PROFESSEUR DOUZI SAMIRA

s.douzi@um5r.ac.ma

What Is Machine Learning?

Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be *predictive* to make predictions in the future, or *descriptive* to gain knowledge from data, or both.

"How do we create computer programs that improve with experience?"

"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. "

Tom Mitchell. Machine Learning 1997.

Examples

1. Handwriting recognition learning problem

- 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

2. A robot driving learning problem

- Task T: Driving on highways using vision sensors
- Performance measure P: Average distance traveled before an error
- training experience: A sequence of images and steering commands recorded while observing a human driver.

3. A chess learning problem

- Task T: Playing chess
- Performance measure P: Percent of games won against opponents
- Training experience E: Playing practice games against itself

Machine Learning has multiple algorithms, techniques, and methodologies that can be used to build models to solve real-world problems using data. Typically, the same Machine Learning methods can be classified in multiple ways under multiple umbrellas. Following are some of the major broad areas of Machine Learning methods:

1. Methods based on the amount of human supervision in the learning process :

a) Supervised learning .
b) Unsupervised learning.
c) Semi-supervised learning
d) Reinforcement learning

2. Methods based on the ability to learn from incremental data samples :

a) Batch learning
b) Online learning

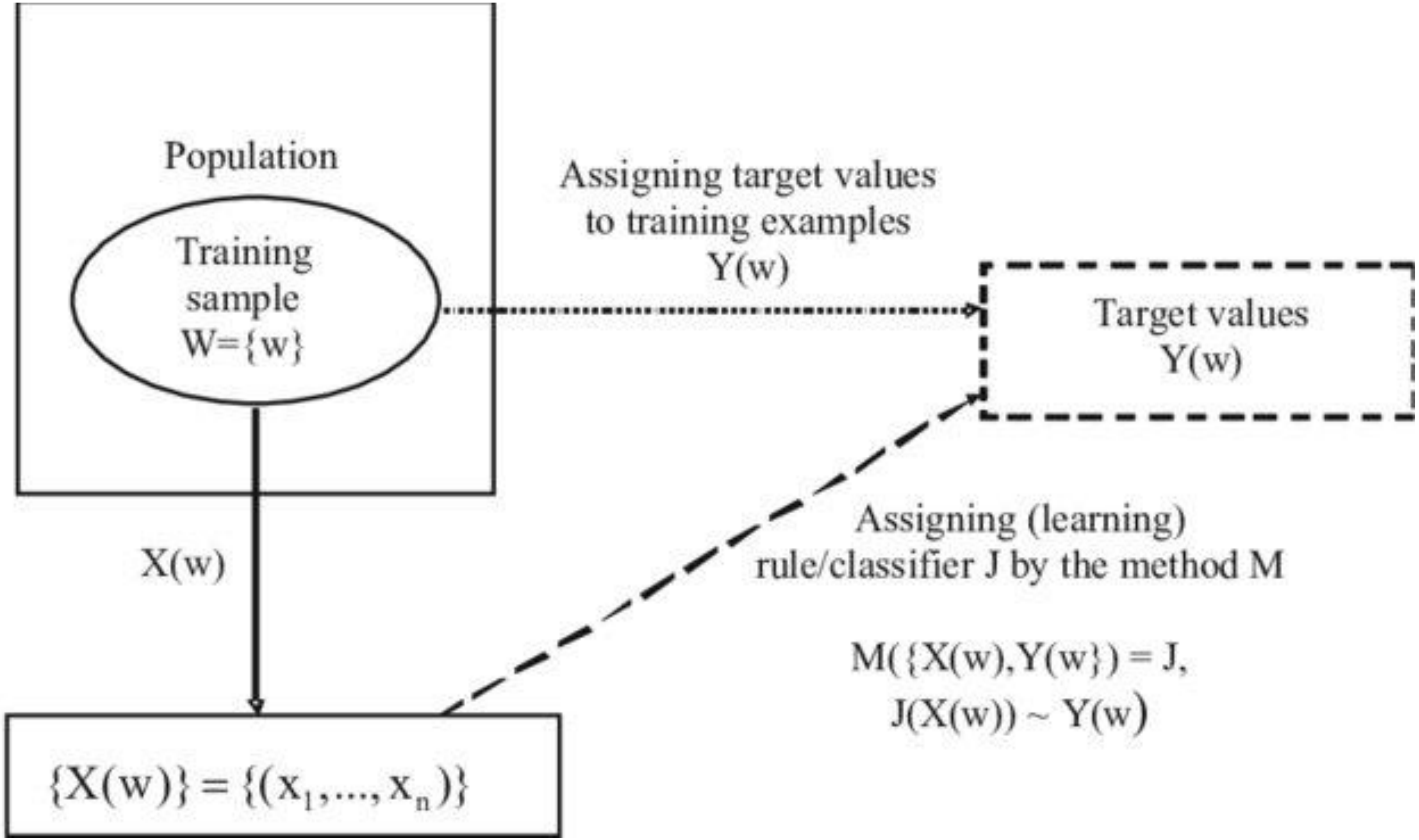
3. Methods based on their approach to generalization from data samples :

a) Instance based learning
b) Model based learning

Supervised Learning

Supervised learning methods or algorithms include learning algorithms that take in data samples (known as training data) and associated outputs (known as labels or responses) with each data sample during the model training process. The main objective is to learn a mapping or association between input data samples x and their corresponding outputs y based on multiple training data instances. This learned knowledge can then be used in the future to predict an output y' for any new input data sample x' which was previously unknown or unseen during the model training process. These methods are termed as supervised because the model learns on data samples where the desired output responses/labels are already known beforehand in the training phase.

Supervised learning basically tries to model the relationship between the inputs and their corresponding outputs from the training data so that we would be able to predict output responses for new data inputs based on the knowledge it gained earlier . This is precisely why supervised learning methods are extensively used in predictive analytics.

