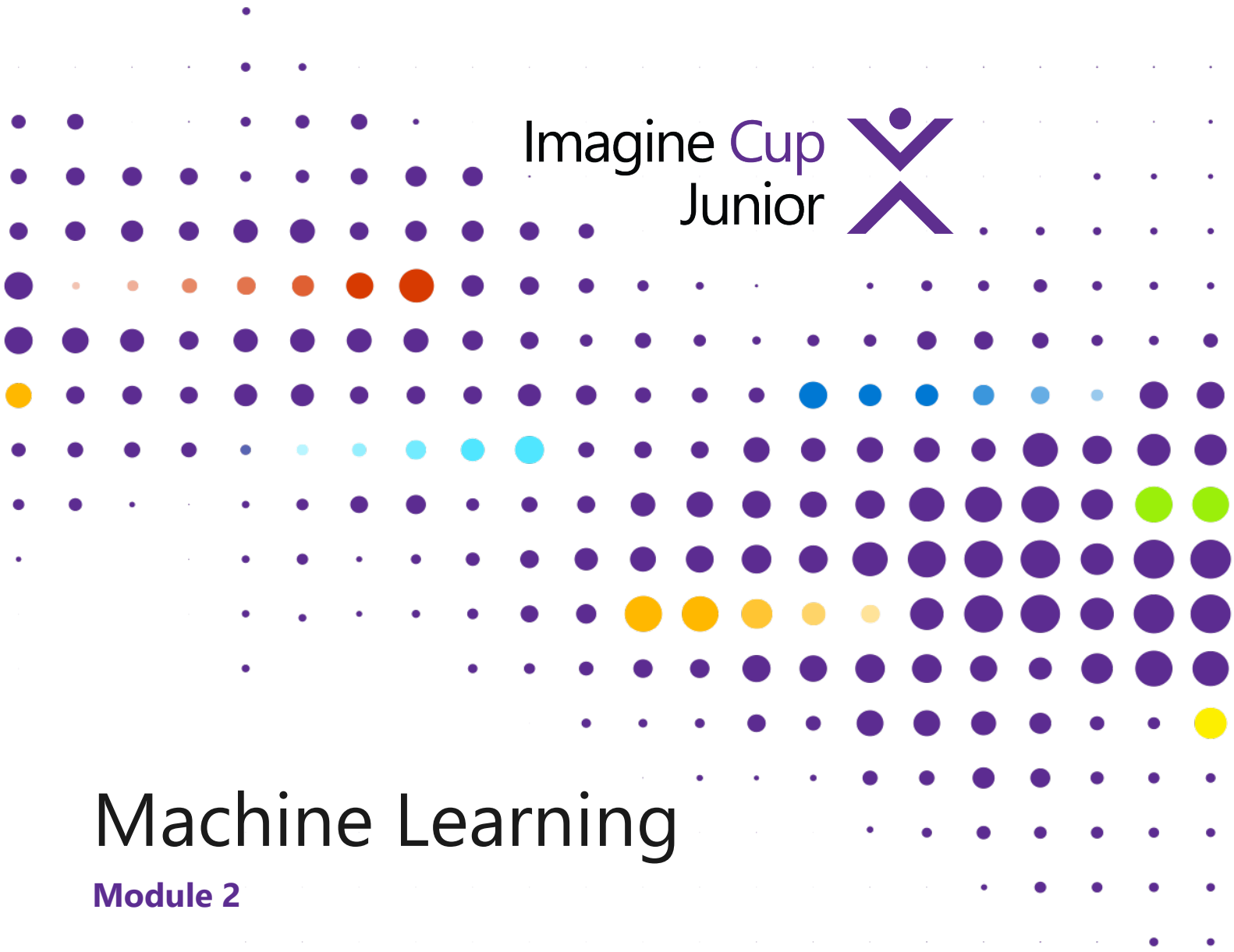


Machine Learning

Module 2

Microsoft

Table of Contents

Learning Objectives **4**
Machine Learning – The foundation of Artificial Intelligence **5**
Machine Learning 6
The Need for Machine Learning 8
Understanding Data and Datasets **10**
Data and Its Utility 10
Use of Data in Machine Learning 12
Different Types of Datasets 12
Sentiment Analysis 14
How machines learn **20**
Machine Learning 22
Major machine learning methods 22
Supervised Learning 23
Data Labeling 24
Machine Learning and classification 26
Semi-Supervised learning 30
Classroom Activity 30
Supervised Learning in Minecraft 31
Unsupervised Learning 32
Reinforcement learning 33
Assessments Questions **34**
Questions to consider **36**
Some Practical Assignments/Lab Work **38**
Practical Assignments **42**
Further Reading **43**
Reference Links **44**
Glossary **46**

Disclaimer:

The Imagine Cup Junior guides and lesson materials are created by Microsoft and our partners and intended to be for guidance only to support with the Imagine Cup Junior Challenge. For the latest on Microsoft AI please visit https://www.microsoft.com/en-us/ai

Learning Objectives

Through this module, students will get an overview of machine learning and understand how it provides the foundation of AI. Students should be able to understand the basics of machine learning and use the concepts as applied to their daily life.

At the conclusion of the module, students should be able to:

- Understand the basics of machine learning.
- Comprehend the basics of using datasets and working with data.
- Understand the machine learning approach to problem-solving, and devise solutions to problems.
- Comprehend the latest design-related perspectives, ideas, concepts, and solutions.
- Understand the importance of data and ways to protect it.
- Execute projects using design thinking principles.
- Understand and analyze data related problems.
- Comprehend the basics of creating a BOT and the related working framework.
- Understand the various challenges of creating a BOT.
- Appreciate the similarities and differences between a machine-driven BOT and a human.
- Understand the concept of cluster algorithms and apply the principle of 'clustering' on data.

Machine Learning – The foundation of Artificial Intelligence

Often used interchangeably with artificial intelligence, machine learning, however, has a different meaning. It is the 'learning' that the machine derives from its experience in processing data. The primary objective of machine learning is to ensure that the machine learns from the data. In other words, machine learning is an application of artificial intelligence (AI) that provides computer systems with the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning as a science focuses on the development of computer programs that can access data and use it to learn for themselves. This is sometimes known as heuristic programming.

Machine learning can also be defined as the study of computer-based algorithms designed to automatically improve the experience through acquired learning. Machines are created with a built-in capability to read and understand human language to comprehend their surroundings and make as many accurate predictions as they can. They can also perform simultaneous real-time assessments of predictions and adapt according to their environment. When a user wants to search a topic, the search engine shows up the most frequently searched related 'search topics'. The search engine looks at past clicks from people around the world in order to understand the pages that are more relevant for those searches than others. It then serves those results a list with the most relevant being at the top. It should be noted that such an exercise is impossible to be performed by humans in the time frame of a few seconds.

The machine learns how to handle search requests and generate a set of instructions to create the expected outcome. Hence, machine learning can also be understood as a set of procedures, which deals with huge amounts of data smartly (using algorithms or a set of logical rules) to derive results.

Machine Learning

Whilst you are all, by now, familiar with artificial intelligence (AI), machine learning is a specific subset of AI which simply trains a machine on how to learn. It is an application of artificial intelligence that provides computer systems with the ability to spontaneously learn and improve based on its experience without being explicitly programmed.

Put simply, machine learning is an application of A.I. that provides computer systems with the ability to automatically learn and improve from experience.

The process of machine learning begins with analyzing observations (data) such as examples, direct experiences, or instructions, and looks for patterns in the data. Based on this analysis and the cumulative data it was provided, it learns to make better decisions in the future. The primary goal is to allow the machine to learn automatically without human intervention or assistance and adjust actions accordingly.

Machine Learning takes place over two phases. In phase one the machine takes data from input devices and pre-process it into a form that it can understand.

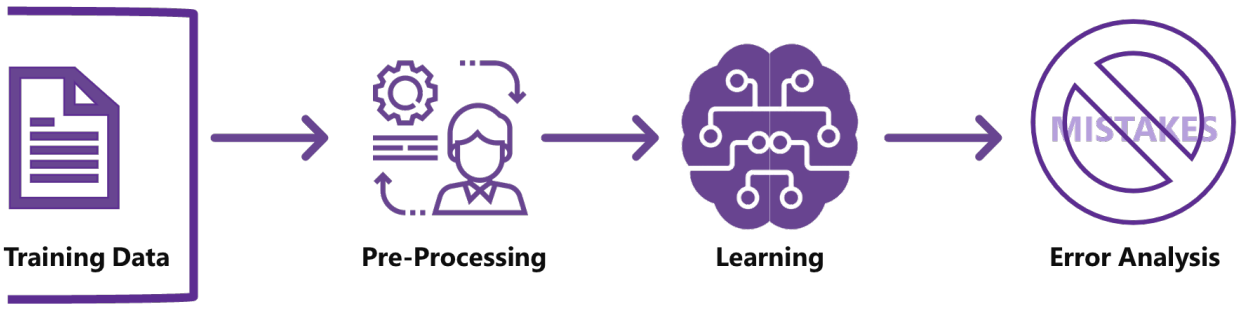


Fig 2.1: Phase 1 of Machine Learning

In phase two machine learning begins. Having pre-processed the data, an analysis takes place and using previous examples, direct experiences, instructions, etc., it looks for patterns which it uses to make better decisions the next time it is used.
The learning is done by predicting the outcome using various models and testing if these outcomes are correct.

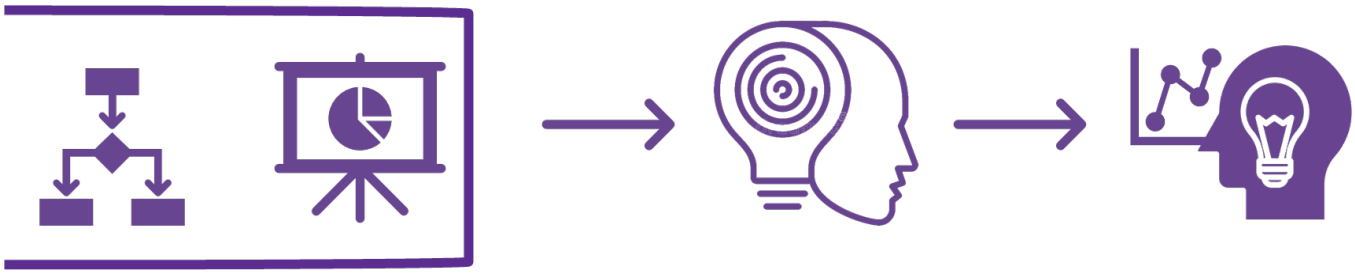


Fig 2.2: Phase 2 of Machine Learning

The Need for Machine Learning

The reason behind machine learning is to automate mundane tasks to the extent that the machine can learn, think and make smart decisions on its own. It is also to minimize human interference and thus bias in various scenarios. The need for machine learning is to complete tasks that are too complex for humans to code computers for directly. Some tasks are so complex that it is impractical, if not impossible, for humans to cater for all the nuances and code for every single instance separately. Instead, a large amount of data is provided to a machine learning algorithm and the algorithm computes the result by exploring the data and constructing a model that will achieve the desired outcome.

Machine learning is also useful for finding relationships between things, especially in exceptionally large datasets which are too big for humans to process efficiently. Its uses here are in object recognition, marketing analytics, analyzing scientific data in labs, and numerous other applications that involve large amounts of data needing to be analyzed.

