Causal Inference

Sanghack Lee

Graduate School of Data Science, Seoul National University

LG AI Research, AI Academy

Prerequisite

- Probability, Chain-Rule, Conditional Independence
- Graphs, Graphical Model, and d-separation

Credit

Many slides are inspired from lecture notes by Prof. Elias Bareinboim at Columbia University

Overview

What is Causality?

Definition

“Causality ... is influence by which **one** event, process, state, or object (a cause) **contributes to the production of another** event, process, state, or object (an effect) where the cause is **partly** responsible for the effect, and the effect is **partly** dependent on the cause.” (emphasis mine)*

- (News) Papers: ‘increases’, ’decreases’ vs. linked to, associated with
- Daily Life: because, hence, thus, due to, ...

* Wikipedia `https://en.wikipedia.org/wiki/Causality`

What is Causality?

Definition

"Causality ... is influence by which **one** event, process, state, or object (a cause) **contributes to the production of another** event, process, state, or object (an effect) where the cause is **partly** responsible for the effect, and the effect is **partly** dependent on the cause." (emphasis mine)*

- (News) Papers: 'increases', 'decreases' vs. linked to, associated with
- Daily Life: because, hence, thus, due to, ...

* Wikipedia https://en.wikipedia.org/wiki/Causality

Why Do We Study Causality?

- Definition of **Science**
  - "Knowledge or a system of knowledge covering general truths or **the operation of general laws** especially as obtained and tested through scientific method."*

- Causality in various academic disciplines
  - Physics, Chemistry, Biology, Climate Science,
  - Psychology, Social Science, Economics,
  - Epidemiology, Public Health
    (COVID-19, mask policy, social distancing, # of vaccination, side effects)

* Merriam-Webster dictionary

Why Do We Study Causality?

- Definition of **Science**
  - "Knowledge or a system of knowledge covering general truths or **the operation of general laws** especially as obtained and tested through scientific method."*
- Causality in various academic disciplines
  - Physics, Chemistry, Biology, Climate Science,
  - Psychology, Social Science, Economics,
  - Epidemiology, Public Health
    (COVID-19, mask policy, social distancing, # of vaccination, side effects)

* Merriam-Webster dictionary

How is Causality related to & {AI, ML, & DS}?

Artificial Intelligence

a rational agent performing actions to achieve a goal
e.g., reinforcement learning $\pi_\theta(\text{action} \mid \text{state})$

Machine Learning

Currently focused on learning correlations, e.g., $\hat{P}_\theta(y|\mathbf{x}) \approx P(y|\mathbf{x})$

Data Science

Capture, Process, Analyze (e.g., Stat, ML), Communicate with Data

How is Causality related to & {AI, ML, & DS}?

Artificial Intelligence

a rational agent performing actions to achieve a goal
e.g., reinforcement learning $\pi_\theta(\text{action} \mid \text{state})$

Machine Learning

Currently focused on learning correlations, e.g., $\hat{P}_\theta(y|\mathbf{x}) \approx P(y|\mathbf{x})$

Data Science

Capture, Process, Analyze (e.g., Stat, ML), Communicate with Data

How is Causality related to & {AI, ML, & DS}?

Artificial Intelligence

a rational agent performing actions to achieve a goal
e.g., reinforcement learning $\pi_\theta(\text{action} \mid \text{state})$

Machine Learning

Currently focused on learning correlations, e.g., $\hat{P}_\theta(y|\mathbf{x}) \approx P(y|\mathbf{x})$

Data Science

Capture, Process, Analyze (e.g., Stat, ML), Communicate with Data

Pearl's Causal Hierarchy

- Level 1: Associational or Observational
- Level 2: Interventional or Experimental
- Level 3: Counterfactual

Pearl's Causal Hierarchy

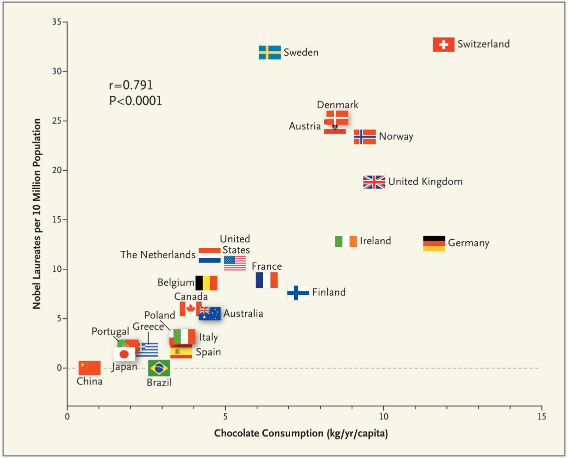
- Level 1: Associational or Observational
- Level 2: Interventional or Experimental
- Level 3: Counterfactual

Pearl's Causal Hierarchy

- Level 1: Associational or Observational
- Level 2: Interventional or Experimental
- Level 3: Counterfactual

Correlation (Level 1) vs. Causation (Level 2)

Chocolate Consumption vs. Nobel Laureates



Franz H. Messerli, Chocolate Consumption, Cognitive Function, and Nobel Laureates, N Engl J Med 2012; 367:1562-1564, `https://www.alexhahnillustrator.com`

Simpson's Paradox

Consider the following scenario:

1. A Patient with Kidney Stone visits a Hospital.
2. A Doctor examines the Patient and provides a Treatment.
3. The Patient's Health Outcome is later reported.

Healthcare Database!

Simpson's Paradox

		Treatment A	Treatment B
Stone	Small	Group 1 93% (81/ 87)	Group 2 87% (234/270)
	Large	Group 3 73% (192/263)	Group 4 69% (55/ 80)

Each cell represents $P(\text{success} \mid \text{treatment}, \text{stone})$

Simpson's Paradox

		Treatment A	Treatment B
Stone	Small	Group 1 93% (81/ 87)	Group 2 87% (234/270)
	Large	Group 3 73% (192/263)	Group 4 69% (55/ 80)
	Aggregated	78% (273/350)	83% (289/350)

Each cell represents $P(\text{success} \mid \text{treatment}, \text{stone})$

Simpson's Paradox

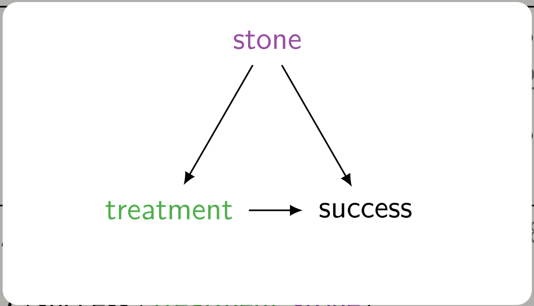
		Treatment A	Treatment B
Stone	Small	Group 1 93% (81/ 87)	Group 2 87% (234/270)
	Large	Group 3 73% (192/263)	Group 4 69% (55/ 80)
	Aggregated	78% (273/350)	83% (289/350)

Each cell represents $P(\text{success} \mid \text{treatment}, \text{stone})$

Aggregated: $P(\text{succ}|\text{treatment})$

Simpson's Paradox

		Treatment	
		A	B
Stone			up 2
			234/270)
			up 4
			55/ 80)
A			89/350)

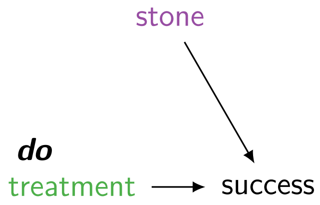


Each cell represents P (success | treatment, stone)

Aggregated: $P(\text{succ}|\text{treatment})$

Simpson's Paradox

		Treatment	
		A	B
Stone			...up 2
			...34/270)
			...up 4
			55/ 80)
A...			...89/350)



Each cell represents P(success | treatment, stone)

Aggregated: $P(\text{succ}|\text{treatment})$

Simpson's Paradox

		Treatment A	Treatment B
Stone	Small	Group 1 93% (81/ 87)	Group 2 87% (234/270)
Stone	Large	Group 3 73% (192/263)	Group 4 69% (55/ 80)
	Aggregated	78% (273/350)	83% (289/350)

What if we administer each treatment *randomly*?

