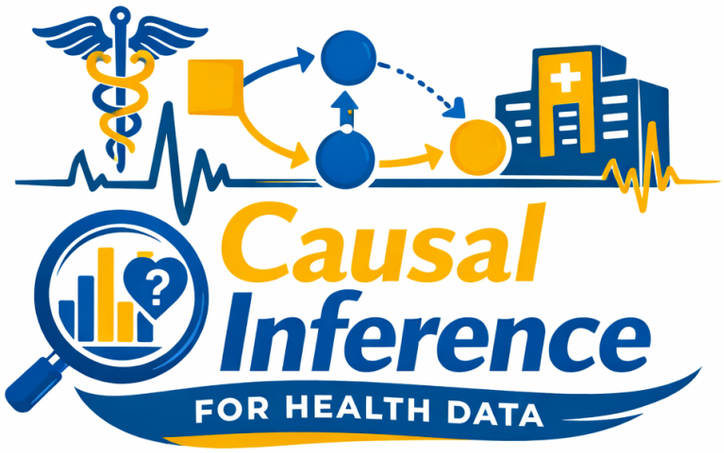
Causal Inference for Health Data

(STATS C160/C260 — Winter 2026)

Lecture 9: Counterfactuals I

Drago Plečko

UCLA College | Physical Sciences Statistics & Data Science



Rules of Do-Calculus

Theorem. The following transformations are valid for any do-distribution induced by a SCM M:

Rule 1: Adding/removing Observations

$$P(y|do(x),z,w)=P(y|do(x),w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}}$$

Rule 2: Action/observation exchange

$$P(y|do(x),do(z),w)=P(y|do(x),z,w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}}$$

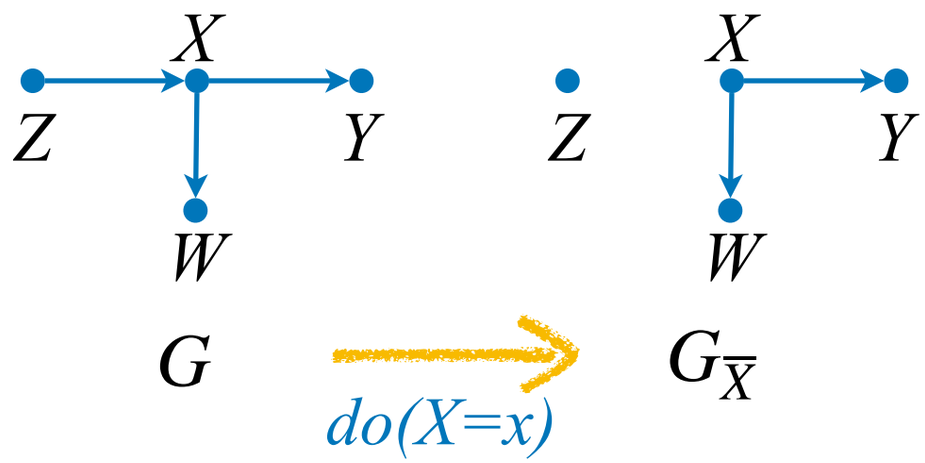
Rule 3: Adding/removing Actions

$$P(y|do(x),do(z),w)=P(y|do(x),w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\overline{Z(W)}}}$$

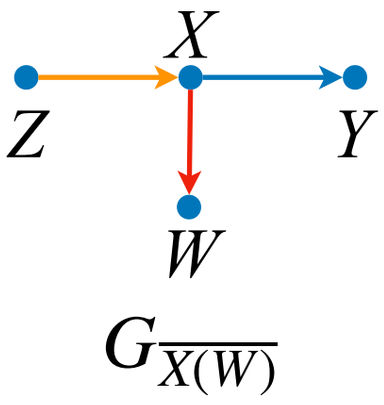
where $Z(W) = Z \setminus \text{an}(W)_{G_{\overline{X}}}$ is the subset of Z that are not ancestors of W in $G_{\overline{X}}$.

A curiosity about R3

- Regarding R3's contingency, note that the probability of Z given W may not be the same with or without intervening on X.



To witness, compare $G_{\overline{X}}$ and $G_{\overline{X(W)}}$:

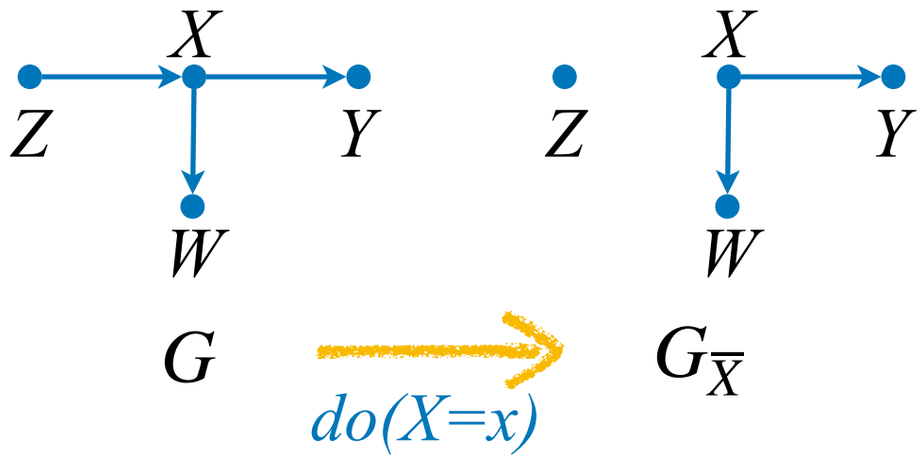


$$(Z \not\perp\!\!\!\perp X \mid W)_{G_{\overline{X(W)}}} \Longrightarrow \exists_M P(z|do(x),w) \neq P(z|w)$$

What? 

