Unit-I
Introduction to Machine Learning

By: Prof. Vasundhara Uchhula

Syllabus

Overview of Human Learning

- Learning is typically referred to as the process of gaining information through observation.
- And why do we need to learn?
  - In our daily life , we need to carry out multiple activities.
  - It may be a task as simple as walking down the street or doing the homework.
  - Or it can be a complex task of deciding the angle of trajectory of a rocket for launching in space.
- As we keep learning more , efficiency in doing tasks keep improving.
- With more knowledge the ability to do homework with less number of mistakes increases
- With more learning, tasks can be performed easily.

Types of Human Learning

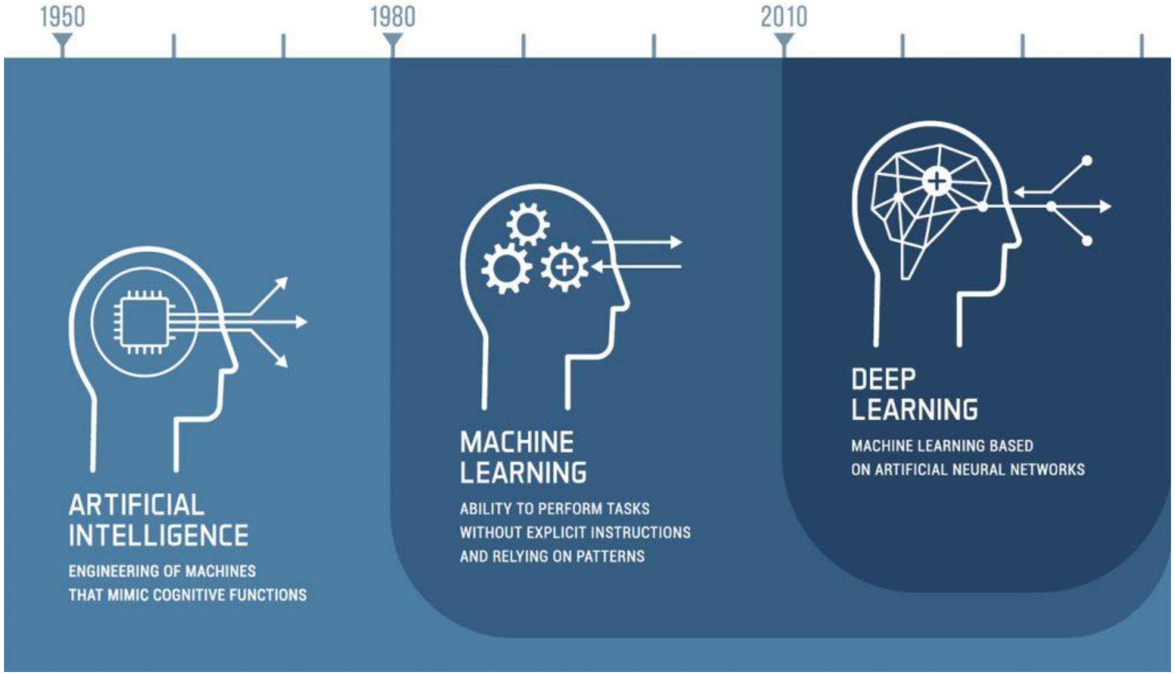
- Learning under expert guidance
    - Like a child taught by parents
    - He calls his hand a ‘hand’ because that is the information he gets from his parents.
    - Sky is blue to him because that is what his parents have taught him
    - Next phase of life is when baby goes to school. He starts with basic familiarization of alphabets and digits
    - Moving to words , sentences, paragraphs, etc.
    - And moves to next phase of life with higher studies, professional life..etc
    - In all phases of life of a human being there is an element of guided learning

      So guided learning is a process of gaining information from a person having sufficient knowledge due to past experience.

Contd..

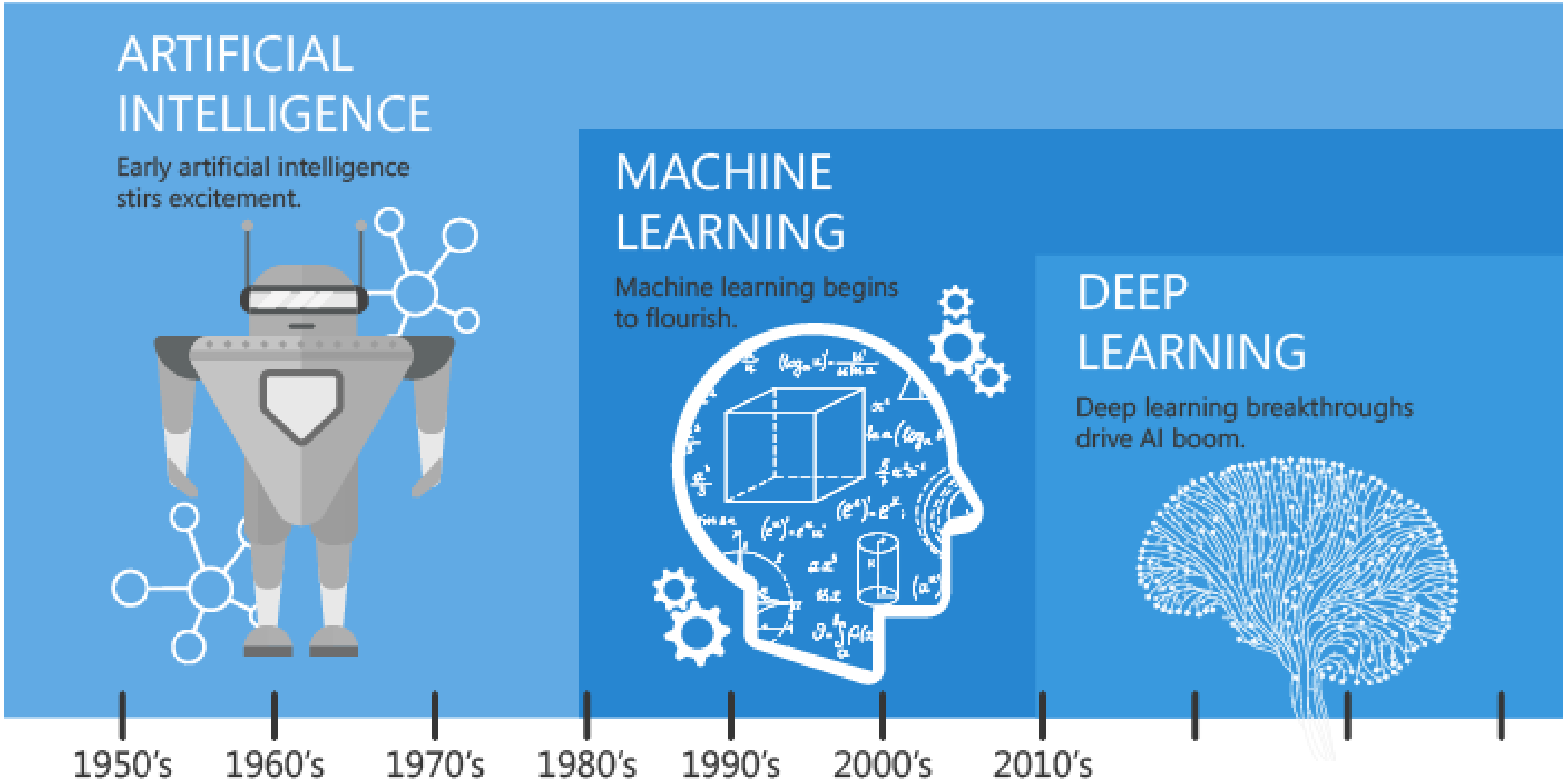
- Learning guided by knowledge gained from experts
  - Knowledge imparted by teacher or mentor at some point of time in some other form or context.
  - Ex: a baby can group together all objects of same color even if his parents have not specifically taught him to do so.
  - There is no direct learning.
  - It is some past information shared on some context which is used as a learning to make decisions.
- Learning by self
  - In many situations, humans are left to learn on their own.
  - A classic example is a baby learning to walk through obstacles.
  - He bumps on to obstacles and falls down multiple times till he learns that whenever there is an obstacle , he needs to cross over it.
  - Not all things are taught by others.
  - A lot of things need to be learnt only from mistakes made in the past.

AI, ML and DL


1950
1980
2010
ARTIFICIAL INTELLIGENCE
ENGINEERING OF MACHINES THAT MIMIC COGNITIVE FUNCTIONS
MACHINE LEARNING
ABILITY TO PERFORM TASKS WITHOUT EXPLICIT INSTRUCTIONS AND RELYING ON PATTERNS
DEEP LEARNING
MACHINE LEARNING BASED ON ARTIFICIAL NEURAL NETWORKS

1. **AI** enables machines to think without any human intervention. It is a broad area of computer science
2. **Machine Learning (ML):** ML is a subset of AI that uses statistical learning algorithms to build smart systems. The ML systems can automatically learn and improve without explicitly being programmed.
3. **Deep Learning (DL)** This subset of AI is a technique that is inspired by the way a human brain filters information.

