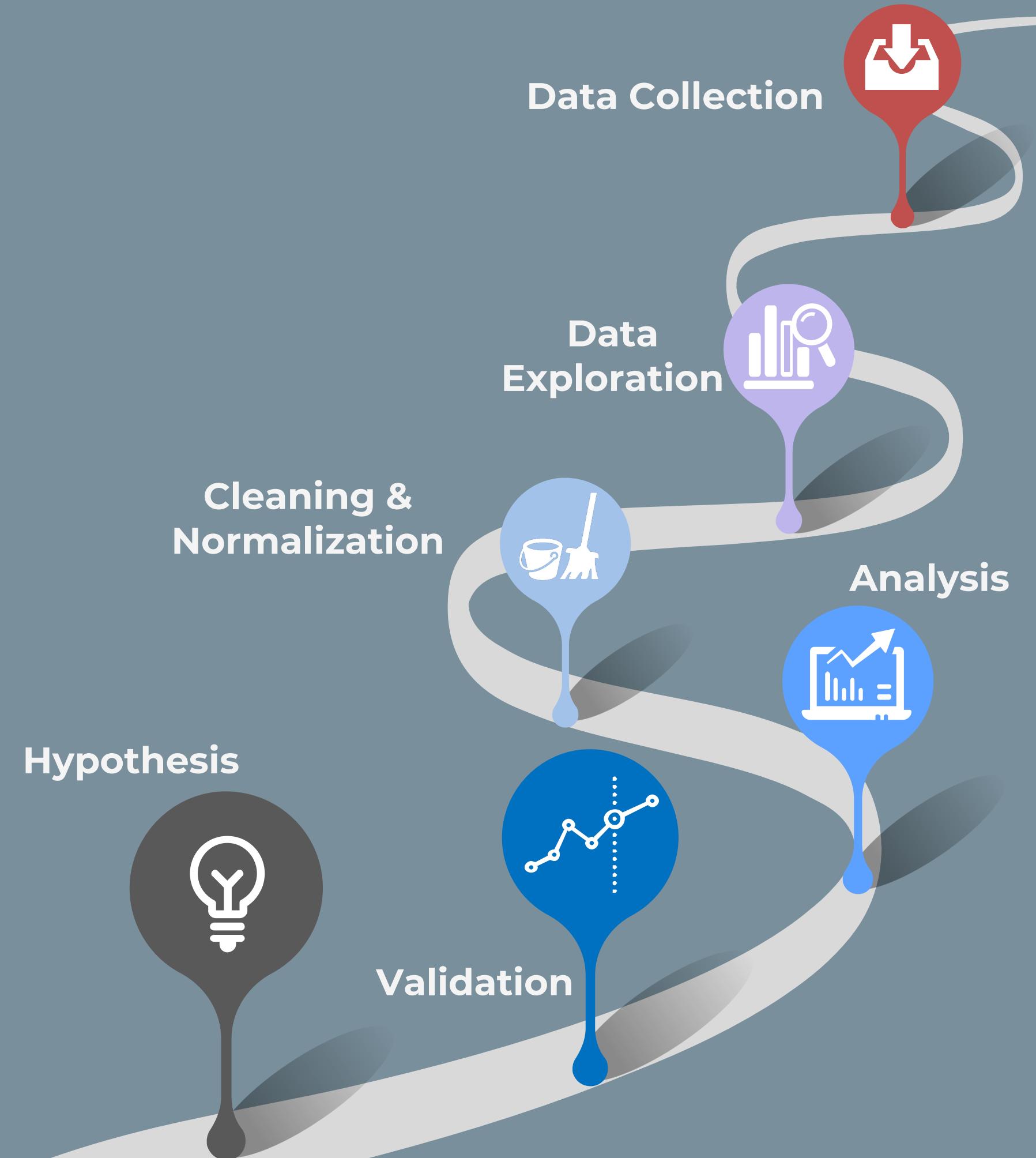
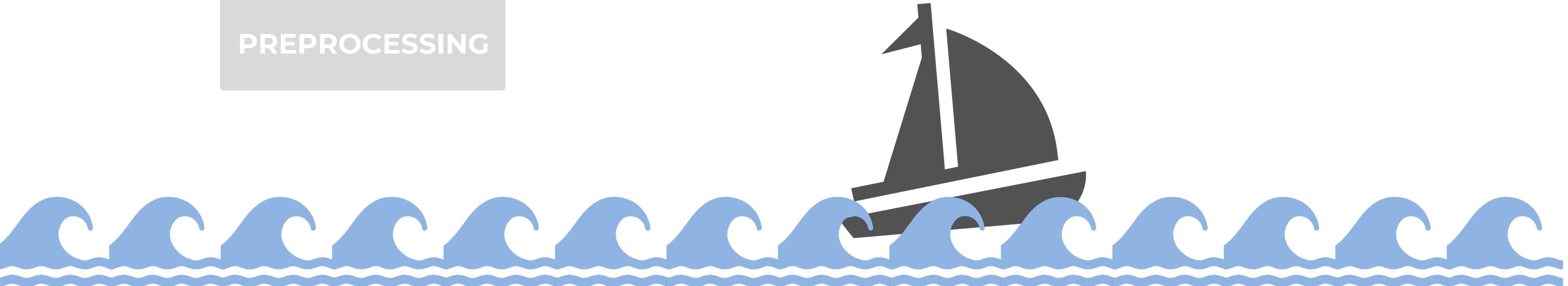
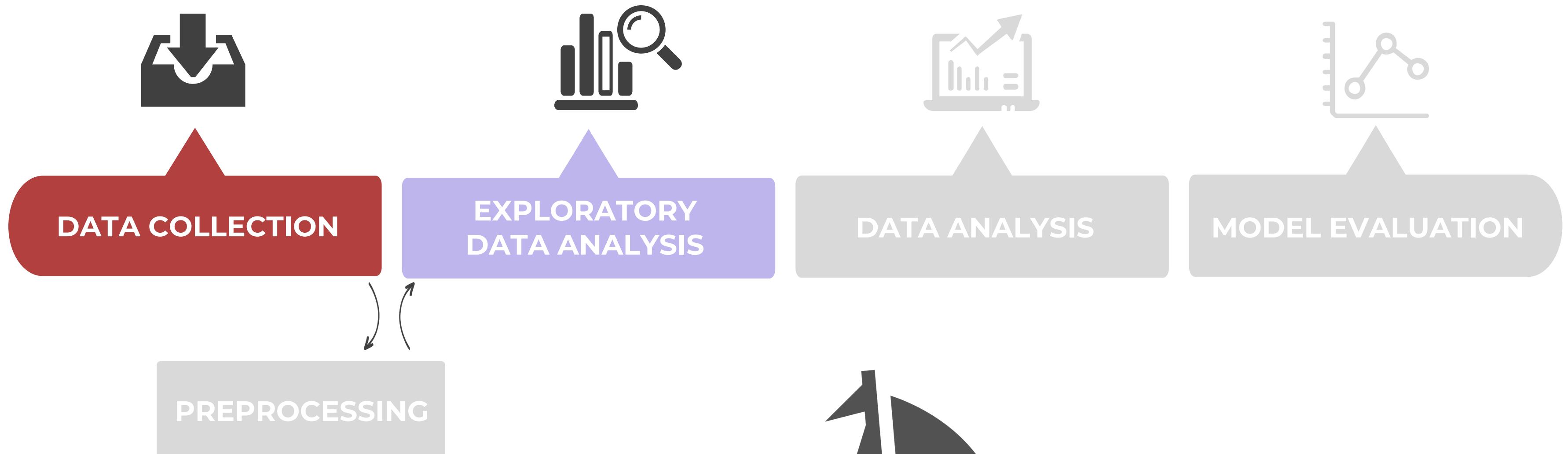


