Machine Learning: Concepts, Maths & Hands-On Practice

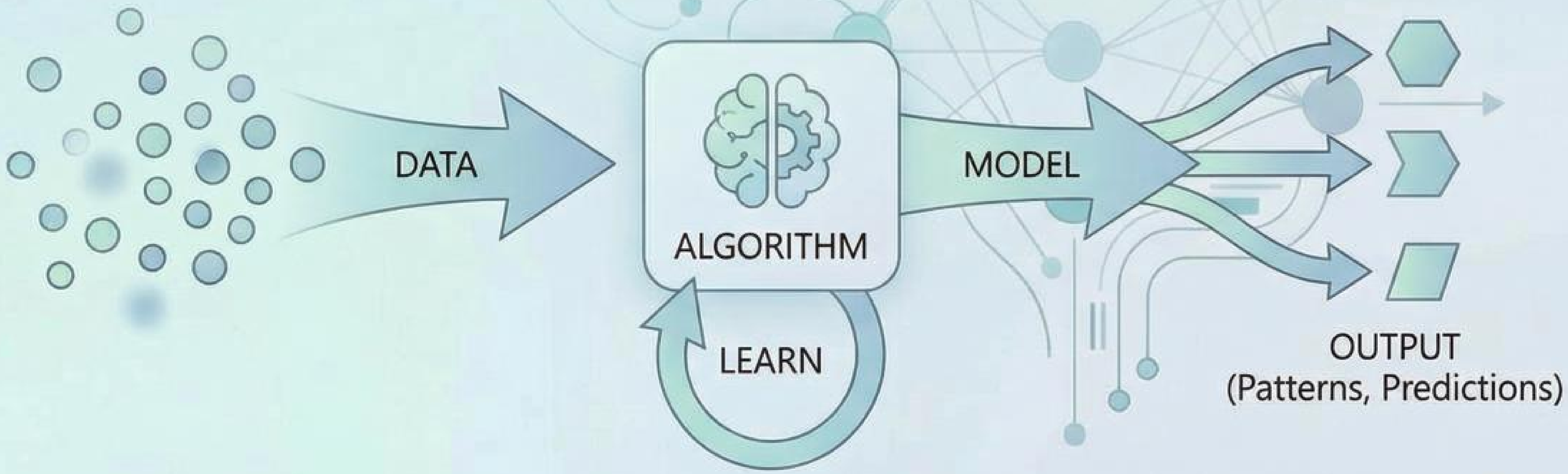
By : Google Genesis Club

Workshop Agenda

- What is Machine Learning?
- Types of ML
- ML Workflow
- Exploratory Data Analysis (EDA)
- Supervised Learning (Regression & Classification)
- Unsupervised Learning (Clustering & Dimensionality Reduction)
- Model Evaluation & Metrics
- Hyperparameter Tuning
- Model Deployment & Architecture
- Hands-On Practice Section
- Closing Summary & Q&A

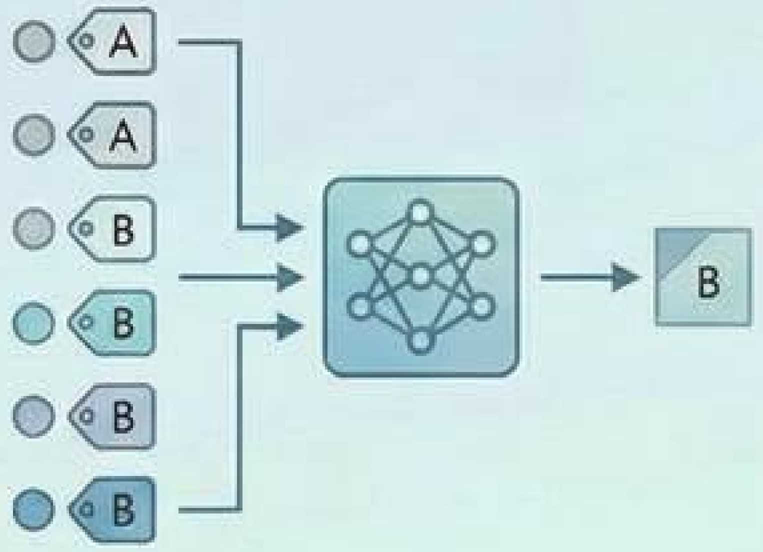
What is Machine Learning?

The science of teaching computers to learn patterns from data **without** explicit programming.



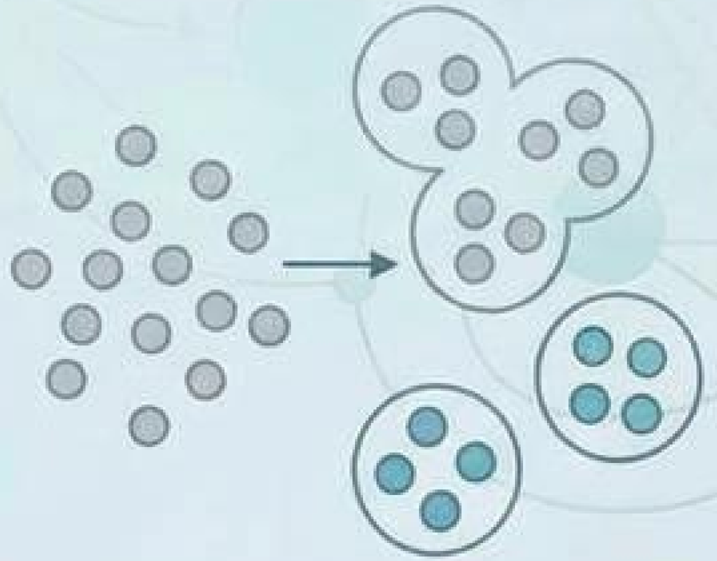
Types of Machine Learning

Supervised Learning



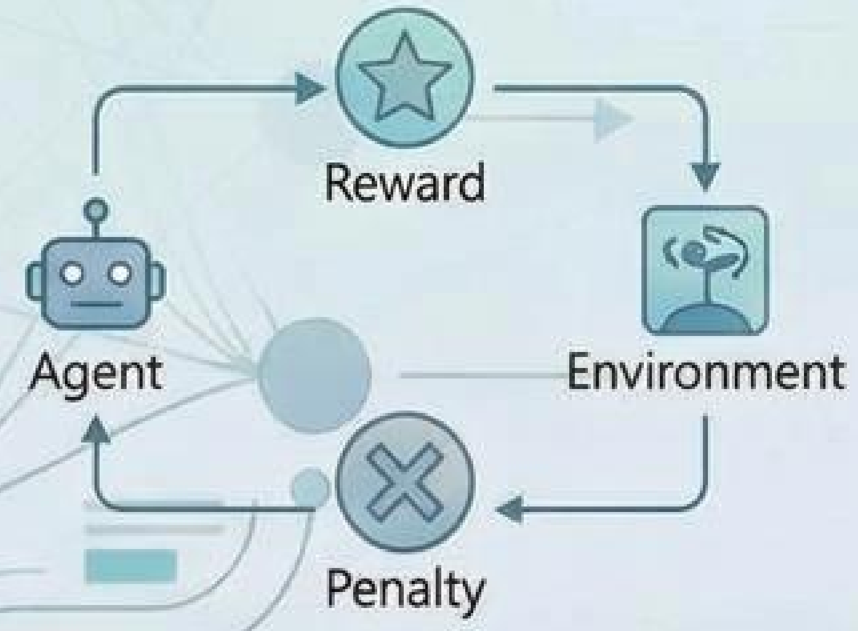
Learning from labeled data to predict outputs.

Unsupervised Learning

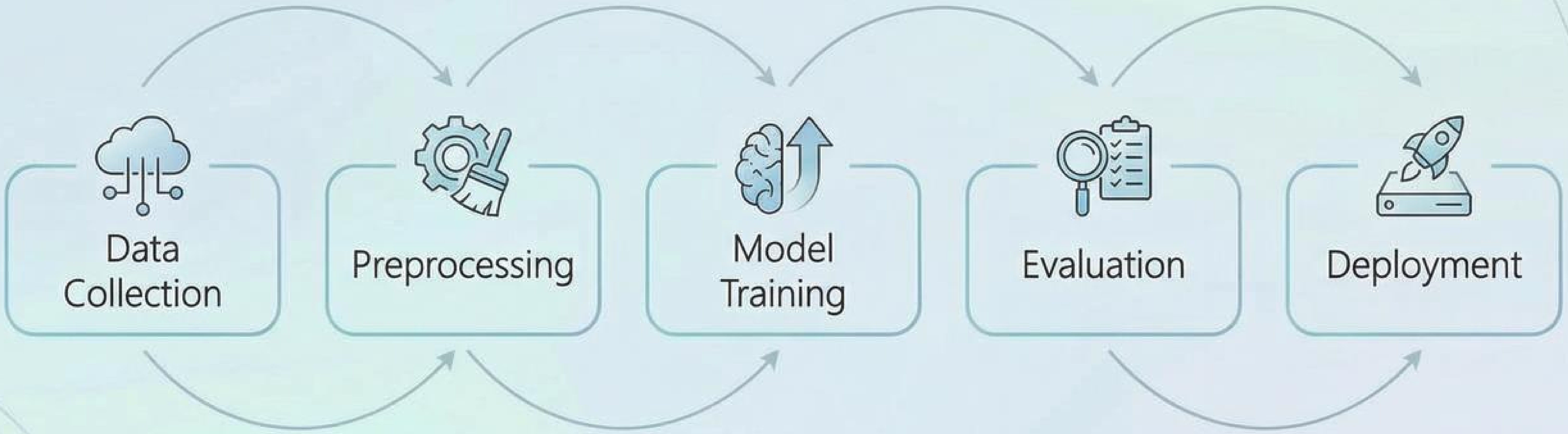


Finding hidden patterns in unlabeled data.

Reinforcement Learning



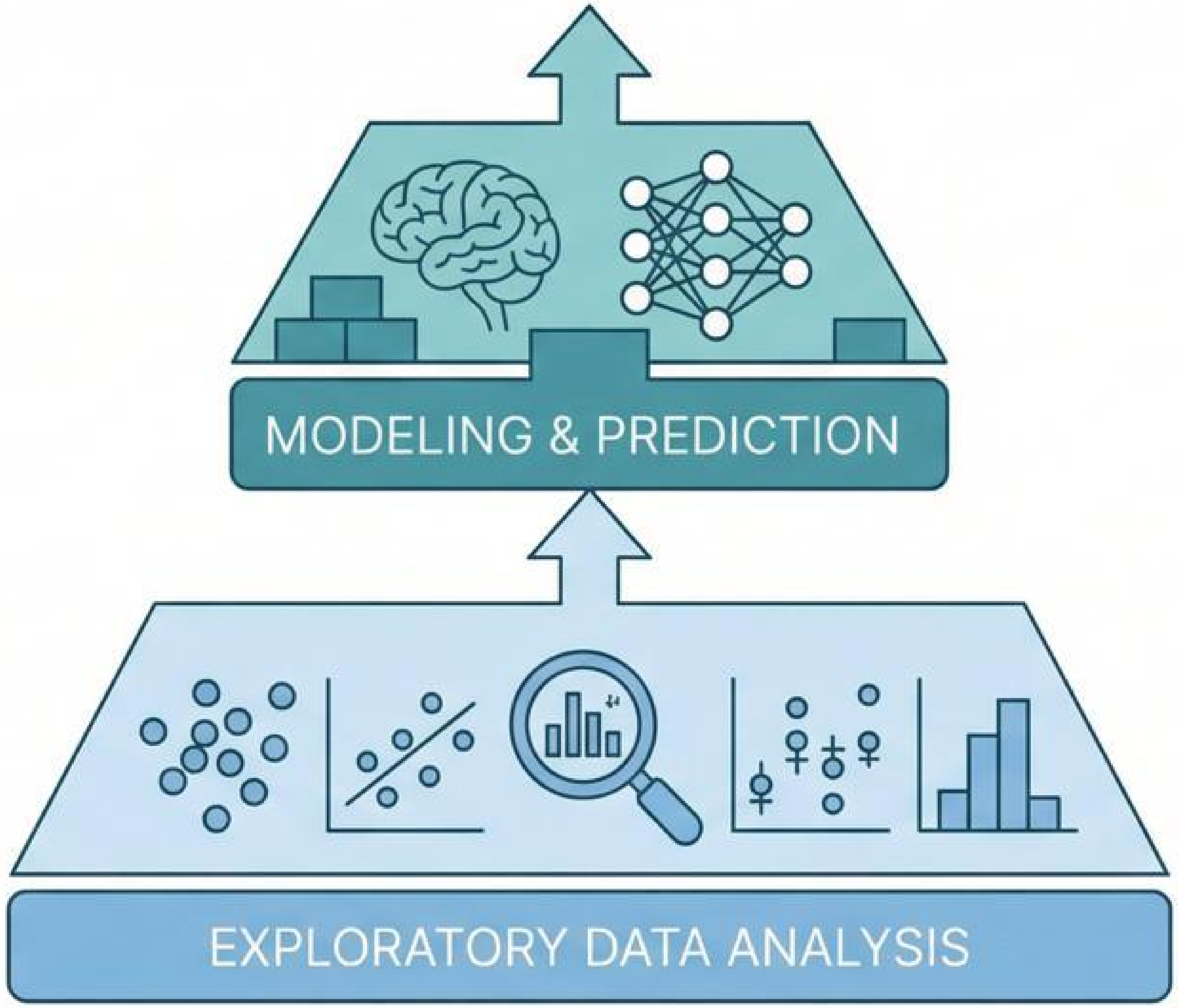
Learning through trial and error using rewards.



