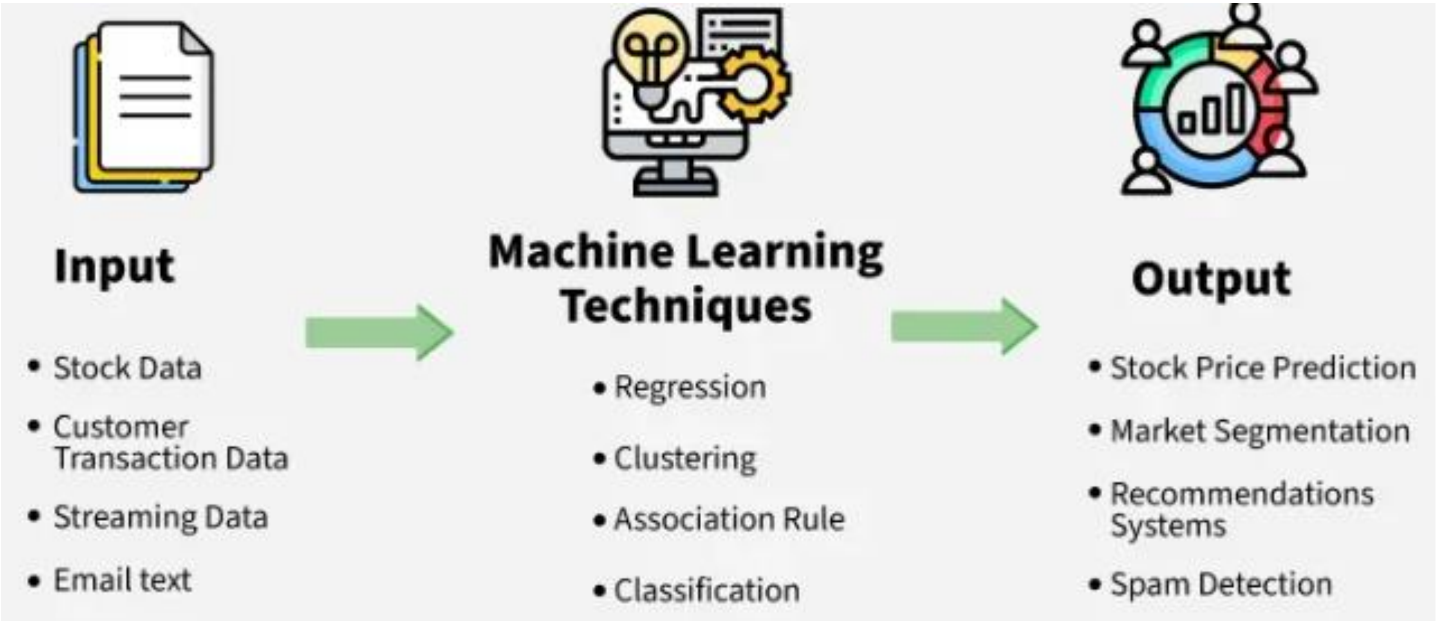
Machine Learning

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that enables computers to learn from data without being explicitly programmed. Instead of following fixed rules, ML models analyze patterns in data and make predictions or decisions based on that data

Example:

- A spam filter in your email learns to separate spam from non-spam by analyzing patterns in previous emails.
- A weather prediction model learns from past climate data to forecast future temperatures.



Traditional Programming vs Machine Learning

Follows predefined rules written by a programmer.

Input → 🔹 Predefined Rules (Logic/Code) → 🔹 Output

Example:

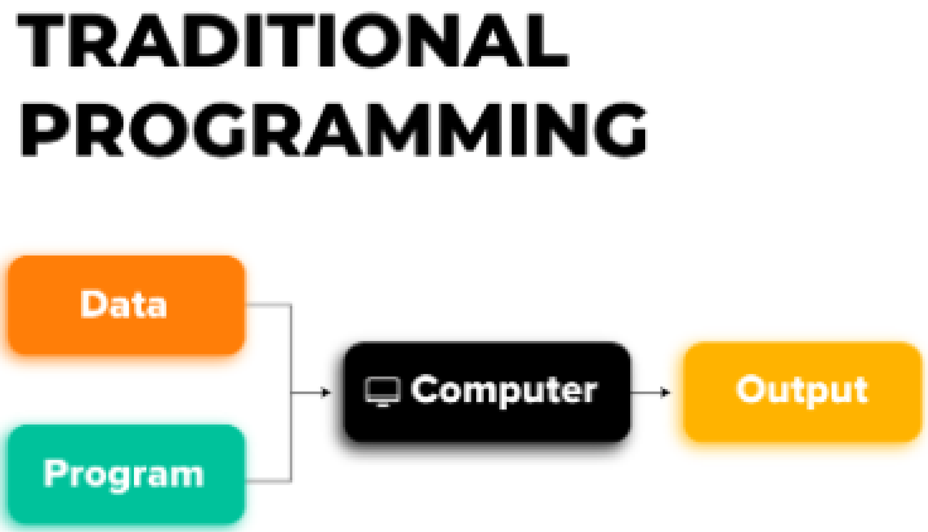
- If email contains "Free Money" → Mark as Spam
- If email is from a trusted sender → Mark as Not Spam

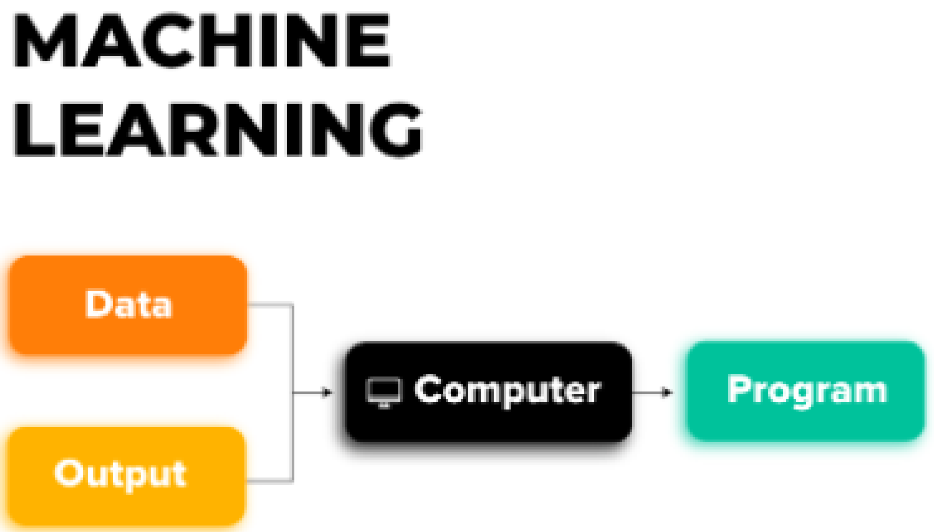
Learns from data and makes predictions.

Input Data → 🔹 Training Model (Learns Patterns) → Trained Model → 🔹 Prediction (Output)

Example:

- Analyzes thousands of emails → Learns spam patterns → Predicts spam/not spam for new emails.





Traditional Programming vs Machine Learning

1. How It Works

Traditional Programming: The programmer writes explicit rules and logic.
Machine Learning: The computer learns from data and identifies patterns.

2. Example: Spam Email Detection

Traditional Programming:
Define rules like:

- If an email contains **"Win a prize"** → Mark as spam.
- If an email contains **"Congratulations"** → Mark as spam.
- Needs manual updates when new spam words appear.

Machine Learning:

- The model analyzes thousands of emails labeled as spam or not spam.
- It learns from patterns (words, sender, links) and predicts spam even if new phrases appear.

3. Flexibility

Traditional Programming: Works well for tasks with clear, fixed rules.
Machine Learning: Works well for complex tasks where rules are difficult to define.

4. Handling New Situations

Traditional Programming: Needs manual updates when new scenarios arise.
Machine Learning: Learns from new data and adapts automatically.

5. Best Used For

Traditional Programming: Simple tasks like calculators, payroll systems, and form validation.
Machine Learning: Complex tasks like image recognition, fraud detection, and recommendation systems

Importance of Machine Learning in the Modern World

Machine Learning is transforming industries by making systems smarter, more efficient, and more autonomous.

Key Reasons Why ML is Important:

1. Automation of Tasks – Reduces human effort (e.g., chatbots, self-driving cars).
2. Better Decision-Making – Helps businesses make data-driven decisions (e.g., fraud detection in banks).
3. Improves Accuracy Over Time – ML models continuously learn and improve (e.g., recommendation systems like Netflix).
4. Personalization – Provides customized experiences (e.g., personalized ads on Google and Facebook).