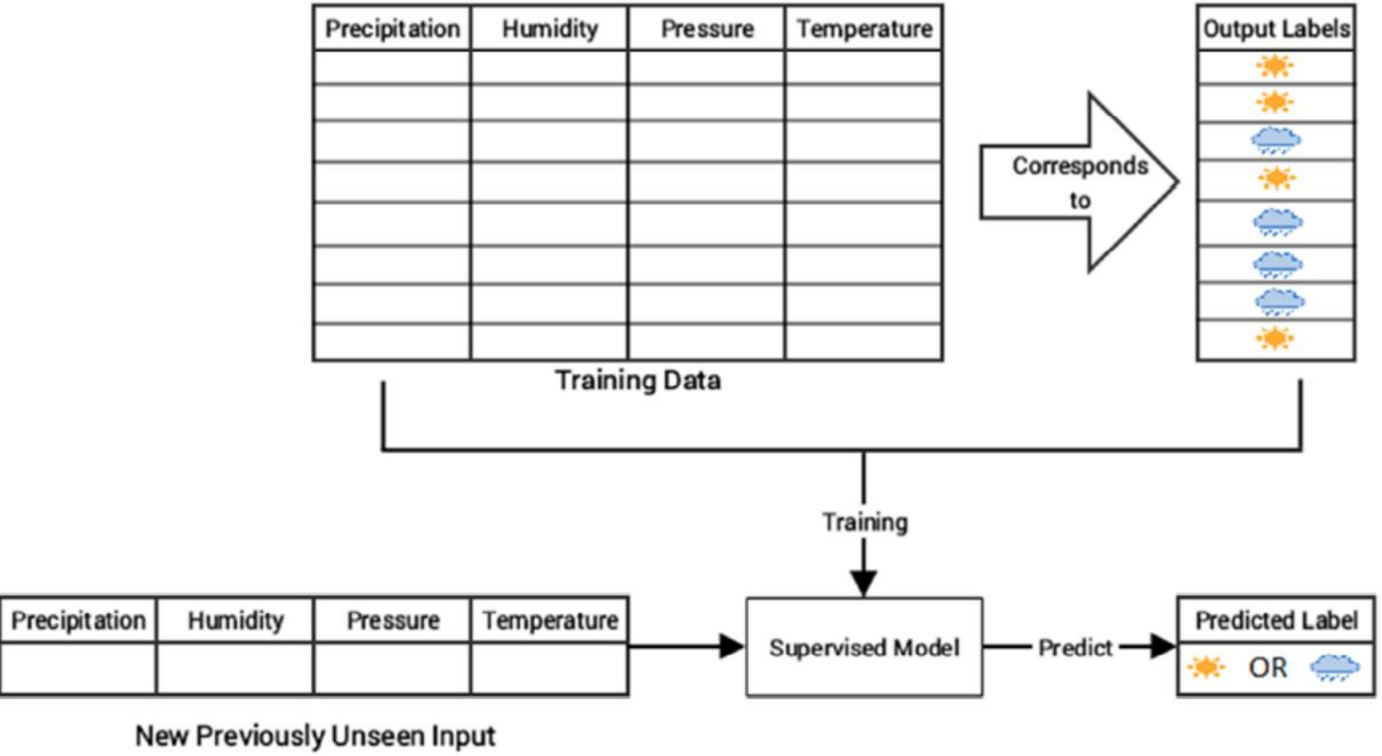

Population
Training
sample
W={w}
Assigning target values
to training examples
Y(w)
Target values
Y(w)
X(w)
Assigning (learning)
rule/classifier J by the method M
M({X(w),Y(w)}) = J,
J(X(w)) ~ Y(w)
{X(w)} = {(x1,...,xn)}

Supervised learning methods are of two major classes based on the type of ML tasks they aim to solve.

- Classification
- Regression

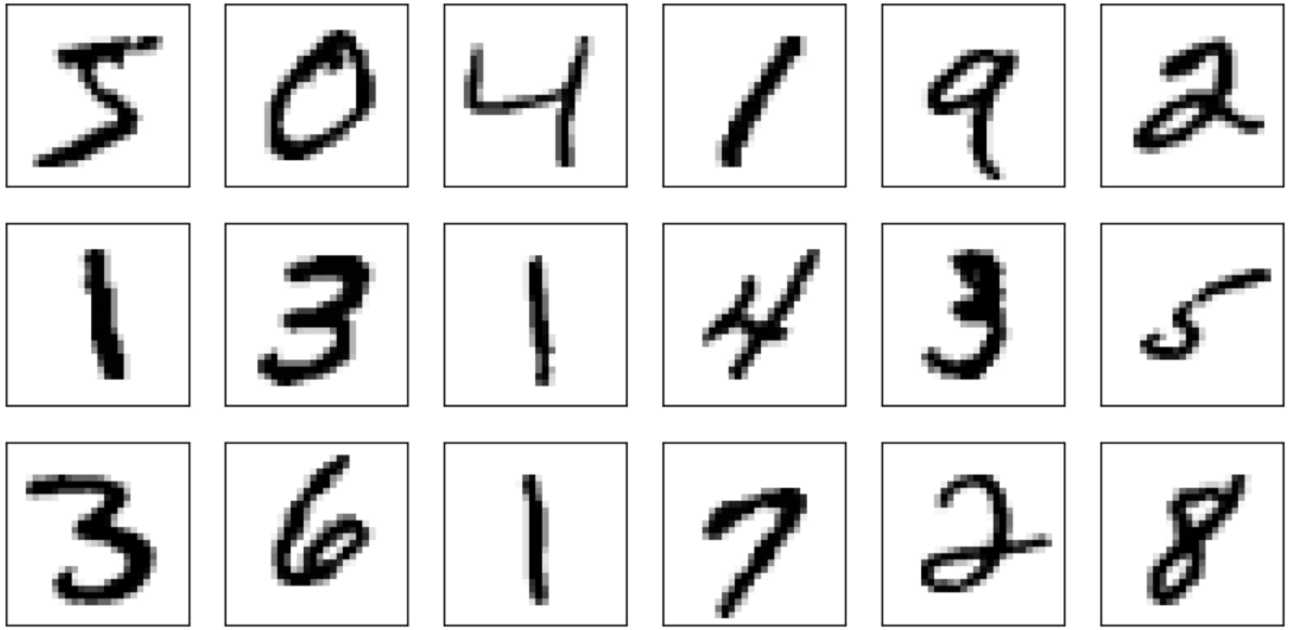
Classification

The classification-based tasks are a sub-field under supervised Machine Learning, where the key objective is to predict output labels or responses that are categorical in nature for input data based on what the model has learned in the training phase. Output labels are also known as classes or class labels are categorical in nature meaning they are unordered and discrete values. Thus, each output response belongs to a specific discrete class or category.



