MODULE 2.1 – INTRODUCTION TO MACHINE LEARNING

PREPARED BY,

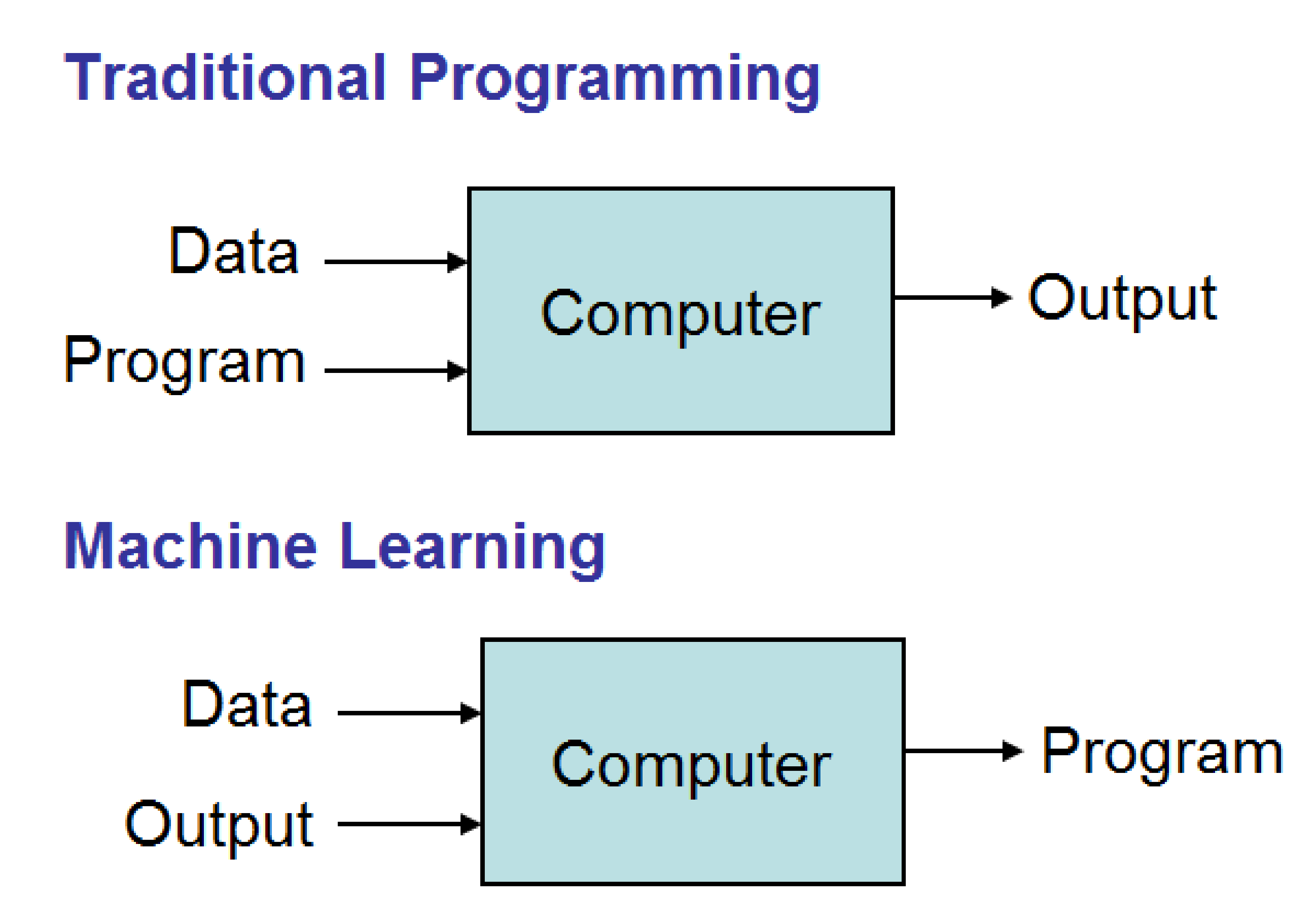
SHELLY SHIJU GEORGE

ASSISTANT PROFESSOR

INTRODUCTION TO MACHINE LEARNING

- The field of machine learning provides a set of algorithms that transform data into actionable knowledge.
- The field of study interested in the development of computer algorithms to transform data into intelligent action is known as **machine learning**.
- A closely related sibling of machine learning, data mining, is concerned with the generation of novel insights from large databases.
- Although there is some disagreement over how widely machine learning and data mining overlap, a potential point of distinction is that machine learning focuses on teaching computers how to use data to solve a problem, while data mining focuses on teaching computers to identify patterns that humans then use to solve a problem.

