Machine Learning (INT_E 14121)

Textbooks

[1]. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

[2]. Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA.

Lesson 1-Introduction

Outline

- What is machine learning?
- Why is machine learning?
- When should use machine learning?
- Types of learning
- Frameworks for building machine learning systems
- Machine Learning Perspective of Data
- Machine Learning Python Packages

What is machine-learning?

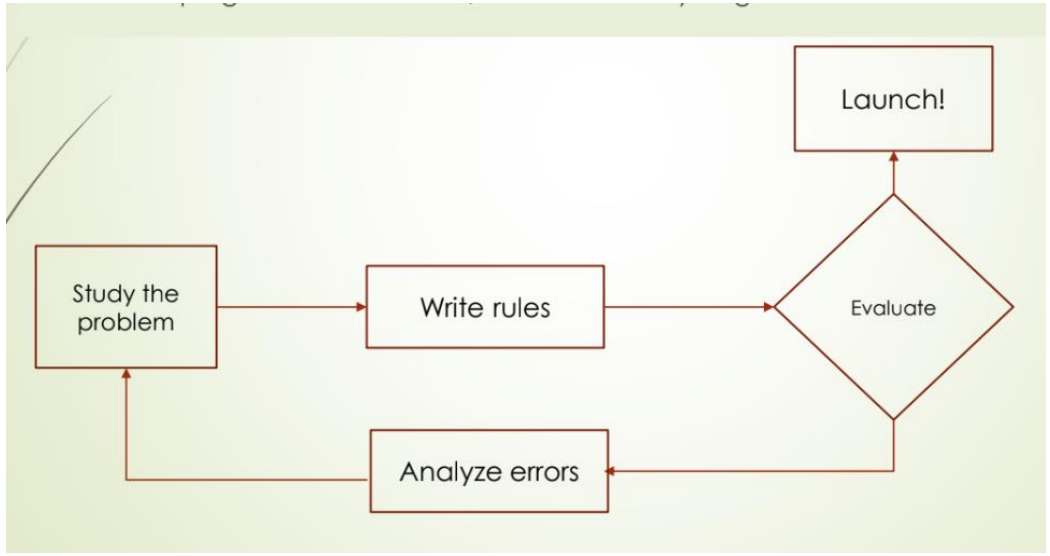
- In 1959, Arthur Samuel, an American pioneer in the field of computer gaming, machine learning, and artificial intelligence has defined machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed."
- Machine learning is the practice of programming computers to learn from data.
- Machine learning is a subfield of computer science that is concerned with building algorithms which, to be useful, rely on a collection of examples of some phenomenon. These examples can come from nature, be handcrafted by humans or generated by another algorithm.
- Machine learning can also be defined as the process of solving a practical problem by
  - 1) gathering a dataset, and
  - 2) algorithmically building a statistical model based on that dataset. That statistical model is assumed to be used somehow to solve the practical problem.

Expert system vs Machine-learning

- Intelligent Sofware 1.0: inputs + rules → results
- Intelligent Sofware 2.0: inputs + results → rules

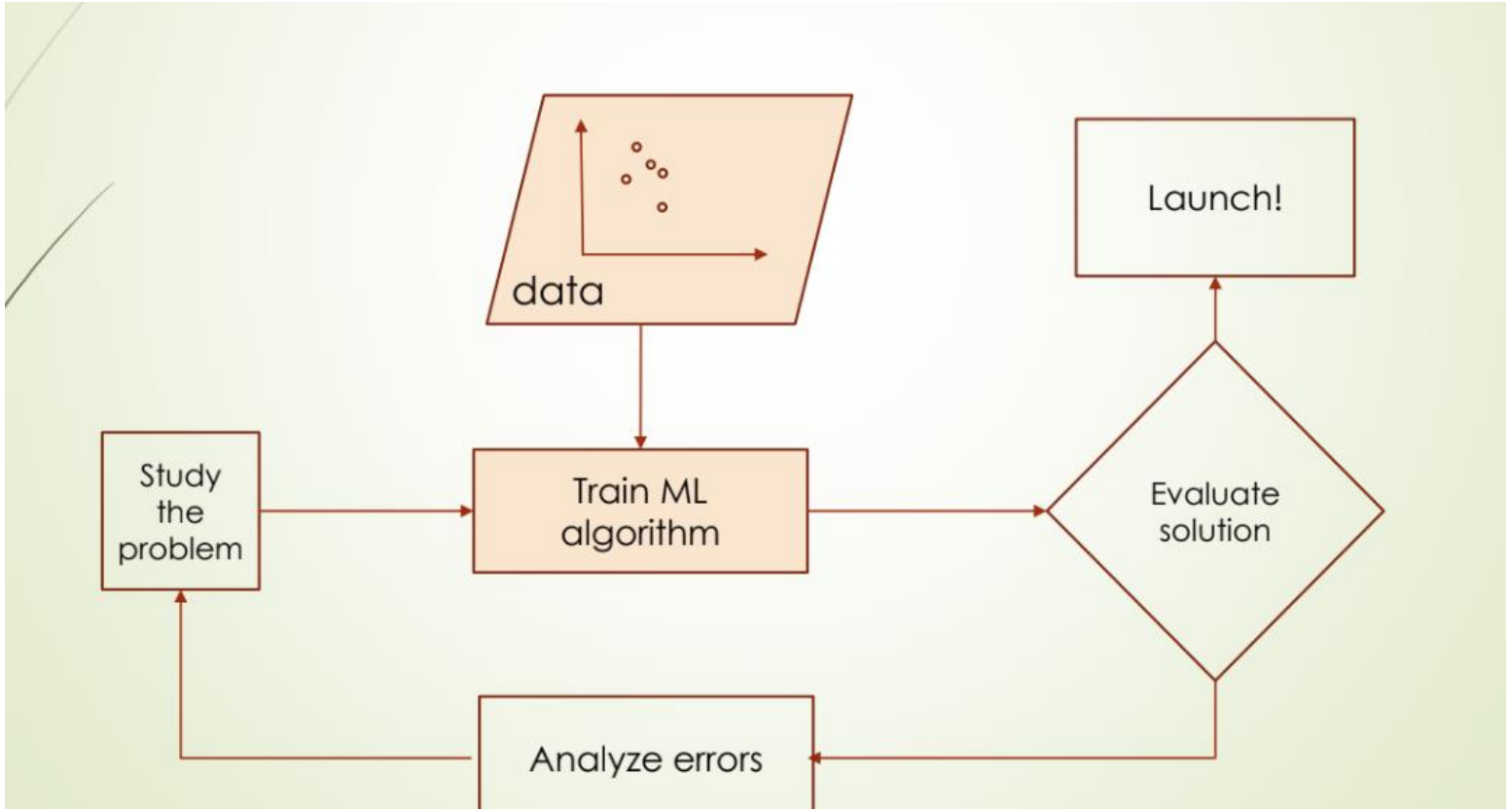
Why is machine learning?

