Data science uses techniques such as machine learning and artificial intelligence to predict future patterns and behaviors

Mr. BHARAT B. DAREKAR
Department of Computer Science, KTHM College, Nashik
bbdarekar67@gmail.com

Abstract

Data science incorporates tools from multiple disciplines to gather a data set, process, and derive insights from the data set, extract meaningful data from the set, and interpret it for decision-making purposes. The disciplinary areas that make up the data science field include mining, statistics, machine learning, analytics, programming, Big data analysis, data driven theory, operations research, random processes, social network analysis, financial technology, quantum computing, intelligent computing, cloud computing, optimization theory, decision-making theory, computer simulation technology, health informatics, medical big data, and health management. Data science combines aspects of different fields with the aid of computation to interpret data for decision-making purposes. Data mining applies algorithms to the complex data set to reveal patterns that are then used to extract useful and relevant data from the set.

Machine learning is an artificial intelligence tool that processes mass quantities of data that a human would be unable to process in a lifetime.

Data Analytics, the data analyst collects and processes the structured data from the machine learning stage using algorithms. The analyst interprets, converts, and summarizes the data into a cohesive language that the decision-making team can understand. Data science is applied to practically all contexts and, as the data scientist's role evolves, the field will expand to encompass data architecture, data engineering.

Key words:

DA (data Analytics, AI (Artificial Intelligence), DA (Data Architecture), BDA (Big Data Analysis)

Introduction:

What is Data Science?

Data science is a field of applied mathematics and statistics that provides useful information based on large amounts of complex data or big data. Data science combines aspects of different fields with the aid of computation to interpret reams of data for decision-making purposes.

Understanding Data Science

Data is drawn from different sectors, channels, and platforms, including cell phones, social media, e-commerce sites, healthcare surveys, and internet searches. The continually increasing access to data is possible due to advancements in technology and collection techniques. Individuals buying patterns and behavior can be monitored and predictions made based on the information gathered.

Data is unstructured and requires parsing for effective decision-making. This process is complex and time-consuming for companies hence, the emergence of data science. It uses big data and machine learning to interpret data for decision-making purposes.

Data science uses techniques such as machine learning and artificial intelligence to extract meaningful information and to predict future patterns and behaviors.

The field of data science is growing as technology advances and big data collection and analysis techniques become more sophisticated.

How Data Science is applied

Data science incorporates tools from multiple disciplines to gather a data set, process, and derive insights from the data set, extract meaningful data from the set, and interpret it for decision-making purposes. The disciplinary areas that make up the data science field include mining, statistics, machine learning, analytics, and programming.

Data mining applies algorithms to the complex data set to reveal patterns that are then used to extract useful and relevant data from the set.

Machine learning is an artificial intelligence tool that processes mass quantities of data that a human would be unable to process in a lifetime. Machine learning perfects the decision model presented under predictive analytics by matching the likelihood of an event happening to what actually happened at a predicted time.

Data science is applied to practically all contexts and, as the data scientist's role evolves, the field will expand to encompass data architecture, data engineering, and data administration.

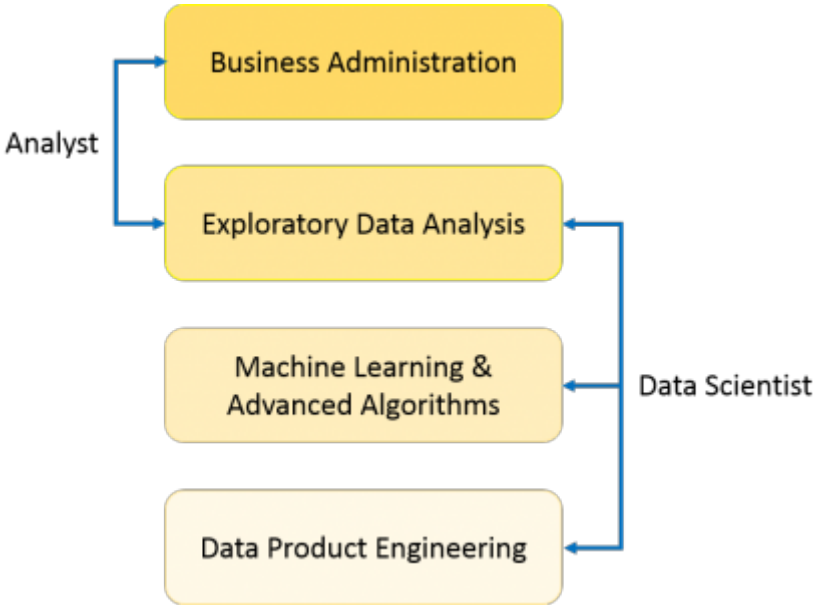
A data scientist collects, analyzes, and interprets large volumes of data, in many cases, to improve a company's operations. Data scientist professionals develop statistical models that analyze data and detect patterns, trends, and relationships in data sets. This information can be used to predict consumer behavior or to identify business and operational risks.

