

Data Science Introduction

Data Science is a combination of multiple disciplines that uses statistics, data analysis, and machine learning to analyze data and to extract knowledge and insights from it.

1.What is Data Science?

Data Science is about data gathering, analysis and decision-making. Data Science is about finding patterns in data, through analysis, and make future predictions.

By using Data Science, companies are able to make:

- Better decisions (should we choose A or B)
- Predictive analysis (what will happen next?)
- Pattern discoveries (find pattern, or maybe hidden information in the data)

Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.

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

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

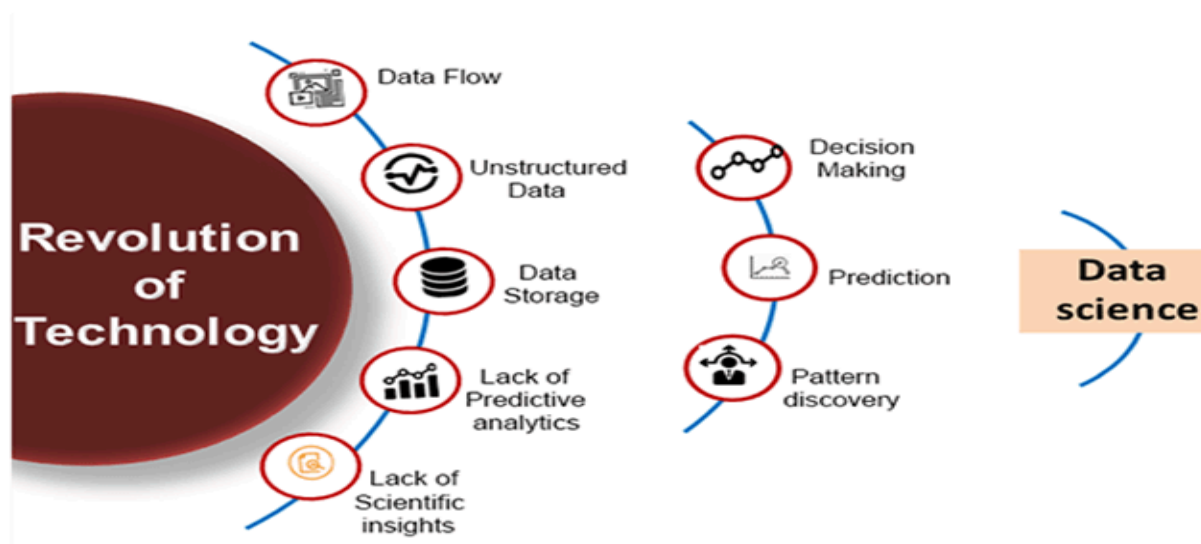
- Asking the correct questions and analyzing the raw data.
- Modeling the data using various complex and efficient algorithms.
- Visualizing the data to get a better perspective.
- Understanding the data to make better decisions and finding the final result.



Example:

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

Need for Data Science:



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

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

Where is Data Science Needed?

Data Science is used in many industries in the world today, e.g. banking, consultancy, healthcare, and manufacturing.

Examples of where Data Science is needed:

- For route planning: To discover the best routes to ship
- To foresee delays for flight/ship/train etc. (through predictive analysis)
- To create promotional offers
- To find the best suited time to deliver goods
- To forecast the next years revenue for a company
- To analyze health benefit of training
- To predict who will win elections

