“Eating pizza increases your IQ!”

Full Orbit pizza session

Philip Boeken

p.a.boeken@uva.nl

[1] University of Amsterdam
The Netherlands

[2] Booking.com
The Netherlands

September 15, 2023

About me

’14 - ’17 BSc. Business Analytics (VU)

’18 - ’20 MSc. Mathematics (UvA)

’21 - ’... PhD Causality and Mathematical Statistics/ML/AI/...

- Supervised by Prof. Dr. Joris Mooij (UvA)
- Co-supervised by Dr. Onno Zoeter
  (Mercury Machine Learning Lab, Booking.com)

Credits

This presentation is heavily inspired by:

- Joris' inaugural lecture [Mooij, 2023];
- the MasterMath Causality course;
- Judea Pearl and Dana Mackenzie's *The Book of Why* [Pearl and Mackenzie, 2018].

Causality in the media

Business insider:

HOME > SCIENCE

Study Links A Country's Chocolate Intake To How Many Nobel Prize Winners It Spawns

Jennifer Welsh Oct 11, 2012, 12:09 AM CEST

The best "brain food" might be chocolate, a new study out in the New England Journal of Medicine suggests. The study links a country's chocolate consumption and the number of Nobel Prize winners that country has created.

Business Insider

Causality in the media

The Guardian:

Diet of fish 'can prevent' teen violence

New study reveals that the root cause of crime may be biological, not social

Gaby Hinsliff, *chief political correspondent*

Sun 14 Sep 2003 09.22 BST

Feeding children a diet rich in fish could prevent violent and anti-social behaviour in their teens, according to research to be announced this week which suggests the root causes of crime may be biological rather than social.

Causality: early history

David Hume (1740):

> *Thus we remember to have seen that species of object we call flame, and to have felt that species of sensation we call heat. We likewise call to mind their constant conjunction in all past instances. Without any farther ceremony, we call the one cause and the other effect, and infer the existence of the one from that of the other.*

Karl Pearson (1892):

> *Beyond such discarded fundamentals as 'matter' and 'force' lies still another fetish amidst the inscrutable arcana of even modern science, namely, the category of cause and effect.*

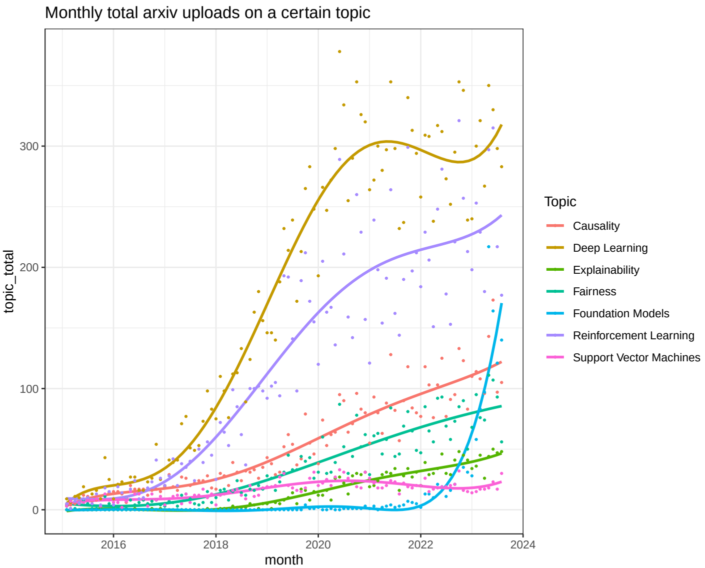
Pearson introduced the correlation coefficient. To him, the slippery concepts of cause and effect seemed outdated and unscientific, compared to the mathematically clear and precise concept of a correlation coefficient.

Causality and statistics

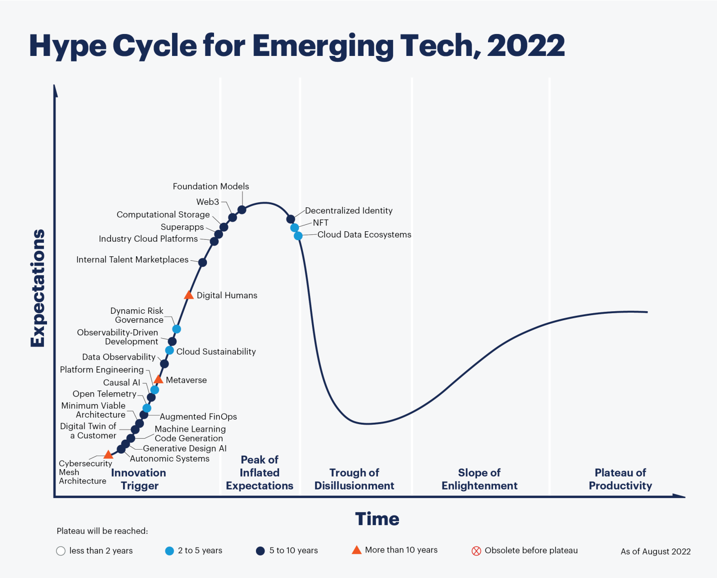
Constructive timeline:

- Wright [1921]: Causal genetics model for guinea pigs (discredited by Pearson)
- Fisher [1925]: Influential advocacy of randomized controlled trials
- Rubin [1974]: Influential mathematical formulation of a causal statistical model
- Dawid [1979]: Proposed the statistical notion of conditional independence
- Robins and Morgenstern [1987]: Estimating causal effects in epidemiology (took 4 years to get published)
- Pearl [1988]: Graphical representation of causal models
- Glymour et al. [1987]: Learning causal structure (graphs) from observational data.

Causal Machine Learning: a hype



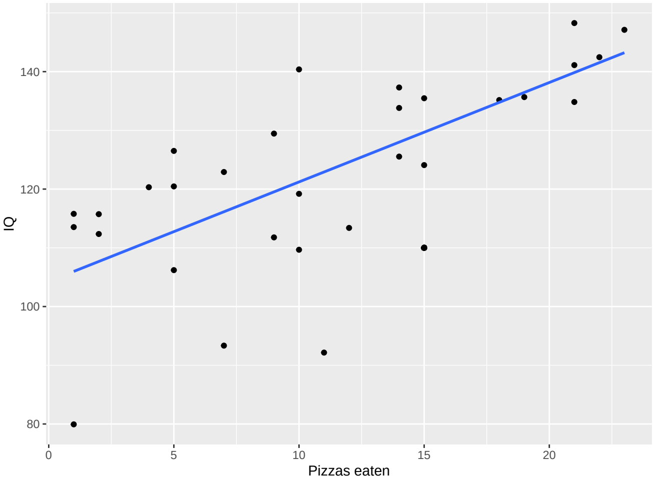
Causal Machine Learning: a hype



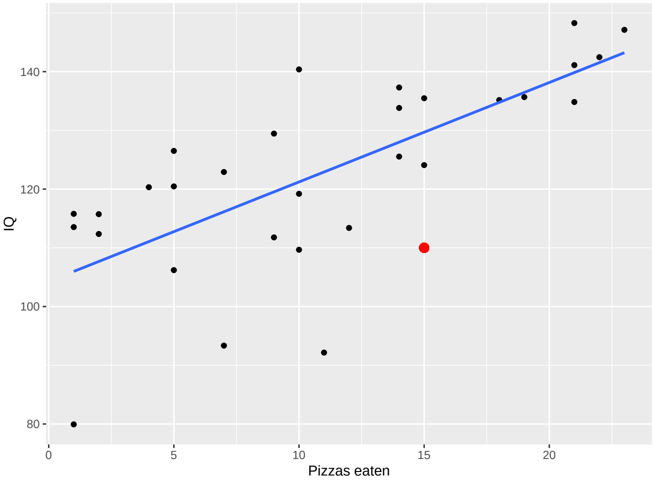
- ‘Neural Causal Models’
- ‘Causal Regression Trees’
- Gartner:
  - greater autonomy
  - robustness
  - adaptability
  - explainability
  - fairness
  - decision support
  - increased AI applicability

