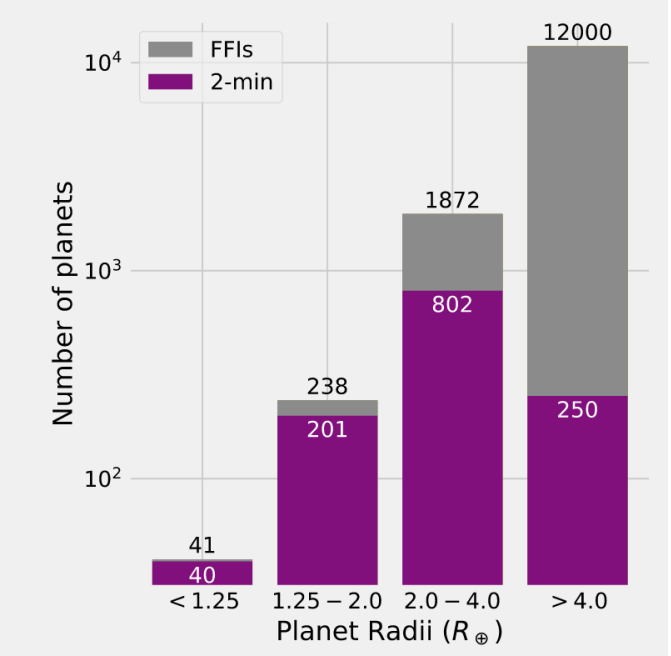
# RAVEN: RAnking and Validation of ExoplaNets

Andreas Hadjigeorghiou, 1st year PhD, Centre for Exoplanets and Habitability, University of Warwick
Supervisor: Dr. David Armstrong

# Motivation

- Simulations suggest potential TESS yield up to 14000 planets from its Primary Mission.
- Vast majority of planets from Full Frame Images (FFI), where False Positive (FP) rate is estimated to 11 per 1 true planet.
- Automated, large scale pipeline needed to identify and separate FPs from planetary candidates.




Predicted TESS planet detections from simulations.
From: Barclay et al. (2018)

# Motivation

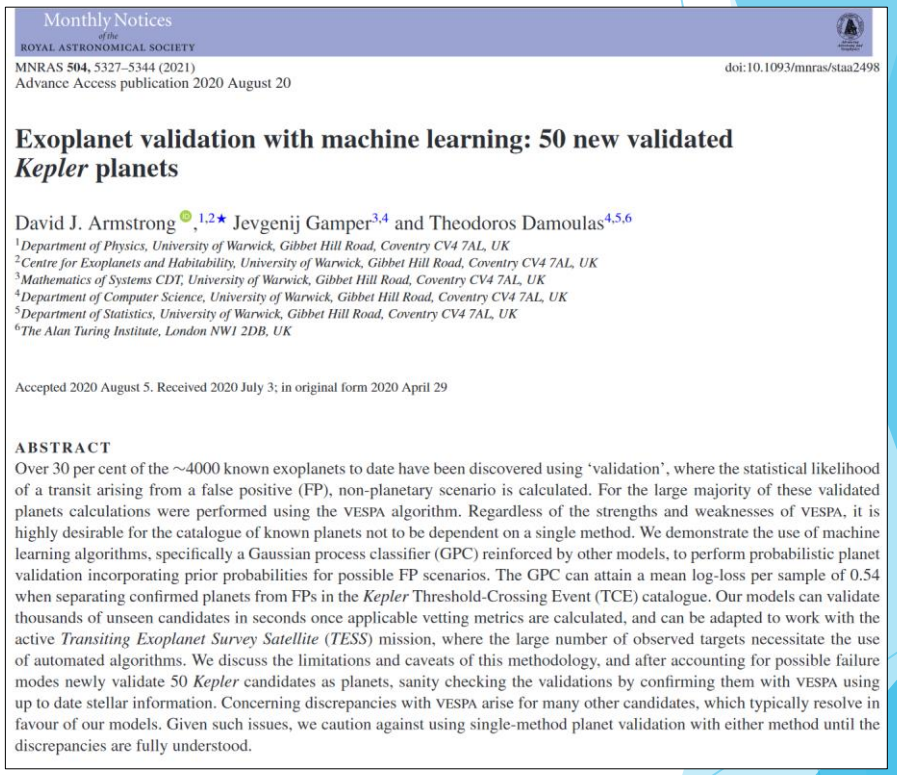
- Planetary candidates still need to be confirmed, ideally through independent follow-up observations.
- Not all candidates accessible to follow-up observations, especially those around faint starts.
- Large number of candidates leads to prioritisation of the most suitable.
- Statistical validation can allow candidates to be classified as “true planets” just from their transit signal.

