unicef | for every child

# Implementing Causal Machine Learning in

## Online Workshop 1: An Introduction to Causality, Potential Outcomes, & Identification

August / September / October 2024


**Michael Lechner**
Professor of Econometrics | University of St. Gallen | Switzerland

unicef for every child


Lechner
Swiss Journal of Economics and Statistics (2023) 159:8
https://doi.org/10.1186/s41937-023-00113-y

Swiss Journal of Economics and Statistics

CONFERENCE KEY NOTE Open Access

Causal Machine Learning and its use for public policy

Michael Lechner[1*]

# The workshop series | 1

4 online workshops

- **Today: Correlation & causation**
- September 4, 13-15: Available identification strategies & classical estimation
- September 10, 13-15: Machine Learning
- September 11, 13-14: The data for the Astana workshop & useful descriptive statistics


Lechner
Swiss Journal of Economics and Statistics (2023) 159:8
https://doi.org/10.1186/s41937-023-00113-y

Swiss Journal of Economics and Statistics

CONFERENCE KEY NOTE Open Access

Causal Machine Learning and its use for public policy

Michael Lechner[1*]

# The workshop series | 2

The Astana Workshop September 30 to October 4, 2024 (10-13, 14:30-17:30)

- Monday
  - Morning: Identification with experiments & selection on observables
  - Afternoon: Discussion of potential programmes to be evaluated
- Tuesday: Causal Machine Learning (theory)
- Wednesday
  - Morning: 2 empirical examples
  - Afternoon: The mcf package – how to use it & how to interpret the results
- Thursday: Doing an empirical study in groups with the data introduced in online workshop 4
- Friday: Discussion of programmes to be evaluated continued (core team only)

# Quick introduction

## Participants

- Professional background?
- Knowledge in the estimation of causal effects, machine learning, Python?

## Myself

- Professor of Econometrics at the University of St. Gallen
- Co-head of The Swiss Institute for Empirical Economic Research at the University of St. Gallen
- [Empirical Economic Research | SEW-HSG | University of St.Gallen (unisg.ch)](), *www.michael-lechner.eu*
- Research interest in Causal Machine Learning, AI, programme evaluation, …

unicef for every child

# Plan for today’s workshop

## Correlation & causation

- Correlation does not imply causation: The role of confounders
- Formalisation of causation in a potential outcome framework: Thought experiments
- Definition of causal effects
- The value of the data & the value of assumptions

- *Recommended reading for today*:

**Beyond prediction: Using big data for policy problems**

**Susan Athey**

Machine-learning prediction methods have been extremely productive in applications ranging from medicine to allocating fire and health inspectors in cities. However, there are a number of gaps between making a prediction and making a decision, and underlying assumptions need to be [illegible]e data-driven decision-making.

Athey, *Science* **355**, 483–485 (2017)

unicef for every child

Example from *rainonomics*

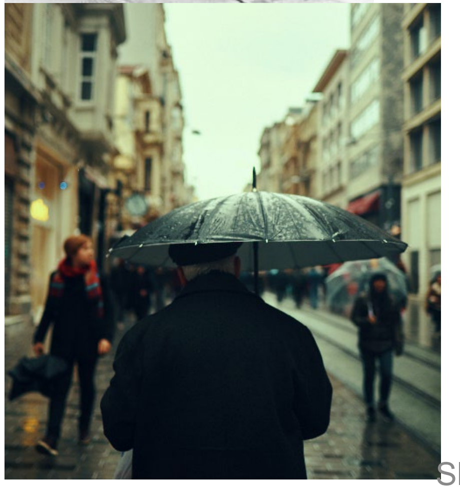
# Predictive vs. causal empirical analysis

Q1: Shall I do a rain dance to increase likelihood of rain?

- Causal problem (estimate effect of rain dance on rain), because performing a rain dance (might) influence whether it rains or not

