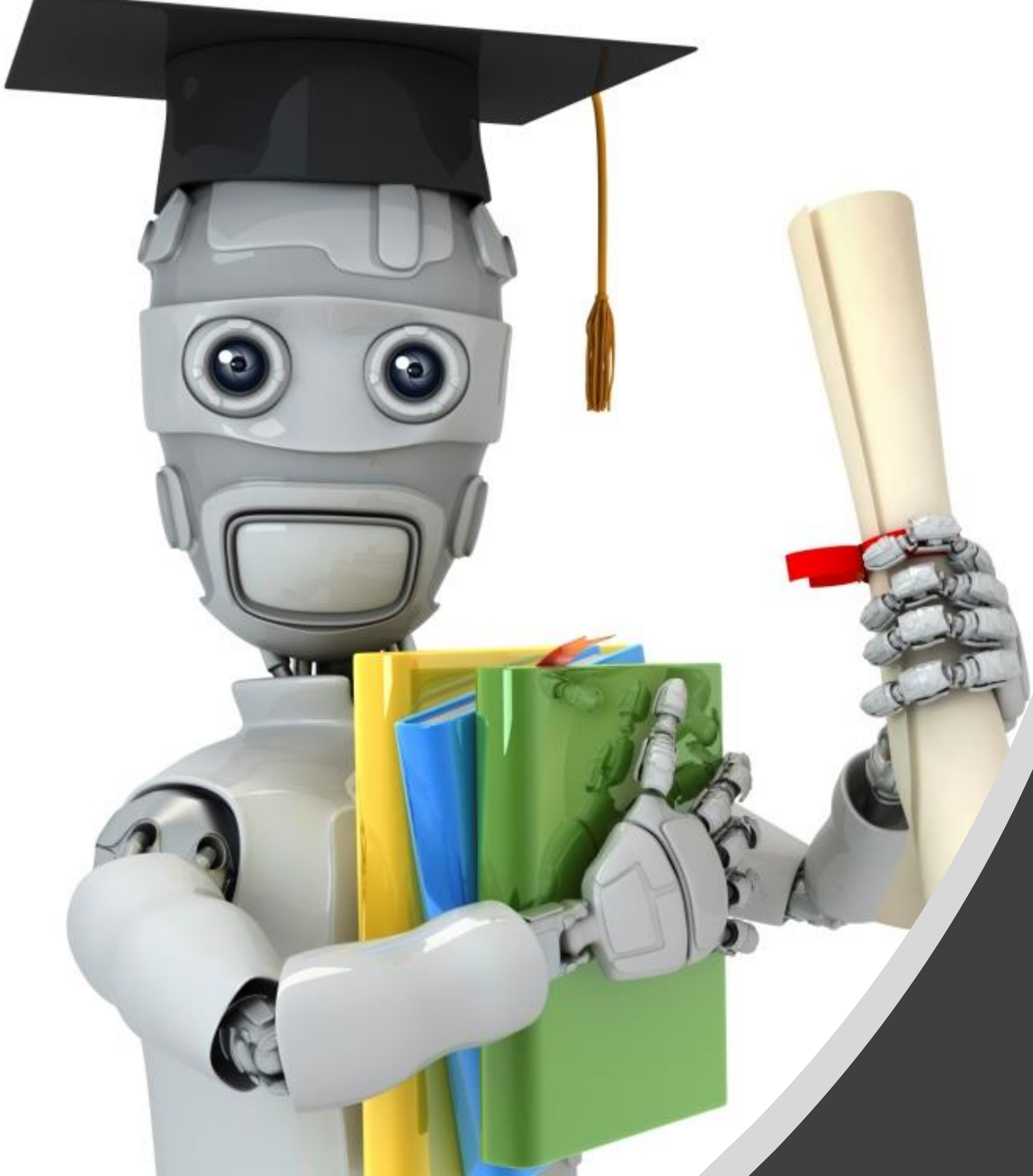
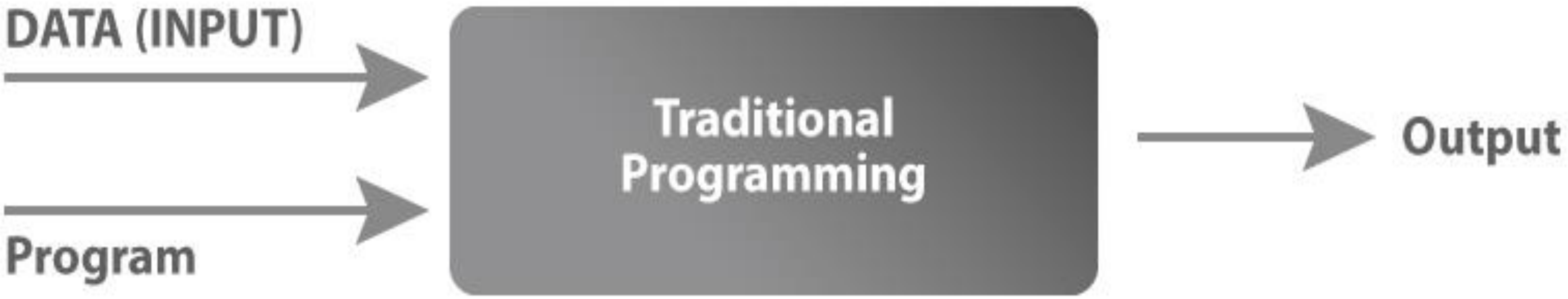


Machine Learning

Prepared By:

Dr. Sara Sweidan

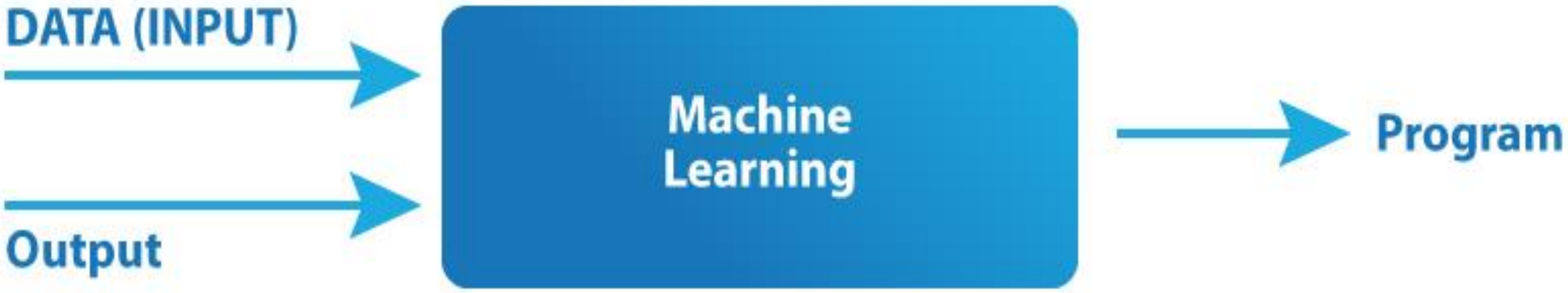


DATA (INPUT)

Program

Traditional Programming

Output

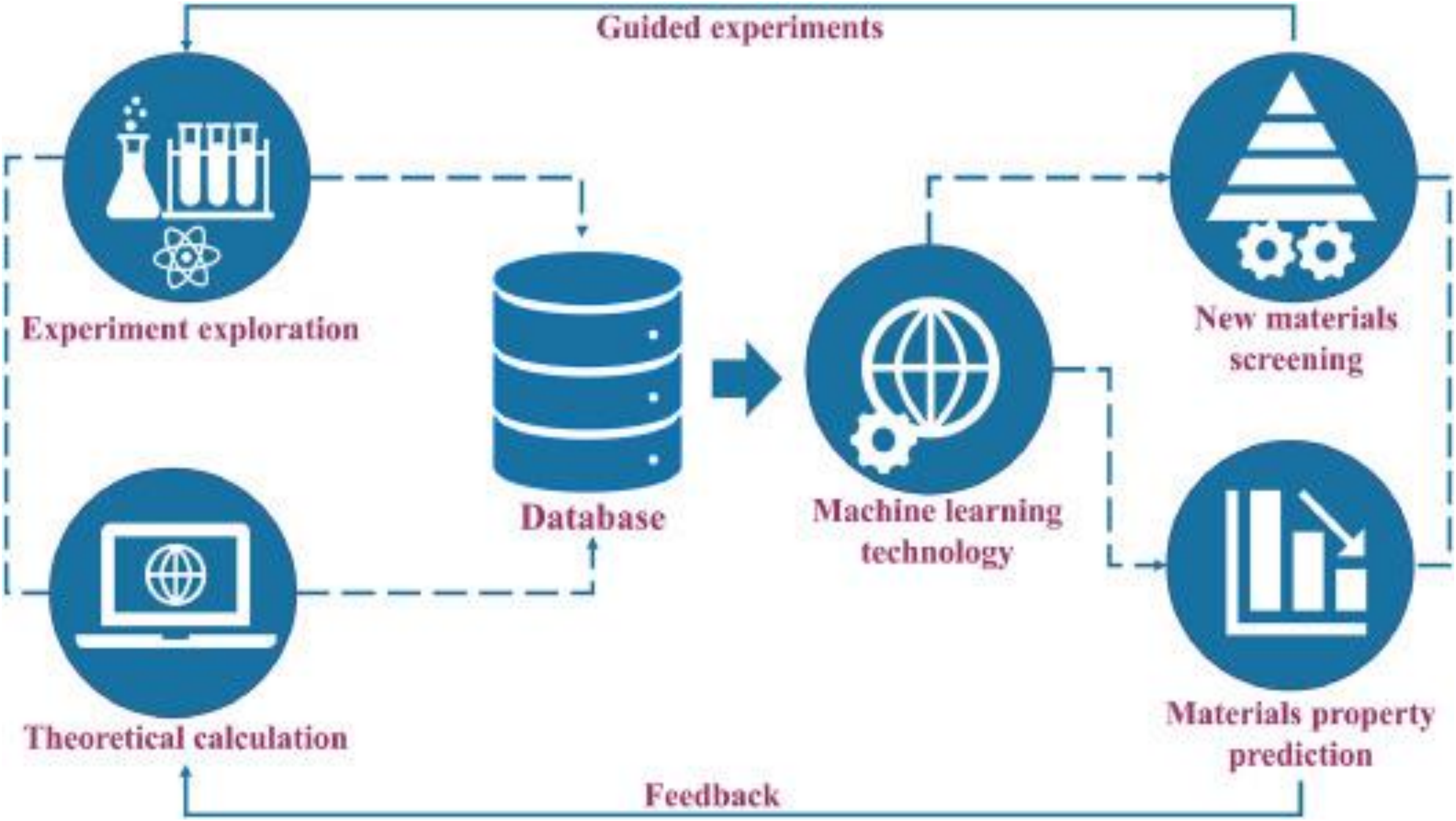


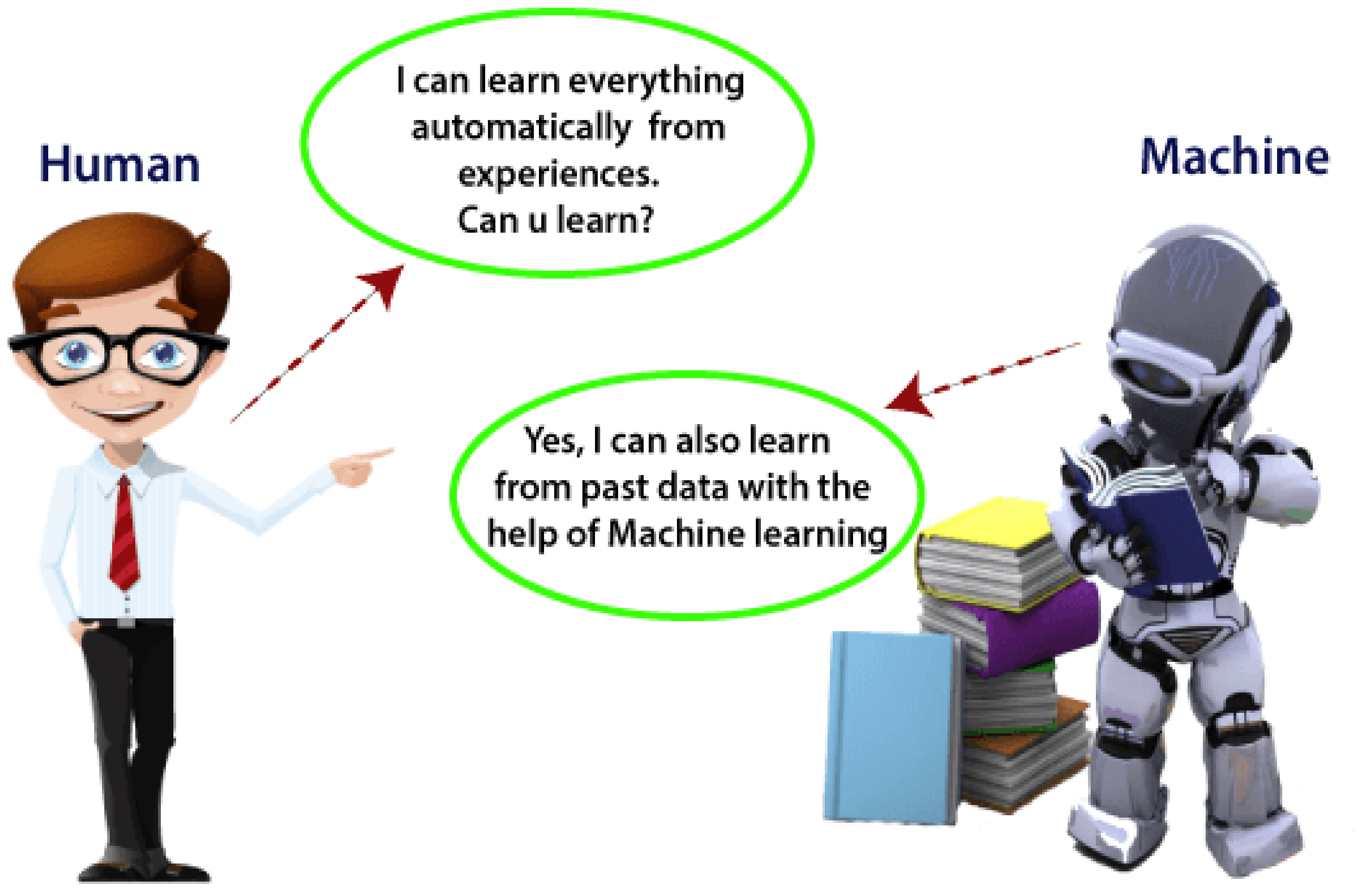
DATA (INPUT)

Output

Machine Learning

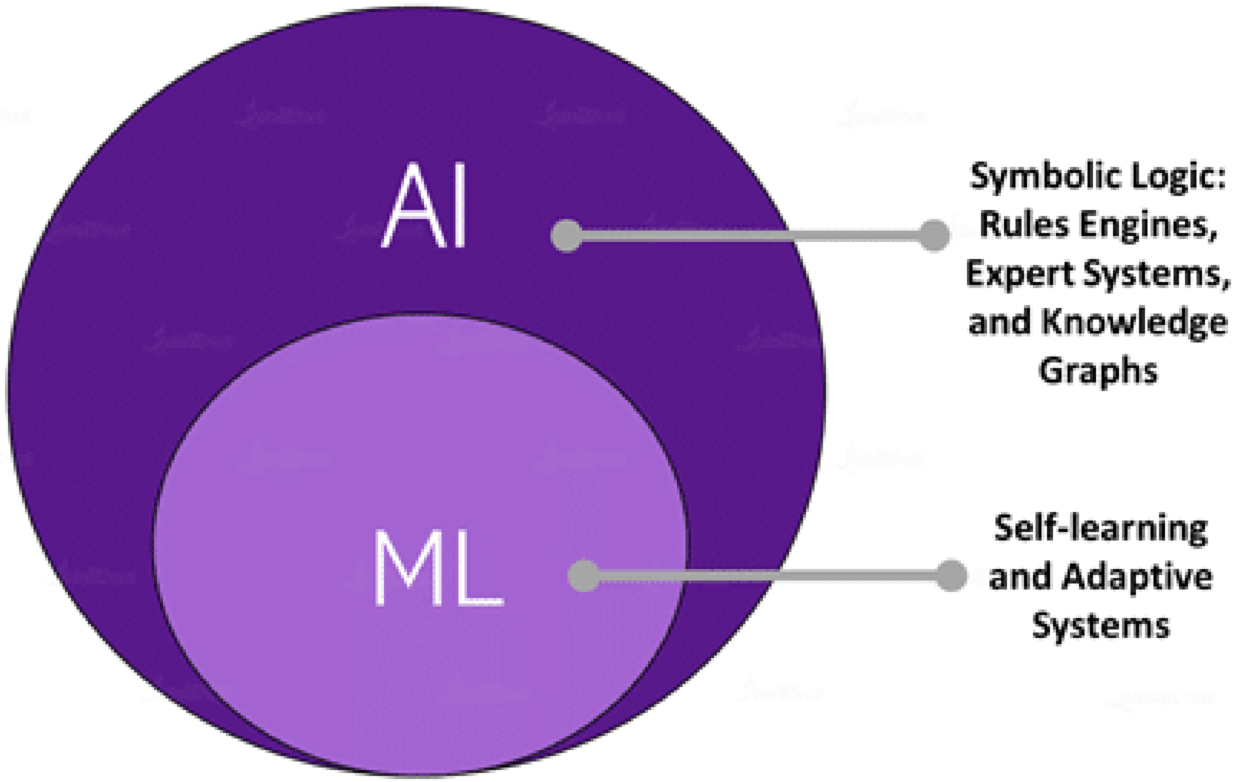
Program


Guided experiments
Experiment exploration
Database
Machine learning technology
New materials screening
Materials property prediction
Theoretical calculation
Feedback


Human
I can learn everything automatically from experiences. Can u learn?
Machine
Yes, I can also learn from past data with the help of Machine learning

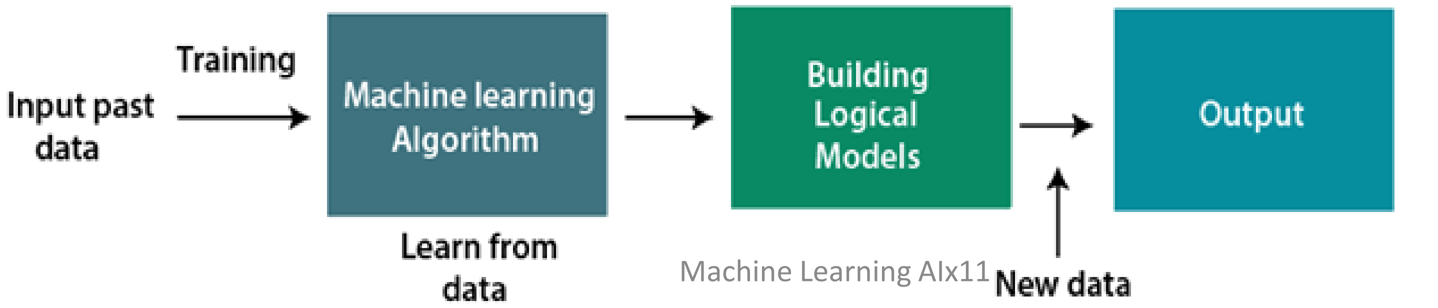
The origins of machine learning

- **Machine learning** is known as th development of computer algorithms to transform data into an intelligent action


AI
Symbolic Logic: Rules Engines, Expert Systems, and Knowledge Graphs
ML
Self-learning and Adaptive Systems

machine learning basics

- Machine Learning is a branch of artificial intelligence concerned with developing algorithms that allow a computer to learn automatically from **past data** and experiences.
- Machine learning builds **models** to make **predictions** using historical data or information.
- Historical data: Known as **training data**.
- **Whenever it receives new data, it predicts its output**.
- The accuracy of predictions depends on the **amount** of data: **More** data → **better** predictions.



