

**DATTA MEGHE COLLEGE OF ENGINEERING**

Plot no. 98, CIDCO, Sector-3, Post Box No.15, Airoli, NaviMumbai-400708.

A Report on

**“DATA SCIENCE”**

Submitted

to

**Dr. Kanungo Prachi**

by

Students of Third Year Information Technology Engineering

Report submitted in fulfilment of requirement of

The University of Mumbai

in

Business Communication and Ethics

October 2019

**SUBMITTED BY:**

|  |  |  |
| --- | --- | --- |
| **SR NO.** | **NAME** | **ROLL NO.** |
| 1 | Harshal Bhoir | 2 |
| 2 | Prasad Deshmukh | 9 |
| 3 | Samruddhi Deshpande | 10 |
| 4 | Pratiksha Gaikwad | 13 |
| 5 | Shubham Gulik | 17 |
| 6 | Pooja Gupta | 18 |

**ACKNOWLEDGEMENT**

Working on the project on “**Data Science”** was a source of immense knowledge to us. However, this would not have been possible without the equal support of all the members.

We would like to express our sincere gratitude to **Prof. Mrs. Kanungo Prachi** for her guidance and valuable support. We would like to extend a sincere thanks to her.

**ABSTRACT**

We live in a world that’s drowning in data. Websites track every user’s every click. Your smartphone is building up a record of your location and speed every second of every day. “Quantified selfers” wear pedometers-on-steroids that are ever recording their heart rates, movement habits, diet, and sleep patterns. Smart cars collect driving habits, smart homes collect living habits, and smart marketers collect purchasing habits. The Internet itself represents a huge graph of knowledge that contains (among other things) an enormous cross-referenced encyclopedia, domain-specific databases about movies, music, sports results, pinball machines, memes, and cocktails; and too many government statistics (some of them nearly true!) from too many governments to wrap your head around.

The volume of data that is generated each day is rising rapidly. There is a need to analyse this data efficiently and produce results quickly. Data science offers a formal methodology for processing and analysing data. It involves a work-flow with multiple stages, such as, data collection, data wrangling, statistical analysis and machine learning.

**INTRODUCTION**

Data scientist has been called “the prettiest job of the 21st century,” presumably by someone who has never visited a fire station. Nonetheless, data science is a hot and growing field, and it doesn’t take a great deal of sleuthing to find analysts breathlessly prognosticating that over the next 10 years, we’ll need billions and billions more data scientists than we currently have.

**1.1 The Ascendance of Data**

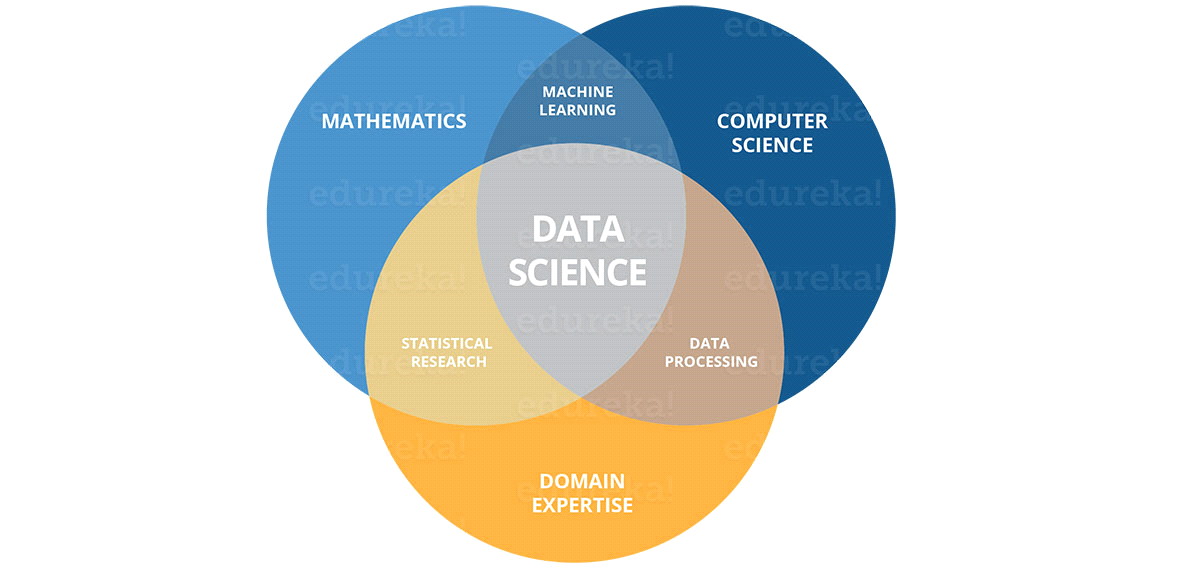
We live in a world that’s drowning in data. Websites track every user’s every click. Your smartphone is building up a record of your location and speed every second of every day. “Quantified selfers” wear pedometers-on-steroids that are ever recording their heart rates, movement habits, diet, and sleep patterns. Smart cars collect driving habits, smart homes collect living habits, and smart marketers collect purchasing habits. The Internet itself represents a huge graph of knowledge that contains (among other things) an enormous cross-referenced encyclopedia, domain-specific databases about movies, music, sports results, pinball machines, memes, and cocktails; and too many government statistics (some of them nearly true!) from too many governments to wrap your head around. Buried in these data are answers to countless questions that no one’s ever thought to ask.

**1.2 What Is Data Science?**

There’s a joke that says a data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician. (I didn’t say it was a good joke.) In fact, some data scientists are — for all practical purposes — statisticians, while others are pretty much indistinguishable from software engineers. Some are machine-learning experts, while others couldn’t machine-learn their way out of kindergarten. Some are PhDs with impressive publication records, while others have never read an academic paper (shame on them, though). In short, pretty much no matter how you define data science, you’ll find practitioners for whom the definition is totally, absolutely wrong.

According to a Venn diagram that is somewhat famous in the industry, data science lies at the intersection of:

* Computer Science
* Math and Statistics Knowledge
* Domain Expertise



**Fig 1.1 : Venn Diagram of Data Science**

**Data science** is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science is the same concept as data mining and big data: "use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems

**Data Mining :** Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use.

**Big Data :** "Big data" is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software

We’ll say that a data scientist is someone who extracts insights from messy data. Today’s world is full of people trying to turn data into insight.

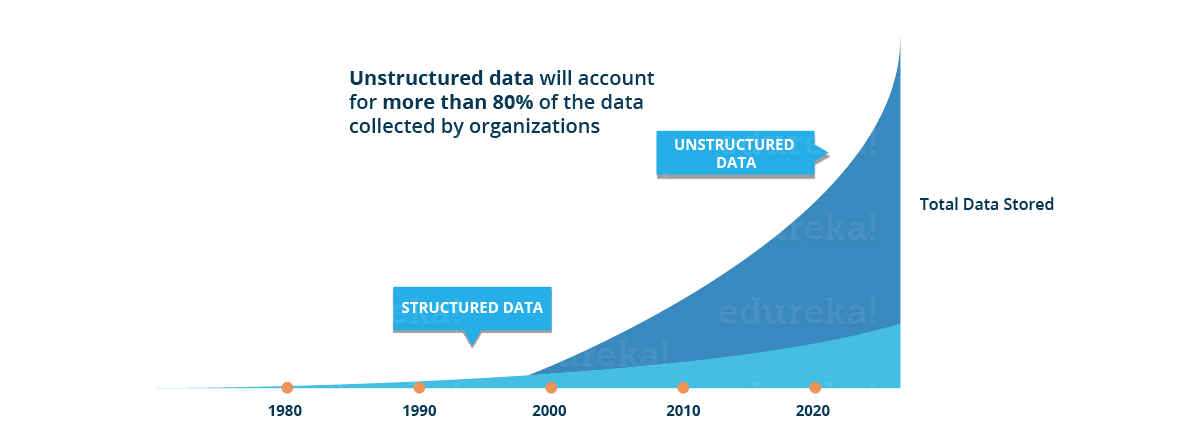
**1.3 Let’s Understand Why We Need Data Science**

Facebook asks you to list your hometown and your current location, ostensibly to make it easier for your friends to find and connect with you. But it also analyzes these locations to identify global migration patterns and where the fanbases of different football teams live.

