As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 8(05):123-132, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0805020
Available online at: http://www.ijmtst.com/vol8issue05.html

Data Science: Concept and Technology

Dr. Swagata Sarkar[1] | Hemanth M[2] | Shrinithi G R[3] |Anjal J Sankar[4]

[1]Head of Department, Artificial Intelligence and Data Science, Sri Sairam Engineering College, Tamil Nadu
[2]Department of Artificial Intelligence and Data Science, Sri Sairam Engineering College, Tamil Nadu
*Corresponding Author Email ID: sec20ad021@sairamtap.edu.in

To Cite this Article
Dr. Swagata Sarkar, Hemanth M, Shrinithi G R and Anjal J Sankar. Data Science: Concept and Technology. International Journal for Modern Trends in Science and Technology 2022, 8(05), pp. 123-132. https://doi.org/10.46501/IJMTST0805020

Article Info
Received: 30 March 2022; Accepted: 02 May 2022; Published: 05 May 2022.

ABSTRACT

A data scientist draws insights and knowledge from noisy, structured, and unstructured data using scientific processes and algorithms, and then applies these insights and knowledge across a wide range of applications. One of the most promising and in-demand career paths for skilled professionals is data science. Today, successful data professionals appreciate that they must move beyond traditional tasks such as analyzing large volumes of data, data mining, and programming.

Data scientists must be adept at mastering all stages of the data science life cycle in order to uncover useful intelligence that will benefit their organizations. Also, they must possess flexibility and understanding to maximize returns at each stage of the process. Data science combines mathematics, statistics, informatics, and their related methods to study and analyze actual phenomena using data. It uses techniques, theories, and concepts from mathematics, statistics, information systems, computer science, information science, and domain knowledge, and other related fields in order to do this.

KEYWORDS:Data Science (DS), Big Data, Pandas, Seaborn, NumPy, Data mining, Data Visualization..

1. INTRODUCTION

The term "Data Science" has been in vogue in the technical, academic, and business realms, as indicated by the increase in vacancies. Critical academics and journalists, however, do not see any distinction between data science and statistics implementation. In addition to capturing unstructured and structured data, Data Science is concerned with cleansing, preparing, and analyzing it. Data is everywhere and is increasing at an exponential rate. The problem of recording, processing, and storing information has become relevant in almost every field. Using satellite data for weather forecasts, traffic jams, or natural disasters is becoming increasingly important. Research is being conducted in new ways, scientists think differently, and research data is being used and shared in new ways as Data Science technologies (also known as Data Intensive Science or Big Data technologies) emerge. As well, it has opened new opportunities for Data Scientists to take on roles in industries and business organizations thanks to the availability of an extensive amount of digital data about business processes. Business and sales are influenced by data science and big data. Big data can also be used for predicting stock fluctuations of a company in government. It is not possible to conduct statistical research on data science usage and the state of big data in science or analyze the data scientists' usage of these tools.

2. METHODOLOGY

Data science integration is proposed to be visualized with networks theory, developed by Euler (1736) for the Konigsberg Bridge problem and presented as mathematical notation of nodes and edges in data science. Qualitative case studies of selected organizations that defined roles for teams that worked on data science projects were conducted to understand how roles are defined today in the field of data science. Two standards bodies, two organizations from industry, and one consulting/advisory firm were explored in order to obtain a point of view of current thinking and usage across the field of data science. Based on the written documentation collected from each organization, we analyzed the defined roles for each case study. Individuals from the identified organizations were also consulted when documentation was not as comprehensive.

TRAINING GOALS

Developing all aspects of one's personality - mind, body, and soul - is essential to cultivate talents. This includes good psychology, good physical quality, and good health. In order to understand the basic skills of students in data science and engineering major, let's start with understanding the Data Science Lifecycle. Data Science Lifecycle includes seven iterative steps, that is business understanding, data mining, data cleaning, data exploration, feature engineering, predictive modeling, and data visualization.

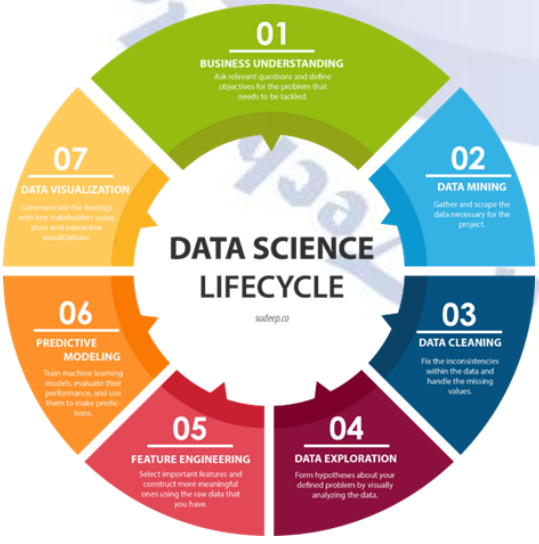

Fig. 1. Data Science Lifecycle, which includes seven iterative steps.

