VAASAN AMMATTIKORKEAKOULU
UNIVERSITY OF APPLIED SCIENCES

Machine Learning

Dr. Ghodrat Moghadampour

mg@vamk.fi

Vaasa University of Applied Sciences

Machine Learning

Contents

- Machine learning
- Data Processing
- Learning Algorithms
- Types of Machine Learning
  - Supervised
  - Unsupervised
  - Reinforcement
- Machine Learning Steps
  - Data Collection
  - Data Exploration and Preparation
  - Model Building and Training
  - Model Evaluation and Validation

Machine Learning

- Machine Learning (ML) is the ability of a computer to learn without explicit programming.
- To predict output values within a satisfactory range, machine learning uses designed algorithms to obtain and interpret input data.
- They learn and optimise their operations as new data is fed into these algorithms to enhance performance and develop intelligence over time.

Machine Learning

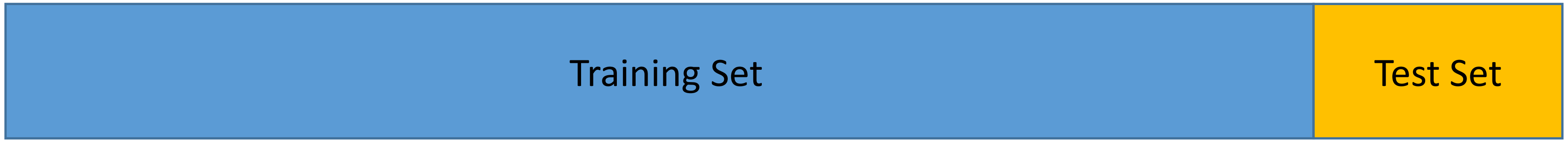
- To predict output values within a satisfactory range, machine learning uses designed algorithms to obtain and interpret input data.
- They learn and optimise their operations as new data is fed into these algorithms to enhance performance and develop intelligence over time.

Machine Learning

- In essence, these tasks all seek to learn from data.
- To address each scenario, we can use a given set of features to train an algorithm and extract insights.
- These algorithms, or learners, can be classified according to the amount and type of supervision needed during training.
- The learning task we hope to accomplish, determines which type of learning we will use.

Data Preprocessing

- Collected data must be divided into two different sets:
  - Training set: a subset to train a model
  - Test set: a subset to test the trained model
- Typically, the ratio of the training set and test set is 80 to 20 percent.
- The model must never be trained on the test set.
- The test set must meet the following two conditions:
  - It must be large enough to yield statistically meaningful results.
  - It must be representative of the data set as a whole.
  - Test set must not have different characteristics than the training set.



Training Set

Test Set

Learning Algorithms

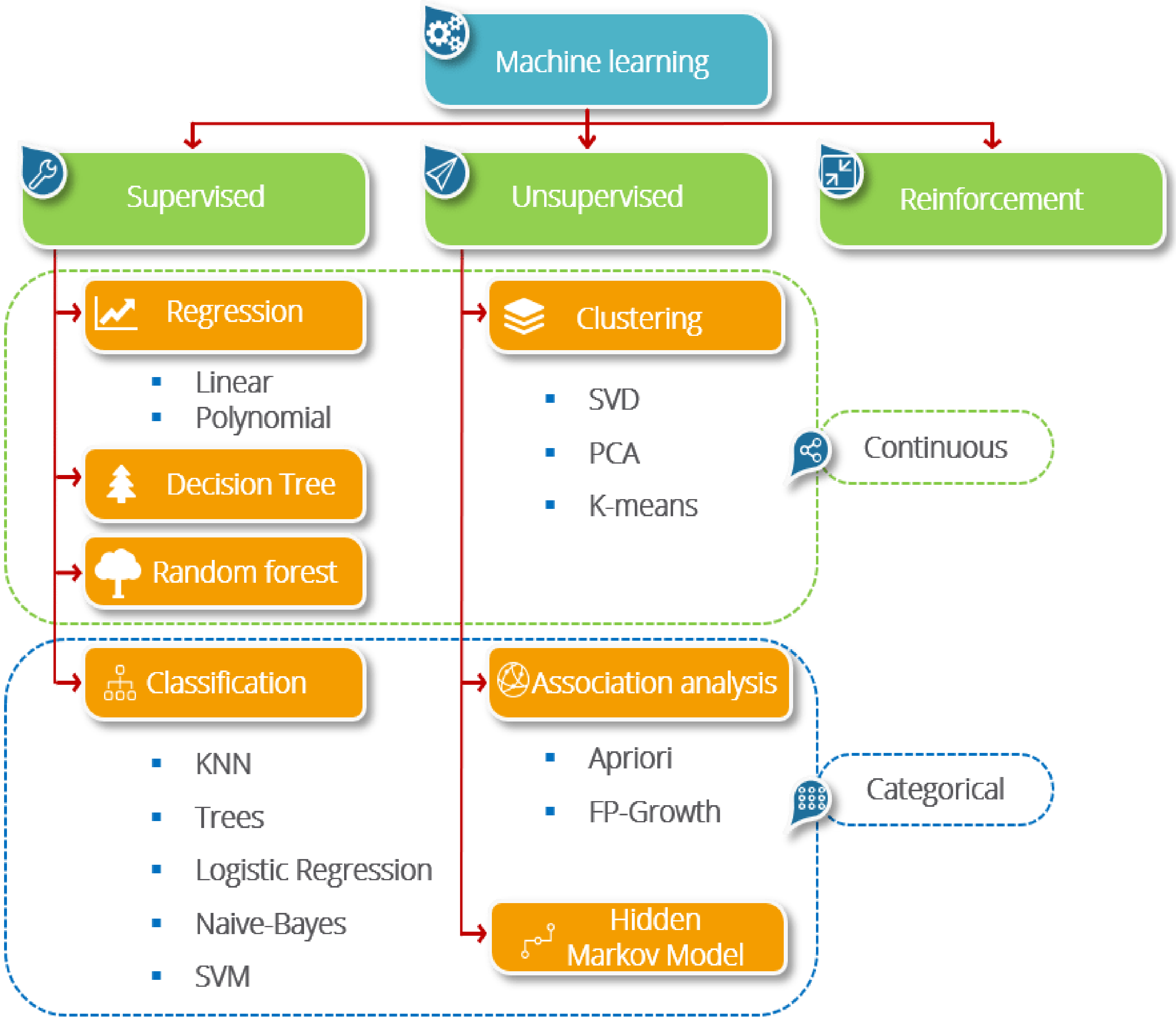
- How does the computer know whether it is getting better or not?
- How does it know how to improve?
- Different answers to these questions produce different types of machine learning.

