Machine Learning Fundamentals

Iván Moreno (ivan@nieveconsulting.com)

What is Machine Learning?

Definition

A field of study that gives computers the ability to learn from data without being explicitly programmed.

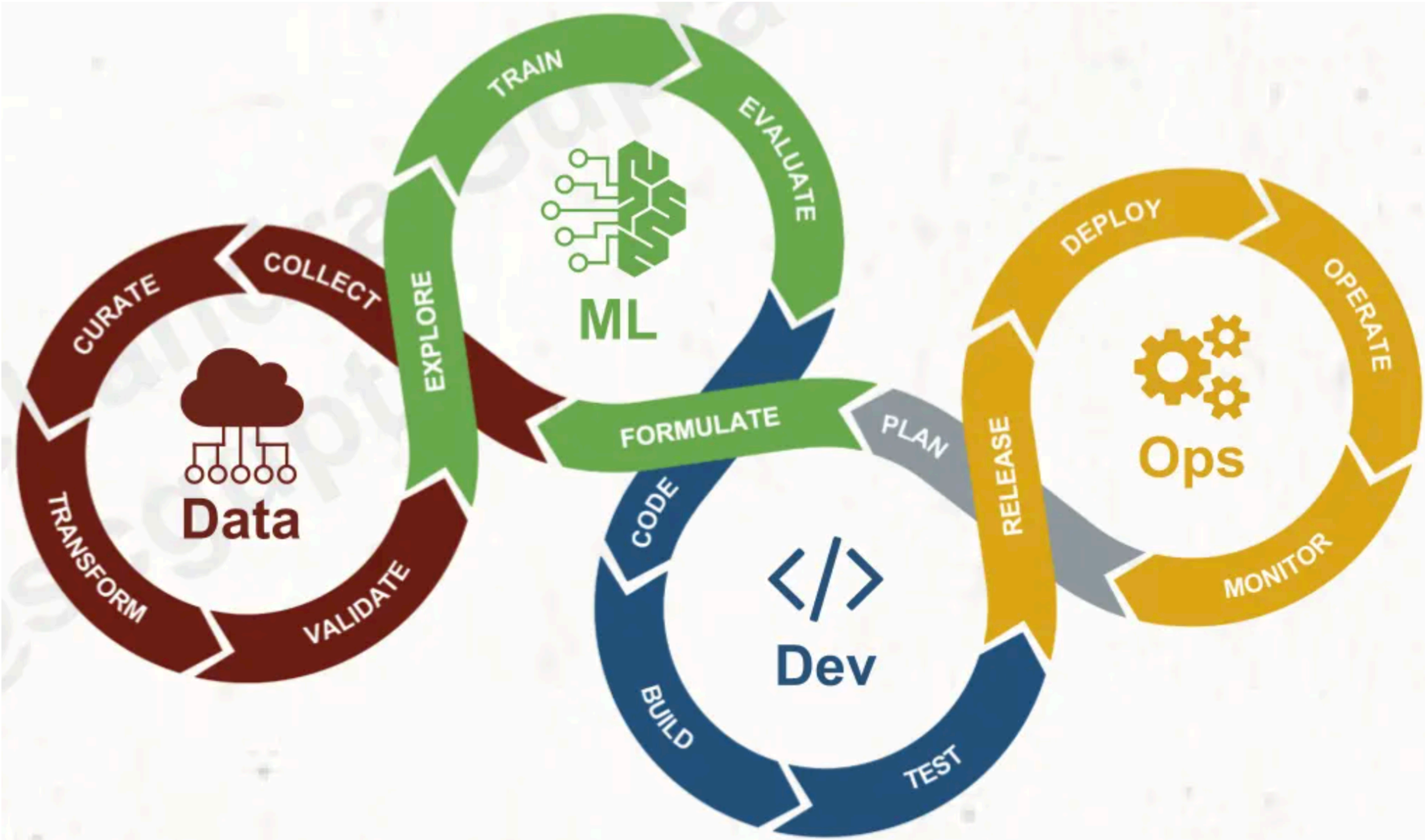
Why Machine Learning Matters

- Scalability in decision-making.
- Automating repetitive tasks.
- Unlocking insights from large datasets.

