Artificial Intelligence, Machine Learning, and Data Science

Samatrix Consulting Pvt Ltd

Samatrix.io

Artificial Intelligence

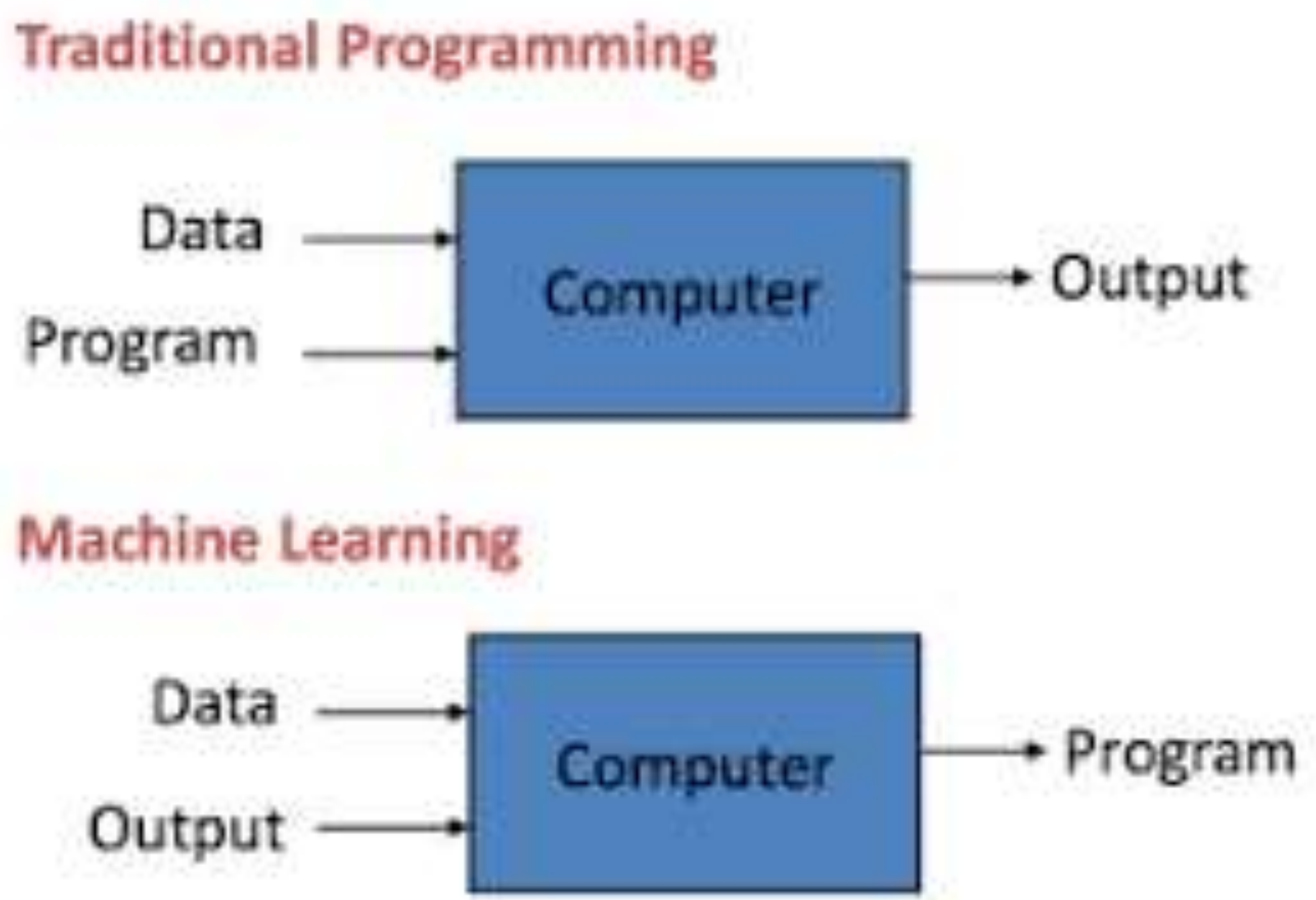
- IBM Deep Blue Chess Program
- Electronic Game Characters (Sims)

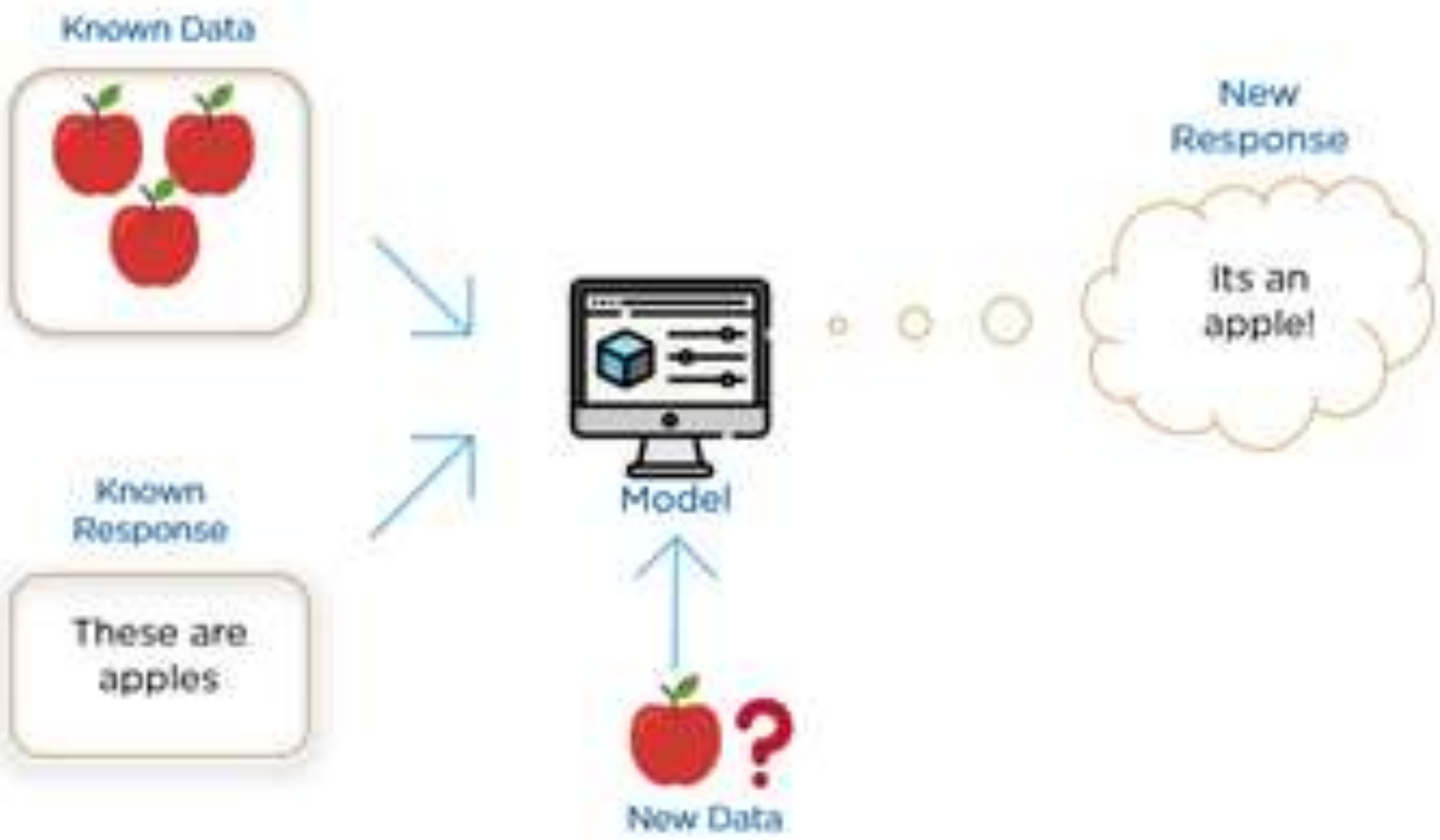
Machine Learning

- IBM Watson
- Google Search Algorithm
- Amazon Recommendations
- Email SPAM filter

Deep Learning

- AlphaGo
- Natural Speech Recognition
- Waymo Level 4 Automated Driving System

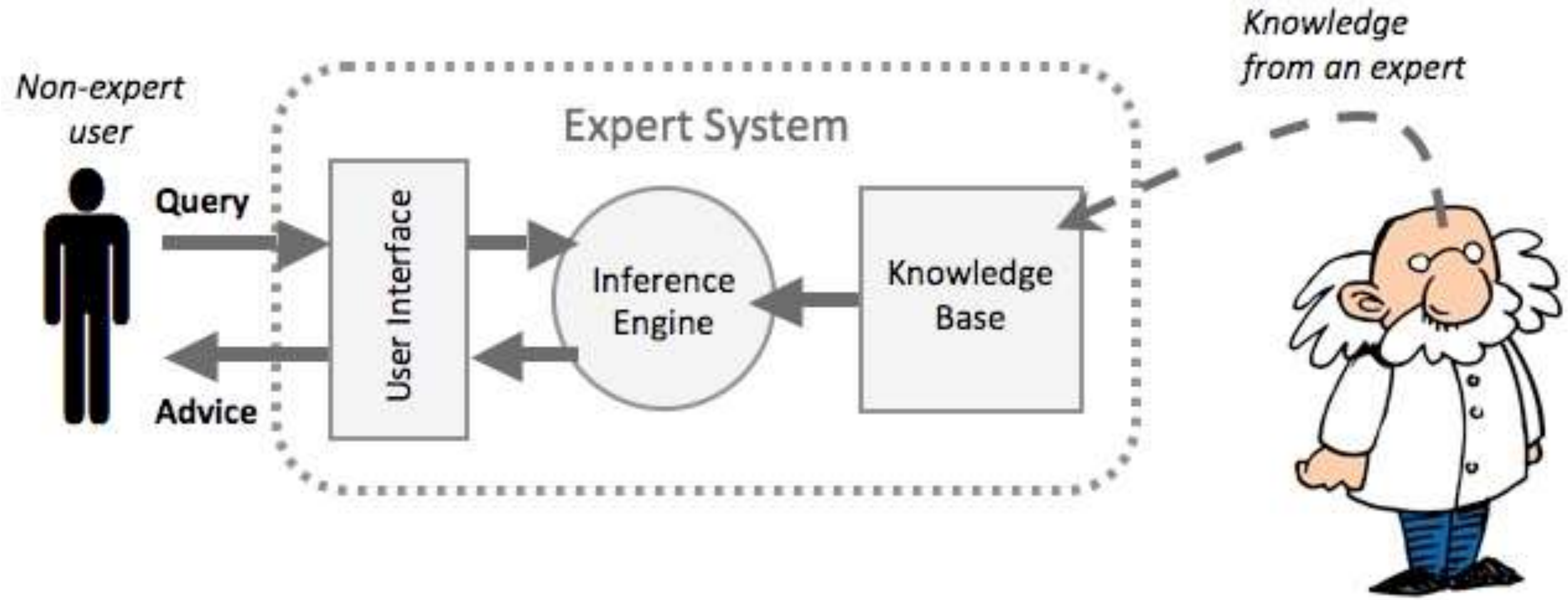

Traditional Programming
Data
Program
Computer
Output
Machine Learning
Data
Output
Computer
Program


Known Data
Known Response
These are apples
Model
New Data
New Response
Its an apple!

Introduction to Machine Learning

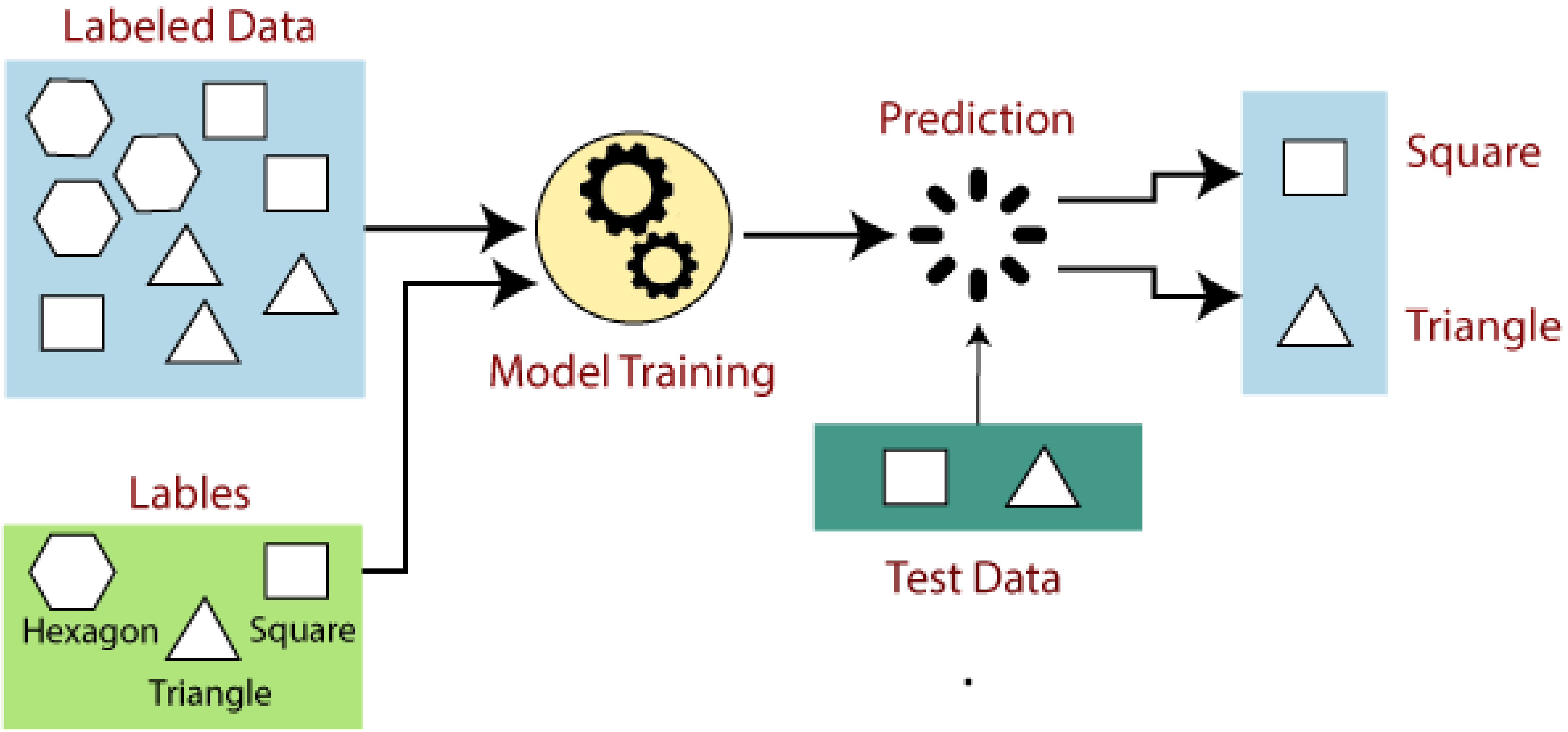
What is Machine Learning

- Machine learning is one of the most important technical approaches to AI and the basis of many recent advances and commercial applications of AI.
- Machine learning is a form of AI that enables a system to learn from data rather than through explicit programming.
- Machine learning uses a variety of algorithms that iteratively learn from data to improve, describe data, and predict outcomes.
- Machine learning helps construct computer systems that automatically improve through experience.
- Machine learning algorithms train a system by showing it examples of desired input-output behaviour than to program it manually by anticipating the desired response for all possible inputs.


Non-expert
user
Query
Advice
Expert System
User Interface
Inference
Engine
Knowledge
Base
Knowledge
from an expert

Machine Learning vs Expert Systems

- Modern machine learning is a statistical process that starts with a body of data and tries to derive a rule or procedure that explains the data or can predict future data.
- This approach—learning from data—contrasts with the older “expert system” approach to AI, in which programmers sit down with human domain experts to learn the rules and criteria used to make decisions and translate those rules into software code.
- An expert system aims to emulate the principles used by human experts, whereas machine learning relies on statistical methods to find a decision procedure that works well in practice.


Labeled Data
Lables
Hexagon
Square
Triangle
Model Training
Prediction
Test Data
Square
Triangle

Samatrix.io

Machine Learning Usages

- Machine learning has progressed dramatically over the past two decades, from laboratory curiosity to a practical technology in widespread commercial use.
- Within artificial intelligence (AI), machine learning has emerged as the method of choice for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications.

Machine Learning Process

- The practitioner of machine learning divides the historical data set into a training set and a test set.
- Then the practitioner chooses a machine learning model or mathematical structure.
- The model consists of a range of possible decision-making rules with adjustable parameters.
- The practitioner also defines an **objective function** or **loss function** that is used to evaluate the quality of the solution that results from the choice of parameters.
- The objective function will reward the model for closely matching the training set and use of simpler rules.

