



# Data Science For Dummies

"It's easy to lie with statistics. It's hard to tell the truth  
without statistics."

JAIMIN PATEL



# Table of Content

## Lesson 1: - Introduction

- Introduction to Data Science
- About this Book

## Lesson 2: - Understanding Data Science

- What is Data Science ?
- How does it work ?
- Defining Big Data
- Types of analytics
- Mind map of Data Science
- Disciplines of Data Science

## Lesson 3: - Understanding Machine learning


- What is Machine learning ?
- Types of Machine Learning
  - Supervised Learning
  - Unsupervised Learning
  - Deep Learning
  - Artificial Neural Network
  - Reinforcement learning
- What is Computer Vision ?
- What is Natural Language Processing ?



## Lesson 4: - Understanding Data Mining

- What is Big Data ?
- What is Data Engineering ?
- What is Data Visualization?

## Lesson 5: - How to get started with Data Science ?

- Math and Statistics
  - Programming Language and coding
  - ML theory
  - Own projects
  - RoadMap
  - Conclusion
- 

# Introduction

## Lesson Outcomes:-

- A Brief Introduction to the field of Data Science.
- About this book

# Introduction

This book's viewpoint is that a data scientist is someone who asks unique, interesting questions of data based on formal or informal theory, to generate rigorous and useful insights. It is likely to be an individual with multi-disciplinary training in computer science, business, economics, statistics, and armed with the necessary quantity of domain knowledge relevant to the question at hand. The potential of the field is enormous for just a few well-trained data scientists armed with big data have the potential to transform organizations and societies. In the narrower domain of business life, the role of the data scientist is to generate applicable business intelligence.

If the past few years hasn't found you living on a desert island without electricity or communication with the outside world, you've likely heard about machine learning (ML). It's hard to miss the trend. Every time we talk about self-driving cars, chatbots, AlphaGo, or predictive analytics, we're discussing some implementation of machine learning techniques. While success stories and evangelists abound, machine learning hasn't become the obligatory for business yet. In the public's perception, algorithms that are applied in ML are close to science fiction, and rolling out a concrete plan for ML adoption is still a high hurdle.

Hence, this whitepaper is aimed at answering practical questions instead of setting the vision and evangelizing the trend. This is about an umbrella term data science and how its subfields interact, the main problems that machine learning can solve, and how these problems can be translated into the language of business. We will also contemplate the main decisions to make concerning talent acquisition and pinpoint the challenges to be considered in advance. Because we've covered data science's potential in articles dedicated to the travel and industries, we will only touch on it briefly today.

# About this book

Data Science for Dummies gives you insights into what Data Science is all about and how it can impact the way you can weaponize data to gain unimaginable insights. As Data is the big for future era. So, Your data is only as good as what you do with it and how you manage it. In this book, you will understand what is Data science, how it works, various disciplines of Data science including machine learning, Artificial intelligence & Deep learning that can help achieve results for your company. This information helps us how to apply Data science to anticipate and predict the future.

# Understanding Data Science

## Lesson Outcomes:-

- What is Data science ?
- How does it work ?
- Types of analysis
- Mind map of Data Science
- Displines of Data Science

# Understanding Data Science

## What is Data Science ?

Data Science is kind of blended with various tools, algorithms, and machine learning principles. Most simply, it involves obtaining meaningful information or insights from structured or unstructured data through a process of analysing, programming and business skills. It is a field containing many elements like mathematics, statistics, computer science, etc. Those who are good at these respective fields with enough knowledge of the domain in which you are willing to work can call themselves as Data Scientist. It's not an easy thing to do but not impossible too. You need to start from data, its visualization, programming, formulation, development, and deployment of your model. In the future, there will be great hype for data scientist jobs. Taking in that mind, be ready to prepare yourself to fit in this world.

## How does it work ?

Data science is not a one-step process such that you will get to learn it in a short time and call ourselves a Data Scientist. It's passes from many stages and every element is important. One should always follow the proper steps to reach the ladder. Every step has its value and it counts in your model. Buckle up in your seats and get ready to learn about those steps.



# Understanding Data Science

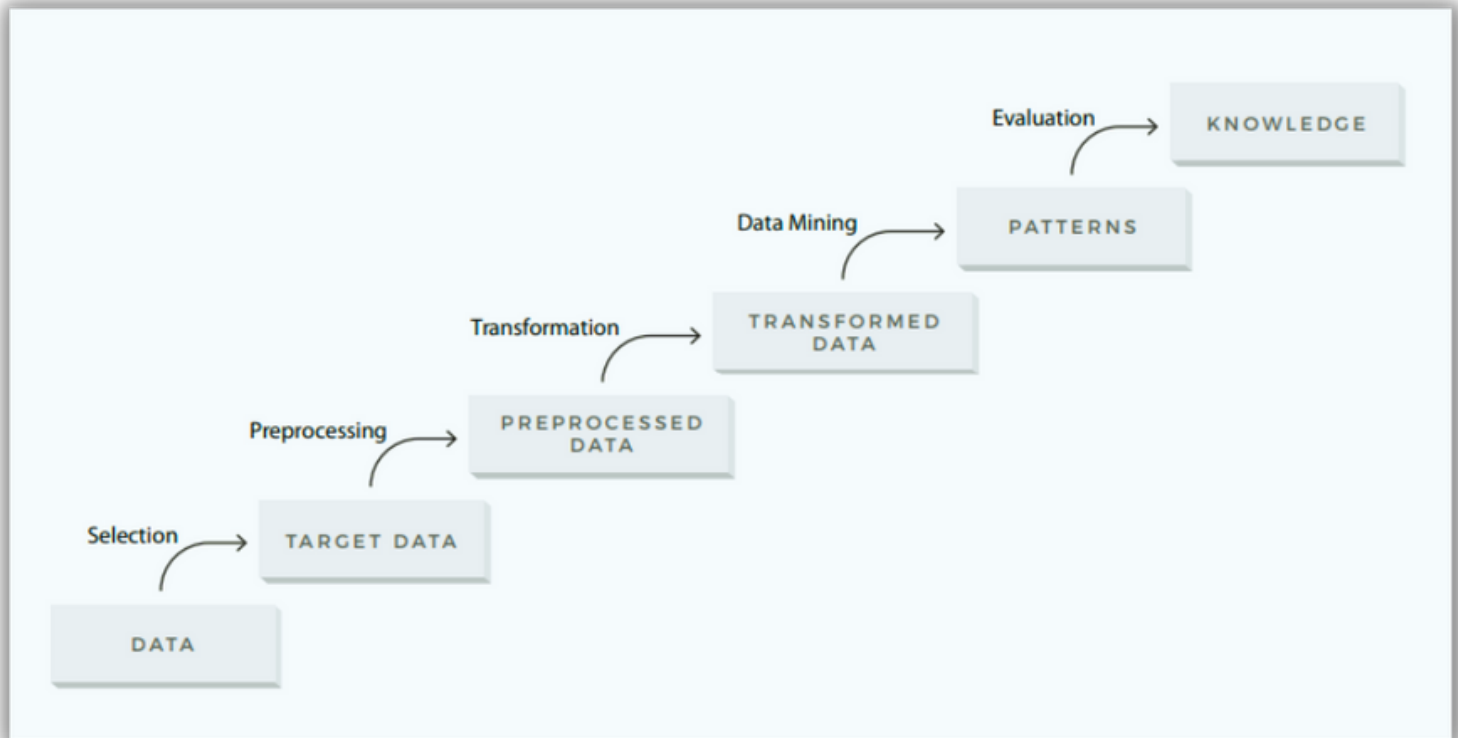


Fig.: Various steps of Data Science

- **Problem Statement:** No work starts without motivation; Data science is no exception though. It's really important to declare or formulate your problem statement very clearly and precisely. Your whole model and it's working depend on your statement. Many scientists consider this as the main and much important step of Date Science. So, make sure what's your problem statement and how well can it add value to business or any other organization.
- **Data Collection:** After defining the problem statement, the next obvious step is to go in search of data that you might require for your model. You must do good research, find all that you need. Data can be in any form i.e., unstructured or structured. It might be in various forms like videos, spreadsheets, coded forms, etc. You must collect all these kinds of sources.