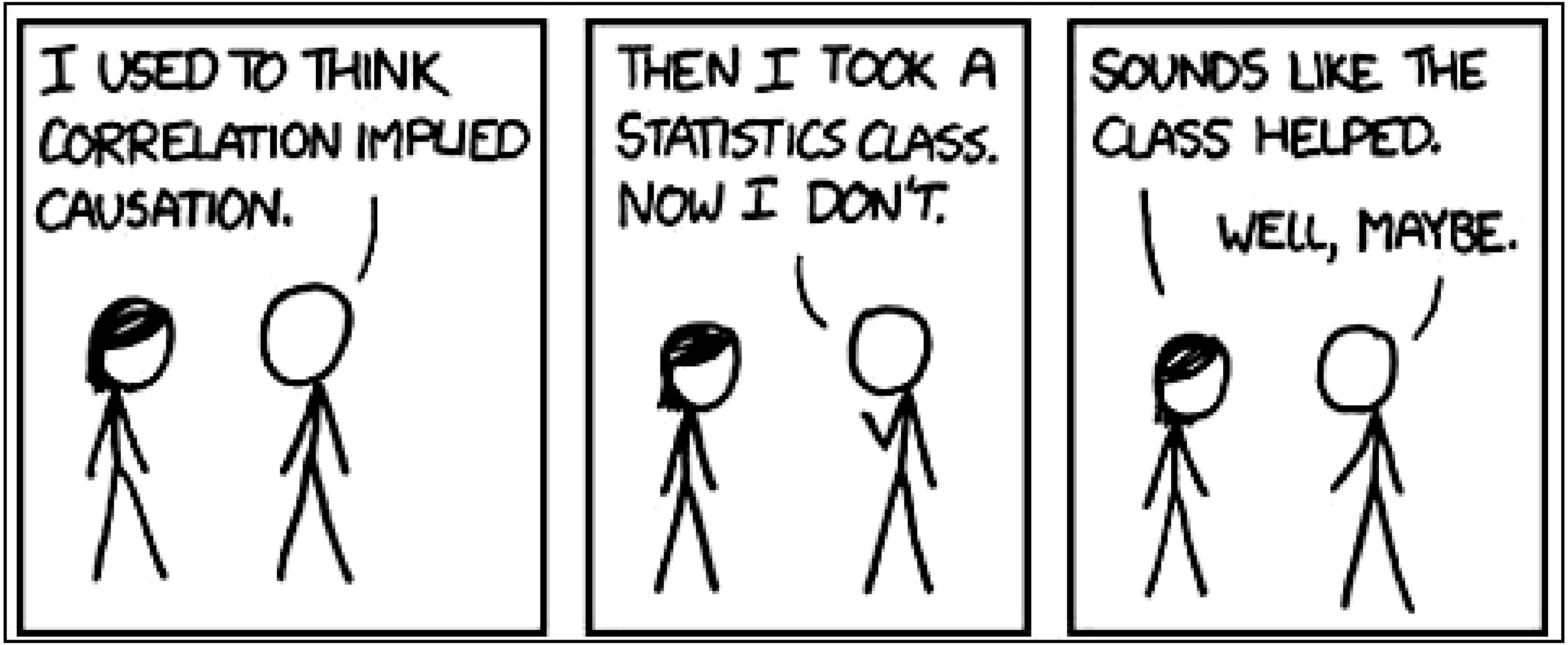
Q2: Shall I use an umbrella when I leave home?

- Predictive problem (estimate likelihood of raining), because taking an umbrella does not influence whether it rains or not, but knowing whether it rains or not is most valuable