Types of Data Science Job

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

1. Data Scientist
2. Data Analyst
3. Machine learning expert
4. Data engineer

5. Data Architect
6. Data Administrator
7. Business Analyst
8. Business Intelligence Manager

Prerequisite for Data Science

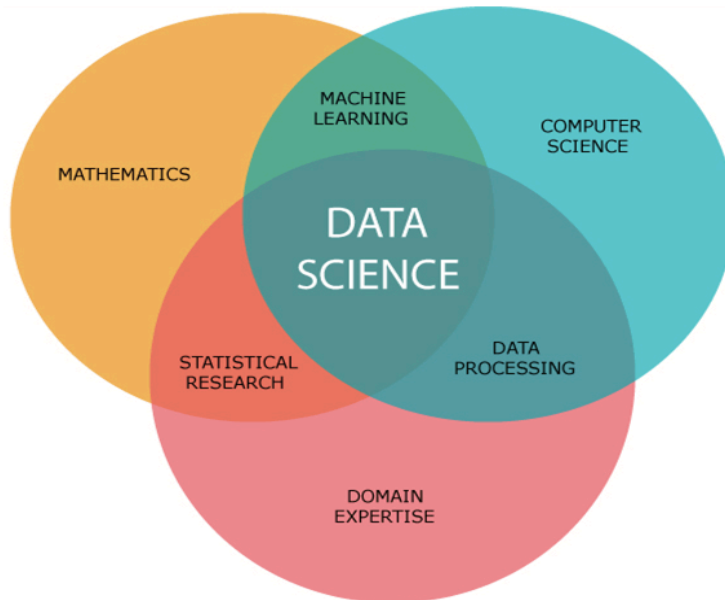
1.Non-Technical Prerequisite:

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

2.Technical Prerequisite:

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

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



Data Science can be applied in nearly every part of a business where data is available. Examples are:

- Consumer goods
- Stock markets
- Industry
- Politics
- Logistic companies
- E-commerce

Applications of Data Science:

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

etc., and these devices respond as per voice control, so this is possible with speech recognition algorithm.

- **Gaming world:** In the gaming world, the use of Machine learning algorithms is increasing day by day. EA Sports, Sony, Nintendo, are widely using data science for enhancing user experience.
- **Internet search:** When we want to search for something on the internet, then we use different types of search engines such as Google, Yahoo, Bing, Ask, etc. All these search engines use the data science technology to make the search experience better, and you can get a search result with a fraction of seconds.
- **Transport:** Transport industries also using data science technology to create self-driving cars. With self-driving cars, it will be easy to reduce the number of road accidents.
- **Healthcare:** In the healthcare sector, data science is providing lots of benefits. Data science is being used for tumour detection, drug discovery, medical image analysis, virtual medical bots, etc.
- **Recommendation systems:** Most of the companies, such as Amazon, Netflix, Google Play, etc., are using data science technology for making a better user experience with personalized recommendations. Such as, when you search for something on Amazon, and you started getting suggestions for similar products, so this is because of data science technology.
- **Risk detection:** Finance industries always had an issue of fraud and risk of losses, but with the help of data science, this can be rescued.

