Nathaniel Campbell, Brendan Gregory, Julian Matos, Matthew Preston

ML Project Proposal

I. Abstract and Goals

This project involves trying to find exoplanets using data generated from Kepler. Kepler is a NASA telescope that collects light data on stars and exoplanets are planets revolving around stars that our not our Sun. This data is publicly available for research and we plan to use it to try to identify exoplanets. This would be a classification project using supervised learning where we look at different light patterns and try to determine whether it is an exoplanet or not. One way to train on the data set would be to use perceptrons but we can use a wide variety of different techniques to see which generates the lowest error.

II. Deliverables

In terms of deliverables for the project we are looking to produce a paper and presentation outlining our approach, findings, and performance. This is a project that will practically apply the machine learning techniques we have explored in class. We expect to be able to deliver code that works to identify different exoplanets with a reasonably low error rate. We can also deliver some graphs pertaining to our actual light samples and what techniques we used to try to smooth the graphs.

III. Challenges

We expect numerous challenges going through this project. Our first challenge is condensing the data into a format that we can do testing on. We are working with raw data from the telescope and that raw data has a lot of noise. While we plan on trying to attempt some machine learning techniques on the raw data we also believe that transforming the data could potentially lead to better results. One possible transformation would be a fourier transform as this is a light signal and we could decompose the different signals to try to improve performance.

In addition, it is possible that there are other interesting things we could possibly find and we could in fact accidentally stumble across new patterns or strange phenomena. This could hamper classification as it could greatly increase our bias and thus our overall error. We don't expect this to be a big problem, but considering how little we know about space and exoplanets, it's possible.

A third problem is that of the amount of data. The Kepler telescope surveyed about 150,000 stars, which is certainly a lot, but it might not be enough for certain models, such as a neural network. Finally, many of the papers we read about describe the problem of finding false

positives with their models, or boast about their model not finding false positives. We expect this might be a problem with our project as well, considering the aforementioned problem of noise. Our references address some of these problems.

IV. Performance Metrics

We believe that looking at the error rate and a confusion matrix might be the best way to determine performance. In terms of error rate, performance using a deep neural net has yielded error rates as low as 3% which is pretty amazing. We believe that with a less complex model an error rate such as 20% should be readily achievable. We also want to look at the false positive rate in the confusion matrix as it could lead to missing exoplanets. We want to keep a low false positive rate but that may not be possible as some papers have said that they have a rather large false positive rate (>60%). Thus we believe a reasonable goal would be 70%. There is a relationship between the false positive rate and the stringency of the testing as allowing for a greater pool of candidates leads to getting more of the planets classified as planets but also a greater pool of noise being classified as planets.

V. Timeline

Nov 1: Create a working model that correctly categorizes planets at with an error lower than 40%.

Nov 15: Attempt different smoothing techniques or transformation techniques to further lower error and try to improve performance and false positive rate.

Stretch Goal: Attempt to modify or rewrite model to display number of planets orbiting star.

VI. References

A. Papers:

1. http://adsabs.harvard.edu/abs/2019AAS...23340507A / https://ui.adsabs.harvard.edu/abs/2019AAS...23340507A/abstract
    a) Author: Ansdell, Megan
    b) Title: How can machine learning contribute to mining Kepler data?
    c) Venue: AA(University of California, Berkeley, Berkeley, CA, United States)
    d) Volume Number: 233
    e) Date: 01/2019
    f) Publisher: The American Astronomical Society

g) This paper describes pros and cons of certain machine learning techniques and models, specifically on the Kepler data, that we would use for inspiration.