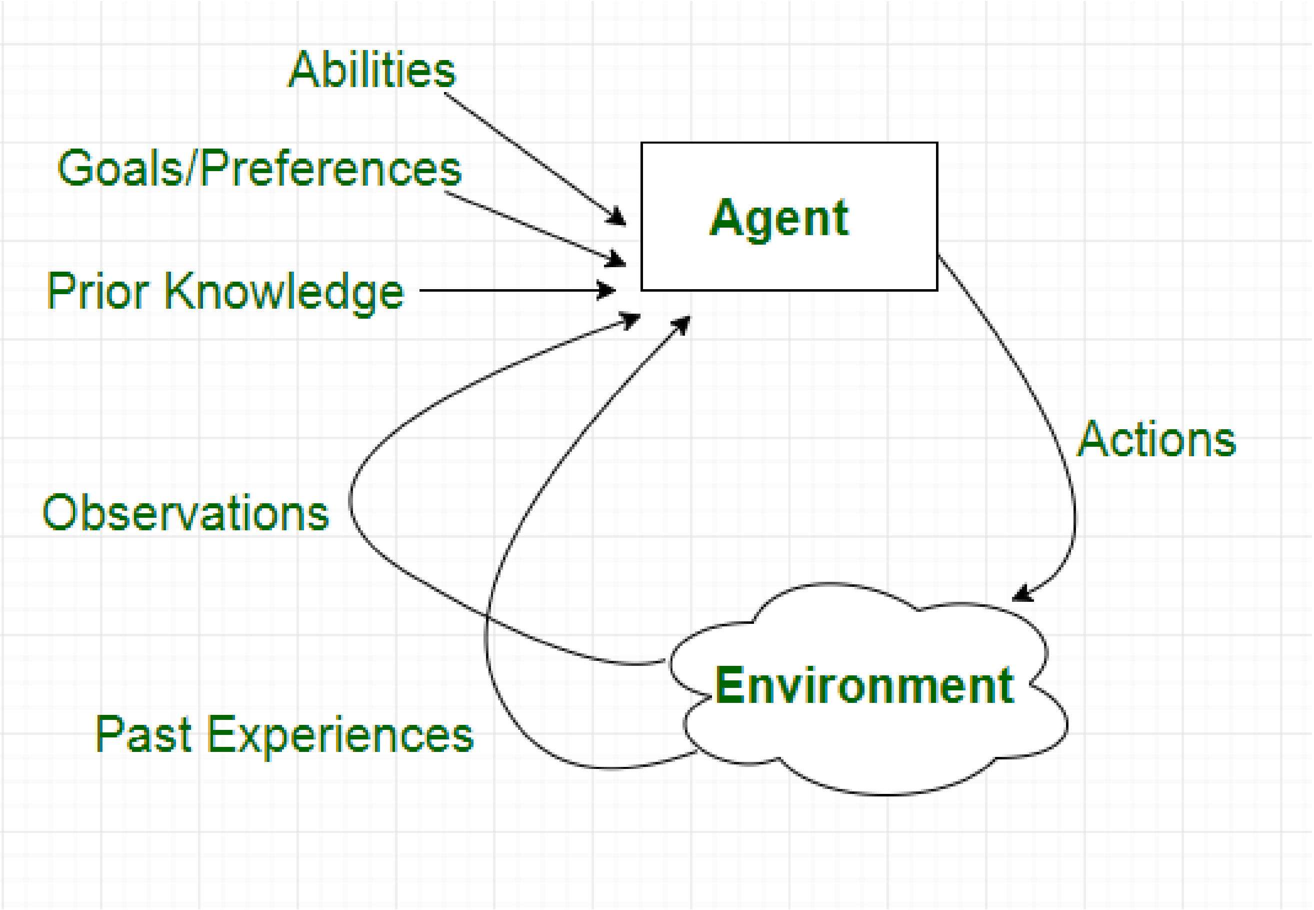
machine learning basics

Why Make Machines Learn?

- Lack of sufficient **human expertise** in a domain (e.g., simulating navigations in unknown territories or even spatial planets).
- Scenarios and **behavior** can keep **changing** over time (e.g., availability of infrastructure in an organization, network connectivity, and so on).
- Humans have **sufficient expertise** in the domain but it is extremely difficult to formally explain or translate this expertise into computational tasks (e.g., speech recognition, translation, scene recognition, cognitive tasks, and so on).
- Addressing **domain** specific **problems** at scale with huge volumes of data with too many complex conditions and constraints.

Machine learning features

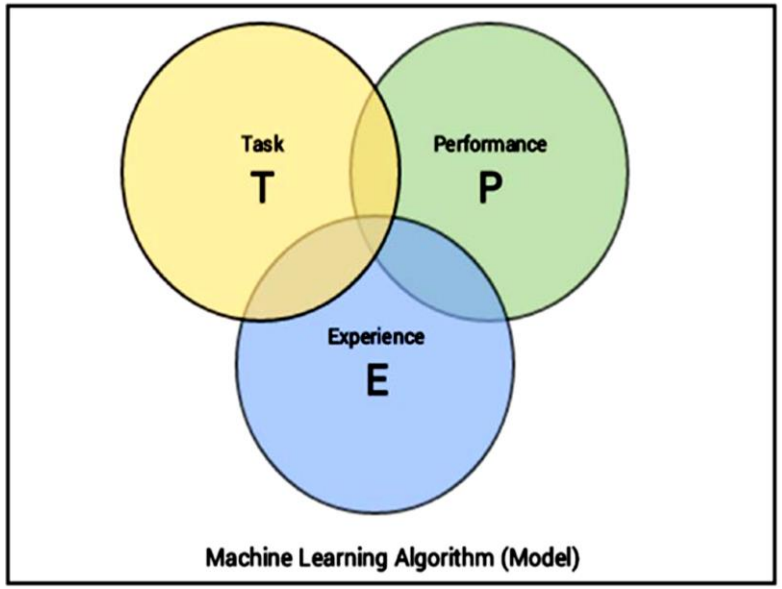
- Learning-based agent.
- It can learn from past data and improve automatically.
- Machine learning can deal with huge amounts of data.



machine learning basics

Formal Definition

The idea of Machine Learning is that there will be some learning algorithm that will help the machine
learn from data. Professor **Mitchell** defined it as follows.
“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.”



Defining task T

Classification: classify animal images into dogs and cats.

Regression: predicting house price.

Anomaly detection: indication of fraud.

Structured annotation: text annotation like grammar, sentiment, named entities, image annotation like annotate specific areas of images.

Translation: natural language translate.

Clustering or grouping: group similar products, events, entities

Transcription: These tasks usually entail various representations of data that are usually continuous and unstructured and converting them into more structured and discrete data elements. Examples include speech to text, optical character recognition, images to text,

Defining experience E

The process of consuming a dataset that consists of data samples or data points such that a learning algorithm or model learns inherent patterns is defined as the experience, *E* which is gained by the learning algorithm.

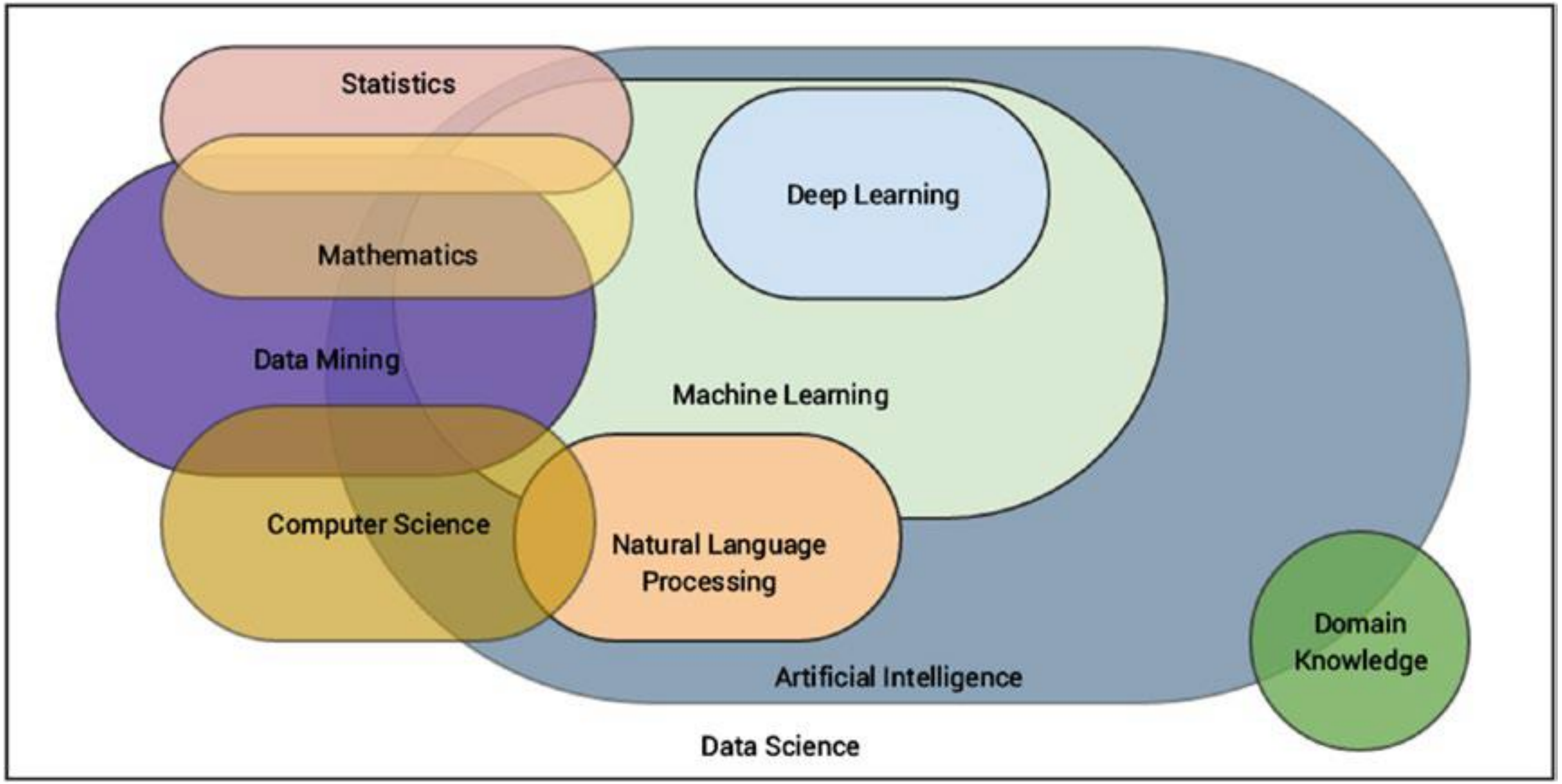
Defining performance P

performance measures include **accuracy, precision, recall, F1 score, sensitivity, specificity, error rate, misclassification rate**.

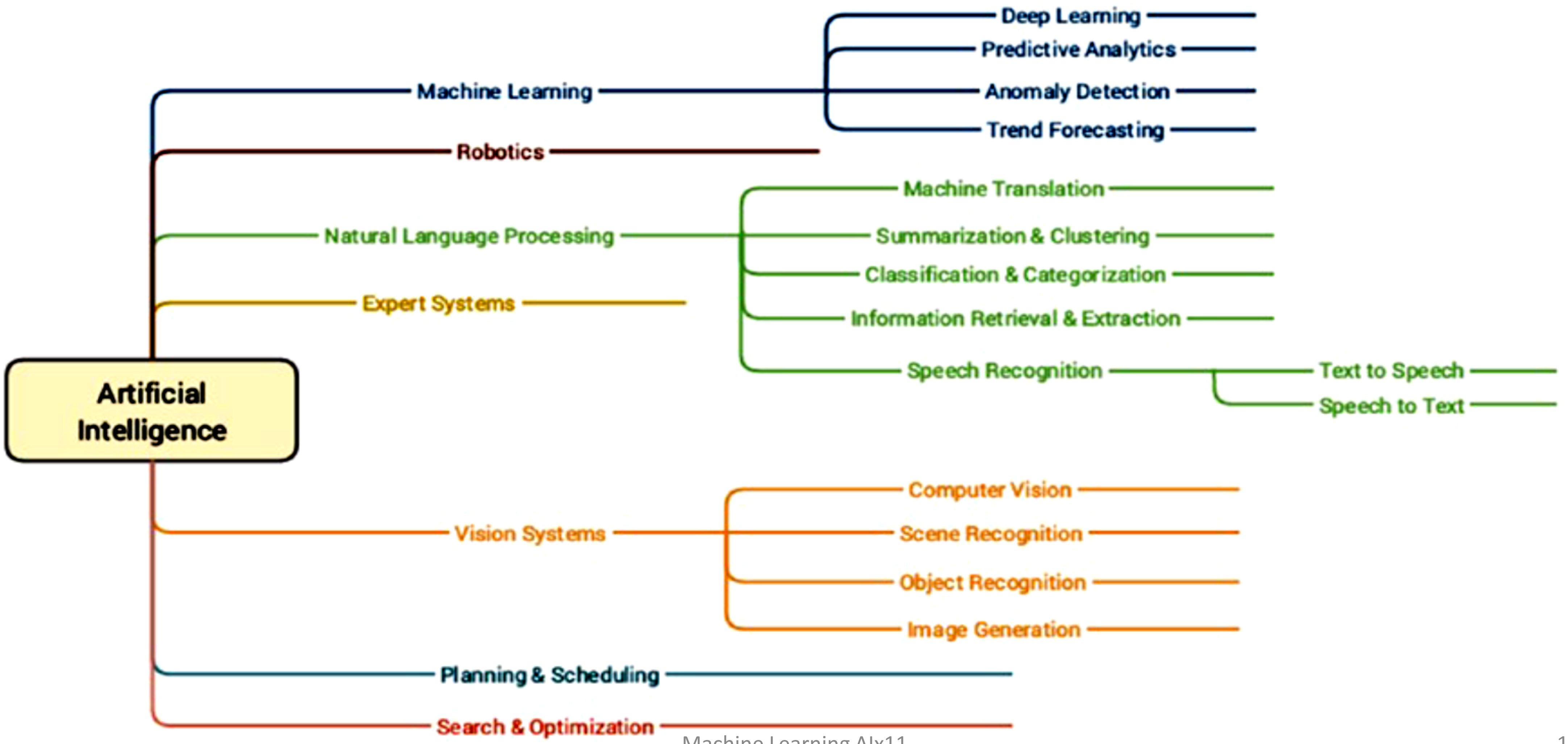
Performance measures are usually evaluated on training data samples (used by the algorithm to gain experience, E) as well as data samples which it has not seen or learned from before, which are usually known as validation and test data samples.