Banking institutions are capitalizing on big data to enhance their fraud detection successes. Asset management firms are using big data to predict the likelihood of a security’s price moving up or down at a stated time. Companies such as Netflix mine big data to determine what products to deliver to their users. Netflix also uses algorithms to create

personalized recommendations for users based on their viewing history. Data science is evolving at a rapid rate, and its applications will continue to change lives into the future.

Methodology (Case Study):

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.



As you can see from the above image, a Data Analyst usually explains what is going on by processing history of the data. On the other hand, Data Scientist not only does the exploratory analysis to discover insights from it, but also uses various advanced machine learning algorithms to identify the occurrence of a particular event in the future.

So, Data Science is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics (predictive plus decision science) and machine learning.

There are several definitions available on Data Scientists. In simple words, a Data Scientist is one who practices the art of Data Science. The term “Data Scientist” has been coined after considering the fact that a Data Scientist draws a lot of information from the scientific fields and applications whether it is statistics or mathematics.

What does a Data Scientist do?

Data scientists are those who crack complex data problems with their strong expertise in certain scientific disciplines. They work with several elements related to mathematics, statistics, computer science, etc. They make a lot of use of the latest technologies in finding solutions and reaching conclusions that are crucial for an organization’s growth and

development. Data Scientists present the data in a much more useful form as compared to the raw data available to them from structured as well as unstructured forms.

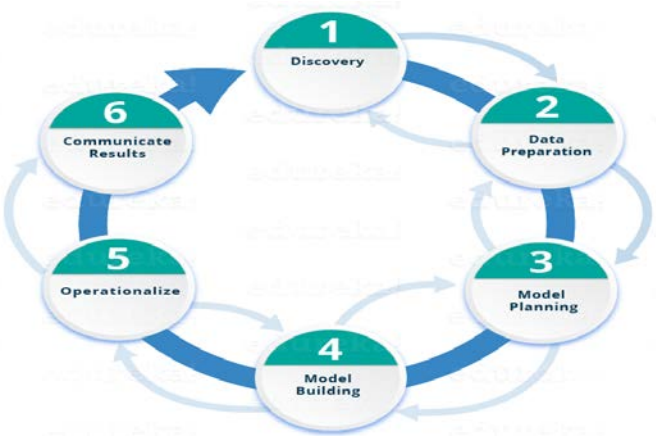
Business Intelligence (BI) vs. Data Science

- Business Intelligence (BI) basically analyzes the previous data to find hindsight and insight to describe business trends. Here BI enables you to take data from external and internal sources, prepare it, run queries on it and create dashboards to answer questions like quarterly revenue analysis or business problems.
- Data Science is a more forward-looking approach, an exploratory way with the focus on analyzing the past or current data and predicting the future outcomes with the aim of making informed decisions.

Features	**Business Intelligence (BI)**	**Data Science**
Data Sources	Structured (Usually SQL, often Data Warehouse)	Both Structured and Unstructured (logs, cloud data, SQL, NoSQL, text)
Approach	Statistics and Visualization	Statistics, Machine Learning, Graph Analysis, Neuro-linguistic Programming (NLP)
Focus	Past and Present	Present and Future
Tools	Pentaho, Microsoft BI, QlikView, R	RapidMiner, BigML, Weka, R

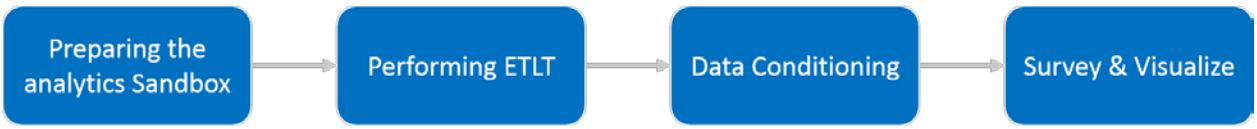
Lifecycle of Data Science

Main phases of the Data Science Lifecycle:

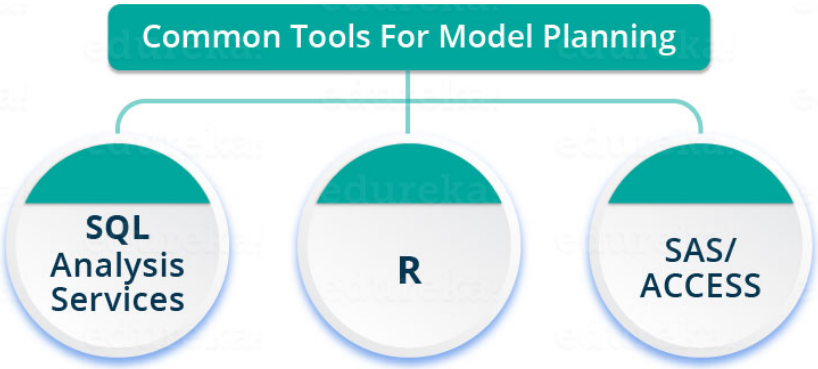


Phase 1—Discovery: Before you begin the project, it is important to understand the various specifications, requirements, priorities and required budget. You must possess the ability to ask the right questions. In this phase, you also need to frame the business problem and formulate initial hypotheses (IH) to test.

Phase 2—Data preparation: It require analytical sandbox in which you can perform analytics for the entire duration of the project. You need to explore, preprocess and condition data prior to modeling. Further, you will perform ETLT (extract, transform, load and transform) to get data into the sandbox.

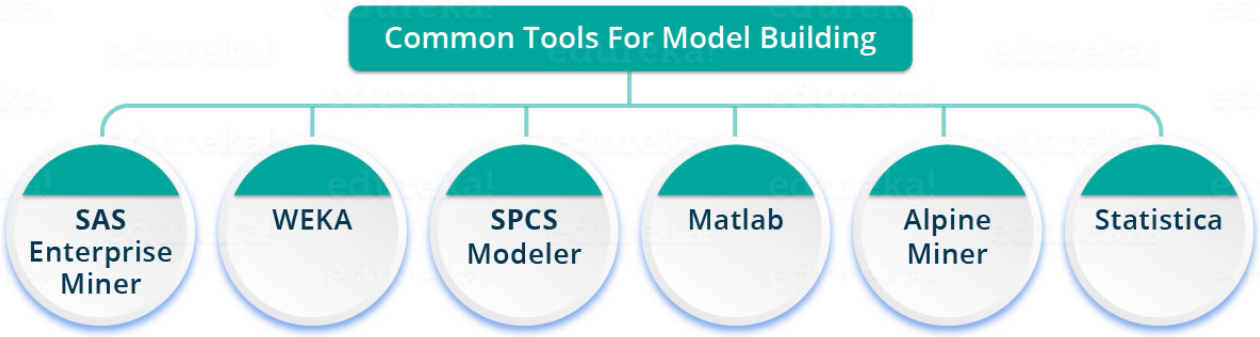


.**Phase 3—Model planning:** We will determine the methods and techniques to draw the relationships between variables. These relationships will set the base for the algorithms which you will implement in the next phase.



1. **R** has a complete set of modeling capabilities and provides a good environment for building interpretive models.
2. **SQL Analysis services** can perform in-database analytics using common data mining functions and basic predictive models.
3. **SAS/ACCESS** can be used to access data from Hadoop and is used for creating repeatable and reusable model flow diagrams.

Phase 4—Model building: In this phase, you will develop datasets for training and testing purposes. We will analyze various learning techniques like classification, association and clustering to build the model.



Phase 5—operationalize: In this phase, you deliver final reports, briefings, code and technical documents. In addition, sometimes a pilot project is also implemented in a real-time production environment.

Phase 6—Communicate results: It is important to evaluate if you have been able to achieve your goal that you had planned in the first phase. So, in the last phase, you identify all the key findings, communicate to the stakeholders and determine if the results of the project are a success or a failure.