- A type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed.
- Machine learning focuses on the development of computer programs that can change when exposed to new data.
- Ability of machines in conducting intelligent tasks – AI
- A branch of artificial intelligence concerned with the design and development of algorithms that allow computers to evolve behaviors based on data.
- As intelligence requires knowledge it is necessary for the computers to acquire knowledge.
- Machine learning is about predicting the future based on the past.



Traditional Programming

Machine Learning

HOW MACHINES LEARN

- A formal definition of machine learning proposed by computer scientist Tom M Mitchell states that
- **“A machine learns whenever it is able to utilize its ‘an experience’ such that its performance improves on similar experiences in the future”.**
- Human brains are naturally capable of learning from birth, but the conditions necessary for computers to learn must be made explicit.
- Regardless of whether the learner is a human or machine the basic learning process is similar.

It can be divided into four interrelated Components -

- Data storage
- Abstraction
- Generalization
- Evaluation

1. Data storage -

- Utilizes observation memory, and recall to provide a factual basis for further reasoning.

2. Abstraction -

- Involves the translation of stored data into broader representations and concepts.

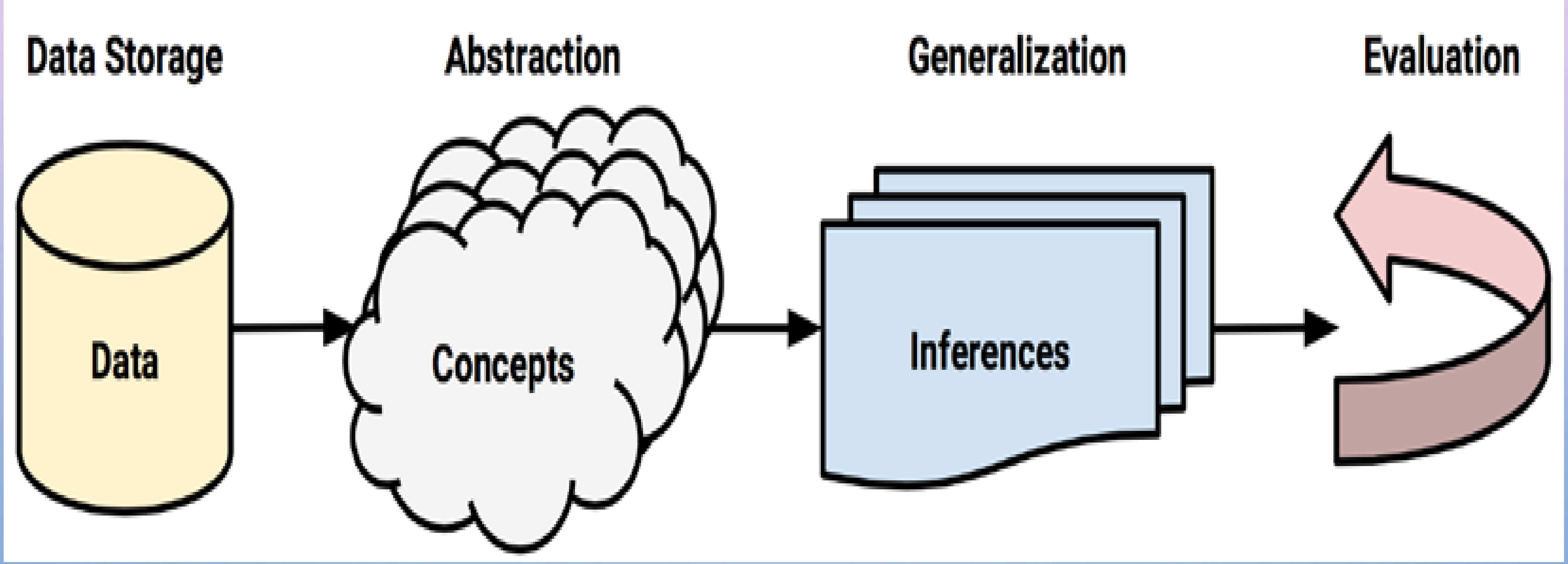
3. Generalization -

- Uses abstracted data to create knowledge and inferences that drive action in new contexts.

