Boldly Going

Identifying New Exoplanets based on the Kepler Telescope Mission

Matt Paterson, hello@hireMattPaterson.com

Data Science Fellow

General Assembly

Artist's rendering of the Kepler Telescope

Photo Credit: NASA/Ames/JPL-Caltech/T Pyle

Data Science Problem

- Astronomy is time consuming, and thus expensive.

- We have gained the ability to compile vast amounts of data but we have a limited number of qualified researchers to interpret that data.

Background

- “NASA’s Kepler spacecraft was launched to search for Earth-like planets orbiting other stars. It discovered more than 2,600 of these "exoplanets"—including many that are promising places for life to exist.”

  - Source: https://solarsystem.nasa.gov/missions/kepler/in-depth/

• Launch Data:	March 7, 2009
• First Planet Found:	December, 2011
• “Out of Gas”	October 30, 2018
• Launch Vehicle:	Delta 7925-10L
• Telescope:	One Photometer (Schmidt Telescope)
• Photographs:	One part of the Cygnus-Lyra Constellations

Cygnus – Lyra

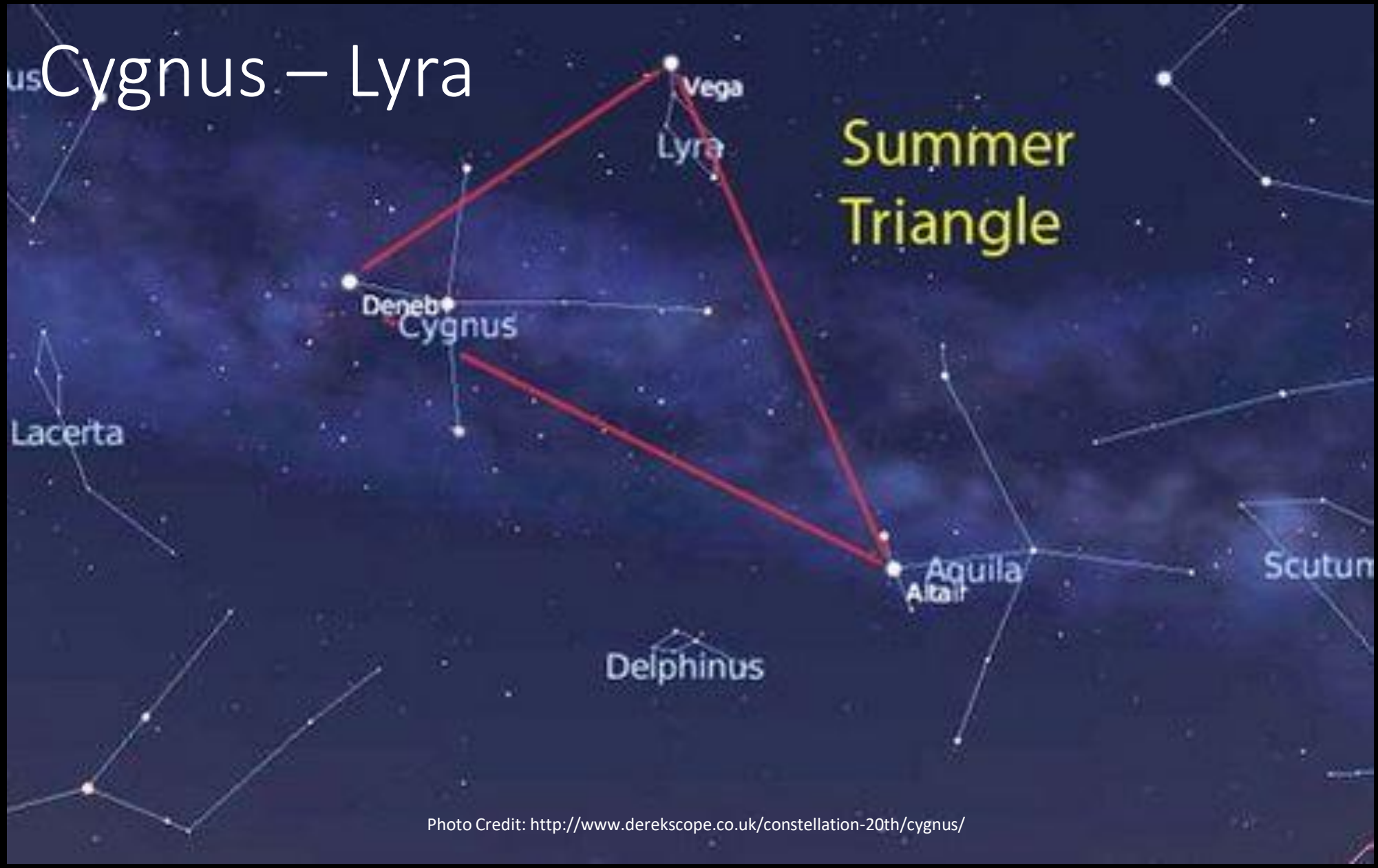


Photo Credit: http://www.derekscope.co.uk/constellation-20th/cygnus/

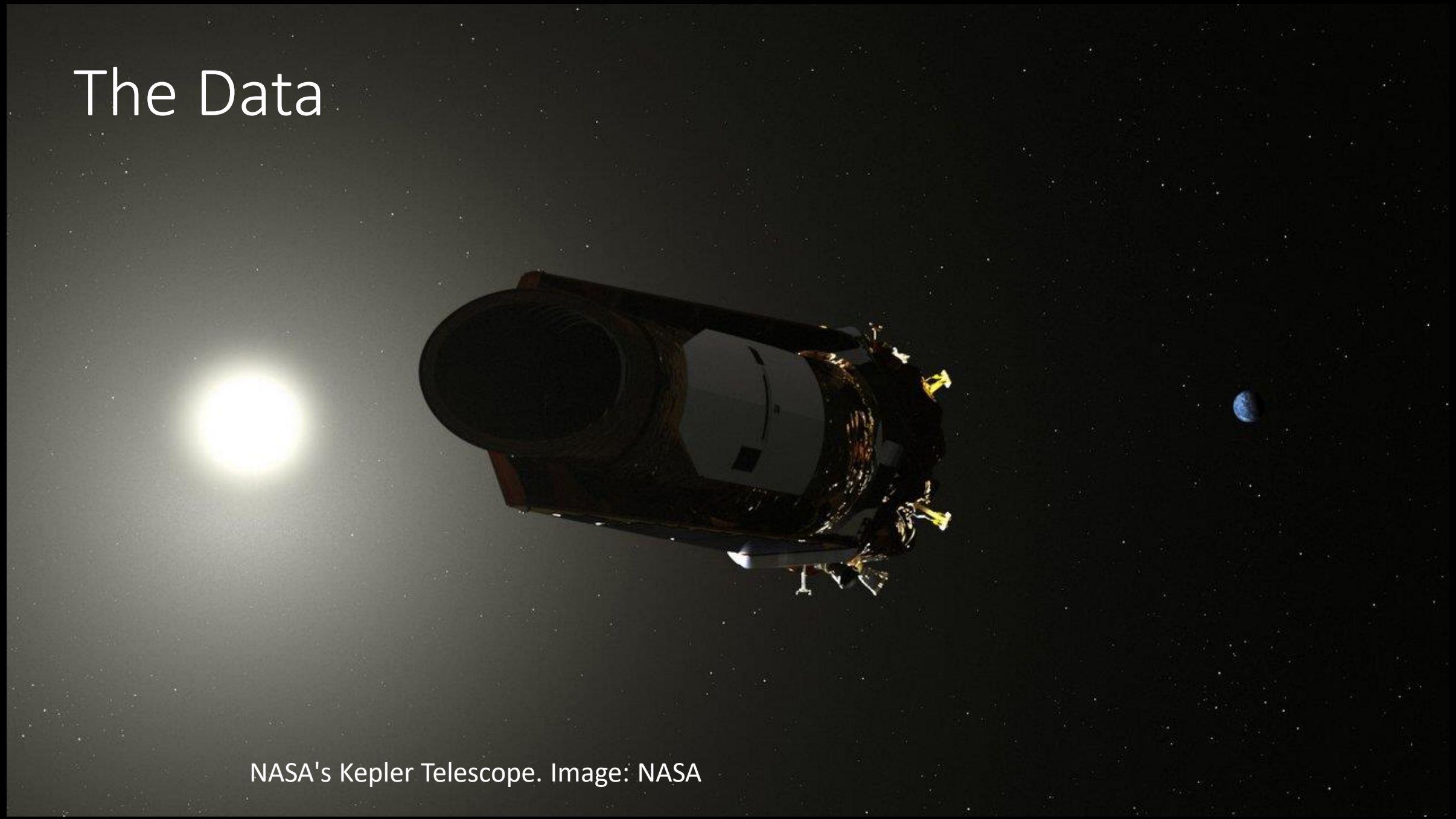
Data Science Solution

- We can compare the data of currently unconfirmed Kepler Objects of Interest to that of Confirmed Exoplanets and Confirmed non-exoplanet observations to predict the existence of planets orbiting nearby stars.

- Further, we can package this identification system to allow us to make the same, faster predictions on data in the K2 and TESS missions.

