Formative presentation feedback

# Group 4

### **General feedback:**

**When performing your preliminary analysis:**

● Familiarise yourself with the relevant domain. You will save a lot of time if you can distinguish between relevant information and noise. Statistics/Machine learning can help sort out details like variable collinearity, but it cannot necessarily help you distinguish the biologically relevant information from noise. You need to remember that high predictive accuracy does not guarantee biologically meaningful results.

● Look for patterns that will help you build your story. What insights can you derive just by looking at your dataset.

● Try to understand the challenges the dataset presents in the context of the problem you are trying to solve. For example, if your dataset is full of outliers or low quality data, how is this going to affect your model’s predictive power?

**When you analyse your data:**

● Familiarise yourself with existing work, not just state of the art, to help you identify the best practices. It is very likely that someone has previously done something similar to what you are trying to do. For example, the standard practice is to perform the clustering prior to dimensionality reduction. In contrast, classification tends to work better after data compression.

● Irrespective of trying to extend or disprove an existing method for analysing data, you need to gain the audience's trust. Formulate your argument in a way that serves this purpose, but be careful not to “lie with statistics”. For example, if you would like to suggest an extension to an existing approach, start by reproducing existing results. Demonstrate the circumstances under which the approach fails and suggest extensions.

● Consider how to test the performance of your approach - what approaches/metrics can you use to measure how well it has done at classifying the data?

● If you are developing your own approach, make sure you compare it with what already exists, including testing its performance.

● When consulting, you have to investigate what the model has learned and try to connect it to existing knowledge. Models that outperform state of the art but can’t offer any meaningful insights are much less likely to be trusted or used.

● It is recommended to start with a simple machine learning approach and see how far you can get. This helps you set the basis for your analysis and gain your audience's trust. For example, if you have to do dimensionality reduction, start with PCA before applying an autoencoder. PCA is a widely accepted tool which can help you set the threshold you have to beat and prevents your audience from pushing back.

**Presenting**

Many teams need to work on their presentation structure. A format which can be effective for communicating is the following:

● A slide with the title of the presentation and names of people involved.

● An overview of the presentation. Bullet points of the main sections. This will summarise what the audience is going to hear through the presentation.

● Introduction. This is perhaps the most important part. Here you introduce the core thesis. In a few sentences you need to argue why what you are talking about is important. (Why should your audience care?)

● Data exploration. Summarise the exploratory analysis you have performed. Focus on things that you found interesting and help you build your story. If exploratory analysis is carefully done, it is very easy to justify preprocessing and modelling steps.

● Data preprocessing. Here you build on your exploratory analysis. Based on patterns you have observed, formulate your filtering/preprocessing criteria and summarise them in a few bullet points. Clearly summarise your dataset after preprocessing.

● Methods. If you are introducing a new method/ presenting an existing one, you can give some high level description. If possible, try to justify your model assumptions based on the insights you have gathered in data exploration.

● Results. Present and analyse your results. Make sure the results are presented in a cohesive order which serves the narrative you are trying to build. Make sure figures are clearly labelled, font size of legends/axes is large enough to be readable. If someone cannot read your figure, they will very likely ignore it.

● Further work (Optional). If your presentation requires further work, clearly state in which direction it must happen. In many cases it is important to include a realistic timeline.

● Conclusion. You should recap your presentation in a couple of bullet points. These might include explanations as to why the problem is interesting, what are the main challenges and how did you address them.

On top of the structure, there are some details teams need to be careful about. Such details are important to maintain audience attention/interest.

● Make sure you understand your audience and adapt your presentation accordingly. For example carefully choose the amount of technical language/details depending on your audience.

● Choose a good presentation layout. White background with black letters guarantees that almost everyone can follow your presentation without being distracted. If you want to use graphics make sure they don’t interfere with your message.

● Keep only necessary information. Multiple images with similar information should be avoided.

● Try to be concise. Bullet points with a lot of text make your life easier, but the audience stops paying attention to read them. Information rich slides make your presentation shorter but your audience cannot follow them.

● Try to stay on time. A rule of thumb is to allow 1 slide per minute. Given the time-limitations you can estimate how long your presentation should be. But don’t be tempted to stay on time by presenting data-dense slides faster than the audience is able to understand them. If you have to do this, you need to think carefully about how to get the key information across concisely.

● Make sure you save your presentation in a format which is unlikely to /have display problems on a different system.

● Prepare before presenting. Practise your presentation and try to think of questions your audience might ask.

### **Specific group feedback**

We liked that you have performed preliminary analysis and have identified some of the challenges the dataset presents. We feel that the preliminary analysis has to go deeper to complete the remaining challenges. Students could give a much more detailed explanation of their preprocessing steps and how the final dataset looks like. Choosing the best approach for representing CDR3s based on how they cluster down the line might be a problematic approach. There is a possibility that KNN performs better because you are biassing the data with the one-hot representation. As CDR3s have different lengths, you have to make sure you are not biasing the data when padding with zeros. It is not clear why combining alpha and beta chains will cause you to lose information during the clustering and classification processes. To understand whether you need the MHC you need to have a look at how the TCR peptide HLA complex looks like. Ensure you explain all essential steps - it wasn’t clear how distances between sequences were being calculated. Regarding the presentation, you showed the results you have obtained so far in a good way. However, there was really small font size on the graphs. You had many graphs on a single slide and it is difficult to focus the attention of the audience on what you want to show when you do that.