THE DATA'S JOURNEY



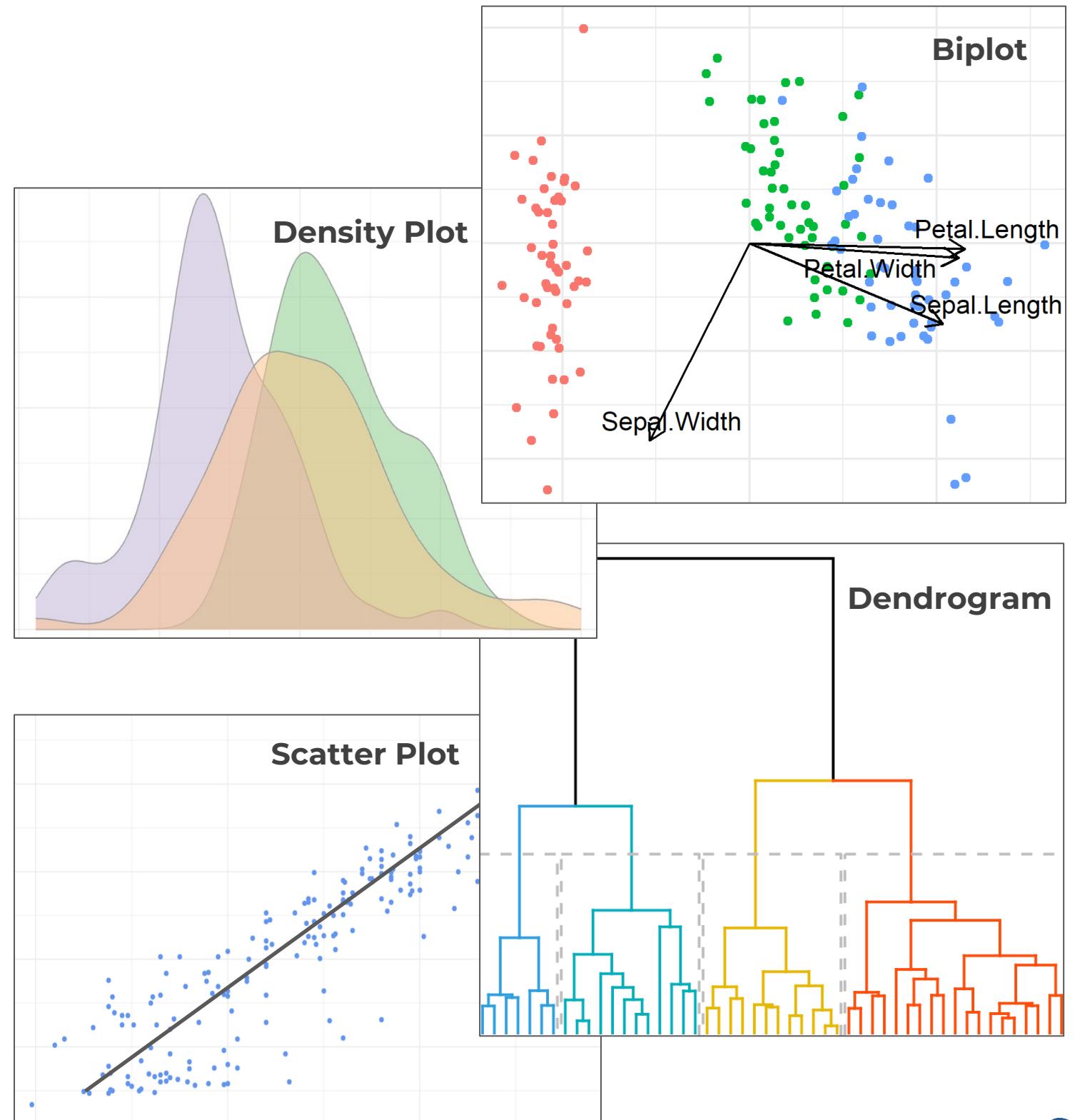
CONTINUING OUR JOURNEY



EXPLORATORY DATA ANALYSIS

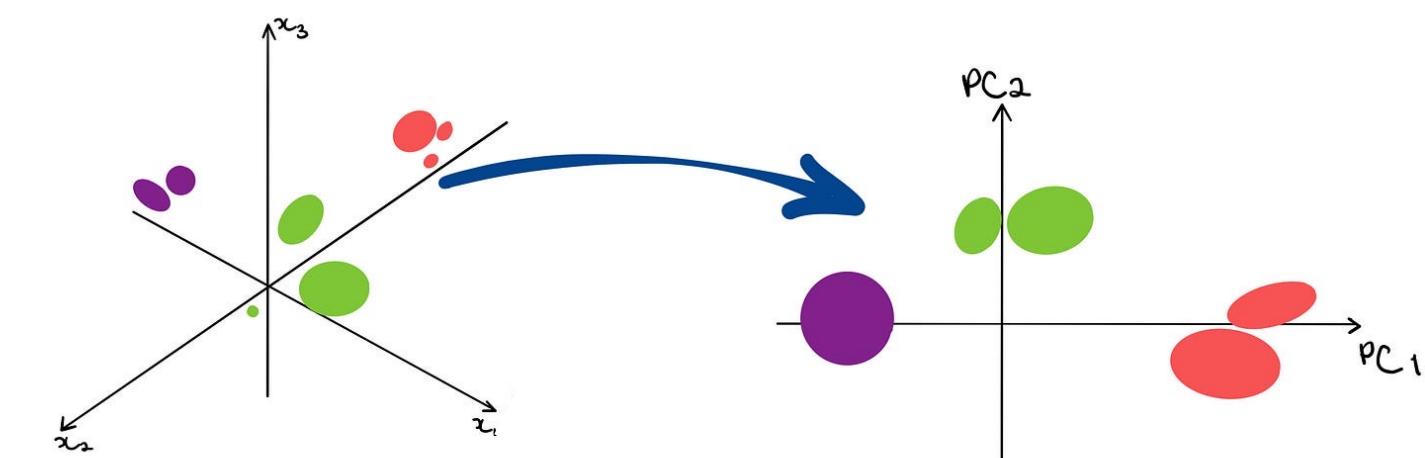
Before we start to do our analysis we need to get familiar with the data.

- 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?



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.



PRINCIPAL COMPONENT ANALYSIS



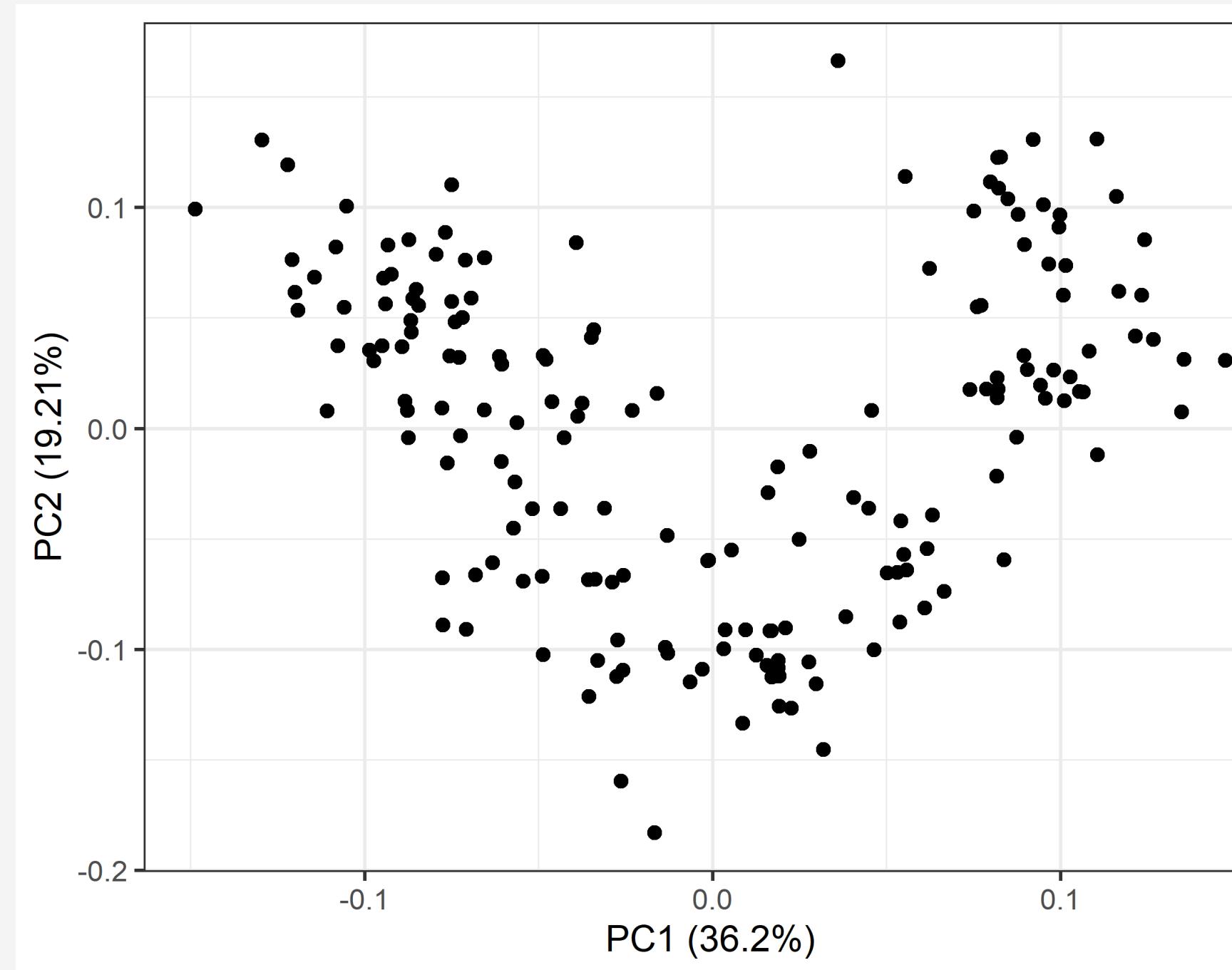
Consider a collection of wine bottles as an example. We have measured 13 different features such as alcohol content, color, alkalinity, and flavonoids.

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.

PRINCIPAL COMPONENT ANALYSIS

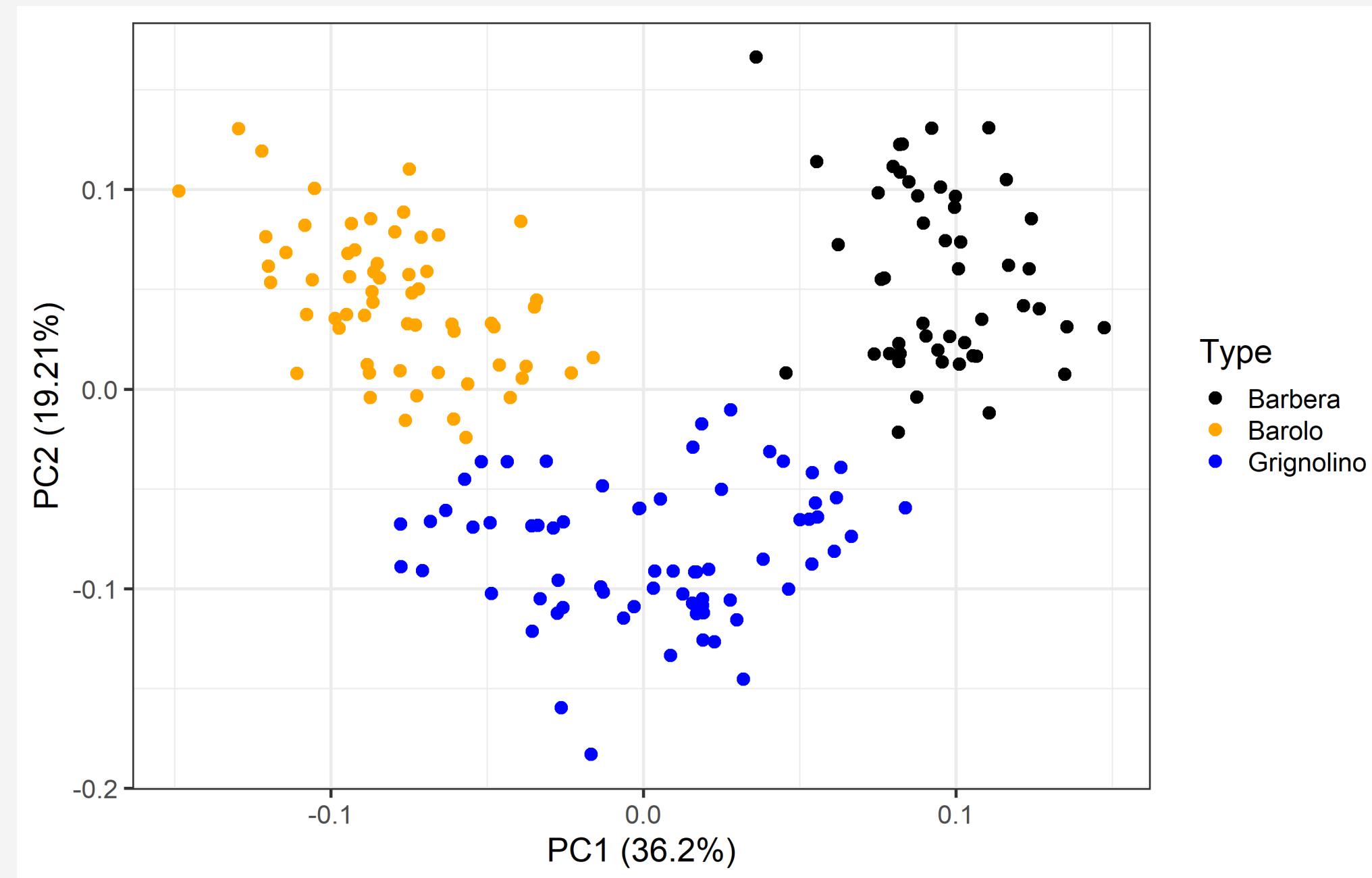
We can now plot all wine bottles in our data by the values of their first two PCs:



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

PRINCIPAL COMPONENT ANALYSIS

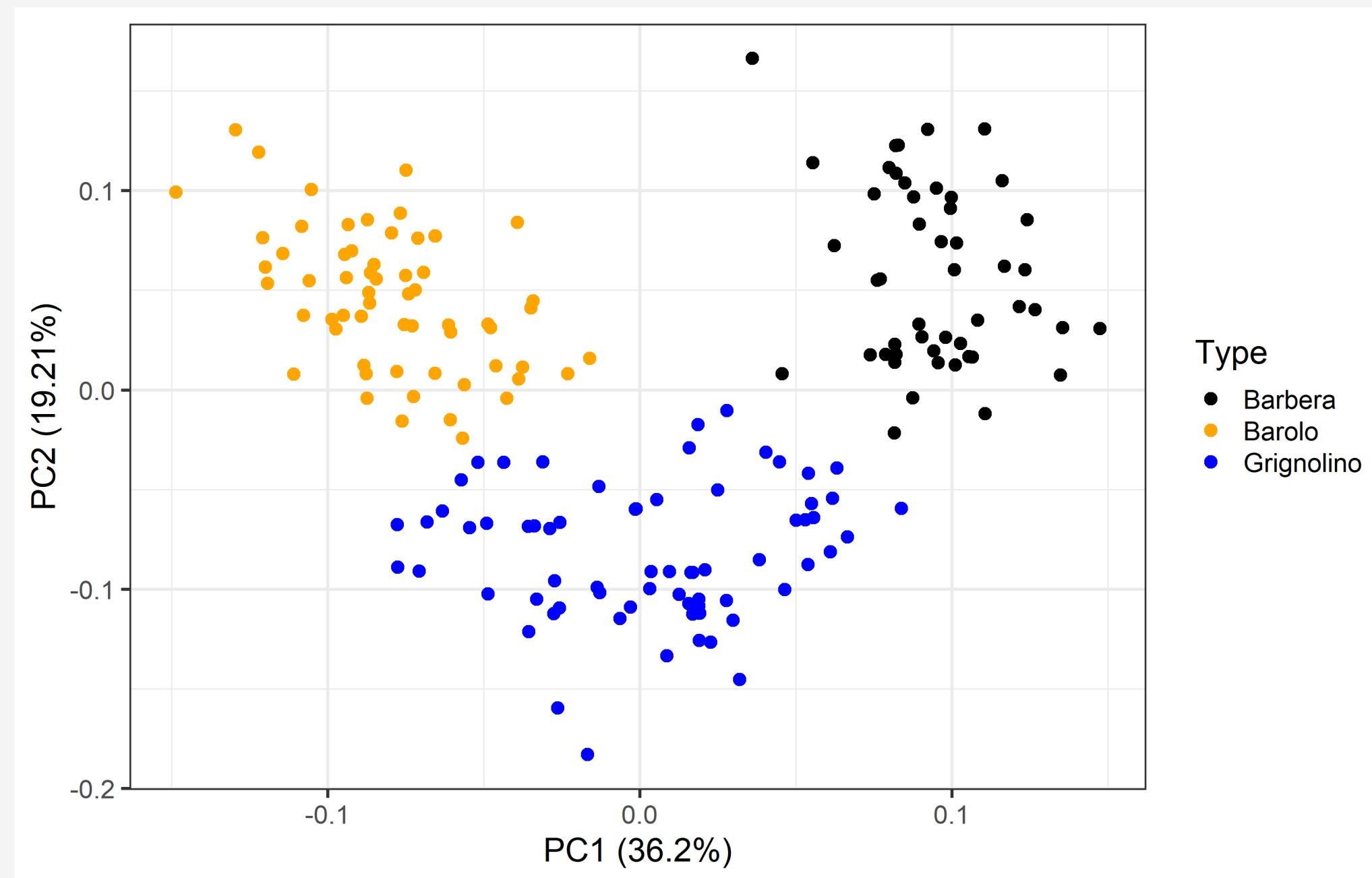
If two wines are close in PC1 and PC2 value they are **similar also in their original features**.



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

PRINCIPAL COMPONENT ANALYSIS

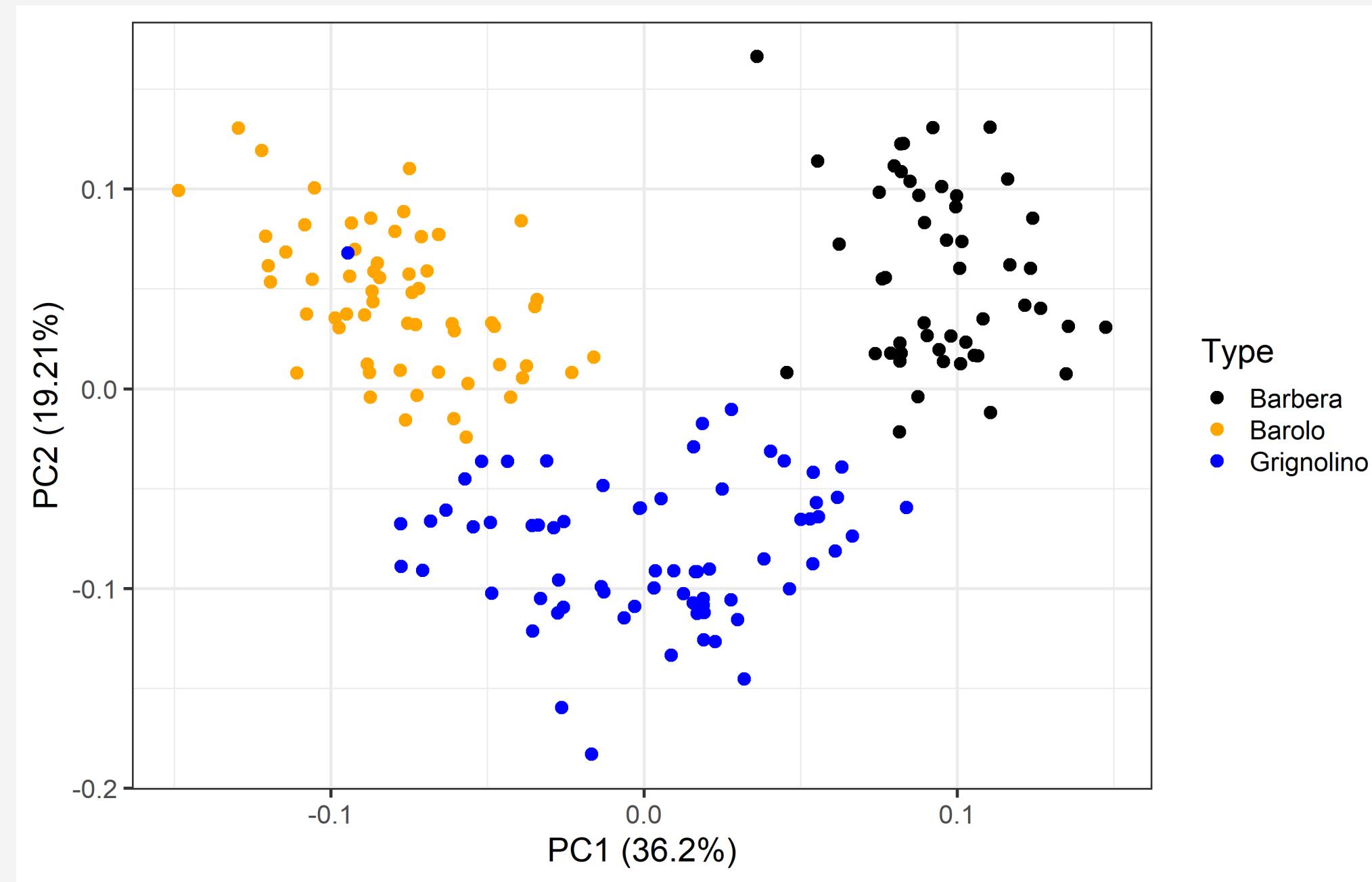
PC's are **linear combinations** of the original features



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

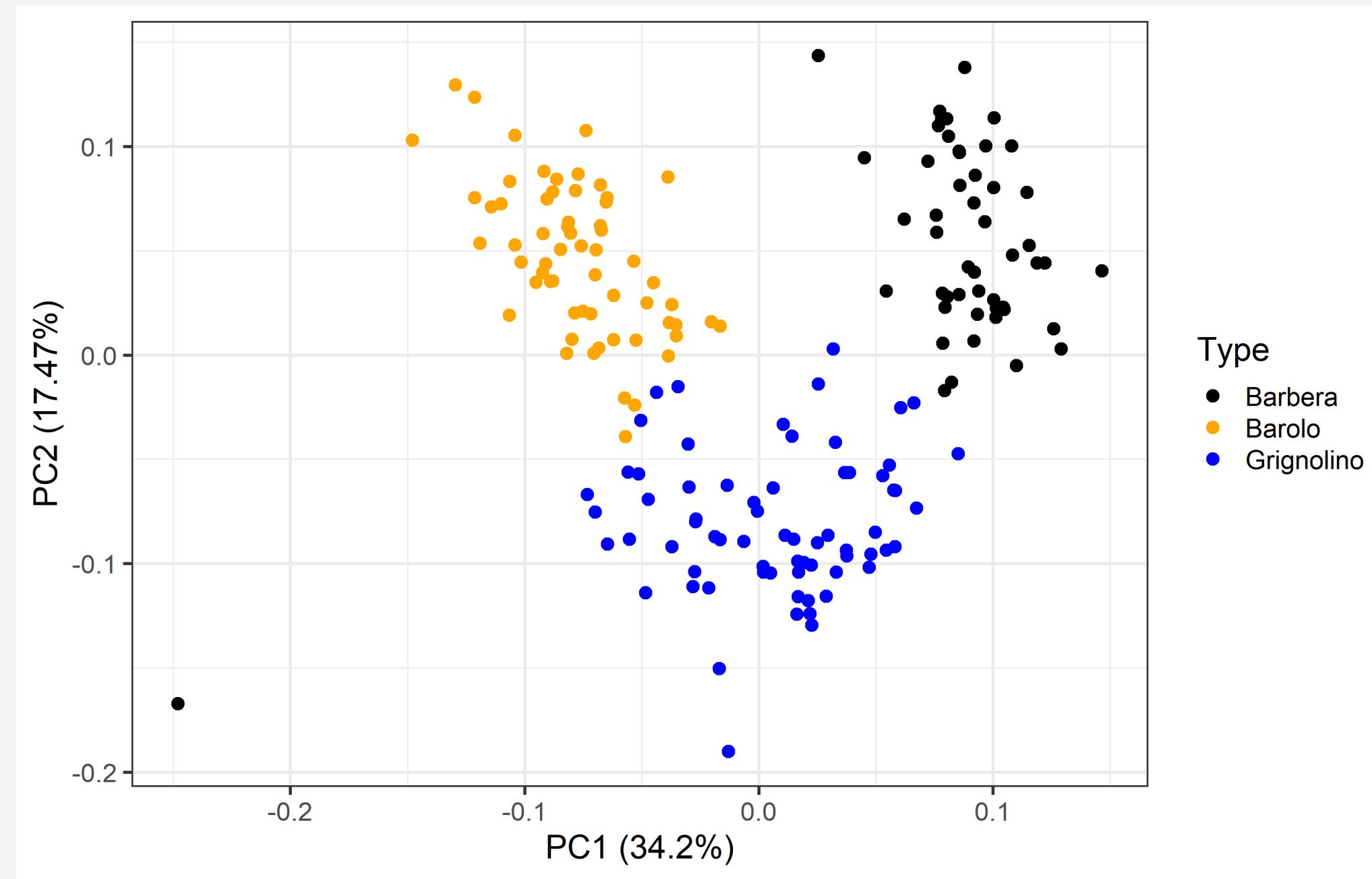
PRINCIPAL COMPONENT ANALYSIS

What do you think is going on in this PCA?



