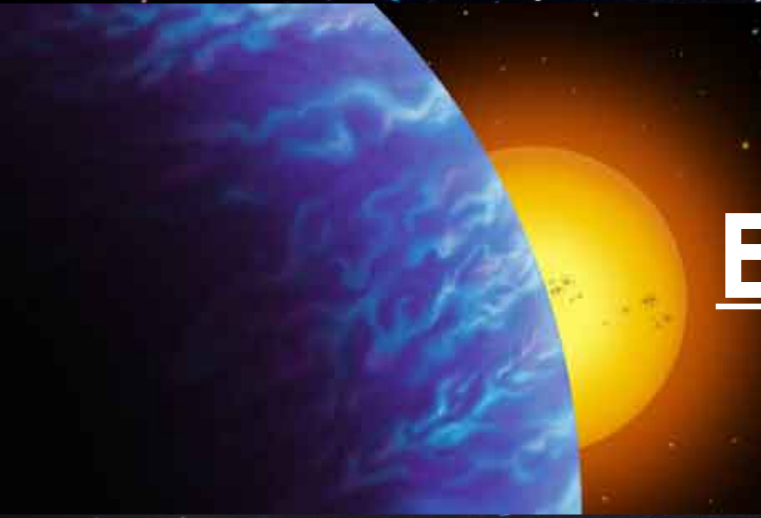
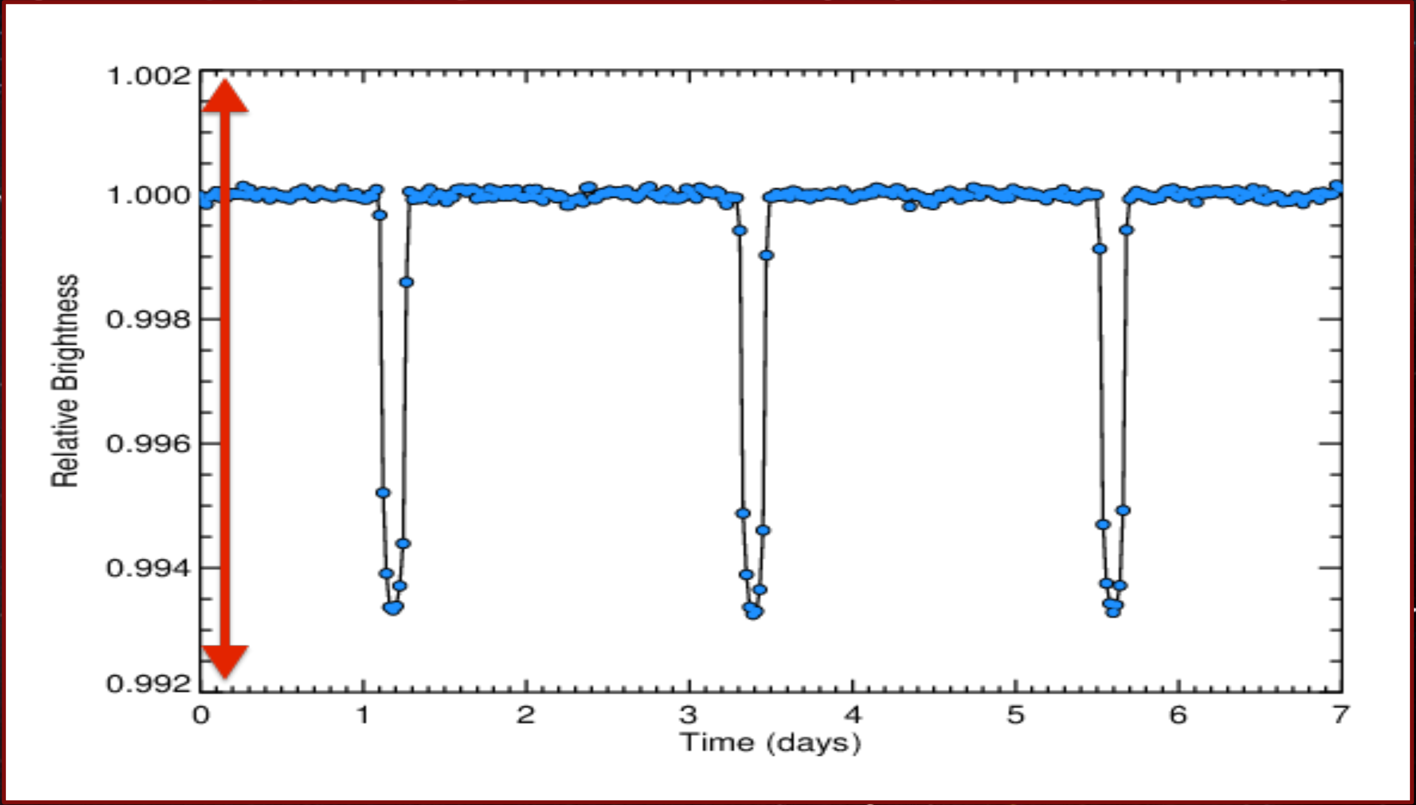