Learning Algorithm Scenarios

1. We can tell the algorithm the correct answer for a problem so that it gets it right next time. We hope that we only have to tell it a few right answers and then it can work out how to get the correct answers for other problems (generalise).
2. Alternatively, we can tell it whether or not the answer was correct, but not how to find the correct answer, so that it has to search for the right answer.
3. A variant of this is that we give a score for the answer, according to how correct it is, rather than just a "right" or "wrong" response.
4. Finally, we might not have any correct answers; we just want the algorithm to find inputs that have something in common.

Types of Machine Learning

- There are several categories of algorithms for machine learning.
- They are largely classified as:
  - Supervised, which constructs predictive models to forecast likely future outcomes
  - Unsupervised, which constructs descriptive models to understand outcomes
  - Reinforcement


Machine learning
Supervised
Unsupervised
Reinforcement
Regression
Linear
Polynomial
Decision Tree
Random forest
Clustering
SVD
PCA
K-means
Continuous
Classification
KNN
Trees
Logistic Regression
Naive-Bayes
SVM
Association analysis
Apriori
FP-Growth
Categorical
Hidden
Markov Model

Algorithms

- In linear algebra, the Singular Value Decomposition (SVD) is a factorization of a real or complex matrix.
- The Principal Component Analysis (PCA) is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss.
- K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.
- The K-Nearest Neighbors algorithm (KNN) is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.
- Decision trees are an approach used in supervised machine learning, a technique which uses labelled input and output datasets to train models. The approach is used mainly to solve classification problems, which is the use of a model to categorize or classify an object.

Algorithms

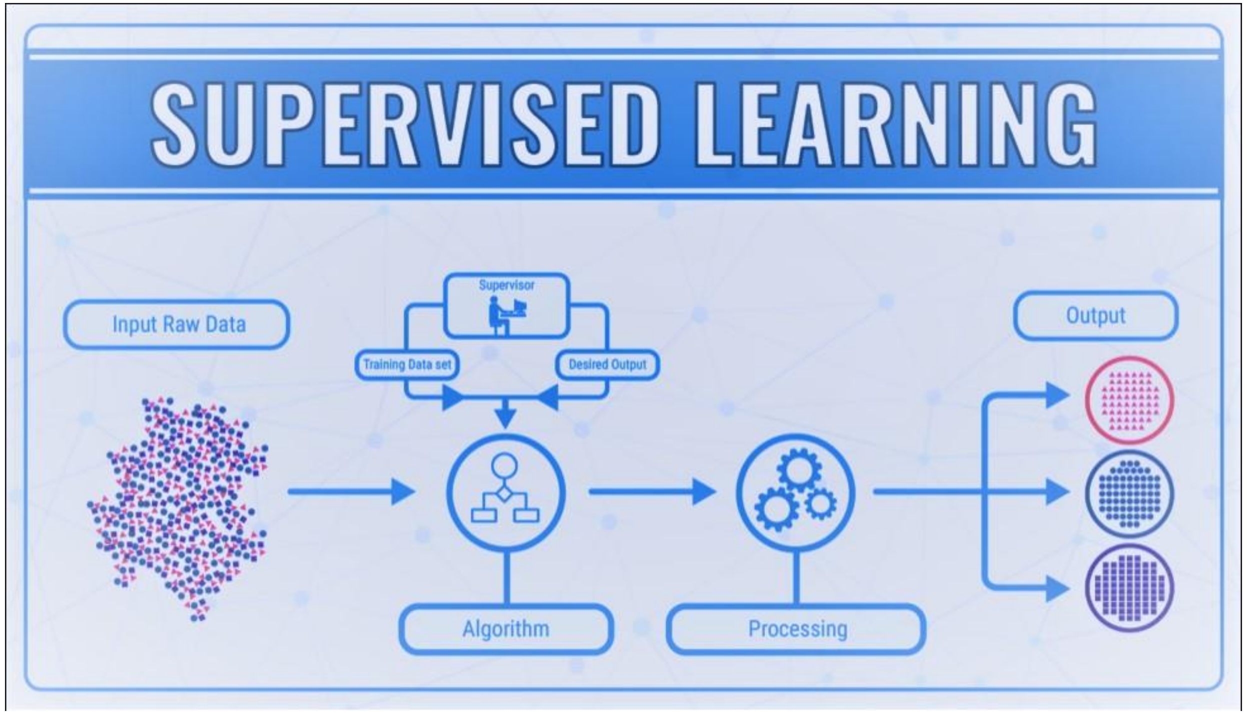
- Logistic Regression is a classification algorithm that comes under the supervised category of machine learning in which machines are trained using "labelled" data, and on the basis of that trained data, the output is predicted.
- Naive Bayes in machine learning is defined as probabilistic model in machine learning technique in the genre of supervised learning that is used in varied use cases of mostly classification, but applicable to regression as well.
- Support Vector Machine (SVM) can be used for both regression and classification tasks. It is highly preferred by many as it produces significant accuracy with less computation power.
- The Apriori algorithm uses frequent item sets to generate association rules, and it is designed to work on the databases that contain transactions. With the help of these association rule, it determines how strongly or how weakly two objects are connected.