- Suppose we take a real-world example of predicting the weather. we are trying to predict if the weather is sunny or rainy based on multiple input data samples consisting of attributes or features like humidity, temperature, pressure, and precipitation.
- Since the prediction can be either sunny or rainy, there are a total of two distinct classes in total; hence this problem can also be termed as a binary classification problem.

A task where the total number of distinct classes is more than two becomes a multi-class classification problem where each prediction response can be any one of the probable classes from this set. A simple example would be trying to predict numeric digits from scanned handwritten images. In this case it becomes a 10-class classification problem because the output class label for any image can be any digit from 0 – 9.



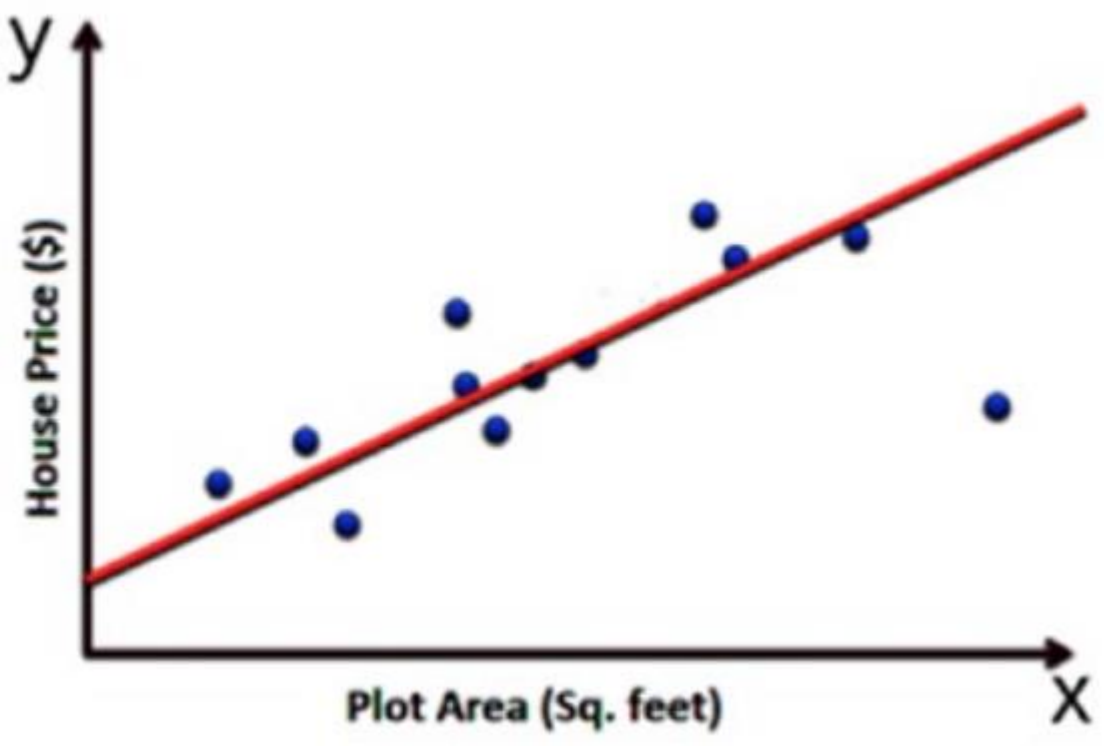
Popular classification algorithms include logistic regression, support vector machines, neural networks, ensembles like random forests and gradient boosting, K-nearest neighbors, decision trees, and many more.

Regression

Machine Learning tasks where the main objective is value estimation can be termed as regression tasks. Regression based methods are trained on input data samples having output responses that are continuous numeric values unlike classification, where we have discrete categories or classes. Regression models make use of input data attributes or features) and their corresponding continuous numeric output values to learn specific relationships and associations between the inputs and their corresponding outputs. With this knowledge, it can predict output responses for new, unseen data instances similar to classification but with continuous numeric outputs.

Let take a simple training dataset with one input feature(*area*) and output (*price*).The following graphs show two possible regression models based on different methods to predict house prices based on plot area

Linear Regression



Estimated (or predicted) y value

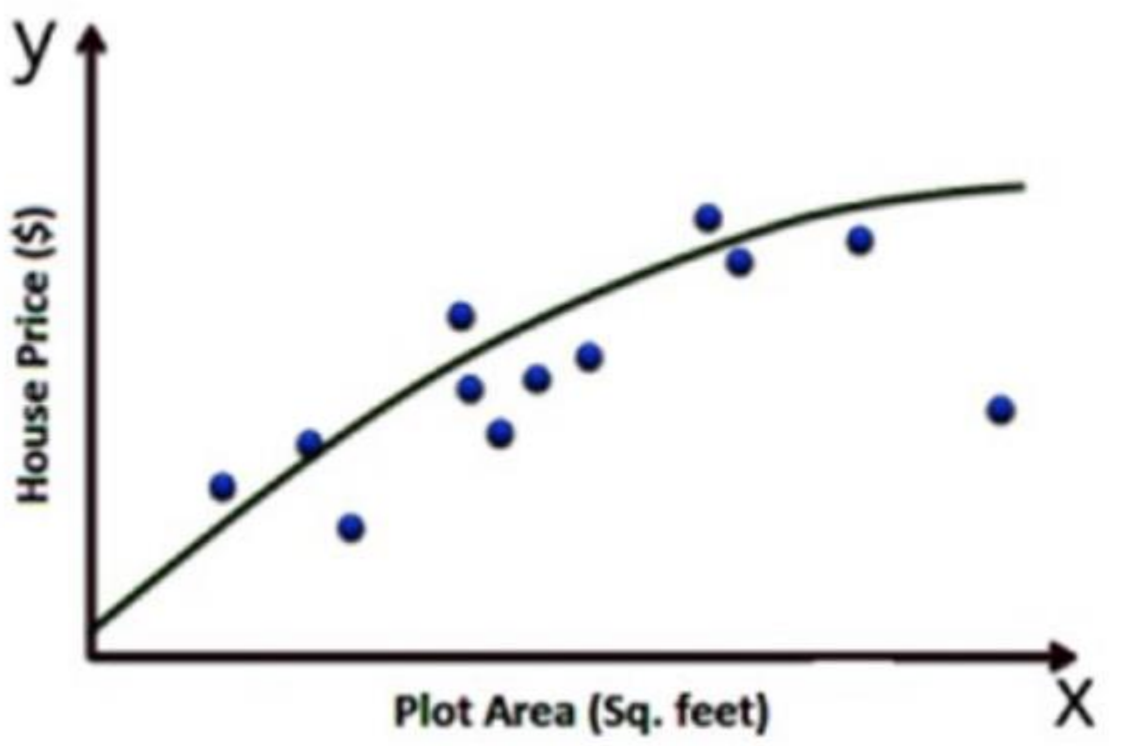
Estimate of the regression y-intercept

Estimate of the regression slope

Independent variable

$$\hat{Y}_i = b_0 + b_1 X_i$$

Multiple Regression (Polynomial)



