Causal Data Science

Roman Kern

Knowledge Discovery and Data Mining 2 (Version 1.0.4)

> **Motivation**: With purely **observational data** we are not able to answer many questions that one would expect data science to deliver. Taking into the **causal perspective**, one may (with assumptions, or domain knowledge) answer these questions.
> **Goal**: Understand the importance and implications of the **data generation process** and its implications of how to tackle a data science analysis.

Causal Data Science

Outline

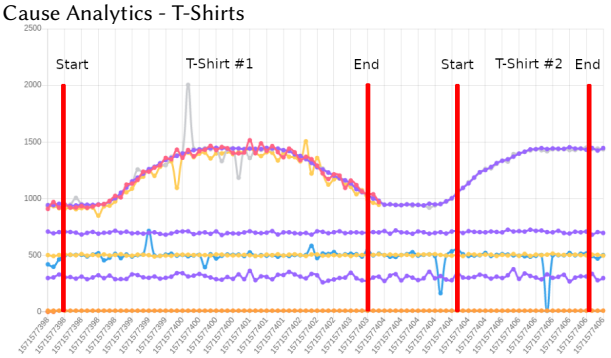
> This lecture can only scratch the surface of causality, so large sections of research are left out.

Overview & Motivation

Gentle introduction to causality, and how we ended up here...

Overview & Motivation

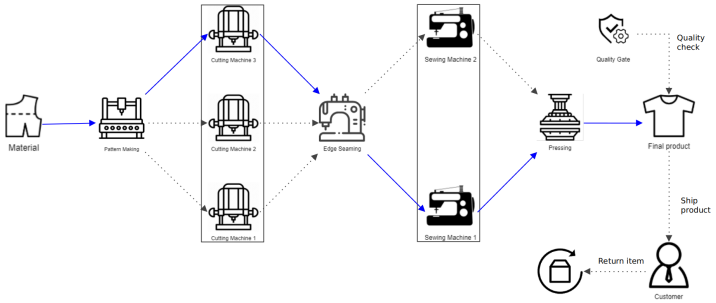
Root Cause Analytics - T-Shirts



> Image a factory that produces **t-shirts**.
> Problem: some of the t-shirts have **defects**.
> Task: **Root cause analytics** to find out, what part of the production process steps is associated (i.e., causally related) with these faults.
> Data to solve this task: **longitudinal** data (mostly time series data) from around the shop floor.
> Spoiler alert: we need **domain knowledge** to better understand the data generation process (e.g., the causal effects).
> We need domain knowledge just to correctly segment our data.

Overview & Motivation

Root Cause Analytics - T-Shirts



> Each shirt is produced in multiple steps, each step may have multiple (semi-)identical machines and each machine provide a number of data (e.g., time series data).
> The arrows present the path a t-shirt takes throughout the production process, this may already be the base for what we will later call a causal graph.
> And already we can use time to our advantage, the root cause need to always precede the effect.
> Knowing the production process will immensely help us in our task!

Overview & Motivation

Starting Point

- Correlation does not imply causation
- Post hoc ergo propter hoc

> We all learnt that we cannot jump to conclusions about the **true nature**, just given observations.
> "Since event Y followed event X, event Y must have been caused by event X".
> In the 20th century we learnt to avoid phrases like "X causes Y", and go for the more vague/safe phrase like "X is associated with Y".
> The "Book of Why" of Judea Pearl gives a nice history lesson.
> Today, we progressed forward and better understand, when (exactly) we are allowed to state "X causes Y" given just observational data.

Overview & Motivation

Regression to the Mean

- The magazine "**Sports Illustrated**" features successful athletes on its cover
- But once they appear on the cover, their performance drops.
  - → "The Sports Illustrated Cover Jinx"
- It can be explained by the regression to the mean
- Or, via reverse causation
  - i.e., good performance caused the cover, and the cover did **not** cause bad performance

> The sports illustrated curse!
> There appears a solid causation (title followed by dip in performance), but in fact the good performance prior to the title page caused the title page.
> There is even a hastag on Instagram: https://www.instagram.com/explore/tags/sicurse/
> And it is mentioned in Kahneman's book, Thinking fast, thinking slow.
> **Initial insight**:
> Correlation is symmetric, causation is directed.

Overview & Motivation

Role of Causality in Data Science

- The gold standard to measure effects are randomised controlled experiments
  - In practice they often **cannot be conducted**
  - **A-B testing** is a form of such experiment
    - Make use, if possible
- **Data-driven causal inference** as next best option

> **Randomised controlled trial** (RCT):
> - Want to study the impact of a treatment
> - Have a (large) number of people
> - Assign people randomly into 2 groups: gets treatment, don't get treatment (without them knowing)
> - Measure the difference
> Since the only difference is the treatment, any change can be attributed to the treatment.
> Many reasons, why randomised controlled trials cannot be conducted: ethical, financial, practical.
> One needs many participants (instances, e.g., t-shirts).
> Data-driven causal inference = causal inference from observational data.

Overview & Motivation

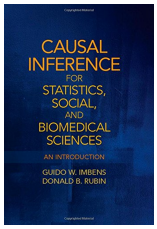
Nomenclature

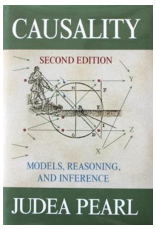
Terminology	Alternatives	Explanation
causality	causal relation, causation	causal relation between variables
causal effect	-	the strength of a causal relation
instance	unit, sample, example	an independent unit of the population
features	covariates, observables, pre-treatment variables	variables describing instances
learning causal effects	forward causal inference, forward causal reasoning	identification and estimation of causal effects
learning causal relations	causal discovery, causal learning, causal search	inferring causal graphs from data
causal graph	causal diagram	a graph with variables as nodes and causality as edges
confounder	confounding variable	a variable causally influences both treatment and outcome

Overview & Motivation

Main Approaches

- Potential Outcomes by *Donald Rubin*
- Structural Causal Models (SCMs) by *Judea Pearl*





Overview & Motivation

Recommended Literature

Suggested reading sequence

1. Glymour, M. M. and Greenland, S. (2008) *'Causal diagrams'*, Modern epidemiology. Lippincott Williams & Wilkins Philadelphia, PA, 3, pp. 183–209.
2. Guo, R. et al. (2020) *'A Survey of Learning Causality with Data'*, ACM Computing Surveys, 53(4), pp. 1–37. doi: 10.1145/3397269.
3. Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect.* Basic Books.
4. Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer.* John Wiley & Sons.