Algorithms

- FP Growth is one of the associative rule learning techniques which is used in machine learning for finding frequently occurring patterns. It is a rule-based machine learning model. It is a better version of Apriori method. This is represented in the form of a tree, maintaining the association between item sets.

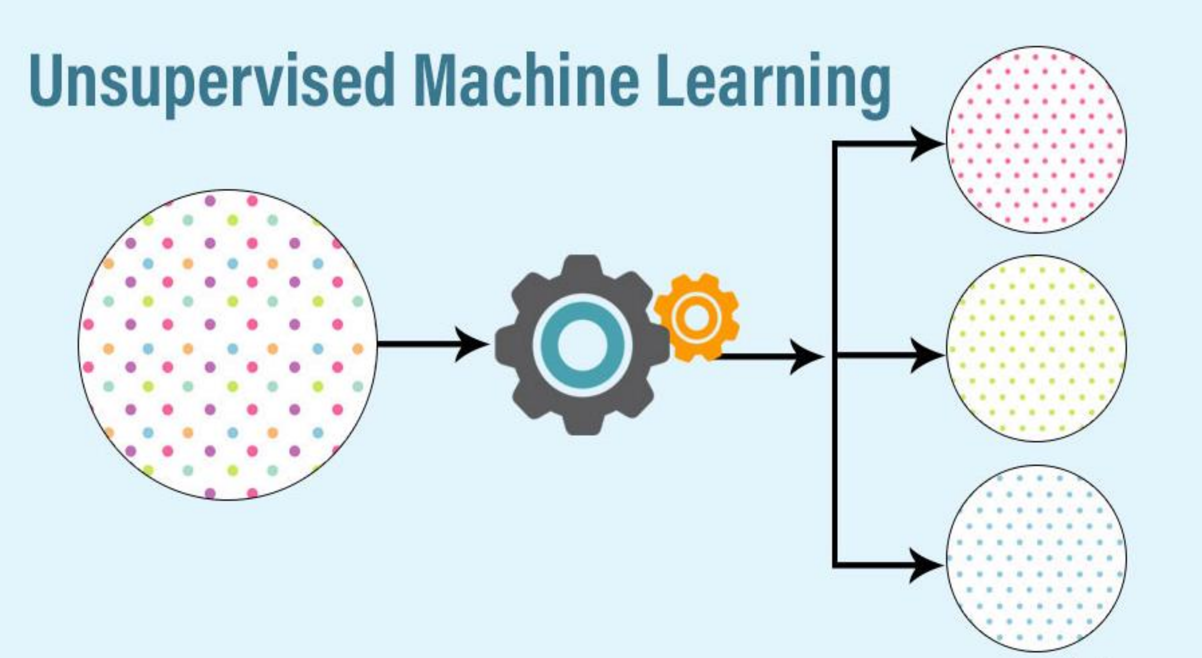
Supervised Learning

- A training set of examples with the correct responses (targets) is provided and, based on this training set, the algorithm generalizes to respond correctly to all possible inputs. This is also called learning from exemplars.



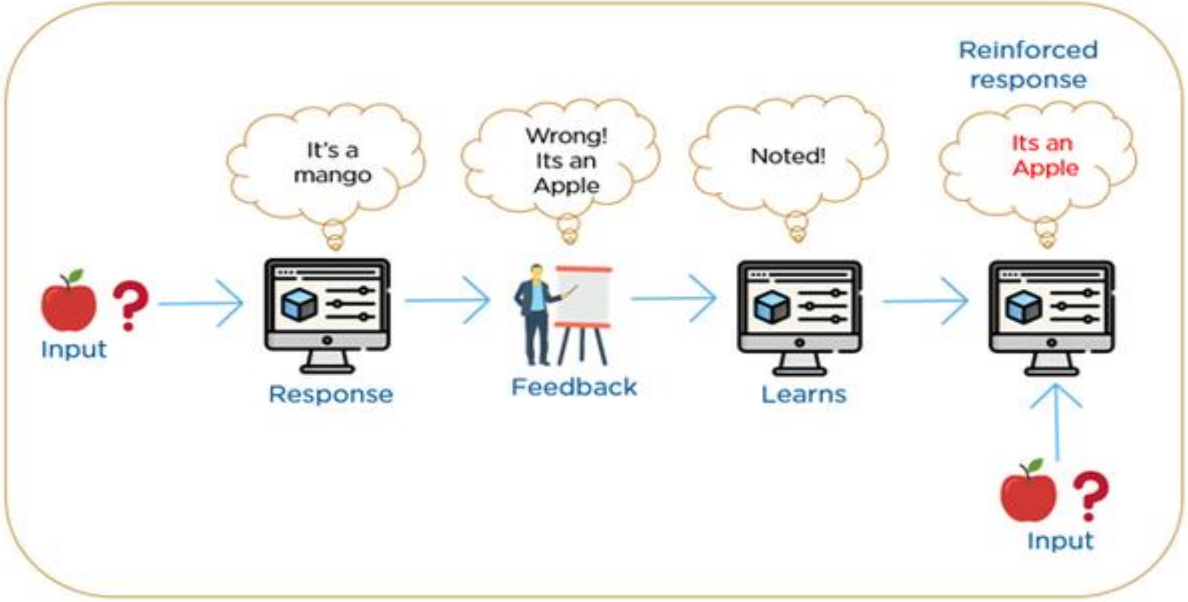
Unsupervised Learning

- Correct responses are not provided, but instead, the algorithm tries to identify similarities between the inputs so that inputs that have something in common are categorized together.