- Positive results could save considerable amounts of money on research and allow those scientists to map our galaxy faster, or to focus on more complexing questions about our celestial neighborhood.

The Data



NASA's Kepler Telescope. Image: NASA

KOI Cumulative Table

- The Kepler Objects of Interest, Cumulative Table
  - Last Updated: September 27, 2018
  - https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=cumulative

  - Gathers info from KOI activity tables describing data from searches of Kepler light curves

  - Contains nearly 10,000 observations, each requiring a minimum 3 observed transits in front of a star in addition to other basic requirements of the mission

Features of the Table

- 69 columns of data, most are quantitative
- 35 Features in our classification models
- 4212 of 8744 have been certified as NOT exoplanets
- BASELINE 35% Positive (exoplanets), 65% Negative, 2245 unknown

- Strongest Correlations:
  - Uncertainty in Photospheric Temperature
  - Uncertainty in Acceleration due to Gravity
  - Uncertainty in Hours of Transit Duration
  - Acceleration due to Gravity
  - Equilibrium Planetary Temperature
  - Transit Depth
  - Star Temperature

Correlated Features – The heat is on!

For those of you here in the live presentation, my apologies for the dense and wordy slide ☺

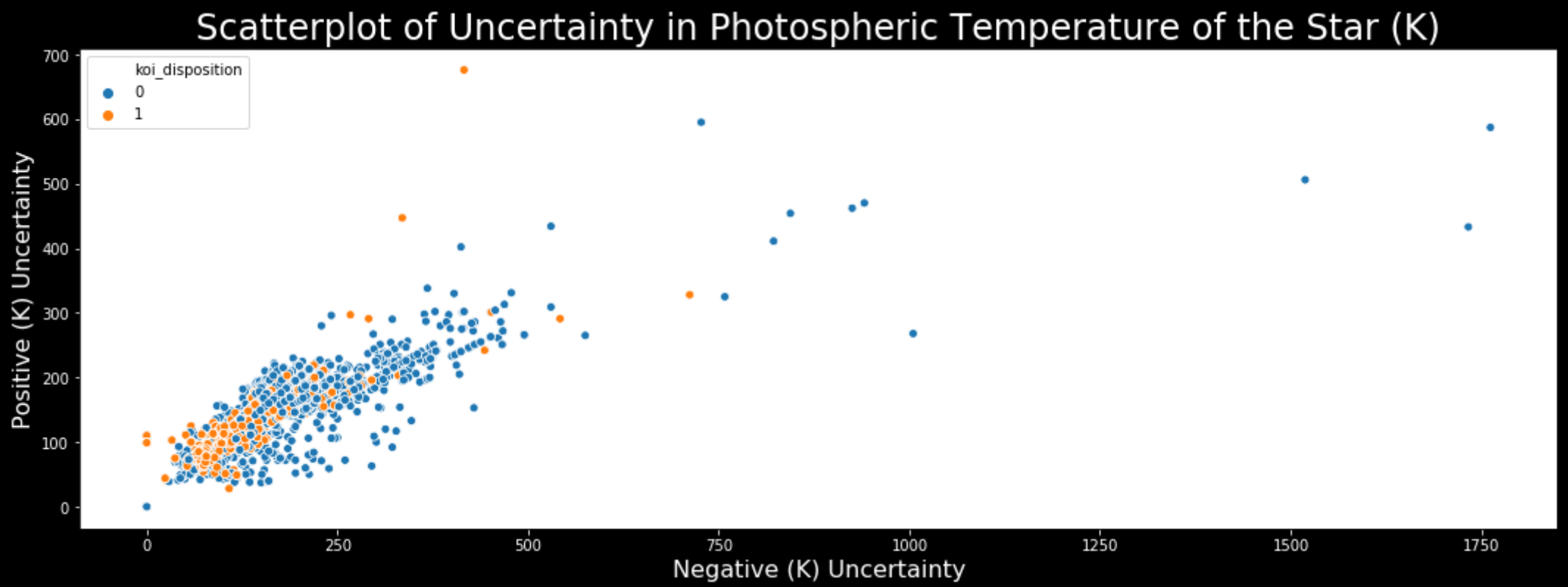
- **We don't have any strong correlations to the koi_disposition**
- **We do have a few strong correlations just the same:**