Evolution of Machine Learning from 1950



ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.

MACHINE LEARNING

Machine learning begins to flourish.

DEEP LEARNING

Deep learning breakthroughs drive AI boom.

1950's 1960's 1970's 1980's 1990's 2000's 2010's

Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

What is Machine Learning?

- Before learning this we should be able to answer more fundamental questions like
  - Do machines really learn?
  - If so , how do they learn?
  - Which problem do we consider as well posed learning problem? What are the important features that are required to well define a learning problem?

Definition of Machine Learning?

- Two definitions of Machine Learning are offered.
- Arthur Samuel described it as: "the field of study that gives computers the ability to learn without being explicitly programmed." This is an older, informal definition.

- Tom Mitchell provides a more modern definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."
- Example: playing checkers.
  - E = the experience of playing many games of checkers
  - T = the task of playing checkers.
  - P = the probability that the program will win the next game.
  - In general, any machine learning problem can be assigned to one of two broad classifications:
    - Supervised learning and Unsupervised learning.

Question?

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?

1. Watching you label emails as spam or not spam.
2. The number (or fraction) of emails correctly classified as spam/not spam.
3. Classify emails as spam or not spam.
4. None of the above, this is not a machine learning algorithm.

??

Suppose we feed a learning algorithm a lot of historical weather data, and have it learn to predict weather.

What would be a reasonable choice for P?

-The process of the algorithm examining a large amount of historical weather data.

-The probability of it correctly predicting a future date's weather.

-The weather prediction task.

??

Handwriting recognition learning problem

What is T, P and E ?

- Task T : Recognizing and classifying handwritten words within images
- Performance P : Percent of words correctly classified
- Training experience E : A dataset of handwritten words with given classifications

How do machines learn?

The basic machine learning process can be divided into three parts

1. Data Input:

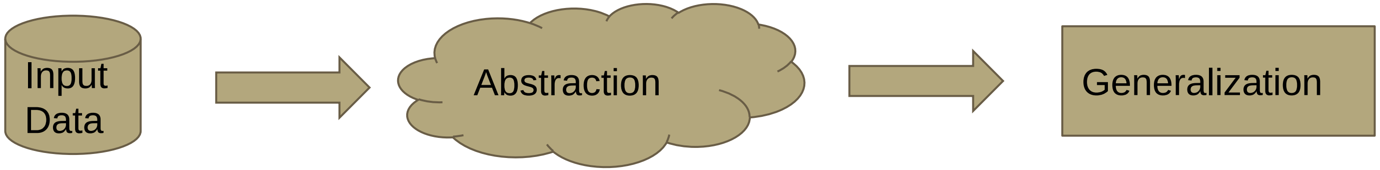
   Past data or past information is utilized as a basis for future decision making

2. Abstraction:

   The input data is represented in a broader way through the underlying algorithm

3. Generalization:

   The abstracted representation is generalized to form a framework for making decisions.



Abstraction

- During the machine learning process, knowledge is fed in the form of input data. However data cannot be used in the original shape and form.
- Abstraction helps in deriving a conceptual map based on input data.
- This map or a model as it is known as in machine learning paradigm is summarized knowledge representation of the raw data.
- The model may be in one of the following forms
    - If/else rules
    - Mathematical equations
    - Data structures like tree or graphs
    - Logical grouping of similar observations

Contd.

- The choice of model to solve a specific learning problem is a human task. The decision related to the choice of model is taken on multiple aspects like
  - The type of problem to be solved
  - Nature of input data
  - Domain of the problem
- Once the model is chosen the next task is to fit the model based on the input data.
- For ex:
  - In a case where the model is represented by a mathematical equation , say **‘y=c1 + c2x’**, based on the input data, we have to find out the values of c1 and c2.
  - Otherwise the equation is of no use.
  - So, fitting the model, in this case, means finding the values of unknown coefficients or constants of the equation or the model.
  - This process of fitting the model based on input data is known as training
  - Also the input data based on which model is being finalized is known as training data .

Generalization

- Next part is to tune up the abstracted knowledge to a form which can be used to take future decisions.
- This is achieved as part of generalization
- This part is quite difficult to achieve.
- This is because the model is trained based on a finite set of data, which may possess a limited set of characteristics.
- But when we want to apply the model to take decision on a set of unknown data, usually termed as test data, we may encounter two problems.
    - The trained model is aligned with the training data too much , hence may not portray the actual trend.
    - The test data possess certain characteristics apparently unknown to the training data.

Well posed Learning Problem

For defining a new problem, which can be solved using machine learning, a simple framework can be used:

1)What is the problem

-Describe the problem informally and formally (using T, P,E)

- List assumptions

-List similar problems

2) Why does the problem need to be solved?

-List the motivation to solve

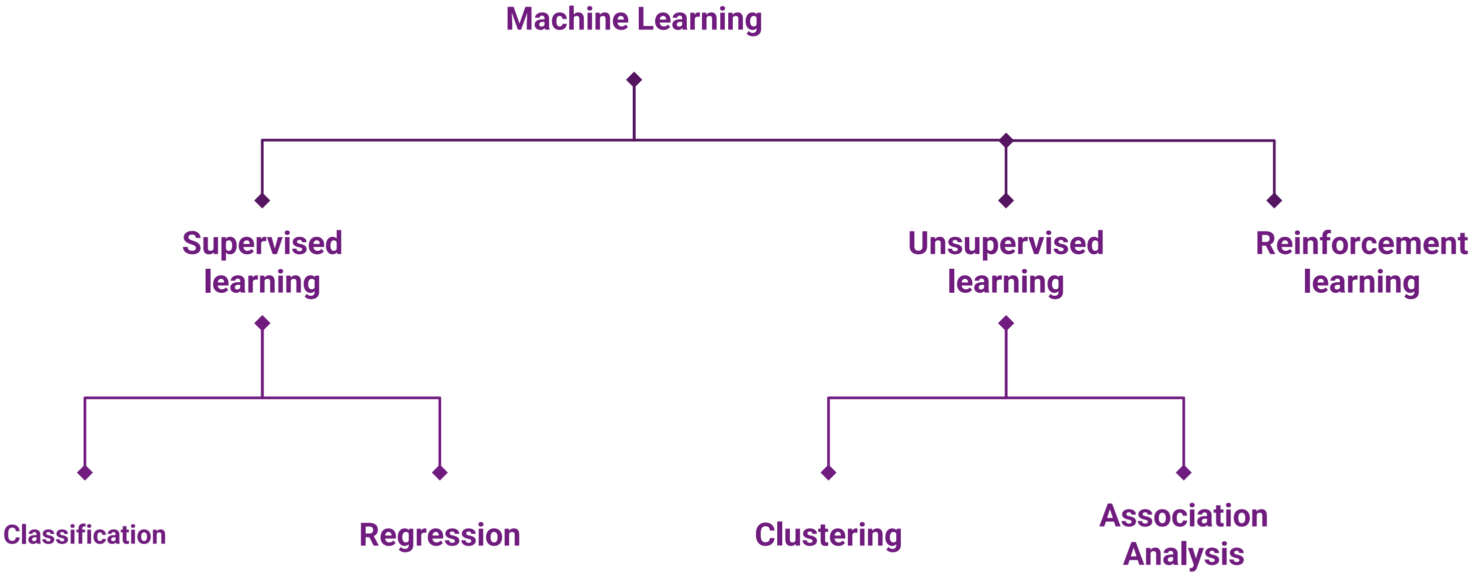
- benefits of solution

- how solution will be used

3)How would I solve the problem?

