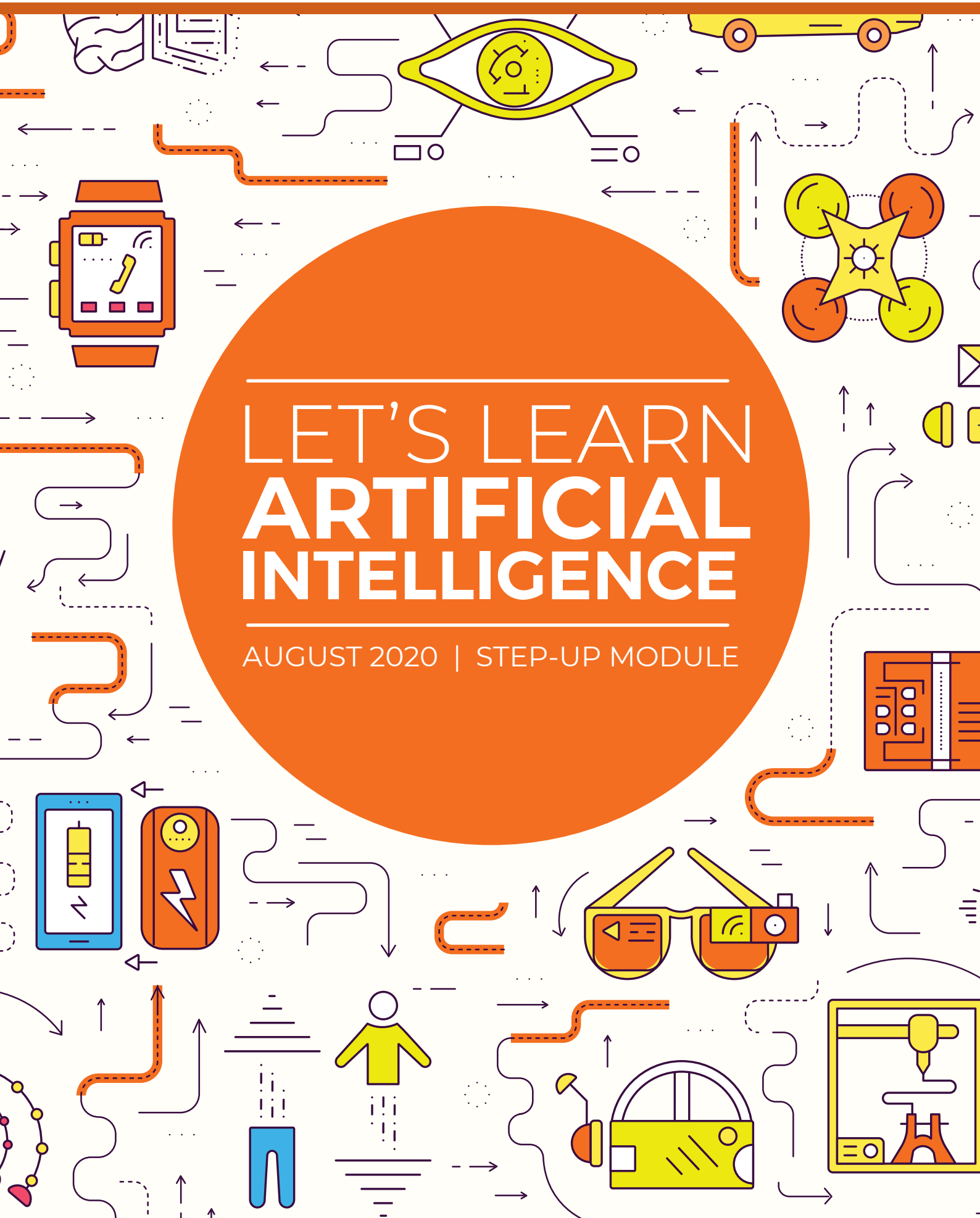
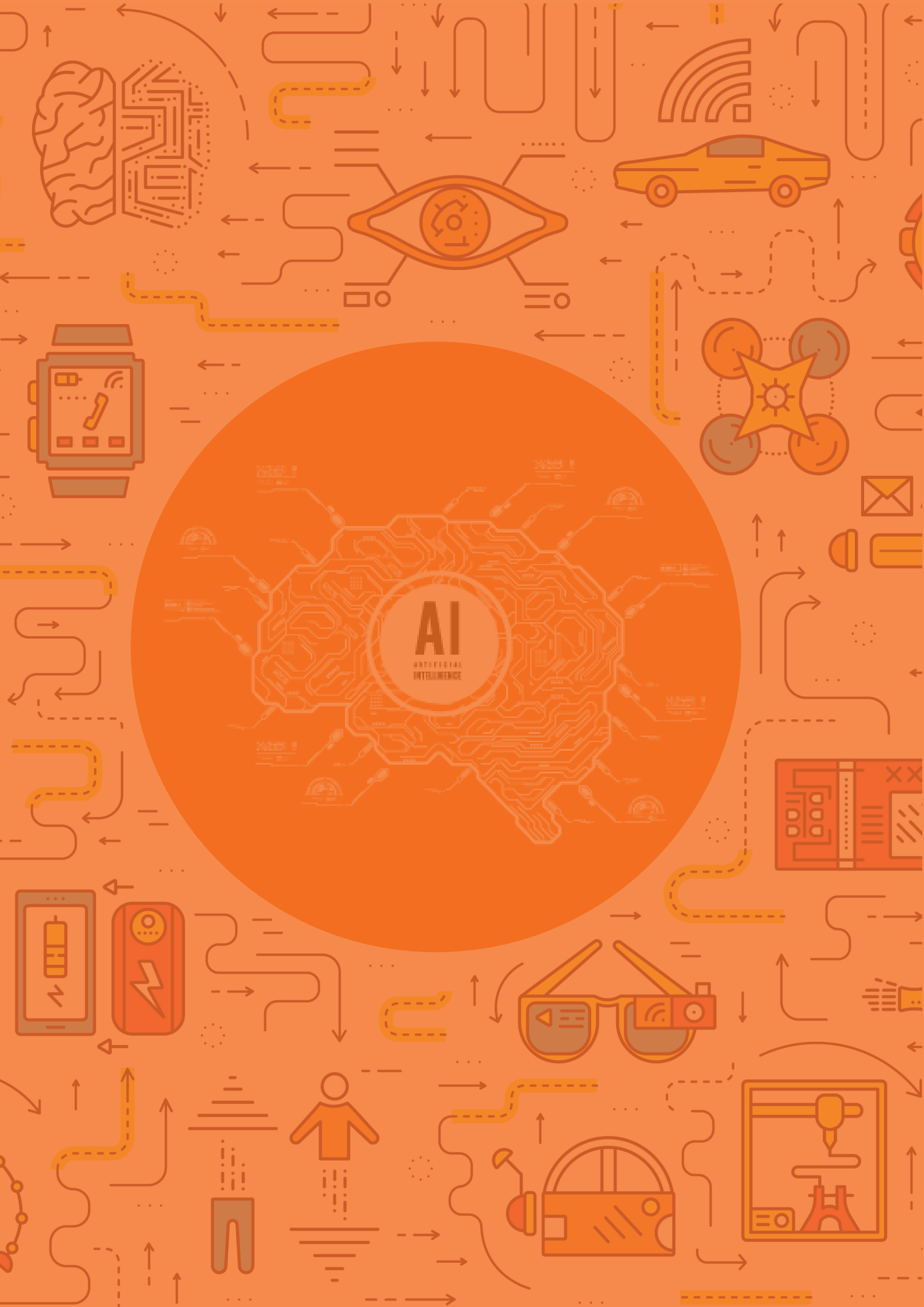
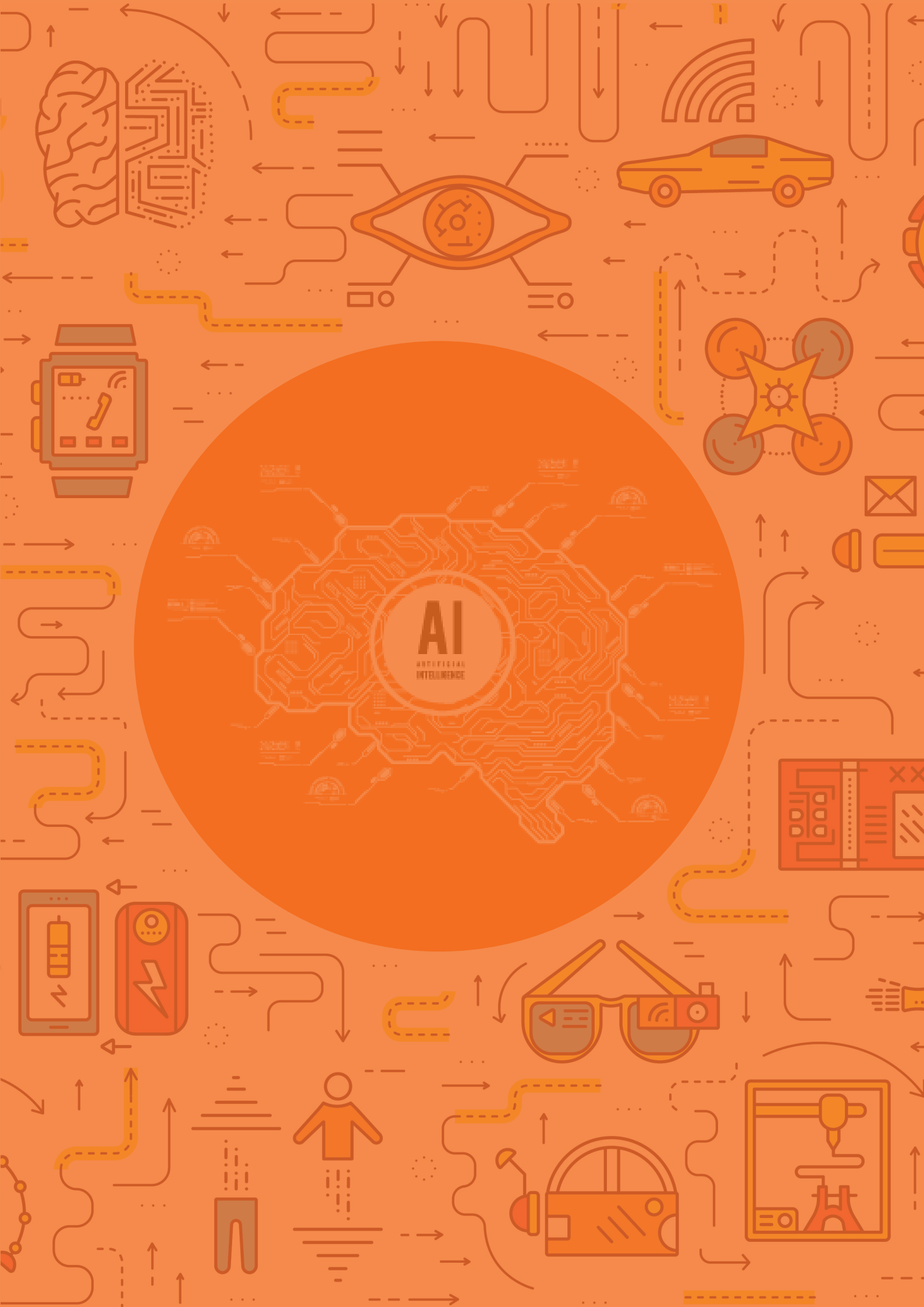
NITI Aayog | ATAL INNOVATION MISSION



LET'S LEARN ARTIFICIAL INTELLIGENCE

AUGUST 2020 | STEP-UP MODULE


AI


AI
ARTIFICIAL
INTELLIGENCE

डॉ. राजीव कुमार
उपाध्यक्ष
DR. RAJIV KUMAR
VICE CHAIRMAN
Phones: 23096677, 23096688
Fax : 23096699
E-mail : vch-niti@gov.in

सत्यमेव जयते

भारत सरकार
नीति आयोग, संसद मार्ग
नई दिल्ली–110 001
Government of India
NATIONAL INSTITUTION FOR TRANSFORMING INDIA
NITI Aayog, Parliament Street
New Delhi-110 001

DR. RAJIV KUMAR
VICE CHAIRMAN,
NITI AAYOG

10th August, 2020

Foreword

India is a country known for ingenious innovations. The sheer heterogeneity and diversity of Indian society makes it a fertile ground for ideas and creativity. Essentially, innovation is a mindset and a vision. An innovative mindset empowers an individual to search for and deliver better solutions. It also unlocks the potential and speeds up the economic growth and development of a country.

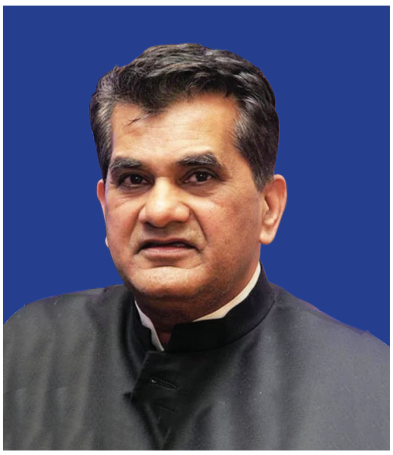
The Atal Innovation Mission (AIM), NITI Aayog is an intervention by the Govt. of India to foster the spirit of innovation and entrepreneurship in the country. The Atal Tinkering Labs (ATL) program is an initiative at the school level transforming India's education system by introducing the students with the latest technologies at a very young age to tinker and come up with innovative solutions, bringing a change in the society and to articulate a vision for creativity and technology.

I appreciate the efforts of AIM Team for leading the innovation movement in the country and introducing the ATL Artificial Intelligence modules for school students. The ATL Step Up module picks up from exactly where the ATL Base Module concluded. It is a concerted effort with an objective of increasing the knowledge and skills of students to make them more employable and acquire global competencies, achieved through creation, dissemination and application of knowledge thereby contributing to the nation building process.

I also congratulate the students, teachers and mentors of the Atal Tinkering labs and it is wonderful to see them put in their best efforts and make the country proud.

(Rajiv Kumar)

स्वच्छ भारत



SHRI AMITABH KANT

CEO,
NITI AAYOG

NITI Aayog | ATAL INNOVATION MISSION

अमिताभ कांत
Amitabh Kant
मुख्य कार्यकारी अधिकारी
Chief Executive Officer

सत्यमेव जयते

भारत सरकार
नीति आयोग, संसद मार्ग,
नई दिल्ली-110 001
Government of India
NATIONAL INSTITUTION FOR TRANSFORMING INDIA
NITI Aayog, Parliament Street,
New Delhi-110001
Tel. : 23096576, 23096574 Fax : 23096575
E-mail : ceo-niti@gov.in, amitabh.kant@nic.in

Message

Education is about learning. If there is no learning, there is no education. One of the key goals of education must be to motivate students who are creative and innovative, encouraging them to approach things in new ways. The economic growth and future of a country depends on its ability to nurture the innovative thinking capability and creativity skills in students. The Atal Tinkering Labs, are bringing a massive disruption in India's education system.

Atal Tinkering Lab established by Atal Innovation Mission, is Government of India's flagship initiative to introduce our youth to the future skills through self-learning, enabling students to develop a creative mindset and transmute the future of the country. With the Atal Innovation Mission, innovation and entrepreneurship have become an integral part of our national mission, and young students are being introduced to the world of technology and innovation.

The rapid advancement of technology like Artificial Intelligence (AI) has impacted all industries, including education. If we hope to utilize AI's full potential for everyone, our focus should be on exposing the next generation to AI early on, and utilizing Artificial Intelligence in the classroom. It also means that the gap between supply and demand of people with AI skills is growing, with better solutions and opportunities for all.

Artificial Intelligence refers to the ability of machines to perform human-like cognitive tasks such as decision making and problem solving. The advent of AI revolution in India, has a potential for inclusive evolution and social development, enabling the country to be a leader and at the forefront of the global innovation activities.

India has always promoted the policy of Education for all, the introduction of the ATL AI Modules is a step in this direction to empower the young students with the latest technologies. The ATL AI Step up module is an extension of the ATL AI Base Module with an objective to further strengthen the skills of the students and challenge them to come up with effective scalable solutions and enhanced implementation.

I congratulate the AIM team who are incessantly facilitating and supporting the students, inspiring and motivating them to embrace the spirit of innovation. My best wishes to all the students, teachers and mentors for their efforts and continuous dedication in transforming the country's future.

(Amitabh Kant)

Place-New Delhi
Dated- 01 August, 2020

SHRI RAMANATHAN RAMANAN

MISSION DIRECTOR
ATAL INNOVATION MISSSION,
NITI AAYOG

The Fourth Industrial Revolution is transforming the global innovation and production systems. The future depends heavily on our individual and collective capacities to use the rapidly emerging technologies. Developing nations like India have a unique opportunity to spring to the next levels of development by investing in the right skills and innovation frameworks and preparing for the future.

Atal Innovation Mission is one such initiative set up by NITI Aayog to promote innovation and entrepreneurship across the length and breadth of the country. Making rapid strides on the innovation front, AIM is setting up Atal Tinkering Labs (ATL) in schools across the country where students are free to experiment, explore ideas and learn future technologies. It encourages out-of-the box thinking and creating scientific temper among the young minds of the country.

It is crucial to understand the importance of education in the country's growth strategy. Introducing students to the latest technologies such as Artificial Intelligence (AI) at an early age sparks optimism and improves their confidence. In order to optimally leverage the full potential of AI, all stakeholders in the innovation ecosystem including students, must work in tandem to build innovative solutions with a long-term perspective that can benefit all.

Launching the ATL Artificial Intelligence modules is and attempt to optimally deploy the technology and has been strategically created as a potent mix of an interactive hands-on resource to encourage a meaningful constructive learning experience and establish a qualitative experience. The ATL AI Step Up module will help them in inculcating a reasoned thinking approach and solving societal problems using technological skills accelerating the sustainable development of the nation.

India follows a unique approach for education, it follows the policy of inclusive education which means access to knowledge for all rather than just introducing and deploying latest technologies. We at AIM believe in supporting and encouraging young students, imparting them with latest skills for the advancement of knowledge and redefining India as a global leader in innovation.

NASSCOM®

DEBJANI GHOSH

President, NASSCOM

The “Techade” or the twenties decade is going to be about advanced technology pervading every aspect of our lives. At NASSCOM, we had even said this before the pandemic struck. And now, with the extreme push in going contactless, the pace of digital adoption has accelerated many times over. The combinatorial power of DeepTech, particularly AI, can transform India and we must ensure that talent continues to stay relevant and at the cutting-edge. A recent study conducted by NITI Ayog & NASSCOM estimates that a unit increase in AI intensity, measured as a ratio of AI to total sales can return 67.25 billion USD or 2.5% of the Indian GDP in the immediate term. Therefore, we are only limited by our imagination of what can be achieved in the long-term.

The Base Module of ATL Artificial Intelligence is a perfect tool that is empowering young minds to tinker with AI while learning the basics through hands-on learning. It has been designed in an easy-paced manner that can be absorbed by students to build a strong foundation. It aims to provide practical experience in a variety of sub-fields while addressing the diverse needs and interests. The subject of AI opens up innumerable options and students must know what they are to make the appropriate choices based on their objectives. The focus on math, programming, and problem-solving is done in a way, keeping in mind the overall construct of the academic curriculum in AI.

And now we have the Step-up Module. It builds on the success of the Base Module. One of the projects included has been done by a student and it is their way of continuing to engage after completing the Base Module. It also gives ample assurance that this approach is effective, and it continues to inspire students in their learning journey by challenging them to tackle bigger blocks along the way.

Any course in AI has to be dynamic, practically-oriented, draw heavily from real-world examples, and be relevant to the industry. I am delighted to say, that these angles have been suitably addressed. Here, I’d like to thank our industry members for their unwavering support towards content creation. While working with the Atal Innovation Mission, they have ensured that theory and practice have been suitably bridged.

We have entered a world where programming will be a mandatory skill irrespective of the stream of higher education one pursues. When I say this, it’s not because I am in the tech sector. It’s the reality and we need to give the right kind of exposure to young minds (in AI) so that as adults they can contribute meaningfully towards nation-building.

Here’s wishing good luck to the creators of the Step-up Module and the students.

futureskills
A NASSCOM initiative

AMIT AGGARWAL

CEO, SSC NASSCOM & CO-ARCHITECT
NASSCOM FUTURESKILLS

A few years from now, we will witness AI becoming larger than the internet revolution.

AI has an unprecedented potential of disruption, the impact of which is far beyond our imagination. It is critical to introduce the world of new technologies to the bright & young minds of India, to fasten the pace of the AI revolution that we are yet to witness.

It is with this ideology, NASSCOM together with NITI Aayog rolled out the AI Base Module, a progressive learning resource, earlier this year to ignite and develop the skills and interests of students in the field of AI and ML.

This Step-Up Module, is an advanced version built upon the success of the base module. Much to our delight, one of the projects included in the module has been designed by a student who completed the base module. It is indeed such a pleasure to see one of the learners get inspired to innovate.

The practical experience, which is an integral part of the program seeks to motivate students to explore various sub-fields or interests that will encourage them to cultivate and pursue disciplines that relate to skills in AI and ML. The module is designed to equip our next generation of talent with basic AI skills and to make data fluency an integral part of their learning journey.

I'd like to take a moment to thank all our Industry members for their whole hearted participation and contribution to the program. At the Tech Sector Skills Council NASSCOM, we have always emphasized on the need for giving real-world and industry-relevant experience to the students. NASSCOM's FutureSkills® initiative believes that building a strong foundation on emerging technologies like AI, is critical for embarking upon a journey towards making India a hub of digital talent.

I urge all students to make the most of this module and think of bold ideas and innovative applications of AI. I am hopeful that this module will groom our students to play an instrumental role in the next wave of innovations and breakthroughs driven through AI in India.

Preface

The collaboration between Atal Innovation Mission, NITI Aayog and NASSCOM is a special one. It is one of the most prolific examples of the collaborative spirit that the Industry, academia and the Government of India share. At its core, we believe that to create a country of highly-skilled workforce and innovators, industry must step-up and introduce cutting-edge knowledge at a young age through the school education ecosystem.

Ronak Jogeshwar
(Innovation Lead - Atal Tinkering Labs, Atal Innovation Mission, NITI Aayog)

Yudhisther Yadav
(Head - Strategic Initiatives, NASSCOM FutureSkills)

Towards this goal, Atal Innovation Mission is setting up state-of-the-art Atal Tinkering Labs (ATls) in districts across India. The ATL ecosystem enables students to become problem solvers by giving access to 21st century technology and skills within their or nearby schools. These makerspaces house various tools such as microprocessors, programmable boards, IoT components, sensors and actuators, rapid prototyping tools, 3D printers and various learning kits.

Artificial Intelligence is one of the key technologies of the 4th Industrial revolution which will impact Nation's sustainable development goals. A firm action to ensure our young minds are equipped to use AI to maximize the benefit is the need of the hour. Training in AI at this young age will result in innovators, researchers, entrepreneurs and the workforce that will be able to successfully create, invent, and improve."

The ATL AI Step-Up module is a continuum to the ATL AI Base Module (Link : https://aim.gov.in/Lets_learn_AI_Base_Module.pdf) where the students who have grown an appetite towards the technology can become experts of the Artificial Intelligence Technology.

This module is a combination of robust exercises to engage with participants and make learning AI a 'Fun' exercise. The hands-on and activity-based approach helps the students hone their skills across various aspects of the AI technology.

Our special acknowledgement to the teams at Atal Innovation Mission-NITI Aayog, NASSCOM-FutureSkills, members of our Work Group and all the individuals who played a key role in development of this module. Our special thanks to the ATL schools, students, teachers, Mentors of Change, and parents who have put their faith in us in this beautiful journey.

Our special gratitude to Mr. S Manish Singh and his team at PROGILENCE Capability Development and Ms. Naba Suroor, Atal Innovation Mission, NITI Aayog for working relentlessly to curate and scribe the AI Base and Step-Up Modules.

We believe that this module shall pave the way for India to become a technology powerhouse of the world and produce a legion of skilled technocrats. So, let's celebrate the spirit of curiosity as the key to innovation and entrepreneurship.

Happy Tinkering ☺

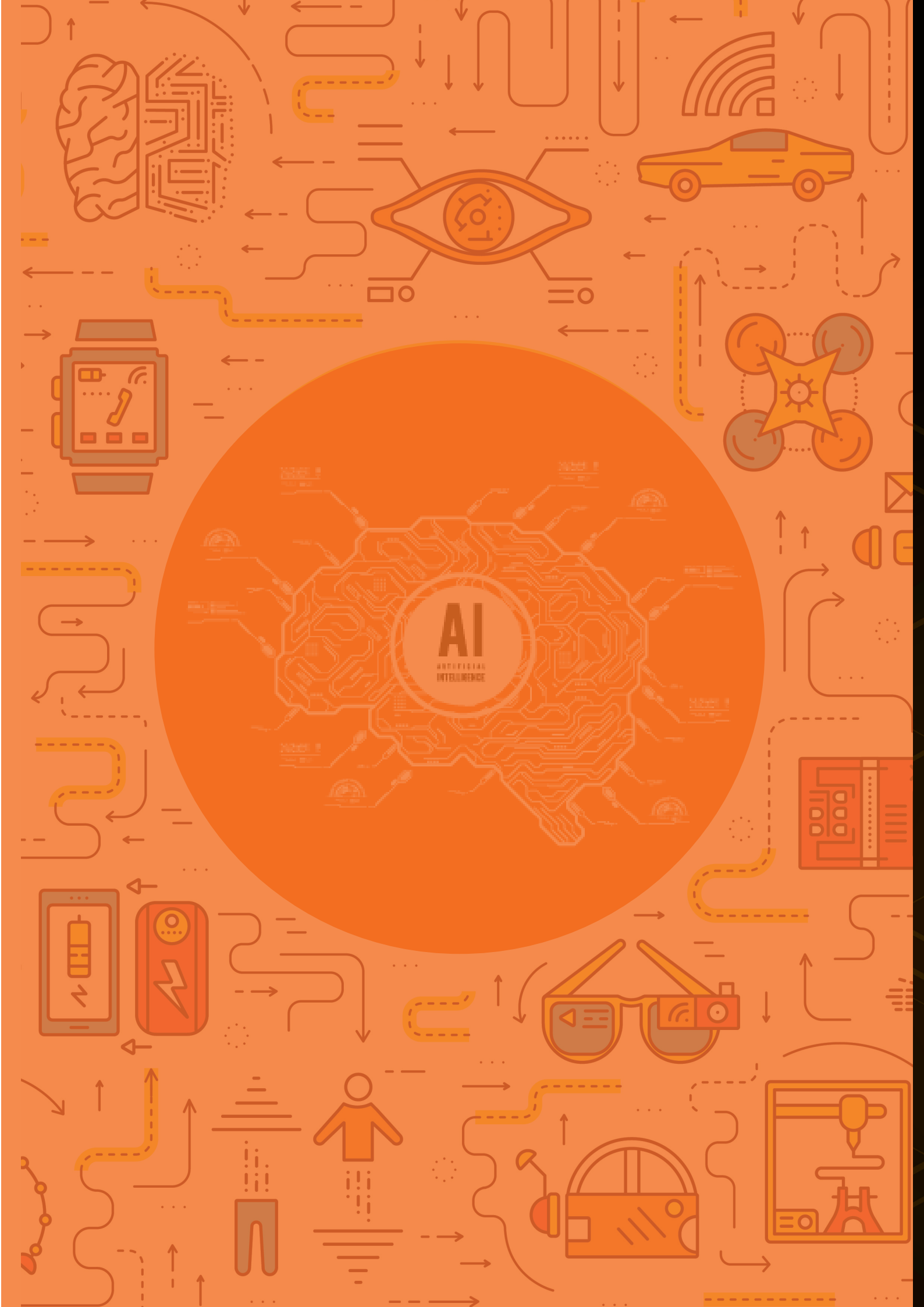

AI
ARTIFICIAL
INTELLIGENCE

TABLE OF CONTENTS

Disclaimer: The content in this handbook has been curated from various online sources. We do not intend to infringe on any copyrights.

Note: This document contains several weblinks and associated QR codes. The relevant documents and video can be accessed either by clicking the weblink or by scanning the QR code using QR code scanner application, which can be downloaded on any smartphone via the application store. Please note that there are third party links in this module and Atal Innovation Mission or NITI Aayog or NASSCOM does not endorse any person(s) or organisations mentioned on or related to these links.


AI
ARTIFICIAL
INTELLIGENCE

UNIT 1

RECONNECTING WITH ARTIFICIAL INTELLIGENCE

Key Learning Outcomes

At the end of this module, you will be able to:

a) Recall the key learnings of the base module of 'Let's learn Artificial Intelligence'
b) State various areas where AI is being applied
c) List the various technologies and tools used in AI
d) State the key steps involved in building AI
e) List the key skills required to build AI solutions
f) Explain how graphical programming works and concepts like loops and conditional statements
g) Use a pre-trained model to perform a simple project using graphical programming

Topic Covered

Base Module Recap | AI Applications | Process of Building an AI | Skills to Build AI Solutions | Project: Graphical Programming and Air Draw

1.1 BASE MODULE RECAP

Welcome back to the fascinating and ever-evolving world of Artificial Intelligence. Most of you would have been through the Base Module for 'Let's Learn AI', by now. This 'Step-Up' module seeks to take you further in your journey to learn how to use and build Artificial Intelligence.

We hope you remember and have retained what we learned in the Base module. Let us have a fun QUIZ to help you recall a few key AI concepts that we have learnt so far.

1.1.1 AI BASE MODULE QUIZ

1. Search for seven things that pertain to what AI machines could understand or do from the following word search puzzle. The words may be given horizontally, vertically or diagonally.

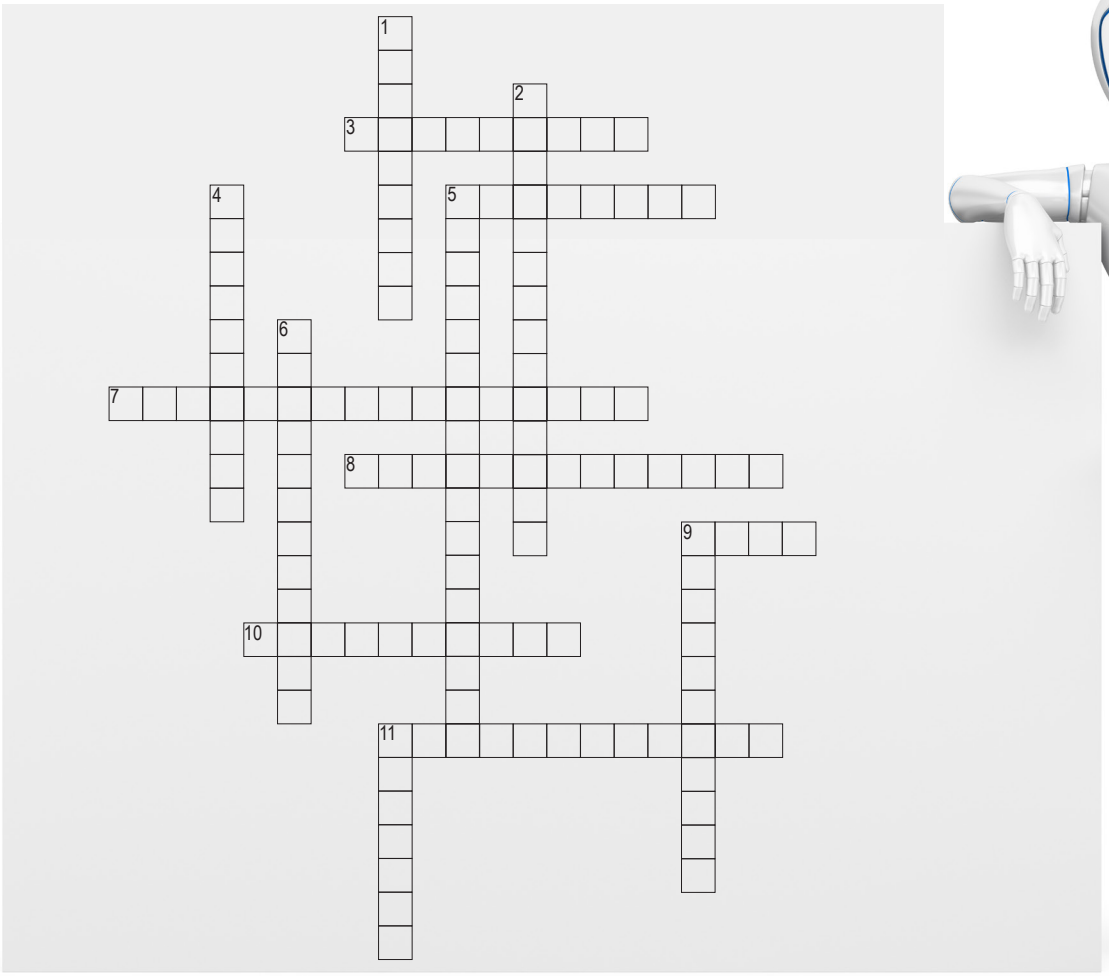
2. Match the terms related to Artificial Intelligence given in Column 'A' with the explanations of the term given in Column 'B'.

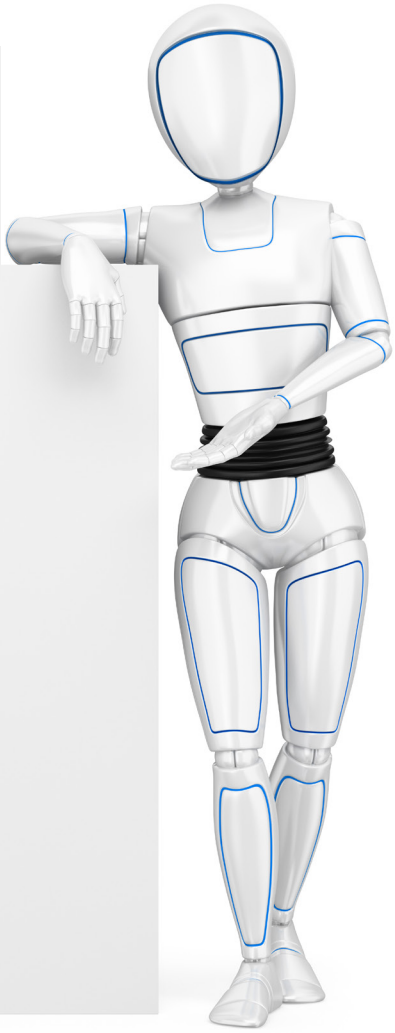
Column A: TERMS	Column B: EXPLANATIONS
a. Supervised Learning	i) AI Machines can be made to respond to naturally spoken and written language.
b. Unsupervised Learning	ii) AI Machines learn by interacting with the environment and receiving rewards for getting it right.
c. Reinforcement Learning	iii) AI Machines figure out how the various inputs relate to each other and then bunch similar inputs together.
d. Natural Language Processing	iv) AI Machines can be made to recognize faces, their gender, age group, ethnicity, and emotions.
e. Facial Recognition	v) AI Machines are fed the features and their corresponding labels in a process called training.
f. Computer Vision	vi) AI Machines are developing to differentiate between images of various objects.

3. Fill the numbered boxes with the correct answers, based on the clues given

CLUES	ANSWERS		
The three basic stages of Machine Learning	i)	ii)	iii)
Three branches of Mathematics that are used in Machine Learning	i)	ii)	iii)
Three types of Qualitative Data	i)	ii)	iii)

4. Complete the crossword given below by using the clues provided on the next page. The clues are given for the words that can be read 'Across' and 'Down'.





CLUES:

ACROSS:	DOWN:
3: It is the process of finding out the relationships between different parts of the data and the patterns that can be recognized in it.	1: It is a set of step-by-step instructions to solve a problem created along with rules that decide which path to take when a decision is faced.
5: This data typically involves integers and is a count that cannot be made more precise.	2: It is a type of complex Algorithm used in Supervised Learning where rules are defined for each training set on the basis of which the objects in the training set are classified.
7: It is a type of Algorithm where we look at each piece of data, one by one, and don't stop until we find what we are looking for.	4: It is a type of data that can take any value within a range, also called Analog data.
8: It is the selection of specific choices of data out of the total data available in a column, in a spreadsheet.	5: It is a method to represent information and data graphically by the use of images like charts, graphs, and maps
9: It is a collection of facts that can be in the form of numbers, words, symbols or even pictures.	6: It is a number system that only uses two digits 0 and 1. Computers process information using this system.
10: It is a mathematical aid that involves gathering, reviewing, analyzing the data to draw important conclusions as per the application.	9: It is the process of arranging the data into a meaningful order so that it can be analyzed or processed easily.
11: It is a type of Algorithm used to find a specific value in an ordered list, by dividing the list into two parts, each time it searches.	11: It refers to a collection of large and complex datasets that may be complex due to its volume or its variety.

5. Fill-in the blanks for the paragraph given below, using the words that are provided in the adjacent box.

Algorithms are _____ instructions to solve a _____, along with _____ that decide which path to take when a _____ is to be made. When trying to solve a problem, we may find more than one _____. A different _____, can be built from each solution. _____ reasoning determines if algorithms will work by _____ what happens when the algorithm's steps and the rules they consist of, are followed.
Algorithm can then be represented in the form of a _____.

solution	decision	flowchart
problem	step-by-step	rules
predicting	algorithm	logical

We hope you enjoyed the QUIZ.

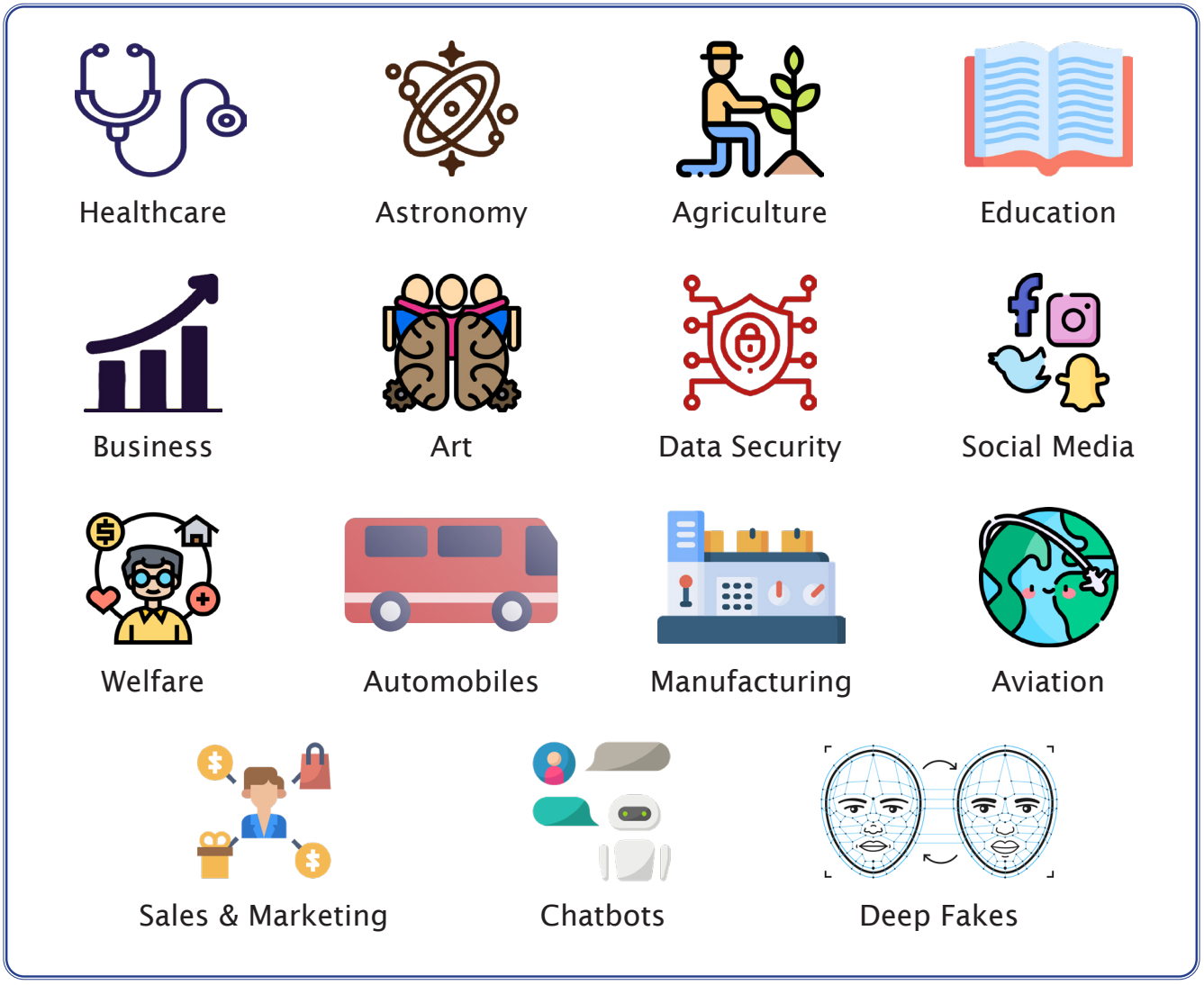
Now you will recall that in the Base module we looked at various aspects of AI, experienced various demo models, initiated programming for AI with an introduction to Scratch and Python Programming. We introduced ourselves to the three different types of Machine learning (Supervised, Unsupervised and Reinforcement Learning). We looked at algorithms including sorting and classification algorithms, We also looked at the importance of algebra, probability and statistics for AI and ML and finally explored some ethical dimensions related to AI use.

Now that you remember the basics, let us revisit how AI is being used in various fields. While you explore the current applications, think about areas where you would like to apply AI.

1.2 AI APPLICATIONS

AI has already penetrated many aspects of our world. It is getting integrated into our lives so seamlessly that often we do not even realize that we are using AI.

Here are some areas where AI is currently being used.



Explore the links under each area of application and then try and list the various aspects of AI being used to solve the problems.

Healthcare

AI has already being used widely in healthcare, whether it is for making a faster and better diagnosis, patient care, clinical research, drug development or insurance. AI is enhancing benefits for healthcare providers, patients and the society at large.

EXPLORE

https://healthcareweekly.com/trend-watch-artificial-intelligence-and-healthcare/

https://indiaai.in/article/how-icmr-leverages-ibm-s-watson-for-rapid-response-on-covid19

https://www.forbes.com/sites/bernardmarr/2017/05/16/how-ai-and-deep-learning-is-now-used-to-diagnose-cancer/#6c8af2dcc783

https://www.youtube.com/watch?time_continue=11&v=zz3rbMvjilU&feature=emb_logo

https://www.ted.com/talks/pratik_shah_how_ai_is_making_it_easier_to_diagnose_disease

Astronomy

Astronomers are using Artificial Intelligence to organize and categorize the images of galaxies taken by the satellites. It is also being used to locate objects such as gravitational lenses in space or craters on the moon.

EXPLORE

https://www.sciencedaily.com/releases/2020/05/200512151951.htm

https://astronomy.com/news/2018/03/artificial-intelligence-identifies-6000-new-craters-on-the-moon

https://phys.org/news/2017-10-artificial-intelligence-gravitational-lens-candidates.html

Video: https://www.youtube.com/watch?time_continue=135&v=ArM7ayiOxew&feature=emb_logo

Agriculture

AI technologies are being used to devise and implement ways to grow healthier crops, with more yield, control pests, monitor the soil and growing conditions, assist farmers with data, help with the workload as well improve the food supply chain.

EXPLORE

https://indiaai.in/article/when-data-surpasses-intuition-the-future-of-precision-agriculture

https://www.youtube.com/watch?v=oot55bK62pk

https://thecsrjournal.in/artificial-intelligence-in-agriculture-in-india/

https://www.youtube.com/watch?v=GSvT940uS_8

Education

AI is being used for personalized testing and development of personalized teaching materials based on each student's needs. It is also being used to take over the administrative work of teachers so that they can spend more quality time on actual teaching.

EXPLORE

https://www.theedadvocate.org/7-ways-that-artificial-intelligence-helps-students-learn/

https://indiaai.in/videos/can-ai-transform-school-education-in-india-last-week-on-indiaai-ep-4

https://cognii.com/

https://www.youtube.com/watch?v=rMYiSnp3-Lg

Entertainment

Entertainment services are using ML/AI algorithms to provide recommendations and use Natural Language Processing to generate automated subtitles. In video games, AI is being used to analyze a player's actions to predict their future moves and adapt to the player's behaviour.

EXPLORE

https://www.bbntimes.com/technology/artificial-intelligence-in-the-entertainment-industry

https://cogent.co/blog/recommendations-with-machine-learning-2/

https://emerj.com/ai-sector-overviews/ai-in-movies-entertainment-visual-media/

https://www.youtube.com/watch?v=YXHxJIGX86E

Business

Where it is to market and sell, give customers a better experience or develop cost-effective and more efficient ways of product manufacture and delivery, AI is transforming businesses.

EXPLORE

https://indiaai.in/article/how-ai-is-helping-the-world-of-hyperlocal-commerce

https://bernardmarr.com/default.asp?contentID=1546

https://blog.griddynamics.com/how-we-built-a-conversational-ai-for-ordering-flowers/

Art

AI is being used for generating art. One such method is the generative adversarial network (GAN), using two AI neural networks: a generator, which produces images, and a discriminator, which judges the paintings based on the data of existing paintings and artistic styles.

EXPLORE

https://www.youtube.com/watch?v=kRS3NZbPRW8

https://www.youtube.com/watch?v=I-EIVIHvHRM

https://www.ted.com/talks/pierre_barreau_how_ai_could_compose_a_personalized_soundtrack_to_your_life/transcript

Data Security

AI is helping CyberSecurity professionals stay ahead of hackers and other security threats. It is being used to make Data secure; for example, AEG bot, AI2 Platforms are used to determine software bug and cyber-attacks in a better way.

EXPLORE

https://emerj.com/ai-sector-overviews/artificial-intelligence-and-security-applications/

https://www.youtube.com/watch?v=MXHfD9a45DU

Social Media

AI can organize and manage massive amounts of data on social media sites. It can analyze lots of data to identify the latest trends, hashtag, and requirements of different users.

EXPLORE

https://medium.com/humansforai/the-impact-of-artificial-intelligence-on-social-media-579345b6f751

https://www.youtube.com/watch?v=udECXLtsBbM

Chatbots

AI-powered Chatbots are software that can simulate a conversation (or a chat) with a user in natural language. This may be done through messaging applications, websites, mobile apps or through the telephone. This helps businesses save time and costs.

EXPLORE

https://www.ibm.com/cloud/learn/chatbots-explained

https://discover.bot/bot-talk/beginners-guide-bots/how-chatbots-work/

Welfare

AI is being put to use for a variety of social welfare works, including health, environmental sustainability, and public welfare.

EXPLORE

https://cra.org/ccc/artificial-intelligence-social-good-speakers/

https://www.aiforpeople.org/projects/

https://www.ted.com/talks/ali_kashani_a_friendly_autonomous_robot_that_delivers_your_food

Automobiles

AI is, of course, being used in self-driving cars. Apart from that, it is also being used for driver safety and assistance, predictive car maintenance, manufacturing process, to name a few.

EXPLORE

https://www.techradar.com/news/self-driving-cars

https://emerj.com/ai-sector-overviews/ai-in-the-automotive-industry-an-analysis-of-the-space/

https://www.ted.com/talks/sajan_saini_how_do_self_driving_cars_see

Aviation

AI is being used as surrogate operators for flight crafts, in combat and training simulators, to support systems for decision making and management of missions. The AI is then used for post-processing of the simulator data and offering summaries.

EXPLORE

https://www.youtube.com/watch?time_continue=9&v=D8NlYPtPgwA&feature=emb_logo

Deepfakes

These are programs that animate the face of a target person, transposing the facial expressions of another person on it. The technology includes animating and synching the lips of people. Some famous past examples include Deepfakes of Barack Obama and Vladimir Putin.

EXPLORE

https://www.youtube.com/watch?v=FCCisdXNqHU&feature=emb_logo

https://www.ted.com/talks/danielle_citron_how_deepfakes_undermine_truth_and_threaten_democracy

ACTIVITY 1

Create an infographic showing how AI might be used in the future. Choose any of the areas mentioned above.

Include the following information about your chosen area on the infographic:

- Current ways AI is used in the industry
- Ways AI can be used in the future for this industry
- Current and future benefits of AI usage to this industry

Students will use at least three different sources and cite each at the bottom of the infographic.

Possible free infographic tools include the following:

https://www.easel.ly

https://www.canva.com/create/infographics/

https://venngage.com

https://spark.adobe.com/make/infographic-maker/

https://www.visme.co/make-infographics/

Word-processing tools (Microsoft Word)

ACTIVITY 2

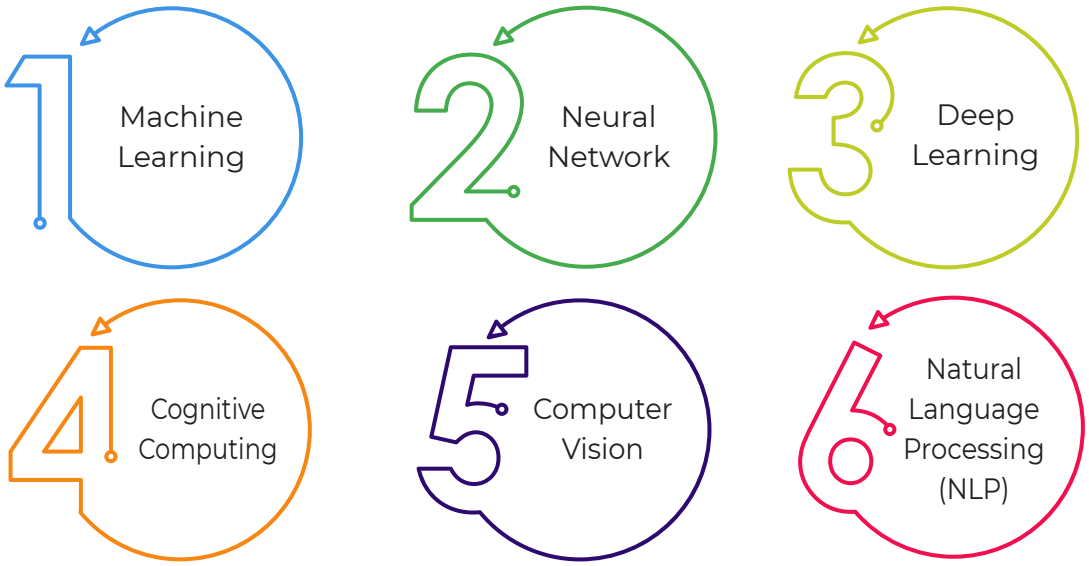
Now pick up at least three applications from the cases we have seen above and answer the following questions for each of them:

1. What was the objective of the project?
2. What AI abilities or capabilities are being used in this example?
3. What is the process you think must have been followed while creating this solution?
4. What are the benefits of using AI in this example?
5. What are the Risks and Considerations do you think that one has to keep in mind while developing and using this AI solution?
6. What are the requirements for developing an AI and categorize them into:
   a. Data **b.** Hardware **c.** Software **d.** Platforms **e.** Others

Once you have answered the questions for the cases, combine your answers for the six questions in a master list. Now compare your list with that of your friends and other course mates. Discuss the cases and answers with them, expand and refine your list.

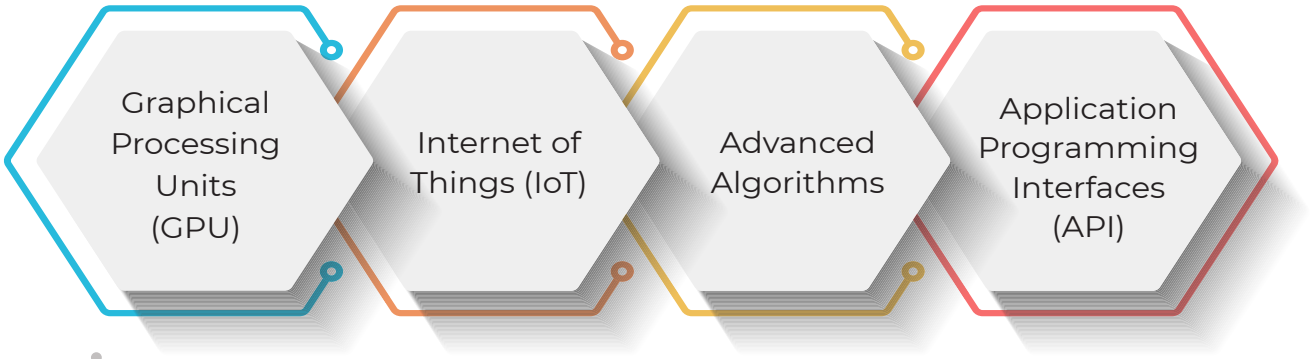
While answering the above, bear in mind that AI working involves fast and continuous processing of large amounts of data and intelligent algorithms, where the software automatically learns from patterns or features in the data.

AI is a broad field that includes several theories, methods and technologies, as well as the following major domains:



- **Machine Learning** is the part of AI that involves automating analytical model building. It uses various methods such as neural networks, statistics, operations research and physics in its attempt to find hidden insights in given data without explicitly being programmed for what conclusions to reach and the path to follow.
- **Neural Network** can be defined as a type of machine learning comprising interconnected units (like human brain neurons) used to process information using an algorithm that allows the computer to learn by incorporating new data. The network of units responds to external inputs, relaying information between each unit. The process is iterative, in that it requires multiple passes at the data to find connections and derive meaning from undefined data.
- **Deep Learning** uses massive neural networks with many layers of processing units, relying on considerable computing power along with improved training techniques to learn complex patterns in large amounts of data. Typical applications include computer vision and speech recognition.
- **Cognitive Computing** is that subfield of AI which strives for natural, convincing, human-like interaction with machines. The primary objective of using AI and cognitive computing is for a machine to simulate human behaviour and mental cognitive processes using its ability to interpret images and speech – and then respond coherently and appropriately.
- **Computer Vision** uses pattern recognition from data and deep learning neural networks for image recognition in pictures or video. As machines process, analyze and understand images, they are expected to capture images or videos in real-time and interpret their surroundings.
- **Natural Language Processing** (NLP) is the subfield of AI that aims to develop the ability of computers to analyze, understand, process and generate human language, including speech for various purposes. The subsequent stage of NLP may be seen as natural language interaction, where humans can communicate with computers using everyday language, intelligently.

Additionally, several technologies enable and support AI, such as:



- **Graphical Processing Units** are fundamental to AI because they provide substantial compute power that is required for continuous and iterative processing. Training neural networks is a data-rich activity that requires big data plus the compute power to process it and the volume of activity involved.
- **Internet of Things (IoT)** generates massive amounts of data from connected devices, most of it remains unanalyzed. IoT enabled with AI allows us to use more of it to achieve greater insight, and develop connected capabilities.
- **Advanced Algorithms** are being developed and used in innovative and efficient ways to analyze complex data, more data faster, and at different levels simultaneously. This intelligent processing is aimed to arm us with capabilities that are key to spot and predict rare and obscure events, understand complex systems and optimize unique scenarios.
- **APIs or Application Programming Interfaces** An API (Application Programming Interface) is essentially a gateway that allows the software to talk to other software while defining the rules of how that conversation needs to take place. APIs allow a programmer to borrow code in a way, where they can work with one and connect to the other software to carry out specific desired actions, without a programmer having to write that software from scratch. It is developed for various specific purposes and can help add computer vision capabilities or language processing capabilities without having to code the entire operation. This is made to simplify processes for developers.

EXPLORE

https://education.microsoft.com/en-us/course/e328f120/8

https://www.youtube.com/watch?time_continue=391&v=OVvTv9Hy91Q&feature=emb_logo

In summary, AI aims to provide a software application that can reason on input and explain the output. AI will have capabilities to provide human-like interactions with machines and offer valuable decision support for specific tasks. Still, it will not be a complete replacement for humans at least not anytime soon.

1.3 PROCESS OF BUILDING AN AI

The core process of developing an AI using machine learning primarily remains similar for various applications. It may involve minor variations based on the context, purpose and approach. Let us look at the process steps:

1) Identify the problem
2) Acquire relevant data
3) Prepare the data
4) Analyze the data using a model
   - Choose the algorithms
   - Train the model
   - Evaluate the model
5) Choose a particular programming language
6) Run on a selected platform

Here are some videos that will give you an explanation on the same. You will see that while different people may name and categorize the operations involved differently, it follows the same core process.

EXPLORE

https://www.youtube.com/watch?time_continue=295&v=SGg5i0kQx-8&feature=emb_logo

https://www.youtube.com/watch?time_continue=152&v=nKW8Ndu7Mjw&feature=emb_logo

1.4 SKILLS TO BUILD AI SOLUTIONS

Now that we have seen the above applications and thought through them, let us think about what are the various skills required to create an AI.

1. Logic and Reasoning for Problem-solving
2. Computer Programming
3. Maths for ML and AI
4. Machine Learning including Neural Networks, Deep Learning, Computer Vision and Natural Language Processing
5. Data Science
6. Algorithms
7. Databases
8. Ethics
9. Cloud Computing
10. Usage of various Platforms, APIs, IDEs used in AI and ML

So, if you are eager to start building AI, you will have to gain a deeper understanding of various fields of AI and develop skills within these areas. Let us now continue our journey to acquire some of these skills.

PROJECT: GRAPHICAL PROGRAMMING AND AIR DRAW

In the base module, you were also introduced to graphical programming using Scratch. Let us revisit some of its concepts before we begin our first project.

GRAPHICAL PROGRAMMING AND ITS ELEMENTS

Graphical programming is one of the easiest ways to begin your programming journey. Due to its intuitive interface and programming blocks, it becomes very easy to make games, animations, program robots, and make AI/ML projects. In this module, we will use graphical programming for doing AI-based projects.

Before we jump on to the project, let's first look at the various elements and understand how it works:

Graphical programming (or visual programming) is a method of programming in which the program(s) are written using graphical elements known as blocks. The blocks are like pieces of a jigsaw puzzle that you must put together to solve the puzzle. This approach makes learning to program easy, interactive, and super fun!

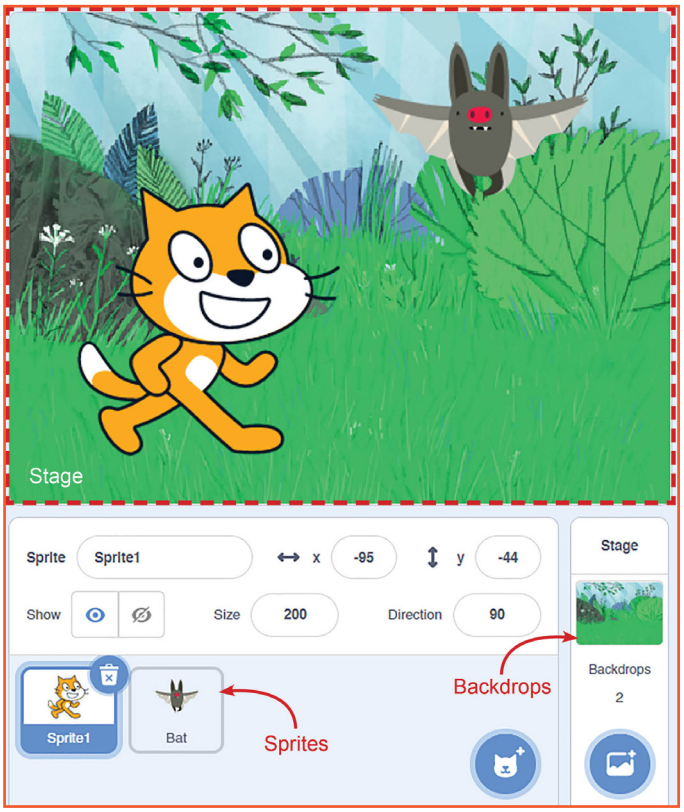
You can find more details about the basics of graphical programming at ATL Gaming Module:

https://atlgamingmodule.in/courseInfo/1

Here's a brief overview of any graphical programming platform:

STAGE

The stage is a background or a backdrop for your projects (the programs you'll create). The Stage is where the Sprite moves, draws, and interacts with other sprites and/or hardware.



SPRITE

A Sprite is an object or a character that performs different actions in the projects. It understands and obeys the instructions that you'll give them in your projects. It can move and can be moved to any place in the stage (you can try by clicking on them and moving them around).

BACKDROP

It is similar to Sprite, but with some limitations like it cannot move. It has its own set of scripts, images, and sounds.

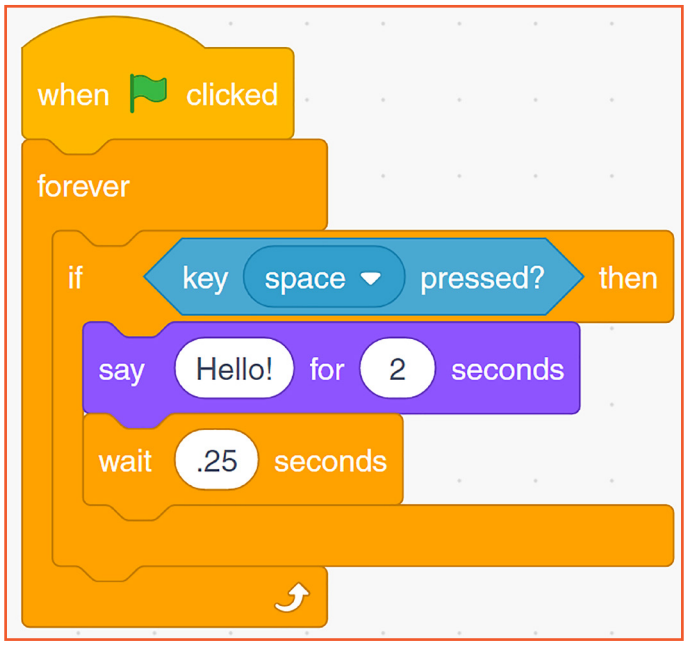
BLOCKS

Blocks are like jigsaw puzzle pieces that fit into each other. They are predefined commands used to create programs by simply dragging and dropping them below one another in the scripting area.

SCRIPT

A script is a program or a code in Scratch based graphical programming lingo. It is a set of 'blocks' that are arranged below one another in a specific order to perform a task or a series of tasks.

The scripting area is where you will create your scripts.



GRAPHICAL PROGRAMMING PLATFORMS WITH AI

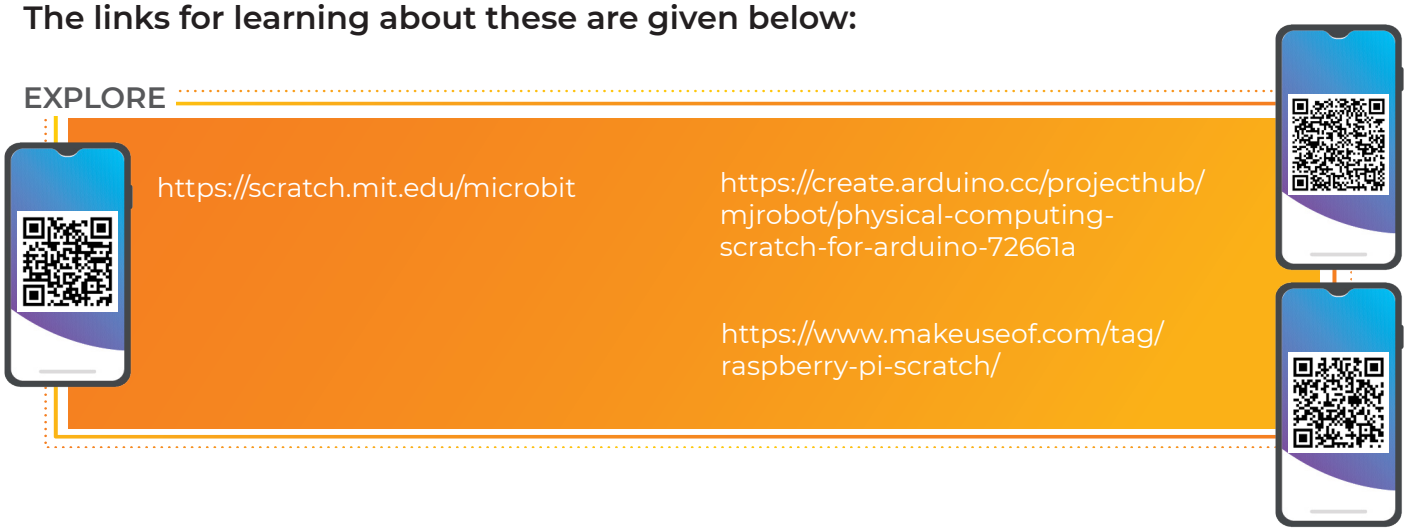
Multiple platforms support graphical programming for implementing artificial intelligence and machine learning. Following are some of the platforms, which you can use for AI projects and activities:

1. Aiconnect - http://aiconnect.cc
2. Cognimates - http://cognimates.me
3. Machine Learning for Kids - https://machinelearningforkids.co.uk/
4. PictoBlox - http://bit.ly/PictoBlox
5. Stretch 3 - https://stretch3.github.io/

You can learn about how to set up the platform, the installation process, and creating the account from the links provided above.

Scratch can also be used for physical computing and can be linked to computers such as Microbit, Arduino and Raspberry Pi.

The links for learning about these are given below:



Besides, we would like you to explore another similar graphical programming language called Snap.

ABOUT SNAP

Snap! (formerly BYOB) is a visual, drag-and-drop programming language. It is an extended reimplementation of Scratch (a project of the Lifelong Kindergarten Group at the MIT Media Lab) that allows you to Build Your Own Blocks. It also features first

class lists, first class procedures, and first class continuations. These added capabilities make it suitable for a serious introduction to computer science for high school or college students.

EXPLORE

To know more visit

https://snap.berkeley.edu/

To experience programs in Snap for AI, visit

https://ecraft2learn.github.io/ai/

PROJECT: AIR DRAW USING ARTIFICIAL INTELLIGENCE

Now let us work on a simple AI project called Air Draw and experience this on various platforms to explore how different programs and platforms can be used to achieve the same result.

PROJECT DESCRIPTION

Control a pen using your hand gesture and draw characters on the stage in real-time.

LEARNING OUTCOMES

1. Understand how graphical programming works.
2. Understand programming concepts like loops and conditional statements.
3. How to use a pre-trained model in your project.

REQUIREMENTS

Laptop/Computer with a camera/webcam along with internet connection

COMPATIBLE PLATFORMS AND APPROACH

1. PictoBlox

While doing this project using PictoBlox, we can use Human Body Detection Extension to trace the fingertips, wrist or any other body part. We can even use the Face Detection Extension to trace nose, mouth or any other facial features.

2. Machine Learning for Kids

While doing this project using Machine Learning for Kids, we can use Pose Detection to trace the wrist or any other body part.

3. Stretch

While doing this project using Stretch 3, we need to add a Facemesh2Scratch extension.

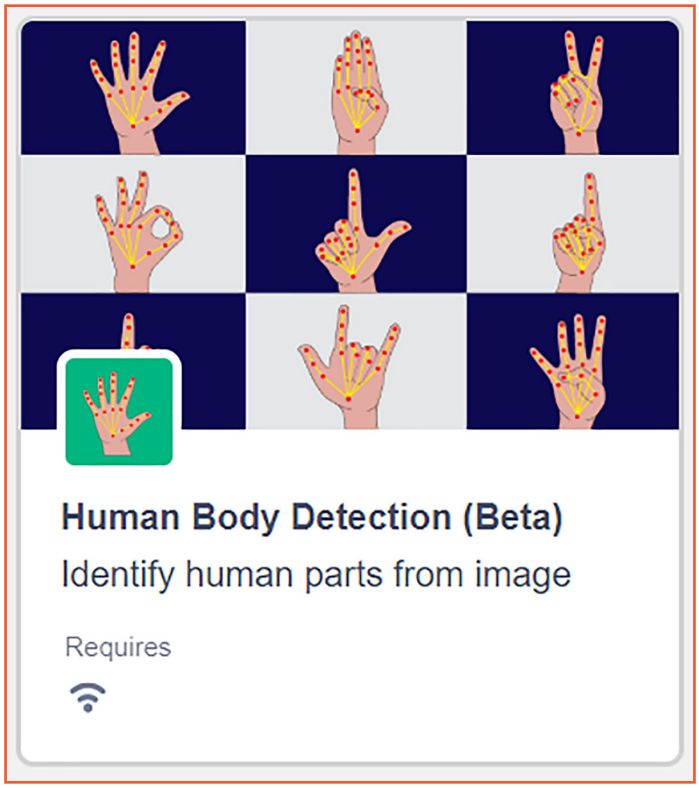
ART is in the AIR; literally. Learn to draw your favorite pictures in the air using the tip of your finger and trace it on your computer screen. Let's begin!

