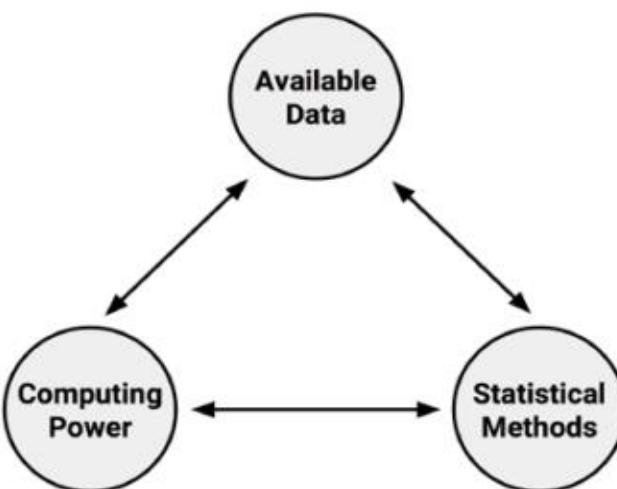


# Machine Learning

Module 2

# Introduction

- The field of study in the development of computer algorithms to transform data into intelligent action is known as machine learning.
- This field is in an environment evolved simultaneously with data, statistical methods, and computing power.
- Growth in data demanded additional computing power, which in turn spurred the development of statistical methods to analyze large datasets. This created a cycle of advancement, allowing even larger and more interesting data to be collected.



# Introduction

- A closely related sibling of machine learning, **data mining**, is concerned with the generation of novel insights from large databases.
- Virtually all data mining involves the use of machine learning, but not all machine learning involves data mining.
- For example, we use machine learning to data mine automobile traffic data for patterns related to accident rates; while if the computer is learning how to drive the car itself, this is purely machine learning without data mining.

# ML Applications

- Identification of unwanted spam messages in e-mail
- Segmentation of customer behavior for targeted advertising
- Forecasts of weather behavior and long-term climate changes
- Reduction of fraudulent credit card transactions
- Actuarial estimates of financial damage of storms and natural disasters
- Prediction of popular election outcomes
- Development of algorithms for auto-piloting drones and self-driving cars
- Optimization of energy use in homes and office buildings
- Projection of areas where criminal activity is most likely
- Discovery of genetic sequences linked to diseases

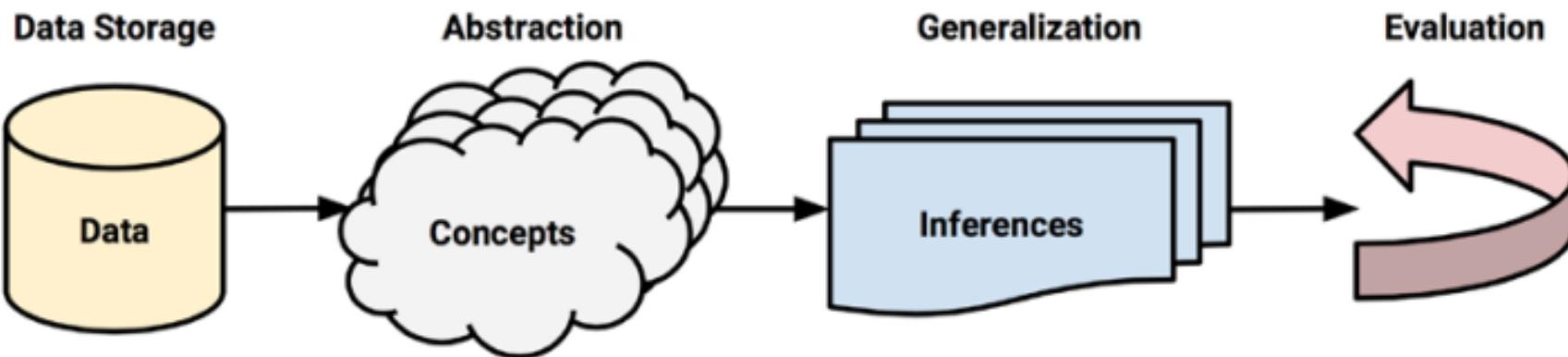
# ML limitations

- It has very little flexibility to extrapolate outside of the strict parameters it learned and knows no common sense.
- These are also limited in their ability to make simple common sense inferences about logical next steps.