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

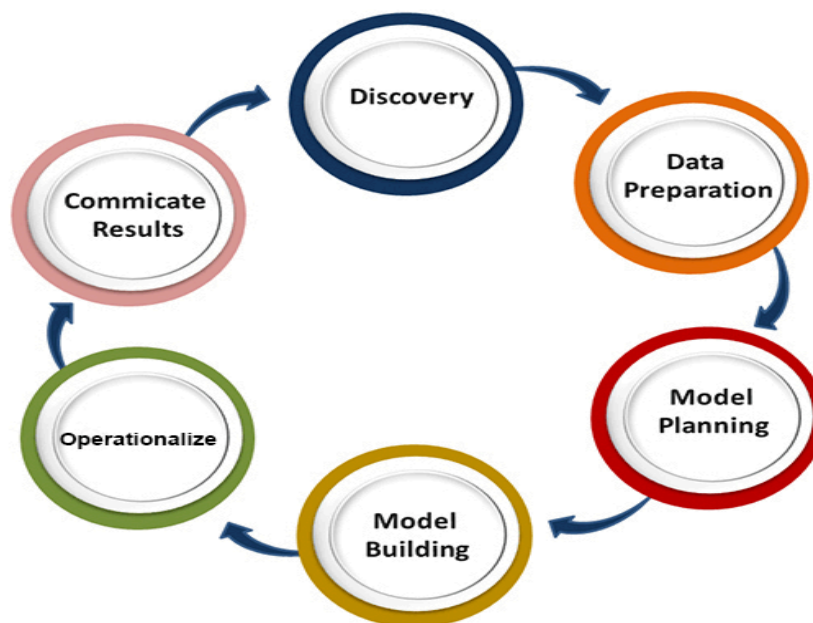
Tools for Data Science

- **Data Analysis tools:** R, Python, Statistics, SAS, Jupyter, R Studio, MATLAB, Excel, RapidMiner.
- **Data Warehousing:** ETL, SQL, Hadoop, Informatica/Talend, AWS Redshift
- **Data Visualization tools:** R, Jupyter, Tableau, Cognos.
- **Machine learning tools:** Spark, Mahout, Azure ML studio.



Data Science Lifecycle

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



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

1. Discovery: The first phase is discovery, which involves asking the right questions. When you start any data science project, you need to determine what are the basic requirements, priorities, and project budget. In this phase, we need to determine all the requirements of the project such as the number of people, technology, time, data, an end goal, and then we can frame the business problem on first hypothesis level.

2. Data preparation: Data preparation is also known as Data Munging. In this phase, we need to perform the following tasks:

- Data cleaning
- Data Reduction
- Data integration
- Data transformation

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

3. Model Planning: In this phase, we need to determine the various methods and techniques to establish the relation between input variables. We will apply Exploratory data analytics(EDA) by using various statistical formula and visualization tools to understand the relations between variable and to see what data can inform us. Common tools used for model planning are:

- SQL Analysis Services
- R
- SAS
- Python

4. Model-building: In this phase, the process of model building starts. We will create datasets for training and testing purpose. We will apply different techniques such as association, classification, and clustering, to build the model.

Following are some common Model building tools:

- SAS Enterprise Miner
- WEKA
- SPCS Modeler
- MATLAB

5. Operationalize: In this phase, we will deliver the final reports of the project, along with briefings, code, and technical documents. This

phase provides you a clear overview of complete project performance and other components on a small scale before the full deployment.

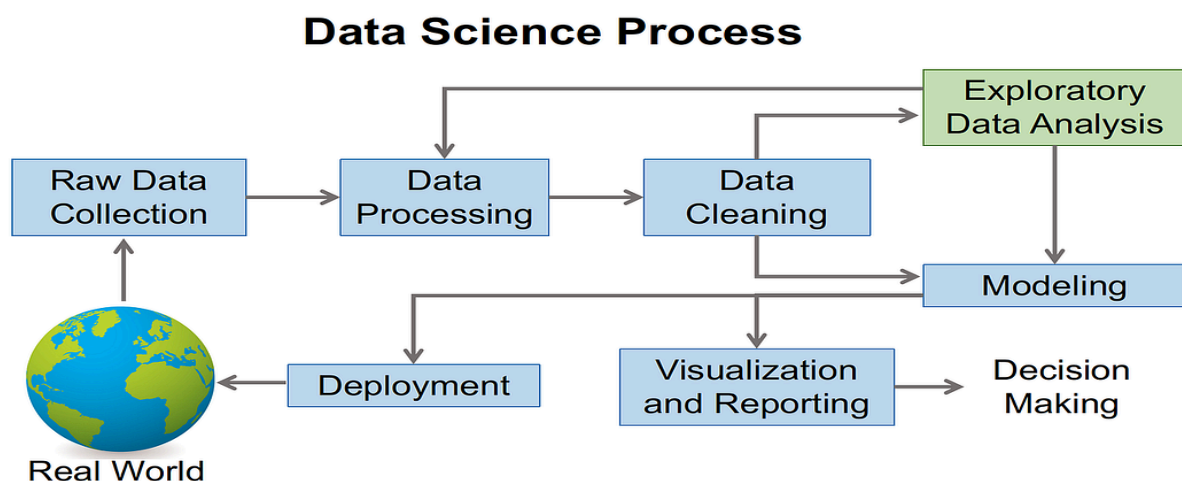
6. Communicate results: In this phase, we will check if we reach the goal, which we have set on the initial phase. We will communicate the findings and final result with the business team.