4. Evaluation -

- Provides a feedback mechanism to measure the utility of learned knowledge and inform potential improvements.

THE FOLLOWING FIGURE ILLUSTRATES THE STEPS IN THE LEARNING PROCESS:



1. DATA STORAGE

- All learning must begin with data.
- Humans and computers 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 uses hard disk drives, flash memory and random access memory (RAM) in combination with a central processing unit (CPU).

2. 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.
- **Knowledge representation**, the formation of logical structures that assist in turning raw sensory information into a meaningful insight.
- 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

2. ABSTRACTION (CONTINUED)

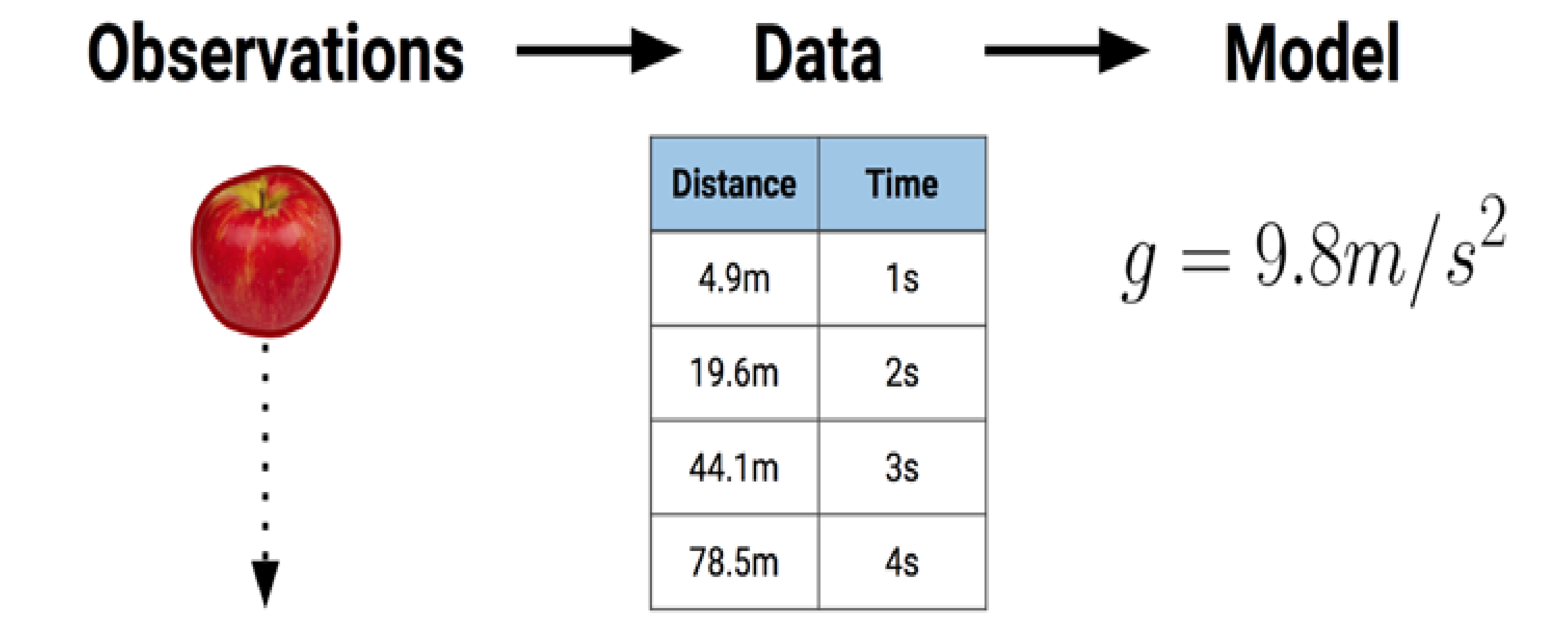
TRAINING

- 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.

