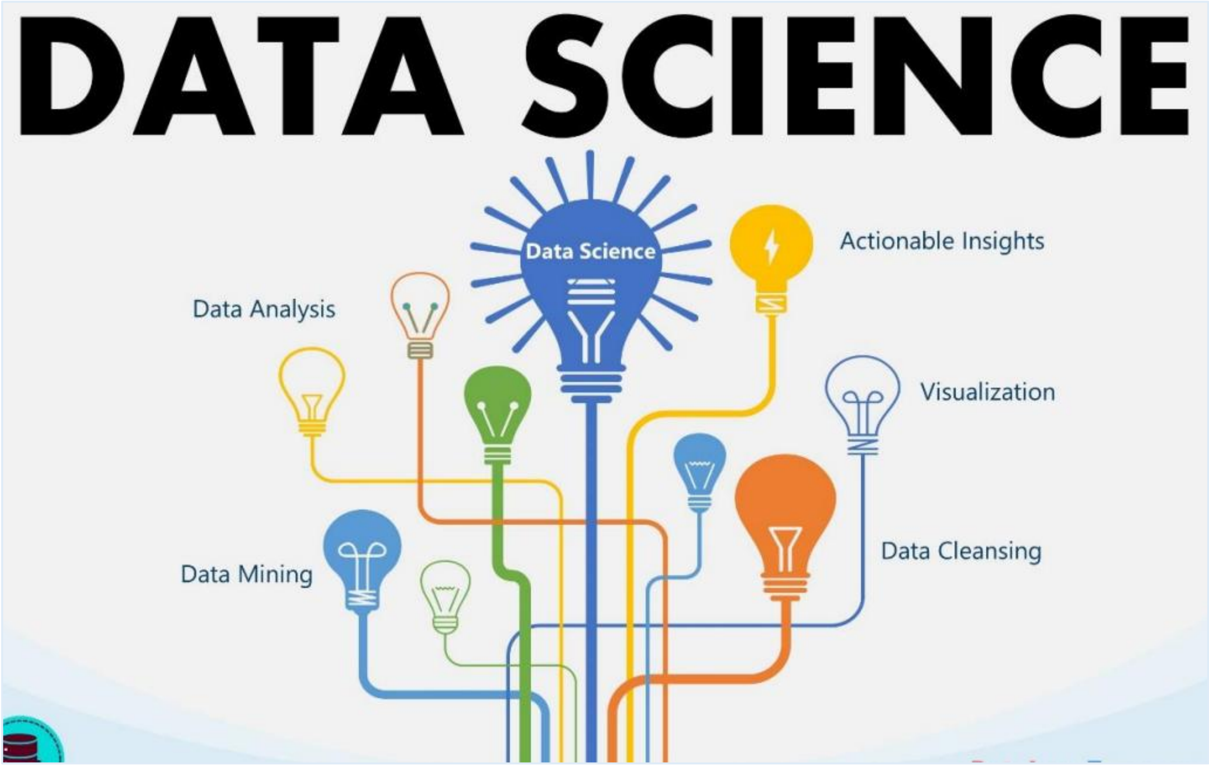
Introduction of Data Science



What is Data Science?

Data Science is the application of a scientific methodology to the study of data for the purpose of extracting insights and making predictions with trained models.