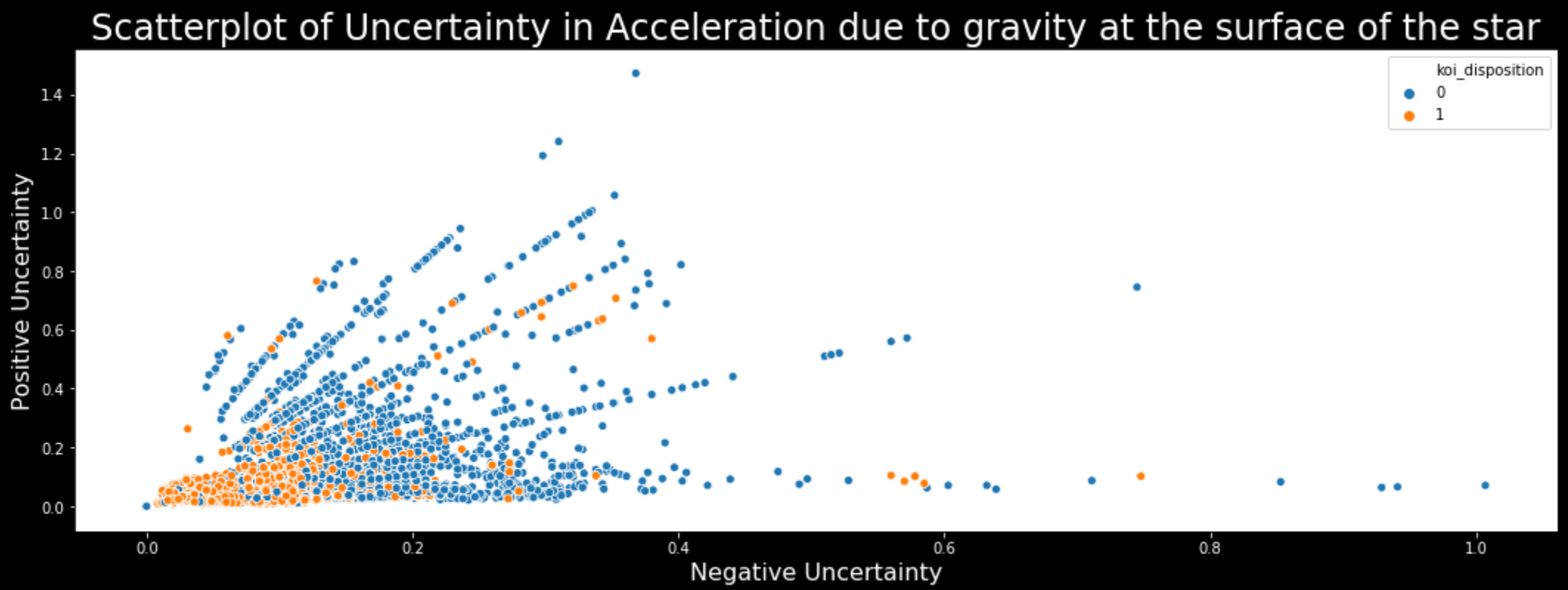
1. There is a very strong correlation between the koi_impact, the projected distance between the center of the stellar disc and the center of the planet disc at conjunction, and the koi_prad, or the planetary radius. This correlation is due to the fact that the Kepler mission is looking for Earth-like planets that orbit in such a distance to allow for surface temperatures that are not much different from earth's temperature range. Thus planets in these zones are most likely to be classified as planets at the outset of the project.
2. There is a very strong correlation between koi_srad, or the photospheric stellar radius, koi_teq (Equilibrium Temperature), and the koi_insol, or the insolation flux, or Earth flux. This is another way to give equilibrium temperature. Again, if a planet's too hot, and the star is too big, or the opposite for that matter, it may not be investigated early and thus not classified yet as an exoplanet, if at all.

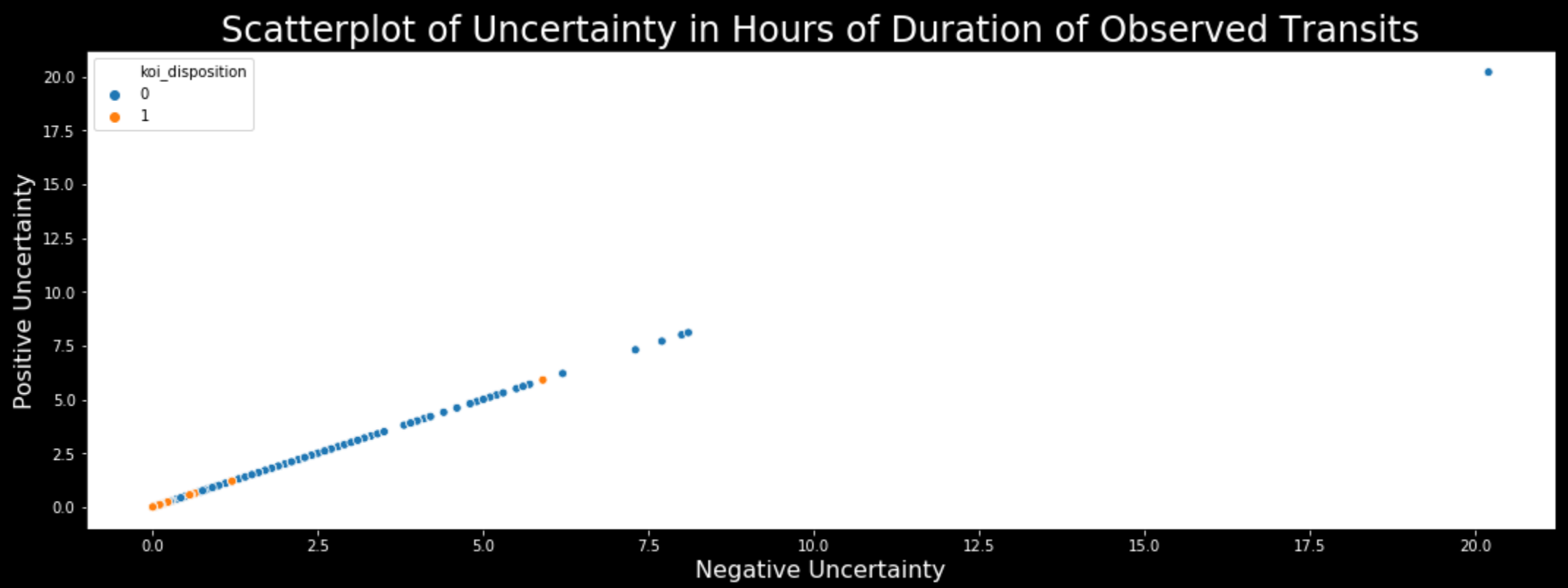
Correlation Matrix of Disqualifier Flags

