KENT STATE UNIVERSITY

Machine Learning (ML) Methods in Construction

Ruoxin Xiong, Ph.D.

Assistant Professor of Construction Management

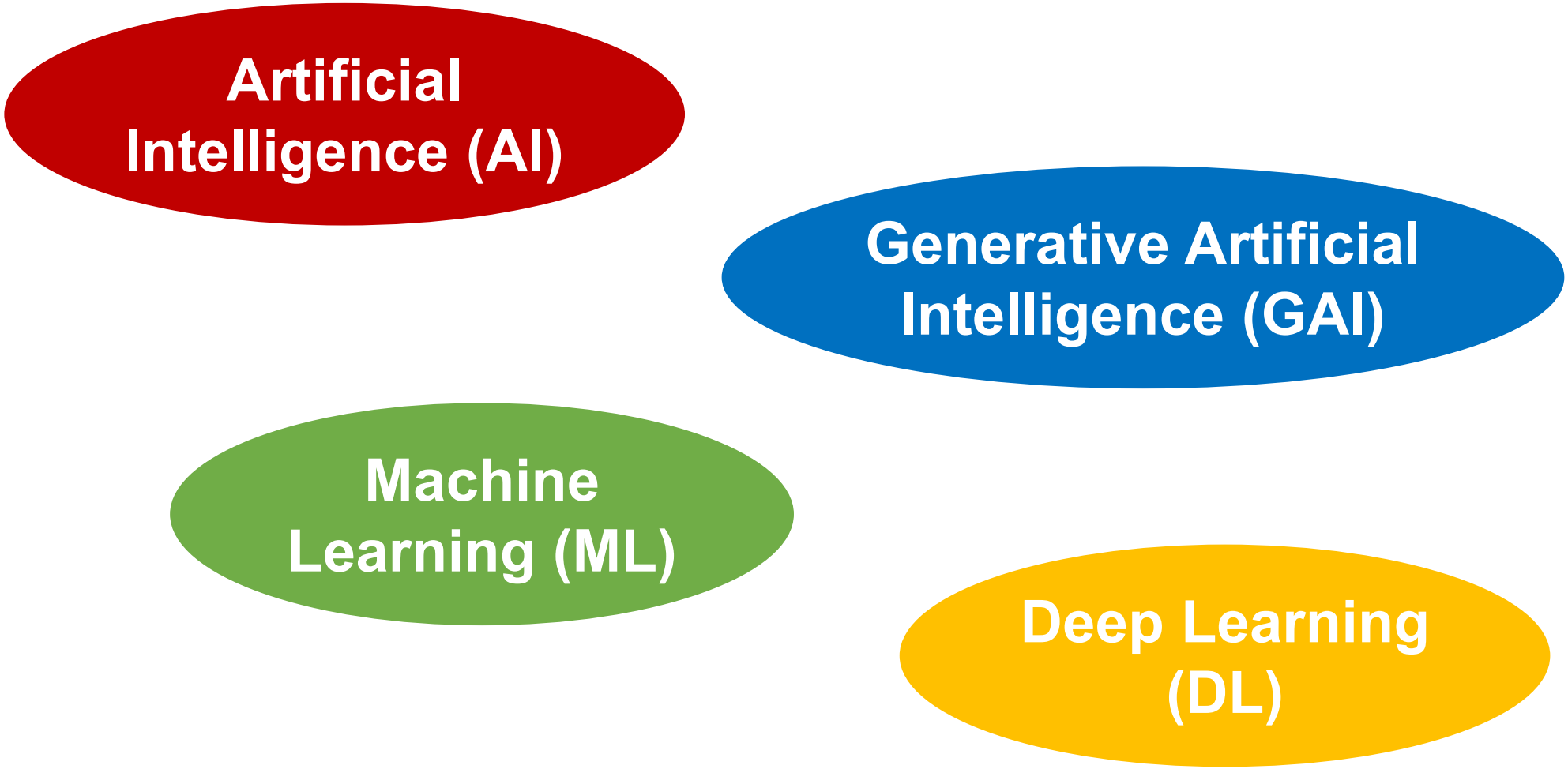
rxiong3@kent.edu

Spring 2026

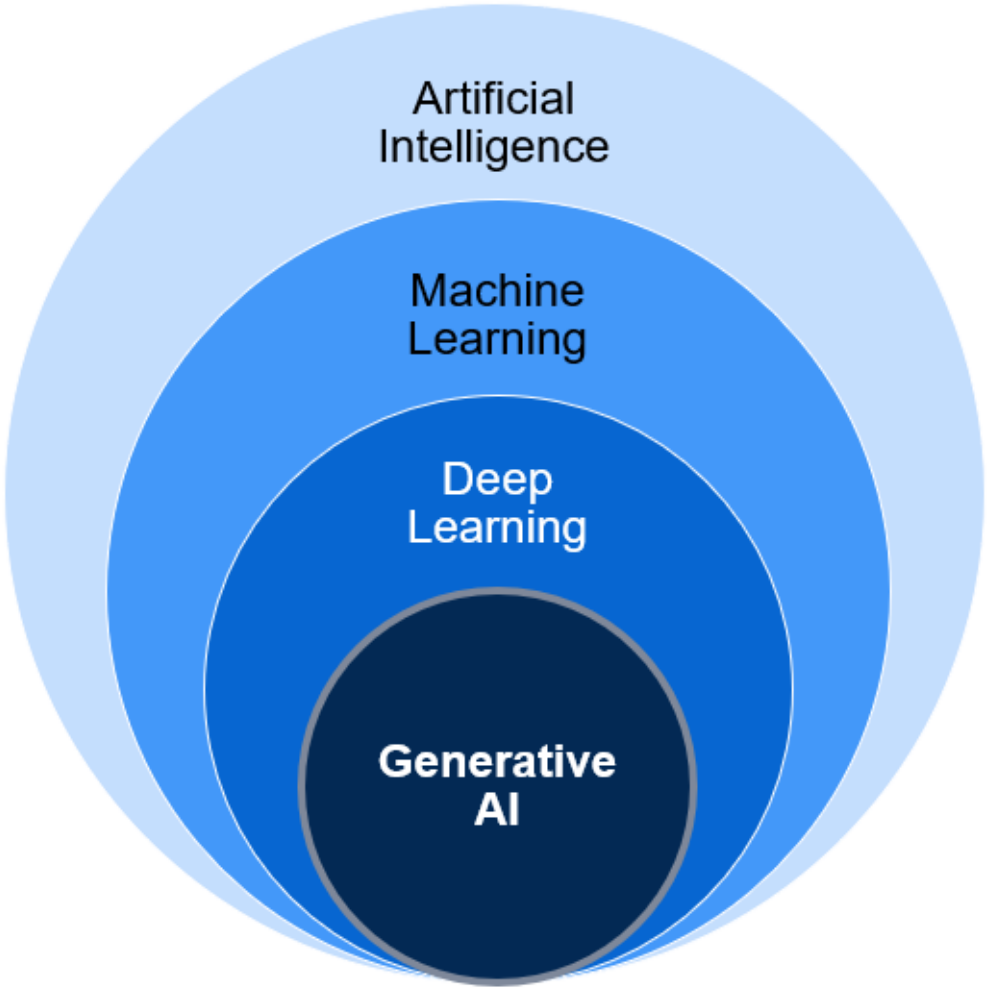
Learning Objectives

- Understand why ML is important in the construction domain.
- Explore popular ML methods and their applications in construction.
- Identify real-world use cases and challenges.
- Learn to approach an ML problem with a construction perspective.

What is AI?