STEPS WITH PICTOBLOX PLATFORM

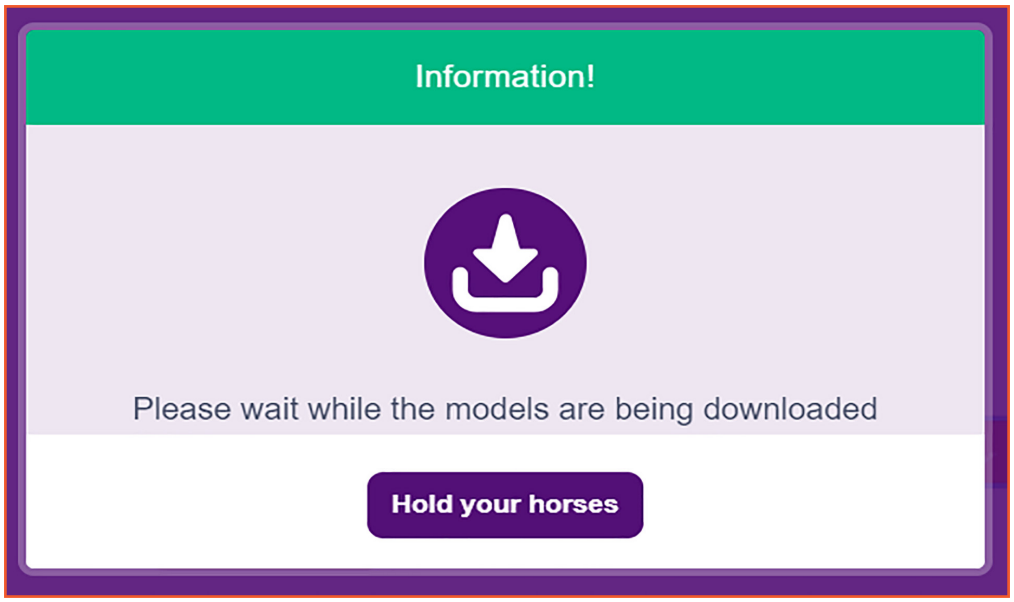
Setting up the Project

1. *Create a new project.*

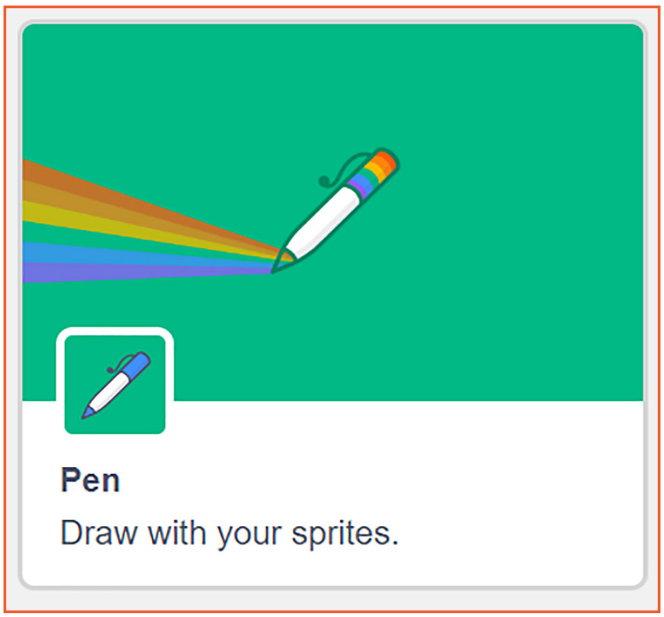
2. *Next, click on the **Add Extension** button and add **Human Body Detection** extension.*



3. *The face detection models will be downloaded, which may take a few seconds to load.*



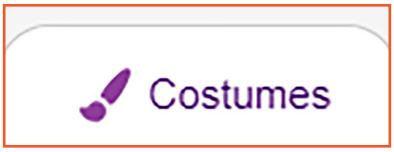
4. *Once downloaded you will get a message saying **"Model loaded successfully"**.*
5. *Next, we need to add the Pen Extension to draw on the stage.*
6. *Again, click on the **Add Extension** button and select **Pen extension**.*



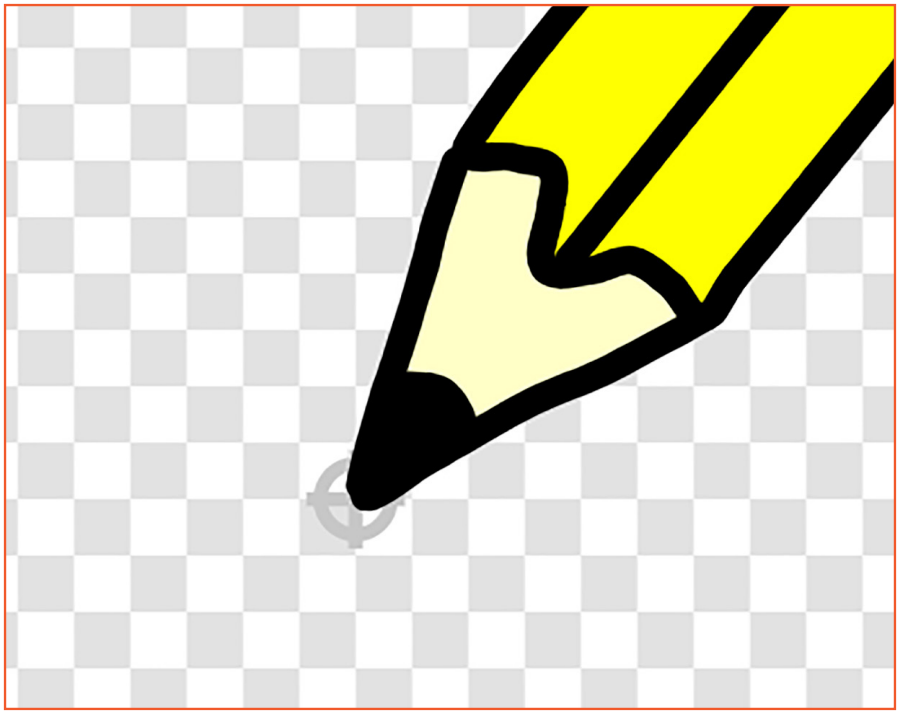
Adding the Sprites

1. *Add a new **Pencil** sprite and delete the **Tobi** Sprite.*

2. *We will need to make a few changes in the costume of the Pencil Sprite. Select the Costume Palette to edit the costume.*



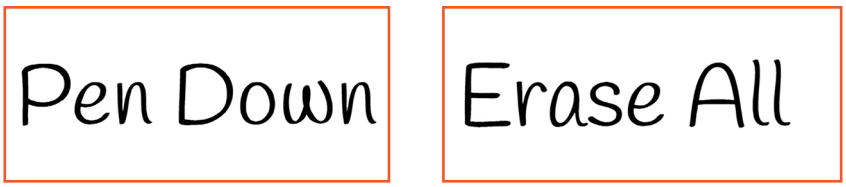
3. *Next, drag the pencil into the costume area, such that the tip of the pencil should be on the center dot.*



4. *Next, we need to make 3 additional sprites*
   a. *Pen Up*
   b. *Pen Down*
   c. *Erase All*

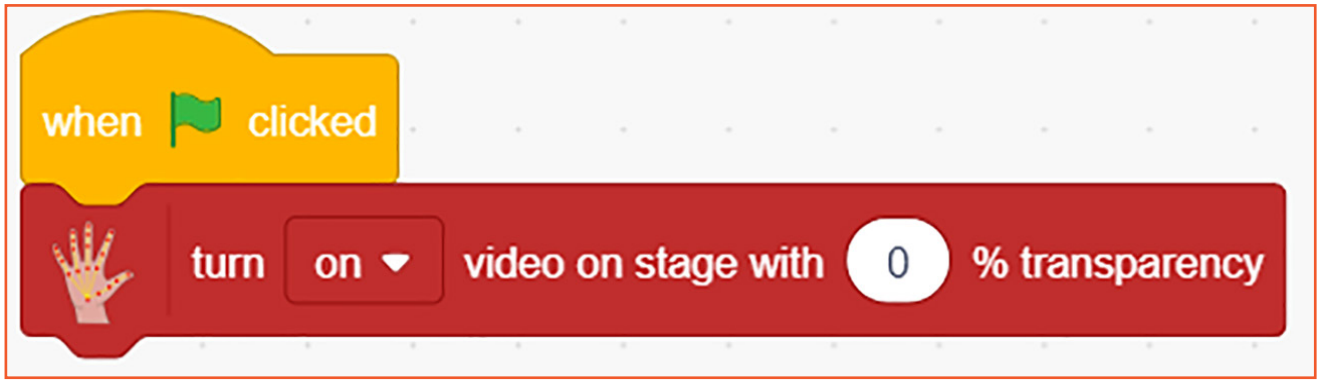
5. *Click on Choose A Sprite button and select the Paint option. Use a rectangle and a text tool to make the sprite.*

6. *Similarly, make the other two sprites.*



Adding the Up the Stage

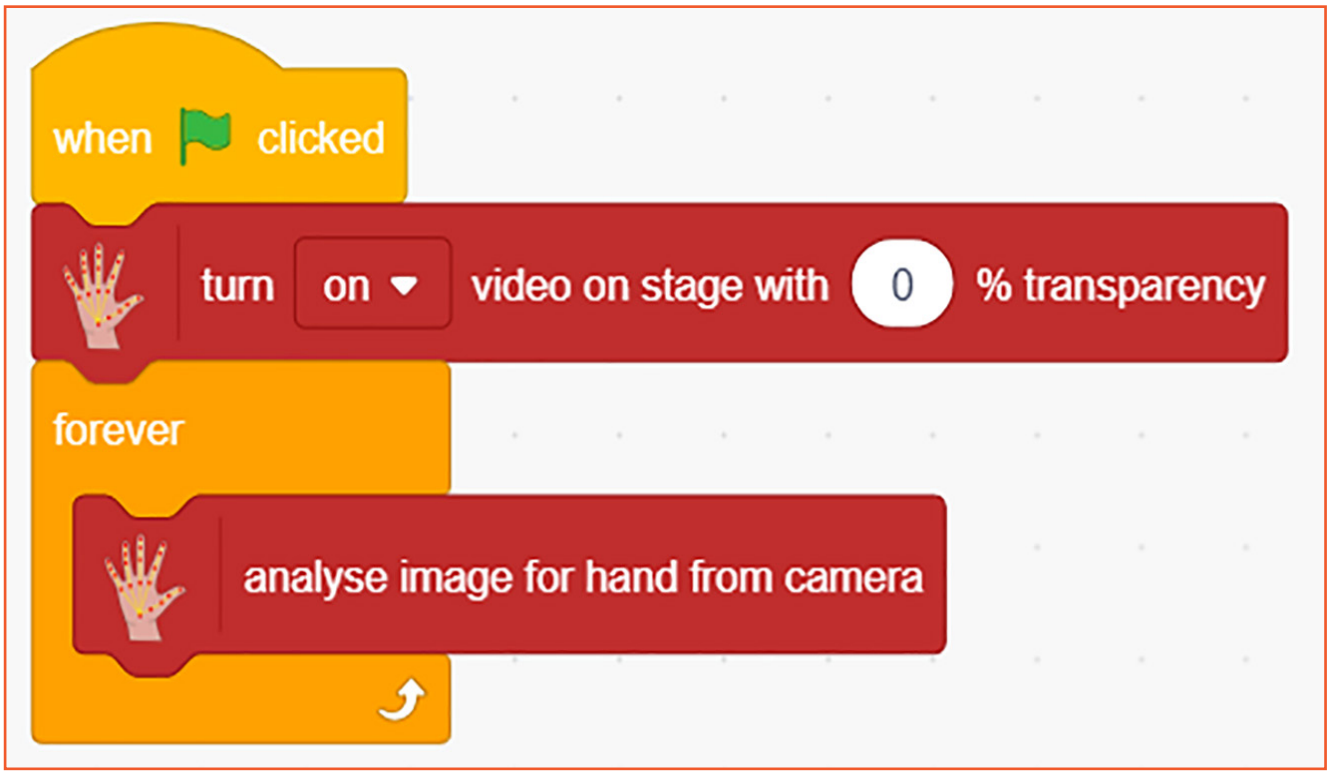
7. *Choose the **Pencil** Sprite.*
8. *Drag and drop a **when flag clicked** block into the scripting area.*
9. *Snap a **hide** block below the when flag clicked block.*
10. *Snap a **turn () video on stage with () transparency** block from the **Human Body Detection** palette. Change the camera option to on and transparency to **0%.***



11. *The stage is set. Click the green flag to get the camera feed on the stage.*

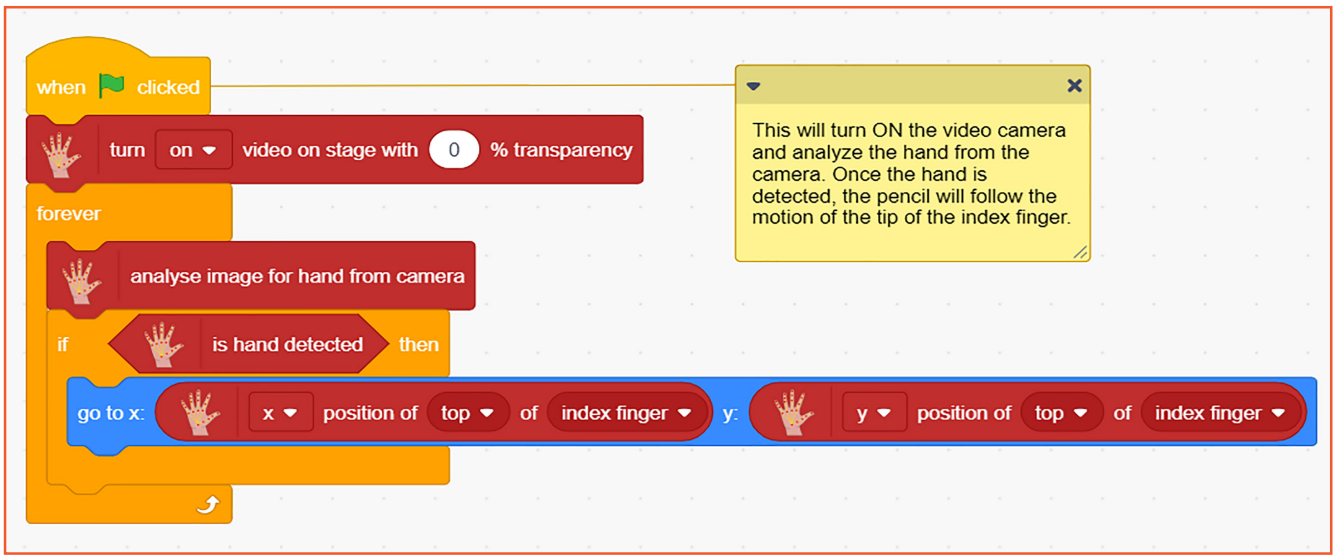
Recognize Image

1. *Add a forever block from the **Control** palette.*
2. *Snap an **analyse image from the ()** block from the **Face Detection** palette. Select the feed as the **camera.***



Detecting the Hand

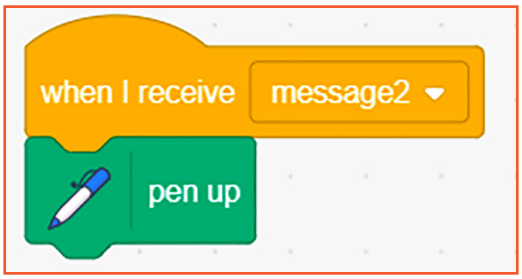
1. *Snap an **if** block below the **analyze image from the ()** block.*
2. *Place the **is hand detected** block from the Human Body Detection extension in the horizontal space of the **if** block.*
3. *Snap a **go to x: () y: ()** block from the Motion Palette into the **if** arm.*
4. *Place **() position of () of ()** block into the first space of the **go to x: () y: ()** block and select **x** from the first dropdown, **top** from the second dropdown and **index finger** from the third dropdown.*
5. *Duplicate the **() position of () of ()** block and place it into the second space. Select **y** from the first*



Activating the Buttons

1. *Click on the **Pen Up** sprite and place **when this sprite clicked** block into the scripting area.*
2. *Next, place a **broadcast ()** block from the events palette below it. Select **message1** from the drop-down.*

3. *Next, go back to the **Pencil** sprite, place the **when I receive ()** block into the scripting area. Select **message1** from the drop-down.*
4. *From the **Pen extension**, snap the **pen up** block below it.*

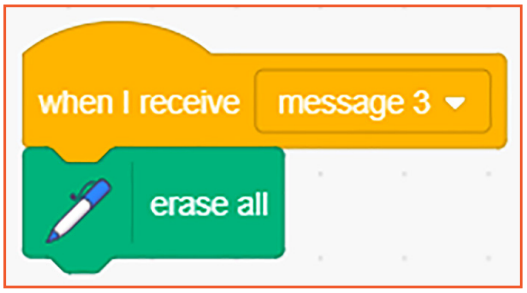


5. *Repeat the steps 1 to 4 for the other two buttons too. There will be a few changes that you'll need to make.*
6. *In the **broadcast ()** block of the **Pen Down** Sprite select New message and type **message2.***

7. *Into the **Pencil** sprite, select **message2** from one of the two **when I receive ()** block.*
8. *Remove the **pen up** block and snap the **pen down** block below it.*

9. *Into the Erase All sprite, select New message and type **message3** in the **broadcast ()** block.*

10. *Finally, go to the **Pencil** sprite, select message3 from the **when I receive ()** block.*
11. *Remove the **pen up** block and snap the **erase all** block below it.*

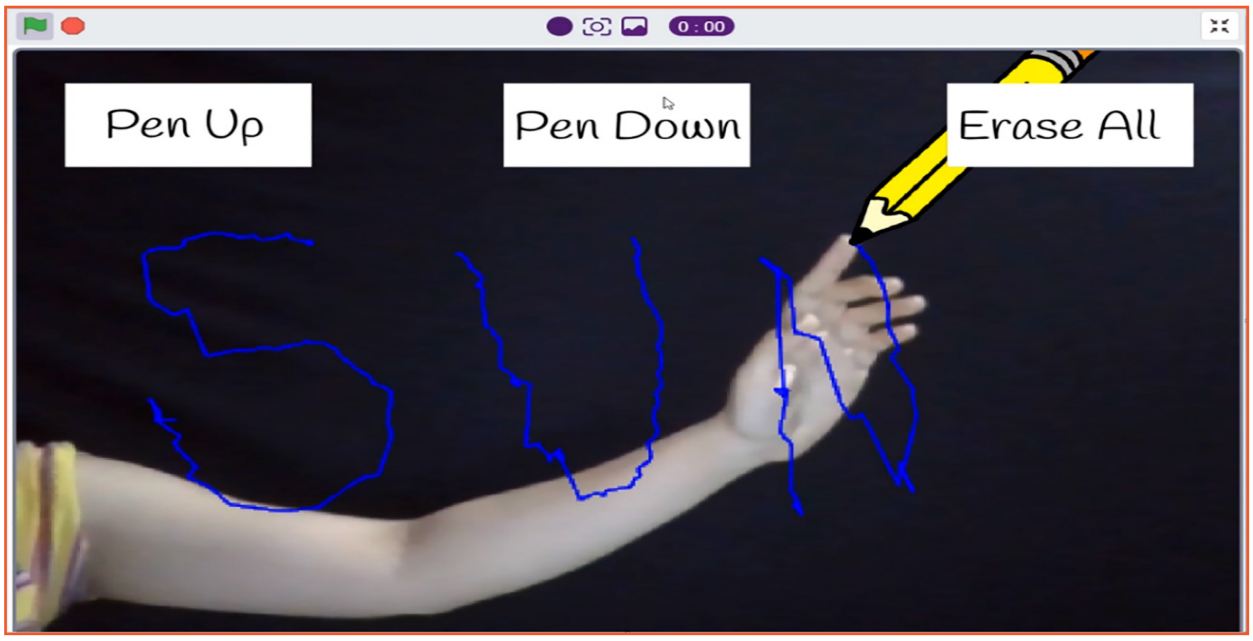


How does it work?

Now, as the script is complete, let's understand the logic behind it, i.e. how it works.

As soon as you press the green flag, the program will start executing and it turns ON the camera/webcam to search the hand. As per our script/code, the Pencil sprite will continuously follow the x and the y coordinates of the top of the index finger. Next, when you press Pen Up, the pen will be deactivated, and thus, the pencil will stop drawing/writing. On pressing Pen Down, the pencil will start writing again and the Erase All button is to clear the mess and keep your screen clean to redraw the wonderful creations.

Now you're all set to make your surroundings your canvas and try making various shapes, numbers, or alphabets. You can also record the output by pressing the Record button and can share your creations with your friends and family.



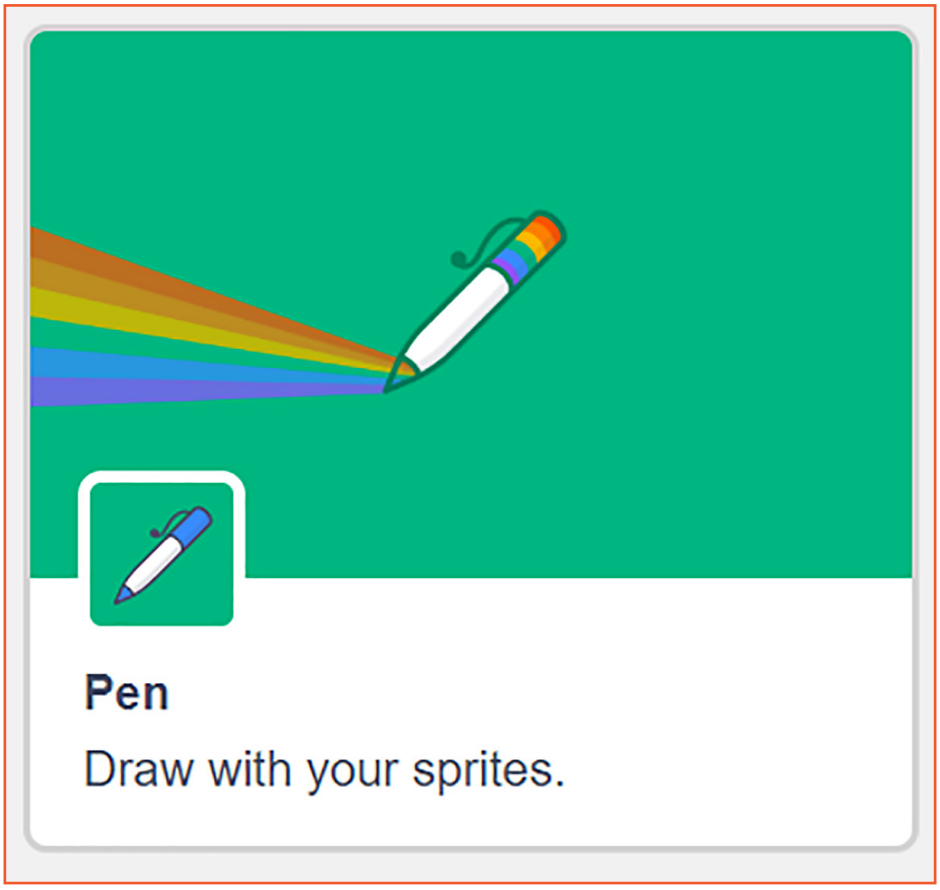
STEPS WITH MACHINE LEARNING FOR KIDS PLATFORM

Similarly, we can create the above project with another platform, i.e. Machine learning for Kids. This time let's try to draw using Nose in place of the Index finger.

Setting up the Project

1. *Visit the platform from here: - https://machinelearningforkids.co.uk/scratch3/*
2. *Create a new project.*
3. *Next, click on the Add Exte**nsion** button and add **Pose Detection** extension.*

4. *Next, we need to add the Pen Extension to draw on the stage.*
5. *Again, click on the **Add Extension** button and select **Pen** extension.*

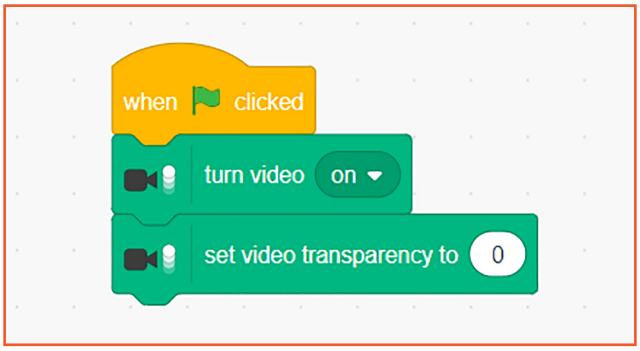


Adding the Sprites

Repeat the steps from 1 to 6 under the heading Adding the Sprites from the first platform to add the pencil sprite and the below 3 buttons.

Setting Up the Stage

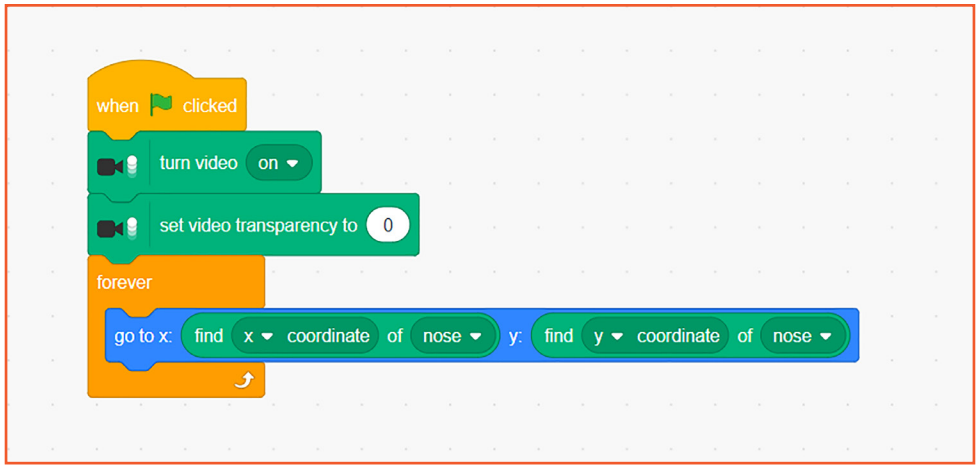
12. *Choose the Pencil Sprite.*
13. *Drag and drop a when flag clicked block into the scripting area.*
14. *Snap a turn video (on) block and set transparency to 0%.*



15. *The stage is set. Click the green flag to get the camera feed on the stage.*

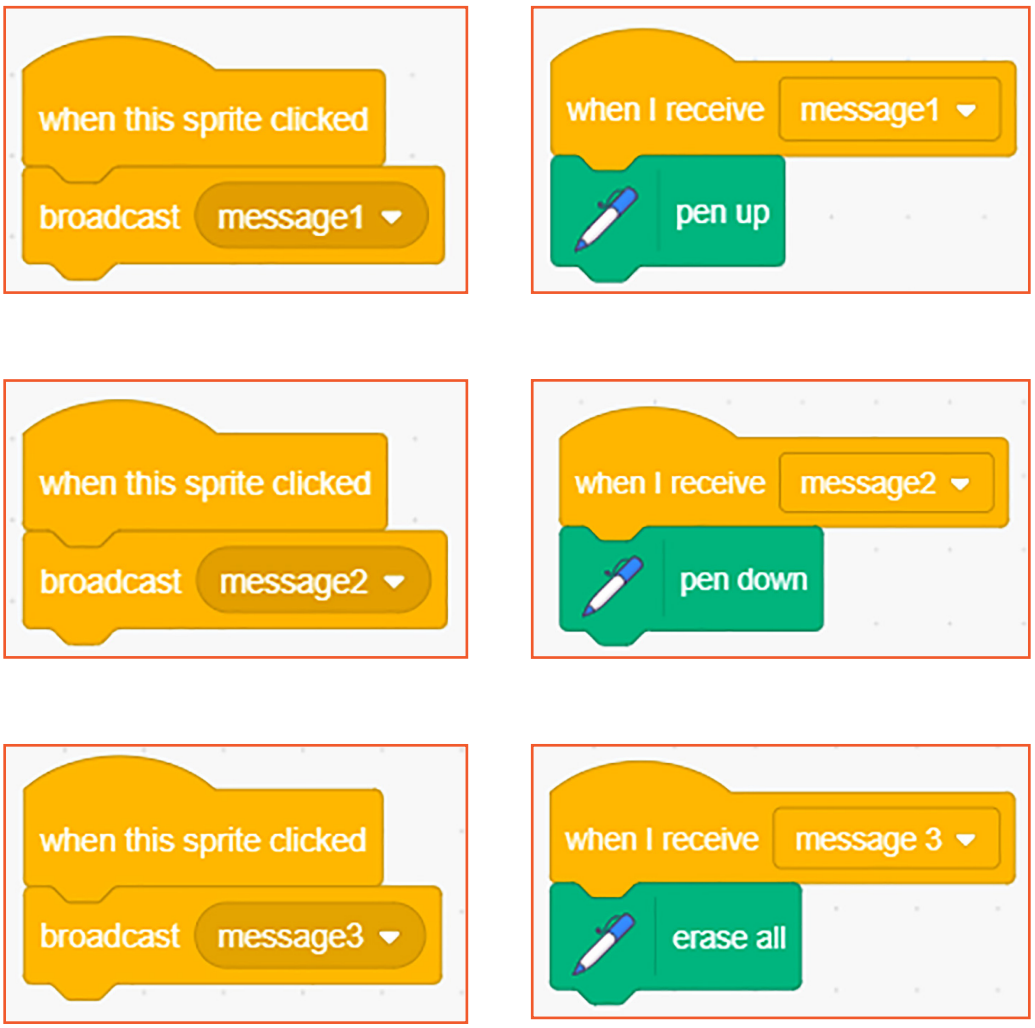
Detecting the Nose

1. *Snap a **forever** block to run further script continuously.*
2. *Inside it put a **go to x: () y: ()** block from the Motion Palette.*
3. *Place **find (x coordinate) of (nose)** and **find (y coordinate) of (nose)** block from the Pose Detection extension.*



Activating the Buttons

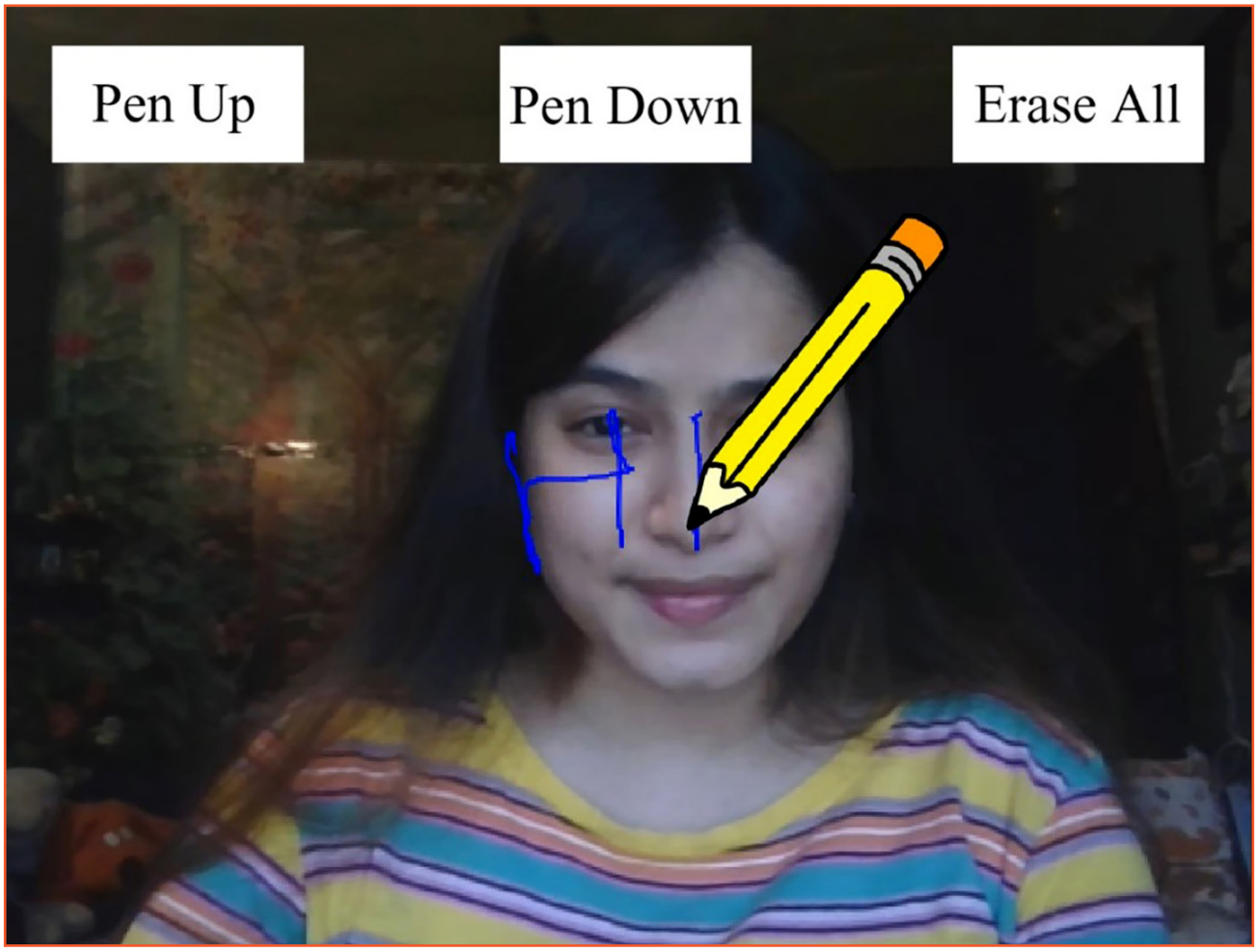
Repeat the steps from 1 to 11 under the heading Activating the Buttons from the first platform to activate, i.e. assign actions to button.



How does it work?

Now, that the script is completed, let's have a look at the logic behind it, i.e. how it works. It's almost similar to the previous project, the only difference is that we are using a different platform; hence the blocks and extensions are different. Also, in this project, we are using our Nose to draw the images.

Now make your nose the brush and try making various shapes, numbers or alphabets on the platform of your choice.



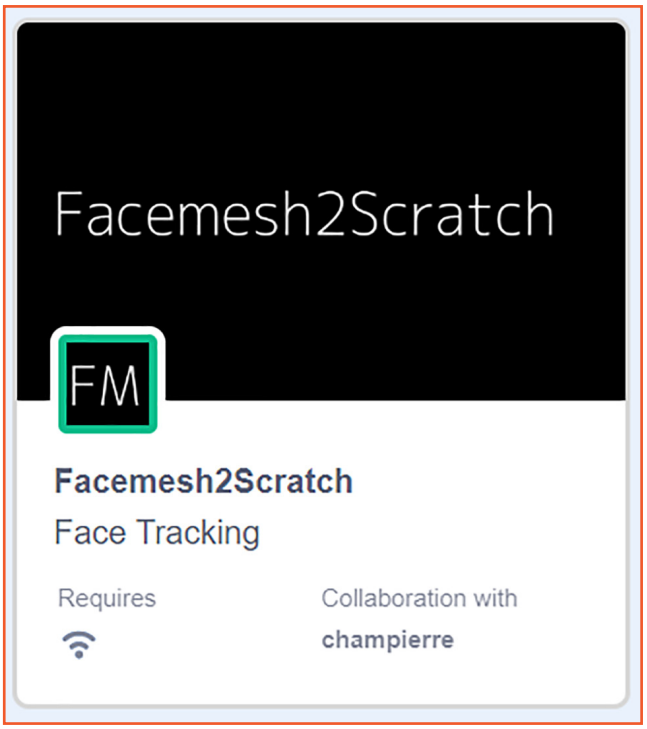
STEPS WITH STRETCH 3 PLATFORM

Similarly, we can create the above project with Stretch 3 platform as well. We will track the nose again and display the path on the screen.

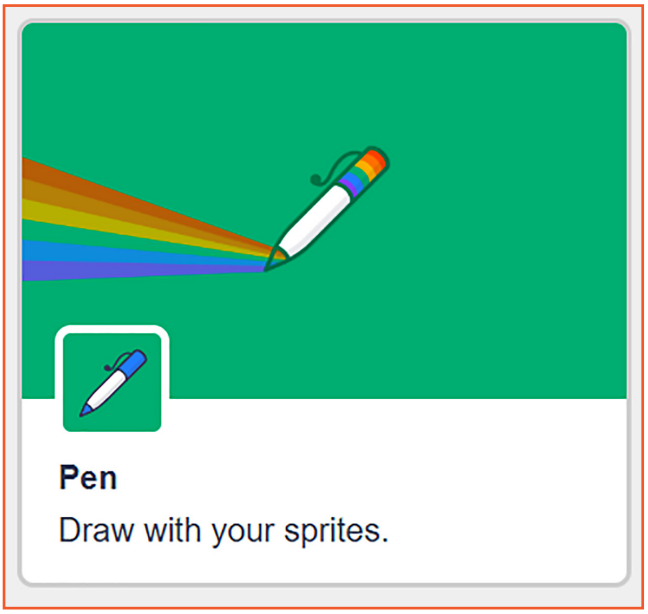
Setting up the Project

1. *Visit the platform from here: - https://stretch3.github.io/*
2. *Create a new project.*

3. Next, click on the **Add Extension** button and add **Facemesh2Scratch** extension.



4. *Next, we need to add the Pen Extension to draw on the stage.*
5. *Again, click on the **Add Extension** button and select **Pen** extension.*

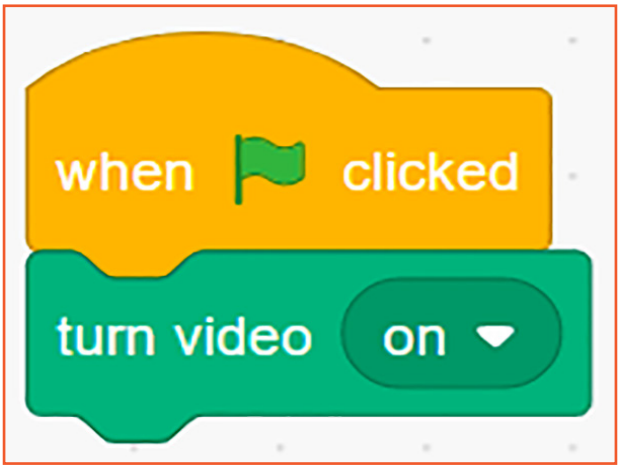


Adding the Sprites

Repeat the steps from 1 to 6 under the heading Adding the Sprites from the first platform to add the pencil sprite and the below 3 buttons.

Setting Up the Stage

16. *Choose the **Pencil** Sprite.*
17. *Drag and drop a **when flag clicked** block into the scripting area.*
18. *Snap a **turn video (on)** block.*

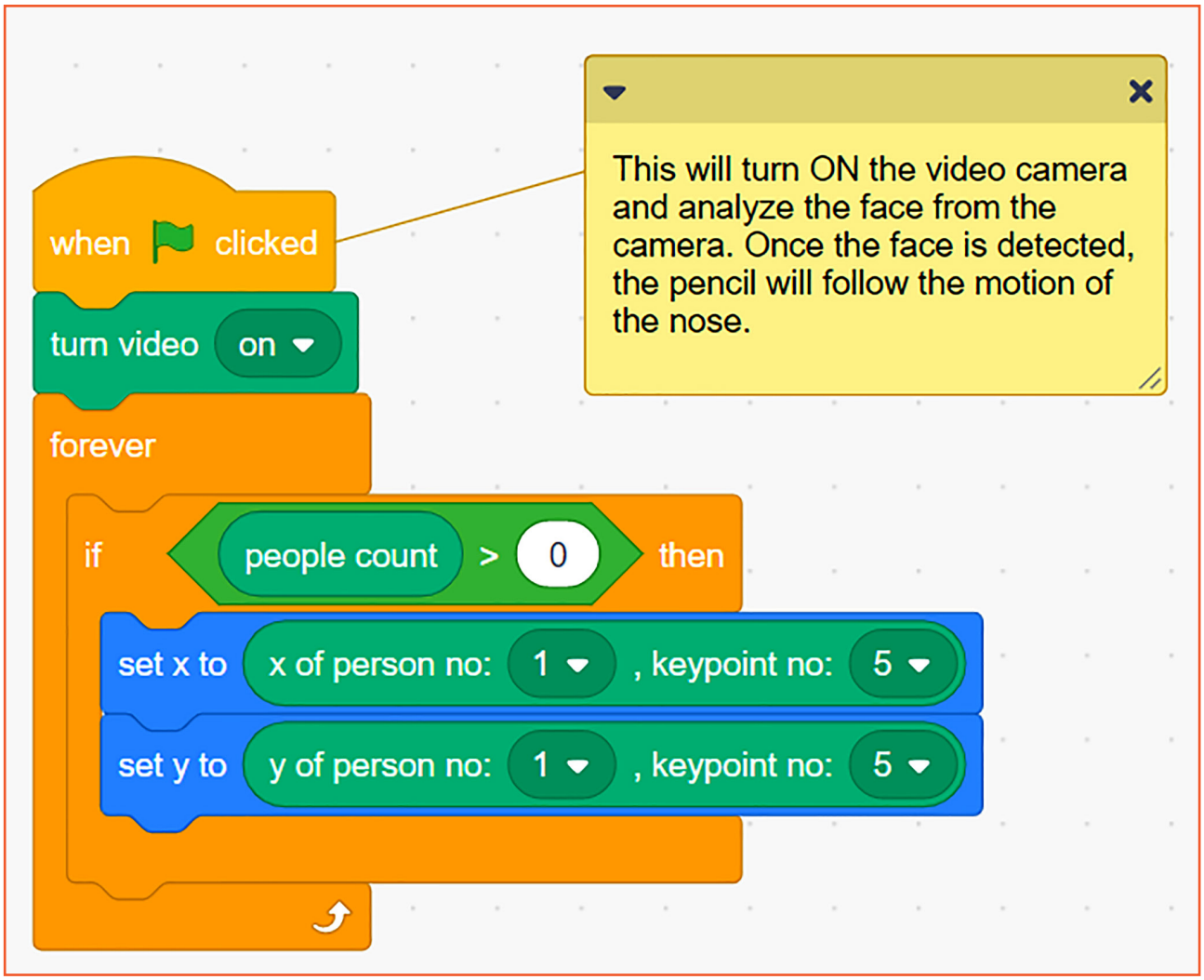


The stage is set. Click the green flag to get the camera feed on the stage.

Detecting the Nose

1. *Snap a **forever** block to run further script continuously.*
2. *Inside it put a **go to x: () y: ()** block from the Motion Palette.*
3. *Place an if block from the Control Palette.*
4. *Into the hexagonal space, place a **greater than** operator block from the Operators Palette.*

5. *Place **people count** reporter block from the Facemesh2Scratch Palette into the first input and write **0** in the second input of the greater block.*
6. *Drag and drop **set x to ()** block from the Motion Palette into the **if** arm.*
7. *Place **x of person no: (), keypoint no: ()** block from the Facemesh2Scratch Palette into the input space. Choose **1** from the first drop-down and **5** from the second.*
8. *Finally, place a **set y to ()** block from the Motion Palette.*
9. *Drag and drop **y of person no: (), keypoint no: ()** block from the Facemesh2Scratch Palette into the input space. Choose 1 from the first drop-down and 5 from the second.*



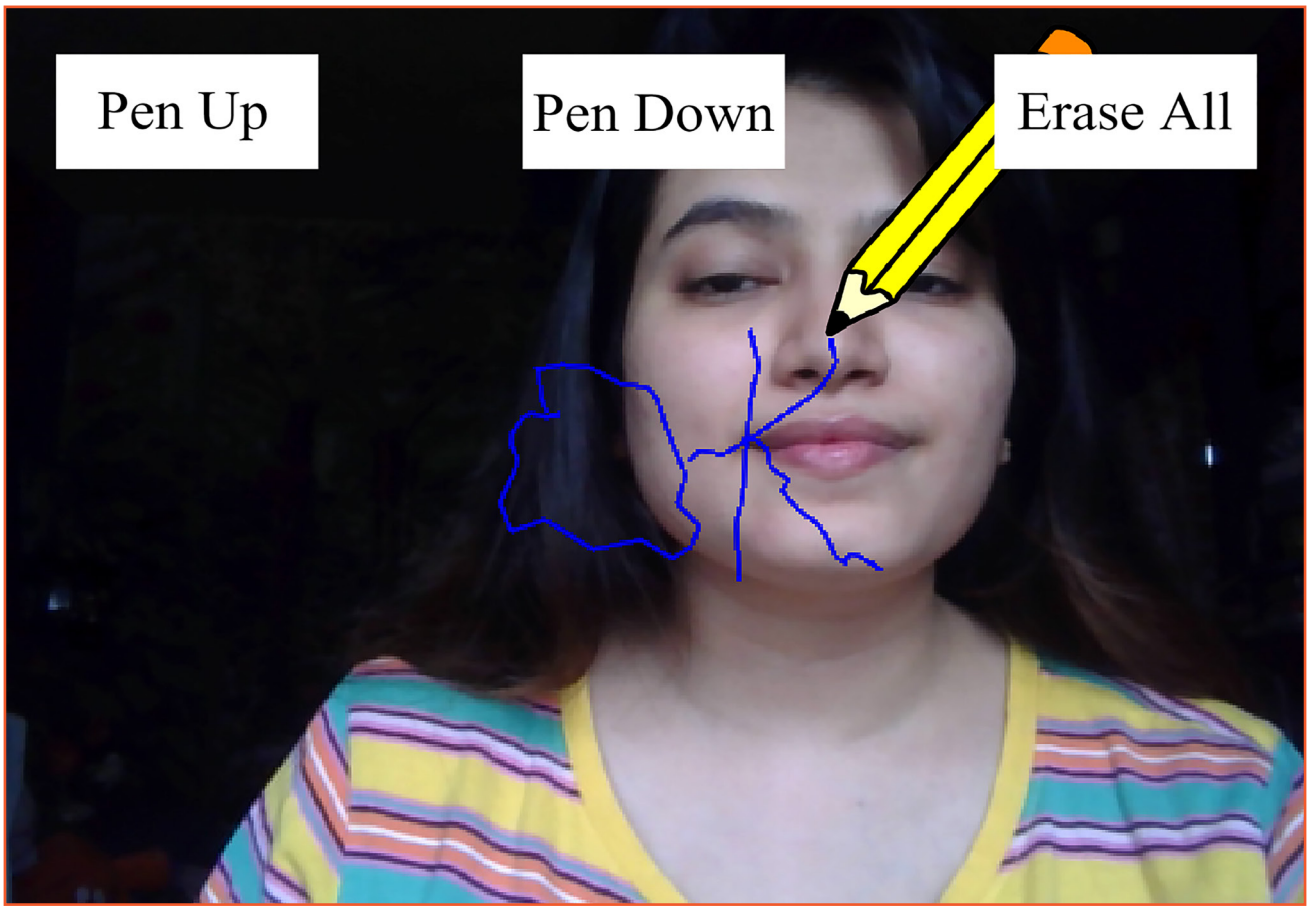
Activating the Buttons

Repeat the steps from 1 to 11 under the heading Activating the Buttons from the first platform, to activate, i.e. assign an action to buttons.

How does it work?

Now, that the script is completed, let's have a look at the logic behind it, i.e. how it works. It's almost similar to the previous two projects, the only difference is that we are using a different platform; hence the blocks and extensions are different.

Now make your nose the brush and try making various shapes, numbers or alphabets on the platform of your choice.



As you can see from this project, the same project can be coded and run with variations, on different platforms to achieve the same outcome. Similarly, there are many different platforms, approaches, service providers, APIs and IDEs, which one can use to achieve the project outcome. It is always advisable to learn about the various options available and develop skills in using these, to pick the ones most suited to the objective and other requirements.

Explore More

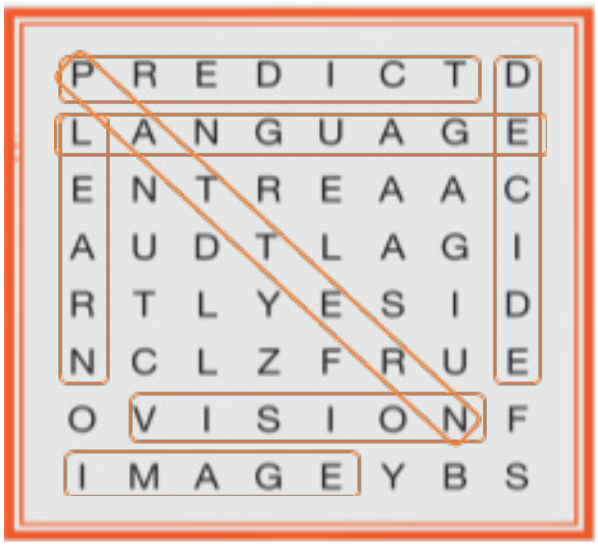
Make scripts to track various body parts one by one and brainstorm some real-world applications you can create by tracking them individually or simultaneously.

Conclusion

In this segment, we first learned about graphical programming fundamentals. Then we explored various graphical programming platforms available. Finally, in the end, we made a fun AI-based Air Draw project using different platforms. Now, before moving onto the next unit, you can explore possibilities with AI and make real-world applications based on learning from this unit.

ANSWERS TO "BASE MODULE RECAP QUIZ"

Answer - 1



Answer - 2

Column A: TERMS	Column B: EXPLANATIONS
Supervised Learning	AI Machines are feed the features and their corresponding labels in a process called training.
Unsupervised Learning	AI Machines figure out how the various inputs relate to each other and then it bunches similar inputs together.
Reinforcement Learning	AI Machines learn from the environment by interacting with it and receive rewards for getting the right results and taking the right actions.
Natural Language Processing	AI Machines can be made to respond to naturally spoken and written language.
Facial Recognition	AI Machines can be made to recognize faces, their gender, age group, ethnicity, and emotions.
Computer vision	AI Machines are developing to differentiate between images of objects like cats, dogs, trees, walls, people, etc.

Answer - 3

The three basic stages of Machine learning
i) Input (Data) (ii) Processing (Algorithms) iii) Output (Prediction)

Three branches of Mathematics that are used in machine learning
i) Algebra ii) Probability iii) Statistics

Three types of Qualitative data
i) Binomial Data ii) Nominal Data iii) Ordinal Data

Answer - 4

	CLUES	ANSWER
ACROSS	It is the process of finding out the relationships between different parts of the data and the patterns that can be recognized in it.	Modelling
ACROSS	This data typically involves integers and is a count that cannot be made more precise.	Discrete
ACROSS	It is a type of Algorithm where we look at each piece of data, one by one, and don't stop until you find what you are looking for.	Sequential search
ACROSS	It is the selection of specific choices of data out of the total data available in a column in a spreadsheet.	Data filtering
ACROSS	It is a collection of facts that can be in the form of numbers, words, symbols or even pictures.	Data
ACROSS	It is a mathematical aid that involves gathering, reviewing, analyzing the data to draw important conclusions as per the application.	Statistics
ACROSS	It is a type of Algorithm used to find a specific value in an ordered list, by dividing the list into two parts, each time it searches.	Binary search

DOWN	It is a set of step-by-step instructions to solve a problem created along with rules that decide which path to take when a decision is faced.	Algorithm
DOWN	It is a type of complex Algorithm used in Supervised Learning where rules are defined for each training set on the basis of which the objects in the training set is classified.	Classification
DOWN	It is a type of data that can take any value within a range, also called Analog data.	Continuous
DOWN	It is a method to represent information and data graphically by the use of images like charts, graphs, and maps	Data visualization
DOWN	It is a number system that only uses two digits 0 and 1. Computers process information using this system.	Binary system
DOWN	It is the process of arranging the data into a meaningful order so that it can be analyzed or processed easily.	Data sorting
DOWN	It refers to a collection of large and complex datasets that may be complex due to its volume or its variety.	Big data

Answer - 5

Algorithms are **step-by-step** instructions to solve a **problem,** along with **rules** that decide which path to take when a **decision** is to be made. When trying to solve a problem, we may find more than one **solutions.** A different **Algorithm,** can be built from each solution. **Logical** reasoning determines if algorithms will work by **predicting** what happens when the algorithm's steps, and the rules they consist of, are followed. Algorithms can then be represented in the form of a **flowchart.**

SUMMARY

Some areas where AI is currently being used are Healthcare, Astronomy, Agriculture, Education, Entertainment, Business, Art, Data Security, Social Media, Chatbots, Welfare, Automobiles, Aviation, Deepfakes.

AI is a broad field that includes several theories, methods and technologies, as well as the following major domains, such as A neural network, Deep learning, Cognitive computing, Computer vision, Natural language processing (NLP).

Additionally, several technologies enable and support AI, such as, Graphical processing units (GPU), Internet of Things (IoT), Advanced algorithms, Application Programming Interfaces (API)

The core process of developing an AI using machine learning primarily remains similar for various applications. It may involve minor variations based on the context, purpose and approach. Let us look at the process steps:

1) Identify the problem
2) Acquire relevant data
3) Prepare the data
4) Analyze the data using a model
    a. Choose the algorithms
    b. Train the model
    c. Evaluate the model
5) Choose a particular programming language
6) Run on a selected platform

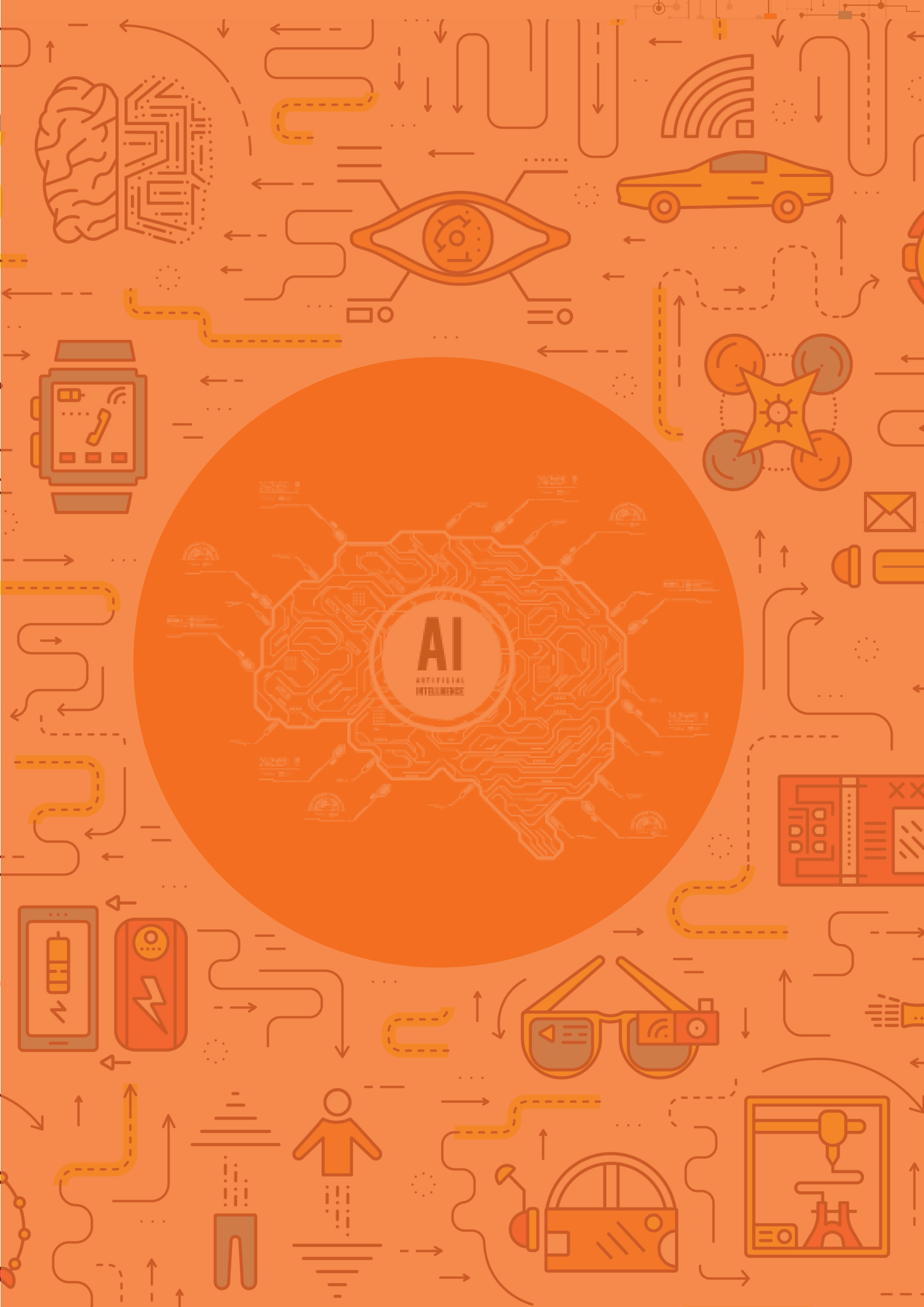
Graphical programming (or visual programming) is a method of programming in which the program(s) are written using graphical elements known as blocks. Due to its intuitive interface and programming blocks, it becomes very easy to make games, animations, program robots, and make AI/ML projects. In this module, we will use graphical programming for doing AI-based projects.

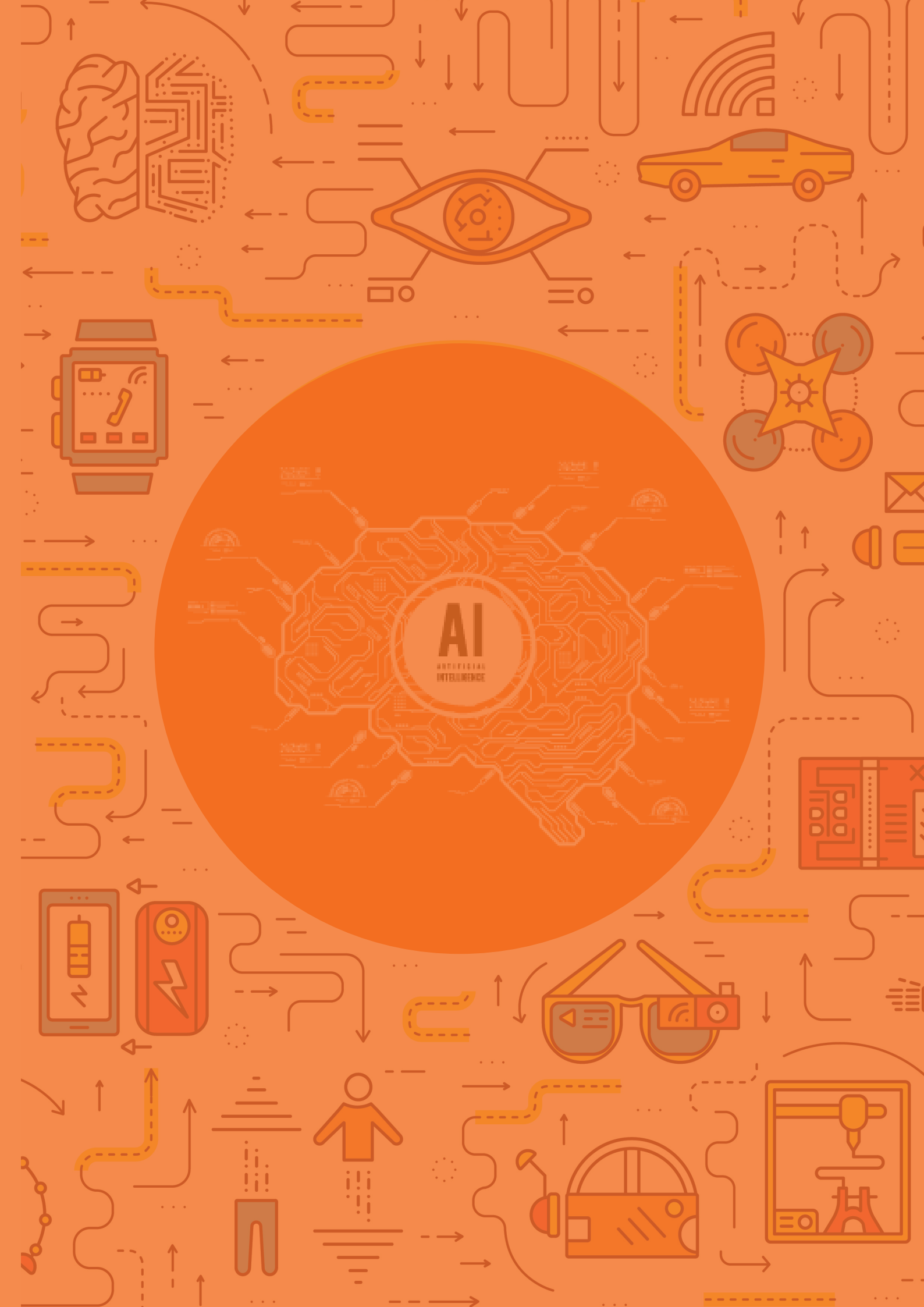
Graphical programming platform consists of the following:

- The stage is a background or a backdrop for your projects (the programs you'll create).

- A Sprite is an object or a character that performs different actions in the projects. It understands and obeys the instructions that you'll give them in your projects. It can move and can be moved
- The Backdrop is similar to Sprite, but with some limitations like it cannot move. It has its own set of scripts, images, and sounds.
- Blocks are like jigsaw puzzle pieces that fit into each other. They are predefined commands used to create programs by simply dragging and dropping them below one another in the scripting area.
- A script is a program or a code in Scratch based graphical programming lingo. It is a set of 'blocks' that are arranged below one another in a specific order to perform a task or a series of tasks.

Multiple platforms support graphical programming for implementing artificial intelligence and machine learning. Scratch can also be used for physical computing and can be linked to computers such as Microbit, Arduino and Raspberry Pi.


AI
ARTIFICIAL
INTELLIGENCE


AI
ARTIFICIAL
INTELLIGENCE

UNIT 2

ETHICS IN AI

Key Learning Outcomes

At the end of this module, you will be able to:

a) State logically the likely impacts of AI technology on society
b) Explain potential sources of bias in AI decision making
c) Critically explore the positive and negative impacts of an AI system
d) Research and present key ethical concerns related to AI development and use in various areas of application
e) Use tools and templates for evaluating ethical considerations and bias in the AI project or application

Topic Covered

What does 'Ethical AI' mean? | Fairness | Robustness | Privacy | Explainability | Governance

2.1 WHAT DOES “ETHICAL AI” MEAN?

Humans have something called “a moral compass”. It is an area of our cognition (mental processes of gaining knowledge and understanding) which tells right from wrong. When you see an injustice, your brain tells you something is not right.

The standards of your moral compass strongly depend on your upbringing and environment.

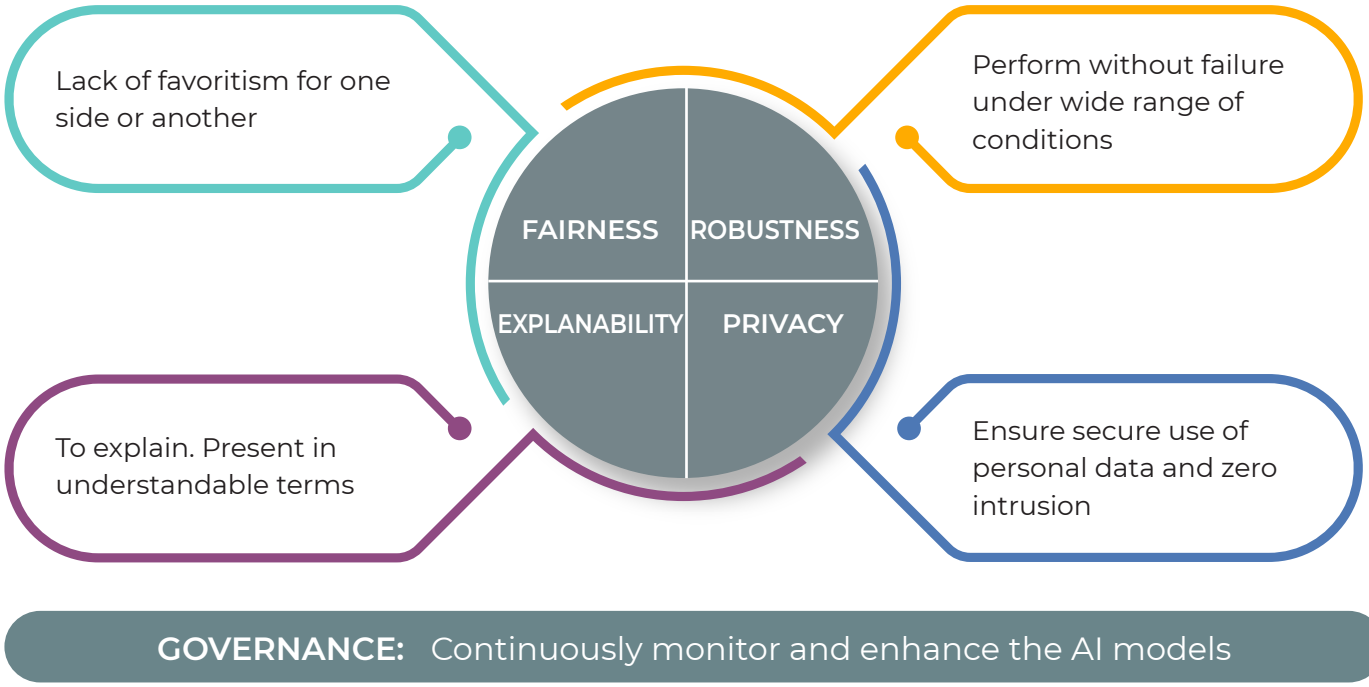
This 'moral compass' and environment is also what companies build their ethics on, i.e. to decide what is right and what is wrong.

AI is a technology that could affect public values, human rights or public decision-making and hence there is a strong need for ethics here.

In the same way that parents raise a child, companies must teach and test their AI systems to learn, communicate, make unbiased decisions and comply with norms and regulations.

Artificial Intelligence is a very vast field, and today most of the AI code developed across the world falls under the bucket of Artificial Narrow Intelligence (ANI). In this Unit, we will discuss the ethical considerations with respect to ANI.

Ethical AI pillars are as below:



GOVERNANCE: Continuously monitor and enhance the AI models

2.2 FAIRNESS

AI Systems should be built such that they are fair and inclusive for all. Impartial behavior of AI model without any discrimination should be our goal. Unfair models can create improper outcomes,

For example, the AI model for face recognition identifies a person as a criminal incorrectly - this will lead to a loss of that person's reputation and false criminal charges against the person.

Unfair models can injure people, damage reputations and become legal liabilities. The consequences of biased AI systems could be an injustice at a large scale.

Ensuring Fairness in AI is hard because of below 3 points:

1. There is pre-existing bias in the data that is collected from the real-world.

   Example: Gender bias - when you search for 'Doctor' on a search website, the results for Doctor images might mostly come up as 'male'. Whereas for 'Nurse' most results would be 'female'.

   This shows inherent bias in the search results. Now, if you train your AI models using this dataset, the AI model will be biased too and learn that to be a Doctor, being male is an important criterion.

Doctor *Nurse*



Hence, we need to ensure that such bias is identified before the AI model is trained.

2. Ensuring that the AI system is fair across all the circumstances and scenarios is a challenge. Even thorough training and testing of the AI systems cannot guarantee this.

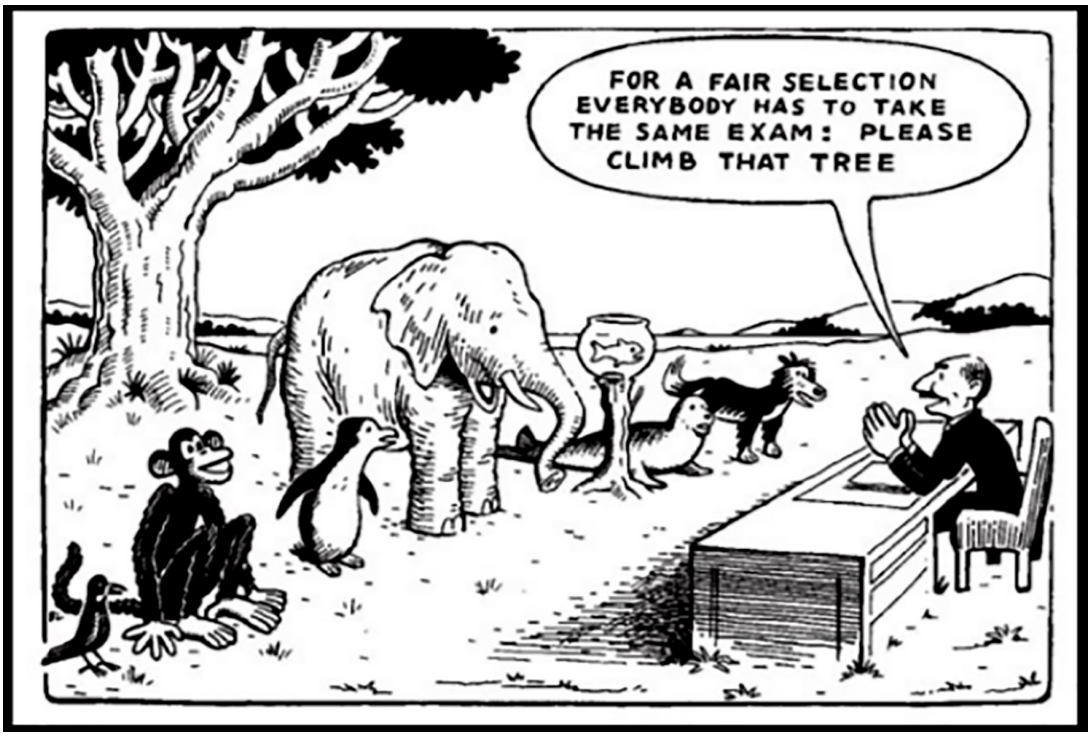
There will always be situations where only after the launch of the AI system, unfair and unintentional blind spots get revealed.

Example: – An AI model is trained on 1000 different scenarios in which a car can meet an accident with another car. It is very good at identifying which car's driver was at fault and caused the accident if one of these 1000 scenarios come up. But after launch, a new scenario comes up that was not foreseen. Hence, the AI model will need retraining and until then, it might be unfair.

3. When it comes to fairness, there is no single approach. What might be fair for one person might not be fair for another. In such cases, the need would be to identify the Fairness Criteria for building the AI system by considering the historical, ethical, social, legal, political and maybe even more scenarios. See picture below to get an understanding of what we mean.

As you can see and understand now, implementing fairness is very important for Artificial Intelligence systems.

And below are some recommended practices that everyone can follow to get a fair AI system in place.



Let's go through an exercise to understand the topic better:

1: Why is Fairness an important criterion for a good AI model?

A) Fair AI models will enhance your reputation as a developer

B) Unfair AI models means you have to rebuild them leading to extra work

C) Unfair AI models will result in biased outcomes that could damage reputation, harm people or even lead to legal issues

D) Fair AI models will always give accurate results

2: Which of the following is an example of an unfair AI model?

A) Model that predicts women having equal probability to be a doctor or a nurse

B) A model that predicts a doctor to be either a man or a woman with equal probability

C) A model that predicts a doctor as a nurse or a nurse as a doctor

D) A model that always predicts nurses to be women and doctors to be men

3: Ensuring fairness in AI is difficult because:

A) Fairness is subjective. So, what appears like a fair model to one may not be acceptable to the others

B) AI models are black box models, so you can't really say whether they are fair

C) Any AI model will eventually surpass human intelligence and starts treating humans as slaves

D) All of the above

4: Which of the following is one of the primary causes for unfairness in any AI model?

A) AI models can remember data about individuals and give subjective results

B) AI models can't be explained easily

C) If the data itself is biased then any model trained on such data will also become biased

D) None of the above

5: Which of the following is one of the methods to ensure fairness in an AI model?

A) By training the model for less duration

B) By looking not only at the historical data but also considering the implications of social, legal, political, and other scenarios related to such data

C) By training the AI model for longer duration

D) All of the above

And below are some recommended practices that everyone can follow to get a fair AI system in place.

RECOMMENDED PRACTICES TO ENSURE FAIRNESS	
Chalk out exact goals for ensuring fairness and inclusion	Take inputs from experts ranging from Industry (Insurance, Healthcare, etc.) to Humanities (lawmakers, ethicists, etc.) to consider different thought processes.
	Set goals for an AI system to be fair for a list of anticipated scenarios. Goals will be monitored, and the list will be expanded over time (based on feedback loop)
	AI models should be built and tested (against defined metric) so that they reflect the fairness goals.
	Update the input data frequently based on usage of the AI system (feedback loop)
Representative datasets should be used	Look for fairness in the data by identifying enough representation across different classes/groups. Identify prejudice, discriminatory correlations between features, groups.
	Data should represent real-world frequencies of people, events etc. that your AI system needs to make predictions about.
	Data will be annotated (text explaining that data will be put on it by humans) before using in AI models. Understand that there could be mistakes in those annotations (because of human error, etc.).
Analyze performance – Real-world scenarios	Take feedback from users after testing your AI model on real-world scenarios. The users and scenarios should be from across different industries, groups etc.

2.3 ROBUSTNESS

AI models are highly susceptible to all sorts of attacks, including many based on adversarial AI methods. In these methods, adversaries exploit AI models for potential gain, in most cases, by predicting the decisions machine learning models will make and then manipulating subsequent sets of data to produce the attacker's desired outcomes—rather than the correct decisions.

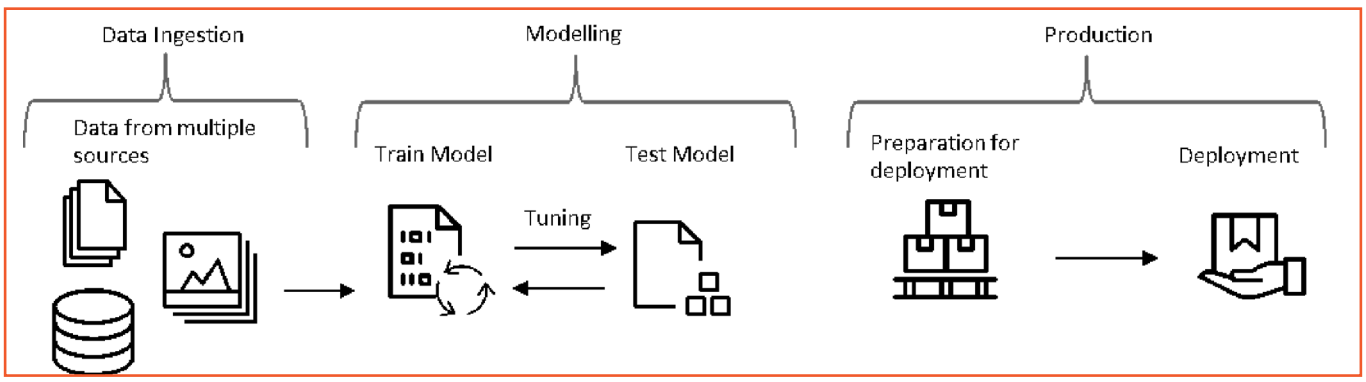
The accuracy of AI models is directly correlated to their vulnerability to small variations on the input dataset. Testing and benchmarking AI models against adversarial attacks is key to establishing trust in AI systems.

For example, the original image below was classified by the AI model as Panda, which is correct. Once an adversary adds an input in the dataset that is random noise, any human being would still classify the image as a Panda. But the AI model, which works at pixel levels for images, gets confused and classifies the image as a 'Gibbon'.

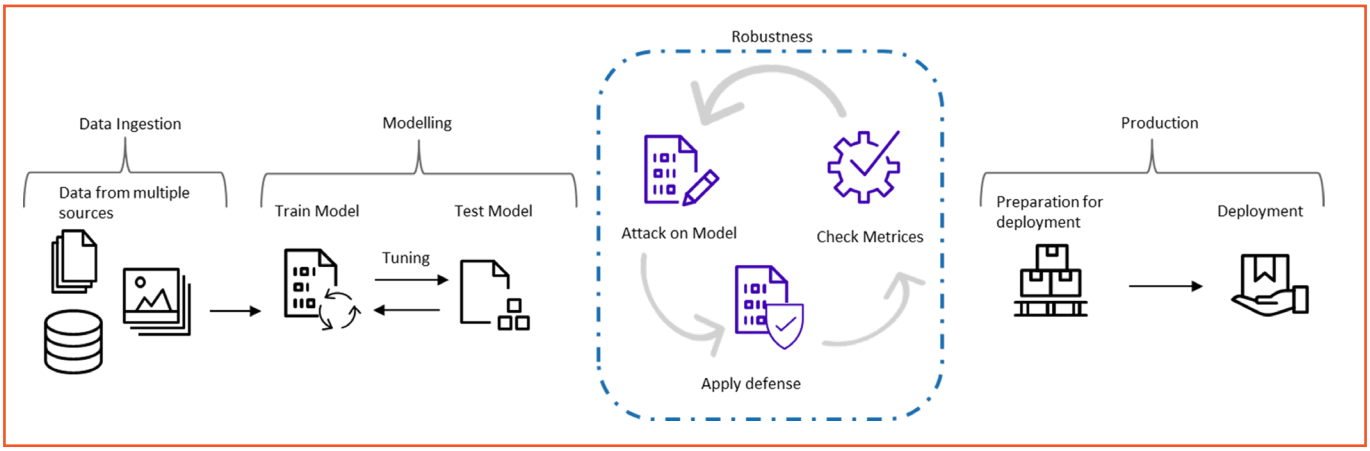
The question is – what can we do to ensure AI models do not break due to such adversarial attacks?

Classified as panda | Small adversarial noise | Classified as gibbon

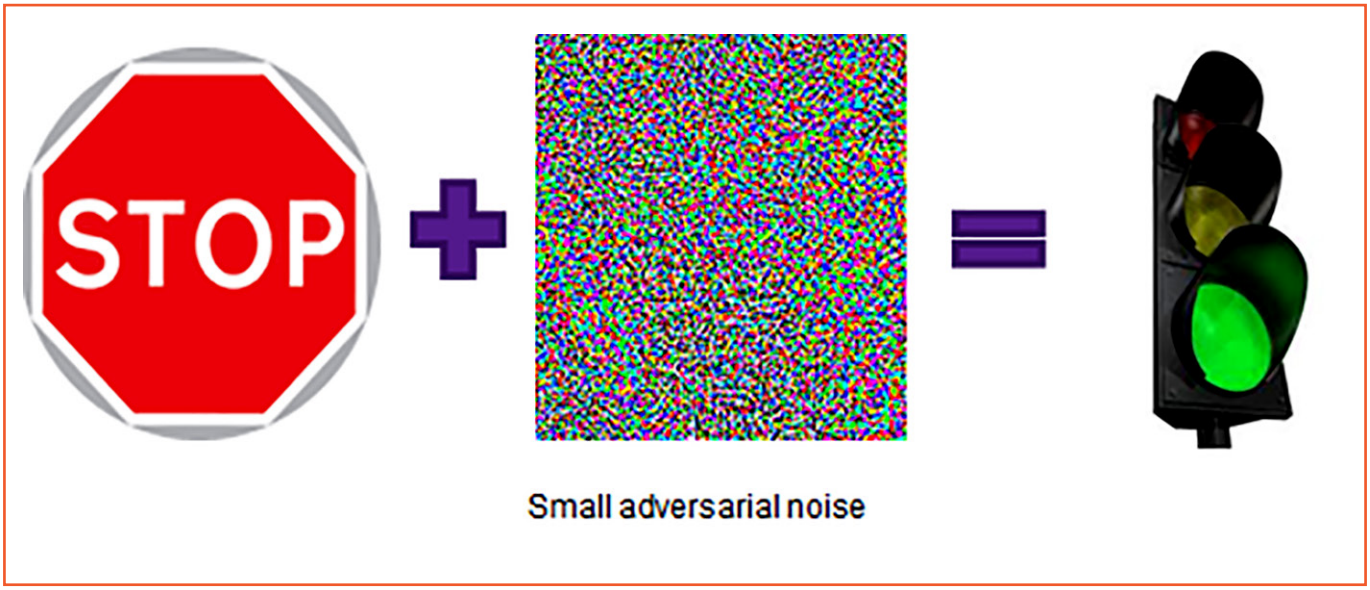
A regular AI model development flow looks like the one below and does not include any checks or fixes to ensure the Robustness of a model.



