A brief introduction to causal inference

Jacqueline Maasch
maasch@cs.cornell.edu

27 March 2025

Presentation overview

1. **Causal inference:** A nonexhaustive overview of the basics

2. **Causal discovery:** Learning structure from data

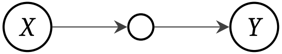
Causal inference:
The basics

Cause-effect relationships

1 Causal inference: The basics

Causation is an influence by which an event or state (**cause** X) contributes to the production of another event or state (**effect** Y).

direct cause



indirect cause

Example: I took the antibiotic, so my infection cleared faster than if I had done nothing.

Definition adapted from Wikipedia. For a formal definition, see Halpern [1].

Pearl's Causal Hierarchy

(1) Associational – (2) Interventional – (3) Counterfactual

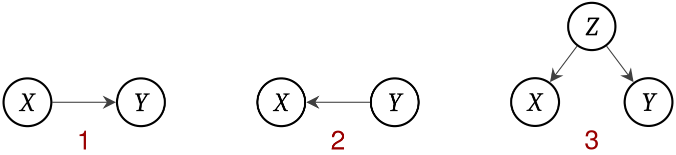
Layer	Activity	Semantics	Example
(1) Associational $p(y \mid x)$	Seeing 👀	How would seeing x change my belief in Y?	What does a symptom tell us about the disease?
(2) Interventional $p(y \mid \mathrm{do}(x), z)$	Doing 💪	What happens to Y if I do x?	What if I take aspirin, will my headache be cured?
(3) Counterfactual $p(y_{x'} \mid x, y)$	Imagining 🤔	Was it x that caused Y?	Was it the aspirin that stopped my headache?

Distributions on one layer underdetermine the higher layers. Figure from [2].

Definition: Reichenbach's Common Cause Principle [3]

A statistical association between two variables X, Y can be explained by any of the following:

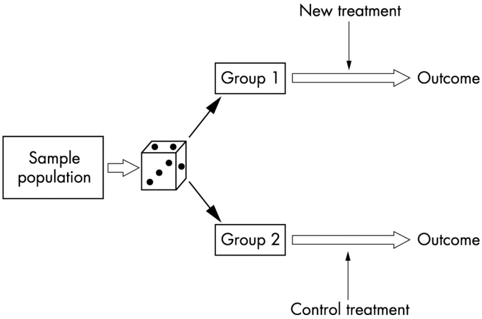
1. X is a cause of Y.
2. Y is a cause of X.
3. X and Y are both caused by a third variable, **confounder** Z.

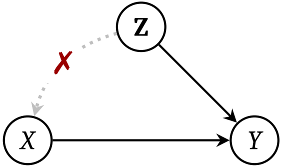


RCTs: Association $==$ causation

1 Causal inference: The basics

Randomized controlled trials (RCTs) are the gold standard (e.g., clinical trials).





Causal graph: $\mathbf{Z} \not\to X$

Left-hand figure from [4].

RCTs: Association == causation

1 Causal inference: The basics

Efficacy and safety of hydroxychloroquine vs placebo for pre-exposure SARS-CoV-2 prophylaxis among health care workers: a randomized clinical trial

BS Abella, EL Jolkovsky, BT Biney… - **JAMA** internal …, 2021 - jamanetwork.com

… This single-health system, double-blind placebo-**controlled randomized trial** was conducted as the prophylaxis substudy of the Prevention and Treatment of COVID-19 With …

☆ Save ❞ Cite Cited by 238 Related articles All 14 versions

Effect of continuous glucose monitoring on glycemic control in patients with type 2 diabetes treated with basal insulin: a randomized clinical trial

T Martens, RW Beck, R Bailey, KJ Ruedy, P Calhoun… - **Jama**, 2021 - jamanetwork.com

… In this **randomized trial** of patients with type 2 diabetes and poor glycemic **control** (mean HbA … To our knowledge, there has not been a prior **randomized trial** that has evaluated CGM in …

☆ Save ❞ Cite Cited by 447 Related articles All 11 versions

Effects of fecal microbiome transfer in adolescents with obesity: the gut bugs randomized controlled trial

KSW Leong, TN Jayasinghe, BC Wilson… - **JAMA** network …, 2020 - jamanetwork.com

… In this **randomized** clinical **trial** of adolescents with obesite, there was no effect of FMT on weight loss in adolescents with obesity, although a reduction in abdominal adiposity was …

☆ Save ❞ Cite Cited by 135 Related articles All 18 versions

RCTs: Association $==$ causation

Why not *always* use RCTs?