Reinforcement Learning

- Reinforcement learning is somewhere between supervised and unsupervised learning.
- The algorithm gets told when the answer is wrong, but does not get told how to correct it.



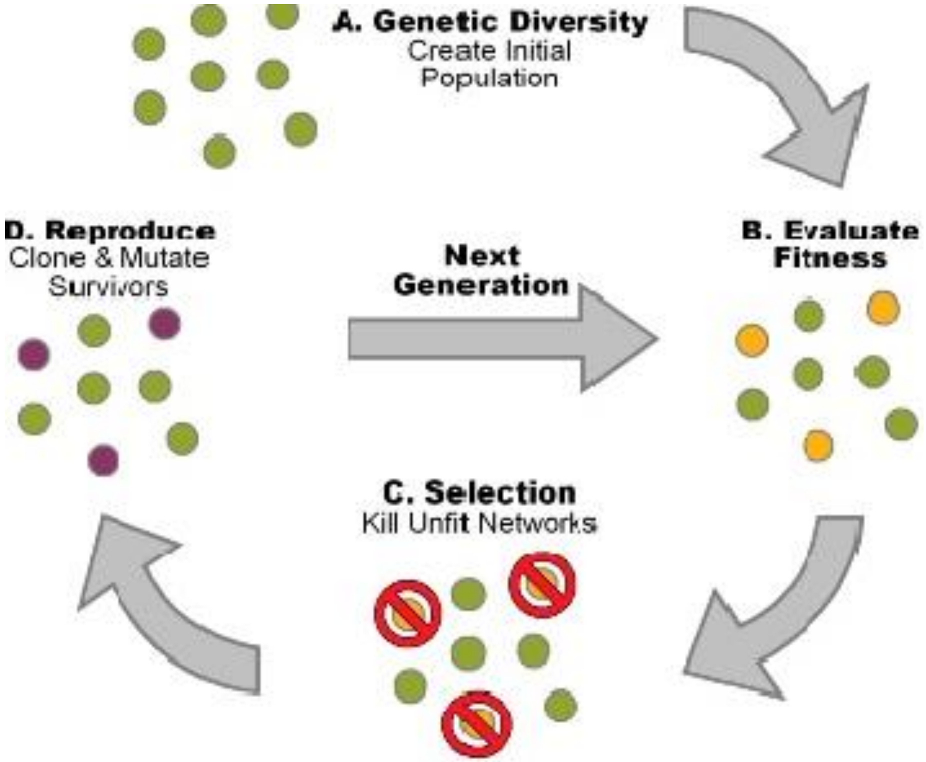
Reinforcement Learning

- The algorithm has to explore and try out different possibilities until it works out how to get the answer right.
- Reinforcement learning is sometime called learning with a critic because of this monitor that scores the answer, but does not suggest improvements.

Evolutionary Learning

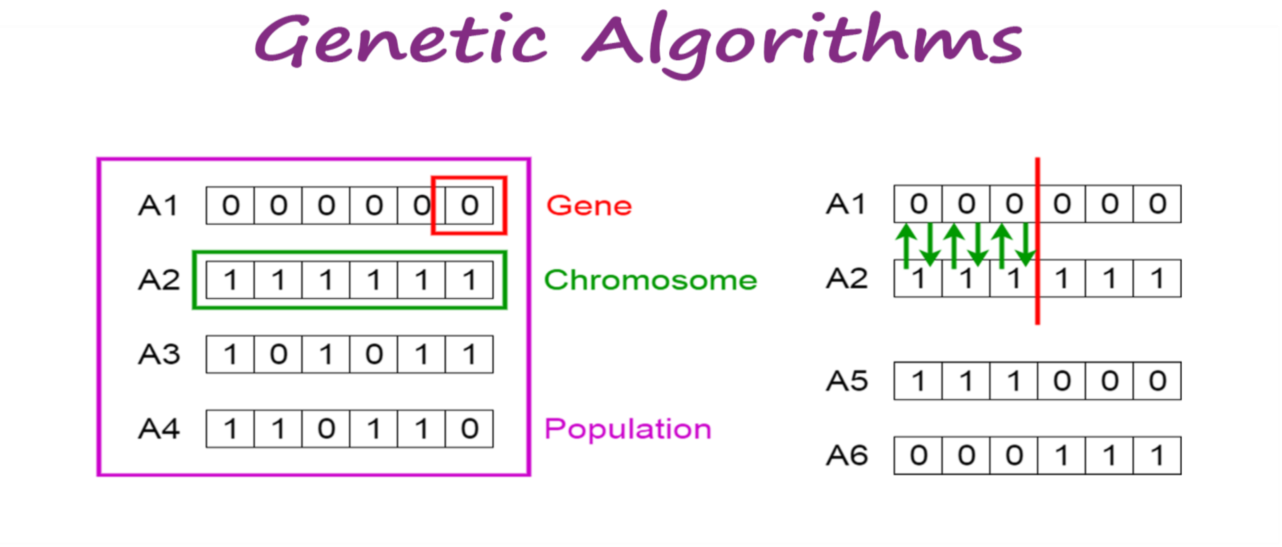
- Biological evolution can be seen as a learning process since biological organisms adapt to improve their survival rates and chance of having offspring in their environment.
- To model this in a computer, we use the idea of fitness, which corresponds to a score for how good the current solution is.

