Modeling data

Introduction and philosophical underpinnings

05.05.2022, Data Science (SpSe 2022): T10

Prof. Dr. Claudius Gräbner-Radkowitsch
Europa-University Flensburg, Department of Pluralist Economics
www.claudius-graebner.com | @ClaudiusGraebner | claudius@claudius-graebner.com

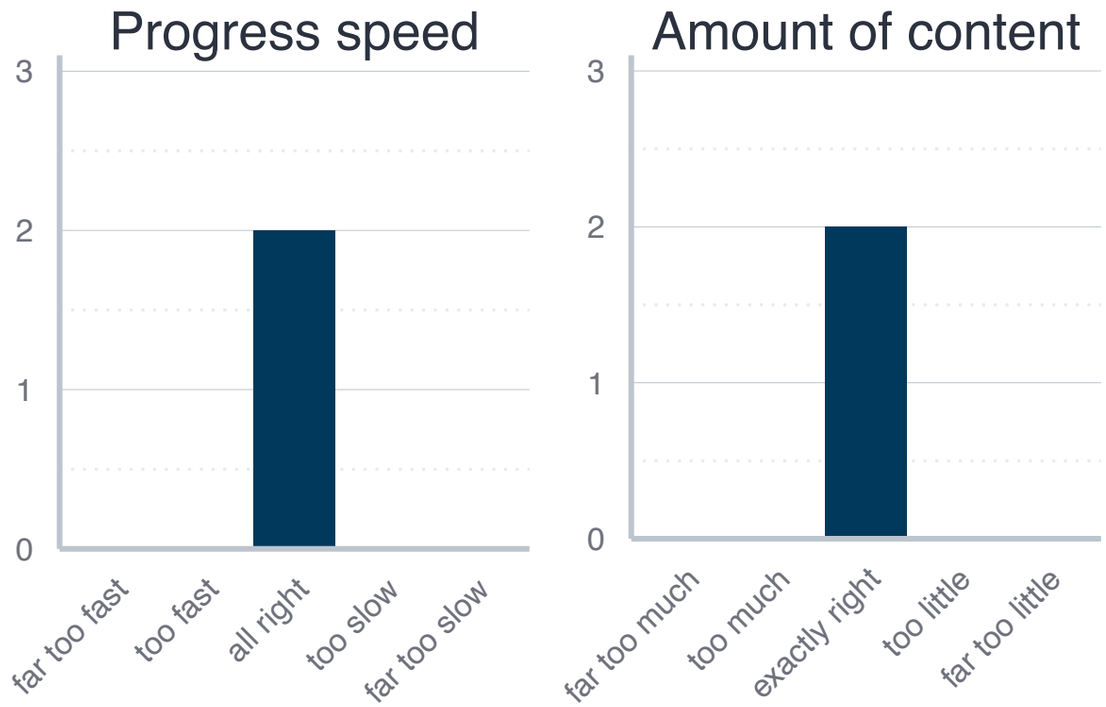
Europa-Universität Flensburg

Europa-Universität Flensburg
International Institute of Management and Economic Education
Department of Pluralist Economics

Prologue:

Prologue
Feedback and exercises

- 2 of you filled out the feedback survey. Main take-aways:
  - You liked the session, despite being a bit more theoretical
- Remark:
  - Good to create and write and knit 3-4 documents to get the necessary practice
  - I put some exercises to start with this on the course page



Goals for today

I. Learn about the different definitions and use cases of models

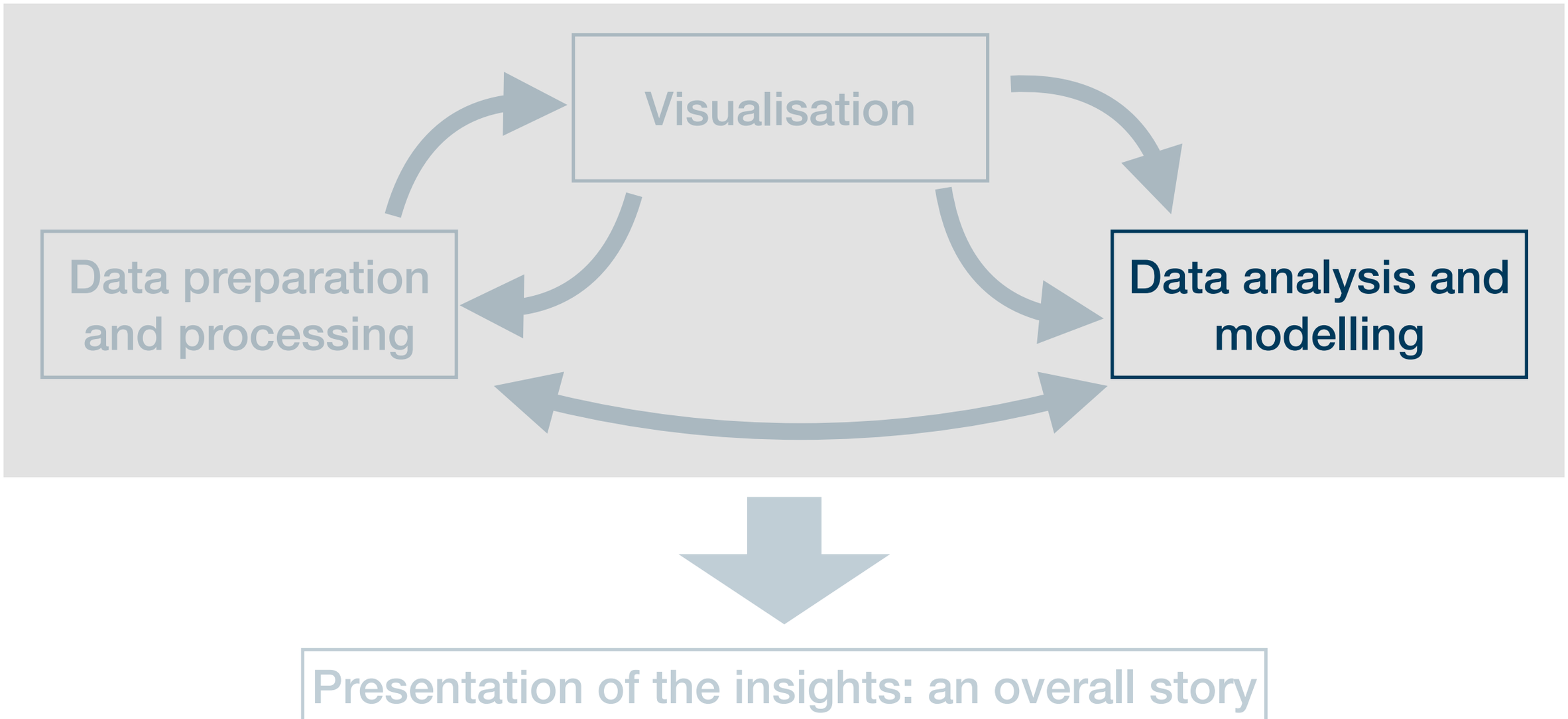
II. Understand relationship of models with the concept of hypotheses and theories of scientific progress

