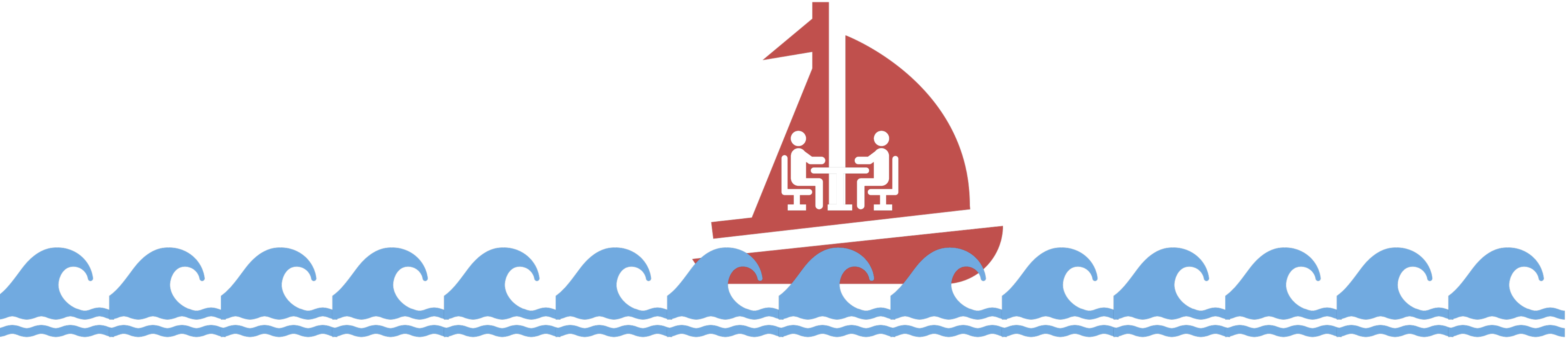
GROUP DISCUSSION - 1.0

Which of the **roles** we have introduced **do you see yourself in**?

Do you have people in your group or among your collaborators to fill the other roles? If not, **what are the alternatives?**



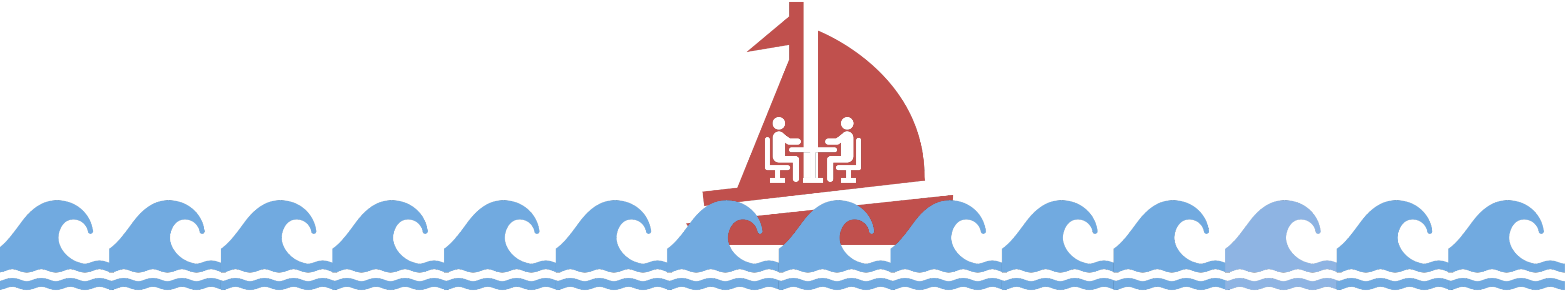
UCPH HeaDS

GROUP DISCUSSION - 1.1

Consider the following **scenario (1)**:

You would like to study whether smokers have an increased risk of heart disease.
For this, you have collected some data from both men and women on whether they smoke and have heart disease.
When you examine the data you see that all of your smokers are men.

Can you answer your research question with this dataset?