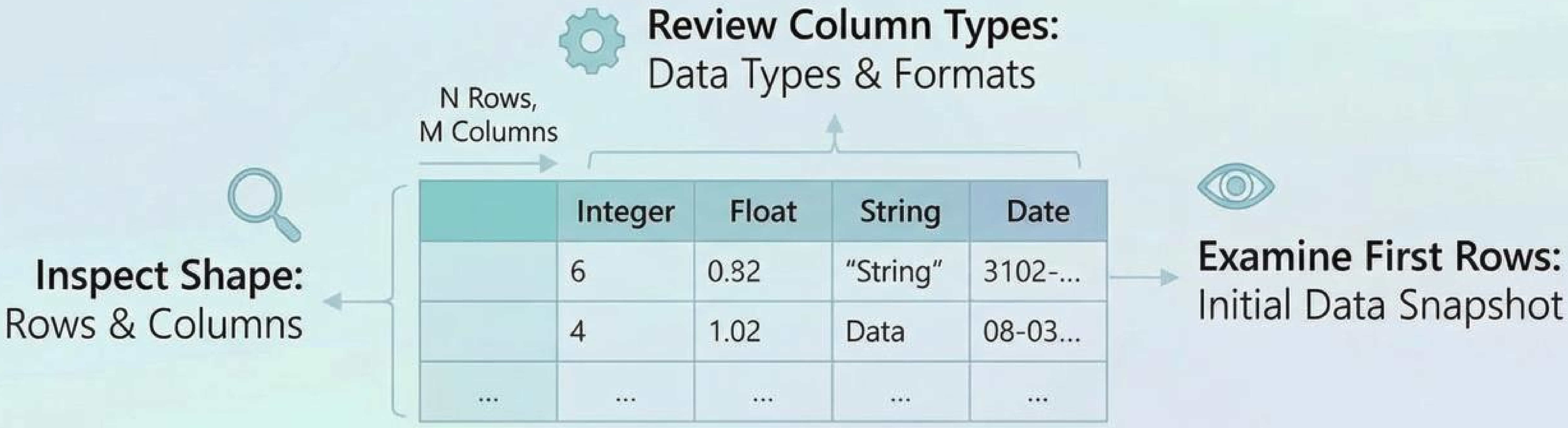
ML Workflow: Data Collection → Preprocessing → Model Training → Evaluation → Deployment

Why EDA Matters

- Reveals hidden patterns
- Detects anomalies
- Informs feature engineering



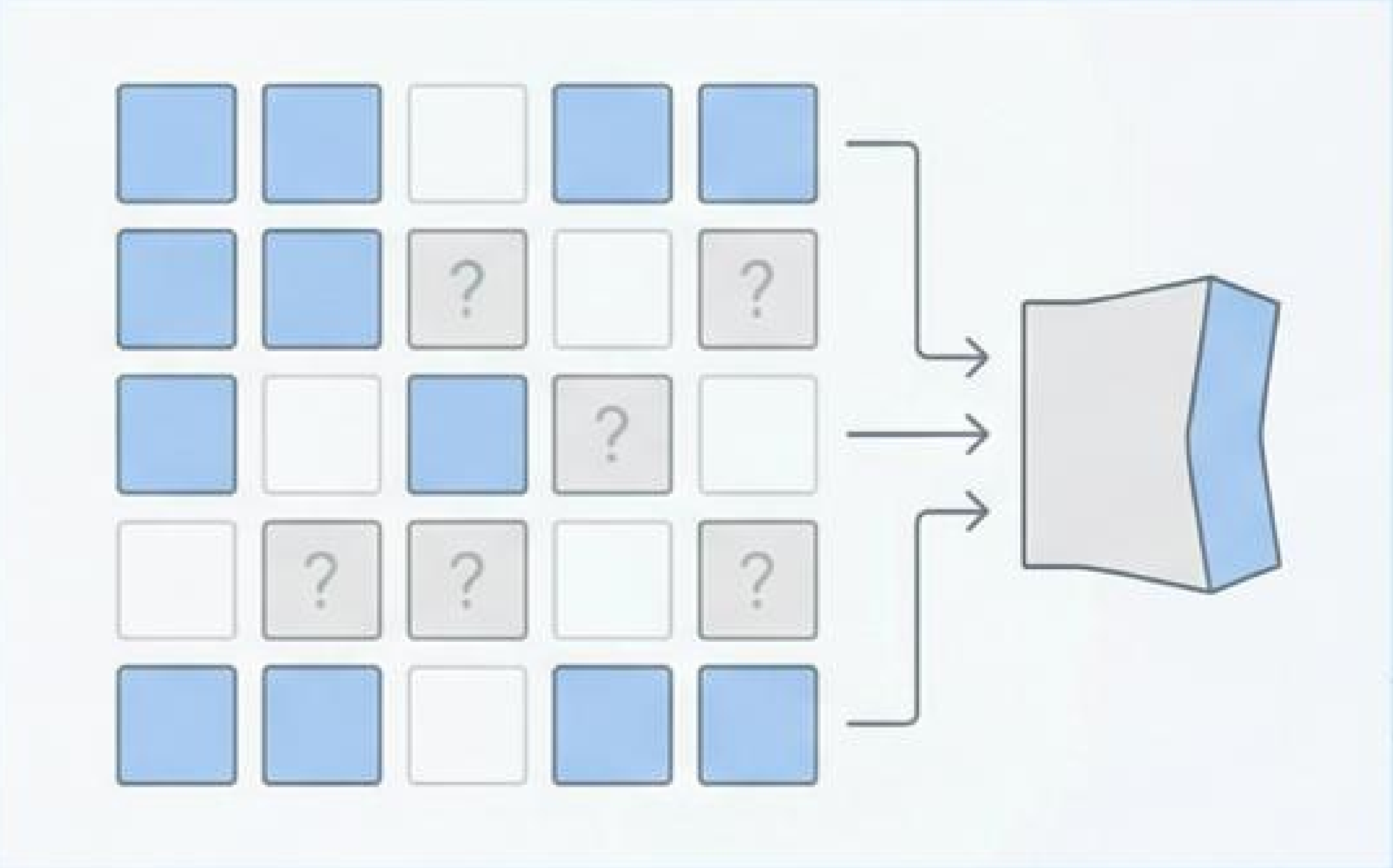
EDA: Checking Data Structure



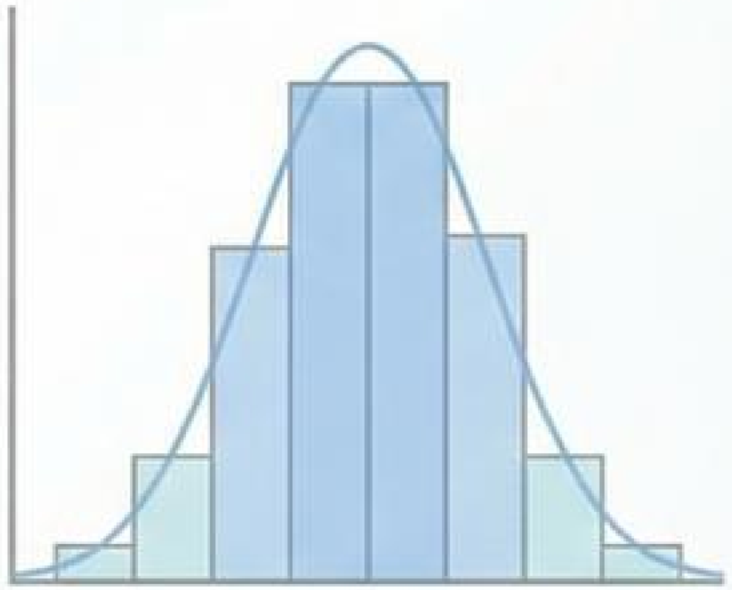
Understanding Data Organization: The crucial first step before deeper analysis or cleaning.

EDA: Missing Values

- Concept: Data points absent for certain features.
- Impact: Can introduce bias and reduce model accuracy.
- Detection: Essential to identify patterns before handling.

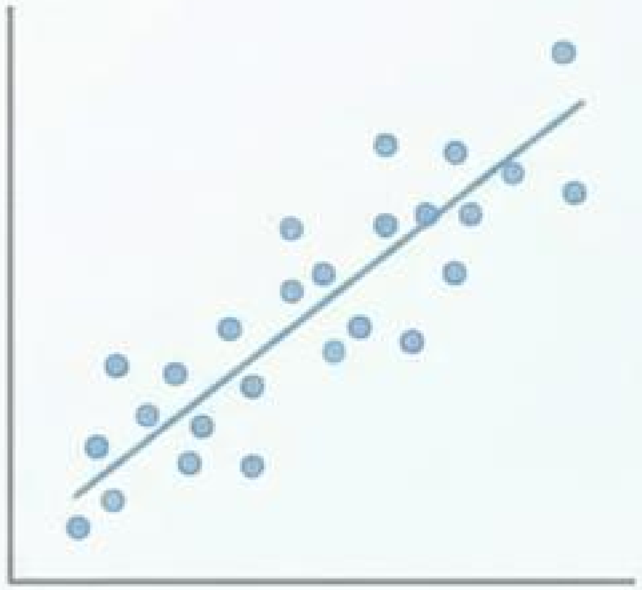


EDA: Data Visualization



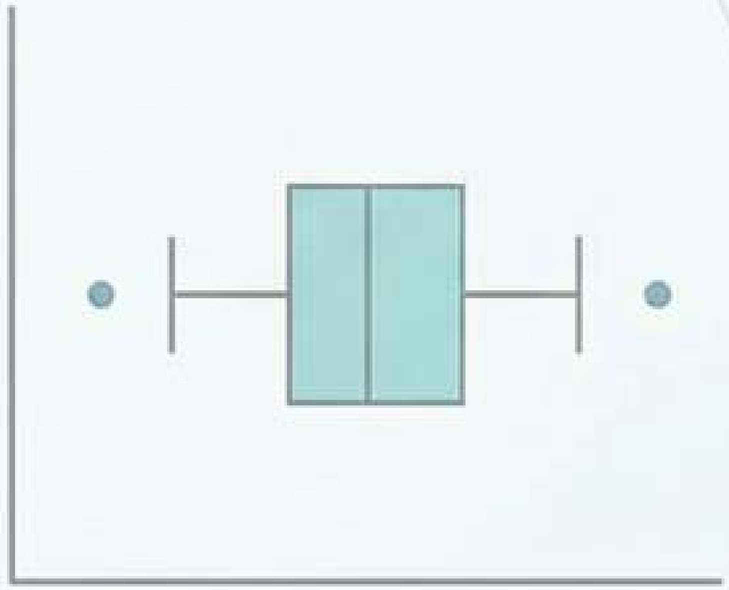
Distributions

Reveals data spread & frequency (e.g., Histograms).



Relationships

Shows correlation between variables (e.g., Scatter Plots).



Outliers & Spread

Identifies anomalies & range (e.g., Boxplots).

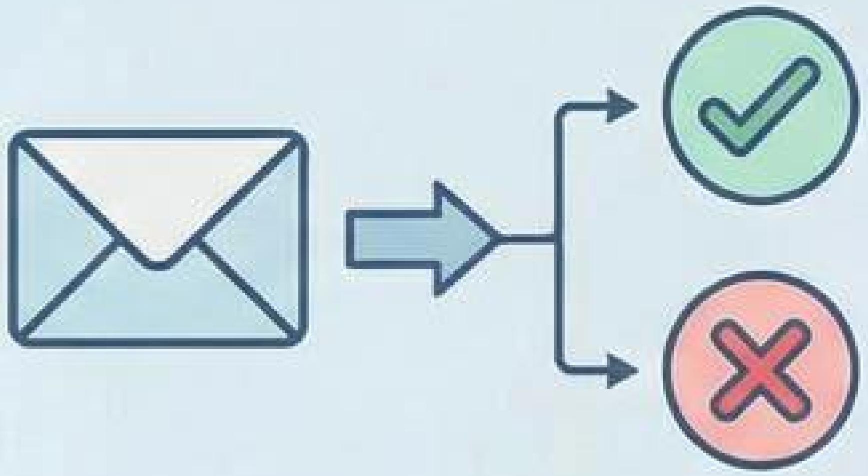
Goal: Generate actionable insights through simple, interpretable visuals.

Supervised Learning: Learning from Labeled Data

A type of machine learning where models learn from a dataset with known inputs and correct outputs to predict future outcomes.

Example 1: House Price Prediction

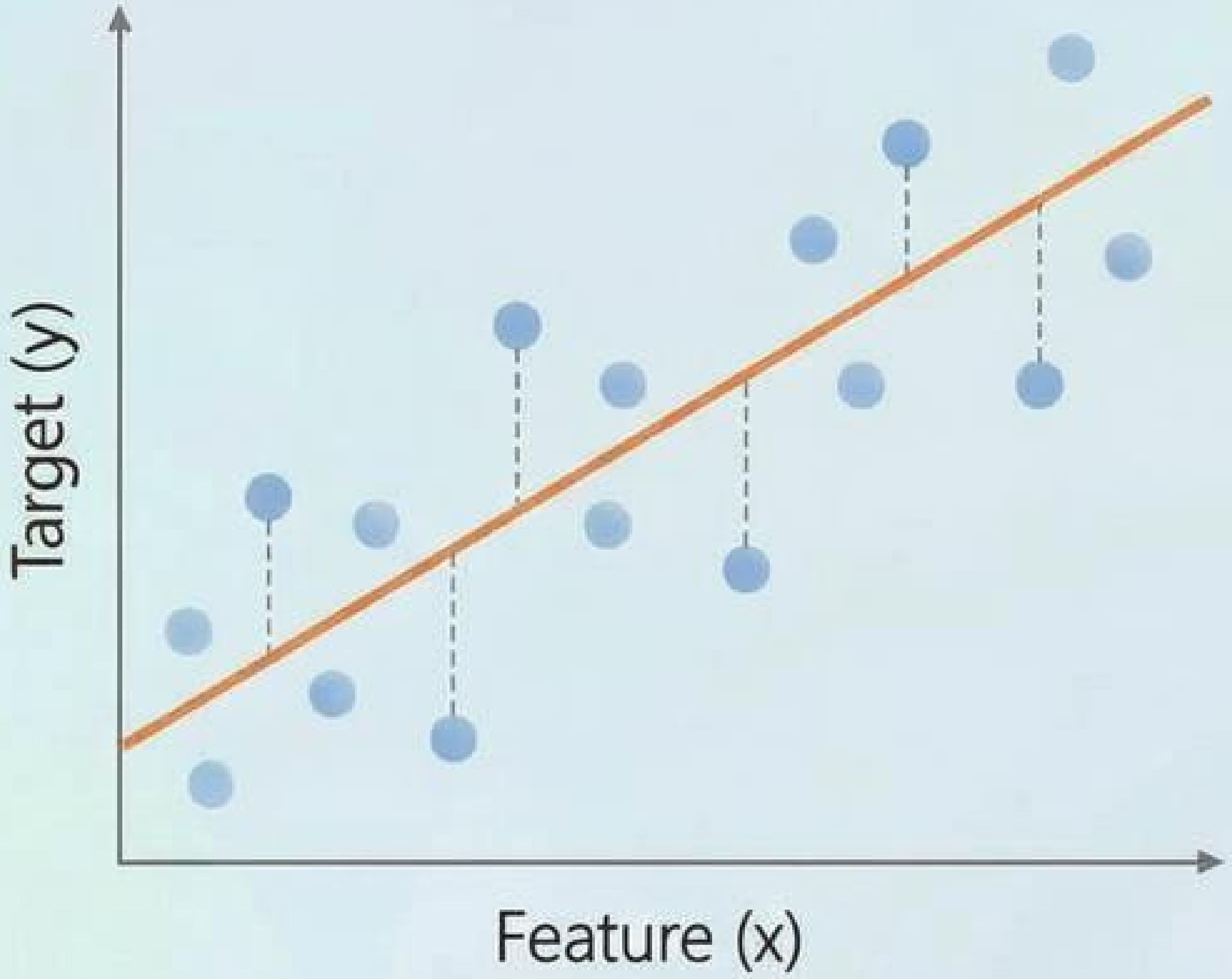
Input: Features like size, location.
Output: Predicted price.