A curiosity about R3

- Verifying whether
  $P(z|do(x),w) \neq P(z|w)$ holds:

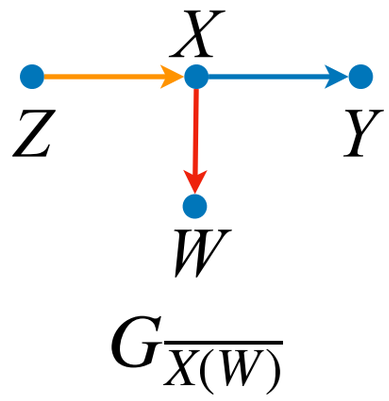


$$P(z \mid do(x), w) = \frac{P(z, w \mid do(x))}{P(w \mid do(x))}$$

$$= \frac{\sum_y P(z)P(w \mid x)P(y \mid x)}{\sum_{y,z} P(z)P(w \mid x)P(y \mid x)}$$

$$= \frac{P(z)P(w \mid x)}{P(w \mid x)}$$

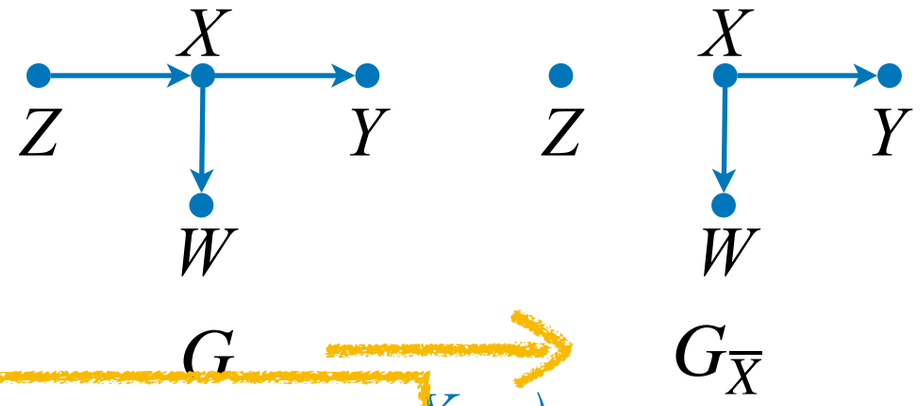
$$= P(z) \neq P(z \mid w) \qquad (Z \not\perp\!\!\!\perp X \mid W)_{G_{\overline{X(W)}}}$$



in almost any model compatible with G.

A curiosity about R3

- Verifying whether
  $P(z|do(x),w) \neq P(z|w)$ holds:

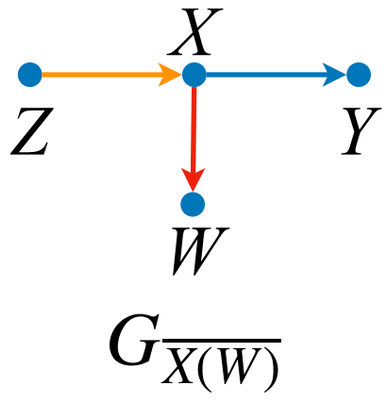


Intuitively:

(i) W is a descendant of X

(ii) in a $do(x)$-world, conditioning on W does not propagate back to $\mathrm{an}(X)$, since incoming arrows into X are removed in the $do(x)$-world

$\Longrightarrow$ (iii) when conditioning on W in normal world, conditioning on W *does* propagate to $\mathrm{an}(X)$



makes the 3rd rule more subtle!

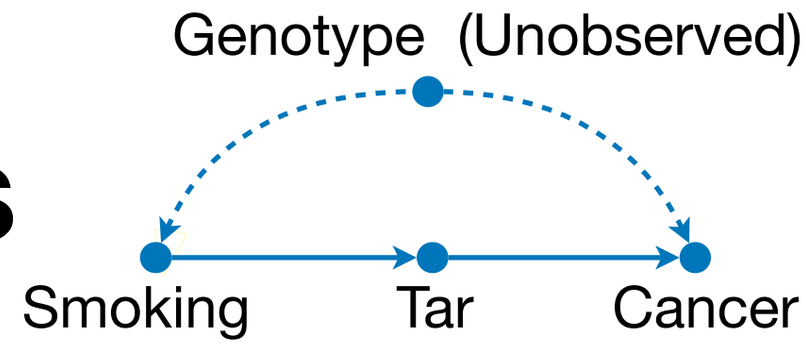
Properties of Do-Calculus

Theorem (soundness & completeness of do-calculus for interventional ID from obs. data).

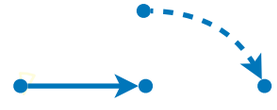
The quantity $Q = P(y|do(x))$ is identifiable from $P(v)$ and G if and only if there exists a sequence of application of the rules of do-calculus and the probability axioms that reduces Q into a do-free expression.

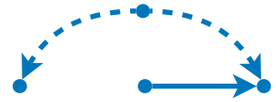
Syntactic goal: Re-express original Q without do()!

Derivation in Do-Calculus



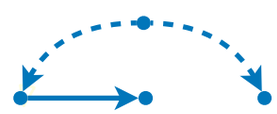
$$P(c \mid do(s)) = \sum_t P(c \mid do(s), t) P(t \mid do(s))$$ Probability Axioms

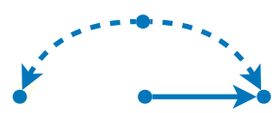
$$= \sum_t P(c \mid do(s), do(t)) P(t \mid do(s))$$ Rule 2 

$$= \sum_t P(c \mid do(s), do(t)) P(t \mid s)$$ Rule 2 

$$= \sum_t P(c \mid do(t)) P(t \mid s)$$ Rule 3 

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

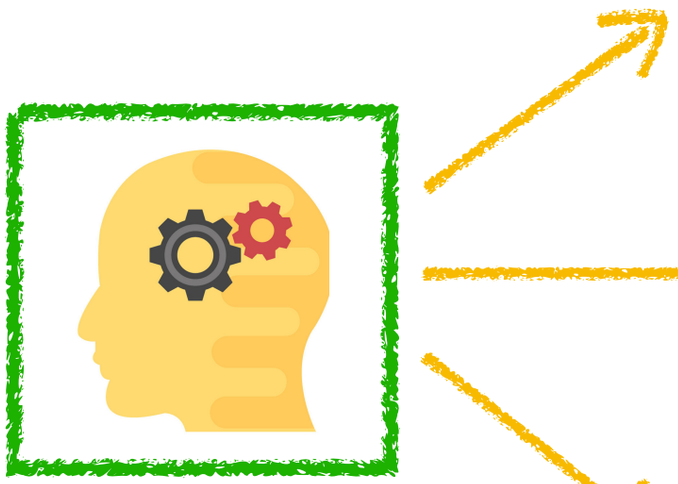
$$= \sum_t \sum_{s'} P(c \mid t, s') P(s' \mid do(t)) P(t \mid s)$$ Rule 2 

