Introduction to Data Science

Relevel
by Unacademy

What is Data Science?

Data science is a cross-functional field of study and practice that is highly focused on obtaining insights & knowledge from data using scientific methods, algorithms, systems & processes.

Data science is rooted in solid foundations of mathematics and statistics, computer science, and domain knowledge.

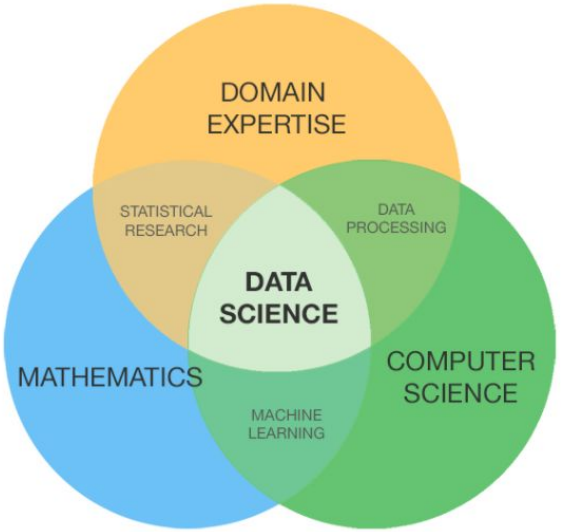
In short, Data Science is a field where we apply ‘science’ to available ‘data’ to get the ‘patterns’ or ‘insights’ that can help a business optimize operations or improvise decisions.

Interdisciplinary fields of Data Science

Data Science is a combination of various fields which are used to extract actionable meaningful insights from the data provided. Few such areas to name here are:

- Mathematics
- Statistical Research
- Computer Science
- Data processing concepts
- Machine Learning
- Natural Language Processing (NLP) concepts
- Domain expertise (Business Knowledge)

A skilled person in all the sub-domains mentioned above of data science is called a Data Scientist.

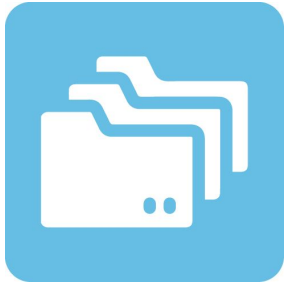


What is Data?

Data are individual facts, statistics, or items of information.

In a more technical sense, data are values of qualitative or quantitative variables about one or more persons or objects.

In the real-world, data is not limited to a small set. It is big data.



Sources of data

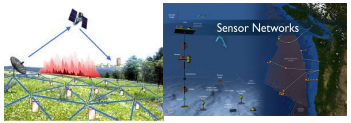
The data come from several sources. They include:

Data generated from sensors used to gather climate information
posts to social media sites,
digital pictures and videos
purchase transaction records
cell phone GPS signals etc.

We create 2.5 quintillion bytes of data every day, so more than 90% of the data in the world today has been created in the last two years alone.

Now data is Big data!

Memory unit	Size	Binary size
kilobyte (kB/KB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}



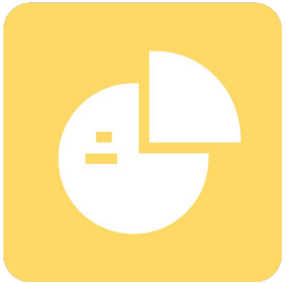
Sensor technology and networks
(Measuring all kinds of data)

Scientific instruments
(Collecting all sorts of data)

Social media and networks
(All of us are generating data)

Mobile devices
(Tracking all objects all the time)

What is Big Data?



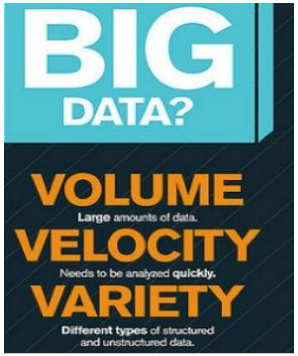
Big data is the data that has greater variety, arriving in increasing volumes and with more velocity.

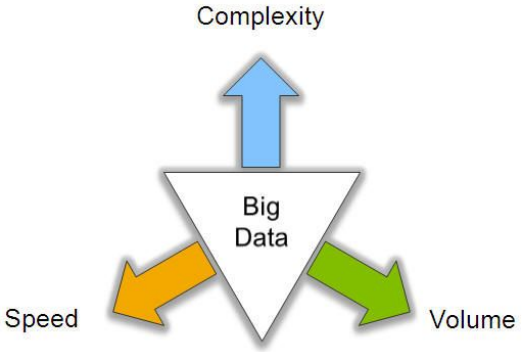
In simple words, Big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software can't manage them. Any small or big business manages a considerable amount of data through its various data points and business processes.