Training Machine Learning Models

- Training the model is the process of adjusting parameters to maximize the objective function.
- Once a model has been trained, the practitioner can use the test set to evaluate the accuracy and effectiveness of the model.
- The goal of machine learning is to create a trained model that will generalize—it will be accurate not only on examples in the training set, but also on future cases that it has never seen before.
- While many of these models can achieve better-than-human performance on narrow tasks such as image labelling, even the best models can fail in unpredictable ways.

Samatrix.io

Training Machine Learning Models

- Another challenge in using machine learning is that it is typically not possible to extract or generate a straightforward explanation for why a particular trained model is effective.
- Because trained models have a very large number of adjustable parameters—often hundreds of millions or more—training may yield a model that "works," in the sense of matching the data, but is not necessarily the simplest model that works.

Learning From Data

- Machine learning helps model to train on data sets before the deployment.
- Some machine learning models are online whereas some models are offline.
- The online models adapt continuously as new data is analysed.
- Offline machine learning models do not change once they are deployed.
- The iterative process of online models helps model improvise the types of associations between data elements.
- Human may overlook the patterns and associations due to complexity and size of the models.
- Once the model is trained, it can be used in real time to learn from data

History of Machine Learning

History of Machine Learning

- AI and machine learning algorithms are old fields.
- The field of AI dates back to 1950s. Arthur Lee Samuels, an IBM researcher, developed a self-learning program for playing checking, which was one of the earliest machine learning programs.
- He published the paper in the IBM Journal of Research and Development in 1959.
- In last few years due to focus on distributed computing models and cheaper compute and storage, there has been a surge in the fields of AI and machine learning.
- A huge amount of money has been invested in start-up software companies, which has led to major advancements and commercial solutions

Key Market Enablers

- The six key market enablers are
  - **Processors**: Modern processors are more powerful and denser. The density to performance ratio has improved significantly
  - **Storage:** The cost of managing and storing large dataset has reduced significantly. New storage innovations have enabled faster performance. The ability to analyse vastly larger data sets have also improved
  - **Distributed compute processing:** The ability to analyse the complex data in record time has also improved due to the ability of distributed compute processing across cluster of computers

Key Market Enablers

- **Commercial data:** More commercial data such as weather data, social media data, and medical sets data is available as cloud services and well-defined Application Programming Interfaces (APIs).
- **Open-source communities:** Machine learning algorithms have been made available through open-source communities with large user bases. Therefore, there are more resources, frameworks, and libraries that have made development easier.
- **Visualization:** Visualization has gotten more consumable. You don’t need to be a data scientist to interpret results, making use of machine learning broader within many industries.

Big Data

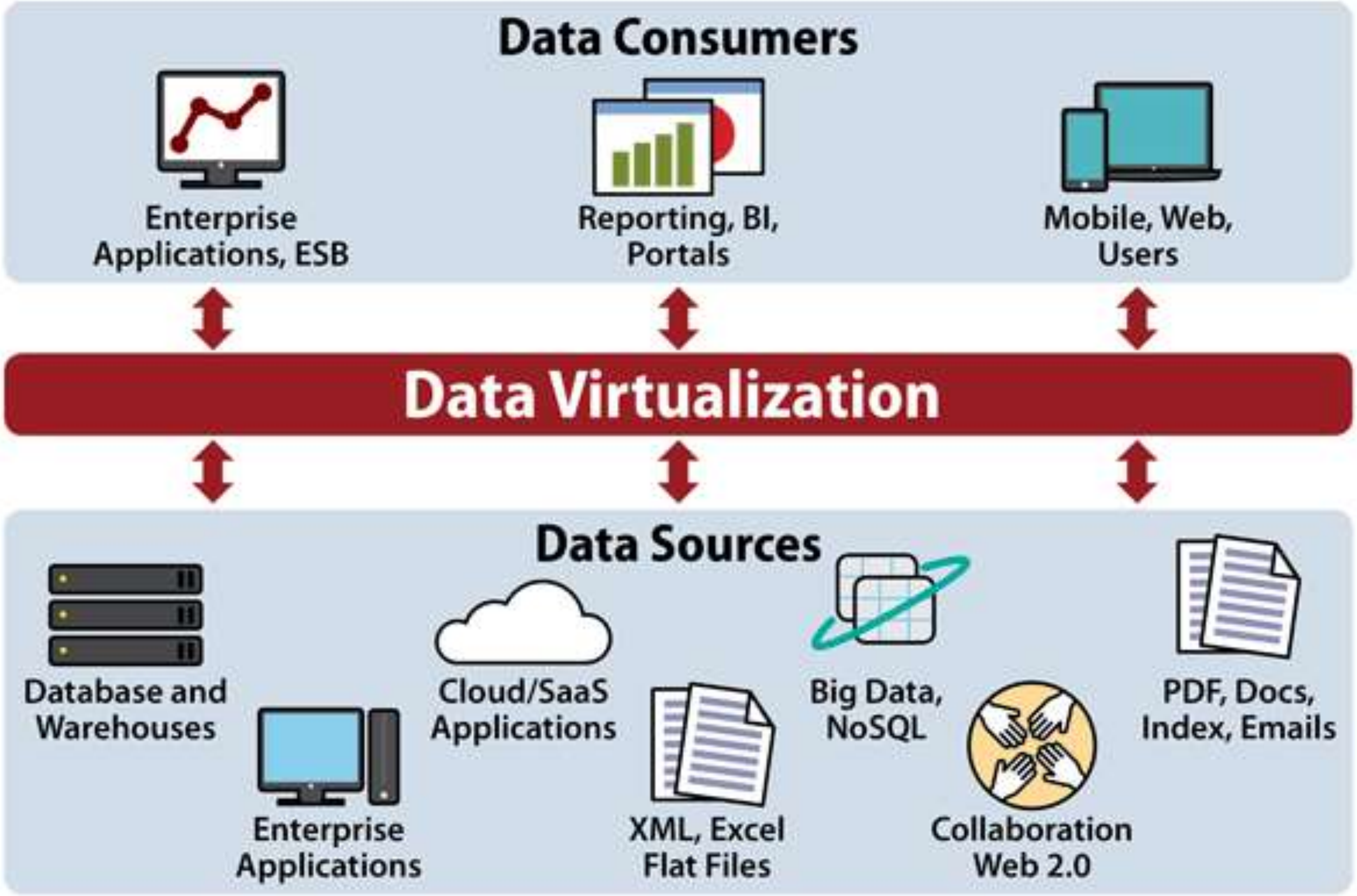
Samatrix.io

Big Data

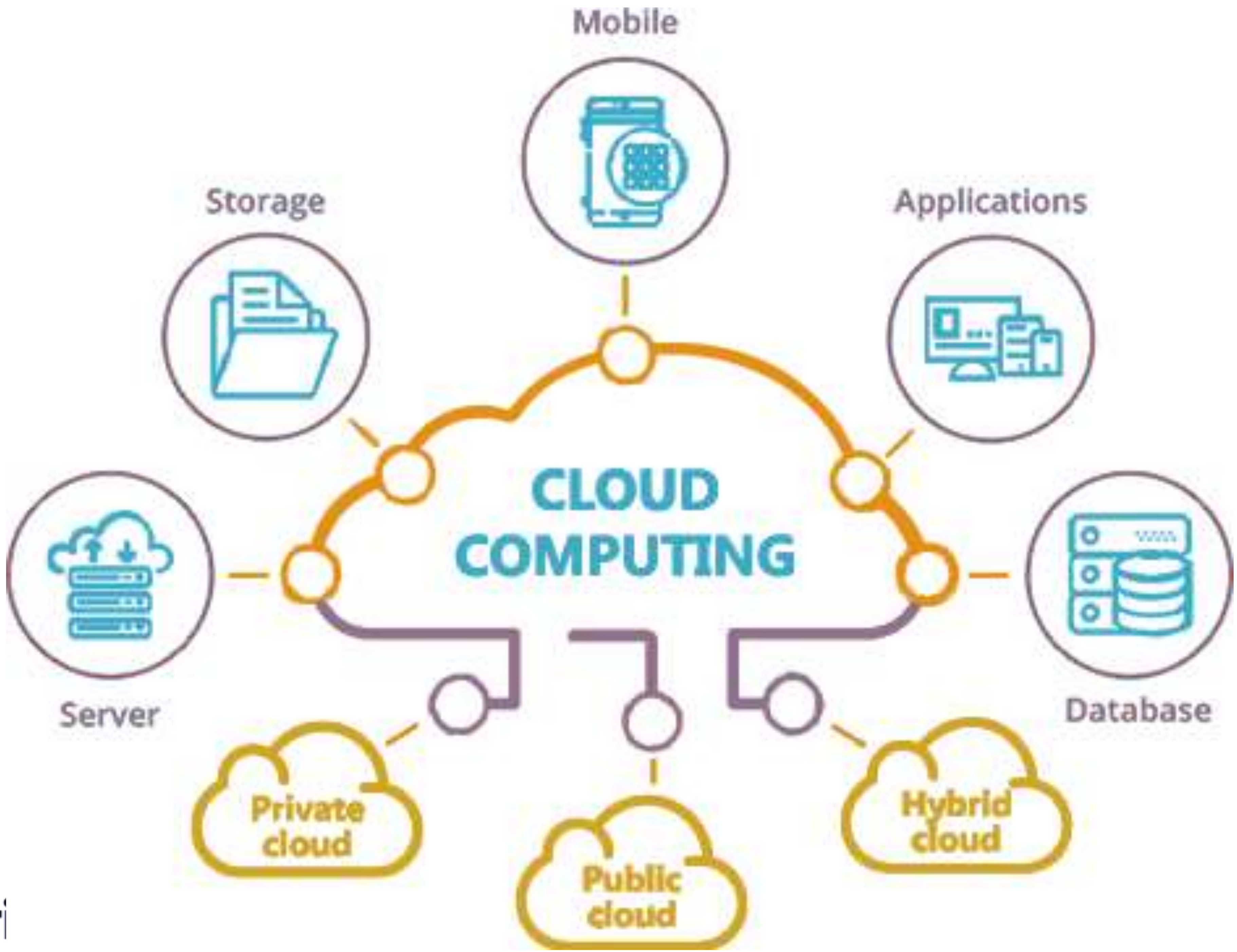
- Any kind of data that has at least one of the following four shared characteristics. They are also known as 4 Vs
  - Extremely large **Volumes** of data
  - The ability to move that data at a high **Velocity** of speed
  - An ever-expanding **Variety** of data sources
  - **Veracity** so that data sources truly represent truth
- Big data incorporates all data, including structured, unstructured, and semi-structured data from email, social media, text streams, images, and machine sensors

Big Data

- To gain right insights using analytics on big data, we need appropriate technology that can gather, store, manage, and manipulate vast amounts data at the right speed and at the right time.
- The evolution of computing technology along with hybrid cloud architectures have enabled the management of immense volumes of data that could have only been handled by supercomputers at great expense.


Data Consumers
Enterprise
Applications, ESB
Reporting, BI,
Portals
Mobile, Web,
Users
Data Virtualization
Data Sources
Database and
Warehouses
Enterprise
Applications
Cloud/SaaS
Applications
XML, Excel
Flat Files
Big Data,
NoSQL
Collaboration
Web 2.0
PDF, Docs,
Index, Emails

Samatrix.io


Mobile
Storage
Applications
CLOUD
COMPUTING
Server
Database
Private
cloud
Public
cloud
Hybrid
cloud

Big Data For Machine Learning

- The big data helps improve the accuracy of machine learning models substantially.
- The low volume of data may lead to misinterpreting a trend or missing an emerging pattern.
- Big data can be very useful for training machine learning models.
- An organization does not have to have big data to use machine learning models.

Big Data For Machine Learning

- But the availability of big data can help improve the accuracy of the models.
- Using big data, the data can be virtualized so that it can be stored in the most efficient and cost-effective manner.
- With the help of big data, the data can be stored on premise or in the cloud.
- The improvements in network speeds and reliability have helped manage massive amount of data at the acceptable speed.

Big Data For Machine Learning

- The big data technologies include data virtualization, parallel processing, distributed file systems, in-memory databases, containerization, and micro-services.
- This combination of technology advances can help organizations address significant business problems.
- Businesses always had large amount of data for decades.
- But the ability to use the richness of data source to gain actionable insights from data was not available.
- Using machine learning models and big data technologies, the organizations can use the data to gain useful insights, anticipate the future, and be prepared for disruption.

Leveraging Machine Learning

Advanced Analytics

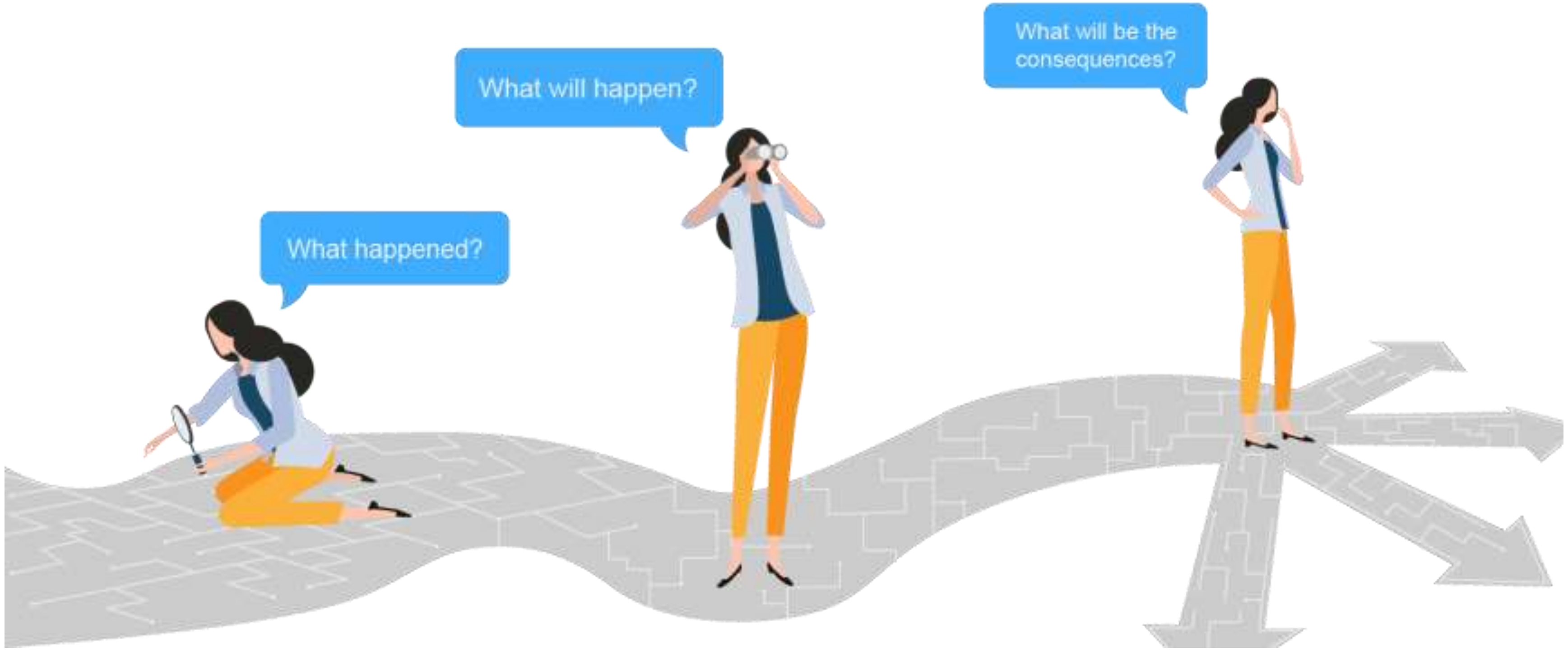
- The role of analytics has been changing in organization's operational processes has been changing for past 30 years.
- The companies have progressed in analytics maturity levels ranging from descriptive analytics to predictive analytics to machine learning and cognitive computing
- Companies have been using analytics to understand both the current status of their business and how they can learn from the past to anticipate the future.
- They can analyse how various actions and events can impact the outcomes. The knowledge from this analysis can help predict future.

Advanced Analytics

- Data scientists and business analysts can make predictions using analytical models that are based on historical data.
- In business environment, unknown factors can impact future outcomes significantly.
- The companies focus on building predictive models that can react and change with the changes in the business environment.

Types of Advanced Analytics

- There are two types of advanced analytics
  - Descriptive analytics
  - Predictive analytics



Samatrix.io

Descriptive Analytics

- Descriptive analytics helps analysts understand the current reality in the business.
- You need to understand the context for historical data in order to understand the current reality of where the business is today.
- Using this approach, the organizations can answer questions such as which product styles are selling better this quarter as compared to last quarter, and which regions are exhibiting the highest/lowest growth.