III. Learn about the different steps involved in modelling data

IV. Recapitulate the difference between correlation and causation

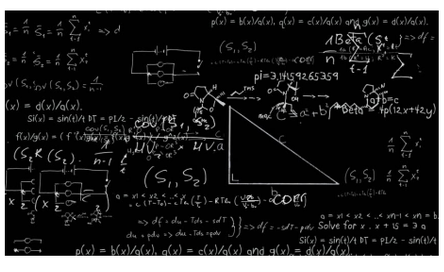
Models - an introduction

The role of modelling in data science



Introduction

- There is no unambiguous and generally accepted definition of a model
  - In any case, models are used to **represent a target**
  - The target is usually the **system under investigation** (SUI)

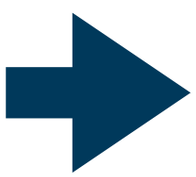


- Sometimes we say that the SUI gets approximated via the data we have collected...
  - ...and sometimes we consider the data itself as the SUI

> "The goal of a model is to provide a simple low-dimensional summary of a dataset."
>
> *R4DS, Chap. 22*

Introduction

- Models can be used for different purposes
  - Most importantly: the distinction between **prediction** and **explanation**
  - **Prediction**: viable statements about future states of the SUI
  - **Explanation**: identification of the mechanisms that have brought about states or dynamics of the SUI
- To explicate how models can generate explanations, we need the concept of a **hypothesis**
  - Hypotheses are also part of many theories of scientific progress



To understand the role of models in science, we need to understand the role of hypotheses for research

Hypotheses in science

The deductive-nomological model

- In the second session we learned about the basic rules of logic

p	Premise 1	*It rains*
$p \rightarrow q$	Premise 2	*If it rains, the street will be wet*
q	Conclusion	*The street is wet*

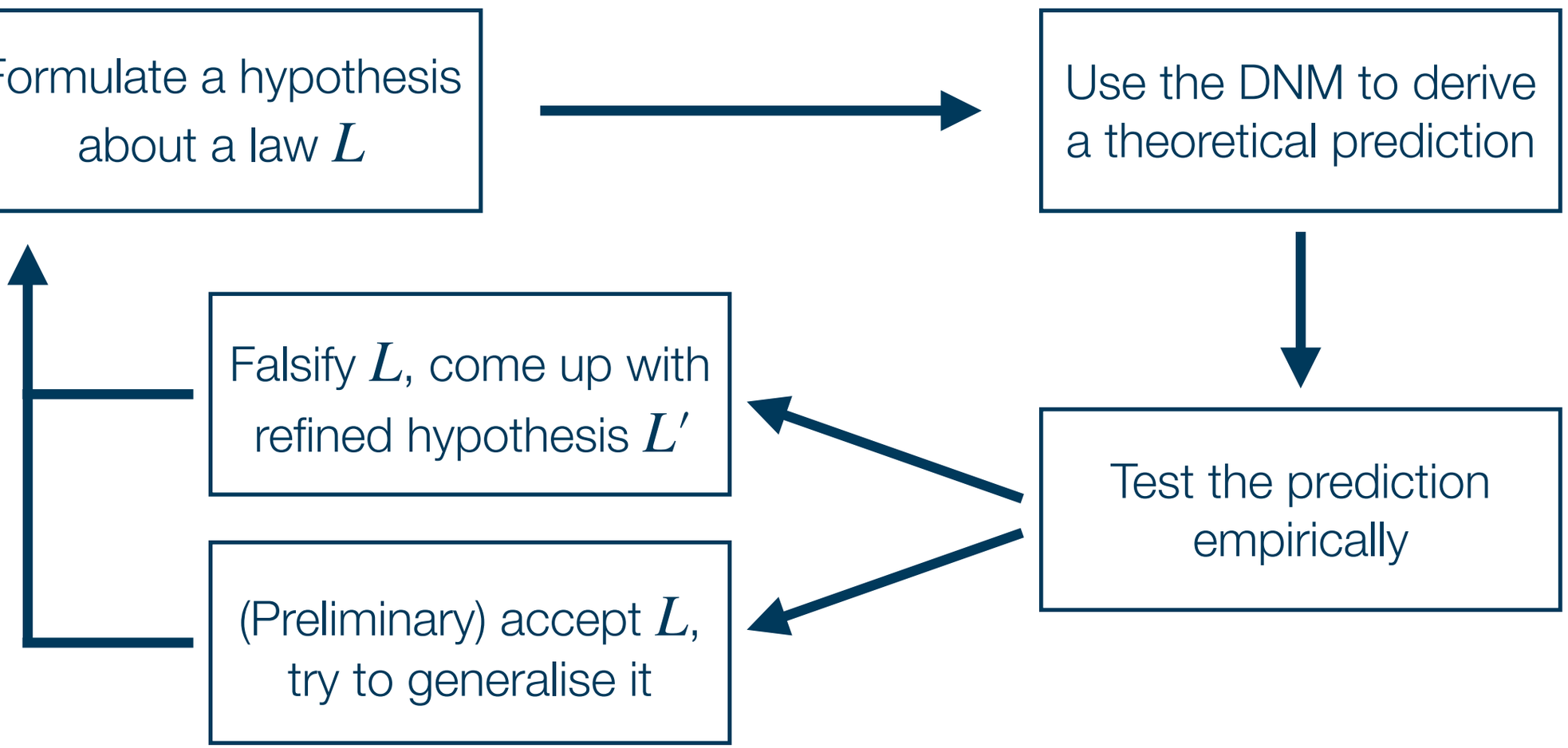
- Adjusting the formal structure a bit yields the famous **deductive-nomological model** of explanation:

p	Condition C	*It rains*
$p \rightarrow q$	Law L	*If it rains, the street will be wet*
q	Phenomenon E	*The street is wet*