Overview & Motivation

Recommended Resources

- Introduction to Causal Inference by *Brady Neal*, https://www.bradyneal.com/causal-inference-course
- Causal Data Science by *Adam Kelleher*, https://medium.com/causal-data-science/causal-data-science-721ed63a4027
- Causal Data Science with Directed Acyclic Graphs by *Paul Hünermund*, https://www.udemy.com/course/causal-data-science/

> See: Guo, R. et al. (2020) 'A Survey of Learning Causality with Data', ACM Computing Surveys, 53(4), pp. 1–37. doi: 10.1145/3397269.
> In data science, we are mostly interested into learning causal effects, i.e, we know (via domain knowledge) the causal relationships, and with observational data we estimate the strength of a relationship (instead of conducting a randomised controlled experiment).
> Often, the cause is called treatment and the effect is called outcome - this is for historic reasons (as causality mostly progressed in these areas).
> Features are often also called independent variables, especially in a setting, where one wants to predict the dependent variable (also called target).
> Relationship to **classical statistics**: see if there is an effect: statistical hypothesis testing, e.g. via p-values → causal discovery, measuring the strength of the effect: effect size, e.g. via correlation → causal inference.
> Two frameworks for causal learning.
> See also: https://blog.methodsconsultants.com/posts/pearl-causality/.
> SCMs are often preferred when learning causal relations among a set of variables, and PO for learning the strength of relations.

Good book for find a match for practical settings:
Hernán MA, Robins JM (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.
https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

> Also interesting, the causal inference tutorial: https://github.com/amit-sharma/causal-inference-tutorial/
> Also good starting point, a four-part lecture on YouTube by *Jonas Peters*: https://www.youtube.com/watch?v=zvrcyqcN9Wo

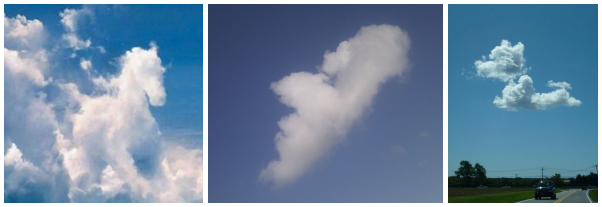
Correlation without Reason

When do we observe correlations that we would not expect?

Correlation without Reason

Motivation

- **Correlation analysis** is a central part of data science
- ... but are there cases, where correlations exists **without proper reason**?



> In data science the correlation analysis (e.g., **pairwise correlation** of all variables, including the target variable) is often one of the first steps of the **exploratory data analysis** phase. Often with the goal to gain a better understanding of the dataset, or to already select (or ignore) certain variables (**feature selection**).
> With correlation analysis we also include notions like **conditional probability**.
> Example: In a production environment one wants to identify **defective items** and wants to understand the root causes, i.e., what sensor data correlates with the defects.
> Even **humans see correlations** (make associations), where there are no real reasons for these correlations, e.g., clouds just happen to look like a horse, dog, etc. Note: Even worse, humans often make causal assumptions starting with purely observational data (correlations).

Correlation without Reason

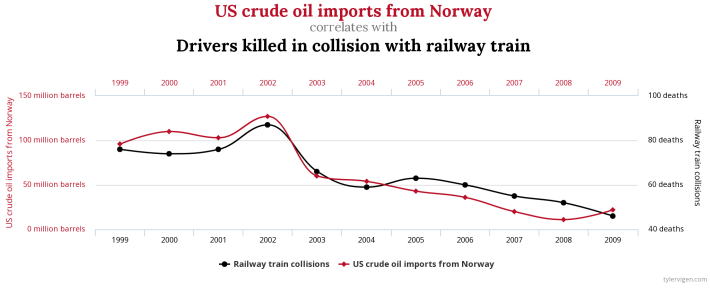
Overview

1. Spurious Correlation
2. Confounders
3. Berkson's Paradox
4. Simpson's Paradox

> Here we will look at **some key scenarios**, which lead to cases, where one may observe correlations in observational data (i.e., a data set), which do not represent the data generation process.

Correlation without Reason

Spurious Correlations



http://www.tylervigen.com/spurious-correlations

> To clarify, there is no reason to believe that "oil imports" and "killed drivers" are somehow connected.
> There are many documented cases where correlations just **happen due to random chance**. In other words, the correlation we see is just bad luck.
> There is a connection to statistical tests, if the p-value falls below a previously defined α-value, we may only assess that the probability of observing a certain phenomenon due to randomness is below our chosen threshold.
> When multiple hypothesis are considered (i.e., each correlation b/w two variables is considered a hypothesis, thus for n variables there are n^2 hypothesis), the chance of observing at least a single spurious correlation **rises quadratically**.

Correlation without Reason

Spurious Correlations

- In big data settings one often **combines different data sets**
- → might be a source of spurious correlations

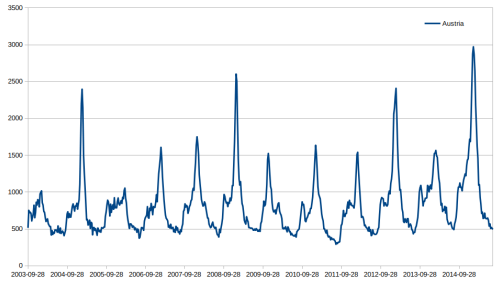
Example

Different **settings** on how the data sets have been collected, one data set with only *bad quality*, and a second data set from only the *night shift* → it will appear as if the night shift produces better quality!

> Besides purely random reasons for spurious correlations, there are cases of **systematically introduced correlations**.
> In fact, some authors consider the fusion of multiple data sets (data sources) and a key component of the **definition of big data**.
> Dataset differ in their **sampling bias** will often cause spurious correlations to happen (as the *distributions* (of many variables) will differ).
> In practice, often "special datasets" are collected with "special properties" (e.g., many outliers)
> This type of spurious correlation is similar to the Berkson's paradox (= Berkson's bias, collider bias).
> To **detect** such correlations: introduce a new "synthetic" variable with the name of the dataset, if now there are some correlations with this variable → indicator for sampling bias.

Correlation without Reason

Spurious Correlations



https://www.google.org/flutrends/about/data/flu/at/data.txt

> In 2008, researchers at Google made a **nowcast** of the flu based on search query terms (the more people search with specific terms, the more flu infections are assumed to be there).
> The idea was published in **Nature**, claiming to be able to accurately predict the flu 2 weeks before the official (assuming based on input for doctors).
> The data was calibrated using ground truth (from the health organisations), to match the past, but due to the high amount of search queries some were considered to be highly predictive, without being related to the target (flu).