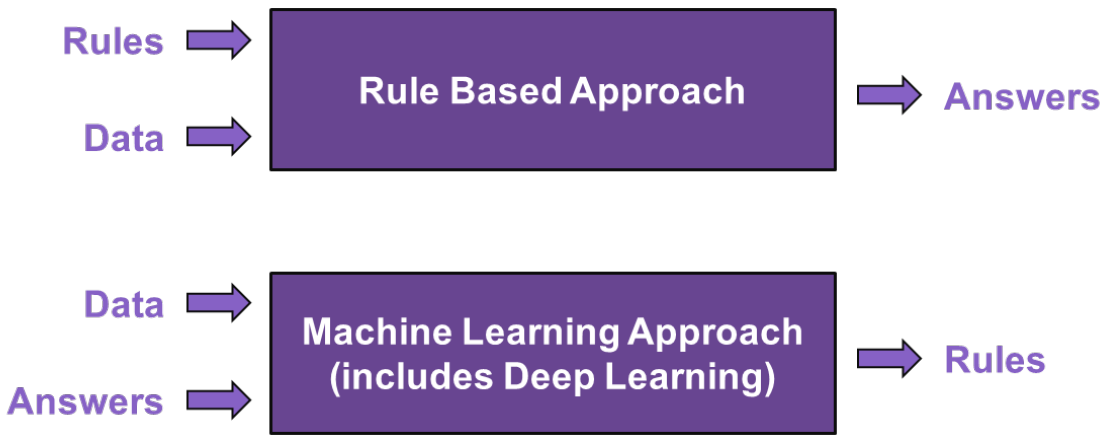


Fig 2.3: Approach towards Traditional Programming and Machine Learning

The key difference between traditional programming and machine learning is that in traditional programming, the data and the rules are run on the computer to produce an answers (output). However, in machine learning, the data and the answers are fed into a computer to create the rules for the program. This program can be then used in the same way as one created by traditional programming.

A few examples of Machine Learning in our day-to-day life are:

- Cortana
- Refined search engine results (as represented in fig 2.4)

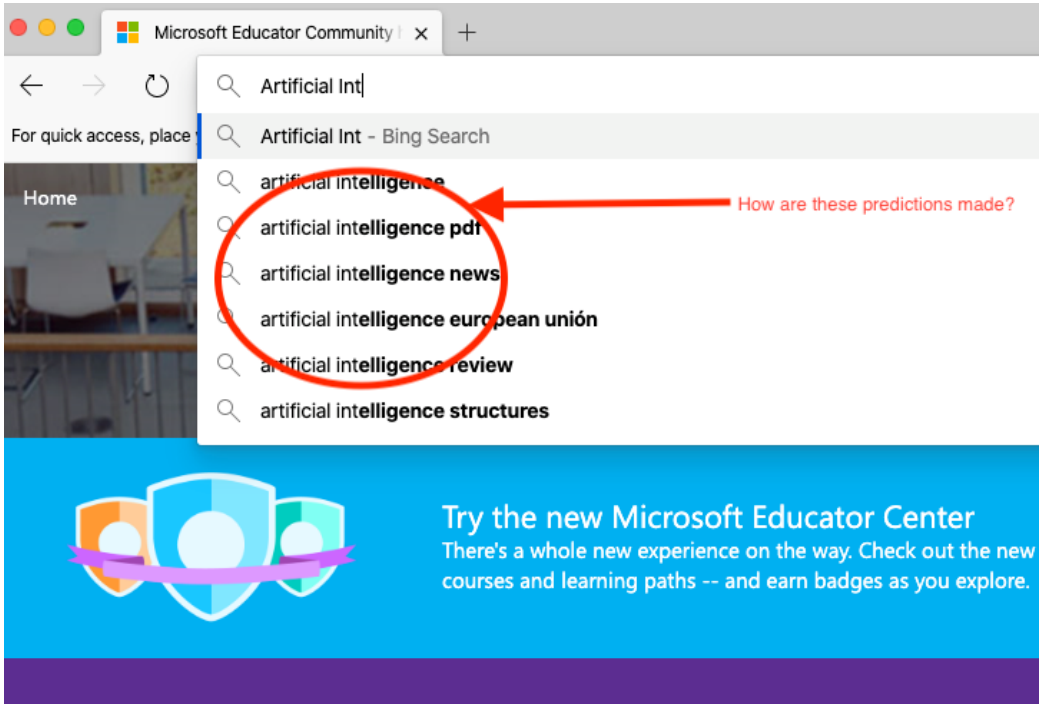
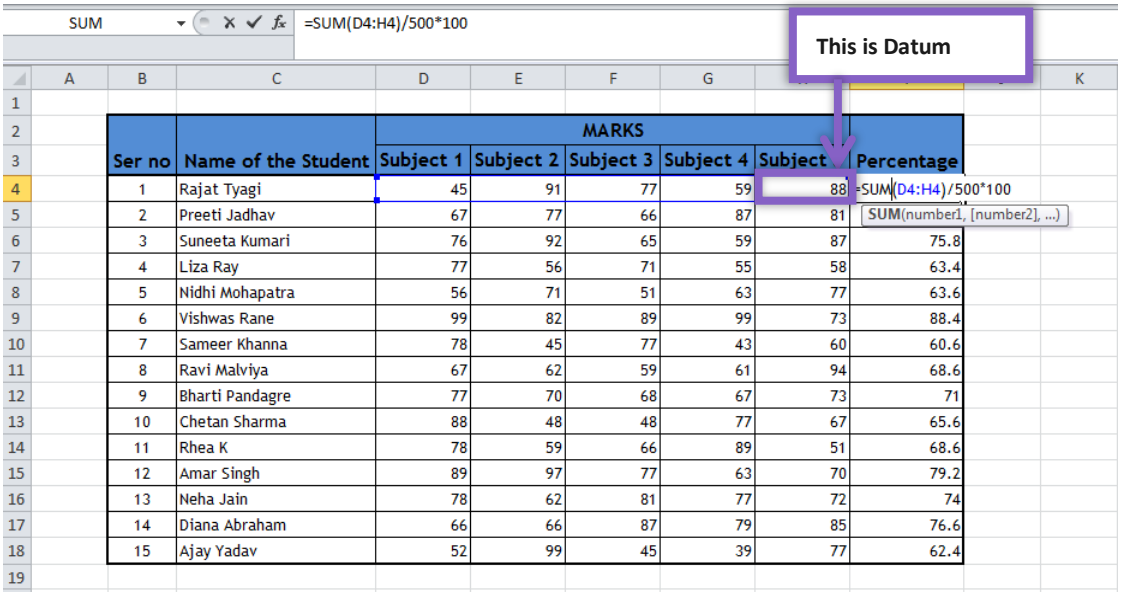


Fig 2.4: Search Engine Result Refining, a typical example of machine learning

Understanding Data and Datasets

Data and Its Utility



Ser no	Name of the Student	MARKS					Percentage
		Subject 1	Subject 2	Subject 3	Subject 4	Subject	
1	Rajat Tyagi	45	91	77	59	88	=SUM(D4:H4)/500*100
2	Preeti Jadhav	67	77	66	87	81	SUM(number1, [number2], ...)
3	Suneeta Kumari	76	92	65	59	87	75.8
4	Liza Ray	77	56	71	55	58	63.4
5	Nidhi Mohapatra	56	71	51	63	77	63.6
6	Vishwas Rane	99	82	89	99	73	88.4
7	Sameer Khanna	78	45	77	43	60	60.6
8	Ravi Malviya	67	62	59	61	94	68.6
9	Bharti Pandagre	77	70	68	67	73	71
10	Chetan Sharma	88	48	48	77	67	65.6
11	Rhea K	78	59	66	89	51	68.6
12	Amar Singh	89	97	77	63	70	79.2
13	Neha Jain	78	62	81	77	72	74
14	Diana Abraham	66	66	87	79	85	76.6
15	Ajay Yadav	52	99	45	39	77	62.4

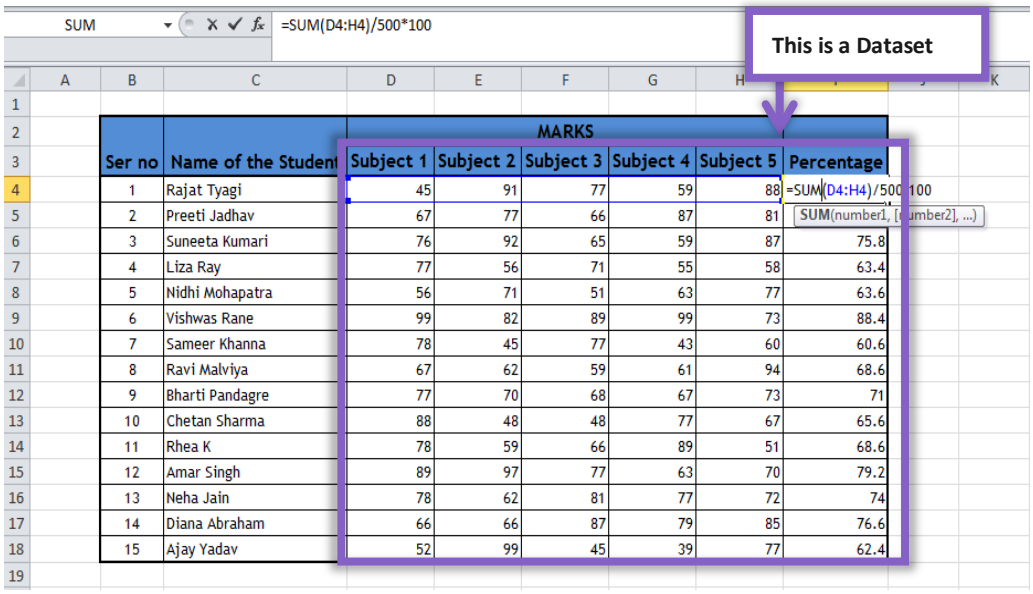
Fig 2.5: Data

Data can be defined as the collection of facts, numbers or other information that are used either for reference, or analysis. The singular form of Data is Datum.

In Figure 2.5, the task is to compute the cumulative percentage of each student's marks. The percentage obtained in an exam by a student is calculated by the sum of all the marks obtained in different subjects divided by the number of subjects. Therefore, the marks are important for calculating the percentage, and to arrive at the result it is important to take into account the marks obtained by each student in each subject.

Dataset

A dataset is defined as a collection, or group, of data where every column denotes a particular variable and each row relates to a specific member of that dataset.



Ser no	Name of the Student	MARKS					
		Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Percentage
1	Rajat Tyagi	45	91	77	59	88	=SUM(D4:H4)/500*100
2	Preeti Jadhav	67	77	66	87	81	
3	Suneeta Kumari	76	92	65	59	87	75.8
4	Liza Ray	77	56	71	55	58	63.4
5	Nidhi Mohapatra	56	71	51	63	77	63.6
6	Vishwas Rane	99	82	89	99	73	88.4
7	Sameer Khanna	78	45	77	43	60	60.6
8	Ravi Malviya	67	62	59	61	94	68.6
9	Bharti Pandagre	77	70	68	67	73	71
10	Chetan Sharma	88	48	48	77	67	65.6
11	Rhea K	78	59	66	89	51	68.6
12	Amar Singh	89	97	77	63	70	79.2
13	Neha Jain	78	62	81	77	72	74
14	Diana Abraham	66	66	87	79	85	76.6
15	Ajay Yadav	52	99	45	39	77	62.4

Fig 2.6: A Collection of Data is called a Dataset

Datasets are needed to create the learning algorithm the machine uses in a particular context of Artificial Intelligence. The method adopted by machines to learn is to use automatic data analysis for building concepts. The whole model of machine learning is built on the premise that systems can be programmed to learn from the data they receive as input. This is done through the identification of patterns to make informed decisions with minimal human intervention.