- The initial condition together with a law explains the phenomenon of interest
  - C and L together form the **explanandum**, E is called the **explanans**

The hypothetico-deductive method

- The DNM forms the basis for a very idealised model of scientific progress: the **hypothetico-deductive method**



The hypothetico-deductive method
An example

"A hungry mob is a angry mob

Robert "Bob" Nesta Marley: *Them Belly Full*

- Bob Marley proposes a (probabilistic) hypothesis about a law:

For all civil protest movements the following is true:
The higher the degree of deprivation DP, the higher the potential of a violent escalation EP

The hypothetico-deductive method
An example

For all civil protest movements the following is true:
The higher the degree of deprivation DP, the higher the potential of a violent escalation EP

- This could serve as an explanation for high frequency of violent uprisings:

Condition C — *The absolute poverty rate in Jamaica increased to high levels*

Law L — $DP \uparrow \rightarrow EP \uparrow$

Phenomenon E — *The potential for violent uprisings in Jamaica is high*

- But we might not be sure the hypothesis about L is true → test it via HDM!

The hypothetico-deductive method
An example

For all civil protest movements the following is true:
The higher the degree of deprivation DP, the higher the potential of a violent escalation EP

- Using suitable indices we might measure C and E and thereby test L

Condition C	*The absolute poverty rate in Jamaica increased to high levels*
Law L	$DP \uparrow \rightarrow EP \uparrow$
Phenomenon E	*The potential for violent uprisings in Jamaica is high*

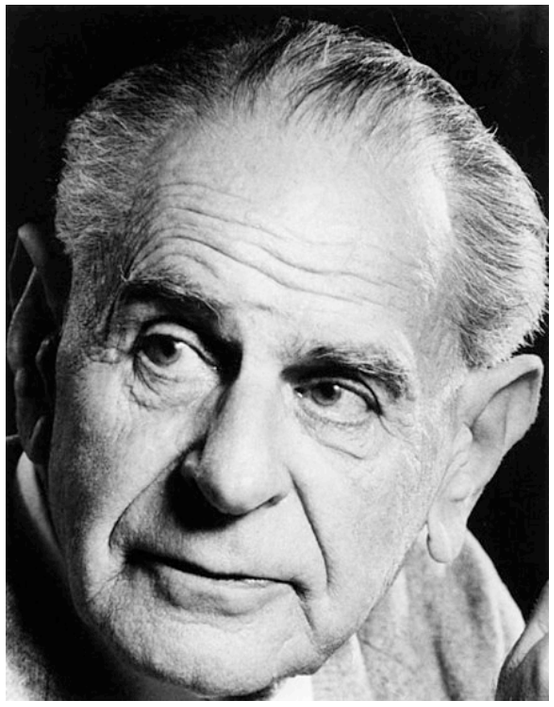
- In this case E becomes a prediction P that we might observe or not
  - If not: reject L, if yes, try to generalise (or break down) L

The hypothetico-deductive method

- The HDM provides a guideline for how to do research
  - Illustrates how we grow knowledge by formulating new hypotheses, and test them
  - **Models**: devices to derive *predictions* and test them
- Ideal: explain as much phenomena as possible with as little laws as possible
  - This was the basis for the philosophical school of **logical positivism**
- There are some fundamental problems with the HDM and logical positivism
  - Theoretical predictions not objectively measurable
  - Every measurement requires us to make additional auxiliary assumptions → isolated hypothesis test is practically impossible (Duhem-Quine problem)
  - We can **never verify a hypothesis** with certainty
- Nevertheless a useful guideline for research, esp. the extension of Karl Popper

Karl Poppers critical rationalism

- Karl Popper was an influential philosopher of science
- He tried to develop a **middle ground** between...
  - **Logical positivism**: knowledge creation through formal logic and empirics
  - **Scepticism/Nihilism/Relativism**: creating reliable knowledge impossible



Karl R. Popper (1902 - 1994)

- Over time he came up with the concept of **critical rationalism**
- This describes an ideal template for scientific inquiry
  - Derive falsifiable hypotheses from theoretical considerations
  - Try to reject the hypotheses using empirical tests

Karl Poppers critical rationalism

- Popper is less strict and optimistic than logical positivists:
  - It is impossible to verify a hypothesis about a law → **Falsificationism**
  - We can never be sure about the truth of a non-falsified law → **Fallibilism**
- At the same time, Popper is not a relativist:
  - We can choose **rationally** among different hypotheses
  - Science is **nothing arbitrary** or fully subjective
- Hypotheses should be tested as critically as possible and should only be accepted preliminary ('not rejected')
  - 'Bolder' hypotheses are better since they contain more information
  - Hypotheses that cannot be falsified theoretically are useless

“

Whenever a theory appears to you as the only possible one, take this as a sign that you have neither understood the theory nor the problem which it was intended to solve

Karl Popper, *Objektive Erkenntnis* (1973)

Discussion

Get together in groups of 2-3 people to conduct the following tasks:

- Summarise the DNM and the HDM and clarify open questions
- What is specific about Karl Poppers critical rationalism?
- What aspects of the critical rationalist approach do you find intuitive and helpful, and where do you see problems or shortcomings?