2. https://ui.adsabs.harvard.edu/abs/2015ApJ...806....6M/abstract / https://iopscience.iop.org/article/10.1088/0004-637X/806/1/6 / https://iopscience.iop.org/article/10.1088/0004-637X/806/1/6/pdf
    a) Authors: McCauliff, Sean D.; Jenkins, Jon M.; Catanzarite, Joseph; Burke, Christopher J.; Coughlin, Jeffrey L.; Twicken, Joseph D.; Tenenbaum, Peter; Seader, Shawn; Li, Jie; Cote, Miles.
    b) Title: AUTOMATIC CLASSIFICATION OF KEPLER PLANETARY TRANSIT CANDIDATES
    c) Venue:
        (1) NASA Ames Research Center, Moffett Field, CA 94035, USA
        (2) SETI Institute/NASA Ames Research Center, Moffett Field, CA 94035, USA
    d) Volume Number 806, Issue Number 1
    e) Pages 806 - 819
    f) Date: 3 June 2015
    g) Publisher: The Astrophysical Journal
    h) “This paper presents the application of machine-learning techniques to the classification of the exoplanet transit-like signals present in the Kepler light curve data”. We hope to take inspiration from the techniques they employ.
3. https://ui.adsabs.harvard.edu/abs/2018MNRAS.474..478P/abstract / https://academic.oup.com/mnras/article/474/1/478/4564439 / https://watermark.silverchair.com/stx2761.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAnswggJ3BgkqhkiG9w0BBwagggJoMIICZAIBADCCAl0GCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMJN6Xh-ang1pDrWsOAgEQgIICLiN4oQhc468ax7hpZYNEpVV578QLFoAJAYkIyc2xORMaLR2zSUjLHXlBmV1xs8Yl261M11i6lK8dNZ3bsRDpf7A4rISJ9ErSt_UFVGUTAjL7AS9Zzv6a4hGpxt58vc57FLBMb7NNNgCGjYTj4QgL8vZ2p-Iixl6O7S3Cx8urbANsCvjQN_1ab-X-ICL1vWvm9YaKkD9qRD8XuwEdD1paCPNH9PhM0V0_XradZNf4qtVVUfu2wHdBpclCWunpZnjT7La-nee74XuQOarme0MUA_-JljdDP9soxL4rrHsdBOwbTbAY69OtNMheNMcstLv8Xz81ROLWfMmpyhZf4H5CmVfr6fvxf2F4hM63we7iJLlQym1WtrkVrIqCqOxvO1Twl2aIm9ylmVzaTzMsoIifyFSGskklc6vvBSgI8FpIRgeJPvLdoG_hn-4VTQkD3bD_zbKazZ0lK6t9xc4bjB4Ooj5RDJcg9ZbhqTqCQP6hgWnVEINNyOlYtAy_hl

3WIZKnF6lbLwNjnEhqPNV_qcDE5_SLkqbxc6kKh5e943CxFpzB9jYp0H5_8wK1PxJHQU1CtXwH0qGOL1q-po-VnF-4ArMYhU9pDBcoBeBFHDiu5RDB2KhgOB0ZCl1Ga9FMoAbIjhdnB50cOmqNCngxK9usNJ3rgq9G93hvR4HyGexXqGmv88BeC5tPymZCvWZmv1xurQECu3F02S5dM4nyxTtQrYrLhtNO-KbgSZMdmlNaow

   a) Authors: Kyle A. Pearson, Leon Palafox, Caitlin A. Griffith.
   b) Title: Searching for exoplanets using artificial intelligence
   c) Venue: Oxford Academic
   d) Volume Number 474, Issue Number 1
   e) Pages 478 - 491
   f) Date: February 2018
   g) Publisher: The Astrophysical Journal
   h) This paper describes techniques to detect planets despite a noise, which is a huge problem that we expect to encounter.

4. https://ui.adsabs.harvard.edu/abs/2018nova.pres.4341K/abstract / https://aasnova.org/2018/12/07/using-machine-learning-to-find-planets/
   a) Author: Kohler, Susanna.
   b) Title: Using Machine Learning to Find Planets
   c) Venue: NOVA
   d) Date: 7 December 2018
   e) Publisher: The Astrophysical Journal
   f) This article describes Exonet, a model that has 97.5% accuracy, meaning that “97.5% of its classifications exoplanet or false-positive are correct” using transitory data. We hope to use the paper (next citation) described in this article to help us cut down false positives as well.
5. https://ui.adsabs.harvard.edu/abs/2018ApJ...869L...7A/abstract / https://iopscience.iop.org/article/10.3847/2041-8213/aaf23b / https://iopscience.iop.org/article/10.3847/2041-8213/aaf23b/pdf
   a) Authors: Megan Ansdell , Yani Ioannou, Hugh P. Osborn, Michele Sasdelli, Jeffrey C. Smith, Douglas Caldwell, Jon M. Jenkins, Chedy Räissi, and Daniel Angerhausen.
   b) Title: Scientific Domain Knowledge Improves Exoplanet Transit Classification with Deep Learning
   c) Venue:
      (1) 2018 NASA Frontier Development Lab Exoplanet Team
      (2) 2018 NASA Frontier Development Lab Exoplanet Mentors
   d) Volume Number 869, Issue Number 1
   e) Pages 869 - 878

f) Date: 10 December 2018
g) Publisher: The Astrophysical Journal Letters
h) Same purpose as above citation.