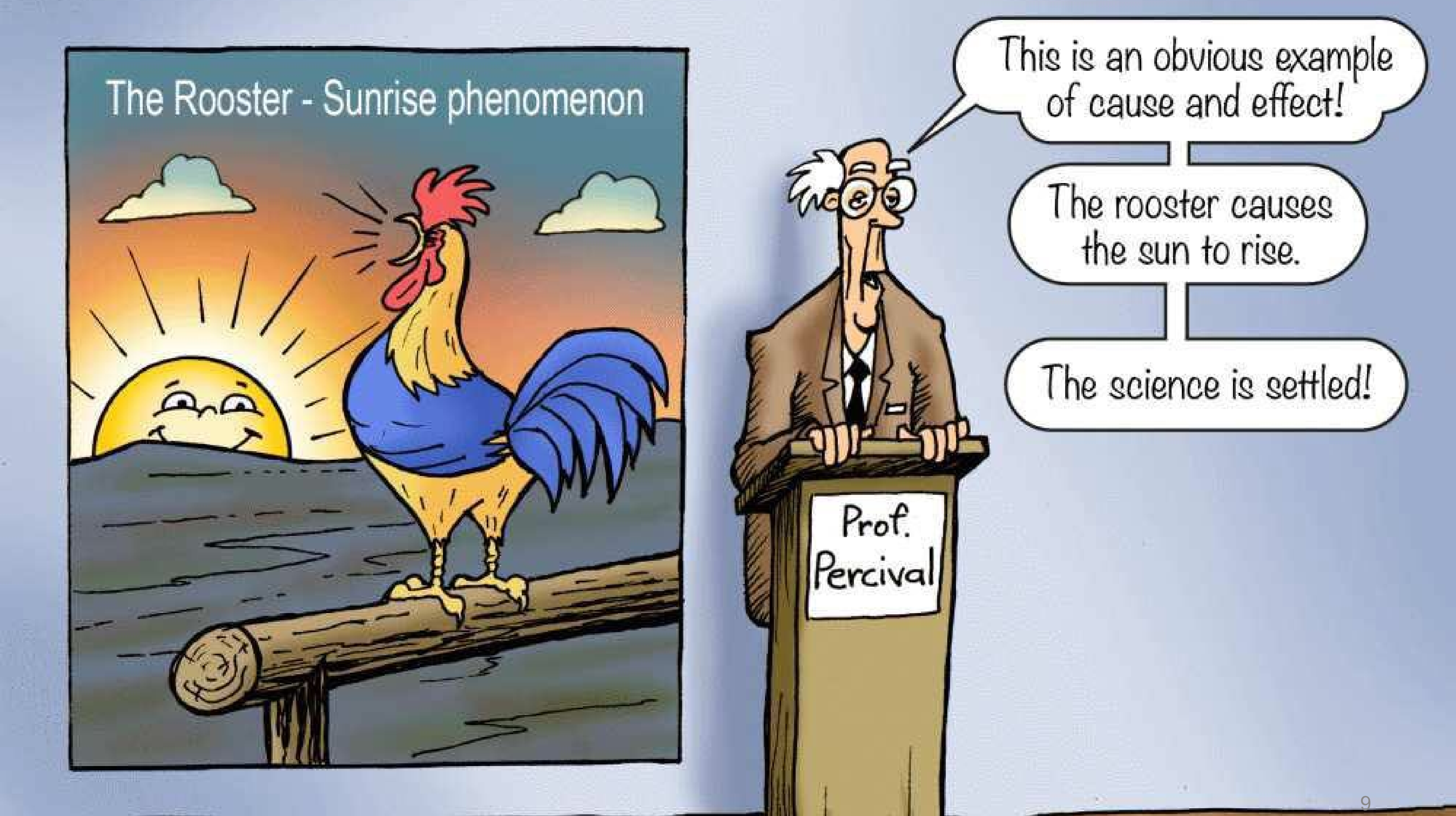


*This example is taken from my colleague Jens Ludwig (Chicago).*

unicef for every child




**Classic - No need to understand this now, ok to understand it after the course 😉**


The Rooster - Sunrise phenomenon
This is an obvious example of cause and effect!
The rooster causes the sun to rise.
The science is settled!
Prof. Percival

unicef for every child


*Maybe, the crew should not turn the seat belt sign on so often, because every time they do, it get's bumpy.*

unicef for every child


Rigby

**Enough with the wind already**

Received April 1

Ever since they installed all those big fans up on the hill it's become even windier. Whose bright idea was that?

I've noticed when they're off, we get a nice calm spell. Please turn them off, at least on weekends. (Word count: 40)

