

# Final Project Part 1

Maximilien Filion - 20598645,      Aneesh Kaura - 66832304,  
Monica Mihailescu - 47409289

2025-11-17

**Analysis roadmap - Note, this is just one way of doing the analysis and it is up to the teams to decide about further analysis steps.**

I have listed quite a few options that you can consider when analyzing the data. Below, I am going to provide some more details regarding mutation analysis that could potentially help you to get on the right track.

- Clinical data
  1. Explore clinical data wrt the distribution based on various variable/attributes such as age, stage, survival, etc
- Mutation data
  1. Build a gene-patient matrix that contains mutation data. Within this matrix, “1” represents whether there is a mutation in a given gene and given patient and “0” represents no mutation. Note, when building this matrix, you can take into account the type of mutations (e.g., synonymous, non-synonymous, etc)
  2. Pick the top 20 or most frequently mutated genes from this matrix.  
Perform clustering on the resulting mutation matrix Perhaps try ward and ward.D2 linkage methods if you see narrow clusters.
  3. Take the clusters of patients and (a) see if outcome data between patient clusters is different, (b) see if any other clinical data is different when you compare patient clusters - how much overlap do you see between the patient clusters?

**Some analysis plans that do not make sense within the context of the project:** • Perform clustering on data\_clinical\_patient • perform survival analysis on top 5 PCAs • suggest sth not requested, like global alignment possibly with Needleman-Wunsch algorithm • Perform PCA for each significant gene