**Project Name:**

Exploring classification models of astronomical objects from SDSS

**Brief Description:**

The Sloan Digital Sky Survey (SDSS) is a pioneering multi-spectral imaging survey covering a large fraction of the sky (see Figure 1). Initially commissioned to develop a uniform, well-calibrated map to study the large-scale structure of the universe, SDSS data acquisition began in 2000, and has since gone through numerous extension phases, characterizing the universe in an increasingly comprehensive way. It remains operational to this day and has contributed to a wide range of basic science including the stellar composition of the Milky Way, supernovae distributions, expansion properties of the universe, exoplanet discoveries by radial velocity, galaxy structure and evolution, and supermassive black holes. Furthermore, as one of the first ‘big data’ astronomical projects, it has contributed to many advances in data storage, reduction, and accessing technologies.

One of the early SDSS studies endeavored to develop an extensive catalog of galaxies to better understand their number, shape, and distribution in the universe. The deluge of data available, however, necessitated a crowdsourcing image classification project called ‘Galaxy Zoo,’ which was launched in 2007 to assign labels (based on a decision tree intended to specify various morphologies) to several hundred thousand galaxy images (see Figure 3). Those same objects, amongst others, were part of other studies including a general survey that captured spectroscopic and photometric features such as light intensity at under 5 filters spanning ultraviolet to infrared (see Figure 2).

**Team members:**

Ahmad Allaou, Daelyn Bergsman, Spencer Morris

**Problem statement:**

The SDSS database provides a rich variety of data for celestial objects, including 0-d scalars (e.g. photometric brightness under a given color filter), 1-d vectors (e.g. detailed spectrum), and 3-d tensors (e.g. multispectral images). Many of these objects are labeled using different methodologies. The SDSS consortium, for example, has used survey data to classify objects as quasars, stars, or galaxies, and has furthermore derived *broad* probabilistic morphologies (e.g. ‘elliptical’ and ‘clockwise spiral’) for galaxies. The Galaxy Zoo project, meanwhile, has classified *specific* galaxy morphologies (e.g. ‘E1’ and ‘Sba’) based on an image classification decision tree.

**Objectives:**

1) We aim to classify images of galaxies according to the specific morphological labels ascribed by the Galaxy Zoo Project decision tree method.

2) We furthermore plan to evaluate a subset of SDSS survey objects and develop models to classify them as either a quasar, a star, or a galaxy.

3) For those survey objects classified as galaxies, we are interested in whether or not we can classify them based on broad morphological labels (such as ‘elliptical’ or ‘edge-on’), or even the specific morphological labels (such as ‘E1’ or ‘Sba’) available from the Galaxy Zoo project, through any combination of SDSS photometric, spectroscopic, or imagery data.

**Approach / Methodology:**

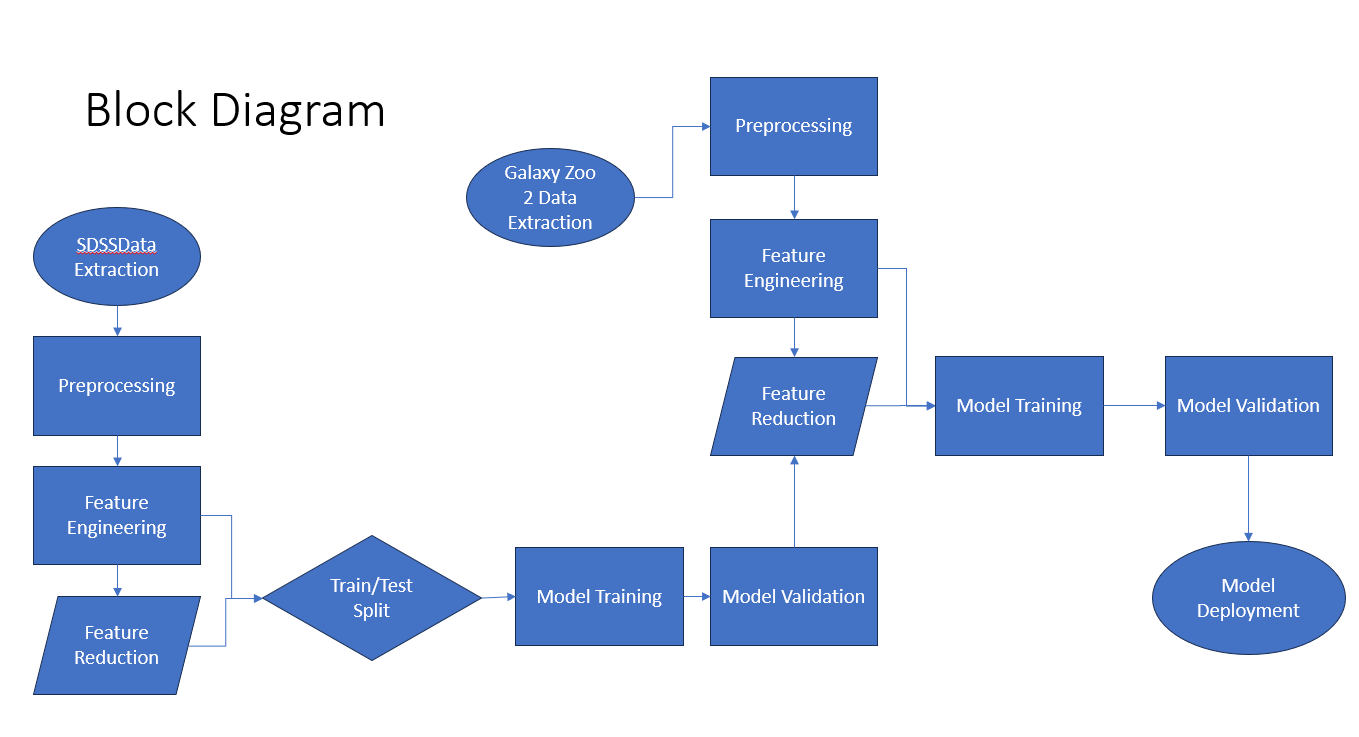
1) For Galaxy Zoo data, we will explore image processing techniques such as compression, reduction, and augmentation. For classification, we will employ convolutional neural networks. Given the large number of classes (over 800) ascribed to specific galaxy morphologies, we will likely develop a mapping schema that projects the original classifications down to a smaller set. If we have the bandwidth, we might investigate how similar hierarchical clustering is to the original decision tree, and might also explore transformer-based models.