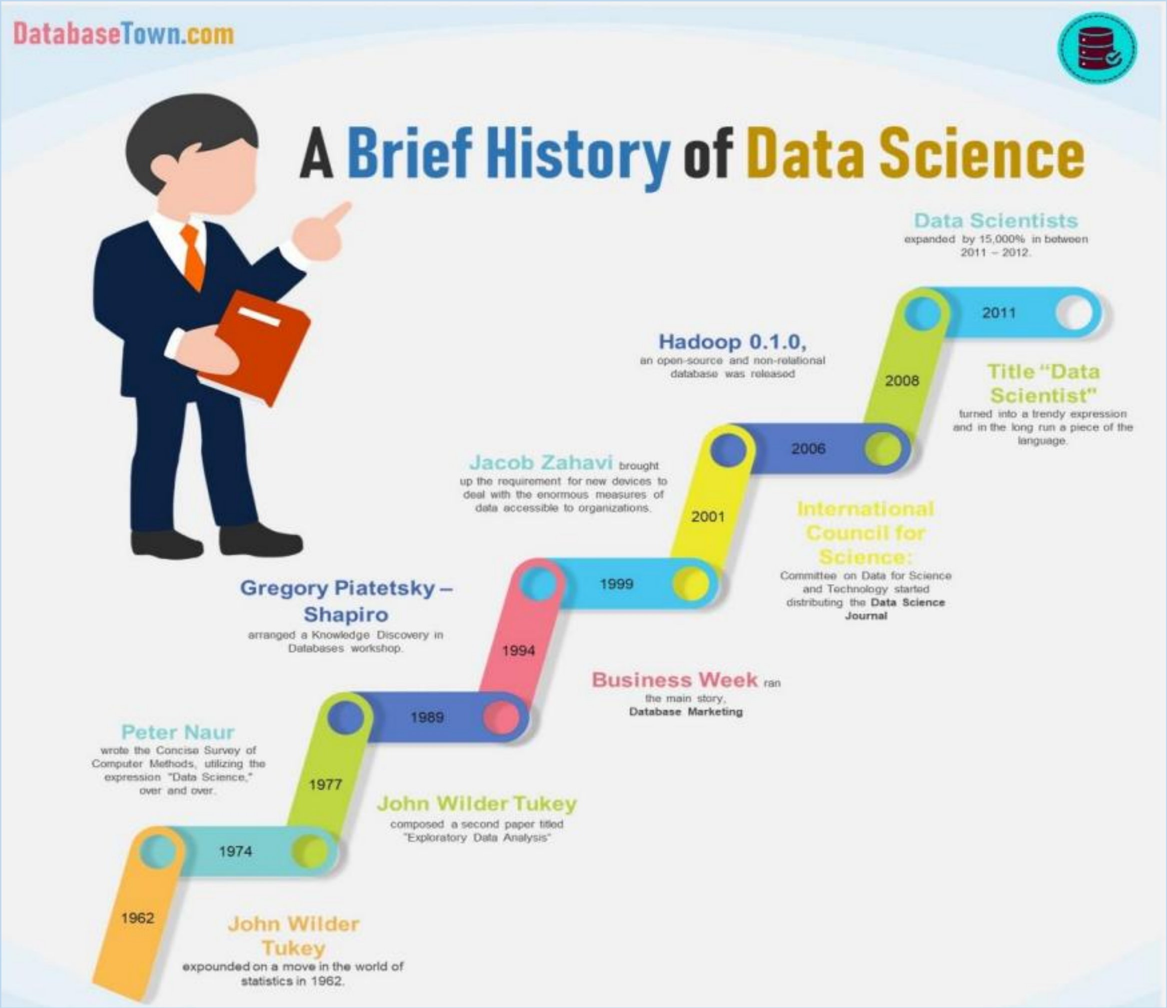
Data Science deals with the processes of data mining, cleansing, analysis, visualization, and actionable insight generation. Data Scientist must have the basic knowledge of mathematics, computer programming and statistics to solve the complex data problems in an efficient way to boost the business revenue.

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 behaviour 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.

As a Data Science professional, you would understand the business goals for conducting the data science work and review the business data landscape. You will devise the plan for collecting business data, prepare and clean this data, and perform its detailed exploration and analysis. Once data validity and reliability are established, work on building the models, their testing, optimization, and evaluation. Finally, deploy the model for use with the newly available data.

Historical Background of Data Science



Data Scientists expanded by 15,000% in between 2011 – 2012.

Hadoop 0.1.0, an open-source and non-relational database was released

Title "Data Scientist" turned into a trendy expression and in the long run a piece of the language.

Jacob Zahavi brought up the requirement for new devices to deal with the enormous measures of data accessible to organizations.

International Council for Science: Committee on Data for Science and Technology started distributing the **Data Science Journal**

Gregory Piatetsky – Shapiro arranged a Knowledge Discovery in Databases workshop.

Business Week ran the main story, **Database Marketing**

Peter Naur wrote the Concise Survey of Computer Methods, utilizing the expression "Data Science," over and over.