Correlation without Reason

Spurious Correlations

- Failed in 2013 by being **140% off**!
- Reason: overfitting to spurious data

Spurious correlations

Among the predictive search queries are seasonal terms like: "*high school basketball*"

> A critical analysis of the Google Flu was then published in **Science**, which shows that the Google model did **overfit on the data** https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Correlation without Reason

Spurious Correlations

- **More data**
  - More instances
  - e.g., held out data to confirm found correlations
- **Less data**
  - Fewer variables
  - e.g., feature selection based on input from domain experts
- **More knowledge**
  - e.g., validation of found patterns

> So, how can we now prevent (or minimize the risk of) spurious correlations, what **strategies** are there?
> More data should also apply to consider **keeping the distributions the same** (of multiple dataset, if they are merged).
> Also, more **diversity** helps, e.g., in the example of the defective items, multiple root causes in the data may help to prevent spurious correlations.
> Here fewer variables (= features) are equivalent to **hypothesis**.
> For example, one could look for all the search terms of the Google Flu prediction and sort out all non-health related queries.
> A variation of more data is: **more (diverse) root causes**, to increase the variability of the phenomenon to study and hence decrease the chance of spurious correlations.

Correlation without Reason
Confounding Factors

Example

- People with *healthy lifestyle*
- ... tend to eat more healthy
- ... exercise more
- ... smoke less
- ... weigh less (lower BMI)

→ correlation between **less smoking** and **low BMI**.

> ... and **all other variables** influenced by healthy lifestyle.
> Many people assume that smoking is associated with lower BMI, but even if this is not true, one would still assume smoking and BMI to be **independent**.
> Why do we see a positive correlation here, if they are independent?
> ... because they have a **common cause**.

Correlation without Reason
Confounding Factors

When is this a problem?

- ... if one is interested on the root cause of low/high BMI
- ... and *healthy lifestyle* is not in the data

→ In this setting, healthy lifestyle is a confounding factor for the relationship between smoking and BMI.

> In short, based on the data one would assume that **quit smoking** might lower the BMI.
> Technically, "healthy lifestyle" is a confounding factor even, if it would be observed (in this case it would be way easier to identify the common cause).
> If it is not observed it is **hard to obtain** the true relationship between smoking and BMI.
> This also applies to many cases in the industry. While one would like to find/identify variables that correlate with e.g. **bad quality**, such confounding factors imply relationships that do not exist (and occlude true relationships).
> Confounders are also called **lurking variables** (if not observed).
> The correlation between smoking and BMI is also called **partial correlation**.

Correlation without Reason
Confounders

- **More data**
  - More variables
  - e.g., include all potential confounders in the dataset
- **More data**
  - More instances, as we need to control for confounders
  - e.g., split into healthy/non-healthy groups
- **Less data**
  - Fewer variables
  - e.g., reduce the collinearity
- **More knowledge**
  - e.g., known confounders (and their influence)

> So, how can we now prevent (or minimize the risk of) spurious correlations, what **strategies** are there?
> More data is related to include all **possible influence factors** (confounders).
> We need to control for each value of the confounder, e.g., healthy and non-healthy instances individually as **bins** (i.e., creating two results, which might be combined via a weighted average, where the weighting needs to be based on the proportion in the population (not in the sample)), see *adjustment formula*.
> Condition on confounders –> smaller bins –> skewed data sets.
, i.e., controlling for variables may create **skewed datasets**.

Correlation without Reason
Berkson's Paradox

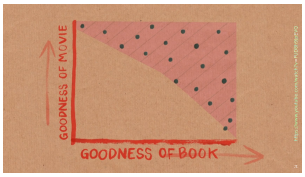
> Typically, if one has read a really good book, then its movie version is disappointing. Vice versa, there are many cases of really good film, where the book is just mediocre. But, there are seemingly **only few examples of good book and good film**!
> This is called the Berkson's Paradox.
> Also called Berkson's bias or collider bias.

Correlation without Reason

Berkson's Paradox

- The selection of books to make movies from is **not random**!
- ... because we rarely observe the combination
  - Bad book, and
  - Bad film

→ creating a skewed distribution



> We do not observe (we did not sample, or they do not end up in out dataset) the **full population**, hence creating an artificial negative association between variables!
> Can be seen as an **opposite of the confounder** example: instead of one common cause and multiple effects, we observe multiple causes and a single effect (=the selection).
> This type of **selection bias** is common in many real-world datasets.
> Another classic example is **wet sidewalk** (pavement), due to rain or sprinkler.
> Also present, in cases of **multiple root causes** within a dataset, each having an internal correlation structure, which due to the sampling bias also mixed up (correlation b/w indicators of different root causes).

Correlation without Reason

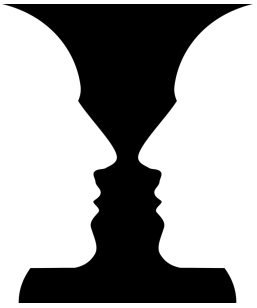
Berkson's Paradox

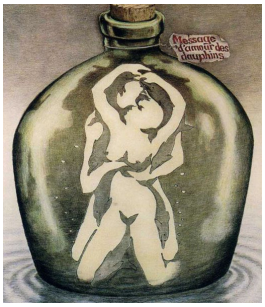
- **More data**
  - More instances, including all combinations (if possible)
  - e.g., also equal amount of bad books/films
- **More knowledge**
  - Document all sampling strategies
  - e.g., identify potential colliders, constraints (not plausible)
- **More data**
  - More instances, to allow for controlling → binning
  - e.g., tread the respective other variable as confounder

> Often, a fair sampling in **not possible**.
> Then, the other (spuriously correlating) variable(s) can be treated as counfounders, and need to be controlled for. Again, by binning with the risk of **small bins and skewed datasets**.
> Another domain knowledge might be the **plausibility check**, e.g., it does not make sense that the better the book, the worse the film!
> In many cases it only allows to detect implausible correlations, but **not correct for**.
> Another example: **Multiple root causes** in the data cause a collider, causing the root causes to be correlated with.

Correlation without Reason

Simpson's Paradox

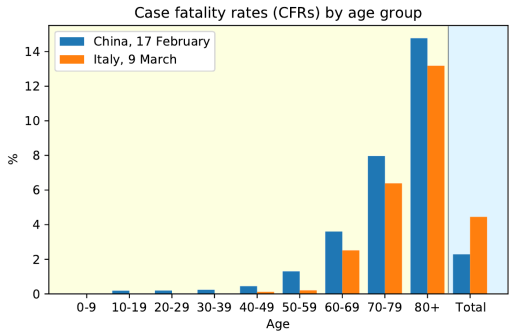




> What do you see? A vase or a 2 faces on the left side, and dolphins or a couple on the right side?
> **Both are correct**, but it depends on the interpretation.
> The same (or at least similar) phenomenon can we **observe in data**, but in data (with the help of additional information), we can even answer **which is (more) correct**.

