TECHNISCHE UNIVERSITÄT DARMSTADT

AIML Lab

Winter Semester 2025/26 Lecture

Causality for AI & ML

“Recap & Exam Info”

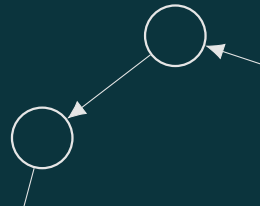
Prof. Dr. Kristian Kersting

Moritz Willig — Today’s speaker

Tim Woydt

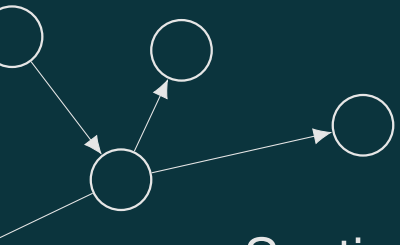
Florian Busch

Matej Zečević



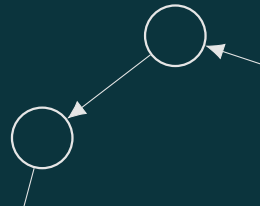
Today's Lecture:

1. Lecture Recap
2. Practice Exam Review
3. Q&A
4. Exam Info

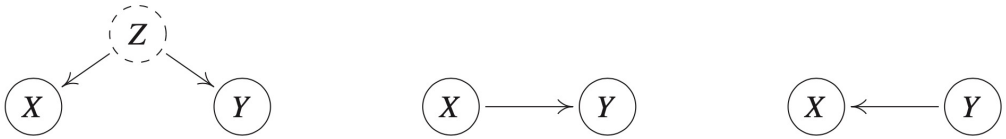


Section 1

Lecture 1: Introduction



Reichenbach's Common Cause Principle



If two variables X, Y are correlated, there are three possible types of causal structure:

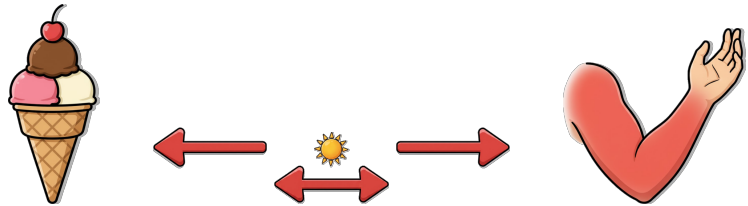
1. X and Y have a common cause Z.
2. X causes Y.
3. Y causes X.

Also: *Conditioning* on a common child of X and Y might correlate variables, leading to collider bias! (More on this in later lectures.)

Why do we care about Causality?

Understanding the structure of the underlying processes is a key component for making robust and reliable decisions.

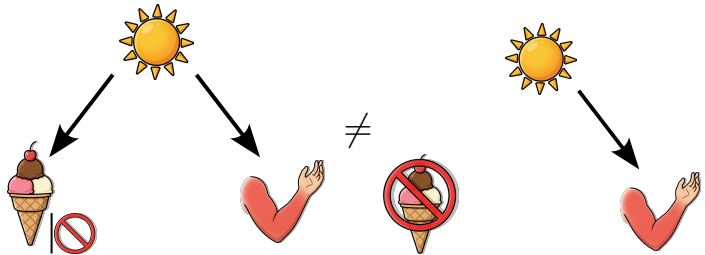
You observe a correlation between ice cream sales and sunburns:



What is the right causal direction?
Observing correlations only is *inconclusive*!
Correlational guessing can be arbitrarily wrong!

Conditioning $\neq$ Intervening

Conditioning and intervening does not yield the same outcome!



Ladder of Causality

The **Ladder of Causality** defines three different levels (or rungs) of causal inference:

1. **Associational**: Observing
2. **Interventional**: Doing
3. **Counterfactual**: Imagining

Levels form a strict hierarchy

Associational $\subset$ **Interventional** $\subset$ **Counterfactual**

with higher levels being strictly more powerful than the lower ones.

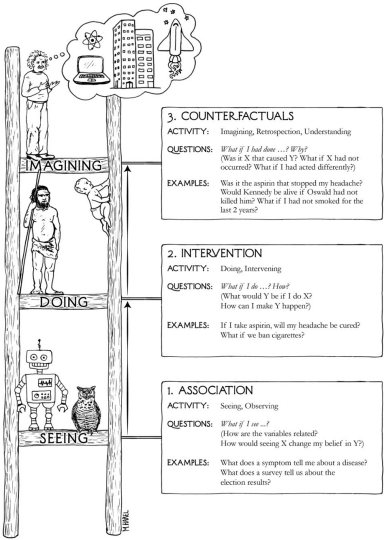
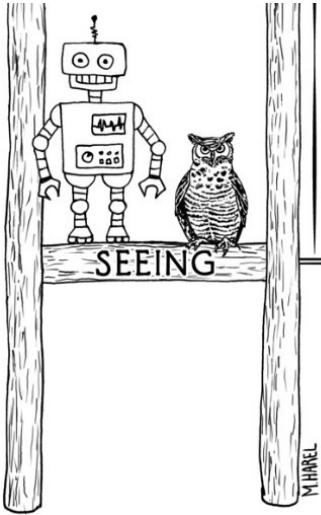


Figure: Pearl and Mackenzie, “The book of why”, Basic books, 2018.

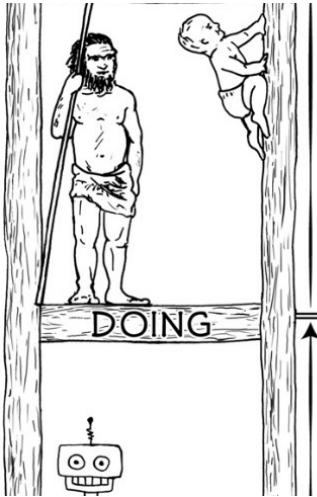
Ladder of Causality - Associational



1. ASSOCIATION

ACTIVITY:	Seeing, Observing
QUESTIONS:	*What if I see ...?* (How are the variables related? How would seeing X change my belief in Y?)
EXAMPLES:	What does a symptom tell me about a disease? What does a survey tell us about the election results?

Ladder of Causality - Interventional



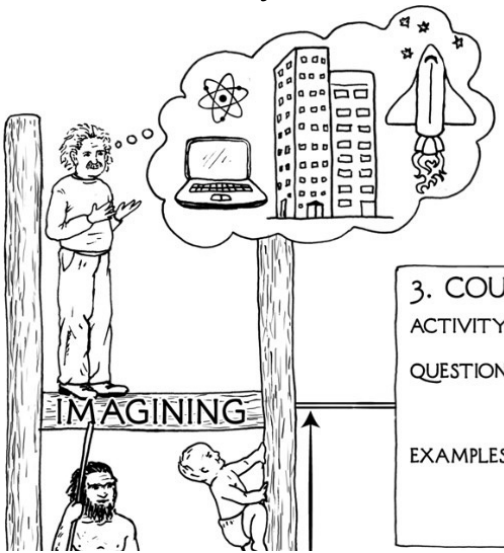
2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

Ladder of Causality - Counterfactual



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

Do-calculus

Important result: Even if we know the correct causal graph, **some queries cannot be solved from pure observations!**

The do-calculus tells us **if** ('completeness') and **how** ('sound') a causal query can be identified from data.

Without making any further assumptions:
For some queries we can only find out by *doing* experiments.

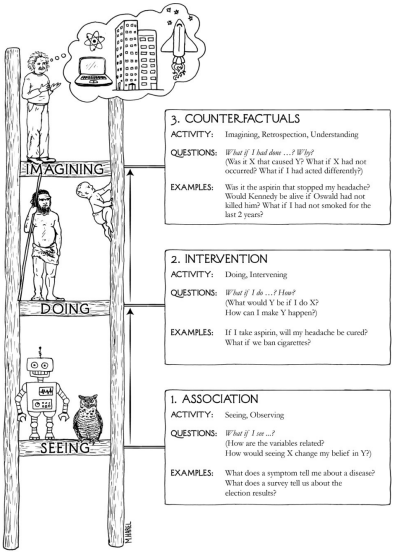
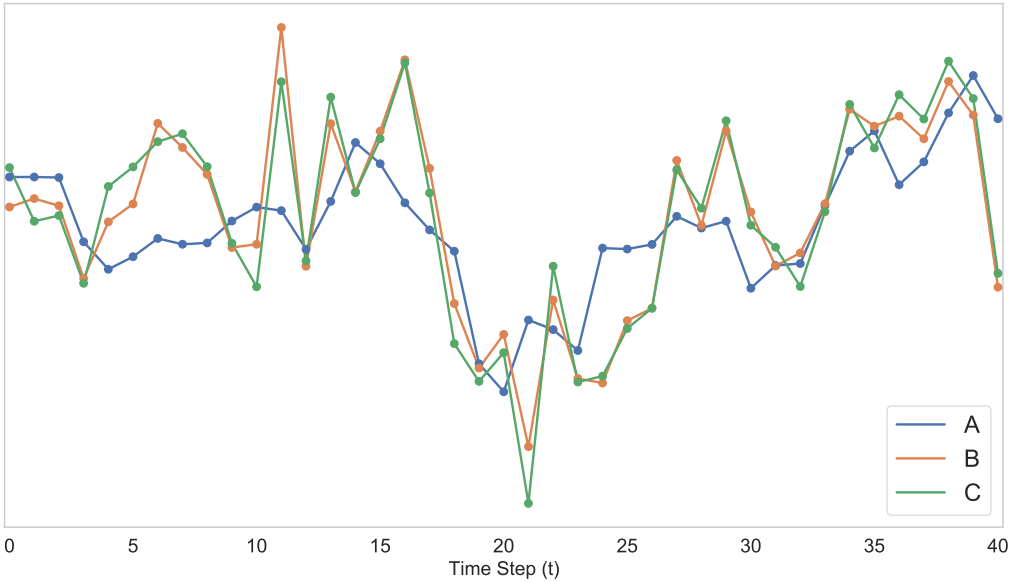


Figure: Pearl and Mackenzie, "The book of why", Basic books, 2018.

Limitations of Correlational Learning

Consider the following setup:
Your task is to train a model that predicts B given A and C.

$$
\begin{aligned}
A &:= \mathcal{N}(0, 1.0) \\
B &:= A + \mathcal{N}(0, 3.1) \\
C &:= B + \mathcal{N}(0, 0.23)
\end{aligned}
$$

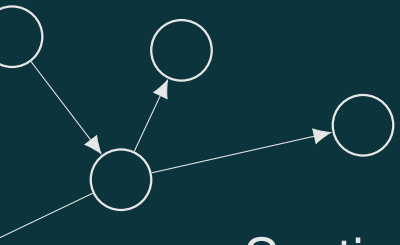

A
B
C
0
5
10
15
20
25
30
35
40
Time Step (t)

Limitations of Correlational Learning

Consider the following setup:
Your task is to train a model that predicts B given A and C.

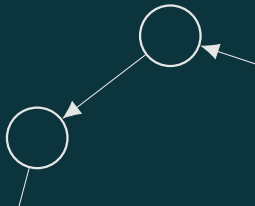
A correlational model will infer B from C because it is more predictive!

```
--- Linear Regression Results ---
Learned coefficients for A: 0.0419
                          C: 0.8919
Intercept: 0.0500
--------------------------------
```



Section 2

Lecture 2: Probabilities & Bayesian Networks



Joint, Conditional and Marginal Distributions

Joint probability $P(X, Y)$

Conditional probability $P(X|Y) := \frac{P(X,Y)}{P(Y)}$ (for $P(Y) > 0$)

Chain Rule $P(X, Y) = P(X)P(Y|X) = P(Y)P(X|Y)$ ($n!$ possibilities)

Marginalization $P(X) = \sum_{y} P(X, Y = y)$

Bayes' rule $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$

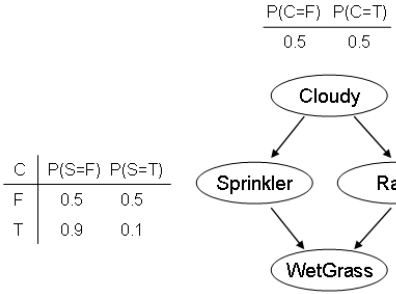
(Marginal) Independence $X \perp Y :\Leftrightarrow P(X|Y) = P(X) \Leftrightarrow P(X, Y) = P(X)P(Y)$

Conditional Independence $X \perp Y \mid Z :\Leftrightarrow P(X|Y, Z) = P(X|Z)$

Probabilities & Bayesian Networks - Exercises

4. Consider the water sprinkler Bayes net with binary nodes.

   a) Provide a minimal factorization for $P(C, S, R, W)$ w.r.t. the DAG.
   b) Compute $P(S = t | W = t)$.
   c) Compute $P(W = t)$.



P(C=F)	P(C=T)
0.5	0.5

C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

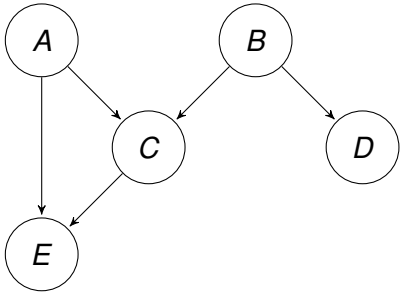
C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S R	P(W=F)	P(W=T)
F F	1.0	0.0
T F	0.1	0.9
F T	0.1	0.9
T T	0.01	0.99

Probabilities & Bayesian Networks - Exercises

5. Apply d-separation to determine which of the following conditional independencies hold for the DAG. For those that do not hold, name an active trail between the nodes.

   a) $A \perp D \mid B, C$
   b) $A \perp B \mid E$
   c) $E \perp D \mid C$



d-separation - Definition

For a DAG $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ and $\mathbf{X}, \mathbf{Y}, W \subseteq V$ we say that $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given W ($d_{\mathcal{G}}(\mathbf{X}, \mathbf{Y}|W) = 1$) if there is no active trail between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ while observing W.

A trail is an undirected path in $\mathcal{G}$ that never visits a node twice and is called active while observing W if for each consecutive triplet $X - Z - Y$ one of the following holds: (We can think of active trails as information flow)

(a) $X \to Z \to Y$ (chain) and $Z \notin W$ (Y unobserved)

(b) $X \leftarrow Z \leftarrow Y$ (chain) and $Z \notin W$ (Y unobserved)

(c) $X \leftarrow Z \to Y$ (fork) and $Z \notin W$ (Y unobserved)

(d) $X \to Z \leftarrow Y$ (collider/v-structure) and $Z \in W$ or $de(Z) \cap W \neq \emptyset$ (Y or some descendent observed)

Markov Equivalence Classes - Definition

We call all DAG $\mathcal{G}'$ Markov equivalent to a DAG $\mathcal{G}$ if $I(\mathcal{G}') = I(\mathcal{G})$ and refer to the set of such equivalent DAG as Markov equivalence class (MEC) .

Using *d*-separation, we can show that (but will not do so in this lecture)

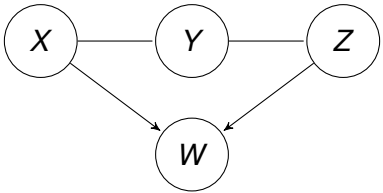
Two DAG $\mathcal{G}, \mathcal{G}'$ are Markov equivalent if and only if

i) they share the same skeleton, i.e. induce the same undirected graph,

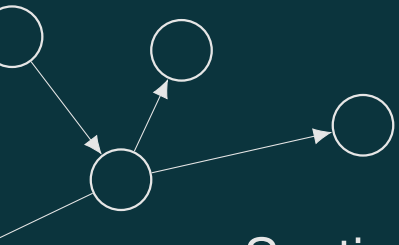
ii) and they have identical v-structures.

Causal PDAGs

If we do not know certain parts of a causal structure due to MEC ambivalence, we can use partially directed acyclic graph (PDAG) that has undirected edges for all non-colliders.

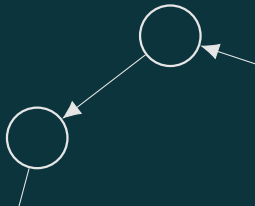


$\longrightarrow$ In the next weeks, we will learn how to infer this graphs from data.



