Detection and Analysis of Exoplanets using Machine Learning Techniques

Lance Dacy[1], Aurian Ghaemmaghami[2], Aniketh Vankina[3], and Jamie Vo[4]

***Abstract*— The Earth's population continues to grow at a steady rate. The natural resources on Earth are in limited supply. The need to find other worlds that could contain life becomes more focused and intense in the past few years. Scientists desire to find exoplanets that have the features similar to Earth for sustaining life, not only for curiosity sake, but for potential new places for humans to thrive. The data collected over the centuries is so large that filtering the data becomes problematic. With new technologies that use a machine's processing power, data science methods could help scientist evaluate this data and its patterns to narrow down exoplanets that could sustain life as we know it. The data needed to model these exoplanets need to be easily consumed; this project will focus on the viability of deploying cloud data stores specifically to aid in exoplanet modeling.**

I. Introduction

Are we alone in the Universe? That is an age-old question that humans having been striving to answer since we first identified that our planet is simply a small part of a larger whole. Earth appears to be one of the planets in the whole Universe that has just the right ingredients to host life. Or is it? The science continues to explore that question "Are we alone?" and have developed a systematic approach of the course of centuries to help compile data that might just answer that question.

We live in a technology age where it is feasible to gather mounds of data about our Universe. Even better, we live in a technology age where computers can assist in stitching that data together to help find the patters that point us in the right direction for an answer. Our Data Science Team joins the ranks of numerous scientist to help analyze the myriad data gathered from the Universe to find the building blocks essential to host life as we know it. Given the quest for life as we know it, scientist narrow down the building blocks to a simple acronym called CHNOPS. CHNOPS stands for Carbon, Hydrogen, Oxygen, Phosphorus, and Sulfur. These are the base elements believed to provide the building blocks for living organisms. Finding a place in the Universe that contains these elements is much like finding a needle in a haystack.

Aside from the fact that the human race has a quest to find others in the Universe, there is also another need that needs to be sustained. We will eventually consume all the natural resources on this precious Earth. The Earth's population continues to grow at a rate of 1.05% [1].

The planet's resources are expected to cap at nearly 8 billion people. The Earth currently inhabits 7.8 billion as of July 2020. Earth's estimated timeline to maintain complex life is anticipated to cease to exist as early as 1.5 billion years. Apart from discovering life at a habitable planet, space exploration allows humankind to gain a greater understanding of the cosmos and potentially answer the question of, "Is there life beyond Earth?" [3] and more importantly, could we catapult the human race to this newfound habitable planet.

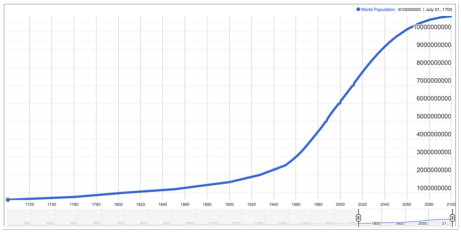

Fig. 1. World Population: Past, Present, and Future. The graph displays the population increase from the 1700's to present day. [2]

As mentioned earlier, the good news is that we have mounds of data that have been collected over the last centuries. What is lacking is a good way to determine patterns in all of the data with computational cycles of the human brain. NASA has collected Kepler data that is a repository of the type of information needed to answer some of the questions of the human race. Unfortunately, as with most big data systems, the repository is data rich, but information poor. It is. becoming increasingly difficult to filter the data to meaningful levels. The data contains information on image sources, light measures, and gravity among other things. In conjunction with the volume of data is the unknown requirements that confirm if a planet is habitable, analyzing the parameters pouring in to confirm a planet, and understanding the cosmic web that connects the galaxies which hold potential habitable planets.

