Information and Data Science in the Order of Nature

Wider Topics in Data Science 905F3

Felipe Martín CandNo: 260774

April 27, 2023

UNIVERSITY
OF SUSSEX

Contents

1 ABSTRACT

This essay explores the significance of information and data science in understanding the natural world, from the microscopic realm of cells to the macroscopic realm of ecosystems and galaxies. Information is encoded in DNA and is essential for the development and function of organisms, while in physics, information can be used to describe the behavior of subatomic particles and the structure of matter and energy. Data science enables researchers to process and analyze this information, generating knowledge and making data-driven decisions, therefore investing in data science research is crucial for advancing scientific discovery and improving our understanding of the world.

2 INTRODUCTION

The order of nature refers to the patterns, regularities, and structures that exist in the natural world, from the microscopic level of cells and molecules to the macroscopic level of ecosystems and galaxies. This order can be seen in the laws of physics, chemistry, and biology, as well as in the behavior of organisms, ecosystems, and the universe as a whole.

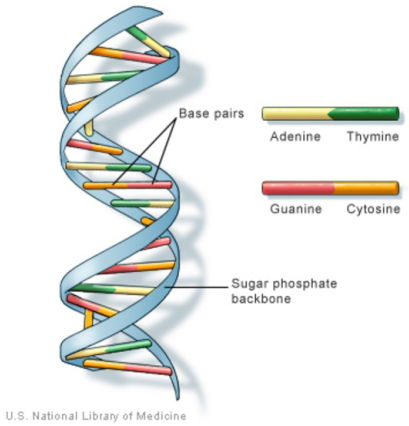
Information is a key aspect of the order of nature, as it provides a way of describing and understanding the relationships and structures within the natural world. In biology, for example, information is encoded in DNA and controls the development, growth, and function of organisms. In physics, information can be used to describe the behavior of subatomic particles, the structure of matter and energy, and the evolution of the universe.

In order to make sense of information and generate knowledge, data science plays an important role in processing data to find hidden patterns and trends. For instance, image processing and machine learning methods are used to look for hidden patterns in DNA, to model the behaviour of neurons, and to predict the outcomes of complex systems. By using data science techniques, researchers and analysts can unlock insights and make data-driven decisions that can improve our understanding of the world around us. With the increasing amount of data being generated every day, the importance of data science will only continue to grow in the coming years.

3 INFORMATION IN BIOLOGY

3.1 DNA as a source of information

DNA (deoxyribonucleic acid) is a vital source of genetic information in living organisms. DNA is a complex molecule composed of nucleotides that are arranged in a specific sequence. Each nucleotide contains a sugar, a phosphate group, and a nitrogenous base (adenine, thymine, guanine, or cytosine).



The specific sequence of nucleotides in DNA contains the instructions for the development, growth, and function of living organisms. This information is encoded in the language of the DNA sequence, which is read and interpreted by various molecular processes within the cell.

Advances in DNA sequencing technology have allowed for the mapping and analysis of entire genomes, providing a wealth of information about the genetic basis of traits and diseases. This has led to new discoveries in fields such as genetics, molecular biology, and biotechnology, and has opened up new avenues for personalized medicine and genetic engineering.

3.2 Data Science in DNA analysis

Data science methods are increasingly used in DNA analysis and research to extract information and patterns from large amounts of genetic data. Among the most commonly used methods are supervised classification learning and unsupervised clustering learning. In both, coding techniques must be used to convert sequences into numbers and make the algorithms more efficient.

3.2.1 DNA Sequence Coding

When processing the DNA sequence, it is necessary to convert the string sequence into a numerical value, so as to form a matrix input model training. There are three methods for sequence encoding [1][2]:

- Secuential encoding: In this approach, each base is a number. For example “ATGC” becomes [0.25, 0.5, 0.75, 1.0].
- One hot encoder: This is widely used in deep learning methods and lends itself well to algorithms like convolutional neural networks. In this example, “ATGC” would become [0,0,0,1], [0,0,1,0], [0,1,0,0], [1,0,0,0].

- K-mer encoding (DNA as a language): First take the long biological sequence and break it down into k-mer length overlapping "words". For example, if we use "words" of length 6 (hexamers), "ATGCATGCA" becomes: 'ATGCAT', 'TGCATG', 'GCATGC', 'CATGCA'. Hence our example sequence is broken down into 4 hexamer words.