We can alter the AI model development pipeline to include attack scenarios when training the AI model. This will teach the defense mechanisms to the AI models at the time of training. So, when the model is deployed in the real world, it will be able to handle the attacks much better. Like shown in the diagram below.



For example, before a model is trained for Robustness, it can be fooled into accepting that the STOP sign on the left is a Green signal. This can create hazardous situations in the real world, where we might have self-driving cars on the road, relying on AI models to read the STOP signs. And will end up reading them as Green signals instead.



But once the AI model is trained with Robustness in mind, it will be able to identify the attack and give a correct result.

Let's go through an exercise to understand the topic better:

1: When should we carry out the Robustness check in the AI model development pipeline?

A) Before we build the AI Model
B) Before the data ingestion phase
C) After we build the AI Model
D) After data ingestion phase

2: Why is it necessary to check the Robustness metrices in the AI model development pipeline?

A) To train the AI Model
B) To validate the defense that has been applied
C) To validate if the input data is correct
D) To test the accuracy of the AI Model

3: Which are the factors on which the concept of AI Robustness is based?

A) Safety and Security
B) Explainability
C) Both A and B
D) None of the above

4: Which of the below scenarios are examples of probable adversarial attacks on the AI system? (Select 2)

A) AI Model recognizing the image of a Panda as a Gibbon
B) AI Model recognizing a Red Traffic signal as a Green Traffic Signal
C) Human recognizing the image of a Panda as a Gibbon
D) None of the above

5: Is Robustness of AI Model dependent on input data?

A) Yes
B) No

Given below are some recommended practices that everyone can follow to achieve a robust AI system.

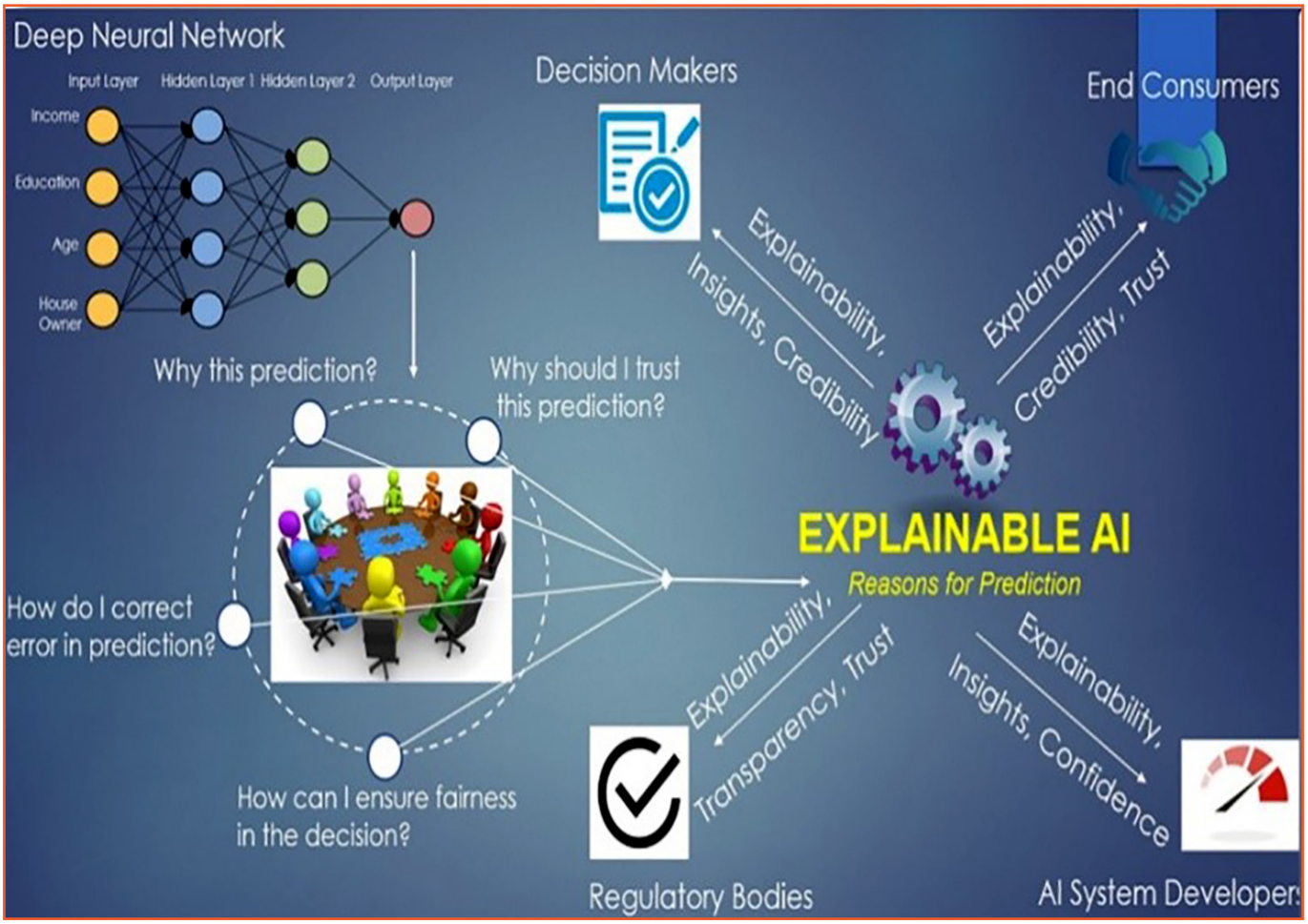
RECOMMENDED ACTIONS TO ENSURE ROBUSTNESS	
Potential threats to the system	Make a list of threats (that could happen) to the AI system and identify the problems if those threats do happen. Understand the likelihood and severity of these problems.
Approach to overcome the threats	Develop a defense approach (from the available defense techniques) to fight the threats. Test the AI system using the chosen approach.
	Perform security tests (at random) to ensure your AI model is prepared in case a real threat happens.
Learn and stay updated on the latest security techniques	Adversarial AI is continually improving. It continues to improve performance for defenses and some defense techniques are offering guarantees of security.

2.4 EXPLAINABILITY

AI models are like a black box. There is little visibility on how AI systems make decisions and the reasons behind the AI model predictions. Due to this lack of transparency, businesses had started to hesitate to utilize this technology.

Explainability has led to the opening of this AI black box leading to transparency of the decision-making process of the AI models. Explainability is vital to being able to question, understand, and trust AI systems.

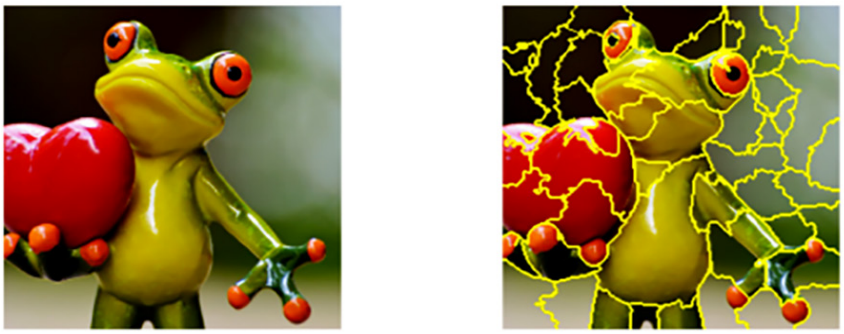
AI systems can now list information about features and the values of these features that went into its prediction/decision-making process. Hence, enabling human beings to look inside the black box and gain insights into the decision-making process of AI models. The below diagram explains the concept very nicely.



Explainable AI systems are available, which can:

- Explain – rationale; the reasoning, whenever needed
- Characterize – strengths and weaknesses
- Compare – with other AI systems
- Convey – an understanding of how they will behave in the future
- Make – decisions smarter by augmenting humans with machines

For example, we want to train an AI model to determine if a species of frog is poisonous or edible for other animals. But, also want an explanation or details on the parameters and values the AI model used to determine the poison level for the frog.

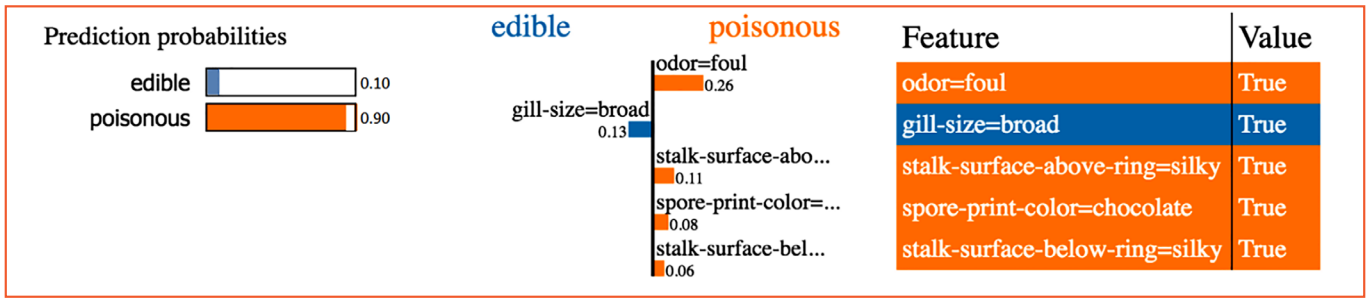


Original Image

Interpretable Components

One of the techniques used for explainability is called LIME (Local Interpretable Model-Agnostic Explanations). It attempts to understand the model by changing the input of data samples and understanding how the predictions change.

It modifies a single data sample by tweaking the feature values and observes the resulting impact on the output/prediction. See picture below.



Let's go through an exercise to understand the topic better:

1: AI models are considered black box models because:

A) They can give unfair results

B) They are susceptible to attacks

C) It is difficult to explain why a particular decision is suggested by AI model and the reason behind the result

D) It is difficult to say if the AI models are ensuring your privacy

2: Why is Explainability an important element for a good AI model?

A) When a model behavior can be explained it builds trust in the users

B) Even if the AI predictions are bad, if you can explain and convince people, they will accept the model

C) If you can explain a model that means it is robust

D) All of the above

3: Which of the following is a technique used for Explainability?

A) Locally Interpretable and Model Aware Explanation

B) Local Interpretable Model-Agnostic Explanations

C) Logical, Interpretable, and Manageable Explanations

D) Logically Interdependent Marginal Explanations

4: One way of explaining model behavior is:

A) Tweak the input features, observe the changes in results. This explains which of the features are impacting the results of the model

B) Take data covering a wide variety of scenarios (e.g., social, political, etc.), then you don't need to provide any explanation, the model speaks for itself

C) There are privacy constraints on the data used for training the models, hence we should not provide any explanation else we violate the privacy terms

D) All of the above

5: Key characteristic of good explainable model is:

A) They can provide the rationale behind the results

B) They are robust and fair

C) They can make better decisions

D) None of the above

And below are some recommended practices that everyone can follow to build an explainable AI system.

RECOMMENDED ACTIONS FOR EXPLAINABLE AI	
Chalk out options to perform Explainability	Engage with the Industry experts (banks, healthcare, etc.) of your AI system and decide on the exact list of features that are needed to be explained. Understand the limitations you may have, e.g. less access to data if working with sensitive data etc.
Explainability is an important part of the user's experience	Explainability is an important part of the user's experience Discuss at regular intervals with the AI system users on their needs, i.e. the list of features they need to be explained etc.
Design AI Model to be Explainable	Use the simplest model with the smallest set of inputs necessary to meet your performance goal of understanding the factors that are affecting your AI model
	Learn causal relationship and not co-relation when possible, e.g. use height and not age to predict if a kid is safe to ride a roller coaster.
Use a metric that shows the performance of the end goal	Select the metric that is suitable and important for your AI system. E.g. a fire alarm system would need to have "high recall" as the metric to be tracked, even if that means the occasional false alarm.
Understand the trained model	Analyze how your AI model behaves for different data, for different subsets of examples. Look at how even minor differences in data impacts your AI model.

2.5 PRIVACY

When AI systems are to be built using sensitive data, we need to consider the privacy implications in using it.

Legal & regulatory requirements, social norms and individual expectations should be obeyed.

If AI models remember or can reveal the sensitive data that they have used, rules should be put in place to ensure that the data is safe.

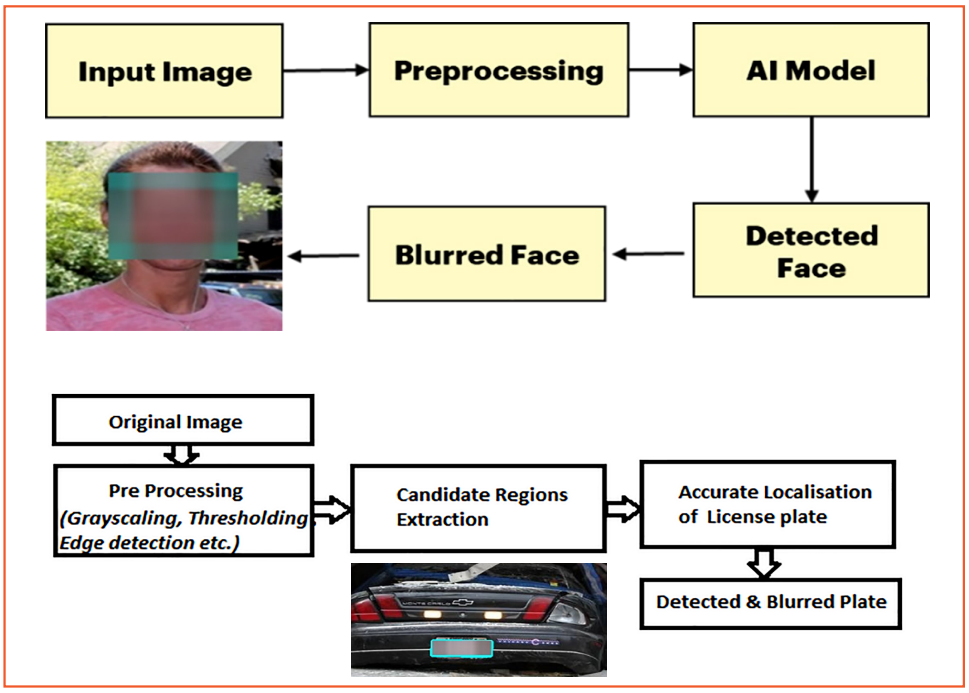
The possibility that AI models reveal data can be decreased by applying various techniques.

For example, personally identifiable information, or PII, is any data that could potentially be used to identify a person. Examples include a person's picture, full name, social security number, driver's license number, bank account number, passport number and email address. Identity theft is a big consumer concern. And identity thieves can piece together PII information from various sources and cause harm to the consumers.

There are various rules and regulations across the globe to protect data privacy. For example, The General Data Protection Regulation (GDPR) is a regulation in European Union (EU) law on data protection and privacy for all individuals within the EU and the European Economic Area (EEA). It requires companies to protect the privacy of their EU customers.

Hence the appropriate measures must be put in place to prevent privacy breaches.

Examples of anonymizing PII data in images – faces and car license plates.



Let's go through an exercise to understand the topic better:

1: Select examples of sensitive data from the options below (Select 2):

A) Images collected of cats and dogs
B) Images collected of people
C) Images collected of trees and plants
D) Images collected of car number plates

2: Why is PII data considered sensitive?

A) The data collected is private and must not be made available for public use
B) The data if stolen can be misused and result in privacy breach
C) The identity of individual persons can be traced
D) All of the above

3: What types of personal information do identity thieves steal?

A) Credit Card Details
B) Bank Account Number
C) Both A and B
D) None of the above

4: How can sensitive data be protected?

A) By enforcing regulations around data privacy like GDPR
B) By encrypting personal details
C) By blurring faces of people in images
D) All of the above

5: Which is the correct sequence of steps to blur a face using AI?

A) Input image -> Preprocessing of Data -> AI Model -> Face Detection -> Blurring
B) Input image -> Face Detection -> Preprocessing of Data -> Blurring -> AI Model
C) Input image -> AI Model -> Face Detection -> Preprocessing of Data -> Blurring
D) Input image -> Preprocessing of Data -> Face Detection -> AI Model -> Blurring

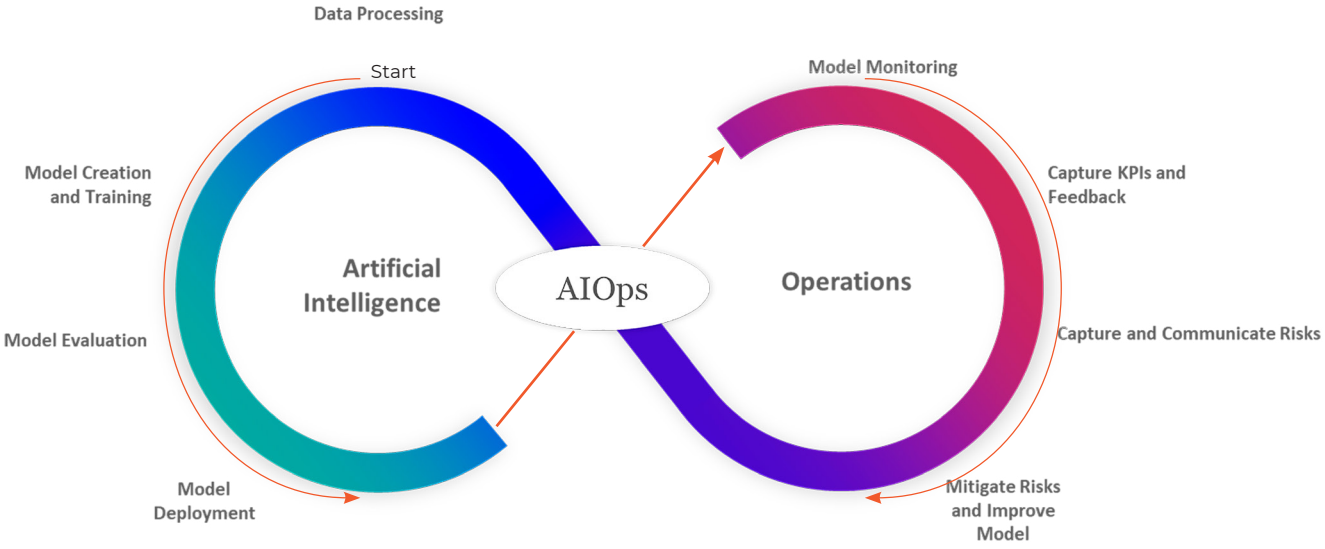
And below are some recommended practices that everyone can follow to ensure Data Privacy.

RECOMMENDED ACTIONS TO ENSURE DATA PRIVACY	
Handle data responsibly	Limit usage of sensitive data as much as possible.
	Obey the laws, standards, privacy principles etc. Provide users with clear notice and give them any necessary controls over data use. Follow best practices such as encryption (the process of converting information or data into a code, mainly to prevent unauthorized access).
	Masking, Anonymizing and/or aggregating the data (i.e. the process of hiding original data with modified data, i.e. characters or other data) using some techniques (data scrubbing, etc.). E.g., consider removing personally identifiable information (PII) like Date of Birth etc
On-device / Different setting processing of data	Apply the technique of federated learning, which is an approach of distributed machine learning wherein data stays in users' devices and is not sent outward/shared.
	Apply transfer learning technique which uses the knowledge gained in one setting/AI system to improve generalization in another setting. Can be used to transfer knowledge from sensitive data without losing data privacy.
Safeguard the privacy of AI Models	Using tests/techniques, find out whether your AI model is unintentionally remembering or exposing sensitive data.
	If minimal/masked/etc. Data is used, experiment with your AI model to find out the most optimal settings of it to ensure that the output prediction is least impacted.
There is no one size fits all approach to privacy protection. In real-world, we must iterate to find an approach that balances privacy and utility for the task we need to work on.	

2.6 GOVERNANCE

Artificial Intelligence algorithms are expected to carry out decision making across several sections of our society. Hence it is vital to ensure that the complete AI Operations lifecycle is continuously monitored and vigilantly governed. The model should also be enhanced on the basis of basis new situations encountered in the real world.

A typical AI Ops (or AI Operations) delivery lifecycle can be seen in the diagram shown below.



Along with the AI model delivery lifecycle, we need to ensure that the AI system is kept accountable. A few of the below points can be put in place to ensure accountability:

1. Plan for doing audits, impact assessments or thorough governance via rules and regulations.
2. For high-risk systems, Governance activities carried out by a human being could also be proposed as a model for accountability.
3. Also, multiple research groups help identify AI systems that are unfair or discriminatory and try to ensure that there is accountability and action taken

APPENDIX

COMMON ETHICAL CONCERNS RAISED BY AI

JOB APOCALYPSE
AI is so ruthlessly efficient that it will lead to massive job loss.

THE SINGULARITY
We will create something that is more intelligent than humans and we will lose control.

LACK OF TRANSPARENCY
True (scary) AI doesn't explain itself.

INCLUSION & DIVERSITY
The power of AI is in the hands of the few and with the traditional power brokers.

ARTIFICIAL STUPIDITY
AI is actually pretty stupid right now, and can lead to an embarrassing PR incident or worse, do something discriminatory.

PRIVACY
AI will erode our notions of data privacy and do things with data that we didn't consent to.

ANSWERS TO QUIZS

Fairness Quiz - Answers

Question	Answer
1	C
2	D
3	A
4	C
5	B

Robustness Quiz – Answers

Question	Answer
1	C
2	B
3	A
4	A and B
5	A

Explainability Quiz – Answers

Question	Answer
1	C
2	A
3	B
4	A
5	A

Privacy Quiz – Answers

Question	Answer
1	B and D
2	D
3	C
4	D
5	A

SUMMARY

Ethical AI pillars are:

- FAIRNESS: Lack of favoritism for one side or another
- ROBUSTNESS: Perform without failure under wide range of conditions
- PRIVACY: Ensure secure use of personal data and zero intrusion
- EXPLANABILITY: To explain. Present in understandable terms
- GOVERNANCE: To continuously monitor and enhance the AI models

AI Systems should be built such that they are fair and inclusive for all. Impartial behavior of AI model without any discrimination should be our goal. Unfair models can create improper outcomes. Unfair models can injure people, damage reputations and become legal liabilities. Recommended practices to ensure fairness are:

- Chalk out exact goals for ensuring fairness and inclusion
- Representative datasets should be used
- Analyze performance using Real-world scenarios

The accuracy of AI models is directly correlated to their vulnerability to small variations on the input dataset. Testing and benchmarking AI models against adversarial attacks is key to establishing trust in AI systems. We can alter the AI model development pipeline to include attack scenarios when training the AI model. And teach the defense mechanisms to the AI models at the time of training.

Recommended actions to ensure robustness

- Potential threats to the system
- Approach to overcome the threats
- Learn and stay updated on the latest security techniques

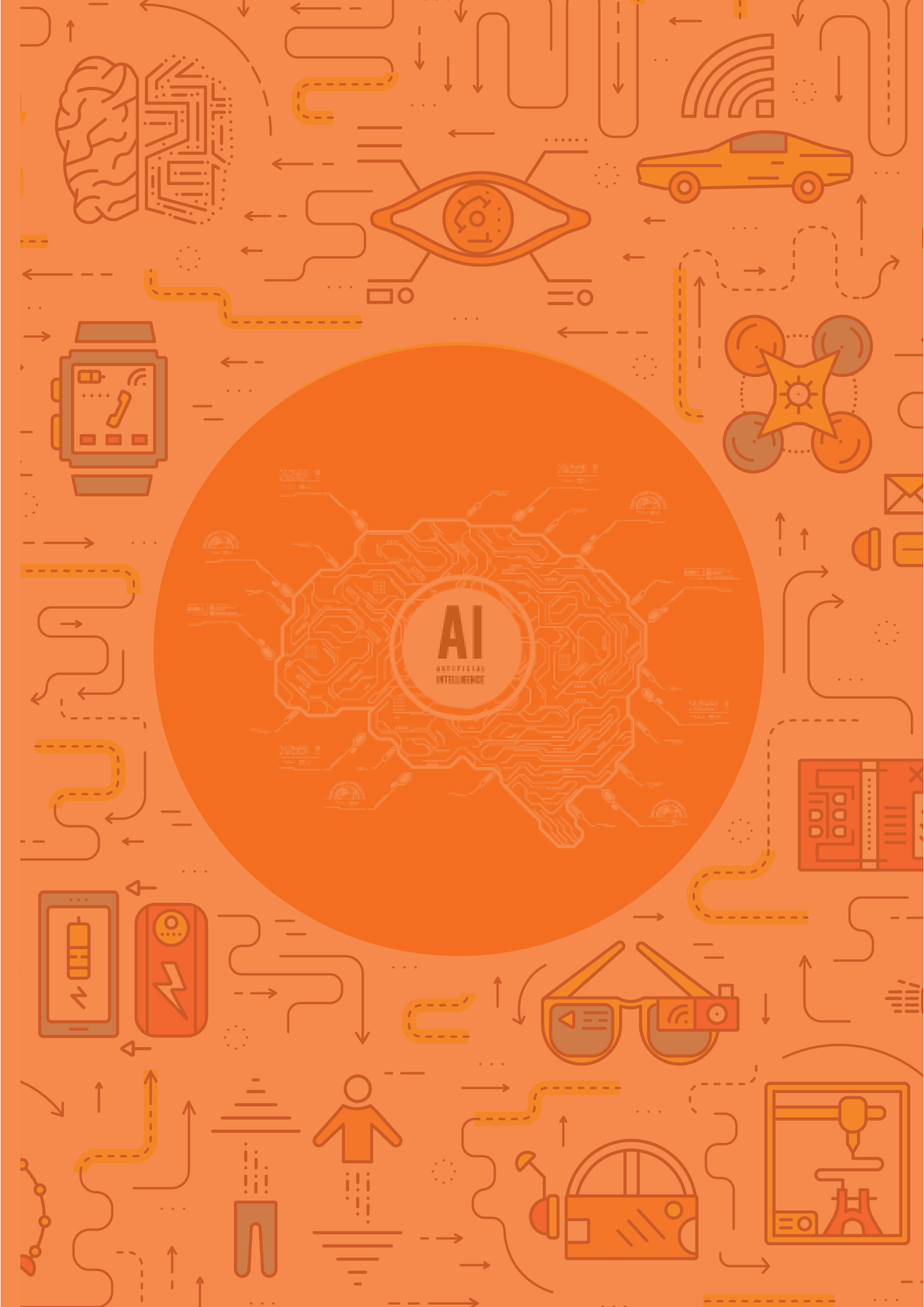
Explainability has led to the opening of this AI black box leading to transparency of the decision-making process of the AI models. Explainability is vital to being able to question, understand, and trust AI systems. Recommended actions for explainable ai

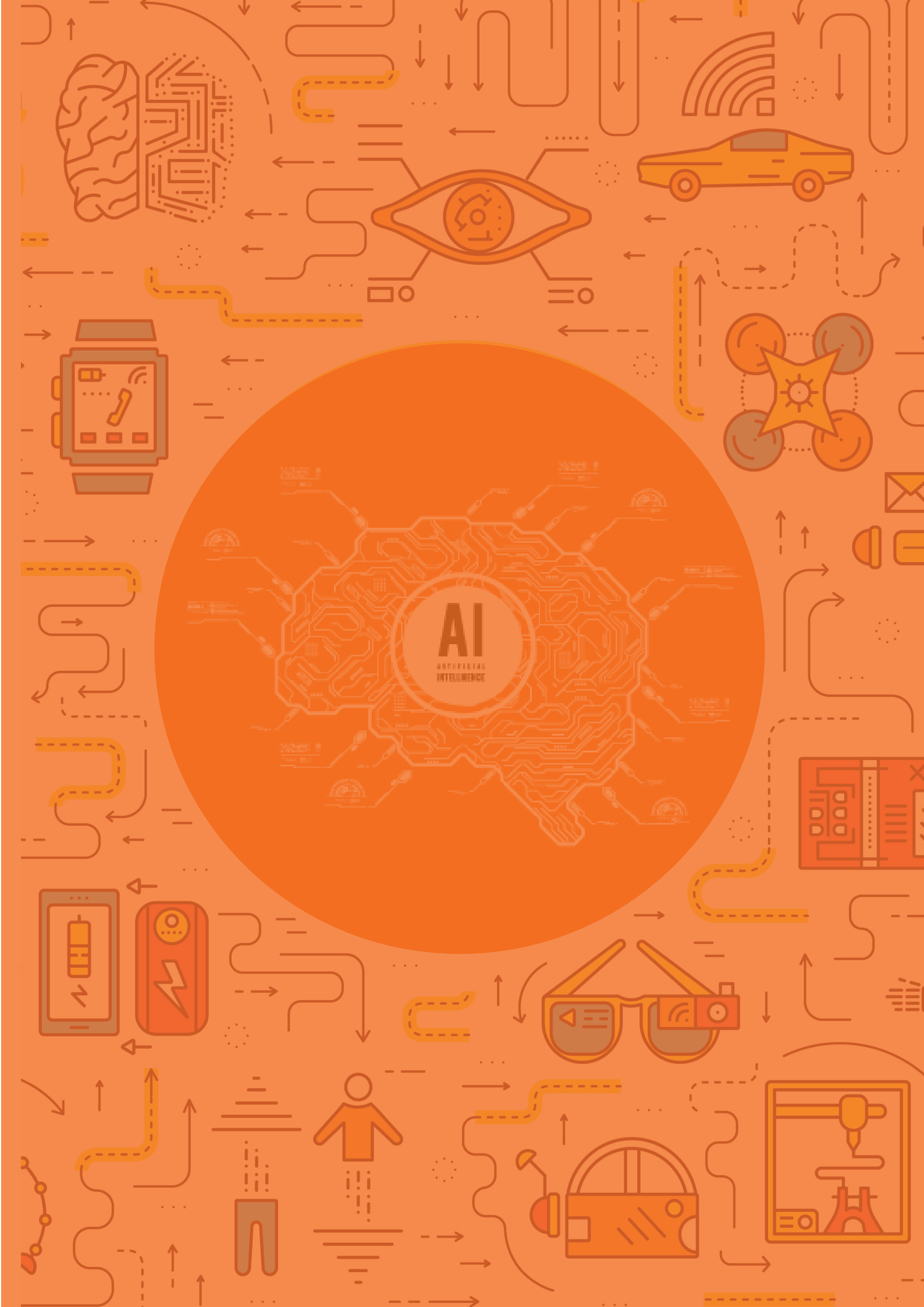
- Chalk out options to perform explainability
- Explainability is an important part of the user's experience
- Design ai model to be explainable
- Use a metric that shows the performance of the end goal
- Understand the trained model

When AI systems are to be built using sensitive data, we need to consider the privacy implications in using it. Legal & regulatory requirements, social norms and individual expectations should be obeyed. There are various rules and regulations across the globe to protect data privacy. Recommended actions to ensure data privacy

- Handle data responsibly
- On-device / different setting
- Processing of data
- Safeguard the privacy of ai models

Along with the AI model delivery lifecycle, we need to ensure that the AI system is kept accountable.


AI
ARTIFICIAL
INTELLIGENCE


AI
ARTIFICIAL
INTELLIGENCE

UNIT 3
MACHINE LEARNING

Key Learning Outcomes

At the end of this module, you will be able to:

a) Differentiate between artificial intelligence and machine learning.
b) Describe the machine learning process
c) List the type of tools used at different stages of the machine learning process
d) Explain how neural networks and deep learning works
e) Build machine learning solutions using MLaaS platforms

Topic Covered

Introduction to Machine Learning | Gathering Data | Preparing the Data Choosing the Model | Neural Networks and Deep Learning | Introduction to Maths for ML | Training | Evaluating | Parameter Tuning | Prediction | Machine Learning as a Service

3.1 INTRODUCTION TO MACHINE LEARNING

Artificial Intelligence has a wide variety of applications. As you engage with AI, you will learn that while some projects are relatively simpler, others can be far more complex. In the Air Draw exercise, we saw how the computer could mimic human action by the following movement. This is a very simple application compared to say a robot following you overcoming physical obstacles and finding their way in the real world.

The capabilities of machines to make decisions, learn and act like humans on their own, is a result of machine learning and deep learning. We shall learn more about it and its various facets in this Unit.

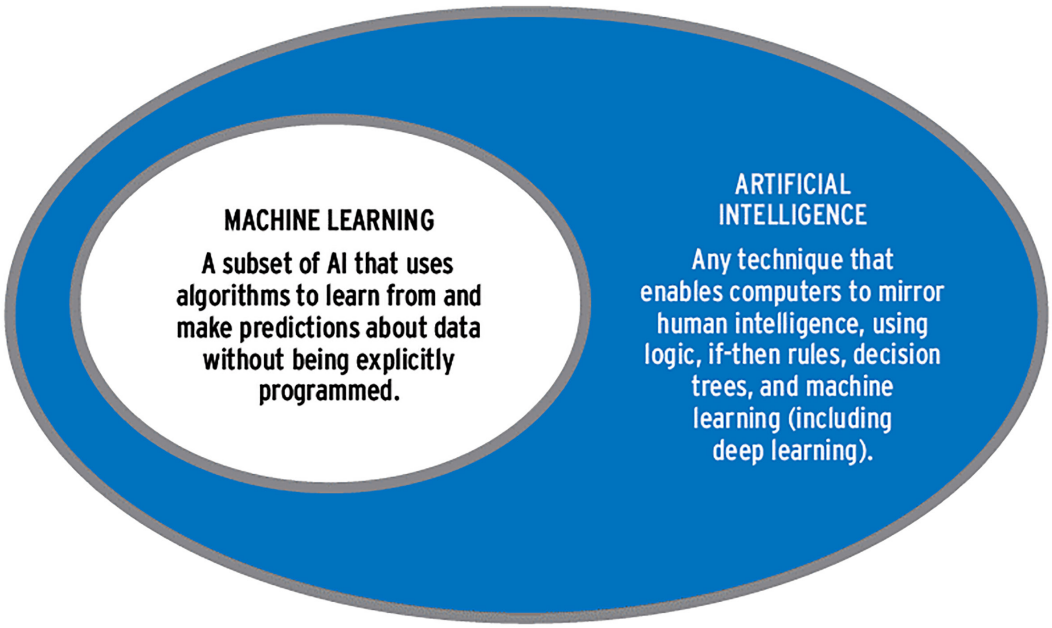
In this Unit, you will learn many new terms that form the core of the Artificial Intelligence building process. It will be an excellent idea to maintain a record of the new terms that you are learning and explore more about them on the Internet.

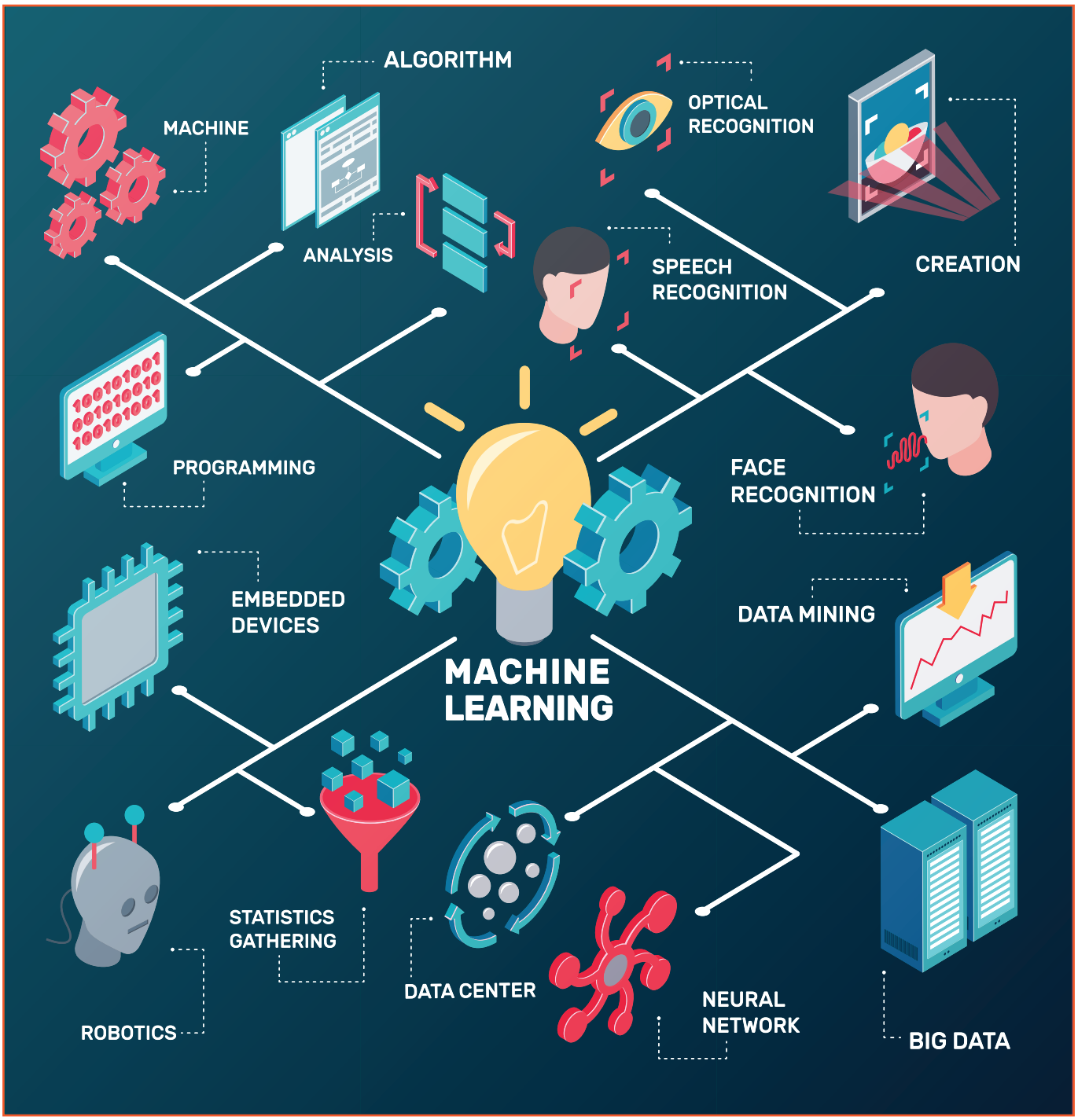
Let us begin now, by understanding what the difference between artificial intelligence and machine learning is.

The relation between Artificial Intelligence (AI) and Machine Learning (ML)

As we know, AI involves machines that can perform tasks akin to human intelligence. Typically this includes planning, recognizing objects and sounds, learning, predicting, understanding language, and problem-solving. There are many techniques through which it can do so. Machine Learning is the technique where algorithm or deep learning is used to build AI rather than the use of pure hard-coded programming.

Machine learning is a subset of AI (see diagram).





Arthur Samuel (an American pioneer in the field of artificial intelligence) coined the phrase Machine Learning in 1959, defining it as “the ability to learn without being explicitly programmed.” Technically we can build an AI without using machine learning, but this would involve writing millions of lines of code with complex rules and decision-trees.

So, by using the method of 'training' an algorithm so that machines can learn, we avoid hard-code software routines where specific instructions are given to accomplish a particular task.

'Training' as per ML involves feeding vast amounts of data to the algorithm and allowing the algorithm to adjust itself and improve.

Example:

Significant improvements in the field of Computer Vision (machine's ability to recognize an object in an image or video) have resulted from using Machine Learning. It involves data-sets of thousands or even millions of pictures that are tagged by humans. For example, you may have experienced captcha where we are asked to tag images with bicycles in them versus those that do not have them. Then, the algorithm attempts to create rules that build a model to accurately identify and tag a new picture as containing a bicycle or not as well as humans. Once required levels of accuracy are achieved by the machine to make the prediction, the machine has now 'learned' what a bicycle looks like.

Expert systems are another one, wherein computers are programmed with rules that enable them to behave like human experts in a particular domain, for example, an autopilot system flying a plane.

EXPLORE

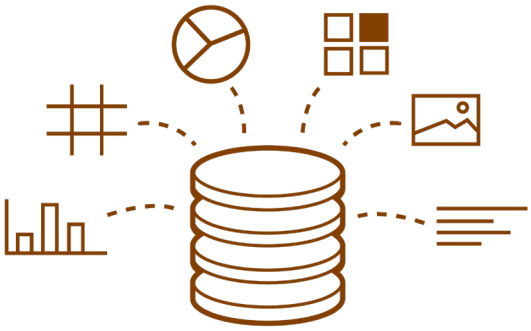
As an introduction to the key steps, you can go through the following video:

https://www.youtube.com/watch?time_continue=103&v=3TD0p_f6bZU&feature=emb_logo

Let us understand each of these steps.

3.2 GATHERING DATA

The machine learning process starts with data. Data could be in various forms, such as numbers, text, images, audios, videos, etc. Data gathering can be an expensive and time-consuming aspect of any machine learning project. Lack of care at this stage can result in failure.



There is a phrase, 'Garbage in, Garbage out' (GIGO) which means if the input is poor, the output will also be poor. GIGO is as valid for machine learning as with any other data-driven effort.

Hence it has to be done with planning, care and in an organized manner.

3.2.1 DATA GATHERING PROCESS

The data gathering process involves:

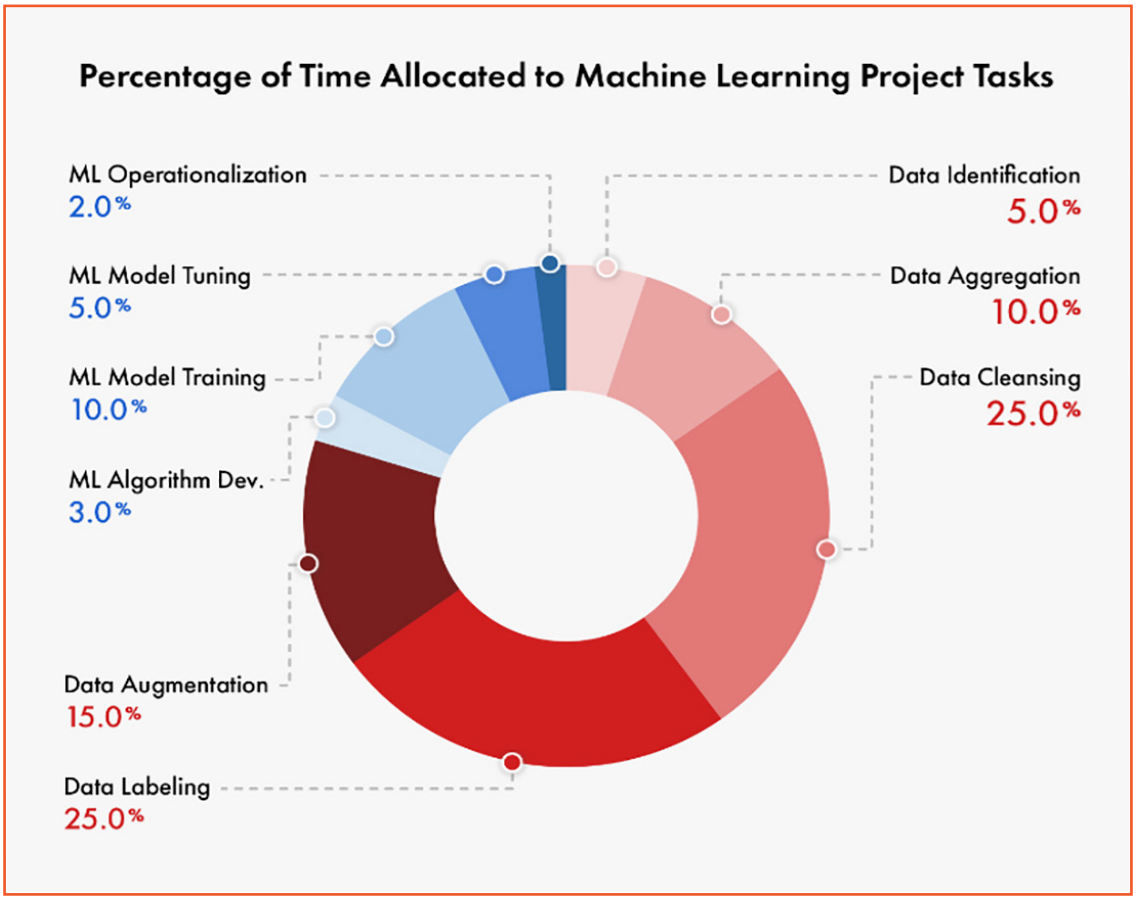
1. **Determining the Information to be collected:** Decide what topics the information will cover, who you want to collect it from, how much data you need and the quality of data.
2. **Setting up a timeframe for data collection:** Formulate a plan for how the data will be collected. Early enough in the planning process, we must establish a timeframe for data collection. There may also be some data that may be gathered continuously.
3. **Identifying data sources:** These could be primary sources (first-hand data from individuals, from observation, raw transactional data, etc.) or secondary sources (Existing data obtained from data repositories with organizations, online sources, offline sources, etc.). Check the authenticity of the source.
4. **Determining the data collection method:** As part of the plan, we must choose our data collection methods that will be used for collecting the data we need. We have to consider the type of information we need, the timeframe, the costs, etc. to select the right collection method along with any other important considerations that may apply.
5. **Collecting the data:** Implement the plan for collecting data and manage the channels for data collection to ensure they are being used as per plan. We can store and organize the data in databases on our data management platform.
6. **Verifying the data:** Once data is collected, we should verify that the data has been collected as per requirement and of the quality required.

3.2.2 DATA COLLECTION ACTIVITIES

Data Collection involves the following Activities:

- **Data Discovery** is the stage when one is searching for new datasets.
- **Data augmentation** is the phase where more data is sought from existing datasets to ensure adequate coverage.
- **Data generation** is used when adequate data is not available in existing datasets, but it is possible to generate crowdsourced or synthetic datasets instead.
- **Data Labeling** is about labeling the data with features, properties, characteristics, or classifications - that can be analyzed for patterns.

Here is a snapshot of the approximate distribution of the time spent in a machine learning project for the stages mentioned above.



(Source: https://www.cloudfactory.com/data-labeling-guide)

3.2.3 WAYS TO COLLECT DATA

Data collection methods could be divided into primary and secondary methods:

- **Methods of collecting primary data:**

1. Direct observation
2. Questioning via interviews, online surveys, offline surveys, etc
3. Online tracking using software tools
4. Data tracking from transaction records

- **Methods of collecting secondary data:**

1. Published sources such as magazines, journals, reports, open-source databases, public domain websites, etc.
2. Unpublished sources such as records maintained by various government and private offices, research studies were done by scholar students or institutions, reports prepared by private entities for internal use, etc.

Data coverage

Data coverage is the distribution of data across target classes or values, as well as across other variables that contribute to variation. Target classes refer to what you are looking to detect or predict. For example, if we are building a model to detect if a machine is performing normally or is displaying a particular kind of error, then 'Normal' and 'Error' are the target classes.

Data has to be collected keeping in mind the sources, variation in data and the different variables.

The purpose of data coverage planning is to make sure that we collect sufficient data to capture and overcome the variation embedded in what we are trying to measure. It requires a statistically significant number of observations taken from a sufficient number of combinations of different target classes and variables.

Iterative Approach to data collection

Iterative mean repeating. In data collection, 'iterative' means collecting data in a continuously repeating process. If we collect too little data, we may have to go back and collect more of it all over again. But collecting too much may be very expensive and time-consuming, and mostly time and budgets are limited. That's why an iterative approach to data collection is found useful. In an iterative approach, a limited set of data is collected at each stage required for answering a specific set of questions. The processing and results at each stage justify the next round of data collection, where more complexities may be added. This process is repeated multiple times, till all objectives are met.

3.2.4 BIG DATA

Humans generate a massive amount of data these days, in a wide variety of areas.

Consider a few examples:

- We send more than 16 million text messages every minute
- More than 100 million spam emails are sent every minute
- We upload more than a billion photos to Google Photos every day
- Around 500 million tweets are sent in a day
- Around 294 billion emails are sent in a day
- Around 65 billion messages are sent on WhatsApp in a day
- Around 5 billion searches are done in a day
- As of June 2019, there were over 4.5 billion internet users in the world

Such data is called big data.

This availability of massive datasets is one of the key reasons that has had machine learning possible. The ability of an algorithm to generate useful solutions from the data is dependent on the availability of a lot of data. More data will provide more opportunity for an algorithm to find associations. As more associations are found, the accuracy of the predictions will be higher. Just like with humans, the more experience a computer has, the better the results will be

However, with respect to big data, good data is certainly better than more data. There are limits to what a clever machine learning algorithm can do with unstructured or poor-quality data. There is no substitute for collecting information correctly at the outset.

3.2.5 VARIOUS TOOLS FOR DATA COLLECTION

Listed below are tools that enable users to source large volumes of raw data quickly.

- **Data Scraping Tools:** Web scraping describes the automated, programmatic use of an application to extract data or perform actions that users would usually perform manually, such as social media posts or images. The following companies offer tools to extract data from the web:
  - **Octoparse:** A web scraping tool that lets users obtain public data without coding.
  - **Mozenda:** A tool that allows people to extract unstructured web data without scripts or developers.

- **Synthetic Data Generators:** Synthetic data can also be programmatically generated to obtain large sample sizes of data. This data can then be used to train neural networks. Below are a few tools for generating synthetic datasets:
  - **Pydbgen:** This is a Python library that can be used to generate a sizeable synthetic database as specified by the user. For example, pydbgen can generate a dataset of random names, credit card numbers, company names and more.
  - **Mockaroo:** Mockaroo is a data generator tool that lets users create custom CSV, SQL, JSON and Excel datasets to test and demo software.

- **Data Augmentation Tools:** In some cases, data augmentation may be used to expand the size of an existing dataset without gathering more data. For instance, an image dataset can be augmented by rotating, cropping, or altering the lighting conditions in the original files.
  - **OpenCV:** This Python library includes image augmentation functions. For example, it includes features for bounding boxes, scaling, cropping, rotation, filters, blur, translation, etc.

Scikit-image: Similar to OpenCV, this is a collection of algorithms for image processing available free of charge and free of restriction. Scikit-image also has features to convert from another color space to another, resizing and rotating, erosion and dilation, filters, and much more.

- **Data Labeling Tools:** If you're building a machine learning model, chances are you're going to need data labelling tools to put together datasets and ensure high-quality data production quickly. It includes data tagging, classification, moderation, transcription, or processing.

Given below are some open-source data labeling tools for various types of data.

EXPLORE

Image & Video Labeling Tools

https://recogito.github.io/annotorious/

https://labelbox.com/

http://labelme.csail.mit.edu/Release3.0/

https://sloth.readthedocs.io/en/latest/

EXPLORE

Audio Labeling Tools

https://www.fon.hum.uva.nl/praat/

https://github.com/felixbur/Speechalyzer

https://aubio.org/

EXPLORE

Text Labeling Tools

https://rasa.com/docs/rasa/nlu/about/

https://github.com/dennybritz/bella

https://stanfordnlp.github.io/CoreNLP/

https://www.tagtog.net/

- **Open-Source Datasets:** Another way to obtain raw data for machine learning is to obtain pre-built, publicly available datasets on the Internet. There are thousands of publicly available datasets spanning a wide range of industries and use cases.

Here are a few of the Indian Government and open-source datasets that you can use. There are many more that you can search for on the Internet.

EXPLORE

Indian Government:

https://data.gov.in/

http://mospi.nic.in/data

https://bhuvan-app3.nrsc.gov.in/data/download/index.php

http://www.surveyofindia.gov.in/

https://dbie.rbi.org.in/DBIE/dbie.rbi?site=home

EXPLORE

Open Source:

https://www.kaggle.com/datasets

https://www.imf.org/en/Data

https://data.worldbank.org/

https://trends.google.com/trends/?q=google&ctab=0&geo=all&date=all&sort=0

3.3 PREPARING THE DATA

Once the data has been collected, it is in a raw form. Raw data is collected from various sources and is usually unsuitable for analysis by itself. For example:

- There might be many duplicate entries.
- The data may have typographical errors.
- There may be missing data.
- The data may be available in different formats.

With an uncleaned dataset, no matter what type of algorithm you try, you will never get accurate results. That is why data scientists spend a considerable amount of time on data cleaning; further, it has to be labeled. This is call pre-processing.

The steps and techniques for data pre-processing will vary for each dataset. But the following steps could be used as a standard approach for any dataset type:

- Identifying relevant data and removing irrelevant data
- Fix Irregular cardinality and structural errors
- Outliers
- Missing data treatment
- Data Transformation

Post the data pre-processing, comes the final data preparation phase of Data Splitting.

3.3.1 IDENTIFYING RELEVANT DATA AND REMOVING IRRELEVANT DATA

Mainly two checks should be carried out for identifying irrelevant data:

1. Duplicate entries

Data is usually collected from many sources and joined together to form one dataset. This might lead to the duplicity of observations. This might not be a problem if the observations are repeated a few times, particularly in the case of Big Data. However, if observation(s) are repeated far too many times, it could lead to an erroneous behaviour. Hence, it is better to remove duplicate observations to have a cleaner dataset.

2. Irrelevant observations:

The dataset would contain observations which might not be useful for your specific task. For example, if you were analyzing the shopping behaviour of women only, you would not require observations for men in your dataset. Similarly, your data might have a whole column of data that is not useful in making predictions and could be dropped.

3.3.2 FIXING IRREGULAR CARDINALITY AND STRUCTURAL ERRORS

Cardinality is the number of elements, or members, in a set of data, i.e. the number of different values there are for a particular feature.

- If in a dataset any feature or column had all entries made with the same type of value (or class), it will have a cardinality of 1. Then that feature or column does not have any effect of the model or predictions. Hence it is suggested that such columns be dropped. The same is the case if the variance is too low to have any impact. I,e, the difference between the highest and lowest value is very small. So it is if the cardinality is 0, i.e. there are no values available in that column.
- The columns could have a high number of types of values (or classes) due to typo-errors or inconsistent capitalization. For example, the gender column might have many classes like male, female, m, f, M, and F, these represent only two levels — Male and Female. Such classes should be mapped to proper levels and the rest of the levels should be dropped. Whitespaces in some of the values entered, some numeral values entered as text or inconsistency in date formats could be some more reasons for an irregularly high number of classes.

3.3.4 DEALING WITH OUTLIERS

An outlier is an element of a dataset that distinctly stands out from the rest of the data. The analyst will have to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

Two activities are essential for characterizing a set of data:

- Examining the overall shape of the graphed data to identify important features, including symmetry and departures, from assumptions.
- Examining the data for unusual observations that are far removed from the mass of data. These points are often referred to as outliers. There are techniques for identifying outliers, such as graphical techniques of scatter plots and box plots, along with an analytic procedure for detecting outliers when the distribution is normal (Grubbs' Test).

Once the outliers are identified, then depending on the objectives, you must decide whether the outliers must be included or excluded.

You should not drop any observations without careful consideration. You must consider the possibility that such information can be valuable to understand the unusual behaviour or anomalies in the data. Do not exclude unless you are sure that the outlier may be due to measurement error or if the presence of outliers influences the fit of a model.

3.3.5 MISSING DATA TREATMENT

Real-world data would almost always have missing values. There can be many reasons for missing data, such as data entry errors or data collection problems. Irrespective of the reasons, it is vital to handle missing data because any statistical results based on a dataset with non-random missing values could be biased. Also, many ML algorithms do not support data with missing values.

There are three types of missing data:

- **MCAR: Missing** Completely At Random. It is the highest level of randomness. This means that the missing values in any features are not dependent on any other features values. This is the desirable scenario in case of missing data.
- **MAR: Missing At Random.** This means that the missing values in any feature or column are dependent on the values of other features or columns.
- **MNAR**: Missing Not At Random. This is when the value that missing is related to the reason it's missing). For example, people who are depressed failed to fill the depression column because it depressed them further to acknowledge it. MNAR is a more serious issue and in this case, it might be wise to check the data gathering process further and try to understand why the information is missing.

We can check for null values in a dataset using some software.

After detecting the missing values, the following course of actions can be followed:

- **Ignore the missing values:** Less than 10% missing data for a single case or observation can generally be ignored, unless the missing data is a MAR or MNAR. However, the number of complete cases, i.e. the observations with no missing data, must be sufficient for the selected analysis technique if removing the observations with missing data.

- **Drop the missing values:** If the data is MCAR or MAR and the number of missing values in a feature is very high, then that feature should be left out of the analysis. If missing data for a particular feature or sample is more than 5%, then you probably should leave that feature or sample out. If the cases or observations have missing values for target variables(s), it is advisable to delete the dependent variable(s) to avoid any artificial increase in relationships with independent variables.
- **Case Deletion:** In this method, cases (the rows) which have missing values for one or more features are deleted. If the cases having missing values are small in number, it is better to drop them. Though this is an easy approach, it might lead to a significant decrease in the sample size. Also, the data may not always be missing completely at random. This may lead to biased estimation of parameters.
- **Imputation:** Imputation is the process of substituting the missing data by some statistical methods. Imputation is useful in the sense that it preserves all cases by replacing missing data with an estimated value based on other available information. But imputation methods should be used carefully as most of them introduce a large amount of bias and reduce variance in the dataset.

3.3.6 DATA TRANSFORMATION

In this final pre-processing phase, a data scientist transforms or consolidates data into a form appropriate for machine learning. This phase is also called feature engineering.

Data can be transformed through scaling (normalization), attribute decompositions, and attribute aggregations.

- **Scaling.** Data may have numeric attributes that span different ranges, for example, millimeters, meters, and kilometers. Scaling is about converting these attributes so that they will have the same scale, such as between 0 and 1, or 1 and 10 for the smallest and biggest value for an attribute.
- **Decomposition.** Sometimes finding patterns in data with features representing complex concepts is more difficult. Decomposition technique can be applied in this case. During decomposition, you can convert higher-level features into lower-level ones. In other words, new features based on the existing ones are being added. Decomposition is mostly used in time series analysis. For example, In order to estimate the monthly demand for computers, a market research analyst converts the data representing the quarterly demand.

- **Aggregation.** Aggregation combines several features into one feature that represents them all. For example, if you have collected basic information about your customers, particularly their age in order to develop a demographic segmentation strategy. You could need to distribute them into categories of age, such as 11-20, 21-30, 31-40, etc. for this you would use aggregation to create large-scale features based on small-scale ones. This technique will allow you to reduce the size of a dataset without the loss of information.

To learn more, you can go through the following link:

https://docs.aws.amazon.com/machine-learning/latest/dg/data-transformations-for-machine-learning.html

3.3.7 DATASET SPLITTING

A dataset used for machine learning should be partitioned into three subsets — training, test, and validation sets.

- **Training set.** A data scientist uses a training set to train a model and define its optimal parameters — parameters it has to learn from data.
- **Test set.** A test set is needed for an evaluation of the trained model and its capability for generalization. The latter means a model's ability to identify patterns in new unseen data after having been trained over the training data. It's crucial to use different subsets for training and testing to avoid model overfitting, which is the incapacity for generalization we mentioned above.
- **Validation set.** The purpose of a validation set is to tweak a model's hyperparameters — higher-level structural settings, that can't be directly learned from data. These settings can express, for instance, how complex a model is and how fast it finds patterns in data.

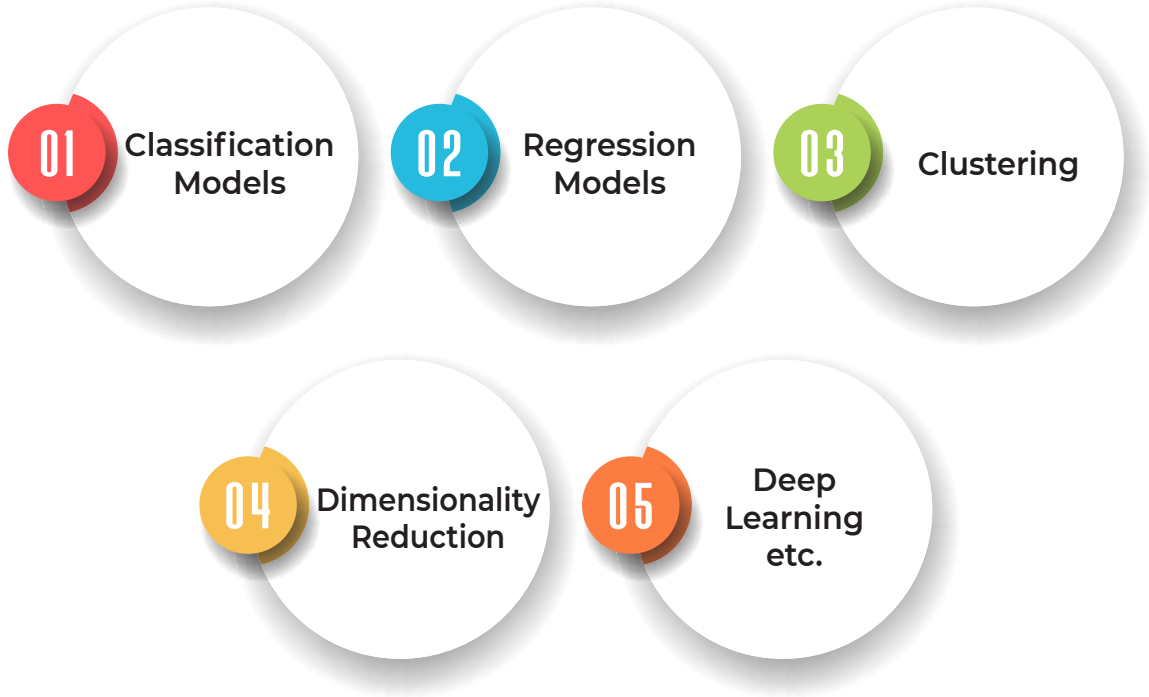
The proportion of training and a test set is usually 80 to 20 percent, respectively. A training set is then split again, and its 20 percent will be used to form a validation set. At the same time, some machine learning practitioners suggest using 66 percent of data for training and 33 percent for testing. Size of each subset depends on the total dataset size.

The more training data you use, the better the potential model will perform. Consequently, more results of model testing data leads to better model performance and generalization capability.

3.4 CHOOSING A MODEL

The next step is selecting an appropriate model. A machine learning model is a set of rules that recognize certain types of patterns. For example, you can train a Machine learning algorithm with a set of data, that can reason over and learn from and become the final model which can make predictions over data that it hasn't seen before.

Based on the type of tasks, we can classify machine learning models in the following types:



3.4.1 CLASSIFICATION

In the context of machine learning, classification is the task of predicting the type or class of an object within a finite number of options. The output variable for classification is always a categorical variable. For example, predicting an email is spam or not is a standard binary classification task.

Classification Model could be Binary or Multi-class.

ML models for binary classification problems predict a binary outcome (one of two possible classes). Examples of Binary Classification Problems

- "Is this email spam or not spam?"
- "Will the customer buy this product?"
- "Is this comment written by a human or a robot?"

ML models for multi-class classification problems allow you to generate predictions for multiple classes (predict one of more than two outcomes). Examples of Multi-class Problems

- "Is this product a book, movie, or clothing?"
- "Is this movie a romantic comedy, documentary, or thriller?"
- "Which category of products is most interesting to this customer?"

We will learn more about this in the UNIT on Algorithms.

3.4.2 REGRESSION

In the machine, learning regression is a set of problems where the output variable can take continuous values. For example, predicting the airline price can be considered as a standard regression task.

Examples of Regression Problems

- "What will the temperature be in Mumbai tomorrow?"
- "How many copies of this book will sell?"
- "What price will this car sell for?"

We will learn more about this in the Unit on Algorithms.

Clustering

In simple words, clustering is the task of grouping similar objects together. It helps to identify similar objects automatically without manual intervention. We can not build effective supervised machine learning models (models that need to be trained with manually curated or labeled data) without homogeneous data. Clustering helps us achieve this more smartly.

We will learn more about this in the Unit on Algorithms.

Dimensionality Reduction

Dimensionality is the number of predictor variables used to predict the independent variable or target.often in the real world datasets the number of variables is too high. Too many variables also bring the curse of overfitting to the models. In practice, among these large numbers of variables, not all variables contribute equally towards the goal and in a large number of cases, we can preserve variances with a lesser number of variables.

Deep Learning

Deep learning is a model of machine learning which deals with neural networks. We shall learn more about it in the next section

3.5 NEURAL NETWORKS AND DEEP LEARNING

3.5.1 NEURAL NETWORKS

Artificial Neural Networks (ANN) are inspired by the central nervous system that we possess. The essential concept is that a network of artificial neurons built out of interconnected threshold switches can learn to recognize patterns in the same way that an animal brain and nervous system (including the retina) does.

ANN's have been successfully applied to several problem domains:

- To classify data by recognizing patterns.
- To detect anomalies or novelties, when test data does not match the usual patterns.
- To process signals, for example, by filtering, separating or compressing.
- To find a function that closely matches the training data (approximate a target function) so that it can predict almost correctly.

Effectively, Neural networks can:

01 Identify faces

02 Recognize speech

03 Read your handwriting (mine perhaps not)

04 Translate texts

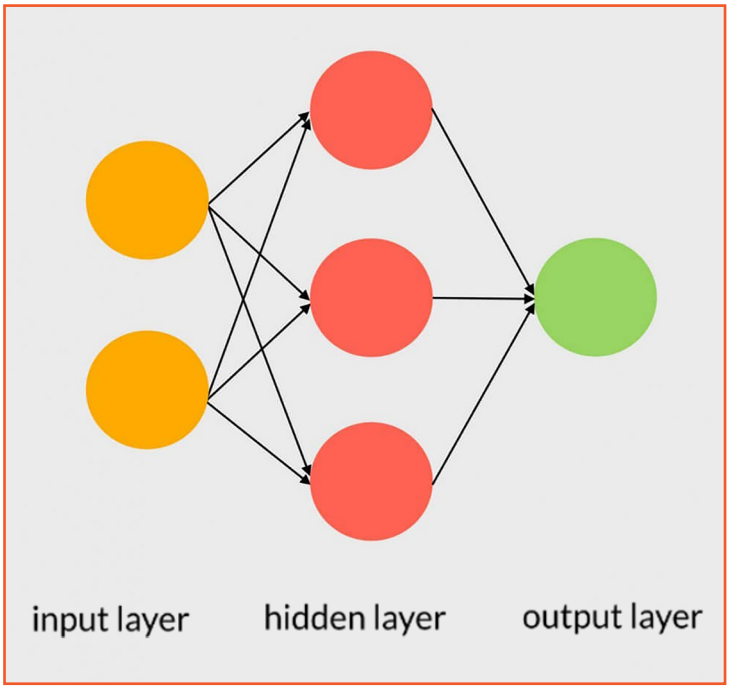
05 Play games (typically board games or card games)

06 Control autonomous vehicles and robots

07 And a few other things

The topology of a neural network

There are many ways of knitting the nodes of a neural network together, and each way results in more or less complex behavior. Possibly the simplest of all topologies is the feed-forward network. Signals flow in one direction only; there is never any loop in the signal paths.



A feed-forward neural network

Typically, ANN's have a layered structure.

- The input layer picks up the input signals
- Then the input layer passes them on to the next layer, the so-called 'hidden' layer. (Actually, there may be more than one hidden layer in a neural network.)
- Last comes the output layer that delivers the result.

Basically, the learning occurs by the strengthening of the connection between two neurons when they are active during training. This is mostly a matter of increasing the weight values of the connections between the neurons. For this a rule called back propagation of error, backprop, or BP is used.

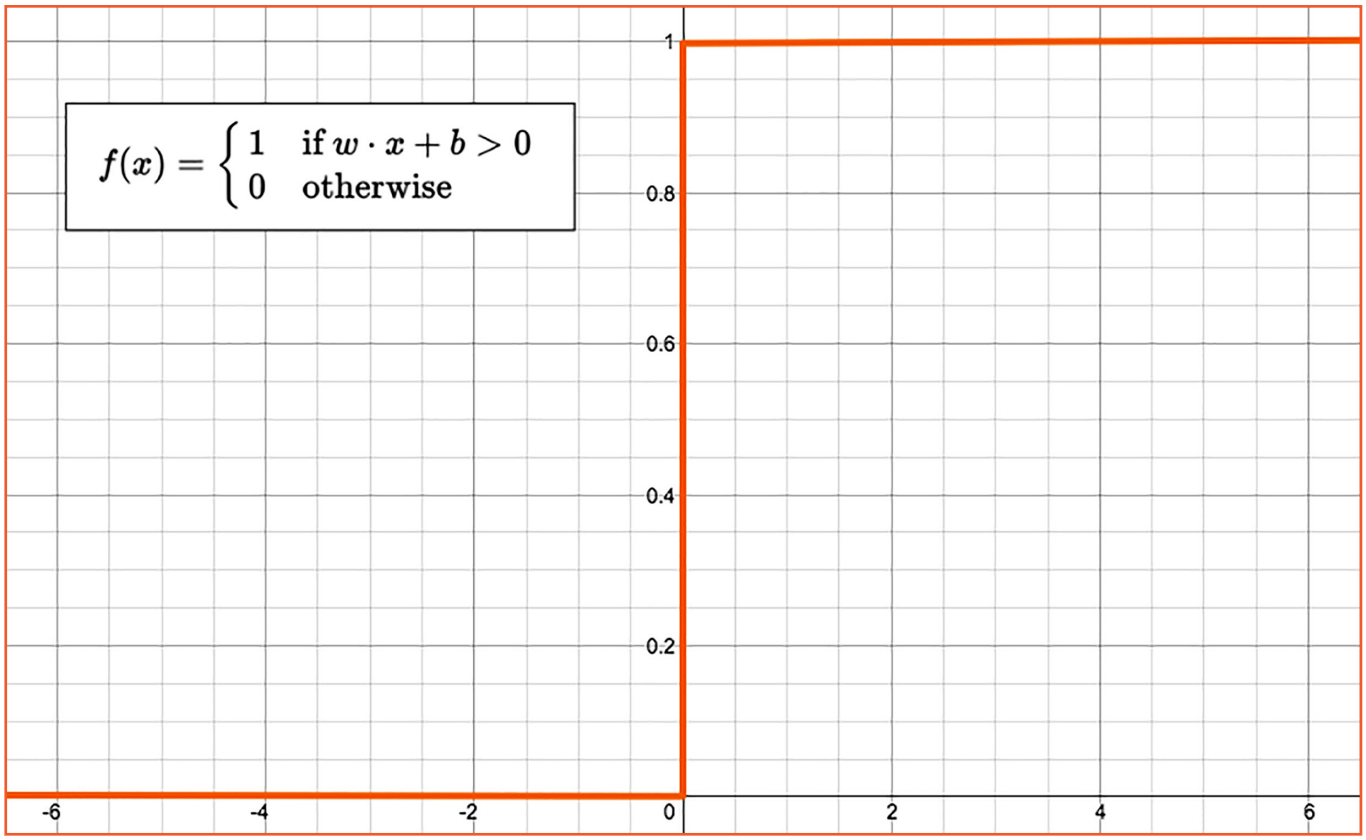
Each has a propagation function that transforms the outputs of the connected neurons, often with a weighted sum. The output of the propagation function goes to an activation function. This activation function fires when input exceeds a certain threshold value.

Perceptrons

In the 1940s and 1950s, a step activation function was used by artificial neurons, called perceptrons. Perceptrons can be said to be the building blocks in a single layer of a neural network. They are made up of four different parts:

- Input Values or One Input Layer
- Weights and Bias
- Net sum
- Activation function

A step activation function is a simple binary function that has only two possible results, which means that it takes several real-valued inputs and gives a single binary output.



The function in the figure given above is called the Heaviside Step function. This function returns 1 if the input is positive or zero, and 0 for any negative input. A neuron whose activation function is a function like this is called a perceptron.

In the perceptron model, every input x has weight w associated with it. The weights indicate the importance of the input in the decision-making process. A threshold value decides the model output. If the weighted sum of the inputs is greater than the threshold value, the output will be 1 else output will be 0. In other words, the model will fire if the weighted sum is greater than the threshold.