As a large retailer, Target tracks your purchases and interactions, both online and in-store. And it uses the data to predictively model which of its customers are pregnant, to better market baby-related purchases to them.

In 2012, the Obama campaign employed dozens of data scientists who data-mined and experimented their way to identifying voters who needed extra attention, choosing optimal donor-specific fundraising appeals and programs, and focusing get-out-the-vote efforts where they were most likely to be useful. It is generally agreed that these efforts played an important role in the president’s re-election, which means it is a safe bet that political campaigns of the future will become more and more data-driven, resulting in a neverending arms race of data science and data collection.

Traditionally, the data that we had was mostly structured and small in size, which could be analyzed by using the simple BI tools. Unlike data in the traditional systems which was mostly structured, today most of the data is unstructured or semi-structured. Let’s have a look at the data trends in the image given below which shows that by 2020, more than 80 % of the data will be unstructured.



**Fig 1.2: Graph of unstructured data**

This data is generated from different sources like financial logs, text files, multimedia forms, sensors, and instruments. Simple BI tools are not capable of processing this huge volume and variety of data. This is why we need more complex and advanced analytical tools and algorithms for processing, analyzing and drawing meaningful insights out of it.

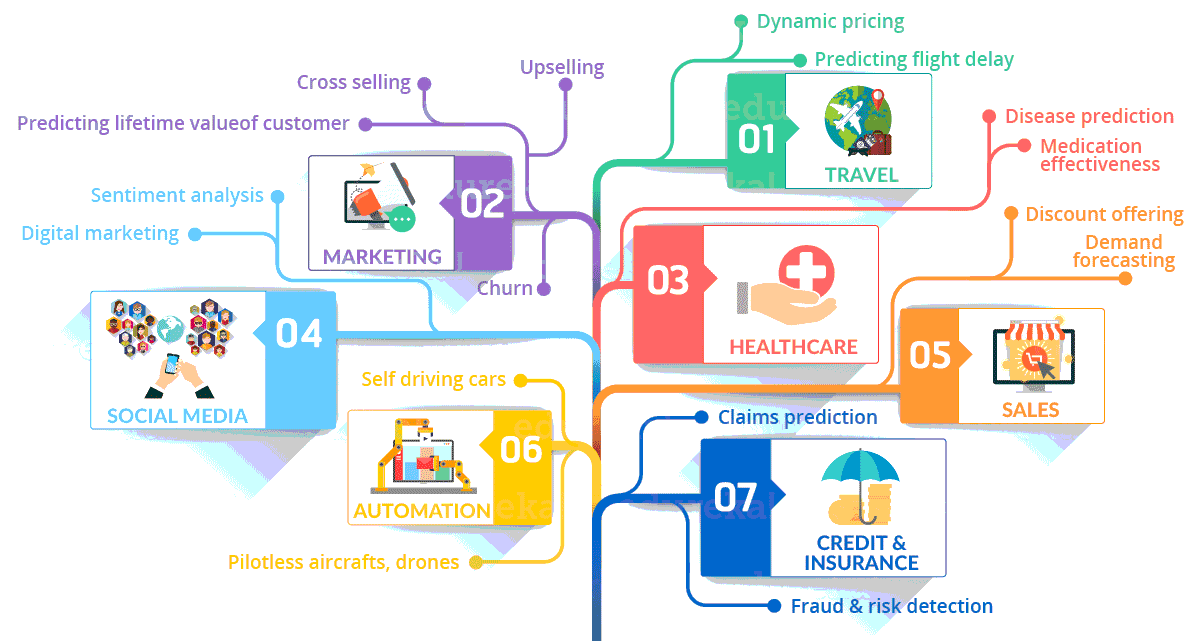
This is not the only reason why Data Science has become so popular. Let’s dig deeper and see how Data Science is being used in various domains.

How about if you could understand the precise requirements of your customers from the existing data like the customer’s past browsing history, purchase history, age and income. No doubt you had all this data earlier too, but now with the vast amount and variety of data, you can train models more effectively and recommend the product to your customers with more precision. Wouldn’t it be amazing as it will bring more business to your organization?

Let’s take a different scenario to understand the role of Data Science in decision making. How about if your car had the intelligence to drive you home? The self-driving cars collect live data from sensors, including radars, cameras and lasers to create a map of its surroundings. Based on this data, it takes decisions like when to speed up, when to speed down, when to overtake, where to take a turn – making use of advanced machine learning algorithms.

Let’s see how Data Science can be used in predictive analytics. Let’s take weather forecasting as an example. Data from ships, aircrafts, radars, satellites can be collected and analyzed to build models. These models will not only forecast the weather but also help in predicting the occurrence of any natural calamities. It will help you to take appropriate measures beforehand and save many precious lives.

Let’s have a look at the below infographic to see all the domains where Data Science is creating its impression.



**Fig 1.3: Domains of data science**

**2. VISUALIZING DATA AND DATABASE**

**2.1 DATA VISUALIZATION**

Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines or bars) contained in graphics. The goal is to communicate information clearly and efficiently to users.  According to Friedman (2008) the "main goal of data visualization is to communicate information clearly and effectively through graphical means. It doesn't mean that data visualization needs to look boring to be functional or extremely sophisticated to look beautiful. To convey ideas effectively, both aesthetic form and functionality need to go hand in hand, providing insights into a rather sparse and complex data set by communicating its key-aspects in a more intuitive way.

Data visualization involves specific terminology, some of which is derived from statistics. For example, author Stephen Few defines two types of data, which are used in combination to support a meaningful analysis or visualization:

* **Categorical:** Text labels describing the nature of the data, such as "Name" or "Age". This term also covers qualitative (non-numerical) data.
* **Quantitative:** Numerical measures, such as "25" to represent the age in years.

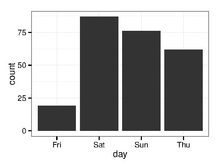
Two primary types of Information displays are tables and graphs.

* A ***table***contains quantitative data organized into rows and columns with categorical labels. It is primarily used to look up specific values. In the example above, the table might have categorical column labels representing the name (a *qualitative variable*) and age (a *quantitative variable*), with each row of data representing one person (the sampled *experimental unit* or *category subdivision*).
* A ***graph***is primarily used to show relationships among data and portrays values encoded as *visual objects* (e.g., lines, bars, or points). Numerical values are displayed within an area delineated by one or more *axes*. These axes provide *scales* (quantitative and categorical) used to label and assign values to the visual objects. Many graphs are also referred to as *charts*.

