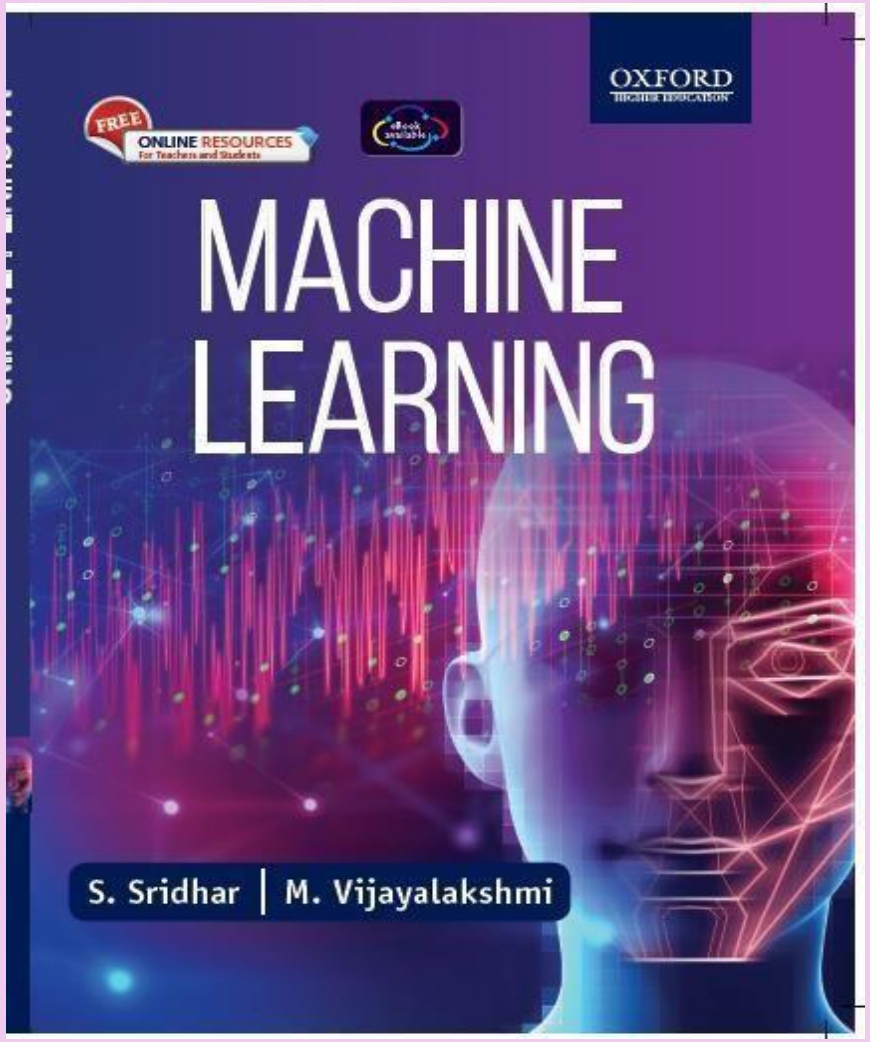


OXFORD
UNIVERSITY PRESS

Machine Learning

S. Sridhar and M. Vijayalakshmi

Chapter 1

Introduction to Machine Learning

Need for Machine Learning?

- BUSINESS ORGANIZATIONS HAVE NUMEROUS DATA
- NEED TO ANALYZE DATA FOR TAKING DECISIONS

1. High volume of available data to manage
2. Cost of storage has reduced: capture, process, store, distribute, and transmit the digital information.
3. Availability of complex algorithms now.

Need for Machine Learning?

