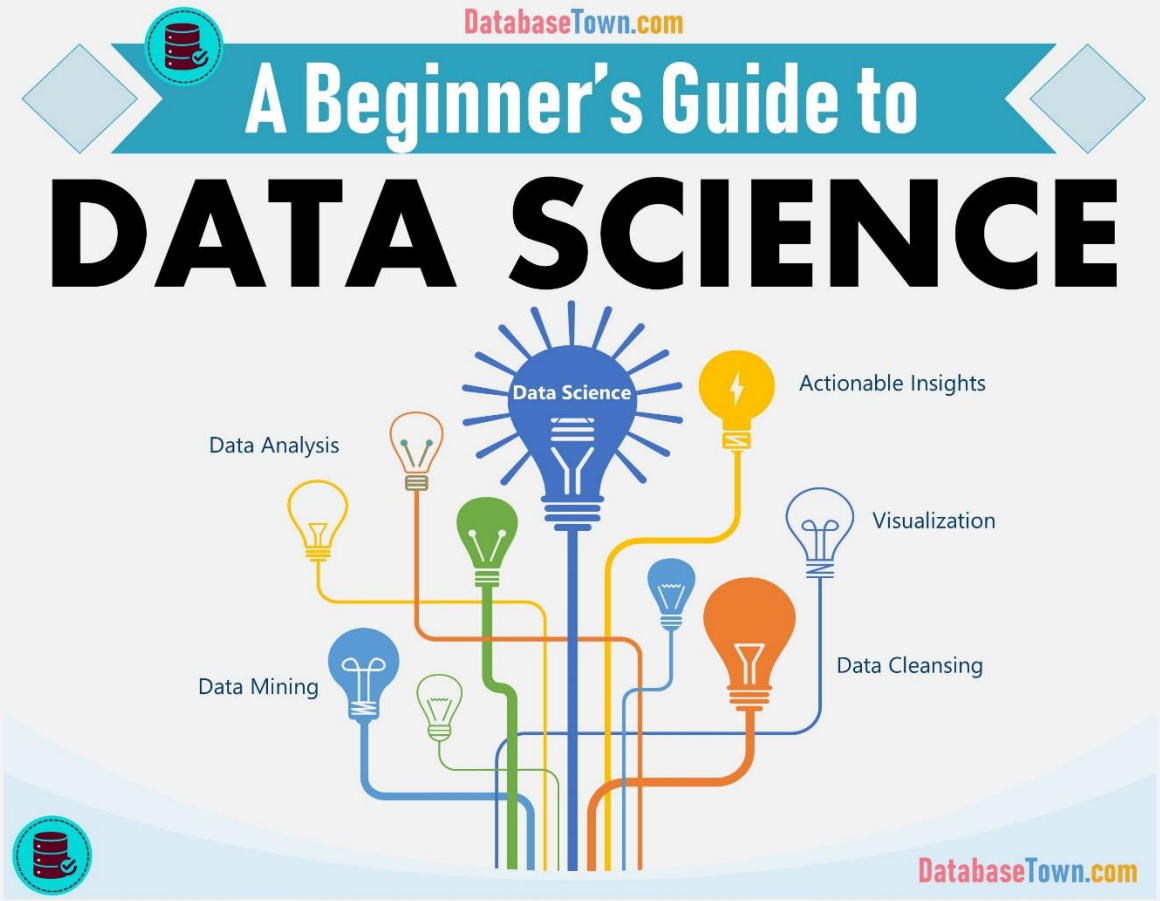
Introduction to Data Science
A Beginner's Guide

In present era, the term Data Science is frequently being used but every book lover or student must want to get the answers of these elementary questions;

- *What is Data Science?*
- *Applications of Data Science*
- *Historical Background of Data Science*
- *Basic Components of Data Science*
- *How Data Science Work?*
- *Main processes of Data Science*
- *Famous data Science tools*
- *Real life usage of Data Science Systems*
- *Top industries players of Data Science*
- *Find your career in Data Science*
- *What challenges are being faced by Data Science?*
- *Helpful Information about Data Science*
    - Data Analysis Vs. Data Analytic
    - Qualitative Analysis Vs. Quantitative Analysis
- *Frequently asked questions about Data Science*

If you are passionate to know the answers of these queries then I assure you that you are at a right place where you will get the valuable insights of these questions, so read this beginner's guide....