# Vetting and Validation

- Vetting: The process of separating FP from true planet candidates.
- Many vetting pipelines, including Machine Learning (ML) implementations (McCauliff et al. 2015; Shallue and Vanderburg 2018; Yu et al. 2019)
- Validation: Statistical confirmation of an exoplanet, by calculating its likelihood probability of being a FP or a true planet.
- Only a handful statistical validation implementations currently exist:
  - BLENDER (Torres et al. 2015)
  - PASTIS (Díaz et al. 2014; Santerne et al. 2015)
  - vespa (Morton et al. 2016)
  - TRICERATOPS (Giacalone & Dressing 2020)
- Validation accounts for ~30% of all confirmed exoplanets.

# Our pipeline: RAVEN

- Automated vetting and validation in a single workflow, specifically developed for TESS.
- Uses four different ML classifiers for vetting.
- New statistical validation implementation: ML classification scores combined with pre-computed prior probabilities to determine the likelihood probability of the candidate being a true planet.
- Candidates with probabilities higher than 99% will be statistically validated.
- The project builds upon previous work on developing such an automated vetting and validation pipeline for Kepler by Armstrong et. al (2020).



Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **504,** 5327–5344 (2021) doi:10.1093/mnras/staa2498
Advance Access publication 2020 August 20

**Exoplanet validation with machine learning: 50 new validated *Kepler* planets**

David J. Armstrong [1,2★] Jevgenij Gamper[3,4] and Theodoros Damoulas[4,5,6]

[1]*Department of Physics, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK*
[2]*Centre for Exoplanets and Habitability, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK*
[3]*Mathematics of Systems CDT, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK*
[4]*Department of Computer Science, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK*
[5]*Department of Statistics, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK*
[6]*The Alan Turing Institute, London NW1 2DB, UK*

Accepted 2020 August 5. Received 2020 July 3; in original form 2020 April 29

**ABSTRACT**

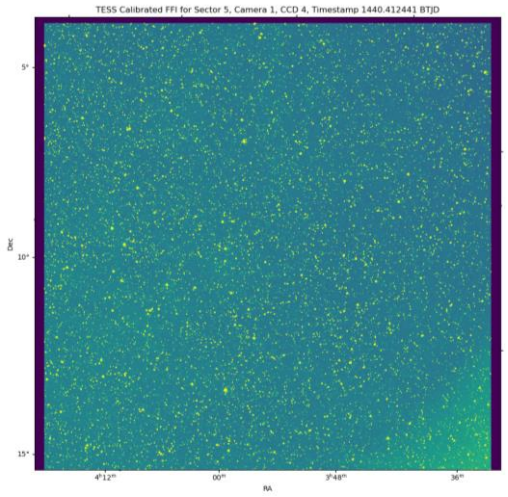
Over 30 per cent of the ~4000 known exoplanets to date have been discovered using 'validation', where the statistical likelihood of a transit arising from a false positive (FP), non-planetary scenario is calculated. For the large majority of these validated planets calculations were performed using the VESPA algorithm. Regardless of the strengths and weaknesses of VESPA, it is highly desirable for the catalogue of known planets not to be dependent on a single method. We demonstrate the use of machine learning algorithms, specifically a Gaussian process classifier (GPC) reinforced by other models, to perform probabilistic planet validation incorporating prior probabilities for possible FP scenarios. The GPC can attain a mean log-loss per sample of 0.54 when separating confirmed planets from FPs in the *Kepler* Threshold-Crossing Event (TCE) catalogue. Our models can validate thousands of unseen candidates in seconds once applicable vetting metrics are calculated, and can be adapted to work with the active *Transiting Exoplanet Survey Satellite* (*TESS*) mission, where the large number of observed targets necessitate the use of automated algorithms. We discuss the limitations and caveats of this methodology, and after accounting for possible failure modes newly validate 50 *Kepler* candidates as planets, sanity checking the validations by confirming them with VESPA using up to date stellar information. Concerning discrepancies with VESPA arise for many other candidates, which typically resolve in favour of our models. Given such issues, we caution against using single-method planet validation with either method until the discrepancies are fully understood.