The idea behind this is to ***generalize*** the algorithm so that it doesn't become too biased only on the

A Multi-Disciplinary Field



major facets under the AI umbrella



Machine learning successes

A survey of recent success stories includes several prominent 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

Machine learning challenges

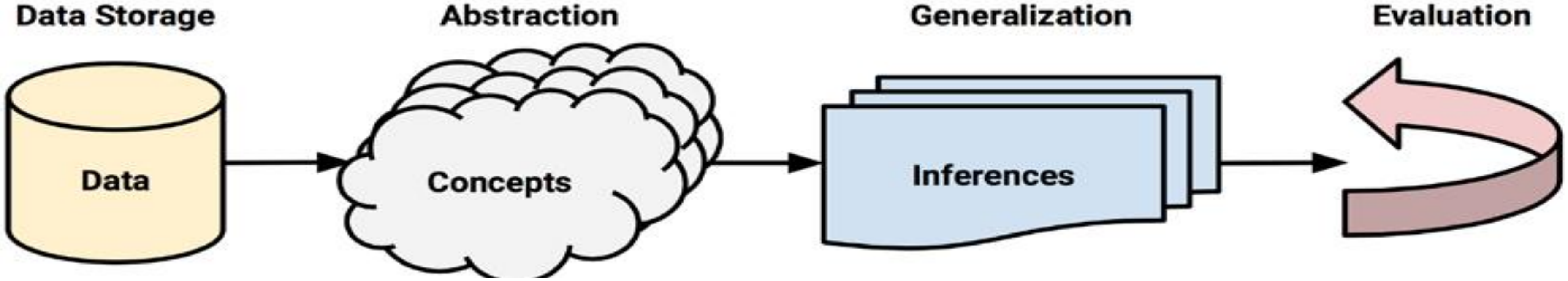
- **Data quality** issues lead to problems, especially with regard to data processing and feature extraction.
- **Data acquisition**, extraction, and retrieval is an extremely tedious and time-consuming process.
- Lack of good quality and sufficient **training data** in many scenarios.
- Formulating business problems clearly with well-defined **goals** and objectives.
- **Feature extraction** and engineering, especially hand-crafting features, is one of the most difficult yet important tasks in Machine Learning.
- **Overfitting** or underfitting models can lead to the model learning poor representations and relationships from the training data leading to detrimental performance.
- Choice of correct **Statistical Model** that fits the data.
- **Complex models** can be difficult to deploy in the real world.

Machine learning limitations

- It has very **little flexibility** to extrapolate outside of the **strict parameters** it learned and knows no common sense.
- Should be extremely careful to **recognize** exactly what the **algorithm** has learned before setting it loose in real-world settings.
- Without a lifetime of **past experiences** to build upon, computers are also limited in their ability to make simple common sense inferences about logical next steps.

How machines learn

- **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.



Data Storage

All learning must begin with **data.**

Humans and computers alike utilize **data storage** as a foundation for more advanced reasoning.

In a **human** being, this consists of a **brain** that uses electrochemical signals in a network of biological cells to store and process observations for short- and long-term future recall.

Computers have similar capabilities of short- and long-term recall using hard disk drives, flash memory, and random-access memory (RAM) in combination with a central processing unit (CPU).

Abstraction

- This work of assigning meaning to stored data occurs during the **abstraction** process, in which raw data comes to have a more abstract meaning. This type of connection, say between an object and its representation
- 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. You may be already familiar with some. Examples include:
    - Mathematical equations
    - Relational diagrams such as trees and graphs
    - Logical if/else rules
    - Groupings of data known as clusters

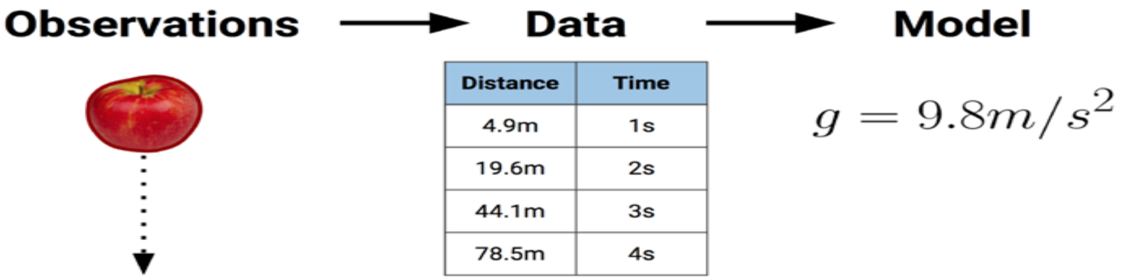
Abstraction

- 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.

➢Note, You might wonder why this step is called training rather than learning.

First, note that the process of learning does not end with data abstraction; the learner must still generalize and evaluate its training.

Second, the word training better connotes the fact that the human teacher trains the machine student to understand the data in a specific way.

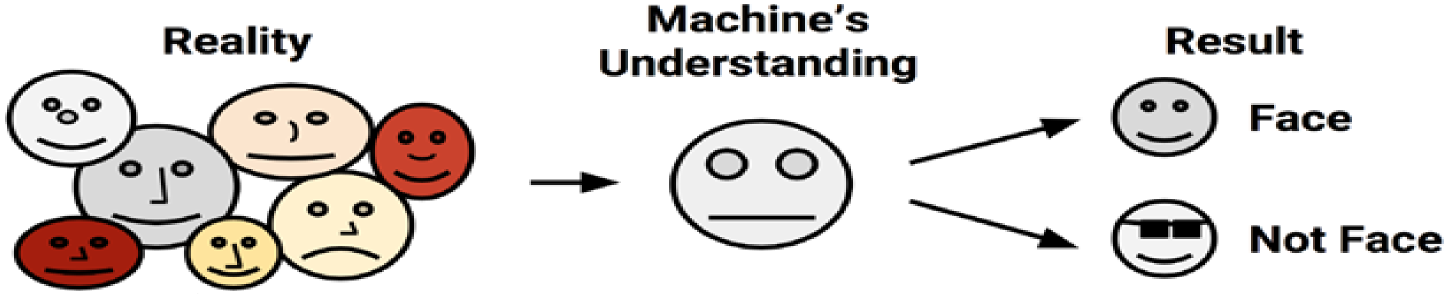


Distance	Time
4.9m	1s
19.6m	2s
44.1m	3s
78.5m	4s

$$g = 9.8m/s^2$$

generalization

- The term **generalization** describes the process of turning abstracted knowledge into a form that can be utilized for future action, on tasks that are similar, but not identical, to those it has seen before.
- In generalization, the learner is tasked with limiting the patterns it discovers to only those that will be most relevant to its future tasks.
- The algorithm is said to have a **bias** if the conclusions are systematically erroneous, or wrong in a predictable manner



Evaluation

- **Bias** is a necessary evil associated with the **abstraction** and **generalization** processes inherent in any learning task.
- 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** to judge how well its characterization of the training data generalizes to new, unseen data. It’s worth noting that it is exceedingly rare for a model to generalize to every unforeseen case perfectly.

Evaluation

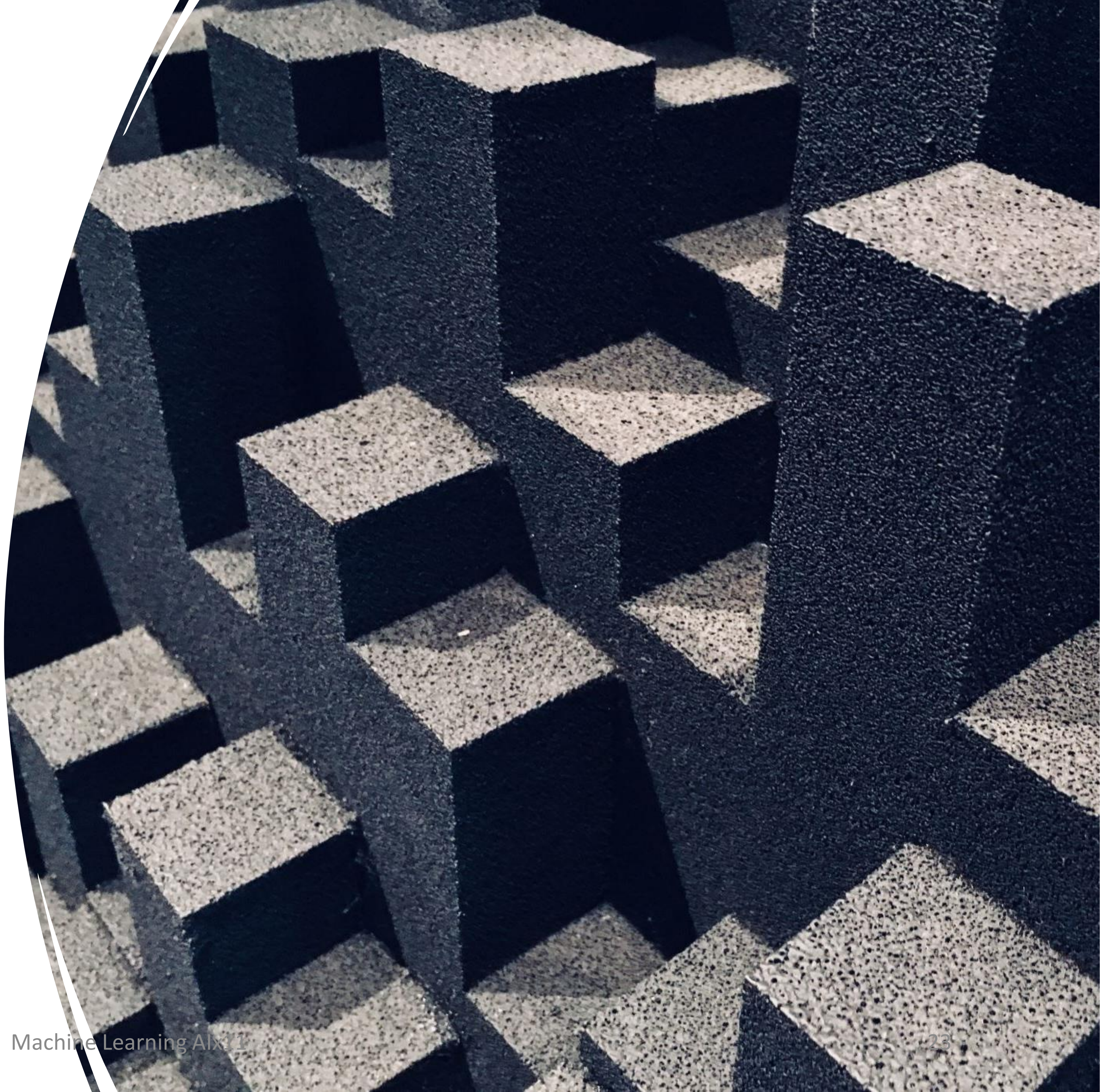
In parts, models fail to perfectly generalize due to the problem of **noise**, Noisy data is caused by seemingly random events, such as:

- • Measurement error due to **imprecise sensors** that sometimes add or subtract a bit from the readings.
- • Issues with human subjects, such as survey **respondents** reporting random answers to survey questions, in order to finish more quickly.
- • **Data quality problems**, including missing, null, truncated, incorrectly coded, or corrupted values.
- • Phenomena that are so **complex** or so little understood that they impact the data in ways that appear to be unsystematic.

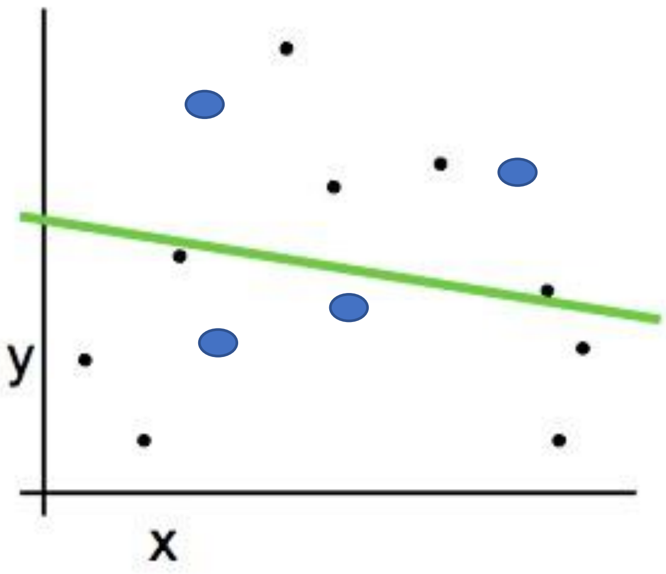
Trying to model noise is the basis of a problem called **overfitting**.

Underfitting & overfitting

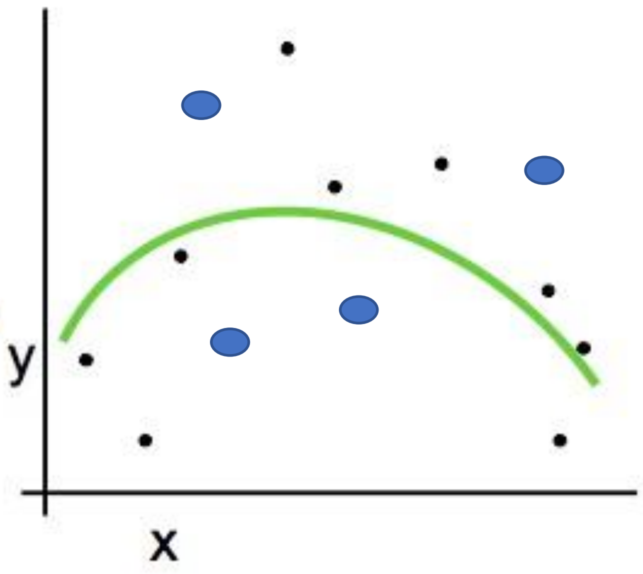
- Underfitting & overfitting



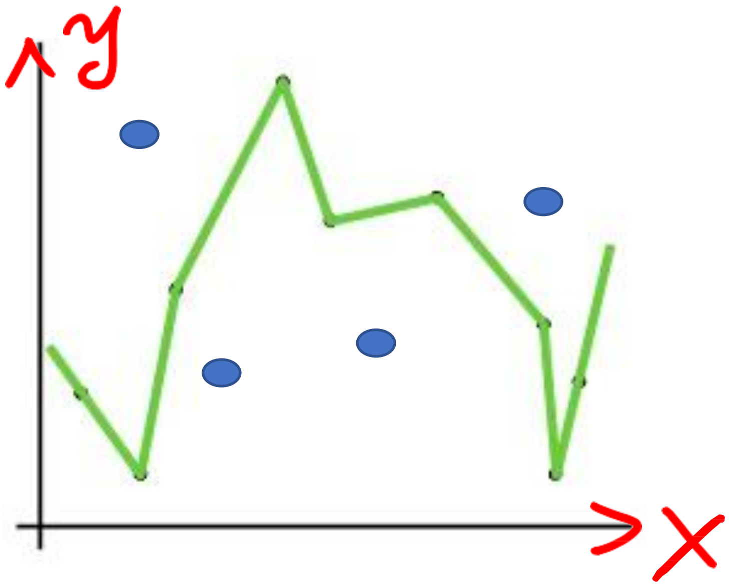
Underfitting & overfitting



Training = 50%
Test = 48%
Bias = 50% (high bias)
Variance= 2 (low var.)
Underfitting model.



Training = 98%
Test = 92%
Bias = 2% (low bias)
Variance= 6 (low var.)
fitting model.

