Chapter - 3

Data Science: A New Door

Anubhav Kumar, Teklay Gebregzabiher,
Solomon Gebremeskel and Haile Misgna

Faculty School of Computing, EIT-M, Mekelle University, Ethiopia
E-mail: *dr.anubhavkumar@gmail.com*

ABSTRACT

Data Science is new term which became popular in 2019 and boom in 2020 worldwide. Many university started course for Data science at different levels certificate / diploma courses to degree level (UG or PG). India is 2nd most populated country on world which means that lots of students who looks for to take admission in university & colleges. In 2019, many university started courses for data science in India but hot & biggest news came from IIT madras which started Bachelor in Data science course in 2020. But people confuse regards the term data science. They are not able to take decision. This chapter will help to understand the data science like what data science is, what type skill is required to take course and what will be the job opportunity and which university took initiative to start course in India.

Keywords: Data Science, Machine Learning, IIT madras

INTRODUCTION

Data Science has become the most demanding job of the 21st century. Every organization is looking for candidates with knowledge of data science.

"Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, deep learning and big data." (Wikipedia, no date).

It is a multidisciplinary field that uses tools and techniques to manipulate the data so that you can find something new and meaningful.

Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems. It is the future of artificial intelligence.

In Summary, data science includes:

- Asking the correct questions and analyzing the raw data.
- Data-Modeling using various complex and efficient algorithms.
- Data-Visualizing.
- Data-Understanding to make better decisions and finding the final result.

For Example

Let suppose we want to travel from station A to station B by car. Now, we need to take some decisions such as which route will be the best route to reach faster at the location, in which route there will be no traffic jam, and which will be cost-effective. All these decision factors will act as input data, and we will get an appropriate answer from these decisions, so this analysis of data is called the data analysis, which is a part of data science.

Need for Data Science

Some years ago, data was less and mostly available in a structured form, which could be easily stored in excel sheets, and processed using BI tools.

But in today's world, data is becoming so vast, i.e., approximately 2.5 quintals bytes of data is generating on every day, which led to data explosion. It is estimated as per researches, that by 2020, 1.7 MB of data will be created at every single second, by a single

person on earth. Every Company requires data to work, grow, and improve their businesses (LearningPoint92, no date).

Now, handling of such huge amount of data is a challenging task for every organization. So to handle, process, and analysis of this, we required some complex, powerful, and efficient algorithms and technology, and that technology came into existence as data Science. Following are some main reasons for using data science technology:

- With the help of data science technology, we can convert the massive amount of raw and unstructured data into meaningful insights.
- Data science technology is opting by various companies, whether it is a big brand or a startup. Google, Amazon, Netflix, etc, which handle the huge amount of data, are using data science algorithms for better customer experience.
- Data science is working for automating transportation such as creating a self-driving car, which is the future of transportation.
- Data science can help in different predictions such as various survey, elections, flight ticket confirmation, etc.

Career & Jobs in Data science

As per various surveys, data scientist job is becoming the most demanding Job of the 21st century due to increasing demands for data science. Some people also called it "the hottest job title of the 21st century". Data scientists are the experts who can use various statistical tools and machine learning algorithms to understand and analyze the data.

The average salary range for data scientist will be approximately $95,000 to $ 165,000 per annum, and as per different researches, about 11.5 millions of job will be created by the year 2026 (Camera, 2020).

Types of Data Science Job

If you learn data science, then you get the opportunity to find the various exciting job roles in this domain. The main job roles are given below:

1. Data Scientist
2. Data Analyst
3. Machine learning expert
4. Data engineer
5. Data Architect
6. Data Administrator
7. Business Analyst
8. Business Intelligence Manager

Below is the explanation of some critical job titles of data science.

