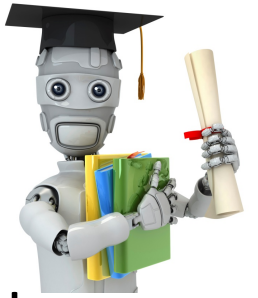
INTRODUCTION TO MACHINE LEARNING

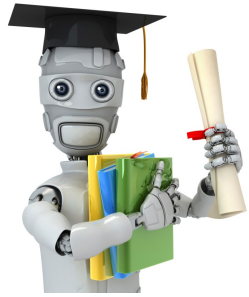
Dr. Debanjan Pathak

Basic Understanding of Machine Learning



- Traditional Programming: Data and program is run on the computer to produce the output.
- Machine Learning: Data and output is run on the computer to create a program. This program can be used in traditional programming.
- Machine Learning is an application of artificial intelligence where a computer/machine learns from the past experiences and makes future predictions.
  - The past experience is developed through the data collected.
  - In order to perform the task, the system learns from the data-set provided.
  - A data-set is a collection of many examples.
  - An example is a collection of features.
- Machine Learning refers to the techniques involved in dealing with vast data in the most intelligent fashion (by developing algorithms) to derive actionable insights.

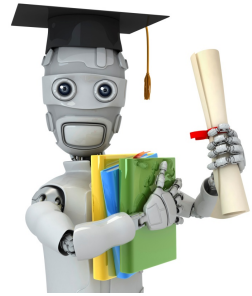
Basic Understanding of Machine Learning



Machine Learning definition

- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.
- Tom Mitchell (1998) Well-posed Learning Problem: A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

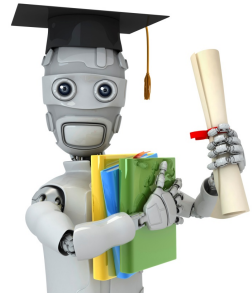
Basic Understanding of Machine Learning



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?

- ◯ Classifying emails as spam or not spam.
- ◯ Watching you label emails as spam or not spam.
- ◯ The number (or fraction) of emails correctly classified as spam/not spam.
- ◯ None of the above—this is not a machine learning problem.