Section 3

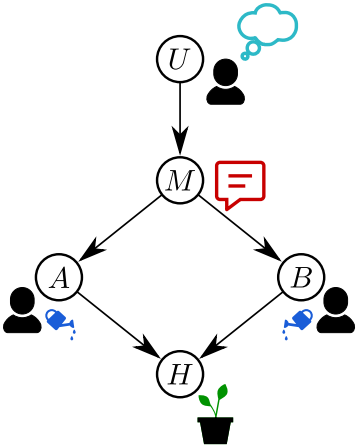
Lecture 3: Structural Causal Models



Causal Modeling

"Tom goes on vacation. Upon remembering that his plant needs to be taken care off while he is away, he sends a message his two friends. In case that the message is sent, both friends will take care of the flower."

$$
\begin{aligned}
U &:= \text{Bernoulli}(0.5) \\
M &:= U \\
A, B &:= M \\
H &:= A \vee B
\end{aligned}
$$



Structural Causal Model

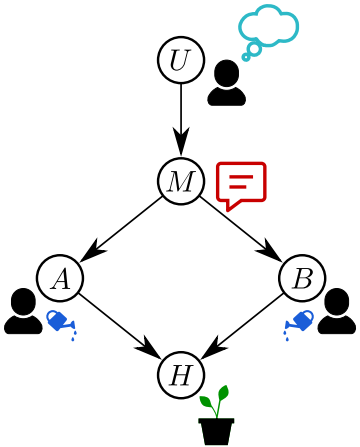
A **Structural Causal Model** (SCM) is a tuple $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, \mathcal{P}_{\mathbf{U}})$.

- $\mathbf{V}$ 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: $\mathbf{X} = \mathbf{V} \cup \mathbf{U}$
  – $\text{pa}(x_i)$ denotes the parents of x_i.

Flowering Plant SCM

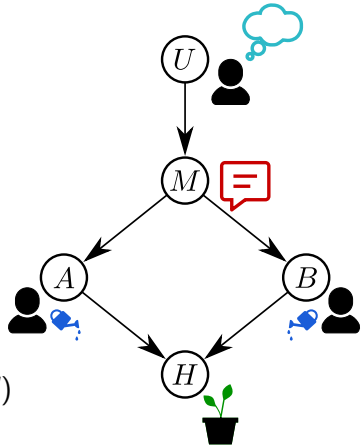
$$\mathcal{M} = \begin{cases} \mathbf{V} & = \{M, A, B, H \in \mathbb{B}\} \\ \mathbf{U} & = \{U \in \mathbb{B}\} \\ \mathbf{F} & = \begin{cases} f_M & := U \\ f_A & := M \\ f_B & := M \\ f_H & := A \vee B \end{cases} \\ \mathcal{P}_{\mathbf{U}} & = \{U = \text{Bernoulli}(0.5)\} \end{cases}$$



Factorization of the Joint Distribution

An SCM $\mathcal{M}$ entails a joint distribution $P_{\mathcal{M}}(X_1, \ldots, X_N)$ over all variables $X_1, \ldots, X_N$ that **factorizes according to the causal graph** $\mathcal{G}$:

$$P_{\mathcal{M}}(X_1, \ldots, X_N) = \prod_{i \in \{1..N\}} P(X_i | \text{pa}(X_i))$$



Example:

$$P(U, M, A, B, H) = P(H|A, B)P(A|M)P(B|M)P(M|U)P(U)$$

Markovianity and Faithfulness

Markovianity: $P_{\mathcal{M}}$ is called Markovian to $\mathcal{G}_{\mathcal{M}}$ if all independencies implied by the graph also hold true in the distribution:

Markov Condition

$$(X \perp\!\!\!\perp Y|Z)_G \Rightarrow (X \perp\!\!\!\perp Y|Z)_P$$

Faithful: $\mathcal{G}_{\mathcal{M}}$ is called faithful to $P_{\mathcal{M}}$ if all independencies implied by the distribution also hold true in the graph:

Faithfulness

$$(X \perp\!\!\!\perp Y|Z)_G \Leftarrow (X \perp\!\!\!\perp Y|Z)_P$$

Independence Maps

I-Map: If a distribution is Faithful to a graph, the graph is a *I-map* (Independency map). It contains at least all the dependencies of the graph, *but might contain more*!

D-Map: If a distribution is Markovian to a graph, the graph is an *D-map* (Dependency map). It lists at least all the independencies of the distribution, *but might contain more*!

If Markovianity and Faithfulness are met it is called a *P-Map* (Perfect map):

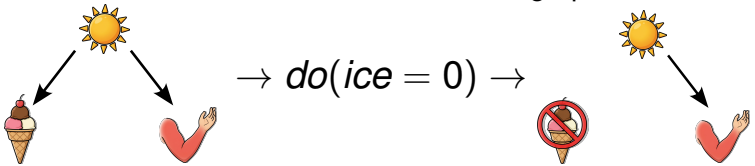
P-Map

$$(X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_P$$

The set of independencies in the graph is exactly identical to the set of independencies in the distribution.

Interventions and do-Operator

From the unintervened to the intervened graph:



The do-operator forcibly sets *ice* to 0, therefore *cutting all influence to the parents* of the intervened node.

do-Operator

For an SCM $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, \mathcal{P}_{\mathbf{U}})$ the do-operator $do(X_i = c)$ replaces the structural equation $f_i \in \mathbf{F}$ with the assignment $f_i := c$.

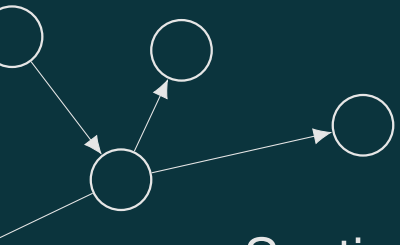
Truncated Factorization

For arbitrary interventions $\forall X_i \in \mathbf{V}.\forall c \in \text{dom}(X_i).do(X_i = c)$, SCM define a family of joint distributions.

Every hard intervention $do(X_j = c)$ entails a particular **Truncated Factorization**:

$$P_{\mathcal{M}}(X_1, \ldots, X_N | do(X_j = c)) = \left(\prod_{i \in \{1..N\} \setminus \{j\}} P(X_i | \text{pa}(X_i)) \right) \cdot P(X_j | do(X_j = c))$$

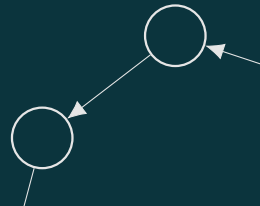
if $X_i = c$ and 0 otherwise.



Section

4

Lecture 4: do-Calculus



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 $X = x$?"

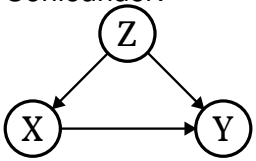
Average Treatment Effect (for binary outcome scenarios):

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

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

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:

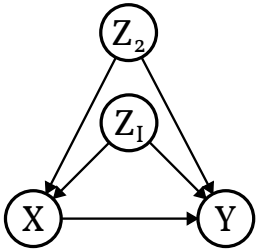
1. No node in $\mathbf{Z}$ is a descendant of X.
2. $\mathbf{Z}$ blocks every path between X and Y that contains an arrow into X.

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 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: $\mathbf{Z} = ?$

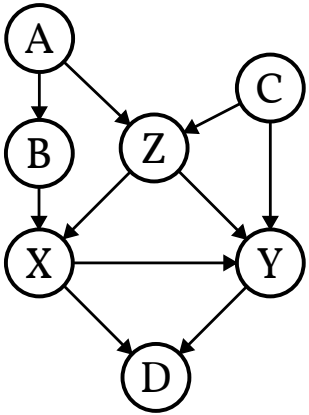
Remember:

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!

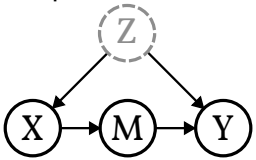
Note: The exists multiple minimal adjustment sets.

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



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.

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:

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

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.

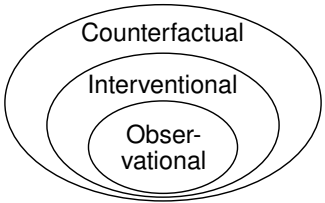
Pearl Causal Hierarchy (PHC)

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:

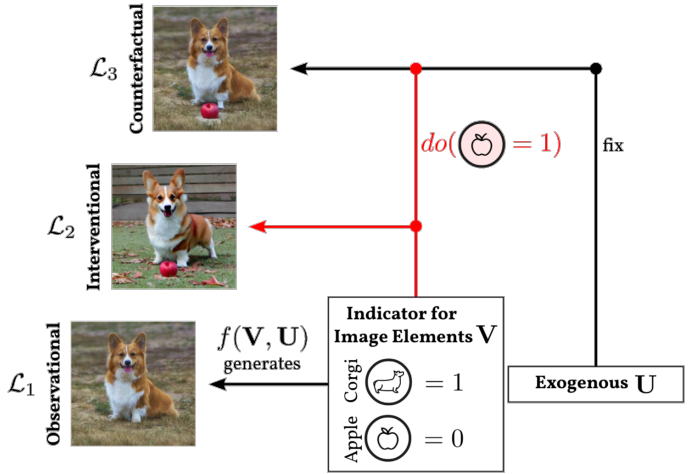


Observational $\subset$ Interventional $\subset$ Counterfactual

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*.

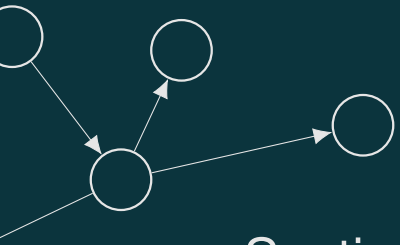
Collapse of Causal Rungs

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.

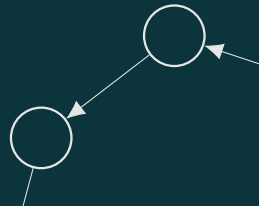
Interventional/Counterfactual: Interventional queries only do Step 2 and 3 on an average, generic population. Counterfactuals are personalized to a specific individual using the evidence from the real world (Step 1).

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*.



Section 5

Lecture 5: Causal Discovery



Causal Discovery (CD)

[1] Zanga, Alessio, Elif Ozkirimli, and Fabio Stella. "A survey on causal discovery: theory and practice." *International Journal of Approximate Reasoning* 151 (2022): 101-129.

Discover the true causal graph from tabular data

Causal Discovery

Let $\boldsymbol{G}$ be the set of graphs defined over the variables $\boldsymbol{V}$ of a dataset $\boldsymbol{D}$ and $G^* \in \boldsymbol{G}$ be the $\textit{true but unknown graph}$ from which $\boldsymbol{D}$ has been generated. The **causal discovery** problem consists in recovering the *true* graph G^* from the given dataset $\boldsymbol{D}$. [1]

→Variables are known and there is data for all variables

(a later lecture: causal representation learning, where the variables must be inferred from higher dimensional and / or entangled data representations)

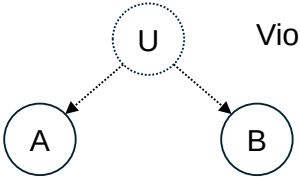
Peter-Clark Algorithm (PC) [2]

[2] Spirtes, Peter, Clark N. Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

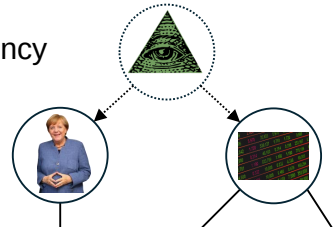
Assumptions: acyclicity, Markov, faithfulness, sufficiency (new)

Causal Sufficiency

A set V of variables is **causally sufficient** for a population if and only if in the population every common cause of any two or more variables in V is in V, or has the same value for all units in the population. [2]



Violations of Sufficiency



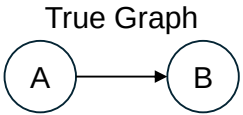
Intuition: If sufficiency does not hold, variables that should be independent in the graph (Markov condition) are correlated. This can not be fixed by drawing a directed edge between those, as either direction would be false.

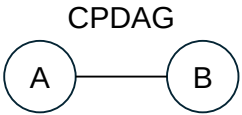
Peter-Clark Algorithm (PC)

Input: Dataset $\boldsymbol{D}^{m \times n}$ (m: number of samples, n: number of variables)

Output: CPDAG (Completed Partially Directed Acyclic Graph)

- A CPDAG represents the Markov Equivalence Class for the true graph
- Some edges can remain undirected
- An undirected edge between A and B means that either A causes B (A → B) or that B causes A (B → A); NOT that they could be confounded
- For example:





Peter-Clark Algorithm (PC)

Phase I: Skeleton discovery (all edges are undirected)

1. Start with fully connected graph
2. Remove all pairwise or conditionally independent variable edges

Phase II: Direct edges wherever possible

3. Direct edges using v-structures

- For all unshielded triplets A—B—C, i.e., A and C are not connected, direct it as A→B←C if B was not part of the conditioning set establishing independence between A and C before (phase I, step 2)

Food for thought: Why does this work?

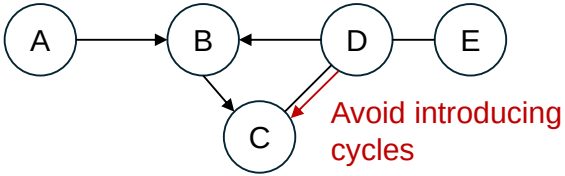
4. Meek rules

- Direct edges that can only be directed in one direction (more details soon)

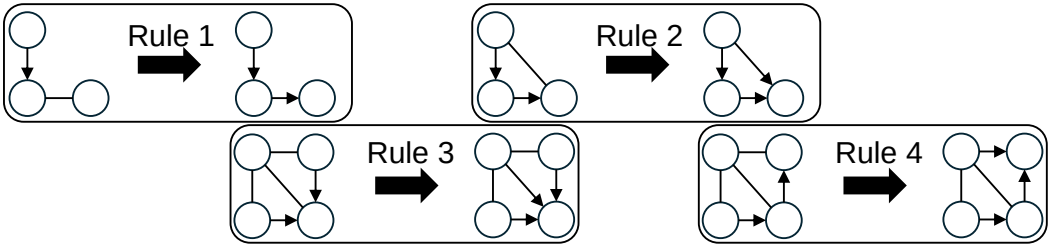
Meek Rules [3] More Generally

[3] Meek, Christopher. "Causal inference and causal explanation with background knowledge." *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. 1995.

- Assuming acyclicity and that all v-structures have been identified
- Meek rules: 4 rules, directing edges based on these assumptions
- Another example for Meek rules:



- All rules:



FCI algorithm (Fast Causal Inference algorithm) [4]

[4] Spirtes, Peter, Christopher Meek, and Thomas Richardson. "Causal inference in the presence of latent variables and selection bias." *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. 1995.

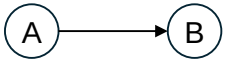
Assumptions: acyclicity, Markov, faithfulness

Input: Dataset $\boldsymbol{D}^{m \times n}$ (m: number of samples, n: number of variables)

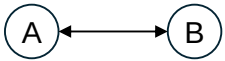
Output: PAG (Partial Ancestral Graph)

PAG: four different edge types

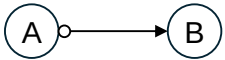
- A is a cause of B:



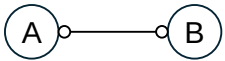
- There is a latent common cause of A and B:



- B is not an ancestor of A:



- No set d-separates A and B:



FCI implements a set of 10 different rules (general idea same as PC)

Score-Based Causal Discovery

Find the causal graph with the best score using a scoring criterion $S(G, \boldsymbol{D})$ [GES]

$$G^* = \text{argmax}_{G \in \mathbb{G}} S(G, \boldsymbol{D})$$

- Brute force all graphs? Super exponential growth (infeasible)
- More efficient simple strategy:

  greedily selecting the next improvement

- Risk of getting stuck in **local optima**

#Variables	#DAGS
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	514394281410443983349417907190839535103
15	237725265553410354992180218286376719253505

Remember lecture 1?

Bayesian Information Criterion (BIC)

$$\mathrm{BIC}(\mathrm{G}, \boldsymbol{X}) = k \ln(n) - 2 \ln\big(p(\boldsymbol{X}|G)\big)$$

Goal: minimize BIC

- k: parameters, i.e., number of edges (**lower** is better)
- n: sample size, i.e., number of data points (not a parameter)
- $p(\boldsymbol{X}|G)$: likelihood, i.e., the probability of observing the data $\boldsymbol{X}$ given the graph G (**higher** is better)

BIC is **decomposable** (can be computed independently for all m nodes):

$$\mathrm{BIC}(\mathrm{G}, \boldsymbol{X}) = k \ln(n) - 2 \ln\Big(\prod_{i=1}^{m} p(\boldsymbol{x}_i|G)\Big) = k \ln(n) - 2\sum_{i=1}^{m} \ln(p(\boldsymbol{x}_i|G))$$

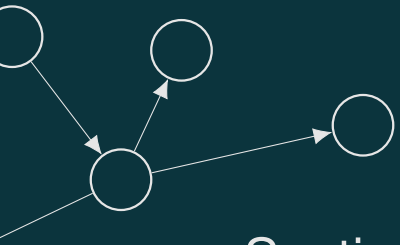
Greedy Equivalence Search (GES)

Chickering Sequence

Let G, H be DAGs such that $I(H) \subseteq I(G)$, then a Chickering sequence from G to H is a sequence of DAGs $G_0, \dots, G_M$ such that $G_0 = G$, $G_M = H$ and each G_m in the sequence is obtained from G_{m-1} either by an edge addition or a covered edge reversal.