# Understanding Data Science

- **Data Cleaning:** As you have formulated your motive and also you did collect your data, the next step to do is cleaning. Yes, it is! Data cleaning is the most favourite thing for data scientists to do. Data cleaning is all about the removal of missing, redundant, unnecessary and duplicate data from your collection. There are various tools to do so with the help of programming in either R or Python. It's totally on you to choose one of them. Various scientist has their opinion on which to choose. When it comes to the statistical part, R is preferred over Python, as it has the privilege of more than 12,000 packages. While python is used as it is fast, easily accessible and we can perform the same things as we can in R with the help of various packages.
- **Data Analysis and Exploration:** It's one of the prime things in data science to do and time to get inner Holmes out. It's about analysing the structure of data, finding hidden patterns in them, studying behaviours, visualizing the effects of one variable over others and then concluding. We can explore the data with the help of various graphs formed with the help of libraries using any programming language. In R, GGplot is one of the most famous models while Matplotlib in Python.
- **Data Modelling:** Once you are done with your study that you have formed from data visualization, you must start building a hypothesis model such that it may yield you a good prediction in future. Here, you must choose a good algorithm that best fit to your model. There different kinds of algorithms from regression to classification, SVM (Support vector machines), Clustering, etc. Your model can be of a Machine Learning algorithm. You train your model with the train data and then test it with test data. There are various methods to do so. One of them is the K-fold method where you split your whole data into two parts, one is Train and the other is test data. On these bases, you train your model.

# Understanding Data Science

- **Optimization and Deployment:** You followed each and every step and hence build a model that you feel is the best fit. But how can you decide how well your model is performing? This where optimization comes. You test your data and find how well it is performing by checking its accuracy. In short, you check the efficiency of the data model and thus try to optimize it for better accurate prediction. Deployment deals with the launch of your model and let the people outside there to benefit from that. You can also obtain feedback from organizations and people to know their need and then to work more on your model.

## Types of analytics

As, Data science is all about collecting and analysing a set of data to extract useful information from them. And use that information to carry out useful concussions and prediction. So, it become important for us to know what are the types of analysis will can perform on data. There are mainly 4 types of analytics: -

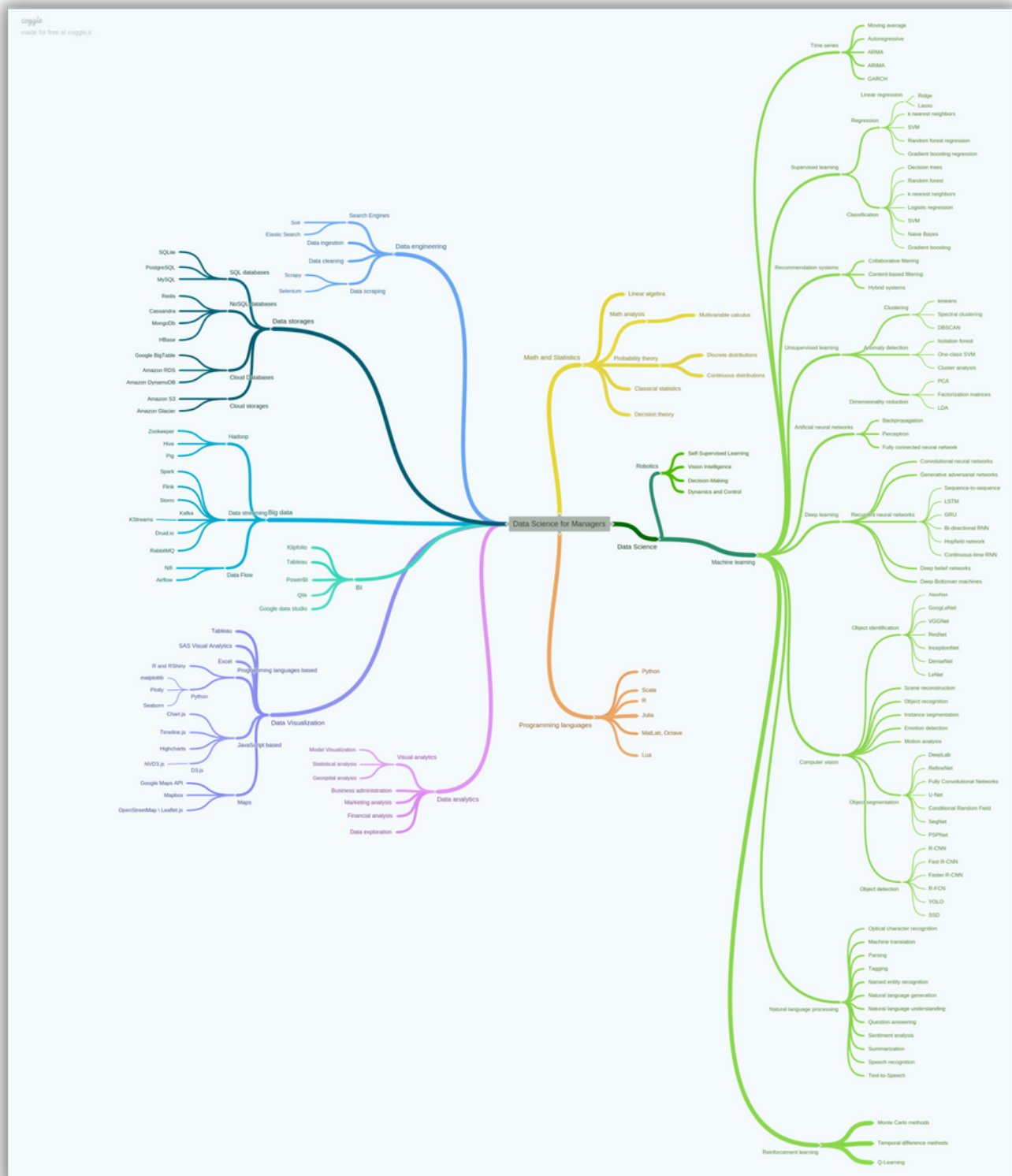
- **Descriptive Analytics:** -Descriptive analytics deals with past trends data, it basically finds out what has happened in the past, and based on past data or historic data it predicts the future outcome. One of the main objectives of descriptive analytics is to look at the trends of past data, summarize it in an innovative way that can be useful for generating insight.
- **Diagnostic Analytics:** -Diagnostic analysis works hand in hand with Descriptive analytics. As descriptive analytics find out what happened in the past, diagnostic analytics, on the other hand, finds out why did that happen or what measures were taken at that time, or how frequent it has happened.it basically gives a detailed explanation of a particular scenario by understanding behavioural patterns.