PRINCIPAL COMPONENT ANALYSIS

What do you think is going on in this PCA?

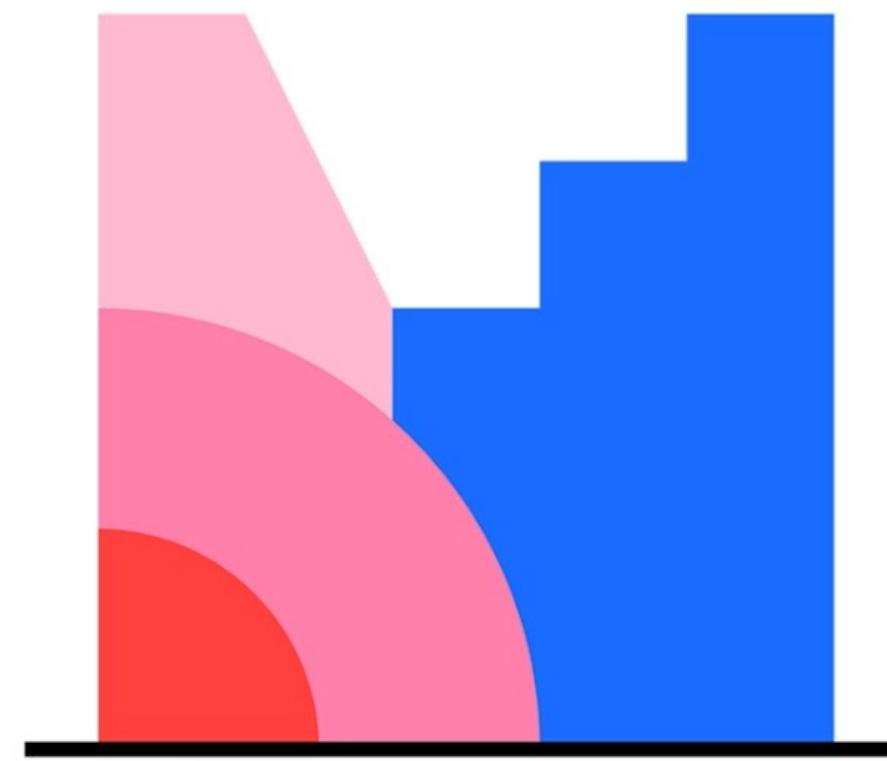


PRINCIPAL COMPONENT ANALYSIS

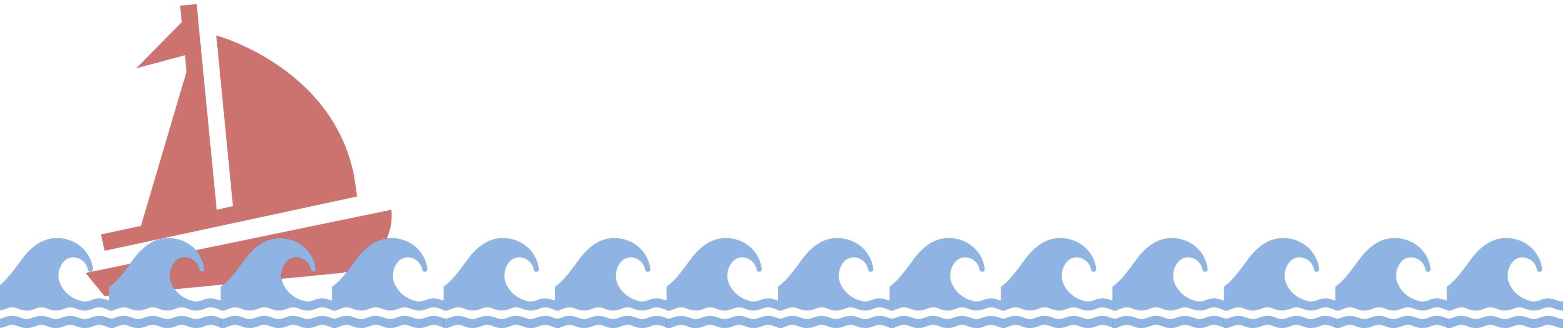
In summary:

- 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



Have you looked at a PCA
before?

GROUP DISCUSSION

In your group discuss the printed PCA plots.

- What can you see?
- What do you think it means?
- What should/could you do about it?



When things go wrong...

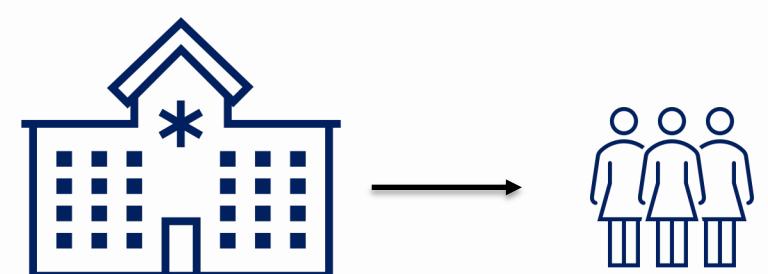
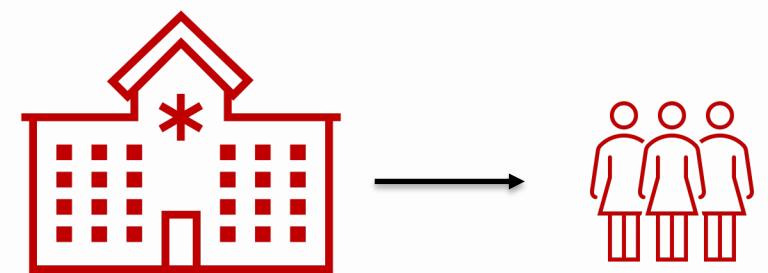
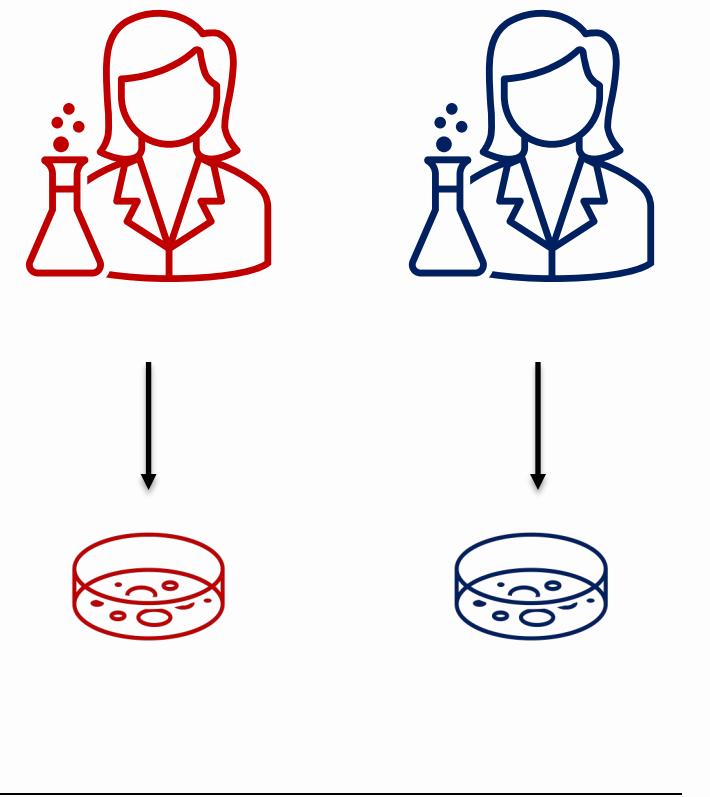


BATCH EFFECTS

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

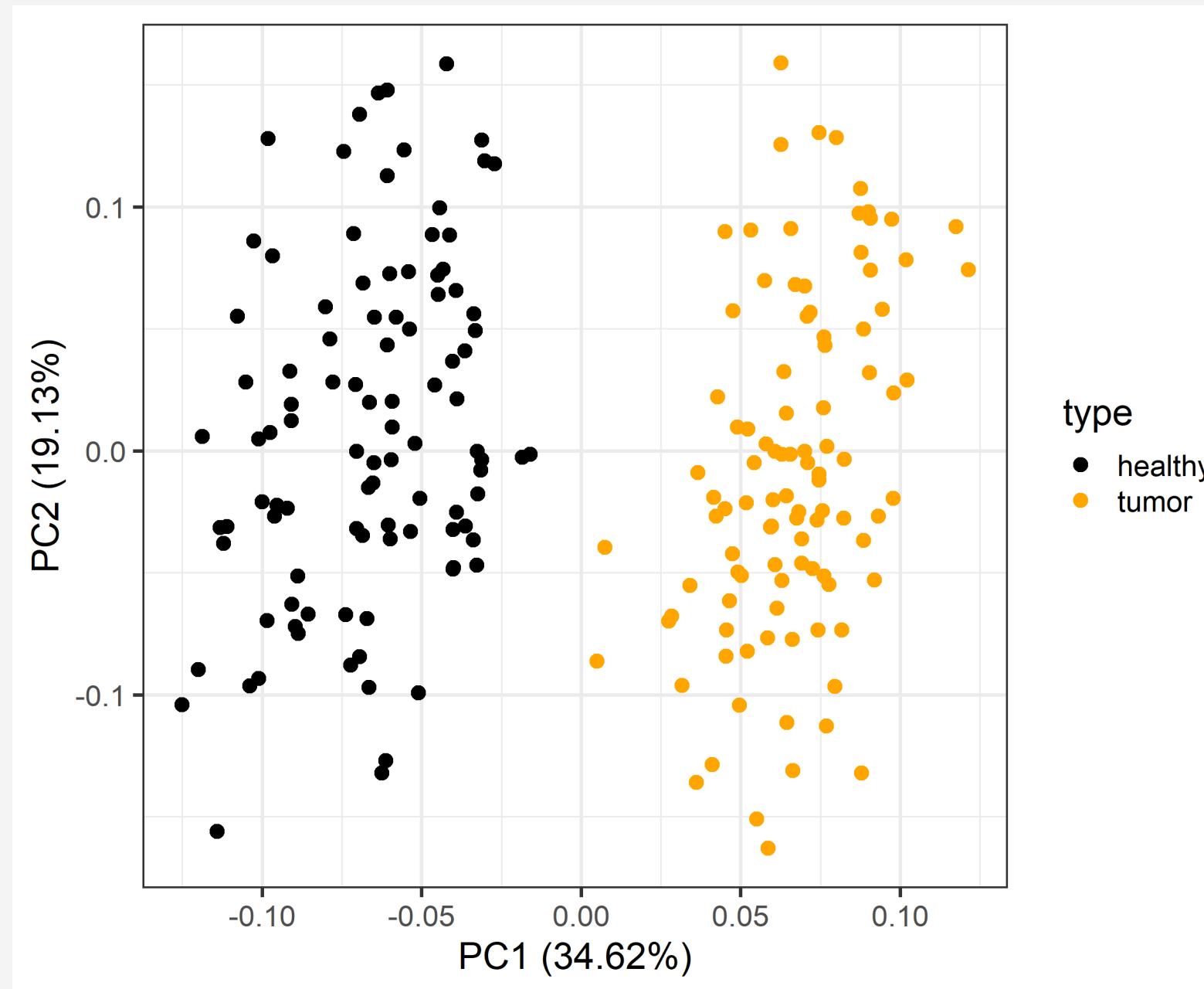
Batch effects can be corrected during analysis, if they are **not** correlated with the outcome.