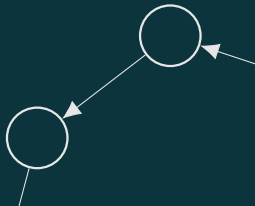
Intuition: H has a subset of G's independencies ($I(H) \subseteq I(G)$). Adding an edge can only "remove" an independence, so $I(H) \subseteq I(G)$ is still true. On the other hand, redirecting covered edges ensure that the independencies do not change, also keeping $I(H) \subseteq I(G)$ intact.

There provably exists such a sequence if $I(H) \subseteq I(G)$.



Section 6

Lecture 6: Dealing with Uncertainty



Precision, Recall and Accuracy

Most simple approach: Count the number of correctly/incorrectly predicted edges.

Accuracy: Acc = $\frac{TP+TN}{\#\text{ total possible edges}}$

Take care: Many graphs are sparse! For 100 nodes there are 10,000 possible edges. But a sparse graph might only have 500 edges.

An algorithm which does not predict a single edge still has an accuracy of 95%!

Precision: "How much percent of the predicted edges are actually correct?"
Prec = $\frac{TP}{TP+FP}$

Recall: How much percent of the true edges are actually predicted?"
Rec = $\frac{TP}{TP+FN}$

Often reported as a combined F_1**-score:**
F_1-score $= 2\frac{prec \cdot recall}{precision+recall}$

Structural Hamming Distance

Context: Hamming Distances are measures that determine the number of locations where two entities differ. (Historically developed for error correction in strings).

Idea: Apply the same to graphs! Check where individual edges differ between the ground truth (GT) and prediction (pred) graph and count up all the errors:

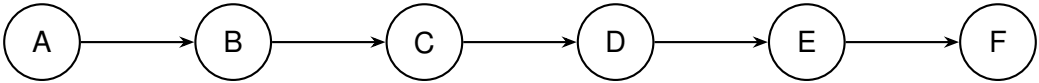
$$SHD(\mathcal{G}_{pred}, \mathcal{G}_{GT}) := \sum_{i \in [0..N]} \sum_{j \in [0..N]} \mathbf{1}((e_{i,j} \in \mathcal{G}_{pred}) \neq (e_{i,j} \in \mathcal{G}_{GT}))$$

Counting Direction Errors Twice: One could consider it a worse mistake to predict an edge in the wrong direction than not to predict it (or leaving it undirected). The above formula counts wrongly directed edges with double the error - once for $e_{i,j}$ and again for $e_{j,i}$.

Acid, Silvia, and Luis M. de Campos. "Searching for Bayesian network structures in the space of restricted acyclic partially directed graphs." Journal of artificial intelligence research 18 (2003): 445-490.

Tsamardinos, Ioannis, Laura E. Brown, and Constantin F. Aliferis. "The max-min hill-climbing Bayesian network structure learning algorithm." Machine learning 65.1 (2006): 31-78.

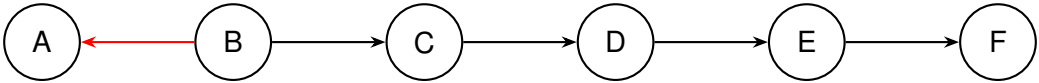
“Not all edges are the same”



Wrongly directing some edges in the above graph might be worse for some than for others.

Wahl, Jonas, and Jakob Runge. "Separation-based distance measures for causal graphs." arXiv preprint arXiv:2402.04952 (2024).

“Not all edges are the same”



Wrongly directing some edges in the above graph might be worse for some than for others.

Flipping A→B: Conditional (in)dependencies between A and all other variables (except B) are affected. Independencies between B,…,F, however, stay intact.

Wahl, Jonas, and Jakob Runge. "Separation-based distance measures for causal graphs." arXiv preprint arXiv:2402.04952 (2024).

“Not all edges are the same”

Wrongly directing some edges in the above graph might be worse for some than for others.

Flipping A→B: Conditional (in)dependencies between A and all other variables (except B) are affected. Independencies between B,…,F, however, stay intact.

Flipping C→D: Conditional (in)dependency statements between variables on both sides of the flipped edge are broken. Predicting this edge wrong has a much stronger impact on the correct prediction of many causal queries.

Wahl, Jonas, and Jakob Runge. "Separation-based distance measures for causal graphs." arXiv preprint arXiv:2402.04952 (2024).

Structural Intervention Distance (SID)

The SID counts the number of incorrectly predicted adjustment sets:

$$\text{SID}(\mathcal{G}, \mathcal{H}) = \# \left\{ (i,j), i \neq j \mid \begin{cases} j \in \mathbf{DE}_i^{\mathcal{G}} & \text{if } j \in \mathbf{PA}_i^{\mathcal{H}} \\ \mathbf{PA}_i^{\mathcal{H}} \text{ is not a valid adj. set for } (\mathcal{G}, i, j) & \text{if } j \notin \mathbf{PA}_i^{\mathcal{H}} \end{cases} \right\}$$

Zero Distance: If two graphs $\mathcal{G}, \mathcal{H}$ are equal, the SID is zero:
$\mathcal{G} = \mathcal{H} \Rightarrow \text{SID}(\mathcal{G}, \mathcal{H}) = 0$

The SID is ***not*** **symmetric**: $\text{SID}(\mathcal{G}, \mathcal{H}) = A \not\Rightarrow \text{SID}(\mathcal{G}, \mathcal{H}) = A$

Bounds: $\text{SHD}(\mathcal{G}, \mathcal{H}) = 1 \Rightarrow \text{SID}(\mathcal{G}, \mathcal{H}) \leq 2 \cdot (p - 1)$

Subgraph Distance: $\mathcal{G} \leq \mathcal{H} \Leftrightarrow \text{SID}(\mathcal{G}, \mathcal{H}) = 0$

CPDAGs: The SID is extended to CPDAGs by iterating over the DAGs of the CPDAG.

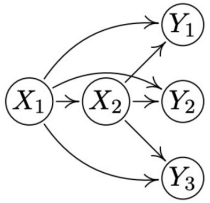
Peters and Bühlmann. “Structural Intervention Distance for Evaluating Causal Graphs”. In: Neural Computation 27.3 2015

SID Example

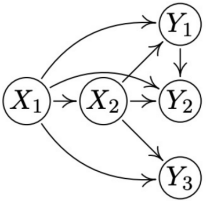
$\mathcal{H}_1$ **adds** an edge over $\mathcal{G}$.
$\mathcal{H}_2$ **reverses** an edge over $\mathcal{G}$.

$\text{SHD}(\mathcal{G}, \mathcal{H}_1) = 1$
$\text{SHD}(\mathcal{G}, \mathcal{H}_2) = 2$
or $\text{SHD}(\mathcal{H}_2) = 1$ depending on whether the reversal error is counted twice or not.

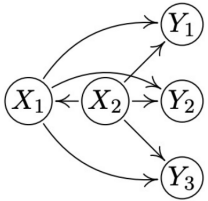
$\text{SID}(\mathcal{G}, \mathcal{H}_1) = 0$ ($\mathcal{G}$ is a subgraph of $\mathcal{H}_1$)
$\text{SID}(\mathcal{G}, \mathcal{H}_2) = 8$



true graph $\mathcal{G}$



graph $\mathcal{H}_1$



graph $\mathcal{H}_2$

Peters and Bühlmann. "Structural Intervention Distance for Evaluating Causal Graphs". In: Neural Computation 27.3 2015

Bounded Effects

Iterate all possible DAGs of a MEC/CPDAG and compute causal effects,
$\mathbf{C} = \{c_1, \ldots, c_m\}$.

Estimate bound by taking minimum and maximum of all estimates:
causal effect bound = $[\min(\mathbf{C}), \max(\mathbf{C})]$.

Deciding based on bounds:

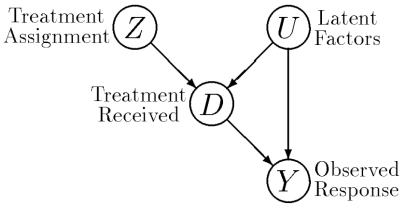
1. Bounds are **strictly positive [negative]** (e.g, $[0.5, 2.3]$)
   $\Rightarrow$ Positive [Negative] treatment effect "The medication always helps [harms]".
2. Bounds are **positive [negative]** but include zero (e.g, $[0.0, 2.3]$)
   $\Rightarrow$ "The medication might help [harm] or have no effect".
3. Bounds span into the **positive and negative domain** (e.g, $[-10.1, 2.1]$)
   $\Rightarrow$ No qualitative decision possible. "Medication might help or harm…".

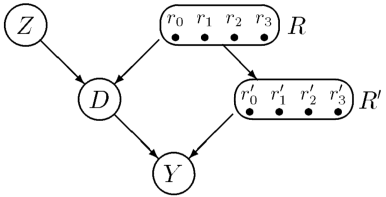
Partial Compliance - cont. II

Crossing all 4×4 different behaviors for treatment and response obtains 16 different latent groups. We write their probabilities as q_{ij}.

Example: Consider the observed sample:
$Z = 1, D = 1, Y = 1$.
We can not distinguish between a 'complier' and an 'always-taker'. Similarly, we can not distinguish between the 'helped' and 'always-recovers' case.

$$P(d, y|z) = p_{dy.z} = \sum_{q_{ij} \text{ consistent with } d,y,z} q_{ij}$$





Balke, Alexander and Pearl, "Judea. Nonparametric bounds on causal effects from partial compliance data". Technical Report R-199, Cognitive Systems Laboratory, UCLA, 1993

Matrix Powers and DAGness

For a weighted adjacency matrix W, an entry $(W^k)_{ij}$ of the k-power of the matrix

$$W^k = \underbrace{W \cdot W \cdot \ldots \cdot W}_{k-times}$$

is the sum of weights along all paths of length k from node i to node j.

The diagonal elements $(W^k)_{ii}$ therefore represent the presence of **self-cycles** $X_i \to \cdots \to X_i$ of length k.

If a graph is a DAG there should be no self-cycles!

graph is DAG $\Leftrightarrow \forall k \in \mathbb{N}.tr(W^k) = 0$

DAGness Score

Step 3. Again, recall

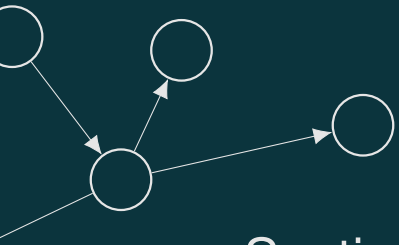
$$e^W = I + W + \frac{1}{2!}W^2 + \frac{1}{3!}W^3 + \dots$$

To avoid cancellation effects between individual negative and positive terms, entries of W are point-wise squared ($W \circ W$) to make them positive.

Step 4. The DAGness score is finally defined as:

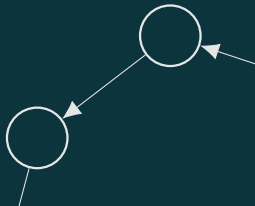
$$h(W) = tr(e^{W \circ W}) - d$$

The matrix exponential (e.g., in `scipy.linalg.expm`) is computed via a 'Scaling and Squaring' approach combined with Padé approximations. Complexity of $\mathcal{O}(d^3)$.

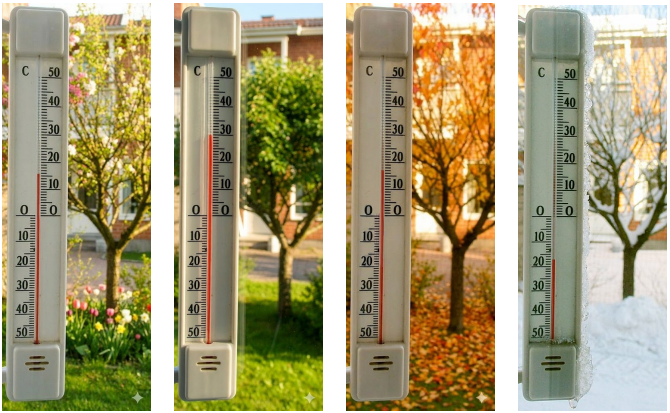


Section 7

Lecture 7: Causal Abstractions



Why Abstractions?: Semantics

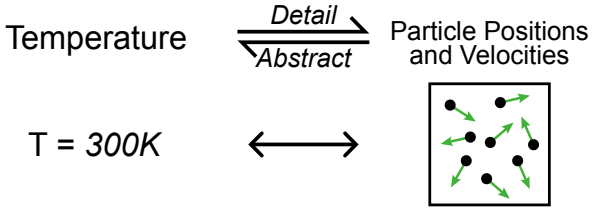


Variables are the 373×1024 pixels of the image.
$\rightarrow$ Is pixel $[23, 825]$ a useful variable?"

Image: https://en.wikipedia.org/wiki/File:Pakkanen.jpg

Levels of Granularity

Systems can be expressed at different levels of detail:



Choose the right level of detail for the right purpose.

→ We want to build causal models that are useful to us.

Transforming Models

We might want to simplify models to

- … reduce model complexity
- … reduce computational effort
- … obtain high-level insights

Typical operations of abstractions:

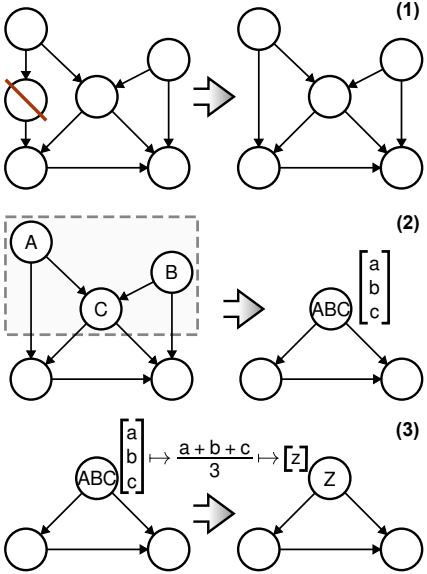
(1) **Marginalization** of variables

(2) **Grouping** of variables

(3) **Transformation** of variables

Making models more *detailed* is also possible but usually requires further observables.

→ Simplifying models is often ‘easier’.



(τ-ω)-Abstraction

An (exact) (τ-ω)-Abstraction is a pair of functions (τ, ω), where

1. $\tau : \mathbf{X}_L \to \mathbf{X}_H$ is a surjective map,
2. ω is a function $\omega : \mathcal{I}_L \to \mathcal{I}_H$ between sets of interventions,
3. and the following diagram commutes:

$$\begin{array}{ccc} \mathbb{P}_X & \xrightarrow{\text{do}(i)} & \mathbb{P}_X^{\text{do}(i)} \\ \downarrow{\scriptstyle \tau} & & \downarrow{\scriptstyle \tau} \\ \mathbb{P}_Y & \xrightarrow{\text{do}(\omega(i))} & \mathbb{P}_Y^{\text{do}(\omega(i))} \end{array}$$

Rubenstein, P.K., Weichwald, S., Bongers, S., Mooij, J.M., Janzing, D., Grosse-Wentrup, M. and Schölkopf, B.. "Causal Consistency of Structural Equation Models". In Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, 2017

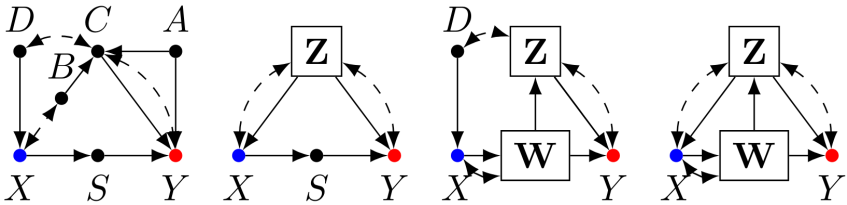
Types of abstractions

More refined notions of $\tau - \omega$-abstractions are given in Beckers et al., 2019:

- **Uniform**: Interventions should not only hold under a specific exogenous probability, but under all possible ones.
- **Strong**: All sets of interventions should be allowed. (Prevents cherry picking of interventions to hide inconsistencies of the mapping.)
- **Constructive**: High-level variables must be formed by grouping low-level variables into disjoint sets (clusters) with no overlap.
- **Natural Intervention Mapping**: τ tells us how to translate low- to high-level values. The ω mapping should reflect that in its translation of intervention values.

