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

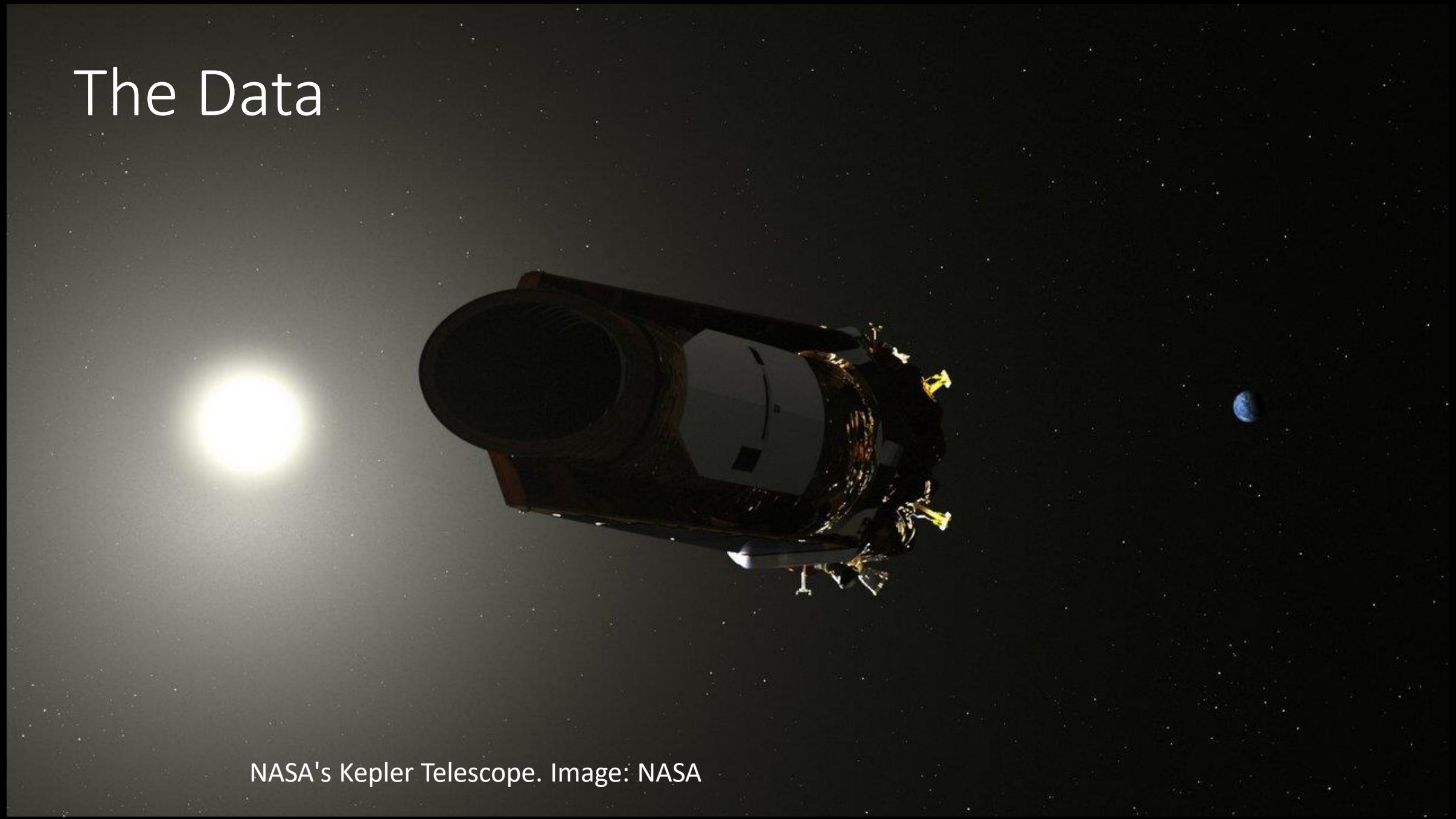
Data Science Solution

- 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 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