- Often prohibitively expensive, challenging, or unethical to conduct.
- Attrition (i.e., loss of study units) can be high.
- Still imperfect: results can still fail to generalize across populations.

The logic and tools of answering causal questions. — Judea Pearl

A scientific framework for inferring the presence and magnitude of cause-effect relationships using observational data...

The process of inferring interventional distributions from observational data...

Etc.

1. **Identify the causal quantity of interest.**
   - *Example:* Average treatment effect (ATE) of a drug on a disease state.
   - We have a sample from our true data distribution.
   - A graphical model of the data generating process (DGP) enables identifiability.
   - We can learn this model with data-driven methods.

2. **Perform inference to estimate this quantity.**
   - Express the parameter as a function of the DGP.
   - Apply estimation methods (e.g., TMLE, doubly robust ML, etc.).

A missing data problem

Fundamental problem of causal inference:

By definition, counterfactuals cannot be observed.

If you cannot observe the effect of more than one treatment on a given subject over a given time frame, **how is causal inference possible**?

Logically equivalent frameworks

There are two main frameworks for causal inference:

1. **Potential outcomes** (PO).
2. **Structural causal models** (SCMs).

> *Formally, the two frameworks are **logically equivalent**; a theorem in one is a theorem in the other, and every assumption in one can be translated into an equivalent assumption in the other. — Judea Pearl*

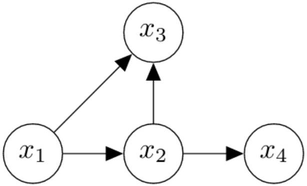
Definition: Structural causal model (SCM) [5]

An SCM is a tuple $\mathcal{M} := \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, p(\mathbf{u}) \rangle$:

- $\mathbf{U} = \{U_i\}_{i=1}^n$ are exogenous variables determined by factors outside $\mathcal{M}$;
- $\mathbf{V} = \{V_i\}_{i=1}^n$ are observed endogenous variables determined by variables in $\mathbf{U} \cup \mathbf{V}$;
- $\mathcal{F} = \{f_i\}_{i=1}^n$ are structural functions such that $v_i = f_i(\mathbf{pa}_{v_i}, u_i)$;
- $p(\mathbf{u})$ is the distribution over $\mathbf{U}$.

SCMs are associated with graphical representations, often **directed acyclic graphs** (DAGs).

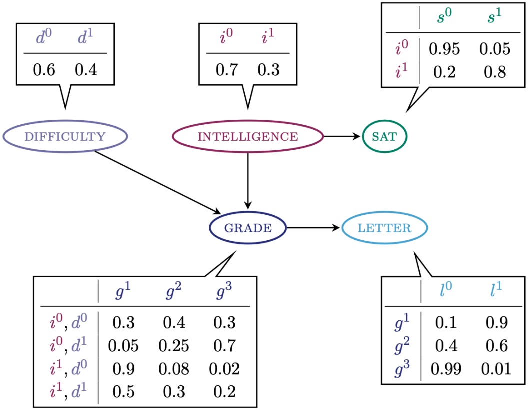
Structural causal models



$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1.3 & 0 & 0 & 0 \\ 0.3 & 2.5 & 0 & 0 \\ 0 & 5.7 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

A DAG and its associated linear-Gaussian structural equations, where noise $\epsilon \sim \mathcal{N}(0, \Sigma)$.

Structural causal models



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

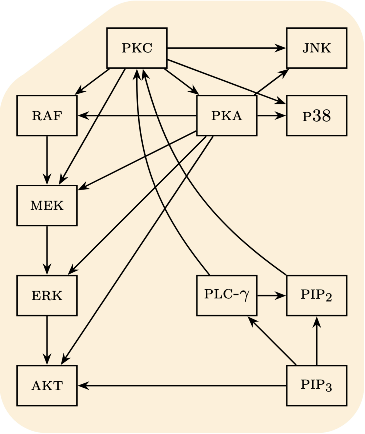
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Figure adapted from [6].