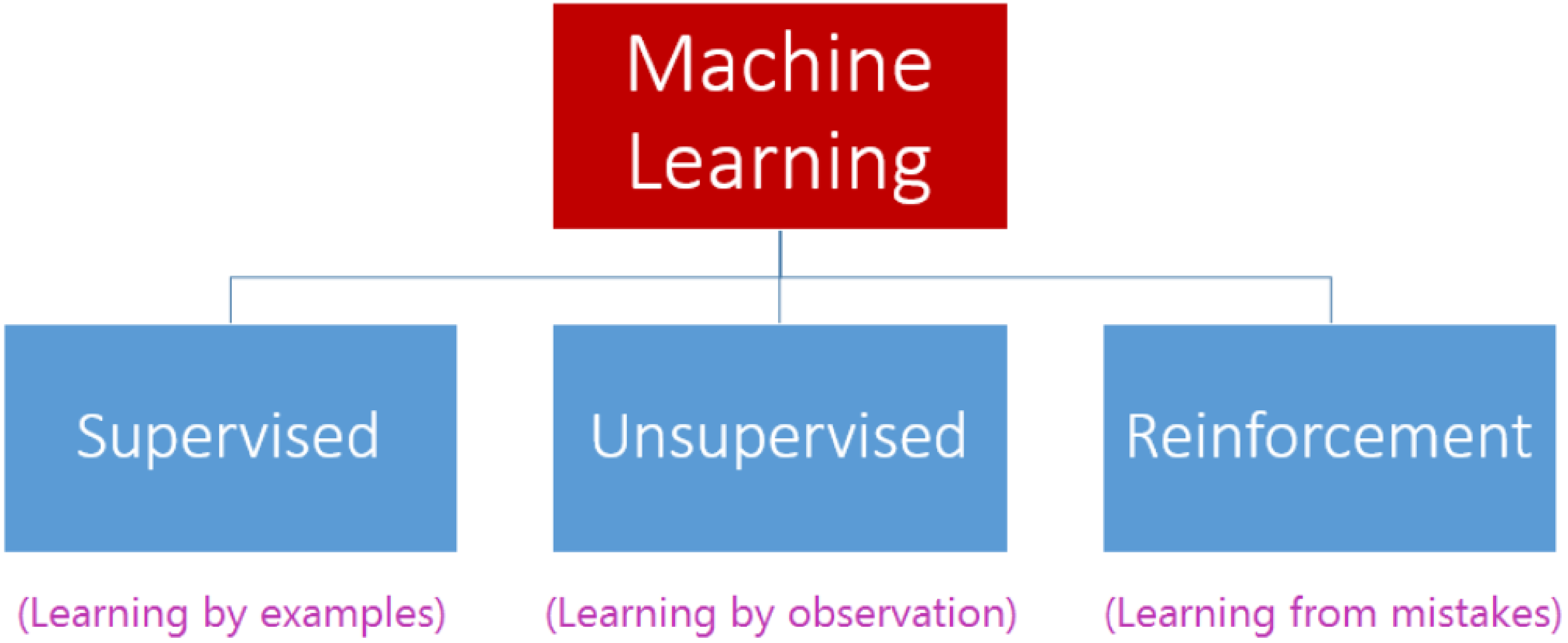
Real-World Impact:

- Healthcare: AI-powered tools detect diseases like cancer from X-ray images.
- Finance: ML detects fraudulent transactions in banking.
- E-commerce: Amazon recommends products based on your browsing history.

Difference Between AI, ML, DL, Data Analytics, and Data Science

Concept	Definition	How It Works	Example
Artificial Intelligence (AI)	The broad field of creating smart machines that can perform tasks like humans.	Uses rules, logic, and learning to make decisions.	A chatbot answering customer questions.
Machine Learning (ML)	A subset of AI where computers learn from data without explicit programming.	Finds patterns in data and makes predictions.	Email spam detection.
Deep Learning (DL)	A subset of ML that uses neural networks to learn complex patterns.	Mimics the human brain with layers of artificial neurons.	Self-driving cars recognizing objects.
Data Analytics	The process of analyzing raw data to find useful insights.	Uses statistics and tools to interpret past data.	Finding which product sells best in a store.
Data Science	A field that combines AI, ML, and analytics to extract knowledge from data.	Involves data collection, processing, and model building.	Predicting stock market trends using historical data.

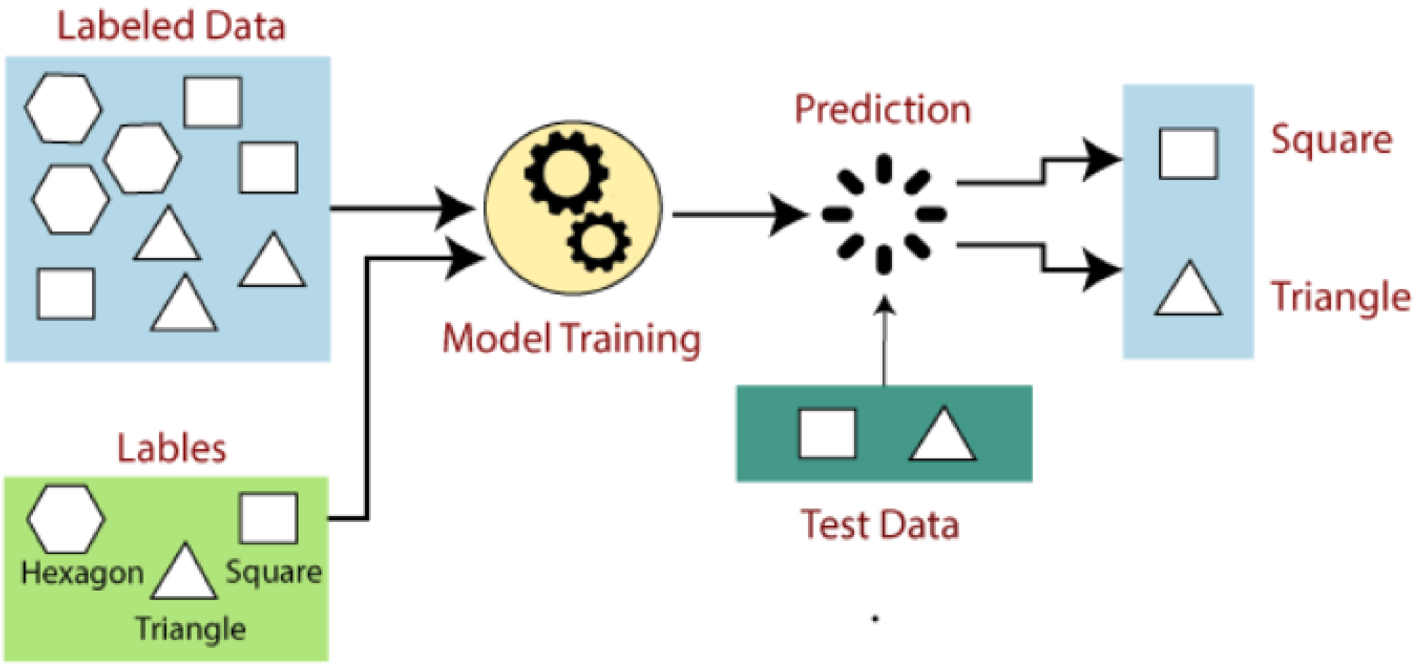
Types of Machine Learning



Supervised Learning

Supervised Learning is a type of Machine Learning where an algorithm learns from labelled data. It means that the dataset used for training has input features (X) and corresponding correct outputs (Y).

The goal is to teach the model to find patterns in the data so that it can predict outcomes for new, unseen data.



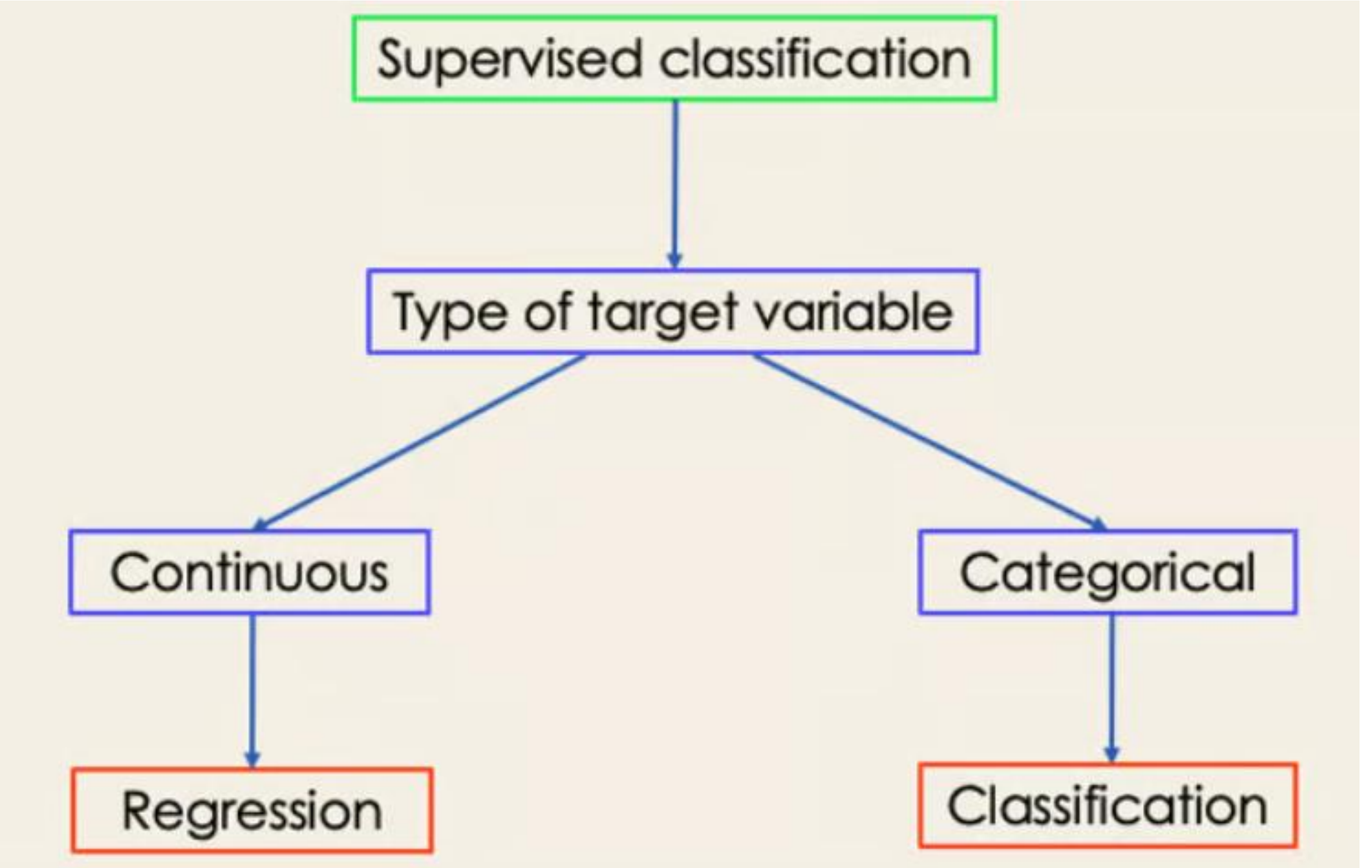
Imagine you are teaching a child to recognize different fruits.