Correlation and causation

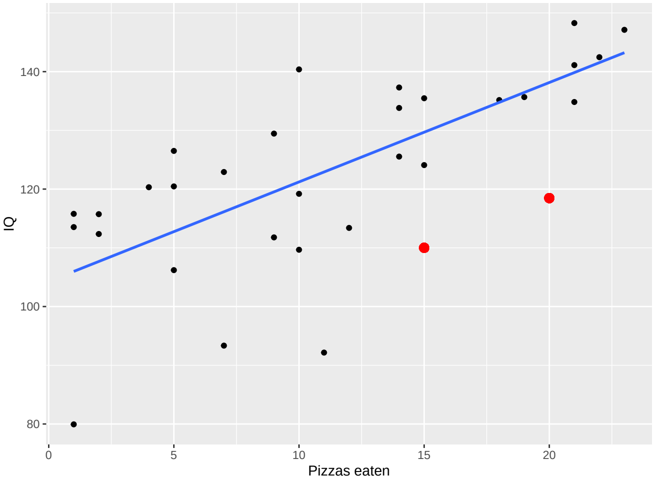
Example: Eating pizza increases your IQ



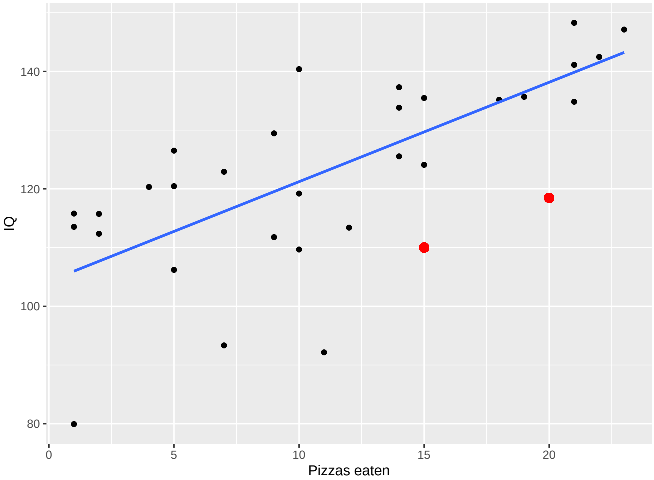
Example: Eating pizza increases your IQ



Example: Eating pizza increases your IQ



Example: Eating pizza increases your IQ

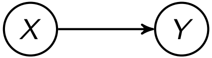


So, eating pizza increases your IQ. But this doesn't seem right, does it?

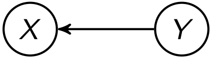
Correlation v.s. Causation

How to explain a correlation between two variables?

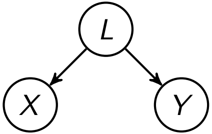
Reichenbach's principle of common cause:[1]
If X and Y are correlated, then we must have one of the following causal relationships:



(a)



(b)



(c)

[1] Reichenbach [1956]

Correlation

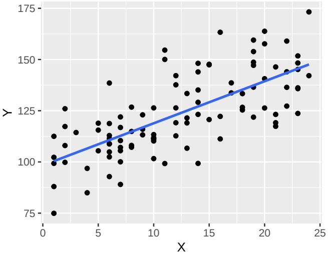
Pearson correlation:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \sqrt{\frac{\mathrm{Var}(X)}{\mathrm{Var}(Y)}} \times \text{the slope of the regression line.}$$

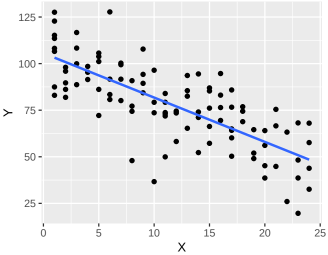
Correlation

Pearson correlation:

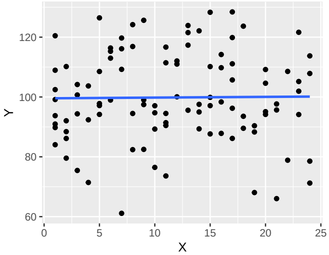
$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \sqrt{\frac{\mathrm{Var}(X)}{\mathrm{Var}(Y)}} \times \text{the slope of the regression line.}$$



(a)



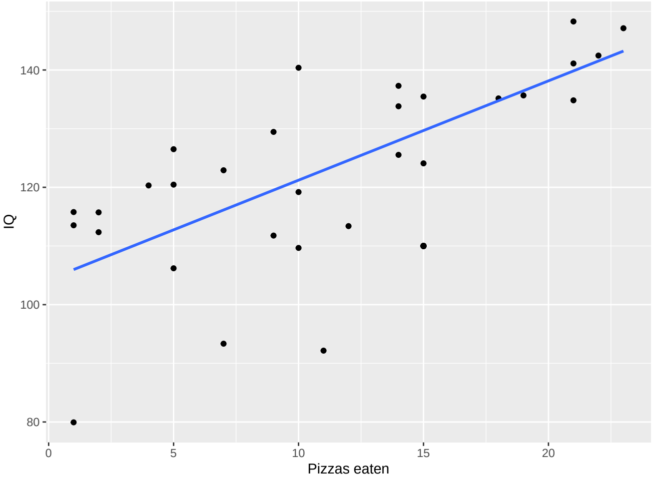
(b)



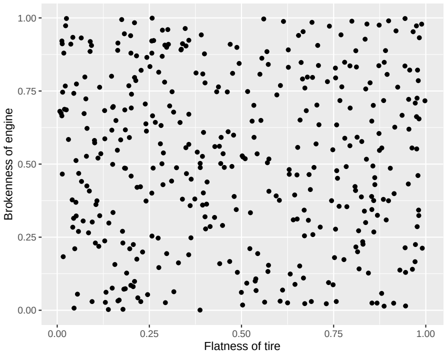
(c)

Example: Eating pizza increases your IQ

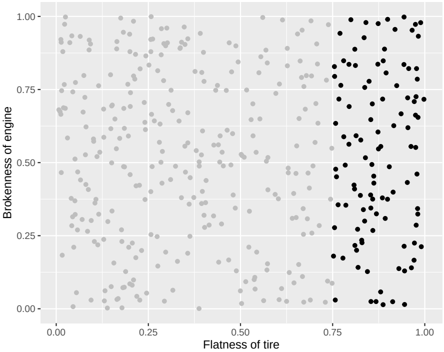
If eating pizza and IQ are correlated, what is the underlying causal mechanism?



Example: Car repair shop

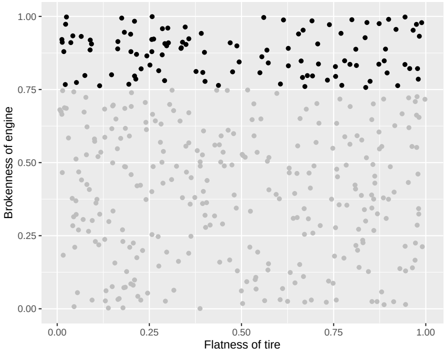


Example: Car repair shop



'flat tire' := 'flatness of tire' > 0.75

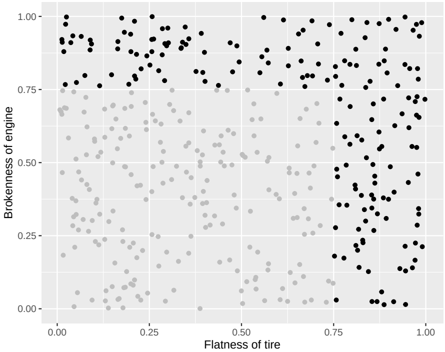
Example: Car repair shop



$$\text{'flat tire'} := \text{'flatness of tire'} > 0.75$$

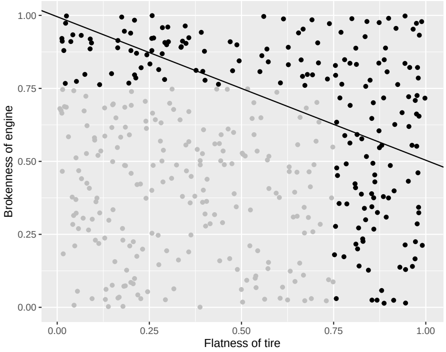
$$\text{'broken engine'} := \text{'brokenness of engine'} > 0.75$$