**2.2 TYPES OF DATA VISUALIZATION**

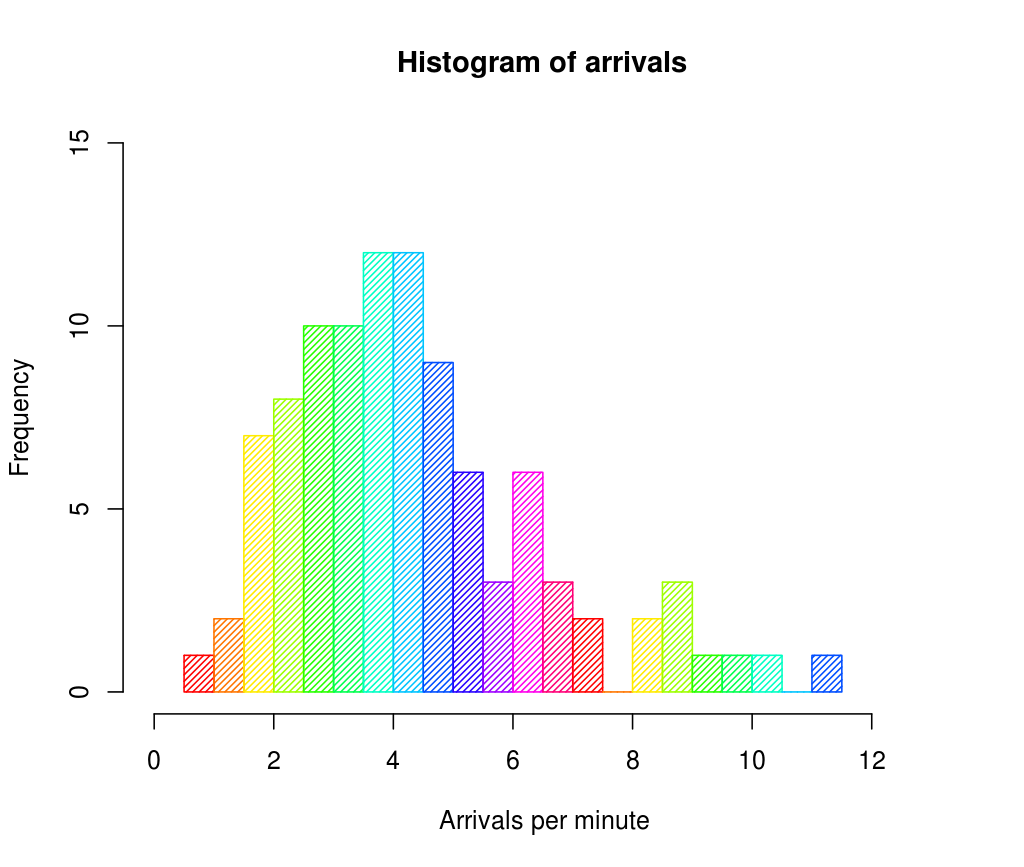
* + 1. **BAR CHART**

A **bar chart** or **bar graph** is a chart or graph that presents a with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a **line graph**.



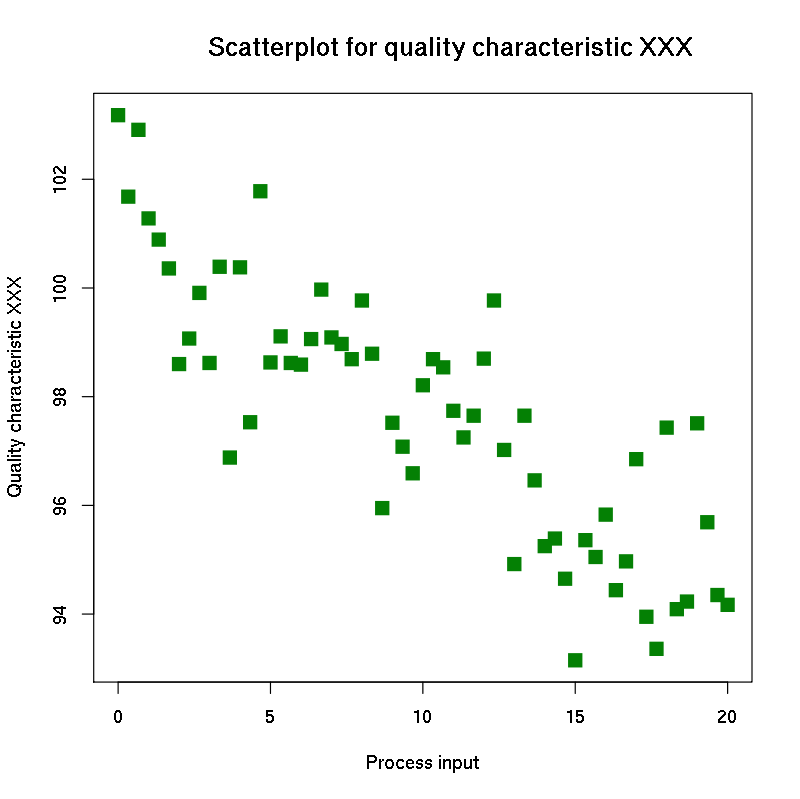
**2.2.2 HISTOGRAM**

A **histogram** is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable and was first introduced by Karl Pearson.



**2.2.3 SCATTER PLOT**

A **scatter plot** (also called a **scatterplot**, **scatter graph**, **scatter chart**, **scattergram**, or **scatter diagram**) is a type of plot or mathematical diagram using Cartesians Coordinates to display values for typically two variables for a set of data.



**2.3 DATABASE AND SQL**

**2.3.1 IMPORTANCE OF DATABASE AND SQL IN DATA SCIENCE**

Data Science is the study and analysis of data. In order to analyse the data, we need to extract it from the database. This is where SQL comes into the picture. Relational Database Management is an important part of Data Science. While many modern industries have geared their product management with NoSQL, SQL remains the ideal choice for many CRM, business intelligence tools and in office operations.Many database platforms are modelled after SQL. This is because it has become a standard for many database systems.

**2.3.2 WHAT IS DATABASE?**

A **database** is an organized collection of data, generally stored and accessed electronically from a computer system. Where databases are more complex they are often developed using formal design and modelling techniques.The database management system (DBMS) is the software that interacts with end users, applications, and the database itself to capture and analyse the data. The DBMS software additionally encompasses the core facilities provided to

**2.3.3 STRUCTURED QUERY LANGUAGE:**

**S**tructured **Q**uery **L**anguage or **SQL** is a standard Database language which is used to create, maintain and retrieve the data from relational databases like MySQL, Oracle, SQL Server, PostGre, etc. The recent ISO standard version of SQL is SQL:2019.As the name suggests, it is used when we have structured data (in the form of tables). All databases that are not relational (or do not use fixed structure tables to store data) and therefore do not use SQL, are called NoSQL databases. Examples of NoSQL are MongoDB, DynamoDB, Cassandra, etc.

* **DDL(Data Definition Language):**DDL or Data Definition Language actually consists of the SQL commands that can be used to define the database schema. It simply deals with descriptions of the database schema and is used to create and modify the structure of database objects in the database.

**Examples of DDL commands:**

[**CREATE**](https://www.geeksforgeeks.org/sql-create/) **–** is used to create the database or its objects (like table, index, function).