Beckers, Sander, and Joseph Y. Halpern. "Abstracting causal models." Proceedings of the aaai conference on artificial intelligence. Vol. 33. No. 01. 2019.

Identifiability in C-DAGs



The effect X on Y in (a) is identifiable via back-door (Adj=$\{B, D\}$) and front-door adjustment (via S).

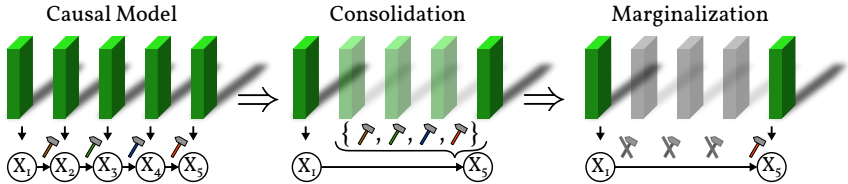
(b) still allows front-door adjustment via S.

(c) is no longer identifiable.

(d) is inadmissible. (Contains a cycle in (X, W, Z)).

Anand, T.V., Ribeiro, A.H., Tian, J. and Bareinboim, E., 2023, June. Causal effect identification in cluster dags. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 10, pp. 12172-12179).

Consolidation - An Intermediate Stage to Marginalization



The state of the last domino simply depends on the state of the first, except if we intervene by holding on to any of the dominos along the way.

- **Low-Level SCM**: Detailed model. All variables intervenable.
- **Marginalized SCM**: Some interventions are no longer directly applicable.
- **Consolidated SCM**: Equations explicitly incorporate interventions.

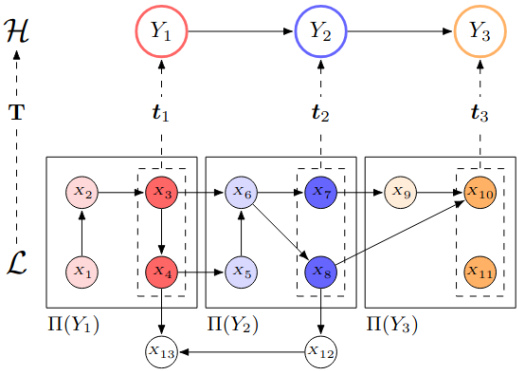
$$X_5 := \begin{cases} c & \textbf{if } do(X_i = c) \in \mathbf{I} \\ X_1 & \textbf{else} \end{cases}$$

Willig, M., Zečević, M., Dhami, D. and Kersting, K., 2023. Do not marginalize mechanisms, rather consolidate!. Advances in Neural Information Processing Systems, 36, pp.60947-60965.

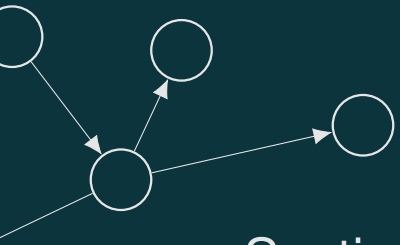
Abs-LiNGAM

Utilize causal abstractions for causal discovery.

1. Learn the abstraction map τ using paired low- and high-level data.
2. Learn the abstract model by abstracting data to the high-level.
3. Infer constraints.
   (here:
   $\blacksquare \not\to \blacksquare$, $\blacksquare \not\to \blacksquare$, $\blacksquare \not\to \blacksquare$, $\blacksquare \not\to \blacksquare$)
4. Discover the low-level model from data and inferred constrains.

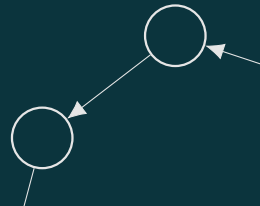


Massidda, Riccardo, Sara Magliacane, and Davide Bacciu, 2024. "Learning Causal Abstractions of Linear Structural Causal Models." The 40th Conference on Uncertainty in Artificial Intelligence.



Section 8

Lecture 8: Neuro-Causal Models



Why Neuro-Causal

Pros of classical ML:

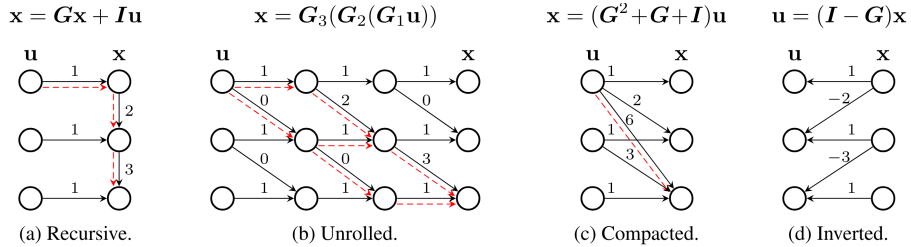
1. Applicable to a wide range of problems.
2. Straightforward training process.
3. Easy large-scale adaption.

Cons of classical ML:

1. Does not generalize well out-of-distribution.
2. Models often stay on the correlational level.
3. Might include confounding factors in decisions.

$\rightarrow$ **Can we counter the cons of ML by leveraging causal methods & models?**
$\leftarrow$ **Can we scale causal methods and models beyond small-scale examples using ML methods?**

SCM Representations



SCM evaluation can be represented in various ways.

Assumption for CNFs: (1) Structural eqs. are diffeomorphisms (‘bijective and smooth’), (2) the model is acyclic and (3) the data is causally sufficient.

Javaloy, Adrián, Pablo Sánchez-Martín, and Isabel Valera. "Causal normalizing flows: from theory to practice." Advances in Neural Information Processing Systems 36 (2023): 58833-58864.

Normalizing Flows

Normalizing Flows (NFs) are *density estimators* that transform distributions into each other, via a series of bijective transforms, $t_\theta^1 \circ \cdots \circ t_\theta^k = T_\theta$.

Every transform t_i needs to

- preserve the probability mass at every step $\forall l \in [1..K]. \int t_\theta^1 \circ \cdots \circ t_\theta^l(P) = 1.0$.
- keep values above (or equal to) 0.

$\rightarrow$ Every step, again, produces a valid density.

Javaloy, Adrián, Pablo Sánchez-Martín, and Isabel Valera. "Causal normalizing flows: from theory to practice." Advances in Neural Information Processing Systems 36 (2023): 58833-58864.

Normalizing Flows

Normalizing Flows (NFs) are *density estimators* that transform distributions into each other, via a series of bijective transforms, $t_\theta^1 \circ \cdots \circ t_\theta^k = T_\theta$.

Every transform t_i needs to

- preserve the probability mass at every step $\forall l \in [1..K]. \int t_\theta^1 \circ \cdots \circ t_\theta^l(P) = 1.0$.
- keep values above (or equal to) 0.

$\rightarrow$ Every step, again, produces a valid density.

Commonly, distributions are transformed into a distribution of independent Gaussians: $T_\theta(\mathbf{x}) =: \mathbf{u} \sim P_\mathbf{u}$ where, e.g., $P_\mathbf{u} = \bigtimes_i \mathbb{N}_i(0, 1)$

Javaloy, Adrián, Pablo Sánchez-Martín, and Isabel Valera. "Causal normalizing flows: from theory to practice." Advances in Neural Information Processing Systems 36 (2023): 58833-58864.

Normalizing Flows

Normalizing Flows (NFs) are *density estimators* that transform distributions into each other, via a series of bijective transforms, $t_\theta^1 \circ \cdots \circ t_\theta^k = T_\theta$.

Every transform t_i needs to

- preserve the probability mass at every step $\forall l \in [1..K]. \int t_\theta^1 \circ \cdots \circ t_\theta^l(P) = 1.0$.
- keep values above (or equal to) 0.

$\rightarrow$ Every step, again, produces a valid density.

Commonly, distributions are transformed into a distribution of independent Gaussians: $T_\theta(\mathbf{x}) =: \mathbf{u} \sim P_\mathbf{u}$ where, e.g., $P_\mathbf{u} = \bigtimes_i \mathbb{N}_i(0,1)$

More generally: $\log p(\mathbf{x}) = \log p(T_\theta(\mathbf{x})) + \log|det(\nabla_\mathbf{x} T_\theta(\mathbf{x}))|$
with parameters θ learned via MLE.

Javaloy, Adrián, Pablo Sánchez-Martín, and Isabel Valera. "Causal normalizing flows: from theory to practice." Advances in Neural Information Processing Systems 36 (2023): 58833-58864.

Causal NFs - Counterfactuals

Having a model that allows us to infer **u** from some observation and intervene on the system directly allows us to compute counterfactuals:

Algorithm 2 Algorithm to sample from the counterfactual distribution, $P(\mathbf{x}^{\mathrm{cf}} \mid do(\mathrm{x}_i = \alpha), \mathbf{x}^{\mathrm{f}})$.

function GETCOUNTERFACTUAL($\mathbf{x}^{\mathrm{f}}$, i, α)
  $\mathbf{u} \leftarrow T_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{f}})$ ▷ Get $\mathbf{u}$ from the factual sample.
  $\mathrm{x}_i^{\mathrm{f}} \leftarrow \alpha$ ▷ Set x_i to the intervened value α.
  $\mathrm{u}_i \leftarrow T_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{f}})_i$ ▷ Change the i-th value of $\mathbf{u}$.
  $\mathbf{x}^{\mathrm{cf}} \leftarrow T_{\boldsymbol{\theta}}^{-1}(\mathbf{u})$
  return $\mathbf{x}^{\mathrm{cf}}$ ▷ Return the counterfactual value.
end function

Javaloy, Adrián, Pablo Sánchez-Martín, and Isabel Valera. "Causal normalizing flows: from theory to practice." Advances in Neural Information Processing Systems 36 (2023): 58833-58864.

CausalVAE

Process:

- Image $x \rightarrow$ Encoder $\rightarrow$ Exogenous Noise (ϵ)
- Exogenous Noise (ϵ) $\rightarrow$ Causal Layer $\rightarrow$ Causal Variables ($\mathbf{z}$)
- Causal Variables ($\mathbf{z}$) $\rightarrow$ Decoder $\rightarrow$ Reconstruction $\hat{x}$.

Yang, Mengyue, et al. "Causalvae: Disentangled representation learning via neural structural causal models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

CausalVAE

Assume the process to be a linear Structural Equation Model (SEM):

$$\mathbf{z} = A^T \mathbf{z} + \epsilon$$

A is a learnable weighted adjacency matrix, representing the graph. A mask is applied (e.g. multiply A by the binary adj matrix) to enforce the correct causal graph.

Solving for $\mathbf{z}$ gives:

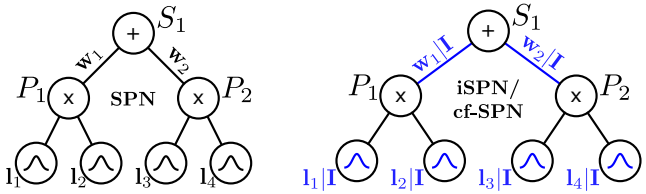
$$\mathbf{z} = (I - A^T)^{-1} \epsilon$$

Given the reconstructed $\mathbf{z}$, we can simply decode the image again: $\hat{\mathbf{x}} := d(\mathbf{z})$.

Interventional Sum-Product Network (iSPN)

SPNs can not compute causal queries without modifications.

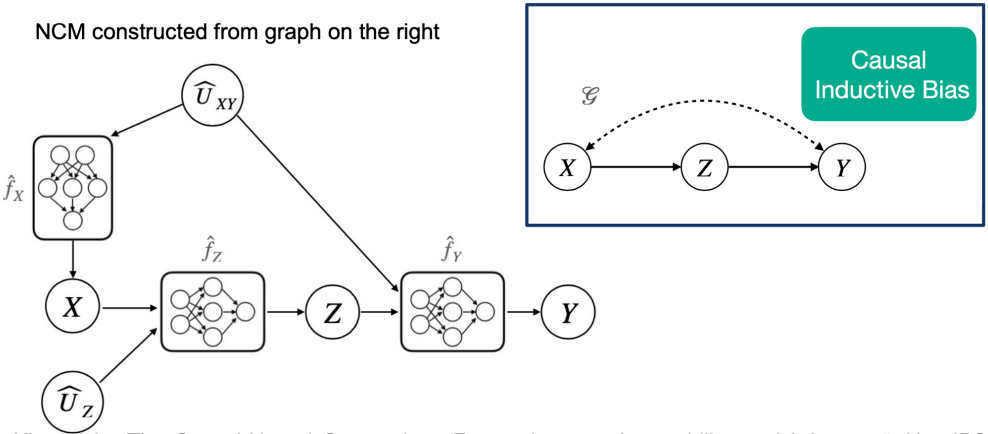
iSPN strategy: parametrize SPN parameter such that they can adapt to the interventional setting **I**.



A neural network is used to learn the mapping between **I** and the SPN parameters.

Zečević, Matej, et al. "Interventional sum-product networks: Causal inference with tractable probabilistic models." Advances in neural information processing systems 34 (2021): 15019-15031.

Example with 3 exogenous vars. and 1 known hidden confounder



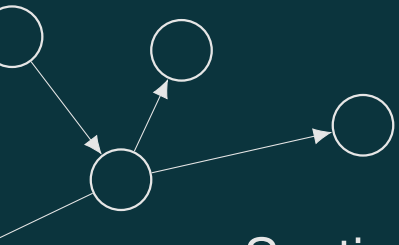
K. Xia et al. "The Causal-Neural Connection: Expressiveness, Learnability, and Inference." NeurIPS 2021.

How to learn an NCM, an algorithm:

Input: causal query $Q = P(y \mid do(x))$, L_1 data $P(v)$, and causal diagram $\mathcal{G}$
Output: $P^{\mathcal{M}^*}(y \mid do(x))$ if identifiable, FAIL otherwise

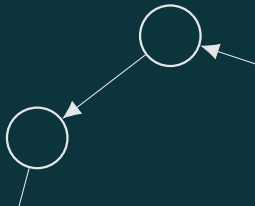
1. $\widehat{M} \leftarrow \mathrm{NCM}(V, \mathcal{G})$ // from Def. on previous slide
2. $\theta^*_{\min} \leftarrow \arg\min_\theta P^{\widehat{M}(\theta)}(y \mid do(x))$ s.t. $L_1(\widehat{M}(\theta)) = P(v)$
3. $\theta^*_{\max} \leftarrow \arg\max_\theta P^{\widehat{M}(\theta)}(y \mid do(x))$ s.t. $L_1(\widehat{M}(\theta)) = P(v)$
4. **if** $P^{\widehat{M}(\theta^*_{\min})}(y \mid do(x)) \neq P^{\widehat{M}(\theta^*_{\max})}(y \mid do(x))$ **then**
5. **return** FAIL
6. **else**
7. **return** $P^{\widehat{M}(\theta^*_{\min})}(y \mid do(x))$ // choose min or max arbitrarily

K. Xia et al. "The Causal-Neural Connection: Expressiveness, Learnability, and Inference." NeurIPS 2021.



Section 9

Lecture 9: Causal Representation Learning



Causality in Images

“What is the causality in this image?”

Causality in Images

Humans have an intuitive ‘high-level’ understanding of the processes depicted in the images.

→ We leverage temporal cues and interventions to disentangle concepts and make sense of the world.

→ We predict underlying dynamics and make inferences about future outcomes.

How can we teach machines to infer the same (causal) dynamics?

Hierarchy of Causal Tasks

Different tasks in causality operate under different knowledge levels:

Task	Variables Known?	Edges Known?
Effect Identification	✓	✓
Causal Discovery	✓	✗
Causal Representation Learning	✗	✗

Factors of Variation

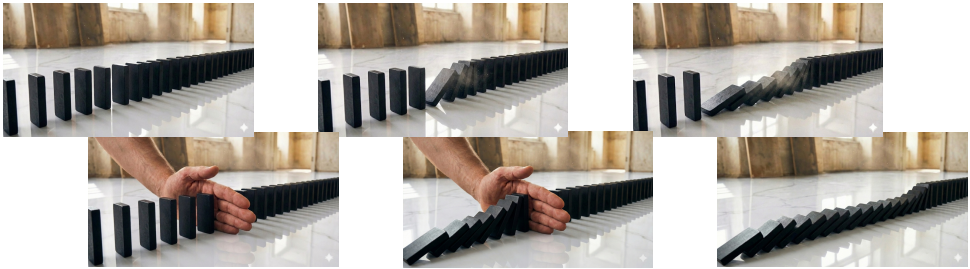
"What's a cause anyway?"

→ A constant can never be inferred to be a cause (nor an effect).

We need to show our models enough *variation* to learn the correct causal relations!

What does this mean for the single images shown before?

→ A model either needs *prior knowledge* or be presented with a *population/series* of images to interpret those.



