# PROJECT: **AMAZON CUSTOMER SENTIMENT PREDICTION USING ML AND DL**

TEAM NO 20

TEAM MEMBERS:

1. PRAKANSHU PATHAK(TEAM HEAD)-(SIRSS2327)
2. ANIKET RAJENDRA DHANE(VICE HEAD) )-(SIRSS2325)
3. TWINKLE )-(SIRSS2321)
4. BADAL )-(SIRSS2334)
5. POOJA KUMARI)-(SIRSS2329)

SUBMISSION DATE:15/10/2021

**WHAT IS DATA SCIENCE?**

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.

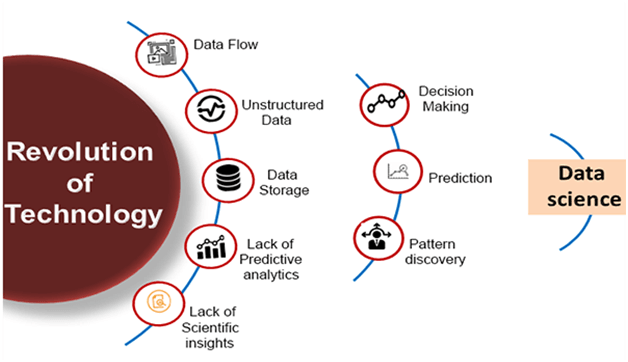
Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems. It is the future of artificial intelligence.



### **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:



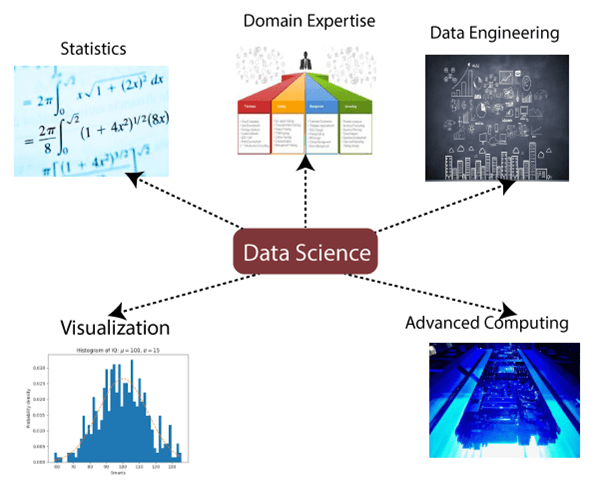
Some years ago, data was less and mostly available in a structured form, which could be easily stored in excel sheets, and processed using BI tools.

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 COMPONENTS:



The main components of Data Science are given below:

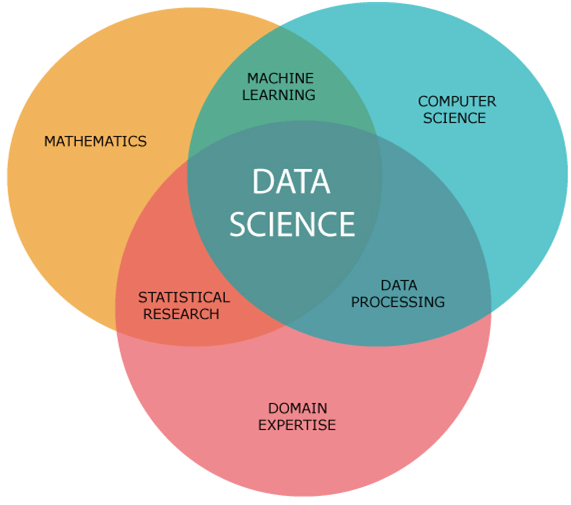
**1. Statistics:** Statistics is one of the most important components of data science. Statistics is a way to collect and analyze the numerical data in a large amount and finding meaningful insights from it.

**Domain Expertise:** In data science, domain expertise binds data science together. Domain expertise means specialized knowledge or skills of a particular area. In data science, there are various areas for which we need domain experts.