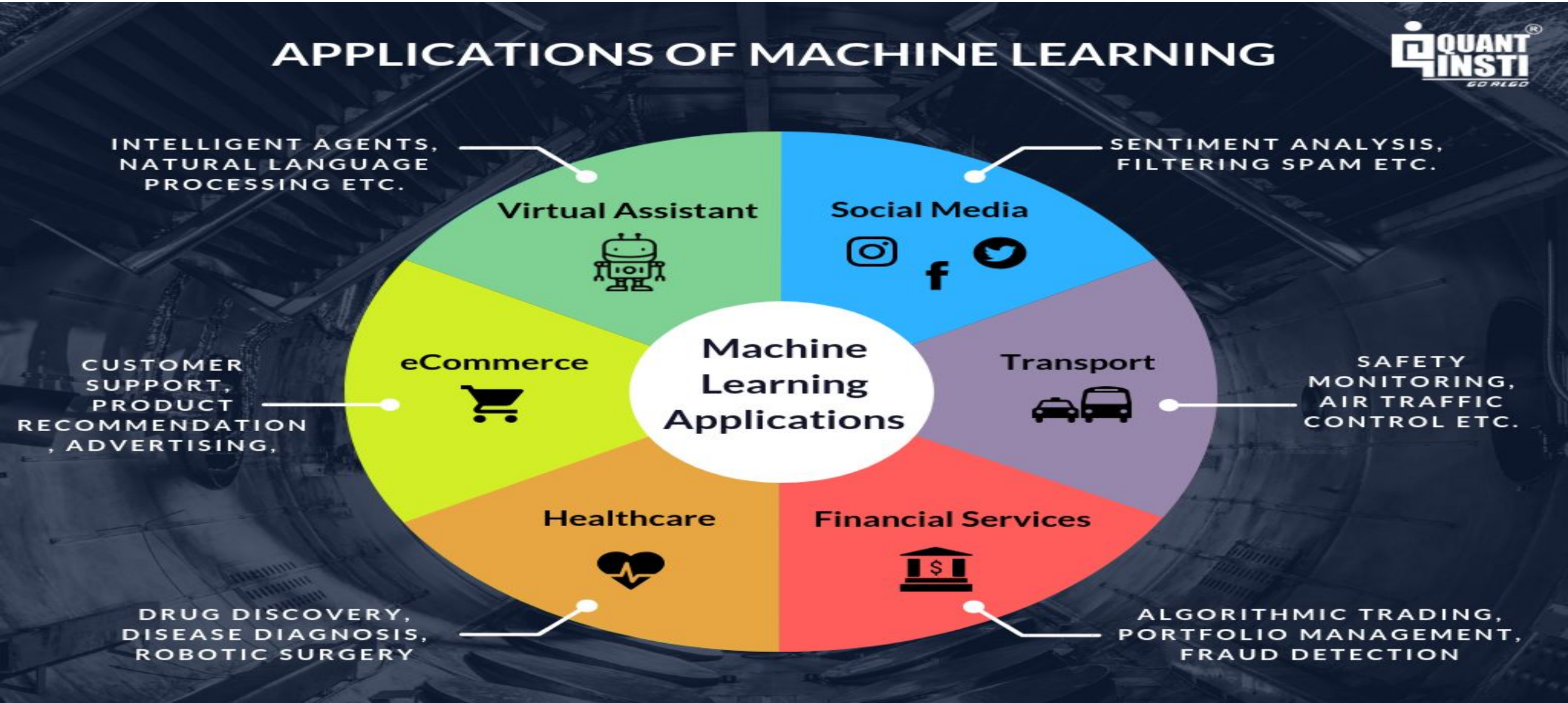
Basic Understanding of Machine Learning



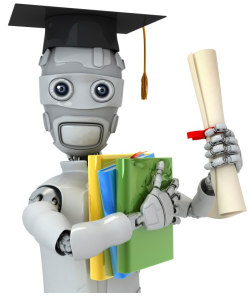
Brief History

- **1950**
- Samuel developed checker playing program and he coined the term Machine learning
- **1960**
- Neural Network- Rosenblatt perceptron
- Pattern Recognition
- Minsky and Papert proved limitation of Perceptron
- **1970**
- Decision tree – J.R Quinlan
- Natural Language processing
- **1980**
- Advanced decision tree and rule based learning
- Resurgence of neural network- Multilayer perceptron and neural network specific back propagation algorithm was developed
- PAC- probably approximate correct learning
- **1990 (Machine learning embraced statistics to a large extent)**
- SVM- 1995
- Data mining
- Adaptive agents and web applications
- Text learning
- Reinforcement learning
- Ensembles
- Bayes learning

Basic Understanding of Machine Learning



Basic Understanding of Machine Learning



Many domains and application

- **Medicine**
- Diagnose a disease
  - Input: Symptoms , Lab measurement, test results, DNA tests,……
  - Output: one of set of possible disease , or none of the above
- Data mine historical medical records to learn which future patients will respond best to which treatment.

- **Robot control**
- Design autonomous mobile robots that learn to navigate from their own experience

- **Financial**
- Predict if a stock will rise or fall in few millisec
- Predict if a user will click on an ad or not in order to decide which ad to show.

- **Application in Business intelligence**
- Robustly forecasting product sale quantities taking seasonality and trend into account
- Identify price sensitivity of a consumer product and identify the optimum price point that maximizes the net profit.

- **Some other applications**
- Fraud detection: Credit card providers determines whether or not someone will default.
- Understanding the consumer sentiment based on unstructured text data.
- Self customized program
  - e.g. Amazon, Netflix product recommendation
  - Algorithm learns by itself to customize.

Basic Understanding of Machine Learning

- Examples of applications in diverse fields
  - Machine Vision
  - Biometric Recognition
  - Handwriting Recognition
  - Medical Diagnosis
  - Alignment of Biological Sequences
  - Drug Design
  - Speech Recognition
  - Text Mining
  - Natural Language Processing
  - Fault Diagnostics
  - Load Forecasting
  - Control and Automation
  - Business Intelligence

Basic Understanding of Machine Learning

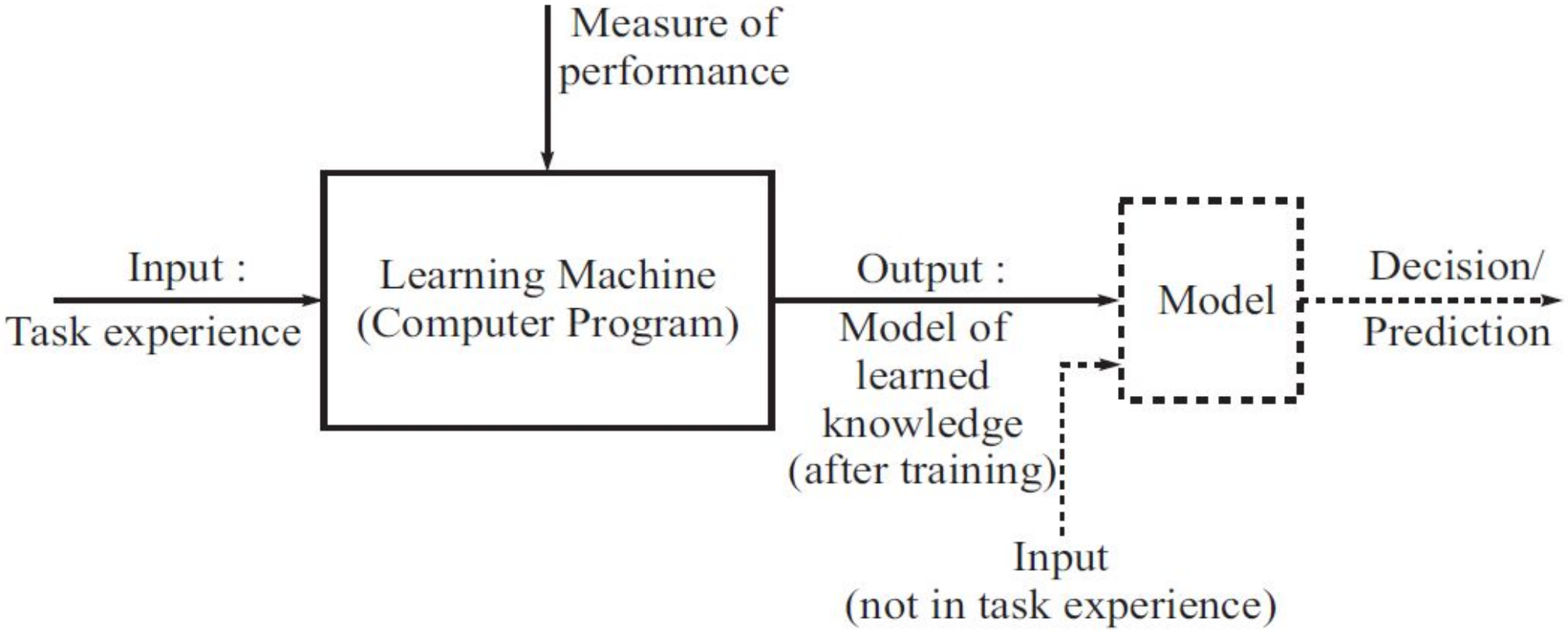
- The field of machine learning is concerned with the question of how to construct computer programs that automatically **improve with experience**.
- In recent years many successful machine learning applications have been developed, ranging from **data-mining programs** that learn to detect fraudulent credit card transactions, to **information-filtering systems** that learn users' reading preferences, to **autonomous vehicles** that learn to drive on public highways.
- At the same time, there have been important **advances** in the theory and algorithms that form the foundations of this field.
- Machine learning **draws on concepts and results from** many fields, including **statistics, artificial intelligence, philosophy, information theory, biology, cognitive science, computational complexity, and control theory.**