3.2.2 Classification algorithm

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data [3]. In genomics, the key issues are genome classification and sequence annotation.

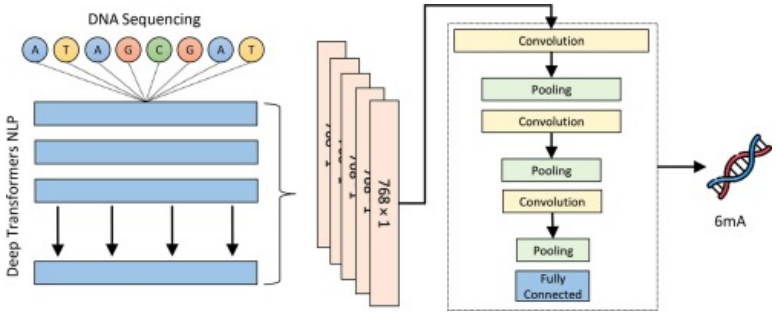
There are several classification machine learning models that are commonly used in genomic sequencing, depending on the specific application and type of data being analyzed. Some common models include Convolutional Neural Networks, Support Vector Machines, Random Forests and Deep Neural Networks.

The following are examples of classification algorithms applied to DNA sequences.

Idetifying epigenetic modifications

Epigenetic modifications play a critical role in regulating gene expression and cellular function, and can influence various biological processes including development, aging, and disease. Identifying epigenetic modifications can provide insights into how these processes are regulated, and may also have important implications for disease diagnosis, prevention, and treatment.

In [4] proposes a novel method to predict DNA 6 mA sites using a combination of deep transformers natural language processing (NLP) and convulutional neural networks (CNNs). The method involves treating DNA sequences as natural sentences, extracting feature vectors using a pre-trained NLP model, and using them as input for a deep learning model to predict 6 mA sites.



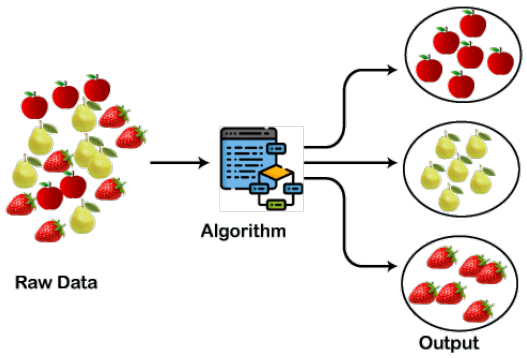
Cancer diagnosis

In [5] describes the use of machine learning algorithms to classify variants detected by next-generation sequencing (NGS) in the diagnosis and treatment of cancer. The authors trained a machine learning model using data from their hospital's NGS database and compared the performance of different algorithms. The best-performing model was a **random forest machine learning model**, which was able to accurately classify pathogenic single nucleotide variants, single nucleotide polymorphisms, multiple nucleotide variants, insertions, and deletions. The authors suggest that artificial intelligence can be valuable in processing large quantities of molecular data in cancer diagnosis and that neural networks show promise in predicting more complex variants in the future.

3.2.3 Clustering

Clustering, is an unsupervised machine learning task that involves automatically discovering natural grouping in data. Unlike supervised learning (like predictive modeling), clustering algorithms only interpret the input data and find natural groups or clusters in feature space [6].

A cluster is often an area of density in the feature space where examples from the domain (observations or rows of data) are closer to the cluster than other clusters. The cluster may have a center (the centroid) that is a sample or a point feature space and may have a boundary or extent.



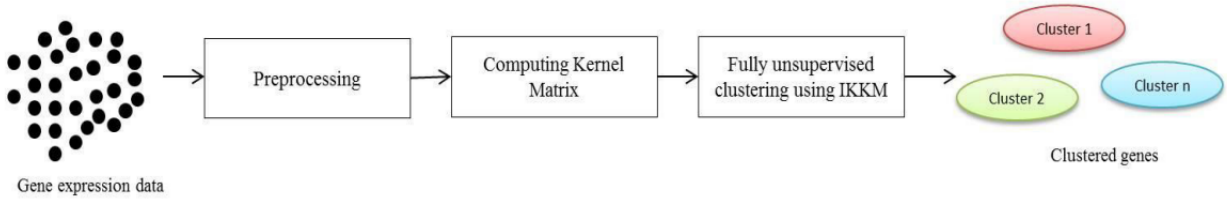
Clustering is a widely used technique in genomics to identify groups of genes or sequences that share similar characteristics or functions. Here are a few examples of how clustering is used in genome analysis:

Grouping the DNA Sequences of Hepatitis B Virus

In [7] discusses the clustering of Hepatitis B virus (HBV) DNA sequences using k-means clustering algorithm and R programming. The data were collected from GenBank and the clustering process was based on the n-mers frequency extraction and normalized using the min-max normalization with interval [0, 1]. The clustering results show that the HBV viruses in the first cluster are more virulent than those in the second cluster, and could potentially evolve with Hepatitis D virus (HDV), which is closely related to HBV infection. The paper also introduces the concept of bioinformatics and the use of clustering methods to process large-scale DNA data. Finally, the paper discusses the importance of preventing hepatitis B as a means of indirectly preventing hepatitis D, given the close relationship between the two diseases.

DNA Clustering for cancer diagnosis

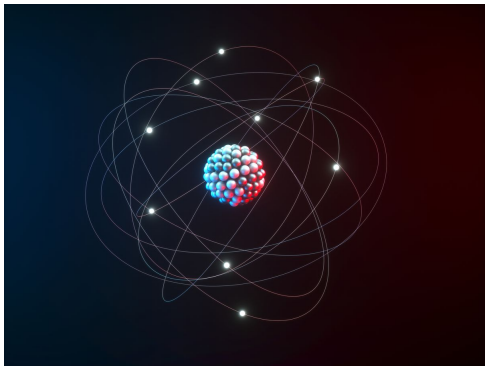
[8] presents a clustering algorithm called Intelligent Kernel K-Means (IKKM) that is fully unsupervised and based on kernels. The aim of the algorithm is to cluster non-linearly separable data such as gene expression. The algorithm is applied to gene expression of human colorectal carcinoma and the resulting clusters are analyzed to determine their trustworthiness and compactness. Correlation ratios between the clustered genes and phenotypes of clinical data are also calculated. The paper concludes that the IKKM algorithm is an effective tool for clustering gene expression data and can be used to identify genes that may contribute to cancer disease.



4 INFORMATION IN PHYSICS

4.1 Information in the behavior of subatomic particles

The behavior of subatomic particles is influenced by a variety of factors, including their physical properties, interactions with other particles, and the environment in which they exist. Information plays a crucial role in describing and predicting the behavior of subatomic particles, particularly at the quantum level.



In quantum mechanics, particles are described by wave functions, which contain information about the particle's position, momentum, and other physical properties. These wave functions are probabilistic in nature, meaning that they describe the probability of finding the particle in a particular state or location.

The study of subatomic particles and their behavior is an active area of research, with many new discoveries and applications being made possible by advances in technology and our understanding of the fundamental principles that govern the behavior of matter at the quantum level. By studying the information contained in wave functions and the interactions between particles, we can gain insights into the nature of matter and the fundamental workings of the universe.

4.1.1 Data Science in the behavior of subatomic particles

Data science is a crucial tool in the study of subatomic particles. The behavior of subatomic particles can be observed and measured through experiments conducted at particle accelerators like the Large Hadron Collider (LHC).

These experiments produce vast amounts of data, which can be difficult to analyze without the aid of data science techniques such as machine learning and statistical analysis. For example, machine learning algorithms can be used to identify specific patterns or features in the data that may indicate the presence of a new particle. Statistical analysis can help to distinguish between different theoretical models based on the experimental data. Sophisticated data visualisation tools are also needed to perform simulations and translate data into graphs.

Machine learning in the search for new particles

The LHC produces millions of collisions between subatomic particles every second, generating huge amounts of data. To make sense of this data, scientists use a variety of detectors to measure the energy and properties of the particles produced in these collisions.

One of the key detectors in the LHC is the ATLAS detector, which measures the energy, momentum, and charge of particles produced in collisions. To improve the accuracy of this detector, scientists have used machine learning algorithms to analyze the data it generates.