# Pipeline Overview

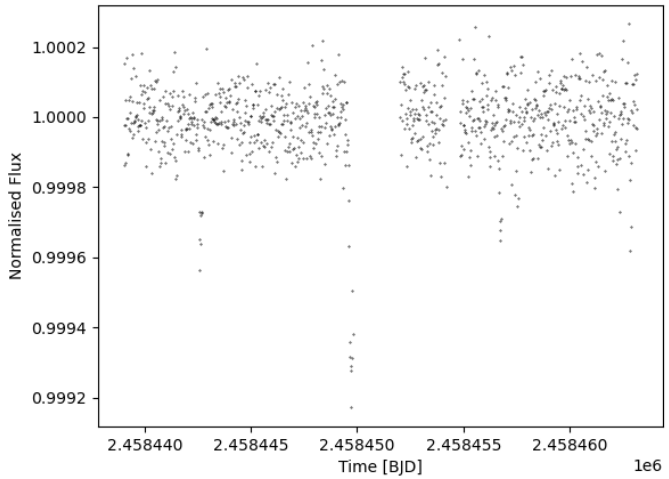
1. **Data ingestion**

# 1. Data ingestion

- Input: List of candidate events with their properties (epoch, period, transit duration, estimated transit depth) and path to lightcurve file.
- Lightcurve reading and detrending.
- Our data:
  - Lighcurves extracted from the FFI images with a pipeline developed by the Next Generation Transit Survey (NGTS) team.
  - Candidate events identified from the NGTS team's BLS survey.




TESS calibrated FFI for Sector 5, Camera 1, CCD 4




Lightcurve for TIC 407966340 extracted using the NGTS FFI pipeline.

# Pipeline Overview

1. Data ingestion
2. **Identification of nearby sources - dilution factors and centroid offset calculation**

# 2. Identification of nearby sources – dilution factors and centroid offset calculation

- Significant blending of light from nearby stars in the target's lightcurve.
- Nearby sources in a 220 arcsec radius identified using the TESS Input Catalog (TIC) and GAIA.
- Flux contribution of each source determined based on their pixel distance from the target, their TESS magnitude and a corresponding Point Spread Function.
- Centroid offset - the pixel shift of the centre of the incoming light, which occurs during the transits of each candidate.

# Pipeline Overview

1. Data ingestion
2. Identification of nearby sources – dilution factors and centroid offset calculation
3. **Identification of possible sources for each candidate event**

# 3. Identification of possible sources for each candidate event

- Nearby sources examined based on:
  - Calculated flux contribution – remove non contributing sources
  - Bright enough to cause observed eclipse – remove faint sources
  - Centroid shift expected if source of eclipse – remove if not consistent with observation
- Identified possible sources for the detected event treated as separate candidates.
- Pipeline computes their likelihood probability – determine Nearby FP probability and potentially identify the actual source of the eclipse.