- Spam e-mails filter Without ML
- The program is not software it contains a very long list of rules that are difficult to maintain.

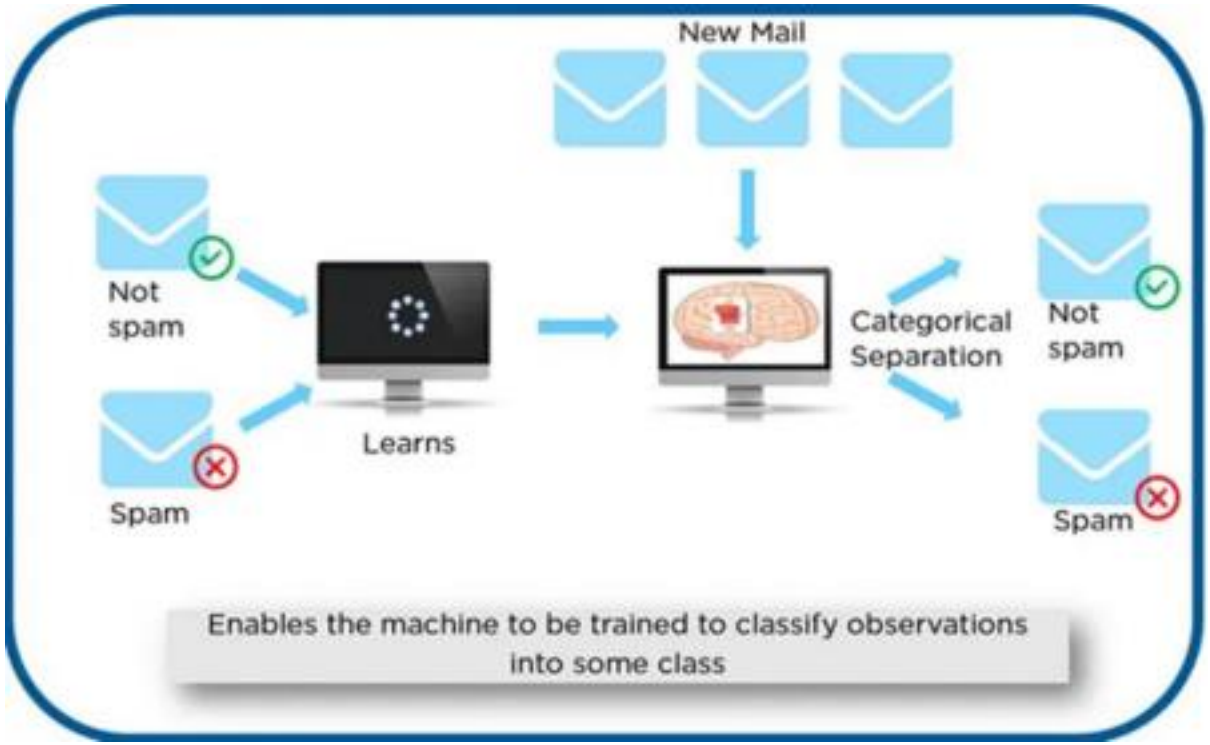


Why is machine learning? (cont.)

- With ML :



Why is machine learning? (cont.)



When should you use machine learning?

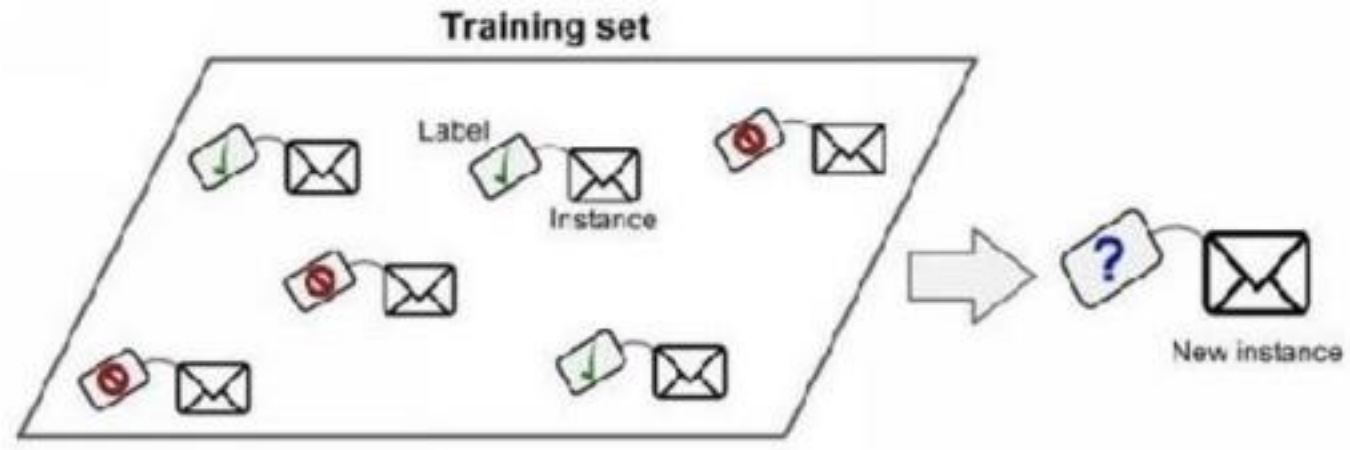
- When you have a problem that requires many long lists of rules to find the solution. In this case, machine-learning techniques can simplify your code and improve performance.
- Very complex problems for which there is no solution with a traditional approach.
- Non- stable environments: machine-learning software can adapt to new data

