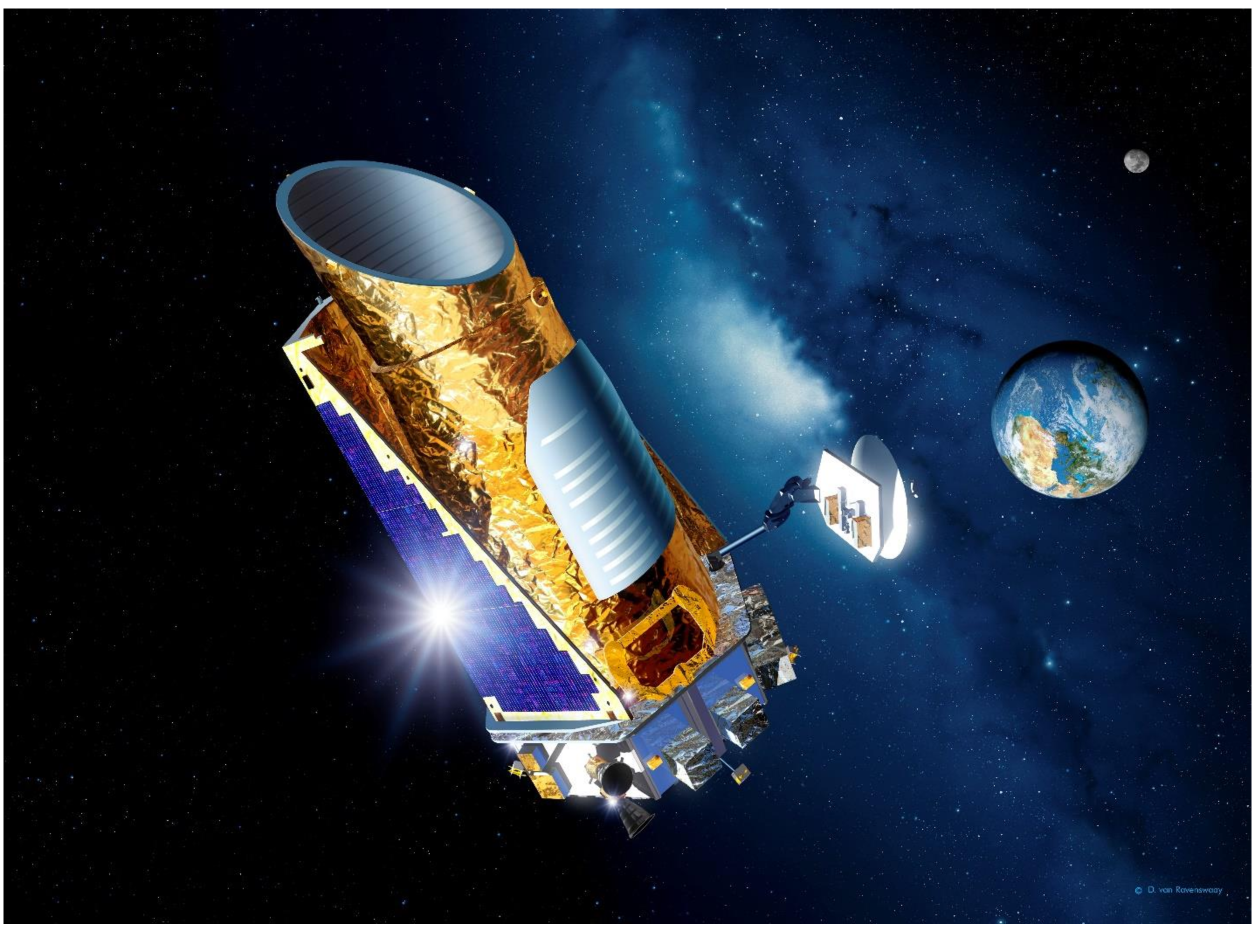
PROJECT PROPOSAL PERSONAL CHALLENGE ARTIFICIAL INTELLIGENCE

Exoplanet detection

By Martin Kiryakov

Student #: 3780090



The Kepler space telescope, image credit: Time magazine.

1. PROJECT DESCRIPTION

This project will be using a database containing thousands of observations made by the NASA's Kepler space telescope. This dataset has a particularly interesting shape – 5087 observations, all divided along a massive 3198 features. This will allow for a lot of very interesting observations and visualizations to be made, as well as posing some challenges due to the extreme number of features. This dataset comes with the benefit of being split into two – a training and testing subset, as well as being a labeled dataset. You can find this dataset, along with a more thorough explanation of the data within it here.

1.1 WHAT?

This dataset includes a massive number of measurements of the light levels emitted from stars. By looking at the periodic nature and seasonality of these light measurements, we can detect patterns. Different patters can mean different things. For example, random dimming of light levels may mean a cluster of small bodies orbiting a star, such as an asteroid belt (much like our solar system's Kuiper belt) or simply a gas cloud. If the light levels of the star do not change drastically, it likely means there's nothing of note orbiting around it. Finally, if the light levels dip and increase periodically, forming a sort of sine wave pattern, then we can confidently say there's an exoplanet orbiting the star.

1.2 HOW?

This database will be explored using the various different methods learned during the semester. This includes, but is not limited to, Pandas, Seaborn, Pyplot and more. I will attempt to visualize the data in a digestible way, as the dataset itself is quite difficult to process by humans. By visualizing different aspects and scales of it, I can help the reader to better understand what we're dealing with.

1.3 WHY?

Exoplanets in general are very important to us, as they contain a vast amount of knowledge about the gigantic cosmos, but also about our tiny little planet – Earth. That's why forming an algorithm that can reliably find exoplanets on distant stars can let our scientist devote their time to studying those exoplanets. From them we can learn how our own planet formed, how life first arose or even finding alien civilizations (although the latter is quite hopeful).

2. CHALLENGES

I think that the biggest challenge here will be to make sense of the features. As there are so many of them, extracting information in a human-readable way could prove quite challenging. Additionally, these values represent stars from every single class – from small white dwarf stars to red hypergiants, meaning that their values may need some complex scaling to ensure proper algorithm training.

3. EXPECTED OUTCOME

The main goal here is to learn how to manage large datasets with complex data. While the final result will likely be an algorithm that can achieve a good prediction score using the test subset, the actual learning outcomes from this challenge will be far greater. This should allow me to improve decision making skills related to ML, as well as prove that my implementation capabilities are advancing.

Finally, being a huge astronomy-nerd, I think that exploring this dataset will not only be educational, but also quite entertaining.