$$y = b_0 + b_1 x_1 \boxed{+ b_2 x_1^2}$$

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

Multiple Linear Regression

Nonlinear Regression

Some popular nonlinear regression models:

1. Exponential model: $(y = ae^{bx})$
2. Power model: $(y = ax^{b})$
3. Saturation growth model: $\left(y = \dfrac{ax}{b+x}\right)$
4. Polynomial model: $(y = a_0 + a_1 x + \ldots + a_m x^{m})$

Popular Regression algorithms include Lasso regression, Ridge regression , and many more.

UNSUPERVISED LEARNING

Unsupervised methods are the models or algorithms that tries to learn inherent latent structures, patterns and relationships from given data without any help or supervision like providing annotations in the form of labeled outputs or outcomes. Unsupervised learning is more concerned with trying to extract meaningful insights or information from data.

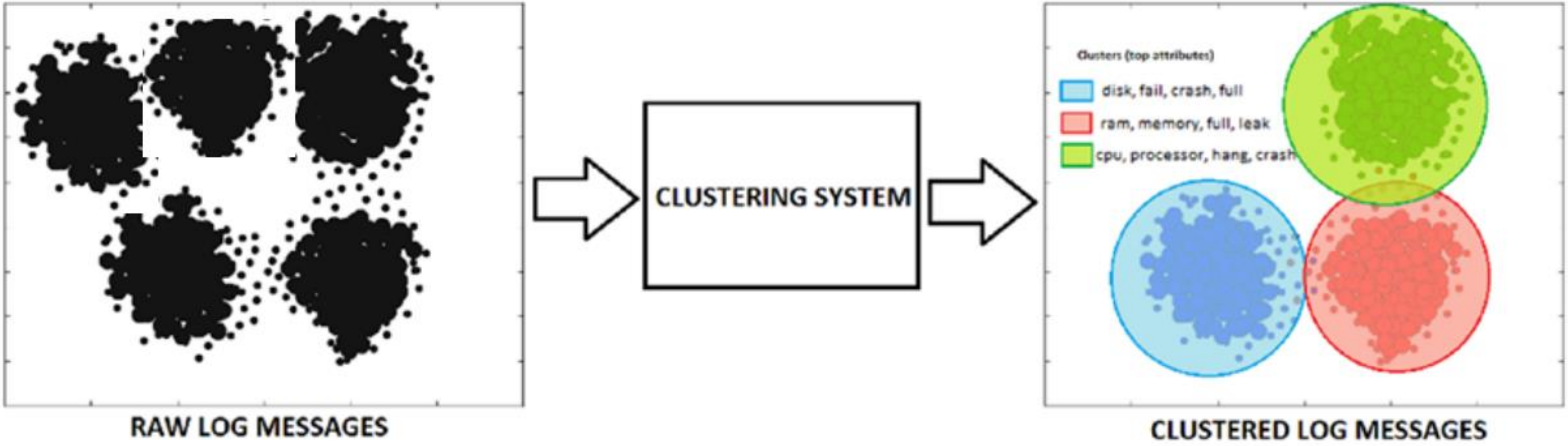
Unsupervised learning methods can be categorized under the following broad areas of ML tasks relevant to unsupervised learning.

1. Clustering

Clustering methods are Machine Learning methods that try to find patterns of similarity and relationships among data samples in our dataset and then cluster these samples into various groups, such that each group or cluster of data samples has some similarity, based on the inherent attributes or features. These methods are completely unsupervised because they try to cluster data by looking at the data features without any prior training, supervision, or knowledge about data attributes, associations, and relationships.

Example :
Consider a real-world problem of running multiple servers in a data center and trying to analyze logs for typical issues or errors. Our main task is to determine the various kinds of log messages that usually occur frequently each week. In simple words, we want to group log messages into various clusters based on some inherent characteristics. A simple approach would be to extract features from the log messages, which would be in textual format and apply clustering on the same and group similar log messages together based on similarity in content.

a clustering algorithm like K-means or hierarchical clustering would be employed to group or cluster messages based on similarity of their inherent features.