# Pipeline Overview

1. Data ingestion
2. Identification of nearby sources – dilution factors and centroid offset calculation
3. Identification of possible sources for each candidate
4. **Prior probability calculation**

# 4. Prior probability calculation

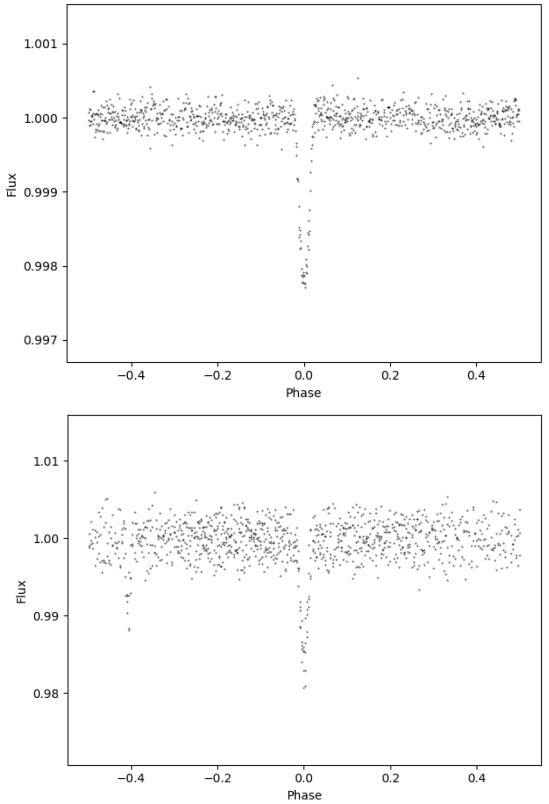
- Prior probabilities calculated for the following scenarios:
  - Transiting Planet
  - Eclipsing Binary (EB)
  - Hierarchical Eclipsing Binary
  - Background Eclipsing Binary
  - Background Transiting Planet
  - Hierarchical Transiting Planet
- Prior probabilities incorporate:
  - the positional probability (target, secondary source, background source)
  - the occurrence rate for planets and EBs
  - the probability to detect the eclipse with TESS
- Prior probability for non-astrophysical FP is set the same as that of a planet.

# Pipeline Overview

1. Data ingestion
2. Identification of nearby sources – dilution factors and centroid offset calculation
3. Identification of possible sources for each candidate
4. Prior probability calculation
5. **Machine Learning Vetting**

# 5. Machine Learning Vetting – Training Sets

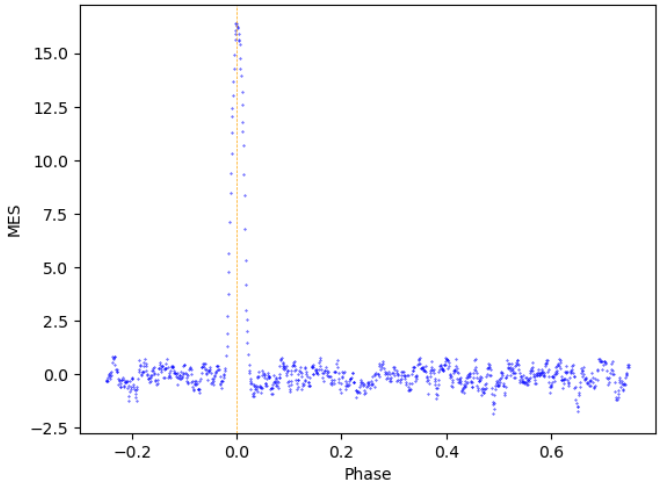
- 4 different ML classifiers, including a Gaussian Process and a Random Forest.
- Synthetic training set:
  - Detailed simulated transits for the planet and astrophysical FP scenarios, generated with the PASTIS software.
  - Tailored to TESS data – targets selected from TIC and simulations injected in the corresponding TESS lightcurves.
- Non-astrophysical FP training set constructed from pre-identified TESS lightcurves with such a signal.