Simpson's Paradox

		Treatment A	Treatment B
Stone	Small	Group 1 93% (81/ 87)	Group 2 87% (234/270)
	Large	Group 3 73% (192/263)	Group 4 69% (55/ 80)
	Efficacy	83.2%	78.2%

What if we administer each treatment *randomly*? The causal effect of A

$$P(\text{succ} \mid A) \neq P(\text{succ} \mid do(A)) = \sum_{\text{stone}} P(\text{succ} \mid A, \text{stone}) P(\text{stone})$$

Lesson's Learned from Simpson's Paradox

- Causal analyses need to be guided by subject-matter knowledge 📚.
- Identical data arising from different causal structures need to be analysed differently.
- No purely statistical rules exist to guide causal analyses.

Hernán et al. (2011) The Simpson's paradox unraveled. International Journal of Epidemiology 2011;40:780–-785

Data & Questions

Data scientists should take care of the types of **data** and **question**:

		Question	
		Non-Causal	Causal
Data	Non-Causal	Observational Study, Machine Learning*	**Causal Inference**
	Causal	**Causal Inference**	Experimental Study, Reinforcement Learning

*ML, in general, does not care about the type of data but the question should match the data type.

Formalizing Causality

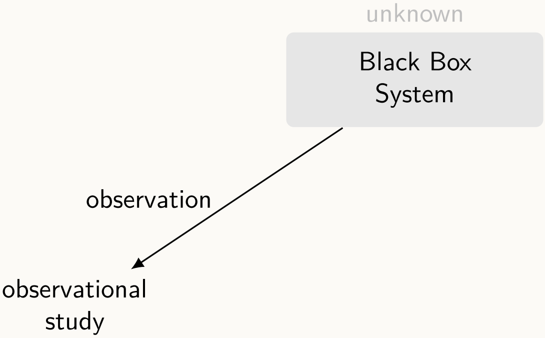
Observation & Intervention (Experiments)

unknown

Black Box
System

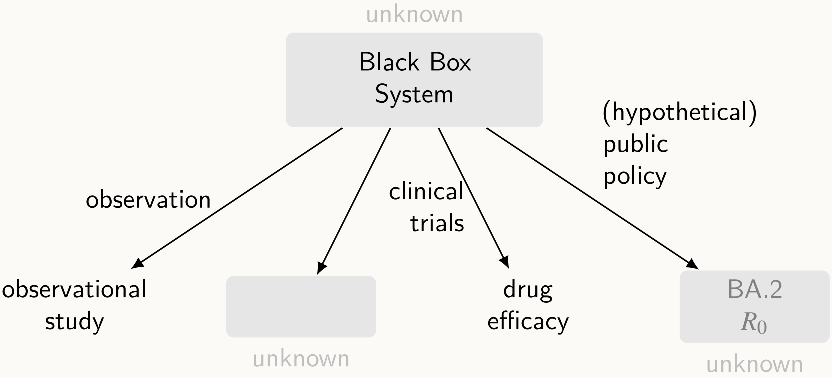
Formalizing Causality

Observation & Intervention (Experiments)



Formalizing Causality

Observation & Intervention (Experiments)



Causal Framework: Structural Causal Model

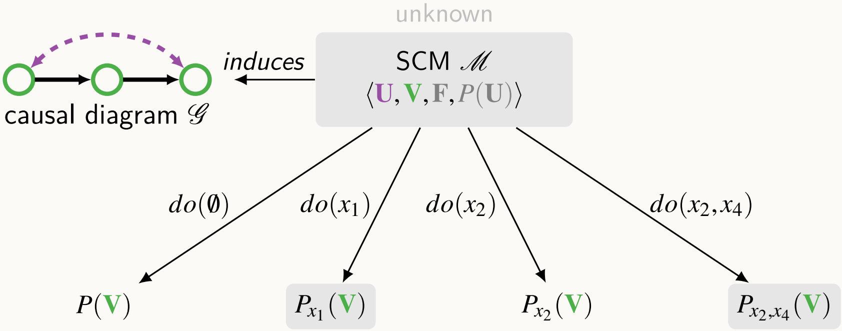
Structural Causal Model (SCM, Pearl [2000]) is a formal framework to study causality.

unknown

$$
\begin{tikzcd}
 & \text{SCM } \mathscr{M} \;\; \langle \mathbf{U},\mathbf{V},\mathbf{F},P(\mathbf{U})\rangle \arrow[dl, "do(\emptyset)"'] \arrow[d, "do(x_1)"] \arrow[dr, "do(x_2)"] \arrow[drr, "do(x_2,x_4)"] & & \\
P(\mathbf{V}) & P_{x_1}(\mathbf{V}) & P_{x_2}(\mathbf{V}) & P_{x_2,x_4}(\mathbf{V})
\end{tikzcd}
$$

Causal Framework: Structural Causal Model

Structural Causal Model (SCM, Pearl [2000]) is a formal framework to study causality.



Causal Framework: Structural Causal Models

Definition (Structural Causal Model)

A structural causal model (SCM) $\mathscr{M}$ is a 4-tuple $\langle \mathbf{V}, \mathbf{U}, \mathbf{F}, P(\mathbf{U}) \rangle$, where

- $\mathbf{U}$ is a set of **exogenous** variables;
- $P(\mathbf{U})$ is a distribution over $\mathbf{U}$;
- $\mathbf{V} = \{V_1, \ldots, V_n\}$ are **endogenous** variables;
- $\mathbf{F} = \{f_1, \ldots f_n\}$ are functions determining $\mathbf{V}$,

$$v_i \leftarrow f_i(\mathbf{pa}_i, \mathbf{u}_i)$$

where $\mathbf{Pa}_i \subseteq \mathbf{V} \setminus \{V_i\}$, $\mathbf{U}_i \subseteq \mathbf{U}$.

Causal Diagram $\mathscr{G}$ is a View for Causal Model $\mathscr{M}$

$$\langle \underbrace{\mathbf{V}}_{\text{observed}}, \underbrace{\mathbf{U}}_{\text{unobserved}}, \underbrace{\mathbf{F}}_{\text{mechanisms for } \mathbf{V}}, P(\mathbf{U})\rangle$$

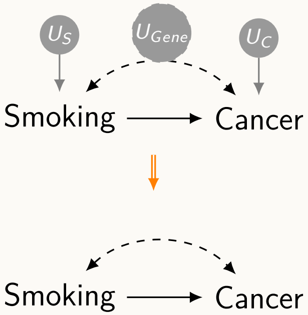
- $\mathbf{V} = \{\text{Smoking}, \text{Cancer}\}$
- $\mathbf{U} = \{U_S, U_C, U_{Gene}\}$
- $\mathbf{F}$:
$$\begin{cases} \text{Smoking} & \leftarrow f_{\text{Smoking}}(U_S, U_{Gene}) \\ \text{Cancer} & \leftarrow f_{\text{Cancer}}(\text{Smoking}, U_C, U_{Gene}) \end{cases}$$

Causal Diagram $\mathscr{G}$ is a View for Causal Model $\mathscr{M}$

$$\langle \underbrace{\mathbf{V}}_{\text{observed}}, \underbrace{\mathbf{U}}_{\text{unobserved}}, \underbrace{\mathbf{F}}_{\text{mechanisms for } \mathbf{V}}, P(\mathbf{U}) \rangle$$

- $\mathbf{V} = \{\text{Smoking}, \text{Cancer}\}$
- $\mathbf{U} = \{U_S, U_C, U_{Gene}\}$
- $\mathbf{F}$:

$$\begin{cases} \text{Smoking} & \leftarrow f_{\text{Smoking}}(U_S, U_{Gene}) \\ \text{Cancer} & \leftarrow f_{\text{Cancer}}(\text{Smoking}, U_C, U_{Gene}) \end{cases}$$



Intervention — $do(\cdot)$ operator

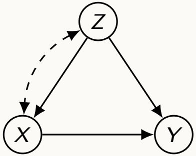
- Given a model $\mathscr{M}$ the action of fixing any observable variable $X \in \mathbf{V}$ to a constant value x is denoted using the $do(\cdot)$ operator as $do(X = x)$.

Intervention — $do(\cdot)$ operator

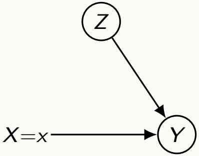
- Given a model $\mathscr{M}$ the action of fixing any observable variable $X \in \mathbf{V}$ to a constant value x is denoted using the $do(\cdot)$ operator as $do(X = x)$.
- This operation gives birth to a **submodel** $\mathscr{M}_x$ that looks exactly like $\mathscr{M}$ but with functions where f_x has been replaced with a constant x.

