Causality for Machine Learning I

Matej Zečević

TECHNISCHE UNIVERSITÄT DARMSTADT

7th Int'l Summer School on Data Science (SSDS 2022),
Day 4, 13th October 2022

Overview of Today's Programme:

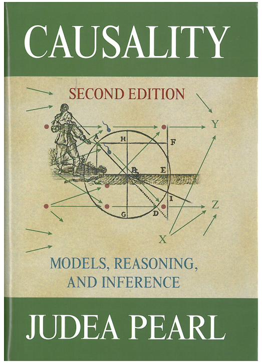
- **Now**: *Causality for Machine Learning 1*, where we explore the basic concepts from causality that are being used in current ML research
  9 am – 10:30 am (including Q&A)

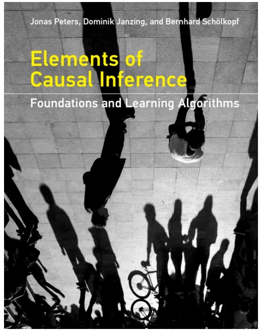
- Following: *Causality for Machine Learning 2*, where we discuss some selected publications and on-going work on the interface of causality and ML
  10:45 am – 12:15 am (including Q&A)

- In the afternoon: *Hands-on session: Causality for Machine Learning,* where you get to experience some code for using causal ML first hand
  1:30 pm – 3 pm

Pointers to Causal Inference References

- Judea Pearl, “**Causality**”, Cambridge University Press, 2009.
- Peters et al., “**Elements of Causal Inference**”, MIT Press, 2017.
- Elias Bareinboim Lecture “**Causal Data Science**”, 2019.
  https://www.youtube.com/watch?v=dUsokjG4DHc
- Brady Neal’s Free Online Course ”**Introduction to Causal Inference**”, 2020.
  https://www.bradyneal.com/causal-inference-course
- Jonas Peters Lecture Series “**Causality**”, 2017.
  https://www.youtube.com/watch?v=zvrcyqcN9Wo





TECHNISCHE
UNIVERSITÄT
DARMSTADT

This Lecture is mostly based on:

Book by Pearl 2009, “**Causality**”

Book by Peters et al. 2017, “**Elements of Causal Inference**”

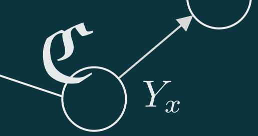
Works from the **Causal AI Lab @ Columbia**

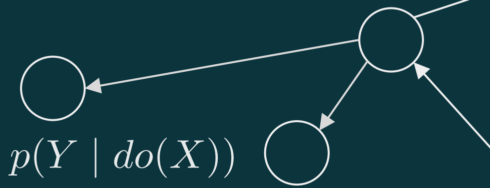
Notation that we will use

X, Y, Z	random variable (RV)	$X \perp Y \mid Z$	conditional independence
x	value of RV	$\mathcal{M}$	structural causal model
$\mathbf{X} = (X_1, \ldots, X_d)$	RV vector length d	$\mathcal{G}$	graph
$X^1, \ldots, X^n \overset{iid}{\sim} P_X$	i.i.d. sample of size n	$\mathbf{PA}_X^{\mathcal{G}}$	parents of RV X in graph G
P	probability measure	$do(X = 3)$	do operator
P_X	prob. distribution of X	$P_{Y \mid X=x}$	conditional distribution
p	density		
$\mathbb{E}[X]$	expectation		
$\mathrm{cov}[X]$	covariance		
$\mathrm{var}[X]$	variance		

TECHNISCHE UNIVERSITÄT DARMSTADT

At least we will try to be consistent, do not worry the community itself has not settled on the best notation yet.

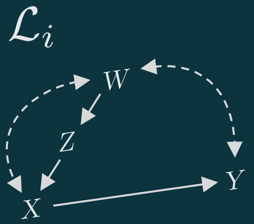




I | Why?

short for:

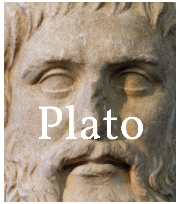
Why does Machine Learning need Causality?



TECHNISCHE UNIVERSITÄT DARMSTADT

What is Causality?

We might want to start here first..

Probably, 

was the first to state the principle of causality:

“Everything that becomes or changes must do so owing to some cause; for nothing can come to be without a cause.”- *Timaeus* 28a

A Counterfactual Theory

The theory of Judea Pearl

Two fundamental laws of causal inference (in words):

1. How **counterfactuals** and probabilities of counterfactuals are deduced from a given SCM

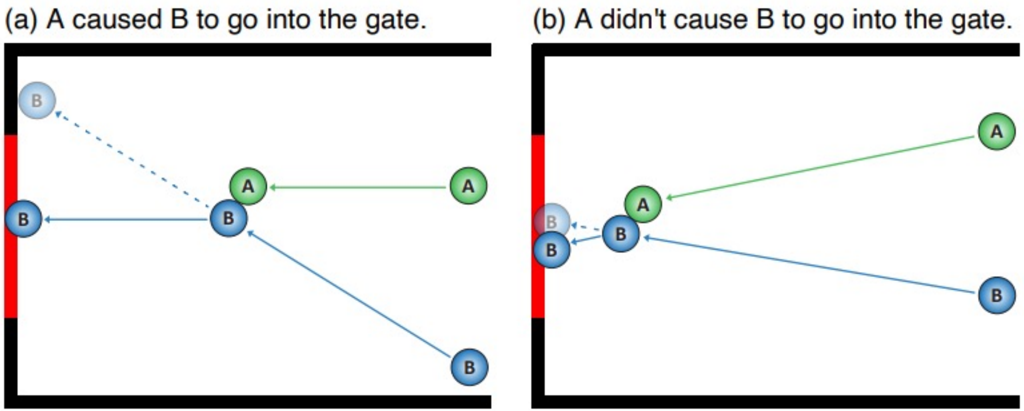
2. How features of the observed data are shaped by the graphical **structure** of a SCM

Pearlian Causality

A success story

The formalization with most success in AI/ML so far.

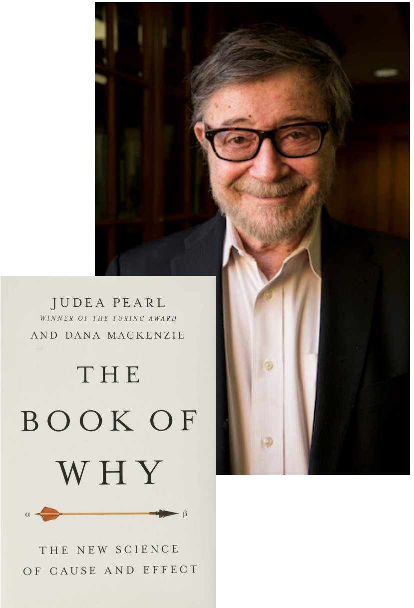
Works in Cognitive Science also in support of the key ideas in the formalism i.e., **humans reason counterfactually**.