1. You show the child pictures of apples, bananas, and oranges (input features).
2. You tell them which fruit is which (labels/output).
3. The child learns by looking at examples.
4. Later, when you show a new fruit, the child can correctly name it based on past learning.

This is exactly how Supervised Learning works- learning from labelled examples and then making predictions

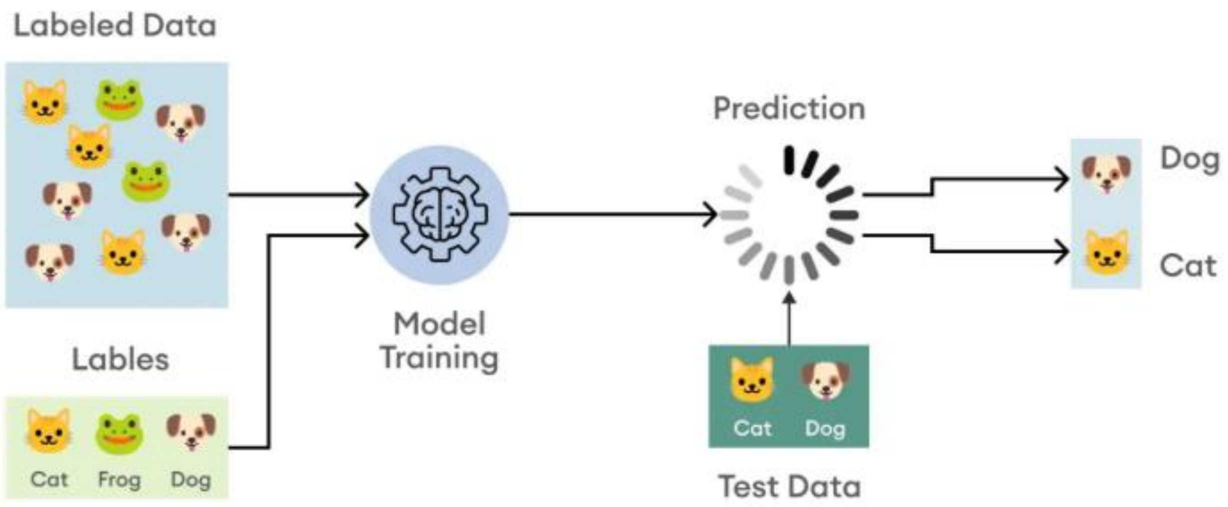
Types of Supervised Machine Learning

Supervised Machine Learning is broadly categorized into two main types



Classification

Classification is a type of supervised learning where the goal is to categorize data into predefined groups or labels. It helps in making decisions based on input data by assigning it to one of several classes.



Tools & Libraries: **Python, Scikit-learn, Pandas, Matplotlib, Seaborn, TensorFlow/Keras**

some commonly used classification algorithms

Linear Classification Algorithms
- Logistic Regression
- Linear Discriminant Analysis (LDA)

Probabilistic Classification Algorithms:
- Naïve Bayes (Gaussian, Multinomial, Bernoulli)

Support Vector-Based Algorithms:
- Support Vector Machine (SVM)

Distance-Based Classification Algorithms
- K-Nearest Neighbors (KNN)

Ensemble Learning Classification Algorithms
- Bagging Classifier (like Random Forest)
- Boosting Classifiers (like XGBoost, LightGBM, CatBoost, AdaBoost)
- Stacking Classifier
- Voting Classifier

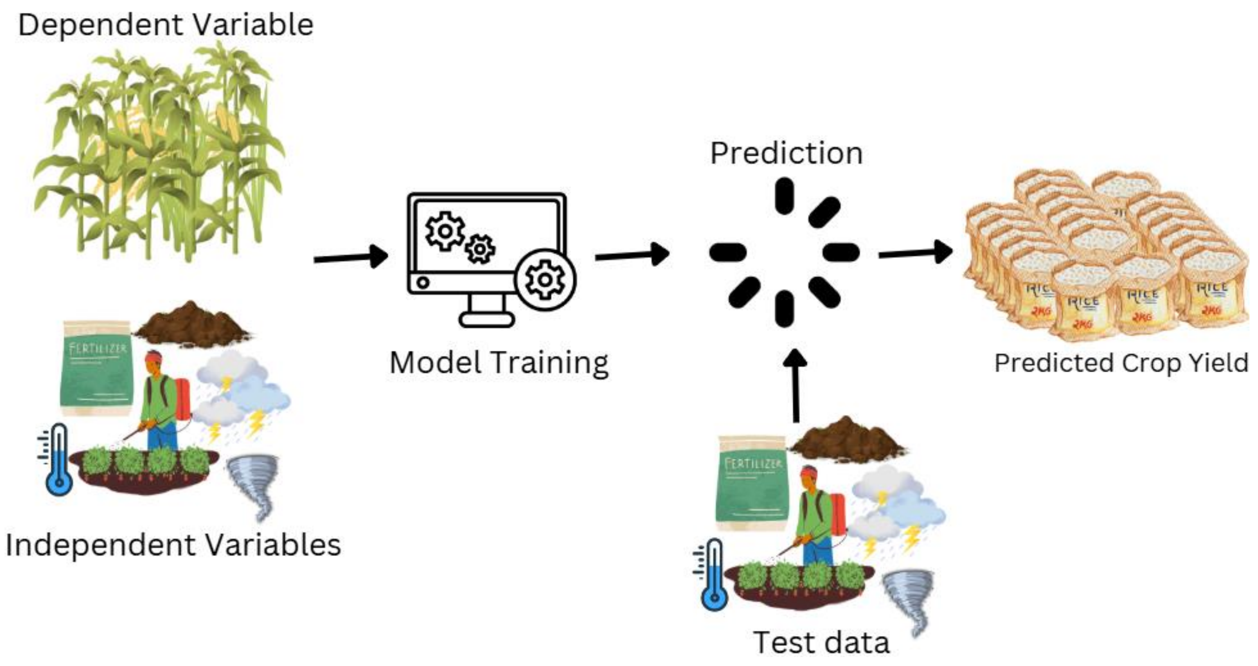
Tree-Based Classification Algorithms
- Decision Tree
- Random Forest
- Gradient Boosting (GBM)
- XGBoost
- LightGBM
- CatBoost

Types of Classification

Type of Classification	Definition	Real-World Examples
Binary Classification	Categorizes data into two groups	1. Spam Detection: Spam (1) or Not Spam (0) 2. Disease Prediction: Positive (1) or Negative (0) 3. Fraud Detection: Fraudulent (1) or Genuine (0)
Multi-Class Classification	Classifies input into three or more categories (only one class per instance)	1. Handwriting Recognition: Digit (0-9) 2. Iris Flower Classification: Setosa, Versicolor, Virginica 3. Animal Classification: Cat, Dog, or Bird
Multi-Label Classification	A single data point can belong to multiple categories	1. Movie Genre Prediction: Action + Comedy 2. News Categorization: Politics + Economy 3. Medical Diagnosis: Diabetes + Hypertension
Imbalanced Classification	One class has significantly fewer instances than the other	1. Fraud Detection: 5% Fraud, 95% Genuine 2. Disease Prediction: Rare diseases in patients 3. Defective Products: Out of 1000 items, only 20 are defective

Regression

Regression is a supervised learning technique focused on predicting continuous outcomes. It helps us understand the relationship between one or more independent variables (features) and a dependent variable (target).



