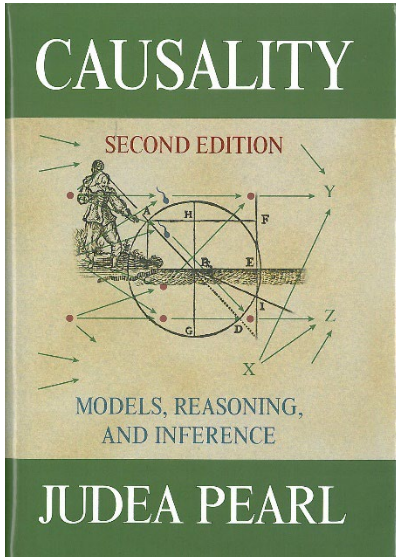
Open-source packages for causal inference

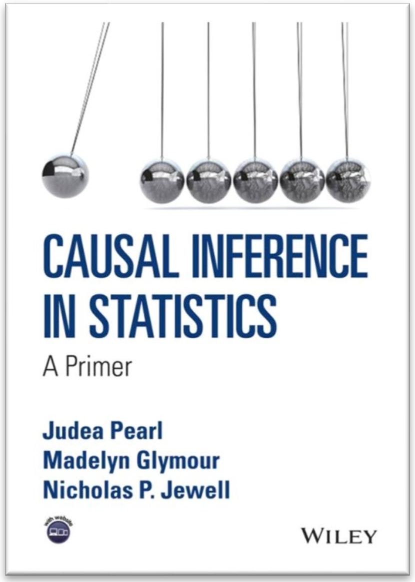
- Microsoft/DoWhy:
- https://github.com/microsoft/dowhy

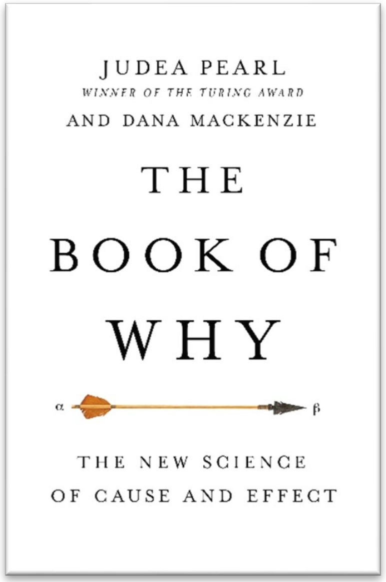
- IBM/causallib
- https://github.com/IBM/causallib

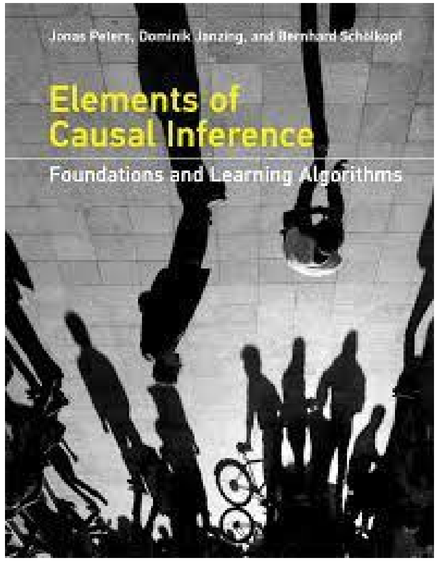
Useful Resources

- Four books









- Website: http://bayes.cs.ucla.edu/