Example 2: Spam Detection

Input: Email content.
Output: Classification (Spam/Not Spam).

Linear Regression: Concept

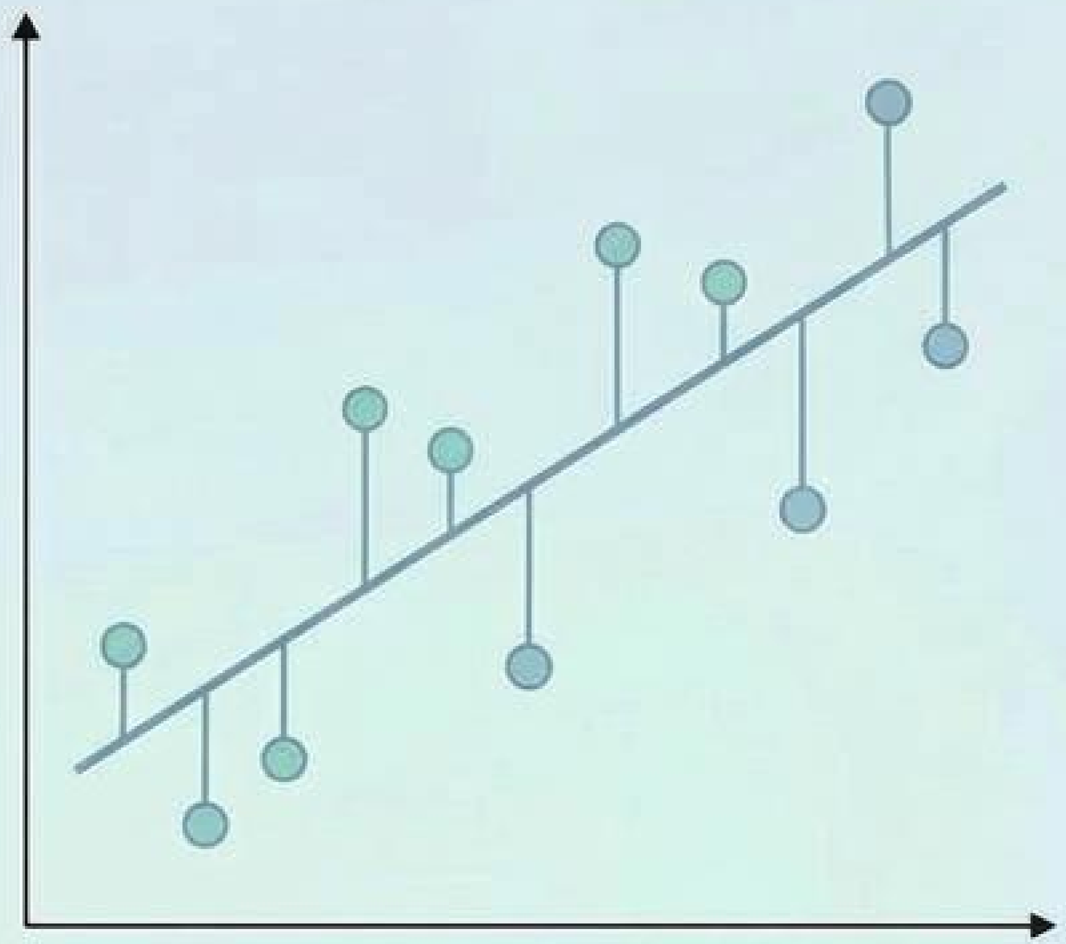


Predicts continuous values by fitting a straight line through data.
The line minimizes the total prediction error.

Linear Regression – Maths

Mean Squared Error (MSE): Loss Function

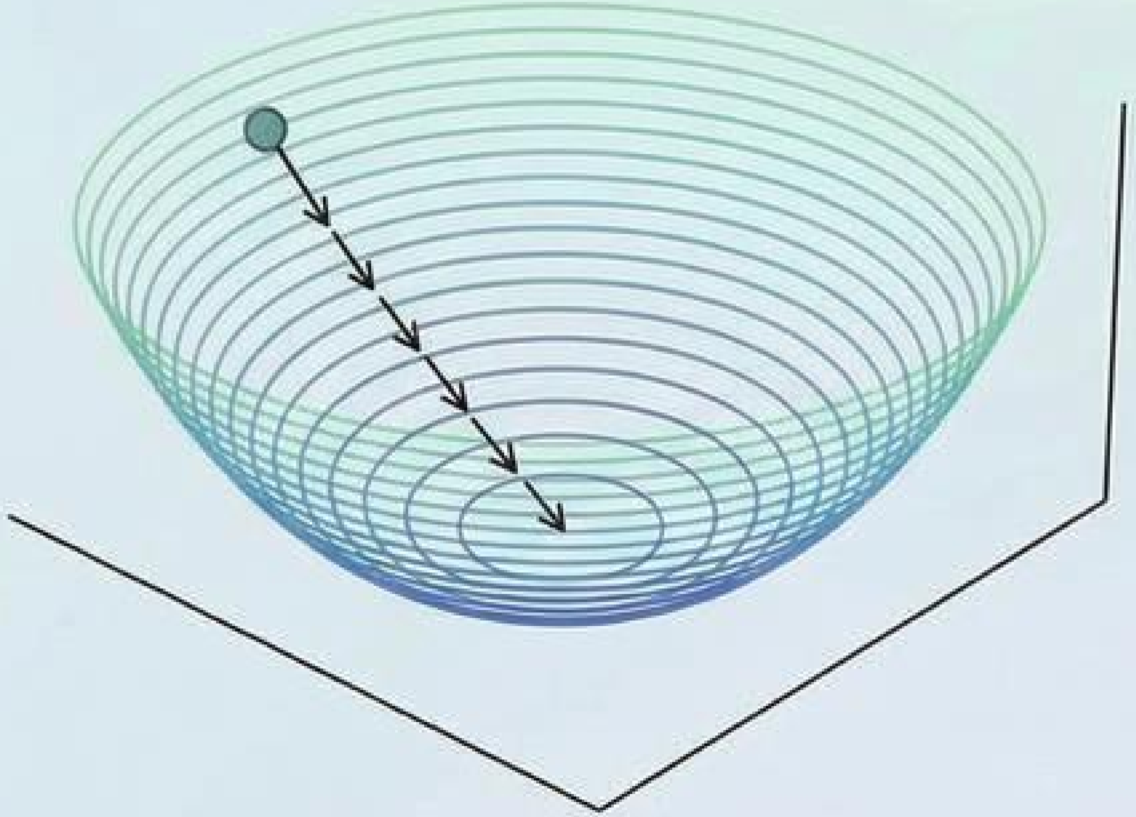
Measures the average squared difference between predicted and actual values.



$$MSE = \frac{1}{n}\sum (y_i - \hat{y}_i)^2$$

Gradient Descent: Optimization Method

Intuition:

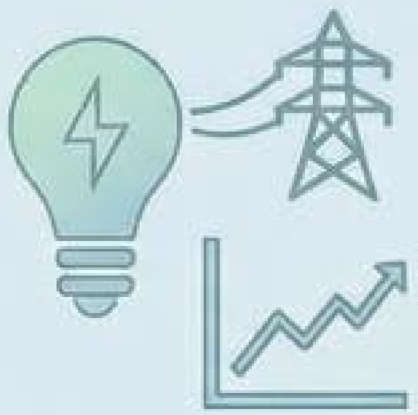


Iteratively updates weights to minimize the loss function (MSE). Gradients guide weight updates.

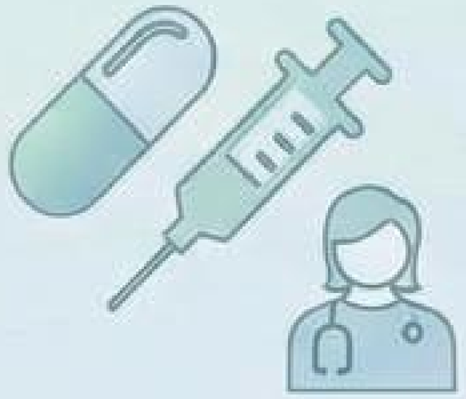
$$w_{new} = w_{old} - learning\_rate * gradient$$

Linear Regression – Where It's Used

Sales
Forecasting



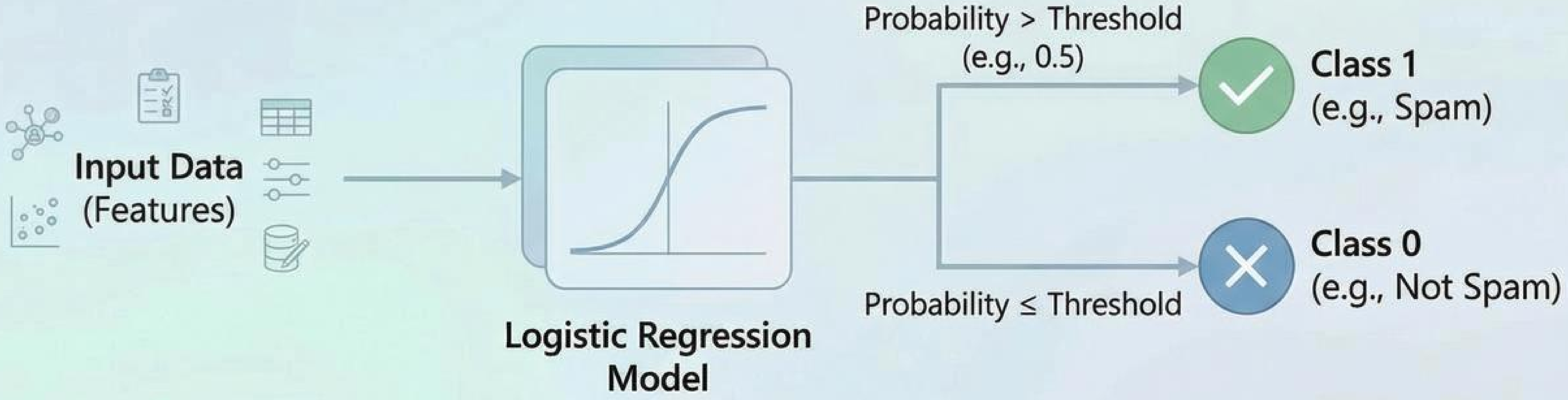
Energy Consumption
Prediction



Medical Dosage
Estimation

Logistic Regression – Concept

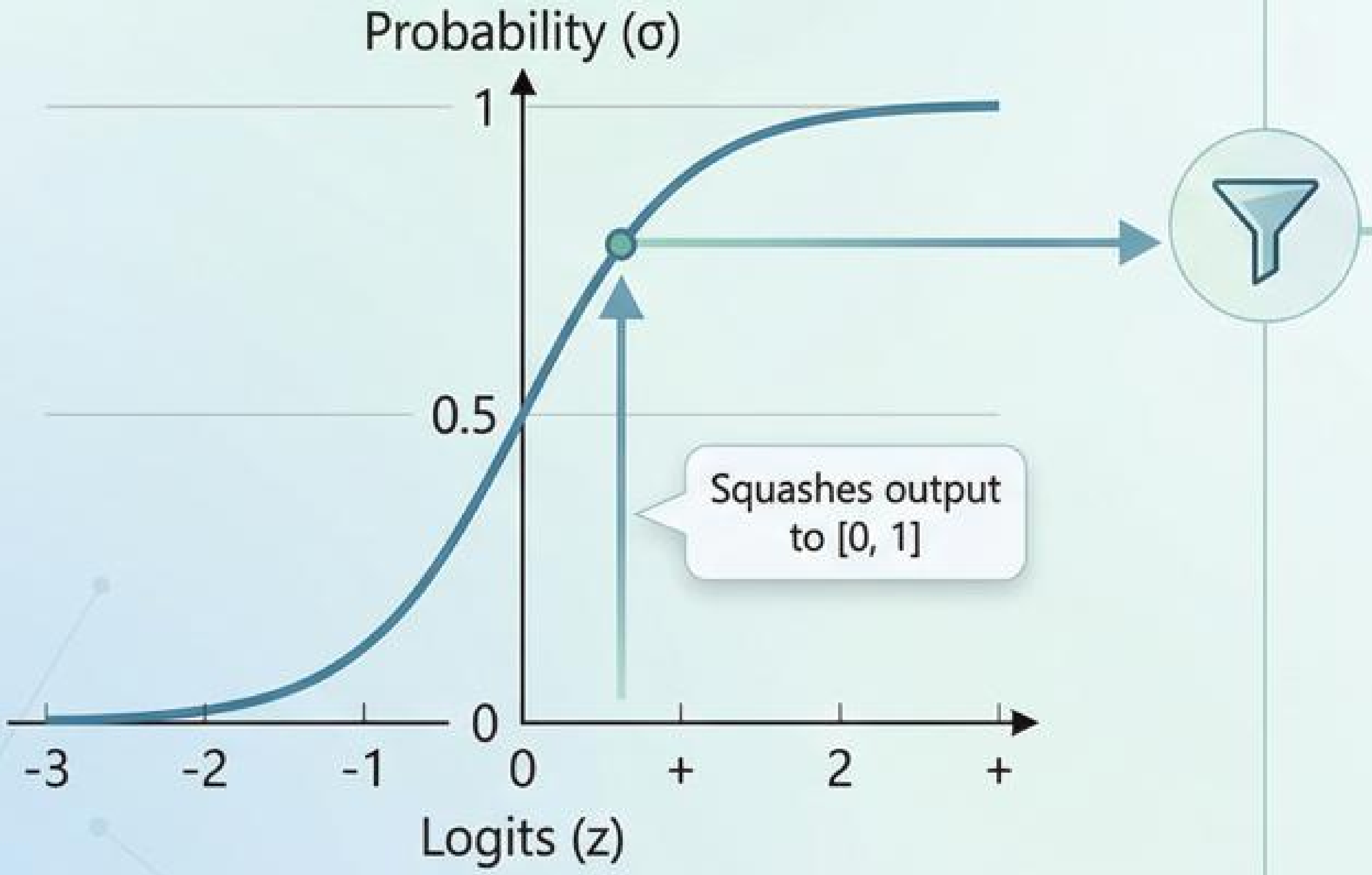
A classifier that predicts the probability of a binary outcome (0 or 1).