# Understanding Data Science

- **Predictive Analytics:** -Whatever information we have received from descriptive and diagnostic analytics, we can use that information to predict future data. It basically finds out what is likely to happen in the future. Now when I say future data doesn't mean we have become fortune-tellers, by looking at the past trends and behavioural patterns we are forecasting that it might happen in the future.
- **Prescriptive Analytics:** -This is an advanced method of Predictive analytics. Now when you predict something or when you start thinking out of the box you will definitely have a lot of options, and then we get confused as to which option will actually work. Prescriptive analytics helps to find which is the best option to make it happen or work. As predictive analytics forecast future data, Prescriptive analytics on the other hand helps to make it happen whatever we have forecasted. Prescriptive analytics is the highest level of analytics that is used for choosing the best optimal solution by looking at descriptive, diagnostic, and predictive data.

# Understanding Data Science

## Mind-Map of Data Science



# Understanding Data Science

## Discipline of Data Science

Summarizing the above figure, Data Science includes the following Disciplines: -

1. Math and statistics
2. Machine learning
  - a. Time series
  - b. Supervised learning
  - c. Recommendation systems
  - d. Unsupervised learning
  - e. Artificial neural network
  - f. Deep learning
  - g. Computer vision
  - h. Natural language processing
  - i. Reinforcement learning
3. Programming language
4. Data engineering
5. Data storage
6. Data streaming Big Data
7. Data visualization
8. Data analytics

In this book, we will discuss all the disciplines of the data science in greater depth in upcoming lessons.

# Understanding Machine Learning

## Lesson Outcomes:-

- What is Machine learning ?
- Types of learning
- What is Computer Vision ?
- What is NLP ?

# Understanding Machine Learning

## What is Machine Learning ?

Data science is also more than “machine learning,” which is about how systems learn from data. Systems may be trained on data to make decisions, and training is a continuous process, where the system updates its learning and (hopefully) improves its decision-making ability with more data. Machine learning has become one of the most important topics within development organizations that are looking for innovative ways to leverage data assets to help the business gain a new level of understanding. Why add machine learning into the mix? With the appropriate machine learning models, organizations have the ability to continually predict changes in the business so that they are best able to predict what’s next. As data is constantly added, the machine learning models ensure that the solution is constantly updated. The value is straightforward: If you use the most appropriate and constantly changing data sources in the context of machine learning, you have the opportunity to predict the future.

Machine learning is a form of AI that enables a system to learn from data rather than through explicit programming. However, machine learning is not a simple process.

Machine learning uses a variety of algorithms that iteratively learn from data to improve, describe data, and predict outcomes. As the algorithms ingest training data, it is then possible to produce more precise models based on that data. A machine learning model is the output generated when you train your machine learning algorithm with data. After training, when you provide a model with an input, you will be given an output. For example, a predictive algorithm will create a predictive model. Then, when you provide the predictive model with data, you will receive a prediction based on the data that trained the model. Machine learning is now essential for creating analytics models.



# Understanding Machine Learning

You likely interact with machine learning applications without realizing. For example, when you visit an e-commerce site and start viewing products and reading reviews, you're likely presented with other, similar products that you may find interesting. These recommendations aren't hard coded by an army of developers. The suggestions are served to the site via a machine learning model. The model ingests your browsing history along with other shoppers' browsing and purchasing data in order to present other similar products that you may want to purchase.

Machine intelligence is re-emerging as the new incarnation of AI (a field that many feel has not lived up to its promise). Machine learning promises and has delivered on many questions of interest, and is also proving to be quite a game-changer, as we will see later on in this chapter, and also as discussed in many preceding examples. What makes it so appealing? Hilary Mason suggests four characteristics of machine intelligence that make it interesting: (i) It is usually based on a theoretical breakthrough and is therefore well grounded in science. (ii) It changes the existing economic paradigm. (iii) The result is commoditization (e.g., Hadoop), and (iv) it makes available new data that leads to further data science.

# Understanding Machine Learning

## Types of learning

Machine learning techniques are required to improve the accuracy of predictive models. Depending on the nature of the business problem being addressed, there are different approaches based on the type and volume of the data. In this section, we discuss the categories of machine learning.

**1. Supervised learning:** -Supervised learning algorithms are used when the output is classified or labelled. These algorithms learn from the past data that is inputted, called training data, runs its analysis and uses this analysis to predict future events of any new data within the known classifications. The accurate prediction of test data requires large data to have a sufficient understanding of the patterns. The algorithm can be trained further by comparing the training outputs to actual ones and using the errors for modification of the algorithms.

List of algorithms in Supervised learning: -

- Regression
  - o Linear regression
  - o K nearest neighbours
  - o SVM (Support Vector Machine)
  - o Random forest regression
  - o Gradient boosting Regression
- Classification
  - o Decision trees
  - o Random forest
  - o K nearest neighbours
  - o Logistics regression
  - o SVM
  - o Navies Bayes

# Understanding Machine Learning

Real-Life Example:

- **Image Classification** – The algorithm is drawn from feeding with labelled image data. An algorithm is trained, and it is expected that the algorithm classifies it correctly in the case of the new image.
- **Market Prediction** – It is also called Regression. Historical business market data is fed to the computer. Then, with analysis and regression algorithm, the new price for the future is predicted depending on variables.

**2. Unsupervised learning:** - Unsupervised learning algorithms are used when we are unaware of the final outputs, and the classification or labelled outputs are not at our disposal. These algorithms study and generate a function to describe completely hidden and unlabeled patterns. Hence, there is no correct output, but it studies the data to give out unknown structures in unlabeled data.

List of algorithms in Supervised learning: -

- Clustering
  - o K-means
  - o Hierarchical clustering
  - o Spectral clustering
  - o DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- Anomaly Detection
  - o Isolation forest
  - o One-class SVM
  - o Cluster analysis
- Dimensionality reduction
  - o PCA (Principal component analysis)
  - o Factorization matrices
  - o Linear Discriminate Analysis

# Understanding Machine Learning

Real-Life Example:

- **Clustering** – Data with similar traits are asked to group together by the algorithm; this grouping is called clusters. These prove helpful in the study of these groups, which can be applied to the entire data within a cluster more or less.
- **High Dimension Data** – High dimension data is normally not easy to work with. With the help of unsupervised learning, visualization of high dimension data becomes possible.
- **Generative Models** – Once your algorithm analyses and comes up with the probability distribution of the input, it can be used to generate new data. This proves to be very helpful in cases of missing data.

**3. Reinforcement learning:** - This type of machine learning algorithm uses the trial-and-error method to churn out output based on the highest efficiency of the function. The output is compared to find out errors and feedback fed back to the system to improve or maximize its performance. The model is provided with rewards which are basically feedback and punishments in its operations while performing a particular goal.

List of algorithms of Reinforcement Learning: -

- Monte carlo methods
- Temporal difference methods
- Q-learning
- Multi-agent

# Understanding Machine Learning

**4. Artificial Neural Network:** - An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron receives signals then processes them and can signal neurons connected to it. The "signal" at a connection is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called edges. Neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons may have a threshold such that a signal is sent only if the aggregate signal crosses that threshold.