At times, businesses can sometimes handle these data using excel sheets, access databases or other similar tools.
Every business, big or small, manages a considerable amount of data generated. However, instances increase above acceptable limits when data cannot fit into such tools and human error due to intensive manual processing. Big Data and Data Science then come into the picture.

Big Data can be defined using the famous 3 Vs – Volume, Velocity and Variety.

Characteristics of Big data: V3

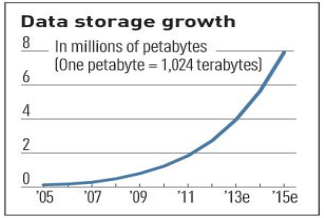


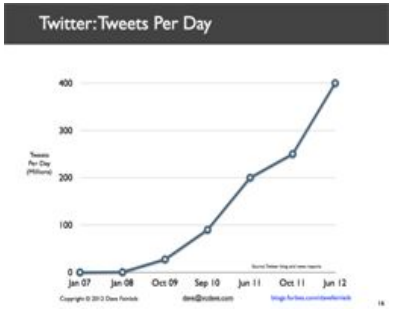


V3 : V for Volume

The volume of raw data is increasing rapidly. Because of this, there is a need of:

- More storage capacity
- More computation
- More tools and techniques





Exponential increase in collected/generated data

V3: V for Velocity

Data is being generated fast and need to be processed quickly.

Big data must be used for time-sensitive processes such as catching fraud, as it streams into your enterprise to maximize its value

Scrutinize 5 million trade events created each day to identify potential fraud

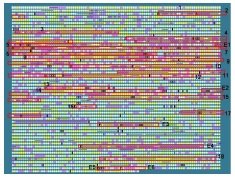
Analyze 500 million daily call detail records in real-time to predict customer churn faster

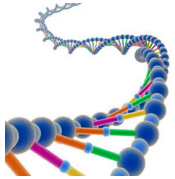
Sometimes, 2 minutes is too late!
The latest news is that the delay of ten ns (nanoseconds) is too much.

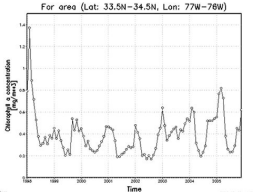
V3: V for Variety

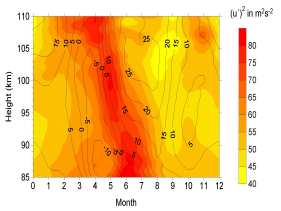
Various formats, types, and structures

- Text
- Numerical data
- Images
- Audio
- Video
- Sequences
- Time series
- Social media data
- Multi-dimensional arrays, etc.
- Streaming data





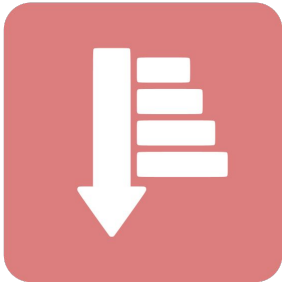




A single application can be generated/collected many types of data.

To extract knowledge , all these types of data need to be linked together.

V3: V for Variety



At a high level, data can be classified into three types:

a. **Structured Data:** The data in the format of relational database and have structured properly in rows and columns format is known as Structured Data.

b. **Unstructured Data:** The data which includes various types of data such as audio, video, XML file, word file, etc. If it is not organised in a proper format, it is said to be Unstructured Data.

c. **Semi-structured Data:** Semi-structured data is self-explanatory in that it is the data that is not fully structured or unstructured. In it, data is partially structured and mixed with the unstructured data format.
For Example, social media contains photos, videos and texts of people in huge figures. This data is nothing but big data; it can be well-structured, unstructured, or semi-structured.

Why is data science important?

According to IDC, the global data will grow to 175 zettabytes by 2025.

Data Science enables companies to understand gigantic data from multiple sources and derive valuable insights to make smarter data-driven decisions accurately.

Data Science is extensively used in various industry domains, including marketing, healthcare, finance, banking, policy work, etc.. That explains why Data Science is important.