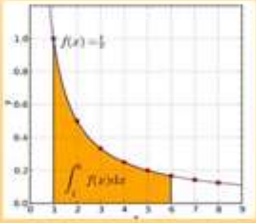
Predictive Analytics

- Predictive analytics helps anticipate changes using the patterns and anomalies within the data.
- Using predictive analytics models, the analyst integrates various related data sources to predict outcome.
- Predictive analytics uses machine learning algorithms to gain insights.
- The predictive analytics tools require new data that can reflect the business changes.
- The addition of new data helps improve the business's ability to anticipate subtle changes in customer preferences, price erosion, market changes, and other factors that will impact the future of business outcomes.

Predictive Analytics

- With a predictive model, you predict future. For example, you can answer the following types of questions
  - How improved web experience can entice a customer to buy frequently?
  - How a stock or a portfolio will perform based on factors such as international news and internal financial factors?
  - Which combination of drugs will provide the best outcome for this cancer patient based on the specific characteristics of the tumor and genetic sequencing?

Machine Learning and Statistics

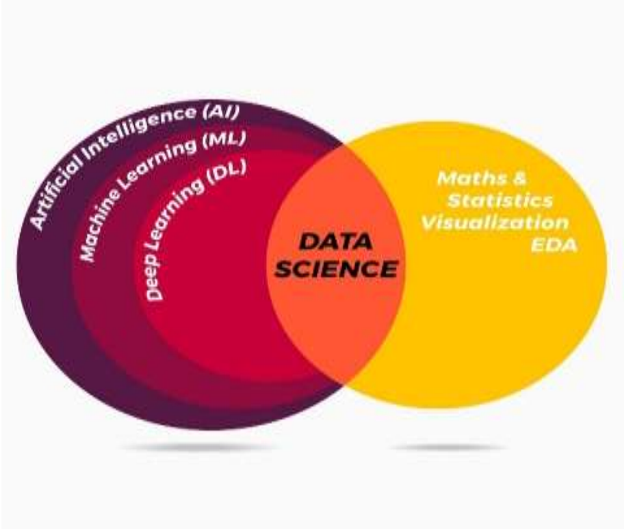


$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

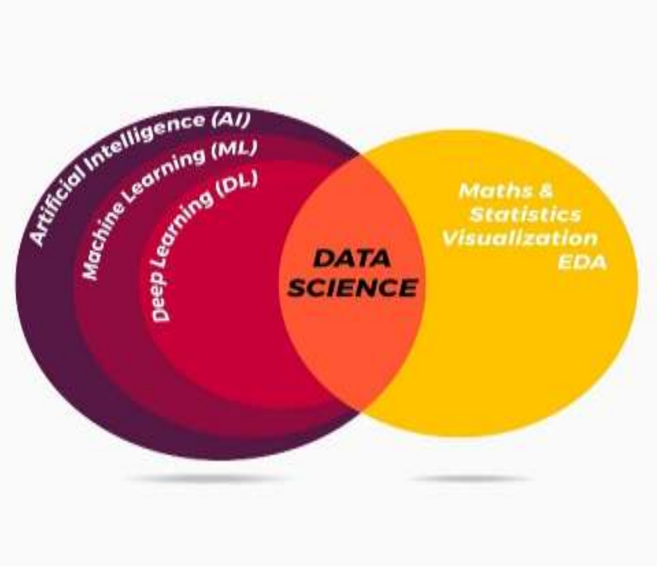
	Statistics +	Machine Learning =	Statistical Learning
Subfield of...	Mathematics	Computer Science (AI)	Statistics & Machine Learning
Focus on...	Building models with explicitly programmed instructions	Creating systems that learn from data	Sets of tools for modeling and understanding complex data
Purpose	Inferences; Relationships between variables	Optimization; Prediction accuracy	Building statistical models for prediction; understanding data
Prior assumptions about data	Some knowledge about population usually required	None	Some knowledge about population may be required
Dimensionality of data	Usually applied to low-dimensional data	Usually applied to high dimensional data; ML learns from data	Usually applied to high dimensional data
Knowledge overlap	No ML knowledge required	Some stats knowledge usually needed: stats is basis for algorithms	Knowledge of statistics and ML required

Machine Learning and Statistics

- Statistics is the discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data. Two main statistical methods used in data analysis are:
    - **Descriptive statistics** summarizes data from a sample using indexes such as the mean or standard deviation
    - **Inferential statistics** draws conclusions from data that are subject to random variation (e.g., observational errors, sampling variation)



Machine Learning and Statistics

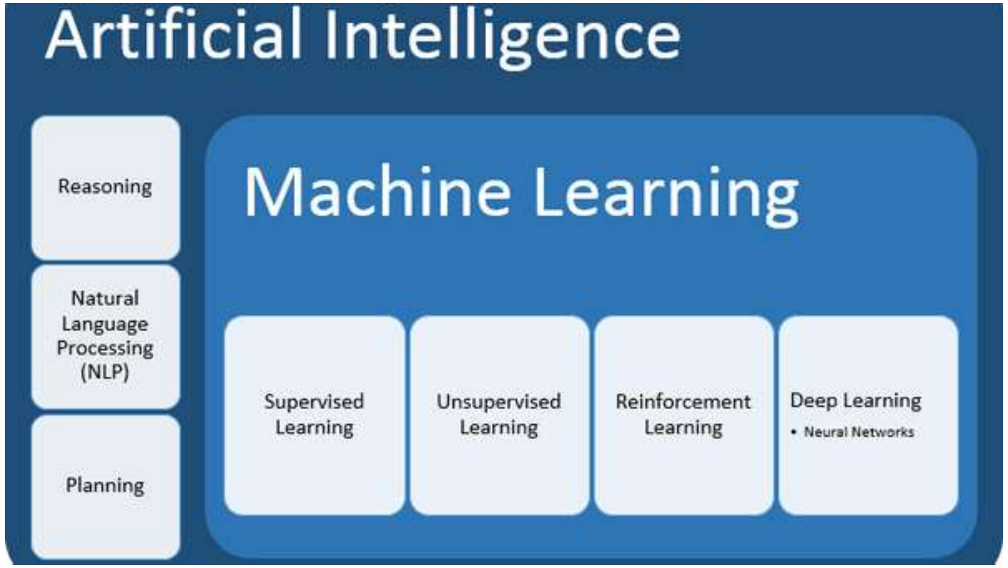


- Machine learning models leverage statistical algorithms to predict analytics.
- The discipline of statistics, data mining, and machine learning have a role in understanding data and characteristics of the data sets.
- They are used in finding relationships and patterns in the data.
- Hence there is a significant overlap.
- The tools and techniques of statistics and machine learning are used to solve business problems
- Many machine learning algorithms are rooted in classical statistical analysis. Data scientists combine the expertise in statistics and machine learning to use all disciplines in collaboration.

AI and Machine Learning

AI and Machine Learning

- AI can be used to describe systems that can "think."
- For example, thermostats that learn your preference or applications that can identify people and what they are doing in photographs can be thought of as AI systems
- There are four main subsets of AI. Reasoning, natural language processing (NLP), planning, and machine learning



Samatrix.io

Reasoning System

- **Reasoning** system is a software system that generates conclusions from available knowledge and data using logical techniques.
- If the data is incomplete, the reasoning helps fill in the blanks using connected data.
- For example, given that the system has enough data and the following question is asked, “What is a safe internal temperature for eating a drumstick?” The system can tell the answer is 165 degrees.

Reasoning System

- The logic chain would be as follows:
  - A drumstick refers to a chicken leg as opposed to a part of a musical instrument
  - A chicken leg contains dark chicken meat
  - The dark chicken meat requires temperature of 165 degrees to cook
  - So, the answer is 165 degree.

- In this example the system was not explicitly trained to predict the safe internal temperature of chicken drumstick.
- The system used the existing data and knowledge to fill in data gaps.

Natural Language Processing

- **Natural Language Processing (NLP)** is the ability of a computer program to understand human language both written text and human speech.
- The idea of giving computers the ability to process human language is as old as the idea of the computer themselves.
- The goal of the NLP is to get computers perform tasks involving human language.

Natural Language Processing

- It includes tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech.
- There are two main reasons why we want our computers to process the natural languages: first, to communicate with humans, and second, to acquire information from written language

Natural Language Processing

- There are over a trillion of pages on web.
- Almost all of them in human language.
- Computer program that wants to do knowledge acquisition, needs to understand ambiguous, messy language that human use.
- We use information seeking tasks such as text classification, information retrieval, and information extraction.
- We use language models to address the tasks.
- Language models predict the probability distribution of language expressions.

Planning

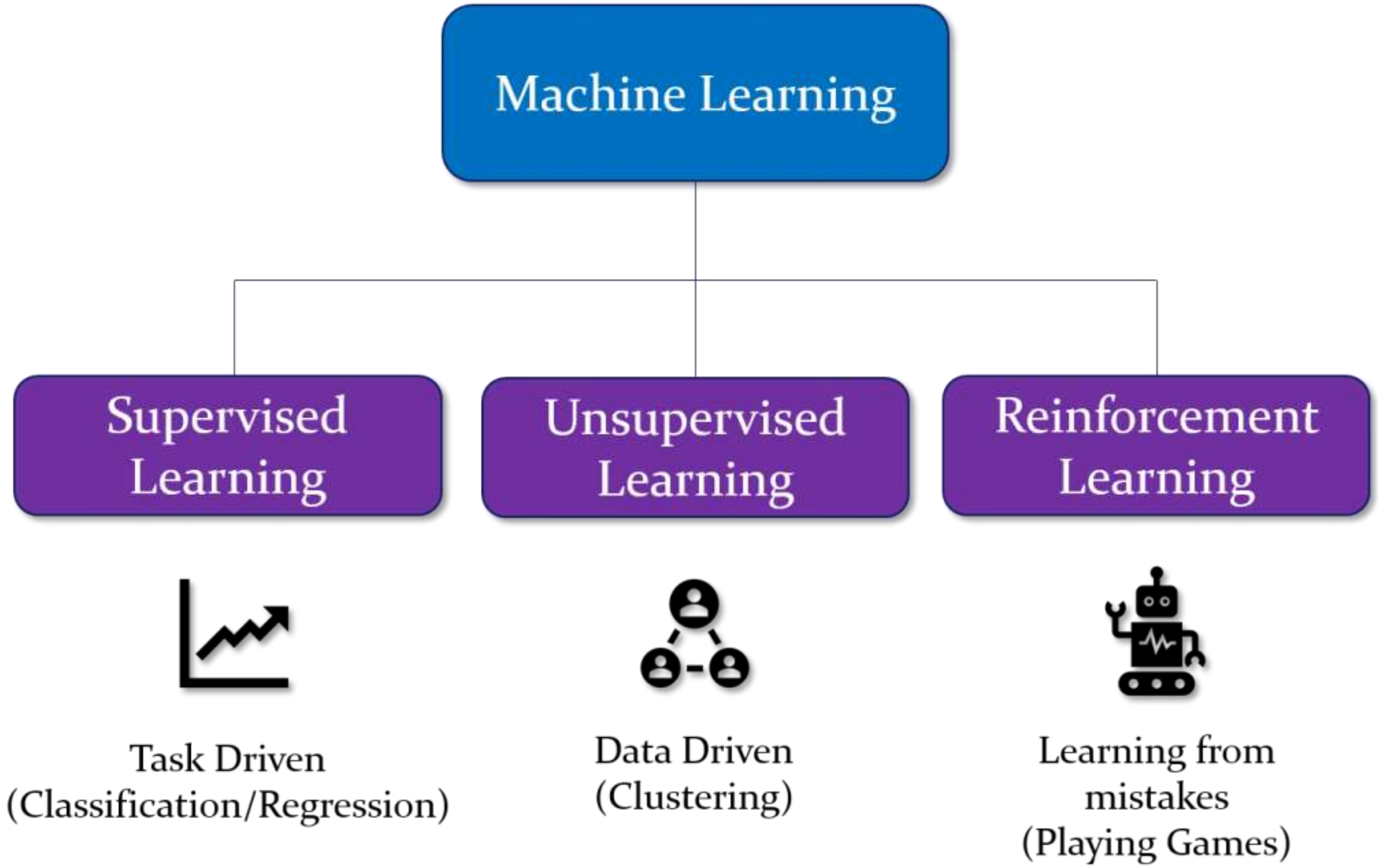
- **Planning** is about how an agent achieves its goals.
- To achieve anything but the simplest goals, an agent must reason about its future.
- Because an agent does not usually achieve its goals in one step, what it should do at any time depends on what it will do in the future.
- What it will do in the future depends on the state it is in, which, in turn, depends on what it has done in the past.

Planning

- Automated planning is the ability of the intelligent system to act autonomously and flexibly to construct the sequence of action to reach the final goal.
- Rather than a pre-programmed decision-making process that goes from A to B to C to reach a final output, automated planning is complex and requires a system to adapt based on the context surrounding the given challenge.

Types of Machine Learning

Types of Machine Learning



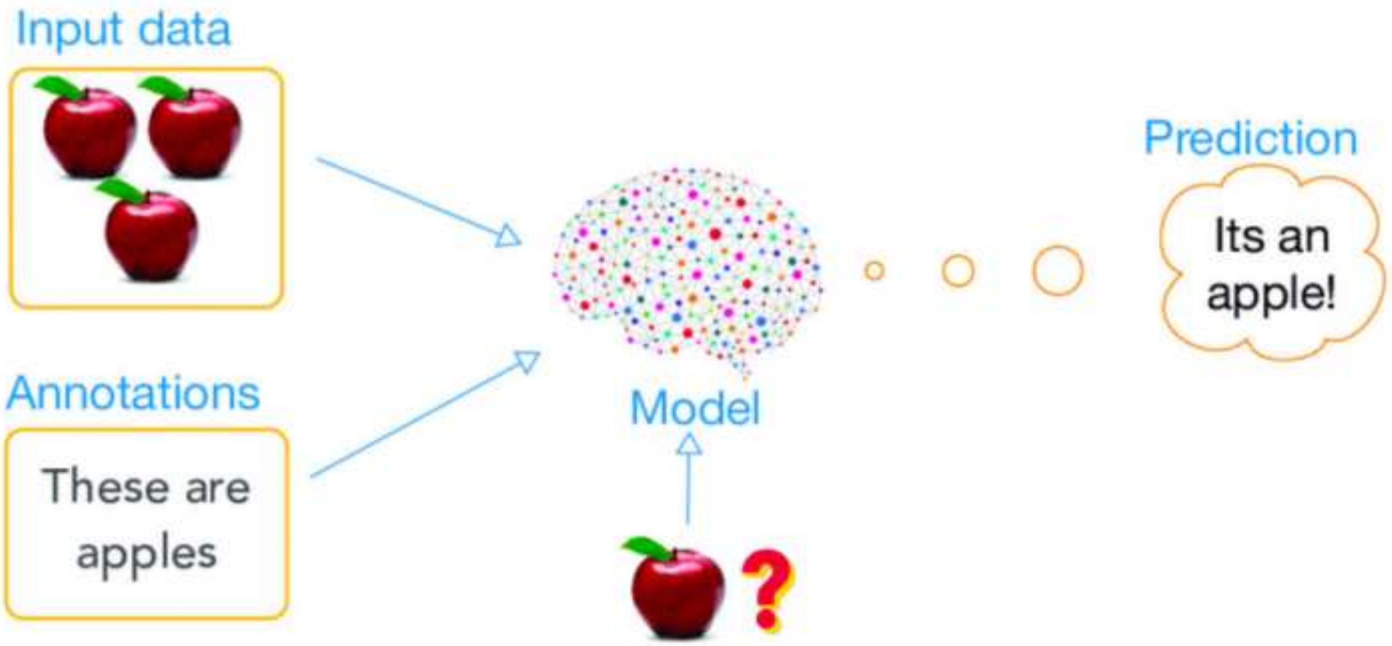
Types of Machine Learning

- Learning is the ability of an agent to improve its behavior based on experience. This could mean the following:
- The range of behaviors is expanded; the agent can do more.
- The accuracy on tasks is improved; the agent can do things better.
- The speed is improved; the agent can do things faster.