JEFF FORBES
Idaho Falls

ds max • Guest columns, solicited: 450 words max • Guest column

# Correlation does not imply causation

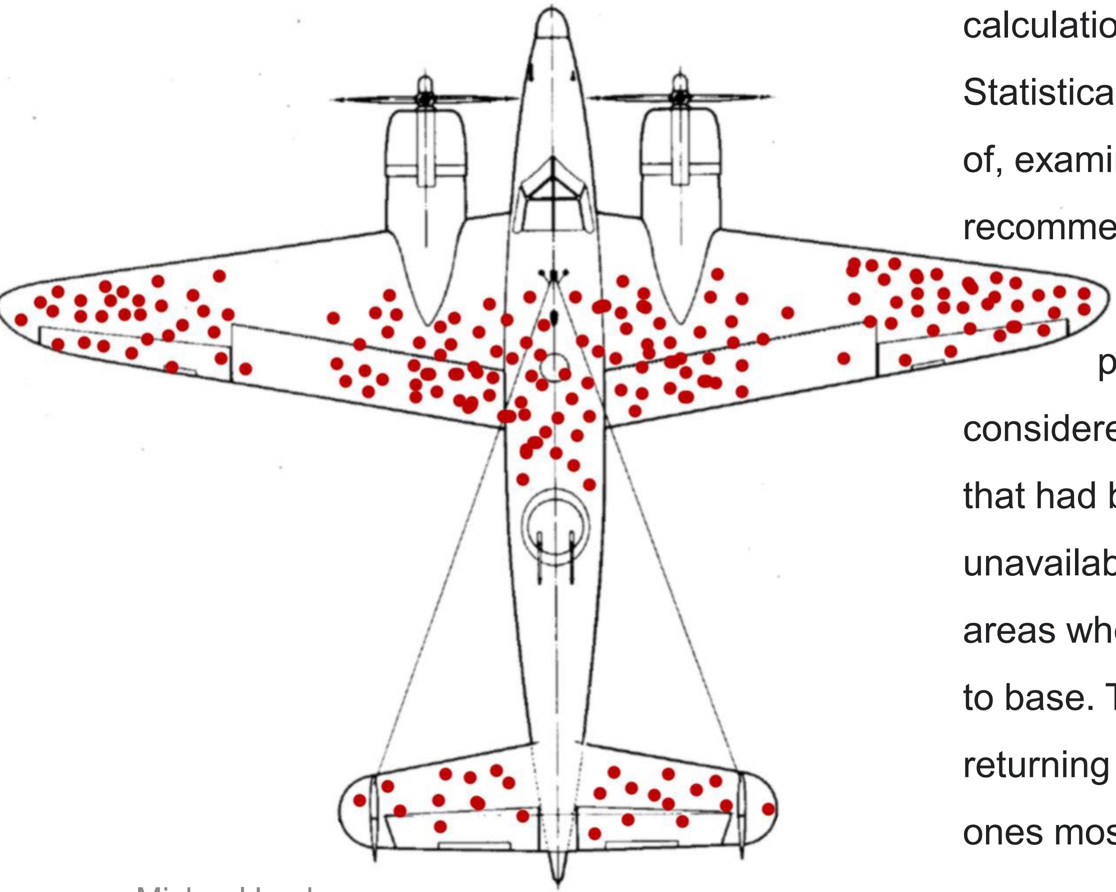
Already clear in these (silly) examples

- To interpret correlational (associative) information causally, **additional non-data knowledge** is needed

Related problems

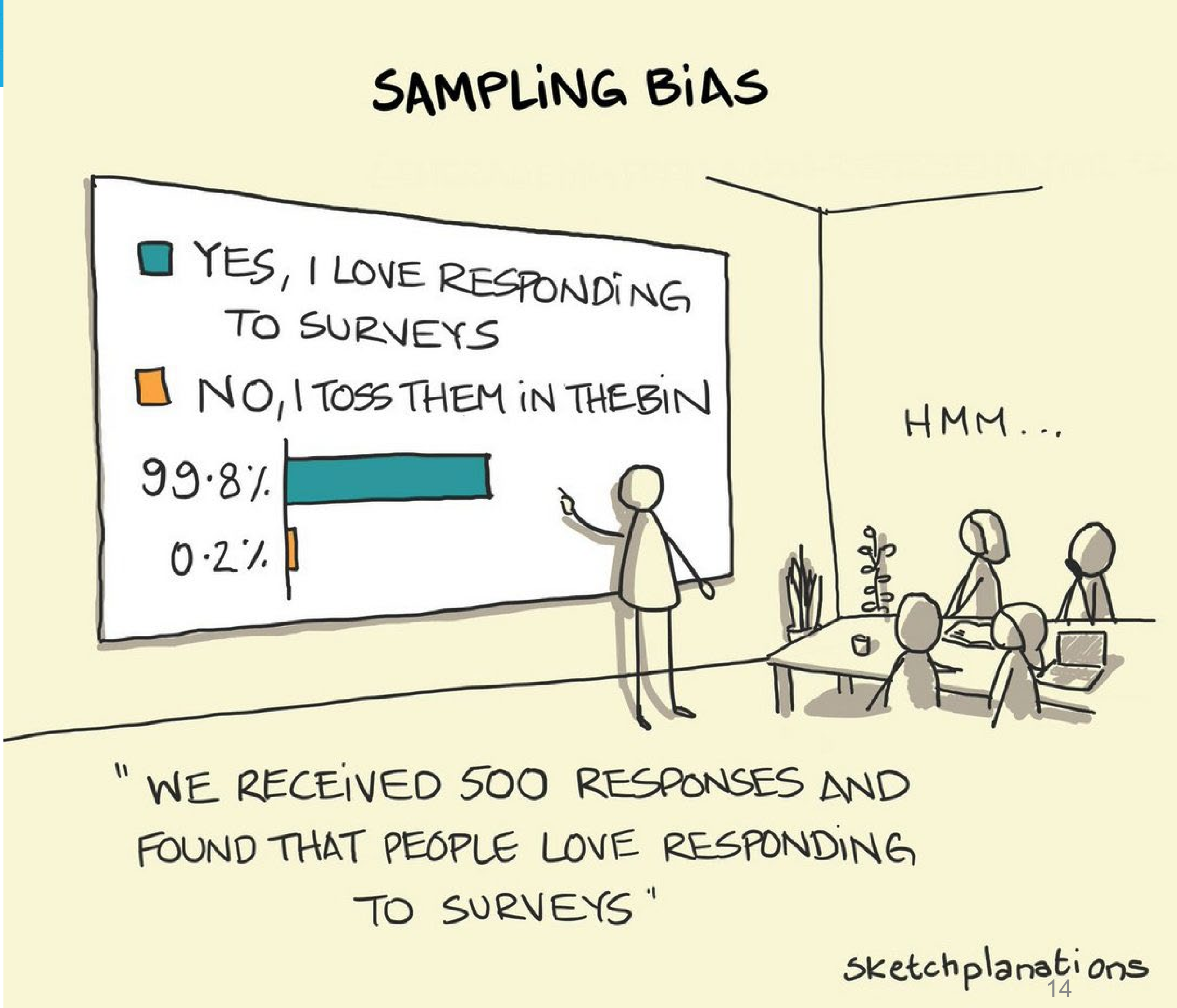
- Sample selection bias
- Non-response bias