After 8 minutes, each group presents their results quickly in 1-2 minutes

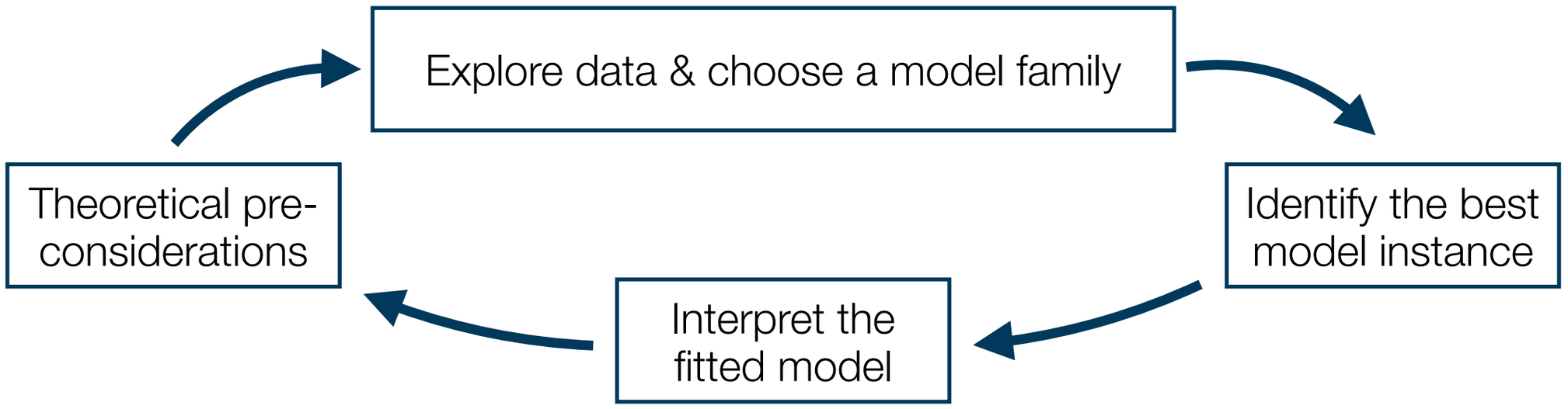
Models and hypotheses

Models and hypotheses

- We distinguish models that help us to come up with new hypotheses...
  - Such models are part of **exploratory data analysis**
  - Never theory-free, but not used to build theory in the strict sense
- ...and those that help use to test existing hypotheses
  - Modeling in this context is part of **explanatory data analysis**
  - This means to follow the idea of Popper's critical rationalism
- Some modelling techniques can be used the one way or the other
  - It is very important that you are explicit with regard to the aim of your model
  - But how do we come up with a useful model?

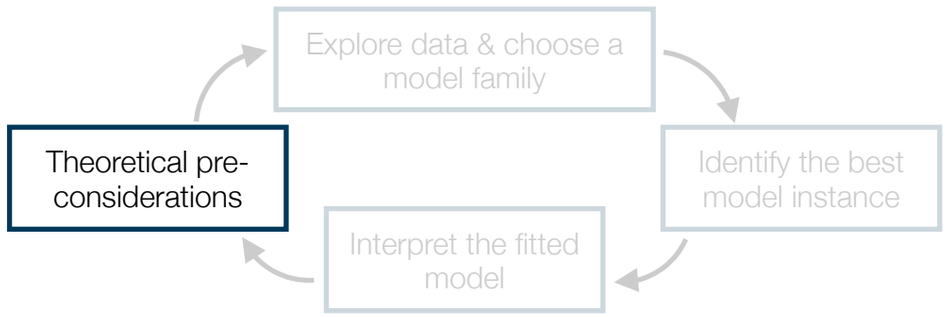
The general sequence of modelling

- In the most general terms, modelling data can be broken down into several steps:



- Applies roughly for both exploratory and explanatory analysis
- Note: sometimes you also choose several models and compare their fit
- Lets illustrate this via a short example

The general sequence of modelling
An example



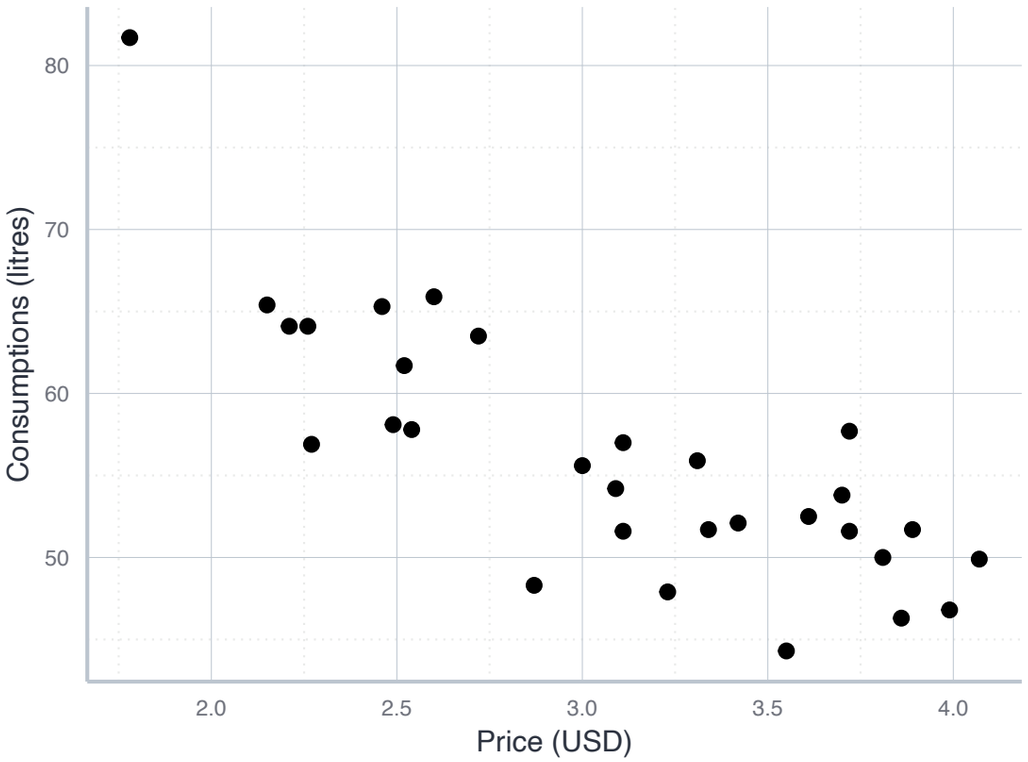
What is the relationship between beer consumption and beer price?

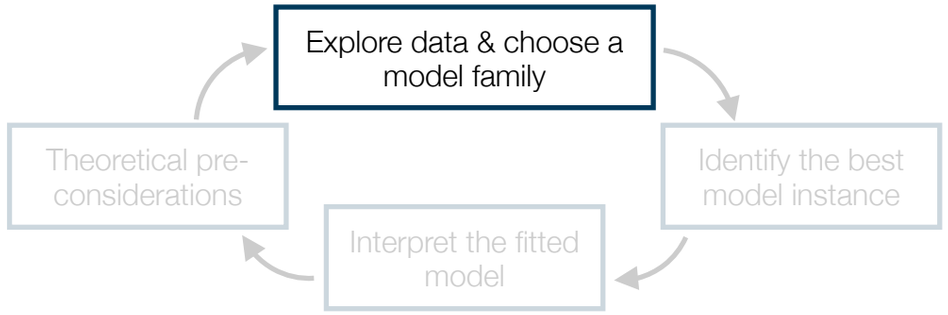
Theoretical law of demand: higher price comes with lower demand

$$D\left(p\right):\frac{\partial D\left(\cdot\right)}{\partial p}<0$$

Obtain survey data on beer consumption and beer prices!

