



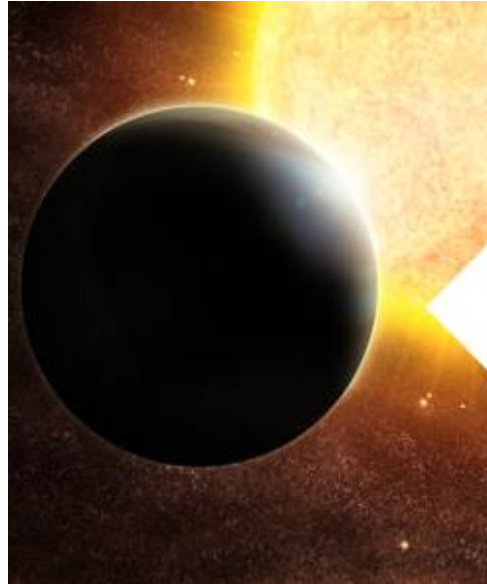
Exoplanet Search

Using Machine Learning to Identify
Exoplanets from Kepler Starship Observations

Dimitri Kourouniotis 2018

How Many Exoplanets Are There?

The search for habitable planets outside our solar system has fascinated us for some time. Once theorized and now confirmed that there are many exoplanets (3,700 and counting), the search for life outside our solar system is now in full swing. Over 2,300 of them from Kepler spaceship.



Exoplanet Count

Kepler:

Candidates: 2,244

Confirmed: 2,327

Small Habitable Zone Confirmed: 30

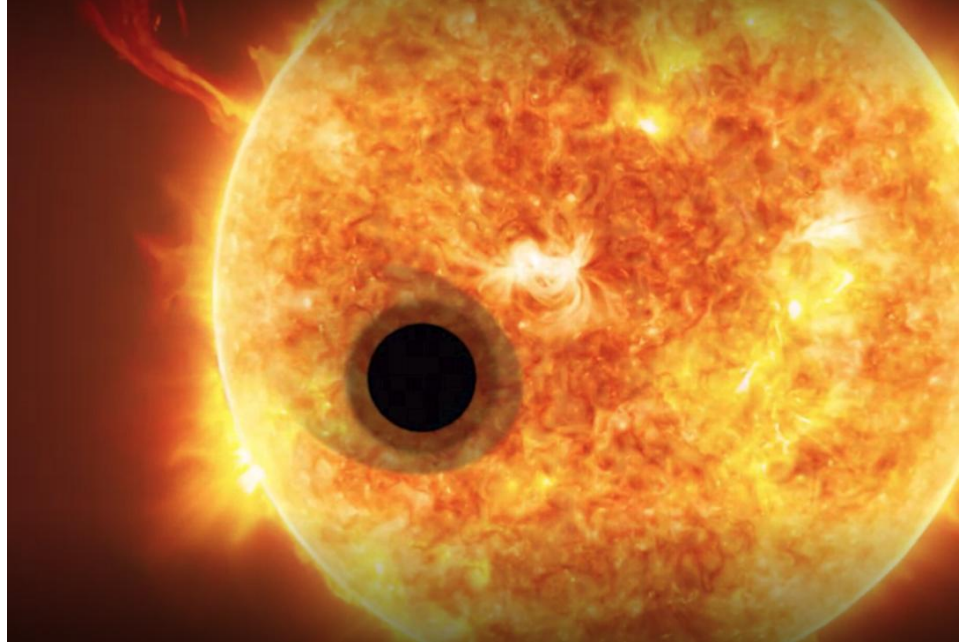
K2:

Candidates: 480

Confirmed: 292

Exoplanets can be identified for the Scientific Community

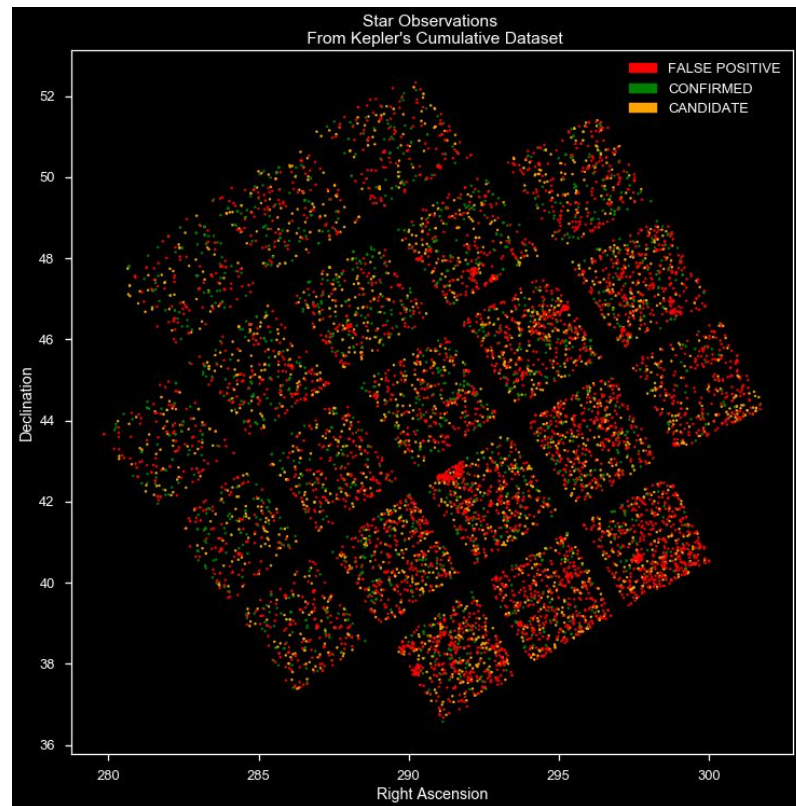
The Transit method identifies possible planets crossing the path of stars from the Kepler camera's point of view. This object of interest can then be directed to other telescopes for further analysis to verify one or more exoplanets. Other details such as chemical composition of the star and by inference the chemical composition and possible atmosphere of the exoplanet can also be determined.



Data from Kepler 'Hunt for Exoplanets' Kaggle competition

Kepler has scanned nearly 10,000 stars over the course of its campaigns, which are ending in the summer of 2018 as it runs out of fuel to send back data.

This dataset contains over 5500 stars with 42 confirmed exoplanets. It is a time series with 3198 measurements of light intensity at 30 minute intervals (80 days). The data had been cleaned for the competition to remove known artifacts from the Kepler camera. For the competition it was split into a training and test set with confirmed exoplanets of 37 and 5 respectively.



Exploratory Data Analysis

With the large range of flux in light intensity I normalized the data.

An exoplanet(s) transiting the star from our point of view will cause the light to dim. Depending on the size of the object, the angle, and duration the light will dim by a certain amount for a certain period. The spaceship is effectively “detecting a flea crossing the beam of a headlight several miles away”. Changes in intensity of the light from a star can also be due to solar flares, sunspots or the rotation of the star.

A single dimming over the 80 day period may be a slower orbiting planet or other star activity. Two low intensity readings provide no additional information as they may not be related to each other, but three dimming equally spaced apart are a strong contender for an exoplanet.

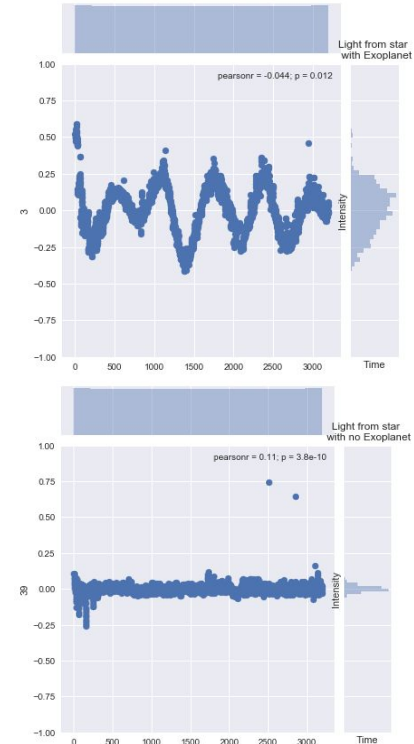
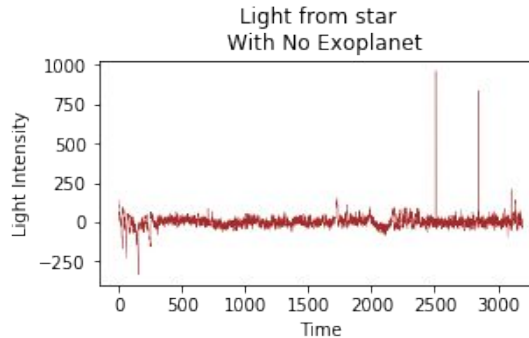
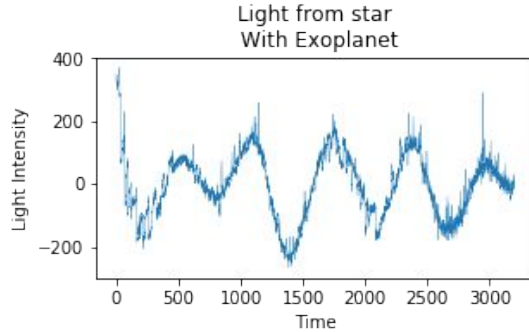
Light Intensity Variation by Star