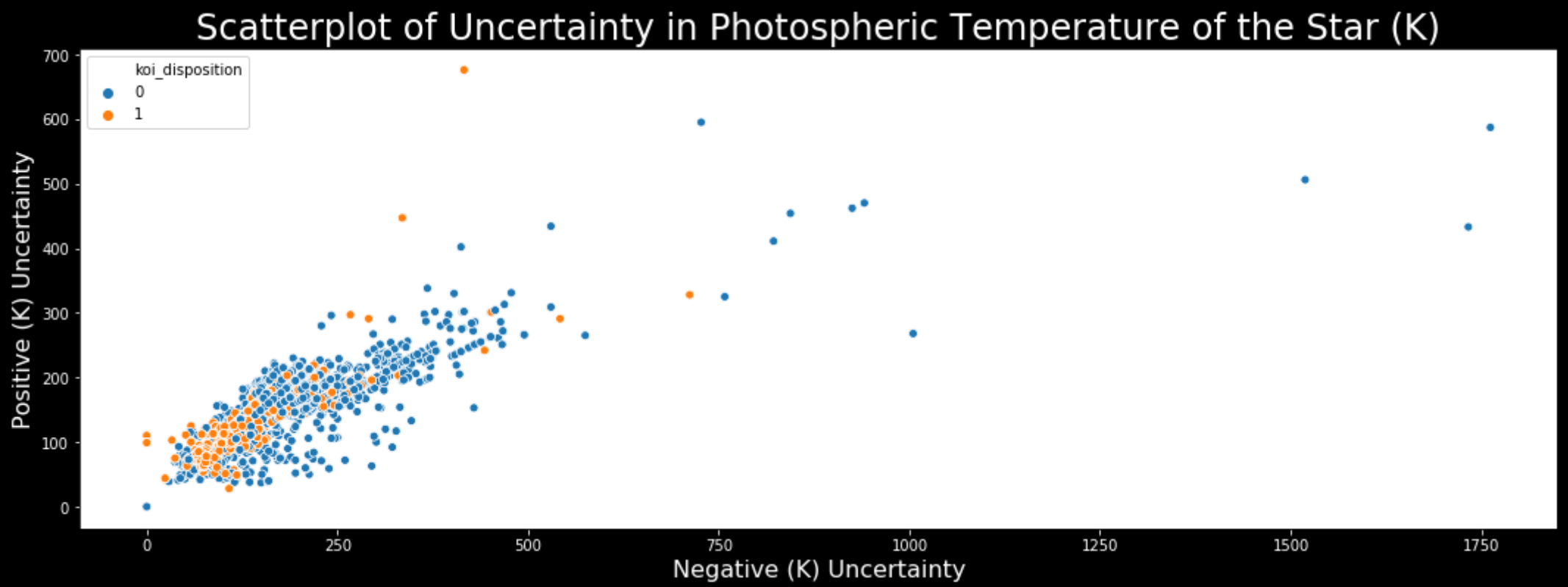
- 69 columns of data, most are quantitative
- 35 Features in our classification models
- 4212 of 8744 have been certified as NOT exoplanets
- BASELINE 35% Positive (exoplanets), 65% Negative, 2245 unknown

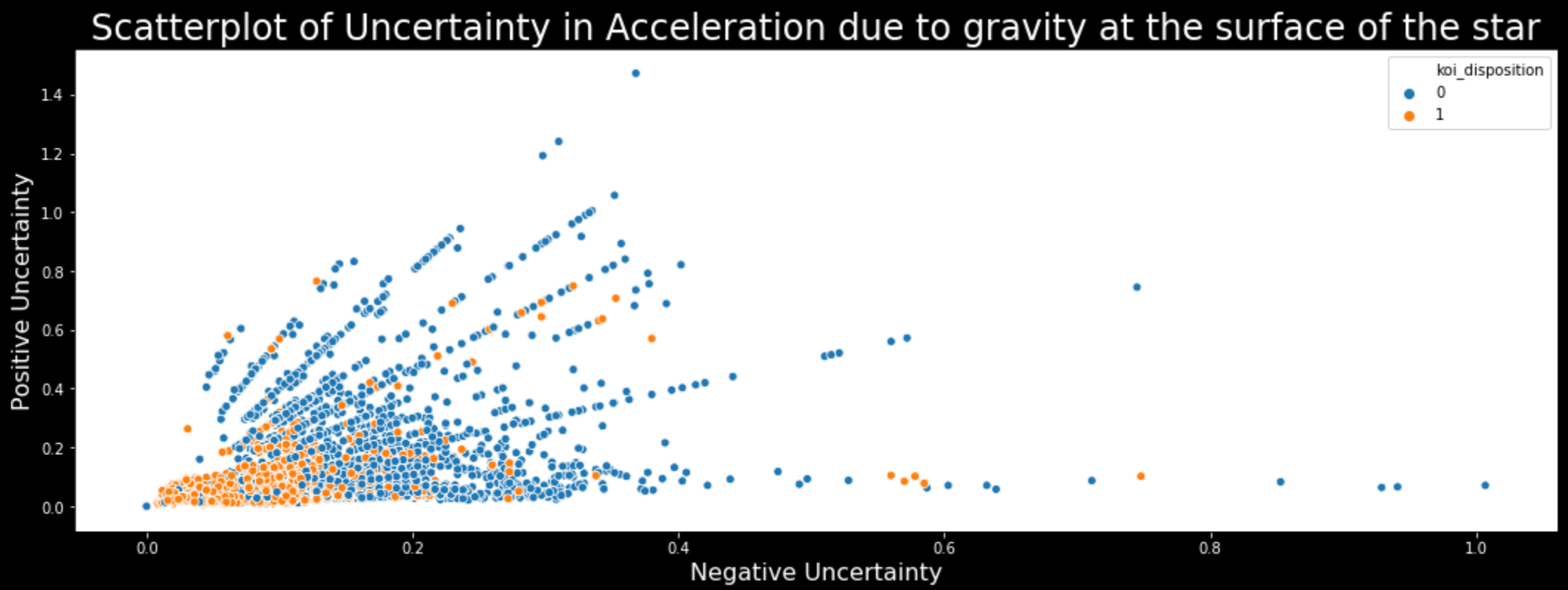
- Strongest Correlations:
  - Uncertainty in Photospheric Temperature
  - Uncertainty in Acceleration due to Gravity
  - Uncertainty in Hours of Transit Duration
  - Acceleration due to Gravity
  - Equilibrium Planetary Temperature
  - Transit Depth
  - Star Temperature