Importance of data science in business

There are several reasons why Data Science is important in business.

- Data Science enables companies to measure, track, and record performance metrics for facilitating enterprise-wide enhanced decision making.
- Companies can examine trends to make certain critical decisions to enhance company performance, engage customers better, and increase profitability.
- Data Science models use existing data that can simulate several actions. Therefore, companies can devise this path to reap the best business outcomes.
- Data Science helps enterprises identify target audiences by combining existing data with other data points for developing useful insights.
- Data Science also helps recruiters by combining several data points to identify candidates that best fit their company needs.

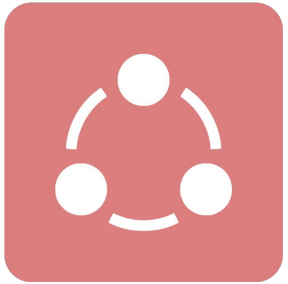
Industrial Sector-wise benefits of data science

The organizational importance of Data Science is constantly increasing.
Some of the many Data Science benefits comprise the following:

- Physicians use Data Science to analyze the data from wearable trackers to ensure their patients' well-being and make vital decisions in the healthcare industry.
- Data Science is broadly used in the banking and finance sectors for personalized financial advice and fraud detection.
- Transportation providers use Data Science to strengthen the transportation journeys of their customers.
- Construction companies use Data Science for effective decision making by tracking activities, including average time for completing tasks, material-based expenses, etc.
- Data Science enables trapping and analyzing huge amounts of data from manufacturing processes, which has gone untapped.

Industrial Sector-wise benefits of data science

- Data Science helps one analyze massive graphical, temporal, and geospatial data to extract insights. It also helps in seismic interpretation and reservoir characterization.
- Data Science allows firms to leverage social media content to obtain real-time media content usage patterns. This also helps the firms to create target audience-specific content, measure content performance, and recommend on-demand content.
- Data Science helps study utility consumption in the energy and utility domain. This study allows for better control of utility use and enhanced consumer feedback.
- Data Science applications in the public service field include health-related research, financial market analysis, fraud detection, energy exploration, environmental protection, etc.



Career options in Data Science

- The career options in Data Science Domain include:

✓ Data Scientist

✓ Data Analyst

✓ Data Engineer

✓ Business Intelligence Developer

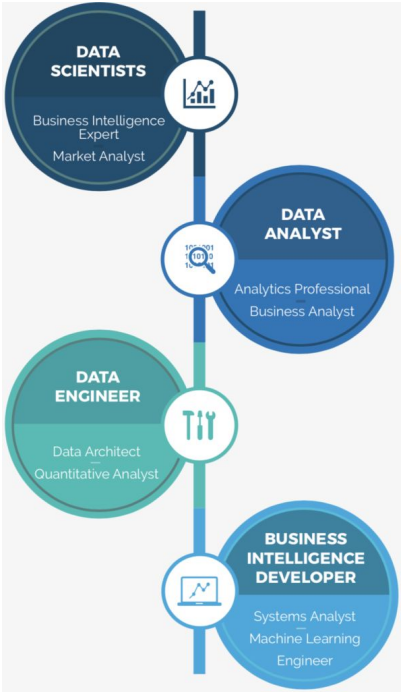
✓ Database Administrator

✓ Business Analyst

✓ Machine Learning Engineer

✓ Data Architect

Let's try to understand these positions a bit more!



Career options in Data Science

1. Data Scientist

A ‘Data Scientist’ is one of the most desired positions in any company.

In the modern workplace, data scientists build machine learning models for prediction, find patterns and trends in data, visualise data, and even pitch in with marketing strategies.

Skillset: Statistics, Mathematics, Data Modelling, Python or R programming,

Other skills: Database skills, Business acumen, Visualization/BI, Presentation skills

2. Data Analyst:

A 'Data Analyst' is in charge to collect, process, and transforming data into usable forms. They help companies make effective business decisions. Depending on the type of organization, an analyst’s job could include tracking web analytics, extracting insights from consumer datasets, analyzing A/B testing, making strategic recommendations based on financial data, or merely organising messy, unstructured data

Skillset: Data Modelling, Python or R programming, HTML, visualization tools, and SQL

Other skills: Business acumen, Database cleaning skills, Presentation skills

Career options in Data Science

3. Data Engineer

A Data Engineer is considered the backbone of any big organization. Companies usually hire data engineers to channel their talents towards software development.