Gerstenberg. What would have happened?[...] PTRBAE 2022.

Judea Pearl’s opinion

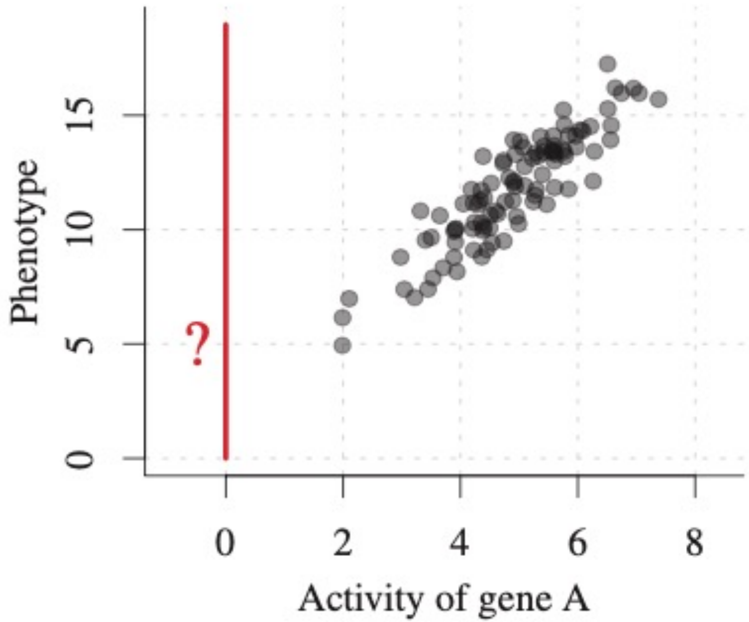
Pioneer of Causality for AI, Turing awardee

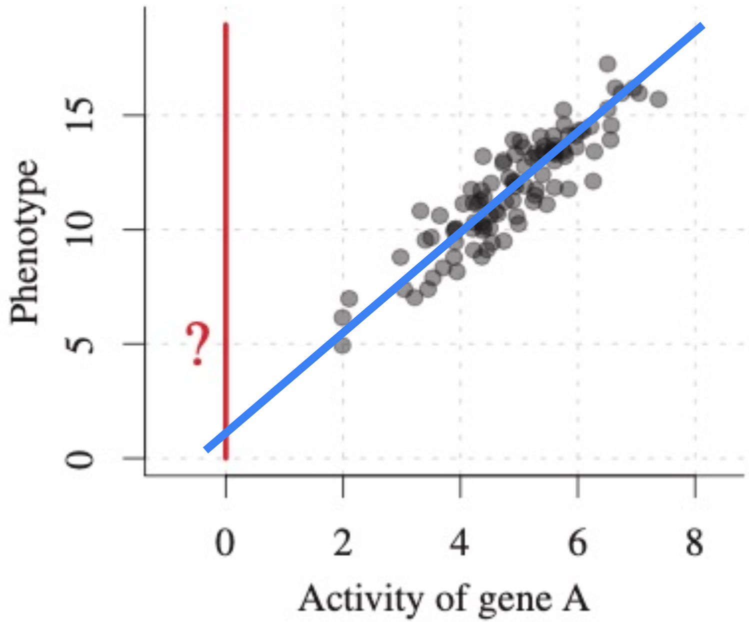


“To Build Truly Intelligent Machines,
Teach Them Cause and Effect”

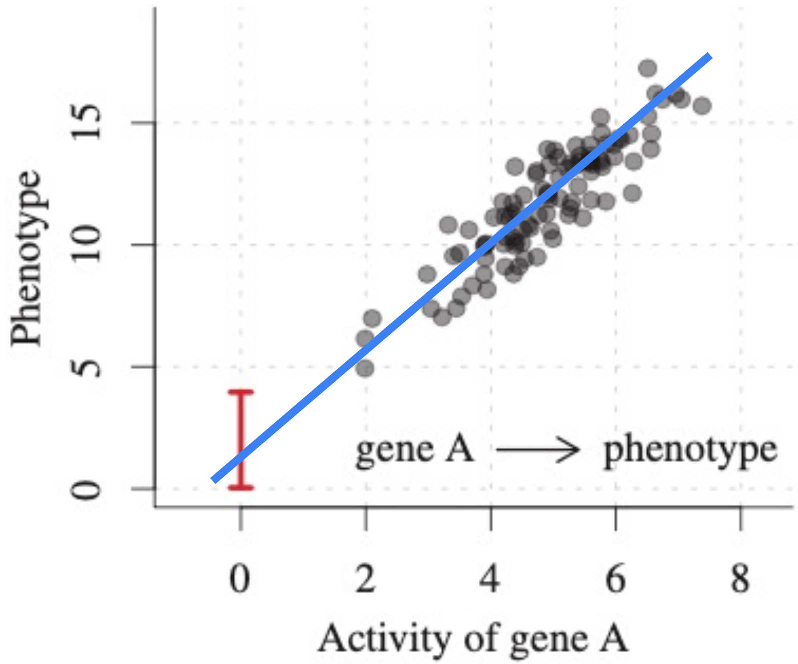
“All the impressive achievements of deep learning amount to just curve fitting”