$$= \sum_t \sum_{s'} P(c \mid t, s') P(s') P(t \mid s)$$ Rule 3 

Pearl’s Causal Hierarchy—

Layer 3: Counterfactuals

SCM → Pearl's Causal Hierarchy

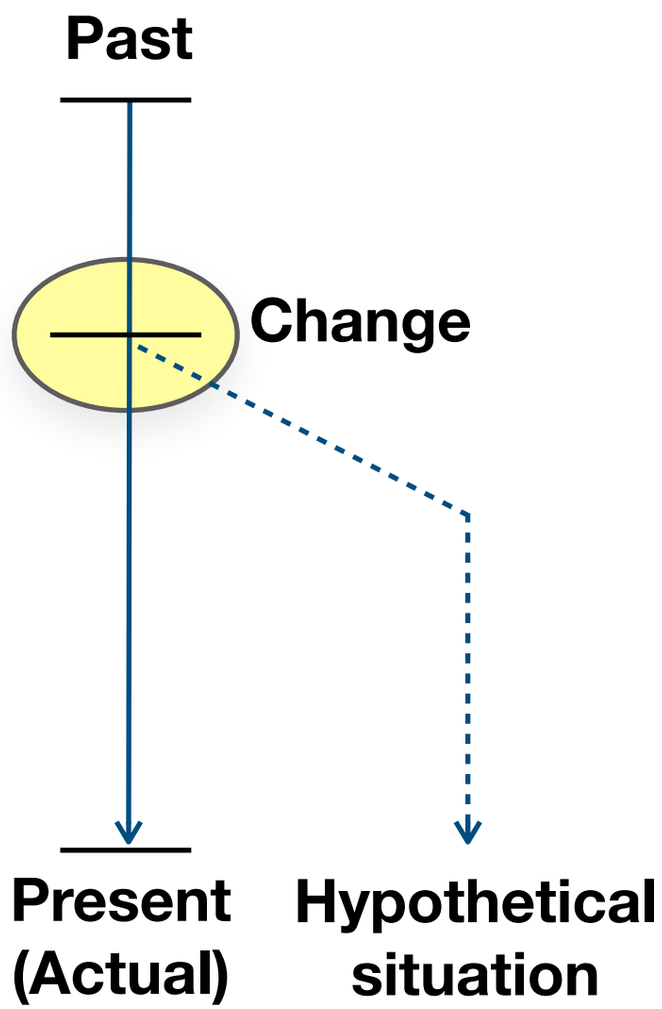


SCM M

Level (Symbol)		Typical Activity
1	Association $P(y \mid x)$	Seeing ML - (Un)Supervised
2	Intervention $P(y \mid do(x), c)$	Doing ML - Reinforcement
3	Counterfactual $P(y_x \mid x', y')$	Imagining, Retrospection

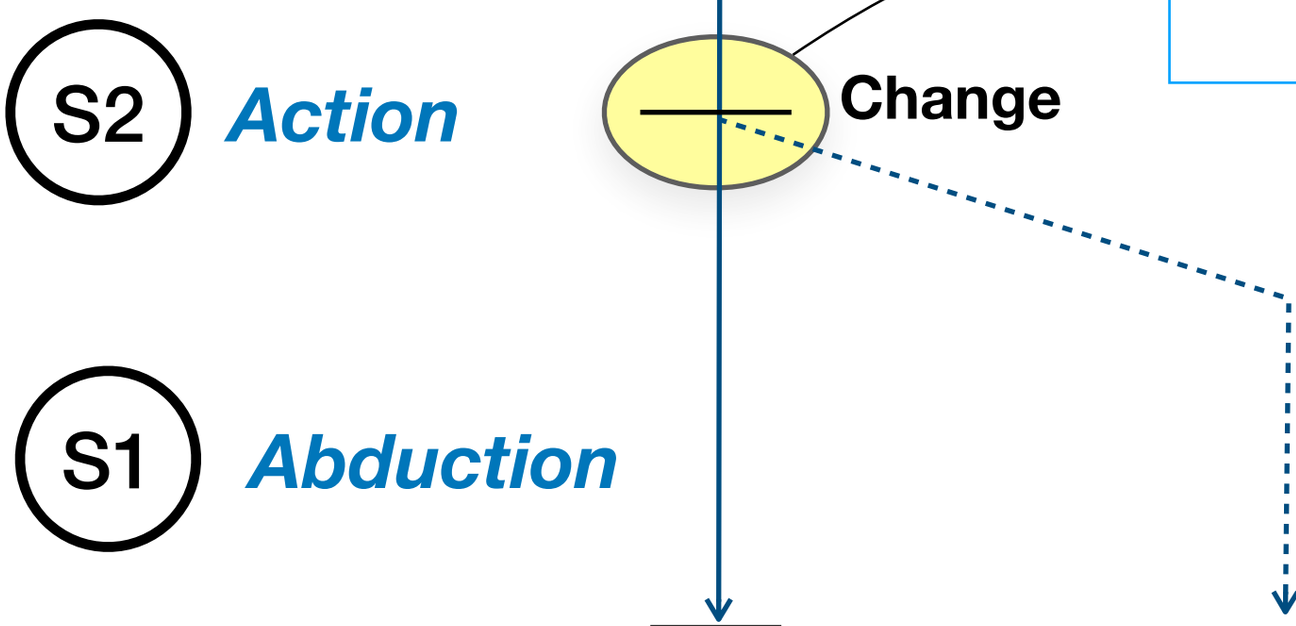
Layer 3 — Semantics

- Layer 3 of the causal hierarchy allows operationalizing the notion of “imagination,” and the closely related activities of retrospection and introspection.
- This entails thinking about alternative ways the world could be, including ways that might conflict with the current state of the world.



Layer 3 — Abduction, Action, Prediction (AAP)

$Z = z_1$



$X = x_1$

$Y = y_1$

we have a way to encode change:
e.g., $X \leftarrow x_0$ mechanism replacement

we can thus compute what would happen under $X \leftarrow x_0$, e.g., the value of Y

S3 ***Prediction***

We need new notation!
$E[Y \mid do(x_0), x_1]$?

$E[Y_{X=x_0} \mid X = x_1]$!