Data engineers are the architects of big data and information pipelines. They strive to create a reliable, interconnected data network for use within an organization.

They design, build and manage systems that analyse and process data for the organisation. They design, build and manage systems that examine and process data for the organization. They also ensure that the systems run effortlessly. This job is unlike typical data science careers because it focuses more on hardware and data warehousing than analysis. Data Engineer is an advanced position and requires a background in software engineering. They must be skilled in SQL, Databases, Big Data, Cloud computing, Java, Python, Ruby, MATLAB, Hive, Pig, SAS, etc.

Other skills: Business acumen, Database cleaning skills, Visualization/BI, Presentation skills

Career options in Data Science

4. Business Intelligence Developer

This position requires extensive knowledge about the business and the potential to convert data into consumable data products. BI developers/analysts gather data, design and develop systems to increase a firm's efficiency, and help management make better decisions. They do this either by mining from a company's software or reviewing competitor data and industry trends. These experts are expected to be tech-savvy. BI developers either use the existing BI tools or develop their own BI analytic applications.

Skillset: Python or R programming, Hadoop, creating models, Notebook, GitHub, data modelling

Other skills: Business acumen, Visualization/BI

Career options in Data Science

5. Database Administrator

Database administrators (DBAs) are in charge of successfully storing and organizing an organization's data. DBAs are often seen as the custodians of a company's data as they – create highly available databases, structure security rules and protocols to safeguard data and perform upgrades to keep databases up to date. DBAs must have at least a bachelor's degree in computer science or information technology. DBAs are skilled in software like SQL, Database management, Big Data, Algorithms, Optimization, UNIX/ Linux, Data Analysis, AWS, Python, etc.

6. Business Analyst

Business Analysis is an excellent career choice for someone with a business educational background and a strong foundation in numbers. Business analysis is a less technical position requiring an alliance between business and IT. A business analyst requires knowledge of business processes, data visualization tools, and data modelling. They are required to understand and map out business processes, identify key business problems that can be solved with analytics, and collaborate with other technical groups to translate business problems into solutions.

Career options in Data Science

7. Machine Learning Engineer

Machine learning engineers work on machine learning-focused software solutions to meet an organization's needs. They design and build machine learning systems that interpret data to make predictions or draw conclusions. In addition to creating data funnels and machine learning software, they also run tests and closely monitor the system's performance to ensure accuracy. These engineers must have advanced skills in statistics, programming, and data science.

8. Data Architect

With the growing importance of big data, this position is becoming more critical. Data architects develop new database systems, design analytics applications, and create blueprints to integrate, centralize, maintain, and protect data. They ensure the performance of data solutions, improve the functionality of existing systems, and provide access to database analysts and administrators. The position requires an in-depth understanding of languages like Hive, XML, SQL, Pig, Spark, systems development, and database architecture skills.

Prerequisite for Data Science

Non-Technical Prerequisite:

- Good Communication skills
- Good listener
- Acceptance for change
- Ability to multitask
- Understanding priorities
- Ability to handle multiple stakeholders
- Identify opportunities to improve processes
- Working as part of a team