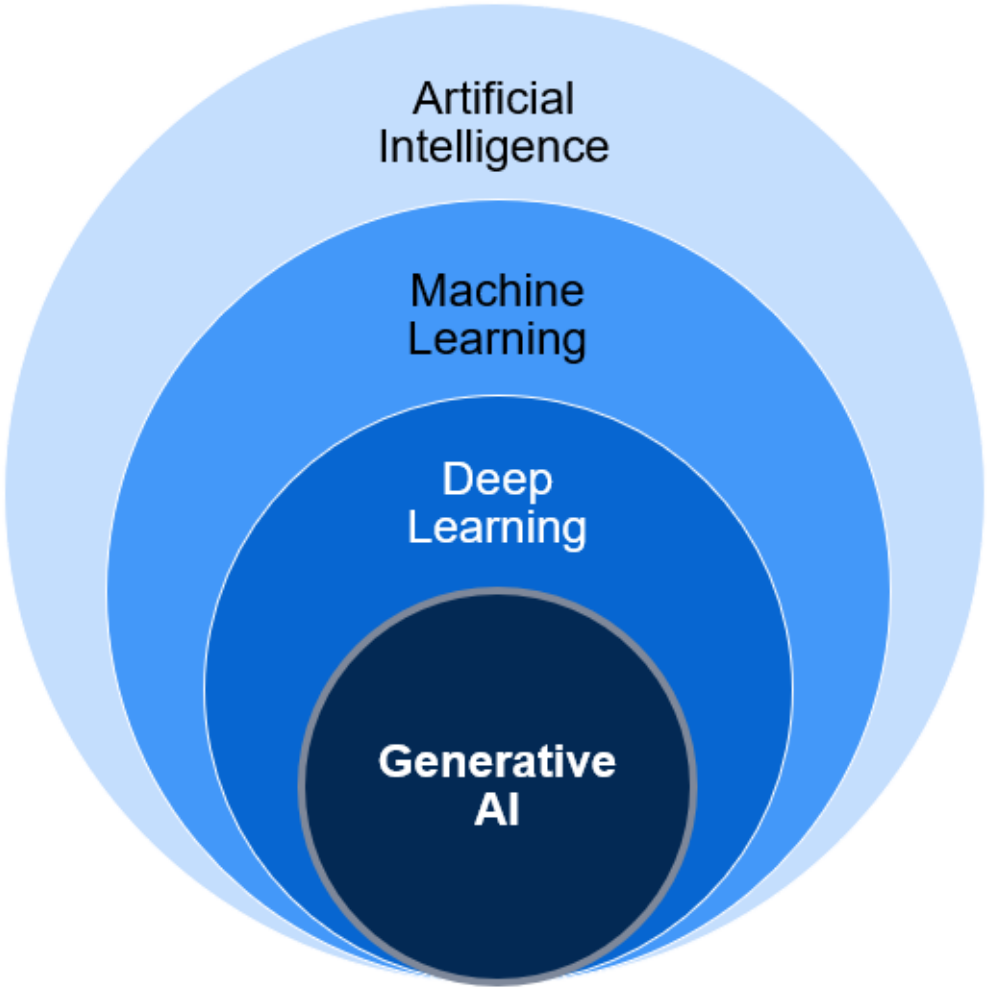

Artificial Intelligence (AI)
Generative Artificial Intelligence (GAI)
Machine Learning (ML)
Deep Learning (DL)

