TECHNISCHE UNIVERSITÄT DARMSTADT

AIML Lab

Winter Semester 2025/26 Lecture

Causality for AI & ML

“do-calculus”

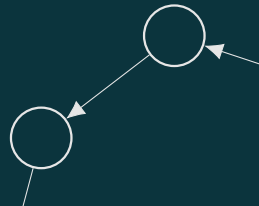
Prof. Dr. Kristian Kersting

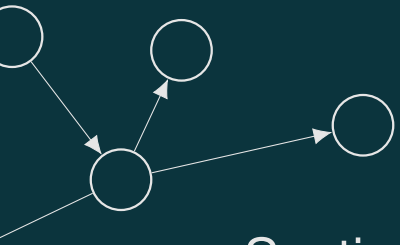
Moritz Willig — Today’s speaker

Tim Woydt

Florian Busch

Matej Zečević

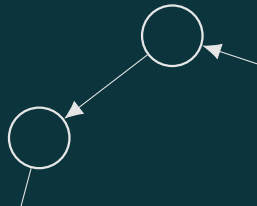




Section

0

Recap: SCM



Structural Causal Models

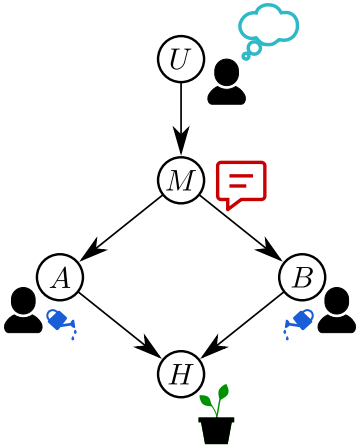
A **Structural Causal Model** (SCM) is a tuple $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, \mathcal{P}_{\mathbf{U}})$.

- $\mathbf{V}$ Set of Endogenous Variables.
- $\mathbf{U}$ Set of Exogenous Variables.
- $\mathbf{F}$ Structural Equations; $x_i := f_i(\text{pa}(x_i))$.
- $\mathcal{P}_{\mathbf{U}}$ Distribution of Exogenous Variables.

– SCM induce a *directed acyclic graph* (DAG) $\mathcal{G}$ with vertices $\mathbf{X}$ and edges $\text{pa}(x_i) \to x_i$.
  – $\mathbf{X}$ is the set of all variables: $\mathbf{X} = \mathbf{V} \cup \mathbf{U}$
  – $\text{pa}(x_i), \text{ch}(x_i), \text{an}(x_i), \text{de}(x_i)$ denote the parents, (direct) children, ancestors and descendants of x_i.

Watering Plant SCM

$$\mathcal{M} = \begin{cases} \mathbf{V} & = \{M, A, B, H \in \mathbb{B}\} \\ \mathbf{U} & = \{U \in \mathbb{B}\} \\ \mathbf{F} & = \begin{cases} f_M & := U \\ f_A & := M \\ f_B & := M \\ f_H & := A \vee B \end{cases} \\ \mathcal{P}_{\mathbf{U}} & = \{U = \text{Bernoulli}(0.5)\} \end{cases}$$

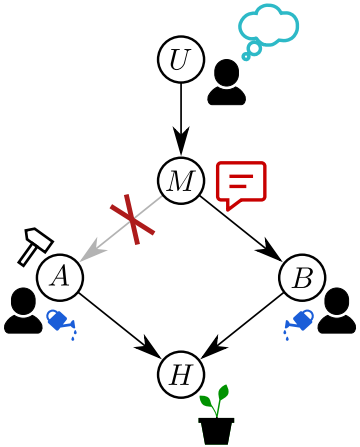


Interventions

An intervention $do(A = a)$ replaces the structural causal equation f_A, with the constant assignment $f_A := a$.

$$\mathcal{M} = \begin{cases} \mathbf{V} & = \{M, A, B, H \in \mathbb{B}\} \\ \mathbf{U} & = \{U \in \mathbb{B}\} \\ \mathbf{F} & = \begin{cases} f_M & := U \\ f_A & := a \\ f_B & := M \\ f_H & := A \vee B \end{cases} \\ \mathcal{P}_{\mathbf{U}} & = \{U = \text{Bernoulli}(0.5)\} \end{cases}$$

In the causal graph this corresponds to cutting all edges into A.



Task of Causal Inference

Task of Causal Inference: Can we answer a causal **query** $P(y|\text{do(x)})$, given the **causal graph** $\mathcal{G}$ and **observational data x**?

