



ARTIFICIAL INTELLIGENCE CURRICULUM

Curated with support from Intel®

CLASS 10
FACILITATOR HANDBOOK

Artificial Intelligence Curriculum

Class 10

Curated with support from Intel®

Acknowledgements

Patrons:

- Sh. Ramesh Pokhriyal 'Nishank', Minister of Human Resource Development, Government of India
- Sh. Dhotre Sanjay Shamrao, Minister of State for Human Resource Development, Government of India

Human Resource Development, Government of India Advisory, Editorial and Creative Inputs:

- Ms. Anita Karwal, IAS, Chairperson, Central Board of Secondary Education
- Ms. Shweta Khurana, Director, Programs, Partnerships and Policy Group, Intel India

Guidance and Support:

- Sh. Anurag Tripathi, IRPS, Secretary, Central Board of Secondary Education
- Dr. Joseph Emmanuel, Director (Academics), Central Board of Secondary Education
- Dr. Biswajit Saha, Director (Skill Education & Training), Central Board of Secondary Education

Education Value adder, Curator and Coordinator:

- Sh. Ravinder Pal Singh, Joint Secretary, Department of Skill Education, Central Board of Secondary Education

Content Curation Team:

- Ms. Sharon E. Kumar, Innovation and Education Consultant, Intel AI4Youth Program
- Ms. Ambika Saxena, Intel AI For Youth Coach
- Mr. Bhavik Khurana, Intel AI For Youth Coach
- Mr. Akshay Chawla, Intel AI For Youth Coach
- Mr. Shivam Agrawal, Intel AI For Youth Coach

Feedback By:

- Ms. Neelam Roy, ITL Public School, Delhi
- Ms. Mehreen Shamim, TGT, DPS Bangalore East, Bengaluru
- Ms. Saswati Sarangi, PGT Computer Science, RCIS Kalyan Nagar, Bengaluru
- Ms. Aayushi Agrawal, Salwan Girls School, Delhi
- Ms. Isha, HOD Computer Science, Salwan Public School, Delhi

Special Thanks To:

- Ms. Indu Khetrpal, Principal, Salwan Public School, Delhi
- Ms. Rekha Vinod, Principal, RCIS Kalyan Nagar, Bengaluru
- Ms. Manilla Carvalho, Principal, Delhi Public School – Bangalore East, Bengaluru
- Ms. Sudha Acharya, Principal, ITL Public School, Delhi
- Ms. Puneet Sardana, Vice-Principal, Salwan Girls School, Delhi

About the book

Artificial Intelligence (AI) is being widely recognised to be the power that will fuel the future global digital economy. AI in the past few years has gained geo-strategic importance and a large number of countries are striving hard to stay ahead with their policy initiatives to get their country ready.

India's own AI strategy identifies AI as a n opportunity and solution provider for inclusive economic growth and social development. The report also identifies the importance of skills-based education (as opposed to knowledge intensive education), and the value of project related work in order to “effectively harness the potential of AI in a sustainable manner” to make India's next generation ‘AI ready’.

As a beginning in this direction, CBSE introduced Artificial Intelligence as an optional subject at Class IX from the Session 2019-2020 onwards. Also, to enhance the multidisciplinary approach in teaching-learning so as to sensitize the new generation, it was decided that schools may start AI “Inspire Module” of 12 hours at class VIII itself. CBSE has extended this subject to class X as well from the Session 2020-2021.

CBSE is already offering various skill subjects at secondary and senior secondary level to upgrade the skills and proficiency of the young generation and also to provide them awareness to explore various career options. Ai secondary level, a skill subject may be offered as additional sixth subject along with the existing five compulsory subjects.

CBSE acknowledges the initiative by Intel India in curating this Facilitator Handbook, the AI training video and managing the subsequent trainings of trainers on the Artificial Intelligence Curriculum.

The aim is to strive together to make our students future ready and help them work on incorporating Artificial Intelligence to improve their learning experience.

Table of Contents

Introduction to AI: Foundational Concepts	9
What is Intelligence?	9
Decision Making	12
How do you make decisions?	12
Make Your Choices!	12
What is Artificial Intelligence?	14
How do machines become Artificially Intelligent?	14
Applications of Artificial Intelligence around us	15
What is not AI?	16
Introduction to AI: Basics of AI	18
AI, ML & DL	20
Introduction to AI Domains	21
Data Sciences	21
Computer Vision	21
Natural Language Processing	22
AI Ethics	23
Moral Issues: Self-Driving Cars	23
Data Privacy	24
AI Bias	26
AI Access	27
AI Project Cycle	29
Introduction	29
Problem Scoping	30
Data Acquisition	34
Data Exploration	35
Modelling	36
Learning Based Approach	37
Evaluation	39
Neural Networks	40
Advance Python	42
Recap	42
Recap 1: Jupyter Notebook	42
Introduction to Virtual Environments	43

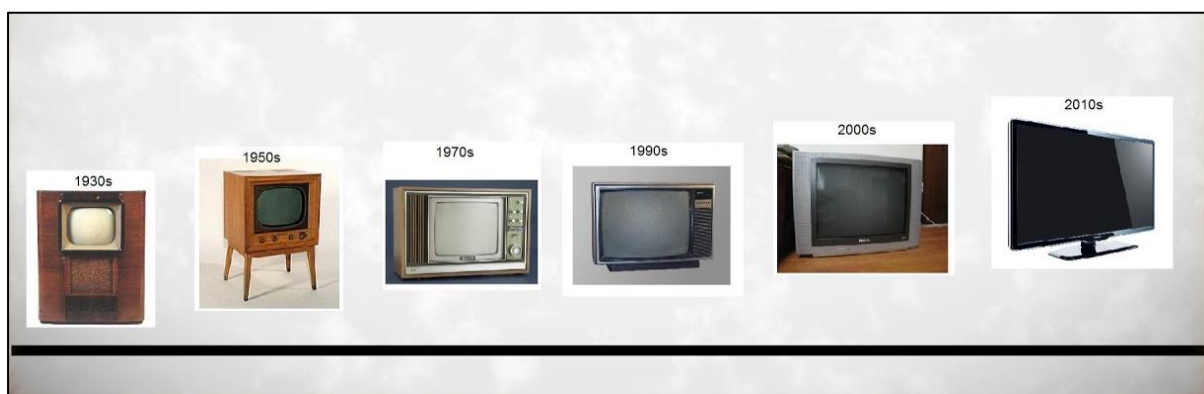
Recap 2: Introduction to Python.....	47
Applications of Python.....	48
Recap 3: Python Basics.....	48
Python Packages	52
Data Sciences	54
Introduction	54
Applications of Data Sciences	55
Getting Started.....	57
Revisiting AI Project Cycle.....	57
Data Collection.....	62
Data Access	63
Basic Statistics with Python	66
Data Visualisation	67
Data Sciences: Classification Model.....	71
Personality Prediction	71
K-Nearest Neighbour: Explained.....	72
Computer Vision	75
Introduction	75
Applications of Computer Vision	76
Computer Vision: Getting Started.....	78
Computer Vision Tasks.....	78
Classification	78
Classification + Localisation	78
Object Detection	78
Instance Segmentation	78
Basics of Images.....	79
Basics of Pixels	79
Image Features.....	84
Introduction to OpenCV.....	85
Convolution.....	86
Convolution : Explained	88
Convolution Neural Networks (CNN)	91
Introduction	91
What is a Convolutional Neural Network ?.....	92
Convolution Layer	93
Rectified Linear Unit Function	94

Pooling Layer.....	95
Fully Connected Layer.....	96
Natural Language Processing.....	99
Introduction.....	99
Applications of Natural Language Processing.....	100
Natural Language Processing: Getting Started.....	101
Revisiting the AI Project Cycle.....	101
Chatbots.....	104
Human Language VS Computer Language.....	105
Arrangement of the words and meaning.....	106
Multiple Meanings of a word.....	107
Perfect Syntax, no Meaning.....	107
Data Processing.....	108
Text Normalisation.....	108
Bag of Words.....	112
TFIDF: Term Frequency & Inverse Document Frequency.....	114
Applications of TFIDF.....	118
DIY – Do It Yourself!.....	118
Evaluation.....	119
Introduction.....	119
What is evaluation?.....	119
Model Evaluation Terminologies.....	119
The Scenario.....	119
Confusion matrix.....	122
Evaluation Methods.....	123
Accuracy.....	123
Precision.....	124
Recall.....	125
Which Metric is Important?.....	126
F1 Score.....	127

Introduction to AI: Foundational Concepts

What is Intelligence?

Humans have been developing machines which can make their lives easier. Machines are made with an intent of accomplishing tasks which are either too tedious for humans or are time consuming. Hence, machines help us by working for us, thereby sharing our load and making it easier for us to fulfil such goals.



Life without machines today is unimaginable, and because of this, humans have been putting efforts into making them even more sophisticated and smart. As a result, we are surrounded by smart devices and gadgets like smartphones, smartwatches, smart TV, etc. But what makes them smart?



For example, how is a smartphone today different from the telephones we had in the last century?

Let us define each term mentioned above to get a proper understanding:

Mathematical Logical Reasoning	•A person's ability to regulate, measure, and understand numerical symbols, abstraction and logic.
Linguistic Intelligence	•Language processing skills both in terms of understanding or implementation in writing or verbally.
Spatial Visual Intelligence	•It is defined as the ability to perceive the visual world and the relationship of one object to another.
Kineasthetic Intelligence	•Ability that is related to how a person uses his limbs in a skilled manilr.
Musical Intelligence	•As the name suggests, this intelligence is about a person's ability to recognize and create sounds, rhythms, and sound patterns.
Intrapersonal Intelligence	•Describes how high the level of self-awareness someone has is. Starting from realizing weakness, strength, to his own feelings.
Existential Intelligence	•An additional category of intelligence relating to religious and spiritual awareness.
Naturalist Intelligence	•An additional category of intelligence relating to the ability to process information on the environment around us.
Interpersonal intelligence	•Interpersonal intelligence is the ability to communicate with others by understanding other people's feelings & influence of the person.

But even though one is more skilled in intelligence than the other, it should be noted that in fact all humans have all 9 of these intelligences only at different levels. One might be an expert at painting, while the other might be an expert in mathematical calculations. One is a musician, the other is an expert dancer.

In other words, we may define intelligence as:

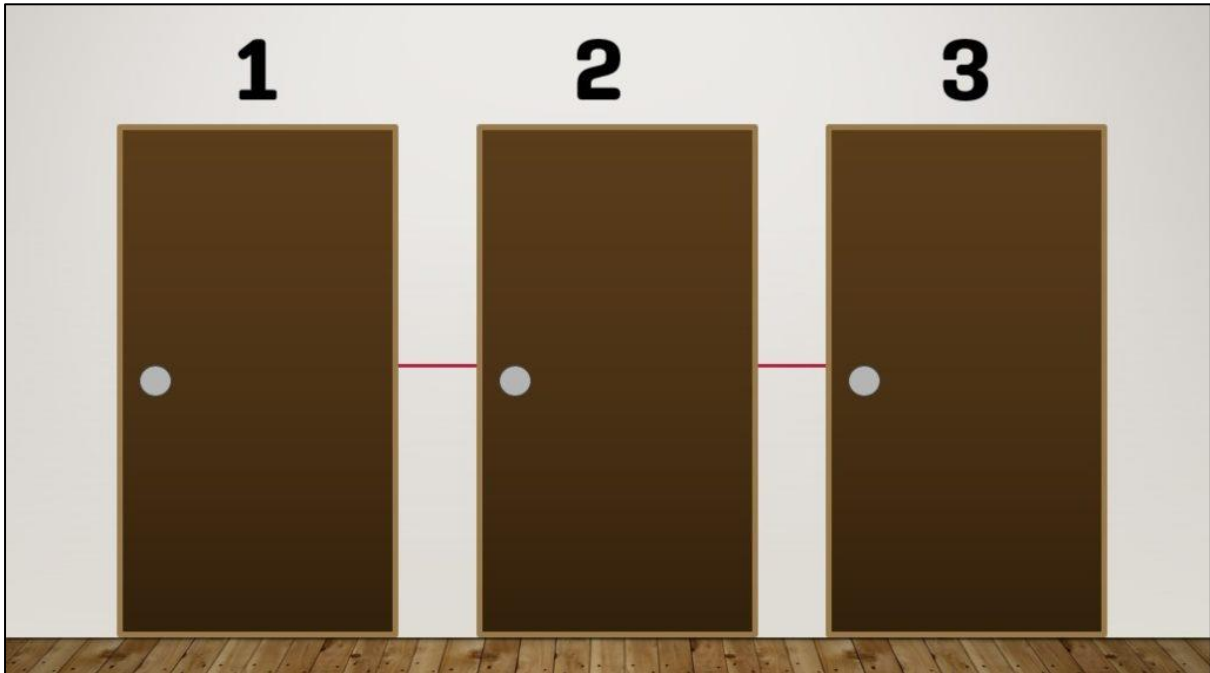
- Ability to interact with the real world
 - To perceive, understand and act
 - Example: Speech Recognition – Understanding and synthesis
 - Example: Image Recognition
 - Example: Ability to take action: to have an effect
- Reasoning and planning
 - Modelling the external world, given input
 - Solving new problems, planning and making decisions
 - Ability to deal with unexpected problems, uncertainties
- Learning and adaptation
 - Continuous learning and adapting graph
 - Our internal models are always being updated
 - Example: Baby learning to categorize and recognise animals

For example, if someone starts talking to us, we know how to keep the conversation going. We can understand what people mean and can reply in the same way. When we are hungry, we can come up

with various options on what to eat depending upon the food we have at our homes. When we read something, we are able to understand its meaning and answer anything regarding it.

While understanding the term intelligence, it must be noticed that decision making comprises of a crucial part of intelligence. Let us delve deeper into it.

Decision Making



You're trapped. All the doors seem to have started shrinking and only one of them leads you out. Which door would you pick?

How do you make decisions?

The basis of decision making depends upon the availability of information and how we experience and understand it. For the purposes of this article, 'information' includes our past experience, intuition, knowledge, and self-awareness.

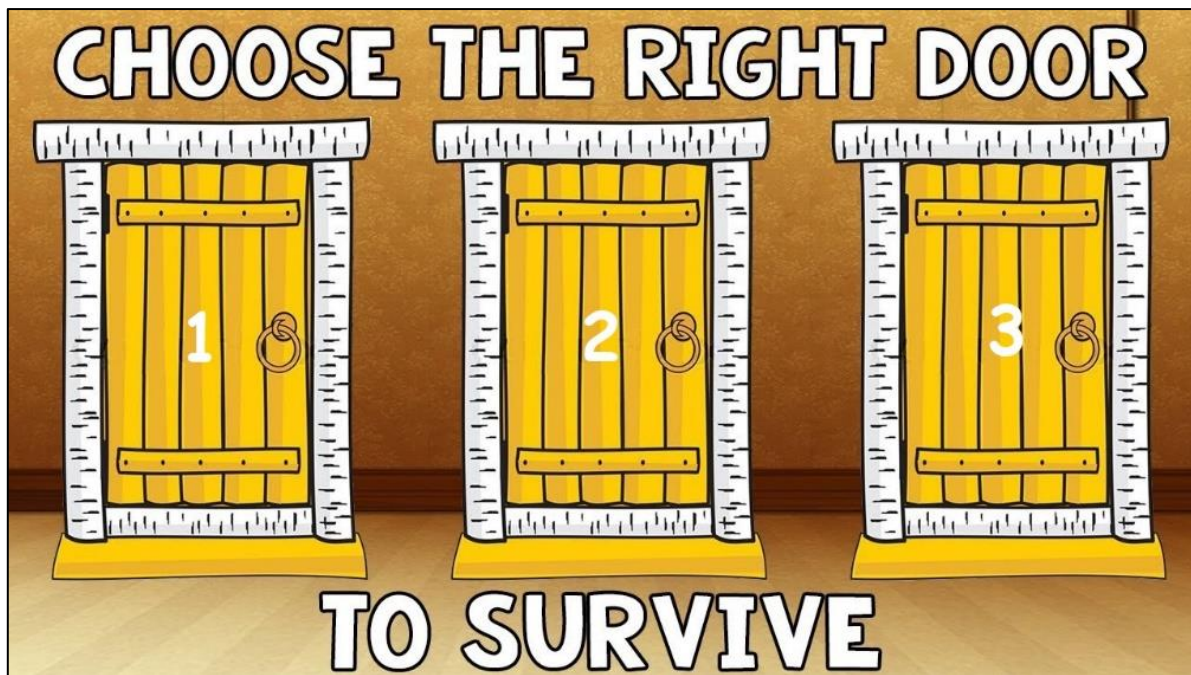
We can't make "good" decisions without information because then we have to deal with unknown factors and face uncertainty, which leads us to make wild guesses, flipping coins, or rolling a dice. Having knowledge, experience, or insights given a certain situation, helps us visualize what the outcomes could be. and how we can achieve/avoid those outcomes.

Make Your Choices!

Scenario 1

You are locked inside a room with 3 doors to move out of the locked room and you need to find a safe door to get your way out. Behind the 1st door is a lake with a deadly shark. The 2nd door has a mad psychopath ready to kill with a weapon and the third one has a lion that has not eaten since the last 2 months.

Which door would you choose? and Why?

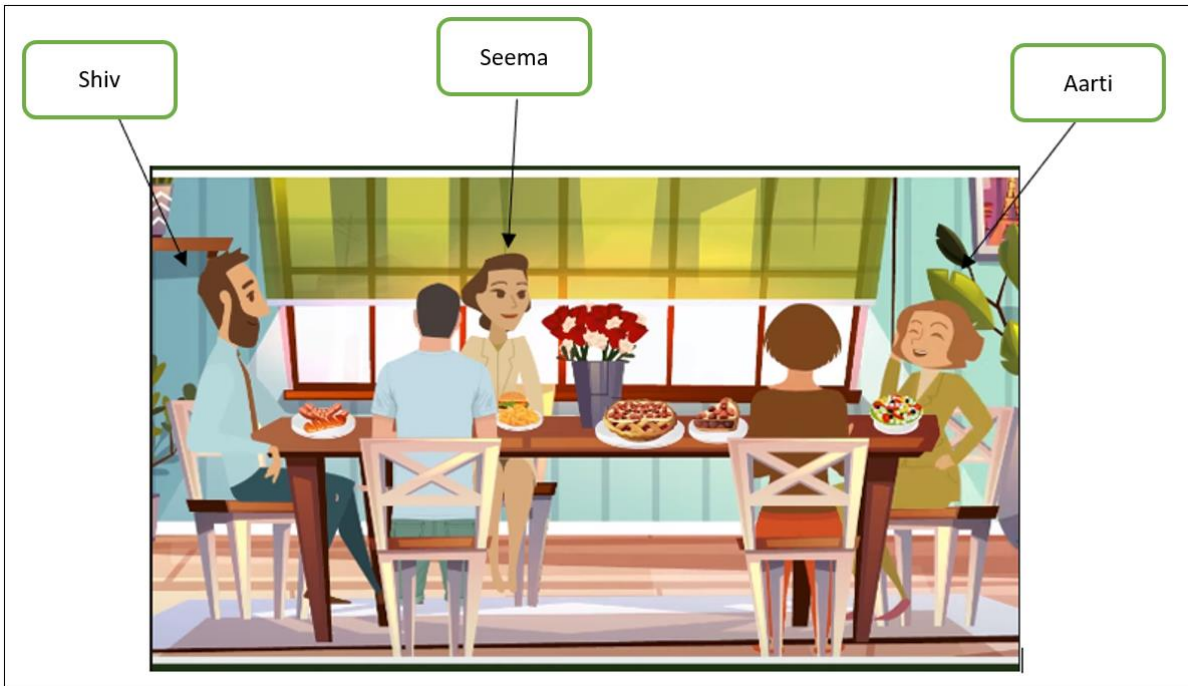


The answer is gate number 3. The reason being that since the lion has not eaten for 2 months, he wouldn't have survived till now and would already be dead . This makes going out from gate 3 the correct option.

Scenario 2

Aarti invited four of her friends to her House.. They hadn't seen each other in a long time, so they chatted all night long and had a good time. In the morning, two of the friends Aarti had invited, died. The police arrived at the house and found that both the friends were poisoned and that the poison was in the strawberry pie. The three surviving friends told the police that they hadn't eaten the pie. The police asked, " Why didn't you eat the pie ?". Shiv said, " I am allergic to strawberries.". Seema said, " I am on a diet." And Aarti said, "I ate too many strawberries while cooking the pie, I just didn't want anymore."

The policemen looked at the pictures of the party and immediately identified the murderer.

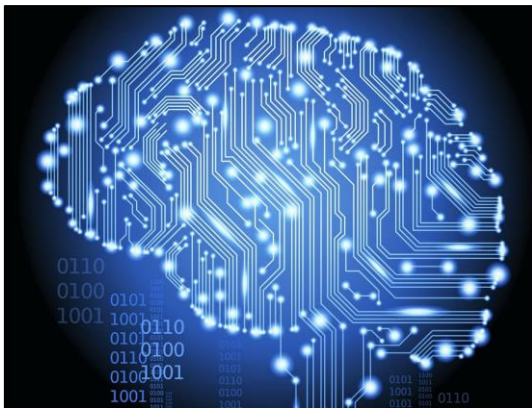


Look at the picture and identify who is the murderer? Also state why do you think this is the murderer?

The answer is Seema, can you guess how the police could tell? It's because she said she is on a diet and in the picture, she is eating a burger and fries which means she lied.

The above scenarios show that it's the information which helps humans take good decisions.

What is Artificial Intelligence?



When a machine possesses the ability to mimic human traits, i.e., make decisions, predict the future, learn and improve on its own, it is said to have artificial intelligence.

In other words, you can say that a machine is artificially intelligent when it can accomplish tasks by itself - collect data, understand it, analyse it, learn from it, and improve it. You will get to know more about it in the next unit.

But, what makes a machine intelligent?

How do machines become Artificially Intelligent?

Humans become more and more intelligent with time as they gain experiences during their lives.

For example, in elementary school, we learn about alphabets and eventually we move ahead to making words with them. As we grow, we become more and more fluent in the language as we keep learning new words and use them in our conversations.

Another example is how we learn walking. Initially a baby struggles to walk. He takes help from others while learning how to walk and once he knows it, he keeps on upgrading it by learning how to run, jump, etc.

Similarly, machines also become intelligent once they are trained with some information which helps them achieve their tasks. AI machines also keep updating their knowledge to optimise their output.



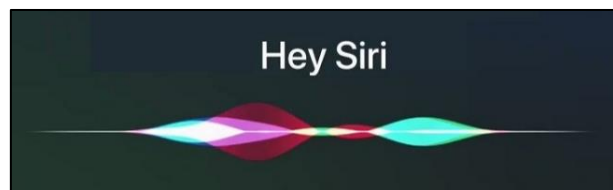
Applications of Artificial Intelligence around us

Whether we notice it or not, we are surrounded by machines that work on AI. They are becoming a crucial part of our everyday life and provide us with an ease of having even some of the most complicated and time-consuming tasks being done at the touch of a button or by the simple use of a sensor.



Every now and then, we surf the internet for things on Google without realizing how efficiently Google always responds to us with accurate answers. Not only does it come up with results to our search in a matter of seconds, it also suggests and auto-corrects our typed sentences.

We nowadays have pocket assistants that can do a lot of tasks at just one command. Alexa, Google Assistant, Cortana, Siri are some very common examples of the voice assistants which are a major part of our digital devices.



To help us navigate to places, apps like UBER and Google Maps come in handy. Thus, one no longer needs to stop repeatedly to ask for directions.

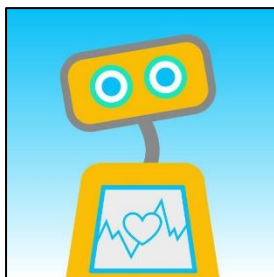
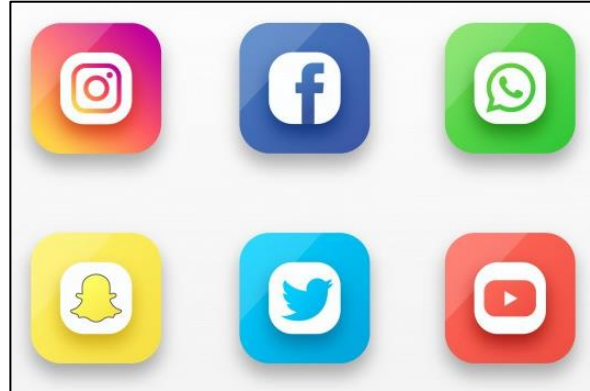
AI has completely enhanced the gaming experience for its users. A lot of games nowadays are backed up with AI which helps in enhancing the graphics, come up with new difficulty levels, encourage gamers, etc.





AI has not only made our lives easier but has also been taking care of our habits, likes, and dislikes. This is why platforms like Netflix, Amazon, Spotify, YouTube etc. show us recommendations on the basis of what we like.

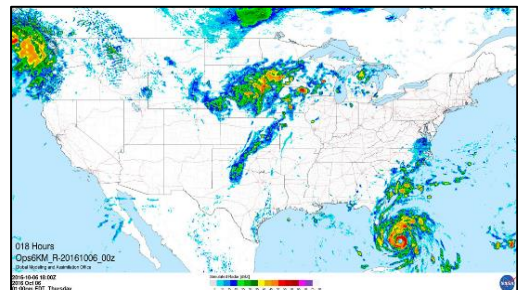
Well, the recommendations are not just limited to our preferences, they even cater to our needs of connecting with friends on social media platforms with apps like Facebook and Instagram. They also send us customized notifications about our online shopping details, auto-create playlists according to our requests and so on. Taking selfies was never this fun as Snapchat filters make them look so cool.



This isn't all. AI is also being used to monitor our health. A lot of chatbots and other health apps are available, which continuously monitor the physical and mental health of its users.



These applications are not limited to smart devices but also vary to humanoids like Sophia, the very first humanoid robot sophisticated enough to get citizenship, biometric security systems like the face locks we have in our phones, real-time language translators, weather forecasts, and whatnot! This list is huge, and this module will go on forever if we keep tabulating them. So, take some time, discuss with a friend and identify more and more AI applications around you!



What is not AI?

Since we have a lot of different technologies which exist around us in today's time, it is very common for us to misunderstand any other technology as AI. That is why, we need to have a clear distinction between what is AI and what is not.

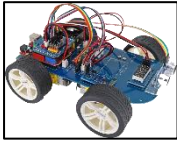
As we discussed earlier, any machine that has been trained with data and can make decisions/predictions on its own can be termed as AI. Here, the term 'training' is important.



A fully automatic washing machine can work on its own, but it requires human intervention to select the parameters of washing and to do the necessary preparation for it to function correctly before each wash, which makes it an example of automation, not AI.



An air conditioner can be turned on and off remotely with the help of internet but still needs a human touch. This is an example of Internet of Things (IoT). Also, every now and then we get to know about robots which might follow a path or maybe can avoid obstacles but need to be primed accordingly each time.

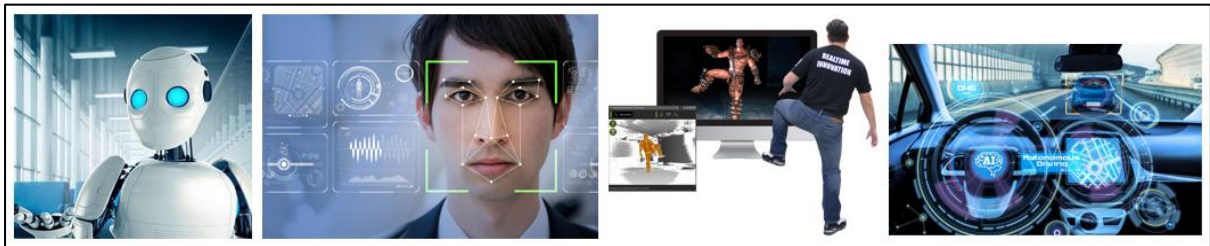


We also get to see a lot of projects which can automate our surroundings with the help of sensors. Here too, since the bot or the automation machine is not trained with any data, it does not count as AI.

Also, it would be valid to say that not all the devices which are termed as "smart" are AI-enabled. For example, a TV does not become AI-enabled if it is a smart one, it gets the power of AI when it is able to think and process on its own.

Just as humans learn how to walk and then improve this skill with the help of their experiences, an AI machine too gets trained first on the training data and then optimises itself according to its own experiences which makes AI different from any other technological device/machine.

But well, surely these other technologies too can be integrated with AI to provide the users with a much better and immersive experience!



Robotics and AI can definitely open the doors to humanoids and self-driving cars, AI when merged with Internet of things can give rise to cloud computing of data and remote access of AI tools, automation along with AI can help in achieving voice automated homes and so on. Such integrations can help us get the best of both worlds!

Introduction to AI: Basics of AI

As discussed in the last chapter, Artificial Intelligence has always been a term which intrigues people all over the world. Various organisations have coined their own versions of defining Artificial Intelligence. Some of them are mentioned below:

NITI Aayog: National Strategy for Artificial Intelligence

AI refers to the ability of machines to perform cognitive tasks like thinking, perceiving, learning, problem solving and decision making. Initially conceived as a technology that could mimic human intelligence, AI has evolved in ways that far exceed its original conception. With incredible advances made in data collection, processing and computation power, intelligent systems can now be deployed to take over a variety of tasks, enable connectivity and enhance productivity.

World Economic Forum

Artificial intelligence (AI) is the software engine that drives the Fourth Industrial Revolution. Its impact can already be seen in homes, businesses and political processes. In its embodied form of robots, it will soon be driving cars, stocking warehouses and caring for the young and elderly. It holds the promise of solving some of the most pressing issues facing society, but also presents challenges such as inscrutable “black box” algorithms, unethical use of data and potential job displacement. As rapid advances in machine learning (ML) increase the scope and scale of AI’s deployment across all aspects of daily life, and as the technology itself can learn and change on its own, multi-stakeholder collaboration is required to optimize accountability, transparency, privacy and impartiality to create trust.

European Artificial Intelligence (AI) leadership, the path for an integrated vision

AI is not a well-defined technology and no universally agreed definition exists. It is rather a cover term for techniques associated with data analysis and pattern recognition. AI is not a new technology, having existed since the 1950s. While some markets, sectors and individual businesses are more advanced than others, AI is still at a relatively early stage of development, so that the range of potential applications, and the quality of most existing applications, have ample margins left for further development and improvement.

Encyclopaedia Britannica

Artificial intelligence (AI), is the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience.

As you can see, Artificial Intelligence is a vast domain. Everyone looks at AI in a different way according to their mindset. Now, according to your knowledge of AI, start filling the KWLH chart:

K	• What I Know?
W	• What I Want to know?
L	• What have I learned?
H	• How I learnt this?

What do you know about Artificial Intelligence (AI)?

What do you want to know about AI?

What have you learnt about AI?

How have you learnt this about AI?

In other words, AI can be defined as:

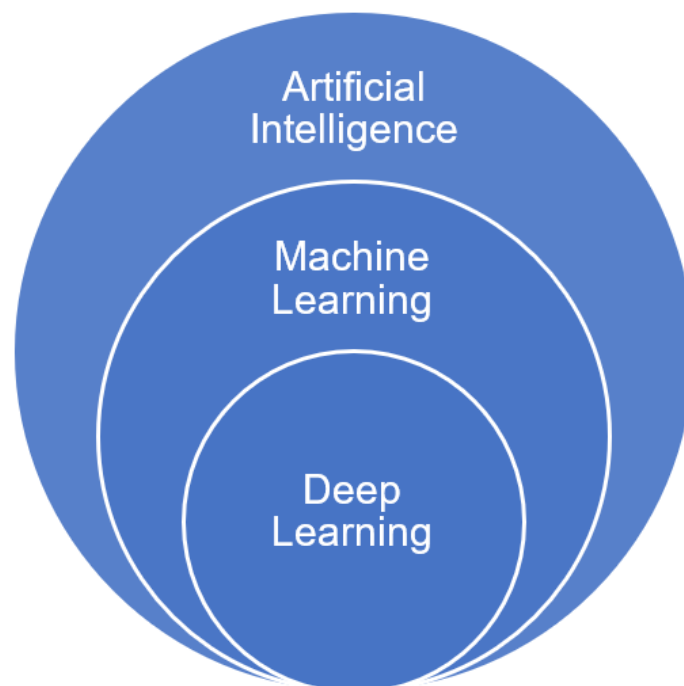
AI is a form of Intelligence; a type of technology and a field of study.

AI theory and development of computer systems (both machines and software) enables machines to perform tasks that normally require human intelligence.

Artificial Intelligence covers a broad range of domains and applications and is expected to impact every field in the future. Overall, its core idea is building machines and algorithms which are capable of performing computational tasks that would otherwise require human like brain functions.

AI, ML & DL

As you have been progressing towards building AI readiness, you must have come across a very common dilemma between Artificial Intelligence (AI) and Machine Learning (ML). Many times, these terms are used interchangeably but are they the same? Is there no difference in Machine Learning and Artificial Intelligence? Is Deep Learning (DL) Also Artificial Intelligence? What exactly is Deep Learning? Let us see.



Artificial Intelligence (AI)

Refers to any technique that enables computers to mimic human intelligence. It gives the ability to machines to recognize a human's face; to move and manipulate objects; to understand the voice commands by humans, and also do other tasks. The AI-enabled machines think algorithmically and execute what they have been asked for intelligently.

Machine Learning (ML)

It is a subset of Artificial Intelligence which enables machines to improve at tasks with experience (data). The intention of Machine Learning is to enable machines to learn by themselves using the provided data and make accurate Predictions/ Decisions.

Deep Learning (DL)

It enables software to train itself to perform tasks with vast amounts of data. In Deep Learning, the machine is trained with huge amounts of data which helps it in training itself around the data. Such machines are intelligent enough to develop algorithms for themselves. Deep Learning is the most advanced form of Artificial Intelligence out of these three. Then comes Machine Learning which is intermediately intelligent and Artificial Intelligence covers all the concepts and algorithms which, in some way or the other mimic human intelligence.