Types of Learning

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
- Reinforcement Learning

Supervised Learning

- In this type of machine-learning system, the data that you feed into the algorithm, with the desired solution, are referred to as “labels.”

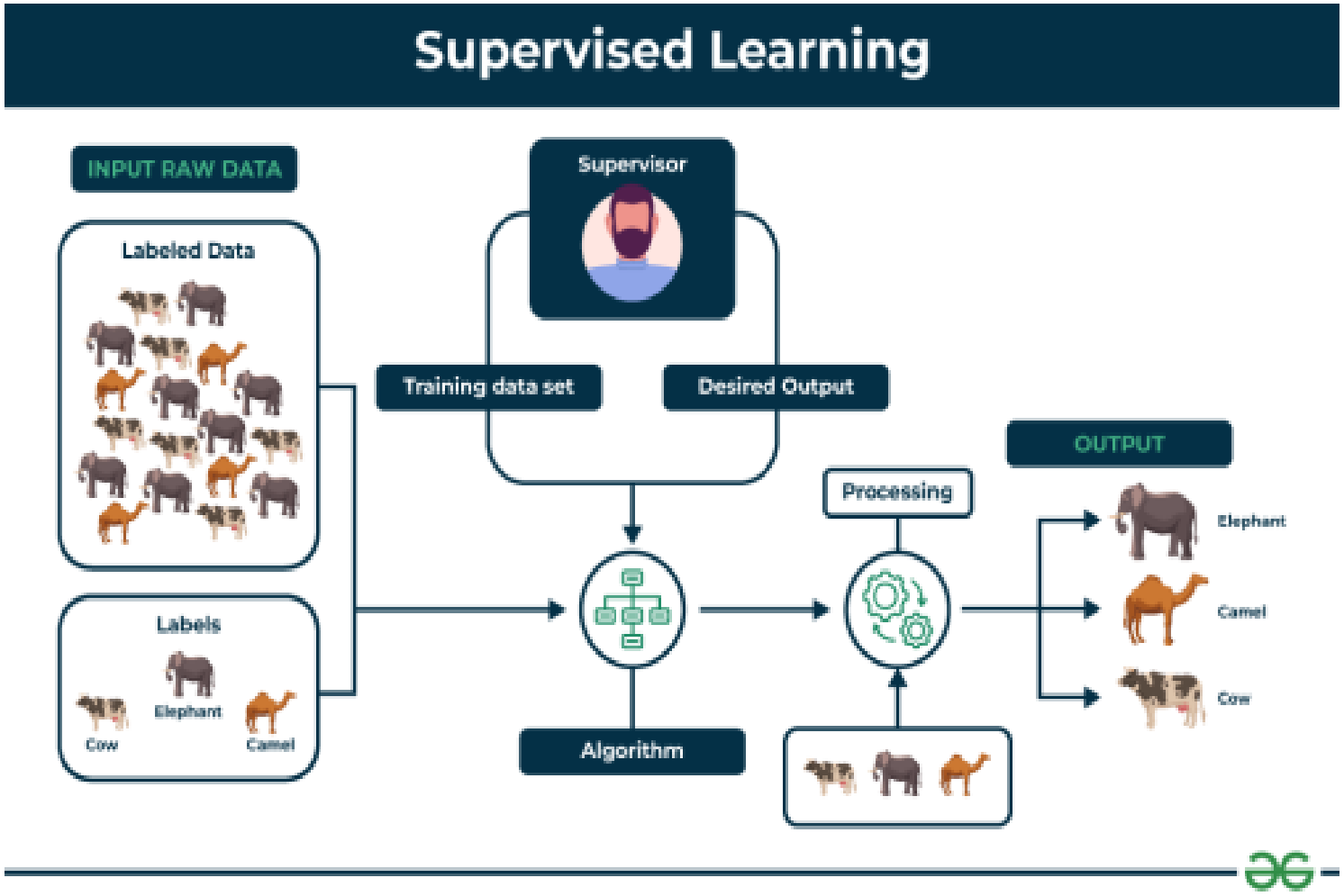


- The dataset is the collection of $\{(x_i, y_i)\}_{i=1}^{N}$ labeled examples
- Each **element x_i** among N is called a **feature vector**. A feature vector is a vector in which each dimension j = 1,...,D contains a value that describes the example somehow. That value is called **a feature** and is denoted as $x^{(j)}$
- The **label y_i** can be either an element belonging to a finite set of classes {1,2,...,C}, or a real number, or a more complex structure, like a vector, a matrix, a tree, or a graph.

Supervised Learning (cont.)

