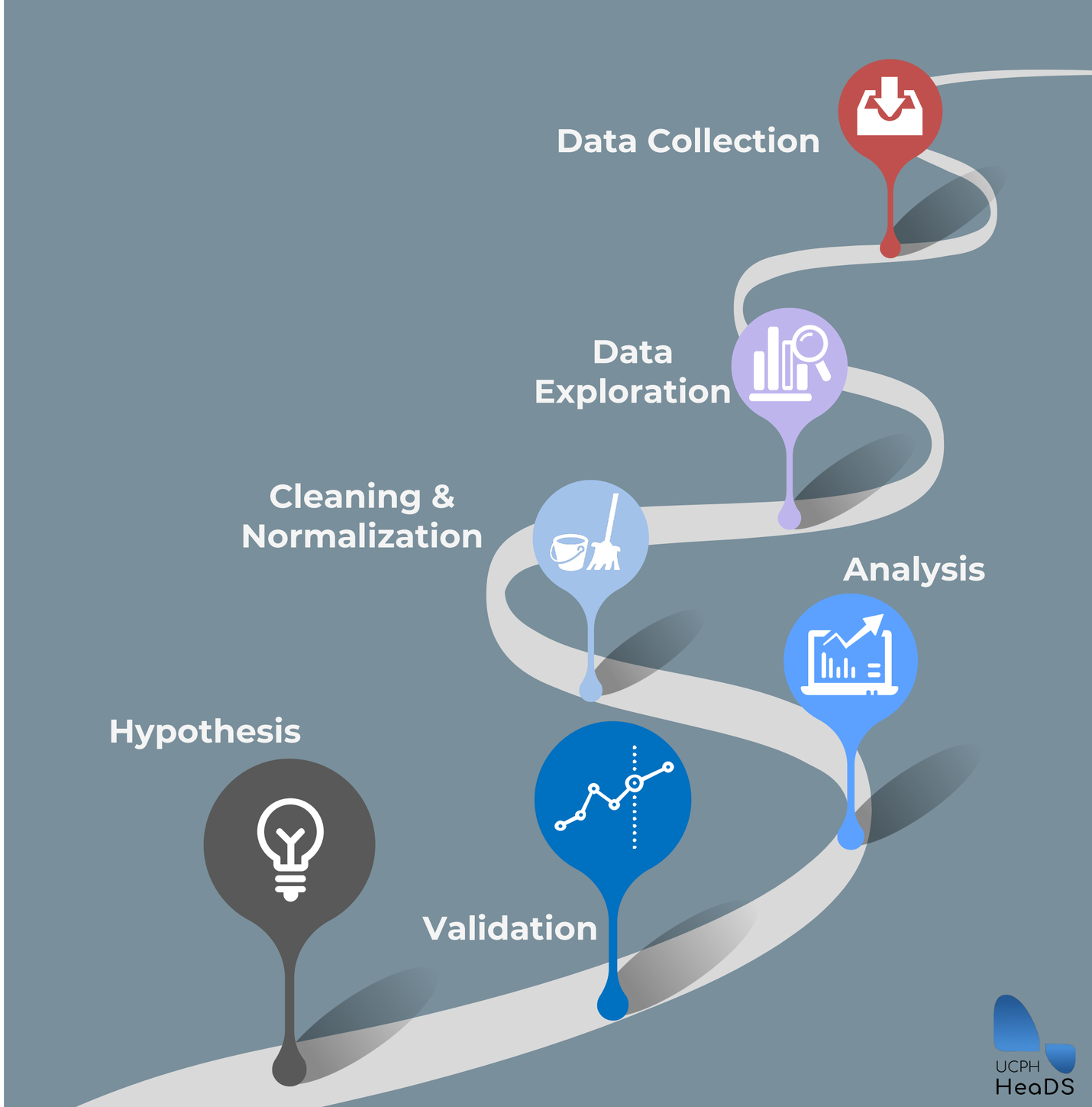
# THE DATA'S JOURNEY



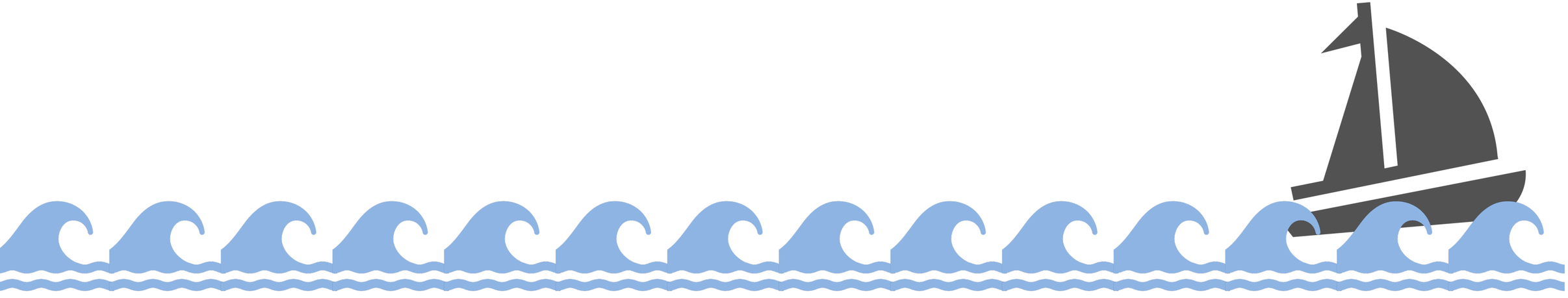
# FINAL PART OF OUR JOURNEY

DATA COLLECTION

EXPLORATORY DATA ANALYSIS

DATA ANALYSIS

MODEL EVALUATION

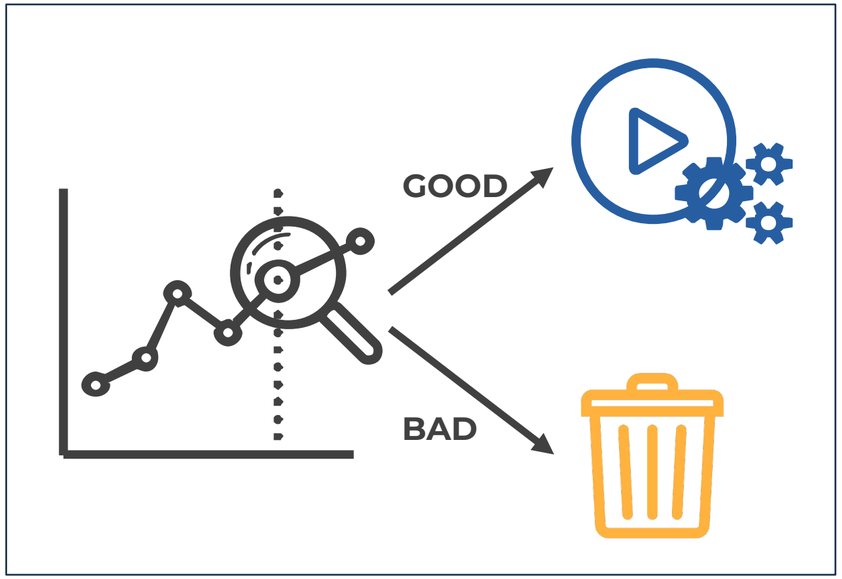


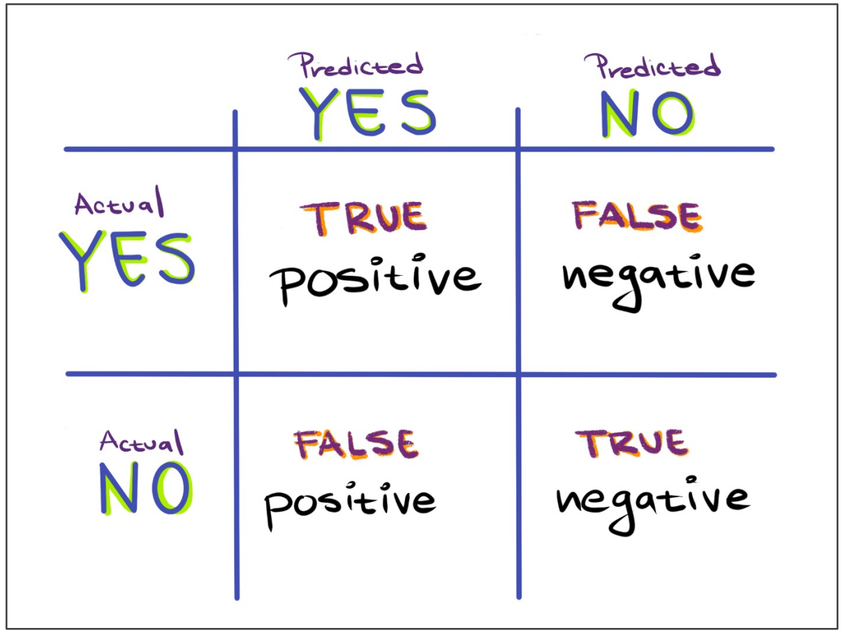
UCPH HeaDS

# MODEL EVALUATION

- How do we measure the performance of a model?
- How much/far can we trust the results we have obtained?
- How do we interpret the results of the evaluation?

*Let the RQ, data and model pick the evaluation metric*



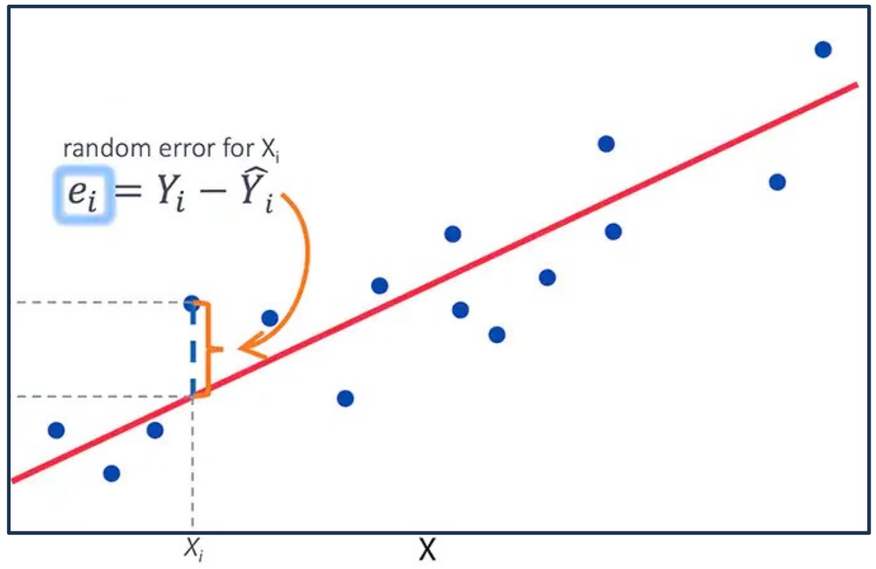




UCPH HeaDS