Data Science Process:

Data Science is a systematic approach to solving data-driven problems, involving the collection, analysis, interpretation, presentation, and communication of data.

The Data Science process is a structured framework used to complete a data science project, and it is essential for both business and research use cases.

Key Steps in the Data Science Process



Key Steps in the Data Science Process

1. Problem Definition: Understand the business problem, its impact, the ultimate goals for addressing it, and the relevant project plan.

2. Data Collection: Gather data from various sources, such as databases, APIs, or web scraping.

3. Data Processing: Perform preliminary data processing, such as handling missing values, encoding categorical variables, and scaling numerical variables.

4. Exploratory Data Analysis (EDA): Explore the data using summary statistics and visualizations to better understand its characteristics, identify patterns, relationships, and outliers.

- *Descriptive Statistics:* Calculate basic statistics, such as mean, median, mode, standard deviation, and variance, to summarize the data.
- *Visualizations:* Create visualizations, such as bar charts, line charts, scatter plots, and histograms, to gain insights into the data.

5. Data Cleaning: Based on the insights from EDA, clean the data by addressing outliers, inconsistencies, and missing values.

6. Modeling: Use the cleaned and understood data to build and train machine learning models.

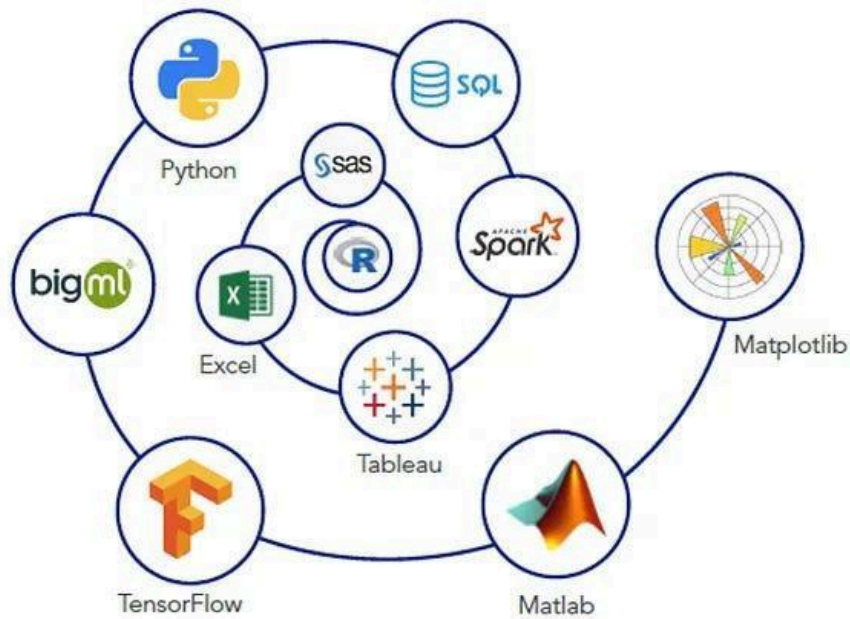
7. Evaluation: Assess the performance of the models using appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score.

8. Iteration: If the model's performance is not satisfactory, return to the EDA step to refine the data understanding and cleaning process, and then rebuild and retrain the models.

9. Reporting: Communicate the results of the analysis, including the insights gained from EDA, the chosen models, and their performance.

Tools for Data Science Process

There are various tools and programming languages used in the Data Science process, such as Mat lab, Tableau/Power BI, Python, and R. These tools provide utility features for different tasks in Data Science, making the process more efficient and effective.



Importance of Following a Well-Defined Data Science Process

Following a well-defined Data Science process has several benefits:

1. **Efficiency:** A structured process helps to ensure that the project is completed efficiently and effectively.
2. **Collaboration:** The process provides a clear way for stakeholders to coordinate and collaborate with the data science team.
3. **Reproducibility:** A well-defined process makes it easier for other data scientists to understand and reproduce the work.
4. **Domain Agnostic:** Data Science is domain agnostic, meaning it can be applied to any industry with data available.