# Selection & survival bias | A famous example



During WWII, the statistician Abraham Wald took survivorship bias into his calculations when considering how to minimize bomber losses to enemy fire. The Statistical Research Group (SRG) at Columbia University, which Wald was a part of, examined the damage done to aircraft that had returned from missions and recommended adding armor to the areas that showed the ***least damage***. This contradicted the US military's conclusion that the ***most-hit* areas** of the plane needed additional armor. Wald noted that the military only considered the aircraft that had *survived* their missions – ignoring any bombers that had been shot down or otherwise lost, and thus also been rendered unavailable for assessment. The bullet holes in the returning aircraft represented areas where a bomber could take damage and still fly well enough to return safely to base. Therefore, Wald proposed that the Navy reinforce areas where the returning aircraft were unscathed, inferring that planes hit in those areas were the ones most likely to be lost. *Downloaded from Wikipedia, Jan, 24, 2022.*

Sampling bias - a related problem



unicef for every child

# ***Artificial*** *example* | Simpson's paradox

New Support Programme for Children in Need *HelpKids25* is tested in a trial

- 40 boys participated: 10 got support, 30 got nothing — *P(support=1|boy)=0.25*
- 40 girls participated: 30 got support, 10 got nothing — *P(support=1|girl)=0.75*
- For boys as well as for girls, getting the support of *HelpKids25* is random
- Outcome is measured as *completing high school* — *P(HS completed|boys) > P(HS completed|girls)*

Mean comparisons reveal that *HelpKids25*

- ... increases high school completion rate for girls
- ... increases high school completion rate for boys
- ... decreases high school completion rate overall

**How is this possible? Does *HelpKids25* work, or not?**

unicef | for every child

| Girls | High School completed | High School not completed | Sum | High School completion rate |
|---|---|---|---|---|
| *HelpKids25* | 9 | 21 | **30** | **30%** |
| Nothing | 2 | 8 | **10** | **20%** |
| All girls | 11 | 29 | **40** | 27% |
| **Boys** | | | | |
| *HelpKids25* | 7 | 3 | **10** | **70%** |
| Nothing | 18 | 12 | **30** | **60%** |
| All boys | 25 | 15 | **40** | 63% |
| **Girls & boys** | | | | |
| *HelpKids25* | 16 | 24 | **40** | **40%** |
| Nothing | 20 | 20 | **40** | **50%** |
| All | 36 | 44 | **80** | 45% |

unicef for every child

# Artificial example | Simpson's paradox | Explanation

The bias in the overall mean comparison is due to **confounding** (*selection bias*)

- Boys are less likely to receive support than girls
- BUT: Boys are more likely to complete HS than girls (without support)

**Gender is a *confounder*!**

True population effect is an increase of HS completion by 10%-points

- (Mean effect of boys x population share of boys) + (Mean effect of girls x population share of girls):

  0.1 x 0.5 + 0.1 x 0.5 = 0.1
- *Non-data knowledge* used for this calculation
  - Data is from experiment stratified by gender

Weighted average of effects in unconfouded subpopulations **instead of mean comparison in full population!**