Core Principles of Machine Learning

- **Data-Driven Decisions:** Leveraging data for predictive modeling.
- **Model Representation:** Choice of models (e.g., linear models, decision trees, neural networks).
- **Generalization:** Balancing model complexity and performance on unseen data.
- **Optimization:** Loss functions, cost functions, and the gradient descent method.
- **Evaluation:** Accuracy, precision, recall, F1-score, ROC-AUC.

The Machine Learning Pipeline



Key Stages in the ML Pipeline

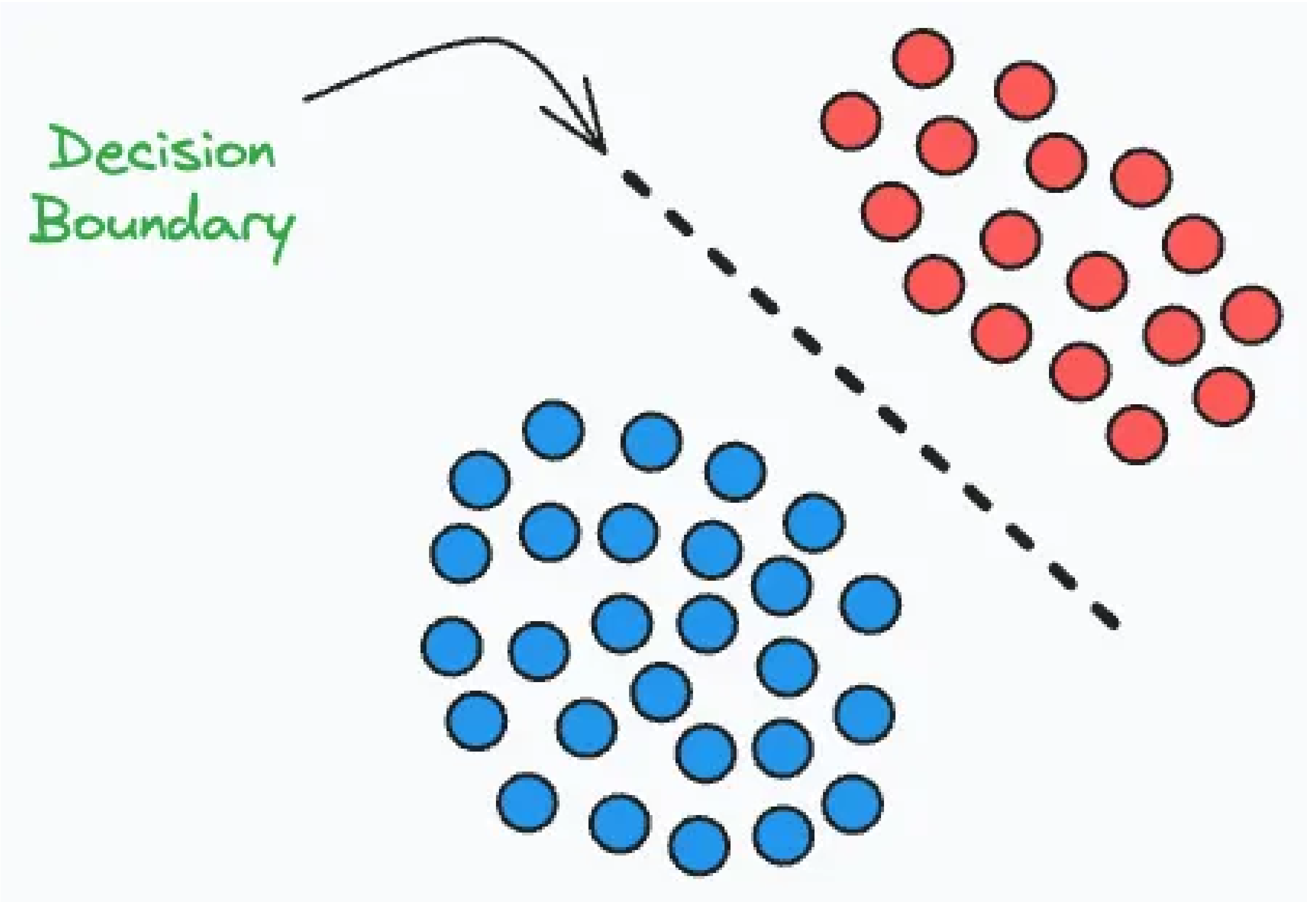
1. **Problem Definition:**
    - Identify the objective (e.g., classification, regression).
2. **Data Collection:**
    - Gather data from reliable sources.
3. **Data Preprocessing:**
    - Handling missing values, data normalization, feature engineering.
4. **Model Selection:**
    - Choose appropriate algorithms based on the problem and data characteristics.