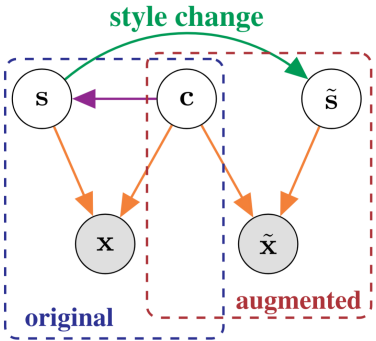
Content-Style Separation

Can we guarantee disentanglement between content and style?

(1) Create pairs of images: $(\mathbf{x}, \tilde{\mathbf{x}})$ with shared content $\mathbf{c}$ but different styles $\mathbf{s}, \tilde{\mathbf{s}}$.

(2) Train with generative self-supervised learning (SSL), e.g. an VAE with contrastive learning.

Intuition: The model learns that 'content' is whatever remains constant when applying augmentations.



Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M. and Locatello, F., 2021. Self-supervised learning with data augmentations provably isolates content from style. Advances in neural information processing systems, 34, pp.16451-16467.

Latent Causal Process Discovery

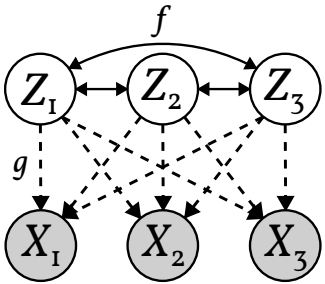
Task of CRL: Recover the latent factors $\mathbf{z} \in \mathbb{R}^n$ (and graph) from an observed signal $\mathbf{x} \in \mathbb{R}^k$.

Assume a **latent process**:

$$\mathbf{z}_i := f_i(\{\mathbf{z}_j \mid \mathbf{z}_j \in \text{pa}(\mathbf{z}_i)\}, \epsilon_i)$$

and a **mixing function**:

$$\mathbf{x}_i := g_i(\mathbf{z})$$



In the simplest form g is an (invertible) linear mixing function, $\mathbf{A} \in R^{k \times n}$:

$$\mathbf{x} = \mathbf{A}\,\mathbf{z}$$

The general case is underdetermined $\rightarrow$ The latent structure is unidentifiable!

Temporal Causal Representation Learning

Task of BSS: Recover the latent factors $\mathbf{z} \in \mathbb{R}^n$ (and graph) from an observed signal $\mathbf{x} \in \mathbb{R}^k$ in a timeseries.

Assume an underlying process:

Latent process: $\mathbf{z}_{it} := f_i(\text{pa}(\mathbf{z}_{it}), \epsilon_{it})$
where $\text{pa}(\mathbf{z}_{it}) \subseteq \mathbf{z}$

Mixing function: $\mathbf{x}_{it} := g_i(\mathbf{z}_t)$

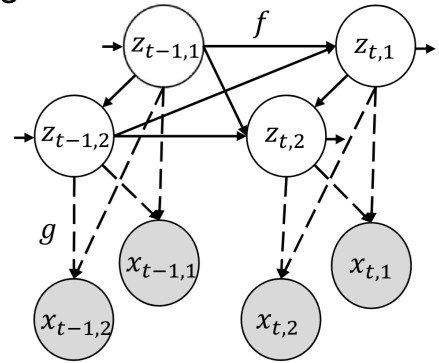


Figure adapted from: Li, Z., Shen, Y., Zheng, K., Cai, R., Song, X., Gong, M., Chen, G. and Zhang, K., On the Identification of Temporal Causal Representation with Instantaneous Dependence. In The Thirteenth International Conference on Learning Representations.

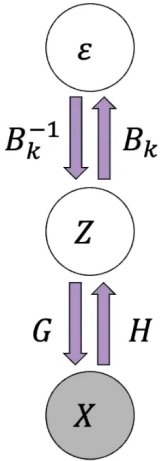
Linear Mixing with Interventions

$\text{Cov}(\epsilon)^{-1} = I_d$
$\text{Cov}_k(Z)^{-1} = B_k^T B_k$
$\Theta_k = \text{Cov}_k(X)^\dagger = H^T B_k^T B_k H$

Input: $\Theta_0, \ldots, \Theta_K$
Task: Recover H and identify all $B_0, \ldots, B_K$.

Squires et al., 2023

One intervention per latent node is sufficient (and in the worst case **necessary**) to recover $H = G^\dagger$ and $B_0, \ldots, B_K$.



Squires, C., Seigal, A., Bhate, S.S. and Uhler, C., 2023, July. Linear causal disentanglement via interventions. In International conference on machine learning (pp. 32540-32560). PMLR.
Figure: https://www.youtube.com/watch?v=ukhEEapKM-E

Temporal Causal Representation Learning

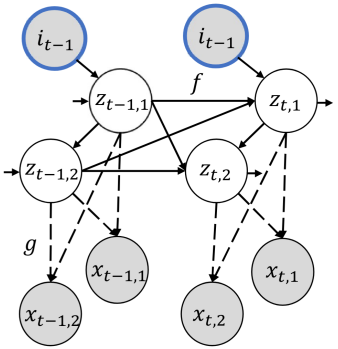
Task of CRL: Recover the latent factors $\mathbf{z} \in \mathbb{R}^n$ (and graph) from an observed signal $\mathbf{x} \in \mathbb{R}^k$ in a timeseries under interventions.

Assume an underlying process:

Latent process: $\mathbf{z}_{it} := f_i(\text{pa}(\mathbf{z}_{it}), \epsilon_{it})$
where $\text{pa}(\mathbf{z}_{it}) \subseteq \mathbf{z}$

Mixing function: $\mathbf{x}_{it} := g_i(\mathbf{z}_t)$

Interventions: $I^t = \{i_1^t, \dots\}$



iCRITIS [1]

Figure: Li, Z., Shen, Y., Zheng, K., Cai, R., Song, X., Gong, M., Chen, G. and Zhang, K., On the Identification of Temporal Causal Representation with Instantaneous Dependence. In The Thirteenth International Conference on Learning Representations.

[1] Lippe, P., Magliacane, S., Löwe, S., Asano, Y.M., Cohen, T. and Gavves, E., Causal Representation Learning for Instantaneous and Temporal Effects in Interactive Systems. In The Eleventh International Conference on Learning Representations.

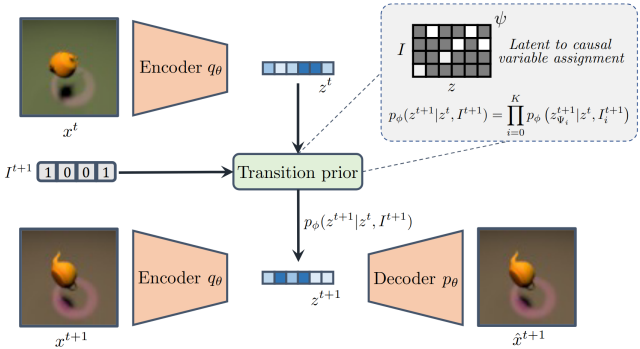
CITRIS: Causal Identifiability from Temporal Intervened Sequences

Loss Term:

1. Reconstruction loss.
2. Transition dynamics between time steps.

Assumes intervention targets are observed!

Assumes multi-valued variables. Requires mapping $\psi : I \to C$.



Lippe, P., Magliacane, S., Löwe, S., Asano, Y.M., Cohen, T. and Gavves, E., Causal Representation Learning for Instantaneous and Temporal Effects in Interactive Systems. In The Eleventh International Conference on Learning Representations.

Temporal Causal Representation Learning

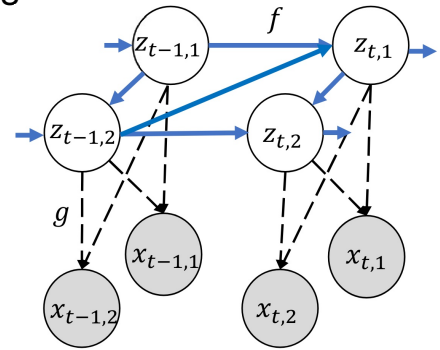
Task of CRL: Recover the latent factors $\mathbf{z} \in \mathbb{R}^n$ from an observed signal $\mathbf{x} \in \mathbb{R}^k$.

Assume an underlying process:

Latent process: $\mathbf{z}_{it} := f_i(\text{pa}(\mathbf{z}_{it}), \epsilon_{it})$
where $\text{pa}(\mathbf{z}_{it}) \subseteq \mathbf{z}$

Mixing function: $\mathbf{x}_{it} := g_i(\mathbf{z}_t)$

Sparse Latent Process: *(blue edges)*

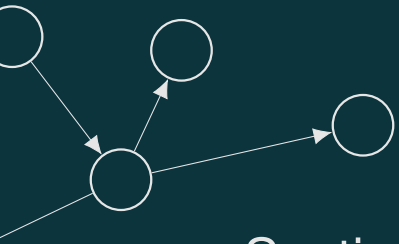


IDOL [1]

Idea: Restrict data in a way that makes the structure become identifiable.

1. Enforce independencies of the shown graph.
2. Require further variability to make the system have a unique solution.

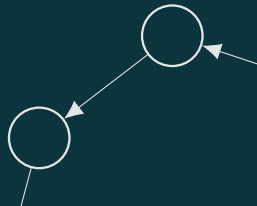
[1] Li, Z., Shen, Y., Zheng, K., Cai, R., Song, X., Gong, M., Chen, G. and Zhang, K., On the Identification of Temporal Causal Representation with Instantaneous Dependence. In The Thirteenth International Conference on Learning Representations.



Section

10

Lecture 10: Causality & LLMs



Why Natural Language?

Natural Language (NL) is a common way to communicate (causal) knowledge.

→ Availability of large and diverse sets of textual data.

→ Texts contain observational, interventional and counterfactual records.

→ Natural language enables explicit reasoning *over* bits of causal information.

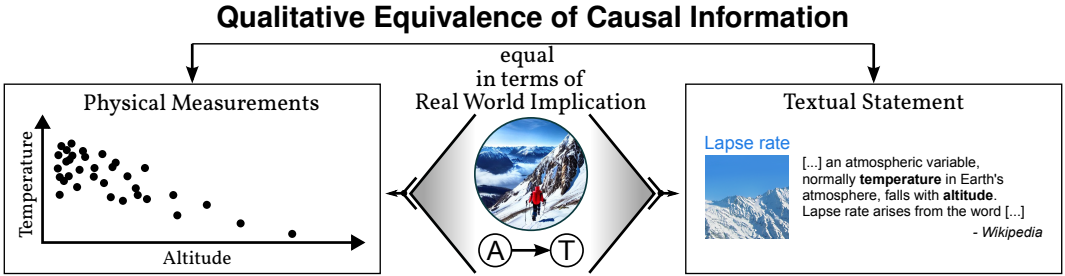
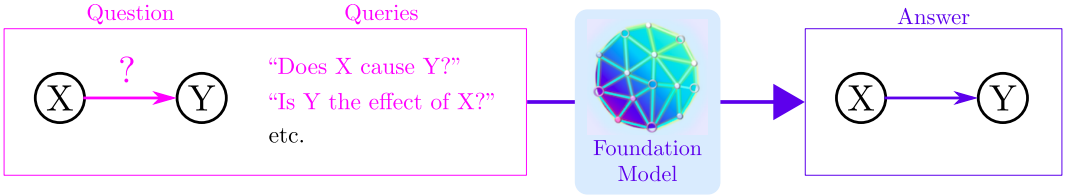


Figure: Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Motivation

LLM are commonly praised as universal tools to solve a variety of tasks.

→ Can they help us with causal reasoning and discovery?



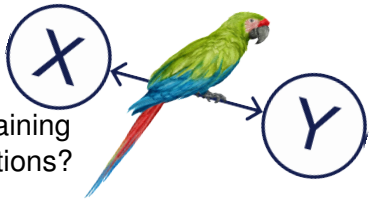
Con: LLMs are trained in an associational manner.

→ They never interact with the world.

→ Can they excel beyond the first rung of Pearls causal ladder?

Willig, M., Zečević, M., Dhami, D.S. and Kersting, K., Can Foundation Models Talk Causality?. In UAI 2022 Workshop on Causal Representation Learning.

Causal Parrots



LLMs might say “Smoking causes cancer”, because
they have seen that sentence thousands of times during training
. . . but do they ‘really understand’ the real-world implications?

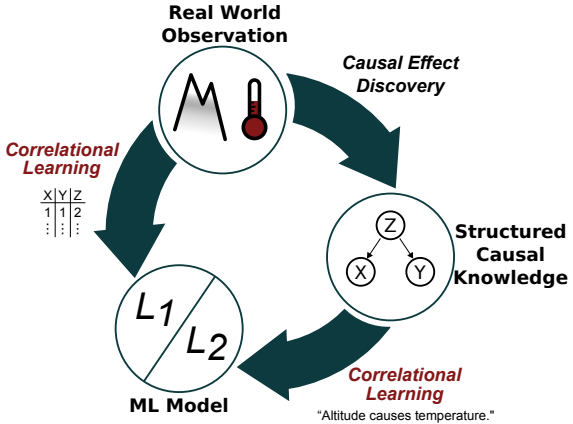
Still, the statement “Smoking causes cancer” is correct, but…
→ text is only a partial description of the real world.
→ models are not free to conduct experiments that break confounding.
→ this opens them up to various biases and fallacies.

Willig, M., Zečević, M., Dhami, D.S. and Kersting, K., Can Foundation Models Talk Causality?. In UAI 2022 Workshop on Causal Representation Learning.
Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Moving up the Ladder: Associational Learning of Causal Knowledge

Texts can contain records *about* interventions and their outcomes.

→ LLMs might use this information to learn about the underlying dynamics!



Temporal Event Ordering - Post-Hoc Fallacy

The Post-Hoc Fallacy:

LLMs frequently assume that because A came before B in the text, A caused B.

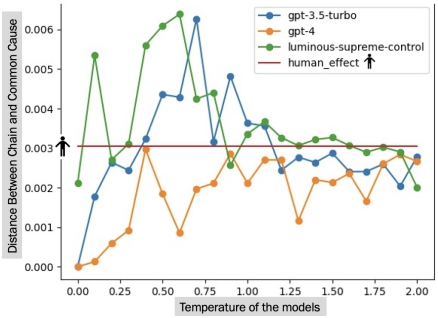
→ *Latin:* "post hoc ergo propter hoc" – "after this, therefore because of this."
→ Reversing event order in text deteriorates LLM's reasoning performance.

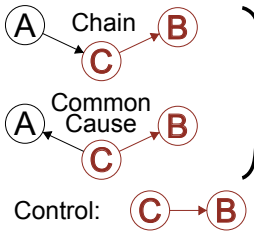
Data	Rel. position in train	Rel. position in eval (X, Y)	Rel. position in eval (Y, X)
causal $X \rightarrow Y$	(X, Y)	92.59%	1.85%
	(Y, X)	0%	100%

"Accuracy of finetuned on temporal relations with different relative event positions."

Joshi, N., Saparov, A., Wang, Y. and He, H., 2024, November. LLMs Are Prone to Fallacies in Causal Inference. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 10553-10569).

LLMs adopt Human Biases in Causal Perception





$$\mathbf{A} \perp \mathbf{B} | \mathbf{C}$$

$$P(\mathbf{A}, \mathbf{B}, \mathbf{C}) = P(\mathbf{B}|\mathbf{C})P(\mathbf{C}|\mathbf{A})P(\mathbf{A}) = P(\mathbf{A}|\mathbf{C})P(\mathbf{B}|\mathbf{C})P(\mathbf{C})$$

“LLM [...] attributing greater causal strength to the intermediate cause in canonical Chains than to the corresponding nodes in Common Cause. [...] With temperatures between 1.0 and 1.9, the observed preference for Chains is remarkably similar to that observed in humans across all three models.”

Keshmirian, A., Willig, M., Hemmatian, B., Kersting, K., Hahn, U. and Gerstenberg, T., 2024. Chain versus common cause: Biased causal strength judgments in humans and large language models. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 46).

Predicting Causal Graphs - Pairwise

The most simple approach of predicting causal graphs is to query LLMs for every variable pair $X_i, X_j \in \mathbf{V}$ and record the answers.

For every query

"What is the causal relationship between <X_i> and <X_j>."

there are 3 possible options:

A) *"<X_i> causes <X_j>."*

B) *"<X_j> causes <X_i>."*

C) *"No edge exists between X_i and X_j."*

The LLM is then tasked to select one of them.

Kiciman, E., Ness, R., Sharma, A. and Tan, C., 2023. Causal reasoning and large language models: Opening a new frontier for causality. Transactions on Machine Learning Research. Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Predicting Causal Graphs - Breadth-First Search

Pairwise querying between any two variables requires $\mathcal{O}(N^2)$ LLM queries.

