# Chapter 2 - Introduction of ML

By Yeonjung Lee

## 1. Machine Learning

Machine Learning (ML) is a central application and subfield of Artificial Intelligence that enables computers to learn from data and improve their performance over time without being explicitly programmed for every task. Instead of relying on fixed rules, machine learning systems analyze data, identify patterns, and make decisions autonomously. This capability allows AI systems to adapt to new information and perform effectively in dynamic and complex environments.

### 1) Definition and Overview

Machine learning allows machines to take input data, analyze it, and generate predictions or decisions based on patterns learned from previous examples. Many machine learning systems rely on neural networks, which are computational models inspired by the structure and functioning of the human brain. These models consist of interconnected units that process data through algorithms, gradually refining their outputs as more data becomes available.

Within the broader field of Artificial Intelligence, machine learning represents a specific approach to intelligence—one that emphasizes learning from experience. Rather than being programmed with explicit instructions for every situation, machine learning systems improve their behavior by adapting to data over time.

### 2) Deep Learning

Deep learning is a specialized subset of machine learning that uses large, multi-layered neural networks. The primary distinction between traditional machine learning and deep learning lies in the depth and complexity of the models employed.

Deep learning networks contain many hidden layers, allowing them to process vast and highly complex datasets. These models are particularly effective in areas such as image recognition, speech processing, natural language understanding, and large-scale data analysis. By leveraging deep learning techniques, AI systems can extract detailed features and uncover intricate patterns that simpler models are unable to detect. This capability has driven many of the recent breakthroughs in modern AI applications.

### 3) Learning and Adaptation

Machine learning and deep learning systems are inherently adaptive. Each prediction or decision made by the system generates feedback, which can be used to improve future performance. Over


© 2025 Yeonjung Lee

time, the system refines its internal parameters and algorithms, becoming increasingly accurate, efficient, and reliable.

There are three primary learning paradigms commonly used in machine learning: supervised learning, unsupervised learning, and reinforcement learning. Each approach differs in how data is presented to the system and how learning takes place.

## 2. Types of Machine Learning

Machine Learning can be broadly classified into three main categories based on the level of guidance provided during the learning process and the nature of the data involved.

### 1) Supervised Learning

Supervised learning is the most structured and guided form of machine learning. In this approach, the system is trained using labeled data, meaning that each input is paired with a known and correct output. The goal is for the machine to learn a mapping between inputs and outputs and to minimize errors during prediction.

Supervised learning is commonly used for classification and regression tasks, such as spam email detection, stock price prediction, and medical diagnosis. Because the correct answers are provided during training, the system receives clear guidance on how to improve its predictions.

For example, when a machine is trained on images of cats and dogs that are correctly labeled, it learns to distinguish features associated with each category. Once trained, the system can accurately classify new, unseen images as either cats or dogs.

### 2) Unsupervised Learning

Unsupervised learning is a more exploratory approach in which the system is given input data without corresponding output labels. Without explicit guidance, the machine must analyze the data to discover hidden patterns, structures, or relationships on its own.

This type of learning is particularly useful for clustering, association analysis, and anomaly detection. Since there are no predefined correct answers, unsupervised learning carries a higher risk of ambiguity or error. However, it is highly valuable when labeled data is unavailable or expensive to obtain.

For instance, a machine may be given a large dataset containing customer purchasing behavior without any predefined categories. Using unsupervised learning techniques, the system can group customers with similar behavior patterns, enabling businesses to tailor marketing strategies and improve customer engagement.

DRAFT


© 2025 Yeonjung Lee

### 3) Reinforcement Learning

Reinforcement learning (RL) focuses on learning through interaction with an environment. In this paradigm, an agent performs actions and receives feedback in the form of rewards or penalties. The objective is to learn a strategy, or policy, that maximizes cumulative rewards over time.

Reinforcement learning does not always rely on predefined outputs, making it distinct from supervised learning. Instead, the system learns which actions are beneficial by evaluating the consequences of its behavior. This approach is widely used in applications such as game playing, robotics, autonomous driving, and recommendation systems.