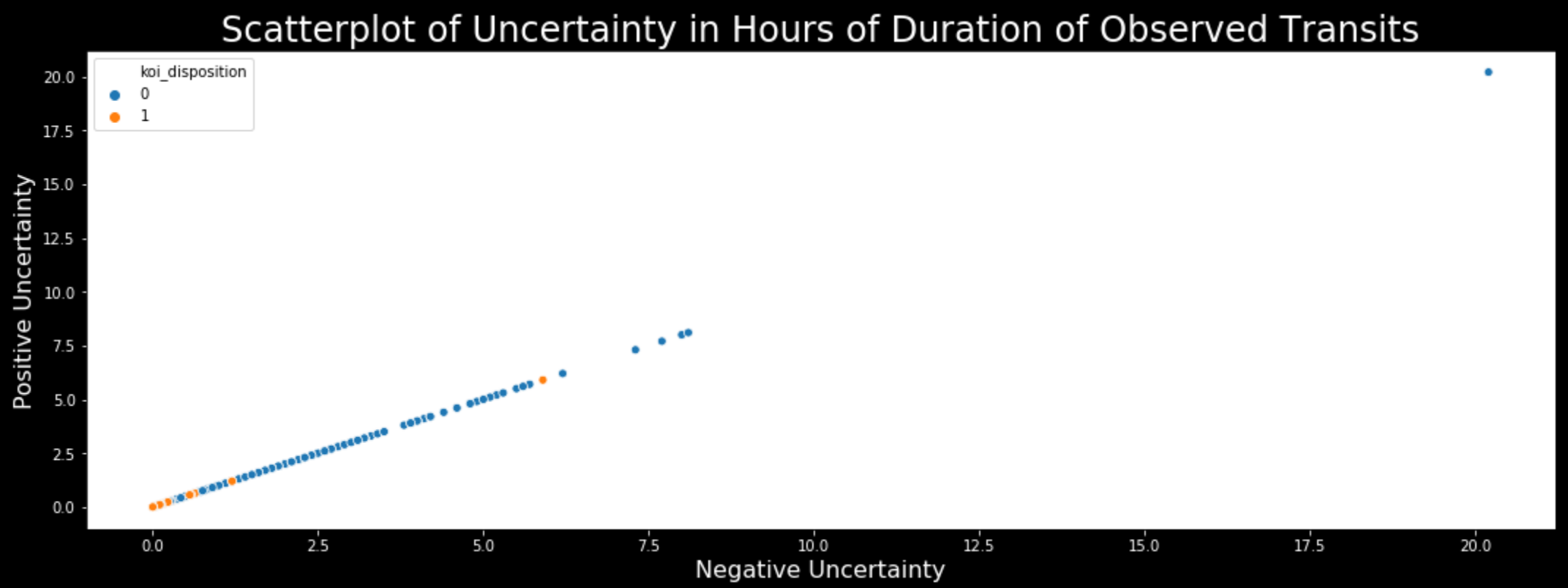
Correlation Matrix of Disqualifier Flags

Self-fulfilling models are bad

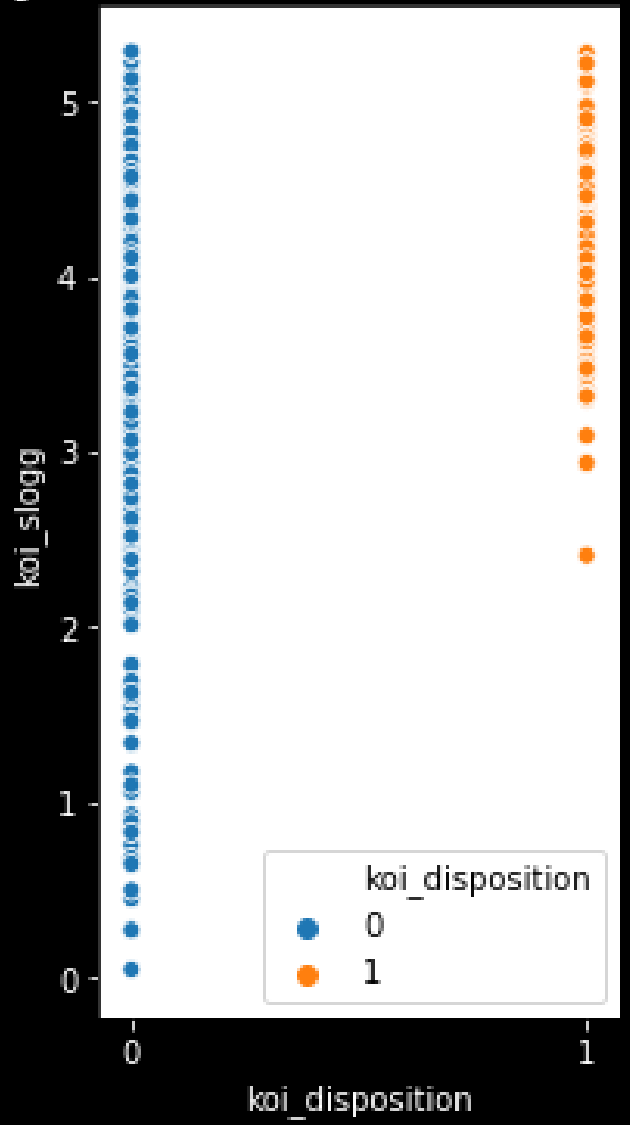
- We must drop these flag columns since our model would not work on unknown or new data if it was trained on these flags.
- We also drop out id numbers and any other non-empirical data
- Look now at the strongest correlations