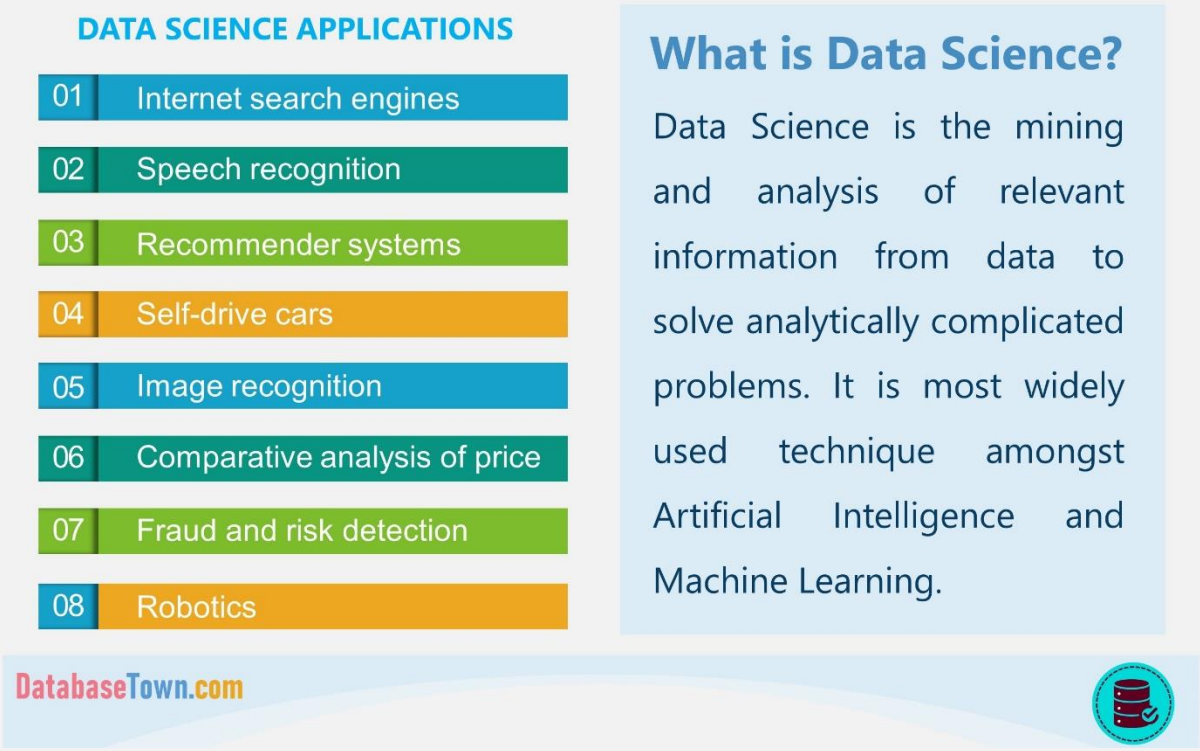


What is Data Science?

Data Science deals with the processes of data mining, cleansing, analysis, visualization, and actionable insight generation. Data Scientist must have the basic knowledge of mathematics, computer programming and statistics to solve the complex data problems in an efficient way to boost the business revenue.

Data Science is the mining and analysis of relevant information from data to solve analytically complicated problems. It is most widely used technique amongst Artificial Intelligence and Machine Learning Engineers. For example, when you logged on any e-commerce website and browsed some categories and products before purchase, you are generating data, which will be helpful for Analysts to know your behavior about purchase.

Data science is about using already stored raw and unstructured data in organization's repository, which process through systematic, programming and business skills in creative ways to generate business worth. Data science keeps on developing as a standout amongst the most encouraging and demanding future career-ways for talented students. Now, experts comprehend that they should progress past the customary abilities of analyzing big data, data mining, and programming expertise. Therefore, there is a dire need for a data scientist to get a full grasp on data science life cycle.