# MODEL PERFORMANCE

How well does the model fit the data?

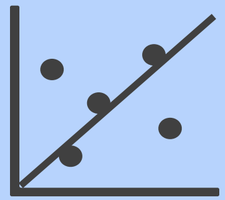
- Criteria depend on model type
- Good models capture the underlying trends and characteristics




## PERFORMACE METRICS:

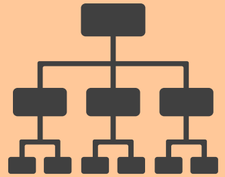
**REGRESSION**

- R-squared
- MSE
- RMSE
- RMSLE



**CLASSIFICATION**

- Accuracy
- Precision
- Recall
- ROC/AUC



**CLUSTERING**

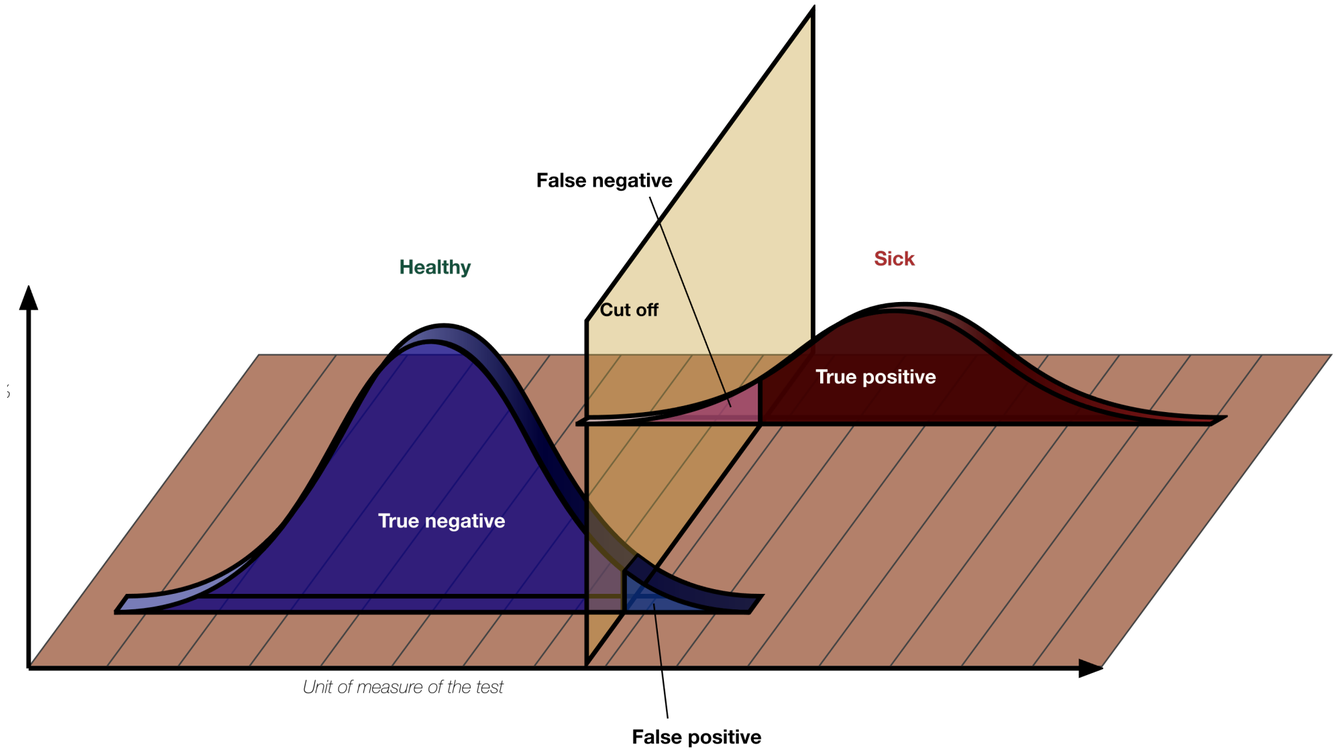
- Silhouette
- Adj. Rand Indx.
- Gap statistic
- Davies-Bouldin

UCPH HeaDS

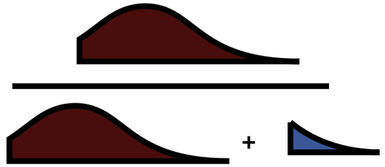
# CLASSIFIER PERFORMANCE

Performance measures of classifiers depend on the chosen cutoff.

