**1. What is the concept of human learning? Please give two examples.**

**Is there any similarity in learning process between humans and machines. Human learning process varies from person to person. Once a learning process is set into the minds of people, it is difficult to change it. But, in Machine Learning (ML), it is easy to change the learning method by selecting a different algorithm. In ML, we have well defined processes to understand and estimate the accuracy in learning. Estimation of human learning is usually done through examinations and it cannot be considered as a measure of intelligence. Let us examine the difference between human and machine learning process in detail in this blog.**

**Humans acquire knowledge through experience either directly or shared by others. Machines acquire knowledge through experience shared in the form of past data. We have the terms, Knowledge, Skill, and Memory being used to define intelligence. Just because you have good memory, that does not mean you are intelligent. And just because you are intelligent, it does not mean you should have a good memory. However, there are exceptions to these rules. Humans begin learning by memorizing. After few years, he realizes that mere capability to memorize is not intelligence. Then he practices on transforming the data stored in memory to knowledge and applies them to develop skills to solve problems faced in real life. A person with good memory and more knowledge without the required skills cannot be considered intelligent. Search engines replaces human memory and these days the focus is on acquiring intelligence by making use of data available on the web. In humans, learning speed depends on individuals and in machines, learning speed depends on the algorithm selected and the volume of examples exposed to it.**

**Skill is a manifestation of intelligence possessed by humans. And intelligence is the ability to apply knowledge. Human intelligence sustains, but his knowledge fades as new technologies emerge. Humans without knowledge in particular subjects can apply their intelligence to solve problems in new domains. But machines can solve new problems only if their intelligence has been updated with retraining on data acquired from the changed scenarios. This is a fundamental difference between human intelligence and machine intelligence.**

**Both humans and machines make mistakes in applying their intelligence in solving problems. In ML, overfitting memorizes all examples and an overfitted model lacks generalization and it fails to work on never seen before examples. In most of the Asian countries, the education system does overfitting of students by over coaching and tuitions on technical subjects enabling them to solve only example problems. These example problems are answered in examinations without need for applying any intelligence. These students can solve the problems which they have already seen and only the problems seen by them in the past. They are not able to handle general problems properly with accuracy because their intelligence is not generalized. This is the major reason for missing skill levels among university recruits. In short, vast majority of students become overfitted learning models and their employability is in question.**

**In ML, Transfer Learning is a technique that reuses a model that was created by machine learning experts and that has already been trained on a large dataset. Transfer learning leverages information extracted from one set of distributions. In humans, transfer of knowledge to students is often done by teachers and tuition providers. This may not make the students intelligent. But in the case of machine learning, transfer learning makes the transferee as intelligent as the transferor. In the case of humans, transfer learning only transfers the knowledge and it depends on the inherent intelligence of the transferee to enhance his/her problem solving skills.**

**To summarize, overfitting is a curse to humans and machine learning systems. Machine intelligence is limited to the areas in which they are trained. But human intelligence is independent of his domain of training. An intelligent human being will be able to solve problems related to unforeseen domains, whereas a machine will not be able to do that.**

* **attention - an observer pays attention to another person's behavior**
* **retention - the observer stores the behavior in their memory**
* **production (or initiation) - the observer must acquire the skills needed to reproduce the behavior**
* **motivation - the observer finds a reason to reproduce the behavior**

**2. What different forms of human learning are there? Are there any machine learning equivalents?**

**This is a difficult question to answer because we don’t know a whole lot yet about all the “variations in human learning”. Also, while machine learning models are loosely based on what we think human learning is , it’s largely a guess and might be completely wrong. Finally,, there’s considerable disagreement about how even to phrase the problem of human learning, and the extent to which “learning” is what we use to acquire, say language skills or visual skills.**

**Two of the most influential thinkers of the 20th century were Jean Piaget, the Swiss developmental psychologist who probably contributed the most to our knowledge of how children learn, and Noam Chomsky, the famous MIT linguist who pioneered the generative grammar framework underlying much work in linguistics.**

**There was a famous debate held in Paris in 1975 where Chomsky and Piaget butted heads, and it was largely around how to formulate human cognition — Piaget was a fan of “learning”, although his notion was far removed from today’s number-crunching statistical paradigm — and Chomsky strongly believed that most of what was needed was built an, and if there was much learning, it was essentially setting parameter values.**

[**Language and Learning: The Debate Between Jean Piaget and Noam Chomsky: Noam Chomsky, Jean Piaget, Massimo Piattelli-Palmarini: 9780674509412: Amazon.com: Books**](https://www.amazon.com/Language-Learning-Debate-Between-Chomsky/dp/0674509412)

**As far as I know, this debate rages on, and every so often, a new book appears, arguing one approach or the other. For example, Pinker’s book argues against much of modern ML and deep learning, arguing that “Tabula Rasa” approaches are doomed to fail:**

[**The Blank Slate: The Modern Denial of Human Nature: Steven Pinker: 8601200961546: Amazon.com: Books**](https://www.amazon.com/Blank-Slate-Modern-Denial-Nature/dp/0142003344)

**There’s a staggeringly large literature on human learning, from psychology to neuroscience to sociology and education. I would conservatively say more than 100,000 papers are probably out there on this topic. Yet, remarkably, even the most questions remain unanswered:**

1. **How is language acquired by a child? In the Chomskyan tradition, language in the form of universal grammar is built in, and a child takes language data to set parameters. In the Piagetian connectionist tradition, language is learned by some “universal learning” method (e.g., imagine some souped up LSTM or some such deep learning architecture). This book on child language acquisition begins by posing the same issues, and doesn’t really answer them.**[**Amazon.com: First Language Acquisition (9781316507605): Eve V. Clark: Books**](https://www.amazon.com/First-Language-Acquisition-Eve-Clark/dp/1316507602/ref=sr_1_5?ie=UTF8&qid=1539831355&sr=8-5&keywords=language%20acquisition)
2. **Regardless of which approach might be correct, no one has yet built a program that learns language with the same fluency that billions of children all over the world manage to do with the first two years of their lives. It’s not the question of data. AI programs have more data by several orders of magnitude. Data is not enough! You need to have the right algorithms, and we obviously don’t. So, this most basic of human abilities — language — remains out of reach of machine learners. We currently fake it by kludges like LSTM and Google Translate, but no one believes such chatbot approaches will scale to learn language like human children do.**
3. **How much of vision is learned and how much is built in? Another great mystery. There is clear evidence that visual filters are learned through experiments on cats and similar animals. But, the structure of the visual cortex and the specific visual pathways (where color and motion are processed separately and in parallel, before they come together, so that when you see a red sports car move, you see it as both something red and moving) seems completely hardwired. Again, there is no convergence of opinion here. An interesting experiment on ferrets by MIT professor Mriganka Sur showed that if the pathway from the lateral geniculate nucleus (LGN) is severed at birth, so that the developing brain connects the output of the eyes to the auditory cortex (an experiment only possible in ferets), the rewired feret is able to “see” with its “ears”, so to speak, since visual information is now being processed by the auditory cortex. But, it’s hard to generalize this experiment to humans, to conclude anything about universality of learning algorithms across sensory modalities.**
4. **How much of behavior is learned? Children seem especially good at learning a vast repertoire of behaviors, from walking to jumping and running, and games of all sorts, well before they ever get to a traditional school. How can they learn so much with so little data? I am amused every time I have to review a deep reinforcement learning paper. A recent paper had the time units in the experiments in billions of steps. We only live for 2 billion seconds. Children learn many interesting behaviors within the first two years of their lives — that is roughly 50 million steps. Of course, more than half that time is gone in sleeping, eating, etc. So, how is it possible that they learn hundreds of behaviors within a few million steps, whereas the best deep RL programs take 1000x as long or more. A mystery.**

**In the last 50 years of ML, clearly much progress has been made, and we now have some rudimentary capabilities in machines to do a few things. My Tesla S can drive me to work, but I chose not to use its autonomous driving mode. I am deeply skeptical that the self driving technology is reliable enough for me to trust my life with. On the same road (101) that I drive every day, an Apple engineer was tragically killed when his Tesla ran into the median boundary at high speed. There are other accidents. Of course, these are probably unavoidable in a technology that is still maturing. But, knowing what I know about how autonomous driving systems work, I don’t choose to trust my life with them.**

**The next decade or two will be exciting, no doubt, as we will see whether the current statistical deep learning approach will continue to scale, or peter out. I don’t think statistical approaches will scale beyond a point. Language is a good example. Many statistical NLP systems seem to make the most basic of errors, despite being trained on billions of sentences, possibly far more than any human sees in a lifetime of reading.**

**Douglas Hofstadter, whose breakthrough book on “Godel, Escher, Bach: An Eternal Golden Braid” got me interested in AI almost four decades ago, recently wrote an impressive article in Atlantic, which shows why Google Translate, despite being trained on terabytes of data, makes the most basic and embarrassing errors in language tranlation:**

[**The Shallowness of Google Translate**](https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/)

**His analysis can be extended to other areas, such as computer vision, where progress has been exaggerated. So, much remains to be done, but most importantly, we are missing crucial insights into how humans and machines learn. Nothing in science comes easily. Physicists and biologists took hundreds of years to work out the basic mysteries of matter, space, and life. Working out the secrets of the mind will undoubtedly take just as long.**

**3. What is machine learning, and how does it work? What are the key responsibilities of machine learning?**

**Machine learning (ML) is a type of artificial intelligence (**[**AI**](https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence)**) focused on building computer systems that learn from data. The broad range of techniques ML encompasses enables software applications to improve their performance over time.**

**Machine learning**[**algorithms**](https://www.techtarget.com/whatis/definition/algorithm)**are trained to find relationships and patterns in data. They use historical data as input to make predictions, classify information, cluster data points, reduce dimensionality and even help**[**generate new content**](https://www.techtarget.com/searchenterpriseai/definition/generative-AI)**, as demonstrated by new ML-fueled applications such as ChatGPT, Dall-E 2 and GitHub Copilot.**

**Machine learning is widely applicable across many industries.**[**Recommendation engines**](https://www.techtarget.com/whatis/definition/recommendation-engine)**, for example, are used by e-commerce, social media and news organizations to suggest content based on a customer's past behavior. Machine learning algorithms and**[**machine vision**](https://www.techtarget.com/searchenterpriseai/definition/machine-vision-computer-vision)**are a critical component of self-driving cars, helping them navigate the roads safely. In healthcare, machine learning is used to diagnose and suggest treatment plans. Other common**[**ML use cases**](https://www.techtarget.com/searchenterpriseai/feature/10-common-uses-for-machine-learning-applications-in-business)**include fraud detection, spam filtering, malware threat detection, predictive maintenance and business process automation.**