# How machines Learn

- According to computer scientist Tom M. Mitchell, a machine learns whenever it is able to utilize its experience such that its performance improves on similar experiences in the future.
- The basic machine learning process can be divided into four interrelated components:
  - **Data storage** utilizes observation, memory, and recall to provide a factual basis for further reasoning.
  - **Abstraction** involves the translation of stored data into broader representations and concepts.
  - **Generalization** uses abstracted data to create knowledge and inferences that drive action in new contexts.
  - **Evaluation** provides a feedback mechanism to measure the utility of learned knowledge and inform potential improvements.

# Activities....



- **(i) Data Storage**
  - All learning must begin with data. Computers use hard disk drives, flash memory, and RAM in combination with a central processing unit (CPU)

# Activities....

- (ii) Abstraction

- During a machine's process of knowledge representation, the computer summarizes stored raw data using a **model**, an explicit description of the patterns within the data.

There are many different types of models. Some of them include:

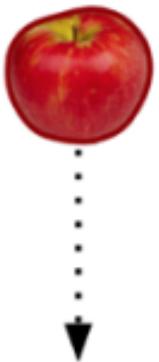
- Mathematical equations
- Relational diagrams such as trees and graphs
- Logical if/else rules
- Groupings of data known as cluster

The choice of model is depends on the learning task and data on hand.

# Activities...Abstraction

- Model creation....

**Observations** → **Data** → **Model**



Distance	Time
4.9m	1s
19.6m	2s
44.1m	3s
78.5m	4s

$$g = 9.8 \text{ m/s}^2$$

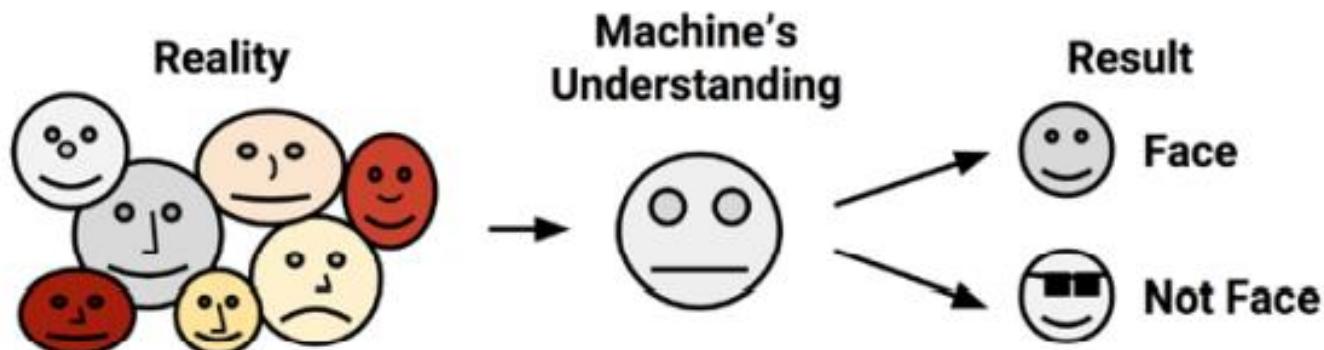
The process of fitting a model to a dataset is known as **training**. When the model has been trained, the data is transformed into an abstract form that summarizes the original information.

# Activities...

- (iii) **Generalization**
  - The term generalization describes the process of turning abstracted knowledge into a form that can be utilized for future action. It has been imagined as a search through the entire set of models (theories or inferences) that could be abstracted during training.
  - In generalization, the learner is tasked with limiting the patterns it discovers to only those that will be most relevant to its future tasks. In its end, the algorithm will employ **heuristics**, which are educated guesses about where to find the most useful inferences.