An example of reinforcement learning is a robot navigating a maze. The robot receives a reward when it reaches the exit and a penalty when it hits a wall. Through repeated trials, it learns the most efficient path to the goal. Because reinforcement learning combines elements of guided feedback and independent exploration, it is sometimes considered a hybrid or semi-supervised learning approach.

## 3. Importance of Machine Learning in AI

Machine learning forms the backbone of many modern AI applications, including autonomous vehicles, virtual assistants, recommendation systems, and medical diagnostic tools. By enabling machines to learn from data and adapt to new situations, machine learning allows AI systems to handle increasingly complex, uncertain, and real-world problems. As data availability and computational power continue to grow, machine learning remains a driving force behind the advancement of Artificial Intelligence.

### 1) Examples of Machine Learning in Real Life

Machine Learning has become an integral part of everyday life, powering many systems and applications that people interact with on a daily basis. By learning from data and user behavior, ML-driven systems continuously improve their performance, efficiency, and personalization. The following examples illustrate how machine learning is applied across various real-world domains.

#### A. Smart Home Assistants

Smart home assistants such as Google Assistant, Siri, and Alexa are among the most familiar examples of machine learning in action. These systems rely on speech recognition to understand spoken commands and natural language processing to interpret user intent. Over time, they learn from user routines and preferences to provide personalized reminders, recommendations, and responses.

In addition, smart assistants integrate with other devices and services to automate everyday tasks, such as controlling smart lights, setting alarms, managing calendars, and playing music. Through


© 2025 Yeonjung Lee

continuous software updates and improvements in deep learning models, these assistants become more accurate and responsive. As a result, they function as intelligent systems that adapt to individual users, making daily life more convenient and efficient.

### B. Social Media Platforms

Social media platforms such as Facebook, Instagram, and Twitter make extensive use of machine learning to enhance user engagement. ML algorithms analyze user behavior, including likes, shares, comments, and browsing patterns, to personalize content feeds.

These platforms also use machine learning to determine which advertisements, posts, and recommendations are most relevant to individual users. As interactions accumulate, the systems continuously refine their models, improving the accuracy of content ranking and recommendations. This demonstrates how machine learning can process massive volumes of data and tailor experiences to millions of users simultaneously.

### C. Translation Services

Machine learning plays a critical role in modern language translation services, such as Google Translate. These systems use neural networks to analyze sentences and predict accurate translations between languages. Rather than translating word by word, ML models consider grammar, vocabulary, sentence structure, and context to produce coherent and meaningful translations.

Deep learning techniques allow translation systems to handle complex language patterns, idiomatic expressions, and variations in syntax. As the system is exposed to more multilingual data, its performance continues to improve. This application highlights the power of machine learning in understanding and processing human language at scale.

### D. Autonomous Vehicles

Autonomous or self-driving vehicles depend heavily on machine learning for safe and efficient operation. ML algorithms enable vehicles to recognize traffic signs, detect pedestrians, identify obstacles, and interpret road conditions. Neural networks process data from cameras, sensors, and radar systems in real time to support driving decisions.

Through continuous learning, autonomous vehicles improve their ability to respond to dynamic and unpredictable traffic environments. This example illustrates how machine learning enables systems to interact with the physical world, combining perception, decision-making, and adaptation in complex real-world scenarios.

DRAFT


© 2025 Yeonjung Lee

# 4. Relationship Between Machine Learning, Artificial Intelligence, and Data Science

Machine Learning, Artificial Intelligence, and Data Science are closely interconnected fields, each playing a distinct but complementary role in the development of intelligent systems. Understanding their relationships is essential for grasping how modern AI applications function.

### 1) Data Science

Data Science is the broad discipline concerned with the entire lifecycle of data. It involves collecting raw data from various sources, cleaning and preprocessing the data to ensure quality, analyzing it using statistical and computational methods, and interpreting the results to support decision-making.