Note: knowing that $X = x_1$ tells us something about the individual

Layer 3 — AAP Example

- Query: $E[Y_{X=x_0} \mid X = x_1]$

S1 ***Abduction***

Computing P(U | evidence):

$$P(U_z, U_x \mid X = 1)$$

$$= \frac{P(U_z, U_x, X = 1)}{P(X = 1)}$$

$0.2 \times 0.7 + 0.8 \times 0.1$

$$= \frac{P(U_z)P(U_x)P(X = 1 \mid U_z, U_x)}{P(X = 1)}$$

$$\mathscr{M}$$

$$Z \leftarrow U_z$$

$$X \leftarrow (U_{x_1} \wedge \neg Z) \vee (U_{x_2} \wedge Z)$$

$$Y \leftarrow 1(0.8 - 0.3Z + 0.2XZ - 0.1X(1 - Z) > U_y)$$

$$U_z \sim B(0.2),\ U_{x_1} \sim B(0.1),$$

$$U_{x_2} \sim B(0.7),\ U_Y \sim \textbf{Unif}[0,1]$$

Distribution $P(u \mid x_1)$!

U_z	U_x	P(U_z, U_x, x_1)	U_z	U_x	P(U_z, U_x, x_1)
0	(0, 0)	0	1	(0, 0)	0
0	(0, 1)	0	1	(0, 1)	0.2 * 0.9 * 0.7
0	(1, 0)	0.8 * 0.1 * 0.3	1	(1, 0)	0
0	(1, 1)	0.8 * 0.1 * 0.7	1	(1, 1)	0.2 * 0.1 * 0.7

Layer 3 — AAP Example

S2 ***Action***

S3 ***Prediction (in M_{x_0})***

$$E[Y_{x_0} \mid X = 1]$$

$$= \sum_u E[Y_{x_0} \mid u] P(u \mid x_1)$$

$$= \frac{32}{220} \times 0.8 + \frac{48}{220} \times 0.8 + \frac{126}{220} \times 0.5 + \frac{14}{220} \times 0.5$$

$U_z = 0, U_x = (1,0)$ $U_z = 0, U_x = (1,1)$ $U_z = 1, U_x = (0,1)$ $U_z = 0, U_x = (1,1)$

$$\approx 0.61$$

$$\mathcal{M}_{x_0}$$

$$Z \leftarrow U_z$$

$$X \leftarrow x_0$$

$$Y \leftarrow 1(0.8 - 0.3Z + 0.2XZ - 0.1X(1 - Z) > U_y)$$

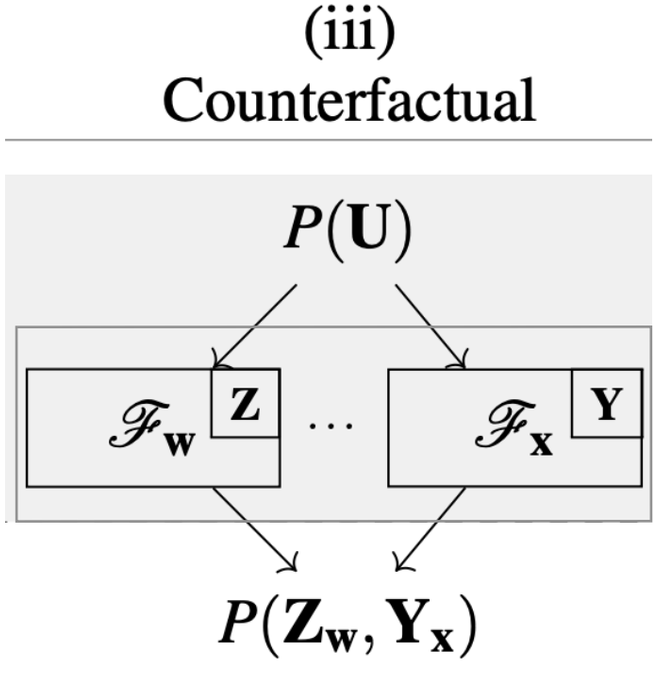
$$U_z \sim B(0.2),\ U_{x_1} \sim B(0.1),$$