$\rightarrow$ Can we be more efficient?

$\rightarrow$ Ask for the direct effects/parents of each variable!

Breadth-First Search (BFS) with LLMs:

1. Find and enqueue all root cause (=parent-less) variables in $\mathbf{V}$.
2. For every variable U in the queue:
   - 2.1 Identify all direct effects of U.
   - 2.2 Add all direct effect variables to the queue.
   - 2.3 Remove U from the list of possible direct effects.

$\rightarrow$ Only requires $\mathcal{O}(N)$ queries.

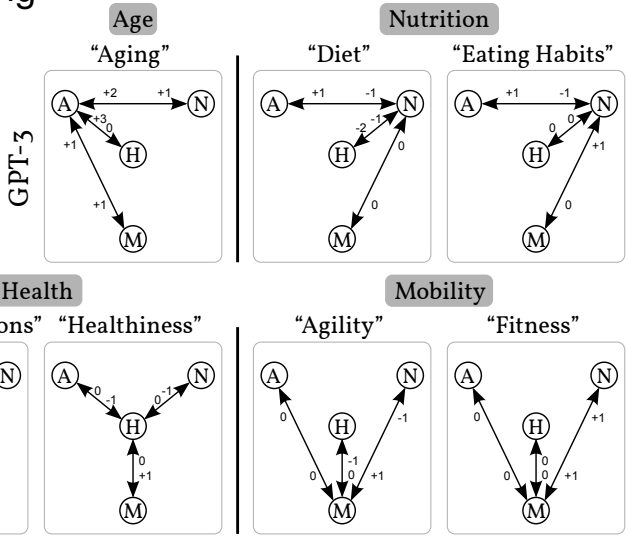
$\rightarrow$ Creates inherently acyclic graphs.

Jiralerspong, T., Chen, X., More, Y., Shah, V. and Bengio, Y., Efficient Causal Graph Discovery Using Large Language Models. In ICLR 2024 Workshop: How Far Are We From AGI.

Robustness to Variable Wording

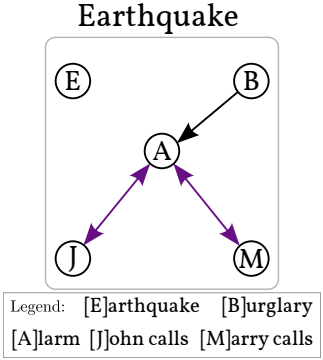
"Does <A> cause <B>?":
LLM answers might be susceptible to variable naming. E.g. “Aging” instead of “Age”.

Different wordings have different implications, or grounding in the training data.



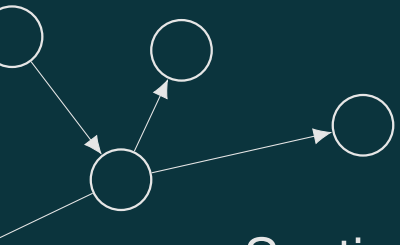
Context and Meta Answers

Context matters: *“Does the alarm cause John to call?”*



LLMs might have the freedom to divert from ‘yes’/‘no’ answers!

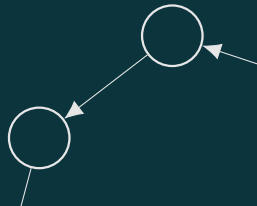
GPT-4: *“The text does not provide information on whether burglaries cause calls from John.”*



Section

11

Lecture 11: Bias & Fairness



Many more Biases…

Hindsight Bias Model Bias Association Bias
Participation Bias Selection Bias Sampling Bias Confirmation Bias
Measurement Bias Observation Bias Stereotyping
Survivorship Bias Gambler's fallacy Anchoring Bias
Gender Bias Outcome bias

Why Fairness?

Simply optimizing models for prediction accuracy might induce or carry over biases from existing data.

Conflict of interest:

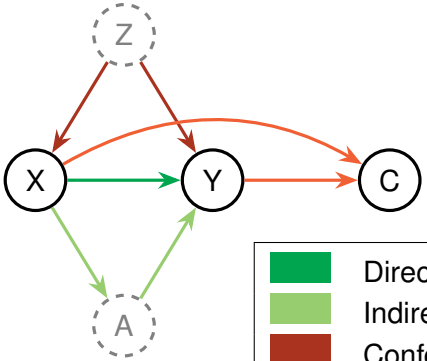
$\rightarrow$ "A bank could boost its profits by denying loans to poor people." ⚡

$\rightarrow$ Boosting performance $\not\leftrightarrow$ being fair

Build models that make accurate predictions while being fair!

Causal Effect Estimation

Remember causal effect estimation: identify the true (total) causal effect?
→ Correct for the influence of all non-causal paths in the estimation of $p(Y|do(X))$.



Failing to adjust for any of the
non-causal paths will bias the estimate!

- Direct Causal Effect ($X \rightarrow Y$)
- Indirect Causal Effect ($X \rightarrow A \rightarrow Y$)
- Confounding Path ($X \leftarrow Z \rightarrow Y$; *Correlational!*)
- Collider Path ($X \rightarrow C \leftarrow Y$; *Correlational!*)

Collider Bias / Berkson's paradox

Scenario: Collecting data for a curated "Top Photos" gallery ($Z = 1$). Images are selected for the gallery if they are either *high-resolution* or show a *rare bird species*.

Training a model on this dataset may learn a *spurious correlation*. The data may imply that "rarity" implies "low resolution".

Deploying the trained model in a real-world setting will likely flag many low-resolution images as rare birds.

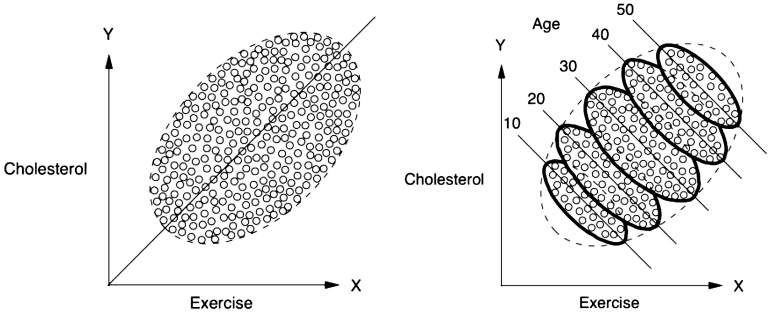
Common bird, high-resolution image.

Rare bird, low-resolution image.

Sparrow image: https://commons.wikimedia.org/wiki/File:House_Sparrow(Passer_domesticus).jpg

Simpson Paradox

Simply regressing the effect of *hours of exercise* on *cholesterol levels* over the whole population predicts a negative effect of exercise on cholesterol levels…



… however, adjusting by *age group* reveals the true *positive* causal effect.

Figure: Pearl, J., Mackenzie D., 2018. The book of why: the new science of cause and effect, Basic Books.

Modeling Decision Processes

A decision process for a particular task T is defined over the following spaces:

- **Construct space** $\mathcal{CS} = (P, d_P)$: space over individuals $p \in P$.
- **Observed space** $\mathcal{OS} = (\hat{P}, \hat{d})$:
  Actual recorded/observed data of individuals, $\hat{p} := g(p)$.
- **Decision Space** $\mathcal{DS} = (O, d_O)$:
  Space representing the actual outcomes, $o := t(p)$.

Decision space	*Construct space*	*Observed space*
Performance in college	Success in High School	GPA
Employee Productivity	Knowledge of job	Number of Years of Experience

Challenge: Properties of the construct space are commonly not directly observed.
$\rightarrow$ Properties of the observed space are used as proxies instead.

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. "Communications of the ACM 64.4 (2021): 136-143.

Protected Attributes

Some attributes might be *protected*.

$\rightarrow$ Protected attributes ($\mathbf{A} \subset \mathbf{X}$) must not be used for making predictions!

Protected attributes might include:

1. Age
2. Cultural/ethnic background
3. Sexual orientation
4. Religion or belief
5. ...

They are usually defined by ethical or legal considerations.

We would like our predictions $\hat{Y}$ of some target label $Y \in \mathbf{X}$ to be invariant to $\mathbf{A}$.

Demographic Parity

Simple idea:

An outcome should be equally likely, independent of the protected attributes.

Demographic Parity

$$\forall a \in \mathcal{A}.\ P(\hat{Y}|A = a) = P(\hat{Y}|A = a')$$

Example:

"Employ people independent of their parental status."
"Grant loans to all ethnic groups equally."
"Hire people independent of their religion."
…

Equality of Opportunity

We would like to have balance the percentage of positive outcomes (same *True Positive Rate*) for all $a \in \mathcal{A}$.

Equality of opportunity

$$\forall a \in \mathcal{A}.\ P(\hat{Y} | Y = 1, A = a) = P(\hat{Y} | Y = 1, A = a')$$

In plain words: "If two students are both qualified for a course ($Y = 1$), do they have the same chance of being approved ($\hat{Y} = 1$) regardless of their origin or gender?"

Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. Advances in neural information processing systems, 29.

Equalized odds

Idea: Enforce an equal true-positive *and false-positive rate* for all values of A.

Equalized odds

$$\forall a \in A, y \in \{0, 1\}.\ P(\hat{Y} | Y = y, A = a) = P(\hat{Y} | Y = y, A = a')$$

Counterfactual Fairness

Idea: Require the same potential outcome for individuals, independent of $\mathbf{A}$.

Counterfactual Fairness

$$P(\hat{Y}_{\mathbf{A}=\mathbf{a}}(\mathbf{U}) | \mathbf{V}=\mathbf{v}, \mathbf{A}=\mathbf{a}) = P(\hat{Y}_{\mathbf{A}=\mathbf{a}'}(\mathbf{U}) | \mathbf{V}=\mathbf{v}, \mathbf{A}=\mathbf{a})$$

Example: People with the same qualifications should be given the same opportunity for college admission, independent of their protected attributes.

Counterfactual Fairness:

$\rightarrow$ compares outcomes *per individual*.

$\rightarrow$ does not depend on some (possibly biased) ground thruth label Y.

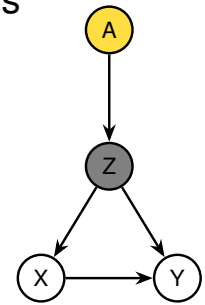
Kusner, M.J., Loftus, J., Russell, C. and Silva, R., 2017. Counterfactual fairness. Advances in neural information processing systems, 30.

Counterfactual Fairness in Graphical Models and NNs

Given causal structural knowledge, one can correct for the effects of the protected attributes via $P(\hat{Y}|do(X))$ and $X \perp\!\!\!\perp \mathbf{A}$.

$\rightarrow$ This, however, requires graph specific estimands, e.g.:

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

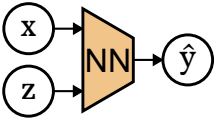


Neural models commonly correspond to a single, joint regression function: $\hat{Y} := f(X, Z)$.

$\rightarrow$ Is oblivious to the causal graphical structure.

$\rightarrow$ Does not differentiate between causes X and the conditioning set Z.

$\rightarrow$ NN do not approximate the true causal estimand. ⚡



Counterfactual Fairness for ML Training

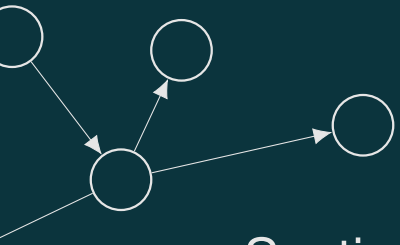
Lemma 1 [Kusner et al, 2017].

Let $\mathcal{G}$ be the causal graph of a given model $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, \mathcal{P}_{\mathbf{U}})$. Then $\hat{Y}$ will be counterfactually fair if it is a function of the non-descendants of A.

$\rightarrow$ Provides a condition which attributes can be safely used for model training.

Proof (informal): For any non-descendants $\mathbf{W}$ we infer the same counterfactual values $\mathbf{W}^*_{A \leftarrow a}$ since they are independent from $\mathbf{A}$. Hence, using $\mathbf{W}$ for training our classifier $\hat{Y}$ is invariant to the counterfactual values of $\mathbf{A}$.

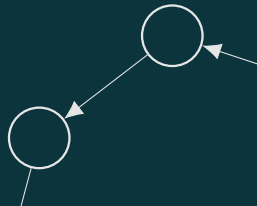
Kusner, M.J., Loftus, J., Russell, C. and Silva, R., 2017. Counterfactual fairness. Advances in neural information processing systems, 30.



Section

12

Lecture 12: Actual Causality



The Need for Actual Causality

Problem: *do*-calculus tells us if X affects Y *in general*.

- Sometimes called '*type causality*' (or '*general causality*').
  $\rightarrow$ In general: "smoking causes cancer".
  $\rightarrow$ Can be used for making predictions / forward-looking.

Commonly does not reason about the *actual* outcome.

$\rightarrow$ 'The fact that Bob smoked for 30 years, did cause his lung cancer.'
$\rightarrow$ 'The fact that a match was dropped, was not responsible for the forest fire.'

Actual causality talks about specific instances/scenarios.

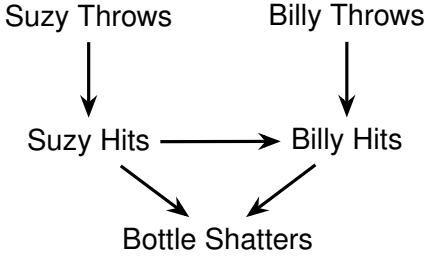
- Sometimes called '*token causality*' or '*specific causality*'.
  $\rightarrow$ Token causality is about explanations / backward-looking.

Causal Models - Rock Throwing

$$\mathcal{M} = (\mathcal{S}, \mathcal{F})$$

$$\mathcal{S} = \begin{cases} \mathcal{U} & = \{U_{ST}, U_{BT}\} \\ \mathcal{V} & = \{ST, BT, SH, BH, BS\} \\ \mathcal{R}(Y) & = \{0, 1\} \text{ for all } Y \in (\mathcal{V} \cup \mathcal{U}) \end{cases}$$

$$\mathcal{F} = \begin{cases} ST & = U_{ST} \\ BT & = U_{BT} \\ SH & = ST \\ BH & = BT \wedge \neg SH \\ BS & = SH \vee BH \end{cases}$$



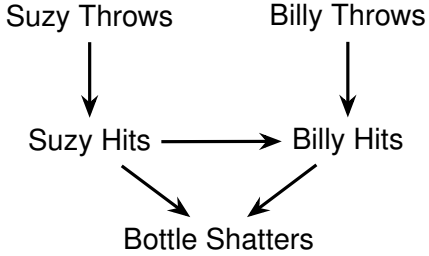
Actual World:

$u = \{U_{ST} = 1, U_{BT} = 1\}$

Entailment in Causal Models - Rock Throwing

With $\mathcal{M}$ as defined before and actual world $u = \{U_{ST} = 1, U_{BT} = 1\}$:

$$(\mathcal{M}, u) \models \underbrace{(ST = 1) \wedge (BT = 1) \wedge}_{\text{Actual World}} \underbrace{(SH = 1) \wedge (BH = 0) \wedge (BS = 1)}_{\text{Entailment}}$$



Interventions: $(\mathcal{M}, \{U_{ST} = 0, U_{BT} = 1\}) \models [BH \leftarrow 0]BS = 0$

"Under context u and intervention do($BH = 0$), BS takes value 0 in model $\mathcal{M}$."

Intervention Semantics:

$$(\mathcal{M}, u) \models [Y \leftarrow y]\psi \text{ iff } (\mathcal{M}^{Y \leftarrow y}, u) \models \psi$$

The Modified HP Definition (mHP)

Modified HP Definition

A primitive event $\vec{X} = \vec{x}$ is an **actual cause** of φ in (M, u) if:

AC1 $(\mathcal{M}, u) \models (\vec{X} = \vec{x})$ and $(\mathcal{M}, u) \models \varphi$.
'The cause and the effect must actually happen.'

AC2 $(\mathcal{M}, u) \models [X \leftarrow x', \vec{W} \leftarrow \vec{w}^*]\neg\varphi$
x' must change φ, while $\vec{w}^*$ must take the values of the actual world.

AC3 $\vec{X}$ is minimal.
No subset of $\vec{X}$ satisfies AC1 and AC2.

Root Causes: In ‘standard’ SCM

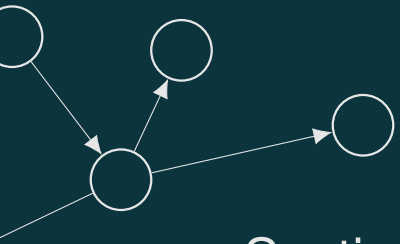
Finding Root Causes $\mathbf{U}_{rc}$:

1.) A standard SCM $\mathcal{M}$ entails an expected distribution $P(\mathbf{V})$ over variables $\mathbf{V}$.
$\rightarrow$ For every observed vector $\mathbf{v}$, we can compute $P(\mathbf{v})$ according to $\mathcal{M}$.