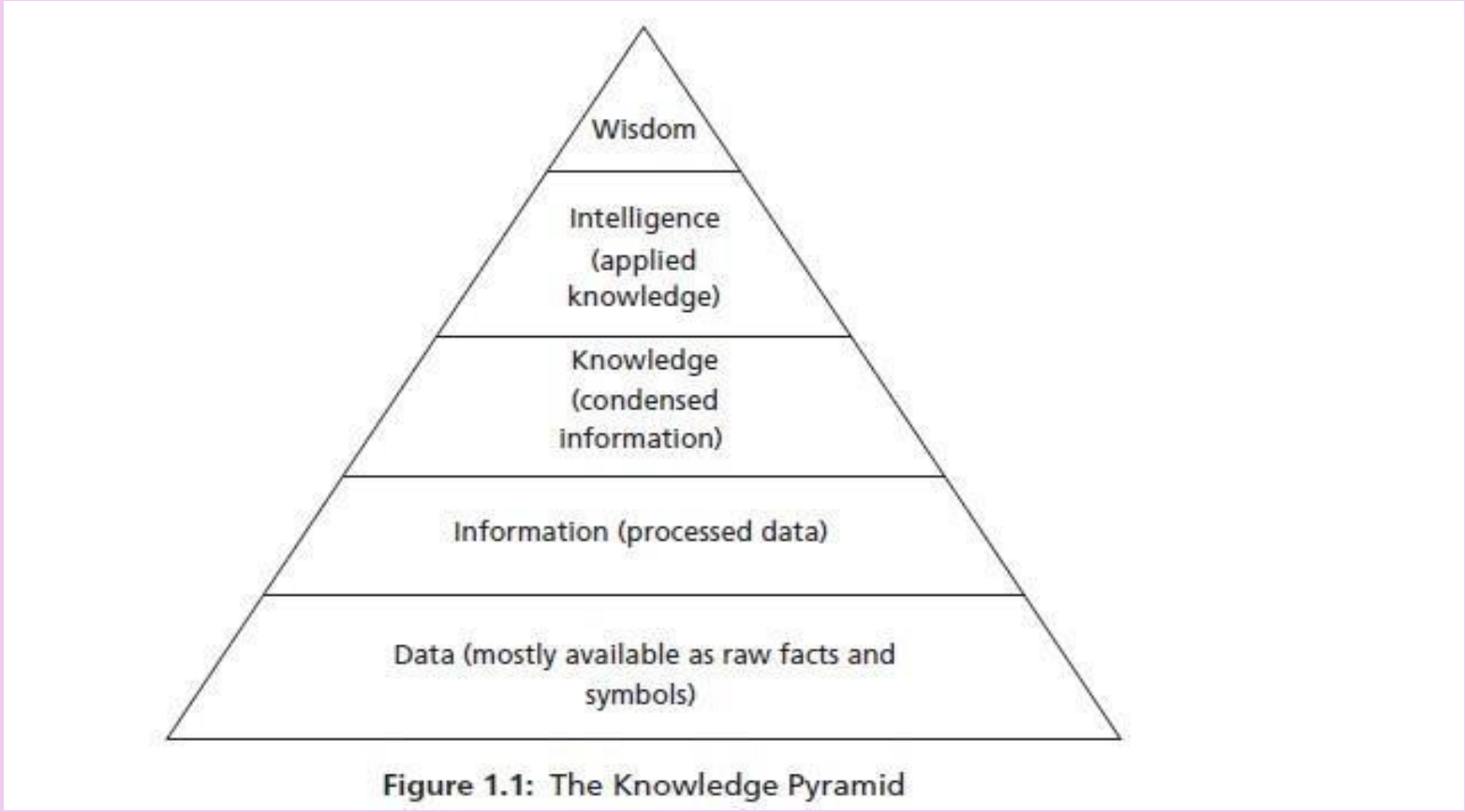


Figure 1.1: The Knowledge Pyramid

What is Machine Learning?

- **"MACHINE LEARNING IS A FIELD OF STUDY THAT GIVES THE COMPUTERS TO LEARN WITHOUT BEING EXPLICITLY PROGRAMMED"**

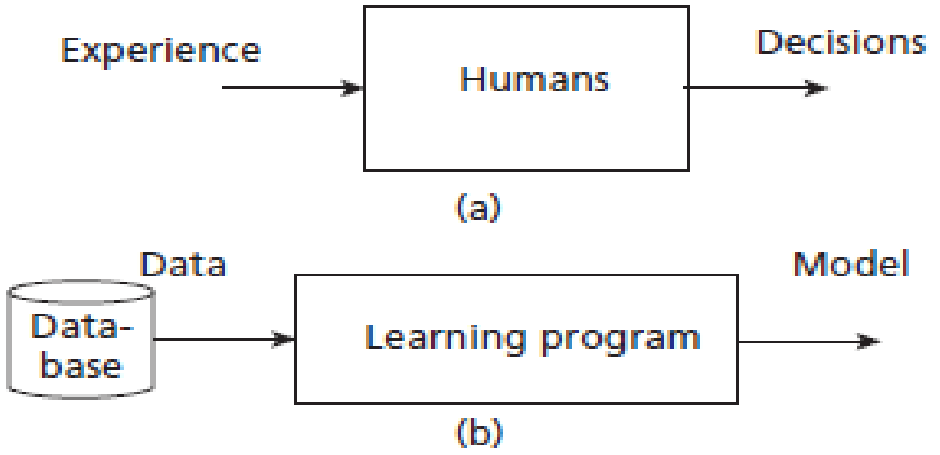


Figure 1.2: (a) A Learning System for Humans (b) A Learning System for Machine Learning

Y=f(x), f is the learning function that maps the input *x* to output *y*. Learning of function *f* is the crucial aspect of forming a model in statistical learning.