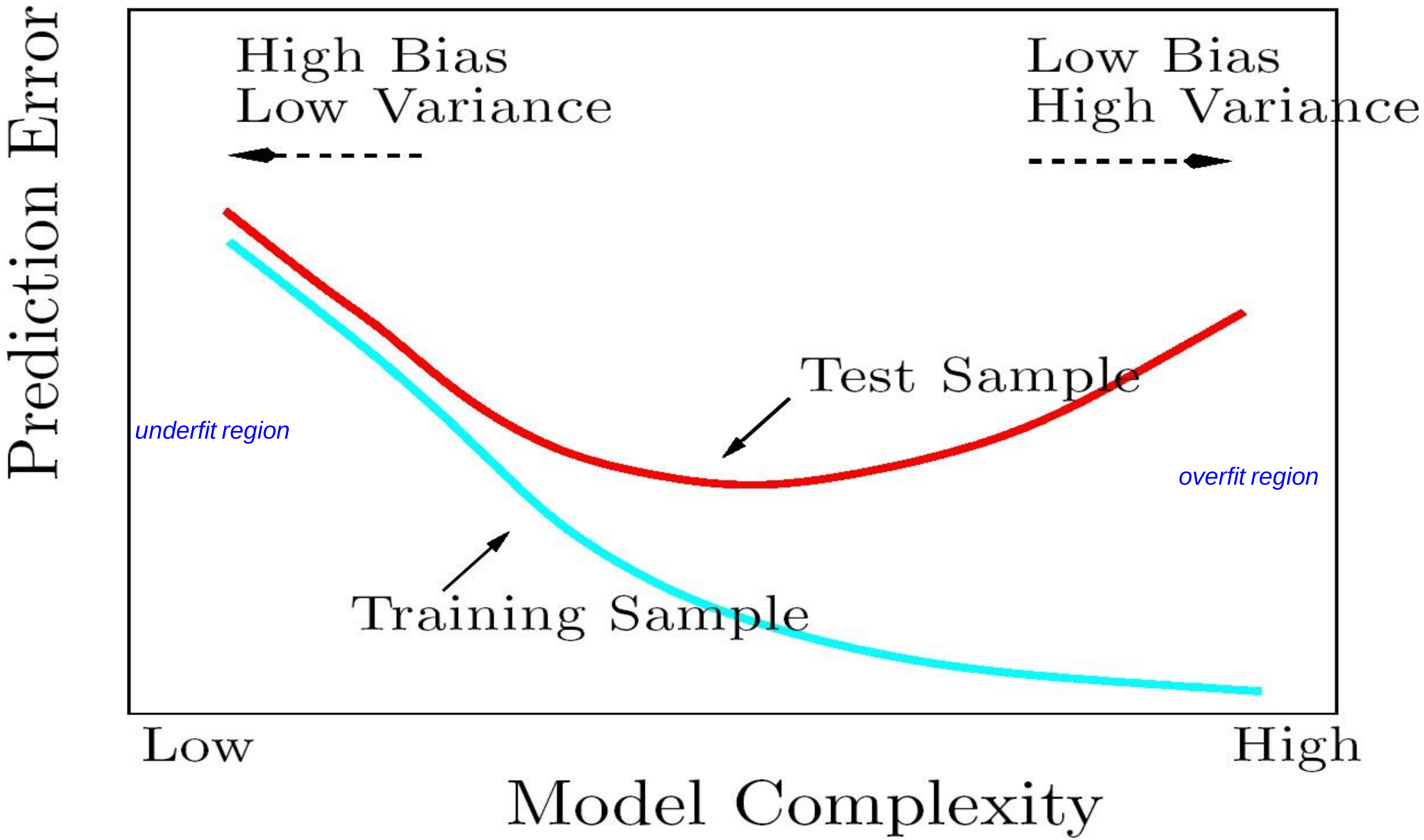


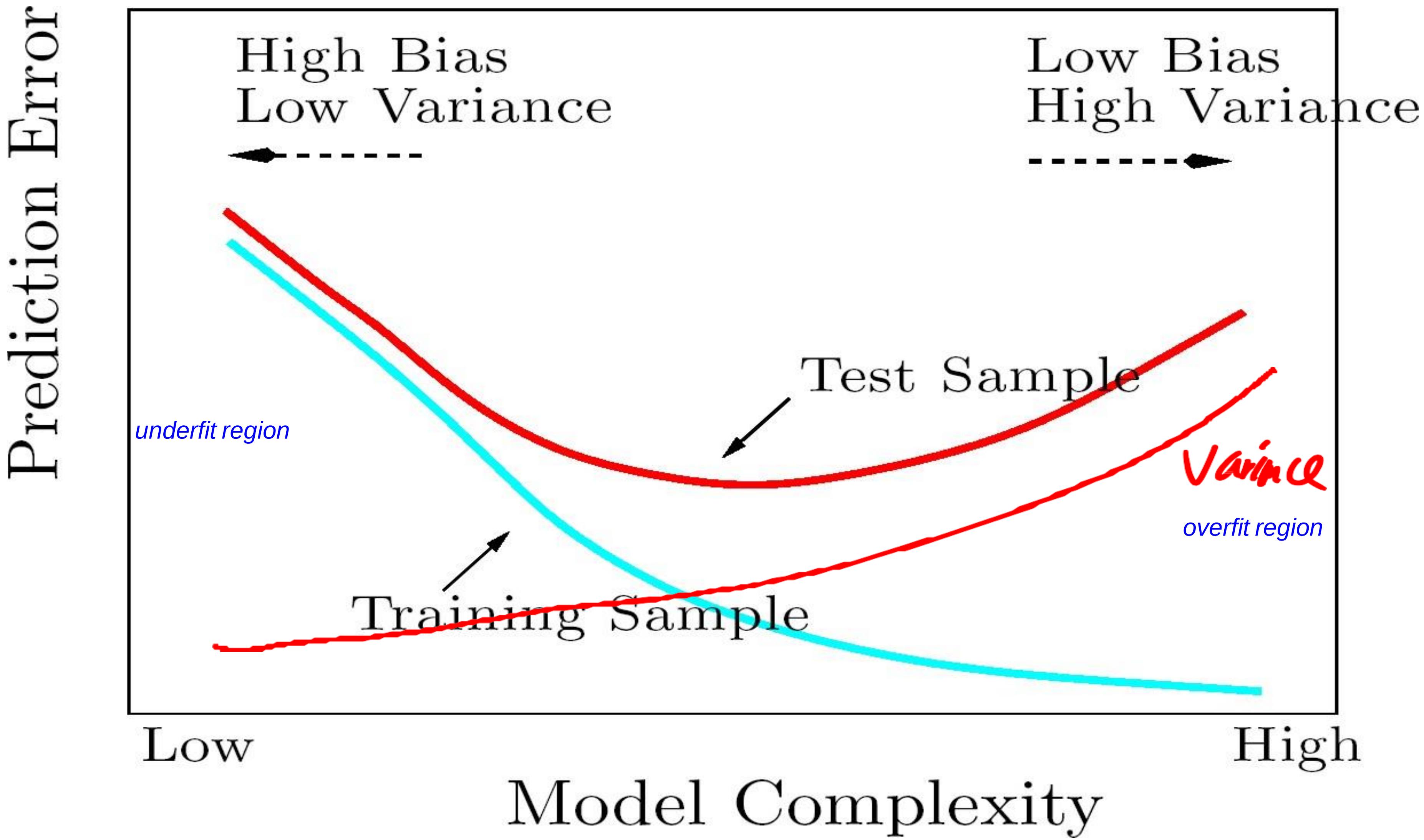
Training = 99%
Test = 60%
Bias = 1% (low bias)
Variance= 39 (high var.)
overfitting model.

Bias & variance

- Bias $E[(\bar{\theta} - \theta)^2]$
  - measures accuracy or quality of the model
  - low bias implies on average we will accurately estimate true parameter from training data

- Variance $E[(\hat{\theta} - \bar{\theta})^2]$
  - Measures precision or specificity of the model
  - Low variance implies the model does not change much as the training set varies


Prediction Error
High Bias
Low Variance
Low Bias
High Variance
Test Sample
Training Sample
underfit region
overfit region
Low
High
Model Complexity

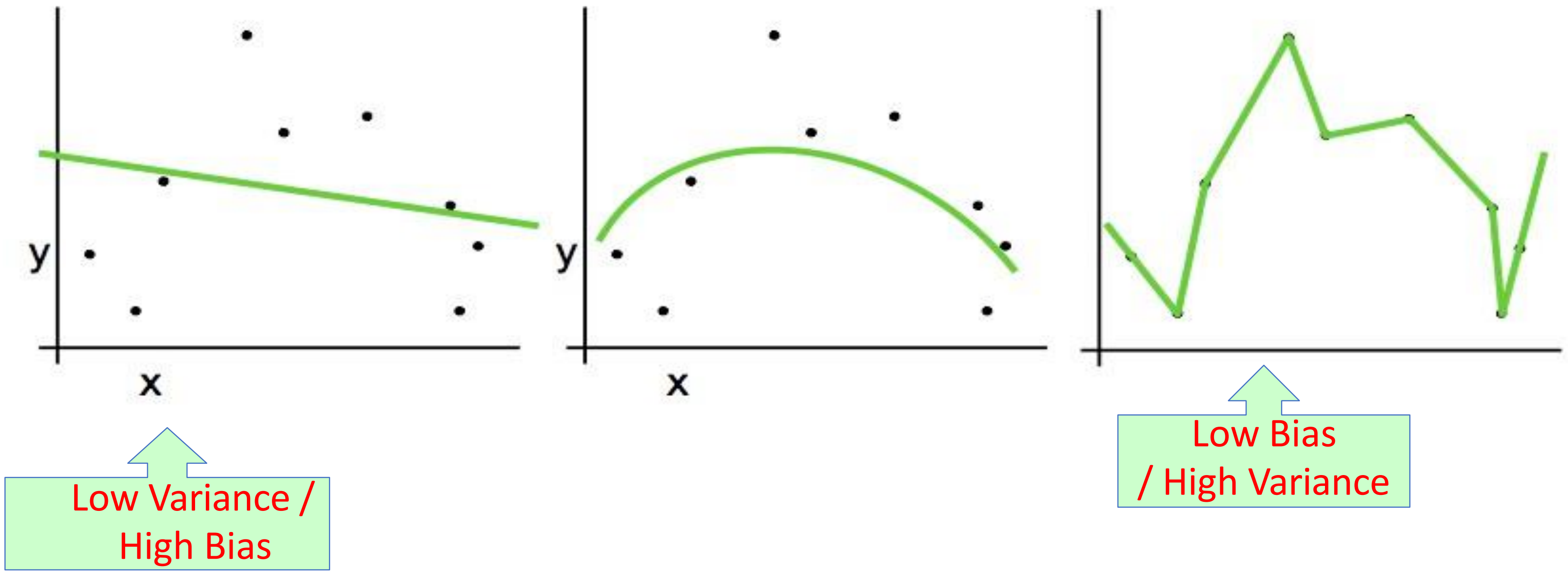

High Bias
Low Variance
Low Bias
High Variance
Test Sample
Training Sample
underfit region
Varince
overfit region
Prediction Error
Low
High
Model Complexity

Underfitting & overfitting

- Models with **too few parameters** are inaccurate because of a **large bias** (not enough flexibility).

- Models with **too many parameters** are inaccurate because of a **large variance** (too much sensitivity to the sample randomness).

Regression: Complexity versus Goodness of Fit



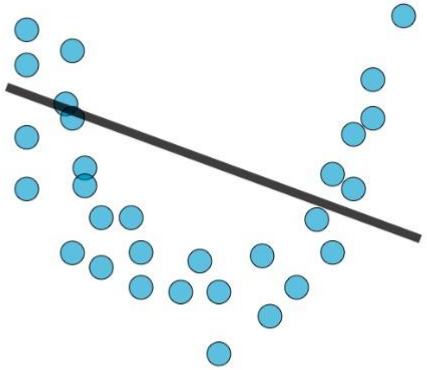
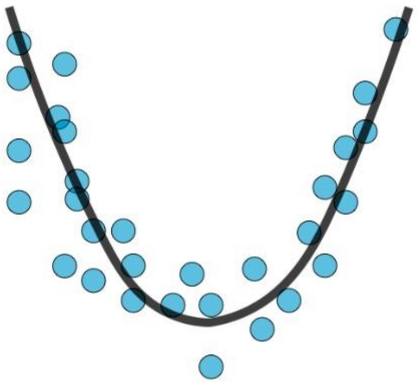
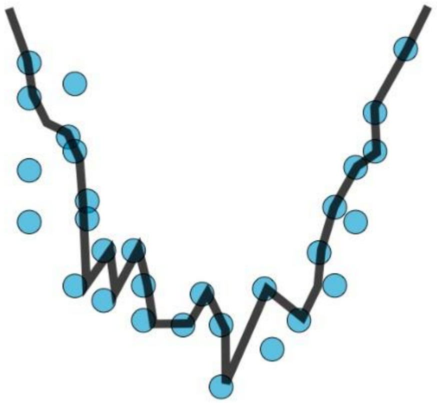
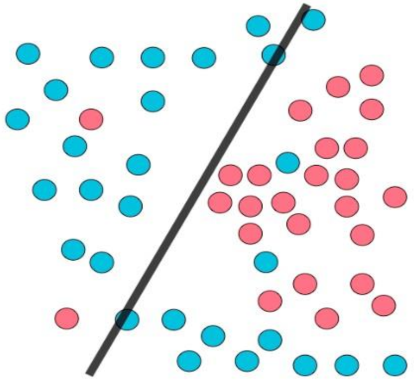
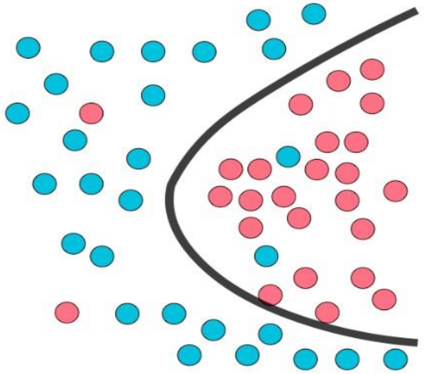
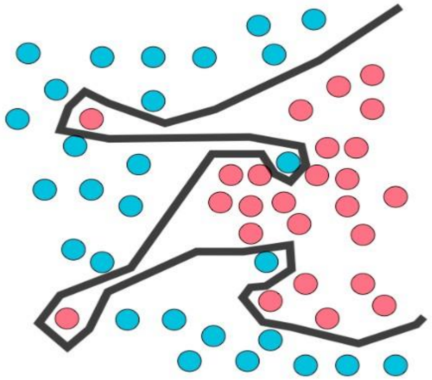
Low Variance /
High Bias

Low Bias
/ High Variance

Highest Bias
Lowest variance
Model complexity = low

low Bias
low Variance
Model complexity = medium

low Bias
High variance
Model complexity = high

	Underfitting	Just right	Overfitting
Symptoms	- High training error - Training error close to test error - High bias	- Training error slightly lower than test error	- Low training error - Training error much lower than test error - High variance
Regression			
Classification			
Remedies	- Complexify model - Add more features - Train longer		- Regularize - Get more data - select feature

Fixes to try:

- Try getting more training examples → fixes high variance.
- Try a smaller set of features → fixes high variance.
- Try a larger set of features → fixes high bias.

Machine Learning Methods

Machine Learning Methods

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

a. Supervised learning (predictive models)

b. Unsupervised learning (descriptive models)