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



Batch effects are visible in the PCA

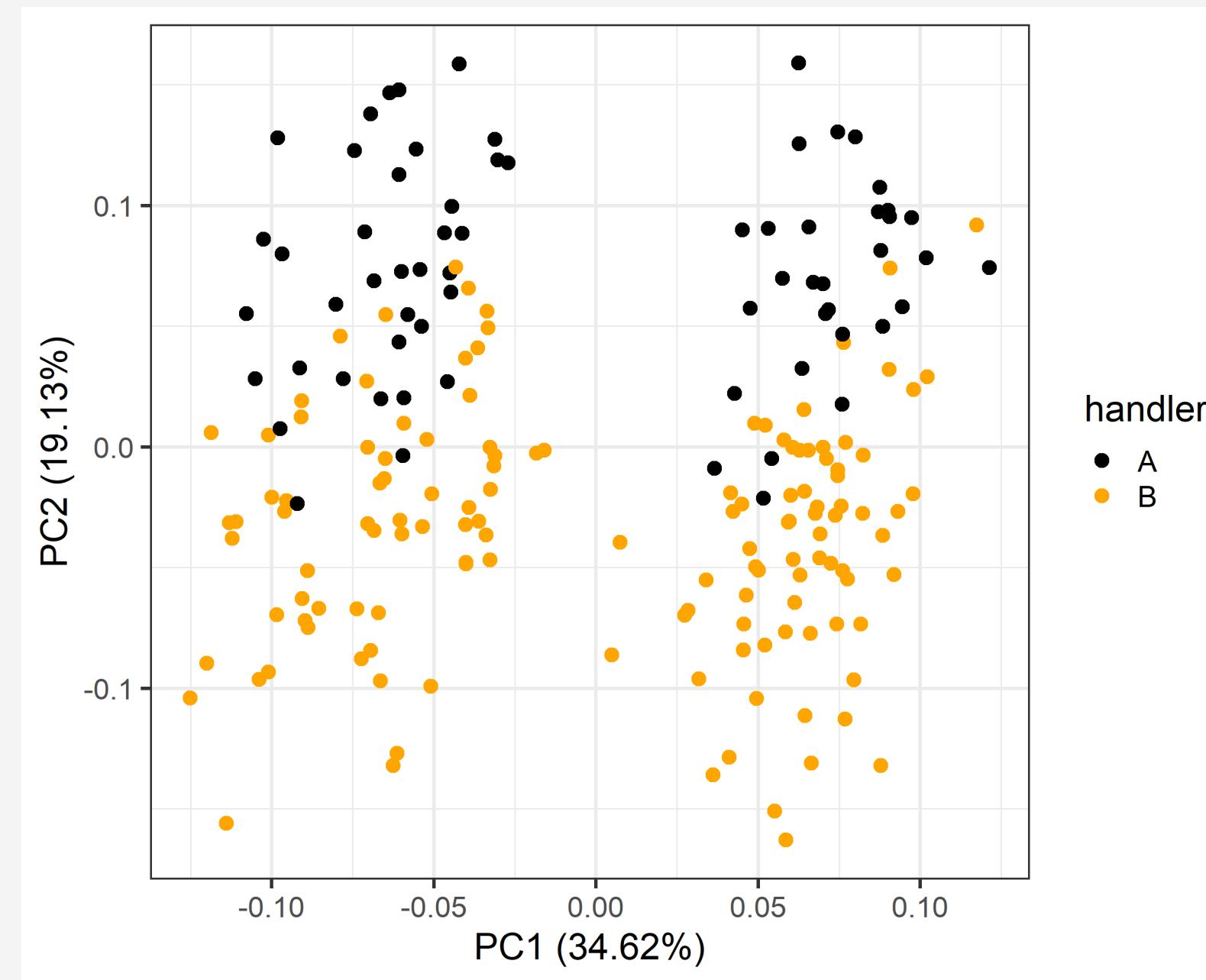
PCA plots can be used to investigate whether batch effects are present



Here we have a PCA of gene expression data from healthy and tumor samples.

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

Batch effects are visible in the PCA

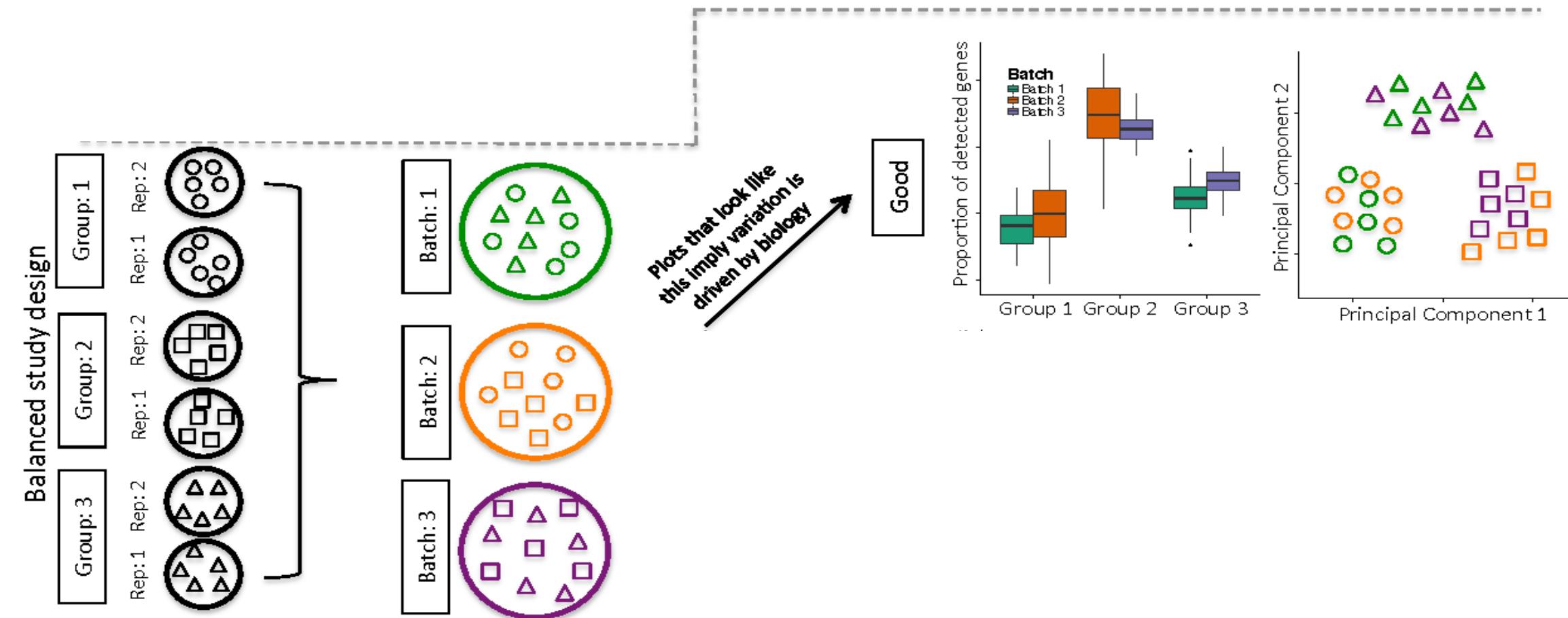


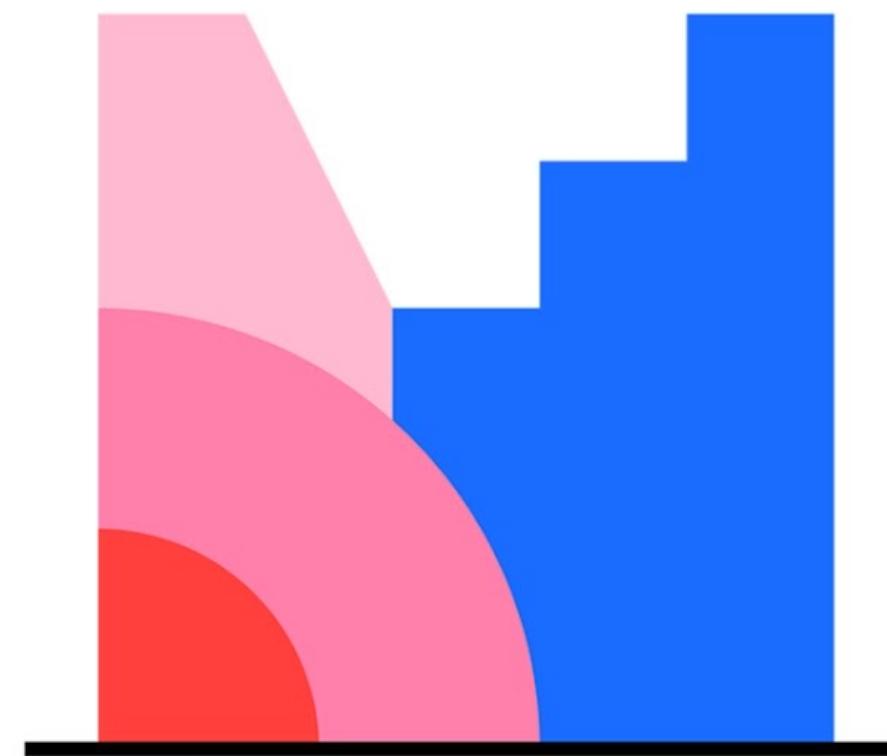
We can see there is another spread in the data along PC2, **not related** to the tissue type.

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

How do you think this would look if you had one person process all tumor samples and another all the healthy samples?

BACK TO EXPERIMENTAL DESIGN





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.