Self-fulfilling models are bad

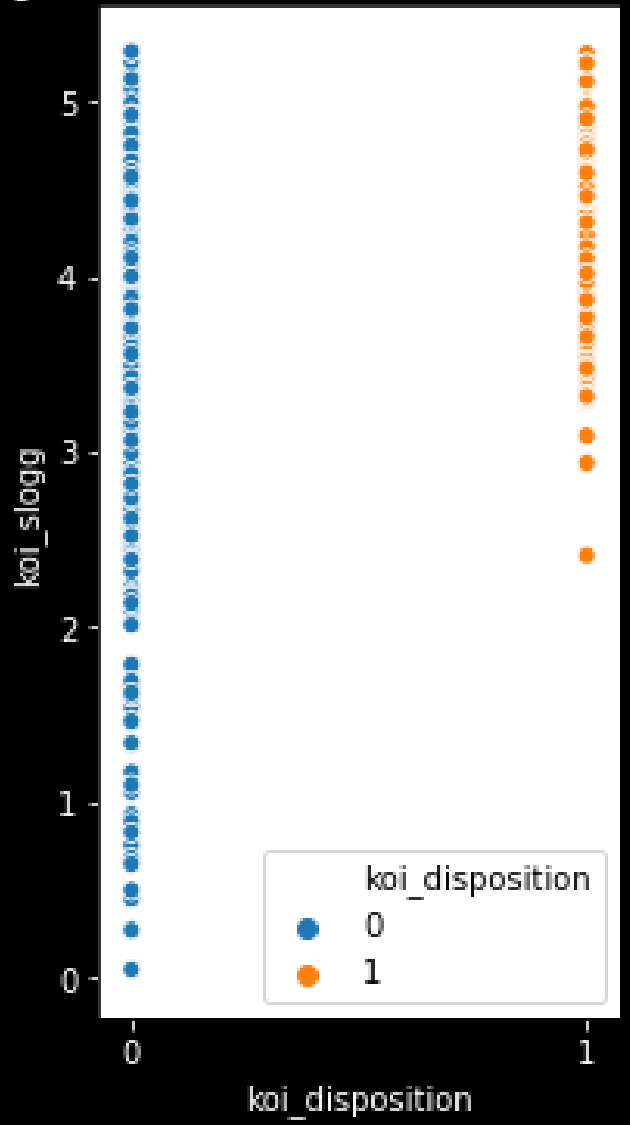
- We must drop these flag columns since our model would not work on unknown or new data if it was trained on these flags.
- We also drop out id numbers and any other non-empirical data
- Look now at the strongest correlations


Scatterplot of Uncertainty in Photospheric Temperature of the Star (K)
koi_disposition
0
1
Positive (K) Uncertainty
Negative (K) Uncertainty
0
100
200
300
400
500
600
700
0
250
500
750
1000
1250
1500
1750

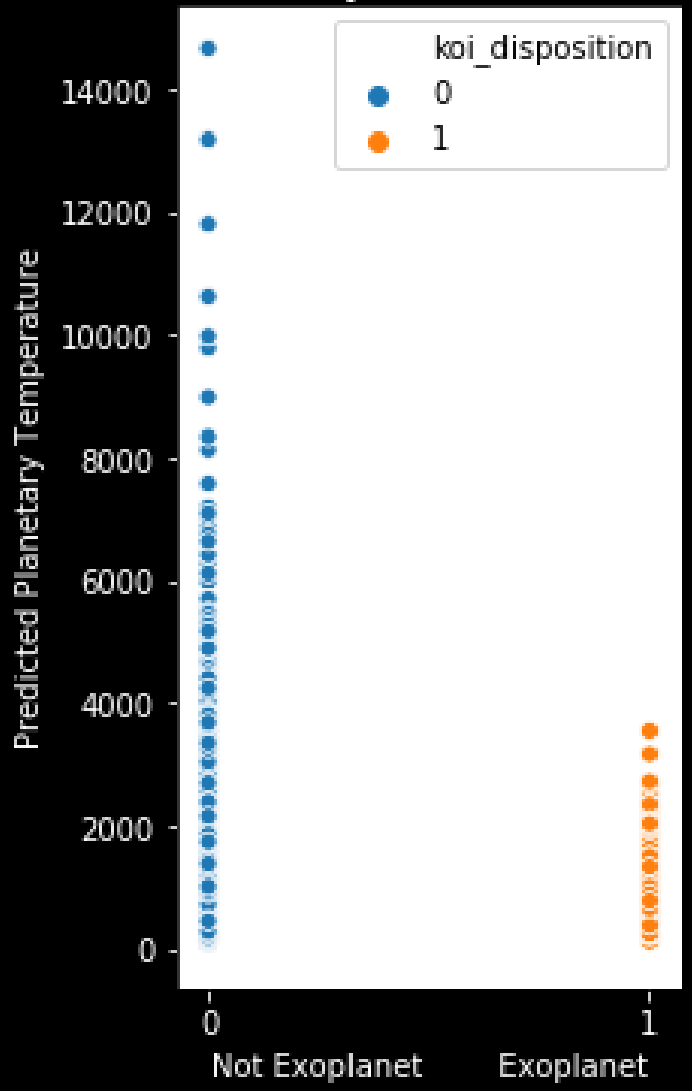

Scatterplot of Uncertainty in Acceleration due to gravity at the surface of the star
koi_disposition
0
1
Positive Uncertainty
Negative Uncertainty
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4


Scatterplot of Uncertainty in Hours of Duration of Observed Transits
koi_disposition
0
1
Positive Uncertainty
Negative Uncertainty
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