What is a Model?

A MODEL CAN BE ANY ONE OF THE FOLLOWING –

- **MATHEMATICAL EQUATION**
- **RELATIONAL DIAGRAMS LIKE GRAPHS/TREES**
- **LOGICAL IF/ELSE RULES**
- **GROUPINGS CALLED CLUSTERS**

- TOM MITCHELL DEFINITION OF MACHINE LEARNING

"A computer program is said to learn from experience E, with respect to task T and some performance measure P, if its performance on T measured by P improves with experience E." The important components of this definition are experience *E*, task *T*, and performance measure *P*.

Machine Learning and AI

- **RELATION BETWEEN MACHINE LEARNING AND AI**

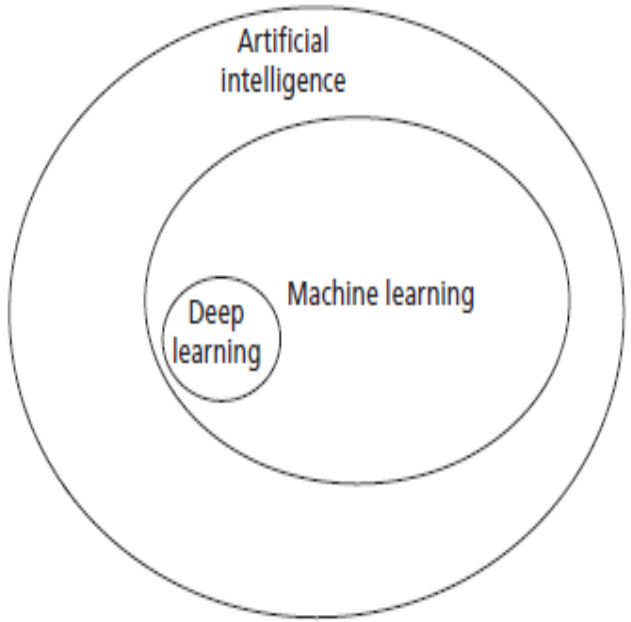


Figure 1.3: Relationship of AI with Machine Learning

Machine Learning and Data Science

- DATA SCIENCE IS AN "UMBRELLA TERM" COVERING FROM DATA COLLECTION TO DATA ANALYSIS.

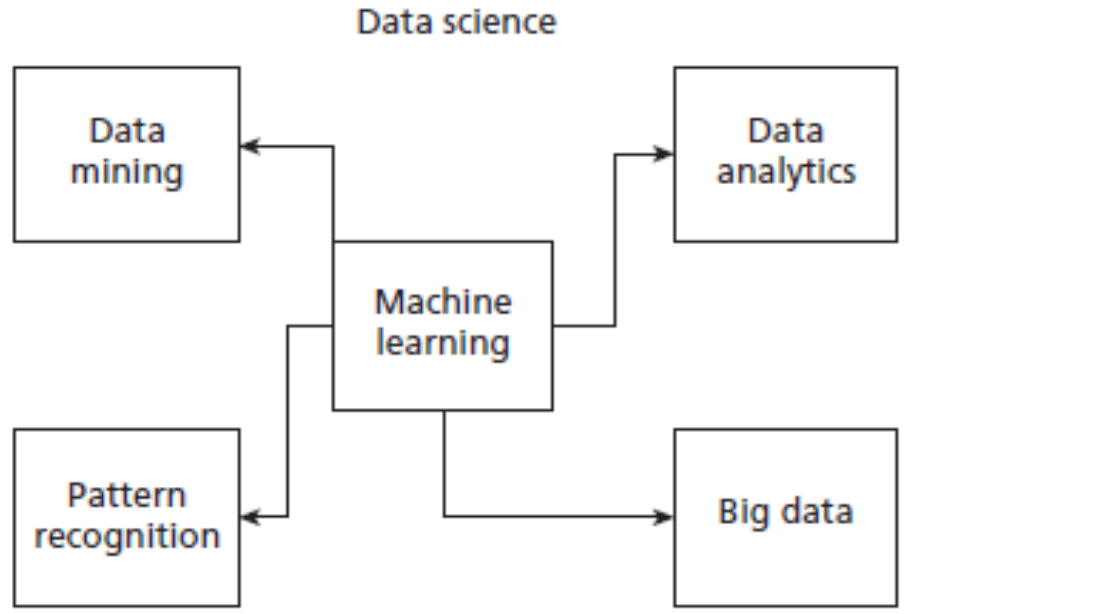


Figure 1.4: Relationship of Machine Learning with Other Major Fields

Machine Learning and Data Science

CHARACTERISTICS OF BIG DATA

Big data is a field of data science that deals with data's following characteristics:

1. **Volume:** Huge amount of data is generated by big companies like Facebook, Twitter, YouTube.
2. **Variety:** Data is available in variety of forms like images, videos, and in different formats.
3. **Velocity:** It refers to the speed at which the data is generated and processed.