| Star | FLUX.1 | FLUX.3 | FLUX.3 | FLUX.4 | FLUX.5 | FLUX.6 | FLUX.7 | FLUX.8 | FLUX.9 |
|------|---------|---------|---------|---------|--------|---------|---------|---------|---------|
| 37 | -141.22 | -81.79 | -52.28 | -32.45 | -1.55 | -35.61 | -23.28 | 19.45 | 53.11 |
| 38 | -35.62 | -28.55 | -27.29 | -28.94 | -15.13 | -51.06 | 2.67 | -5.21 | 9.67 |
| 39 | 142.40 | 137.03 | 93.65 | 105.64 | 98.22 | 99.06 | 86.40 | 60.78 | 45.18 |
| 40 | -167.02 | -137.65 | -150.05 | -136.85 | -98.73 | -103.14 | -107.70 | -123.19 | -125.65 |
| 41 | 207.74 | 223.60 | 246.15 | 224.06 | 210.77 | 189.56 | 172.68 | 170.31 | 148.79 |

Time Series and Density Plot comparisons

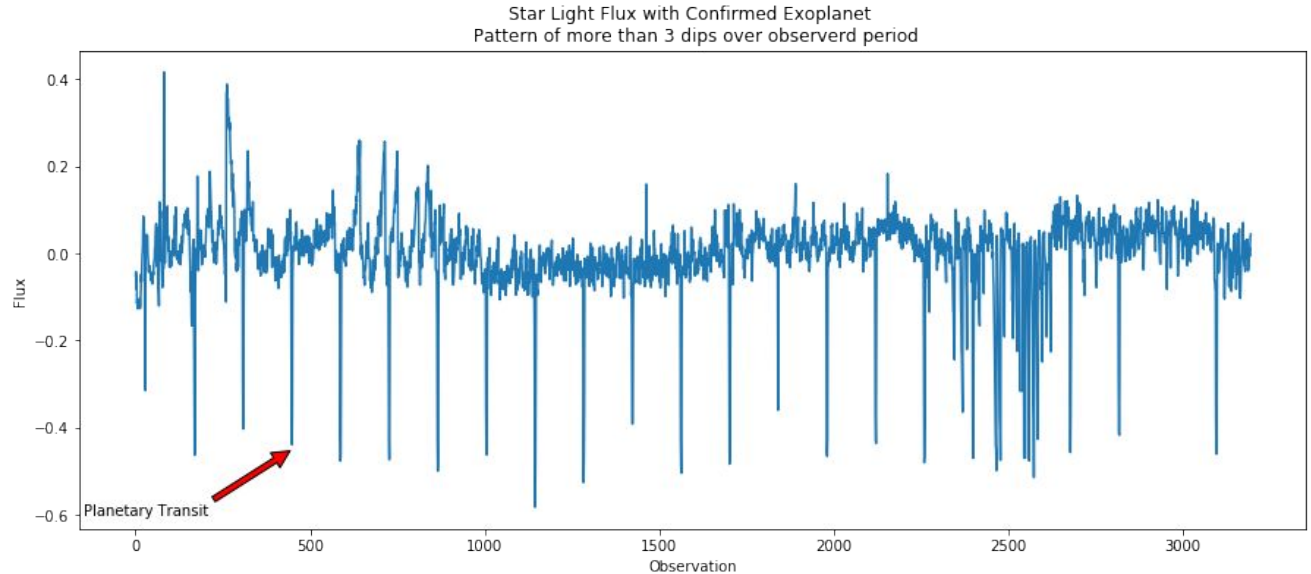


Stars with exoplanets have a slightly different distribution (closer to two peaks) than those without, which have a clear single high peak. The peaks and troughs of the time series are also more distinct.