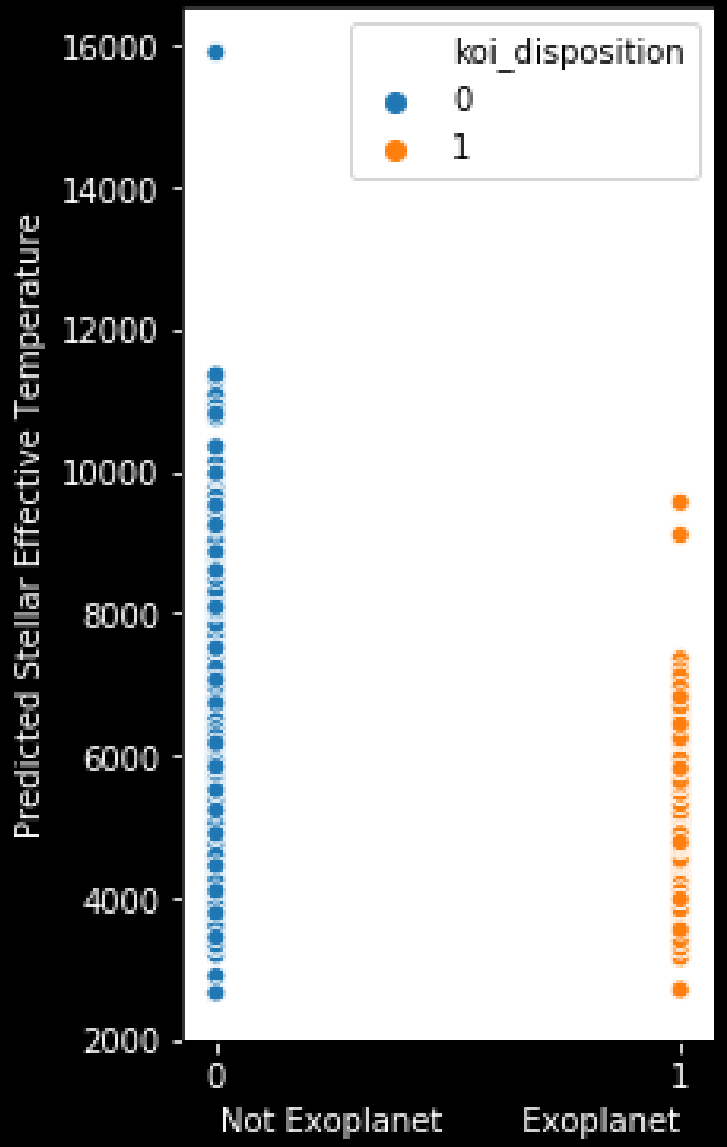
Surface Gravity of the stars in the Kepler Study



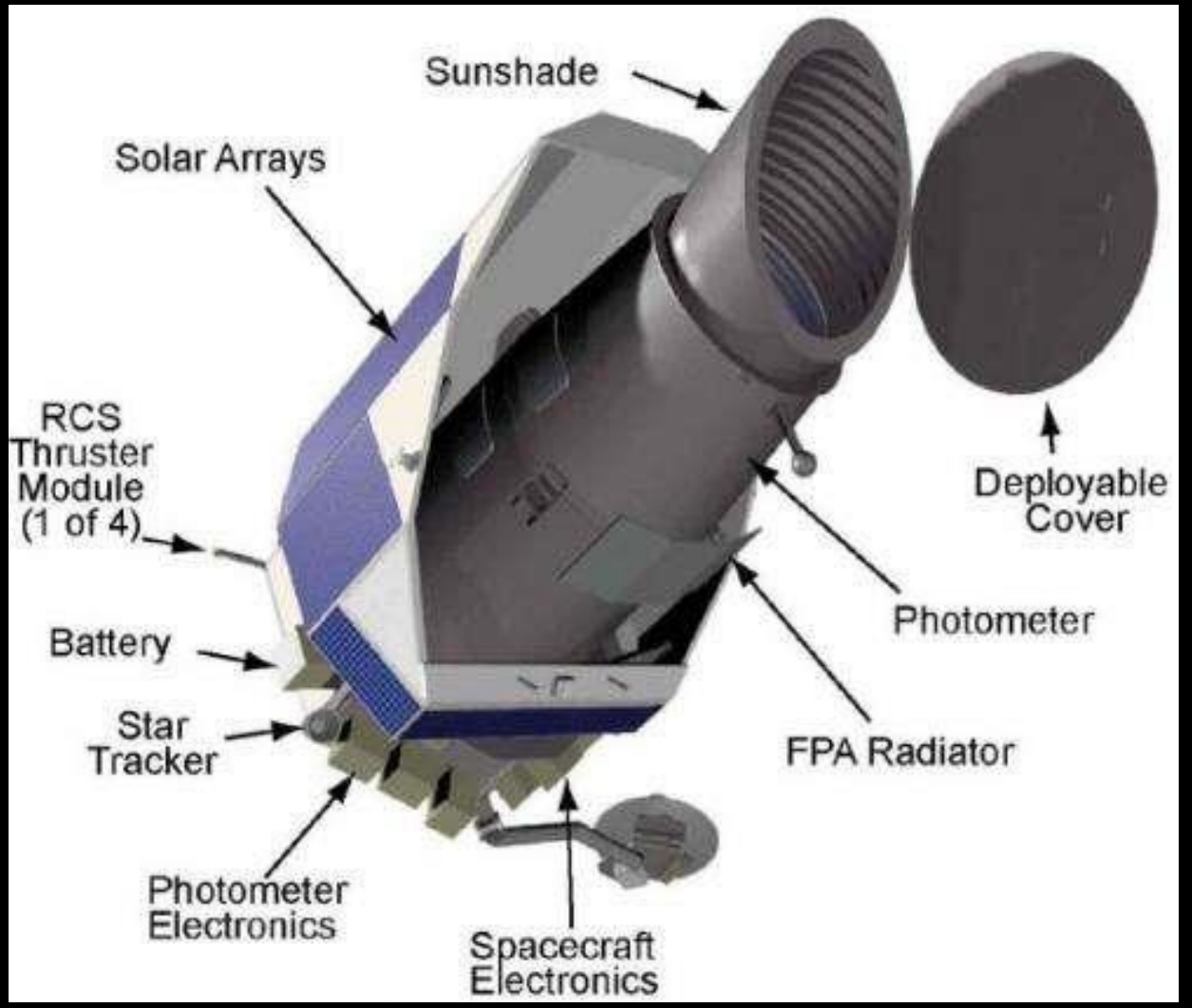
Predicted Equilibrium Temperature on possible Exoplanet