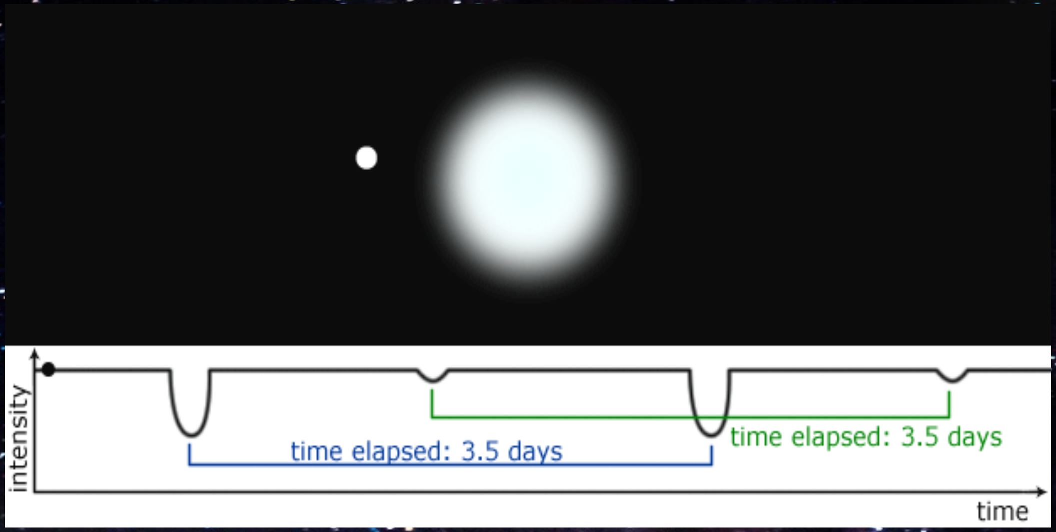
Exoplanet Detection

Current Status & Problem

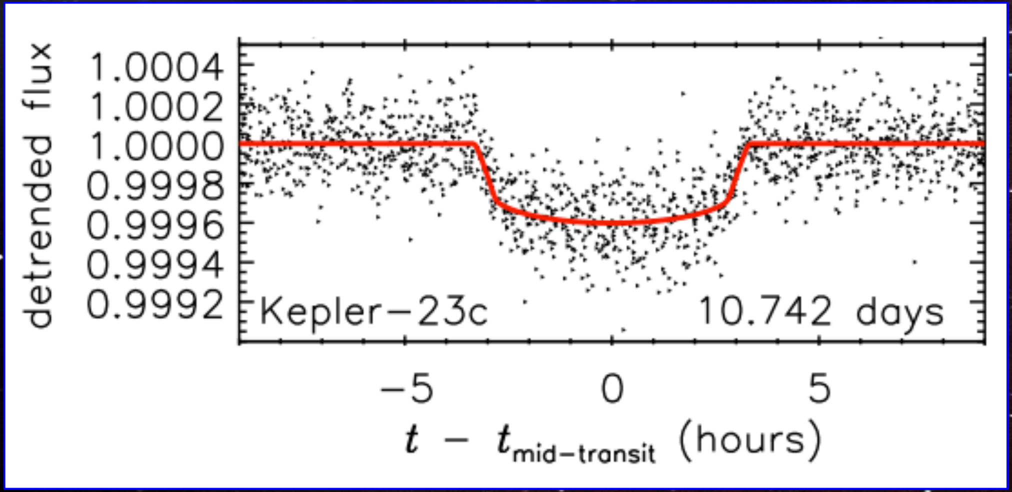
More than 1000 confirmed exoplanets
3000 unconfimed candidates
Noise in the Flux data has led to high false-positive rates for detecting transits
Need to **reduce the error rate** in exoplanet candidate identification.



Transit Method



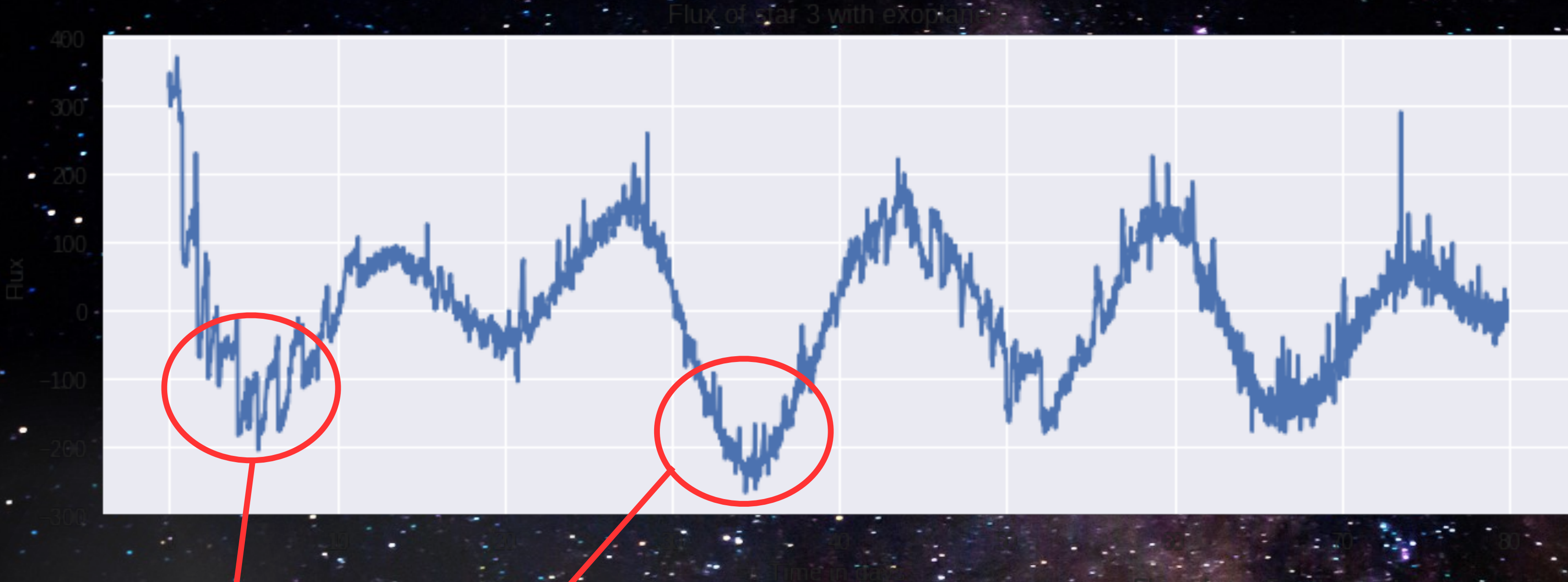
Transit - The planet is in front of the star, eclipsing the star.
Occultation - The planet is behind the star, being eclipsed by the star.



This method detects distant planets by measuring the **minute dimming of a star** as an orbiting planet passes between it and the Earth.