[**DROP**](https://www.geeksforgeeks.org/sql-drop-truncate/) **–** is used to delete objects from the database.

[**ALTER**](https://www.geeksforgeeks.org/sql-alter-add-drop-modify/)**-**is used to alter the structure of the database.

[**TRUNCATE**](https://www.geeksforgeeks.org/sql-drop-truncate/)**–**is used to remove all records from a table, including all spaces allocated for the records are removed.

[**COMMENT**](https://www.geeksforgeeks.org/sql-comments/) **–**is used to add comments to the data dictionary.

[**RENAME**](https://www.geeksforgeeks.org/sql-alter-rename/)**–**is used to rename an object existing in the database.

* **DQL (Data Query Language):**

DML statements are used for performing queries on the data within schema objects. The purpose of DQL Command is to get some schema relation based on the query passed to it.

**Example of DQL:**

[**SELECT**](https://www.geeksforgeeks.org/sql-select-clause/) **–** is used to retrieve data from the database.

* **DML(Data Manipulation Language):**The SQL commands that deals with the manipulation of data present in the database belong to DML or Data Manipulation Language and this includes most of the SQL statements.

**Examples of DML:**

[**INSERT**](https://www.geeksforgeeks.org/sql-insert-statement/) **–** is used to insert data into a table.

[**UPDATE**](https://www.geeksforgeeks.org/sql-update-statement/) **–** is used to update existing data within a table.

[**DELETE**](https://www.geeksforgeeks.org/sql-delete-statement/) **–** is used to delete records from a database table.

* **DCL(Data Control Language):**DCL includes commands such as GRANT and REVOKE which mainly deals with the rights, permissions and other controls of the database system.

**Examples of DCL commands:**

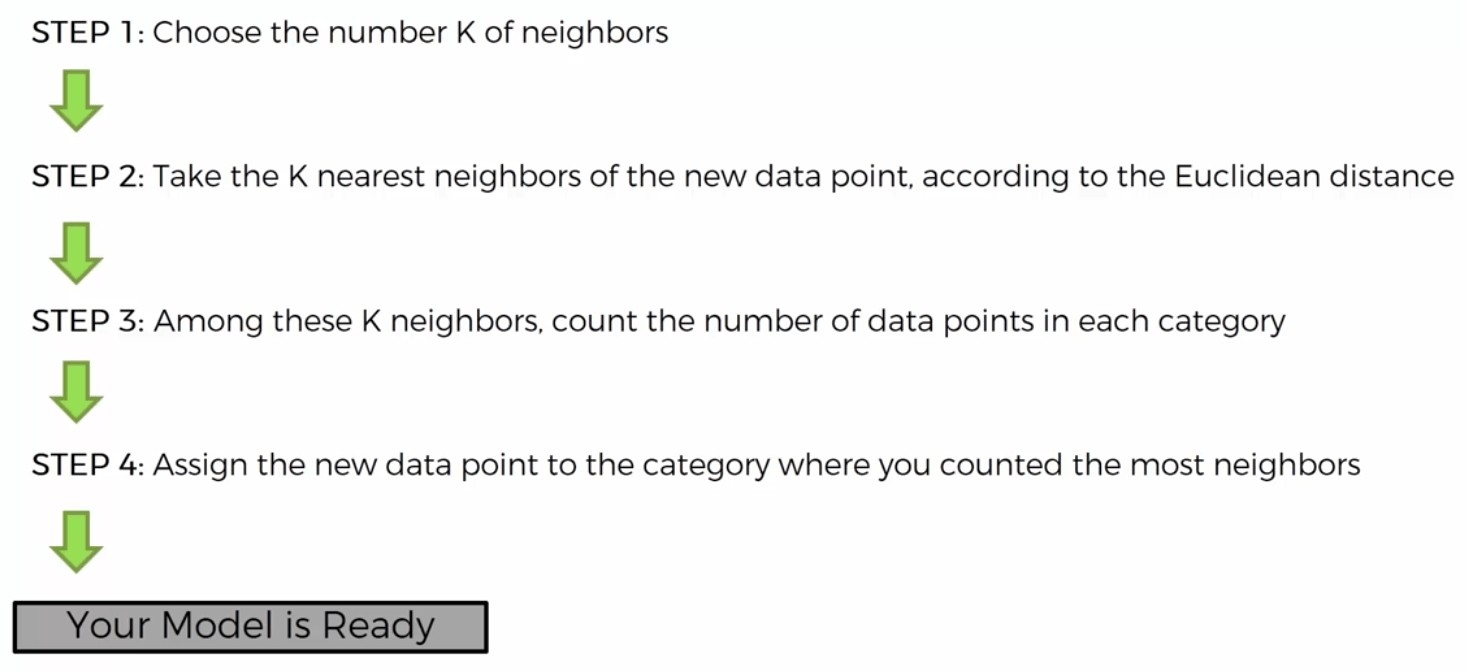
**GRANT**-gives user’s access privileges to database.

**REVOKE**-withdraw user’s access privileges given by using the GRANT command.

**3. ALGORITHMS**

**3.1 K – Nearest Neighbour**

Imagine that you’re trying to predict how I’m going to vote in the next presidential election. If you know nothing else about me (and if you have the data), one sensible approach is to look at how my neighbours are planning to vote. Living in downtown Seattle, as I do, my neighbours are invariably planning to vote for the Democratic candidate, which suggests that “Democratic candidate” is a good guess for me as well. Now imagine you know more about me than just geography — perhaps you know my age, my income, how many kids I have, and so on. To the extent my behaviour is influenced (or characterized) by those things, looking just at my neighbours who are close to me among all those dimensions seems likely to be an even better predictor than looking at all my neighbours. This is the idea behind nearest neighbour classification. Nearest neighbour is one of the simplest predictive models there is. It makes no mathematical assumptions, and it doesn’t require any sort of heavy machinery.



**Fig 3.1 : Working of k-Nearest Neighbour Algorithm**

The only things it requires are:

* Some notion of distance.
* An assumption that points that are close to one another are similar.

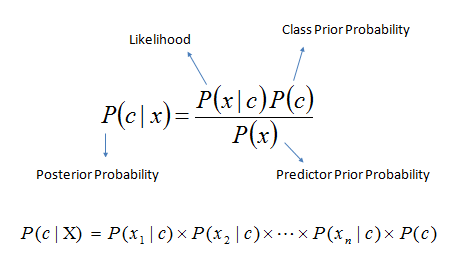
What’s more, nearest neighbour is probably not going to help you understand the drivers of whatever phenomenon you’re looking at. Predicting my votes based on my neighbours’ votes doesn’t tell you much about what causes me to vote the way I do, whereas some alternative model that predicted my vote based on (say) my income and marital status very well might. In the general situation, we have some data points and we have a corresponding set of labels. The labels could be True and False, indicating whether each input satisfies some condition like “is spam?” or “is poisonous?” or “would be enjoyable to watch?” Or they could be categories, like movie ratings (G, PG, PG-13, R, NC-17). Or they could be the names of presidential candidates. Or they could be favourite programming languages.

**3.2 Naive Bayes**

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the colour, roundness, and diameter features.

For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers. Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests. An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification.