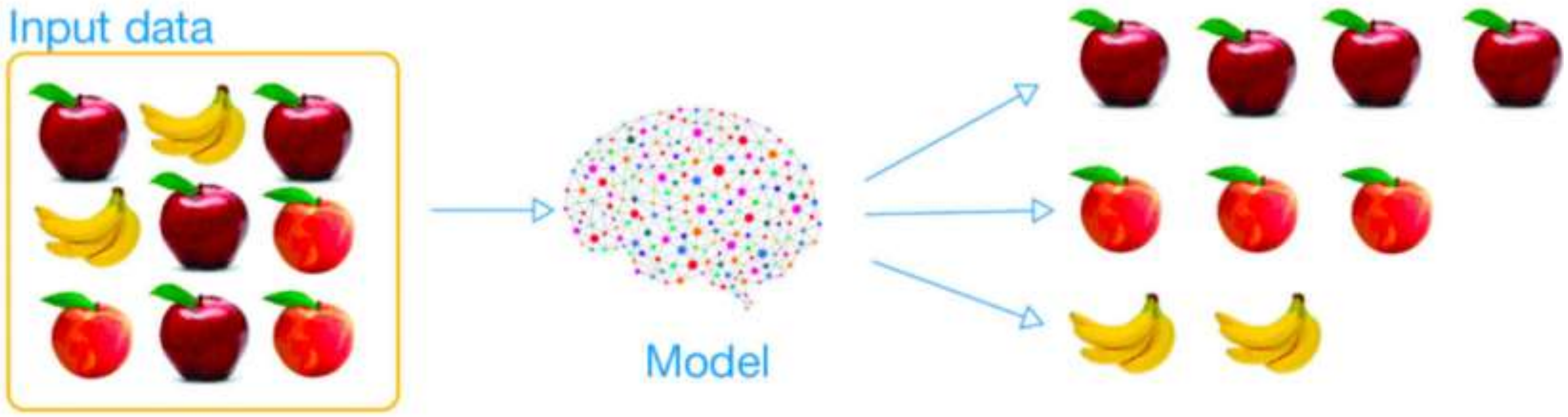
Types of Machine Learning

- There are four main types of learning
  - Supervised learning
  - Unsupervised learning
  - Self-supervised learning
  - Reinforcement learning

supervised learning



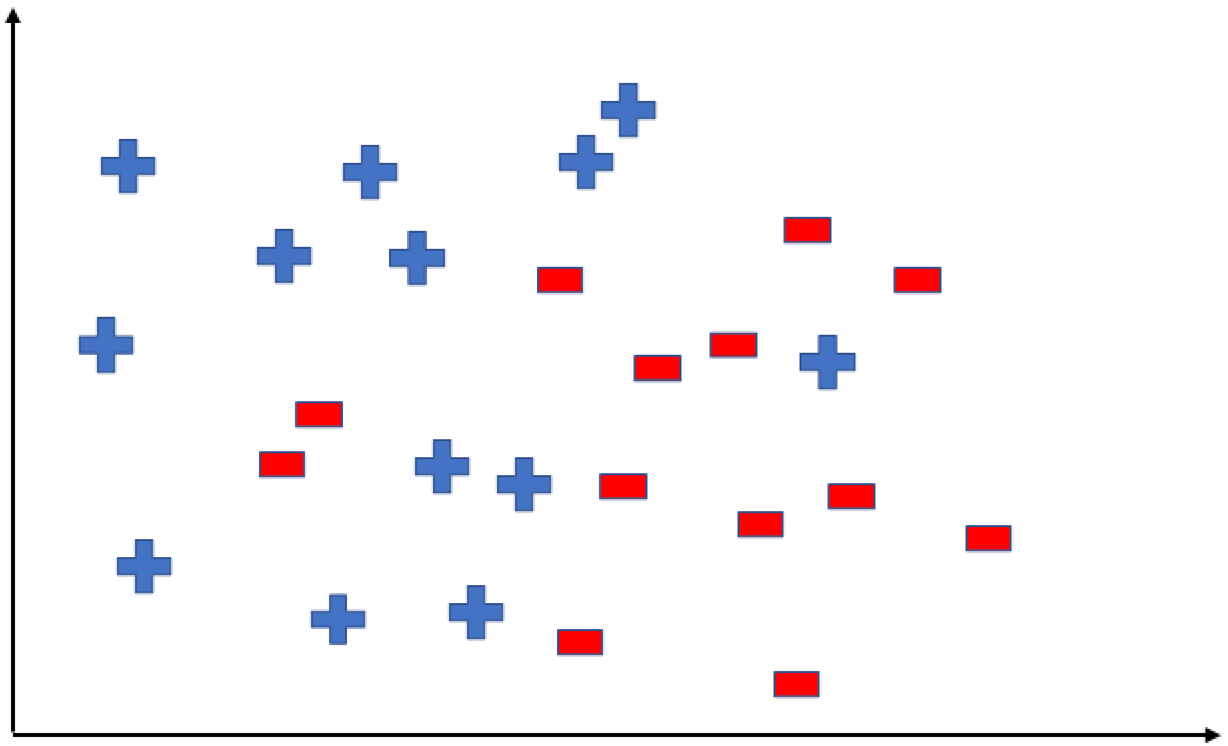
unsupervised learning

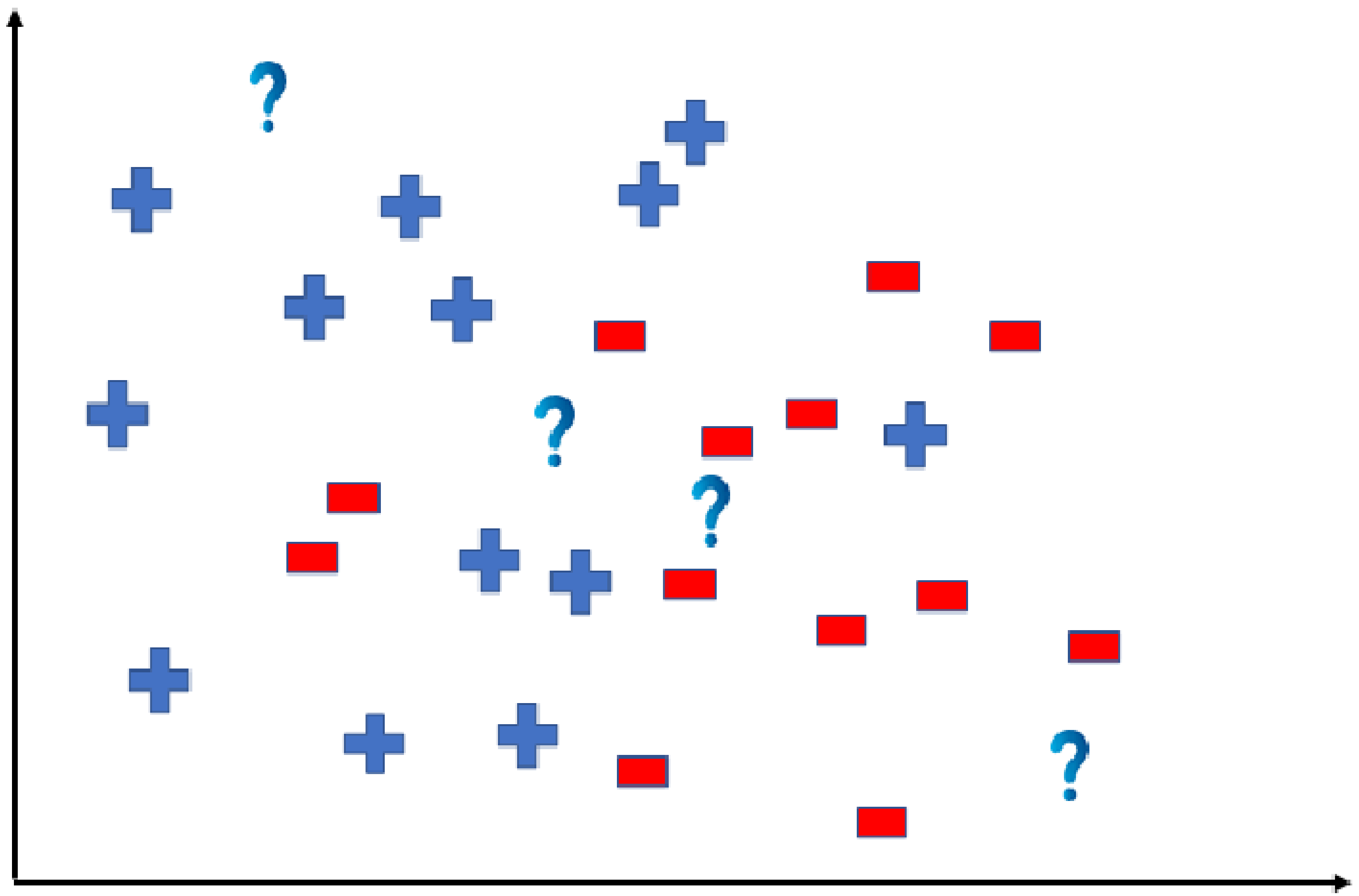


Unsupervised Learning	Supervised Learning
The machine is given huge sets of data that are not labelled as inputs to analyse.	The input is in the form of raw data that is labelled.
The machine needs to figure out the output on its own by identifying patterns in the raw data provided to it.	The machine is already fed with the required feature set to classify between inputs (hence the term 'supervised').
Divided into two types of problems – Association (where we want to find a set of rules that describe our data) and Clustering (where we want to find groups in our data).	Divided into two types of problems – Regression (outputs are real values) and Classification (outputs are categories).
K-means for clustering problems and Apriori algorithm for association rule learning problems.	Linear regression for regression problems, Random Forest for classification and regression problems, Support Vector Machines for classification problems.

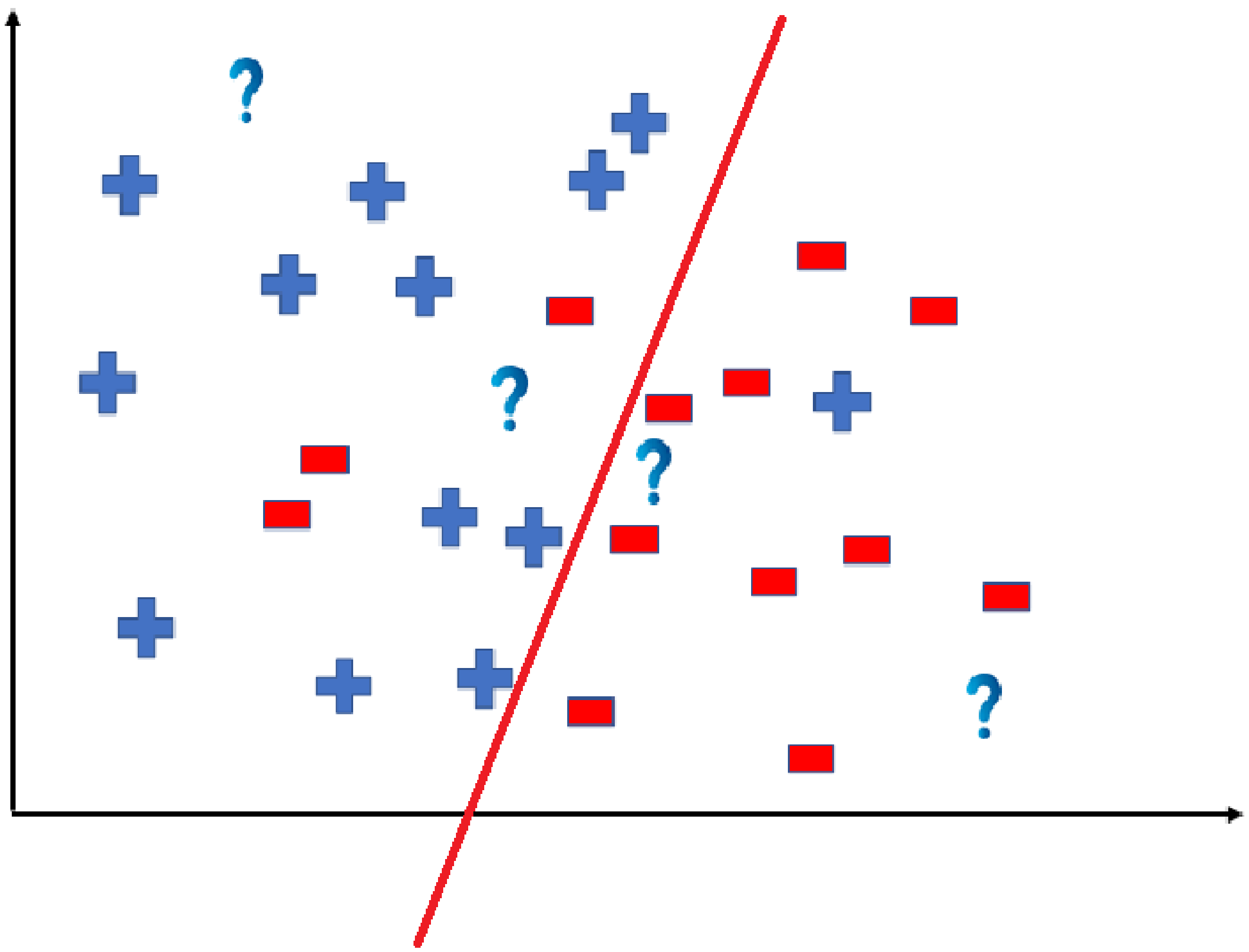
Hypothesis Space (H):
Hypothesis space is the set of all the possible legal hypothesis. This is the set from which the machine learning algorithm would determine the best possible (only one) which would best describe the target function or the outputs.

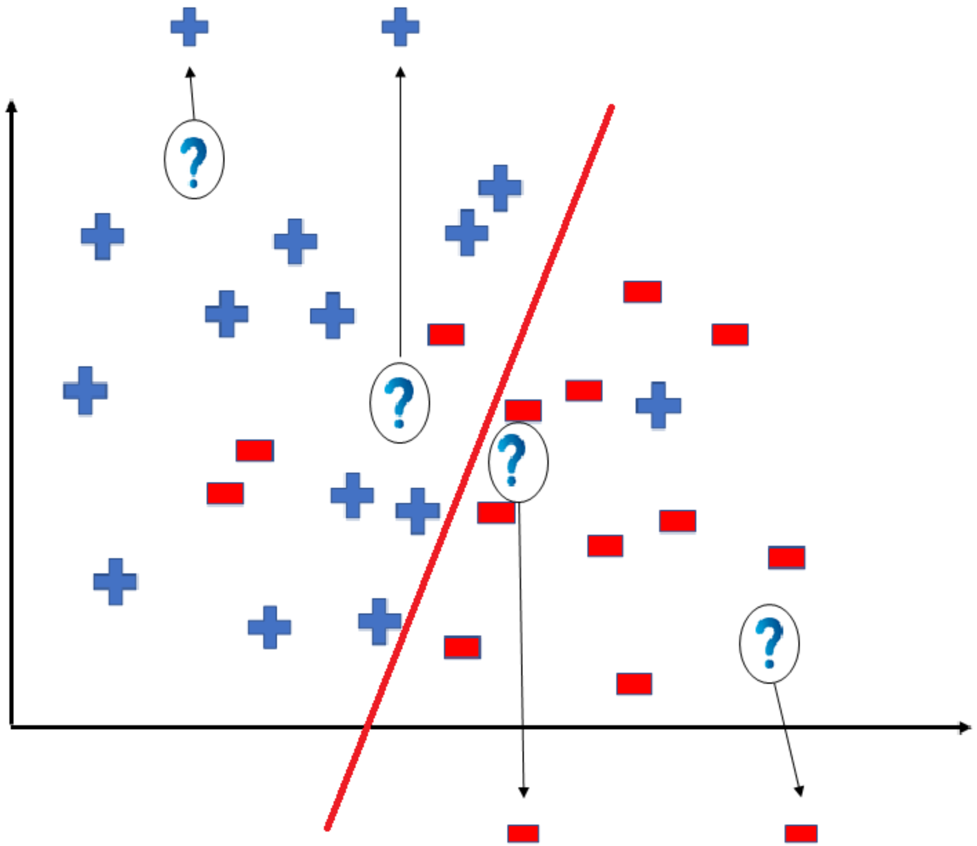
Hypothesis (h):
A hypothesis is a function that best describes the target in supervised machine learning. The hypothesis that an algorithm would come up depends upon the data and also depends upon the restrictions and bias that we have imposed on the data