Correlation without Reason

Simpson's Paradox



> **Question**: Is the CRF higher in Italy than in China? I.e., is the total with the blue background correct, or the individual age groups with the yellow background? (CFR = Case Fatality Rate = What fraction of people die being diagnosed with Covid-19)
> First approach: It depends on the question!
> For the question: I am an Italian, are my chances better than a Chinese, **the answer is no**.
> For the question: I am an Italian and 33 years old, are my chances better than a Chinese of equal age, **the answer is yes**.
> We cannot answer the question: I am an Italian and 33 years old, would be chances be better, if I would live in China.
> With the help of domain knowledge we can answer the question: To answer the original question, do we need to **control for age**?

Correlation without Reason
Simpson's Paradox

Explanation

- **Both variables** (country, age) have an influence on CFR
- The **relations** between the variables and their **strength**
- ... determine what we see

Solution

- Country influences age more...
- ... the total (blue = Italy worse than China) is correct.

> The difference is whether the data generating process conforms to a mediator or confounder.
> The solution is actually an **assumption**, i.e., **age is not a confounder**.
> It would be vice versa, if the age would be more influential, i.e., if old people decide to move to Italy.

Correlation without Reason
Simpson's Paradox

Observations

1. **Correlation**, where there should be **none**
2. **No correlation**, where there **should be**
3. **Reversal of outcomes**

> Not only limited to correlations, but also for other types of associations (e.g., Italy better than China, treatment A better than treatment B, ...)

Correlation without Reason
Simpson's Paradox

- **More knowledge**
  - Understand/document data generation
  - e.g., identify potential mediators, confounders, etc.
- **More data**
  - More instances per variable value
  - e.g., enough people per age group

> The domain knowledge is most important here.

Correlation without Reason
Summary

Reasons

- Randomness
  e.g., too many variables
- Data generation
  e.g., confounder
- Data collection
  e.g., sampling
- Data processing
  e.g., fusion of datasets

Solutions

- Domain knowledge
  e.g., implausible, dependencies
- More data
  e.g., fair sampling, more (controlled) experiments
- Assumptions
  e.g., smoothness, complete dataset
- Constrains
  e.g., time

> Assumptions are then typically made by the data scientist.
> The assumption of complete dataset (there are no unobserved confounders), is also called **sufficiency** in literature.

Potential Outcomes

Causal Framework proposed by Donald Rubin

Potential Outcomes
Motivation

Based on the notion of treatment and outcome

- With the **treatment** $T_i \in \{0, 1\}$
  - ... and i indicating the instance, e.g., patient
- Then the **outcome** y_i^T consists of
  - y_i^0 outcome, not receiving the treatment
  - y_i^1 outcome, if received the treatment

> https://blog.methodsconsultants.com/posts/pearl-causality/
> For example, the treatment could be a drug (potential cure) a patient receives (vs. a placebo).
> The outcome is here, if the patient recovered.
> Recommended literature:
> Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association, 100(469), 322-331.

Potential Outcomes
Definition

Potential outcome

- Given the treatment and outcome: t, y
- The **potential outcome** for instance i: y_i^t
  - Outcome one would had observed, if i received treatment t

> Potential outcomes is modelled after randomised controlled experiments.
> Hard part: isolate the individual effect of the treatment.

Potential Outcomes
Average Treatment Effect

Causal effect of intervention

- Difference in outcome(s)
- Individual Treatment Effect (ITE)
  - $\tau_i = y_i^1 - y_i^0$
- Average Treatment Effect (ATE)
  - $\text{ATE} = \mathbb{E}[\tau_i] = \mathbb{E}\left[y_i^1 - y_i^0\right]$

> Since we want to measure how well our drug performs.
> ITE is on **patient level**, as the difference between potential outcomes of a certain instance under two different treatments.
> ATE is defined on **population level**, since we cannot administer a drug to a patient and not doing it at the same time.
> There is also conditional average treatment effect (CATE) for analysis of specific sub-populations.
> Please note, the ATE is **not specific** to potential outcomes.

Potential Outcomes
Assumptions

> SUTVA can be split into two assumptions

Stable unit treatment value assumption (SUTVA)

- Well-defined treatment levels
  - Same treatment value → same treatment
- No interference
  - The potential outcome is not influenced by other instances' treatment

Potential Outcomes
Assumptions (cont.)

> Additional assumption that needs to hold.

Consistency

- Outcome is independent of treatment assignment process

Potential Outcomes
Assumption - Ignorability

> We should not select the treatment based on the patient (or her condition) - see Wikipedia example on Simpson's Paradox on kidney stones!
> Also called unconfoundedness.

- Treatment should be independent from outcome
  - $Y_i^0, Y_i^1 \perp\!\!\!\perp T_i$
  - This assumption is called **ignorability** (unconfoundedness)
- There are **different ways** to achieve this
  - Randomised controlled experiment
  - Propensity score matching
  - Regression discontinuity
  - Instrumental variables
  - ...

Potential Outcomes
Matching methods

> Propensity score matching being a special case of matching methods.

- Divide the data into **groups** (strata, bins)
  - Grouping is defined via a function, $f(\mathbf{x})$
  - ... with $\mathbf{x}$ being the features
  - ... to create homogeneous groups
  - i.e., they differ just in treatment and potential outcome
- Each group is treated as randomised controlled experiment
  - Compute the ATE between groups

Potential Outcomes
Propensity score matching

- The **propensity score** is such grouping function
  - $f(\mathbf{x}) := P(t \mid \mathbf{x})$
  - The probability of receiving a treatment
- Needs to be estimated
  - e.g., via Logistic Regression

Structural Causal Model

Causal Framework proposed by Judea Pearl

Structural Causal Model
Structural Equation Models (SEM)

- Structural Equation Models (SEM)
  - e.g., $Z = b_0 + b_1 X + b_2 Y$
- **Structural Causal Models** (SCM)
  - Without assuming a functional form
  - Consider random variables $Z_1, \ldots Z_n$, and for each
  - $Z_i = f_i(PA_i, U_i)$
  - where PA_i are the direct parents,
  - and U_i is a noise term

> SEMs are often assumed to be linear, and parametric (there are important exceptions).
> SEMs are used in statistics since a long time, more popular in the 70ties.
> X_i can also be seen as observables.
> Noise/unexplained terms U_i are often **omitted** for brevity, but typically assumed to be there.
> Furthermore, the U_is are jointly **independent** from each other.
> There is a single noise term for each variable, which represent **all influences** outside of the model (confounders, measurement noise, ...).
> The dependencies given by the structural causal model (i.e., which parent each node X_i has can also be represented by a graph, the **causal graph**).
> For computer scientists: A SCM is a program to generate data (following the respective distributions).