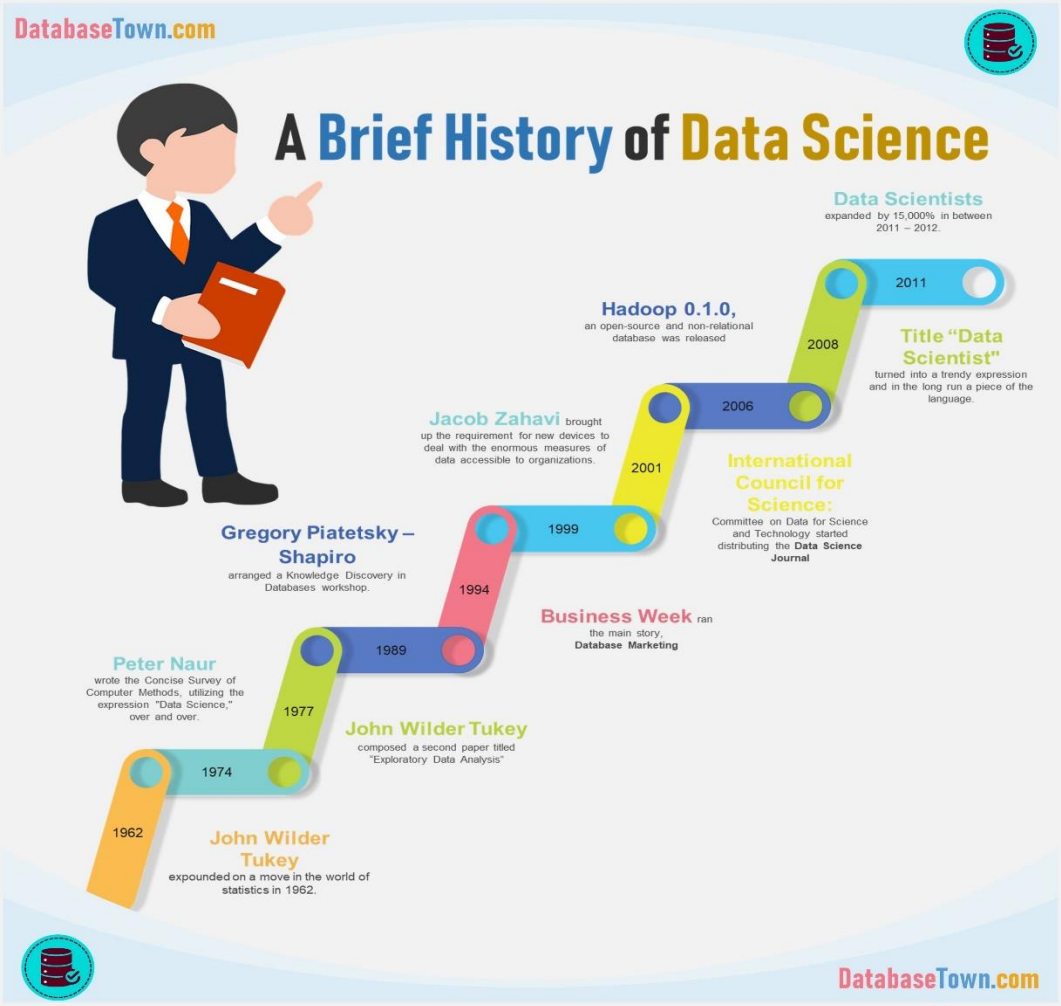
Applications of data science

Presently application of data science is very vast. You can see it everywhere in your daily life. Some prominent examples are given here.

- Internet Search Engines
- Speech Recognition

- Recommender Systems (YouTube, Netflix, Amazon)
- Self-driving Cars
- Image Recognition
- Comparative analysis of Price
- Fraud and risk detection
- Gaming
- Robotics
- Airline route planning

Historical Background of Data Science



History of data goes back to 1500s when the Latin originated word "datum" was used. But the work started on it during the period from 1940 to 1950. Claude Elwood Shannon, an American Mathematical Engineer published a paper "A Mathematical Theory of Communication" in 1948. Although he was not a data scientist but his information theory formed the basis of machine learning algorithms.

John Wilder Tukey wrote a book **Exploratory Data Analysis** in 1977. The concept of Exploratory Data Analysis was promoted by him to explore the data. The exploratory data analysis (EDA) technique is used to analyze datasets mainly with the visual methods.

Peter Naur wrote the Concise Survey of Computer Methods in 1974 where he utilized the expression "Data Science" first time. He used this term repeatedly in his book.

In 1999, **Jacob Zahavi** brought up the requirement for new devices to deal with the enormous measures of data accessible to organizations, in “Mining Data for Nuggets of Knowledge”.

In 2001, **William Cleveland** published a paper, “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics”. You can find the paper here.

The International Council for Science: Committee on Data for Science and Technology started distributing the ***Data Science Journal*** in 2001, concentrated on issues like the portrayal of data systems, their production on the web, applications and legitimate issues.

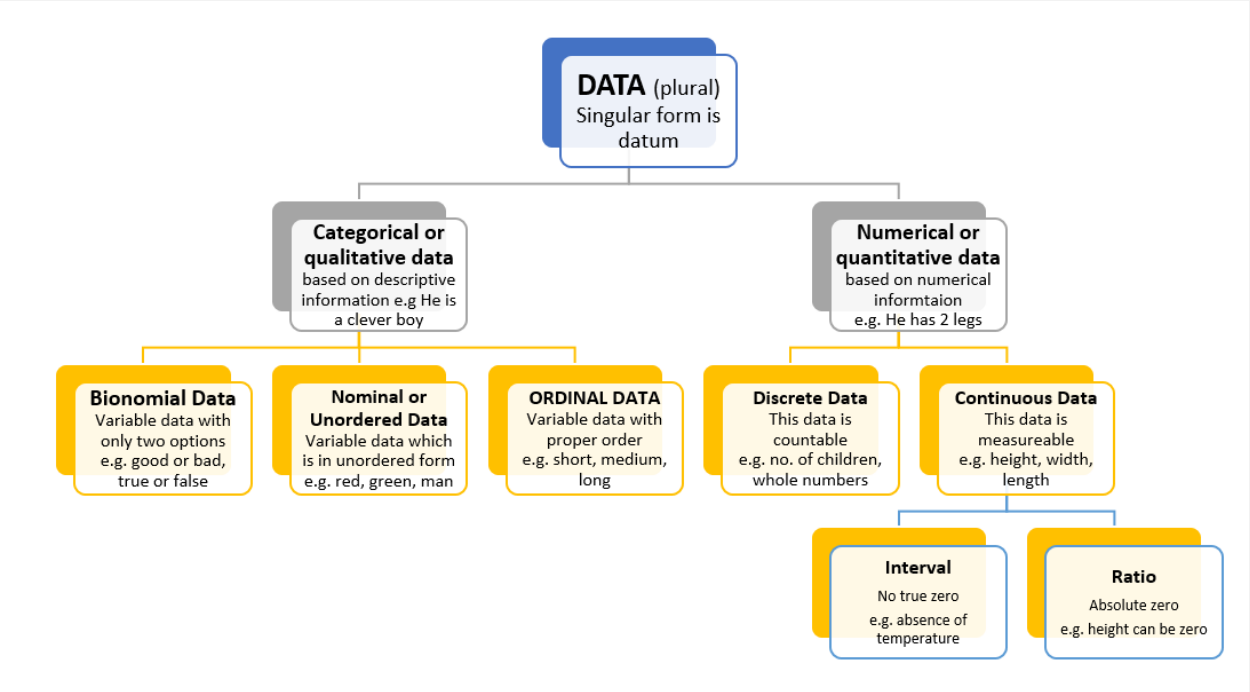
In 2008, the title, "**Data Scientist**" turned into a trendy expression and in the long run a piece of the language. Jeff Hammerbacher and DJ Patil of Facebook and LinkedIn are given acknowledgment for starting its utilization as a trendy expression. **Johan Oskarsson** was reintroduced the term NoSQL