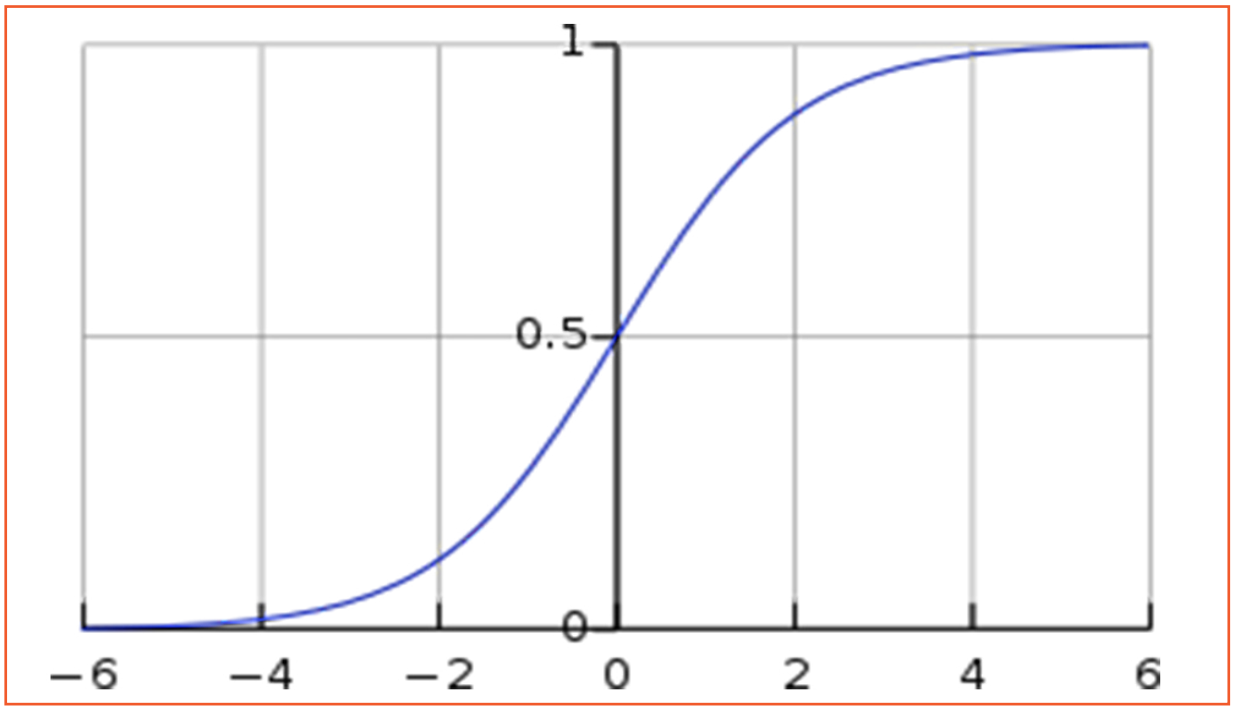
A neural network, which is made up of perceptrons, can be perceived as a complex logical statement (neural network) made up of very simple logical statements.

Some other activation functions

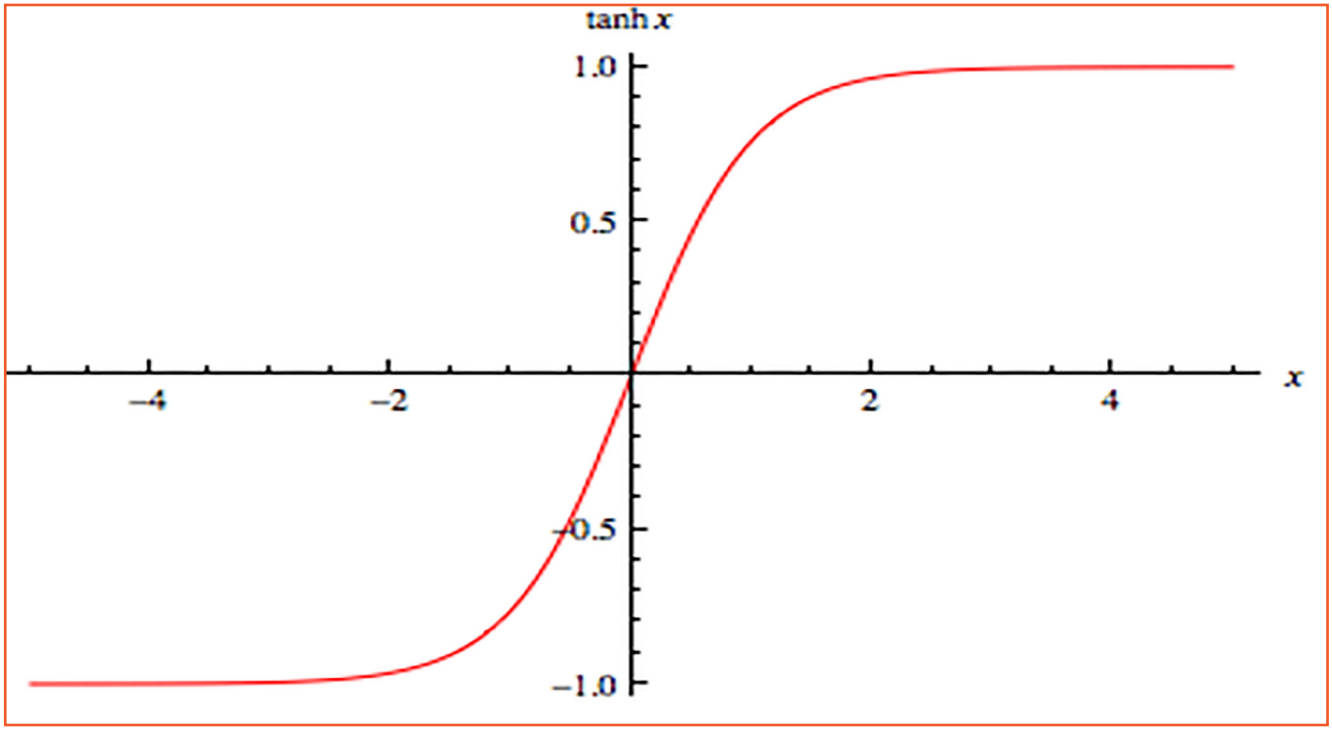
Modern neural networks may reference perceptrons but have smooth activation functions, such as:

The logistic or sigmoid function: Sigmoid neurons are similar to perceptrons, but they are slightly modified such that the output from the sigmoid neuron is much smoother than the step functional output from a perceptron. A small change in the input causes a small change in the output as opposed to the stepped output. There are many functions with the characteristic of an "S" shaped curve known as sigmoid functions.

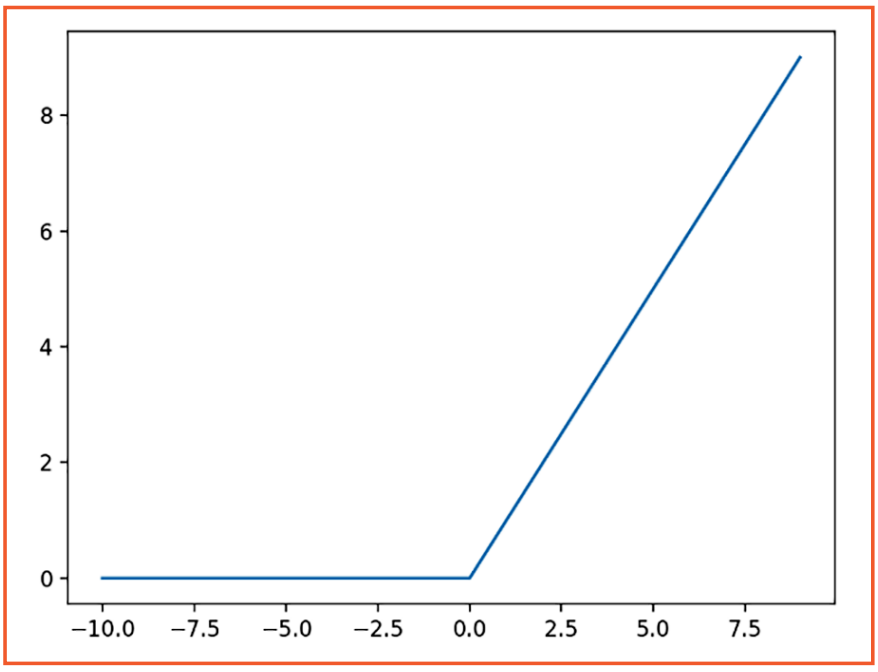
For example:



The Hyperbolic Tangent: In neural networks, as an alternative to sigmoid function, the hyperbolic tangent function could be used as an activation function. When you backpropagate, the derivative of activation function would be involved in the calculation for error effects on weights. The derivative of the hyperbolic tangent function has a simple form just like a sigmoid function.

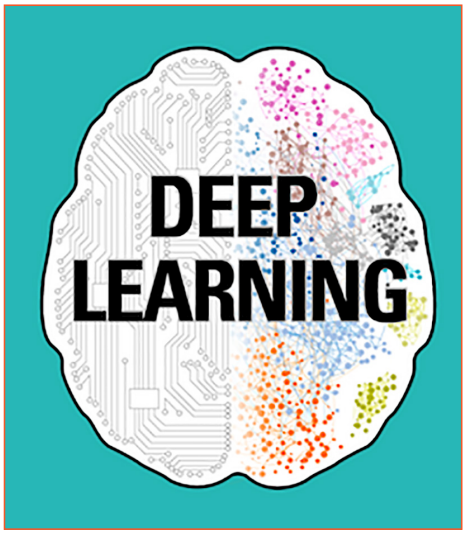


The Rectified Linear Unit (ReLU): The rectifier function mostly looks and acts like a linear activation function. In general, a neural network is easier to optimize when its behavior is linear or close to linear. Hence, ReLU is usually the best choice for fast convergence, although it has an issue of neurons 'dying' during training if the learning rate is set too high.

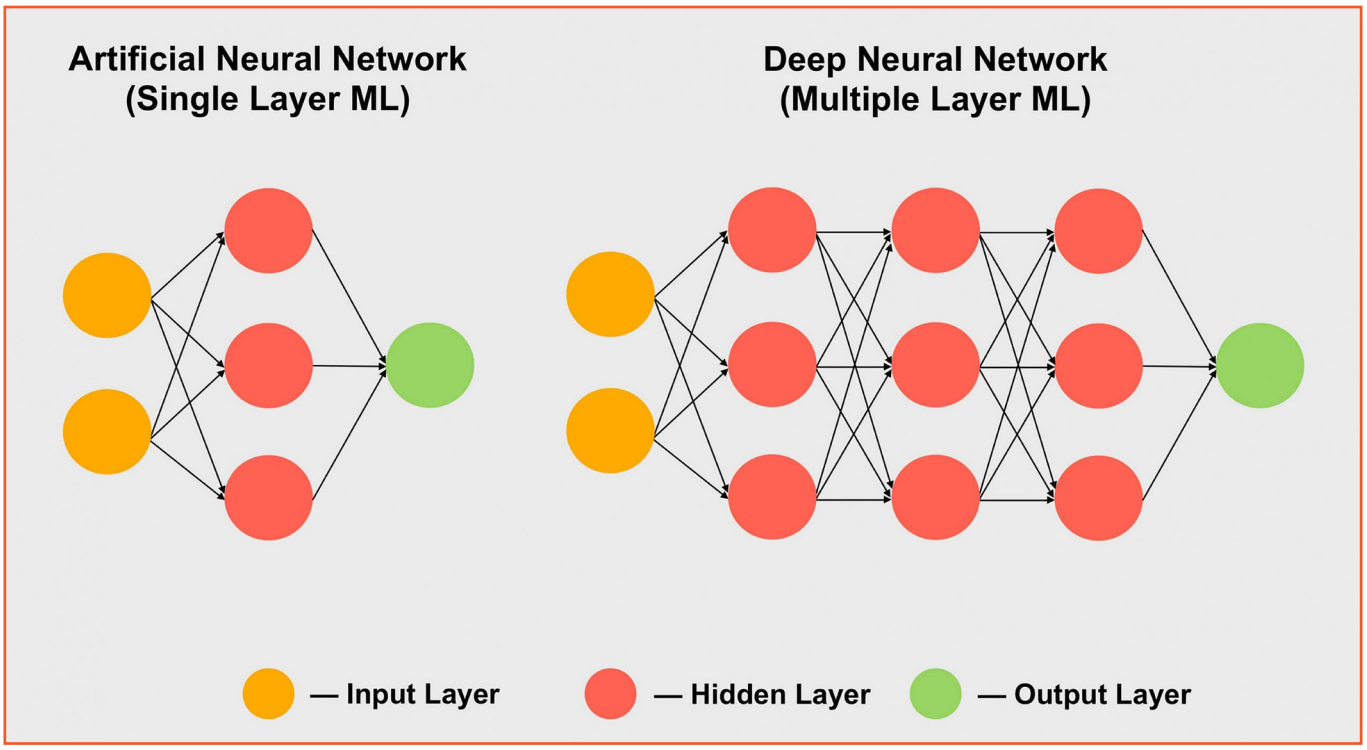


3.5.2 DEEP LEARNING

Now that you know something about machine learning and neural networks, it's only a small step to understanding the nature of deep learning algorithms.



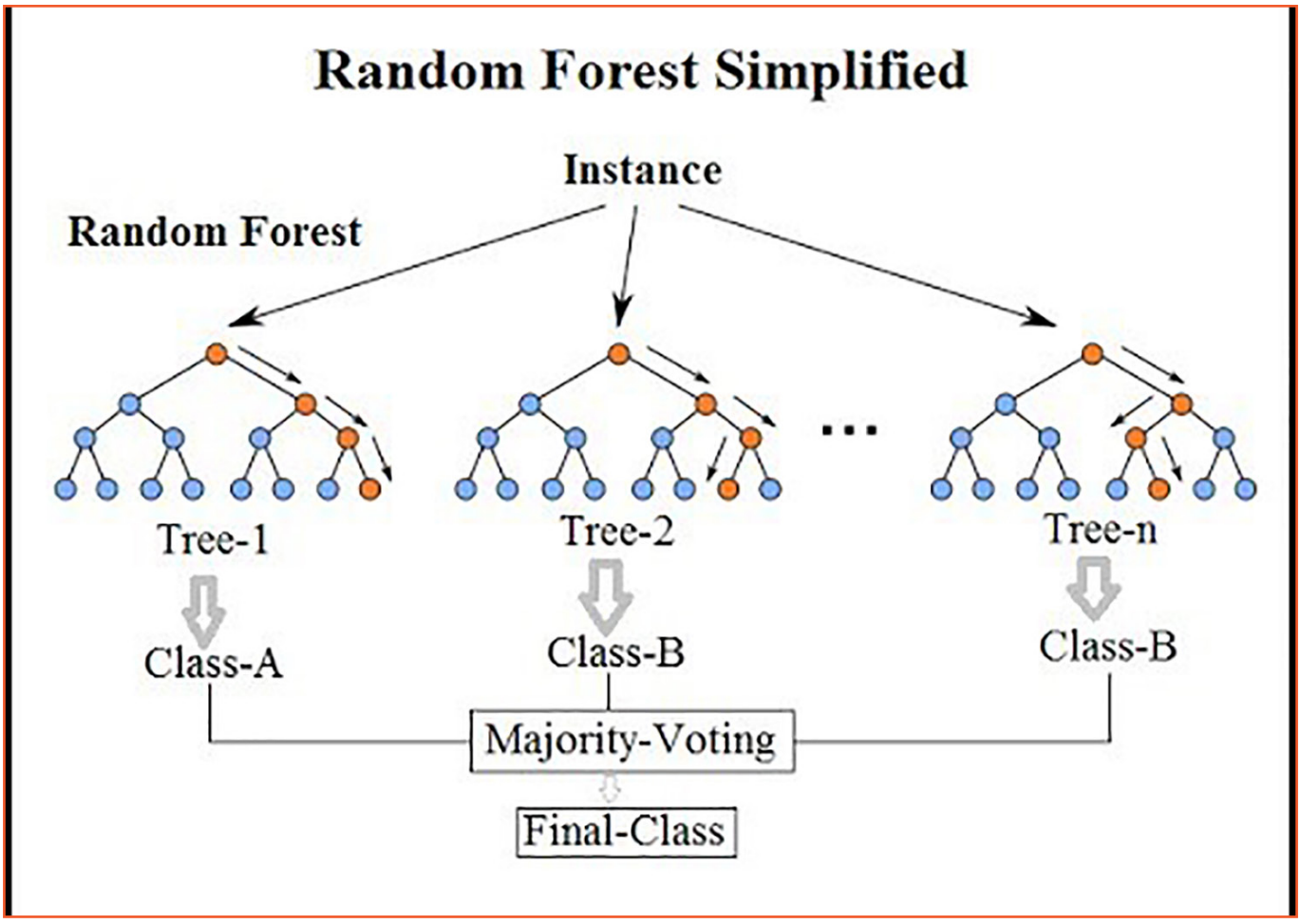
Deep learning is a subset of AI and machine learning. These constructs consist of multiple layers of ML algorithms. Thus, they are often referred to as 'Deep Neural Networks' (DNN). Input is passed through the layers, with each adding qualifiers or tags. So deep learning does not require pre-classified data to make interpretations.



The layers are of alternating linear and nonlinear processing units and are trained using large-scale algorithms and massive amounts of training data. A deep neural network might have 10 to 20 hidden layers, whereas a typical neural network may have only a few.

The more layers in the network, the more characteristics it can recognize. Unfortunately, the more layers in the network, the longer it will take to calculate, and the harder it will be to train.

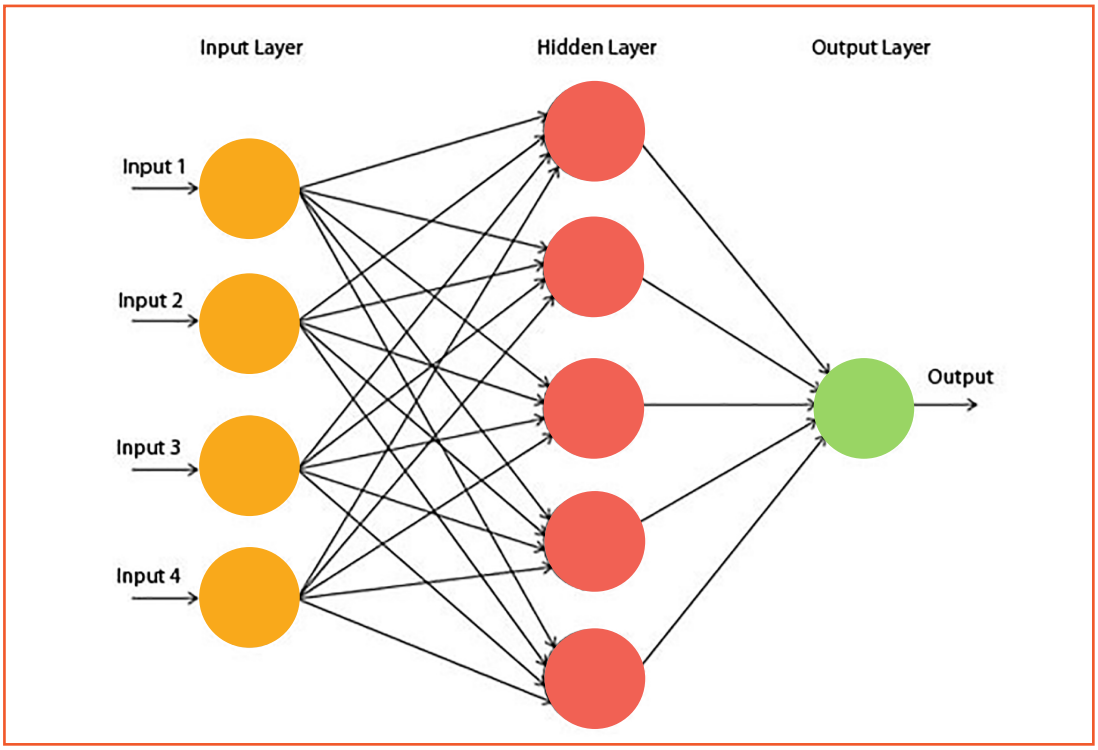
Another kind of deep learning algorithm is Random Decision Forests (RDFs). Again, they are constructed from many layers, but instead of neurons, the RDF is constructed from decision trees and outputs a statistical average (mode or mean) of the predictions of the individual trees. The randomized aspects of RDFs are the use of bootstrap aggregation (bagging) for individual trees and taking random subsets of the features.



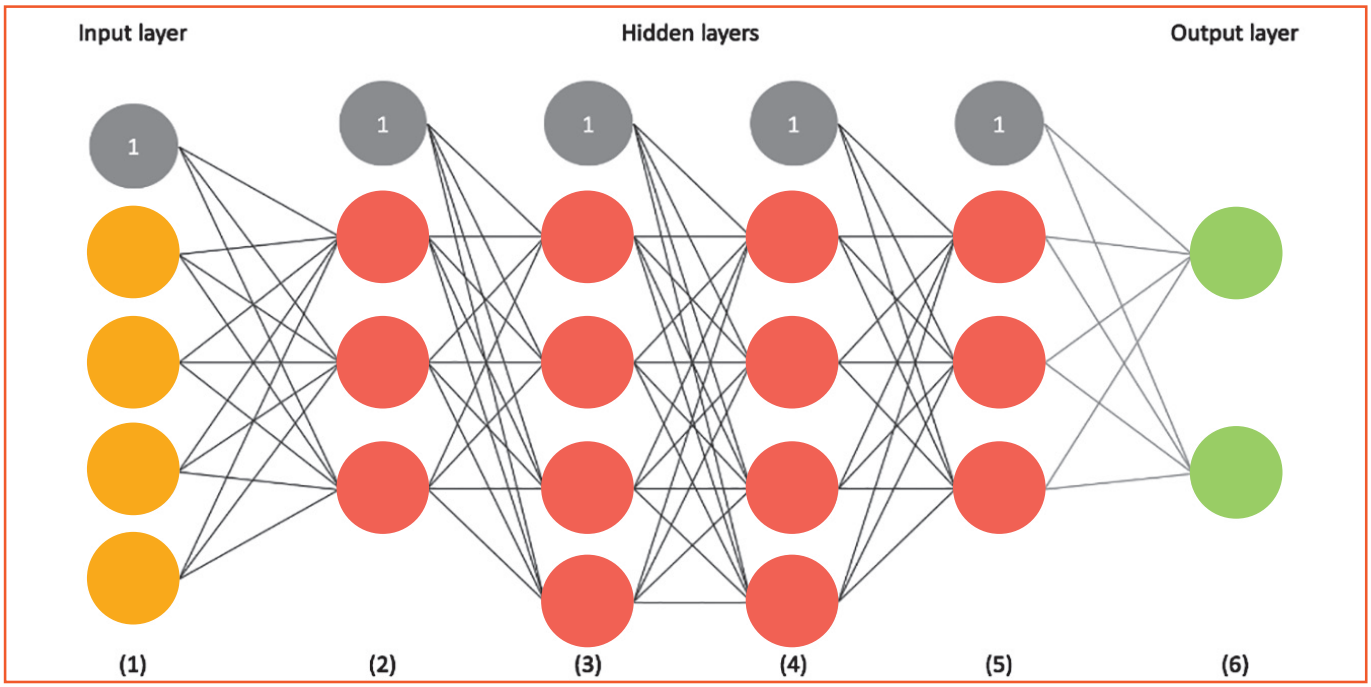
There are many ways to approach deep learning, but none are perfect, at least not yet. There are only better and worse strategies for each application.

Based on the architecture of neural networks, some important deep learning models are:

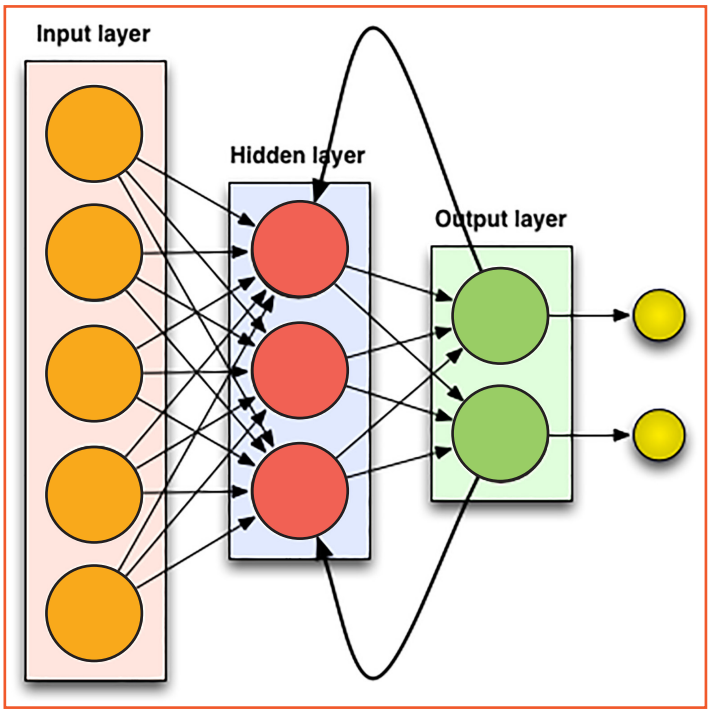
Multi-Layer perceptron



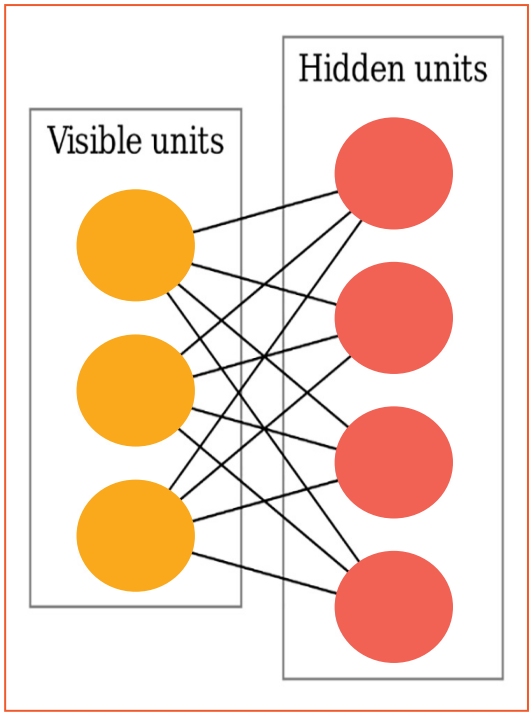
Convolution Neural Networks



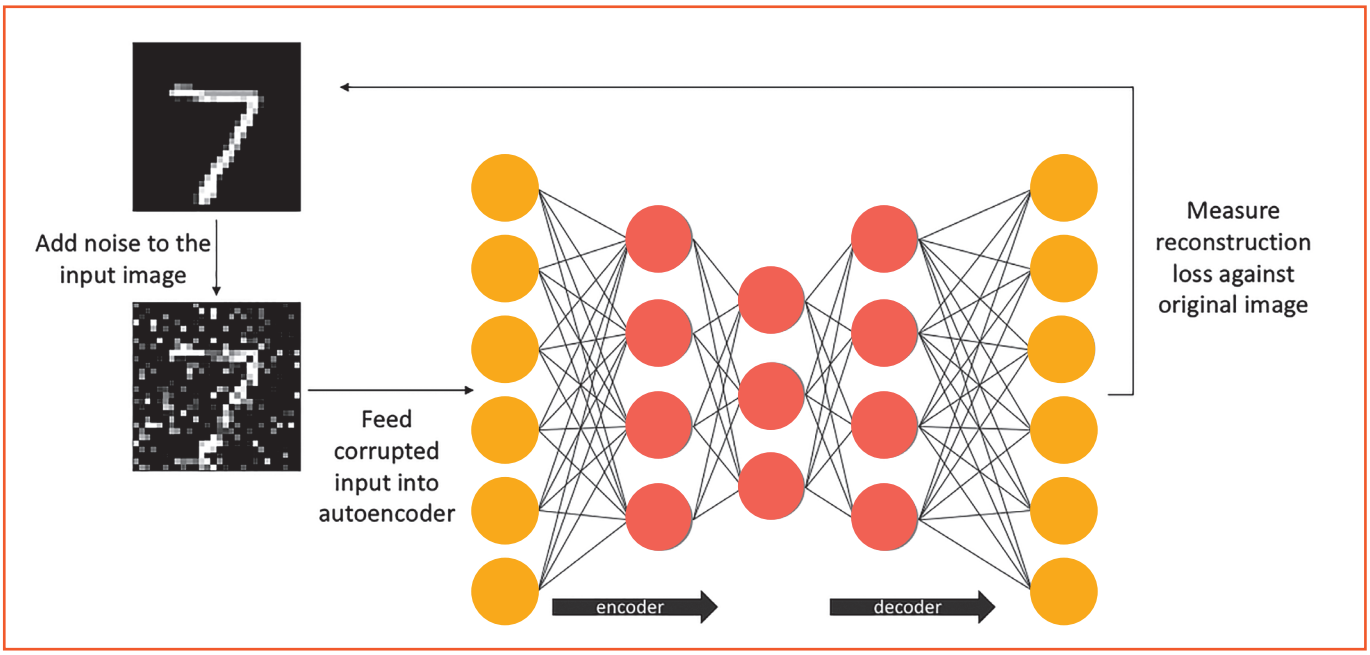
Recurrent Neural Networks



Boltzmann machine



Autoencoders etc.



EXPLORE

To learn more about deep learning and neural networks, you can go through the following links:

https://www.freecodecamp.org/news/deep-learning-neural-networks-explained-in-plain-english/

https://pathmind.com/wiki/neural-network

3.6 TRAINING

Training process entails 'feeding' the algorithm with training data. An algorithm will process data and output a model that can find a target value (attribute) in new data — an answer you want to get with predictive analysis. The purpose of model training is to develop a model.

Two model training styles are most common — supervised and unsupervised learning. The choice of each style depends on whether you want to forecast specific attributes or group data objects by similarities.

- Supervised learning - Supervised learning allows for processing data with target attributes or labeled data. These attributes are mapped in historical data before the training begins. With supervised learning, you can solve classification and regression problems.
- Unsupervised learning - During unsupervised learning, an algorithm analyzes data that is unlabeled. The goal is to find hidden interconnections, similarities or differences in the data. It is used for solving problems such as clustering, association rule learning, and dimensionality reduction.

The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The term ML model refers to the model artifact that is created by the training process. The training data must contain the correct answer, which is known as a target or target attribute. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict), and it outputs an ML model that captures these patterns.

You can use the ML model to get predictions on new data for which you do not know the target. To train an ML model, you need to specify the following:

- Input training data source
- Name of the data attribute that contains the target to be predicted
- Training parameters to control the learning algorithm

Training Parameters

Typically, machine learning algorithms accept parameters that can be used to control certain properties of the training process and the resulting ML model. In Amazon Machine Learning, these are called training parameters. You can set these parameters using the Amazon ML console, API, or Command-Line Interface (CLI). If you do not set any parameters, Amazon ML will use default values that are known to work well for a large range of machine learning tasks.

You can specify values for the following training parameters:

1. Maximum model size
2. Maximum number of passes over training data
3. Shuffle type
4. Regularization type
5. Regularization amount

EXPLORE

To learn more, you can go through the following link:

https://docs.aws.amazon.com/machine-learning/latest/dg/training-parameters.html

3.7 EVALUATING

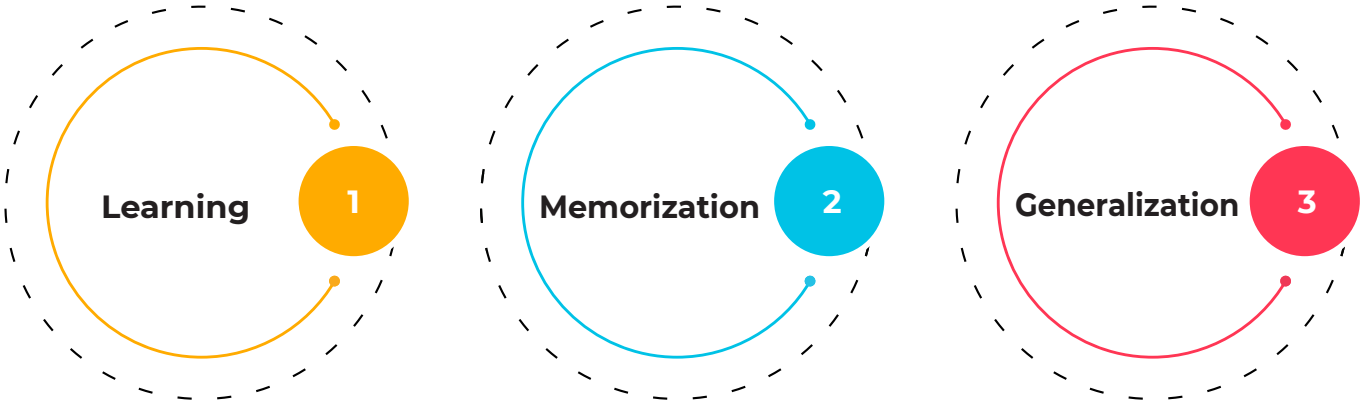
Once training is complete, it's time to see if the model is any good, using Evaluation. This is where that dataset that we set aside, for testing, comes into play. Evaluation allows us to test our model against data that has never been used for training. This metric allows us to see how the model might perform against data that it has not yet seen. This is meant to be representative of how the model might perform in the real world.

By using different metrics for performance evaluation, we should be in a position to improve the overall predictive power of our model before we roll it out for production on unseen data.

Without doing a proper evaluation of the ML model using different metrics, and depending only on accuracy, can lead to a problem when the respective model is deployed on unseen data and can result in poor predictions.

This happens because, in cases where the model does not learn but instead memorizes; hence, it is not able to generalize well on unseen data.

To get started, let's define these three important terms:

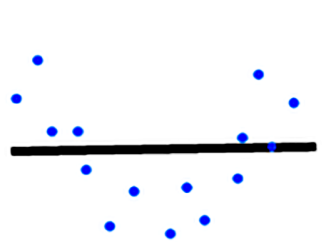


Learning: ML model learning is concerned with the accurate prediction of future data, not necessarily the accurate prediction of training/available data.

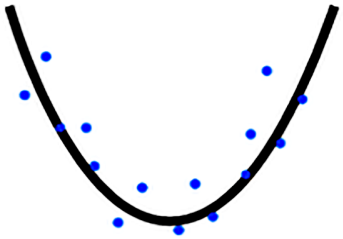
Memorization: ML Model performance on limited data; in other words, overfitting on the known training dataset.

Generalization: Can be defined as the capability of the ML model to apply learning to previously unseen data. Without generalization, there's no learning, just memorization. But note that generalization is also goal specific —for instance, a well-trained image recognition model on zoo animal images may not generalize well on images of cars and buildings.

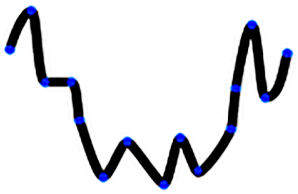
The images below depict how solely relying on model accuracy during training, leads to poor performance during validation.



Underfitting-perform badly on both training and validation data set



Generalized-perform good on both training and validation data set



Memorized (Overfitting)-perform good on training but fails on validation data set

Evaluation Metrics

Evaluation metrics are tied to machine learning tasks. There are different metrics for the tasks of classification, regression, ranking, clustering, topic modeling, etc.

Some key metrics are as follows:

Model Accuracy

Model accuracy in terms of classification models can be defined as the ratio of correctly classified samples to the total number of samples:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Or for binary classification models, the accuracy can be defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

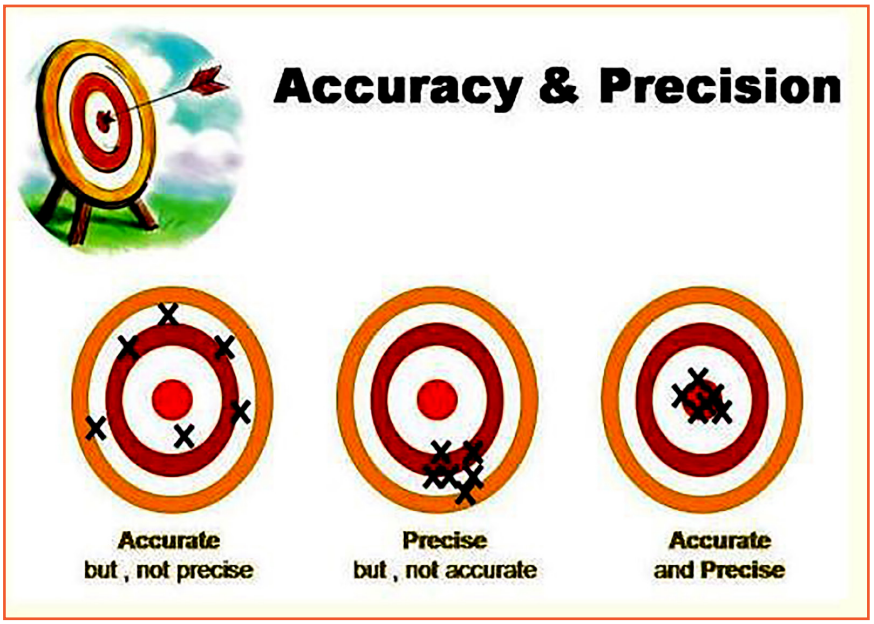
Confusion Matrix

- **True Positive (TP) —** A true positive is an outcome where the model correctly predicts the positive class.
- **True Negative (TN) —** A true negative is an outcome where the model correctly predicts the negative class.
- **False Positive (FP) —** A false positive is an outcome where the model incorrectly predicts the positive class.
- **False Negative (FN) —** A false negative is an outcome where the model incorrectly predicts the negative class.

Though highly-accurate models are what we aim to achieve, accuracy alone may not be sufficient to ensure the model's performance on unseen data.

Precision and Recall

In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class).



$$\text{Precision} = \frac{TP}{(TP+FP)}$$

High precision means that an algorithm returned substantially more relevant results than irrelevant ones.

In this context, recall is defined as the number of true positives divided by the total number of elements that belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

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

High recall means that an algorithm returned most of the relevant results

To fully evaluate the effectiveness of a model, it's necessary to examine both precision and recall. Unfortunately, precision and recall are often in conflict. That is, improving precision typically reduces recall and vice versa.

F1 Score

The F1 score is the harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

Why harmonic mean? Since the harmonic mean of a list of numbers skews strongly toward the least elements of the list, it tends (compared to the arithmetic mean) to mitigate the impact of large outliers and aggravate the impact of small ones:

$$F_1 = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

An F1 score punishes extreme values more. Ideally, an F1 Score could be a useful evaluation metric in the following classification scenarios:

When FP and FN are equally costly—meaning they miss on true positives or find false positives— both impact the model almost the same way, as in our cancer detection classification example

Adding more data doesn't effectively change the outcome effectively

TN is high (like with flood predictions, cancer predictions, etc.)

ROC Curve

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

The ROC curve is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The true positive rate is also known as sensitivity, recall, or probability of detection in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm.

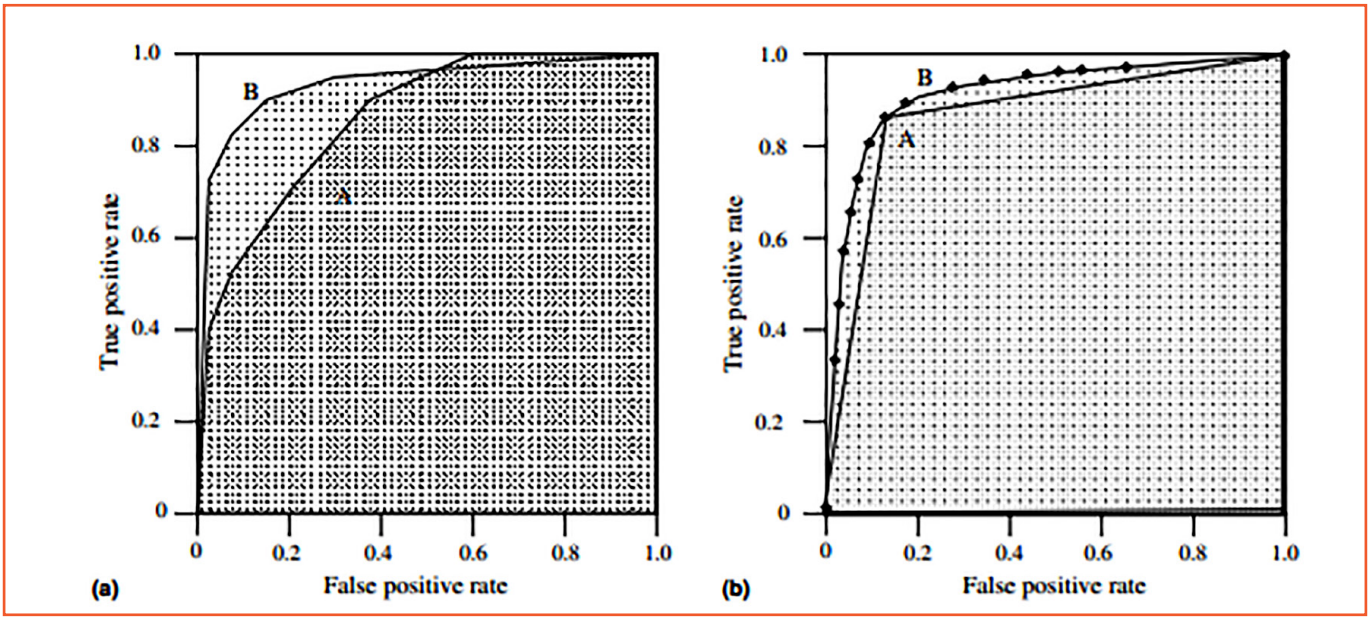
$$\text{True Positive Rate (TPR)} = \frac{TP}{(TP+FN)}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{(FP+TN)}$$

AUC — Area under the ROC Curve

A ROC curve is a two-dimensional depiction of classifier performance. To compare classifiers, we may want to reduce ROC performance to a single scalar value representing expected performance. A common method is to calculate the area under the ROC curve, abbreviated AUC.

Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1.0. However, because random guessing produces the diagonal line between (0, 0) and (1, 1), which has an area of 0.5, no realistic classifier should have an AUC less than 0.5



4a shows the areas under two ROC curves, A and B. Classifier B has a greater area and therefore better average performance. Fig. 4b shows the area under the curve of a binary classifier A and a scoring classifier B.

It is possible for a high-AUC classifier to perform worse in a specific region of ROC space than a low-AUC classifier. Fig. 4a shows an example of this: classifier B is generally better than A except at FP rate > 0.6, where A has a slight advantage. But in practice, AUC performs very well and is often used when a general measure of prediction is desired.

AUC is desirable for the following two reasons:

- AUC is scale-invariant. It measures how well predictions are ranked, rather than their absolute values.
- AUC is a classification-threshold-invariant. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

Multi-Class ROC

With more than two classes, the classifications problem become much more complicated if the entire space is to be managed. With n classes, the confusion matrix becomes an n · n matrix containing the n correct classifications (the major diagonal entries) and n*n — n possible errors (the off-diagonal entries). Instead of managing trade-offs between TP and FP, we have n benefits and n*n — n errors. With only three classes, the surface becomes a 3*3- 3 = 6-dimensional polytope.

One method for handling n classes is to produce n different ROC graphs, one for each class.

Specifically, if C is the set of all classes, ROC graph i plots the classification performance using class ci as the positive class and all other classes as the negative class:

Pi = Ci

Ni = Union(Cj) for j≠i

Multi-Class AUC: Similarly, AUC can be calculated for multi-class problems by generating each class reference ROC curve, in turn, measuring the area under the curve and then summing the AUCs weighted by the reference class's prevalence in the data.

EXPLORE

To learn more go through the following link:

https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/evaluate-model#bkmk_classification

3.8 PARAMETER TUNING

Once you've done evaluation, you may want to further improve your training in some way. You can do this by tuning your parameters. There may have been a few parameters that were implicitly assumed during training. Now is a good time to go back and test those assumptions and try other values.

One example is how many times we run through the training dataset during training. What we mean by that is we can 'show' the model our full dataset multiple times, rather than just once. This can sometimes lead to higher accuracies.

Another parameter is **'learning rate'**. This defines how far we shift the line during each step, based on the information from the previous training step. These values all play a role in how accurate our model can become and how long the training takes.

For more complex models, initial conditions can play a significant role in determining the outcome of training. Differences can be seen depending on whether a model starts off training with values initialized to zeroes versus some distribution of values, which leads to the question of which distribution to use.

As you can see there are many considerations at this phase of training, and it's important that you define what makes a model "good enough", otherwise you might find yourself tweaking parameters for a very long time.

These parameters are typically referred to as **'hyperparameters'**. The adjustment, or tuning, of these hyperparameters, remains a bit of an art and is more of an experimental process that heavily depends on the specifics of your dataset, model, and training process.

Ensemble Methods

Data scientists mostly create and train one or several dozen models to be able to choose the optimal model among well-performing ones. Models usually show different levels of accuracy as they make different errors on new data points. There are ways to improve analytic results. Model ensemble techniques allow for achieving a more precise forecast by using multiple top-performing models and combining their results. The accuracy is usually calculated with mean and median outputs of all models in the ensemble. Mean is a total of votes divided by their number. Median represents a middle score for votes rearranged in order of size.

The common ensemble methods are stacking, bagging, and boosting.

- **Stacking -** Also known as Stacked Generalization, this approach suggests developing a meta-model or higher-level learner by combining multiple base models. Stacking is usually used to combine models of different types, unlike bagging and boosting. The goal of this technique is to reduce generalization error.

- **Bagging (bootstrap aggregating) -** This is a sequential model ensembling method. First, a training dataset is split into subsets. Then models are trained on each of these subsets. After this, predictions are combined using mean or majority voting. Bagging helps reduce variance error and avoid model overfitting.
- **Boosting -** According to this technique, the work is divided into two steps. A data scientist first uses subsets of an original dataset to develop several averagely performing models and then combines them to increase their performance using majority vote. Each model is trained on a subset received from the performance of the previous model and concentrates on misclassified records.

Once you're happy with your training and hyperparameters, guided by the evaluation step, it's time to finally use your model to do something useful!

3.9 PREDICTION

Machine learning is using data to answer questions. So Prediction, or inference, is the step where we get to answer some questions. This is the point of all this work, where the value of machine learning is realized.

However, to get the results, the model has to be deployed. To do so, you will have to translate the final model from high-level programming languages (i.e. Python and R) into low-level languages such as C/C++ and Java. The distinction between two types of languages lies in the level of their abstraction in reference to hardware. A model that's written in low-level or a computer's native language, therefore, better integrates with the production environment.

After translating a model into an appropriate language, you can measure its performance with testing. Testing can show how many customers engaged with a model used for a personalized recommendation, for example, correlates with a business goal.

Machine learning as a service (MaaS) is an automated or semi-automated cloud platform with tools for data preprocessing, model training, testing, and deployment, as well as forecasting. The top three MLaaS are Google Cloud AI, Amazon Machine Learning, and Azure Machine Learning by Microsoft. ML services differ in a number of provided ML-related tasks, which, in turn, depends on these services' automation level.

Deployment on MLaaS platforms is automated. For example, the results of predictions can be bridged with internal or other cloud corporate infrastructures through APIs.

The predictions from the model can be real-time or in set intervals.

Batch prediction

This deployment option is appropriate when you don't need your predictions continuously. When you choose this type of deployment, you get one prediction for a group of observations. A model is trained on a static dataset and outputs a prediction. You can deploy a model on your server, on a cloud server if you need more computing power or use MLaaS for it. Deployment is not necessary if a single forecast is needed or you need to make sporadic forecasts. For example, you can solve a classification problem to find out if a certain group of customers accept your offer or not.

Web service

Such machine learning workflow allows for getting forecasts almost in real-time. A model, however, processes one record from a dataset at a time and makes predictions on it. It's possible to deploy a model using MLaaS platforms, in-house, or cloud servers.

Real-time prediction (real-time streaming or hot path analytics)

This type of deployment speaks for itself. With real-time streaming analytics, you can instantly analyze live streaming data and quickly react to events that take place at any moment. Real-time prediction allows for processing of sensor or market data, data from IoT or mobile devices, as well as from mobile or desktop applications and websites. As this deployment method requires processing large streams of input data, it would be reasonable to use Apache Spark or rely on MlaaS platforms. Apache Spark is an open-source cluster-computing framework. A cluster is a set of computers combined into a system through software and networking. Due to a cluster's high performance, it can be used for big data processing, quick writing of applications in Java, Scala, or Python.

Web service and real-time prediction differ in the amount of data for analysis a system receives at a time.

3.10 MACHINE LEARNING AS A SERVICE

Machine learning as a service (MLaaS) covers the various cloud-based platforms that provide most of the machine learning services such as data pre-processing, model training, and model evaluation, with further prediction. Prediction results can be bridged with your internal IT infrastructure through REST APIs.

The four leading cloud MLaaS services that allow for fast model training and deployment are:

These platforms are most useful for Machine Learning enthusiasts. We shall be working with some of these in the activities and projects in subsequent units.

EXPLORE

Hence explore the following websites to familiarize yourself with the services they provide.

https://aws.amazon.com/machine-learning/

https://cloud.google.com/products/ai/

https://azure.microsoft.com/en-us/services/machine-learning/

https://www.ibm.com/watson

SUMMARY

Machine Learning is the technique where algorithm or deep learning is used to build AI rather than the use of programming. It uses the method of 'training' an algorithm so that machines can learn.

The machine learning process involves the following steps:

- Gathering data
- Preparing that data
- Choosing a model
- Training
- Evaluation
- Hyperparameter tuning
- Prediction.

The data gathering process involves:

1. Determining the Information to be collected
2. Setting up a timeframe for data collection
3. Identifying data sources
4. Determining the data collection method
5. Collecting the data
6. Verifying the data

Data Collection involves Data discovery, Data augmentation, Data generation and Data Labeling

Big Data is the name given to massive collections of data generated by humans. This availability of massive datasets is one of the key reasons that has had machine learning possible. More data means more opportunity for a computer algorithm to find associations and hence, higher accuracy of predictions.

There are tools that enable users to source large volumes of raw data quickly such as, Data Scraping Tools, Synthetic Data Generators, Data Augmentation Tools, Data Labelling Tools and Open-Source Datasets.

Once the data has been collected, it is in a raw form. That is why a large amount of time is spent on data cleaning which involves:

- Identifying relevant data and removing irrelevant data
- Fix Irregular cardinality and structural errors
- Outliers
- Missing data treatment
- Data Transformation

Post the data pre-processing, comes the final data preparation phase of Data splitting.

The next step is selecting an appropriate model. A machine learning model is a set of rules that recognizes certain types of patterns. Based on the type of tasks, we can classify machine learning models in the following types:

- Classification Models
- Regression Models
- Clustering
- Dimensionality Reduction
- Deep Learning etc.

Artificial Neural Networks (ANN) are inspired by the central nervous system that we possess and are applied to several problem domains:

- To classify data by recognizing patterns.
- To detect anomalies or novelties, when test data does not match the usual patterns.
- To process signals, for example, by filtering, separating or compressing.
- To find a function that closely matches the training data (approximate a target function) so that it can predict almost correctly.

Effectively, Neural networks can identify faces, recognize speech, read handwriting, translate texts, play games (typically board games or card games), control autonomous vehicles and robots and a few other things.

Typically, ANN's have a layered structure. The input layer picks up the input signals. Then the input layer passes them on to the next layer, the so-called 'hidden' layer. There may be more than one hidden layer in a neural network. Last comes the output layer that delivers the result.

Perceptrons can be viewed as building blocks in a single layer in a neural network, made up of four different parts, Input Values or One Input Layer; Weights and Bias; Net sum and Activation function.

A step activation function is a simple binary function that has only two possible results. Some other activation functions are the logistic or sigmoid function, hyperbolic tangent and Rectified Linear Unit (ReLU).

Deep learning is a subset of AI and machine learning that consist of multiple layers of ML algorithms. They are often referred to as "deep neural networks" (DNN). Input is passed through the layers, with each adding qualifiers or tags.

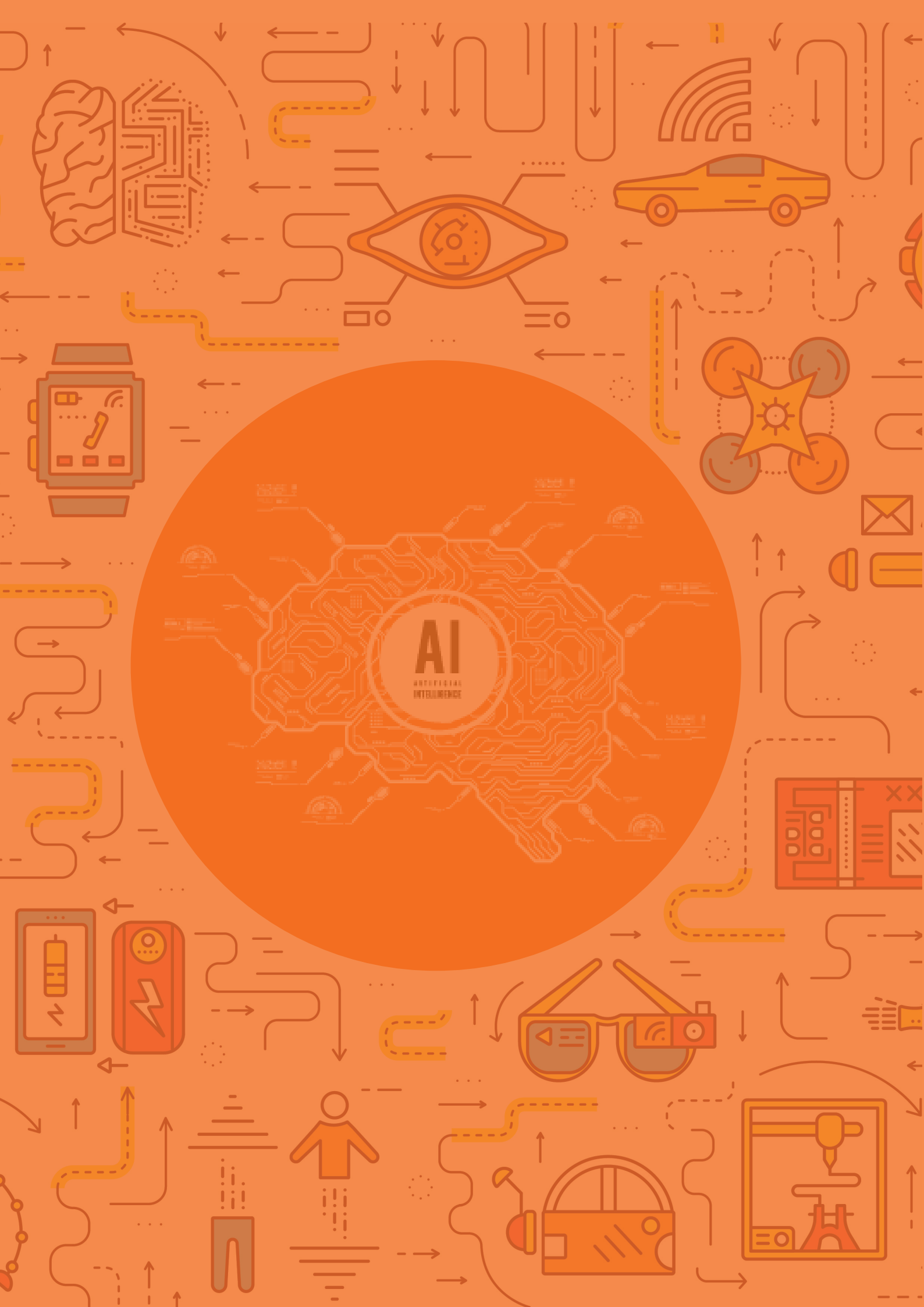
Training process entails “feeding” the algorithm with training data. An algorithm will process data and output a model that can find a target value (attribute) in new data — an answer you want to get with predictive analysis. The purpose of model training is to develop a model. Two model training styles are most common — supervised and unsupervised learning.

Evaluation allows us to test our model against data that has never been used for training. This metric allows us to see how the model might perform against data that it has not yet seen. This is meant to be representative of how the model might perform in the real world. Some key metrics are Model Accuracy, Precision and Recall, F1 Score and ROC Curve

Once you’ve done evaluation, you may want to further improve your training in some way. You can do this by tuning your parameters.

Machine learning is using data to answer questions. So Prediction, or inference, is the step where we get to answer some questions. This is the point of all this work, where the value of machine learning is realized.

Machine learning as a service (MaaS) is an automated or semi-automated cloud platform with tools for data preprocessing, model training, testing, and deployment, as well as forecasting.


AI
ARTIFICIAL
INTELLIGENCE

UNIT 4

NATURAL LANGUAGE PROCESSING (NLP)

Key Learning Outcomes

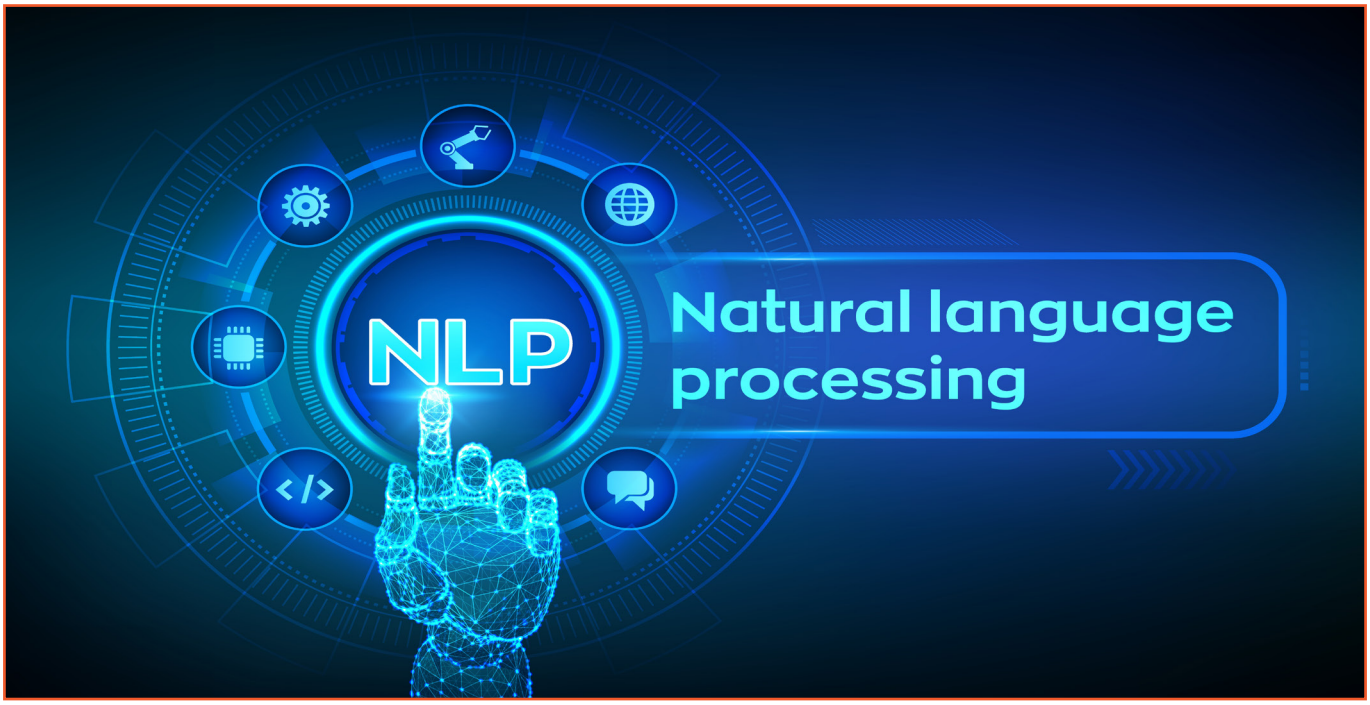
At the end of this module, you will be able to:

a) Explain what is NLP and how it works
b) Differentiate between NLP and NLU
c) List common applications of NLP

Topic Covered

Introduction to Natural Language Processing (NLP) | Natural Language Understanding I Application of NLP | Use Cases | How Does NLP work | Elements of NLP

4.1 INTRODUCTION TO NATURAL LANGUAGE PROCESSING (NLP)



Natural Language Processing (NLP) is the part of machine learning that helps computers interpret and process natural human language.

Hey Alexa, set the alarm for 5 AM every day

Done — your alarm is set for 5 AM every day

Alexa, can you play the Earth Song for me?

Sorry, I cannot find the Bird Song, would like me to play a song from your playlist?

Ever wondered how devices like Siri, Google Assistant and Alexa understand and interpret your voice? Have you also been irritated when they have not been able to pick up specific terms or confuse them with others?

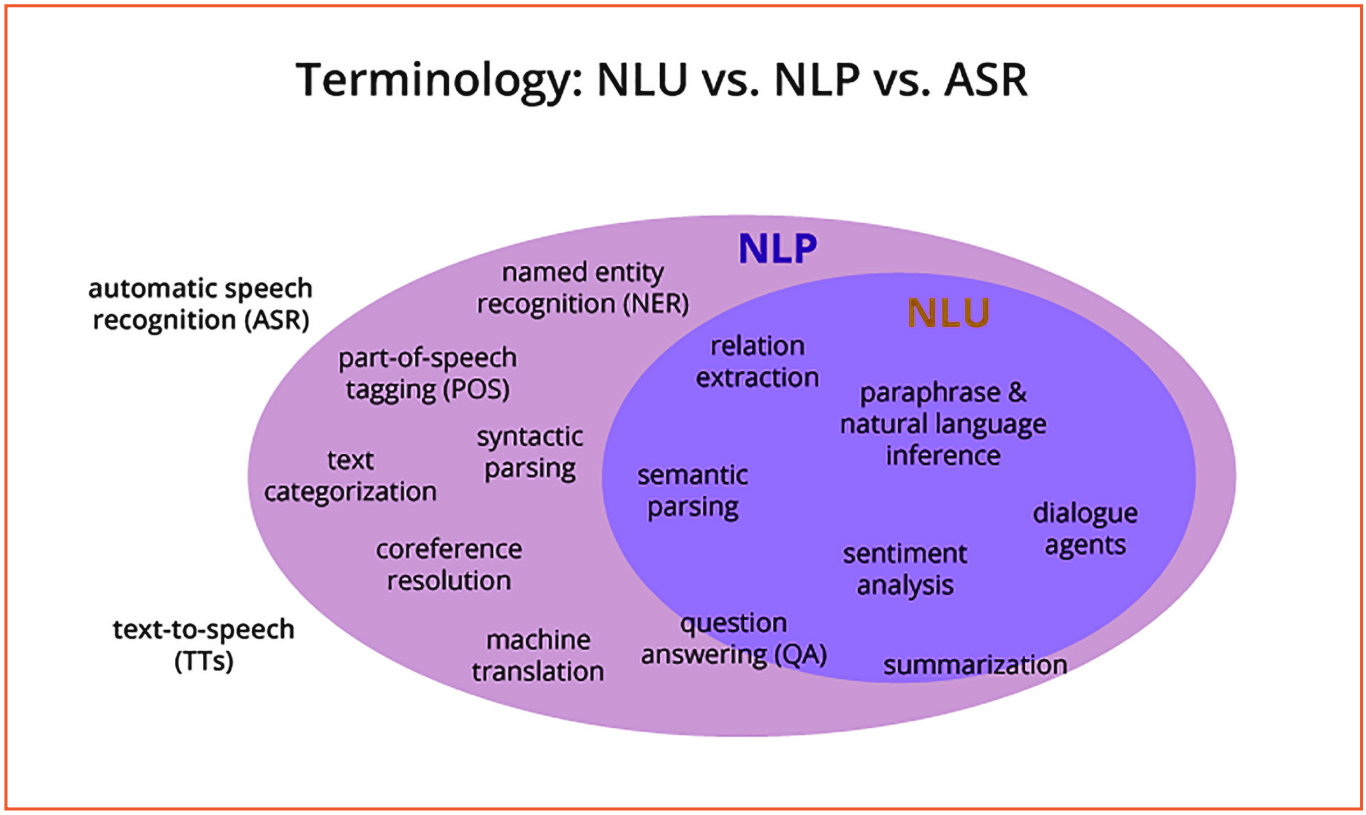
Well, this is the result of Natural Language Processing (NLP), the potential and challenges of it.

NLP is a subfield of artificial intelligence. NLP uses AI and ML tools, techniques, algorithms (computer science) and linguistics to help computers understand unstructured 'human language' data and derive meaning from it. NLP bridges the gap between the human language and the command line interface of a computer.

Within the domain of computer-aided processing of natural languages, there are concepts of natural language processing and natural language understanding, which are sometimes used interchangeably. But are they the same? Also, what is Automatic Speech Recognition?

Let us understand these.

Both NLP and NLU relate natural language to artificial intelligence. Both aim to derive meaning from unstructured data, like language, as opposed to structured data like statistics, actions, etc. However, there are differences between NLU and NLP when it comes to data mining techniques.



NLP involves computer science, artificial intelligence, linguistics and data mining. The objective of NLP is to help machines read, decipher, understand, and make sense of the human languages. It aims at substituting humans for doing various tasks. Common real-world examples of such tasks include online chatbots, text summarizers, auto-generated keyword tabs, and tools that analyze text to identify the sentiment of that given text.

4.2 NATURAL LANGUAGE UNDERSTANDING

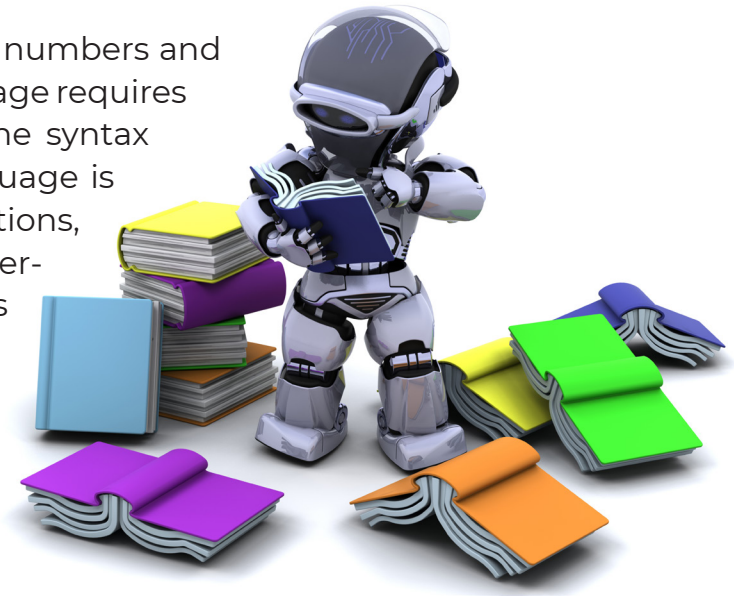
Considered a subfield of NLP, natural language understanding is narrower in purpose, that focuses mainly on machine reading comprehension, i.e. getting the computer to comprehend what a text means. Without understanding the meaning of the text, there is little expectation to respond intelligently. Similar to NLP, NLU uses algorithms to reduce human speech into a structured form. Then these AI algorithms detect such things as sentiment, intent, timing, locations, etc.

Natural language understanding can be applied in several tasks such as gathering news, classifying text, archiving pieces of text and analyzing content. Real-world examples of NLU range from simpler to more complex applications. A simpler task could be issuing short commands based on comprehending text to a limited degree such as rerouting an e-mail to the right person based on basic syntax and a reasonably sized lexicon. More complex applications may include, fully comprehending news articles or shades of meaning within poetry or novels.

Automatic speech recognition is used to convert spoken words into computer text. It is also used for biometric authentication (authenticating users via their voice) and following human instructions to perform an action. Typically, automatic speech recognition requires preconfigured or saved voices of the primary user(s). The human needs to train the automatic speech recognition system by storing speech patterns and their vocabulary into the system.

Nowadays, big data can sort, organize, and provide insights into data unthinkable to humans in the past. Yet, in the pursuit of artificial intelligence, we realise that it takes a lot more for computers to make sense of the data of language.

Machines are used to finding patterns in numbers and statistics, but the understanding of language requires a lot more, including comprehending the syntax of a language, differences in how a language is spoken or written, the context and definitions, changing language patterns and ever-evolving new definitions, the subtleties and nuances that result in sarcasm,



which aren't inherently readable from the text or understanding the real purpose or goal of a body of content.

This is where NLP and NLU are required and these applications can be quite complex. With so many different languages in the world, communication is a complex phenomenon varying between cultures and geographies. Add to it the diversity in the languages, such as writing styles, syntax, and grammar rules, varying accents and dialects; all make a significant difference in language comprehension.

4.3 APPLICATION OF NLP

Thus, NLP is still an ever-evolving and developing field that still involves a vast amount of research, experimentation and innovation to cater to the various needs it serves.

As Language understanding is complex, it is differentiated into three linguistic levels:

Syntax – understanding the rules and arrangement of words and phrases followed to create well-formed and meaningful sentences in a language.

Semantics – understanding the meaning of words, phrases, sentences and other text.

Pragmatics – understanding the purpose or what the text is trying to achieve.

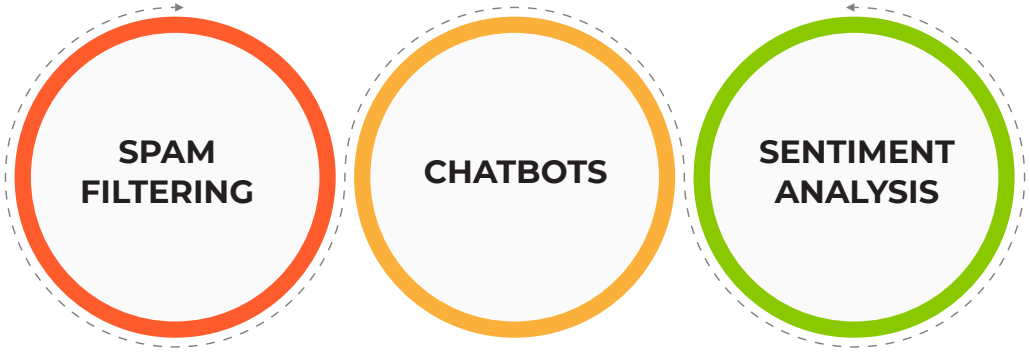
There isn't a single way to ensure language acquisition, even for humans, a new language is hard to learn. What makes it more complicated is that languages are always being modified, shifting, adding and subtracting from an ever-widening lexicon. It also incorporates new forms of usage such as e-mails, texts, social media impacting language.

NLP refers to a wide range of applications, including speech recognition, natural language recognition and natural language generation.

Tasks that NLP is used for include text summarization, sentiment analysis, topic extraction, parts-of-speech tagging, stemming, text mining, machine translation, named entity recognition, relationship extraction, and automated question answering, language modeling and almost all other language-related tasks.

4.3.1 USE CASES

Given below are some real-world applications of NLP that you may be familiar with.



Spam Filtering

SPAM filtering uses NLP extensively. A spam filter identifies unwanted e-mails and sends them to a separate folder instead of the inbox.

Simple spam filtering can be achieved using classification models in machine learning. However, NLP provides better capabilities to distinguish between useful e-mails and real spam. NLP techniques such as n-gram modelling are applied to e-mails to classify them as spam or ham (not spam) with higher accuracy than traditional classification models.

Chatbots

A chatbot with artificial intelligence (AI) is a conversational piece of software used to answer questions without the need of a human operator. It can simulate a conversation with a user in natural language through voice and text. It is usually customized for use on mobile devices, web browsers, and chat platforms such as Facebook Messenger, or Slack.

A chatbot is an advanced application of natural language processing, that takes us closer to communicating with computers in a way like human-to-human conversations. Without NLP, chatbots will not deliver any value to its users. The NLP models allow the chatbot to understand your query or comment and reply to you. Businesses these days are increasingly using chatbots across all domains, especially customer service, sales and marketing, etc. It is being used to respond to customer queries, analyse feedback and provide solutions to the customer through product or service recommendations.

Chatbots help businesses in saving time and money. Since most questions asked by customers are frequently repeated, they can be handled by chatbots. This helps customer service agents prioritize relevant customer queries, thereby ensuring overall customer satisfaction.

Chatbots use a combination of Natural Language Processing, Natural Language Understanding, and Natural Language Generation to achieve the desired Conversational User Interface.

Sentiment Analysis

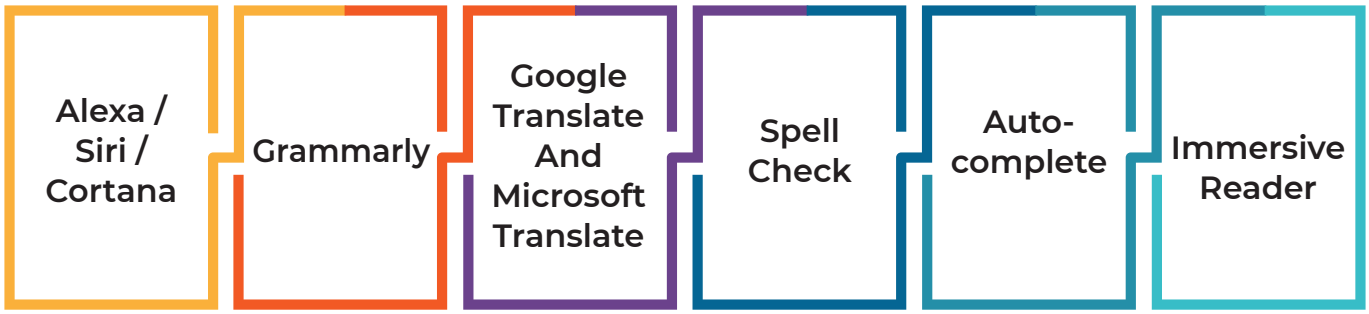
Sentiment analysis is used when we have to identify the sentiment of a text to interpret and classify emotions to get feedback or understand intent. We may use this to gauge how customers feel when they use a service or to classify comments made on a site.

With many people expressing their feedback through large scale surveys or social media platforms, organizations rely on software to process large volumes of such information to save time by automating responses or collecting and analyzing data. NLP is used to understand human emotions expressed via text to understand the customer's feedback. These analyses may also be used to adapt products and services to meet customer expectations.

Sentiment Analysis may identify opinions expressed as positive, neutral, or negative. One version of this is Real-Time Sentiment analysis. This helps organisations identify critical issues in real-time. For example, a constant analysis of social media comments can help identify the break-out of a crisis or a scandal that may affect an organization due to growing protests. Organizations may use real-time sentiment analysis models to quickly identify these issues, even before they happen and address it before it becomes wide-spread and unmanageable.