There are a lot of applications of AI out of which few are those which come under ML out of which very few can be labelled as DL. Therefore, Machine Learning (ML) and Deep Learning (DL) are part of Artificial Intelligence (AI), but not everything that is Machine learning will be Deep learning.

Introduction to AI Domains

Artificial Intelligence becomes intelligent according to the training which it gets. For training, the machine is fed with datasets. According to the applications for which the AI algorithm is being developed, the data which is fed into it changes. With respect to the type of data fed in the AI model, AI models can be broadly categorised into three domains:

Data Sciences

Computer Vision

Natural Language Processing

Data Sciences

Data sciences is a domain of AI related to data systems and processes, in which the system collects numerous data, maintains data sets and derives meaning/sense out of them.

The information extracted through data science can be used to make a decision about it.

Example of Data Science



Price Comparison Websites

These websites are being driven by lots and lots of data. If you have ever used these websites, you would know, the convenience of comparing the price of a product from multiple vendors at one place. PriceGrabber, PriceRunner, Jungle, Shopzilla, DealTime are some examples of price comparison websites. Now a days, price comparison website can be found in almost every domain such as technology, hospitality, automobiles, durables, apparels etc.

Computer Vision

Computer Vision, abbreviated as CV, is a domain of AI that depicts the capability of a machine to get and analyse visual information and afterwards predict some decisions about it. The entire process involves image acquiring, screening, analysing, identifying and extracting information. This extensive processing helps computers to understand any visual content and act on it accordingly. In computer vision, Input to machines can be photographs, videos and pictures from thermal or infrared sensors, indicators and different sources.

Computer vision related projects translate digital visual data into descriptions. This data is then turned into computer-readable language to aid the decision-making process. The main objective of this domain of AI is to teach machines to collect information from pixels.

Examples of Computer Vision



Self-Driving cars/ Automatic Cars

CV systems scan live objects and analyse them, based on whether the car decides to keep running or to stop.



Face Lock in Smartphones

Smartphones nowadays come with the feature of face locks in which the smartphone's owner can set up his/her face as an unlocking mechanism for it. The front camera detects and captures the face and saves its features during initiation. Next time onwards, whenever the features match, the phone is unlocked.

Natural Language Processing

Natural Language Processing, abbreviated as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. *Natural language* refers to language that is spoken and written by people, and natural language processing (NLP) attempts to extract information from the spoken and written word using algorithms.

The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable.

Examples of Natural Language Processing



Email filters

Email filters are one of the most basic and initial applications of NLP online. It started out with spam filters, uncovering certain words or phrases that signal a spam message.



Smart assistants

Smart assistants like Apple's Siri and Amazon's Alexa recognize patterns in speech, then infer meaning and provide a useful response.

AI Ethics

Nowadays, we are moving from the Information era to Artificial Intelligence era. Now we do not use data or information, but the intelligence collected from the data to build solutions. These solutions can even recommend the next TV show or movies you should watch on Netflix.

We can proudly say that India is leading in the AI usage trends, so we need to keep aspects relating to ethical practices in mind while developing solutions using AI. Let us understand some of the ethical concerns in detail.

Moral Issues: Self-Driving Cars

Scenario 1:

Let us imagine that we are in year 2030. Self-Driving cars which are just a concept in today's time are now on roads. People like us are buying them for ease and are using it for our daily transits. Of-course because of all the features which this car has, it is expensive. Now, let us assume, one day your father is going to office in his self-driving car. He is sitting in the back seat as the car is driving itself. Suddenly, a small boy comes in front of this car. The incident was so sudden that the car is only able to make either of the two choices:

1. Go straight and hit the boy who has come in front of the car and injure him severely.
2. Take a sharp right turn to save the boy and smash the car into a metal pole thus damaging the car as well as injuring the person sitting in it.

With the help of this scenario, we need to understand that the developer of the car goes through all such dilemmas while developing the car's algorithm. Thus, here the morality of the developer gets transferred into the machine as what according to him/her is right would have a higher priority and hence would be the selection made by the machine.

If you were in the place of this developer and if there was no other alternative to the situation, which one of the two would you prioritise and why?

Scenario 2:

Let us now assume that the car has hit the boy who came in front of it. Considering this as an accident, who should be held responsible for it? Why?

1. The person who bought this car
2. The Manufacturing Company
3. The developer who developed the car's algorithm
4. The boy who came in front of the car and got severely injured

Here, the choices might differ from person to person and one must understand that nobody is wrong in this case. Every person has a different perspective and hence he/she takes decisions according to their moralities.

Data Privacy

The world of Artificial Intelligence revolves around Data. Every company whether small or big is mining data from as many sources as possible. More than 70% of the data collected till now has been collected in the last 3 years which shows how important data has become in recent times. It is not wrongly said that *Data is the new gold*. This makes us think:

Where do we collect data from?

Why do we need to collect data?

One of the major sources of data for many major companies is the device which all of us have in our hands all the time: Smartphones. Smartphones have nowadays become an integral part of our lives. Most of us use smartphones more than we interact with people around us. Smartphones in today's era provide us with a lot of facilities and features which have made our lives easier. Feeling hungry? Order food online. Want to shop but don't have time to go out? Go shopping online. From booking tickets to watching our favourite shows, everything is available in this one small box loaded with technology.

Another feature of smartphones nowadays is that they provide us with customised recommendations and notifications according to our choices. Let us understand this with the help of some examples:

1. When you are talking to your friend on a mobile network or on an app like WhatsApp. You tell your friend that you wish to buy new shoes and are looking for suggestions from him/her. You discuss about shoes and that is it. After some time, the online shopping websites start giving you notifications to buy shoes! They start recommending some of their products and urge you to you buy some.
2. If you search on Google for a trip to Kerala or any other destination, just after the search, all the apps on your phone which support advertisements, will start sending messages about packages that you can buy for the trip.
3. Even when you are not using your phone and talking to a person face-to-face about a book you've read recently while the phone is kept in a locked mode nearby, the phone will end up giving notifications about similar books or messages about the same book once you operate it.

In all such examples, how does the smartphone get to know about the discussions and thoughts that you have? Remember whenever you download an app and install it, it asks you for several permissions to access your phone's data in different ways. If you do not allow the app these permissions, you normally cannot access it. And to access the app and make use of it, we sometimes don't even give it a thought and allow the app to get all the permissions that it wants. Hence every now and then, the app has the permission to access various sensors which are there in your smartphone and gather data about you and your surroundings. We forget that the smartphone which we use is a box full of sensors which are powered all the time while the phone is switched on.

This leads us to a crucial question: Are we okay with sharing our data with the external world?

Why do these apps collect data?

We need to understand that the data which is collected by various applications is ethical as the smartphone users agree to it (by clicking on allow when it asks for permission and by agreeing to all the terms and conditions). But at the same time if one does not want to share his/her data with anyone, he/she can opt for alternative applications which are of similar usage and keep your data private. For example, an alternative to WhatsApp is the Telegram app which does not collect any data from us. But since WhatsApp is more popular and used by the crowd, people go for it without thinking twice.

AI Bias

Another aspect to AI Ethics is bias. Everyone has a bias of their own no matter how much one tries to be unbiased, we in some way or the other have our own biases even towards smaller things. Biases are not negative all the time. Sometimes, it is required to have a bias to control a situation and keep things working.

When we talk about a machine, we know that it is artificial and cannot think on its own. It can have intelligence, but we cannot expect a machine to have any biases of its own. Any bias can transfer from the developer to the machine while the algorithm is being developed. Let us look at some of the examples:

1. Majorly, all the virtual assistants have a female voice. It is only now that some companies have understood this bias and have started giving options for male voices but since the virtual assistants came into practice, female voices are always preferred for them over any other voice. Can you think of some reasons for this?

2. If you search on Google for salons, the first few searches are mostly for female salons. This is based on the assumption that if a person is searching for a salon, in all probability it would be a female. Do you think this is a bias? If yes, then is it a Negative bias or Positive one?

Various other biases are also found in various systems which are not thought up by the machine but have got transferred from the developer intentionally or unintentionally.

AI Access

Since Artificial Intelligence is still a budding technology, not everyone has the opportunity to access it. The people who can afford AI enabled devices make the most of it while others who cannot are left behind. Because of this, a gap has emerged between these two classes of people and it gets widened with the rapid advancement of technology. Let us understand this with the help of some examples:

AI creates unemployment

AI is making people's lives easier. Most of the things nowadays are done in just a few clicks. In no time AI will manage to be able to do all the laborious tasks which we humans have been doing since long. Maybe in the coming years, AI enabled machines will replace all the people who work as labourers. This may start an era of mass unemployment where people having little or no skills may be left without jobs and others who keep up with their skills according to what is required, will flourish.

This brings us to a crossroads. On one hand where AI is advancing and improving the lives of people by working for them and doing some of their tasks, the other hand points towards the lives of people who are dependent on laborious jobs and are not skilled to do anything else.

Should AI replace laborious jobs? Is there an alternative for major unemployment?

Should AI not replace laborious jobs? Will the lives of people improve if they keep on being unskilled?

Here, we need to understand that to overcome such an issue, one needs to be open to changes. As technology is advancing with time, humans need to make sure that they are a step ahead and understand this technology with its pros and cons.

AI for kids

As we all can see, kids nowadays are smart enough to understand technology from a very early age. As their thinking capabilities increase, they start becoming techno-savvy and eventually they learn everything more easily than an adult. But should technology be given to children so young?

Consider this: A young boy in class 3 has got some Maths homework to finish. He is sitting at a table which has the Google chat bot - Alexa on it, and he is struggling with his homework. Soon, he starts asking Alexa to answer all his questions. Alexa replies with answers and the boy simply writes them down in his notebook.

While this scenario seems funny, it still has some concerns related to it. On one hand where it is good that the boy knows how to use technology effectively, on the other hand he uses it to complete his

homework without really learning anything since he is not applying his brain to solve the Math problems. So, while he is smart, he might not be getting educated properly.

Is it ethical to let the boy use technology to help in this manilr?

Conclusion

Despite AI's promises to bring forth new opportunities, there are certain associated risks that need to be mitigated appropriately and effectively. To give a better perspective, the ecosystem and the socio-technical environment in which the AI systems are embedded needs to be more trustworthy.

AI Project Cycle

In this chapter, we will revisit the concept of AI Project Cycle.

Introduction

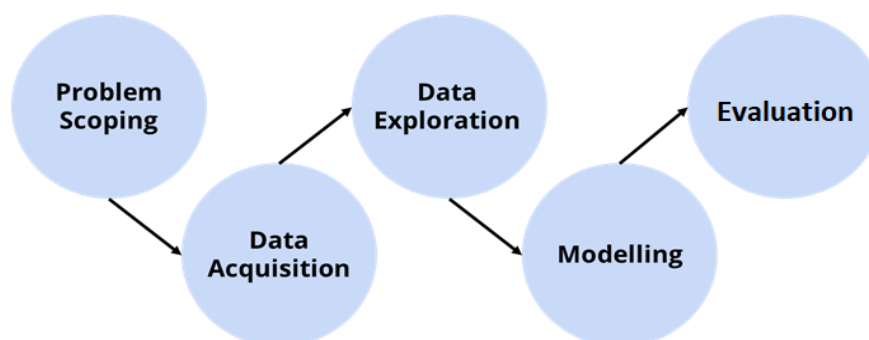
Let us assume that you have to make a greeting card for your mother as it is her birthday. You are very excited about it and have thought of many ideas to execute the same. Let us look at some of the steps which you might take to accomplish this task:

1. Look for some cool greeting card ideas from different sources. You might go online and checkout some videos or you may ask someone who has knowledge about it.
2. After finalising the design, you would make a list of things that are required to make this card.
3. You will check if you have the material with you or not. If not, you could go and get all the items required, ready for use.
4. Once you have everything with you, you would start making the card.
5. If you make a mistake in the card somewhere which cannot be rectified, you will discard it and start remaking it.
6. Once the greeting card is made, you would gift it to your mother.

Are these steps relatable?

Do you think your steps might differ? If so, write them down!

These steps show how we plan to execute the tasks around us. Consciously or Subconsciously our mind makes up plans for every task which we have to accomplish which is why things become clearer in our mind. Similarly, if we have to develop an AI project, the AI Project Cycle provides us with an appropriate framework which can lead us towards the goal. The AI Project Cycle mainly has 5 stages:



Starting with Problem Scoping, you set the goal for your AI project by stating the problem which you wish to solve with it. Under problem scoping, we look at various parameters which affect the problem we wish to solve so that the picture becomes clearer.

To proceed,

- You need to acquire data which will become the base of your project as it will help you in understanding what the parameters that are related to problem scoping are.
- You go for data acquisition by collecting data from various reliable and authentic sources. Since the data you collect would be in large quantities, you can try to give it a visual image of different types of representations like graphs, databases, flow charts, maps, etc. This makes it easier for you to interpret the patterns which your acquired data follows.
- After exploring the patterns, you can decide upon the type of model you would build to achieve the goal. For this, you can research online and select various models which give a suitable output.
- You can test the selected models and figure out which is the most efficient one.
- The most efficient model is now the base of your AI project and you can develop your algorithm around it.
- Once the modelling is complete, you now need to test your model on some newly fetched data. The results will help you in evaluating your model and improving it.
- Finally, after evaluation, the project cycle is now complete and what you get is your AI project.

Let us understand each stage of the AI Project Cycle in detail.

Problem Scoping

It is a fact that we are surrounded by problems. They could be small or big, sometimes ignored or sometimes even critical. Many times, we become so used to a problem that it becomes a part of our life. Identifying such a problem and having a vision to solve it, is what Problem Scoping is about. A lot of times we are unable to observe any problem in our surroundings. In that case, we can take a look at the Sustainable Development Goals. 17 goals have been announced by the United nations which are termed as the Sustainable Development Goals. The aim is to achieve these goals by the end of 2030. A pledge to do so has been taken by all the member nations of the UN.



* Images shown here are the property of individual organisations and are used here for reference purpose only.

Here are the 17 SDGs. Let's take a look:



As you can see, many goals correspond to the problems which we might observe around us too. One should look for such problems and try to solve them as this would make many lives better and help our country achieve these goals.

Scoping a problem is not that easy as we need to have a deeper understanding around it so that the picture becomes clearer while we are working to solve it. Hence, we use the 4Ws Problem Canvas to help us out.

4Ws Problem Canvas

The 4Ws Problem canvas helps in identifying the key elements related to the problem.



Let us go through each of the blocks one by one.

Who?

The "Who" block helps in analysing the people getting affected directly or indirectly due to it. Under this, we find out who the 'Stakeholders' to this problem are and what we know about them. Stakeholders are the people who face this problem and would be benefitted with the solution. Here is the Who Canvas:

Who is having the problem?

1. Who are the stakeholders?

2. What do you know about them?

What?

Under the “What” block, you need to look into what you have on hand. At this stage, you need to determine the nature of the problem. What is the problem and how do you know that it is a problem? Under this block, you also gather evidence to prove that the problem you have selected actually exists. Newspaper articles, Media, announcements, etc are some examples. Here is the What Canvas:

What is the nature of the problem?

1. What is the problem?

2. How do you know it is a problem?

Where?

Now that you know who is associated with the problem and what the problem actually is; you need to focus on the context/situation/location of the problem. This block will help you look into the situation in which the problem arises, the context of it, and the locations where it is prominent. Here is the Where Canvas:

Where does the problem arise?

1. What is the context/situation in which the stakeholders experience the problem?

Why?

You have finally listed down all the major elements that affect the problem directly. Now it is convenient to understand who the people that would be benefitted by the solution are; what is to be solved; and where will the solution be deployed. These three canvases now become the base of why you want to solve this problem. Thus, in the “Why” canvas, think about the benefits which the stakeholders would get from the solution and how it will benefit them as well as the society.

Why do you believe it is a problem worth solving?

1. What would be of key value to the stakeholders?

2. How would it improve their situation?

After filling the 4Ws Problem canvas, you now need to summarise all the cards into one template. The Problem Statement Template helps us to summarise all the key points into one single Template so that in future, whenever there is need to look back at the basis of the problem, we can take a look at the Problem Statement Template and understand the key elements of it.

Our	[stakeholder(s)]	Who
has /have a problem that	[issue, problem, need]	What
when / while	[context, situation]	Where
An ideal solution would	[benefit of solution for them]	Why

Data Acquisition

As we move ahead in the AI Project Cycle, we come across the second element which is : **Data Acquisition**. As the term clearly mentions, this stage is about acquiring data for the project. Let us first understand what is Data. Data can be a piece of information or facts and statistics collected together for reference or analysis. Whenever we want an AI project to be able to predict an output, we need to train it first using data.

For example, If you want to make an Artificially Intelligent system which can predict the salary of any employee based on his previous salaries, you would feed the data of his previous salaries into the machine. This is the data with which the machine can be trained. Now, once it is ready, it will predict his next salary efficiently. The previous salary data here is known as **Training Data** while the next salary prediction data set is known as the **Testing Data**.

For better efficiency of an AI project, the Training data needs to be relevant and authentic. In the previous example, if the training data was not of the previous salaries but of his expenses, the machine would not have predicted his next salary correctly since the whole training went wrong. Similarly, if the previous salary data was not authentic, that is, it was not correct, then too the prediction could have gone wrong. Hence....

For any AI project to be efficient, the training data should be authentic and relevant to the problem statement scoped.

Data Features

Look at your problem statement once again and try to find the data features required to address this issue. **Data features refer to the type of data you want to collect.** In our previous example, data features would be salary amount, increment percentage, increment period, bonus, etc.

After mentioning the Data features, you get to know what sort of data is to be collected. Now, the question arises- From where can we get this data? There can be various ways in which you can collect data. Some of them are:



Sometimes, you use the internet and try to acquire data for your project from some random websites. Such data might not be authentic as its accuracy cannot be proved. Due to this, it becomes necessary to find a reliable source of data from where some authentic information can be taken. At the same time, we should keep in mind that the data which we collect is open-sourced and not someone's property. Extracting private data can be an offence. One of the most reliable and authentic sources of information, are the open-sourced websites hosted by the government. These government portals have general information collected in suitable format which can be downloaded and used wisely.

Some of the open-sourced Govt. portals are: data.gov.in, india.gov.in

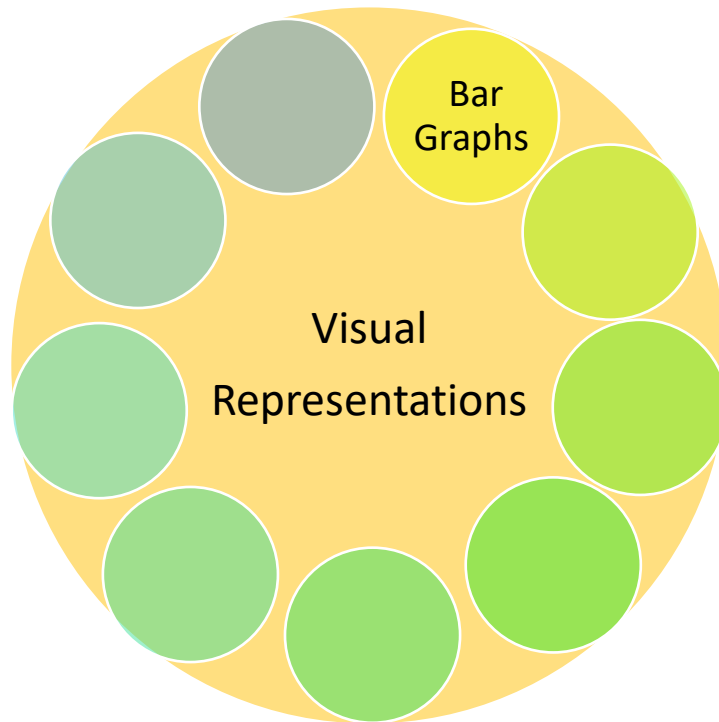
Data Exploration

In the previous modules, you have set the goal of your project and have also found ways to acquire data. While acquiring data, you must have noticed that the data is a complex entity – it is full of numbers and if anyone wants to make some sense out of it, they have to work some patterns out of it. For example, if you go to the library and pick up a random book, you first try to go through its content quickly by turning pages and by reading the description before borrowing it for yourself, because it helps you in understanding if the book is appropriate to your needs and interests or not.

Thus, to analyse the data, you need to visualise it in some user-friendly format so that you can:

- Quickly get a sense of the trends, relationships and patterns contained within the data.
- Define strategy for which model to use at a later stage.
- Communicate the same to others effectively. To visualise data, we can use various types of visual representations.

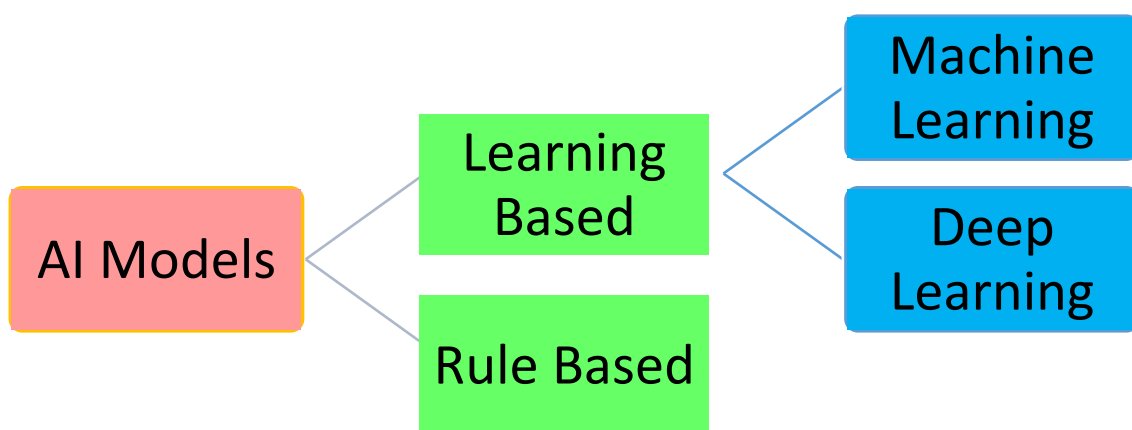
Are you aware of visual representations of data? Fill them below:



Modelling

In the previous module of Data exploration, we have seen various types of graphical representations which can be used for representing different parameters of data. The graphical representation makes the data understandable for humans as we can discover trends and patterns out of it. But when it comes to machines accessing and analysing data, it needs the data in the most basic form of numbers (which is binary – 0s and 1s) and when it comes to discovering patterns and trends in data, the machine goes in for mathematical representations of the same. The ability to mathematically describe the relationship between parameters is the heart of every AI model. Thus, whenever we talk about developing AI models, it is the mathematical approach towards analysing data which we refer to.

Generally, AI models can be classified as follows:



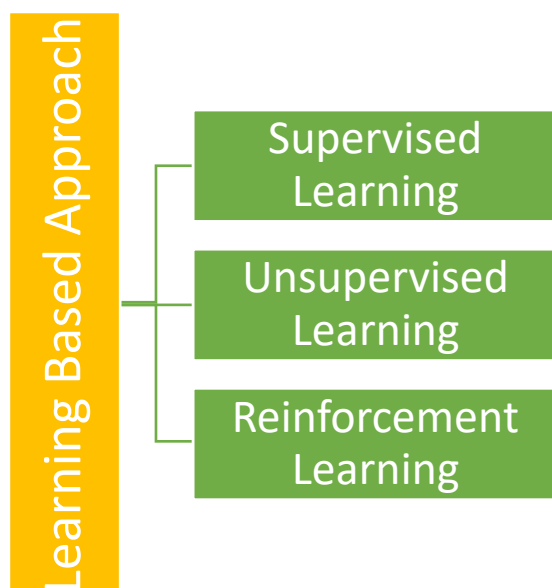
Rule Based Approach

Refers to the AI modelling where the rules are defined by the developer. The machine follows the rules or instructions mentioned by the developer and performs its task accordingly. For example, we have a dataset which tells us about the conditions on the basis of which we can decide if an elephant may be spotted or not while on safari. The parameters are: Outlook, Temperature, Humidity and Wind.

Now, let's take various possibilities of these parameters and see in which case the elephant may be spotted and in which case it may not. After looking through all the cases, we feed this data in to the machine along with the rules which tell the machine all the possibilities. The machine trains on this data and now is ready to be tested. While testing the machine, we tell the machine that Outlook = Overcast; Temperature = Normal; Humidity = Normal and Wind = Weak. On the basis of this testing dataset, now the machine will be able to tell if the elephant has been spotted before or not and will display the prediction to us. This is known as a rule-based approach because we fed the data along with rules to the machine and the machine after getting trained on them is now able to predict answers for the same. A drawback/feature for this approach is that the learning is static. The machine once trained, does not take into consideration any changes made in the original training dataset. That is, if you try testing the machine on a dataset which is different from the rules and data you fed it at the training stage, the machine will fail and will not learn from its mistake. Once trained, the model cannot improvise itself on the basis of feedbacks. Thus, machine learning gets introduced as an extension to this as in that case, the machine adapts to change in data and rules and follows the updated path only, while a rule-based model does what it has been taught once.

Learning Based Approach

Refers to the AI modelling where the machine learns by itself. Under the Learning Based approach, the AI model gets trained on the data fed to it and then is able to design a model which is adaptive to the change in data. That is, if the model is trained with X type of data and the machine designs the algorithm around it, the model would modify itself according to the changes which occur in the data so that all the exceptions are handled in this case. For example, suppose you have a dataset comprising of 100 images of apples and bananas each. These images depict apples and bananas in various shapes and sizes. These images are then labelled as either apple or banana so that all apple images are labelled 'apple' and all the banana images have 'banana' as their label. Now, the AI model is trained with this dataset and the model is programmed in such a way that it can distinguish between an apple image and a banana image according to their features and can predict the label of any image which is fed to it as an apple or a banana. After training, the machine is now fed with testing data. Now, the testing data might not have similar images as the ones on which the model has been trained. So, the model adapts to the features on which it has been trained and accordingly predicts if the image is of an apple or banana. In this way, the machine learns by itself by adapting to the new data which is flowing in. This is the machine learning approach which introduces the dynamicity in the model.



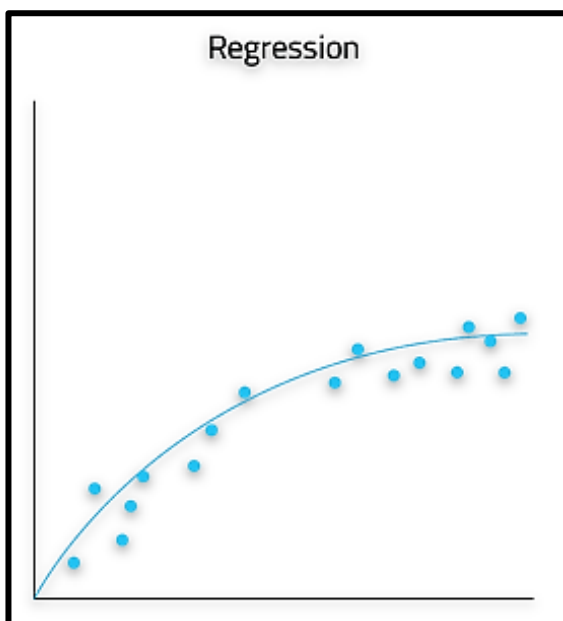
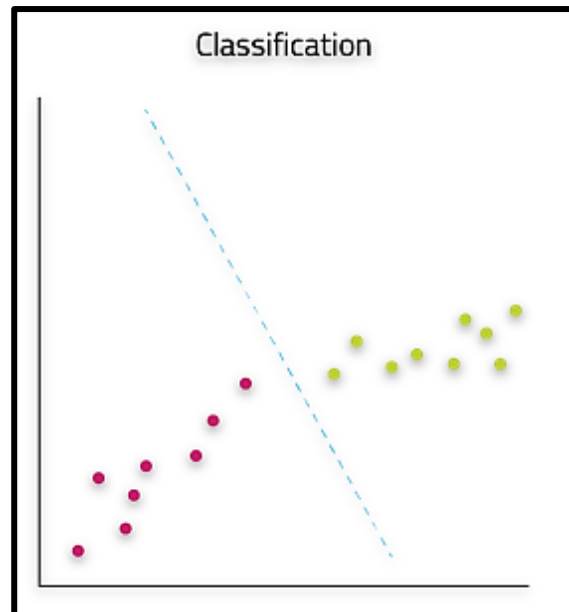
The learning-based approach can further be divided into three parts:

Supervised Learning

In a supervised learning model, the dataset which is fed to the machine is labelled. In other words, we can say that the dataset is known to the person who is training the machine only then he/she is able to label the data. A label is some information which can be used as a tag for data. For example, students get grades according to the marks they secure in examinations. These grades are labels which categorise the students according to their marks.

There are two types of Supervised Learning models:

Classification: Where the data is classified according to the labels. For example, in the grading system, students are classified on the basis of the grades they obtain with respect to their marks in the examination. This model works on discrete dataset which means the data need not be continuous.



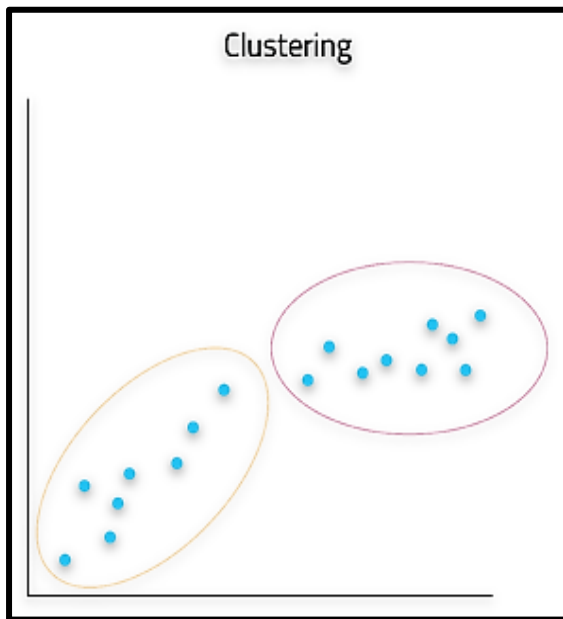
Regression: Such models work on continuous data. For example, if you wish to predict your next salary, then you would put in the data of your previous salary, any increments, etc., and would train the model. Here, the data which has been fed to the machine is continuous.

Unsupervised Learning

An unsupervised learning model works on unlabelled dataset. This means that the data which is fed to the machine is random and there is a possibility that the person who is training the model does not have any information regarding it. The unsupervised learning models are used to identify relationships, patterns and trends out of the data which is fed into it. It helps the user in understanding what the data is about and what are the major features identified by the machine in it.

For example, you have a random data of 1000 dog images and you wish to understand some pattern out of it, you would feed this data into the unsupervised learning model and would train the machine on it. After training, the machine would come up with patterns which it was able to identify out of it. The Machine might come up with patterns which are already known to the user like colour or it might even come up with something very unusual like the size of the dogs.

Unsupervised learning models can be further divided into two categories:



Clustering: Refers to the unsupervised learning algorithm which can cluster the unknown data according to the patterns or trends identified out of it. The patterns observed might be the ones which are known to the developer or it might even come up with some unique patterns out of it.

Dimensionality Reduction: We humans are able to visualise upto 3-Dimensions only but according to a lot of theories and algorithms, there are various entities which exist beyond 3-Dimensions. For example, in Natural language Processing, the words are considered to be N-Dimensional entities. Which means that we cannot visualise them as they exist beyond our visualisation ability. Hence, to make sense out of it, we need to reduce their dimensions. Here, dimensionality reduction algorithm is used.

As we reduce the dimension of an entity, the information which it contains starts getting distorted. For example, if we have a ball in our hand, it is 3-Dimensions right now. But if we click its picture, the data transforms to 2-D as an image is a 2-Dimensional entity. Now, as soon as we reduce one dimension, at least 50% of the information is lost as now we will not know about the back of the ball. Whether the ball was of same colour at the back or not? Or was it just a hemisphere? If we reduce the dimensions further, more and more information will get lost.

Hence, to reduce the dimensions and still be able to make sense out of the data, we use Dimensionality Reduction.