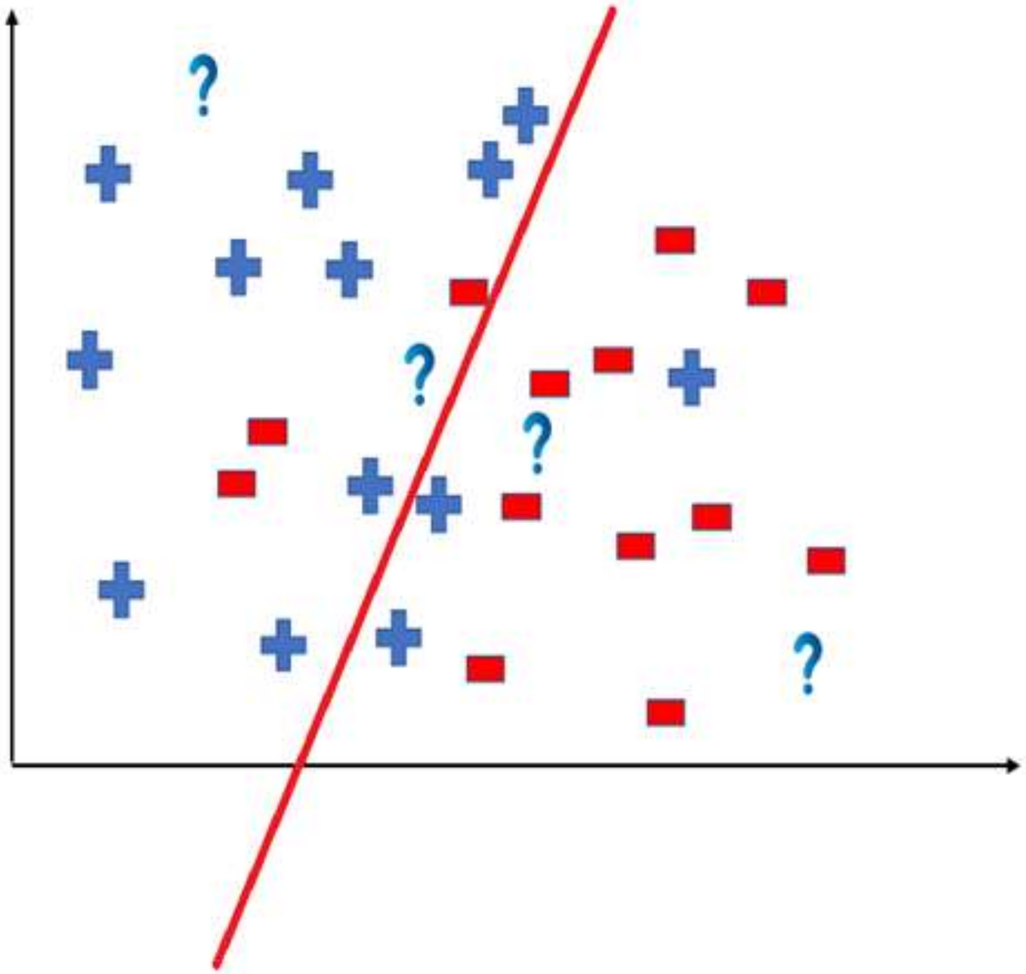




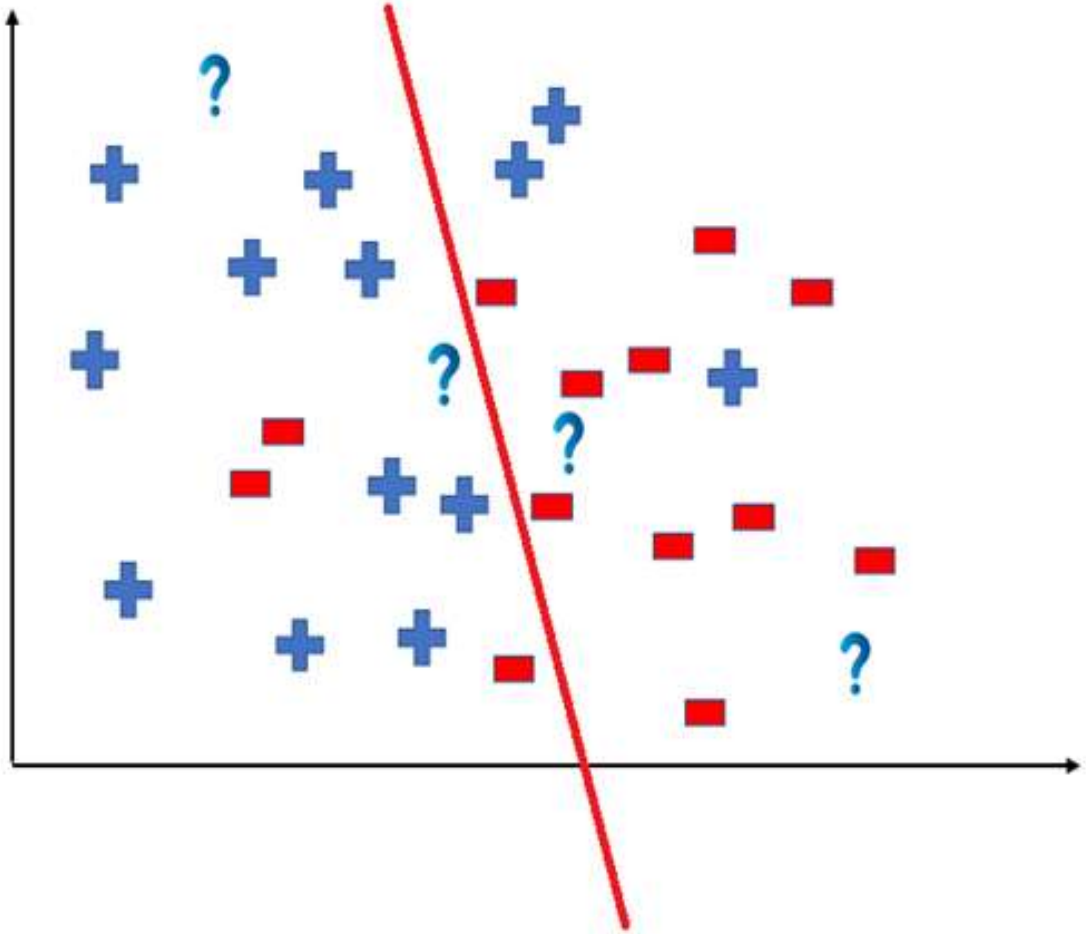
Samatrix.io



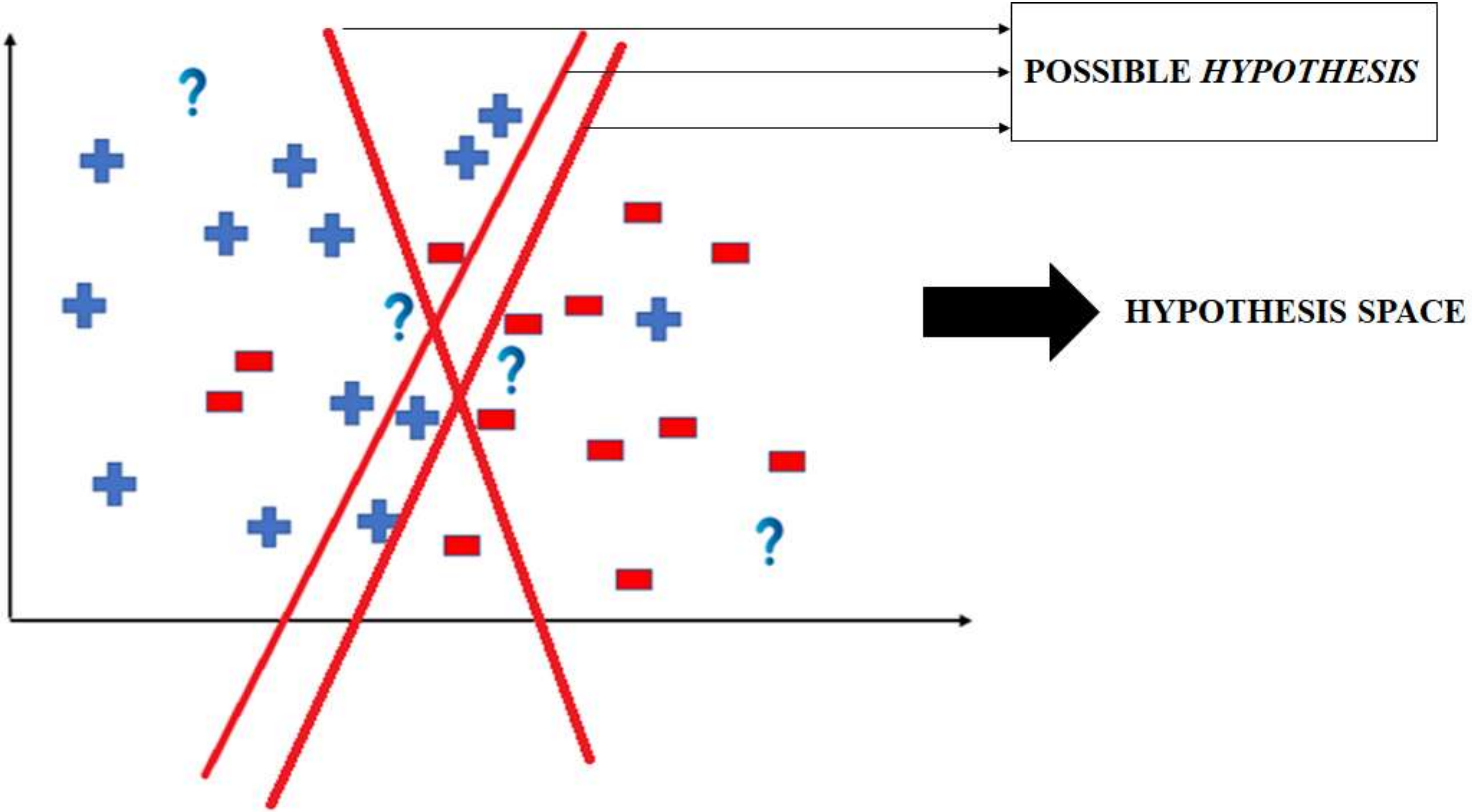




OR



Samatrix.io


POSSIBLE HYPOTHESIS
HYPOTHESIS SPACE

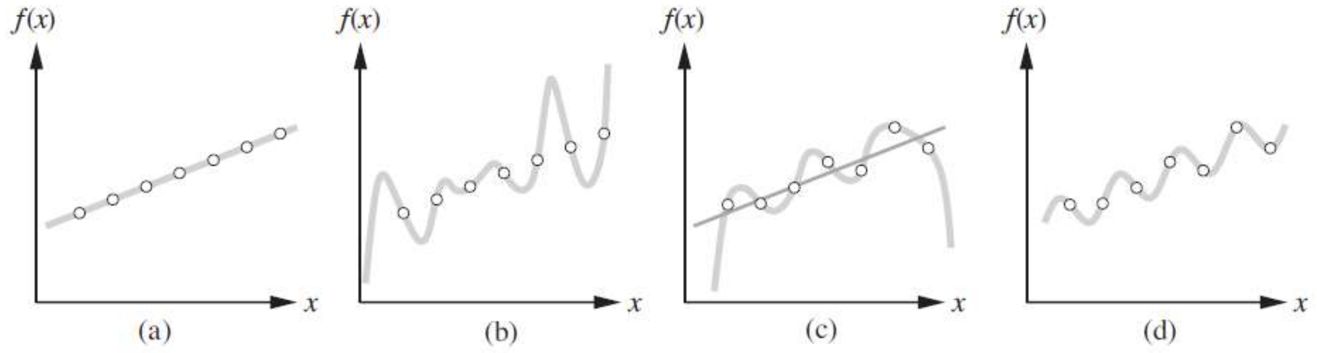
Supervised Learning

- In **predictive** or supervised learning, the agent observes some example input-output pairs and learn a function that map input and output.

  Given a training set of Nexample input-output pairs

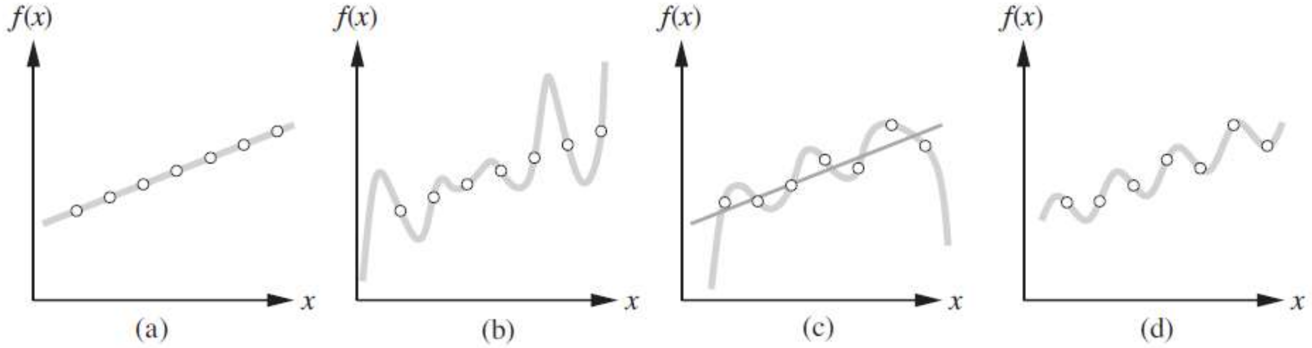
  $$(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_N, y_N),$$

  Where each y_j was generated by unknown function $y = f(x)$, discover a function h that approximates the true function f.



Samatrix.io

Supervised Learning

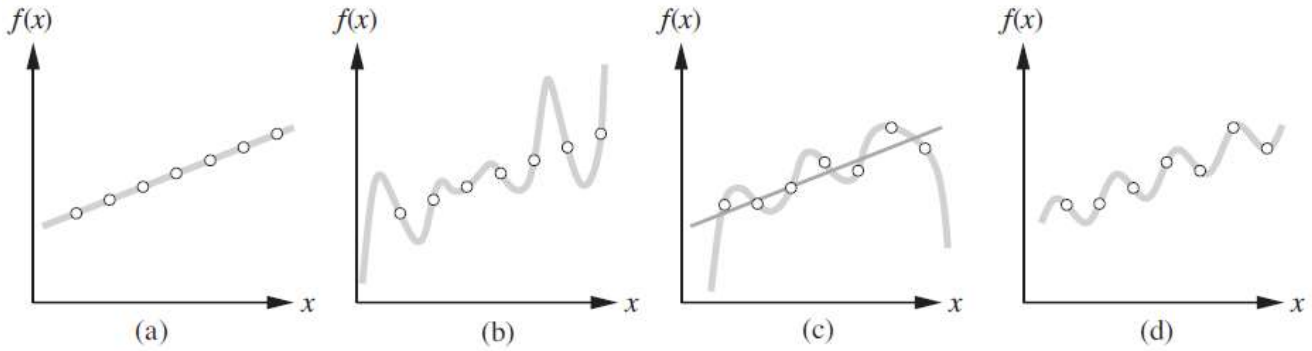


- Here x and y can be of any value not necessarily be the numbers.
- Each training input x_j can also represent the height and weight of a person.
- These are called **features**, **attributes**, or **covariates**.
- The x_j can also be a complex structured object such as an image, a sentence, an email message, a time series, a molecular shape, a graph etc.
- The function h is a **hypothesis**.
- Learning is the search through the space of possible hypothesis for one that will perform well, even on new examples that are beyond the **training set**.
- To measure the accuracy of the training set, we give the **test set** of examples that are separate from the training set.
- We say that the hypothesis **generalizes** well, if it correctly predicts the value of y for new examples.

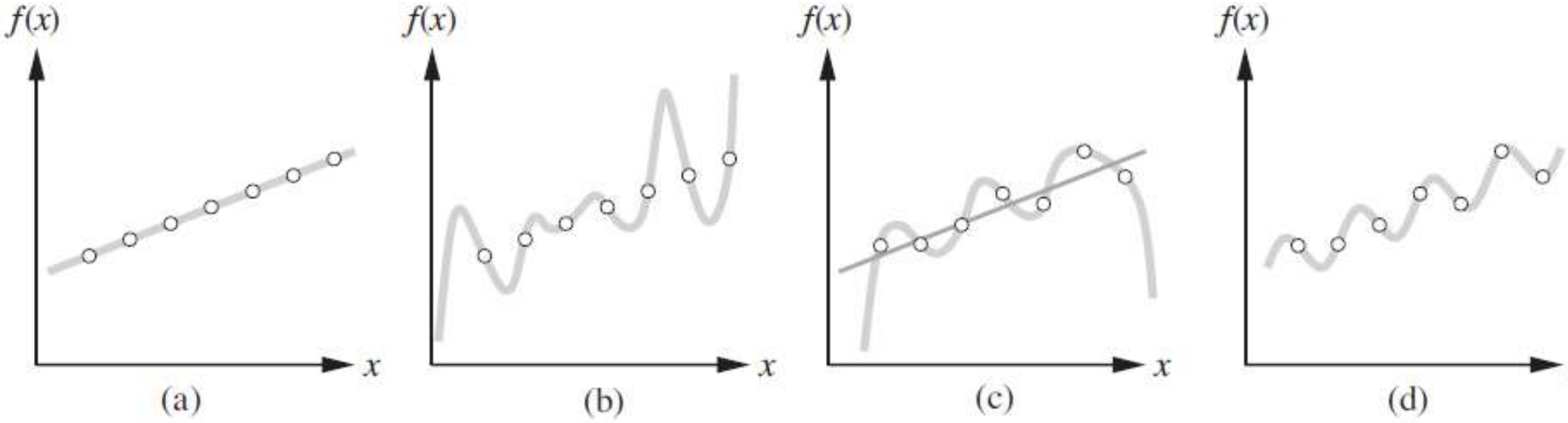
Samatrix.io

Supervised Learning

- When the output y is one among the finite set of **categorical** or **nominal** values (such as sunny, cloudy, or rainy), the learning problem is called **classification or pattern recognition**.
- When y is a number (such as income level), the learning problem is called **regression**.