The entire process begins with the input of data into the machine. Data can be accumulated either through datasets or real-time data through physical sensors such as cameras, temperature sensors, microphones etc. The machine is equipped to understand and analyze patterns and then perform certain tasks using those patterns as references. The machine works iteratively, which as the model is exposed to various kinds of data makes it capable of adapting itself independently.

The machine learning, like humans, comes from earlier results on similar scenarios and thus the application learning improves the more it is used. This leads to more trustworthy results, but remember, the trustworthiness of the output is only as robust as the datasets it has been provided with over time.

Use of Data in Machine Learning

Data can be in the form of text, images, numbers, and even sound or video. The datasets are analyzed to create an experience which in turn is used to create a form of machine learning program.

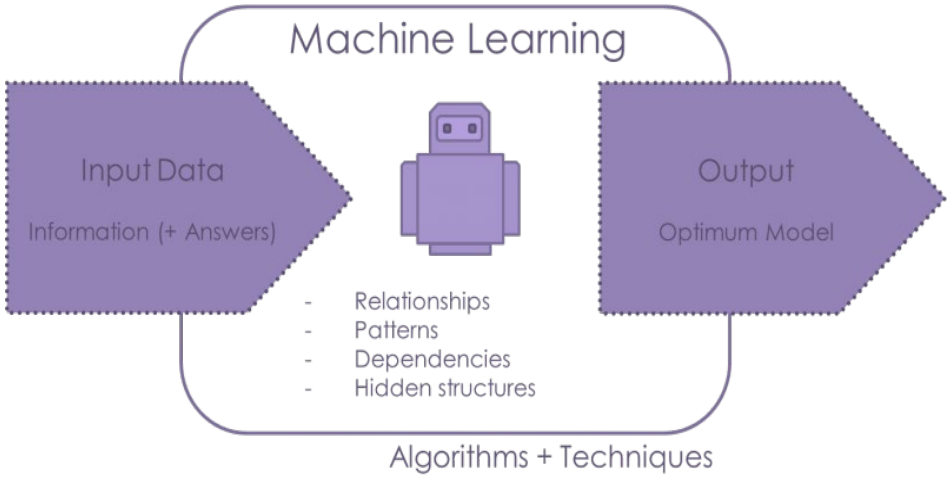


Fig 2.7: Machine Learning
Image Source - https://quantdare.com/machine-learning-a-brief-breakdown/

Different Types of Datasets

Different kinds of datasets are used to achieve a particular machine learning objective. Here are a few of them:

- Image Processing
- Sentiment Analysis
- Natural Language Processing
- Video Processing
- Speech Recognition
- Internet of Things (IoT)

Image Processing

Datasets for image processing can be used for object captioning, detection and segmentation of the dataset (Maj, 2019).

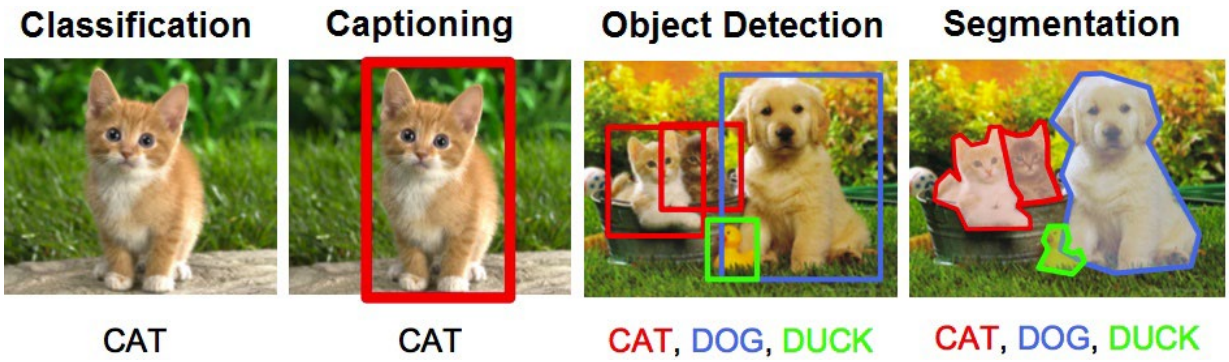


Fig 2.8: Datasets for image processing
Image Source - https://www.kdnuggets.com/2018/09/object-detection-image-classification-yolo.html

Fig 2.9: Segmentation and Captioning through Machine Learning
Image Source - https://towardsdatascience.com/faster-r-cnn-object-detection-implemented-by-keras-for-custom-data-from- a browsers-open-images-125f62b9141a

A Dataset of a variety of facial expressions is used to understand expression and caption the image accordingly. Such as a happy or sad face.

Sentiment Analysis

Whilst one parameter may be Object Recognition, another is that of the human sentiment. This layer of 'sentiment analysis' when put into context can categorize the various human emotions as a datatype and its intensity. The algorithms used for the analysis of the human sentiments are advanced and designed to generate accurate and useful results. Examples of the use of this can be to analyze the sentiment of a customer.

Fig 2.10: Emotions depicted in a smiley

By analyzing sentiment accurately, and in particular when people are unhappy, the application can focus on actions that could alter the person's emotion. This could be used for good in supporting people with certain mental illnesses, or for a not so good purpose such as convincing someone to purchase a product they may not wish too. Remember it is not the technology but how it is used that is important.

How is this achieved? Two `Polarity' nodes are created; one for a positive sentiment, the other for a negative sentiment. This is done to assist in identifying the right sentiment of a person. The words associated with a given polarity node are then re-submitted to the algorithm for more accurate sentiment analysis.

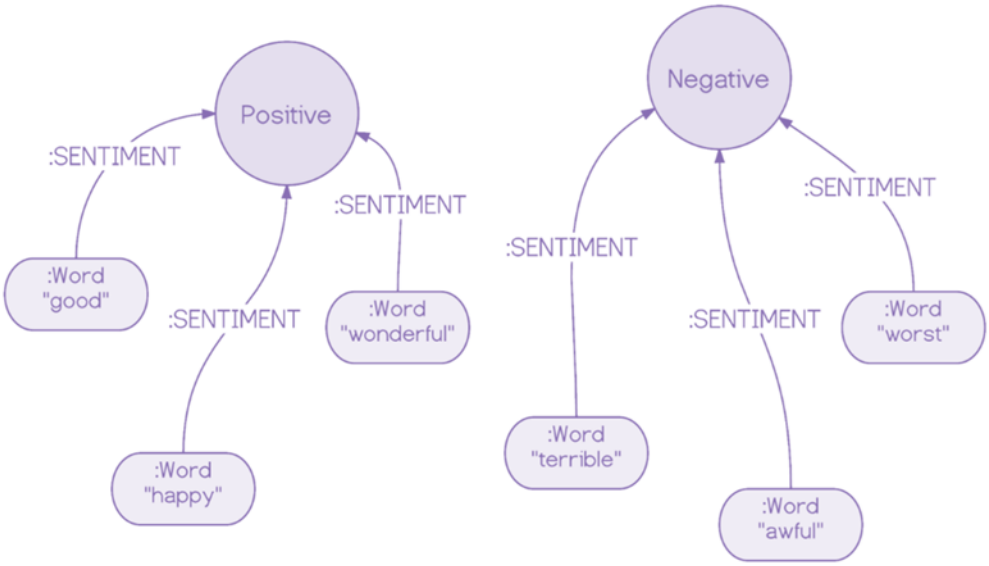


Fig 2.11: Polarity Nodes for Sentiment Analysis

Natural Language Processing (NLP)

NLP uses language comprehension to train the machine to adapt to natural changes in language such as the addition of new words, editing of words to suit the current context, labeling new words according to usage, and deleting words that have become obscure and no longer in use.

Natural language processing can be defined as a technology which enables the machine to comprehend human language in the way it is spoken and understood by humans. One of the most striking aspects is that many natural languages such as English, French, German, and Mandarin Chinese, etc. keep evolving and the learning element of the machines adapts to this. There is no fixed or permanent structure to language, thus making it very flexible. Different kinds of dialects and sub-dialects are spoken in different regions and each one of these is slightly different from the other. This includes the use of slang and other urban or sub-cultural languages.

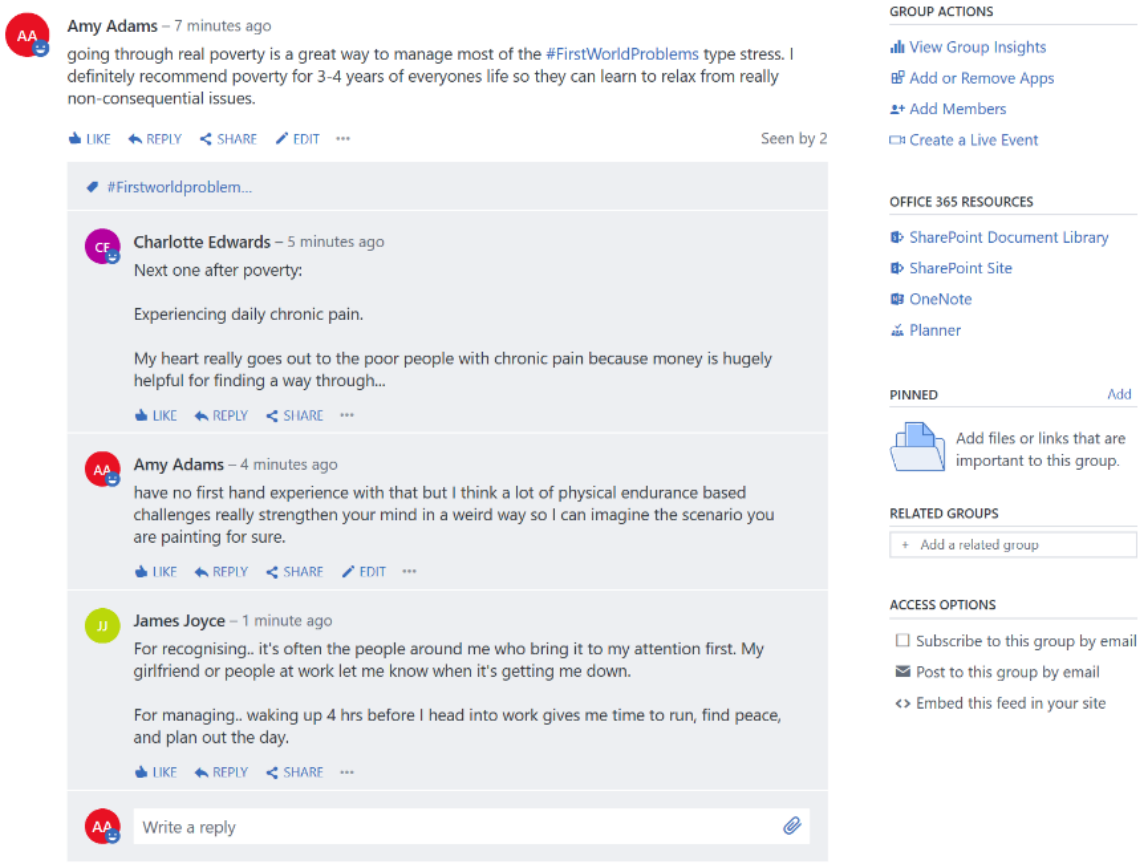


Fig 2.12: Example of Sentiment Analysis of a Yammer discussion

In Fig 2.12, the AI software takes words from a micro blogging website and associates them with a particular sentiment. This enables it to identify the overall tone of the conversation.

