# Classification of Sloan Digital Sky Survey Objects Using Machine Learning Algorithms

Candice Stauffer and Gabriel Casabona

May 5, 2020

With the advent of technological advancements in the design and implementation of telescopes, astronomers need more sophisticated tools to analyze the immense amounts of data. Machine learning has allowed astronomers to improve on the classification of objects observed in the sky. Algorithms developed through machine learning allow astronomers to dissect up to petabytes of data, in order to learn about the vast mysteries of the universe.

In this project, data will be taken from a Sloan Digital Sky Survey (SDSS) release and processed through machine learning algorithms to classify stars from galaxies. The SDSS is an international collaboration which uses a 2.5-m wide-angle optical telescope at the Apache Point Observatory in New Mexico, US. With this telescope, the SDSS creates a multi-spectral imaging and spectroscopic redshift survey which has so far been able to observe over 1 billion objects.

For the scope of this class, a training data set will be used, since the real data set releases tend to be larger and more difficult to use. The primary goal of this project is to build and implement an algorithm that will distinguish objects as stars or galaxies. Planets and other objects are not included in this training set. This algorithm will begin by separating the images into equal zones using a set arcsecond parameter. It will then try and resolve distinct objects, such that each object emits its own light. Objects will be categorized by their different features given in the SDSS dataset (i.e. absolute magnitudes, PSF Flux, etc). A random forest (RF) classifier can provide which feature(s) is the most important. Based on the most important feature, we can build a model that will predict the classification of the object as either a star or galaxy.

Adding more trees to the RF model can be chosen to optimize the classification and it can finally be tested by withholding a fraction ($\sim 0.2$) of the training set (i.e. the test set) to determine the accuracy of the model predictions on new data.