- **Data Science :** It is branch of computer Science where we analyse large amount of data apply techniques(algorithms) on it and extract insights that help us to take decisions.

  15 Python Libraries for Data Science You Should Know – Dataquest

  E.g :

  1)Lets say we have hotels in india and usa ,we have sales data for each month and want to anlyze this data .

Month	India(Sales)	US(Sales)
Jan	1000	800
Feb	1500	1000
Mar	890	900
Apr	220	700
May	180	1100
Jun	200	900
July	900	1100
Aug	700	2000
Sept	1000	900
Oct	990	890
Nov	999	700
Dec	890	900

If we draw bar chart for every moth it will show that during summer (Apr,May,Jun) sale is less india so will carried out insight that we need to do advertisement to boost sales in this month .this can be done in excel but data is growing in millions or billions due to internet, social media so it is not possible to handel this data using excel so we use Python ,R langugue for the same.

2)Sales Prediction : suppose we have data that we spent on advertisement and sales .and wanna predict how much will be sale when will spend for amount next year.this can be done by using linear regression algorithm and predict next year sale.

Year	Amount paid for Advertisement	Sales
2020	120 $	2300$
2021	100$	2000$
2022	200$	1700$
2023	250$	?

Need Of Data Science :

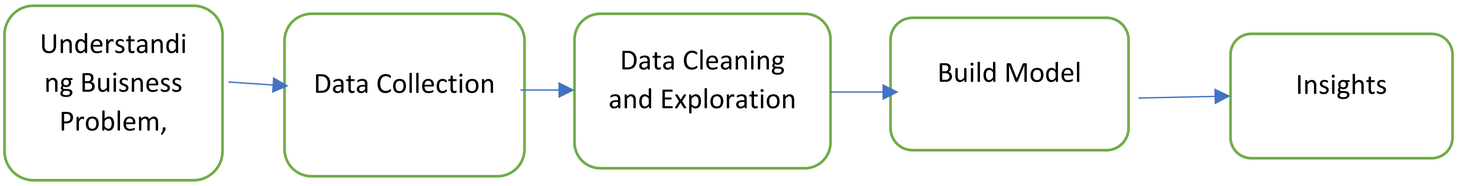
Some years ago, data was less and mostly available in a structured form, which could be easily stored in excel sheets, and processed using BI tools.

But in today's world, data is becoming so vast, i.e., approximately **2.5 quintals bytes** of data is generating on every day, which led to data explosion. It is estimated as per researches, that by 2020, 1.7 MB of data will be created at every single second, by a single person on earth. Every Company requires data to work, grow, and improve their businesses.

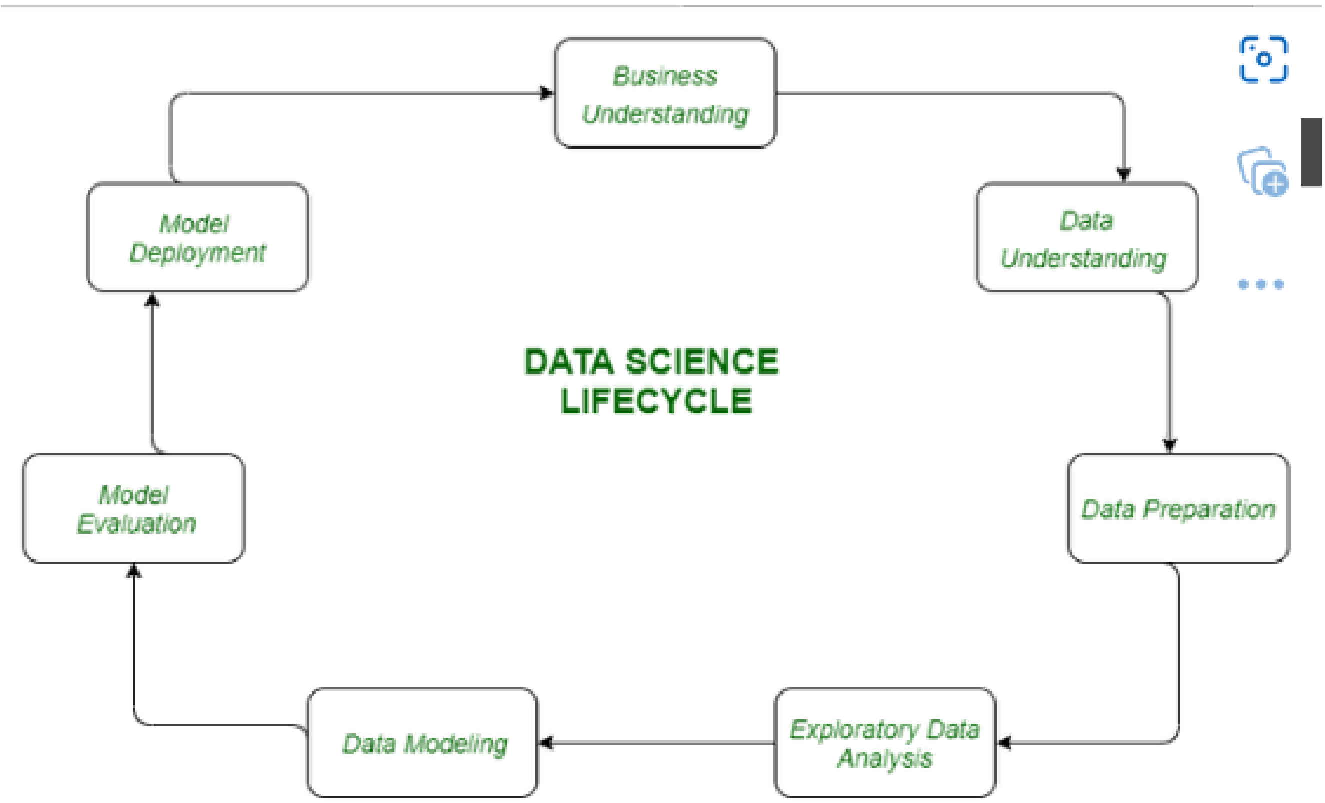
Now, handling of such huge amount of data is a challenging task for every organization. So to handle, process, and analysis of this, we required some complex, powerful, and efficient algorithms and technology, and that technology came into existence as data Science. Following are some main reasons for using data science technology:

- 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.

Data Science Process :



Understanding Buisness Problem, → Data Collection → Data Cleaning and Exploration → Build Model → Insights



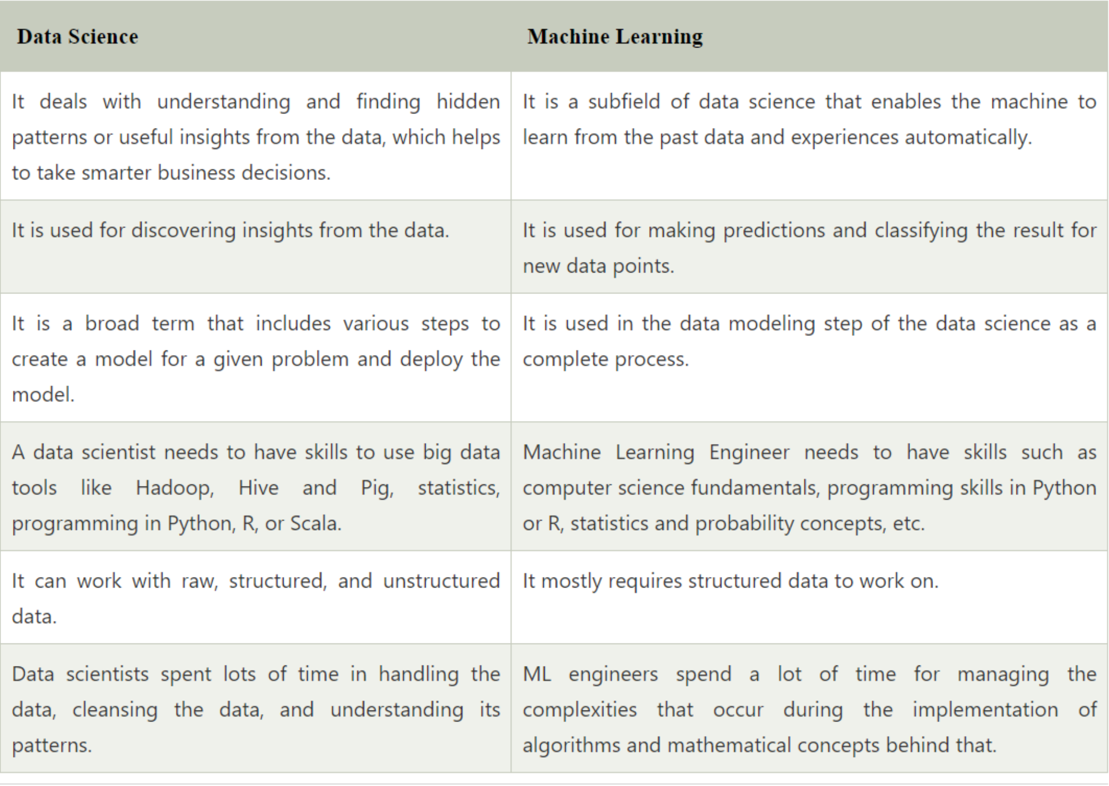
- **Three Types of Data Science :**

  - **Data Mining :** also called web scraping.it means extracting useful data from huge set of data.BeautifulSoup,Scrapy libraries are used for this.
  - **Data Processing and Modelling :** Clean data and build the model.Numpy,pandas,Scipy,sklearn,kearas ,Tensorflow.
  - **Data Visualization :** representinfg data in the form of chart ,plots.Matplotlib,Seaborn.
  - **Data Wrangling :** also called data mugging .It is the process of transforming and mapping data from one "raw" data form into another format to make it more appropriate and valuable for various downstream purposes such as analytics. The goal of data wrangling is to assure quality and useful data. Data analysts typically spend the majority of their time in the process of data wrangling compared to the actual analysis of the data.
  - **Data Warehousing :** Combing data from various resourses at single database.