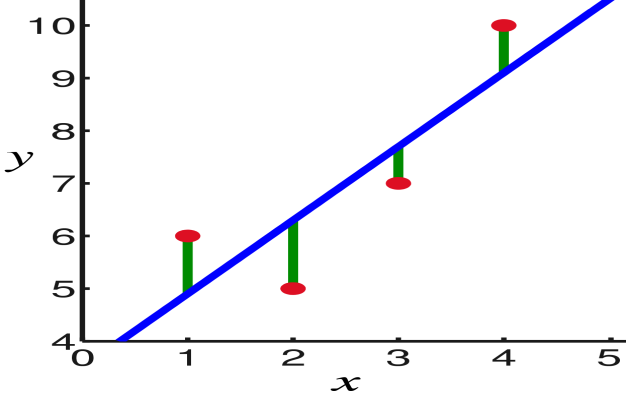


**Fig 3.2 : Naïve Bayes Algorithm**

**3.3 Linear Regression**

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variablesor independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.  Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

In linear regression, the observations (**red**) are assumed to be the result of random deviations (**green**) from an underlying relationship (**blue**) between a dependent variable (*y*) and an independent variable (*x*).

**Fig 3.3 : Graph of linear regression**

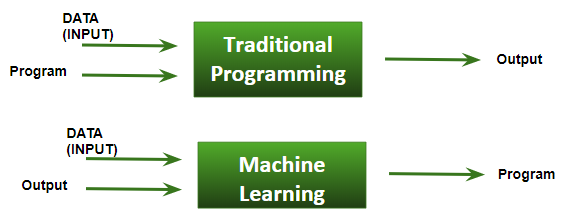
Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

* If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
* If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables.

**4. MACHINE LEARNING**

**4.1 An introduction to Machine Learning:**

The term Machine Learning was coined by Arthur Samuel in 1959, an American pioneer in the field of computer gaming and artificial intelligence and stated that “it gives computers the ability to learn without being explicitly programmed”.  
And in 1997, Tom Mitchell gave a “well-posed” mathematical and relational definition that “A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E”.



**Fig 4.1 : Basic Difference in ML and Traditional Programming**

* **Traditional Programming:**We feed in DATA (Input) + PROGRAM (logic), run it on machine and get output.
* **Machine Learning:**We feed in DATA(Input) + Output, run it on machine during training and the machine creates its own program(logic), which can be evaluated while testing.

**4.2 What does exactly learning means for a computer?**

A computer program is said to learn from experience **E**with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**

**Example:** playing checkers.  
**E** = the experience of playing checkers.  
**T** = the task of playing checkers.  
**P** = the probability that the program will win the next game

## **4.3 How Does Machine Learning Work?**

To get the maximum value from big data, businesses must know exactly how to pair the right algorithm with a particular tool or process and build machine learning models based on iterative learning processes. Some of the key machine learning algorithms are –

1. Random forests 2. Neural networks

3. Discovery of sequence and associations 4. Decision trees

5. Mapping of nearest neighbour 6. Supporting vector machines

7. Boosting and Bagging gradient 8. Self-organizing maps

9. Multivariate adaptive regression 10. Analysis of principal components

As mentioned above, the secret to successfully harnessing the applications of ML lies in not just knowing the algorithms, but in pairing them accurately with the right tools and processes, which include -

* Data exploration followed by visualization of model results
* Overall data quality and management
* Easy model deployment to quickly get reliable and repeatable results
* Developing graphical user interface for creating process flows and building models
* Comparing various machine learning models and identifying the best
* Identify best performers through automated ensemble model evaluation
* Automated data-to-decision process

**4.4 Types of machine learning problems:**

**4.4.1 On basis of the nature of the learning “signal” or “feedback” available to a learning system**

**Supervised learning**: The computer is presented with example inputs and their desired outputs, given by a “teacher”, and the goal is to learn a general rule that maps inputs to outputs. The training process continues until the model achieves the desired level of accuracy on the training data. Some real-life examples are:

* **Image Classification:**You train with images/labels. Then in the future you give a new image expecting that the computer will recognize the new object.
* **Market Prediction/Regression:**You train the computer with historical market data and ask the computer to predict the new price in the future.

**Unsupervised learning**: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. It is used for clustering population in different groups. Unsupervised learning can be a goal in itself (discovering hidden patterns in data).

* **Clustering:** You ask the computer to separate similar data into clusters, this is essential in research and science.
* **High Dimension Visualization:** Use the computer to help us visualize high dimension data.
* **Generative Models:** After a model captures the probability distribution of your input data, it will be able to generate more data. This can be very useful to make your classifier more robust.

**Semi-supervised learning**: Problems where you have a large amount of input data and only some of the data is labelled, are called semi-supervised learning problems. These problems sit in between both supervised and unsupervised learning. For example, a photo archive where only some of the images are labelled, (e.g. dog, cat, person) and the majority are unlabelled.

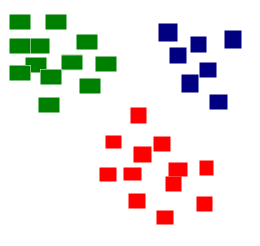
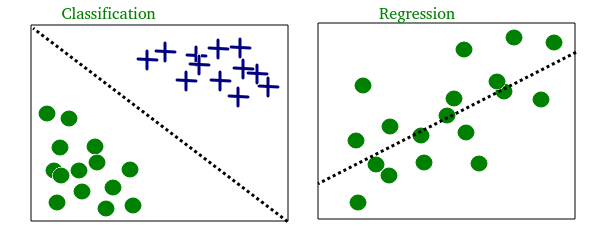
**Reinforcement learning**: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). The program is provided feedback in terms of rewards and punishments as it navigates its problem space.

* + 1. **On the basis of “output” desired from a machine learned system**

**Classification**: Inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are “spam” and “not spam”.

**Regression**: It is also a supervised learning problem, but the outputs are continuous rather than discrete. For example, predicting the stock prices using historical data.

**Clustering:** When a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.



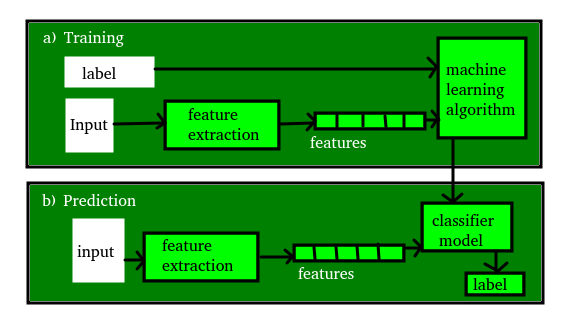
**Fig 4.2 :Classification, Regression and Clustering**

**4.5 Terminologies of Machine Learning:**

* **Model**  
  A model is a specific representation learned from data by applying some machine learning algorithm. A model is also called hypothesis.
* **Feature**  
  A feature is an individual measurable property of our data. A set of numeric features can be conveniently described by a feature vector. Feature vectors are fed as input to the model. For example, in order to predict a fruit, there may be features like colour, smell or taste.  
  **Note:** Choosing informative, discriminating and independent features is a crucial step for effective algorithms. We generally employ a feature extractor to extract the relevant features from the raw data.
* **Target**

A target variable or label is the value to be predicted by our model. For the fruit example discussed in the features section, the label with each set of input would be the name of the fruit like apple, orange, banana, etc.