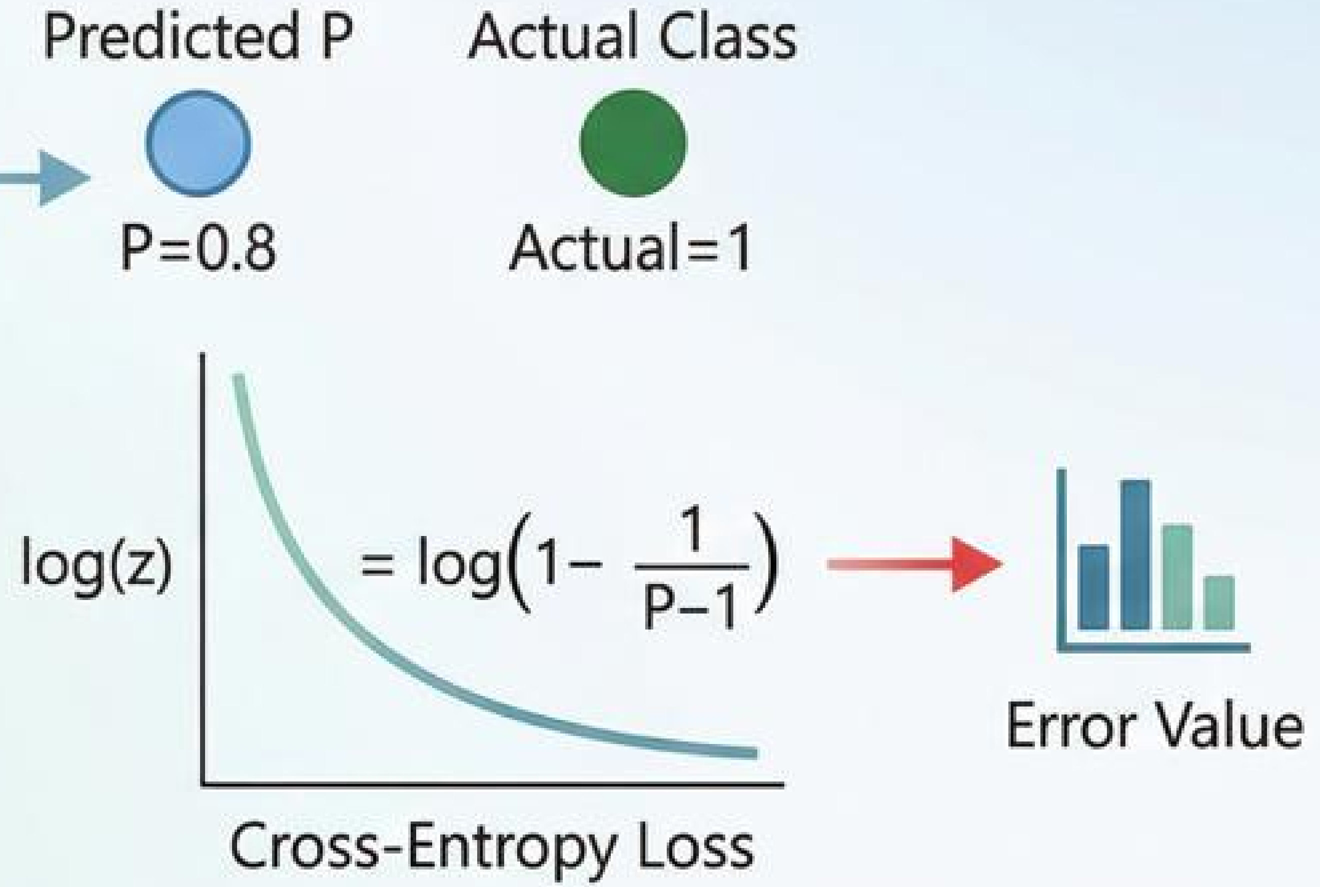
Output is a probability score between 0 and 1, utilized for binary classification.

Logistic Regression – Maths

Sigmoid Function: Logits to Probabilities

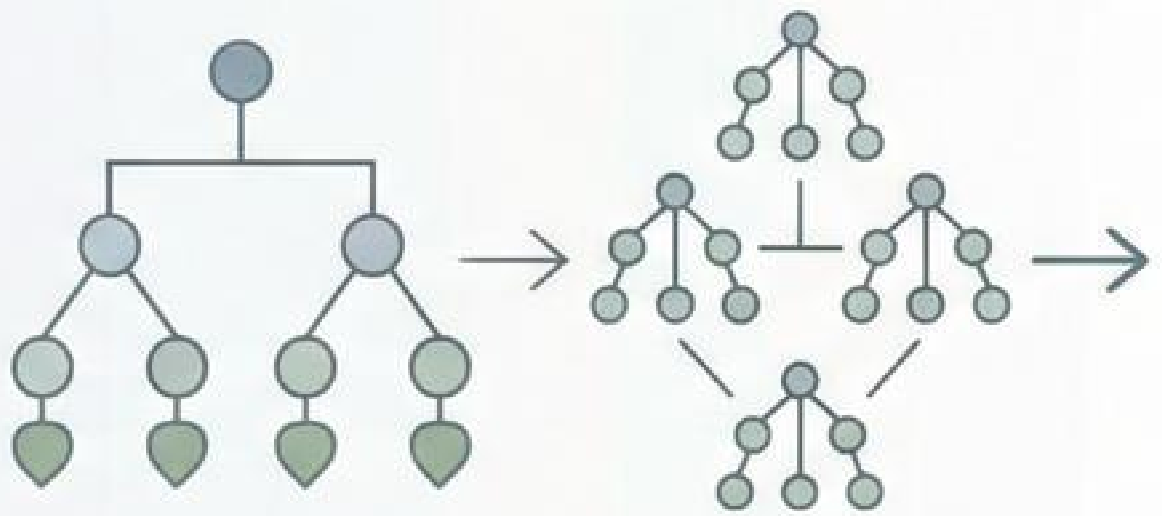


Cross-Entropy: Measuring Error



DECISION TREES & RANDOM FORESTS

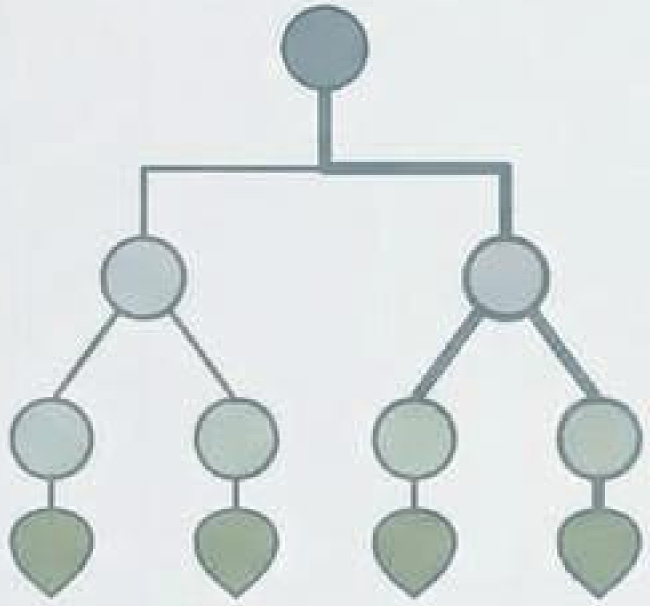
Concept, Strengths, and Application in Data Analysis



SINGLE DECISION TREES: BALANCE & CAUTION

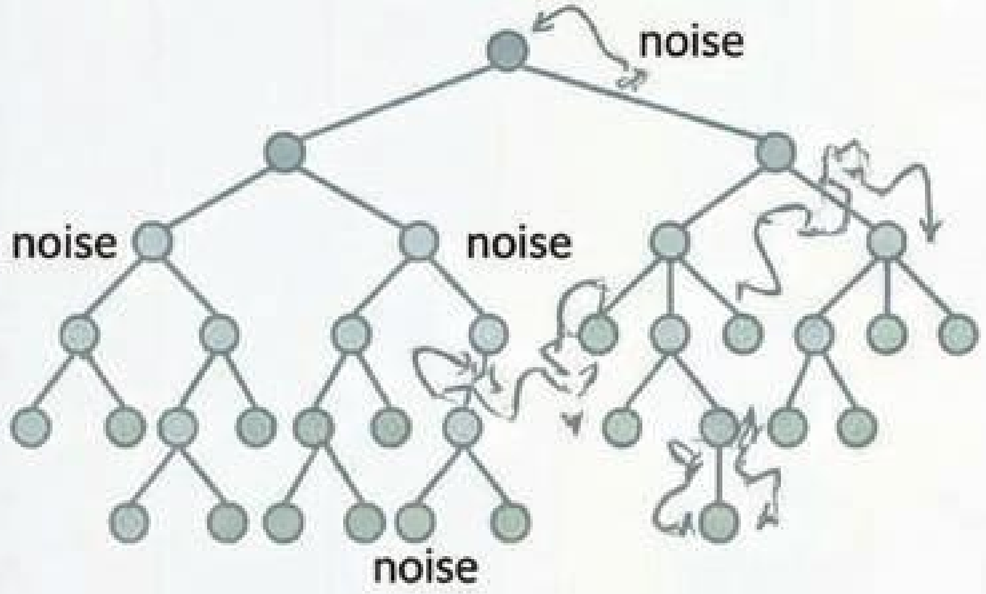
Intuition vs. Overfitting

TRANSPARENT & INTUITIVE RULES



- Easy to interpret.
- Rules anyone can trace.
- Excel on tabular data.

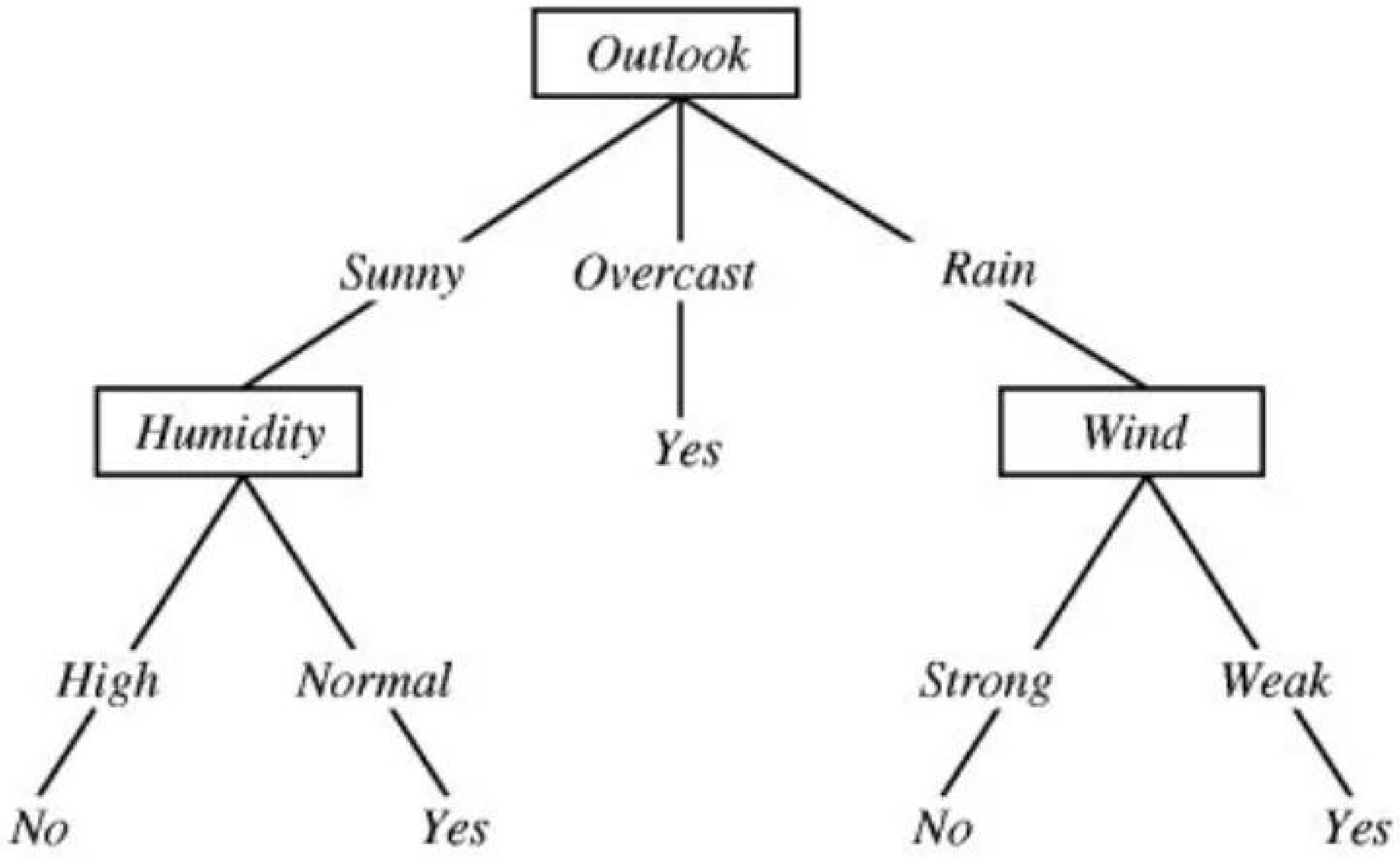
OVERFITTING & SENSITIVITY



- Flexible, but memorize noise.
- High variance on new data.
- Wilt under slight changes.