Evolutionary Learning



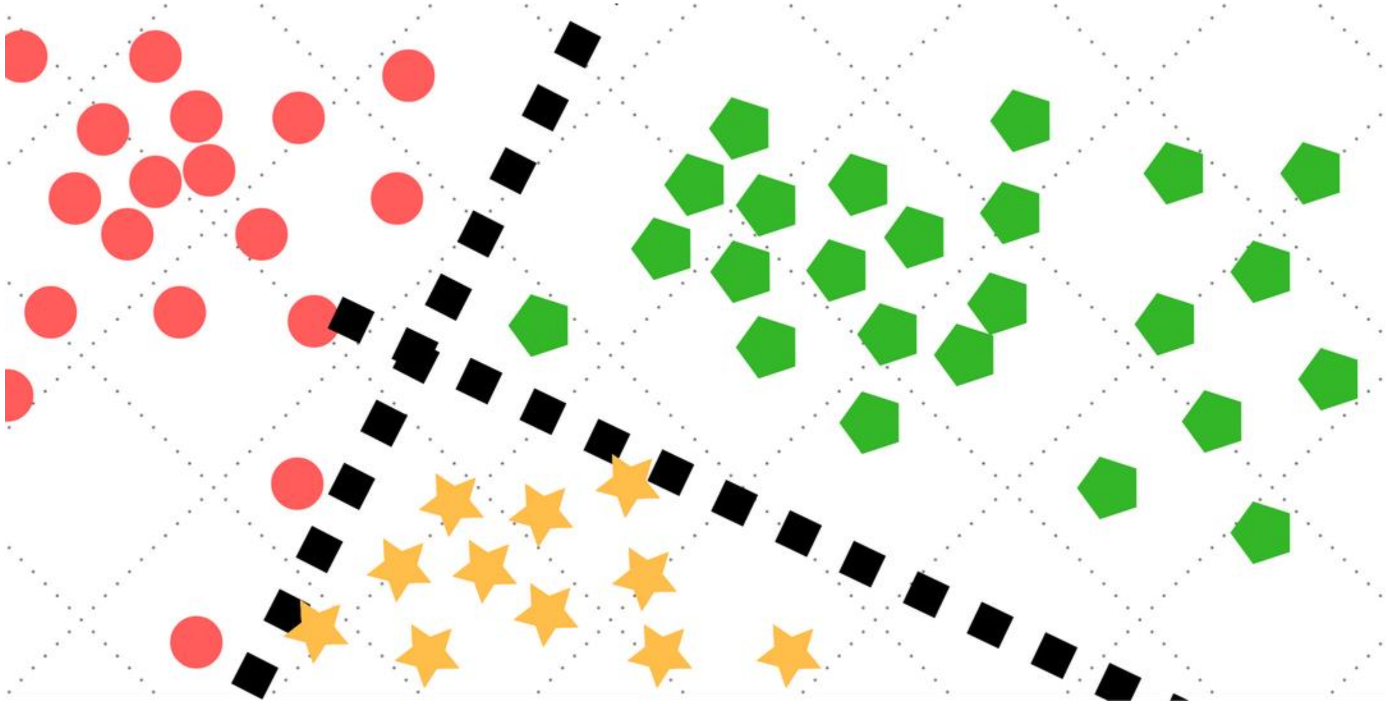
Evolutionary Learning

- The evolutionary algorithms use the idea of populations, which will be refined through breeding and applying genetic operators, like crossover and mutation in the hope of generating better individuals.



Classification

- Classification is the supervised learning process where classes are sometimes referred to as targets/labels or categories to predict the class of given data points.

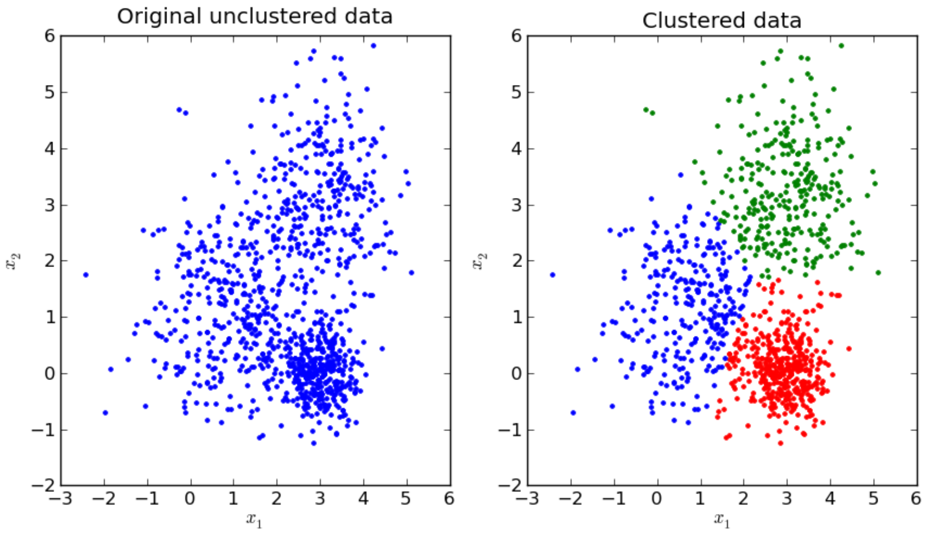


Classification

- The machine learning programs draw conclusions in classification from given values and find the category to which new data points refer to.
- For example, a Bottle-Return device examines inserted things, classifies them and filters them out as "bottle" and "not bottle".