Time Series and Density Plot comparisons

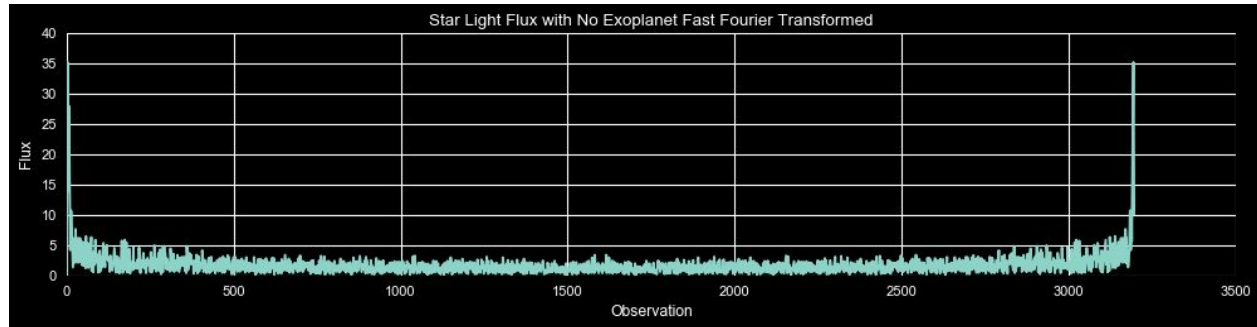
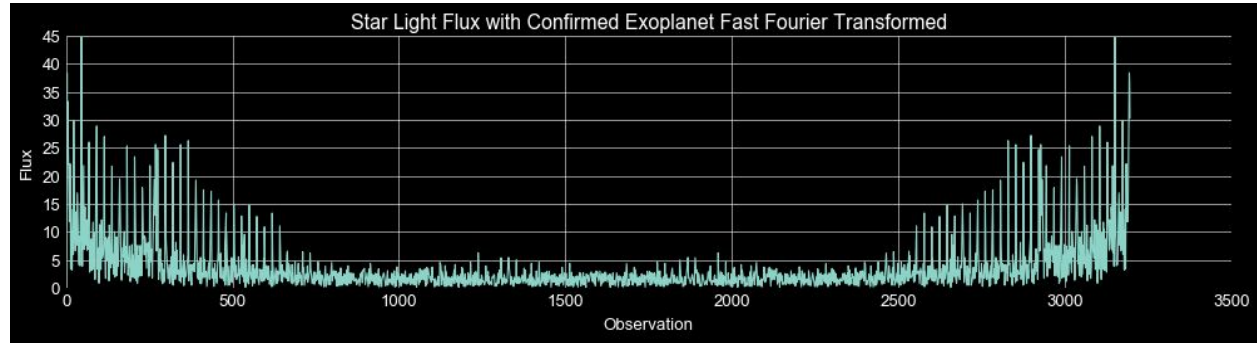
The plot of this star's light shows 23 dips/transits over 80 days, which equates to an exoplanet with an orbital period of 3-4 days. The size of the dip also indicates a large very (and fast!) planet.



Applying Fast Fourier Transformation

Applying the Fast Fourier Transformation allows the signal to be broken down to its visible frequencies.