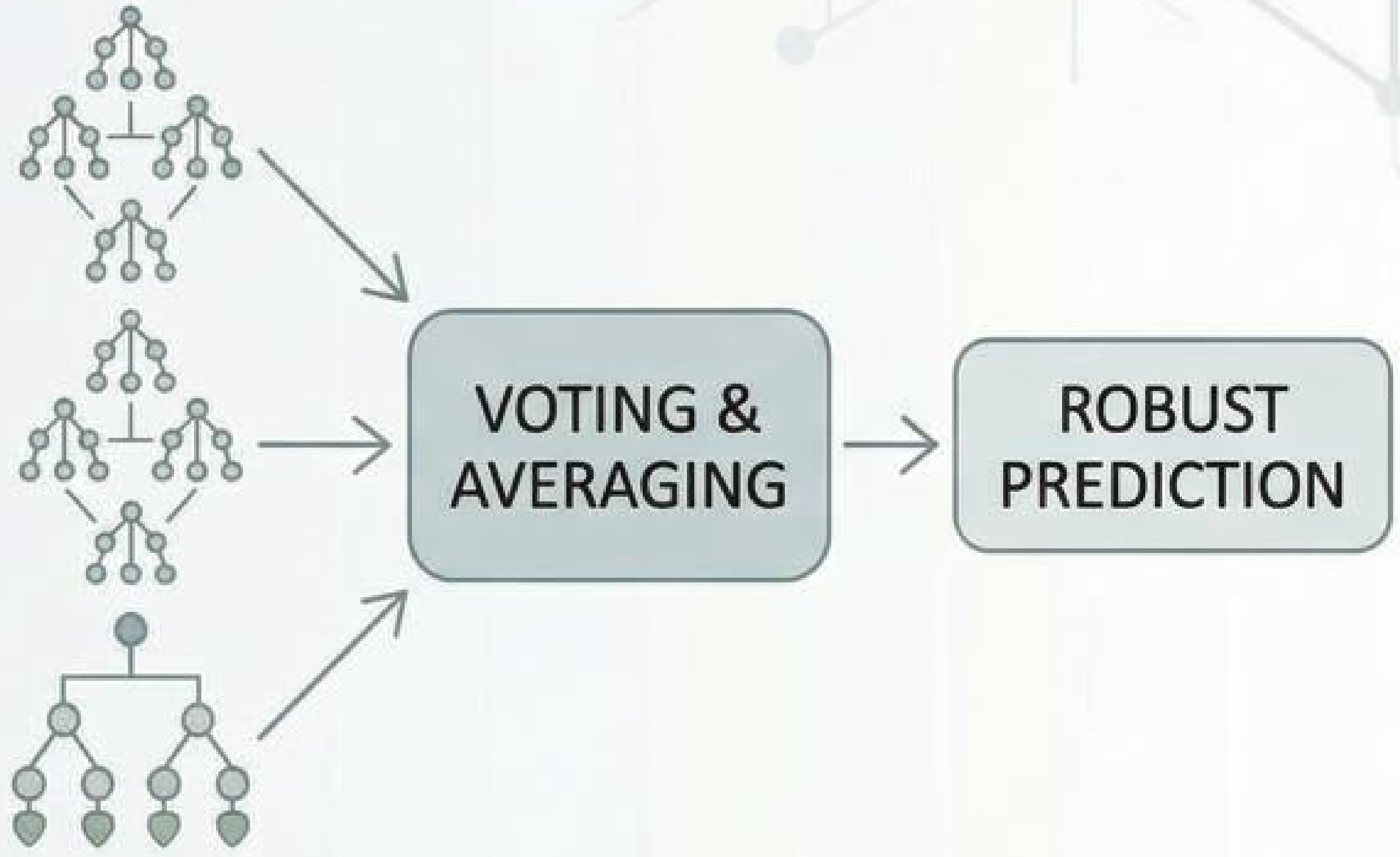
Intuition must be balanced against caution.

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



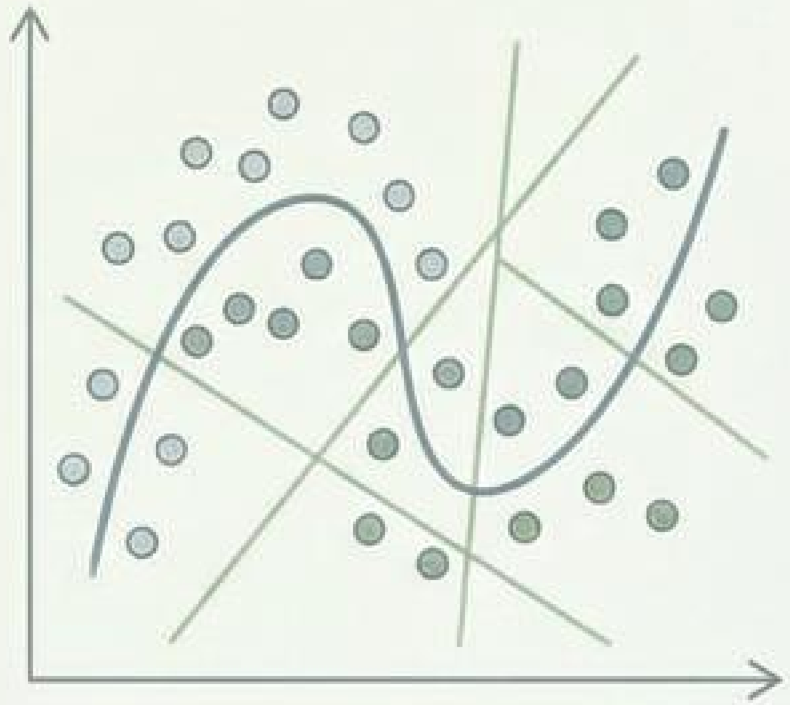
RANDOM FOREST – CONCEPT

A random forest grows numerous deep trees on bootstrapped samples and random subsets of features, then lets them vote; averaging these noisy, independent predictions cancels errors, tames variance, and keeps the decision boundary smooth without complicated math, turning overfitting into robustness.



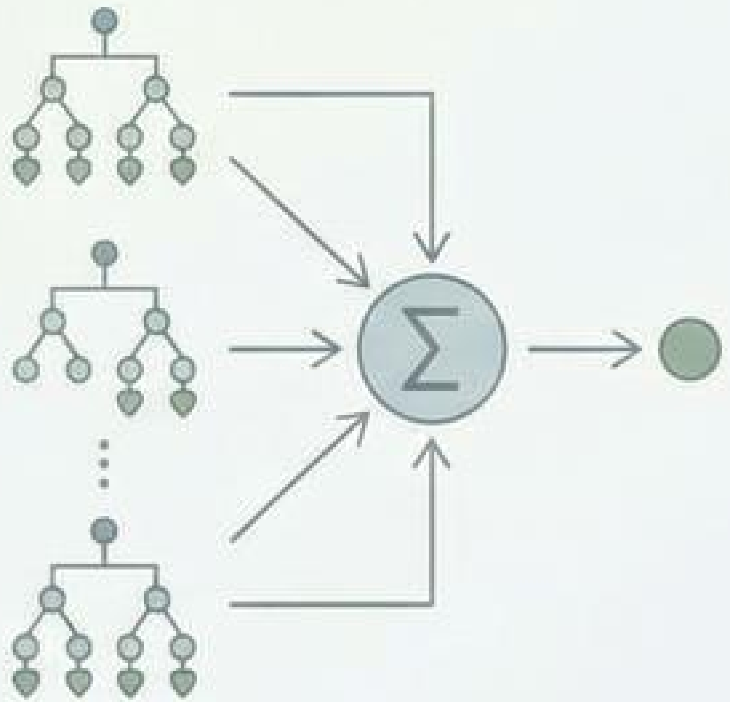
Random Forest – Why It Works Well

Handles Non-Linear Patterns



Ensemble captures curved, irregular patterns that stump a single tree.

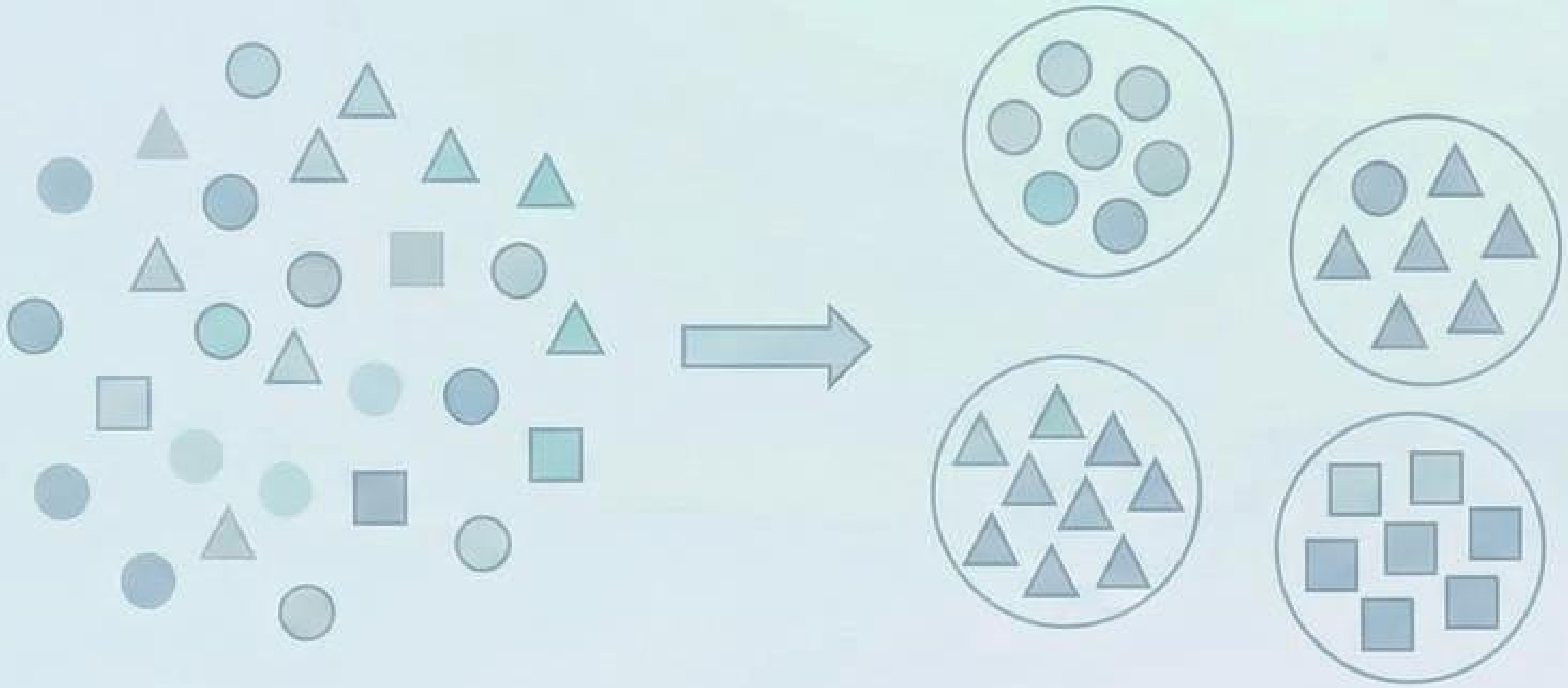
More Stable & Reliable



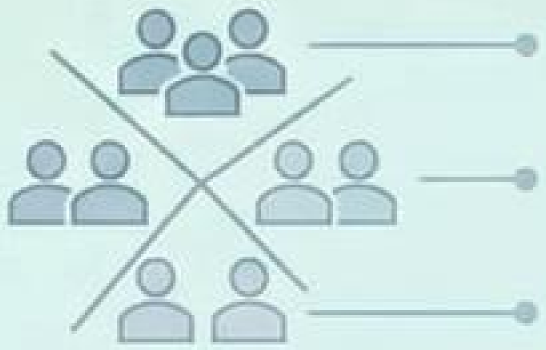
Aggregation shields predictions from outliers and small data shifts for higher accuracy.

Unsupervised Learning Overview

Finding hidden patterns in unlabeled data.



Customer Segmentation



Group A: spending habits

Group B: medium habits

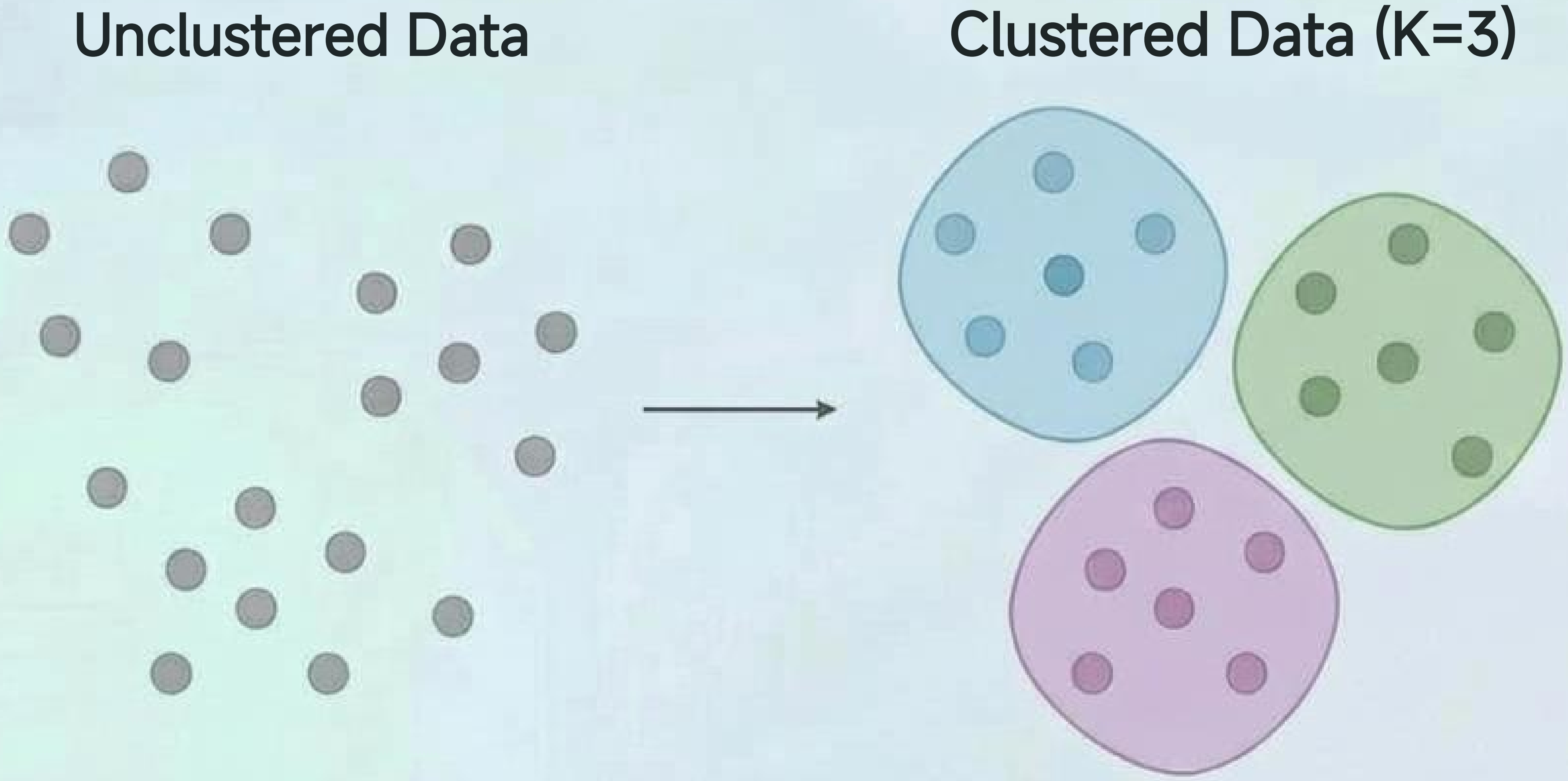
Group C: low

Anomaly Detection

Anomaly/Outlier

K-Means—Concept

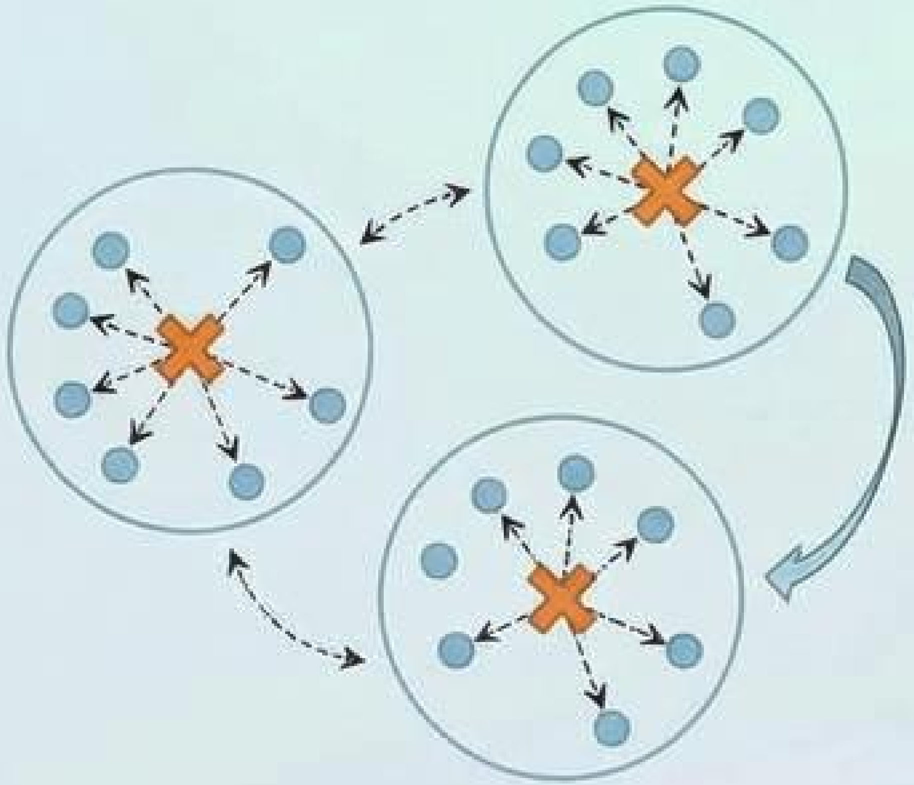
Grouping similar points into K clusters.