- **Applications of Data Science :**
  1)Finacial Industries : Stock Market Analyze past data and make future predictions.
  2)Health care : Detecting Tumor ,Medical image Analysis it provide data in terms of mathematical values for different features and we can train model for it and can use this model for prediction.

3)Search engine : when we search on search enging it provide mostly visited content by analysing past data.
4)Autocomplete Facility in email by analysing past text written.
5)Image Recognition : when we upload image on social media like facebook it suggestion of tagging. When image is recognized ,if it matched with picture available it provide autotagging.
6)Transportation : For self driving cars data is fed to algorithm that analyses what should be speed on highway,busy Street ,Narrow Roads.
7)E-Commerce : when we search any product on social media ,our interaction of that product recorded in terms of data points and abnalysed and it provide suggestion to us.

Tools : 1)Apache Hadoop/spark : java based framework to process and analyse large amount of data
2Tablu : data visualization
3)Rapidminer :data mining and ML
4)Tensorflow : for ML and deep learning.
5)Hive and Pig :



Data Science	Machine Learning
It deals with understanding and finding hidden patterns or useful insights from the data, which helps to take smarter business decisions.	It is a subfield of data science that enables the machine to learn from the past data and experiences automatically.
It is used for discovering insights from the data.	It is used for making predictions and classifying the result for new data points.
It is a broad term that includes various steps to create a model for a given problem and deploy the model.	It is used in the data modeling step of the data science as a complete process.
A data scientist needs to have skills to use big data tools like Hadoop, Hive and Pig, statistics, programming in Python, R, or Scala.	Machine Learning Engineer needs to have skills such as computer science fundamentals, programming skills in Python or R, statistics and probability concepts, etc.
It can work with raw, structured, and unstructured data.	It mostly requires structured data to work on.
Data scientists spent lots of time in handling the data, cleansing the data, and understanding its patterns.	ML engineers spend a lot of time for managing the complexities that occur during the implementation of algorithms and mathematical concepts behind that.

STEP 2: Data Collection and Data Cleaning

- Data Preparation :- It is the process of making data ready for analysis it include cleaning data .
- Data Cleaning :-

When working with multiple data sources, there are many chances for data to be incorrect, duplicated, or mislabeled. If data is wrong, outcomes and algorithms are unreliable, even though they may look correct. *Data cleaning* is the process of changing or eliminating garbage, incorrect, duplicate, corrupted, or incomplete data in a dataset.

- The motive of data cleaning services is to construct uniform and standardized data sets that enable easy access to data analytics tools and business intelligence and perceive accurate data for each problem.
- If we provide uncleaned ,missing data to ml algorithm it can give unreliable or incorrect result so it is very essential to clean data .

- Reasons of Data Corrption :-

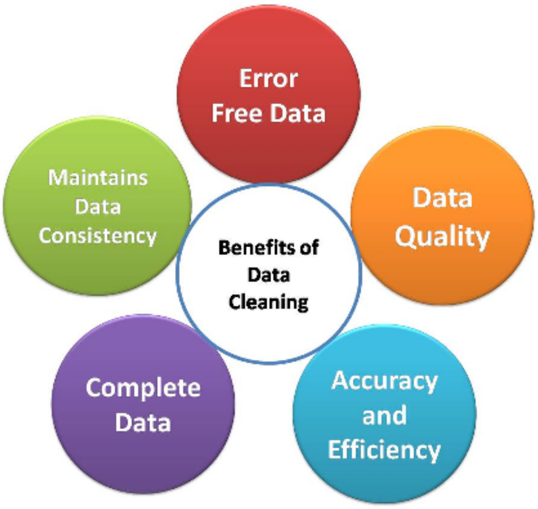
  1)Data is collected from various structured and unstructured sources and then combined, leading to duplicated and mislabeled values.

  2)Different data dictionary definitions for data stored at various locations.

  3) Manual entry error/Typos.

  4)Mislabelled categories/classes.