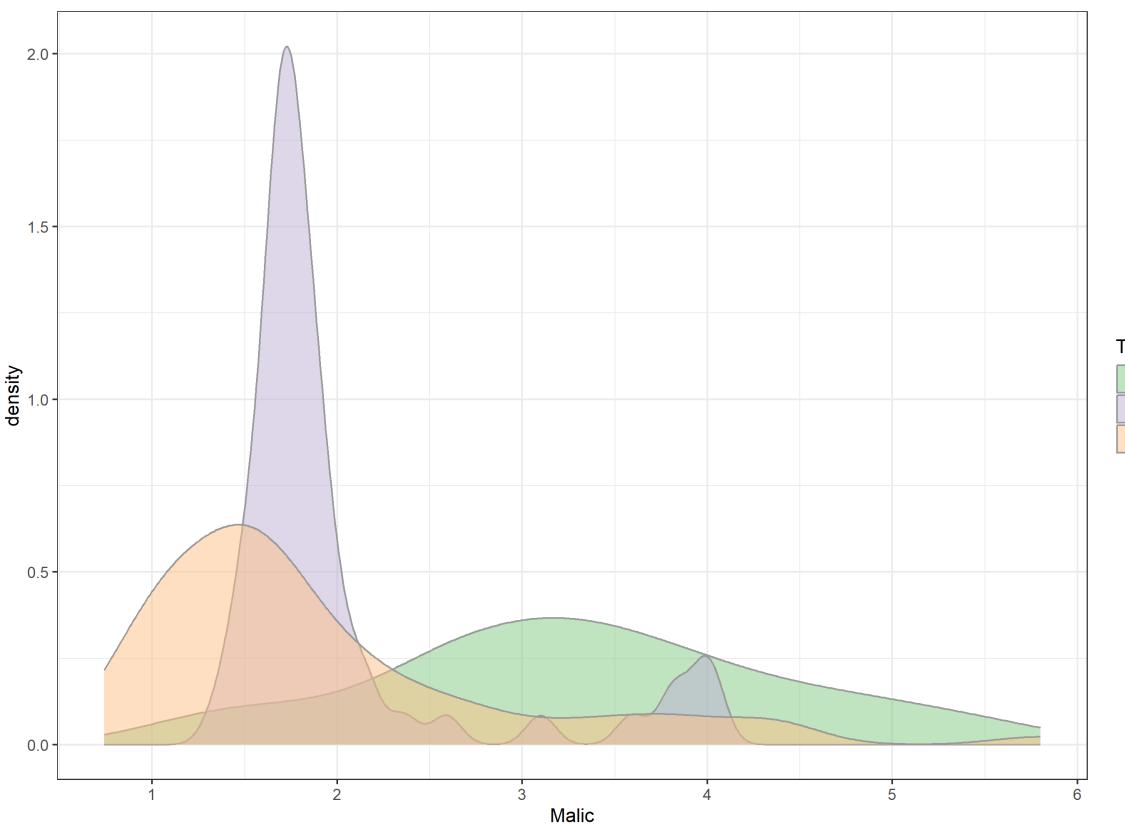
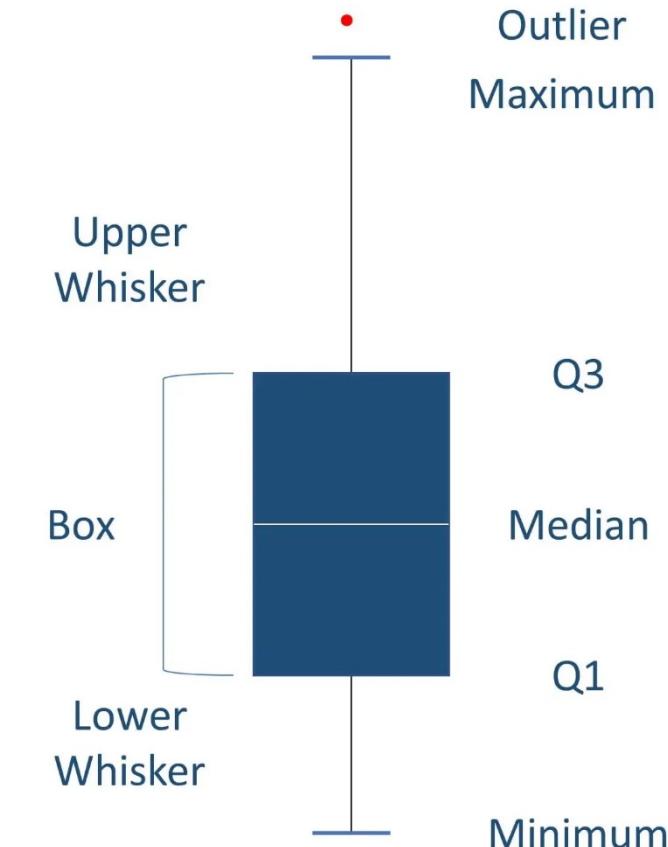
Which of the two workflows will lead to a confounded dataset?

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



OTHER CHECKS

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

- **Check the data:**
 - **Unreasonable values:** If some data points are in “out of scale” you should be able to see that in the PCA plot, but if you have few features it can also be a good idea to 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 - more on this later...

THE ROLE OF DOMAIN EXPERTS

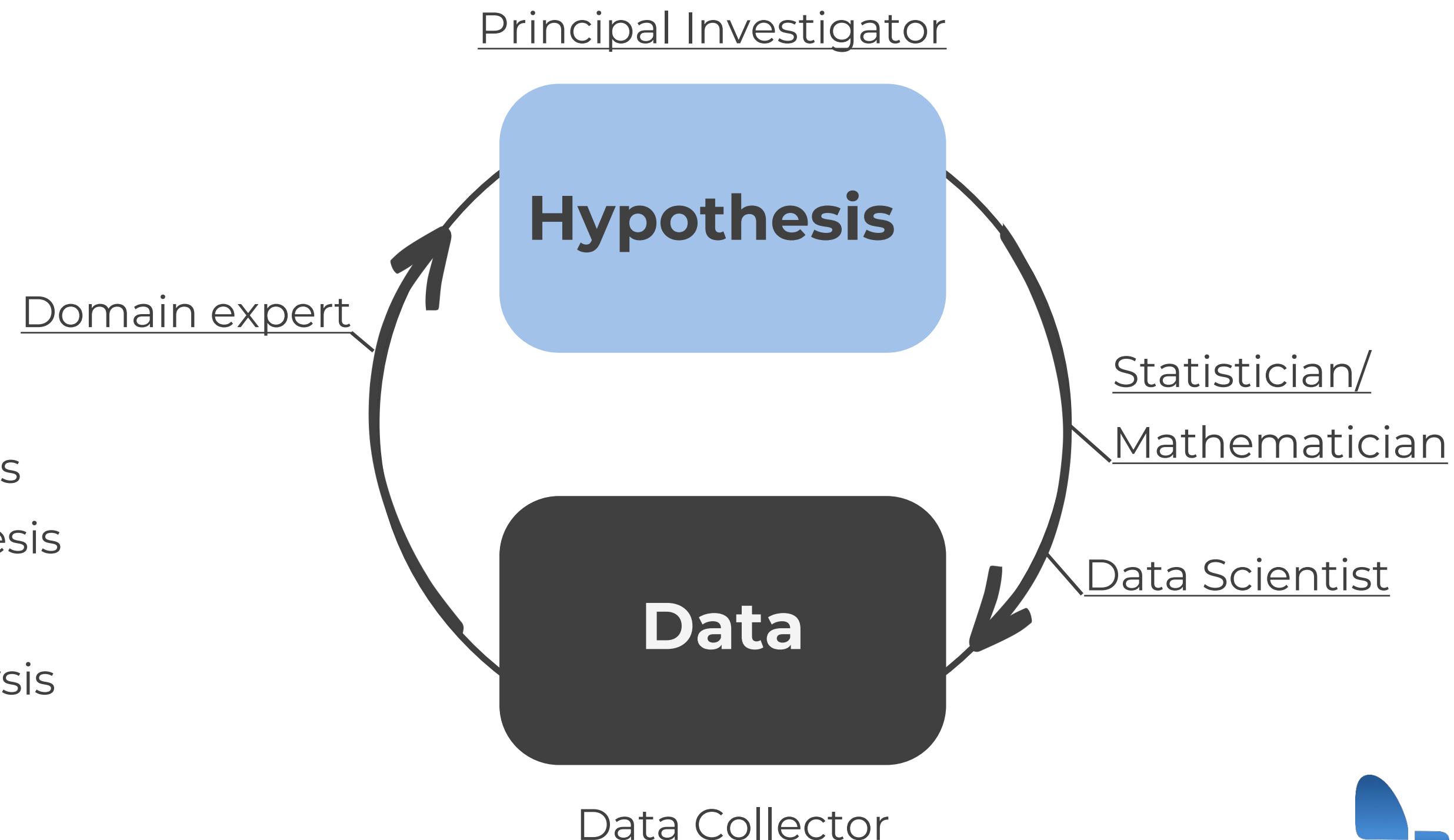
Domain knowledge plays a big role in EDA.

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

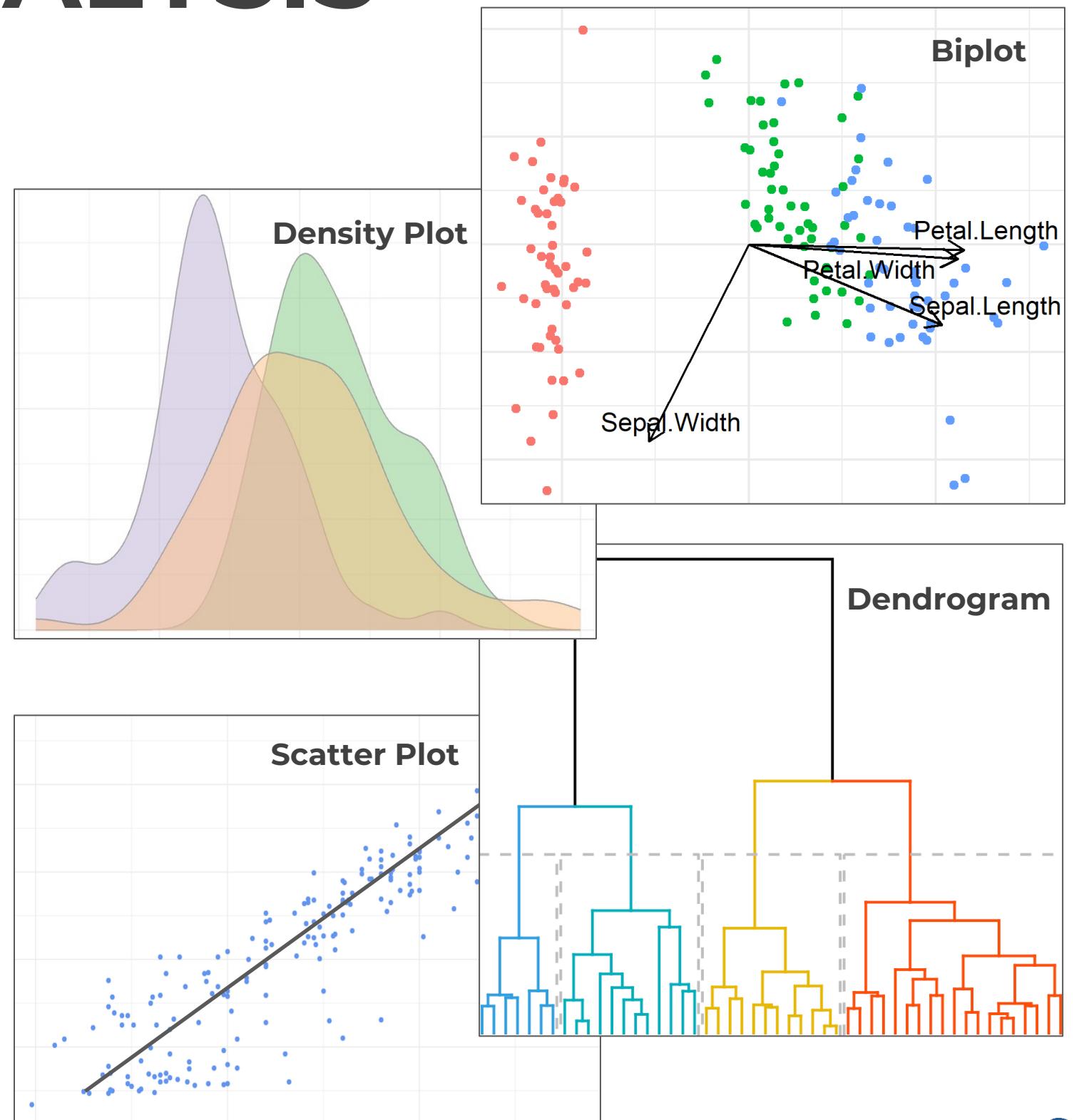


EXPLORATORY DATA ANALYSIS

This was a quick intro to EDA.