Tools & Libraries: Python, Scikit-learn, Pandas, Matplotlib, Seaborn, TensorFlow/Keras

some commonly used Regression algorithms

Linear Regression Models:

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Ridge Regression (L2 Regularization)
- Lasso Regression (L1 Regularization)
- Elastic Net Regression (Combination of L1 & L2)

Support Vector-Based Regression Models:

- Support Vector Regression (SVR)

Probabilistic & Bayesian Regression Models:

- Bayesian Linear Regression
- Gaussian Process Regression (GPR)

Tree-Based Regression Models:

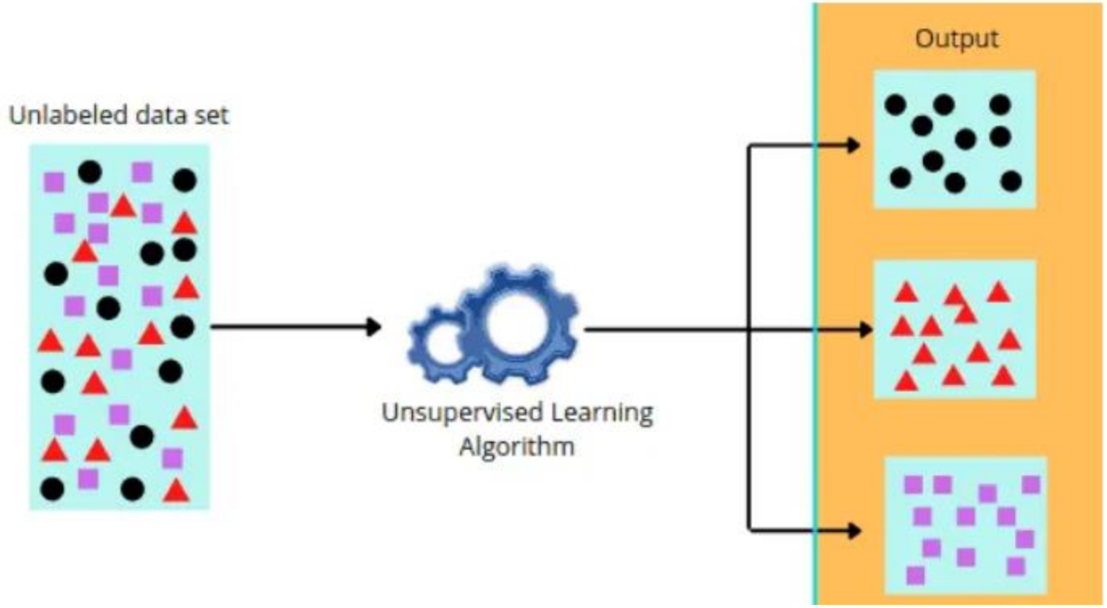
- Decision Tree Regression
- Random Forest Regression
- Gradient Boosting Regression (GBR)
- XGBoost Regression
- LightGBM Regression
- CatBoost Regression

Types of Regression

Type of Regression	Definition	Real-World Examples
Linear Regression	Establishes a linear relationship between input (X) and output (Y)	• House Price Prediction based on area • Sales Forecasting using marketing spend
Multiple Linear Regression	Extends linear regression with multiple input variables	• Car Price Prediction based on mileage, age, and brand • Medical Cost Estimation based on age, BMI, and smoking status
Polynomial Regression	Model non-linear relationships by adding polynomial terms (X^2, X^3)	• Predicting Growth Trends • Weather Forecasting
Ridge Regression	Adds L2 regularization to prevent overfitting	• Predicting Company Revenue • High-dimensional Data Analysis
Lasso Regression	Adds L1 regularization, reducing less important features to zero	• Feature Selection for machine learning models • Gene Expression Analysis

Unsupervised Learning

Unsupervised learning is a type of machine learning where the model learns patterns from unlabeled data without any human supervision. Unlike supervised learning, where we provide labeled examples, unsupervised learning identifies hidden patterns, structures, or relationships in data.
Imagine you are organizing a pile of books without knowing their categories. You observe patterns like genre, author, or color of the cover and group them accordingly. This is similar to unsupervised learning, where the model finds patterns without predefined labels.

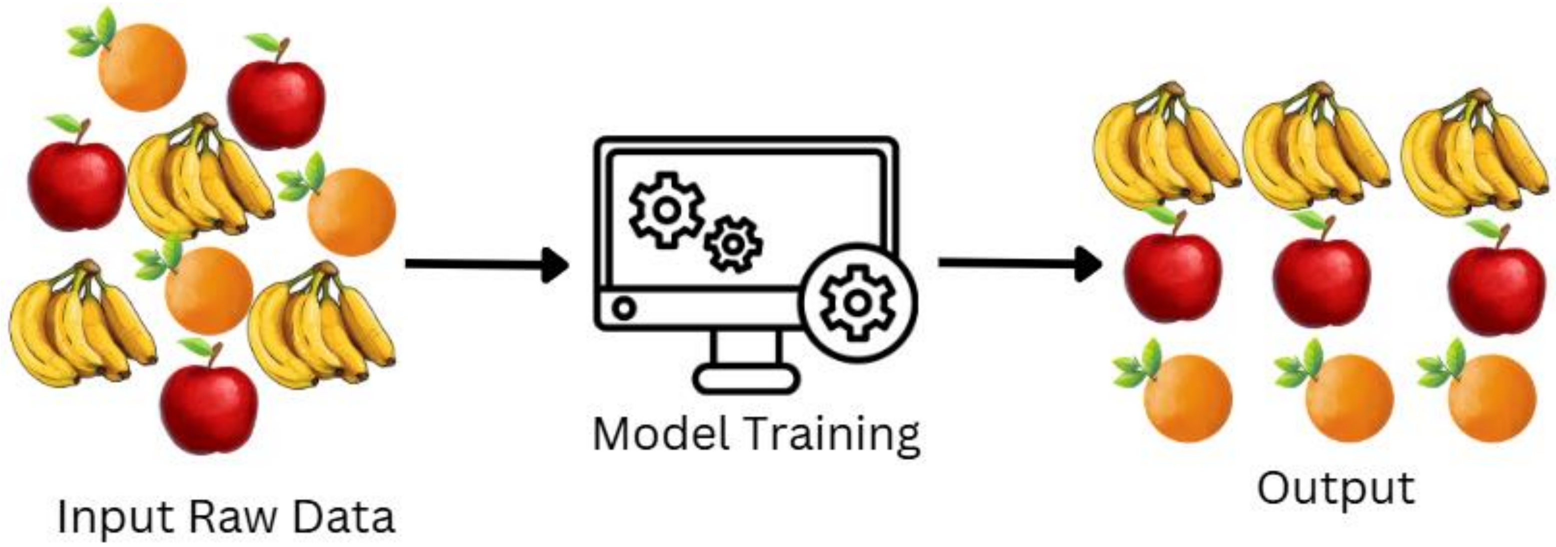


Sorting Fruits Without Knowing Their Names

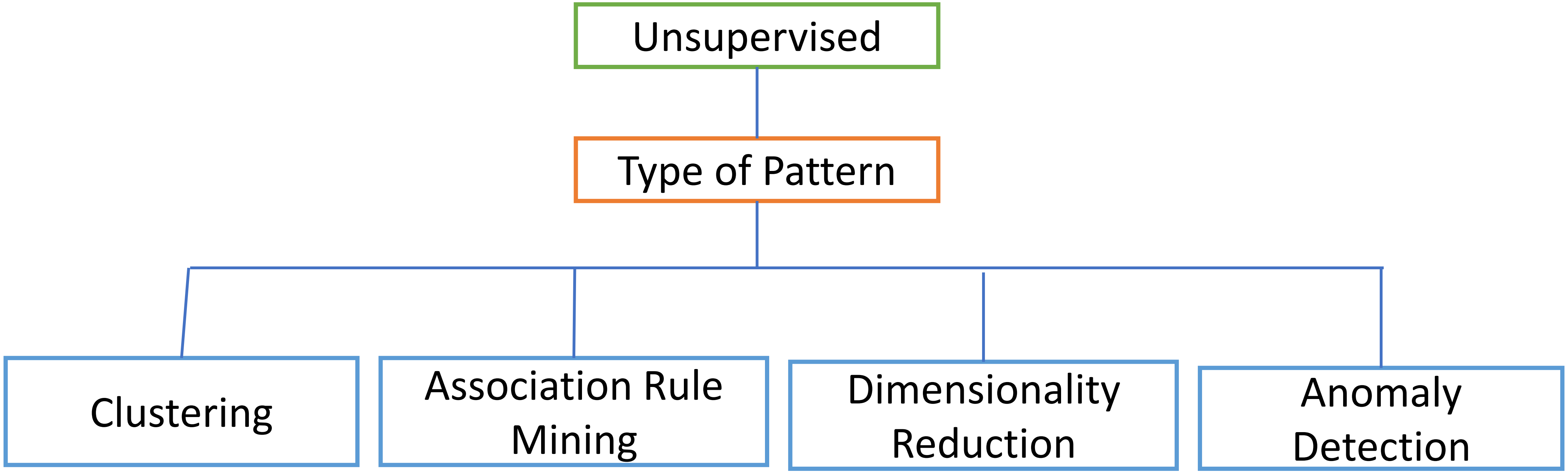
Imagine you have a basket full of different fruits, but you don’t know their names.

How can you group them?