Evaluation

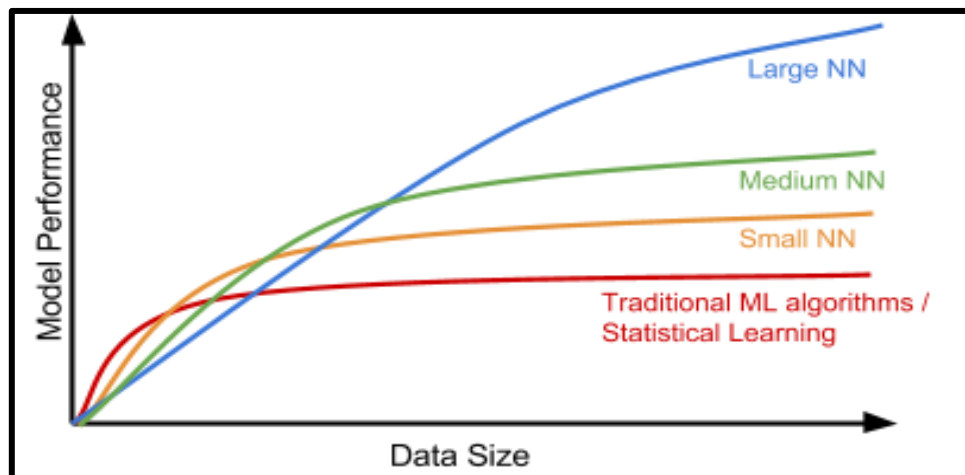
Once a model has been made and trained, it needs to go through proper testing so that one can calculate the efficiency and performance of the model. Hence, the model is tested with the help of Testing Data (which was separated out of the acquired dataset at Data Acquisition stage) and the efficiency of the model is calculated on the basis of the parameters mentioned below:



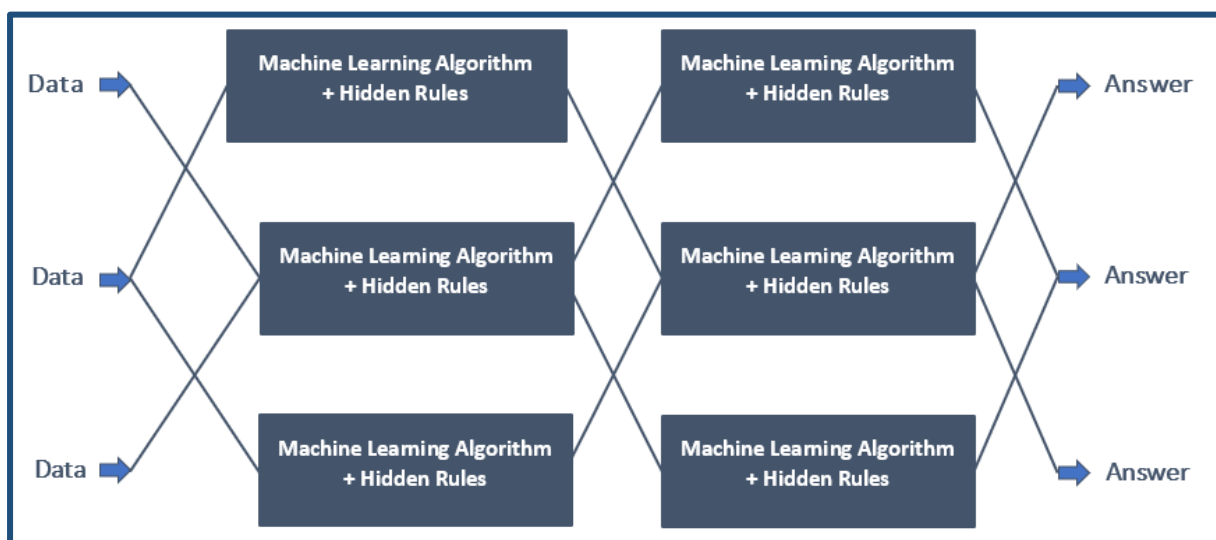
You will read more about this stage in Chapter 7.

Neural Networks

Neural networks are loosely modelled after how neurons in the human brain behave. The key advantage of neural networks are that they are able to extract data features automatically without needing the input of the programmer. A neural network is essentially a system of organizing machine learning algorithms to perform certain tasks. It is a fast and efficient way to solve problems for which the dataset is very large, such as in images.



As seen in the figure given, the larger Neural Networks tend to perform better with larger amounts of data whereas the traditional machine learning algorithms stop improving after a certain saturation point.



This is a representation of how neural networks work. A Neural Network is divided into multiple layers and each layer is further divided into several blocks called nodes. Each node has its own task to accomplish which is then passed to the next layer. The first layer of a Neural Network is known as the input layer. The job of an input layer is to acquire data and feed it to the Neural Network. No processing occurs at the input layer. Next to it, are the hidden layers. Hidden layers are the layers in which the whole processing occurs. Their name essentially means that these layers are hidden and are not visible to the user.

Each node of these hidden layers has its own machine learning algorithm which it executes on the data received from the input layer. The processed output is then fed to the subsequent hidden layer

of the network. There can be multiple hidden layers in a neural network system and their number depends upon the complexity of the function for which the network has been configured. Also, the number of nodes in each layer can vary accordingly. The last hidden layer passes the final processed data to the output layer which then gives it to the user as the final output. Similar to the input layer, output layer too does not process the data which it acquires. It is meant for user-interface.

Some of the features of a Neural Network are listed below:



Neural Network systems are modelled on the human brain and nervous system.



They are able to automatically extract features without input from the programmer.



Every neural network node is essentially a machine learning algorithm.



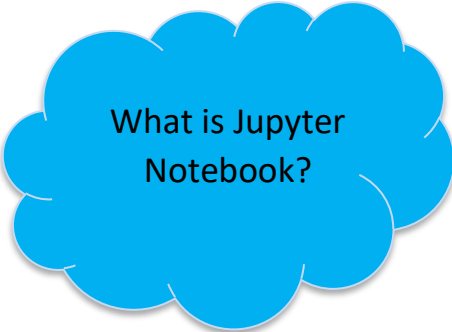
It is useful when solving problems for which the data set is very large.

Advance Python

Recap

In this section, we will go through a quick refreshing session around Python concepts and Jupyter notebook. Along with this we will talk about newer concepts like packages, virtual environments, etc.

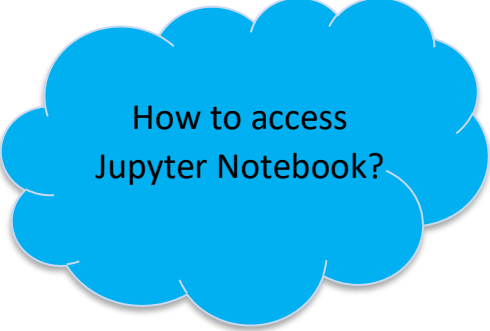
Recap 1: Jupyter Notebook



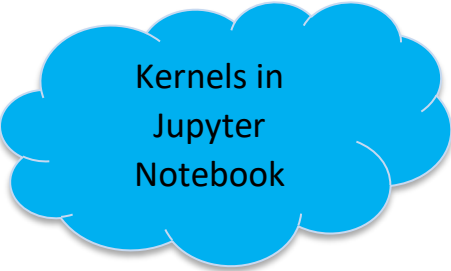
What is Jupyter Notebook?

The Jupyter Notebook is an incredibly powerful tool for interactively developing and presenting AI related projects. The Jupyter project is the successor to the earlier IPython Notebook, which was first published as a prototype in 2010. Although it is possible to use many different programming languages within Jupyter Notebooks, Python remains the most commonly used language for it. In other words, we can say that the Jupyter Notebook is an open source web application that can be used to create and share documents that contain live code, equations, visualizations, and text.

The easiest way to install and start using Jupyter Notebook is through Anaconda. Anaconda is the most widely used Python distribution for data science and comes pre-loaded with all the most popular libraries and tools. With Anaconda, comes the Anaconda Navigator through which we can scroll around all the applications which come along with it. Jupyter notebook can easily be accessed using the Anaconda Prompt with the help of a local host.



How to access Jupyter Notebook?



Kernels in Jupyter Notebook

To work with Jupyter Notebook, it is necessary to have a kernel on which it operates. A kernel provides programming language support in Jupyter. IPython is the default kernel for Jupyter Notebook. Therefore, whenever we need to work with Jupyter Notebook in a virtual environment, we first need to install a kernel inside the environment in which the Jupyter notebook will run.

Introduction to Virtual Environments

What?

A virtual environment is a tool that helps to keep dependencies required by different projects separated, by creating isolated Python virtual environments for them. This is one of the most important tools that most of the Python developers use.

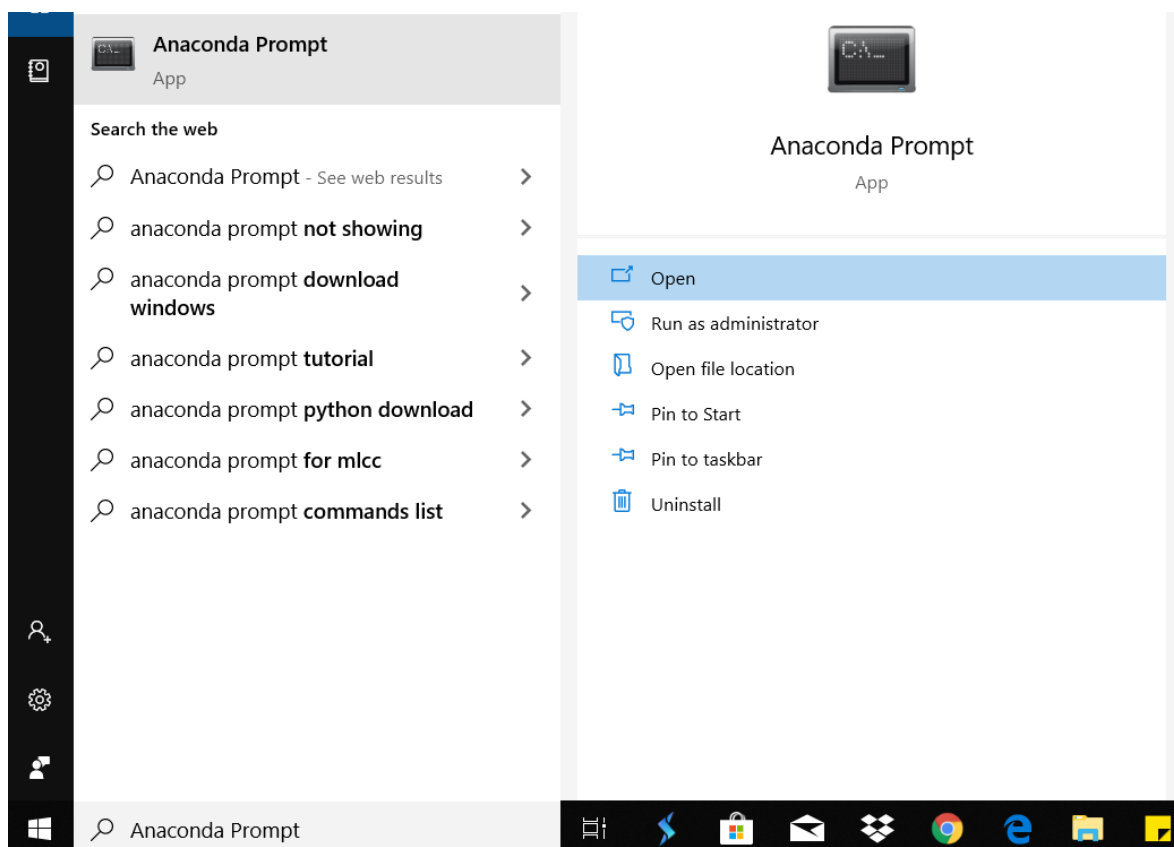
Why?

Imagine a scenario where we are working on two Python-based projects and one of them works on Python 2.7 and the other uses Python 3.7. In such situations virtual environment can be really useful to maintain dependencies of both the projects as the virtual environments will make sure that these dependencies are not conflicting with each other and no impact reaches the base environment at any point in time. Thus, different projects developed in the system might have another environment to keep their dependencies isolated from each other.

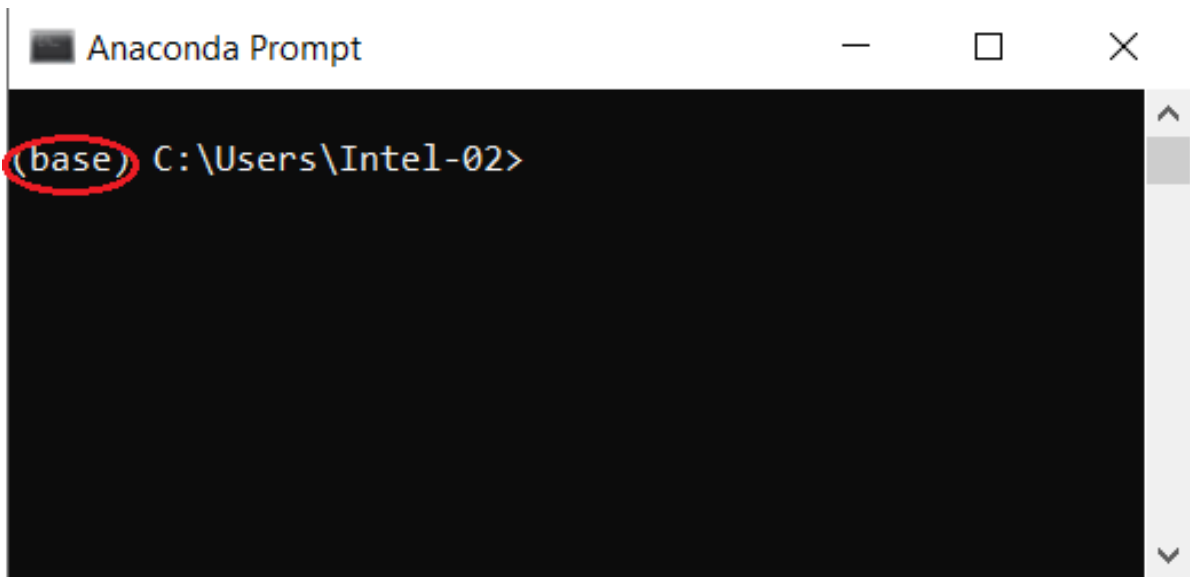
How?

Creating virtual environments is an easy task with Anaconda distribution. Steps to create one are:

1. Open Anaconda Prompt.

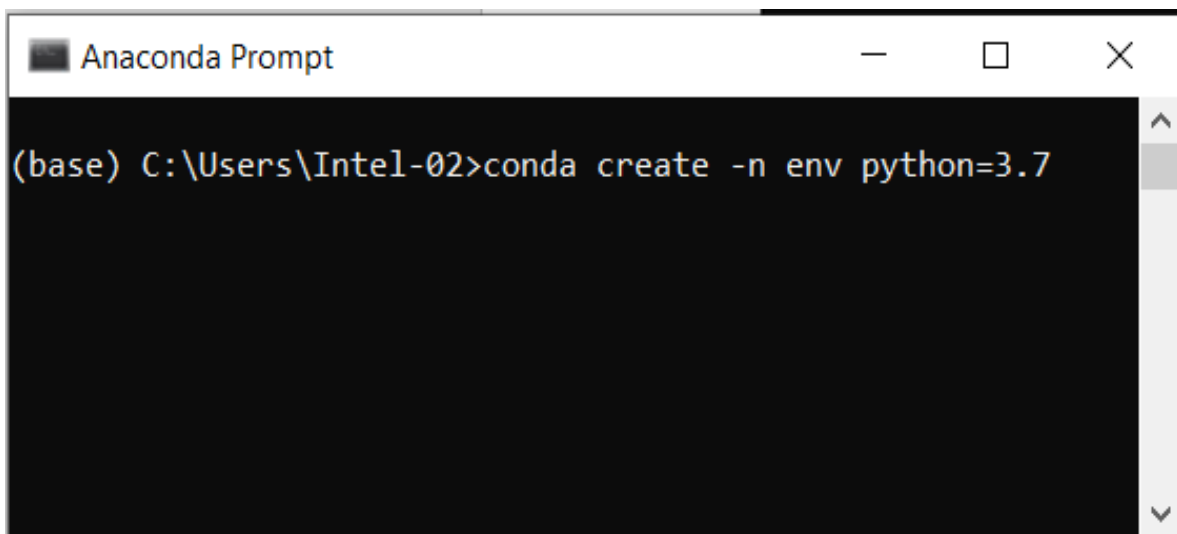


- As we open the Anaconda prompt, we can see that in the beginning of the prompt message, the term **(base)** is written. This is the default environment in which the anaconda works. Now, we can create our own virtual environment and use it so that the base does not get affected by anything that is done in the virtual environment.



```
Anaconda Prompt
(base) C:\Users\Intel-02>
```

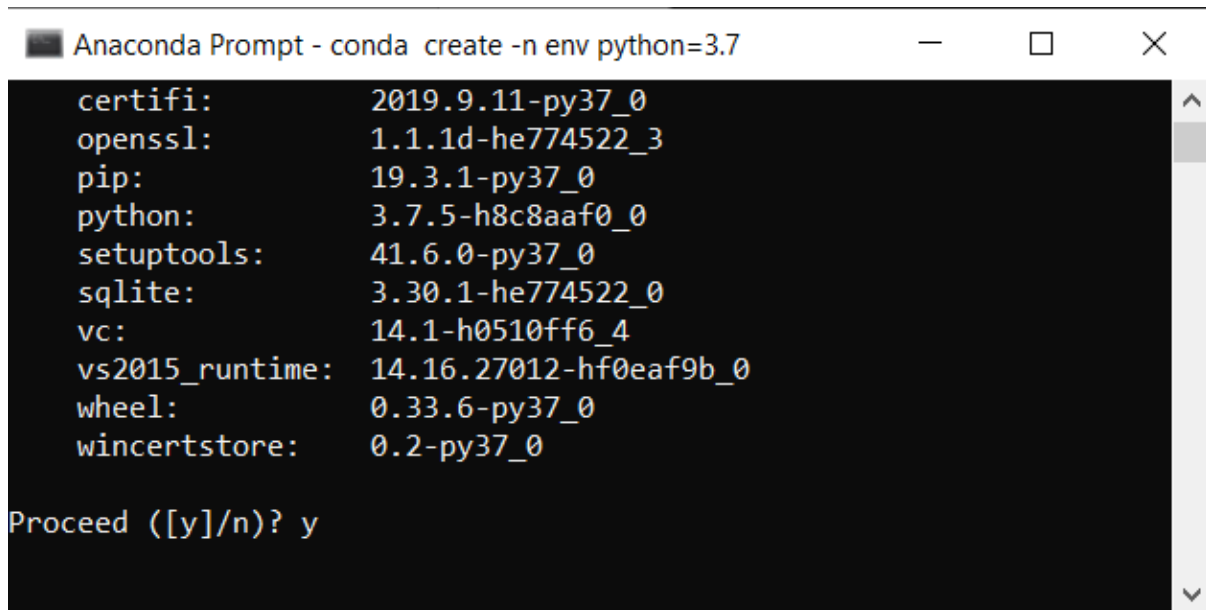
- Let us now create a virtual environment named **env**. To create the environment, write `conda create -n env python=3.7`



```
Anaconda Prompt
(base) C:\Users\Intel-02>conda create -n env python=3.7
```

This code will create an environment named **env** and will install Python 3.7 and other basic packages into it.

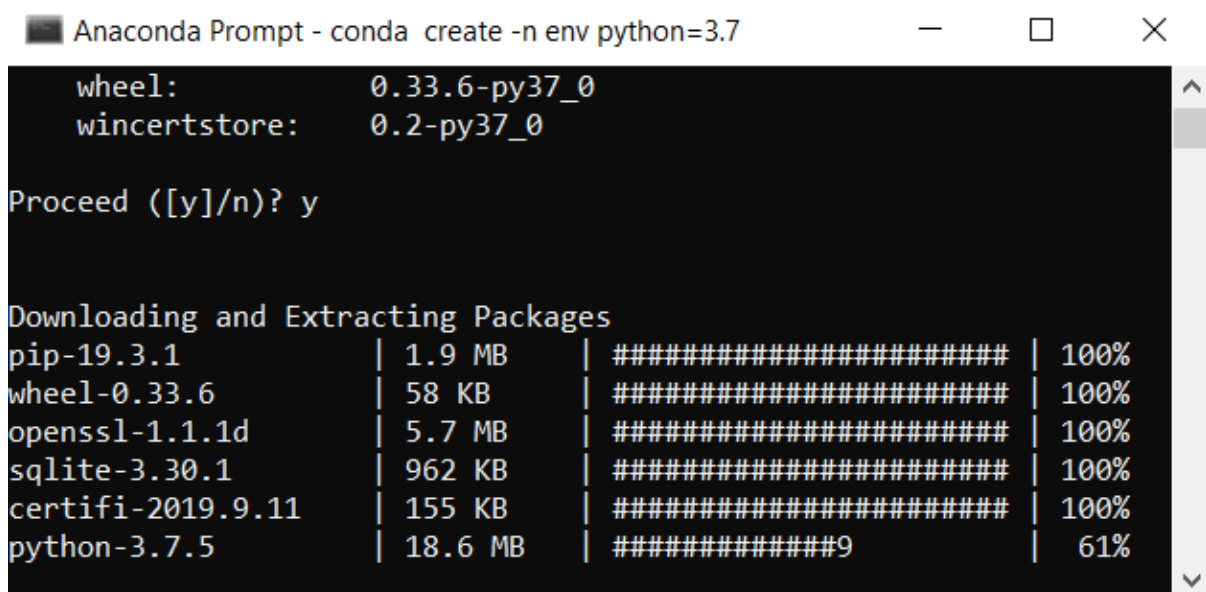
4. After some processing, the prompt will ask if we wish to proceed with installations or not. Type **Y** on it and press Enter. Once we press Enter, the packages will start getting installed in the environment.



```
Anaconda Prompt - conda create -n env python=3.7
certifi:      2019.9.11-py37_0
openssl:     1.1.1d-he774522_3
pip:         19.3.1-py37_0
python:      3.7.5-h8c8aaf0_0
setuptools:  41.6.0-py37_0
sqlite:     3.30.1-he774522_0
vc:         14.1-h0510ff6_4
vs2015_runtime: 14.16.27012-hf0eaf9b_0
wheel:      0.33.6-py37_0
wincertstore: 0.2-py37_0

Proceed ([y]/n)? y
```

5. Depending upon the internet speed, the downloading of packages might take varied time. The processing screen will look like this:

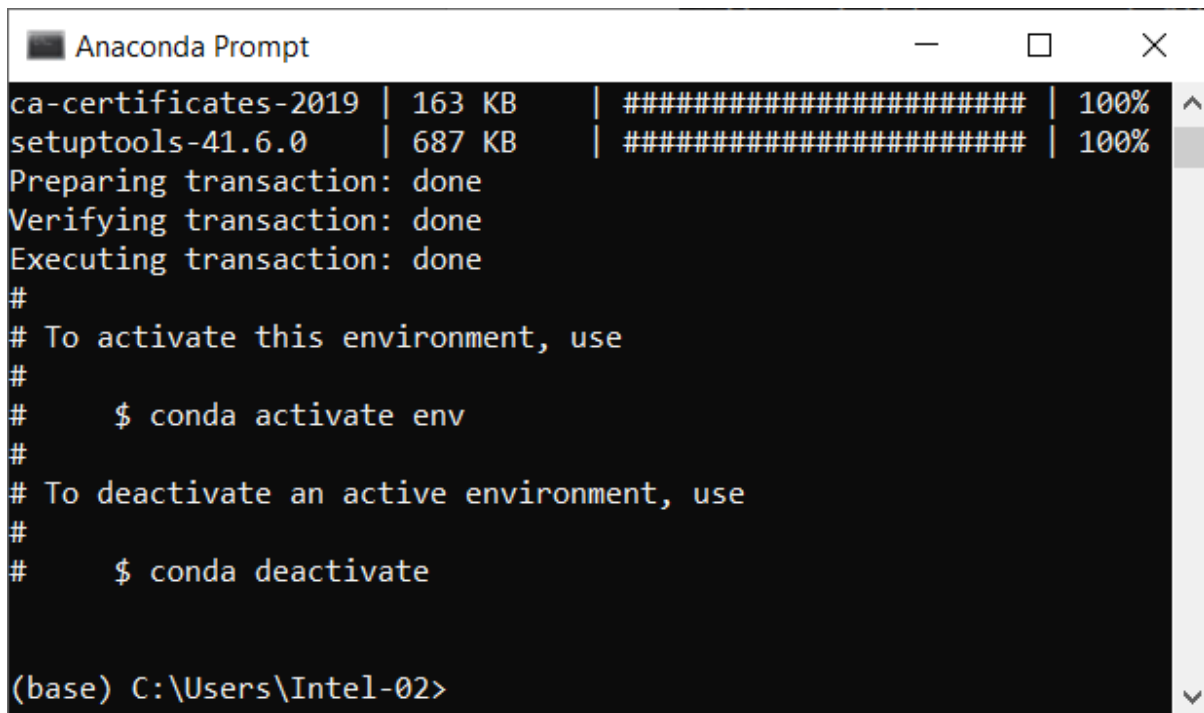


```
Anaconda Prompt - conda create -n env python=3.7
wheel:      0.33.6-py37_0
wincertstore: 0.2-py37_0

Proceed ([y]/n)? y

Downloading and Extracting Packages
pip-19.3.1      | 1.9 MB | ##### | 100%
wheel-0.33.6   | 58 KB | ##### | 100%
openssl-1.1.1d | 5.7 MB | ##### | 100%
sqlite-3.30.1  | 962 KB | ##### | 100%
certifi-2019.9.11 | 155 KB | ##### | 100%
python-3.7.5   | 18.6 MB | #####9 | 61%
```

6. Once all the packages are downloaded and installed, we will get a message like this:

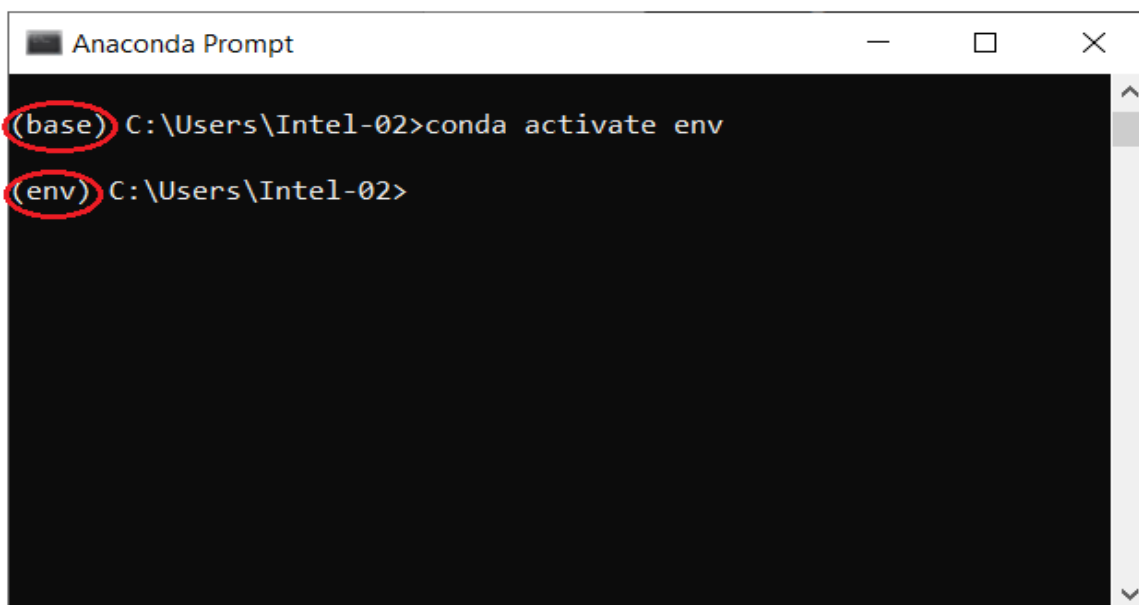


```
Anaconda Prompt
ca-certificates-2019 | 163 KB | ##### | 100%
setuptools-41.6.0 | 687 KB | ##### | 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
#
# To activate this environment, use
#
#     $ conda activate env
#
# To deactivate an active environment, use
#
#     $ conda deactivate

(base) C:\Users\Intel-02>
```

7. This shows that our environment called **env** has been successfully created. Once an environment has been successfully created, we can access it by writing the following:

conda activate env



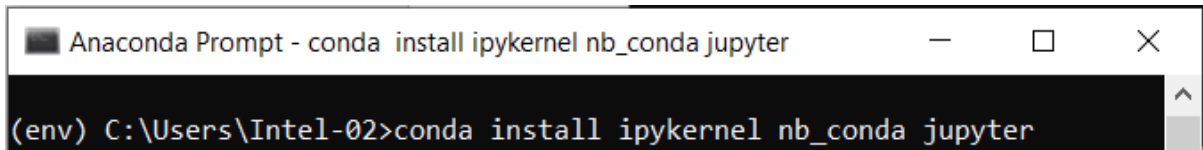
```
Anaconda Prompt
(base) C:\Users\Intel-02>conda activate env
(env) C:\Users\Intel-02>
```

This would activate the virtual environment and we can see the term written in brackets has changed from **(base)** to **(env)**. Now our virtual environment is ready to be used.

But, to open and work with Jupyter Notebooks in this environment, we need to install the packages which help in working with Jupyter Notebook. These packages get installed by default in the base environment when Anaconda gets installed.

To install Jupyter Notebook dependencies, we need to activate our virtual environment **env** and write:

```
conda install ipykernel nb_conda jupyter
```



```
Anaconda Prompt - conda install ipykernel nb_conda jupyter
(env) C:\Users\Intel-02>conda install ipykernel nb_conda jupyter
```

It will again ask if we wish to proceed with the installations, type **Y** to begin the installations. Once the installations are complete, we can start working with Jupyter notebooks in this environment.

Recap 2: Introduction to Python

In class 9, we were introduced to Python as the programming language which will be used for working around AI. Let us recall the basics of Python.

What?

Python is a programming language which was created by Guido Van Rossum in Centrum Wiskunde & Informatica. The language was publicly released in 1991 and it got its name from a BBC comedy series from 1970s – ‘Monty Python’s Flying Circus’. It can be used to follow both procedural approach and object-oriented approach of programming. Python has a lot of functionalities which makes it so popular to use.

Why?

Artificial intelligence is the trending technology of the future. We can see so many applications around us. If we as individuals would also like to develop an AI application, we will need to know a programming language. There are various programming languages like Lisp, Prolog, C++, Java and Python, which can be used for developing applications of AI. Out of these, Python gains a maximum popularity because of the following reasons:

Easy to learn, read and maintain

Python has few keywords, simple structure and a clearly defined syntax. Python allows anyone to learn the language quickly. A program written in Python is fairly easy-to-maintain.

A Broad Standard library

Python has a huge bunch of libraries with plenty of built-in functions to solve a variety of problems.

Interactive Mode

Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.

Portability and Compatibility

Python can run on a wide variety of operating systems and hardware platforms, and has the same interface on all platforms.

Extendable

We can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.

Databases and Scalable

Python provides interfaces to all major open source and commercial databases along with a better structure and support for much larger programs than shell scripting.

Applications of Python

There exist a wide variety of applications when it comes to Python. Some of the applications are:



Recap 3: Python Basics

In class 9, as Python was introduced, we also discussed about some basic Python syntaxes which can help us in writing codes in Python language. Let us brush up all the concepts once and see how we can use them in coding.

1. Printing Statements

We can use Python to display outputs for any code we write. To print any statement, we use **print()** function in Python.

2. Python Statements and Comments

Instructions written in the source code to execute are known as statements. These are the lines of code which we write for the computer to work upon. For example, if we wish to print the addition of two numbers, say 5 and 10, we would simply write:

```
print(5+10)
```

This is a Python statement as the computer would go through it and do the needful (which in this case would be to calculate 5+10 and print it on the output screen)

On the other hand, there exist some statements which do not get executed by the computer. These lines of code are skipped by the machine. They are known as comments. Comments are the statements which are incorporated in the code to give a better understanding of code statements to the user. To write a comment in Python, one can use # and then write anything after it. For example:

```
# This is a comment and will not be read by the machine.  
print(5+10) # This is a statement and the machine will print the  
summation.
```

Here, we can see that the first line is a comment as it starts with #. In the second line, we have an executable statement followed by a comment which is written to explain the code. In this way, we can add comments into our code so that anyone can understand the gist of it.

3. Keywords & Identifiers

In Python, there exist some words which are pre-defined and carry a specific meaning for the machine by default. These words are known as keywords. Keywords cannot be changed at any point in time and should not be used any other way except the default one, otherwise they create confusion and might result in ambiguous outputs. Some of the Keywords are mentioned below:

Keywords in Python

False	class	finally	is	return
None	continue	for	lambda	try
True	def	from	nonlocal	while
and	del	global	not	with
as	elif	if	or	yield
assert	else	import	pass	
break	except	in	raise	

Note that keywords are case-sensitive.

An identifier is any word which is variable. Identifiers can be declared by the user as per their convenience of use and can vary according to the way the user wants. These words are not defined and can be used in any way. **Keywords cannot be used as identifiers.** Some examples of keywords can be: **count, interest, x, ai_learning, Test**, etc. Identifiers are also case-sensitive hence an identifier named as **Test** would be different from an identifier named **test**.

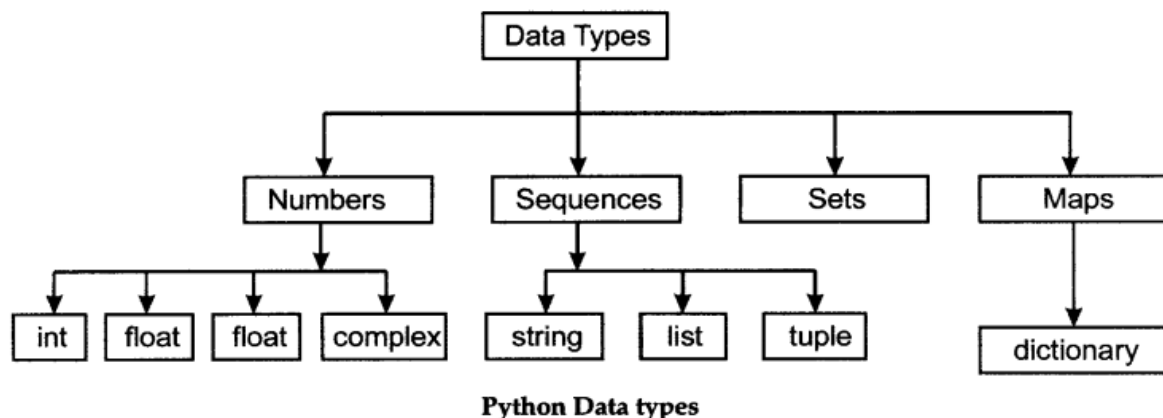
4. Variables & Datatypes

A variable is a named location used to store data in the memory. It is helpful to think of variables as a container that holds data which can be changed later throughout programming. Just like in Mathematics, in Python too we can use variables to store values in it. The difference here is, that in Python, the variables not only store numerical values, but can also contain different types of data.

For example:

```
X = 10 # X variable contains numerical data
Letters = 'XYZ' # Letters variable contains alphabetic data
number = 13.95 # number variable contains a decimal value
word = 'k' # word variable contains a character
```

All of these variables contain different types of data in them. The type of data is defined by the term datatype in Python. There can be various types of data which are used in Python programming. Hence, the machine identifies the type of variable according to the value which is stored inside it. Various datatypes in Python can be:



5. Python inputs

In Python, not only can we display the output to the user, but we can also collect data from the user and can pass it on to the Python script for further processing. To collect the data from the user at the time of execution, **input()** function is used. While using the input function, the datatype of the expected input is required to be mentioned so that the machine does not interpret the received data in an incorrect manner as the data taken as input from the user is considered to be a string (sequence of characters) by default.