## Activities...Generalization

- The algorithm is said to have a **bias** if the conclusions are systematically erroneous, or wrong in a predictable manner.
- For example, suppose that a machine learning algorithm learned to identify faces by finding two dark circles representing eyes, positioned above a straight line indicating a mouth. The algorithm might then have trouble with, or be biased against faces that do not conform to its model like faces with glasses, turned at an angle, looking sideways, or with various skin tones.

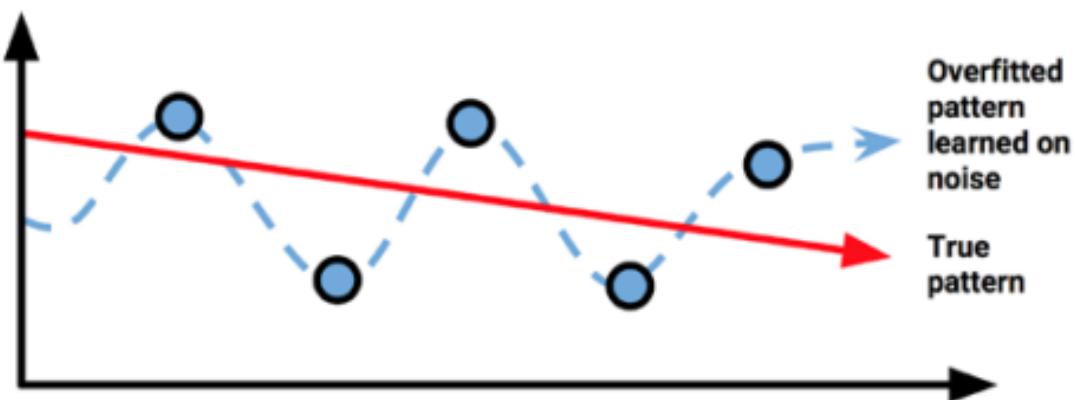


# Activities...Evaluation

- **Evaluation**
  - Bias occurs frequently with the abstraction and generalization processes inherent in any learning task. There is no single learning algorithm to rule them all. Therefore, the final step in the generalization process is to evaluate or measure the learner's success in spite of its biases and use this information to inform additional training if needed.
  - Generally, evaluation occurs after a model has been trained on an initial training dataset. Then, the model is evaluated on a new test dataset in order to judge how well it generalizes to new, unseen data.

## Activities...Evaluation

- Trying to model noise is the basis of a problem called **overfitting**. It occurs when a classifier fits the training data too tightly and doesn't generalize well to independent test data.
- A model that seems to perform well during training, but does poorly during evaluation, is said to be overfitted to the training dataset, as it does not generalize well to the test dataset.



# Machine learning in Practice

- To apply the learning process to real-world tasks, we'll use a five-step process.
- **1. Data collection:** The data collection step involves gathering the learning material an algorithm will use to generate actionable knowledge. In most cases, the data will need to be combined into a single source like a text file, spreadsheet, or database.
- **2. Data exploration and preparation:** The quality of any machine learning project is based largely on the quality of its input data. Thus, it is important to learn more about the data and its tones during a practice called data exploration. Additional work is required to prepare the data for the learning process. This involves fixing or cleaning noisy data, eliminating unnecessary data, and recoding the data to conform to the learner's expected inputs.

# Machine learning in Practice

- **3. Model training:** By the time the data has been prepared for analysis, you are likely to have a sense of what you are capable of learning from the data. The specific machine learning task chosen will inform the selection of an appropriate algorithm, and the algorithm will represent the data in the form of a model.
- **4. Model evaluation:** Because each machine learning model results in a biased solution to the learning problem, it is important to evaluate how well the algorithm learns from its experience. Depending on the type of model used, you might be able to evaluate the accuracy of the model using a test dataset or you may need to develop measures of performance specific to the intended application.