- ❑ You start by looking at similarities like shape, size, and color.
- ❑ Without knowing their actual names, you automatically separate them into groups:
    - •Round & Red → Apples
    - •Long & Yellow → Bananas
    - •Small & Orange → Oranges
- ❑ This is exactly what unsupervised learning does!
    - •It finds hidden patterns without labels.
    - •No one tells the model what the groups are in advance.

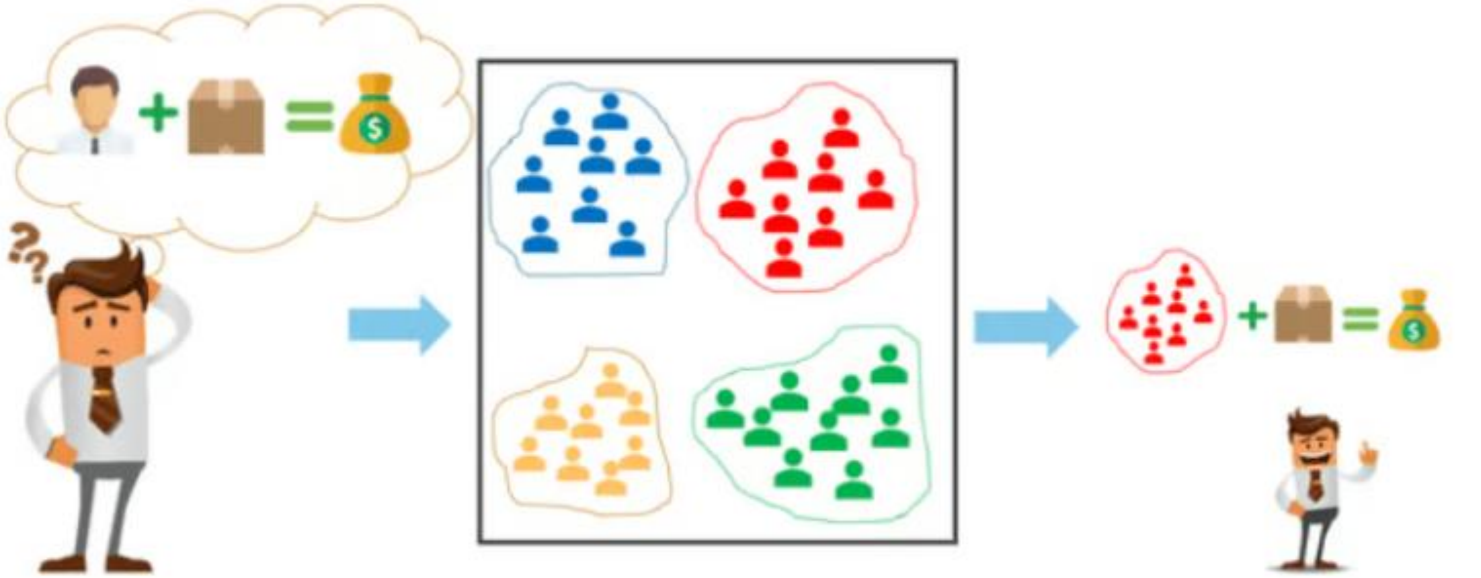


Types of Unsupervised Machine Learning



Clustering

Clustering is a machine learning technique used to group similar data points together without predefined labels. It helps in identifying patterns and structures in data by organizing objects into clusters based on similarity.



Trying to determine the appropriate audience for the product

Using clustering algorithms on the customer base

Selling the product to the targeted audience

some commonly used clustering algorithms

Partitioning Clustering (Centroid-Based)

- K-Means
- K-Medoids

Distribution-Based Clustering

- Gaussian Mixture Models (GMM)
- Expectation-Maximization (EM)

Hierarchical Clustering

- Agglomerative Clustering
- Divisive Clustering

Density-Based Clustering

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- OPTICS (Ordering Points To Identify the Clustering Structure)

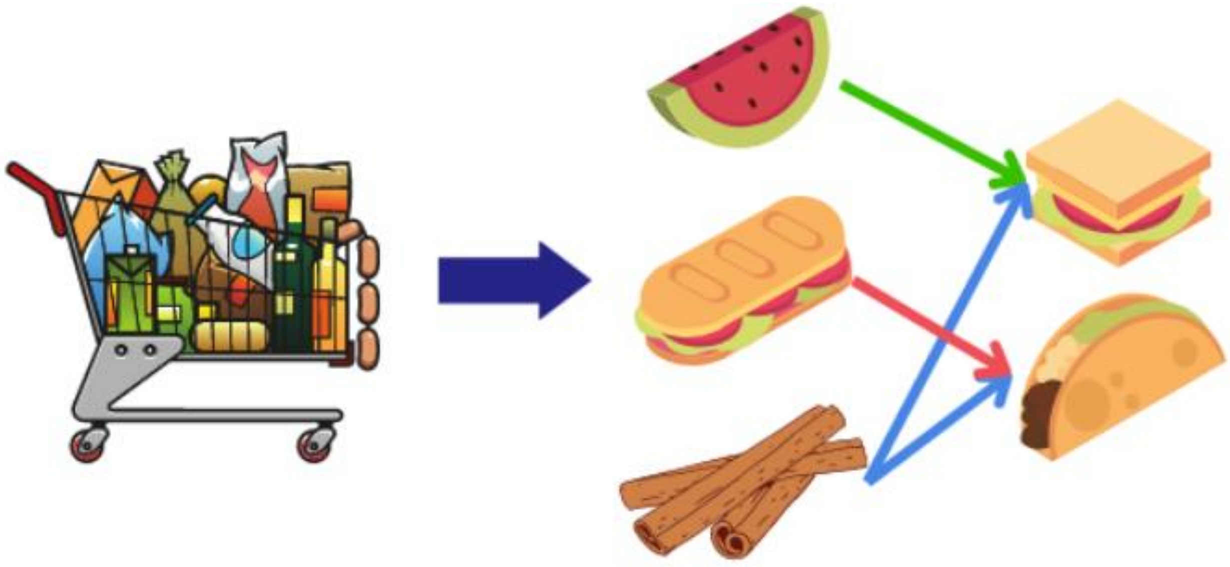
Types of clustering

Type of Clustering	Description	Common Algorithms	Real-Life Example
Partitioning Clustering	Divides data into K fixed clusters based on similarity.	K-Means, K-Medoids	Customer segmentation in e-commerce
Hierarchical Clustering	Creates a tree-like structure (dendrogram) to form clusters.	Agglomerative, Divisive	Family tree classification
Density-Based Clustering	Groups data based on dense regions and ignores noise.	DBSCAN, OPTICS	Fraud detection in banking
Distribution-Based Clustering	Assumes data follows a probabilistic distribution and assigns clusters accordingly.	Gaussian Mixture Models (GMM), Expectation-Maximization (EM)	Identifying handwritten digits (MNIST)
Grid-Based Clustering	Divides space into a grid structure and clusters within each grid.	STING, CLIQUE	Geographic data clustering (weather patterns)

Association Rule Mining

Association Rule Mining is a data mining technique used to find hidden patterns, relationships, or associations between different items in a large dataset. It helps in understanding how items are related to each other based on how frequently they appear together.

- ➢ It is unsupervised learning, meaning the algorithm discovers patterns without predefined labels.
- ➢ It is widely used in market basket analysis, where businesses analyze customer purchase behavior.
- ➢ The discovered patterns are in the form of IF-THEN rules (e.g., "If a customer buys bread, they are likely to buy butter").



some commonly used association rule mining algorithms

Sequential Pattern Mining Algorithms

- GSP (Generalized Sequential Pattern Mining)
- SPADE (Sequential Pattern Discovery using Equivalence classes)

Tree-Based Algorithms

- FP-Growth (Frequent Pattern Growth)
- H-Mine (Hyper-structure Mining of Frequent Patterns)

Apriori-Based Algorithms

- Apriori
- AIS (Artificial Immune System)
- SETM (Set-Oriented Mining Algorithm)

Vertical Data Format-Based Algorithms

- ECLAT (Equivalence Class Clustering and Bottom-Up Lattice Traversal)
- OPUS (Optimal Unique Subset)