- Describe how the problem would be solved manually

Types of Machine Learning



Machine Learning

Supervised learning

Unsupervised learning

Reinforcement learning

Classification

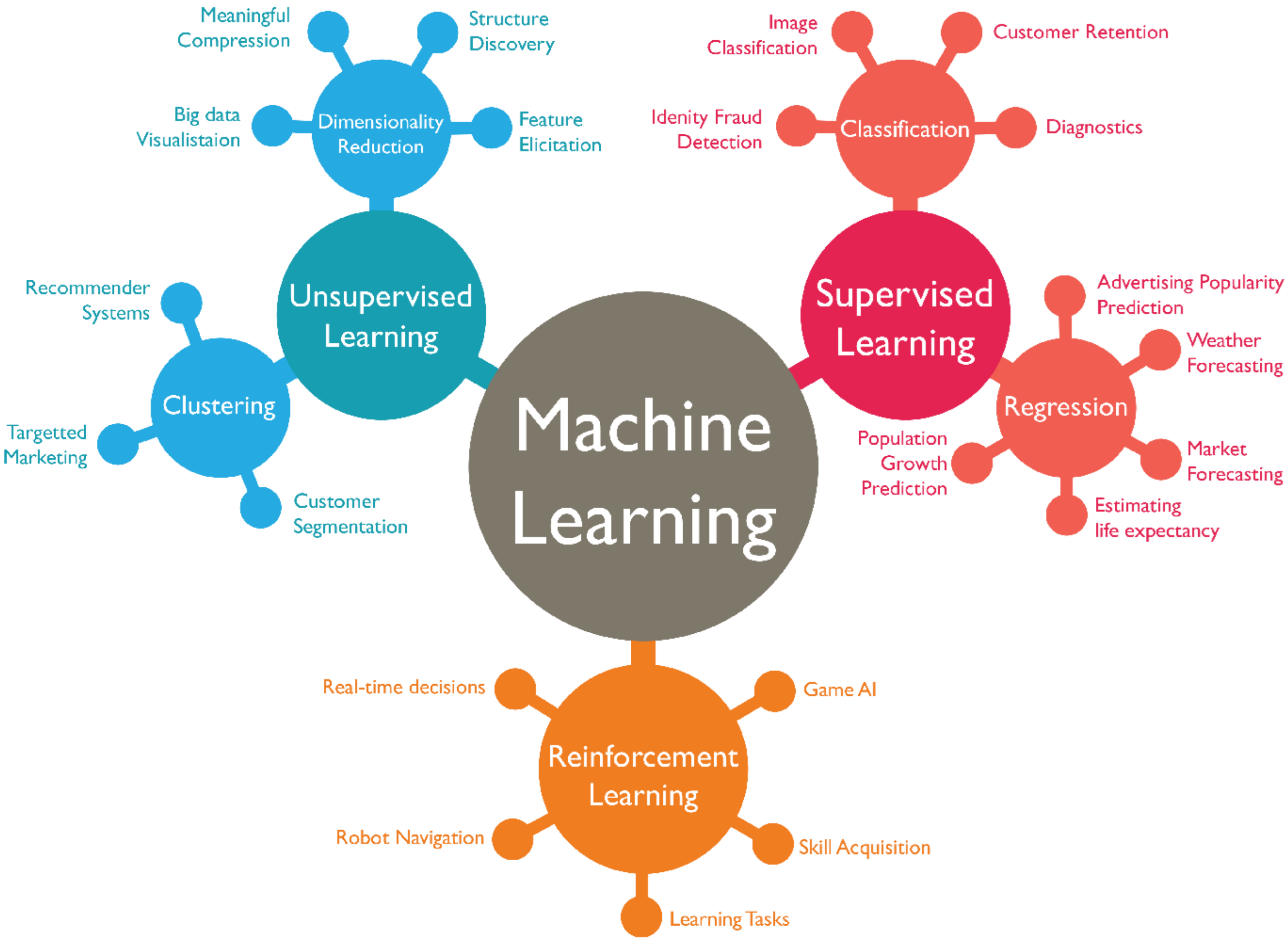
Regression

Clustering

Association Analysis

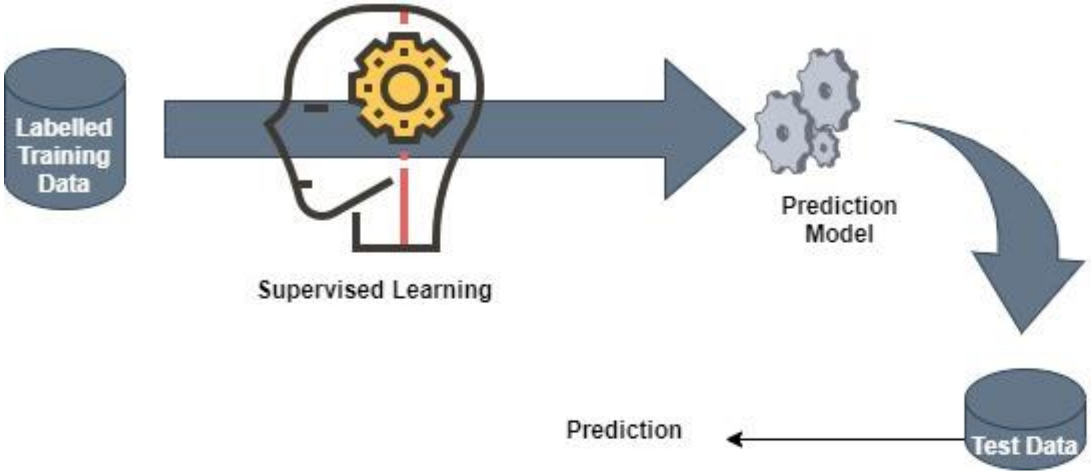
Types of Machine Learning

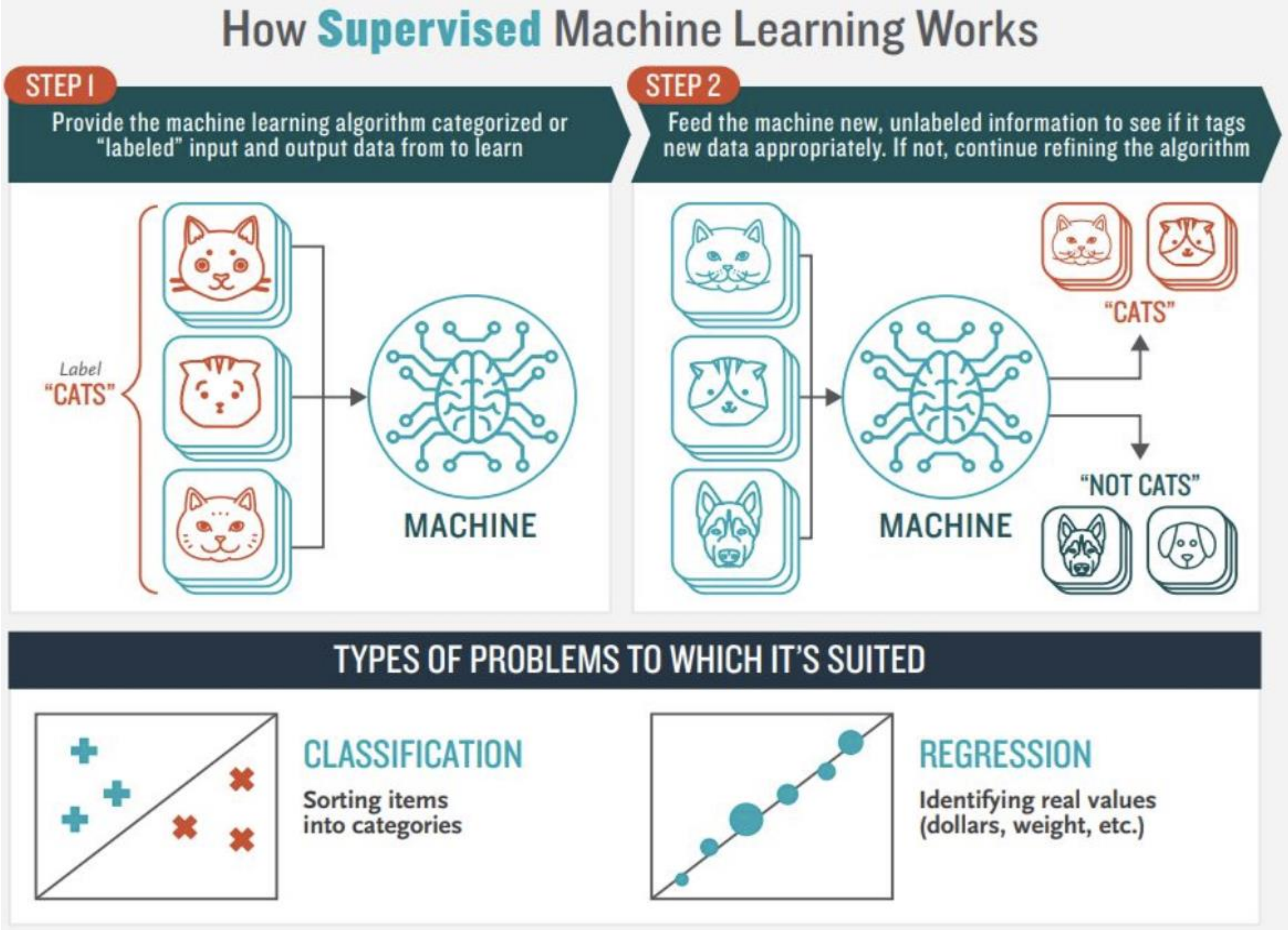
- Supervised learning:
  - Also called predictive learning . A machine predicts the class of unknown objects based on prior class-related information of similar objects.
- Unsupervised learning:
  - Also called descriptive learning. A machine finds patterns in unknown objects by grouping similar objects.
- Reinforcement learning:
  - A machine learns to act on its own to achieve the given goals.