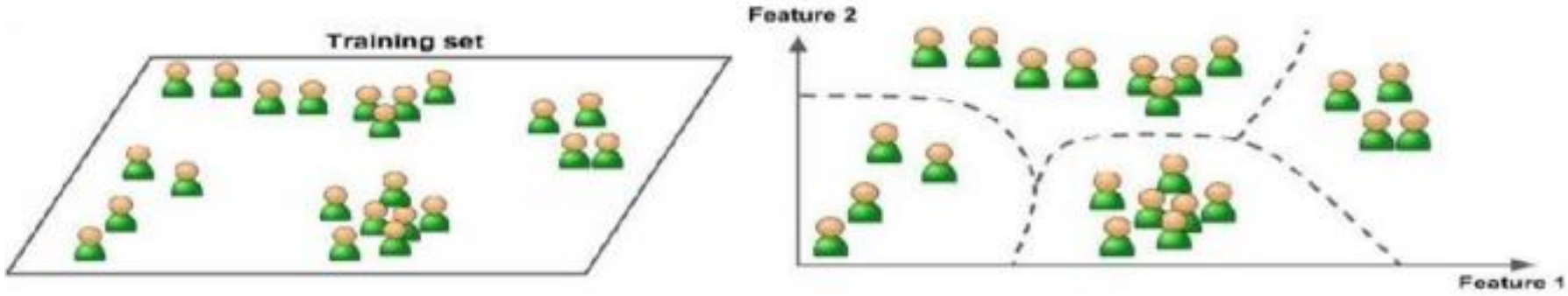
- The goal of the algorithm is to learn patterns in the data and build a general set of rules to map input to the class or event.
- There are two types commonly used as supervised learning algorithms.

 Regression Classification



Unsupervised Learning

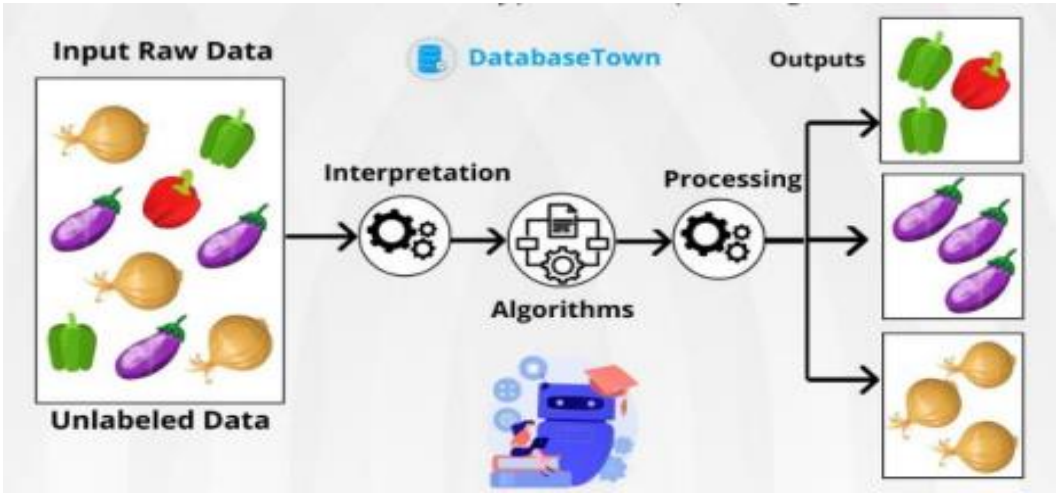
- In this type of machine-learning system, can guess that the data is **unlabeled**.
  - the dataset is a collection of unlabeled examples $\{x_i\}^N_{i=1}$. Again, x is a feature vector, and the goal of an unsupervised learning algorithm is to create a model that takes a feature vector x as input and either transforms it into another vector or into a value that can be used to solve a practical problem.



Unsupervised Learning

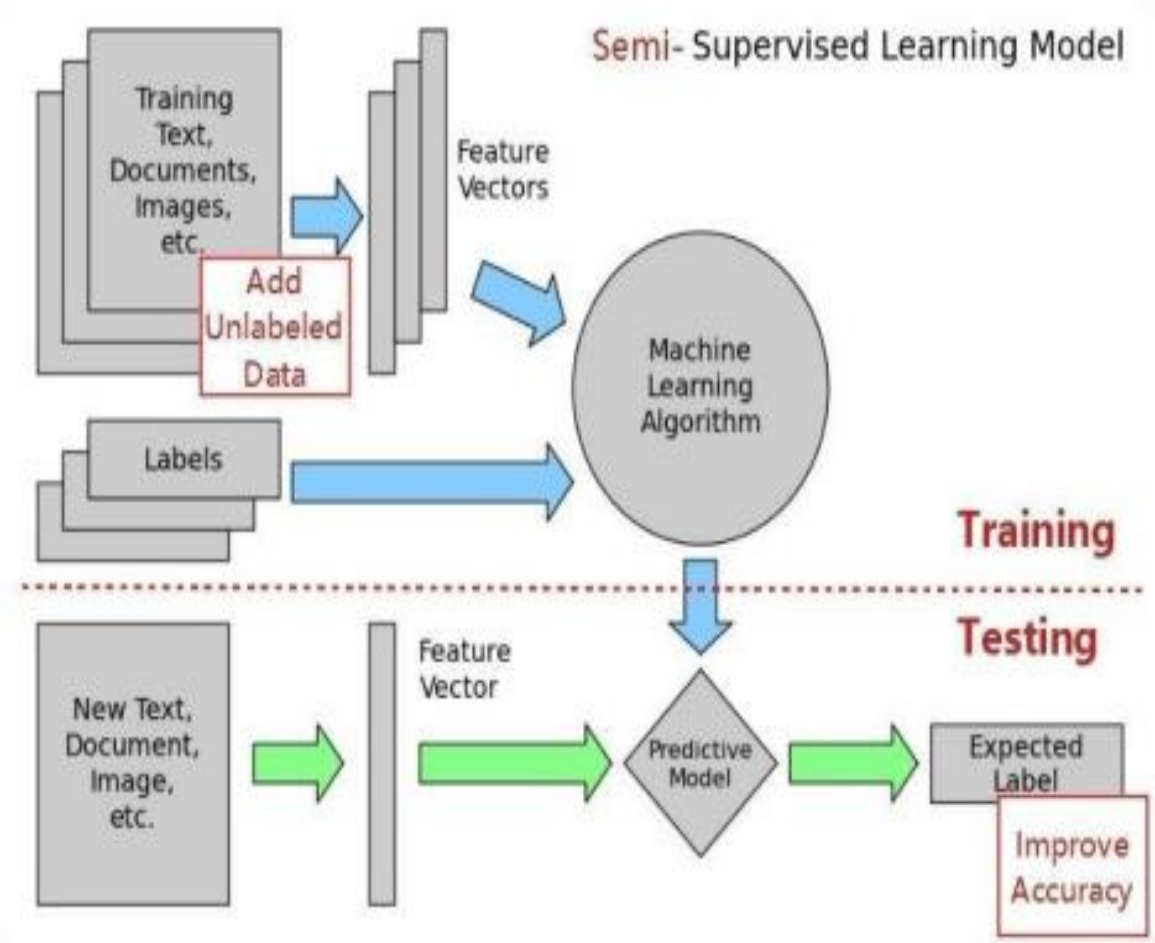
(cont.)