2. ABSTRACTION (CONTINUED)

Eg: -discovery of gravity.

- By fitting equations to observational data, Sir Isaac Newton inferred the concept of gravity. But the force we now know as gravity was always present. It was not recognized until Newton recognized it as an abstract concept that relates some data to others.
- i.e, by becoming the ‘g’ term in a model that explains observations of falling objects.



Observations → Data → Model

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

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

3. 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.
- Process is a bit difficult to describe.
- It has been imagined as a search through the entire set of models (inferences) that could be abstracted during training.
- In other words, if you can imagine a hypothetical set containing every possible theory that could be established from the data, generalization involves the reduction of this set into a manageable number of important findings.

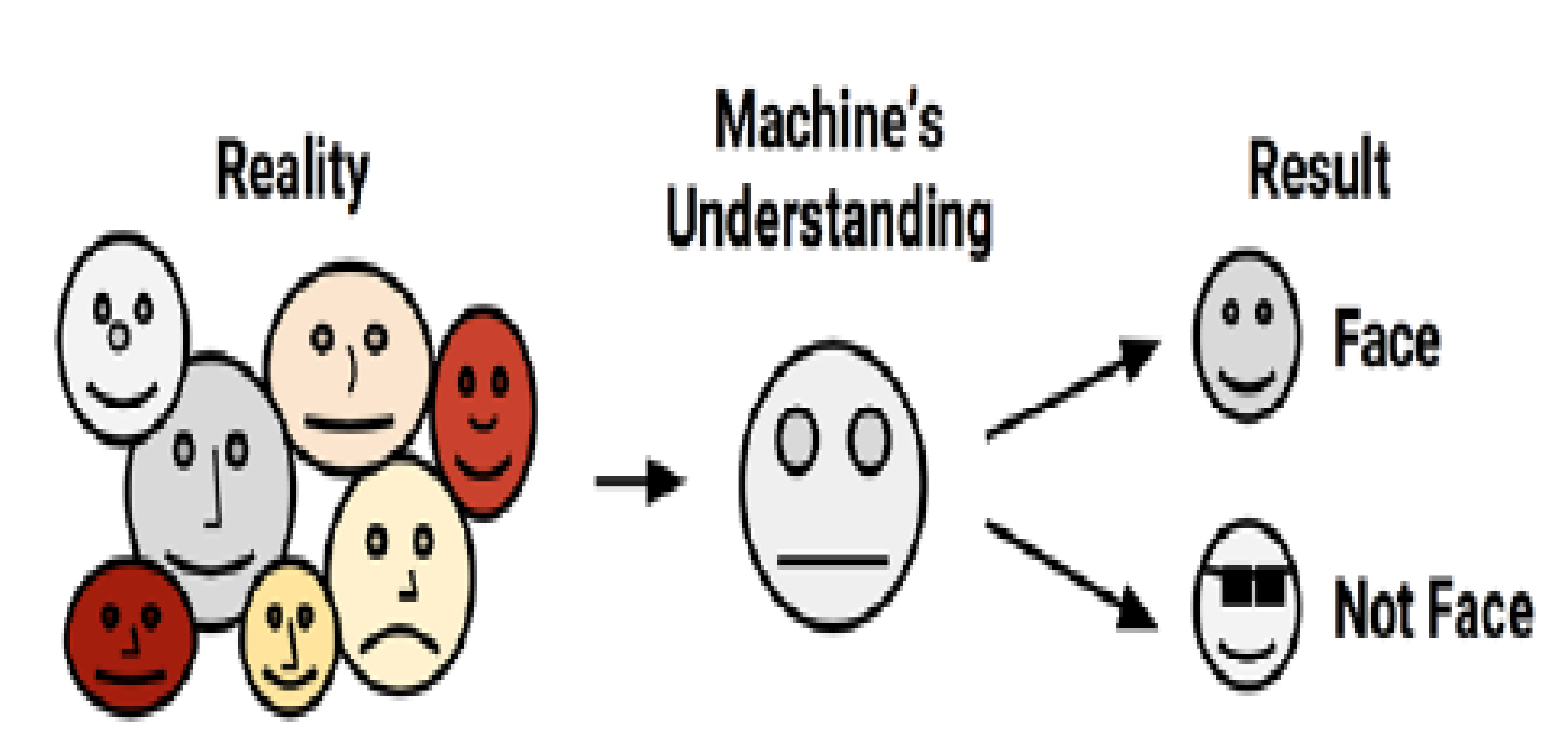
3. GENERALIZATION (CONTINUED)