- Why Data Cleaning is Essential :



1. Error-Free Data: When multiple sources of data are combined, there may be a chance of so much error. Through Data Cleaning, errors can be removed from data

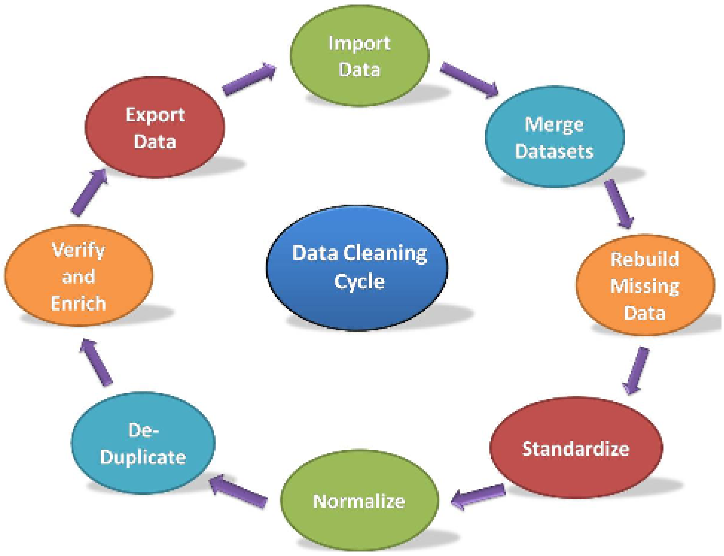
2. Data Quality: The quality of the data is the degree to which it follows the rules of particular requirements. For example, if we have imported phone numbers data of different customers, and in some places, we have added email addresses of customers in the data. But because our needs were straightforward for phone numbers, then the email addresses would be invalid data.Data cleaning will help us simplify this process and avoid useless data values.

3. Accurate and Efficient: Ensuring the data is close to the correct values. We know that most of the data in a dataset are valid, and we should focus on establishing its accuracy. Even if the data is authentic and correct, it doesn't mean it is accurate. Determining accuracy helps to figure out whether the data entered is accurate or not. For example, a customer's address is stored in the specified format; maybe it doesn't need to be in the right one. The email has an additional character or value that makes it incorrect or invalid.

4. Complete Data: Completeness is the degree to which we should know all the required values. Completeness is a little more challenging to achieve than accuracy or quality. Because it's nearly impossible to have all the info we need, only known facts can be entered. We can try to complete data by redoing the data-gathering activities like approaching the clients again, re-interviewing people, etc. For example, we might need to enter every customer's contact information. But a number of them might not have email addresses. In this case, we have to leave those columns empty. If we have a system that requires us to fill all columns, we can try to enter missing or unknown there. But entering such values does not mean that the data is complete. It would still be referred to as incomplete.

5. Maintains Data Consistency: To ensure the data is consistent within the same dataset or across multiple datasets, we can measure consistency by comparing two similar systems. We can also check the data values within the same dataset to see if they are consistent or not. Consistency can be relational. For example, a customer's age might be 25, which is a valid value and also accurate, but it is also stated as a senior citizen in the same system. In such cases, we have to cross-check the data, similar to measuring accuracy, and see which value is true. Is the client a 25-year-old? Or is the client a senior citizen? Only one of these values can be true. There are multiple ways to for your data consistent.

- Data Cleaning Life Cycle :



1) Import Data : import the data.
2) Merge data : Merging the dataset is the process of combining two datasets in one and lining up rows based on some particular or common property for data analysis. We can do this by using the merge() function of the dataframe.
3) Rebuild Missing data + Drop rows having missing valiues(dropna()) : To find and fill in the missing data in the dataset, we will use another function. There are 4 ways to find the null values if present in the dataset.
   i)isnull() : return true if missing value is there.
   ii)fillna() : fill given value at null value .[it can be mean ,median ,mode of given column]
   iii)isna().any() ,isna().sum()
   iv)isna().any().sum()

4)Standardization and Normalization : Data Standardization and Normalization is a common practices in machine learning.

Standardization: is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.

Normalization : is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

It aslo refers to changing the distribution of the data so that it can represent a bell curve where the values of the attribute are equally distributed across the mean

Scaling: refers to transforming the range of data and shifting it to some other value range. This is beneficial when we want to compare different attributes

5)De-Duplicate : De-Duplicate means removing all duplicate values. There is no need for duplicate values in data analysis. These values only affect the accuracy and efficiency of the analysis result.

Functions used : 1)Duplicated() : return true if duplicate data is there.

2) DataFrame_name.drop_duplicates(subset=None, keep='first', inplace=False, ignore_index=False)

6)Verify and Enrich data : After removing null, duplicate, and incorrect values, we should verify the dataset and validate its accuracy.

7) Export data : This is the last step of the data-cleaning process. After performing all the above operations, the data is transformed into a clean dataset, and it is ready to export for the next process in Data Science or Data Analysis.

- **Data Preparation Steps :**

1)Load data in Pandas.

2)Drop columns that aren't useful.

3) Drop rows with missing values.

4)Create dummy variables.

5)Take care of missing data.

6)Convert the data frame to NumPy.

7)Divide the data set into training data and test data.