To complete data processing, in summary, graduates of data science and engineering major should have the following basic skills:

1. Have mature data thinking ability
2. Have skills in data processing

i) Information mining ability
ii) Information processing capability
iii) Computer coding ability

3. Have ability to work in teams
4. Some interdisciplinary knowledge is needed

ALGORITHM

Conceptually speaking, we refer to a meaningful verbal unit, or a combination of verbal units, which is described as the framework for scientific interpretation of the meaning of a specific concept relevant to one or more subject areas. A network of subject areas is just a way to define generalized descriptions of subject areas by giving them names and subordinate units, and depicting the domain as a whole, Describes the subject area that is designated by the information system or by the proposed systematization of the given topic area which defines the given network.

3. DATA SCIENCE FRAMEWORK

Data Science Framework has several libraries which provide data mining functionality which includes methods for exploring the data, cleaning it up, and transforming the data into some more useful for-mat that can be used for data processing, as well as machine learning. The complete process sums up in building visualizations to gain knowledge from your data.

Data is divided into two distinct parts that can take an array of different forms. One is structured da-ta, which has discrete and categorical variables (finite sets of possible values, like gender) or continuous and numerical variables, which have integers or re-al numbers. One of the special case of categorical variables are Boolean or binary variables which are yes/no or true/false. We can convert numeric to discrete by binning. Some algorithms work with categorical variables but cannot handle numeric variables, or we would like to purposely unspecify data. This Categorical variable can be converted numeric values by encoding.

ANALYZING DATA

Once the data has been collected and cleaned, we are now ready to start the analysis or conduct data mining.

Data mining (sometimes called knowledge discovery in data) is extracting implicit and potentially useful information from data. This is done by finding patterns and structure in data and summarizing it into useful information which can be used for decision making. It is the union of statistics, artificial intelligence & machine learning, and databases.

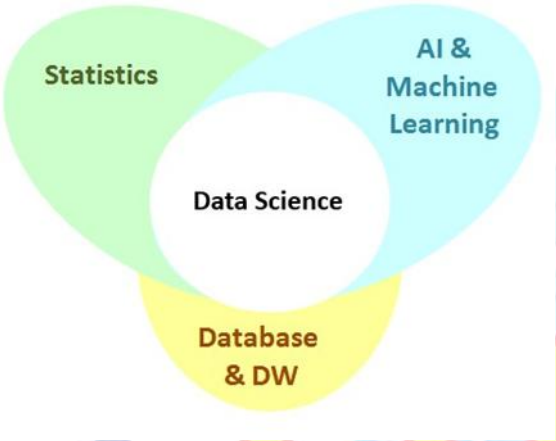

Fig. 2. Data Science Framework

Association Having an association is useful because one can use it for product placement, to lower the price of another product and determine which the products would be affected if one production was not available. When buying together or in a document, you can determine which items are related (and how strong the relation is). An example is a database with supermarket transactions, a common association could be that bread is associated with butter.

Clustering The clustering method uses an algorithm that organizes the data into clusters and determines center clusters and clusters. The method is called unsupervised because the data is not labelled and the natural groups are unknown.

Regression is fitting a line (y = mx+b) or a curve to a set of data to conduct numeric prediction. In regression, a line is fitted to a set of data to perform numeric analysis. Regression can also be used to do predictive causality analysis, which means assessing the relationship between two features. For example, for class grades, a line can be fitted to describe the relationship between the grades and against the number of hours of study and grade.

An outlier Outlier is a data point or a set of data points with observations that deviate so much from normal observations that there is a possibility they may have been generated by another mechanism. Outlier detections are a typical method used to determine whether such observations are outliers. This shows that it can be used to identify a problem with the data and provide an interesting observation.

Classification In classification, given a set of inputs or features and a discrete class variable or attribute, classification determines which class or label it belongs to for new data. This is determined by a set of rules based on past data. To generate rules, there are many techniques to use and even a combination of techniques that each have their advantages and disadvantages. The rules are generated by using a subset of the labelled data that the algorithm has not used to generate rules.