Key Stages in the ML Pipeline (cont.)

5. **Training:**
   - Split data into training/validation/test sets. Train the model.
6. **Evaluation:**
   - Use performance metrics to assess model quality.
7. **Deployment:**
   - Integrate the model into production environments.
8. **Monitoring & Maintenance:**
   - Continuously monitor model performance and update as needed.

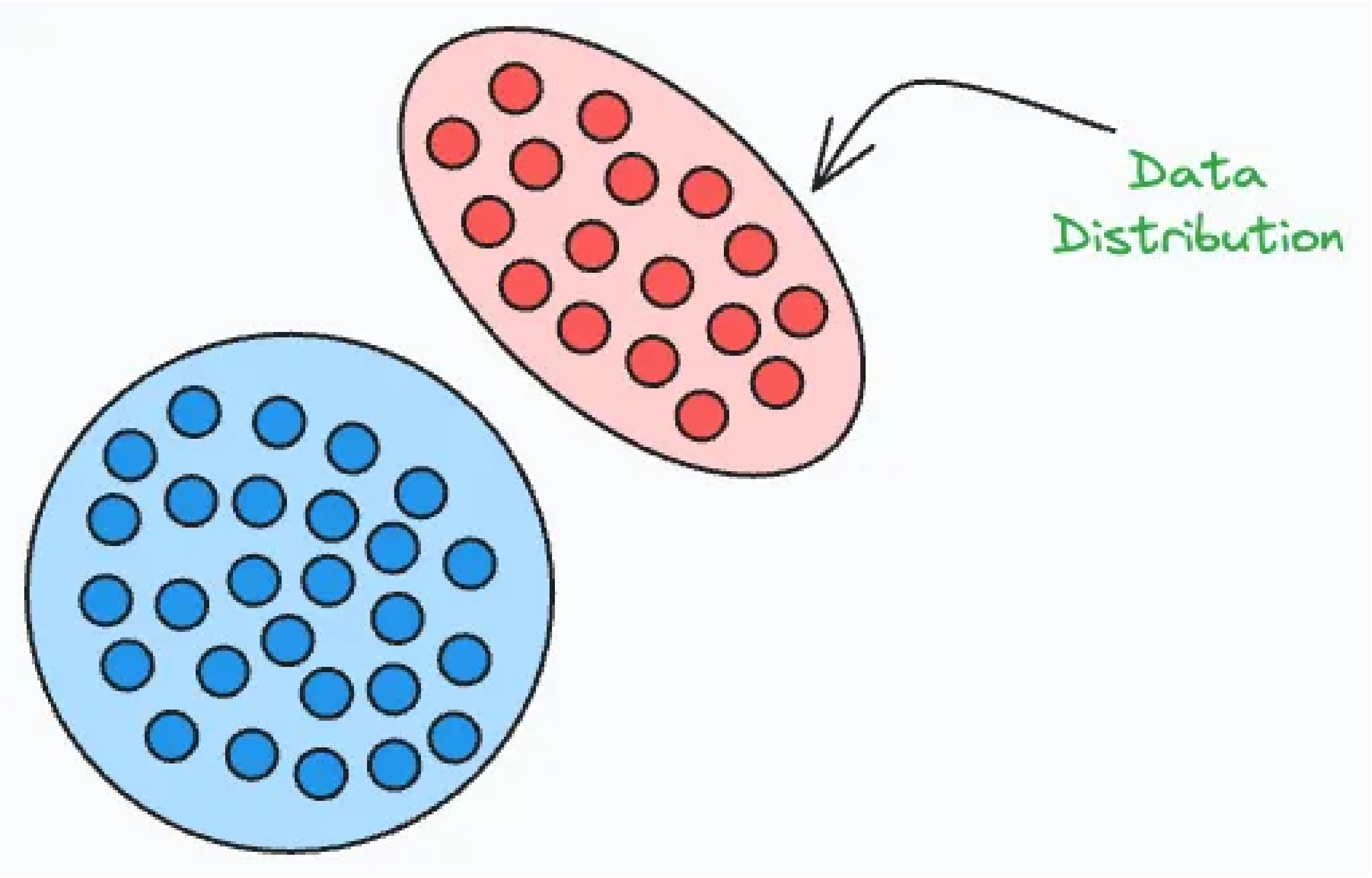
Discriminative vs. Generative Models

Discriminative Models



- Focus on predicting the target variable directly (e.g., $\mathrm{P(y|X)}$).
- **Examples:** Logistic Regression, Support Vector Machines (SVMs), Neural Networks.
- **Advantages:** Often provide better predictive accuracy.