2.) If for some $V \in \mathbf{V}$, $P(v)$ drops below a certain threshold we consider them out of distribution.

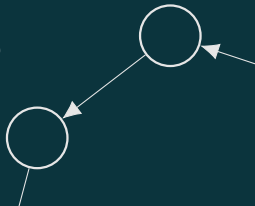
3.) Structural equations induce a form of ‘normality’.
$\rightarrow$ Consider the conditional probability $p(v_i|pa(V_i))$.
$\rightarrow$ It is *high*, if the observed v_i follows the structural equation.
$\rightarrow$ It is *low*, if the observed v_i diverts from f_i due to noise.

- Marginal probabilities $P(v_i)$ let us detect anomalous variable values.
- Conditional probabilities $p(v_i|pa(V_i))$ tell us the locations of outlier noise.



Section 13

Lecture 13: Causal Perspectives



What Defines a Fair Attribution?

1. Efficiency. The value of the whole coalition is distributed to the individual actors:

$$\sum_{i \in N} \varphi_i(v) = v(N)$$

2. Symmetry. Actors i, j that contribute equally should receive equal attribution:

$$\forall S \in (N \setminus \{i,j\}).(v(S \cup \{i\}) = v(S \cup \{j\})) \Rightarrow (\varphi_i(v) = \varphi_j(v))$$

3. Additivity/Linearity.: When a game can be decomposed into two independent value functions v, w, attribution should behave linearly:

$$\varphi_i(v + w) = \varphi_i(v) + \varphi_i(w)$$

4. Dummy/Null Player. An actor i that does not contribute to any coalition gets zero attribution:

$$(\forall S \in N.v(S \cup \{i\}) = v(S)) \Rightarrow (\varphi_i(v) = 0)$$

Shapley Values

An actor joining a coalition improves reward by $\Delta v(S, i) = v(S \cup \{i\}) - v(S)$, assuming that $v(S \cup \{i\}) \geq 0$.

Shapley Value for a player i in a coalition game (v, N):

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \Delta v(S, i) = \frac{1}{|N|} \sum_{S \subseteq N \setminus \{i\}} \binom{|N| - 1}{|S|}^{-1} \Delta v(S, i)$$

Interpretation

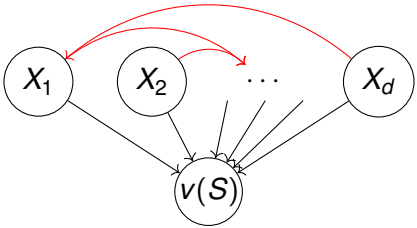
1. Consider actors joining a coalition S one-by-one.
2. Every player takes its respective reward $v(S \cup \{i\}) - v(S)$.
3. Shapley values compute the average contribution to every possible way (permutation) that the respective coalition could have been formed.

Shapley values are the only solution that fulfill all properties on efficiency, symmetry, additivity and null players.

Shapley, Lloyd S. "A value for n-person games." (1953): 307-317.

Considerations

Independence: Standard SHAP often assumes feature independence:



What if individual factors cause each other: education $\rightarrow$ income $\rightarrow$ buying power.

Spurious Explanations: SHAP might attribute all influence to ‘income’, but miss the stronger total causal effects of ‘education’.

Off-Manifold Estimates: Sampling features independently can lead to “unrealistic” data points that the model never saw during training.

Interventional Expectations

Causal Shapley Values replace the conditional expectation with the *interventional expectation* using the *do*-operator.

→ "Explain *mechanisms* rather than just *data statistics*."

$$v_{causal}(S) = \mathbb{E}[f(x) \mid do(X_S = x_S)] = \int f(x_S, x_{\bar{S}}) \cdot \underbrace{p(x_{\bar{S}} \mid do(x_S))}_{\text{Interventional Density}} \, dx_{\bar{S}}$$

Distinction:

- **Conditional:** $P(y \mid x)$. Propagates information up and down the causal graph (allows for confounding and mediators).
- **Interventional:** $P(y \mid do(x))$. Cuts upstream links. Only propagates effects to descendants.

Heskes, T., Sijben, E., Bucur, I.G. and Claassen, T., 2020. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. Advances in neural informationprocessing systems, 33, pp.4778-4789.

Incorporating Structure

Some permutations of actors i, j, k joining a coalition $\{i, j, k\}$ might violate the underlying causal graph.

- If $x_i \rightarrow x_j \leftarrow x_k$, then j joining before i and k might not be allowed.

If the DAG is known we can use **Asymmetric Shapley Values**.

Instead of all permutations π being weighted equally $w(\pi) = 1/(N!)$, only consider permutations that **respect the causal ordering** of the DAG.

$\rightarrow$ Incompatible orderings receive zero weight.

$$\leq_{\mathcal{G}} \subseteq \pi_S \Leftrightarrow w(\pi_S) \neq 0$$

Frye, C., Rowat, C. and Feige, I., 2020. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. Advances in neural information processing systems, 33, pp.1229-1239.

The Fundamental Problem of Causal Inference

In the Potential Outcomes framework, every individual i has two 'potential' outcomes of their future:

- $Y_i(1)$: The outcome if individual i receives the treatment ($D = 1$).
- $Y_i(0)$: The outcome if individual i does *not* receive the treatment ($D = 0$).

$\rightarrow$ *"The drug was either administered to the patient ($D = 1$) or not ($D = 0$)."*

We only ever observe either $Y_i(0)$ or $Y_i(1)$:

$$Y_i = \begin{cases} Y_i(0) & \textbf{if } D = 0 \\ Y_i(1) & \textbf{otherwise.} \end{cases}$$

The unobserved outcome is the **counterfactual**.

Causal inference is essentially framed as a 'missing data' problem.

Assumptions of PO

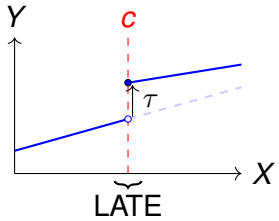
To estimate ATE from observational data, the setting must satisfy:

1. **SUTVA:** "Stable Unit Treatment Value Assumption".
   "One units treatment does not affect another's outcome."
2. **Ignorability (Unconfoundedness):** $\{Y(0), Y(1)\} \perp D|X$.
   "Given covariates X, the treatment assignment is as good as random."
3. **Positivity (Overlap):** $0 < P(D = 1|X) < 1$.
   "Every person has a non-zero chance of being in either group."

Regression Discontinuity Design

Assumption: Units just above and just below the cutoff are effectively 'randomized'.
→ For units extremely close to the cutoff (e.g., 69.9 and 70.1 test score), their position is determined by idiosyncratic noise.
→ RDD estimates the *Local Average Treatment Effect* (**LATE**) near the cut-off.

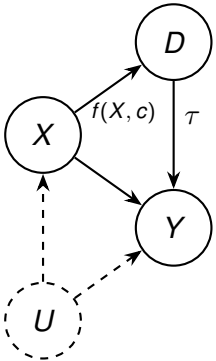
- **Running Variable (X):** The score determining assignment (e.g., Test Score).
- **Outcome (Y):** Long-term performance of the student.
- **Cutoff (c):** The threshold (e.g., Score ≥ 70).
- **Treatment (D):** Assigned if $X \geq c$.

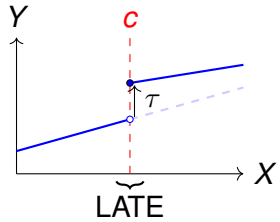


Local Average Treatment Effect (LATE) at the cutoff c:

$$\tau = \lim_{x \downarrow c} E[Y_i \mid X_i = x] - \lim_{x \uparrow c} E[Y_i \mid X_i = x]$$

Graphical Perspective



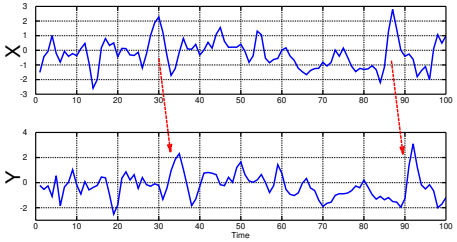


Insight:

U does not point to D directly. By conditioning on X (at the limit $X \to c$), we block the back-door path through U.

Granger Causality: Probabilistic Perspective

"Does X_i provide unique information about Y?"



Select feature X_i if $P(Y^{t+1} | \mathbf{X}^t) \neq P(Y^{t+1} | \mathbf{X}^t \setminus X_i^t)$

Figure: https://en.wikipedia.org/wiki/File:GrangerCausalityIllustration.svg, Creative Commons Attribution-Share Alike 3.0 Unported license.

Autoregression

"Does some X_i provide unique information about Y?"

Consider the time series X_i^t and Y^t.

Compare prediction error of two linear autoregressive models:

Baseline (Y onto itself):

$$Y^t = \sum_{i=1}^{k} \alpha_i Y^{t-i} + \epsilon_1^t$$

Extended Model (with additional X_i):

$$Y^t = \sum_{i=1}^{k} \alpha_i Y^{t-i} + \sum_{i=1}^{k} \beta_i X^{t-i} + \epsilon_2^t$$

X_i Granger-causes Y if the variance of the prediction error $\sigma^2(\epsilon_2)$ is *significantly smaller* than that of $\sigma^2(\epsilon_1)$.

$\rightarrow$ Commonly determined via an F-Test.

Downsides of Granger Causality

Granger Causality is a prediction driven notion of causality.

Easy to apply, but
does not handle confounders!

'Granger causality does *not* imply that if X is intervened, Y will change.'

Why Cycles?

Most of this course assumed **Acyclic** Causal Models. But in nature we often encounter feedback cycles:

- **Biological Pathways:** Gene regulation networks where A inhibits B and B inhibits A.
- **Economic Equilibria:** Price affects Demand, which in turn affects Price.
- **Climate Systems:** Surface albedo and temperature create self-reinforcing cycles.

Problem: If $X \to Y$ and $Y \to X$ exist, the standard Markov Property and topological ordering of DAGs break down.

$\to$ We move from *recursive* models (step-by-step generation) to *non-recursive* models (simultaneous constraints).

Converging Systems

Where do cyclic SCMs actually come from?

→ Cycles can be framed as projections of **Ordinary Differential Equations**.

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}(t), \mathbf{u}, t)$$

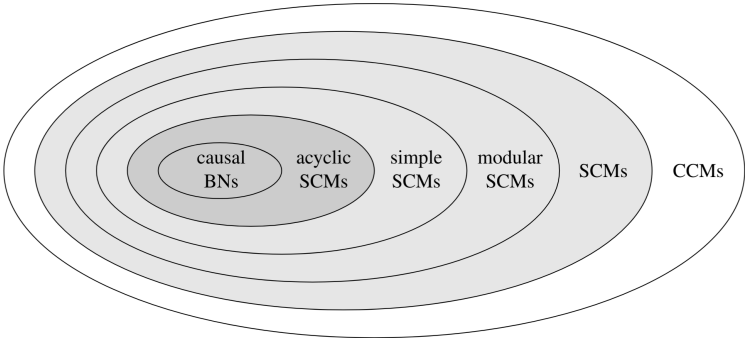
Equilibrium Points: If the system reaches a steady state ($\frac{d\mathbf{x}}{dt} = 0$), the differential equations collapse into Structural Equations.

$$0 = f(\mathbf{x}^*, \mathbf{u})$$

→ The fixed point $\mathbf{x}^*$ 'solves' the equations of $\mathcal{M}$.

Bongers, S. and Mooij, J.M., 2018. From random differential equations to structural causal models: The stochastic case. arXiv preprint arXiv:1803.08784, 3.
Iwasaki, Y. and Simon, H.A., 1994. Causality and model abstraction. Artificial intelligence, 67(1), pp.143-194.

Hierarchy of Structural Causal Models



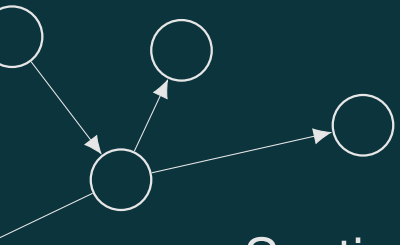
Simple SCM are uniquely solvable for (subsets of) the endogenous variables.

Modular SCM are closed under marginalization and intervention. Yield unique solutions under subsets of variables.

SCMs: Arbitrary structural equations between variables.

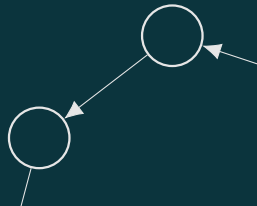
Constraint Causal Models: Contain 'instantaneous' equality constraints among variables.

Figure: Bongers, S., Forré, P., Peters, J. and Mooij, J.M., 2021. Foundations of structural causal models with cycles and latent variables. The Annals of Statistics, 49(5), pp.2885-2915.



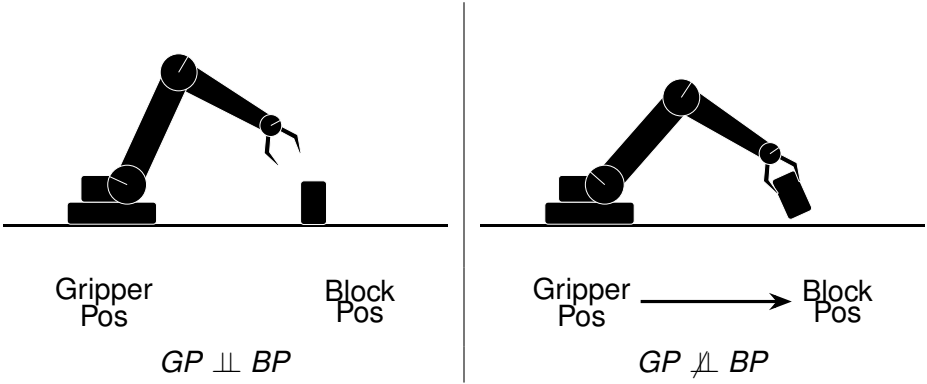
Section 14

Lecture 14: Meta-Causal Models



Contextual Independence: Robot Arm

(In)dependencies can materialize under more complex conditions:

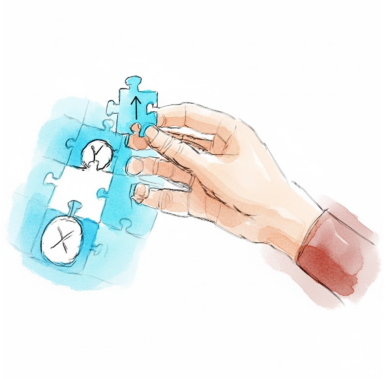


$$\rightarrow GP \perp\!\!\!\perp BP \mid (||GP, BP|| > 0)$$

Figure inspired by: Seitzer, M., Schölkopf, B. and Martius, G., 2021. Causal influence detection for improving efficiency in reinforcement learning. Advances in Neural Information Processing Systems, 34, pp.22905-22918.

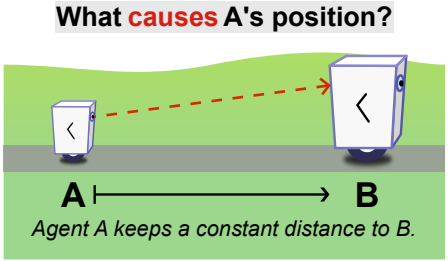
“Causal Understanding”

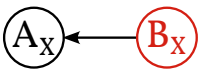
We would like to have a framework that allows us to reason about and manipulate causal relations in dynamic environments.



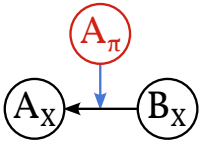
- Predict the stability of causal links.
- Reason over system dynamics.
- Attribution beyond static causal graphs.

Attributing Responsibility: A Meta-Causal Perspective





Classical Attribution
A_X is caused by the structural equation $A_X := f(B_X)$.



Meta-Causal Attribution
But the relation $B_X \to A_X$ only *exists* due to A's policy A_π.

Meta-Causality considers factors that lead to the emergence of causal edges.

Willig, M., Tobiasch, T., Busch, F.P., Seng, J., Dhami, D.S. and Kersting, K., Systems with Switching Causal Relations: A Meta-Causal Perspective. In The Thirteenth International Conference on Learning Representations.

