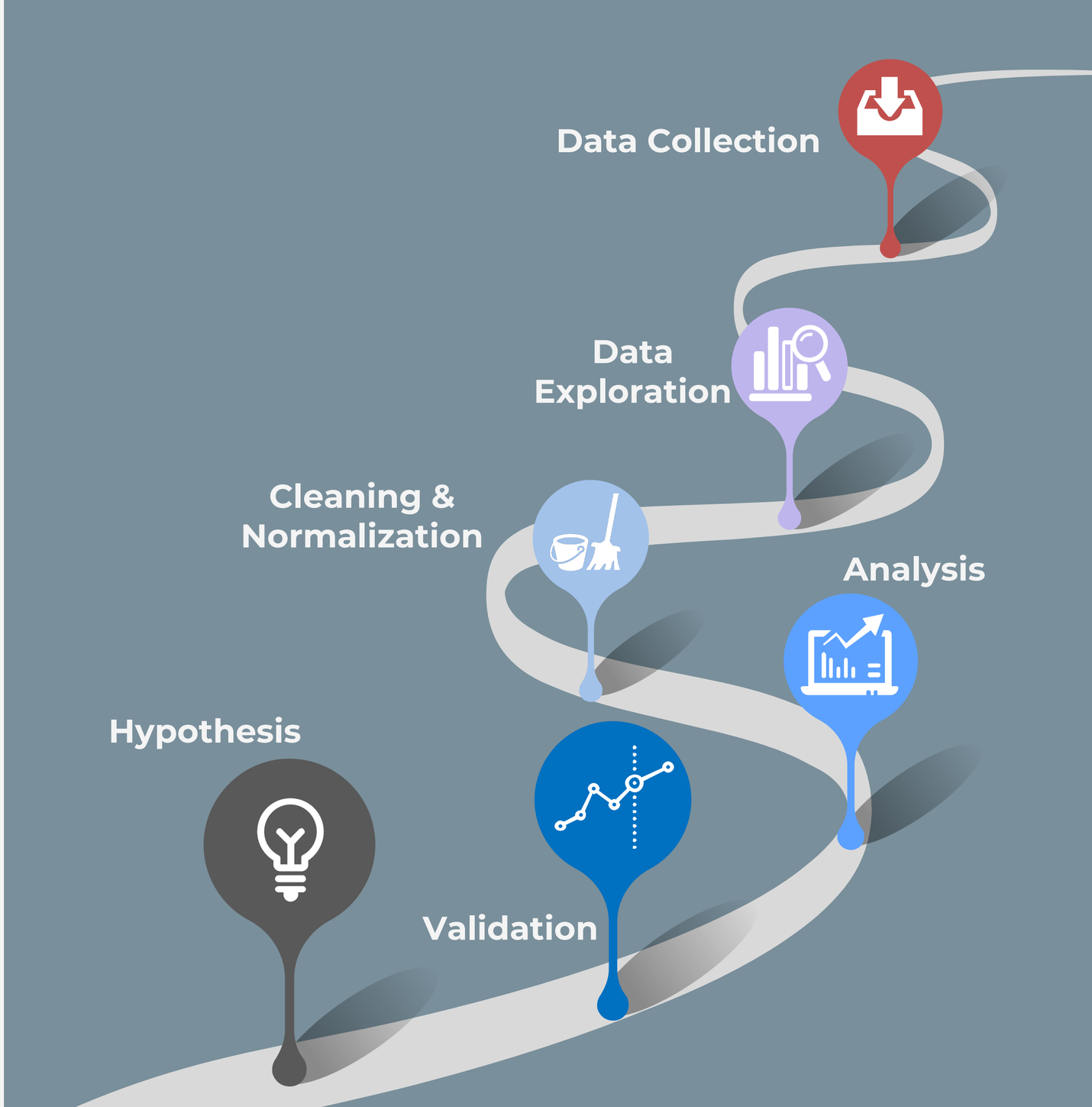
THE DATA'S JOURNEY



UCPH HeaDS

CONTINUING OUR JOURNEY

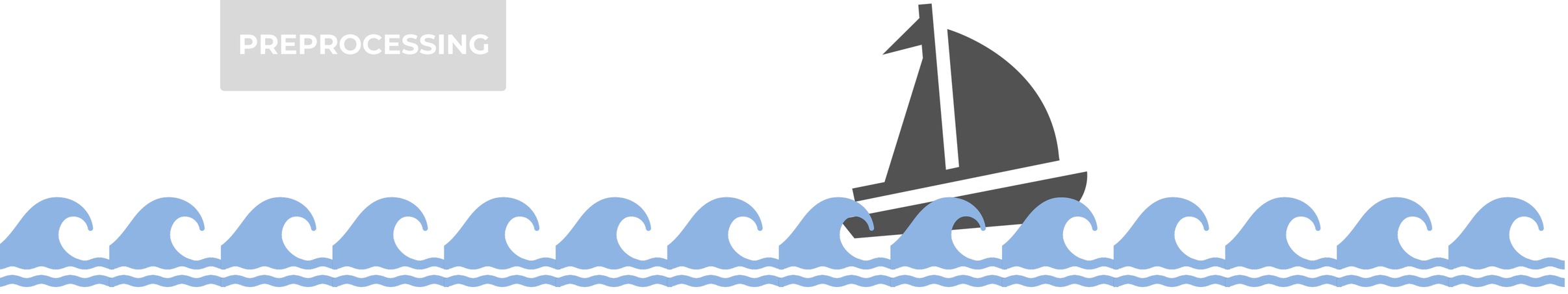
DATA COLLECTION

EXPLORATORY DATA ANALYSIS

DATA ANALYSIS

MODEL EVALUATION

PREPROCESSING

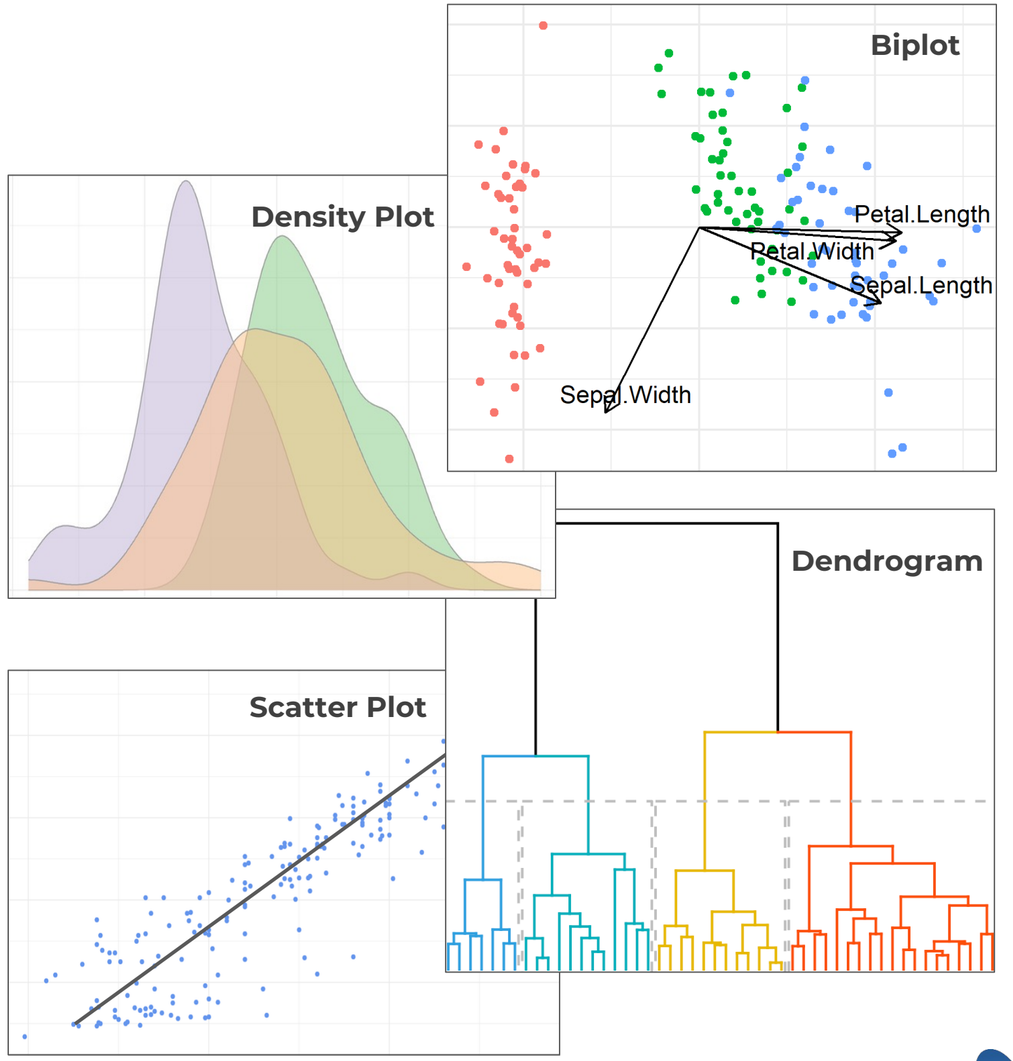


UCPH HeaDS

EXPLORATORY DATA ANALYSIS

Before we start our analysis we need to familiarize with the data.

- What data do we have?
- Does it look as you expected?
- Is it suitable for the planned analysis?
- How do we need to prepare the data for analysis?