The AI Landscape



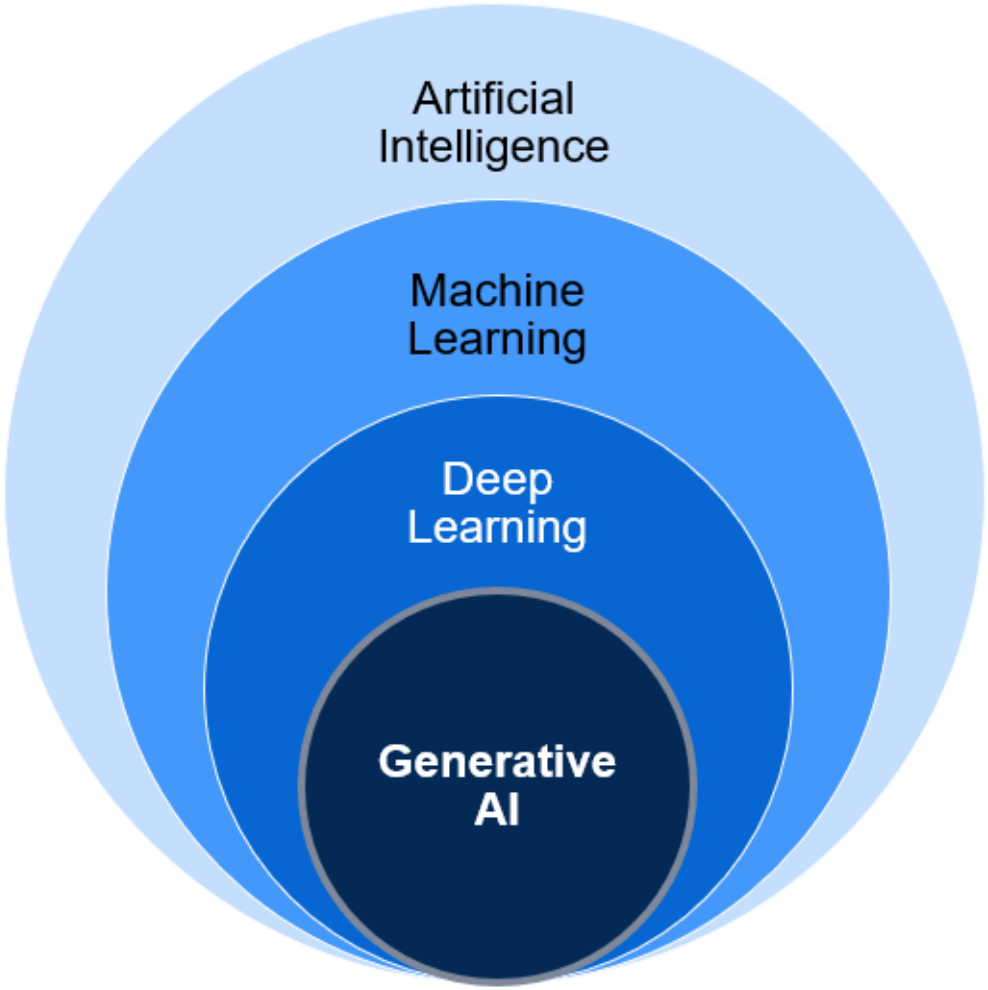
- AI is the broad field of computer science focused on creating **machines capable of performing tasks that typically require human intelligence**

The AI Landscape



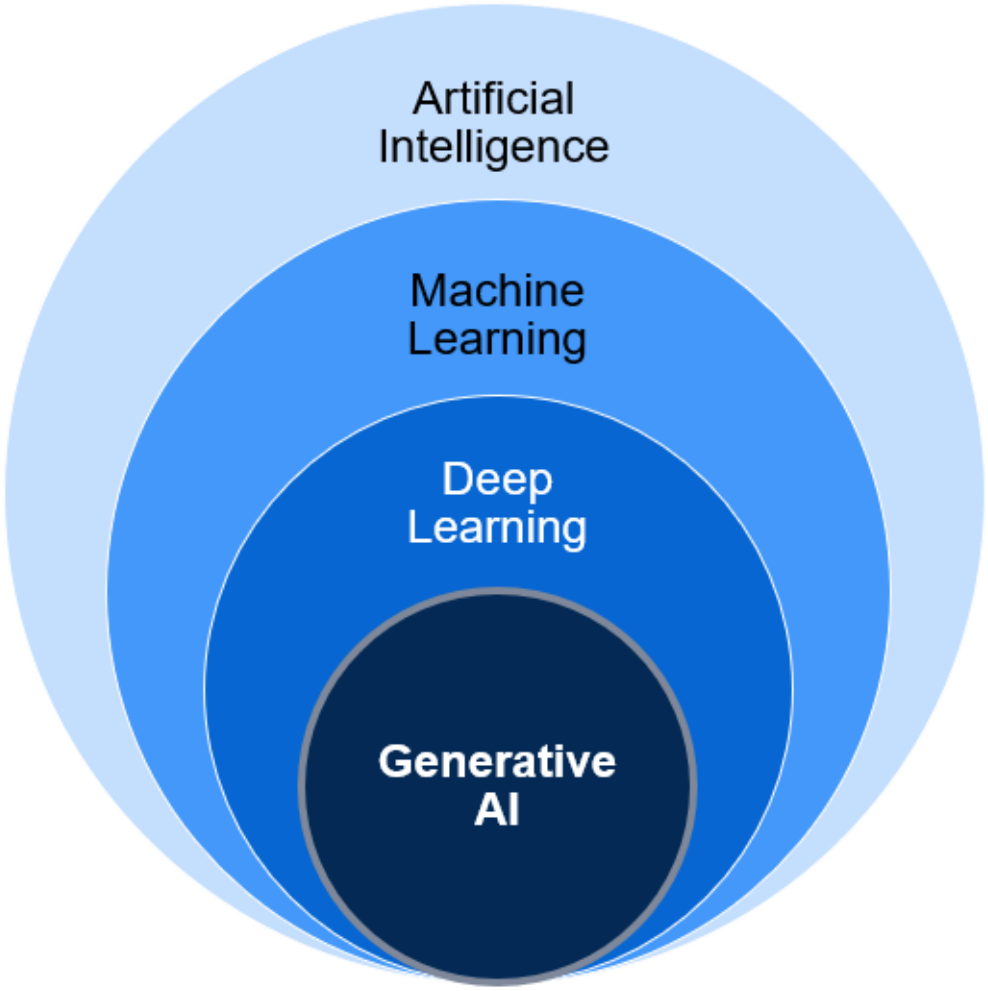
- ML is a subset of AI involving algorithms and statistical models that enable computers to **improve their performance on a task through examples**