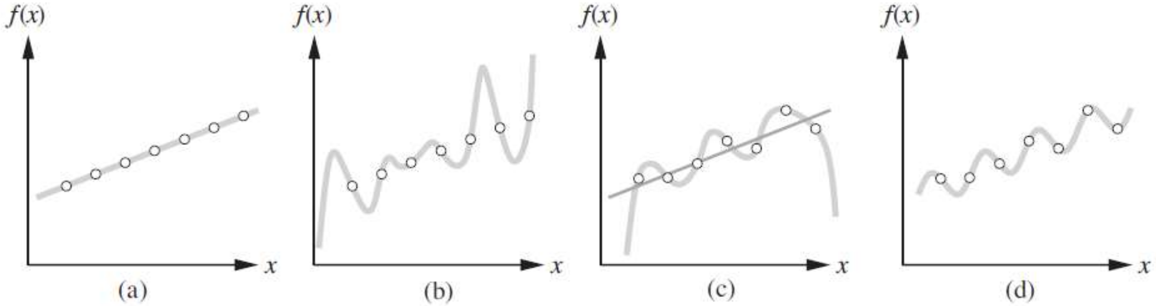
Supervised Learning



- Fig shows the example of fitting a function of a single variable to some data points. (a) Example $(x, f(x))$ pairs and a consistent, linear hypothesis. (b) A consistent, degree-7 polynomial hypothesis for the same data set. (c) A different data set, which admits an exact degree-6 polynomial fit or an approximate linear fit. (d) A simple, exact sinusoidal fit to the same data set.

Samatrix.io

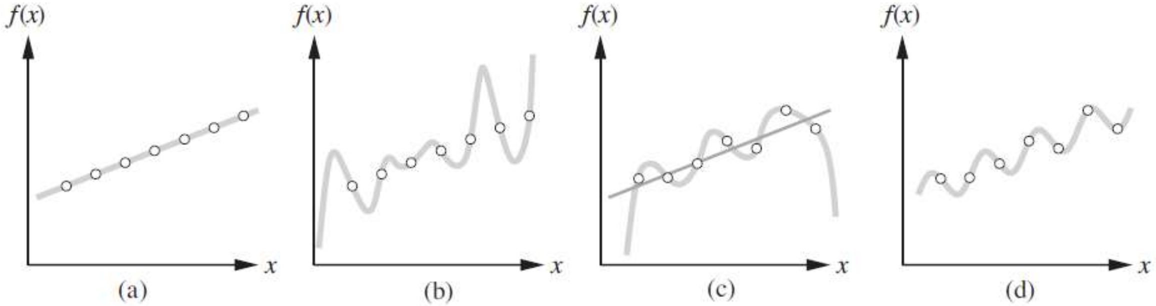
Supervised Learning



- The fig shows the example of fitting a function of a single variable to some data points.
- The example of the points in the (x, y) plane where $y = f(x)$.
- We do not know what f is, but we will approximate it with a function h selected from **hypothesis space**.
- Fig (a) shows some data with an exact fit by a straight line (the polynomial $0.4x + 3$).
- The line is called **consistent** hypothesis because it agrees with all the data.

Samatrix.io

Supervised Learning

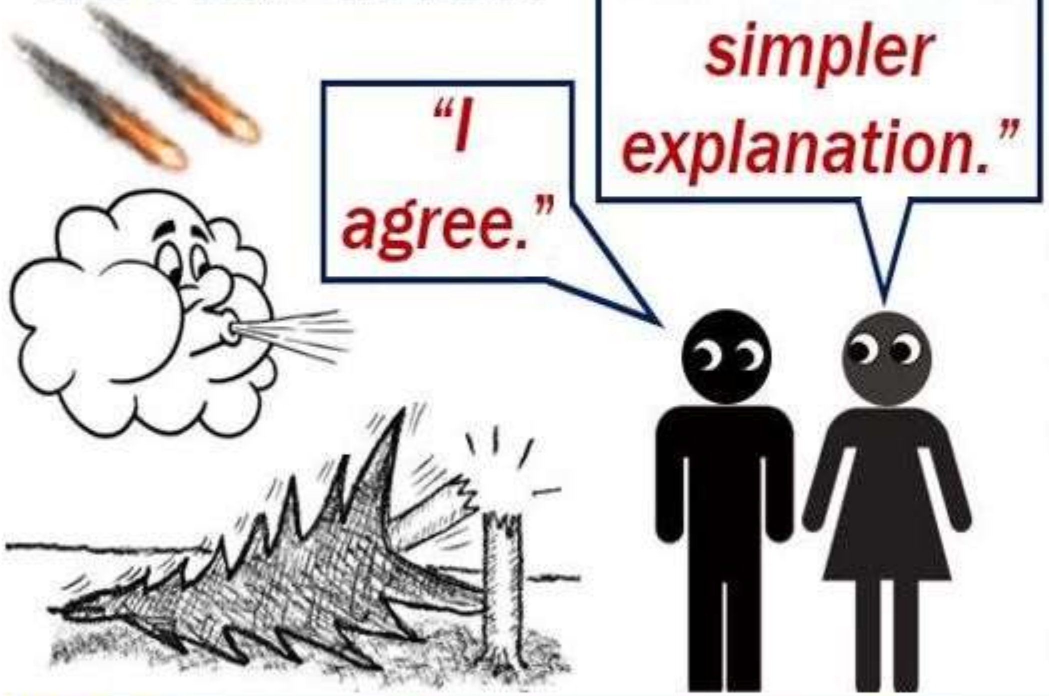


- Figure (b) shoes a high degree polynomial that is also consistent with the same data.
- This illustrates the fundamental problem: *how do we choose from among multiple consistent hypothesis?*
- One answer is to prefer the simplest hypothesis consistent with the data.
- This principle is known as **Ockham's razor**, after 14th century English philosopher William of Ockham, who used it to argue against all sorts of complications.
- Since degree-1 polynomial is simpler than degree-7 polynomial, so (a) should be preferred over (b).

Samatrix.io

Ockham's Razor

Why did the tree fall down?

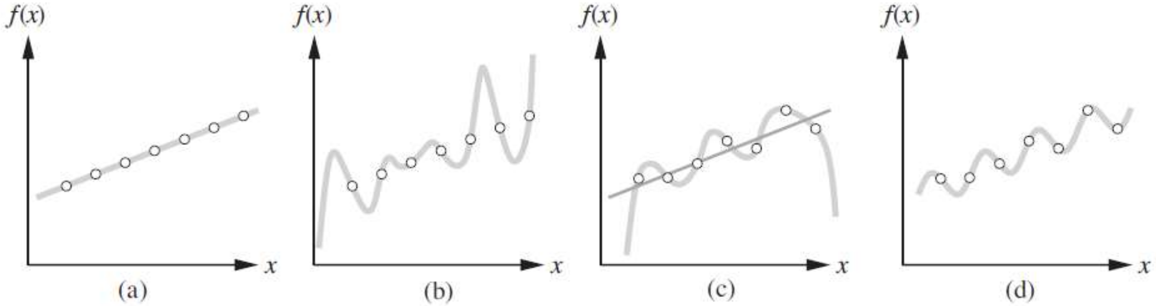


Two Explanations

1. The wind knocked down the tree.
2. Two meteorites. One hit the tree and knocked it down. Then it hit the other meteorite, thus obliterating evidence of its existence.

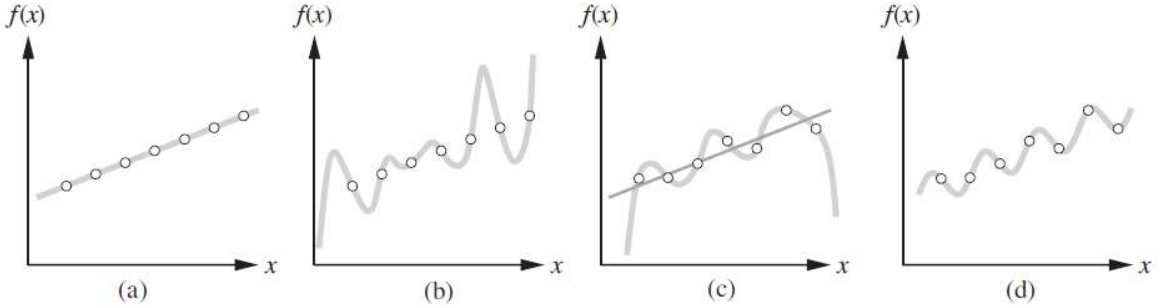
When there are two explanations, choose the simpler one

Supervised Learning



- Figure (c) shows a second data set.
- There is no consistent straight line for this data set.
- It requires degree-6 polynomial for an exact fit.
- There are just 7 data points, so a polynomial with 7 data points does not find any pattern in the data, so we do not expect it to generalize well.
- A straight line that is not consistent with any of the data points, may generalize well for the unseen values of x.

Supervised Learning

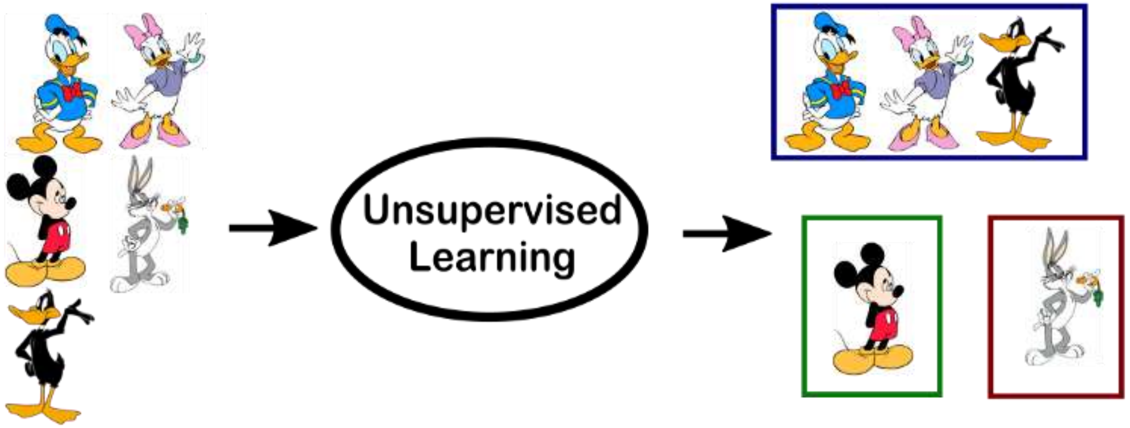


- *In general, there is a trade-off between complex hypothesis that fit the training data well and simpler hypothesis that may generalize better.*
- In figure (d), we expand the hypothesis space to allow polynomials over both x and $sin(x)$, and find that the data in (c) can be fitted better by a simpler function of the form $ax + b + csin(x)$. This shows the importance of hypothesis space.
- The learning problem is **realizable** if the hypothesis space contains the true function. Unfortunately, we cannot always tell whether the given learning problem is realizable, because the true function is not known

Samatrix.io

Unsupervised Learning

- The second main type of machine learning is **descriptive** or unsupervised learning approach.
- Here we are only given a training set of N features or inputs $x_1, x_2, \dots, x_N$.
- We are not interested in prediction, because we do not have associated response variable y.
- Rather the goal is to identify the interesting things about the measurements on $x_1, x_2, \dots, x_N$.
- Is there an informative way to visualize the data? Can we discover subgroups among the variables or among the observations?



Samatrix.io

Unsupervised Learning

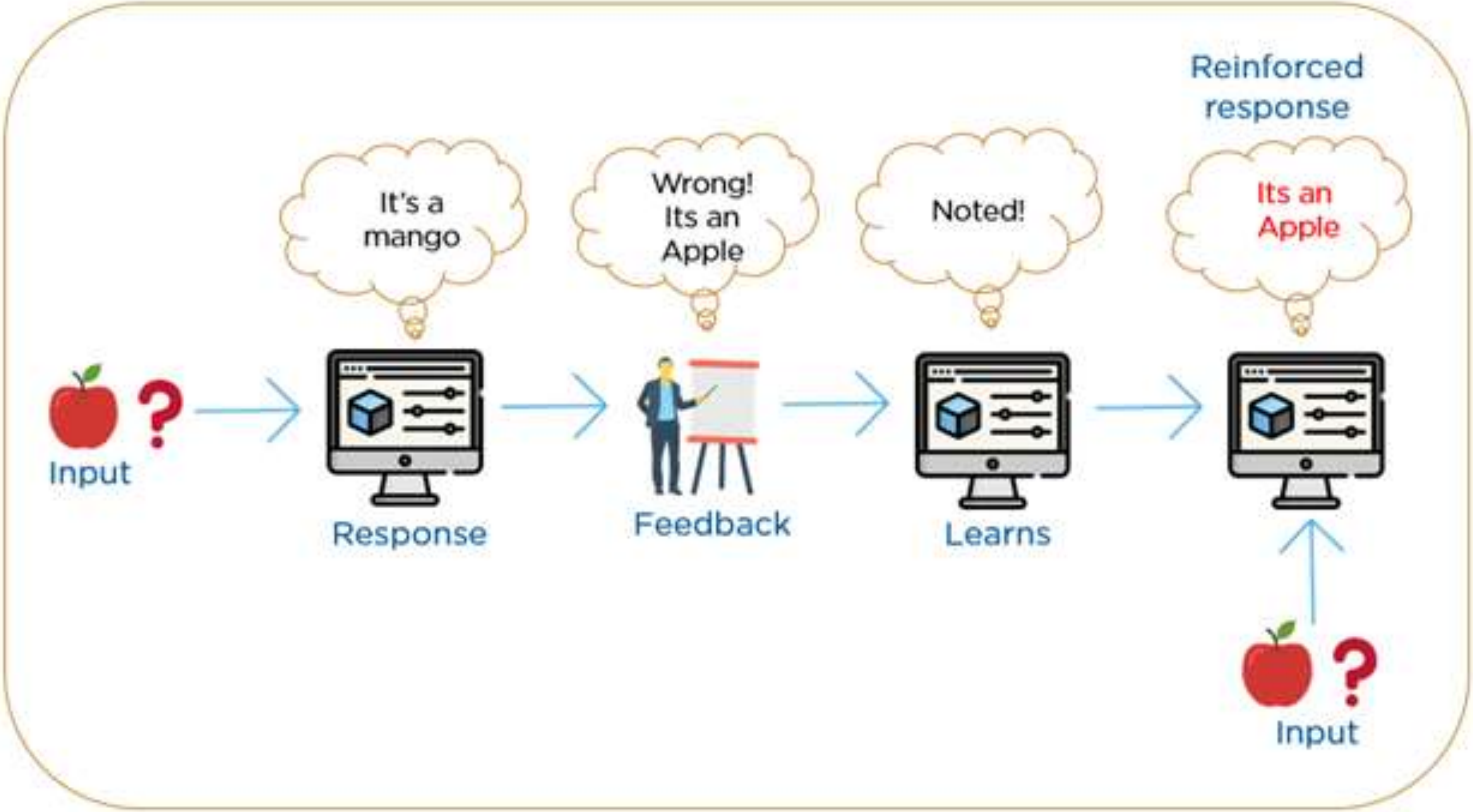
- Unsupervised learning refers to a diverse set of techniques to answer such questions.
- The goal of the unsupervised learning is to find “interesting patterns” in the data.
- This is also known as **knowledge discovery**.
- The purpose of unsupervised learning is data visualization, data compression, or data denoising, or to better understand the correlation present in the data at hand.
- Two most common unsupervised learning types are **principal component analysis**, a tool used for data visualization or data pre-processing before supervised techniques are applied and **clustering**, that consists of dividing the dataset into clusters of similar examples.
- Unsupervised learning is the bread and butter of data analytics, and it’s often a necessary step in better understanding a dataset before attempting to solve the supervised learning problem.

Self - supervised Learning

- This is a specific instance of supervised learning, but it is different enough that it deserves its own category.
- Self-supervised learning is supervised learning without human-annotated label or response variable y.
- You can think of it as supervised learning without any human in the loop.
- There are still labels or response variables involved (because the learning must be supervised by something), but they are generated from the input data, typically using a heuristic algorithm. Input data can be labelled by finding and exploiting the relations (or correlations) between different input signals.