The primary goal of Data Science is to transform raw data into actionable knowledge. This knowledge serves as the foundation upon which machine learning models are trained and AI systems operate.

### 2) Machine Learning

Machine Learning is a subset of Artificial Intelligence that focuses on creating systems capable of learning from data. Unlike traditional software programs that follow fixed rules, ML systems adapt automatically by identifying patterns in data.

Machine learning algorithms analyze input data to make predictions, detect trends, or support decision-making without being explicitly programmed for each task. In this sense, machine learning acts as the engine that enables AI systems to learn from experience and improve performance over time.

### 3) Artificial Intelligence

Artificial Intelligence is the broadest field, aiming to create machines that simulate aspects of human intelligence. AI encompasses reasoning, problem-solving, perception, language understanding, and decision-making.

Machine learning plays a critical role within AI by providing the mechanisms that allow systems to adapt and learn. While not all AI systems rely on machine learning, most modern AI applications are powered by ML techniques.

To sum up, Data Science provides the data and analytical insights needed to train models. Machine Learning uses this data to learn patterns and improve performance. Artificial Intelligence integrates these learned capabilities to perform intelligent tasks. Together, these fields form the backbone of modern intelligent systems.

DRAFT


© 2025 Yeonjung Lee

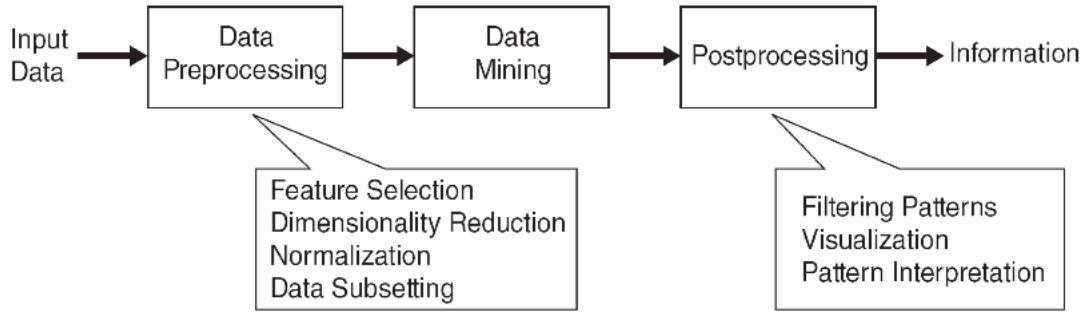
# 5. Data Mining in the Machine Learning

## 1) Workflow

Data mining is a key component of the machine learning workflow and focuses on extracting valuable information from large datasets. It involves discovering patterns, relationships, and trends that are not immediately obvious.

Data mining has been defined as the non-trivial extraction of implicit, previously unknown, and potentially useful information from data. It also refers to the exploration and analysis of large datasets, using automatic or semi-automatic methods, to uncover meaningful patterns.

The purpose of data mining is to provide structured insights that feed machine learning algorithms, enabling them to learn, predict outcomes, and optimize decisions. Common examples include analyzing sales data to identify customer purchasing patterns or detecting fraudulent transactions in banking systems.




# 6. What Is Data?

Data forms the foundation of Artificial Intelligence, Machine Learning, and Data Science. A clear understanding of what data is and how it is structured is essential for building effective intelligent systems.

### 1) Definition

Data can be defined as a collection of data objects and their associated attributes. Data objects represent the entities or items of interest, while attributes describe the properties or characteristics of those objects.

### 2) Attributes of Data

An attribute is a specific property or characteristic that provides meaningful information about a data object. Attributes can take many forms depending on the context, such as a person's eye


© 2025 Yeonjung Lee

color, temperature measurements, or a product's price and weight. Attributes are also commonly referred to as variables, fields, characteristics, dimensions, or features. Together, attributes define the information used by machines to analyze patterns and make predictions.