Types of Association Rule Mining

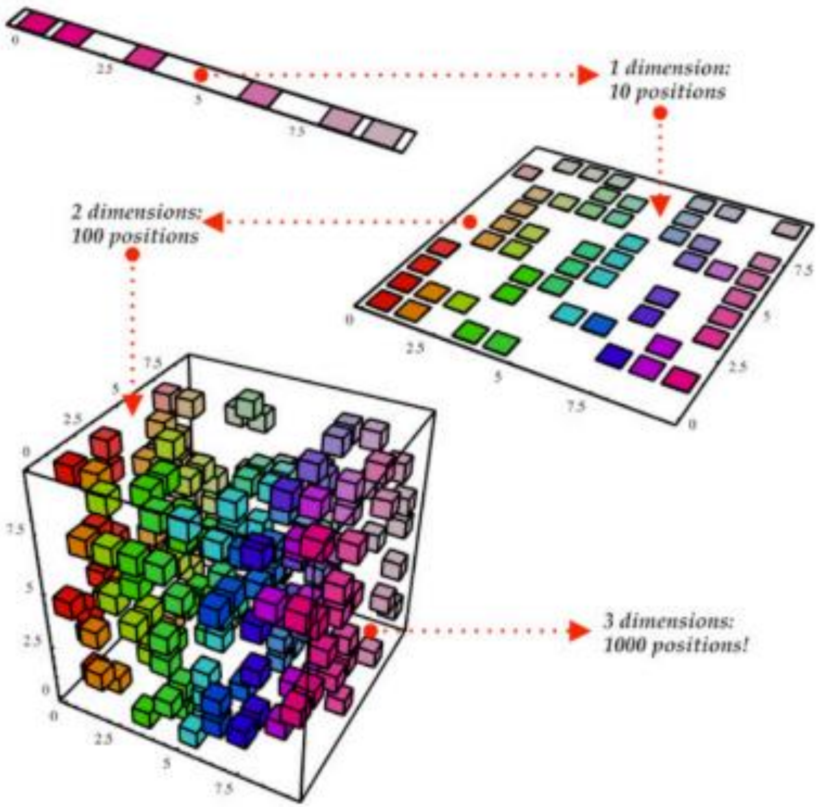
Type of Association Rule Mining	Algorithm	Description	Used For
Single-Dimensional Association Rule Mining	Apriori, FP-Growth	Finds associations based on a single attribute like product purchases.	Market basket analysis, retail sales
Multi-Dimensional Association Rule Mining	Apriori, SETM	Considers multiple attributes (e.g., age, location) to generate rules.	Customer segmentation, targeted marketing
Quantitative Association Rule Mining	Apriori, FP-Growth	Uses numerical attributes (e.g., price, quantity) in mining.	Retail pricing strategies, revenue analysis
Frequent Pattern-Based Association Rule Mining	FP-Growth, ECLAT	Identifies frequently occurring itemset in transactions.	Fraud detection, recommendation systems
Sequential Pattern Mining	GSP (Generalized Sequential Pattern), SPADE	Finds time-based or sequential patterns in data.	Predicting customer behavior, stock market trends

Dimensionality Reduction

Dimensionality reduction is a technique in **unsupervised machine learning** that reduces the number of input features (dimensions) in a dataset while preserving essential patterns and structures. It helps in eliminating redundant, irrelevant, or highly correlated features, making data analysis more efficient without losing significant information.

Why is Dimensionality Reduction Needed?

- As dimensions increase, data becomes sparse and difficult to analyze.
- Reducing dimensions speeds up machine learning models.
- Helps in visualizing high-dimensional data in 2D or 3D.
- Removes irrelevant variations in data.



some commonly used Dimensionality reduction algorithms

Linear Dimensionality Reduction Algorithms

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Singular Value Decomposition (SVD)

Non-Linear Dimensionality Reduction Algorithms

- t-Distributed Stochastic Neighbour Embedding (t-SNE)
- Uniform Manifold Approximation and Projection (UMAP)
- Isomap (Isometric Mapping)
- Locally Linear Embedding (LLE)

Deep Learning-Based Dimensionality Reduction

- Autoencoders

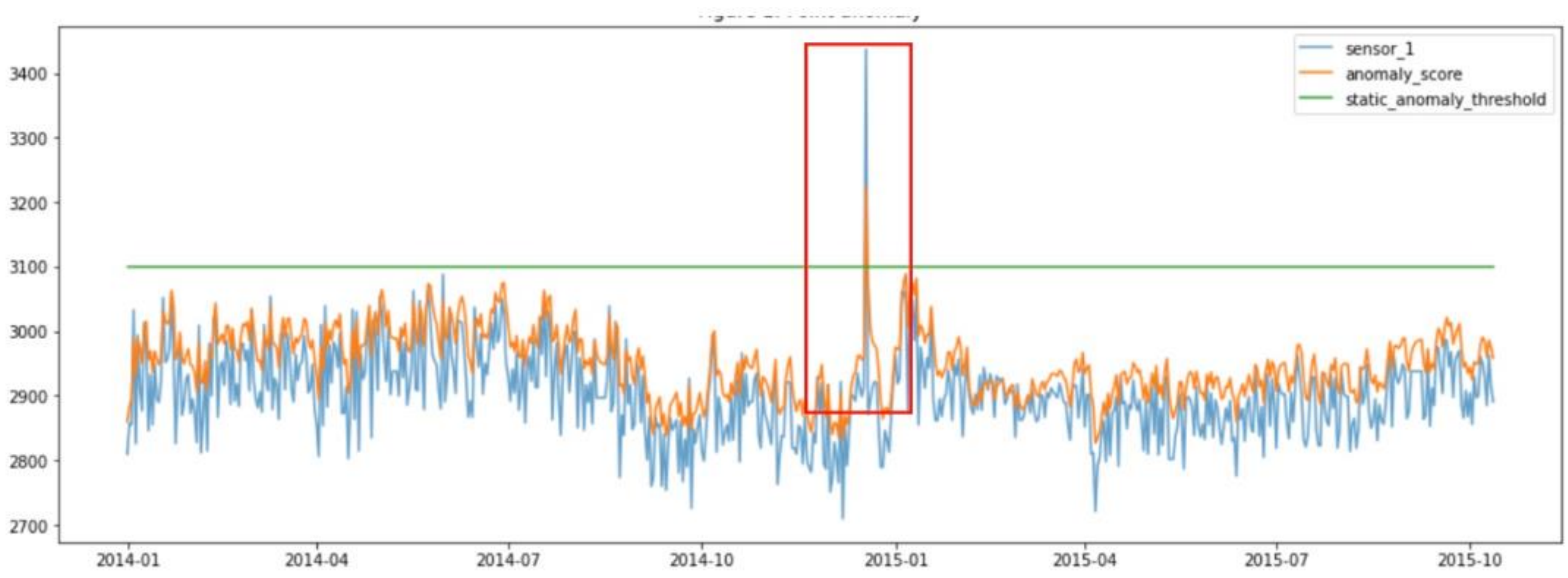
Types of Dimensionality Reduction

Type	Algorithm	Description	Used For
Feature Selection	Filter Methods, Wrapper Methods, Embedded Methods	Selects the most relevant features and removes redundant ones.	Preprocessing in machine learning models
Feature Extraction	PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis), t-SNE, Autoencoders	Creates new features by transforming existing ones.	Image compression, text data analysis
Matrix Factorization	SVD (Singular Value Decomposition), NMF (Non-Negative Matrix Factorization)	Breaks down a large matrix into lower-rank representations.	Recommendation systems, topic modeling
Manifold Learning	t-SNE, Isomap, UMAP	Finds low-dimensional representations of non-linear data.	Visualizing high-dimensional data.

Anomaly Detection

Anomaly detection is the process of identifying data points, events, or observations that deviate significantly from the expected or normal patterns in a dataset. These deviations, known as anomalies or outliers, may indicate fraud, defects, cyber threats, system failures, or other rare and important events.

Anomaly detection is mostly unsupervised, but it can also be supervised or semi-supervised depending on the availability of labeled data.



some commonly used Anomaly Detection algorithms

Statistical-Based Methods
- Z-Score
- Grubbs' Test
- Chi-Square Test

Density-Based Methods
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- Kernel Density Estimation (KDE)

Distance-Based Methods
- k-Nearest Neighbors (k-NN)
- Local Outlier Factor (LOF)
- Mahalanobis Distance

Classification-Based Methods
- Support Vector Machine (SVM)
- Decision Trees
- Random Forest

Time-Series Anomaly Detection
- ARIMA
- LSTM
- Isolation Forest

Clustering-Based Methods
- K-Means
- Gaussian Mixture Model (GMM)

Reconstruction-Based Methods
- Autoencoders (Neural Networks)
- Principal Component Analysis (PCA)