Self – supervised Learning

- For instance, **autoencoders** are a well-known instance of self-supervised learning, that learns to copy its input to its output.
- The purpose of the autoencoder is to reconstruct its inputs by minimizing the difference between the input and the output instead of predicting the target value Y given inputs X.
- Therefore, autoencoders do not require labelled inputs to enable learning.
- In the same way, trying to predict the next frame in a video, given past frames, or the next word in a text, given previous words, are instance of self-supervised learning (**temporally supervised learning**, in this case: supervision comes from future input data).


Reinforced
response
It's a
mango
Wrong!
Its an
Apple
Noted!
Its an
Apple
Input
Response
Feedback
Learns
Input

Samatrix.io

Reinforcement Learning

- In reinforcement learning, an agent receives information about its environment and learns to choose actions that will maximize some **rewards** or **reinforcement**.
- In a reinforcement learning problem, a robot can act in a world, receiving rewards and punishments and determining from these what it should do.
- Reinforcement learning differs from other types of supervised learning because the system isn’t trained with the sample data set.
- Rather, the system learns through trial and error. Therefore, a sequence of successful decisions will result in the process being “reinforced” because it best solves the problem at hand.

Reinforcement Learning

- Consider, for example, the problem of learning to play chess.
- A supervised learning agent needs to be told the correct move for each position it encounters, but such feedback is seldom available.
- In the absence of feedback from a teacher, an agent can learn a transition model for its own moves and can perhaps learn to predict the opponent's moves, but without some feedback about what is good and what is bad, the agent will have no grounds for deciding which move to make.
- The agent needs to know that something good has happened when it (accidentally) checkmates the opponent, and that something bad has happened when it is checkmated—or vice versa, if the game is suicide chess.

Reinforcement Learning

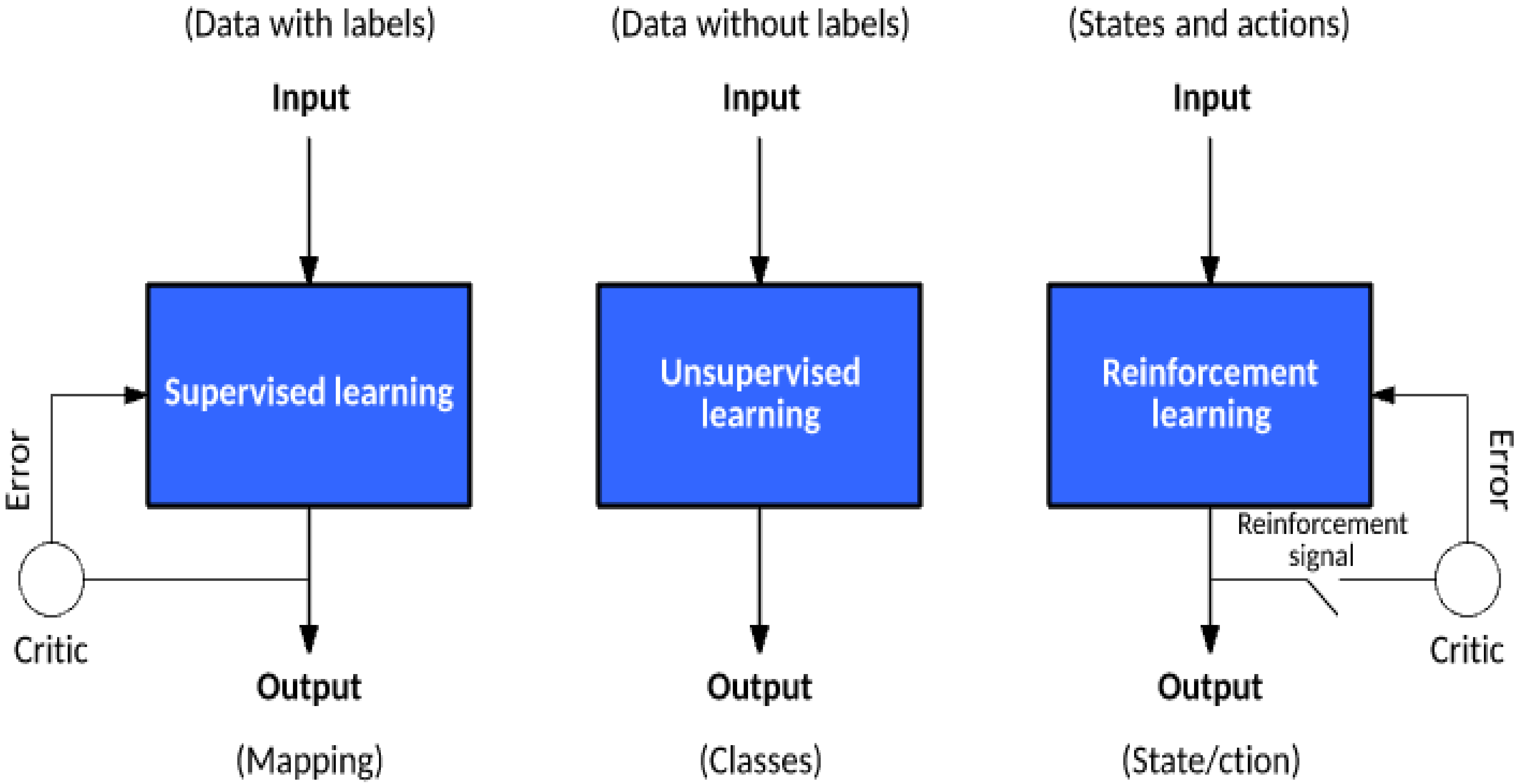
- This kind of feedback is called a reward, or reinforcement.
- In games like chess, the reinforcement is received only at the end of the game.
- In other environments, the rewards come more frequently.
- In ping-pong, each point scored can be considered a reward; when learning to crawl, any forward motion is an achievement.

Reinforcement Learning

- One of the most common applications of reinforcement learning is in robotics or game playing.
- Take the example of the need to train a robot to navigate a set of stairs.
- The robot changes its approach to navigating the terrain based on the outcome of its actions.
- When the robot falls, the data is recalibrated, so the steps are navigated differently until the robot is trained by trial and error to understand how to climb stairs.
- In other words, the robot learns based on a successful sequence of actions.

Reinforcement Learning

- Reinforcement learning is also the algorithm that is being used for self-driving cars.
- In many ways, training a self-driving car is incredibly complex because there are so many potential obstacles.
- If all the cars on the road were autonomous, trial and error would be easier to overcome.
- However, in the real world, human drivers can often be unpredictable.
- Even with this complex scenario, the algorithm can be optimized over time to find ways to adapt to the state where actions are rewarded.
- One of the easiest ways to think about reinforcement learning is the way an animal is trained to take actions based on rewards.
- If the dog gets a treat every time he sits on command, he will take this action each time.



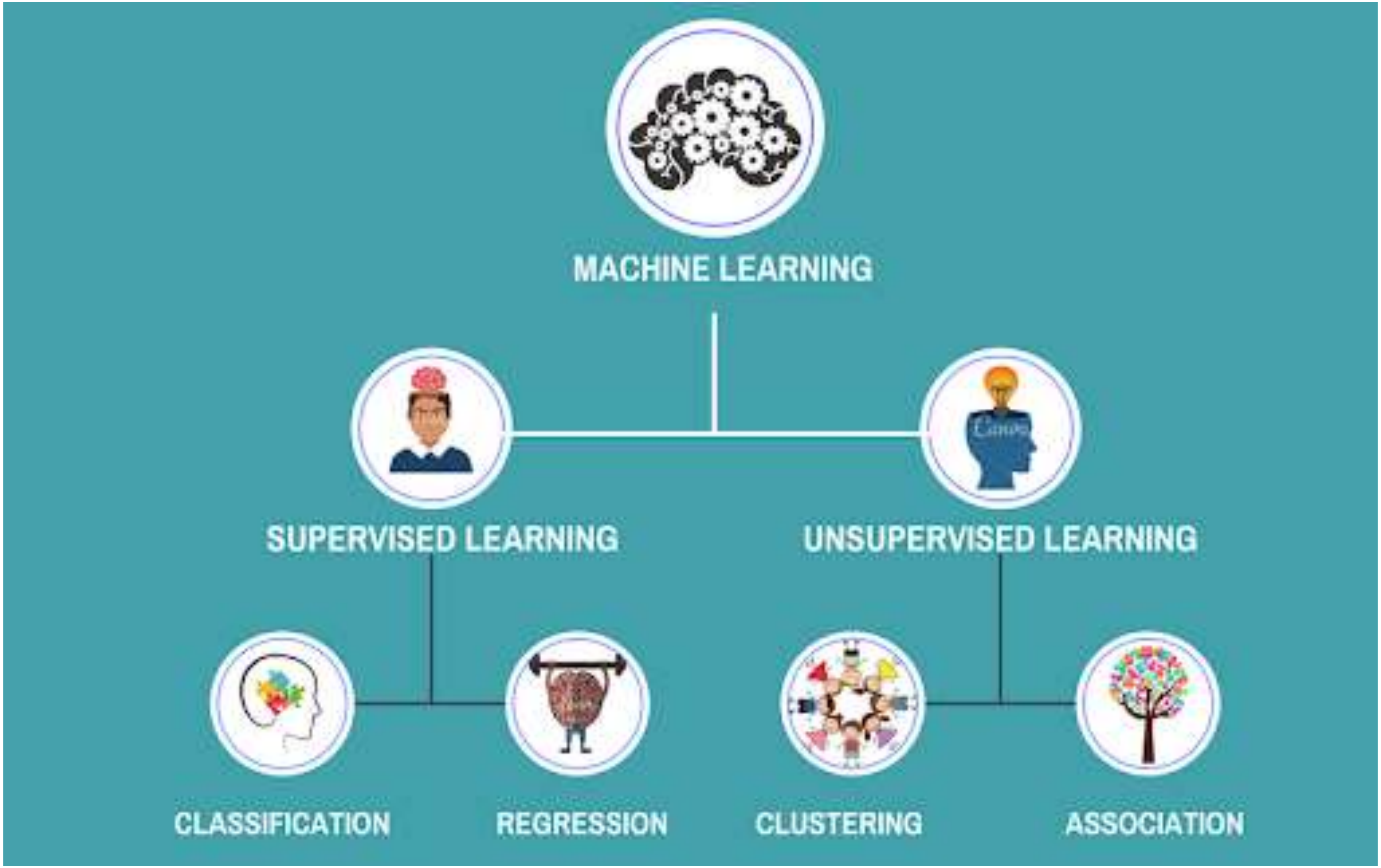
(Data with labels)
Input
Supervised learning
Error
Critic
Output
(Mapping)
(Data without labels)
Input
Unsupervised learning
Output
(Classes)
(States and actions)
Input
Reinforcement learning
Error
Reinforcement signal
Critic
Output
(State/ction)

Samatrix.io

Type of Machine Learning Algorithms

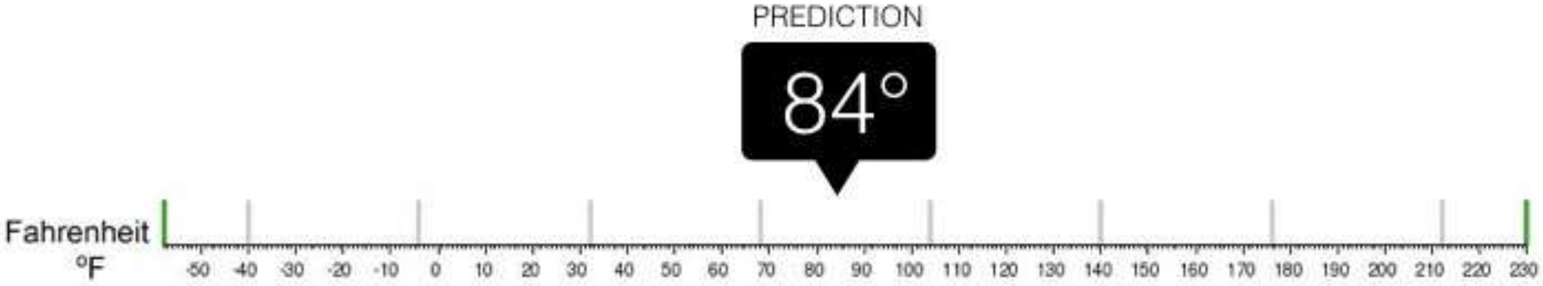
Type of Machine Learning Algorithms

- Selecting the right machine learning algorithms is part art and part science.
- Two data scientists can use the two different machine learning algorithms solving the same business problem using the same data sets.
- Hence, understanding different machine learning algorithms help the data scientists select the best types of algorithms to solve a given business problem


MACHINE LEARNING
SUPERVISED LEARNING
UNSUPERVISED LEARNING
CLASSIFICATION
REGRESSION
CLUSTERING
ASSOCIATION

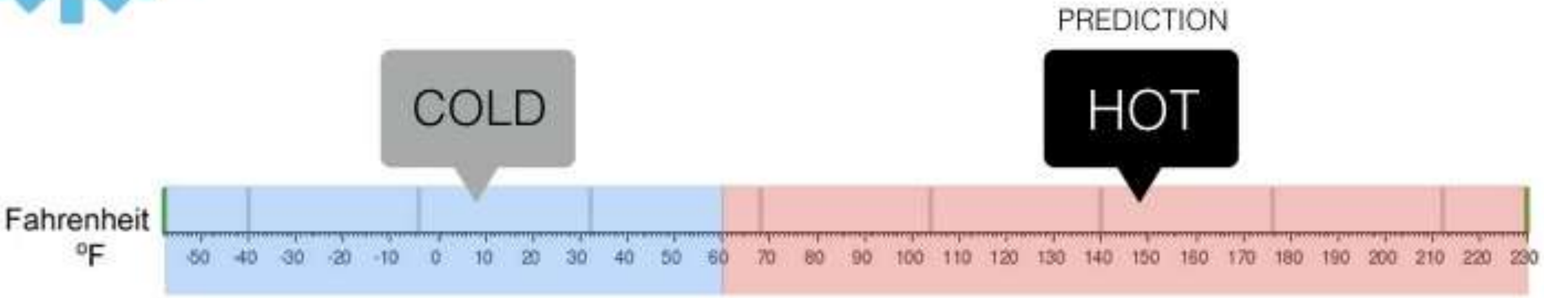
Regression

What is the temperature going to be tomorrow?



Classification

Will it be Cold or Hot tomorrow?



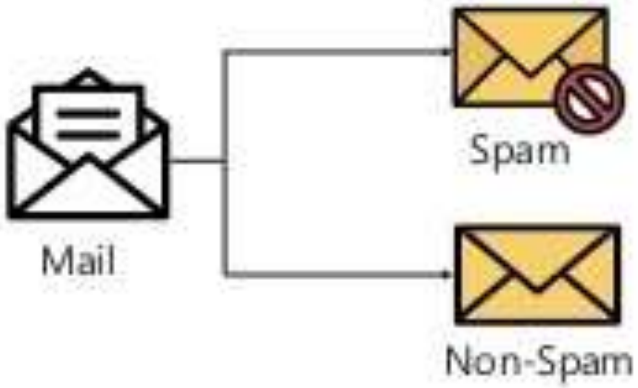
Samatrix.io

Regression Vs Classification

Classification

Classification is the task of predicting a discrete class label

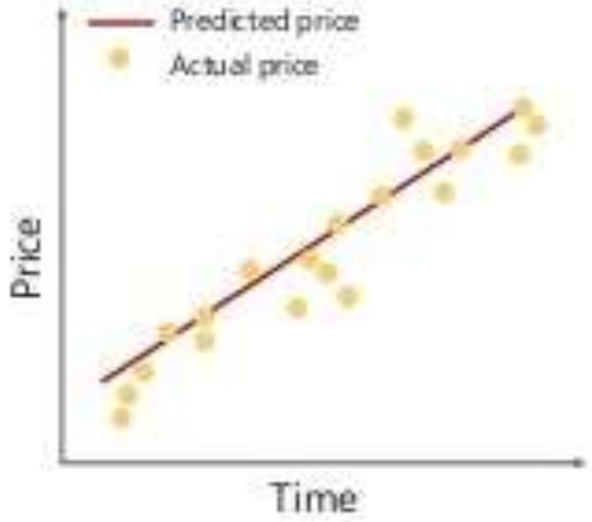
- In a classification problem data is classified into one of two or more classes
- A classification problem with two classes is called binary, more than two classes is called a multi-class classification



Regression

Regression is the task of predicting a continuous quantity

- A regression problem requires the prediction of a quantity
- A regression problem with multiple input variables is called a multivariate regression problem



Regression vs Classification

- Variables can be characterised as either *quantitative* or *qualitative* (also known as categorical).
- The quantitative variables take on numerical values. Example includes a person's age, height, or income.
- The qualitative variables take on values in one of K different classes or categories.
- Example of qualitative variables include a person's gender (male or female), the brand of the product purchased (brand X, Y, or Z), whether the person defaults on a debt (yes or no), or a cancer diagnosis (Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, or No Leukemia).