Predicted Stellar Effective Temperature



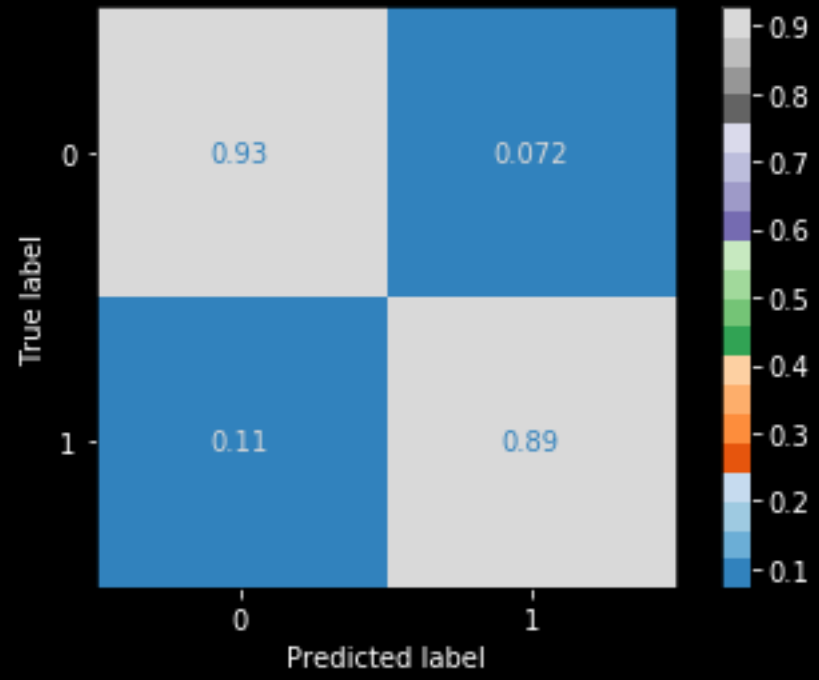
The Models



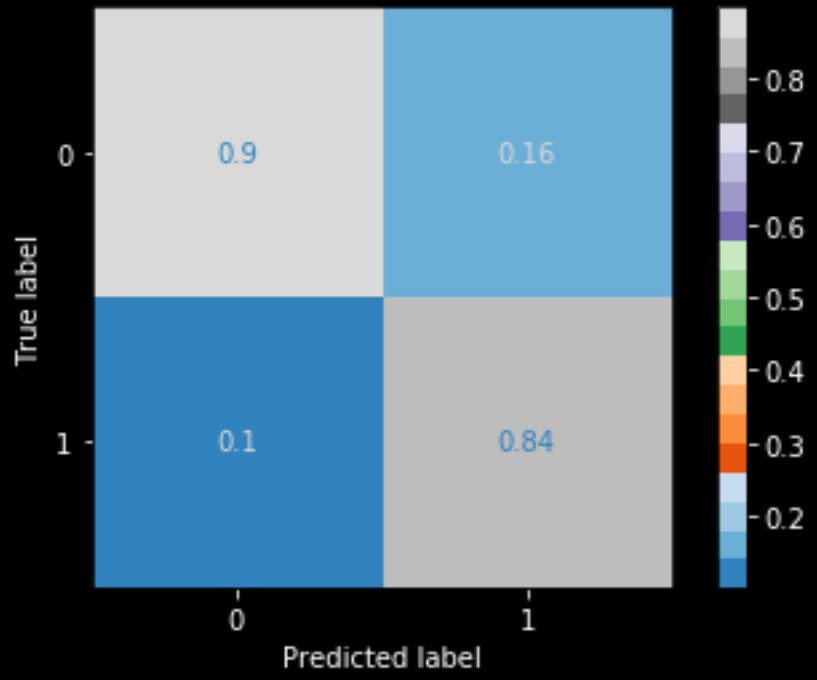
A diagram of the Kepler space telescope. Credit: NASA

Two Logisitc Regression Models

All Features…91.38% Validation Accuracy



Limited…87.92% Validation Accuracy



Logit Function

In the simplest terms, a Logistic Regression model will classify a KOI as Exoplanet if its probability is more than 50% based on the model.

A similar decision is used in other classifier models, and the Sigmoid function that we'll use in our Recurrent Neural Network also works in a similar manner.

I created a function that reads in these probabilities and the model's predictions, and then adjusts the predictions in line with the requested probability threshold. When adjusted to 99% certainty, the model eliminates false positives in the classification models.