Video Processing

In this example the AI Software takes screenshots at regular intervals from a live video stream and analyses it to count the number of people who are about to get onto the bus. By recognizing other objects it can also calculate other parameters such as the frequency of buses arriving at the bus stop, or automatically identifying crowd rush hours across the day. This data can then be used to manage more efficient public transport.

Fig 2.13: Video Processing

Speech Recognition

Speech recognition can be defined as a technology which enables the recognition of the spoken word and subsequent translation into text. The machine learns ways of identifying and analyzing various human voices, both live and recorded, and processes them accordingly, such as the conversion into text for dictation purposes in word-processors such as Microsoft Word. It is also used to understand the user in voice-activated modules in automated cars, and in the important role of assisting those people with disabilities.

Internet of Things (IoT)

The Internet of Things (IoT) makes reference to the countless networked devices we use to make our lives easier. These devices rely on the Internet to gather and share data from responsible sources in order to provide 'smart' services. With these huge datasets and massive amounts of data sources becoming a reality, machine learning has become an integral part of our daily lives.

Machine learning can be applied in almost all scenarios where the outcome is known. It can however also be applied where the datasets are unknown and in situations where there are repeated forms of the same sort of data which can be used to reinforce the machine learning. For example, machine learning can help in understanding and analyzing the patterns of waves and oceanic currents in order to predict future sea temperatures, monsoon patterns, and even the potential for a cyclone or other natural disaster in a specific geographical location.

Capturing IoT and Sensor Data

The Internet of Things (IoT) is more of a concept than an actual thing. The concept is to allow us to interpret data from networked sensors or devices in the most meaningful ways possible. The aim is to measure, analyze, visualize, predict, and react to the data accumulated from these sensors. One form of IoT most people are familiar with is a smart thermostat, smart switches, or other internet-connected devices and appliances in your house. These are generally considered part of Consumer IoT. Then there's Industrial IoT, or IIoT. This includes things like the use of IoT devices in smart buildings, industrial automation, and monitoring of industrial processes.

Processing IoT Data

Processing the data from connected IoT sensors requires time and many interactions with sub-procedures such as:

- Standardizing or transforming the data into a uniform format to ensure it is compatible with your application.
- Creating and Storing a backup of the newly transformed data.
- Removing any repetitive, outdated, or unwanted data to help improve accuracy.
- Integration with additional structured (or unstructured) data from other sources to help enrich the dataset.

IoT Data Analytics

When we apply data analysis tools or procedures to different types of IoT data, the process is called IoT analytics. This process is performed on huge datasets to improve the efficiency of procedures, applications, business processes, and production. Several types of data analytics can be used on IoT data:

Prescriptive analytics

Prescriptive analytics is used to analyze what steps to take in a specific situation. It's often described as being a combination of descriptive and predictive analysis. When used in commercial applications, prescriptive analytics helps decipher large amounts of information to obtain more precise conclusions.

Spatial analytics

This is used to analyze location-based data. Spatial analytics deciphers various geographic patterns, determining any type of spatial relationship between various physical objects. Parking

applications, smart cars, and crop management are all examples of applications that benefit from spatial analytics.

Streaming analytics

Streaming analytics, sometimes referred to as event stream processing, is the analysis of massive datasets of moving images. These real-time data streams can be analyzed to detect emergency or urgent situations, facilitating an immediate response. The types of IoT applications that benefit from streaming analytics include those used in traffic analysis and air traffic control, and CCTV by Police.

Time series analytics

Time series analytics is based on time-based data, which is analyzed to show any anomalies, patterns, or trends. Two systems that greatly benefit from time series analytics are health and weather-monitoring systems.

We are surrounded by IoT data in our homes, our cars, and in our schools. The amount of data that IoT technology produces is massive. By collecting, processing, and analyzing this data, we can gain valuable insights to help us make better decisions about their future.

The following links give access to free datasets of IoT and sensor-based data for you to download.

- https://data.world/datasets/iot
- https://hub.packtpub.com/25-datasets-deep-learning-iot/
- https://www.kaggle.com/uciml/biomechanical-features-of-orthopedic-patients
- https://www.datasciencecentral.com/profiles/blogs/great-sensor-datasets-to-prepare-your-next-career-move-in-iot-int

How machines learn

Difference between A.I., ML and DL

- A.I. (Artificial Intelligence) - Any technique that enables computers to mimic human intelligence.
- ML (Machine Learning) - A subset of A.I. that enables machines to improve the execution of tasks with experience.
- Deep Learning - A subset of machine learning that enables software to train itself to perform tasks with vast amounts of data.

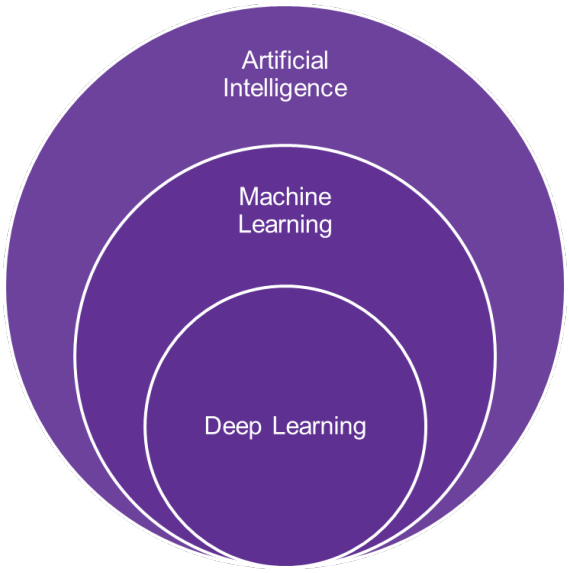


Fig 2.14: Ways that machines learn

Rules-based approach

The first ,and perhaps the most simple, approach to training bots is the rules-based approach (also known as a decision-tree bot). These bots are the most common, and many of us have likely interacted with one either through Live Chat features, on e-commerce sites, or via social media.

In rule-based systems, the data and rules used to make the decisions are carefully constructed based on the knowledge of human experts. Expert systems are generally used in situations where the level of unknown and/or variance is low.

Below is an example of a traditional 'rule-based' chatbot. Notice how it resembles the structure of a decision tree!

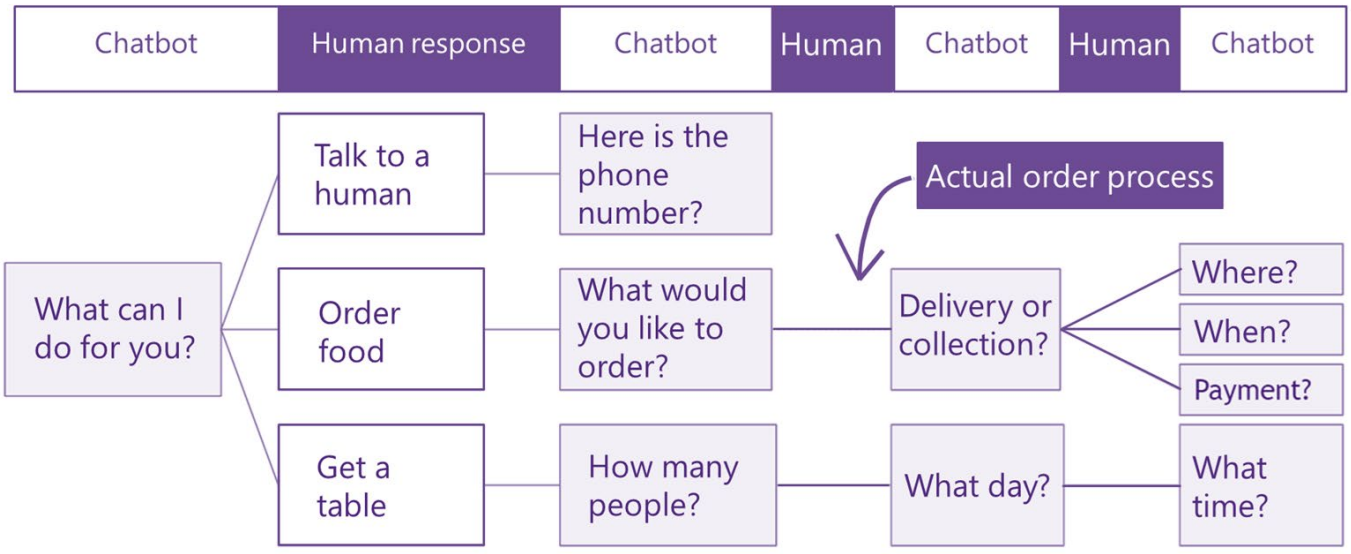


Fig 2.15: Rule-based chatbot

Machine Learning

Unlike rule-based algorithms, AI-powered bots learn as they go. AI bots that use machine learning (as opposed to traditional rule-based systems):

- learn from information gathered
- continuously improve as more data comes in
- understand patterns of behaviour
- have a broader range of decision-making skills
- can understand many languages

Major machine learning methods

There are many widely adopted machine learning methods but the three main methods include supervised learning, unsupervised learning and reinforcement learning:

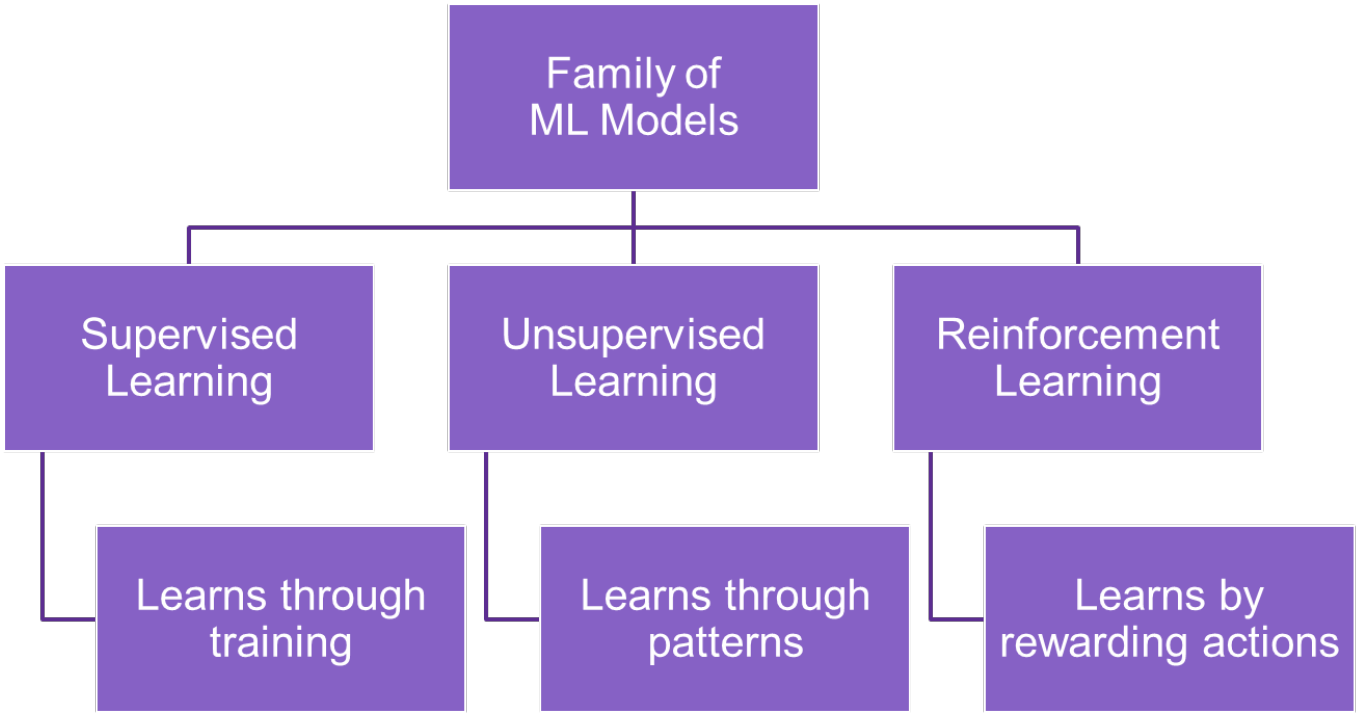


Fig 2.16: Family of machine learning models

Supervised Learning

Supervised Learning algorithms are trained using labeled examples, from an input where the desired output is known. The learning algorithm receives a collection of inputs and the corresponding correct outputs and learns by comparisons between its actual output and previous correct outputs in order to identify errors, and modifies its model accordingly. Using strategies such as classification, regression, prediction and others, supervised learning uses patterns to predict the values of the label on unlabeled data. Supervised learning is usually employed in applications wherever historical knowledge can easily predict future events.

Supervised Learning can be thought of as a teacher providing students with the correct answers to a set of known questions upon which they can develop a strategy / learn how to answer similar questions.

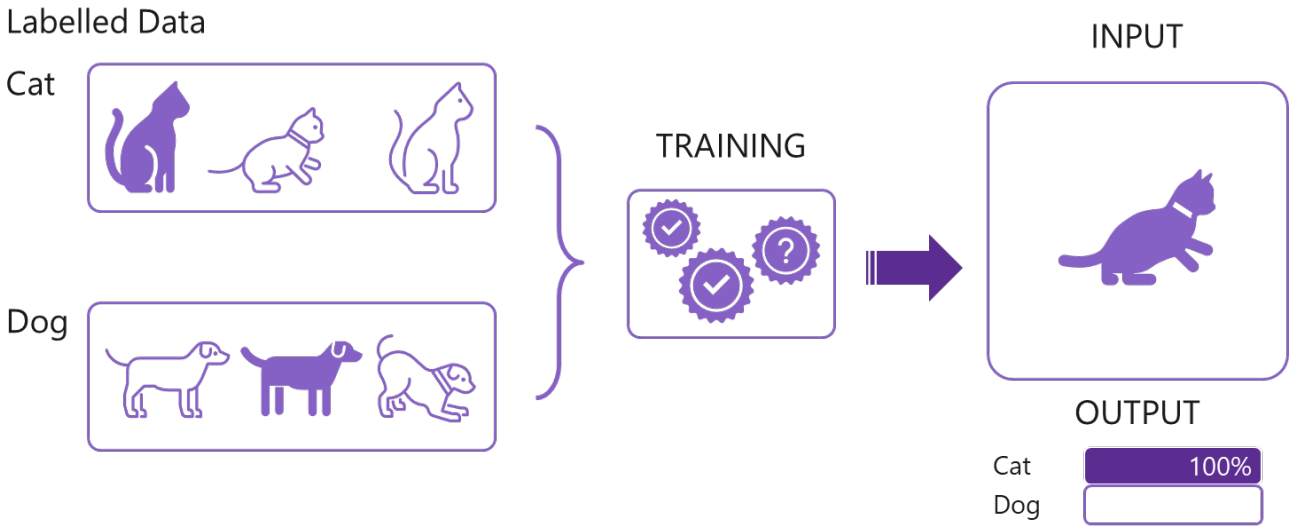


Fig 2.17: Pictorial example of Supervised Learning

Data Labeling

One of the most important requisites of supervised learning is the labeling of data. With artificial intelligence having more of an impact on our daily routines, there is a constant need to upgrade the machines in order to continue to provide results with ever enhanced accuracy. To accomplish this the data input into the algorithm must be precisely labeled.

Look at the image in Figure 2.18 closely. What do you see? In it, unlabeled data gives a learning machine no information about what is in the image.

Fig 2.18: Unlabeled data is difficult for Machine Learning

As such the machine cannot learn much about it and therefore the outcomes are inaccurate. From a machine point of view the need for accurately labeled data is of the utmost priority in order for it to understand what the image shows, what is written in a piece of text, and even what a sound recording contains.

In the subsequent image (Figure 2.19), the data has been labeled. The machine can now easily identify what is understood from the image and can find similar patterns from other images when fed similar data. Data labeling is a process that involves putting electronic boundary boxes on image files and tagging them with keywords that are both related and relevant to the item within the boundary. It can also involve many other processes such as marking a human face with points to analyze facial features for use in person identification search engines such as those used by the police. Another important aspect is the categorization of texts, audio files and videos, based on their content. In our example, the tag would be a 'car' as the traffic image shows many varieties of vehicles including cars, mini trucks, open vans, two-wheelers, buses etc.

Fig 2.19: Labeled data is easy for Machine Learning

As mentioned earlier, labeling of data may also involve the identification and marking of certain points on the face such as the nose, eyes etc. Data marked like this is done from various angles in order that the machine can recognize the human face more appropriately. The labeling occurs repetitively in the image, a car in our example, is done to teach the machine that the label applies to the car irrespective of how it looks, from what position, what color, and the angle the image of the car was captured.

Similarly, the machine needs to learn the analysis of both text and the sentiments in order to produce accurate textual outcomes. In a text scenario, the natural language is structured in such a manner that the algorithm can understand and compute the relevant meaning of the text. For spoken text, the machine needs to understand not only the word but also the tone and context in which the word is spoken to correctly gauge the true meaning of the word and attribute the same to an emotion.

Machine Learning and classification

Classification, also known as categorization, is a machine learning technique that uses known data to determine how the new data should be classified into a set of existing categories.

Imagine that we have weighed and measured 200 cats (100 Norwegian Forest cats and 100 domestic cats) and plotted those results in a graph. Using our eye, we could draw two lines to roughly determine the boundary between a Forrest cat and a domestic cat. We call the space inside this boundary the 'decision space' (see figure 2.20).

Norwegian Forest Cat

Domestic Moggy

Labelled data

	Mass (Kg)	Length (cm)
Forest cat	8	40
Moggy	3	31
Moggy	2.8	29
Forest Cat	7.6	37

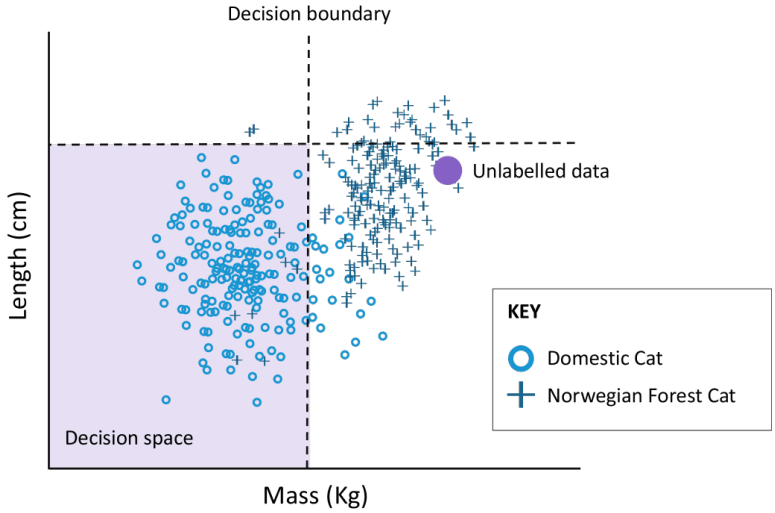


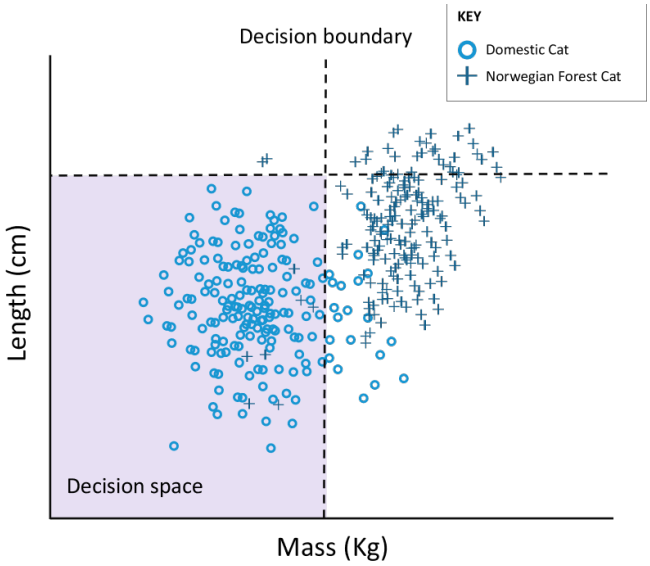
Fig 2.22: Using machine learning to classify different species of domestic cat

Question: Which type of cat represented by the 'unlabelled data'. Answer: Norwegian Forest cat.

Confusion Matrix

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualisation of the performance of an algorithm.

In this confusion matrix (fig 2.23), of the 100 forest cat pictures, the system judged that 21 were domestic cats, and of the 100 domestic cat pictures, it predicted that 9 were forest cats. All correct predictions are located in the diagonal of the table (highlighted in bold), so it is easy to visually inspect the table for prediction errors, as they will be represented by values outside the diagonal.



Training data: 200		Actual class	
		Forest Cat	Domestic
Predicted class	Forest Cat	**79**	21
	Domestic	9	**91**

Fig 2.23: Example of a confusion matrix

Decision algorithms

In the previous example (fig 2.23), a horizontal and vertical line was used to plot the decision boundary, but a computer can create any path it sees fit (e.g., diagonals or curves) in order to find the most efficient decision boundary.