Judea Pearl in “The Book of Why”
and in an interview with quantamagazine in 2018

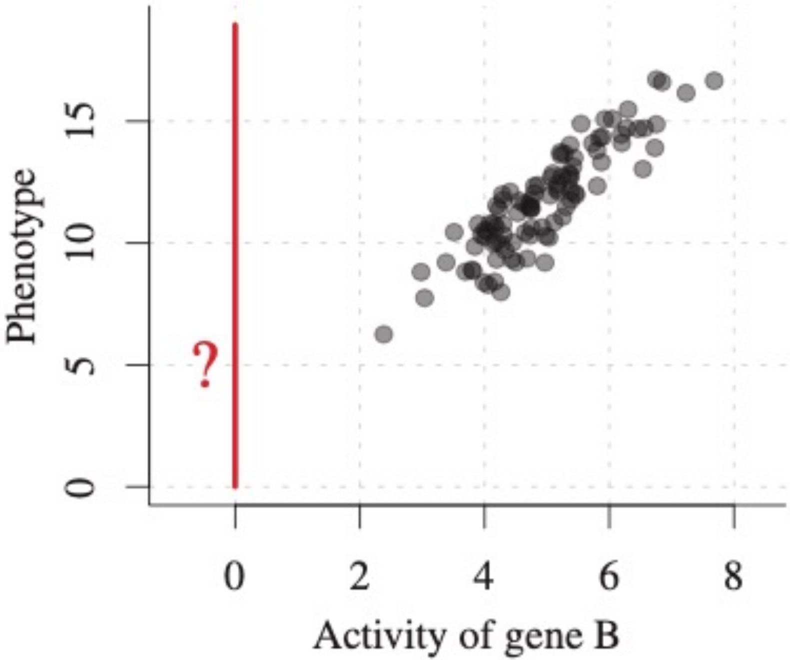

Phenotype
0
5
10
15
?
0
2
4
6
8
Activity of gene A



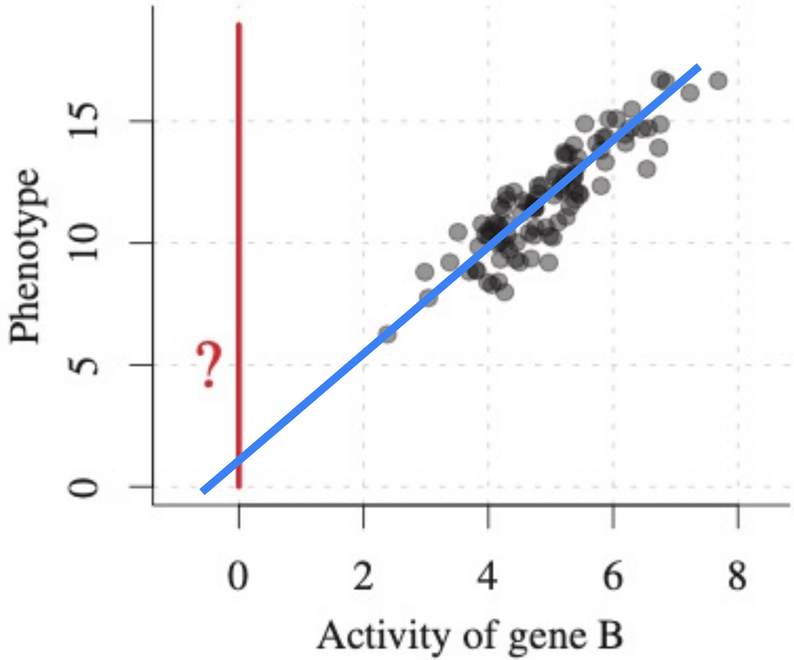
TECHNISCHE UNIVERSITÄT DARMSTADT



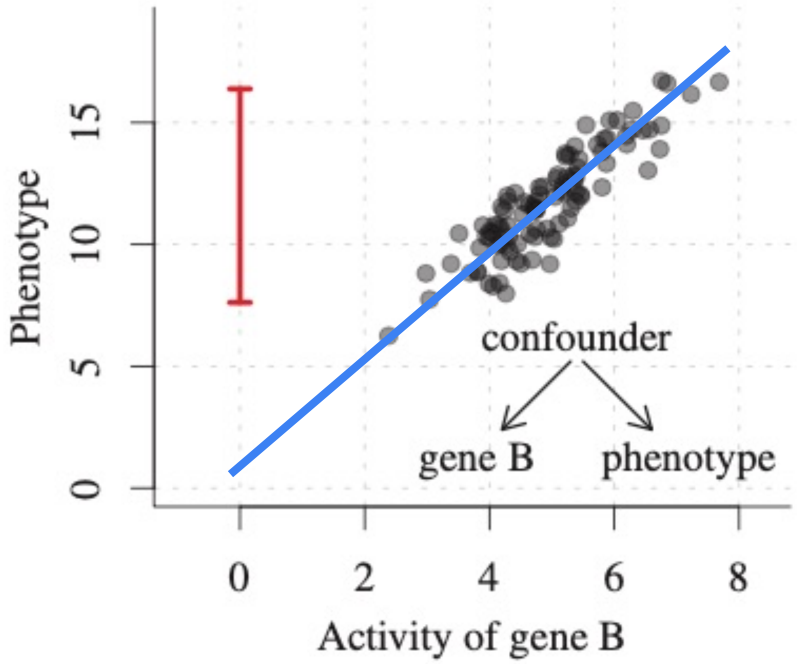
TECHNISCHE UNIVERSITÄT DARMSTADT



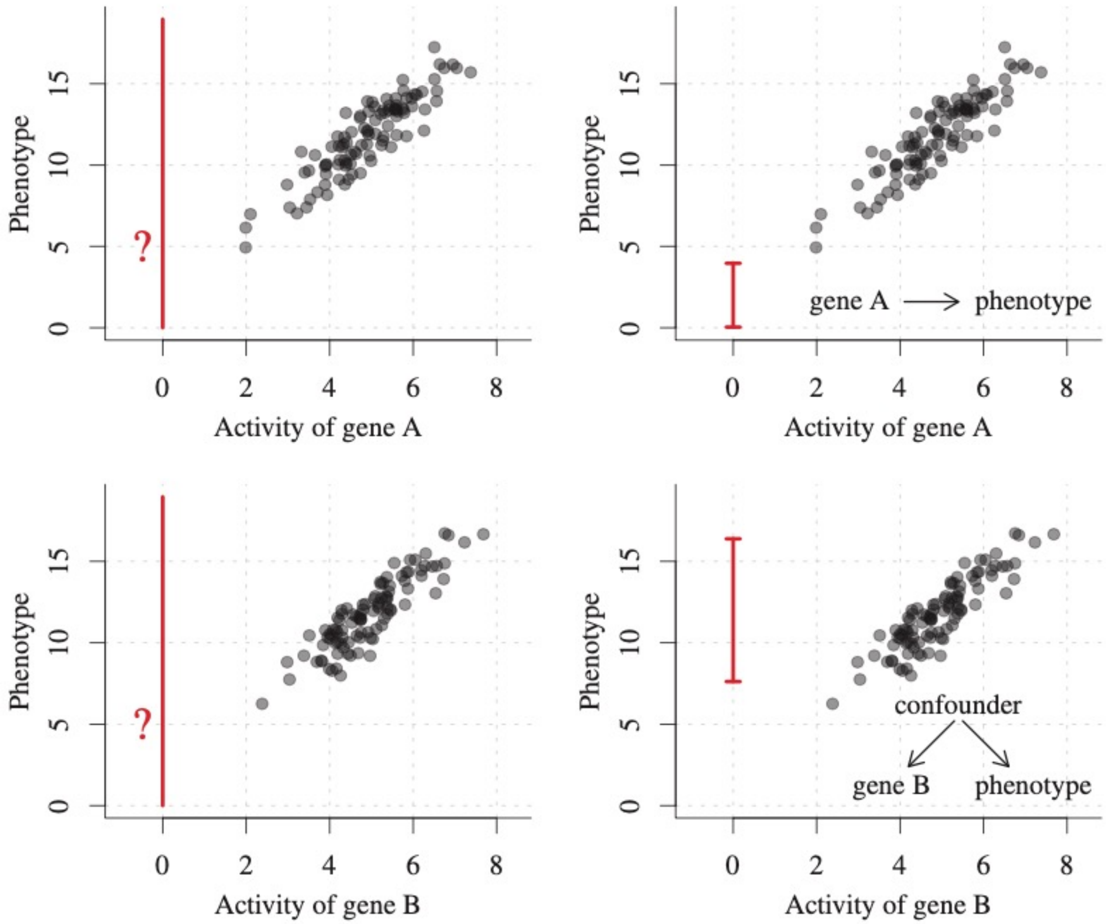
Phenotype
0
5
10
15
?
0
2
4
6
8
Activity of gene B