Regression vs Classification

- The problems with quantitative response are referred as **regression** problems.
- Those involved in qualitative response are referred as **classification** problems.
- However, the distinction is always not very crisp.
- Least square linear regression is used with a quantitative response.
- Logistic regression is typically used for qualitative (two-class or binary) response.

Regression vs Classification

- We tend to select the learning model on the basis whether the response is quantitative or qualitative.
- However, whether the predictors are quantitative or qualitative is considered less important.

Likelihood

How probable is the evidence given that our hypothesis is true?

Prior

How probable was our hypothesis *before* observing the evidence?

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

Posterior

How probable is our hypothesis given the observed evidence? (Not directly computable)

Marginal

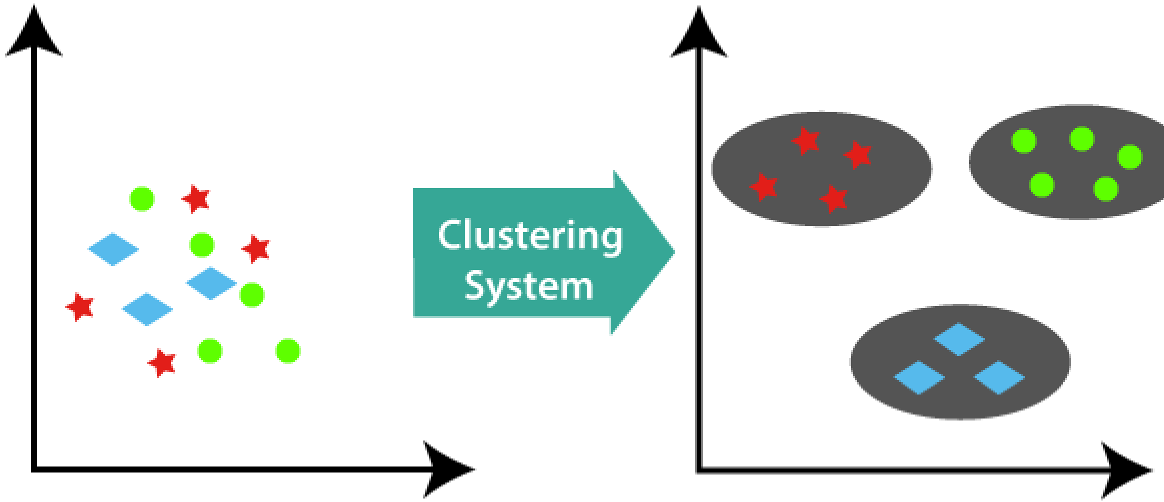
How probable is the new evidence under all possible hypotheses?

$P(e) = \sum P(e \mid H_i)\, P(H_i)$

Bayesian

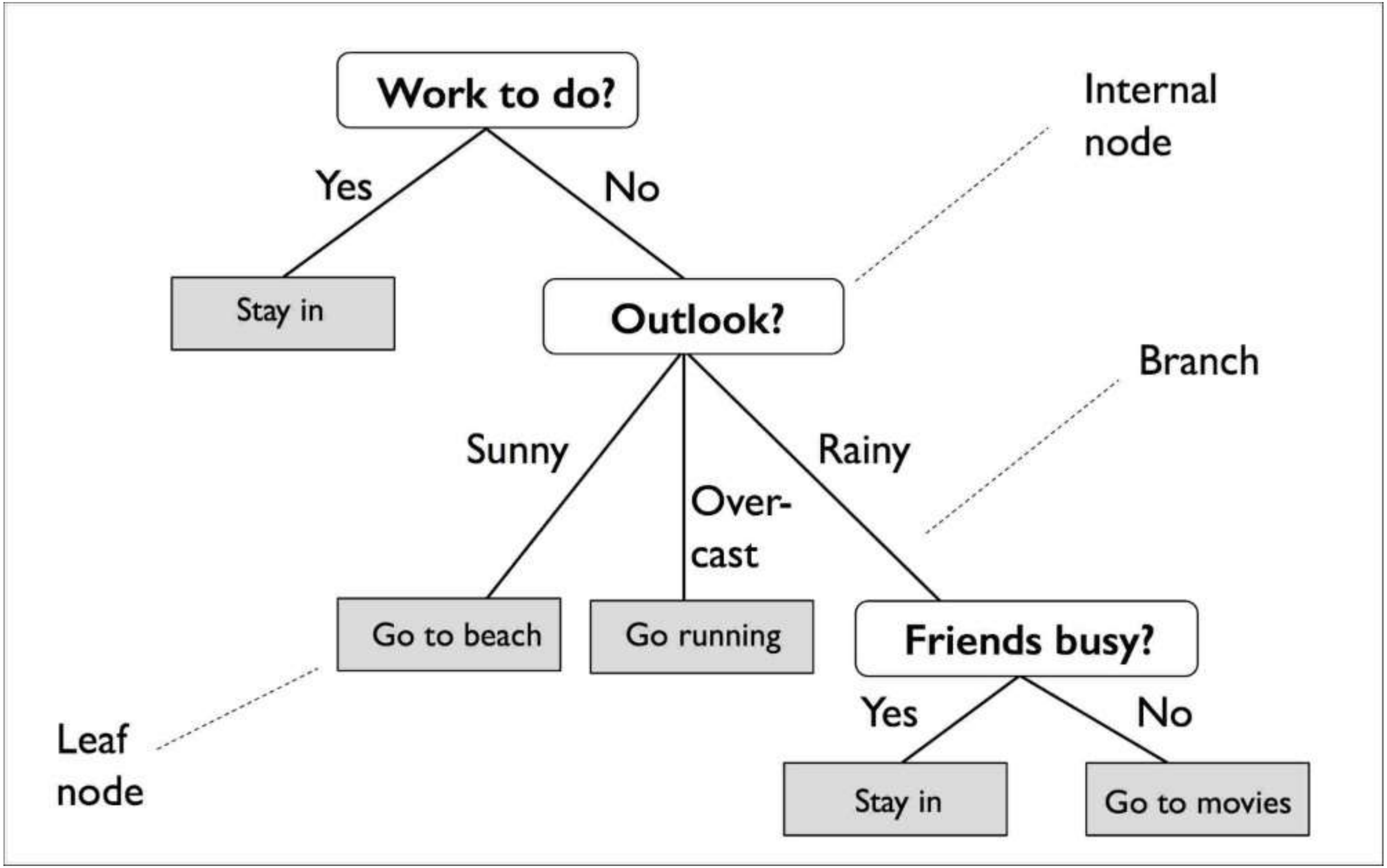
- Bayesian algorithms help the data analysts encode their prior beliefs about what the model look like, independent of the data set.
- These algorithms are especially useful when you do not have massive amount of data to train the model confidently.
- The Bayesian algorithm would be helpful if you have prior knowledge to some part of the model and you can code that directly.
- For example, if you want to model a medical imaging diagnosis system that looks for lung disorder.
- If a published journal study estimates the probability of different lung disorders based on lifestyle, those probabilities can be encoded into the model.

Clustering

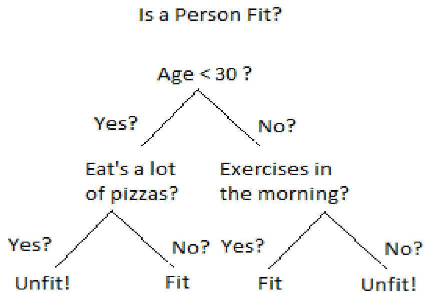


- In clustering objects with similar parameters are grouped together in a cluster.
- All objects in a cluster are more like each other than objects in other cluster.
- The clustering is a type of unsupervised learning because the data is not labelled.
- The clustering algorithm interprets the parameter that make up each item and then groups them accordingly.

Samatrix.io


Work to do?
Yes
No
Stay in
Outlook?
Internal node
Branch
Sunny
Over-cast
Rainy
Go to beach
Go running
Friends busy?
Leaf node
Yes
No
Stay in
Go to movies

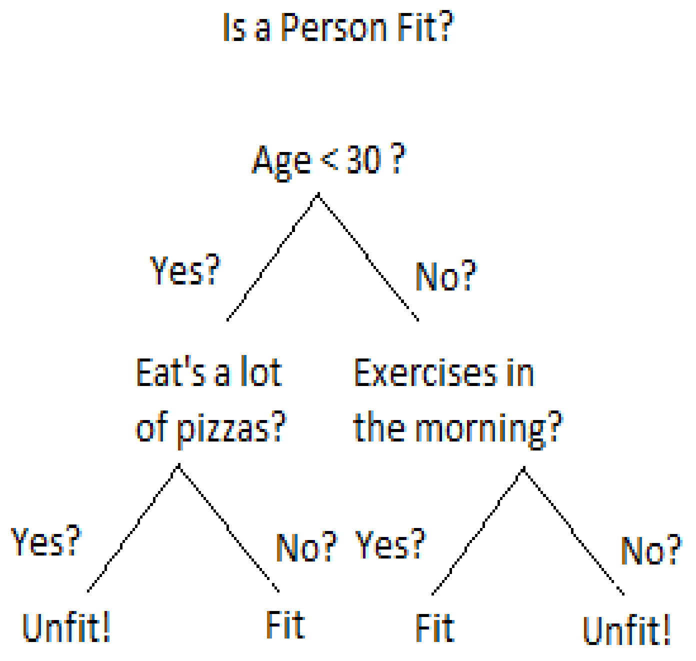
Decision Tree

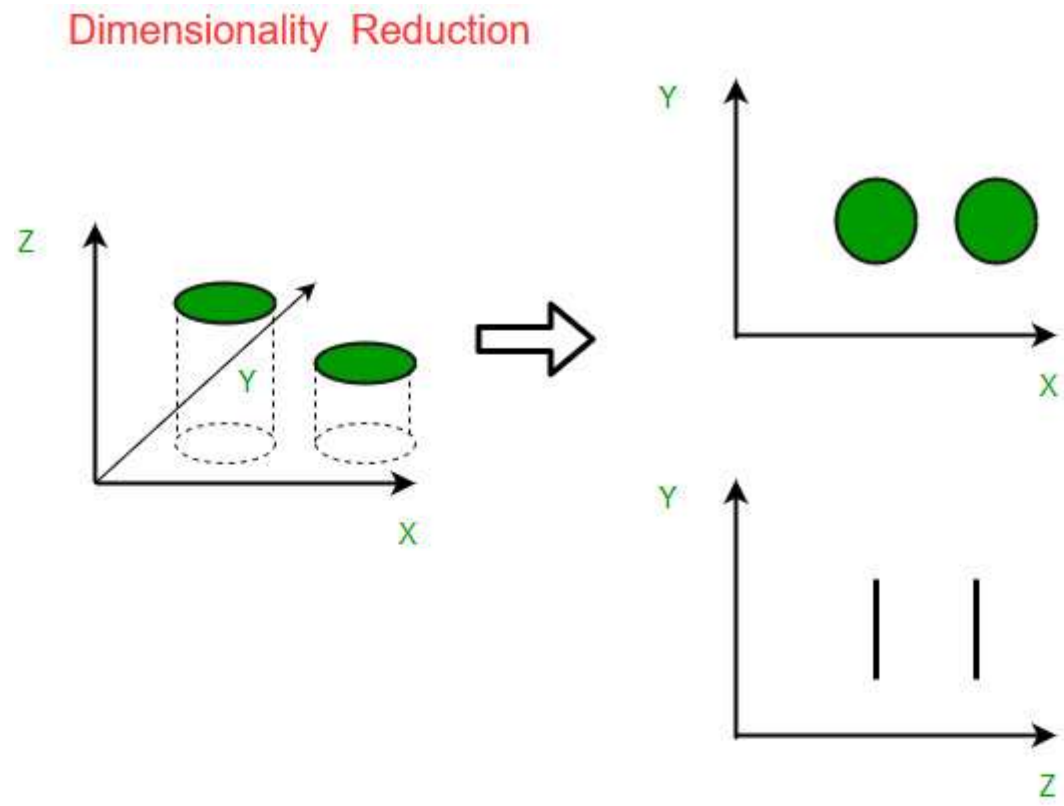


- Decision tree algorithms uses a tree-like graph or branching structure to illustrate the event outcomes, resource costs, and utility.
- The decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether the object is cat or a dog), each branch represents the outcome of the test.
- Each leaf node represents a possible outcome.
- The paths from root to leaf represent classification rule.
- Percentages are assigned to nodes based on the likelihood of the outcome occurring.

Decision Tree

- Decision tree algorithms are the one of the most widely used supervised learning methods.
- Tree based algorithms empower predictive models with high accuracy, stability, and ease of interpretation.
- They can easily solve both classification and regression problems.
- Decision Tree algorithms are referred to as CART (Classification and Regression Trees).



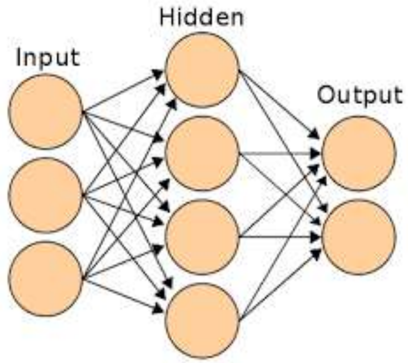


- Dimensionality reduction helps systems remove data that’s not useful for analysis.
- This group of algorithms is used to remove redundant data, outliers, and other non-useful data.
- Dimensionality reduction can be helpful when analysing data from sensors and other Internet of Things (IoT) use cases.

Dimensionality Reduction

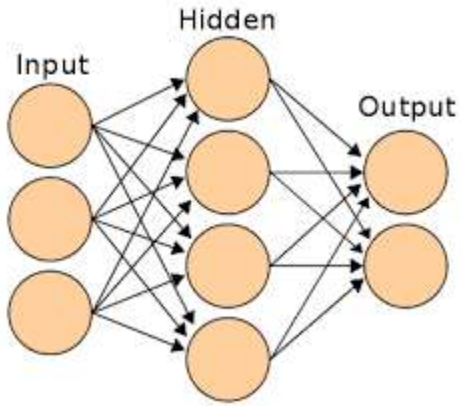
- In IoT systems, there might be thousands of data points simply telling you that a sensor is turned on.
- Storing and analyzing that "on" data is not helpful and will occupy important storage space.
- In addition, by removing this redundant data, the performance of a machine learning system will improve.
- Finally, dimensionality reduction will also help analysts visualize the data.

Neural Network and Deep Learning



- A neural network attempts to mimic the way a human brain approaches problem and uses layers of interconnected units to learn and infer relationships based on observed data.
- A neural network can have several connected layers.
- When there is more than one hidden layer in a neural network, it is sometimes called deep learning.
- Neural network models can adjust and learn as data changes.
- Neural networks are often used when data is unlabeled or unstructured.
- One of the key use cases for neural networks is computer vision.

Neural Network and Deep Learning



- Deep learning is being leveraged today in a variety of applications.
- Self-driving cars use deep learning to help the vehicle understand the environment around the car.
- As the cameras capture images of the surrounding environment, deep learning algorithms interpret the unstructured data to help the system make near real-time decisions.
- Likewise, deep learning is embedded in applications that radiologists use to help interpret medical images

Training Machine Learning Systems

Training Machine Learning Systems

- The process of developing and refining the model is an iterative process.
- The steps include selecting the correct algorithm, training, and testing a system.
- Training is a critical step in machine learning process.

Training Machine Learning Systems

- To train your machine learning systems, you need to know the inputs (for example: Income level, size of the house, location of the house and so on), you should know your desired goal (the price of the house in the area).
- However, the mathematical function to transform the row data to the price of the house is unknown.
- As the learning algorithm is exposed to more and more data, the model will become more accurate in predicting the price of the house.

Steps

- There are three steps in training the machine learning algorithm:
  - Representation
  - Evaluation
  - Optimization

Representation

- The machine learning algorithm creates a model to transform the input data into the desired results.
- As the model is exposed to more data, it will learn the relationship between the raw data and point out data points that are strong predictors for the desired output

Evaluation

- The algorithm creates multiple models.
- Once the models are ready, we need to evaluate and score the models based and find out which model produces the most accurate predictions.
- After the model is operationalize, it is will be exposed to unknown data.
- As a result, we need to ensure that the model is generalized and not overfit to training data.

Optimization

- After the algorithm creates and scores various models, we need to select the best performing model and algorithm.
- As the algorithm is exposed to more diverse sets of input data, we need to select the most generalized model.

Training Process

- For the training process, it is very important to have enough data so that you can also test the model.
- The first pass of the training process provides mixed results.
- That means either you need to refine the model or provide more data.

Thanks

Samatrix Consulting Pvt Ltd