* **Training**  
  The idea is to give a set of inputs(features) and it’s expected outputs(labels), so after training, we will have a model (hypothesis) that will then map new data to one of the categories trained on.
* **Prediction**  
  Once our model is ready, it can be fed a set of inputs to which it will provide a predicted output(label).



**Fig 4.3 : Machine Learning Terminology**

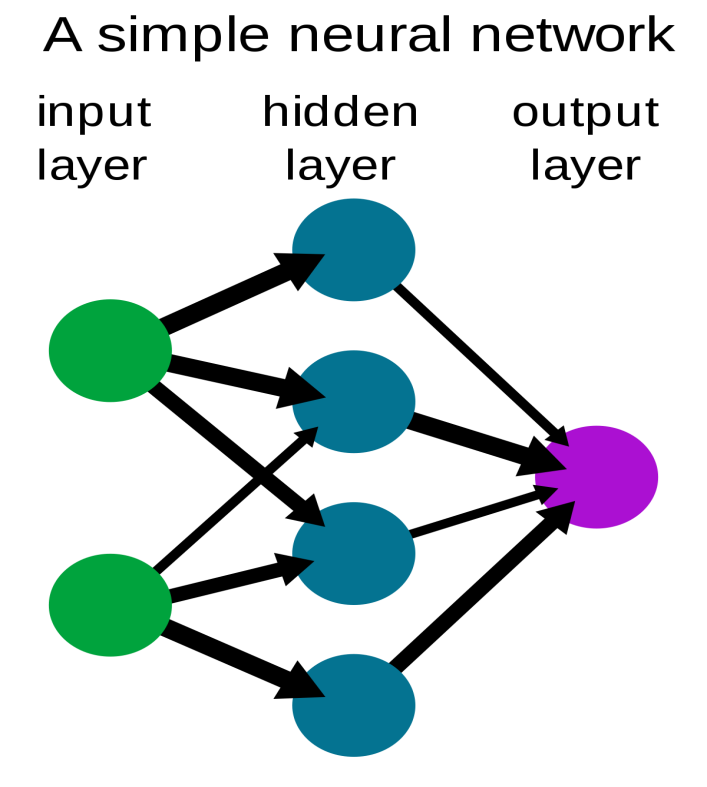
We probably use a learning algorithm, dozens of time without even knowing it. Applications of Machine Learning include:

* Web Search Engine
* Photo tagging Applications
* Spam Detector
* Financial Service
* Marketing and Sales
* Government
* Healthcare

**5. NEURAL NETWORKS**

**5.1 What are neural networks?**

An artificial neural network (or neural network for short) is a predictive model motivated by the way the brain operates. **Artificial Neural Networks** (ANNs) or connectionist systems are computing systems vaguely inspired by the biological neural networks that constitute animal brains. Such systems “learn” to perform tasks by considering examples, generally without being programmed with any task-specific rules. Think of the brain as a collection of neurons wired together. Each neuron looks at the outputs of the other neurons that feed into it, does a calculation, and then either fires (if the calculation exceeds some thresh‐ hold) or doesn’t. Artificial neural networks consist of artificial neurons, which perform similar calculations over their inputs. Neural networks can solve a wide variety of problems like handwriting recognition and face detection, and they are used heavily in deep learning, one of the trendiest subfields of data science.



**Fig 5.1: Simplified view of a feedforward ANN**

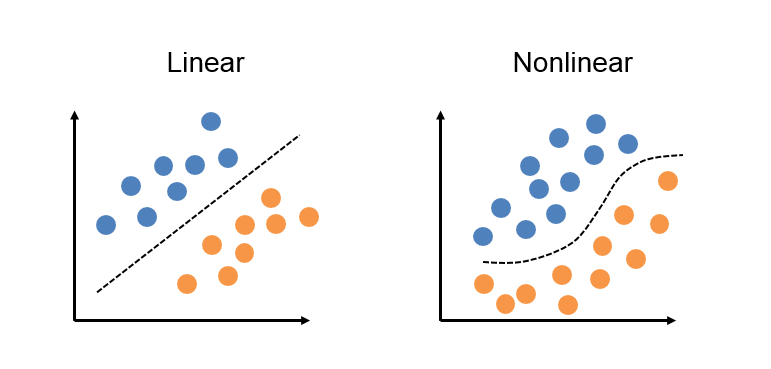
The architecture of an Artificial Neural Network (ANN) is a set of connected neurons organized in layers that are described as follows:

* Input Layer: Brings the initial data into the system for further processing by subsequent layers of artificial neurons.
* Hidden Layer: A layer in between input layers and output layers, where artificial neurons take in a set of weighted inputs and produce an output through an activation function.
* Output Layer: The last layer of neurons that produces given outputs for the program.

# 5.2 Types of Artificial Neural Networks:

# Perceptron:

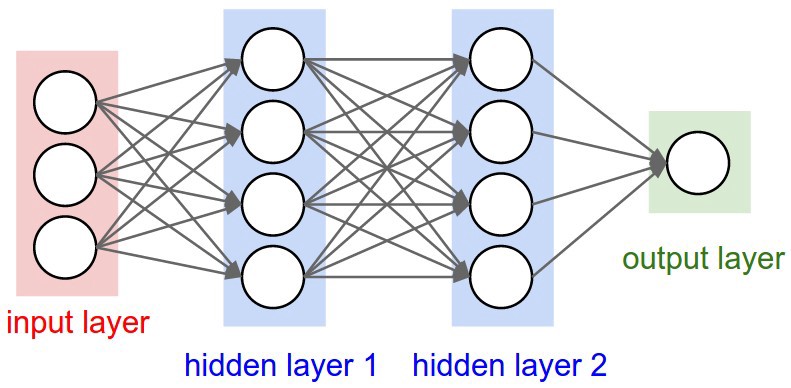
The simplest and oldest model of an ANN, the Perceptron is a linear classifier used for binary predictions. This means that in order for it to work, the data must be linearly separable.



**Fig 5.2: Linear vs Non-linear data**

# Multi-layer ANN:

More sophisticated than the perceptron, a Multi-layer ANN (e.g: Convolutional Neural Networks, Recurrent Neural Networks, etc) is capable of solving more complex classification and regression tasks thanks to its hidden layer(s).



**Fig 5.3: Multi-layer ANN architecture**

**5.3 Activation Function:**

Activation function decides whether a neuron should be activated or not by calculating weighted sum and further adding bias with it. The purpose of the activation function is to **introduce non-linearity** into the output of a neuron. A neural network without an activation function is essentially just a linear regression model. The activation function does the non-linear transformation to the input making it capable to learn and perform more complex tasks.

The activation function is a mathematical “gate” in between the input feeding the current neuron and its output going to the next layer. It can be as simple as a step function that turns the neuron output on and off, depending on a rule or threshold. Or it can be a transformation that maps the input signals into output signals that are needed for the neural network to function.

The Activation Functions can be basically divided into 2 types:

1. Linear Activation Function - This function is a line or linear. Therefore, the output of the functions will not be confined between any range.

**Equation :**f(x) = x**Range :** (-infinity to +infinity)

1. **Non-linear Activation Function -**The Nonlinear Activation Functions are the most used activation functions. They allow back-propagation because they have a derivative function which is related to the inputs. They allow “stacking” of multiple layers of neurons to create a deep neural network.