2) For survey data, we will investigate k-means clustering and principal component analysis to understand the complexity of the data, employ feature selection to reduce dimensionality of the primary dataset, investigate if any derived features are appropriate, and impose appropriate scaling and transformations. For classification, we will investigate a variety of modeling approaches suitable for tabular data, such as logistic regression, random forest classification, k-nearest neighbors, and feedforward neural networks. We may investigate clustering and principal component analysis to understand the complexity of the data.

3) If we have the bandwidth, we will combine SDSS and Galaxy Zoo datasets and use knowledge gained in training performant models thereof to explore more complex classification models incorporating multiply-shaped data.

For all datasets, random train/validation/test subsets will be generated according to the standard approaches, and we will strive to implement cross-validation for at least one model. We will explore a variety of optimizers, loss functions, neural network depths, learning rates, training epochs, regularization approaches, performance metrics, etc.

**Block Diagram:**



**Datasets (potential):**

In order to ensure data provenance, we will retrieve our datasets from their original sources. The Galaxy Zoo data (including the images, assignment statistics, and final labels) will be obtained from the [project’s website](https://data.galaxyzoo.org/). The photometric and spectroscopic survey data will be queried directly from the [SDSS Skyserver database](https://skyserver.sdss.org/dr1/en/sdss/data/data.asp).

**What is considered success / failure?**

Given the large size of the datasets and the data engineering and computational challenges with developing our first machine learning models, we will be happy if we can develop any models that are reasonably performant. We hope to achieve a high accuracy (at or above 90%) for our first two objectives. We have identified many things we can try, and realistically are unlikely to succeed in (or even attempt) them all given our limited experience and resources – but we’re going to try our best! We will consider the project a success if we can really understand how to apply the modeling approaches we’ve learned in class to an original problem.

**Evaluation parameters (potential):**

For all approaches, we will stratify the datasets as necessary for establishing intuitive performance metrics. We will evaluate models against each other and in general based on loss, accuracy, precision, and recall.

**Figures**

Figure 1 – [SDSS Legacy sky coverage](https://classic.sdss.org/legacy/)

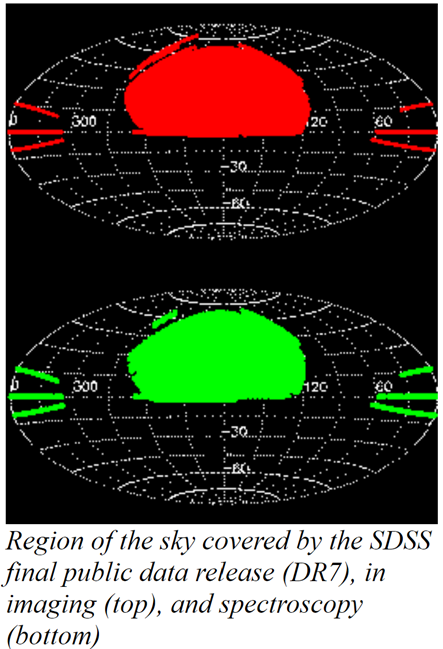


Figure 2 – [SDSS Color Filters](https://skyserver.sdss.org/dr1/en/proj/advanced/color/sdssfilters.asp)

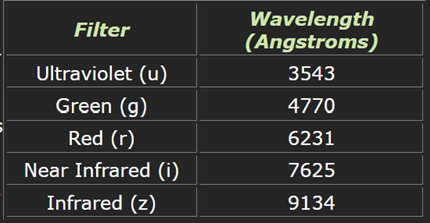


Figure 3 – subset of specific galaxy classifications assigned to Galaxy Zoo objects