The general sequence of modelling
An example

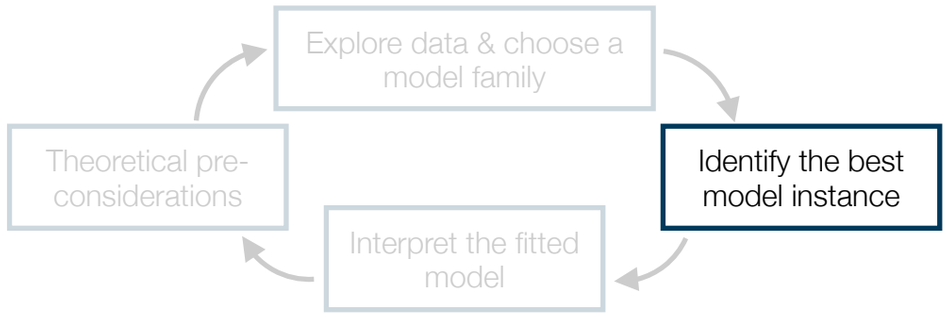




Seems to be a linear relationship → work with the family of linear models:

$$C = a + b \cdot p$$

The general sequence of modelling
An example

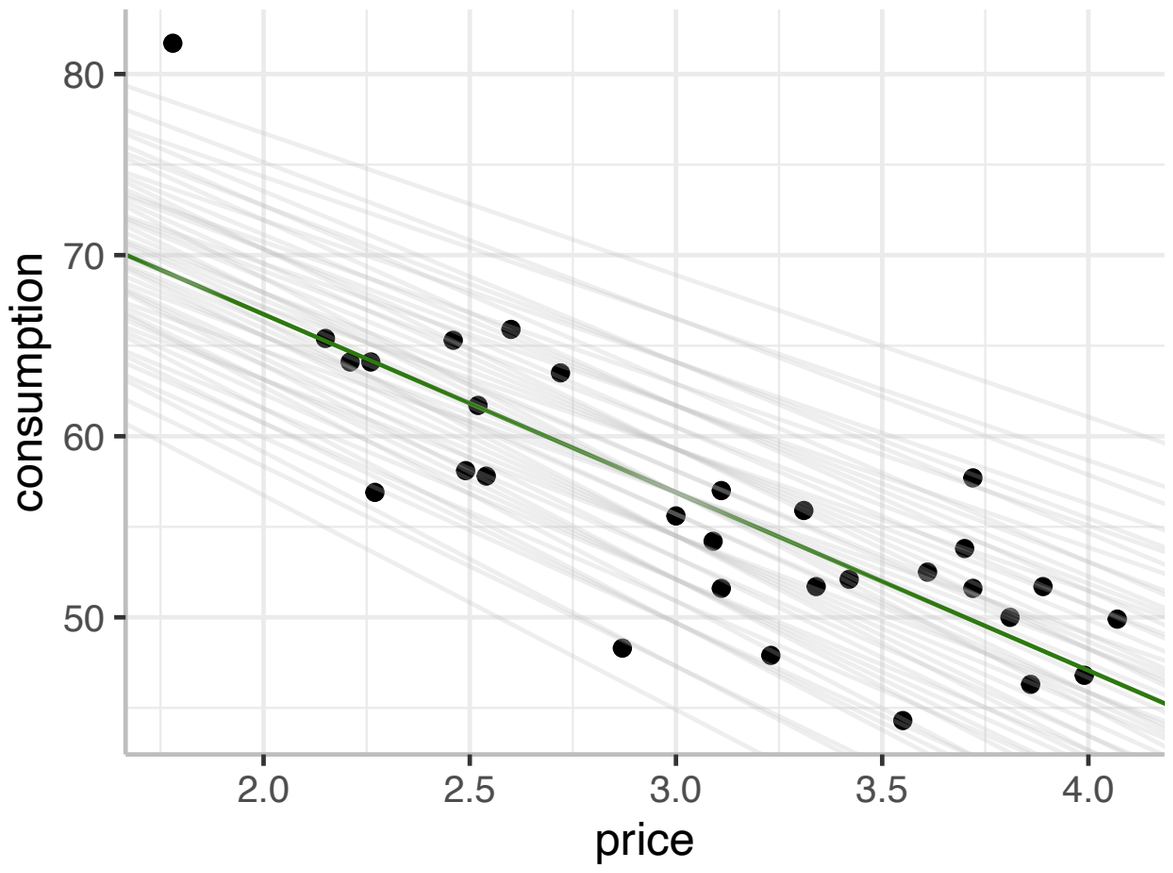


Two parameters:

$$C = a + b \cdot P$$

a and b

Choose parameter such that model describes data best

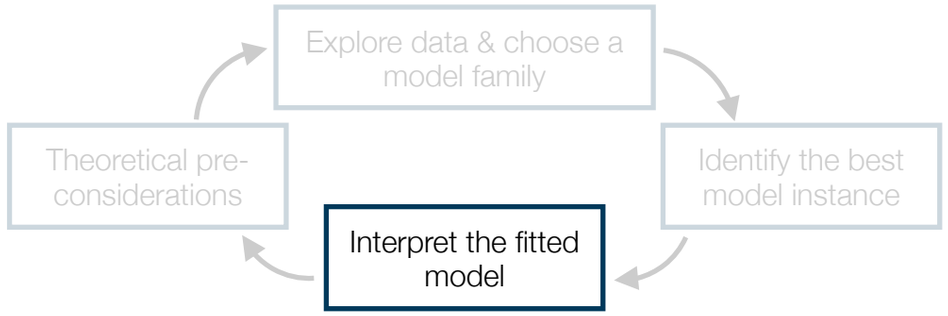


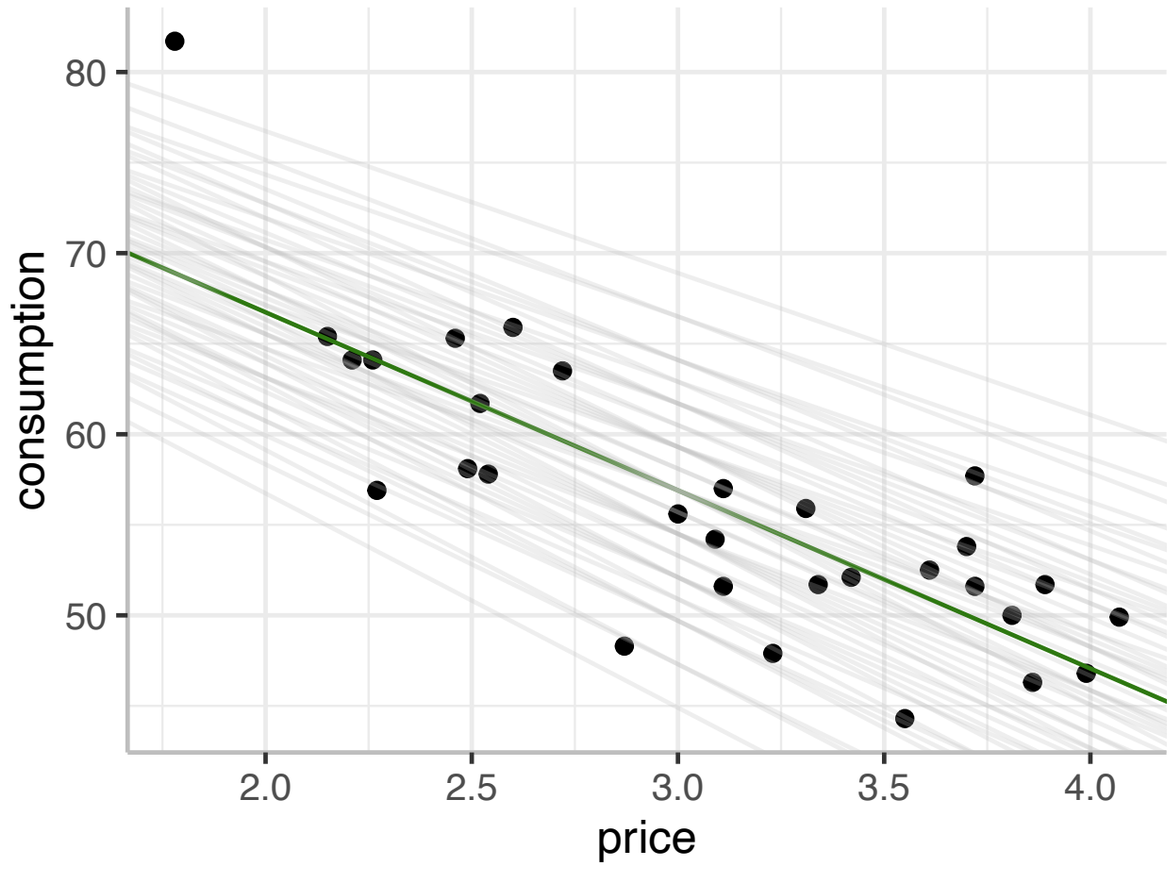

```
Call:
lm(formula = consumption ~ price, data = beer_data_red)

Coefficients:
(Intercept)        price
     86.406       -9.835
```

The general sequence of modelling
An example

```
> linmod_c_price <- lm(
+   formula = consumption~price, data = beer_data_red)
> moderndive::get_regression_table(linmod_c_price)
# A tibble: 2 × 7
  term      estimate
  <chr>        <dbl>
1 intercept    86.4
2 price        -9.84
```

For every increase of 1 unit in `price`, there is an **associated decrease** of, **on average**, 9.84 units of `consumption`.