In summary it helps us to:

- Check if the data looks as expected
- Identify obvious errors: outliers, label swaps
- Establish what kind of analysis we can do
- Find out how we need to prepare the data for analysis (more on this later)



EXPLORATORY DATA ANALYSIS

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

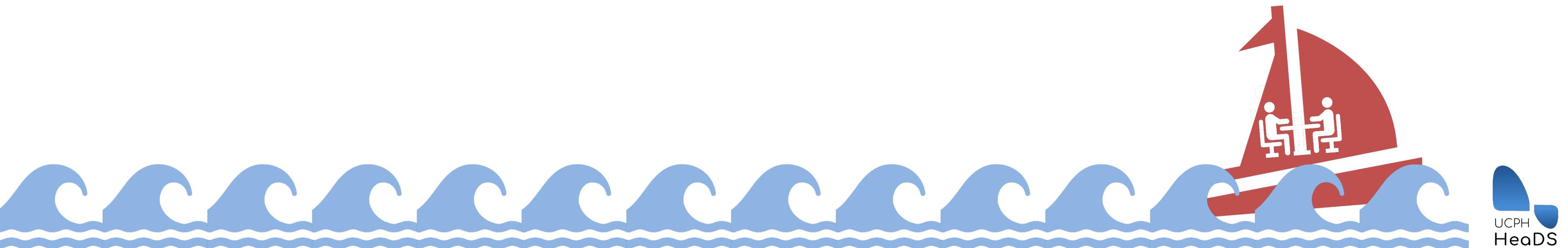
- Consult us in the Data Lab for a sparring/discussion session
- Consult the literature, specifically review papers. Find out what is known to work and what other people do
- Consult colleagues and collaborators
- Take a course/send a group member to a course

GROUP DISCUSSION

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?).

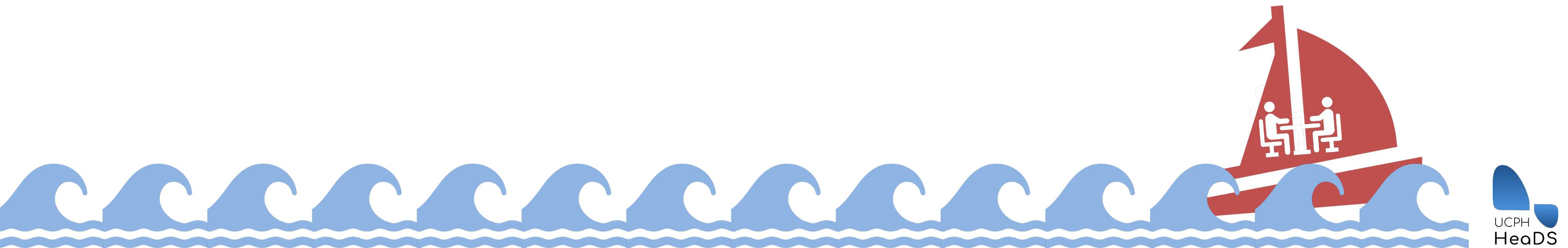


GROUP DISCUSSION

In your group discuss:

The density plot:

- What does the plot tell you about the distribution of gene counts.
Are the data normally distributed?



A BREAK IN OUR JOURNEY



A BREAK IN OUR JOURNEY

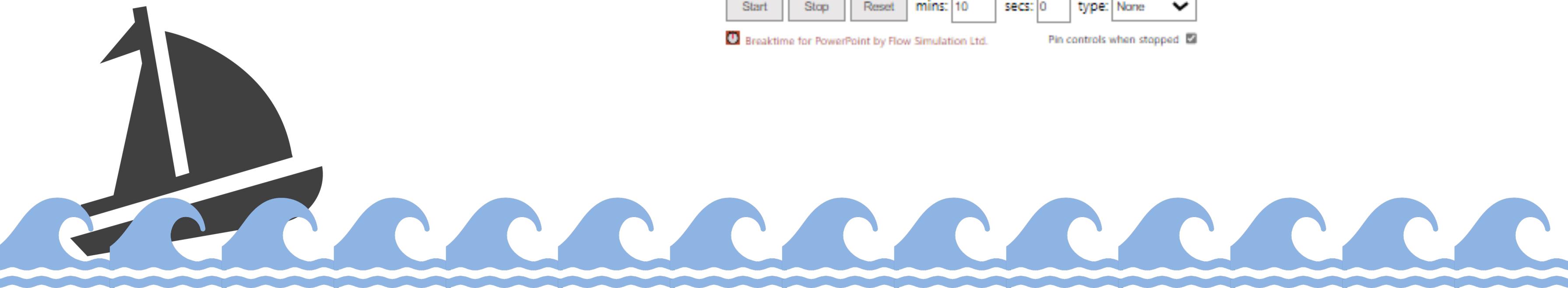


10:00

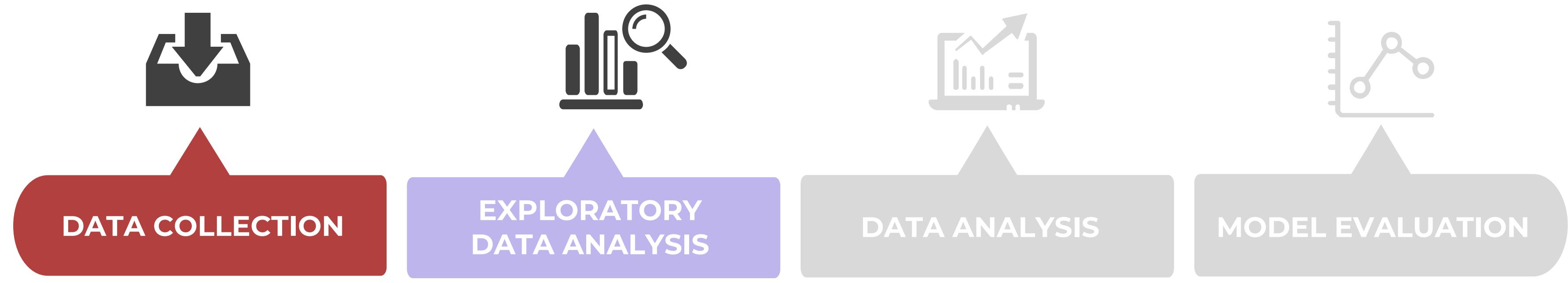
mins: 10 secs: 0 type: None ▾

Breaktime for PowerPoint by Flow Simulation Ltd.

Pin controls when stopped



DETOUR ON OUR JOURNEY



DATA CLEANING AND SET-UP

We might 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 (simplest version of this is substitution)

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

GROUP DISCUSSION

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 ?



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).

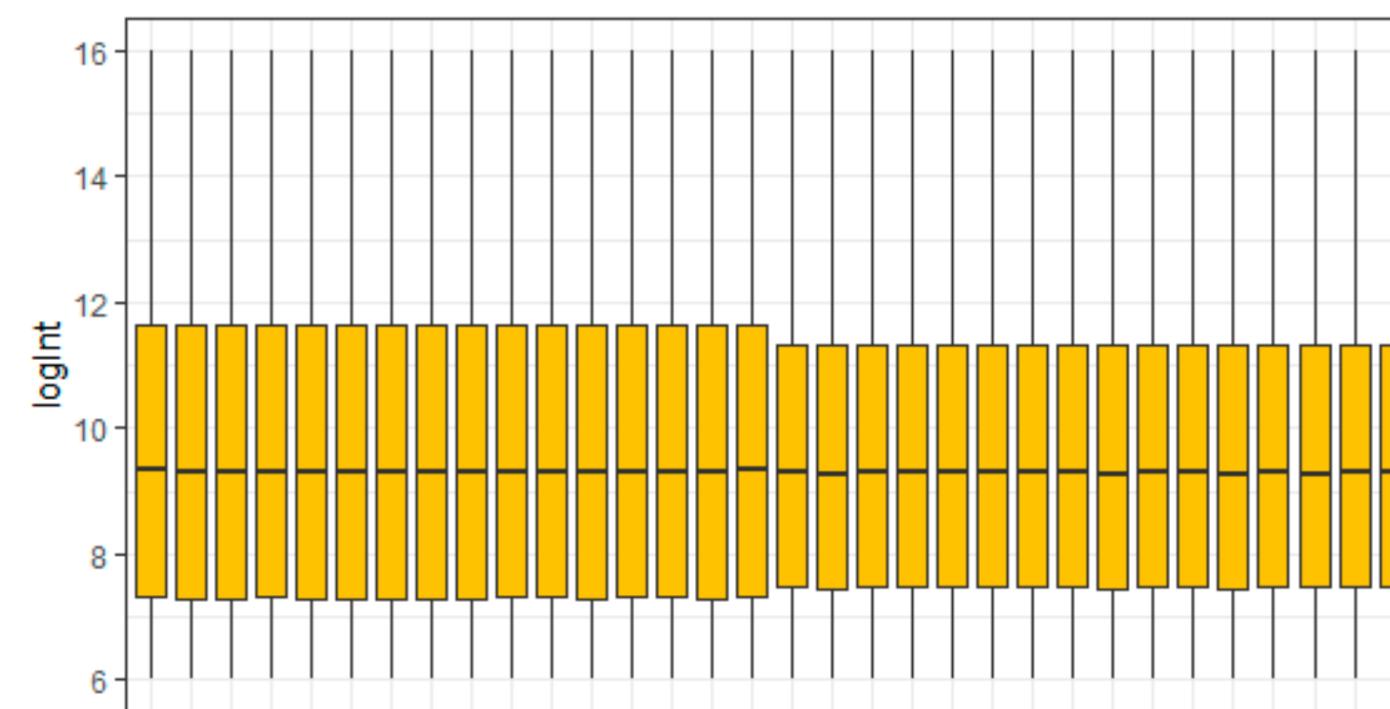
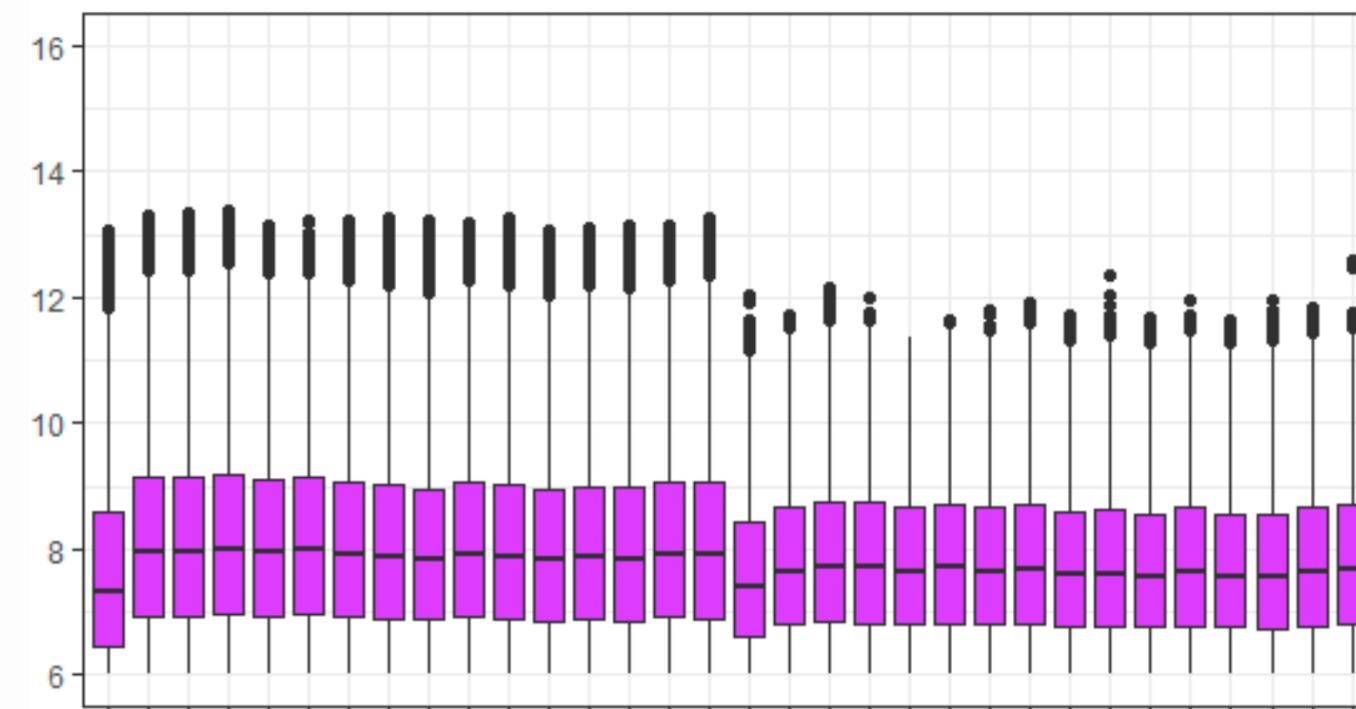
UNWANTED VARIATION