Structural Causal Model
Pearl's do() Notation

- Conditional distribution for **intervention** on X
  - $P(Z|do(X = x))$
- If we can compute this, we can compute the **causal effect**
  - $P(Z = z \mid do(X = 1)) - P(Z = z \mid do(X = 0))$
- Average treatment effect
  - $\text{ATE} = \mathbb{E}[y \mid do(t = 1)] - \mathbb{E}[y \mid do(t = 0)]$

> The do() operator represents the intervention on the variable,
> e.g., in a dataset of smokers and non-smokers, artificially set all to non-smokers.
> So, we only need to be able to compute P(Z | do(X)).
> ... it turns out that this is possible (in certain cases).
> Important take away: the **interventional** distribution, $P(Z|do(X = x))$, is not the same as **conditional** distribution $P(Z|X = x)$ (even if there are cases, where there are identical)

Structural Causal Model
Pearl's do() Notation

- We need to map the do() operator to s/t we can compute
  - $P(Z|do(X = x))$
- It might be
  - $P(Z|do(X = x)) = P(Z)$, or
  - $P(Z|do(X = x)) = P(Z|X = x)$, or
  - $P(Z|do(X = x)) = \sum_{y \in Y} P(Z|X = x, Y = y)P(Y = y)$
  - ... even more complex

Structural Causal Model
Pearl's do() Notation

> Additionally, Pearl makes the suggestion to prefer causal graphs, instead of SEMs/SCMs since humans better cope with graphical representations.

How to map?

The mapping of the interventional space to the observational space, i.e., the realisation of $P(Z|do(X = x))$, depends on the **causal structure**!

Causal Graph

Simple graphical language to capture (relevant aspects of) the data generation process

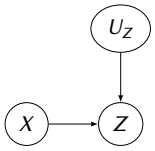
Causal Graph
Causal Graph

> Often unobservable and observable variables are included, typically the ones we can measure are grey.
> The graphs do not need to be acyclic, but in practice it is hard to model cyclic dependencies.
> Note: in difference with Bayesian networks, causal graphs can be **manipulated via interventions**.
> Note: The causal graph might **be part** of a causal model, but a causal graph alone is not a complete causal model.

- Causal graph
  - Extension of **Bayesian networks**
- Nodes represent variables/observables/...
- Edges represent causal relationships
  - With an arrow pointing from the **cause to the effect**
- → Directed acyclic graphs (DAG)

Causal Graph

Causal Graph

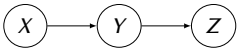


Represents the SCM: $Z := f_Z(X, U_Z)$
The combination of X and U_Z cause Z
$\rightarrow X \not\perp\!\!\!\perp Z$

> Typically, the noise term is omitted (and in the following slide it will not be shown).
> To be more clear: Any change in X will likely to effect changes in Z, but not vice versa (as X and U_Z are independent, $X \perp\!\!\!\perp U_Z$).
> If we would have a dataset that conforms to the causal graph, we would **(faithfully) expect** that X and Z correlation, but purely from the data alone we could not infer a causal relationship (in general).

Causal Graph

Causal Graph - Chain

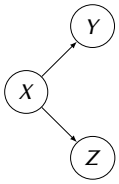


Chain (cascade): X causes Y, Y causes Z
$\rightarrow X \not\perp\!\!\!\perp Z, X \perp\!\!\!\perp Z \mid Y$

> No coincidence this looks like a **Markov chain**.
> In fact, $P(X, Y, Z) = P(X)P(Y \mid X)P(Z \mid Y)$.
> X and Z are not independent, but once we "know" Y, we no longer need X, since all we could learn about Z (from X) is already "contained" in Y.
> Another relationship is the **data processing inequality**, here X might be the raw data, Y is the preprocessed dataset, and Y the processing results.
> The data processing inequality states, that Y cannot "invent" new data that is helpful for Y.
> In data science, from a purely theoretical standpoint, we can only loose information while (pre-)processing the data.

Causal Graph

Causal Graph - Fork

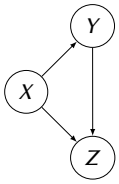


Fork, or common cause: X causes Y and Z
$\rightarrow Y \not\perp\!\!\!\perp Z$ (spurious), $Y \perp\!\!\!\perp Z \mid X$

> One cause, with **multiple effects**.
> With X as the common cause, Y and Z are no longer independent (i.e., we will expect them to correlate), also known as **partial correlation**.
> But conditioned on X, the will be independent (i.e., knowing X will render them independent).
> For example, healthy lifestyle causes exercise and causes healthy diet.
> See also **Reichenbach's common cause principle**, which combines the fork with time:
> "If an improbable coincidence has occurred, there must exist a common cause"

Causal Graph

Causal Graph - Confounder

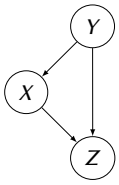


Confounder: X causes Y and Z, and Y causes Z
Assuming Z is the depended variable, X is a confounder for the relationship between Y and Z

> In this case, Y might be the treatment and Z the effect.
> The presence of X modifies the relationship between treatment and effect.
> For example: X are genes, Y is smoking, and Z is lung cancer.
> If we want to compute the influence of smoking on lung cancer, we have to remove the influence of genes.

Causal Graph

Causal Graph - Mediator

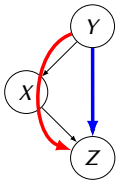


Mediator: Y causes Z directly and via X

> For example: Y is smoking, X is tar in the lungs, and Z is lung cancer.
> For completeness sake, there is also a moderator (similar to the mediator), and mediated moderation, and moderated mediation (can be seen as part of the noise variable, but then the noise is no longer independent from the causal parent).

Causal Graph

Causal Graph - Mediator

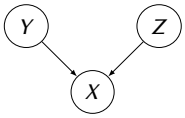


Mediator: Y causes Z directly and indirectly via X

> If we want to compute the influence of smoking on lung cancer, we have to inspect **two causal pathways**.
> In practice often challenging to estimate.

Causal Graph

Causal Graph - Collider

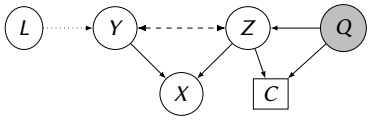


Collider: Y and Z causes X.
$\rightarrow Y \perp\!\!\!\perp Z, Y \not\perp\!\!\!\perp Z \mid X$

> For example: Y is smoking, Z is air pollution by cars, and X is lung cancer.
> Smoking and air pollution are independent, but once we observe lung cancer, they longer longer are.
> i.e., **conditioning** on the collider will cause the causes to correlate.
> Note: If Y and Z were not independent, we need to see an edge between them.
> Same as the example with the good film, good book.
> Collider bias is created via sample selection, stratification, or covariate adjustments.

