**GRAPH EXPLORER WRITE-UP**

**Sanjana Rao, Gina El Nesr, Milind Agarwal**

**Scienceinthe.cloud:**

Scienceinthe.cloud docker demo was run in the notebook provided on the website. The demo took approximately 3 minutes to run and was swift while running. Results were reproducible and the instructions were easy to follow. The demo was beneficial because it wrote out the steps that it was performing in real time.

1. Automated links like the one provided at the end of the notebook so we could see the files referenced to
2. Better explanations the significance of the results from the graphs.
3. Lastly, while running the first cell, the following error resulted that would have been nice to suppress:

/usr/local/lib/python2.7/dist-packages/matplotlib/font\_manager.py:273: UserWarning: Matplotlib is building the font cache using fc-list. This may take a moment.  
 warnings.warn('Matplotlib is building the font cache using fc-list. This may take a moment.')

Three things done well

1. The graphs provided at the end were visually appealing
2. It was easy to run the docker and notebook file.
3. Information on the graphs was clearly displayed

**Computing the Mean Connectomes:** In order to find the mean connectomes, the list, g1, must be traversed. It holds the data of each of the individuals. The data is aggregated and then averaged in the end. The example connectome visualization helped show how to portray the data. To find the mean connectomes of males and females, the sex list must be checked for each g1 value to see if the individual is female or male. Then the same method as before is used to find the average graphs based on the genders of the individuals.

**Covariates:** It is important to choose appropriate covariates that are possibly predictive of the outcome under study i.e. the accuracy. Covariates available on the networkx algorithms page are very helpful in terms of calculating features related to graph theory but most of them weren’t applicable to a connectomic study! Choosing features that increase the score the most and are also neurologically relevant was a challenging part of the project.

*Degree Centralit*y = the number of links incident upon a node. It can can be interpreted in terms of the immediate risk of a brain region for catching whatever information is flowing through the neuronal pathways!

*Edges and nodes* = this will tell us about the number of brain regions that we’re looking at and the number of pathways that connect them to each other which can be interpreted as the strength of the network’s connectivity!

*Connectivity* = It asks for the [minimum](https://en.wikipedia.org/wiki/Minimum) number of elements (nodes or edges) that need to be removed to disconnect the remaining nodes from each other. This should be fairly high if a network has a lot of strong connections, and can prove to be a good covariate in our study!

*Density* = The number of connections in the neuronal network is going to closer to the maximal number of connections in a brain region network. Previous research has shown that male graphs usually have higher densities than female graphs, and so this would be a good covariate to implement!

**Classifier:** We tried to implement Naive Bayes, K-means clustering, SVM models but none of them got the accuracy to be greater than 0.5 with the set of covariates we had decided to implement. So we decided to stick to Random Forest Classifier which got the accuracy up to 0.63!