For example:

```
Str = input(<String>) # Python expects the input to be of string
datatype
Number = int(input(<string>)) # Input string gets converted to an
integer value before assignment
Value = float(input(<String>)) # Input string gets converted to a
decimal value before assignment
```

6. Python Operators

Operators are special symbols which represent computation. They are applied on operand(s), which can be values or variables. Some operators can behave differently on different data types. Operators when applied on operands form an expression. Operators are categorized as Arithmetic, Relational, Logical and Assignment. Value and variables when used with operators are known as operands.

a. Arithmetic Operators

Operator	Meaning	Expression	Result
+	Addition	10 + 20	30
-	Subtraction	30 - 10	20
*	Multiplication	30 * 100	300
/	Division	30 / 10	20.0
//	Integer Division	25 // 10	2
%	Remainder	25 % 10	5
**	Raised to power	3 ** 2	9

b. Conditional Operators

Operator	Meaning	Expression	Result
>	Greater Than	20 > 10	True
		15 > 25	False
<	Less Than	20 < 45	True
		20 < 10	False
==	Equal To	5 == 5	True
		5 == 6	False
!=	Not Equal to	67 != 45	True
		35 != 35	False
>=	Greater than or Equal to	45 >= 45	True
		23 >= 34	False
<=	Less than or equal to	13 <= 24	True
		13 <= 12	False

c. Logical Operators

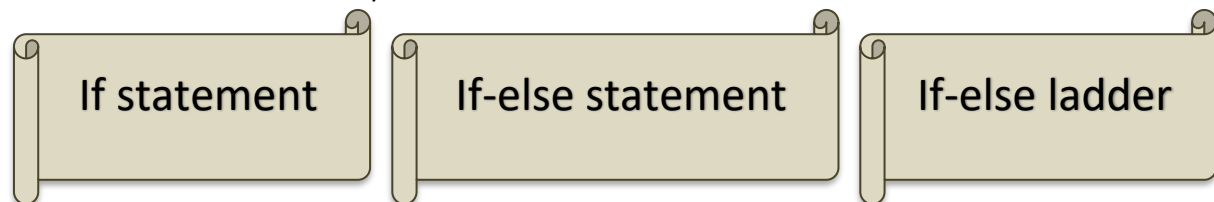
Operator	Meaning	Expression	Result
and	And operator	True and True	True
		True and False	False
or	Or operator	True or False	True
		False or False	False
not	Not Operator	not False	True
		not True	False

d. Assignment Operators

Operator	Expression	Equivalent to
=	X=5	X = 5
+=	X +=5	X = X + 5
-=	X -= 5	X = X - 5
*=	X *= 5	X = X * 5
/=	X /= 5	X = X / 5

7. Conditional Statements

While coding in Python, a lot of times we need to take decisions. For example, if a person needs to create a calculator with the help of a Python code, he/she needs to take in 2 numbers from the user and then ask the user about which function he/she wishes to operate. Now, according to the user's choice, the selection of function would change. In this case, we need the machine to understand what should happen when. This is where conditional statements help. Conditional statements help the machine in taking a decision according to the condition which gets fulfilled. There exist different types of conditional statements in Python. Some of them are:



According to the number of conditions and their dependency on each other, the relevant type of conditional statement is used.

8. Looping

A lot of times, it happens that a task needs to be executed multiple number of times. For example, we need to print hello 10 times on the output screen. One way of doing this is writing 10 print statements. But this is time and space consuming. The other way, which is more efficient, is to use loop statements. The loop statements help in iterating statements or a group of statements as many times as it is asked for. In this case, we will simply write a loop which would start counting from 1 to 10. At every count, it will print hello once on the screen and as soon as it reaches 10, the loop will stop executing. All this can be done by just one loop statement.

Various types of looping mechanisms are available in Python. Some of them are:



These were some of the basic concepts for writing a code in Python. We can explore these concepts further by going through the experiential Jupyter notebook for this chapter. In that notebook, we will get to explore Python basic concepts and we can also work around them to develop better understanding around it.

Python Packages

A package is nothing but a space where we can find codes or functions or modules of similar type. There are various packages readily available to use for free (perks of Python being an open-sourced language) for various purposes.

To use any package in Python, we need to install it. Installing Python packages is easy. Steps for package installation are:

1. Open Anaconda Navigator and activate your working environment.
2. Let us assume we wish to install the numpy package. To install this package, simply write:

```
conda install numpy
```

3. It will ask us to type Y if we wish to proceed with the installations. As soon as we type Y, the installations will start and our package will be installed in our selected environment.
4. We can also install multiple packages all at once by mentioning all of them in one line. For example, if we wish to install numpy, pandas and matplotlib package in our working environment. For this, simply write:

```
conda install numpy pandas matplotlib
```

This code will install these three packages altogether in our environment.

Now, once the packages are installed, we can start using them by importing them in the file where they are required. As soon as we open our Jupyter Notebook, include the package in the notebook by writing the import command. Importing a package can be done in various ways:

```
import numpy
```

Meaning: Import numpy in the file to use its functionalities in the file to which it has been imported.

```
import numpy as np
```

Meaning: Import numpy and refer to it as np wherever it is used.

```
from numpy import array
```

Meaning: import only one functionality (array) from the whole numpy package. While this gives faster processing, it limits the package's usability.

```
from numpy import array as arr
```

Meaning: Import only one functionality (array) from the whole numpy package and refer to it as arr wherever it is used. Some of the readily available packages are:

NumPy

- A package created to work around numerical arrays in python.
- Handy when it comes to working with large numerical databases and calculations around it.

OpenCV

- An image processing package which can explicitly work around images and can be used for image manipulation and processing like cropping, resizing, editing, etc.

Matplotlib

- A package which helps in plotting the analytical (numerical) data in graphical form.
- It helps the user in visualizing the data thereby helping them in understanding it better.

NLTK

- NLTK stands for Natural Language Tool Kit and it helps in tasks related to textual data.
- It is one of the most commonly used package for Natural Language Processing.

Pandas

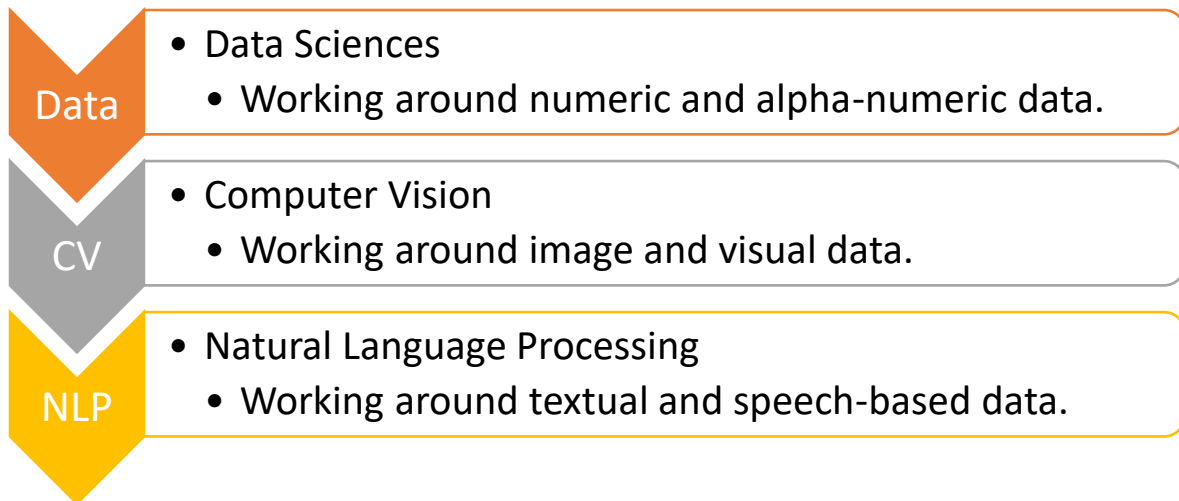
- A package which helps in handling 2-dimensional data tables in python.
- It is useful when we need to work with data from excel sheets and other databases.

To develop a better understanding around these packages, let us go through the Jupyter Notebook of package exploration and see how these packages can be used in Python.

Data Sciences

Introduction

As we have discussed earlier in class 9, Artificial Intelligence is a technology which completely depends on data. It is the data which is fed into the machine which makes it intelligent. And depending upon the type of data we have; AI can be classified into three broad domains:



Each domain has its own type of data which gets fed into the machine and hence has its own way of working around it. Talking about Data Sciences, it is a concept to unify statistics, data analysis, machine learning and their related methods in order to understand and analyse actual phenomena with data. It employs techniques and theories drawn from many fields within the context of Mathematics, Statistics, Computer Science, and Information Science.

Now before we get into the concepts of Data Sciences, let us experience this domain with the help of the following game:



* **Rock, Paper & Scissors:** <https://www.afiniti.com/corporate/rock-paper-scissors>

Go to this link and try to play the game of Rock, Paper Scissors against an AI model. The challenge here is to win 20 games against AI before AI wins them against you.

Did you manage to win?

What was the strategy that you applied to win this game against the AI machine?

Was it different playing Rock, Paper & Scissors with an AI machine as compared to a human?

What approach was the machine following while playing against you?

Applications of Data Sciences

Data Science is not a new field. Data Sciences majorly work around analysing the data and when it comes to AI, the analysis helps in making the machine intelligent enough to perform tasks by itself. There exist various applications of Data Science in today's world. Some of them are:



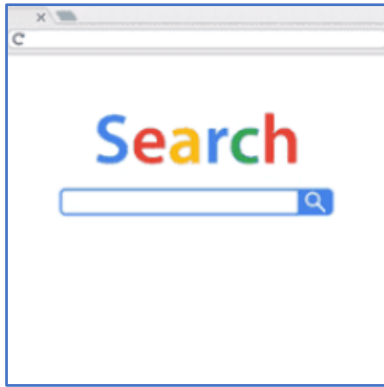
Fraud and Risk Detection*: The earliest applications of data science were in Finance. Companies were fed up of bad debts and losses every year. However, they had a lot of data which use to get collected during the initial paperwork while sanctioning loans. They decided to bring in data scientists in order to rescue them from losses.

Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures, and other essential variables to analyse the probabilities of risk and default. Moreover, it also helped them to push their banking products based on customer's purchasing power.

Genetics & Genomics*: Data Science applications also enable an advanced level of treatment personalization through research in genetics and genomics. The goal is to understand the impact of the DNA on our health and find individual biological connections between genetics, diseases, and drug response. Data science techniques allow integration of different kinds of data with genomic data in disease research, which provides a deeper understanding of genetic issues in reactions to particular drugs and diseases. As soon as we acquire reliable personal genome data, we will achieve a deeper understanding of the human DNA. The advanced genetic risk prediction will be a major step towards more individual care.

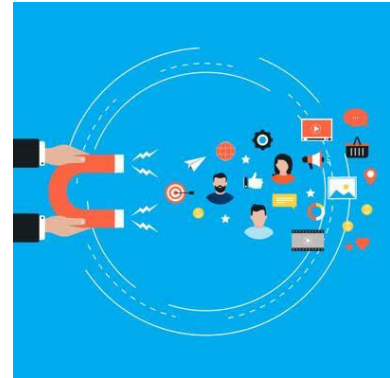


* Images shown here are the property of individual organisations and are used here for reference purpose only.



Internet Search*: When we talk about search engines, we think 'Google'. Right? 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 the fraction of a second. Considering the fact that Google processes more than 20 petabytes of data every day, had there been no data science, Google wouldn't have been the 'Google' we know today.

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 much higher CTR (Call-Through Rate) than traditional advertisements. They can be targeted based on a user's past behaviour.



Website Recommendations*: Aren't we all used to the suggestions about similar products on Amazon? They not only help us find relevant products from billions of products available with them but also add a lot to the user experience. A lot of companies have fervidly used this engine to promote their products in accordance with the user's interest and relevance of information. Internet giants like Amazon, Twitter, Google Play, Netflix, LinkedIn, IMDB and many more use this system to improve the user experience. The recommendations are made based on previous search results for a user.

Airline Route Planning*: The Airline Industry across the world is known to bear heavy losses. Except for a few airline service providers, companies are struggling to maintain their occupancy ratio and operating profits. With high rise in air-fuel prices and the need to offer heavy discounts to customers, the situation has got worse. It wasn't long before airline companies started using Data Science to identify the strategic areas of improvements. Now, while using Data Science, the airline companies can:



* Images shown here are the property of individual organisations and are used here for reference purpose only.

- Predict flight delay
- Decide which class of airplanes to buy
- Whether to directly land at the destination or take a halt in between (For example, A flight can have a direct route from New Delhi to New York. Alternatively, it can also choose to halt in any country.)
- Effectively drive customer loyalty programs

Getting Started

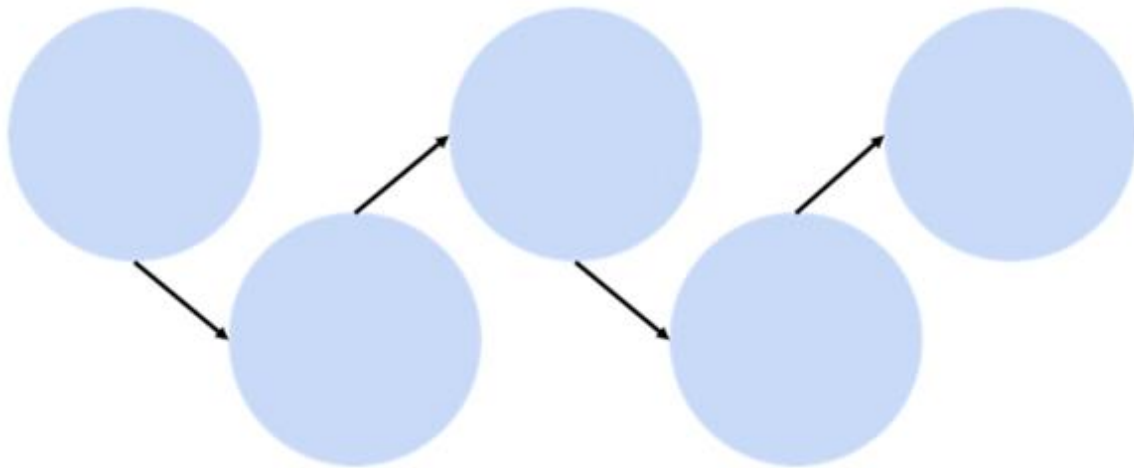
Data Sciences is a combination of Python and Mathematical concepts like Statistics, Data Analysis, probability, etc. Concepts of Data Science can be used in developing applications around AI as it gives a strong base for data analysis in Python.

Revisiting AI Project Cycle

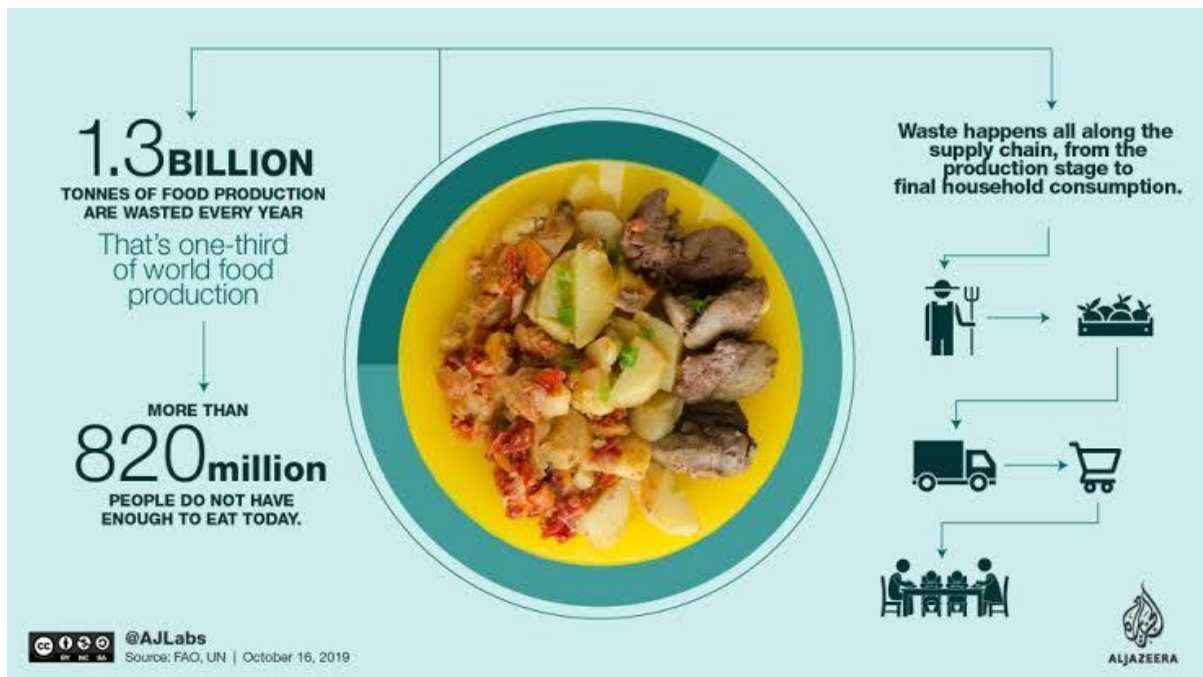
But, before we get deeper into data analysis, let us recall how Data Sciences can be leveraged to solve some of the pressing problems around us. For this, let us understand the AI project cycle framework around Data Sciences with the help of an example.

Do you remember the AI Project Cycle?

Fill in all the stages of the cycle here:



The Scenario*



Humans are social animals. We tend to organise and/or participate in various kinds of social gatherings all the time. We love eating out with friends and family because of which we can find restaurants almost everywhere and out of these, many of the restaurants arrange for buffets to offer a variety of food items to their customers. Be it small shops or big outlets, every restaurant prepares food in bulk as they expect a good crowd to come and enjoy their food. But in most cases, after the day ends, a lot of food is left which becomes unusable for the restaurant as they do not wish to serve stale food to their customers the next day. So, every day, they prepare food in large quantities keeping in mind the probable number of customers walking into their outlet. But if the expectations are not met, a good amount of food gets wasted which eventually becomes a loss for the restaurant as they either have to dump it or give it to hungry people for free. And if this daily loss is taken into account for a year, it becomes quite a big amount.

Problem Scoping

Now that we have understood the scenario well, let us take a deeper look into the problem to find out more about various factors around it. Let us fill up the 4Ws problem canvas to find out.

Who Canvas – Who is having the problem?

<i>Who are the stakeholders?</i>	<ul style="list-style-type: none"> ○ Restaurants offering buffets ○ Restaurant Chefs
<i>What do we know about them?</i>	<ul style="list-style-type: none"> ○ Restaurants cook food in bulk every day for their buffets to meet their customer needs. ○ They estimate the number of customers that would walk into their restaurant every day.

* Images shown here are the property of individual organisations and are used here for reference purpose only.

What Canvas – What is the nature of their problem?

<i>What is the problem?</i>	<ul style="list-style-type: none"> ○ Quite a large amount of food is leftover everyday unconsumed at the restaurant which is either thrown away or given for free to needy people. ○ Restaurants have to bear everyday losses for the unconsumed food.
<i>How do you know it is a problem?</i>	<ul style="list-style-type: none"> ○ Restaurant Surveys have shown that restaurants face this problem of food waste.

Where Canvas – Where does the problem arise?

<i>What is the context/situation in which the stakeholders experience this problem?</i>	<ul style="list-style-type: none"> ○ Restaurants which serve buffet food ○ At the end of the day, when no further food consumption is possible
---	--

Why? – Why do you think it is a problem worth solving?

<i>What would be of key value to the stakeholders?</i>	<ul style="list-style-type: none"> ○ If the restaurant has a proper estimate of the quantity of food to be prepared every day, the food waste can be reduced.
<i>How would it improve their situation?</i>	<ul style="list-style-type: none"> ○ Less or no food would be left unconsumed. ○ Losses due to unconsumed food would reduce considerably.

Now that we have noted down all the factors around our problem, let us fill up the problem statement template.

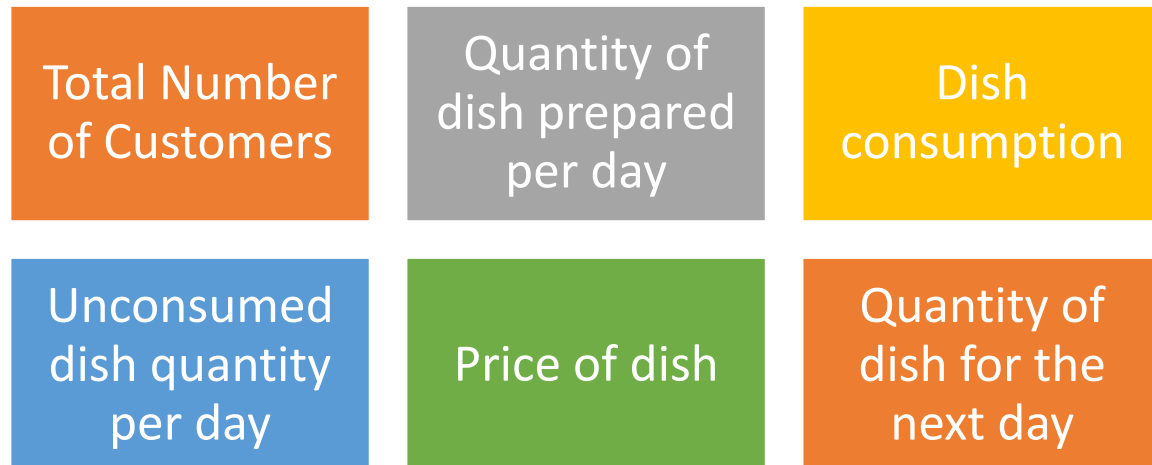
<i>Our</i>	Restaurant Owners	Who?
<i>Have a problem of</i>	Losses due to food wastage	What?
<i>While</i>	The food is left unconsumed due to improper estimation	Where?
<i>An ideal solution would</i>	Be to be able to predict the amount of food to be prepared for every day consumption	Why

The Problem statement template leads us towards the goal of our project which can now be stated as:

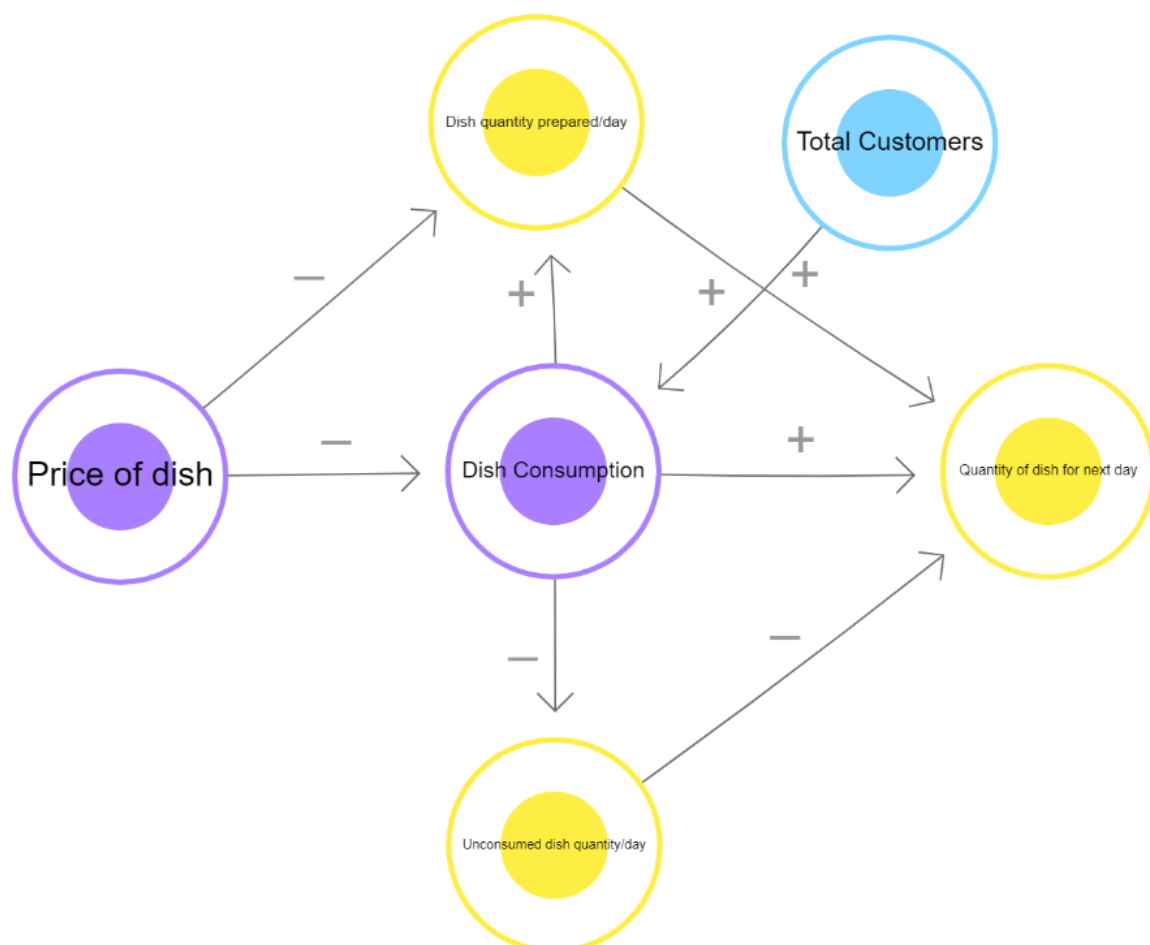
“To be able to predict the quantity of food dishes to be prepared for everyday consumption in restaurant buffets.”

Data Acquisition

After finalising the goal of our project, let us now move towards looking at various data features which affect the problem in some way or the other. Since any AI-based project requires data for testing and training, we need to understand what kind of data is to be collected to work towards the goal. In our scenario, various factors that would affect the quantity of food to be prepared for the next day consumption in buffets would be:



Now let us understand how these factors are related to our problem statement. For this, we can use the System Maps tool to figure out the relationship of elements with the project's goal. Here is the System map for our problem statement.



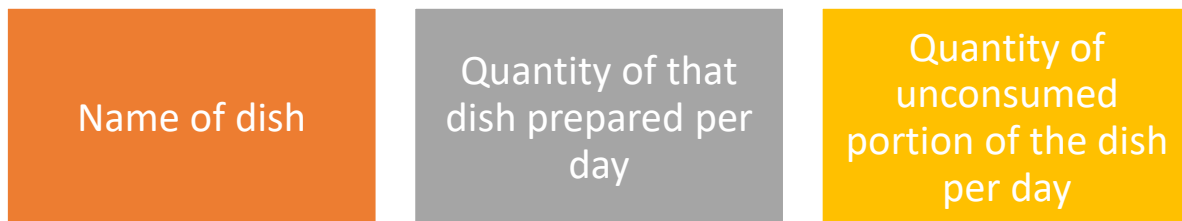
In this system map, you can see how the relationship of each element is defined with the goal of our project. Recall that the positive arrows determine a direct relationship of elements while the negative ones show an inverse relationship of elements.

After looking at the factors affecting our problem statement, now it's time to take a look at the data which is to be acquired for the goal. For this problem, a dataset covering all the elements mentioned above is made for each dish prepared by the restaurant over a period of 30 days. This data is collected offline in the form of a regular survey since this is a personalised dataset created just for one restaurant's needs.

Specifically, the data collected comes under the following categories: Name of the dish, Price of the dish, Quantity of dish produced per day, Quantity of dish left unconsumed per day, Total number of customers per day, Fixed customers per day, etc.

Data Exploration

After creating the database, we now need to look at the data collected and understand what is required out of it. In this case, since the goal of our project is to be able to predict the quantity of food to be prepared for the next day, we need to have the following data:



Thus, we extract the required information from the curated dataset and clean it up in such a way that there exist no errors or missing elements in it.

Modelling

Once the dataset is ready, we train our model on it. In this case, a regression model is chosen in which the dataset is fed as a dataframe and is trained accordingly. Regression is a Supervised Learning model which takes in continuous values of data over a period of time. Since in our case the data which we have is a continuous data of 30 days, we can use the regression model so that it predicts the next values to it in a similar manner. In this case, the dataset of 30 days is divided in a ratio of 2:1 for training and testing respectively. In this case, the model is first trained on the 20-day data and then gets evaluated for the rest of the 10 days.

Evaluation

Once the model has been trained on the training dataset of 20 days, it is now time to see if the model is working properly or not. Let us see how the model works and how it is tested.

Step 1: The trained model is fed data regarding the name of the dish and the quantity produced for the same.

Step 2: It is then fed data regarding the quantity of food left unconsumed for the same dish on previous occasions.

Step 3: The model then works upon the entries according to the training it got at the modelling stage.

Step 4: The Model predicts the quantity of food to be prepared for the next day.

Step 5: The prediction is compared to the testing dataset value. From the testing dataset, ideally, we can say that the quantity of food to be produced for next day's consumption should be the total quantity minus the unconsumed quantity.

Step 6: The model is tested for 10 testing datasets kept aside while training.

Step 7: Prediction values of testing dataset is compared to the actual values.

Step 8: If the prediction value is same or almost similar to the actual values, the model is said to be accurate. Otherwise, either the model selection is changed or the model is trained on more data for better accuracy.

Once the model is able to achieve optimum efficiency, it is ready to be deployed in the restaurant for real-time usage.

Data Collection

Data collection is nothing new which has come up in our lives. It has been in our society since ages. Even when people did not have fair knowledge of calculations, records were still maintained in some way or the other to keep an account of relevant things. Data collection is an exercise which does not require even a tiny bit of technological knowledge. But when it comes to analysing the data, it becomes a tedious process for humans as it is all about numbers and alpha-numerical data. That is where Data Science comes into the picture. It not only gives us a clearer idea around the dataset, but also adds value to it by providing deeper and clearer analyses around it. And as AI gets incorporated in the process, predictions and suggestions by the machine become possible on the same.

Now that we have gone through an example of a Data Science based project, we have a bit of clarity regarding the type of data that can be used to develop a Data Science related project. For the data domain-based projects, majorly the type of data used is in numerical or alpha-numerical format and such datasets are curated in the form of tables. Such databases are very commonly found in any institution for record maintenance and other purposes. Some examples of datasets which you must already be aware of are:

Banks

Databases of loans issued, account holder, locker owners, employee registrations, bank visitors, etc.

ATM Machines

Usage details per day, cash denominations transaction details, visitor details, etc.

Movie Theatres

Movie details, tickets sold offline, tickets sold online, refreshment purchases, etc.

Now look around you and find out what are the different types of databases which are maintained in the places mentioned below. Try surveying people who are responsible for the designated places to get a better idea.

Your classroom

Your school

Your city

As you can see, all the type of data which has been mentioned above is in the form of tables. Tables which contain numeric or alpha-numeric data. But this leads to a very critical dilemma: are these datasets accessible to all? Should these databases be accessible to all? What are the various sources of data from which we can gather such databases? Let's find out!

Sources of Data

There exist various sources of data from where we can collect any type of data required and the data collection process can be categorised in two ways: Offline and Online.

Offline Data Collection	Online Data Collection
Sensors	Open-sourced Government Portals
Surveys	Reliable Websites (Kaggle)
Interviews	World Organisations' open-sourced statistical websites
Observations	

While accessing data from any of the data sources, following points should be kept in mind:

1. Data which is available for public usage only should be taken up.
2. Personal datasets should only be used with the consent of the owner.
3. One should never breach someone's privacy to collect data.
4. Data should only be taken from reliable sources as the data collected from random sources can be wrong or unusable.
5. Reliable sources of data ensure the authenticity of data which helps in proper training of the AI model.

Types of Data

For Data Science, usually the data is collected in the form of tables. These tabular datasets can be stored in different formats. Some of the commonly used formats are:

1. CSV: CSV stands for comma separated values. It is a simple file format used to store tabular data. Each line of this file is a data record and each record consists of one or more fields which are separated by commas. Since the values of records are separated by a comma, hence they are known as CSV files.
2. Spreadsheet: A Spreadsheet is a piece of paper or a computer program which is used for accounting and recording data using rows and columns into which information can be entered. Microsoft excel is a program which helps in creating spreadsheets.
3. SQL: SQL is a programming language also known as Structured Query Language. It is a domain-specific language used in programming and is designed for managing data held in different kinds of DBMS (Database Management System) It is particularly useful in handling structured data.

A lot of other formats of databases also exist, you can explore them online!

Data Access

After collecting the data, to be able to use it for programming purposes, we should know how to access the same in a Python code. To make our lives easier, there exist various Python packages which help us in accessing structured data (in tabular form) inside the code. Let us take a look at some of these packages:

NumPy

NumPy, which stands for Numerical Python, is the fundamental package for Mathematical and logical operations on arrays in Python. It is a commonly used package when it comes to working around numbers. NumPy gives a wide range of arithmetic operations around numbers giving us an easier approach in working with them. NumPy also works with arrays, which is nothing but a homogenous collection of Data.

An array is nothing but a set of multiple values which are of same datatype. They can be numbers, characters, booleans, etc. but only one datatype can be accessed through an array. In NumPy, the arrays used are known as ND-arrays (N-Dimensional Arrays) as NumPy comes with a feature of creating n-dimensional arrays in Python.