**Types of Non**-linear Activation Functions :

* Sigmoid / Logistic
* TanH / Hyperbolic Tangent
* ReLU(Rectified Linear Unit)
* Leaky ReLU
* Softmax
* Swish

CHOOSING THE RIGHT ACTIVATION FUNCTION:

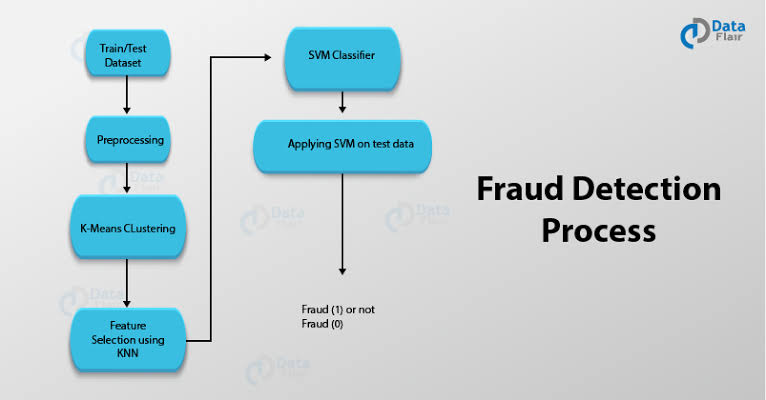
* The basic rule of thumb is if you really don’t know what activation function to use, then simply use *RELU* as it is a general activation function and is used in most cases these days.
* If your output is for binary classification, then *sigmoid function* is a very natural choice for output layer.

**6.APPLICATIONS OF DATA SCIENCE**

Data Science has dominated almost all the industries of the world today. There is no industry in the world today that does not use data. As such, data science has become fuel for industries. There are various industries like banking, finance, manufacturing, transport, e-commerce, education, etc. that use data science. As a result, there are several Data Science Applications related to it.

**6.1 Banking**

Banking is one of the biggest applications of Data Science. Big Data and Data Science have enabled banks to keep up with the competition. With Data Science, banks can manage their resources efficiently, furthermore, banks can make smarter decisions through fraud detection, management of customer data, risk modelling, real-time predictive analytics, customer segmentation, etc.Banks also assess the customer lifetime value that allows them to monitor the number of customers that they have. It provides them with several predictions that the business bank will derive through their customers. In case of fraud detection, banks allow the companies to detect frauds that involve a credit card, insurance, and accounting. Banks are also able to analyse investment patterns and cycles of customers and suggest you several offers that suit you accordingly.In real-time and predictive analytics, banks use Machine learning algorithms to improve their analytics strategy. Furthermore, banks use real-time analytics to understand underlying problems that impede their performance.



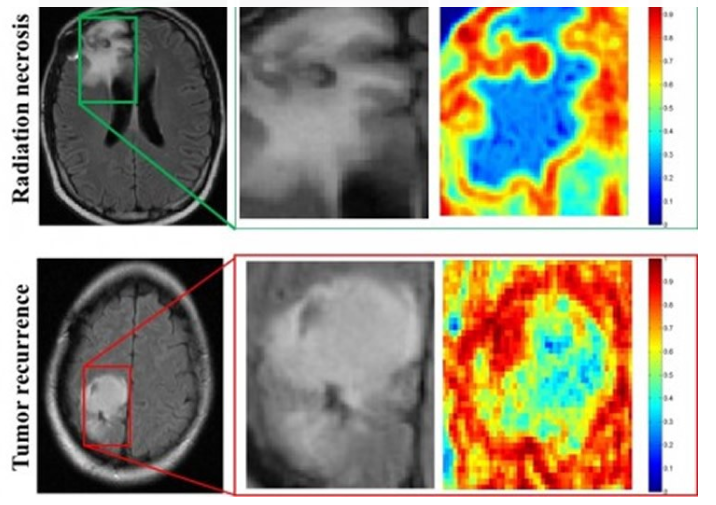
**Fig 6.1: Fraud Detection Process**

### **6.2 Healthcare**

In the health-care industry, data science is making great leaps. The various industries in health-care making use of data science are

#### **i. Medical Image Analysis**

In the medical image analysis, data science has created a strong sphere of influence for analyzing medical images such as X-rays, MRIs, CT-Scans, etc. Previously, doctors and medical examiners would have to manually search for clues in the medical images. However, with the advancements in computing technologies and surge in data, it is possible to create machines that can automatically detect flaws in the imagery. Data Scientists have created powerful **image recognition tools** that allow doctors to have an in-depth understanding of complex medical imagery.



**Fig 6.2.i : Medical image Analysis**

#### **ii. Genomic Data Science**

Genomic Data Science applies the statistical techniques to genomic sequences, allowing the bioinformaticians and geneticists to understand the defects in genetic structures. It is also helpful in classifying diseases that are genetic in nature. With data science, we can analyze how genes react to varying kinds of medicines. Also, several big data technologies like MapReduce have significantly reduced the processing time for genome sequencing.



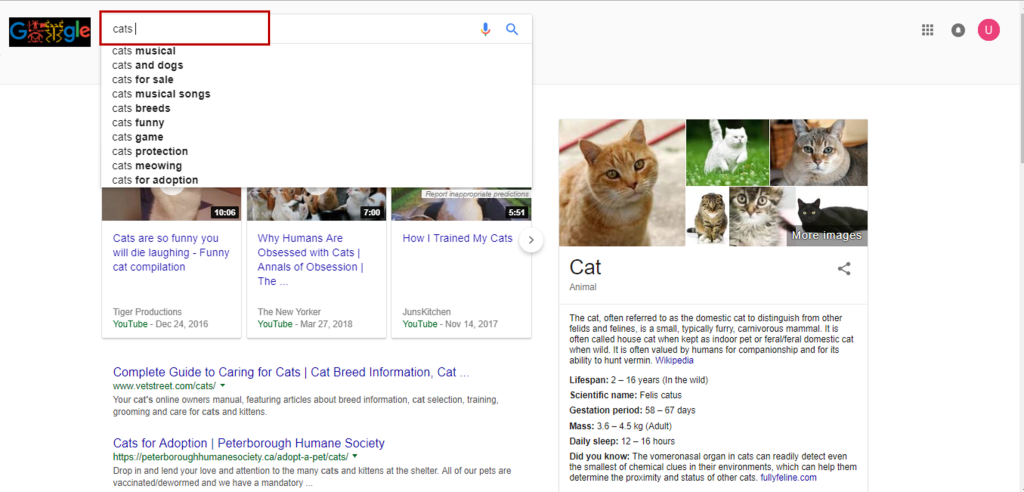
**Fig 6.2.ii: Genomics**

#### **iii. Drug Discovery**

Another important field making use of data science is drug discovery. In drug discovery, new candidate medicines are formulated. Drug Discovery is a tedious and often complex process. Data Science can help us to simplify this process and provide us with an early insight into the success rate of the newly discovered drug. With Machine Learning, we can also analyse several combinations of drugs and their effect on different gene structure to predict the outcome.Data science applications and machine learning algorithms simplify and shorten this process, adding a perspective to each step from the initial screening of drug compounds to the prediction of the success rate based on the biological factors. Such algorithms can forecast how the compound will act in the body using advanced mathematical modelling and simulations instead of the “lab experiments”.