John Wilder Tukey composed a second paper titled "Exploratory Data Analysis"

John Wilder Tukey expounded on a move in the world of statistics in 1962.

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.

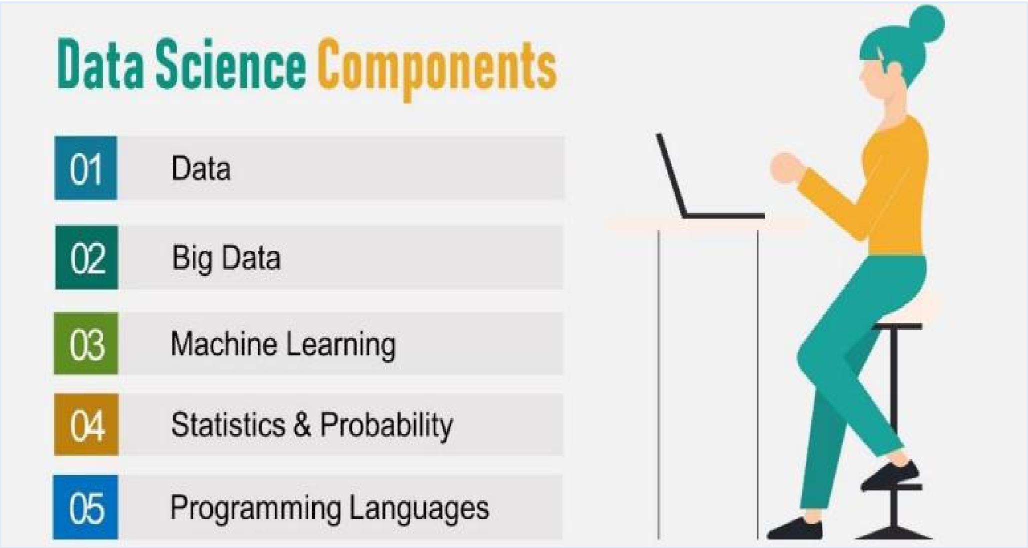
In 1999, Jacob Zahavi brought up the requirement for new devices to deal with the enormous measures of data accessible to organizations, in “Mining Data for Nuggets of Knowledge”.

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.

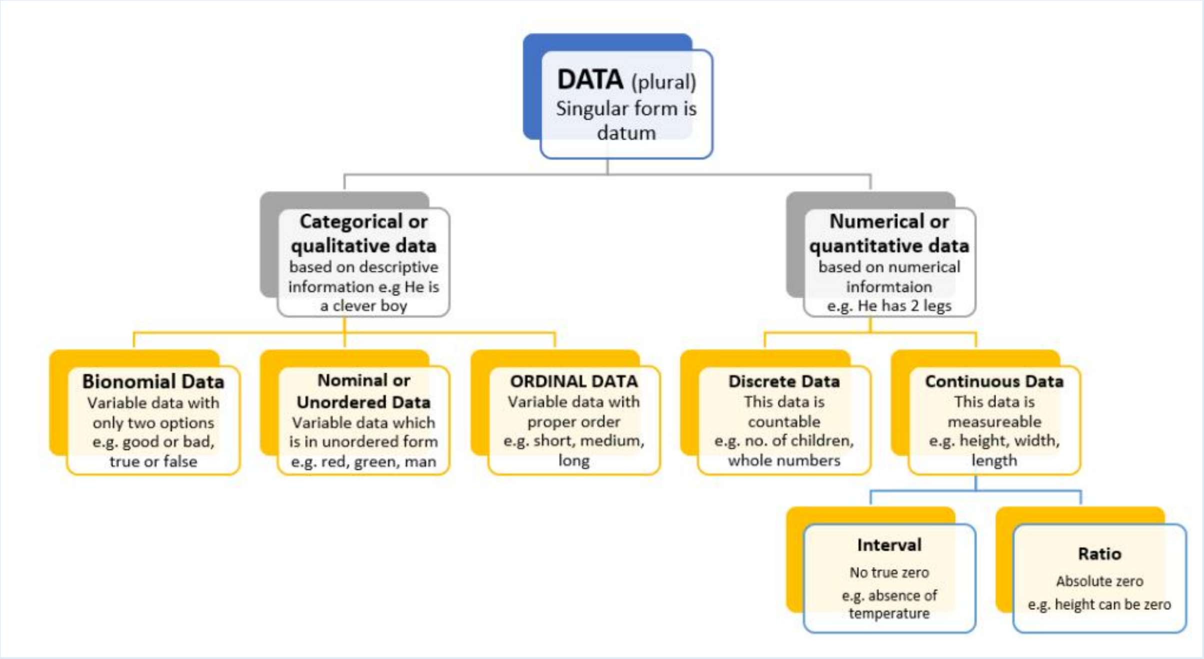
In 2008, the title, "Data Scientist" turned into a trendy expression and in the long run a piece of the language. Jeff Hammerbacher and DJ Patil of Facebook and LinkedIn are given acknowledgment for starting its utilization as a trendy expression. Johan Oskarsson was reintroduced the term NoSQL in 2009 when he sorted out a dialog on "open-source, non-relational databases".