Initial Data Exploration

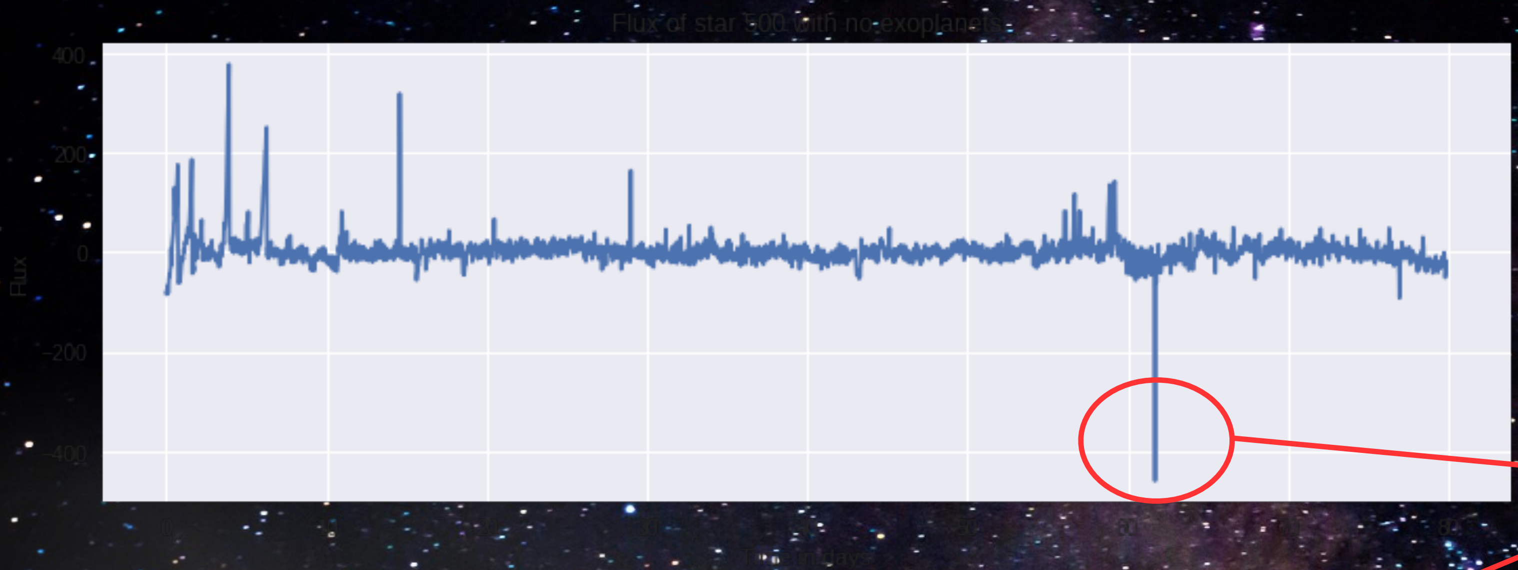
Flux of stars with Confirmed Exoplanets



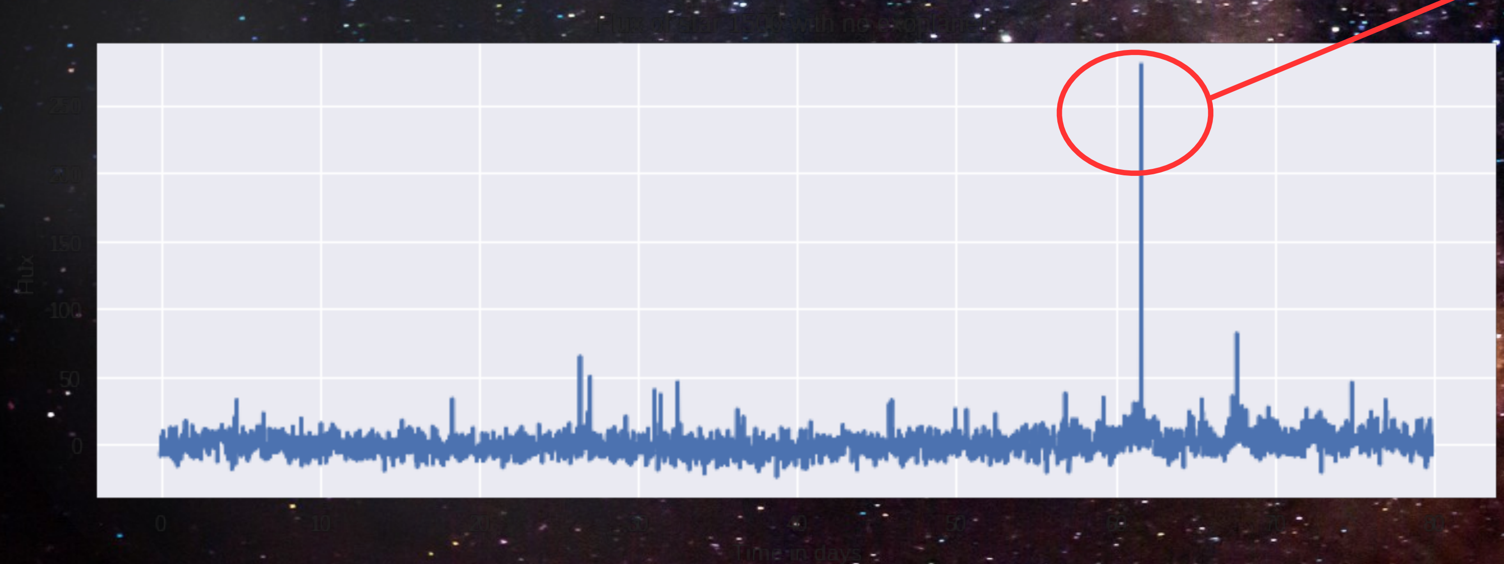
Dips Observed Due to Change in Brightness

Initial Data Exploration

Flux of stars with **Confirmed lack of** Exoplanets



No significant Dips
Only **Outliers** observed

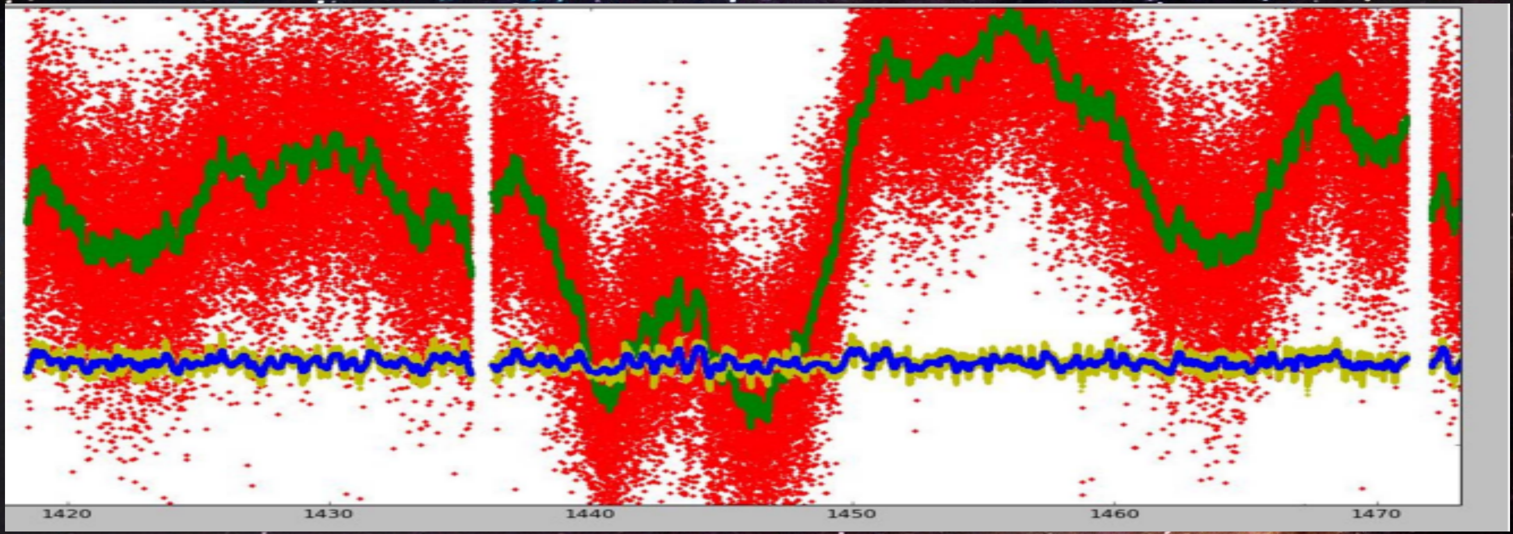


Detrending

Detrending the Data

Applying Gaussian Filter (5σ)