Machine Learning and Data Science

1. **Data Mining**

   Data mining's original genesis is in the business

2. **Data Analytics**: aims to extract useful knowledge from crude data

3. **Pattern Recognition**: It uses machine learning algorithms to extract the features for pattern analysis and pattern classification.

Machine Learning and Statistics

1. It is mathematics intensive and models are often complicated equations and involve many assumptions.
2. Statistical methods are developed in relation to the data being analyzed

Machine Learning Types

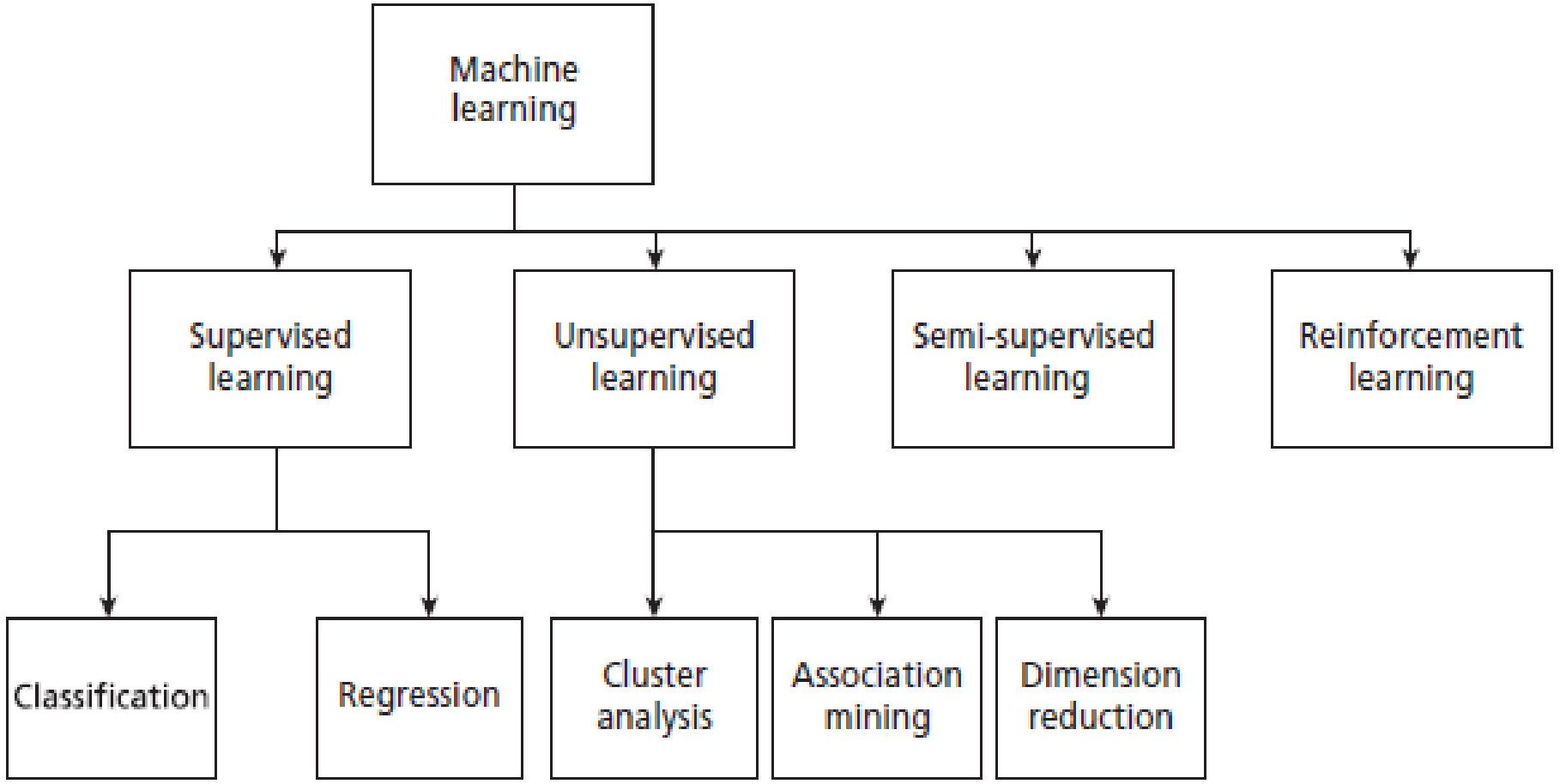


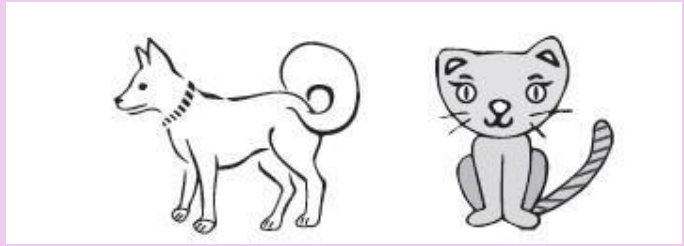
Figure 1.5: Types of Machine Learning

Labelled Data

1. Data is a raw fact.
2. Data is represented in the form of a table.
3. Data also can be referred to as a data point, sample, or an example.
4. Each row of the table represents a data point.
5. Features are attributes or characteristics of an object.
6. Columns are attributes. Out of all attributes, one attribute is important and is called a label

Unlabelled Data

DATA THAT IS NOT ASSOCIATED WITH LABELS IS CALLED UNLABELLED DATA.



Supervised Learning

1. Supervised algorithms use labelled dataset.

2. There is a supervisor or teacher component in supervised learning.
3. A supervisor provides labelled data so that the model is constructed and generates test data

Supervised Learning

Supervised learning has two methods:

1. Classification
2. Regression

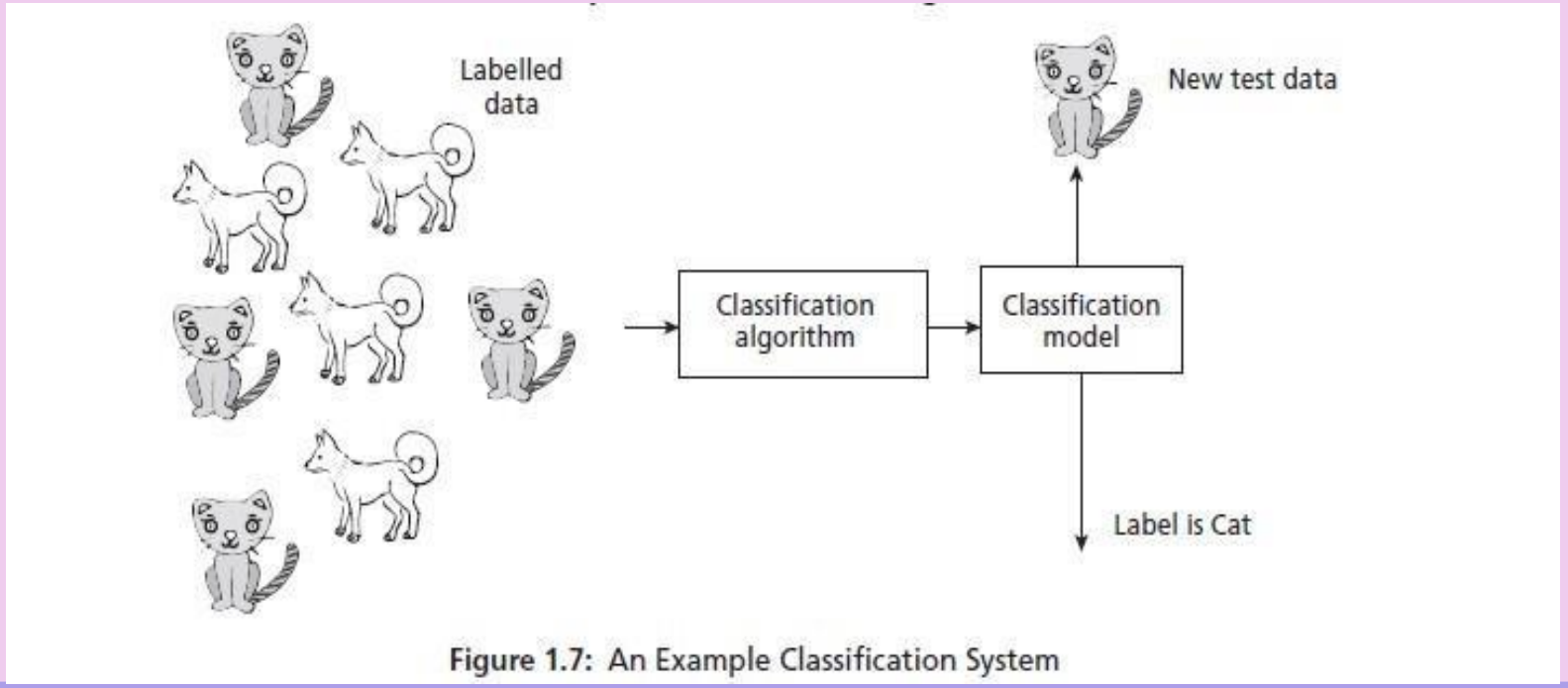


Figure 1.7: An Example Classification System

Supervised Learning

The input attributes of the classification algorithms are called independent variables.

The target attribute is called label or dependent variable.

The relationship between the input and target variable - classification model.

In classification, learning takes place in two stages.

- **Training stage**, the learning algorithm takes a labelled dataset and starts learning. After the training set, samples are processed and the model is generated.
- **Constructed model** is tested with test or unknown sample and assigned a label. This is the classification process.

Supervised Learning

S.No.	Length of Petal	Width of Petal	Length of Sepal	Width of Sepal	Class
1.	5.5	4.2	1.4	0.2	Setosa
2.	7	3.2	4.7	1.4	Versicolor
3.	7.3	2.9	6.3	1.8	Virginica

Supervised Learning

KEY ALGORITHMS

Some of the key algorithms of classification are:

- Decision Tree
- Random Forest
- Support Vector Machines