Specifically, they have used a technique called "deep learning" to identify and classify the different particles produced in collisions, such as electrons, muons, and photons. Deep learning involves training a neural network on large amounts of data, allowing it to learn patterns and relationships in the data and make predictions. By applying deep learning to the data generated by the ATLAS detector, scientists have been able to improve the accuracy of particle identification, reducing the number of misidentified particles and improving overall data quality. This has helped researchers to better understand the behavior of subatomic particles and probe the fundamental laws of physics.

In [9] explains the development of a new model-independent search for new particles using a novel machine learning technique called "weak supervision". Traditionally, searches for new particles start with a specific theoretical model, and physicists simulate how new particles would be produced and decay in the detector, and then they develop classifiers that separate signals from the background. However, ATLAS' new search uses neural networks trained on data using the CWoLa technique to differentiate between background and potential signals. This method does not require per-event labels, and it exploits structures in the data to reduce dependence on a specific model. The new search combines the bump hunt and weak supervision to enhance the sensitivity to a wide variety of hypothetical particles without specifying their properties ahead of time.

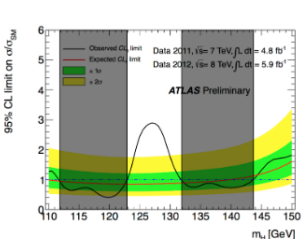
This approach results in an analysis that is mostly free of signal-model and background-model dependence. The new search is the first application of fully data-driven machine-learning-enhanced anomaly detection.

ROOT software for data analysis and visualisation

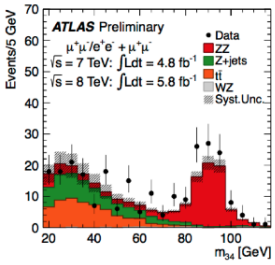
In the field of data analysis and visualisation, ROOT software developed by CERN has been a game-changer. Designed primarily for data analysis in particle physics, this C++ based framework offers a comprehensive suite of tools to analyze, visualise, and store data [10]. However, its applications have expanded beyond particle physics, making it a widely used tool in various scientific fields, including astrophysics and medical imaging.

On 4 July 2012, the Higgs boson was discovered by the ATLAS experiment and the CMS experiment at the Large Hadron Collider (LHC). This particle was first proposed in the 1960s by theoretical physicist Peter Higgs as an elementary particle that could explain why other particles have mass. Its discovery was a major milestone in particle physics and helped confirm the Standard Model theory of particle physics.

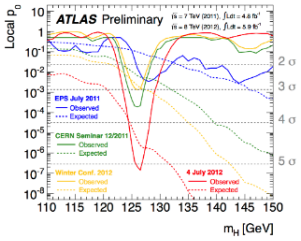
ROOT played an important role in this event by generating data visualisations that allowed a better understanding of the discovery. The following images show some of the graphics produced with ROOT when the discovery of the Higgs boson was announced at CERN [11].



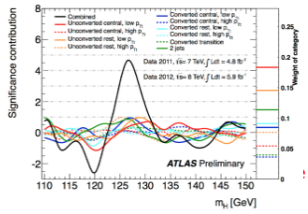
Atlas CLs Exclusion



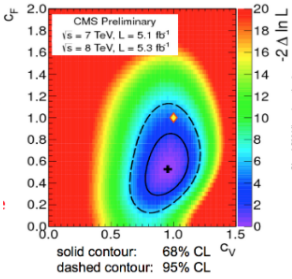
Atlas Data MC Comparison

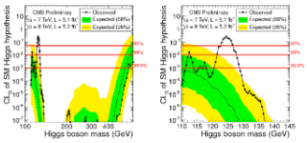


Atlas P0 Trends



Atlas Signifiance Channel by Channel



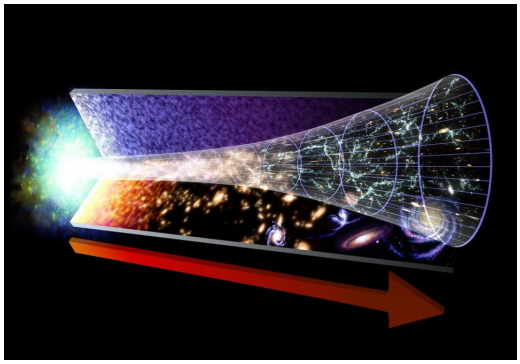


CMS CLs Exclusion Plot

5 INFORMATION IN COSMOLOGY

5.1 The evolution of the Universe

Information is fundamental to understanding the evolution of the universe. The universe began with the Big Bang, a highly energetic event that created matter, energy, and space-time. As the universe expanded and cooled, particles such as protons, neutrons, and electrons began to form, eventually forming atoms and stars.