Basic Understanding of Machine Learning

- **How is machine learning different from ____**
  - **Artificial Intelligence:** It refers to the procedure of programming a computer (machine) to take rational decision.
    - to make the machine behave in an excellent fashion in lieu of human guidance.
    - ML is a subset of AI.
  - **Statistics:** utilizes data to carry out the analysis and present inferences.
    - regression,variance, standard deviation, conditional probability etc.
    - Machine learning algorithms uses statistical concepts to execute machine learning
  - **Deep Learning:** associated with a ML algorithm (ANN) which uses the concept of human brain to facilitate the modeling of arbitrary functions.
    - ANN requires a vast amount of data and this algorithm is highly flexible when it comes to model multiple outputs simultaneously.
  - **Data Mining:** deals with searching specific information
    - ML solely concentrates on performing a given task

Basic Understanding of Machine Learning

- Block diagrammatic representation of a learning machine



Basic Understanding of Machine Learning

- Steps used in Machine Learning

1. Collecting data: Be it the raw data from excel, access, text files etc., this step (gathering past data) forms the foundation of the future learning.
   - The better the variety, density and volume of relevant data, better the learning prospects for the machine becomes.
2. Preparing the data: Any analytical process thrives on the quality of the data used.
   - One needs to spend time determining the quality of data and then taking steps for fixing issues such as missing data and treatment of outliers.
3. Training a model: This step involves choosing the appropriate algorithm and representation of data in the form of the model.
   - The cleaned data is split into two parts – train and test (proportion depending on the prerequisites);
   - the first part (training data) is used for developing the model.
   - The second part (test data), is used as a reference.