TECHNISCHE UNIVERSITÄT DARMSTADT


Phenotype
0
5
10
15
confounder
gene B
phenotype
0
2
4
6
8
Activity of gene B

The Full Picture



Peters et al. 2017, Fig.1.4

Reichenbach's Common Cause Principle

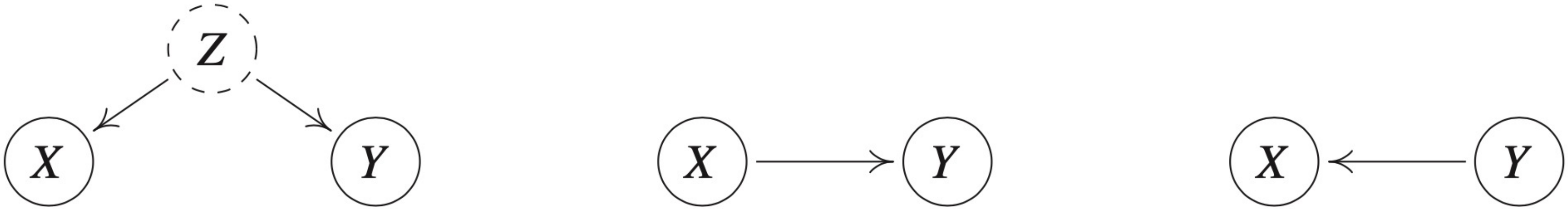
Formalism

Principle 1. *If two random variables X and Y are statistically dependent ($X \not\perp Y$), then there exists a third variable Z that causally influences both. (As a special case, Z may coincide with either X or Y.) Furthermore, this variable Z screens X and Y from each other in the sense that given Z, they become independent, $X \perp Y \mid Z$.*

TECHNISCHE UNIVERSITÄT DARMSTADT

Reichenbach's Common Cause Principle

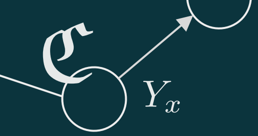
Three cases

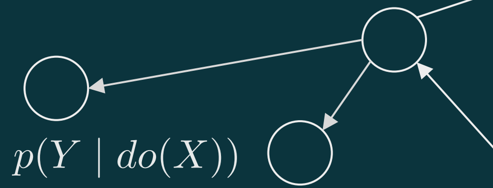


Peters et al. 2017, Fig.1.2

Causality allows us to talk about **modelling assumptions**

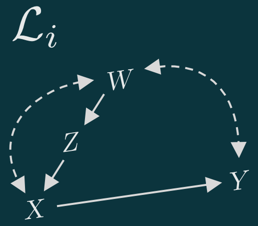
Causality allows us to consider not just the joint distribution but the **data generating process** which induces said distribution





2 | What?

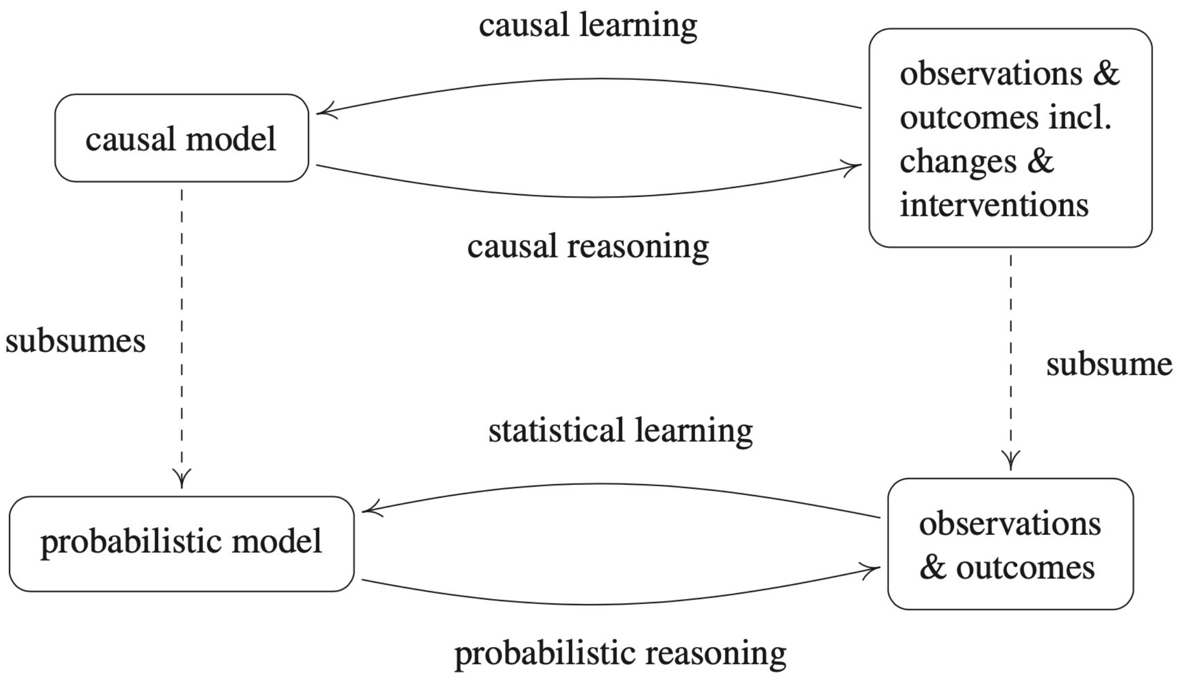
short for:
What does Pearlian Causality look like?



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Causal versus Probabilistic Inference

Another consequence from the previous example



Peters et al. 2017, Fig.1.1

Pearl Causal Hierarchy

Level I

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?

Pearl Causal Hierarchy

Level II

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?