Basic Components of Data Science



1. Data

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


DATA (plural)
Singular form is datum
Categorical or qualitative data
based on descriptive information e.g He is a clever boy
Numerical or quantitative data
based on numerical informtaion e.g. He has 2 legs
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
Discrete Data
This data is countable e.g. no. of children, whole numbers
Continuous Data
This data is measureable e.g. height, width, length
Interval
No true zero
e.g. absence of temperature
Ratio
Absolute zero
e.g. height can be zero

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:-

- Binomial 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)

2. Big Data

Big data refers to extremely large and complex datasets that are beyond the capacity of traditional data processing tools and methods to handle efficiently. These data sets are analysed and visualized to unveil the trends, human behaviour, and interactions.

The examples of big data

Social media site Facebook where hundreds of terabytes data is added daily in the form of text, audio, video, images etc.

GPS and Location Data: Navigation apps collect real-time location data from millions of users globally.

Online Streaming: Services like Netflix analyze user viewing habits and preferences to recommend content.

Credit Card Transactions: Banks process millions of transactions daily, detecting fraud through data analysis.

3. 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.

Machine learning has three types.

- **Supervised machine learning** (labelled data sets are used, here input and output variables are used to produce outcome)
- **Unsupervised machine learning** (un-labelled 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.)

4. 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.

5. Programming Languages

Programming languages especially 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.

Data Science applications

Data Science applies to every field and situation where data either already exist or can be obtained, essentially every industry and subject. This is because the Data Science deals with the study of data for the purpose of extracting meaningful insights, finding answers to interesting questions, and being able to make predictions.

The advancement in automatic data capturing, low-cost storage, and compute power has opened opportunities for applying Data Science in an unprecedented scenario. Let's take a brief look at some of those scenarios:

Health services will have data about patient's visits to clinics, medical diagnostics data, medications prescribed by doctors, improvements in the patients' health, data about diseases, and patient demographics. Data Science is applied here for a large variety of scenarios like building insights about the effectiveness of various treatments, predicting the high-risk patients for specific diseases, and preventive care that can be adopted.

Educational services like schools, colleges, universities possess data about different educational programs, student's socioeconomic data, teachers, enrolment data, student and teacher performance, dropouts, job placements. Data Science is applied for building insights about student's engagement, teacher's effectiveness, improvements in educational programs, determining dropout risk to help plan some preventive measures. Performance evaluations and effective formation of educational programs.

Sales and services companies like telecom providers, insurance providers, hospitality industry, manufacturing. Service providers have data about various services and products they offer, data about existing customers, data about the bundling of services, promotions, customer service consumption, feedback, and so on. Data Science is applied for driving customer satisfaction campaigns, proactive maintenance of infrastructure, profitable grouping of services, adjustments, and alterations for different geographies and seasons, effective stock management, early identification of risk of customer churn.

Financial sector is another important field where a very large number of financial transactions take place daily. In daily average transactions in the Forex market were to the tune of 6.6 trillion U.S. dollars. NPCI platform for retail payments and settlements in India recorded 4,428 million transactions in 2014-15 which reached 42,660 million transactions in extracting insights and patterns from this huge volume of transactions is possible only with the use of Data Science. The determination of new opportunities, crossselling, and detection of fraudulent transactions and much more can be achieved with the application of Data Science.