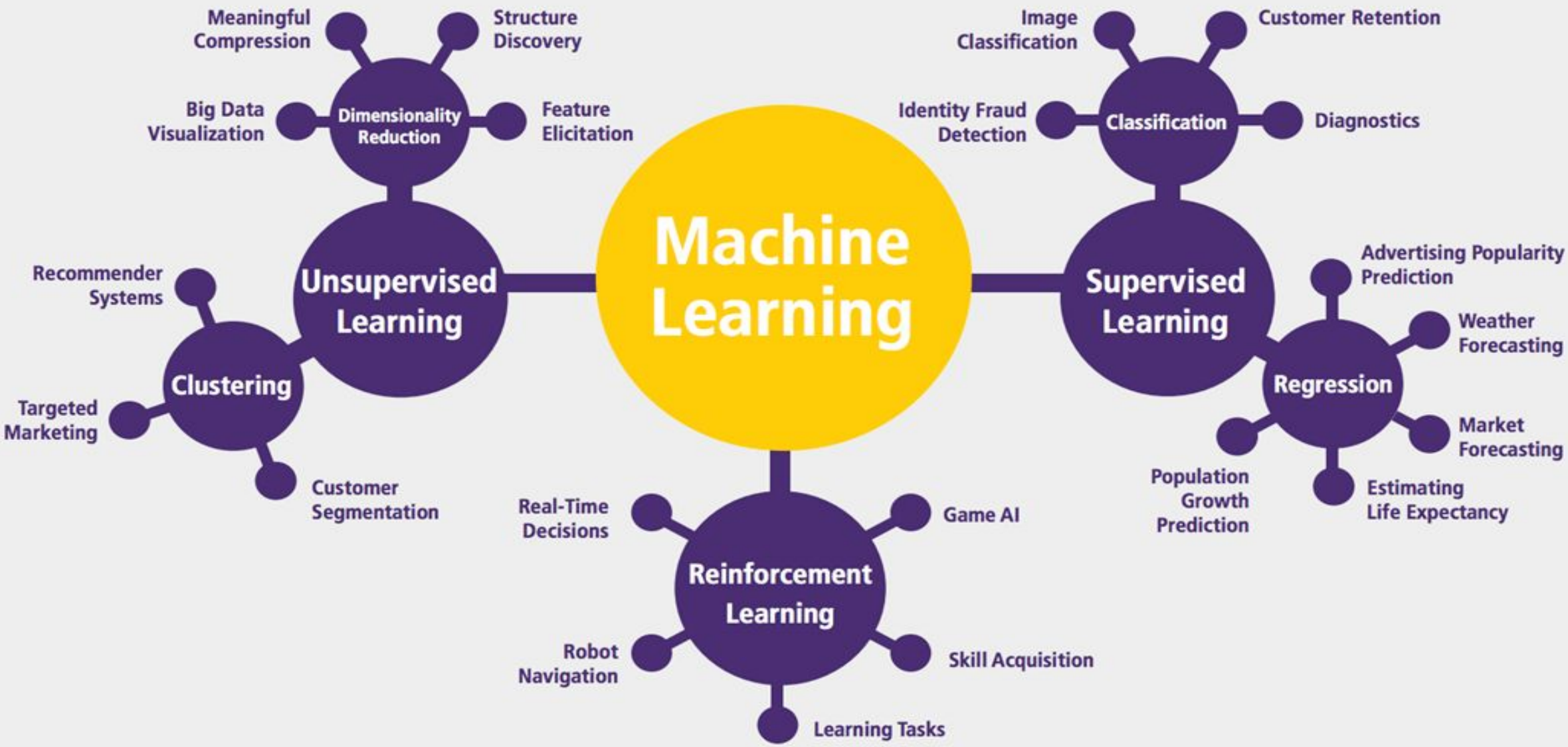
Basic Understanding of Machine Learning

- Steps used in Machine Learning

4. Evaluating the model: To test the accuracy, the second part of the data (holdout / test data) is used.
    - This step determines the precision in the choice of the algorithm based on the outcome.
    - A better test to check accuracy of model is to see its performance on data which was not used at all during model build.
5. Improving the performance: This step might involve choosing a different model altogether or introducing more variables to augment the efficiency.
    - That's why significant amount of time needs to be spent in data collection and preparation.

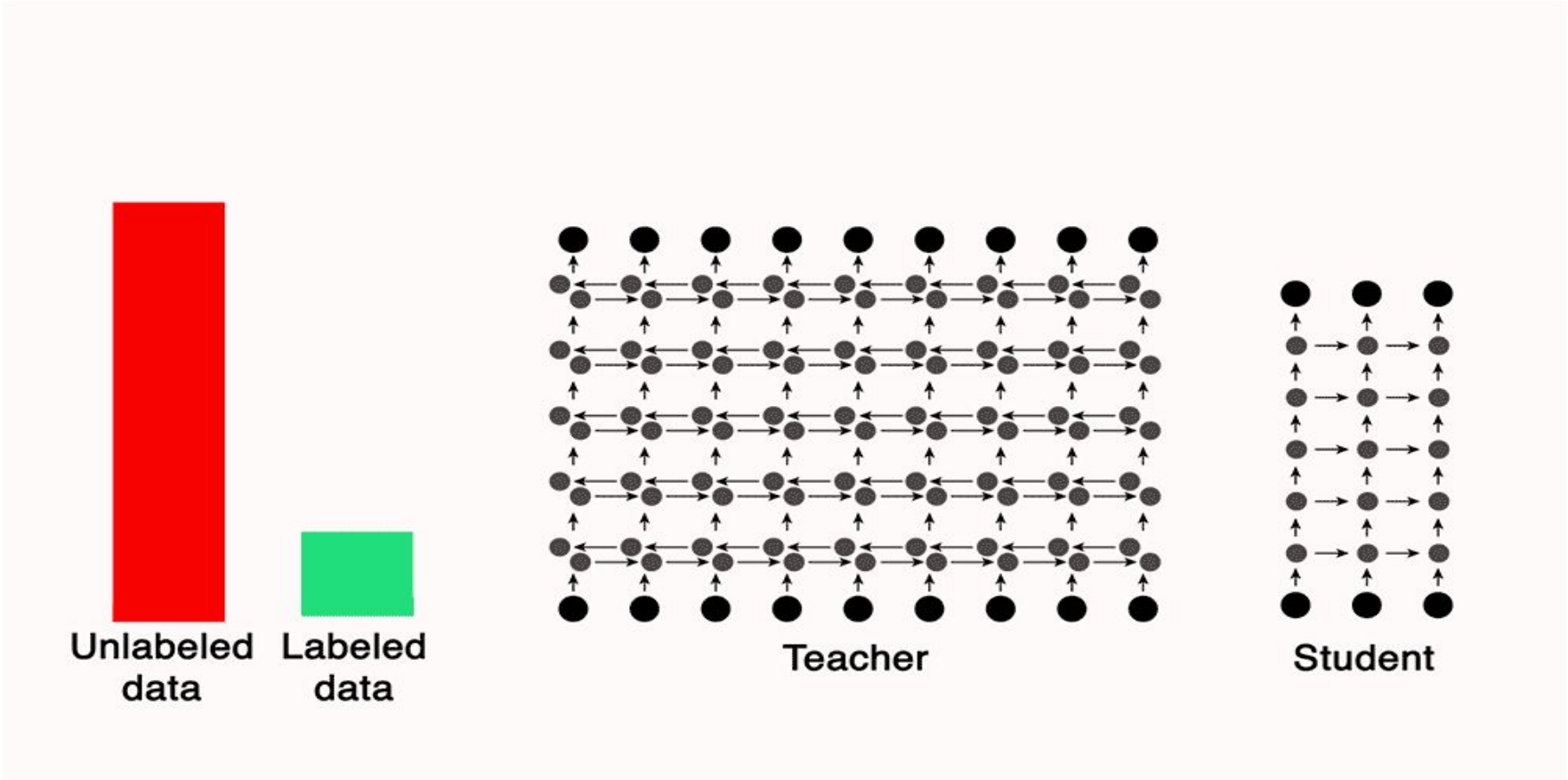
Type of Machine Learning Problem: Supervised, Unsupervised and Reinforced

