

# Introduction to data analysis

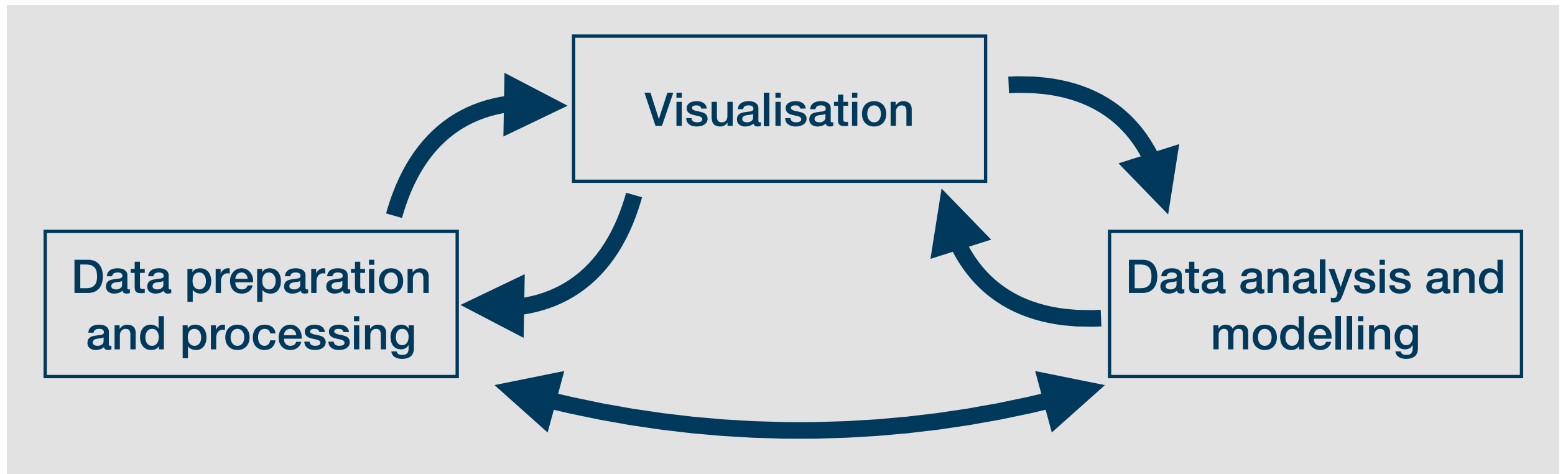
Applied data science with R

**Prof. Dr. Claudius Gräbner-Radkowsch**

**Europa-University Flensburg, Department of Pluralist Economics**

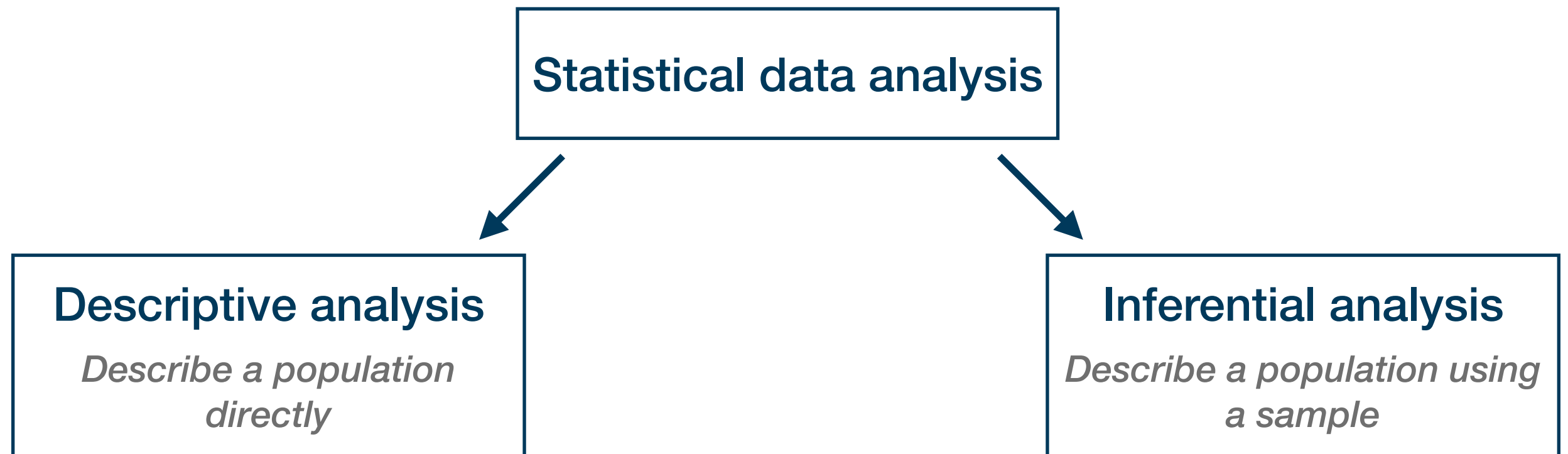
[www.claudius-graebner.com](http://www.claudius-graebner.com) | [@ClaudiusGraebner](https://twitter.com/ClaudiusGraebner) | [claudius@claudius-graebner.com](mailto:claudius@claudius-graebner.com)

# Learning goals



1. Get an overview about different approaches to data analysis and modelling
2. Recap the difference between correlation and causation
3. Familiarise yourself with common notations

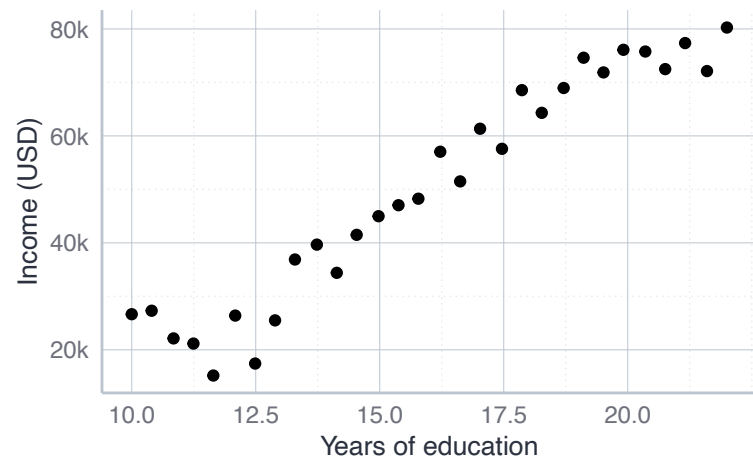
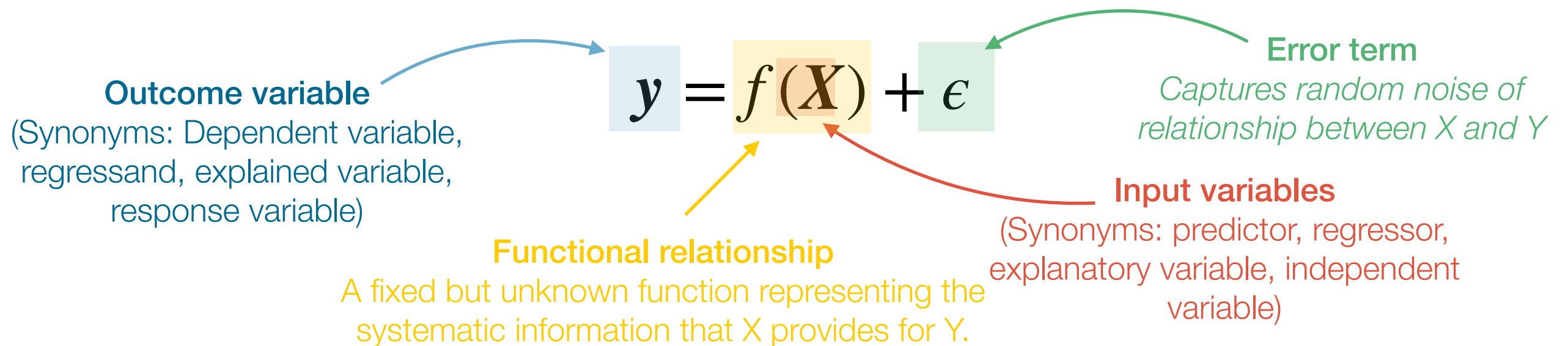
# Data analysis: an overview