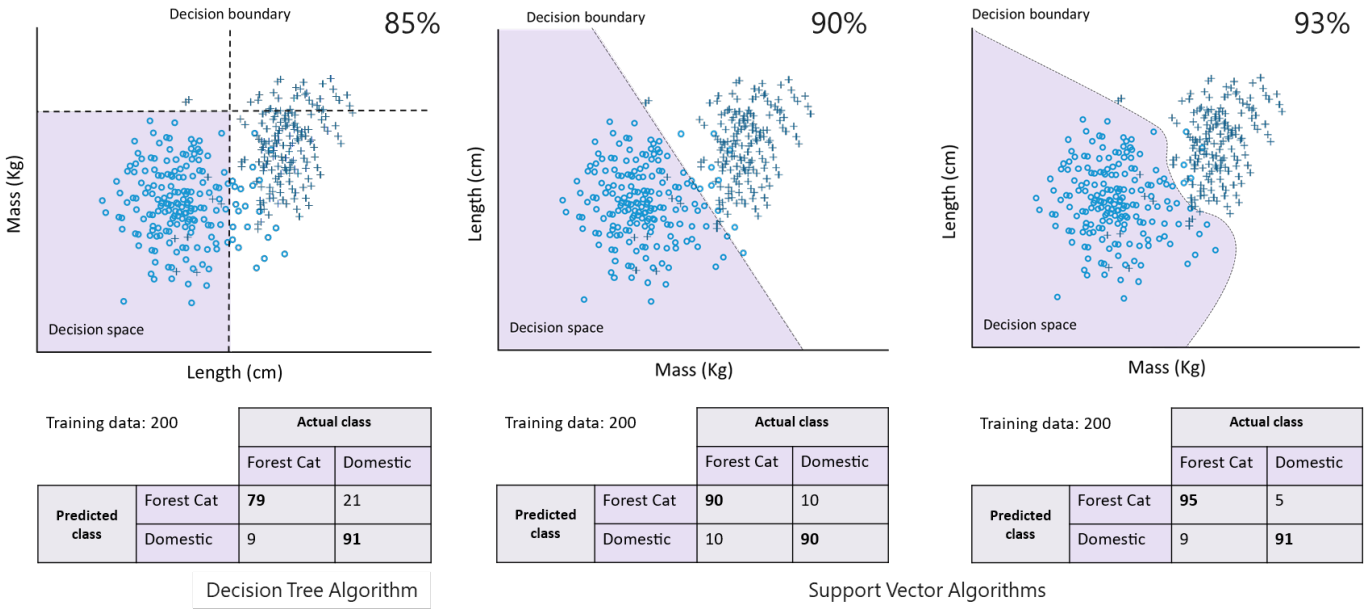


Fig 2.24: Decision algorithms

So, how does a machine tell the difference between a Forest Cat and an overweight cat? The answer lies in the number of features the algorithm uses.

It is possible to represent 3 features using a graph but beyond 3 features it becomes difficult for us to visualize. This is where we need computers to step in.

	Mass (Kg)	Length (cm)	Ear length (cm)
Forest cat	8.0	40.0	2.0
Moggy	3.0	31.0	0.9

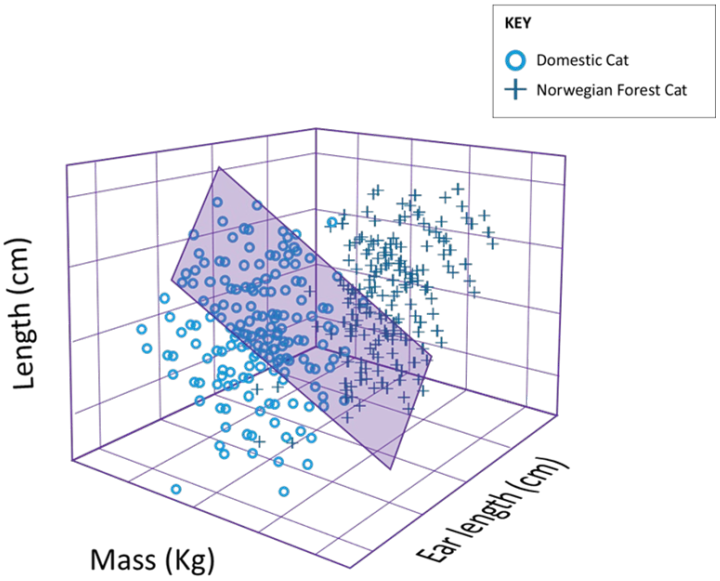


Fig 2.25: A decision algorithm which uses 3 features represented as a graph.

A Machine Learning algorithm receives data represented as numbers. In the example of the forest cat vs the domestic cat (fig 2.22), those numbers included the mass (kg) and length (cm) but pretty much anything can be converted to a number. For example, a sound wave can be represented as the amplitude taken at different samples and an image can be represented by the brightness levels of each individual pixel.

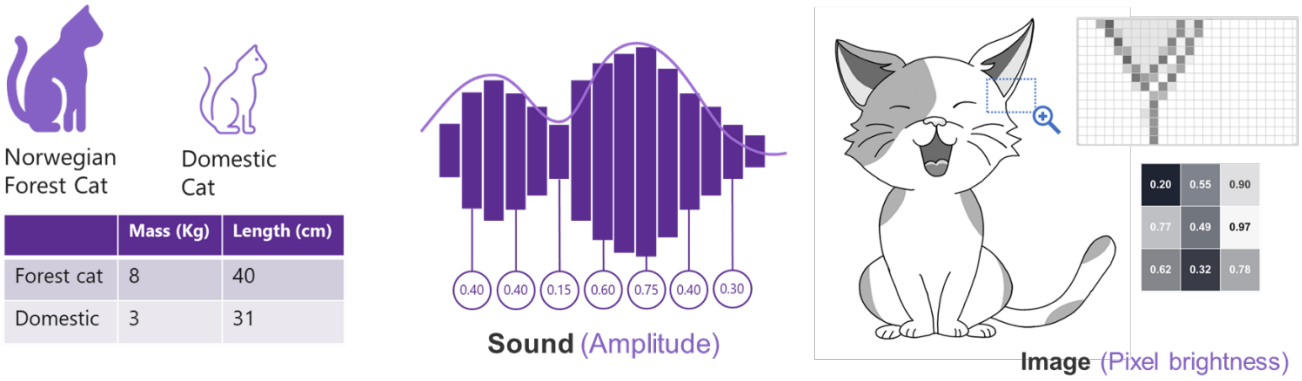


Fig 2.26: How a Machine Learning algorithm might receive data represented as numbers.

Semi-Supervised learning

This type of learning is used for the same types of applications as supervised learning, but uses a mix of labeled and unlabeled data. Semi-supervised learning is useful when the cost associated with labeling is too high to allow for a fully labeled training process. Early samples of this is seen in systems that distinguish an individual's face on a web-based camera from other faces.

Classroom Activity

- Suppose you had a basket filled with a number of different shaped blocks (circle, square, triangle and rectangular).
- Your task is to arrange them into groups.
- To understand the task first assign names to these shapes.
- We have four types of block called circle, square, triangle and rectangular. But they could be called anything.

You learn from previous information about the physical characters of the blocks, so arrange some of the blocks from the basket of the same type together. In data mining terminology the earlier work is called training the data.

Now take a new block from the basket, note its size and shape and put it in the right group based on what you have learn from an analysis of previous blocks.

This is Supervised Learning. The dataset you will have used to classify the blocks will be as follows:

Ser.	Description of the block	Block Name
1.	Round without corners	Circle
2.	With 4 straight sides of equal length and 4 corners	Square
3.	With 3 straight sides of equal or unequal length and 3 corners	Triangle
4.	With 4 straight sides (opposite sides are equal) and 4 corners	Rectangle

Table 1: Dataset for Classroom Activity

Supervised Learning in Minecraft

Explore basic coding concepts and learn about Artificial Intelligence (AI) and Supervised Learning in this free Hour of Code lesson in Minecraft: Education Edition! Help the Agent prevent forest fires with Minecraft and MakeCode: https://education.minecraft.net/hour-of-code

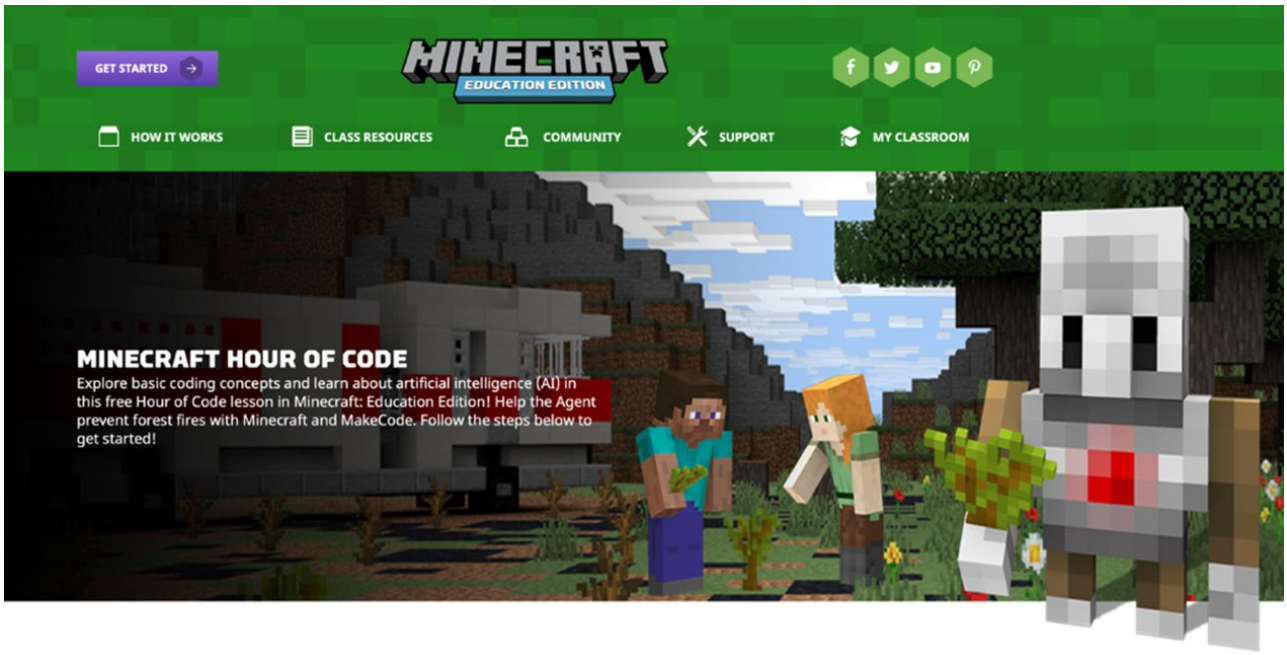


Fig 2.27: Supervised Learning in Minecraft using MakeCode.

Unsupervised Learning

This kind of learning is used with data that has no historical labels and therefore cannot use these to help it learn. However, the dataset may contain a few data points that are labeled. The algorithm needs to compute and analyze based purely on this limited amount of labelled data available and learn how to automatically label those that are not. The goal is to explore the data and recognize some patterns contained within.

Unsupervised learning works well on transactional data. For example, it can identify customers with similar attributes who can then be treated similarly in marketing campaigns. Or it can notice the most prominent attributes that a particular group of customers have. Popular techniques which use this form of learning include self-organizing maps, nearest-neighbor mapping, marketing analysis etc.

Unsupervised Learning is similar to a teacher asking students to solve a problem with no known expertise.