4.3.2 SOME PRODUCTS THAT USE NLP

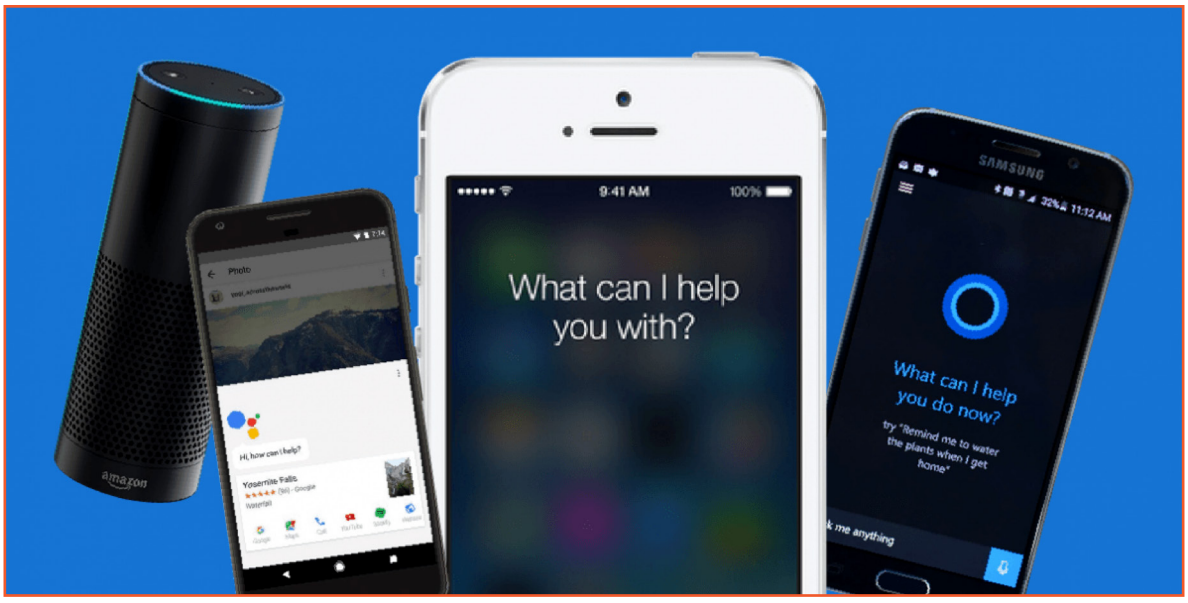
Now let us look at some products that use NLP. There are numerous products that we use daily without realizing the fact that NLP powers them; here is a short list.



Alexa / Siri / Cortana

Several technology companies have produced virtual assistant to provide an exciting and engaging service experience to its users. Alexa, Cortana, Siri, and Google Assistant are examples of Amazon, Microsoft, Apple and Google, respectively and now widely used across the globe. Many other products are now linking to these offering voice control options, for example, smartphone, refrigerators, smart TVs and air-conditioners, etc. One can even program security systems, home power systems, to these devices.

NLP is a core enabler of these virtual assistants. NLP operations help to detect, interpret, and understand the text or voice commands to perform the requested function. All these assistants are continually evolving through AI and machine learning to expand the accuracy of their responses.

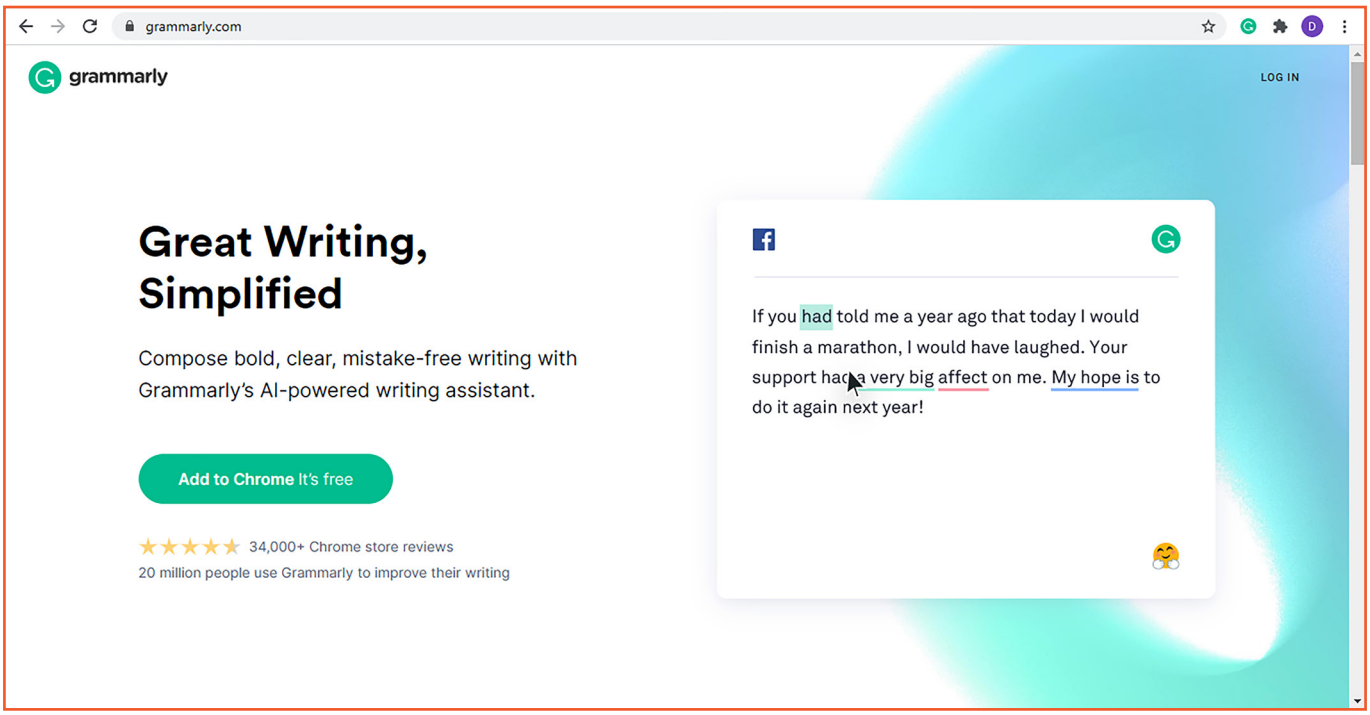


Grammarly

Grammarly is one of the popular writing/editing tools used around the world for all types of writing. It provides options for different writing styles, audiences and other capabilities to help edit your text. One can use auto-detect capabilities for spellings and grammar, passive voice writings, styles of English, tone of the document, etc. to get suggestions for changes according to the selected writing goals.

AI, ML and NLP are the primary enablers of Grammarly's capabilities. Grammarly's AI system relies on a wide range of NLP algorithms that address each of the options that one can see in its offering. NLP allows Grammarly to perform various tasks, including writing improvement, sentiment analysis, readability scoring, suggestions to use alternate words, phrases, and sentence structure, etc.

The machine learning system of Grammarly is continuously updated even as it is used by the many users across the world and it will only get better with time as it has more data to work with.

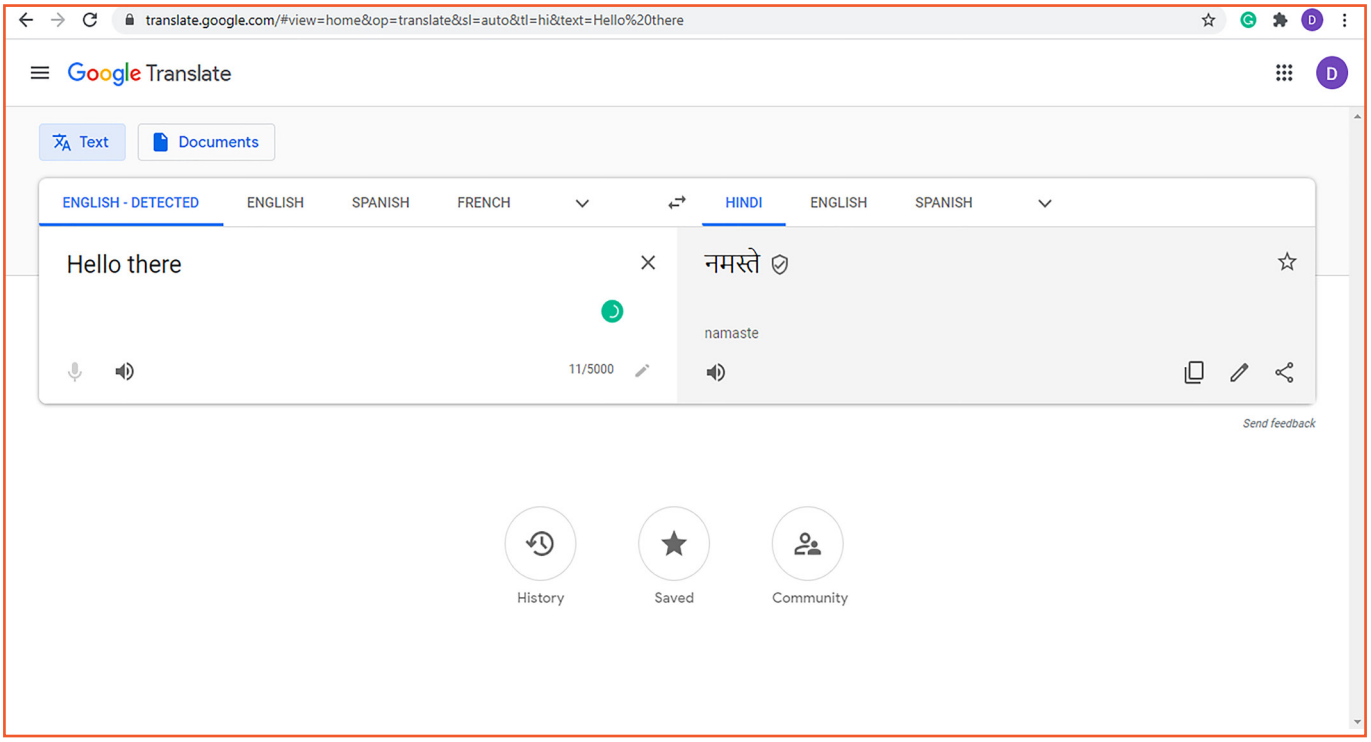


Google Translate And Microsoft Translator

Language translation is a core application of Natural Language Processing. Google translate is one such application that uses NLP. It provides capabilities required to translate documents and audio from one language to the other across many different languages.

Google Translate or Microsoft Translate is an incredibly efficient system that makes use of AI, machine learning, and NLP to translate text from one language to another. Google Translate supports over 100 languages.

As discussed earlier, Language translation can be very complicated because the software, given the different rules (Syntax) followed by different languages, must still understand words, phrases, and sentence structures accurately for correct translation. Google Translate too, is continuously updated to improve the quality and accuracy of the language-translation, using an ever-increasing amount of data and iterations to improve.



Spell Check

Spell Check is a much-used tool that corrects spelling or even some grammar errors. The spell check feature is usually combined with an auto-suggestion feature used in smartphones that can help save time and enables people to overcome language and accurate typing limitations. It helps to eradicate the spelling errors from your messages and improves communication. This feature uses NLP to identify errors and auto-suggest the right word or spelling you are looking for, by analysing the text and predicting the words/spellings you are looking for and is consistent with the writing.

Autocomplete

Autocomplete is another extended application of NLP that is used along with spell check often. It is b used by almost every web and mobile application, including emails and search engines like Internet Explorer and Google.

Autocomplete using previous searches performed by other customers, helps us to find what we are looking for quickly. Thanks to the advancements in NLP word combinations and text analysis, we can experience the benefits of these.

Immersive Reader

This is an application of NLP that is used to capture information from documents, whiteboards, business cards, receipts, menus, signs, handwritten memos, or anything else containing text and have it read by the software. This is used for teaching students how to read, for helping visually challenged individuals and anyone else who wants their documents auto-read.

4.3.3 OTHER APPLICATIONS OF NLP

Let us now see how NLP may be used in a few other applications:

Marketing

Natural language processing is being used in marketing efforts to boost efficiency. From using chatbots to smart AI copywriters generating slogans, NLP models are extending marketing capabilities and creating new forms of marketing as well.

Banking

NLP is helping AI software process documentation in Banking institutions faster. It is also helping to mitigate risks, automate business processes, or check the quality of customer services. For example, NLP enabled AI is being used by banks to extract key data and clauses to help loan officers review commercial loan agreements.

Fake News Detection

Fake news is inundating social media these days. It is creating a lot of confusion and complications and becoming a huge problem. NLP systems are frequently used to detect fake information and provide statistics on their exposure. NLP models are being used by detecting the topic and the facts of a news article, then comparing the topic of the given article with other trustworthy articles. The model accords weights based on the trustworthiness, then sets a threshold. If the weight is greater than the threshold, the news will be labeled as real news, if not, then it will be labeled as fake news.

Healthcare

NLP models are currently helping medical workers to process patient data more efficiently. It is also helping to improve the quality of medical care and identify patients who need special care. NLP is being used to map data elements present in unstructured text to structured fields in an electronic health record (EHR) to improve clinical data integrity. Health workers are also using speech recognition to dictate clinical notes or other information that is then turned into text.

4.4 FUTURE OF NLP

The ultimate target of NLP AI is to develop a chatbot that would be able to carry out meaningful, unsupervised interactions with humans and pass the Turing test. We are likely to see more investments from businesses and academia in NLP and NLU for the many products and services it enables.

There is much ground to be covered and people from different fields such as language, linguistics, cognitive science, data science, computer science are combining efforts and collaborating to produce successful NLP AI. We hope you will keep track of the development of this exciting front too. Now let's get into some activity to produce a chatbot.

4.4.1 How Does NLP Work

NLP derives meaning from human languages by analysis of the text using semantics and syntax.

Syntactic analysis is used to understand how grammatical rules apply within a language and how these can be used to derive meaning from a group of words. To do this, several techniques are used; these include:

1. **Lemmatization:** reducing the inflected forms of a word into a single form for easy analysis
2. **Stemming:** cutting the inflected words to their root form
3. **Morphological segmentation:** dividing words into morphemes
4. **Word segmentation:** dividing a continuous text into distinct units

5 **Parsing:** grammatical analysis of a sentence

6 **Part-of-speech tagging:** identifying the part of speech for every word

7 **Sentence breaking:** placing sentence boundaries on a continuous text

Humans have hundreds of languages like English, Spanish, Hindi, Chinese, or French. Computers, on the other hand, have only one native language, which is called the machine language. All the processes in your computers and smart devices communicate via millions of zeros and ones to perform a specific function. The machine code is incomprehensible to humans, which makes NLP a critical part of human-computer interactions.

4.5 ELEMENTS IN NATURAL LANGUAGE PROCESSING

Let us briefly explore the process and main elements of natural language processing:

If we run through the NLP basics, there are seven essential steps you need to undertake to help your computer understand natural language:

Sentence Segmentation

The first element involves splitting sentences into separate objects. In this stage, a smart AI algorithm evaluates the data sets and defines punctuation marks like periods (full-stops) to identify where a sentence finishes. This stage is vital the NLP model here has to meaningfully segments the sentence and then derive the meaning by analysis of the whole paragraph.

Under ideal circumstances, this should be easy, but it is not. It is made complicated as the real-world data comes without accurately used punctuation marks or has other errors in the text. Therefore, data scientists need to apply sophisticated algorithms and techniques to identify the correct segments.

Word Tokenization

After sentence segmentation, words are separated to determine various parts of speech. While it is easy to do in some languages like English by identifying spaces between the words or tokens, in other languages it may be complicated as words may be strung together, etc. Punctuation marks in themselves are also considered separate tokens as they carry particular meaning that can change the meaning of the overall text.

To determine its part of speech, AI algorithms analyze each word and based on selected criteria categorize these into adverbs, adjectives, nouns, verbs, etc. The machine uses this to determine the role of each token in the sentence or text.

A pre-trained parts-of-speech classification model enables this function. Millions of texts in the given language that are previously tagged and marked are used to train this model. Algorithms use this essential data in large data sets to analyze and develop statistical models that determine which part of speech do the words belongs to.

Text Lemmatization

Lemma refers to the common part or the basic form of related words used in different grammatical forms. For example, the words ‘greater’ and ‘greatly’ will have ‘great’ as the lemma. Many texts or sentences may contain the root word as well as its different grammatical forms. NLP figures out the lemma from each token (word) and uses it to identify the words, categorise them and find their meaning.

Stop Words

Stop words are words that do not add to the meaning of the text. These are filler words, or casually inserted words that add noise to the text and only complicate the text. NLP includes processes to identify and skip these when trying to understanding the text.

Dependency Parsing in NLP

Dependency parsing is discovering the relations between all the words in a text. A 'Parse Tree' is created by AI algorithms that identifies the root word in the sentence and bridge the gap between other tokens. A parent word for each token may also be defined to gain more insight and understand the core concept.

Named Entity Recognition (NER)

Named Entity Recognition (NER) is the stage where data scientists start extracting ideas from the text relating real-life objects to tokens. In the NER process, an algorithm takes a string of text (sentence or paragraph) as input and identifies relevant nouns (people, places, and organizations) that are mentioned in that string.

Through this information, the NLP model extracts extra meaning from the text that is used to conduct a thorough analysis of the text to arrive at its meaning.

Coreference Resolution

Coreference resolution refers to the task of finding all expressions in the text that refer to the same entity. Every language has many entities, such as pronouns and other parts of speech, that may relate to another word in the sentence and expand its meaning. Coreference resolution is used to cluster all mentions in the text, which refer to a real-life concept or entity.

For example, He said, "Mary and I are going to see my other apartment so that she can choose, which living quarters may be more suited to her needs".

Here Mary, Her and She, refer to one cluster and I and My belong to the second cluster.

NLP Algorithms using coreferences commonly look for the nearest preceding mention that is compatible with the referring expression. Neural networks can also be trained to account for word embeddings and distance between mentions as features, to identify these clusters to find the meaning of the whole text.

EXPLORE

Use this link to explore this dimension of NLP in a demo.

https://huggingface.co/coref/

ACTIVITY

Go through the following links and try replicating the instructions provided using Scratch.

EXPLORE

1. Text to speech using Scratch

https://www.youtube.com/watch?v=KSr5bLXaoig

EXPLORE

2. Chatbot in Scratch

https://www.youtube.com/watch?time_continue=53&v=gZAdYyjxRcg&feature=emb_logo

SUMMARY

Natural Language Processing (NLP) is the part of machine learning that helps computers interpret and process natural human language. It uses AI and ML tools, techniques, algorithms (computer science) and linguistics to help computers understand unstructured 'human language' data and derive meaning from it.

Natural Language Understanding (NLU) is considered a subfield of NLP, and focuses mainly on machine reading comprehension, i.e. getting the computer to comprehend what a text means. Natural language understanding can be applied in several tasks such as gathering news, classifying text, archiving pieces of text and analyzing content.

As Language understanding is complex, it is differentiated into three linguistic levels:

- Syntax – understanding the rules and arrangement of words and phrases followed to create well-formed and meaningful sentences in a language.
- Semantics – understanding the meaning of words, phrases, sentences and other text
- Pragmatics – understanding the purpose or what the text is trying to achieve

Tasks that NLP is used for include text summarization, sentiment analysis, topic extraction, parts-of-speech tagging, stemming, text mining, machine translation, named entity recognition, relationship extraction, and automated question answering, language modeling and almost all other language-related tasks.

Some familiar real-world applications of NLP are Spam Filtering, Chatbots, Sentiment Analysis.

Some products that we use daily that ae powered by NLP are Alexa / Siri / Cortana, Grammarly, Google Translate or Microsoft Translate, Spell Check, Autocomplete, Immersive Reader, etc.

Syntactic analysis is used to understand how grammatical rules apply within a language and how these can be used to derive meaning from a group of words. Techniques used for this include:

- Lemmatization: reducing the inflected forms of a word into a single form for easy analysis
- Stemming: cutting the inflected words to their root form
- Morphological segmentation: dividing words into morphemes

- Word segmentation: dividing a continuous text into distinct units
- Parsing: grammatical analysis of a sentence
- Part-of-speech tagging: identifying the part of speech for every word
- Sentence breaking: placing sentence boundaries on a continuous text

Seven essential steps to help a computer understand natural language are:

- Sentence Segmentation
- Word Tokenization
- Text Lemmatization
- Stop Words
- Dependency Parsing in NLP
- Named Entity Recognition (NER)
- Coreference Resolution
- Sentence Segmentation


AI
ARTIFICIAL
INTELLIGENCE

UNIT 5

COMPUTER VISION

Key Learning Outcomes

At the end of this module, you will be able to:

a) Explain what is Computer Vision (CV) and how it works.
b) List common applications of CV
c) Describe the working and application of Optical Character Recognition (OCR)
d) Use OpenCV to read and image, display an image, resize an image
e) Implement a facial recognition AI project using a platform and OpenCV

Topic Covered

Introduction to computer vision | Computer Vision Applications | Challenges of Computer Vision | Optical Character Recognition | OpenCV | Project: Face Recognition using Scratch | Project: Attendance System using Face Recognition

5.1 INTRODUCTION TO COMPUTER VISION

In the base module, we introduced you to the exciting field of Computer Vision (CV). CV is the field of computer science that aims to replicate parts of the human vision system, to enable computers to identify and process objects in images and videos in the same way that humans do. We must not confuse Computer Vision with Image processing though.

EXPLORE

Learn about the difference here.

https://www.youtube.com/watch?time_continue=39&v=9-8Js62wzQs&feature=emb_logo

Thanks to advances and innovation in artificial intelligence, deep learning and neural networks, computer vision has taken great leaps in recent times. CV has been able to better humans in some tasks related to detecting and labeling objects in images and videos.

But despite our aim to mimic the human brain and its functions, we do not know how exactly do our brains work? Therefore, it is difficult to know how we can approximate that with our algorithms. The reality is that brain computation is still not completely understood, so although AI and Neural Nets including those in Computer Vision are supposed to “mimic the way the brain works,” nobody can say if that is true. We are not sure how the brain and eyes process images, therefore it’s difficult to say how well the algorithms used in CV and AI processes approximate our own internal mental processes.

But despite that over the past decade, systems have reached an accuracy of 99 percent from 50 percent. This makes them even more accurate than humans at reacting quickly to visual inputs.

5.1.1 HOW DOES COMPUTER VISION WORK?

Core to Computer Vision is pattern recognition. To enable pattern recognition, the first step is to train a computer to understand visual data. This is done by feeding it images, lots of images (thousands, millions if possible). These images need to have been labeled.

Then these images are subjected to various software techniques, or algorithms, that allows the computer to identify patterns in all the elements of these images that relate to those labels. It is an example of supervised learning.

Deep learning is a somewhat fundamentally different approach. Deep learning relies on neural networks that can extract common patterns between examples of labeled data provided to them and transform it into a mathematical equation that will help in identifying and classifying future images or pieces of data. It is an example of unsupervised learning. The difference between the two approaches is the kind of algorithms used.

For instance, to create a facial recognition application with deep learning, one is required to develop or choose a pre-defined algorithm and train it with images of people's faces, we want it to detect. Once it is trained with sufficient examples, the neural network used will be able to detect these faces without any direct programming or instructions that provide information on features or measurements of these faces.

Deep learning is very effective when the deep learning algorithm works on a large amount of labeled training data. Further tuning the parameters such as the type and number of layers of neural networks and training epochs (a measure of the number of times all of the training vectors are used once to update the weights). Deep learning is considered easier and faster to develop and deploy, compared to other types of machine learning.

Due to its ease and effectiveness, current computer vision applications mostly use deep learning; these include applications such as self-driving cars, tuberculosis or cancer detection and facial recognition. Deep learning and deep neural networks have been enabled in large measure due to the availability and more recent developments in hardware and cloud computing.

Another driving factor behind computer vision's growth is the sheer amount of data we generate and capture, that is then used to train and make computer vision better. Along with the massive amounts of visual graphic data (for example the billions of images shared online every day), the vast amount of computing power required to analyze the data is now available at low costs. All these developments, along with extensive and widespread use, new algorithms and developments in hardware have resulted in massive improvements in accuracy rates for object identification.

The advantage offered by ML and Deep learning is that developers no longer needed to hard-code every single rule into their vision applications manually. Now they program 'features' and smaller applications that could detect specific patterns in images. These combined with statistical learning algorithm such as linear regression, logistic regression, decision trees or support vector machines (SVM) help them to detect patterns, classify images and identify objects in them.

Machine learning has provided an alternative to traditional software development that required dozens of engineers, programmers and domain experts to work together and code every scenario and detail into the software. Also, such programs were often unable to handle situations where new situations arose. Machine learning has helped to overcome many such problems. For instance, in the past to predict a disease such as a lung or breast cancer, machine learning these days can do it without the need for dozens of engineers and breast cancer experts. It can be done with relative ease by training the computers using Computer Vision and loads of data.

EXPLORE

See this video to know more:

https://www.youtube.com/watch?v=ALQ_RNSRE40

Not only is the problem solved by Machine Learning, but it can also happen in lightning-fast speed today. This is because of ultra-fast chips and related hardware, along with a fast, reliable internet and cloud networks. One vital factor in the development of these solutions is that big companies doing AI research like Amazon, Google, Microsoft, Facebook, Intel and IBM are sharing their work, that allows others to use that to build on their previous work rather than start from the beginning. As a result, the AI industry is making rapid progress, and processes that took weeks or months some time ago can be done in minutes or even in real-time today. For these real-time applications of computer vision, this process is carried out continuously and in microseconds. This is why computer today can be termed to be what scientists call 'situationally aware'. Here is an example of how Computer Vision is being used in real-time in cancer prediction.

EXPLORE

See this video to know more:

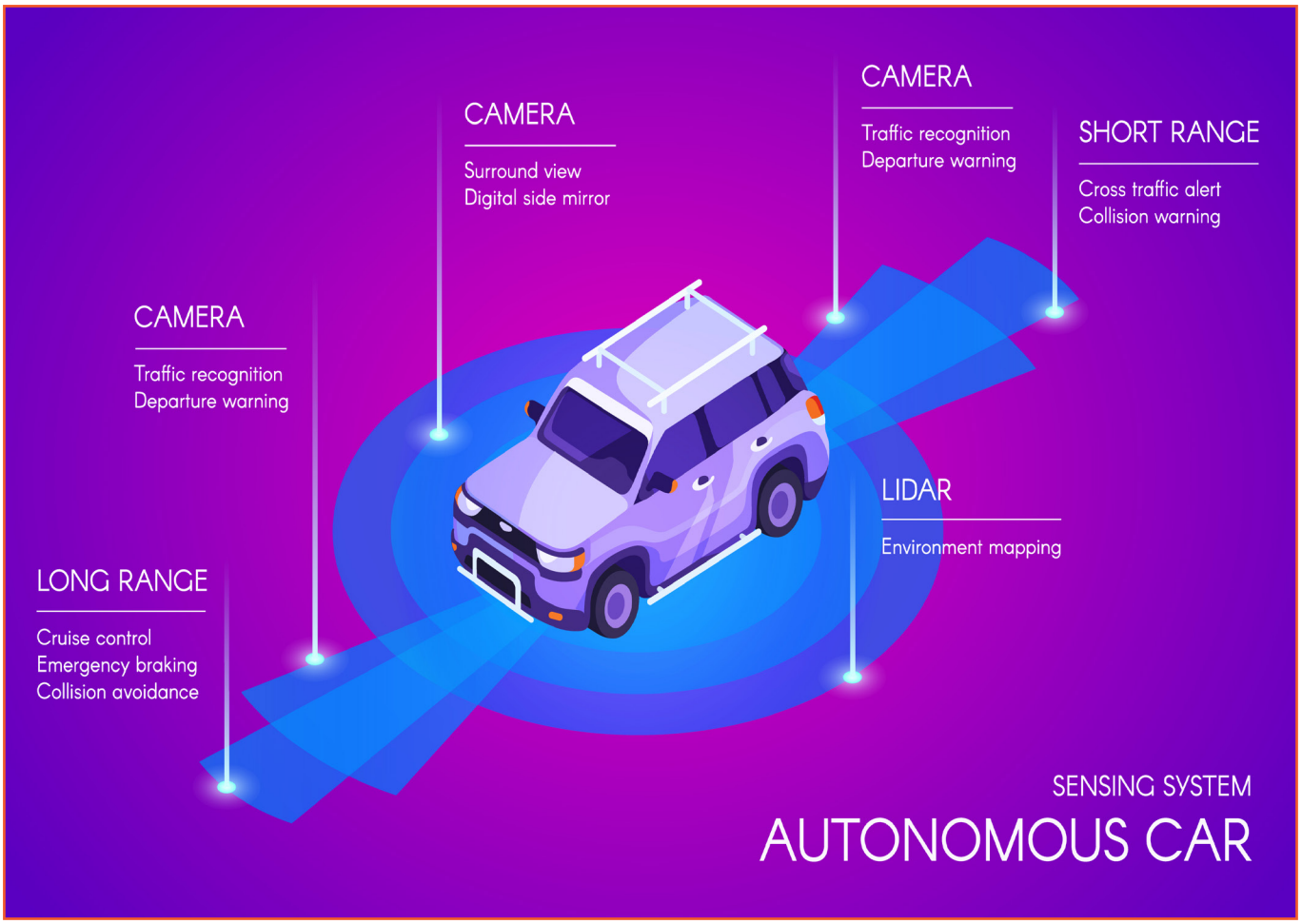
https://www.youtube.com/watch?v=9Mz84cwVmS0

5.2 COMPUTER VISION APPLICATIONS

Computer Vision is already being integrated into many major products that surround us every day.

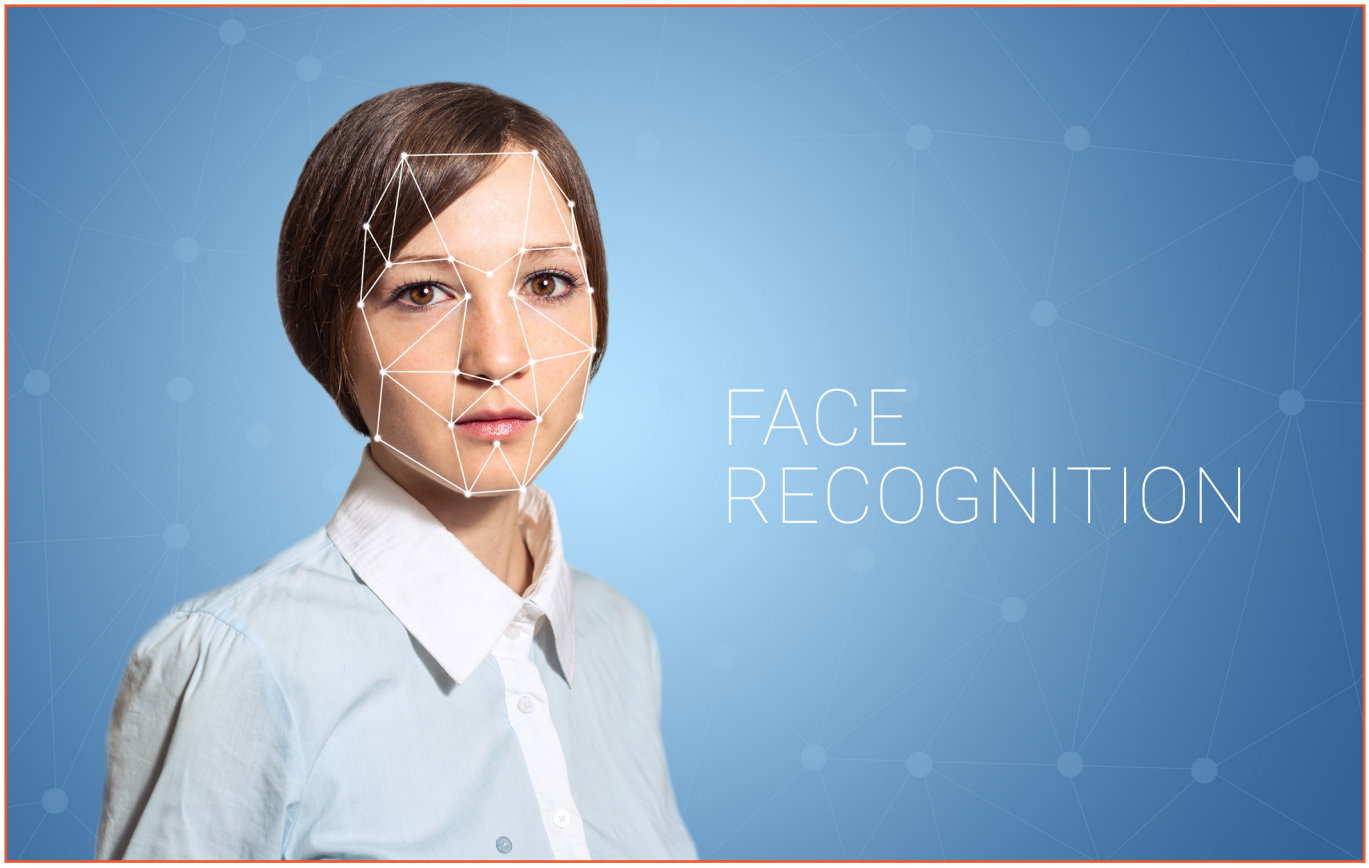
Self-Driving Cars

Self-driving cars make sense of their surroundings by using computer vision. Many different cameras around the car capture video from different angles and feed it to a computer vision software. These images are processed in real-time to interpret traffic signs, detect other cars, find the extremities of roads, locate obstacles, detect pedestrians, etc. This information enables the self-driving car to steer its way on the road, following traffic rules, avoiding obstacles, and drive its passengers safely to their destination.



CV In Facial Recognition

Facial recognition applications are another Computer Vision enabled technology that helps computers to identify people using images of people's faces. This is done through Computer Vision algorithms that detect facial features in images and compare them with stored databases of face profiles. Smartphone and other consumer devices authenticate the identities of their owners or authorized users relying on facial recognition technologies. Social media apps detect and tag users using facial recognition, similar to how law enforcement agencies rely on facial recognition to identify criminals in video feeds.



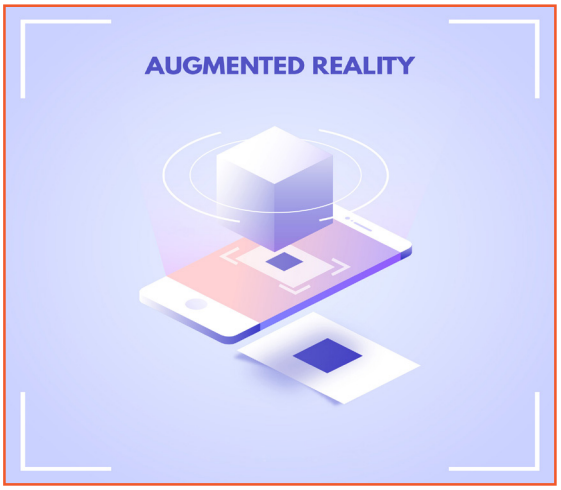
EXPLORE

See this video to know more:

https://video.sas.com/detail/videos/advanced-analytics/video/5410893556001/face-recognition-using-sas-viya?autoStart=true

CV In Augmented Reality & Mixed Reality

Computer vision also has an integral role in augmented and mixed reality. CV enables computing devices such as smartphones, tablets and smart glasses to overlay and embed virtual objects on real-world imagery. Using computer vision, AR gear detects objects in the real-world to determine locations on a device's display to place a virtual object.



EXPLORE

Read this article to see how AR works

https://www.clickz.com/how-ar-works/241718/

Computer Vision algorithms enable AR applications to detect surfaces such as tabletops, walls and floors, an integral part of establishing depth and dimensions and then placing virtual objects in the physical world.

CV In Healthcare

Computer Vision has also been an essential part of advances in health-tech as we have seen in the earlier cancer detection model. Computer vision algorithms can help automate many tasks, just like detecting various types of cancer and other diseases such as tuberculosis; another example is detecting symptoms in X-rays and MRI scans.

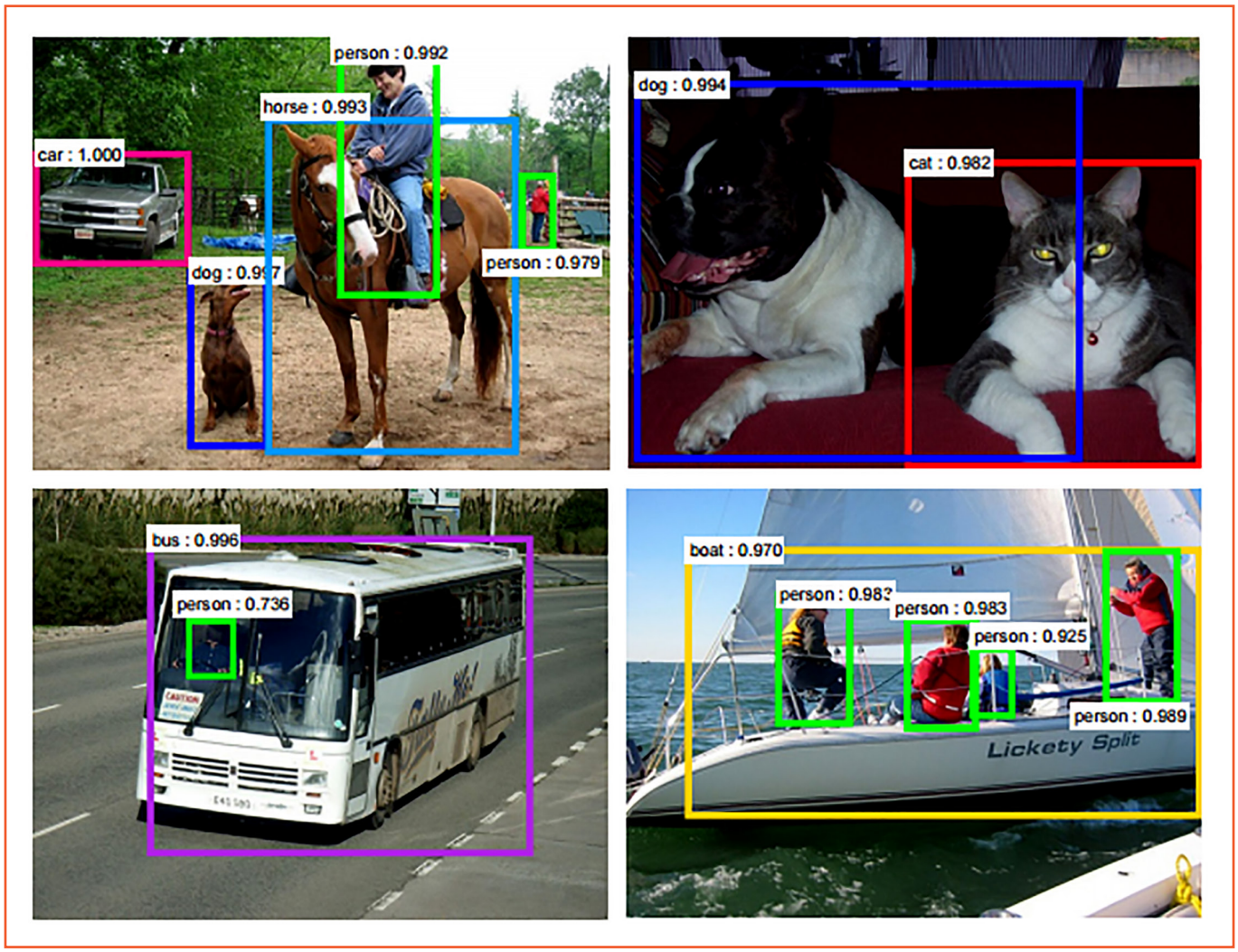
Here is an exciting video for seeing how CV is being used through smartphone devices in healthcare.

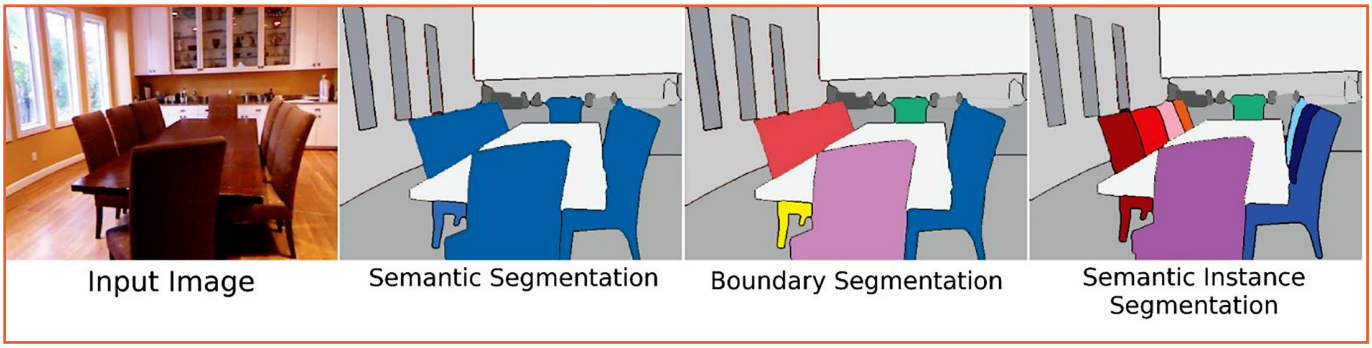
EXPLORE

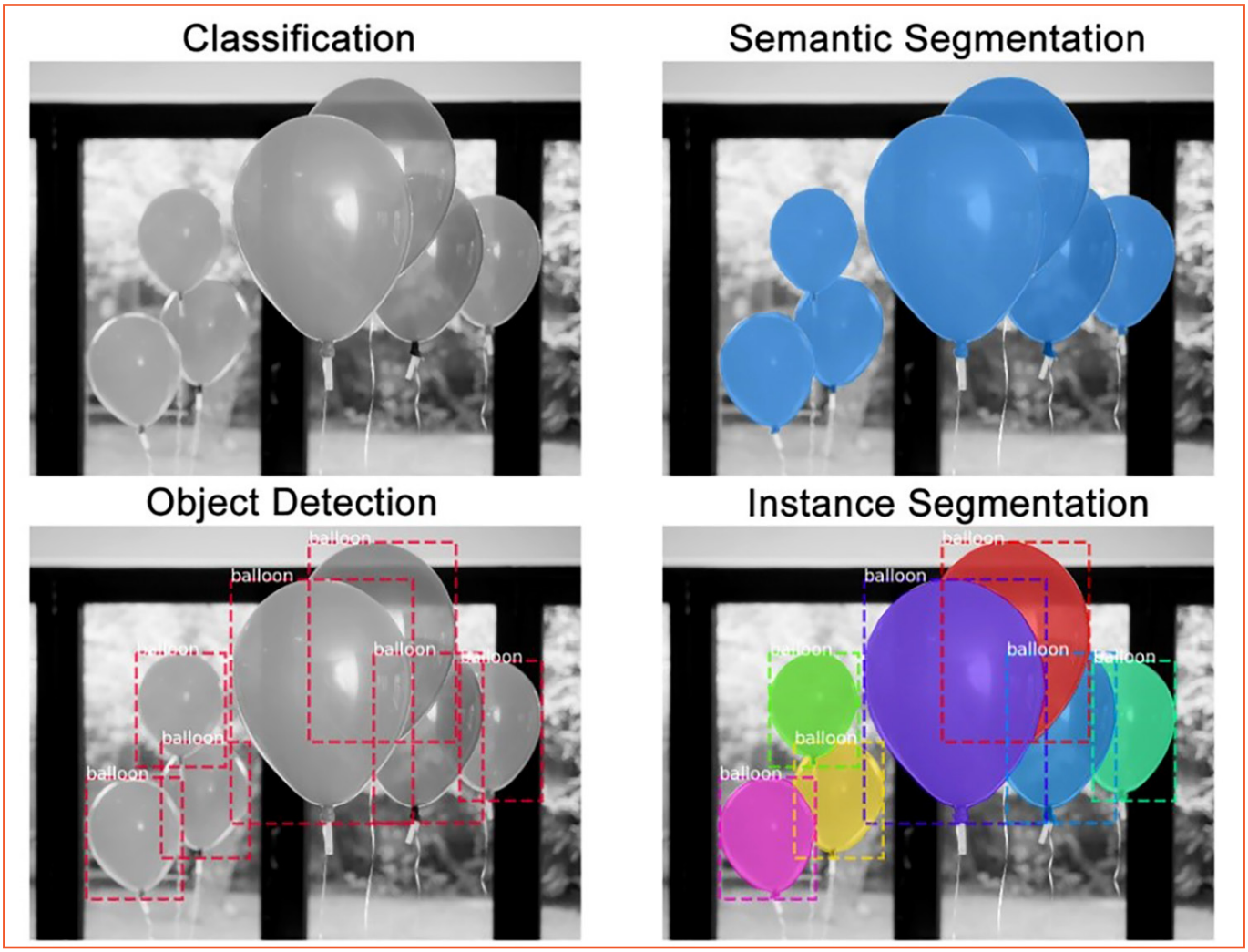
See this video to know more:

https://www.youtube.com/watch?time_continue=16&v=Dt-rC5nR-FE&feature=emb_logo

5.3 CHALLENGES OF COMPUTER VISION







Many popular computer vision applications involve operations to Detect, Recognize, Classify and Identify (DCRI) things in images; for example:

- **Image Classification –** which broad category does this image belong to? Involves assigning a label to the overall image.
- **Image Classification With Localization –** It involves not only classifying the image but also locating the object in the image with a bounding box.
- **Object Detection –** This involves locating different objects in the image and classifying them, using different bounding boxes and labels, for instance.
- **Object Landmark Detection –** This identifies critical points for the object in the image rather than use a bounding box.

- **Object Verification –** This involves finding out if a specific object is in the image.
- **Object Segmentation –** This involves a more closer defining of the object marking every pixel that belongs to the object.
- **Image Segmentation -** This involves the computer segmenting different sections of the image.
- **Image Style Transfer –** This involves the task of learning the style from one or more images and applying that style to a new image.
- **Image Colorization –** This involves converting a grayscale image to a full-color image.
- **Image Reconstruction –** This involves filling in missing or corrupt parts of an image.
- **Image Super-Resolution –** This involves generating a new version of an image with a higher resolution and detail than the original image.
- **Image Synthesis –** This involves generating targeted modifications of existing images or entirely new images.
- **Video Motion Analysis –** This involves using computer vision to estimate the velocity of objects in a video, or the camera itself.

EXPLORE

To know more about these read this article:

https://machinelearningmastery.com/applications-of-deep-learning-for-computer-vision/

EXPLORE

Here are some related videos:

https://www.youtube.com/watch?v=BMFAbYAxpII

https://www.youtube.com/watch?v=08myARQYXXY

https://www.youtube.com/watch?v=st_U59Yv3a8

5.4 OPTICAL CHARACTER RECOGNITION (OCR)

Let us now look at another application of AI which is OCR

Optical Character Recognition (OCR) is a technology used to identify characters – such as letters and numbers in an image. Also known as Text Recognition, OCR recognizes patterns and classifies information for Artificial Intelligence (AI) to use.

OCR helps institutions improve their productivity by processing large volumes of printed documents. After processing the pages and the text, these can be edited, searched, indexed, and retrieved quickly and efficiently.

5.4.1 HOW DOES OPTICAL CHARACTER RECOGNITION WORK

OCR goes through a series of steps:

First, a scanner is used that converts the paper document image to binary data. The scanned image's light and dark areas are analyzed by the machine, separating the writing from the background. The light areas are segmented and classified as background while the dark ones are classified as handwriting. This step is called Image Acquisition.

The image color or grayscale is converted to binary. During segmentation, feature extraction, and classification OCR finds the letters, numbers and symbols inside the dark areas, using a variety of techniques.

Then pattern recognition helps to identify text across various fonts and formats to compare characters from the scan. Various assigned rules help the OCR system to identify a single letter's different forms such as typed vs handwritten or upper vs lower case.

Then the characters are converted to ASCII (American Standard for Information Exchange), or similar codes that make it easy to identify and read the text. In the ASCII system, every letter is represented by a different 7-bit binary number.

OCR is widely used in several places. It is used for data entry. OCR is used for data classification, such as in post offices to sort letters and banks to deposit checks electronically.

OCR-driven apps include smartphone camera applications that scan business cards and other documents to digitally store this information on the phone or in the cloud.

OCR is also be used to convert documents into text for the visually impaired. OCR is used in translation apps, online text databases like Google Books and security cameras that recognize license plates.OCR is used in the examination system to mark large volumes of answer sheets.

5.4.2 HOW DO MACHINES SEE?

Machines analyze images based on various features, including colors, edges, etc. The techniques and approaches to analyze images are integrated within algorithms that help to detect these features and classify or identify them based on trained data. Some of the processes involved are:

Images are broken down into colors by using hexadecimal numbers to identify each color. A similar group of colors identified and then used to segment the image, for example, distinguish the foreground from background. Color gradient technique is used to find edges of different objects. The computer then uses algorithms to then search for specific features, also called as corners. These are lines that meet or intersect at an angle and cover a specific part of the image with a single color shade. These are considered the building blocks that enable the search for more detailed information within the image. Another vital aspect of identifying images or parts within it correctly requires determining textures in the image. The variance in textures between different objects helps to categorize an object correctly. Using these steps, the machine then makes a guess and matches the image with those stored in the database. The machine then based on algorithms and provided instructions verifies if the image and the objects within were correctly classified, then repeats the process as per requirement till a sufficiently good and reliable accuracy level is achieved.

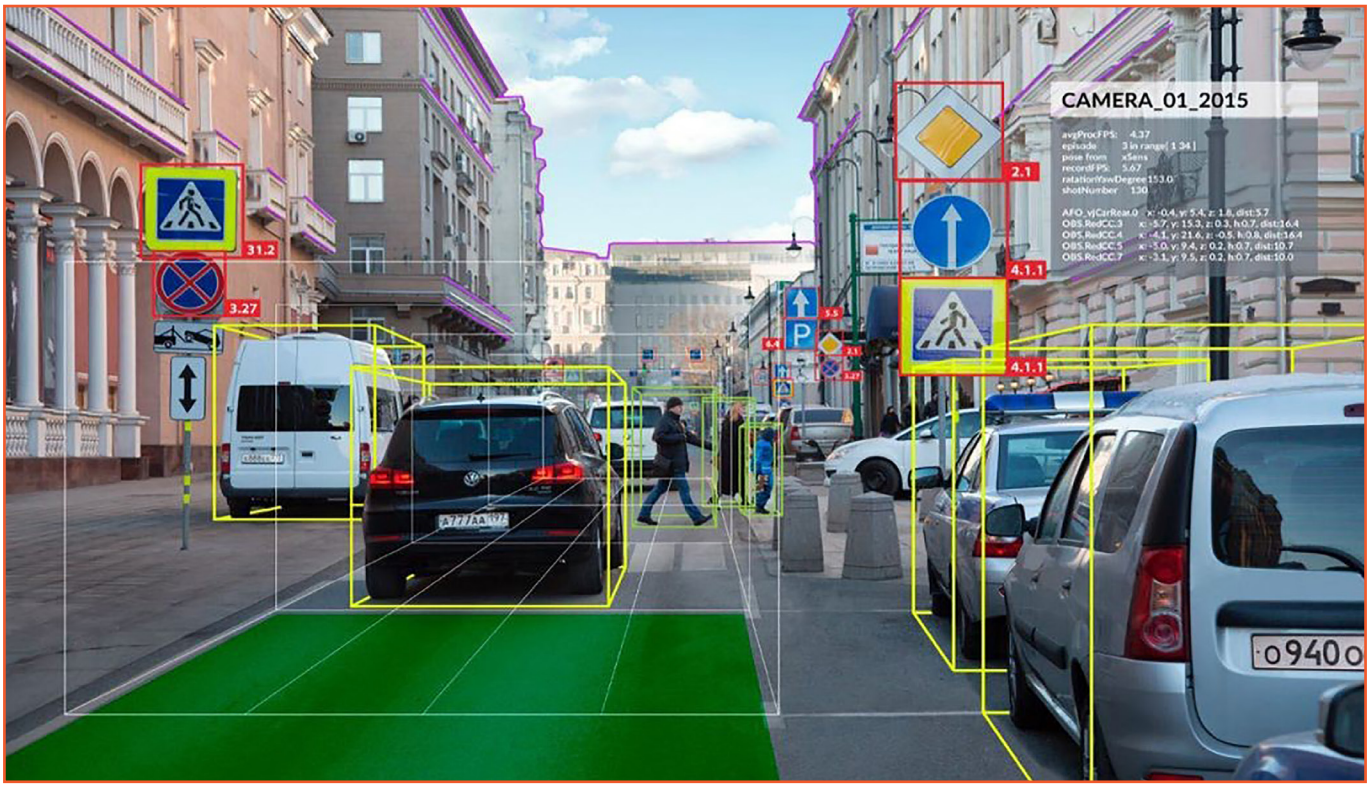
Despite the recent progress, which has been impressive, we're still not even close to solving computer vision. However, there are already multiple institutions and enterprises that have found ways to apply CV systems, powered by CNNs, to real-world problems.

So where can one practice computer vision?

There are many options and APIs available to learn and practice CV. OpenCV is one such option.

5.5 OPENCV

OpenCV (Open Source Computer Vision Library) is an open-source Computer Vision and Machine Learning software library. It mainly focuses on image processing, video capture, and analysis, including features like face detection and object detection. OpenCV was built to provide a common infrastructure for computer vision applications and to accelerate the use of machine perception in commercial products. Being a BSD-licensed product, OpenCV makes it easy for businesses to utilize and modify the code.



Let's play

What do you need before we start?

- Python
- An IDE
- OpenCV basics
- Import library
- import cv2

Read an image

We can use cv2.imread function to do this. First, make a variable and use the function to read the image using its path as the first parameter.

#For a colored image

img = cv2.imread(“Man.jpg”, 1)#For a grayscale image

img = cv2.imread(“Man.jpg”, 0)

In here, the ‘1’ and ‘0’ represent the type of the image. If we put ‘1’, then it will be read in RGB (Colored), or if we put ‘0’, then it will be read as grayscale or black and white image. OpenCV will read the read as a NumPy array or in English a matrix of numbers. So if the image is colored, it will be a 3d matrix, or if it is a grayscale one, it will be a 2d matrix.

If you print the img variable you will be able to see the image is now converted to a matrix! And if you print the shape of the image, you can see the Numpy array with the number of rows and columns.

print(img.shape)

Displaying the image

We can use cv2.imshow function to display the image you just read.

cv2.imshow(“Man”, img)

Here the first parameter is the name of the window and the second one is the object we read before. When we are using the imshow function, we have to use waitkey to wait until a user presses a key, unless the window gets closed as soon as it opens.

cv2. waitKey(0)

cv2.destroyAllWindows()

Resize an image

If you want to resize the image you read you can simply use cv2.resize function.

resized = cv2.resize(img, (650,650))

The first parameter is the object we read and the second one is the size we want it to be resized. You can also use the original image properties for the parameters as well. Here is how you can do it.

#New image shape = Old image shape/2

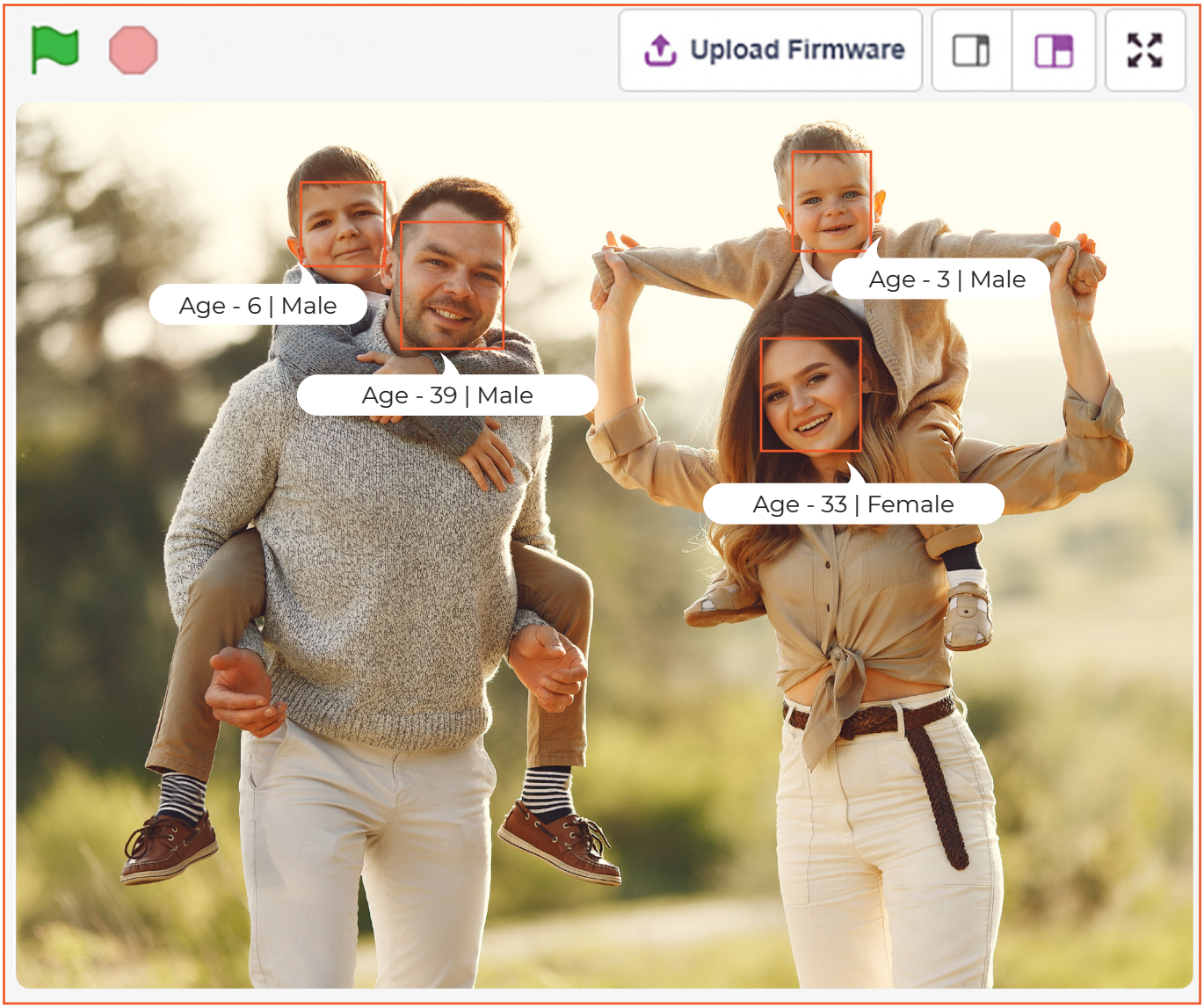
resized = cv2.resize(img,(int(img.shape[1]/2),int(img.shape[0]/2)))

So here you go with the super basic functionalities of computer vision using OpenCV. Let’s meet again to continue with the session later! Godspeed.

PROJECT: FACE RECOGNITION USING SCRATCH

Can you identify the age and gender of the people in this image?

Yes, you can! You can identify that there are 3 males and 1 female in the photo. You can also perhaps guess their age.



But can a machine recognize if there are any faces in the image and guess their gender, age, and emotion?

Well, in the past few years, face recognition has become one of the most promising applications of Computer Vision. Face detection can be considered as a substantial part of face recognition operations.

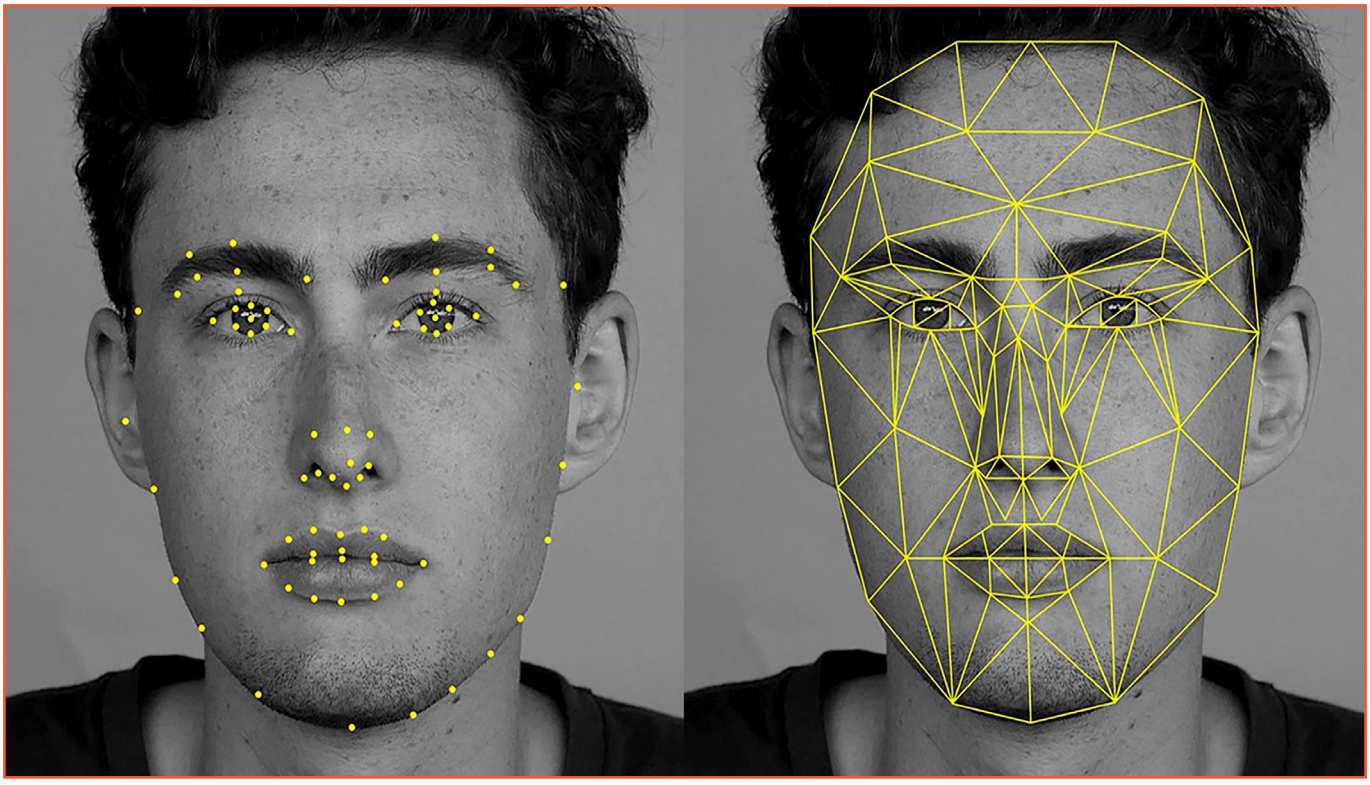
The method of face detection in pictures is complicated because, well, human faces are widely different from one another! They can have different poses, expressions, position, and orientation, skin color, have glasses or facial hair or not, etc. Then there are also differences in camera settings, lighting conditions, and image resolution.

Face detection is the action of locating human faces in an image and optionally returning different kinds of face-related data.

HOW DO WE DETECT FACES?

There are a few methods that you can use to detect a face. But here we will use feature-based detection. Feature-based detection uses what is known as face landmarks to detect faces.

Face landmarks are a set of easy-to-find points on a face, such as the pupils or the tip of the nose.



The feature-based method detects faces by extracting the face landmarks of the face.

Pros:

1. Features are invariant to pose and orientation change.
2. This method has a higher chance of face detection.

Cons:

1. Difficult to locate facial features due to noise in the images.
2. Difficult to detect features in complex backgrounds.

Once you have detected the face, you can classify the face for the age, gender, and emotion using machine learning algorithms and techniques.

Now it is time for you to do a face recognition project in Scratch. But do you think you can do one on your own? Give it a try!

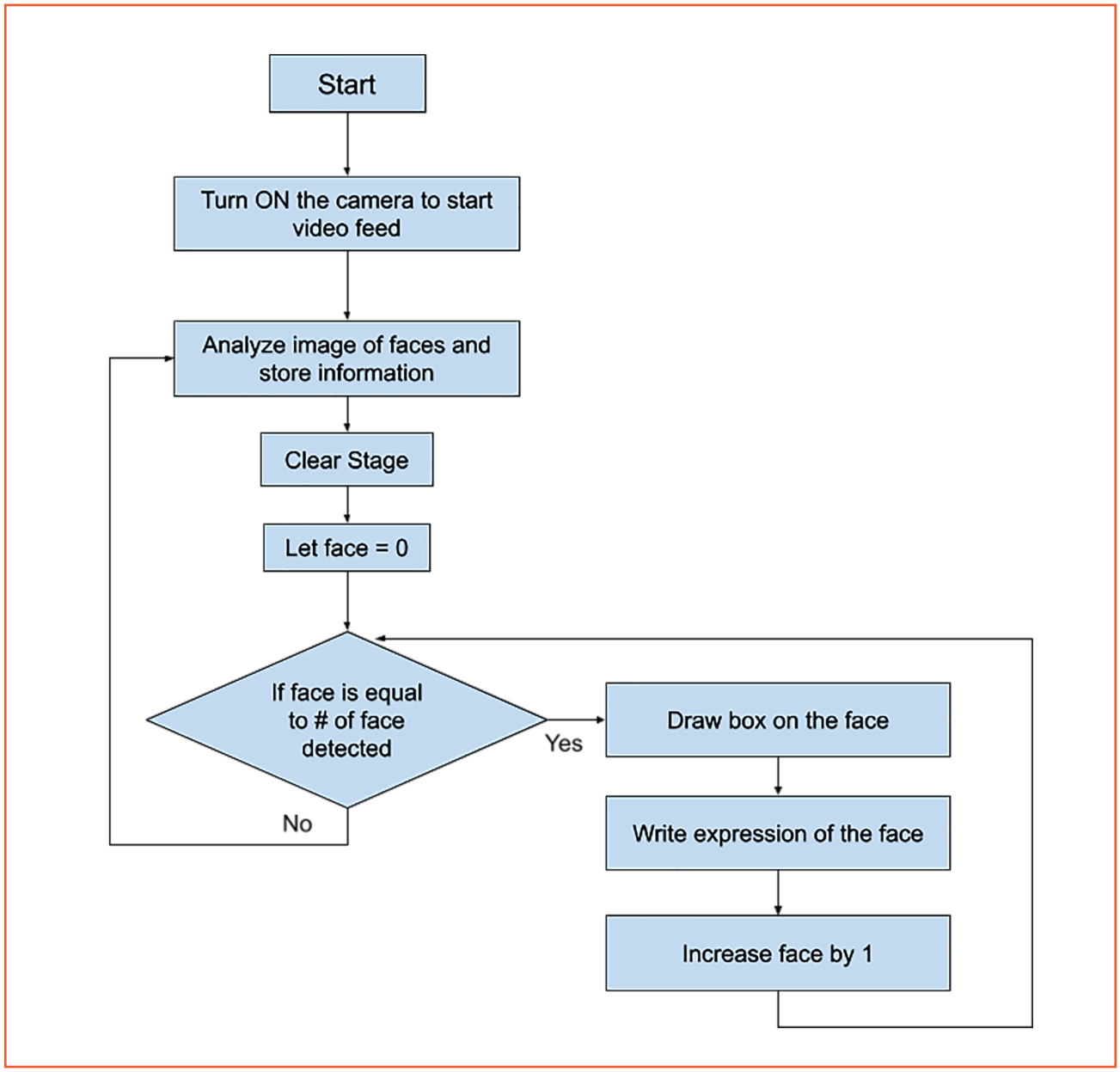
In this project, we will make a script that detects the face using the camera and reports the expression of all the faces detected on the stage. Let's begin!

COMPATIBLE PLATFORMS

1. Machine Learning for Kids
2. Stretch 3

FLOW CHART

The following flow chart shows the overview of the program required to implement Face Expression Recognition in the compatible platforms. You can find various graphical blocks in the compatible platforms to make the face expression recognizer. Now go ahead and try the project.



EXPLORE

Use the following video to get some guidance for the project.

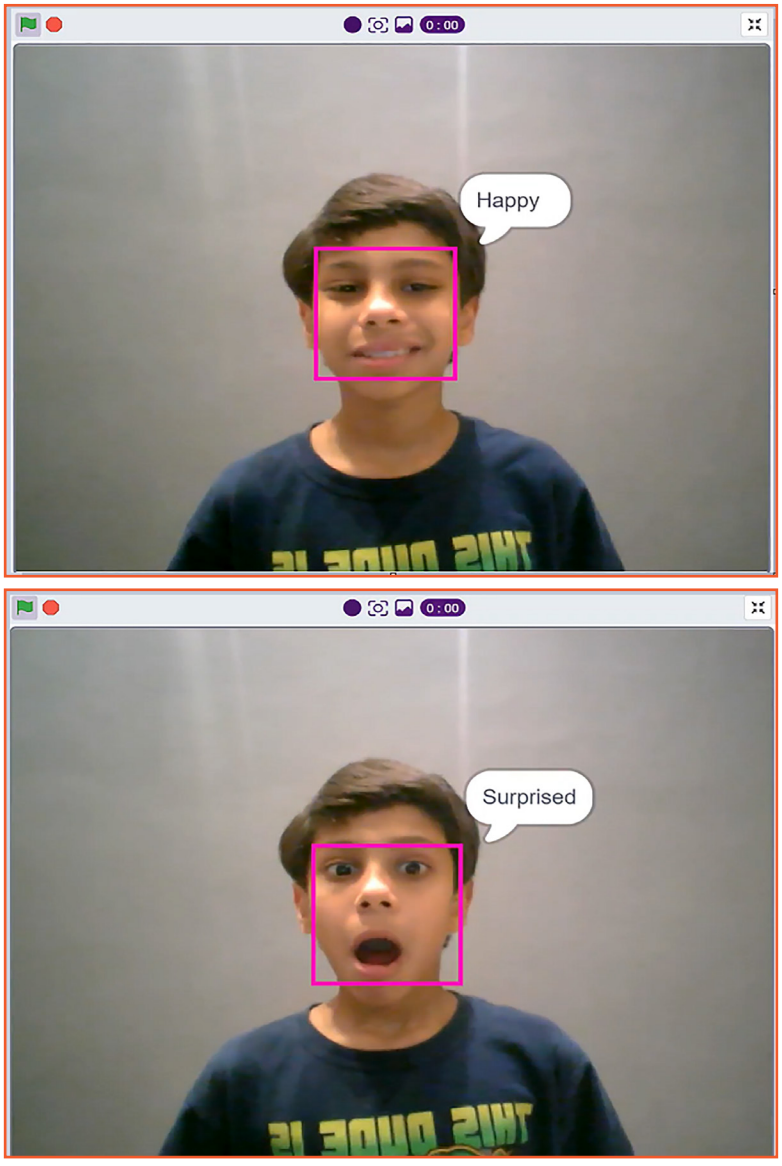
https://www.youtube.com/watch?v=6w9kvc_pN4o

HOW DOES IT WORK?

Let's understand the logic behind it i.e. how it works.

First, we will turn on the video feed and analyze the faces available in the frame and storing the information. Then we are setting initial value of face variable as zero and then we are comparing value of face variable as face detected inside the frame and if no face is detected then we are again analysing the video feed else we one by one mark the detected face by a square box and write its expression until face variable value becomes equal to the detected faces.

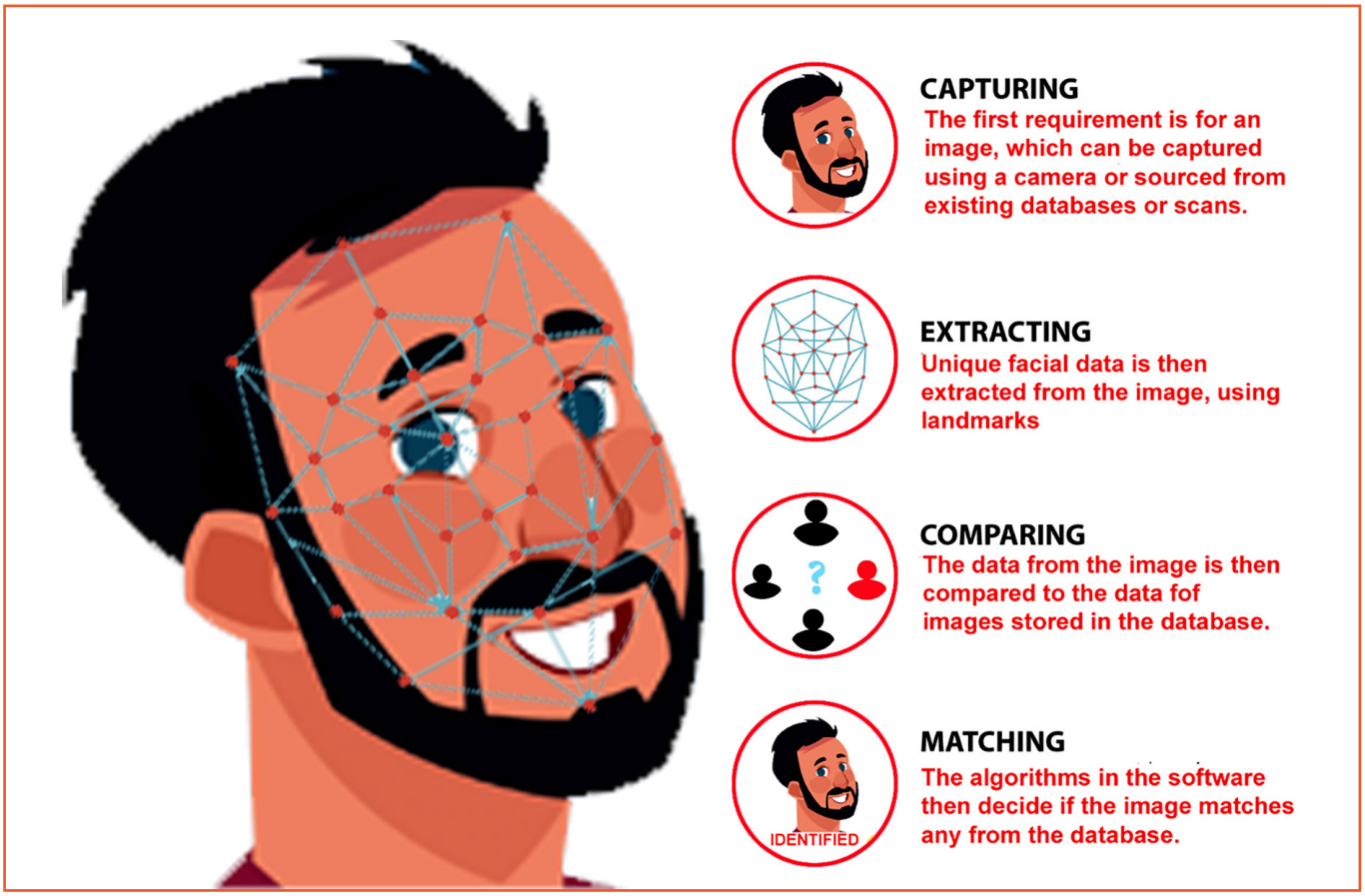
So, now it is time to test your face mimicking skills and simultaneously testing whether the machine is perfect to recognize it.