- Inferential analysis built upon central insights from **sampling theory**
  - How to infer features of a population from the features of a sample?
  - See separate session with more details

# Kinds of data analysis: notation

- Inferential analysis is **model-based**:



$y$  : Yearly income (in USD)

$X$  : Years of education

$f(\cdot)$ ,  $\epsilon$  : not observable

➡ Goal: learn about (or ‘estimate’)  $f(\cdot)$  and thereby obtain  $\hat{f}(\cdot)$

# Detour: how $X$ and $Y$ look like in practice...

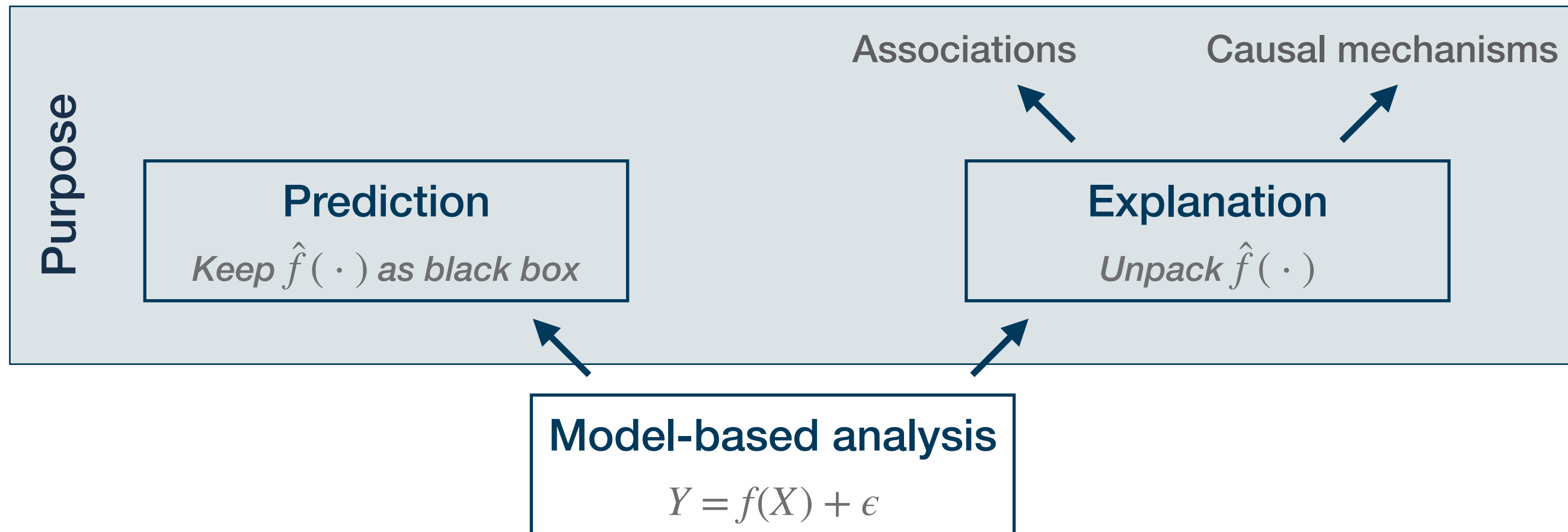
- In equations,  $X$  and  $Y$  are often considered vectors or matrices:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

- In R we can use matrices, but more conveniently operate with **tibbles**:

```
# A tibble: 5 × 4
      Y      X1 X2      X3
  <dbl> <dbl> <chr> <dbl>
1  0.713  0.594 Group A  0.0347
2 -0.731 -2.36  Group A  1.28
3  0.945 -1.42  Group B  2.92
4 -0.352  0.516 Group B -0.342
5  1.07   0.694 Group B  0.562
```

# Types of inferential data analysis I



# Types of inferential data analysis I

## Examples for predictive and explanatory approaches

- Common prediction case:
  - For some time/space we have data for both  $\mathbf{X}$  and the associated  $\mathbf{y}$
  - For a different time/space we have only data on  $\mathbf{X}$
  - We want to know what values for  $\mathbf{y}$  we can expect
- Procedure: assume relationship  $\mathbf{y} = f(\mathbf{X}) + \epsilon$  and estimate  $\hat{f}(\mathbf{X})$ 
  - Then obtain fitted/predicted values  $\hat{\mathbf{y}} = \hat{f}(\mathbf{X})$

A company has information about how many sales occurred when a certain amount of money was spent on advertising. It wants to know how many sales are to be expected when it doubles its expenses for advertising.

# Types of inferential data analysis I

## Examples for predictive and explanatory approaches

- Common explanation case:
  - We have data for both  $X$  and the associated  $y$
  - We want to know the strength and direction of the relative association of different variables in  $X$  and  $y \rightarrow$  focus on **associations**

The company wishes to understand how the association between advertising expenses and sales differs for different kinds of advertising.

- We want to know whether there is a causal mechanism connecting  $X$  and  $y$  focus on **mechanisms**