In this use case, we will predict the occurrence of diabetes making use of the entire lifecycle that we discussed earlier. Let's go through the various steps.

Step 1:

- we will collect the data based on the medical history of the patient as discussed in Phase

Step 2:

- Now, once we have the data, we need to clean and prepare the data for data analysis.
- Here, we have organized the data into a single table under different attributes – making it look more structured.

This data has a lot of inconsistencies.

Step 3: Now let's do some analysis as discussed earlier in Phase 3.

- First, we will load the data into the analytical sandbox and apply various statistical functions on it. For example, R has functions like *describe* which gives us the number of missing values and unique values. We can also use the summary function which will give us statistical information like mean, median, range, min and max values.

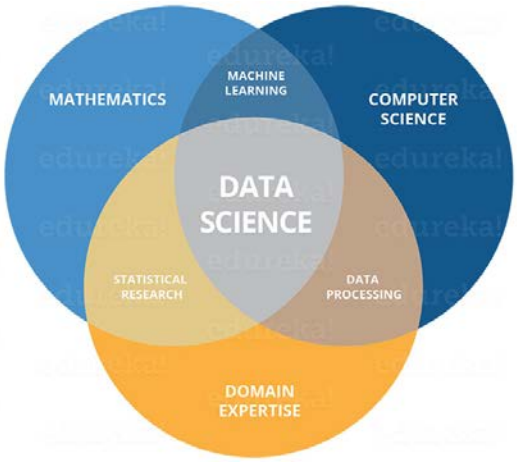
	npreg	glu	bp	skin	bmi	ped	age
1	6	148	72	35	33.6	0.627	50
2	1	85	66	29	26.6	0.351	31
3	1	89	80	23	28.1	0.167	21
4	3	78	50	32	31	0.248	26
5	2	197	70	45	30.5	0.158	53
6	5	166	72	19	25.8	0.587	51
7	0	118	84	47	45.8	0.551	31
8	1	103	30	38	43.3	0.183	33
9	3	126	88	41	39.3	0.704	27
10	9	119	80	35	29	0.263	29
11	1	97	66	15	23.2	0.487	22
12	5	109	75	26	36	0.546	60
13	3	88	58	11	24.8	0.267	22
14	10	122	78	31	27.6	0.512	45
15	4	97	60	33	24	0.966	33
16	9	102	76	37	32.9	0.665	46
17	2	90	68	42	38.2	0.503	27
18	4	111	72	47	37.1	1.39	56
19	3	180	64	25	34	0.271	26
20	7	106	92	18	39	0.235	48
21	9	171	110	24	45.4	0.721	54

Step 4: Now, based on insights derived from the previous step, the best fit for this kind of problem is the decision tree. Let's see how?

Step 5: In this phase, we will run a small pilot project to check if our results are appropriate. We will also look for performance constraints if any. If the results are not accurate, then we need to replan and rebuild the model.

Step 6: Once we have executed the project successfully, we will share the output for full deployment.

A Data Scientist requires skills basically from three major areas as shown below.



In the end, it won't be wrong to say that the future belongs to the Data Scientists. It is predicted that by the end of the year 2018, there will be a need of around one million Data Scientists. More and more data will provide opportunities to drive key business decisions. It is soon going to change the way we look at the world deluged with data around us. Therefore, a Data Scientist should be highly skilled and motivated to solve the most complex problems.

Types of Data Science Jobs: Data Scientist Data Analyst

Business Analyst	Data engineer
Business Intelligence Analyst	Machine learning Engineer
Statistician	

Data scientist responsibilities

1. Solving business problems through explored research and constructing open-ended industry questions.
2. Collect huge volumes of unstructured and structured data. They have to query structured data from relational databases using programming languages such as SQL.
3. Employ polished analytical methods, statistical and machine learning methods to prepare data.
4. Rigorously clean data to discard irrelevant information and prepare the data for modeling and preprocessing.
5. Carry out exploratory data analysis (EDA) for understanding how to handle missing data and to look for trends and/or opportunities.

Data Scientist Job Roles

Some of the prominent Data Scientist job titles are:

Data Scientist	Data Engineer	Data/Analytics Manager
Data Architect	Data Administrator	Business Intelligence Manager
Data Analyst	Business Analyst	

The time to up-skill in Data Science and Big Data Analytics to take advantage of the Data Science career opportunities that come your way. This brings us to the end of Data Science tutorial blog. I hope this blog was informative and added value to you. Now is the time to enter the Data Science world and become a successful Data Scientist.

