**Capstone Project I Proposals**

Wenbin Liu

1. Lymphography classification using the data set on UCI. Various classification methods can be tried. The beauty of the data set is that there are a decent number of features (19) and there are 4 levels in the class variable, and is within my domain. The drawback is that there are only 148 cases and therefore the total amount of data to be handled might not be “impressive” to employers.
2. Breast cancer clustering using proteomic data (data set is on Kaggle). There are nearly 20000 features in the data. It is not only in my domain but I’m very familiar with the disease. The challenge is that there are so many evaluation methods to choose from, but interesting findings could emerge e.g. more clusters may be found or suggested than the prevailing opinion of the subtypes of this disease, thanks to the large number of features.
3. Predicting breast cancer using mammogram images from normal and cancer breasts. Data from over 1000 cases can be downloaded from Digital Database for Screening Mammography (DDSM). The biggest challenges are (1) quantification of the images because the images are in different sizes and shapes, and (2) the images look rather diverse and the difference between the normal and cancer images look very similar other than the small marked areas on the cancer ones.