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Supervised learning

- Learn from past information
  - It is the information about the task the machine has to execute.
- In context of definition of machine learning, this past information is the experience.





How Supervised Machine Learning Works

STEP 1

Provide the machine learning algorithm categorized or "labeled" input and output data from to learn

STEP 2

Feed the machine new, unlabeled information to see if it tags new data appropriately. If not, continue refining the algorithm

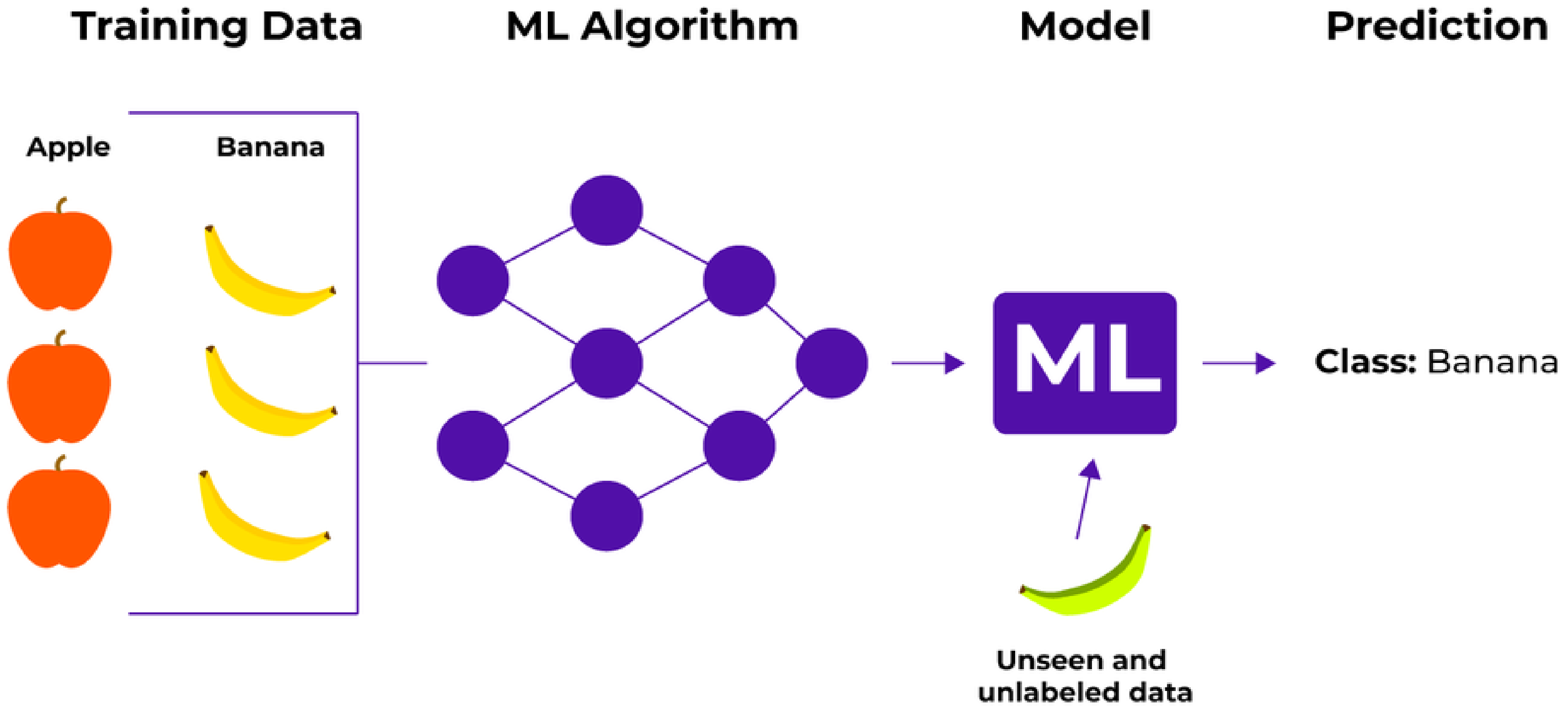
TYPES OF PROBLEMS TO WHICH IT'S SUITED

CLASSIFICATION

Sorting items into categories

REGRESSION

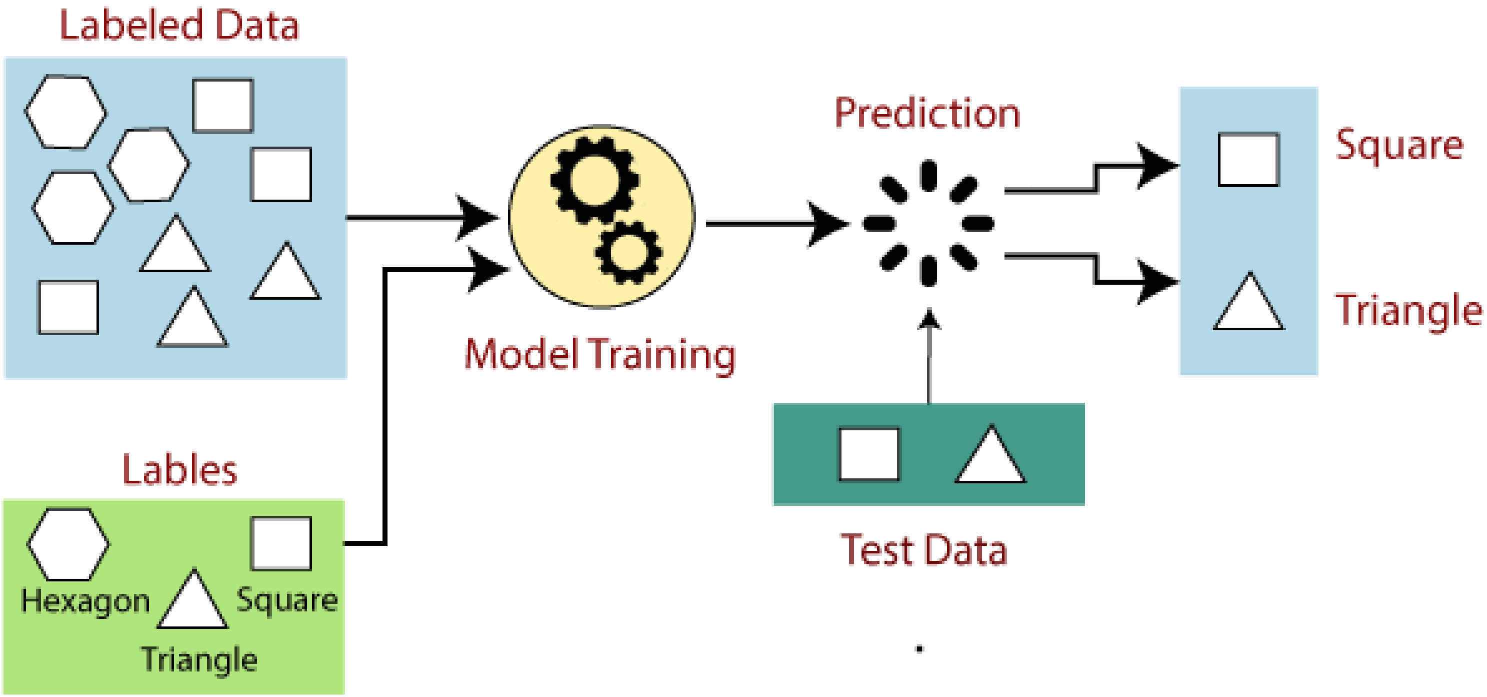
Identifying real values (dollars, weight, etc.)



Training Data
ML Algorithm
Model
Prediction
Apple
Banana
ML
Class: Banana
Unseen and unlabeled data

Supervised learning - example

- Say a machine is getting images of different objects as input and the task is to segregate the images by either shape or colour.
  - How can a machine know what is round shape or triangular shape?
  - How can a machine distinguish image of an object based on whether it is blue or green in color?
- A machine needs the basic information to be provided to it.
- The basic input is given in the form of **training data.**
- **Training data** will have past data on different aspects or features on a number of images along with the tag on whether the image is round or rectangular or blue or green in color.
- The tag is called **'label'** and we say training data is labelled in case of supervised learning.


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

Examples of supervised learning

- Predicting the results of a game
- Predicting whether the tumor is malignant(cancerous) or benign(non cancerous)
- Predicting the price of domains like real estate , stocks, etc
- Classifying texts such as classifying a set of emails as spam or not.

- When we are trying to predict a categorical or nominal variable, the problem is known as **classification problem.**
- Whereas when we are trying to predict a real values variable, the problem falls under the category of **regression.**

Question?

You're running a company, and you want to develop learning algorithms to address each of two problems. **Problem 1:You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.**

Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised. Should you treat these as classification or as regression problems?

1. Treat both as classification problems.
2. Treat problem 1 as a classification problem, problem 2 as a regression problem.
3. Treat problem 1 as a regression problem, problem 2 as a classification problem.
4. Treat both as regression problems.