Generative Models

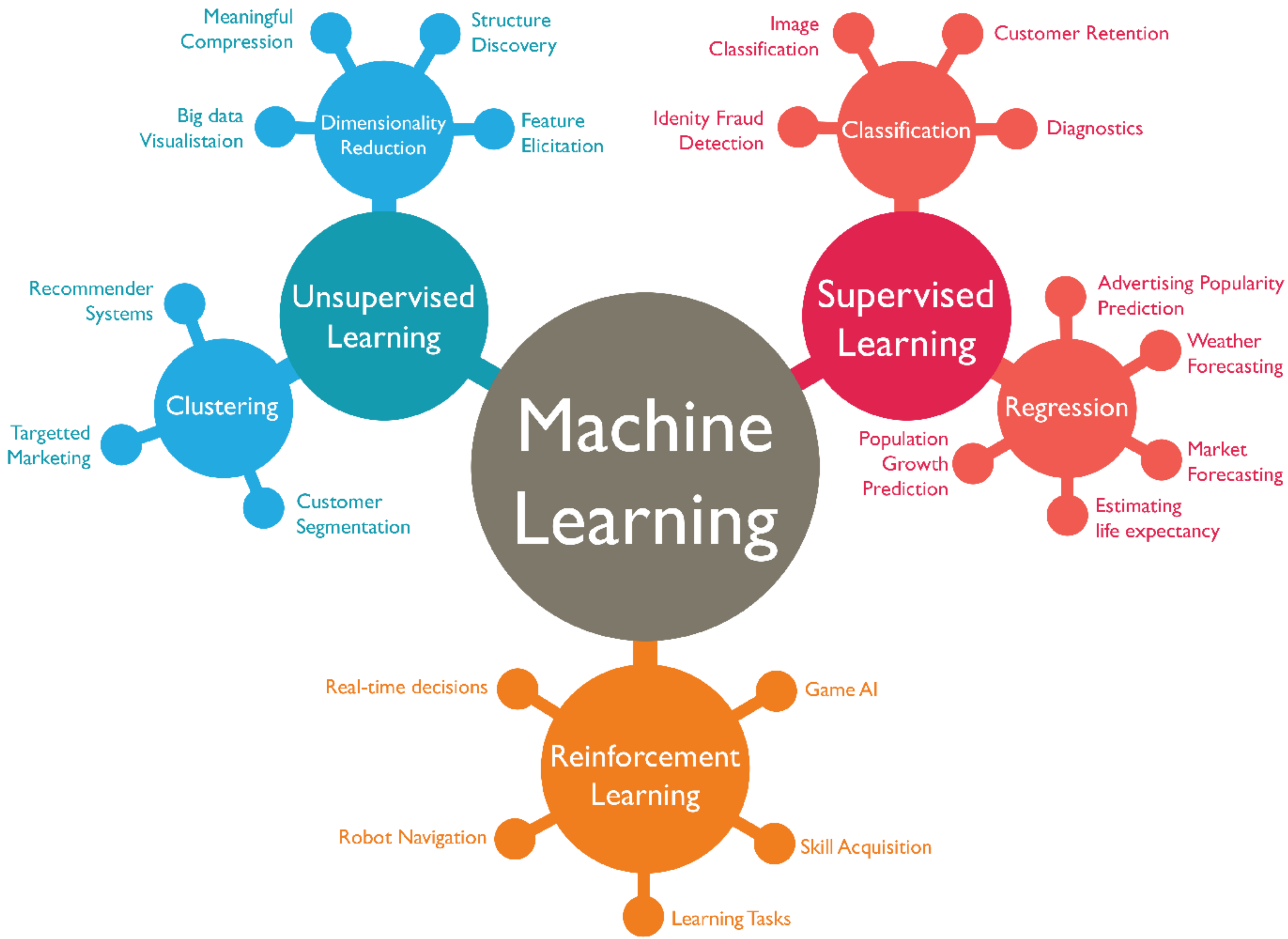


When to Use Each Type of Model

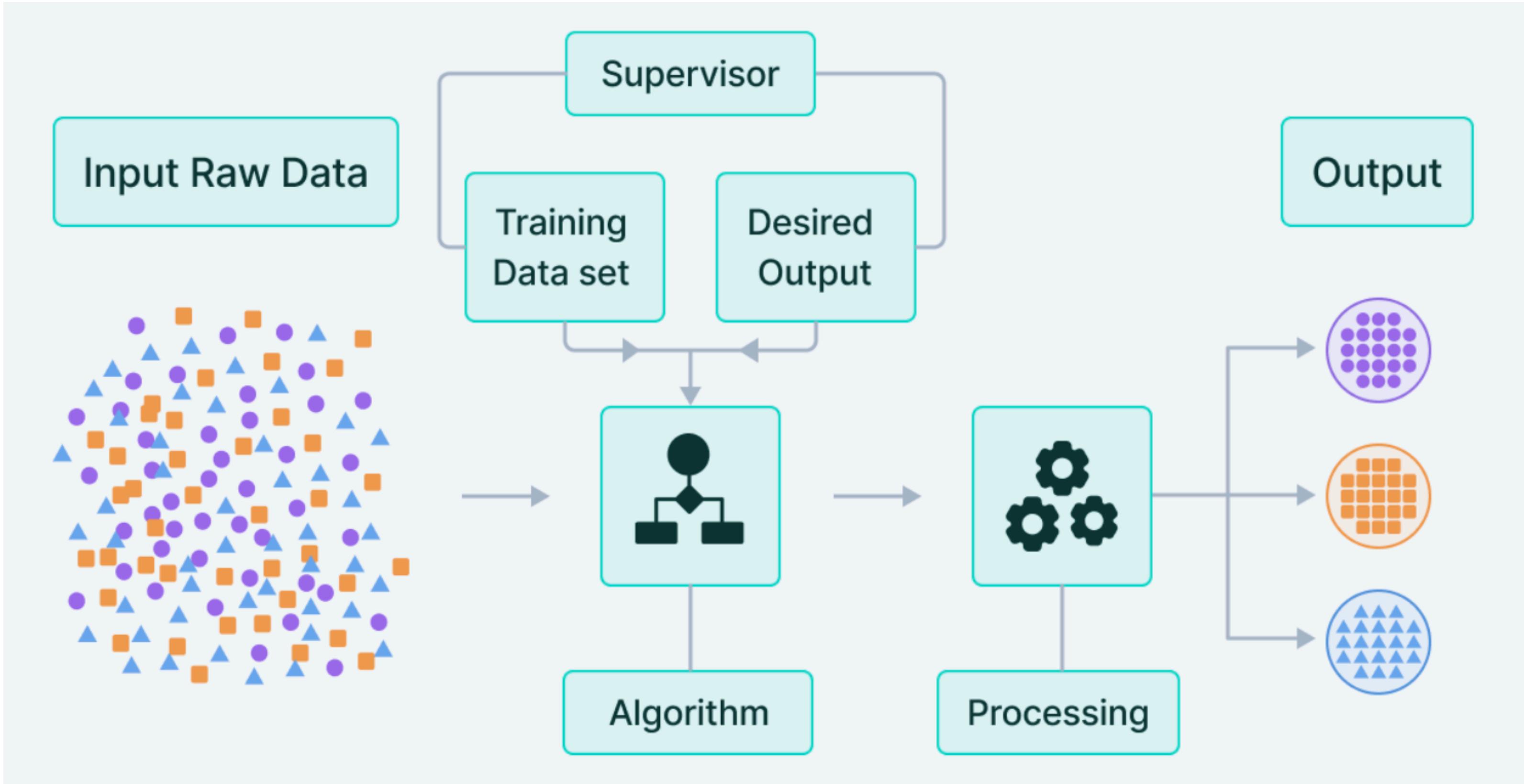
Tip

Choose based on the problem context—discriminative models for classification accuracy, generative models for data understanding and synthesis.