An array can easily be compared to a list. Let us take a look at how they are different:

NumPy Arrays	Lists
<ol style="list-style-type: none">1. Homogenous collection of Data.2. Can contain only one type of data, hence not flexible with datatypes.3. Cannot be directly initialized. Can be operated with Numpy package only.4. Direct numerical operations can be done. For example, dividing the whole array by 3 divides every element by 3.5. Widely used for arithmetic operations.6. Arrays take less memory space.7. Functions like concatenation, appending, reshaping, etc are not trivially possible with arrays.8. Example: To create a numpy array 'A': <pre>import numpy A=numpy.array([1,2,3,4,5,6,7,8,9,0])</pre>	<ol style="list-style-type: none">1. Heterogenous collection of Data.2. Can contain multiple types of data, hence flexible with datatypes.3. Can be directly initialized as it is a part of Python syntax.4. Direct numerical operations are not possible. For example, dividing the whole list by 3 cannot divide every element by 3.5. Widely used for data management.6. Lists acquire more memory space.7. Functions like concatenation, appending, reshaping, etc are trivially possible with lists.8. Example: To create a list: <pre>A = [1,2,3,4,5,6,7,8,9,0]</pre>

Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

Pandas is well suited for many different kinds of data:

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet
- Ordered and unordered (not necessarily fixed-frequency) time series data.
- Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels
- Any other form of observational / statistical data sets. The data actually need not be labelled at all to be placed into a Pandas data structure

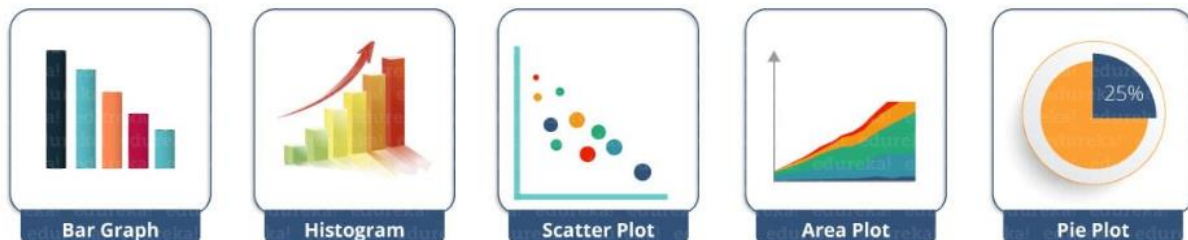
The two primary data structures of Pandas, Series (1-dimensional) and DataFrame (2-dimensional), handle the vast majority of typical use cases in finance, statistics, social science, and many areas of engineering. Pandas is built on top of NumPy and is intended to integrate well within a scientific computing environment with many other 3rd party libraries.

Here are just a few of the things that pandas does well:

- Easy handling of missing data (represented as NaN) in floating point as well as non-floating point data
- Size mutability: columns can be inserted and deleted from DataFrame and higher dimensional objects
- Automatic and explicit data alignment: objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let Series, DataFrame, etc. automatically align the data for you in computations
- Intelligent label-based slicing, fancy indexing, and subsetting of large data sets
- Intuitive merging and joining data sets
- Flexible reshaping and pivoting of data sets

Matplotlib*

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib comes with a wide variety of plots. Plots helps to understand trends, patterns, and to make correlations. They're typically instruments for reasoning about quantitative information. Some types of graphs that we can make with this package are listed below:



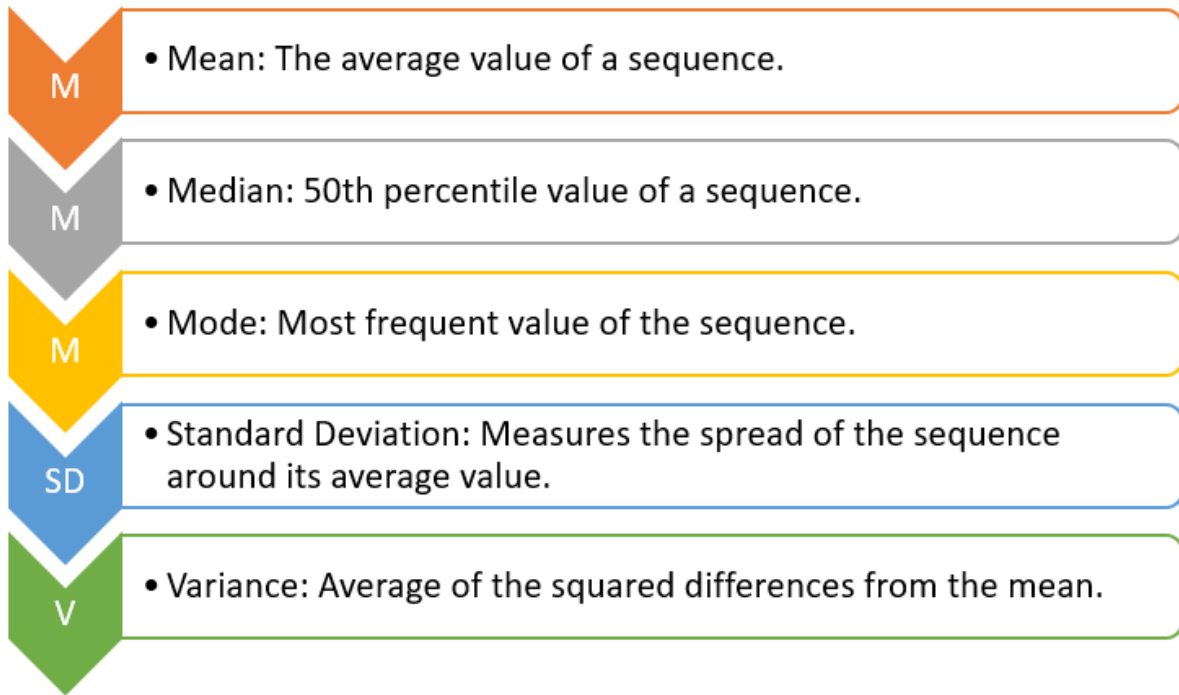
Not just plotting, but you can also modify your plots the way you wish. You can stylise them and make them more descriptive and communicable.

These packages help us in accessing the datasets we have and also in exploring them to develop a better understanding of them.

* Images shown here are the property of individual organisations and are used here for reference purpose only.

Basic Statistics with Python

We have already understood that Data Sciences works around analysing data and performing tasks around it. For analysing the numeric & alpha-numeric data used for this domain, mathematics comes to our rescue. Basic statistical methods used in mathematics come quite handy in Python too for analysing and working around such datasets. Statistical tools widely used in Python are:



Do you remember using these formulas in your class? Let us recall all of them here:

1. What is Mean? How is it calculated?

2. What is Median? How is it calculated?

3. What is Mode? How is it calculated?

4. What is Standard Deviation? How is it calculated?

5. What is Variance? How is it calculated?

Advantage of using Python packages is that we do not need to make our own formula or equation to find out the results. There exist a lot of pre-defined functions with packages like NumPy which reduces this trouble for us. All we need to do is write that function and pass on the data to it. It's that simple!

Let us take a look at various Python syntaxes that can help us with the statistical work in data analysis. Head to the Jupyter Notebook of Basic statistics with Python and start exploring! You may find the Jupyter notebook here: http://bit.ly/data_notebook

Data Visualisation

While collecting data, it is possible that the data might come with some errors. Let us first take a look at the types of issues we can face with data:

1. Erroneous Data: There are two ways in which the data can be erroneous:

- Incorrect values: The values in the dataset (at random places) are incorrect. For example, in the column of phone number, there is a decimal value or in the marks column, there is a name mentioned, etc. These are incorrect values that do not resemble the kind of data expected in that position.
- Invalid or Null values: At some places, the values get corrupted and hence they become invalid. Many times you will find NaN values in the dataset. These are null values which do not hold any meaning and are not processible. That is why, these values (as and when encountered) are removed from the database.

2. Missing Data: In some datasets, some cells remain empty. The values of these cells are missing and hence the cells remain empty. Missing data cannot be interpreted as an error as the values here are not erroneous or might not be missing because of any error.

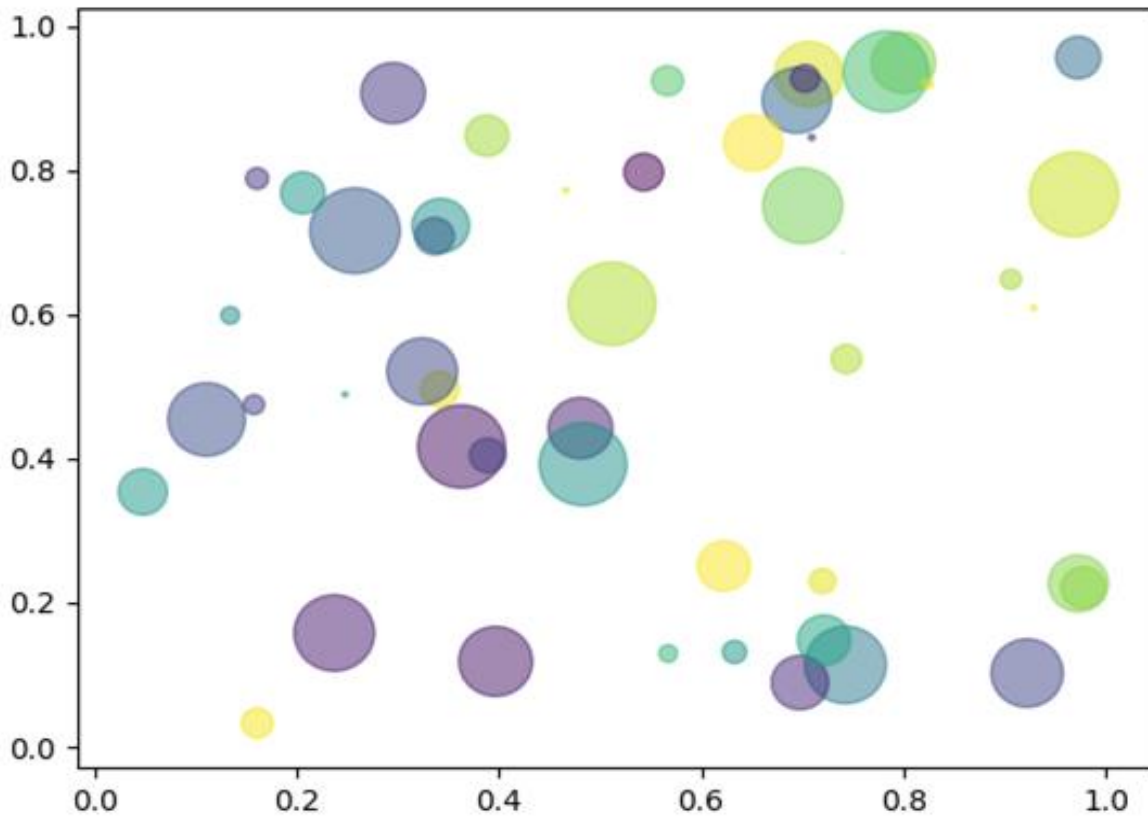
3. Outliers: Data which does not fall in the range of a certain element are referred to as outliers. To understand this better, let us take an example of marks of students in a class. Let us assume that a student was absent for exams and hence has got 0 marks in it. If his marks are taken into account, the whole class's average would go down. To prevent this, the average is taken for the range of marks from highest to lowest keeping this particular result separate. This makes sure that the average marks of the class are true according to the data.

Analysing the data collected can be difficult as it is all about tables and numbers. While machines work efficiently on numbers, humans need visual aid to understand and comprehend the information passed. Hence, data visualisation is used to interpret the data collected and identify patterns and trends out of it.

In Python, Matplotlib package helps in visualising the data and making some sense out of it. As we have already discussed before, with the help of this package, we can plot various kinds of graphs. Let us discuss some of them here:

Scatter Plot

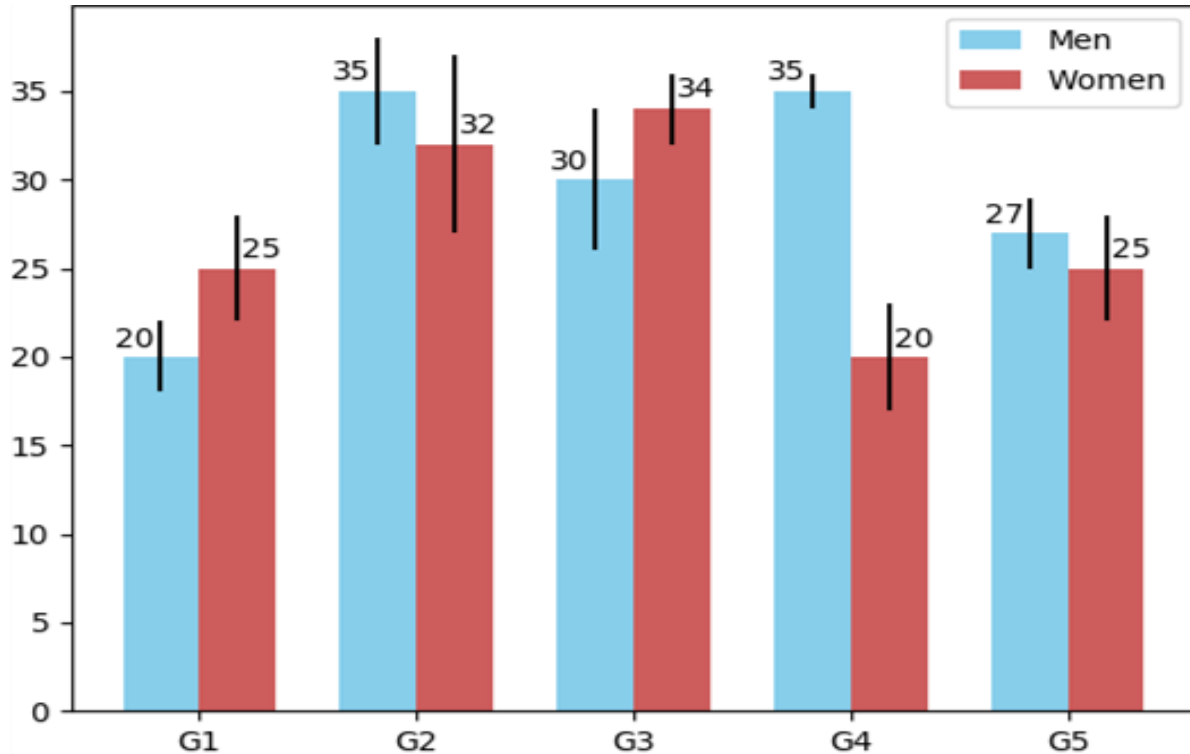
Scatter plots are used to plot discontinuous data; that is, the data which does not have any continuity in flow is termed as discontinuous. There exist gaps in data which introduce discontinuity. A 2D scatter plot can display information maximum upto 4 parameters.



In this scatter plot, 2 axes (X and Y) are two different parameters. The colour of circles and the size both represent 2 different parameters. Thus, just through one coordinate on the graph, one can visualise 4 different parameters all at once.

Bar Chart

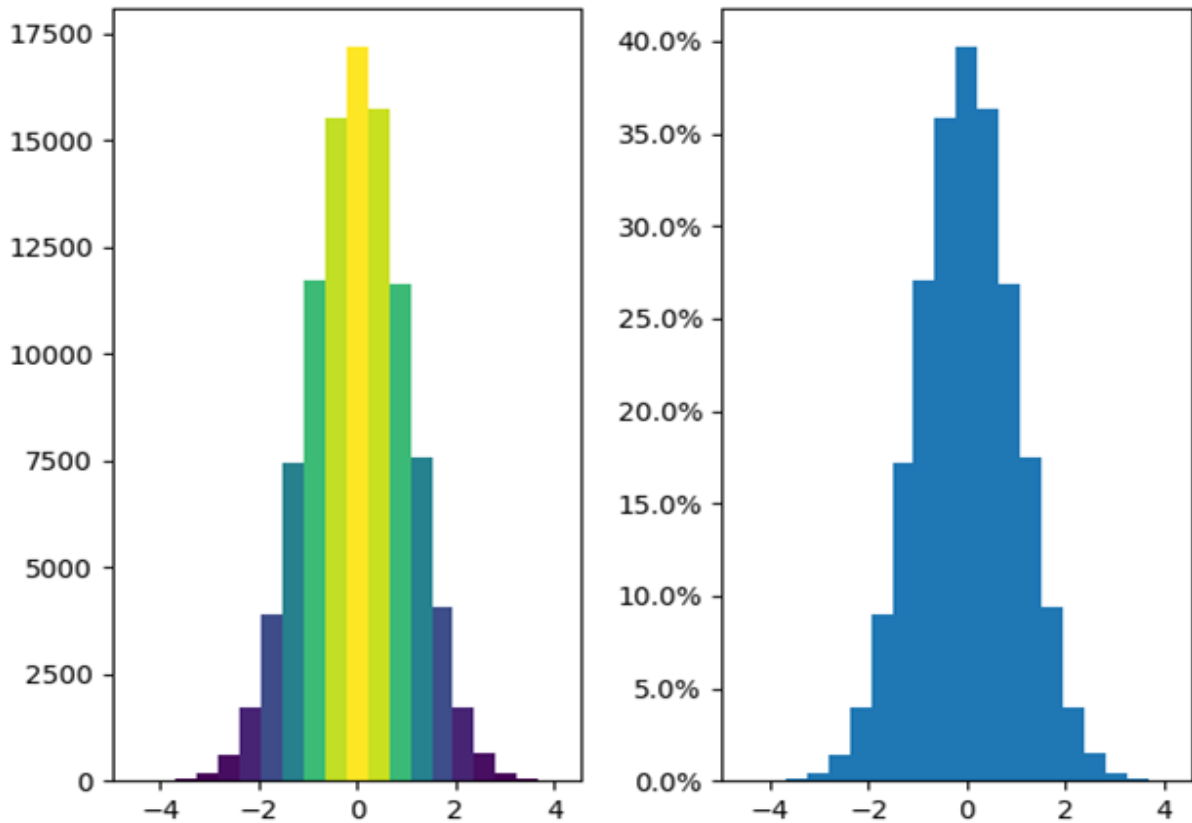
It is one of the most commonly used graphical methods. From students to scientists, everyone uses bar charts in some way or the other. It is a very easy to draw yet informative graphical representation. Various versions of bar chart exist like single bar chart, double bar chart, etc.



This is an example of a double bar chart. The 2 axes depict two different parameters while bars of different colours work with different entities (in this case it is women and men). Bar chart also works on discontinuous data and is made at uniform intervals.

Histogram

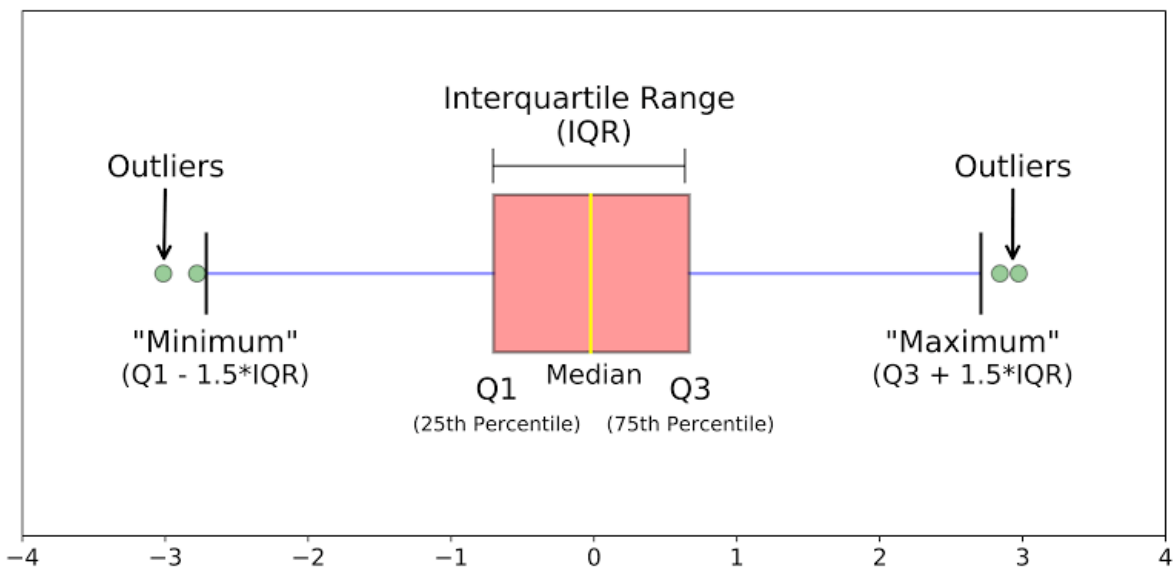
Histograms are the accurate representation of a continuous data. When it comes to plotting the variation in just one entity of a period of time, histograms come into the picture. It represents the frequency of the variable at different points of time with the help of the bins.



In the given example, the histogram is showing the variation in frequency of the entity plotted with the help of XY plane. Here, at the left, the frequency of the element has been plotted and it is a frequency map for the same. The colours show the transition from low to high and vice versa. Whereas on the right, a continuous dataset has been plotted which might not be talking about the frequency of occurrence of the element.

Box Plots

When the data is split according to its percentile throughout the range, box plots come in handy. Box plots also known as box and whiskers plot conveniently display the distribution of data throughout the range with the help of 4 quartiles.



Here as we can see, the plot contains a box and two lines at its left and right are termed as whiskers. The plot has 5 different parts to it:

Quartile 1: From 0 percentile to 25th percentile – Here data lying between 0 and 25th percentile is plotted. Now, if the data is close to each other, lets say 0 to 25th percentile data has been covered in just 20-30 marks range, then the whisker would be smaller as the range is smaller. But if the range is large that is 0-30 marks range, then the whisker would also get elongated as the range is longer.

Quartile 2: From 25th Percentile to 50th percentile – 50th percentile is termed as the mean of the whole distribution and since the data falling in the range of 25th percentile to 75th percentile has minimum deviation from the mean, it is plotted inside the box.

Quartile 3: From 50th percentile to 75th percentile – This range is again plotted in the box as its deviation from the mean is less. Quartile 2 & 3 (from 25th percentile to 75th percentile) together constitute the Inter Quartile Range (IQR). Also, depending upon the range of distribution, just like whiskers, the length of box also varies if the data is less spread or more.

Quartile 4: From 75th percentile to 100th percentile – It is the whiskers plot for top 25 percentile data.

Outliers: The advantage of box plots is that they clearly show the outliers in a data distribution. Points which do not lie in the range are plotted outside the graph as dots or circles and are termed as outliers as they do not belong to the range of data. Since being out of range is not an error, that is why they are still plotted on the graph for visualisation.

Let us now move ahead and experience data visualisation using Jupyter notebook. Matplotlib library will help us in plotting all sorts of graphs while Numpy and Pandas will help us in analysing the data.

Data Sciences: Classification Model

In this section, we would be looking at one of the classification models used in Data Sciences. But before we look into the technicalities of the code, let us play a game.

Personality Prediction

Step 1: Here is a map. Take a good look at it. In this map you can see the arrows determine a quality. The qualities mentioned are:

1. Positive X-axis – People focussed: You focus more on people and try to deliver the best experience to them.
2. Negative X-axis – Task focussed: You focus more on the task which is to be accomplished and try to do your best to achieve that.
3. Positive Y-axis – Passive: You focus more on listening to people and understanding everything that they say without interruption.
4. Negative Y-axis – Active: You actively participate in the discussions and make sure that you make your point in-front of the crowd.

Think for a minute and understand which of these qualities you have in you. Now, take a chit and write your name on it. Place this chit at a point in this map which best describes you. It can be placed anywhere on the graph. Be honest about yourself and put it on the graph.

Step 2: Now that you have all put up your chits on the graph, it's time to take a quick quiz. Go to this link and finish the quiz on it individually: <https://tinyurl.com/discanimal>

On this link, you will find a personality prediction quiz. Take this quiz individually and try to answer all the questions honestly. Do not take anyone's help in it and do not discuss about it with anyone. Once

the quiz is finished, remember the animal which has been predicted for you. Write it somewhere and do not show it to anyone. Keep it as your little secret.

Once everyone has gone through the quiz, go back to the board remove your chit, and draw the symbol which corresponds to your animal in place of your chit. Here are the symbols:

Lion	Otter	Golden Retriever	Beaver
♥	●	😊	★

Place these symbols at the locations where you had put up your names. Ask 4 students not to do so and tell them to keep their animals a secret. Let their name chits be on the graph so that we can predict their animals with the help of this map.

Now, we will try to use the nearest neighbour algorithm here and try to predict what can be the possible animal(s) for these 4 unknowns. Now look that these 4 chits one by one. Which animal is occurring the most in their vicinity? Do you think that if the m lion symbol is occurring the most near their chit, then there is a good probability that their animal would also be a lion? Now let us try to guess the animal for all 4 of them according to their nearest neighbours respectively. After guessing the animals, ask these 4 students if the guess is right or not.

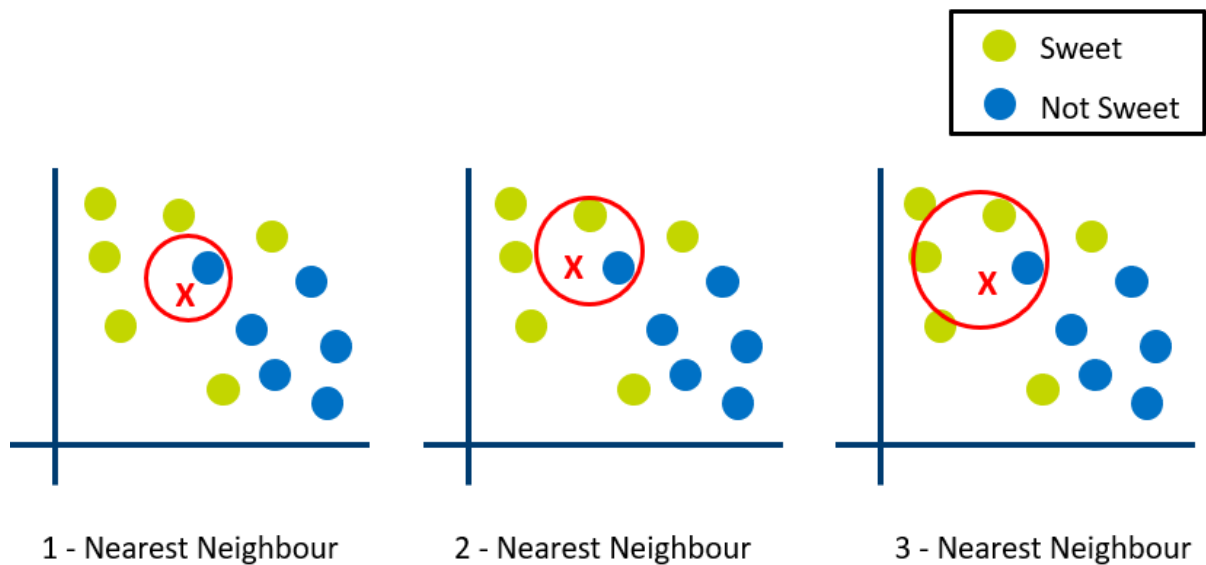
K-Nearest Neighbour: Explained

The k-nearest neighbours (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other as the saying goes "*Birds of a feather flock together*". Some features of KNN are:

- The KNN prediction model relies on the surrounding points or neighbours to determine its class or group
- Utilises the properties of the majority of the nearest points to decide how to classify unknown points
- Based on the concept that similar data points should be close to each other

The personality prediction activity was a brief introduction to KNN. As you recall, in that activity, we tried to predict the animal for 4 students according to the animals which were the nearest to their points. This is how in a lay-man's language KNN works. Here, K is a variable which tells us about the number of neighbours which are taken into account during prediction. It can be any integer value starting from 1.

Let us look at another example to demystify this algorithm. Let us assume that we need to predict the sweetness of a fruit according to the data which we have for the same type of fruit. So here we have three maps to predict the same:



Here, X is the value which is to be predicted. The green dots depict sweet values and the blue ones denote not sweet.

Let us try it out by ourselves first. Look at the map closely and decide whether X should be sweet or not sweet?

Now, let us look at each graph one by one:

- 1

Here, we can see that K is taken as 1 which means that we are taking only 1 nearest neighbour into consideration. The nearest value to X is a blue one hence 1-nearest neighbour algorithm predicts that the fruit is not sweet.
- 2

In the 2nd graph, the value of K is 2. Taking 2 nearest nodes to X into consideration, we see that one is sweet while the other one is not sweet. This makes it difficult for the machine to make any predictions based on the nearest neighbour and hence the machine is not able to give any prediction.
- 3

In the 3rd graph, the value of K becomes 3. Here, 3 nearest nodes to X are chosen out of which 2 are green and 1 is blue. On the basis of this, the model is able to predict that the fruit is sweet.

On the basis of this example, let us understand KNN better:

KNN tries to predict an unknown value on the basis of the known values. The model simply calculates the distance between all the known points with the unknown point (by distance we mean to say the different between two values) and takes up K number of points whose distance is minimum. And according to it, the predictions are made.

Let us understand the significance of the number of neighbours:

1. As we decrease the value of K to 1, our predictions become less stable. Just think for a minute, imagine K=1 and we have X surrounded by several greens and one blue, but the blue is the single nearest neighbour. Reasonably, we would think X is most likely green, but because K=1, KNN incorrectly predicts that it is blue.

2. Inversely, as we increase the value of K , our predictions become more stable due to majority voting / averaging, and thus, more likely to make more accurate predictions (up to a certain point). Eventually, we begin to witness an increasing number of errors. It is at this point we know we have pushed the value of K too far.
3. In cases where we are taking a majority vote (e.g. picking the mode in a classification problem) among labels, we usually make K an odd number to have a tiebreaker.

Computer Vision

Introduction

In the previous chapter, you studied the concepts of Artificial Intelligence for Data Sciences. It is a concept to unify statistics, data analysis, machine learning and their related methods in order to understand and analyse actual phenomena with data.

As we all know, artificial intelligence is a technique that enables computers to mimic human intelligence. As humans we can see things, analyse it and then do the required action on the basis of what we see.

But can machines do the same? Can machines have the eyes that humans have? If you answered Yes, then you are absolutely right. The Computer Vision domain of Artificial Intelligence, enables machines to see through images or visual data, process and analyse them on the basis of algorithms and methods in order to analyse actual phenomena with images.

Now before we get into the concepts of Computer Vision, let us experience this domain with the help of the following game:



* Emoji Scavenger Hunt : <https://emojiscavengerhunt.withgoogle.com/>

Go to the link and try to play the game of Emoji Scavenger Hunt. The challenge here is to find 8 items within the time limit to pass.

Did you manage to win?

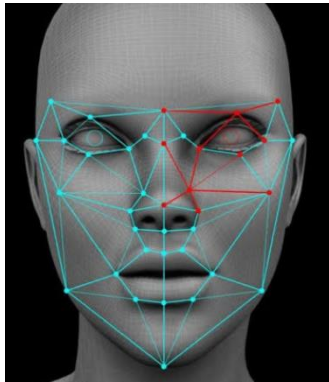
What was the strategy that you applied to win this game?

Was the computer able to identify all the items you brought in front of it?

Did the lighting of the room affect the identifying of items by the machine?

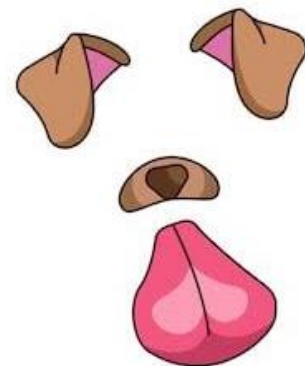
Applications of Computer Vision

The concept of computer vision was first introduced in the 1970s. All these new applications of computer vision excited everyone. Having said that, the computer vision technology advanced enough to make these applications available to everyone at ease today. However, in recent years the world witnessed a significant leap in technology that has put computer vision on the priority list of many industries. Let us look at some of them:



Facial Recognition*: With the advent of smart cities and smart homes, Computer Vision plays a vital role in making the home smarter. Security being the most important application involves use of Computer Vision for facial recognition. It can be either guest recognition or log maintenance of the visitors.

It also finds its application in schools for an attendance system based on facial recognition of students.



Face Filters*: The modern-day apps like Instagram and snapchat have a lot of features based on the usage of computer vision. The application of face filters is one among them. Through the camera the machine or the algorithm is able to identify the facial dynamics of the person and applies the facial filter selected.

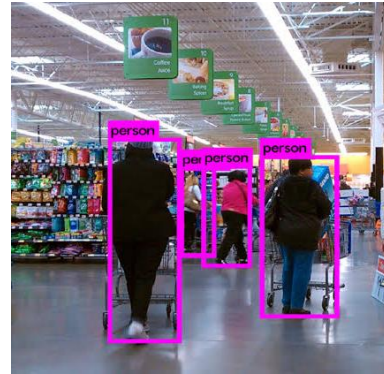


Google's Search by Image*: The maximum amount of searching for data on Google's search engine comes from textual data, but at the same time it has an interesting feature of getting search results through an image. This uses Computer Vision as it compares different features of the input image to the database of images and give us the search result while at the same time analysing various features of the image.

* Images shown here are the property of individual organisations and are used here for reference purpose only.

Computer Vision in Retail*: The retail field has been one of the fastest growing fields and at the same time is using Computer Vision for making the user experience more fruitful. Retailers can use Computer Vision techniques to track customers' movements through stores, analyse navigational routes and detect walking patterns.

Inventory Management is another such application. Through security camera image analysis, a Computer Vision algorithm can generate a very accurate estimate of the items available in the store. Also, it can analyse the use of shelf space to identify suboptimal configurations and suggest better item placement.



Self-Driving Cars: Computer Vision is the fundamental technology behind developing autonomous vehicles. Most leading car manufacturers in the world are reaping the benefits of investing in artificial intelligence for developing on-road versions of hands-free technology.

This involves the process of identifying the objects, getting navigational routes and also at the same time environment monitoring.

Medical Imaging*: For the last decades, computer-supported medical imaging application has been a trustworthy help for physicians. It doesn't only create and analyse images, but also becomes an assistant and helps doctors with their interpretation. The application is used to read and convert 2D scan images into interactive 3D models that enable medical professionals to gain a detailed understanding of a patient's health condition.