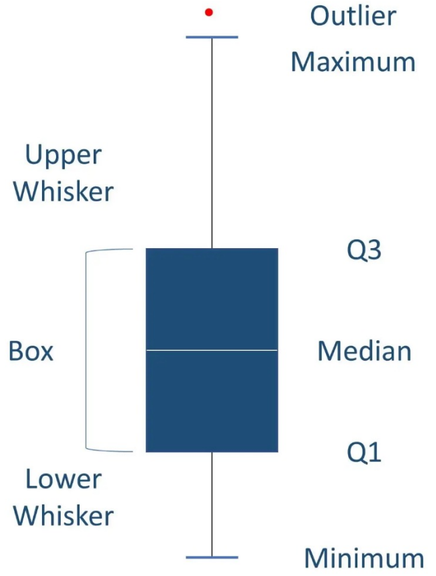
UCPH HeaDS

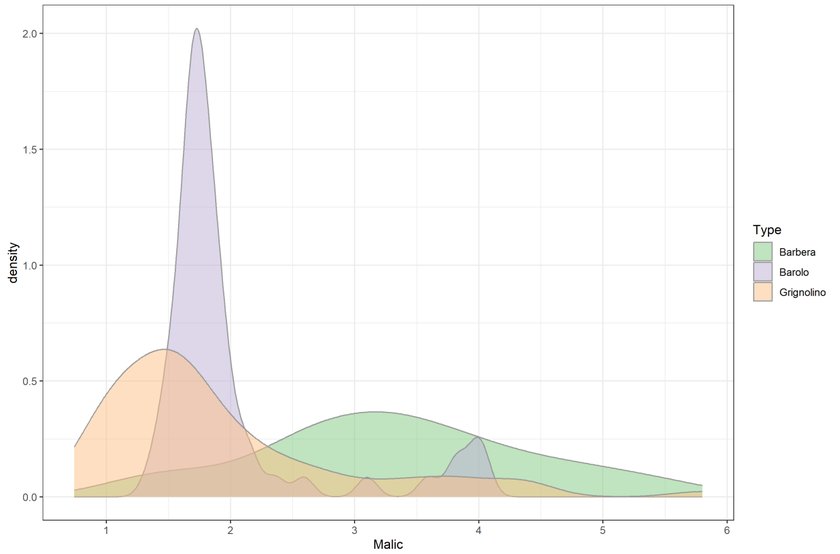
SUMMARY STATISTICS

Summary statistics are used to distill a dataset into some key characteristics.

They often include the following measures:

- Central tendency (mean, median, mode)
- Spread (standard deviation)
- Minimum and Maximum
- Quartiles
- Shape of the distribution
- Correlation between features



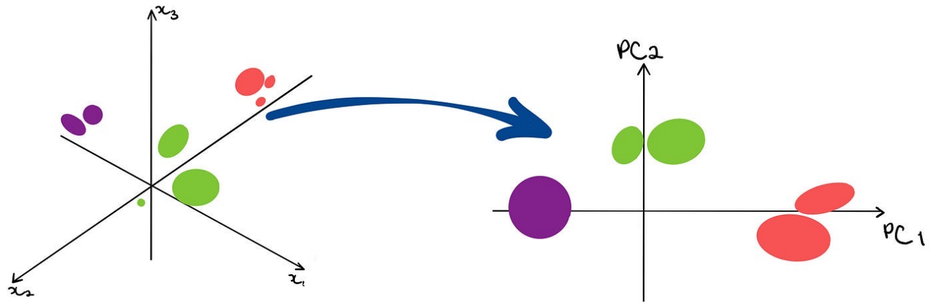


UCPH HeaDS

DOES THE DATA LOOK AS EXPECTED?

- Single variables may be checked by their **distribution**, (density plot) but what about **high-dimensional data**?

- One way to inspect the structure of your data is with a **Principal Component Analysis (PCA) Plot**



- PCA is a **dimensionality reduction technique**. It transforms data (linearly) from the original high dimensional space into a low (2 or 3) dimensional space, which the human eye can view and interpret.

UCPH HeaDS

PRINCIPAL COMPONENT ANALYSIS

Collection of wine bottles as an example. We have measured 13 different features such as alcohol content, color, alcalinity, and flavonoids.

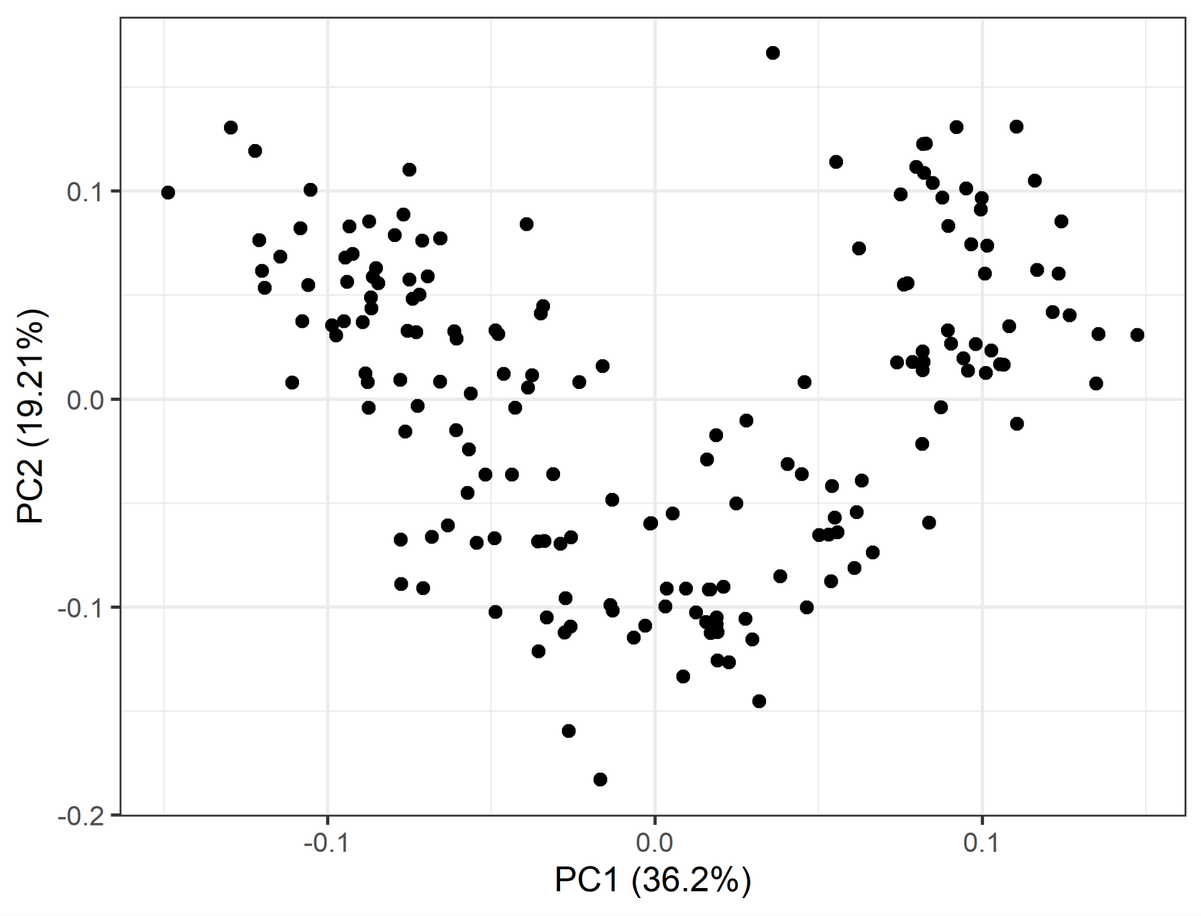
PCA lets us derive a set of new dimensions that best describe all the original wine features while also being much lower dimensional.

These derived dimensions are the **principal components** and they will help us understand the structure of our data.

UCPH HeaDS

PRINCIPAL COMPONENT ANALYSIS

We can now plot all wine bottles in our data by the values of their first two PCs:

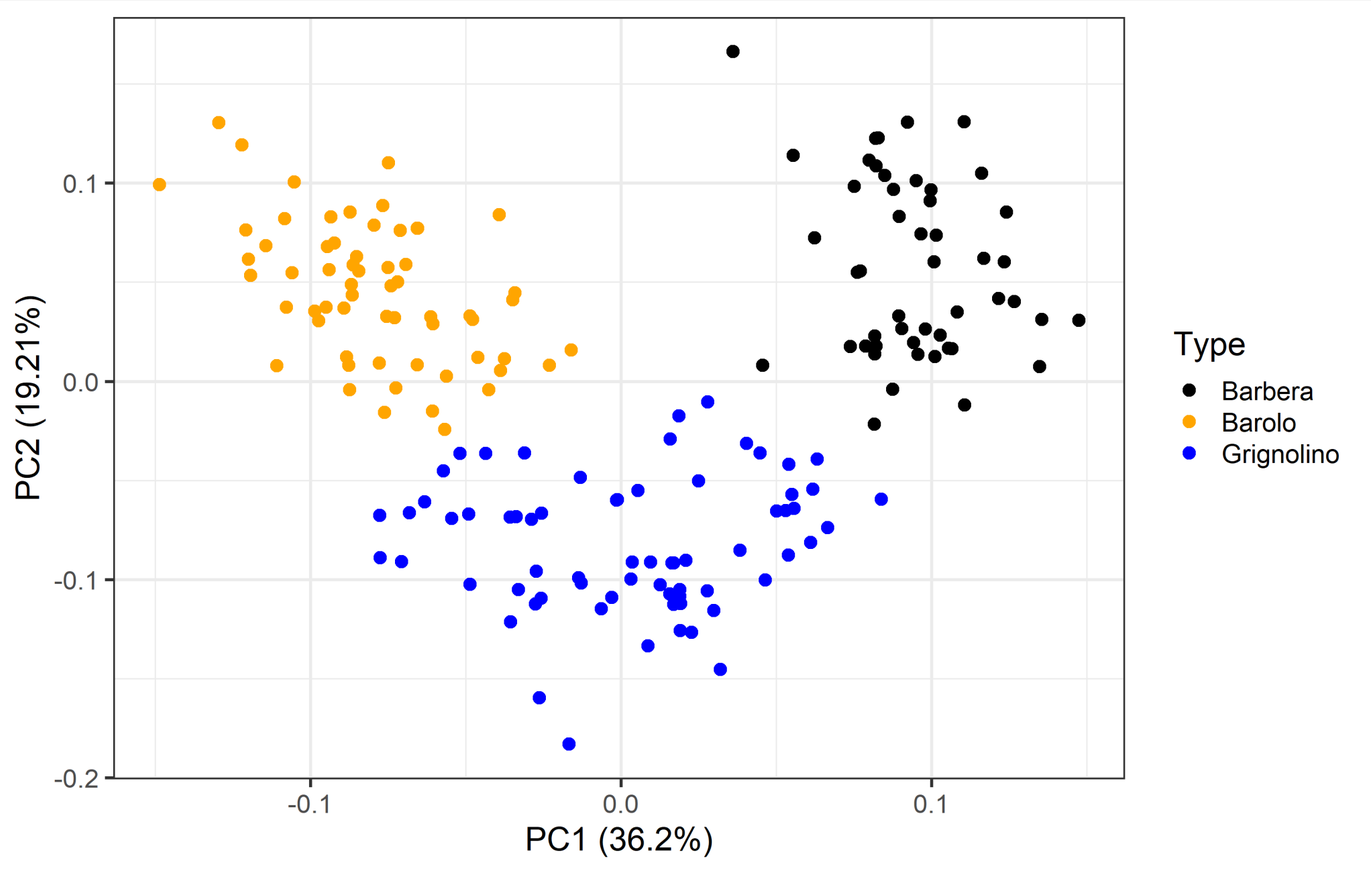


- Each dot is one wine bottle.
- If two wines are close in their PC1 and PC2, they are similar in the original feature space.
- This helps us to visually understand the structure of our data (we cannot plot in 13 dimensions!).