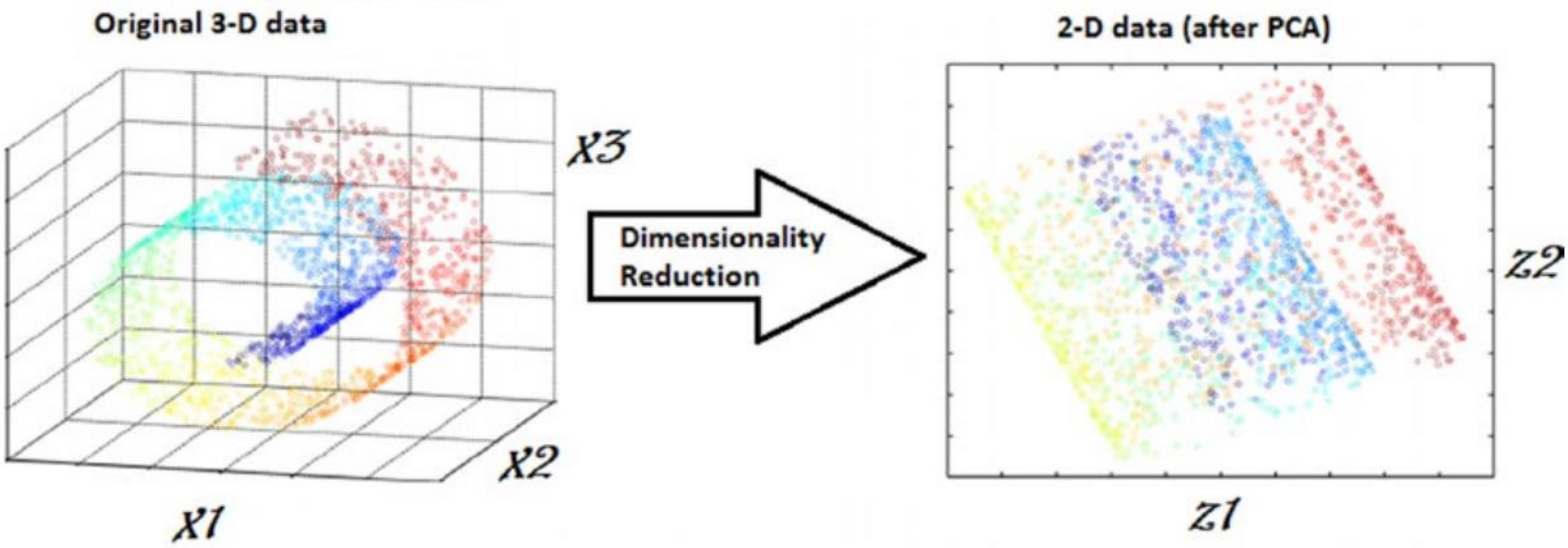
It is quite clear that our systems have three distinct clusters of log messages where the first cluster depicts disk issues, the second cluster is about memory issues, and the third cluster is about processor issues. Top feature words that helped in distinguishing the clusters and grouping similar data samples (logs) together are also depicted in the figure.

There are various types of clustering methods that can be classified under the following major approaches.

- Centroid based methods such as K-means and K-medoids
- Hierarchical clustering methods such as agglomerative and divisive (Ward's, affinity propagation)
- Distribution based clustering methods such as Gaussian mixture models • Density based methods such as dbscan and optics.

2. Dimensionality Reduction

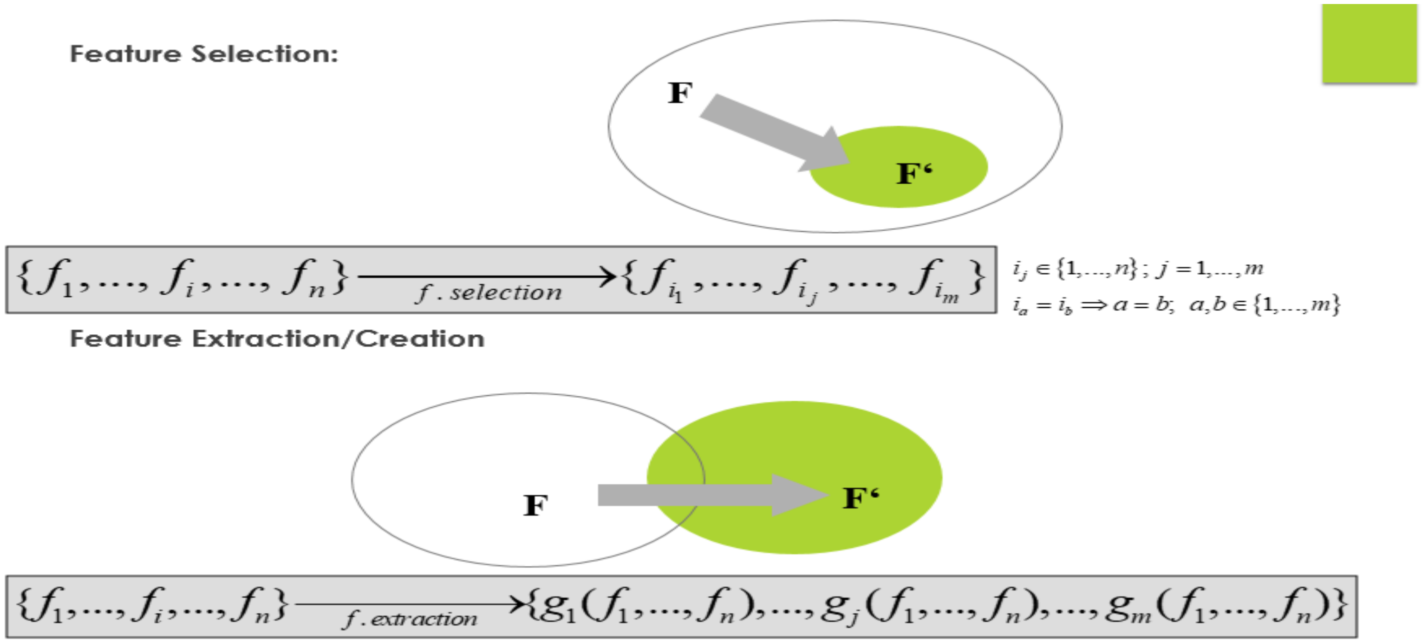
- sometimes our feature space gets bloated up with a humongous number of features. This poses multiple challenges including analyzing and visualizing data with thousands or millions of features, which makes the feature space extremely complex posing problems regarding training models, memory, and space constraints. In fact, this is referred to as the "curse of dimensionality". Unsupervised methods can also be used in these scenarios, where we reduce the number of features or attributes for each data sample. These methods reduce the number of feature variables by extracting or selecting a set of principal or representative features. There are multiple popular algorithms available for dimensionality reduction like Principal Component Analysis (PCA), nearest neighbors, and discriminant analysis.


Original 3-D data
X3
X2
X1
Dimensionality
Reduction
2-D data (after PCA)
Z2
Z1

Dimensionality reduction techniques can be classified in two major approaches as follows.