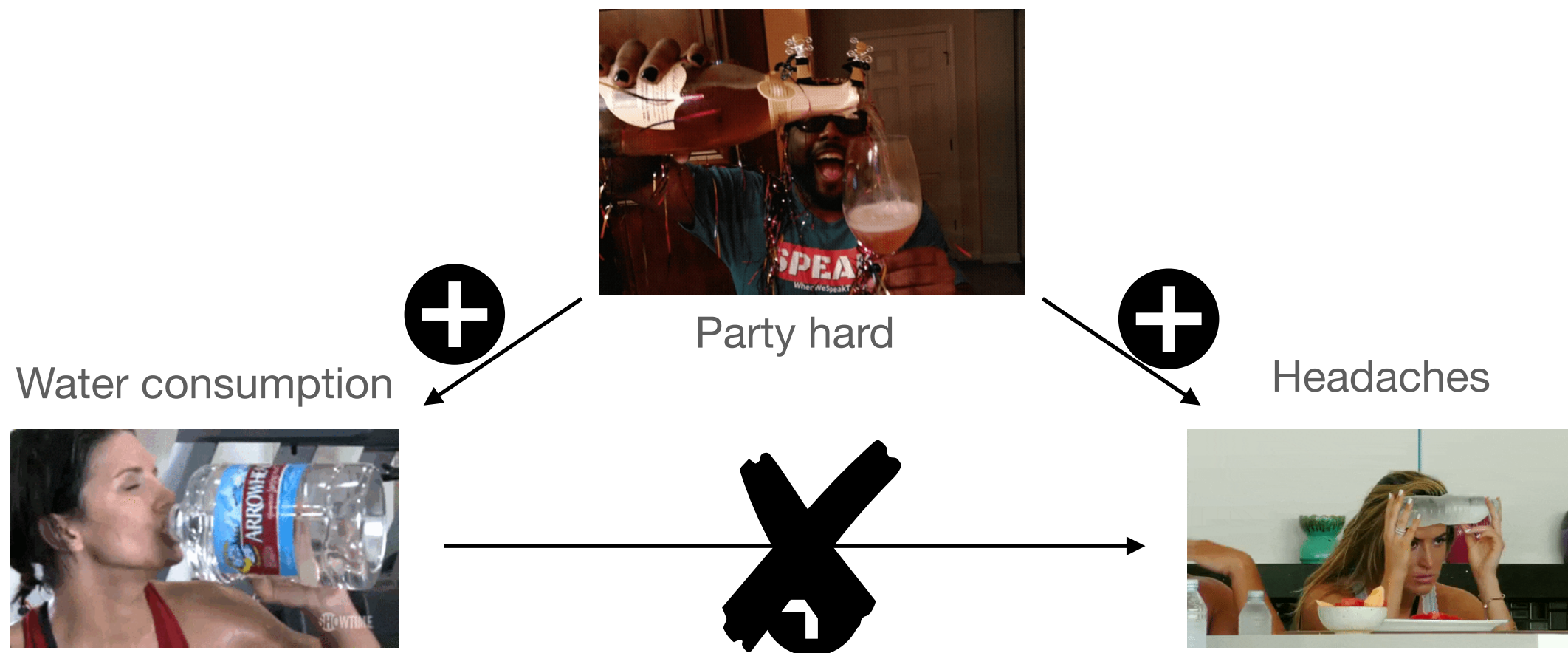
The consumer advice center wants to understand whether TV advertising causes people to buy things they otherwise would not have bought.



# Detour: 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**



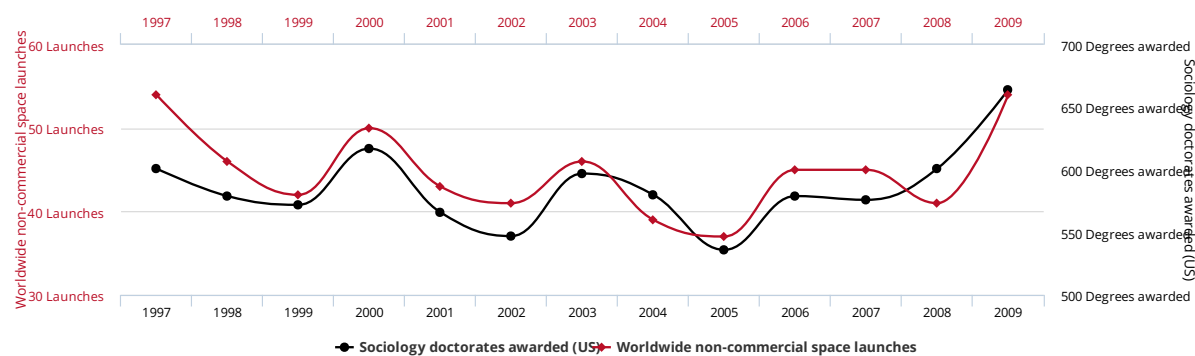
# Correlation and causation

- 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**

Worldwide non-commercial space launches

correlates with

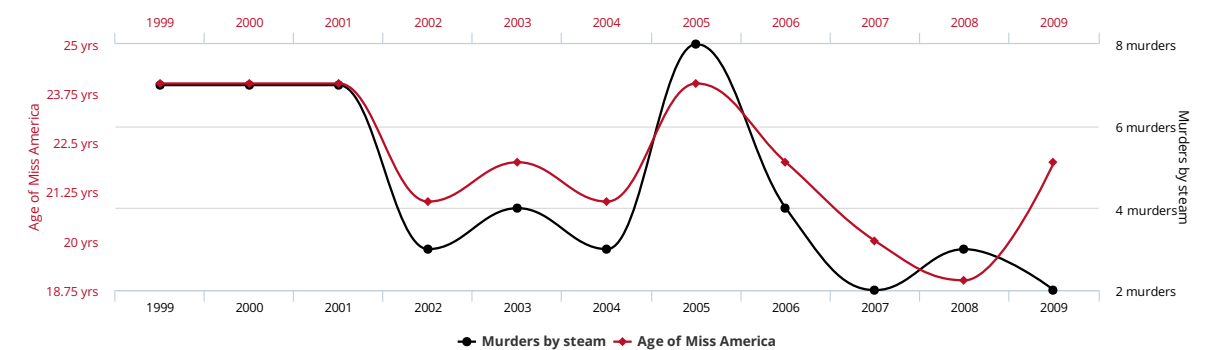
Sociology doctorates awarded (US)



Age of Miss America

correlates with

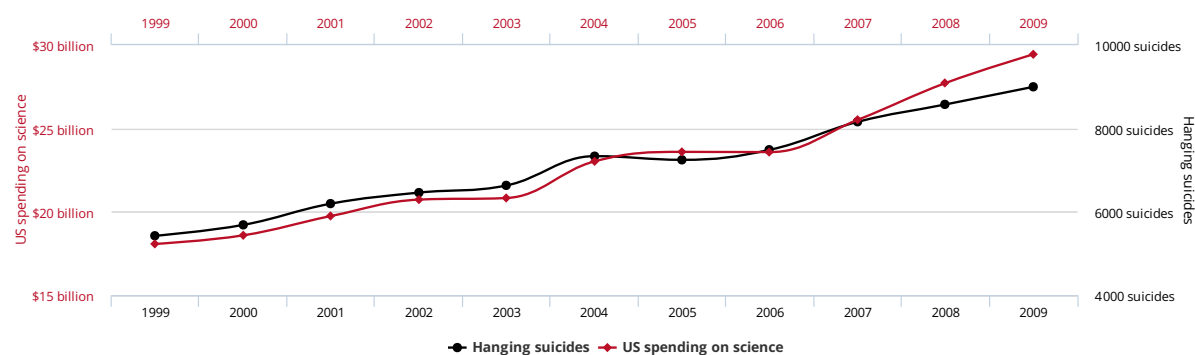
Murders by steam, hot vapours and hot objects