in 2009 when he sorted out a dialog on "open-source, non-relational databases".

Basic Components of Data Science

Data

Data is a very basic element of data science. There are different types of data. This picture shows you what are different kinds.



Data is divided into categorical or qualitative data and numerical or quantitative data.

Categorical or qualitative data is based on descriptive information e.g. He is a cleaver boy. It has further three types:-

- **Bionomial Data** (Variable data with only two options e.g. good or bad, true or false)
- **Nominal or Unordered Data** (Variable data which is in unordered form e.g. red, green, man)
- **Ordinal Data** (Variable data with proper order e.g. short, medium, long)

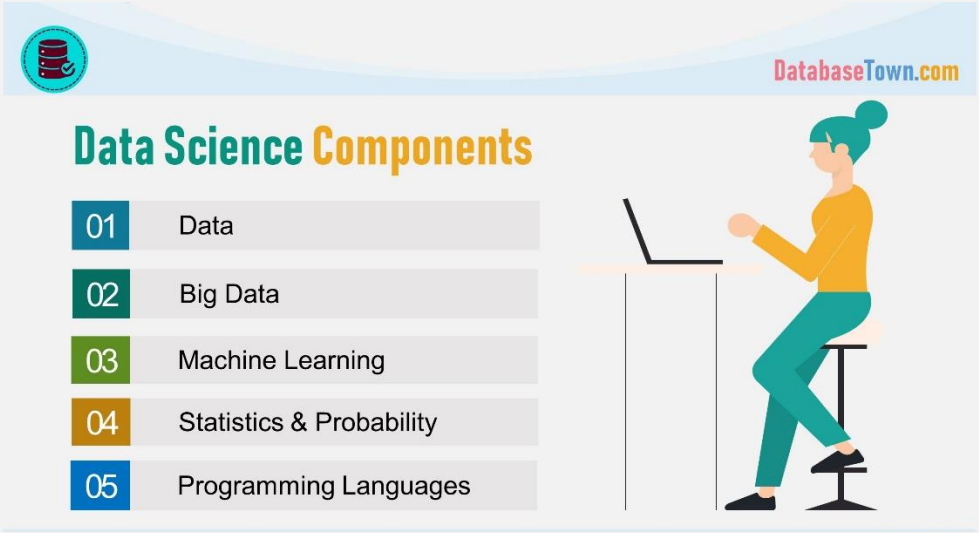
Numerical or quantitative data is based on numerical information e.g. He has 2 legs. It is further divided into:

- **Discrete data** (This data is countable e.g. no. of children, whole numbers) and
- **Continuous data** (This data is measurable e.g. height, width, length). Continuous data has further two types.
    - **Interval** (No true zero e.g. absence of temperature)
    - **Ratio** (Absolute zero e.g. height can be zero)

Big Data

Big data consists of huge data sets. These data sets are analyzed and visualized to unveil the trends, human behavior, and interactions.

The great example of big data is social media site Facebook where hundreds of terabytes data is added daily in the form of text, audio, video, images etc.



Machine Learning

Machine Learning is a part of **Data Science** that enables the system to process data sets without any human interference (autonomously). It utilizes different algorithms to work on massive volume of data generated from various sources and makes prediction, analysis patterns and gives recommendations. The real life example of Machine learning is its use in fraud detection and client retention.

Machine learning has three types.