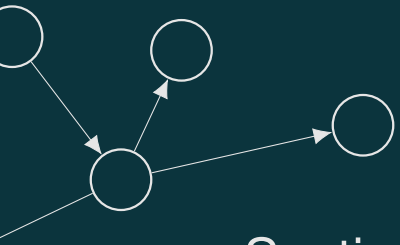
Is there an purely observational **estimand**?

"What is the probability of the outcome $Y = y$ if I do set $X = x$?"

Average Treatment Effect (for binary outcome scenarios):

$$\text{ATE} = \mathbb{E}[P(y|do(X = 1)) - P(y|do(X = 0))]$$

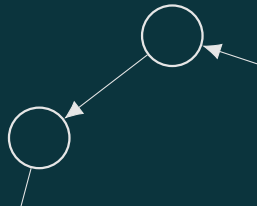
The ATE is the *expected difference in outcome* that would result from, e.g., treating an individual compared to not treating them.



Section

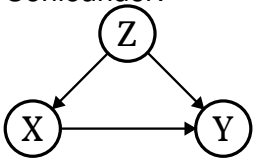
1

Back-Door Adjustment



Back-Door Adjustment

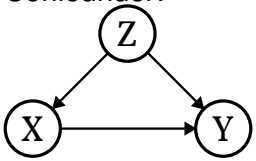
Confounder:



Z is biasing X and Y. We need to adjust for the effects of Z!

Back-Door Adjustment

Confounder:



Z is biasing X and Y. We need to adjust for the effects of Z!

The shown graph leads to the most simple application of back-door adjustment.

Back-Door Criterion

Back-Door Criterion

Consider a causal graph $\mathcal{G}$ and a causal query $P(y|do(x))$. A set of variables $\mathbf{Z}$ satisfies the back-door criterion iff:

1. No node in $\mathbf{Z}$ is a descendant of X.
2. $\mathbf{Z}$ blocks every path between X and Y that contains an arrow into X.

Back-Door Criterion

Back-Door Criterion

Consider a causal graph $\mathcal{G}$ and a causal query $P(y|do(x))$. A set of variables $\mathbf{Z}$ satisfies the back-door criterion iff:

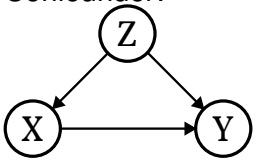
1. No node in $\mathbf{Z}$ is a descendant of X.
2. $\mathbf{Z}$ blocks every path between X and Y that contains an arrow into X.

If $\mathbf{Z}$ satisfies the back-door criterion relative to X and Y in $\mathcal{G}$ and if $P(x, z) > 0$, then the causal query is identifiable by:

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

Back-Door Adjustment I

Confounder:

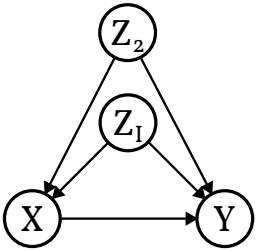


Adjustment Set $\mathbf{Z} = \{Z\}$

$$P(Y|do(X = x)) = \sum_{z \in \mathcal{Z}} P(Y|X = x, Z = z)P(Z = z)$$

Back-Door Adjustment II

Adjustment Set: $\mathbf{Z} = \{Z_1, Z_2\}$



$$P(Y|do(X=x)) =$$

$$\sum_{\mathbf{z}\in\mathcal{Z}} P(Y|X=x,\mathbf{Z}=\mathbf{z})P(\mathbf{Z}=\mathbf{z})$$

$$\sum_{z_1\in\mathcal{Z}_1}\sum_{z_2\in\mathcal{Z}_2} P(Y|X=x,Z_1=z_1,Z_2=z_2)P(Z_1=z_1,Z_2=z_2)$$

$$\sum_{z_1\in\mathcal{Z}_1}\sum_{z_2\in\mathcal{Z}_2} P(Y|X=x,Z_1=z_1,Z_2=z_2)P(Z_1=z_1)P(Z_2=z_2)$$

Back-Door Adjustment III

Adjustment Set: $\mathbf{Z} = ?$

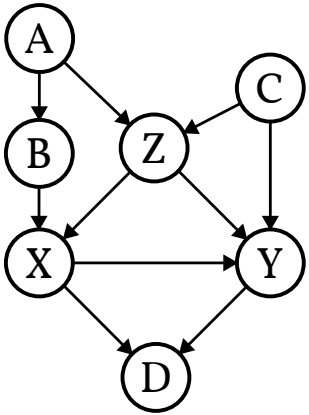
Remember:

1. No node in $\mathbf{Z}$ is a descendant of X.
2. $\mathbf{Z}$ blocks every path between X and Y that contains an arrow into X.