Types of Machine Learning algorithms



Supervised/Directed Learning

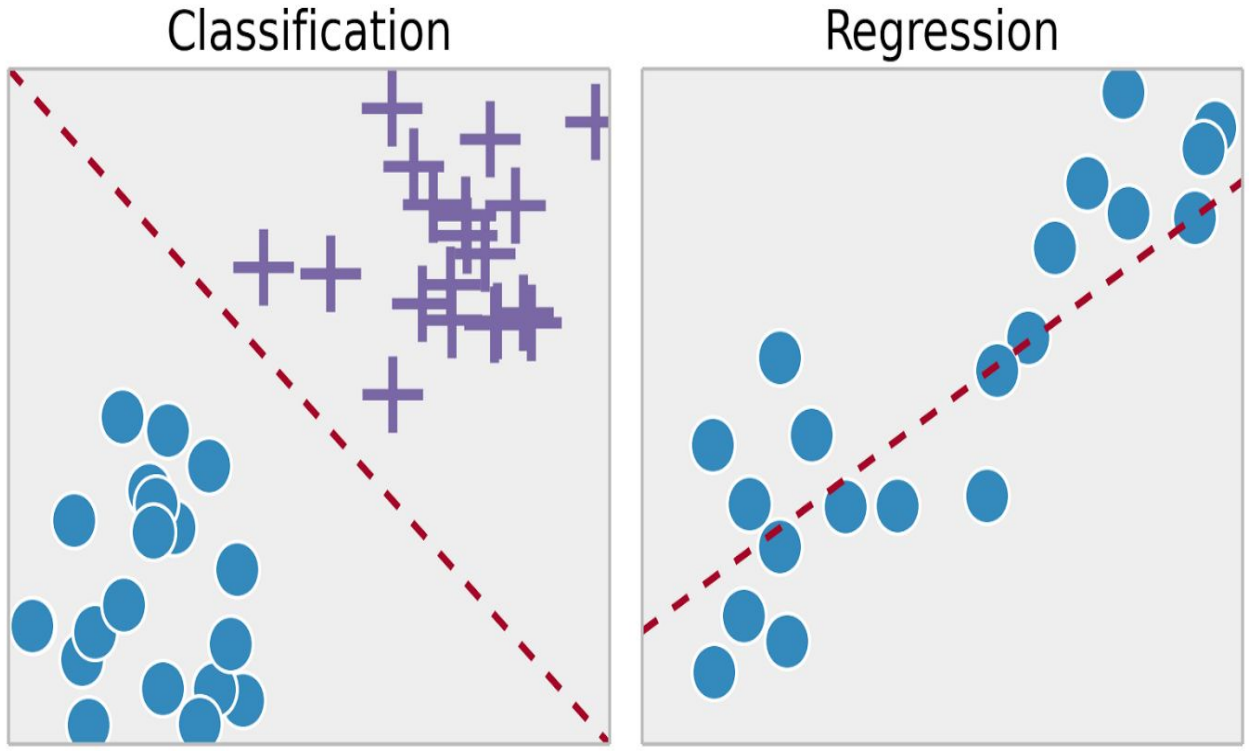
- In supervised learning the machine experiences the examples along with the labels or targets for each example.
- The labels in the data help the algorithm to correlate the features.



Supervised/Directed Learning

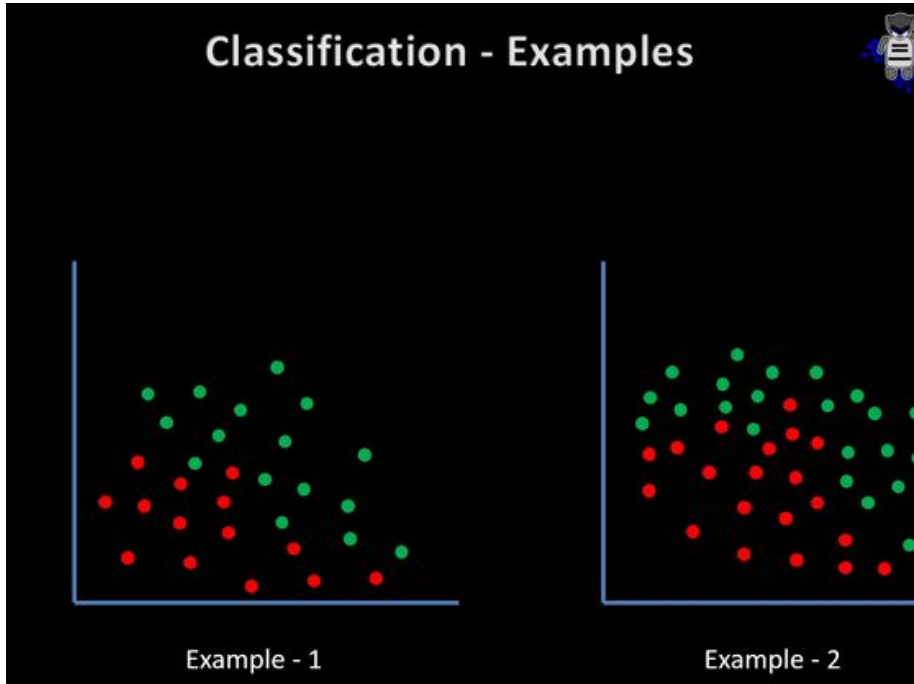
- Two of the most common supervised machine learning tasks are:

- **Classification:** machine learns to predict discrete values
  - predicting whether a stock's price will rise or fall
  - deciding if a news article belongs to the politics or leisure section
- **Regression:** machine predicts the value of a continuous response variable
  - predicting the sales for a new product
  - the salary for a job based on its description



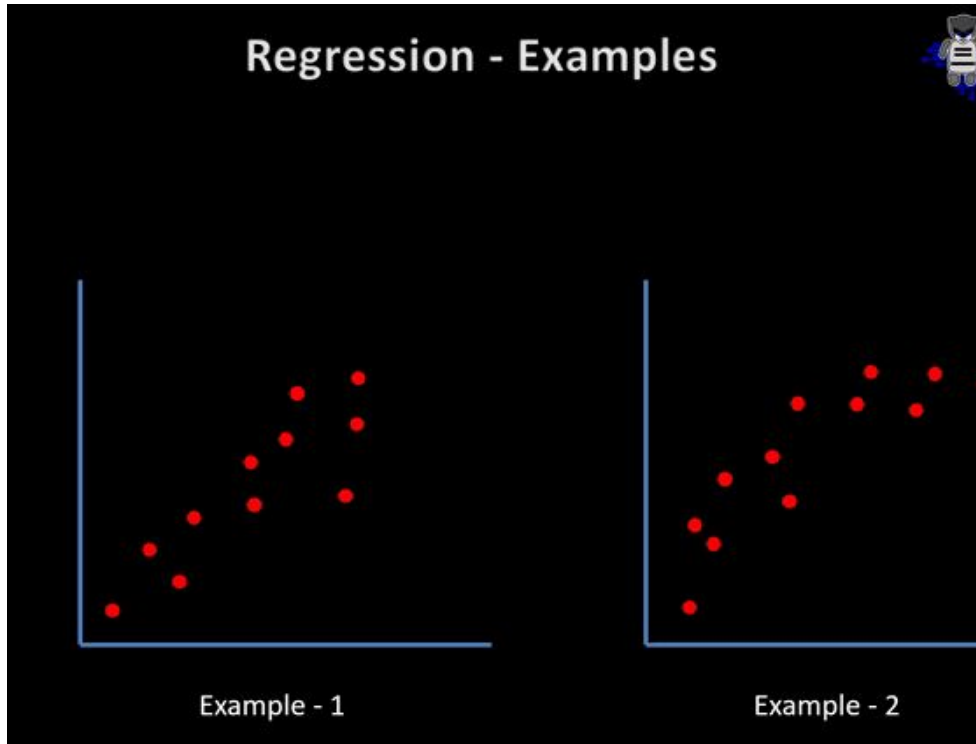
Supervised/Directed Learning

- Classification:
- A classification problem is when the output variable is a category/class and the goal is to classify the input into the known categories/classes.
- It basically aims to predict which class the input corresponds to.
- Some examples include systems where we seek a yes-or-no prediction, such as:
  - “Is this tumour cancerous?”,
  - “Does this product meet our quality standards?”