**While machine learning is a powerful tool for solving problems, improving business operations and automating tasks, it's also a complex and challenging technology, requiring deep expertise and significant resources. Choosing the right algorithm for a task calls for a strong grasp of mathematics and statistics. Training**[**machine learning algorithms**](https://www.techtarget.com/whatis/definition/machine-learning-algorithm)**often involves large amounts of good quality data to produce accurate results. The results themselves can be difficult to understand -- particularly the outcomes produced by complex algorithms, such as the deep learning**[**neural networks**](https://www.techtarget.com/searchenterpriseai/definition/neural-network)**patterned after the human brain. And**[**ML models**](https://www.techtarget.com/searchenterpriseai/tip/What-are-machine-learning-models-Types-and-examples)**can be costly to run and tune.**

**Still, most organizations either directly or indirectly through ML-infused products are embracing machine learning. According to the "2023 AI and Machine Learning Research Report" from Rackspace Technology, 72% of companies surveyed**[**said**](https://www.rackspace.com/sites/default/files/2023-02/AI-Machine-Learning-Research-Report.pdf)**that AI and machine learning are part of their IT and business strategies, and 69% described AI/ML as the most important technology. Companies that have adopted it reported using it to improve existing processes (67%), predict business performance and industry trends (60%) and reduce risk (53%).**

**TechTarget's guide to machine learning is a primer on this important field of computer science, further explaining what machine learning is, how to do it and how it is applied in business. You'll find information on the various types of machine learning algorithms, the challenges and best practices associated with developing and**[**deploying ML models**](https://www.techtarget.com/whatis/definition/machine-learning-operations-MLOps)**, and what the future holds for machine learning. Throughout the guide, there are hyperlinks to related articles that cover the topics in greater depth.**

**Machine learning has played a progressively central role in human society since its**[**beginnings in the mid-20th century**](https://www.techtarget.com/whatis/A-Timeline-of-Machine-Learning-History)**, when AI pioneers like Walter Pitts, Warren McCulloch, Alan Turing and John von Neumann laid the groundwork for computation. The training of machines to learn from data and improve over time has enabled organizations to automate routine tasks that were previously done by humans -- in principle, freeing us up for more creative and strategic work.**

**Machine learning also performs manual tasks that are beyond our ability to execute at scale -- for example, processing the huge quantities of data generated today by digital devices. Machine learning's ability to extract patterns and insights from vast data sets has become a competitive differentiator in fields ranging from finance and retail to healthcare and scientific discovery. Many of today's leading companies, including Facebook, Google and Uber, make machine learning a central part of their operations.**

**As the volume of data generated by modern societies continues to proliferate, machine learning will likely become even more vital to humans and essential to machine intelligence itself. The technology not only helps us make sense of the data we create, but synergistically the abundance of data we create further strengthens ML's data-driven learning capabilities.**

**What will come of this continuous learning loop? Machine learning is a pathway to**[**artificial intelligence**](https://www.techtarget.com/searchenterpriseai/Ultimate-guide-to-artificial-intelligence-in-the-enterprise)**, which in turn fuels advancements in ML that likewise improve AI and progressively blur the boundaries between machine intelligence and human intellect.**

**DEFINITION**

**What is machine learning and how does it work? In-depth guide**

**By**

* [**Linda Tucci,**](https://www.techtarget.com/contributor/Linda-Tucci)**Industry Editor -- CIO/IT Strategy**
* [**Ed Burns**](https://www.techtarget.com/contributor/Ed-Burns)

**Machine learning (ML) is a type of artificial intelligence (**[**AI**](https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence)**) focused on building computer systems that learn from data. The broad range of techniques ML encompasses enables software applications to improve their performance over time.**

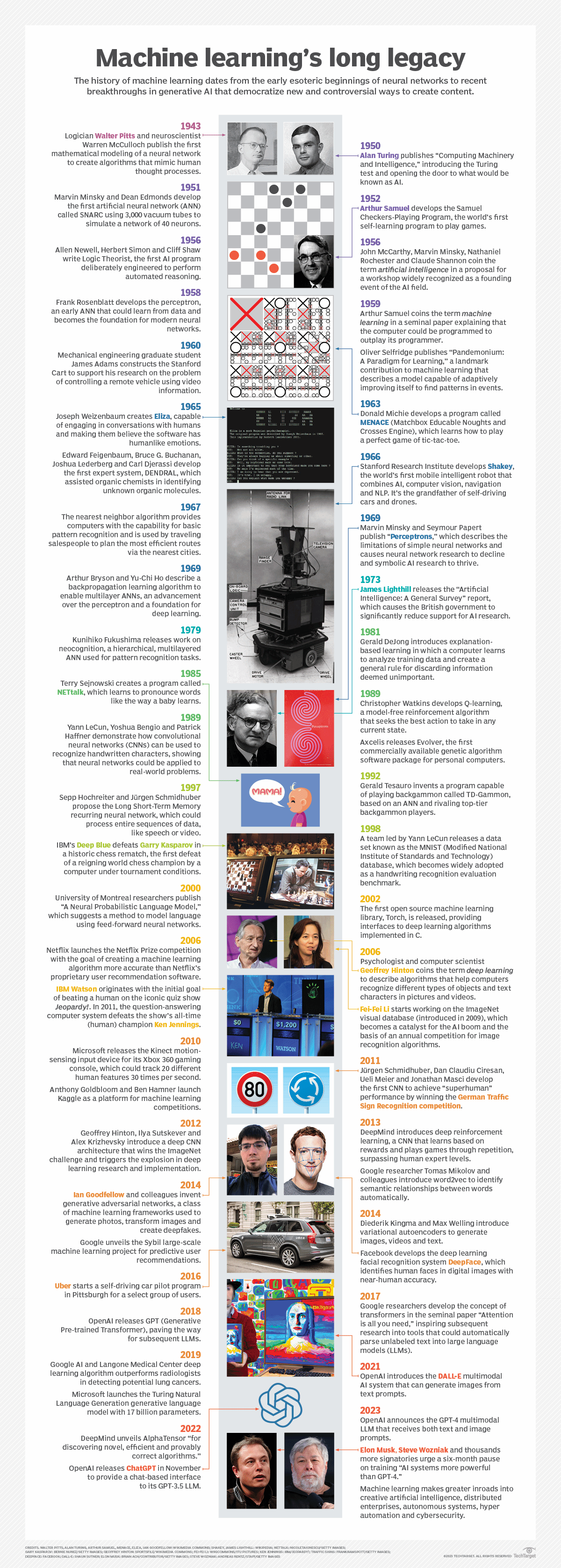
**Machine learning**[**algorithms**](https://www.techtarget.com/whatis/definition/algorithm)**are trained to find relationships and patterns in data. They use historical data as input to make predictions, classify information, cluster data points, reduce dimensionality and even help**[**generate new content**](https://www.techtarget.com/searchenterpriseai/definition/generative-AI)**, as demonstrated by new ML-fueled applications such as ChatGPT, Dall-E 2 and GitHub Copilot.**

**Machine learning is widely applicable across many industries.**[**Recommendation engines**](https://www.techtarget.com/whatis/definition/recommendation-engine)**, for example, are used by e-commerce, social media and news organizations to suggest content based on a customer's past behavior. Machine learning algorithms and**[**machine vision**](https://www.techtarget.com/searchenterpriseai/definition/machine-vision-computer-vision)**are a critical component of self-driving cars, helping them navigate the roads safely. In healthcare, machine learning is used to diagnose and suggest treatment plans. Other common**[**ML use cases**](https://www.techtarget.com/searchenterpriseai/feature/10-common-uses-for-machine-learning-applications-in-business)**include fraud detection, spam filtering, malware threat detection, predictive maintenance and business process automation.**

**While machine learning is a powerful tool for solving problems, improving business operations and automating tasks, it's also a complex and challenging technology, requiring deep expertise and significant resources. Choosing the right algorithm for a task calls for a strong grasp of mathematics and statistics. Training**[**machine learning algorithms**](https://www.techtarget.com/whatis/definition/machine-learning-algorithm)**often involves large amounts of good quality data to produce accurate results. The results themselves can be difficult to understand -- particularly the outcomes produced by complex algorithms, such as the deep learning**[**neural networks**](https://www.techtarget.com/searchenterpriseai/definition/neural-network)**patterned after the human brain. And**[**ML models**](https://www.techtarget.com/searchenterpriseai/tip/What-are-machine-learning-models-Types-and-examples)**can be costly to run and tune.**

**Still, most organizations either directly or indirectly through ML-infused products are embracing machine learning. According to the "2023 AI and Machine Learning Research Report" from Rackspace Technology, 72% of companies surveyed**[**said**](https://www.rackspace.com/sites/default/files/2023-02/AI-Machine-Learning-Research-Report.pdf)**that AI and machine learning are part of their IT and business strategies, and 69% described AI/ML as the most important technology. Companies that have adopted it reported using it to improve existing processes (67%), predict business performance and industry trends (60%) and reduce risk (53%).**

**TechTarget's guide to machine learning is a primer on this important field of computer science, further explaining what machine learning is, how to do it and how it is applied in business. You'll find information on the various types of machine learning algorithms, the challenges and best practices associated with developing and**[**deploying ML models**](https://www.techtarget.com/whatis/definition/machine-learning-operations-MLOps)**, and what the future holds for machine learning. Throughout the guide, there are hyperlinks to related articles that cover the topics in greater depth.**

****

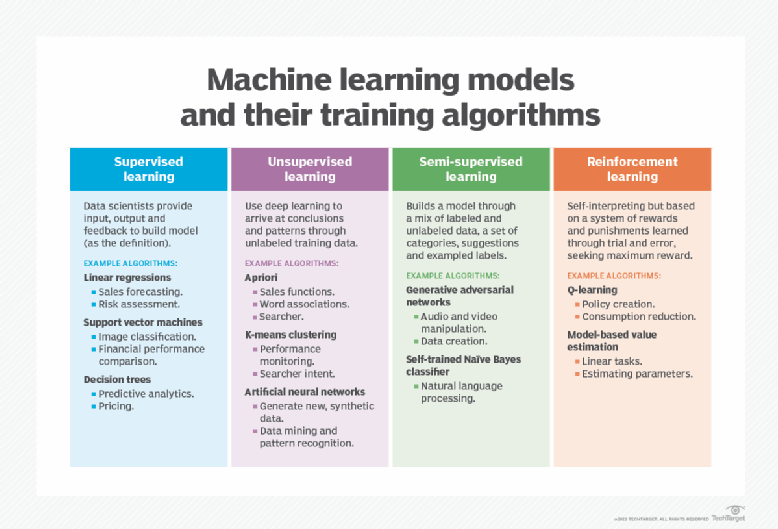
**Why is machine learning important?**

**Machine learning has played a progressively central role in human society since its**[**beginnings in the mid-20th century**](https://www.techtarget.com/whatis/A-Timeline-of-Machine-Learning-History)**, when AI pioneers like Walter Pitts, Warren McCulloch, Alan Turing and John von Neumann laid the groundwork for computation. The training of machines to learn from data and improve over time has enabled organizations to automate routine tasks that were previously done by humans -- in principle, freeing us up for more creative and strategic work.**