- **Supervised machine learning** (labeled data sets are used, here input and output variables are used to produce outcome)
- **Unspervised machine learning** (un-labeled data sets are used, here only input variables are used and no output variable is used)
- **Reinforcement learning** (It is different from supervised machine learning. It is about taking appropriate action in particular situation to maximize the reward.)

Statistics and Probability

Statistics and Probability are assumed essential elements in data science as they make the numerical foundation of data science and likelihood. It is difficult to do data science without the basic knowledge of statistics and probability.

Programming Languages

Programming languages specially **Python** and **R** play vital role in data organization, visualization and data investigation. Python is high level programming language which provides free libraries for data analysis. It is popular amongst the data scientists.

R is another popular language. The best feature of R is data visualization. This language is mostly used for social media post analysis.

There are another languages that provide support for data science like Java 8 with Lambdas and Scala. SQL is used for structured data and NoSQL for unstructured data.

How Data Science Work?

Data science integrates devices from multi disciplines to accumulate a data set, process and get experiences from the data collection, obtain requisite information from the set, and decipher it for basic leadership purposes.

Data science field incorporates statistics, data mining, Artificial Intelligence, programming and analytics.

Data mining applies algorithm in the perplexing informational collection to uncover designs that are then used to separate usable and pertinent information from the set.

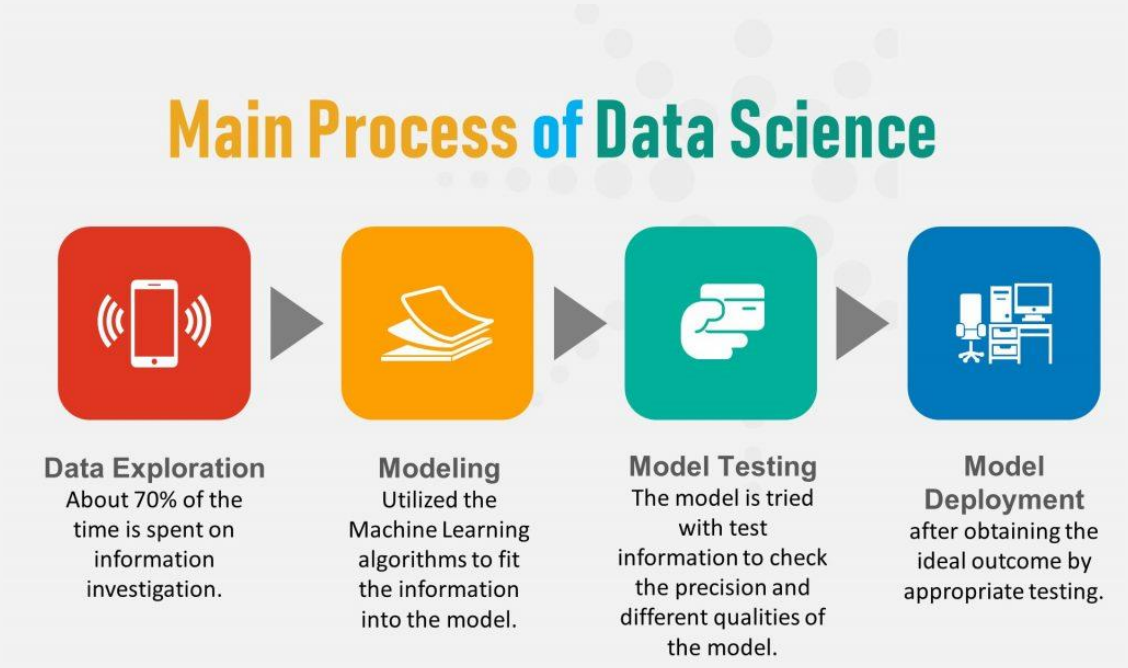
Factual estimates like prescient examination use this separated information to check occasions that are probably going to occur later on dependent on what the information indicates occurred before.

Artificial Intelligence is a man-made reasoning instrument that forms mass amounts of information that a human would not be able to process in whole life.

The data examiner gathers and processes the organized information from the AI by using various algorithms under analytics. Data analyst translates, changes over and summarizes the data to a understandable language that the basic leadership group can understand it easily.

Main processes of Data Science

Main processes of Data Science are as follows:



Data Exploration

It is an essential step as it consumes most amount of time span. About 70% of the time is spent on information investigation.

The principle element for data science is information, so, when we get information, it is only from time to time that information is in a right organized structure.

There is a ton of commotion present in the information which means a great amount of undesirable information that isn't required. So what we do in this progression? This progression includes examining and change of information in which we check the perceptions (lines) and highlights (segments) and expel the commotion by utilizing measurable techniques.

This progression is likewise used to check the relationship among different features(columns) in the informational index, by the relationship we mean whether the features(columns) are subject to one another or autonomous of one another, regardless of whether there are missing qualities in the information or not. So, essentially the information is changed and prepared for further use.

Modeling:

At this point, our information is arranged and prepared to go ahead. This is the second step where we really utilized the Machine Learning algorithms to fit the information into the model.