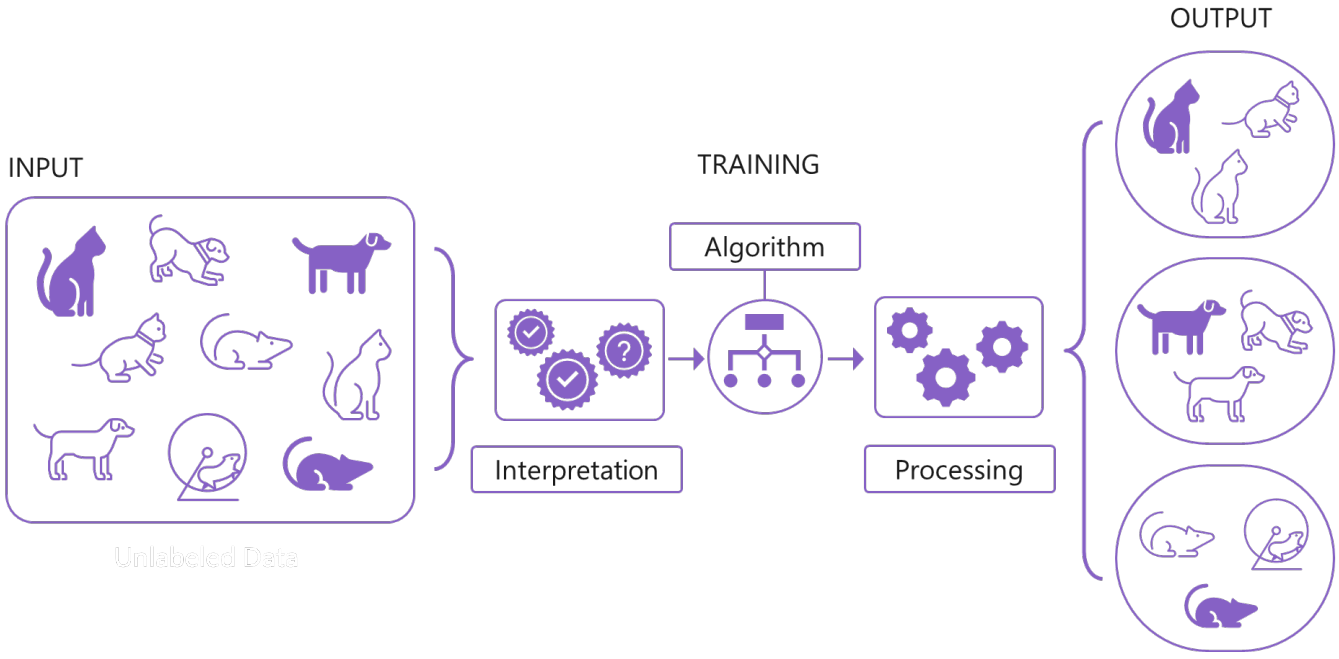


Fig 2.28: Pictorial example of Unsupervised Learning

Reinforcement learning

This form of learning is used in robotics, gaming and navigation applications. In reinforcement learning the algorithm discovers through trial and error which actions yield the greatest rewards. This type of learning has three primary components: the agent (the learner or decision-maker), the environment (everything the agent interacts with) and actions (what the agent can do).

The objective is for the agent to choose actions that maximize the expected reward over a given amount of time. The agent will reach the goal much faster by following a good strategy.

The goal in reinforcement learning is to learn the best strategy.

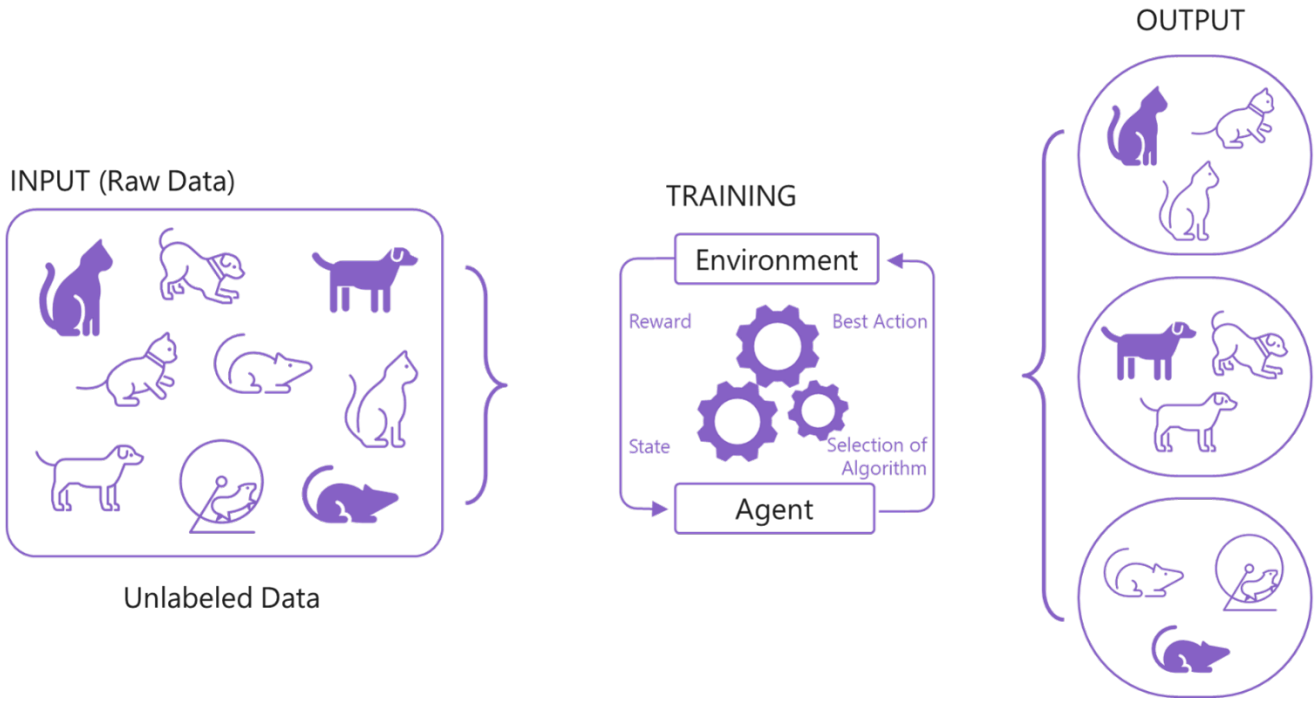


Fig 2.29: Pictorial example of Reinforcement Learning

Assessments Questions

1. Define data and list examples of data that you can think your school would gather from others or prepare itself and how it could be used.

2. What do you associate as sentiment in a human being?

3. How is machine learning useful in meteorology?

4. How can customer services be improved by using sentiment analysis?

5. Write down the correct term for each of these definitions:

6. Why Machine learning is the need of today's world?

7. Give at least 4 real-life examples of how machines are replacing humans.

8. Using any example from your daily life to explain what you have understood from unsupervised and supervised learning.

9. Explain the use of reinforcement learning in gaming.

10. What is the process of marking human face related data points?

11. What is Machine Learning?

12. What are the probable applications of Supervised Machine Learning?

13. Why is labeled data easier for a machine to learn from?

14. Fill in the Blanks

- In Data mining terminology, the earlier work is called as _____.
- Supervised learning uses _____.
- A generic BOT framework consists of a _____.
- A CHATBOT translates the text written by the user using a _____

True or False

- Unlabeled data is less expensive and takes less effort to acquire.
- Semi supervised learning uses only unlabeled data.
- Reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards.

Questions to consider

- Name the major differences between alarms and reminders.
  https://medium.com/truemd/whats-the-difference-between-an-alarm-and-a-reminder-a73c11dc1a73

- What probable challenges will Cortana face when playing a song that has a remake version too?
  https://www.howtogeek.com/402579/i-used-a-cortana-smart-speaker-all-weekend.-heres-why-it-failed/

- Probe the various aspects of the debate of 'human vs. BOT interaction'.
  https://www.intercom.com/blog/bots-versus-humans/
  https://www.retaildive.com/news/70-of-consumers-still-want-human-interaction-versus-bots/543324/
  https://cyfuture.com/blog/the-great-bot-battle-ai-chatbots-vs-human-powered-live-chat

- Investigate the various data labeling approaches and find out the pros and cons with suitable examples.
  https://www.kdnuggets.com/2018/05/data-labeling-machine-learning.html

- Can a BOT replace humans?
  https://www.bbntimes.com/en/technology/the-rise-of-chatbots-will-they-replace-humans

- Can BOTs turn malicious?
  https://www.webroot.com/us/en/resources/tips-articles/what-are-bots-botnets-and-zombies
  https://www.symantec.com/blogs/feature-stories/malicious-bot-attacks-why-theyre-more-dangerous-ever

- Elaborate the steps for data protection? Name the steps to annotate the data?
  https://www.richardsandrichards.com/6-steps-to-complete-data-protection-for-your-small-business/
  https://resources.infosecinstitute.com/how-to-implement-a-data-privacy-strategy-10-steps/#gref
  https://medium.com/thelaunchpad/spinning-up-an-annotation-team-c74c6765531b

- Deduce how much data is required for analysis by the machine?
  https://machinelearningmastery.com/much-training-data-required-machine-learning/
  https://towardsdatascience.com/how-do-you-know-you-have-enough-training-data-ad9b1fd679ee

- Find out communities which give free data for research?
  https://www.nature.com/sdata/policies/repositories

- Design steps for 'Tic Tac Toe' so that winning chances of computer is maximum.
  https://www.wikihow.com/Win-at-Tic-Tac-Toe

- Calculate the best moves to solve the 'Tower of Hanoi' problem. https://en.wikipedia.org/wiki/Tower_of_Hanoi

- Write down the different appliances that generate datasets, in a smart home
    - Security camera related data
    - Thermostat related data
    - Electricity consumption data

Some Practical Assignments/Lab Work

Assignment 1

Use the Bing search engine to prepare a report on the following.

A. Types of machine learning.
B. Heuristic search in AI.
C. Knowledge representation technique.
D. AI based games for competitive entertainment such as Chess.

Assignment 2

Outline the challenges that can be encountered in problem identification.

https://www.toolshero.com/problem-solving/problem-definition-process/

Assignment 3

Evaluate the importance and role of an identifier in the dataset.

https://www.ngdc.noaa.gov/wiki/index.php/Data_Set_Identifiers_and_other_Unique_IDs

https://www.dataone.org/best-practices/provide-identifier-dataset-used

Assignment 4

Create datasets for the following:

- **Image processing** – Create a dataset of at least 100 images of natural scenes
- **Sentiment Analysis** – Chose a product of your choice and search more than one ecommerce website. Write down all the reviews (not less than 100) written for the product in a document under the website name.
- **Video processing** – Shoot or download birthday party videos (not less than 50) and collect them in a folder.
- **IoT** – Download (any three) IoT data online. Suggested – Weather data, traffics data, agriculture data and smart phone data.

Assignment 5

Suggested questions for different scenarios:

Club membership enquiry

- What is your age (check for eligibility)?
- What games do you play?
- What are your play timings?
- Duration for membership – quarterly, bi-monthly, half yearly or yearly?

Career guidance or higher education

- Percentage scored in class 10th exams?
- Present options for subjects?

What is the user preference – (present streams – science, arts, commerce, commerce with mathematics etc.)

Design a BOT interaction possible and questions and answers based on the same for Club membership enquiry scenario.

Design a BOT interaction possible questions and answers for career guidance or higher education in AI scenario.