Normalizing the data

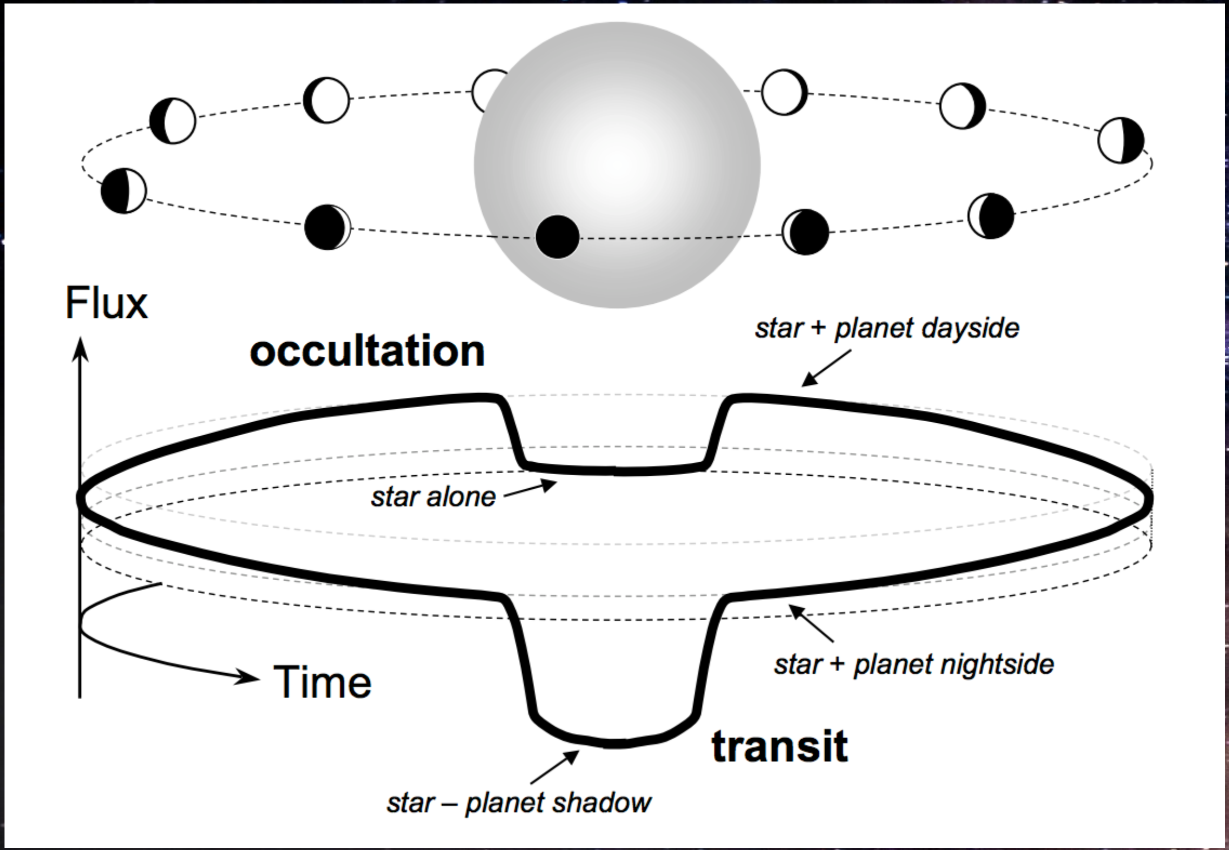


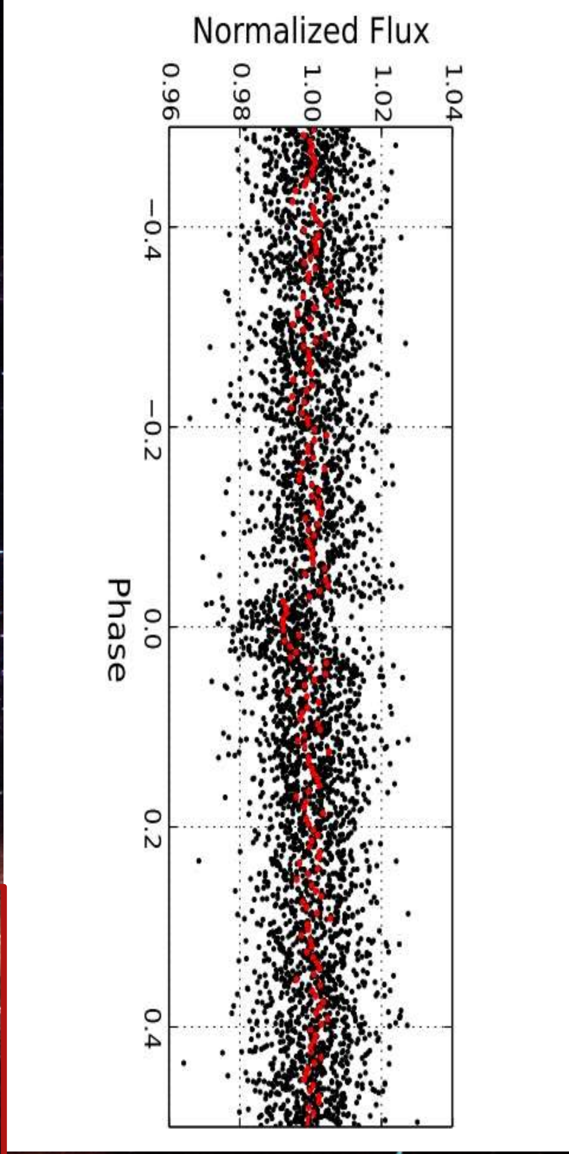
Detrended Data= Flux1 – Flux 2

Flux1: Raw Data

Flux2: Data after applying Gaussian Filter of 5σ cut-off

Box-Least Squares Method





- Geometric Criterion: Short Transit Duration **< 5% Period**
- Fits the input time series to periodic **"box"-shaped** functions
- Periodic box-shaped functions represent the behaviour of a light curve during a transit better than sines and cosines