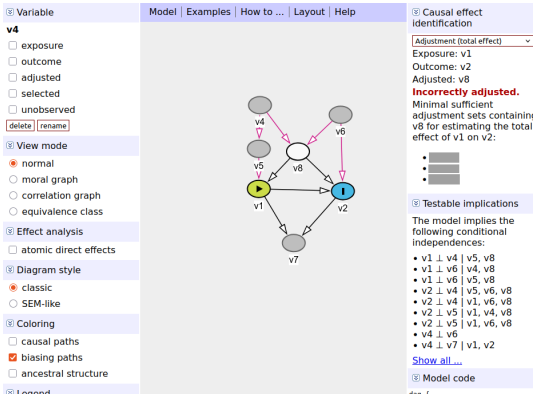
General rules of d-separation apply!

Note: The exists multiple minimal adjustment sets.

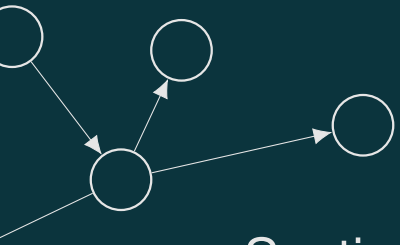
Excercise: Try to (1) find all adjustment sets, and additionally (2) all minimal adjustment sets.



DAGitty

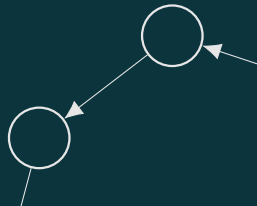


DAGitty: Very helpful library and online tool!
Available online: https://www.dagitty.net/dags.html



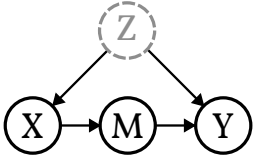
Section 2

Front-Door Adjustment



Front-Door Adjustment

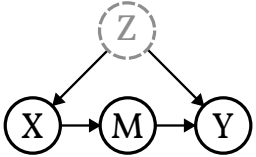
Graph with an unobserved confounder (marked gray/dashed):



The confounder Z is unobserved. We can no longer condition on it!

Front-Door Adjustment

Graph with an unobserved confounder (marked gray/dashed):

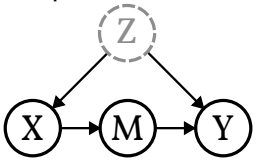


The confounder Z is unobserved. We can no longer condition on it!

Idea: The variables in the pairs (X, M) and (M, Y) are not jointly confounded by Z. Estimate the individual terms and piece together the full causal effect!

Front-Door Adjustment

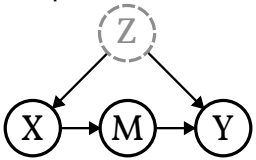
Graph with an unobserved confounder:



First: Estimate $P(m|do(x))$.
There are no open backdoor paths from X to M, so:
$P(m|do(x)) = P(m|x)$.

Front-Door Adjustment

Graph with an unobserved confounder:

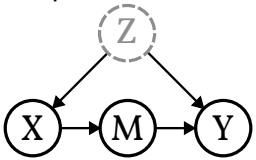


First: Estimate $P(m|do(x))$.
There are no open backdoor paths from X to M, so:
$P(m|do(x)) = P(m|x)$.

Second: Estimate $P(y|do(m))$.
There is an open backdoor path from M to Y!
But we can block it by conditioning on X:
$P(y|do(m)) = \sum_{x' \in \mathcal{X}} P(y|m, x')P(x')$.

Front-Door Adjustment

Graph with an unobserved confounder:



Third: Join the two parts together and marginalize out M.

$$P(y|do(x)) = \sum_{m \in \mathcal{M}} P(m|do(x))P(y|do(m))$$

$$\sum_{m \in \mathcal{M}} P(m|x) \sum_{x' \in \mathcal{X}} P(y|m, x')P(x')$$

Front-Door Criterion

Font-Door Criterion

Consider a causal graph $\mathcal{G}$ and a causal query $P(y|do(x))$. A set of variables $\mathbf{Z}$ satisfies the front-door criterion iff:

1. $\mathbf{Z}$ intercepts all directed paths from X to Y.
2. There is no back-door path from X to $\mathbf{Z}$;
3. All back-door paths from $\mathbf{Z}$ to Y are blocked by X.

Front-Door Criterion

Font-Door Criterion

Consider a causal graph $\mathcal{G}$ and a causal query $P(y|do(x))$. A set of variables $\mathbf{Z}$ satisfies the front-door criterion iff:

1. $\mathbf{Z}$ intercepts all directed paths from X to Y.
2. There is no back-door path from X to $\mathbf{Z}$;
3. All back-door paths from $\mathbf{Z}$ to Y are blocked by X.