- There are situations where the desired output class/event is unknown for historical data. The objective in such cases would be to study the patterns in the input dataset to get better understanding and identify similar patterns that can be grouped into specific classes or events.
- There are three types commonly used as unsupervised learning algorithms.
  - Clustering
  - Dimension Reduction
  - Anomaly Detection



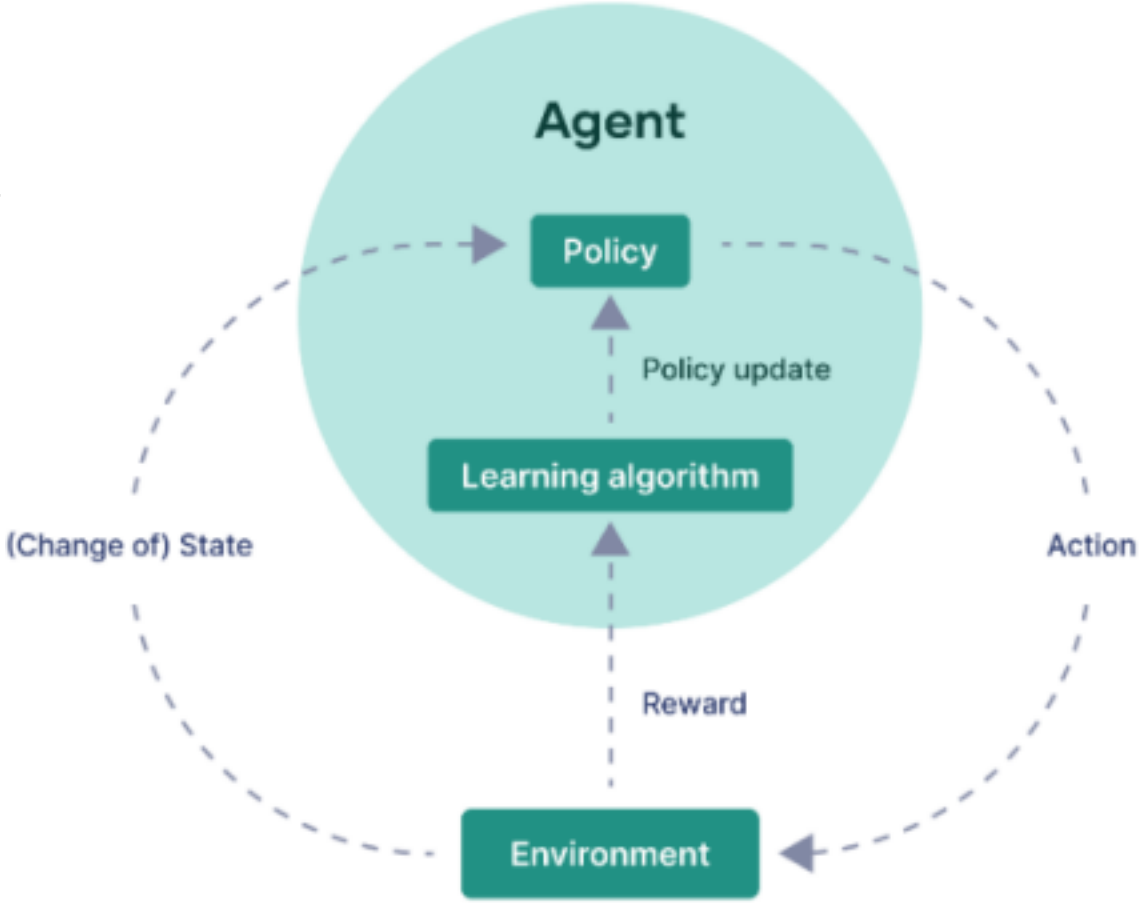
Semi-Supervised Learning

- The dataset contains both labeled and unlabeled examples. Usually, the quantity of unlabeled examples is much higher than the number of labeled examples.
- The goal of a semi-supervised learning algorithm is the same as the goal of the supervised learning algorithm. The hope here is that using many unlabeled examples can help the learning algorithm to find a better model.