Learning Paradigms

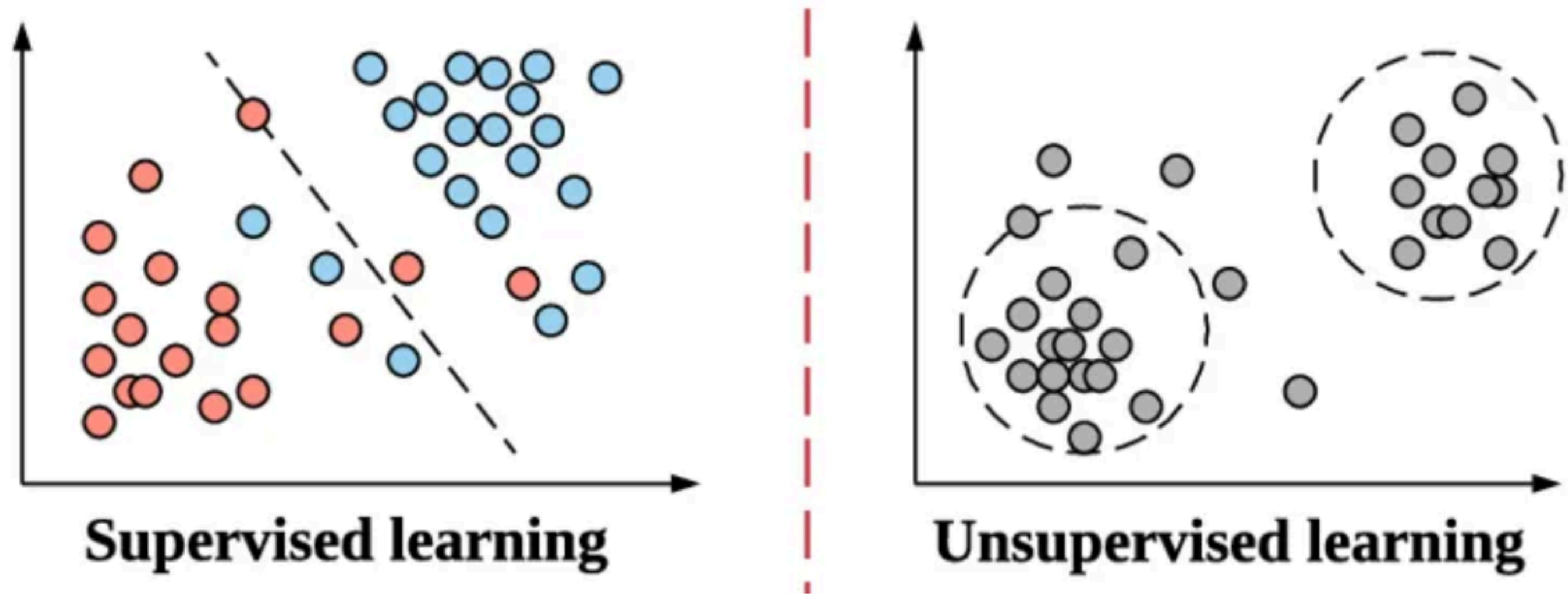


Supervised Learning



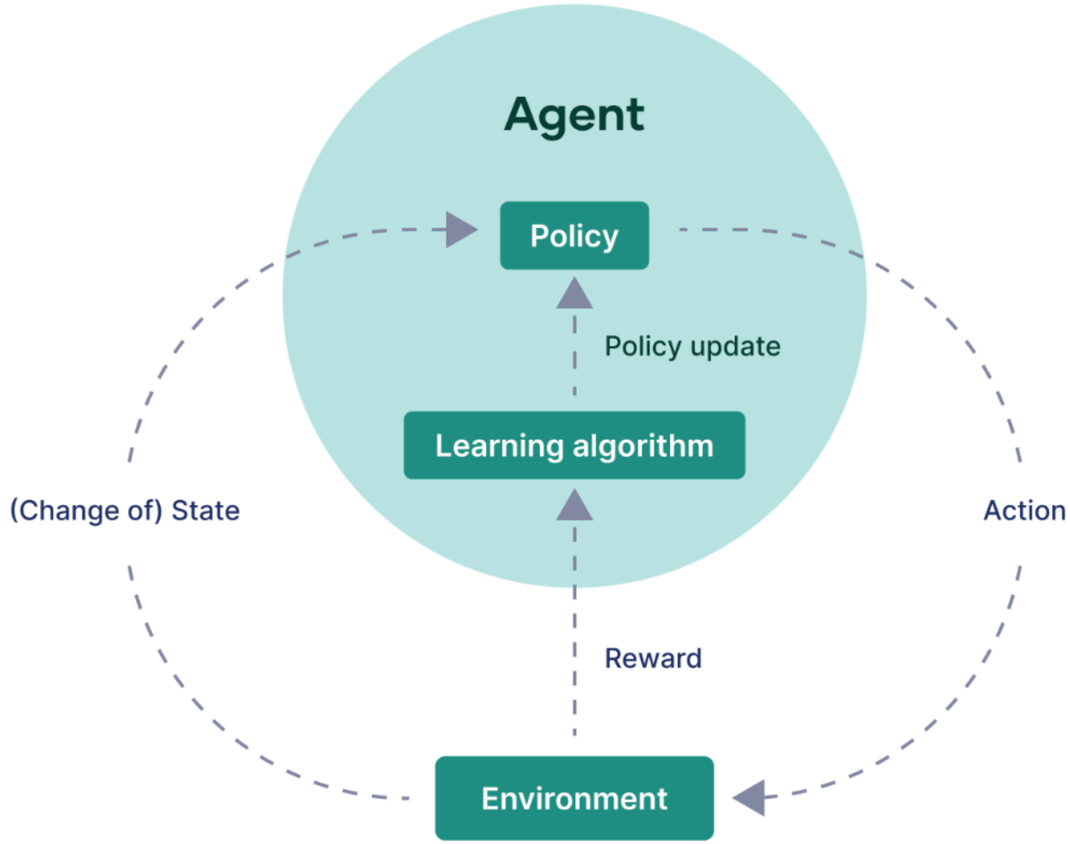
- **Goal:** Learn a function that maps inputs to outputs using labeled data.
- **Common Algorithms:** Linear Regression, Decision Trees, Support Vector Machines (SVMs), k-Nearest Neighbors (k-NN).
- **Applications:** Spam detection, medical diagnosis, fraud detection.