Intervention — $do(\cdot)$ operator

- Given a model $\mathscr{M}$ the action of fixing any observable variable $X \in \mathbf{V}$ to a constant value x is denoted using the $do(\cdot)$ operator as $do(X = x)$.
- This operation gives birth to a **submodel** $\mathscr{M}_x$ that looks exactly like $\mathscr{M}$ but with functions where f_x has been replaced with a constant x.
- These two graphs represent the world *before* and *after* an intervention $do(X = x)$.



Causal Graph $\mathscr{G}$



Causal Graph under Intervention $\mathscr{G}_{\overline{X}}$

Intervention — Causal Effects

Definition (Causal Effect)

Given two disjoint sets of variables, $\mathbf{X}$ and $\mathbf{Y}$, the causal effect of $\mathbf{X}$ on $\mathbf{Y}$, denoted as $P(\mathbf{y}|do(\mathbf{x}))$ or $P_{\mathbf{x}}(\mathbf{y})$, is a function from $\mathbf{X}$ to the space of probability distributions of $\mathbf{Y}$.

Intervention — Causal Effects

Definition (Causal Effect)

Given two disjoint sets of variables, $\mathbf{X}$ and $\mathbf{Y}$, the causal effect of $\mathbf{X}$ on $\mathbf{Y}$, denoted as $P(\mathbf{y}|do(\mathbf{x}))$ or $P_{\mathbf{x}}(\mathbf{y})$, is a function from $\mathbf{X}$ to the space of probability distributions of $\mathbf{Y}$.

Researchers may be interested in

- Expectation: $\mathbb{E}[Y|do(x)]$
- Difference: $\mathbb{E}[Y|do(X=1)] - \mathbb{E}[Y|do(X=0)]$ (Average Treatment Effect, ATE)
- Conditional: $\mathbb{E}[Y|do(X=1),\mathbf{Z}] - \mathbb{E}[Y|do(X=0),\mathbf{Z}]$ (Conditional ATE)

Reading Conditional Independence from Causal Diagram

'Separation' in graph $\mathscr{G}$ implies 'Conditional Independence' in distribution P:

$$(X \perp\!\!\!\perp_{\mathscr{G}} Y \mid Z) \implies (X \perp\!\!\!\perp_{P} Y \mid Z).$$

Reading Conditional Independence from Causal Diagram

'Separation' in graph $\mathscr{G}$ implies 'Conditional Independence' in distribution P:

$$(X \perp\!\!\!\perp_{\mathscr{G}} Y \mid Z) \Longrightarrow (X \perp\!\!\!\perp_{P} Y \mid Z).$$

$$\begin{tikzcd}
 & \text{season} \arrow[dl] \arrow[dr] & \\
\text{sprinkler} \arrow[dr] & & \text{rain} \arrow[dl] \\
 & \text{wet} \arrow[d] & \\
 & \text{slippery} &
\end{tikzcd}$$

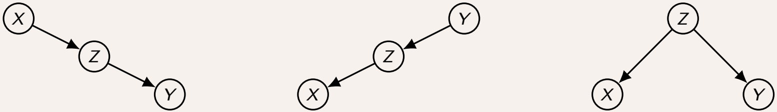
1. ✗ CI: (Wet $\perp\!\!\!\perp$ Sprinkler)
2. ✗ CI: (Wet $\perp\!\!\!\perp$ Season | Sprinkler)
3. ✓ CI: (Rain $\perp\!\!\!\perp$ Slippery | Wet)
4. ✓ CI: (Season $\perp\!\!\!\perp$ Wet | Sprinkler, Rain)
5. ✗ CI: (Sprinkler $\perp\!\!\!\perp$ Rain | Season, Wet)

Reading Conditional Independence from Causal Diagram

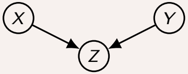
Definition (d-separation)

Two vertices X, Y are said to be d-separated by a set $\mathbf{Z}$ in a directed acyclic graph $\mathscr{G}$, denoted by $(X \perp\!\!\!\perp_{\mathscr{G}} Y \mid \mathbf{Z})$, if every path $\mathbf{p}$ from X to Y are blocked where blockage occurs when one of the following holds:

1. $\mathbf{p}$ contains at least one arrow-emitting node that is in $\mathbf{Z}$, or



2. $\mathbf{p}$ contains at least one collider that is outside $\mathbf{Z}$ and has no descendant in $\mathbf{Z}$.



Generalized to two sets $\mathbf{X}$ and $\mathbf{Y}$ and with bidirected edges.

Summary for Part 1

- **Structural Causal Model** $\mathscr{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ provides a formal framework.
- SCM induces observational, interventional, and counterfactual distributions.
- SCM *induces* a **causal graph** $\mathscr{G}$, which implies **conditional independencies** testable via **d-separation** (blockage).
- The underlying model $\mathscr{M}$ is unknown but the causal graph $\mathscr{G}$ can be given from *common sense* or *domain expertise*.
- **Intervention** $do(\mathbf{X} = \mathbf{x})$ as a **submodel** $\mathscr{M}_{\mathbf{x}}$, which induces a manipulated causal graph $\mathscr{G}_{\overline{\mathbf{X}}}$.
- **Causal effect** of $\mathbf{X} = \mathbf{x}$ on $\mathbf{Y} = \mathbf{y}$ is defined as $P(\mathbf{y} | do(\mathbf{x}))$.

Summary for Part 1

- **Structural Causal Model** $\mathscr{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ provides a formal framework.
- SCM induces observational, interventional, and counterfactual distributions.
- SCM *induces* a **causal graph** $\mathscr{G}$, which implies **conditional independencies** testable via **d-separation** (blockage).
- The underlying model $\mathscr{M}$ is unknown but the causal graph $\mathscr{G}$ can be given from *common sense* or *domain expertise*.
- **Intervention** $do(\mathbf{X} = \mathbf{x})$ as a **submodel** $\mathscr{M}_{\mathbf{x}}$, which induces a manipulated causal graph $\mathscr{G}_{\overline{\mathbf{X}}}$.
- **Causal effect** of $\mathbf{X} = \mathbf{x}$ on $\mathbf{Y} = \mathbf{y}$ is defined as $P(\mathbf{y}|do(\mathbf{x}))$.

Summary for Part 1

- **Structural Causal Model** $\mathscr{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ provides a formal framework.
- SCM induces observational, interventional, and counterfactual distributions.
- SCM *induces* a **causal graph** $\mathscr{G}$, which implies **conditional independencies** testable via **d-separation** (blockage).
- The underlying model $\mathscr{M}$ is unknown but the causal graph $\mathscr{G}$ can be given from *common sense* or *domain expertise*.
- **Intervention** $do(\mathbf{X} = \mathbf{x})$ as a **submodel** $\mathscr{M}_{\mathbf{x}}$, which induces a manipulated causal graph $\mathscr{G}_{\overline{\mathbf{X}}}$.
- **Causal effect** of $\mathbf{X} = \mathbf{x}$ on $\mathbf{Y} = \mathbf{y}$ is defined as $P(\mathbf{y} | do(\mathbf{x}))$.

Summary for Part 1

- **Structural Causal Model** $\mathscr{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ provides a formal framework.
- SCM induces observational, interventional, and counterfactual distributions.
- SCM *induces* a **causal graph** $\mathscr{G}$, which implies **conditional independencies** testable via **d-separation** (blockage).
- The underlying model $\mathscr{M}$ is unknown but the causal graph $\mathscr{G}$ can be given from *common sense* or *domain expertise*.
- **Intervention** $do(\mathbf{X} = \mathbf{x})$ as a **submodel** $\mathscr{M}_{\mathbf{x}}$, which induces a manipulated causal graph $\mathscr{G}_{\overline{\mathbf{X}}}$.
- **Causal effect** of $\mathbf{X} = \mathbf{x}$ on $\mathbf{Y} = \mathbf{y}$ is defined as $P(\mathbf{y} | do(\mathbf{x}))$.