US spending on science, space, and technology

correlates with

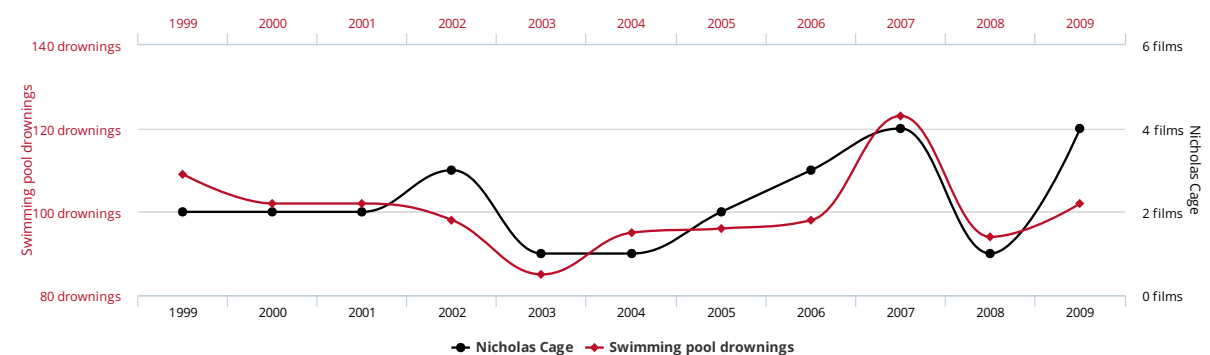
Suicides by hanging, strangulation and suffocation



Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in



# Correlation and causation

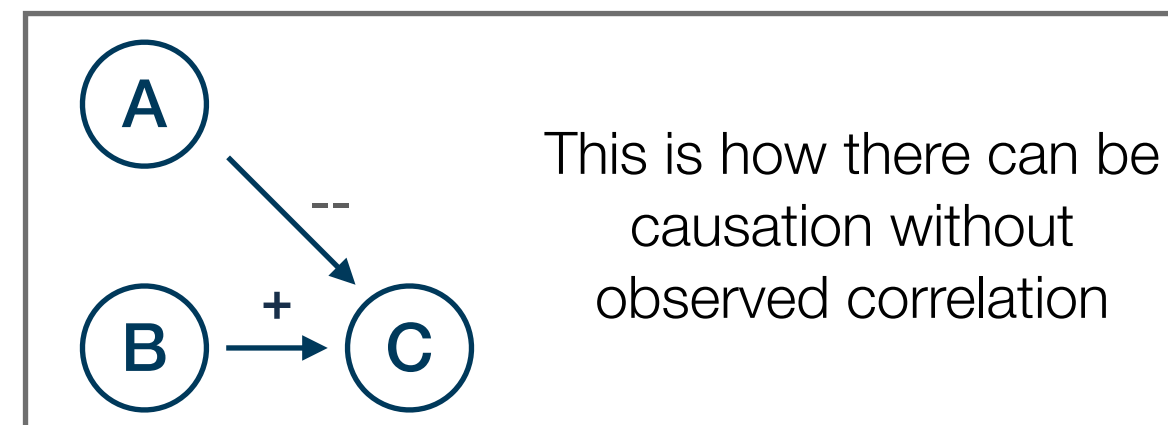
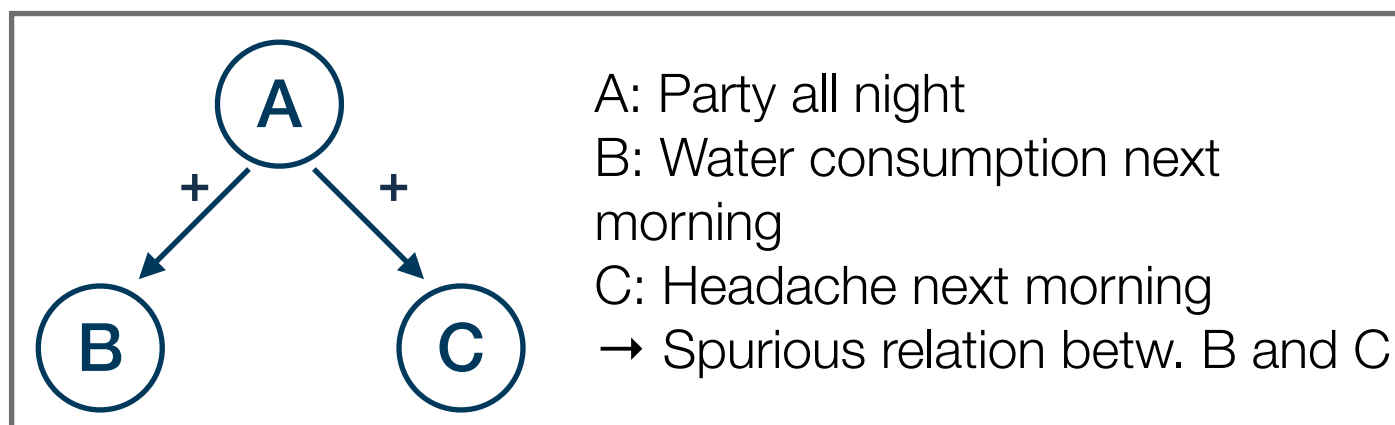
- The distinction between correlation and causation is central for any applied (data) scientist
  - Correlation describes an **observed relationship**
  - Causation refers to an (unobservable) **cause-effect relationship**
- If we observe correlation without causation as in the example we speak of a **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)
- We do not engage in causal modelling, but note that event simply directed cycling graphs (DAGs) help you to sort your thoughts about causation



Nancy Cartwright





# Correlation and causation

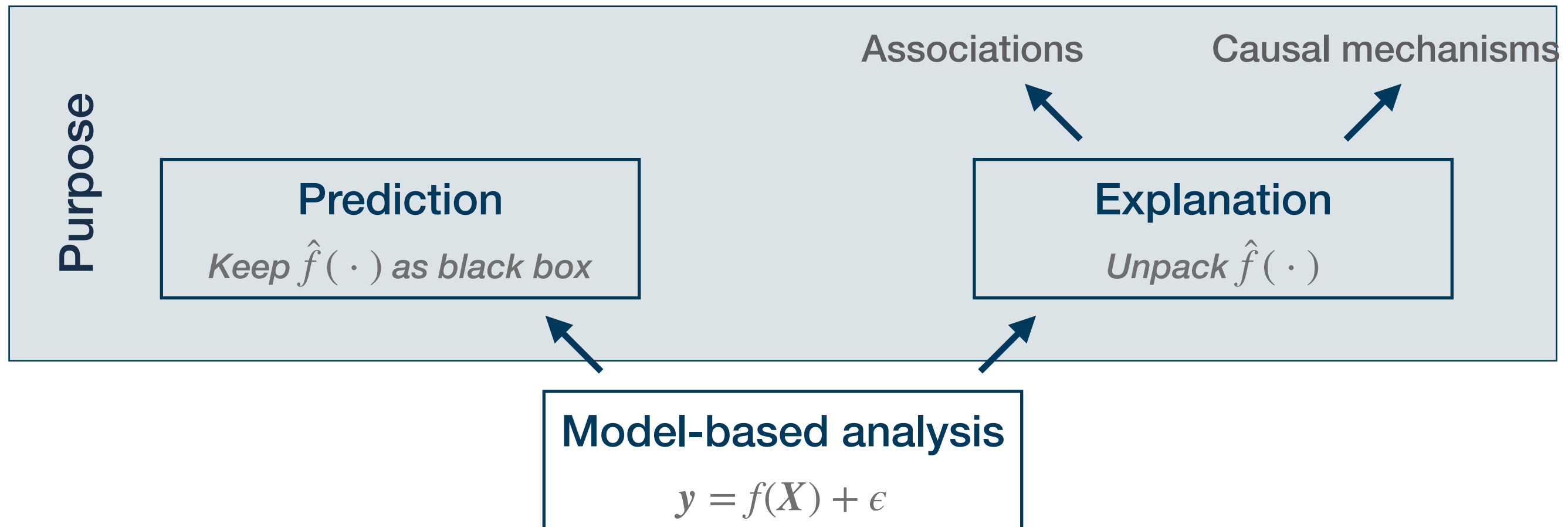
- Identifying causation is attractive but very hard
- It requires us to add theoretical hypotheses about cause-effect-relationships into a model
  - "No causes in, no causes out!"
  - This gives rise to **causal models** (which are often represented graphically)
- We do not engage in causal modelling, but note that event simply directed cycling graphs (DAGs) help you to sort your thoughts about causation
- What you will do is how to 'sort out' or 'control for' variables being related to your variable of interest
  - Example: What is the relative association of migration background and income with criminal activity?