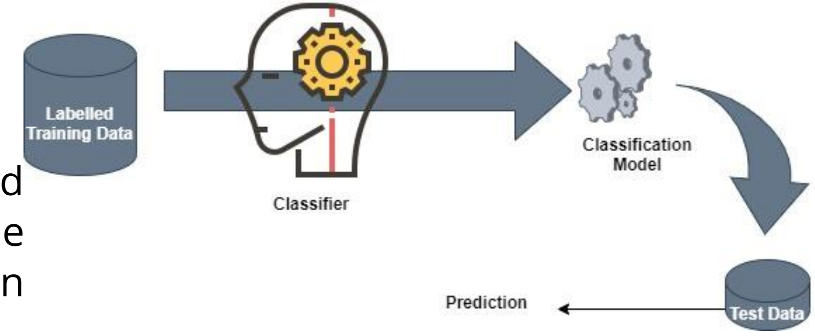
??

Suppose you are working on weather prediction, and use a learning algorithm to predict tomorrow's temperature (in degrees Centigrade/Fahrenheit). Would you treat this as a classification or a regression problem?

Suppose you are working on stock market prediction. You would like to predict whether or not a certain company will win a patent infringement lawsuit (by training on data of companies that had to defend against similar lawsuits). Would you treat this as a classification or a regression problem?

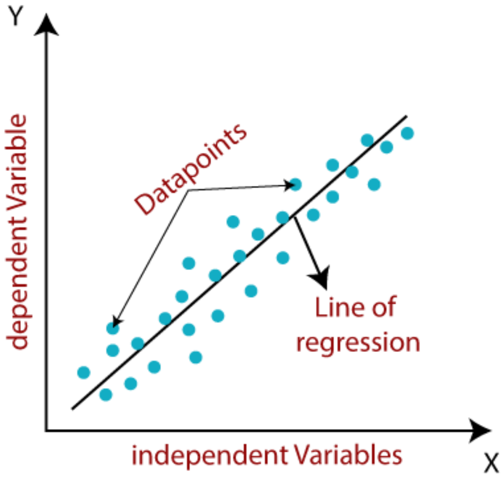
Machine learning algorithms for classification

- Some Machine learning algorithms for classification
  - Naive Bayes
  - Decision tree
  - k-Nearest Neighbour algorithm
- In summary, classification is a type of supervised learning where a target feature which is of type categorical is predicted for test data based on information imparted by training data.
- Some typical classification problems include:
  - Image classification
  - Prediction of disease
  - Win loss prediction of games
  - Prediction of natural calamity
  - Recognition of handwriting



Regression

- In linear regression the objective is to predict numerical features like real estate or stock price, temperature, marks in an examination , sales revenue etc.
- The underlying variables are continuous in nature.
- In case of linear regression,a straight line relationship is fitted between the predictor variables and target variables using statistical concept of least squares method.
- In least squares method, the sum of square of error between actual and predicted values of the target variable is tried to be minimized.
- In case of simple linear regression, there is only one predictor variable whereas in case of multiple linear regression, multiple predictor variables can be included in the model.



Some real life examples

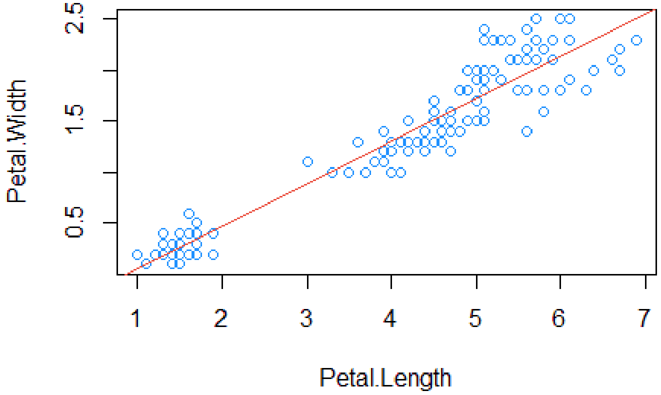
Example 1:

Given data about the size of houses on the real estate market, try to predict their price. Price as a function of size is a continuous output, so this is a regression problem.

Example 2:

Given a picture of a person, we have to predict their age on the basis of the given picture

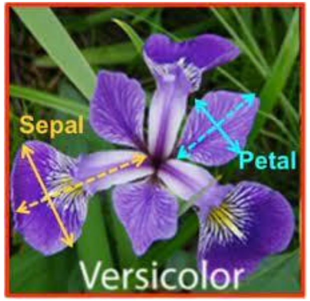
Regression- Contd



- A typical linear regression model can be represented in the form-
  - y= $\alpha + \beta$x
  - Where x is predictor variable and y is target variable
- Given figure here, the input data comes from a famous multivariate data set names Iris introduced by the British statistician and biologist Ronald Fisher.
- The data set consists of 50 samples from each of three species of Iris- Iris setosa, Iris virginica and Iris versicolor.
- Four features were measured for each sample to distinguish different species of flower-
  - Sepal length
  - Sepal width
  - Petal length
  - Petal width

Regression- Contd.

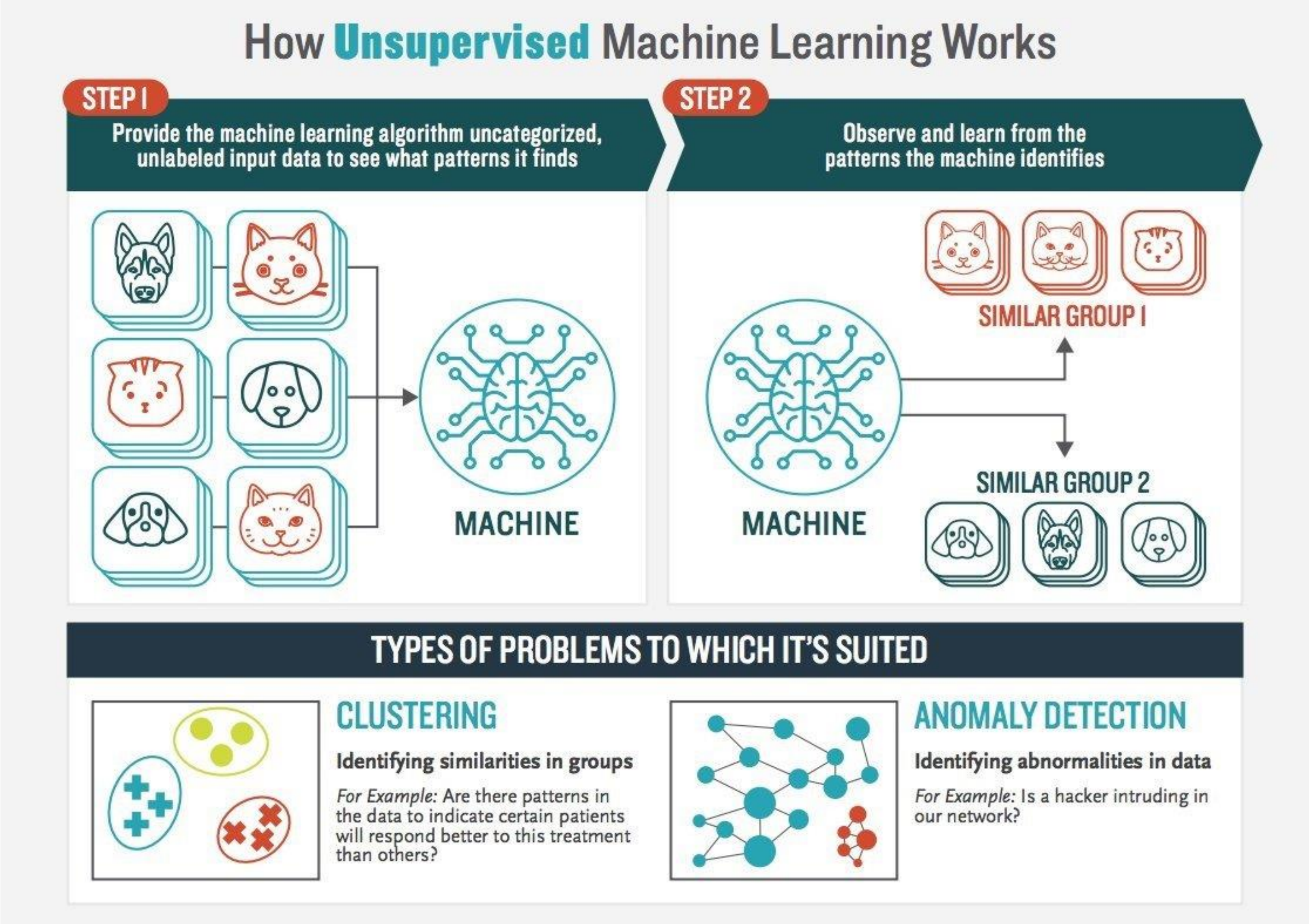
- The iris dataset is typically used as a training data for solving the classification problem of predicting the flower species based on feature values.
- But we can also demonstrate regression using this data set by predicting the value of one feature using another feature as predictor.
- In the figure given previously, petal length is predictor variable , which helps in predicting the value of target variable petal width.
- Typical applications of regression
  - Forecasting in retails
  - Sales prediction
  - Price prediction
  - Weather forecasting
  - Skill demand forecasting