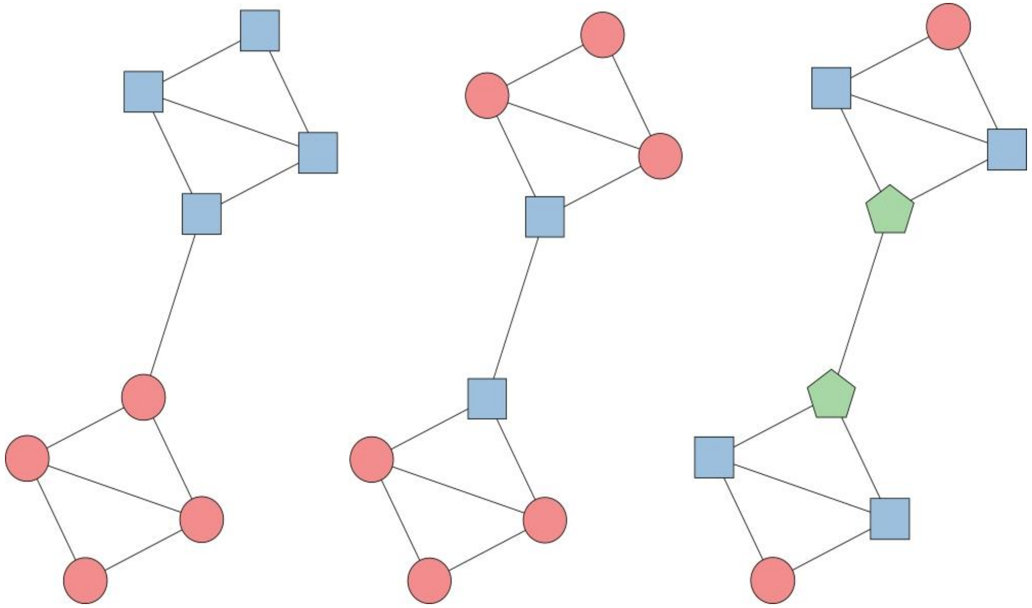
Clustering

- Grouping unlabeled data is called clustering.
- Clustering unlabeled data relies on unsupervised machine learning.
- Clustering is used to group together similar instances, and to see whether this allows fewer features to be used.



Clustering

- We can measure similarity between instances by combining the instances' feature data into a metric, called a similarity measure.
- When each instance is defined by one or two features, it's easy to measure similarity.