Causal Graph

Causal Graph - Variations



Some additional notation: Explicit connection for collider causes, unobserved variable via dotted line, conditioned variable via boxes, observed variables via grey nodes

> There is currently no universal agreement on the actual graphical notation.

Causal Graph
d-Separation

About

- The **d-separation** helps to identify
  - All influence factors
  - For example: building a prediction model
    - e.g., We want to predict y, and x is d-separated
    - $\rightarrow$ we do not need to include x in the model
- Its counterpart is the **d-connectedness**

> For the example, the two sets might be (i) the depended variable (we we would like to predict), and (ii) all independent variables (with the goal to find these that actually have an influence).
> Applications: feature selection, parsimonious models.
> See also: http://bayes.cs.ucla.edu/BOOK-2K/d-sep.html

Causal Graph
d-Separation

- Given two nodes (or set of nodes): A, B
  1. Identify all paths between the nodes, ignoring the direction
  2. Identify all nodes on the paths that are conditioned on, add to set C
  3. Use the direction to identify colliders

> All independent variables that are d-separated, can be ignored.

Causal Graph
d-Separation

d-sep(A, B, C) - the covariates b/w A and B will are zero, if C is given

1. A node in C separates, but
2. A collider in C does not, but
3. A collider not in C does separate

> In short, conditioned/observed nodes to separate (recall condition on the "middle" variable in the causal chain) and unconditioned node do not.
> For colliders it is just the opposite.
> e.g., for a regression the coefficients will be zero and the variables will be C (one leave out)

Causal Graph
Causal Graph

Causal graphs and causal discovery

- Given some observational data, can one infer the causal graph?
- ... and only with some (strong) assumptions
- $\rightarrow$ only up to a point, i.e., **Equivalence classes**
  - Different graphs, but cannot be distinguished

> This directly addresses the question, of how the observational data (e.g., conditional probabilities, correlations) and the causal graph are related.
> e.g., If two variables correlate, do we expect them to be connected via an edge in the causal graph?
> e.g., If two variables are connected in the causal graph, do we expect them to correlate in the observations?
> Cannot distinguish chain and forks.

Causal Graph
Causal Graph - Summary

- Causal graph represents the **data generation process**
  - One part of a causal model (e.g., SCM)
- **Intuitive** for humans
- Many cases sufficient to conduct **causal inference**
  - Allows to assess the "causal influences"

> The d-separation guides us, which variables are expected to be independent (assuming we already observe others).

Causal Inference

Given data and a causal model, how do we estimate the causal effects?

Causal Inference
Motivation

- Want to measure effect of **treatment**, T, on the **outcome** Z
- Depending on the **causal structure**[1][2]
  - ... it will be easy
  - ... it will be possible
  - ... it will be impossible

> For example, what is the impact of smoking on lung cancer?
> What is the impact of pressure of the printing machine on the quality of the t-shirts?
> What is the impact of increase the "buy" button on my shopping web site on the purchase behaviour?
> Recall the causal effect:
$P(Z = z \mid do(X = 1)) - P(Z = z \mid do(X = 0))$

[1] Given sufficiently many observations
[2] Given some assumptions

Causal Inference
Trivial Case

- If there is **no connection** and there are **no confounders**
- $P(Z|do(T = t)) = P(Z)$, i.e., there is no effect

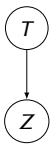
No edge/path between T and Z, no confounders between T and Z

> For example, an intervention on smoking (getting a smoker to quit smoking), is not expected to change the quality of the t-shirt factory.
> While this may sound trivial, the causal graph and the d-sep() guides us to infer, which relations are independent.

Causal Inference

Simple Case

- If there are **no confounders** and **no causal parents**
- $P(Z|do(T = t)) = P(Z|T = t)$, only the direct effect



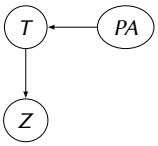
No incoming edges on T, no confounders between T and Z

> Then our observations directly match the effects.
> There might even be some mediators in the path between T and Y.
> Or there might also many other outgoing edges from T, which can all ignore here.
> Note: Variables, which have no incoming edges are also called **exogenous variables**.

Causal Inference

Backdoor Case

- If there are **no confounders**, but **causal parents**
- We **close/block the backdoor** of PA onto T
  - $P(Z|do(T = t)) = \sum_{pa \in PA} P(Z|T = t, PA = pa)P(PA = pa)$



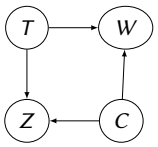
Incoming edges on T from its causal PA, the backdoor

> Since we are interested on the isolated effect of the treatment, but not on the combined influences that go into the treatment, we need to remove (debias) their influence.
> If all peers smoke, influencing the decision on smoking, but one is not interested to learn the influence of the peers on lung cancer.
> Note: Consider a more complex causal graph, where PA also has a cause - we can close the backdoor with any variable on the "backdoor path", (i.e., the grand parents).

Causal Inference

Backdoor Collider Case

- If there is a **backdoor**, which is **collider**
- We do not **close/block the backdoor** of the collider W
  - $P(Z|do(T = t)) = P(Z|T = t)$, since the unconditioned collider blocks



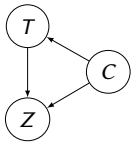
Collider W creates an additional path between T and Z ($T \rightarrow W \leftarrow C \rightarrow Z$)

> If we in this case control for the collider, we would introduce an unwanted bias.

Causal Inference

Confounded Case

- If there are **observed confounders** (not blocked)
- For a single confounder C
  - $P(Z|do(T = t)) = \sum_{c \in C} P(Z|T = t, C = c)P(C = c)$

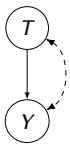


C is a confounder influencing both, the treatment T, and the outcome Z

> This formula is also known as **adjustment formula**.
> We adjust for the bias introduced by the confounder, i.e., we seek to remove its influence.
> Classical example: effect of smoking on lung cancer, with the genes being the confounder.

Causal Inference
Instrumental Variables

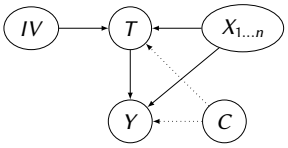
- If there are **unobserved confounders**
- Idea: introduce variation independent from confounders



> See also: https://p-hunermund.com/2018/10/30/you-cant-test-instrument-validity/.
> The IV can be seen as experiment, also called **surrogate experiment** and **surrogate variable**.