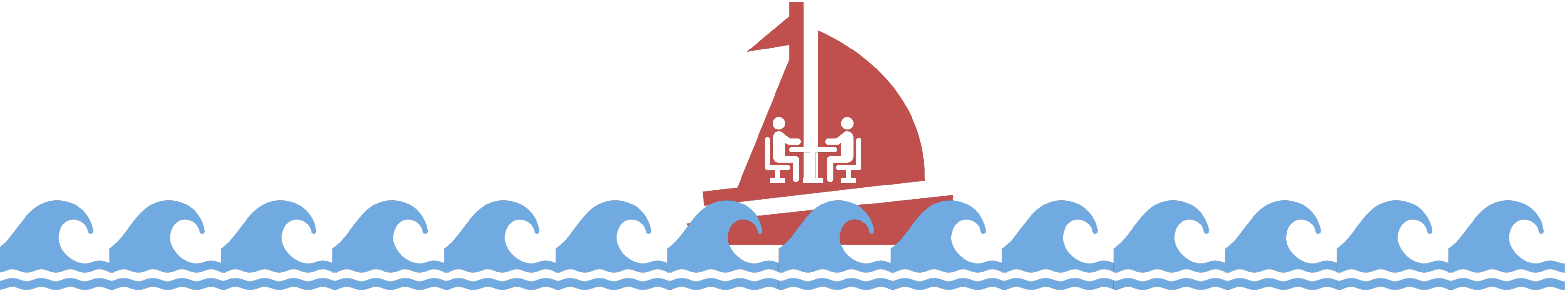


GROUP DISCUSSION - 1.2

Consider the following **scenario (2)**:

You would like to study differences in gene expression between tumor and healthy tissue. Since you have a lot of samples you ask two lab techs to each process half.

What are potential biases and confounders in this set-up?

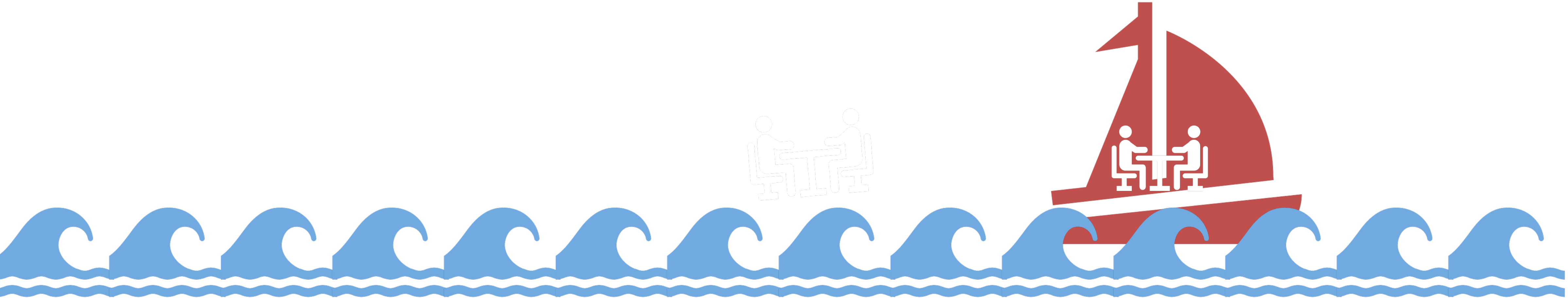


GROUP DISCUSSION - 1.3

In your groups discuss:

- What **data types** do you and your collaborators currently work with and/or what are you interested in working with in the future?