This makes the Data Science process a valuable tool for solving problems across various industries.

What is Datafication?

It means is this- from our actions to our thoughts, everything is getting transformed into a numerically quantified format or “Data”.

Datafication enables the transformation of business operations, behaviors, and actions, in addition to those of its clients and consumers, into quantifiable, usable, and actionable data. This

information can then be tracked, processed, monitored, analyzed, and utilized to improve an organization and the products and services it offers to customers. To put them into perspective.

- Google transforms our searches into data
- Facebook transforms our friendships into data
- LinkedIn transforms our professional life into data
- Netflix or Amazon Prime transforms our watched TV shows and films into data
- Tinder transforms our dating activities into data
- Amazon transforms our shopping into data

Benefits of Datafication

Datafication is a technique that is financially advantageous to pursue since it provides great opportunity for streamlining corporate procedures. Datafication is a cutting-edge process for creating a futuristic framework that is both secure and inventive.

1. Actionable Insights

Datafication converts unstructured, incomprehensible data into usable insights, allowing you to get insight into your processes and procedures – the basis of any organization.

Datafication suggests that you will be better able to comprehend the advantages, disadvantages, potential, and future of your business. Additionally, it gives you knowledge about the results and effects of your efforts, allowing you to evaluate what you're doing and how you're doing it.

2. Digital Transformation

In order for organisations to stay relevant and up-to-date in a constantly shifting ecosystem, digital transformation is no longer just a passing trend but rather a necessity.

You should have useable data if you want to benefit from the most recent and cutting-edge technologies. It holds the key to enhancing the effectiveness and efficiency of corporate operations. You will be

better able to comprehend the organization's current situation and the necessary next actions to advance.

3. Improve Productivity and Efficiency

Datafication will be able to understand what you're doing and how you're improving. Streamlining operations will improve the utilisation of all assets, including staff members, to increase output and efficiency overall and turn your company into a prosperous corporation.

4. Manage Information

Any business generates a lot of data, which is constantly collected and stored. A well-managed data set will produce superior outcomes. Otherwise, it could become overwhelming or useless information.

By correctly organising it through datafication, you can use the information to make judgements. You will be able to access and analyse the data in addition to storing it. To contextualise the consumer experience, several organisations are experimenting with integrating user-sourced data and putting it into apps.

4. Exploratory Data Analysis in python

Exploratory data analysis is one of the basic and essential steps of a data science project. A data scientist involves almost 70% of his work in doing the EDA of the dataset.

What is Exploratory Data Analysis (EDA)?

Exploratory Data Analysis (EDA) is a crucial initial step in data science projects.

It involves analysing and visualizing data to understand its key characteristics, uncover patterns, and identify relationships between variables refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables.

Key aspects of EDA include:

- **Distribution of Data:** Examining the distribution of data points to understand their range, central tendencies (mean, median), and dispersion (variance, standard deviation).
- **Graphical Representations:** Utilizing charts such as histograms, box plots, scatter plots, and bar charts to visualize relationships within the data and distributions of variables.
- **Outlier Detection:** Identifying unusual values that deviate from other data points. Outliers can influence statistical analyses and might indicate data entry errors or unique cases.
- **Correlation Analysis:** Checking the relationships between variables to understand how they might affect each other. This includes computing correlation coefficients and creating correlation matrices.
- **Handling Missing Values:** Detecting and deciding how to address missing data points, whether by imputation or removal, depending on their impact and the amount of missing data.
- **Summary Statistics:** Calculating key statistics that provide insight into data trends.
- **Testing Assumptions:** Many statistical tests and models assume the data meet certain conditions (like normality or homoscedasticity). EDA helps verify these assumptions.

Why Exploratory Data Analysis is Important?