Box-Least Squares Method

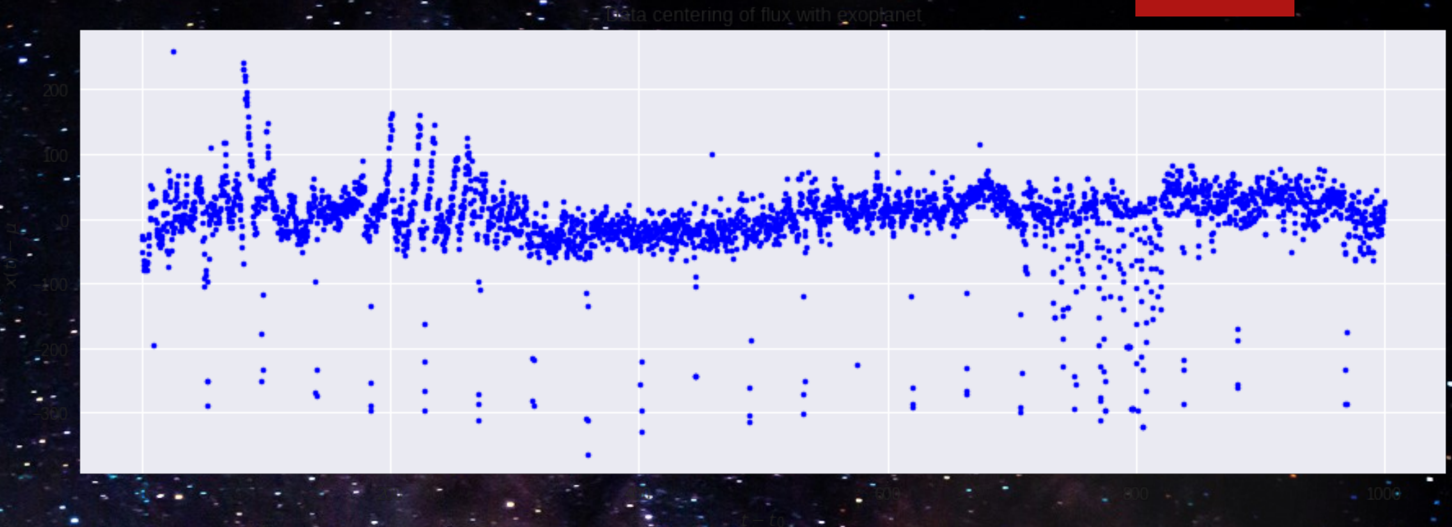
Steps

1. Data Centering

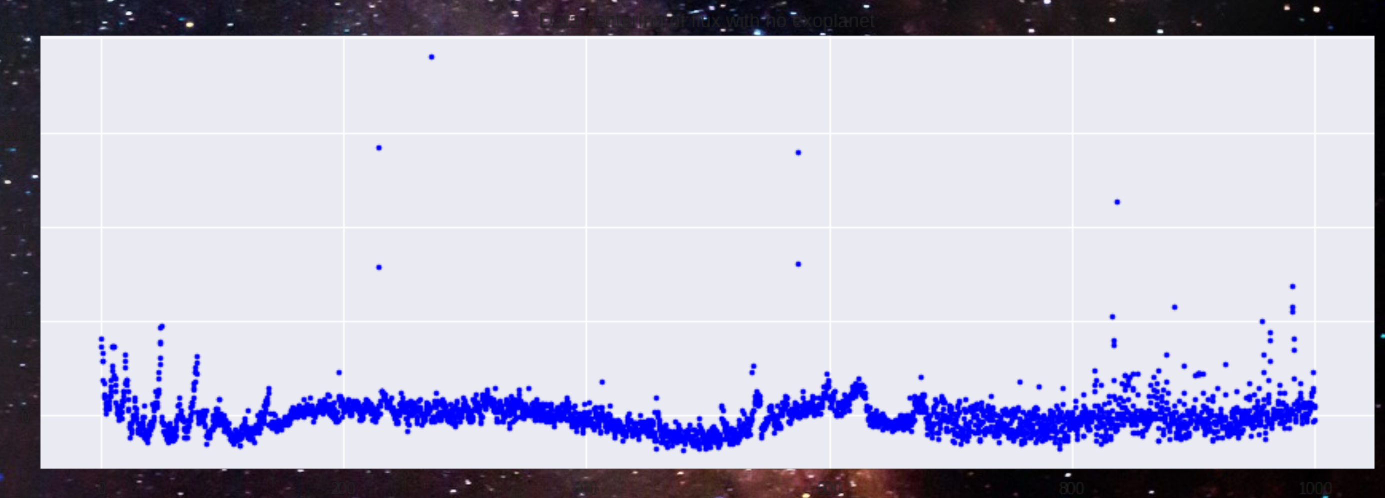
$$x_i = x_i - \mu$$

$$\mu = \frac{1}{n}\sum_{i=1}^{i=n} x_i$$

Centered Time Series with arithmetic average $\mu = 0$



With confirmed Exoplanet



With No confirmed Exoplanet

Box-Least Squares Method

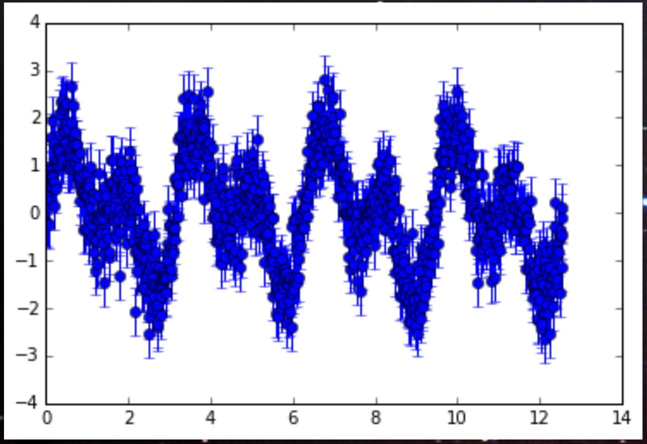
2. Data Folding and Binning

Folding is useful when a source has periodic variability; the data is plotted in terms of phase, such that all the data are plotted together as a single period, in order to see what the repeated pattern of variability is.

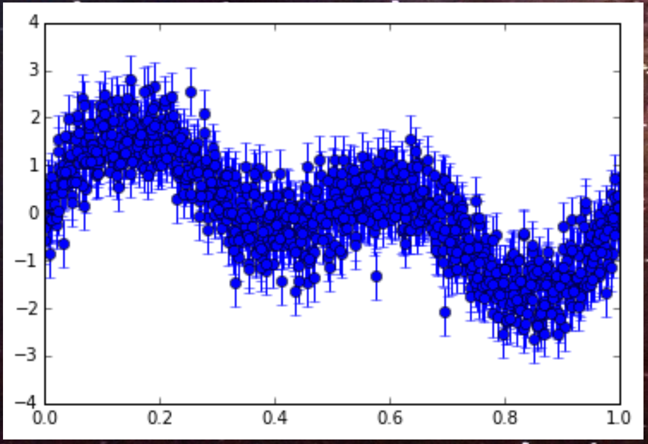
Random variations from this pattern can be reduced by binning the data in time, which involves splitting the phase range into steps (bins) in which all the data are averaged, using a weighted mean.

Large deviation to the correct Period results in very scattered folded time series.

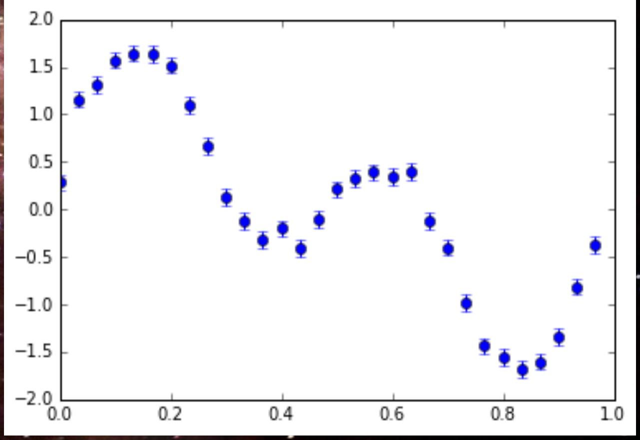
Raw Data



Folded



Folded + Binned

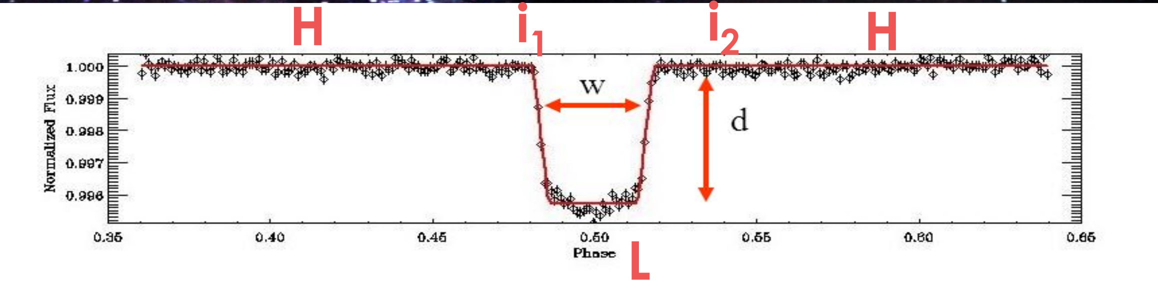


Box-Least Squares Method

3. Fitting

The algorithm applies the **weighted least squares** and fits the box shaped function to get parameters like Period(P), High, Low, fractional transit length(q)

$$\tilde{w}_i = \frac{\frac{1}{\sigma_i^2}}{\sum_{j=1}^{j=n} \sigma_j^{-2}}$$



For any test period, *i* in a trial period we minimize the *residual sum of squares*