Pearl Causal Hierarchy

Level III

3. COUNTERFACTUALS

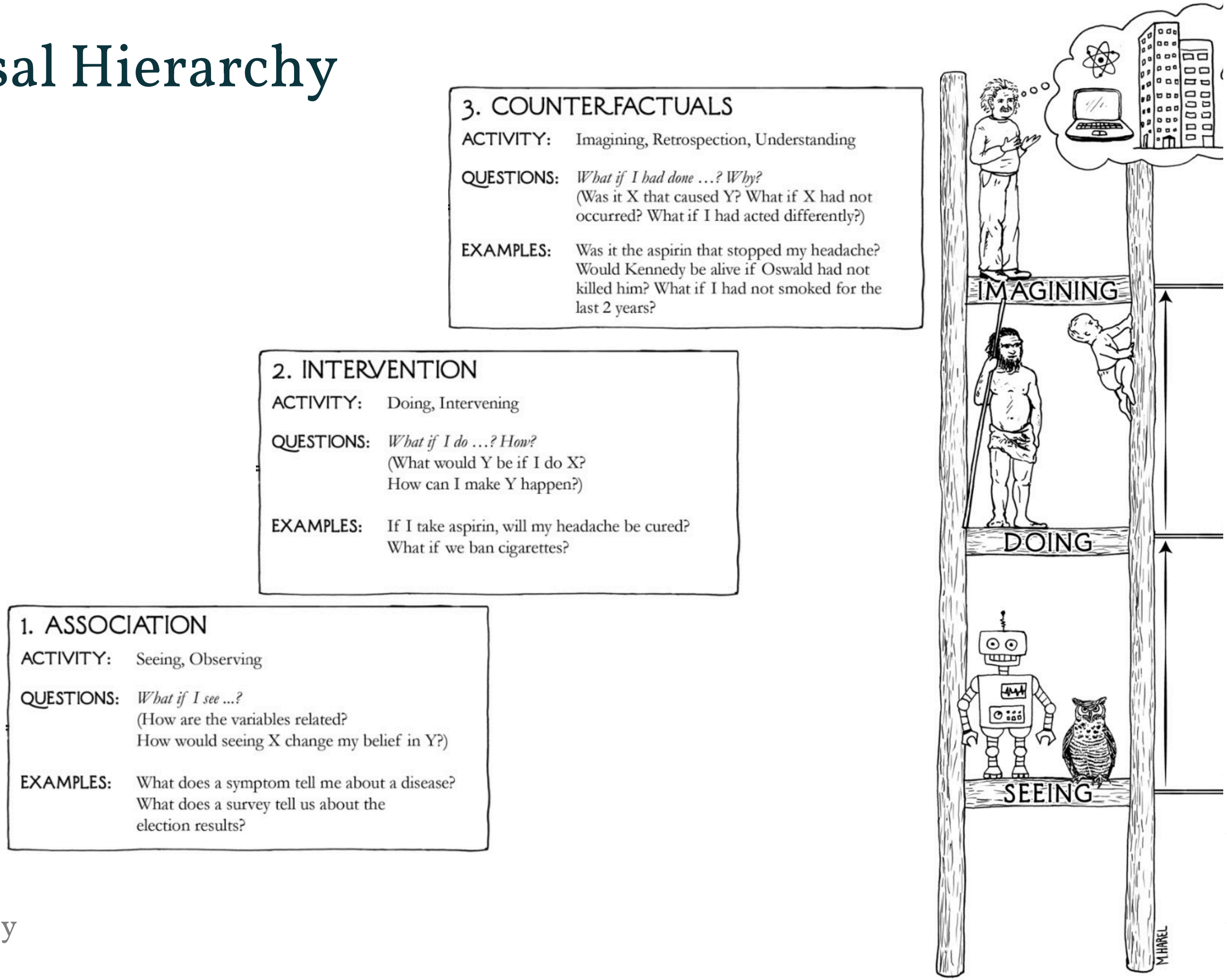
ACTIVITY: Imagining, Retrospection, Understanding

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?

Pearl Causal Hierarchy

Level 1



Pearl, The Book of Why

TECHNISCHE UNIVERSITÄT DARMSTADT

Structural Causal Model

Definition

A structural causal model $\mathcal{M}$ (or data generating model) is a tuple $\langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P_{\mathbf{U}} \rangle$, where

$\mathbf{V}$ are endogenous variables

$\mathbf{U}$ are exogenous variables

$\mathcal{F}$ are functions determining V i.e., $v_i = f_i(\mathbf{pa}_i, \mathbf{u}_i)$

$P_{\mathbf{U}}$ is the probability distribution over U.

Assumption: $\mathcal{M}$ is recursive i.e., there are no feedback (cyclic) mechanisms

TECHNISCHE UNIVERSITÄT DARMSTADT

Interventions

Changing the SCM's structural equations

$$\mathcal{M} \begin{cases} \mathbf{V} = \{X, Y\} \quad \mathbf{U} = \{U_{XY}, U_X, U_Y\} \quad P_{\mathbf{U}} \\ \mathcal{F} = \begin{cases} X = f_X(U_X, U_{XY}) \\ Y = f_Y(X, U_Y, U_{XY}) \end{cases} \end{cases}$$

$$P(Y = y \mid X = x) \neq P(Y = y)$$

$$\Big\downarrow \; do(X = x)$$

$$\mathcal{M}_x \quad \mathcal{F} = \begin{cases} X = x \\ Y = f_Y(x, U_Y, U_{XY}) \end{cases}$$

$$P(Y = y \mid do(X = x)) \neq P(Y = y)$$

Counterfactuals

A 3-step procedure

$$\begin{aligned}
&\text{Abduction}: &&\text{Update belief in } P_{\mathbf{U}} \text{ given evidence } \mathbf{E}\\
&\text{Action}: &&\text{Change equations accordingly, } do(X = x)\\
&\text{Prediction}: &&\text{Look at variable of interest } P(Y = y)
\end{aligned}$$

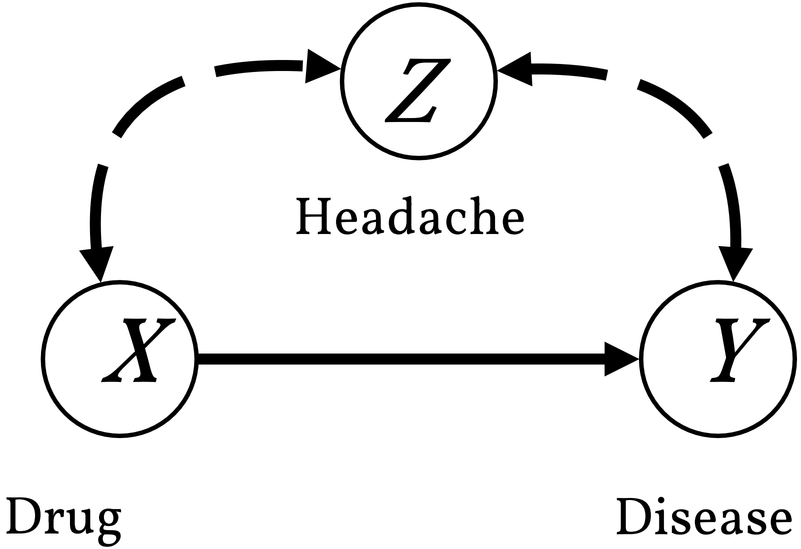
The Causal Graph

An induced property of the SCM

latent

$$\mathcal{F} = \begin{cases} X = f_X(U_X, U_{XZ}) \\ Y = f_Y(X, U_Y, U_{YZ}) \\ Z = f_Z(U_Z) \end{cases}$$

$\longrightarrow$



Causal Hierarchy Theorem

Impossibility Results

cross-layer inference

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?