This project's goal is aimed at answering the following problem: Based on the myriad data available about space exploration, is it possible to provide an estimate of the number of other exoplanets in the Universe that are able to maintain life as we know it? In addition, could a repeatable hosted data center be built that allows for the community to consume and build algorithms for this specific purpose; easing the issue of data filtering? The main contributions to this paper are mentioned in the references section which supports the understanding of the science behind this data. In the early stages of the paper, the team will discuss the motivation behind the problems statement and then continue on into the methodologies considered. More focus will be aimed at the hosting and replication of the data using Amazon Web Services (AWS) than the actual algorithms

used to determine habitual planets. Nonetheless, the team will demonstrate a few algorithm techniques to prove out the hosting platform and its feasibility. The team will then explore methodologies in deep learning and machine learning that are widely used in the application of exoplanet scientific research. This will entail creating a working model to detect exoplanet feasibility based on configurable parameters set by the scientists and showcase the platform of data that will be served against those models. The team will then move on to the various milestones of the project which are the ability to collect the data, store the data in the Cloud (AWS), configure ETL processes that can be consumed, and then deploy a few machine learning models that will consume the data store.

II. Motivation

Based on algorithms and programs provided in research, it is possible to reign in the available data and provide a predictable and repeatable method to determine whether exoplanets in the Universe are habitable. Having an affinity of space, data science, and the hosting of large data sets provides the background for such experiments to thrive. As mentioned in the Introduction, the Astronomy data community find themselves data rich, information poor. While mounds of data exists, there is so much that even filtering the data is problematic. The best solution would be to find ways to federate the data to specific fields of studies. If a scientist is trying to determine where the next Super Nova will occur, they need specific pieces of data. Another scientist looking for the best environments for nebulae will look at another set of data. The mission of this project is to find what data elements would be best suited for helping scientist determine exoplanets and find a predictable and repeatable method to host and consume that data.

The only race on Earth is the human race, we must band together, using all technologies available to help sustain the human race. Earth is on borrowed time. Population growth and natural resource consumption will end tragically for Earth at some point in the future. Humans must invest in ways to either minimize natural resource consumption or look for other parts of the Universe that we could colonize. Armageddon comes in many forms: climate breakdown, asteroid strike, zombie apocalypse, and so on. There are scientist that choose to use the Copernican Principle with statistics to figure out how long anything will last. We could use this to determine within a 95% confidence how long humans will be around. Right now, that number is somewhere between 5,130 to 7.8 million years if you assume that humans have been around for about 200,000 years [4]. This is in close correlation with the mean duration of a mammal species; which is about 2million years.

While the motivation might not be there for our current generation; technology evolves as stepping stones from generation to generation. The need to support scientists in taking the next big steps of finding exoplanets is now. The technology is ripe, the desires are there, all that is needed are tools to move us in the right direction. The next question in generations to come might be, "how do we get to said exoplanet", but technology isn't really at a point to answer that question yet. The first step is to equip scientists with data techniques that allow them to pin point areas that could provide colonization options. The sooner we do this, the sooner we can move on to the next stepping stone of technology needs.

$$(\frac{z}{1-z}) * t_{current} \leq t_{future} \leq (\frac{1-z}{z}) * t_{current}$$

$$z = \frac{1 - \text{confidence interval}}{2}$$

Fig. 2. The Copernican Lifetime Equation [4]

III. Methodology

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua [2]. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

IV. Milstones-Data Collection

The data was collected from NASA's exoplanet site which contains 287 columns, ranging from the planet's name to data concerning the galaxy where exo-planet lives in. Originally pulled from an API, due to the recent changes in the NASA site, the data can be exported to a csv. All data in the table are reviewed by a team of astronomers, verifying its accuracy. The data is collected through a number of means, including various telescopes (such as the kepler mission) or satelites such as the TESS, which is part of the NASA's explorer program. The data also includes focuses on the host stars of the planets, for deeper understandings of how the planet's access to sunlight, temperature, etc. [5].

V. Milestones-Data Storage in the Cloud

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor

sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua [6]. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum [7].

- Item 1
- Item 2
- Item 3

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur [8]. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum [9].

Lorum ipsum and another table:

TABLE I
DEANONYMIZE CLASSIFICATION

	client ACK	server data	server ACK	server data
detection rate	96%	94%	96%	94%
false negative	4%	6%	4%	6%
false positive	0%	0%	0%	0%

Lorem ipsum dolor sit amet [10], consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

VI. MILESTONES-ETL (EXTRACT, TRANSFORM, AND LOAD)

Lorem ipsum dolor sit amet, consectetur adipiscing elit [11], sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum:

1) Client: a client of the Tor network is targeted to identify it.
2) Server: the Tor onion (hidden) service is targeted to reveal its identity or to weaken it.
3) Network: the broader Tor network itself is targeted, usually via multiple malicious Tor nodes.