- In generalization, the learner is tasked with limiting the patterns it discovers, to only those that will be most relevant to its future tasks.
- Generally it is not feasible to reduce the number of patterns by examining them one by one and ranking them by future utility.
- So, machine learning algorithms generally employ shortcuts that reduce the search space more quickly.
- The algorithm will uses **heuristics technique** designed for solving a problem more quickly.

3. GENERALIZATION (CONTINUED)

- Heuristics are routinely used by human beings to quickly generalize experience to new scenarios.
- If you have ever utilized your gut instinct to make a snap decision prior to fully evaluating your circumstances, you were intuitively using mental heuristics.
- The incredible human ability to make quick decisions often relies not on computer-like logic, but rather on heuristics guided by emotions.
- Sometimes, this can result in illogical conclusions.

3. GENERALIZATION (CONTINUED)



4. EVALUATION

- 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 its characterization of the training data generalizes to new, unseen data.
- It’s worth noting that it is exceedingly rare for a model to perfectly generalize to every unforeseen case.

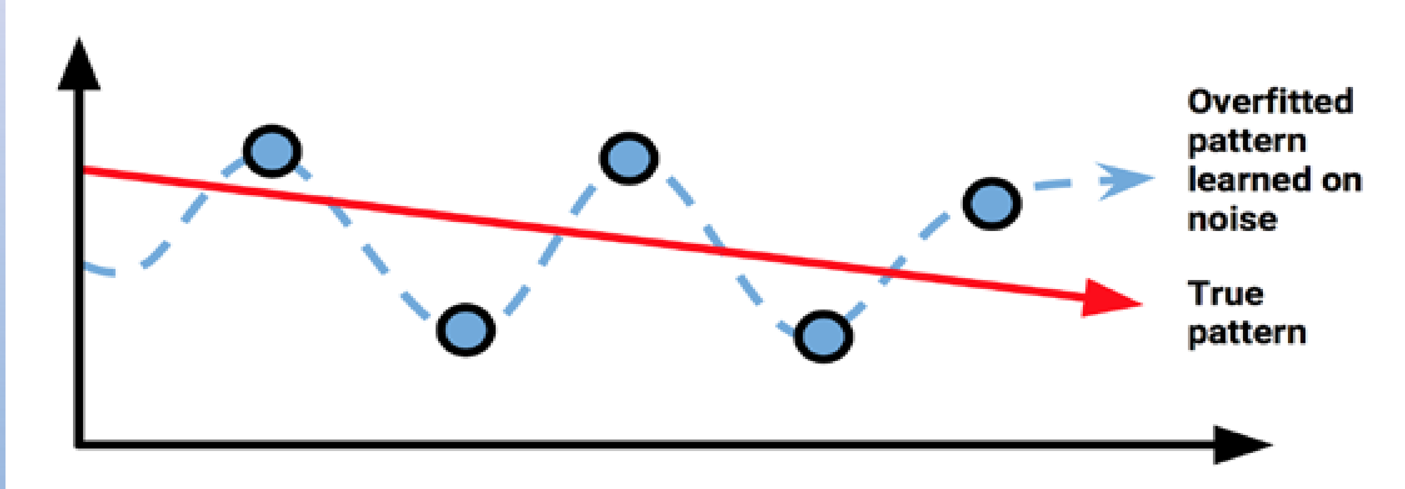
4. EVALUATION (CONTINUED)

In parts, models fail to perfectly generalize due to the problem of **noise**, a term that describes unexplained or unexplainable variations in data. 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.