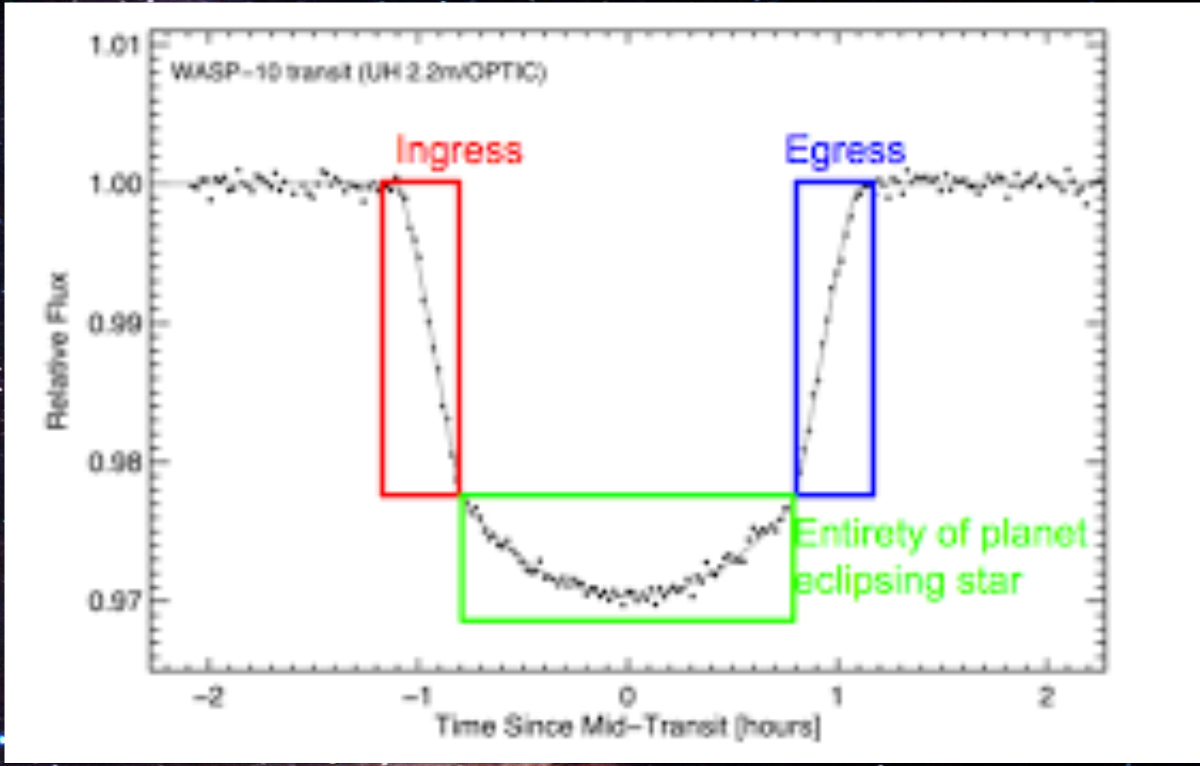
$$R(i_1, i_2) = \sum_{i=1}^{i=i_1-1} \tilde{w}_i(\tilde{x}_i - H)^2 + \sum_{i=i_1}^{i=i_2} \tilde{w}_i(\tilde{x}_i - L)^2 + \sum_{i=i_2+1}^{i=n} \tilde{w}_i(\tilde{x}_i - H)^2$$

Features Extracted from BLS

Orbital Period - The interval between consecutive planetary transits(in days)

Transit Duration- The duration of the observed transits. Duration is measured from first contact between the planet and star(in hours)

Transit Depth- The fraction of stellar flux lost at the minimum of the planetary transit.



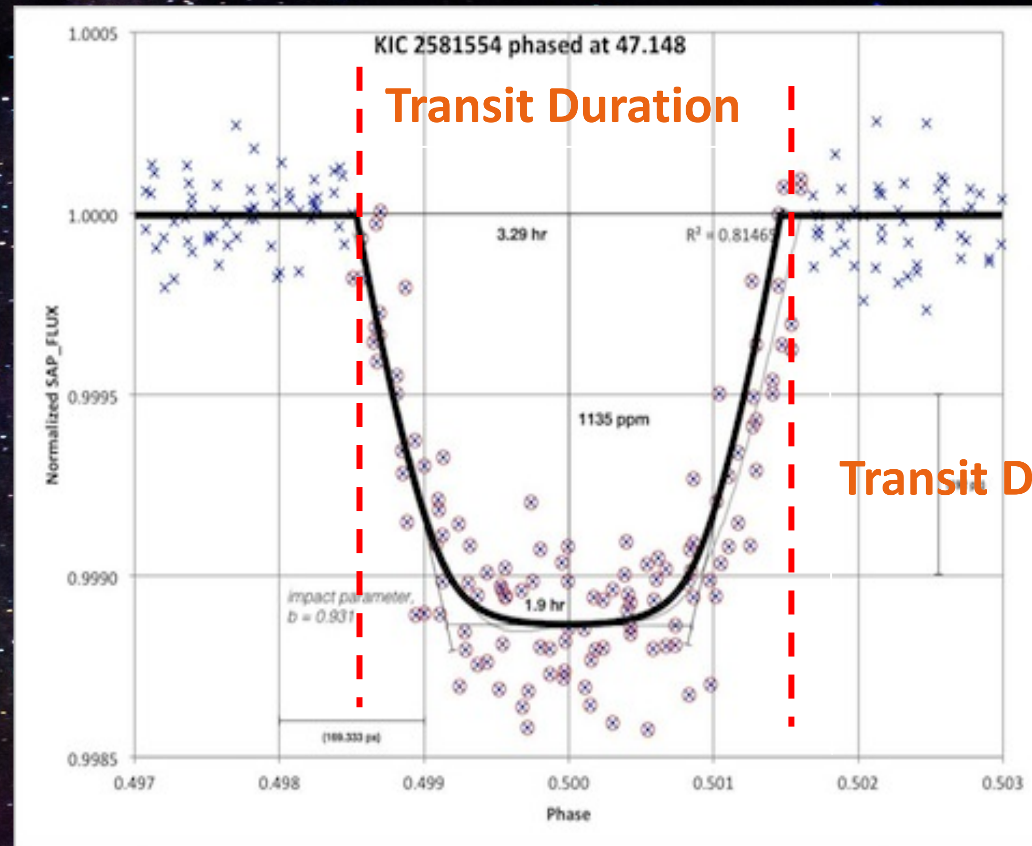
Features Extracted from BLS

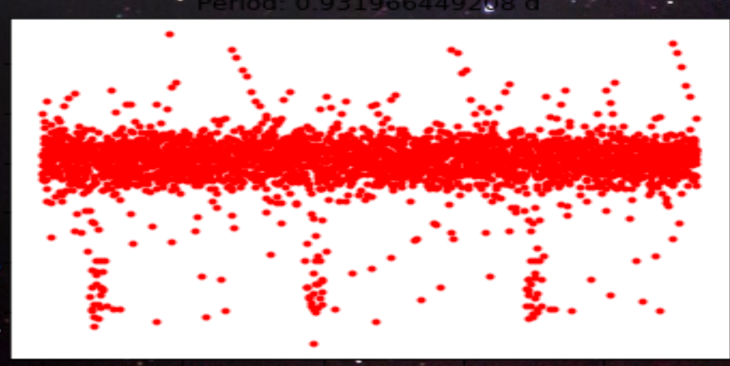
<u>Number of Transits-</u> **The number of expected transits or partially-observed transits associated with the planet candidate occurring within the searched light curve.**

<u>Other Features :</u>

Ingress- Ingress lasts from when the planet meets the solar limb (contact I) until the instant at which the planetary disk is totally encompassed (contact II)