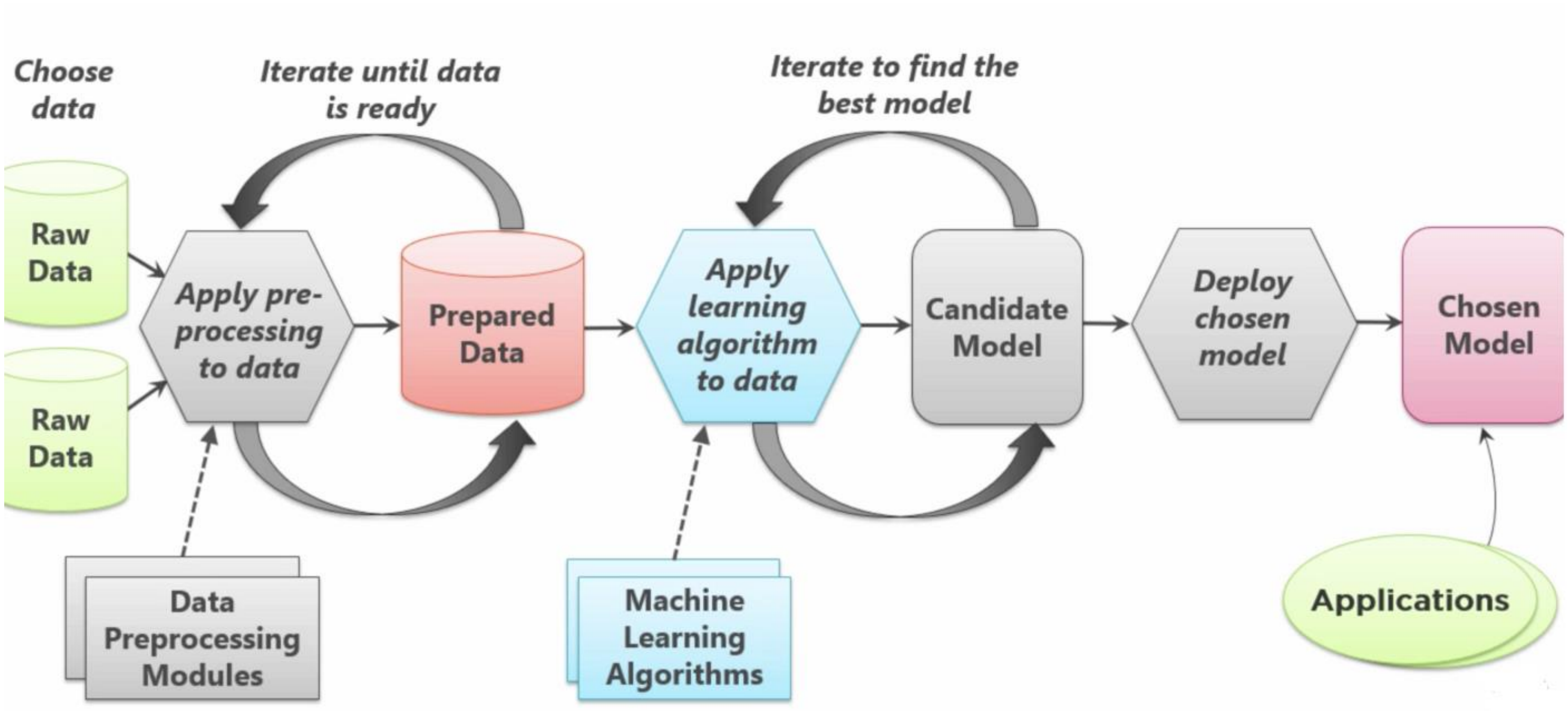
Clustering

- For example, we can find similar products by their producers.
- As the number of features increases, creating a similarity measure becomes more complex.
- Clustering has a numerous uses in a variety of industries, like:
  - Anomaly detection (identifies data points, events, and/or observations that deviate from a dataset's normal behavior)
  - Image segmentation (partitioning a digital image into multiple segments; sets of pixels, also known as image objects)
  - Market segmentation (grouping future buyers into segments with common needs and who respond similarly to a marketing action)
  - Social network analysis (mapping and measuring social relations to discover, analyze and visualize the social networks of criminal suspects)

Machine Learning Steps

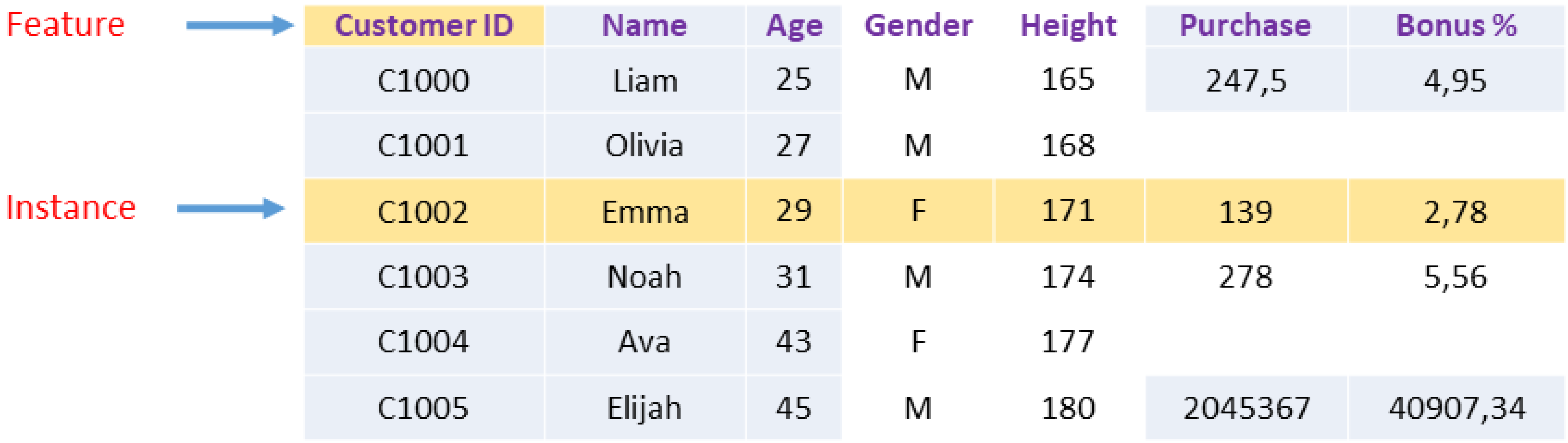
- A machine learning task occurs through the following steps:
  - Data collection, where raw data required for analysis is gathered from various sources.
  - Data exploration and preparation, where data is studied deeply and data is pre-processed using different techniques so as to prepare a high-quality data.
  - Model building and training, where the machine is trained using the algorithms and other machine learning techniques.
  - Model evaluation and implementation, where built model is evaluated and validated for accuracy and other performance measures, like precision.

Machine Learning Steps


Choose data
Iterate until data is ready
Iterate to find the best model
Raw Data
Raw Data
Apply pre-processing to data
Prepared Data
Apply learning algorithm to data
Candidate Model
Deploy chosen model
Chosen Model
Data Preprocessing Modules
Machine Learning Algorithms
Applications

Data Collection

- In data collection, raw data required for analysis is gathered from various sources and serves as the input to machine learning algorithms, to make intuitions from it.
- The input data is in the form of instances and features.



Feature →

Instance →

Customer ID	Name	Age	Gender	Height	Purchase	Bonus %
C1000	Liam	25	M	165	247,5	4,95
C1001	Olivia	27	M	168		
C1002	Emma	29	F	171	139	2,78
C1003	Noah	31	M	174	278	5,56
C1004	Ava	43	F	177		
C1005	Elijah	45	M	180	2045367	40907,34

Data Collection

- A feature is a characteristic or attribute that uniquely identifies the instance, and might be useful for learning the desired concept.
- A feature could be numeric, nominal or categorical.
  - A numeric variable is measured in numbers, such as the height of a person or the temperature.
  - A categorical variable is represented by a set of various levels, for example job (*librarian, writer, engineer, etc.*) is a nominal and education (*bachelor, master, doctor*) is an ordinal and age group (*0-6,7-15, 16-19*) is an interval variable.
- Type of the features determines the kind of machine learning algorithm to model.