Sl. No.	Job Title	Description	Skill Required	Tools Used
1	Data Analyst (DA)	DA performs mining of huge amount of data, models the data, looks for patterns, relationship, trends, and so on	Mathematics, business intelligence, data mining, and basic knowledge of statistics	MATLAB, Python, SQL, Hive, Pig, Excel, SAS, R, JS, Spark, etc.
2	Machine Learning Expert (MLE)	MLE works with various machine learning algorithms used in data science such as regression, clustering, classification, decision tree, random forest, etc.	Understanding of various algorithms, problem-solving analytical skill, probability, and statistics	Python, C++, R, Java, MATLAB and Hadoop
3	Data Engineer (DE)	DE works with massive amount of data and responsible for building and maintaining the data architecture of a data science project	Depth knowledge of SQL, MongoDB, Cassandra, HBase, Apache Spark, Hive, MapReduce	Python, C/C++, Java, Perl, MySQL, NoSQL, etc

4	Data Scientist (DS)	DS works with an enormous amount of data to come up with compelling business insights through the deployment of various tools, techniques, methodologies, algorithms, etc.	Must have an understanding of Statistics, Mathematics, Database, visualization, AI, NLP and communication skills	R, SAS, SQL, Python, Hive, Pig, Apache spark, MATLAB
5	Business Intelligence (BI) Manager	BI manager manages a team of analysts or developers who facilitate business intelligence processes and procedures development and implementation	Data Analysis, Decision Making, Communication Skills, Specific Industry domain knowledge	Tableau, SaaS, Sisense, MicroStrategy

Data Science road map

Data science is a broad field of study pertaining to data systems and processes, aimed at maintaining data sets and deriving meaning out of them. Data scientists use a combination of tools, applications, principles and algorithms to make sense of random data clusters. Since almost all kinds of organizations today are generating exponential amounts of data around the world, it becomes difficult to monitor and store this data. Data science focuses on data modelling and data warehousing to track the ever-growing data set. The information extracted through data science applications are used to guide business processes and reach rganizational goals. Data scientists primarily deal with huge chunks of data to analyse the patterns, trends and more (Learning, no date).

Artificial intelligence, or AI for short, has been around since the mid 1950s. It's not necessarily new. But it became super popular recently because of the advancements in processing capabilities. Back in the 1900s, there just wasn't the necessary computing power to realise AI. Today, we have some of the fastest computers the world has ever seen. And the algorithm implementations

have improved so much that we can run them on commodity hardware, even your laptop or smartphone that you're using to read this right now. And given the seemingly endless possibilities of AI, everybody wants a piece of it (Medium, 2019).

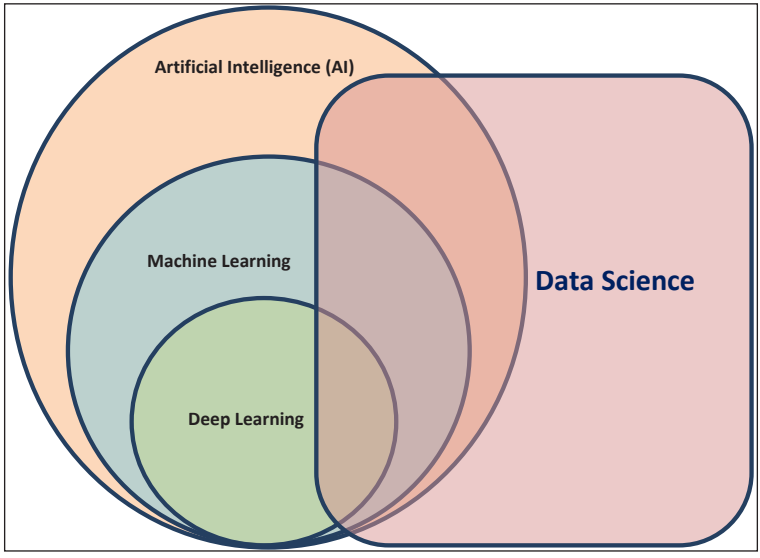

Fig. 1: Scope of Data Science

Machine Learning (ML) is considered a sub-set of AI. You can even say that ML is an implementation of AI. So whenever you think AI, you can think of applying ML there. As the name makes it pretty clear, ML is used in situations where we want the machine to learn from the huge amounts of data we give it, and then apply that knowledge on new pieces of data that streams into the system. But how does a machine learn, you might ask.