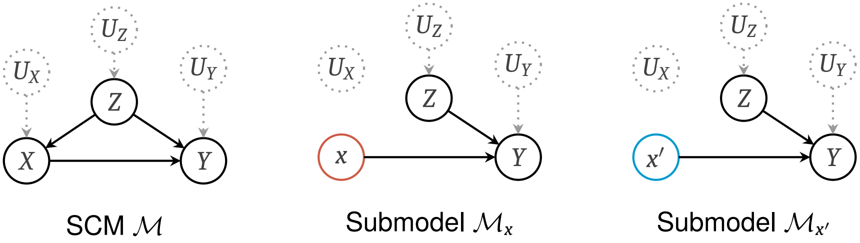
Real causal graphs are usually complicated



The Sachs protein-signaling network, a famous causal DAG [7].

Counterfactual submodels and *do* notation

1 Causal inference: The basics



Interventions $do(x)$ and $do(x')$ replace the true causal mechanism that generates X with a constant function evaluating to x or x'.

Figure adapted from [8]

The *do* notation and potential outcomes

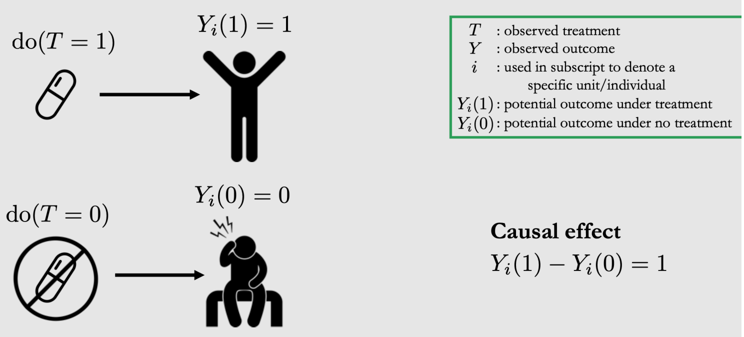


Figure from Brady Neal's *Introduction to Causal Inference*.

From observational to interventional probabilities

- Sampling from $\mathcal{M}$ provides evidence for **observational probabilities** of events, e.g.:

$$p(x, y).$$

- To reason about the effect of interventions, we want **interventional probabilities**e.g.:

$$p(y \mid do(x)) \quad \text{(also denoted } p(y_x) \text{) (or } p_x(y)\text{)}.$$

- **Problem:** Observational data *alone* cannot be used to uniquely determine $p(y \mid do(x))$.

Slide adapted from Alexis Bellot.

Assumptions are necessary

Causal inference without assumptions is an ill-posed problem.

> *"...all causal inference is based on assumptions that cannot be derived from observations alone." – Greenland, Pearl & Robins (1999)*

Definition: The Law of Decreasing Credibility [9]

The credibility of inference decreases with the strength of the assumptions maintained.

Assumptions are necessary

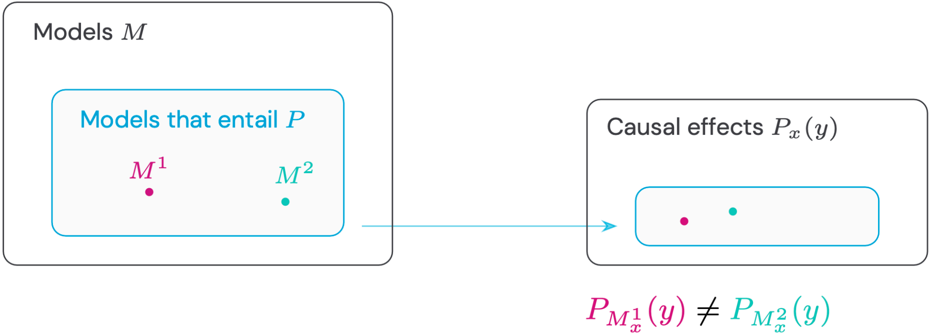
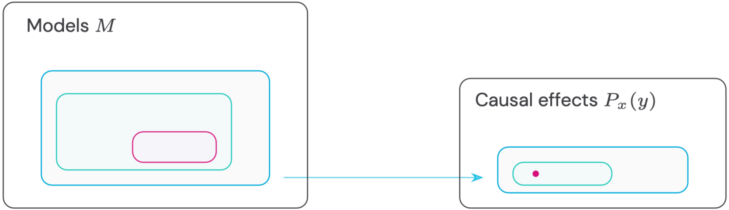