6. https://ui.adsabs.harvard.edu/abs/2019AJ....157..169D/abstract / https://iopscience.iop.org/article/10.3847/1538-3881/ab0e12 / https://iopscience.iop.org/article/10.3847/1538-3881/ab0e12/pdf / https://arxiv.org/abs/1903.10507
   a) Authors: Anne Dattilo, Andrew Vanderburg, Christopher J. Shallue, Andrew W. Mayo, Perry Berlind, Allyson Bieryla, Michael L. Calkins, Gilbert A. Esquerdo, Mark E. Everett, Steve B. Howell, David W. Latham, Nicholas J. Scott, and Liang Yu.
   b) Title: Identifying Exoplanets with Deep Learning. II. Two New Super-Earths Uncovered by a Neural Network in K2 Data
   c) 9 Venues:
      (1) Department of Astronomy, The University of Texas at Austin, Austin, TX 78712, USA
      (2) Google Brain, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
      (3) Astronomy Department, University of California, Berkeley, CA 94720, USA
      (4) Harvard–Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
      (5) National Optical Astronomy Observatory, 950 North Cherry Avenue, Tucson, AZ 85719, USA
      (6) Space Science and Astrobiology Division, NASA Ames Research Center, Moffett Field, CA 94035, USA
      (7) Department of Physics and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
      (8) NASA Sagan Fellow.
      (9) NSF Graduate Research Fellow.
   d) Volume Number 157, Issue Number 5
   e) Pages 157 - 169
   f) Date: 9 April 2019
   g) Publisher: The American Astronomical Society
   h) This paper describes a neural network model that this team used to identify exoplanets from the Kepler data. They claim it has an accuracy of 98% on their test set and “culls false positives”.

7. https://ui.adsabs.harvard.edu/abs/2019MNRAS.483.5534S/abstract / https://academic.oup.com/mnras/article/483/4/5534/5199219 / https://academic.oup.com/mnras/article-pdf/483/4/5534/27496899/sty3146.pdf
    a) Authors: N. Schanche, A. Collier Cameron, G. Hébrard, L. Nielsen, A. H. M. J. Triaud, J. M. Almenara, K. A. Alsubai, D. R. Anderson, D. J. Armstrong, S. C. C. Barros, F. Bouchy, P. Boumis, D. J. A. Brown, F. Faedi, K. Hay, L. Hebb, F. Kiefer, L. Mancini, P. F. L. Maxted, E. Palle, D. L. Pollacco, D. Queloz, B. Smalley, S. Udry, R. West and P. J. Wheatley.
    b) Title: Machine-learning approaches to exoplanet transit detection and candidate validation in wide-field ground-based surveys
    c) Venue: Monthly Notices of the Royal Astronomical Society
    d) Volume Number 483, Issue Number 4
    e) Pages 5534 - 5547
    f) Date: 22 November 2018
    g) Publisher: The American Astronomical Society
    h) This paper describes “a combination of machine-learning methods including Random Forest Classifiers (RFCs) and convolutional neural networks (CNNs) to distinguish between the different types of signals” to parse “stellar light curves” identifying exoplanets. It could help us either in inspiring our use of similar models, or in generating more data if we need it, as that is what they did.

B. Data:
    1. All the Kepler data:
    2. https://www.cfa.harvard.edu/~avanderb/k2.html
    3. https://www.cfa.harvard.edu/~avanderb/tutorial/tutorial3.html
    4. https://exoplanetarchive.ipac.caltech.edu/docs/data.html

C. Articles:
    1. http://www.planetary.org/explore/space-topics/exoplanets/transit-photometry.html
    2. https://www.cfa.harvard.edu/~avanderb/tutorial/tutorial.html
        a) These articles are from where our initial inspiration drew.