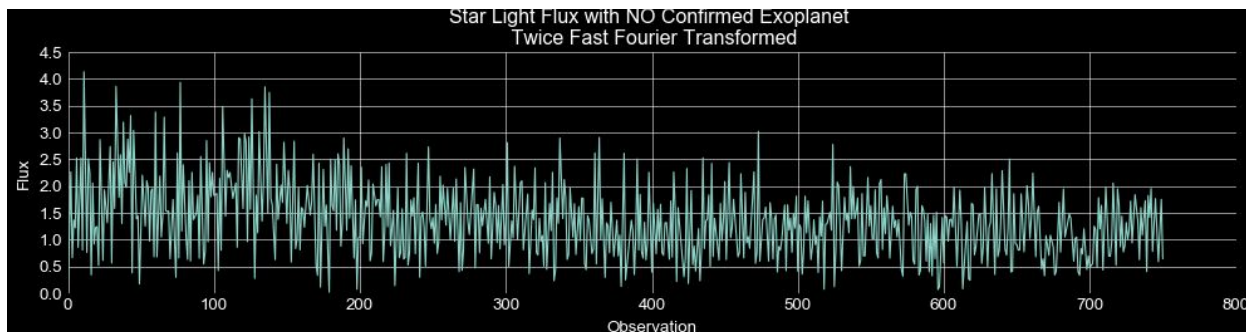
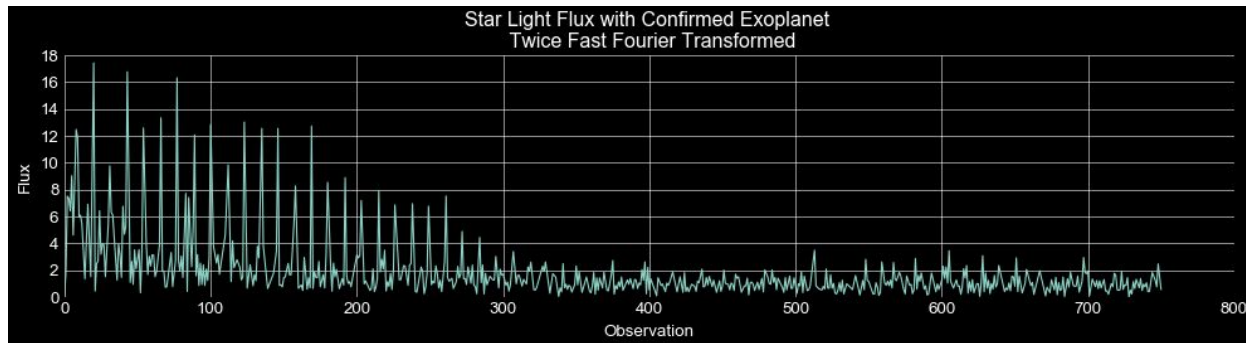
The first peaks are the overall indicator of the frequency of the series. Subsequent harmonics (peaks) indicate the presence of an exoplanet.



Removing Symmetrical Data and Initial Outlier Data

To reduce the number of features I removed half of them leaving ~1600. In addition to see if more information could be highlighted I re-applied the Fast Fourier Transformation and culled half the features again.

Since the first readings are the basic frequency of the star, I dropped those 50 features. This leaves a dataset of only 750 features down from 3178.



Gridsearch Using 6 Classifiers



I used six models and their hyper-parameters range for the Gridsearch.


I used SMOTE (Synthetic Minority Oversampling Technique) to counter the imbalance of the labeled planets.

The scoring I thought was most important was recall to get the correct number of True Positives for exoplanets.

- Random Forest
- AdaBoost
- XGBoost
- Decision Tree
- K Neighbors
- Bagging

Gridsearch Results Sorted by Mean Recall Score

Highest mean score is KNeighbors



| | estimator | min score | max score | mean score | max depth | min samples leaf | min samples split | n estimators | n neighbors |
|----|---------------|-----------|-----------|--------------|-----------|------------------|-------------------|--------------|-------------|
| 69 | KN | 0.538 | 0.833 | 0.707 | N/A | N/A | N/A | N/A | 7 |
| 68 | KN | 0.461 | 0.833 | 0.681 | N/A | N/A | N/A | N/A | 5 |
| 67 | KN | 0.384 | 0.75 | 0.600 | N/A | N/A | N/A | N/A | 3 |
| 49 | AdaBoost | 0.153 | 0.667 | 0.384 | N/A | N/A | N/A | 18 | N/A |
| 11 | Random Forest | 0.153 | 0.667 | 0.384 | 3 | 10 | 15 | 30 | N/A |

Recall Score 80%

The strongest mean score was K Neighbors

Running the model with Neighbors = 7:

Recall score: 80.00%

F1 score: 14.55%

Classification Report: KNeighbors

| | Precision | Recall | f1-score | Support |
|-------------------------|-----------|--------|----------|---------|
| No Exoplanet | 1.00 | 0.92 | 0.96 | 565 |
| Exoplanet | 0.08 | 0.80 | 0.15 | 5 |
| avg / total | 0.99 | 0.92 | 0.95 | 570 |