$$U_{x_2} \sim B(0.7),\ U_Y \sim \textbf{Unif}[0,1]$$

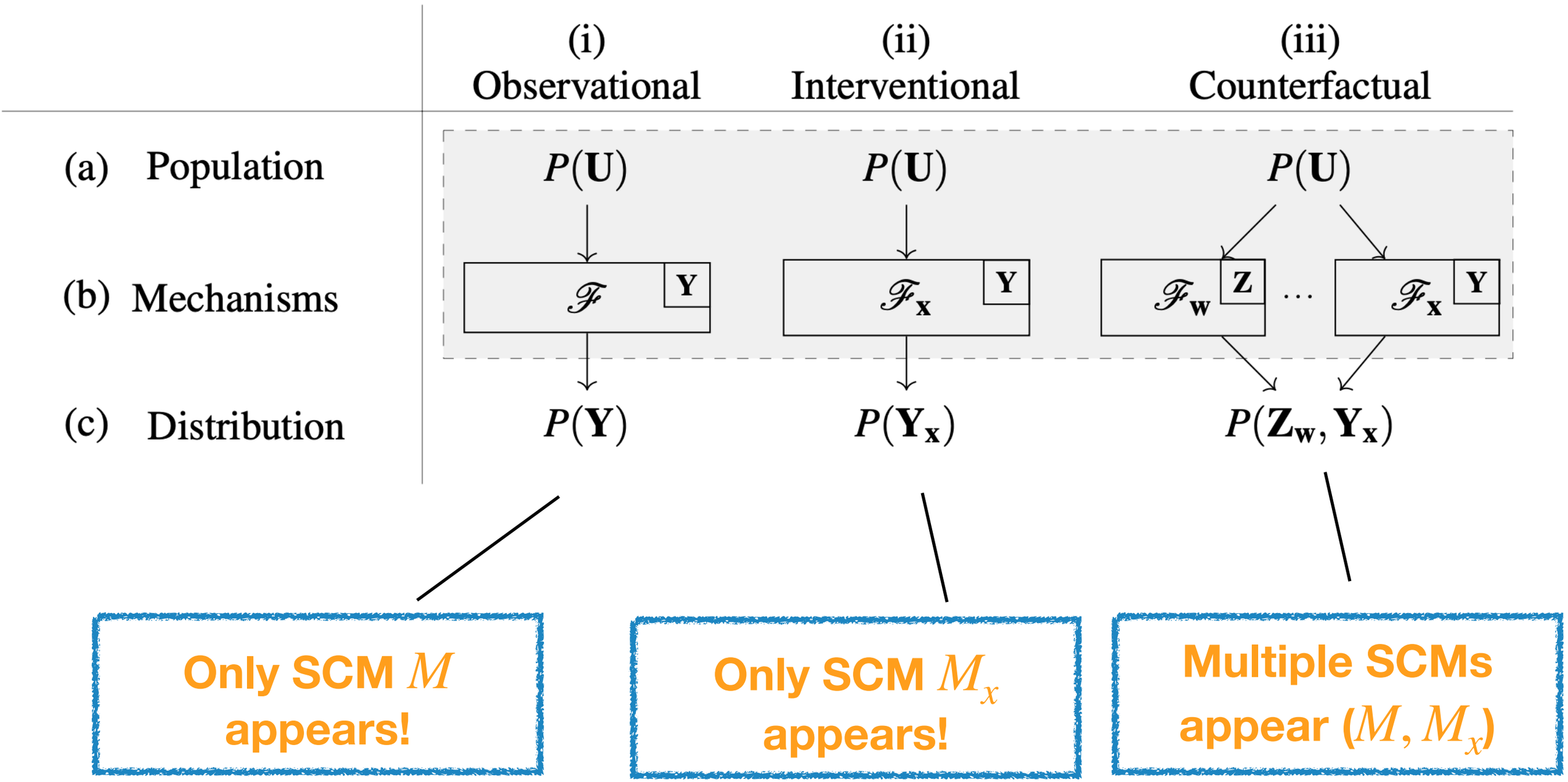
Layer 3 — Valuation

Definition 2.2.6 (Layer 3 Valuation). An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ induces a family of joint distributions over counterfactual events $Y_x, \ldots, Z_w$, for any $Y, Z, \ldots, X, W \subseteq \mathbf{V}$:

$$P^{\mathcal{M}}(\mathbf{y_x}, \ldots, \mathbf{z_w}) = \sum_{\mathbf{u}} \mathbf{1}\Big(\mathbf{Y_x}(\mathbf{u}) = \mathbf{y}, \quad \ldots, \mathbf{Z_w}(\mathbf{u}) = \mathbf{z} \Big) P(\mathbf{u}) .$$



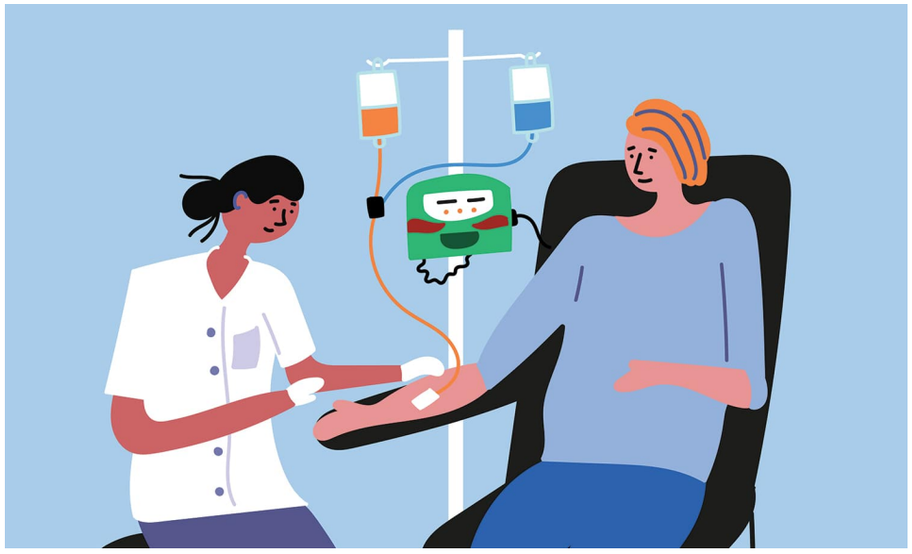
Layer 3 vs. the other layers



II. Why Use Counterfactuals (Motivation)

Example 5.1. Oncologist and Chemotherapy

- An oncologist needs to decide whether to treat her patients, suffering from breast cancer, with chemotherapy,
- For this purpose, she wants to analyze the data collected in her hospital,
- Let X represent the treatment, where $X = 0$ encodes no chemotherapy, and $X = 1$ for chemotherapy.



- Let Y represent the patient's health, where $Y = 0$ means that the patient has died, and $Y = 1$ that the patient is still alive two years after taking the treatment; Z represents how advanced the cancer is ($Z = 0$ for Stage I-II, $Z = 1$ for Stage III-IV).
- Using the data, how should the physician decide to treat her patients?

(1) Observational distribution

- She extracts the preliminary data from the hospital EHR, and gets the observational distribution P(v),
- She then decides to compute the metric

$$E[Y \mid X = 1] - E[Y \mid X = 0],$$

- Based on the table she finds that

$$E[Y \mid X = 1] = 70\,\%$$
$$E[Y \mid X = 0] = 77.7\,\%\,,$$

P(v)	Z	X	Y
0.144	0	0	0
0.03	1	0	0
0.024	0	1	0
0.576	0	0	1
0.042	1	1	0
0.03	1	0	1
0.056	0	1	1
0.098	1	1	1

- The results imply that treated patients are less likely to survive. At first, they seem puzzling, but…
- Recalling the class in causal inference, she realizes that this quantity is not informative, since data it does not adjusted for confounding bias — more sick patients (Stage III-IV) are more likely to receive chemotherapy in the first place.

(2) Causal Effect Computation

- She then decides to use the adjustment formula based on Z, based on the back-door criterion, $E[y \mid do(x)] = \sum_z E[y \mid x, z]P(z)$
- From BD-adjust, one can compute

$$E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)],$$

- Based on the table she finds that

$$E[Y \mid do(X = 1)] = 70\,\%$$
$$E[Y \mid do(X = 0)] = 74\,\%\,,$$

P(v)	Z	X	Y
0.144	0	0	0
0.03	1	0	0
0.024	0	1	0
0.576	0	0	1
0.042	1	1	0
0.03	1	0	1
0.056	0	1	1
0.098	1	1	1

- She finds that even after adjustment, the causal effect of chemotherapy results in worse survival. Does this imply the treatment is futile, and should be abandoned entirely?
- How do make sense of this finding?

(3) Oracle — Effect on Treated

- Let's look at the SCM generating the data, on the left side,
- The physician wants to understand whether the treatment has helped the group of patients who were actually given it,

$$\mathscr{M}$$

$$Z \leftarrow \textbf{Bern}(0.2)$$

$$X \leftarrow \textbf{Bern}(0.1 + 0.6Z)$$

$$Y \leftarrow \textbf{Bern}(0.8 - 0.3Z + 0.2XZ - 0.1X(1 - Z))$$

Would a treated patient (X=1) survive without treatment?