Factors of Change: Meta-Causal Variables

$$\mathbf{C} := \{X_k \in \mathbf{X} \mid \exists X_i, X_j \in \mathbf{X}.\, \exists \mathbf{x}, \mathbf{x}' \in \boldsymbol{\mathcal{X}} \text{ s.t.}$$
$$(\mathbf{x}_{\bar{k}} = \mathbf{x}'_{\bar{k}}) \ \wedge \ (x_k \neq x'_k) \wedge (\mathcal{I}(\mathbf{x}, X_i, X_j) \neq \mathcal{I}(\mathbf{x}', X_i, X_j))\}$$

"With all other variables fixed, a different value of X_k can induce a different edge type"

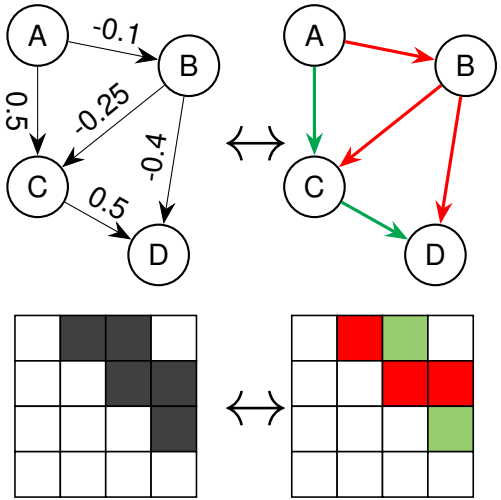
Willig, M., Tobiasch, T., Dhami, D.S. and Kersting, K., When Causal Dynamics Matter: Adapting Causal Strategies through Meta-Aware Interventions. In The Thirty-ninth Annual Conference on Neural Information Processing Systems.

Qualitative Causal Relations

In every-day life we communicate *qualitative* properties of causal relations.

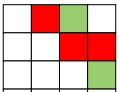
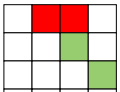
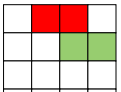
→ Abstract away from specific structural equations.

- "Temperature increases with altitude."
- "Caffeine intake increases alertness, but deprives sleep quality."
- "Time spent studying, improves exam performance, but reduces spare time."
- "Regulations reduce company gains, but improve environmental impact."
- …



Meta-Causal States

A **Meta-Causal State** (MCS) is a particular configuration of a qualitative causal graph.

$T_1=$  , $T_2=$  , $T_3=$  , $\cdots$ $\quad T_i \in \mathcal{T}^{N \times N}$

The system follows an underlying mediation process $\mathcal{E} = (\mathcal{S}, \sigma)$ where

- $\mathcal{S}$ is the domain of the process.
- $\sigma : \mathcal{S} \to \mathcal{S}$ is the transition function (e.g. a Markov process).

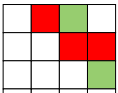
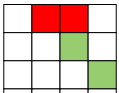
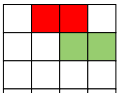
Causal variables **X** are identified via a causal abstraction: $\varphi : \mathcal{S} \to \boldsymbol{\mathcal{X}}$.

$\to$ Causal relations follow by abstracting the process $\varphi \circ \sigma$.

Willig, M., Tobiasch, T., Busch, F.P., Seng, J., Dhami, D.S. and Kersting, K., Systems with Switching Causal Relations: A Meta-Causal Perspective. In The Thirteenth International Conference on Learning Representations.

Meta-Causal State Identification

A **Meta-Causal State** is a particular configuration of a qualitative causal graph.

$T_1=$  , $T_2=$  , $T_3=$  , $\cdots$ $\quad T_i \in \mathcal{T}^{N \times N}$

Types $T_{s,ij}$ are identified via the **Identification Function** $\mathcal{I} : \mathcal{S} \times \mathbf{X} \times \mathbf{X} \to \mathcal{T}$,
with $T_{s,ij} := \mathcal{I}(s, X_i, X_j) = \tau_{ij}(\varphi(s), \varphi \circ \sigma)$ and

$\rightarrow$ **Type encoders**, $\tau_{ij} : \mathcal{X} \times \boldsymbol{\mathcal{X}}^{\mathcal{S}} \to \mathcal{T}$, that identify the relation type between any two variables X_i, X_j from the context $s \in \mathcal{S}$ and the causal relations $\varphi \circ \sigma$.

Type encoders can be freely chosen by the user.

$\rightarrow$ A **Meta-Causal Frame** is a combination of a mediation process with a particular set of type encoders.

Willig, M., Tobiasch, T., Busch, F.P., Seng, J., Dhami, D.S. and Kersting, K., Systems with Switching Causal Relations: A Meta-Causal Perspective. In The Thirteenth International Conference on Learning Representations.

Meta-Causal Models

Given a Meta-Causal Frame,
Meta-Causal Models (MCM) model the change in MCS:

$$\delta : \mathcal{T}^{N \times N} \times \mathcal{S} \to P(\mathcal{T}^{N \times N})$$

MCM are (probabilistic) state machines

$\to$ 'Given the current MCS and the system state, what is next MCS?'

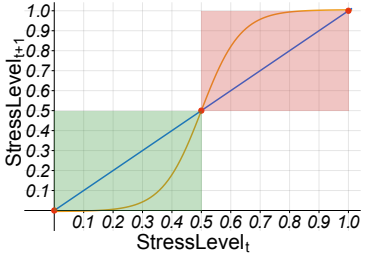
MCM model the qualitative change in cause-effect relations.
$\to$ MCM allow to reason about the emergence and vanishing of causal edges.

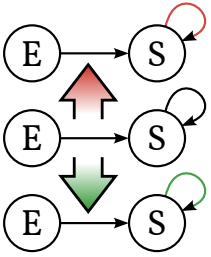
Dynamics Systems - Self-Reinforcing Stress

'Is the person stressing themselves at the moment?'

→ Types can change without the structural equations changing.

→ The state of the system matters!



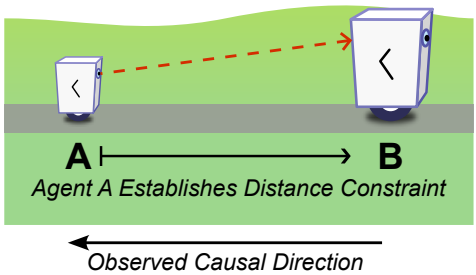


Influence of S onto itself identifies suppressing or self-reinforcing dynamics.

→ $(T_s = \text{supressing}) \Leftarrow (\ddot{f}_s < 0)$.

$$\tau\left(\begin{bmatrix} e_t \\ s_t \end{bmatrix}\right) := \begin{bmatrix} 0 & 0 \\ 1 & a \end{bmatrix} \text{ with } a := \text{sign}(\ddot{f}_s) = \text{sign}(t_s - 0.5) \in \{-1, 0, 1\}$$

Meta-Causal State Inference

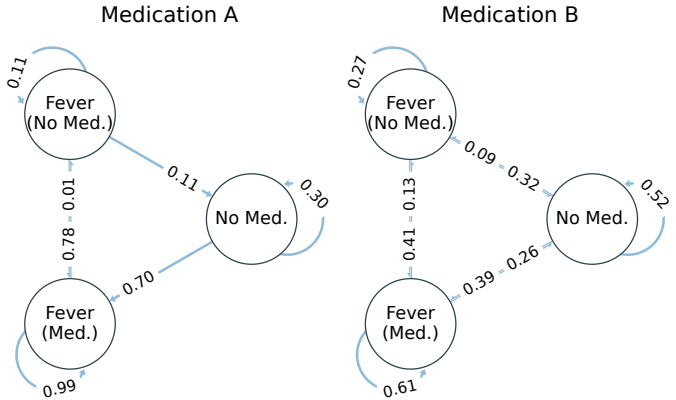


Some meta-causal states can be inferred from observations:

$$(t_{B \to A} = \text{“A chasing”}) \Leftrightarrow (\dot{A}_{pos} \cdot (B_{pos} - A_{pos}) > 0)$$

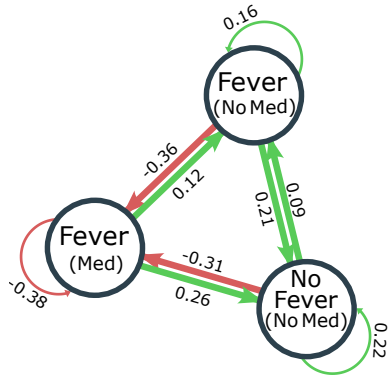
Medicating Flu: MCA

Record MCM Transition Statistics



Meta-Causal ATE

$\text{sMCATE}(P_A, P_B) := P_B - P_A$

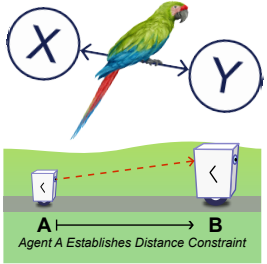


From Parroting to Understanding: A Meta-Causal Path

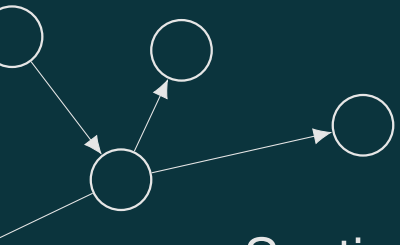
Reflection & Adaptation: Genuine understanding isn't just about knowing that 'A causes B', but understanding the conditions under which that relationship holds, and to adapt when it changes.

Meta-Causal Models allow to explicitly model and reason about *how* and *why* causal relationships change.

Future AI Systems should not just produce due to their intrinsic weights, but deliberately think about the underlying mechanisms at play.

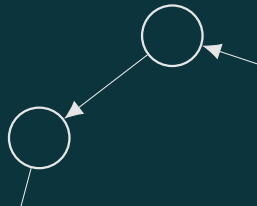


Meta-Causality may be the dividing line between systems that merely describe the world from those that truly understand it.



Section 15

Practice Exam



Practice Exam

$\rightarrow$ *published solution.*

128 concepts to know - I

Probability & Bayesian Networks
Joint/Conditional/Marginal Distribution
Chain Rule of Probability
Bayes' Rule
Marginal/Conditional Independence
Conditional Prob. Tables
Bayesian Networks
Directed Acyclic Graph (DAG)
(Local) Markov Assumption
I-Map/D-Map/P-Map
d-separation
Chain/Forks/Colliders
Markov Condition
Faithfulness
Markov Equivalence Class (MEC)

Fundamentals & Causal Models
Reichenbach's Common Cause Principle
Spurious Correlation
Confounding
Correlation vs. Causation
Conditioning vs. Intervening
Structural Causal Model (SCM)
Endogenous/Exogenous Variables
Structural Equations
Truncated Factorization
do-operator
Independent Mechanisms Principle

128 concepts to know - II

Interventions & do-Calculus
Causal Query
Average Treatment Effect (ATE)
Identifiability
Back-door Criterion
Back-door Adjustment
Front-door Criterion
Front-door Adjustment
do-calculus
- Rule 1 (Insertion/Deletion of Observations)
- Rule 2 (Action/Observation Exchange)
- Rule 3 (Insertion/deletion of actions)

Soundness of do-calculus
Completeness of do-calculus

Counterfactuals
Pearl Causal Hierarchy (PCH)
Counterfactual Probability
Counterfactual Inference
1. Abduction
2. Action
3. Prediction

Causal World
Twin-Networks
Collapse of Causal Rungs

128 concepts to know - III

Causal Discovery & Evaluation
Task of Causal Discovery
Causal Sufficiency
Peter-Clark Algorithm (PC)
Skeleton
V-Structures
Unshielded Triplets
Meek Rules (towards Acyclicity)
Greedy Equivalence Search (GES)
Bayesian Information Criterion
Completed Partially Directed Acyclic Graph (CPDAG)
Precision/Recall/F1-Score (on Graphs)
Structural Hamming Distance (SHD)
Structural Intervention Distance (SID)
Bounded Effects
Partial Compliance
Matrix Powers and 'DAGness'

Causal Abstractions
Levels of Granularity
Marginalization (Abstraction)
Grouping (Abstraction)
$(\tau - \omega)$-Abstractions
Commuting Diagram (Abstraction)
Approximate Abstractions
Types of Abstractions
Cluster DAGs (C-DAGs)
Consolidation

128 concepts to know - IV

Neuro-Causal Models & CRL
Causal Normalizing Flows (CNFs)
CausalVAE
Interventions in CNFs/CausalVAEs
Interv. Sum-Product Network (iSPN)
Neural Causal Models (NCM)
Identifiability in NCM
Task of Causal Representation Learning
Hierarchy of Causal Tasks
Content-Style Separation
Latent Causal Process
CRL under Intervention
CRL under Sufficient Variation

Causality & LLM
Qualitative Equiv. of Causal Information
Post-Hoc Fallacy
Pairwise Causal Graph Prediction
Breadth-First Search (BFS) Discovery
Order and Ancestral Constraints
Causal Parrots

Fairness & Bias
Berkson's Paradox (Collider Bias)
Simpson Paradox
Construct/Observed/Decision Spaces
Protected Attributes
Demographic Parity

128 concepts to know - V

Equality of Opportunity
Equalized Odds
Counterfactual Fairness
CF Fairness for ML Training
'Fairness through Unawareness'

Actual Causality
Type Causality
Token Causality
Signature, Causal World & Entailment
But-For Test
Modified HP Definition (mHP)
- AC1 (Relevant Worlds)
- modified AC2
- AC3 (Minimality)

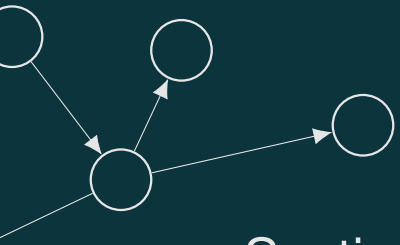
Root Causes in Standard SCM

Causal Perspectives
Shapley Values
SHapley Additive exPlanations (SHAP)
Causal Shapley Values
Potential Outcomes (PO) Framework
Fundamental Problem of Causal Inference
PO Assumptions:
- SUTVA (Stable Unit Treatment Value Assumption)
- Ignorability (Unconfoundedness)
- Positivity (Overlap)

128 concepts to know - VI

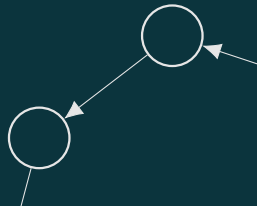
Regression Discontinuity Design (RDD)
Granger Causality
Cyclic SCM
Hierarchy of Structural Causal Models

Meta-Causality
Contextual Independence
Meta-Causal Models (MCMs)
Meta-Causal Variables (MCVs)
Meta-Causal States (MCS)
Meta-Causal Predictability



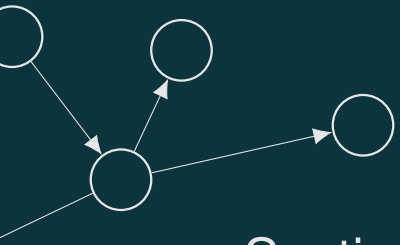
Section 16

Q&A



Q&A

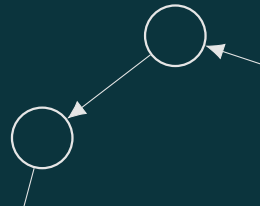
$\rightarrow$ *Your questions here.*



Section

17

Final Exam



Formalities

Exam date: 18. March 12:30-13:30 (60min.)
Rooms: *To be determined...*

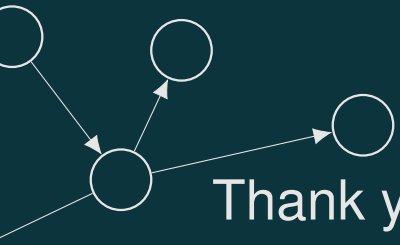
The exam will be held in English.

Allowed materials:

- Single, one-sided handwritten A4 sheet.
- Standard scientific (non-graphic, non-programmable, non-CAS) calculator.
- German-English dictionary.

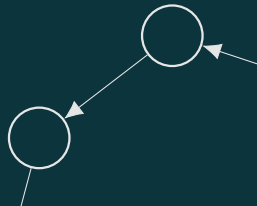
"To Build Truly Intelligent Machines,
Teach Them Cause and Effect"

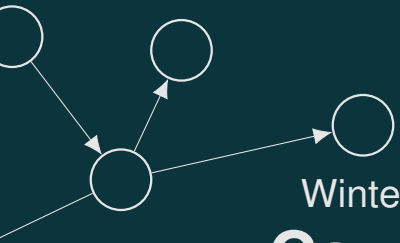
- Judea Pearl, Dana Mackenzie *"The Book of Why"* (2018)



Thank you for a great semester!

Wishing you a successful
completion of the module.





TECHNISCHE UNIVERSITÄT DARMSTADT

AIML Lab

Winter Semester 2025/26 Lecture

Causality for AI & ML

Feel free to reach out:

Moritz Willig https://moritz-willig.de/

Computer Science Department
Technical University of Darmstadt
moritz.willig@cs.tu-darmstadt.de