Unsupervised Learning



- **Goal:** Find hidden patterns or intrinsic structures in unlabeled data.
- **Techniques:** Clustering (K-means, Hierarchical Clustering), Dimensionality Reduction (PCA, t-SNE).
- **Applications:** Market segmentation, anomaly detection, gene expression analysis.

Reinforcement Learning



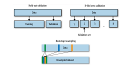
- **Goal:** Learn a policy to maximize cumulative reward in an environment.
- **Key Concepts:** Agent, Environment, Actions, Rewards, Policy, Value Functions.
- **Applications:** Robotics, game playing (e.g., AlphaGo), autonomous driving.

Evaluating Model Performance

Evaluation Metrics

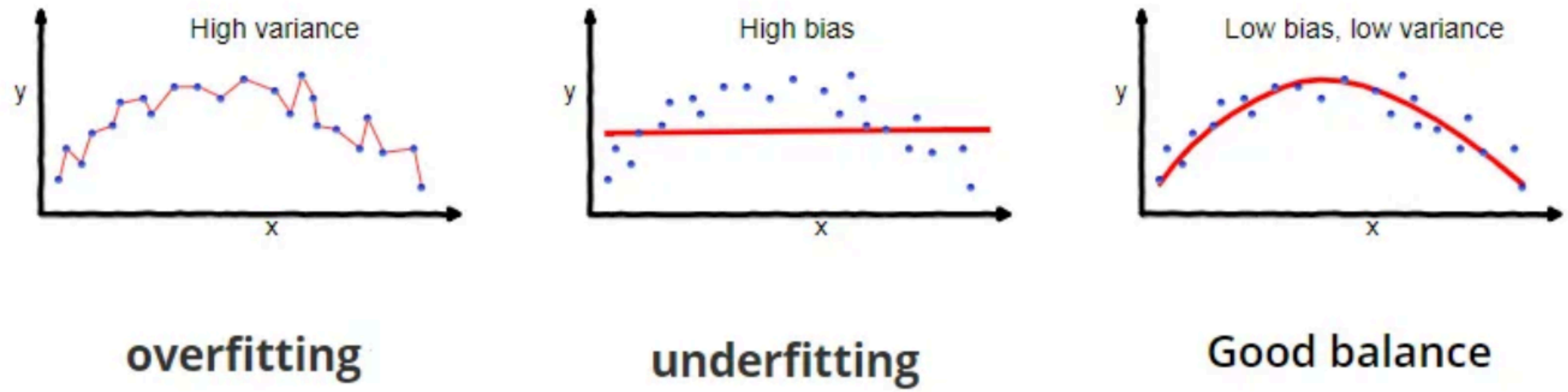
- Depending on the problem type, different metrics are used:
  - **Classification:** *Accuracy*, *Precision*, *Recall*, *F1-Score*, *ROC-AUC*.
  - **Regression:** *Mean Squared Error (MSE)*, *Mean Absolute Error (MAE)*, *R-Squared*.

Evaluation Techniques



- In order to assess model performance on unseen data, there are several techniques.
- **Holdout Method:** Split data into training and test sets.
  - Common split ratios: 70/30, 80/20, 90/10.
  - Disadvantage: Sensitive to the random split. Not all data is used for training.
- **Cross-Validation:** Split data into training and validation sets multiple times.
  - Common types: k-Fold, Stratified k-Fold, Leave-One-Out.
  - All data is used for training and validation. Reduces variance in performance estimates.
  - Disadvantage: Computationally expensive.
- **Bootstrapping:** Repeatedly sample the dataset with replacement, creating multiple “new” datasets.
  - Useful for estimating the uncertainty of a model’s performance.
  - Disadvantage: Computationally expensive. Possible bias in resampling.

The Bias-Variance Tradeoff



Represents the tradeoff between model complexity and generalization.

- **Overfitting:** Model captures noise instead of underlying patterns.
- **Underfitting:** Model is too simple to capture the data’s complexity.

> **Goal**
>
> Find the right balance to minimize error on unseen data.