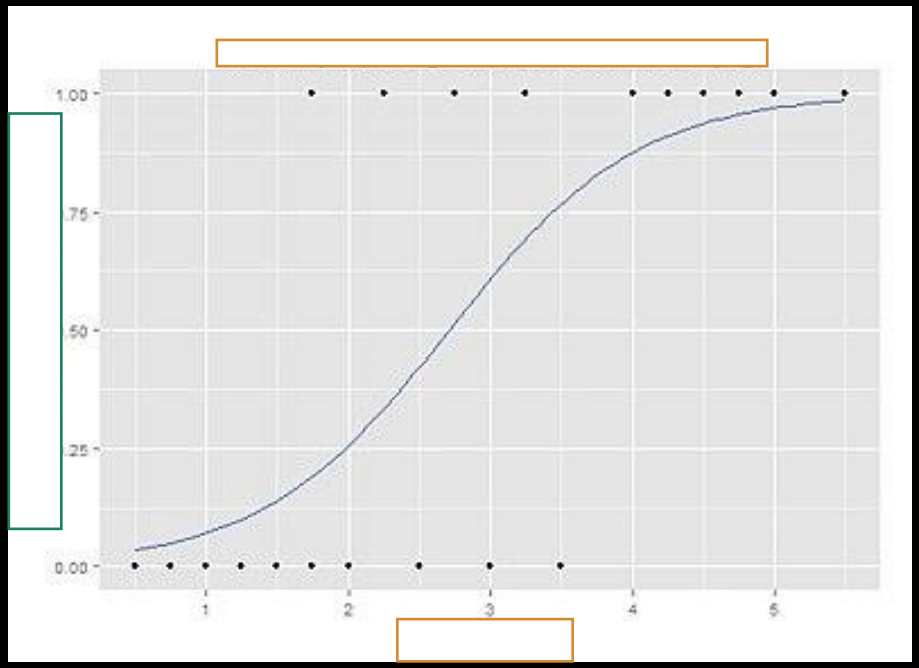
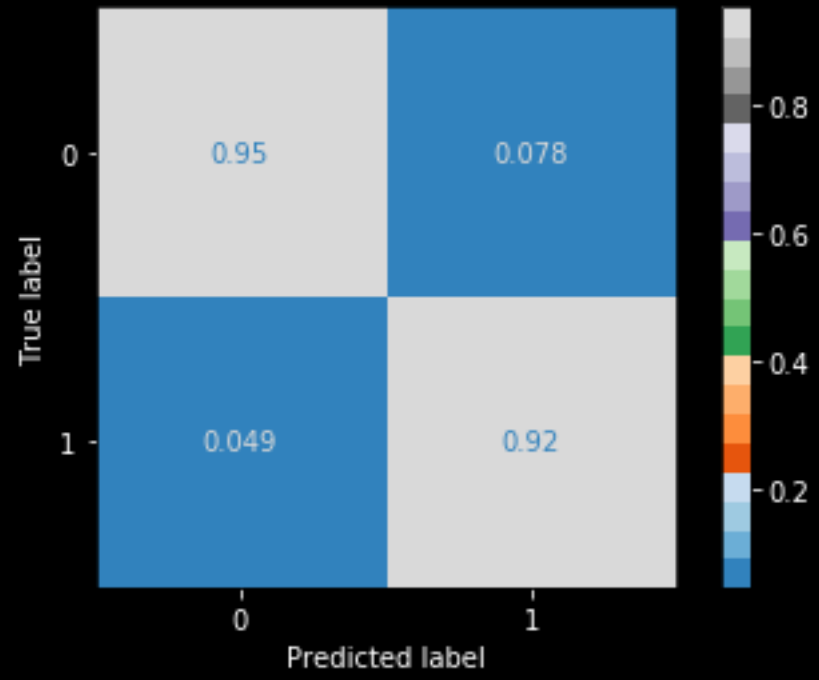