4. EVALUATION (CONTINUED)

- Trying to model noise is the basis of a problem called **overfitting.**
- Because most noisy data is unexplainable by definition, attempting to explain the noise will result in erroneous conclusions that do not generalize well to new cases.
- Efforts to explain the noise will also typically result in more complex models that will miss the true pattern that the learner tries to identify.
- 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. Regardless of the task at hand, any machine learning algorithm can be deployed by following these steps:

- Data collection
- Data exploration and preparation
- Model training
- Model evaluation
- Model improvement

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 nuances during a practice called data exploration.
- Additional work is required to prepare the data for the learning process.
- This involves fixing or cleaning so-called "messy" data, eliminating unnecessary data, and recoding the data to conform to the learner's expected inputs.

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.

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 INPUT DATA

UNIT OF OBSERVATION

- It is used to describe the smallest entity with measured properties of interest for a study.
- Commonly, the unit of observation is in the form of persons, objects or things, transactions, time points, geographic regions, or measurements.
- Sometimes, units of observation are combined to form units such as person-years, which denote cases where the same person is tracked over multiple years; each person-year comprises of a person's data for one year.

TYPES OF INPUT DATA (CONTINUED)

UNIT OF ANALYSIS

- It is the smallest unit from which the inference is made.
- Although it is often the case, the observed and analyzed units are not always the same.
- For example, data observed from people might be used to analyze trends across countries.

TYPES OF INPUT DATA (CONTINUED)

DATASETS

That store the units of observation and their properties can be imagined as collections of data consisting of:

- **Examples:** instances of the unit of observation for which properties have been recorded.
- **Features:** recorded properties or attributes of examples that may be useful for learning.

TYPES OF INPUT DATA (CONTINUED)

E.g to build a learning algorithm to identify spam e mail

- Unit of observation – e-mail messages
- Examples - specific messages
- Features - consist of the words used in the messages

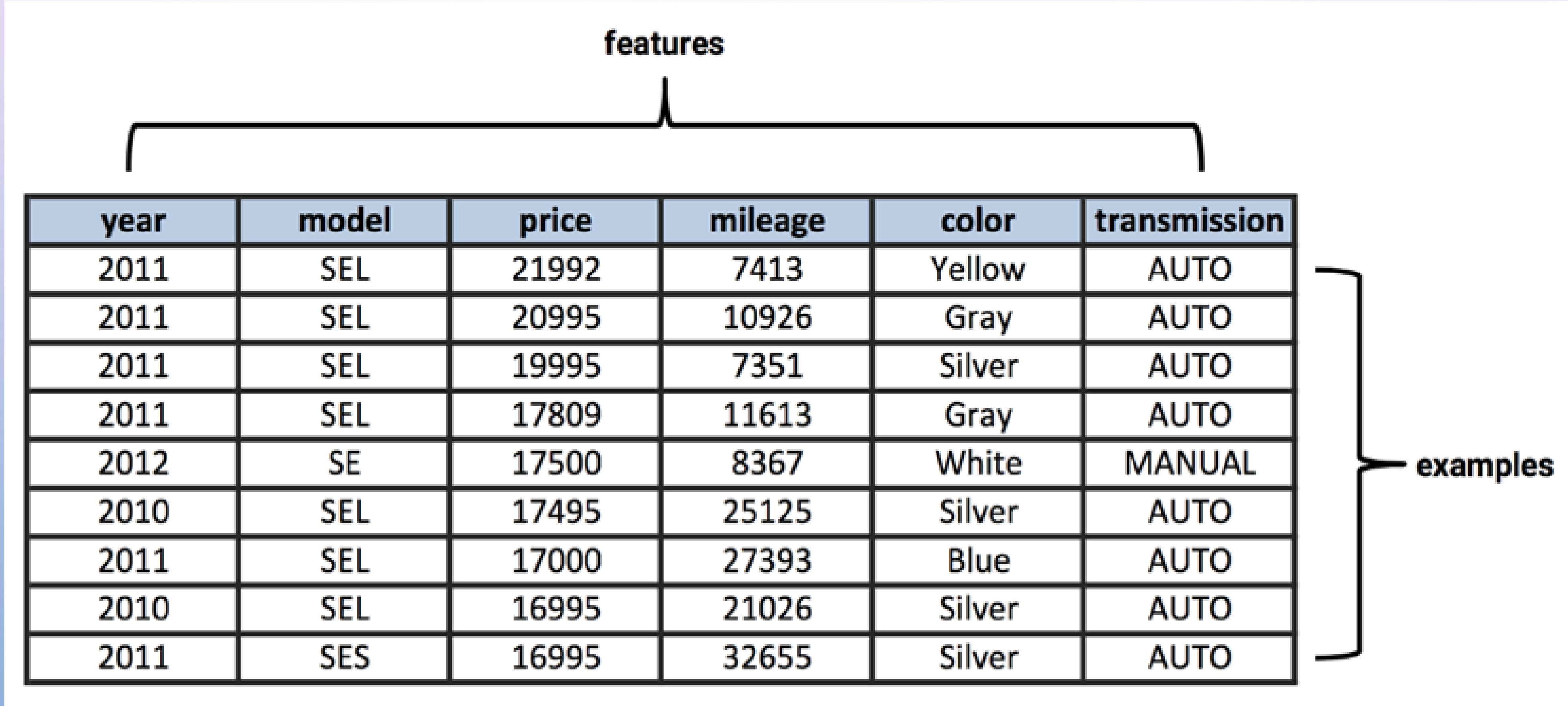
TYPES OF INPUT DATA (CONTINUED)