The determination of a model relies upon the sort of information we have and the business prerequisite. For instance, the model choice for prescribing an article to a client will be not quite the same as the model required for

foreseeing the quantity of articles that will be sold on a specific day. When the model is chosen, we fit the information into the model.

Model Testing:

Model deployment is the subsequent stage and critical for the execution of the model. The model is tried with test information to check the precision and different qualities of the model and roll out the required improvements in the model to get the ideal outcome.

In the event that we don't get the ideal precision we can again go to previous Step-II i.e. modeling, select an alternate model and afterward rehash a similar Step-III i.e. model testing and pick the model which gives the best outcome according to the business necessity.

Model deployment:

When we obtain the ideal outcome by appropriate testing according to the business prerequisites, we conclude the model, which gives us the best outcome according to testing results and send the model in the manufacturing location.

Famous data Science tools

The main purpose of using data science tools is to avoid the programming aspect and provide user-friendly GUI. So, a person with less knowledge of algorithms can easily use them in building machine learning models.

Some famous tools are discussed here.

RapidMiner

RapidMiner is a gadget for the complete life-cycle of forecast modeling. In excess of 300,000 clients in more than 150 nations use RapidMiner items to drive income, diminish costs, and keep away from dangers. They make information progressively gainful through an extremely quick platform that binds together information prep, AI, and model deployment.

They made-up their platform on three noteworthy parts, RapidMiner Studio is the Visual Workflow Designer for Data Science Teams. It is a platform with code-discretionary with guided examination with in excess of 1500 capacity, it enables clients to mechanize predefined associations, worked in formats, and repeatable work processes. RapidMiner Server share and works together on each progression and part of the information mining process. It permits to upgrade with the progressed lining instrument: RapidMiner Server can cut out assets and devote to groups, use cases or ventures. The platform makes it conceivable to get perceive-ability into information science collaboration and administration. RapidMiner Radoop expels the multifaceted nature of information prep and AI on Hadoop and Spark. The platform is utilized in numerous enterprises with various sorts of arrangements.

You can get the RapidMiner here.

Data Robot:

It is the platform for automated Machine Learning that can be utilized by data scientists, software engineers, IT professionals and executives. Data Robot has Python SDK and APIs. It ensure an easy development process and parallel processing.

Data Robot site.

Apache Hadoop:

It is a java based open source framework which can perform distributed processing of immense data sets across computer clusters. Apache Hadoop runs in parallel on a cluster, so, it has the capability to permit you to process data across all the nodes. It has many modules, such as HDFS, Hadoop Map Reduce, Hadoop Common, Hadoop YARN, Hadoop Ozone. HDFS splits immense data and allocate across many nodes in a cluster to ensure high accessibility.

Download link

Matlab

Matlab is available for personal use as well as for students which provides you the solution for evaluating data, developing algorithms and producing models. Matlab is also utilized for wireless communications and data analytics. The best ability of Matlab is its scalability. Its algorithms can easily be converted to HDL, CUDA & C/C++ code.

Download link

KNIME

KNIME is free and open source platform which is helpful for data scientists in blending tools and data types. It also permit to utilize your desire's gadgets and expend to Apache Spark and Big Data. KNIME can easily work with various data sources and various types of platforms.

Download link

Trifacta

Basically, it has 03 pricing plans such as Wrangler, Wrangler Pro and Wrangler Enterprise but their main product is Wrangler that is helpful in sightseeing, converting, scrubbing and joining the desktop files together. You just import your datasets to Wrangler and the application will automatically start to shape and structure your data. Its algorithms help you to make your data by telling common changes and accumulations. Its advance self-service platform for data training is Trifacta Wrangler Pro and Trifacta Enterprise is more helpful for the predictor staff.

Alteryx

Alteryx provide end-to-end analytics platform which permits the data scientists and business analysists to break data hurdles and bring game-changing insights which are helpful in solving big corporate hitches. Alteryx determine the data and collaborate across the group. It has the ability to make and investigate the model. It also permitted you to implant Python, R and Alteryx models into your processes.

Excel

Microsoft Excel can be utilized as data science tool as it is easier and best analyzing data tool for non-professional people. You can easily organize, sort, filter and summarize data with the help of Microsoft Excel.

Tableau

Tableau can be utilized by anyone due to its drag and drop functionality. In some basic versions, data visualization tool is free of cost. It can work with any database and also support various format data, such as, xml, csv, xls, etc.

Kubernates

It is an open source tool for handling clusters of containers. It provides combination of features which are very helpful for data scientists. Kubernates provides tools for installing applications, variations to existing container type applications, scaling those applications and help in enhancing the usage of the existing hardware under your containers.