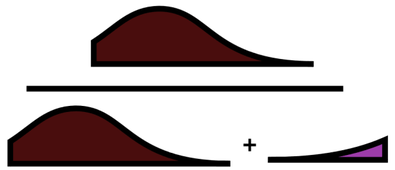
One can easily optimize on the number of false positives or false negatives, but not both at the same time.




Precision :



Recall/
True positive rate

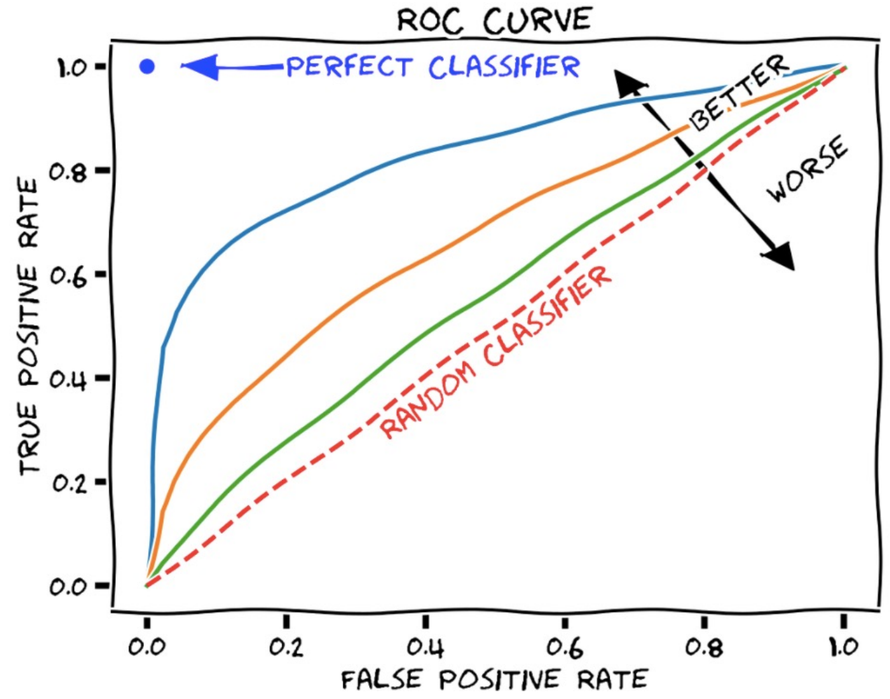


UCPH HeaDS

# AREA UNDER THE CURVE (AUC)

This is why we use the AUC to report on the goodness of the classification.

How many of the true cases have we discovered?




- Receiver Operator Curve (ROC)
- Plotted by using every possible cutoff.
- Each True Positive Rate has an associated False Positive rate.

UCPH HeaDS

# GROUP DISCUSSION 4.1

For a medical screening procedure to detect a disease, what does each of these terms correspond to:

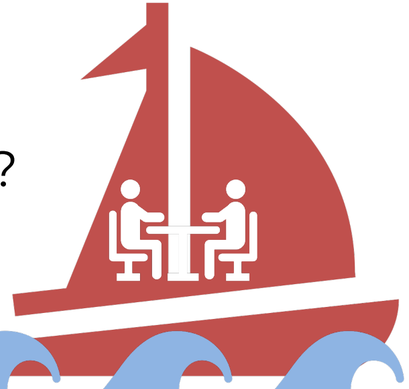
- True positive
- False positive
- True negative
- False negative

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

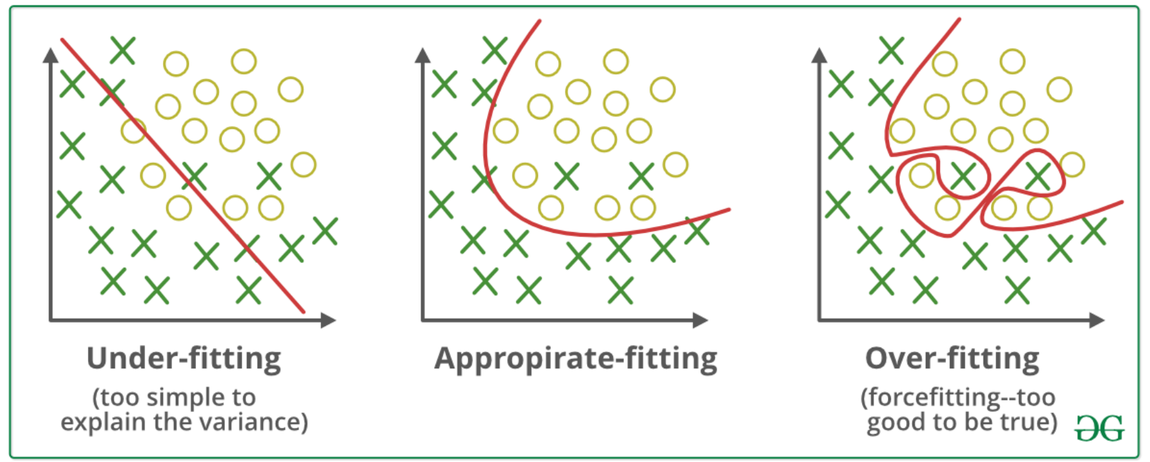
On the right you have the formulas for precision and recall.

Discuss what are the benefits and drawbacks of optimizing one over the other in relation to a disease screening program?