Iris Versicolor Iris Setosa Iris Virginica

Unsupervised learning

- There is no labelled training data to learn from and no prediction to be made.
- The objective is to take a dataset as input and try to find natural groupings or patterns within the data
- It is often termed as descriptive model and the process of unsupervised learning is referred to as pattern discovery or knowledge discovery.
- Clustering is the main type of unsupervised learning.
    - It intends to group or organize similar objects together.
    - Objects belonging to the same cluster are quite similar to each other while objects belonging to different clusters are quite dissimilar.
    - Objective of clustering is to discover the intrinsic grouping of unlabelled data and form clusters .
    - Different measures of similarity can be applied for clustering



How Unsupervised Machine Learning Works

STEP 1

Provide the machine learning algorithm uncategorized, unlabeled input data to see what patterns it finds

STEP 2

Observe and learn from the patterns the machine identifies

TYPES OF PROBLEMS TO WHICH IT'S SUITED

CLUSTERING

Identifying similarities in groups

For Example: Are there patterns in the data to indicate certain patients will respond better to this treatment than others?

ANOMALY DETECTION

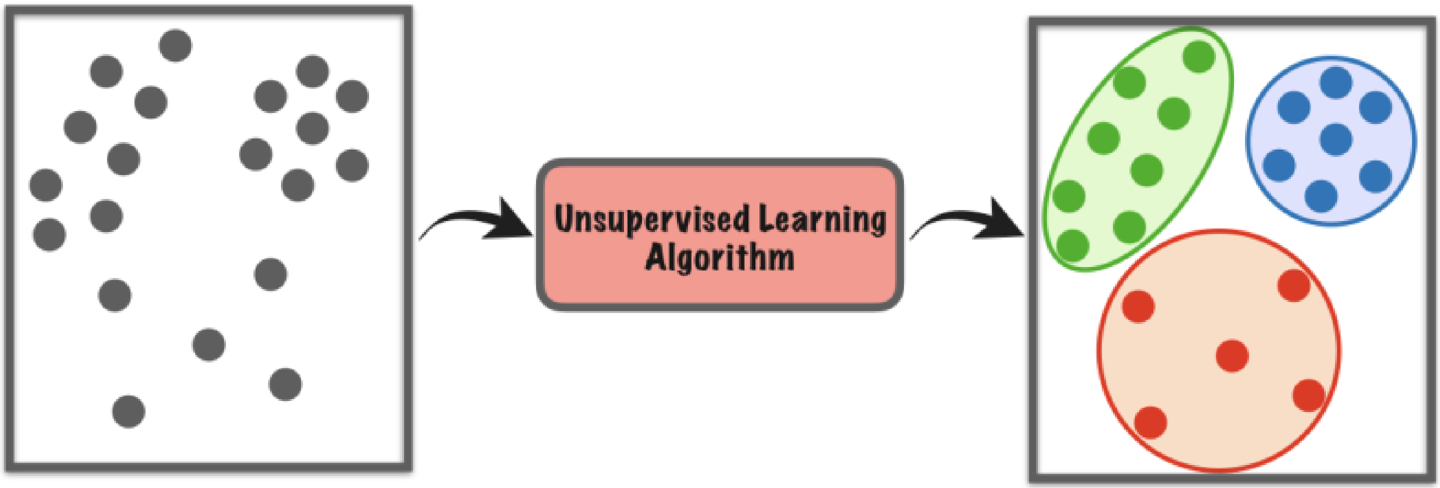
Identifying abnormalities in data

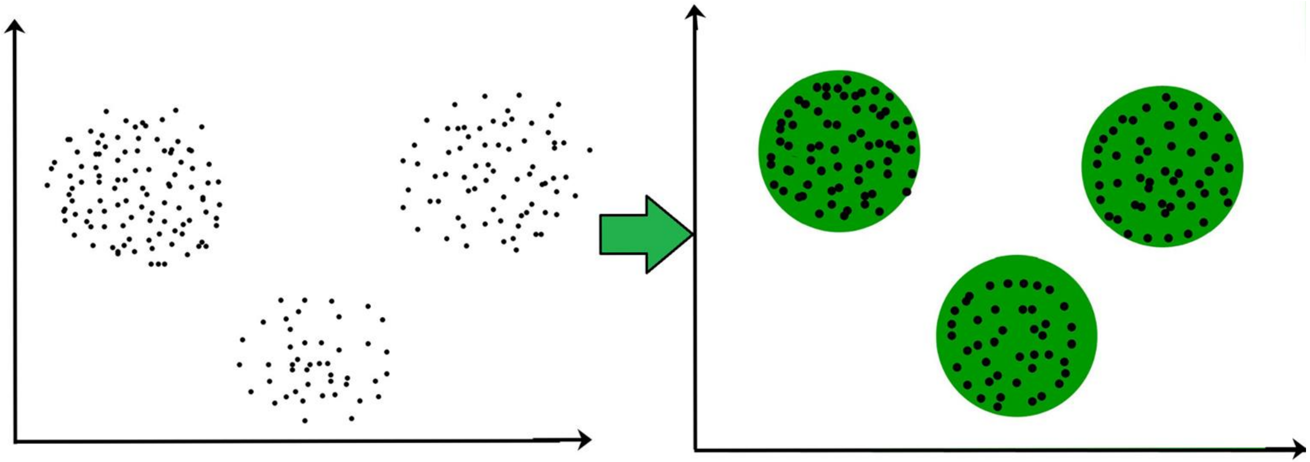
For Example: Is a hacker intruding in our network?

Question?

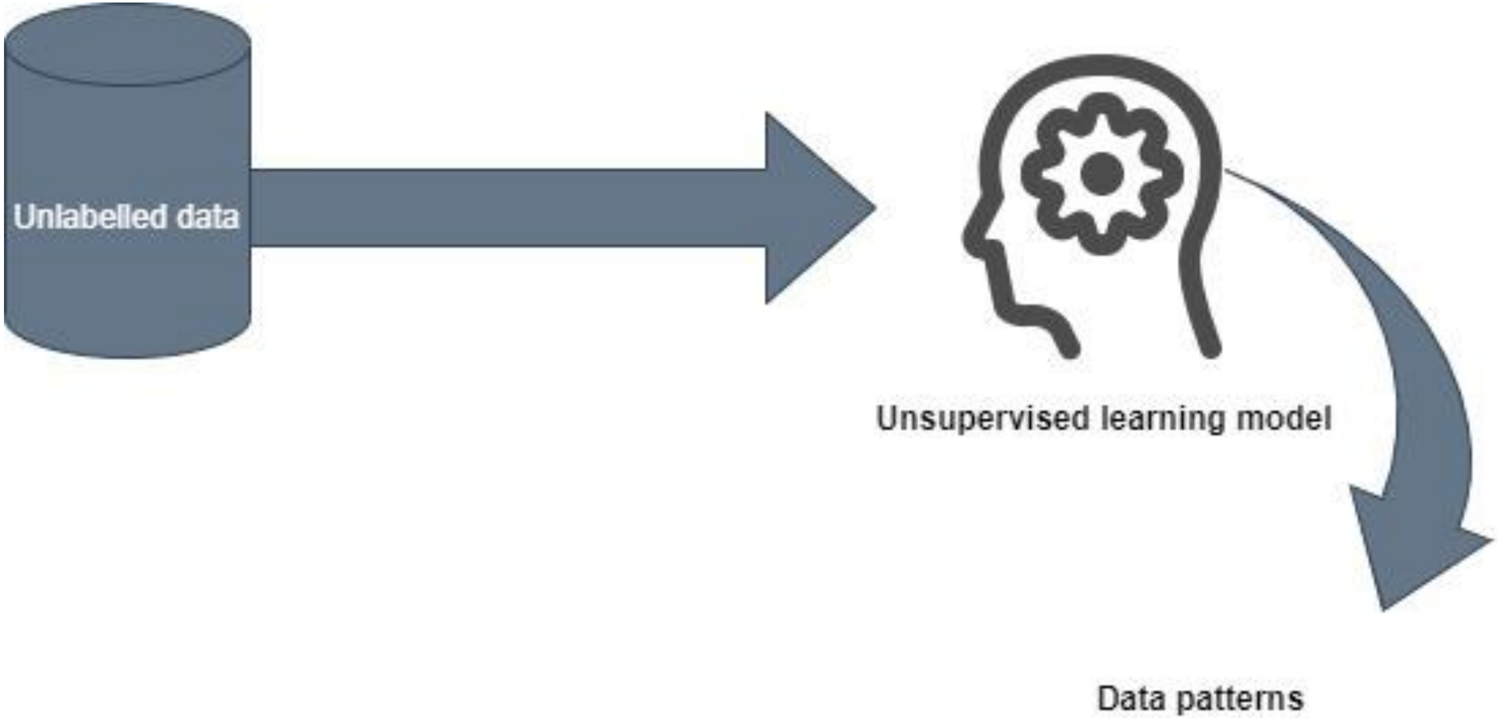
Of the following examples, which would you address using an unsupervised learning algorithm? (Check all that apply.)