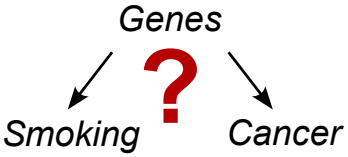
If $\mathbf{Z}$ satisfies the front-door criterion relative to X and Y in $\mathcal{G}$ and if $P(x, z) > 0$, then then causal query $P(y|do(x))$ is identifiable by:

$$P(y|do(x)) = \sum_{m \in \mathcal{M}} P(m|x) \sum_{x' \in \mathcal{X}} P(y|m, x')P(x')$$

Recap to Lecture 1

“Does Smoking cause Cancer?”

The tobacco industry claimed for a long time that genes might be a confounding factor between adopting smoking and cancer.

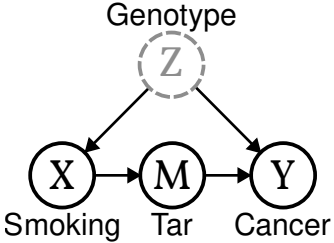


Recap to Lecture 1

"Does Smoking cause Cancer?"

The tobacco industry claimed for a long time that genes might be a confounding factor between adopting smoking and cancer.

Idea: Measure tar in lungs as a mediator.



Soundness and Completeness

If there exists an adjustment set $\mathbf{Z}$ that satisfies the back-door criterion for $P(y|do(x))$ then $P(y|do(x))$ *is identifiable*.

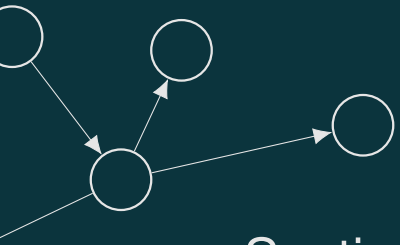
If there exists an adjustment set $\mathbf{Z}$ that satisfies the front-door criterion for $P(y|do(x))$ then $P(y|do(x))$ *is identifiable*.

Soundness and Completeness

If there exists an adjustment set **Z** that satisfies the back-door criterion for $P(y|do(x))$ then $P(y|do(x))$ *is identifiable*.

If there exists an adjustment set **Z** that satisfies the front-door criterion for $P(y|do(x))$ then $P(y|do(x))$ *is identifiable*.

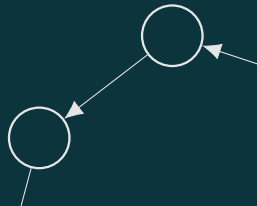
If there exists **no adjustment set Z** that satisfies the back- or front-door criterion for $P(y|do(x))$, $P(y|do(x))$ **might still be identifiable** via the do-calculus!



Section

3

Do-Calculus



Do-Calculus

Do-calculus provides a sound and complete set of rules for determining (total) causal effects for non-parametric models.

- **Soundness:** If you can derive an effect using the rules of do-calculus, it is *guaranteed* to be the correct formula.
- **Completeness:** If a causal effect is identifiable at all, then the three rules of do-calculus are *sufficient* to find that formula.

Pearl, Judea. "Causal diagrams for empirical research." Biometrika 82.4 (1995): 669-688.

Huang, Yimin, and Marco Valtorta. "Pearl's calculus of intervention is complete." Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence. 2006.

Shpitser, Ilya, and Judea Pearl. "Identification of joint interventional distributions in recursive semi-Markovian causal models." AAAI. 2006.

Rule 1: Insertion/Deletion of Observations

$$P(y \mid do(x), z, w) = P(y \mid do(x), w) \text{ if } (Y \perp\!\!\!\perp Z \mid X, W) \text{ in } \mathcal{G}_{\overline{X}} \quad (1)$$

X, Y, W, Z can all be sets of nodes. (W can simply be empty set.)

$G_{\overline{X}}$ means that all edges that go into X are deleted.

Interpretation: “If the intervention makes Y and Z independent (d-separated) in the graph we can remove it from the conditioning set”

Pearl, Judea. "Causal diagrams for empirical research." Biometrika 82.4 (1995): 669-688.

Rule 2: Action/Observation Exchange

$$P(y \mid do(x), do(z), w) = P(y \mid do(x), z, w) \text{ if } (Y \perp\!\!\!\perp Z \mid X, W) \text{ in } \mathcal{G}_{\overline{X},\underline{Z}} \qquad (2)$$

$G_{\underline{Z}}$ means that we delete all edges emerging from Z.