Cloud Dataflow

Cloud Dataflow is a best gadget for data scientist as it offers fully managed environment that can easily measure the massive data sets and enables data science crews to own more of the creation process. It exposes transformational use cases across businesses, including:

- Point-of-sale and segmentation analysis in marketing.
    - Fraud exposure in economic facilities
    - Personalized user experience in gaming
    - IoT analytics in healthcare, logistics and engineering

Real life usage of Data Science Systems

Amazon.com, Inc., a multinational e-commerce company used the following data science systems for business optimization:-

- Selection of ideal routes, plans, and products groupings to reduce the delivering cost.
- Selection of warehouses to minimize the distribution cost.
- Best traffic prediction in order to curtail the time spent by drivers in traffic jams.

- Product's price may fluctuate which requires it proper monitoring, so, these systems are helpful to place the products in appropriate categories.
- Detect system interferences and hacking attempts to prevent from stealing personal and confidential data.
- Appropriate algorithms are used in these systems to detect the fake reviews.
- Products are properly categories by utilizing tagging and indexing algorithms.
- Provide efficient search engine technology to its customers.
- Also provide multivariate testing and detect artificial sale.
- Provide competitive real-time analysis.
- Increase marketing and advertising efficiency by providing correct customer segmentation.
- Provide advertising optimization and inventory forecasting
- Check market trend and sales forecasting.
- Detect employees at risk of committing fraud and leaving jobs without any intimation.
- Optimize redundancy with budget constraints, generate email alert and prioritize the messages automatically.
- Maximizing profit with reduce cost on publisher, author, vendor, etc.
- Advertisement relevancy algorithm to select and rank Ads to be exhibited on a particular webpage to fascinate the invitees and to grow the profit.

Top industries players of Data Science

IT companies have the dire need to address their complex and extending information situations, so, as to recognize new esteem sources, to use future chances, and to develop or advance proficiently. The following absolute

greatest and best organizations that are employing information researchers at top rate salaries:

Google – Google hire best data scientists from all over the world and offers the absolute best data science pay rates.

Amazon – Amazon is a worldwide online business and distributed computing mammoth that is contracting data scientists on a major scale. They hire data scientist to get some answers concerning the client mentality, upgrade the geographical contact of both the web based business area and cloud space among different business-driven objectives.

VISA – It is online money related portal for the majority of the organizations and Visa does exchanges in the scope of several millions throughout a day. Because of this, the necessity for data scientists is colossal at Visa to create more income, check false exchanges, and alter the items and administrations according to the client prerequisites.

Find your career in Data Science

Data Architect:

Data Architect is responsible for building and maintaining an organization's database with the assistance of database administrators and analysts. They create database solutions, evaluate requirements and plan design reports.

Data Engineer:

Data Engineer is responsible for real-time processing on stored or collected business data, so that, the data could be ready for analysis by data scientists.

Database Administrator:

Database Administrator utilized various software tools to store and organize conventional data for further examination.

Data Scientist:

Data Scientists used conventional statistical methods or machine learning techniques for making strategic business decisions.

Data Analyst:

Data analysts perform advance types of analysis for companies and they may also be responsible for tracking web analytics and analyzing A/B testing.

Data Visualizer:

They translate data analytics into clear and concise information for business communication.

Machine Learning Scientist:

Machine Learning Scientist explores new data approaches and algorithms.

Machine Learning Engineer:

Machine Learning Engineer applies state of the art computational models and delivers software solutions.

Statistician:

Statisticians must have the solid knowledge of statistics and probability. They are responsible for analyzing and report statistical information for business point of view.

Business Intelligence Analyst:

They mainly focus on analyzing market trends.

Business Intelligence Consultant:

Business Intelligence Consultants provide their expertise in designing, developing and implementation of BI and analytics systems. They also examine the business feat and prepare reports on performance metrics.

Business Intelligence Developer:

They are responsible for designing and developing strategies to support business consumer in rapidly searching the requisite information for better business assessments.

What challenges are being faced by Data Science?

In spite of the guarantee of data science and tremendous interests in data science groups, numerous organizations do not understand the full estimation of their information. Without progressively taught, focal administration, officials probably won't see a full profit for their ventures. This riotous condition presents numerous difficulties.

Data Researchers can't perform excellently

Since access to information must be allowed by an IT head, data researchers frequently have long waits for information and the assets they have to analyze it properly. When they approach, the data science group may analyze the information by utilizing extraordinary and perhaps inconsistent instruments. For instance, a researcher may build up a model utilizing the R language; however, the application it will be utilized in is written in an alternate language. This is the reason it can take weeks, or even months, to arrange the models into helpful applications.

Business Administrators are excessively expelled from data science

Data science work processes are not constantly coordinated into business basic leadership procedures and frameworks, making it troublesome for business chiefs to team up knowledgably with data researchers. Without better joining, Business Administrators think that it's hard to comprehend why it takes such a long time to go from model to creation.

Application Engineers can't access to usable AI

Once in a while, the AI models that designers get must be recoded or are not prepared to be sent in applications. Also, in light of the fact that passages can be resolute, models can't be conveyed in all situations and versatility is left to the Application Designer. **IT Directors invest a lot of energy in help –** Due to the spread of open source instruments, IT has a regularly developing apparatuses to help. A Data Researcher in selling, for instance, may utilize various instruments in comparison to a Data Researcher in economics. Groups may likewise have distinctive work processes, which implies IT should consistently modify and refresh conditions.