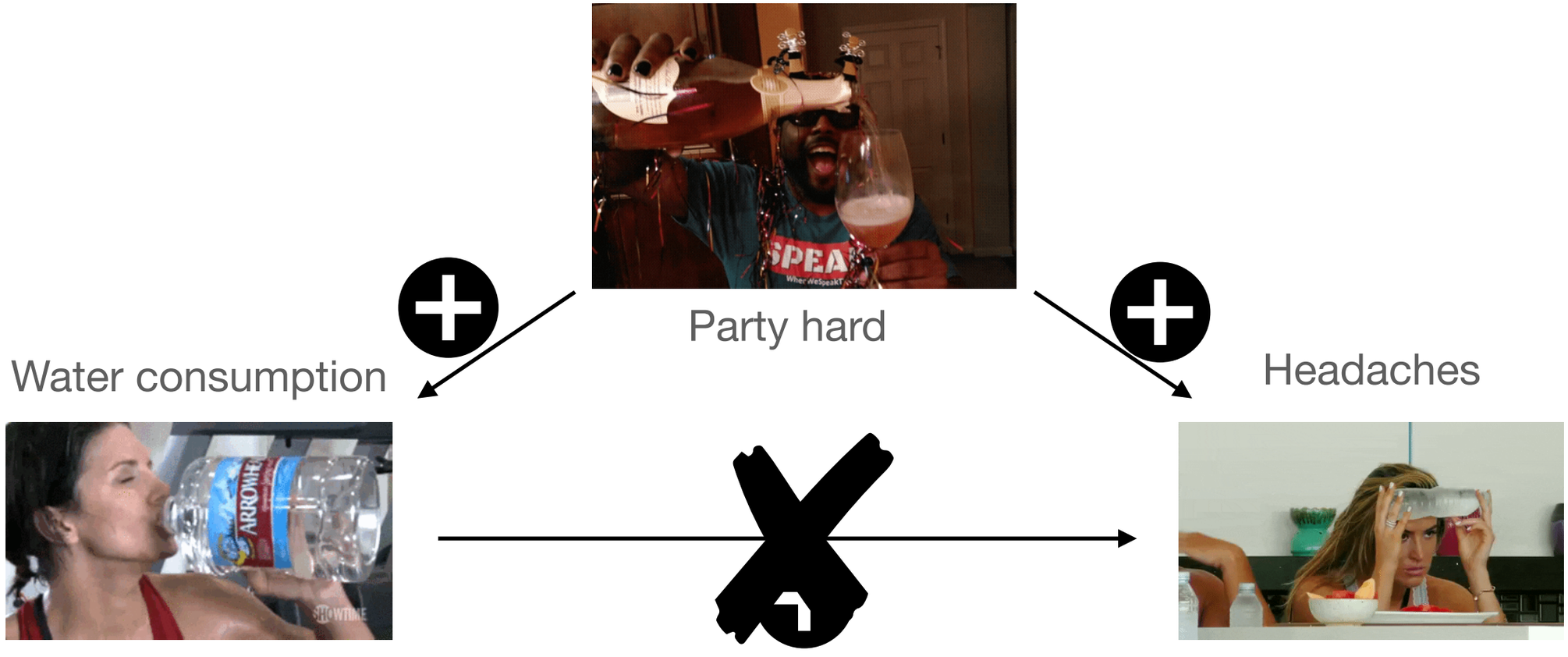


- We expand upon this example in the next session

Correlation & causation

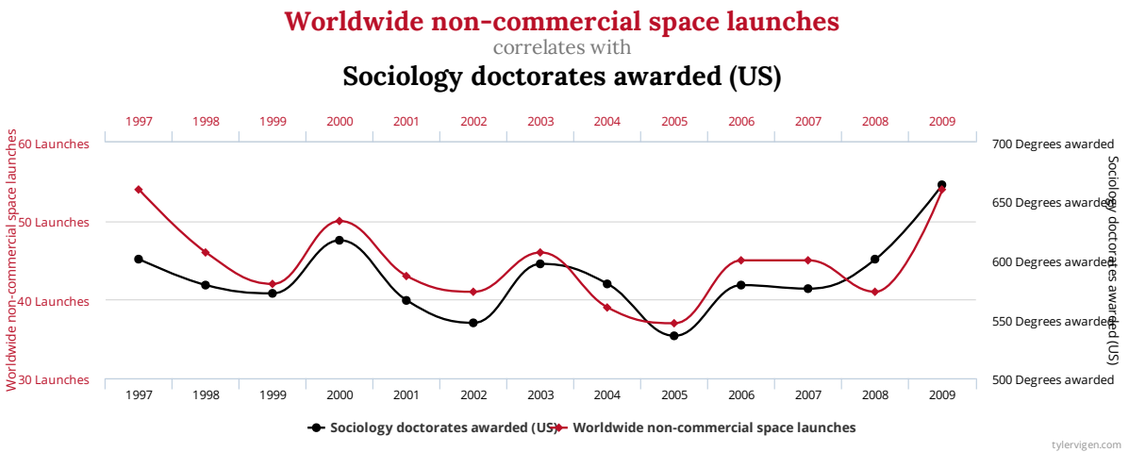
Correlation and causation

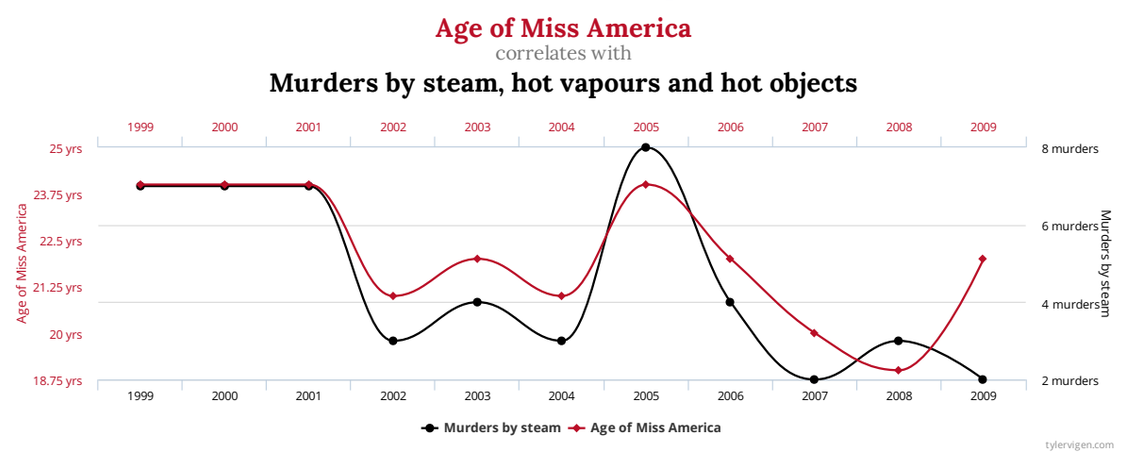
- The distinction between correlation and causation is central for any applied (data) scientist
  - Correlation describes an **observed relationship**
  - Causation refers to an (unobservable) **cause-effect relationship**