Reinforcement Learning

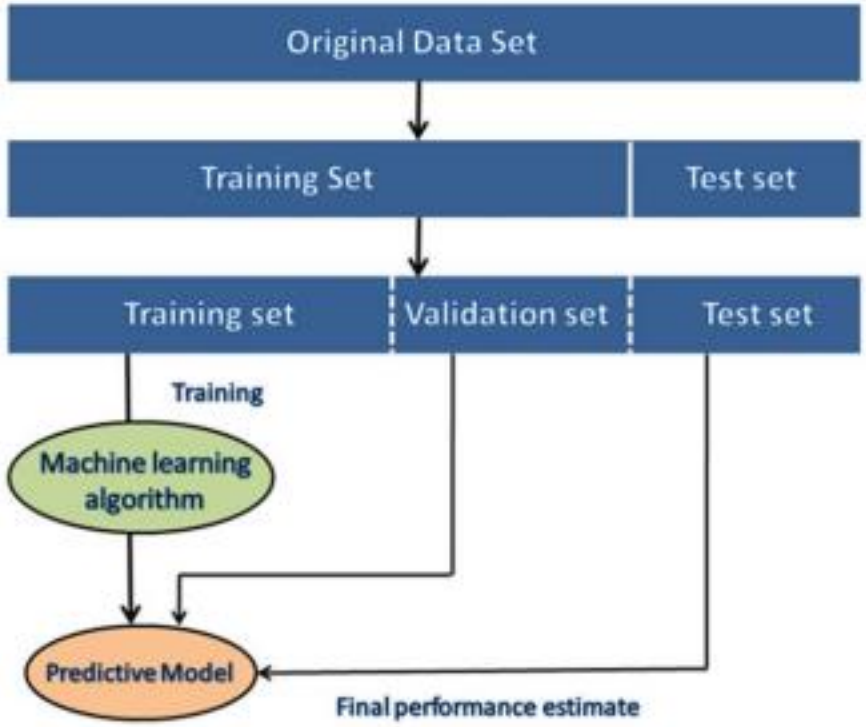
- The machine "lives" in an environment and is capable of perceiving the state of that environment as a vector of features.
- The machine can execute actions in every state. Different actions bring different rewards and could also move the machine to another state of the environment. The goal of a reinforcement learning algorithm is to learn a policy.



Reinforcement Learning (cont.)

- The goal of a reinforcement learning algorithm is **to learn a policy**.
- A policy is a function (similar to the model in supervised learning) that takes the feature vector of a state as input and outputs an optimal action to execute in that state.
- The action is optimal if it maximizes the expected average reward.
- Reinforcement learning solves a particular kind of problem where decision making is sequential, and the goal is long-term
- Examples of reinforcement learning techniques are the following:
  - Markov decision process
  - Q-learning/Q deep learning
  - Temporal Difference methods
  - Monte-Carlo methods

Basic operations of machine learning system

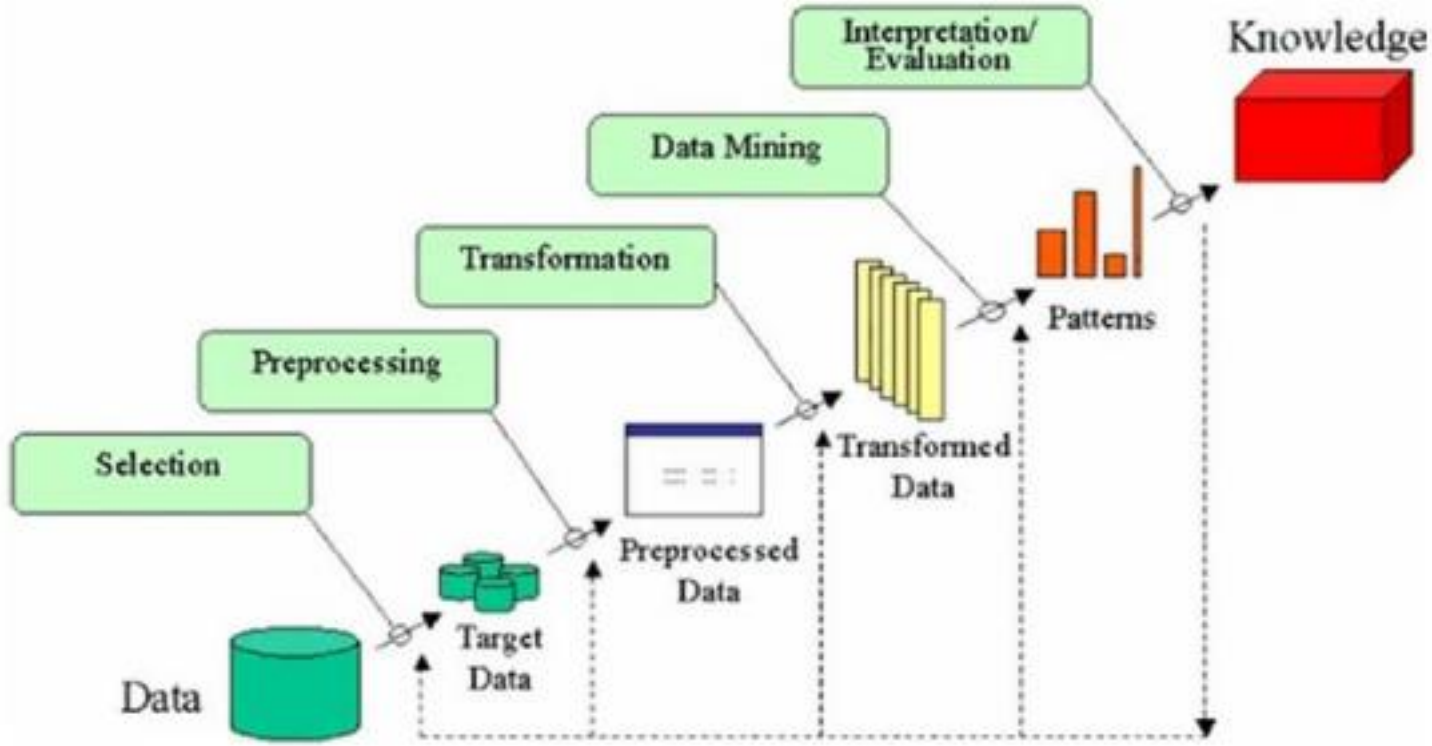


Frameworks for Building Machine Learning Systems

- Knowledge Discovery Databases (KDD) process model
- CRoss Industrial Standard Process for Data Mining (CRISP – DM)
- Sample, Explore, Modify, Model and Assess (SEMMA)