Nancy Cartwright

# Back to categories...

# Types of inferential data analysis I



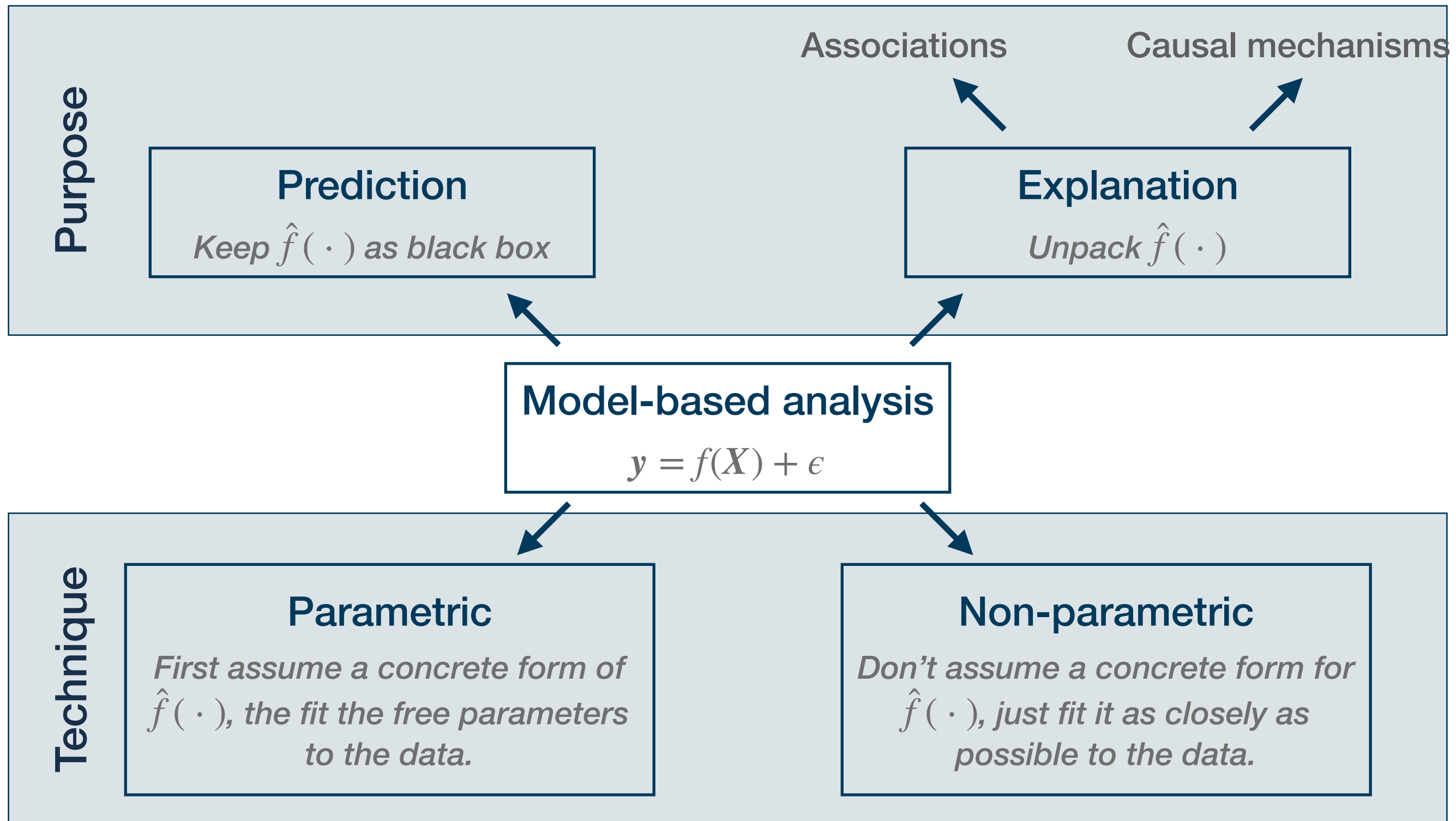
## Group work:

*Think about when you used or referred to data analysis in your studies so far. Were predictive or explanatory analyses more common? What would you consider more interesting for your final thesis?*



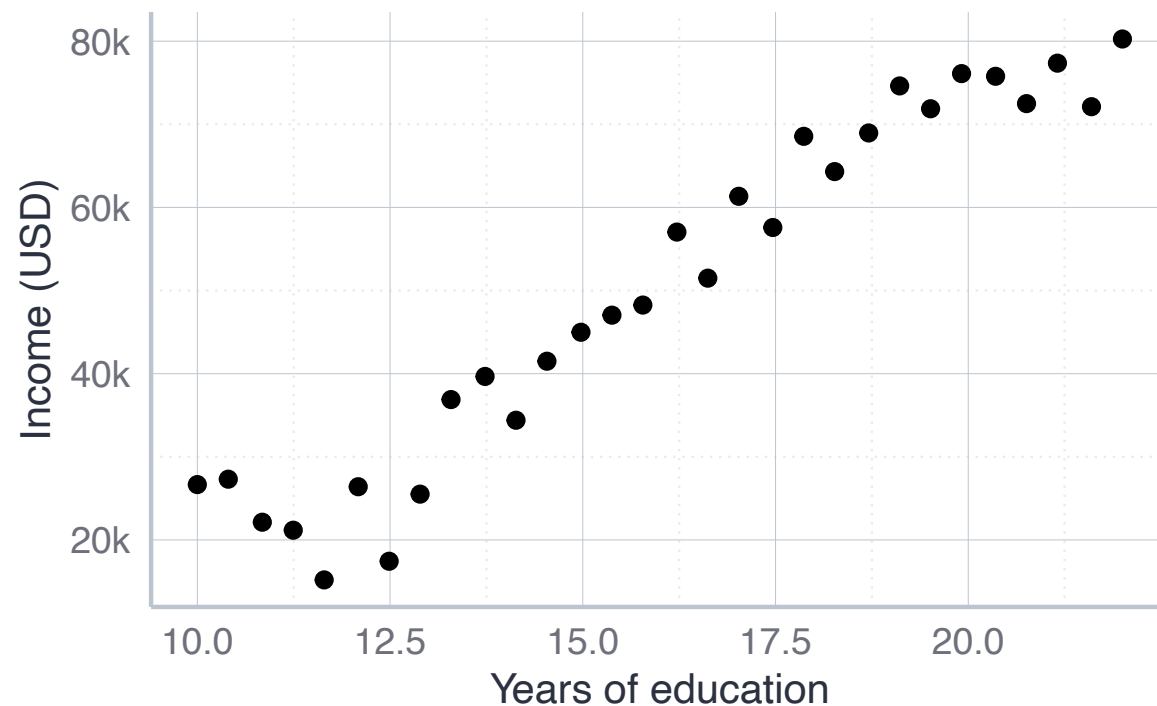


# Types of inferential data analysis I



# Types of inferential data analysis I

## Examples for parametric and non-parametric approach



What is the association  
between years of education and  
annual income?

