**DS4300 Project Proposal**

Jacqueline Daniel, Zoe Morris, Vivek Kanpa, and Amanda Bell

**The NoSQL database(s) you intend to use:**

We would be interested in using a graph database like neo4j. Not only are we curious about graph databases and wish to explore its tradeoffs and functionality, but it seems best suited to represent the interconnectivity of human disease pathogenicity and microbial neurotransmitter synthesis, which is not always a direct or 1-to-1 relationship.

**A list of data sources (if applicable)**

This link is for a github repo full of links for different public datasets which we could use for our project:

<https://github.com/awesomedata/awesome-public-datasets#biology>

<https://www.hmpdacc.org/hmp/catalog/grid.php?dataset=genomic>

<https://www.nature.com/articles/s41467-020-15457-9#MOESM4>

<https://www.nature.com/articles/s41586-020-2395-5#data-availability>

<https://www.kaggle.com/antaresnyc/human-gut-microbiome-with-asd>

<https://disbiome.ugent.be/experimentoverview>

We could potentially use the GAD dataset too for information about which diseases are associated with different neurotransmitters.

**A high-level description of your application and its significance. In other words, explain  
to me why this is an interesting non-trivial project and something you would enthusiastically want to show to a potential employer.**

The human microbiome is vast and diverse, containing various bacterial species throughout the body that are crucial for disease defense and metabolism. More specifically, recent research has determined that the gut microbiome contains bacteria which can synthesize GABA, serotonin, norepinephrine, dopamine, acetylcholine and melatonin (Carpenter, 2012). These neurotransmitters play important roles in regulating psychological disorders and neurological development.

For our project, we want to investigate the relationship between the volume of bacteria in our microbiome, the neurotransmitters they synthesize, and the diseases caused by malfunctions in these neurotransmitters. A graph database would visualize the relationships and allow deeper analysis of these connections.

Neo4J Structure

* Node: Disease {disease\_name}
* Node: Organism {organism\_name, sample\_name}
* Node: Host {host\_id}
* Edge: Organism associated with disease {num\_publications, method\_name}
* Edge: Organism lives in host {}