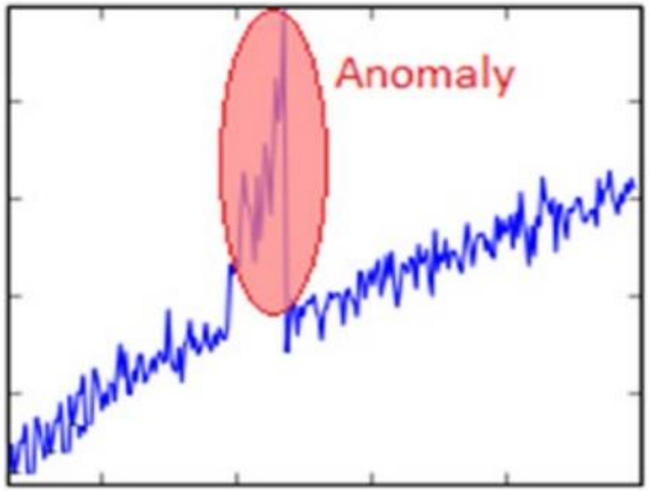
- Feature Selection methods: Specific features are selected for each data sample from the original list of features and other features are discarded. No new features are generated in this process.

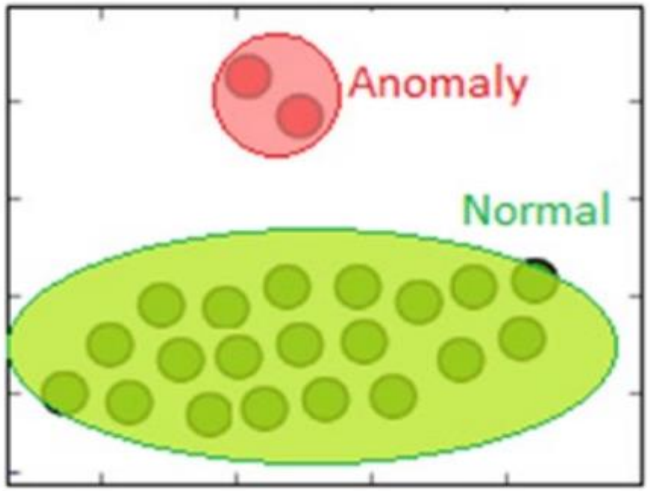
- Feature Extraction methods: We engineer or extract new features from the original list of features in the data. Thus, the reduced subset of features will contain newly generated features that were not part of the original feature set. PCA falls under this category.



3. Anomaly Detection

The process of anomaly detection is also termed as outlier detection, where we are interested in finding out occurrences of rare events or observations that typically do not occur normally based on historical data samples. Unsupervised learning methods can be used for anomaly detection such that we train the algorithm on the training dataset having normal, non-anomalous data samples. Once it learns the necessary data representations, patterns, and relations among attributes in normal samples, for any new data sample, it would be able to identify it as anomalous or a normal data point by using its learned knowledge. Unsupervised methods like clustering , K-nearest neighbors, auto-encoders , and so on can detect anomalies based on data and its features.