FACE RECOGNITION

Face recognition systems allow programs to identify a person by analyzing the face landmarks of her/his face in a digital image.

1. When a face is analyzed, its geometrical properties, size, and position are obtained.
2. A structure is generated that forms a pattern that is later stored in a database.
3. Once stored, the user's facial record is compared with the biometric features of each face that is scanned and allows adding new templates to make the recognition more effective.
4. If the image matches a search, the system will alert us that a match has been found.



PROJECT: ATTENDANCE SYSTEM USING FACE RECOGNITION

Learn how to make an attendance management system based on Face Recognition using Scratch and store data in real-time.

LEARNING OUTCOMES

1. Understand face recognition cycle
2. Training the AI model using image data sets
3. Data logging

Requirements: Laptop with a camera. There is no need for any additional hardware for this project.

COMPATIBLE PLATFORMS

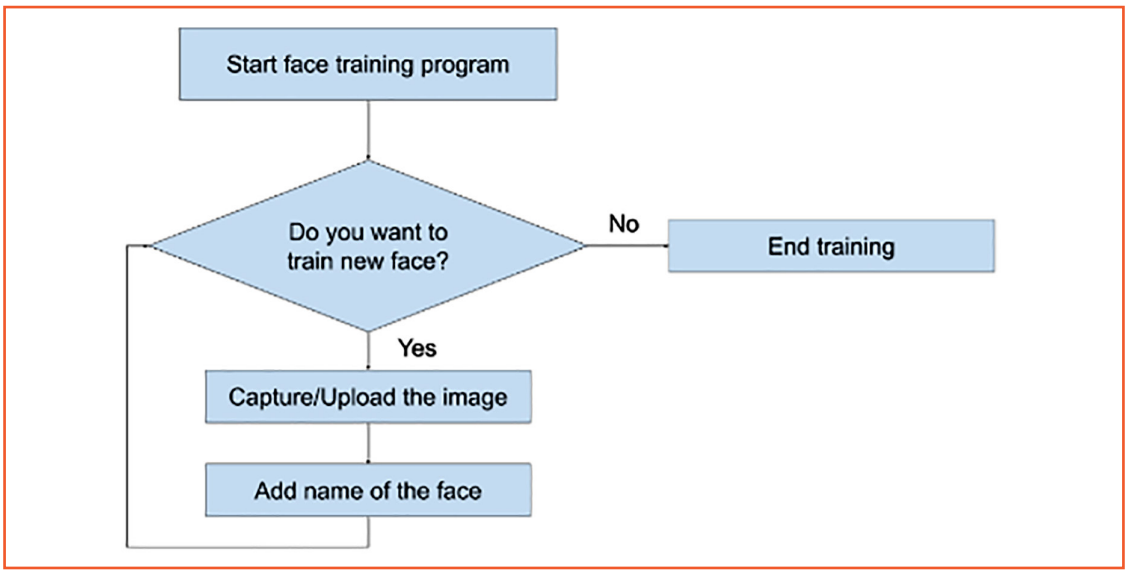
1. Stretch 3
2. PictoBlox

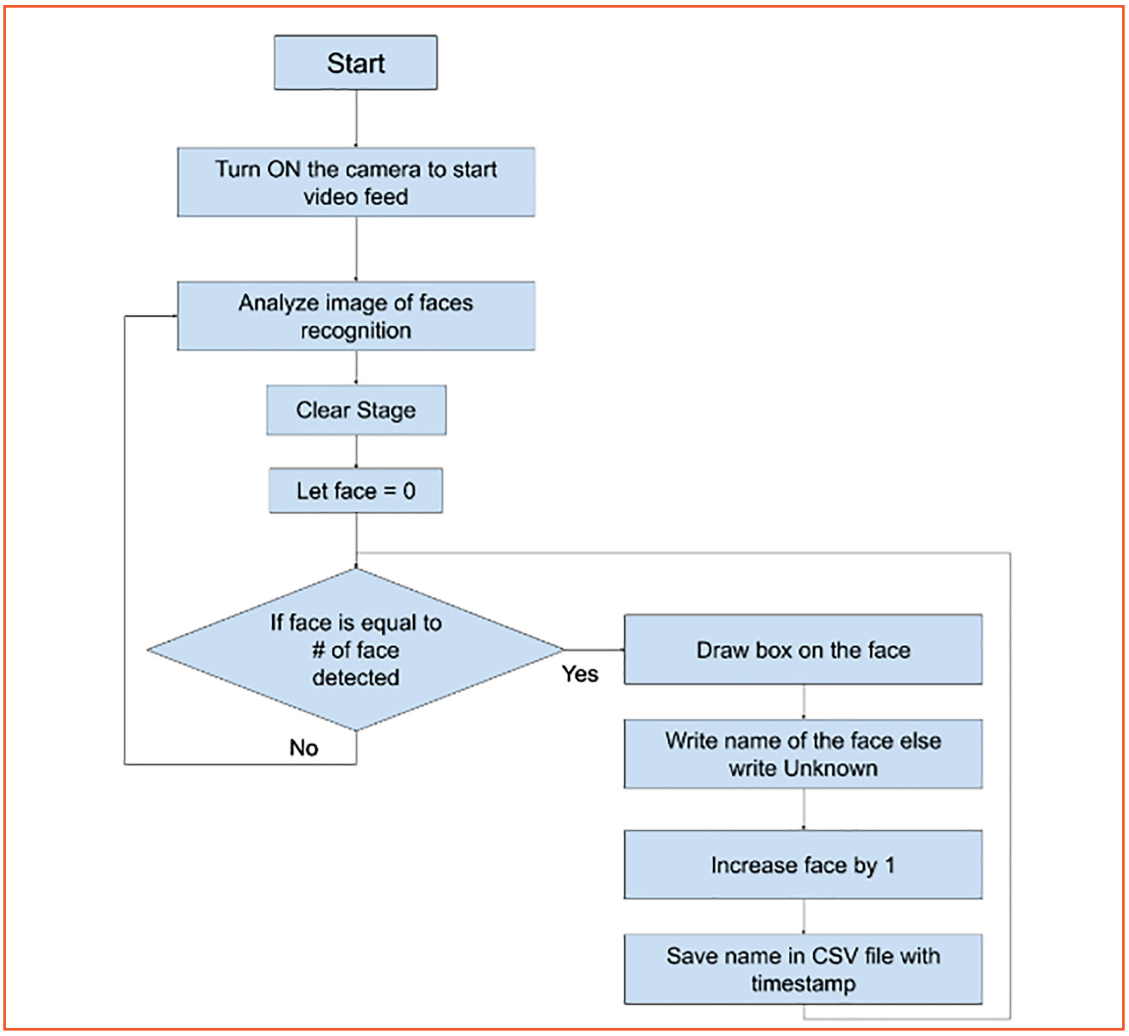
Because of the COVID-19 pandemic which started in early 2020, a lot of things are becoming contactless – from lift buttons to doors to sanitizer dispensers. So, why should the fingerprint-based attendance systems, widely used in offices and schools, be left behind? Thus, to make sure that we do not touch, we'll develop a contact-free, face recognition-based attendance system. Our system will first take a sample image of the person whose attendance needs to be tracked. The next time when the system scans the person's face, it will match the image with the sample image stored and will note down the time stamp for that person.

FLOW CHART

The following flow chart shows the overview of the program required to implement the **Attendance system using Face Recognition** in the compatible platforms. You can find various graphical blocks in the compatible platforms to make the face expression recognizer. Go ahead!

Face recognition





SUMMARY

Computer Vison (CV) is the field of computer science aims to replicate parts of the human vision system, to enable computers to identify and process objects in images and videos. Core to CV is pattern recognition.

To enable pattern recognition, the first step is to train a computer to understand visual data by feeding it lots of images. These images should be labelled. Then these images are subjected to various software techniques, or algorithms, that allows the computer to identify patterns in all the elements of these images that relate to those labels. It is an example of supervised learning.

Deep learning can also be employed for this, which relies on neural networks that extract common patterns between examples of labelled data and transform it into a mathematical equation that will help in identifying and classifying future images. It is an example of unsupervised learning.

Some applications of computer vision in the real world are self-driving cars, facial recognition by law enforcement agencies, Augmented Reality & Mixed Reality projects, disease detection, etc.

Many computer vision applications involve Detect, Recognize, Classify and Identify (DCRI).

Optical Character Recognition (OCR) is a technology used to identify characters – such as letters and numbers in an image. It is also known as Text Recognition. It helps institutions improve their productivity by processing large volumes of printed documents so that they can be edited, searched, indexed, and retrieved efficiently.

OCR goes through a series of steps. First, a scanner converts the paper document image to binary data. The scanned image's light and dark areas are analyzed by the machine, separating the writing from the background. The light areas are segmented and classified as background while the dark ones are classified as handwriting. This step is called Image Acquisition. The image color or grayscale is converted to binary. During segmentation, feature extraction, and classification, OCR finds the letters, numbers and symbols inside the dark areas, using a variety of techniques. Pattern recognition helps to identify text across various fonts and formats to compare characters from the scan. Then the characters are converted to ASCII (American Standard for Information Exchange), or similar codes.

OCR is widely used in post offices to sort letters, in banks to deposit checks electronically, to convert documents into text for the visually impaired, in translation apps, in online text databases like Google Books, with security cameras that recognize license plates and in the examination system to mark large volumes of answer sheets.

OpenCV (Open Source Computer Vision Library) is an open-source computer vision and machine learning software library. It mainly focuses on image processing, video capture, and analysis, including features like face detection and object detection.

Face detection is the action of locating human faces in an image and optionally returning different kinds of face-related data.

Face recognition systems allow programs to identify a person by analyzing the face landmarks of her/his face in a digital image.


AI
ARTIFICIAL
INTELLIGENCE

UNIT 6

DATABASES AND SQL

Key Learning Outcomes

At the end of this module, you will be able to:

a) Create a relational database with tables filled with Fields and their values
b) Identify primary key, cardinality and parent and child in relational databases
c) Use SQL to create, update, delete, query and modify data from a relational database

Topic Covered

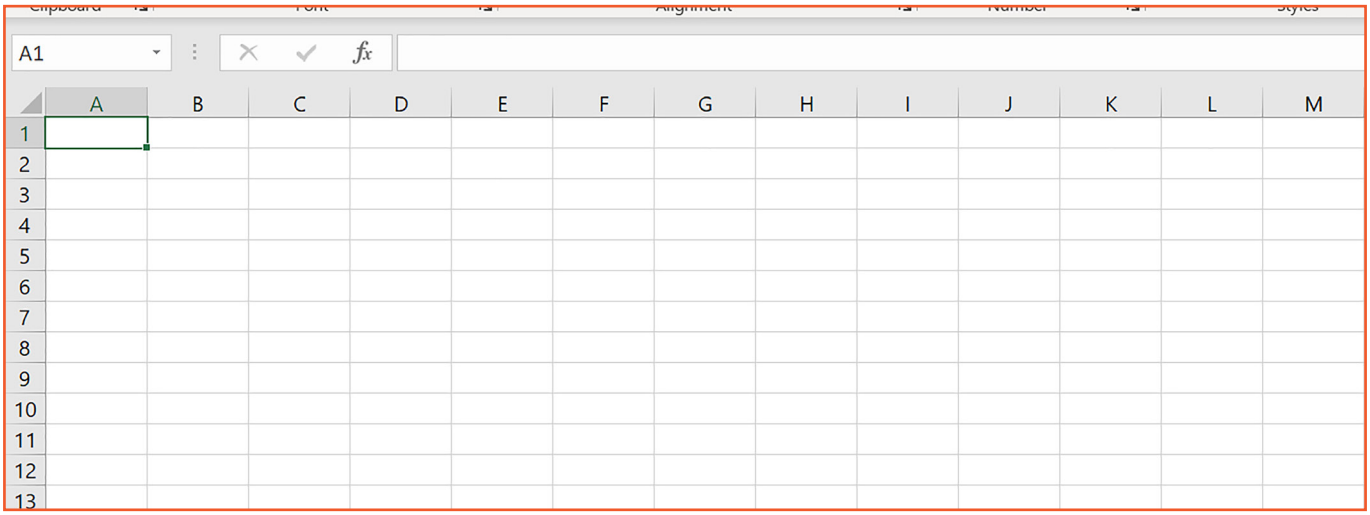
Databases | SQL

6.1 DATABASES

Databases offer a convenient and powerful way to store and organise information.

6.1.1 SPREADSHEET

A Spreadsheet is a computer program that looks like a large sheet of paper with boxes. It has a grid of cells that form rows and columns.



Users can write words or numbers in the cells. They can make it in the form of headings with their values given against them. These files then can be saved. Usually, the rows in a table are called records and the columns in a table are called fields or attributes.

	NAME	ENGLISH	HINDI	MATHS	SCIENCE	SOCIAL STUDIES	TOTAL
3	APPU	64	85	98	88	72	407
4	AYESHA	82	88	85	90	79	424
5	BIJU	92	75	92	89	63	411
6	DEEPA	75	90	62	59	81	367
7	KEVIN	69	61	75	79	53	337
8	KIARA	75	91	54	69	78	367
9	MANJEET	70	82	78	81	72	383
10	PARI	88	90	99	90	75	442
11	RAGHU	79	80	97	95	70	421
12	SAMAR	88	85	63	66	69	371
13	SOHINI	90	91	91	90	78	440
14	SRINIVAS	75	76	89	92	78	410

We can also perform many actions on it, such as:

- Move the grid lines to make some cells are bigger than others.
- Change the color or size of text/font or cell.
- Put pictures in the worksheets, make multiple pages.
- Print out their results on real paper.
- Sort the data in increasing or decreasing order.
- Filter data matching with certain criteria.

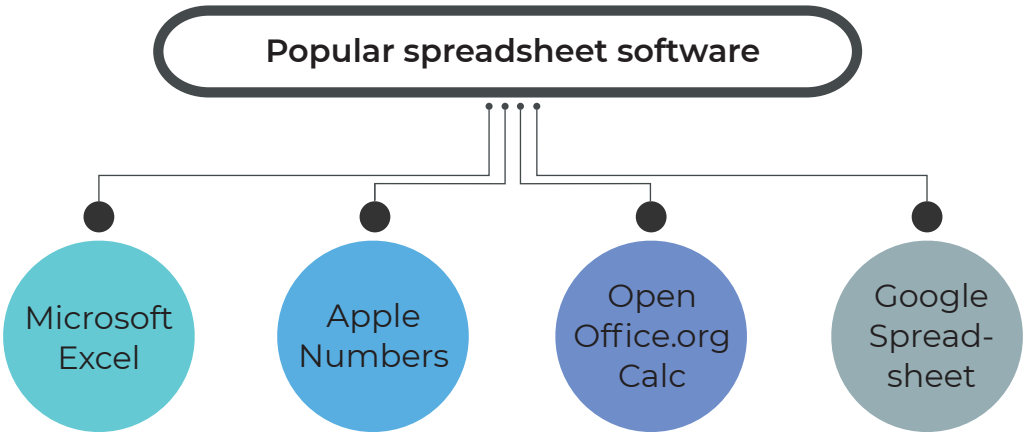
The main reason most people use spreadsheets is for performing various types of automatic calculations such as:

- Count items, identify maximum or minimum.
- Perform various arithmetic calculations such as addition, subtraction, multiplication, division, etc.
- Calculate averages, percentages, square roots, sine, cosine, tangents, logs, etc.
- Apply standard formulas that can be used like that of interest rate, price value, yield, etc.

One can even write a formula and apply on the data. This means the computer does all the math, and people only need to give the correct information.

Using spreadsheets, one can make calculations on a large set of numbers in seconds, that would take a very long time if done manually.

Another strength of spreadsheets is that they can produce diagrams, like graphs and pie-charts, based on the data the user enters. Sometimes the numbers make more sense when the computer turns them into a picture.



ACTIVITY 1 (Part 1)

Capturing and Recording Data

Given below are some open sources of data shared by the Indian Government regarding Economic activity, population, weather, etc.

Select a set of data that you would like to analyze and copy that data along with the headings in a spreadsheet such as Excel.

Save the sheet and give it a name.

After each of the next few sections, we will apply our learnings to this database.

EXPLORE

Indian Government:

https://data.gov.in/

http://www.surveyofindia.gov.in/

http://mospi.nic.in/data

https://bhuvan-app3.nrsc.gov.in/data/download/index.php

https://dbie.rbi.org.in/DBIE/dbie.rbi?site=home

6.1.2 SORTING DATA

What is data sorting?

Data sorting is the process of arranging the data into a meaningful order to enable the viewer to analyze the information more effectively. The choice of data sorting is selected as part of the query. There are two methods of data sorting available:

- Data (numbers or characters) sorted into ascending order is when they are arranged from the smallest to the largest number, e.g. 7, 12, 19, 23 and 37 or A, C, E, H, M, X, and Z.

- Data (numbers or characters) sorted into descending order is when they are arranged from the largest to the smallest number, e.g. 37, 23, 19, 12 and 7 or Z, X, M, H, E, C, and A.

How does the computer perform the sort?

The computer sorts the elements of a list into the desired order using a Sorting Algorithm. There are many different types of sorting algorithm, some more or less efficient than others, but the most commonly taught in schools are the Bubble, Insertion and Merge sorts.

Let's say you had the spreadsheet above and wanted to sort by age. This process is fairly simple. You can click on the first cell in the column that has to be sorted to get started. Then you will:

- Right-click to open the menu
- Go down to the Sort option – when hovering over Sort, the sub-menu will appear
- Click on Largest to Smallest
- Select Expand the selection
- Click OK

The whole table has now adjusted for the sorted column. Note: when the data in one column is related to the data in the remaining columns of the table, you want to select Expand the selection. This will ensure the data in that row carries over with sorted column data.

ACTIVITY 1 (Part 2)

If the data provided is not sorted. Apply a sort to at least two columns. If it is already sorted then change the sort, i.e. make it ascending instead of descending or vice-versa.

You will notice that all the data in the table rearranges when you sort on one heading.

If you select the whole column and the sort it, then only the data in that column gets sorted, but now it will have changed it's row headings and row-wise alignment with rest of the data.

Make a note of information that you can extract from the data using the sorting function.

6.1.3 FILTERING DATA

The filter feature applies a dropdown menu to each column heading, allowing you to select specific choices to narrow a table. Using the above example, let's say you wanted to filter your table by gender.

To do this using the filter you would:

- Go to the Data tab on the Excel ribbon
- Select the Filter tool
- Click on the arrow next to Gender to bring down the dropdown menu
- Select Female and ensure all other boxes are unticked

ACTIVITY 1 (Part 3)

Now select all the column headings of your database and activate the filtering option for them.

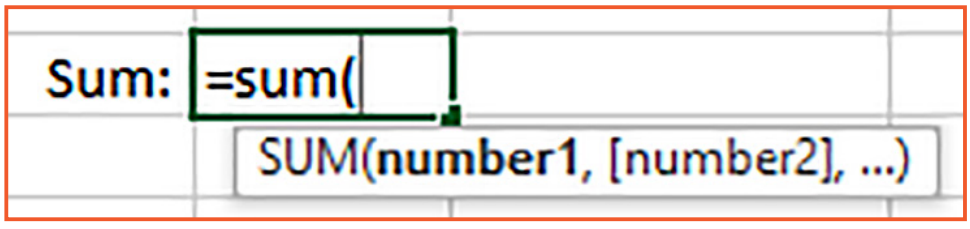
Now select a certain type of data from any of the columns by using the filter.

You can apply filters for greater than a value, less than a value, a value containing something or value not containing something as well. Find out how.

Now write down three new findings that you were able to make from your database, using filters.

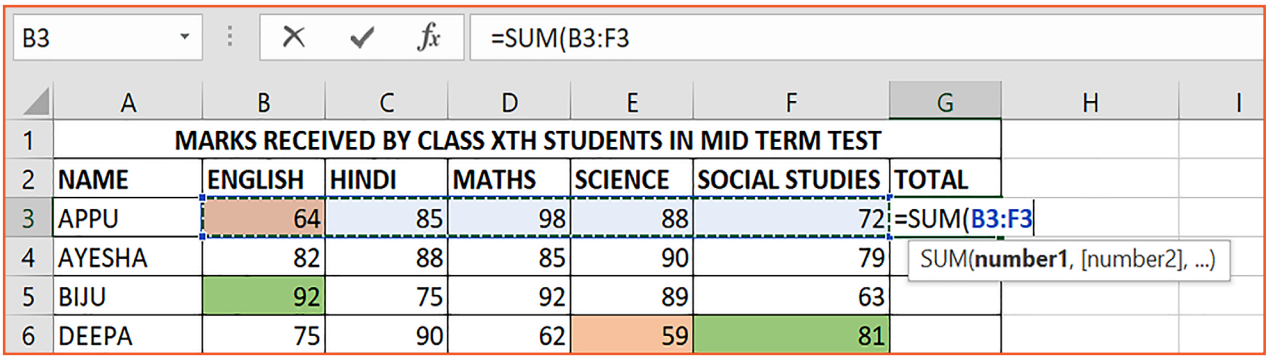
6.1.4 APPLYING SIMPLE FORMULAE

Excel formula can be calculated in any empty cell, but preferably at the end of the data in the relevant column or row. Excel formulas all start with an = sign written in the empty cell, followed by the function name (In this case 'sum') and a section in parentheses to provide ranges or arguments. Start out by typing =sum(



After the start parentheses enter the range of cells that contain the data you want to add together. You can this in the following ways:

- Type your range in by writing the starting cell(Column and row name) followed by a colon, followed by the ending cell (column and row name). E.g. D2:D20
- Select the respective cells using your mouse or input device. In the following case, we will select B3:F3.



	A	B	C	D	E	F	G	H	I
1	MARKS RECEIVED BY CLASS XTH STUDENTS IN MID TERM TEST								
2	NAME	ENGLISH	HINDI	MATHS	SCIENCE	SOCIAL STUDIES	TOTAL		
3	APPU	64	85	98	88	72	=SUM(B3:F3		
4	AYESHA	82	88	85	90	79			
5	BIJU	92	75	92	89	63			
6	DEEPA	75	90	62	59	81			

Next you can close the parenthesis, although newer versions of Excel will auto close it for you.

By applying the same method, you can calculate the average, minimum value, maximum value, and count:

Sum → =SUM(B3:F3)

Average → =AVERAGE(B3:F3)

Maximum Value → =MAX(B3:F3)

Minimum Value → =MIN(B3:F3)

Count of entries → =COUNT(B3:F3)

ACTIVITY 1 (Part 4)

In your database, apply the Count, Sum, Average, Minimum and Maximum formulae for at least three different ranges.

Extract three new insights that you got by using these formulae on your database.

6.1.5 RELATIONAL DATABASES

A database can contain one or more tables of information.

- A database that contains only one table is called a flat database.
- A database that contains two or more related tables is called a relational database.

There are other more complex kinds of databases, but here we will focus on relational databases.

Here's an easy way to understand the benefits of dividing the data into multiple tables:

Imagine that you are responsible for keeping track of all the books being checked out of a library. You could use a single table (a flat database) to track all the critical information:

First Name	Last Name	Address	Phone	Book Title	Due Date
Pari	Kumari	ABC, D1, EF STREET	4111111	The Guide	30-08-20

This table meets the basic need to keep track of who has checked out which book, but does have some serious flaws in terms of efficiency, space required and maintenance time. For example, as voracious reader Pari checks out more books over time, you will have to re-enter all of her contact information for every book.

First Name	Last Name	Address	Phone	Book Title	Due Date
Pari	Kumari	ABC, D1, EF STREET	4132549	The Guide	30-08-20
Biju	Joseph	DEF, A1, XY STREET	4156772	David Copperfield	30-08-20
Pari	Kumari	ABC, D1, EF STREET	4132549	Wings of Fire	15-09-20
Samar	Shareef	JKL, A1, XY STREET	4254661	The Story of My Experiments With Truth	15-09-20

Sohini	Saha	MNO, P1, UV STREET	4321589	Mill on the Floss	15-09-20
Pari	Kumari	MNO, D1, EF STREET	4132549	Malgudi Days	29-09-20

To re-enter Pari's contact information wastes time and increases the opportunity for error. Moreover, when an update is necessary (e.g. Pari's phone number changes), each of Pari's records must be located and corrected. If one of Pari's records has a different address from the rest, is it a correction, a record overlooked during the last update, or a data-entry mistake?

These problems can be decreased by normalizing our data – in other words, dividing the information into multiple tables with the goal of having 'a place for everything, and everything in its place.' Each piece of information should appear just once, simplifying data maintenance and decreasing the storage space required.

STUDENTS TABLE				
Student Id	First Name	Last Name	Address	Phone
0001	Pari	Kumari	ABC, D1, EF STREET	4132549
0002	Biju	Joseph	DEF, A1, XY STREET	4156772
0003	Samar	Shareef	JKL, A1, XY STREET	4254661
0004	Sohini	Saha	MNO, P1, UV STREET	4321589

CHECKOUT TABLE		
Student Id	Book Title	Due Date
0001	The Guide	30-08-20
0002	David Copperfield	30-08-20
0001	Wings of Fire	15-09-20
0003	The Story of My Experiments With Truth	15-09-20
0004	Mill on the Floss	15-09-20
0001	Malgudi Days	29-09-20

Now that the data are arranged efficiently, we need a way to show which records in the STUDENTS table correspond to which records in the CHECKOUT table – in other words, who checked out which book. Instead of repeating everything we know about a patron whenever he checks out a book, we will instead give each library patron an ID, and repeat only the ID whenever we want to associate that person with a record in a different table.

Now the PATRONS and CHECKOUT tables can be related (how relationships are formally declared in various database software is beyond the scope of this paper).

Let's learn some new terms to talk about our related tables.

The primary key:

It is a field whose values are unique in this table, and so can be used as identifiers for the records.

In table STUDENTS, the Student ID field is the primary key and so its values must remain unique. For example, the value "0002" can appear only in one record - Biju's - and Biju can have only one Student ID - "0002."

Is the Student ID field in table CHECKOUT the primary key?

We can see that it contains duplicate values, so the answer is No. If Student ID were the primary key for CHECKOUT, each person would only be permitted to check out one book, and afterwards would be forbidden to check out any more books, ever.

So if Student ID is not the primary key for table CHECKOUT, which field is?

We can't make Book Title the primary key, or we'd have a similar problem – each book could only be checked out once, and afterwards, no one would be permitted to check it out ever again. We can't make 'Due Date' the primary key, or else only one book could be due each day. Since none of the existing fields works as a primary key, we will add a new field to hold an identifier for each record. We could name this field Checkout ID.

CHECKOUT TABLE			
Checkout Id	Student Id	Book Title	Due Date
0001	0001	The Guide	30-08-20
0002	0002	David Copperfield	30-08-20
0001	0001	Wings of Fire	15-09-20
0003	0003	The Story of My Experiments With Truth	15-09-20
0004	0004	Mill on the Floss	15-09-20
0001	0001	Malgudi Days	29-09-20

The ID fields can be related. For example, Checkout ID in CHECKOUT is related to Student ID in STUDENTS.

To further increase efficiency, we could also use the Book ID here and have another table where the Book ID is connected to Book Title and Author Name.

Parent and Child tables

When two tables have an unequal relationship, we call the independent table the parent and the dependent table the child. You can identify the parent table by determining which table could contain a record without needing a corresponding record in the table on the other side of the relationship.

For Example:

Is it possible to have a record of a student who has never checked out any book? Yes. If the student is not yet started taking books.

Is it possible to check out a book to a student, who doesn't have his or her records of address and phone number captured? No, the library will not issue to a student who is not registered.

This implies that the STUDENT Table is the Parent Table and the CHECKOUT TABLE is the Child table.

If somehow the child table contains a record that does not have a corresponding record in the parent table, that record is called an orphan. Orphaned records are a problem that generally requires attention from the database administrator.

Cardinality in a related database

It describes how many records in one table can be related to records in another table. Two tables might have a cardinality of:

- 1 - 1 (one to one)
- 1 - oo (one to many)
- 1 - 3 (one to three)
- oo - oo (many to many), etc.

The STUDENTS – CHECKOUT relationship has a 1 - oo cardinality, because 1 patron may have any number of checkouts, from zero to infinity.

The CHECKOUT – STUDENTS relationship has a cardinality of oo - 1.

If we really were designing the data model (tables, fields, relationships, etc.) for a library, we would continue by separating even more data (such as the authors) into other tables until the model was as efficient as possible.

6.1.6 DATABASE MANAGEMENT

You have just learnt some basic concepts of databases, using the Spreadsheets. However, Spreadsheets can hold only a limited amount of data. AI requires the use of very large amounts of data that could carry thousands or millions of records.

As businesses are expanding, large enterprises need to respond to a lot of queries instantly. Therefore, the database handlers or administrators focus on improving the performance, which involves employing various techniques. While these techniques are being implemented, enterprises face various challenges, such as:

- Increase in the volume of data to be stored
- Keeping the stored data secured
- Meet the increased demand for new data and opportunities
- Maintaining the database and the related infrastructure
- Ensuring continuous growth of the business by predicting organizational capacity

It is a tedious task to address all the challenges mentioned above. The process can be very time-consuming and prevent database administrators from focusing their attention to more strategic tasks.

Hence, specialized database development and management tools such as like SQL Server Management Studio (SQL Server), Oracle SQL Developer (Oracle), MySQL Workbench (MySQL), MS Access, Oracle, Sybase, Informix, Postgres and SQL Server are used to perform various types of operations in a given database. These operations can be adding, deleting, updating, inserting and displaying information as stored in a database.

6.2 SQL

6.2.1 INTRODUCTION TO SQL

Structured Query Language (SQL) is a language used to create databases objects such as tables or views as well as to view or change data in the databases. It is an important database development and management tool, used by most data analysts and it is supported by various other DBMS products as well.

SQL was developed at IBM by the joint effort of Donald D. Chamberlin and Raymond F. Boyce during the early 1970s. Initially, it was called a SEQUEL(Structured English QUEry Language). SQL language was invented for creating, deleting, fetching and modifying rows for the data present in a tabulated format. SQL is an **ANSI** (American National Standards Institute) standard language although there are many more versions of this language such as:

- MySQL Database Server
- MS SQL Server using T-SQL,
- Oracle using PL/SQL,
- MS Access version of SQL is called JET SQL (native format) etc.

Some common relational database management systems that use SQL are Oracle, Sybase, Microsoft SQL Server, Access, Ingres, etc. Although most database systems use SQL, most of them also have their own additional proprietary extensions that are usually only used on their system.

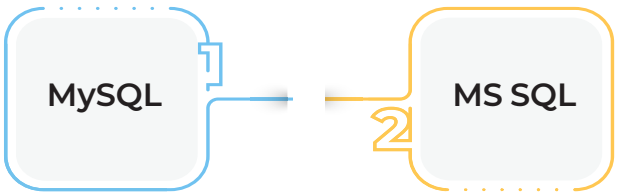
SQL is used to communicate with a database. According to ANSI (American National Standards Institute), it is the standard language for relational database management systems.

SQL statements or queries are used to perform tasks such as:

- Create, modify and delete database objects such as tables of views
- Update data on a database
- Retrieve data from a database
- Grant permissions and connect tables, etc.

6.2.2 INSTALLING THE SQL SERVER

We will learn to install 2 versions of SQL:



The SQL syntax for both are more or less the same.

Installing MySQL Server

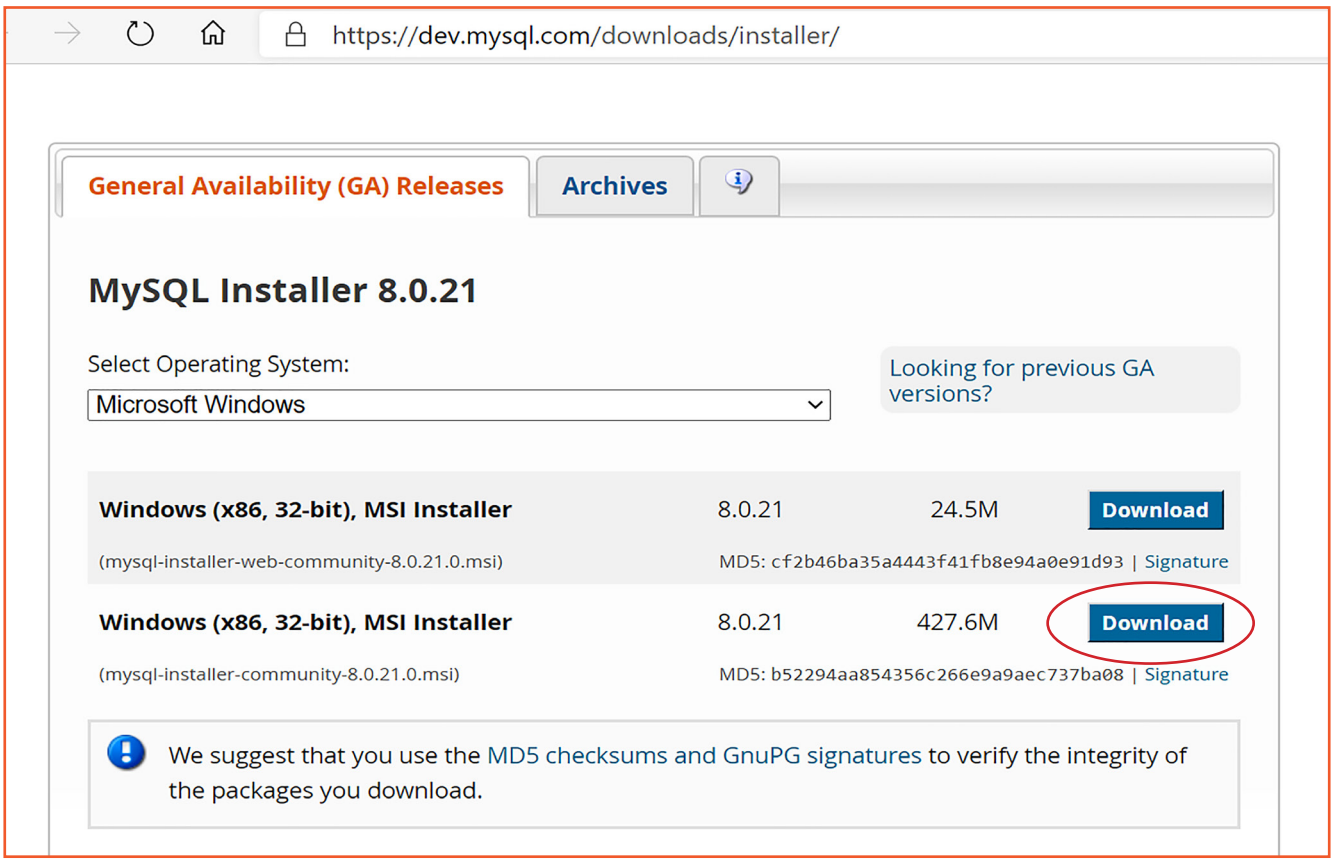
EXPLORE

MySQL is a well-established open-source relational database management system. To install a MySQL server, first download the installer from the following link:

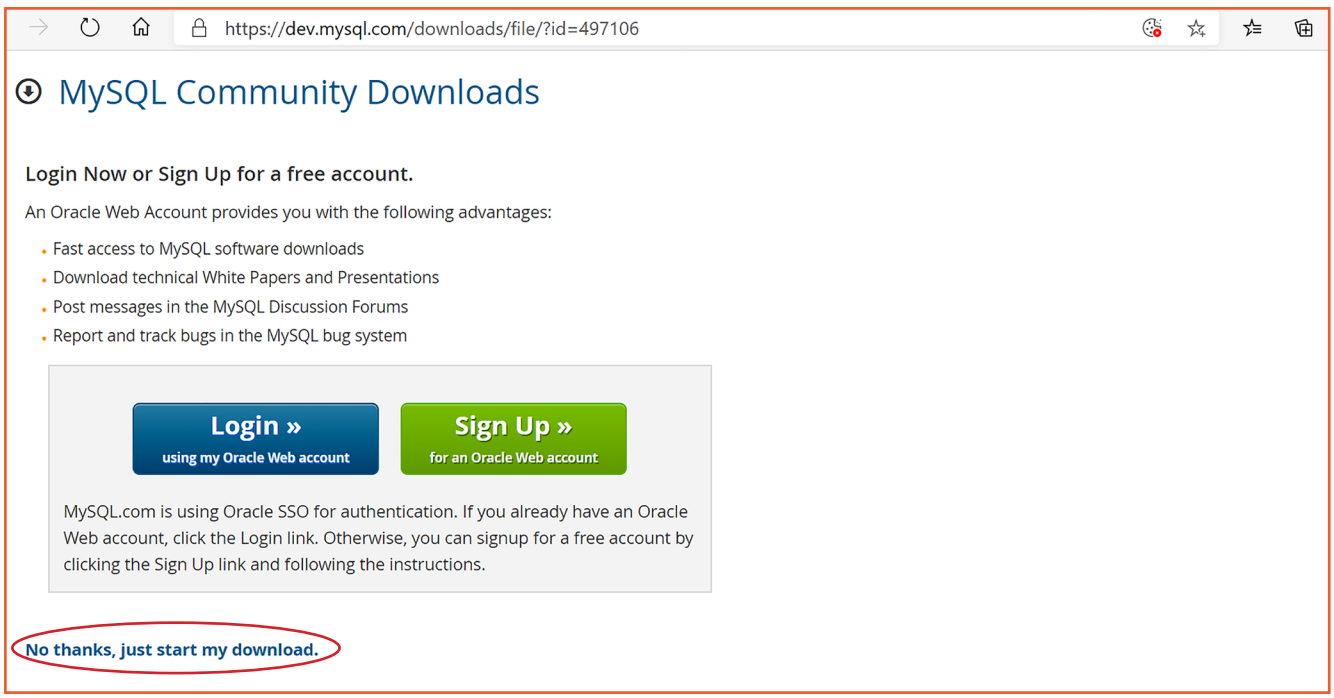
https://dev.mysql.com/downloads/installer/

You may select the web version of the full version.

- The Web version (A) contains the MySQL Installer and configuration files. The other MySQL files will have to be added later.
- The Full version (B) contains all MySQL Windows products, including the MySQL Server. This is preferable.



You will be asked to register and sign-in. You also have the option to skip this by clicking on “No Thanks, just start my download”



This will download the installer. After the download process is complete, you can execute the the MySQL installer file from its location. You can select the full developer package from the options it provides.

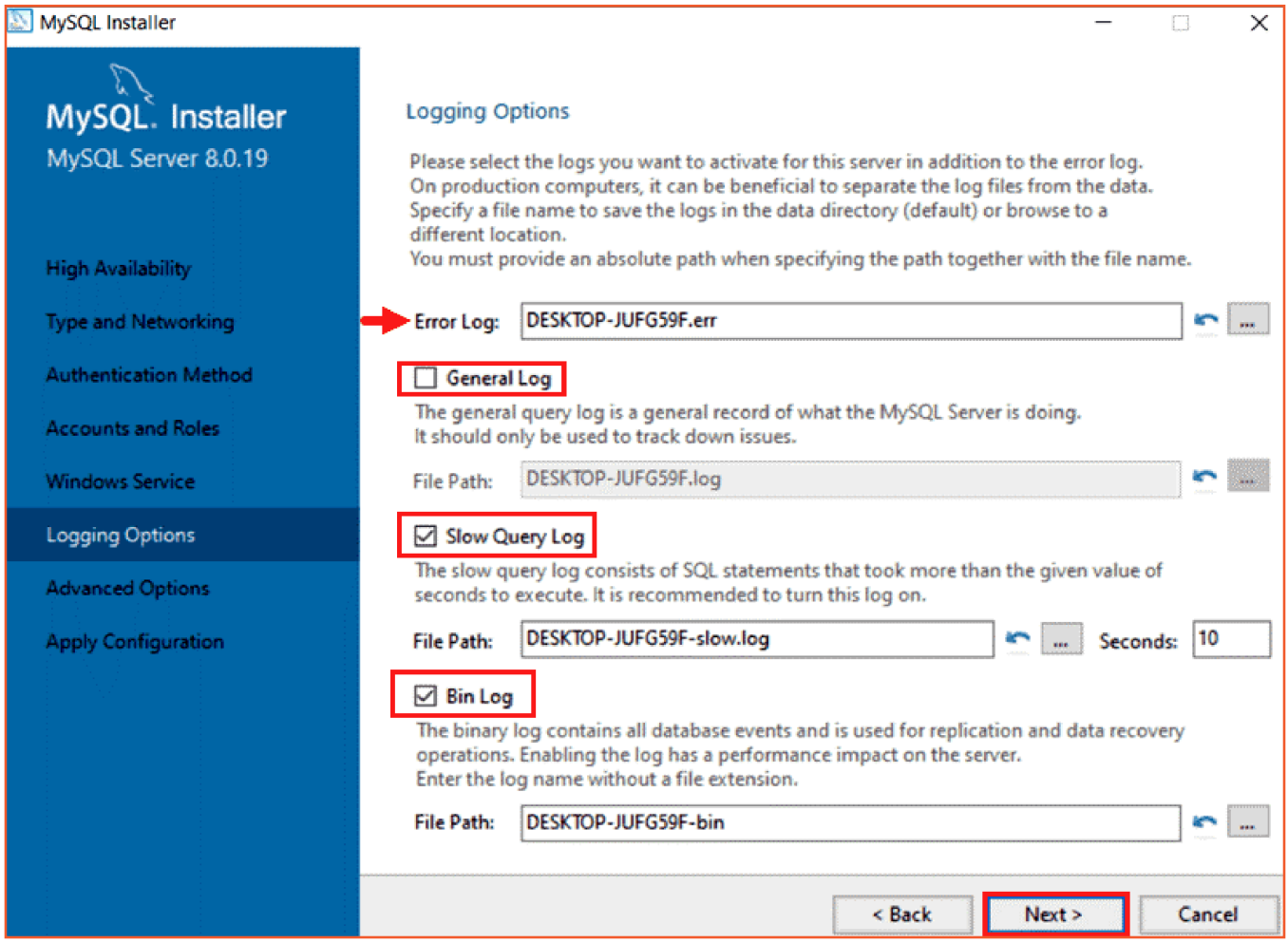
At this point, the system tries to resolve possible inconsistencies. It might inform you that additional packages need to be installed for the process to continue (e.g., Microsoft Visual C++ 2019 Redistributable Package). You can also run into Path installation inconsistencies if you have previous MySQL installations on your Windows Server.

Luckily the MySQL Installer auto-resolves issues and installs the latest binary compatible version of missing software. You are now ready to start the installation process in earnest. Click Execute to begin the installation process.

Once the status of the installation status is labeled as Complete, you are ready to configure the MySQL database.

You can do the following selections during the configuration:

- Selected the classic, single server option
- Select the Dedicated Computer
- Select the default setting is port number 3306 and can be changed to suit your needs
- Select “Use Strong Password Authentication” option
- You are now prompted to select a password for your MySQL root user.
- By selecting MySQL as a Windows Service, it can now start automatically whenever the Windows system boots.
- Make the following selections in the ‘Logging Options’



You have successfully configured the MySQL server and need to confirm for the MySQL Installer to apply the configuration.

Installing MS SQL Server

MS SQL requires a .Net Framework, 1GB of recommended memory and an NTFS system.

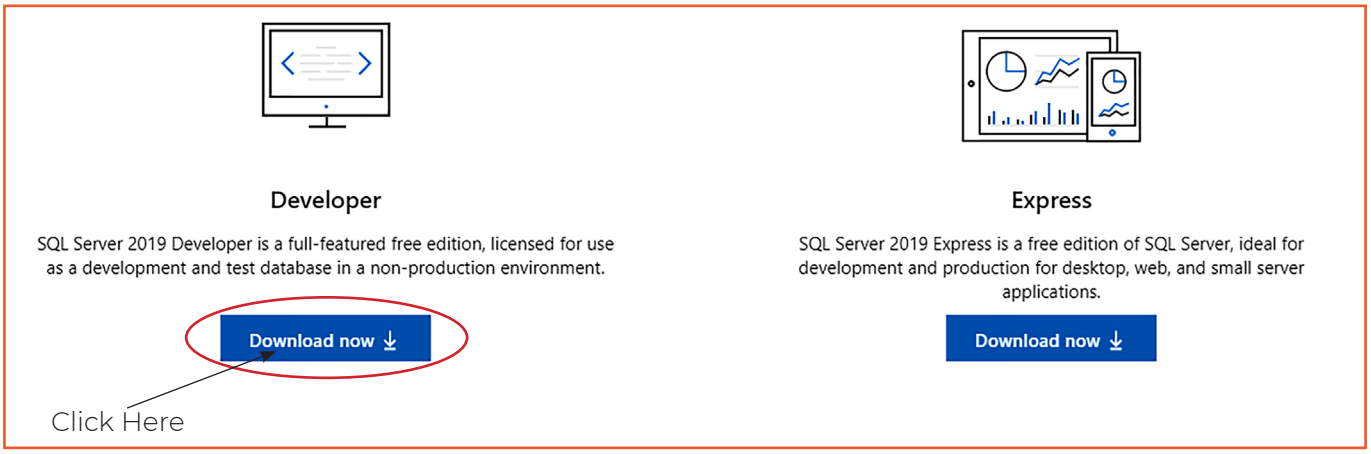
Downloading the SQL Server

EXPLORE

- Open the URL

https://www.microsoft.com/en-in/sql-server/sql-server-downloads

- Microsoft has provided two specialized free editions to work on MS SQL server.
- Select the Developer Edition.
- Click on the **"Download Now"** tab to start the downloading process.



- Save the **'SQLServer2017-SSEI-Dev.exe'** file in your desired location and double click to run the setup.
- Once you have double-clicked on the **'SQLServer2017-SSEI-Dev.exe'** setup, a screen will appear. The screen will contain three options, i.e. Basic, Custom and Download Media.

- Click on the **"Basic"** option. This option has the default configuration for learning MS SQL.
- Next, you will see **"Microsoft Server License Terms"** on the screen. Read it and click **"Accept"**.
- Server install location - By default, the file will be saved in the main drive, but you can change the installation location by clicking on 'Browse'
- Once you have selected the installation location, click on the 'Install' button.
- Select the **Connect now** option and a separate command line window will open to test the connection of the installed software. The system will run by default **'select @@Version'** statement to confirm that we can connect to new MSSQLSERVER instance successfully.
- Select **Install SSMS.** This option takes us to Microsoft SSMS download link. To run the SQL commands, the SQL Server should be connected with SQL Server Management Studio (SSMS).

SQL Server Management Studio (SSMS) is an integrated environment for managing any SQL infrastructure. SSMS platform provides various tools for configuring and monitoring SQL Server and databases. SSMS is used to query, design, and manage the databases and data warehouses. These can present on the local computer or in the cloud.

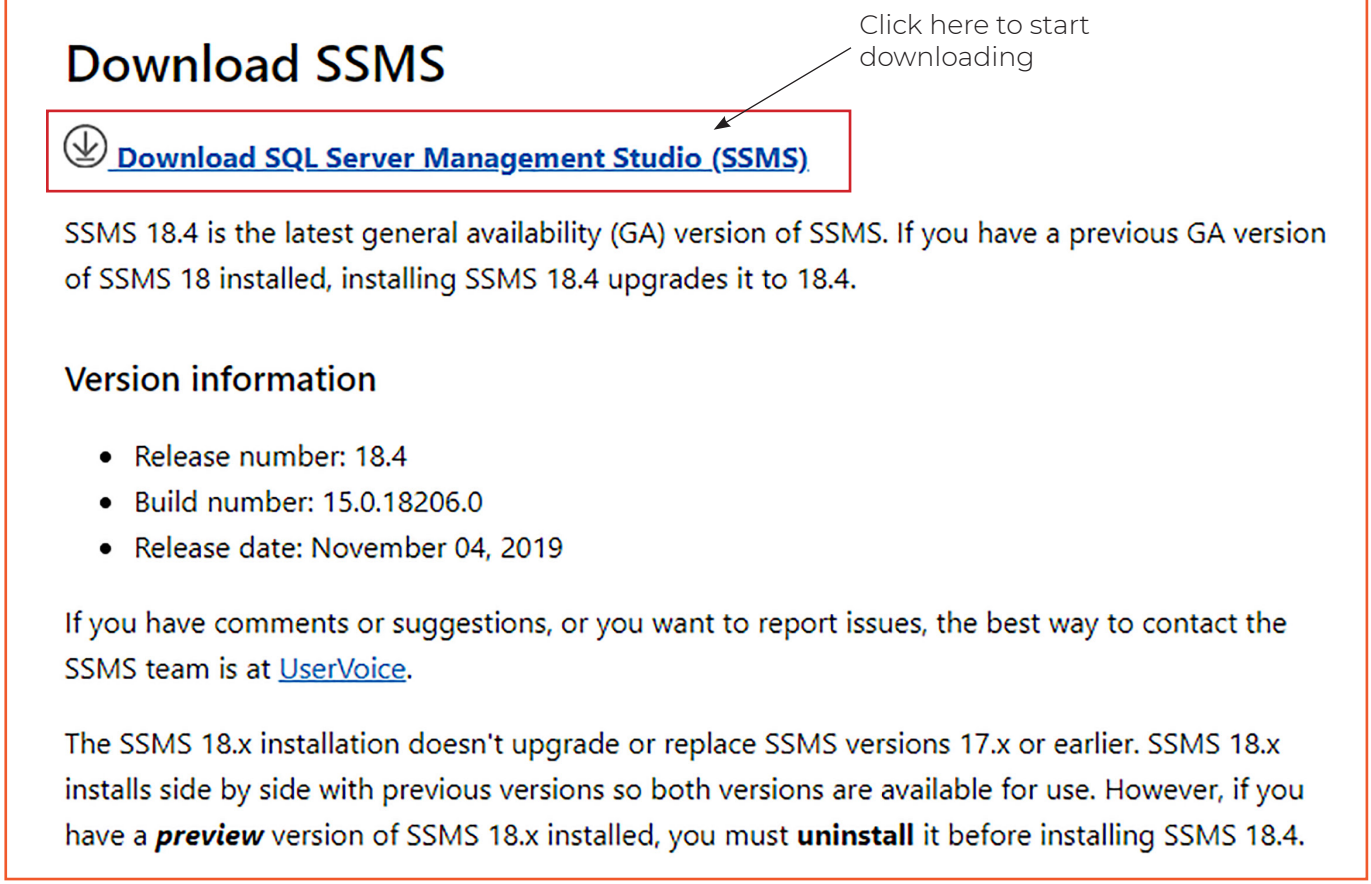
Steps involved in downloading SSMS:

EXPLORE

- Go to the website

https://docs.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver15

and click on **"Download SQL Server Management Studio (SSMS)"** to start the downloading process.



- Save the setup.exe file in the desired location. Run the ".exe" file and click on "Install" button as shown in.
- Once you have clicked on installation, the packages will start loading.

ACTIVITY 2 (Part 1)

Follow the steps of any one of the installation methods mentioned above and install the SQL Server of your choice into your computer.

Let us now read about the various types of operations that take place for manipulating data in the database using SQL commands.

6.2.4 BASIC SQL COMMANDS

Here are some basic commands used in SQL and their functions:

Data Definition Language (DDL)

Statements in DDL can be used to create, modify, and delete database objects such as databases, tables, columns, and constraints. The commonly used DDL statements are:

CREATE: Used to create database objects

ALTER: Used to modify database objects

DROP: Used to delete database objects

Data Control Language (DCL)

Statements in DCL can be used to grant or revoke permissions on database objects. The commonly used DCL statements are:

GRANT: Used to grant permissions for a user to use a database object or execute some SQL commands

REVOKE: Used to remove previously granted permissions

Data Manipulation Language (DML)

Statements in DML can be used to select, insert, update, and delete data in a database object. The commonly used DML statements are:

SELECT: Used to query information from a database

INSERT: Used to insert new rows of data into database tables

UPDATE: Used to modify data values in database tables

DELETE: Used to delete existing rows in database tables

you will learn the basics of each of these commands and practice some Queries using the SQL Interpreter.

Let's apply these in the next few sections.

6.2.5 DATABASE

Creating A Database

For creating a new database, you can use the CREATE DATABASE command. The syntax is given below:

Syntax:

```
CREATE DATABASE databasename;
```

Example:

```
CREATE DATABASE MLProject;
```

After you have entered this statement, a database of the name “MLProject” will be created.

Deleting A Database

For deleting an existing database in SQL (such as School in the given example), one can use the DROP command as given below:

Syntax:

```
DROP DATABASE databasename;
```

Example:

```
DROP DATABASE MLProject;
```

This SQL command will ensure that the database with the name MLProject will be deleted.

6.2.6 TABLE

As the name suggests, a database consists of data which is organised in the form of tables. Table, as we all know, consists of rows and columns. The data present in the tables can be called as data element or values. These values are organised in the form of horizontal rows and vertical columns. The point where the rows and columns intersect is called a cell. The table can consist of any number of rows but should have a specified number of columns.

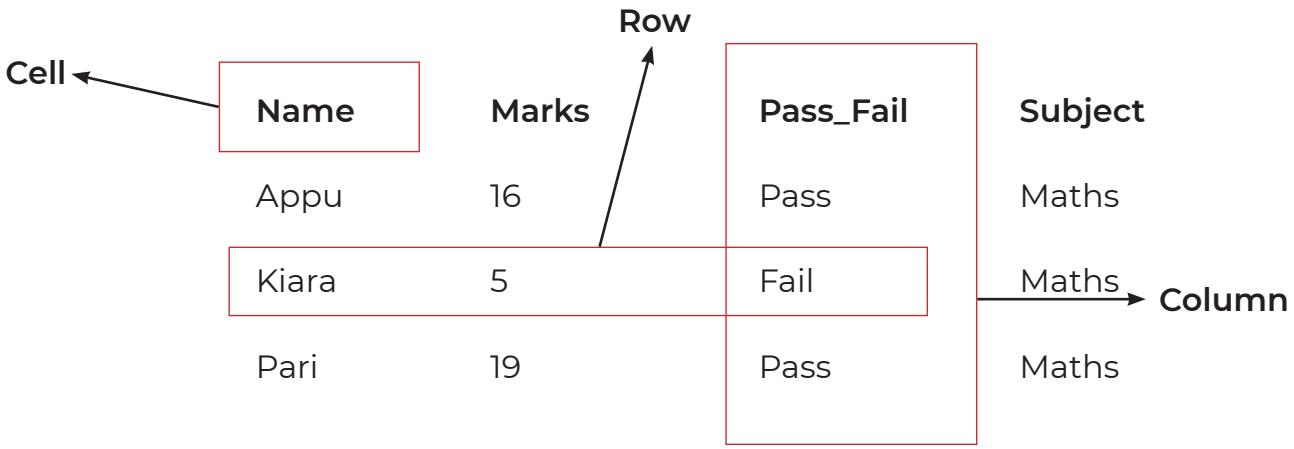


Table 2. A Standard Tabular Format

As can be seen in the given example, this table consists of three rows and four columns. The **columns** are the vertical components under the subject name as “Name”, “Marks”, “Pass_Fail” and “Subject”. The horizontal component is the **rows** that contain the names such as “Appu, Kiara and Pari”. Each entry in the table is present in a space called a **cell.** This table consists of twelve cells. All the entries present in the table (such as Name, Marks, Pass_Fail and Subject) are the **values** or simply **data elements.**

Creating a Table in SQL

For creating a table in SQL, we use CREATE TABLE command. One can understand this from the given syntax:

Syntax:

```
CREATE TABLE table_name (column1 datatype, column2 datatype, column3 datatype, ....);
```

Example:

Suppose we wish to create a table under the name “Month”. As an example, we have taken three columns as the datatypes for the table representing months. These can be:

num – to represent the number of the month being referred. This will be of the type integer.

name – to represent the name of the month. This will be a character.

days – to represent the number of days in a given month. This will also be an integer type.

Hence, the command would become:

```
CREATE TABLE months (num int, name varchar(8), days int);
```

In the given statement, the datatypes are int and varchar, which represent integer and character data, respectively. The data entries are represented by num, name and days, which will be the headings for the database.

By using the keyword CREATE TABLE, we are telling the database that we are going to create a new table. Then we need to enter the table name, which has to be unique. Note that SQL is case insensitive, but the data stored inside the table will be case sensitive. The columns for the table will be entered inside the open and close brackets. Each column has to be specified with a certain data type.

ACTIVITY 2 (Part 2)

Follow the steps mentioned above and create the database.

Create a table named “Class_Test_August” for the data provided in the table given above.

6.2.7 INSERTING ROWS IN A TABLE IN SQL

Now that we have chosen the headings for the database, there is a need to enter the values or data elements. For entering the information in the table, we use the INSERT INTO keyword. The syntax to use this statement is as follows:

Syntax:

```
INSERT INTO table_name (column1, column2, column3, ...) VALUES (value1, value2, value3, ...);
```

Example:

Let us use this syntax in the table that we formed under the heading **"months".**

```
INSERT INTO months (num, name, days) VALUES (8,'August',31);
```

If you want to add more entries in the table, simply enter the details of the data element you wish to put in the database. The changes will automatically get reflected in the existing table.

You can write an INSERT statement to add more months in the given table.

The table data will look as follows:

num	name	days
1	January	31
2	February	28
3	March	31
4	April	30
5	May	31
6	June	30
7	July	31
8	August	31
9	September	30
10	October	31
11	November	30
12	December	31

ACTIVITY 2 (Part 3)

Follow the steps mentioned above, enter the rows as given in the table with the data mentioned in the following table:

Num	Name	Days
Appu	16	Pass
Kiara	5	Fail
Pari	19	Pass
Samar	10	Pass
Sohini	6	Fail
Biju	18	Pass
Kevin	14	Fail

6.2.8 TRUNCATING A TABLE IN SQL

Using SQL, it is possible to delete the data present in a particular field rather than deleting the entire table. You must be wondering what keyword is used to perform such an operation. This is called truncating a query in SQL. This operation can be performed by using the TRUNCATE keyword. The syntax depicting the use of the keyword is given below:

Syntax:

```
TRUNCATE TABLE table_name;
```

Example:

If we wish to delete an entry under the table name “months”, we should use the following statement:

```
TRUNCATE TABLE months;
```

Once this statement is executed, the data inside the table “months” will be deleted, but the table will still remain.

6.2.9 DELETING ROWS IN SQL

When we want to delete a particular row in a given table, we use DELETE keyword. Let us understand the usage of this keyword for deleting a particular record or entry.

Syntax:

```
DELETE FROM table_name
WHERE condition;
```

Example:

If we wish to delete an entry under the table name “months”, we should use the following statement:

```
DELETE FROM months
WHERE days = 31;
```

After the execution, the given code, the entries where the days of the are 31 will be deleted. Also, the table “months” will be modified as follows:

num	name	days
2	February	28
4	April	30
6	June	30
9	September	30
11	November	30

ACTIVITY 2 (Part 4)

From the table Class_Test_August, delete the entries that are less than 12.

6.2.10 QUERIES

SELECTING DATA FROM A TABLE IN SQL

If we want to fetch data from a database, we use the SELECT keyword. This statement is commonly used in for performing various operations using SQL. Let us now understand how this keyword is used for handling databases.

Syntax:

```
SELECT column 1, column 2.....column N
FROM table_name;
```

Example:

If you want to select the data present in the columns “name” and “days” in the table name “months”, you should enter the following statements:

```
SELECT name, days
FROM months;
```

The above example will display all the entries, i.e. months and days as present under the table name “months”. In case we wish to display all the entries present in a given table, we use * (Star) operator.

```
SELECT *
FROM months;
```

If we wish to display certain fields without any duplication, then we would use the DISTINCT keyword along with SELECT.

```
SELECT DISTINCT name
FROM months;
```

The above entry will display all the data entries under the column “name” without any duplication. Every data entry will be unique.

Using 'WHERE' COMMAND for DATA SELECTION in SQL

There can be a situation where we need to select certain records from a given table. The WHERE keyword acts as a filtering mechanism for the data present in a certain table. For extracting the data, there is a need to follow certain conditions. Only if those conditions are met, the data will be extracted.

Syntax:

```
SELECT column1, column2, ...column N
FROM table_name
WHERE condition;
```

Example:

```
SELECT name, days
FROM months
WHERE days = 30;
```

name	days
April	30
June	30
September	30
November	30

Once these statements are executed, all the names of the months that have 30 days will be displayed. This is because we had selected the entries under the category "Name" and "Age" and applied a condition that the age of the people should be 18 years. As per the example, the names Shashi, Tarun and Akhil will be displayed along with their age.

Adding MULTIPLE CONDITIONS in SQL

In case you wish to include multiple conditions in the where clause, we use AND, OR, NOT operators in SQL.

AND Operator:

This operator displays the record if all the conditions that have been entered and separated by AND are true.

Syntax:

```
SELECT column1, column2, ...column N
FROM table_name
WHERE condition1 AND condition2 AND condition3...;
```

Example:

```
SELECT *
FROM months
WHERE days = '30' AND name = 'April';
```

num	name	days
4	April	30

OR Operator:

This operator displays a record if even one of the conditions that have been entered are true. The conditions will be separated by an OR keyword.

Syntax:

```
SELECT column1, column2, ...column N
FROM table_name
WHERE condition1 OR condition2 OR condition3...;
```

Example:

```
SELECT *
FROM People
WHERE days = '28' OR num = '10';
```

num	name	days
2	February	28
10	October	31

NOT Operator:

NOT operator displays a record if the condition/conditions mentioned in the code are not true. Let us understand the function of a NOT operator through an example.

Syntax:

```
SELECT column1, column2, ...column N
FROM table_name
WHERE NOT condition;
```

Example:

```
SELECT *
FROM People
WHERE NOT days = '31';
```

num	name	days
2	February	28
4	April	30
6	June	30
9	September	30
11	November	30

6.2.11 AGGREGATE FUNCTIONS IN SQL

An **aggregate function** is the one wherein the values of multiple rows are grouped as input based on a specific criterion which returns a single value. Some of the commonly used aggregate functions in SQL are COUNT, SUM, AVG, MIN, MAX. Let us discuss each of these aggregate functions for understanding their usage in practical applications.

COUNT ( ) Function:

COUNT function displays the number of rows that fulfil the specified criteria.

Syntax:

```
SELECT COUNT (column_name)
FROM table_name
```

Example:

```
SELECT COUNT (days)
FROM months
```

The above statements will display the following result as output:

COUNT(days)

12

The given code asks the compiler to count all the records under the column “days” and display the number of records present as the output.

AVG ( ) Function:

The AVG function displays the average of the data entries under a particular numeric column.

Syntax:

```
SELECT AVG(column_name)
FROM table_name
```

Example:

```
SELECT AVG(days)
FROM months
```

The given code will display the average of the entries present under the column “days”. Hence, the output would be:

AVG(Age)

30.41

SUM ( ) Function:

The SUM function will display the sum of all the data entries present under a particular numeric column.

Syntax:

```
SELECT SUM(column_name)
FROM table_name
```

Example:

```
SELECT SUM(days)
FROM People
```

The code will add up the age of all the seven records present in the table. Therefore, the output will be:

SUM(Age)

365

MIN ( ) Function:

MIN function returns the smallest value as present in a numeric column.

Syntax:

```
SELECT MIN(column_name)
FROM table_name
```

Example:

```
SELECT MIN(num)
FROM months
```

The code will display the minimum age among all the records as present in the column "num". In this case, the output will be:

MIN(num)

1

Note: If we want to modify the heading in the output, we use the "AS" function. For example, if we want to change the heading "MIN(days)" to "Smallest_Month", we will execute the following code:

```
SELECT MIN(days) AS Smallest_Month
FROM People
```

The previous output will be modified and displayed as:

Smallest_Month

16

MAX ( ) Function:

MAX function will return the largest value as present in a particular numeric column.

Syntax:

```
SELECT MAX(column_name)
FROM table_name
```

Example:

```
SELECT MAX(num)
FROM months
```

The output will be:

MAX(num)
12

6.2.12 UPDATING RECORDS IN SQL

Sometimes, there is a need to modify the data present in certain rows. This operation is performed using the UPDATE keyword. The steps involved in using the UPDATE keyword are:

1. Choose the table where the record needs to be modified
2. Set new values for the desired column(s)
3. Select the rows using WHERE keyword where the modification needs to be done

Code	Country	Capital
1	India	New Delhi
2	Sri Lanka	Colombo
3	Nepal	Kathmandu
4	Pakistan	Islamabad
5	Bangladesh	Dddhaka
6	Bhutan	Thimphu

Database “Capital Cities”, Table “Asia”

6.2.13 BETWEEN AND IN OPERATORS IN SQL

BETWEEN Condition:

The SQL BETWEEN condition allows the user to test if an expression is within a range of values (inclusive). These values can be in the form of text, date, or numbers. Also, it can be used in statements containing SELECT, INSERT, UPDATE, or DELETE keywords.

Syntax:

```
SELECT column_name(s)
FROM table_name
WHERE column_name BETWEEN value1 AND value2;
```

ID	SName	Maths	Physics	OverallPercentage
342	Appu	76	54	65
271	Pari	84	90	88
562	Kiara	56	82	74
875	Sohini	92	94	90
198	Biju	89	60	81
974	Samar	78	54	63

Table 13. Database "Result"

Consider the table given. Suppose if we want to display the name of the students who scored marks between 80-95 in Mathematics, then we would use the BETWEEN keyword and the code will be written as:

```
SELECT SName
FROM Result
WHERE Maths
BETWEEN 80 AND 95;
```

The output after the code is executed would be:

SName
Pari
Kiara
Sohini

Similarly, if we want to display the ID and names of the students who scored between 80-95 in Physics, then we would write the following code:

```
SELECT ID, SName
FROM Result
WHERE Physics
BETWEEN 80 AND 95;
```

ID	SName
271	Pari
562	Kiara
875	Sohini

IN Condition:

The IN operator is useful in testing if the expression matches any value in the existing list of values. By using IN operator, there is no need of using multiple OR condition in SELECT, INSERT, UPDATE or DELETE.

Syntax:

```
SELECT column_name(s)
FROM table_name
WHERE column_name IN (list_of_values);
```

Example:

Refer to the table "Result" given below. Write an SQL code to display the ID and the name of students who have scored 75 marks in Mathematics.

ID	SName	Maths	Physics	OverallPercentage
342	Appu	75	54	65
271	Pari	80	90	88
562	Kiara	75	82	74
875	Sohini	75	90	90
198	Biju	90	60	81
974	Samar	84	90	63

Table 14. Database "Result"

```
SELECT ID, SName
FROM Result
WHERE Maths IN (75);
```

Output:

ID	SName
342	Appu
562	Kiara
875	Shi

Similarly, if we want to display the ID and the name of students who have scored 90 marks in Physics, we can write the following SQL code:

```
SELECT ID, SName
FROM Result
WHERE Physics IN (90);
```

Output:

ID	SName
271	Harsh
875	Shalini
974	Mehak

Similarly, if we want to display the ID and the name of students who have scored 90 marks in Physics, we can write the following SQL code:

ACTIVITY 2 (Part 5)

1. Create a table by the name 'PLAYERS'
2. Add the following entry in the given table:

ID	Name	Matches	Runs	Wickets
3	Kevin	12	120	7

3. Add more entries for Pari, Sohini, Samar, Kiara, Appu and Biju:
4. Display the records who have played a specified number of matches and taken a specified number of wickets.
5. Display the names of the players who have scored maximum runs, played maximum matches and taken maximum wickets. Also, write a code to display the players who have scored minimum runs, played minimum matches and taken minimum wickets.
6. Display the average of the total runs scored.
7. Replace the entry in row number 5 (ID 5) with the given data:

ID	Name	Matches	Runs	Wickets
5	Goldie	14	198	10

8. Display the records who have scored runs between 110-150.
9. Display the records who have played more than 10 matches.

SUMMARY

Databases offer a convenient and powerful way to store and organise information.

A Spreadsheet is a computer program that looks like a large sheet of paper with boxes. It has a grid of cells that form rows and columns. Usually, the rows in a table are called records and the columns in a table are called fields or attributes.

Data sorting is the process of arranging the data into a meaningful order to enable the viewer to analyse the information more effectively. It can be done in ascending and descending order. The computer sorts the elements of a list into the desired order using a Sorting Algorithm.

The filter feature applies a dropdown menu to each column heading, allowing you to select specific choices to narrow a table.

Excel formulas all start with an = sign written in the empty cell, followed by the function name (In this case 'sum') and a section in parentheses to provide ranges or arguments.

A database with only one table is called a flat database. A database that contains two or more related tables is called a relational database. Some important concepts of relational databases are:

- The primary key is a field whose values are unique in this table, and so can be used as identifiers for the records.
- When two tables have an unequal relationship, we call the independent table the parent and the dependent table the child.
- Cardinality describes how many records in one table can be related to records in another table. Two tables might have a cardinality of 1-1, 1- oo, 1-3, oo – oo

Specialized database development and management tools are used to perform various types of operations in a given database. These operations can be adding, deleting, updating, inserting and displaying information as stored in a database.

Structured Query Language (SQL) is a language used to create databases objects such as tables or views as well as to view or change data in the databases. It is an important database development and management tool.

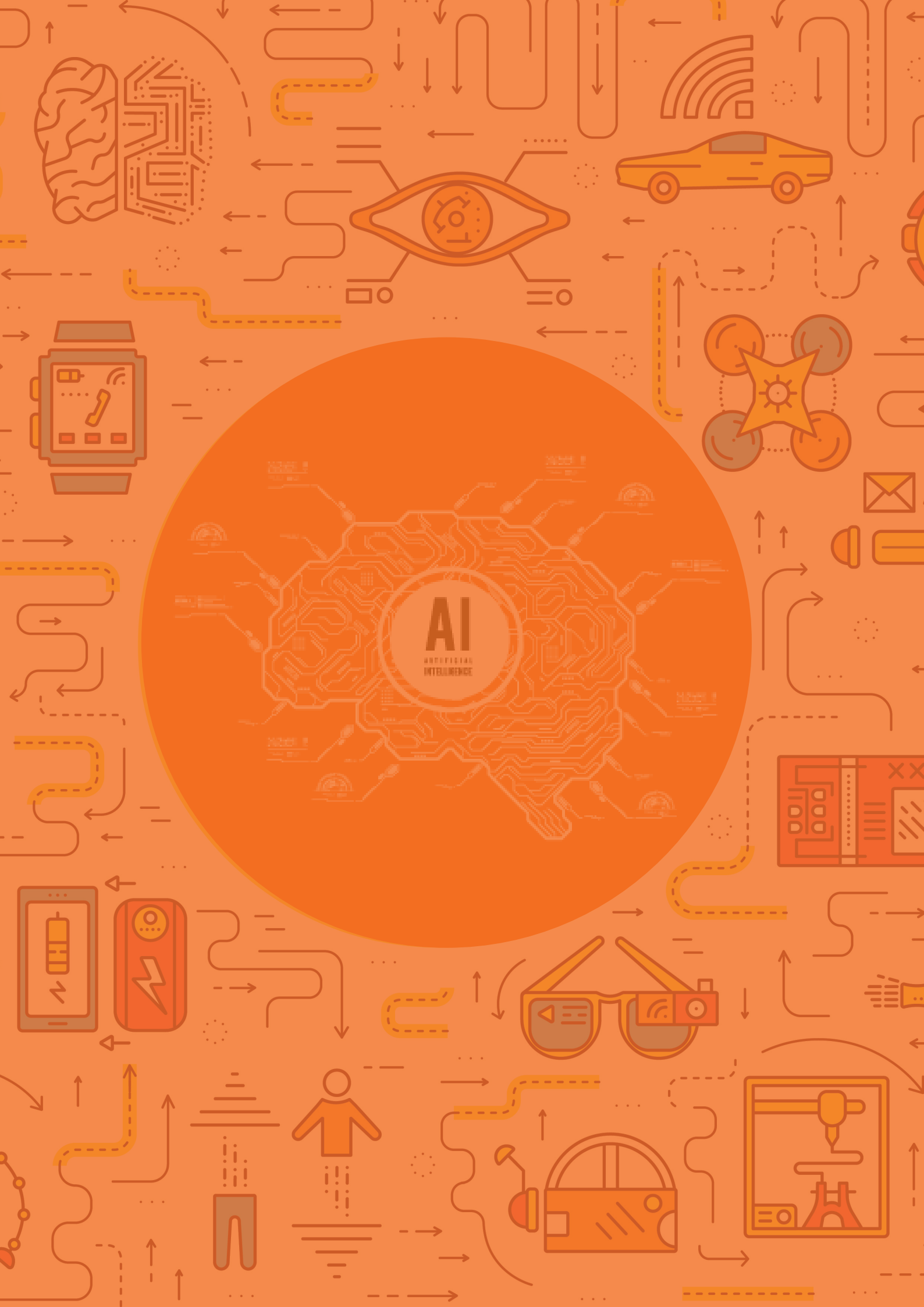
SQL statements or queries are used to perform tasks such as:

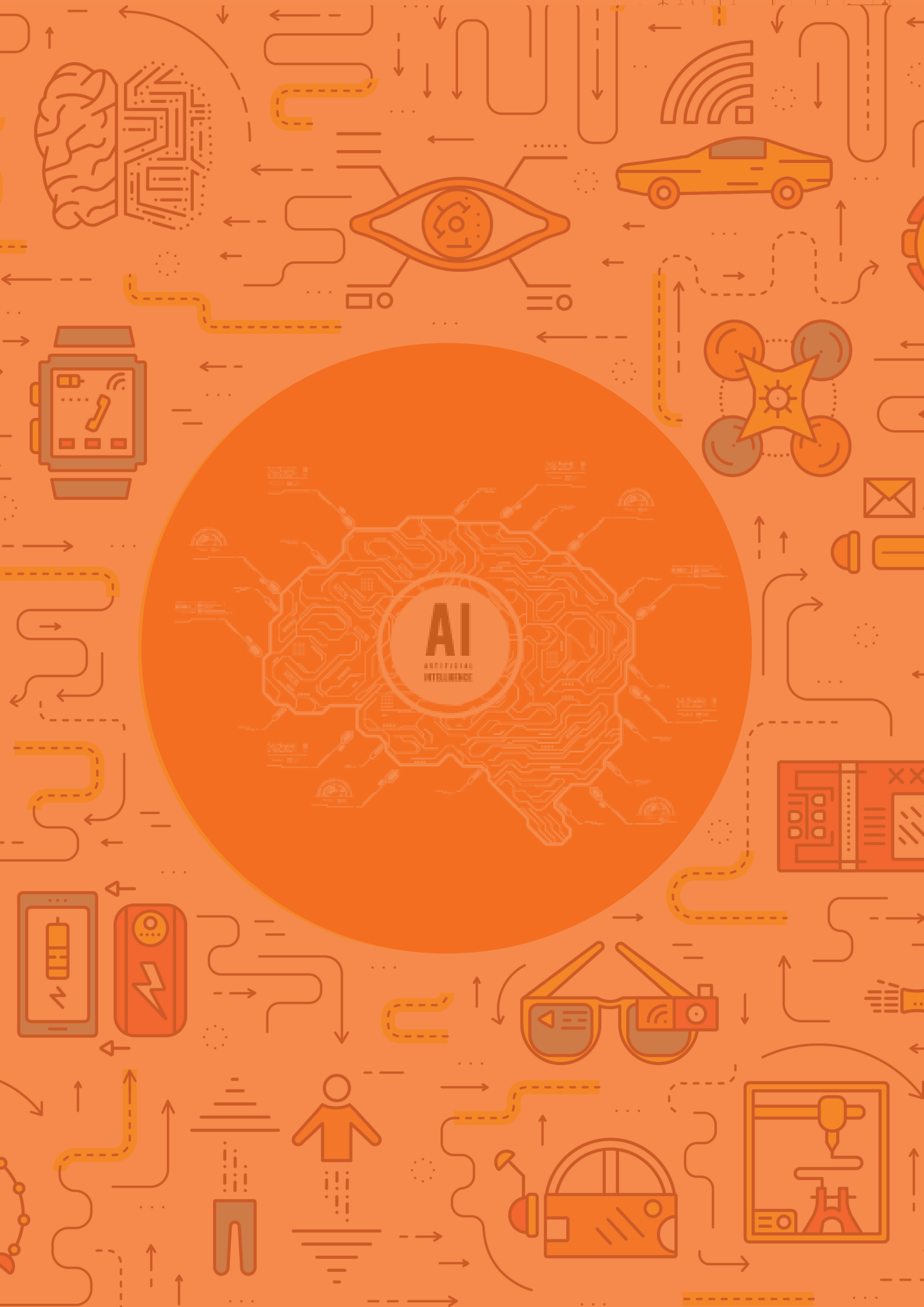
- Create, modify and delete database objects such as tables of views
- Update data on a database

- Retrieve data from a database
- Grant permissions and connect tables, etc.