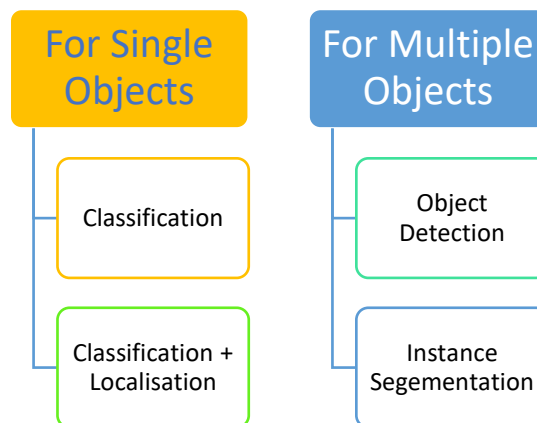
Google Translate App*: All you need to do to read signs in a foreign language is to point your phone's camera at the words and let the Google Translate app tell you what it means in your preferred language almost instantly. By using optical character recognition to see the image and augmented reality to overlay an accurate translation, this is a convenient tool that uses Computer Vision.

Computer Vision: Getting Started

Computer Vision is a domain of Artificial Intelligence, that deals with the images. It involves the concepts of image processing and machine learning models to build a Computer Vision based application.

Computer Vision Tasks

The various applications of Computer Vision are based on a certain number of tasks which are performed to get certain information from the input image which can be directly used for prediction or forms the base for further analysis. The tasks used in a computer vision application are :



Classification

Image Classification problem is the task of **assigning an input image one label from a fixed set of categories**. This is one of the core problems in CV that, despite its simplicity, has a large variety of practical applications.

Classification + Localisation

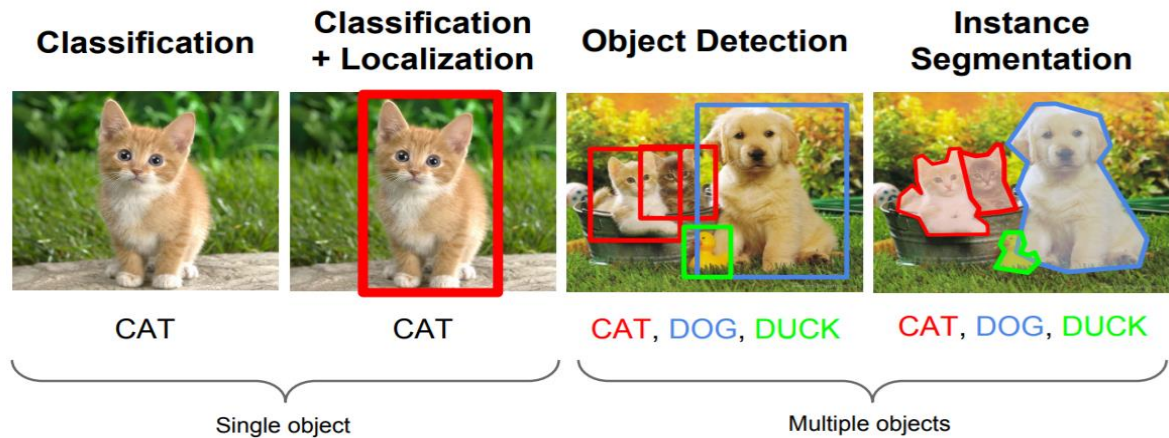
This is the task which involves both processes of **identifying what object is present** in the image and at the same time **identifying at what location** that object is present in that image. It is used only for single objects.

Object Detection

Object detection is the process of **finding instances of real-world objects such as faces, bicycles, and buildings in images or videos**. Object detection algorithms typically use extracted features and learning algorithms to recognize instances of an object category. It is commonly used in applications such as image retrieval and automated vehicle parking systems.

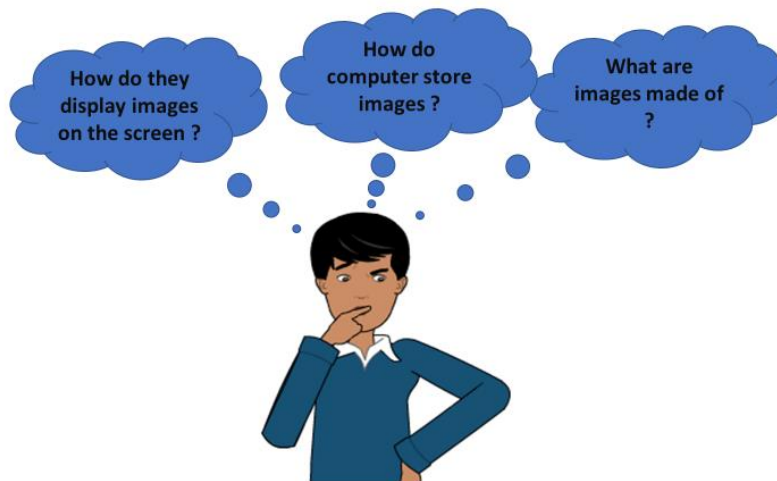
Instance Segmentation

Instance Segmentation is the process of detecting instances of the objects, giving them a category and then giving each pixel a label on the basis of that. A segmentation algorithm takes an image as input and outputs a collection of regions (or segments).



Basics of Images

We all see a lot of images around us and use them daily either through our mobile phones or computer system. But do we ask some basic questions to ourselves while we use them on such a regular basis.

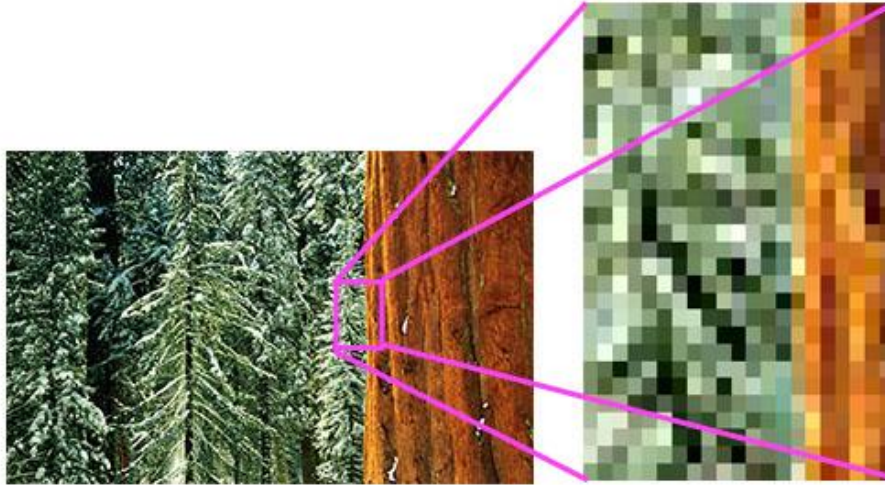


Don't know the answer yet? Don't worry, in this section we will study about the basics of an image:

Basics of Pixels

The word "pixel" means a picture element. Every photograph, in digital form, is made up of pixels. They are the smallest unit of information that make up a picture. Usually round or square, they are typically arranged in a 2-dimensional grid.

In the image below, one portion has been magnified many times over so that you can see its individual composition in pixels. As you can see, the pixels approximate the actual image. The more pixels you have, the more closely the image resembles the original.



Resolution

The number of pixels in an image is sometimes called the *resolution*. When the term is used to describe pixel count, one convention is to express resolution as the width by the height, for example a monitor resolution of 1280×1024. This means there are 1280 pixels from one side to the other, and 1024 from top to bottom.

Another convention is to express the number of pixels as a single number, like a 5 mega pixel camera (a megapixel is a million pixels). This means the pixels along the width multiplied by the pixels along the height of the image taken by the camera equals 5 million pixels. In the case of our 1280×1024 monitors, it could also be expressed as $1280 \times 1024 = 1,310,720$, or 1.31 megapixels.

Pixel value

Each of the pixels that represents an image stored inside a computer has a *pixel value* which describes how bright that pixel is, and/or what colour it should be. The most common *pixel format* is the *byte image*, where this number is stored as an 8-bit integer giving a range of possible values from 0 to 255. Typically, zero is to be taken as no colour or black and 255 is taken to be full colour or white.

Why do we have a value of 255 ? In the computer systems, computer data is in the form of ones and zeros, which we call the binary system. Each bit in a computer system can have either a zero or a one.

Since each pixel uses 1 byte of an image, which is equivalent to 8 bits of data. Since each bit can have two possible values which tells us that the 8 bit can have 255 possibilities of values which starts from 0 and ends at 255.

Number of bits	Different patterns	No. of patterns	No. of patterns
1	0 1	2^1	2
2	00 01 10 11	2^2	4
3	000 001 010 100 011 101 110 111	2^3	8

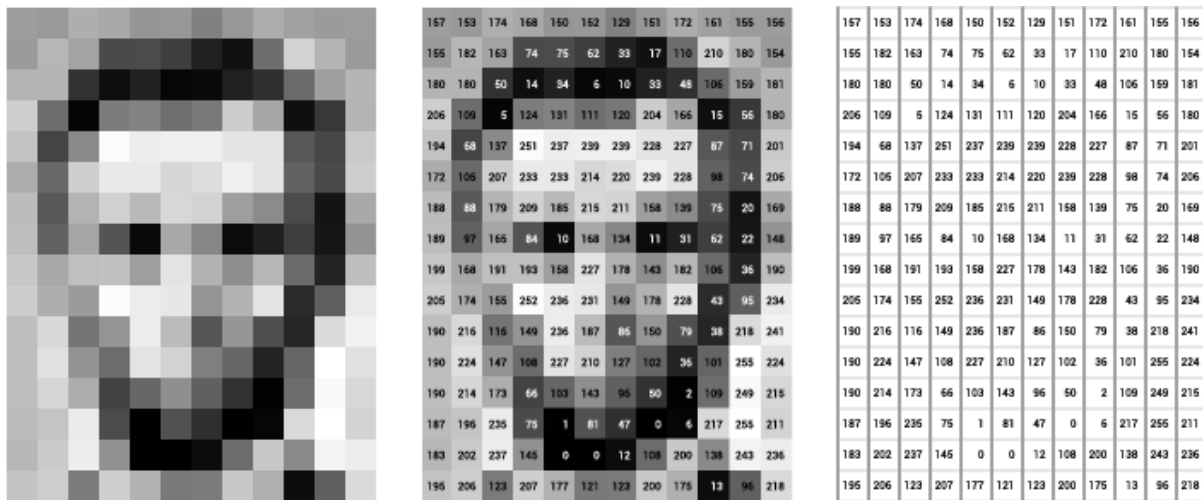
$$2^8 = 256$$

Grayscale Images

Grayscale images are images which have a range of shades of gray without apparent colour. The darkest possible shade is black, which is the total absence of colour or zero value of pixel. The lightest possible shade is white, which is the total presence of colour or 255 value of a pixel . Intermediate shades of gray are represented by equal [brightness](#) levels of the three primary colours.

A grayscale has each pixel of size 1 byte having a single plane of 2d array of pixels. The size of a grayscale image is defined as the Height x Width of that image.

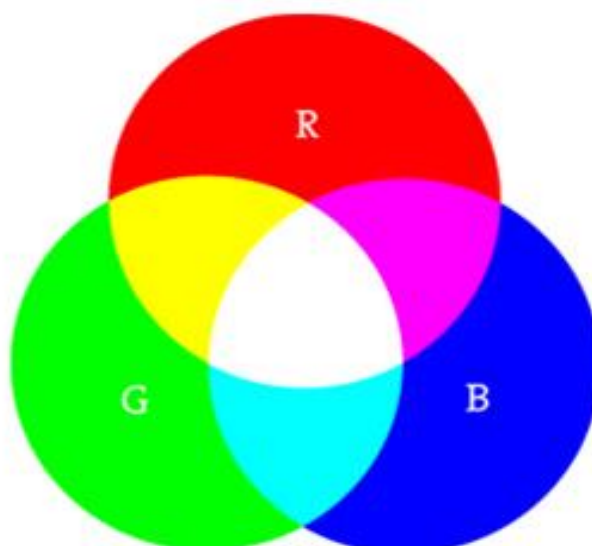
Let us look at an image to understand about grayscale images.



Here is an example of a grayscale image. as you check, the value of pixels are within the range of 0-255. The computers store the images we see in the form of these numbers.

RGB Images

All the images that we see around are coloured images. These images are made up of three primary colours Red, Green and Blue. All the colours that are present can be made by combining different intensities of red, green and blue.



* Images shown here are the property of individual organisations and are used here for reference purpose only.

Let us experience!

Go to this online link https://www.w3schools.com/colors/colors_rgb.asp. On the basis of this online tool, try and answer all the below mentioned questions.

- 1) What is the output colour when you put $R=G=B=255$?

- 2) What is the output colour when you put $R=G=B=0$?

- 3) How does the colour vary when you put either of the three as 0 and then keep on varying the other two?

- 4) How does the output colour change when all the three colours are varied in same proportion ?

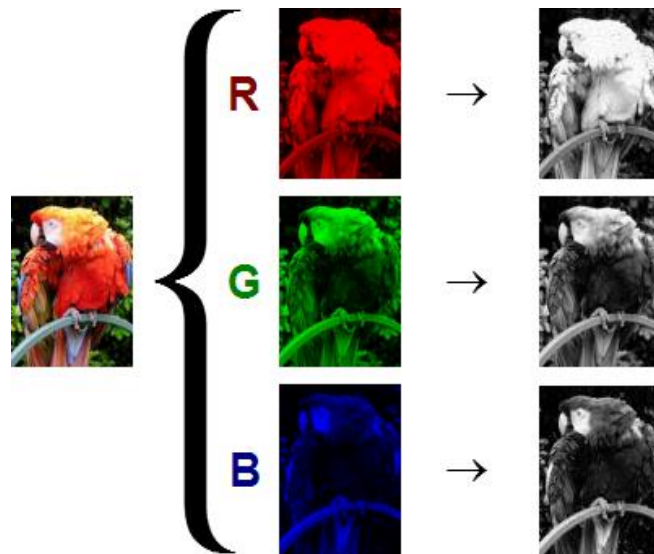
- 5) What is the RGB value of your favourite colour from the colour palette?

Were you able to answer all the questions? If yes, then you would have understood how every colour we see around is made.

Now the question arises, how do computers store RGB images? Every RGB image is stored in the form of three different channels called the R channel, G channel and the B channel.

Each plane separately has a number of pixels with each pixel value varying from 0 to 255. All the three planes when combined together form a colour image. This means that in a RGB image, each pixel has a set of three different values which together give colour to that particular pixel.

For Example,



As you can see, each colour image is stored in the form of three different channels, each having different intensity. All three channels combine together to form a colour we see.

In the above given image, if we split the image into three different channels, namely Red (R), Green (G) and Blue (B), the individual layers will have the following intensity of colours of the individual pixels. These individual layers when stored in the memory looks like the image on the extreme right. The images look in the grayscale image because each pixel has a value intensity of 0 to 255 and as studied earlier, 0 is considered as black or no presence of colour and 255 means white or full presence of colour. These three individual RGB values when combined together form the colour of each pixel.

Therefore, each pixel in the RGB image has three values to form the complete colour.

Task :

Go to the following link www.piskelapp.com and create your own pixel art. Try and make a GIF using the online app for your own pixel art.

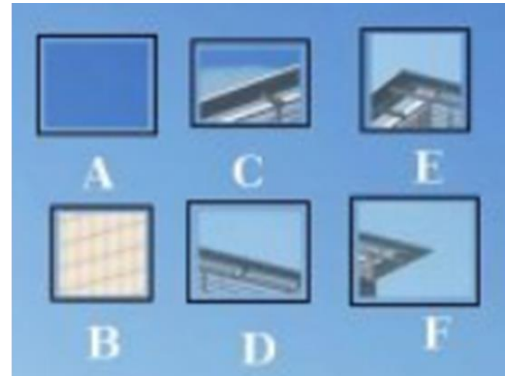
Image Features

In computer vision and image processing, a **feature** is a piece of information which is relevant for solving the computational task related to a certain application. Features may be specific structures in the image such as points, edges or objects.

For example:

Imagine that your security camera is capturing an image. At the top of the image we are given six small patches of images. Our task is to find the exact location of those image patches in the image.

Take a pencil and mark the exact location of those patches in the image.



Were you able to find the exact location of all the patches?

Which one was the most difficult to find?

Which one was the easiest to find?

Let's Reflect:

Let us take individual patches into account at once and then check the exact location of those patches.

For Patch A and B: The patch A and B are flat surfaces in the image and are spread over a lot of area. They can be present at any location in a given area in the image.

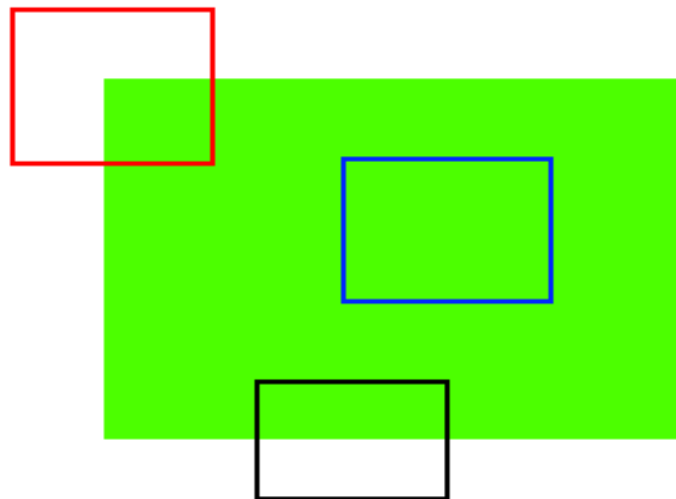
For Patch C and D: The patches C and D are simpler as compared to A and B. They are edges of a building and we can find an approximate location of these patches but finding the exact location is still difficult. This is because the pattern is the same everywhere along the edge.

For Patch E and F: The patches E and F are the easiest to find in the image. The reason being that E and F are some corners of the building. This is because at the corners, wherever we move this patch it will look different.

Conclusion

In image processing, we can get a lot of features from the image. It can be either a blob, an edge or a corner. These features help us to perform various tasks and then get the analysis done on the basis of the application. Now the question that arises is which of the following are good features to be used? As you saw in the previous activity, the features having the corners are easy to find as they can be found only at a particular location in the image, whereas the edges which are spread over a line or an edge look the same all along. This tells us that the corners are always good features to extract from an image followed by the edges.

Let's look at another example to understand this. Consider the images given below and apply the concept of good features for the following.



In the above image how would we determine the exact location of each patch?

The blue patch is a flat area and difficult to find and track. Wherever you move the blue patch it looks the same. The black patch has an edge. Moved along the edge (parallel to edge), it looks the same. The red patch is a corner. Wherever you move the patch, it looks different, therefore it is unique. Hence, corners are considered to be good features in an image.

Introduction to OpenCV

Now that we have learnt about image features and its importance in image processing, we will learn about a tool we can use to extract these features from our image for further processing.

OpenCV or Open Source Computer Vision Library is that tool which helps a computer extract these features from the images. It is used for all kinds of images and video processing and analysis. It is capable of processing images and videos to identify objects, faces, or even handwriting.



In this chapter we will use OpenCV for basic image processing operations on images such as resizing, cropping and many more.

To install OpenCV library, open anaconda prompt and then write the following command:

```
pip install opencv-python
```

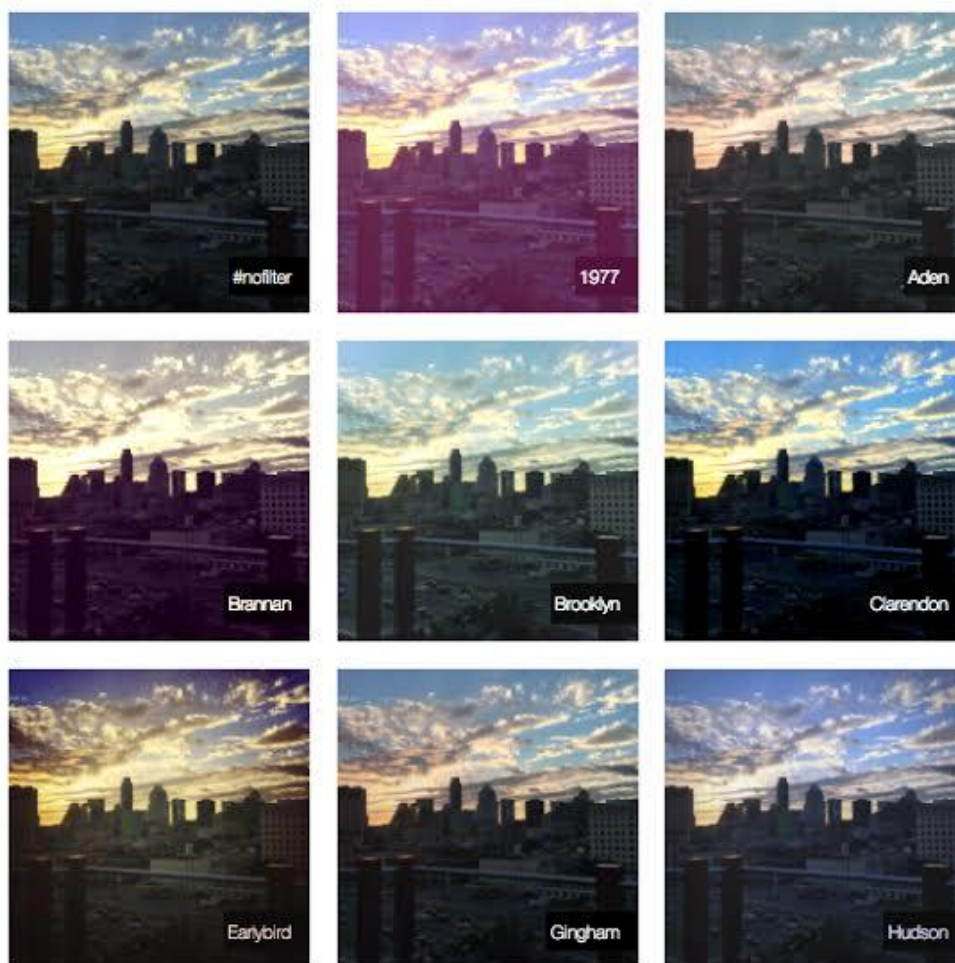
Now let us take a deep dive on the various functions of OpenCV to understand the various image processing techniques. Head to Jupyter Notebook for introduction to OpenCV given on this link: http://bit.ly/cv_notebook

Convolution

We have learnt that computers store images in numbers, and that pixels are arranged in a particular manner to create the picture we can recognize. These pixels have value varying from 0 to 255 and the value of the pixel determines the color of that pixel.

But what if we edit these numbers, will it bring a change to the image? The answer is yes. As we change the values of these pixels, the image changes. This process of changing pixel values is the base of image editing.

We all use a lot of image editing software like photoshop and at the same time use apps like Instagram and snapchat, which apply filters to the image to enhance the quality of that image.



As you can see, different filters applied to an image change the pixel values evenly throughout the image. How does this happen? This is done with the help of the process of convolution and the convolution operator which is commonly used to create these effects.

Before we understand how the convolution operation works, let us try and create a theory for the convolution operator by experiencing it using an online application.

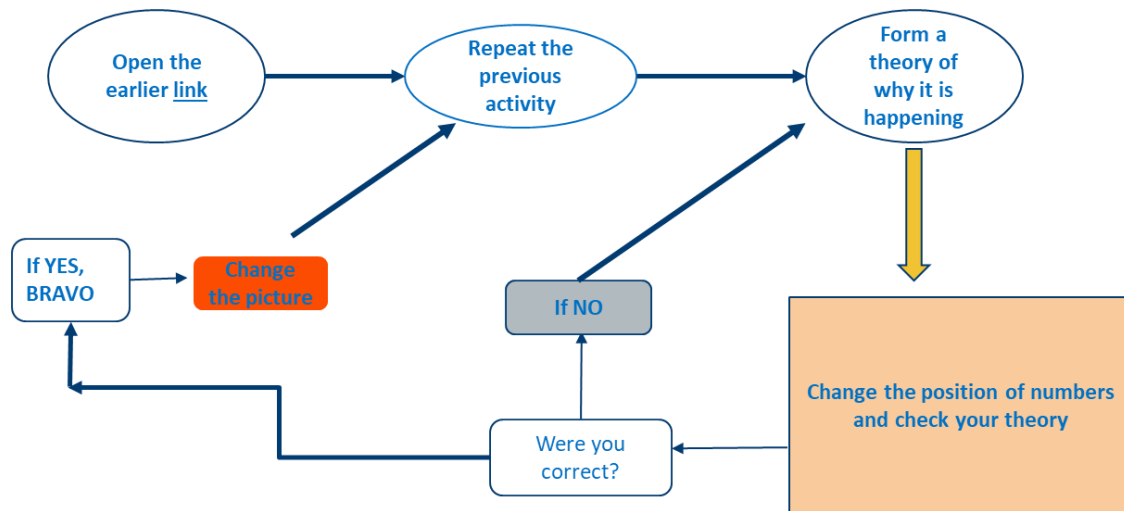
Task

Go to the link <http://matlabtricks.com/post-5/3x3-convolution-kernels-with-online-demo> and at the bottom of the page click on load "Click to Load Application"

Once the application is loaded try different filters and apply it on the image. Observe how the value of the kernel is changing for different filters. Try these steps

- 1) Change all to positive values
- 2) Change all to negative values
- 3) Have a mixture of negative and positive values

Let us follow the following steps to understand how a convolution operator works. The steps to be followed are:



Try experimenting with the following values to come up with a theory:

- 1) Make 4 numbers negative. Keep the rest as 0.
- 2) Now make one of them as positive.
- 3) Observe what happens.
- 4) Now make the second positive.

What theory do you propose for convolution on the basis of the observation?

It is time to test the theory. Change the location of the four numbers and follow the above mentioned steps. Does your theory hold true?

If yes, change the picture and try whether the theory holds true or not. If it does not hold true, modify your theory and keep trying until it satisfies all the conditions.

Let's Discuss

What effect did you apply?

How did different kernels affect the image?

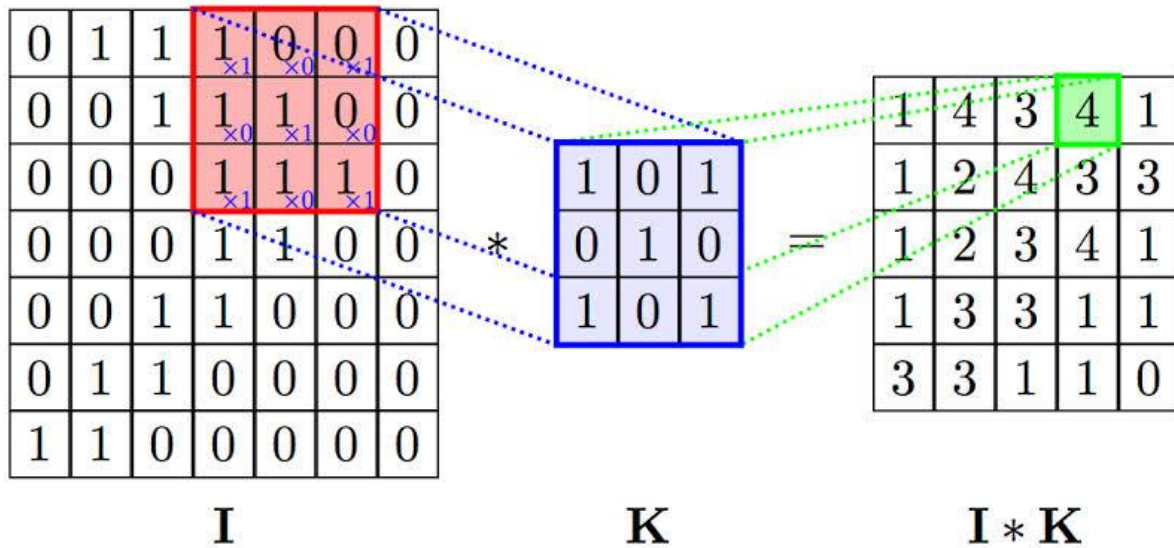
Why do you think we apply these effects?

How do you think the convolution operator works?

Convolution : Explained

Convolution is a simple Mathematical operation which is fundamental to many common image processing operators. Convolution provides a way of 'multiplying together' two arrays of numbers, generally of different sizes, but of the same dimensionality, to produce a third array of numbers of the same dimensionality.

An (image) **convolution is simply an element-wise multiplication of image arrays and another array called the kernel followed by sum.**



As you can see here,

I = Image Array

K = Kernel Array

I * K = Resulting array after performing the convolution operator

Note: The Kernel is passed over the whole image to get the resulting array after convolution.

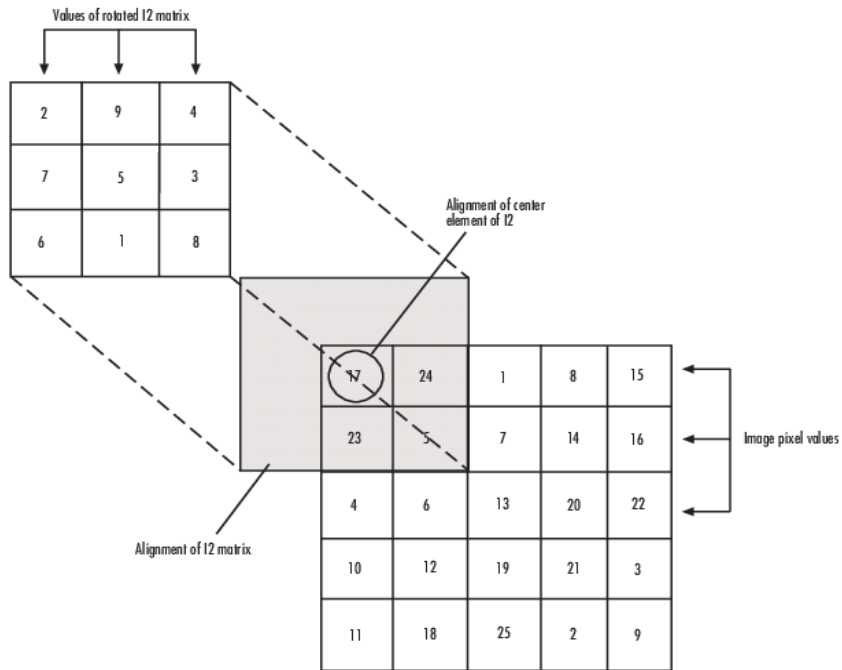
What is a Kernel?

A Kernel is a matrix, which is slid across the image and multiplied with the input such that the output is enhanced in a certain desirable manner. Each kernel has a different value for different kind of effects that we want to apply to an image.

In Image processing, we use the convolution operation to extract the features from the images which can be later used for further processing especially in Convolution Neural Network (CNN), about which we will study later in the chapter.

In this process, we overlap the centre of the image with the centre of the kernel to obtain the convolution output. In the process of doing it, the output image becomes smaller as the overlapping is done at the edge row and column of the image. What if we want the output image to be of exact size of the input image, how can we achieve this?

To achieve this, we need to extend the edge values out by one in the original image while overlapping the centres and performing the convolution. This will help us keep the input and output image of the same size. While extending the edges, the pixel values are considered as zero.



Let's try

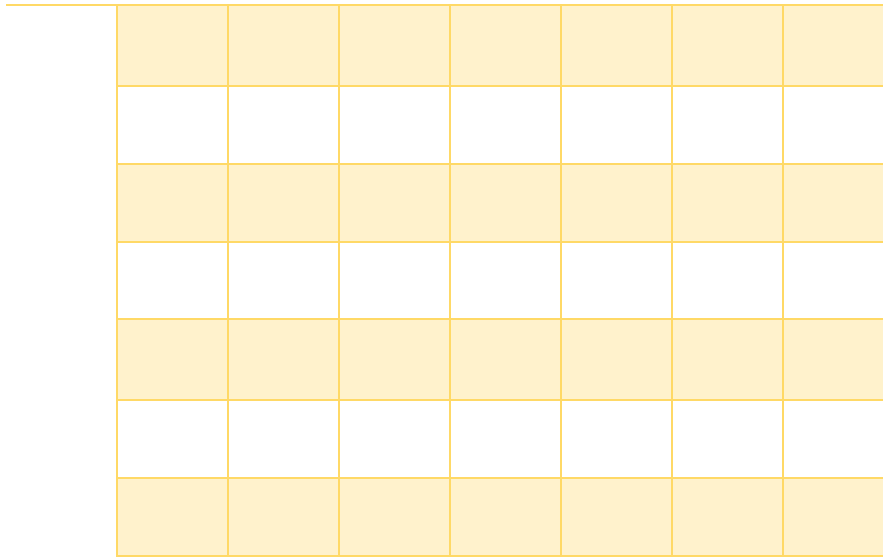
In this section we will try performing the convolution operator on paper to understand how it works. Fill the blank places of the output images by performing the convolution operation.

150	0	255	240	190	25	89	255
100	179	25	0	200	255	67	100
155	146	13	20	0	12	45	0
100	175	0	25	25	15	0	0
120	156	255	0	78	56	23	0
115	113	25	90	0	80	56	155
135	190	115	116	178	0	145	165
123	255	255	0	255	255	255	0



-1	0	-1
0	-1	0
-1	0	-1

Write Your Output Here :



Summary

1. Convolution is a common tool used for image editing.
2. It is an element wise multiplication of an image and a kernel to get the desired output.
3. In computer vision application, it is used in Convolutional Neural Network (CNN) to extract image features.