# Machine learning in Practice

- **5. Model improvement:** If better performance is needed, it becomes necessary to utilize more advanced strategies to augment the performance of the model. Sometimes, it may be necessary to switch to a different type of model altogether. You may need to supplement your data with additional data or perform additional preparatory work as in step two of this process.

# Types of Data

features

The diagram illustrates the concept of features and examples. A horizontal bracket spans the width of the table, with the word "features" centered above it. Another bracket on the right side of the table, pointing downwards, is labeled "examples".

year	model	price	mileage	color	transmission
2011	SEL	21992	7413	Yellow	AUTO
2011	SEL	20995	10926	Gray	AUTO
2011	SEL	19995	7351	Silver	AUTO
2011	SEL	17809	11613	Gray	AUTO
2012	SE	17500	8367	White	MANUAL
2010	SEL	17495	25125	Silver	AUTO
2011	SEL	17000	27393	Blue	AUTO
2010	SEL	16995	21026	Silver	AUTO
2011	SES	16995	32655	Silver	AUTO

- Features come in various forms. If a feature **uses numbers, it is numeric**. If a feature is an attribute that consists of a **set of categories**, the feature is called **categorical or nominal**. A **special case of categorical variables** is called **ordinal**, which designates a nominal variable with categories **falling in an ordered list**. Some examples of ordinal variables include clothing sizes such as small, medium, and large.

# Types of ML Algorithms

- Machine learning algorithms are divided into categories according to their purpose. They are classified into two categories:
  - **Predictive**
  - **Descriptive**
- A **predictive model** is used for tasks that involve the prediction of one value using other values in the dataset. The learning algorithm attempts to discover and model the relationship between the target feature (the feature being predicted) and the other features. The process of training a predictive model is known as **supervised learning**. **Predictive** mining tasks **perform inference** on the current data in order **to make predictions**.

# Types of ML Algorithms...

- The often used supervised machine learning task is **classification**. In classification, the **target feature** to be predicted is a categorical feature known as the **class**, and is divided into categories called **levels**. A class can have two or more levels, and the levels may or may not be ordinal. Some classification problems are: Prediction on whether
  - An e-mail message is spam
  - A person has cancer
  - A football team will win or lose
  - An applicant will default on a loan

Supervised learners can **also be used to predict numeric data** such as income, laboratory values, test scores, or counts of items. To predict such **numeric values**, a common form of numeric prediction fits linear **regression models** to the input data.

# Types of ML Algorithms...

A **descriptive model** is used for tasks that would benefit from the insight gained from summarizing data in new and interesting ways. In a descriptive model, no single feature is more important than any other. In fact, **because there is no target to learn, the process of training a descriptive model is called unsupervised learning.**

- The related problems are:
  - pattern discovery
  - Clustering
  - Segmentation Analysis

# Types of ML Algorithms

- Commonly, ML algorithms can be divided into categories according to their purpose. The main categories are the following:
  - Supervised learning
  - Unsupervised Learning
  - Semi-supervised Learning
  - Reinforcement Learning
- (i) **Supervised learning** algorithms try to *model relationships and dependencies between the target prediction output and the input features* such that we can predict the output values for new data based on those relationships which it learned from the previous data sets. The main types of supervised learning problems include regression and classification problems

# Types of ML Algorithms

- List of Common Supervised Learning Algorithms
  - Nearest Neighbor
  - Naive Bayes
  - Decision Trees
  - Linear Regression
  - Support Vector Machines (SVM)
  - Neural Networks

# Types of ML Algorithms...

- (ii) **Unsupervised Learning**
- The computer is trained with unlabeled data.
- In this actually the computer need to learn patterns in data, and these algorithms are particularly useful in cases where the human expert doesn't know what to look for in the data.
- These are mainly used in ***pattern detection and descriptive modeling***. There are no output categories and these algorithms try to use techniques on the input data to *mine for rules, detect patterns, and summarize and group the data points* which help in deriving meaningful insights and describe the data better to the users.
- The main types of unsupervised learning algorithms include ***Clustering algorithms and Association rule learning algorithms***.