Exploratory Data Analysis (EDA) is important for several reasons, especially in the context of data science and statistical modelling. Here are some of the key reasons why EDA is a critical step in the data analysis process:

1. **Understanding Data Structures:** EDA helps in getting familiar with the dataset, understanding the number of features, the type of data in each feature, and the distribution of data points. This understanding is crucial for selecting appropriate analysis or prediction techniques.
2. **Identifying Patterns and Relationships:** Through visualizations and statistical summaries, EDA can reveal hidden patterns and intrinsic relationships between variables. These insights can guide further analysis and enable more effective feature engineering and model building.

3. **Detecting Anomalies and Outliers:** EDA is essential for identifying errors or unusual data points that may adversely affect the results of your analysis. Detecting these early can prevent costly mistakes in predictive modeling and analysis.
4. **Testing Assumptions:** Many statistical models assume that data follow a certain distribution or that variables are independent. EDA involves checking these assumptions. If the assumptions do not hold, the conclusions drawn from the model could be invalid.
5. **Informing Feature Selection and Engineering:** Insights gained from EDA can inform which features are most relevant to include in a model and how to transform them (scaling, encoding) to improve model performance.
6. **Optimizing Model Design:** By understanding the data's characteristics, analysts can choose appropriate modelling techniques, decide on the complexity of the model, and better tune model parameters.
7. **Facilitating Data Cleaning:** EDA helps in spotting missing values and errors in the data, which are critical to address before further analysis to improve data quality and integrity.
8. **Enhancing Communication:** Visual and statistical summaries from EDA can make it easier to communicate findings and convince others of the validity of your conclusions, particularly when explaining data-driven insights to stakeholders without technical backgrounds.

Types of Exploratory Data Analysis

EDA, or Exploratory Data Analysis, refers back to the method of analysing and analysing information units to uncover styles, pick out relationships, and gain insights. There are various sorts of EDA strategies that can be hired relying on the nature of the records and the desires of the evaluation. Depending on the number of columns we are analysing we can divide EDA into three types:

[Univariate, bivariate and multivariate.](#)

1. Univariate Analysis

Univariate analysis focuses on a single variable to understand its internal structure. It is primarily concerned with describing the data and finding patterns existing in a single feature. This sort of

evaluation makes a speciality of analysing character variables inside the records set. It involves summarizing and visualizing a unmarried variable at a time to understand its distribution, relevant tendency, unfold, and different applicable records. Common techniques include:

- **Histograms:** Used to visualize the distribution of a variable.
- **Box plots:** Useful for detecting outliers and understanding the spread and skewness of the data.
- **Bar charts:** Employed for categorical data to show the frequency of each category.
- **Summary statistics:** Calculations like mean, median, mode, variance, and standard deviation that describe the central tendency and dispersion of the data.

2. Bivariate Analysis

Bivariate evaluation involves exploring the connection between variables. It enables find associations, correlations, and dependencies between pairs of variables. Bivariate analysis is a crucial form of exploratory data analysis that examines the relationship between two variables. Some key techniques used in bivariate analysis:

- **Scatter Plots:** These are one of the most common tools used in bivariate analysis. A scatter plot helps visualize the relationship between two continuous variables.
- **Correlation Coefficient:** This statistical measure (often Pearson's correlation coefficient for linear relationships) quantifies the degree to which two variables are related.
- **Cross-tabulation:** Also known as contingency tables, cross-tabulation is used to analyze the relationship between two categorical variables. It shows the frequency distribution of categories of one variable in rows and the other in columns, which helps in understanding the relationship between the two variables.
- **Line Graphs:** In the context of time series data, line graphs can be used to compare two variables over time. This helps in identifying trends, cycles, or patterns that emerge in the interaction of the variables over the specified period.
- **Covariance:** Covariance is a measure used to determine how much two random variables change together. However, it is sensitive to the scale of the variables, so it's often

supplemented by the correlation coefficient for a more standardized assessment of the relationship.

3. Multivariate Analysis

Multivariate analysis examines the relationships between two or more variables in the dataset. It aims to understand how variables interact with one another, which is crucial for most statistical modeling techniques. Techniques include:

- **Pair plots:** Visualize relationships across several variables simultaneously to capture a comprehensive view of potential interactions.
- **Principal Component Analysis (PCA):** A dimensionality reduction technique used to reduce the dimensionality of large datasets, while preserving as much variance as possible.