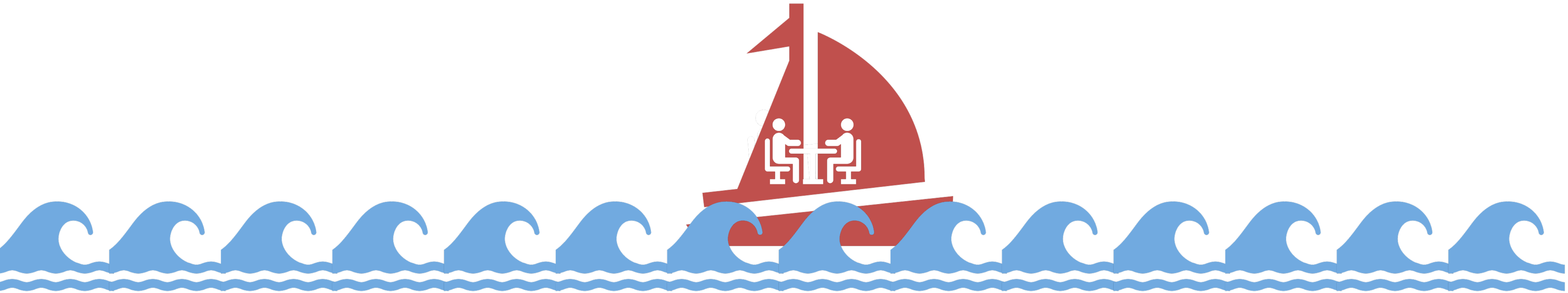
- What considerations are there in terms of **experimental design, data collection** and/or **data management/set-up**?



GROUP DISCUSSION - 2.1

In your group discuss the printed PCA plots.

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

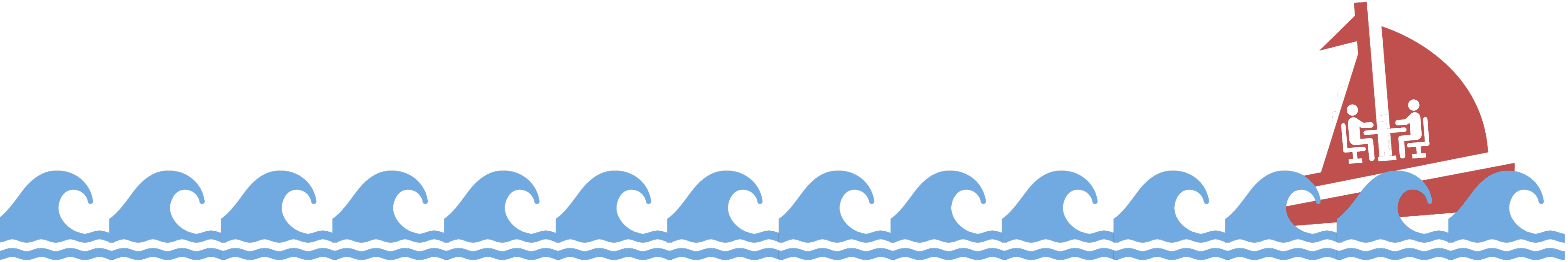


GROUP DISCUSSION - 2.2

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?).