Lorem ipsum dolor sit amet, consectetur adipiscing elit [12], sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

VII. MILSTONES-MACHINE LEARNING MODELS

Lorem ipsum dolor sit amet, consectetur adipiscing elit [13], sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco [14] laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

TABLE II

TRAFFIC ANALYSIS THROUGH FLOW CORRELATION

	In-Lab	Tor Relay
De-anonymized	100%	81.4%
false negative	0%	12.2%
false positive	0%	6.4%

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

VIII. CONCLUSIONS

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. [15]. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

REFERENCES

[1] S. Basak, S. Saha, A. Mathur, K. Bora, S. Makhija, M. Safonova, and S. Agrawal, “Ceesa meets machine learning: A constant elasticity earth similarity approach to habitability and classification of exoplanets,” *Astronomy and Computing*, vol. 30, p. 100335, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213133719300319

[2] Worldometer, *World Population Clock: 7.8 Billion People (2020)*, 2020 (July 26, 2020). [Online]. Available: https://www.worldometers.info/world-population/

[3] S. M. Directorate, *Exoplanet Exploration*, 2020 (July 26, 2020). [Online]. Available: https://science.nasa.gov/astrophysics/programs/ExEP

[4] W. Koehrsen, *The Copernican Principle and How to Use Statistics to Figure Out How Long Anything Will Last*, 2018 (Feburary 28, 2021). [Online]. Available: https://towardsdatascience.com/the-copernican-principle-and-how-to-use-statistics-to-figure-out-how-long-anything-w

[5] “About the nasa exoplanet archive.” [Online]. Available: https://exoplanetarchive.ipac.caltech.edu/docs/intro.html

[6] T. Liška, T. Sochor, and H. Sochorová, “Comparison between normal and tor-anonymized web client traffic,” *Procedia-Social and Behavioral Sciences*, vol. 9, pp. 542–546, 2010.

[7] E. Hjelmvik, *Detecting TOR Communication in Network Traffic*, 2020 (accessed October 7, 2020). [Online]. Available: https://www.netresec.com/?page=Blog&month=2013-04&post=Detecting-TOR-Communication-in-Network-Traffic

[8] K. Kasunic, *How To Use Tor Browser: Everything You MUST Know*, 2020 (accessed October 17, 2020). [Online]. Available: https://www.vpnmentor.com/blog/tor-browser-work-relate-using-vpn/

[9] *How Tails Works*, 2020 (accessed October 18, 2020). [Online]. Available: https://tails.boum.org/about/index.en.html

[10] *Tor Browser Leaks Secure Cookies Into Insecure Backend Channels*, 2020 (accessed October 27, 2020). [Online]. Available: https://github.com/alecmuffett/eotk/blob/master/docs.d/security-advisories.d/001-torbrowser.md

[11] A. Biryukov, I. Pustogarov, F. Thill, and R.-P. Weinmann, “Content and popularity analysis of tor hidden services,” in *2014 IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW)*. IEEE, 2014, pp. 188–193.

[12] *Court Throws out Lawsuit Against Tor for Providing Anonymous Routing*, 2019 (accessed Novemeber 3, 2020). [Online]. Available: https://reason.com/volokh/2019/05/21/court-throws-out-lawsuit-against-tor-for-providing-anonymous-routing/

[13] D. Balaban, “Best Operating Systems for Anonymity: Comparing Titans,” 2019 (accessed November 6, 2020). [Online]. Available: https://hackernoon.com/best-operating-systems-for-anonymity-comparing-titans-3501fd5cba3b

[14] *Tor Project: Top Relays*, 2020 (accessed October 1, 2020). [Online]. Available: https://metrics.torproject.org/rs.html#toprelays

[15] *Tor Project: Top Relay2*, 2020 (accessed October 1, 2020). [Online]. Available: https://metrics.torproject.org/rs.html#toprelays