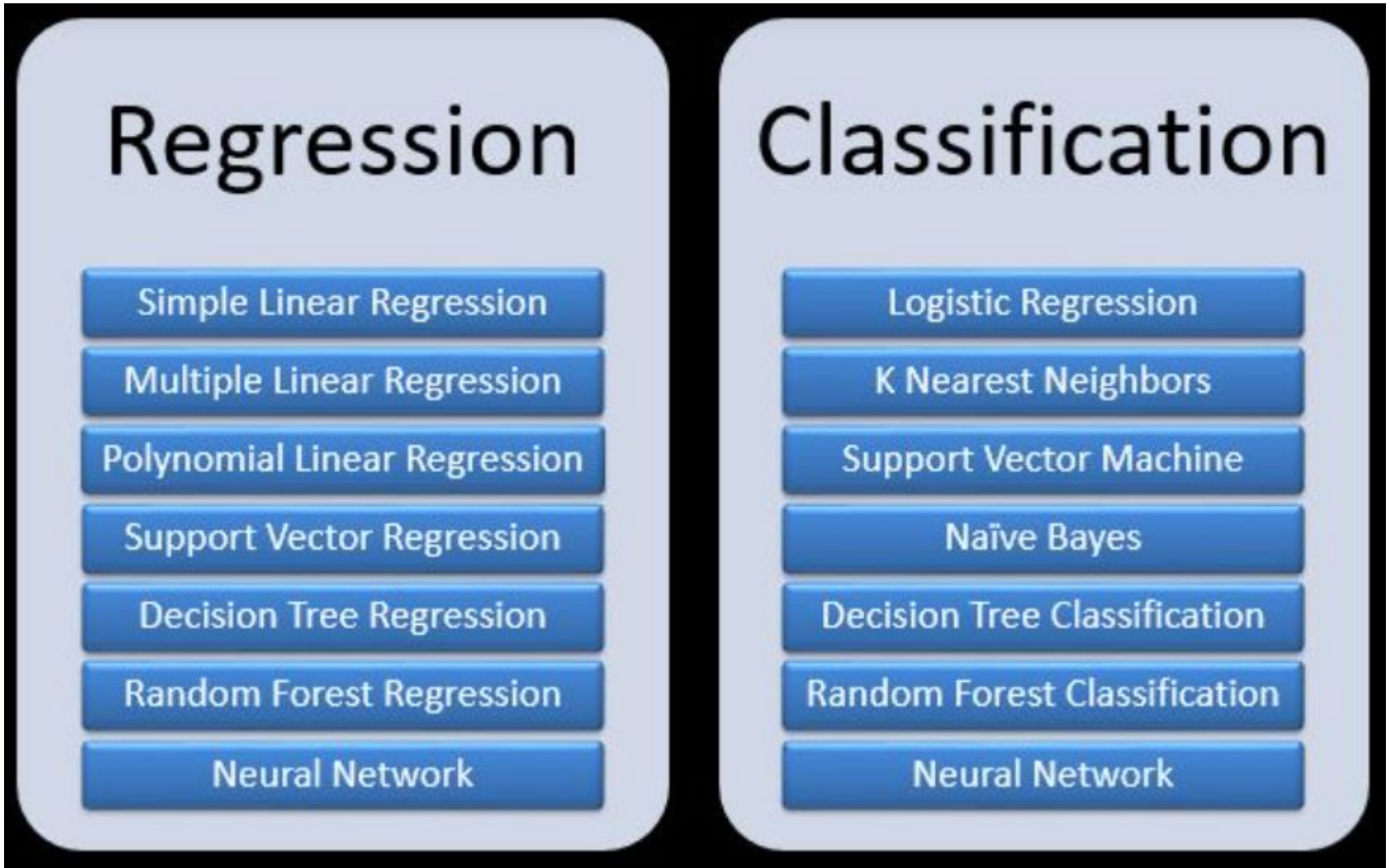
Supervised/Directed Learning

- Regression:
- A regression problem is when the output variable is a real and continuous value. Here, the goal is to predict the output value given an input x.
- Some examples include systems where the value being predicted falls somewhere on a continuous spectrum.
  - predicting the stock price of a company
  - predicting the temperature tomorrow based on historical data.



Supervised/Directed Learning

- **Commonly used algorithms for classification and regression**



Unsupervised/Undirected Learning

- When we have unclassified and unlabeled data, the system attempts to uncover patterns from the data .
- There is no label or target given for the examples.
- One common task is to group similar examples together called clustering.
- clustering and association learning belong to this category

Unsupervised/Undirected Learning

- Clustering:
- Here the goal is to find similarities in the dataset and group similar data points together.
- Cluster is the collection of data objects which are similar to one another within the same group (class or category) and are different from the objects in the other clusters.
- This method can also be used to detect anomalies that do not fit to any group.

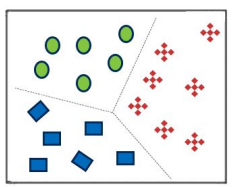
- In marketing, customers are segmented according to similarities to carry out targeted marketing.
- Given a collection of text, we need to organize them, according to the content similarities to create a topic hierarchy.
- Detecting distinct kinds of pattern in image data (Image processing).
- It’s effective in biology research for identifying the underlying patterns.

Unsupervised/Undirected Learning

- Association learning:
- In association learning, any relation between observations is required, not merely association capable of predicting a specific class value.
- Here the aim is to discover rules that describe large portions of our data, such as people that buy X also tend to buy Y.
- A classic example of association rules would be market basket analysis, like, a person who buys biryani and burgers usually buys a soft drink too.

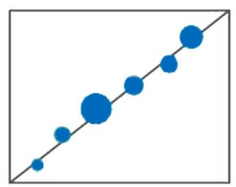
SUPERVISED LEARNING

EXAMPLES



Classification

Is this data input red, blue or green?

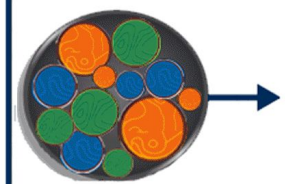


Regression

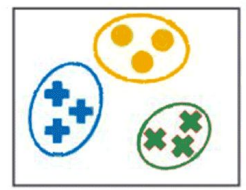
What is the impact of product price on number of sales?
What is the impact of years of experience on salary?

UNSUPERVISED LEARNING

Raw Data



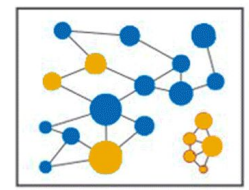
EXAMPLES



CLUSTERING

Identifies similarities in groups:

Are there patterns in the data that indicate which groups to target?