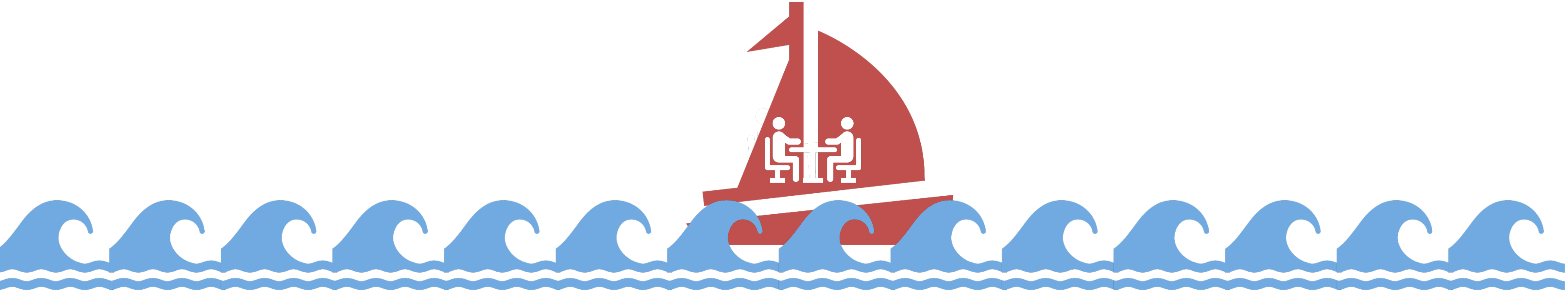


GROUP DISCUSSION - 2.3

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

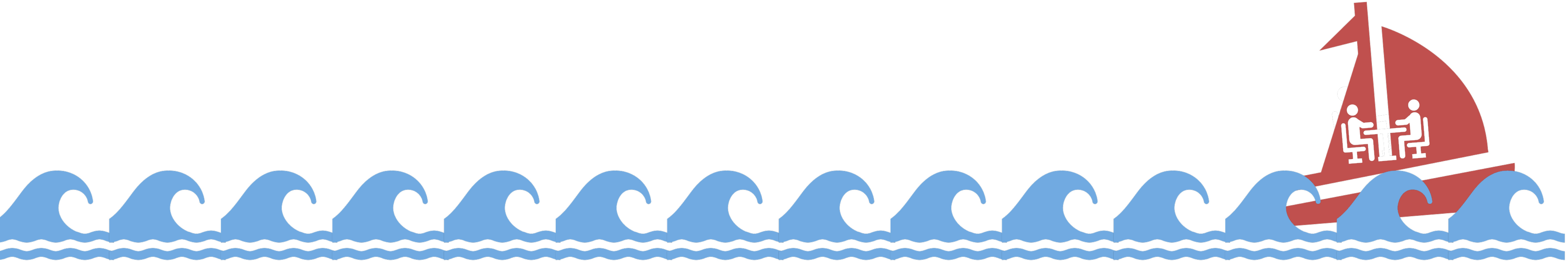
- Identify the different data **types** it contains (categorical, numerical, integer, binary, factors).

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



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?