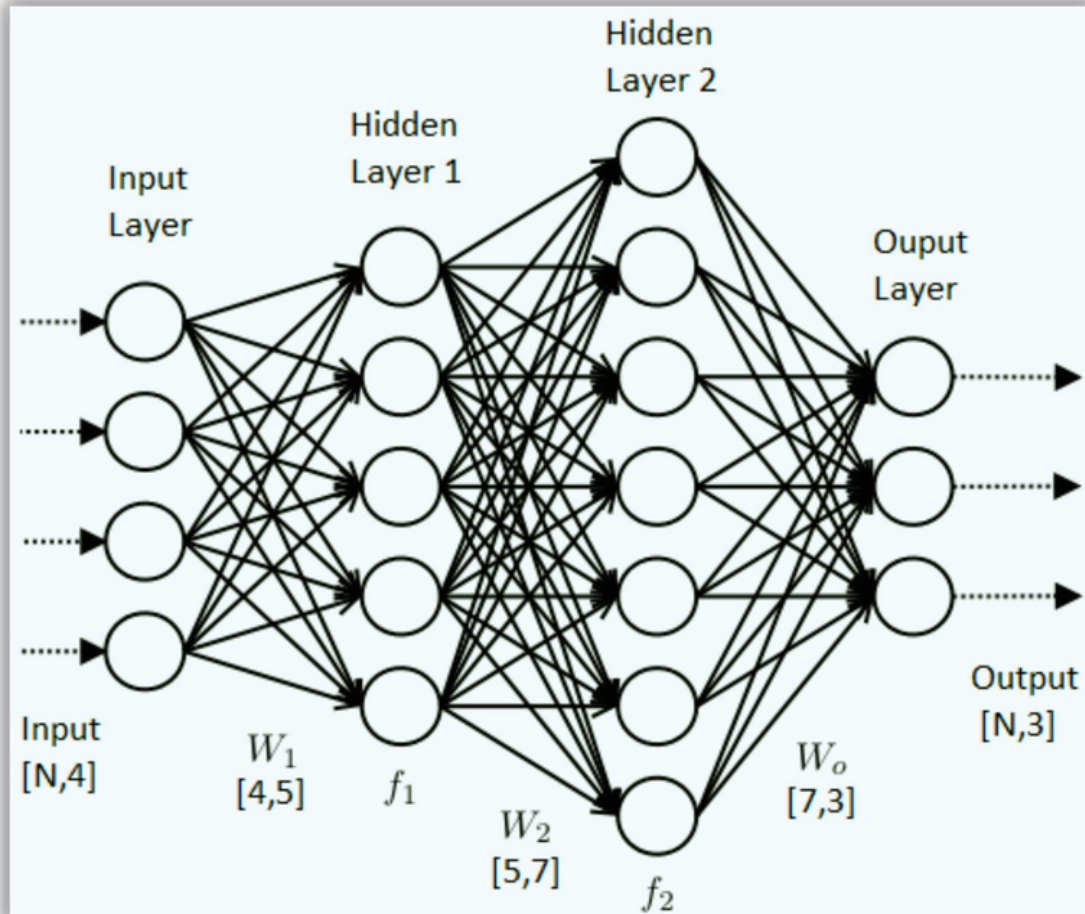


Fig:.. Neural Network

# Understanding Machine Learning

List of algorithms in Artificial Neural Network: -

- Autoencoder
- Deep learning
- RNN (Recurrent Neural Network)
  - LSTM (Long Short Term Memory)
  - GRU (Gated recurrent Unit)
  - ESN (Echo state network)
- CNN (Convolutional Neural Network)
  - U-net
- Multilayer perceptron
- GAN (Generative Adversarial Network)
- Restricted Boltzmann Machine
- Spiking Neural Network

## What is Computer Vision ?

Computer vision is a field of artificial intelligence (AI) that enables computers and systems to derive meaningful information from digital images, videos and other visual inputs — and take actions or make recommendations based on that information. If AI enables computers to think, computer vision enables them to see, observe and understand.

Computer vision works much the same as human vision, except humans have a head start. Human sight has the advantage of lifetimes of context to train how to tell objects apart, how far away they are, whether they are moving and whether there is something wrong in an image.

# Understanding Machine Learning

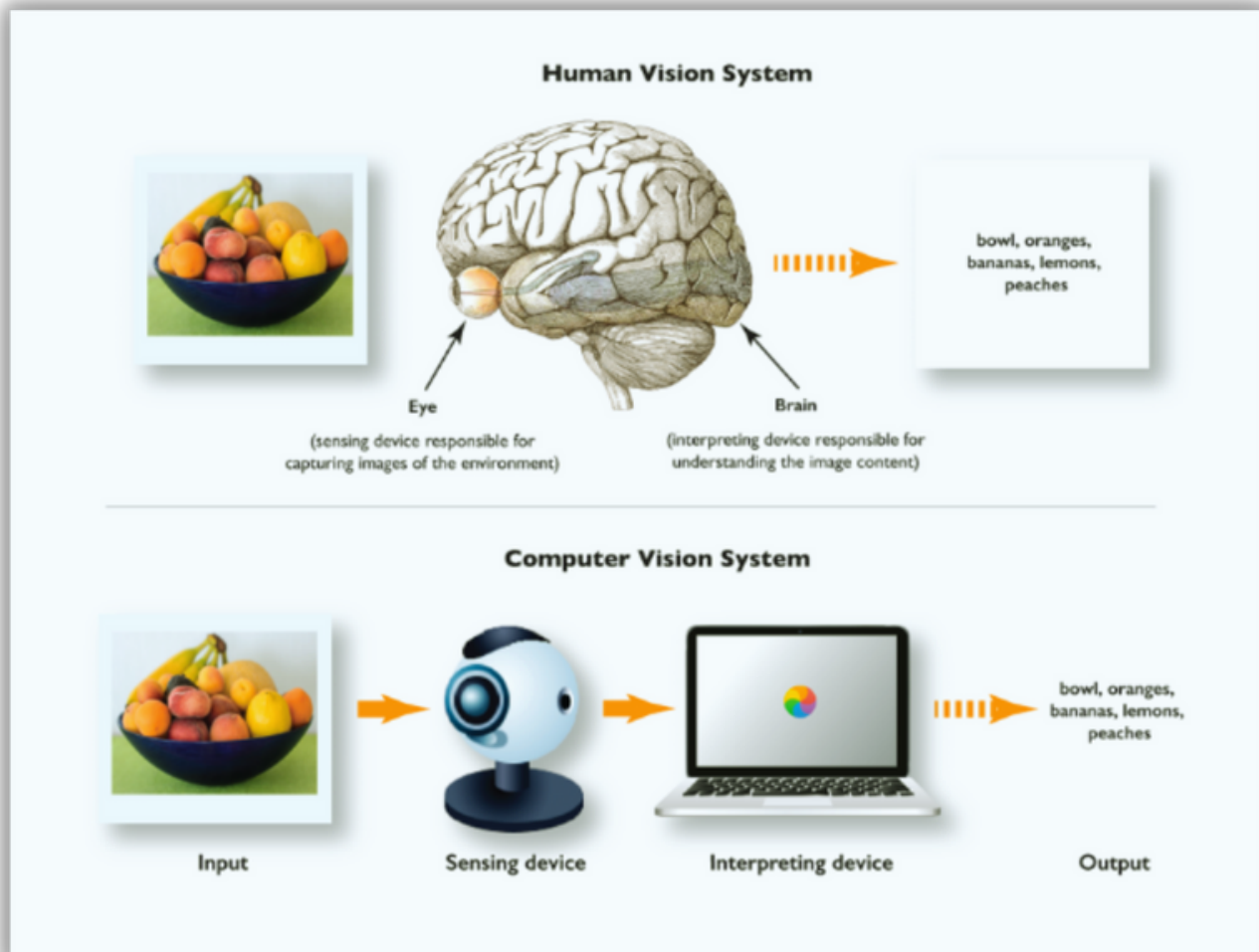


Fig.: Computer Vision System in Analog to Human Vision System

Computer vision trains machines to perform these functions, but it has to do it in much less time with cameras, data and algorithms rather than retinas, optic nerves and a visual cortex. Because a system trained to inspect products or watch a production asset can analyse thousands of products or processes a minute, noticing imperceptible defects or issues, it can quickly surpass human capabilities.