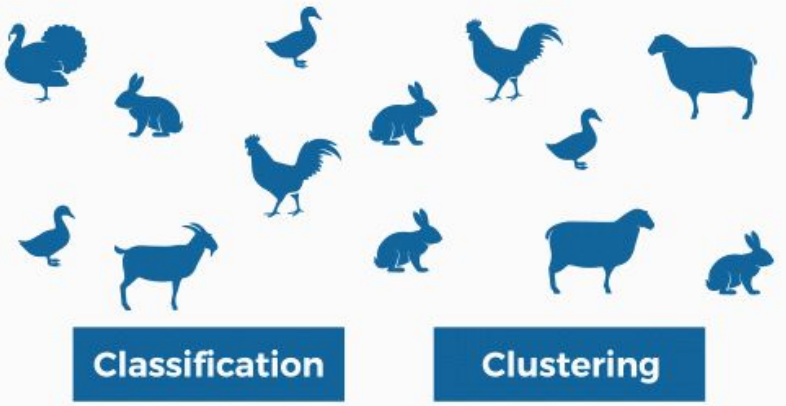
ANOMALY DETECTION

Identifies abnormalities in dataset:

Is the user behaving as it should? Is a hacker intruding the network?

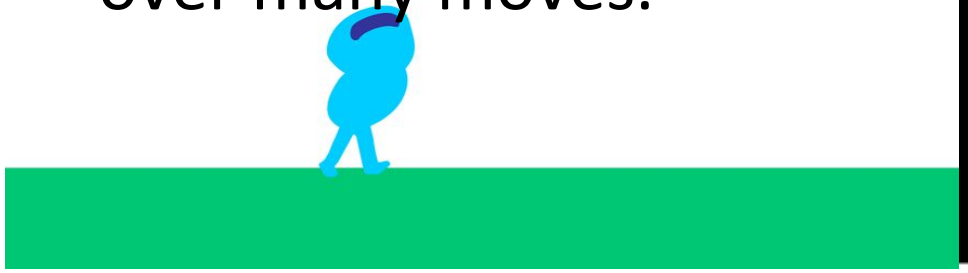
Clustering vs clasification

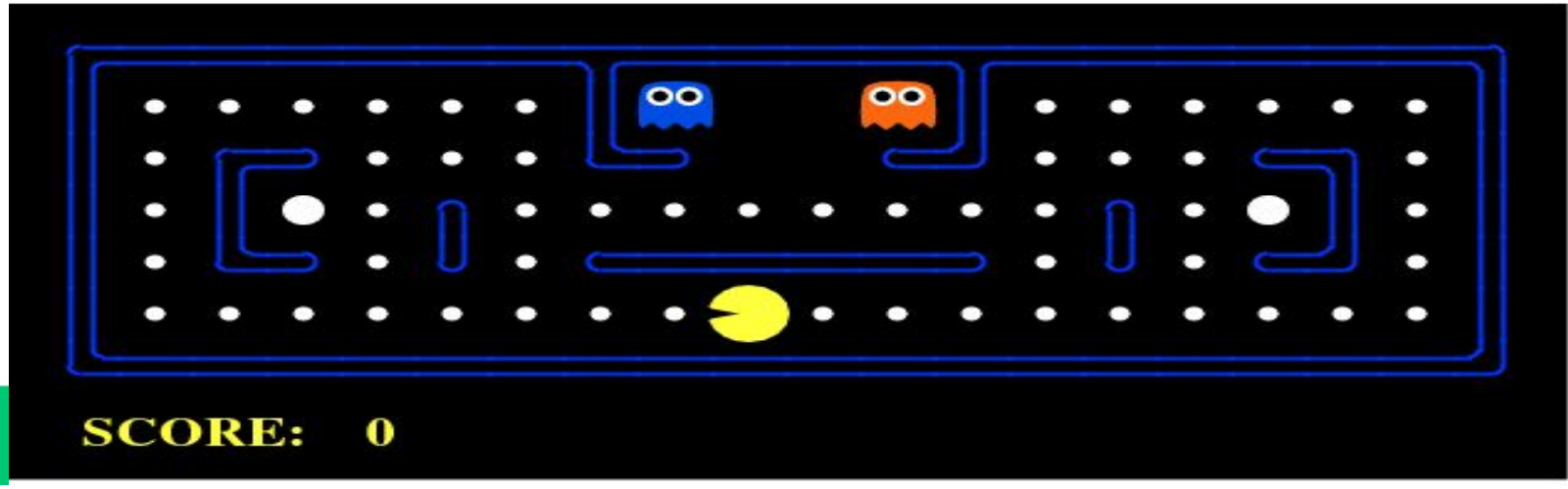
- In the case of Classification, there are predefined labels assigned to each input instance according to their properties
  - whereas in clustering those labels are missing.
- The process of classifying the input instances based on their corresponding class labels is known as classification
  - whereas grouping the instances based on their similarity without the help of class labels is known as clustering.



Reinforcement Learning

- Reinforcement learning refers to goal-oriented algorithms, which learn how to attain a complex objective (goal) or maximize along a particular dimension over many steps.
- This method allows machines to automatically determine the ideal behavior within a specific context in order to maximize its performance.
- Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.
- For example, maximize the points won in a game over many moves.





Semi-Supervised Learning

- Problems where you have a large amount of input data (X) and only some of the data is labeled (Y) are called semi-supervised learning problems.
- Many real world machine learning problems fall into this area.
- This is because it can be expensive or time-consuming to label data as it may require access to domain experts, whereas, unlabeled data is cheap and easy to collect and store.
- You can also use supervised learning techniques to make best guess predictions for the unlabeled data, feed that data back into the supervised learning algorithm as training data and use the model to make predictions on new unseen data.