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

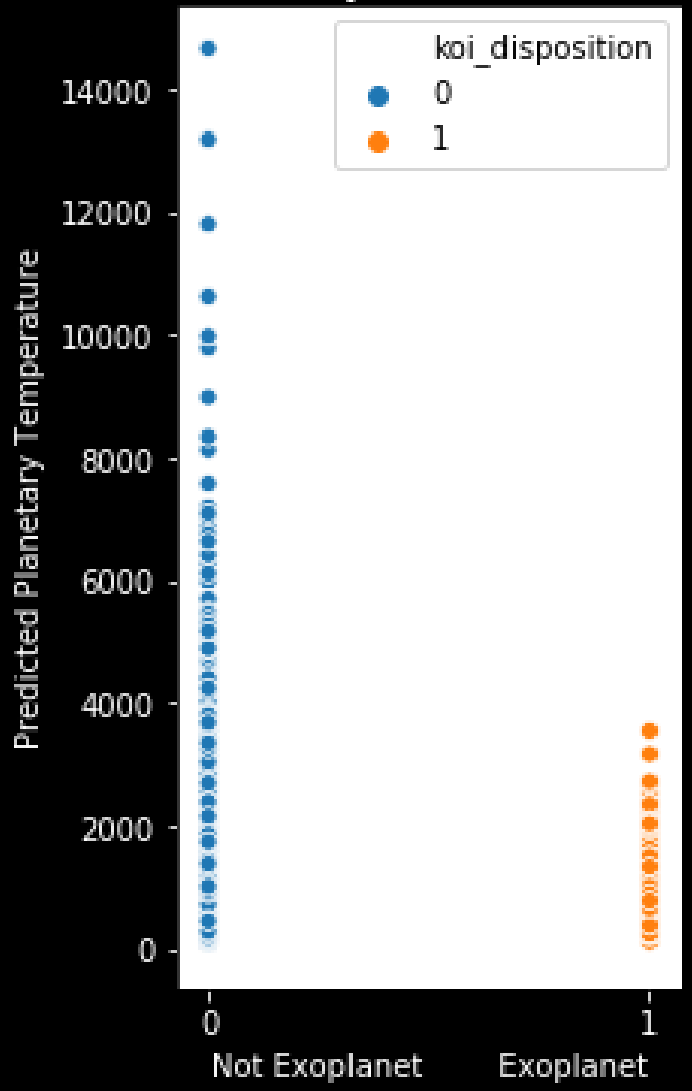

Scatterplot of Uncertainty in Acceleration due to gravity at the surface of the star
koi_disposition
0
1
Positive Uncertainty
Negative Uncertainty
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4


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

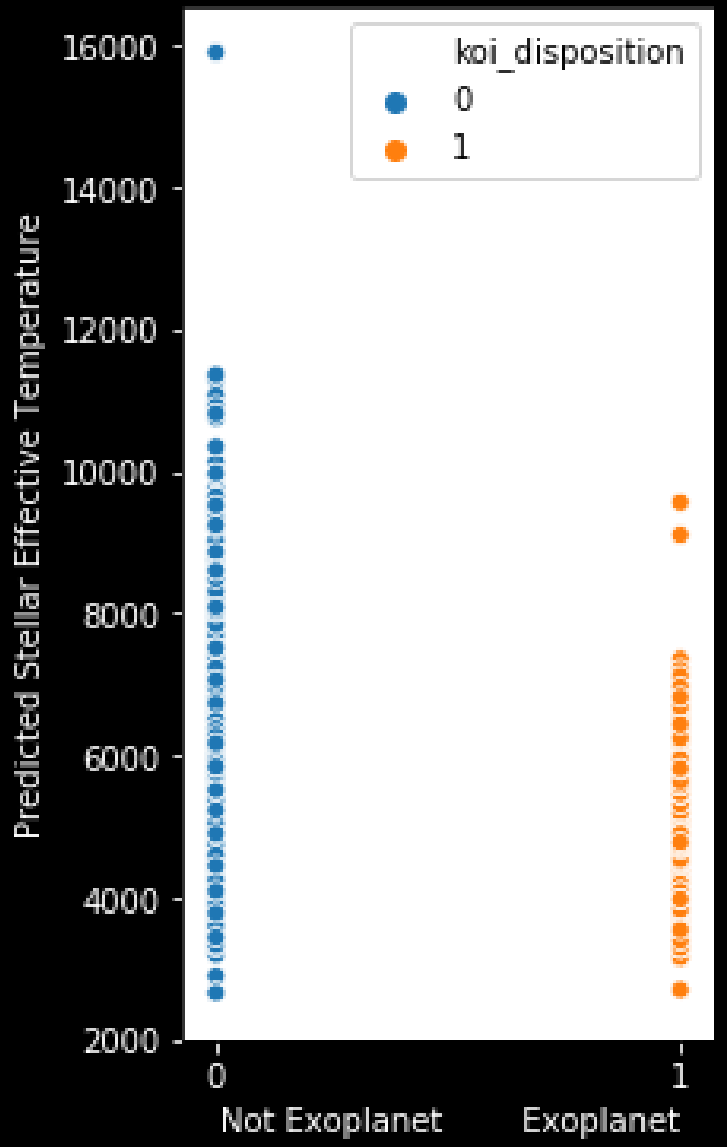
Surface Gravity of the stars in the Kepler Study