**Machine learning also performs manual tasks that are beyond our ability to execute at scale -- for example, processing the huge quantities of data generated today by digital devices. Machine learning's ability to extract patterns and insights from vast data sets has become a competitive differentiator in fields ranging from finance and retail to healthcare and scientific discovery. Many of today's leading companies, including Facebook, Google and Uber, make machine learning a central part of their operations.**

**As the volume of data generated by modern societies continues to proliferate, machine learning will likely become even more vital to humans and essential to machine intelligence itself. The technology not only helps us make sense of the data we create, but synergistically the abundance of data we create further strengthens ML's data-driven learning capabilities.**

**What will come of this continuous learning loop? Machine learning is a pathway to**[**artificial intelligence**](https://www.techtarget.com/searchenterpriseai/Ultimate-guide-to-artificial-intelligence-in-the-enterprise)**, which in turn fuels advancements in ML that likewise improve AI and progressively blur the boundaries between machine intelligence and human intellect.**

****

**The four main types of machine learning and their most common algorithms.**

**What are the different types of machine learning?**

**Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are**[**four basic types of machine learning**](https://www.techtarget.com/searchenterpriseai/tip/Types-of-learning-in-machine-learning-explained)**: supervised learning, unsupervised learning, semisupervised learning and reinforcement learning.**

**The type of algorithm data scientists choose depends on the nature of the data. Many of the algorithms and techniques aren't limited to just one of the primary ML types listed here. They're often adapted to multiple types, depending on the problem to be solved and the data set. For instance, deep learning algorithms such as [convolutional neural networks and recurrent neural networks](https://www.techtarget.com/searchenterpriseai/feature/CNN-vs-RNN-How-they-differ-and-where-they-overlap) are used in supervised, unsupervised and reinforcement learning tasks, based on the specific problem and availability of data.**

**Machine learning vs. deep learning neural networks**

[**Deep learning**](https://www.techtarget.com/searchenterpriseai/definition/deep-learning-deep-neural-network)**is a subfield of ML that deals specifically with neural networks containing multiple levels -- i.e., deep neural networks. Deep learning models can automatically learn and extract hierarchical features from data, making them effective in tasks like image and speech recognition.**

**How does supervised machine learning work?**

**In**[**supervised learning**](https://www.techtarget.com/searchenterpriseai/definition/supervised-learning)**, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and output of the algorithm are specified in supervised learning. Initially, most machine learning algorithms worked with supervised learning, but unsupervised approaches are becoming popular.**

**Supervised learning algorithms are used for several tasks, including the following:**

* **Binary classification. Divides data into two categories.**
* **Multiclass classification. Chooses between more than two types of answers.**
* **Ensembling. Combines the predictions of multiple ML models to produce a more accurate prediction.**
* **Regression modeling.**[**Predicts continuous values**](https://www.techtarget.com/searchenterpriseai/feature/What-is-regression-in-machine-learning)**based on relationships within data.**

**4. Define the terms "penalty" and "reward" in the context of reinforcement learning.**

**What is Reinforcement Learning?**

* **Reinforcement Learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty.**
* **In Reinforcement Learning, the agent learns automatically using feedbacks without any labeled data, unlike**[**supervised learning.**](https://www.javatpoint.com/supervised-machine-learning)
* **Since there is no labeled data, so the agent is bound to learn by its experience only.**
* **RL solves a specific type of problem where decision making is sequential, and the goal is long-term, such as game-playing, robotics, etc.**
* **The agent interacts with the environment and explores it by itself. The primary goal of an agent in reinforcement learning is to improve the performance by getting the maximum positive rewards.**
* **The agent learns with the process of hit and trial, and based on the experience, it learns to perform the task in a better way. Hence, we can say that *"Reinforcement learning is a type of machine learning method where an intelligent agent (computer program) interacts with the environment and learns to act within that."* How a Robotic dog learns the movement of his arms is an example of Reinforcement learning.**
* **It is a core part of**[**Artificial intelligence**](https://www.javatpoint.com/artificial-intelligence-tutorial)**, and all**[**AI agent**](https://www.javatpoint.com/agents-in-ai)**works on the concept of reinforcement learning. Here we do not need to pre-program the agent, as it learns from its own experience without any human intervention.**
* **Example: Suppose there is an AI agent present within a maze environment, and his goal is to find the diamond. The agent interacts with the environment by performing some actions, and based on those actions, the state of the agent gets changed, and it also receives a reward or penalty as feedback.**
* **The agent continues doing these three things (take action, change state/remain in the same state, and get feedback), and by doing these actions, he learns and explores the environment.**
* **The agent learns that what actions lead to positive feedback or rewards and what actions lead to negative feedback penalty. As a positive reward, the agent gets a positive point, and as a penalty, it gets a negative point.**

**Terms used in Reinforcement Learning**

* **Agent(): An entity that can perceive/explore the environment and act upon it.**
* **Environment(): A situation in which an agent is present or surrounded by. In RL, we assume the stochastic environment, which means it is random in nature.**
* **Action(): Actions are the moves taken by an agent within the environment.**
* **State(): State is a situation returned by the environment after each action taken by the agent.**
* **Reward(): A feedback returned to the agent from the environment to evaluate the action of the agent.**
* **Policy(): Policy is a strategy applied by the agent for the next action based on the current state.**
* **Value(): It is expected long-term retuned with the discount factor and opposite to the short-term reward.**
* **Q-value(): It is mostly similar to the value, but it takes one additional parameter as a current action (a).**

**Key Features of Reinforcement Learning**

* **In RL, the agent is not instructed about the environment and what actions need to be taken.**
* **It is based on the hit and trial process.**
* **The agent takes the next action and changes states according to the feedback of the previous action.**
* **The agent may get a delayed reward.**
* **The environment is stochastic, and the agent needs to explore it to reach to get the maximum positive rewards.**

**Approaches to implement Reinforcement Learning**

**There are mainly three ways to implement reinforcement-learning in ML, which are:**

1. **Value-based:  
   The value-based approach is about to find the optimal value function, which is the maximum value at a state under any policy. Therefore, the agent expects the long-term return at any state(s) under policy π.**
2. **Policy-based:  
   Policy-based approach is to find the optimal policy for the maximum future rewards without using the value function. In this approach, the agent tries to apply such a policy that the action performed in each step helps to maximize the future reward.  
   The policy-based approach has mainly two types of policy:**
   * **Deterministic: The same action is produced by the policy (π) at any state.**
   * **Stochastic: In this policy, probability determines the produced action.**
3. **Model-based: In the model-based approach, a virtual model is created for the environment, and the agent explores that environment to learn it. There is no particular solution or algorithm for this approach because the model representation is different for each environment.**
4. **Explain the term "learning as a search"?**

**Applied machine learning is challenging because the designing of a perfect learning system for a given problem is intractable.**

**There is no best training data or best algorithm for your problem, only the best that you can discover.**

**The application of machine learning is best thought of as search problem for the best mapping of inputs to outputs given the knowledge and resources available to you for a given project.**

**In this post, you will discover the conceptualization of applied machine learning as a search problem.**

**After reading this post, you will know:**

* **That applied machine learning is the problem of approximating an unknown underlying mapping function from inputs to outputs.**
* **That design decisions such as the choice of data and choice of algorithm narrow the scope of possible mapping functions that you may ultimately choose.**
* **That the conceptualization of machine learning as a search helps to rationalize the use of ensembles, the spot checking of algorithms and the understanding of what is happening when algorithms learn.**

**Kick-start your project with my new book**[**Optimization for Machine Learning**](https://machinelearningmastery.com/optimization-for-machine-learning/)**, including *step-by-step tutorials* and the *Python source code* files for all examples.**

**Problem of Function Approximation**

**Applied machine learning is the development of a learning system to address a specific learning problem.**

**The learning problem is characterized by observations comprised of input data and output data and some unknown but coherent relationship between the two.**

**The goal of the learning system is to learn a generalized mapping between input and output data such that skillful predictions can be made for new instances drawn from the domain where the output variable is unknown.**

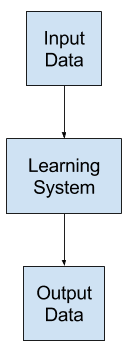
**In statistical learning, a statistical perspective on machine learning, the problem is framed as the learning of a mapping function (*f*) given input data (*X*) and associated output data (*y*).**

|  |  |
| --- | --- |
| **1** | **y = f(X)** |

**We have a sample of *X* and *y* and do our best to come up with a function that approximates *f*, e.g. *fprime*, such that we can make predictions (*yhat*) given new examples (*Xhat*) in the future.**

|  |  |
| --- | --- |
| **1** | **yhat = fprime(Xhat)** |

**As such, applied machine learning can be thought of as the problem of function approximation.**

****

**The learned mapping will be imperfect.**

**The problem of designing and developing a learning system is the problem of learning a useful approximate of the unknown underlying function that maps the input variables to the output variables.**

**We do not know the form of the function, because if we did, we would not need a learning system; we could specify the solution directly.**

**Because we do not know the true underlying function, we must approximate it, meaning we do not know and may never know how close of an approximation the learning system is to the true mapping.**

1. **What are the various goals of machine learning? What is the relationship between these and human learning?**

**Need for machine learning:**

**Machine learning is important because it allows computers to learn from data and improve their performance on specific tasks without being explicitly programmed. This ability to learn from data and adapt to new situations makes machine learning particularly useful for tasks that involve large amounts of data, complex decision-making, and dynamic environments.**

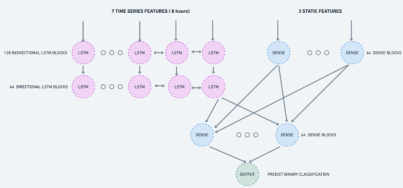
**Here are some specific areas where machine learning is being used:**