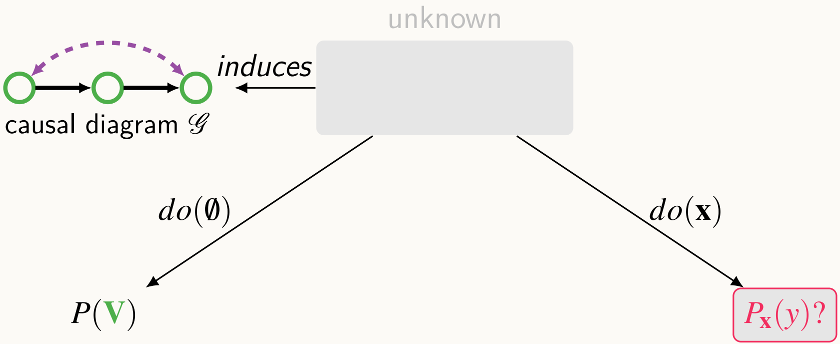
Summary for Part 1

- **Structural Causal Model** $\mathscr{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ provides a formal framework.
- SCM induces observational, interventional, and counterfactual distributions.
- SCM *induces* a **causal graph** $\mathscr{G}$, which implies **conditional independencies** testable via **d-separation** (blockage).
- The underlying model $\mathscr{M}$ is unknown but the causal graph $\mathscr{G}$ can be given from *common sense* or *domain expertise*.
- **Intervention** $do(\mathbf{X} = \mathbf{x})$ as a **submodel** $\mathscr{M}_{\mathbf{x}}$, which induces a manipulated causal graph $\mathscr{G}_{\overline{\mathbf{X}}}$.
- **Causal effect** of $\mathbf{X} = \mathbf{x}$ on $\mathbf{Y} = \mathbf{y}$ is defined as $P(\mathbf{y} | do(\mathbf{x}))$.

Summary for Part 1

- **Structural Causal Model** $\mathscr{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ provides a formal framework.
- SCM induces observational, interventional, and counterfactual distributions.
- SCM *induces* a **causal graph** $\mathscr{G}$, which implies **conditional independencies** testable via **d-separation** (blockage).
- The underlying model $\mathscr{M}$ is unknown but the causal graph $\mathscr{G}$ can be given from *common sense* or *domain expertise*.
- **Intervention** $do(\mathbf{X} = \mathbf{x})$ as a **submodel** $\mathscr{M}_{\mathbf{x}}$, which induces a manipulated causal graph $\mathscr{G}_{\overline{\mathbf{X}}}$.
- **Causal effect** of $\mathbf{X} = \mathbf{x}$ on $\mathbf{Y} = \mathbf{y}$ is defined as $P(\mathbf{y} | do(\mathbf{x}))$.

Next Part ...



* illustrative purposes

Overview

Part 1: Causality

Part 2: Causal Effect Identification

- Back-door Criterion
- Do-Calculus

Part 3: Modern Identification

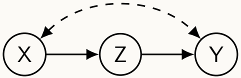
- Generalized Identification
- Transportability
- Recovering from Selection Bias
- Recovering from Missing Data

Causal Effect Identifiability

1. Query

$P_{\mathbf{x}}(\mathbf{y}) = P(\mathbf{y} | do(\mathbf{x}))$

2. Causal Diagram

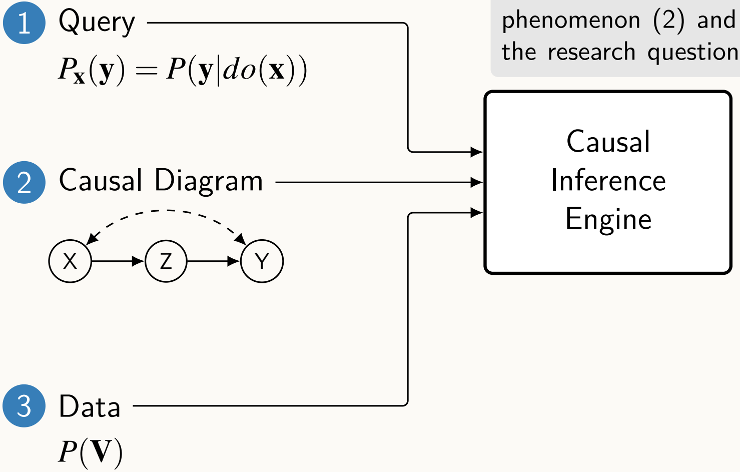


3. Data

$P(\mathbf{V})$

Causal Effect Identifiability

Based on the current knowledge about the phenomenon (2) and the available data (3), is the research question (1) identifiable?



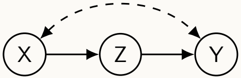
Causal Effect Identifiability

Based on the current knowledge about the phenomenon (2) and the available data (3), is the research question (1) identifiable?

1. Query

   $P_{\mathbf{x}}(\mathbf{y}) = P(\mathbf{y}|do(\mathbf{x}))$

2. Causal Diagram



3. Data

   $P(\mathbf{V})$

Causal Inference Engine

Solution

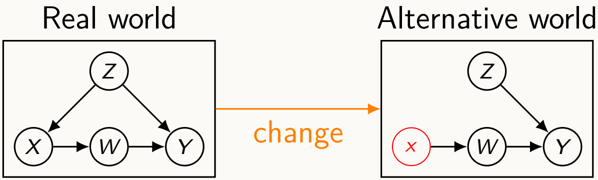
Yes / No

formula

$P_{\mathbf{x}}(\mathbf{y}) = f_{\mathscr{G}}(P(\mathbf{V}))$

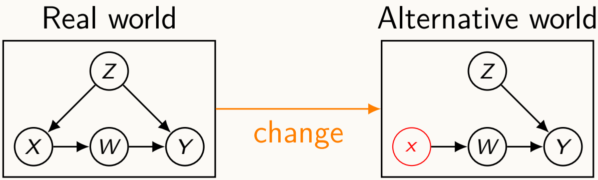
evidence

Computing Causal Effects from Observational Data



Z: age, sex
X: action
W: mediator
Y: outcome

Computing Causal Effects from Observational Data



Z: age, sex
X: action
W: mediator
Y: outcome

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

$$\xrightarrow{do(x)}$$

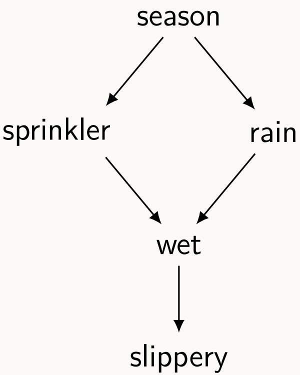
$$P_x(\mathbf{v} \backslash \{X\}) = P(z) \times \cancel{P(x|z)} \times P(w|x) \times P(y|w,z)$$

$\longleftarrow$ equal to 1 in $\mathscr{M}_x$

Computing Causal Effects from Observational Data

This distribution decomposes as

$$P(\mathbf{V}) = P(\mathsf{Sl}|W)P(W|\mathsf{Sp},R)P(\mathsf{Sp}|\mathsf{Sn})P(R|\mathsf{Sn})P(\mathsf{Sn})$$



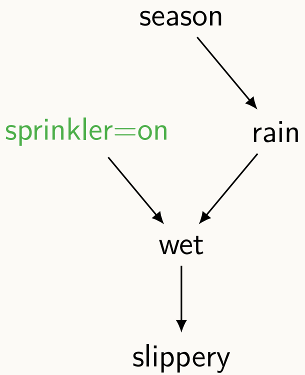
sn:season, sp:sprinkler, sl:slippery

Computing Causal Effects from Observational Data

This distribution decomposes as

$$P(\mathbf{V}) = P(\mathsf{Sl}|W)P(W|\mathsf{Sp},R)P(\mathsf{Sp}|\mathsf{Sn})P(R|\mathsf{Sn})P(\mathsf{Sn})$$

$$P(W \mid do(\mathsf{Sp} = \mathsf{on}))$$

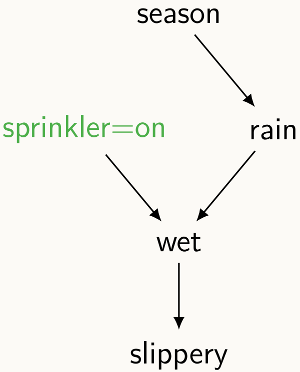


sn:season, sp:sprinkler, sl:slippery

Computing Causal Effects from Observational Data

This distribution decomposes as

$$P(\mathbf{V}) = P(\mathsf{Sl}|W)P(W|\mathsf{Sp},R)P(\mathsf{Sp}|\mathsf{Sn})P(R|\mathsf{Sn})P(\mathsf{Sn})$$



$$P(W \mid do(\mathsf{Sp} = \mathsf{on}))$$
$$= \sum_{\mathsf{sn},r,\mathsf{sl}} P(W, \mathsf{sn}, r, \mathsf{sl} \mid do(\mathsf{Sp} = \mathsf{on}))$$
$$= \sum_{\mathsf{sn},r,\mathsf{sl}} P(\mathsf{sl}|W)P(W|\mathsf{Sp} = on, r)P(r|\mathsf{sn})P(\mathsf{sn})$$

sn:season, sp:sprinkler, sl:slippery

Adjustment by Direct Parents for Singleton Intervention

Theorem

The causal effect $Q = P(\mathbf{y}|do(x))$ *is identifiable whenever* $X, \mathbf{Y}, \mathbf{Pa}_X \subseteq \mathbf{V}$ *(all parents of* X*) are measured.*[1]