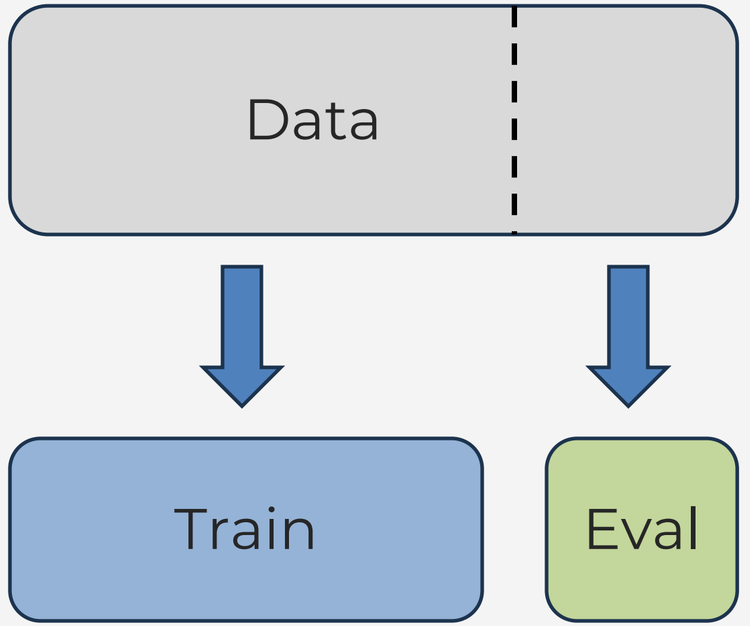


UCPH HeaDS

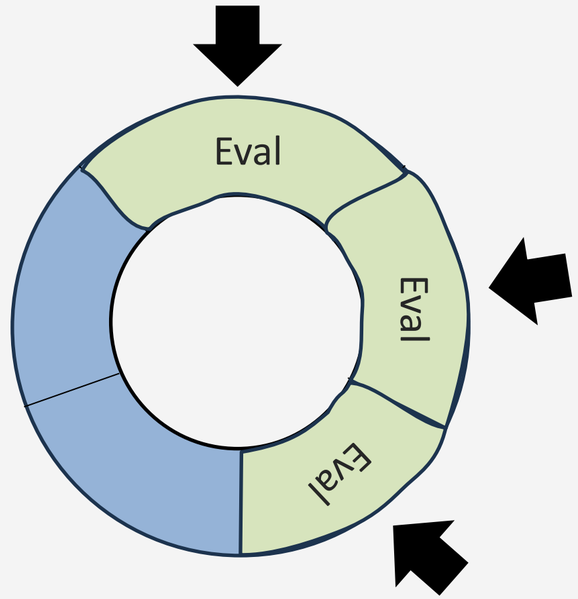
# OVERFITTING



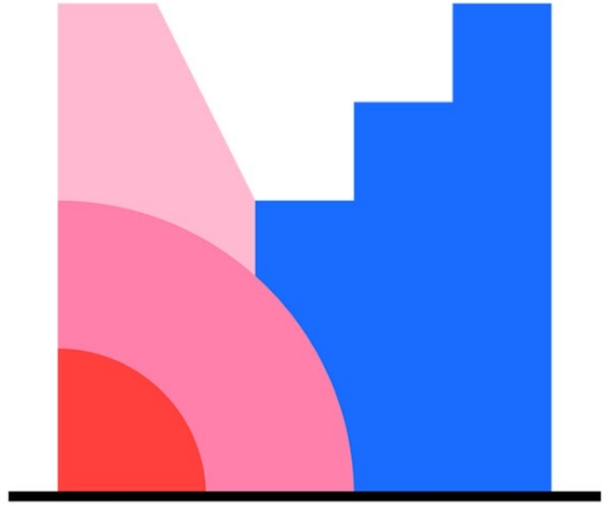

- When a model follows the data too closely we get an effect known as overfitting.
- We avoid this by splitting the data into training and evaluation sets.




Cross Validation




UCPH HeaDS



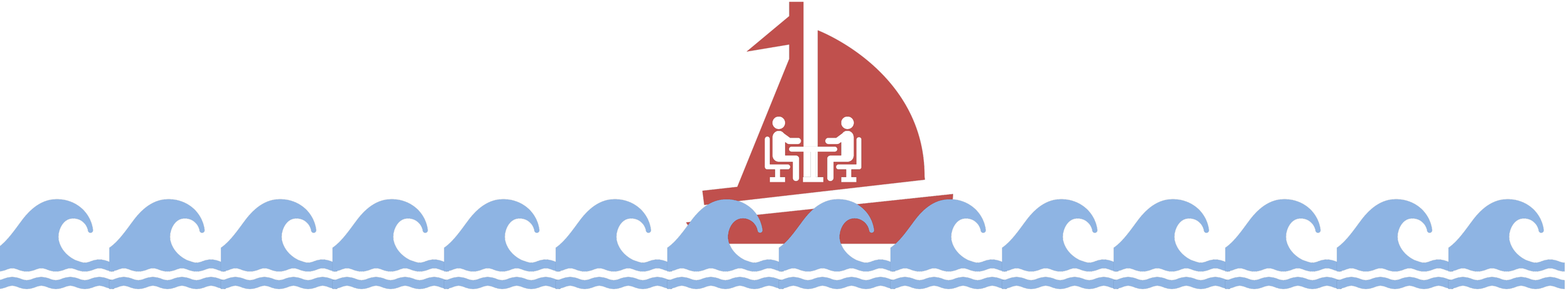

UCPH HeaDS

# GROUP DISCUSSION 4.2

On your table you have a print-out in which the same data has been fitted with three different models (the red lines).

In your group, discuss which of these you think is the **most suitable** and **why.**

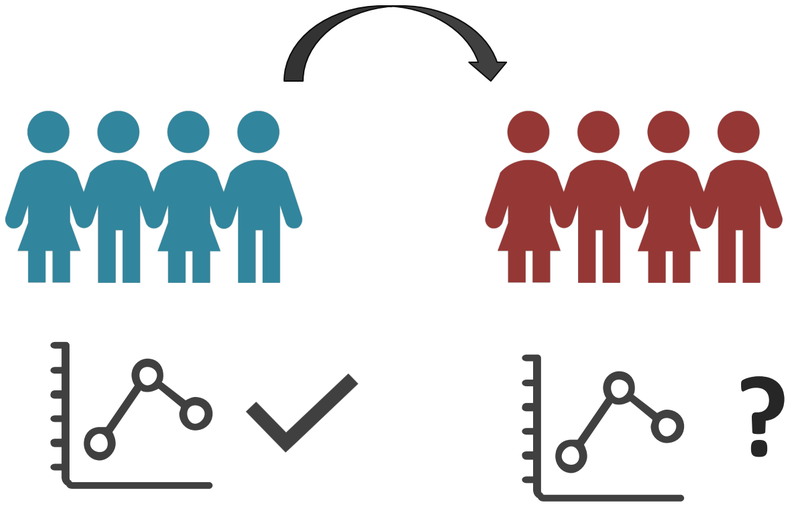
In general, what is the problem with high-dimensional model with many parameters?



UCPH HeaDS

# BIASES

- What biases could have influenced the results?
- Bias arises primarily from how the data was gathered and processed.
- Bias can reduce model validity, i.e. a model trained in one population may not work in another.

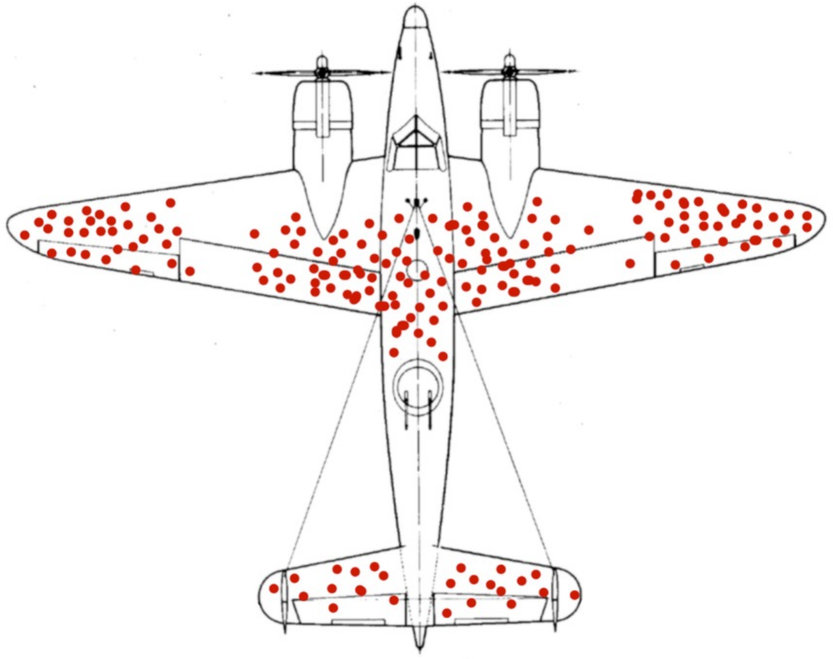




UCPH HeaDS

# GROUP DISCUSSION 4.3

Consider the following schematic of bullet holes on returning WW2 planes.
It shows with red circles where each plane was struck.