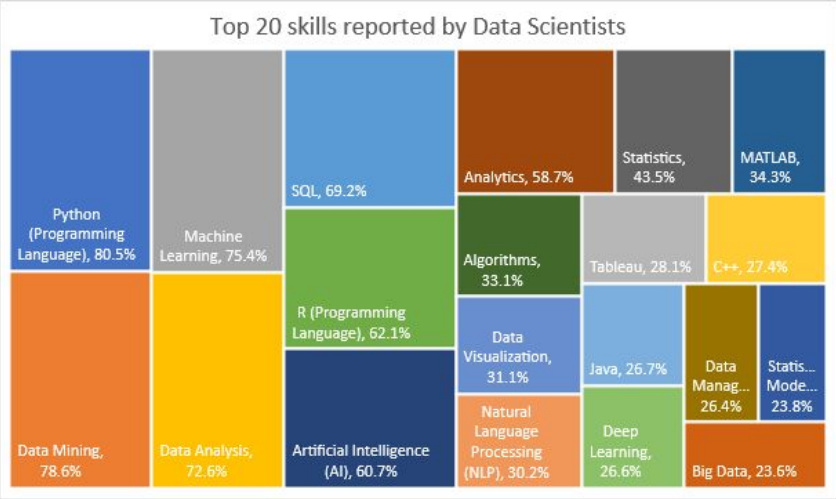
Prerequisite for Data Science

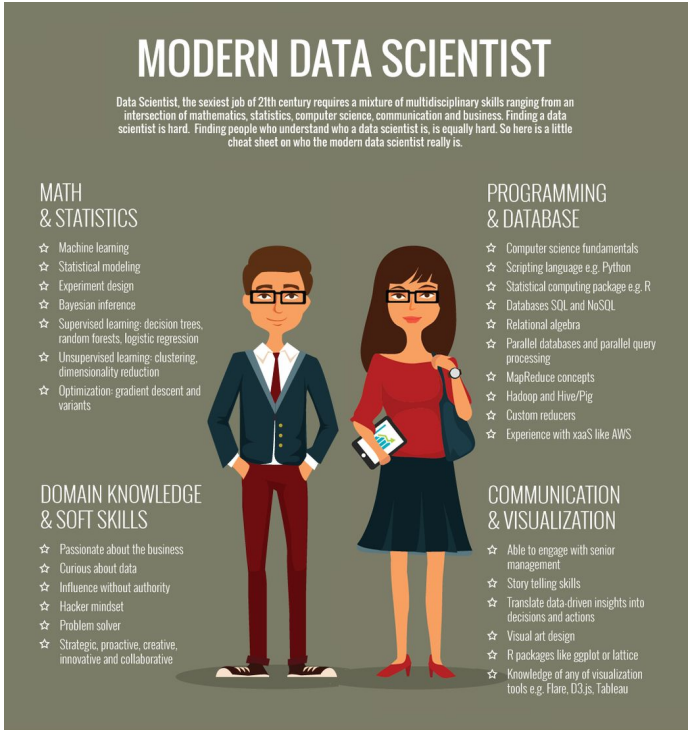
Technical Prerequisite:

- Machine learning: One needs to understand the concept of machine learning To understand data science. Data science uses machine learning algorithms to solve several problems.
- Mathematical modelling: Mathematical modelling is required to make quick mathematical calculations and predictions from the available data.
- Statistics: Basic understanding of statistics is important, such as mean, median, or standard deviation. It is required 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 in-depth understanding of Databases such as SQL is essential for data science to get the data and work with data.

Prerequisite for Data Science

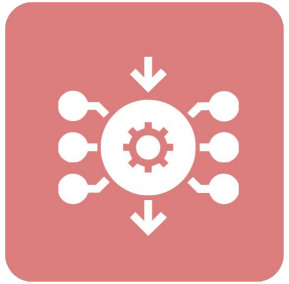
According to one of the studies, the skills of 200 data scientists from leading firms have these common skillset. They are listed as per the percentage of data scientists holding the respective skill.





Five reasons why you must learn data science.

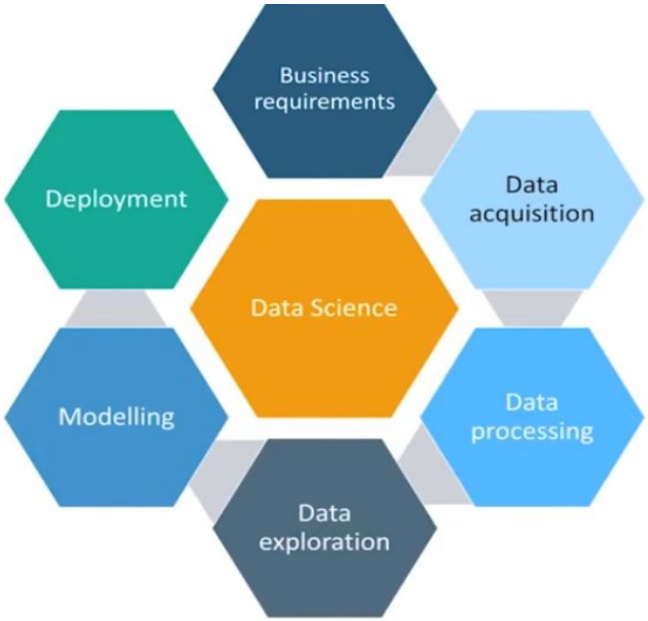
- Great career trajectory with data science
- Significant potential to branch out with different options
- Highest salary takeaway quotient
- Become a decision-maker
- Less competitive because it is a highly analytical role