# Types of inferential data analysis I

## Examples for parametric and non-parametric approach

### The parametric approach

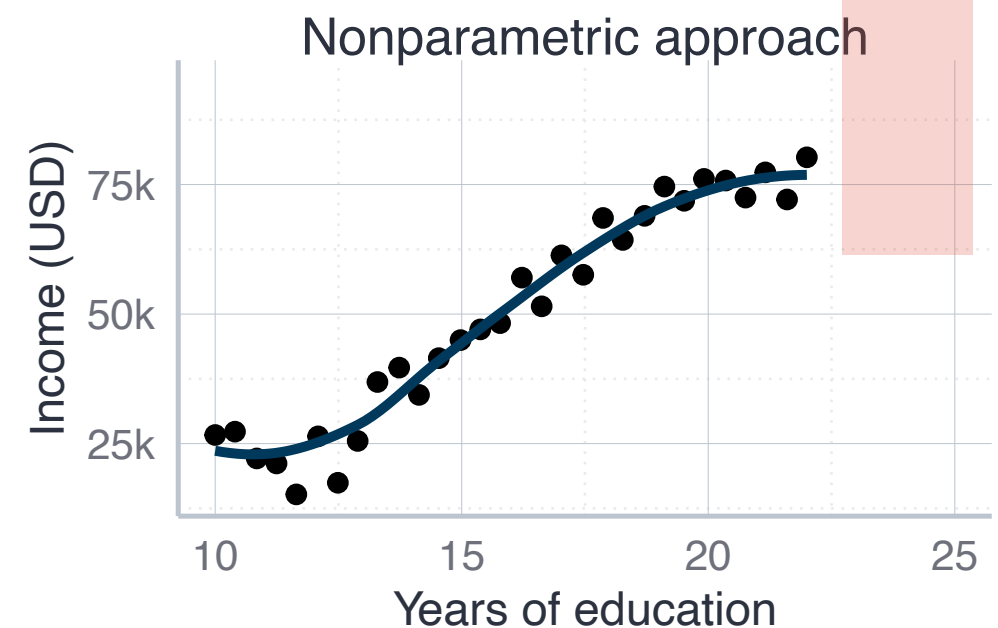
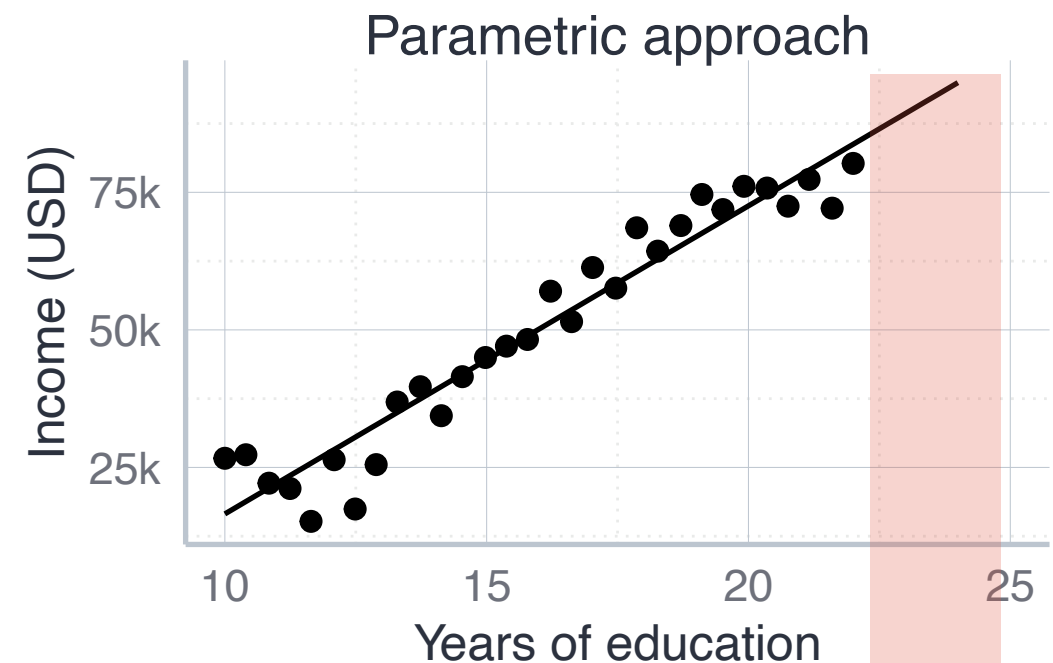
1. Assume a particular functional form:

$$f(X) = \beta_0 + \beta_1 \cdot EDUC + \epsilon$$

2. Estimate the free parameters  $\beta_0$  and  $\beta_1$ :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot EDUC$$