Bases on this data, which part of the plane do you think should be reinforced to better protect the plane from being damaged?

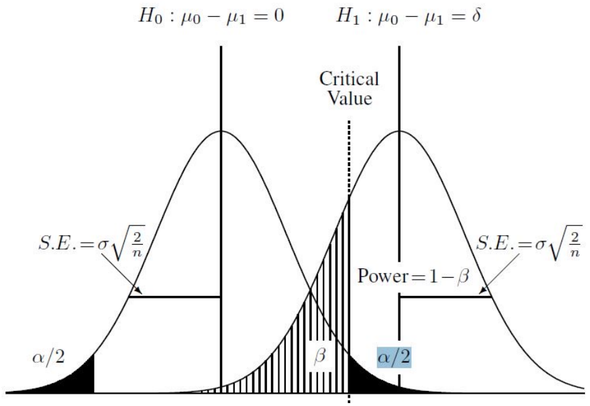


UCPH HeaDS

# WHAT DO THE RESULTS MEAN?

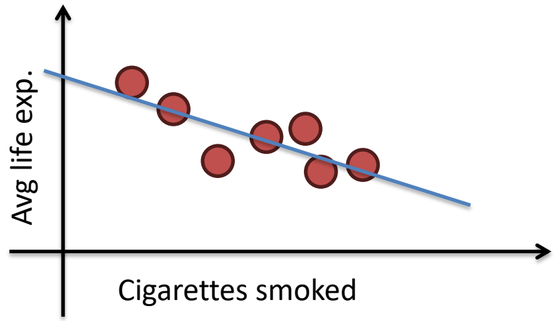
## Hypothesis testing

A significant p-value supports the rejection of the null hypothesis.




## Classification / Prediction

- Poor model performance, the model does not capture the data pattern.
- Interpretable parameters?
- Coefficients/Feature importance tell which predictor variables have a large influence in outcome




**Regression model:**

birth weight ~ mother's weight + smoking + mother's age

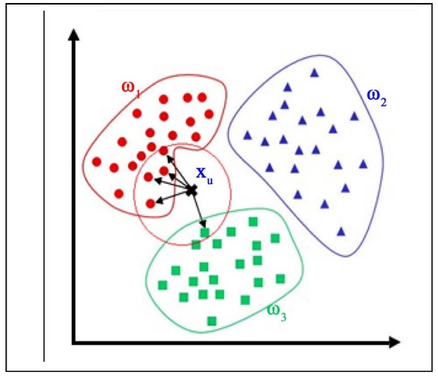
UCPH HeaDS

# WHAT DO THE RESULTS MEAN?

## Unsupervised Learning/Clustering

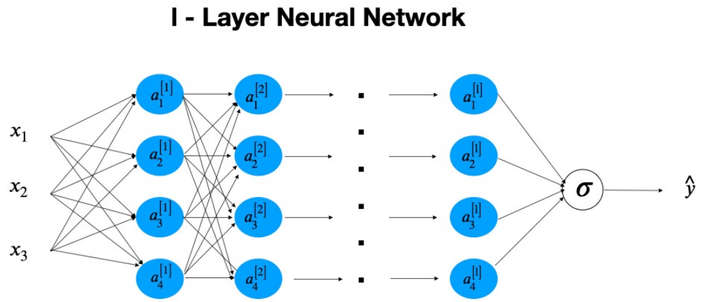
- Insights about the structure of the data, esp. number of clusters
- Further analyse properties of discovered clusters (**biological meaning**)

- Predict cluster membership of new data points




## Black Box models

- Highly non-linear models with many parameters are difficult to interpret
- We mostly use the performance instead of trying to gain insights



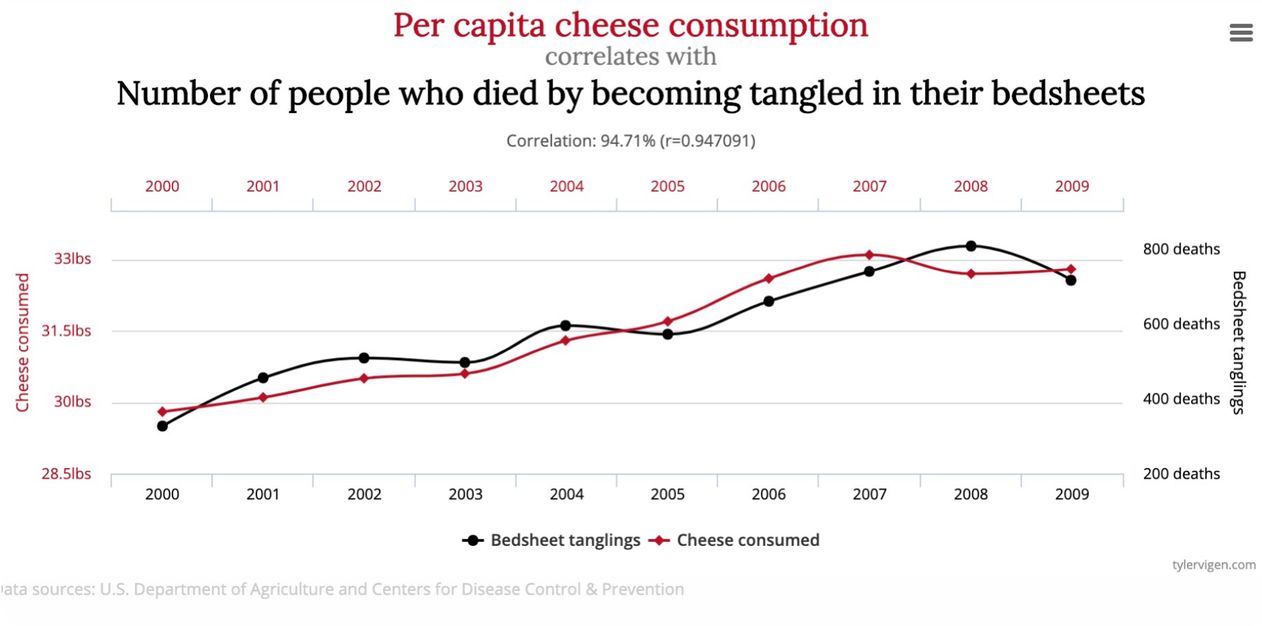

UCPH HeaDS

# CORRELATION ≠ CAUSATION

One goal of modelling is to discover relationships between predictor and outcome variable.

However, modelling is **not primarily concerned with causality**.

If our model shows that a relationship exists, that does not mean the relationship is causal.



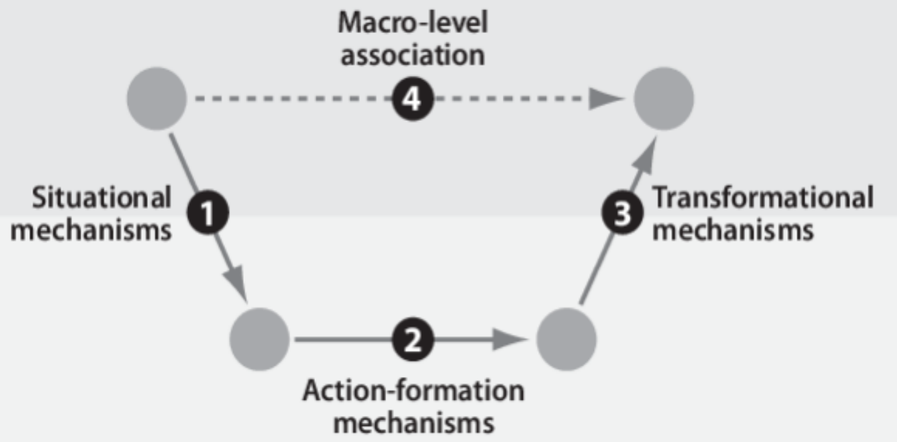

UCPH HeaDS

# THE QUESTION OF CAUSALITY

Where does this leave us?