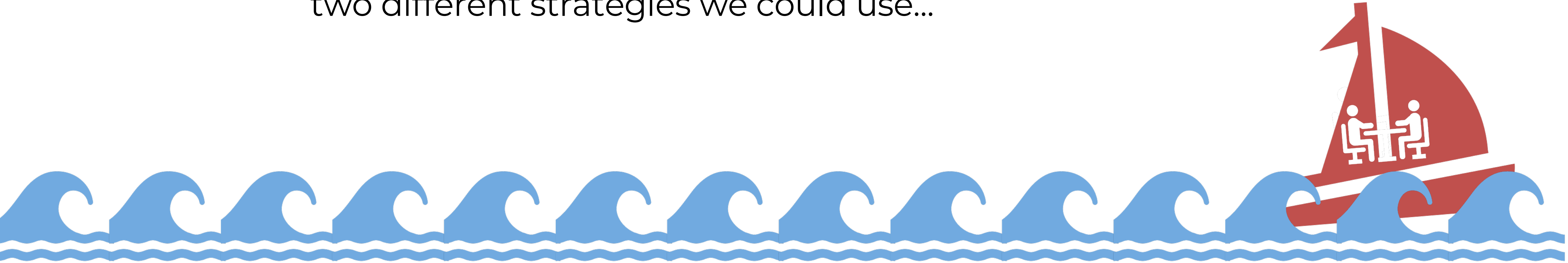


GROUP DISCUSSION - 2.5

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…

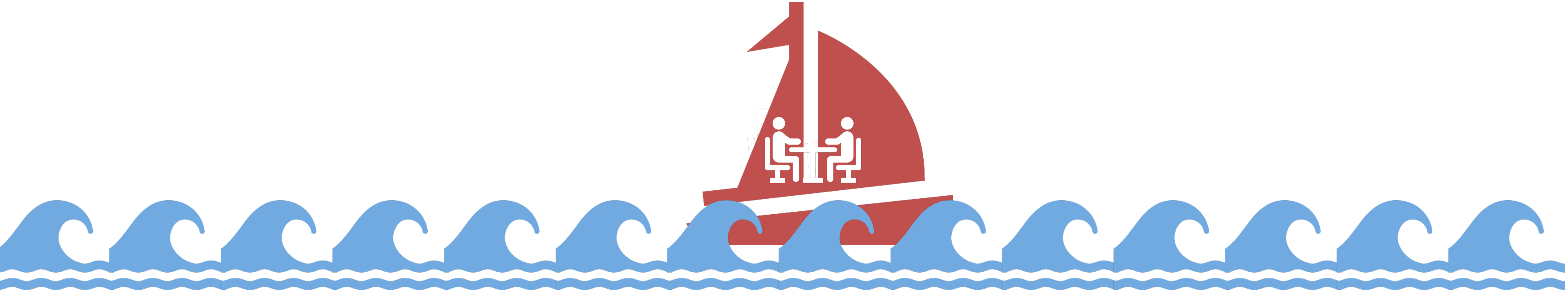


GROUP DISCUSSION - 3.1

In your groups discuss:

- Within your field what type(s) of model(s) is most often used? (slide 8 for inspiration). Why do you think this model is favored, i.e. what scientific question are you trying to answer (look at slide 3).

- If you had no ‘ground truth’ (no labels) to use for model training, do you think you could investigate a classification problem regardless? If so, what would the scientific question be?



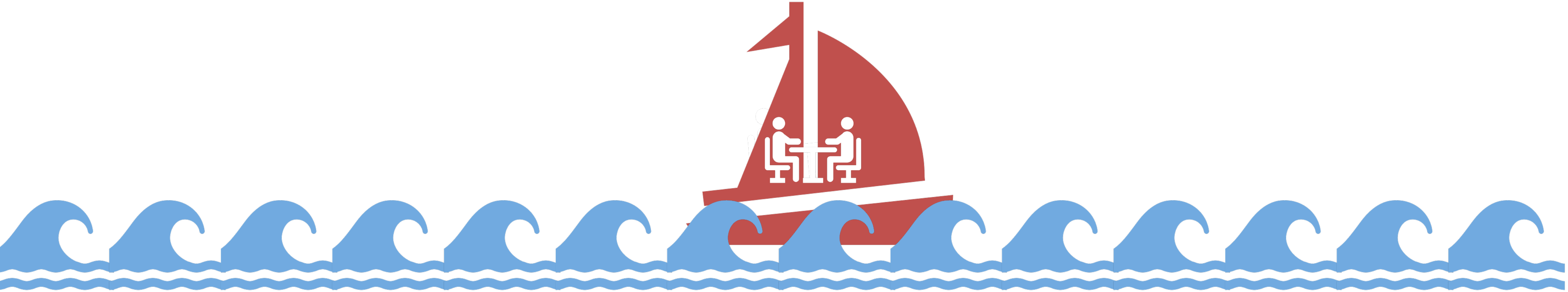
GROUP DISCUSSION - 3.2

In your groups discuss which of the three areas of DS analysis we talked about best applies to this question.

Is there a difference in weight between mice of strain A and strain B?

Further discuss:

What types of variables (data types) do we have? Do you know of any test you could use to answer this question?