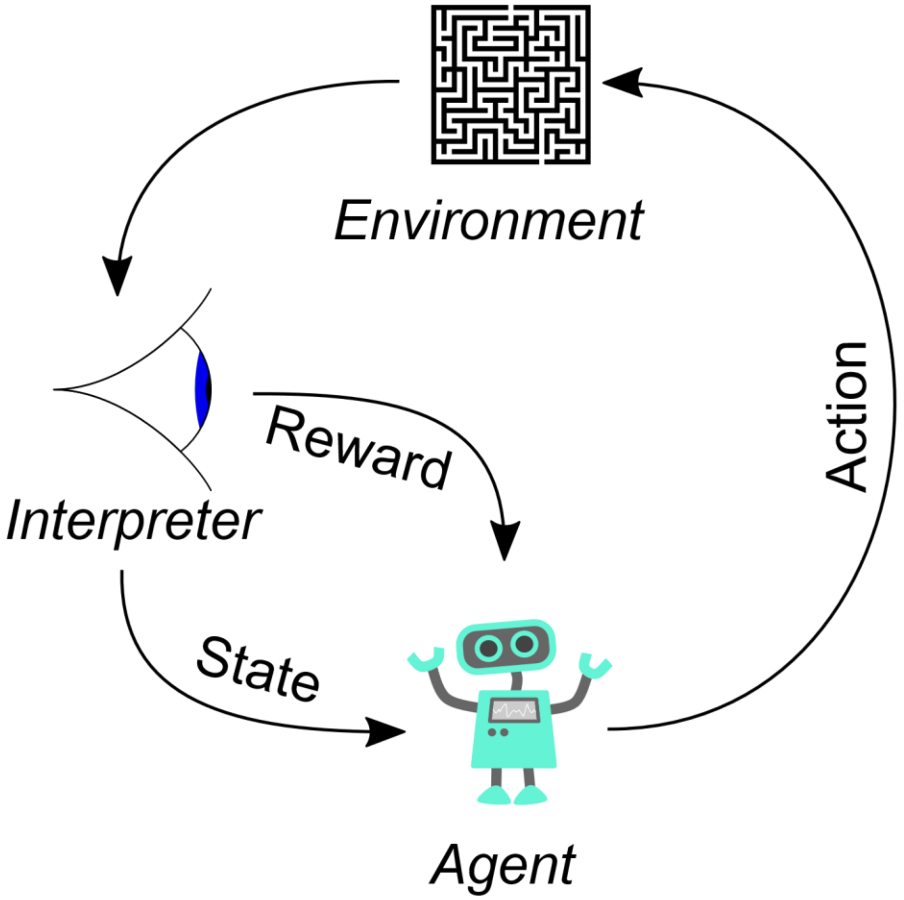
Types of Anomaly Detection

Type of Anomaly Detection	Algorithm	Description	Used For
Statistical-Based Anomaly Detection	Z-Score, Grubbs' Test, Chi-Square Test	Uses statistical methods to detect outliers based on distribution.	Fraud detection in banking, network intrusion detection
Distance-Based Anomaly Detection	k-Nearest Neighbors (k-NN), LOF (Local Outlier Factor)	Identifies anomalies by measuring distance from other points.	Credit card fraud detection, healthcare anomaly detection
Density-Based Anomaly Detection	DBSCAN, KDE (Kernel Density Estimation)	Detects anomalies in sparse regions with lower density.	Malware detection, intrusion detection in cybersecurity
Clustering-Based Anomaly Detection	K-Means, Gaussian Mixture Model (GMM), DBSCAN	Treats anomalies as points that do not fit into clusters.	Image processing, manufacturing defect detection
Classification-Based Anomaly Detection	Decision Trees, SVM (Support Vector Machine), Random Forest	Uses supervised learning to classify normal vs. anomalous data.	Spam filtering, medical diagnosis
Reconstruction-Based Anomaly Detection	Autoencoders, PCA (Principal Component Analysis)	Uses dimensionality reduction or neural networks to reconstruct normal data and detect anomalies.	Sensor failure detection, industrial IoT monitoring
Time-Series Anomaly Detection	ARIMA, LSTM, Isolation Forest	Detects anomalies in time-sequenced data like trends or patterns.	Stock market anomaly detection, server downtime prediction

Reinforcement Learning

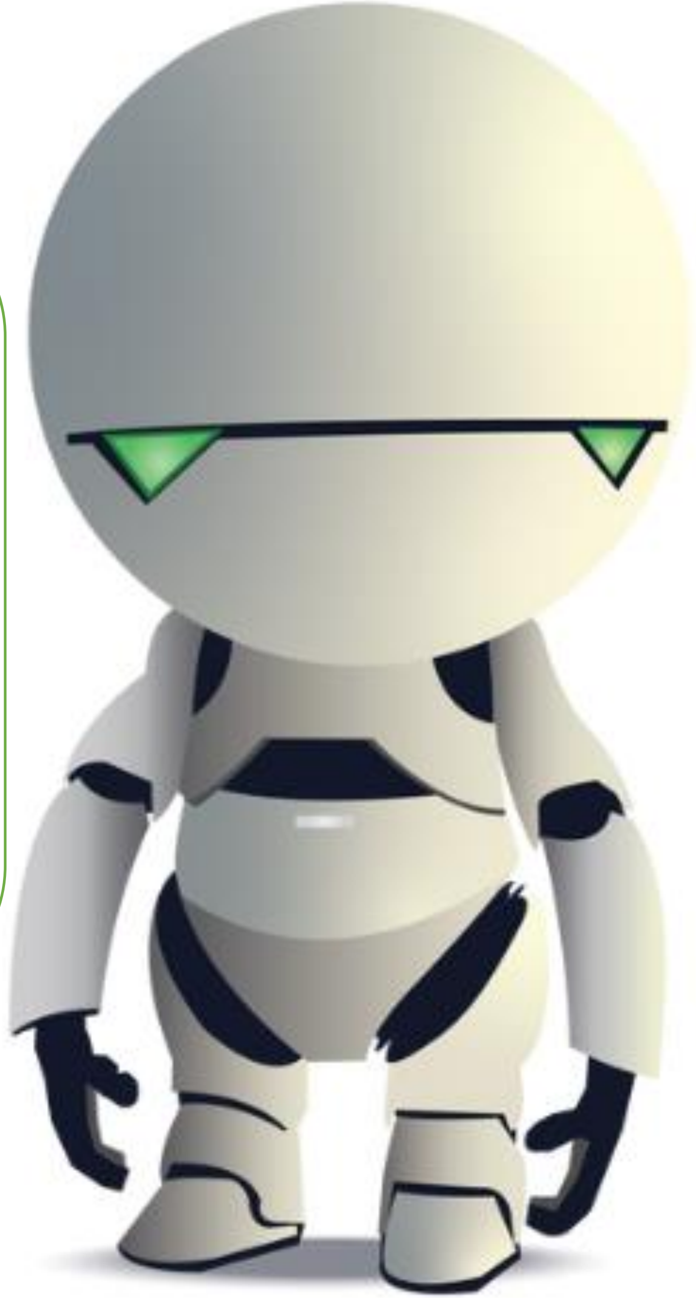
Reinforcement Learning (RL) is a type of machine learning where an agent (learner) interacts with an environment to learn the best actions through trial and error. The agent receives rewards for good actions and penalties for bad actions, helping it refine its decision-making over time.

For example, imagine a robot learning to walk. At first, it may fall often, but with each attempt, it adjusts its movements based on the feedback (whether it stays balanced or falls). Over time, it learns the optimal way to walk efficiently.



Define the RL Problem

1. **Agent:** The robot.
2. **Environment:** The floor on which the robot moves.
3. **Actions:** The robot can move left, right, forward, or balance itself.
4. **State:** The robot's current position and posture (standing, falling, balanced).
5. **Reward:**
   - +10 points for successfully moving forward.
   - -5 points for stumbling.
   - -10 points for falling.



Initial Exploration (Random Actions)

At the beginning, the robot does not know how to walk, so it randomly selects actions.

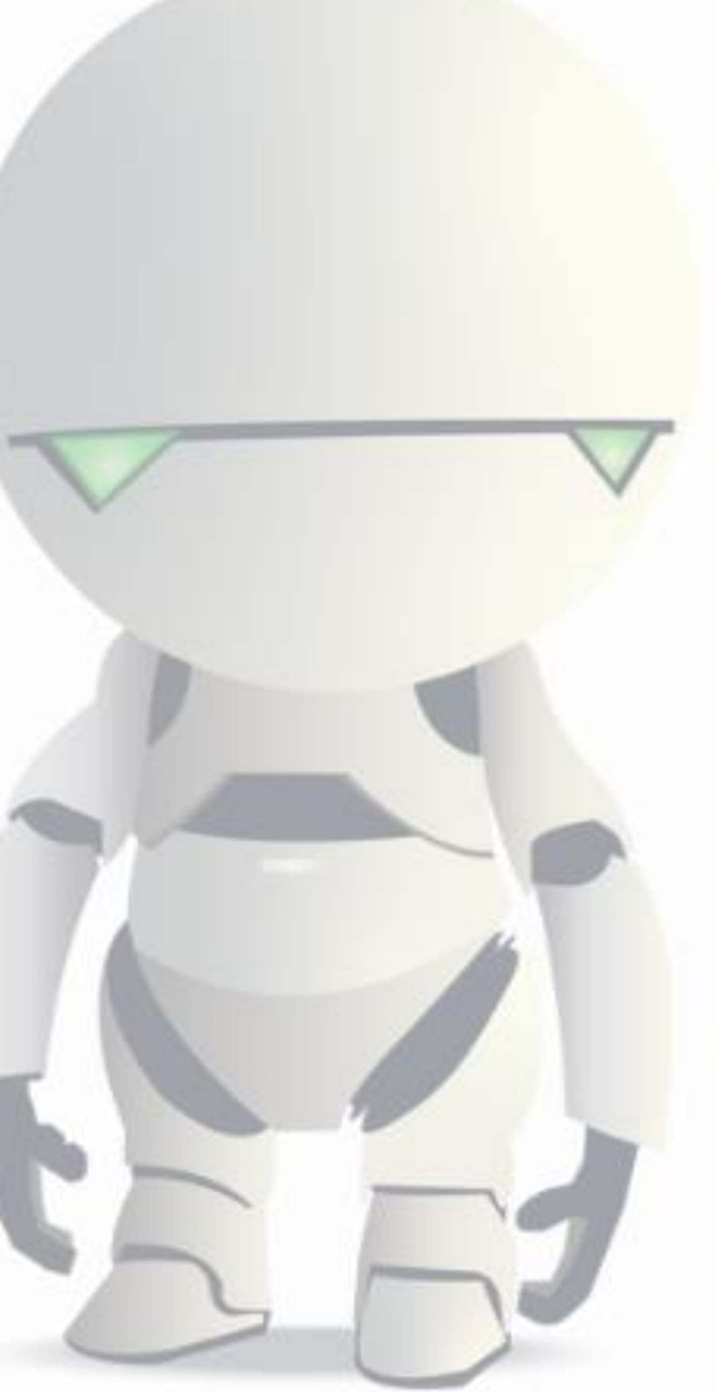
Example of Random Actions:

1. The robot moves one leg → Falls down (-10 points)
2. The robot moves both legs together → Falls down (-10 points)
3. The robot moves one leg slightly and adjusts balance → Does not fall (No penalty)
4. The robot moves one leg forward while balancing → Moves forward (+10 points)

Learning from Rewards (Updating Policy)

The robot **remembers** that:

1. Falling gives a penalty, so it should avoid those actions.
2. Balancing and moving forward gives a reward, so it should repeat those actions.