The evolution of the universe is described by physical laws and mathematical models that incorporate information about the properties and behavior of matter and energy. The study of cosmology involves understanding the large-scale structure and evolution of the universe, including the formation of galaxies, the distribution of dark matter and dark energy, and the eventual fate of the universe.

Overall, the study of the evolution of the universe involves understanding the fundamental principles that govern the behavior of matter and energy, and how these principles have influenced the large-scale structure and evolution of the universe over billions of years. By incorporating information from a wide range of sources, including astronomical observations and theoretical models, scientists are gaining a deeper understanding of the origins and evolution of the universe.

5.1.1 Data Science in the evolution of the universe

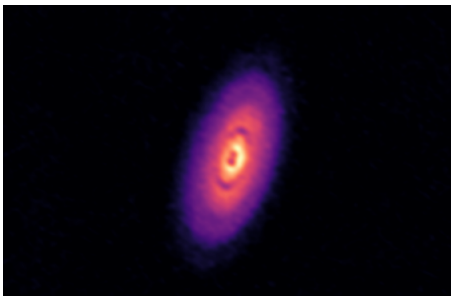
Data science plays a crucial role in the study of the evolution of the universe. By analyzing large datasets, scientists can better understand the structure, composition, and behavior of the cosmos.

Data science techniques, such as machine learning, image processing and statistical analysis, are used to identify patterns and relationships in astronomical data, and to test theoretical models of the universe's evolution.

One of the primary sources of data in the study of the universe's evolution is astronomical observations. Astronomers use telescopes to observe and measure the properties of celestial objects, such as stars, galaxies, and black holes. These observations produce vast amounts of data, which can be difficult to analyze without the aid of data science techniques.

Machine learning in exoplanet discovery

The use of machine learning techniques such as artificial neural networks has allowed astronomers to analyse large datasets more efficiently and detect patterns that can be difficult to detect manually. The discovery of the exoplanet is just one of many examples where machine learning is being used to advance scientific research.

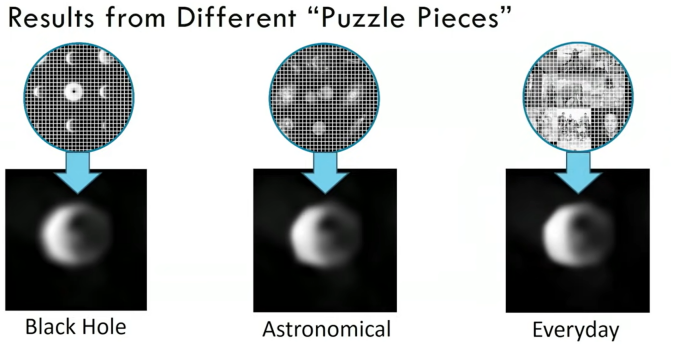


Astronomers collected data on the luminosity of a star (the Sun) at different times and then used machine learning algorithms to look for patterns in the data that could indicate the presence of an exoplanet orbiting the star.

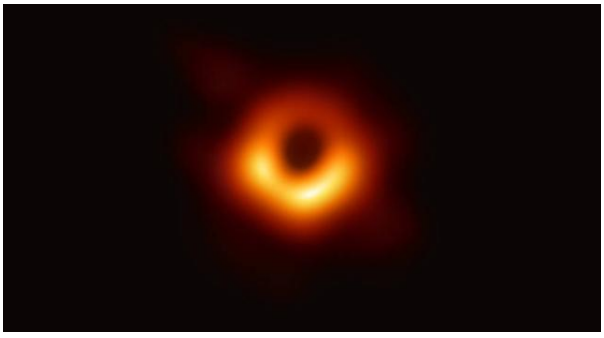
Machine learning algorithms are capable of detecting complex patterns in large data sets and, in this case [12], the astronomers used an algorithm called "Artificial Neural Networks" (ANN) to analyse the data. Once they identified patterns in the data that could be caused by the presence of an exoplanet, the astronomers made additional observations to confirm the discovery.