Figure from Alexis Bellot.

Assumptions are necessary



Models that entail P

Models that entail P and induce causal diagram $\mathcal{G}$

Models that entail P and induce causal diagram $\mathcal{G}$ and are linear with non-Gaussian error terms

Figure from Alexis Bellot.

Linearity: A common modeling assumption

Example: Decomposition of total causal effects in linear SCMs [5]

Let TE be the total effect, NDE the natural direct effect, and NIE the natural indirect effect. When causal functions are linear,

$$\underbrace{\mathrm{TE}}_{\textit{global}} = \underbrace{\mathrm{NDE}}_{\textit{local}} + \underbrace{\mathrm{NIE}}_{\textit{local}}.$$

$$\begin{tikzcd} & Z \arrow[dr, "c"] & \\ X \arrow[ur, "b"] \arrow[rr, "a"] & & Y \end{tikzcd}$$

$$\mathrm{TE}_{XY} = a + bc$$

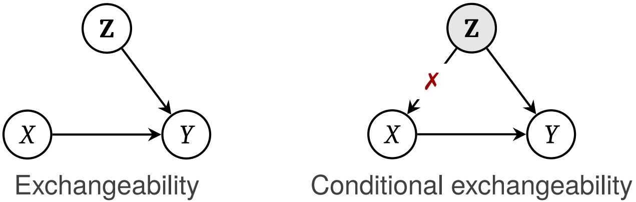
(Conditional) exchangeability

1. **Exchangeability:** Potential outcomes $Y(1)$ and $Y(0)$ are independent of treatment X.

$$Y(1), Y(0) \perp\!\!\!\perp X.$$

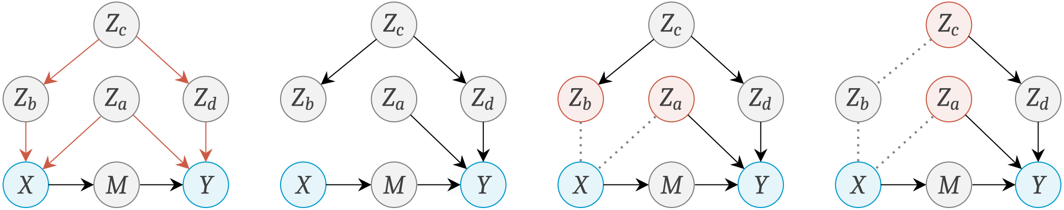
2. **Conditional exchangeability:** Exchangeable given a valid subset of confounders **Z**.

$$Y(1), Y(0) \perp\!\!\!\perp X \mid \mathbf{Z}.$$



(Conditional) exchangeability

1 Causal inference: The basics

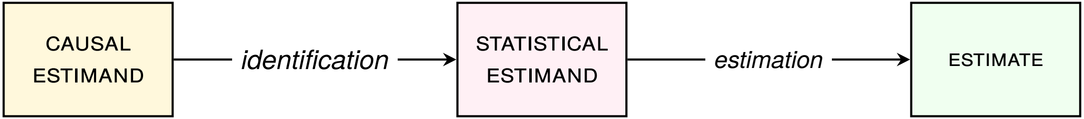


A: $p(y \mid x)$ **B**: $p(y \mid do(x))$ **C**: $p(y \mid x, z_a, z_b)$ **D**: $p(y \mid x, z_a, z_c)$

Conditional exchangeability simulates **unconfoundedness**:
all **backdoor paths** for X, Y are blocked.