* **Predictive modeling: Machine learning can be used to build predictive models that can help businesses make better decisions. For example, machine learning can be used to predict which customers are most likely to buy a particular product, or which patients are most likely to develop a certain disease.**
* **Natural language processing: Machine learning is used to build systems that can understand and interpret human language. This is important for applications such as voice recognition, chatbots, and language translation.**
* **Computer vision: Machine learning is used to build systems that can recognize and interpret images and videos. This is important for applications such as self-driving cars, surveillance systems, and medical imaging.**
* **Fraud detection: Machine learning can be used to detect fraudulent behavior in financial transactions, online advertising, and other areas.**
* **Recommendation systems: Machine learning can be used to build recommendation systems that suggest products, services, or content to users based on their past behavior and preferences.**

**Overall, machine learning has become an essential tool for many businesses and industries, as it enables them to make better use of data, improve their decision-making processes, and deliver more personalized experiences to their customers.**

1. **Illustrate the various elements of machine learning using a real-life illustration.**

**1. Healthcare and medical diagnosis**

****

**Source Omdena**

**Machine learning deals with prognostic and diagnostic issues in medicine and**[**healthcare**](https://www.omdena.com/ai-in-healthcare)**. Disease breakthroughs, patient monitoring and management, medical data analysis, and management of inappropriate medical data are just some of many machine learning examples in healthcare.**

[**Omdena has utilized recurrent neural networks**](https://www.omdena.com/blog/time-series-classification-model-tutorial/)**(RNNs) to combine sequential and static feature modeling to predict cardiac arrest.**

**RNNs are proven to work exceptionally well with time-series-based data. Often in actual life data, supplementary static features may be available, which cannot get directly incorporated into RNNs because of their non-sequential nature. The method described involves adding static features to RNNs to influence the learning process. A previous approach to the problem was implementing several models for each modality and combining them at the prediction level. Combining these two methods into the same model architecture allows the model to learn simultaneously from the static and temporal features.**

**We conclude that the addition of the static features improves the performance of the RNN than would otherwise by using the sequential and static features alone.**

****

**Image Source: Omdena**

[**In the challenge of predicting biological age through AI**](https://www.omdena.com/projects/ai-aging/)**, Humanity and the Omdena team compressed high throughput markers such as activity and other lifestyle action data from the user (e.g. diet, weight, socio-economic status) to develop weighted algorithms predictive of the biological age outcome.**

**The team has built a system that takes in the user attributes and lifestyle actions that are being monitored on one side (activity rates, sleep, meditation, diet, etc.) and uses the ongoing increases or decreases in the user’s Biological Age measure to decide which actions were most effective and in what combinations and when. The system then also matched across users with similar attributes to use the insights and weightings set for one user to affect the weightings given to actions and the combination of actions to another user.**

**Read full case study here:**[**How can AI help people slow their aging down using causal inference**](https://www.omdena.com/blog/causal-inference/)

**2. Face detection in images**

****

[**Image Source**](https://miro.medium.com/max/2000/1*j5V50ennD9MDdbFTEPYlzA.jpeg)

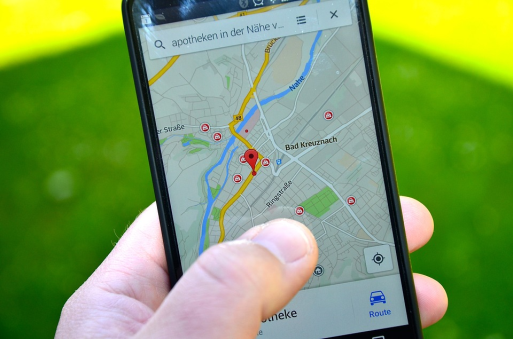
**Machine learning finds its application in face detection amidst non-face objects such as buildings, landscapes, or other human body parts, such as legs or hands. It plays a crucial role in fortifying surveillance techniques by tracking down terrorists and criminals, making the world a safer place.**

****

**Image Source: Demo of Child Growth Monitor application**

**Child Growth Monitor (CGM) is a game-changer application in this space as it replaces traditional methods of anthropometric measurement which are complex, slow, and expensive, frequently resulting in poor data and wrong assessments of the situation. In the challenge**[**Identifying Malnutrshed Children through Computer Vision**](https://www.omdena.com/projects/ai-malnutrition/)**, Omdena and CGM predict the measurement of height, weight, and mid-upper arm circumference (MUAC) of children under age 5 using its open-sourced state-of-the-art**[**neural network algorithms**](https://www.omdena.com/blog/types-of-neural-network-algorithms-in-machine-learning/)**to determine if a child is malnourished or not. The goal of this challenge was to increase the accuracy of CGM’s neural networks’ prediction, so that 90% of children get a height measurement with less than 1cm error.**

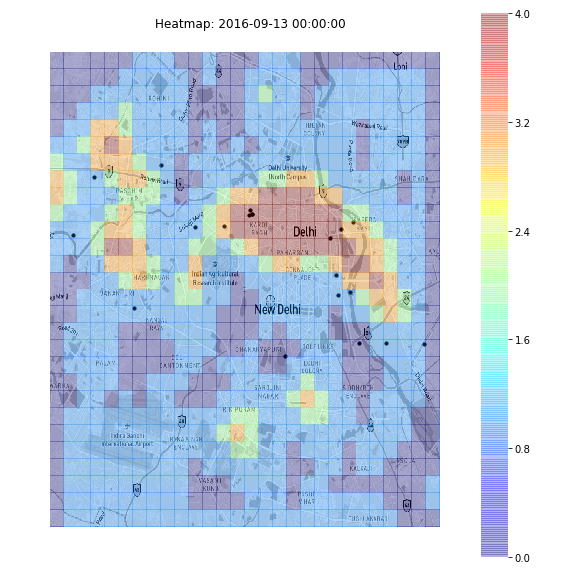
**3. Commute predictions**

****

[**Image Source**](https://149695847.v2.pressablecdn.com/wp-content/uploads/2019/02/maps.png)

**Machine Learning in platforms that use maps and routing ensures punctuality through ML algorithms to calculate the quickest route having less traffic, arrival time, the pick-up location, and the best optimal route to a destination. Machine learning techniques have incorporated a deep learning model to explore transportation traffic, intricate roadway interactions, and environmental elements. It has helped address many traffic bottlenecks, thereby enhancing a nation’s safety, economy, and quality of life. Emergency vehicles like ambulances can find the shortest and quickest way to reach a hospital, saving lives. Besides, people can save time rather than getting stuck in traffic and have a more productive day.**

**4. Public safety**

**[](https://www.omdena.com/blog/heatmap-machine-learning/attachment/figure-2-heatmap-of-crime-intensity-against-women-in-delhi/)**

**Sexual Abuse Hotspots, Source: Omdena**

**Machine learning can improve community safety by preventing, reducing, and responding to crimes.  30 data scientists and machine learning engineers collaborated with an award-winning NGO, Safecity, to predict sexual harassment hotspots through**[**machine learning-driven heatmaps**](https://www.omdena.com/blog/heatmap-machine-learning/)**.**

**5. Agriculture**

****

**Crop Yield Predication, Source: Omdena**

**Machine learning in**[**agriculture**](https://www.omdena.com/ai-in-agriculture)**enables precise and efficient farming with less manpower for high-quality production. Machine learning also provides invaluable insights and recommendations about crops so that farmers can minimize their losses.**

**Using satellite imagery for Google Earth Engine (GEE) images and Jupyter, Omdena built**[**an app for crop yield prediction**](https://www.omdena.com/blog/yield-prediction/)**in Senegal, Africa, that helps to**[**improve food security**](https://www.omdena.com/blog/improving-food-security-and-crop-yield-through-machine-learning/)**.**

**[embedyt] https://www.youtube.com/watch?v=4OnBGkhA4jc[/embedyt]**

****

**Image Source: Omdena**

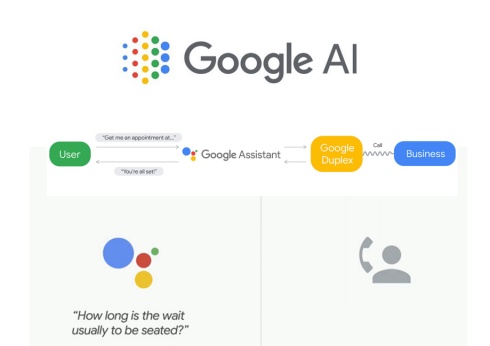
**Another use case in Agriculture is the Omdena challenge with OKO**[**using satellite imagery to detect and assess the damage of armyworms in farming**](https://www.omdena.com/projects/armyworms-assessment/)**, where the team developed an AI pipeline for generating, preprocessing, and training classification algorithms; with a developed web application connected to the deployed model solving the problem of**[**damage assessment of armyworms attack on plants**](https://www.omdena.com/blog/using-satellite-imagery-to-detect-and-assess-the-damage-of-armyworms-in-farming/)**.**

**Using satellite images, the team was able to detect and identify the damage assessment of either or all together (depending on data availability):**

* **Fall Armyworm**
* **Africa Armyworm**
* **Locust Desert, with surge/outbreak in Mali (or Ivory Coast or Ouganda)**

**Learn more about this challenge in our article**[**supervised machine learning for damage assessment in agriculture**](https://www.omdena.com/blog/supervised-machine-learning-damage-assessment-farming/)

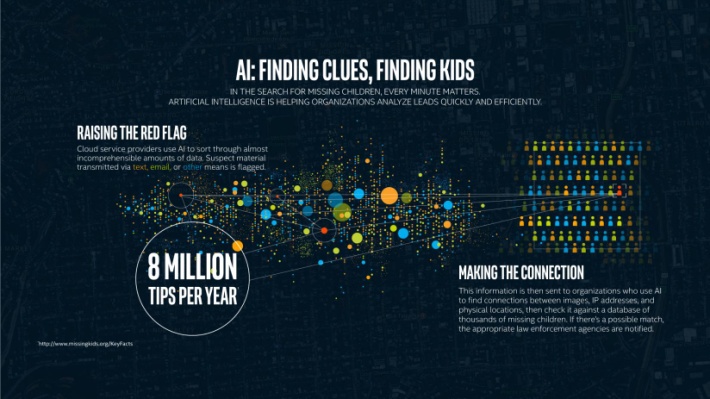
**6. Smart assistants**

****

[**Image Source**](https://miro.medium.com/max/1838/0*CpYsbph_HUKFIVWN.jpg)

**Siri, Alexa, and Google Assistants are just some of the smart assistants used in everyday life to carry out activities like setting reminders, alarms, checking the weather, etc. Voice-based smart assistants have numerous societal benefits as they bring people together, make visually or physically challenged people independent, and help them become more independent. Smart assistants also render a sense of companionship to people who live alone.**