UCPH HeaDS

PRINCIPAL COMPONENT ANALYSIS

If two wines are close in PC1 and PC2 value they are **similar also in their original features**.

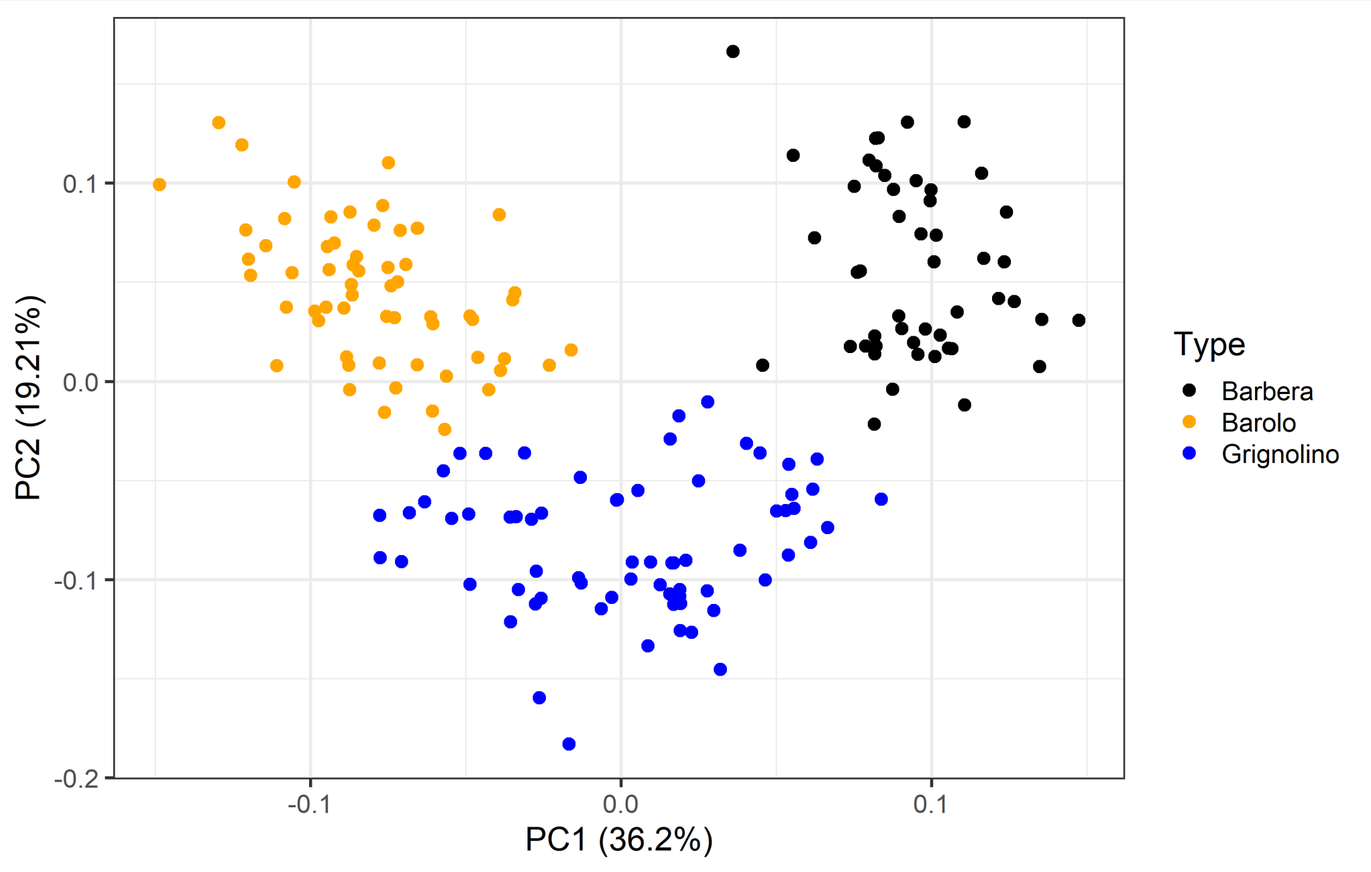


- Color by wine type
- Types of wine cluster together
- Most Barbera wines are similar to other Barbera wines even in only 2 dimensions.

UCPH HeaDS

PRINCIPAL COMPONENT ANALYSIS

PC's are **linear combinations** of the original features



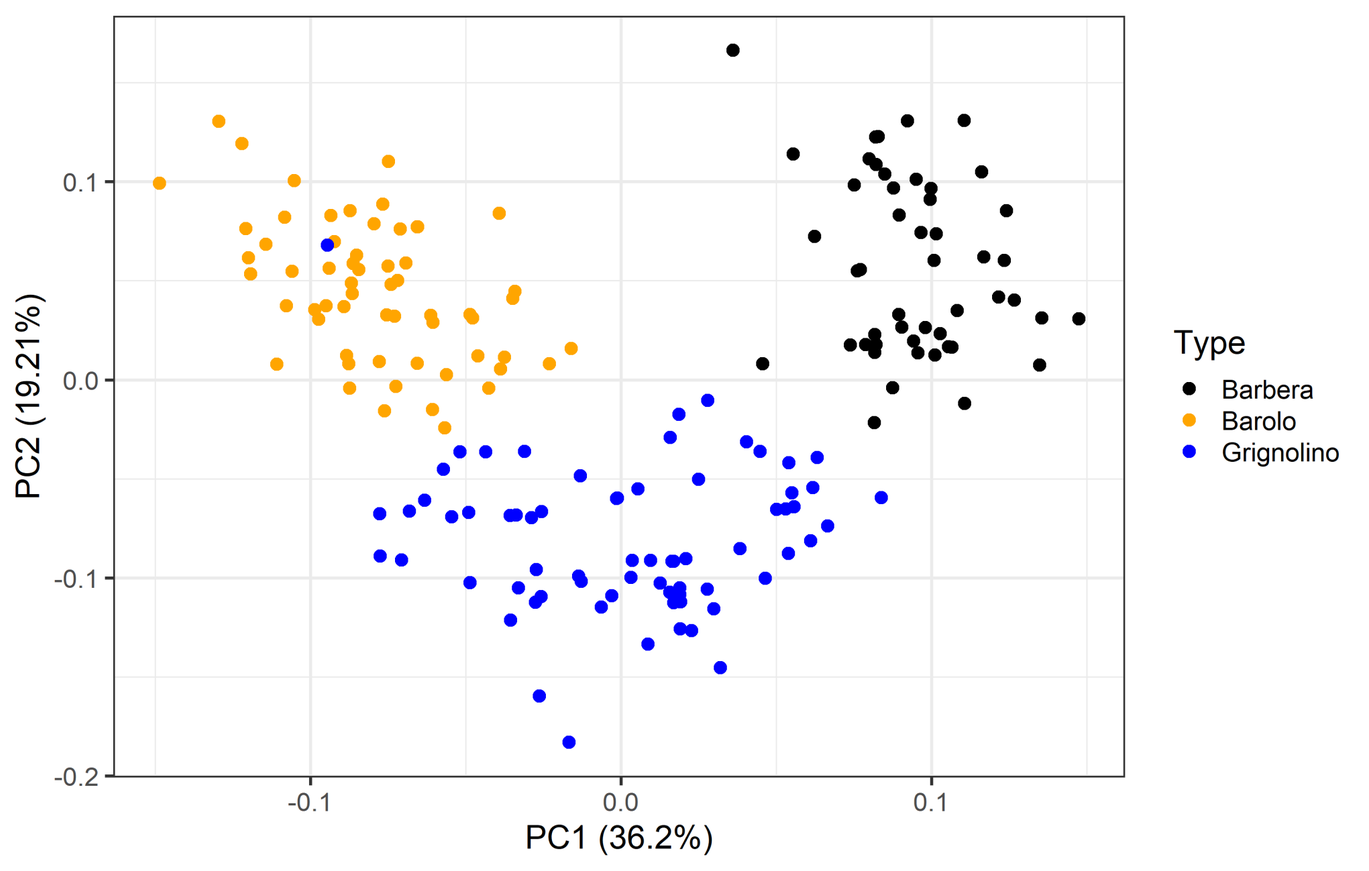
If two clusters are well separated in the PCA that means there is a combination of original features that **explains** which cluster a data point belongs to.

We can use this to predict the cluster for new data points.

