# Creating an Automated Classifier to Approximate the By-eye Selection of HAT Exoplanet Candidates

**Will Coulton**
Department of Physics
Princeton University
wcoulton@princeton.edu

**Joshua Wallace**
Department of Astrophysical Sciences
Princeton University
joshuajw@princeton.edu

## Abstract

A proposal to use machine learning techniques to aid in the binary classification of data from the Hungarian Automated Telescope (HAT) arrays. The HAT arrays are designed to detect periodic dimmings of stars, potentially caused by orbiting planets around those stars. There are many signals, both physical and spurious, that can approximate planet signals, so it is important to weed out as many false positives as possible. After an automated pipeline removes probable false planets, a final manual by-eye selection is made to create the final set of planet cadidates. We propose to train classifier models to approximate this by-eye selection with the hope of finding a sufficiently well-performing model that can stand as a proxy for or even entirely replace the manual portion of the selection pipeline.

## 1 Introduction

The field of exoplanets ("extra-solar planets", or planets around other stars) has established itself as the hip new field in astrophysics. Since the discovery of the first exoplanets in the early 90's [11], nearly 3000 exoplanets have been discovered [6], with potentially thousands more exoplanets remaining unconfirmed or thus far undiscovered in existent data. This explosion in exoplanet discovery has been fueled by the funding and construction of many exoplanet detection surveys. The majority of exoplanets have been found by the space-based *Kepler* telescope and its primary and "K2" surveys [3], [9]. Ground-based surveys, while not as sensitive to smaller planets as *Kepler* (due to atmospheric distortion of image quality and other problems associated with observing from the ground), are sufficiently sensitive to discover Jupiter-sized planets on short-period orbits. Such planets are called "hot Jupiters" because they are the same size as Jupiter but are sufficiently close in to the stars they orbit (much closer than Mercury is to our sun) that the star heats them up to several thousands of degrees Celsius.

Since ground-based surveys are much cheaper to run per area of sky monitored than spaced-based surveys, the majority of known hot Jupiters have been discovered by large-scale ground-based survesy. Of the few hundred hot Jupiters that have been found to date, the plurality ($\sim$100) have been found by Princeton's own Hungarian Automated Telescope (HAT) collaboration, headed by Prof. Gaspar Bakos of the Department of Astrophysical Sciences. This collaboration runs two surveys. The longest-running of the two surveys is HATNet, a collection of five telescopes in Arizona and two telescopes in Hawaii [7]. The other survey is HATSouth, which consists of three locations with eight telescopes each: Chile, Namibia, and Australia [2]. A third survey, HATPI, which will consist of 63 telescopes in Chile, is currently being constructed [4]. The telescopes themselves are small as telescopes go: only about 10 cm across, they are more similar to camera lenses than the "normal" telescopes usually associated with astronomy. Figure 1 shows a picture of four of the HAT-South telescopes.

Figure 1: The four HATSouth telescopes at Siding Springs Observatory, Australia. Photo credit Gaspar Bakos. From [5].

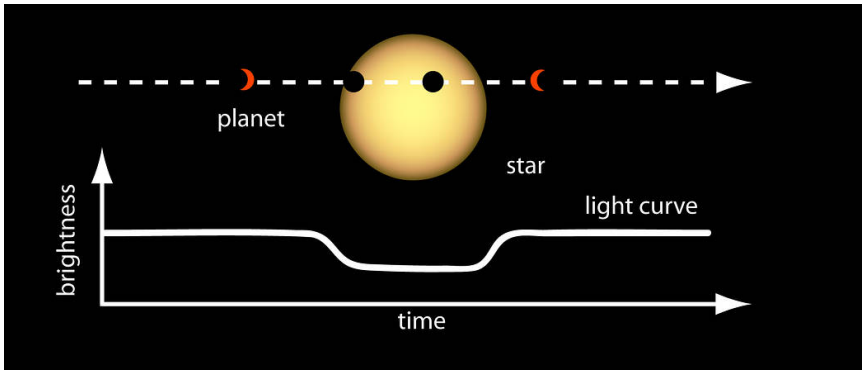


Figure 2: An example "light curve" (brightness as a function of time) from a star that hosts a transiting exoplanet. As the planet crosses in front of the star, the measured brightness from the star is less than the out-of-transit value. Partial eclipses as the planet is transitioning from being out-of-transit to fully in transit put a slope in the light curve between in-transit and out-of-transit magnitudes. Public domain image from [1].