There are different ways of making a machine learn. Different methods of machine learning are supervised learning, non-supervised learning, semi-supervised learning, and reinforced machine learning. In some of these methods, a user tells the machine what are the features or independent variables (input)

and which is the dependent variable (output). So, the machine learns the relationship between the independent and dependent variables present in the data that is provided to the machine. This data which is provided is called the training set. And once the learning phase or the training is complete, the machine, or the ML model, is tested on a piece of data which the model has not encountered before. This new dataset is called the test dataset. There are different ways in which you can split your existing dataset between the training and the test dataset. Once the model is mature enough to give reliable and high accuracy results, the model will be deployed to a production setup where it will be used against absolutely new datasets for problems such as predictions or classification.

Deep Learning (DL) is an advancement of ML. Even though ML is super powerful for most applications, there are situations where ML leaves a lot to be desired. That is where deep learning steps in. It is generally believed that if your training dataset is relatively small, you go with ML. But if you have huge amounts of data on which you can train a model, and if the data has too many features, and if accuracy is super important (accuracy is always important though), you take the deep learning route (Srinidhi, 2019).

It is also important to note that deep learning requires much powerful hardware to run on (mostly GPUs are used), it takes significantly more time to train your models, and it is generally more difficult to implement compared to ML. But these are some of the compromises that you have to live with when the problem you're trying to solve is that much more complex.

You might have heard of TensorFlow, which is a neural network that Google is extensively using and pushing to developers. Well, that's using deep learning, as neural network is a kind of deep learning model. The self driving cars we started seeing in the last few years, they are self driving thanks to deep learning. There are many such applications of deep learning in the modern world

that are kind of behind the scenes. For example, entertainment services such as Netflix are using deep learning extensively to improve their recommendations for you, and also to decide, based on user engagement, which shows are worth continuing production, and which shows need to be axed because they're wasting time and money.

Prerequisite for Data Science

Non-Technical Prerequisite:

- **Curiosity:** To learn data science, one must have curiosities. When you have curiosity and ask various questions, then you can understand the business problem easily.
- **Critical Thinking:** It is also required for a data scientist so that you can find multiple new ways to solve the problem with efficiency.
- **Communication skills:** Communication skills are most important for a data scientist because after solving a business problem, you need to communicate it with the team.

Technical Prerequisite:

- **Machine learning:** To understand data science, one needs to understand the concept of machine learning. Data science uses machine learning algorithms to solve various problems.
- **Mathematical modeling:** Mathematical modeling is required to make fast mathematical calculations and predictions from the available data.
- **Statistics:** Basic understanding of statistics is required, such as mean, median, or standard deviation. It is needed to extract knowledge and obtain better results from the data.
- **Computer programming:** For data science, knowledge of at least one programming language is required. R,

Python, Spark are some required computer programming languages for data science.

- **Databases:** The depth understanding of Databases such as SQL, is essential for data science to get the data and to work with data.