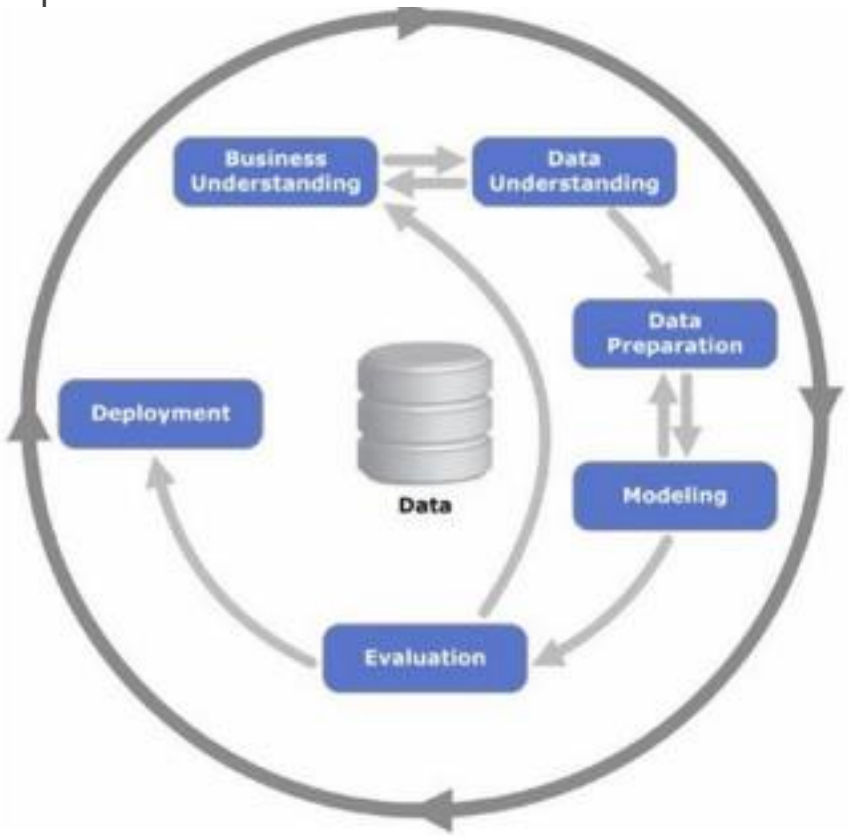
Knowledge Discovery Databases (KDD)

- This refers to the overall process of discovering useful knowledge from data, which was presented by **a book by Fayyad et al., 1996**


Interpretation/
Evaluation
Knowledge
Data Mining
Transformation
Preprocessing
Selection
Patterns
Transformed
Data
Preprocessed
Data
Target
Data
Data

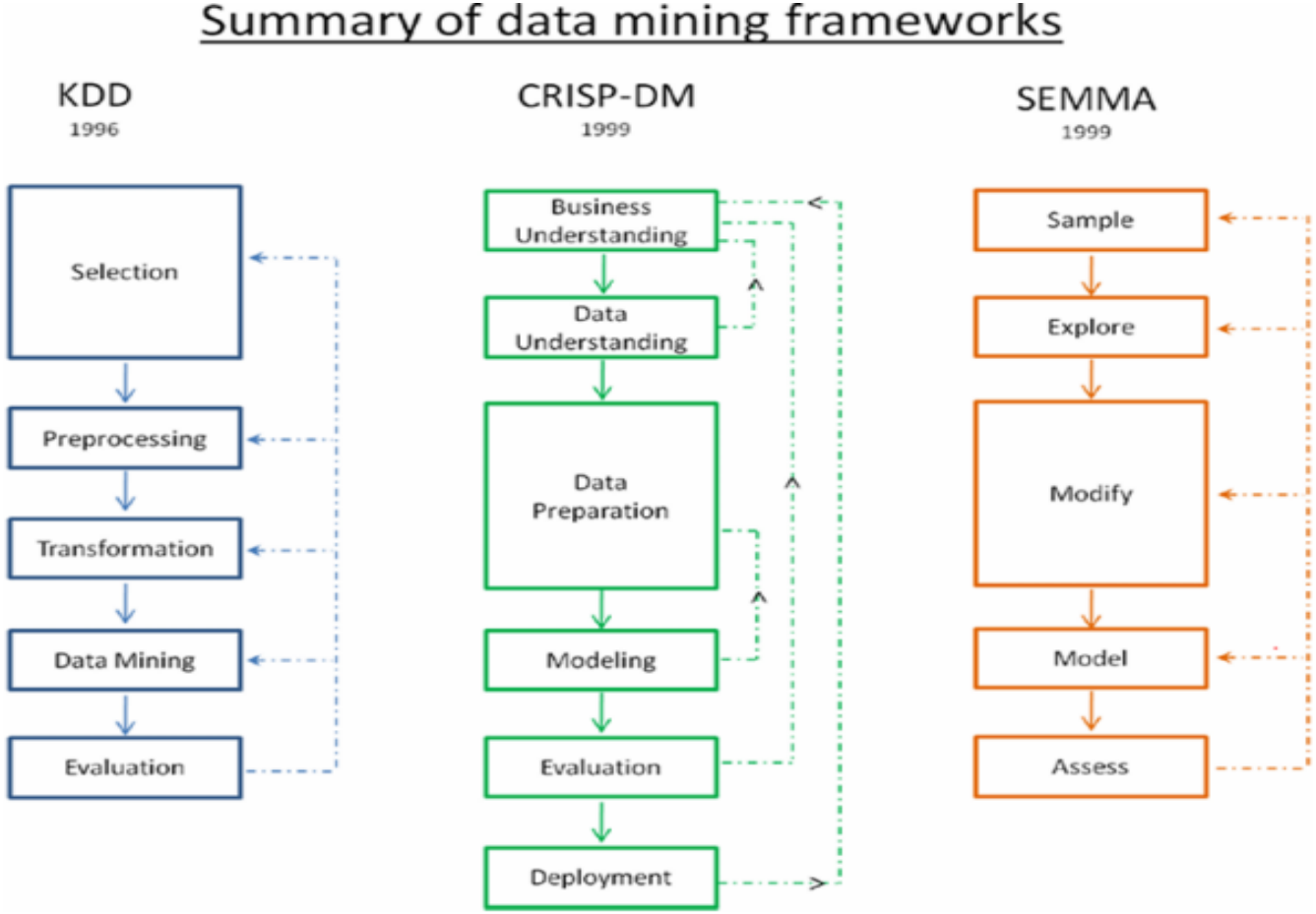
Cross-Industry Standard Process for Data Mining