The detection technique used by HAT*, *Kepler*, and many other exoplanet surveys is the "transit" technique. Planets are far too faint compared to their host stars to be able to detect directly around most stars. Instead, in the transit technique, the brightness of stars is monitored as a function of time (usually by taking periodic long-exposure photographs of large areas of the sky). When a planet crosses in front the star it orbits, the amount of light detected from that star is decreased. This is shown in Figure 2. Transits are found in data using matched filtering (with either a box- or trapezoid-shaped filter, matching the signal from a transiting exoplanet), folding the data on a variety of periods and phases. When a star is found to have periodic dips in brightness, it is labeled as a potentially planet-hosting star. However, transiting planets are not the only source of periodic signals like that seen in Figure 2. Starspots rotating in and out of view also provide periodic dips in brightness and binary stars that eclipse each other in a grazing way (so the amount of light that goes missing during an eclipse is about the same as a planet) are two examples of astrophysical false positive signals. Various cuts are made to help ensure that a prospective planet is neither one of these false positives nor a spurious signal. The various parameters of the fit to the transit signal (depth, time of ingress, length of transit) are often good indicators to separate between planet and not. These cuts are automated. However, after the cuts, there are still false positives that are not obvious to the pipeline but are obvious to a trained human eye. Thus, the final step in the HAT pipeline is a manual, by-eye examination of the signal. Planet candidates that pass this step then are sent for follow up (and hopefully confirmation!) at bigger telescopes.

Other than this manual step, the vetting of possible planet candidates is completely automatic. This single manual step is quite labor intensive, and prevents a fully automatic characterization of planet detection efficiency, which is necessary for the calculation of statistics related to the occurrence rate of these planets. Thus, it would be nice to have an automated approximation to the by-eye work that has occurred. This is the purpose of this project.

## 2 Related Work

The premier (and only statistically robust) calculation of a hot Jupiter occurrence rate from transit data is that from the primary *Kepler* mission [8]. Because of the exquisite quality of the data from the *Kepler* telescope, their detection efficiency for hot Jupiters is near 100% and they are better able to automatically identify false positives than the HAT collaboration is from our surveys. Because of this, the *Kepler* pipeline needs much less manual vetting of planet candidates and thus the *Kepler* team has not needed to worry about the manual step as teh HAT collaboration does. However, as mentioned before, the HAT collaboration has found more hot Jupiters than any other collaboration, and so if we can get a good handle on our detection efficiency we have the potential to provide the most precise measure of the occurrence rate of these planets, which will help inform planet formation theory.

Unfortunately, since the *Kepler* work did not need to worry so much about detection efficiencies and pipeline misclassifications, they did not develop the hard statistical/machine learning tools that would be useful for this work. However, a recent astrophysics study [10] that used a very similar time-series data set as the HAT surveys used a random forest classifier to classify 450,000 eclipsing binaries (binary stars that transit each other, similar to how a planet transits a star) from their data.

## 3 Data

The data used are unpublished and (currently) proprietary. They were extracted from the HAT database using a script provided by Joel Hartman[1], Research Scienctist at Princeton University. The database itself currently has no reference.

The data consists of 1024 positive cases and 29,421 negative cases of potential planet candidates that passed the manual vetting process. The number of positive cases is perhaps too low to train a truly robust model, but it is unfortunately all we have to work with. The process of discovering new planets is one of constant attrition, and there are unfortunately more cases of false positive signals than of real signals. There are 92 features that consist of such measured quantities as the fit parameters to the transit dip (transit depth, length, etc.), quantities related to the power spectrum from the matched filtering, measured quantities for the star (brightness in various colors), and signal to noise measures. All of these quantities are continuous, not categorical. There are some missing values: 284 "inf" (0.01% of all values) and 4461 "nan" (0.2% of all values). The "inf" values are mostly concentrated in a single column (feature) of the data, while the "nan" values tend to cluster in the same rows (objects).