Data Science Components

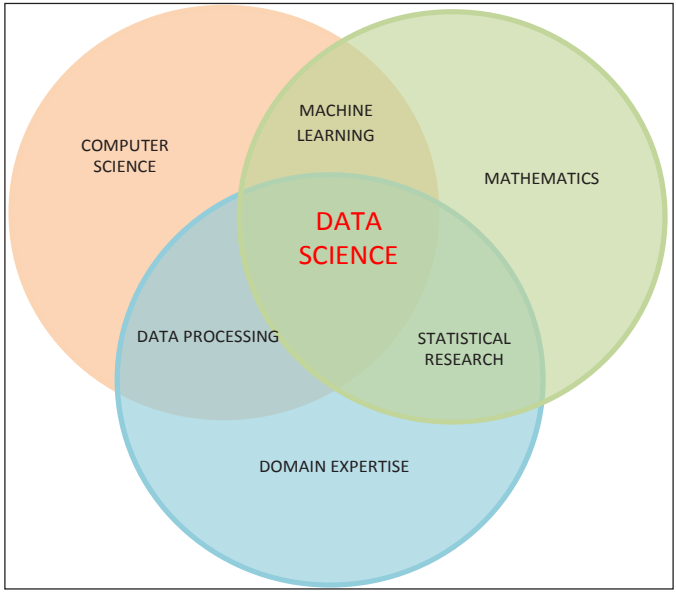

Fig. 2: Different Components of Data Science

1. **Statistics:** Statistics is one of the most important components of data science. Statistics is a way to collect and analyze the numerical data in a large amount and finding meaningful insights from it.
2. **Domain Expertise:** In data science, domain expertise binds data science together. Domain expertise means specialized knowledge or skills of a particular area. In data science, there are various areas for which we need domain experts.

3. **Data engineering:** Data engineering is a part of data science, which involves acquiring, storing, retrieving, and transforming the data. Data engineering also includes metadata (data about data) to the data.
4. **Visualization:** Data visualization is meant by representing data in a visual context so that people can easily understand the significance of data. Data visualization makes it easy to access the huge amount of data in visuals.
5. **Advanced computing:** Heavy lifting of data science is advanced computing. Advanced computing involves designing, writing, debugging, and maintaining the source code of computer programs.
6. **Mathematics:** Mathematics is the critical part of data science. Mathematics involves the study of quantity, structure, space, and changes. For a data scientist, knowledge of good mathematics is essential.
7. **Machine learning:** Machine learning is backbone of data science. Machine learning is all about to provide training to a machine so that it can act as a human brain. In data science, we use various machine learning algorithms to solve the problems.

Applications of Data Science

Image recognition and speech recognition: Data science is currently using for Image and speech recognition. When you upload an image on Facebook and start getting the suggestion to tag to your friends. This automatic tagging suggestion uses image recognition algorithm, which is part of data science.

When you say something using, "Ok Google, Siri, Cortana", etc., and these devices respond as per voice control, so this is possible with speech recognition algorithm.

Gaming world: In the gaming world, the use of Machine learning algorithms is increasing day by day. EA Sports, Sony, Nintendo, are widely using data science for enhancing user experience.

Internet search: When we want to search for something on the internet, then we use different types of search engines such as Google, Yahoo, Bing, Ask, etc. All these search engines use the data science technology to make the search experience better, and you can get a search result with a fraction of seconds.

Transport: Transport industries also using data science technology to create self-driving cars. With self-driving cars, it will be easy to reduce the number of road accidents.

Healthcare: In the healthcare sector, data science is providing lots of benefits. Data science is being used for tumor detection, drug discovery, medical image analysis, virtual medical bots, etc.

Recommendation systems: Most of the companies, such as Amazon, Netflix, Google Play, etc., are using data science technology for making a better user experience with personalized recommendations. Such as, when you search for something on Amazon, and you started getting suggestions for similar products, so this is because of data science technology.

Risk detection: Finance industries always had an issue of fraud and risk of losses, but with the help of data science, this can be rescued.

Most of the finance companies are looking for the data scientist to avoid risk and any type of losses with an increase in customer satisfaction.

Data Science Life-cycle

The life-cycle of data science is explained as below diagram.

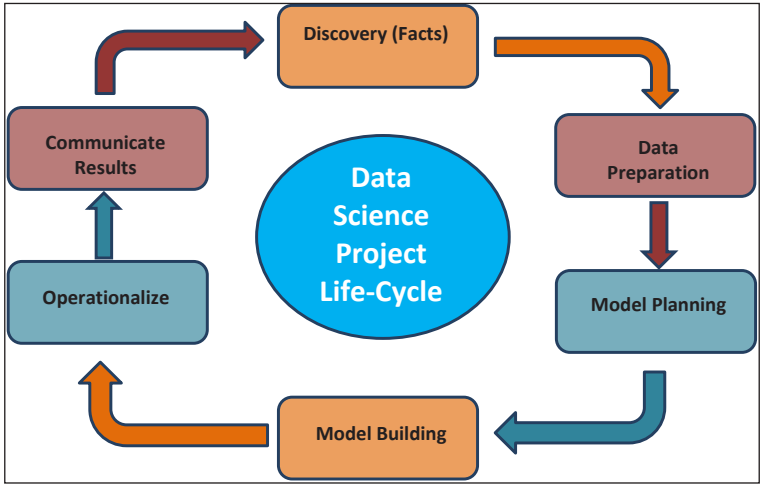

Fig. 3: Life Cycle of Data Science Project

The main phases of data science life cycle are given below:

1. **Discovery:** The first phase is discovery, which involves asking the right questions. When you start any data science project, you need to determine what are the basic requirements, priorities, and project budget. In this phase, we need to determine all the requirements of the project such as the number of people, technology, time, data, an end goal, and then we can frame the business problem on first hypothesis level.
2. **Data preparation:** Data preparation is also known as Data Munging. In this phase, we need to perform the following tasks:
   - Data cleaning
   - Data Reduction
   - Data integration
   - Data transformation,

   After performing all the above tasks, we can easily use this data for our further processes.

3. **Model Planning:** In this phase, we need to determine the various methods and techniques to establish the relation between input variables. We will apply Exploratory data analytics (EDA) by using various statistical formula and visualization tools to understand the relations between variable and to see what data can inform us. Common tools used for model planning are:
   - SQL Analysis Services
   - R
   - SAS
   - Python
4. **Model-building:** In this phase, the process of model building starts. We will create datasets for training and testing purpose. We will apply different techniques such as association, classification, and clustering, to build the model.

   Following are some common Model building tools:
   - SAS Enterprise Miner
   - WEKA
   - SPCS Modeler
   - MATLAB
5. **Operationalize:** In this phase, we will deliver the final reports of the project, along with briefings, code, and technical documents. This phase provides you a clear overview of complete project performance and other components on a small scale before the full deployment.
6. **Communicate results:** In this phase, we will check if we reach the goal, which we have set on the initial phase. We will communicate the findings and final result with the business team.

Courses & Universities

Table 2, there are 12 different courses which are started in 12 different universities are listed. The courses belongs to different level (UG or PG) but mostly are degree level, a few are diploma level. Course duration and proposed fee structure is also mention there. Out of this list IIT Madras tool bold look, along with new guidelines IITM started first ever B.Sc. course in Data Science with different exit scheme (few Input also). It also seems cheapest undergraduate course. More details are in next section

Table 2: Course & Universities Who started in 2019

Study of Data Science Course offered by Indian Universities				
University Name	**Private/Government**	**Course offered**	**Course Duration**	**Current Fee**
Poorina University, Jaipur	Private	B.Tech CSE (Data-Science)	4 Years	145000/Year
Techno India University, Kolkata	First Private Uni in WB	B.Sc. data Science	3 Years	410000
National Institute of Securities Markets (NISM), Mumbai	Government	PG Diploma in Data Science	9 months (PT)	1,25000
NSHM campus at Durgapur & Kolkata	Private	M.Sc.	2 Years	200000+
University of Technology, Jaipur	Self-finance	B.Tech Data science	4 Years	**90000/Year (minimum Fee)**
St Xavier's College - Autonomous , Mumbai (MOU with TCS)	Reputed management college	M.Sc Big Data & Analytics	2 Years	90000/Year; Next year will revise
Amrita Vidyapeetham, Coimbatore	Private	Integrated M.Sc (Data Science)	5 Years	200000/Years
PSG College of Technology, Coimbatore	(government aided) Private	Integrated M.Sc. (Data Science)	5 Years	2,26000+other
Coimbatore institute of technology	Government	Integrated M.Sc. (Data Science)	5 Years	130000/Year
IISC Bangalore	Government	M.Sc. Computational & Data Science	2 Years	34000/Year (lowest)
Bits-Pilani	Aided	M.Tech Data Science	2 Years	2,27000 (Total

B.Sc. Course from IIT Madras

For the first time, in India online BSc Degree program in Programming and Data Science was started by IIT Madras. Which also follows the new guidelines of HRM which as suggested by in 2020. It has 3 stages namely foundation level, diploma level & degree level. If student will left course after 1 year he/she

will entitle for certificate course and if he exit after 2nd year than student will entitle for diploma certificate and after completion of 3rd year he/she will get the B.Sc. degree. The fees also differ, for 1st year it is only INR 3200 which in 2nd year it is INR 1, 10,000 (which is again dividing in 2 parts 55,000 each) and for 3rd year it is 100000. This course have 3 exit scheme but again student also rejoin the course when they feel free to join and can start at same place where they exit. If they exit after certificate than again they can join for diploma and if they left after diploma than they can join again for degree level (IITMadras, 2020).

