Likely Planet Candidates Identified by Machine Learning Applied to Four Years of Kepler Data

Jon Jenkins[1,3], Sean McCauliff[2,3], Joe Catanzarite[1,3], Joe Twicken[1,3], and Jennifer Campbell[2,3]

[1]SETI Institute, [2]Orbital Sciences corporation, [3]NASA Ames Research Center

Abstract

Over 3600 transiting planet candidates, 156 confirmed planets, and ~2,400 eclipsing binaries have been identified by the Kepler Science pipeline since launch in March 2009. Compiling the list of candidates is an intensive manual effort as over 18,000 transit-like signatures are identified for a run across 34 months. The vast majority are caused by artifacts that mimic transits. While the pipeline provides diagnostics that can reduce the initial list down to ~5,000 light curves, this effort can reject valid planetary candidates. The large number of diagnostics (~100) makes it difficult to examine all available information. The effort required for vetting all threshold-crossing events (TCEs) takes several months by many individuals associated with the Kepler Threshold Crossing Event Review Team (TCERT).

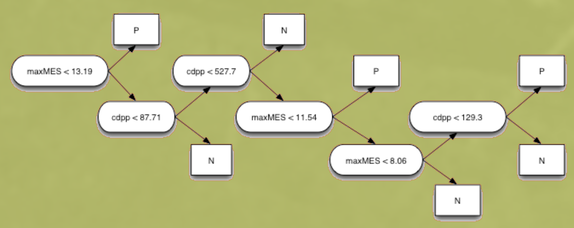
We have developed a random-forest classifier that classifies each TCE as `planet candidate', `astrophysical false positive', or `non-Keplerian phenomena'. Ideally the algorithm will generate a list of candidates that approximates those generated by human review, thereby allowing the humans to focus on the most interesting cases. By using a machine learning-based auto-vetting process, we have the opportunity to identify the most important metrics and diagnostics for separating signatures of transiting planets and eclipsing binaries from instrument-induced features, thereby improving the efficiency of the manual effort.

A random forest trained on the Q1-Q12 TCERT results was applied to Q1-Q16 TCEs (Tenenbaum et al. 2013). A total of 1312 likely new planet candidates are identified, with 452 smaller than 2 Re. In addition, there are 1220 likely new astrophysical false positives. We present characteristics of the likely planet candidates identified by the auto-vetter as well as those objects classified as astrophysical false positives (eclipsing binaries and background eclipsing binaries). We examine the auto-vetter's performance through receiver operating characteristic curves for each of three classes: planet candidate, astrophysical false positive, and non-transiting phenomena.

Funding for this mission is provided by NASA's Science Mission Directorate.

1. A Falling Tree in a Random Forest

A random forest is a supervised machine learning algorithm that grows a forest of Classification and Regression Trees (CART) to classify objects based on a set of attributes for each object (Breiman 2001). Each CART is trained on a subset of the training set (a set of objects with known classifications), and on a subset of the available attributes and the trees are "grown" iteratively in order to minimize the misclassifications of the training data. A random forest will typically have 1000s of decision trees. These key design aspects of the algorithm yield a automated classifier that is robust against over-fitting, against errors in the input training data, and against missing information in the input set. The importance of each attribute can be measured by scrambling that attribute in the training set and retraining to measure the degradation in the performance. The random forest records the number of votes for each class for each object, allowing the results to be "scored" according to credibility of the classification for each category. Once a random forest is trained, it can be applied to new and unknown objects.



A sample decision tree from the Q1-Q12 random forest classifier that uses the maximum multiple event statistic (SNR) and the photometric noise on transit time scales (CDPP) to classify Threshold Crossing Events.