1. Given email labeled as spam/not spam, learn a spam filter.
2. Given a set of news articles found on the web, group them into sets of articles about the same stories.
3. Given a database of customer data, automatically discover market segments and group customers into different market segments.
4. Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.


Unsupervised Learning
Algorithm



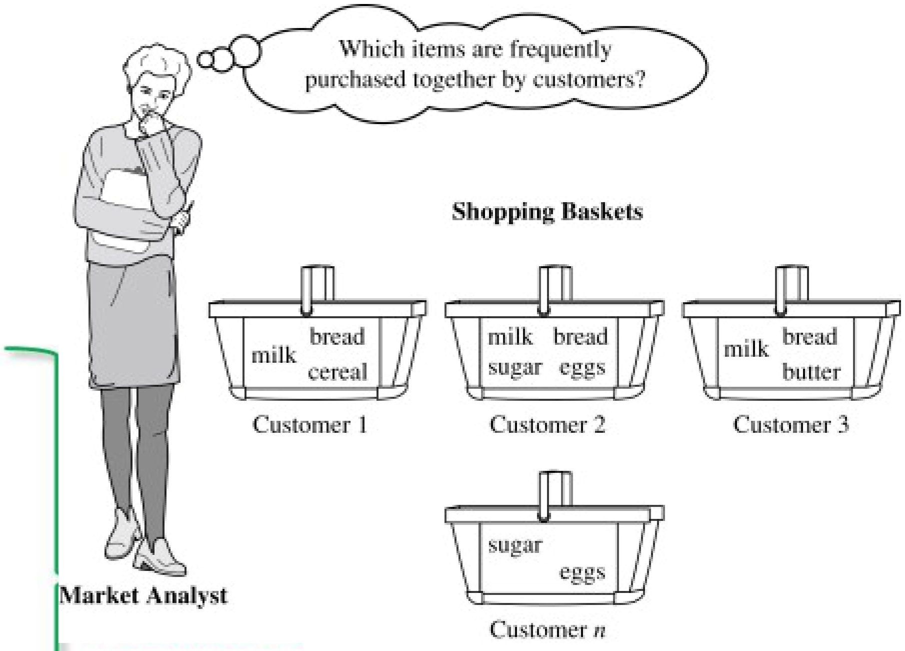
- Common similarity measure is distance.
- Two data items are considered a part of the same cluster if the distance between them is less.
- If the distance between the data items is high, the items do not generally belong to the same cluster.
- This is known as distance based clustering.


Unlabelled data
Unsupervised learning model
Data patterns

Association analysis

- One more variant of unsupervised learning.
- Association between data items is identified.
- Examples: Market basket analysis.
    - From past transaction data in a grocery store, it may be observed that most of the customers who have bought item A, have also bought item B and item C or atleast one of them.
    - This means that there is a strong association of the event ‘purchase of item A’ with the event ‘purchase of item B’ or ‘purchase of item C’ .
    - Identifying these sort of associations is the goal of association analysis.
- Applications:
    - Market basket analysis
    - Recommender systems

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
...	...



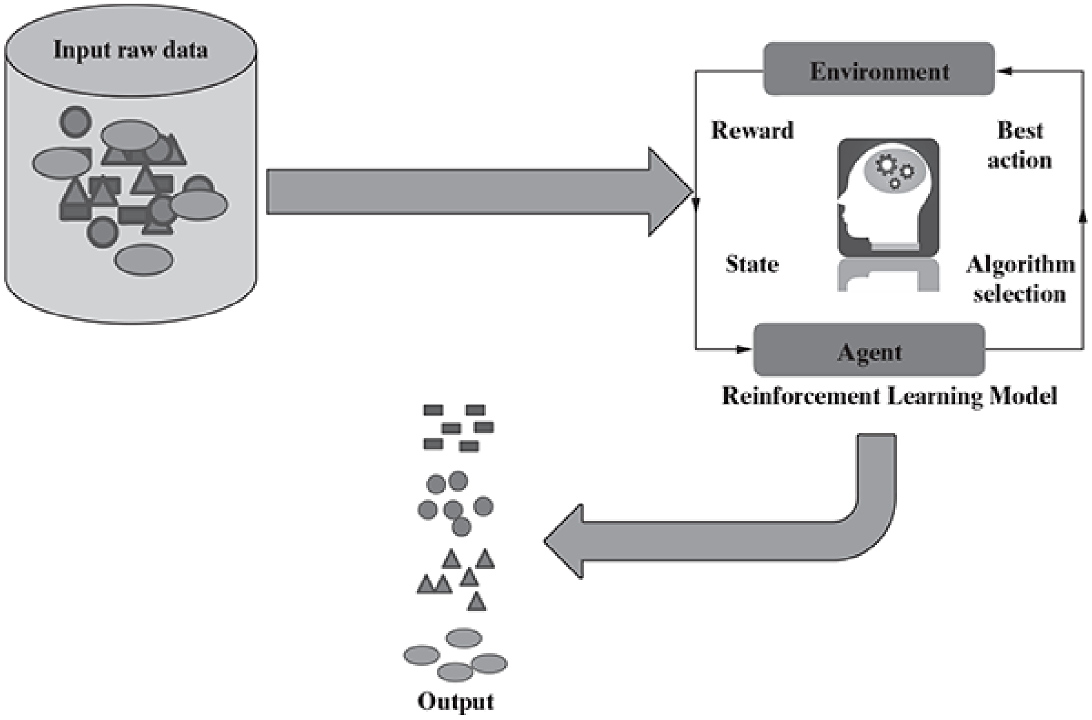
{Diapers, Beer} Example of a frequent itemset

{Diapers} → {Beer} Example of an association rule

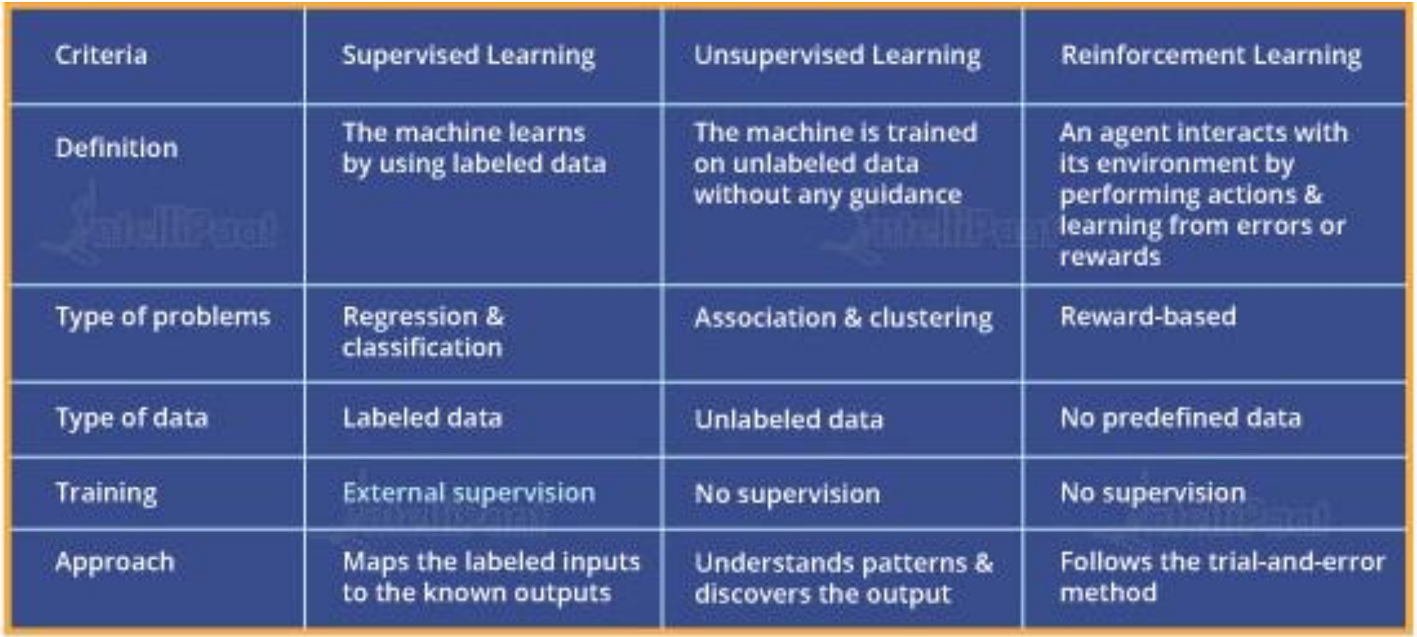
Reinforcement learning

- Example: We have seen babies learning to walk without any prior knowledge of how to do it.
  - First they notice how others do it.
  - They understand that legs have to be used, one at a time, to take a step
  - While walking, sometimes they fall down hitting an obstacle, whereas other times they are able to walk smoothly
  - Babies might get a reward like clapping of hands by parents or chocolates.
  - Obviously no claps when baby falls.
  - Slowly a time comes when the babies learn from mistakes and are able to walk with much ease
- In the same way, machines often learn to do tasks automatically.
- Machine is given a task with hurdles.
- It tries to improve its performance of doing task .
- When a sub task is completed successfully, a reward is given.
- When a sub task is not performed successfully no reward is given
- This continues until the task is completed successfully.
- This process of learning is called reinforcement learning
- Applications
  - Self driving cars