Causal Inference
Instrumental Variables (IV)

- An instrumental variable, IV, satisfies
  - $IV \not\perp\!\!\!\perp T \mid \mathbf{X}$
  - $IV \perp\!\!\!\perp Y \mid \mathbf{X}, do(T = t)$
  - with T being the treatment, Y the outcome, and X the features, and C unobserved confounders



> The IV influence the outcome only via the treatment:
> ... no direct path
> ... no unobserved confounders between IV and outcome
> The shown causal graph is just an example, the relation b/w IV and T could also be a chain, or a conditioned collider.
> The second assumption is important, since we use IVs to compute the effect of T on Y.
> Note: There is no way to judge if the assumptions for IV are fulfilled by the data alone, we require domain knowledge (a single unobserved confounder may render our results useless).
> Tools like Dagitty allow to automatically find IVs, https://cran.r-project.org/web/packages/dagitty/vignettes/dagitty4semusers.html

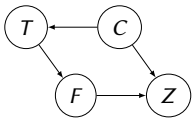
Causal Inference
Instrumental Variables

- Instrumental variables allow to estimate the **local average treatment effect** (LATE) of T and Z
  - Assumption: relationship between IV and T needs to be monotone
  - Specific to the chosen IV
- Estimator needed
  - e.g., Wald estimator for binary treatment and instrument

> Not always clear, if the LATE is representative for the full population.

Causal Inference
Front-Door Case

- If there are **unobserved backdoors**, and **confounders**
- The **font-doors** F blocks all direct paths b/w T and Z
  - $P(Z \mid do(T = t)) = \sum_f P(f \mid t) \sum_{t'} P(z \mid t', f) P(t')$



All backdoor path (of F) are blocked by T, no unblocked path b/w T and F

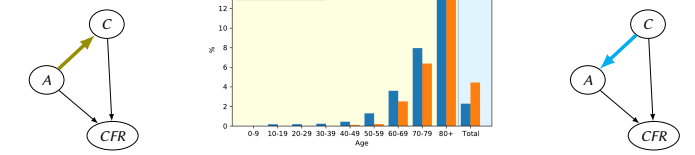
> For example, T is smoking, C are genes (unobservable confounder), Z is lung cancer, and our front-door F would be tar deposits in the lung.
> We assume, that genes do not play a role in tar disposition.
> This is also called the front-door adjustment.

Causal Inference

Simpson's Paradox

Recall Covid'19 case

A - age, *C* - country, *CRF* - case fatality rate



Age (A) is a **confounder**

Age (A) is a **mediator**

We assumed age to be a mediator → *CFR* is higher in Italy
(and the total causal effect (TCE) is the difference in CFRs).

> Here the **intervention** can be seen as change the country from China (no treatment) to Italy (treatment).
> What is the **correct question** if we want to find out: what is the effect of the country (Italy) on CFR?
> Possible questions:
> - What is the average effect of the country? (mediator)
> - What is the age group effect of the country? (confounder)
> In our case we assume age to be the mediator, as Italy causes people to get old → we now know, which is the right question to ask.

Causal Inference

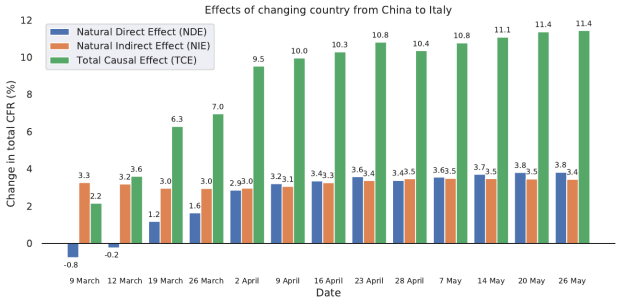
Causal Effects

- Total causal effect (TCE)
  - "What would be the effect on mortality of changing the country from China to Italy?"
- Controlled direct effect (CDE)
  - "For 50–59 year-olds, is it safer to get the disease in China or in Italy?"
  - Controlling for a value of the mediator (i.e., different for each age group)
- Natural direct effect (NDE)
  - "For the Chinese case demographic, would the Italian approach have been better?"
- Natural indirect effect (NIE)
  - "How would the overall CFR in China change if the case demographic had instead been that from Italy, while keeping all else (i.e., the CFR's of each age group) the same?"

> Measuring the causal effect in various ways to learn about the **causal implications**.
> Mediation analysis to split the total causal effect into direct and indirect effect.
> Real world scenarios it is often difficult or even impossible to control both the treatment and the mediator
> Much more, e.g. Sample Average Treatment Effect (SATE), Population Average Treatment Effect (PATE), Population Average Treatment Effect for the Treated (PATT), Conditional Average Treatment Effect, ...
> For mediation analysis also see: https://david-salazar.github.io/2020/08/26/causality-mediation-analysis/

Causal Inference

Causal Effects



> *Evolution of TCE, NDE, and NIE of changing country from China to Italy on total CFR over time. We compare static data from China [27] with different snapshots from Italy reported by [10]. The direct effect initially was negative, meaning that age-specific mortality in Italy was lower; however, it changes sign around mid-March when an overloaded health system in northern Italy was reported [1]. The indirect effect remains mostly constant at a substantial +3–3.5%.*

> von Kügelgen, J., Gresele, L. and Schölkopf, B. (2020) 'Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects', pp. 10–19.

Causal Inference

Simpson's Paradox #2

Example: Kidney Stone Treatment

Stone size vs Treatment	Treatment A	Treatment B	Interpretation
Small stones	93% (81/87)	87% (234/270)	$A > B$
Large stones	73% (192/263)	69% (55/80)	$A > B$
Both	78% (273/350)	83% (289/350)	$A < B$

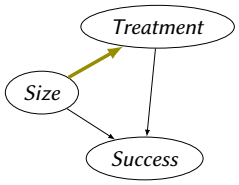
Patients, who suffer from kidney stones receive either treatment A or B, and then the success of the treatment is measured, for multiple patients then a success rate can be computed.

We are interested to know: Which treatment is better?

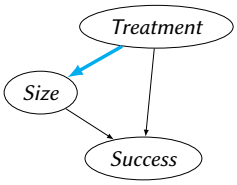
> Example taken from Wikipedia: https://en.wikipedia.org/wiki/Simpson's_paradox
> There are two treatments (A, B) for kidney stones, where the stones have different sizes (small, large).
> The outcome is the success rate of the treatment (in percent).
> When conditioned on the stone size, treatment A appears to be better than B (for both small and large stones), but in total the direction is reversed.
> Note: The numbers in the brackets specify the size of the groups, where we can observe a skewed distribution (while there as many receiving the two treatments, in this case 350 patients for each treatment).