2. Training Results for Q1-Q12 and for Q1-Q16

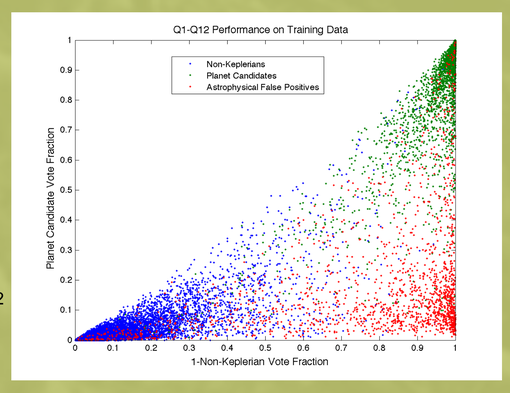
The graphics below show the results of training our random forests for Q1-Q12 and for Q1-Q16 based on the results of the TCERT process for Q1-Q12 (Rowe et al. in prep). The first step is to correlate the TCEs identified in a run of the *Kepler* transit search pipeline (TPS/DV) with exisiting planet candidates (and planets), astrophysical false positives, and to identify a set of false positives from artifacts (non-Keplerians). The training for Q1-Q12 included 3,425 planet candidates, 1,698 astrophysical false positives, and 11,301 false positives from artifacts. The Q1-Q12 non-Keplerians were the TCEs that failed the first step of the TCERT vetting process ("triage"). The training for Q1-Q16 included 3,193 planet candidates, 1,143 astrophysical false positives, and 1,110 artifacts that were a subset of the 11,301 identified in the Q1-Q12 TCEs. The behavior of the artifacts yielded by the Q1-Q16 TPS/DV run differed markedly from that of the Q1-Q12 data set, due to significant changes in the TPS and DV code, especially as to internal vetting logic applied in TPS (Seader 2013, Tenenbaum 2013). The set of artifacts that were re-identified in the Q1-Q16 run was insufficient, and we added a set of 100 objects chosen at random from a large cluster of "long period artifacts" to improve the performance.

We illustrate the performance of the Q1-Q12 RF and the Q1-Q16 RF on the training sets below. Note the improved performance for Q1-Q16.

Q1-Q12

True Class	Classification PC	AFP	NK
PC	3275	82	68
AFP	207	1161	330
NK	60	102	11139

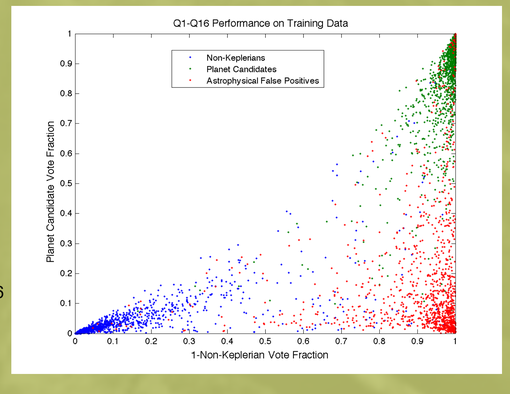
Misclassification rate = 0.051
PC False Discovery Rate = 0.082



Q1-Q16

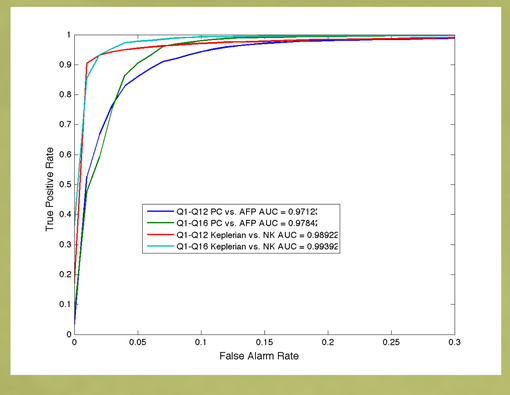
True Class	Classification PC	AFP	NK
PC	3150	34	9
AFP	126	975	42
NK	18	72	1120

Misclassification rate = 0.054
PC False Discovery Rate = 0.046



3. Comparing Q1-Q12 and Q1-Q16 Performance

The graph below shows the Receiver Operating Characteristic (ROC) curves for both random forests. In general, the Q1-Q16 RF performs better than the Q1-Q12 RF on the training data. In both cases, however, the ability of the auto-vetter to distinguish between artifacts and astrophysical objects is quite good, with areas under the curve (AUC) of ~0.99. It is harder for the autovetter to distinguish between planet candidates and astrophysical false positives, but the AUCs are still high, ~.98.



4. New Results for Q1-Q16

In this section we examine the new results from Q1-Q16. The graphic below shows the results of applying the RF to the new and unknown TCEs from the Q1-Q16 run.