**7. Government industry and policymaking**

****

[**Image Source**](https://www.intel.com/content/dam/www/public/us/en/images/infographics/ai-helps-find-kids-infographic.jpg.rendition.intel.web.1648.927.jpg)

**The use of machine learning helps authorities track and manage the huge amount of data generated by public surveillance devices. Data analysis in real-time for anomalies and threats by law enforcement agencies helps track criminals and missing children. Thus, internet service providers are more successful in identifying instances of suspicious online activity pointing to child exploitation.**

**Another example is where a team of data scientists and ML engineers at, Omdena successfully applied**[**machine learning to enhance public sector transparency**](https://www.omdena.com/projects/public-sector-democratizing/)**by enabling increased access to government contract opportunities.**

**8. Workplace safety**

****

[**Image Source**](https://venturebeat.com/wp-content/uploads/2020/09/Hazzard_Outline.jpg?fit=750%2C362&strip=all)

**Machine learning applications enhance workplace safety by reducing workplace accidents, helping companies detect potentially ill employees as they arrive on-site, and aiding organizations in managing natural disasters.**

**9. Safeguarding the environment**

****

[**Image Source**](https://techieloops.com/wp-content/uploads/2020/12/machine-learning-and-environment-protection.jpg)

**Machine learning algorithms can help in boosting**[**environmental sustainability**](https://www.omdena.com/ai-in-environment-and-conservation/)**. A good example is IBM’s Green Horizon Project, wherein environmental statistics from varied assets and sensors are leveraged to produce pollution forecasts. The aim is to bring down the environmental impact.**

**10. Cyber security**

****

[**Image Source**](https://miro.medium.com/max/5126/1*cm-7cFHBnEKv_RceFXAv3Q.jpeg)

**Applications like PayPal and GPay use machine learning for tracking transactions and differentiating between illegitimate and legitimate transactions. In this way, machine learning maximizes cyber security by preventing online monetary fraud.**

**Besides the above-listed applications, substantial other sectors and areas implement ML technologies.**

[**Omdena**](https://www.omdena.com/)**runs AI Projects with organizations that want to get started with AI, solve a real-world problem, or build deployable solutions within two months.**

**If you want to learn more about us, you can check out all**[**AI projects**](https://www.omdena.com/projects/)**and**[**real-world case studies on Omdena’s blog.**](https://www.omdena.com/blog/)

1. **Provide an example of the abstraction method.**
2. **abstract class arithmetic\_operation {**
3. **abstract void printInfo();**
4. **}**
5. ***// Class add***
6. **class add extends arithmetic\_operation {**
7. ***// class add must override printInfo() method***
8. ***// otherwise, compile-time***
9. ***// exception will be thrown***
10. **void printInfo()**
11. **{**
12. **int a = 3;**
13. **int b = 4;**
14. **System.out.println(a + b);**
15. **}**
16. **}**
17. ***// Class sub***
18. **class sub extends arithmetic\_operation {**
19. ***// class sub must override printInfo() method***
20. ***// otherwise, compile-time***
21. ***// exception will be thrown***
22. **void printInfo()**
23. **{**
24. **int c = 4;**
25. **int d = 5;**
26. **System.out.println(c - d);**
27. **}**
28. **}**
29. ***// Driver Class***
30. **class abstraction {**
31. ***// Main Function***
32. **public static void main(String args[])**
33. **{**
34. **arithmetic\_operation n = new add();**
35. **n.printInfo();**
36. **arithmetic\_operation y = new sub();**
37. **y.printInfo();**
38. **}**
39. **}**
40. **Output**
41. **7**
42. **-1**

**9. What is the concept of generalization? What function does it play in the machine learning process?**

**The term ‘generalization’ refers to the model’s capability to adapt and react properly to previously unseen, new data, which has been drawn from the same distribution as the one used to build the model. In other words, generalization examines how well a model can digest new data and make correct predictions after getting trained on a training set.**

**How well a model is able to generalize is the key to its success. If you train a model too well on training data, it will be incapable of generalizing. In such cases, it will end up making erroneous predictions when it’s given new data. This would make the model ineffective even though it’s capable of making correct predictions for the training data set. This is known as overfitting. The inverse (underfitting) is also true, which happens when you train a model with inadequate data. In cases of underfitting, your model would fail to make accurate predictions even with the training data. This would make the model just as useless as overfitting.**

***You* would ideally want to choose a model that stands at the sweet spot between [overfitting](https://en.wikipedia.org/wiki/Overfitting) and underfitting. To achieve this goal, you can track the performance of a machine learning algorithm over time as it’s working with a set of training data. You can plot both the skill on the training data and the skill on a test dataset that you’ve held back from the training process. As the algorithm learns over time, the level of error for the model on the training data would decrease and so would the error on the test dataset. Training the model for too long would cause a continual decrease in the performance on the training dataset due to overfitting. At the same time, due to the model’s decreasing ability for generalization, the error for the test set would start to increase again. The sweet spot is the point just before the error on the test dataset begins to rise where the model shows good skill on both the training dataset as well as the unseen test dataset.**

**To limit overfitting in a machine learning algorithm, two additional techniques that you can use are:**

* **Using a resampling method to estimate the accuracy of the model**
* **Holding back a validation dataset**

**So, during your machine learning training, keep an eye on generalization when estimating your model accuracy on unseen data.**

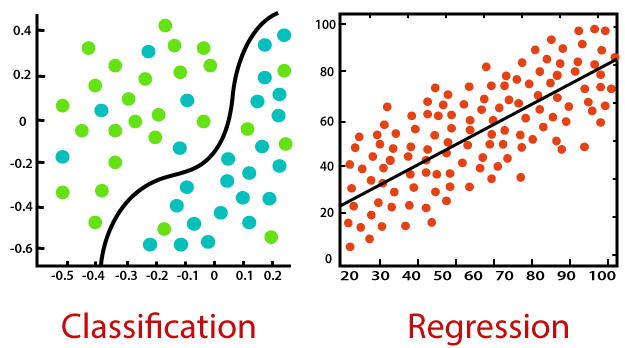
**10. What is classification, exactly? What are the main distinctions between classification and regression?**

**Regression vs. Classification in Machine Learning**

**Regression and Classification algorithms are Supervised Learning algorithms. Both the algorithms are used for prediction in Machine learning and work with the labeled datasets. But the difference between both is how they are used for different machine learning problems.**

**The main difference between Regression and Classification algorithms that Regression algorithms are used to predict the continuous values such as price, salary, age, etc. and Classification algorithms are used to predict/Classify the discrete values such as Male or Female, True or False, Spam or Not Spam, etc.**

**Consider the below diagram:**

****

**Classification:**

**Classification is a process of finding a function which helps in dividing the dataset into classes based on different parameters. In Classification, a computer program is trained on the training dataset and based on that training, it categorizes the data into different classes.**

**PauseNext**

**Mute**

**Current TimeÂ 0:00**

**/**

**DurationÂ 18:10**

**Loaded: 0.37%**

**Â**

**Fullscreen**

**The task of the classification algorithm is to find the mapping function to map the input(x) to the discrete output(y).**

**Example: The best example to understand the Classification problem is Email Spam Detection. The model is trained on the basis of millions of emails on different parameters, and whenever it receives a new email, it identifies whether the email is spam or not. If the email is spam, then it is moved to the Spam folder.**

**Types of ML Classification Algorithms:**

**Classification Algorithms can be further divided into the following types:**

**ADVERTISEMENT**

* **Logistic Regression**
* **K-Nearest Neighbours**
* **Support Vector Machines**
* **Kernel SVM**
* **Naïve Bayes**
* **Decision Tree Classification**
* **Random Forest Classification**

**Regression:**

**Regression is a process of finding the correlations between dependent and independent variables. It helps in predicting the continuous variables such as prediction of Market Trends, prediction of House prices, etc.**

**The task of the Regression algorithm is to find the mapping function to map the input variable(x) to the continuous output variable(y).**

**Example: Suppose we want to do weather forecasting, so for this, we will use the Regression algorithm. In weather prediction, the model is trained on the past data, and once the training is completed, it can easily predict the weather for future days.**

**Types of Regression Algorithm:**

* **Simple Linear Regression**
* **Multiple Linear Regression**
* **Polynomial Regression**
* **Support Vector Regression**
* **Decision Tree Regression**
* **Random Forest Regression**

**Difference between Regression and Classification**

|  |  |
| --- | --- |
| **Regression Algorithm** | **Classification Algorithm** |
| **In Regression, the output variable must be of continuous nature or real value.** | **In Classification, the output variable must be a discrete value.** |
| **The task of the regression algorithm is to map the input value (x) with the continuous output variable(y).** | **The task of the classification algorithm is to map the input value(x) with the discrete output variable(y).** |
| **Regression Algorithms are used with continuous data.** | **Classification Algorithms are used with discrete data.** |
| **In Regression, we try to find the best fit line, which can predict the output more accurately.** | **In Classification, we try to find the decision boundary, which can divide the dataset into different classes.** |
| **Regression algorithms can be used to solve the regression problems such as Weather Prediction, House price prediction, etc.** | **Classification Algorithms can be used to solve classification problems such as Identification of spam emails, Speech Recognition, Identification of cancer cells, etc.** |
| **The regression Algorithm can be further divided into Linear and Non-linear Regression.** | **The Classification algorithms can be divided into Binary Classifier and Multi-class Classifier.** |

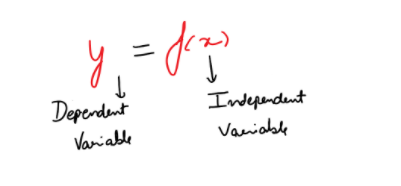
**11. What is regression, and how does it work? Give an example of a real-world problem that was solved using regression.**

**What is Regression?**

**It is a statistical method that can help to analyze and understand the relationship b/w two or more variables of interest. Regression Analysis adopts the process that helps to understand which factors are mandatory, which factors can be ignored, and how these factors are influencing each other.**

**For the regression analysis there are two basic terms that are required to understand:**

* **Dependent Variable: It is the variable that we are trying to understand or forecast.**
* **Independent Variable: It is a variable that acts as the factor that can influence the overall analysis or target variable and provide us with information regarding the relationship of the variables with the target variable.**

****

**Key types of Regressions**

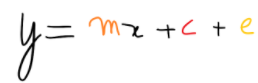
**There are different types of Regression Analysis techniques and these techniques can be selected on the basis of many other factors. These factors are the types of the target variable, the number of independent variables, and shape of the regression line.  
Key types of regression:**