Helpful Information about Data Science

See the difference between the similar terminologies frequently used in data science. Difference between data analysis and data analytics & difference between qualitative analysis and quantitative analysis.

Data Analysis Vs. Data Analytics

Data Analysis	Data Analytics
Data definition, cleaning, investigation and transformation into meaningful results	Data collection & its inspection
Used in businesses to analyze data and extract useful insights from the data	Used in businesses to make verdicts from data which are data-driven
Used to perform predictive analysis, descriptive analysis, exploratory analysis, inferential analysis	Used to find market trends, customer preferences, masked patterns, anonymous correlations
RapidMiner, KNIME, Google Fusion Tables, Tableau Public, NodeXL, WolframAlpha are utilized for data analysis.	For data analytical purpose, Python, SAS, R, Tableau Public, Apache Spark, Excel are frequently used

Qualitative Analysis Vs Quantitative Analysis

Qualitative Analysis	Quantitative Analysis
Subjective analysis due to absence of statistical data	Objective analysis due to presence of statistical data
Classification of data on the bases of attributes and properties such as, color, gender, etc.	Classification of data based on measureable quantities like volume, weight, length, density, etc.
Small data collection	Large data collection
Methodology of qualitative analysis is investigative	Methodology of quantitative analysis is decisive
Result are particular to the objects being examined	Obtained results can be applicable on the general population
Indefinite questions, observations and interviews are conducted by researches	Measurements, surveys, observation and experiments are made by researchers
Qualitative analysis is performed to obtain profound knowledge about occurrence of certain objects	Quantitative analysis is performed to test hypotheses and furnish the future forecast

Frequently Asked Questions about Data Science

Who can learn data science?

The person who is passionate about data science can learn it. There are some prerequisites like mathematics, statistics, programming knowledge, use of data science tools like RapidMiner and SQL/NoSQL knowledge.

If you are passionate to dive into the vast sea of data science and ready learn all the necessary elements and tools, then data science is right choice for you.

Is Data Science hard or easy to learn?

To become a data scientist is not an easy task. You have invest your energies and time to become data scientist. You may have learned online that the data science and easy to learn but it is not the actual fact.

To learn data science you have to understand the statistics and mathematical concepts, programming languages (like python and R) to organize and visualize the data. To understand the concepts of machine learning, SQL for structured data and NoSQL for unstructured data.

If you analyse all the above mentioned topics with cool mind, you will come to the point that it is not an uphill and straight forward task. You have to build real-world models to prove yourself a data scientist.

To understand the all processes from building models to testing and deployment requires a lot of work. We can say that learning data science is hard.

How do I get a data science job?

Before getting data science job, you have to learn the skills that are necessary for a data scientist. There are different types of jobs i.e. Data Analyst, Machine Learning Engineer, Data Engineer, Data Scientist and many more.

You have to build specific skill set for the job. You can go to Glassdoor and search data science jobs and see what kind of job description is required by the companies. Prepare yourself for the interview as per job description.

Can I become a data scientist without a degree?

You may know that there are many examples that people without a degree created their space in tech. A great example is Bill Gates.

However, you can learn data science without a degree but the degree has its own value. It helps in getting job and relevant position in a company.

To be a data scientist, you need to learn statistics and mathematics for analysis purpose, a good programming skills and some knowledge of business. You should also learn how to build the projects based on real-life examples.

To be a good data scientist, you have to prove data science skills attained through online resources, through a degree or by learning at your own.

What is data scientist salary?

According to glassdoor average salary of Data Scientist in United States is $117,345 per year while it ranges from $86,000 to $157,000 per year depending upon the expertise, experience and skills as well as job nature. Salaries of other data science related posts differ due to different nature of job for each post.

Average year salaries of different posts are

- **Senior Data Scientist:** $137,000
- **Data Analyst:** $67,000
- **Quantitative Analyst:** $116,000
- **Data Engineer:** $117,000
- **Machine Learning Engineer:** $121,000

Conclusion

In modern era, organizations are taking assistance in their decision-making processes with data science technologies like machine learning, measure the aftereffects of the choices they make to realize what worked paramount and boost their revenue and market position.

Advanced machine learning models are being utilized that provide valuable insights into the composed data. Data Science is a wide-ranging field as it does not depend only on algorithms and statistics but pay attention on whole data processing methods, whereas, machine learning is a part of data science.

Data Science utilizes machine learning to analyze the data and formulate possible predictions. In order to get the optimum outcomes, some data scientists, use data science technologies combine with other disciplines, such as cloud computing and big data analytics.

Author:

Tariq Aziz Rao, MS(CS), Data Scientist

Co-Author:

Muhammad Zubair Akhtar, Master of Computer Science