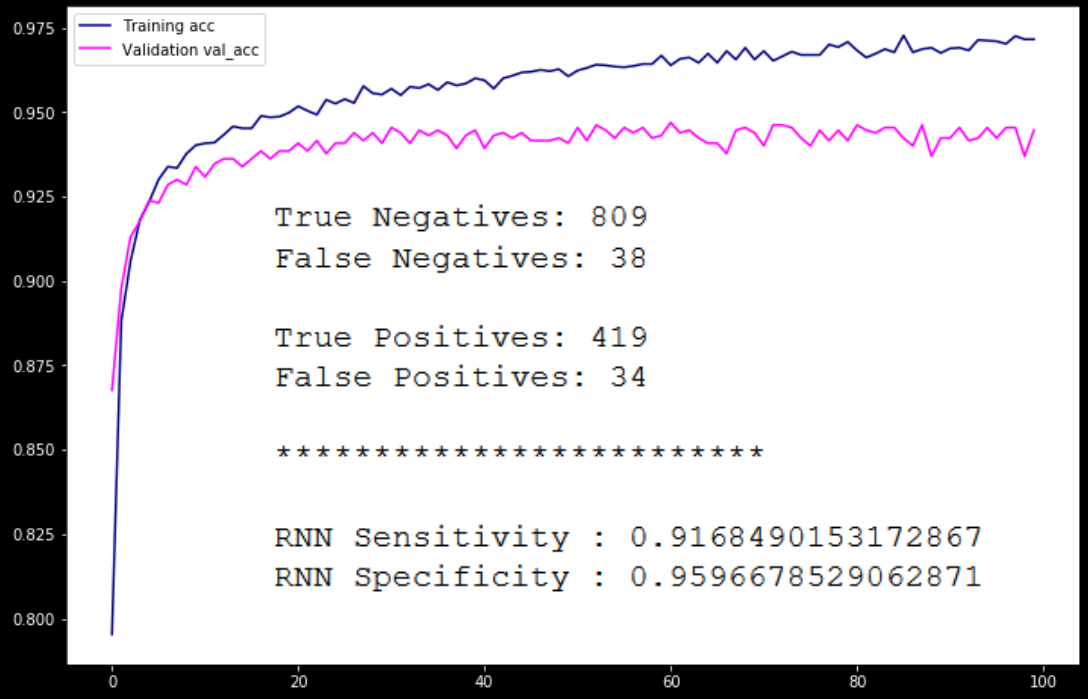
Photo Credit: Wikipedia

Random Forest VERSUS Neural Network

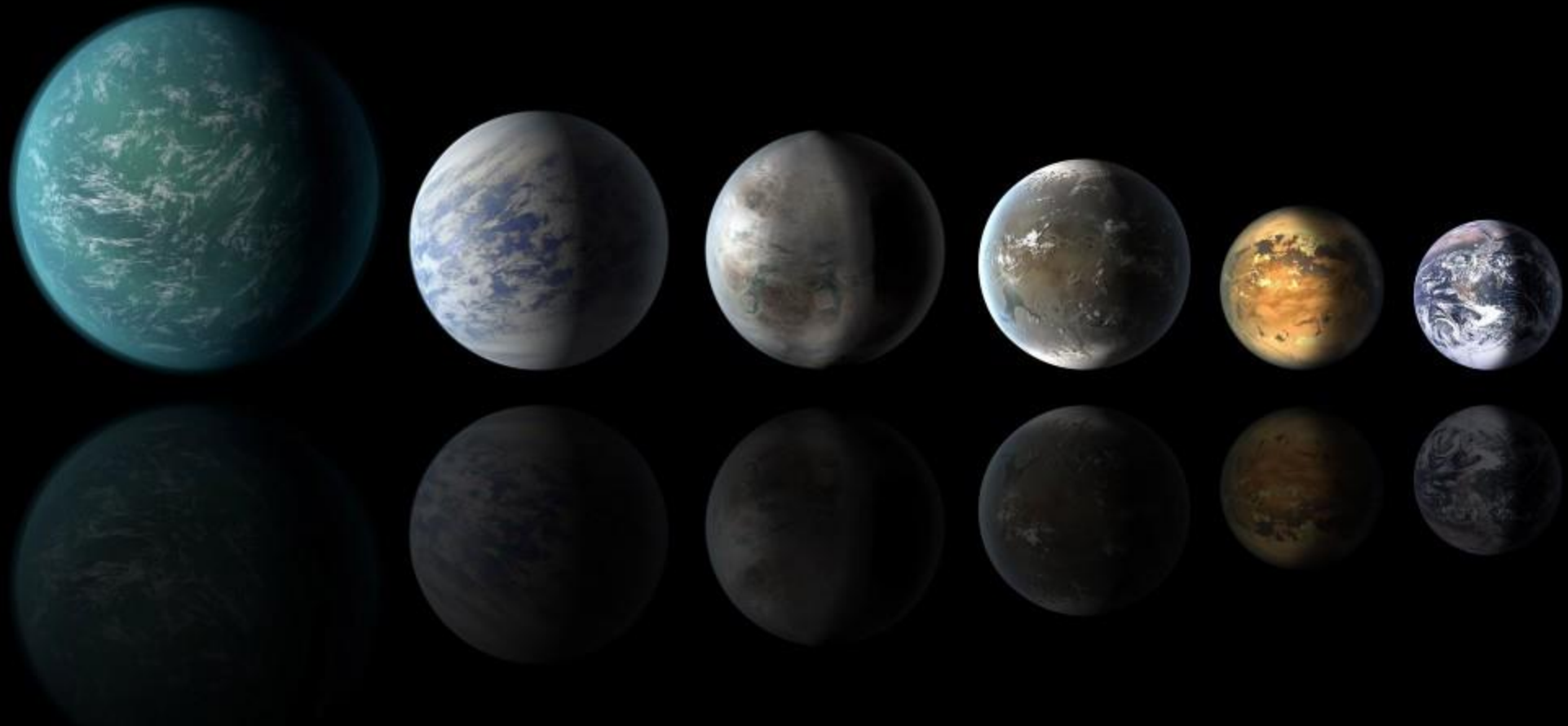
Random Forest Classifier…94.08% Acc



Recurrent Neural Network…95.23% Acc



So…Is there anybody out there?



Habitable-zone planets with similarities to Earth: from left, Kepler-22b, Kepler-69c, the just announced Kepler-452b, Kepler-62f and Kepler-186f. Last in line is Earth itself. Illustration by NASA/Ames/JPL-Caltech

Prediction Time

- There are 2,245 Kepler Objects of Interest yet to be classified

	Logistic Regression	Random Forest Classifier	Recurrent Neural Network
Accuracy	91%	94%	95%
Sensitivity	89%	92%	92%
Specificity	93%	95%	96%
Predicted New Exoplanets	1029	811	955

Which of these things is not like the other?

- For a sanity check, we created a table, a pandas DataFrame, containing the predictions from each model
- We then asked for only the rows containing KOI's where all three models predicted the existence of an Earth-like exoplanet
- This resulted in 664 newly-identified exoplanets, about 630 of which should definitely be planets based on a roughly 95% specificity and 90% sensitivity in the model

Conclusion – WE FOUND 39 NEW PLANETS!

- By adjusting our probability threshold to 99% using the new function created in the lebowski.py library called tell_the_truth(), we see results of 0 false positives on the validation data, and 39 predicted exoplanets on the unknown data. These 39 planets exist on all three of our model outputs as well.

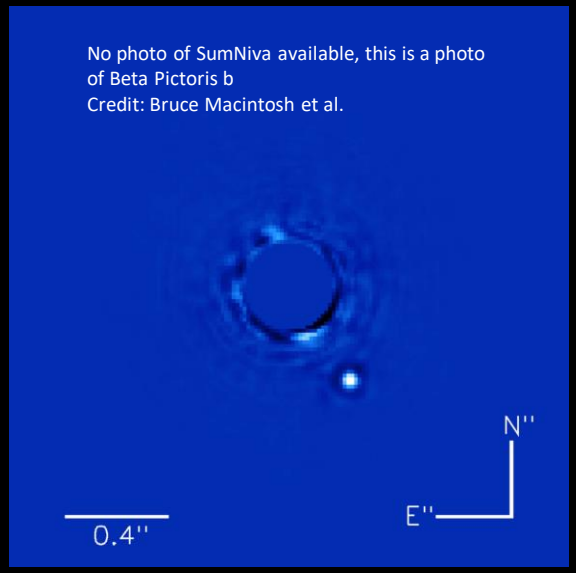
Next Steps

- This study is certain that there are about 630 more exoplanets out of the 2245 unknown candidates yet to be confirmed by CalTech or NASA.

- Next I will be creating a Python package that can be employed in classifying possible exoplanets observed in the TESS project using knowledge from Kepler and K2.

- I hope to publish findings in the future, and I hope that my TESS-assistant library can be used by NASA and CalTech much like the K2 library known as EDI-Vetter is used today.

And finally, I'd like to introduce



- The newest Identified Exoplanet, SumNiva

- Named after my loving partner, Summer Nicole Vanslager, this planet is designated with kepler id 5709725, has a timespan of about 86 ½ Earth days between transits, when it passes between Earth and its star, which is located about 3/4 of the distance from SumNiva as our Sun is to Earth.
- SumNiva's parent star is about 9/10 the size of the Sun, and SumNiva's transit between Earth and that star, the period of time in which we can see it from our solar system lasts for about 7 ½ hours.
- Declination 40.934769, Right Ascention 293.123410

Questions?

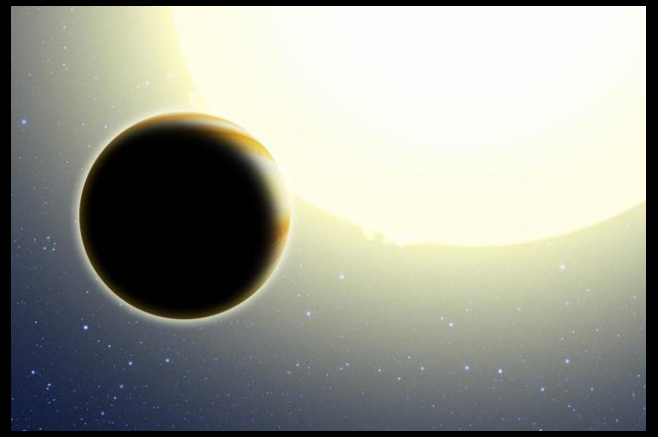


Photo Credit : David A. Aguilar / CfA

In [91]:

```
y_pred = rf.predict(test_sc)
y_probas = rf.predict_proba(test_sc)

ambit = .99
exoplanet_truths = []

for i in range(len(y_pred)):
    exoplanet_truths.append(dude.tell_the_truth(y_pred[i],
                                                y_probas[i],
                                                ambit, 1))

np.sum(exoplanet_truths)

```

Out[91]: 39