Interpretation: "If Z and Y are not confounded via a backdoor path (under $do(x), w$), conditioning and intervening on Z are the same".

Pearl, Judea. "Causal diagrams for empirical research." Biometrika 82.4 (1995): 669-688.

Rule 3: Insertion/deletion of actions

$$P(y \mid do(x), do(z), w) = P(y \mid do(x), w) \text{ if } (Y \perp\!\!\!\perp Z \mid X, W) \text{ in } \mathcal{G}_{\overline{X},\overline{Z(W)}} \qquad (3)$$

where $Z(W)$ is the set of Z-nodes that are *not* anchestors of any W-node in $G_{\overline{X}}$.

Interpretation: We can only mimic the cutting of edges of the do-operator if no causal effects into Z reach Y in the $do(X), w$-only case. This is either true, if there simply are no such effects into Z, or that W blocks all chains and does not activate any colliders that could propagate the effect to Y.

In short: “If the intervention $do(Z)$ is irrelevant to Y, it can be deleted from the equation”.

Pearl, Judea. "Causal diagrams for empirical research." Biometrika 82.4 (1995): 669-688.

Do-Calculus

do-calculus

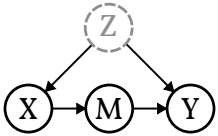
$$P(y \mid do(x), z, w) = P(y \mid do(x), w) \quad \text{if } (Y \perp\!\!\!\perp Z \mid X, W) \text{ in } \mathcal{G}_{\overline{X}}$$

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

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

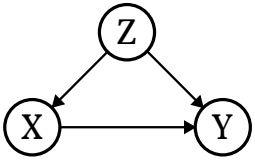
Pearl, Judea. "Causal diagrams for empirical research." Biometrika 82.4 (1995): 669-688.

Derivation of Front-Door Adjustment



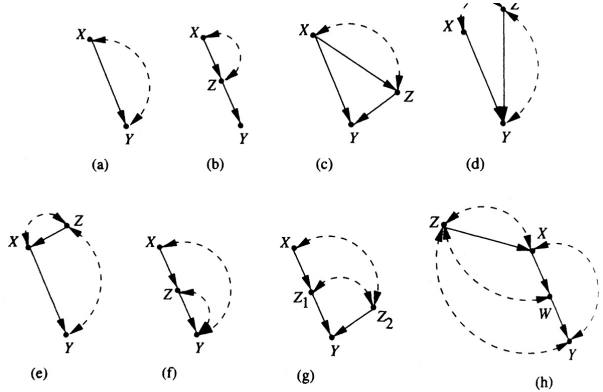
$$
\begin{aligned}
& P(Y|do(X)) \\
&= \sum_M P(Y|do(X), M) P(M|do(X)) && \text{Law of total prob.} \\
&= \sum_M P(Y|do(X), \text{do(M)}) P(M|do(X)) && \text{Rule 2} \\
&= \sum_M P(Y|do(X), do(M)) P(M|\text{X}) && \text{Rule 2} \\
&= \sum_M P(Y|\ , do(M)) P(M|X) && \text{Rule 3} \\
&= \sum_X' \sum_M P(Y|do(M), \text{X'}) \text{P(X'|do(M))} P(M|X) && \text{Prob. ax.} \\
&= \sum_X' \sum_M P(Y|\text{M}, X') P(X'|do(M)) P(M|X) && \text{Rule 2} \\
&= \sum_X' \sum_M P(Y|M, X') P(X', \ ) P(M|X) && \text{Rule 3}
\end{aligned}
$$

Exercise



Try to derive the estimand of the backdoor criterion for the query $P(Y|do(X = x))$ on the above graph, using the rules of do-calculus.

Non-Identifiable Graphs



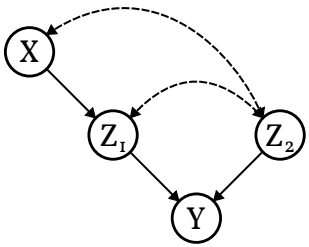
Not all queries can be identified. A sufficient criterion for non-identifiability of $P(y|do(x))$ is confounding of X and any of its children on a path from X to Y (e.g. in (b),(c)).

Pearl, Judea. Causality. Cambridge university press, 2009.

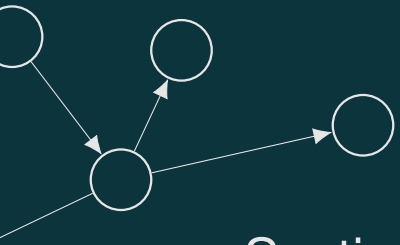
Parametric Identifiability

We can identify $P(z_1|do(x))$ and $P(y|do(z_1))$ but not $P(y|do(x))$.