Example: Car repair shop



$$\text{‘flat tire’} := \text{‘flatness of tire’} > 0.75$$
$$\text{‘broken engine’} := \text{‘brokenness of engine’} > 0.75$$
$$\text{‘car in shop’} := \text{‘flat tire’ OR ‘broken engine’}$$

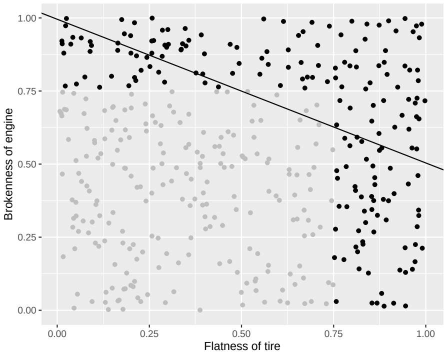
Example: Car repair shop



$$
\begin{aligned}
\text{'flat tire'} &:= \text{'flatness of tire'} > 0.75 \\
\text{'broken engine'} &:= \text{'brokenness of engine'} > 0.75 \\
\text{'car in shop'} &:= \text{'flat tire' OR 'broken engine'}
\end{aligned}
$$

Among the cars brought to the shop, 'flat tire' and 'broken engine' are negatively correlated!

Example: Car repair shop

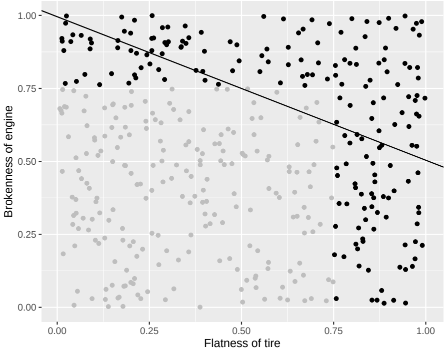


$$\text{‘flat tire’} := \text{‘flatness of tire’} > 0.75$$
$$\text{‘broken engine’} := \text{‘brokenness of engine’} > 0.75$$
$$\text{‘car in shop’} := \text{‘flat tire’ OR ‘broken engine’}$$

Among the cars brought to the shop, ‘flat tire’ and ‘broken engine’ are negatively correlated!

What is the underlying causal mechanism?

Example: Car repair shop



$$\text{'flat tire'} := \text{'flatness of tire'} > 0.75$$
$$\text{'broken engine'} := \text{'brokenness of engine'} > 0.75$$
$$\text{'car in shop'} := \text{'flat tire' OR 'broken engine'}$$

Among the cars brought to the shop, ‘flat tire’ and ‘broken engine’ are negatively correlated!

What is the underlying causal mechanism?

None of Reichenbach’s systems apply. Instead, this is a case of *selection bias*!

Correlation and causation

If X and Y are correlated, then this is explained either by

- $X \to Y$

Correlation and causation

If X and Y are correlated, then this is explained either by

- $X \to Y$
- $X \leftarrow Y$

Correlation and causation

If X and Y are correlated, then this is explained either by

- $X \to Y$
- $X \leftarrow Y$
- $X \leftarrow L \to Y$

Correlation and causation

If X and Y are correlated, then this is explained either by

- $X \to Y$
- $X \leftarrow Y$
- $X \leftarrow L \to Y$
- selection bias

Correlation and causation

If X and Y are correlated, then this is explained either by

- $X \to Y$
- $X \leftarrow Y$
- $X \leftarrow L \to Y$
- selection bias
- functional constraints

Correlation and causation

If X and Y are correlated, then this is explained either by

- $X \to Y$
- $X \leftarrow Y$
- $X \leftarrow L \to Y$
- selection bias
- functional constraints
- $\ldots$?

Correlation and causation

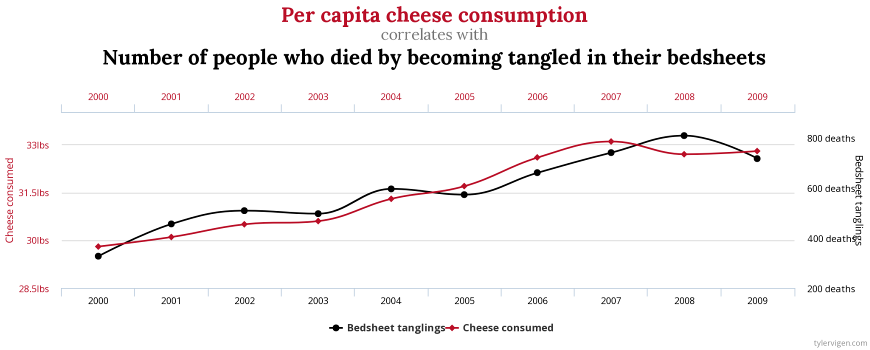
If X and Y are correlated, then this is explained either by

- $X \to Y$
- $X \leftarrow Y$
- $X \leftarrow L \to Y$
- selection bias
- functional constraints
- $\ldots$?

So correlation $\not\Longrightarrow$ causation

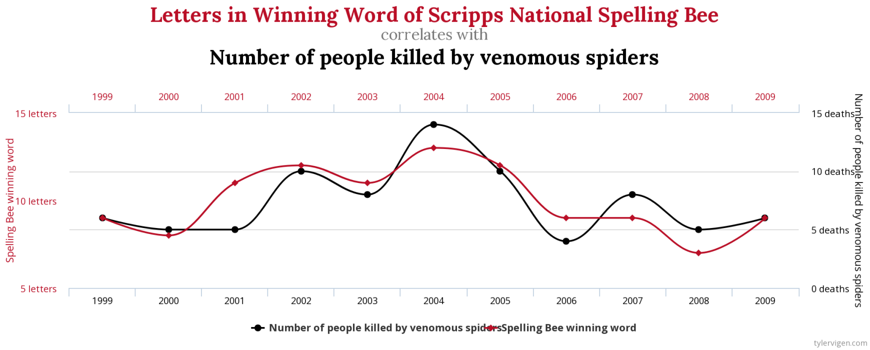
(My current research: how typical is causation without correlation?)
(So causation $\not\Longrightarrow$ correlation)

Spurious correlations



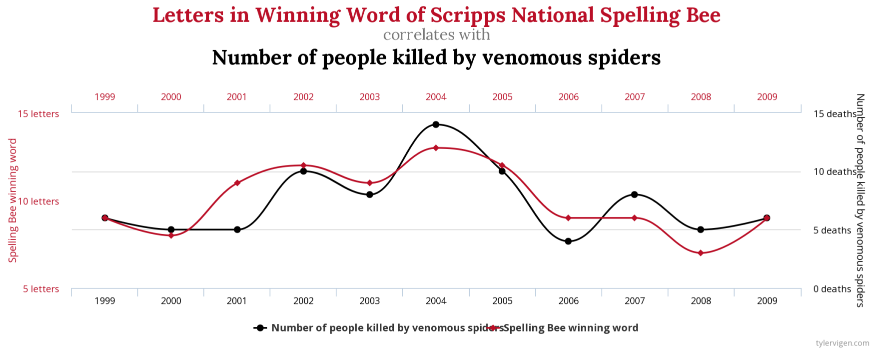
[1] tylervigen.com

Spurious correlations



1 tylervigen.com

Spurious correlations



So, what is going on here?

[1] tylervigen.com

Now, we've seen how correlation can relate to causation.