The AI Landscape



- DL is a subset of ML based on **artificial neural networks**, where algorithms learn from **large amounts of data** to identify patterns and make decisions

The AI Landscape



- GAI refers to AI technologies that can **generate new content, ideas, or data** that are coherent and plausible, often resembling human-generated outputs

Advantages of AI/ML

Machine learning gives computers the ability to learn by example without explicit programming

Advantages of AI/ML

Machine learning gives computers the ability to learn by example without explicit programming

```
if email contains “Nigerian
prince”:
    then mark as spam;
if email contains “limited
offer”:
    then mark as spam;
if email contains “internship
offer”:
    then do not mark as spam;
...
```

Traditional Programming

```
while accuracy low:
    classify some emails;
    check errors;
    change to reduce errors;
```

Machine Learning Programs

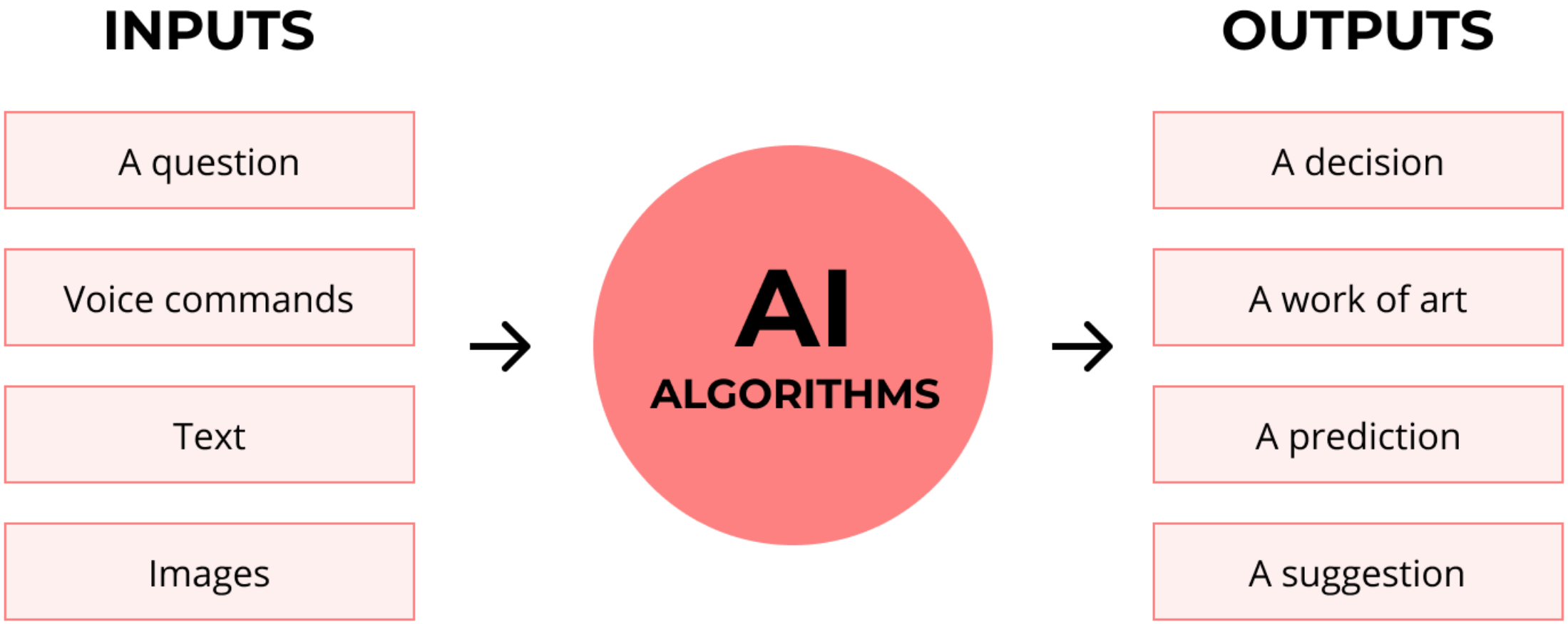
Without Explicit Programming

Machine learning **takes some data to train the system**