Convolution Neural Networks (CNN)

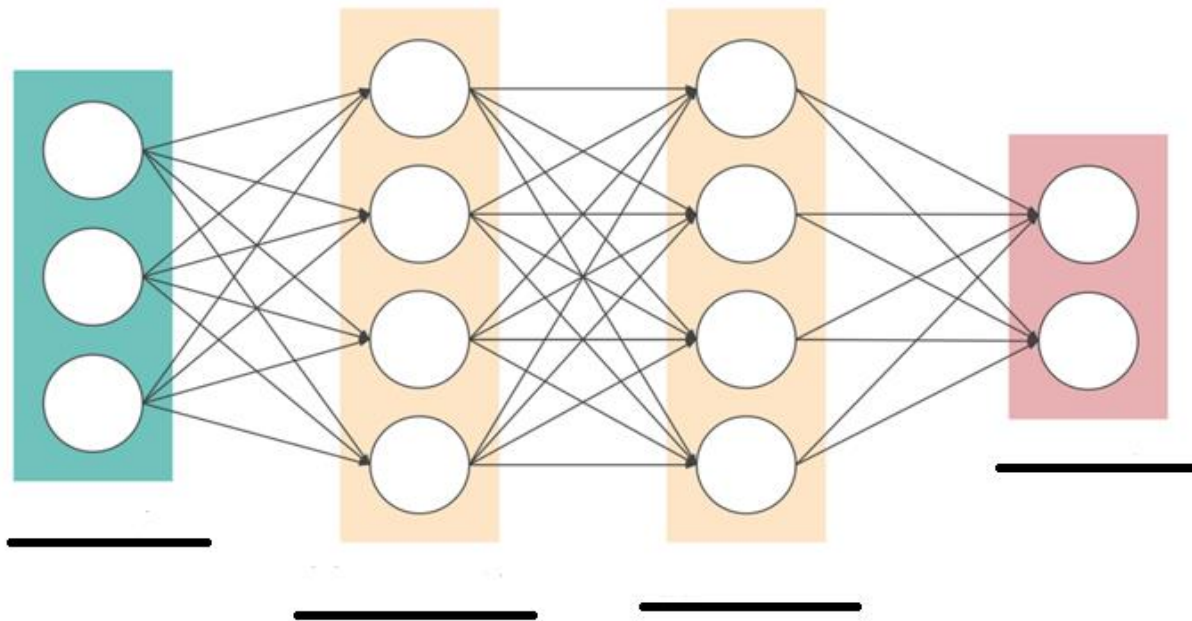
Introduction

In class 9, you studied about the concepts of Neural Network. You played a neural network game to understand how a neural network works.

Let's recall

What is a Neural Network?

Fill in the names of different layers of Neural Network.

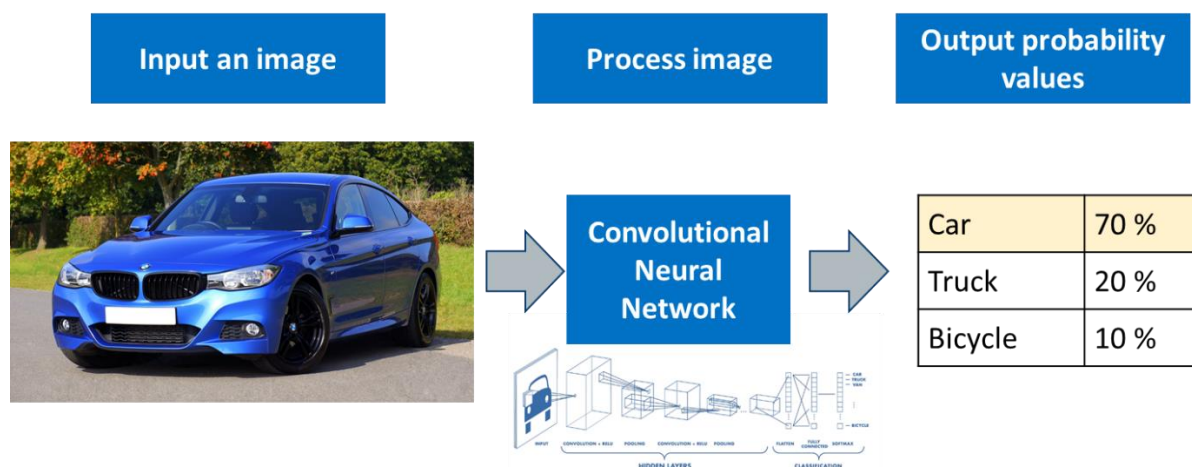


Did you get the answers right? In this section, we are going to study about one such neural network which is Convolutional Neural Network (CNN). Many of the current computer vision applications use a powerful neural network called the convolutional neural network.

What is a Convolutional Neural Network?

A **Convolutional Neural Network (CNN)** is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other.

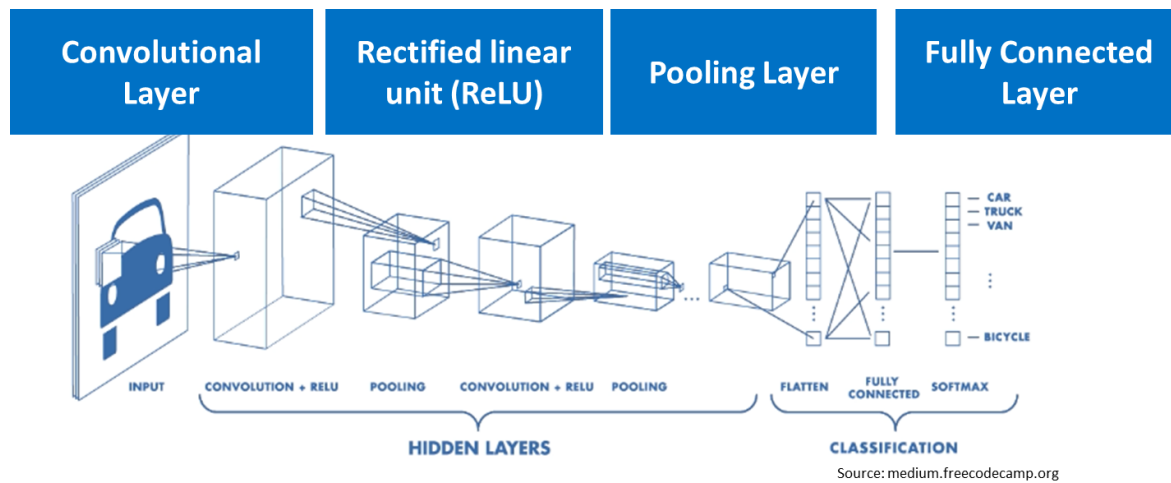
The process of deploying a CNN is as follows:



* Images shown here are the property of individual organisations and are used here for reference purpose only.

In the above diagram, we give an input image, which is then processed through a CNN and then gives prediction on the basis of the label given in the particular dataset.

The different layers of a Convolutional Neural Network (CNN) is as follows:



A convolutional neural network consists of the following layers:

- 1) Convolution Layer
- 2) Rectified linear Unit (ReLU)
- 3) Pooling Layer
- 4) Fully Connected Layer

Convolution Layer

It is the first layer of a CNN. The objective of the Convolution Operation is to extract the **high-level features** such as edges, from the input image. CNN need not be limited to only one Convolutional Layer. Conventionally, the first Convolution Layer is responsible for capturing the Low-Level features such as edges, colour, gradient orientation, etc. With added layers, the architecture adapts to the High-Level features as well, giving us a network which has the wholesome understanding of images in the dataset.

It uses convolution operation on the images. In the convolution layer, there are several kernels that are used to produce several features. The output of this layer is called the feature map. A feature map is also called the activation map. We can use these terms interchangeably.

There's several uses we derive from the feature map:

- We **reduce the image size** so that it can be processed more efficiently.
- We only focus on the features of the image that can help us in processing the image further.

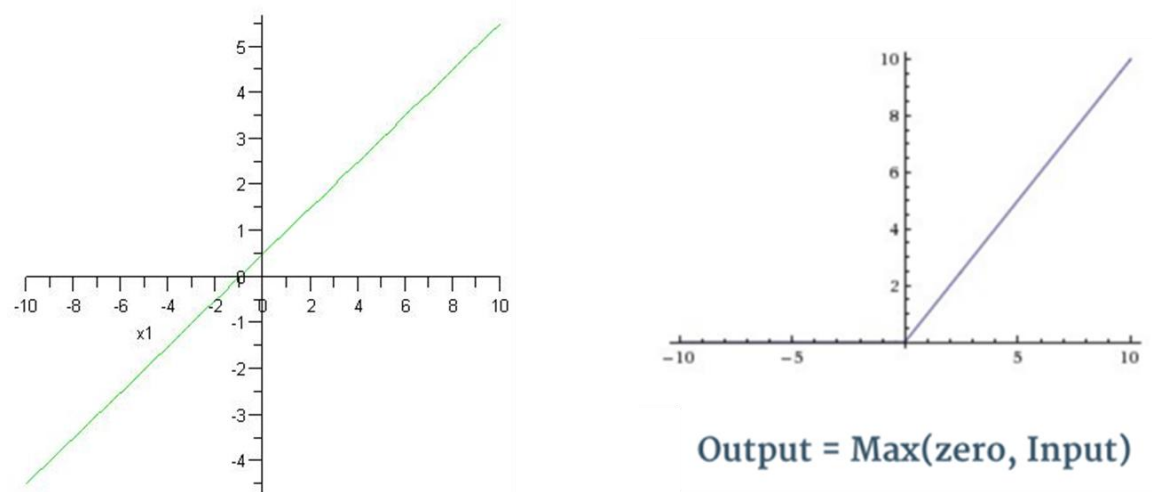
For example, you might only need to recognize someone's eyes, nose and mouth to recognize the person. You might not need to see the whole face.



Rectified Linear Unit Function

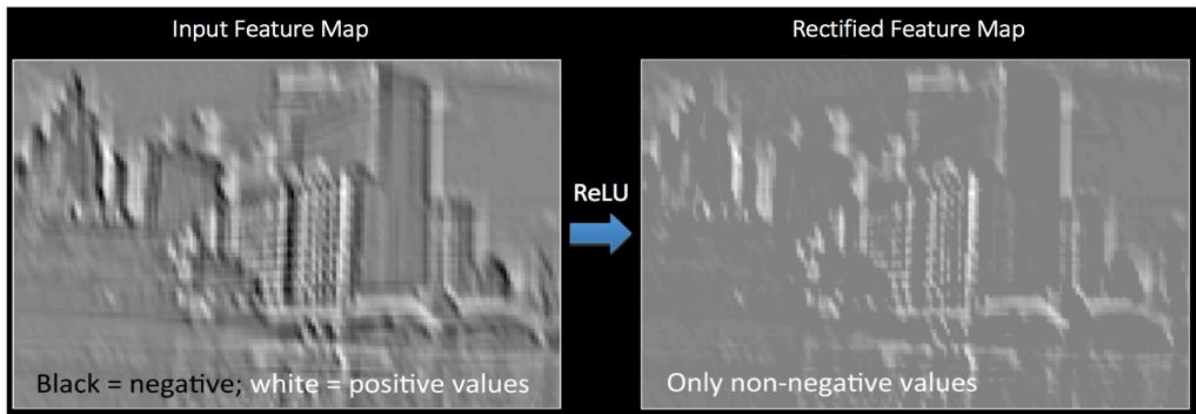
The next layer in the Convolution Neural Network is the Rectified Linear Unit function or the ReLU layer. After we get the feature map, it is then passed onto the ReLU layer. This layer simply gets rid of all the negative numbers in the feature map and lets the positive number stay as it is.

The process of passing it to the ReLU layer introduces non – linearity in the feature map. Let us see it through a graph.



If we see the two graphs side by side, the one on the left is a linear graph. This graph when passed through the ReLU layer, gives the one on the right. The ReLU graph starts with a horizontal straight line and then increases linearly as it reaches a positive number.

Now the question arises, why do we pass the feature map to the ReLU layer? it is to make the colour change more obvious and more abrupt?



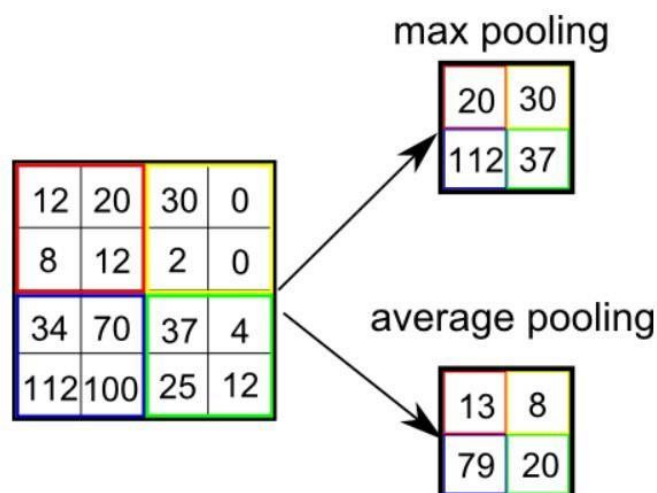
As shown in the above convolved image, there is a smooth grey gradient change from black to white. After applying the ReLU function, we can see a more abrupt change in color which makes the edges more obvious which acts as a better feature for the further layers in a CNN as it enhances the activation layer.

Pooling Layer

Similar to the Convolutional Layer, the Pooling layer is responsible for reducing the spatial size of the Convolved Feature while still retaining the important features.

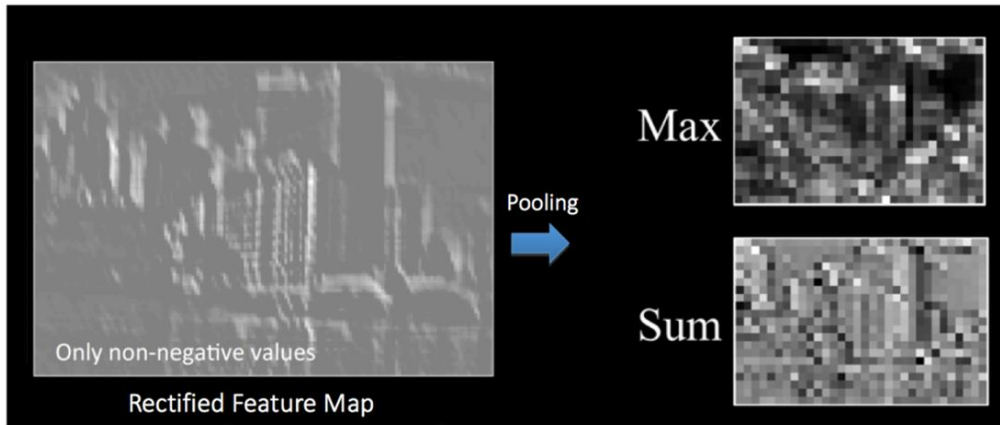
There are two types of pooling which can be performed on an image.

- 1) Max Pooling : Max Pooling returns the maximum value from the portion of the image covered by the Kernel.
- 2) Average Pooling: Average Pooling returns the average value from the portion of the image covered by the Kernel.

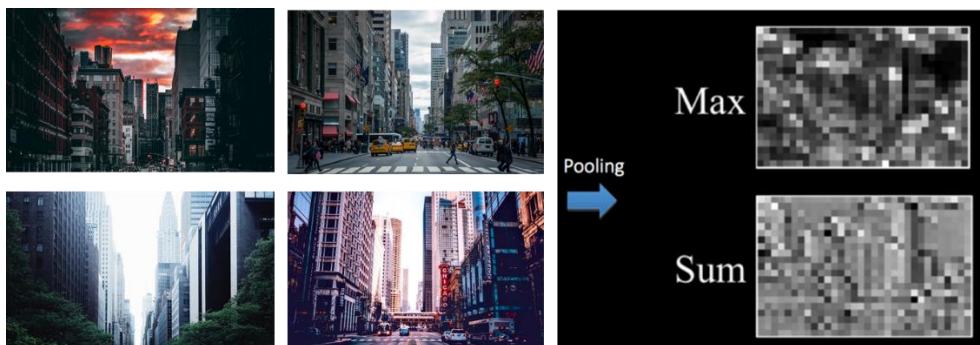


The pooling layer is an important layer in the CNN as it performs a series of tasks which are as follows :

- 1) Makes the image smaller and more manageable
- 2) Makes the image more resistant to small transformations, distortions and translations in the input image.



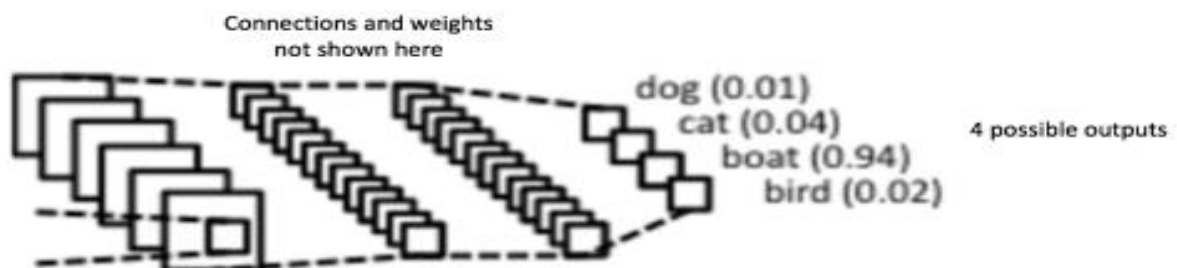
A small difference in input image will create very similar pooled image.



Fully Connected Layer

The final layer in the CNN is the Fully Connected Layer (FCP). The objective of a fully connected layer is to take the results of the convolution/pooling process and use them to classify the image into a label (in a simple classification example).

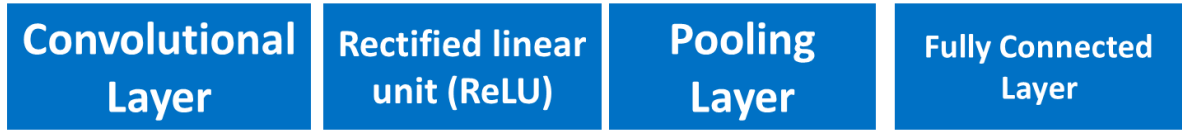
The output of convolution/pooling is flattened into a single vector of values, each representing a probability that a certain feature belongs to a label. For example, if the image is of a cat, features representing things like whiskers or fur should have high probabilities for the label "cat".



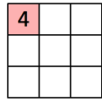
* Images shown here are the property of individual organisations and are used here for reference purpose only.

Let's Summarize:

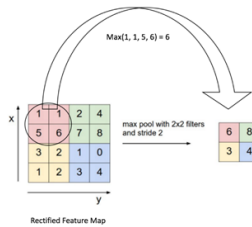
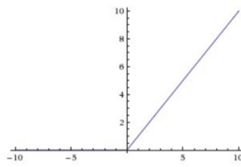
Convolutional Neural Network



Image



Convolved
Feature



Rectified Feature Map

Car	70 %
Truck	20 %
Bicycle	10 %

Reduce size, improve feature, give probability value

Write the whole process of how a CNN works on the basis of the above diagram.

* Images shown here are the property of individual organisations and are used here for reference purpose only.

Let's Experience

Now let us see how this comes into practice. To see that, go to the link <http://scs.ryerson.ca/~aharley/vis/conv/flat.html>

This is an online application of classifying different numbers. We need to analyse the different layers in the application on the basis of the CNN that we have studied in the previous section.

Natural Language Processing

Introduction

Till now, we have explored two domains of AI: Data Science and Computer Vision. Both these domains differ from each other in terms of the data on which they work. Data Science works around numbers and tabular data while Computer Vision is all about visual data like images and videos. The third domain, Natural Language Processing (commonly called NLP) takes in the data of Natural Languages which humans use in their daily lives and operates on this.

Natural Language Processing, or NLP, is the sub-field of AI that is focused on enabling computers to understand and process human languages. AI is a subfield of Linguistics, Computer Science, Information Engineering, and Artificial Intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyse large amounts of natural language data.

But how do computers do that? How do they understand what we say in our language? This chapter is all about demystifying the Natural Language Processing domain and understanding how it works.

Before we get deeper into NLP, let us experience it with the help of this AI Game:



Identify the mystery animal: <http://bit.ly/iai4yma>

Go to this link on Google Chrome, launch the experiment and try to identify the Mystery Animal by asking the machine 20 Yes or No questions.

Were you able to guess the animal?

If yes, in how many questions were you able to guess it?

If no, how many times did you try playing this game?

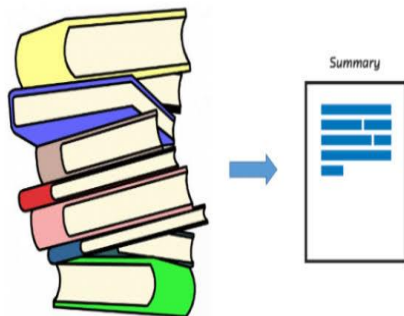
What according to you was the task of the machine?

Were there any challenges that you faced while playing this game? If yes, list them down.

What approach must one follow to win this game?

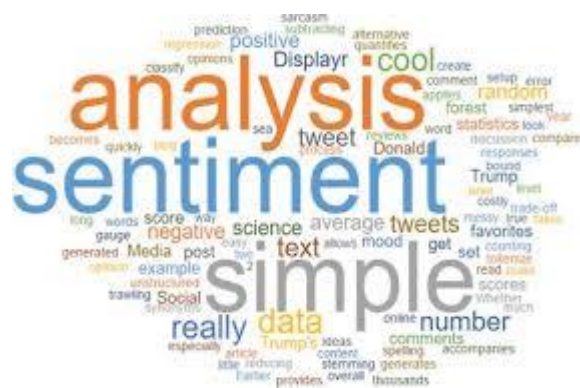
Applications of Natural Language Processing

Since Artificial Intelligence nowadays is becoming an integral part of our lives, its applications are very commonly used by the majority of people in their daily lives. Here are some of the applications of Natural Language Processing which are used in the real-life scenario:



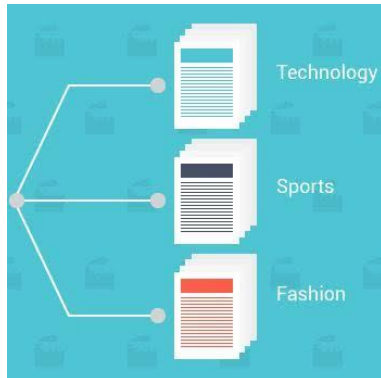
Automatic Summarization: Information overload is a real problem when we need to access a specific, important piece of information from a huge knowledge base. Automatic summarization is relevant not only for summarizing the meaning of documents and information, but also to understand the emotional meanings within the information, such as in collecting data from social media. Automatic summarization is especially relevant when used to provide an overview of a news item or blog post, while avoiding redundancy from multiple sources and maximizing the diversity of content obtained.

Sentiment Analysis: The goal of sentiment analysis is to identify sentiment among several posts or even in the same post where emotion is not always explicitly expressed. Companies use Natural Language Processing applications, such as sentiment analysis, to identify opinions and sentiment online to help them understand what customers think about their products and services (i.e., “I love the new iPhone” and, a few lines later “But sometimes it doesn’t work well” where the person is still talking about the iPhone) and overall



* Images shown here are the property of individual organisations and are used here for reference purpose only.

indicators of their reputation. Beyond determining simple polarity, sentiment analysis understands sentiment in context to help better understand what's behind an expressed opinion, which can be extremely relevant in understanding and driving purchasing decisions.



Text classification: Text classification makes it possible to assign predefined categories to a document and organize it to help you find the information you need or simplify some activities. For example, an application of text categorization is spam filtering in email.

Virtual Assistants: Nowadays Google Assistant, Cortana, Siri, Alexa, etc have become an integral part of our lives. Not only can we talk to them but they also have the abilities to make our lives easier. By accessing our data, they can help us in keeping notes of our tasks, make calls for us, send messages and a lot more. With the help of speech recognition, these assistants can not only detect our speech but can also make sense out of it. According to recent researches, a lot more advancements are expected in this field in the near future.



Natural Language Processing: Getting Started

Natural Language Processing is all about how machines try to understand and interpret human language and operate accordingly. But how can Natural Language Processing be used to solve the problems around us? Let us take a look.

Revisiting the AI Project Cycle

Let us try to understand how we can develop a project in Natural Language processing with the help of an example.

The Scenario



The world is competitive nowadays. People face competition in even the tiniest tasks and are expected to give their best at every point in time. When people are unable to meet these expectations, they get stressed and could even go into depression. We get to hear a lot of cases where people are depressed due to reasons like peer pressure, studies, family issues, relationships, etc. and they eventually get into something that is bad for them as well as for others. So, to overcome this, cognitive behavioural therapy (CBT) is considered to be one of the best methods to address stress as it is easy to implement on people and also gives good results. This therapy includes

understanding the behaviour and mindset of a person in their normal life. With the help of CBT, therapists help people overcome their stress and live a happy life.

To understand more about the concept of this therapy, visit this link:

https://en.wikipedia.org/wiki/Cognitive_behavioral_therapy

Problem Scoping

CBT is a technique used by most therapists to cure patients out of stress and depression. But it has been observed that people do not wish to seek the help of a psychiatrist willingly. They try to avoid such interactions as much as possible. Thus, there is a need to bridge the gap between a person who needs help and the psychiatrist. Let us look at various factors around this problem through the 4Ws problem canvas.

Who Canvas – *Who has the problem?*

Who are the stakeholders?	<ul style="list-style-type: none"> ○ People who suffer from stress and are at the onset of depression.
What do we know about them?	<ul style="list-style-type: none"> ○ People who are going through stress are reluctant to consult a psychiatrist.

What Canvas – *What is the nature of the problem?*

What is the problem?	<ul style="list-style-type: none"> ○ People who need help are reluctant to consult a psychiatrist and hence live miserably.
How do you know it is a problem?	<ul style="list-style-type: none"> ○ Studies around mental stress and depression available on various authentic sources.

Where Canvas – *Where does the problem arise?*

What is the context/situation in which the stakeholders experience this problem?	<ul style="list-style-type: none"> ○ When they are going through a stressful period of time ○ Due to some unpleasant experiences
--	--

Why Canvas – *Why do you think it is a problem worth solving?*

What would be of key value to the stakeholders?	<ul style="list-style-type: none"> ○ People get a platform where they can talk and vent out their feelings anonymously ○ People get a medium that can interact with them and applies primitive CBT on them and can suggest help whenever needed
How would it improve their situation?	<ul style="list-style-type: none"> ○ People would be able to vent out their stress ○ They would consider going to a psychiatrist whenever required

Now that we have gone through all the factors around the problem, the problem statement templates go as follows:

Our	People undergoing stress	Who?
Have a problem of	Not being able to share their feelings	What?
While	They need help in venting out their emotions	Where?
An ideal solution would	Provide them a platform to share their thoughts anonymously and suggest help whenever required	Why

This leads us to the goal of our project which is:

“To create a chatbot which can interact with people, help them to vent out their feelings and take them through primitive CBT.”

Data Acquisition

To understand the sentiments of people, we need to collect their conversational data so the machine can interpret the words that they use and understand their meaning. Such data can be collected from various means:

1. Surveys
2. Observing the therapist’s sessions
3. Databases available on the internet
4. Interviews, etc.

Data Exploration

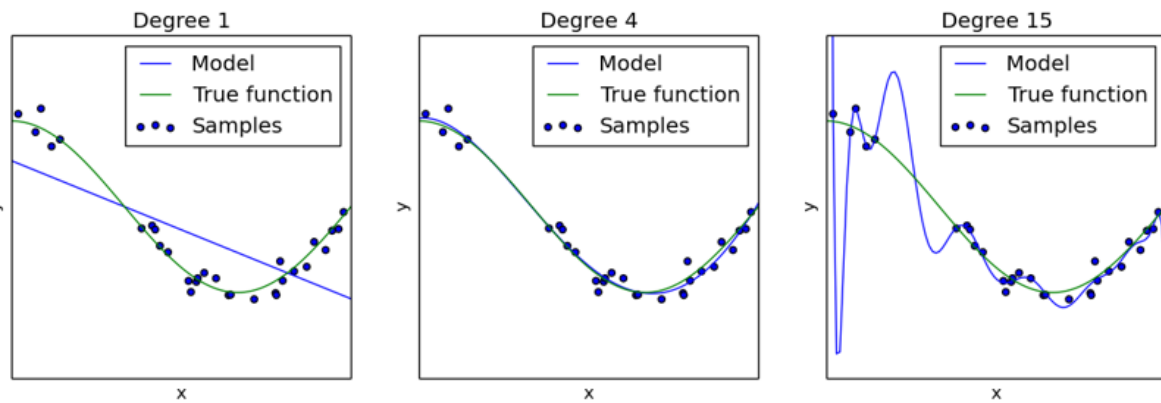
Once the textual data has been collected, it needs to be processed and cleaned so that an easier version can be sent to the machine. Thus, the text is normalised through various steps and is lowered to minimum vocabulary since the machine does not require grammatically correct statements but the essence of it.

Modelling

Once the text has been normalised, it is then fed to an NLP based AI model. Note that in NLP, modelling requires data pre-processing only after which the data is fed to the machine. Depending upon the type of chatbot we try to make, there are a lot of AI models available which help us build the foundation of our project.

Evaluation

The model trained is then evaluated and the accuracy for the same is generated on the basis of the relevance of the answers which the machine gives to the user’s responses. To understand the efficiency of the model, the suggested answers by the chatbot are compared to the actual answers.



As you can see in the above diagram, the blue line talks about the model's output while the green one is the actual output along with the data samples.

Figure 1

The model's output does not match the true function at all. Hence the model is said to be underfitting and its accuracy is lower.

Figure 2

In the second one, the model's performance matches well with the true function which states that the model has optimum accuracy and the model is called a perfect fit.





Figure 3

In the third case, model performance is trying to cover all the data samples even if they are out of alignment to the true function. This model is said to be overfitting and this too has a lower accuracy.



Once the model is evaluated thoroughly, it is then deployed in the form of an app which people can use easily.

Chatbots

As we have seen earlier, one of the most common applications of Natural Language Processing is a chatbot. There are a lot of chatbots available and many of them use the same approach as we used in the scenario above.. Let us try some of the chatbots and see how they work.

	<ul style="list-style-type: none"> • Mitsuku Bot* https://www.pandorabots.com/mitsuku/
	<ul style="list-style-type: none"> • CleverBot* https://www.cleverbot.com/
	<ul style="list-style-type: none"> • Jabberwacky* http://www.jabberwacky.com/
	<ul style="list-style-type: none"> • Haptik* https://haptik.ai/contact-us

* Images shown here are the property of individual organisations and are used here for reference purpose only.

	<ul style="list-style-type: none"> • Rose* <p>http://ec2-54-215-197-164.us-west-1.compute.amazonaws.com/speech.php</p>
	<ul style="list-style-type: none"> • Ochatbot* <p>https://www.ometrics.com/blog/list-of-fun-chatbots/</p>

Let us discuss!

- Which chatbot did you try? Name any one.
- What is the purpose of this chatbot?
- How was the interaction with the chatbot?
- Did the chat feel like talking to a human or a robot? Why do you think so?
- Do you feel that the chatbot has a certain personality?

As you interact with more and more chatbots, you would realise that some of them are scripted or in other words are traditional chatbots while others were AI-powered and had more knowledge. With the help of this experience, we can understand that there are 2 types of chatbots around us: Script-bot and Smart-bot. Let us understand what each of them mean in detail:

Script-bot	Smart-bot
Script bots are easy to make	Smart-bots are flexible and powerful
Script bots work around a script which is programmed in them	Smart bots work on bigger databases and other resources directly
Mostly they are free and are easy to integrate to a messaging platform	Smart bots learn with more data
No or little language processing skills	Coding is required to take this up on board
Limited functionality	Wide functionality

The story speaker activity which was done in class 9 can be considered as a script-bot as in that activity we used to create a script around which the interactive story revolved. As soon as the machine got triggered by the person, it used to follow the script and answer accordingly. Other examples of script bot may include the bots which are deployed in the customer care section of various companies. Their job is to answer some basic queries that they are coded for and connect them to human executives once they are unable to handle the conversation.

On the other hand, all the assistants like Google Assistant, Alexa, Cortana, Siri, etc. can be taken as smart bots as not only can they handle the conversations but can also manage to do other tasks which makes them smarter.

Human Language VS Computer Language

Humans communicate through language which we process all the time. Our brain keeps on processing the sounds that it hears around itself and tries to make sense out of them all the time. Even in the classroom, as the teacher delivers the session, our brain is continuously processing everything and storing it in some place. Also, while this is happening, when your friend whispers something, the focus of your brain automatically shifts from the teacher's speech to your friend's conversation. So now, the brain is processing both the sounds but is prioritising the one on which our interest lies.

* Images shown here are the property of individual organisations and are used here for reference purpose only.

Multiple Meanings of a word

Let's consider these three sentences:

His face turned red after he found out that he took the wrong bag

What does this mean? Is he feeling ashamed because he took another person's bag instead of his? Is he feeling angry because he did not manage to steal the bag that he has been targeting?

The red car zoomed past his nose

Probably talking about the color of the car

His face turns red after consuming the medicine

Is he having an allergic reaction? Or is he not able to bear the taste of that medicine?

Here we can see that context is important. We understand a sentence almost intuitively, depending on our history of using the language, and the memories that have been built within. In all three sentences, the word red has been used in three different ways which according to the context of the statement changes its meaning completely. Thus, in natural language, it is important to understand that a word can have multiple meanings and the meanings fit into the statement according to the context of it.

Think of some other words which can have multiple meanings and use them in sentences.

Perfect Syntax, no Meaning

Sometimes, a statement can have a perfectly correct syntax but it does not mean anything. For example, take a look at this statement:

Chickens feed extravagantly while the moon drinks tea.

This statement is correct grammatically but does this make any sense? In Human language, a perfect balance of syntax and semantics is important for better understanding.

In CBT, we learn to decipher the lies we are undermining ourselves with— based on the bias embedded in the things we say. For example, “I’m never going to make any friends” is an example of all-or-nothing thinking and we feel bad because we buy into this thought.



1. In CBT, we learn to decipher the lies we are undermining ourselves with— based on the bias embedded in the things we say.
2. For example, “I’m never going to make any friends” is an example of all-or-nothing thinking and we feel bad because we buy into this thought.

Tokenisation

After segmenting the sentences, each sentence is then further divided into tokens. Tokens is a term used for any word or number or special character occurring in a sentence. Under tokenisation, every word, number and special character is considered separately and each of them is now a separate token.

In CBT, we learn to decipher the lies we are undermining ourselves with— based on the bias embedded in the things we say.