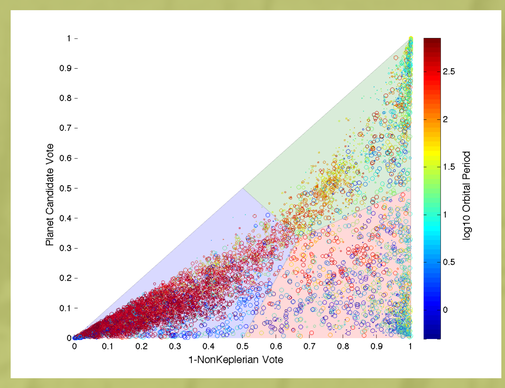


Figure 4a. Auto-vetting results for the 10,790 unknown TCEs identified in the Q1-Q16 run. The classification regions for PC, AFP and NK are indicated by the green, blue and red backgrounds partitioning the triangle. The points are colored by the $\log_{10}$(orbital period). There is a dense cluster of long period artifacts in the non-Keplerian "triant".

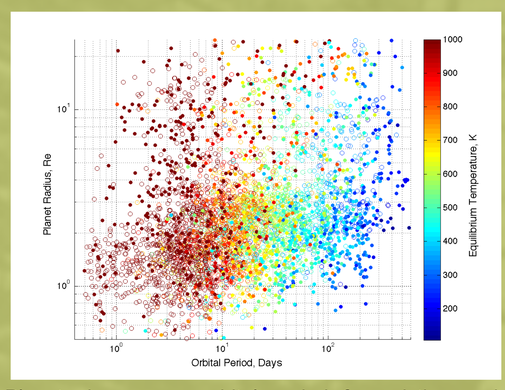


Figure 4b. Planet size versus orbital period for existing and likely planet candidates, colored by equilibrium temperature (max of 1000 K). Likely planet candidates identified in the Q1-Q16 run are indicated by the red filled circles. Planet candidates identified through Q1-Q12 are presented in blue. The new candidates extend the orbital periods spanned by the Kepler Mission as well as pushing down the size of the likely candidates.

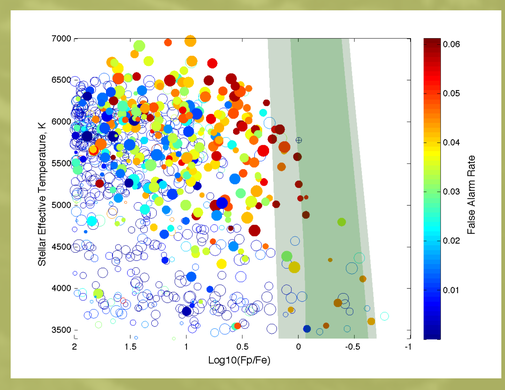
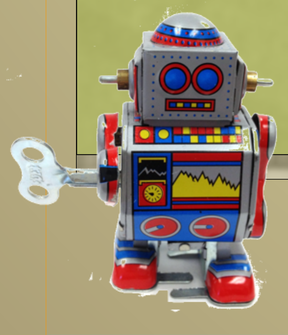
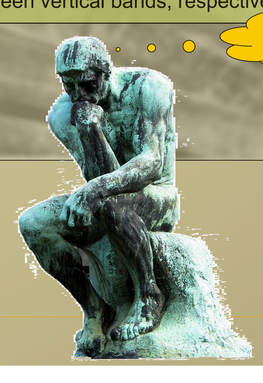


Figure 4c. Relative insolation for the 1,945 likely planet candidates smaller than 2 Re as a function of stellar effective temperature are plotted as circles whose size is proportional to the planet size. The circles are colored according to the false alarm rate predicted by the random forest for these candidates based on the training data. Planet candidates identified up through Q1-Q12 are indicated by unfilled circles. 452 new likely candidates identified in the Q1-Q16 data are indicated by filled circles. The "wide" and "narrow" habitable zones are indicated by the light and dark green vertical bands, respectively.





Leonid Kulik Expedition

Discovery Mission 10 - Launched 2009 - http://Kepler.NASA.gov