- Causality can only be surmised in the statistical framework of causal inference.

- **Mode of action** is vital.
- If A causes B, how so?
- Does that make sense in the framework of existing domain knowledge?

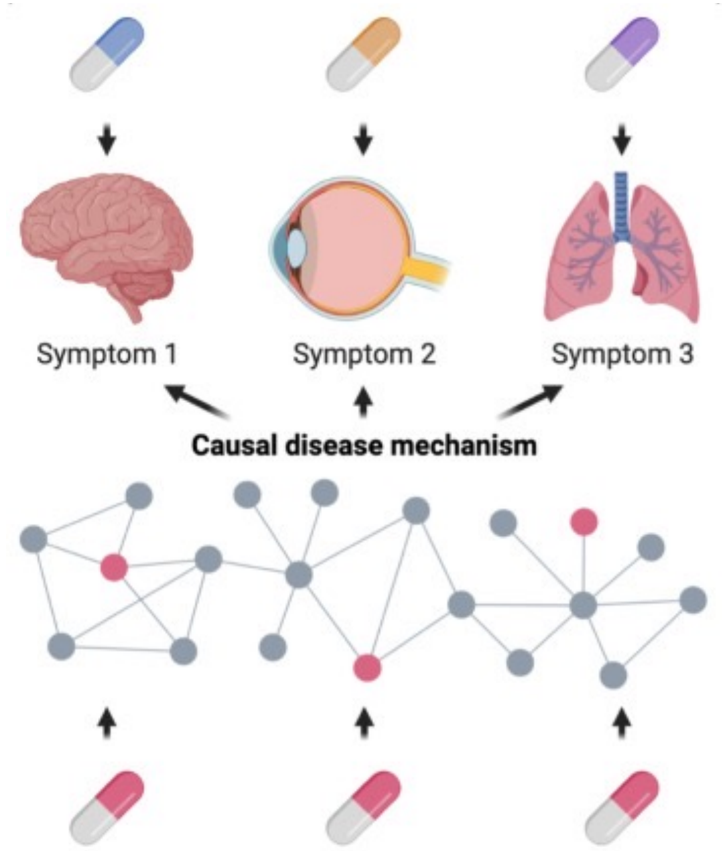



Hedström, P. and Petri Y. " *Annual review of sociology* 36 (2010): 49-67.

- Causality can be shown by gathering convergent evidence from different lines of inquiry and direct cause-effect experiments (animal/cell studies).

UCPH HeaDS

# THE QUESTION OF CAUSALITY




Nogales, Cristian, et al. *Trends in Pharmacological Sciences* 43.2 (2022): 136-150.

Is causality always central in (Health) Data Science Analysis?

The answer is that it depends on what we are trying to achieve:

**A question of causality:**

- **YES:** Underlying biological mechanism -> true treatment of disease
- **NO:** Diagnosis of disease (causality is nice but not necessary)

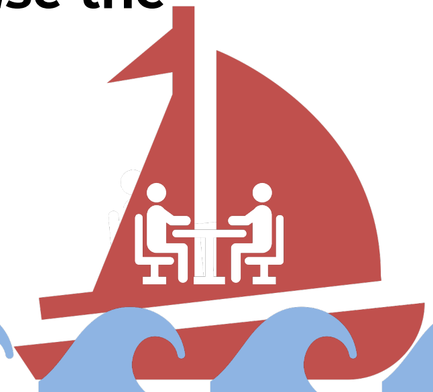
UCPH HeaDS

# GROUP DISCUSSION 4.4

1. In a cross-sectional study, researchers observed that the number of vaccinations a population received was positively associated with the prevalence of certain diseases.

2. In a cohort study, researchers found an inverse (negative) relationship between exercise and prevalence of arthritis in a group of individuals, i.e. less exercise more likely to be diagnosed with arthritis.

**Discuss the two studies above and their conclusion, does vaccination *cause* an increase in disease development and does lack of exercise *cause* the development of arthritis? Think about confounders...**

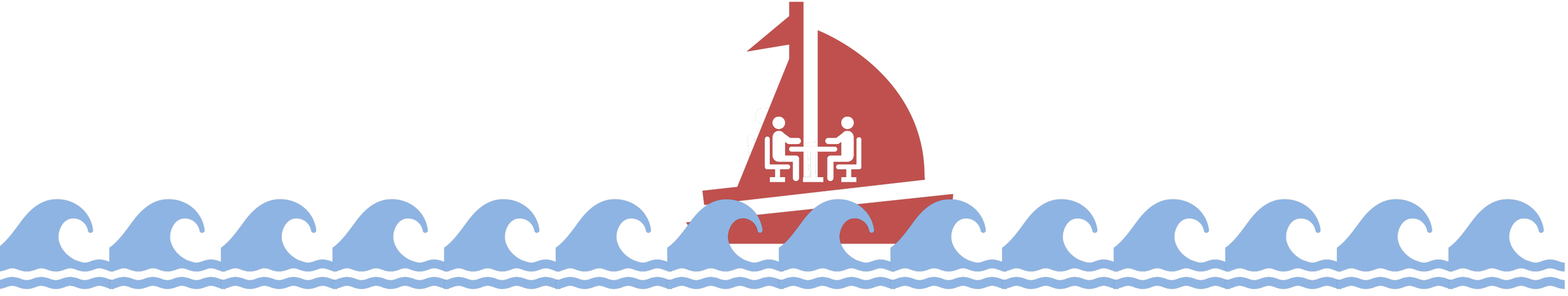


UCPH HeaDS

# GROUP DISCUSSION 4.5

1. In a recent population study on alcohol consumption and cancer development, analysis revealed that people who often drink wine, have an increased risk of developing lung cancer compared to people who do not.