Machine learning in the first-ever image of a black hole

Machine learning played a crucial role in creating the first-ever image of a black hole. The image was created by combining data from 8 telescopes across the globe to fill in the image parts. The computer vision AI algorithm was trained on galactic and other images to learn how things look in the universe. To avoid bias, the algorithm was trained on three different sets of data: expected sizes of a black hole, other galactic images, and general images of everyday objects [13].



Once the data from telescopes were fed to all three trained algorithms, they all created almost similar images, reaffirming the assumptions and giving us the first-ever image of a black hole.



6 CONCLUSIONS

In conclusion, the study of information and data science in the order of nature is of utmost importance in unlocking the secrets of the natural world. This report has shown how information is fundamental to understanding the regularities and patterns that exist in biology, physics and cosmology. Specifically, we explore three different areas of the natural world, DNA, the behaviour of subatomic particles and the evolution of the universe.

These three areas, despite having great differences, share one thing in common, they are all sources of information. In DNA, the information contains the instructions for the development, growth and functioning of living organisms; in subatomic particles, the information contains the particle's position, momentum and other physical properties that are described by wave functions; and in the evolution of the universe, the information tells a story about the initial conditions of the universe, such as the distribution of matter and energy or the cosmic microwave background radiation, which is a product of the Big Bang.

In this context, Data science plays a crucial role, as it allows this information to be processed and analysed, enabling researchers to generate knowledge and make data-driven decisions. The significance of data science in advancing scientific discovery cannot be overstated. Therefore, it is essential to continue investing in data science research to further our understanding of the order of nature and improve our knowledge of the world.

7 REFERENCES

[1] Yang, A. et al. (2020) Review on the application of machine learning algorithms in the sequence data mining of DNA, Frontiers. Frontiers. Available at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.01032/full (Accessed: April 27, 2023).

[2] Singhakash (2021) DNA sequencing with Machine Learning, Kaggle. Kaggle. Available at: https://www.kaggle.com/code/singhakash/dna-sequencing-with-machine-learning (Accessed: April 27, 2023).

[3] Classification algorithm in Machine Learning - Javatpoint www.javatpoint.com. Available at: https://www.javatpoint.com/classification-algorithm-in-machine-learning (Accessed: April 27, 2023).

[4] Nguyen Quoc Khanh Le a b c et al. (2021) Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes, Methods. Academic Press. Available at: https://www.sciencedirect.com/science/article/pii/S1046202321002747 (Accessed: April 27, 2023).

[5] Pellegrino, E. et al. (2021) Machine learning random forest for predicting oncosomatic variant NGS analysis, Nature News. Nature Publishing Group. Available at: https://www.nature.com/articles/s41598-021-01253-y (Accessed: April 27, 2023).

[6] Ali, M. (2022) Clustering in machine learning: 5 essential clustering algorithms, DataCamp. DataCamp. Available at: https://www.datacamp.com/blog/clustering-in-machine-learning-5-essential-clustering-algorithms (Accessed: April 27, 2023).

[7] Aip.scitation.org. Available at: https://aip.scitation.org/doi/abs/10.1063/1.4991238 (Accessed: April 27, 2023).

[8] Teny Handhayani a et al. (2015) Intelligent kernel K-means for clustering gene expression, Procedia Computer Science. Elsevier. Available at: https://www.sciencedirect.com/science/article/pii/S1877050915020736 (Accessed: April 27, 2023).

[9] Machine learning qualitatively changes the search for new particles (no date) ATLAS. Available at: https://atlas.cern/updates/briefing/search-new-particles-machine-learning (Accessed: April 27, 2023).

[10] team, R.O.O.T. (no date) Analyzing petabytes of data, scientifically., ROOT. Available at: https://root.cern/ (Accessed: April 27, 2023).

[11] team, R.O.O.T. (no date) Galleries of images produced with Root, ROOT. Available at: https://root.cern/gallery/ (Accessed: April 27, 2023).

[12] Flurry, A. (2023) UGA researchers discover new planet outside Solar System, UGA Today. Available at: https://news.uga.edu/uga-researchers-discover-new-planet-outside-solar-system/ (Accessed: April 27, 2023).

[13] Singh, D. (2021) The AI behind getting the first-ever picture of a 'black hole', Medium. Medium. Available at: https://deepaksingh-rv.medium.com/the-ai-behind-getting-the-first-ever-picture-of-a-black-hole-c483e8eb6a21 (Accessed: April 27, 2023).