DATA SCIENCE FRAMEWORKS FOR PYTHON

Here are the top data science frameworks for Python. The list is based on insights and experience from practicing data scientists.

1. TensorFlow and Keras

TensorFlow plays a powerful role in machine learning framework that is based on Python. It can do everything from creating a simple calculation to building complicated neural networks. TensorFlow is backed by Google and has been around since 2007, though it only became open source in 2015. In 2017, TensorFlow released an add-on package called Keras that provides high-level APIs and building blocks (like MATLAB) for creating Deep Learning models

2. NumPy

NumPy library provides a package that is built on top of the Python language providing efficient numerical operations. It's suitable for manipulating matrices and performing many other numerical calculations. It can be used on its own or with other frameworks like TensorFlow or Theano.

3. Panda

Python package Pandas provides high-level data structures and analysis tools. It also used to load CSV or excel files to manipulate the data, visualize it using graphs or charts, etc. The main concept of pandas is that everything you do with your data happens in a Series (1D array) or a Data Frame (2D Array). Pandas make working with Data Frames extremely easy.

4. Matplotlib

Matplotlib is a python library used to visualize data. Some of the most used Python libraries to visualize scientific and numerical data are provided by this package so that you can create graphs like R or MATLAB. You can also choose from different back-ends like Qt, WX, etc., for your visualizations.

5. Scikit-learn

Scikit-learn defines a collection of Python modules for machine learning built on top of SciPy. Pipelines are also supported, which are steps composed of transformers and estimators connected to form a model.

6. SpaCy

SpaCy is an excellent Natural Language Processing (NLP) library in Python. It provides tools and models to process text to compute meaning of words, sentences, or entire texts. In addition, you can easily tokenize and parse natural language with SpaCy.

7. Natural Language Toolkit

NLTK is another collection of Python modules for processing natural languages. Its features include part-of-speech tagging, parsing trees, named entity recognition, classification, etc. In addition, it can be used to process text to compute the meaning of words, sentences, or entire texts.

8. Theano

Theano is a Python library that is efficiently used to define, optimize, and evaluate mathematical expressions involving multi-dimensional arrays. With a result, it can be used for machine learning applications that work with computationally intensive calculations.

9. Pytorch

The Pytorch, a machine learning framework is based on Torch, it provides high-level APIs and helps in building blocks for creating Deep Learning models. It offers maximum flexibility and the ability to use Python code to define, load, transform, and manipulate data.

10. Caffe/Caffe2

Caffe is a machine learning/deep learning framework modelled with speed and modularity in mind. caffe2 is a lightweight, modular, and scalable library built to provide easy-to-use, extensible building blocks for fast prototyping of machine intelligence algorithms such as neural networks. In addition, Caffe/Caffe2 can be used for machine vision, speech and audio processing, and reinforcement learning.

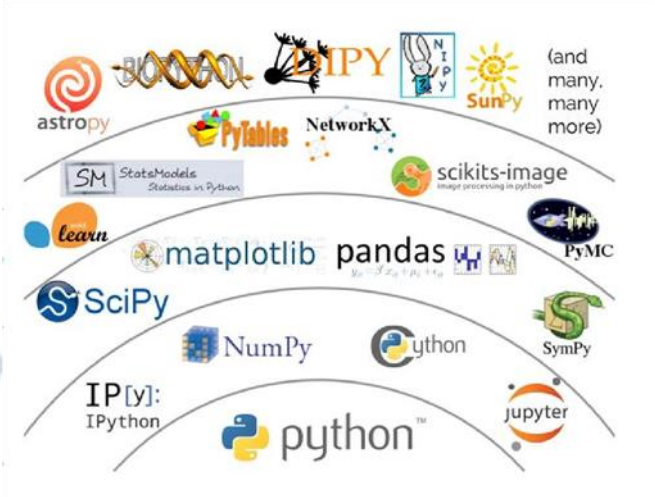

Fig. 3. Data Science Framework for Python

HOW TO CHOOSE THE RIGHT FRAMEWORK?

There are some frameworks used in machine learning. However, it is difficult to choose the proper framework without understanding its capabilities, limitations, and use cases. There are some factors considered while choosing a framework are:

Ease of Use: A good library helps to make it easy to get started with your dataset, whether images, text, or anything else. We should be able to load and save data in memory efficiently.

