LSST Zooniverse Progress Update (June 2017)

This writeup details the progress of the LSST Zooniverse project up to June 2017. It is presented in order of events and will highlight key milestones.

Project Summary:
In order to meet one of the main science goals of the LSST, we are developing a portion of the data pipeline for real/bogus classification of candidate transient events. Classification must be autonomous and able to move through all candidate transients detected over the lifetime of the LSST. This invokes a need for supervised machine learning. We plan to generate a robust training set for ML using the Zooniverse, a platform for citizen science. This necessitates a concise, visually distinct and well-documented classification schema for citizen scientists to provide a consistent training set.

Project Start (September 2016)
Work began in September 2016, the first assigned task was to get a handle of the Zooniverse and understand its application and limitations for our specific project needs. A basic project was created using the Zooniverse's web-based project builder. Code for pushing data up to the Zooniverse was pushed up to Github as ZooPipe.py

Introducing Data (October 2016-December 2016)
We pulled asteroid tracklets data for use as a test set for the Zooniverse. This data was pushed up to the Zooniverse with a basic classifier in place for classifying each image as "real" or "bogus". In December, the project was presented to the LSST data management group and it became clear that the tracklet data was messy and useless for our specific needs.

Generating Cutouts Manually (January 2017- April 2017)
Taking the feedback from the project presentation in December, we opted to build a testing dataset from the LSST stack itself, taking code from Colin Slater that generated co-adds of asteroid tracklets. We decomposed these co-adds into a template/science/difference format. From here, we spent much time figuring out the correct scaling method for the dataset. We settled on including two scaling options, Arcsinh through the luptonRGB code on Astropy.Visualization and a z-scale method.

First Classification Run (May 2017)
We generated 1000 cutouts for use as a subject set to attempt a trial run of the Zooniverse classifier. We built a Difference Imaging Classifier, which featured a classification schema of Transient Event, Variable Star, Dipole, Subtraction Error, Pixel Artifact, Noise and Other. We set a retirement limit of 5 (each image can be classified 5 times before being declared 'classified'). Within a few days we had ~1600 unique classifications from within the LSST DM group (Thanks Bob Abel!). We produced some statistics on the run in the TrainingSet_Statistics.ipynb jupyter notebook. In short, we found that our classification scheme was unclear and needed refinement. Especially in the distinction between Variable Star, Dipole and Subtraction Error. This will be an ongoing discussion for not only this project but the LSST DM effort in general. We will possible use Twinkles to simulate expected objects and refine classification schema towards them.

Machine Learning (May 2017)
Based on the classifications received from the first run, we built a Random Forest classifier through Sci-kit Learn. Using a 80:20 train-test split, we reported 75-80% agreement with the training set. This is an encouraging result as the classifications being contentious likely is the explanation for the roughly 20% of disagreement between ML and the training set. The Random Forest gave us feature importances

on the roughly 100 continuous features measured by the stack on each candidate event. The next step here will be to attempt to decompose these features into a smaller more representative feature set that allows a physical intuition as to what each classification depends on.