**3. Data engineering:** Data engineering is a part of data science, which involves acquiring, storing, retrieving, and transforming the data. Data engineering also includes metadata (data about data) to the data.

**4. Visualization:** Data visualization is meant by representing data in a visual context so that people can easily understand the significance of data. Data visualization makes it easy to access the huge amount of data in visuals.

**5. Advanced computing:** Heavy lifting of data science is advanced computing. Advanced computing involves designing, writing, debugging, and maintaining the source code of computer programs.



**6.Mathematics:** Mathematics is the critical part of data science. Mathematics involves the study of quantity, structure, space, and changes. For a data scientist, knowledge of good mathematics is essential.

**7. Machine learning:** Machine learning is backbone of data science. Machine learning is all about to provide training to a machine so that it can act as a human brain. In data science, we use various machine learning algorithms to solve the problems.

**APPLICATIONS OF DATA SCIENCE:**

1. Image recognition and speech recognition:
2. Gaming world:
3. Internet search:
4. Transport:
5. Healthcare:
6. Recommendation systems:
7. Risk detection:

**PROJECT DESCRIPTION**

# **AMAZON CUSTOMER SENTIMENT PREDICTION USING ML AND DL**

# **AMAZON REVIEWS DATASET**

This dataset contains several million reviews of Amazon products, with the reviews separated into two classes for positive and negative reviews. The two classes are evenly balanced here.

This is a large dataset, and the version that I am using here only has the text as a feature with no other metadata. This makes this an interesting dataset for doing NLP work. It is data written by users, so it's like that there are various typos, nonstandard spellings, and other variations that you may not find in curated sets of published text.

In this notebook, I will do some very simple text processing and then try out two fairly unoptimized deep learning models:

1. A convolutional neural net
2. A recurrent neural net These models should achieve results that are within a couple percent of state of the art at predicting the binary sentiment of the reviews.

**SENTIMENT ANALYSIS**

Sentiment analysis, also known as opinion mining. The goal of this research field is to analyse people's views, feelings, evaluations, attitudes and emotions towards entities and their attributes from the text, and these entities can be a variety of products, services, institutions, individuals, events, issues, or topics

**DEPENDENCIES**

Following are the dependencies of the project

* pandas
* numpy
* matplotlib.pyplot
* seaborn

# **BASIC NLP COUNT BASED FEATURES :**

A number of basic text based features can also be created which sometimes are helpful for improving text classification models. Some examples are:

* **Word Count**: total number of words in the documents
* **Character Count**: total number of characters in the documents
* Average Word Density: average length of the words used in the documents
* **Puncutation Count:** total number of punctuation marks in the documents
* **Upper Case Count**: total number of upper count words in the documents
* **Title Word Count**: total number of proper case (title) words in the documents

# **LEVERAGING TEXT SENTIMENT**

Reviews are pretty subjective, opinionated and people often express stong emotions, feelings. This makes it a classic case where the text documents here are a good candidate for extracting sentiment as a feature.

The general expectation is that highly rated and recommended products (label 1) should have a positive sentiment and products which are not recommended (label 0) should have a negative sentiment.

TextBlob is an excellent open-source library for performing NLP tasks with ease, including sentiment analysis. It also an a sentiment lexicon (in the form of an XML file) which it leverages to give both polarity and subjectivity scores.

* The polarity score is a float within the range [-1.0, 1.0].
* The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.

# **FEATURES FROM SENTIMENT ANALYSIS :**

Remember this is unsupervised, lexicon-based sentiment analysis where we don't have any pre-labeled data saying which review migth have a positive or negative sentiment. We use the lexicon to determine this.

## LIMITATION AND FUTURE IMPROVEMENTS

Currently the model in only 81% accurate. So we have a target to increase the accuracy to 90%.

* Due to less physical storage, we are not able to get the model accuracy upto 90%.