### 3) Objects in Data

A data object represents a single entity described by a set of attributes. Data objects are often referred to as records, points, cases, samples, entities, or instances. For example, in a dataset containing student information, each student represents a data object, while attributes may include name, age, grade, gender, and hometown. The combination of objects and attributes enables machines to analyze data systematically, identify patterns, and perform intelligent tasks.

## 7. Types of Attributes (Data)

Attributes describe the properties or characteristics of data objects. Depending on the nature of their values and the operations that can be meaningfully applied to them, attributes can be classified into several types.

### 1) Nominal Attributes

Nominal attributes represent categories or labels with no inherent ordering. For this type of attribute, only distinctness matters, and mathematical operations such as addition, subtraction, or comparison are not meaningful. Nominal attributes are commonly used to represent names, labels, or categories. Examples include ID numbers, eye color, and ZIP codes.

### 2) Ordinal Attributes

Ordinal attributes represent categories that have a meaningful order, but the differences between values cannot be precisely measured. In ordinal attributes, both distinctness and order are meaningful, while differences and ratios are not. Ordinal data allows comparison of relative positions but does not support precise measurement. Examples include rankings such as taste ratings on a scale from 1 to 10, letter grades (A, B, C), and height categories such as short, medium, and tall.

### 3) Interval Attributes

Interval attributes consist of ordered values where the difference between values is meaningful, but there is no true zero point. These attributes support distinctness, order, and meaningful differences, but ratios are not meaningful. Examples include calendar dates and temperature measured in Celsius or Fahrenheit. For instance, a temperature of 20°C is not considered twice as hot as 10°C.

DRAFT


© 2025 Yeonjung Lee

4) Ratio Attributes

Ratio attributes have all the properties of interval attributes, along with a meaningful zero point. This means that distinctness, order, meaningful differences, and meaningful ratios are all supported. Ratio attributes allow all standard mathematical operations. Examples include temperature measured in Kelvin, length, weight, counts, and elapsed time, such as the time taken to complete a race.

## 8. Properties of Attribute Values

The type of an attribute is determined by which properties or operations are meaningful for its values. These properties define how the data can be compared, ordered, and analyzed mathematically.

- Distinctness refers to the ability to determine whether two values are equal or not equal ( $=, \neq$ ).
- Order indicates whether values can be ranked or compared using greater-than or less-than relationships ( $<, >$ ).
- Meaningful differences indicate whether subtraction or addition operations produce meaningful results ( $+, -$ ).
- Meaningful ratios indicate whether multiplication or division operations are meaningful ( $\times, \div$ ).

| Attribute Type | Distinctness | Order | Differences | Ratios |
|---|---|---|---|---|
| Nominal | ✓ | ✗ | ✗ | ✗ |
| Ordinal | ✓ | ✓ | ✗ | ✗ |
| Interval | ✓ | ✓ | ✓ | ✗ |
| Ratio | ✓ | ✓ | ✓ | ✓ |

Based on these properties, attributes are classified into four main types.

Nominal attributes support only distinctness. Values can be identified as the same or different, but they cannot be ordered or used in arithmetic operations.

Ordinal attributes support both distinctness and order. Values can be ranked, but the differences between them and their ratios are not meaningful.

Interval attributes support distinctness, order, and meaningful differences. However, they do not support meaningful ratios because there is no true zero point.


© 2025 Yeonjung Lee

Ratio attributes support all four properties: distinctness, order, meaningful differences, and meaningful ratios. These attributes have a true zero, allowing all standard mathematical operations to be applied meaningfully.

## 9. Discrete and Continuous Attributes

Attributes can also be classified based on the nature of the values they take. Two important categories are discrete attributes and continuous attributes.

### 1) Discrete Attributes

Discrete attributes take on a finite or countably infinite set of values. These values are often represented as integers. A special case of discrete attributes is binary attributes, which can take only two possible values, such as yes/no or true/false.

Examples of discrete attributes include ZIP codes, counts such as the number of purchases, and the set of words in a document collection.