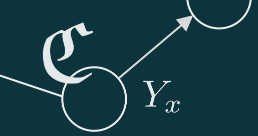
2. INTERVENTION

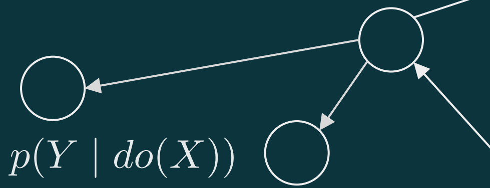
ACTIVITY: Doing, Intervening

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

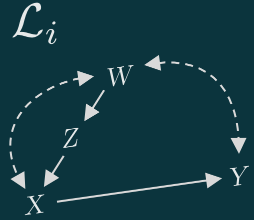
EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

Bareinboim et al., On Pearl's Causal Hierarchy





3 | Machine Learning for Causal Effects



TECHNISCHE
UNIVERSITÄT
DARMSTADT

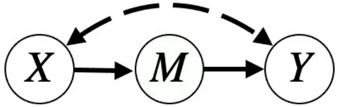
Feasible Cross-layer Inference

Via Constraints on Causal Graph

Using Observational Data From One Population/Domain

1. **Query**
   $P(y \mid do(x))$

2. **Causal Contraints**



3. **Data**
   $P(x, m, y)$

Inference Engine

Solution

yes / no

$$P(y \mid do(x)) = \sum_{m} P(m \mid x) \sum_{x'} P(y \mid m, x') P(x')$$

Interventional Distribution ← Available Distributions

Pearl's do-calculus

"The Inference Engine"

Rule 1 (Insertion/deletion of observations):

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

Rule 2 (Action/observation exchange):

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

Rule 3 (Insertion/deletion of actions):

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

✅

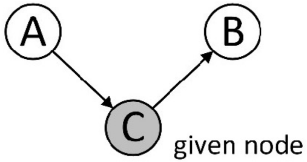
complete set of rules

Pearl 2009, Causality

d-Separation & Conditional Independence

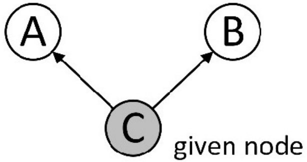
Graphical Tools

G



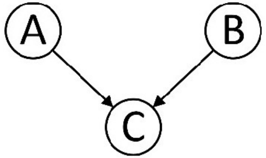
$(A \perp B \mid C)_G$

G



$(A \perp B \mid C)_G$

G



$(A \perp B \mid \emptyset)_G$

global Markov property

$$(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})_G \implies (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})_P$$

faithfulness

$$(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})_P \implies (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})_G$$

local factorization

$$P(\mathbf{v}) = \sum_{\mathbf{u}} P(\mathbf{u}) \prod_i P(v_i \mid \mathbf{pa}_i, u_i)$$

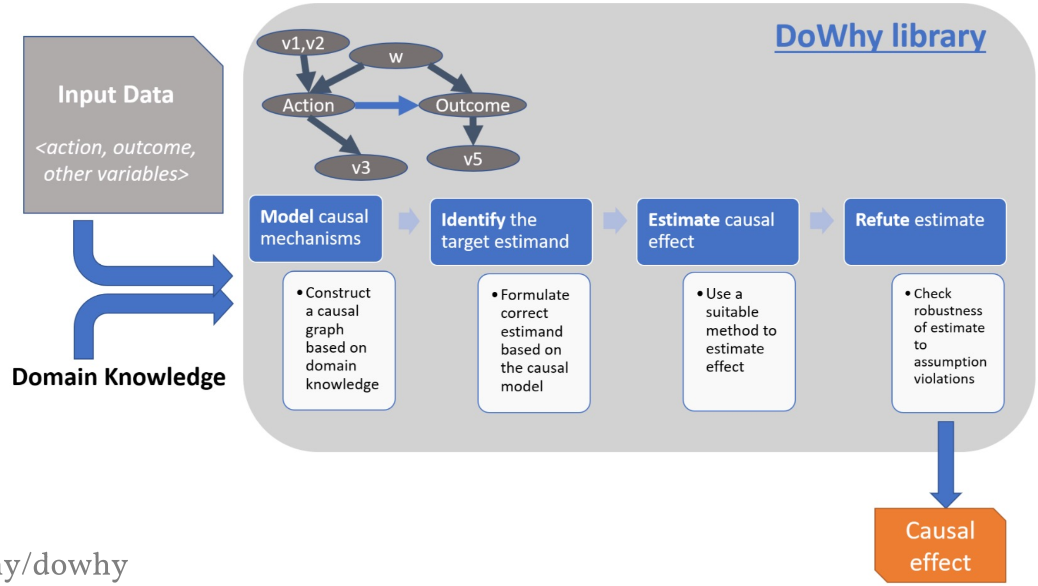
Free Code Libraries

Do it for you

DoWhy | An end-to-end library for causal inference

> Introducing DoWhy and the 4 steps of causal inference | Microsoft Research Blog | Video Tutorial | Arxiv Paper | Arxiv Paper (GCM-extension) | Slides
>
> Read the docs | Try it online! launch binder