- Naïve Bayes
- Artificial Neural Network and Deep Learning networks like CNN

Supervised Learning

REGRESSION ALGORITHMS

The regression model takes input x and generates a model in the form of a fitted line of the form $y = f(x)$. Here, x is the independent variable that may be one or more attributes and y is the dependent variable

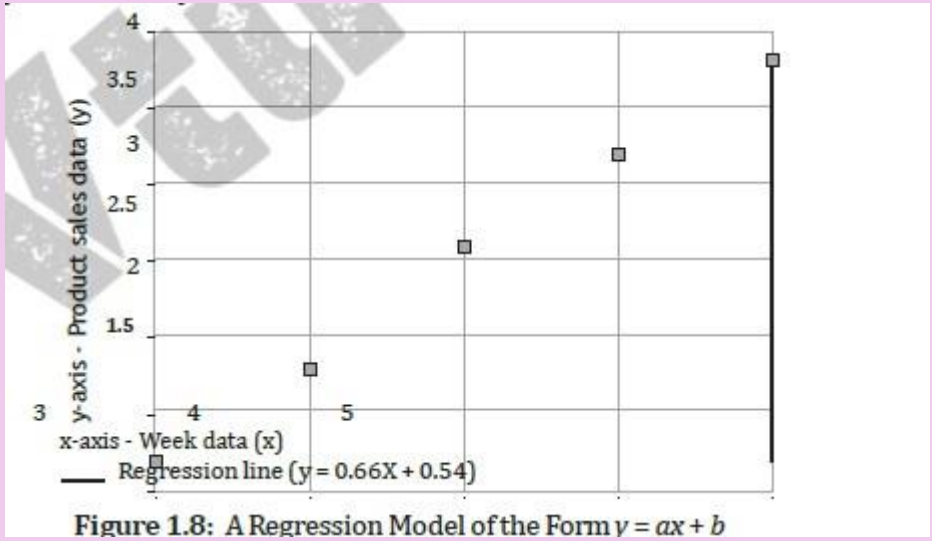


Figure 1.8: A Regression Model of the Form $y = ax + b$

product sales = 0.66 × Week + 0.54.

Unsupervised Learning

1. learning is by self-instruction.
2. There are no supervisor or teacher components.
3. In the absence of a supervisor or teacher, self-instruction is the most common kind of learning process.
4. This process of self-instruction is based on the concept of trial and error

Unsupervised Learning

Cluster Analysis

1. group objects into disjoint clusters or groups.
2. Clusters objects based on its attributes.
3. All the data objects of the partitions are similar in some aspect and vary from the data objects in the other partitions significantly.