From identification to estimation

- **Estimand**: a quantity of interest.
  - Causal: $\mathbb{E}[Y(1) - Y(0)]$
  - Statistical: $\mathbb{E}_{\mathbf{Z}}\left[\mathbb{E}[Y \mid X = 1, \mathbf{Z}] - \mathbb{E}[Y \mid X = 0, \mathbf{Z}]\right]$
- **Estimate**: an approximation obtained from data.



Slide adapted from Brady Neal's *Introduction to Causal Inference*

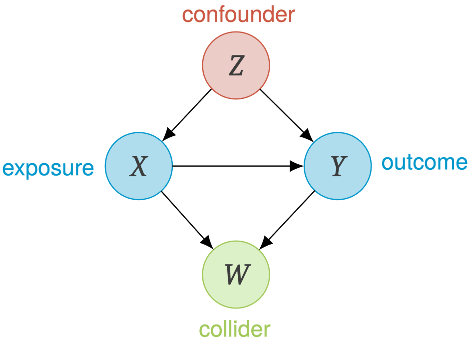
Definition: Average treatment effect (ATE)

Let X denote a binary treatment variable and Y an outcome. We can define a measure that estimates the **average difference in outcome** between treated and untreated individuals.

We express the ATE as the following difference of expectations:

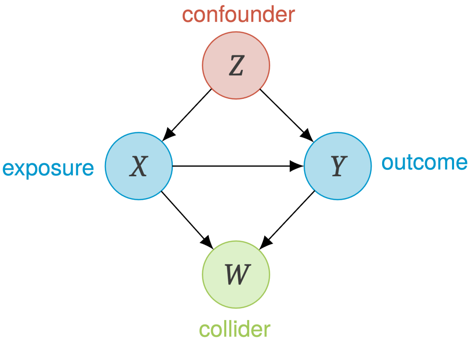
$$\text{ATE} := \mathbb{E}[Y \mid do(X = 1)] - \mathbb{E}[Y \mid do(X = 0)]. \tag{1}$$

Estimating the average treatment effect



Blocking all **backdoor paths** for $\{X, Y\}$ by adjusting for confounder Z: $Y(1), Y(0) \perp\!\!\!\perp X \mid Z$.

This removes *noncausal association* for unbiased ATE estimation.

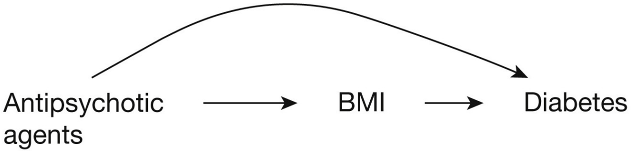


The correct directed acyclic graph (DAG) enables unique identification of the true ATE:

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_Z\left[\mathbb{E}[Y \mid X = 1, Z] - \mathbb{E}[Y \mid X = 0, Z]\right]$$

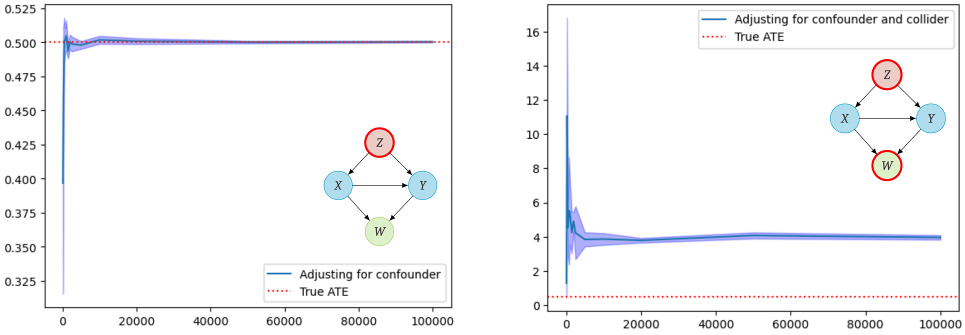
Estimating the average treatment effect

Real-world example: diabetes risk



Whether the patient takes certain antipsychotics is a confounder for BMI and risk of developing diabetes [10].