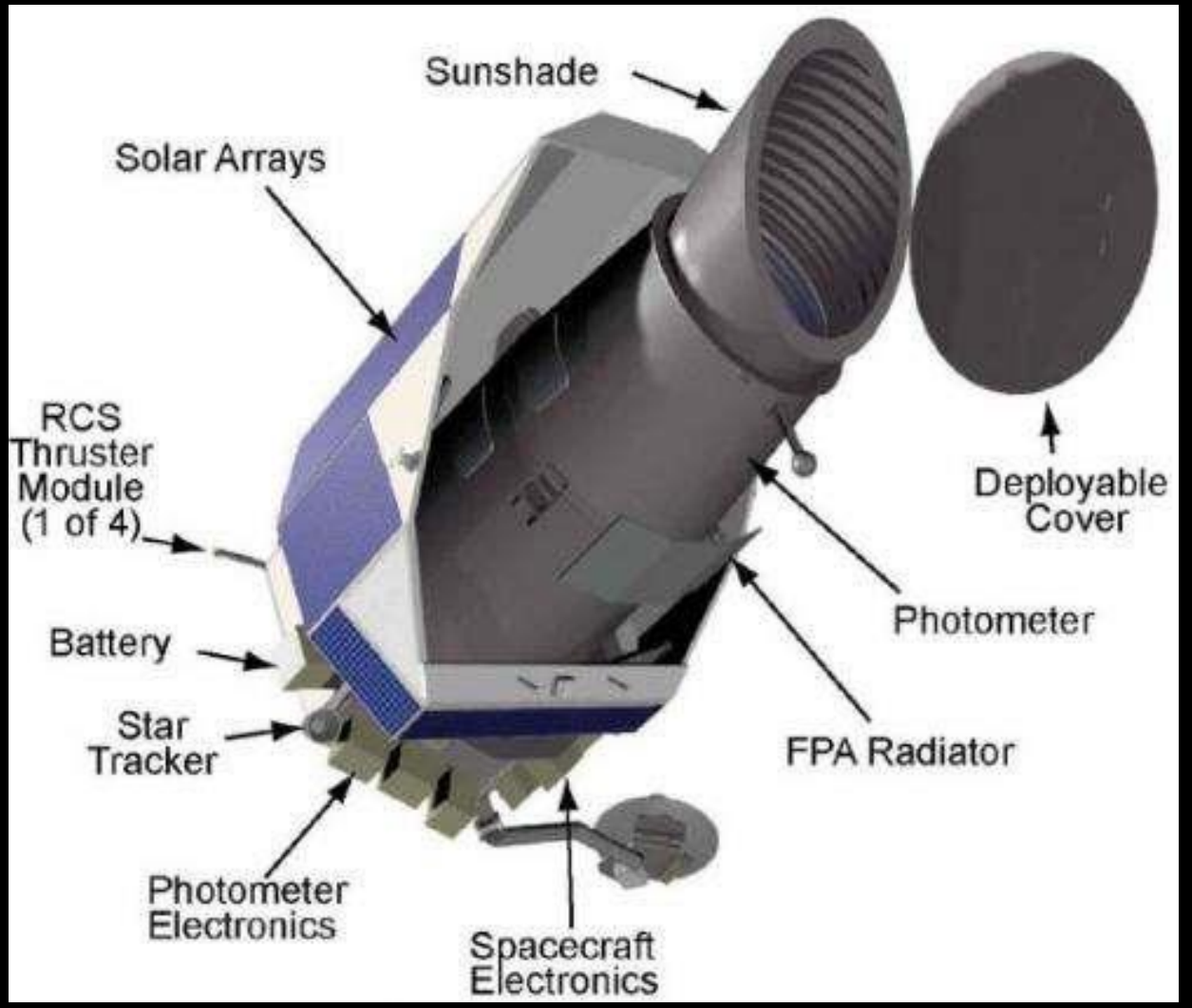
Predicted Equilibrium Temperature on possible Exoplanet