1. **Linear Regression**
2. **Logistic Regression**
3. **Polynomial Regression**
4. **Ridge Regression**
5. **Lasso Regression**

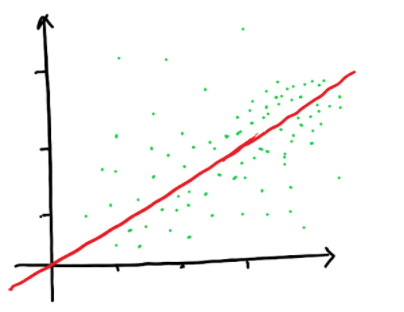
**Linear Regression Model**

**It is one of the most basic Regression models which contains the predictable and a dependent variable related linearly to each other. In case the data you have contains more than one independent variable, then linear regression models become the multi-linear regression model.**

**Equation that is used to denote the Linear Regression:**

****

**Where m is the slope of the line, c is an intercept and e is the error in the model.**

****

***(Linear Regression Example Illustration)***

**Pros and Cons of Regression Model**

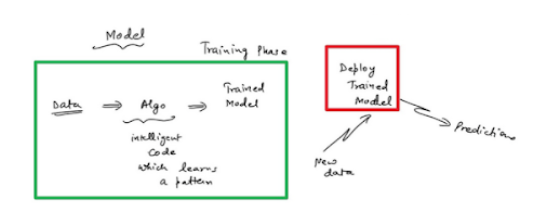
**Pros of the Linear Regression model are:**

1. **It is really simple to implement**
2. **It is Highly Interpretable**
3. **Predictions can be done fastly**
4. **It is easy to train the data**
5. **Less resources are required to implement this model**

**Cons of the Linear Regression Model:**

1. **We made the fundamental assumption that there is a linear relationship between feature and label**
2. **This model doesn’t work with the outliers**
3. **Many features than interpretability is low**
4. **Categorical data is inappropriate for this model**
5. **Handling missing data creates problem for this model**
6. **Multicollinearity is a big issue**

**Model Training and Evaluation**

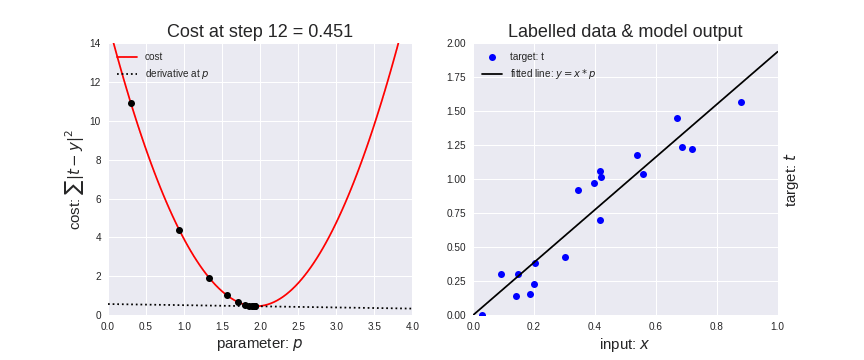
****

**We have different models, which can be developed by supplying the data to intelligent codes i.e. Algorithms which can help to learn the patterns. Then we have the Trained models with us. This whole process of supplying data to algorithms and getting the model is done under the Training Phase. Once we have our deployed Training Model, we supplied the new data (which was not used earlier) to that model so as to make the predictions.**

**This is how models are created and made the predictions.**

**Cost Evaluation using Gradient Descent**

**It enables a model to learn the gradient or *direction* that the model should take in order to reduce errors (differences between actual y and predicted y). Direction in the simple linear regression refers to how the model parameters b0 and b1 should be tweaked or corrected to further reduce the cost function. As the model iterates, it gradually converges towards a minimum where further tweaks to the parameters produce little or zero changes in the loss — also referred to as convergence.**

****

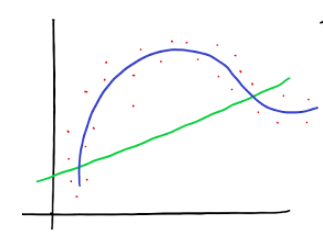
**At this point, the model has optimized the weights such that they minimize the cost function. This process is integral to the ML process because it greatly expedites the learning process — you can think of it as a means of receiving corrective feedback on how to improve upon your previous performance.**

**Overfitting and Underfitting**

**Overfitting**

**A model is said to be overfitted when we train it with a lot of data and now it starts learning from the noise and inaccurate data entries. Because this model does not categorize the data accurately due to many details and noise. This problem occurs when the model is too complex. In regression models, overfitting can possibly produce the R-Squared values, regression coefficients, and p-values.**

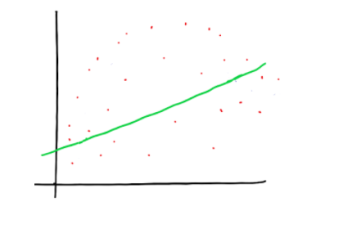
**In nutshell, Overfitting is the High variance and Low bias**

****

**Underfitting**

**A model is said to be underfitt when it cannot capture the underlying trend of data. It destroys the accuracy, as its occurrence signifies that our algorithm or model does not fit the data very well. It usually happens when we have less amount of data for training the model.**

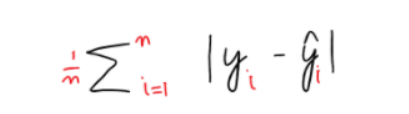
**In nutshell, Underfitting is High bias and Low variance**

****

**Model Evaluation Metrics**

**Here are the three common model evaluation metrics for regression problems:**

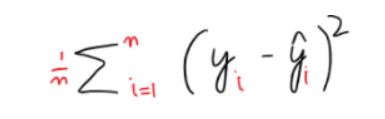
1. **Mean Absolute Error (MAE): It is the mean of absolute value of errors. It can be denoted as**

****

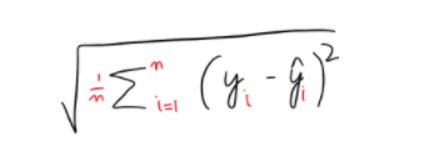
* **MAE is the easiest to understand because it’s the average error.**

**2. Mean Square Error (MSE): It is the mean of the squared error. It can be denoted as:**

* **MSE is more popular than MAE, because MSE “punishes” larger errors, which tends to be useful in the real world. Also, MSE is continuous and differentiable, making it easier to use than MAE for optimization.**

****

**3. Root Mean Square Error (RMSE): It is the square root of the mean of the squared root. It can be denoted as:**

****

**RMSE is even more popular than MSE, because RMSE is interpretable in the “y” units.**

**12. Describe the clustering mechanism in detail.**

**his kind of data cannot be analyzed using supervised learning algorithms. We need the help of unsupervised algorithms. One of the most popular type of analysis under unsupervised learning is**[**Cluster analysis**](https://www.geeksforgeeks.org/data-mining-cluster-analysis/)**. When the goal is to group similar data points in a dataset, then we use cluster analysis. In practical situations, we can use cluster analysis for**[**customer segmentation**](https://www.geeksforgeeks.org/customer-segmentation-using-unsupervised-machine-learning-in-python/)**for targeted advertisements, or in medical imaging to find unknown or new infected areas and many more use cases that we will discuss further in this article.**

**Table of Content**

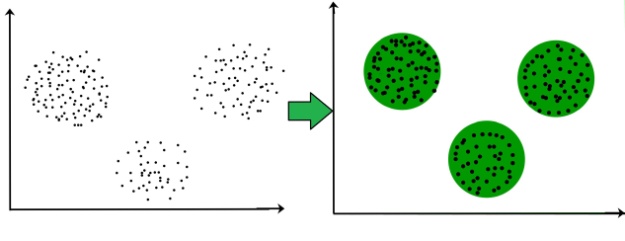
* [**What is Clustering ?**](https://www.geeksforgeeks.org/clustering-in-machine-learning/#what-is-clustering-)
* [**Types of Clustering**](https://www.geeksforgeeks.org/clustering-in-machine-learning/#types-of-clustering)
* [**Uses of Clustering**](https://www.geeksforgeeks.org/clustering-in-machine-learning/#uses-of-clustering)
* [**Types of Clustering Algorithms**](https://www.geeksforgeeks.org/clustering-in-machine-learning/#types-of-clustering-algorithms)
* [**Applications of Clustering in different fields:**](https://www.geeksforgeeks.org/clustering-in-machine-learning/#applications-of-clustering-in-different-fields)
* [**Frequently Asked Questions (FAQs) on Clustering**](https://www.geeksforgeeks.org/clustering-in-machine-learning/#frequently-asked-questions-faqs-on-clustering)

**What is Clustering ?**

**The task of grouping data points based on their similarity with each other is called Clustering or Cluster Analysis. This method is defined under the branch of**[**Unsupervised Learning**](https://www.geeksforgeeks.org/supervised-unsupervised-learning/)**, which aims at gaining insights from unlabelled data points, that is, unlike**[**supervised learning**](https://www.geeksforgeeks.org/supervised-unsupervised-learning/)**we don’t have a target variable.**

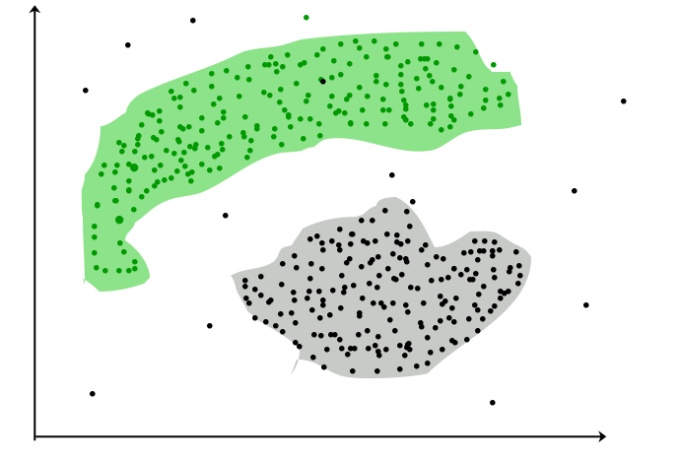
**Clustering aims at forming groups of homogeneous data points from a heterogeneous dataset. It evaluates the similarity based on a metric like Euclidean distance, Cosine similarity, Manhattan distance, etc. and then group the points with highest similarity score together.**

**For Example, In the graph given below, we can clearly see that there are 3 circular clusters forming on the basis of distance.**

****

**Now it is not necessary that the clusters formed must be circular in shape. The shape of clusters can be arbitrary. There are many algortihms that work well with detecting arbitrary shaped clusters.**