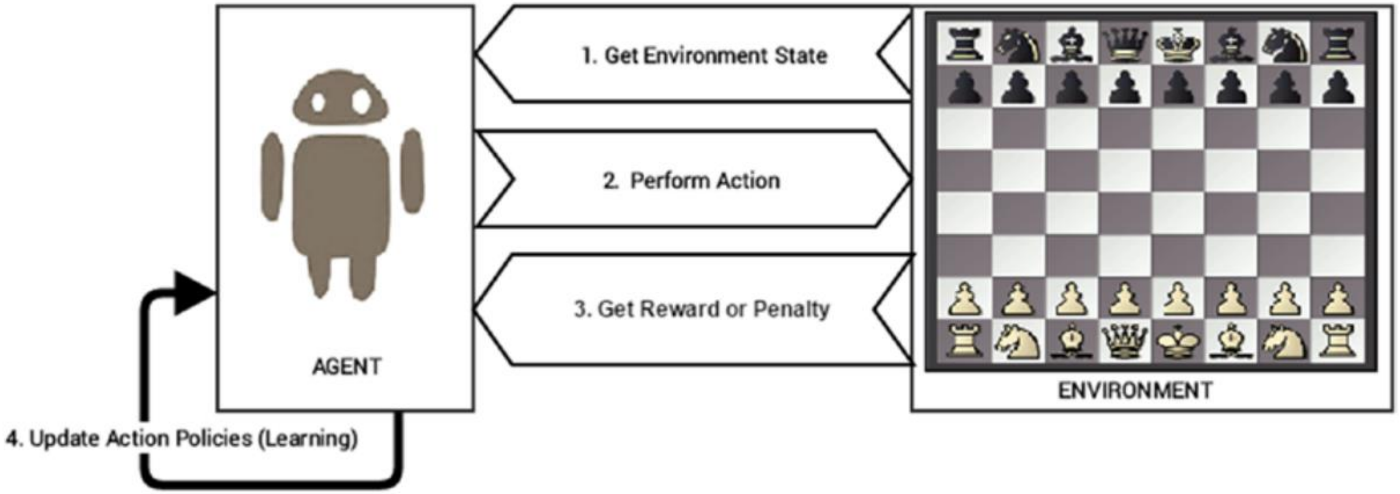


Reinforcement Learning

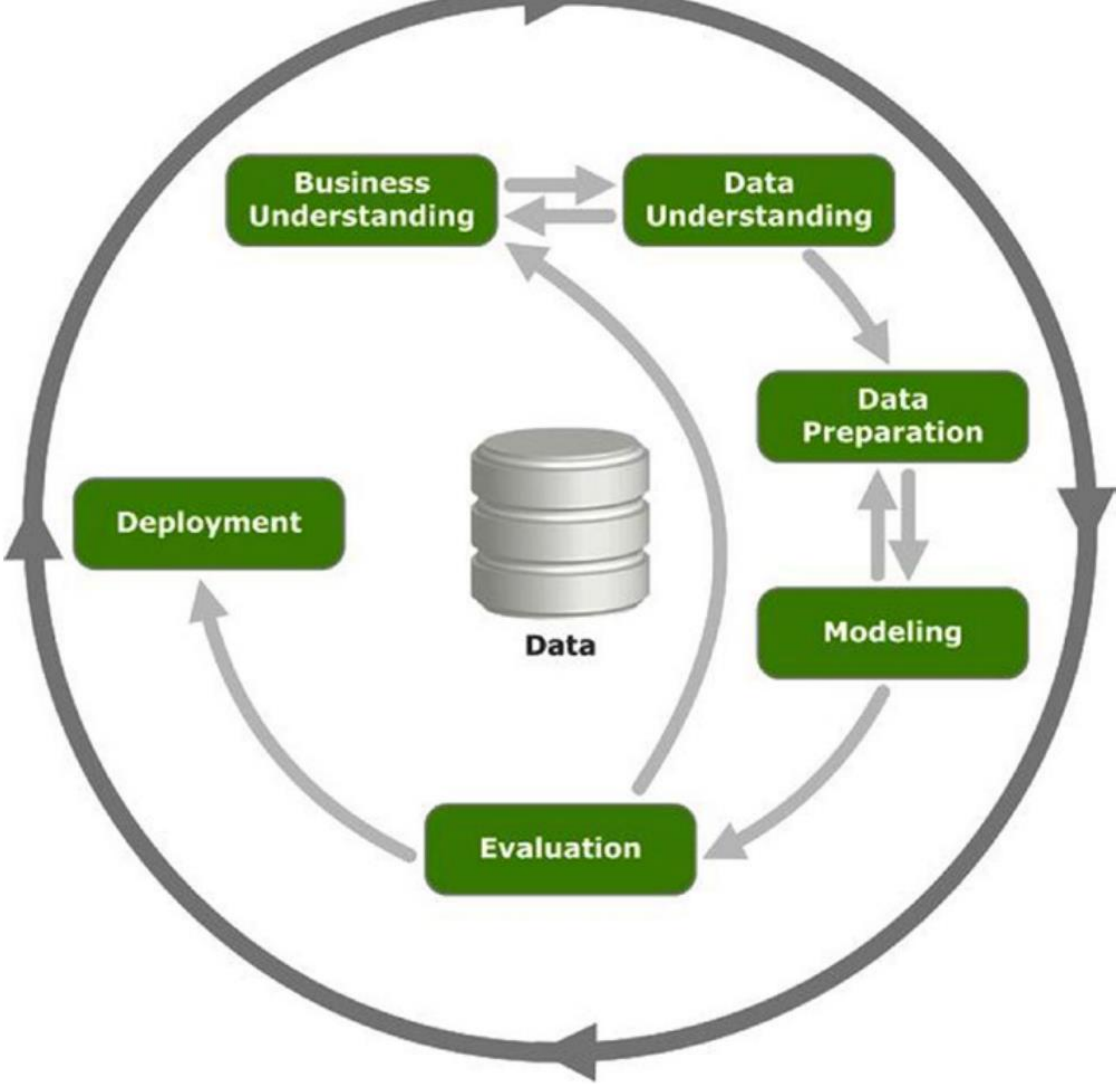
The reinforcement learning methods are a bit different from conventional supervised or unsupervised methods. In this context, we have an agent that we want to train over a period of time to interact with a specific environment and improve its performance over a period of time with regard to the type of actions it performs on the environment. Typically the agent starts with a set of strategies or policies for interacting with the environment. On observing the environment, it takes a particular action based on a rule or policy and by observing the current state of the environment. Based on the action, the agent gets a reward, which could be beneficial or detrimental in the form of a penalty. It updates its current policies and strategies if needed and this iterative process continues till it learns enough about its environment to get the desired rewards. The main steps of a reinforcement learning method are mentioned as follows.

Consider a real-world problem of trying to make a robot or a machine learn to play chess. In this case the agent would be the robot and the environment and states would be the chessboard and the positions of the chess pieces. A suitable reinforcement learning methodology is depicted in Figure

- Prepare agent with set of initial policies and strategy
- Observe environment and current state
- Select optimal policy and perform action
- Get corresponding reward (or penalty)
- Update policies if needed
- Repeat Steps 2 - 5 iteratively until agent learns the most optimal policies.



GOOGLE'S DEEPMIND BUILT THE ALPHAGO AI WITH COMPONENTS OF REINFORCEMENT LEARNING TO TRAIN THE SYSTEM TO PLAY THE GAME OF GO.



Building Machine Intelligence

To solve real-world problems by building machine intelligence using a structured process ,we establish full-fledged end-to-end Machine Learning pipelines based on the CRISP-DM model.

The CRISP-DM model tells us that for building an end-to-end solution for any analytics project or system, there are a total of six major steps or phases, some of them being iterative.

1. Business Understanding : this is one of the most important phases .The main objective here starts with understanding the business context and requirements for the problem to be solved at hand.
2. Data Understanding : The second phase in the CRISP-DM process involves taking a deep dive into the data available and understanding it in further detail before starting the process of analysis. This involves the following steps:

- Data Collection : This task is undertaken to extract, curate, and collect all the necessary data needed for your business objective. This can be obtained from the web, i.e., open data sources or it can be obtained from other channels like surveys, purchases, experiments and simulations.
- Exploratory Data Analysis : Exploratory data analysis, also known as EDA, is one of the first major analysis stages in the lifecycle. Here, the main objective is to explore and understand the data in detail. You can make use of descriptive statistics, plots, charts, and visualizations to look at the various data attributes, find associations and correlations

- Data Quality Analysis : is the final stage in the data understanding phase where we analyze the quality of data in our datasets ,the focus on data quality analysis involves the following :
  - Missing values
  - Inconsistent values
  - Wrong information due to data errors (manual/automated)

3. Data Preparation

The third phase in the CRISP-DM process takes place after gaining enough knowledge on the business problem and relevant dataset. Data preparation is mainly a set of tasks that are performed to clean, wrangle, curate, and prepare the data before running any analytical or Machine Learning methods and building models. An important point to remember here is that data preparation usually is the most time-consuming phase and often takes 60% to 70% time in the overall project, because, bad data will lead to bad models and poor performance and results. This phase involves the following :

- Data Integration : it is mainly done when we have multiple datasets that we might want to integrate or merge. This can be done in two ways. Appending several datasets by combining them, which is typically done for datasets having the same attributes. Merging several datasets together having different attributes or columns, by using common fields like keys.
- Data wrangling : involves data processing, cleaning, normalization, and formatting. Data in its raw form is rarely consumable by Machine Learning methods to build models. Hence we need to process the data based on its form, clean underlying errors and inconsistencies, and format it into more consumable formats for ML algorithms.