Predicted Stellar Effective Temperature



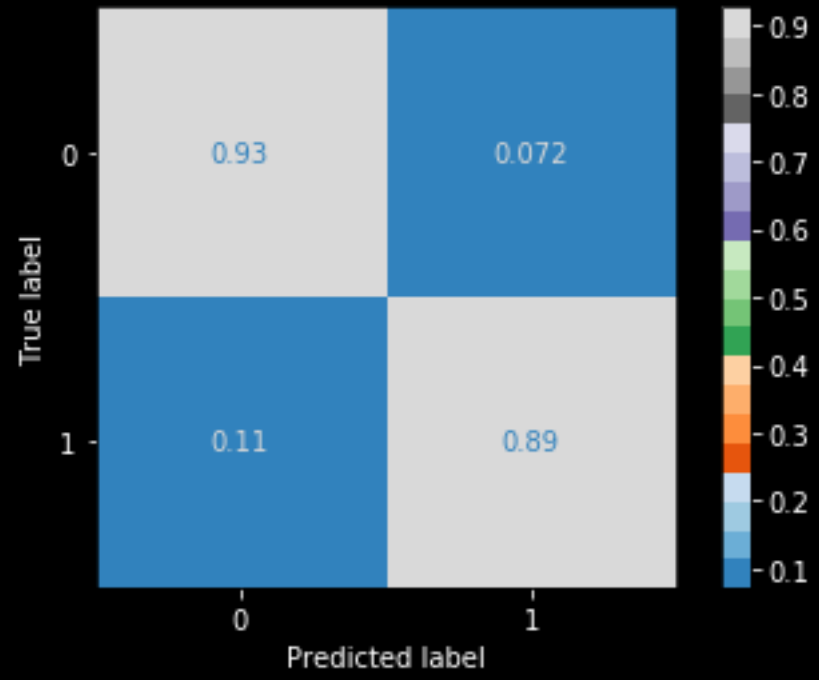
The Models



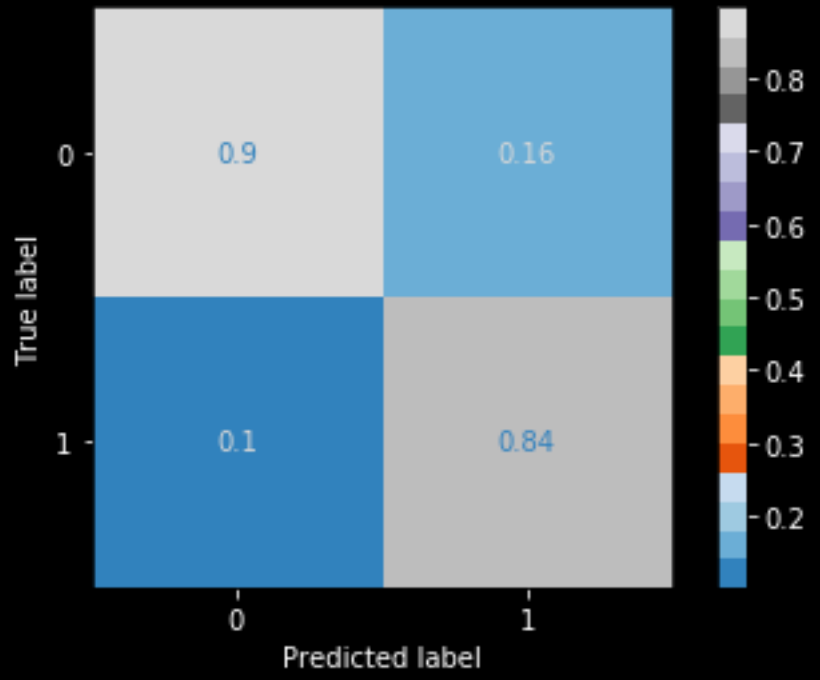
A diagram of the Kepler space telescope. Credit: NASA

Two Logisitc Regression Models

All Features…91.38% Validation Acc



Limited … 87.92% Validation Acc



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.

We can either accept that our model will have about 89% Sensitivity, or 11% False Positives returned, or we can write our own function to limit the return by adjusting the probability threshold for a positive classification.