UCPH HeaDS

PRINCIPAL COMPONENT ANALYSIS

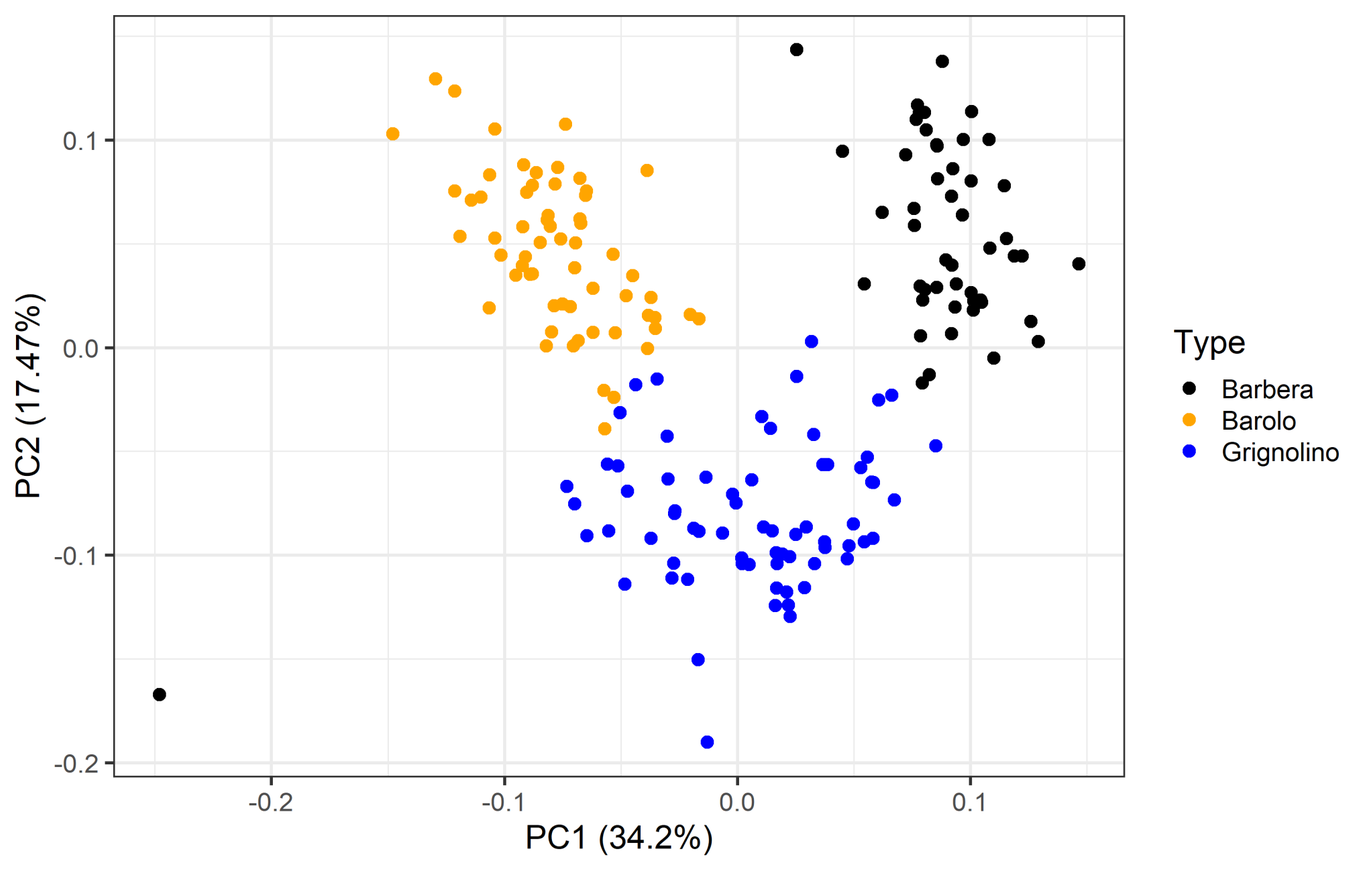
What do you think is going on in this PCA?



UCPH HeaDS

PRINCIPAL COMPONENT ANALYSIS

What do you think is going on in this PCA?



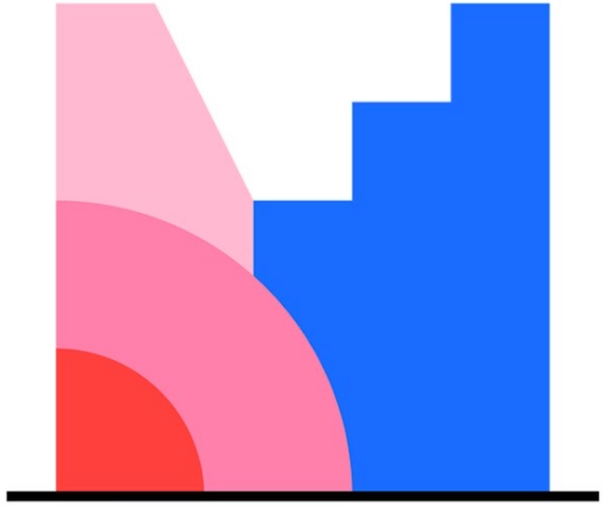
UCPH HeaDS

PRINCIPAL COMPONENT ANALYSIS

In summary:

- PCA is a ***linear*** **dimensionality reduction technique**
- We can see if there are clusters in the data
- We can see whether these clusters correspond to a variable we are interested in
- We can see if data points are outliers
- We can see if data point behave strangely, i.e. if they are in the wrong cluster.

OBS: PCA works only on numeric data!



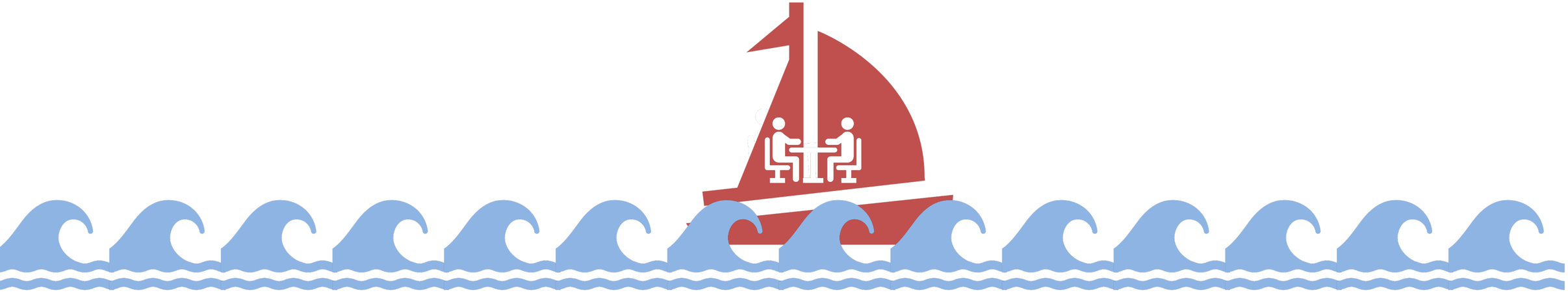
Mentimeter

UCPH
HeaDS

GROUP DISCUSSION 2.1

In your group discuss the printed PCA plots.

- What can you see?
- What do you think it means?



UCPH HeaDS

When things go wrong...

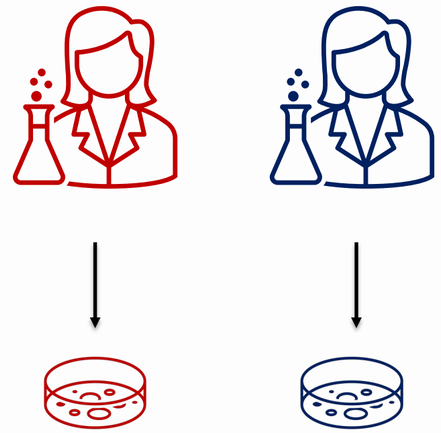
UCPH HeaDS

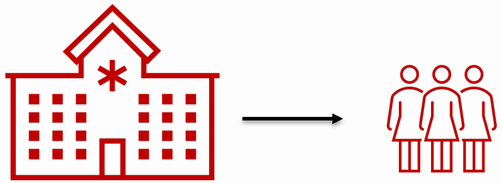
BATCH EFFECTS

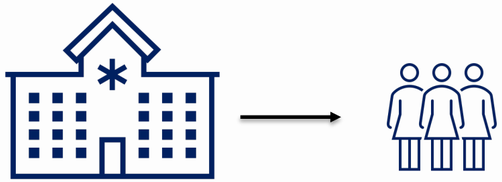
Batch effect == unwanted variation introduced by **technical procedures,** i.e. collection, handling, storage, or experimental protocol.

Batch effects can be corrected during analysis, if they are **not** *fully/mainly* correlated with the outcome.

If the batch effect **is fully correlated** with the outcome you now have a **confounded** dataset (cannot be fixed!).