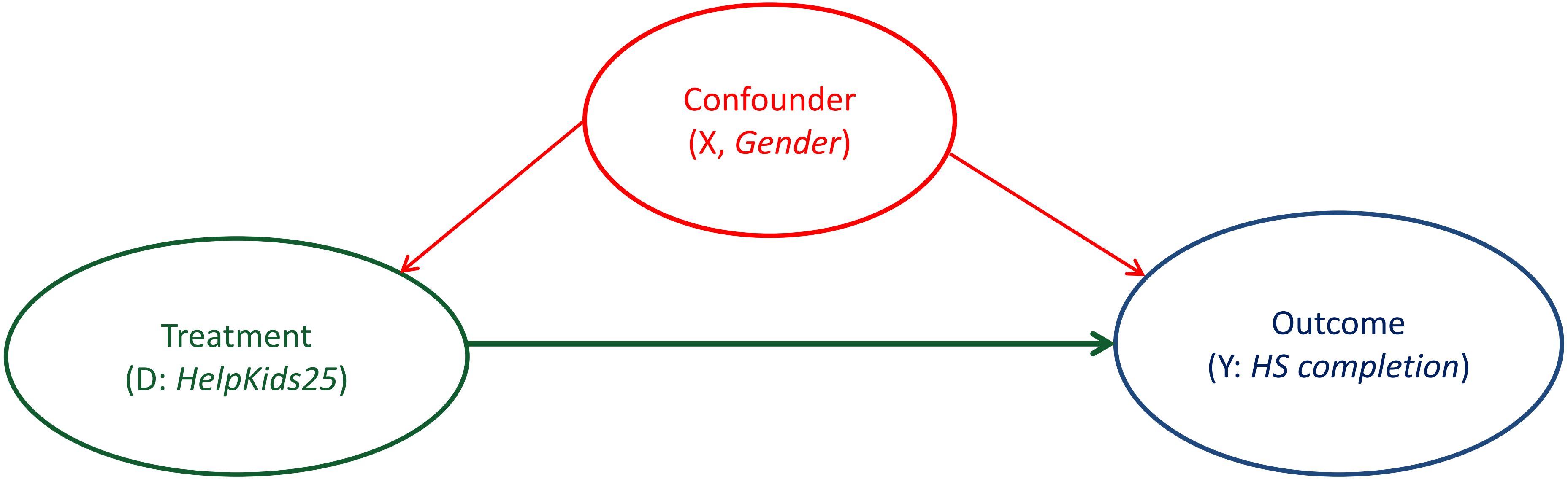
*HelpKids25* **works nicely**

- Mean comparison in full population is misleading, because of *selection bias (ignoring confounder)*

# Confounders

A confounder is variable that *confounds* the comparison of treatment & outcomes ➔ selection bias

- More formal definition will follow



unicef for every child

# Challenge of causal analysis

How to deal with …

- observed
- unobserved

… confounders to adjust associations such that they reveal causal effects?

Conceptional framework: Counterfactual worlds

- What would be the value of the outcome if treated?
- What would be the value of the outcome if not treated?

Formal frameworks

- Potential outcomes
  - Neyman (1923)
  - Rubin (1974): Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701.
- Very similar alternative: DAG (Directed Acyclic Graphs, Pearl, 2000)
  - Pearl, J. (2000), *Causality - Models, Reasoning, and Inference*, Cambridge: Cambridge University Press

*Statistical Science*
1990, Vol. 5, No. 4, 465–480

**On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.**

**Jerzy Splawa-Neyman**

**Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczych Tom X (1923) 1–51 (Annals of Agricultural Sciences)**

unicef for every child

# Notation of the potential outcome model

We need a notation that reflect the *counterfactuals*

| | | | |
|---|---|---|---|
| Treatment: | $D$ | (for simplicity: 0, 1) | observable |
| Observed outcome: | $Y$ | | observable |
| Potential outcome when treated: | $Y(1)$ | *observable if $D=1$*, unobservable if D=0 | |
| Potential outcome when not treated: | $Y(0)$ | unobservable if D=1, *observable if $D=0$* | |

Connection of observable to potential outcomes: $Y = D\,Y(1) + D\,Y(0)$

# Define causal effects with potential outcomes | 1

*Individual* causal effect: Causal effect for a single unit

- $ITE = Y(1) - Y(0)$
- Fundamentally unidentifiable as no unit can be simultaneously be treated & non-treated

*Average* causal effects: Averages of ITEs over groups

- Usual objects of interest in evaluation studies
  - We may find plausible assumptions to identify these effects (e.g., experiment)
- Average treatment effect: $ATE = E(Y(1)-Y(0))$
- Average treatment effect on the treated: $ATET = E(Y(1)-Y(0)|D=1)$

unicef for every child

# Define causal effects with potential outcomes | 1

Average effects can also be identified for finer subgroups with observed features

- Main theme of Astana workshop

Beyond the average: Distributional effects may also of interest

- More difficult to identify & estimate
- Do not play much of a role in the current Causal Machine Learning literature

# Machine Learning (ML) versus Causal Machine Learning (CML)

## ML & CML have different goals / targets

- Supervised, predictive, regression ML should make a good prediction of *Y*
  - The quality of this prediction can be checked as observed values of *Y* are in the data
  - Statistical properties of ML algorithms are therefore 2nd order
- CML should make a good prediction of a causal effect (ATE, …)
  - The quality of this prediction cannot be checked (directly)
    - ATE is a theoretical thought construct (parameter) that is fundamentally unobservable
  - Statistical properties (guarantees) of CML algorithms are therefore crucial

## CML & ML algorithms will be different

1 | Introduction

2 | Correlation does not imply causation

3 | Confounders

4 | Potential outcomes & causal effects

5 | The role of assumptions & the role of data

6 | Conclusions & outlook

# Definition of identification

*A parameter is identified if it can be expressed in terms of random variables for which observations can be sampled*

- This is usually unproblematic for prediction problems, but not for causal problems (because of the counterfactual)