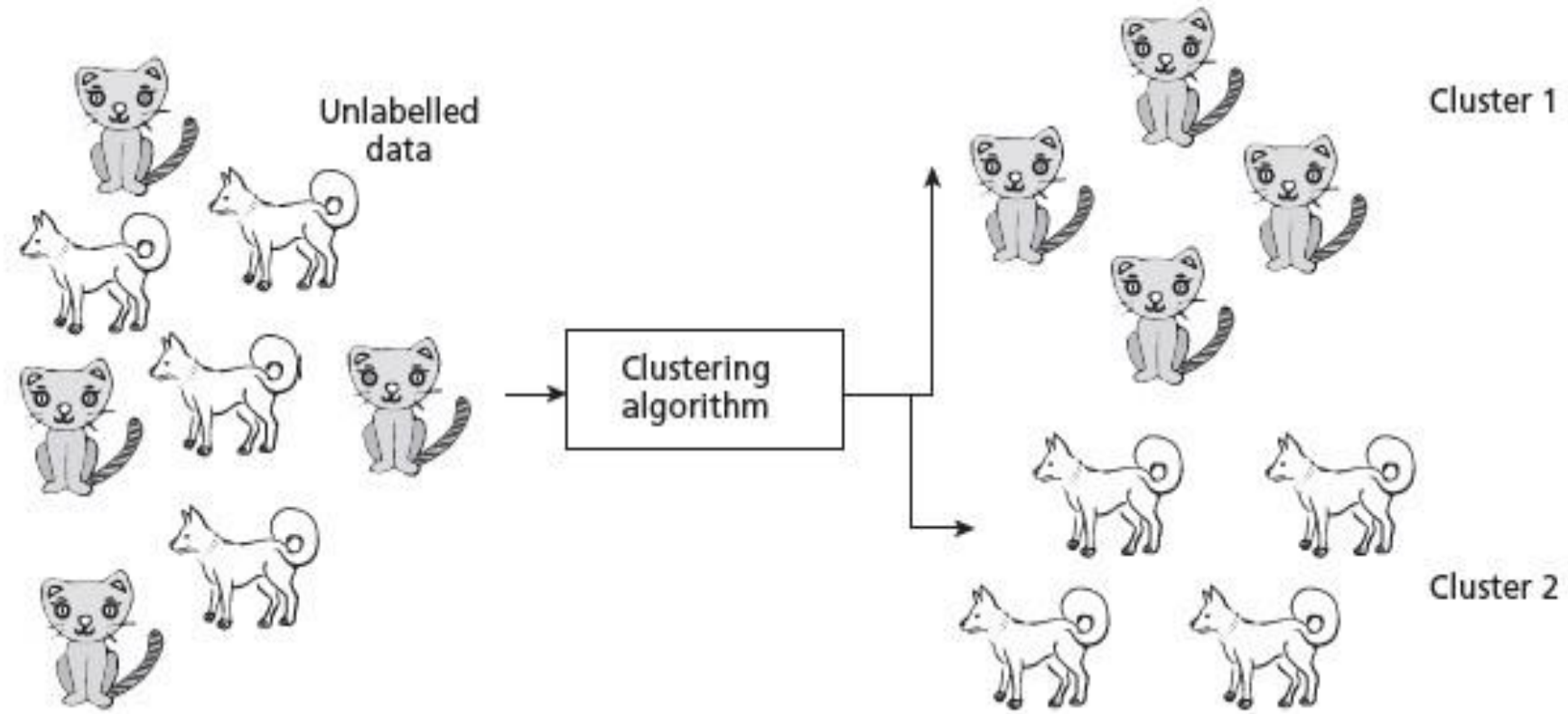


Figure 1.9: An Example Clustering Scheme

Unsupervised Learning

KEY ALGORITHMS OF UNSUPERVISED LEARNING

- k-means algorithm
- Hierarchical algorithms

Key Differences

S.No.	Supervised Learning	Unsupervised Learning
1.	There is a supervisor component	No supervisor component
2.	Uses Labelled data	Uses Unlabelled data
3.	Assigns categories or labels	Performs grouping process such that similar objects will be in one cluster

Semi-supervised Learning

There are circumstances where the dataset has a huge collection of unlabelled data and some labelled data. Labelling is a costly process and difficult to perform by the humans. Semi-supervised algorithms use unlabelled data by assigning a pseudo-label. Then, the labelled and pseudo-labelled dataset can be combined.

Reinforcement Learning

1. Reinforcement learning mimics human beings.
2. Allows the agent to interact with the environment to get rewards.
3. The agent can be human, animal, robot, or any independent program. The rewards enable the agent to gain experience. The agent aims to maximize the reward.
4. The reward can be positive or negative (Punishment). When the rewards are more, the behavior gets reinforced and learning becomes possible.