K-Means – Maths: Distance Minimization & Centroid Update

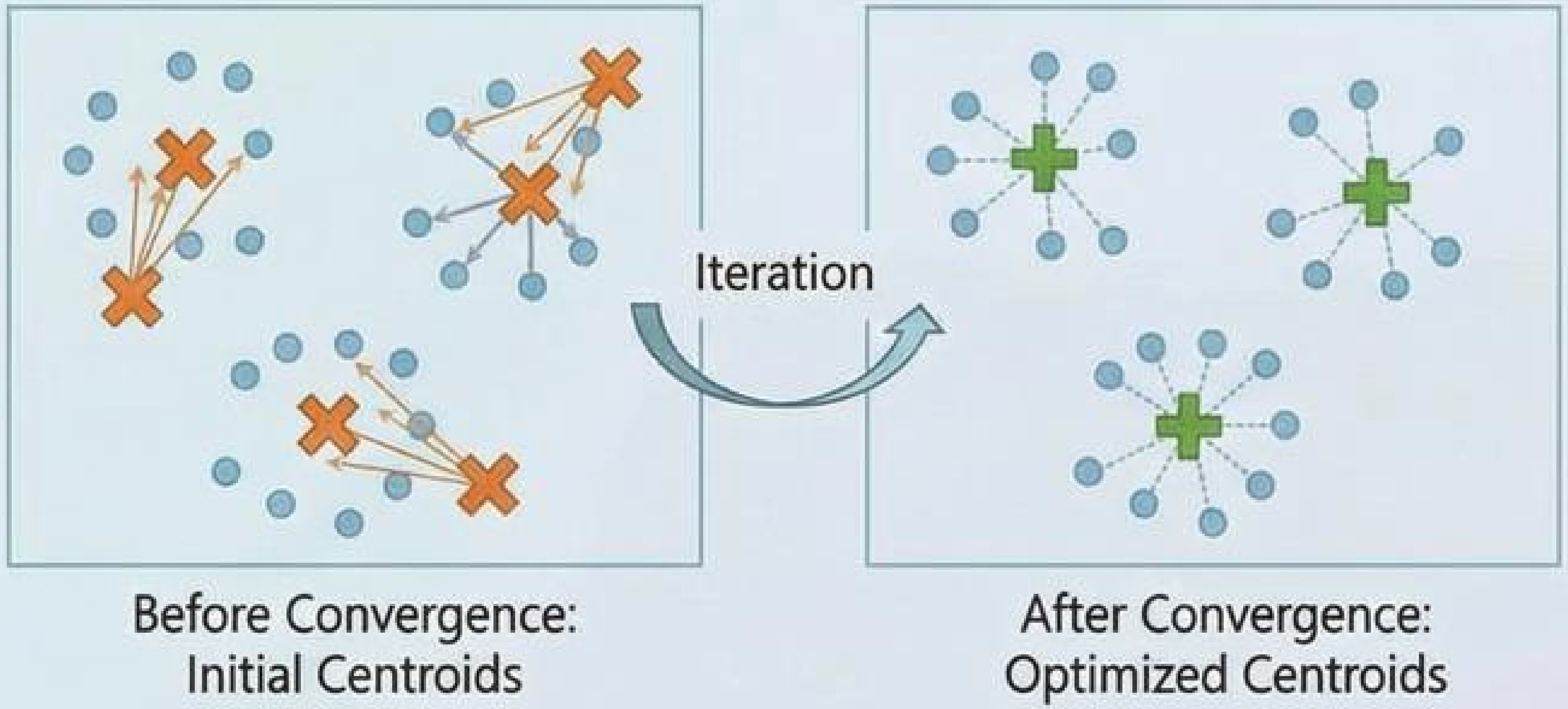
Distance Minimization

Goal: Minimize the sum of squared distances between each data point and its assigned cluster centroid.



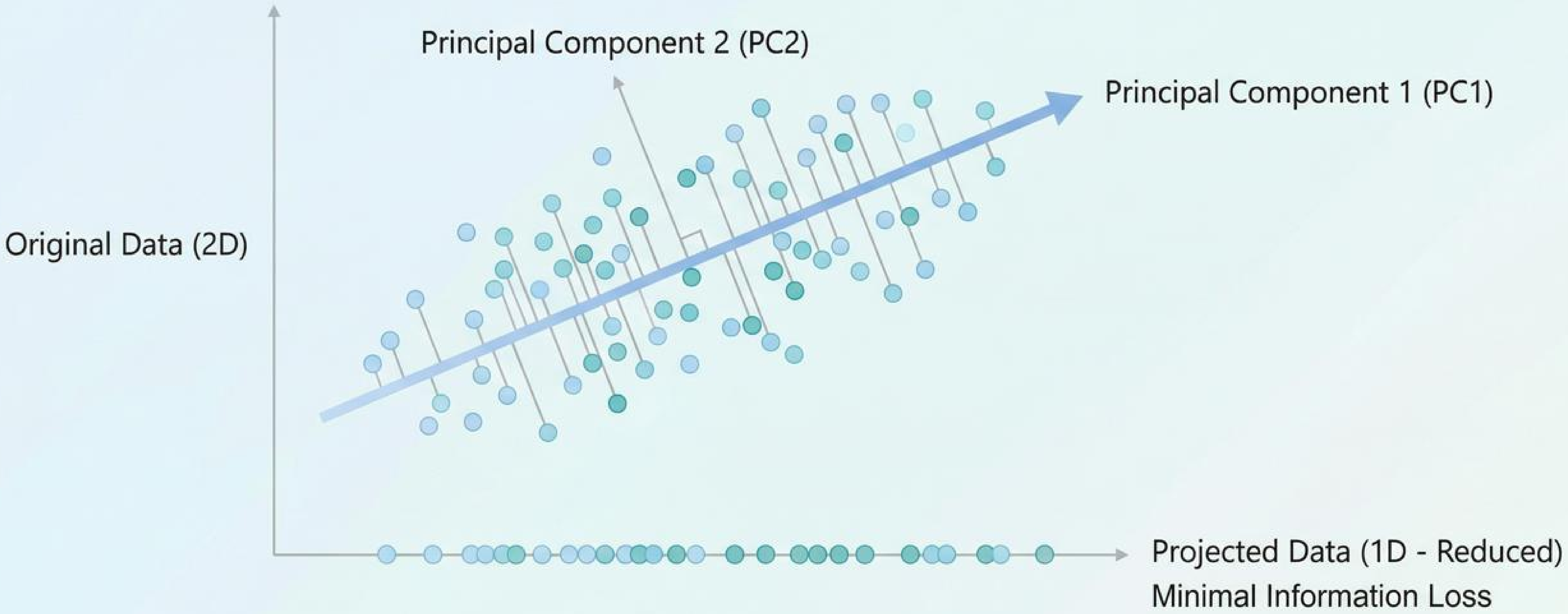
Centroid Update (Iterative Process)

Recalculate centroid positions by averaging all points within each cluster. Repeat until convergence.



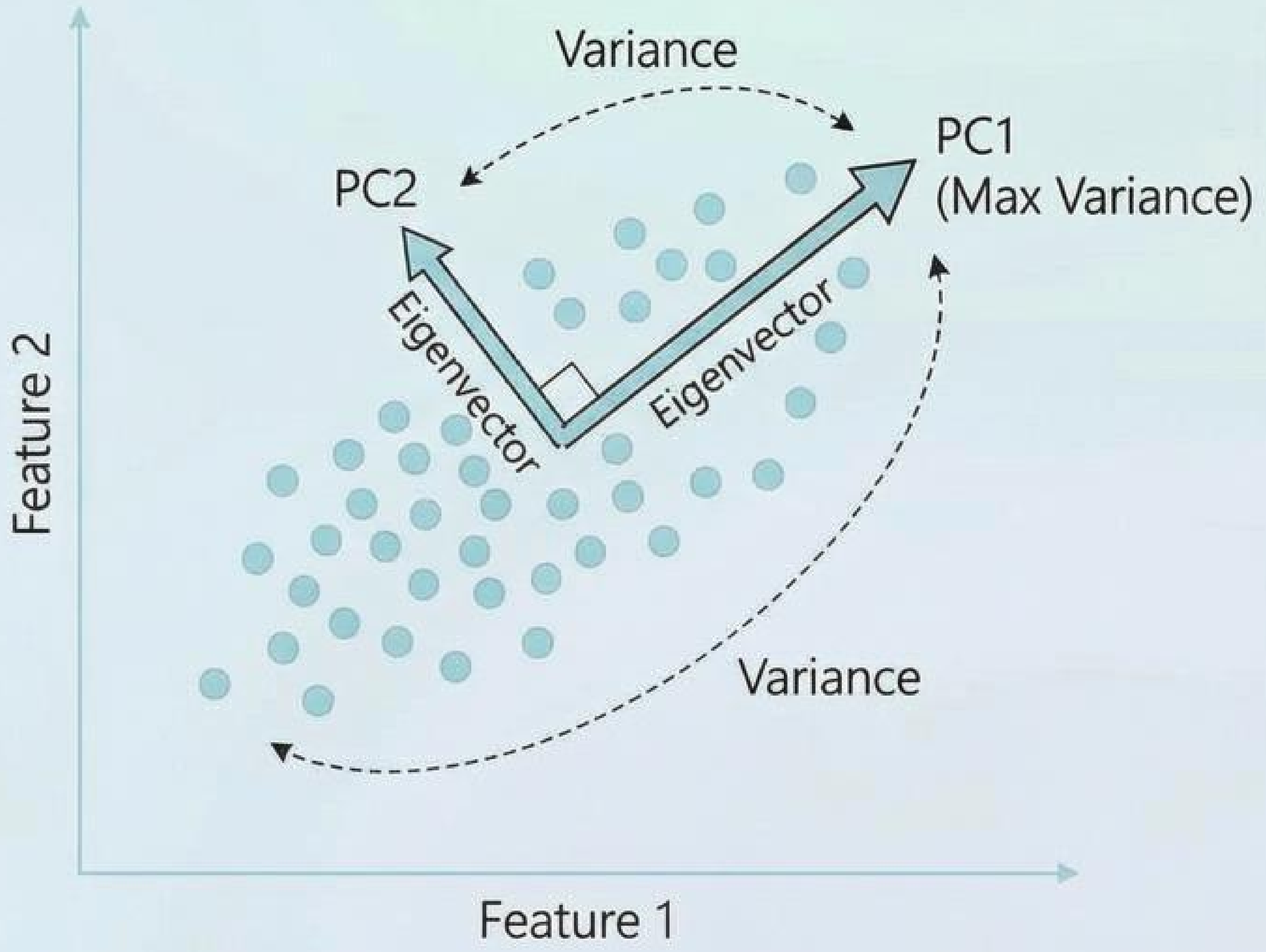
PCA – Concept

Dimensionality Reduction: Simplifying data by reducing the number of variables while retaining essential information (variance).



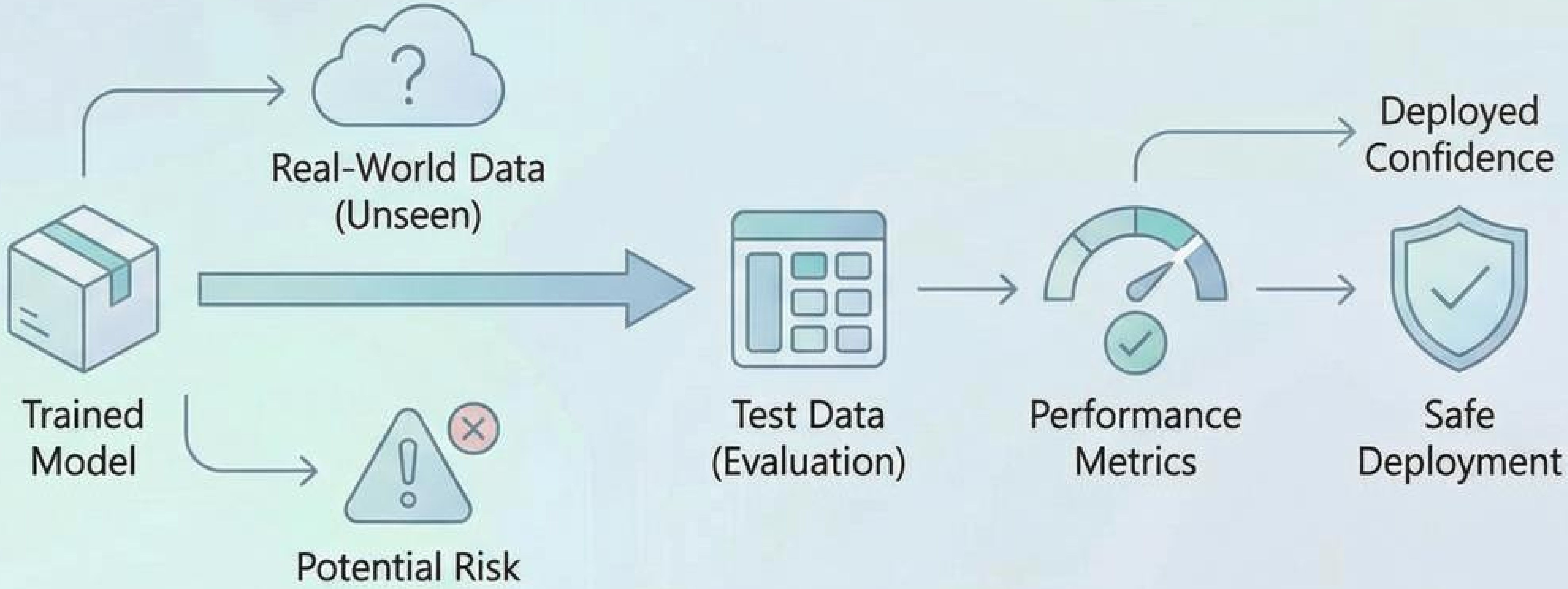
PCA – Maths (Light)

PCA finds directions of maximum variance. These directions are called **eigenvectors**.