**For example, In the below given graph we can see that the clusters formed are not circular in shape.**

****

**Types of Clustering**

**Broadly speaking, there are 2 types of clustering that can be performed to group similar data points:**

* **Hard Clustering: In this type of clustering, each data point belongs to a cluster completely or not. For example, Let’s say there are 4 data point and we have to cluster them into 2 clusters. So each data point will either belong to cluster 1 or cluster 2.**

| **Data Points** | **Clusters** |
| --- | --- |
| **A** | **C1** |
| **B** | **C2** |
| **C** | **C2** |
| **D** | **C1** |

* **Soft Clustering: In this type of clustering, instead of assigning each data point into a separate cluster, a probability or likelihood of that point being that cluster is evaluated. For example, Let’s say there are 4 data point and we have to cluster them into 2 clusters. So we will be evaluating a probability of a data point belonging to both clusters. This probability is calculated for all data points.**

| **Data Points** | **Probability of C1** | **Probability of C2** |
| --- | --- | --- |
| **A** | **0.91** | **0.09** |
| **B** | **0.3** | **0.7** |
| **C** | **0.17** | **0.83** |
| **D** | **1** | **0** |

**Uses of Clustering**

**Now before we begin with types of clustering algorithms, we will go through the use cases of Clustering algorithms. Clustering algorithms are majorly used for:**

* [**Market Segmentation**](https://www.geeksforgeeks.org/customer-segmentation-using-unsupervised-machine-learning-in-python/)**– Businesses use clustering to group their customers and use targeted advertisements to attract more audience.**
* [**Market Basket Analysis**](https://www.geeksforgeeks.org/market-basket-analysis-in-data-mining/)**– Shop owners analyze their sales and figure out which items are majorly bought together by the customers. For example, In USA, according to a study diapers and beers were usually bought together by fathers.**
* [**Social Network Analysis**](https://www.geeksforgeeks.org/social-network-analysis-using-r-programming/)**– Social media sites use your data to understand your browsing behaviour and provide you with targeted friend recommendations or content recommendations.**
* **Medical Imaging – Doctors use Clustering to find out diseased areas in diagnostic images like X-rays.**
* [**Anomaly Detection**](https://www.geeksforgeeks.org/machine-learning-for-anomaly-detection/)**– To find outliers in a stream of real-time dataset or forecasting fraudulent transactions we can use clustering to identify them.**
* **Simplify working with large datasets – Each cluster is given a cluster ID after clustering is complete. Now, you may reduce a feature set’s whole feature set into its cluster ID. Clustering is effective when it can represent a complicated case with a straightforward cluster ID. Using the same principle, clustering data can make complex datasets simpler.**

**There are many more use cases for clustering but there are some of the major and common use cases of clustering. Moving forward we will be discussing Clustering Algorithms that will help you perform the above tasks.**

**Types of Clustering Algorithms**

**At the surface level, clustering helps in the analysis of unstructured data. Graphing, the shortest distance, and the density of the data points are a few of the elements that influence cluster formation. Clustering is the process of determining how related the objects are based on a metric called the similarity measure. Similarity metrics are easier to locate in smaller sets of features. It gets harder to create similarity measures as the number of features increases. Depending on the type of clustering algorithm being utilized in data mining, several techniques are employed to group the data from the datasets. In this part, the clustering techniques are described. Various types of clustering algorithms are:**

1. **Centroid-based Clustering (Partitioning methods)**
2. **Density-based Clustering (Model-based methods)**
3. **Connectivity-based Clustering (Hierarchical clustering)**
4. **Distribution-based Clustering**

**We will be going through each of these types in brief.**

**1. [Centroid-based Clustering (Partitioning methods)](https://www.geeksforgeeks.org/partitioning-method-k-mean-in-data-mining/)**

**Partitioning methods are the most easiest clustering algorithms. They group data points on the basis of their closeness. Generally, the similarity measure chosen for these algorithms are Euclidian distance, Manhattan Distance or Minkowski Distance. The datasets are separated into a predetermined number of clusters, and each cluster is referenced by a vector of values. When compared to the vector value, the input data variable shows no difference and joins the cluster.**

**The primary drawback for these algorithms is the requirement that we establish the number of clusters, “k,” either intuitively or scientifically (using the Elbow Method) before any clustering machine learning system starts allocating the data points. Despite this, it is still the most popular type of clustering.**[**K-means**](https://www.geeksforgeeks.org/k-means-clustering-introduction/)**and**[**K-medoids**](https://www.geeksforgeeks.org/ml-k-medoids-clustering-with-example/)**clustering are some examples of this type clustering.**

**2.**[**Density-based Clustering (Model-based methods)**](https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/)

**Density-based clustering, a model-based method, finds groups based on the density of data points. Contrary to centroid-based clustering, which requires that the number of clusters be predefined and is sensitive to initialization, density-based clustering determines the number of clusters automatically and is less susceptible to beginning positions. They are great at handling clusters of different sizes and forms, making them ideally suited for datasets with irregularly shaped or overlapping clusters. These methods manage both dense and sparse data regions by focusing on local density and can distinguish clusters with a variety of morphologies.**

**In contrast, centroid-based grouping, like k-means, has trouble finding arbitrary shaped clusters. Due to its preset number of cluster requirements and extreme sensitivity to the initial positioning of centroids, the outcomes can vary. Furthermore, the tendency of centroid-based approaches to produce spherical or convex clusters restricts their capacity to handle complicated or irregularly shaped clusters. In conclusion, density-based clustering overcomes the drawbacks of centroid-based techniques by autonomously choosing cluster sizes, being resilient to initialization, and successfully capturing clusters of various sizes and forms. The most popular density-based clustering algorithm is**[**DBSCAN**](https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/)**.**

**3.**[**Connectivity-based Clustering (Hierarchical clustering)**](https://www.geeksforgeeks.org/ml-hierarchical-clustering-agglomerative-and-divisive-clustering/)

**A method for assembling related data points into hierarchical clusters is called hierarchical clustering. Each data point is initially taken into account as a separate cluster, which is subsequently combined with the clusters that are the most similar to form one large cluster that contains all of the data points.**

**Think about how you may arrange a collection of items based on how similar they are. Each object begins as its own cluster at the base of the tree when using hierarchical clustering, which creates a dendrogram, a tree-like structure. The closest pairings of clusters are then combined into larger clusters after the algorithm examines how similar the objects are to one another. When every object is in one cluster at the top of the tree, the merging process has finished. Exploring various granularity levels is one of the fun things about hierarchical clustering. To obtain a given number of clusters, you can select to cut the [dendrogram](https://www.geeksforgeeks.org/scipy-cluster-hierarchy-dendrogram/)at a particular height. The more similar two objects are within a cluster, the closer they are. It’s comparable to classifying items according to their family trees, where the nearest relatives are clustered together and the wider branches signify more general connections. There are 2 approaches for Hierarchical clustering:**

* [**Divisive Clustering**](https://www.geeksforgeeks.org/difference-between-agglomerative-clustering-and-divisive-clustering/)**: It follows a top-down approach, here we consider all data points to be part one big cluster and then this cluster is divide into smaller groups.**
* [**Agglomerative Clustering**](https://www.geeksforgeeks.org/difference-between-agglomerative-clustering-and-divisive-clustering/)**: It follows a bottom-up approach, here we consider all data points to be part of individual clusters and then these clusters are clubbed together to make one big cluster with all data points.**

**4. Distribution-based Clustering**

**Using distribution-based clustering, data points are generated and organized according to their propensity to fall into the same probability distribution (such as a Gaussian, binomial, or other) within the data. The data elements are grouped using a probability-based distribution that is based on statistical distributions. Included are data objects that have a higher likelihood of being in the cluster. A data point is less likely to be included in a cluster the further it is from the cluster’s central point, which exists in every cluster.**

**A notable drawback of density and boundary-based approaches is the need to specify the clusters a priori for some algorithms, and primarily the definition of the cluster form for the bulk of algorithms. There must be at least one tuning or hyper-parameter selected, and while doing so should be simple, getting it wrong could have unanticipated repercussions. Distribution-based clustering has a definite advantage over proximity and centroid-based clustering approaches in terms of flexibility, accuracy, and cluster structure. The key issue is that, in order to avoid [overfitting](https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/), many clustering methods only work with simulated or manufactured data, or when the bulk of the data points certainly belong to a preset distribution. The most popular distribution-based clustering algorithm is**[**Gaussian Mixture Model**](https://www.geeksforgeeks.org/gaussian-mixture-model/)**.**

**Applications of Clustering in different fields:**

1. **Marketing: It can be used to characterize & discover customer segments for marketing purposes.**
2. **Biology: It can be used for classification among different species of plants and animals.**
3. **Libraries: It is used in clustering different books on the basis of topics and information.**
4. **Insurance: It is used to acknowledge the customers, their policies and identifying the frauds.**
5. **City Planning: It is used to make groups of houses and to study their values based on their geographical locations and other factors present.**
6. **Earthquake studies: By learning the earthquake-affected areas we can determine the dangerous zones.**
7. **Image Processing: Clustering can be used to group similar images together, classify images based on content, and identify patterns in image data.**
8. **Genetics: Clustering is used to group genes that have similar expression patterns and identify gene networks that work together in biological processes.**
9. **Finance: Clustering is used to identify market segments based on customer behavior, identify patterns in stock market data, and analyze risk in investment portfolios.**
10. **Customer Service: Clustering is used to group customer inquiries and complaints into categories, identify common issues, and develop targeted solutions.**
11. **Manufacturing: Clustering is used to group similar products together, optimize production processes, and identify defects in manufacturing processes.**
12. **Medical diagnosis: Clustering is used to group patients with similar symptoms or diseases, which helps in making accurate diagnoses and identifying effective treatments.**
13. **Fraud detection: Clustering is used to identify suspicious patterns or anomalies in financial transactions, which can help in detecting fraud or other financial crimes.**
14. **Traffic analysis: Clustering is used to group similar patterns of traffic data, such as peak hours, routes, and speeds, which can help in improving transportation planning and infrastructure.**
15. **Social network analysis: Clustering is used to identify communities or groups within social networks, which can help in understanding social behavior, influence, and trends.**
16. **Cybersecurity: Clustering is used to group similar patterns of network traffic or system behavior, which can help in detecting and preventing cyberattacks.**
17. **Climate analysis: Clustering is used to group similar patterns of climate data, such as temperature, precipitation, and wind, which can help in understanding climate change and its impact on the environment.**
18. **Sports analysis: Clustering is used to group similar patterns of player or team performance data, which can help in analyzing player or team strengths and weaknesses and making strategic decisions.**
19. **Crime analysis: Clustering is used to group similar patterns of crime data, such as location, time, and type, which can help in identifying crime hotspots, predicting future crime trends, and improving crime prevention strategies.**