## List of Common Unsupervised Learning Algorithms

- k-means clustering, Association Rules

# Types of ML Algorithms...

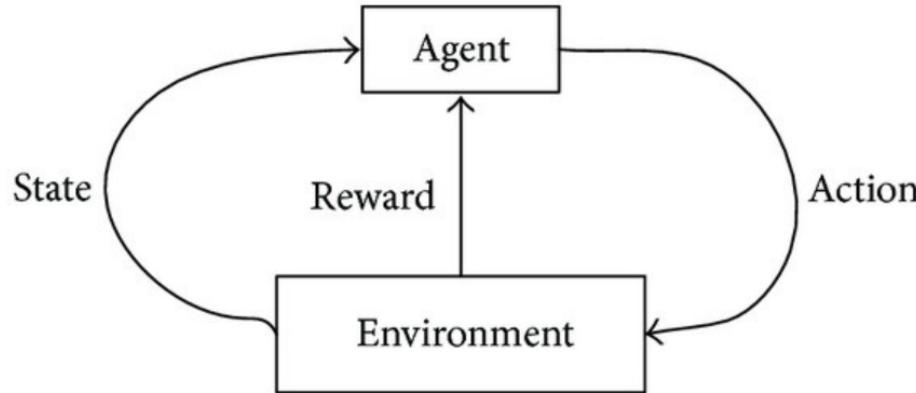
- (iii) **Semi-Supervised Learning**
- In the previous two types, either there are no labels for all the observation in the dataset or labels are present for all the observations.
- Semi-supervised learning falls in between these two. In the absence of labels in the majority of the observations but present in few, semi-supervised algorithms are the best candidates for the model building. These methods exploit the idea that even though the group memberships of the unlabeled data are unknown, this data carries important information about the group parameters.

# Types of ML Algorithms...

- (iv) **Reinforcement Learning**
- This method aims at using observations gathered from the interaction with the environment to take actions that would maximize the reward or minimize the risk. Reinforcement learning algorithm (called the agent) continuously learns from the environment in an iterative fashion. In the process, the agent learns from its experiences of the environment until it explores the full range of possible states.
- Reinforcement Learning is a type of *Machine Learning*, and also a branch of *Artificial Intelligence*. It allows machines and software agents 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 its behavior; this is known as the reinforcement signal.

# Types of ML Algorithms...

- **Reinforcement Learning**



In the problem, an agent is supposed decide the best action to select based on his current state. When this step is repeated, the problem is known as a *Markov Decision Process*.