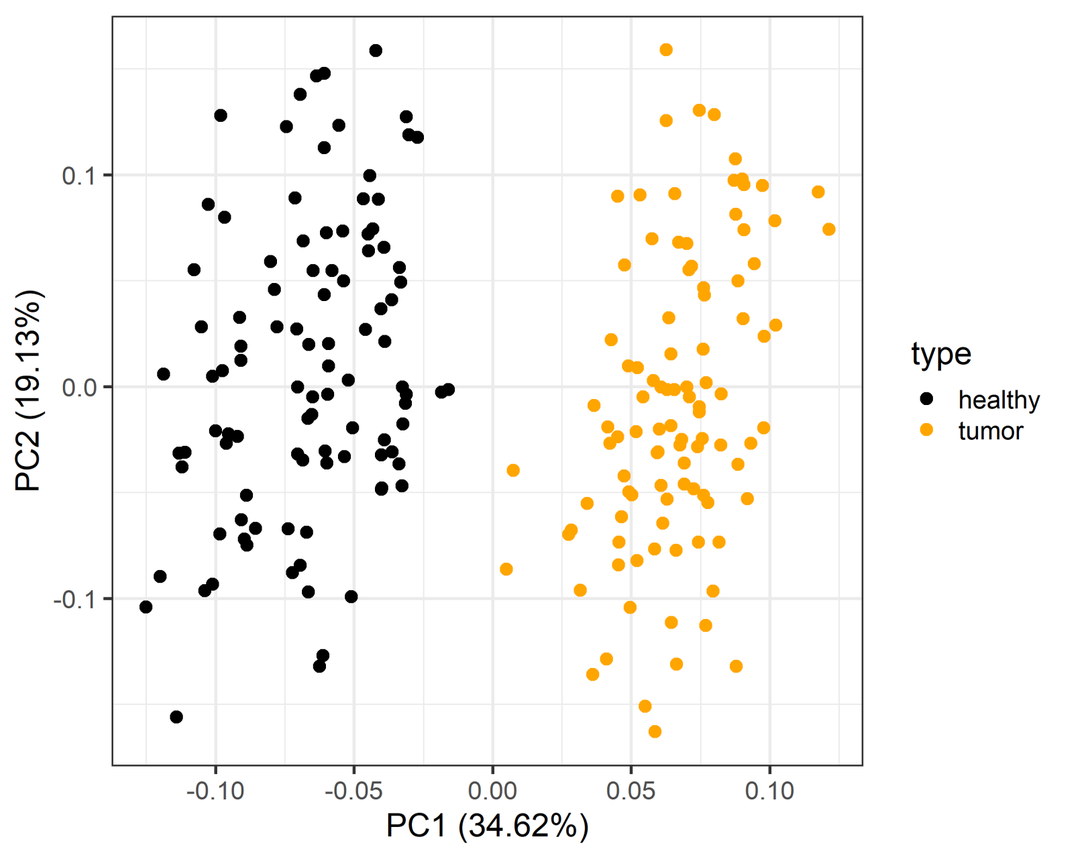






BATCH EFFECTS ON PCA PLOT

PCA plots can be used to investigate whether batch effects are present

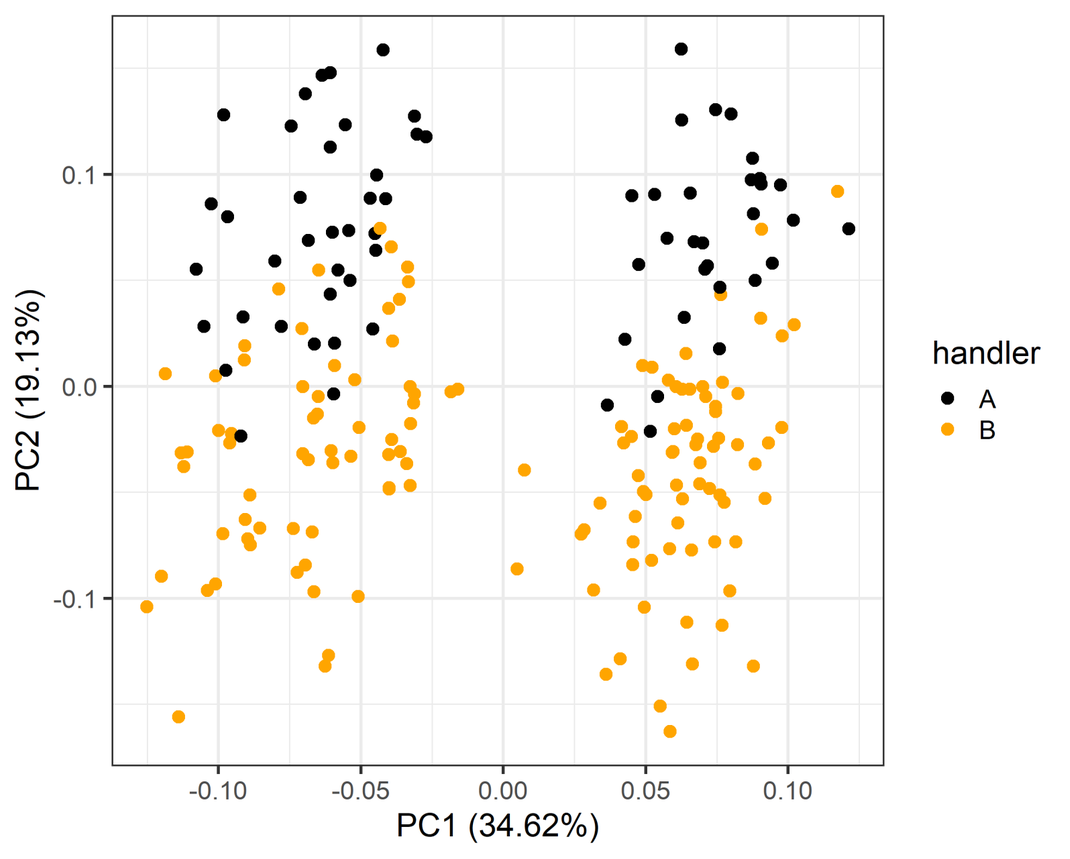


PCA of gene expression data from healthy and tumor samples.

Separation of samples depending on the tissue type along PC1 (explains ~ 35% of the variation in our dataset).

UCPH HeaDS

BATCH EFFECTS ON PCA PLOT



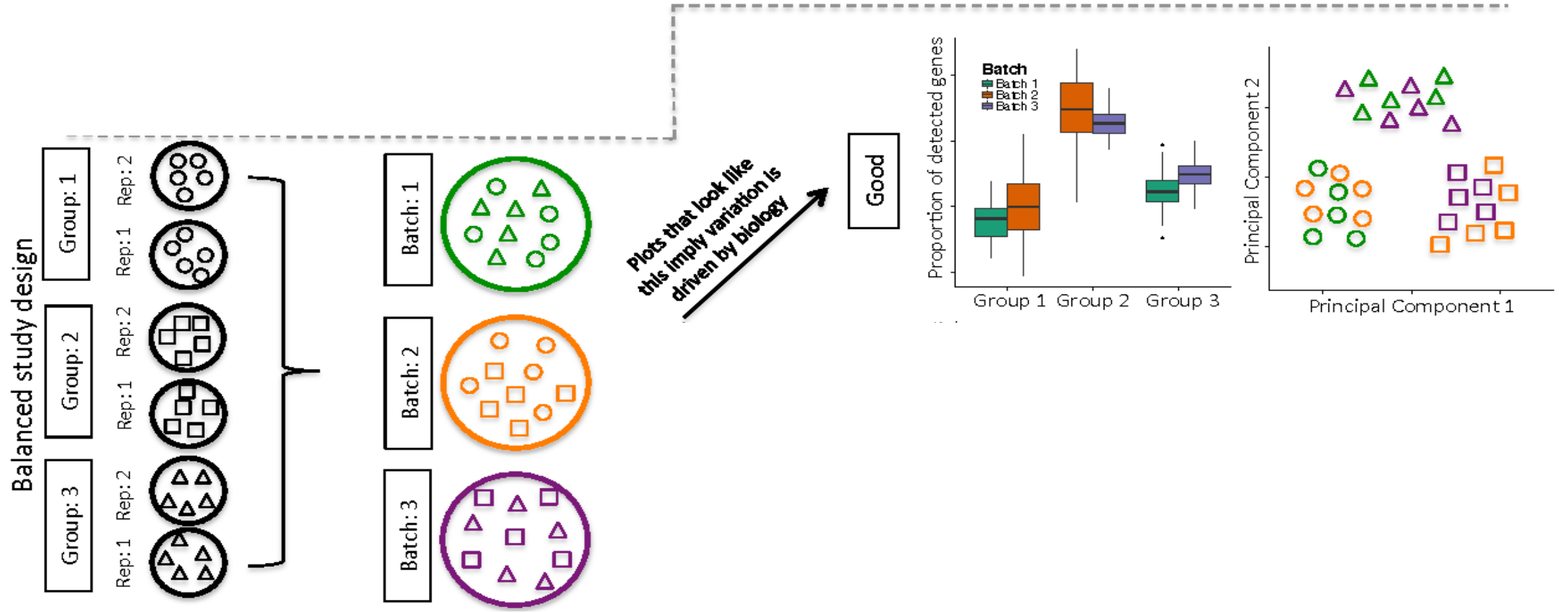
PC2, **not related** to the tissue type.

Color by who has processed the samples we can see this is a **processing batch effect**.

IF *you had one person process all tumor samples and another person all the healthy samples, how would this plot look?*

UCPH HeaDS

BACK TO EXPERIMENTAL DESIGN



Hicks, et al. *BioRxiv* 10 (2015): 025528.

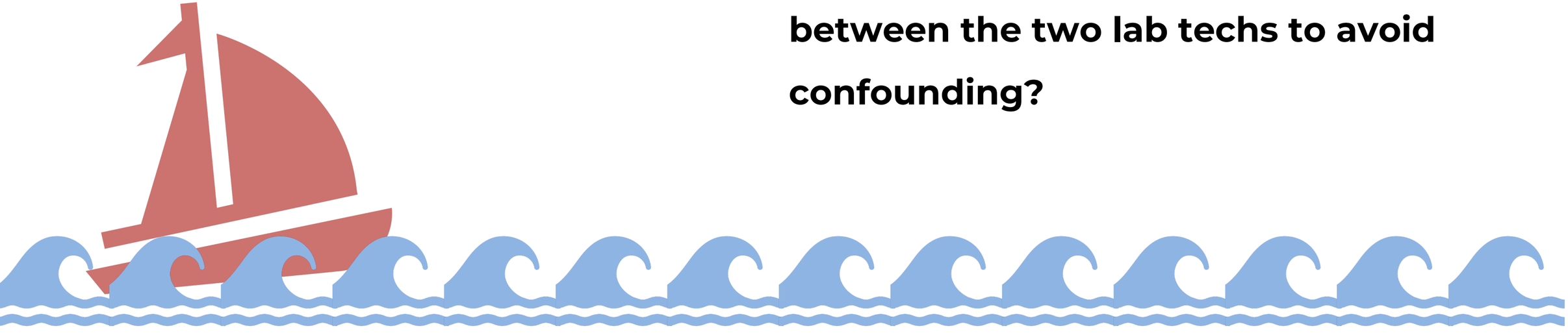
UCPH HeaDS

Mentimeter

Remember our study of gene expression in tumor and healthy tissue from part 1.

Since you have a lot of samples you ask two lab techs to each process half.

How would you split the samples between the two lab techs to avoid confounding?



UCPH HeaDS

OTHER CHECKS

There are many things one should check before DS analysis. Here we mention a just few:

- **Check the data:**
  - **Unreasonable values:** If a data point is “out of scale” you should be able to see that in the PCA plot (if you have few features you could also check their range).

  - **Unreasonable combinations:** Sometimes each variable is fine by itself but they are combined in a way that strike you as odd, i.e. children who are also former smokers

- **Check model/test assumptions:**
  - Most tests have assumptions, such as that data is normally distributed and that the variance is homogenous between groups (iid) - more on this later…

UCPH HeaDS

THE ROLE OF DOMAIN EXPERTS

Domain knowledge plays a big role in EDA

The person doing the data processing may not know what values are reasonable for a given variable or which combinations of variables look suspicious.