# Understanding Machine Learning

Here are a few common tasks that computer vision systems can be used for:

- Object classification. The system parses visual content and classifies the object on a photo/video to the defined category. For example, the system can find a dog among all objects in the image.
- Object identification. The system parses visual content and identifies a particular object on a photo/video. For example, the system can find a specific dog among the dogs in the image.
- Object tracking. The system processes video finds the object (or objects) that match search criteria and track its movement.
- Content-based image retrieval uses computer vision to browse, search and retrieve images from large data stores, based on the content of the images rather than metadata tags associated with them. This task can incorporate automatic image annotation that replaces manual image tagging. These tasks can be used for digital asset management systems and can increase the accuracy of search and retrieval.
- Scene Reconstruction
- Emotion detection
- Motion analysis
- Instance segmentation



# Understanding Machine Learning

## How does Computer Vision works ?

Computer vision technology tends to mimic the way the human brain works. Two essential technologies are used to accomplish this: a type of machine learning called deep learning and a convolutional neural network (CNN).

Machine learning uses algorithmic models that enable a computer to teach itself about the context of visual data. If enough data is fed through the model, the computer will “look” at the data and teach itself to tell one image from another. Algorithms enable the machine to learn by itself, rather than someone programming it to recognize an image.

A CNN helps a machine learning or deep learning model “look” by breaking images down into pixels that are given tags or labels. It uses the labels to perform convolutions (a mathematical operation on two functions to produce a third function) and makes predictions about what it is “seeing.” The neural network runs convolutions and checks the accuracy of its predictions in a series of iterations until the predictions start to come true. It is then recognizing or seeing images in a way similar to humans.

Much like a human making out an image at a distance, a CNN first discerns hard edges and simple shapes, then fills in information as it runs iterations of its predictions. A CNN is used to understand single images. A recurrent neural network (RNN) is used in a similar way for video applications to help computers understand how pictures in a series of frames are related to one another.

# Understanding Machine Learning

List of algorithms of Computer Vision: -

- Object identification: -
  - o AlexNet
  - o GoogLeNet
  - o VGGNet
  - o ResNet
  - o InceptionNet
  - o DenseNet
  - o LeNet
- Object segmentation: -
  - o DeepLab
  - o RefineNet
  - o Fully Convolutional network
  - o U-Net
  - o PSPNet
  - o SegNet
  - o Conditional random field
- Object detection: -
  - o R-CNN
  - o Fast R-CNN
  - o R-FCN
  - o YOLO
  - o SSD

# Understanding Machine Learning

## What is Natural language Processing ?

Natural language processing (NLP) concerns itself with the interaction between natural human languages and computing devices. NLP is a major aspect of computational linguistics, and also falls within the realms of computer science and artificial intelligence.

NLP combines computational linguistics—rule-based modelling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to ‘understand’ its full meaning, complete with the speaker or writer’s intent and sentiment. Challenges in natural language processing frequently involve speech recognition, natural-language understanding, and natural-language generation.

### NLP Task:-

Several NLP tasks break down human text and voice data in ways that help the computer make sense of what it's ingesting. Some of these tasks include the following:

- **Speech recognition**, also called speech-to-text, is the task of reliably converting voice data into text data. Speech recognition is required for any application that follows voice commands or answers spoken questions. What makes speech recognition especially challenging is the way people talk—quickly, slurring words together, with varying emphasis and intonation, in different accents, and often using incorrect grammar.
- **Part of speech tagging**, also called grammatical tagging, is the process of determining the part of speech of a particular word or piece of text based on its use and context. Part of speech identifies ‘make’ as a verb in ‘I can make a paper plane,’ and as a noun in ‘What make of car do you own?’
- **Word sense disambiguation** is the selection of the meaning of a word with multiple meanings through a process of semantic analysis that determine the word that makes the most sense in the given context. For example, word sense disambiguation helps distinguish the meaning of the verb 'make' in ‘make the grade’ (achieve) vs. ‘make a bet’ (place).

# Understanding Machine Learning

- **Named entity recognition**, or NEM, identifies words or phrases as useful entities. NEM identifies 'Kentucky' as a location or 'Fred' as a man's name.
- **Co-reference resolution** is the task of identifying if and when two words refer to the same entity. The most common example is determining the person or object to which a certain pronoun refers (e.g., 'she' = 'Mary'), but it can also involve identifying a metaphor or an idiom in the text (e.g., an instance in which 'bear' isn't an animal but a large hairy person).
- **Sentiment analysis** attempts to extract subjective qualities—attitudes, emotions, sarcasm, confusion, suspicion—from text.
- **Natural language generation** is sometimes described as the opposite of speech recognition or speech-to-text; it's the task of putting structured information into human language.

The Python programming language provides a wide range of tools and libraries for attacking specific NLP tasks. Many of these are found in the Natural Language Toolkit, or NLTK, an open source collection of libraries, programs, and education resources for building NLP programs.

Application of NLP: -

Natural language processing is the driving force behind machine intelligence in many modern real-world applications. Here are a few examples:

- **Spam detection:** You may not think of spam detection as an NLP solution, but the best spam detection technologies use NLP's text classification capabilities to scan emails for language that often indicates spam or phishing. These indicators can include overuse of financial terms, characteristic bad grammar, threatening language, inappropriate urgency, misspelled company names, and more. Spam detection is one of a handful of NLP problems that experts consider 'mostly solved' (although you may argue that this doesn't match your email experience).

# Understanding Machine Learning

- **Machine translation:** Google Translate is an example of widely available NLP technology at work. Truly useful machine translation involves more than replacing words in one language with words of another. Effective translation has to capture accurately the meaning and tone of the input language and translate it to text with the same meaning and desired impact in the output language. Machine translation tools are making good progress in terms of accuracy. A great way to test any machine translation tool is to translate text to one language and then back to the original. An oft-cited classic example: Not long ago, translating “The spirit is willing but the flesh is weak” from English to Russian and back yielded “The vodka is good but the meat is rotten.” Today, the result is “The spirit desires, but the flesh is weak,” which isn’t perfect, but inspires much more confidence in the English-to-Russian translation.
- **Virtual agents and chatbots:** Virtual agents such as Apple's Siri and Amazon's Alexa use speech recognition to recognize patterns in voice commands and natural language generation to respond with appropriate action or helpful comments. Chatbots perform the same magic in response to typed text entries. The best of these also learn to recognize contextual clues about human requests and use them to provide even better responses or options over time. The next enhancement for these applications is question answering, the ability to respond to our questions—anticipated or not—with relevant and helpful answers in their own words.
- **Social media sentiment analysis:** NLP has become an essential business tool for uncovering hidden data insights from social media channels. Sentiment analysis can analyze language used in social media posts, responses, reviews, and more to extract attitudes and emotions in response to products, promotions, and events—information companies can use in product designs, advertising campaigns, and more.
- **Text summarization:** Text summarization uses NLP techniques to digest huge volumes of digital text and create summaries and synopses for indexes, research databases, or busy readers who don't have time to read full text. The best text summarization applications use semantic reasoning and natural language generation (NLG) to add useful context and conclusions to summaries.