Examples and features

- Do not have to be collected in any specific form.
- Commonly gathered in **matrix format** which means that each example has exactly the same features.

TYPES OF INPUT DATA (CONTINUED)

- Each row – example
- Each column - feature
- Eg : examples of automobiles



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

VARIOUS FORM OF “FEATURES”

- Numeric
- Categorical / nominal
- Ordinal

1. NUMERIC FEATURE

The feature which represents a characteristic measured in numbers

- Number of black pixels
- Noise ratios, length of sounds, relative power
- Frequency of specific terms
- Height, weight, etc..

2. CATEGORICAL/NOMINAL FEATURE

- A feature which is an attribute that consists of a set of categories.
- Allows you to assign categories.

E.g

- Gender
- Colour of a ball

3. ORDINAL FEATURE

- A special case of categorical variables is called ordinal variable.
- A nominal variable with categories falling in an ordered list.
- An ordinal variable is a categorical variable for which the possible values are ordered.
- E.g clothing sizes small medium, and large
- E.g a measurement of customer satisfaction on a scale from “not at all happy” to “very happy”.
- E.g : ordinal variable - Educational qualification – SSLC, Plus Two, degree, PG

- It is important to consider “**what the features represent**”**,** as the **“type”** and **“number of features”** in your dataset which will assist in determining an appropriate machine learning algorithm for your task.

TYPES OF MACHINE LEARNING ALGORITHMS

Machine learning algorithms are divided into categories according to their purpose

- Predictive model
- Descriptive model

Learning algorithms

- Supervised learning
- Unsupervised learning

PREDICTIVE MODEL

- 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.
- E.g. model is used to predict whether an email is spam or “ham”
- E.g. supervised learning algorithms are used.

DESCRIPTIVE MODEL

- No single feature is more important than any other
- No target to learn
- Process of training a descriptive model is called unsupervised learning

Descriptive modeling task

- Pattern discovery (market basket analysis)
- Clustering (segmentation analysis)

If user buys A e.g. bread) then machine should automatically give him a suggestion to buy B(e.g. Jam)

DESCRIPTIVE MODEL

- **Pattern discovery** - used to identify useful associations within data.
- **Clustering** - dividing a dataset into homogeneous groups.

SUPERVISED LEARNING

- Process of training a predictive model is known as supervised learning.
- Supervised learning algorithm attempts to optimize a function (model) to find the combination of feature values that result in the target output.
- Supervised learning is where you have input variables (X) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output

$Y = f(X)$

SUPERVISED LEARNING (CONTINUED)

- The often used supervised machine learning task of predicting which category an example belongs to is known as **classification**.
- Classifier

E.g. we could predict 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

UNSUPERVISED LEARNING

- Have input data (X) and no corresponding output variables.