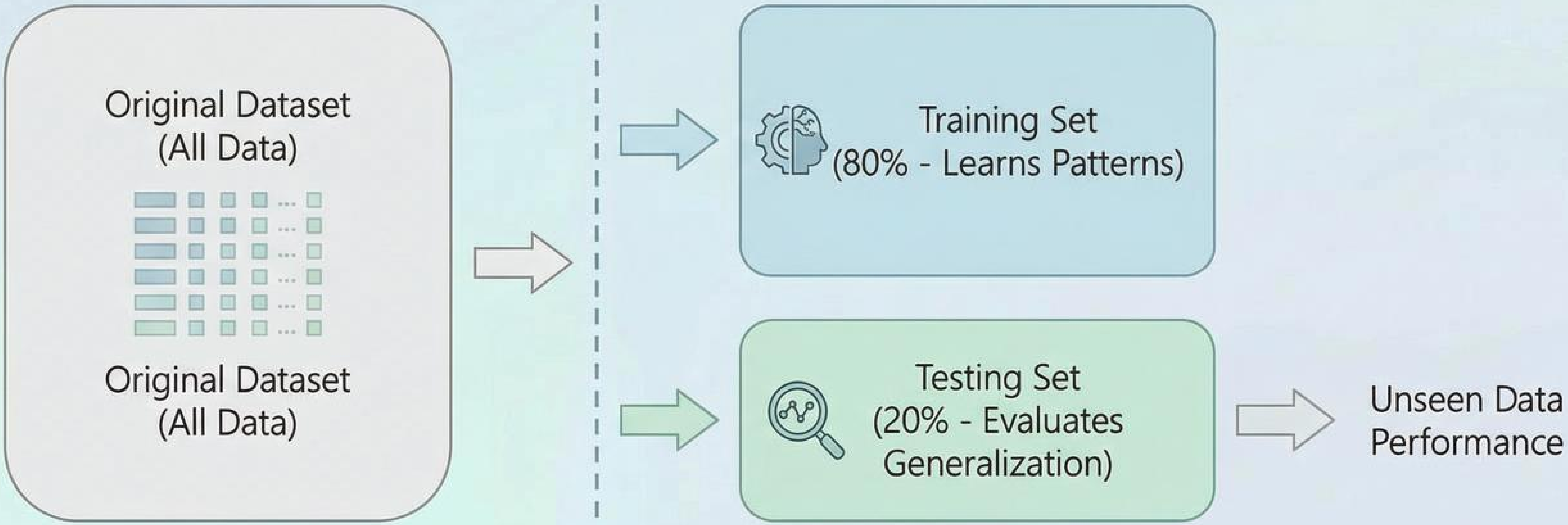
Model Evaluation: Why It Matters

Quantifies performance and guards against overfitting—essential before deployment.



Model Evaluation: Train/Test Split

Splitting data simulates real-world performance and reveals true generalization ability.



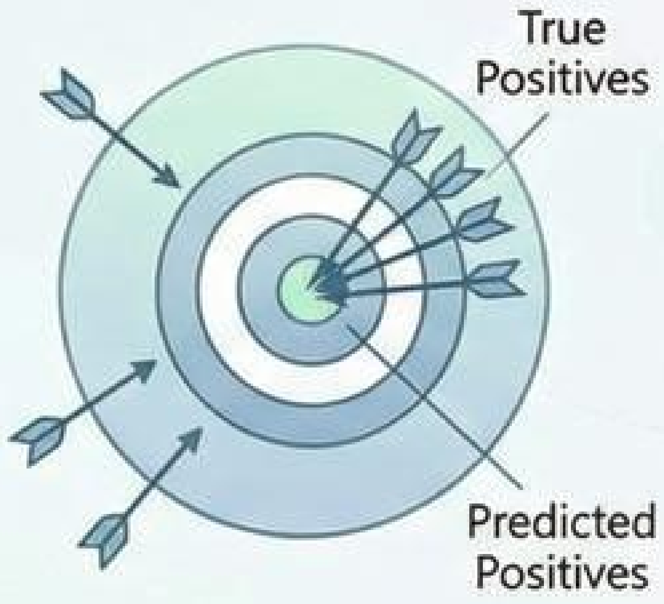
Model Evaluation: Classification Metrics

Accuracy

$$\frac{\text{Correct}}{\text{Total}}$$

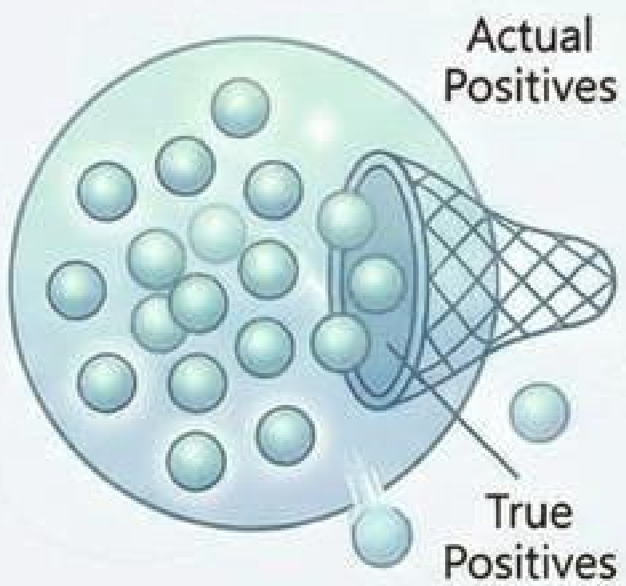
Overall correctness of predictions.

Precision



Correctness of positive predictions.

Recall



Ability to find all positive instances.

F1-Score

Harmonic mean of Precision and Recall, balancing both.

Model Evaluation: Regression Metrics

MSE (Mean Squared Error)

Measures average squared difference between predicted and actual values.

RMSE (Root Mean Squared Error)

Square root of MSE, interpretable in the same units as the target variable.

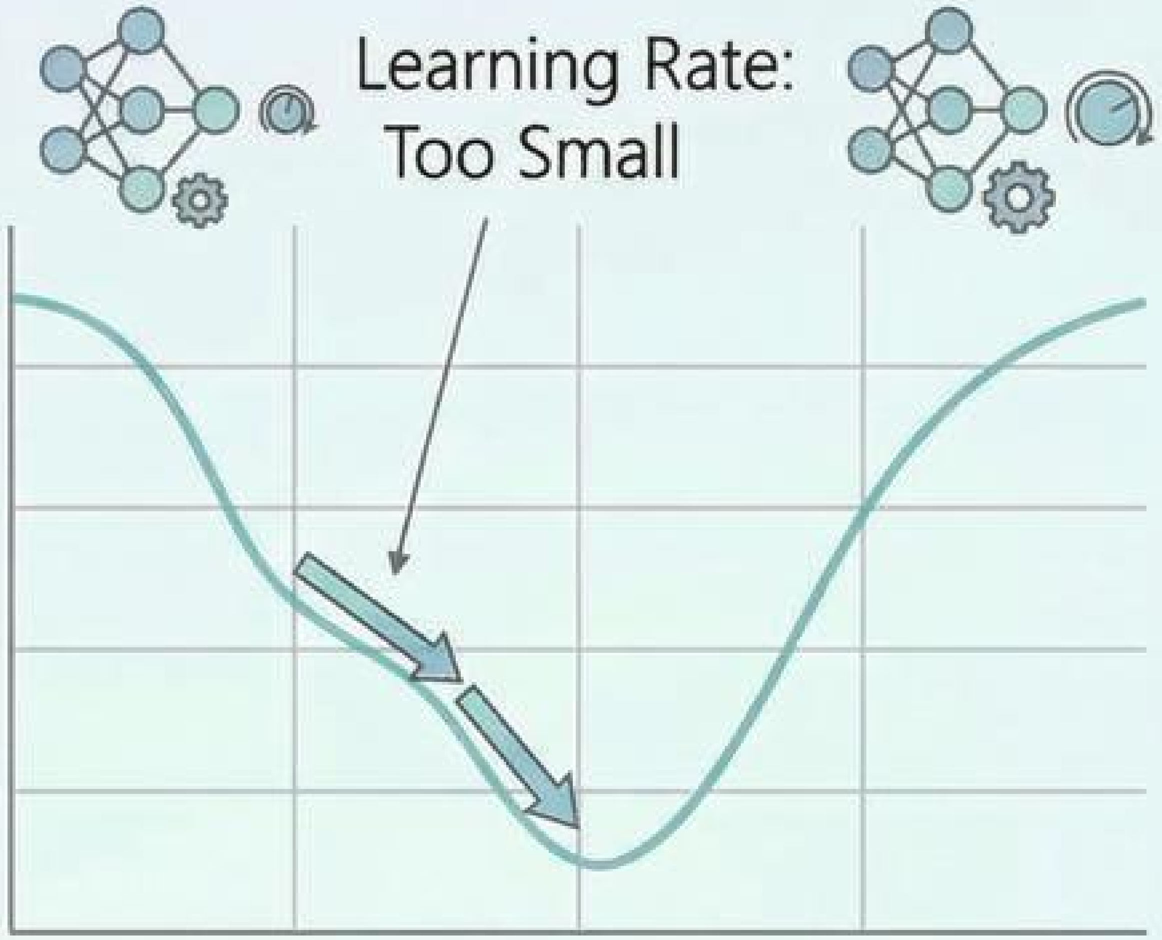
R^2 (R-squared)

Proportion of variance in the dependent variable that is predictable from the independent variables

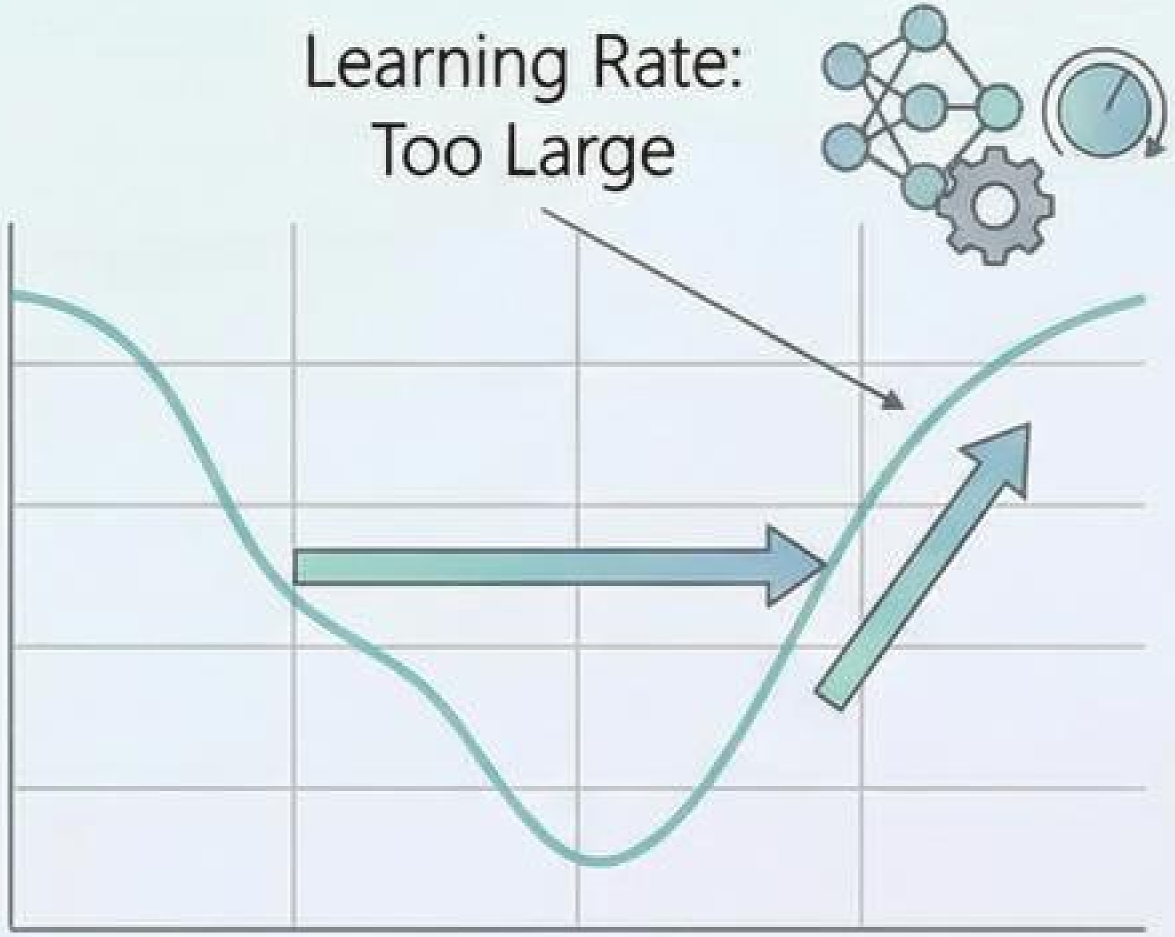
Clarifies that each measures prediction error differently, guiding model selection for continuous targets.

Hyperparameter Tuning – Concept

Settings chosen before training that control model behavior.