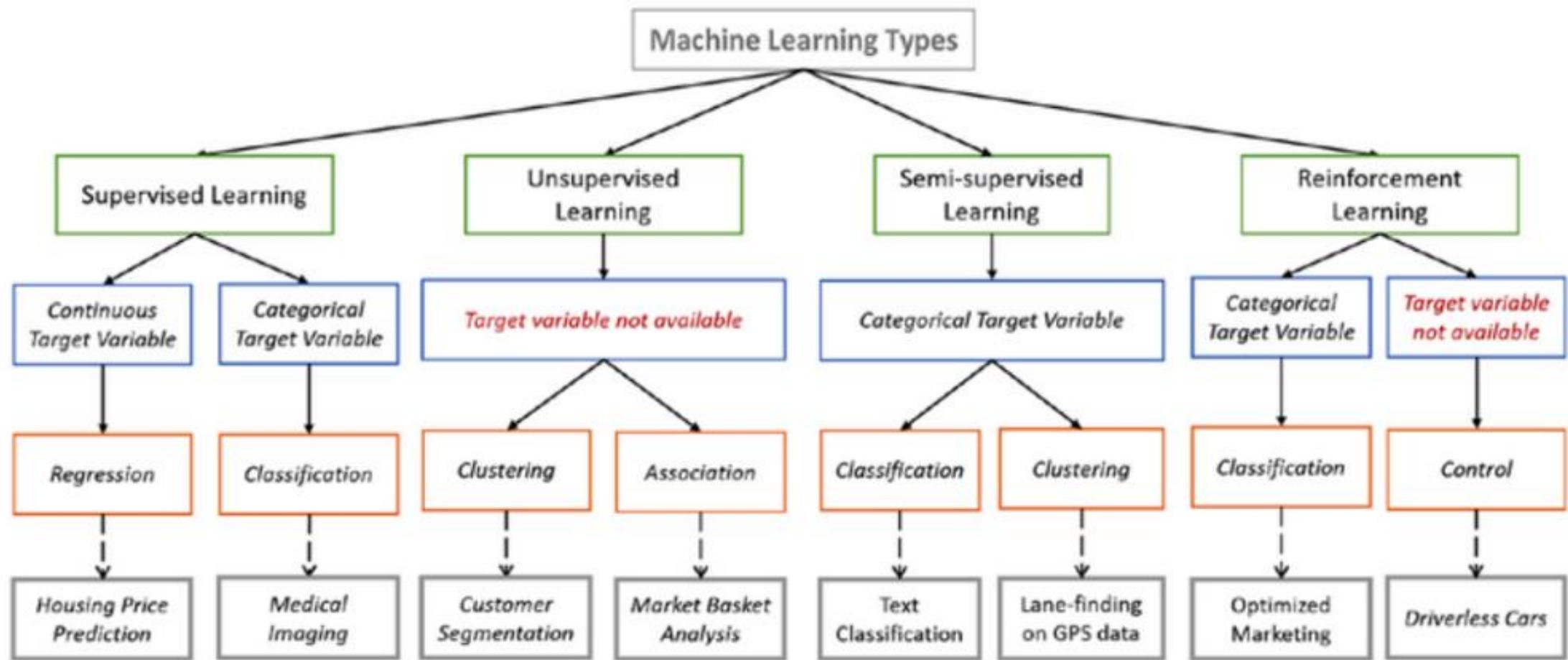
In order to produce intelligent programs (also called agents), reinforcement learning goes through the following steps:

1. Input state is observed by the agent.
2. Decision making function is used to make the agent perform an action.
3. After the action is performed, the agent receives reward or reinforcement from the environment.
4. The state-action pair information about the reward is stored.

# Types of ML Algorithms...

- **Reinforcement Learning**
- This method aims at using observations gathered from the interaction with the environment to take actions that would maximize the reward or minimize the risk. Reinforcement learning algorithm (called the agent) continuously learns from the environment in an iterative fashion. In the process, the agent learns from its experiences of the environment until it explores the full range of possible states.

# Types of ML Algorithms



- Features

# Types of ML Algorithms...

- Lastly, a class of machine learning algorithms known as **meta-learners** is not tied to a specific learning task, but is rather focused on learning how to learn more effectively.
- A meta-learning algorithm uses the result of some learnings to inform additional learning. This can be beneficial for very challenging problems or when a predictive algorithm's performance needs to be as accurate as possible.

# Types of ML Algorithms...

Model	Learning task
<b>Supervised Learning Algorithms</b>	
Nearest Neighbor	Classification
Naive Bayes	Classification
Decision Trees	Classification
Classification Rule Learners	Classification
Linear Regression	Numeric prediction
Regression Trees	Numeric prediction
Model Trees	Numeric prediction
Neural Networks	Dual use
Support Vector Machines	Dual use
<b>Unsupervised Learning Algorithms</b>	
Association Rules	Pattern detection
k-means clustering	Clustering
<b>Meta-Learning Algorithms</b>	
Bagging	Dual use
Boosting	Dual use
Random Forests	Dual use