Causal Inference

Simpson's Paradox #2



Stone size is a **confounder**, i.e., $A > B$



Stone size is a **mediator**, i.e., $A < B$

Since the doctors already assume treatment A to be better, they assign more severe cases (i.e., larger stones) to treatment A (and less severe cases to treatment B) → the size of the stone has a causal effect on the treatment (stone size is a confounder), $A > B$.

Causal Inference

Causal Inference - Summary

- We need to study the **causal relationships**
  - Which we assume to be given (and correct)
- We study, which variables are **observed** (conditioned)
- We **select** the approach and **derive** the matching formula
- Finally, apply on the **data**

Causal Discovery

Given data, can we infer the causal model?

Causal Discovery

Assumptions

- **Sufficiency**, there are no hidden confounders
- **Markov assumption**, an event is independent from non-descendants, if conditioned on its parents
- **(Weak) faithfulness**, there might observe a correlation if there is a causal relationships
- **Faithfulness**, there expect to observe a correlation if there is a causal relationships

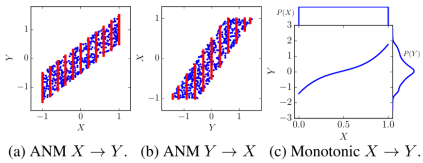
> Crucially, the size has an influence on the outcome as well, in fact in this case we expect that the "influence" of the stone size if bigger than the influence of the treatment alone, known as **Cornfield's conditions**:
> $P(success \mid small\ stone) - P(success \mid large\ stone) > P(success \mid treatment\ B) - P(success \mid treatment\ A)$
> In this case: $0.16 > 0.05$
> Schield, M. and Milo Schield (1999) 'Simpson's paradox and cornfield's conditions', ASA Proceedings of the Section on Statistical Education, 1999, pp. 106–111.
> This is a classical example of **bias in data science** and we often assume that the treatment to be randomised, while in practice it often is not.
> e.g., if the workers/engineers in the t-shirt manufacturing plant already assume a certain machine to provide better/worse results, this may bias the results.
> Practical problem of data science: **how do we find these potential confounders?**
> In many case it may happen that we require binning.
> Is is important that the bins are then sufficiently large, i.e., enough data points/instances available per bin.

> Often, faithfulness is required, because, if due to spurious reasons we do not observe a correlation, even if the data generation process (causality) may indicate so, there is little chance to successfully recover the correct causal structure.

Causal Discovery
Additive Noise Model

- Recall SCM
  - $Y_i = f_i(X_i, U_i)$
- For the **Additive Noise Model** (ANM) we assume
  - $Y_i = f_i(X_i) + U_i$, assuming, $X_i \perp\!\!\!\perp U_i$

Causal Discovery
Additive Noise Model



(a) ANM $X \to Y$. (b) ANM $Y \to X$ (c) Monotonic $X \to Y$.

- We further assume a non-linear function and a "bounded" noise
  - $\to$ expect the noise on Y to be independent from X
  - While, vice-versa, do not expect this behaviour

> Image taken from: Lopez-Paz, D. et al. (2017) 'Discovering causal signals in images', Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-Janua, pp. 58–66. doi: 10.1109/CVPR.2017.14..

Causal Discovery
Additive Noise Model

Application

- Regress Y on X
  - Non-linear, good-fit
- Compute residuals E
- If X and E are not independent, then X causes Y

Causal Discovery
Granger Causality

- Identify causal relationships in multivariate time series
- Intuition
  - If X_i is uniquely helpful to predict future values of X_j,
  - in the presence of other predictive time series X_k (may be multiple)
  - ... then we assume $X_i \to X_j$, i.e., X_i forecasts X_j
- Conditional ignorability assumption is not satisfied
  - $\to$ assumes no hidden confounders
  - e.g., for stock market prediction need to know influence of other stocks, economy, ...

> Extended intuition: ablation study to single out the predictive power of X_i on X_j
> e.g., compare prediction using $P(X_j \mid X_k)$ vs $P(X_j \mid X_k, X_i)$
> Quote: "it cannot be used to discover real causality", as "the values of both treatment variable X and control variable Y maybe driven by a third variable"
> Tsapeli, F., Musolesi, M. and Tino, P. (2017) 'Non-parametric causality detection: An application to social media and financial data', Physica A: Statistical Mechanics and its Applications. Elsevier B.V., 483, pp. 139–155. doi: 10.1016/j.physa.2017.04.101.
> "Transfer entropy is a model-free equivalent of Granger causality."
> Does not detect causality, but mere temporally related phenomena.

Causal Discovery
Granger Causality

- Typically solved via (linear) regression
- Individual coefficients for lagged causality
  - i.e., The causal relationship may manifest itself only after some observations

> The lag is typically assumed to be constant and independent from other factors.
> Another typical assumption is stationarity of the time series.

Conclusions

Practical Aspects and Conclusion

Conclusions
Causal Data Science Process

- Gain an understanding of the **domain**
  - e.g., causal graph, what happens when, five whys, Ishikawa diagram, FMEA
- Gain an understaning of the **data**
  - e.g., correlation anlaysis, visual tools, dimensionality reduction, EDA
- Formulate **questions**/hypothesis
  - Find **answers** (by following the causal pathways)
- **Iterate**! (e.g., refine questions, gather more data)

Conclusions
Causal Data Science Process
Best practice

- **Check**, if the causal relationships (from the domain expert) hold true in the data (faithfulness)
- Collect **constraints** not available as causal relationship
- Keep **held-out data** to validate findings
- If possible, run **controlled experiments** to experimentally validate findings
- Be aware of **assumptions** and their implications

Confounders

- Unobserved confounding nonetheless remains a major obstacle in practice
- → include more data (more variables) into the dataset

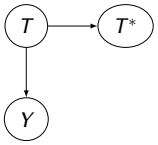
> https://fairmlbook.org/causal.html.

- Building causal graphs is not trivial[3]
  - Important: the causal graph is the **statistical, causal interpretation** of the underlying data generation process
  - When to **merge**/**split** nodes?
  - What confounders realistically **exists**? Which can be simply ignore?
  - Which causal relationships are **transitive**?
  - Mediator vs. moderator?

[3] Even with domain knowledge

> The data scientist should draw the causal graph, not the domain expert.

- **Split** the event/variable into an unobservable and observable
  - Often assumed a measurement equals the treatment



T is the true treatment (not observed), T^* is a measurement (observed)

- Causality is a powerful tool
- Guides the data scientist to ask the **correct questions**
  - … and to **correctly answer** them (e.g., unbiased)
- But, purely data-driven causal inference is not possible
  - … domain knowledge (e.g., via causal graphs) is always needed
  - … if not available, **avoid** jumping to (causal) conclusions

The End

Thank you for your attention!