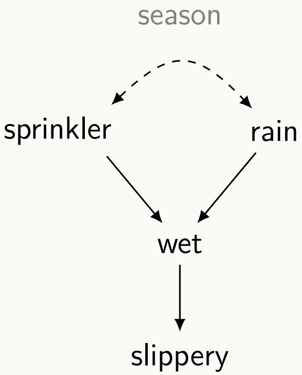
The expression of Q *is then obtained by adjustment for* $\mathbf{Pa}_X$*, or*

$$P(\mathbf{y}|do(x)) = \sum_{\mathbf{pa}_X} P(\mathbf{y}|x, \mathbf{pa}_X)P(\mathbf{pa}_X).$$

e.g.,

$$\sum_{\mathsf{sn}} P(W|\mathsf{Sp} = on, \mathsf{sn})P(\mathsf{sn})$$

If Season is latent, is the effect still computable?



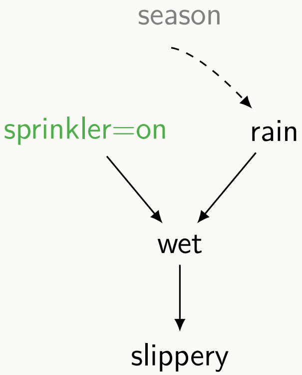
If Season is latent, is the effect still computable?

$$P(W \mid do(\text{Sp} = \text{on}))$$

$$= \sum_{\text{sn},r} P(W \mid \text{Sp} = \text{on}, r)P(\text{sn})P(r|\text{sn})$$

$$= \sum_{r} P(W \mid \text{Sp} = \text{on}, r)\sum_{\text{sn}} P(r, \text{sn})$$

$$= \sum_{r} P(W \mid \text{Sp} = \text{on}, r)P(r)$$



Overview

Back-door Criterion

$$\begin{tikzcd} & \text{kidney stone } Z \arrow[dl] \arrow[dr] & \\ \text{treatment } X \arrow[rr] & & \text{health } Y \end{tikzcd}$$

Definition (Back-door)

Find a set $\mathbf{Z}$ such that it can sufficiently explain 'confounding' between $\mathbf{X}$ and $\mathbf{Y}$. Then,

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

Back-door Criterion

Definition (Back-door Criterion)

A set $\mathbf{Z}$ satisfies the back-door criterion with respect to a pair of variables $\mathbf{X}$, $\mathbf{Y}$ in a causal diagram $\mathscr{G}$ if;

(i) no node in $\mathbf{Z}$ is a descendant of $\mathbf{X}$; and

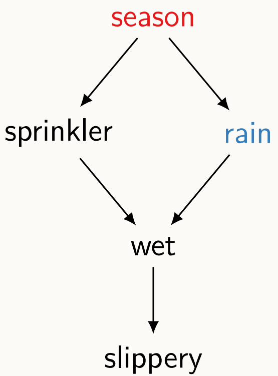
(ii) $\mathbf{Z}$ blocks every path between $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ that contains an arrow into X.

Back-door sets as substitutes of the direct parents of X

Rain satisfies the back-door criterion relative to Sprinkler and Wet:

(i) Rain is not a descendant of Sprinkler, and

(ii) Rain blocks the only back-door path from Sprinkler to Wet.

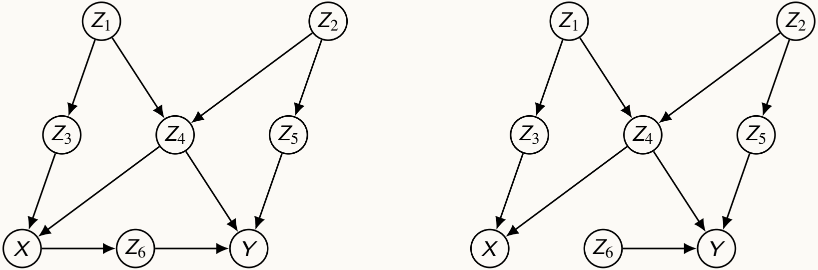


Adjusting for the direct parents of Sprinkler, we have:

$$\begin{aligned} P(\mathsf{wt}|do(\mathsf{sp})) &= \sum_{\mathsf{sn}} P(\mathsf{wt}|\mathsf{sp},\mathsf{sn})P(\mathsf{sn}) \\ &\vdots \\ &= \sum_{\mathsf{rn}} P(\mathsf{wt}|\mathsf{sp},\mathsf{rn})P(\mathsf{rn}) \end{aligned}$$

A Graphical Condition for Back-door Admissible Sets

$P(\mathbf{y}|do(\mathbf{x}))$ is identifiable if (i) & (ii) there is a set that d-sep. $\mathbf{X}$ from $\mathbf{Y}$ in $\mathcal{G}_{\underline{\mathbf{X}}}$.



$$P(y|do(x)) = \sum_{z_1,z_4} P(y|x,z_1,z_4)P(z_1,z_4)$$

Overview

Part 1: Causality

Part 2: Causal Effect Identification

- Back-door Criterion
- Do-Calculus

Part 3: Modern Identification

- Generalized Identification
- Transportability
- Recovering from Selection Bias
- Recovering from Missing Data

Rules of Do-calculus

- Backdoor criterion results in a very specific form of identification formula.
- **Do-Calculus** [Pearl 1995] provides general machinery to manipulate observational and interventional distributions.

$$\text{Level 1 Associational} \Leftrightarrow \text{Level 2 Experimental}$$

Rules of Do-calculus

Theorem (Rules of Do-calculus (simplified))

Rule 1: *Adding/removing Observations*

$$P(\mathbf{y}|do(\mathbf{x}),\mathbf{z}) = P(\mathbf{y}|do(\mathbf{x})) \qquad \textit{if } (\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}) \textit{ in } \mathscr{G}_{\overline{\mathbf{X}}}.$$

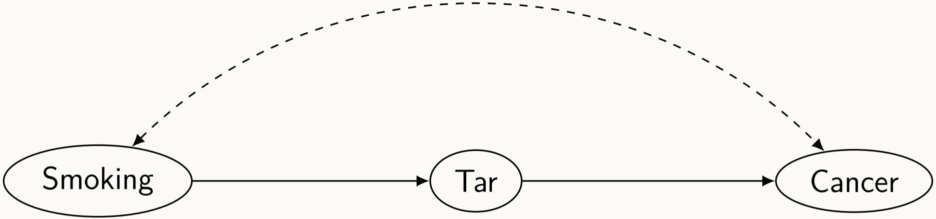
Rule 2: *Action/observation Exchange*

$$P(\mathbf{y}|do(\mathbf{x}),do(\mathbf{z})) = P(\mathbf{y}|do(\mathbf{x}),\mathbf{z}) \qquad \textit{if } (\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}) \textit{ in } \mathscr{G}_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}.$$

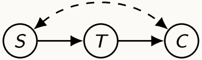
Rule 3: *Adding/removing Actions*

$$P(\mathbf{y}|do(\mathbf{x}),do(\mathbf{z})) = P(\mathbf{y}|do(\mathbf{x})) \qquad \textit{if } (\mathbf{Z} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}) \textit{ in } \mathscr{G}_{\overline{\mathbf{X}\mathbf{Z}}}$$

Do-calculus in Action



Do-calculus in Action

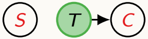


$$P(c|do(s))$$

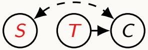
$$= \sum_t P(c|do(s),t)P(t|do(s))$$ Probability Axioms

$$= \sum_t P(c|do(s),do(t))P(t|do(s))$$ Rule 2 $(T \perp\!\!\!\perp C \mid S)_{\mathscr{G}_{\overline{S}\underline{T}}}$

$$= \sum_t P(c|do(t))P(t|do(s))$$ Rule 3 $(S \perp\!\!\!\perp C \mid T)_{\mathscr{G}_{\overline{T,S}}}$



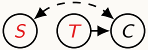
$$= \sum_t P(c|do(t))P(t|s)$$ Rule 2 $(T \perp\!\!\!\perp S)_{\mathscr{G}_{\underline{S}}}$



$$= \sum_t P(t|s) \sum_{s'} P(c|s',do(t))P(s'|do(t))$$ Probability Axioms

$$= \sum_t P(t|s) \sum_{s'} P(c|s',t)P(s'|do(t))$$ Rule 2 $(T \perp\!\!\!\perp C \mid S)_{\mathscr{G}_{\underline{T}}}$

$$= \sum_t P(t|s) \sum_{s'} P(c|s',t)P(s')$$ Rule 3 $(T \perp\!\!\!\perp S)_{\mathscr{G}_{\overline{T}}}$



Smoking, **T**ar deposit in lung, and **C**ancer

Algorithmic Identification

- **Do-calculus** is sound and complete but it has no algorithmic insight.
- A graphical condition and an efficient **algorithmic** procedure have developed for identifiability.