Is this distinction really important?

Example: drug efficacy

	Recovery	No recovery	Total	Recovery rate
Drug	20	20	40	$\ldots\%$
No drug	16	24	40	$\ldots\%$
Total	36	44	80	

Example: drug efficacy

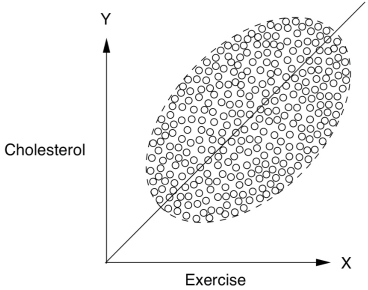
Males	Recovery	No recovery	Total	Recovery rate
Drug	18	12	30	. . . %
No drug	7	3	10	. . . %
Total	25	15	40	

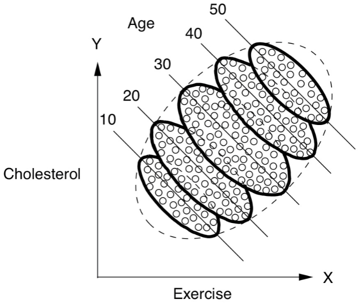
Females	Recovery	No recovery	Total	Recovery rate
Drug	2	8	10	. . . %
No drug	9	21	30	. . . %
Total	11	29	40	

Example: drug efficacy

For the entire population it's better to take the drug, but for any subgroup of the population it's better not to take the drug ?

Simpson's paradox[2]





[2] Simpson [1951]

Okay, so correlation and causation are related, and the latter is more subtle than the former.

When do we care about all this?

Causal effect estimation

Selection bias

Causal discovery

Counterfactuals

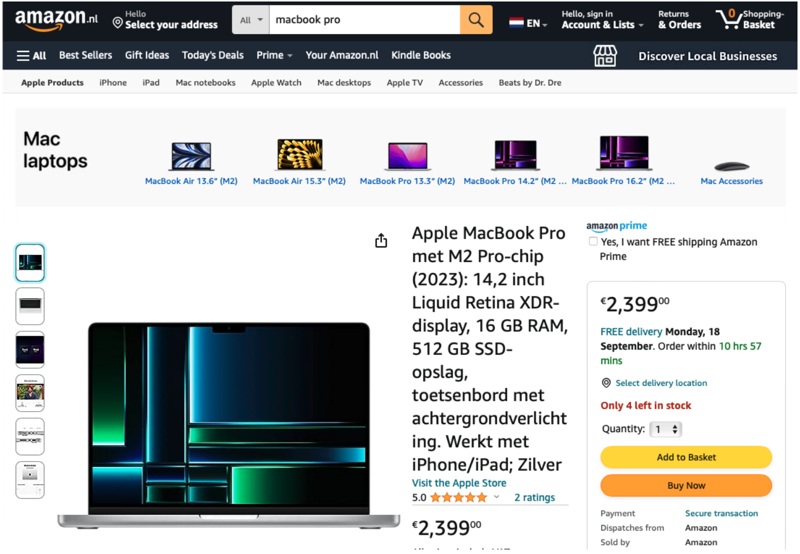
Causal effect estimation

Selection bias

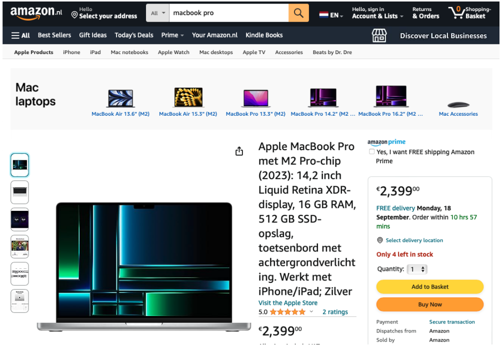
Causal discovery

Counterfactuals

Example: optimizing a webpage



Example: optimizing a webpage



- Decide which color X the "Buy now" button should be
- to maximize the probability that the user will buy the product, Y.

$$X = \arg\max_{x} \mathbb{P}(Y = 1 | X = x)$$

Example: optimizing a webpage

We might have

$$\mathbb{P}(\text{buy}|\text{color} = \text{orange}) = 0.1 < 0.15 = \mathbb{P}(\text{buy}|\text{color} = \text{blue}),$$

so should we always show the blue button?

Example: optimizing a webpage

We might have

$$\mathbb{P}(\text{buy}|\text{color} = \text{orange}) = 0.1 < 0.15 = \mathbb{P}(\text{buy}|\text{color} = \text{blue}),$$

so should we always show the blue button?

This might be a case of Simpson's paradox, where

$$\mathbb{P}(\text{buy}|\text{color} = \text{orange}, \text{dep't} = \text{electr.}) = 0.2 > 0.15 = \mathbb{P}(\text{buy}|\text{color} = \text{blue}, \text{dep't} = \text{electr.}).$$

Example: optimizing a webpage

We might have

$$\mathbb{P}(\text{buy}|\text{color} = \text{orange}) = 0.1 < 0.15 = \mathbb{P}(\text{buy}|\text{color} = \text{blue}),$$

so should we always show the blue button?

This might be a case of Simpson's paradox, where

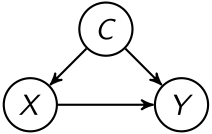
$$\mathbb{P}(\text{buy}|\text{color} = \text{orange}, \text{dep't} = \text{electr.}) = 0.2 > 0.15 = \mathbb{P}(\text{buy}|\text{color} = \text{blue}, \text{dep't} = \text{electr.}).$$

We want to predict the outcome Y if we *intervene* on the color X of the button. Thus, we want to estimate *the causal effect of X on Y*.

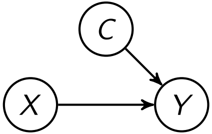
Definition: Causal effect

'Definition': Intervention

When we *intervene* on X, we determine its value without any dependence on other variables.



(a) Graph G

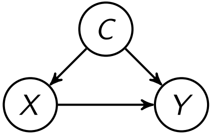


(b) Graph $G_{\mathrm{do}(X)}$

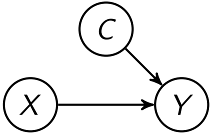
Definition: Causal effect

'Definition': Intervention

When we *intervene* on X, we determine its value without any dependence on other variables.



(a) Graph G



(b) Graph $G_{\mathrm{do}(X)}$

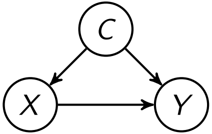
'Definition': Causal effect

The *causal effect* of X on Y is the conditional probability of Y given an intervened value of X, denoted with $\mathbb{P}(Y \,|\, \mathrm{do}(X))$.

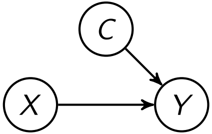
Definition: Causal effect

'Definition': Intervention

When we *intervene* on X, we determine its value without any dependence on other variables.



(a) Graph G



(b) Graph $G_{\mathrm{do}(X)}$

'Definition': Causal effect

The *causal effect* of X on Y is the conditional probability of Y given an intervened value of X, denoted with $\mathbb{P}(Y | \mathrm{do}(X))$.

Rule of thumb:

If $X \leftarrow Y$ or if X and Y are *confounded*, we have $\mathbb{P}(Y|X) \neq \mathbb{P}(Y| \mathrm{do}(X))$.

'Seeing $\neq$ doing'

Seeing $\neq$ doing: exercise 1

Explain why

$$\mathbb{P}(\text{rain}|\text{barometer} = \text{`rain'}) \neq \mathbb{P}(\text{rain}|\,\mathrm{do}(\text{barometer} = \text{`rain'}))$$

Seeing $\neq$ doing: exercise 2

Explain why

$$\mathbb{P}(\text{hair length yesterday}|\text{visit barber today} = 1)$$
$$\neq \mathbb{P}(\text{hair length yesterday}|\,\mathrm{do}(\text{visit barber today} = 1))$$

Seeing $\neq$ doing: exercise 2

Explain why

$$\mathbb{P}(\text{hair length yesterday}|\text{visit barber today} = 1) \neq \mathbb{P}(\text{hair length yesterday}|\,\mathrm{do}(\text{visit barber today} = 1))$$

We don't always want to predict the effect of a cause! E.g. predict nano scale properties from micro scale measurements.