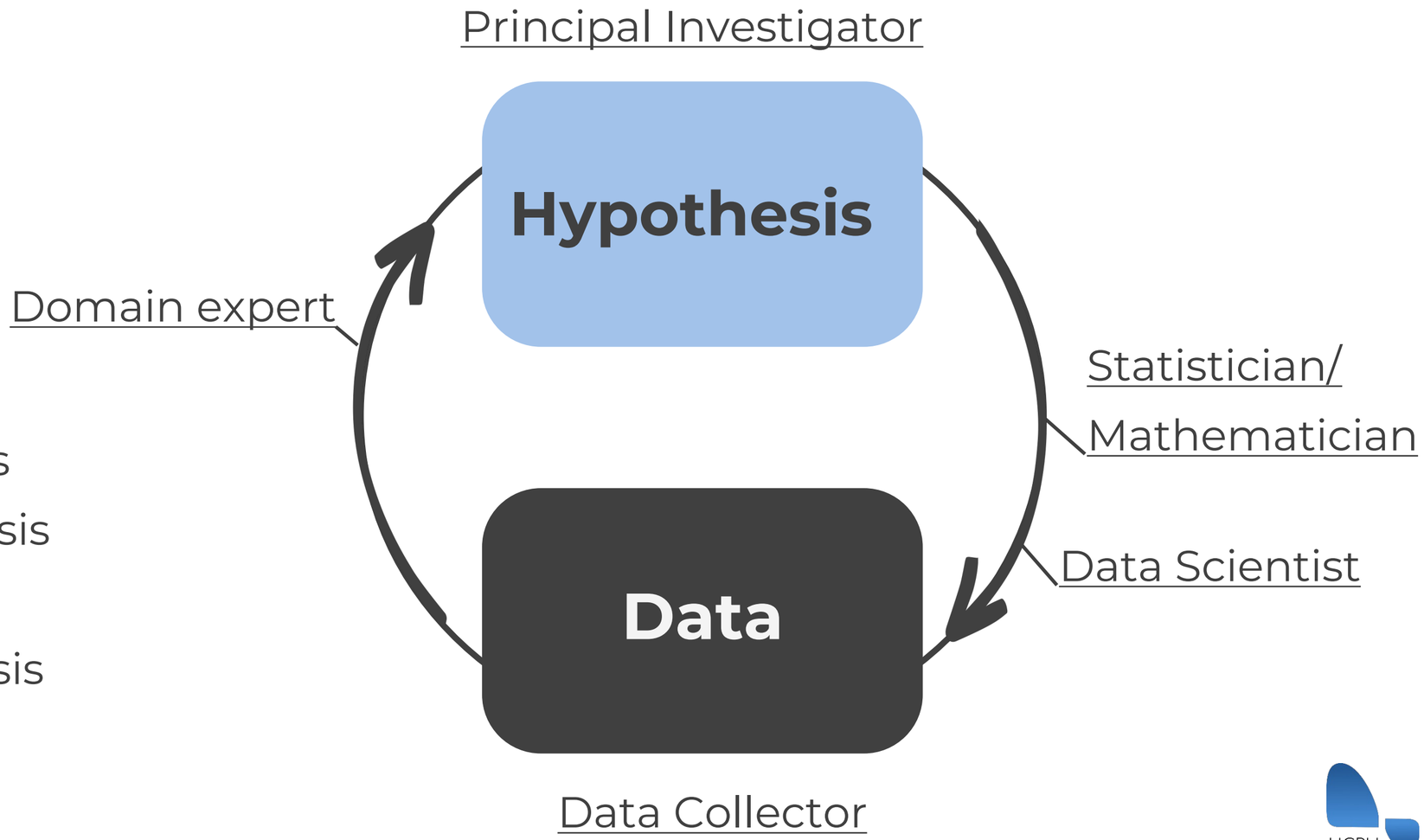
It is a good idea to have a person who understands the data check plots, ask questions and suggest things to test.

BACK TO THE ROLES

- **Idea generation:**
  - Patterns
  - Relationship between variables

- Confirm whether the data is suitable to analyze hypothesis

- Perform data science analysis



UCPH HeaDS

EXPLORATORY DATA ANALYSIS

If you are unfamiliar with some data types and/or analysis **we recommend you to:**

- Consult us in the Data Lab for a sparring/discussion session
- Consult the literature, specifically review papers. Find out what is known to work and what other people do
- Consult colleagues and collaborators
- Take a course/send a group member to a course

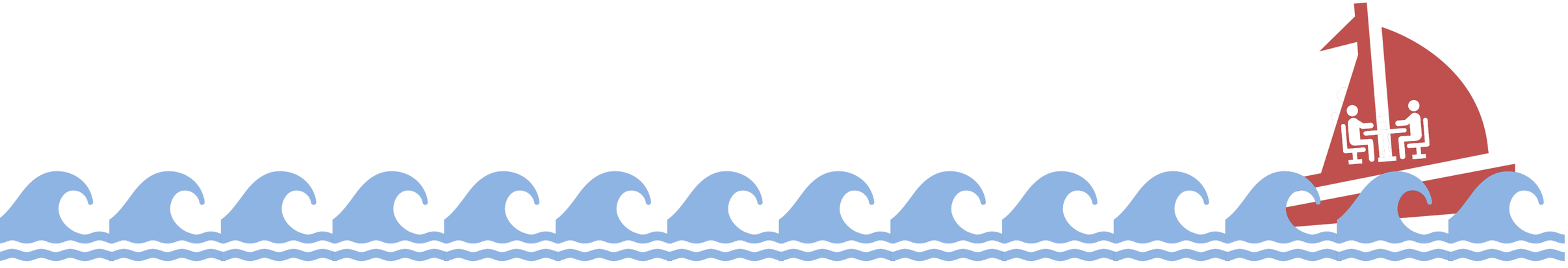
UCPH HeaDS

GROUP DISCUSSION 2.2

In your group discuss:

Boxplot:

- Does the plot display a pattern worth noting. If so, what is the cause of it?
- Are the data confounded?
- Are there any outliers? If so, do you have any theory as to what gave rise to them (i.e. biological or technical reason?).



UCPH HeaDS

DETOUR ON OUR JOURNEY

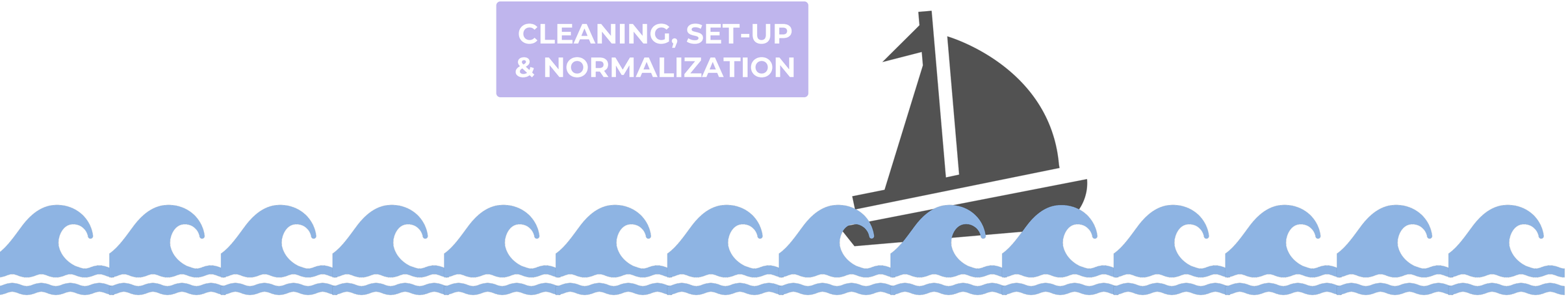
DATA COLLECTION

EXPLORATORY DATA ANALYSIS

CLEANING, SET-UP & NORMALIZATION

DATA ANALYSIS

MODEL EVALUATION



UCPH HeaDS

DATA CLEANING AND SET-UP

We want to clean up our data:

- Remove data outliers or duplicates
- Ensure categorical variables are spelled the same
- Change one type of variable to another type
- Merge, add or remove variables
- Layout, long VS wide format

Missing values:

- Filter out some/all missing values
- Impute missing values (Regression, Maximum Likelihood, ML)

Wide w. missing values

Treatment	Age	height	weight
A	55.0	**NA**	65.7
B	31.0	172.0	69.4
C	39.0	161.0	**NA**

Imputation

Treatment	Age	height	weight
A	55.0	170.0	65.7
B	31.0	172.0	69.4
C	39.0	161.0	58.1