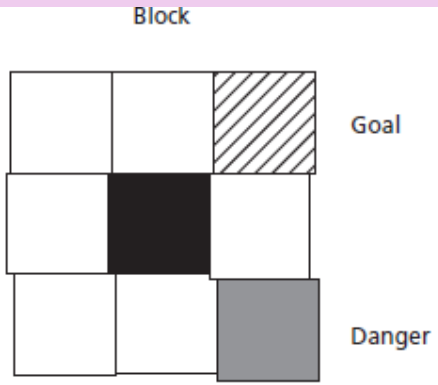


Figure 1.10: A Grid game

Challenges of Machine Learning

1. **Problems** –
2. **Huge data** –
3. **High computation power .**
4. **Complexity of the algorithms** –
5. **Bias/Variance** – Variance is the error of the model..

Machine Learning Process

MACHINE LEARNING/DATA MINING PROCESS

CRISP-DM stands for Cross Industry Standard Process – Data Mining

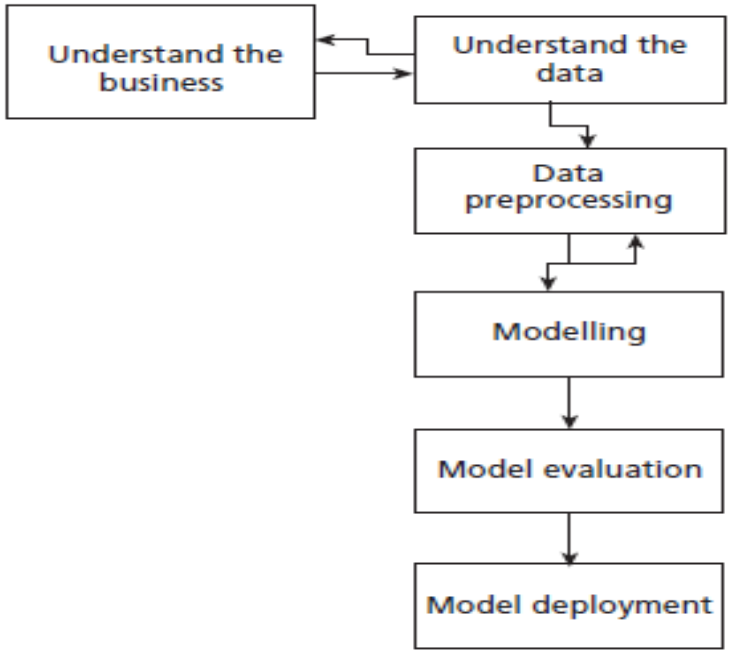


Figure 1.11: A Machine Learning/Data Mining Process

Machine Learning Applications

MACHINE LEARNING MAJOR APPLICATIONS

1. **Sentiment analysis:** NLP words of documents are converted to sentiments like happy, sad, and angry.
2. **Recommendation systems** – These are systems that make personalized purchases possible.
3. **Voice assistants** –Amazon Alexa, Microsoft Cortana, Apple Siri, and Google Assistant are all examples of voice assistants. They take speech commands and perform tasks.
4. **Technologies like Google Maps** and those used by Uber are all examples of machine learning which offer to locate and navigate shortest paths to reduce time.

S.No.	Problem Domain	Applications
1.	Business	Predicting the bankruptcy of a business firm
2.	Banking	Prediction of bank loan defaulters and detecting credit card frauds
3.	Image Processing	Image search engines, object identification, image classification, and generating synthetic images
4.	Audio/Voice	Chatbots like Alexa, Microsoft Cortana. Developing chatbots for customer support, speech to text, and text to voice
5.	Telecommuni-cation	Trend analysis and identification of bogus calls, fraudulent calls and its callers, churn analysis
6.	Marketing	Retail sales analysis, market basket analysis, product performance analysis, market segmentation analysis, and study of travel patterns of customers for marketing tours
7.	Games	Game programs for Chess, GO, and Atari video games
8.	Natural Language Translation	Google Translate, Text summarization, and sentiment analysis
9.	Web Analysis and Services	Identification of access patterns, detection of e-mail spams, viruses, personalized web services, search engines like Google, detection of promotion of user websites, and finding loyalty of users after web page layout modification
10.	Medicine	Prediction of diseases, given disease symptoms as cancer or diabetes. Prediction of effectiveness of the treatment using patient history and Chatbots to interact with patients like IBM Watson uses machine learning technologies.
11.	Multimedia and Security	Face recognition/identification, biometric projects like identification of a person from a large image or video database, and applications involving multimedia retrieval
12.	Scientific Domain	Discovery of new galaxies, identification of groups of houses based on house type/geographical location, identification of earthquake epicenters, and identification of similar land use