# Understanding Data mining

## Lesson Outcomes:-

- What is Big Data ?
- What is Data Engineering ?
- What is Data Visualization ?

# Understanding Data Mining

## What is Big Data ?

Big data is any kind of data source that has at least one of four shared characteristics, called the four Vs: -

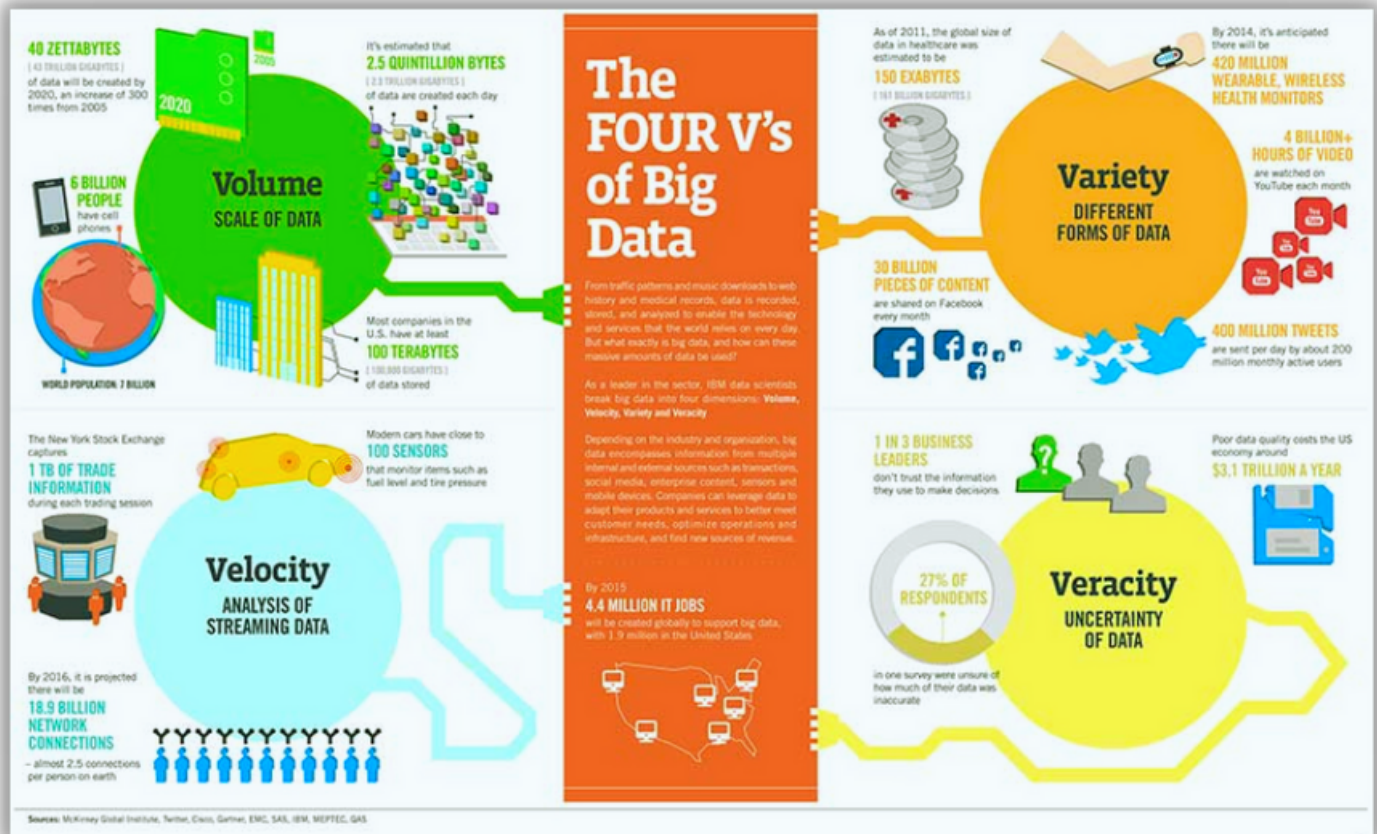


Fig.: Four V's of Big Data

# Understanding Data Mining

**1. Volume:** - Scale of data is known as Volume. Big data exceeds the storage capacity of conventional databases. Today we generate 5 exabytes of data every two days. That it's Volume aspect of data.

**2. Velocity:** -Data velocity is accelerating. Streams of tweets, Facebook entries, financial information, etc., are being generated by more users at an ever-increasing pace. Whereas velocity increases data volume, often exponentially, it might shorten the window of data retention or application. For example, high-frequency trading relies on micro-second information and streams of data, but the relevance of the data rapidly decays.

**3. Variety:** - Data variety is much greater than ever before. Models that relied on just a handful of variables can now avail of hundreds of variables, as computing power has increased. The scale of change in volume, velocity, and variety of the data that is now available calls for new econometrics, and a range of tools for even single questions. This book aims to introduce the reader to a variety of modelling concepts and econometric techniques that are essential for a well-rounded data scientist.

**4. Veracity:** - The veracity of big data relates to the nature of the data that is being examined. The high veracity of big data has numerous records that are important to break down and that contribute in a significant manner to the general outcomes. The low veracity of big data, then again, contains a high level of pointless data.

Data science is more than the mere analysis of large data sets. It is also about the creation of data. The field of “text-mining” expands available data enormously, since there is so much more text being generated than numbers. The creation of data from varied sources, and its quantification into information is known as “datafication.”



# Understanding Data Mining

The definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs. Big data can help us address a range of business activities, from customer experience to analytics. Here are just few applications of big data:-

- Product development
- Predictive maintenance
- Customer experience
- Fraud and compliance
- Machine learning
- Operational efficiency
- Drive innovation

## How Big Data Work ?

Big data gives you new insights that open up new opportunities and business models. Getting started involves three key actions:

### **1. Integrate:-**

Big data brings together data from many disparate sources and applications. Traditional data integration mechanisms, such as extract, transform, and load (ETL) generally aren't up to the task. It requires new strategies and technologies to analyze big data sets at terabyte, or even petabyte, scale.

During integration, you need to bring in the data, process it, and make sure it's formatted and available in a form that your business analysts can get started with.

# Understanding Data Mining

## **2.Manage:-**

Big data requires storage. Your storage solution can be in the cloud, on premises, or both. You can store your data in any form you want and bring your desired processing requirements and necessary process engines to those data sets on an on-demand basis.

Many people choose their storage solution according to where their data is currently residing. The cloud is gradually gaining popularity because it supports your current compute requirements and enables you to spin up resources as needed.

## **3.Analyze:-**

Your investment in big data pays off when you analyze and act on your data. Get new clarity with a visual analysis of your varied data sets. Explore the data further to make new discoveries. Share your findings with others. Build data models with machine learning and artificial intelligence. Put your data to work.

As we know that data is the next big thing, so if we succeed in creating new strategies and technologies on how big data works, it will take the game to whole another level.

# Understanding Data Mining

## What is Data Engineering ?

The key to understanding what data engineering lies in the “engineering” part. Engineers design and build things. “Data” engineers design and build pipelines that transform and transport data into a format wherein, by the time it reaches the Data Scientists or other end users, it is in a highly usable state. These pipelines must take data from many disparate sources and collect them into a single warehouse that represents the data uniformly as a single source of truth.