Egress- The time period between Contact III and Contact IV

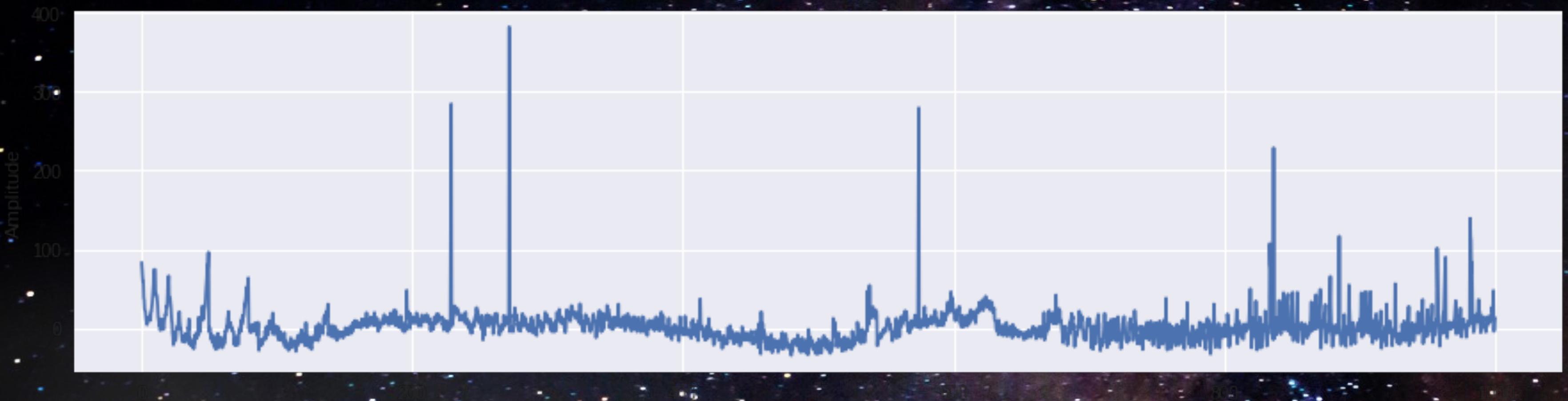




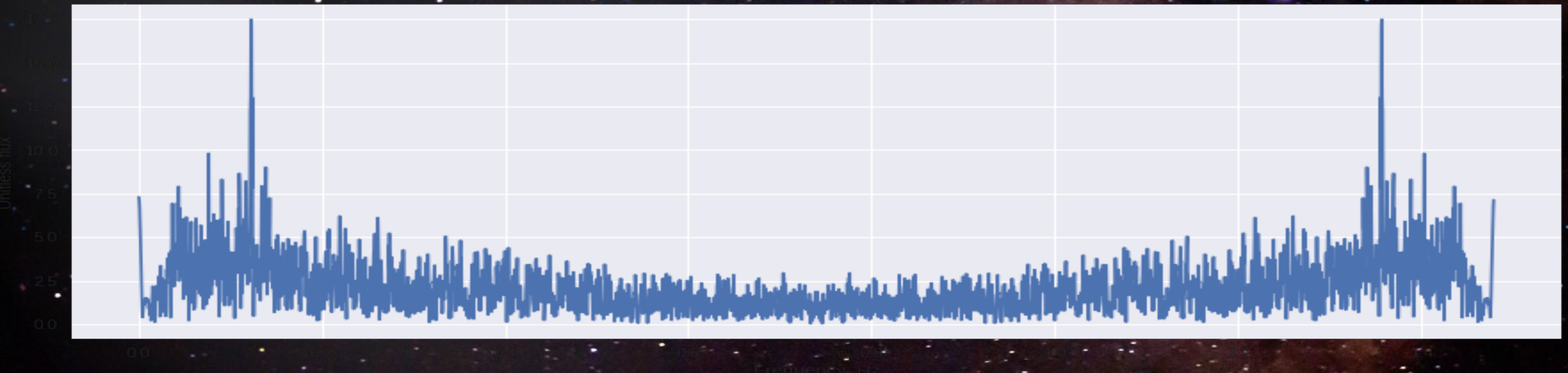
Ingress:	147
Egress:	152
q:	0.0297153581483
Depth:	69.2617852659
Period:	0.931966449208
SNR:	9.56999866395

Fast-Fourier Transform

When FFT is applied on the detrended data the plots were found to be symmetrical. So we took only half of the features.



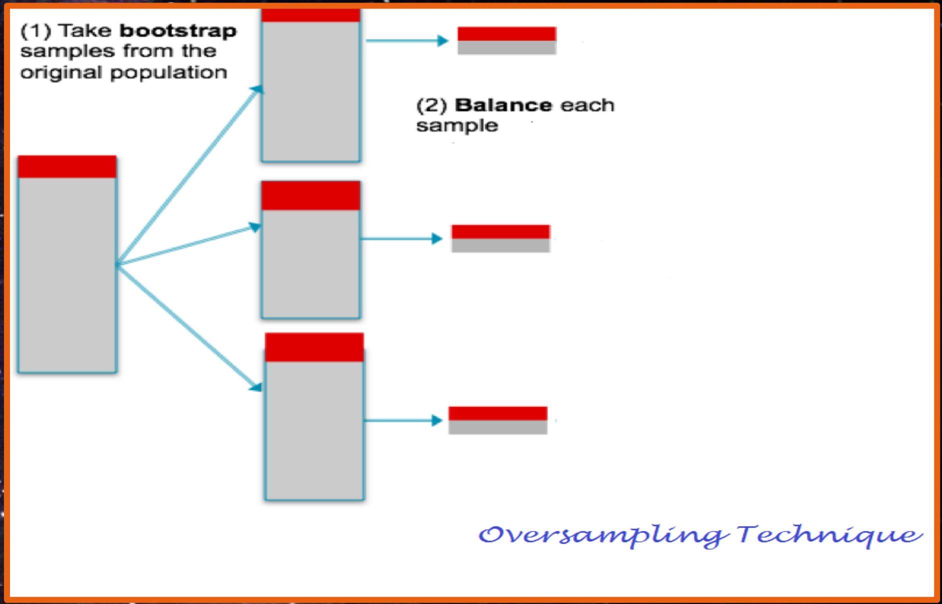
Time Domain



Frequency Domain

Modelling

- The data is highly imbalanced so **bootstrapping** was used to balance the classes during oversampling.



- Here we use **stratified K fold(K=5) cross validation** because during stratification each fold contains roughly the same proportions of the two types of class labels.
- Baseline predictive modelling on raw data using **Linear SVC** gives the following results:

```
Average Accuracy:  0.9880095293
Average Precision: 0.2522619048
Average Recall:    0.3416666667
Average F1:        0.2781313131
```

Modelling

Input Data (3960,3198)