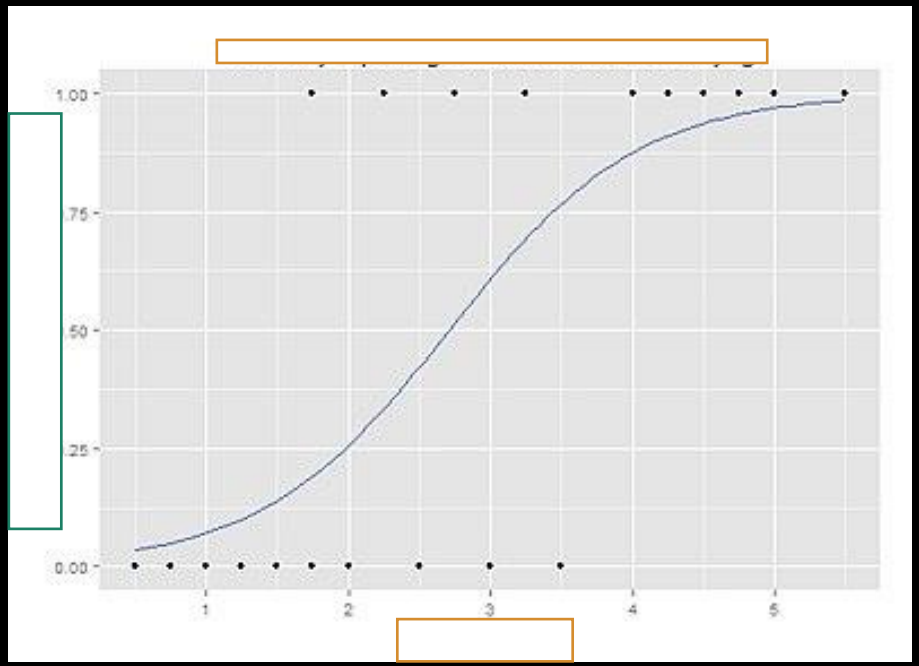
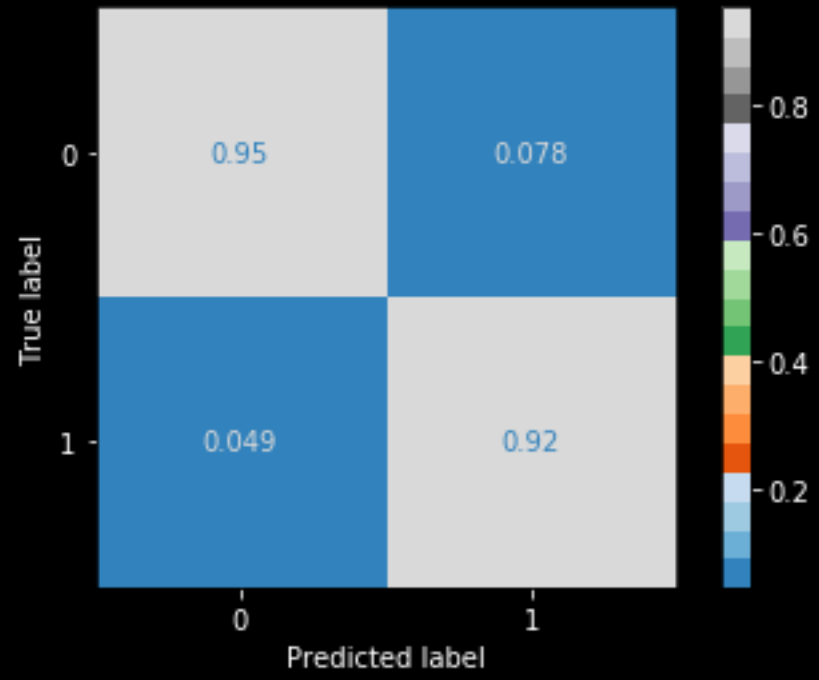