# Information available & information missing

The data contains information on expectations (& marginal distributions) of observed outcomes:

$$E(Y \mid D=1),\ E(Y \mid D=0)$$

We can learn the following expectations of the *potential* outcomes from the data:

$$E\big(Y(1) \mid D=1\big)\,[=E(Y \mid D=1)], \qquad E\big(Y(0) \mid D=0\big)\,[=E(Y \mid D=0)]$$

We cannot learn the counterfactual expectations from the data:

$$E\big(Y(1) \mid D=0\big)=?, \quad E\big(Y(0) \mid D=1\big)=?$$

The data helps us *partly* with the unconditional expectations of the potential outcomes:

$$E(Y(1)),\quad E(Y(0)) \qquad E\big(Y(d)\big)=E\big(Y(d) \mid D=d\big)P(D=d)+E\big(Y(d) \mid D=1-d\big)\big)(1-P(D=d)\big)$$

# Partial identification of causal parameters | Example ATET

$$\boxed{\begin{aligned} ATET &= E(Y(1) \mid D=1) - E(Y(0) \mid D=1) \\ &= \underbrace{E(Y \mid D=1)}_{identified} - \underbrace{E(Y(0) \mid D=1)}_{counterfactual} \end{aligned}}$$

- Similar for other parameters

Data can tell only a part of the causal story

# Related: Selection bias of mean comparison | 1

$$
\begin{aligned}
E(Y \mid D=1) - E(Y \mid D=0) = \\
&= E(Y(1) \mid D=1) - E(Y(0) \mid D=0) \\
&= E(Y(1) \mid D=1) - E(Y(0) \mid D=0) + E(Y(0) \mid D=1) - E(Y(0) \mid D=1) \\
&= \underbrace{E(Y(1) \mid D=1) - E(Y(0) \mid D=1)}_{\text{causal effect (ATET)}} + \underbrace{E(Y(0) \mid D=1) - E(Y(0) \mid D=0)}_{\substack{\text{selection bias (for ATET) due to confounding} \\ \text{average difference of Y(0) between treated and non-treated}}}
\end{aligned}
$$

Additional assumptions will be needed to either estimate or remove selection bias

# Selection bias of mean comparison | Example | 1

An example from sports economics

## Set-up

- In big events, for some disciplines, like long jump, there are qualifying trials in the morning & a main event ('finals') in the evening (with, e.g., the 12 best athletes from the qualifying trial).
- 'Marginal' athletes perform consistently better in the qualifying than in the finals.

## Explanations

- **Explanation 1 (psychology)**: These athletes have no chance for a medal anyway & already achieved their goal by making it to the final ➔ reduced effort.
- **Explanation 2 (statistics):** This is selection bias.

# Selection bias of mean comparison | Example | 2

The selection bias argument

- Suppose there are 12 spots in the final, 30 athletes attempt to qualify.
- Suppose: 8 athletes are much better than the rest. They qualify.
- Suppose: 22 are equally bad. They compete for the 4 remaining spots.
- Suppose: There is some randomness in the individual outcome (luck, how the athlete 'feels' this day, …).
- Therefore, the 4 most lucky athletes in the ‘bad’ group qualify.
- If luck is random & scarce, it is unlikely that all those 4 athletes are lucky again in the finals.
- They are thus likely to do worse in the finals than in the qualification. **No deeper substantive story needed.**

This is related to the *1st Fundamental Law of Causal Econometrics: Never ever select your sample on the basis of the values of the outcome variable.*

# What is the value of data in causal analysis? | An example | 1

Suppose *outcome* is binary (0,1) ➔ expectations of the counterfactuals bounded in [0,1]

$$\text{If } Y \in \{0,1\} \Rightarrow E(Y^d \mid D = 1-d) \in [0,1]$$

What can learn from the data without making further assumptions?

# What is the value of data in causal analysis? | An example

## Bounds for ATET *without any data*

$$ATET \in [\text{worst case } Y(1) - Y(0), \text{ best case } Y(1) - Y(0)] = [\text{worst } Y(1) - \text{best } Y(0), \text{ worst } Y(1) - \text{best } Y(0)]$$

$$ATET \in [0-1, +1-0] = [-1, +1]$$

**Width of ATET = 2**

## Bounds for ATET *with data*

$$E(Y \mid D=d) \quad \text{(can be estimated from the data)}$$

$$ATET = E(Y^1 \mid D=1) - E(Y^0 \mid D=1) = \underbrace{E(Y \mid D=1)}_{\text{identified}} - \underbrace{E(Y^0 \mid D=1)}_{\textit{bounded}} \quad \Rightarrow$$

$$ATET \in \left[ E(Y \mid D=1) - 1, E(Y \mid D=1) \right]$$

**Width = 1**

# Take-aways

Data reduces the uncertainty about the causal effects by half (only?)