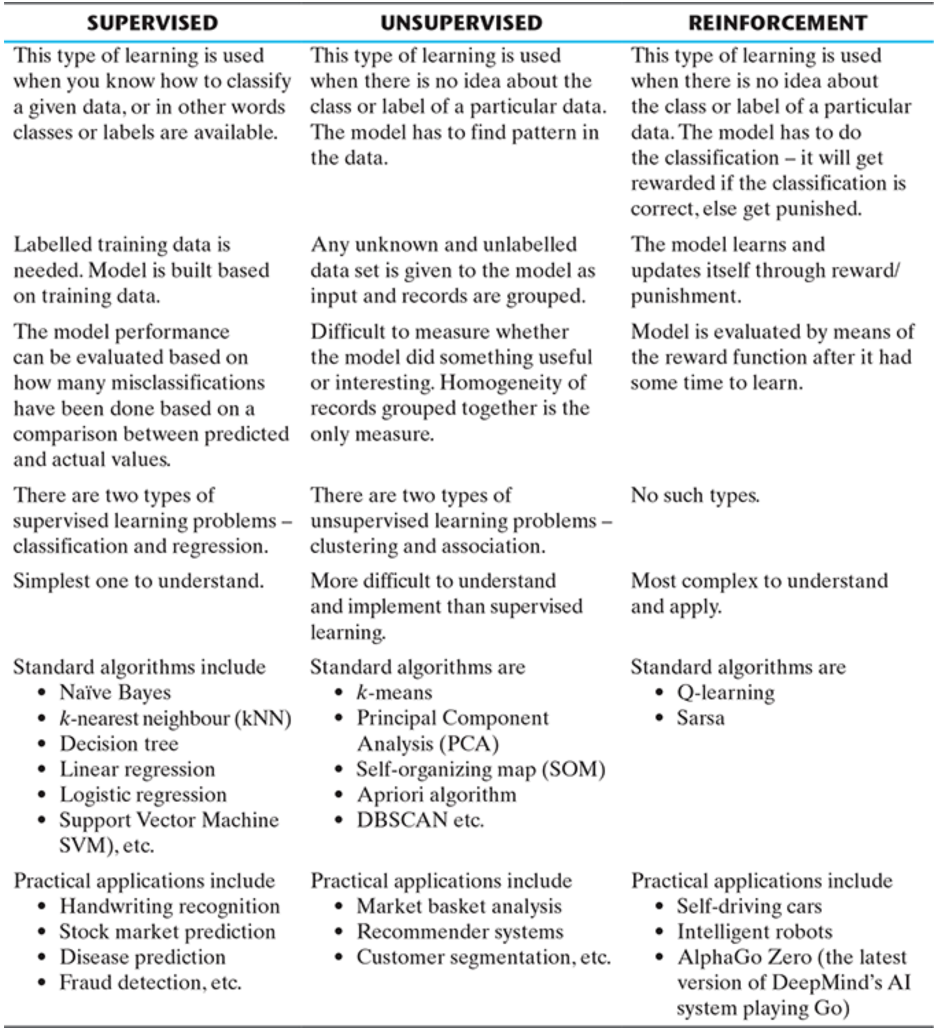
Reinforcement learning



Comparison - supervised , unsupervised and reinforcement learning



Criteria	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Definition	The machine learns by using labeled data	The machine is trained on unlabeled data without any guidance	An agent interacts with its environment by performing actions & learning from errors or rewards
Type of problems	Regression & classification	Association & clustering	Reward-based
Type of data	Labeled data	Unlabeled data	No predefined data
Training	External supervision	No supervision	No supervision
Approach	Maps the labeled inputs to the known outputs	Understands patterns & discovers the output	Follows the trial-and-error method



SUPERVISED	UNSUPERVISED	REINFORCEMENT
This type of learning is used when you know how to classify a given data, or in other words classes or labels are available.	This type of learning is used when there is no idea about the class or label of a particular data. The model has to find pattern in the data.	This type of learning is used when there is no idea about the class or label of a particular data. The model has to do the classification – it will get rewarded if the classification is correct, else get punished.
Labelled training data is needed. Model is built based on training data.	Any unknown and unlabelled data set is given to the model as input and records are grouped.	The model learns and updates itself through reward/ punishment.
The model performance can be evaluated based on how many misclassifications have been done based on a comparison between predicted and actual values.	Difficult to measure whether the model did something useful or interesting. Homogeneity of records grouped together is the only measure.	Model is evaluated by means of the reward function after it had some time to learn.
There are two types of supervised learning problems – classification and regression.	There are two types of unsupervised learning problems – clustering and association.	No such types.
Simplest one to understand.	More difficult to understand and implement than supervised learning.	Most complex to understand and apply.
Standard algorithms include • Naïve Bayes • *k*-nearest neighbour (kNN) • Decision tree • Linear regression • Logistic regression • Support Vector Machine SVM), etc.	Standard algorithms are • *k*-means • Principal Component Analysis (PCA) • Self-organizing map (SOM) • Apriori algorithm • DBSCAN etc.	Standard algorithms are • Q-learning • Sarsa
Practical applications include • Handwriting recognition • Stock market prediction • Disease prediction • Fraud detection, etc.	Practical applications include • Market basket analysis • Recommender systems • Customer segmentation, etc.	Practical applications include • Self-driving cars • Intelligent robots • AlphaGo Zero (the latest version of DeepMind's AI system playing Go)

Question?

Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. Which of the following would you apply **supervised learning** to? (Select all that apply.) In each case, assume some appropriate dataset is available for your algorithm to learn from.

1. Given a large dataset of medical records from patients suffering from heart disease, try to learn whether there might be different clusters of such patients for which we might tailor separate treatments.
2. Have a computer examine an audio clip of a piece of music, and classify whether or not there are vocals (i.e., a human voice singing) in that audio clip, or if it is a clip of only musical instruments (and no vocals).
3. Given data on how 1000 medical patients respond to an experimental drug (such as effectiveness of the treatment, side effects, etc.), discover whether there are different categories or "types" of patients in terms of how they respond to the drug, and if so what these categories are.
4. Given genetic (DNA) data from a person, predict the odds of him/her developing diabetes over the next 10 years.

Applications of Machine learning

Three major domains where machine learning is applied:

- Banking and finance
  - Identifying fraudulent transactions
  - To maintain the customers so that they don’t leave the bank.
  - To identify customers who are vulnerable to leave
- Insurance
  - Risk prediction during new customer onboarding
  - Claims management- whether fraudulent ?
- Healthcare
  - Predict health conditions
  - Person is alerted to take preventive actions
  - Machine learning with computer vision also plays an important role in disease diagnosis from medical imaging

Tools in Machine learning

1. Python
   a. Most popular open source programming languge
   b. Numpy - mathematical functions
   c. Matplotlib - numerical plotting
   d. Scipy- mathematical tools
   e. Scikit-learn- for classification, regression and clustering algorithms
2. R
   a. Used for Statistical computing and data analysis
   b. Open source
3. Matlab
   a. Matrix laboratory
   b. Licensed commercial software
   c. Used for numerical computing
4. SAS- Statistical Analysis System
   a. Licensed commercial software
   b. Strong support for machine learning functionalities

??

This type of learning to be used when there is no idea about the class or label of particular data

A. Supervised learning
B. Unsupervised learning
C. Reinforcement learning

The model learns and updates itself through reward/punishment in case of

A. Supervised learning
B. Unsupervised learning
C. Reinforcement learning

Which of the following is NOT an example of regression?

A. Estimating the amount of rain
B. Predicting the demand for a product
C. Determining whether power usage will rise or fall
D. Predicting the price of a stock