It then learns patterns from the data so that it can **predict for future data**

The best part is that this “code” (not patterns) can be **re-used for many different problems!**

Examples of AI Capabilities



Why ML Matters in Construction

- Construction industry challenges: cost overruns, schedule delays, safety risks
- ML can enhance **forecasting, optimization, and risk detection**
- **Emergence of big data** from IoT devices, BIM, site sensors, and project management systems

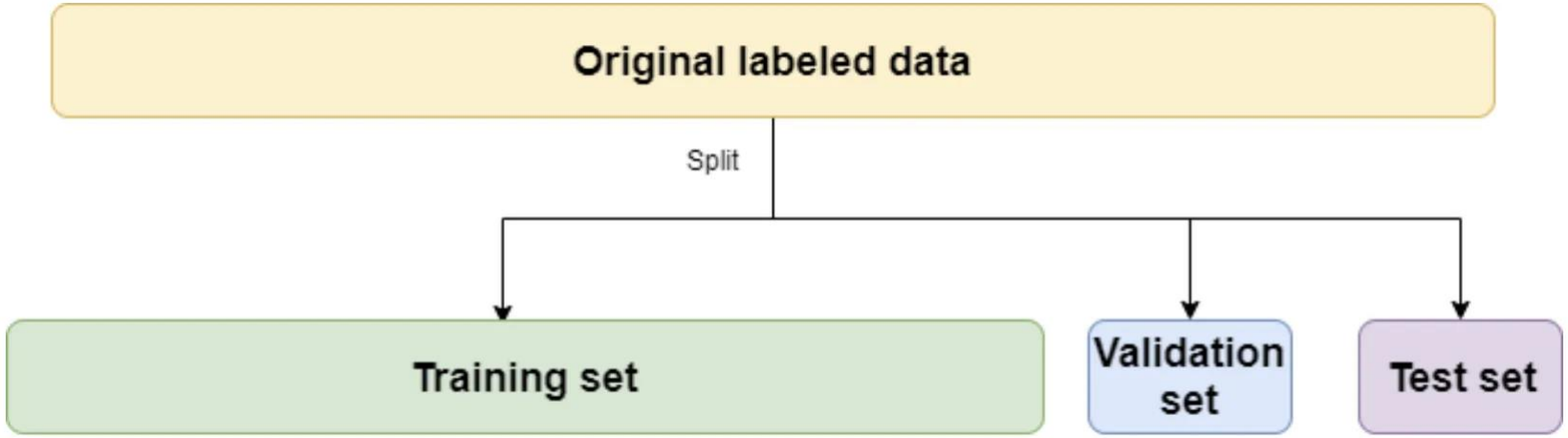
The ML Pipeline

The ML Pipeline in Construction

- **Data Collection** – from sensors, historical records, BIM.
- **Data Cleaning & Preprocessing** – handle missing values, outliers, ensure consistent naming.
- **Feature Engineering** – domain expertise to create useful features (e.g., weather delays, labor hours).
- **Model Selection** – choose algorithms (linear models, tree-based, neural networks).
- **Training & Validation** – split data, tune hyperparameters.
- **Testing & Evaluation** – measure performance on a held-out set.
- **Deployment & Monitoring** – real-time dashboards, iterative improvements.