Refining the Strategy (Exploitation vs. Exploration)

Now the robot tries to improve:

- ❑ Exploration: It tests new walking patterns (e.g., longer steps, bending knees).
- ❑ Exploitation: It reuses the best movements it has learned so far.

Example:

- It moves one leg first, then the other (this was the best action from previous attempts).
- It learns how to shift its weight to avoid falling.
- Over time, it optimizes the movements to walk smoothly.

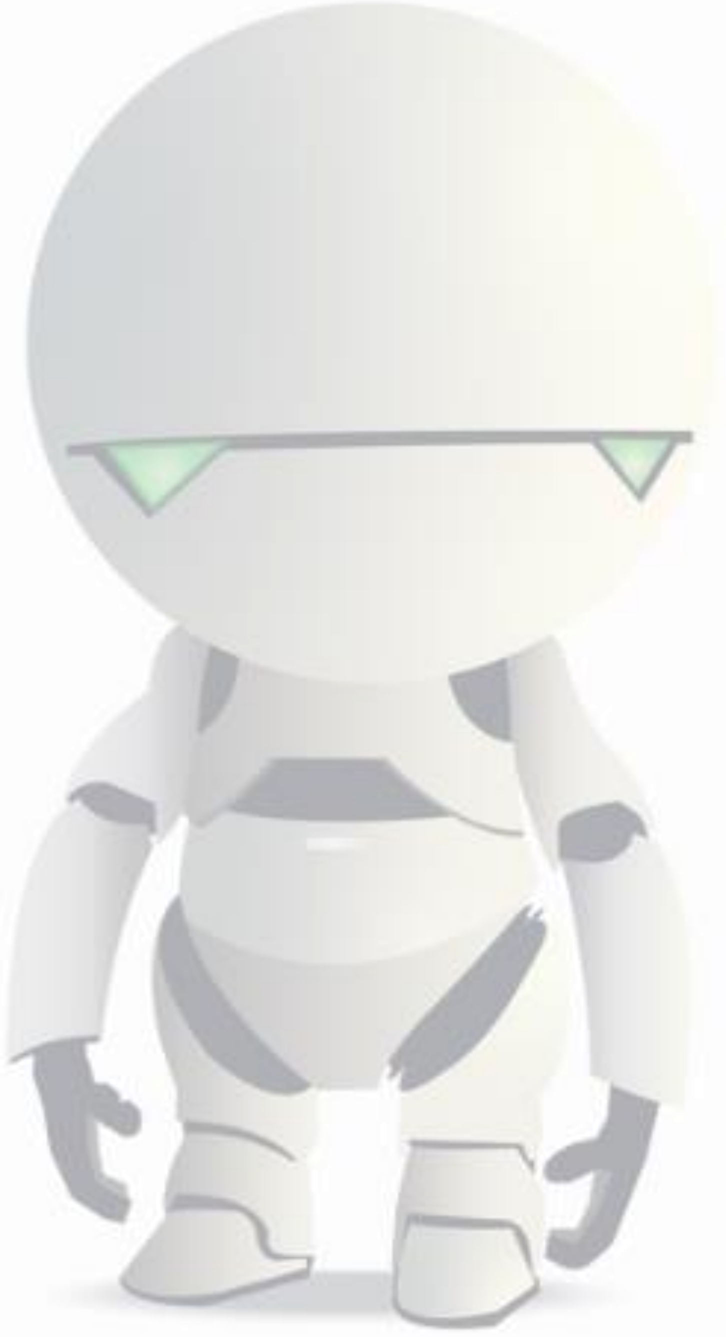
Achieving the Goal (Optimal Walking Policy)

After many trials, the robot learns the **best walking sequence** to move forward without falling.

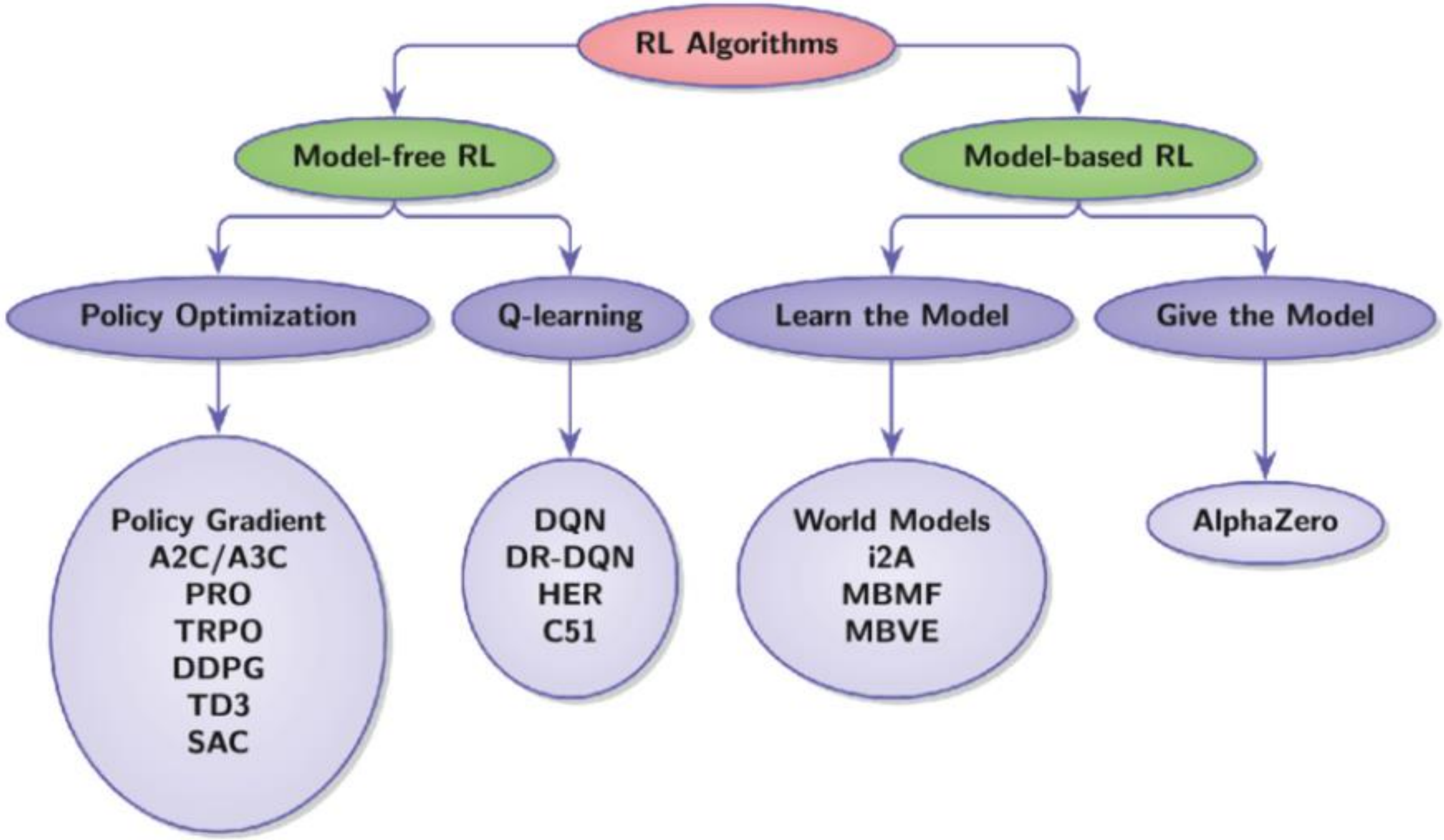
Final Learned Strategy:

1. Move left leg slightly forward.
2. Shift weight onto the left leg.
3. Move right leg slightly forward.
4. Maintain balance.
5. Repeat the cycle.

The robot can now walk efficiently!



Types of Reinforcement Learning



Real-World Use Cases

- Autonomous Vehicles – Used for self-driving cars to learn safe driving in real-world traffic.
- Robotics – Helps robots learn tasks like walking, grasping objects, and industrial automation.
- Game Playing AI – Used in games like Chess, Go, and Atari (e.g., AlphaGo, AlphaZero).
- Stock Market Trading – RL optimizes buy/sell strategies based on market conditions.
- Healthcare & Drug Discovery – Used for treatment planning, robotic surgery, and drug design.
- Energy Management – Optimizes power grid efficiency and smart home energy consumption.
- Recommendation Systems – Improves personalized content recommendations (Netflix, YouTube, Amazon).
- Traffic Signal Control – Optimizes traffic lights for reducing congestion in smart cities.