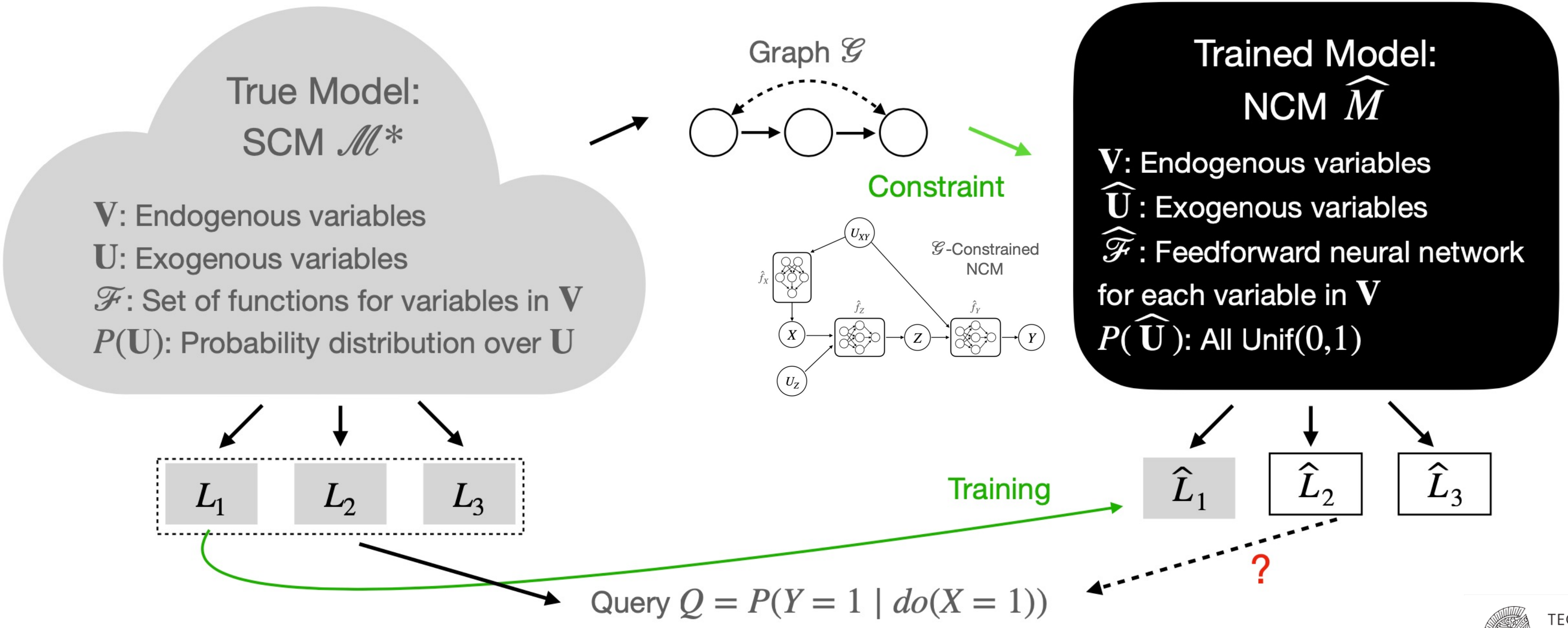
Case Studies using DoWhy: Hotel booking cancellations | Effect of customer loyalty programs | Optimizing article headlines | Effect of home visits on infant health (IHDP) | Causes of customer churn/attrition



DoWhy, https://github.com/py-why/dowhy

Neural Networks for SCM

Training an SCM



Xia et al. 2021, The Causal Neural Connection

4 Machine Learning for Causal Structures

TECHNISCHE UNIVERSITÄT DARMSTADT

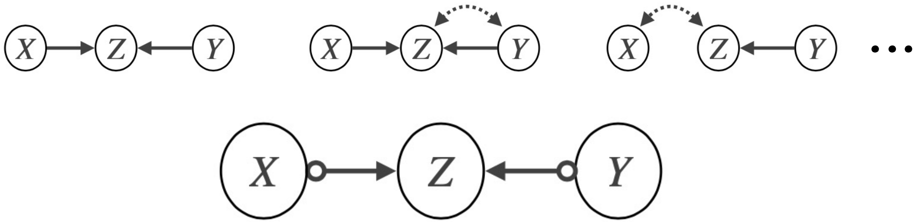
Causal Structure Learning using FCI

Old school conditional independencies

$$P(x,y,z) = P(z\,|\,x,y)\color{red}{P(x\,|\,y)}\color{black}{P(y)}$$
$$= P(z\,|\,x,y)\color{red}{P(x)}\color{black}{P(y)}$$

latent

$$\mathcal{F} = \begin{cases} X = f_X(U_X) \\ Y = f_Y(U_Y) \\ Z = f_Z(X, Y, U_Z) \end{cases}$$



$\color{red}{X \perp Y}$

$X \not\perp Z$
$Y \not\perp Z$
$X \not\perp Y \mid Z$

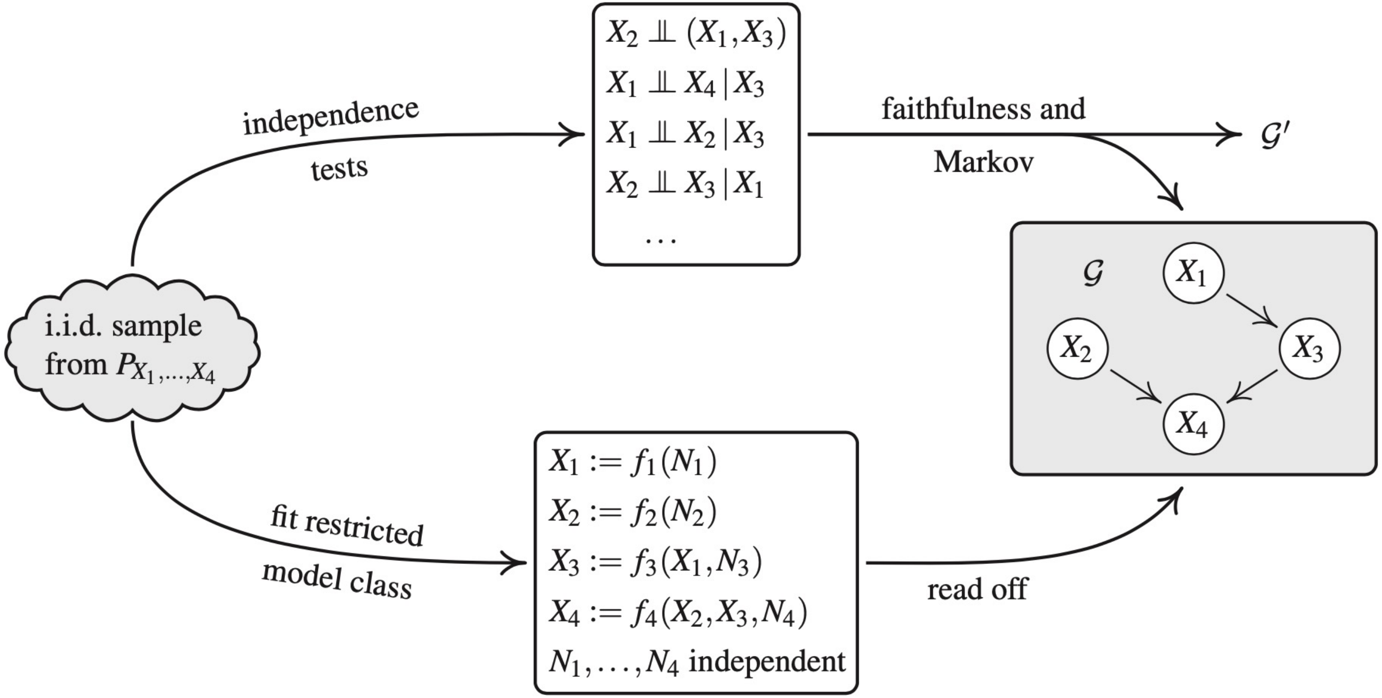
Fast Causal Inference

TECHNISCHE UNIVERSITÄT DARMSTADT

Zhang 2008, On the completeness of orientation rules for causal discovery (...)