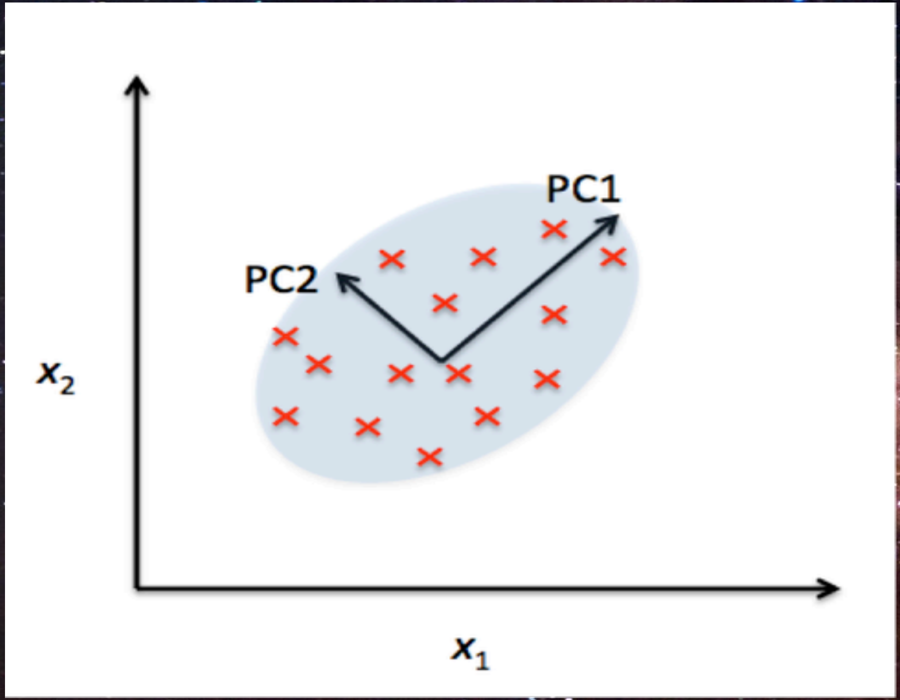
Dimensional Reduction

Features from BLS Algorithm (9)

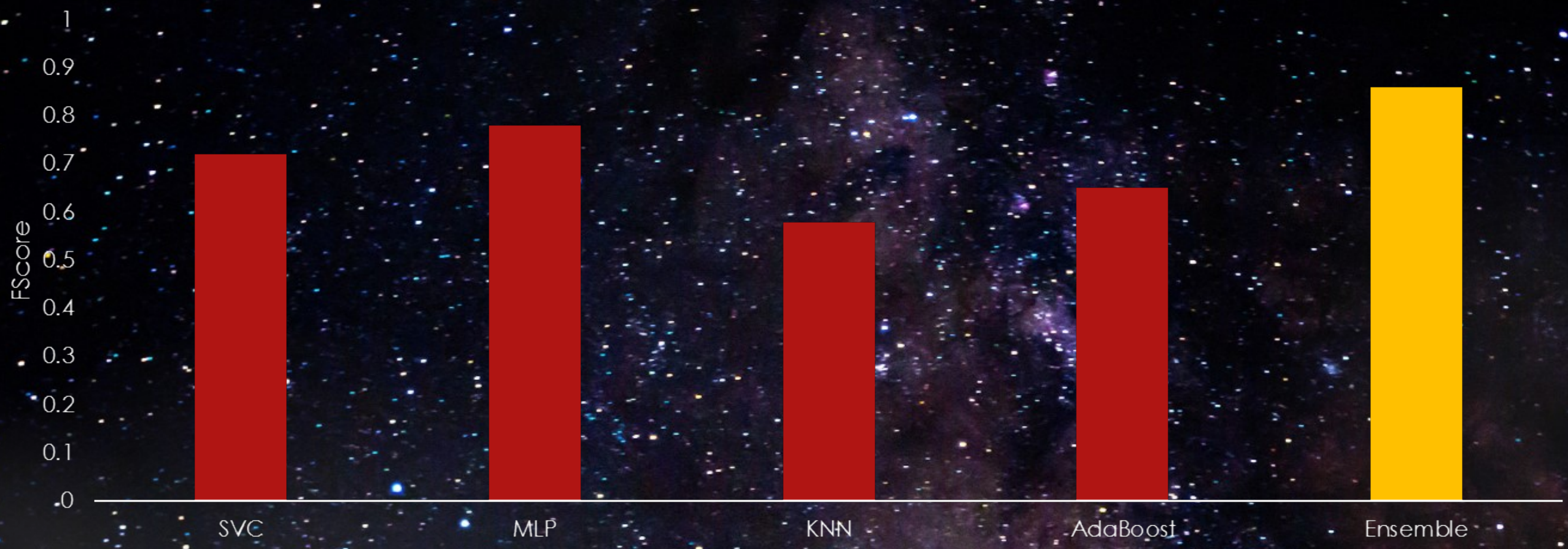
Symmetrical Plots-Data reduced to Half(3960,1599)

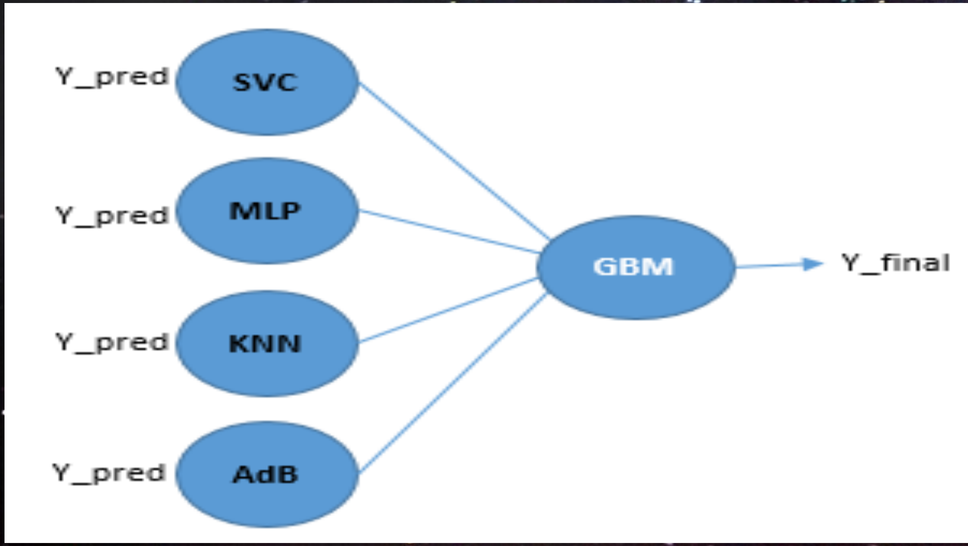
Incremental PCA (3960,165)

Combined Data(3960,174)



Modelling





Ensembling

Average Accuracy:	0.994
Average Precision:	0.894
Average Recall:	0.832
Average F1:	0.862

Novelty in our approach

- When the signal becomes distorted by higher harmonics, the sinusoidal representation fails. The BLS algorithm assumes only two levels of the periodic light curve- “high” and “low”.
- The Detrended data is combined with the features extracted from BLS Algorithm without any machine learning algorithm. This led to a significant improvement in the Fscore.
- For predicting the periodicity of the dips converting the time domain to frequency domain was necessary. We have cross-checked our period from the BLS algorithm using Auto-correlation and zero crossing. Also we have used Incremental PCA in our method since there were too many features.
- For classification purpose we have used an ensemble of learners since the aggregate opinion of a multiple models is less noisy than individual models. They average out biases and reduce the variance.

References

Kovács, G., Zucker, S. and Mazeh, T., 2002. A box-fitting algorithm in the search for periodic transits. Astronomy & Astrophysics, 391(1), pp.369-377.

McCauliff, S.D., Jenkins, J.M., Catanzarite, J., Burke, C.J., Coughlin, J.L., Twicken, J.D., Tenenbaum, P.,Seader, S., Li, J. and Cote, M., 2015. Automatic classification of Kepler planetary transit candidates. The Astrophysical Journal, 806(1), p.6.

http://www.stat.tamu.edu/~jlong/astrostat/fall2015/astr689.pdf

Jin, X. and Glass, D., 2014. Searching for exoplanets in the Kepler public data

http://www.planetary.org/explore/space-topics/exoplanets/transit-photometry.html