1. X to Z_1 is identifiable with an empty adjustment set.
2. Z_1 to Y is only identifiable with Z_2 adjusted.



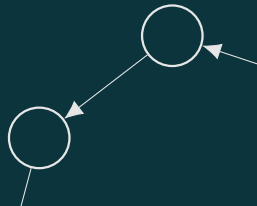
If we were in a *parametric* setting, e.g. linear additive models, we could infer the coefficients of the intermediate edges individually, and thus identify the total effect!



Section

4

Pearl Causal Hierarchy



Pearl Causal Hierarchy (PHC) I

Given the soundness and completeness of the Pearlian do-calculus, we found some queries which can not be answered with interventional data.

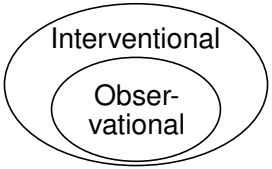
(Meaning, we can not transform them into expressions of purely observational terms.)

Pearl Causal Hierarchy (PHC) I

Given the soundness and completeness of the Pearlian do-calculus, we found some queries which can not be answered with interventional data.

(Meaning, we can not transform them into expressions of purely observational terms.)

$\Rightarrow$ there exists a distinct subset of queries that *require* interventions.



Observational $\subset$ Interventional

Pearl Causal Hierarchy (PHC) II

Observational and interventional queries give insights on the probabilities of population statistics.

E.g. what would happen *on average* if I do prescribe this medicine to patients?

Sometimes we want to *retrospectively* reason about the *individual outcome of a particular scenario*.

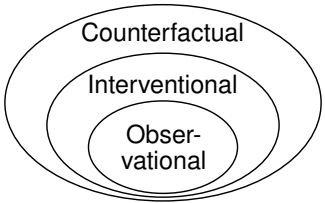
Pearl Causal Hierarchy (PHC) II

Observational and interventional queries give insights on the probabilities of population statistics.

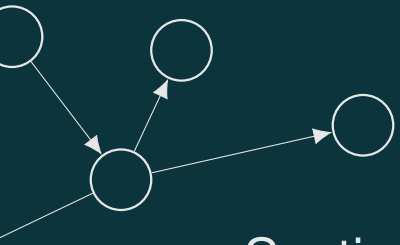
E.g. what would happen *on average* if I do prescribe this medicine to patients?

Sometimes we want to *retrospectively* reason about the *individual outcome of a particular scenario*.

Counterfactuals form a third class of queries that is distinct from the previous two:

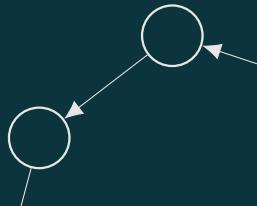


Observational $\subset$ Interventional $\subset$ Counterfactual



Section 5

Counterfactuals



What are Counterfactuals

Counterfactuals answer questions of *"What if…?"* and *"Why…?"*.
Observational: "How many people that go to parties catch a flu?"
Interventional: "How many will catch a flu, if we order them to join the party?"
Counterfactual: "Would John have caught a flu if he didn't go to go the party?"

What are Counterfactuals

Counterfactuals answer questions of *"What if…?"* and *"Why…?"*.
Observational: "How many people that go to parties catch a flu?"
Interventional: "How many will catch a flu, if we order them to join the party?"
Counterfactual: "Would John have caught a flu if he didn't go to go the party?"

Counterfactuals let us reason about the potential outcomes of individuals, given some observed evidence.

What are Counterfactuals

Counterfactuals answer questions of *"What if…?"* and *"Why…?"*.
Observational: "How many people that go to parties catch a flu?"
Interventional: "How many will catch a flu, if we order them to join the party?"
Counterfactual: "Would John have caught a flu if he didn't go to go the party?"

Counterfactuals let us reason about the potential outcomes of individuals, given some observed evidence.

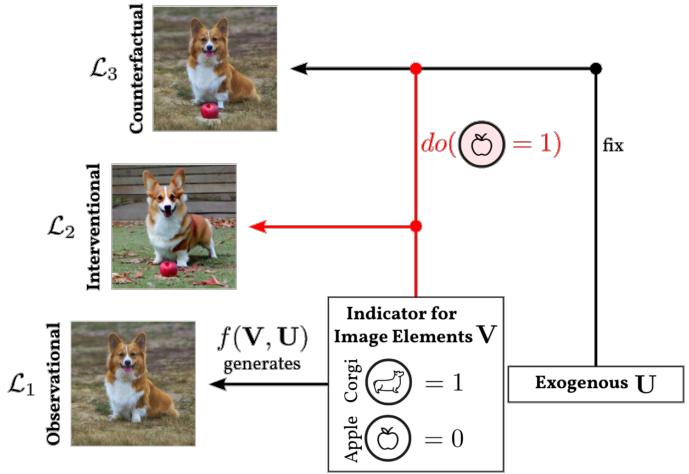
Not a generic person, *but John* who went to the party and got sick.
What does the evidence tell us about the strength of his immune system?