Data science is an essential part of many industries today, given the massive amounts of data that are produced, and is one of the most debated topics in IT circles. Its popularity has grown over the years, and companies have started implementing data science techniques to grow their business and increase customer satisfaction. In this article, we'll learn what data science is, and how you can become a data scientist.

Conclusion:

A data scientist analyzes business data to extract meaningful insights. In other words, a data scientist solves business problems through a series of steps, including:

- Before tackling the data collection and analysis, the data scientist determines the problem by asking the right questions and gaining understanding.
- The data scientist then determines the correct set of variables and data sets.

- The data scientist gathers structured and unstructured data from many disparate sources—enterprise data, public data, etc.
- Once the data is collected, the data scientist processes the raw data and converts it into a format suitable for analysis. This involves cleaning and validating the data to guarantee uniformity, completeness, and accuracy.
- After the data has been rendered into a usable form, it's fed into the analytic system—ML algorithm or a statistical model. This is where the data scientists analyze and identify patterns and trends.
- When the data has been completely rendered, the data scientist interprets the data to find opportunities and solutions.
- The data scientists finish the task by preparing the results and insights to share with the appropriate stakeholders and communicating the results.

Data Scientist:

- Job role: Determine what the problem is, what questions need answers, and where to find the data. Also, they mine, clean, and present the relevant data.
- Skills needed: Programming skills (SAS, R, Python), storytelling and data visualization, statistical and mathematical skills, knowledge of Hadoop, SQL, and Machine Learning.

Data Analyst:

- Job role: Analysts bridge the gap between the data scientists and the business analysts, organizing and analyzing data to answer the questions the organization poses. They take the technical analyses and turn them into qualitative action items.
- Skills needed: Statistical and mathematical skills, programming skills (SAS, R, Python), plus experience in data wrangling and data visualization.

Data Engineer:

- Job role: Data engineers focus on developing, deploying, managing, and optimizing the organization's data infrastructure and data pipelines. Engineers support data scientists by helping to transfer and transform data for queries.

Data Science Tools

The data science profession is challenging, but fortunately, there are plenty of tools available to help the data scientist succeed at their job.

- Data Analysis: SAS, Jupiter, R Studio, MATLAB, Excel, RapidMiner
- Data Warehousing: Informatica/ Talend, AWS Redshift
- Data Visualization: Jupiter, Tableau, Cognos, RAW

- Machine Learning: Spark MLib, Mahout, Azure ML studio

The disciplinary areas that make up the data science field include mining, statistics, machine learning, analytics, programming, Big data analysis, data driven theory, operations research, random processes, social network analysis, financial technology, quantum computing, intelligent computing, cloud computing, optimization theory, decision-making theory, computer simulation technology, health informatics, medical big data, and health management. Data science combines aspects of different fields with the aid of computation to interpret data for decision-making purposes.

References:-

1. www.investopedia.com/terms/d/data-science.asp
2. en.wikipedia.org/wiki/Data science
3. www.edureka.co/blog/what-is-data-science/
4. https://www.edureka.co/blog/data-science-tutorial/
5. Aden so-Diaz, B., Laguna, M.: Fine-tuning of algorithms using fractional experimental designs and local search. Oper. Res. 54(1), 99–114 (2006)
6. Aggarwal, C.C. (ed.): Data Classification: Algorithms and Applications. CRC Press, Boca Raton (2014)
7. Allen, E., Allen, L., Arciniega, A., Greenwood, P.: Construction of equivalent stochastic differential equation models. Stoch. Anal. Appl. 26, 274–297 (2008)
8. Anderson, C.: The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine https://www.wired. com/2008/06/pb-theory/ (2008)
9. Aue, A., Horváth, L.: Structural breaks in time series. J. Time Ser. Anal. 34(1), 1–16 (2013) 6. Berger, R.E.: A scientific approach to writing for engineers and scientists. IEEE PCS Professional Engineering Communication Series IEEE Press, Wiley (2014)
10. Bischl, B., Mersmann, O., Trautmann, H., Weihs, C.: Resampling methods for meta-model validation with recommendations for evolutionary computation. Evol. Comput. 20(2), 249–275 (2012)