The Alternative to CSL

Restricting the function class



Peters et al. 2017, Fig. 7.1

Identifiability Results

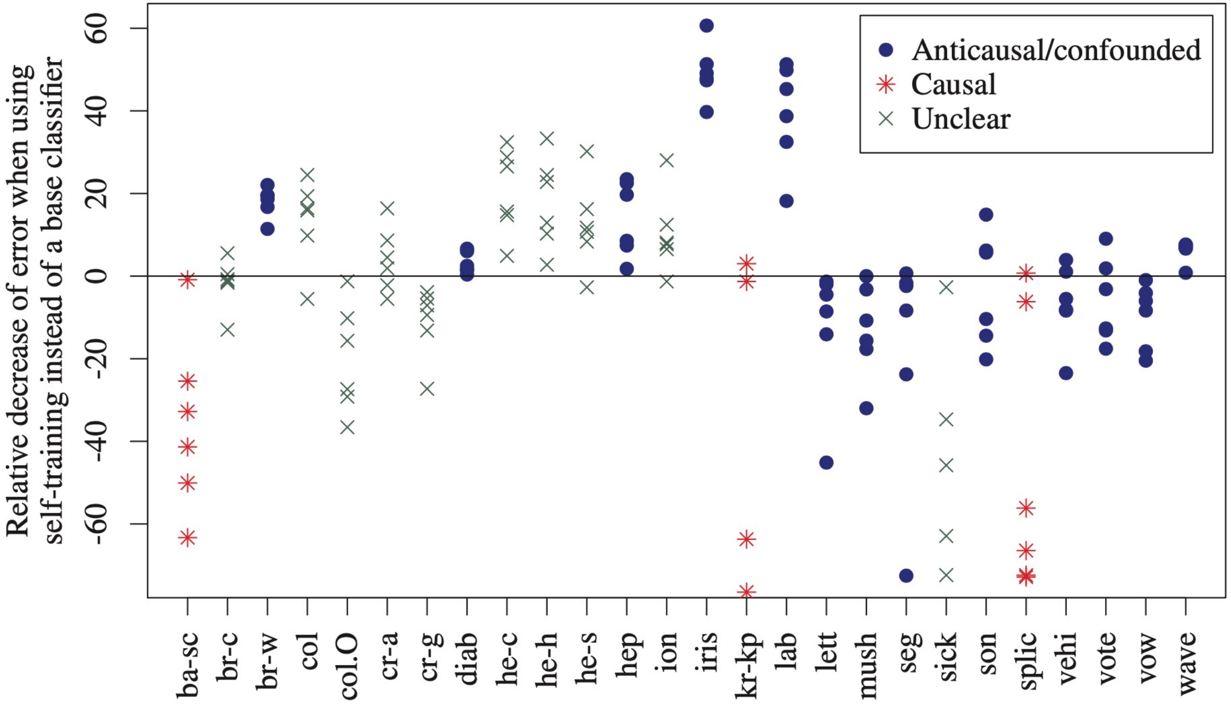
Taxonomy of Results

Type of structural assignment		Condition on funct.	DAG identif.
(General) SCM:	$X_j := f_j(X_{\mathbf{PA}_j}, N_j)$	—	✗
ANM:	$X_j := f_j(X_{\mathbf{PA}_j}) + N_j$	nonlinear	✓
CAM:	$X_j := \sum_{k \in \mathbf{PA}_j} f_{jk}(X_k) + N_j$	nonlinear	✓
Linear Gaussian:	$X_j := \sum_{k \in \mathbf{PA}_j} \beta_{jk} X_k + N_j$	linear	✗
Lin. G., eq. error var.:	$X_j := \sum_{k \in \mathbf{PA}_j} \beta_{jk} X_k + N_j$	linear	✓

Peters et al. 2017, Tab. 7.1

Anti-Causal Learning

Semi-supervised Learning depends on Causal Structure



Schölkopf et al. 2012, On causal and anticausal learning.

After having seen all this, we realize..

“As X-rays are to the surgeon, <u>graphs</u> are for causation.”

-Judea Pearl in Causality (2009)

"As graphs are to the causality, causal nets are for AI."

"As X-rays are to the surgeon, graphs are for causation."

-Judea Pearl in Causality (2009)

A | Announcements

Opportunities following SSDS 2022

TECHNISCHE UNIVERSITÄT DARMSTADT

Join the weekly discussion

(or simply listen..)

Causality Discussion Group

Join the community via discuss.causality.link

Join the virtual event
ncsi.cause-lab.net

neuro (Causal ∧ Symbolic) AI

Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS)

December 2022

Visit **ncsi.cause-lab.net**

TECHNISCHE UNIVERSITÄT DARMSTADT

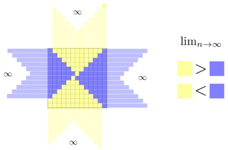
Stay in Touch via Social Media

Twitter
@matej_zecevic

Open-access articles on my Blog
(pretty much on anything but mostly on nerdy topics)
matej-zecevic.de

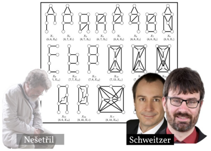
The Infinite is Useful

Talking about the foundation of mathematics usually involves at some initial point the discussion on set theory, which one arguably considers to be a "theory...



Proving Nešetřil's Conjecture: Minimal Asymmetric Graphs

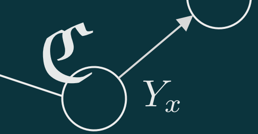
In 1988 at a seminar in Oberwolfach (located in Germany and by many mathematicians considered as a sort of "mecca for mathematicians" [1]) the czech...

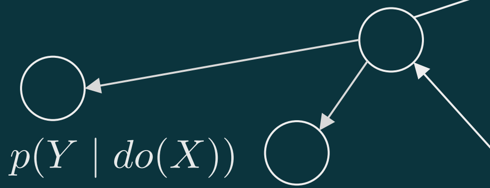


Sports Excellence: The Greatest Shohei Ono

Judo (柔道, japanese for "gentle way") is a martial art centered around throwing techniques for close-quarters combat (opposed to for instance Karate, another japanese martial...

TECHNISCHE
UNIVERSITÄT
DARMSTADT





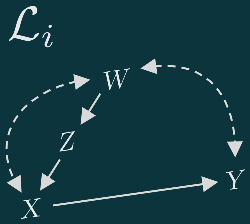
That's a wrap for CML I!

Questions?

See you at CML II later!

matej.zecevic@tu-darmstadt.de | https://www.matej-zecevic.de

Further, my gratitude and thanks go out to Kristian Kersting, Devendra Dhami, all our collaborators, the AIML lab @ TU Darmstadt and finally the organizers of SSDS 2022



TECHNISCHE UNIVERSITÄT DARMSTADT