Some basic commands used in SQL are:

- For Data Definition Language (DDL): CREATE, ALTER, DROP
- For Data Control Language (DCL): GRANT, REVOKE
- For Data Manipulation Language (DML): SELECT, INSERT, UPDATE, DELETE


AI
ARTIFICIAL
INTELLIGENCE


AI
ARTIFICIAL
INTELLIGENCE

UNIT 7
PYTHON

Key Learning Outcomes

At the end of this module, you will be able to:

a) Download an Integrated development environment (IDE)
b) State core programming concepts
c) Apply essential data science libraries.
d) Write clean, logical code
e) Apply python codes to simple machine learning projects

Topic Covered

Introduction To Python | Integrated Development Environment (IDE) | Core Programming Concepts | Essential Data Science Libraries | Troublesh Ooting Defects In Python Code | Practice For Writing Code

7.1 INTRODUCTION TO PYTHON

As we have learnt in the base module, Python is one of the various languages used to communicate with computers.

Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991. It is mainly used for Data Science since Python provides easy and simple tools for data mining and data analysis.

Because of it's precise and efficient syntax, Python is fast becoming one of the most favoured languages used for Artificial Intelligence and Machine Learning. Python also has various packages called 'libraries' for data analysis and machine learning, which drastically reduce the time it takes to produce results.

A python program is a sequence of Python statements (code) written to perform some computational task. Like the word document has an extension of .doc or .docx, the python scripts are stored in files and these files are named with the extension .py. For example, hello.py.

While the world of Python is quite vast, to start one does not have to master everything and learn all the syntax. You can start with the following:

- Download an Integrated development environment (IDE)
- Learn the core programming concepts
- Gaining a working knowledge of essential data science libraries
- Practicing to write clean, logical code
- Practicing with actual projects

With this approach, you can build mastery over time, as well as enjoy the process.

7.2 INTEGRATED DEVELOPMENT ENVIRONMENT (IDE)

7.2.1 Need for IDEs

An integrated development environment is a software application that provides comprehensive facilities to computer programmers for software development. It helps programmers by providing tools for various common activities of writing software in a single application. These could be for editing source code, highlighting syntax, building executables files or debugging.

- **Editing Source Code:** Writing code is an important part of programming. We start with a blank file, write a few lines of code, and a program is born! IDEs facilitate this process with features like syntax highlighting and autocomplete.
- **Syntax Highlighting:** An IDE that knows the syntax of your language can provide visual cues. Keywords, words that have special meaning or errors are highlighted with different colors.
- **Compiler:** Compilers are components that translate programming language into a form machines can process, such as binary code. The machine code is analyzed to ensure its accuracy. The compiler then parses and optimizes the code to optimize performance.
- **Debugging:** No programmer can avoid writing bugs and programs with errors. When a program does not run correctly, IDEs provide debugging tools that allow programmers to examine different variables and inspect their code in a deliberate way. IDEs also provide hints while coding to prevent errors before compilation.

It is much easier working with an IDE rather than downloading separate applications for each software development activity.

There are many IDE available on the internet for different types of software. Out of the various IDEs that support Python Language, we need to use the ones where to write analytics and apply data science. Some such IDEs that are commonly used for AI and Machine learning are:

Out of these, we shall help you work with Jupyter, but you can explore the others as well.

7.2.2 Downloading Jupyter Notebook

What is Jupyter notebook?

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

What is Jupyter notebook used for?

It is used for data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Who uses Jupyter notebook?

Jupyter is mostly used by Data Scientists since it can easily be used for analyzing, processing, and modeling data and then interpret the results. Data scientists are big data wranglers, gathering and analyzing large sets of structured and unstructured data. A data scientist's role combines computer science, statistics, and mathematics.

How do I set up Jupyter Notebook?

There are many ways to set up a Jupyter Notebook. Here we will discuss the setup of Jupyter Notebook using the Python Distributer, Anaconda.

The first step to install it is to install Python through Anaconda.

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS.

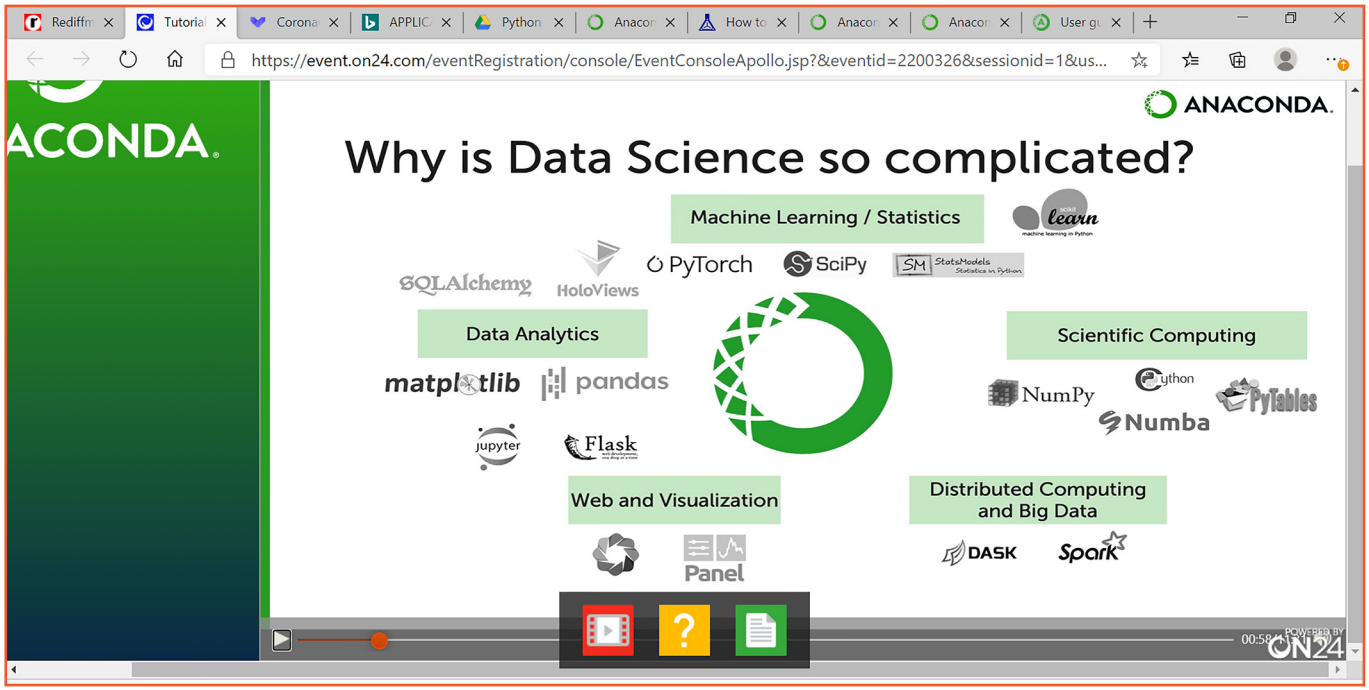
The Anaconda bundle comes with the libraries you'll need for data science.

Step 1: The open-source version can be downloaded from the following site:

https://www.anaconda.com/products/individual

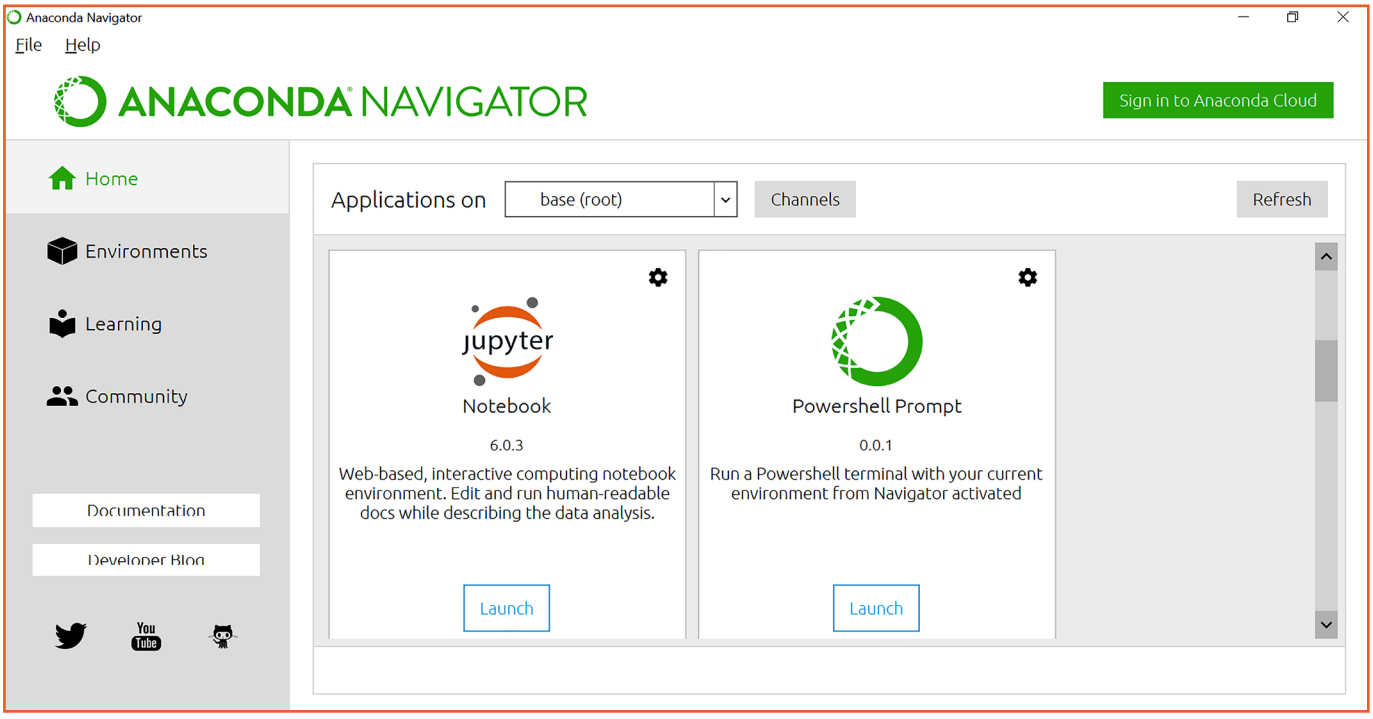
Anaconda Individual Edition is the industry standard for data scientists developing, testing and training on a single machine.

Step 2: After installing it. Go through the brief introductory webinar to understand the kind of packages that Anaconda offers and how to install them.



Step 3: Now open the application 'Anaconda Navigator' from your computer.

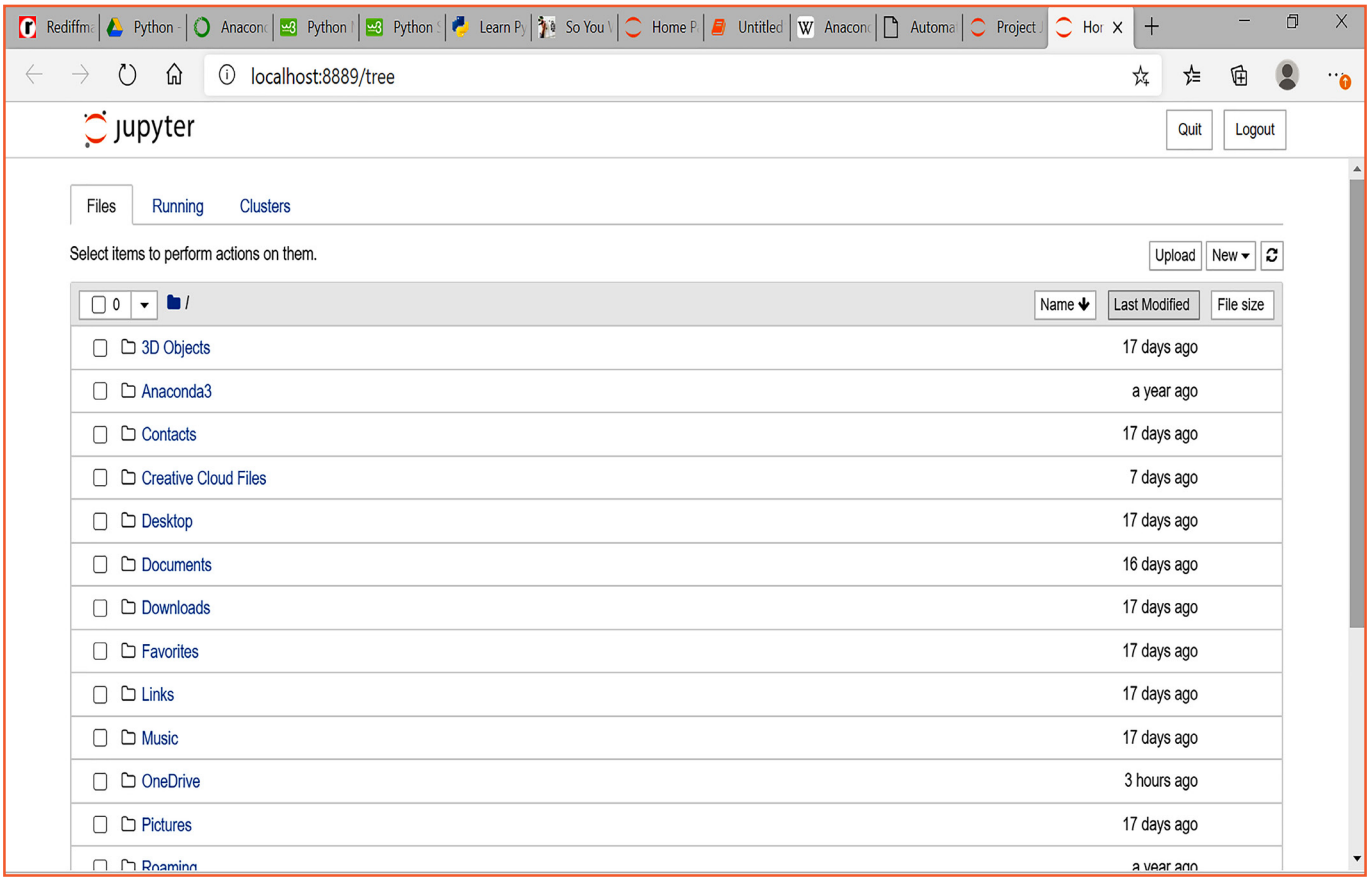
You will see all the packages included in Anaconda on the home screen, one of which will be Jupyter Notebook.



Step 4: Launch Jupyter Notebook.

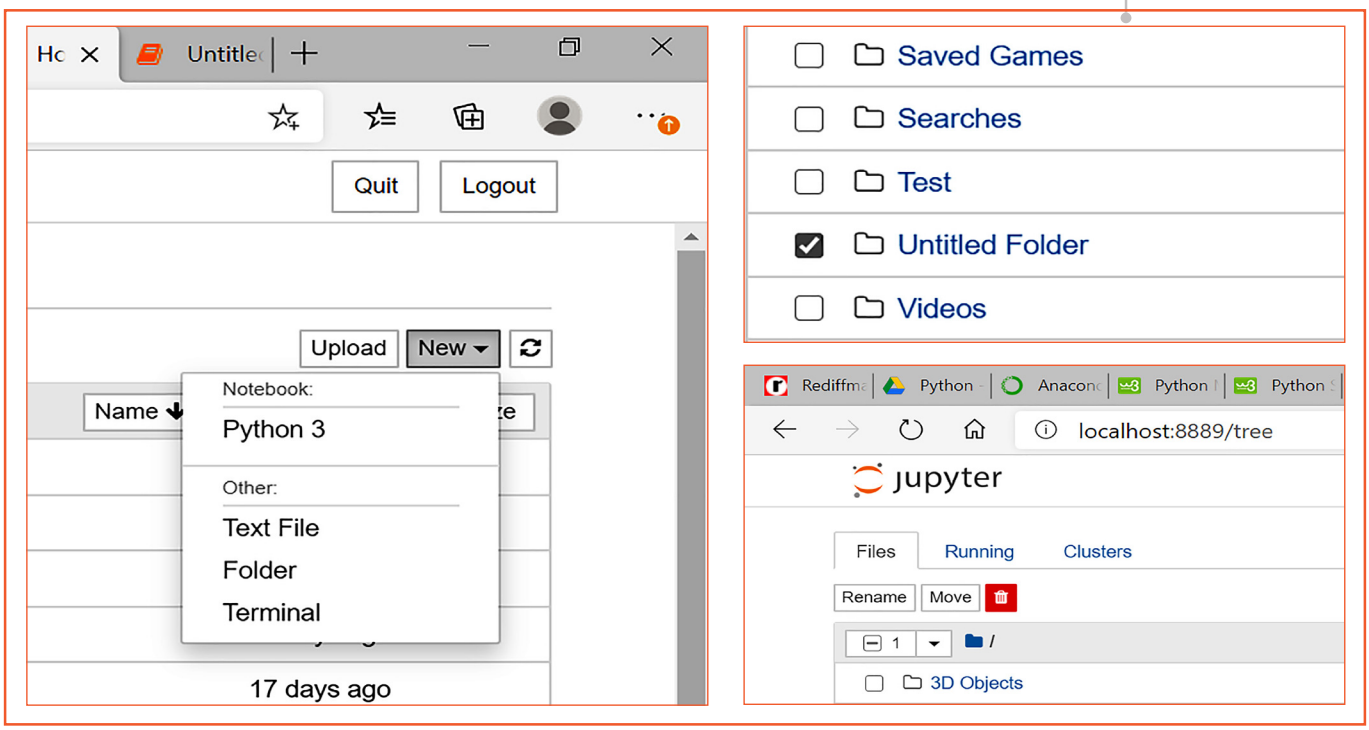
Step 5: After completing the instalment procedure, open the Jupyter Notebook application. It will open in the browser. The preferred browsers are Chrome, Safari and firefox. It may also open in Opera and Edge; however, if it doesn't, one should use Chrome, Safari and Firefox.

You will see a similar-looking screen to the one given below:

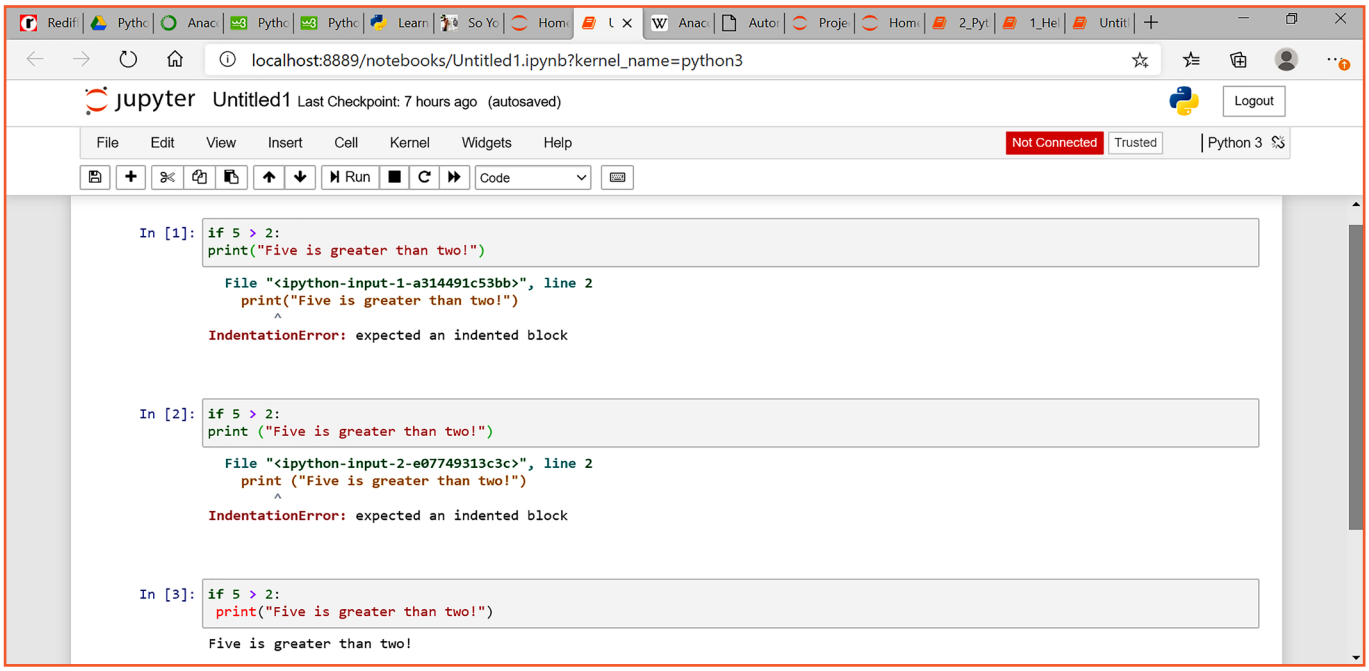


Step 6: To write and run code, click on the 'New' button and then select Python 3. These files can be saved and named. All the files created by you in Jupyter will have the extension '.ipynb'.

You can create a Folder too, by selecting Folder, where you can save all your program files. The folders can be renamed by selecting it and then clicking the rename button on the top.



Step 7: After selecting New → Python 3, the following screen will open, wherein you can write the code and click on 'Run' to view the output. If there is any error, it will show the error so that you can rectify it.



To learn more, you can go through the User Guide provided.

https://docs.anaconda.com/anaconda/user-guide/getting-started/

Installing Jupyter Notebook using Amazon Sagemaker

An Amazon SageMaker notebook instance is a fully managed ML compute instance running the Jupyter Notebook App. Use Jupyter notebooks in your notebook instance to prepare and process data, write code to train models, deploy models to Amazon SageMaker hosting, and test or validate your models.

A best practice when using an Amazon SageMaker notebook is to use the notebook instance to orchestrate other AWS services.

Below steps ask for necessary information required to create SageMaker notebook instance. For other optional fields that are not covered in below steps, SageMaker will choose the default values.

To create an Amazon SageMaker notebook instance:

EXPLORE

Step 1: Open the Amazon SageMaker console at:

https://console.aws.amazon.com/sagemaker/

Step 2: Choose Notebook instances, then choose Create notebook instance.

Step 3: On the Create notebook instance page, provide the following information:

Step 4: For Notebook instance name, type a name for your notebook instance.

- **a.** For Notebook instance type, choose an instance type for your notebook instance.
- **b.** For IAM role, ask your faculty/instructor for the role to choose from an existing IAM role in your account that has the necessary permissions to access Amazon SageMaker resources.
- **c.** For Root access, to enable root access for all notebook instance users, choose Enable. To disable root access for users, choose Disable. If you enable root access, all notebook instance users have administrator privileges and can access and edit all files on it.
- **d.** Choose Create notebook instance.

In a few minutes, Amazon SageMaker launches an ML compute instance (a notebook instance) and attaches an ML storage volume to it. The notebook instance has a preconfigured Jupyter notebook server and a set of Anaconda libraries.

When the status of the notebook instance is **InService,** in the console, the notebook instance is ready to use.

- Choose **Open Jupyter** next to the notebook name to open the classic Jupyter dashboard.
- You can choose **Open JupyterLab** to open the JupyterLab dashboard.

Installing Jupyter notebook using PIP

For this, you'll have to make sure that Python and pip is already setup in your PC before we start the setup of the Jupyter Notebook. PIP is a package manager for Python packages or modules.

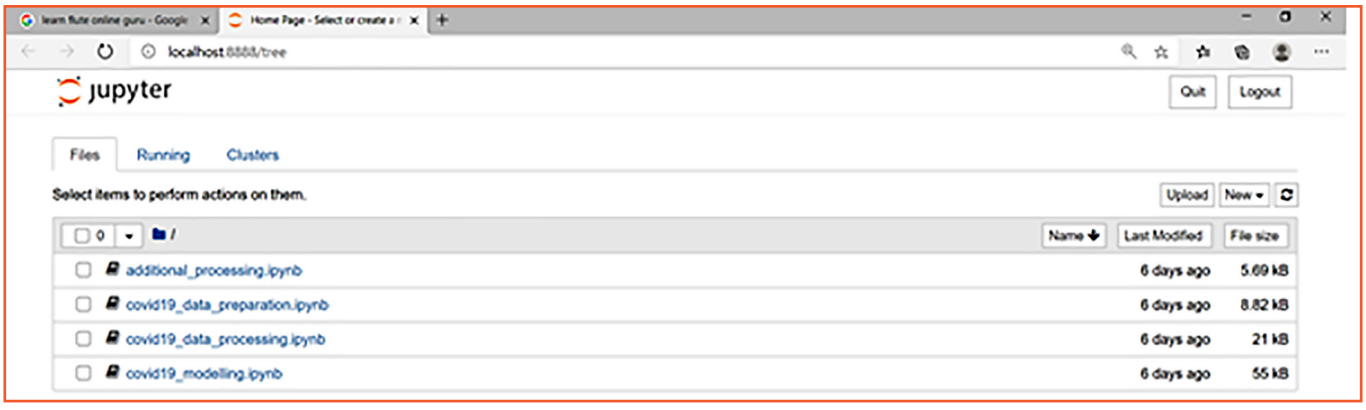
The Steps:

Step 1: Install Jupyter notebook with the command "pip install Jupyter notebook" in your command prompt

Once it is installed successfully, it can be launched using the command' jupyter notebook' from your folder of choice where you want to store your work

```
C:\user\myprojects\btcat>jupyter notebook
```

Multiple notebooks can be created and used. Here is a screenshot of jupyter with 4 BTCAT notebook (. ipynb) files



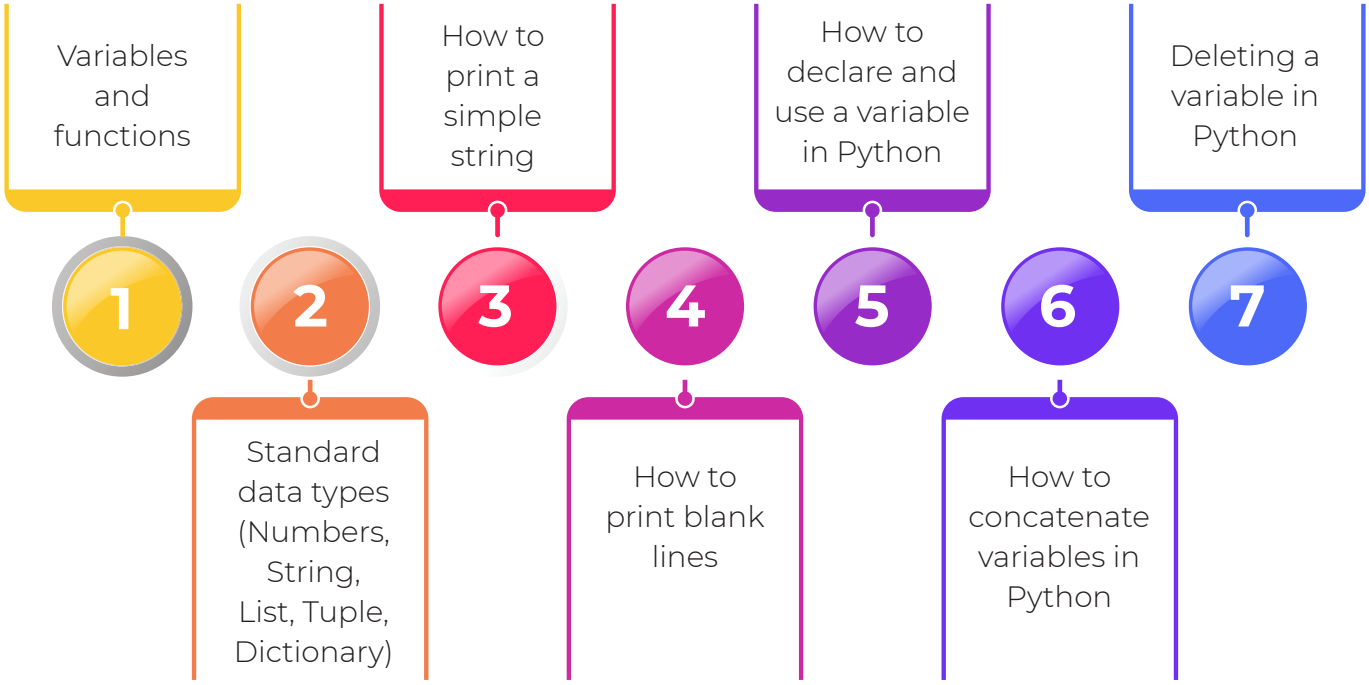
Installing Jupyter Notebook using Visual Studio Code (VS Code)

Note that this is one of the many IDEs for Python programming, that is available out there. So, feel free to use the one you like. VS Code can be downloaded from

https://code.visualstudio.com/download

7.3 CORE PROGRAMMING CONCEPTS

You must have learnt the fundamental concepts of Python language from the Base Module, which are:



Let us go deeper into a few of these concepts and learn some more.

7.3.1 Python Language Elements

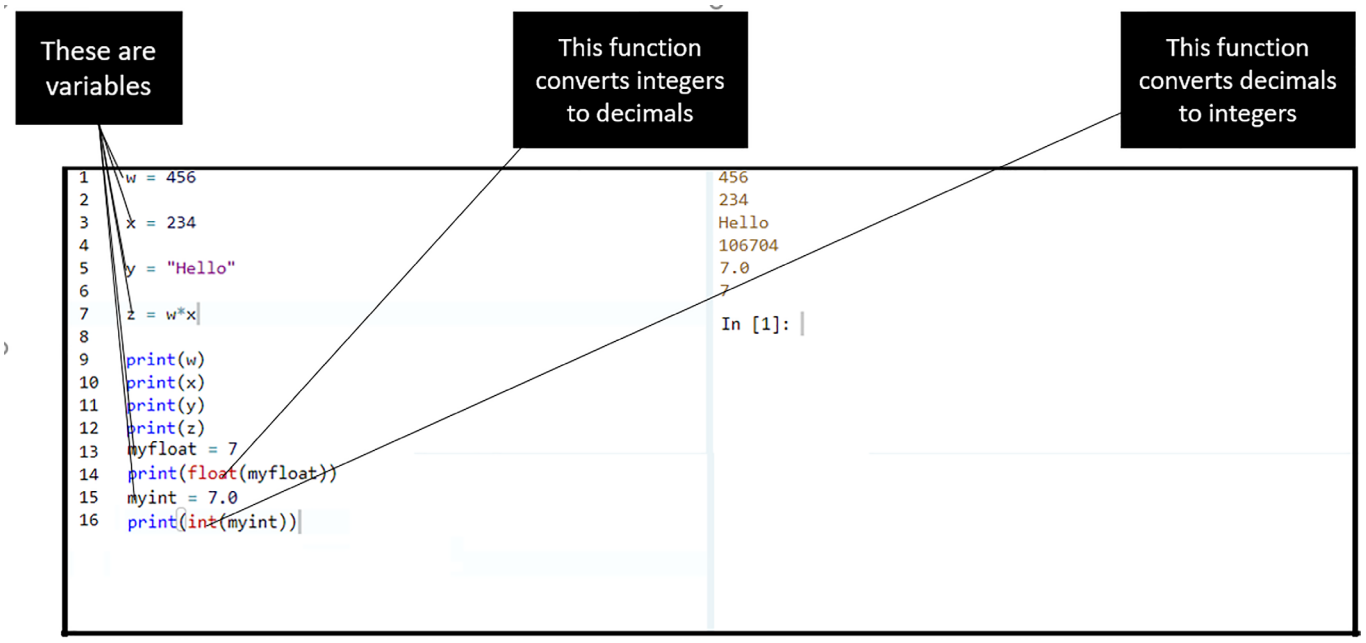
Fundamentals of Python consists of discussion of basic building blocks of Python programming language.

1. Variables

Variables are names to store data and use in computations. In Python, variables can be a sequence of characters, including special character underscore ('_ ').

Variables are containers for storing data values. Unlike other programming languages, Python has no command for declaring a variable. A variable is created the moment you first assign a value to it. Variables can be integers, decimals (Floats), strings (We will see what they are in a moment) and we can also perform calculations and much more inside variables.

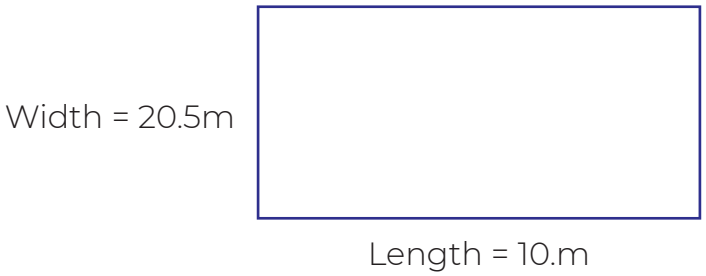
Example 1:



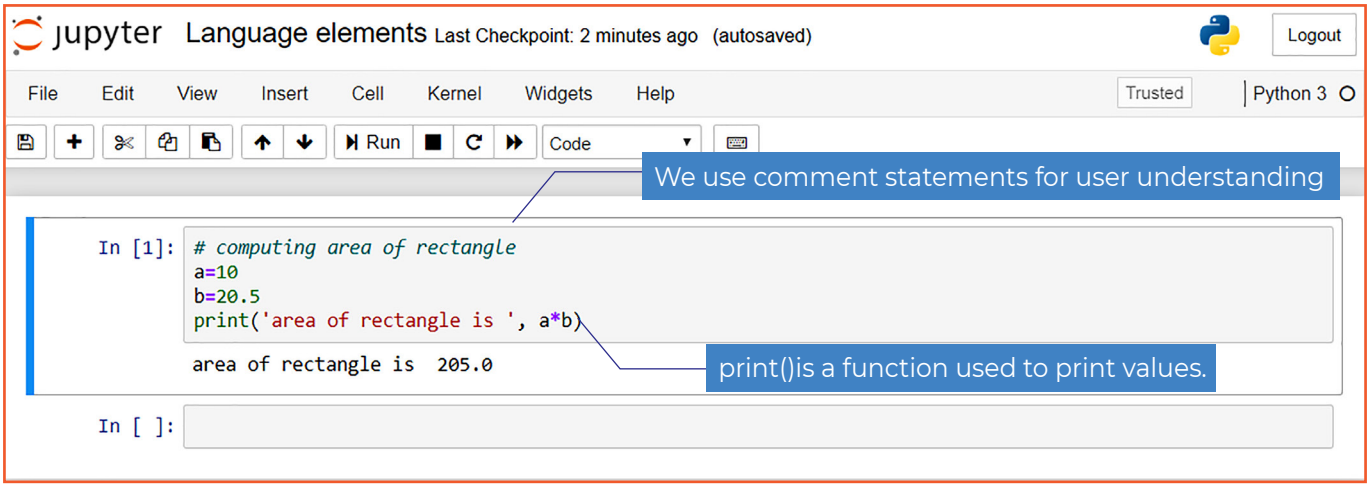
To run this program in jupyter, we press the shift key and enter key together Enter Shift on the keyboard or click on run.

Example 2:

Consider calculating the area of a rectangle in Python: Let the sides of the rectangle are given by length =10m, width=20.5 m:



We store the value of length and width in variables, say a and b respectively:



In programming, we choose names that improve the readability of the program to the outside person. Hence we define naming conventions.

Example 3:

The equivalent program to compute the area of a rectangle can be written as:

```
# improved version with good naming conventions
# this program calculates area of a rectangle
length=10
width=20.5
rectangle_area=length*width
print('Area of rectangle is', rectangle_area)
```

```
Area of rectangle is 205.0
```

From these three programs, we understand that variables are used to store the values for future calculations and print() is a function, in which we can write the meaningful statement in 'single quotes followed by a comma (,) and the result variable. In the above example, the result variable name is 'rectangle_area'. Though both these two programs perform the same computation –computing area of the rectangle, the second version above is more readable. However, the above programs have a limitation.

☞ How do we compute the area of a rectangle of 'any length' and 'any width'?

2. Functions (Methods)

One way to answer the above question is by giving a name to the piece of the program so that the values can be passed to that and that piece of the program uses the values to perform the computation and return the resulting value. For example, say computing area of the rectangle, in the above example, we can write a function in Python that takes length and width as arguments so that whenever we pass the values of length and width of a rectangle as arguments, the function computes the area for given arguments (or parameters) and the function returns the result.

Let us see how it works here.

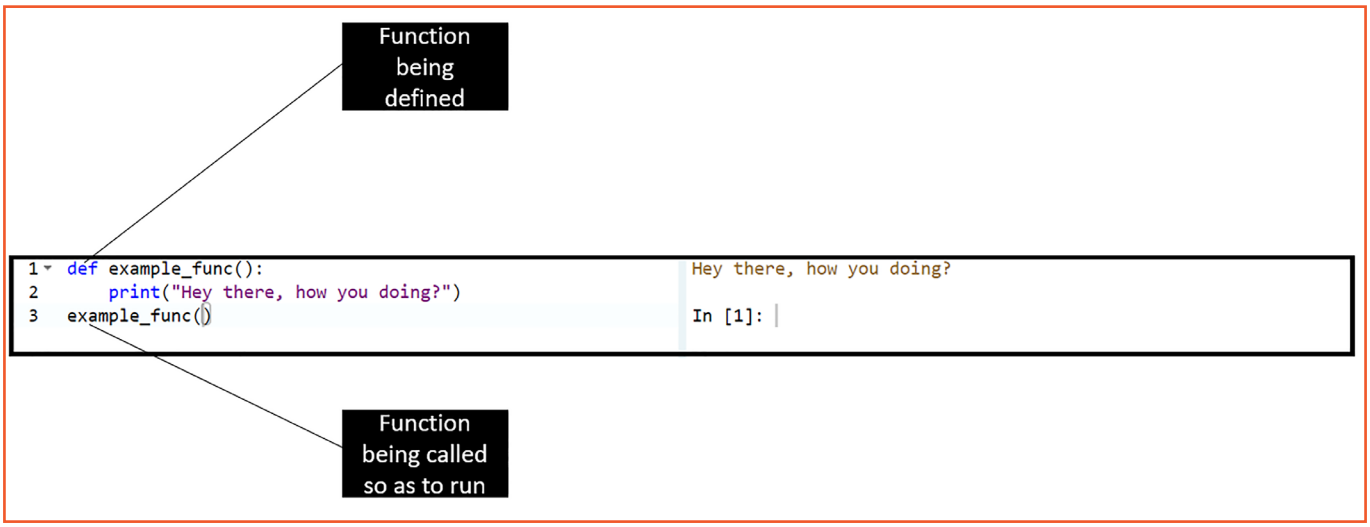
Example1:

```
In [4]: def compute_area(length, width):
            return length*width

In [5]: #computing area of a rectangle given length and width
        x=10
        y=20.5
        result=compute_area(x,y)
        print('Area is', result )

        Area is 205.0
```

In Python, a function is defined using a special word (we call key word: def), followed by the name of the function and the place holders (arguments) for the values when we call this function. This can be seen from cells [4] and [5] above. In cell [5], we declared two variables x,y and stored length and width values. Here we call the function using the statement: result =compute_area(x,y)where the values 10, 20.5 get copied to length and width respectively, and the function returns the value 10*20.5 by executing the statement return length*width in the function compute_area.

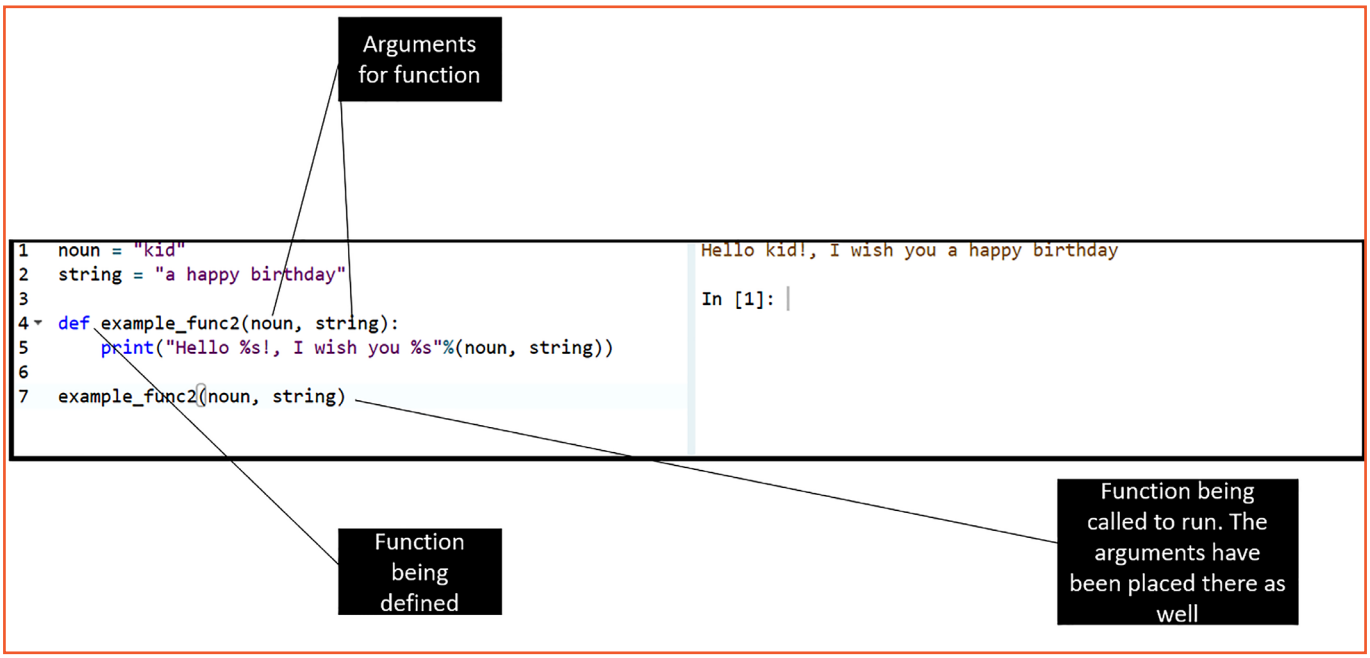
Functions are a convenient way to divide your code into useful blocks, allowing you to order your code and make it more readable, reuse it and save some time. Also, functions are a key way to define interfaces so programmers can share their code.

Example 2:



Some functions may have some arguments which are variables passed into the functions.

Example 3:

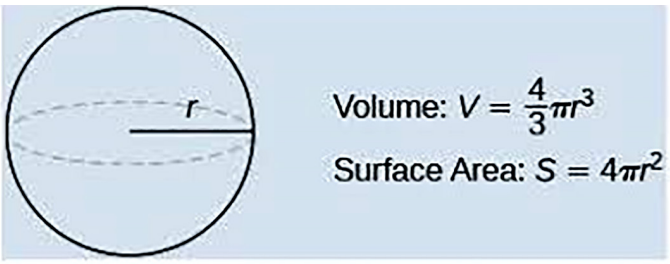


Note: The arguments have to be placed while calling the function otherwise a "MissingPositionalArgument" error would come up

⚠ Remember that the type and order of parameters that we pass must be in the same order as that of function defined.

Example 4:

Let us consider we would like to write a function to return surface area and volume of a sphere, given its radius. The formulae for surface area and volume are given by: We name this function: get_SA_and_Volume:




```
In [6]: def get_SA_and_Volume(radius):
            PI=22/7
            area=(4/3)*PI*radius*radius*radius
            volume=4*PI*radius*radius
            return area, volume

In [10]: r=1.00
         area_r, vol_r=get_SA_and_Volume(r)
         print('Area and volume of sphere of radius 1 is', area_r, vol_r)

         Area and volume of sphere of radius 7 is 4.19047619047619 12.571428571428571
```

A function in Python can return more than one value. So it is important that we collect these returned values in the order in which the function returns them. In the function above, **get_SA_and_Volume(radius)** contains a sequence of statements (i.e., four statements) which are indented with some column margin (called indentation). Such a group of statements with indentation is called a block and the statements within the block can be called as compound statements. Python allows us to create such blocks with 4 space indentation. As you see, the jupyter notebooks by default we get this indentation instead of we typing these blank space.

3. Indentation

Indentation refers to the spaces at the beginning of a code line. Where in other programming languages the indentation in code is for readability only, the indentation in Python is very important. Python uses indentation to indicate a block of code.

Example 1:

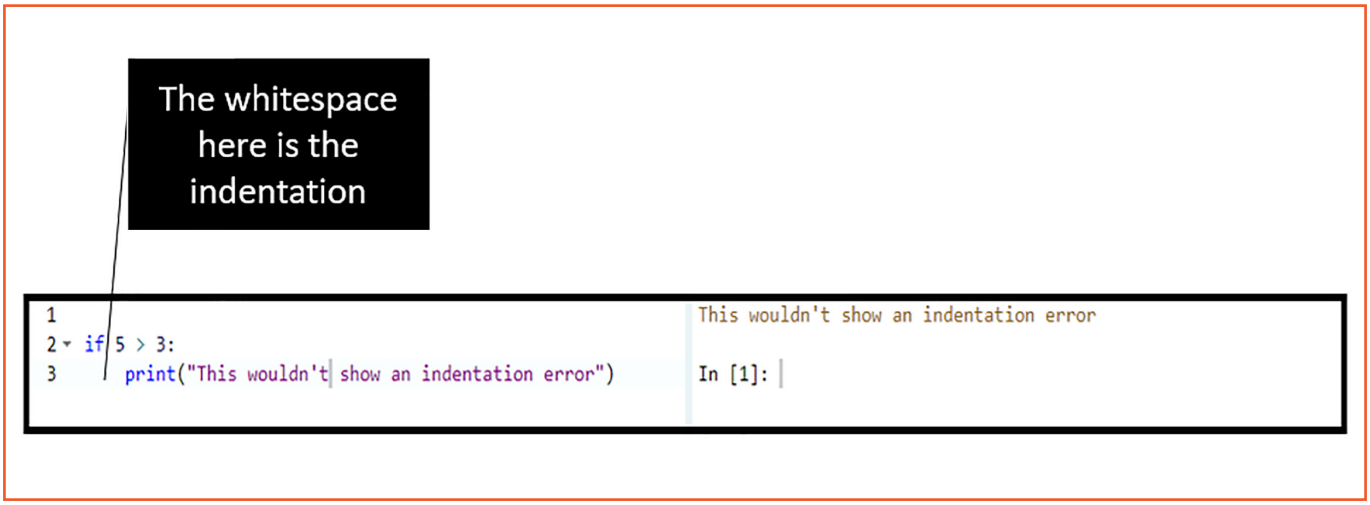
script.py

```
1
2 if 5 > 3:
3   print("This would show an indentation error")
```

IPython Shell

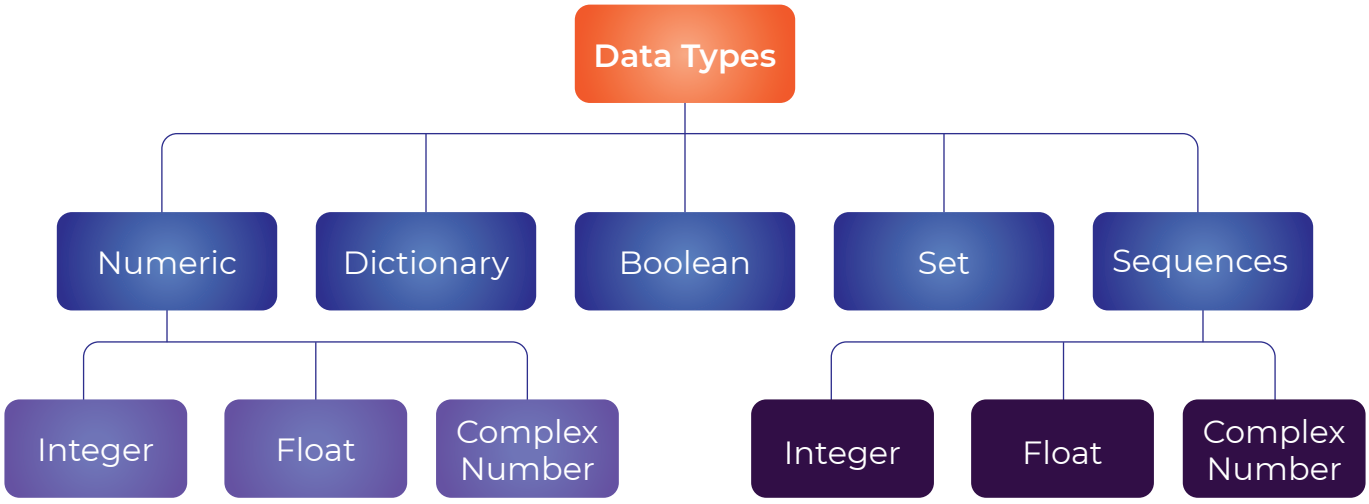
```
  File "<stdin>", line 3
    print("This would show an indentation error")
        ^
IndentationError: expected an indented block
```

But if we add indentation;



4. Data Types

We select a suitable data type based on the computation that we are performing. For example, to perform arithmetic operations such as the above examples where we used floating-point (real numbers for the performing area of the rectangle. Similarly, for processing character-based sequences (strings), it will be unrealistic to think operations such as multiplication of strings and division of strings. However, there are specific computations defined for each of these data types. For example, string reversal, finding a character in a string at a particular location, finding the number of characters in the string (length) etc. are valid operations on the strings. Below is a high level view of various data types we commonly use in Python.



Integers:

Integers are whole numbers – for example, 5, 10, ...

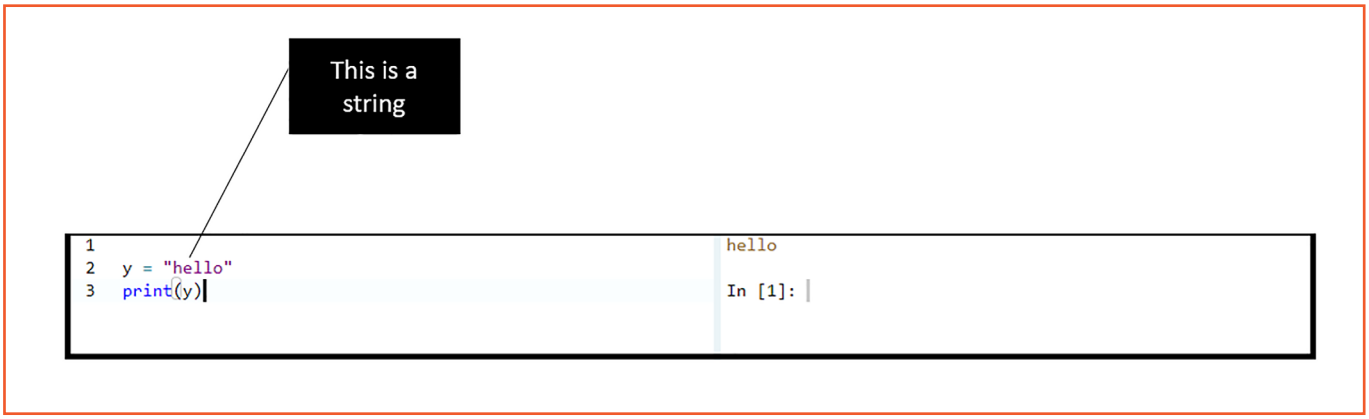
Floating-point numbers:

Unlike integers floating point numbers will have two components, namely integer part and fractional part. For example, 24.8973.

Strings:

Strings are a sequence of characters enclosed between either single quotes or double-quotes.

Example:



Boolean:

Boolean in Python takes two values True, False. Notice the first character written in capital Letter. The expressions which are part of checking conditions in branch statement and iteration statements are evaluated to True or False. The statement types you will see in the subsequent section.

In Python, the type of the variable used in computation is based on the value we assign on the right-hand side of the assignment statement. The following example illustrates this where we use the built-in function: type() to check the type of the variable –write the value within these parentheses. For example, a string value can be assigned with single quotes '-' or double quotes "-".

```
name='Narendra Bahubali'
print( type(name))
```

```
<class 'str'>
```

```
class_average=54.50
print( type(class_average))
```

```
<class 'float'>
```

```
class_total=150
print(type(class_total))
```

```
<class 'int'>
```

5. Statements

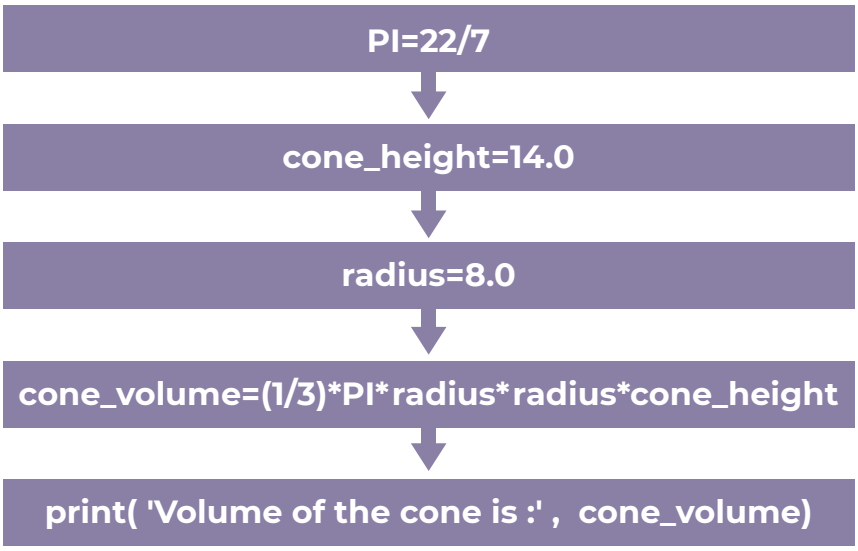
In programming, there are three types of statements – namely compound statements, conditional or branching statements and iterative statements. Few considerations for selecting these statements include:

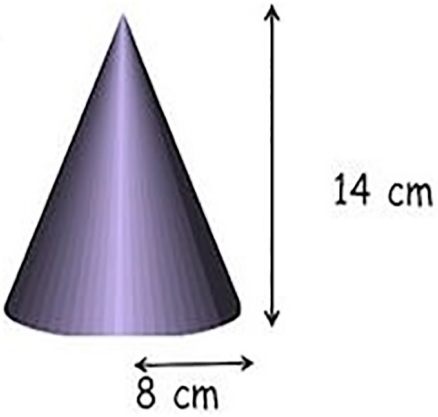
- ☞ Statements are to be executed in a sequence? (see previous examples)
- ☞ Do we require conditional execution of these statements?
- ☞ Is there an iteration/loop required for repeated execution of the same sequence of the statements for a given number of times?

For example, printing the name of a student 10 times involves usage of iterative statement.

Sequential Statements:

A flow chart for computing volume of a cone (V), given the radius r, height h is given by $V=1/3\ \pi r^2\ h$:





Conditional Statements:

Conditional statements help us define the change in behaviour of the program – for example, execution of a few set of statements under one condition. Hence the condition checking is performed through Boolean and logical expressions using suitable operators appropriately and the results of these expressions are either **True or False.**

```
# examples for Boolean expression
val=5
print(val == 5)
```

```
True
```

```
print(val ==7)
```

```
False
```

```
# example for Logical expressions
val >3 and val < 10
```

```
True
```

```
val % 2 ==1
```

```
True
```

Boolean Operation	Result	Logical (**and, or, not**)	Result
x != y	x is **not equal to** y ?	**(x%3==0) and (x%2==0)**	x is divisible by 3 **and** x is divisible by 2?
x == y	x **equals** y ?	**x >3 and x <=10**	x lies between 4 and 10?
x < y	x **is less than** y?	**x %2 == 0 or x%5 == 0**	x is a multiple of 2 or multiple of 5?
x <= y	x **is less than or equals** y?	not x	It is the complement of x-If x is true, not x evaluated as false
x >= y	x **is greater than or equals** y?		
x is y	x **is the same as** y?		

Example for not:

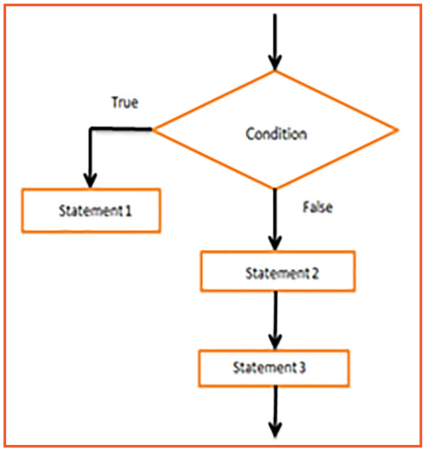
```
is_married =True
not is_married
```

```
False
```

Example of conditional execution using if-else

```
# printing whether given number is even or odd

number =int(input('Enter an integer'))

if number%2 == 0:
    print('you entered an even number')
else:
    print('you entered an odd number')
    print('Have a nice day')
```

```
Enter an integer5
you entered an odd number
Have a nice day
```



You can also have if statement without else block. Also, Python provides chained conditionals -through **if – elif – else** programming construct as explained below:

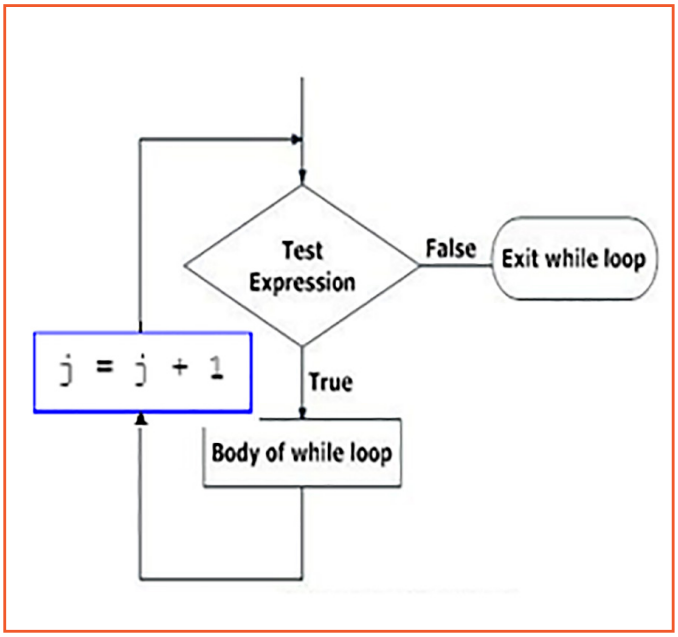
```
# using chained conditionals
x=10
y=10

if x < y :
    print('x is less than y')
elif x > y:
    print('x is greater than y')
else:
    print(' x equals y')
```

```
 x equals y
```

Observe that the above construct is helpful, particularly in the case where operands x and y are common in each branch of test evaluation.

Iteration Statements:

Iteration statements are helpful when a piece of program code to be executed for a specific number of times, say as a repetitive task.



This is where the computations performed by programs are error-free compared to human beings because we as humans might make mistakes, especially the performing the same task again and again due to overlooking certain aspects.

As shown in the flow chart, the body of the look gets executed until the test condition fails. Once the test condition fails, we call it a termination condition for the loop, and the statements after the loop get executed as usual. Python provides while statement as two iteration constructs namely while loop and for a loop as given below: One of the vital aspect for iteration statements I update of the loop variable. The following example illustrates the points discussed here.

```
# Example for while construct -printing 6 table upto 10x6=60
number =6
count=1
while (count <=10):
    print(count,'x',number,'=', count*number)
    count=count+1
```

Initialization of iteration counter

Value after every each iteration

Statement for updating the counter variable

```
1 x 6 = 6
2 x 6 = 12
3 x 6 = 18
4 x 6 = 24
5 x 6 = 30
6 x 6 = 36
7 x 6 = 42
8 x 6 = 48
9 x 6 = 54
10 x 6 = 60
```

⚠ The termination condition (i.e., when to stop iterations) is provided as part of the body of while either through the count update till desired no of iterations or when a particular result is obtained during these iterations.

⚠ The loop without termination condition gets executed indefinitely and hence it is called infinite while. Hence the programmer needs to keep in mind regarding this termination condition while developing iterative constructs.

- ☞ break statement can be used for early termination of the loop as shown in the below:

- ☞ Also, multi-line comments can be written by enclosing the text within the three single quotes (‘ ‘ ‘) in the beginning and at the end.

Strings: A sequence of characters terminated by a newline \n character or any other character. Individual characters in a string can be accessed using index positions as illustrated below:

Consider a string: My_str=’Python’; Let us initialize index variable, say, index=0. My_str[0] contains character ‘P’, My_str[3] contains character ‘h’ and so on. We increment index and retrieve the characters from the string. The following example illustrates this.

P	y	t	h	o	n
0	1	2	3	4	5

The following function takes an input string (a sequence of characters) that are ended with ! and returns the string till $. In this example, the input string variable is input_string,

```
'''
function that takes input_string terminated by !;groups sequence of characters till dollar
and returns new string so formed
'''

def get_string_until_dollar(input_string):
    current_pos=0
    output=''
    while (input_string[current_pos]!='!'):
        output=output+mystring[current_pos]
        if mystring[current_pos]=='$':
            break
        current_pos=current_pos+1
    return output
```

The first character in python strings is indexed at 0

Initialize output string output with empty string string

while loop gets terminated when the character at current_pos is '!'

break the loop if the character at current_pos is $

Increment current_pos by 1

```
my_string='I ordered few books online. Amount spent was 200$. The books are very good!'
s1=get_string_until_dollar(my_string)
s1
```

```
'I ordered few books online. Amount spent was 200$'
```

- **while** loop can be used when the termination condition is not known prior to the execution. In the case of the previous example, the loop reads character by character until it encounters the character '!'. This indicates, prior to the execution of the loop, we may not always know the number of characters in the input string. Hence while the loop is called the indefinite loop.
- Sometimes we want to iterate through a collection of items such as a list of numbers, words, etc., where the size of the collection is finite. In such cases, we use for loop, as shown below:

We know in mathematics, a prime number is the one which has 1 and itself as factors. We have may such prime integers in the set of natural numbers, N. For example, 3, 7, 11, 29 ... these are prime numbers as they don't have any other factors except 1 and the number itself. Note that 2 is only even prime number.

Consider a program that checks whether a given number n is a prime or not. One way of checking it is to divide the number with all the numbers from 1 to N and check whether there are any factors between 1 and the number, n itself. So we generate a sequence of integers in the range of 2 to n-1 and divide n with each number from this sequence. We initially assume that the given number is prime. During the division on the first encounter of the factor number from the sequence, we flag that the number is not prime. Let us see how it works:

Let n = 6; The sequence: 2, 3, 4, 5 and 2 is a factor of 6 and the function should return that 6 is not prime OR **is_prime(6)** is False. Similarly, the function **is_prime(5)** should return true.

Consider the program snippet for this problem below:

```
In [14]: # function to check whether a number is prime or not
         def is_prime(n):
         # let n is prime. this is represented by variable is_prime
             is_prime=True
             for i in range(2, n):
                 if n%i ==0:
                     is_prime=False
                     break
             return is_prime

In [15]: is_prime(4)
Out[15]: False

In [16]: is_prime(5)
Out[16]: True
```

range(start, end) is python provided function that returns a list of values from **start to end-1**

There are other efficient ways of implementing this function. For example, instead of checking for factors from 1 to n, even 1 to √n which you can try on your own and see.

- The statement for i in range(2,n) is an iterative statement and all the statements within its scope (body of the loop) are executed n-1 times
- Python provides in-built function range(2,n) that returns list of values [2, 3, ...,n-1]
- The statement if n%i == 0 is a branching statement. The operator % returns reminder when n is divided by i

7.3.2 Python Data Structures

Lists

Lists can contain any type of variable, and they can contain as many variables as you wish. Lists can also be iterated over in a very simple manner.

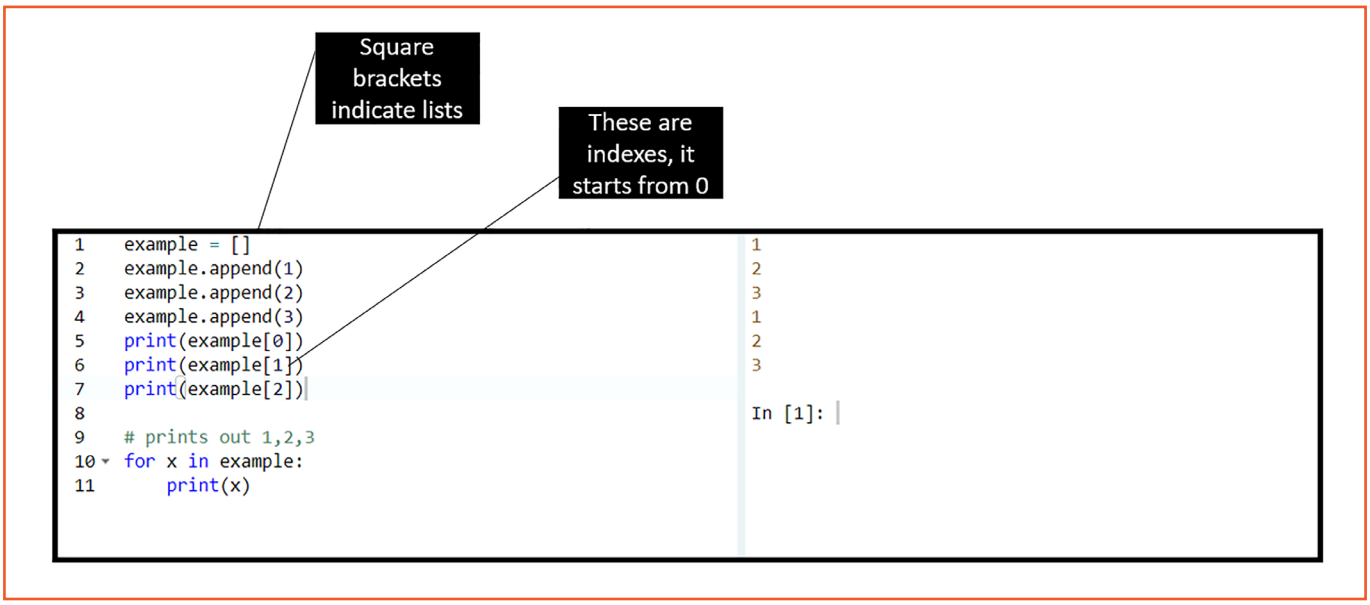
Example 1

Note: The index of python starts from 0, and so

0 = 1

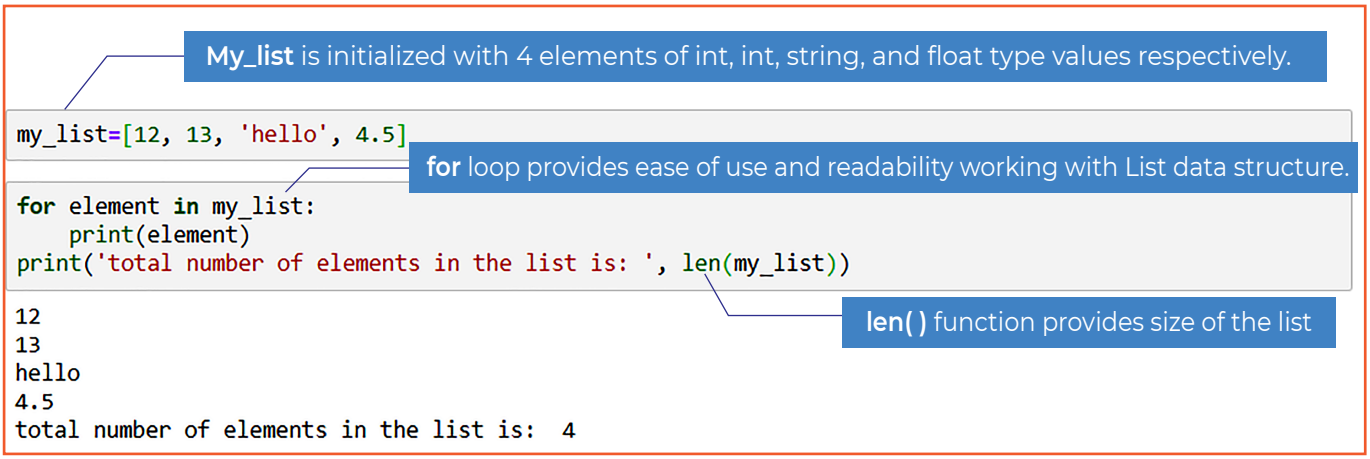
1 = 2

2 = 3



A List is a heterogeneous collection of elements separated by a comma. The number of elements in the list can be obtained using the **len( )** function.

Example 2



my_list	12	13	'hello'	4.5
index positions →	0	1	2	3

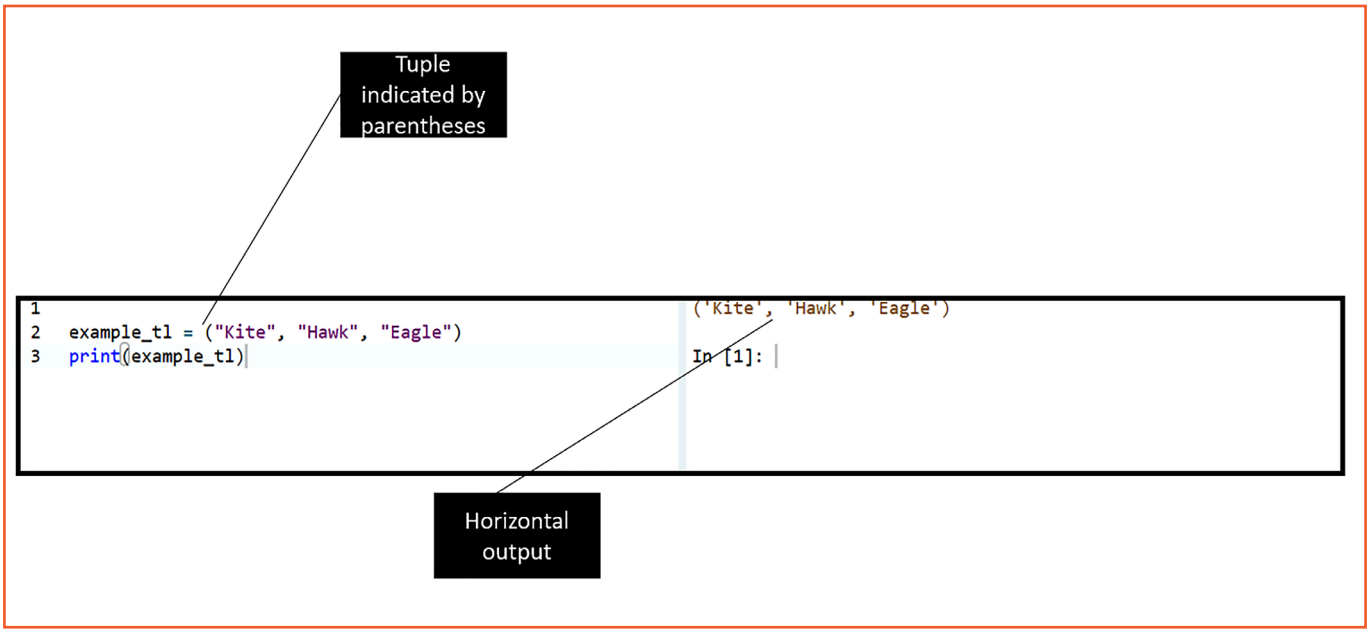
Elements of the list can be accessed using index position List methods and usage:

- Index positions help in accessing elements of the list. In Python index starts from 0.
- In the above example, my_list[0] returns the first element in my_list, i.e., 12, my_list[1] returns the second element in my_list, ie., 13 etc.
- The last index position of the list will be one less than the number of elements in the list. In the above example, there are 4 elements and the last index is 3 and hence my_list[3]is the last element with value 4.5.