- This query has special meaning in the language of layer 3 (L_3) of the PCH, and is written as the counterfactual statement:

$$P(Y_{X=0} = 1 \mid X = 1),$$

This expression is called the *effect of treatment on the treated (ETT)*.

(3) Oracle — Effect on Treated

- What is the probability of survival under a certain treatment, given that the patient's natural treatment regime?
- From the SCM, we can compute:

$$P(Y_{X=1} = 1 \mid X = 1) = 70\,\% \qquad P(Y_{X=0} = 1 \mid X = 1) = 61\,\%$$

- Therefore, among those patients who are treated ($X = 1$ behind the conditioning bar), being treated ($X = 1$) is actually be better since

$$P(Y_{X=1} = 1 \mid X = 1) > P(Y_{X=0} = 1 \mid X = 1),$$

- For those who are not treated, we can show the opposite:

$$P(Y_{X=1} = 1 \mid X = 0) < P(Y_{X=0} = 1 \mid X = 0),$$

- How is this possible?

(3) Oracle — Effect on Treated

- We first look at the mechanism for Y:

$$Y \leftarrow \textbf{Bern}(0.8 - 0.3Z + 0.2XZ - 0.1X(1 - Z))$$

- baseline survival of 80%
- 30% survival decrease for late stage
- for late stage, chemo increases survival 20%
- for early stage, chemo decreases survival 10%

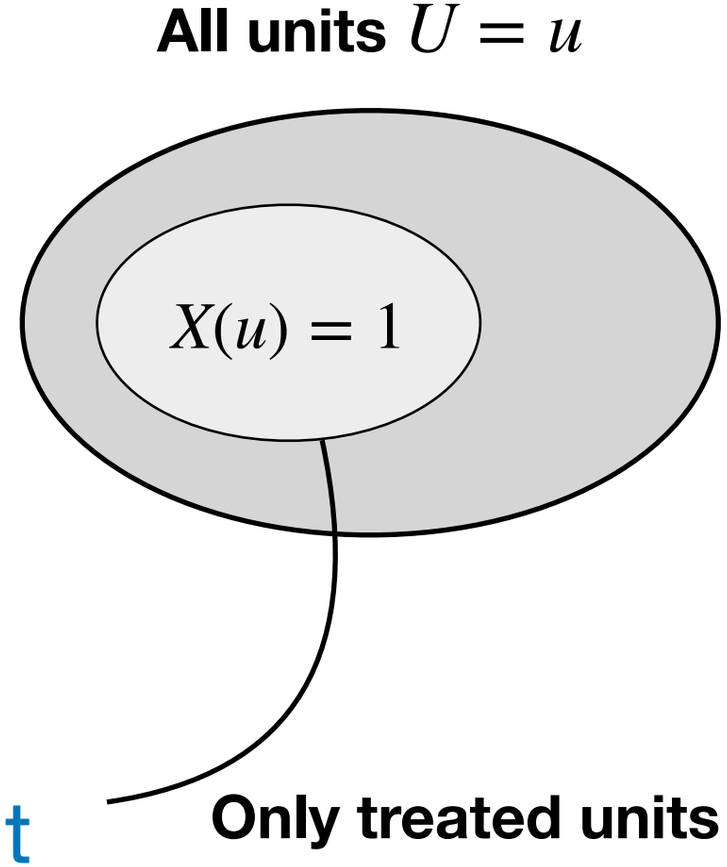
- Therefore, treatment is harmful for early stage, but helpful for late,
- Furthermore, the mechanism for X shows us that

$$X \leftarrow \textbf{Bern}(0.1 + 0.6Z)$$

among treated (X=1), there are more late-stage patients than usual!

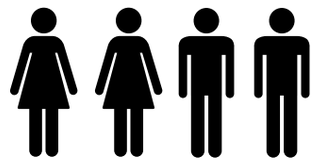
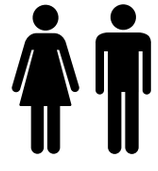
Example — Takeaways

- Causal effects may exhibit *heterogeneity*, meaning that different parts of the population are affected differently,
- The average causal effect
  $E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)]$
  is an average over the entire, possibly heterogenous, population of individuals,
- To better understand heterogeneity, we may use conditional effects, which sometimes require Layer 3 of PCH,
- This is the case for the Effect of Treatment on the Treated (ETT).



(1) Use of Counterfactuals: Granularity

Population Axis

population $P(u)$		entire population
x-subset $P(u \mid x)$		treated population
z, x-subset $P(u \mid x, z)$		treated, Stage I-II
z, x, w-subset $P(u \mid x, z, w)$		treated, Stage I-II, with side-effects
v-subset $P(u \mid v)$		treated, Stage I-II, side-effects, survived
unit-level u		Catherine, treated, Stage I-II, side-effects, survived

Example. Semaglutide Effects on Diabetes

- A diabetologist is treating his patients with semaglutide, to stabilize their HbA1c values,
- However, he realized that there are different causal pathways through which semaglutide may affect HbA1c levels: (1) through weight-loss and associated insulin sensitivity, and (2) direct GLP-1 receptor effects,