Many similar questions:
What if I had not smoked in the past?
What if I had learned for the exam?
What if Boris the Animal would have made all the lights on Bowery and got there early and were just about to discharge a weapon through the doorway?

Counterfactuals in Images

Interventions enforce a particular action, but do not control for the remaining variables. Outcomes still vary due to the sampling of the latent noise factors U.

Counterfactuals infer the latent noise factors from a given observation and only then apply the intervention!



Zečević*, M., Willig*, M., Singh Dhami, D. and Kersting, K., 2023. Identifying challenges for generalizing to the pearl causal hierarchy on images. ICLR 2023 Workshop on Domain Generalization.

Counterfactuals

"A counterfactual $P(B_A|e)$ is the probability of B given the evidence e under $do(A)$."

Task of Counterfactual Inference: Given a model $\langle \mathcal{M}, P(u) \rangle$, compute a counterfactual $P(B_A|e)$.

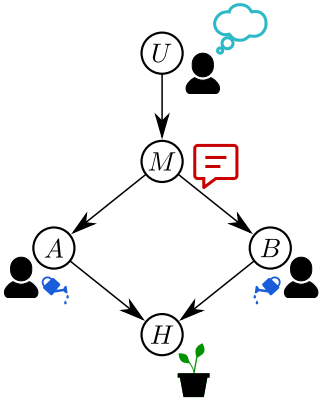
Counterfactual Inference

1. **Abduction:** Update $P(u)$ by the evidence e to obtain $P(u|e)$.
2. **Action:** Apply $do(A)$ to obtain the graph $\mathcal{M}_A$.
3. **Prediction:** Use the modified model $\langle \mathcal{M}_A | P(u|e) \rangle$ to compute $P(B|u, e)$.

Sometimes the pair $\langle \mathcal{M}, u \rangle$ is called a *causal world*.

Example: Watering Plant Counterfactuals

$$\mathcal{M} = \begin{cases} \mathbf{V} & = \{M, A, B, H \in \mathbb{B}\} \\ \mathbf{U} & = \{U \in \mathbb{B}\} \\ \mathbf{F} & = \begin{cases} f_M & := U \\ f_A & := M \\ f_B & := \neg M \\ f_H & := A \vee B \end{cases} \\ \mathcal{P}_{\mathbf{U}} & = \{U = \text{Bernoulli}(0.5)\} \end{cases}$$

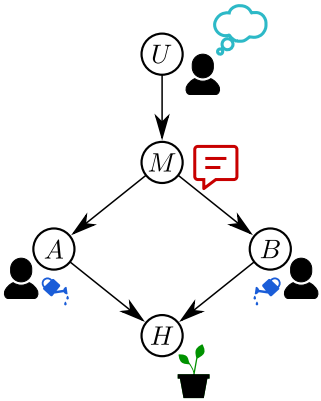


Modified scenario: Tom now always messages either friend A or friend B to take care of the plant.

"We saw friend A watering the plant yesterday. What would have happened if we prevented A from watering the plant?"

Example: Watering Plant Counterfactuals

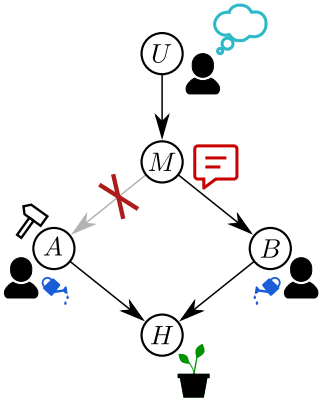
$$\mathcal{M}_{A=a} = \begin{cases} \mathbf{V} & = \{M, A, B, H \in \mathbb{B}\} \\ \mathbf{U} & = \{U \in \mathbb{B}\} \\ \mathbf{F} & = \begin{cases} f_M & := U \\ f_A & := M \\ f_B & := \neg M \\ f_H & := A \vee B \end{cases} \\ \mathcal{P}_{\mathbf{U}} & = \{U = \mathit{true}\} \end{cases}$$



1 Abduction: Update P(U) by the evidence.
Evidence: Friend A was watering the plant: e = (A=true).
Therefore $M = \mathit{true}$ and therefore $U = \mathit{true}$.