c. Semi-supervised learning

d. Reinforcement learning

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

a. Batch learning

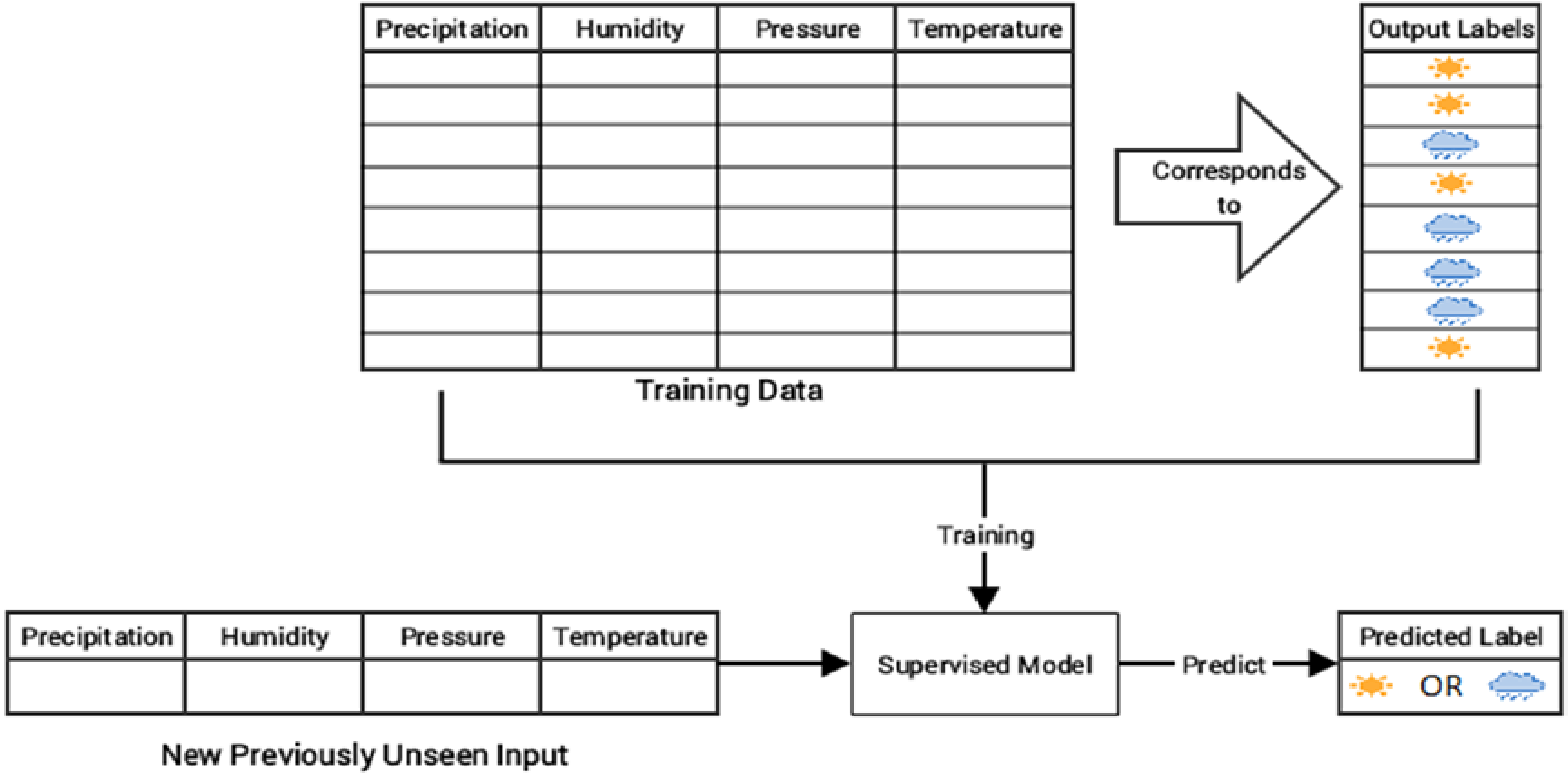
b. Online learning

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

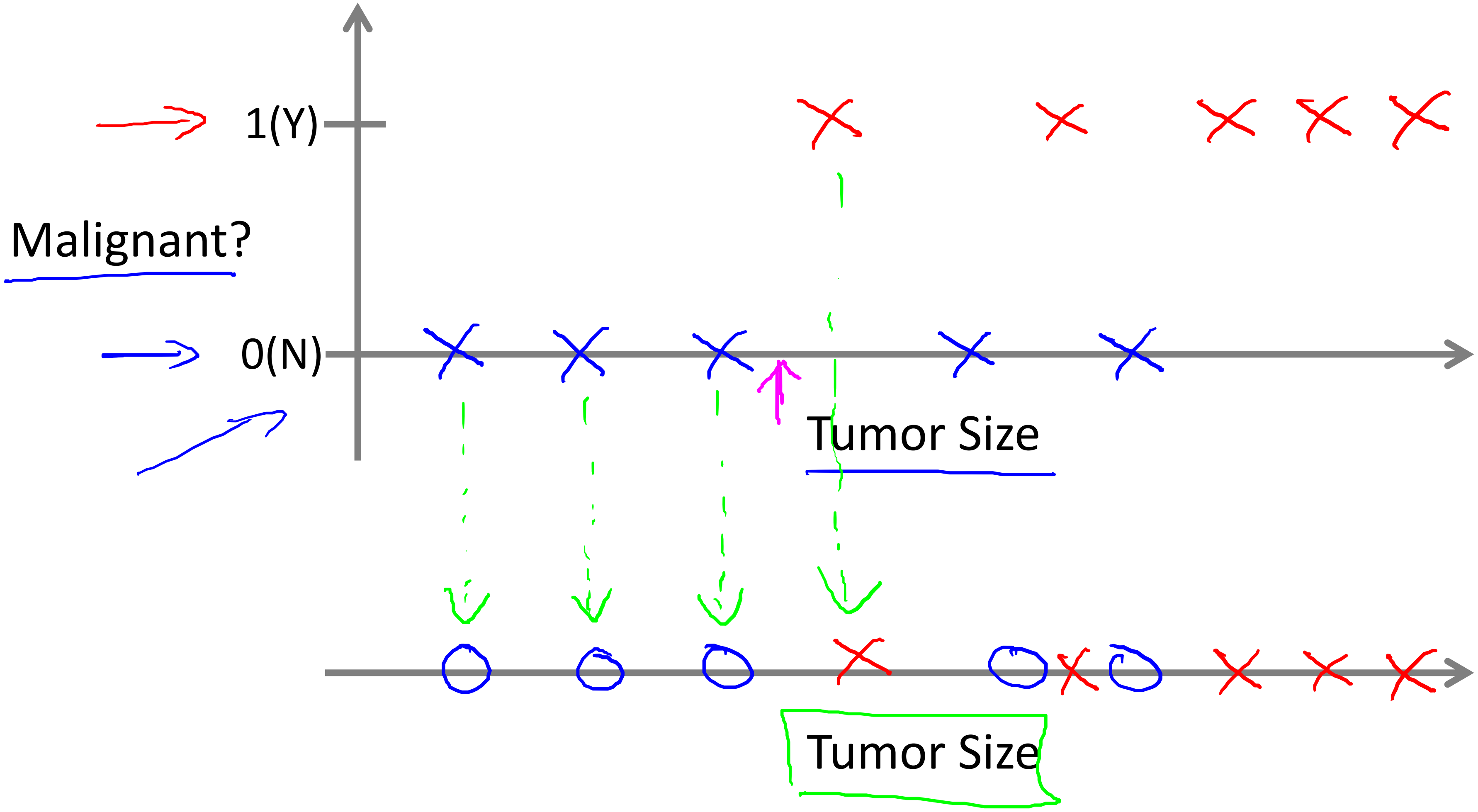
a. Instance based learning

b. Model based learning

a. Supervised learning (classification)



Breast cancer (malignant, benign)



1(Y)

Malignant?

0(N)

Tumor Size

Tumor Size

Classification

Discrete valued output (0 or 1)

0, 1, 2, 3

benign

type 1 cancer

Supervised Learning(Classification)

Classification aims to identify group membership.

Input: $\{x_1, x_2, \ldots, x_n\}$ *categorical values*, called ***features***

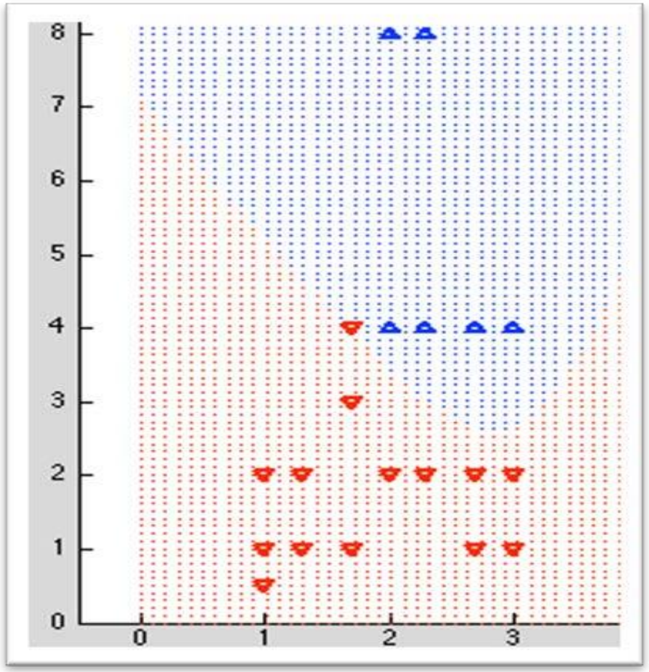
Output: *y categorical values*, called ***Target Value***

Ex.: data about computers (training data) Find a model to predict status of unseen cases

features *Target Value*

Processor (GHz)	Memory (GB)	Status
1.0	1.0	Bad
2.3	4.0	Good
2.6	4.0	Good
3.0	8.0	Good
2.0	4.0	Bad
2.6	0.5	Bad

➔ 3.0 4.0 ???



Prediction tasks in Supervised Learning

Binary classification (e.g., email ⇒ spam/not spam):

classification: the label is a discrete variable

- e.g., the task of predicting the types of residence

$$x \longrightarrow f \longrightarrow y \in \{0,1\}$$

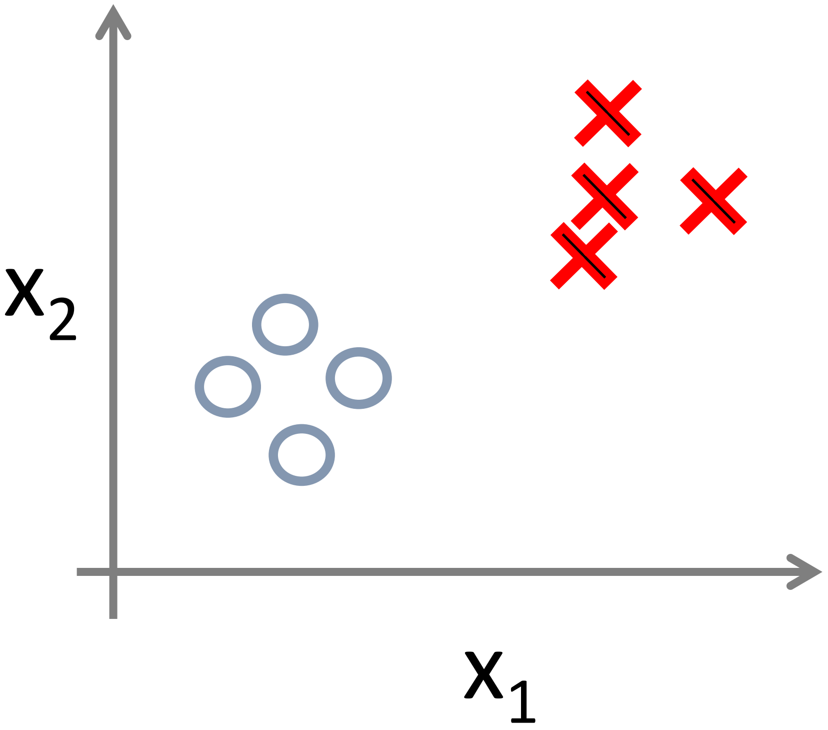
Regression (e.g., location, year ⇒ housing price):

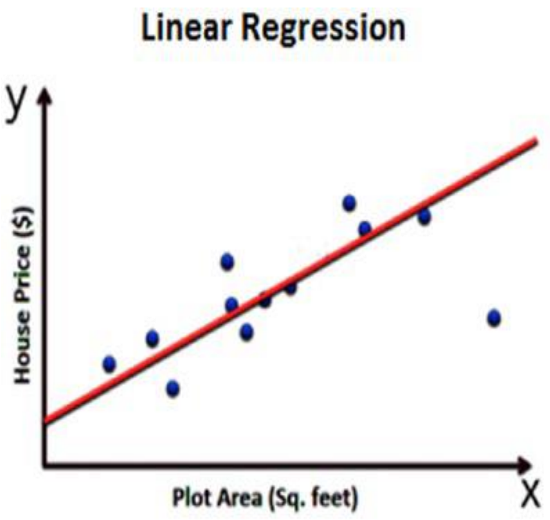
regression: if y is a continuous variable

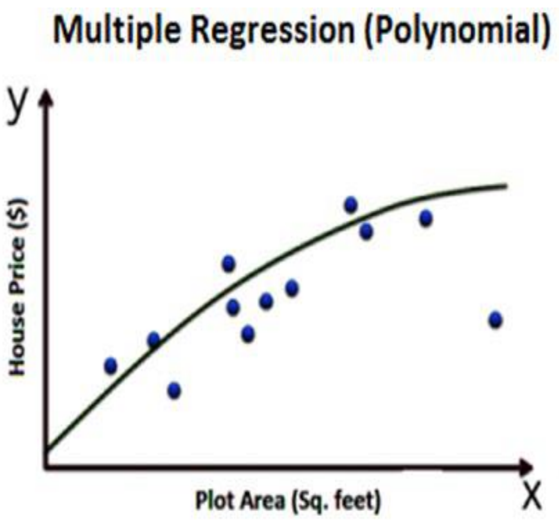
- e.g., price prediction

$$x \longrightarrow f \longrightarrow y \in \mathrm{R}$$

Supervised Learning







b- Unsupervised Learning

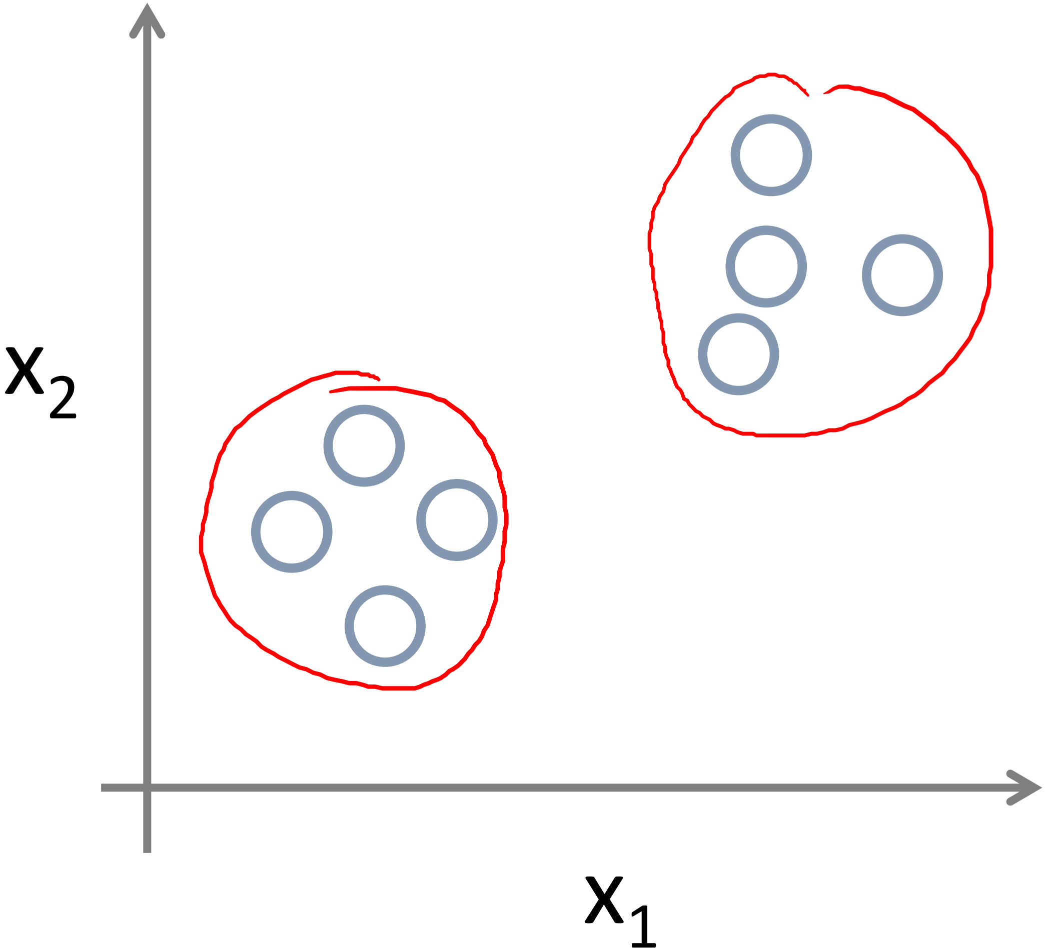
Unsupervised learning is a learning method in which a machine learns without any supervision from data that has no labels.

- The algorithm needs to act on the data without any supervision.
- The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.
- In unsupervised learning, **we don't have a predetermined result.** The machine tries to find useful insights from a huge amount of data.

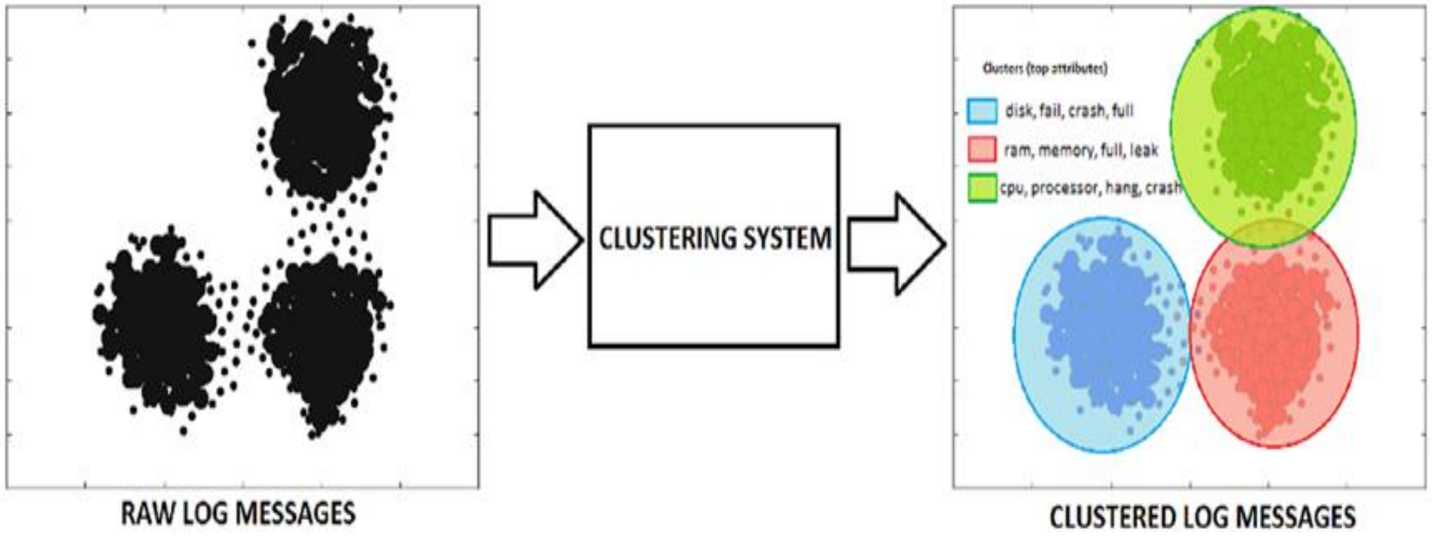
Unsupervised Learning

- Training data contain only the input vectors (No labeled data)
- Definition of training data: $\{x_1, x_2,\ldots, x_n\}$
- Goal: Learn some structures in the inputs.
- Can be divided into two categories: Clustering and Dimensionality Reduction

Unsupervised Learning



unsupervised learning (clustering)

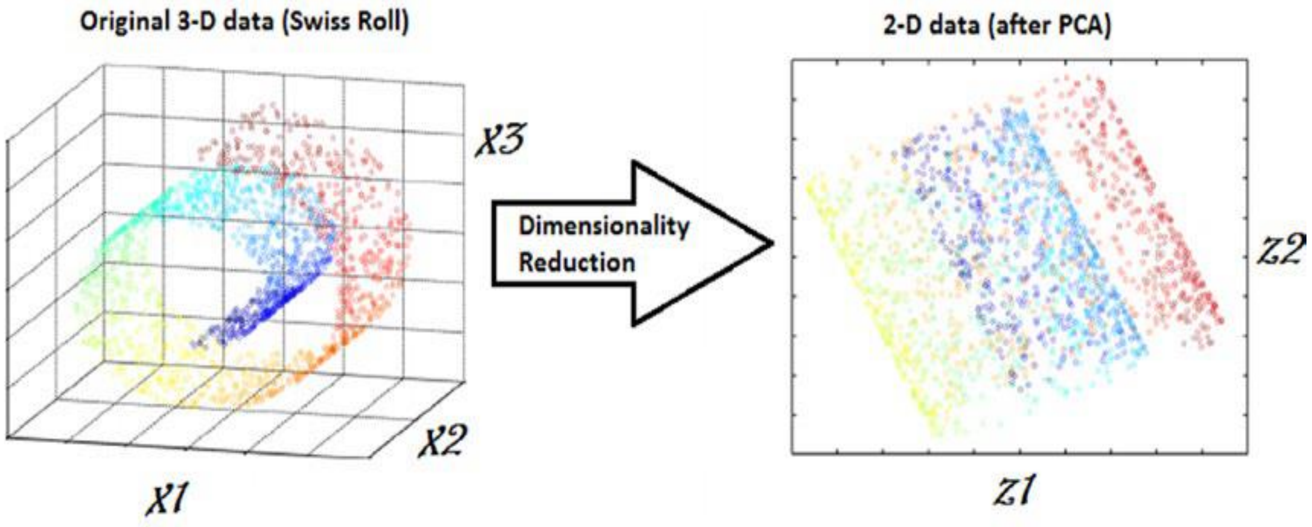

Clusters (top attributes)
🟦 disk, fail, crash, full
🟥 ram, memory, full, leak
🟩 cpu, processor, hang, crash
RAW LOG MESSAGES
CLUSTERING SYSTEM
CLUSTERED LOG MESSAGES

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

- Centroid based methods such as K-means
- Hierarchical clustering methods
- Distribution based clustering methods such as Gaussian mixture models
- Density based methods such as dbscan and optics.