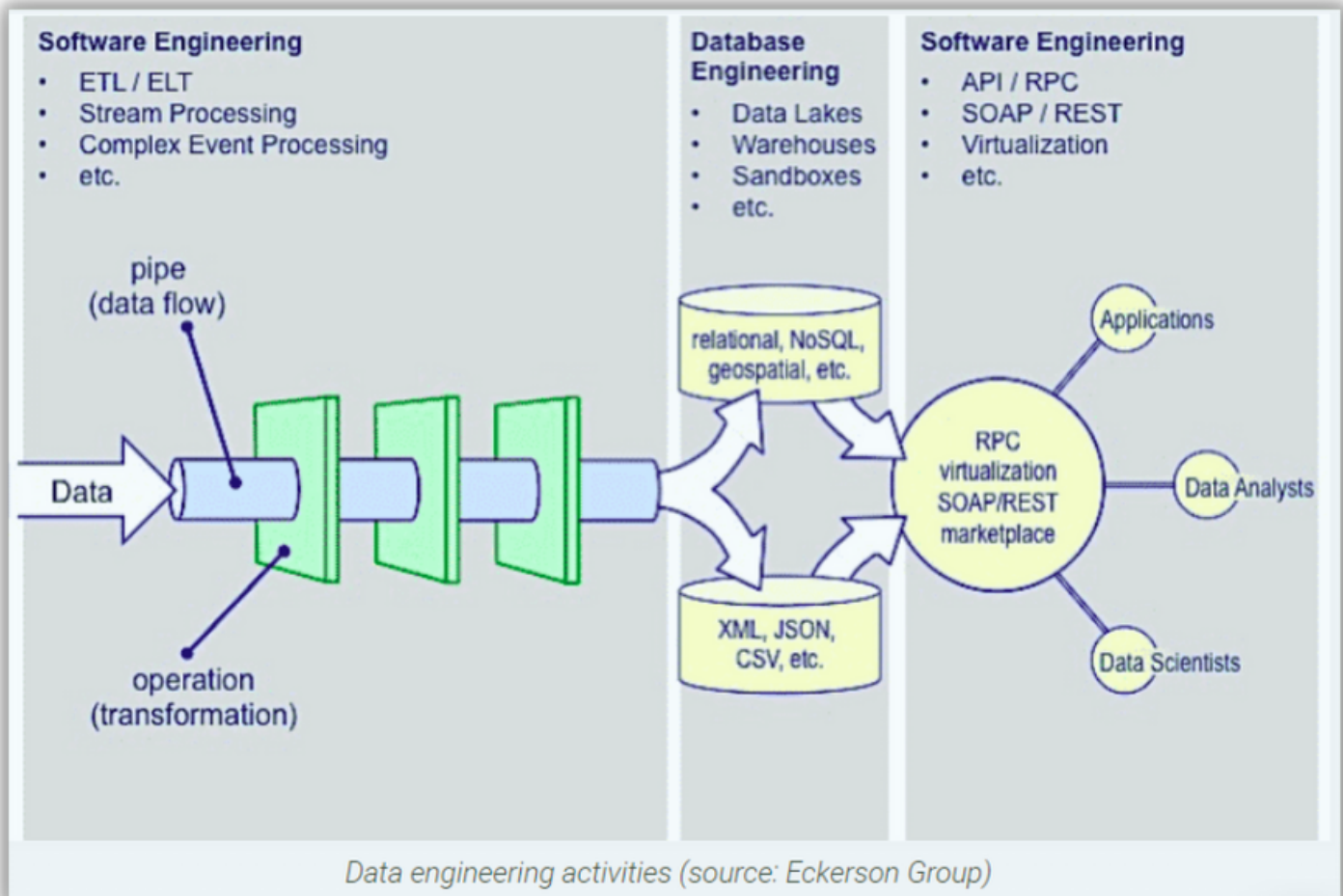


Fig.: Data Engineering activities

# Understanding Data Mining

## What is Data Visualization?

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

### **Idea generation**

Data visualization is commonly used to spur idea generation across teams. They are frequently leveraged during brainstorming or Design Thinking sessions at the start of a project by supporting the collection of different perspectives and highlighting the common concerns of the collective. While these visualizations are usually unpolished and unrefined, they help set the foundation within the project to ensure that the team is aligned on the problem that they're looking to address for key stakeholders.

### **Idea illustration**

Data visualization for idea illustration assists in conveying an idea, such as a tactic or process. It is commonly used in learning settings, such as tutorials, certification courses, centers of excellence, but it can also be used to represent organization structures or processes, facilitating communication between the right individuals for specific tasks. Project managers frequently use Gantt charts and waterfall charts to illustrate workflows. Data modeling also uses abstraction to represent and better understand data flow within an enterprise's information system, making it easier for developers, business analysts, data architects, and others to understand the relationships in a database or data warehouse.

### **Visual discovery**

Visual discovery and every day data viz are more closely aligned with data teams. While visual discovery helps data analysts, data scientists, and other data professionals identify patterns and trends within a dataset, every day data viz supports the subsequent storytelling after a new insight has been found.

# Understanding Data Mining

## Types of Data Visualization?

Dashboards are effective data visualization tools for tracking and visualizing data from multiple data sources, providing visibility into the effects of specific behaviour by a team or an adjacent one on performance. Dashboards include common visualization techniques, such as:

- **Tables:** This consists of rows and columns used to compare variables. Tables can show a great deal of information in a structured way, but they can also overwhelm users that are simply looking for high-level trends.
- **Pie charts and stacked bar charts:** These graphs are divided into sections that represent parts of a whole. They provide a simple way to organize data and compare the size of each component to one other.
- **Line charts and area charts:** These visuals show change in one or more quantities by plotting a series of data points over time and are frequently used within predictive analytics. Line graphs utilize lines to demonstrate these changes while area charts connect data points with line segments, stacking variables on top of one another and using color to distinguish between variables.
- **Histograms:** This graph plots a distribution of numbers using a bar chart (with no spaces between the bars), representing the quantity of data that falls within a particular range. This visual makes it easy for an end user to identify outliers within a given dataset.
- **Scatter plots:** These visuals are beneficial in revealing the relationship between two variables, and they are commonly used within regression data analysis. However, these can sometimes be confused with bubble charts, which are used to visualize three variables via the x-axis, the y-axis, and the size of the bubble.
- **Heat maps:** These graphical representation displays are helpful in visualizing behavioural data by location. This can be a location on a map, or even a webpage.
- **Tree maps,** which display hierarchical data as a set of nested shapes, typically rectangles. Tree-maps are great for comparing the proportions between categories via their area size.

# How to get started with Data Science

## Lesson Outcomes:-

- Math and Statistics
- Programming and coding
- ML Theory
- Own Projects
- Roadmap

# How to get started with Data Science

When getting started with data science it is important to learn strategically in order to make your learning faster and effective. So in this book I have provided you the step by step guide to learn data science.

## Step by Step Guide:-

- **Learn Python:-** The First and Foremost Step Towards Data Science should learning be a programming language ( i.e. Python). Python is the most common coding language, used by the Majority of Data Scientist, because of its simplicity, versatility and being pre-equipped with powerful libraries ( like NumPy, SciPy, and Pandas) useful in data analysis and other aspects in Data Science. Python is an open-source language and supports various libraries.
- **Learn Statistics:-** If Data Science is a language, then statistics is basically the grammar. Statistics is basically the method of analyzing, interpretation of large data sets. When it comes to data analysis and gathering insights, statistics is as noteworthy as air to us. Statistics help us understand the hidden details from large datasets.
- **Data Collection:-** This is one of the key and important steps in the field of Data Science. This skill involves knowledge of various tools to import data from both local systems, as CSV files, and scraping data from websites, using BeautifulSoup Python library. Scraping can also be API-based. Data collection can be managed with knowledge of Query Language or ETL pipelines in Python.
- **Data Cleaning:-** This is the Step where most of the time is being spent as a Data Scientist. Data cleaning is all about obtaining the data, fit for doing work & analysis, by removing unwanted values, missing values, categorical values, outliers, and wrongly submitted records, from the Raw form of Data. Data Cleaning is very important as real-world data is messy in nature and achieving it with help of various Python libraries (Pandas and NumPy) is really important for an aspirant Data Scientist.