Seeing $\neq$ doing: exercise 3

Explain why:

$$\mathbb{P}(\text{buy}|\text{color} = \text{blue}) \neq \mathbb{P}(\text{buy}|\,\mathrm{do}(\text{color} = \text{blue}))$$

Seeing $\neq$ doing: exercise 4

Explain why:

$$\mathbb{P}(\text{IQ} > 120 | \text{pizza's eaten} = 20) \neq \mathbb{P}(\text{IQ} > 120 | \operatorname{do}(\text{pizza's eaten} = 20))$$

Seeing $\neq$ doing: exercise 5

Explain why:

$$\mathbb{P}(\text{sunshine}|\text{ice cream consumption} = \text{`high'}) \\ \neq \mathbb{P}(\text{sunshine}|\,\mathrm{do}(\text{ice cream consumption} = \text{`high'}))$$

Seeing $\neq$ doing: exercise 6

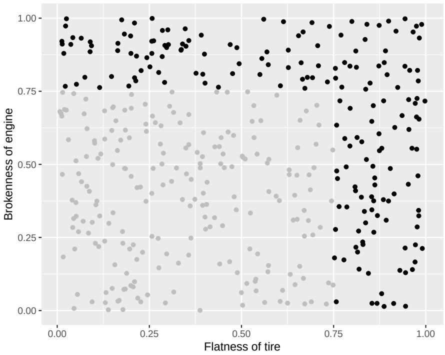
Explain why:

$$\mathbb{P}(\text{recovery}|\text{drug} = 1) \neq \mathbb{P}(\text{recovery}|\,\mathrm{do}(\text{drug} = 1))$$

Seeing $\neq$ doing: exercise 7

Prove that:

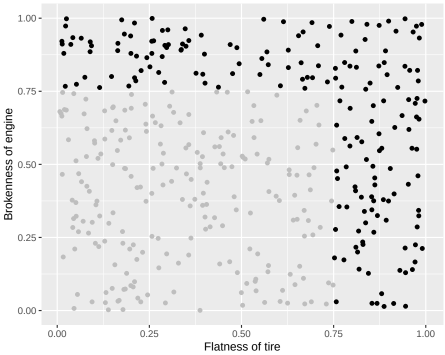
$$\mathbb{P}(\text{broken engine}|\text{Car in shop}) \neq \mathbb{P}(\text{broken engine}|\,\mathrm{do}(\text{Car in shop}))$$



Seeing $\neq$ doing: exercise 7

Prove that:

$$\mathbb{P}(\text{broken engine}|\text{Car in shop}) \neq \mathbb{P}(\text{broken engine}|\operatorname{do}(\text{Car in shop}))$$

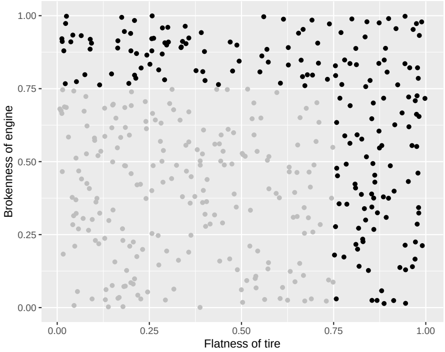


1. Give $\mathbb{P}(\text{broken engine})$
2. Give $\mathbb{P}(\text{broken engine}|\text{Car in shop})$
3. Draw a causal graph G with variables 'broken engine', 'Car in shop', 'flat tire'.
4. Draw the causal graph $G_{\operatorname{do}(\text{Car in shop})}$, i.e. the graph where we intervene on 'Car in shop'.
5. Motivate what is $\mathbb{P}(\text{broken engine}|\operatorname{do}(\text{Car in shop}))$

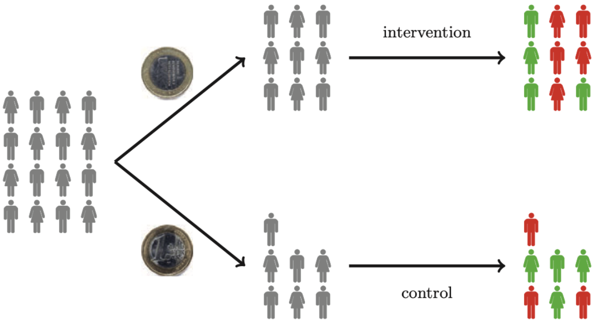
Seeing $\neq$ doing: exercise 7

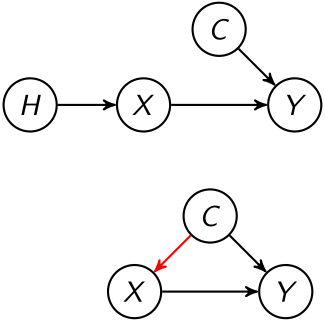
Prove that:

$$\mathbb{P}(\text{broken engine}|\text{Car in shop}) \neq \mathbb{P}(\text{broken engine}|\operatorname{do}(\text{Car in shop}))$$

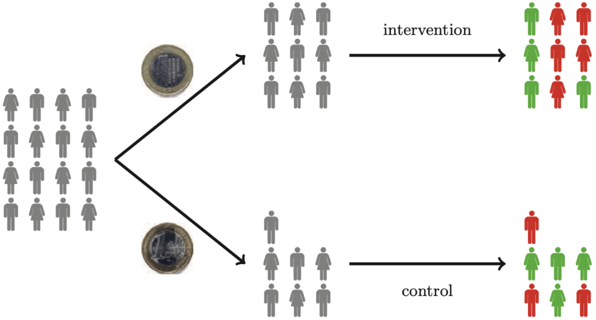


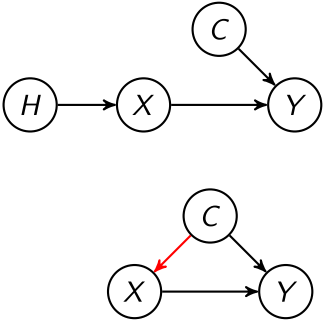
Randomized Controlled Trials





Randomized Controlled Trials





Then there are no common causes of X and Y and Y is not a cause of X, hence $\mathbb{P}(Y = 1 | \operatorname{do}(X = 1)) = \mathbb{P}(Y = 1 | X = 1)$.

Randomized Controlled Trials

Flemish physician Jan Baptista van Helmont [Van Helmont, 1646]:

> *Let us take from the itinerants' hospitals, from the camps or from elsewhere 200 or 500 poor people with fevers, pleurisy etc. and divide them in two: let us cast lots so that one half of them fall to me and the other half to you. I shall cure them without blood-letting or perceptible purging, you will do so according to your knowledge (nor do I even hold you to your boast of abstaining from phlebotomy or purging) and we shall see how many funerals each of us will have: the outcome of the contest shall be the reward of 300 florins deposited by each of us.*

Popularized by Fisher [1925] for smaller confidence intervals of the t-test.

RCT / Causal Effect Estimation

- In software engineering known as A/B testing[3]

[3] Amazon offers their vendors an A/B testing platform.

RCT / Causal Effect Estimation

- In software engineering known as A/B testing[3]
- RCT is not always feasible or ethical: smoking causes lung cancer, eating ultra-processed foods causes obesity, etc.

[3] Amazon offers their vendors an A/B testing platform.

RCT / Causal Effect Estimation

- In software engineering known as A/B testing[3]
- RCT is not always feasible or ethical: smoking causes lung cancer, eating ultra-processed foods causes obesity, etc.
- In such cases, try to estimate the causal effect from observational data by correcting for confounding bias.