**Does drinking wine cause lung cancer? Can you think of any confounding factors which could explain this correlation, i.e. life style choices, socioeconomic group, etc.?**



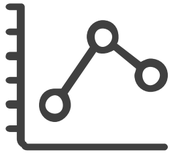
UCPH HeaDS

# OVERVIEW OF OUR JOURNEY

DATA COLLECTION

EXPLORATORY DATA ANALYSIS

DATA ANALYSIS



MODEL EVALUATION

UCPH HeaDS

# HEALTH DATA SCIENCE – YOUR NEXT STEP

## COURSES

**HeaDS:**

- ***Programming (Python, R)***
- ***High performance compute***
- ***Applied health DS analysis***

**Biostatistics**
**NNF Centers (ReNew, CPR)**
**SUND faculty**

## DATA SCIENCE COLLABORATION

- **Conferences & seminars on DS**
- **Matchmaking events**
- **Contact relevant department, group, support - and core facility**

## ONLINE RESOURCES

- **https://www.coursera.org/**
- **https://www.codecademy.com**
- **https://www.datacamp.com/**
- **https://towardsdatascience.com/**
- **Stack Overflow & Reddit**
- ***Books:***

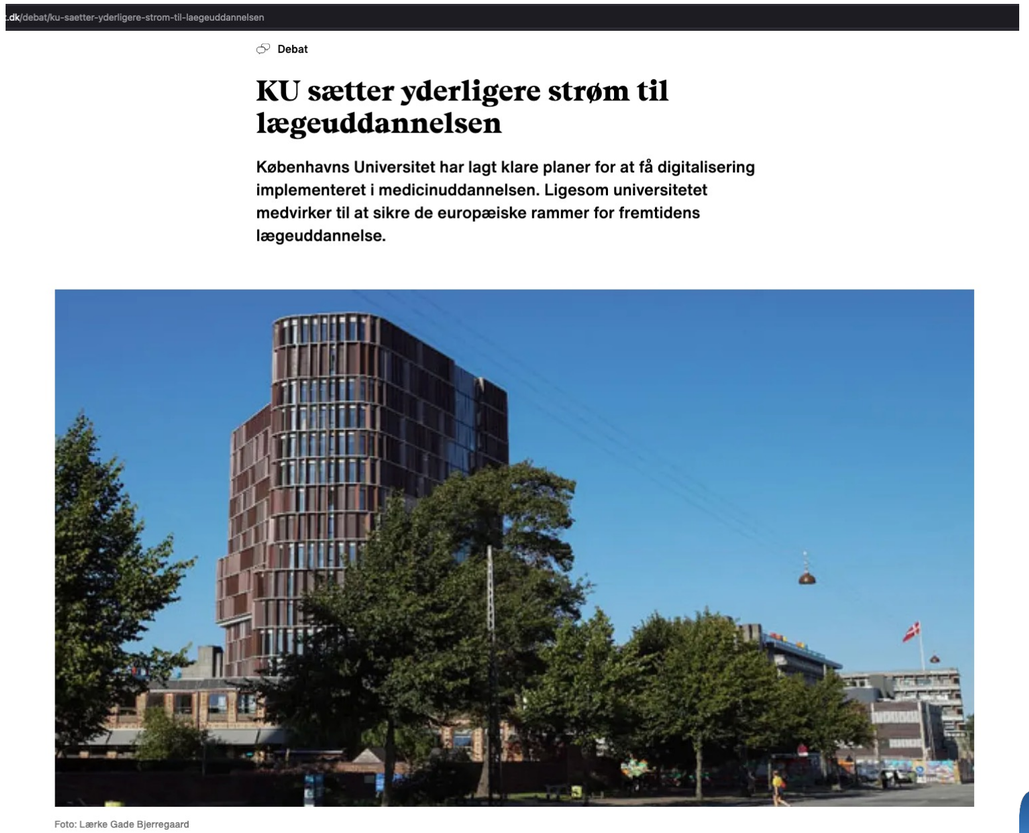
***”R for Data Science”***
***“Python Data Science Handbook”,***
***“Statistics for Health Data Science”***

UCPH HeaDS

# BACK TO THE FUTURE

The Digital Core Curriculum (**DCC**) initiative:

- KU educations to include digital literacy and data science competences
- DS in Medical Research is the future!
- As a domain expert, so what is your role?
    - Form a collaboration with a data scientist
    - Encourage your staff to take DS courses
    - Know where to get help with DS analysis

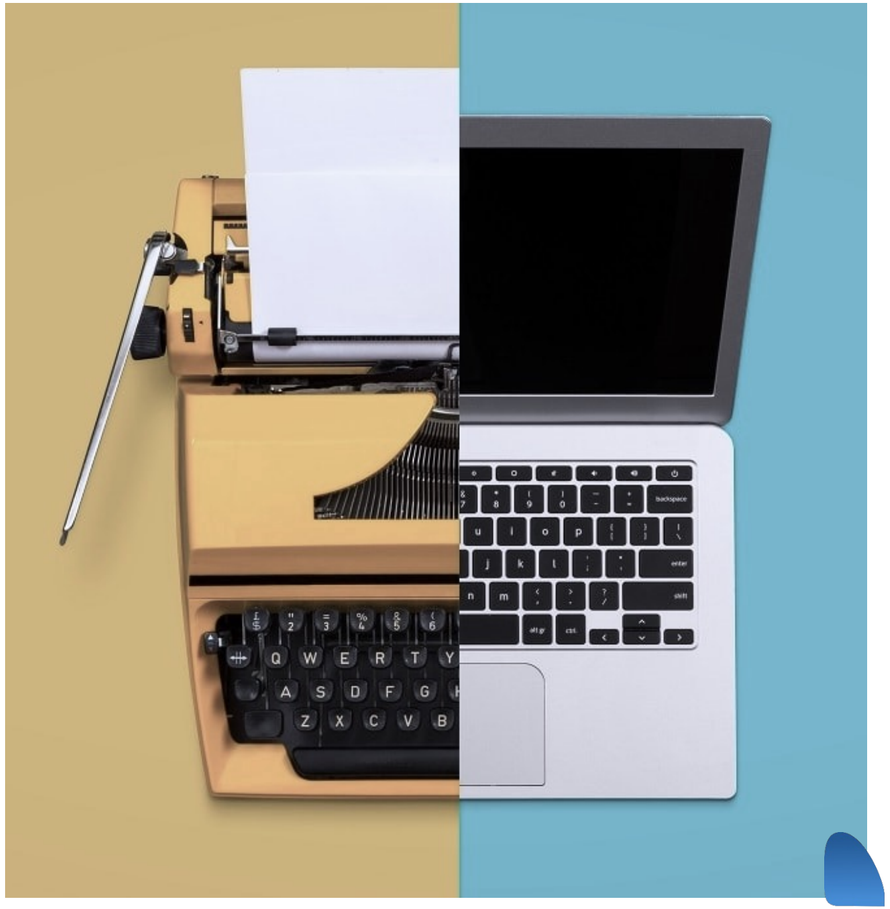



UCPH HeaDS

# DATA SCIENCE IN YOUR TEACHING

- **Could you incorporate more data science in your teaching?**
  - Take courses and learn yourself
  - Get inspired by colleagues
  - Collaborate with colleagues in DS on course materials

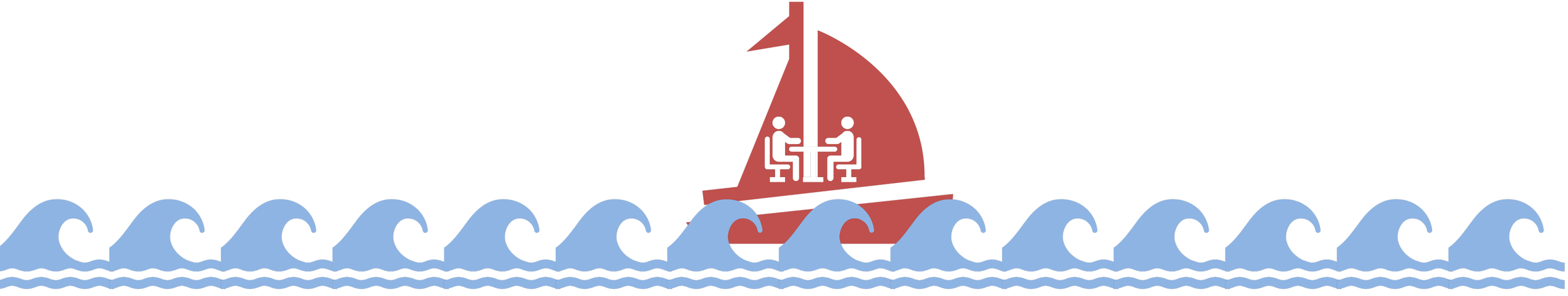
- **Start a little at a time:**
  - Move from Excel → R or Python
  - Think about old analysis and RQs in a new ‘DS light’
  - Encourage your students to explore DS tools (Github, Programming, HPC)
  - Get familiar with ChatGPT