The below figure have the clear view of in & out possibilities of course.

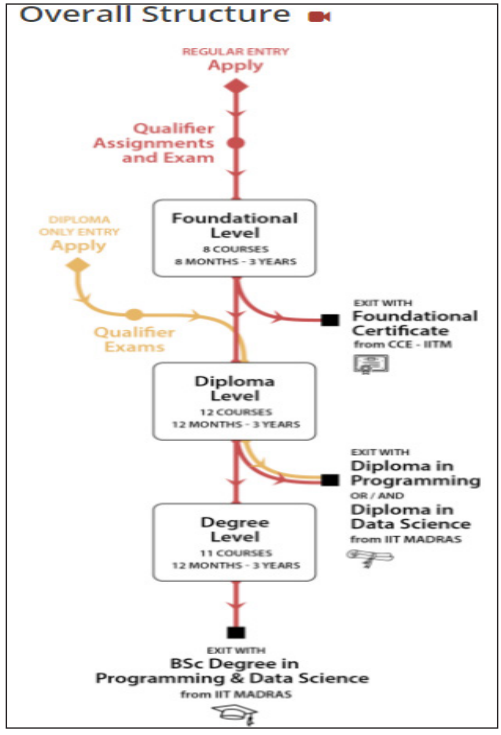

Fig. 4: Course structure with different in & out provisions

There are minimum 3 years for BSc. Course to complete while maximum time is 6 years. Course contains 31 credits to earn to complete the course. There different set of elective student can select later as per their interest. Evaluation is based on Quizzes during studies and end term exam later the course study at other hand to clear a particular subject student have to secure 50% marks in Quizzes and 40% marks in end term exam (more details are available at https://onlinedegree.iitm.ac.in/).

CONCLUSION

In few time Data Science is most popular among students. And all universities took advantage of this new trend and started different course at different level and IIT Madras is the best example of this. This chapter gives a details introduction to Data Science with all required information. It will help to understand the term Data science and one will take decision to choose right course and correct path for their career.

EXERCISE

1. What do you mean by Data science, explain in detail?
2. What are different data sciences course available?
3. What are different applications of Data science?
4. What is life-cycle of Data science project?
5. What are different components of data science
6. What are different job profiles related to data science?

REFERENCES

1. Camera, D. 2020. *Data Science Quick Start*. Available at: https://medium.com/datadriveninvestor/data-science-quick-start-408ddf05b63d?source=rss------machine_learning-5.
2. IIT Madras. 2020. *Programming & Data Science*. Available at: https://onlinedegree.iitm.ac.in/.

3. Learning, G. (no date) *Data Science vs Machine Learning and Artificial Intelligence*. Available at: https://www.mygreatlearning.com/blog/difference-data-science-machine-learning-ai/.

4. LearningPoint92 (no date) *Why Data Science?* Available at: https://learningpoint92.blogspot.com/2019/09/why-data-science.html.

5. Medium, D.L. 2019. *Data Science vs. Artificial Intelligence vs. Machine Learning vs. Deep Learning*. Available at: https://mc.ai/data-science-vs-artificial-intelligence-vs-machine-learning-vs-deep-learning/.

6. Srinidhi, S. 2019 *Data Science vs. Artificial Intelligence vs. Machine Learning vs. Deep Learning*. Available at: https://towardsdatascience.com/data-science-vs-artificial-intelligence-vs-machine-learning-vs-deep-learning-9fadd8bda583.

7. Wikipedia (no date) *Data science*. Available at: https://en.wikipedia.org/wiki/Data_science.