Research and analysis is closely related to the scientific methodology of Data Science. Research work revolves around building theories and hypotheses, conducting experiments, collecting, and analysing data to prove or reject the hypothesis. Hence, the research and data science are very closely related.

Intelligent applications are another big area for applying Data Science. User's historical and contextual data is utilized in building the models that help in delivering the personalized application behaviour, preferences, and choices. This produces highly adaptive applications with higher user satisfaction and retention.

The Steps in Doing Data Science

Problem Definition:

- Clearly define the problem or question you want to address using data.
- Understand the business context and objectives to ensure alignment.

Data Collection:

- Gather relevant data from various sources, which may include databases, APIs, spreadsheets, or data scraping.
- Ensure data quality, accuracy, and completeness.
- Organize and store the data in a suitable format.

Data Cleaning and Pre-processing:

- Clean and pre-process the data to handle missing values, outliers, and inconsistencies.
- Transform and reshape the data as needed for analysis.
- Feature engineering: Create new features or modify existing ones to enhance the data's relevance.

Exploratory Data Analysis (EDA):

- Explore and visualize the data to understand its characteristics, distributions, and relationships.
- Identify patterns, trends, and potential insights.
- EDA often involves creating plots, charts, and summary statistics.

Feature Selection and Engineering:

- Select the most relevant features (variables) for modelling to improve model performance and reduce complexity.
- Continue feature engineering as needed based on EDA findings.

Model Selection:

- Choose appropriate machine learning or statistical models based on the problem type (classification, regression, clustering, etc.) and data characteristics.
- Consider different algorithms and techniques and evaluate their suitability.

Data Splitting:

- Split the data into training, validation, and test sets to train and evaluate models separately.
- Cross-validation may also be used to assess model performance more robustly.

Model Training:

- Train the selected models using the training data.
- Optimize model hyper parameters through techniques like grid search or random search.

Model Evaluation:

- Evaluate model performance using appropriate metrics (e.g., accuracy, precision, recall, F1-score, RMSE).
- Compare multiple models and choose the best-performing one.

Model Interpretation:

- Interpret model results to understand the factors driving predictions.
- Use techniques like feature importance, SHAP values, or partial dependence plots.

Deployment:

- Deploy the chosen model into production if it provides value to the business.
- Implement necessary infrastructure and monitoring for real-time predictions.

Communication and Reporting:

- Communicate findings and insights to stakeholders using data visualizations, reports, and presentations.
- Explain the implications of the analysis and its impact on decision-making.

Feedback Loop:

- Continuously monitor model performance and retrain as needed with new data.
- Adapt to changing business requirements and data dynamics.

Skills Needed to Do Data Science - Storing Data and Combining Bits into Larger Structures

Database Management:

- Proficiency in working with databases like SQL and NoSQL.
- Ability to design, create, and manage databases to store structured data.

Data Storage Solutions:

- Familiarity with data storage technologies such as data lakes, data warehouses, and cloud storage.
- Knowledge of data storage formats like Parquet, Avro, or JSON.

Data Integration:

- Skills to integrate data from various sources into a unified dataset.
- Use of ETL (Extract, Transform, Load) processes and tools for data integration.

Big Data Technologies:

- Understanding of big data technologies like Hadoop and Spark for handling large-scale data.

Data Versioning and Cataloguing:

- Use of version control systems to track changes in data and code.
- Cataloguing data assets for easy discovery and access.

Skills Needed to Identify Data Problems

Data Profiling:

- Ability to assess data quality, including identifying missing values, duplicates, and outliers.
- Profiling data to understand its structure and characteristics.

Statistical Analysis:

- Proficiency in statistical techniques to analyse data distributions and relationships.
- Hypothesis testing and statistical modelling skills.

Domain Knowledge:

- Understanding of the specific industry or domain that the data pertains to.
- Knowledge of domain-specific data problems and challenges.

Data Visualization:

- Skills to create visual