### 2) Continuous Attributes

Continuous attributes can take any real-valued number within a given range. These attributes are typically represented using floating-point numbers. In practice, continuous values are measured with limited precision due to measurement constraints.

Examples of continuous attributes include temperature, height, and weight.

## 10. Origins of Data Mining

Data mining originates from the integration of ideas and techniques from multiple disciplines. These include machine learning and artificial intelligence, pattern recognition, statistics, and database systems.

Traditional data analysis techniques often become inadequate when dealing with modern data challenges. Such challenges include large-scale datasets, high dimensionality, heterogeneous and complex data, and data that is distributed across multiple sources. To address these challenges, data mining has emerged as a key component of data science and data-driven discovery, enabling the extraction of meaningful patterns and knowledge from complex and large datasets.

## 11. Data Mining Tasks

Data mining tasks are generally divided into two major categories:


© 2025 Yeonjung Lee

- Prediction Methods - Prediction methods use known variables to predict unknown or future values.
- Description Methods - Description methods focus on discovering human-interpretable patterns that describe the data.

1) Predictive Modeling: Classification

Classification involves learning a model that predicts a *class attribute* based on other attributes.

- Examples of Classification Tasks - Classification is widely applied across many domains. Common examples include detecting fraudulent versus legitimate credit card transactions, classifying land cover types using satellite imagery, categorizing news articles into topics such as sports or finance, identifying network intrusions, predicting tumors as benign or malignant, and classifying protein secondary structures.

- Classification Applications

a. Fraud Detection

The goal of fraud detection is to predict fraudulent credit card transactions. This is achieved by using transaction details and customer behavior as attributes, labeling historical transactions as fraudulent or legitimate, and training a model to detect fraud in new transactions.

b. Customer Churn Prediction

Customer churn prediction aims to identify whether a customer is likely to leave for a competitor. The approach involves analyzing call frequency, call duration, time of calls, financial status, and demographic information. Customers are labeled as loyal or disloyal, and a predictive model is built to estimate customer loyalty.

c. Sky Survey Cataloging

The goal of sky survey cataloging is to classify celestial objects as stars or galaxies. Large telescope images are analyzed, visual features are extracted from segmented images, and classification models are built. This approach has successfully led to the discovery of new high red-shift quasars.

2) Regression

Regression predicts the value of a continuous variable based on the values of other variables using linear or nonlinear dependency models. Typical applications include predicting product sales based on advertising expenditure, predicting wind speed from temperature and pressure, and time-series prediction of stock market indices.


© 2025 Yeonjung Lee

3) Clustering

Clustering is the task of grouping objects such that objects within the same group are similar to one another, while objects in different groups are dissimilar.

- Applications of Clustering - Clustering is commonly used for customer profiling in targeted marketing, document organization and browsing, gene and protein analysis, and stock market analysis.
- Clustering Applications
  a. Market Segmentation

     The goal of market segmentation is to identify distinct customer groups. This is achieved by collecting geographic and lifestyle attributes, clustering similar customers, and evaluating cluster quality using purchasing behavior.

  b. Document Clustering

     Document clustering aims to group similar documents. The process involves identifying frequently occurring terms, measuring document similarity, and applying clustering algorithms.

4) Association Rule Discovery

Association rule discovery identifies dependency rules that predict the occurrence of an item based on the occurrence of other items within a set of records.

- Applications - Association analysis is widely used in market-basket analysis, shelf and inventory management, telecommunication alarm diagnosis, and medical diagnosis through symptom and test result analysis.

5) Deviation, Anomaly, and Change Detection

This task focuses on identifying significant deviations from normal behavior. Applications include credit card fraud detection, network intrusion detection, sensor network monitoring, and environmental change detection such as deforestation.

## 12. Motivating Challenges in Data Mining

Modern data mining faces several key challenges, including scalability, high dimensionality, heterogeneous and complex data, data ownership and distribution, and the need for non-traditional analysis techniques.


© 2025 Yeonjung Lee