Assignment 6

Suggested keywords

- **Staff communication mails** – staff, teacher, educator, subject, class teacher, discipline, permission, class, etc.
- **Parent complaint mails –** ward, mother, father, guardian, class, student, complaint, unaware etc.
- **Educational bodies mails** – authority, board, school, inspection, requirements etc.
- **Vendor mails** – vendor, issue, payment, permission, principal, office, dated etc.
- **Co-Curricular notification mails** – notification, circular, school, district, state, level, competitions, class, participation, participate etc.

It is to be noted that there could be certain keywords that would be common to more than one communication type. The machine is expected to focus on both similar and dissimilar keywords and labels to identify and segregate.

Consider the scenario of a school where the principal needs help from a machine-based application with mail segmentation. Students are to consider segregation of the mails in the following categories

- Staff communication mails
- Parent complaint mails
- Educational bodies mails
- Vendor mails

- Co-Curricular notification mails

Assignment 7

Prepare a detailed report on how machines develop intelligence and learn from reinforcement methodology in a game of chess.

https://www.infoworld.com/article/3400876/reinforcement-learning-explained.html

Assignment 8

Design a student's assistance program for students with low performance. How can AI assist in identifying the weak students?

Assignment 9

Refer to the website below to understand the IRIS dataset and answer the following questions.

https://archive.ics.uci.edu/ml/datasets/iris

A. What are the features/attributes of the dataset?
B. What are the targets/classes of the dataset?
C. How many rows are there in the dataset?
D. Are there any missing values in the dataset?
E. Is the data univariate or multivariate?
F. If we follow 60:20:20 pattern for train, validate and test, how many rows will be there in each of the dataset?

Assignment 10

Perform classification of students to understand who would be interested in joining the sports club of the school.

Lab Session -1: Create a dataset or use any existing free dataset

Lab Session -2: Study the dataset of the students

Lab Session -3: Dataset should include:

- Student ID/Admission number
- Interest in the sports (name of the sport in which the student is interested to participate)
- Achievement in the sports
- Academic scores
- Distance from school to home
- Height and shoe size

Practical Assignments

Assignment 1

Imagine a situation at home where your family is expecting guests. You have lights at various locations both indoors and outdoors. There are lights at the doorway, near the gate, along the pathway, in the garden and also in the interior rooms of the house. Each family member has a different option about which light to switch on. Write down the problem statement and alternative solutions.

Assignment 2

Imagine a hypothetical situation where you are looking forward for various applicable career choices with a help of a CHATBOT assisting you in the process. Prepare a set of question/answer trails.

Assignment 3

Create a bank of images (more than 50). It should contain images of people with emotions (various ages, color, expressions etc.). Using Microsoft's online 'Face and Emotion Recognition' application, run the images to predict the emotion and analyze visual content.
https://aidemos.microsoft.com/face-recognition

Further Reading

- https://www.forbes.com/sites/willemsundbladeurope/2018/10/18/data-is-the-foundation-for-artificial-intelligence-and-machine-learning/#3eccba5251b4
- https://towardsdatascience.com/role-of-data-science-in-artificial-intelligence-950efedd2579
- http://www.dbta.com/BigDataQuarterly/Articles/The-Importance-of-Data-for-Applications-and-AI-129316.aspx
- https://www.technative.io/data-quality-vs-data-quantity-whats-more-important-for-ai/
- https://pjreddie.com/darknet/yolo/
- https://www.houseofbots.com/news-detail/3581-4-understand-the-machine-learning-from-scratch-for-beginners
- https://www.minigranth.com/artificial-intelligence/problem-solving-in-artificial-intelligence/
- Problem Solving in Artificial Intelligence by Prof Philippe Codognet Link - http://webia.lip6.fr/~codognet/PSAI/1-introduction.pdf
- Introduction to Artificial Intelligence: Problem Solving and Search by by Berhard Beckert 2004. Link - https://formal.iti.kit.edu/~beckert/teaching/Einfuehrung-KI-WS0304/04ProblemSolving.pdf
- Learning problem solving (artificial intelligence, machine) by Bruce Walter Porter by University of California, Irvine 1984.
- Learning problem solving strategies using refinement and macro generation by HA Güvenir, GW Ernst, Elsevier Science Publishers B.V. (North-Holland) 1990. Link - http://repository.bilkent.edu.tr/bitstream/handle/11693/26215/bilkent-research-paper.pdf?sequence=1&isAllowed=y
- Microsoft Power BI Dashboards Step by Step 1st Edition by Errin O'Connor
- The 5 Clustering Algorithms Data Scientists Need to Know - https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
- Clustering Introduction & different methods of clustering - https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/
- What is Data Labeling? - https://www.youtube.com/watch?v=_BasmAAub7w
- Why Smart Labeling is the Future of Data Annotation - https://www.youtube.com/watch?v=V33Ut36eUsY
- Four Mistakes You Make When Labeling Data - https://towardsdatascience.com/four-mistakes-you-make-when-labeling-data-7e431c4438a2
- Practical Machine learning problems - https://machinelearningmastery.com/practical-machine-learning-problems/
- https://www.messengerpeople.com/chatbots-what-is-a-whatsapp-bot-actually/
- https://www.geeksforgeeks.org/what-is-reinforcement-learning/
- https://deepsense.ai/what-is-reinforcement-learning-the-complete-guide/

Reference Links

- Algorithmia (2018). Introduction to Unsupervised Learning | Algorithmia Blog. [online] Algorithmia Blog. Available at: https://blog.algorithmia.com/introduction-to-unsupervised-learning/ [Accessed 10 Sep. 2019].
- Al-Masri, A. (2019). What Are Supervised and Unsupervised Learning in Machine Learning? [online] Medium. Available at: https://towardsdatascience.com/what-are-supervised-and-unsupervised-learning-in-machine-learning-dc76bd67795d [Accessed 6 Sep. 2019].
- Author (2019). Data labeling service: training data for machine learning | Clickworker. [online] Clickworker.com. Available at: https://www.clickworker.com/crowdsourcing-glossary/data-labeling/ [Accessed 6 Sep. 2019].
- Automationanywhere.com. (2019). TAKE CHARGE OF THE BOT LIFECYCLE. [online] Available at: https://www.automationanywhere.com/in/solutions/enterprise-bot-lifecycle-management [Accessed 12 Jul. 2019].
- Brownlee, J. (2015). Basic Concepts in Machine Learning. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/basic-concepts-in-machine-learning/ [Accessed 29 Jun. 2019].
- Chen, J. (2019). Neural Network Definition. [online] Investopedia. Available at: https://www.investopedia.com/terms/n/neuralnetwork.asp [Accessed 27 Sep. 2019].
- Chris, (2009) How To Write A Problem Statement | Ceptara. 2009. How To Write A Problem Statement | Ceptara. [online] Available at: http://www.ceptara.com/blog/how-to-write-problem-statement. [Accessed 04 July 2019].
- Decypher. (2018). Machine Learning: What it is and Why it Matters - Decypher. [online] Available at: https://www.decypher.com/machine-learning-matters/ [Accessed 4 Jul. 2019].
- Dietrich, D., Heller, B. and Yang, B. (2015). Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. [ebook] Indianapolis: John Wiley & Sons, Inc., pp.29-30. Available at: http://index-of.co.uk/Big-Data-Technologies/Data%20Science%20and%20Big%20Data%20Analytics.pdf [Accessed 13 Sep. 2019].
- Guru99team (2019). Supervised Machine Learning: What is, Algorithms, Example. [online] Guru99.com. Available at: https://www.guru99.com/supervised-machine-learning.html [Accessed 6 Sep. 2019].
- Kaushik, S. (2016). Clustering Introduction & different methods of clustering. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/ [Accessed 10 Sep. 2019].
- Loon, R. (2018). Machine learning explained: Understanding supervised, unsupervised, and reinforcement learning. [online] Big Data Made Simple. Available at: https://bigdata-madesimple.com/machine-learning-explained-understanding-supervised-unsupervised-and-reinforcement-learning/ [Accessed 6 Sep. 2019].

- Maj, M. (2019). Object Detection and Image Classification with YOLO. [online] Kdnuggets.com. Available at: https://www.kdnuggets.com/2018/09/object-detection-image-classification-yolo.html [Accessed 29 Jun. 2019].
- McFadin, P. (2019). *Internet of Things: Where Does the Data Go?* [Online] WIRED. Available at: https://www.wired.com/insights/2015/03/internet-things-data-go/ [Accessed 15 Nov. 2019].
- Sarah Mitroff. 2019. What is a BOT? - CNET. [ONLINE] Available at: https://www.cnet.com/how-to/what-is-a-bot/. [Accessed 05 July 2019].
- Sheth, B. (2016). The BOT Lifecycle. [online] CHATBOTs Magazine. Available at: https://chatbotsmagazine.com/the-bot-lifecycle-1ff357430db7 [Accessed 12 Jul. 2019].
- Shoemaker, C. (2019). *IoT Data: How to Collect, Process, and Analyze Them*. [Online] Tech. Available at: https://it.toolbox.com/blogs/carmashoemaker/iot-data-how-to-collect-process-and-analyze-them-032619 [Accessed 15 Nov. 2019].
- Simmons, D. (2019). *Pushing IoT Data Gathering, Analysis, and Response to the Edge - DZone IoT*. [Online] dzone.com. Available at: https://dzone.com/articles/pushing-iot-data-gathering-analysis-and-response-to-the-edge [Accessed 15 Nov. 2019].
- Smith, A. (2018). Understanding Architecture Models of CHATBOT and Response Generation Mechanisms - DZone AI. [online] dzone.com. Available at: https://dzone.com/articles/understanding-architecture-models-of-chatbot-and-r [Accessed 12 Jul. 2019].
- University of Bath, (2019) Data access statements - Archiving and sharing data - Library at University of Bath. 2019. Data access statements - Archiving and sharing data - Library at University of Bath. [online] Available at: https://library.bath.ac.uk/research-data/archiving-and-sharing/data-access-statements. [Accessed 04 July 2019].
- University of Nebraska-Lincoln, (2019) Remember the 5 W's | IT Best Practices | Nebraska. 2019. Remember the 5 W's | IT Best Practices | Nebraska. [online] Available at: https://its.unl.edu/bestpractices/remember-5-ws. [Accessed 04 July 2019].

Glossary

Ancient - belonging to the very distant past and no longer in existence.

Logic - a system or set of principles underlying the arrangements of elements in a computer or electronic device so as to perform a specified task.

Algorithms - a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.

Perceptions - The way in which something is regarded, understood, or interpreted.

Intervention - The action or process of intervening.

Complex - A group or system of different things that are linked in a close or complicated way; a network.

Segmentation - Division into separate parts or sections.

Sentiment - Feelings of tenderness, happiness, sadness, or nostalgia.

Emotion - a strong feeling deriving from one's circumstances, mood, or relationships with others.

Polarity - The state of having two opposite or contradictory tendencies, opinions, or aspects.

Parameter - a numerical or other measurable factor forming one of a set that defines a system or sets the conditions of its operation.

Non-linear - Not arranged in a straight line.

Crux - The decisive or most important point at issue.

Application - A program or piece of software designed to fulfil a particular purpose.

Data mining - The practice of examining large pre-existing databases in order to generate new information.

Stakeholder - A person with an interest or concern in something

Narrative - A spoken or written account of connected events.

Untagged - Of a piece of text or data not identified or categorized by a tag.