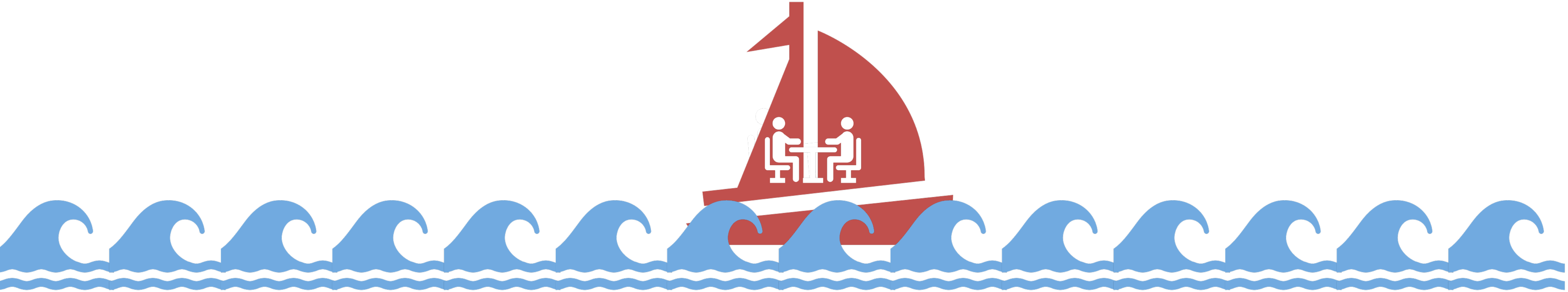


GROUP DISCUSSION - 3.3

In your groups discuss which of the three areas of DS analysis we talked about best applies to this question.

You have protein abundance data from skin samples (~ 10.000 different protein species). These samples were collected from patients with psoriasis (normal adjacent -and affected skin) and from healthy controls.

Is protein abundance predictive of the skin phenotype? And if so, are the levels of all proteins equally predictive/informative?

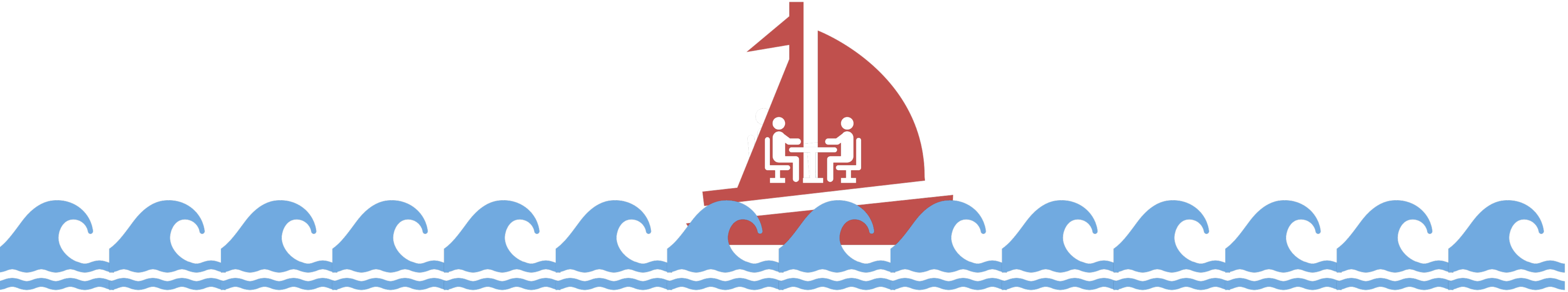


GROUP DISCUSSION - 3.4

In your groups discuss which of the three areas of DS analysis we talked about best applies to this question.

Is the amount of bacterial load in a swap of the oral epithelium (gums) based on skin type, diet and whether the person recently had antibiotic treatment?

Further discuss:
What are the outcome variable(s) and the explanatory variable(s) in this scenario?