Wide to long

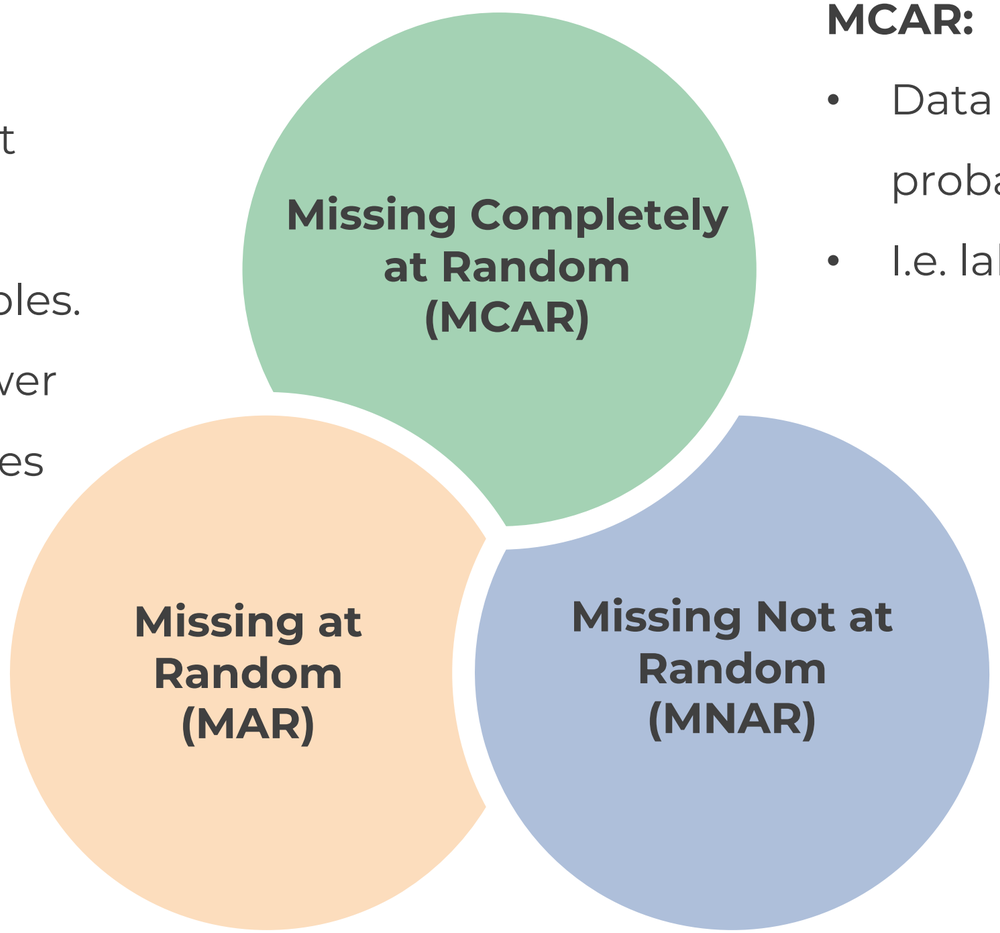
Treatment	variable	value
A	Age	55.0
A	height	170.0
A	weight	65.7
B	Age	31.0
B	height	172.0
B	weight	69.4
C	Age	39.0
C	height	161.0
C	weight	58.1

UCPH HeaDS

TYPES OF MISSING VALUES

MAR:

- Missing data point is not related to missing data itself, but to other variables.
- I.e. Health records in lower median income countries



MCAR:

- Data points have the same probability of being missing
- I.e. lab tool error

MNAR:

- Missing data points are related to the reason why data are missing.
- I.e. Assessing depression through questionnaires

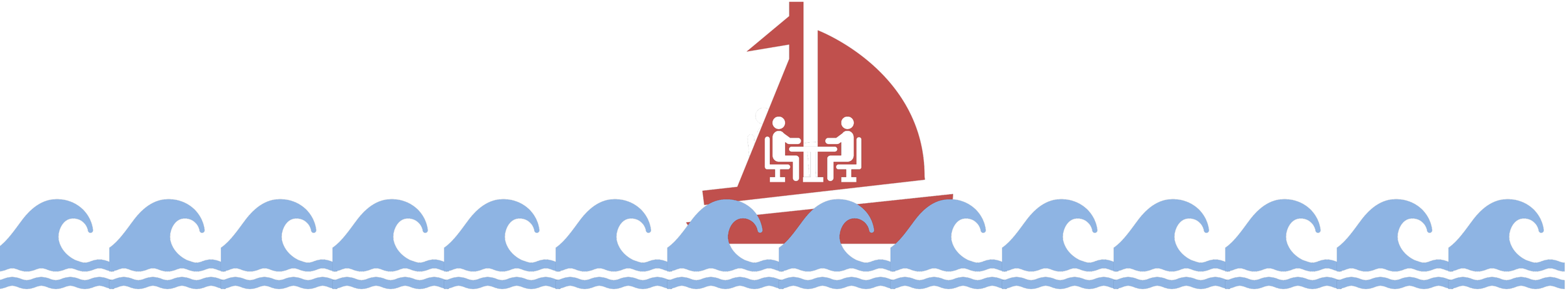
UCPH HeaDS

GROUP DISCUSSION 2.3

In your group discuss the **data table** we have handed out:

- Identify the different data **types** it contains (categorical, numerical, integer, binary, factors).

- Can you find any **errors/problems** within the data table which would have to be fixed before data analysis ?



UCPH HeaDS

DATA NORMALIZATION

In applied statistics the word '**normalization'** can have a range of meanings. Here we're going to use the following definition:

> Normalization is a process intended to reduce unwanted variation and make samples more easily comparable

Unwanted variation should be low between samples within a group (i.e. all the tumor samples) as we are interested in differences between groups (tumor VS healthy).

UCPH HeaDS

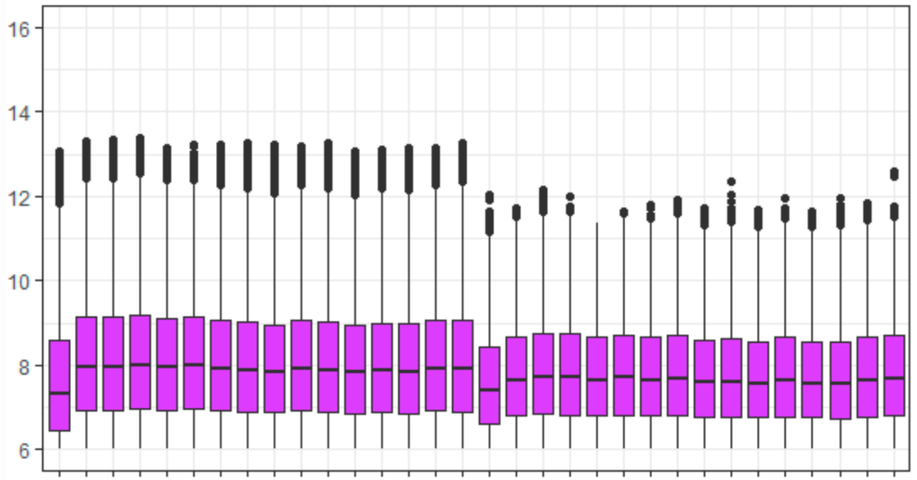
UNWANTED VARIATION

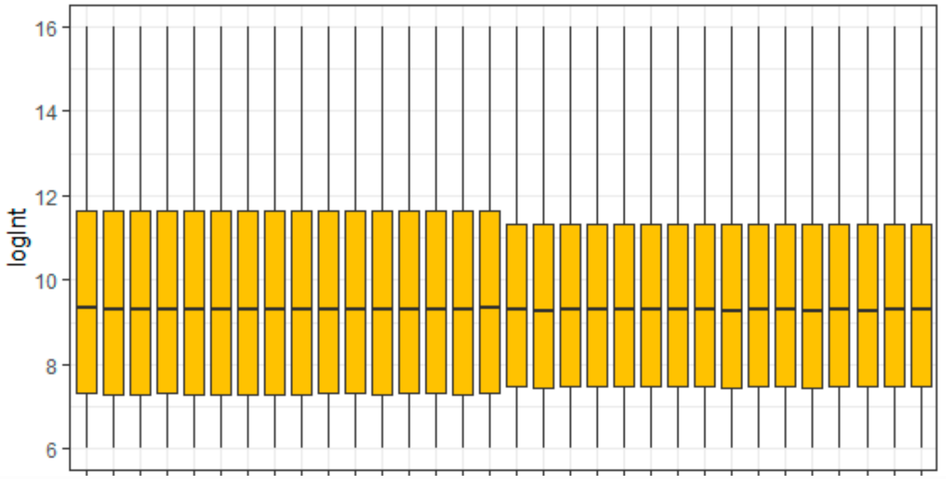
Technical variation:

Introduced by i.e. sample handling, data batches, device calibrations, ect.

Non-technical variation not related to the outcome:

i.e. gene length and library size (number of reads) in RNAseq



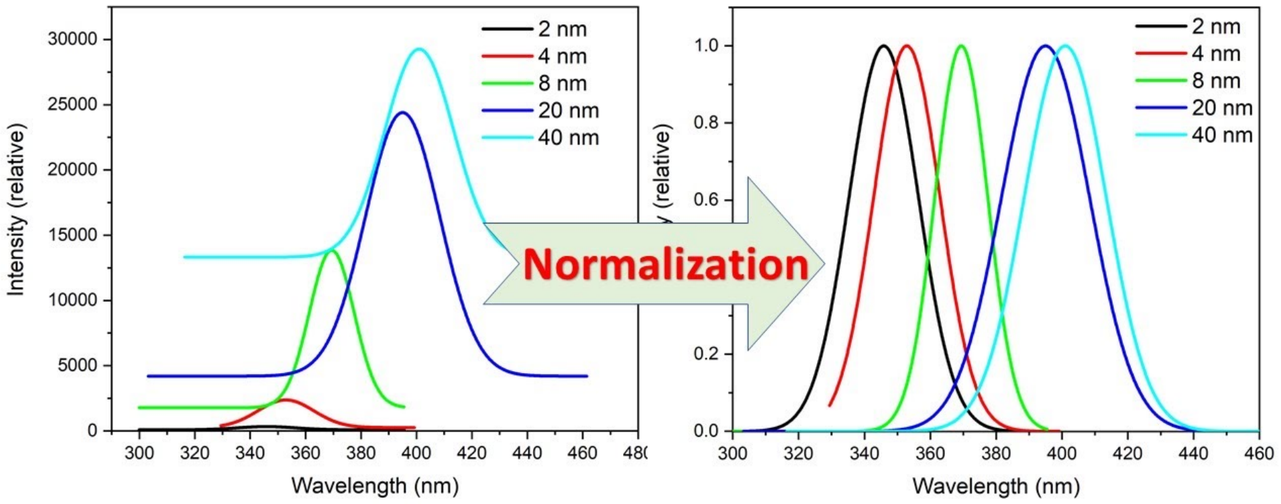


UCPH HeaDS

DATA NORMALIZATION

Normalization is not a trivial task!

- Different data types have different suitable normalization procedures.
- Sometimes one type of data/experiment can be normalized in multiple ways
- New and improved normalization procedures are introduced regularly



Important that one uses the same normalization with same parameters for the entire dataset.

Often, we need to consult literature and/or an expert.

UCPH HeaDS

STANDARDIZATION & TRANSFORMATION

NORMALIZATION

- Can use for not normally distributed data.
- Variables do not get zero centered.
- Normalization within a range (max, min).
- Affected by outliers.

STANDARDIZATION

- For normally distributed data
- Scaled to a mean of 0 and standard dev. of 1.
- Scaling is not con-strained to a particular data range.
- Not affected by outliers.