 It was established by the **European Strategic Program on Research in Information Technology** initiative with an aim to create an unbiased methodology that is not domain dependent.



SEMMA (Sample, Explore, Modify, Model, Assess)

- SEMMA are the sequential steps to build machine learning models incorporated in 'SAS Enterprise Miner', a product by **SAS Institute Inc**.



Machine Learning Perspective of Data

- Data is the fact and figures (can also be referred as raw data) that we have available with respect to the business context. Data are made up of these two aspects:
  - Objects such as people, tree, animals, etc.
  - Attributes that were recorded for objects such as age, size, weight, cost, etc.
- The things we measure, control, or manipulate for objects are the variables
- The amount of information that can be provided by a variable is determined by its type of measurement scale.

Scales of Measurement

- Variables can be measured on four different scales.
- Mean, median, and mode are the way to understand the central tendency, that is, the middle point of data distribution
- Standard deviation, variance, and range are the most commonly used dispersion measures used to understand the spread of the data.

Nominal Scale of Measurement

- Data are measured at the nominal level when each case is classified into one of a number of discrete categories.

Variable Name	Example Measurement Values
Color	Red, Green, Yellow, etc.
Gender	Female, Male
Football Players Jersey Number	1, 2, 3, 4, 5, etc.

Ordinal Scale of Measurement

- Data are measured on an ordinal scale if the categories imply order.

Variable Name	Example Measurement Values
Military rank	Second Lieutenant, First Lieutenant, Captain, Major, Lieutenant Colonel, Colonel, etc.
Clothing size	Small, Medium, Large, Extra Large. Etc.
Class rank in an exam	1,2,3,4,5, etc.

Interval Scale of Measurement

- If the differences between values have meanings, the data are measured at the interval scale

Variable Name	Example Measurement Values
Temperature	10, 20, 30, 40, etc.
IQ rating	85 - 114, 115 - 129, 130 - 144, 145 – 159, etc.

Ratio Scale of Measurement

- Data measured on a ratio scale have differences that are meaningful, and relate to some true zero point.

Variable Name	Example Measurement Values
Weight	10, 20, 30, 40, 50, 60, etc.
Height	5, 6, 7, 8, 9, etc.
Age	1, 2, 3, 4, 5, 6, 7, etc.

Machine Learning Python Packages

- There is a rich number of opensource libraries available to facilitate practical machine learning.
- These are mainly known as scientific Python libraries and are generally put to use when performing elementary machine learning tasks.
- At a high level we can divide these libraries into data analysis and core machine learning libraries based on their usage/purpose
  - Data analysis packages
  - Core Machine learning packages

Data Analysis Packages

- There are four key packages that are most widely used for data analysis
  - NumPy
  - SciPy
  - Matplotlib
  - Pandas

Machine Learning Core Libraries

- Python has a plethora of opensource machine learning libraries

	Contributors				
Project Name	**2015**	**2016**	**Change (%)**	**License Type**	**Source**
Scikit-learn	404	732	81%	BSD 3	www.github.com/scikit-learn/scikit-learn
Pylearn2	117	115	-2%	BSD 3	www.github.com/lisa-lab/pylearn2
NuPIC	60	75	25%	AGPL 3	www.github.com/numenta/nupic
Nilearn	28	46	64%	BSD	www.github.com/nilearn/nilearn
PyBrain	27	31	15%	BSD 3	www.github.com/pybrain/pybrain
Pattern	20	20	0%	BSD 3	www.github.com/clips/pattern
Fuel	12	29	142%	MIT	www.github.com/mila-udem/fuel
Bob	11	13	18%	BSD	www.github.com/idiap/bob
Skdata	10	11	10%	N/A	www.github.com/jaberg/skdata
MILK	9	9	0%	MIT	www.github.com/luispedro/milk

The End

Homework

1. Give 2 real-world examples that fit the application of machine learning.

2. Find 2 real-world examples that not fit the application of machine learning.

Solution for Homework 1:

- Here are two real-world examples in which machine learning is effectively applied:

1. Fraud Detection in Banking:

Banks use machine learning models to analyze transaction data in real time. These models learn patterns of normal behavior and can quickly flag unusual transactions that may indicate fraud. This helps banks reduce financial losses and protect customers.

2. Personalized Recommendations in E-commerce:

Online retailers like Amazon and Netflix use machine learning to analyze user behavior, such as past purchases or viewing history. The algorithms then suggest products or movies that align with the user's interests, improving user engagement and sales.

Solution for Homework 2:

1. Deterministic Calculations:

Consider a basic calculator that performs arithmetic operations like addition, subtraction, multiplication, or division. These operations follow clear, deterministic rules with known outcomes, making it unnecessary to apply a learning-based approach when a simple algorithm will always yield the correct result.

2. Fixed Rule-Based Systems:

Many systems rely on hard-coded business rules where conditions and outcomes are explicitly defined. For instance, determining tax brackets or calculating shipping rates based on predetermined rules doesn’t require machine learning since the logic is static and easily implemented with conventional programming techniques.