Data Science Life Cycle

The Data Science project cycle includes the following steps:

1. Understanding Business requirement
2. Data Acquisition
3. Data Processing
4. Data Exploration (Exploratory Data Analysis)
5. Data Modelling
6. Deployment



1. Understanding Business Requirements

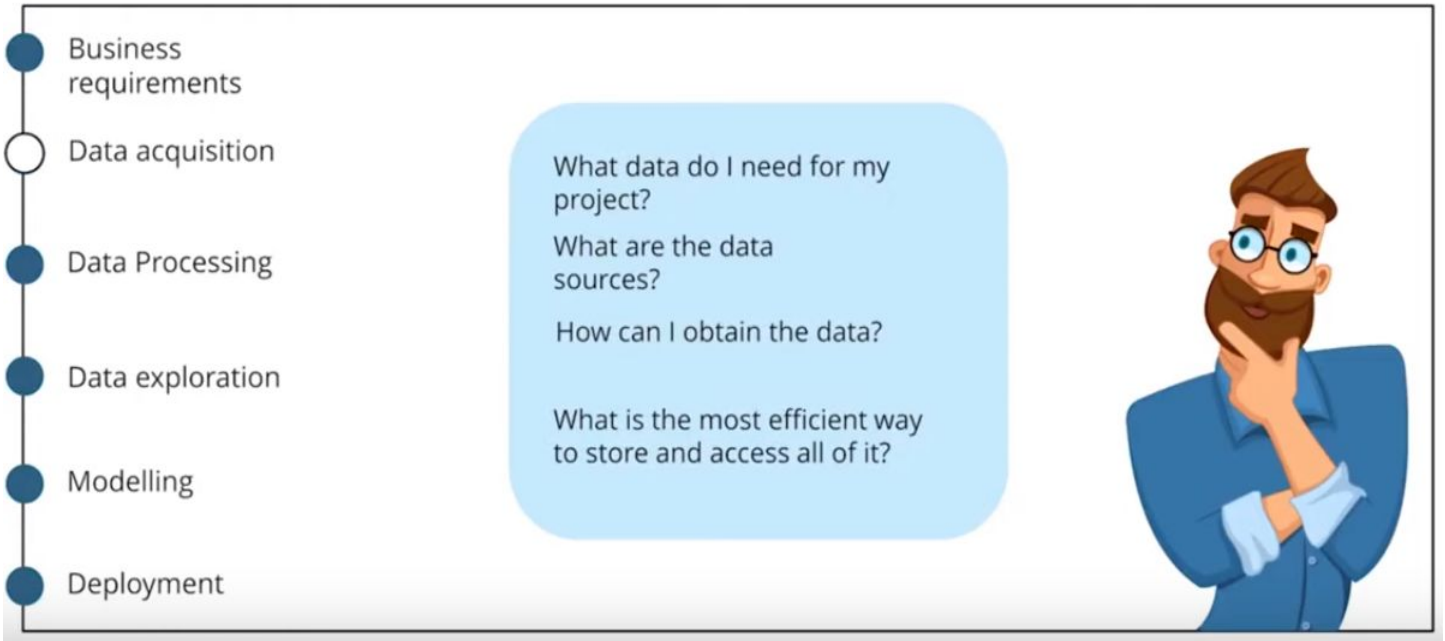
- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Understand the problem

Identify central objectives

Identify variables that need to be predicted

2. Data Acquisition



3. Data Processing

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Transform data into desired format

Data cleaning

- Missing values
- Corrupted data
- Remove unnecessary data

4. Data Exploration

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

understand the patterns in the data

Retrieve useful insight

form hypotheses

5. Data Modelling

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Determine optimal data features for the machine-learning model

Create a model that predicts the target most accurately

Evaluate & test the efficiency of the model

6. Deployment

- Business requirements
- Data acquisition
- Data Processing
- Data exploration
- Modelling
- Deployment

Check the deployment environment for dependency issues

Deploy the model in a pre-production/ test environment

Monitor the performance

Data Science Course Structure

As we are aware now that procuring a career in the data science domain requires knowledge in diverse domains, including mathematics/ statistics, Data Mining, Data Visualization, Programming skills (mainly Python), Data handling (Excels, DBMS, SQL), Machine Learning concepts, NLP, Cloud computing, Text mining etc.

We have a similarly designed data science course structure.

You will now understand everything required to work in the corporate sector in any data science role.