Hardware Deployment: Some frameworks target hardware deployment, and they might provide a way to speed up your models by using GPUs, TPUs, etc.

Multi-Language Support: Many libraries support Python, R, Scala, C++, etc., where-as some modules are restricted to one or two languages. If you want to research building prototypes for your startup, consider the multi-language support.

Flexibility: Some frameworks are more rigid, forcing you to use their pre-defined architectures for building models. However, the ability to define your model is vital if you want to expand beyond the present capabilities of the framework.

Ecosystem Support: A sound library should have documentation, tutorials, examples, Stack Overflow questions, etc., available online. This lets you get started quickly and efficiently.

4. DATA MODELLING AND DATA VISUALISATION

Data Modeling involves the analysis of data objects and their relationships to other objects. It is used to simplify the analysis of data requirements for a business process. The data models are then used to generate the database to store the data. Rather than focusing on what operations we need to perform, a Data Model includes what data is needed and how we have to organize it. Its purpose is to display how data flows in a form that is easy to understand, similar to an architect's blueprint. A diagram showing the data flow across a complex software system will have text and symbols.

Data models can be divided into conceptual models, logical models, and physical models, each with a specific purpose. The data models are used to represent the data and how it is stored in the database, and they are also used to establish the relationships between data items.

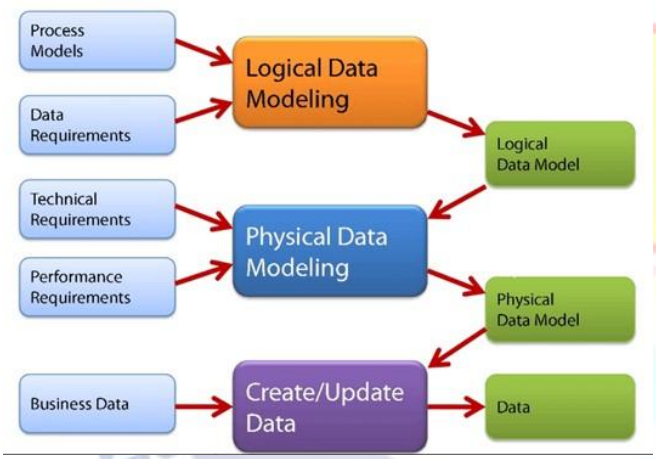

Fig. 4. Data Modelling Types

Data Visualization is the act of visualizing information and data using visual elements such as charts, graphs, and maps. Such visual representations aid in the understanding of patterns, trends, and outliers within data. A data visualization is an interdisciplinary field that deals with the visual representation of data. This form of communication is particularly efficient when large quantities of data are being communicated, for example a time series. It is often considered a branch of descriptive statistics since its roots are in the field of statistics. Yet, due to the fact that both design skills and statistical and computing skills are needed to visualize effectively, data visualization is actually a branch of descriptive statistics.

Here are eight important data visualization tools to help data scientists make better decisions. Data visualization tools can effectively increase the efficiency of data scientists.

1) Tableau

It allows users to connect to different data sources and create visualizations from different data sources with ease. Tableau is used to visualize data and create interactive graphs, charts, and maps.

2) QlikView

By accelerating analytics, revealing new business insights, and increasing the accuracy of results, QlikView is more than just another data visualization tool. It is a platform for data discovery that helps users make better decisions faster. In many organizations around the world, it has been used for many years as an intuitive software development kit, which can acquire various kinds of data sources and visualize them using color-coded tables, bar graphs, line graphs, pie charts, and sliders. The data visualization interface was developed so that users can easily drag and drop data from various sources, such as databases or spreadsheets, into the application without writing any code. These characteristics also make the tool relatively easier to learn and grasp.

3) Microsoft Power BI

The Microsoft Power BI is designed for data visualizations and report generation, but it can also be used as a self-service analytics and predictive analytics tool. As a centralized repository for all your business data, it is accessible by all users of your business. It permits users to create reports and share insights with others within their organization.

4) Datawrapper

Datawrapper is a powerful online data visualization tool that can be used for many different situations. As part of its easyto-use user interface, it has a clear and intuitive interface. Through Datawrapper, users can create charts and maps directly in the browser by uploading data files. The charts and maps created in Datawrapper, no matter what kind of device they're using, will look great in the browser no matter how they're viewed.

5) Plotly

Plotly is an online tool for creating data visualizations such as graphs, charts, maps, and more. People can use

Plotly to visualize a dataset and share the visualization on social media or through blogs.

