Intro to Data Mining Project: **The Strength of Support Vector Machines, Decision Trees, and K-Nearest Neighbor in Image Classification**

Participant Presenters: **Luis Ahumada, Sam Cohen, Rich Gude**

Purpose Statement:
Image processing in data science has many uses in the medical field including the areas of radiography, topography, and other visual testing methods. This study will identify amongst three testing methods, support vector machines, decision trees, and k-nearest neighbor, the best method for image classification by comparing them using a dataset of retinal fundus images.

Dataset:
The dataset under consideration is a collection of ~300 jpeg files showing images of the foveal avascular zone of the retina. The images show patients with a diagnosis of myopia (near-sightedness), diabetes, or neither of the two ("normal"). The images are already presented in black and white, and there is a clear difference in the contrast and shape of the features between the three diagnoses, so little preprocessing may be necessary.

The data is presented from the link below:
https://www.openicpsr.org/openicpsr/project/117543/version/V1/view
And the images are available here:
https://www.dropbox.com/sh/7zlxred89zghhl3/AAArgYPdOvW36xIngBl24cG2a?dl=0

Data Mining Algorithms:
As stated in the purpose statement, this study will explore three methods of data classification, namely support vector machines, decision trees, and k-nearest neighbor. It is expected that these methods will be implemented in their standard forms, with only changes to the preprocessing of image data to improve characterization results.

Metric for Success:
The success of a data science algorithm is measured by two metrics, accuracy of the model and the computational time to perform to the desired results. Some variables within the data science algorithms can be increased, such as with the k variable in k-nearest neighbor, or decreased, such as the gamma variable in support vector machines, indefinitely in order to improve performance of the algorithm, albeit with diminishing returns of added value compared to the increase computational time or load on servers. For the purposes within this study, the best data science method will have the highest model accuracy with variables that keep the computational time equivalent between the three methods.

Project Schedule:
In order to meet the needs of the presentation schedule, this project will be completed within three (3) weeks time, by April 20, 2020.