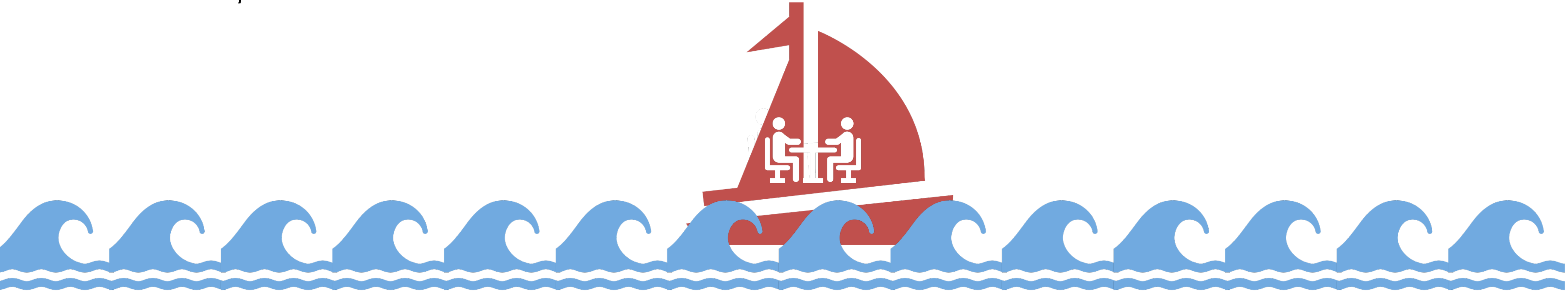
UCPH HeaDS

GROUP DISCUSSION - 3.5

In your groups discuss which of the three areas of DS analysis we talked about best applies to this question.

Gene expression data and biometrics (height, weight, age, etc.) from patients with colorectal cancer. You are interested in exploring if there are any potential subgroups of cancer patients within your dataset, in order to pair each subgroup with the appropriate healthy controls.

What type of analysis could you use for this? *N.B while avoiding a fishing-expedition*



GROUP DISCUSSION - 4.1

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

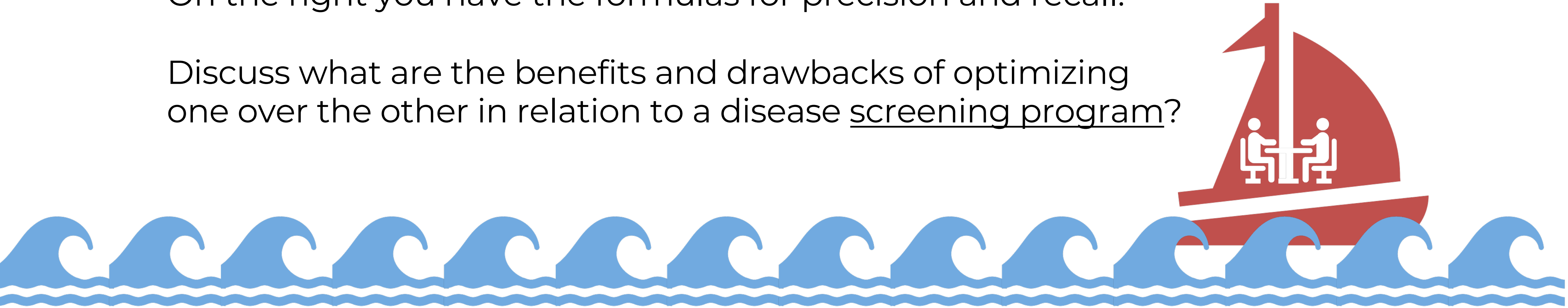
- True positive
- False positive
- True negative
- False negative

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

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

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?