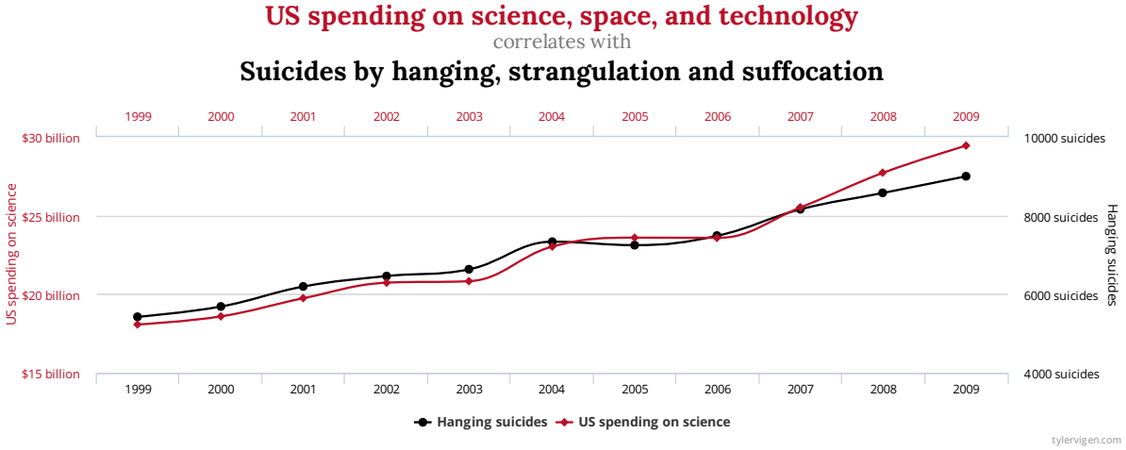


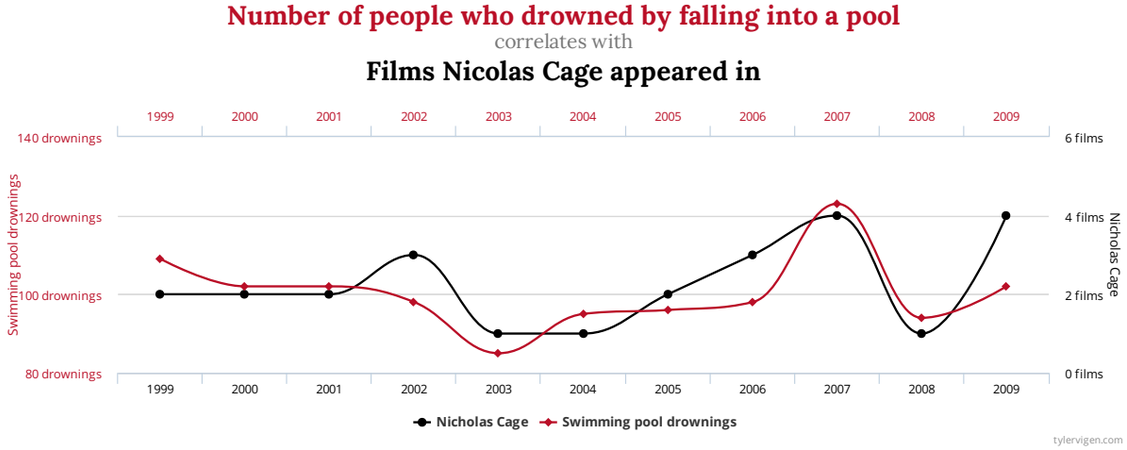
Correlation and causation

- The distinction between correlation and causation is central for any applied (data) scientist
  - Correlation describes an **observed relationship**
  - Causation refers to an (unobservable) **cause-effect relationship**









Correlation and causation

- The distinction between correlation and causation is central for any applied (data) scientist
  - Correlation describes an **observed relationship**
  - Causation refers to an (unobservable) **cause-effect relationship**

- If we observe correlation without causation as in the example we speak of a **spurious relationship** and (potentially) a **confounding variable**

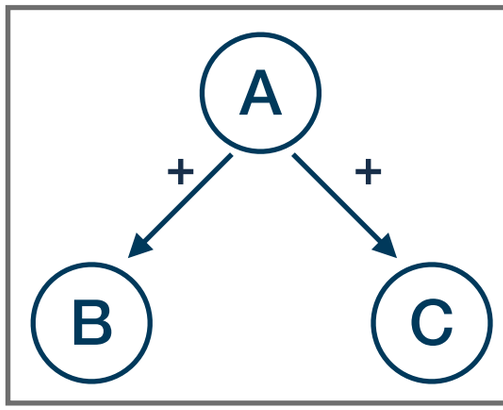
- Knowledge about causality is important whenever we think about the effect of interventions
  - Here we need knowledge that goes **beyond our ability to predict**
  - We might be able to predict suicides by hanging or strangulation via US spending on aircraft, but cannot think about how to reduce them like this...

Correlation and causation

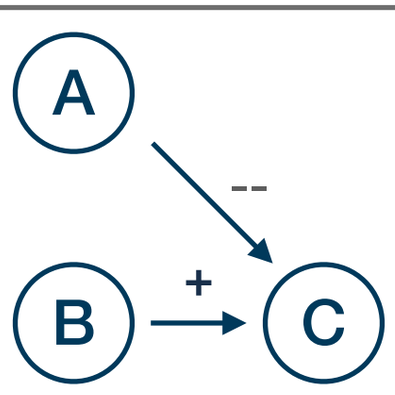
- Identifying causation is attractive but very hard
- It requires us to add theoretical hypotheses about cause-effect-relationships into a model
  - "No causes in, no causes out!"
  - This gives rise to **causal models** (which are often represented graphically)

Nancy Cartwright

- We do not engage in causal modelling, but note that event simply directed cycling graphs (DAGs) help you to sort your thoughts about causation



A: Party all night
B: Water consumption next morning
C: Headache next morning
→ Spurious relation betw. B and C

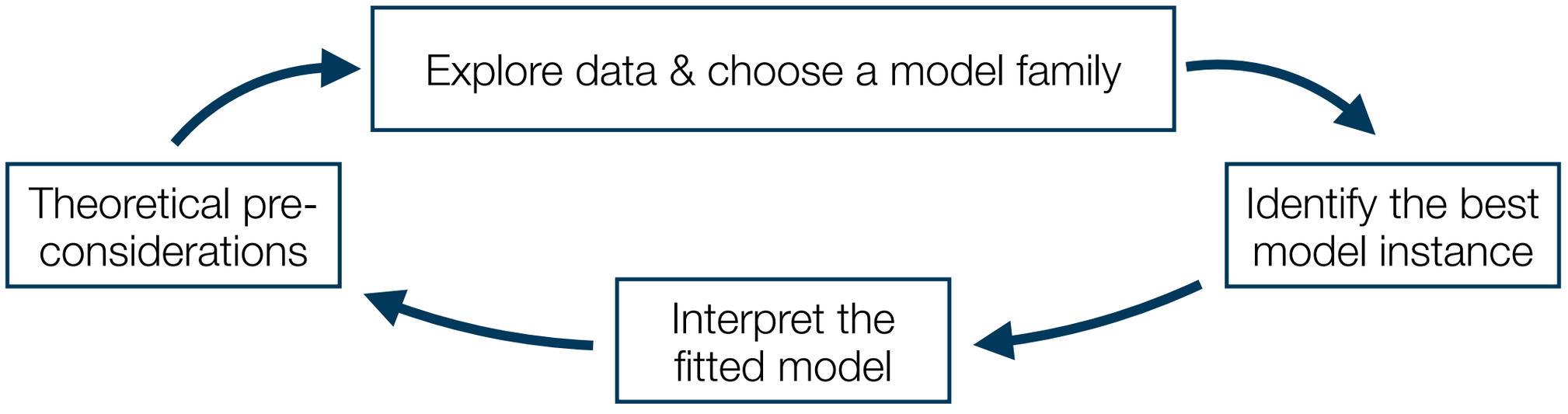


This is how there can be causation without observed correlation

Summary & outlook

Summary and outlook

- We learned what it means to analyse data using models
  - We clarified the role of **hypotheses** and theoretical pre-considerations by introducing Karl Popper's **critical rationalism** as a guiding meta-theory
  - We discussed model purposes, especially **explanation** vs. **prediction**
  - Correspondingly, models can be part of **explanatory** or **exploratory** data analysis
  - We also clarified the important distinction between **correlation** and **causation**
- We introduced the general **workflow** of empirical modelling:



Summary and outlook

- Next session we will operationalise this process using a concrete modelling technique: **simple linear regression**
  - Here we choose the family of linear models with one variable
  - The we identify the family member that best describes the data
  - First and fundamental example for supervised machine learning

Tasks until next week:

1. Fill in the **quick feedback survey** on Moodle
2. Read the **tutorials** posted on the course page
3. Do the **exercises** provided on the course page and **discuss problems** and difficulties via the Moodle forum