Specialized EDA Techniques

In addition to univariate and multivariate analysis, there are specialized EDA techniques tailored for specific types of data or analysis needs:

- **Spatial Analysis:** For geographical data, using maps and spatial plotting to understand the geographical distribution of variables.
- **Text Analysis:** Involves techniques like word clouds, frequency distributions, and sentiment analysis to explore text data.
- **Time Series Analysis:** This type of analysis is mainly applied to statistics sets that have a temporal component. Time collection evaluation entails inspecting and modeling styles, traits, and seasonality inside the statistics through the years. Techniques like line plots, autocorrelation analysis, transferring averages, and ARIMA (AutoRegressive Integrated Moving Average) fashions are generally utilized in time series analysis.

Tools for Performing Exploratory Data Analysis

Exploratory Data Analysis (EDA) can be effectively performed using a variety of tools and software, each offering unique features suitable for handling different types of data and analysis requirements.

1. Python Libraries

- **Pandas**: Provides extensive functions for data manipulation and analysis, including data structure handling and time series functionality.
- **Matplotlib**: A plotting library for creating static, interactive, and animated visualizations in Python.
- **Seaborn**: Built on top of Matplotlib, it provides a high-level interface for drawing attractive and informative statistical graphics.
- **Plotly**: An interactive graphing library for making interactive plots and offers more sophisticated visualization capabilities.

2. R Packages

- **ggplot2**: Part of the tidyverse, it's a powerful tool for making complex plots from data in a data frame.
- **dplyr**: A grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges.
- **tidyr**: Helps to tidy your data. Tidying your data means storing it in a consistent form that matches the semantics of the dataset with the way it is stored.

Real World Business Cases



Machine learning and data science have been applied in various real-world scenarios, providing valuable insights and driving decision-making across different industries. Here are some examples:

1. **Recommendation Systems:** Machine learning algorithms are used to analyze user behavior and predict preferences, enabling personalized recommendations on e-commerce websites and streaming services[6].

2. **Social Media Connections:** Machine learning helps to identify and suggest potential connections on social media platforms based on user behavior and interests[6].

3. **Image Recognition:** Machine learning algorithms can identify objects in images, enabling applications such as facial recognition and self-driving cars[7].

4. **Natural Language Processing:** Machine learning is used to analyze and understand human language, enabling applications such as speech recognition and automated translation[6].

5. **Medical Diagnosis:** Machine learning can analyze medical data to identify patterns and predict diseases, helping healthcare professionals to make more accurate diagnoses and treatment plans[8].

6. **Financial Fraud Detection:** Machine learning algorithms can analyze transaction data to identify unusual patterns and discrepancies, helping banks and financial institutions to detect fraudulent activities[7].

7. **Predictive Analytics:** Machine learning can classify available data into groups and calculate the probability of specific outcomes, enabling applications such as predicting customer churn and optimizing product pricing[7].

8. **Extraction:** Machine learning can extract structured information from unstructured data, helping organizations to manage and analyze large volumes of data from customers[7].

9. **Statistical Arbitrage:** Machine learning can analyze large data sets and identify real-time arbitrage opportunities in financial markets, optimizing trading strategies and enhancing results[7].

These examples demonstrate the versatility and impact of machine learning and data science in various industries, helping businesses and organizations make data-driven decisions and improve overall efficiency[6].



A Data Scientist's Role in This Process

This model so far seems to suggest this will all magically happen without human intervention. By “human” here, we mean “data scientist.” Someone has to make the decisions about what data to collect, and why. That person needs to be formulating questions and hypotheses and making a plan for how the problem will be attacked. And that someone is the data scientist or our beloved data science team.

Let's revise or at least add an overlay to make clear that the data scientist needs to be involved in this process throughout, meaning they are involved in the actual coding as well as in the higher-level process, as shown in [Figure 2-3](#).