UCPH HeaDS

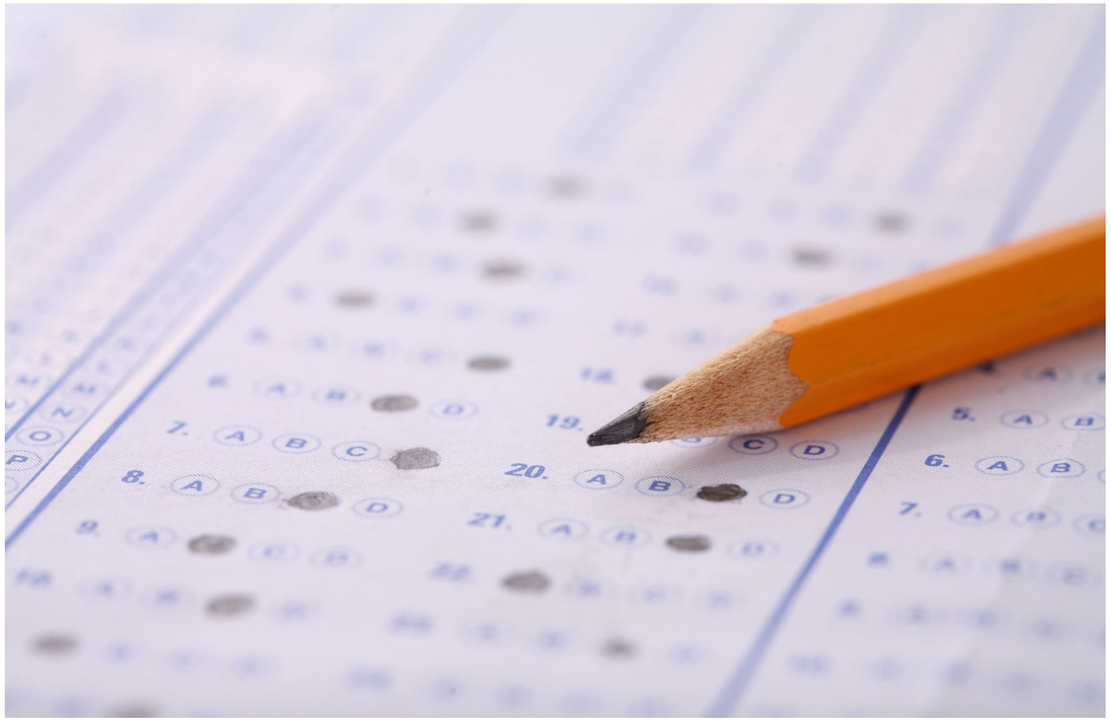
# GROUP DISCUSSION 4.6

- What is your main take away from today?
- Has what your have learned today changed your perspective on Data Science?
- How do you think your future relationship to Data Science will be?
- In what ways can you incorporate Data Science thinking into your teaching?



UCPH HeaDS

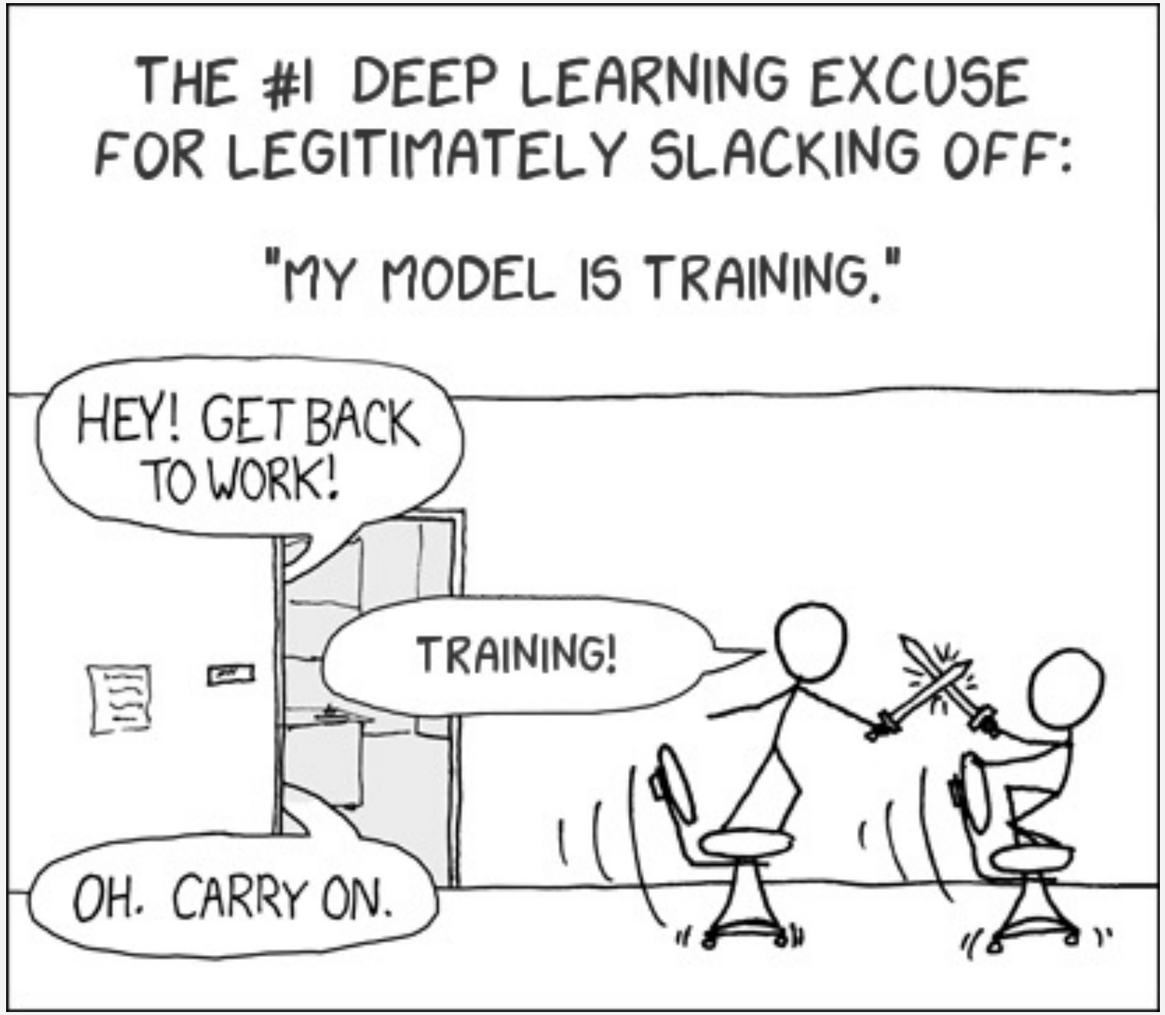
# COURSE EVALUATION



https://forms.office.com/e/Ffc2csjk7R

UCPH HeaDS

# THANK YOU




UCPH HeaDS