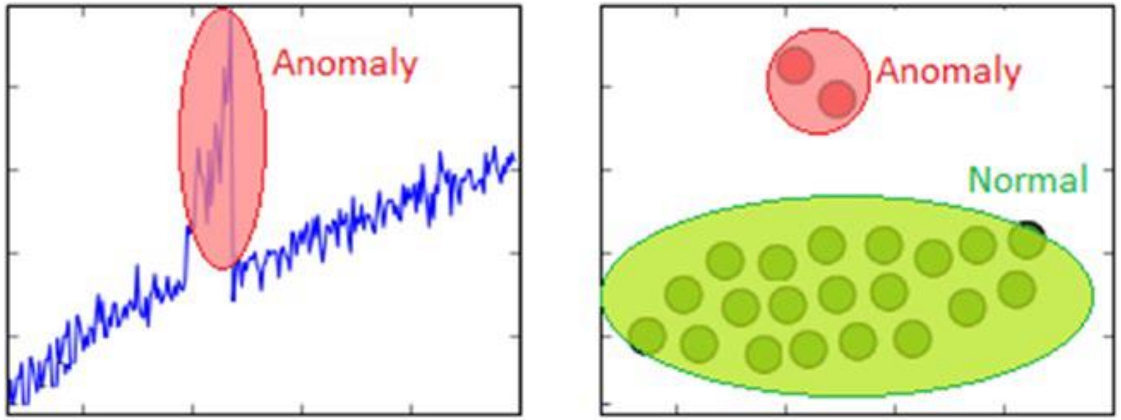
unsupervised learning (dimensionality reduction)

- **Feature selection methods**
- **Feature extraction methods**



- 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.

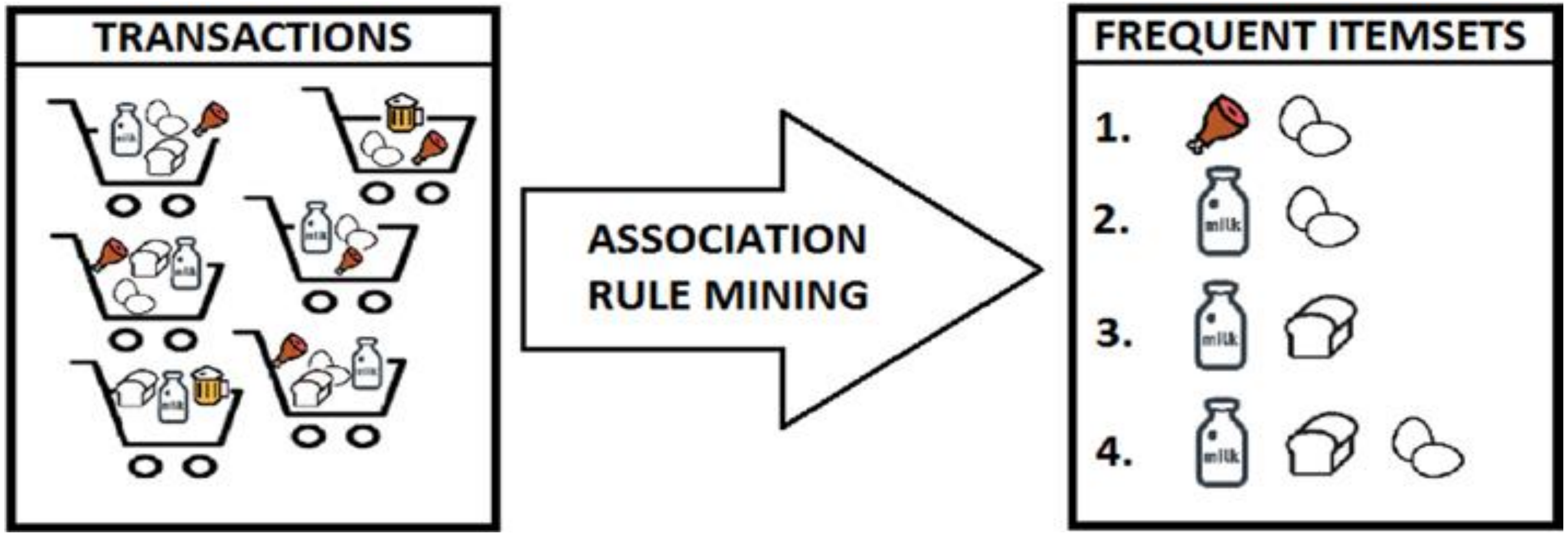
unsupervised learning (anomaly detection)



Pattern discovery

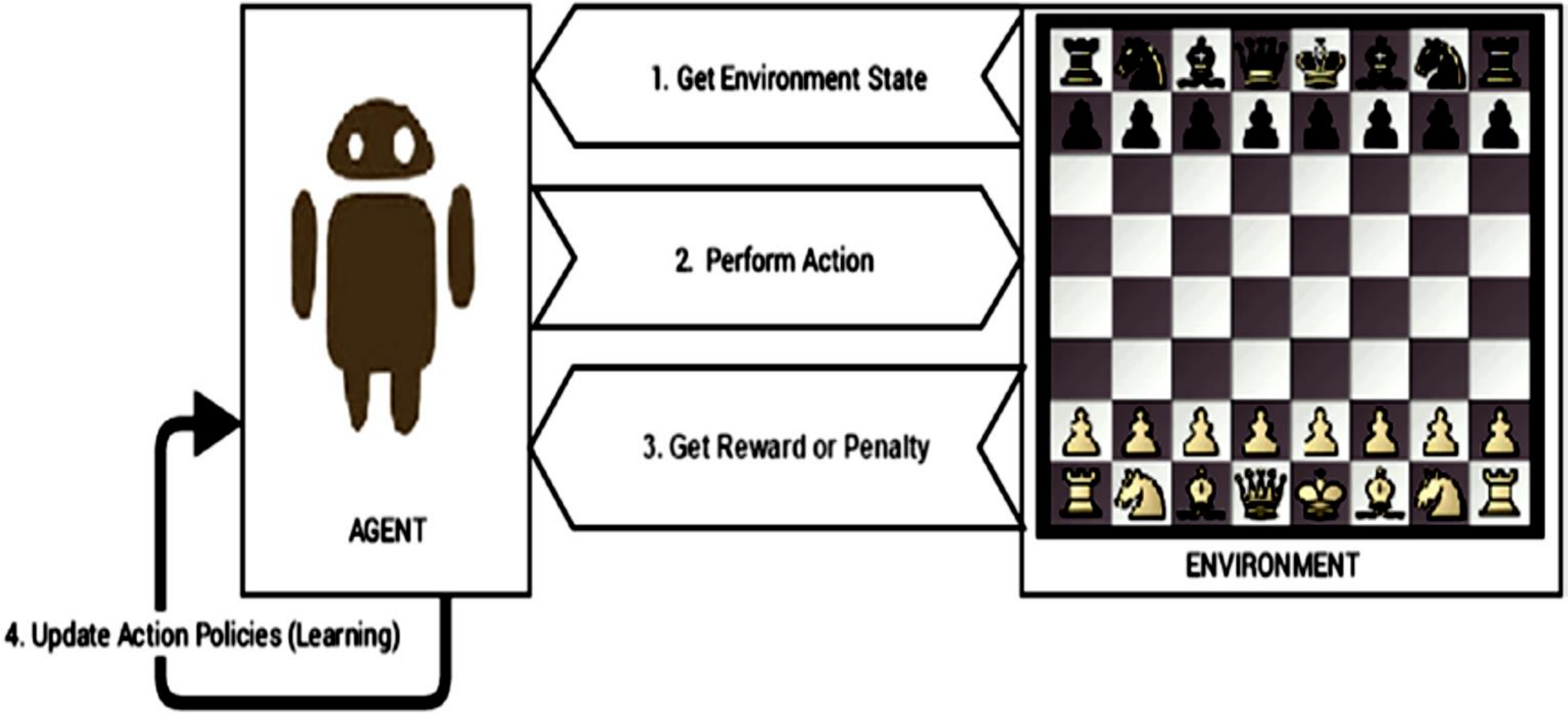
- Unsupervised learning methods can be used for anomaly detection such that we train the algorithm on the training dataset having normal, non-anomalous data samples.
- Once it learns the necessary data representations, patterns, and relations among attributes in normal samples, for any new data sample, it would be able to identify it as anomalous or a normal data point by using its learned knowledge.

unsupervised learning (Association Rule-Mining)



Association rules help in detecting and predicting transactional patterns based on the knowledge it gains from training transactions.

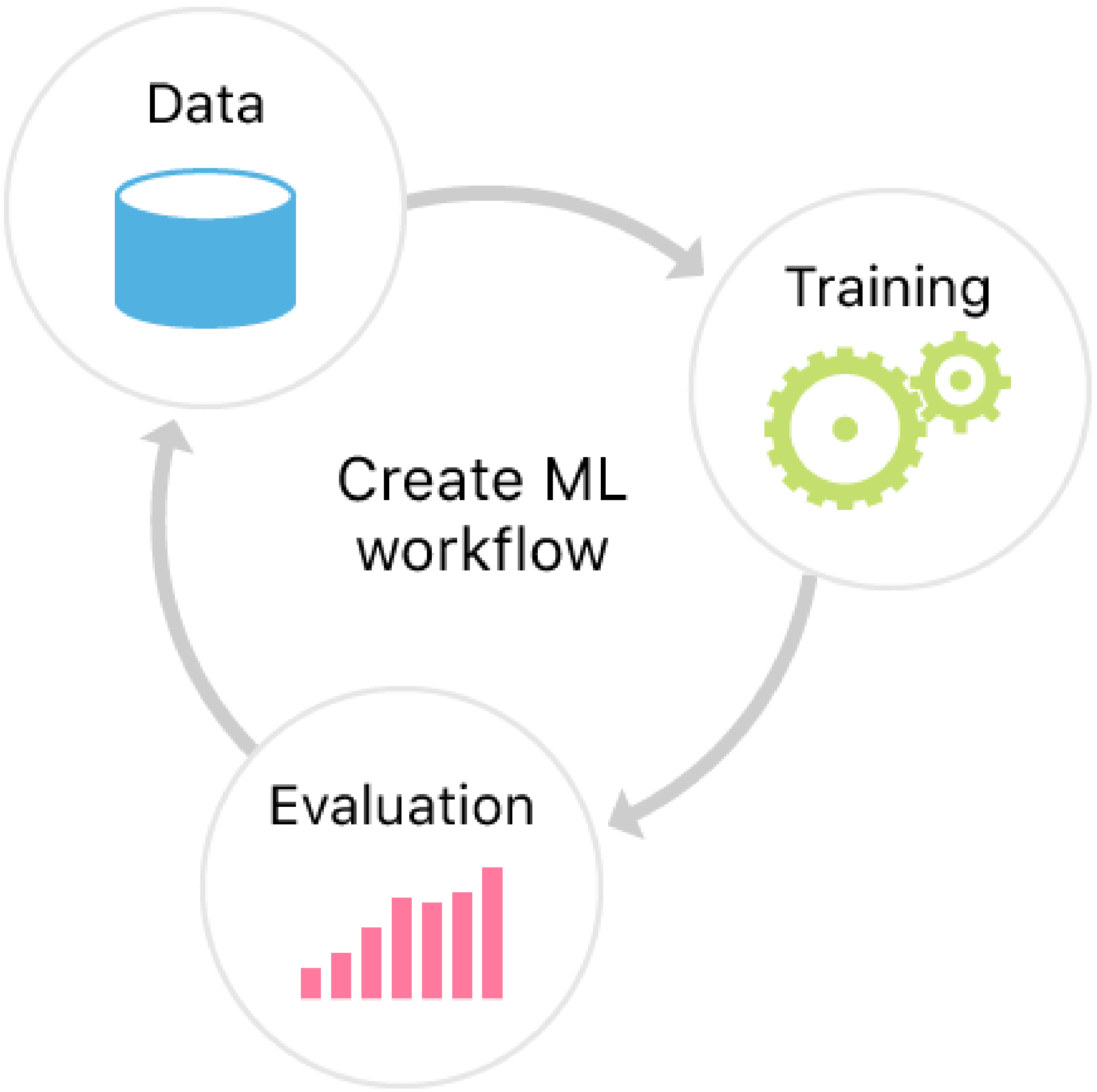
d. Reinforcement learning



Instance-Based learning vs model-Based learning

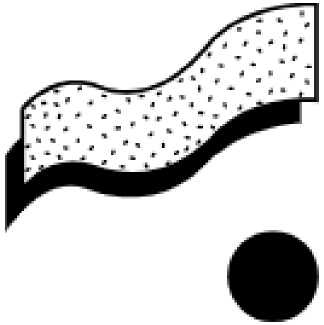
instance-based learning	model-based learning
The instance-based learning works by looking at the input data points and using a similarity metric to generalize and predict for new data points. A simple example would be a K-nearest neighbor algorithm	The model-based learning methods are a more traditional ML approach toward generalizing based on training data. Typically an iterative process takes place where the input data is used to extract features and models are built based on various model parameters (known as *hyperparameters*). These hyperparameters are optimized based on various model validation techniques to select the model that generalizes best on the training data and some amount of validation and test data (split from the initial dataset).

Machine learning in practice



Machine learning in practice

Machine Learning AIx11



Types of input data

Types of learning algorithms

Matching data to algorithm.

Machine learning in practice

- To apply the learning process to real-world tasks, we'll use a life-cycle development project. any machine learning algorithm can be deployed by following these steps:

1- **data retrieval**

2- data preparation

3- modeling

4- model evaluation and tuning

5- deployment

Data retrieval

1- **Data collection**: collect all the necessary data needed for your business objective. Such as: historical data warehouses, data marts, data lakes and so on.

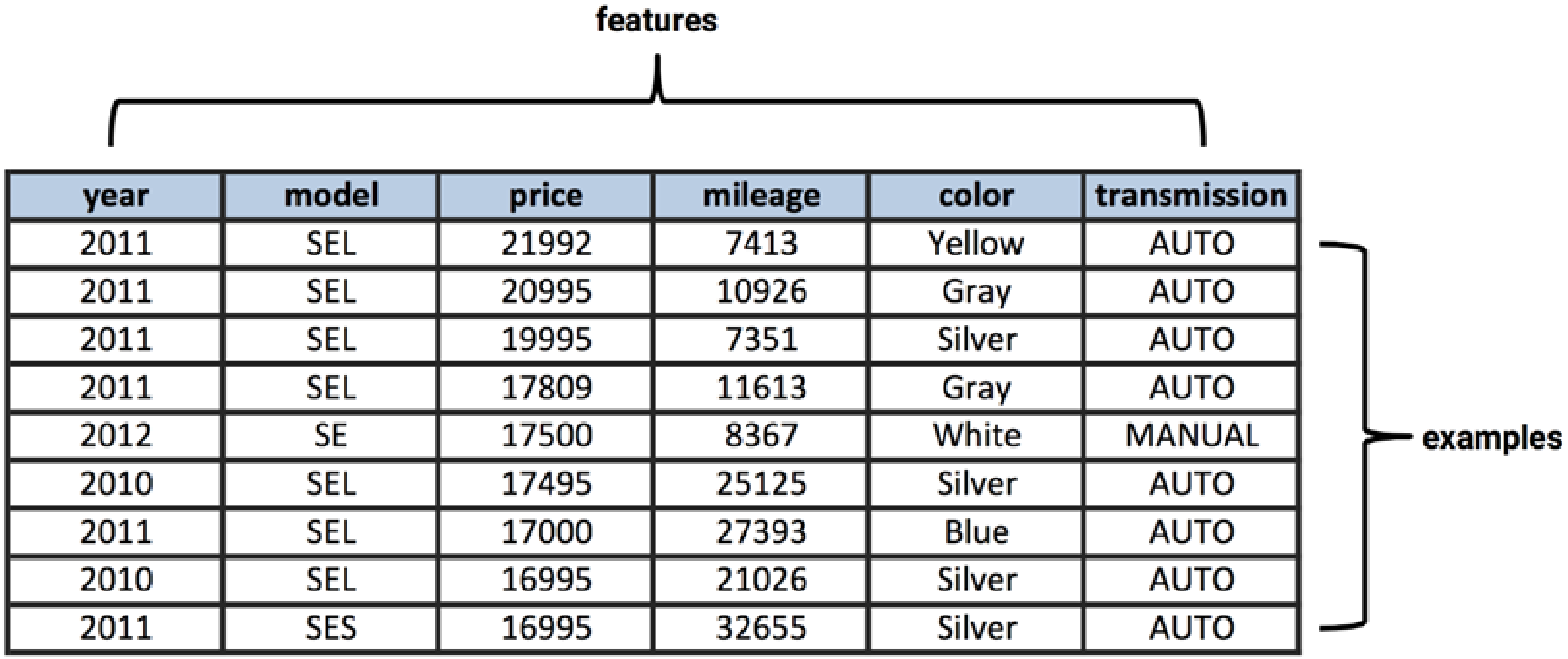
2- **Data description**: Analysis the data to understand the nature of data

such as: Data sources (SQL, NoSQL, Big Data), Data volume (size, number of records, total databases, tables), Data attributes and their description (variables, data types), Relationship and mapping schemes (understand attribute representations), Basic descriptive statistics (mean, median, variance), focuse on most important attributes.