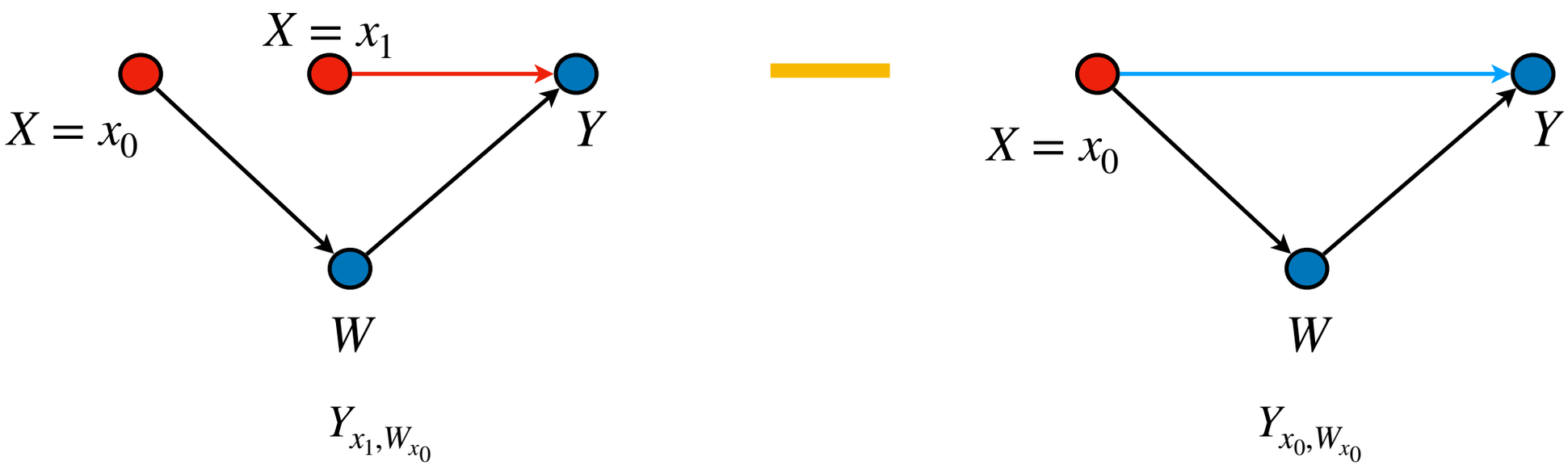
- The doctor is interested in how strong each of these pathways is, and he wonders if he can obtain an answer from data:
- $X = 1$ denotes semaglutide treatment, $X = 0$ standard treatment, $Y = 1$ denote normal HbA1c levels ($Y = 0$ otherwise); $W = 1$ for weight loss ($W = 0$ otherwise).

Gedankenexperiment (NDE)

- For investigating this, the doctor performs following thought experiments,

- For an untreated individual ($X = x_0$), how would his HbA1c level (Y) change had he been treated ($X = x_1$), while keeping the weight loss unchanged (at the natural level $X = x_0$)?

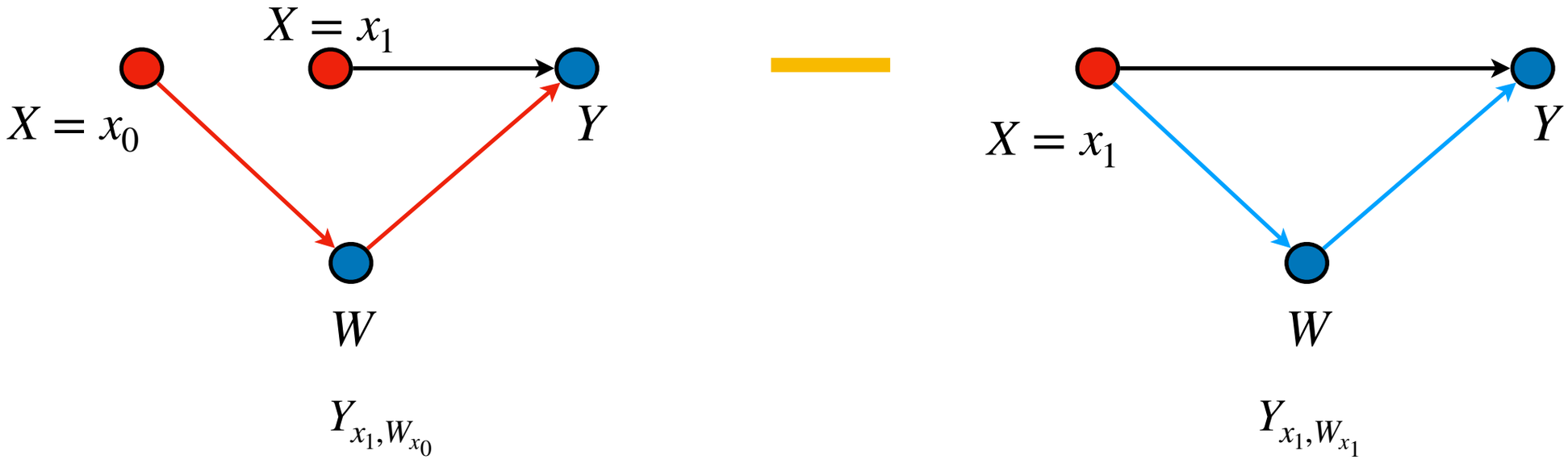
$$\mathbf{NDE}_{x_0,x_1}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_0,W_{x_0}})$$



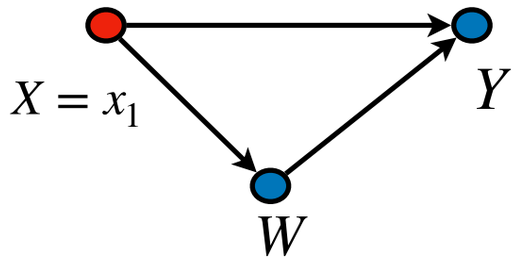
Gedankenexperiment (NIE)

- For a treated individual ($X = x_1$), how would their HbA1c level (Y) change had they not been treated ($X = x_0$), while keeping treatment unchanged along the direct causal pathway (at the natural level $X = x_1$)?

$$\textbf{NIE}_{x_1,x_0}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_1,W_{x_1}})$$

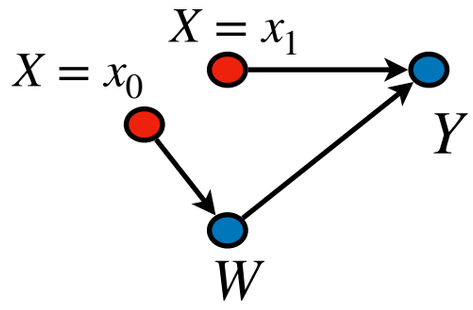


ATE Decomposition



$Y \mid do(X = x_1)$

$-$



$Y_{x_1, W_{x_0}}$

$-\text{NIE}_{x_1, x_0}(y)$

Theorem. *The average total effect can be decomposed into its direct and indirect parts:*

$$ATE_{x_0,x_1}(y) = NDE_{x_0,x_1}(y) - NIE_{x_1,x_0}(y)\,.$$

(Y)