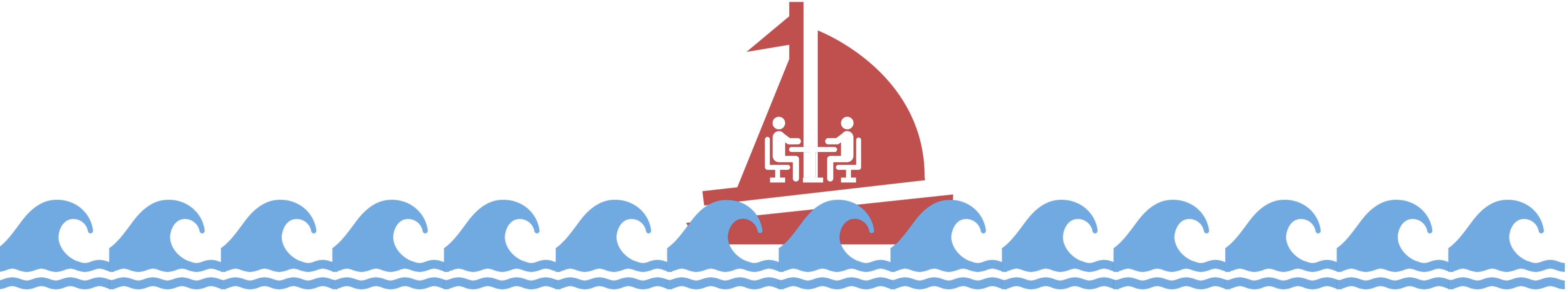


GROUP DISCUSSION - 4.2

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.**

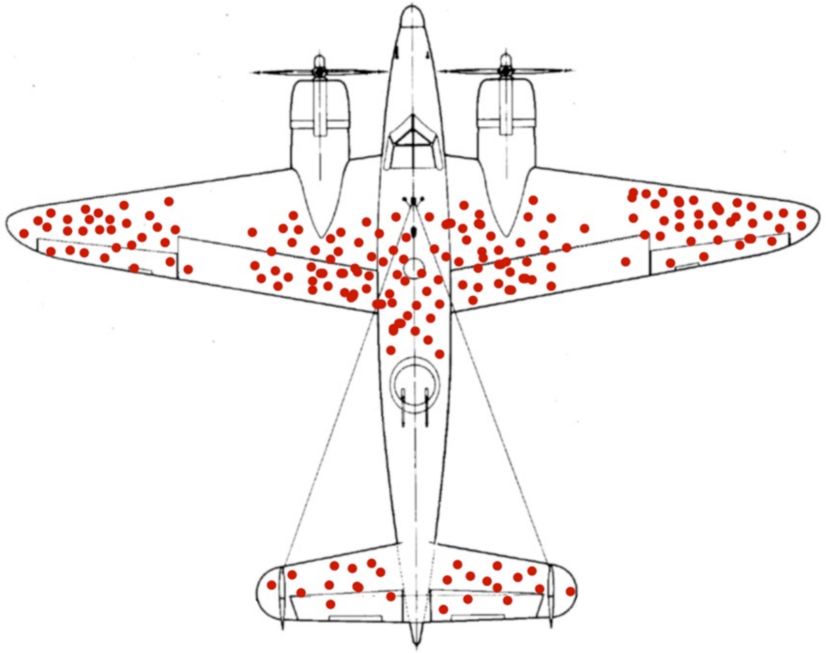
In general, what is the problem with high-dimensional model with many parameters?



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?



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.

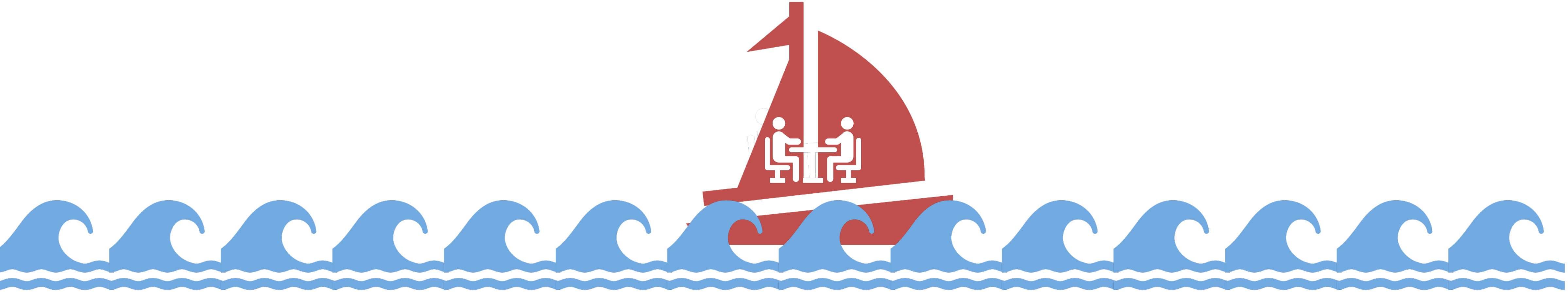
2. In a cohort study, researchers found an inverse (negative) relationship between exercise and prevalence of arthritis in a group of individuals, i.e. less exercise more likely to be diagnosed with arthritis.

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...

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.?



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?