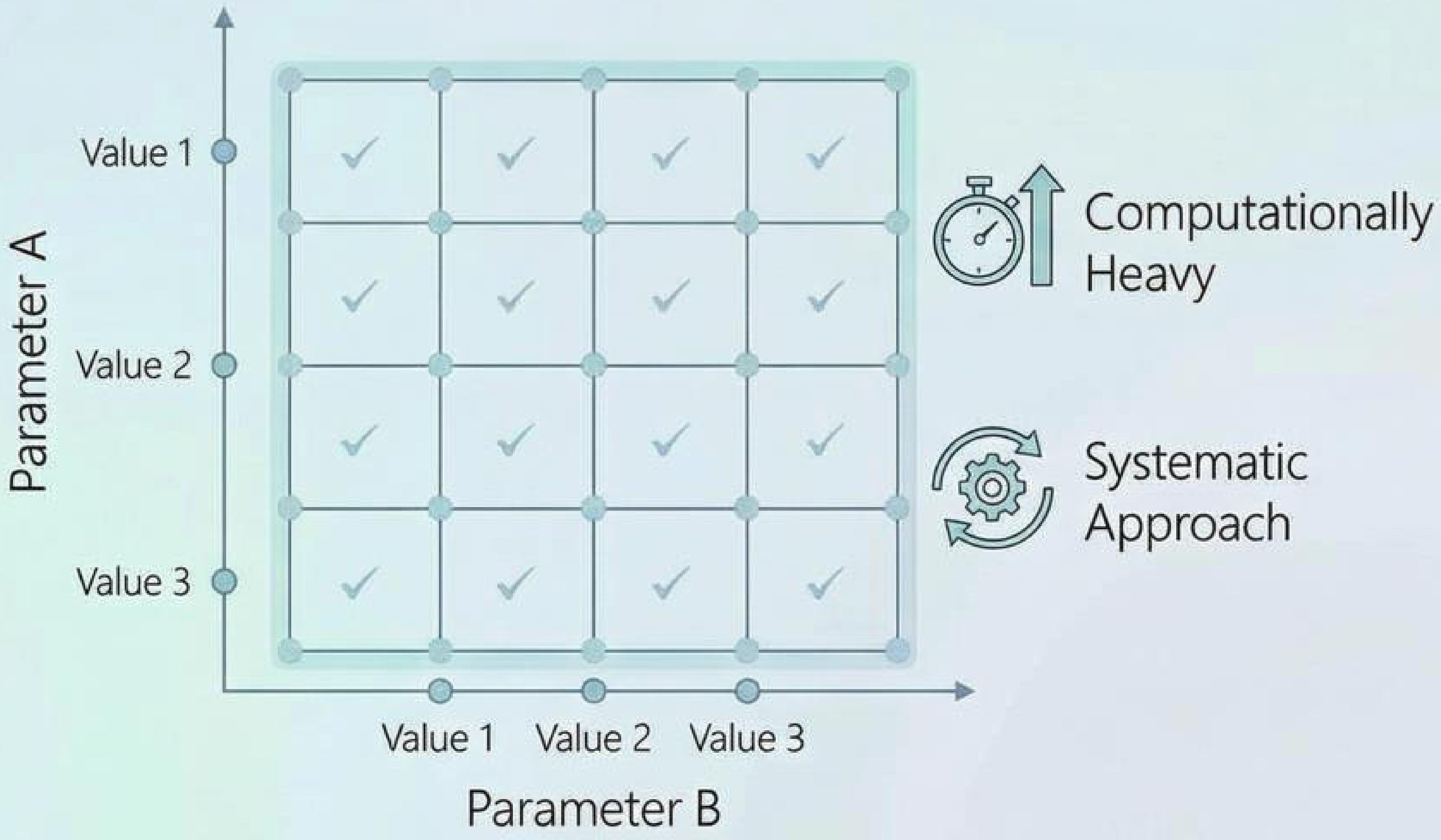


A simple example contrasts learning rates to illustrate impact on performance.



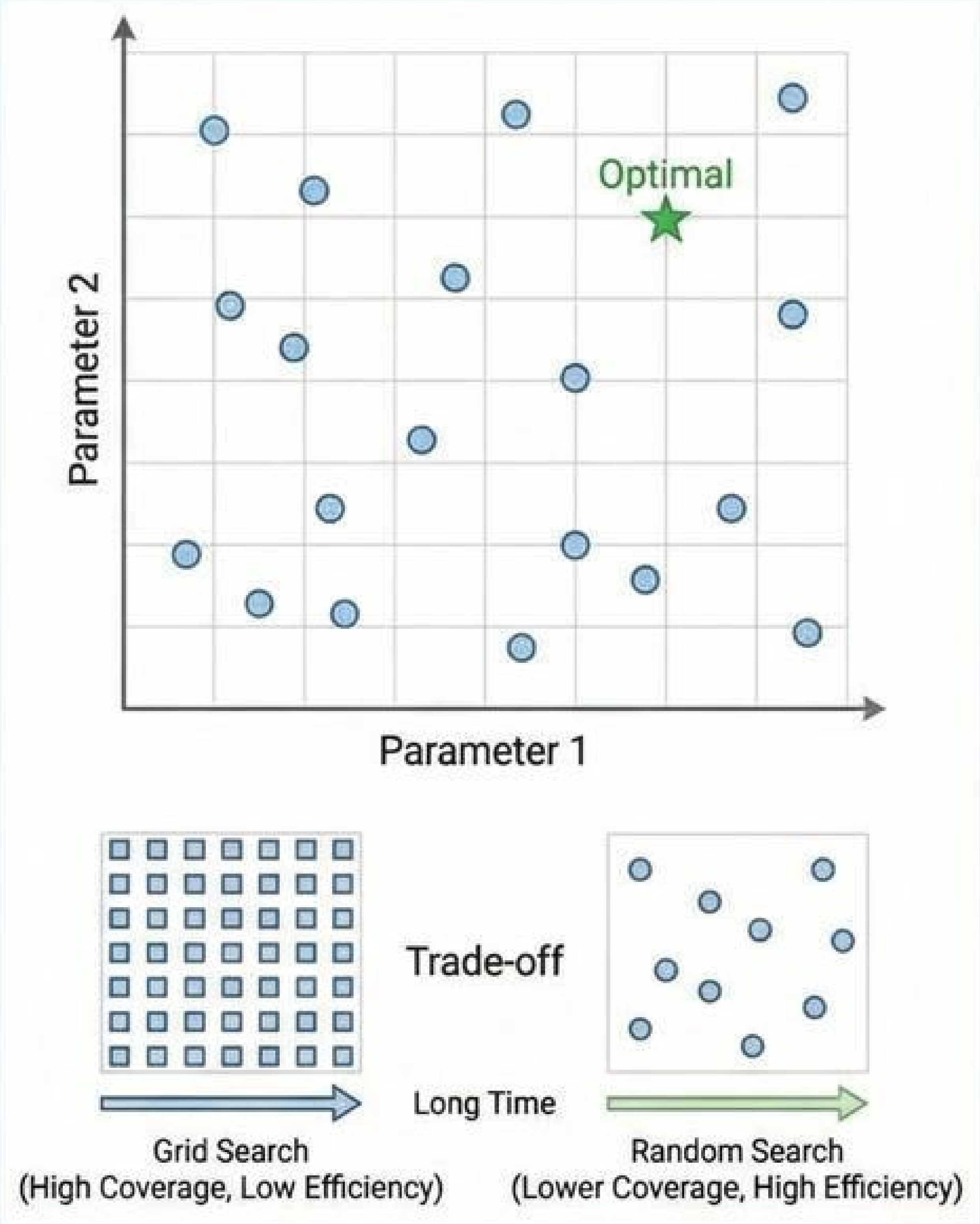
Grid Search

Exhaustive evaluation of all parameter combinations. Systematic but computationally heavy.

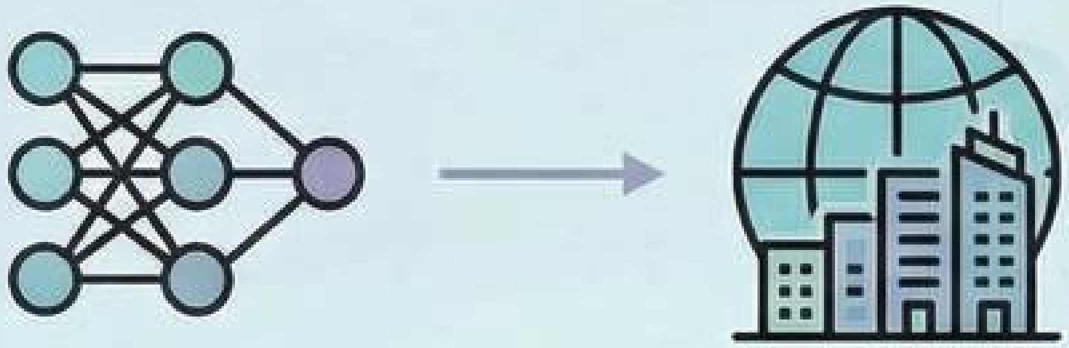


Random Search

- Samples parameter combinations randomly from defined distributions.
- Often faster and more efficient than grid search for high-dimensional spaces.



Model Deployment: Overview



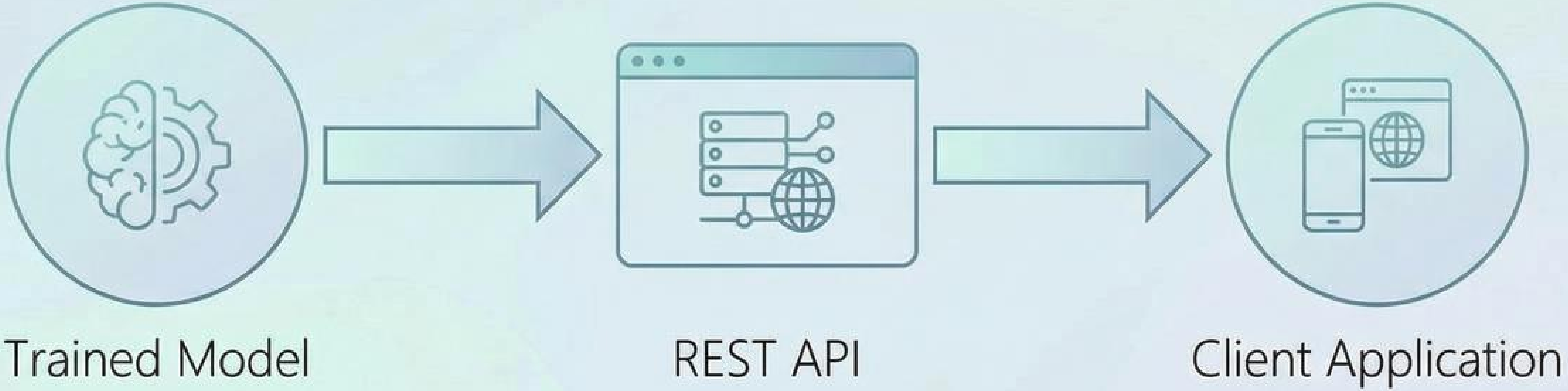
Definition

Making a trained model available for real-world predictions.

Purpose

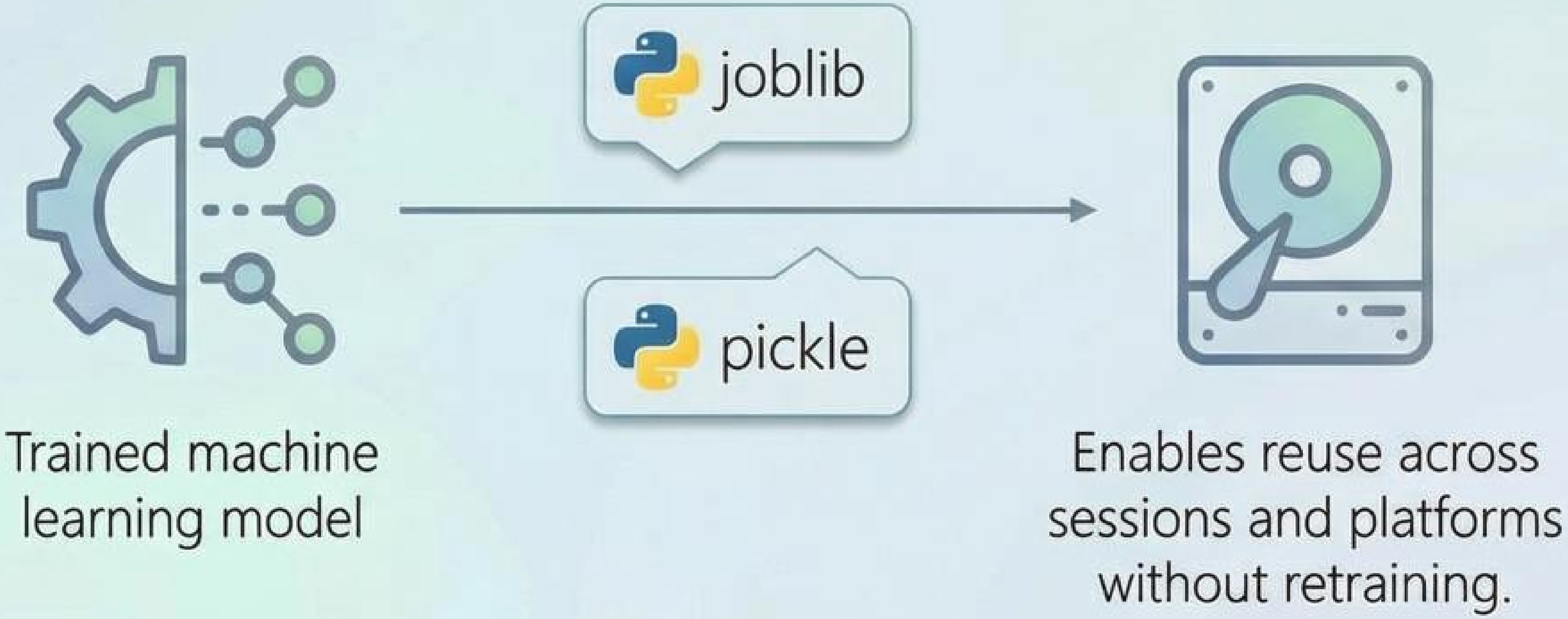
Bridges the gap between data science and tangible business value.

Deployment Architecture



Clarifies the flow from model artifact to end-user prediction without technical clutter.

Saving & Loading Models



Recap & Summary

- What Machine Learning is and how it works
- Explored data through EDA (Exploratory Data Analysis)
- Built supervised models (Linear & Logistic Regression)
- Understood unsupervised learning (K-Means & PCA)
- Evaluated models with key metrics
- Tuned models using Grid Search & Random Search
- Learned Decision Trees and Random Forests

Thank you for attending this workshop.

Feel free toreachout viathecontactdetails providedandand scan theQR code for additional notes,code,andresourcestocontinueexploringensemblemethods on your own.

Instagram: ggc_hits
LinkedIn: Google Genesis Club

Scan for Resources & Code