## 4 Methods

Our first approach to this problem will be to investigate the simple binary classifiers that we have covered in class so far. These include Naive Baye's, random forests, support vector machine (SVM), and logistic regression. With logistic regression will investigate two types of regularization using $l_1$ and $l_2$ penalties. We will also explore a set of feature selection methods such as mutual information.

Whilst our main goal is to predict which objects would be manually selected, we are also interested in running a set of unsupervised learning algorithms to view any hidden structure in the data set. These methods will include K-means and gaussian mixture models. This will be useful for potentially identifying new approaches to automating this process as well as potentially probing some new relationships.

[1] jhartman@astro.princeton.edu

# 5 Results

# 6 Discussion and Conclusion


### Acknowledgments

We are grateful to Joel Hartman for providing a script to extract the necessary data from the HAT database. We are also grateful to the entire HAT team (PI: Gaspar Bakos) for their many dedicated years of constant observations and the huge pile of astronomical data they've collected.


# References


[1] NASA Ames. Light curve of a planet transiting its star. `https://www.nasa.gov/mission_pages/kepler/multimedia/images/transit-light-curve.html`. Accessed: 2017-04-27.

[2] G Bakos, C Afonso, T Henning, A Jordán, M Holman, RW Noyes, PD Sackett, D Sasselov, Gábor Kovács, Z Csubry, et al. Hat-south: a global network of southern hemisphere automated telescopes to detect transiting exoplanets. *Proceedings of the International Astronomical Union*, 4(S253):354–357, 2008.

[3] William J Borucki, David Koch, Gibor Basri, Natalie Batalha, Timothy Brown, Douglas Caldwell, John Caldwell, Jørgen Christensen-Dalsgaard, William D Cochran, Edna DeVore, et al. Kepler planet-detection mission: introduction and first results. *Science*, 327(5968):977–980, 2010.

[4] HATPI collaboration. The hatpi project. `https://hatpi.org/`. Accessed: 2017-04-27.

[5] HATSouth collaboration. The hatsouth exoplanet survey. `https://hatsouth.org/operations.html`. Accessed: 2017-04-27.

[6] exoplanets.org. Exoplanets.org main page. `http://exoplanets.org/`. Accessed: 2017-04-21.

[7] JD Hartman, G Bakos, KZ Stanek, and RW Noyes. Hatnet variability survey in the high stellar density kepler field with millimagnitude image subtraction photometry. *The Astronomical Journal*, 128(4):1761, 2004.

[8] Andrew W Howard, Geoffrey W Marcy, Stephen T Bryson, Jon M Jenkins, Jason F Rowe, Natalie M Batalha, William J Borucki, David G Koch, Edward W Dunham, Thomas N Gautier III, et al. Planet occurrence within 0.25 au of solar-type stars from kepler. *The Astrophysical Journal Supplement Series*, 201(2):15, 2012.

[9] Steve B Howell, Charlie Sobeck, Michael Haas, Martin Still, Thomas Barclay, Fergal Mullally, John Troeltzsch, Suzanne Aigrain, Stephen T Bryson, Doug Caldwell, et al. The k2 mission: Characterization and early results. *Publications of the Astronomical Society of the Pacific*, 126(938):398, 2014.

[10] I Soszyński, M Pawlak, P Pietrukowicz, A Udalski, MK Szymański, Ł Wyrzykowski, K Ulaczyk, R Poleski, S Kozlowski, DM Skowron, et al. The ogle collection of variable stars. over 450 000 eclipsing and ellipsoidal binary systems toward the galactic bulge. *arXiv preprint arXiv:1701.03105*, 2017.

[11] A. Wolszczan and D. A. Frail. A planetary system around the millisecond pulsar PSR1257 + 12. *Nature*, 355:145–147, January 1992.