6) Sisense

Using Sisense, you can create interactive dashboards that illustrate your data in an understandable way. Sisense lets you build extensive, informative dashboards that help you learn more about your data.

7) Excel

Data visualization can be done easily with Microsoft Excel, since it has an easy interface, so you don't have to be an expert to use it. Scatter plots are a visual representation of the relationship between two datasets. They are often used to compare data in Excel.

8) Zoho analytics

Zoho Analytics is a powerful tool that helps you visualize and manage your data with ease while creating custom reports and dashboards. You can make custom reports and dashboards in just a few clicks and get access to all the insights into your data with interactive charts and graphs.

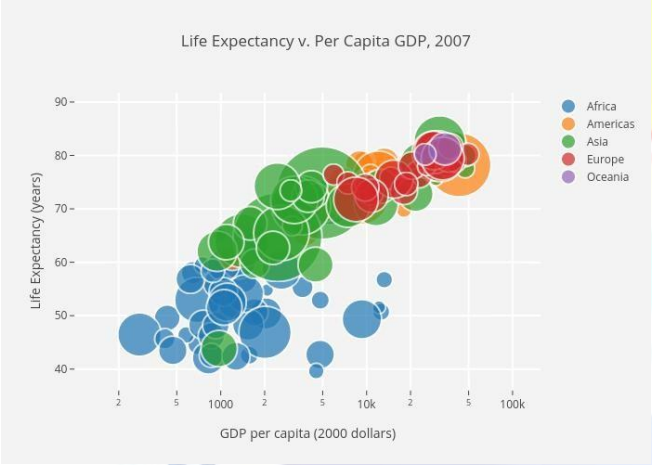

Fig. 5. Data Visualization foe Life expectancy vs. Per CapitaDGP

5. DATA SCIENCE APPLICATIONS AND CHALLENGES

Data Science Applications

Data scientists can perform predictive modeling, pattern recognition, anomaly detection, classification, categorization, sentiment analysis, and develop technologies such as recommendation engines, personalization systems, and artificial intelligence tools like chatbots and autonomous vehicles and machines.

These applications drive a wide variety of business processes in organizations, including:

i. Customer Analytics
ii. Fraud Detection
iii. Risk Management
iv. Stock Trading
v. Targeted Advertising
vi. Customer Service
vii. Website Personalization
viii. Logistics & Supply Chain Management
ix. Predictive Maintenance x. Image Recognition
xi. Natural Language Processing
xii. Speech Recognition
xiii. Cybersecurity
xiv. Medical Diagnosis

Challenges in Data Science

Its advanced analytics make data science inherently challenging due to the large amounts of data typically involved. The sheer amount of data typically being analyzed adds to the challenge and takes a lot of time to complete a project. The process of analyzing big data is further complicated when data scientists deal with pools of structured, unstructured, and semi structured data.

Data sets and analytics applications present a number of major challenges, including issues with the underlying data and ones that data scientists unconsciously incorporate into algorithms and predictive models. In the absence of an awareness and correction of such biases, inaccurate findings can skew analytics results, leading to faulty business decisions. At the worst, they can hurt certain groups -- for example, racial bias in AI systems. Gartner analyst AfrazJaffri and four of his co-workers cited finding the right dataset to analyze in a report published in January 2020. The report also mentioned choosing the right tools, managing deployments of analytical models, as well as quantifying business value and maintaining models as significant hurdles.

6. BENEFITS OF DATA SCIENCE

An important benefit of data science is to empower and facilitate better decision-making. Through data science, organizations can factor quantitative, evidence-based evidence into their business decisions, therefore leading to stronger business performance, cost savings and

smoother business processes and workflows. Data science has specific business benefits for different companies and industries. For customer-facing organizations, for example, it allows them to identify and refine target audiences. Marketers and salespeople can use customer data to improve conversion rates and create customized marketing campaigns and promotional offers that boost sales. For example, reducing fraud, improving risk management, improving financial outcomes, extending the lifespan of manufacturing equipment, improving supply chain performance, and ensuring cybersecurity protection are all benefits. As a result of the benefits. Real-time analytics provides, data science also allows the analysis of data in real-time, facilitating faster decision-making and increasing business agility.

• **Multiple Job Options:** Since data science is in such high demand, it has given rise to many career opportunities in various fields. Some of them include Data Scientists, Data Analysts, Research Analysts, Business Analysts, Analytics Managers, and Big Data Engineers, for instance.