The goal for unsupervised learning is

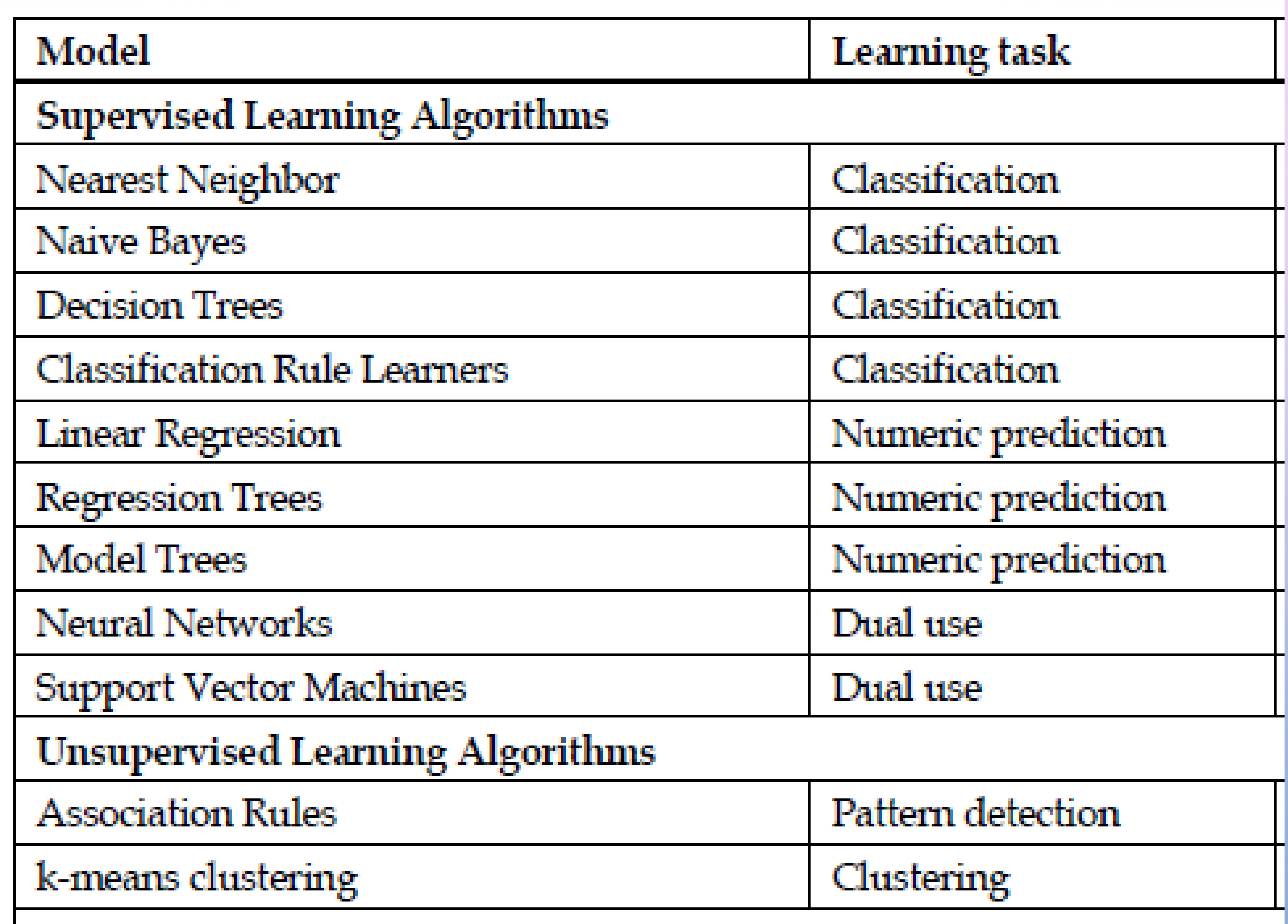
- To “model the underlying structure or distribution in the data” in order to learn more about the data

Clustering

Association rules

- K means for clustering problems
- Apriori algorithm for association rule learning problems

GENERAL TYPES OF MACHINE LEARNING ALGORITHMS



Model	Learning task
Supervised Learning Algorithms	
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
Unsupervised Learning Algorithms	
Association Rules	Pattern detection
k-means clustering	Clustering