**13. Make brief observations on two of the following topics:**

**i. Machine learning algorithms are used**

### ****1. Linear Regression****

**Linear regression is a simple algorithm used to map the linear relationship between input features and a continuous target variable. It works by fitting a line to the data and then using the line to predict new values.**

### ****2. Logistic Regression****

**Logistic regression is an extension of linear regression that is used for classification tasks to estimate the likelihood that an instance belongs to a specific class.**

### ****3. SVM (Support Vector Machine)****

**SVMs are supervised learning algorithms that can perform classification and regression tasks. It finds a hyperplane that best separates classes in feature space.**

### ****4. KNN (K-nearest Neighbour)****

**KNN is a non-parametric technique that can be used for classification as well as regression. It works by identifying the k most similar data points to a new data point and then predicting the label of the new data point using the labels of those data points.**

### ****5. Decision Tree****

**Decision trees are a type of supervised learning technique that can be used for classification as well as regression. It operates by segmenting the data into smaller and smaller groups until each group can be classified or predicted with high degree of accuracy.**

### ****6. Random Forest****

**Random forests are a type of ensemble learning method that employs a set of decision trees to make predictions by aggregating predictions from individual trees. It improves the precision and resilience of single decision trees. It can be used for both classification and regression tasks.**

### ****7. Naive Bayes****

**Naive Bayes is a probabilistic classifier based on Bayes’ theorem that is used for classification tasks. It works by assuming that the features of a data point are independent of each other.**

### ****8. PCA (Principal Component Analysis)****

**PCA is a dimensionality reduction technique used to transform data into a lower-dimensional space while retaining as much variance as possible. It works by finding the directions in the data that contain the most variation, and then projecting the data onto those directions.**

### ****9. Apriori algorithms****

**Apriori algorithm is a traditional data mining technique for  association rules mining in transactional databases or datasets. It is designed to uncover links and patterns between things that regularly co-occur in transactions. Apriori detects frequent itemsets, which are groups of items that appear together in transactions with a given minimum support level.**

### ****10. K-Means Clustering****

**K-Means clustering is an unsupervised learning approach that can be used to group together data points. It works by finding k clusters in the data so that the data points in each cluster are as similar to each other as feasible while remaining as distinct as possible from the data points in other clusters.**

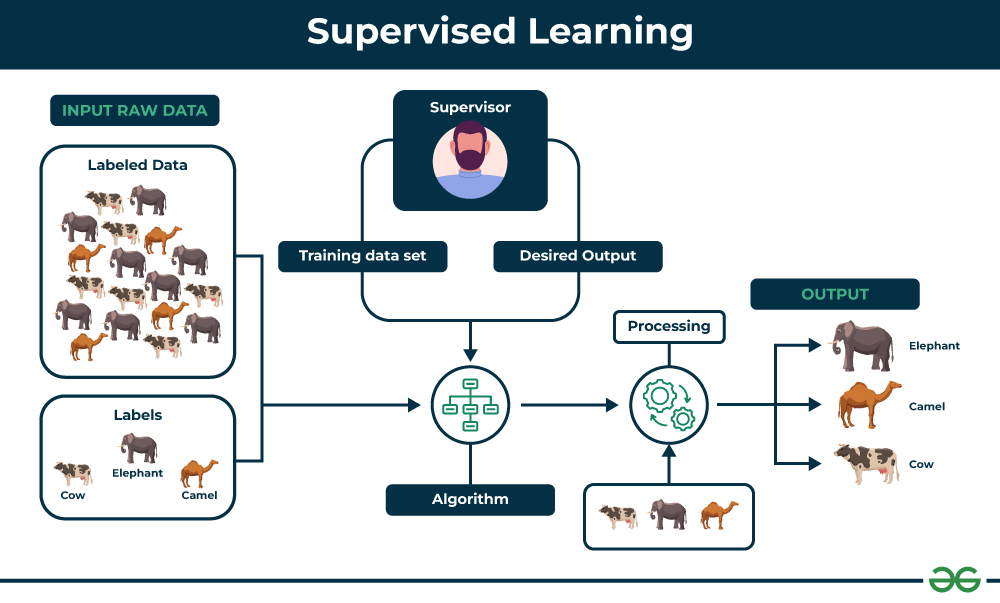
**ii. Studying under supervision**

**What is Supervised learning?**

**Supervised learning is a type of**[**machine learning algorithm**](https://www.geeksforgeeks.org/machine-learning-algorithms/)**that learns from labeled data. Labeled data is data that has been tagged with a correct answer or classification.**

**Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Supervised learning is when we teach or train the machine using data that is well-labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.**

**For example, a labeled dataset of images of Elephant, Camel and Cow would have each image tagged with either “Elephant” , “Camel”or “Cow.”**

****

**Key Points:**

* **Supervised learning involves training a machine from labeled data.**
* **Labeled data consists of examples with the correct answer or classification.**
* **The machine learns the relationship between inputs (fruit images) and outputs (fruit labels).**
* **The trained machine can then make predictions on new, unlabeled data.**

**Example:**

**Let’s say you have a fruit basket that you want to identify. The machine would first analyze the image to extract features such as its shape, color, and texture. Then, it would compare these features to the features of the fruits it has already learned about. If the new image’s features are most similar to those of an apple, the machine would predict that the fruit is an apple.**

**For instance, suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all the different fruits one by one like this:**

* **If the shape of the object is rounded and has a depression at the top, is red in color, then it will be labeled as –Apple.**
* **If the shape of the object is a long curving cylinder having Green-Yellow color, then it will be labeled as –Banana.**

**Now suppose after training the data, you have given a new separate fruit, say Banana from the basket, and asked to identify it.**

**Since the machine has already learned the things from previous data and this time has to use it wisely. It will first classify the fruit with its shape and color and would confirm the fruit name as BANANA and put it in the Banana category. Thus the machine learns the things from training data(basket containing fruits) and then applies the knowledge to test data(new fruit).**

**iii. Studying without supervision**

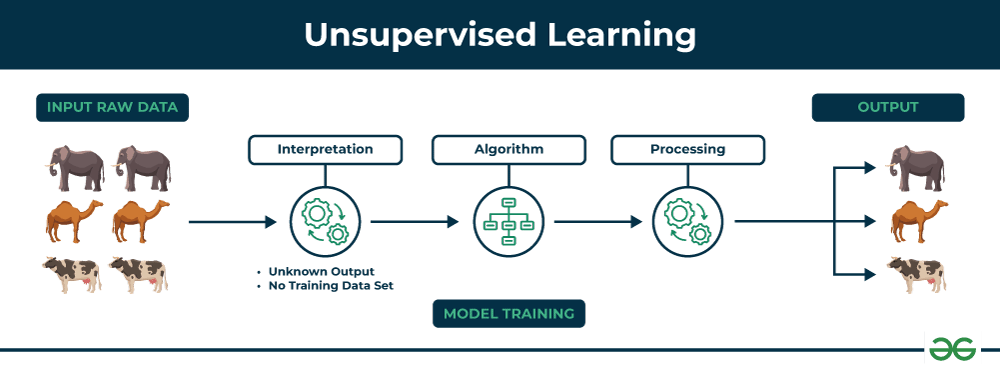
**What is Unsupervised learning?**

**Unsupervised learning is a type of machine learning that learns from unlabeled data. This means that the data does not have any pre-existing labels or categories. The goal of unsupervised learning is to discover patterns and relationships in the data without any explicit guidance.**

**Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.**

**Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.**

**You can use unsupervised learning to examine the animal data that has been gathered and distinguish between several groups according to the traits and actions of the animals. These groupings might correspond to various animal species, providing you to categorize the creatures without depending on labels that already exist.**

****

**Key Points**

* **Unsupervised learning allows the model to discover patterns and relationships in unlabeled data.**
* **Clustering algorithms group similar data points together based on their inherent characteristics.**
* **Feature extraction captures essential information from the data, enabling the model to make meaningful distinctions.**
* **Label association assigns categories to the clusters based on the extracted patterns and characteristics.**

**Example**

**Imagine you have a machine learning model trained on a large dataset of unlabeled images, containing both dogs and cats. The model has never seen an image of a dog or cat before, and it has no pre-existing labels or categories for these animals. Your task is to use unsupervised learning to identify the dogs and cats in a new, unseen image.**

**For instance, suppose it is given an image having both dogs and cats which it has never seen.**

**Thus the machine has no idea about the features of dogs and cats so we can’t categorize it as ‘dogs and cats ‘. But it can categorize them according to their similarities, patterns, and differences, i.e., we can easily categorize the above picture into two parts. The first may contain all pics having dogs in them and the second part may contain all pics having cats in them. Here you didn’t learn anything before, which means no training data or examples.**

**It allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with unlabelled data.**

**iv. Reinforcement learning is a form of learning based on positive reinforcement.**

**Reinforcement learning is an area of Machine Learning. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation. Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience.**

**Reinforcement Learning (RL) is the science of decision making. It is about learning the optimal behavior in an environment to obtain maximum reward. In RL, the data is accumulated from machine learning systems that use a trial-and-error method. Data is not part of the input that we would find in supervised or unsupervised machine learning.**

**Reinforcement learning uses algorithms that learn from outcomes and decide which action to take next. After each action, the algorithm receives feedback that helps it determine whether the choice it made was correct, neutral or incorrect. It is a good technique to use for automated systems that have to make a lot of small decisions without human guidance.**

**Reinforcement learning is an autonomous, self-teaching system that essentially learns by trial and error. It performs actions with the aim of maximizing rewards, or in other words, it is learning by doing in order to achieve the best outcomes.**

**Example:**

**The problem is as follows: We have an agent and a reward, with many hurdles in between. The agent is supposed to find the best possible path to reach the reward. The following problem explains the problem more easily.**

**The above image shows the robot, diamond, and fire. The goal of the robot is to get the reward that is the diamond and avoid the hurdles that are fired. The robot learns by trying all the possible paths and then choosing the path which gives him the reward with the least hurdles. Each right step will give the robot a reward and each wrong step will subtract the reward of the robot. The total reward will be calculated when it reaches the final reward that is the diamond.   
Main points in Reinforcement learning –**

* **Input: The input should be an initial state from which the model will start**
* **Output: There are many possible outputs as there are a variety of solutions to a particular problem**
* **Training: The training is based upon the input, The model will return a state and the user will decide to reward or punish the model based on its output.**
* **The model keeps continues to learn.**
* **The best solution is decided based on the maximum reward.**