Data Selection Proposal – MAIS 202 Deliverable 1
Ritchie Yu

Project idea
A COVID-19 vaccine is urgently needed. However, a problem facing vaccines is that viruses evolve. Eventually, the virus may evolve to evade immunological memory, rendering the previous vaccine obsolete. This project aims to predict next-generation mutations in SARS-CoV-2. Such information would help pre-emptively identify novel mutations that lead to evasion of immunological memory.

Dataset: https://www.gisaid.org/
Contains ~98 000 fully sequenced SARS-CoV-2 genomes (~135 000 total) generated by labs around the world since the outbreak began. Data is publicly accessible, though it does require a registration process which I've completed and now have access to all of the data. I picked this dataset because of its sheer size, containing global data. Each sequence is documented very well, labelled with collection date, accession ID, sequence length, originating lab, authors, etc. The data also seems reliable, as GISAID has a team maintaining data quality round-the-clock. NCBI has an alternative dataset, but it is much smaller.

Methodology
Data Preprocessing
The genome sequences will be the most important information

1. Download GISAID sequences in FASTA format, collection date labels will be important
2. Remove trailing adenine nucleotides
3. Multiple sequence alignment using MAFFT or BioPython, and order sequences by collection date
4. Translate RNA collection into amino acid sequences using BioPython
5. Convert sequences into vector representations (similar to word embedding)

I am not sure yet if I will divide the sequences by geographic region, as information online has not clearly suggested whether that is an important factor. This is something I will be testing. As well, steps 4 and 5 may change, rather than translating the RNA I may opt to use the original nucleotide sequences.

Machine learning model
The objective is to predict next-generation mutations in the entire SARS-CoV-2 virus given historical genomic data. Another option is to predict mutations in certain parts of the virus. Based on past papers, I plan on using a recurrent neural network with LSTM, however a downside is that large input sequences may result in deteriorating performance. If this becomes a problem, there is a paper which proposed a method called "Tempel" to fix this, which I may consider using. Rough set technique is a method that has been proposed by another paper, but I am less sure about how it works.

Evaluation metric
The model will be evaluated by the accuracy in which it successfully predicts next-generation sequences.

Final conceptualization
I plan to build a webapp in which the user can view predicted sequences vs actual sequences, with annotations of predicted mutations and other useful information. DeepMind has been developing AlphaFold, a method to predict COVID-19 structure from amino acid sequences. Though pretty unlikely, it would be awesome if this becomes open source, as then I might be able to display the 3D structure of the predictions on the webapp.