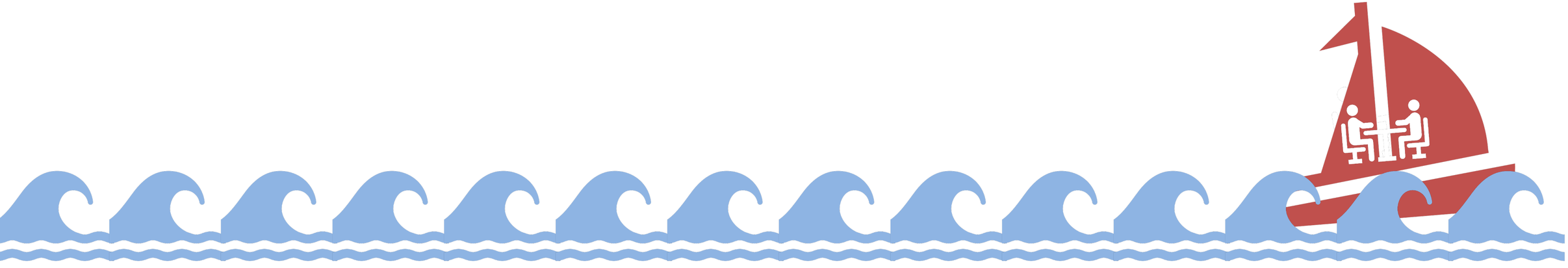
TRANSFORMATION

- Attempt to make data normally distributed (required for some tasks)
- Often logarithmic transformations
- Squeezes outliers for less impact on model.

UCPH HeaDS

GROUP DISCUSSION 2.4

Thinking of the data that **you (or your students) work with**, what are potential sources of unwanted variance, technical or non-technical?



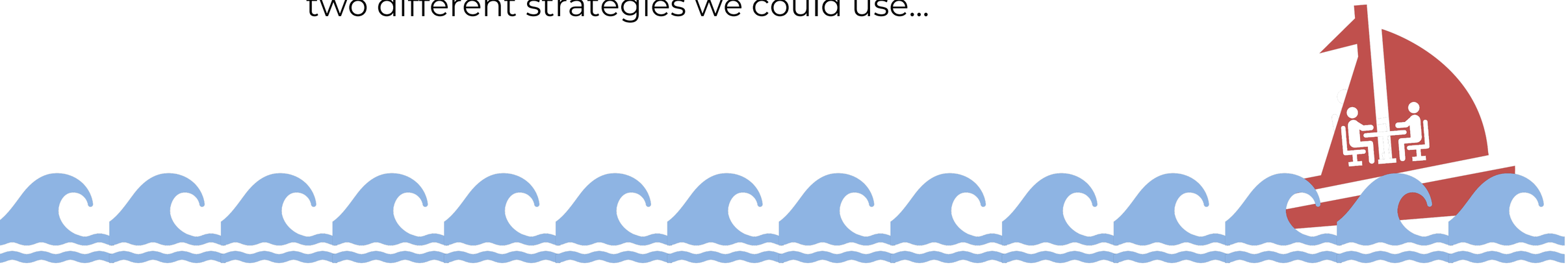
UCPH HeaDS

GROUP DISCUSSION 2.5

In your group discuss:

The density plot:

- What does the plot tell you about the distribution of gene counts.
- Are the data normally distributed? Why do we often like our data to be normally distributed?
- If data are not normally distributed what could we do? There are two different strategies we could use...



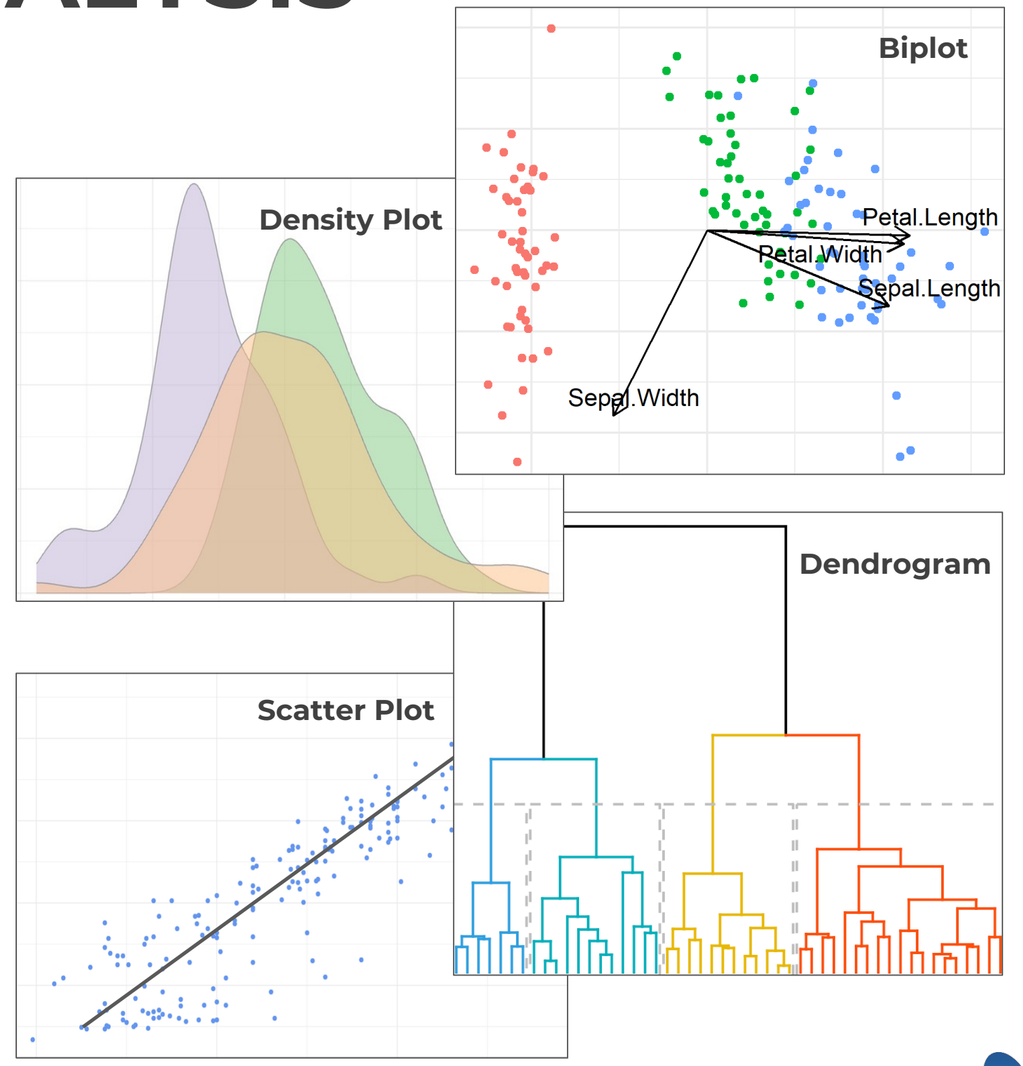
UCPH HeaDS

EXPLORATORY DATA ANALYSIS

This was a quick intro to EDA.

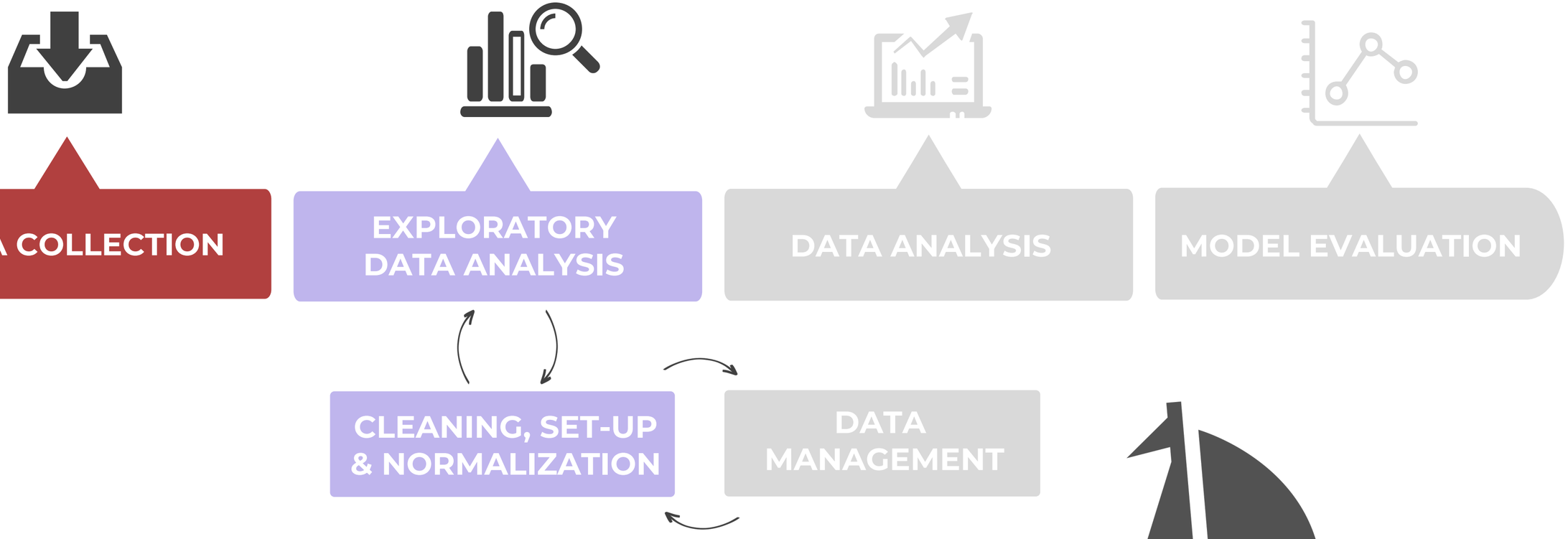
In summary it helps us to:

- Do the data look as expected
- Identify obvious errors: outliers, label swaps, mislabeling, etc.
- Constraints: Missing values, power
- How should the data be prepared for analysis – setup, clean
- Normalization, standardization, transformation



UCPH HeaDS

DETOUR ON OUR JOURNEY



UCPH HeaDS