3. Use  $\hat{\beta}_0$  and  $\hat{\beta}_1$  to obtain  $\hat{Y}$  for hypothetical values of  $EDUC$

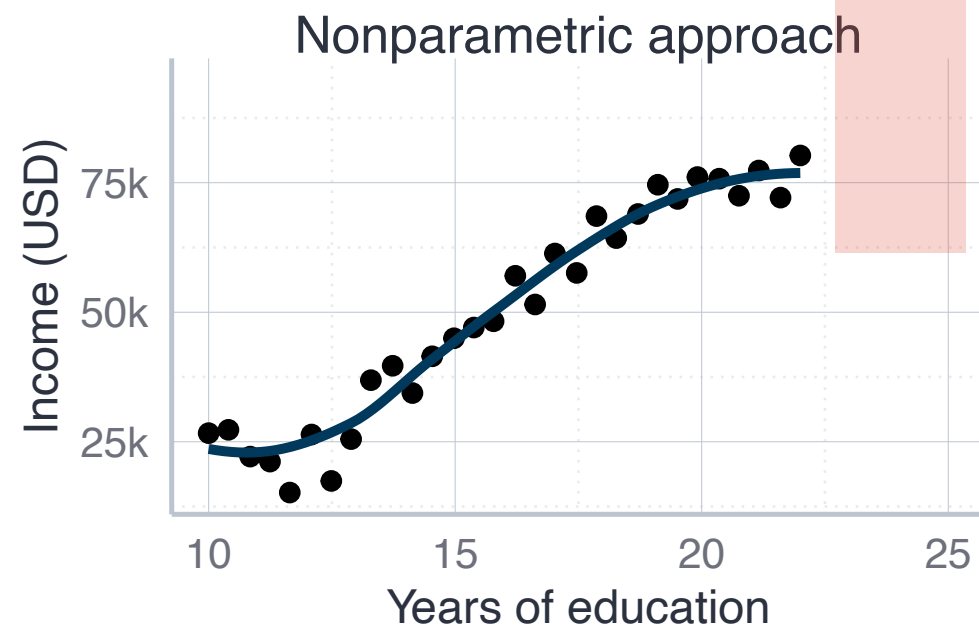
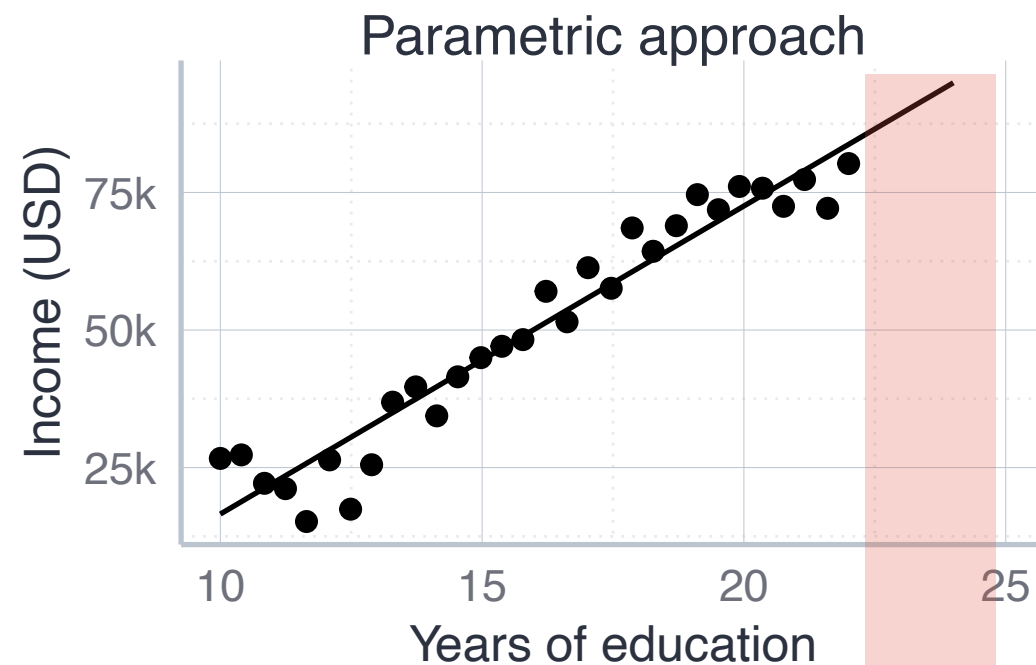


### The non-parametric approach

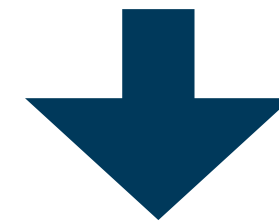
1. Make some regularity assumptions
2. Fit a high-dimensional polynomial or similar

# Types of inferential data analysis I

## Examples for parametric and non-parametric approach



Why did the non-parametric model not extrapolate beyond the data?



Source: [xkcd](#)

# Types of inferential data analysis II

## Different approaches to machine learning (ML)

### Supervised ML

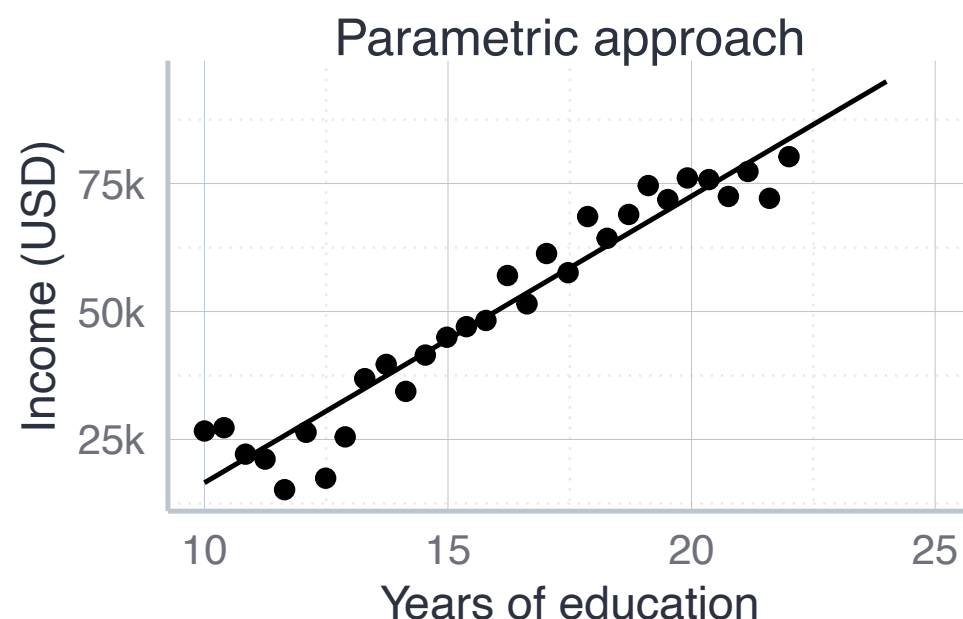
*For each input variable  $x_i \in X$ , there is an output variable  $y_i \in y$*

*We say input data is labelled*

*We are interested in the relationship*

$$y = f(X)$$

*Clear quality criteria for the output*



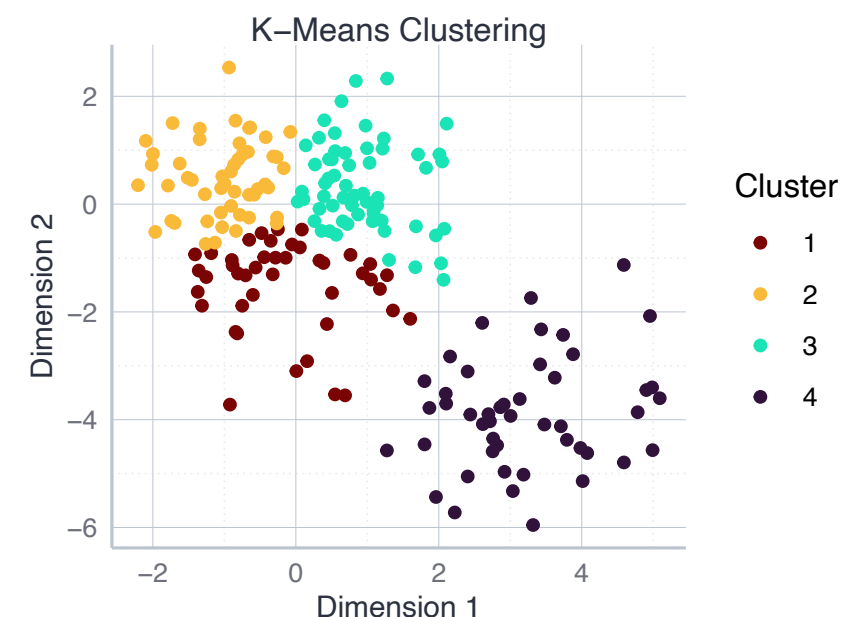
### Unsupervised ML

*We only have input variables  $x_i \in X$ , but there is no output variable*

*We say data is not labelled*

*We are interested in deeper structures in  $X$ , not in a relationship*

*Few (if any) quality criteria for the output available*



# Types of inferential data analysis II

## Different approaches to machine learning (ML)

### Supervised ML

*For each input variable  $x_i \in X$ , there is an output variable  $y_i \in y$*

