Capstone 1 Proposal:

Determining malignant tumors based on topological data

What is the problem you want to solve?

The task of diagnosing a patient for cancers can be an arduous and laboring task on healthcare professionals. Yet, not by coincidence, time is of the essence for the patient as well. The use of machine learning algorithms in healthcare is absolutely an active area of research, especially in medical imaging. The problem here that we plan on solving is to use machine learning techniques to help speed up the decision making process for a diagnosis based on topological data and other data that might be relevant.

Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?

Obviously one who studies cancers for a living most likely could make diagnoses on their own. The idea here is not to be the sole deciding factor in determination, but to be another valuable tool to a doctor making the conclusive statement: Does this patient have cancer or not? The clients for this project, therefore, aren't necessarily the healthcare professionals themselves, but also the patients who are looking for an accurate diagnosis. The use of a number of machine learning techniques would help speed the time it takes to make a diagnosis as well as provide another tool to help make the decision.

What data are you going to use for this? How will you acquire this data?

The data that we are using are coming from possibly multiple sources, and are acquired from public repositories. First, the UCI Breast Cancer Dataset provides us with topological data of the tumor itself from clinical cases. This will ultimately be the main dataset to be used, however we might like to extend our algorithm to include other information such as imaging data or other miscellaneous information. There is another dataset that includes imaging data for tumors as well, found at the Cancer Imaging Archive. Finally, our last source of data might not be used as well, but the Cancer Statistics Center also hosts a wealth of data that could be used. We do note, however, data found at this website are mostly summary statistics, and might not actually be useful in making a diagnosis.

Outline and Deliverables:

I plan on solving my problem presented above by:

- Cleaning the Data, deciding which could be used in making the decision of diagnosis
- Using Machine Learning techniques and Statistics to develop said decision maker

The form of these deliverables will come in the form of Python code, supplemented with a paper/report summarizing the technique used and it's accuracy.