Finite sample and estimation is another important issue.

Summary for Part 2

- Identifiability: Causal Effect may be computable from existing observational data for some causal graphs.
- In a Markovian case and singleton X, a causal effect can be easily derivable by canceling out $P(x|\mathbf{pa}_x)$.
- A back-door adjustment formula is simple and widely used but limited.
- Do-calculus is a set of rules to manipulate observational or interventional probabilities. (Do-calculus is complete)
- There exists a polynomial time algorithm to yield a causal effect formula (whenever identifiable) given an arbitrary causal diagram.

Overview

Various Data Sources

Target ◎

$Q = P^*(y|do(x))$

Various Data Sources

Target $Q = P^*(y\|do(x))$	Dataset 1	Dataset 2	Dataset n

Various Data Sources

Target ◎ $Q = P^*(y \mid do(x))$	Dataset 1	Dataset 2	Dataset n

d_1	Population	**Los Angeles**	**New York**	**Seoul**
d_2	Obs./Exp.	Experimental	Observational	Experimental
	Treatment Assignment	Randomized Z_1	-	Randomized Z_2
d_3	Sampling	Selection on Age	Selection on SES	-
d_4	Measured	$\{X_1, Z_1, W, M, Y_1\}$	$\{X_1, X_2, Z_1, N, Y_2\}$	$\{X_2, Z_1, W, L, M, Y_1\}$

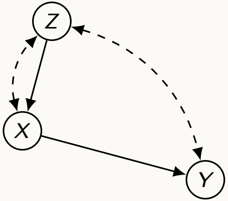
SES = Socio-Economic Status

Modern Identification Tasks

1. Experimental conditions
   Generalized Identification
2. Environmental conditions
   Transportability
3. Sampling conditions
   Recovering from Selection Bias
4. Responding conditions
   Recovering from Missingness

Overview

Generalized Identifiability

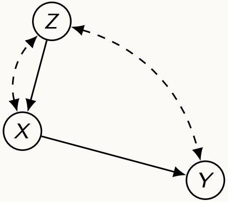


Z: Diet
X: Cholesterol Level
Y: Heart Attack

Measured:
Observational study: $P(X, Y, Z)$

Needed: $Q = P(y|do(x))$?

Generalized Identifiability

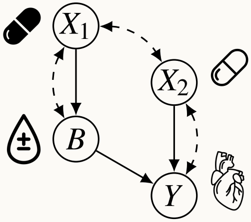


Z: Diet
X: Cholesterol Level
Y: Heart Attack

Measured:
Observational study: $P(X,Y,Z)$
Experimental Study: $P(X,Y|do(Z))$

Needed: $Q = P(y|do(x)) = \dfrac{P(x,y|do(z))}{P(x|do(z))}$

Generalized Identifiability: Drug-Drug Interactions

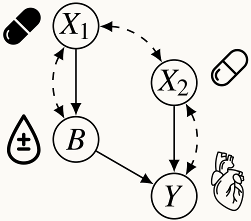


$$P_{x_1,x_2}(y) \Leftarrow \{P_{x_1}(\mathbf{V}), P_{x_2}(\mathbf{V})\}?$$

Y *cardiovascular disease*; B *blood pressure*; X_1 taking an *antihypertensive drug*; and X_2 the use of an *anti-diabetic drug*.

Goal: assess the effect of prescribing **both** treatments ( ) on the risk of cardiovascular diseases from **individual** drug experiments, either or .

Generalized Identifiability: Drug-Drug Interactions



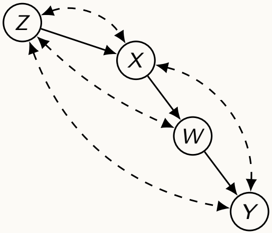
$$P_{x_1,x_2}(y) = \sum_b P_{x_2}(y|b)P_{x_1}(b)$$

Y *cardiovascular disease*; B *blood pressure*; X_1 taking an *antihypertensive drug*; and X_2 the use of an *anti-diabetic drug*.

Goal: assess the effect of prescribing **both** treatments ( ) on the risk of cardiovascular diseases from **individual** drug experiments, either or .

General Identifiability reduced to Calculus

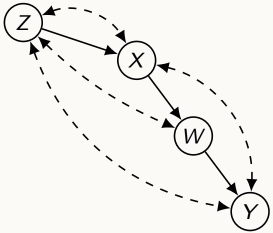
$$
\begin{aligned}
P(y|do(x)) &= \sum_w P(y|do(x), w)P(w|do(x)) \\
&= \sum_w P(y|do(x, w))P(w|do(x)) \\
&= \sum_w \underbrace{P(y|do(w))}_{Q[Y]} \underbrace{P(w|do(x))}_{Q[W]}
\end{aligned}
$$



Both effects are not identifiable from $P(\mathbf{V})$.

General Identifiability reduced to Calculus

$$
\begin{aligned}
Q[Y] &= P(y|do(w)) \\
&= P(y|do(w,z)) \\
&= \sum_x P(y|do(w,z),x)P(x|do(w,z)) \\
&= \sum_x P(y|do(w,z),x)P(x|do(z)) \\
&= \sum_x P(y|do(z),w,x)P(x|do(z)).
\end{aligned}
$$



$$
\begin{aligned}
Q[W] &= P(w|do(x)) \\
&= P(w|do(x,z)) \\
&= P(w|do(z),x).
\end{aligned}
$$

Available from $P(\mathbf{V}|do(z))$!

Summary for General Identifiability

The identifiability of any expression of the form

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

can be determined given any causal graph $\mathscr{G}$ and an arbitrary **combination of observational and experimental studies**.

If the query is identifiable, then its estimand can be derived in *polynomial time*.

Lee, Correa, and Bareinboim (2019) *General Identifiability with Arbitrary Surrogate Experiments*, UAI.

Overview

Transportability

Is it possible to compute the effect of $\mathbf{X}$ on $\mathbf{Y}$ in a target environment ◎, using **observational and experimental findings** from different populations?

e.g., applying education policies of U.S. to South Korea.

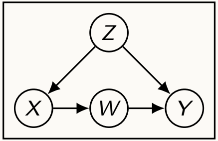
Environment can be understood as population, domain, etc.

How is this Problem seen in other Sciences?