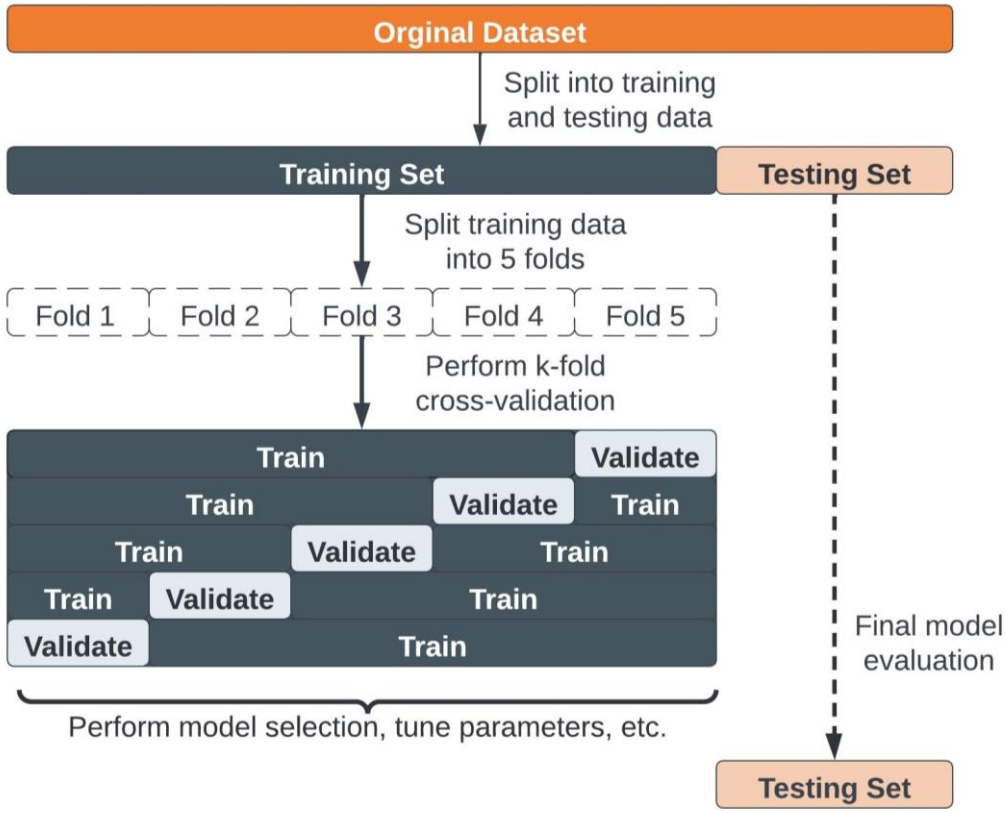
Data Splitting

- **Common Splitting Strategies:**
  - Train/Validation/Test split (e.g., 70%-15%-15%)



Data Splitting

- **Common Splitting Strategies:**
  - Train/Validation/Test split (e.g., 70%-15%-15%)
  - Cross-validation (k-fold for smaller datasets)