3- **Exploratory data analysis**: Explore, describe, and visualize data attributes

4- **Data quality analysis**: missing values, inconsistent values, wrong information and metadata.

Data retrieval


features
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

Data retrieval

Types of input data: Features also come in various forms.

- If a feature represents a characteristic measured in numbers, it is unsurprisingly called **numeric**.
- Alternatively, 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; or a measurement of customer satisfaction on a scale from "not at all happy" to "very happy."
- It is important to consider what the features represent, as the type and number of features in your dataset will assist in determining an appropriate machine learning algorithm for your task.

Data preparation

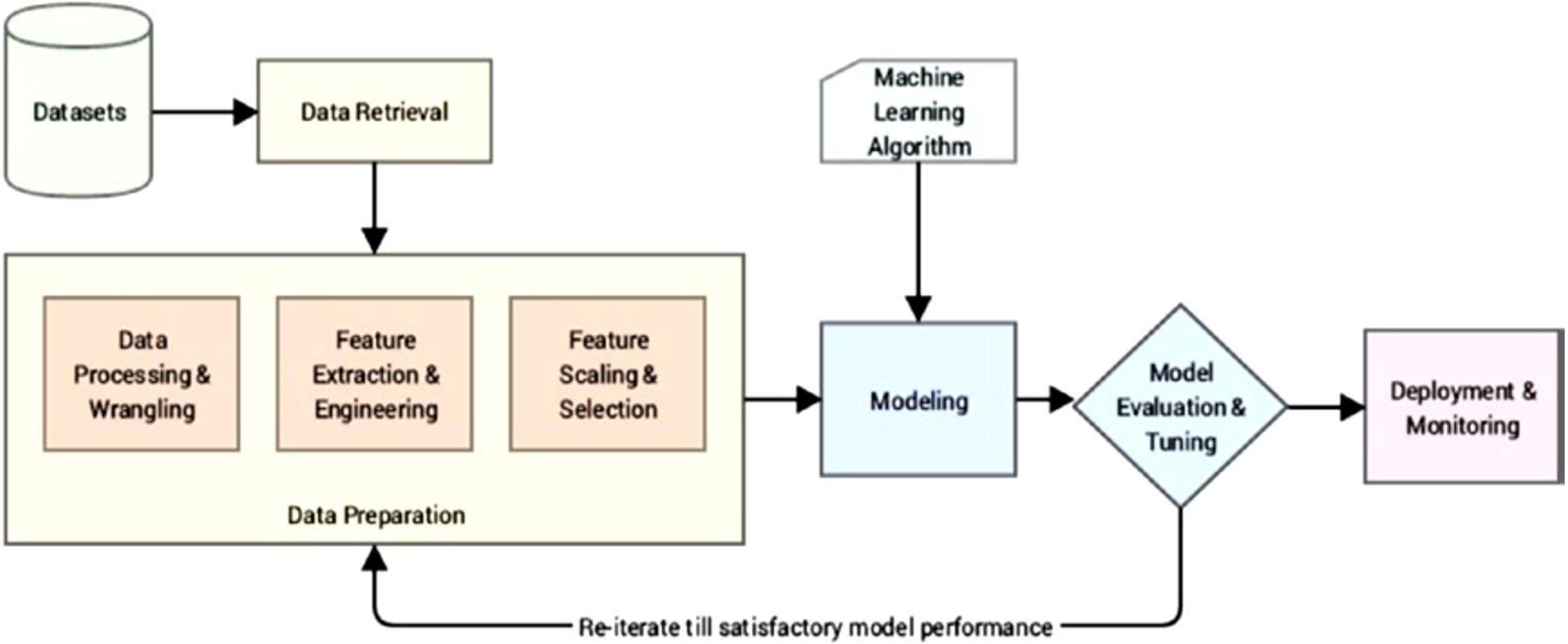
1. **Data integration**: is mainly done when we have multiple datasets that we might want to integrate or merge.
2. **Data wrangling**: data pre-processing, cleaning, normalization, and formatting.
3. **Attribute generation and selection**: is basically selecting a subset of features or attributes from the dataset based on parameters like attribute importance, quality, relevancy, assumptions, and constraints.

3-Modeling: In the process of modeling, we usually feed the data features to a Machine Learning method or algorithm and train the model, typically to optimize a specific cost function in most cases with the objective of reducing errors and generalizing the representations learned from the data.

4-Model evaluation and tuning: Built models are evaluated and tested on validation datasets and, based on metrics like accuracy, F1 score, and others, the model performance is evaluated. Models have various parameters that are tuned in a process called hyperparameter optimization to get models with the best and optimal results.

5- **Deployment and monitoring**: Selected models are deployed in production and are constantly monitored based on their predictions and results.

Standard machine learning pipeline


Datasets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

Machine learning in practice

- **Matching dataset to algorithms:**

Model	Learning task
Supervised learning algorithms	
Nearest neighbor	Classification
Naïve bayes	
Decision tree	
Classification rule learner	
Linear regression	Numeric prediction
Regression tree	
Model trees	
Neural network	Dual use
Support vector machine	

Machine learning in practice

- **Matching input data to algorithms:**

Model	Learning task
Unsupervised learning algorithms	
Association rules	Pattern detection
K-means clustering	Clustering

> “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.”

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.

> “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.”

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.

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?

○ Treat both as classification problems.

○ Treat problem 1 as a classification problem, problem 2 as a regression problem.

○ Treat problem 1 as a regression problem, problem 2 as a classification problem.

○ Treat both as regression problems.

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?

○ Treat both as classification problems.

○ Treat problem 1 as a classification problem, problem 2 as a regression problem.

● Treat problem 1 as a regression problem, problem 2 as a classification problem.

○ Treat both as regression problems.

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

○ Given email labeled as spam/not spam, learn a spam filter.

○ Given a set of news articles found on the web, group them into set of articles about the same story.

○ Given a database of customer data, automatically discover market segments and group customers into different market segments.

○ Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.

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

○ Given email labeled as spam/not spam, learn a spam filter.

● Given a set of news articles found on the web, group them into set of articles about the same story.

● Given a database of customer data, automatically discover market segments and group customers into different market segments.

○ Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.

The k-NN algorithm

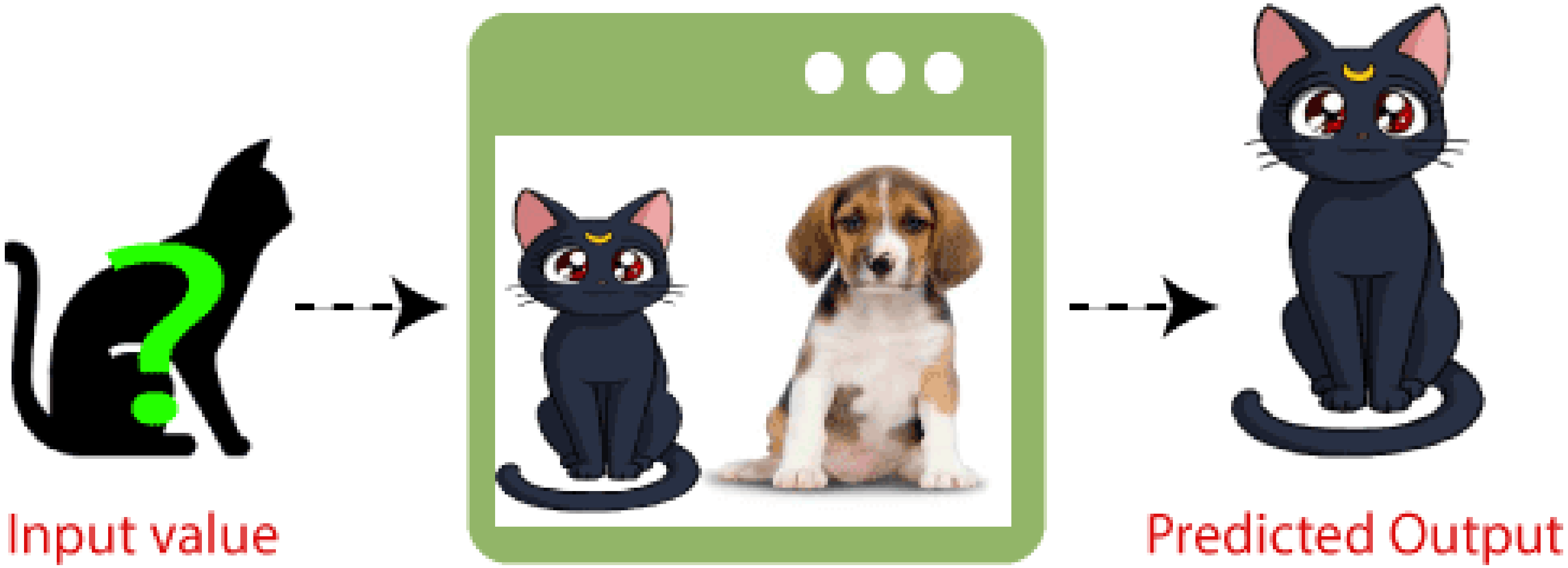
- The nearest neighbors approach to classification is exemplified by the **k-nearest neighbors algorithm** (**k-NN**).
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- The strengths and weaknesses of this algorithm are as follows:

Strengths	**Weaknesses**
• Simple and effective. • Makes no assumptions about the underlying data distribution. • Fast training phase.	• the model is limited ability to understand how the features are related to the class. • Requires selection of an appropriate *k.* • Slow classification phase. • Nominal features and missing data require additional processing.

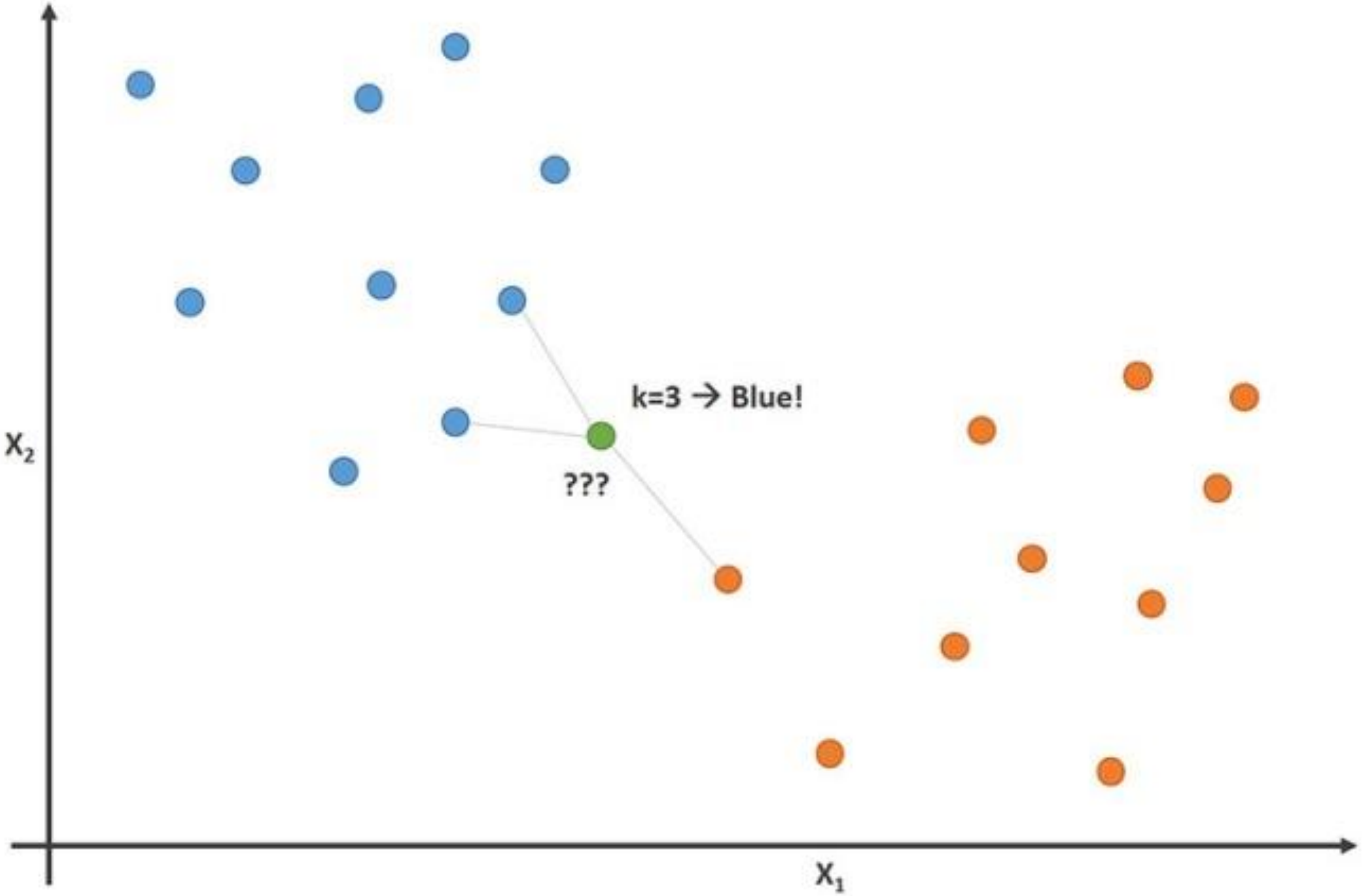
The k-NN algorithm

- The k-NN algorithm gets its name from the fact that it uses information about an example's k-nearest neighbors to classify unlabeled examples.
- The letter *k* is a variable term implying that any number of nearest neighbors could be used.
- After choosing *k*, the algorithm requires a training dataset of examples classified into several categories, as labeled by a nominal variable.
- Then, for each unlabeled record in the test dataset, k-NN identifies *k* records in the training data that are the "nearest" in similarity.
- The unlabeled test instance is assigned the class of most of the k nearest neighbors.

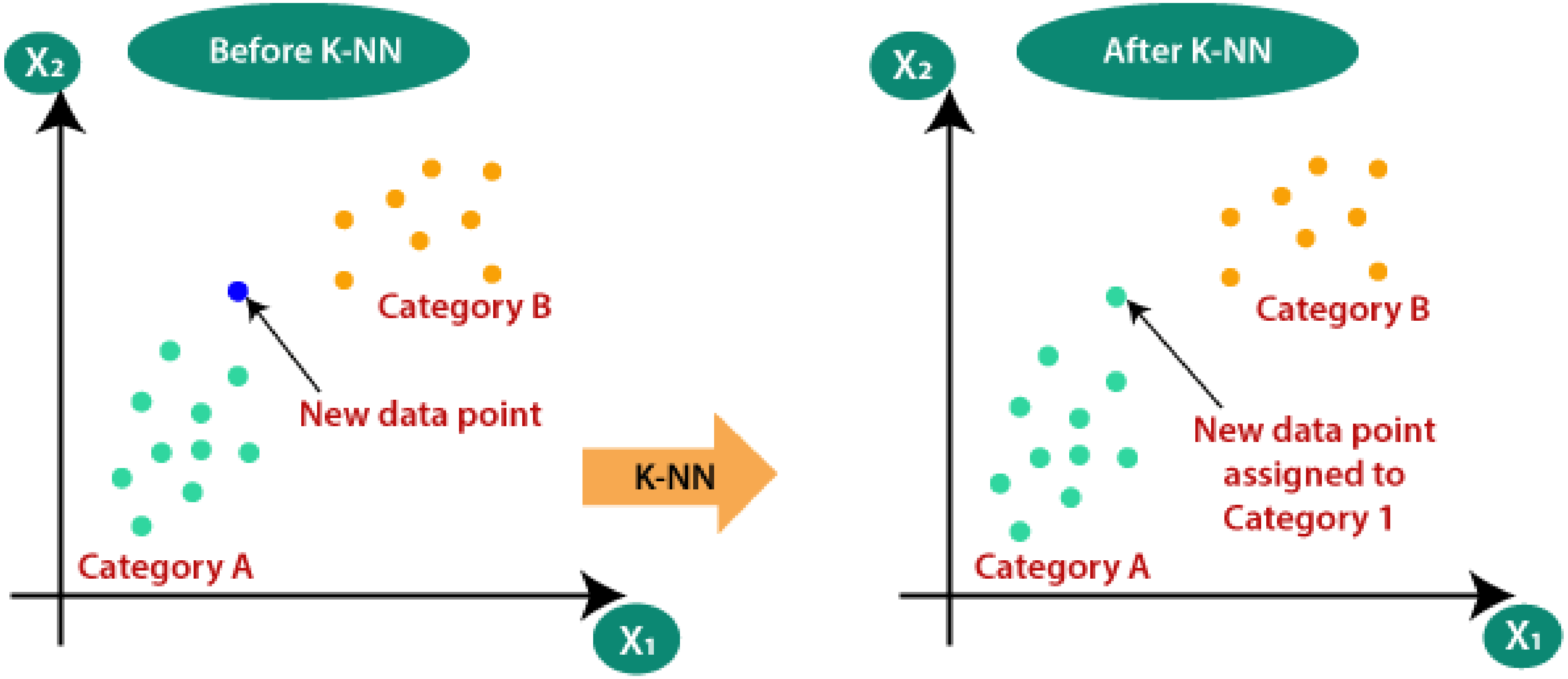
KNN Classifier



The k-NN algorithm



The k-NN algorithm



The K-NN algorithm:

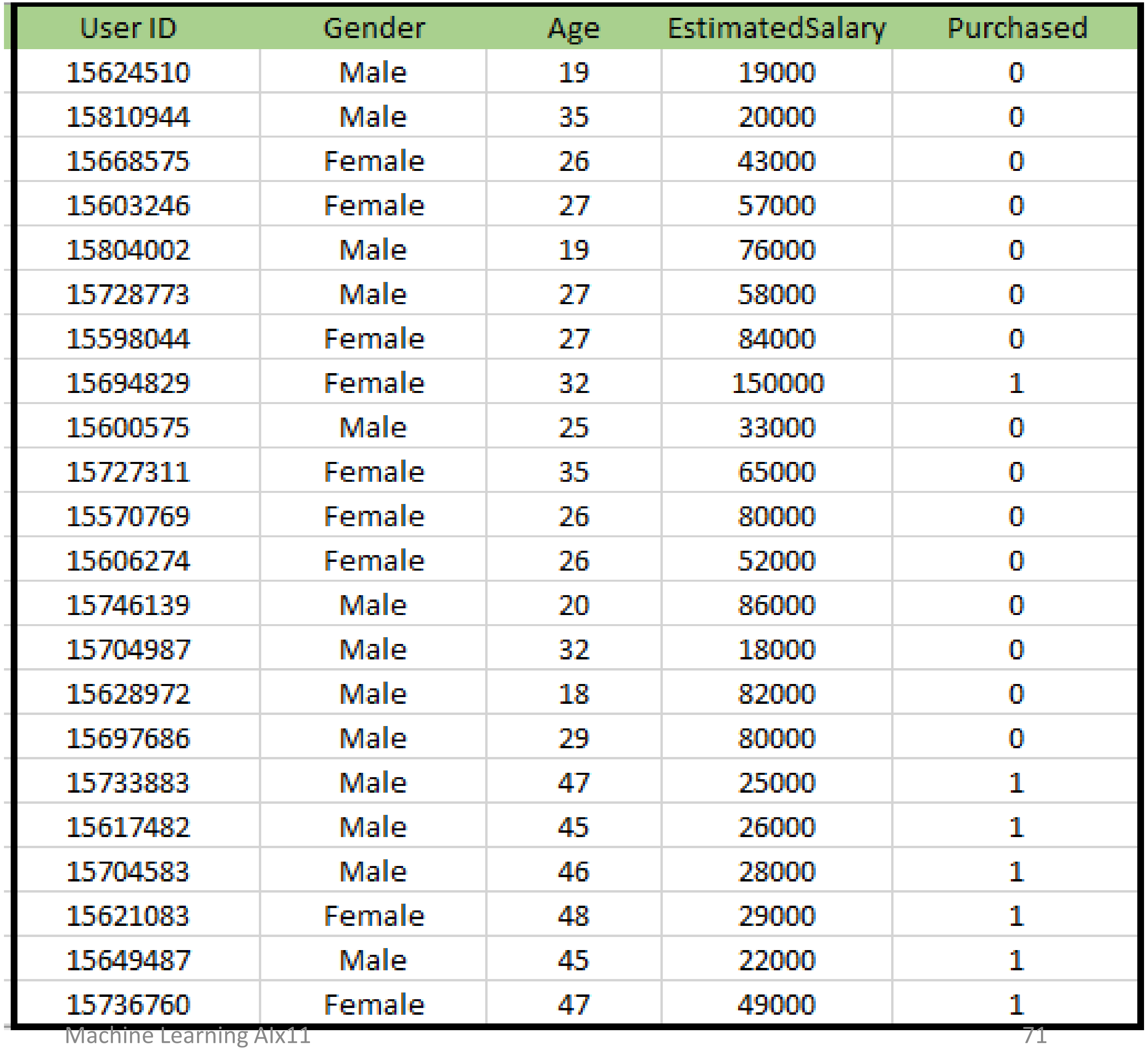
- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
    - $dist(p,q) = \sqrt{(p1 - q1)^2 + (p2 - q2)^2 + \cdots \ldots + (p_n \ - q_n)^2}$
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.

- Example:

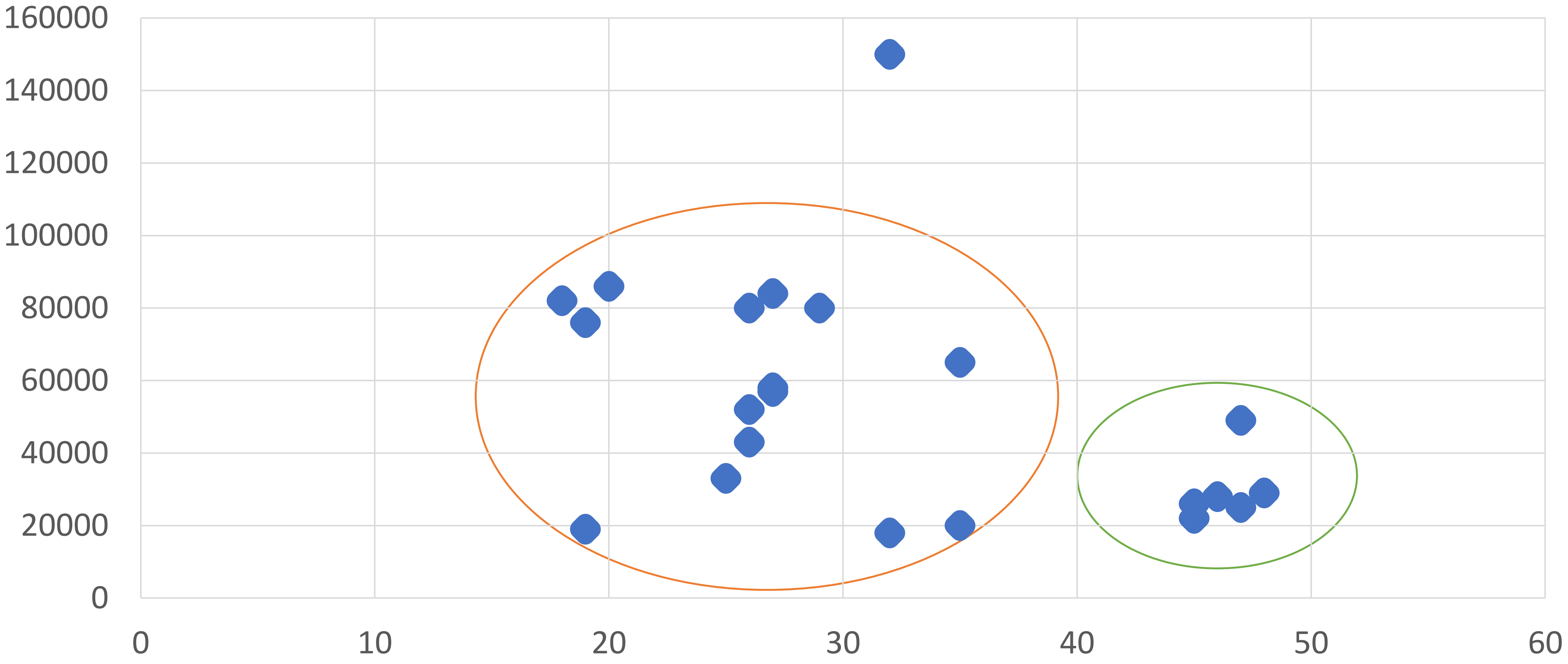
There is a Car manufacturer company that has manufactured a new SUV car.

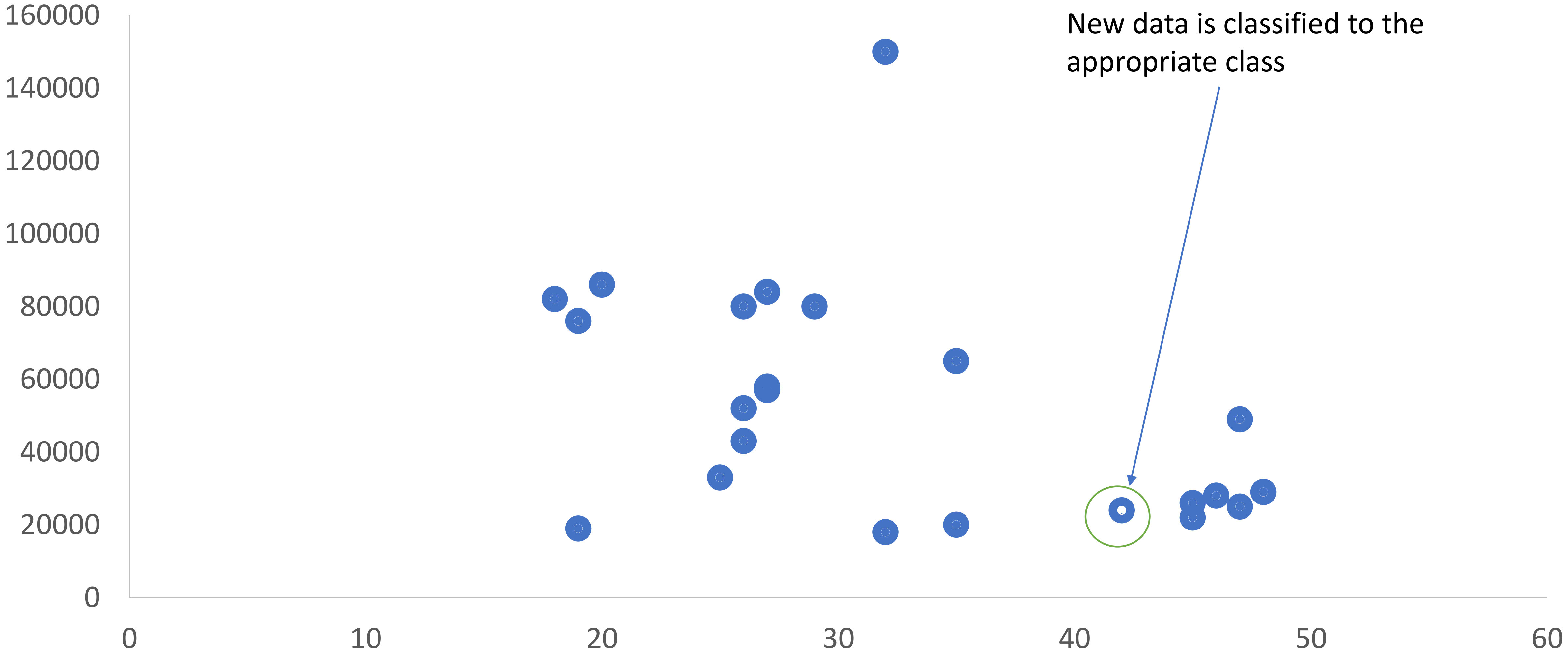
The company wants to give the ads to the users who are interested in buying that SUV.

So for this problem, we have a dataset that contains multiple user's information through the social network.

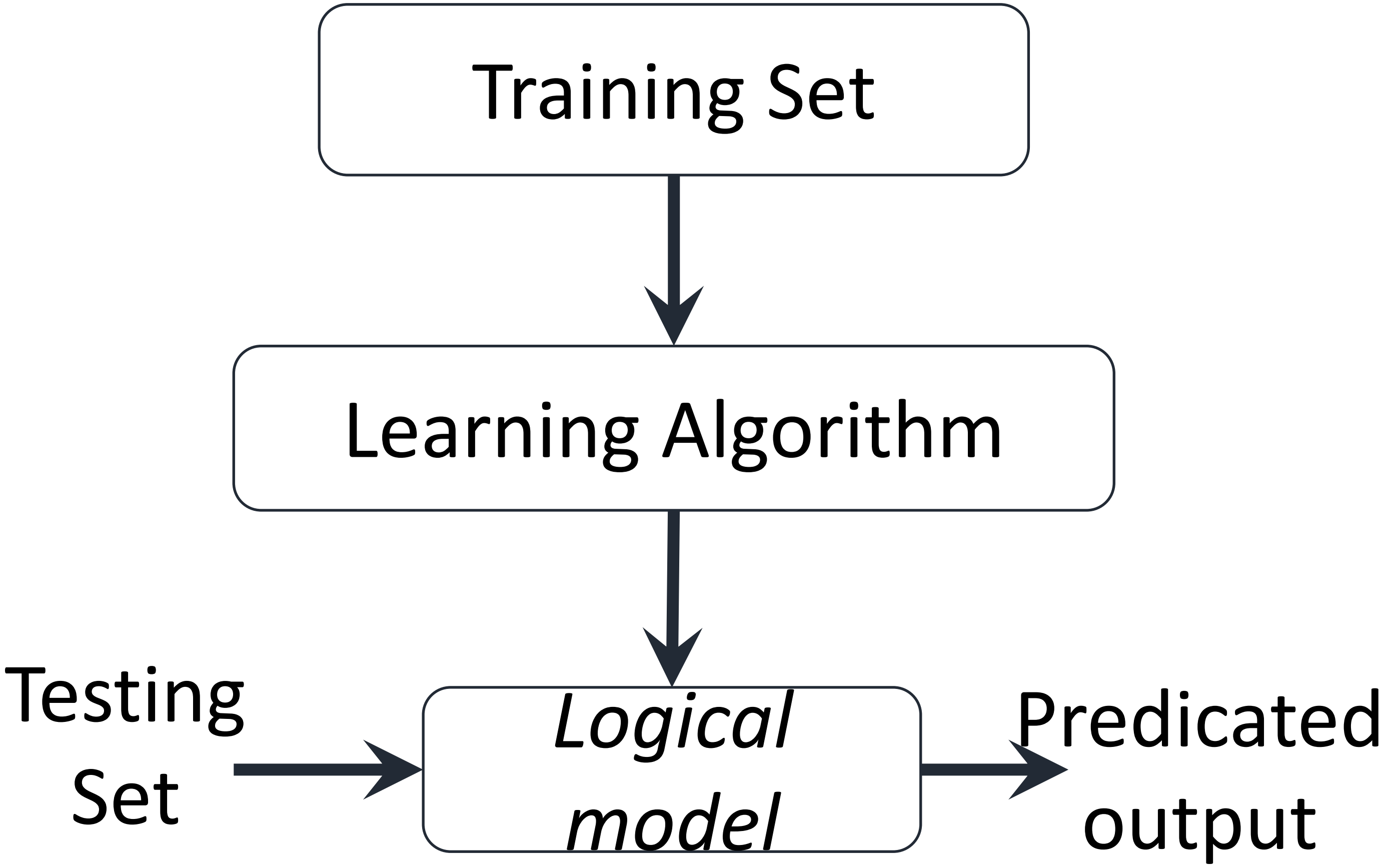


User ID	Gender	Age	EstimatedSalary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0
15570769	Female	26	80000	0
15606274	Female	26	52000	0
15746139	Male	20	86000	0
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1
15617482	Male	45	26000	1
15704583	Male	46	28000	1
15621083	Female	48	29000	1
15649487	Male	45	22000	1
15736760	Female	47	49000	1


160000
140000
120000
100000
80000
60000
40000
20000
0
0
10
20
30
40
50
60



New data is classified to the appropriate class



Training Set

Learning Algorithm

Testing Set

Logical model

Predicated output

- In case of a very large value of k, we may include points from other classes in the neighborhood.
- In case of too small value of k, the algorithm is very sensitive to noise

The k-NN algorithm

- k-NN algorithm can be used for imputing missing values of both categorical and continuous variables.

- For numerical values, Euclidean distance is a good choice. You might want to try Manhattan distance , which is sometimes used as well. For text analytics, cosine distance can be another good alternative worth trying.

- The algorithm's training phase consists only of storing the feature vectors and class labels of the training samples.
- In the testing phase, a test point is classified by assigning the labels that are most frequent among the k training samples nearest to that query point – hence, higher computation.

- **k-NN algorithm does more computation on test time rather than train time.**

(a) True **(b) false**

- **Which of the following statement is true about k-NN algorithm?**

1. k-NN performs much better if all of the data have the same scale
2. k-NN works well with a small number of input variables (p), but struggles when the number of inputs is very large
3. k-NN makes no assumptions about the functional form of the problem being solved

(a)1and 3 **(b)1 and 2** **(c) all of above**

- **Which of the following will be Euclidean Distance between the two data point A(1,3) and B(2,3)?**

(a)1 (b) 2 (c) 4 (d) 8

Which of the following will be true about k in k-NN in terms of Bias?

(A)When you increase the k the bias will be increases
(B) When you decrease the k the bias will be increases
(C) Can’t say
(D) None of these

- **A company has build a kNN classifier that gets 100% accuracy on training data. When they deployed this model on client side it has been found that the model is not at all accurate. Which of the following thing might gone wrong?**
- **Note: Model has successfully deployed and no technical issues are found at client side except the model performance**

- A) It is probably an overfitted model
  B) It is probably an underfitted model
  C) Can’t say
  D) None of these

Dr. Sara Sweidan

Sweidan_ds@fci.bu.edu.eg