Uses or Applications of Data Science

Internet Search:
Google search uses Data science technology to search for a specific result within a fraction of a second.

Recommendation Systems:
To create a recommendation system. For example, "suggested friends" on Facebook or suggested videos" on YouTube, everything is done with the help of Data Science.

Image & Speech Recognition:
Speech recognises systems like Siri, Google Assistant, Alexa runs on the Data science technique. Moreover, Facebook recognises your friend when you upload a photo with them, with the help of Data Science.

Gaming world:
EA Sports, Sony, Nintendo are using Data science technology that enhances your gaming experience. Games are now developed using the Machine Learning technique. It can update itself when you move to higher levels.

Online Price Comparison:
PriceRunner, Junglee, Shopzilla work on the Data science mechanism. Here, data is fetched from the relevant websites using APIs.

Case-studies of Data Science

1. Data Science in Bio-Tech

The human gene is comprised of four building blocks – A, T, C and G. Our features and traits are determined by the three billion permutations of these four building blocks. While there are genetic defects acquired during lifestyle, their consequences can lead to chronic diseases.
Identifying these defects at an early stage can help doctors and professionals to take preventive measures.

Helix is a genome analysis company that provide customers with their genomic details. Also, various medicines tailored for specific genetic designs have become increasingly popular due to the advent of new computational methodologies.

Due to the data explosion, we can better understand complex genomic sequences and analyze them on a vast scale.
Data Scientists can use modern computing power to handle large datasets and understand genomic sequence patterns to determine the defects and provide insights to physicians and researchers.

Furthermore, using wearable devices, data scientists can use the relationship between genetic characteristics and medical visits to develop a predictive modelling system.

Case-studies of Data Science

2. Data Science in Education

Data Science has also changed how students interact with teachers and assess their performance. Instructors/teachers can use data science to examine the feedback received from the students and use it to enhance their teaching.

Data Science can also be used to create predictive modelling that can predict the drop-out rate of students based on their performance and inform the instructors to take necessary precautions.

IBM analytics has designed a project for schools to assess students based on their performance. Universities use data to avoid retention supplement the performance of their students.

For example, the University of Florida uses IBM Cognos Analytics to keep track of student performance and make necessary predictions.

Also, MOOCs and online education platforms use data science to keep track of the students, automate the assignment evaluation, and better the course based on student feedback.

Case-studies of Data Science

3. Netflix

Impressive customer retention rate - Netflix has a retention rate of more than 90%, which is much more than its prime competitors, such as Hulu's 63% and Amazon Prime's 70%. This retention and addition of users on Netflix can be majorly attributed to Netflix's extensive usage of data analytics.

Netflix can capture the smallest of information of the user; for Eg, At what point in the horror movie did the user pause, what type of content do a user watch, which other users watch similar kind of content and all such information

This helped them to extract multiple profitable insights such as:

- Movie recommendation system
- User segment that overlaps the most with their current user base and hence would have more possibility to get converted into a user
- Demand for content for a particular genre, Netflix can create the show and increase its user base because that show would only be available on Netflix itself.
- Create the right subscription package based on the financial information of its user segment

Case-studies of Data Science

4. Data Science: Case Study Elections:

The Obama campaigns (2008 and 2012) are credited for their successful use of social media and data mining.

Micro-targeting in 2012 –

http://www.theatlantic.com/politics/archive/2012/04/thecreepiness-factor-how-obama-and-romney-are-getting-to-knowyou/255499/

Micro-profiles built from multiple sources accessed by apps, real-time updating data based on door-to-door visits, focused media buys, emails and Facebook messages highly targeted.

One million people installed the Obama Facebook app that gave access to info on “friends”.

Case-studies of Data Science

5. Data Science: Case Study - Customer Analytics

Leveraging customer data to move ever closer to the elusive goal of truly personalized marketing: the right offer, at the right time, in the right location and context, to the right person.

By capturing and analyzing the data from customer touch points within an organization, companies can identify customer pain points and issues proactively and update their customer service FAQs or other communications with existing customers.

Using customer data and analytics, these companies deploy and refine predictive models that help them retain customers with proactive approaches. Investments, in terms of offers and upgrades, can be made at the right time to increase the likelihood of retaining desirable customers.

The experience that customers have with companies matters a great deal. Other recent research has highlighted the critical connection between experience and company financial performance.

THANK YOU