- Attribute Generation and Selection:

Data is comprised of observations or samples (rows) and attributes or features (columns). The process of attribute generation is also known as feature extraction and engineering in Machine Learning terminology. Attribute generation is basically creating new attributes or variables from existing attributes based on some rules, logic, or hypothesis. Attribute selection is basically selecting a subset of features or attributes from the dataset based on parameters like attribute importance, quality, relevancy, assumptions, and constraints. Sometimes even Machine Learning methods are used to select relevant attributes based on the data. This is popularly known as feature selection in Machine Learning terminology.

4. Modeling

The fourth phase in the CRISP-DM process is the core phase in the process where most of the analysis takes place with regard to using clean, formatted data and its attributes to build models to solve business problems. This is an iterative process, as depicted in Figure earlier. The basic idea is to build multiple models iteratively trying to get to the best model that satisfies our success criteria.

5. Evaluation

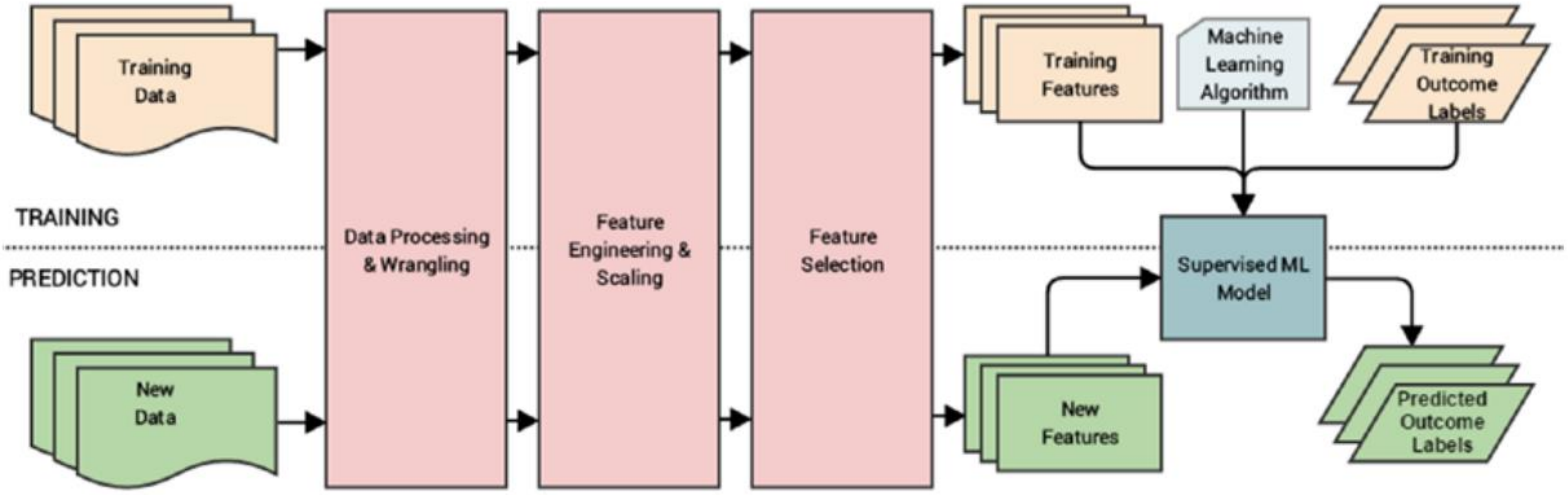
The fifth phase in the CRISP-DM process takes place once we have the final models from the modeling phase that satisfy necessary success criteria goals and have the desired performance and results with regard to model evaluation metrics like accuracy.

6. Deployment

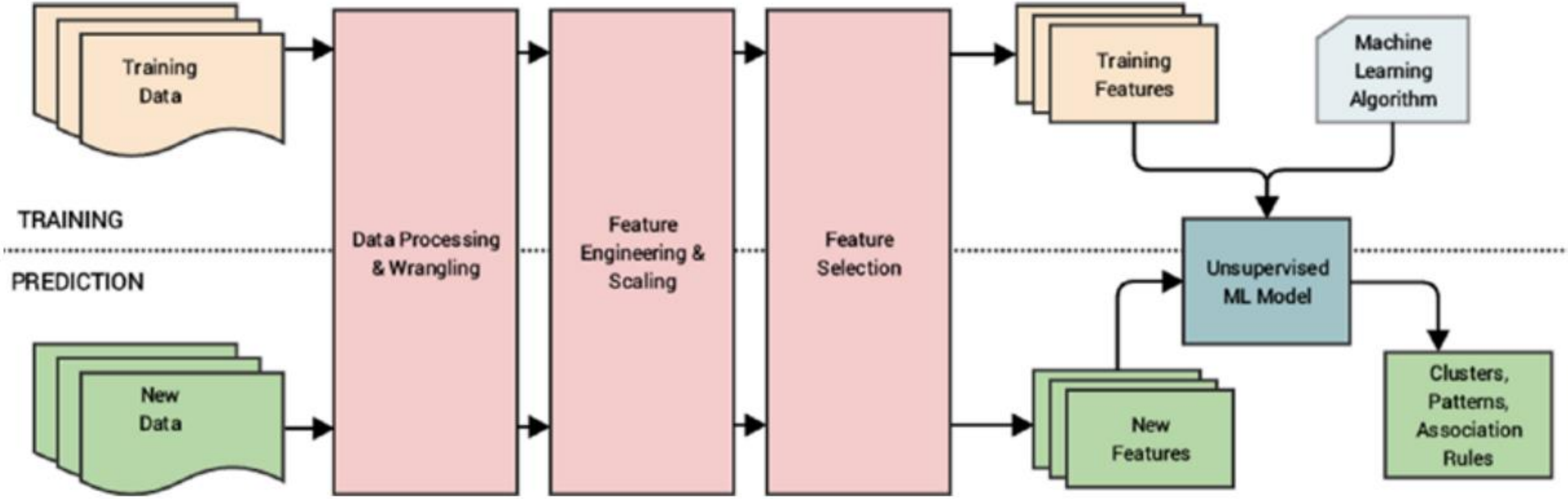
Deployment The final phase in the CRISP-DM process is all about deploying your selected models to production and making sure the transition from development to production is seamless.

Machine Learning Pipelines

The best way to solve a real-world Machine Learning or analytics problem is to use a Machine Learning pipeline starting from getting your data to transforming it into information and insights.



Supervised Machine Learning pipeline



Unsupervised Machine Learning pipeline