• **Business benefits:** Therefore, the products are always delivered to the right place and right time since data science helps organizations know when and how their products sell best. In turn, firms make quicker and smarter decisions which improve efficiency and increase profits.

• **Highly Paid jobs & career opportunities:** The Data Scientist job continues to be the sexiest, and the salaries are also huge. According to a Dice Salary Survey, a Data Scientist earns an average of $106,000 a year.

• **Hiring benefits:** With Big Data and data mining, the recruitment teams can quickly sort data and select the best candidates for their organization.CVs, aptitude tests and games are easier to process and select with Big Data and data mining.

7. ROLE OF DATA SCIENTIST

The role of a data scientist combines computer science, statistics, and mathematics in gathering and analyzing data. Data scientists gather and analyze large sets of structured and unstructured data. Data scientists use technology and social science skills to find trends and manage data to create actionable plans for companies and other organizations. They analyze, model, and process data to create actionable plans. Data scientists use industry knowledge, contextual understanding, and a skepticism of existing assumptions - to uncover solutions to business problems. Their work occurs when they make sense of messy, unstructured data, such as from smart devices, social media feeds, and emails that don't neatly fit into a database etc.

However, technical skills are not the only thing important. Data scientists are often working in business environments and are involved in communicating complex ideas and making data-driven organizational decisions.

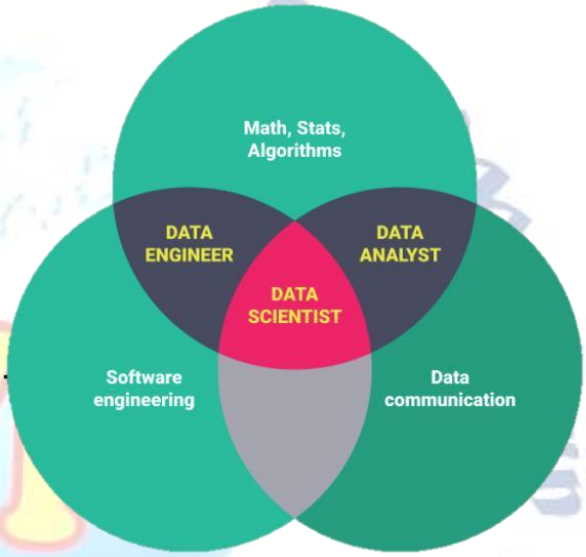

Fig. 6. Roles and Responsibilities of a Data Scientist

8. FUTURE SCOPE OF DATA SCIENCE

The field of data science consists of many revolutionary technologies, such as Artificial Intelligence, Internet of Things, and Deep Learning to name a few. As scientific and technological breakthroughs have increased, data science's influence has grown tremendously. Collecting and analyzing data is crucial because retailers can use it to find out our purchasing habits and, thus, influence them. They can also exert control by exercising the power of their buying power.

Several factors point to the future of data science, thusdemonstrating compelling reasons for its importance

• **Companies' inability to manage data:** Several companies face the same challenge of analyzing and categorizing the data they collect and store. Businesses and companies collect data through transactions and through website interactions. Data scientists become the saviors in situations like this. They can help companies

progress with better and more efficient handling of data, which ultimately enhances productivity.

• **Regulations for protecting personal data revised:** The European Union passed the General Data Protection Regulation (GDPR) in May of 2018. California will pass a similar data protection law in 2020. Data scientists and companies will be increasingly dependent on each other for the purpose of storing data appropriately and responsibly. Generally, people today are more cautious and alert about sharing their personal data with businesses and giving up a portion of control to them, as there is rising awareness about data breaches and their malefic consequences. In the coming future, companies cannot allow their data to be carelessly and irresponsibly captured. The GDPR will ensure some level of data privacy.

• **Data Science is constantly evolving:** There is a risk of stagnating career areas that have no growth potential, indicating that the respective fields need to constantly evolve and undergo changes in order for new opportunities to arise and flourish in the field. It is expected that data science job roles will become more specific in the future, which will lead to specializations in the field. Data science is a broad field that is undergoing development and is thus promising plenty of opportunities in the future. Through these specifications and specializations, people who are inclined to this stream can make use of their opportunities.

• **An astonishing incline in data growth:** The amount of data generated every day is increasing without our knowledge. As time passes, we will have more and more interactions with data. Additionally, worldwide data production will increase at an explosive rate. This will create a great demand for data scientists to help enterprises make the most of this wealth of information.