Effect estimation with a misspecified model



ATE estimates converge to the true value when controlling for Z only (left), but remain biased when controlling for $\{W, Z\}$ (right).

Causal discovery:
Learning structure from data

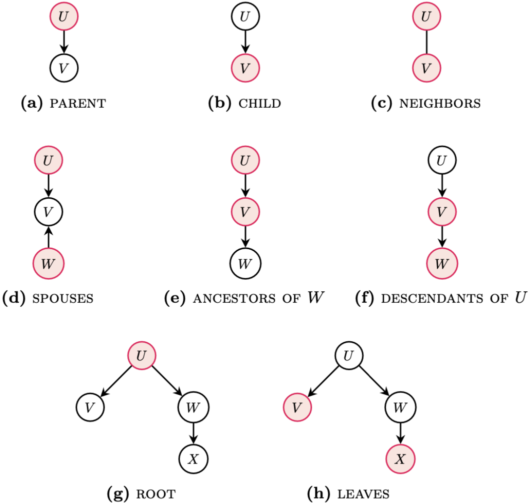
DAGs: A superexponential search space

2 Causal discovery: Learning structure from data

TOTAL NODES	TOTAL POSSIBLE DAGS	POWER OF 10
1	1	
2	3	
3	25	
4	543	$> 10^2$
5	29281	$> 10^4$
6	3781503	$> 10^6$
7	1138779265	$> 10^9$
8	783702329343	$> 10^{11}$
8	1213442454842881	$> 10^{15}$
10	4175098976430598143	$> 10^{18}$

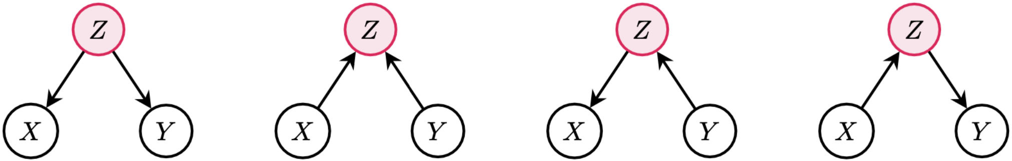
Structural relationships

2 Causal discovery: Learning structure from data



Structural primitives

2 Causal discovery: Learning structure from data



(a) FORK — $p(z)p(x|z)p(y|z)$ — $X \not\perp\!\!\!\perp Y; X \perp\!\!\!\perp Y \mid Z$

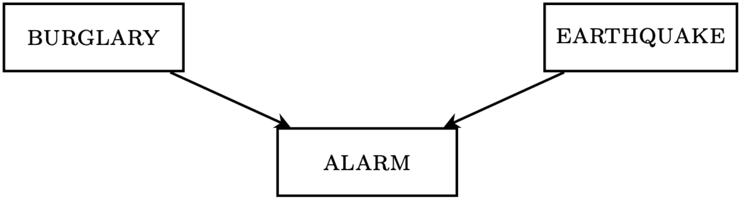
(b) v-STRUCTURE — $p(x)p(y)p(z|x,y)$ — $X \perp\!\!\!\perp Y; X \not\perp\!\!\!\perp Y \mid Z$

(c) CASCADE — $p(y)p(z|y)p(x|z)$ — $X \not\perp\!\!\!\perp Y; X \perp\!\!\!\perp Y \mid Z$

(d) CASCADE — $p(x)p(z|x)p(y|z)$ — $X \not\perp\!\!\!\perp Y; X \perp\!\!\!\perp Y \mid Z$

The v-structure

2 Causal discovery: Learning structure from data



Assume all variables are binary and no variables are latent. Knowledge of the value that ALARM and one of its parents takes can provide knowledge of the value that its other parent takes. For example, if ALARM = 1 and BURGLARY = 0, then we know that earthquake must equal 1.