Technical variation:

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

Non-technical variation not related to the outcome:

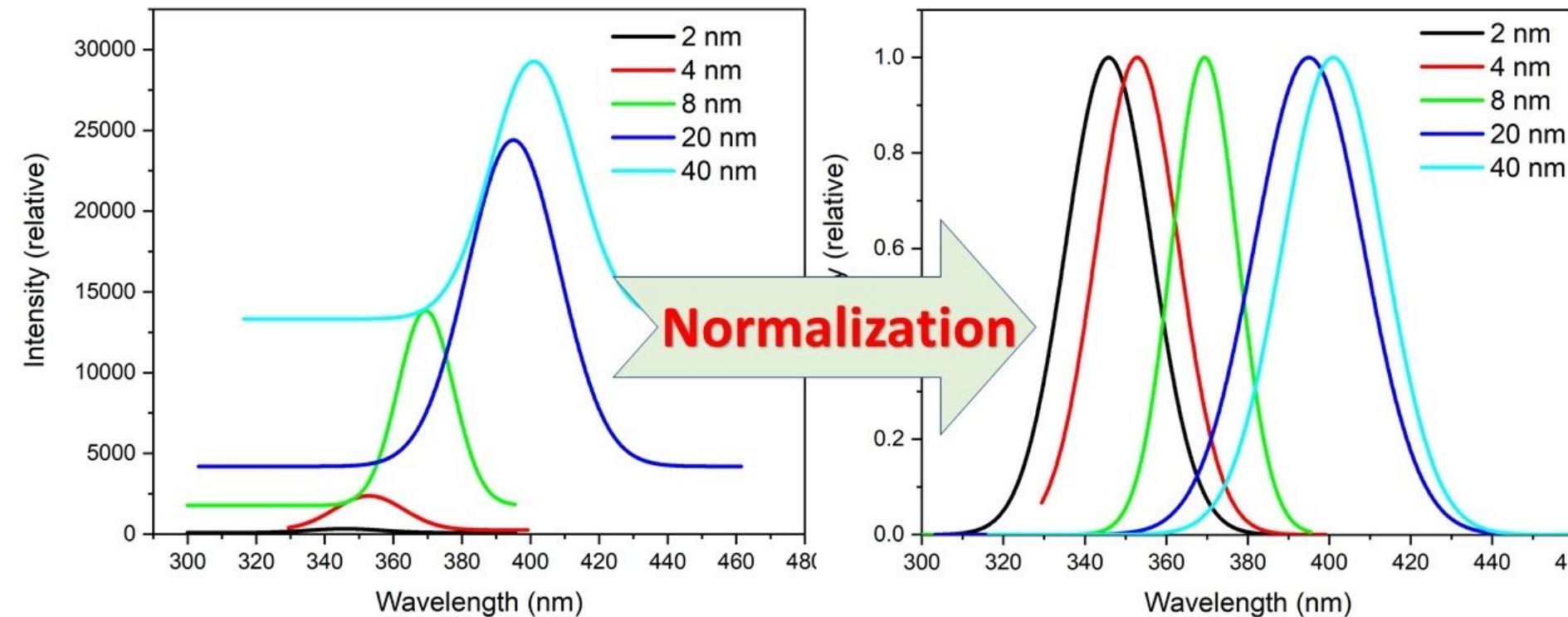
i.e. gene length and library size (number of reads) in RNAseq



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

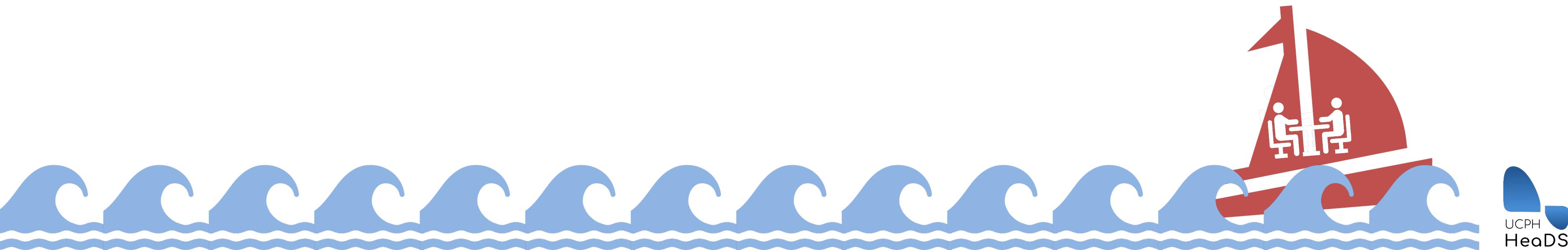


Important that one uses the same normalization with same parameters for the entire dataset.

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

GROUP DISCUSSION

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



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.

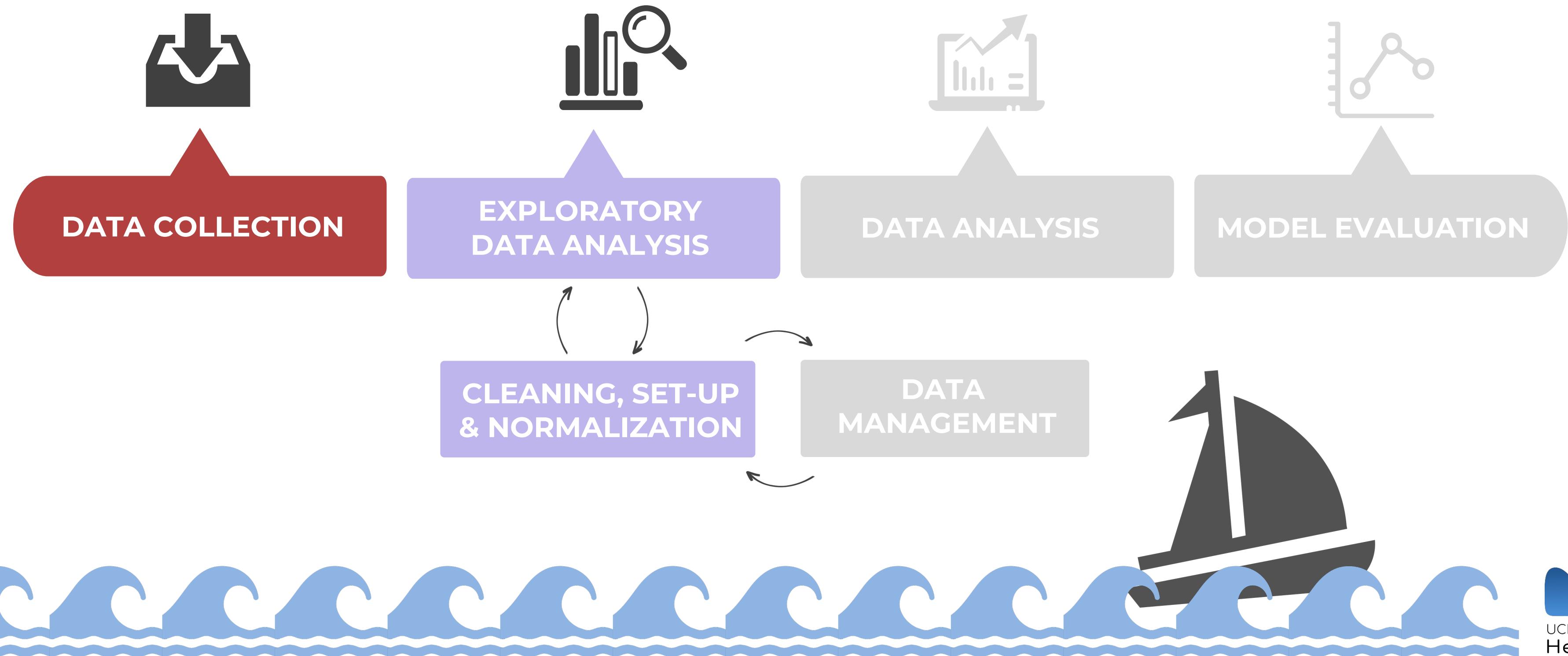
STANDARDIZATION

- For normally distributed data
- Scaled to a mean of 0 and standard dev. of 1.
- Scaling is not constrained 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.

DETOUR ON OUR JOURNEY



LUNCH BREAK IN OUR JOURNEY

45:00

mins: 45 secs: 0 type: None

 Breaktime for PowerPoint by Flow Simulation Ltd. Pin controls when stopped