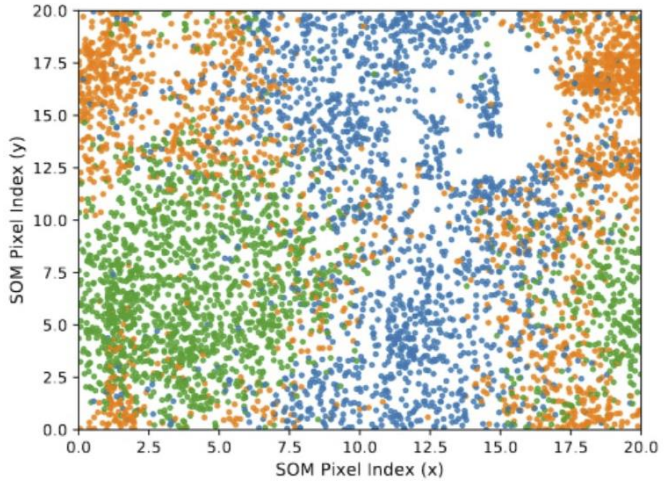

Top: Simulated Planetary Transit
Bottom: Simulated Background Eclipsing Binary

# 5. Machine Learning Vetting - Metrics

- ML models trained based on a set of metrics derived from the lightcurves.
- Incorporate stellar and planetary characteristics and features specific to the observed transits.
- Two of the most important metrics:
  - Multiple Event Statistic (MES) – relates to the significance of the detected signal,
  - Self-Organising Map statistic - reduces the transit shape into a single metric value based on unsupervised ML clustering.




MES for simulated planetary transit injected in a TESS lightcurve.




SOM clustering for differentiating planets (green), astrophysical (orange) and non-astrophysical (blue) FPs. From: Armstrong et. al (2020)

# Pipeline Overview

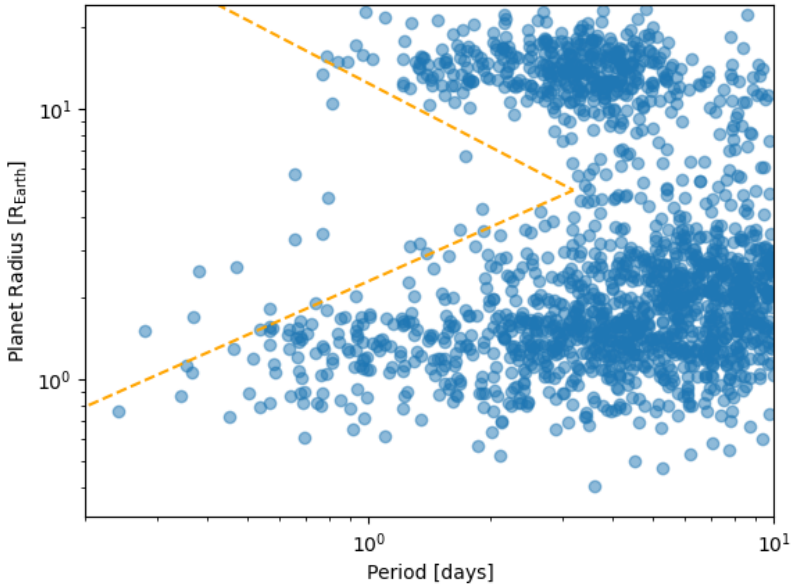
1. Data ingestion
2. Identification of nearby sources – dilution factors and centroid offset calculation
3. Identification of possible sources for each candidate
4. Prior probability calculation
5. Machine Learning Vetting
6. **Statistical Validation Implementation**

# 6. Statistical Validation Implementation

- ML classification calibrated and transformed into a probability.
- Gaussian Process Classifier naturally probabilistic and thus exempt.
- Probability from ML classification combined with the pre-computed prior probabilities for the planet and the FPs scenarios.
- Posterior probability for each scenario.
- Candidates with probability higher than 99% to be planets can be validated.
- Use of four ML classifiers provides independent check within the pipeline – a candidate must be validated by all four, thus no reliance on a single method.

# Science Outcomes

- Validation of many TESS candidates inaccessible to independent confirmation, but with strong evidence from transit signal.
- Allows more true exoplanets to be used in population studies, such as the study of the Neptunian Desert.
- Potential “real time” ranking of candidates, helpful follow-up observation prioritisation.




Radius-period plot for confirmed Planets from the NASA Exoplanet Archive, showing the Neptunian Desert with arbitrarily drawn boundaries

# THANK YOU FOR LISTENING!