*We say input data is labelled*

*We are interested in the relationship*

$$y = f(X)$$

*Clear quality criteria for the output*

### Unsupervised ML

*We only have input variables  $x_i \in X$ , but there is no output variable*

*We say data is not labelled*

*We are interested in deeper structures in  $X$ , not in a relationship*

*Few (if any) quality criteria for the output available*

### Semi-Supervised ML

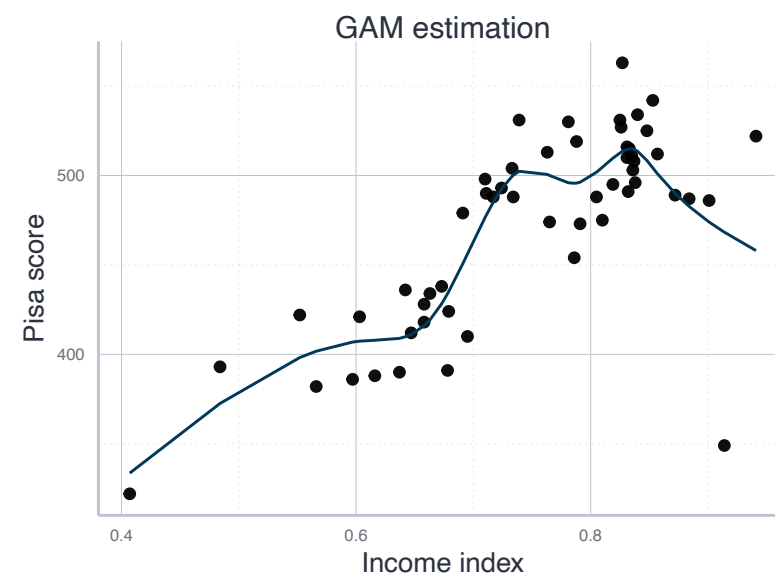
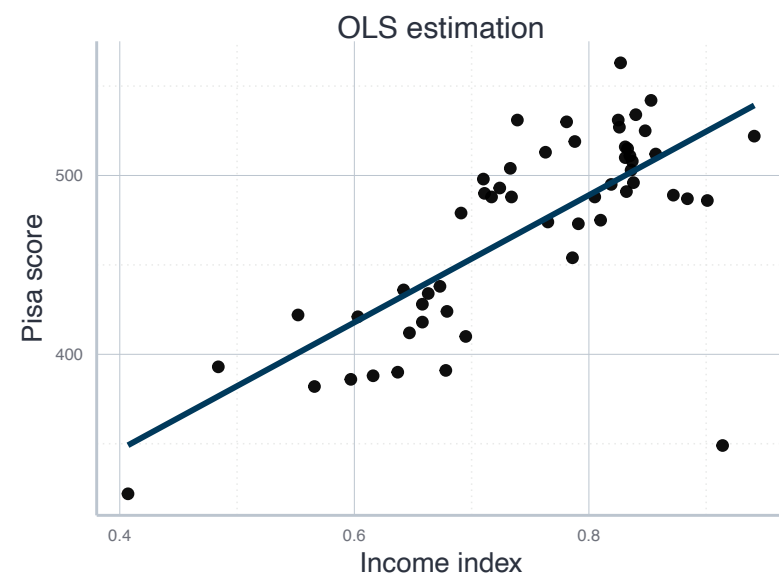
*For some input variables an associated output variable exists, for others not*

### Reinforcement learning

*Algorithms receive constant feedback on their performance and adapt accordingly.*

# A trade-off between flexibility and interpretability

- Compare the analysis of wages using a simple linear regression model estimated by ordinary least squares, and a generalised additive model:

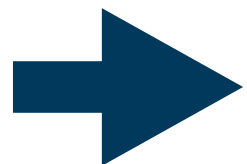
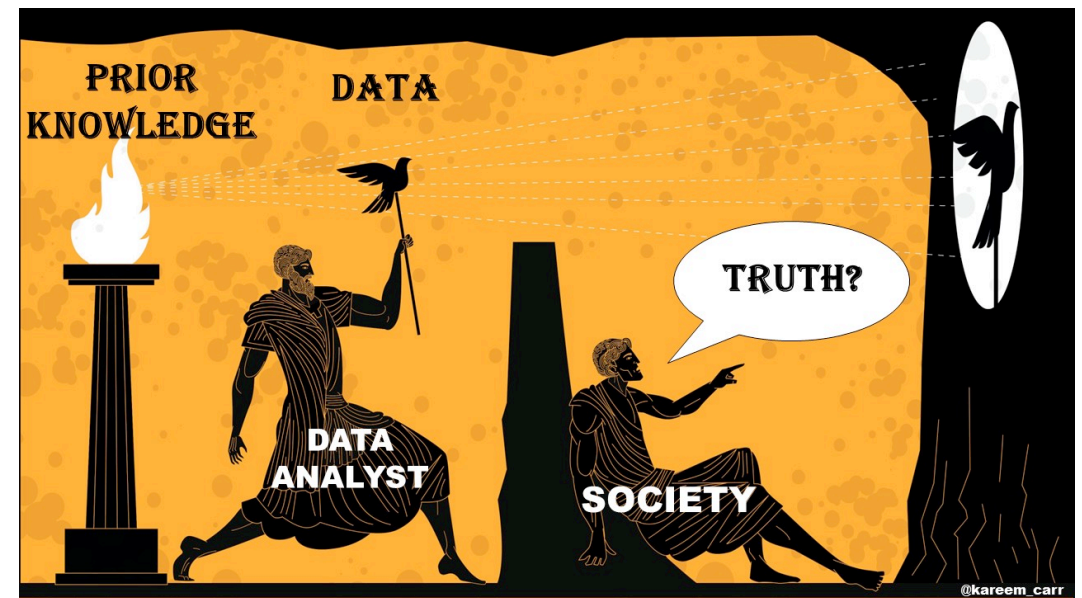


- Why would you choose the less flexible regression approach?
- More flexibility often comes with the cost of lesser interpretability:
  - The slope in the OLS model can be interpreted as the marginal association of income and the Pisa score
  - I have no idea how to interpret the slope in the GAM model...



# Final remarks on the categories

- The separations between categories is not always clear-cut
- But they are a very useful general guidance when choosing a method
- No approach is entirely superior → it always depends on your interest and the purpose of your analysis
- None of the approaches can yield a fully objective analysis
  - Due to the **theory-laddeness of observation** no empirical method can



Before choosing a method, think about what you want to achieve. And always be explicit and aware of your assumptions, they are never neutral!



# Recap questions

- What distinguishes a descriptive and inferential analysis?
- Explain the different parts of this general model formulation:  $Y = f(X) + \epsilon$
- What's the difference between  $Y$  and  $f(X)$  on the one, and  $\hat{Y}$  and  $\hat{f}(X)$  on the other hand?
- Explain the difference between correlation and causation
- What are the four different kinds of machine learning that you encountered?
- What distinguishes a parametric from a non-parametric approach to estimate  $f(\cdot)$ ? Which one is better?
- Explain what is meant by the *theory-laddeness of observation* and what this implies for data analysis.