• **Virtual Reality will be friendlier:** Nowadays, we can see and are in fact seeing how artificial intelligence (AI) is spreading across the globe, and how companies are relying on it. Machine learning is being introduced and implemented in almost every application right now, which means that big data prospects will thrive more with advanced concepts like Deep Learning. As a result, Virtual Reality (VR) and Augmented Reality (AR) are undergoing significant modifications as well. Additionally, the relationship between machines and humans is likely to improve and expand drastically over time.

• **Blockchain updating with Data science:** Blockchain is the core technology pertains to cryptocurrencies like Bitcoin. Consequently, data security will serve its function here as well as it will be able to record and secure the detailed transactions. Big data will be a major factor driving the IoT growth and popularity since it will be responsible for addressing data issues. Edge computing will be responsible for handling big data issues.

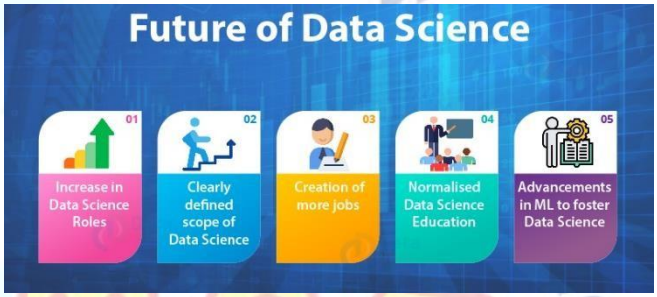

Fig. 7. Future Scope of Data Science

9. CONCLUSION

Education in the data sciences is at the beginning of its developmental stage; it is transforming into a self-sustaining field and producing professionals with distinct and complementary skills when compared to the computer, information, and statistical sciences. The term "data science" is still evolving, and it's best understood from the perspective of a data scientist, who uses programming to conduct deeper, more flexible data analysis. Data science education is currently in its formative stage in the country, where educational institutions are pioneering their own programs based on a variety of depth approaches. As a growing field, data science has become a key component of almost every industry. It provides the best solutions that help to meet the increasing needs and ensure the sustainability of the future. Data scientists are the future of the world, so they must be capable of providing great solutions that address any challenge in any field. As data science's importance is growing day by day, the demand for a data scientist increases too. Data scientists are the future of the world. They need to set up appropriate systems and resources that will enable them to accomplish this. A few of these examples illustrate the impressive capabilities of Data

science, which are far beyond our wildest imaginations, on which Data scientists and experts are engaged and on which further research is ongoing that will definitely contribute to the rapid development of the world.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] M.M. Najafabadi, F. Villanustre, T.M. Khoshgoftaar et al., "Deep learning applications and challenges in big data analytics," Journal of Big Data, Springer, vol. 2 , iss. 1, pp. 1-21, 2015. https://doi.org/10.1186/s40537-014-0007-7

[2] P. Thakuria, N. Y. Tilahun, and M. Zellner, "Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery," Seeing Cities Through Big Data, Springer, NY, pp.11-45, 2017.

[3] Q. Li, Y. Chen, J. Wang, Y. Chen, and H. Chen, "Web Media and Stock Markets : A Survey and Future Directions from a Big Data Perspective," IEEE Transactions on Knowledge and Data Engineering, vol. 30, pp.381-399, Feb. 1 2018.

[4] Eitan D. Hersh, and Brian F. Schaffner, "Targeted Campaign Appeals and the Value of Ambiguity," The Journal of Politics, The University

[5] of Chicago Press, vol. 75, iss. 2, pp.520-534, April 2013.

[6] [5] SjaakWolfert, Lan Ge, CorVerdouw, and Marc-JeroenBogaardt, "Big Data in Smart Farming – A review," Agricultural Systems, vol.153, pp. 69-80, 2017.

[7] M. MazharRathore, Awais Ahmad, Anand Paul, and Seungmin Rho, "Urban planning and building smart cities based on the Internet of Things using Big Data analytics," Computer Networks, vol. 101, pp. 63-80, 2016.

[8] U. Sivarajah, M. Mustafa Kamal, Z. Irani, and V. Weerakkody,"Critical analysis of Big Data challenges and analytical methods," Journal of Business Research, vol. 70, pp. 263-2862, 2017. [8] V. Kale, Big data computing: A Guide for Business and Technology Managers. Taylor and Francis Group, CRC Press, 2017.

[9] [9] D. V. Lande, I. V. Balagura, and V. B. Andrushchenko, "The detection of actual research topics using co-word networks," Open Semantic Technologies for Intelligent Systems : proceedings, Minsk, BNUIR, pp. 207-210, 2018.