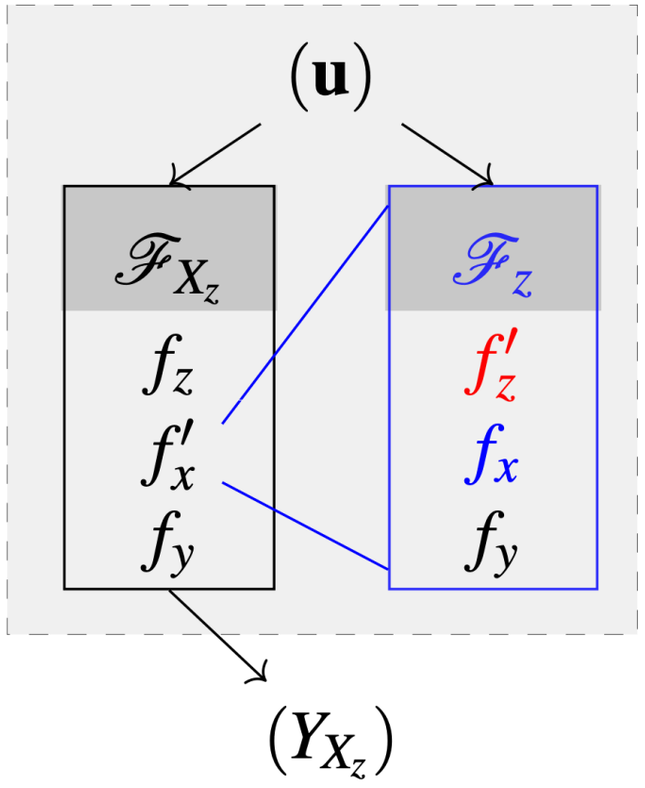
$Y_{x_1, W_{x_0}}$

$Y \mid do(X = x_0)$

$\text{NDE}_{x_0, x_1}(y)$

Nested Counterfactuals

- We used the notion of *nested counterfactuals*.
- Y_x refers to Y under intervention $do(X = x)$, that is, when X is fixed to a constant value x.
- Here, we consider an intervention $do(X \leftarrow X_z)$, where X is not fixed to a constant but is supposed to behave as another counterfactual variable X_z.

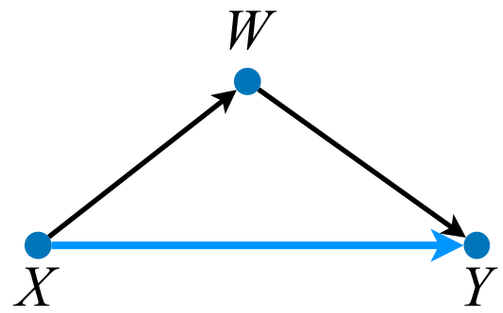


- For such intervention, we first need to evaluate X_z (from $\mathcal{M}_z$), and then consider a model $\mathcal{M}_{X_z}$ where X is given the value dictated by X_z.

Natural Direct Effects

- The Natural Direct Effect (NDE) is defined as

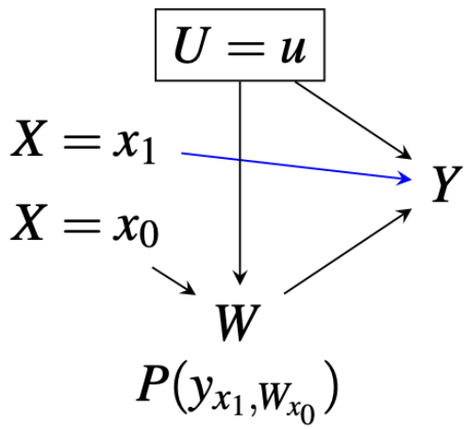
$$\boxed{\text{NDE}_{x_0,x_1}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_0,W_{x_0}})} \quad (5.101)$$



- $Y_{x_1,W_{x_0}}$ refers to the outcome Y under $X = x_1$ and $W = W_{x_0}$, the value that W would attain had X been x_0.

- Y_{x_0} represents a baseline where Y perceives $X = x_0$ in all causal paths. It is also equivalent to $Y_{x_0 W_{x_0}}$, that is, a situation where W also gets $X = x_0$.

- Taking the difference of those two quantities keeps the path $X \rightarrow W \rightarrow Y$ constant while changing the level of X from x_0 to x_1 in the path $X \rightarrow Y$, effectively measuring the direct impact of X on Y.



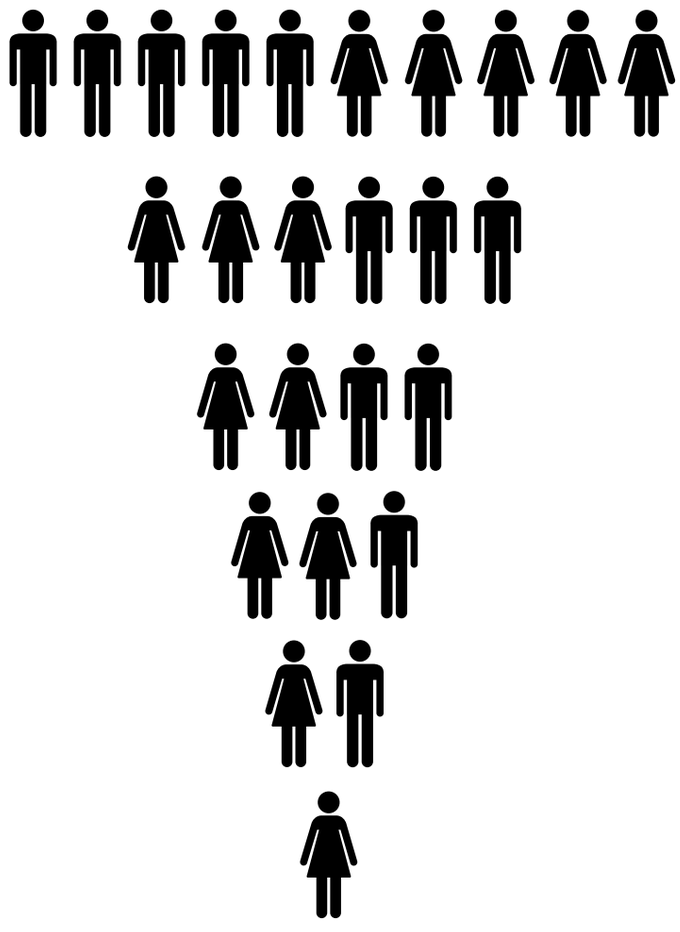
Summary: Using of Counterfactuals

Total Effect — TE

$$E[Y \mid do(x_1)] - E[Y \mid do(x_0)]$$

We spent the first 8 lectures building foundations to understand this quantity

Ext. 1: Granularity



Ext. 2: Mechanisms

DE

IE