# How to get started with Data Science

- **Acquaintance With EDA( Exploratory Data Analysis):-** EDA( Exploratory data analysis) is the most important aspect in the vast field of data science. It includes analyzing various data, variables, various data patterns, trends and extracting useful insights from them with help of various graphical and statistic l methods. EDA identifies various pattern which Machine learning algorithm might fail to identify. It includes all Data Manipulation, Analysis, and Visualization.
- **Machine Learning & Deep Learning:-** Machine learning is the core skill required to be a Data Scientist. Machine learning is used to build various predictive models, classification models, etc., and is being used By big firms, Companies to Optimize their planning as per the predictions. For example Car Price prediction.
- **Deep Learning** on the other hand is and an advanced version of Machine Learning which deploys the use of Neural Network, a framework that combines various machine learning algorithms for solving various tasks, for training data. Various Neural networks are recurrent neural network (RNN) or a convolutional neural network (CNN) etc. For Example: Face Recognition.
- **Learn Deploying of ML model:-** Deployment is basically the process of making your Machine Learning Model available to end-users for use. This is achieved by the integration of the model with various existing production environments thus implementing the practical use of the ML model for various Business solutions. There are many services for deploying your ML model like Flask, Pythoneverywhere, MLOps , Microsoft Azure, Google Cloud, Heroku, etc.
- **Real-World Testing:-** Testing and Validation of the Machine Learning Model after Deployment Should Be done In order to check its effectiveness and accuracy. Testing is an Important Step In Data Science for keeping the efficiency and effectiveness of the ML model In check. There is Various Type Of Testing like A/B, AAB Testing, etc.



# How to get started with Data Science

## Master your path

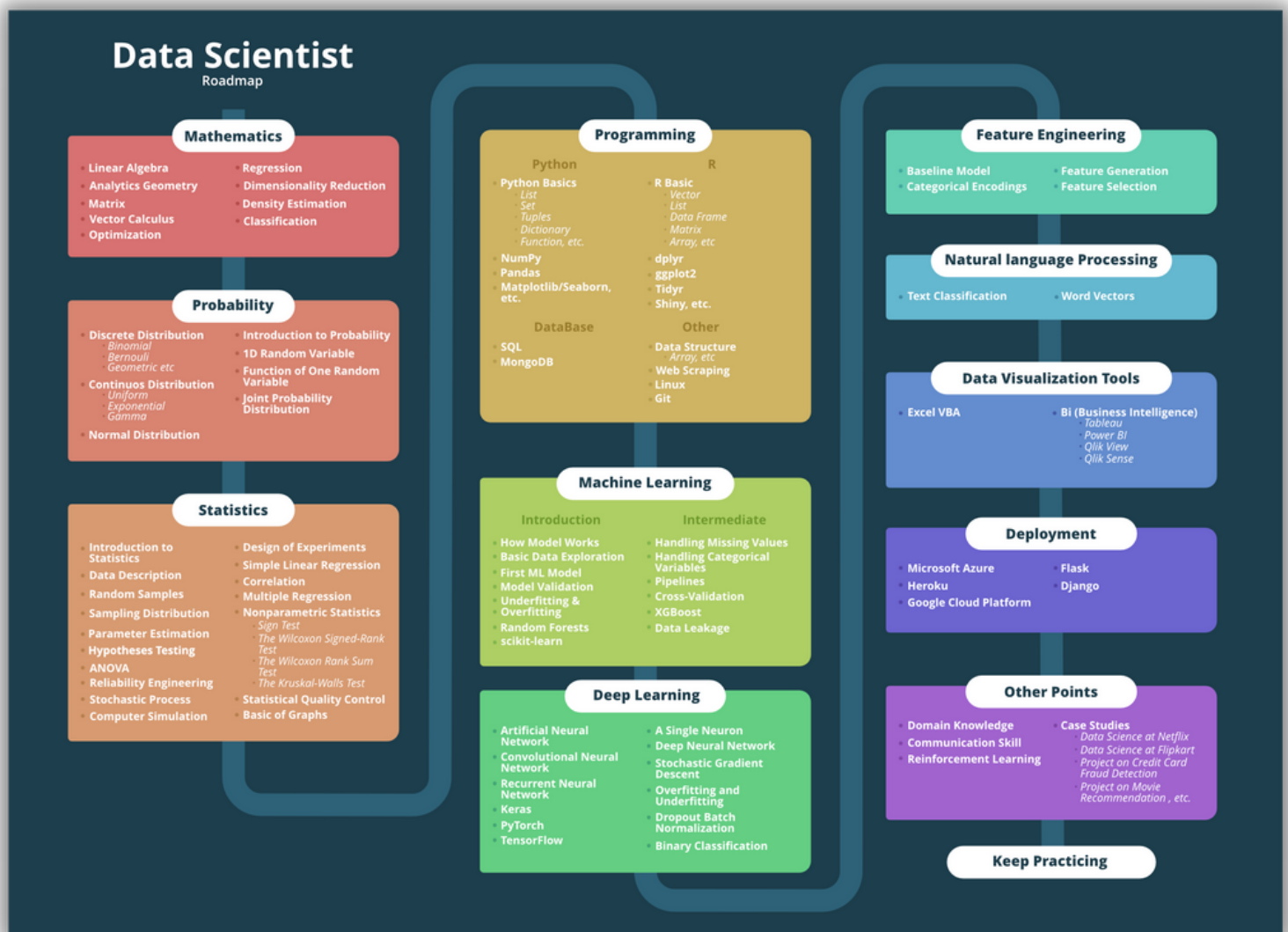
To become an expert in machine learning, you first need a strong foundation in four learning areas : coding, math, ML theory, and how to build your own ML project from start to finish.

The four areas of machine learning education:-

- **Coding skills:** Building ML models involves much more than just knowing ML concepts—it requires coding in order to do the data management, parameter tuning, and parsing results needed to test and optimize your model.
- **Math and stats:** ML is a math heavy discipline, so if you plan to modify ML models or build new ones from scratch, familiarity with the underlying math concepts is crucial to the process.
- **ML theory:** Knowing the basics of ML theory will give you a foundation to build on, and help you troubleshoot when something goes wrong.
- **Build your own projects:** Getting hands on experience with ML is the best way to put your knowledge to the test, so don't be afraid to dive in early with a simple colab or tutorial to get some practice.

# How to get started with Data Science

## Roadmap to Data Science



# Conclusion

Data Science is a developing field with evolving technologies and tools. Jobs for data science professionals seem far from saturation in the next decade, securing the future of data scientists. The ever-increasing opportunities and ever-changing data landscape in the industry can be marked as a reason for the same.

Digital transformation and technological advancements have led to data-driven businesses. The humongous growth in data being gathered from multiple sources has necessitated the growth of data science. Thus, the demand for skilled data science professionals with the ability to handle massive zettabytes of data is going to rise.