[3] Amazon offers their vendors an A/B testing platform.

RCT / Causal Effect Estimation

- In software engineering known as A/B testing[3]
- RCT is not always feasible or ethical: smoking causes lung cancer, eating ultra-processed foods causes obesity, etc.
- In such cases, try to estimate the causal effect from observational data by correcting for confounding bias.
- 2021 Nobel Prize in Economics is won by Angrist and Imbens for estimating causal effects from observational data.

[3] Amazon offers their vendors an A/B testing platform.

RCT / Causal Effect Estimation

- In software engineering known as A/B testing[3]
- RCT is not always feasible or ethical: smoking causes lung cancer, eating ultra-processed foods causes obesity, etc.
- In such cases, try to estimate the causal effect from observational data by correcting for confounding bias.
- 2021 Nobel Prize in Economics is won by Angrist and Imbens for estimating causal effects from observational data.
- Which correction method to apply depends on the causal graph.

[3] Amazon offers their vendors an A/B testing platform.

RCT / Causal Effect Estimation

- In software engineering known as A/B testing[3]
- RCT is not always feasible or ethical: smoking causes lung cancer, eating ultra-processed foods causes obesity, etc.
- In such cases, try to estimate the causal effect from observational data by correcting for confounding bias.
- 2021 Nobel Prize in Economics is won by Angrist and Imbens for estimating causal effects from observational data.
- Which correction method to apply depends on the causal graph.

Knowledge of the causal graph is instrumental for causal effect estimation from observational data.

[3] Amazon offers their vendors an A/B testing platform.

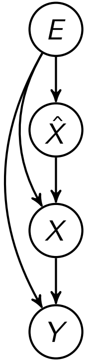
Applications: Decision Support Systems

Non-automated decision making

- For context E
- *advise* action $\hat{X} \in \{x_1, ..., x_n\}$ to optimize the expected outcome of Y

$$\hat{X} = \arg\max_x \mathbb{P}(Y = 1 | E, \mathrm{do}(X = x))$$

- after which the 'user' takes action X
- and we observe outcome Y.



Examples: decision support in healthcare (e.g. PacMed), decision support in legal cases (recidivism risk), child welfare screening, bank loan applications, etc.

[3] Boeken et al. [2023b], Evaluating the Performative Effects of Decision Support Systems

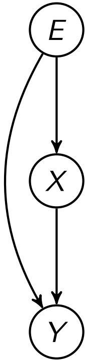
Applications: Contextual Bandits

Automated decision making:

- For context E
- *pick* action $X \in \{x_1, ..., x_n\}$ to optimize the expected outcome of Y

$$X = \arg\max_x \mathbb{P}(Y = 1 | E, \mathrm{do}(X = x))$$

- after which we observe outcome Y.



Examples: layout of online platforms, automated fraud detection, ranking of news items on a webpage.

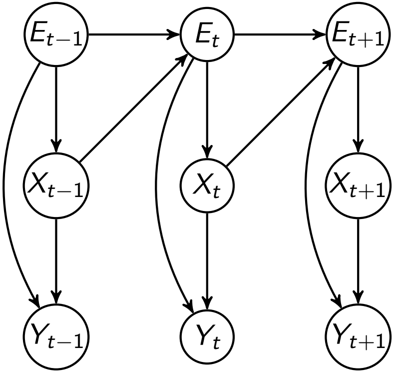
Applications: Reinforcement Learning

Sequential automated decision making:

- At time t
- for context E_t
- pick action $X_t \in \{x_1, \ldots, x_m\}$ to optimize the expected outcome of Y_{t+1}

$$X_t = \arg\max_x \mathbb{P}(Y_{t+1} = 1 | E_t, \mathrm{do}(X_t = x))$$

- after which we observe outcome Y_t
- and we continue to $t + 1$...



Examples: self driving cars, Roomba's, treatment regimes in healthcare, wind farm optimization, cooling Google's data centers, etc.

Summary

We've seen:

- How to draw a causal graph
- What an intervention is
- What a causal effect is
- How to apply causal reasoning to practical cases
- How to estimate a causal effect with an RCT (A/B testing)
- ML problems that can leverage causal effect estimation

Causal effect estimation

Selection bias

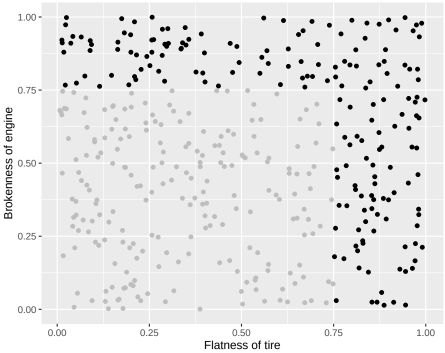
Causal discovery

Counterfactuals

Causal discovery

- To identify a causal effect from observational data, we must know the causal graph of the data generating process.
- In many cases, this graph is not readily available.
  Notable exception: when we are *learning from controlled sources* (e.g. at Booking.com)
- Can we, from observing a system at rest (i.e. not intervening on it), infer the underlying causal structure?
- At the heart of the controversy surrounding causality in statistics, with Pearson and Fisher as strong opponents.
- Since 1980's a serious field of research.

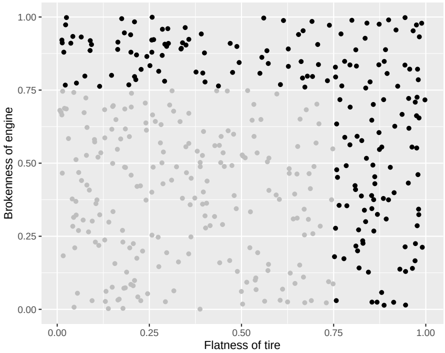
Conditional dependence example: Car repair shop



$$\mathbb{P}(\text{broken engine}|\text{Car in shop, flat tire}) = \ldots$$
$$\mathbb{P}(\text{broken engine}|\text{Car in shop, no flat tire}) = \ldots$$

[4] Dawid [1979]

Conditional dependence example: Car repair shop



$$\mathbb{P}(\text{broken engine}|\text{Car in shop, flat tire}) = \ldots$$
$$\mathbb{P}(\text{broken engine}|\text{Car in shop, no flat tire}) = \ldots$$

So, given information about Z, any information about X provides information about Y as well, written $X \not\perp\!\!\!\perp Y|Z$.[4]

What is the underlying causal mechanism?

[4] Dawid [1979]

Causal discovery: V-structures

- Given data from variables X, Y, Z,

[5] assuming acylicity and no latent confounding

Causal discovery: V-structures

- Given data from variables X, Y, Z,
- if X and Y are statistically independent ($\approx$ not correlated) ('$X \perp\!\!\!\perp Y$')

[5] assuming acylicity and no latent confounding

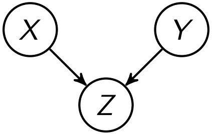
Causal discovery: V-structures

- Given data from variables X, Y, Z,
- if X and Y are statistically independent ($\approx$ not correlated) ('$X \perp\!\!\!\perp Y$')
- but conditioned on Z, they are statistically dependent ('$X \not\perp\!\!\!\perp Y|Z$')

[5] assuming acylicity and no latent confounding

Causal discovery: V-structures

- Given data from variables X, Y, Z,
- if X and Y are statistically independent ($\approx$ not correlated) ('$X \perp\!\!\!\perp Y$')
- but conditioned on Z, they are statistically dependent ('$X \not\perp\!\!\!\perp Y | Z$')
- then the causal graph must be a v-structure:[5]



[5] assuming acylicity and no latent confounding

Constraint-based causal discovery

X_1	X_2	X_3	X_4
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

$\longrightarrow$

$X_2 \not\!\perp\!\!\!\perp X_4$

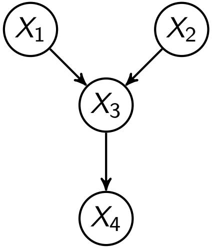
$X_2 \perp\!\!\!\perp X_4 | X_3$

$X_1 \perp\!\!\!\perp X_2$

$X_1 \not\!\perp\!\!\!\perp X_2 | X_3$

etc.