Connection to the Scientific Method

We can think of the data science process as an extension or variation of the scientific method:

- Ask a question.
- Do background research.
- Construct a hypothesis.
- Test your hypothesis by doing an experiment.
- Analyze your data and draw a conclusion.
- Communicate your results.

In both the data science process and the scientific method, not every problem requires one to go through all the steps, but almost all problems can be solved with some combination of the stages. For example, if your end goal is a data visualization (which itself could be thought of as a data product), it's possible you might not do any machine learning or statistical modeling, but you'd want to get all the way to a clean dataset, do some exploratory analysis, and then create the visualization.

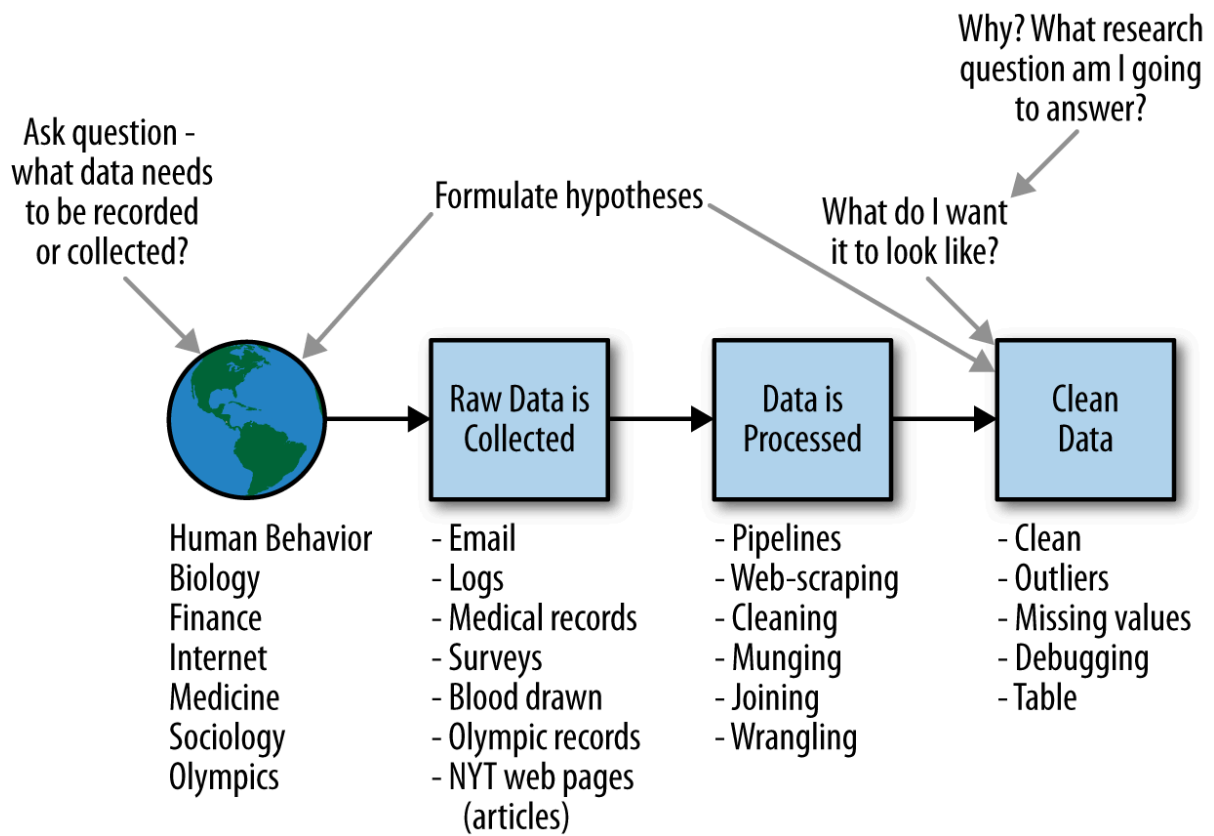


Figure 2-3. The data scientist is involved in every part of this process