**6.3 Internet Search**

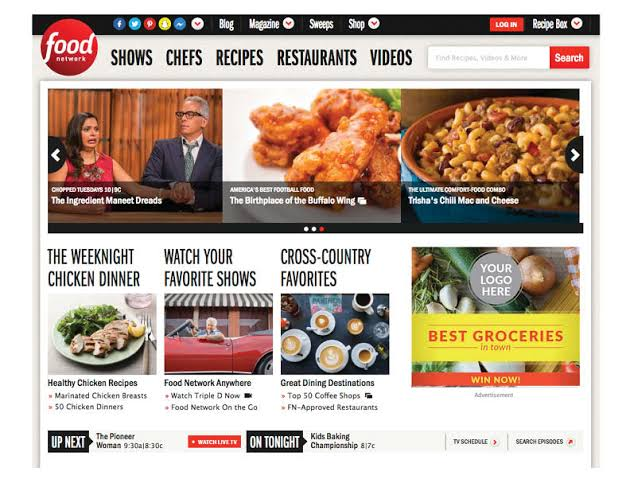
Now, this is probably the first thing that strikes your mind when you think Data Science Applications.When we speak of search, we think ‘Google’. But there are many other search engines like Yahoo, Bing, Ask, AOL, and so on. All these search engines (including Google) make use of data science algorithms to deliver the best result for our searched query in a fraction of seconds. Considering the fact that, Google processes more than 20 petabytes of data every day.



**Fig 6.3: Internet Search**

## **6.4 Targeted Advertising**

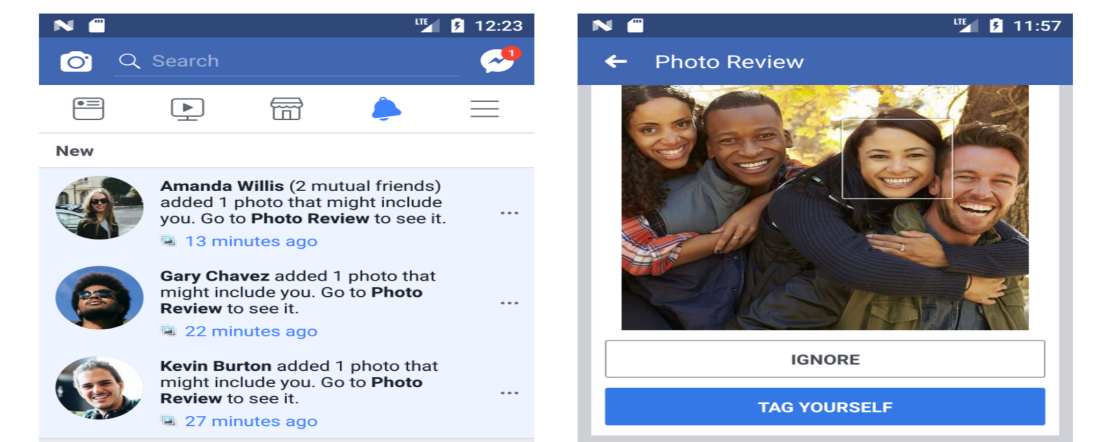
If you thought Search would have been the biggest of all data science applications, here is a challenger – the entire digital marketing spectrum. Starting from the display banners on various websites to the digital billboards at the airports – almost all of them are decided by using data science algorithms.This is the reason why digital ads have been able to get a lot higher CTR (Call-Through Rate) than traditional advertisements. They can be targeted based on a user’s past behaviour.This is the reason why you might see ads of Data Science Training Programs while I see an ad of apparels in the same place at the same time.



**Fig 6.4: Targeted Advertisements**

**6.5 Advance Image Recognition**

Image recognition will continue to help society in areas such as the automobile industry, medical industry, and manufacturing industry. A vehicles ability to recognize an object in the road and automatically apply the brakes, or a factory machine that can recognize a defect, is in part due to the advancement in technology in the field of image recognition. The technology leads to a computer’s ability to not only locate objects, but accurately identify objects within a photo. Over the past several years the technology and potential for real world application of image recognition has seen a tremendous increase in accuracy. You upload your image with friends on Facebook and you start getting suggestions to tag your friends. This automatic tag suggestion feature uses face recognition algorithm. n their latest update, Facebook has outlined the additional progress they’ve made in this area, making specific note of their advances in image recognition accuracy and capacity.



**Fig 6.4:ImageRecognition**

**CONCLUSION**

Data Science has become a revolutionary technology that everyone seems to talk about.It is a [multi-disciplinary](https://en.wikipedia.org/wiki/Multi-disciplinary) field that uses scientific methods, processes, algorithms and systems to extract [knowledge](https://en.wikipedia.org/wiki/Knowledge) and insights from structured and unstructured [data](https://en.wikipedia.org/wiki/Data). Data science is a concept to unify [statistics](https://en.wikipedia.org/wiki/Statistics), [data analysis](https://en.wikipedia.org/wiki/Data_analysis), [machine learning](https://en.wikipedia.org/wiki/Machine_learning)and their related methods in order to understand and analyse actual phenomena with data. This field uses the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems.It has made the computing faster and storage has also become cheaper. We can now predict outcomes in minutes, what could take several human hours to process. Data Science has helped various industries to automate redundant tasks. This is a vastly abundant field and a versatile technology and also there are **numerous applications of Data Science. This** **is a field with many lucrative advantages and**is an ever-evolving field that will take years to gain proficiency.

**BIBLIOGRAPHY**

**1. INTRODUCTION**

* <https://en.wikipedia.org/wiki/Data_science>
* <https://www.edureka.co/blog/what-is-data-science/>
* <https://en.wikipedia.org/wiki/Data_mining>
* <https://en.wikipedia.org/wiki/Big_data>
* Joel Grus, 2015: Data Science from Scratch. United States of America: O’Reilly Media

**2. VISUALIZING DATA AND DATABASE**

* <https://en.wikipedia.org/wiki/Scatter_plot>
* <https://en.wikipedia.org/wiki/Data_visualization>
* <https://en.wikipedia.org/wiki/Database>
* <https://data-flair.training/blogs/sql-in-data-science/>
* <https://www.geeksforgeeks.org/sql-ddl-dql-dml-dcl-tcl-commands/>
* <https://www.w3schools.com/sql/sql_join.asp>

**4. MACHINE LEARNING**

* <https://www.expertsystem.com/machine-learning-definition/>
* [https://www.geeksforgeeks.org/machine-learning/](https://www.geeksforgeeks.org/ml-machine-learning/)
* <https://media.geeksforgeeks.org/wp-content/uploads/ML-vs-Programming.png>
* <https://media.geeksforgeeks.org/wp-content/uploads/unsupervised_learning-.png>
* <https://media.geeksforgeeks.org/wp-content/uploads/classification_regression.png>
* <https://media.geeksforgeeks.org/wp-content/uploads/training.png>

**5. NEURAL NETWORKS**

* <https://en.wikipedia.org/wiki/Artificial_neural_network>
* <https://blog.goodaudience.com/artificial-neural-networks-explained-436fcf36e75>
* <https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6>
* <https://missinglink.ai/guides/neural-network-concepts/7-types-neural-network-activation-functions-right/>
* [https://geeksforgeeks.org/activation-functions-neural-networks/](https://www.geeksforgeeks.org/activation-functions-neural-networks/)

**6. APPLICATIONS OF DATA SCIENCE**

* <https://www.edureka.co/blog/data-science-applications/>
* <https://www.analyticsvidhya.com/blog/2015/09/applications-data-science/>