[10] J. C. Hayes, and D. J. M. Kraemer, "Grounded understanding of abstract concepts: The case of STEM learning." Cognitive Research, vol. 2, iss.1, 2017. doi:10.1186/s41235-016-0046-z.

[11] D. V. Lande, V. B. Andrushchenko, and I. V. "Balagura Formation of the Subject Area on the Base of Wikipedia Service," Open Semantic Technologies for Intelligent Systems : proceedings, Minsk, BNUIR, pp. 211-214, 2017.

[12] European eCompetences Framework http://www.ecompetences.eu/CEN ICT skills Workshophttp://www.ecompetences.eu/cen-ictskills workshop/

[13] Research Data Alliance (RDA) IG Education and Training on handling of research data (IG-ETRD) [online] https://www.rdalliance.org/ig education-and-training-handling-research-data.html

[14] EIT ICT Labs Master Program on Data Science(DSC)[online]http://www.masterschool.eitictlabs.eu/programmes/data-science/Bloom's taxonomy: the 21st century version. [online] http://www.educatorstechnology.com/2011/09/blooms-taxonomy21st century-version.html

[15] John Biggs, Constructive alignment in university teaching, HERDSA Review of Higher Education,Vol.1[online]http://www.herdsa.org.au/wp content/uploads/HERDSARHE2014v01p05. pdf

[16] Descriptors defining levels in the European Qualifications Framework (EQF) [online] http://ec.europa.eu/ploteus/en/content/descriptors-page.

[17] Demchenko, Y., E.Gruengard, S.Klous, Instructional Model for Building effective Big Data Curricula for Online and Campus Education. In Proc. 6th IEEE International Conference and Workshops on Cloud Computing Technology and Science (CloudCom2014), 15-18 December 2014, Singapore

[18] Wiktorski-Wlodarczyk, T. Hacker, T.J. "Problem-based Learning Approach to a Course in Data Intensive Systems", In Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on (pp. 942-948). IEEE.

[19] Demchenko, Yuri, David Bernstein, A.S.Z. Belloum, Ana Oprescu, Tomasz W. Wlodarczyk, and Cees de Laat, New Instructional Models for Building Effective Curricula on Cloud, Compting Technologies and Engineering, IEEE International Conference on Cloud Computing Technology and Science, Bristol, Dec 2013.

[20] Demchenko, Y., Bernstein, D., Belloum, A., Oprescu, A., WiktorskiWlodarczyk, T. &Laat, C. D. "New Instructional Models for Building Effective Curricula on Cloud Computing Technologies and Engineering". In Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on (Vol. 2, pp. 112-119). IEEE.

[21] DRAFT NIST Special Publication 1500-1: NIST Big Data Interoperability Framework, Volume 1-7. [Online]. Available: http://bigdatawg.nist.gov/V1_output_docs.ph

[22] Saltz, J. (2015). The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectiveness, Big Data Conference

[23] NIST SP 1500-1, "Big Data Interoperability Framework: Volume 1, Definitions" version 2 (201x), available at https://bigdatawg.nist.gov/.

[24] Saltz, J. S., &Shamshurin, I. (2016). Big data team process methodologies: A literature review and the identification of key factors for a project's success. In Big Data (Big Data), 2016 IEEE International Conference on (pp. 2872-2879). IEEE.

[25] NIST Draft SP 800-181, "NICE Cybersecurity Workforce Framework (NCWF), National Initiative for Cybersecurity Education (NICE)", (2016).

[26] National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework.(2017). available at https://niccs.us cert.gov/workforce-development/cyber-securityworkforce framework

[27] NIST SP 800-16, "A Role-Based Model for Federal Information Technology/ Cyber Security Training", revision 1, second draft

version 2, retrieved September 19, 2017, from https://www.nist.gov/sites/default/files/documents/2017/09/06/draft_sp800_16_rev1_2nd-draft.pdf

[28] Saltz, J., Shamshurin, I., and Connors, C. (2017). Predicting data science sociotechnical execution challenges by categorizing data science projects. Journal of the Association for Information Science and Technology.

[29] Grady, N. W. (2016). KDD meets Big Data. In Big Data (Big Data), IEEE International Conference on. IEEE.

[30] Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining," Journal of Data Warehousing, 5(4)

[31] Piatetsky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Available:

[32] http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology analyticsdata-mining-data-science-projects.html [32] NIST 1500 series; available at https://bigdatawg.nist.gov/V2_output_docs.php