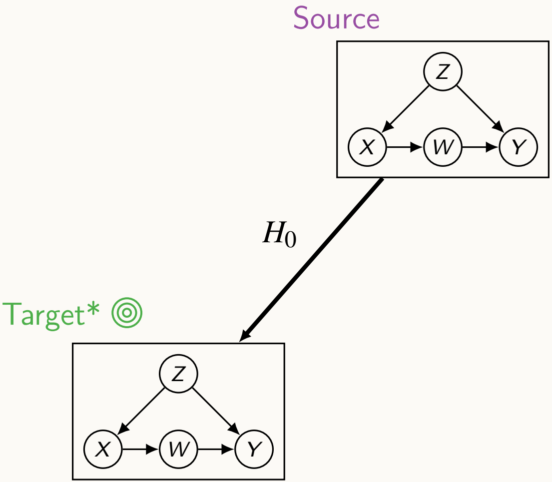
- "**External Validity** asks the question of generalizability: To what populations, settings, treatment variables, and measurement variables can this effect be generalized?" (Shadish, Cook, and Campbell, 2002)
- "**Extrapolation** across studies requires 'some understanding of the reasons for the differences.' " (Cox, 1958)
- "An experiment is said to have "**external validity**" if the distribution of outcomes realized by a treatment group is the same as the distribution of outcome that would be realized in an actual program." (Manski, 2007)

Transportability: the Spectrum

Source

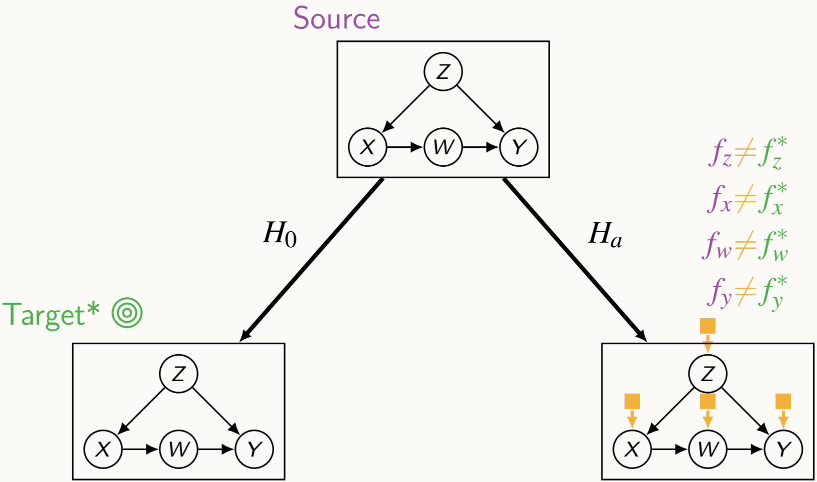


Transportability: the Spectrum



Everything is assumed to be
the same, trivially transportable!

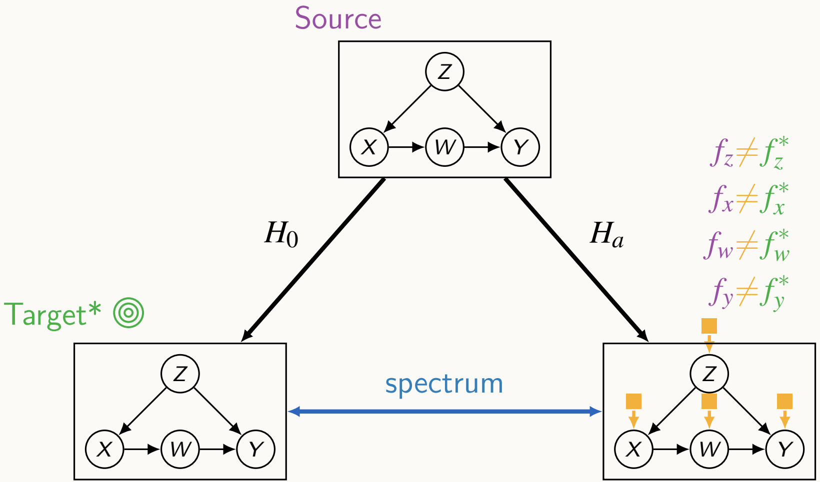
Transportability: the Spectrum



Everything is assumed to be
the same, trivially transportable!

Everything is assumed to be
different, *not* transportable!

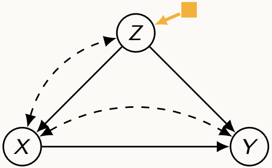
Transportability: the Spectrum



Everything is assumed to be
the same, trivially transportable!

Everything is assumed to be
different, *not* transportable!

Transportability: Formulas Depend on the Causal Story



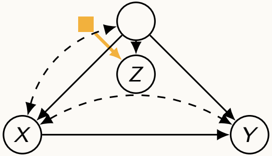
(a)

(a) Z represents **age**

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

P^* target domain & P source domain

Transportability: Formulas Depend on the Causal Story



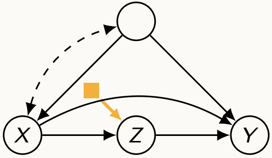
(b)

(b) Z represents **language skill**

$$P^*(y|do(x)) = P(y|do(x))$$

P^* target domain & P source domain

Transportability: Formulas Depend on the Causal Story



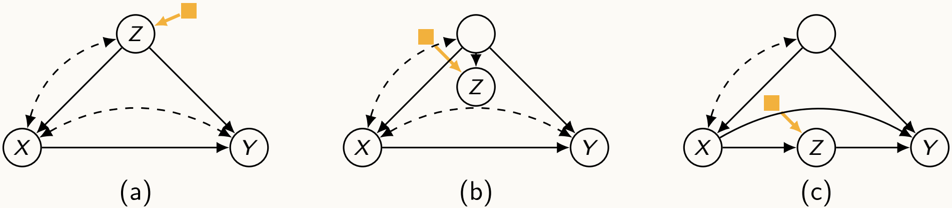
(c)

(c) Z represents **bio-marker**

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

P^* target domain & P source domain

Transportability: Formulas Depend on the Causal Story



(a) Z represents **age**

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

(b) Z represents **language skill**

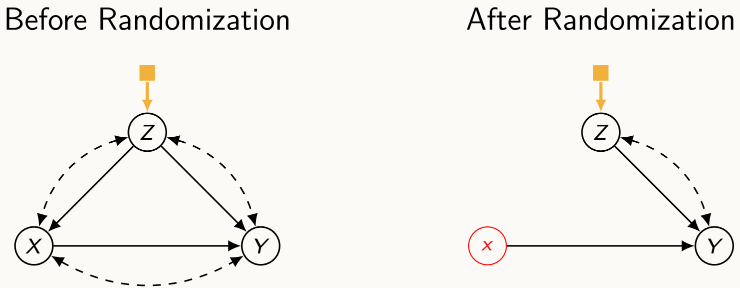
$$P^*(y|do(x)) = P(y|do(x))$$

(c) Z represents **bio-marker**

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

P^* target domain & P source domain

Is the Gold Standard Golden? (Generalizability from Trials)



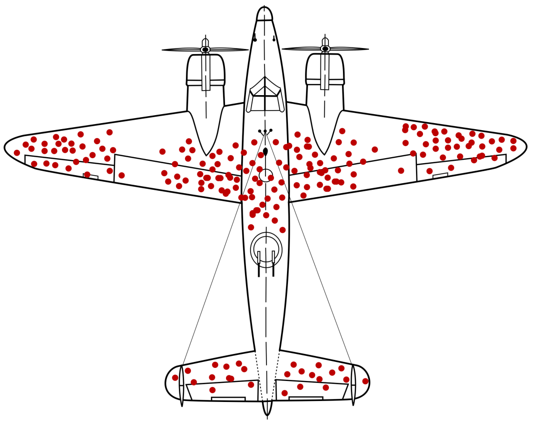
Lesson. Even if we have a perfect RCT, one still needs to exercise transportability.

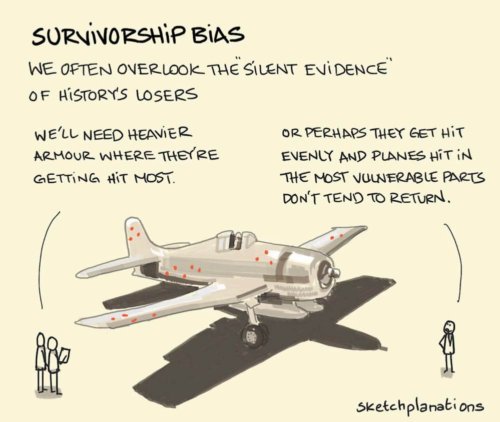
Summary for Transportability

- Non-parametric transportability can be determined provided that the problem instance is encoded in selection diagrams ($= \mathscr{G} + \blacksquare$).
- When transportability is feasible, the transport formula can be derived in polynomial time.
- The causal calculus and the corresponding transportation algorithm are complete.