Tuples

A tuple is an immutable sequence of Python objects. Tuples are sequences, just like lists. Except parentheses indicate tuples, while square brackets indicate lists.

Example 1



As you notice, there are updated methods available in the list data structure (like append, which helps you adding a new object to the existing list). Unlike lists, tuples in Python are immutable, indicating that as the tuple is defined, we will not be able to change the contents of it through addition or deletion of its elements. The tuple is useful in the situations where you would like to provide the data to the outside world and at the same time, you prevent modification of the data. The common example of such a situation is to read employee records from a file and get the relevant details without actually modifying any of these records. Following Python example illustrates storing and retrieving student records which are stored in a list. The retrieval is based on the grade **'A+'.**

Example 2

Step1: Define five tuples storing the data of each student:

```
# each student record values are stored in tuple where % and grade obtained
student1 =('Sita', 'X class', 'E section', 89.5, 'A')
student2= ('Rajesh', 'X class', 'B section', 91.5, 'A+')
student3 =('Kiran', 'IX class', 'C section', 72.5, 'B+')
student4=('Anil', 'X class', 'A section', 62.5, 'B')
student5=('Joseph', 'X class', 'B section', 96.5, 'A+')
```

Step2: Store these records in a list with name students:

```
# store them in a list
students=[]
students.append(student1)
students.append(student2)
students.append(student3)
students.append(student4)
students.append(student5)
```

Step3: Retrieve the student records who scored 'A+' grade:

```
print('Student records who scored A+ grade')
for s in students:
    if 'A+' in s:
        print(s)
```

```
Student records who scored A+ grade in class X
('Rajesh', 'X class', 'B section', 91.5, 'A+')
('Joseph', 'X class', 'B section', 96.5, 'A+')
```

Exercise: Re-implement this using set data structure creating a set object S =set( )

Which of the above to data structures is appropriate in the given context and why?

Dictionary

The data structures you have learned so far, i.e., strings, lists and tuples use integers as indices and you retrieve the value at a particular index position by [<index_position>]. You will notice an error If you use any other data type. Also frequently you require to store the data based on a key-value which can be a string such as the name of a person. In such cases, dictionary data structure comes handy. Dictionaries allow you to use any other immutable type as an index, such as string, tuple. Each entry in the dictionary contains an index and a value separated by a colon (:). We can add a new object to dictionary using [ ] operator.

Example 1: The following code example illustrates the usage of the string as an index. In this example, we create a dictionary depicting the number of fruits in a basket:

```
# Make an empty basket
basket ={}
basket['Apple']=50
basket['Mango']=20
basket['Orange']=15
basket['pears']=10
print(basket)
```

```
{'Apple': 50, 'Mango': 20, 'Orange': 15, 'pears': 10}
```

Alternatively, you can also initialize all fruits in one statement in a dictionary:

```
#Alternatively you can also store the objects all at once
basket={'Apple':50, 'Mango':20, 'Orange': 15 }
# adding 10 pears to basket
```

☞ What happens if you write a statement: basket['Mango']=50

Example 2: Suppose, we would like to compute the number of occurrences of each character in a word. This can be implemented using the dictionary where each letter in the input word is used as a key and the number of occurrences of that letter (i.e., character frequency) as its value:

```
# finding the frequency of words in a sentence
letter_freq={}
word='Mississippi'
for current_character in word:
    letter_freq[current_character]=letter_freq.get(current_character,0) + 1

print('letter frequency', letter_freq)
```

```
letter frequency {'M': 1, 'i': 4, 's': 4, 'p': 2}
```

Some more Python concepts

To learn how to write codes, you will also have to learn some more concepts of Python.

- Python Syntax
- Python Comments
- Python Casting
- Python Booleans
- Python Operators
- Python Sets
- Python Dictionaries

- Python If...Else
- Python While Loops
- Python For Loops
- Python Functions
- Python Lambda
- Python Arrays
- Python Scope
- Python Modules
- Python Iterator
- Python Dates
- Python User Input
- Python String Formatting

EXPLORE

You can learn these from any one of the following links:

https://www.w3schools.com/python/default.asp

https://www.programiz.com/python-programming

https://www.techbeamers.com/python-tutorial-step-by-step/#tutorial-list

You can practice writing lines of code using the concepts learnt in Jupyter Notebook/ Python 3.

7.4 ESSENTIAL DATA SCIENCE LIBRARIES

A collection of core modules can be called libraries. A Library can also be called a module. Some of the most important modules of Python for Data Science include:

To learn these:

- Open up a new Jupyter Notebook.
- Read the library's training materials from the website provided after each.
- Import the library into your Jupyter Notebook.
- Practice the programs from the training material using Python 3.

We recommend diving much deeper into a library later, as one may forget most of what has been learned by the time it's time for using it in the projects. Instead, right now one can aim to discover what each library is capable of.

7.4.1 Pandas

Pandas is a software library created for the Python programming language. It is aimed at data manipulation and analysis. It primarily offers data structures and operations for manipulating numerical tables and time series. Given below are examples of some pandas functions;

```
import pandas as pd

example = pd.read_csv('./random_numbers.csv')
#We have read the file
#Now let's perform some operations and save it as a csv
x = 5
y = example['Numbers'] #We are accessing the column here
z = x*y
print(z)
z.to_csv('Example.csv')
#Great the calculation worked and is now stored as a csv inside our current working directorty
```

```
0       5
1      10
2      15
3      20
4      25
5      30
6      35
7      40
8      45
9      50
10     55
11     60
12     65
13     70
14     75
15     80
16     85
17     90
18     95
19    100
20    105
Name: Numbers, dtype: int64
```

You can learn the following using Pandas by visiting the website provided.

- Creating a DataFrame
- Indexing DataFrames
- Reshaping Data: Change the layout of a data set
- Combining Datasets
- Working with text data
- Working with missing data
- Computational Tools
- Visualization

EXPLORE

https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html#user-guide

https://www.w3resource.com/pandas/index.php

7.4.2 Numpy

Numpy is a library for the **Python** programming language, adding support for large, multi-dimensional arrays and matrices. It has an extensive collection of high-level mathematical functions to operate on these arrays. Let's look at some Numpy functions

```
import numpy as np

x = -134567
x = np.abs(x)
print(x)
```

```
134567
```

Learn the following concepts about Numpy:

- Python NumPy
- NumPy Intro
- NumPy Getting Started
- NumPy Creating Arrays
- NumPy Array Indexing
- NumPy Array Slicing
- NumPy Data Types

- NumPy Copy vs View
- NumPy Array Shape
- NumPy Array Reshape
- NumPy Array Iterating
- NumPy Array Join
- NumPy Array Split
- NumPy Array Search
- NumPy Array Sort
- NumPy Array Filter
- NumPy Random
- NumPy ufunc

EXPLORE

https://numpy.org/doc/stable/user/

https://www.w3schools.com/python/default.asp

7.4.3 Scikit-Learn

Scikit-learn is a free software or machine learning library for the Python programming language. It features various classification, regression and clustering algorithms

EXPLORE

https://scikit-learn.org/0.18/_downloads/scikit-learn-docs.pdf

https://scikit-learn.org/stable/index.html

7.4.4 Modules for Data Visualization

Data visualization is the graphic representation of data. It involves producing images that communicate relationships among the represented data to viewers of the images.

1. Dash

Dash is a Python framework for building web applications. It is built on top of Flask, Plotly. It is used to build dashboards based on Python.

2. Plotly

Plotly provides online graphing, analytics, and statistics tools for individuals and collaboration.

3. Matplotlib

Matplotlib is a plotting library, used with the Python programming language and its numerical mathematics extension NumPy.

7.4.5 CSV files(Comma Separated Values)

A CSV is a comma-separated values file, which allows data to be saved in a tabular format. CSVs look like a garden-variety spreadsheet but with a .csv extension. We have already seen how we can read and write a csv but for remembrance's sake let's look at it again.

```
import pandas as pd

example = pd.read_csv('./random_numbers.csv')

y = example['Numbers'] #We are accessing the column here

y.to_csv('Example.csv')
```

7.4.6 Heroku and deploying applications

Heroku

Heroku is a container-based cloud Platform as a Service. Developers use Heroku to deploy, manage, and scale modern apps.

We have talked about Heroku and know that it is used for deploying Python-based applications. So how do we deploy it? It is actually pretty simple for you to deploy a python application, but first, you'll have to make sure that you have a Heroku account. Sign up for your account on Heroku. So now let's see on how we can deploy it?

1). Type the command "Heroku login" to login to your Heroku account

2). Initialize your directory as a git repository

3). Activate your virtual environment

4). Install gunicorn with pip (I already have it, so the output says 'requirement already satisfied')

5). Next, since we can't run 'pip install ...' commands on the server, we will have to make a requirements file which consists of all the modules we require (Changes can be made by going to the file explorer and add the modules missing into the file)

6). Use the normal git commands to add and commit the code

7). Specify the app name to the application you have created

8). Before we start pushing the code into 'flaskexampleproject', let's make sure that we will be pushing the code to the right place. Do so by running this command

9). Now push the code by running this command

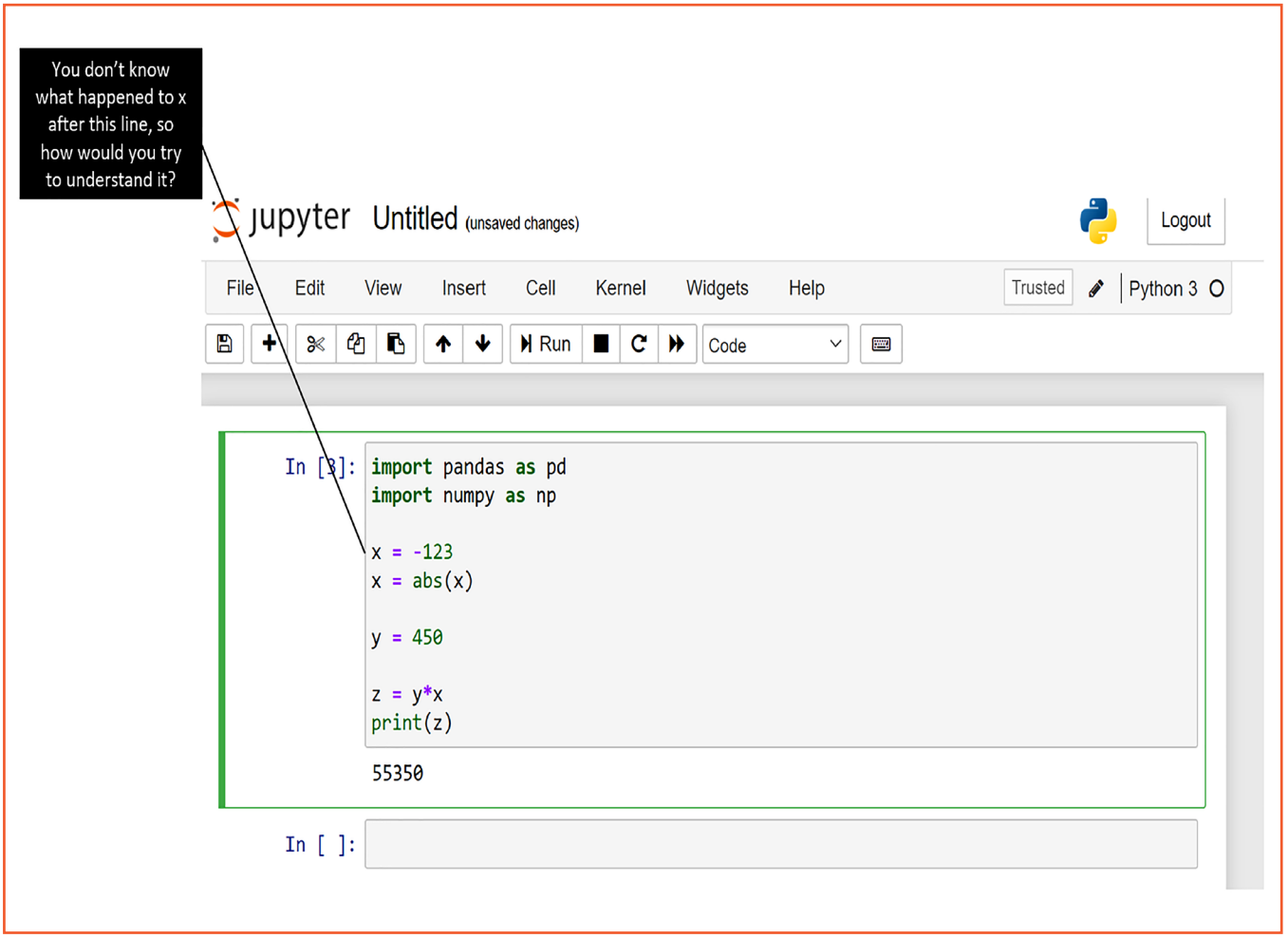
10). Scale the web and open your application

11). Open your application

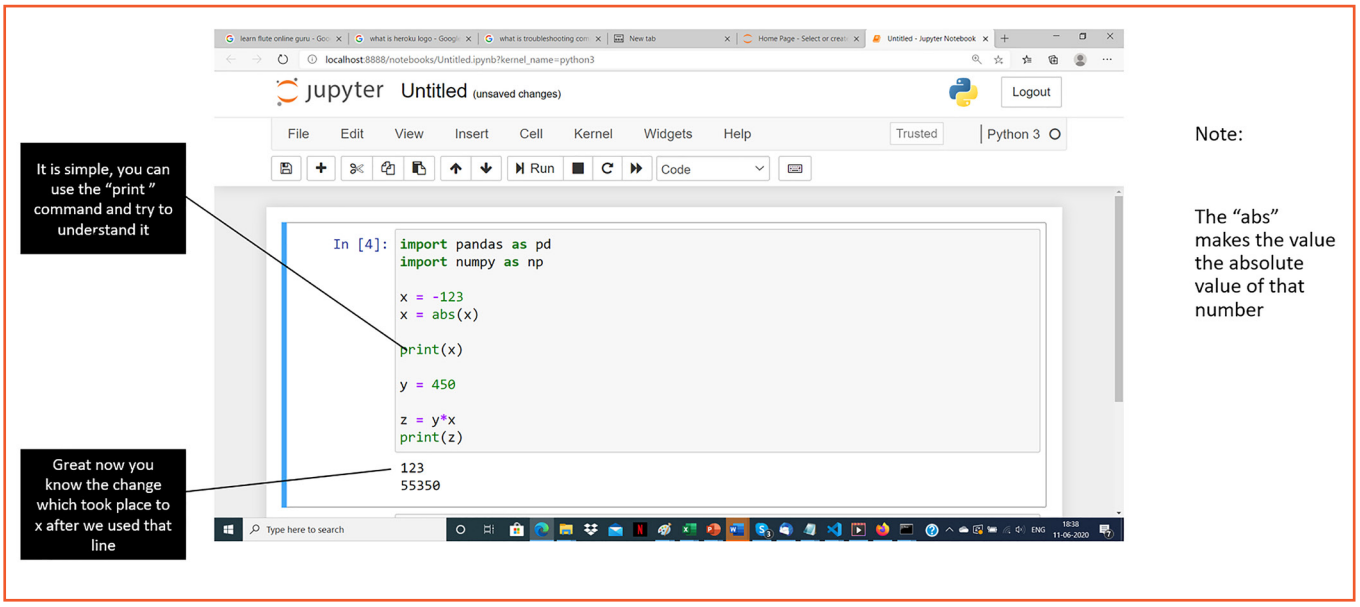
7.5 TROUBLESHOOTING DEFECTS IN PYTHON CODE

In the jupyter notebook, print() and other functions to inspect the objects can be used. For instance, head() of dataframe, shape() of dataframe, etc

Here is a simple example to inspect variable and code statements



And the below screen shows the results



By now, you'll have a basic understanding of programming and a working knowledge of essential libraries. This covers most of the Python you'll need to get started with data science.

7.6 PRACTICE FOR WRITING CODE

Now is the time to solidify your knowledge through plenty of practice.

For Beginners:

Beginners can use the following websites to get very simple exercises and their solution. Try to develop the solution to the exercises on your own using Python 3. After you are done, you can compare it with the solution provided. Always remember that there are more than ways to get the desired output.

EXPLORE

https://w3resource.com/python-exercises/heap-queue-algorithm/index.php

https://www.w3schools.com/python/python_exercises.asp

After mastering basic codes:

EXPLORE

Now try the following practice exercises.

http://www.practicepython.org/

In later Units, you will be given projects that you can do where you will have to use Python extensively. Hence, it is a good idea that you practice with as many exercises as you can.

ACTIVITY

Try this simple activity for creating a Chatbot using Python

EXPLORE

Go to the following link and follow all the steps in it.

http://www.practicepython.org/

SUMMARY

Python is an interpreted, high-level, general-purpose programming language. A python program is a sequence of Python statements (code) written to perform some computational task.

An integrated development environment is a software application that provides comprehensive facilities to computer programmers for software development. It helps programmers by providing tools for various common activities of writing software in a single application. Some such IDEs that are commonly used for AI and Machine learning are Jupyter, Spyder, PyCharm, PyDev and Visual Studio.

Fundamentals of Python consists of a discussion of basic building blocks of Python programming language.

1. **Variables:** Variables are names to store data and use in computations. In Python, variables can be a sequence of characters, including special character underscore ('_ ').
2. **Functions (Methods):** One way to answer the above question is by giving a name to the piece of the program so that the values can be passed to that and that piece of the program uses the values to perform the computation and return the resulting value.
3. **Indentation:** Indentation refers to the spaces at the beginning of a code line. Where in other programming languages the indentation in code is for readability only, the indentation in Python is very important. Python uses indentation to indicate a block of code.
4. **Data Types:** These are Numeric (integer, float, complex number); Dictionary; Boolean; Set; Sequences (string, list, tuple)
5. **Statements:** In programming, there are three types of statements –namely compound statements, conditional or branching statements and iterative statements.

Python data structures involves the following:

- **Lists:** Lists can contain any type of variable, and they can contain as many variables as you wish. Lists can also be iterated over in a very simple manner.
- **Tuples:** A tuple is an immutable sequence of Python objects. Tuples are sequences, just like lists. Except parentheses indicate them while square brackets indicate lists.

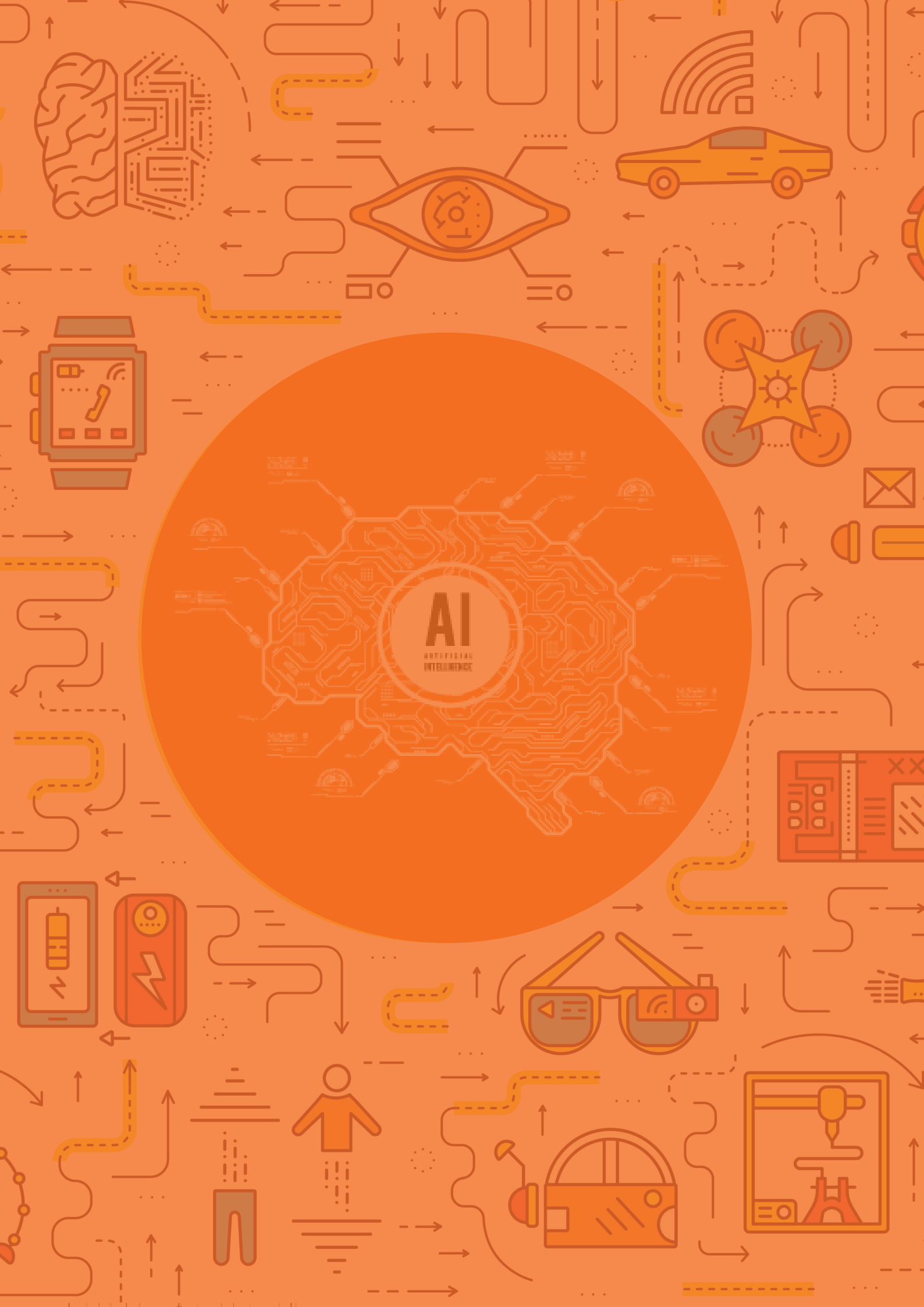
- **Dictionary:** The data structures you have learned so far, i.e., strings, lists and tuples use integers as indices and you retrieve the value at a particular index position by [<index_position>].

A collection of core modules can be called libraries. It can also be called as a module. Some of the most important modules of Python for Data Science include:

- Pandas is a software library created for the Python programming language, aimed at data manipulation and analysis.
- Numpy: Numpy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices.
- Scikit-Learn: Scikit-learn is a free software or machine learning library that features various classification, regression and clustering algorithms

Data visualization is the graphic representation of data. Python allows the same using Dash, Plotly and Matplotlib.

Heroku is a container-based cloud Platform as a Service. Developers use Heroku to deploy, manage, and scale modern apps.


AI
ARTIFICIAL
INTELLIGENCE

UNIT 8
ALGORITHMS

Key Learning Outcomes

At the end of this module, you will be able to:

a) List the key machine learning algorithms for supervised, unsupervised and reinforcement learning
b) State applications of key machine learning algorithms
c) Train and evaluate a classification, regression and clustering model algorithm

Topic Covered

Introduction to Machine Learning Algorithms |
Algorithms Used in Supervised Learning |
Algorithms Used in Unsupervised Learning |
Algorithms Used in Reinforcement Learning

8.1 INTRODUCTION TO MACHINE LEARNING ALGORITHMS

Algorithms are at the heart of Artificial Intelligence. It is the model that ultimately does the intelligent work.

In the Base module, you had dabbled a bit with Algorithms, so you would remember that Algorithms are step-by-step instructions to solve a problem, created along with rules that decide which path to take when a decision is faced. Numerous Algorithms are used in various fields of Data Science. In this Unit, you will learn about the Algorithms very commonly used in Artificial Intelligence and Machine Learning.

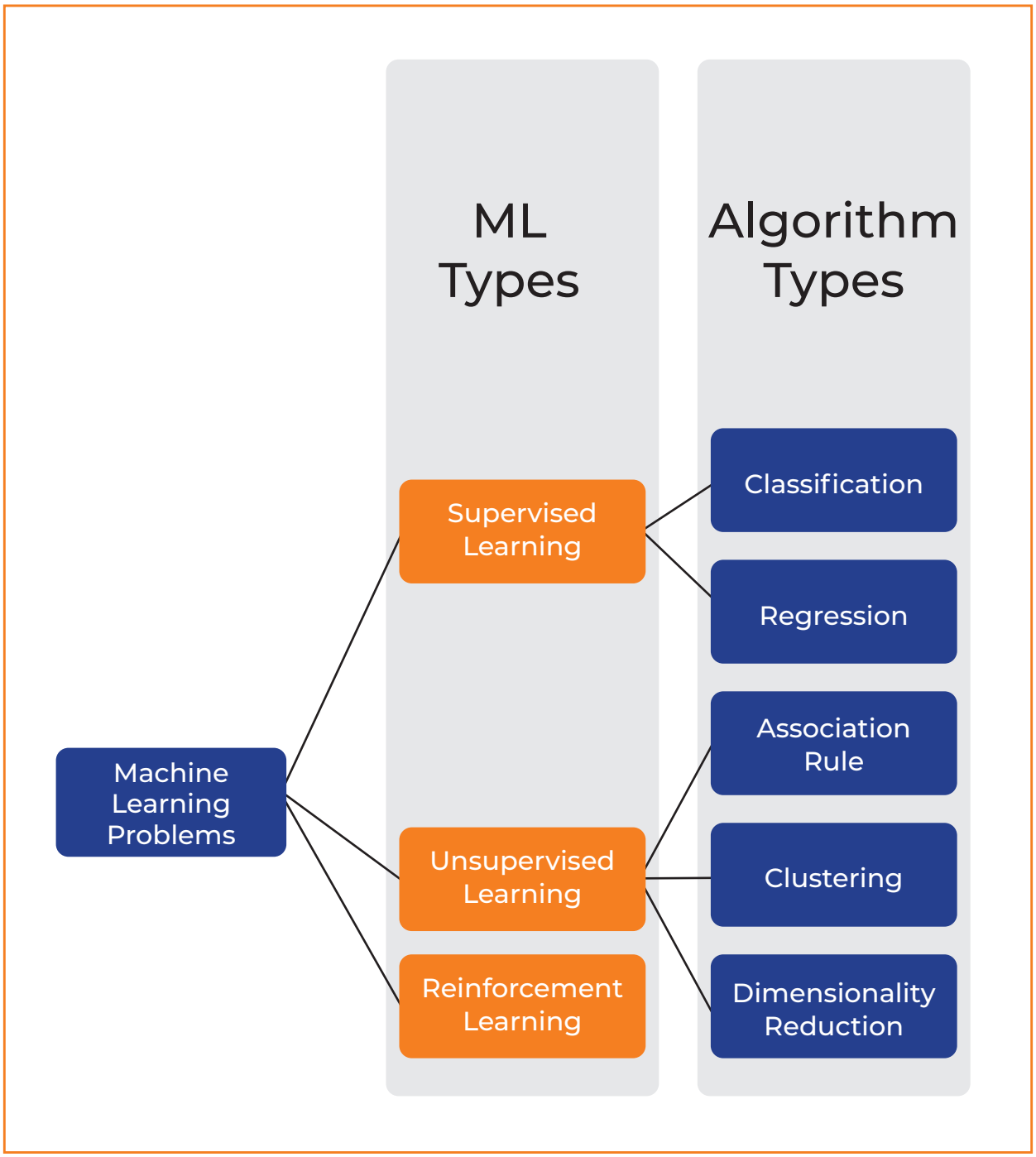
In Machine Learning, the Algorithm is a program that can alter its own parameters based on given responses on the past predictions of the data set. Based on the type of problems to be solved, we can divide Machine Learning Algorithms into three main categories:

Supervised Learning – Where the algorithms learn based on existing labels/target to make better predictions.

Unsupervised Learning – Where algorithms learn without labels/target to identify insights/clusters.

Reinforcement Learning – Where algorithms learn based on trials and errors to maximize rewards.

Each of these threeMachine Learning Algorithm categories also has a breakdown of sub-categories. Here is a chart showing the ML types.



Let's look at them one by one!

8.2 ALGORITHMS USED IN SUPERVISED LEARNING

In supervised learning problems, the machines are provided with a labeled training dataset, which means that the values to be predicted are known for the algorithm from the very beginning.

The process is called 'supervised' learning since it is 'supervised' by the human-labeled output variable, i.e. the input variables (x) and an output variable (y) are known.

The objective of the problem is to find a suitable mapping function f from x to y. Then, when new inputs come in, we can apply the function f to predict the corresponding output.

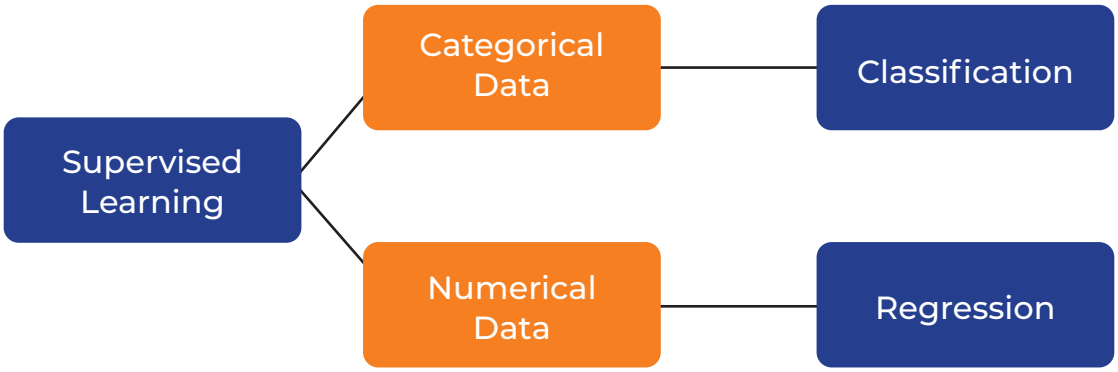
y = f(X) or Output = f(Input)

The machine learns and improves the function f through iterative optimization, where an initial guess is used to generate a sequence of improving approximate solutions for a class of problems. This process involves minimizing the difference between the estimated and the actual output.

We can divide supervised learning into regression and classification based on whether the output variable/target is Numerical or Categorical.

- **Numerical data** are represented by numbers such as body weight, the number of dogs.
- **Categorical data** are divided into categories such as gender (male/female), competition levels (low/medium/high).

Choice of the algorithms for Categorical or Numerical data, within Supervised Learning, is as follows:



8.2.1 CLASSIFICATION

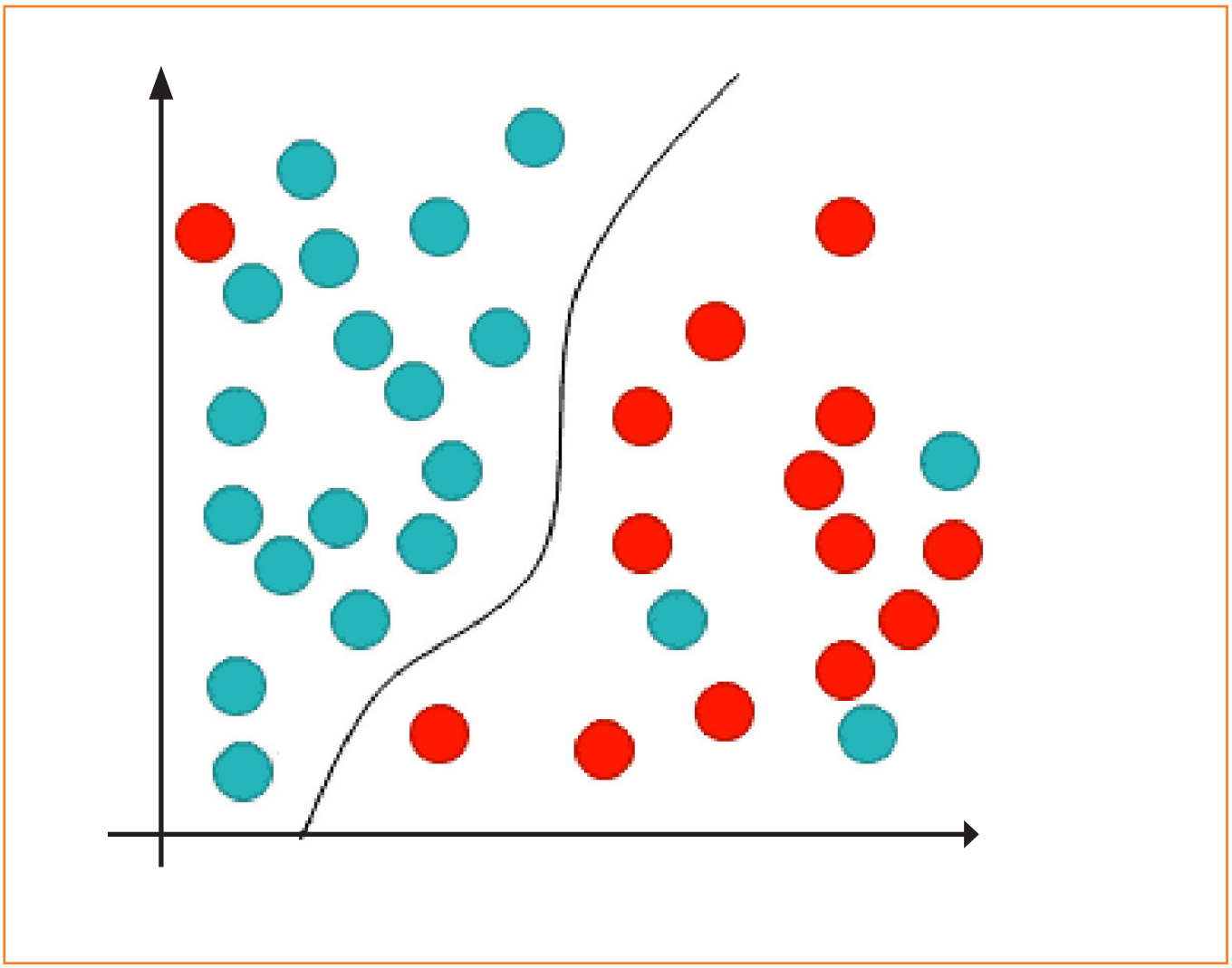
Classification algorithms are used when the target is a categorical variable.

Classification is the act of dividing the variable that we are trying to predict into output categories or classes and then predicting a class for a given input.

This is done based on the training observations with known labeled categories.

It is about learning the patterns among observations based on experience.

It falls into the category of Supervised Machine Learning, where the data set needs to have the classes, to begin with.



Thus, classification comes into play at any place where we need to predict an outcome, from a set number of fixed, predefined outcomes.

Below are some examples of classification applications:

- Classify customers of banks into different credit profiles based on existing customers' status.
  - Input variables include income, age, gender, location, expenses, etc.
  - Output variable is, whether a customer defaults on the payment to banks.
- Identify fraud among bank transactions based on fraudulent historical tags.
  - Input variables include transaction amount, date/time, customer history, merchants, etc.
  - Output variable is, whether a transaction is a fraud or not.
- Apply sentiment analysis on Yelp review data based on historical sentiment labels.
  - Input variables include the text of reviews.
  - Output variable is, whether the review is positive, negative, or neutral.

Further, classification can broadly be Binary, Multi-Class or Multi-Label

- **Binary (Two Class) classification** algorithms are used to divide the data into two categories.

  For example:

  - Are these symptoms suggesting cancer - Target Class is YES or NO
  - Is this a picture of a male or female - Target Class is MALE or FEMALE
- **Multiclass (multinomial) classification** algorithms divide the data into three or more categories. However, every input can be allocated to one class only.

  For example:

  - In which month do the majority of travelers purchase airline tickets?
  - What emotion is the person in this photo displaying?

- Multilabel classification where each sample is mapped to a set of target labels, i.e. more than one class.

  For example
  - A news article can be about sports, a person, and location at the same time

Types of Classification Algorithms

The various types of Classification Algorithms used in machine learning are as follows:

- Linear Classifiers
  - Logistic regression
  - Naive Bayes classifier
  - Fisher's linear discriminant
- Support vector machines
  - Least squares support vector machines
- Quadratic classifiers
- Kernel estimation
  - k-nearest neighbor
- Decision trees
  - Random forests
- Neural networks
- Learning vector quantization

EXPLORE

To learn more about classification algorithms, you can go through the following link:

https://analyticsindiamag.com/7-types-classification-algorithms/

https://dzone.com/articles/introduction-to-classification-algorithms

ACTIVITY

Try to train and evaluate a regression model by using 2 popular Machine Learning Platforms, AWS SageMaker and Microsoft Azure.

Using Microsoft Azure

Prerequisites:

- Scikit-learn
- Microsoft Azure subscription
- Visual Studio Codespace based on the MicrosoftDocs/ml-basics GitHub repository.

EXPLORE

To do that, visit the following link:

https://docs.microsoft.com/en-us/learn/modules/train-evaluate-classification-models/

Using AWS SageMaker

Prerequisites:

- Jupyter Notebook
- AWS SageMaker Studio Subscription

EXPLORE

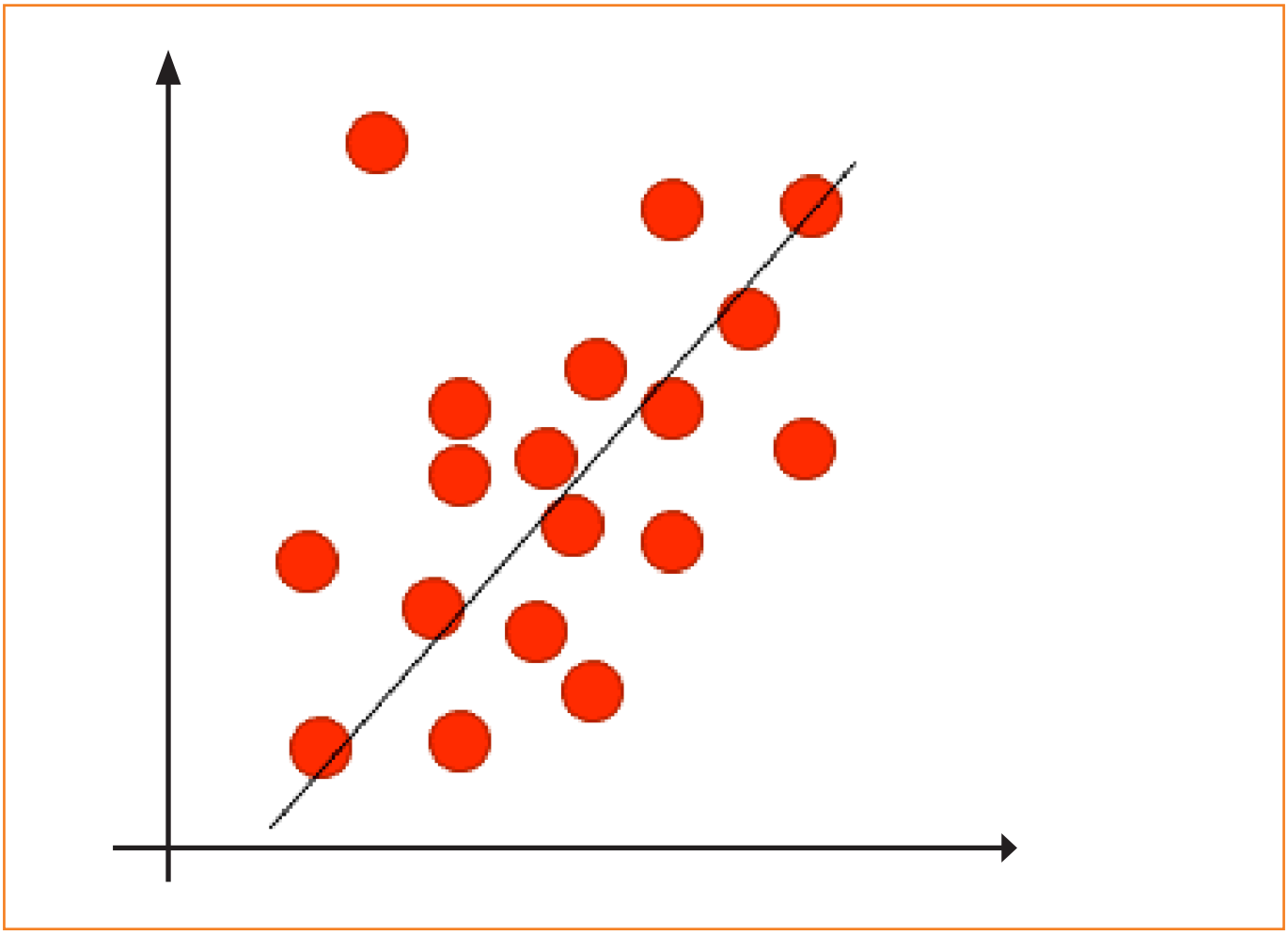
To do that, visit the following link:

(Structured Data Classification) :
https://aws.amazon.com/getting-started/hands-on/build-train-deploy-machine-learning-model-sagemaker/

(Image Classification) :
https://www.getstartedonsagemaker.com/workshop-studio-training/trainingbuilt-in/imgclassification/

8.2.2 REGRESSION

When the output variable is numerical, we have a regression problem.
It is about estimating the relationship between the target (dependent variable) based on at least one input variable (independent variable/feature). It is often used to predict or forecast based on experience.



Below are some examples of regression problems:

- Predict housing prices based on historical sales.
  - Input variables may include the size and age of the property, number of bathrooms, property tax, etc.
  - Output variable: the price of the house.

- Time series forecasting.
  - Input variables are the values of the time series (the historical data).
  - Output variable: the value of the time series at a later time.
- Customer Spending based on a sample of existing customers.
  - Input variables can have customer demographics, income, etc.
  - Output variable: how much a customer spends.

Machine Learning Regression Algorithms are being used in:

- Forecasting or Predictive analysis
- Optimization
- Error correction
- Economics – to predict supply, demand, consumption, inventory investment etc.
- Finance – predict the risk portfolio and factors that affect the customers, etc.

Types of Regression Models

The most common regression models used in Machine Learning are of following two types –

- **Linear Regression**
  - **Simple linear regression model** – This is the most basic regression model in which predictions are formed from a single, univariate feature of the data.
  - **Multiple linear regression model** – As the name implies, in this regression model, the predictions are formed from multiple features of the data.
- **Logistic Regression:** When the label is that of categorical or discrete nature, we use log odds ratio to optimize business problems such as – scoring customers and then predict those that are most likely to default on the loan payment, or predict higher response rates to a certain marketing strategy.

EXPLORE

To learn more about this, go through the following link:

https://www.youtube.com/watch?v=nk2CQITm_eo

https://www.youtube.com/watch?v=ylYKR4sgzI8

https://www.justintodata.com/linear-regression-machine-learning-python-tutorial/

https://www.justintodata.com/logistic-regression-for-machine-learning-tutorial/

ACTIVITY 2

Try to train and evaluate a regression model by using two popular Machine Learning Platforms, AWS SageMaker and Microsoft Azure.

Using Microsoft Azure

Prerequisites:

- Scikit-learn
- Microsoft Azure subscription
- Visual Studio Codespace based on the MicrosoftDocs/ml-basics GitHub repository.

EXPLORE

To do that, visit the following link:

https://docs.microsoft.com/en-us/learn/modules/train-evaluate-regression-models/

Using AWS SageMaker

Prerequisites:

- Jupyter Notebook
- AWS SageMaker Studio Subscription

EXPLORE

To do that, visit the following link:

https://www.getstartedonsagemaker.com/workshop-studio-training/trainingbuilt-in/housepricepred/

8.3 ALGORITHMS USED IN UNSUPERVISED LEARNING

The term 'unsupervised' refers to the fact that the algorithm is not guided like Supervised Learning Algorithm.

Machine Learning Algorithms that use unsupervised learning bring order to the dataset and make sense of data. They are used to group unstructured data according to their similarities and distinct patterns in the dataset.

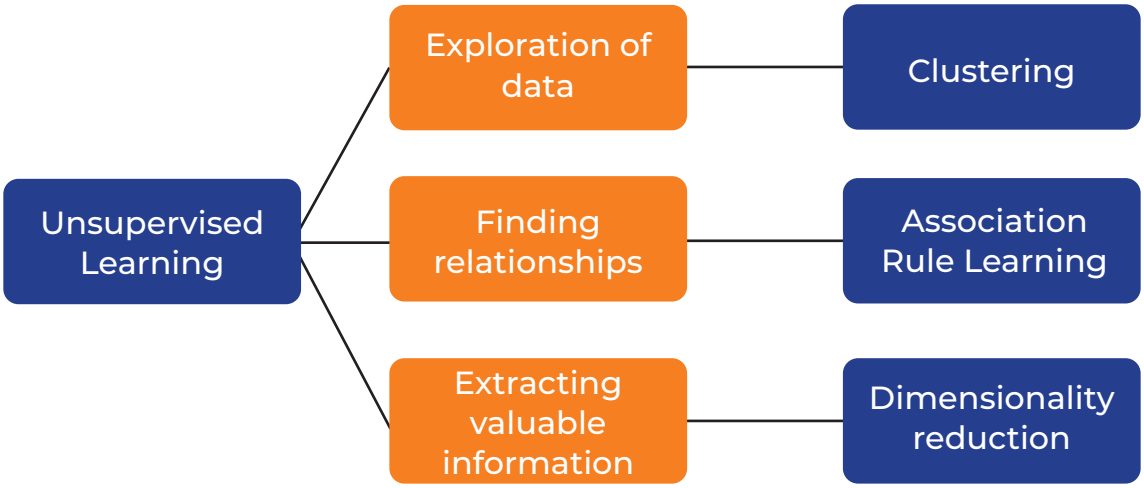
When we only have input variables (X) but no corresponding output variable, the training dataset is unlabeled. The machine can still extract structures within the dataset, such as grouping or clustering of data points.

This becomes an unsupervised learning problem. You can also think of it as 'Data mining' without any existing knowledge/labels.

The Unsupervised Machine Learning Algorithm is used to:

- Explore the structure of the information and detect distinct patterns
- Extract valuable insights
- Implement this into its operation to increase the efficiency of the decision-making process

To make that happen, unsupervised learning applies three major techniques:



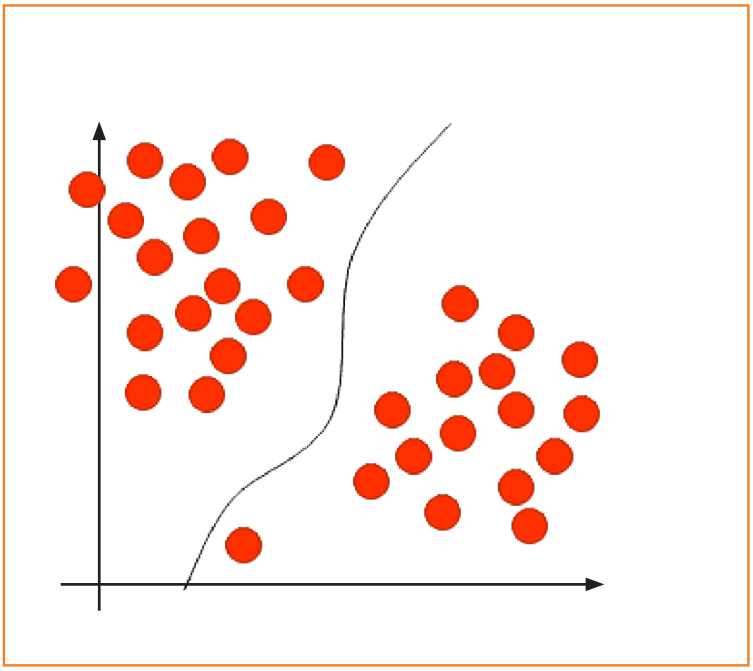
8.3.1 CLUSTERING

'Clustering' is the term used to describe the exploration of data, where the similar pieces of information are grouped. There are several steps to this process:

- Defining the credentials that form the requirement for each cluster. The credentials are then matched with the processed data and thus the clusters are formed.
- Breaking down the dataset into the specific groups (known as clusters) based on their common features.

Clustering techniques are simple yet effective. They require some intense work yet can often give us some valuable insight into the data.

When the training dataset doesn't have labeled categories, the clustering algorithm helps grouping observations into categories based on some measure of inherent similarity or distance.



In other words, clustering is about separating a set of data objects into clusters. The observations in the same group are more like each other than the observations in different groups.

It is used to recognize patterns among clusters in different fields such as biology, marketing, etc.

For example, we can segment customers into groups such that customers in the same group behave similarly.

Then we can target specific groups with customized marketing campaigns to generate more sales.

Types of Clustering Algorithms

Some common types of Clustering Algorithms that are used in machine learning are:

- k-Means Clustering Algorithm
- Hierarchical Clustering Algorithm
- Fuzzy C Means Algorithm
- Mean Shift Clustering
- Density-based Spatial Clustering
- Maximization Clustering

EXPLORE

To learn more about these, you can go through the following links:

https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/

https://www.youtube.com/watch?v=7xHsRkOdVwo

https://www.youtube.com/watch?v=4b5d3muPQmA

ACTIVITY 3

Try to train and evaluate a clustering model by using two popular Machine Learning Platforms, AWS SageMaker and Microsoft Azure.

Using Microsoft Azure

Prerequisites:

- Scikit-learn
- Microsoft Azure subscription
- Visual Studio Codespace based on the MicrosoftDocs/ml-basics GitHub repository

EXPLORE

To do that, visit the following link:

https://docs.microsoft.com/en-us/learn/modules/train-evaluate-cluster-models/

Using AWS SageMaker

Prerequisites:

- Jupyter Notebook
- AWS SageMaker Studio Subscription

EXPLORE

To do that, visit the following link:

https://aws.amazon.com/blogs/machine-learning/k-means-clustering-with-amazon-sagemaker/

8.3.2 ASSOCIATION RULE LEARNING

Association rules are used to analyze transactions to find relationships among different variables. It is mostly used in market basket analysis.

For Example:

What other products are likely to be purchased when a customer buys both laptop and a mouse?

From the algorithms, we may find a rule that there's a 90% probability the customer will also buy a laptop cover.

We can then design marketing strategies to make it convenient for the customers to shop.

EXPLORE

To learn more about this, go through the following links:

https://www.tutorialandexample.com/association-rule-learning-algorithm/

https://www.aionlinecourse.com/tutorial/machine-learning/association-rule-learning

8.3.3 DIMENSIONALITY REDUCTION

Dimensionality reduction is the transformation of data from a high dimensional space (read as - high feature space) to a low dimensional space (read as - low feature space) so that the new low dimensional space retails most of the intrinsic properties of the original data.

When there are a large number of features, it gets harder to visualize and work on the training set. However, sometimes, certain features are correlated. I.e is if one feature has an increased value, the corelated feature with also increase or decrease proportionately. This make the corelated feature unnecessary. This is where dimensionality reduction algorithms are used.

It is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables.

There are several advantages of reducing the features:

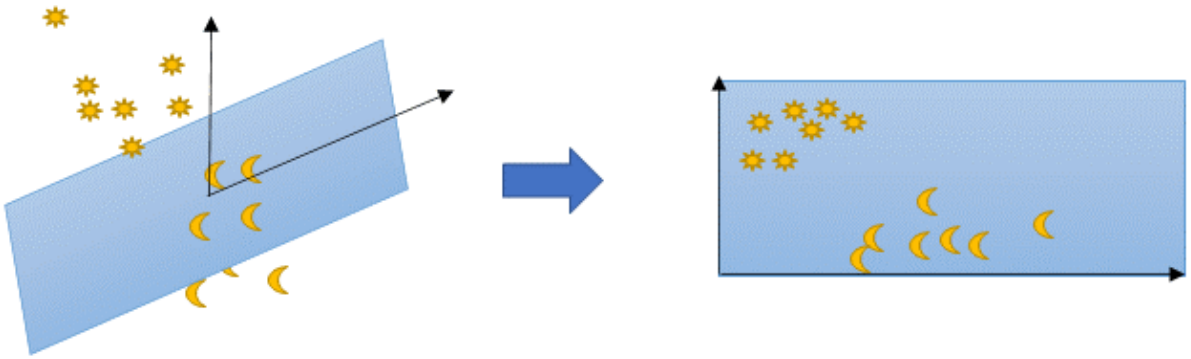
- Less processing time/storage space is needed.
- The potential of removing multi-collinearity.
- It is easier to visualize the data.

It can be divided into feature selection and feature extraction.

- In Feature selection we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem.

- Feature extraction reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.

For example, reducing 3 dimensions/variables to 2 dimensions



EXPLORE

To learn more about this, go through the following links:

https://courses.analyticsvidhya.com/courses/dimensionality-reduction-for-machine-learning

https://www.geeksforgeeks.org/dimensionality-reduction/

8.4 ALGORITHMS USED IN REINFORCEMENT LEARNING

Reinforcement learning develops a self-sustained system that, through contiguous sequences of tries and fails, improves itself based on labeled data and its interactions with the incoming data.

Reinforcement learning algorithms can be taught to exhibit one or both types of the following experimentation learning styles:

- Exploration
- Exploitation

Exploration is the process of the algorithm pushing its learning boundaries, assuming more risk, to optimize towards a long-run learning goal. In an example of Tic-Tac-Toe, this could take the form of running simulations that assume more risk, or purposefully place pieces unconventionally to learn the outcome of a given move. Armed with a greater possibility of manoeuvres, the algorithm becomes a much more fierce opponent to match against.

Exploitation is the process of the algorithm leveraging information it already knows to perform well in an environment with short term optimization goals in mind. Back to the Tic-Tac-Toe example, the algorithm would perform the same moves it knew to produce a nominal probability of winning. However, in this learning mode, the ML algorithm will not develop beyond elementary sophistication.

It is usually a hybrid of exploration and exploitation styles that produces the optimal algorithm. In the same way that a human must branch out of comfort zones to increase their breadth of learning, but at the same time cultivate their given resources to increase their depth of learning.

In the hybrid exploration/exploitation technique:

- ☞ an action takes place
- ☞ the consequences are observed
- ☞ the next action considers the results of the first action.

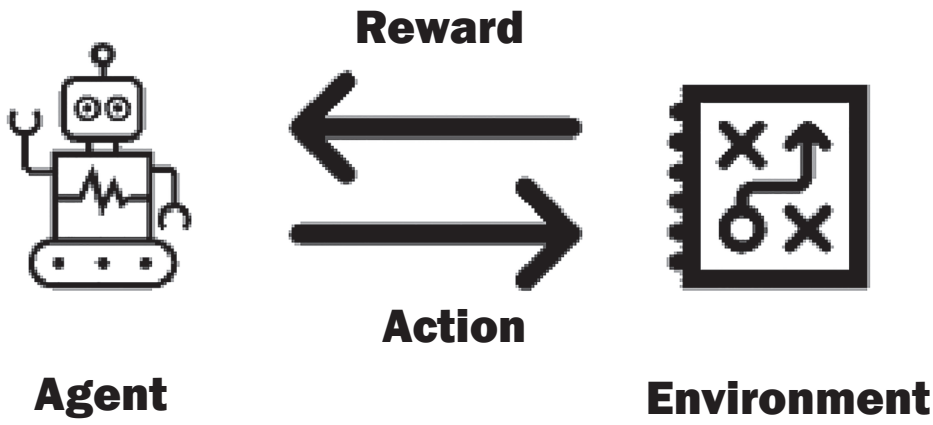
Further, reward signals occur upon performing specific tasks serving as a navigation tool for the reinforcement algorithms. They give it an understanding of right and wrong course of action.

Two main types of reward signals are:

- **Positive reward signal** encourages continuing performance a particular se quence of action
- **Negative reward signal** penalizes for performing certain activities and urges to correct the algorithm to stop getting penalties.

The training dataset is a set of rules/environment, and we want the machine/agent to maximize the rewards within this environment, allowing the machines to learn by trial and error.

This is often expected in games.



With feedback/labels from the environment, reinforcement learning is somewhat similar to supervised learning.

Most common reinforcement learning algorithms used in machine learning include:

- Q-Learning;
- Markov Decision Process
- Temporal Difference (TD);
- Monte-Carlo Tree Search (MCTS);
- Asynchronous Actor-Critic Agents (A3C).

EXPLORE

To learn more about this, you can go through the following links:

https://www.guru99.com/reinforcement-learning-tutorial.html#4

https://www.geeksforgeeks.org/ml-monte-carlo-tree-search-mcts/

https://www.youtube.com/watch?v=-fVu_bj8Os8

ACTIVITY 4

Using the learnings from this Unit as well as the previous units on Machine Learning, NLP, Computer Vision and Python, try to do a project by using the MLaaS platform, Amazom Sagemaker.

To do that, go through the steps mentioned in the following document:

AWS Production Platform-ao-Step-Up-Module.doc

EXPLORE

You can also refer to multiple other projects references at the following link :

https://www.getstartedonsagemaker.com/resources/sampleproblems/

SUMMARY

Algorithms are step-by-step instructions to solve a problem, created along with rules that decide which path to take when a decision is faced.

In machine learning, the algorithm is a program that can alter its own parameters based on given responses on the past predictions of the data set. ML algorithms can be divided into three types:

- Supervised Learning: Classification and Regression algorithms.
- Unsupervised Learning – Association Rule, Clustering and Dimensional Reduction Algorithms
- Reinforcement Learning

We can divide supervised learning into regression and classification based on whether the output variable/target is Numerical (Classification) or Categorical (Regression).

Classification algorithms are used when the target is a categorical variable. Classification is the act of dividing the variable that we are trying to predict into output categories or classes and then predicting a class for a given input. Classification can broadly be Binary, Multi-Class or Multi-Label.

When the output variable is numerical, we have a regression problem. It is about estimating the relationship between the target (dependent variable) based on at least one input variable (independent variable/feature). It is often used to predict or forecast based on experience. The most common regression models used in Machine Learning are of two types, 1) Linear Regression (Simple linear regression model and Multiple linear regression model); 2) Logistic Regression.

Unsupervised learning applies three major techniques – 1) Exploration of data (Clustering); 2) Finding relationships (Association Rule Learning); Extracting valuable information (Dimensionality reduction).

Clustering is about separating a set of data objects into clusters. The observations in the same group are more like each other than the observations in different groups. It is used to recognize patterns among clusters in different fields such as biology, marketing, etc.

Association rules are used to analyze transactions to find relationships among different variables. It is mostly used in market basket analysis.

Dimensionality reduction is the transformation of data from a high dimensional space (read as - high feature space) to a low dimensional space (read as - low feature space) so that the new low dimensional space retails most of the intrinsic properties of the original data.

Reinforcement learning develops a self-sustained system that, through contiguous sequences of tries and fails, improves itself based on labeled data and its interactions with the incoming data.

Reinforcement learning algorithms can be taught to exhibit one or both types of experimentation learning styles i.e Exploration and/or Exploitation. In the hybrid exploration/exploitation technique an action takes place, the consequences are observed and the next action considers the results of the first action.

ACKNOWLEDGEMENT

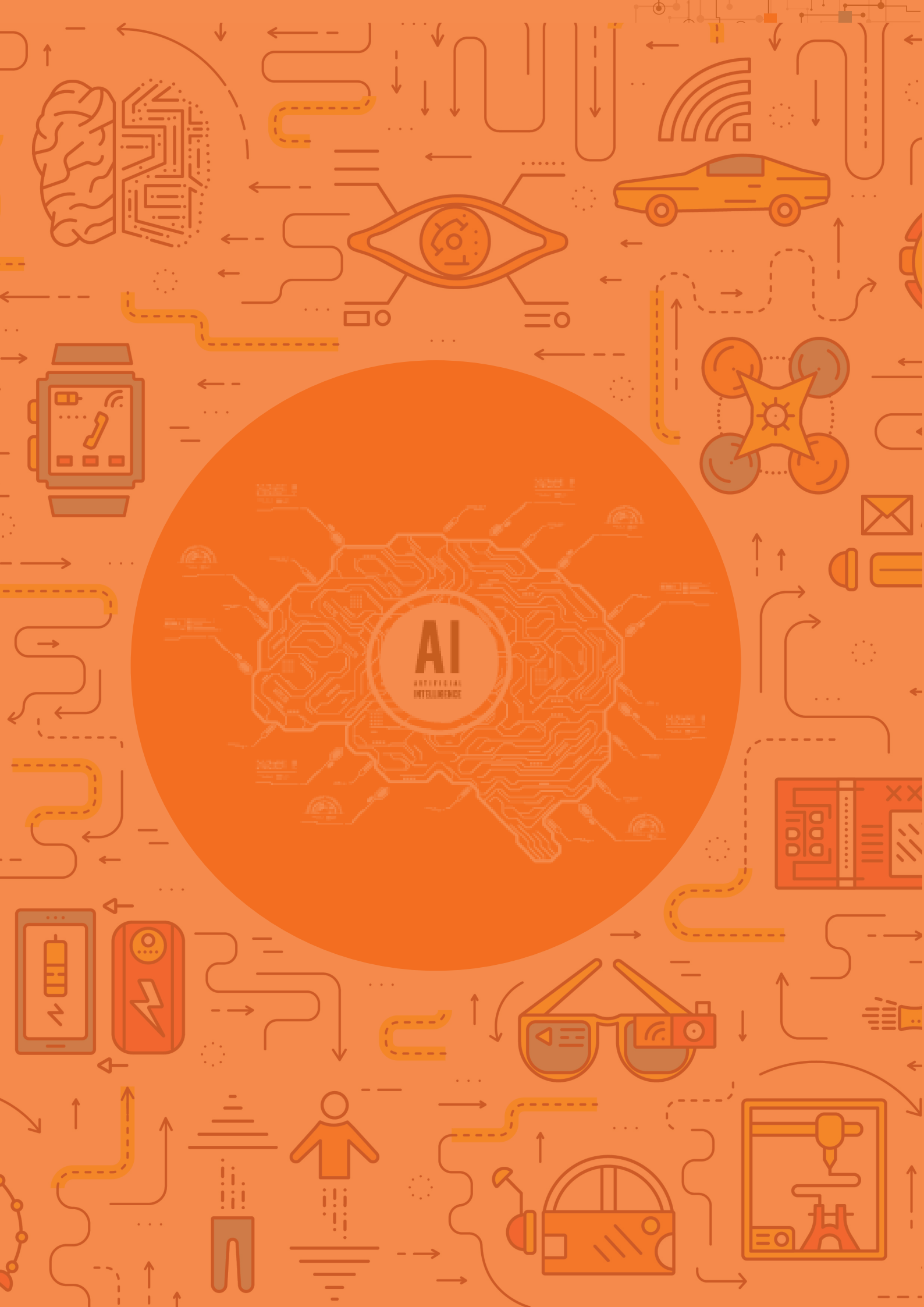
We thank all contributors and facilitators who have supported the development of this Module. We thank the **NASSCOM** team for facilitating the Industry-Academia workgroup in this endeavor.

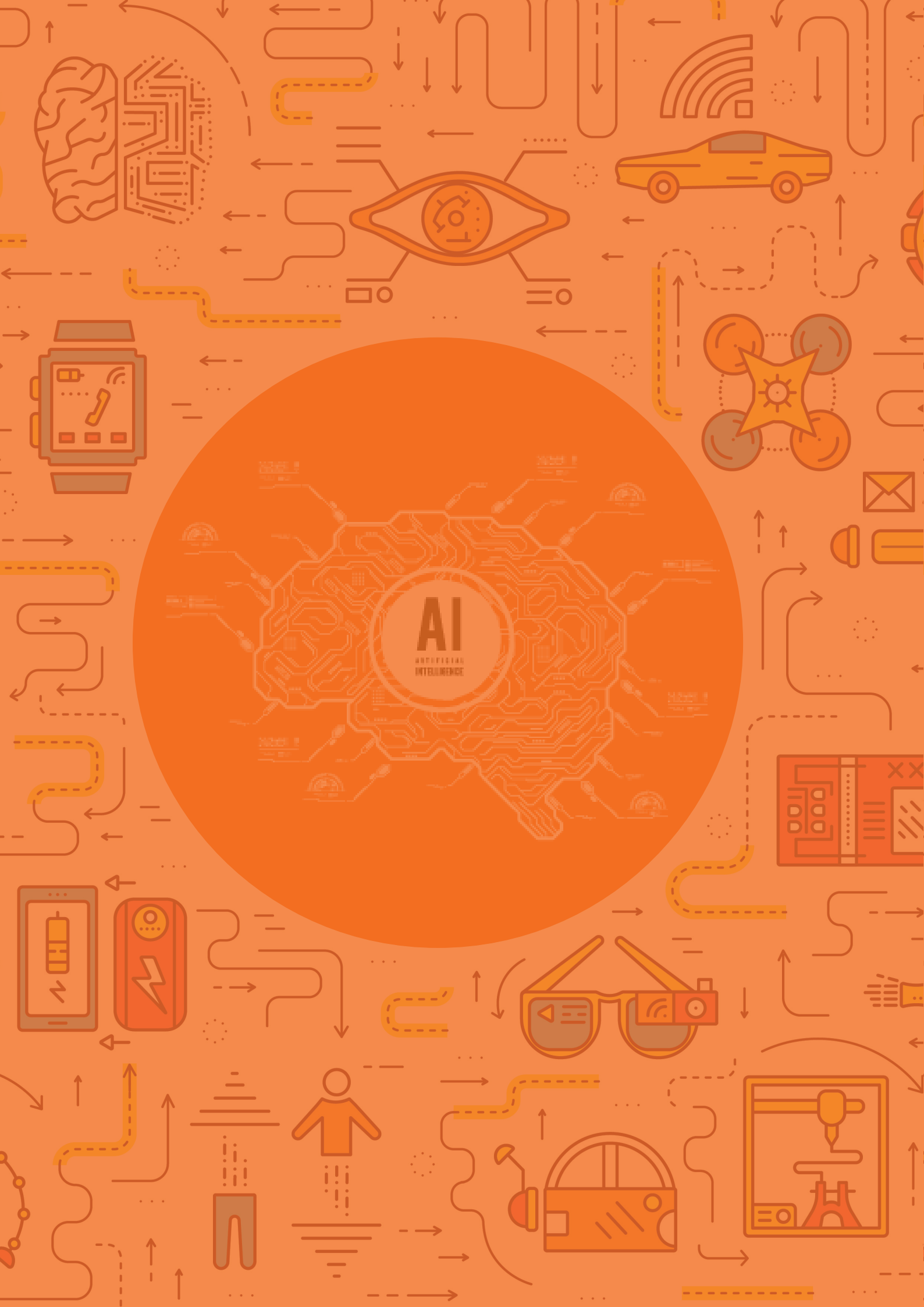
Companies Represented

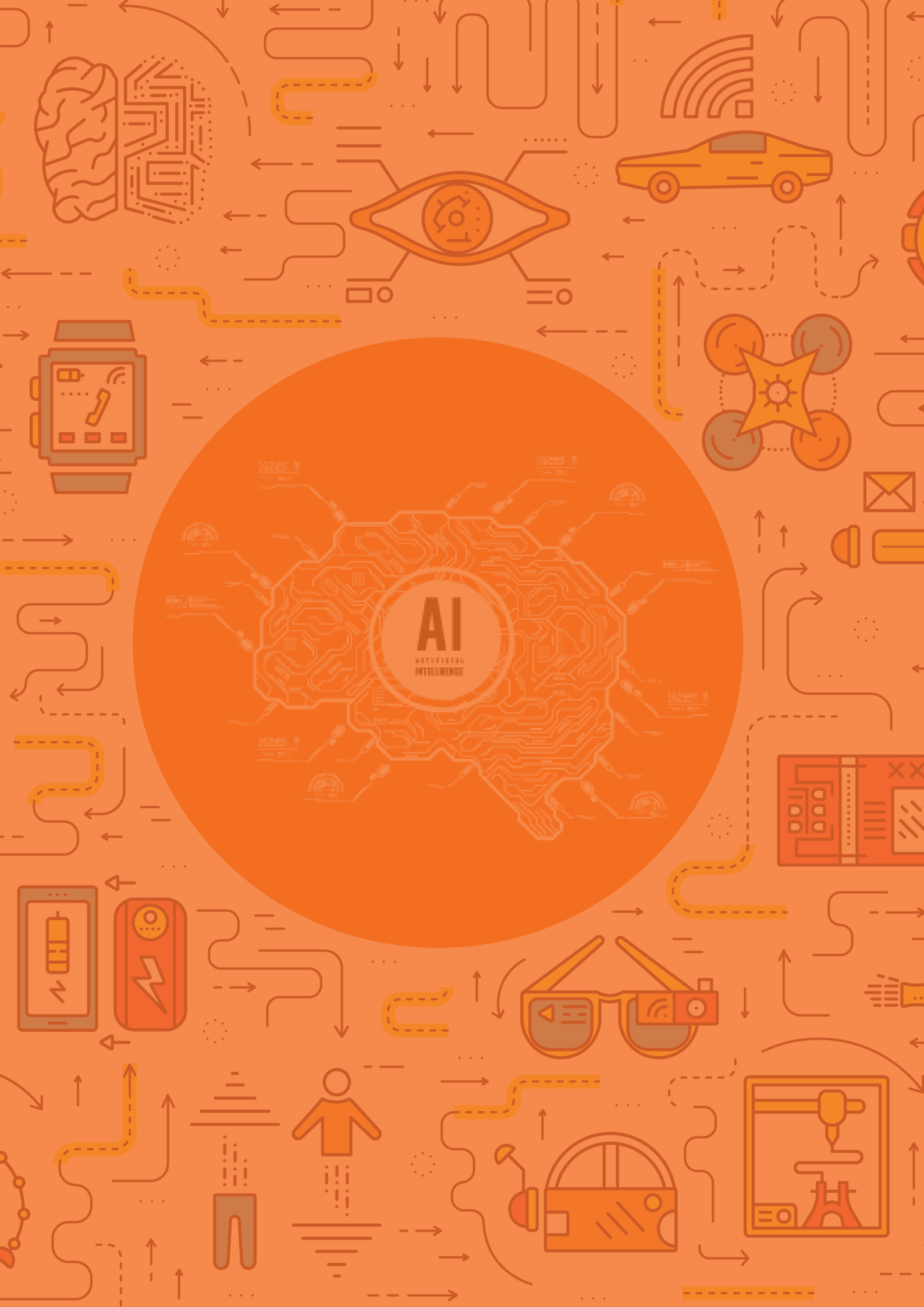
Workgroup Members

01. **Amit Nevatia**	Educate Program Lead, India/South Asia	Amazon Internet Services Pvt Ltd
02. **Arokiasamy Stephen Raj**	Head of Data Science and AI	Robert Bosch Engineering & Business Solutions Pvt. Ltd.
03. **Ashish Goyal**	SME, Computer Vision, AI Practice	Accenture
04. **Ashish Mital**	Solutions Architect, India/South Asia	Amazon Internet Services Pvt Ltd
05. **Ashutosh Raina**	Doctoral Research Scholar - Ed Tech - IDP	IIT-Bombay
06. **Asna Siddiqui**	Program Lead - INDIAai	NASSCOM
07. **Bhushan Jagyasi**	Lead, Speech Processing, AI Practice	Accenture
08. **Bolaka Mukherjee**	SME, Computer Vision, AI Practice	Accenture
09. **Deepa Shah**	SME, Computer Vision, AI Practice	Accenture
10. **Dhrupal R Shah**	Chief Executive Officer (CEO)	STEMpedia
11. **Ganesh Vaitheeswaran**	Senior Architect - Artificial Intelligence	SAP Labs
12. **Ganeshan Rajan**	Senior Data and Information Architect	GE Healthcare
13. **Gopali Contractor**	Managing Director, AI Practice	Accenture
14. **Jassi Dimple**	Director	PROGILENCE Capability Development Pvt. Ltd.
15. **Jennifer Christian**	Assistant Program Manager	STEMpedia
16. **Jibu Elias**	Content & Research Lead, INDIAai	NASSCOM
17. **Kanishka Agiwal**	Head, AI/ML & Analytics, India/South Asia	Amazon Internet Services Pvt Ltd
18. **Kishore Garg**	Program Manager - Education	Microsoft Corporation India Pvt Ltd.
19. **Mukhvinder Singh Bains**	Lead Learning and Leadership Development for India	Accenture
20. **Naba Suroor**	Young Professional	Atal Innovation Mission, Niti Aayog

21. **Omjiwan Gupta**	Director Education	Microsoft Corporation India Pvt Ltd
22. **Pankaj Kumar Verma**	Chief Technology Officer (CTO)	STEMpedia
23. **Prakash Goteti Principal**	Consultant, AI and Analytics	TechMahindra
24. **Rahul Lodhe**	Senior Director - Artificial Intelligence	SAP Labs
25. **Ramswaroop Mishra**	Data Analytics & AI Competency Head	WIPRO
26. **Ravinder Singh Bidhan**	Education Program Manager	STEMpedia
27. **Ronak Jogeshwar**	Innovation Lead	Atal Innovation Mission, Niti Aayog
28. **S Manish Singh**	Managing Director	PROGILENCE Capability Development Pvt. Ltd.
29. **Sachin Punyani**	Technical Business Development – AI/ML, India South Asia	Amazon Internet Services Pvt Ltd
30. **Saket Saurabh**	Business Head	Tevatron Technologies Pvt. Ltd.
31. **Sridhar Iyer**	Convener Ed Tech - IDP	IIT-Bombay
32. **Syama Sundar Cheruvu**	SME, Predictive Analytics and NLP, AI Practice	Accenture
33. **Vinnie Jauhari**	Director Education Industry Expert	Microsoft Corporation India Pvt Ltd.
34. **Vipin Bhasin**	Lead, Computer Vision, AI Practice	Accenture
35. **Waris Candra**	Head of Asia Pacific	Micro:bit Educational Foundation
36. **Yudhisther Yadav**	Lead Strategic Initiatives – FutureSkills	NASSCOM


AI
ARTIFICIAL
INTELLIGENCE


AI


AI
ARTIFICIAL
INTELLIGENCE


HAPPY
TINKERING