- On top of this there will also be estimation error

The other half of the reduction has to come from assumptions

unicef for every child

1 | Introduction

2 | Correlation does not imply causation

3 | Confounders

4 | Potential outcomes & causal effects

5 | The role of assumptions & the role of data

6 | Conclusions & outlook

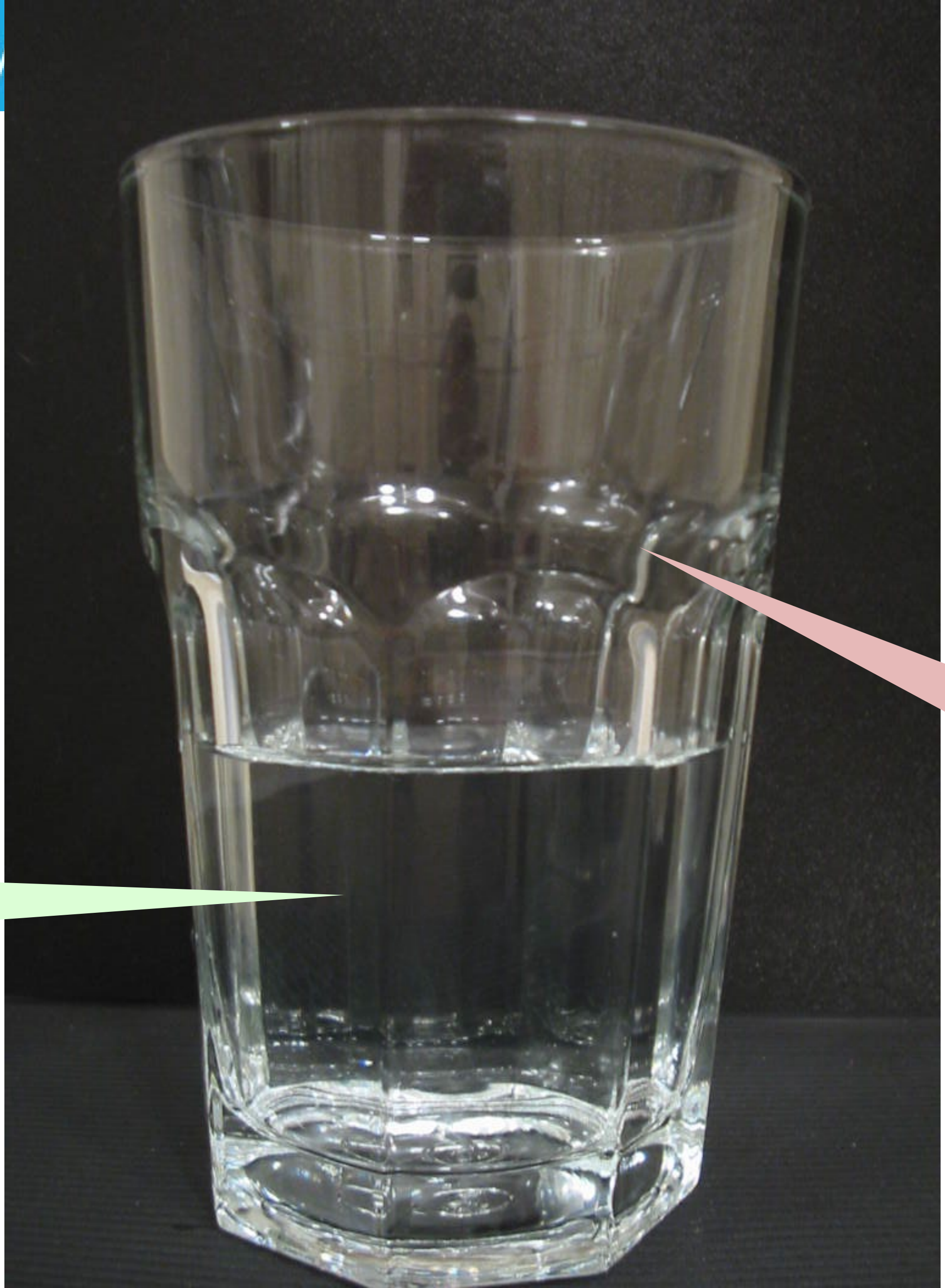
# Summary

Data alone can never answer causal questions

- *Identifying assumptions* are required that cannot be tested by the data
  - A set of connected identifying assumptions defines a *Research Design (RD)*
- *Credibility* of a particular RD is important for the relevance of the empirical results
  - Substantive knowledge of the phenomenon under investigation needed
  - Knowledge of statistics / data science is necessary but not sufficient

Choosing a **convincing research design** is a very important task in any empirical analysis

- It is (very) different from choosing a suitable estimator
- Suitable estimator depends on research design chosen



Research Design

Data

## *Next week:*
## *An overview of different ways to identify causal effects & related estimation strategies*


**Michael Lechner**

Swiss Institute for Empirical Economic Research (SEW)
University of St. Gallen | Switzerland