$\longrightarrow$



[5]Actually, the algorithm outputs an equivalence class of graphs, but this is beyond the scope of this presentation.

Application: feature selection

Dataset:

X_1	$\ldots$	X_{12}	Y
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Task:
Make a model to predict Y.

Which features should you use?

[6] Yaramakala and Margaritis [2005]

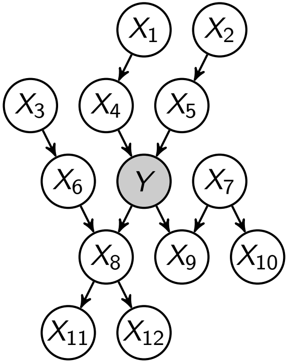
Application: feature selection

Dataset:

X_1	$\dots$	X_{12}	Y
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Task:
Make a model to predict Y.

Which features should you use?



[6] Yaramakala and Margaritis [2005]

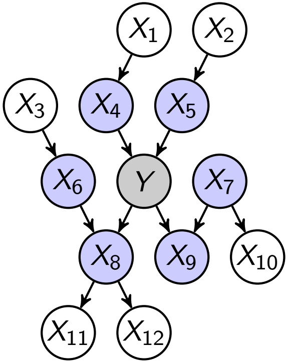
Application: feature selection

Dataset:

X_1	$\dots$	X_{12}	Y
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Task:
Make a model to predict Y.

Which features should you use?



[6] Yaramakala and Margaritis [2005]

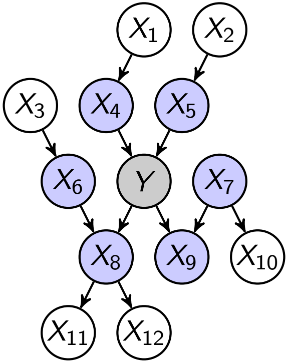
Application: feature selection

Dataset:

X_1	$\dots$	X_{12}	Y
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Task:
Make a model to predict Y.

Which features should you use?



Select the *Markov Boundary*.[6]

[6] Yaramakala and Margaritis [2005]

Applications of Causal Discovery

- Broad Institute of MIT and Harvard (world leading biomedical research center) is betting on causal discovery to predict a genetic modification of human T-cells to improve the cells endurance in fighting cancer.
- London based data consultancy CausaLens leverages Causal Discovery to validate their assumptions of an underlying causal graph for causal effect estimation.

However, it is not (yet) robust:

- General conditional independence testing is a provably 'unsolvable' problem, and
- there is a lack of real-world datasets with a known ground-truth causal graph to validate our algorithms.

Causal effect estimation

Selection bias

Causal discovery

Counterfactuals

Example: Cervical cancer screening

- We have data from Hospital Universitario de Caracas, Venezuela:[7]

[7] Available at https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+(Risk+Factors).

[8] Boeken et al. [2023a], Correcting for Selection Bias and Missing Response in Regression Using Privileged Information

Example: Cervical cancer screening

- We have data from Hospital Universitario de Caracas, Venezuela:[7]
  - X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)

[7] Available at https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+(Risk+Factors).

[8] Boeken et al. [2023a], Correcting for Selection Bias and Missing Response in Regression Using Privileged Information

Example: Cervical cancer screening

- We have data from Hospital Universitario de Caracas, Venezuela:[7]
  - X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)
  - Y: Presence of cervical cancer

[7] Available at https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+(Risk+Factors).

[8] Boeken et al. [2023a], Correcting for Selection Bias and Missing Response in Regression Using Privileged Information

Example: Cervical cancer screening

- We have data from Hospital Universitario de Caracas, Venezuela:[7]
  - X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)
  - Y: Presence of cervical cancer
- Patients in this dataset are self-selected: their own initiative caused them to be recorded in this dataset.

[7] Available at https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+(Risk+Factors).

[8] Boeken et al. [2023a], Correcting for Selection Bias and Missing Response in Regression Using Privileged Information

Example: Cervical cancer screening

- We have data from Hospital Universitario de Caracas, Venezuela:[7]
  - X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)
  - Y: Presence of cervical cancer
- Patients in this dataset are self-selected: their own initiative caused them to be recorded in this dataset.
- Suppose we train a model to predict Y from digitally available features X.

[7] Available at https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+(Risk+Factors).

[8] Boeken et al. [2023a], Correcting for Selection Bias and Missing Response in Regression Using Privileged Information

Example: Cervical cancer screening

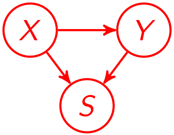
- We have data from Hospital Universitario de Caracas, Venezuela:[7]
  - X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)
  - Y: Presence of cervical cancer
- Patients in this dataset are self-selected: their own initiative caused them to be recorded in this dataset.
- Suppose we train a model to predict Y from digitally available features X.
- Can we use this model in a large-scale, automated screening of the population?[8]

[7] Available at https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+(Risk+Factors).

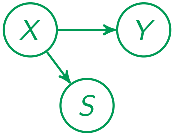
[8] Boeken et al. [2023a], Correcting for Selection Bias and Missing Response in Regression Using Privileged Information

Example: Cervical cancer screening

- We have data from Hospital Universitario de Caracas, Venezuela:[7]
  - X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)
  - Y: Presence of cervical cancer
- Patients in this dataset are self-selected: their own initiative caused them to be recorded in this dataset.
- Suppose we train a model to predict Y from digitally available features X.
- Can we use this model in a large-scale, automated screening of the population?[8]



v.s.



[7] Available at https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+(Risk+Factors).

[8] Boeken et al. [2023a], Correcting for Selection Bias and Missing Response in Regression Using Privileged Information

Causal effect estimation

Selection bias

Causal discovery

Counterfactuals

Example: Groninger HIV case

- I: Victim got injected with HIV infected blood

Example: Groninger HIV case

- I: Victim got injected with HIV infected blood
- U: Victim had unprotected intercourse with potentially HIV infected men

Example: Groninger HIV case

- I: Victim got injected with HIV infected blood
- U: Victim had unprotected intercourse with potentially HIV infected men
- H: The victim contracted HIV

Example: Groninger HIV case

- I: Victim got injected with HIV infected blood
- U: Victim had unprotected intercourse with potentially HIV infected men
- H: The victim contracted HIV
- $\mathbb{P}(H = 1|I = 1) = 1/30$

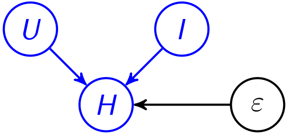
Example: Groninger HIV case

- I: Victim got injected with HIV infected blood
- U: Victim had unprotected intercourse with potentially HIV infected men
- H: The victim contracted HIV
- $\mathbb{P}(H = 1 | I = 1) = 1/30$
- $\mathbb{P}(H = 1 | U = 1) = 1/300$

Example: Groninger HIV case

- I: Victim got injected with HIV infected blood
- U: Victim had unprotected intercourse with potentially HIV infected men
- H: The victim contracted HIV
- $\mathbb{P}(H = 1 | I = 1) = 1/30$
- $\mathbb{P}(H = 1 | U = 1) = 1/300$

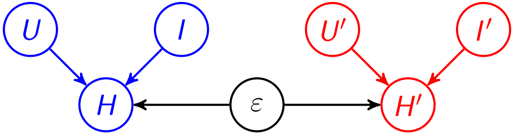
What was the cause of H? The unprotected intercourse U or the injection I?