Overview

Identification under Selection



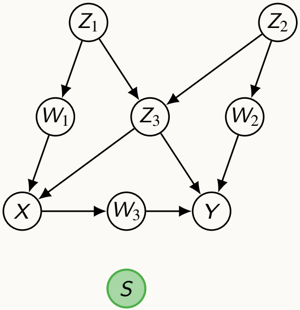


Wikipedia, https://en.wikipedia.org/wiki/Survivorship_bias, sketchplanations, https://sketchplanations.com/survivorship-bias-silent-evidence

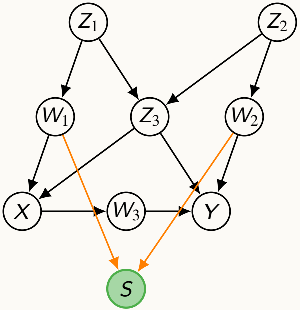
Identification under Selection

- Selection bias, caused by preferential inclusion S of samples from the data, is a major obstacle to valid **causal** and **statistical** inferences;

Without Selection Bias



With Selection Bias

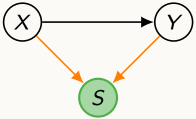


Selection Bias without External Information

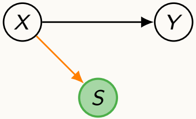
Theorem

$Q = P(y|x)$ is recoverable from selection biased data if and only if

$$(S \perp\!\!\!\perp Y \mid X).$$



$P(y|x)$ is not recoverable



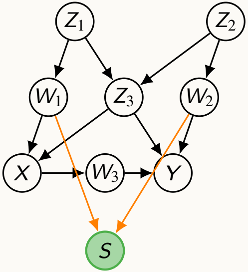
$P(y|x)$ is recoverable

Identification under Selection (with External Data)

Theorem

$P(y|x)$ is recoverable if there is a set $\mathbf{C}$ such that $(Y \perp\!\!\!\perp S \mid \mathbf{C}, X)$ holds in $\mathcal{G}$ and $P(\mathbf{C}, X)$ is estimable. Moreover,

$$P(y|x) = \sum_{\mathbf{c}} P(y|x, \mathbf{c}, S=1) P(\mathbf{c}|x)$$



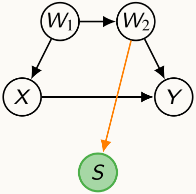
$\mathbf{C} = \{W_1, W_2\}$? Yes

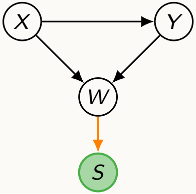
$\mathbf{C} = \{W_1, Z_1, Z_2\}$? No

$\mathbf{C} = \{W_2, Z_3\}$? Yes

Identification under Selection (with External Data)

Goal: recover a causal effect $P(y|do(x))$.





$$P(y|do(x)) = \sum_{w_2} P(y|x, w_2) P(w_2) = \sum_{w_2} P(y|x, w_2, S=1) P(w_2).$$

$$P(y|do(x)) = P(y|x) = \sum_{w} P(y|w, x) P(w|x) = \sum_{w} P(y|w, x, S=1) P(w|x).$$

Summary for Selection Bias

- Nonparametric recoverability of selection bias from causal and statistical settings can be determined provided that an augmented causal graph (w/ the selection mechanism S) is available.
- When recoverability is feasible, the estimand can be derived in **polynomial time**.
- The result is complete for pure recoverability, and sufficient for recoverability with external information.

Correa, Tian, and Bareinboim (2019) *Identification of Causal Effects in the Presence of Selection Bias*. AAAI.

Overview

Identification under Missing Data

Missing data present a challenge in many academic disciplines.

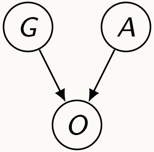
- *Sensors* do not always work reliably.
- *Respondents* do not fill out every question in the questionnaire.
- Medical *patients* are often unable to recall treatments or outcomes.

#	Age	Gender	Obesity*
1	16	F	Obese
2	15	F	N/A
3	15	M	N/A
4	14	F	Not Obese
5	13	M	Not Obese
6	15	M	Obese
7	14	F	Obese

Excerpted mainly from Mohan, K. and Pearl, J. (2021) *Graphical models for processing missing data*, JASA.

Identification under Missing Data: Example

Consider a study conducted in a school with
Age (A), **G**ender (G) and **O**besity (O).



- **Age** and **Gender** are fully observed (obtained from school records).
- **Obesity** however is corrupted by missing values due to some students not reporting their weight.

Identification under Missing Data: Proxy Variable

Modelling the **missingness process** using

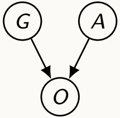
- Obesity O (true, partly-observed),
- a missingness mechanism R_O, and
- a proxy variable O^* (what's observed)

#	Age	Gender	Obesity*	R_O
1	16	F	Obese	0
2	15	F	m	1
3	15	M	m	1
4	14	F	Not Obese	0
5	13	M	Not Obese	0
6	15	M	Obese	0
7	14	F	Obese	0

$$O^* = \begin{cases} O & \text{if } R_O = 0 \\ m & \text{if } R_O = 1 \end{cases}$$

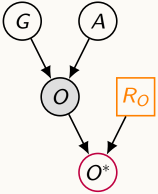
Identification under Missing Data: Reasons for Missingness

Missingness can be *caused* by random processes or can depend on other variables.



Identification under Missing Data: Reasons for Missingness

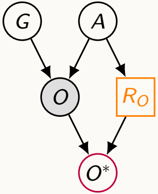
Missingness can be *caused* by random processes or can depend on other variables.



- Students *accidentally* *losing* their questionnaires.

Identification under Missing Data: Reasons for Missingness

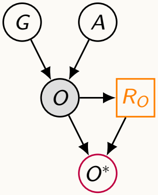
Missingness can be *caused* by random processes or can depend on other variables.



- Teenagers rebelling and not reporting their weight.

Identification under Missing Data: Reasons for Missingness

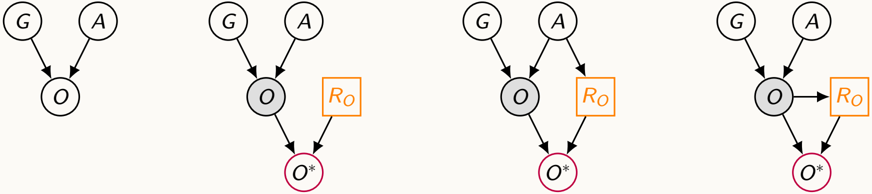
Missingness can be *caused* by random processes or can depend on other variables.



- Obese students who are embarrassed of their obesity and hence reluctant to reveal their weight.

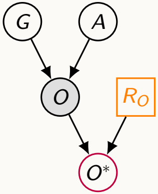
Identification under Missing Data: Reasons for Missingness

Missingness can be *caused* by random processes or can depend on other variables.

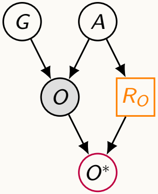


- Students *accidentally losing* their questionnaires.
- Teenagers rebelling and not reporting their weight.
- Obese students who are embarrassed of their obesity and hence reluctant to reveal their weight.

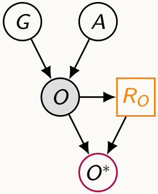
Three Categories of Missingness



MCAR



MAR



MNAR

$$O, A, G \perp\!\!\!\perp R_O$$

$$O \perp\!\!\!\perp R_O \mid A, G$$

$$O \not\!\perp\!\!\!\perp R_O \mid A, G$$

Identification under Missing Data: Example

Factorization:

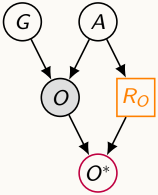
$$P(G,O,A) = P(G,O|A)P(A)$$

Transformation into observables:

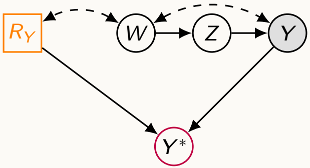
$$= P(G,O|A,R_O = 0)P(A)$$

Conversion

$$= P(G,O^*|A,R_O = 0)P(A).$$



Identification under Missing Data: Example



$$
\begin{aligned}
P(y|do(z)) &= P(y|do(z), r_y) \\
&= P(y^*|do(z), r_y) \\
&= \sum_w P(y^*|w, do(z), r_y) P(w|do(z), r_y) \\
&= \sum_w P(y^*|w, z, r_y) P(w|r_y).
\end{aligned}
$$

Summary for Part 3

Modern Identification

1. **General** Identification: combining data sets of different experimental conditions
2. **Transportability**: combining data sets from different sources ■
3. Identification under **Selection** S
4. Identification under **Missingness** $\boxed{R_O}$

Summary for Causal Inference Lecture

This lecture focused mainly on a basic causal effect identification task (SCM, do-operator, Causal Graph, Conditional Independence ...)

There are many interesting future research directions

- Causal Data Science
- Causal Discovery
- Causal Decision Making
- Causality + Machine Learning