**DATA SCIENCE PROJECT 1 (MAST90106)   
Semester 1 2021**

**GROUP 17**

Meeting No. 12

**Date & Time:** 19th May 2021, 3 PM - 4PM

**Agenda:** Update on the project with clients

**Attendees:** Kartika Waluyo, Vrinda Rajanahally, Roberto Bonelli, Brendan Ansell

Some models to try and work out:

- gene x, y and z in one sample is used to predict gene x in another sample

- neural network

- traditional machine learning: using linear regression and creating one model per gene in every tissue (not an accurate solution)

- Decision tree/Random Forests

- Penalised Linear Modelling

- Speak to someone who is an expert on neutral networks.

Input – will remain the same, genes from blood

Output – will differ, the genes from all the tissues we've chosen

Rough difference between test and validation sets:

Training: trailing the models and parameters

Validation: helps us determine which parameters work, which works the best

After choosing the best model method and parameters, test set comes in.

Test: to valid our model and parameters chosen on data that the model has never seen before. (make sure it has 30 samples at least!) 

Read column that Brenden sent

Come up with proportions

Make sure test set has at least 30 samples

Couple of plots/table - row is tissue (incld blood) - table gives the count of samples per set.

Speak to Michael on how we can model multidimensional correlated data