Data Exploration and Preparation

- Any machine learning project is based on the quality of data it uses.
- The next step, data exploration and preparation is concerned with a deep study of data so as to prepare high-quality data through the following operations on the data set:
  - Cleaning
  - Removing null values
  - Detecting outliers and any suspicious value
  - Removing unwanted features

Data Exploration and Preparation

Customer ID	Name	Job	Age	Gender	Height	Purchase	Bonus %
C1000	Liam	Marketing Coordinator	25	M	165	247,5	4,95
C1001	Olivia	Medical Assistant	27	M	168	🟥	🟥
C1002	Emma	Web Designer	29	F	171	139	2,78
C1003	Noah	President of Sales	31	M	174	278	5,56
C1004	Ava	Marketing Coordinator	43	F	177	🟥	🟥
C1005	Elijah	Medical Assistant	45	M	180	2045367	40907,34
C1006	Oliver	Web Designer	47	M	183	235	2,35
C1007	Sophia	President of Sales	49	F	186	30	0,3
C1008	Amelia	Marketing Coordinator	51	F		301234	3012,34
C1009	Lucas	Medical Assistant	53	M	183	🟥	🟥
C1010	Isabella	Web Designer	55	F	186	1139	11,39
C1011	Mason	President of Sales	57	M	185	20278	202,78
C1012	Ethan	Marketing Coordinator	59	M	188	20000	2
C1013	Mia	Medical Assistant	61	F	194	106473	10,6473

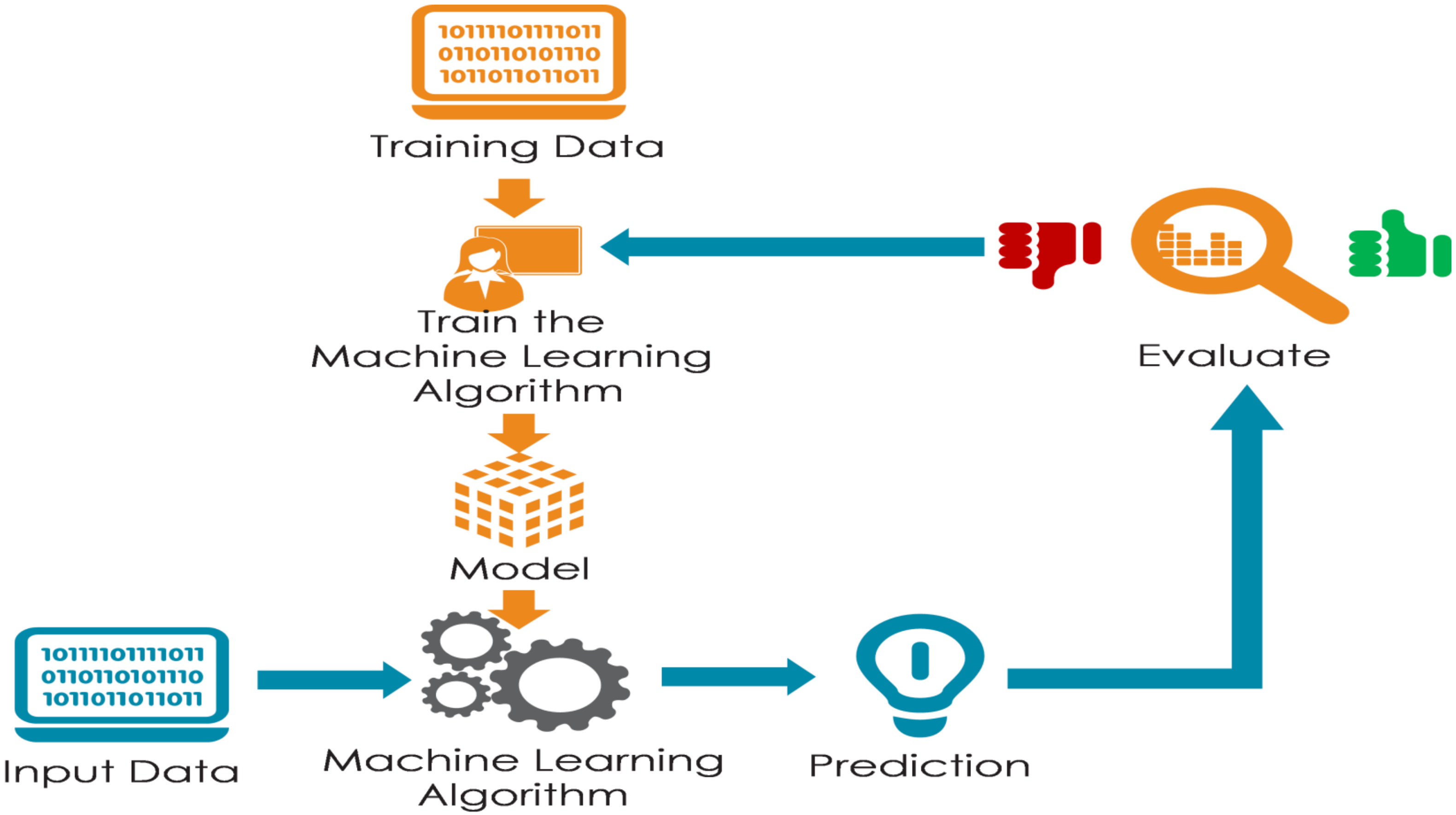
Model Building and Training

- The model is then built by training the machine using the algorithms and other machine learning techniques depending on what kind of analysis is required: descriptive, predictive or prescriptive.
  - Descriptive data analysis tries to provide insight into the past and find out what has happened.
  - Predictive data analysis tries to understand the future and find out what could happen.
  - Prescriptive analysis tries to advise on possible outcomes and answer what should be done.

Model Evaluation and Validation

- The built model must be then evaluated and validated for accuracy and other performance measures, like precision.
- If the model performance is not acceptable, a different model will be built.

Model Building and Evaluation



Thanks for your attention ☺