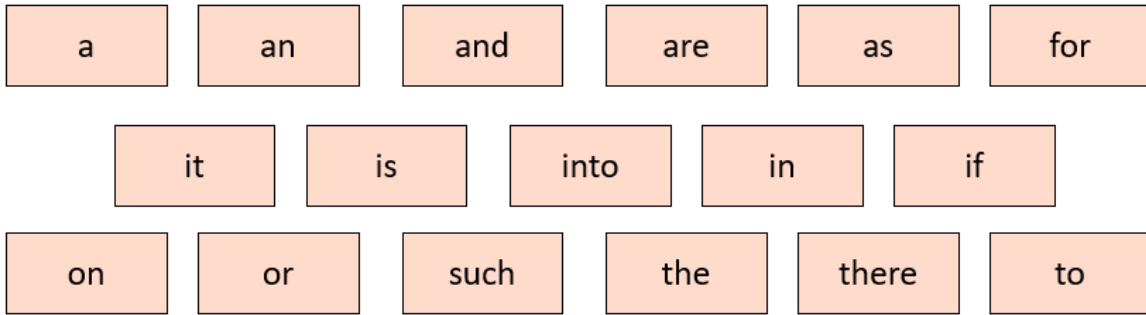


In CBT, we learn to decipher the lies we are undermining ourselves with— based on the bias embedded in the things we say.

Removing Stopwords, Special Characters and Numbers

In this step, the tokens which are not necessary are removed from the token list. What can be the possible words which we might not require?

Stopwords are the words which occur very frequently in the corpus but do not add any value to it. Humans use grammar to make their sentences meaningful for the other person to understand. But grammatical words do not add any essence to the information which is to be transmitted through the statement hence they come under stopwords. Some examples of stopwords are:

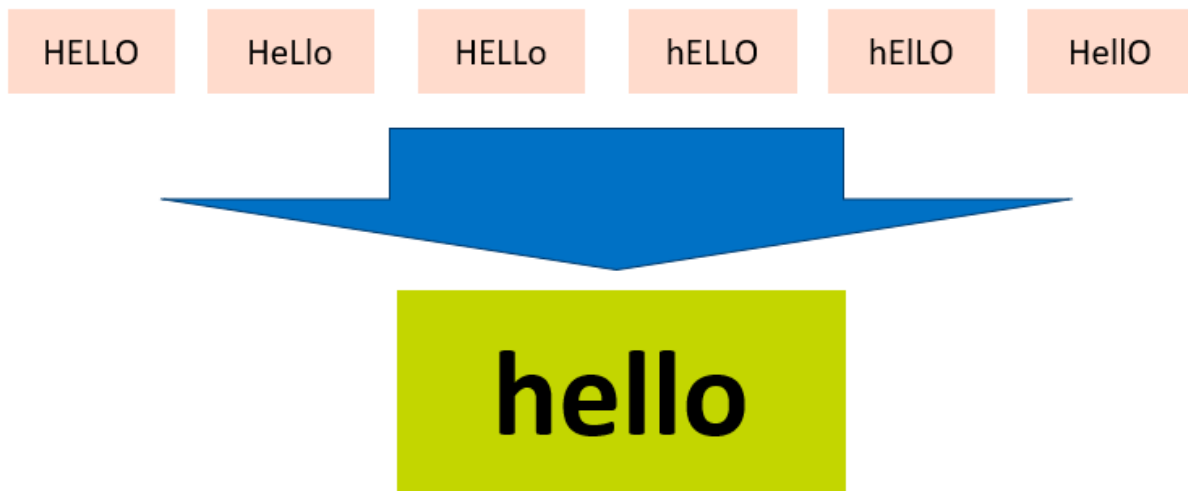


These words occur the most in any given corpus but talk very little or nothing about the context or the meaning of it. Hence, to make it easier for the computer to focus on meaningful terms, these words are removed.

Along with these words, a lot of times our corpus might have special characters and/or numbers. Now it depends on the type of corpus that we are working on whether we should keep them in it or not. For example, if you are working on a document containing email IDs, then you might not want to remove the special characters and numbers whereas in some other textual data if these characters do not make sense, then you can remove them along with the stopwords.

Converting text to a common case

After the stopwords removal, we convert the whole text into a similar case, preferably lower case. This ensures that the case-sensitivity of the machine does not consider same words as different just because of different cases.



Here in this example, the all the 6 forms of hello would be converted to lower case and hence would be treated as the same word by the machine.

Stemming

In this step, the remaining words are reduced to their root words. In other words, stemming is the process in which the affixes of words are removed and the words are converted to their base form.

Word	Affixes	Stem
healed	-ed	heal
healing	-ing	heal
healer	-er	heal
studies	-es	studi
studying	-ing	study

Note that in stemming, the stemmed words (words which are we get after removing the affixes) might not be meaningful. Here in this example as you can see: healed, healing and healer all were reduced to heal but studies was reduced to studi after the affix removal which is not a meaningful word. Stemming does not take into account if the stemmed word is meaningful or not. It just removes the affixes hence it is faster.

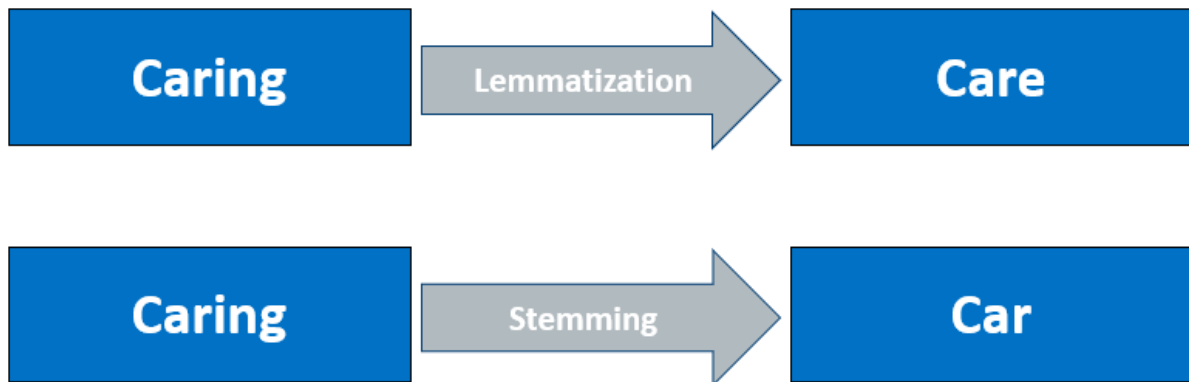
Lemmatization

Stemming and lemmatization both are alternative processes to each other as the role of both the processes is same – removal of affixes. But the difference between both of them is that in lemmatization, the word we get after affix removal (also known as lemma) is a meaningful one. Lemmatization makes sure that lemma is a word with meaning and hence it takes a longer time to execute than stemming.

Word	Affixes	lemma
healed	- <u>ed</u>	heal
healing	-ing	heal
healer	-er	heal
studies	-es	study
studying	-ing	study

As you can see in the same example, the output for studies after affix removal has become study instead of studi.

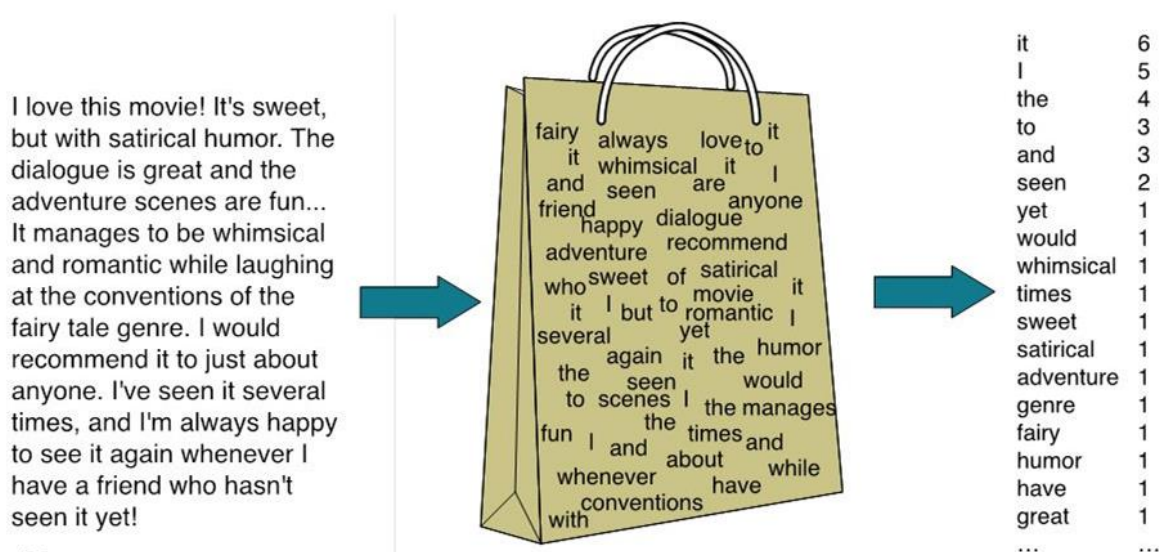
Difference between stemming and lemmatization can be summarized by this example:



With this we have normalised our text to tokens which are the simplest form of words present in the corpus. Now it is time to convert the tokens into numbers. For this, we would use the Bag of Words algorithm

Bag of Words

Bag of Words is a Natural Language Processing model which helps in extracting features out of the text which can be helpful in machine learning algorithms. In bag of words, we get the occurrences of each word and construct the vocabulary for the corpus.



This image gives us a brief overview about how bag of words works. Let us assume that the text on the left in this image is the normalised corpus which we have got after going through all the steps of text processing. Now, as we put this text into the bag of words algorithm, the algorithm returns to us the unique words out of the corpus and their occurrences in it. As you can see at the right, it shows us a list of words appearing in the corpus and the numbers corresponding to it shows how many times the word has occurred in the text body. Thus, we can say that the bag of words gives us two things:

1. A vocabulary of words for the corpus
2. The frequency of these words (number of times it has occurred in the whole corpus).

Here calling this algorithm "bag" of words symbolises that the sequence of sentences or tokens does not matter in this case as all we need are the unique words and their frequency in it.

Here is the step-by-step approach to implement bag of words algorithm:

1. Text Normalisation: Collect data and pre-process it
2. Create Dictionary: Make a list of all the unique words occurring in the corpus. (Vocabulary)
3. Create document vectors: For each document in the corpus, find out how many times the word from the unique list of words has occurred.
4. Create document vectors for all the documents.

Let us go through all the steps with an example:

Step 1: Collecting data and pre-processing it.

Document 1: *Aman and Anil are stressed*

Document 2: *Aman went to a therapist*

Document 3: *Anil went to download a health chatbot*

Here are three documents having one sentence each. After text normalisation, the text becomes:

Document 1: [aman, and, anil, are, stressed]

Document 2: [aman, went, to, a, therapist]

Document 3: [anil, went, to, download, a, health, chatbot]

Note that no tokens have been removed in the stopwords removal step. It is because we have very little data and since the frequency of all the words is almost the same, no word can be said to have lesser value than the other.

Step 2: Create Dictionary

Go through all the steps and create a dictionary i.e., list down all the words which occur in all three documents:

Dictionary:

aman	and	anil	are	stressed	went
download	health	chatbot	therapist	a	to

Note that even though some words are repeated in different documents, they are all written just once as while creating the dictionary, we create the list of unique words.

Step 3: Create document vector

In this step, the vocabulary is written in the top row. Now, for each word in the document, if it matches with the vocabulary, put a 1 under it. If the same word appears again, increment the previous value by 1. And if the word does not occur in that document, put a 0 under it.

aman	and	anil	are	stressed	went	to	a	therapist	download	health	chatbot
1	1	1	1	1	0	0	0	0	0	0	0

Since in the first document, we have words: aman, and, anil, are, stressed. So, all these words get a value of 1 and rest of the words get a 0 value.

Step 4: Repeat for all documents

Same exercise has to be done for all the documents. Hence, the table becomes:

aman	and	anil	are	stressed	went	to	a	therapist	download	health	chatbot
1	1	1	1	1	0	0	0	0	0	0	0
1	0	0	0	0	1	1	1	1	0	0	0
0	0	1	0	0	1	1	1	0	1	1	1

In this table, the header row contains the vocabulary of the corpus and three rows correspond to three different documents. Take a look at this table and analyse the positioning of 0s and 1s in it.

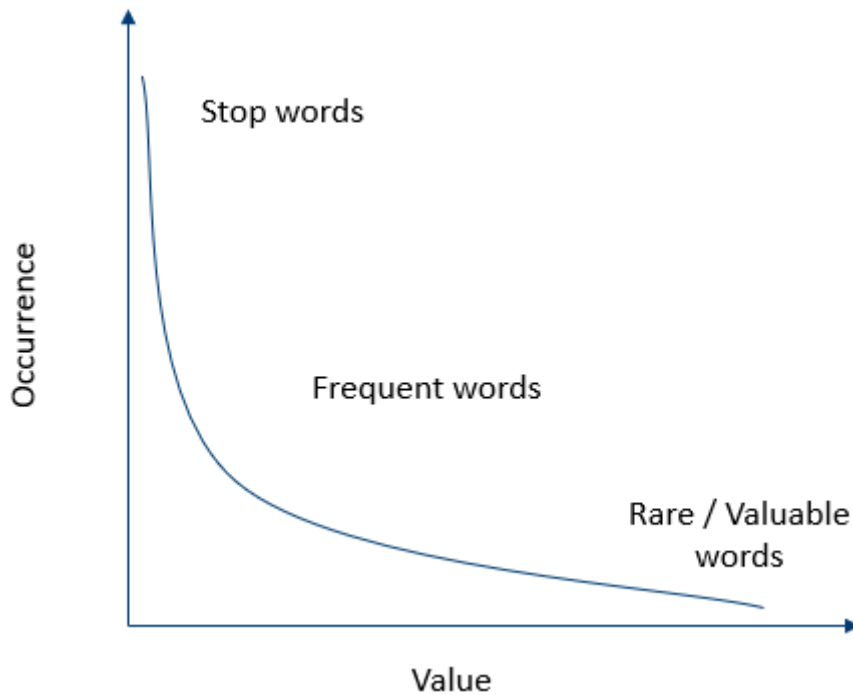
Finally, this gives us the **document vector table** for our corpus. But the tokens have still not converted to numbers. This leads us to the final steps of our algorithm: TFIDF.

TFIDF: Term Frequency & Inverse Document Frequency

Suppose you have a book. Which characters or words do you think would occur the most in it?

Bag of words algorithm gives us the frequency of words in each document we have in our corpus. It gives us an idea that if the word is occurring more in a document, its value is more for that document. For example, if I have a document on air pollution, air and pollution would be the words which occur many times in it. And these words are valuable too as they give us some context around the document. But let us suppose we have 10 documents and all of them talk about different issues. One is on women empowerment, the other is on unemployment and so on. Do you think air and pollution would still be one of the most occurring words in the whole corpus? If not, then which words do you think would have the highest frequency in all of them?

And, this, is, the, etc. are the words which occur the most in almost all the documents. But these words do not talk about the corpus at all. Though they are important for humans as they make the statements understandable to us, for the machine they are a complete waste as they do not provide us with any information regarding the corpus. Hence, these are termed as stopwords and are mostly removed at the pre-processing stage only.



Take a look at this graph. It is a plot of occurrence of words versus their value. As you can see, if the words have highest occurrence in all the documents of the corpus, they are said to have negligible value hence they are termed as stop words. These words are mostly removed at the pre-processing stage only. Now as we move ahead from the stopwords, the occurrence level drops drastically and the words which have adequate occurrence in the corpus are said to have some amount of value and are termed as frequent words. These words mostly talk about the document's subject and their occurrence is adequate in the corpus. Then as the occurrence of words drops further, the value of such words rises. These words are termed as rare or valuable words. These words occur the least but add the most value to the corpus. Hence, when we look at the text, we take frequent and rare words into consideration.

Let us now demystify TFIDF. TFIDF stands for Term Frequency and Inverse Document Frequency. TFIDF helps un in identifying the value for each word. Let us understand each term one by one.

Term Frequency

Term frequency is the frequency of a word in one document. Term frequency can easily be found from the document vector table as in that table we mention the frequency of each word of the vocabulary in each document.

aman	and	anil	are	stressed	went	to	a	therapist	download	health	chatbot
1	1	1	1	1	0	0	0	0	0	0	0
1	0	0	0	0	1	1	1	1	0	0	0
0	0	1	0	0	1	1	1	0	1	1	1

* Images shown here are the property of individual organisations and are used here for reference purpose only.

Here, you can see that the frequency of each word for each document has been recorded in the table. These numbers are nothing but the Term Frequencies!

Inverse Document Frequency

Now, let us look at the other half of TFIDF which is Inverse Document Frequency. For this, let us first understand what does document frequency mean. Document Frequency is the number of documents in which the word occurs irrespective of how many times it has occurred in those documents. The document frequency for the exemplar vocabulary would be:

aman	and	anil	are	stressed	went	to	a	therapist	download	health	Chatbot
2	1	2	1	1	2	2	2	1	1	1	1

Here, you can see that the document frequency of 'aman', 'anil', 'went', 'to' and 'a' is 2 as they have occurred in two documents. Rest of them occurred in just one document hence the document frequency for them is one.

Talking about inverse document frequency, we need to put the document frequency in the denominator while the total number of documents is the numerator. Here, the total number of documents are 3, hence inverse document frequency becomes:

aman	and	anil	are	stressed	went	to	a	therapist	download	health	chatbot
3/2	3/1	3/2	3/1	3/1	3/2	3/2	3/2	3/1	3/1	3/1	3/1

Finally, the formula of TFIDF for any word W becomes:

$$TFIDF(W) = TF(W) * \log(IDF(W))$$

Here, log is to the base of 10. Don't worry! You don't need to calculate the log values by yourself. Simply use the log function in the calculator and find out!

Now, let's multiply the IDF values to the TF values. Note that the TF values are for each document while the IDF values are for the whole corpus. Hence, we need to multiply the IDF values to each row of the document vector table.

aman	and	anil	are	stress	went	to	a	therapist	download	health	chatbot
1*log(3/2)	1*log(3)	1*log(3/2)	1*log(3)	1*log(3)	0*log(3/2)	0*log(3/2)	0*log(3/2)	0*log(3)	0*log(3)	0*log(3)	0*log(3)
1*log(3/2)	0*log(3)	0*log(3/2)	0*log(3)	0*log(3)	1*log(3/2)	1*log(3/2)	1*log(3/2)	1*log(3)	0*log(3)	0*log(3)	0*log(3)
0*log(3/2)	0*log(3)	1*log(3/2)	0*log(3)	0*log(3)	1*log(3/2)	1*log(3/2)	1*log(3/2)	0*log(3)	1*log(3)	1*log(3)	1*log(3)

Here, you can see that the IDF values for Aman in each row is the same and similar pattern is followed for all the words of the vocabulary. After calculating all the values, we get:

aman	and	anil	are	stress	went	to	a	therapist	download	health	chatbot
0.176	0.477	0.176	0.477	0.477	0	0	0	0	0	0	0
0.176	0	0	0	0	0.176	0.176	0.176	0.477	0	0	0
0	0	0.176	0	0	0.176	0.176	0.176	0	0.477	0.477	0.477

Finally, the words have been converted to numbers. These numbers are the values of each for each document. Here, you can see that since we have less amount of data, words like 'are' and 'and' also have a high value. But as the IDF value increases, the value of that word decreases. That is, for example:

Total Number of documents: 10

Number of documents in which 'and' occurs: 10

Therefore, $IDF(\text{and}) = 10/10 = 1$

Which means: $\log(1) = 0$. Hence, the value of 'and' becomes 0.

On the other hand, number of documents in which 'pollution' occurs: 3

$IDF(\text{pollution}) = 10/3 = 3.3333\dots$

Which means: $\log(3.3333) = 0.522$; which shows that the word 'pollution' has considerable value in the corpus.

Summarising the concept, we can say that:

1. Words that occur in all the documents with high term frequencies have the least values and are considered to be the stopwords.
2. For a word to have high TFIDF value, the word needs to have a high term frequency but less document frequency which shows that the word is important for one document but is not a common word for all documents.
3. These values help the computer understand which words are to be considered while processing the natural language. The higher the value, the more important the word is for a given corpus.

Applications of TFIDF

TFIDF is commonly used in the Natural Language Processing domain. Some of its applications are:

Document Classification	Topic Modelling	Information Retrieval System	Stop word filtering
Helps in classifying the type and genre of a document.	It helps in predicting the topic for a corpus.	To extract the important information out of a corpus.	Helps in removing the unnecessary words out of a text body.

DIY – Do It Yourself!

Here is a corpus for you to challenge yourself with the given tasks. Use the knowledge you have gained in the above sections and try completing the whole exercise by yourself.

The Corpus

Document 1: We can use health chatbots for treating stress.

Document 2: We can use NLP to create chatbots and we will be making health chatbots now!

Document 3: Health Chatbots cannot replace human counsellors now. Yay >< !! @1nteLA!4Y

Accomplish the following challenges on the basis of the corpus given above. You can use the tools available online for these challenges. Link for each tool is given below:

1. Sentence Segmentation: <https://tinyurl.com/y36hd92n>
2. Tokenisation: <https://text-processing.com/demo/tokenize/>
3. Stopwords removal: <https://demos.datasciencedojo.com/demo/stopwords/>
4. Lowercase conversion: <https://caseconverter.com/>
5. Stemming: <http://textanalysisonline.com/nltk-porter-stemmer>
6. Lemmatisation: <http://textanalysisonline.com/spacy-word-lemmatize>
7. Bag of Words: Create a document vector table for all documents.
8. Generate TFIDF values for all the words.
9. Find the words having highest value.
10. Find the words having the least value.

Evaluation

Introduction

Till now we have learnt about the 4 stages of AI project cycle, viz. Problem scoping, Data acquisition, Data exploration and modelling. While in modelling we can make different types of models, how do we check if one's better than the other? That's where Evaluation comes into play. In the Evaluation stage, we will explore different methods of evaluating an AI model. Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future

What is evaluation?

Evaluation is the process of understanding the reliability of any AI model, based on outputs by feeding test dataset into the model and comparing with actual answers. There can be different Evaluation techniques, depending of the type and purpose of the model. Remember that It's not recommended to use the data we used to build the model to evaluate it. This is because our model will simply remember the whole training set, and will therefore always predict the correct label for any point in the training set. This is known as overfitting.

Firstly, let us go through various terms which are very important to the evaluation process.

Model Evaluation Terminologies

There are various new terminologies which come into the picture when we work on evaluating our model. Let's explore them with an example of the Forest fire scenario.

The Scenario

Imagine that you have come up with an AI based prediction model which has been deployed in a forest which is prone to forest fires. Now, the objective of the model is to predict whether a forest fire has broken out in the forest or not. Now, to understand the efficiency of this model, we need to check if the predictions which it makes are correct or not. Thus, there exist two conditions which we need to ponder upon: Prediction and Reality. The prediction is the output which is given by the machine and the reality is the real scenario in the forest when the prediction has been made. Now let us look at various combinations that we can have with these two conditions.

Case 1: Is there a forest fire?



Prediction: Yes

Reality: Yes

True Positive

Here, we can see in the picture that a forest fire has broken out in the forest. The model predicts a Yes which means there is a forest fire. The Prediction matches with the Reality. Hence, this condition is termed as **True Positive**.

Case 2: Is there a forest fire?



Prediction: No

Reality: No

True Negative

Here there is no fire in the forest hence the reality is No. In this case, the machine too has predicted it correctly as a No. Therefore, this condition is termed as **True Negative**.

Case 3: Is there a forest fire?



Prediction: Yes

Reality: No

False Positive

Here the reality is that there is no forest fire. But the machine has incorrectly predicted that there is a forest fire. This case is termed as **False Positive**.

Case 4: Is there a forest fire?



Prediction: No

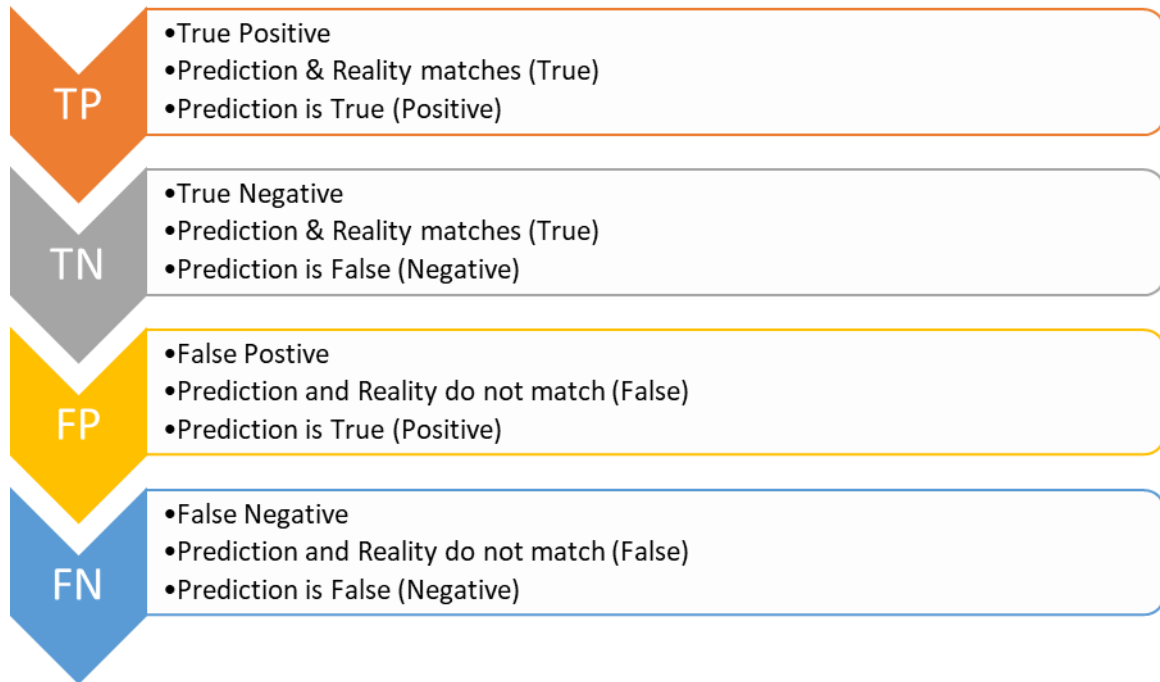
Reality: Yes

False Negative

Here, a forest fire has broken out in the forest because of which the Reality is Yes but the machine has incorrectly predicted it as a No which means the machine predicts that there is no Forest Fire. Therefore, this case becomes **False Negative**.

Confusion matrix

The result of comparison between the prediction and reality can be recorded in what we call the confusion matrix. The confusion matrix allows us to understand the prediction results. Note that it is not an evaluation metric but a record which can help in evaluation. Let us once again take a look at the four conditions that we went through in the Forest Fire example:



Let us now take a look at the confusion matrix:

The Confusion Matrix		Reality	
		Yes	No
Prediction	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

Prediction and Reality can be easily mapped together with the help of this confusion matrix.

* Images shown here are the property of individual organisations and are used here for reference purpose only.

Evaluation Methods

Now as we have gone through all the possible combinations of Prediction and Reality, let us see how we can use these conditions to evaluate the model.

Accuracy

Accuracy is defined as the percentage of correct predictions out of all the observations. A prediction can be said to be correct if it matches the reality. Here, we have two conditions in which the Prediction matches with the Reality: True Positive and True Negative. Hence, the formula for Accuracy becomes:

$$\text{Accuracy} = \frac{\text{Correct prediction}}{\text{Total cases}} * 100\%$$

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100\%$$

Here, total observations cover all the possible cases of prediction that can be True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

As we can see, Accuracy talks about how true the predictions are by any model. Let us ponder:

Is high accuracy equivalent to good performance?

How much percentage of accuracy is reasonable to show good performance?

Let us go back to the Forest Fire example. Assume that the model always predicts that there is no fire. But in reality, there is a 2% chance of forest fire breaking out. In this case, for 98 cases, the model will be right but for those 2 cases in which there was a forest fire, then too the model predicted no fire.

Here,

True Positives = 0

True Negatives = 98

Total cases = 100

Therefore, accuracy becomes: $(98 + 0) / 100 = 98\%$



Prediction: Always No

Reality: 2% probability of Yes

98% accurate
But is it usable?

This is a fairly high accuracy for an AI model. But this parameter is useless for us as the actual cases where the fire broke out are not taken into account. Hence, there is a need to look at another parameter which takes account of such cases as well.

Precision

Precision is defined as the percentage of true positive cases versus all the cases where the prediction is true. That is, it takes into account the True Positives and False Positives.

$$\text{Precision} = \frac{\text{True Positive}}{\text{All Predicted Positives}} * 100\%$$

$$\text{Precision} = \frac{TP}{TP + FP} * 100\%$$

Going back to the Forest Fire example, in this case, assume that the model always predicts that there is a forest fire irrespective of the reality. In this case, all the Positive conditions would be taken into account that is, True Positive (Prediction = Yes and Reality = Yes) and False Positive (Prediction = Yes and Reality = No). In this case, the firefighters will check for the fire all the time to see if the alarm was True or False.

You might recall the story of the boy who falsely cries out that there are wolves every time and so when they actually arrive, no one comes to his rescue. Similarly, here if the Precision is low (which means there are more False alarms than the actual ones) then the firefighters would get complacent and might not go and check every time considering it could be a false alarm.

This makes Precision an important evaluation criteria. If Precision is high, this means the True Positive cases are more, giving lesser False alarms.

But again, is good Precision equivalent to a good model performance? Why?



Prediction: 10 cases of TP

Reality: 20 cases of yes

100% precise
But is it usable?

Let us consider that a model has 100% precision. Which means that whenever the machine says there's a fire, there is actually a fire (True Positive). In the same model, there can be a rare exceptional case where there was actual fire but the system could not detect it. This is the case of a False Negative condition. But the precision value would not be affected by it because it does not take FN into account. Is precision then a good parameter for model performance?

Recall

Another parameter for evaluating the model's performance is Recall. It can be defined as the fraction of positive cases that are correctly identified. It majorly takes into account the true reality cases where in Reality there was a fire but the machine either detected it correctly or it didn't. That is, it considers True Positives (There was a forest fire in reality and the model predicted a forest fire) and False Negatives (There was a forest fire and the model didn't predict it).

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Now as we notice, we can see that the Numerator in both Precision and Recall is the same: True Positives. But in the denominator, Precision counts the False Positives while Recall takes False Negatives into consideration.

Let us ponder... Which one do you think is better? Precision or Recall? Why?

Which Metric is Important?

Choosing between Precision and Recall depends on the condition in which the model has been deployed. In a case like Forest Fire, a False Negative can cost us a lot and is risky too. Imagine no alert being given even when there is a Forest Fire. The whole forest might burn down.

Another case where a False Negative can be dangerous is Viral Outbreak. Imagine a deadly virus has started spreading and the model which is supposed to predict a viral outbreak does not detect it. The virus might spread widely and infect a lot of people.

On the other hand, there can be cases in which the False Positive condition costs us more than False Negatives. One such case is Mining. Imagine a model telling you that there exists treasure at a point and you keep on digging there but it turns out that it is a false alarm. Here, False Positive case (predicting there is treasure but there is no treasure) can be very costly.

Similarly, let's consider a model that predicts that a mail is spam or not. If the model always predicts that the mail is spam, people would not look at it and eventually might lose important information. Here also False Positive condition (Predicting the mail as spam while the mail is not spam) would have a high cost.

Cases with high FN cost

Cases with high FP cost

Forest fire

Viral

Spam

Mining

Which one is more important? Recall or Precision?

Think of some more examples having:

- High False Negative cost

- High False Positive cost

Both measures are important

High Precision,

High Recall,

Precision = $(TP) / (TP + FP)$

Recall = $(TP) / (TP + FN)$

We need something that account for the 2 metrics

To conclude the argument, we must say that if we want to know if our model's performance is good, we need these two measures: Recall and Precision. For some cases, you might have a High Precision but Low Recall or Low Precision but High Recall. But since both the measures are important, there is a need of a parameter which takes both Precision and Recall into account.

F1 Score

F1 score can be defined as the measure of balance between precision and recall.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Take a look at the formula and think of when can we get a perfect F1 score?

An ideal situation would be when we have a value of 1 (that is 100%) for both Precision and Recall. In that case, the F1 score would also be an ideal 1 (100%). It is known as the perfect value for F1 Score. As the values of both Precision and Recall ranges from 0 to 1, the F1 score also ranges from 0 to 1.

Let us explore the variations we can have in the F1 Score:

Precision	Recall	F1 Score
Low	Low	Low
Low	High	Low
High	Low	Low
High	High	High

In conclusion, we can say that a model has good performance if the F1 Score for that model is high.

Let's practice!

Let us understand the evaluation parameters with the help of examples.

Challenge

Find out Accuracy, Precision, Recall and F1 Score for the given problems.

Scenario 1:

In schools, a lot of times it happens that there is no water to drink. At a few places, cases of water shortage in schools are very common and prominent. Hence, an AI model is designed to predict if there is going to be a water shortage in the school in the near future or not. The confusion matrix for the same is:

The Confusion Matrix	Reality: 1	Reality: 0
Predicted: 1	22	12
Predicted: 0	47	118

Scenario 2:

Nowadays, the problem of floods has worsened in some parts of the country. Not only does it damage the whole place but it also forces people to move out of their homes and relocate. To address this issue, an AI model has been created which can predict if there is a chance of floods or not. The confusion matrix for the same is:

The Confusion Matrix	Actual: 1	Actual: 0
Predicted: 1	0	3
Predicted: 0	3	94

Scenario 3:

A lot of times people face the problem of sudden downpour. People wash clothes and put them out to dry but due to unexpected rain, their work gets wasted. Thus, an AI model has been created which predicts if there will be rain or not. The confusion matrix for the same is:

The Confusion Matrix	Actual: 1	Actual: 0
Predicted: 1	5	0
Predicted: 0	45	50

Scenario 4:

Traffic Jams have become a common part of our lives nowadays. Living in an urban area means you have to face traffic each and every time you get out on the road. Mostly, school students opt for buses to go to school. Many times the bus gets late due to such jams and students are not able to reach their school on time. Thus, an AI model is created to predict explicitly if there would be a traffic jam on their way to school or not. The confusion matrix for the same is:

The Confusion Matrix	Actual: 1	Actual: 0
Predicted: 1	50	50
Predicted: 0	0	0