Example: Groninger HIV case

- I: Victim got injected with HIV infected blood
- U: Victim had unprotected intercourse with potentially HIV infected men
- H: The victim contracted HIV
- $\mathbb{P}(H = 1 | I = 1) = 1/30$
- $\mathbb{P}(H = 1 | U = 1) = 1/300$

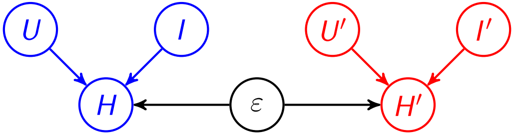
What was the cause of H? The unprotected intercourse U or the injection I?



Example: Groninger HIV case

- I: Victim got injected with HIV infected blood
- U: Victim had unprotected intercourse with potentially HIV infected men
- H: The victim contracted HIV
- $\mathbb{P}(H = 1 | I = 1) = 1/30$
- $\mathbb{P}(H = 1 | U = 1) = 1/300$

What was the cause of H? The unprotected intercourse U or the injection I?



Probability of causation (in a possibly unrealistic model, see Vragovic [2023]):

$$0.9 \leq \mathbb{P}(H' = 0 | U = 1, I = 1, H = 1, U' = 1, I' = 0) \leq 0.91$$

Example: Groninger HIV case

- In 2010 the court of appeal found the defendants guilty of aggravated assault. It is argued that
$$\mathbb{P}(H = 1|I = 1) = 1/30 > 1/300 = \mathbb{P}(H = 1|U = 1),$$
hence I must be the cause of H.

Example: Groninger HIV case

- In 2010 the court of appeal found the defendants guilty of aggravated assault. It is argued that
$$\mathbb{P}(H = 1 | I = 1) = 1/30 > 1/300 = \mathbb{P}(H = 1 | U = 1),$$
hence I must be the cause of H.
- In 2012 the court of cassation ordered a re-trail of the case because of insufficient evidence of I being the actual cause of H.

Example: Groninger HIV case

- In 2010 the court of appeal found the defendants guilty of aggravated assault. It is argued that
$$\mathbb{P}(H = 1 | I = 1) = 1/30 > 1/300 = \mathbb{P}(H = 1 | U = 1),$$
hence I must be the cause of H.
- In 2012 the court of cassation ordered a re-trail of the case because of insufficient evidence of I being the actual cause of H.
- In this re-trail, the defendants are convicted for *attempted* aggravated assault.

Example: Groninger HIV case

- In 2010 the court of appeal found the defendants guilty of aggravated assault. It is argued that
$$\mathbb{P}(H = 1|I = 1) = 1/30 > 1/300 = \mathbb{P}(H = 1|U = 1),$$
hence I must be the cause of H.
- In 2012 the court of cassation ordered a re-trail of the case because of insufficient evidence of I being the actual cause of H.
- In this re-trail, the defendants are convicted for *attempted* aggravated assault.

$$0.9 \leq \mathbb{P}(H' = 0|U = 1, I = 1, H = 1, U' = 1, I' = 0) \leq 0.91$$

Example: Groninger HIV case

- In 2010 the court of appeal found the defendants guilty of aggravated assault. It is argued that
$$\mathbb{P}(H=1|I=1)=1/30>1/300=\mathbb{P}(H=1|U=1),$$
hence I must be the cause of H.
- In 2012 the court of cassation ordered a re-trail of the case because of insufficient evidence of I being the actual cause of H.
- In this re-trail, the defendants are convicted for *attempted* aggravated assault.

$$0.9 \leq \mathbb{P}(H'=0|U=1, I=1, H=1, U'=1, I'=0) \leq 0.91$$

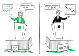
In the process of causal modelling we noticed that pieces of information are missing, making the bounds on the probability of causation uninformative. It seems that causal modelling could be a suitable methodology for gathering and processing statistical evidence in court cases.

Take-aways

Take-aways

- Different ways to explain correlation (some are non-causal).
- What is selection bias.
- Causal effect estimation: seeing $\neq$ doing.
- Randomized controlled trials (A/B testing).
- Applications of causal effect estimation in ML problems.
- The basic concepts behind causal discovery.
- (When to correct for selection bias)
- (What are counterfactuals, and how they can be used to determine the actual cause)

Data Fallacies to Avoid



Cherry Picking

Selecting results that fit your claim and excluding those that don't.

Data Dredging

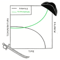
Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.

Survivorship Bias

Drawing conclusions from an incomplete set of data, because that data has 'survived' some selection criteria.

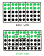
Cobra Effect

Setting an incentive that accidentally produces the opposite result to the one intended. Also known as a Perverse Incentive.



False Causality

Falsely assuming when two events appear related that one must have caused the other.



Gerrymandering

Manipulating the geographical boundaries used to group data in order to change the result.

Sampling Bias

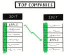
Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.

Gambler's Fallacy

Mistakenly believing that because something has happened more frequently than usual, it's now less likely to happen in future (and vice versa).

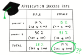
Hawthorne Effect

The act of monitoring someone can affect their behaviour, leading to spurious findings. Also known as the Observer Effect.



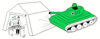
Regression Towards the Mean

When something happens that's unusually good or bad, it will revert back towards the average over time.



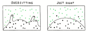
Simpson's Paradox

When a trend appears in different subsets of data but disappears or reverses when the groups are combined.



McNamara Fallacy

Relying solely on metrics in complex situations and losing sight of the bigger picture.

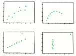


Overfitting

Creating a model that's overly tailored to the data you have and not representative of the general trend.

Publication Bias

Interesting research findings are more likely to be published, distorting our impression of reality.



Danger of Summary Metrics

Only looking at summary metrics and missing big differences in the raw data.

geckoboard

Read more at
geckoboard.com/data-fallacies

References I

P. Boeken, N. de Kroon, M. de Jong, J. M. Mooij, and O. Zoeter. Correcting for Selection Bias and Missing Response in Regression using Privileged Information. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (To appear)*. PMLR, 2023a.

P. Boeken, O. Zoeter, and J. M. Mooij. Evaluating the performative effects of decision support systems, May 2023b.

A. P. Dawid. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979. ISSN 0035-9246. URL https://www.jstor.org/stable/2984718.

R. A. Fisher. *Statistical Methods for Research Workers*. Number 3. Oliver and Boyd, 1925.

C. Glymour, R. Scheines, P. Spirtes, and K. Kelly. *Discovering causal structure: Artificial intelligence*. Philosophy of science, and Statistical Modeling, 394, 1987.

J. M. Mooij. Causality: from data to science. *Nieuw Archief voor Wiskunde*, 24:76–87, 6 2023.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. The Morgan Kaufmann Series in Representation and Reasoning. Morgan Kaufmann, San Francisco, Calif, rev. 2. ed., transferred to digital printing edition, 1988. ISBN 978-1-55860-479-7.

References II

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

H. Reichenbach. *The direction of time*. 1956.

J. M. Robins and H. Morgenstern. The foundations of confounding in epidemiology. *Computers & Mathematics with Applications*, 14(9):869–916, Jan. 1987. ISSN 0898-1221. doi: 10.1016/0898-1221(87)90236-7. URL https://www.sciencedirect.com/science/article/pii/0898122187902367.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, Oct. 1974. ISSN 1939-2176, 0022-0663. doi: 10.1037/h0037350. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/h0037350.

E. H. Simpson. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):238–241, 1951. ISSN 0035-9246. URL https://www.jstor.org/stable/2984065.

J. B. Van Helmont. *Ortus Medicinae, Id Est Initia Physicae Inaudita: Progressus medicinae novus, In Morborum Ultionem ad Vitam longam*. Apud Ludovicum Elzevirium, 1646.

M. Vragovic. *Counterfactual Probabilities in Law: "the Groninger HIV case"*. PhD thesis, 2023.

References III

Wright. Correlation and Causation. 1921.

S. Yaramakala and D. Margaritis. Speculative markov blanket discovery for optimal feature selection. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 4 pp.–, 2005. doi: 10.1109/ICDM.2005.134.