Example: Watering Plant Counterfactuals

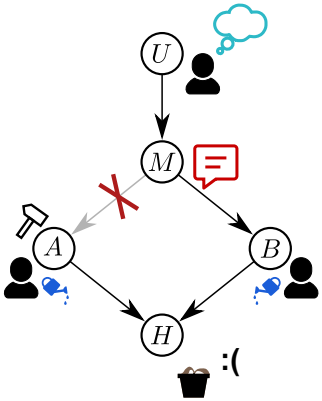
$$\mathcal{M}_{A=a} = \begin{cases} \mathbf{V} & = \{M, A, B, H \in \mathbb{B}\} \\ \mathbf{U} & = \{U \in \mathbb{B}\} \\ \mathbf{F} & = \begin{cases} f_M & := U \\ f_A & := \textit{false} \\ f_B & := \neg M \\ f_H & := A \vee B \end{cases} \\ \mathcal{P}_{\mathbf{U}} & = \{U = \textit{true}\} \end{cases}$$



1 Abduction: Update P(U) by the evidence.
2 Action: Apply $do(A = \textit{false})$ to obtain the graph $\mathcal{M}_{A=a}$.

Example: Watering Plant Counterfactuals

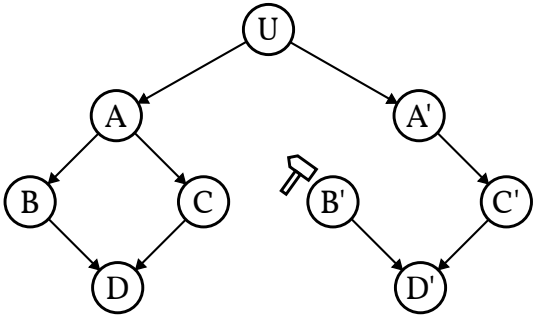
$$\mathcal{M}_{A=a} = \begin{cases} \mathbf{V} & = \{M, A, B, H \in \mathbb{B}\} \\ \mathbf{U} & = \{U \in \mathbb{B}\} \\ \mathbf{F} & = \begin{cases} f_M & := U \\ f_A & := \mathit{false} \\ f_B & := \neg M \\ f_H & := A \vee B \end{cases} \\ \mathcal{P}_{\mathbf{U}} & = \{U = \mathit{true}\} \end{cases}$$



1 Abduction: Update P(U) by the evidence.
2 Action: Apply $do(A = \mathit{false})$ to obtain the graph $\mathcal{M}_{A=a}$.
3 Prediction: Use the modified model $\langle \mathcal{M}_A | P(u|e) \rangle$ to compute $P(h_{A=a}|u, e)$:

$P(H = \mathit{false}|U = \mathit{true}) = 100\%$

Twin-Networks



Twin-networks contain the ‘factual’ (left) and ‘counterfactual’ (right) world in a single graph. Both worlds are connected via the same shared exogenous variables.

Collapse of Causal Rungs

When queries of a higher level of the PHC become identifiable from lower levels of the hierarchy, we say that rungs of the PHC *collapse*.

Collapse of Causal Rungs

When queries of a higher level of the PHC become identifiable from lower levels of the hierarchy, we say that rungs of the PHC *collapse*.

Interventional/Observational: We have seen that some interventional queries could be transformed into purely observational ones. If every interventional query in a particular graph becomes observationally identifiable, we say that levels $\mathcal{L}_1$ and $\mathcal{L}_2$ collapse.

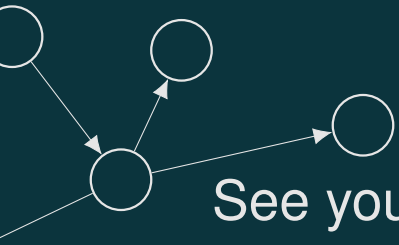
Collapse of Causal Rungs

When queries of a higher level of the PHC become identifiable from lower levels of the hierarchy, we say that rungs of the PHC *collapse*.

Interventional/Observational: We have seen that some interventional queries could be transformed into purely observational ones. If every interventional query in a particular graph becomes observationally identifiable, we say that levels $\mathcal{L}_1$ and $\mathcal{L}_2$ collapse.

Interventional/Counterfactual: Interventional queries only do Step 2 and 3 on an average, generic population. Counterfactuals are personalized to a specific individual using the evidence from the real world (Step 1).

Counterfactual queries collapse to interventional queries, when the first 'abduction' step provides no information on U. In simpler terms, *when the given evidence does not help us learn anything new about a systems hidden state*.



See you next week!