Three major paradigms

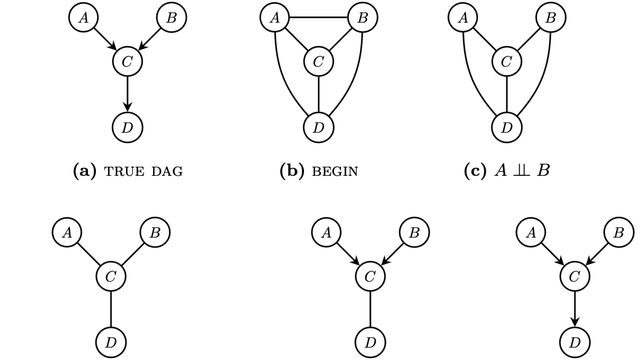
1. **Constraint-based**: Structure determined by conditional independence testing.

2. **Score-based**: Structure determined by a goodness-of-fit metric and search procedure.

3. **Functional causal models**: Parametric assumptions enable unique identification.

PC algorithm: Worked example

2 Causal discovery: Learning structure from data



(a) TRUE DAG **(b)** BEGIN **(c)** $A \perp\!\!\!\perp B$

(d) $A \perp\!\!\!\perp D \mid C, B \perp\!\!\!\perp D \mid C$ **(e)** v-STRUCTURE **(f)** OUTPUT

Figure adapted from [11].

Concluding remarks

Applications: AI/ML + causality

1. AI/ML for causal effect estimation (double ML, causal forests, etc).
2. AI/ML for causal discovery [12].
3. Causal representation learning [13].
4. Causal reasoning in LLMs: elicitation and evaluation [14, 15].
5. Causality for AI intepretability + explainability [16, 17].

Recommended reading

1. The free online course Introduction to Causal Inference by Brady Neal.
2. Applied Causal Inference Powered by ML & AI, Chernozhukov et al [18].
3. Most books by Judea Pearl.

Thank you! Any questions?

`maasch@cs.cornell.edu`

References

[1] J. Y. Halpern. “A modification of the Halpern-Pearl definition of causality”. In: *Proceedings of the 24th International Conference on Artificial Intelligence*. 2015, pp. 3022–3033.

[2] J. Kaddour et al. “Causal Machine Learning: A Survey and Open Problems”. In: (2022), p. 191.

[3] H. Reichenbach. *The Direction of Time*. University of California Press, 1956.

[4] J. Kendall. “Designing a research project: randomised controlled trials and their principles”. In: *Emergency medicine journal* 20.2 (2003), pp. 164–168.

[5] J. Pearl. “Direct and Indirect Effects”. In: *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence*. San Francisco. 2001.

[6] D. Koller et al. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[7] K. Sachs et al. “Causal protein-signaling networks derived from multiparameter single-cell data”. In: *Science* 308.5721 (2005), pp. 523–529.

[8] A. Darwiche. “Causal inference using tractable circuits”. In: *arXiv preprint arXiv:2202.02891* (2022).

[9] C. F. Manski. *Partial Identification and Credible Inference*. Springer, 2003.

[10] M. Etminan et al. “Using causal diagrams to improve the design and interpretation of medical research”. In: *Chest* 158.1 (2020), S21–S28.

References

[11] C. Glymour et al. “Review of causal discovery methods based on graphical models”. In: *Frontiers in genetics* 10 (2019), p. 524.

[12] X. Zheng et al. “DAGs with no tears: Continuous optimization for structure learning”. In: *Advances in neural information processing systems* 31 (2018).

[13] B. Schölkopf et al. “Toward causal representation learning”. In: *Proceedings of the IEEE* 109.5 (2021), pp. 612–634.

[14] E. Kıcıman et al. “Causal reasoning and large language models: Opening a new frontier for causality”. In: *arXiv preprint arXiv:2305.00050* (2023).

[15] A. Hüyük et al. “Reasoning Elicitation in Language Models via Counterfactual Feedback”. In: *Proceedings of the 13th International Conference on Learning Representations (ICLR)*. 2025.

[16] A. Geiger et al. “Causal abstraction: A theoretical foundation for mechanistic interpretability”. In: *arXiv preprint arXiv:2301.04709* (2023).

[17] L. Sharkey et al. “Open Problems in Mechanistic Interpretability”. In: *arXiv preprint arXiv:2501.16496* (2025).

[18] V. Chernozhukov et al. “Applied causal inference powered by ML and AI”. In: *arXiv preprint arXiv:2403.02467* (2024).