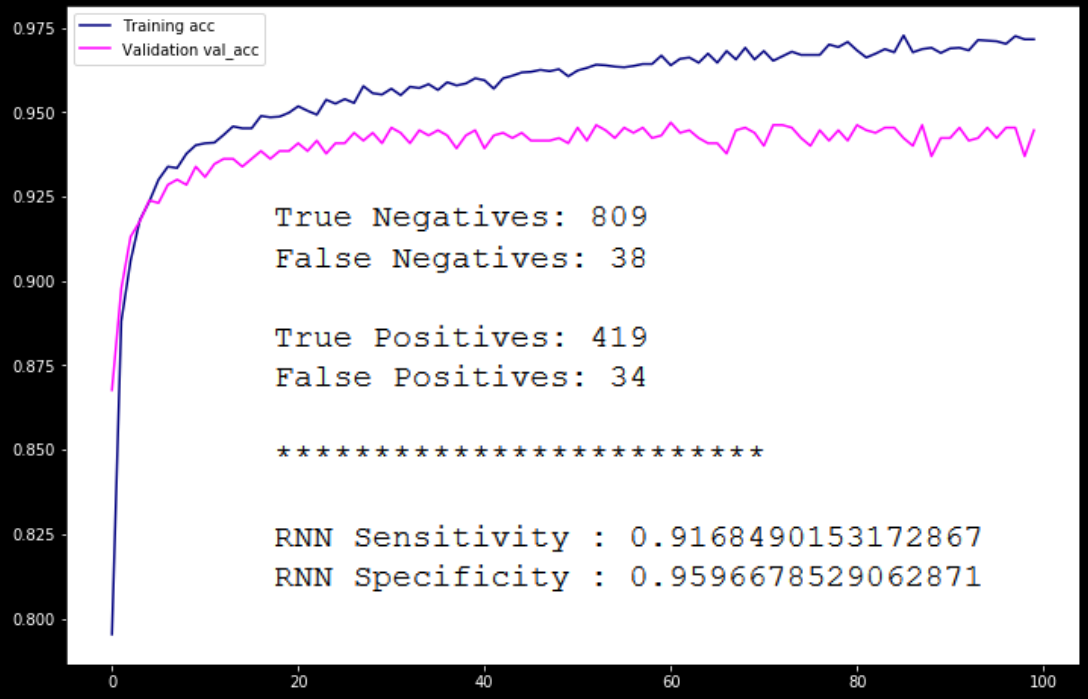
Photo Credit: Wikipedia

Random Forest VERSUS Neural Network

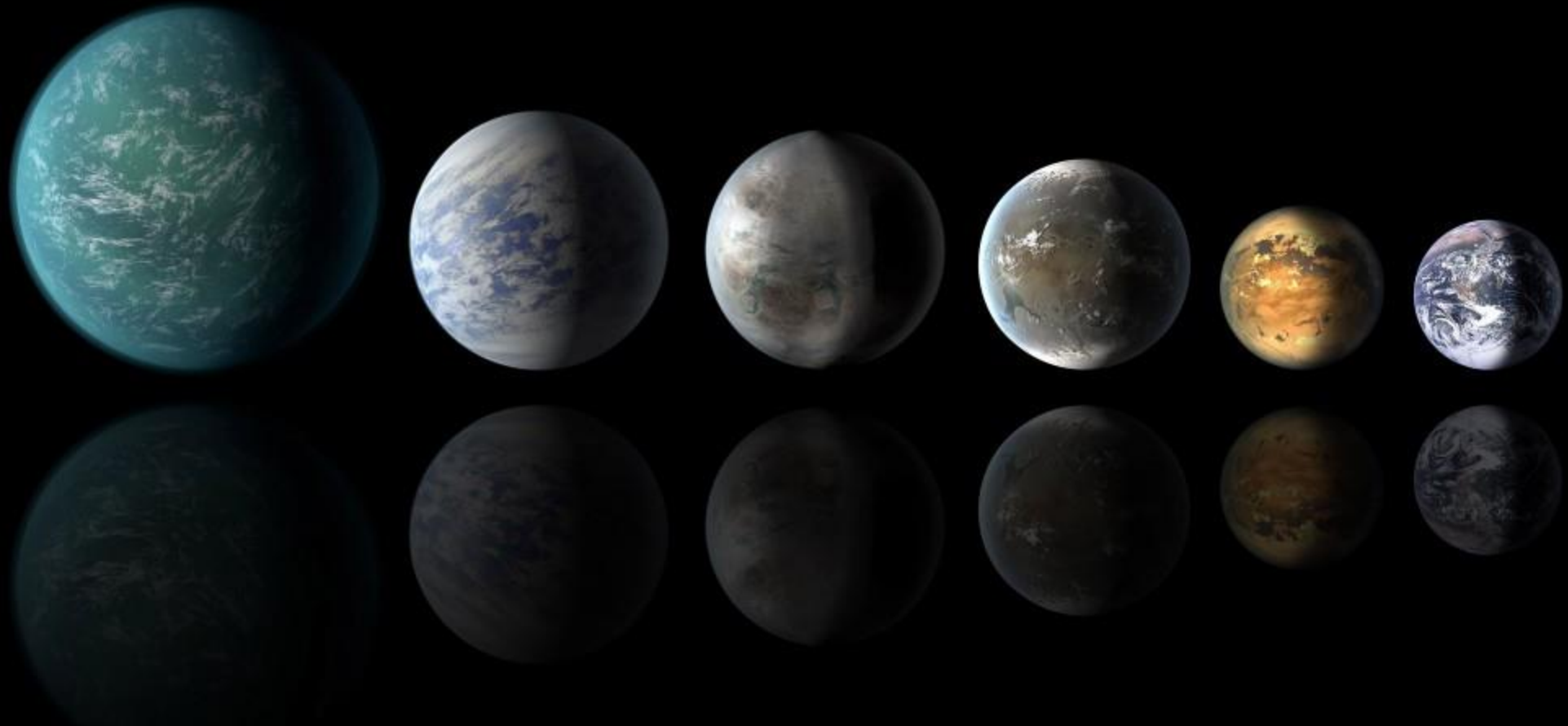
Random Forest Classifier…94.08% Acc



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 682 newly-identified exoplanets, about 613 of which should definitely be planets

Conclusion

- More tweaking should be done to make the highest confidence predictions against specific exoplanet locations, including limiting the models to zero false-positive predictions on training and validation data.

- Regardless, this study is certain that there are about 600 more exoplanets out of the 2245 unknown candidates yet to be confirmed by CalTech or NASA.

- A packaged version of this model is now available in Python and can be employed in classifying objects of interest in the K2 project and the TESS project next. We hope to publish findings in the future.

Questions?

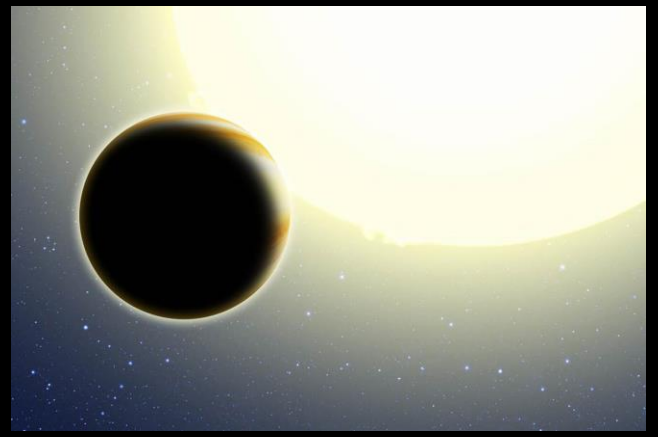


Photo Credit : David A. Aguilar / CfA

	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