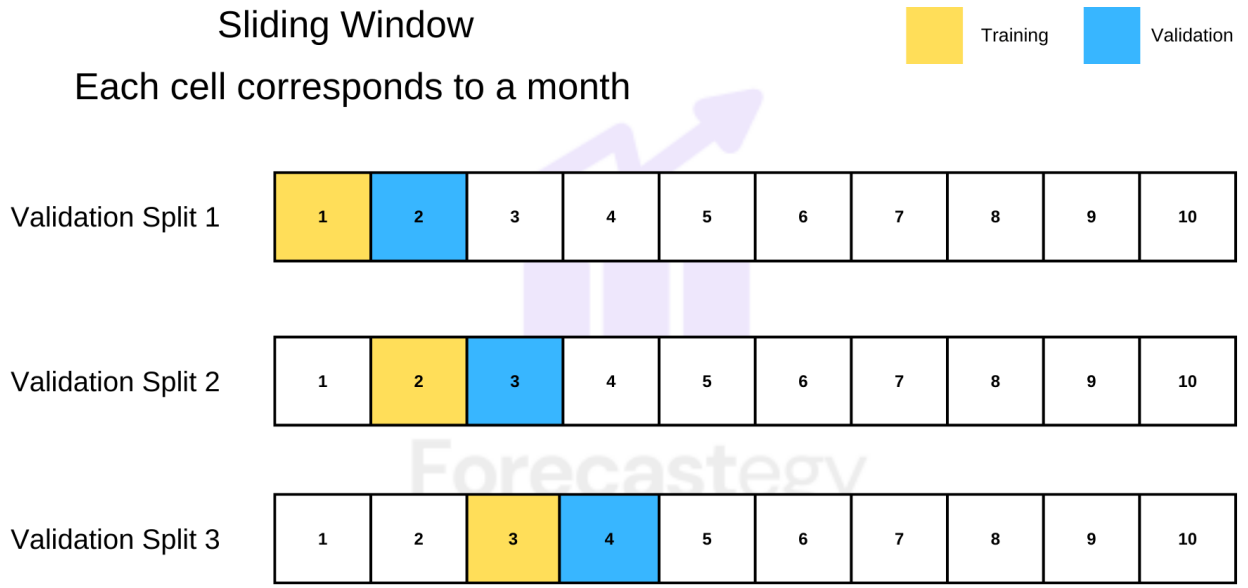


Data Splitting

- **Time-Series Considerations:**
  - Can we use the common strategy: Random Train/Validation/Test split (e.g., 70%-15%-15%)?

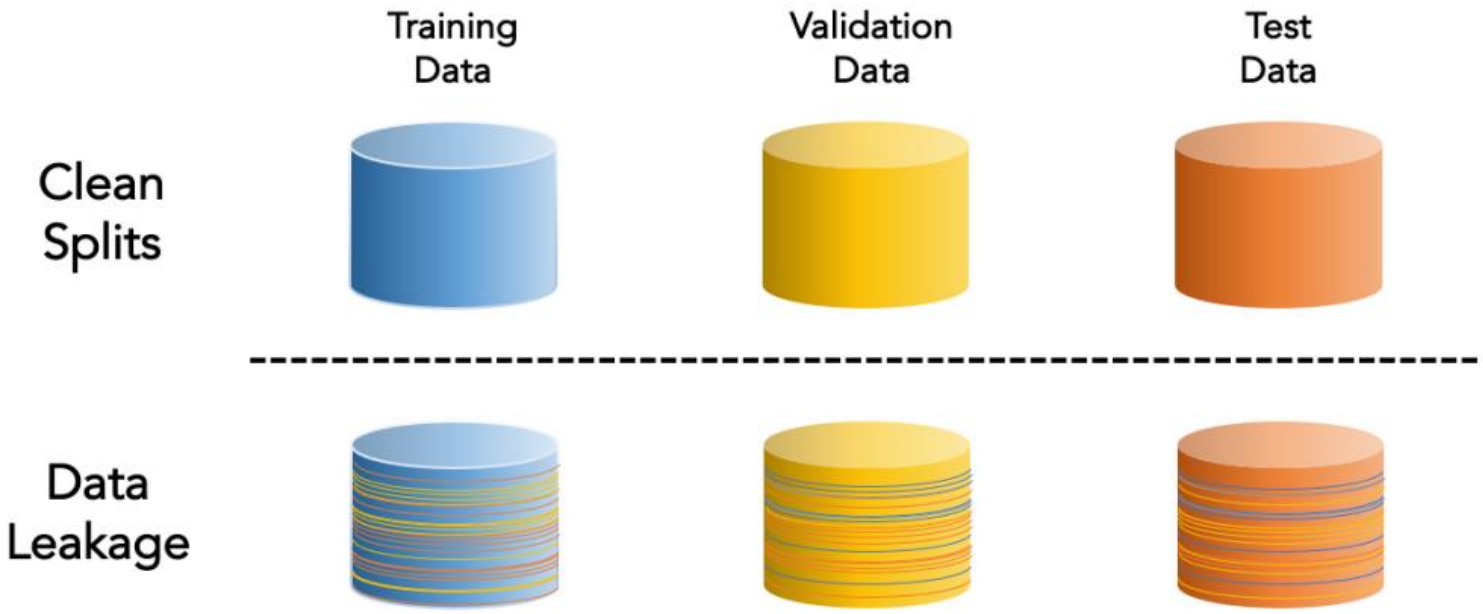
Data Splitting

- **Time-Series Considerations (e.g., schedule/cost forecasting, time-sensitive):**
  - Can we use the common strategy: Train/Validation/Test split (e.g., 70%-15%-15%)?

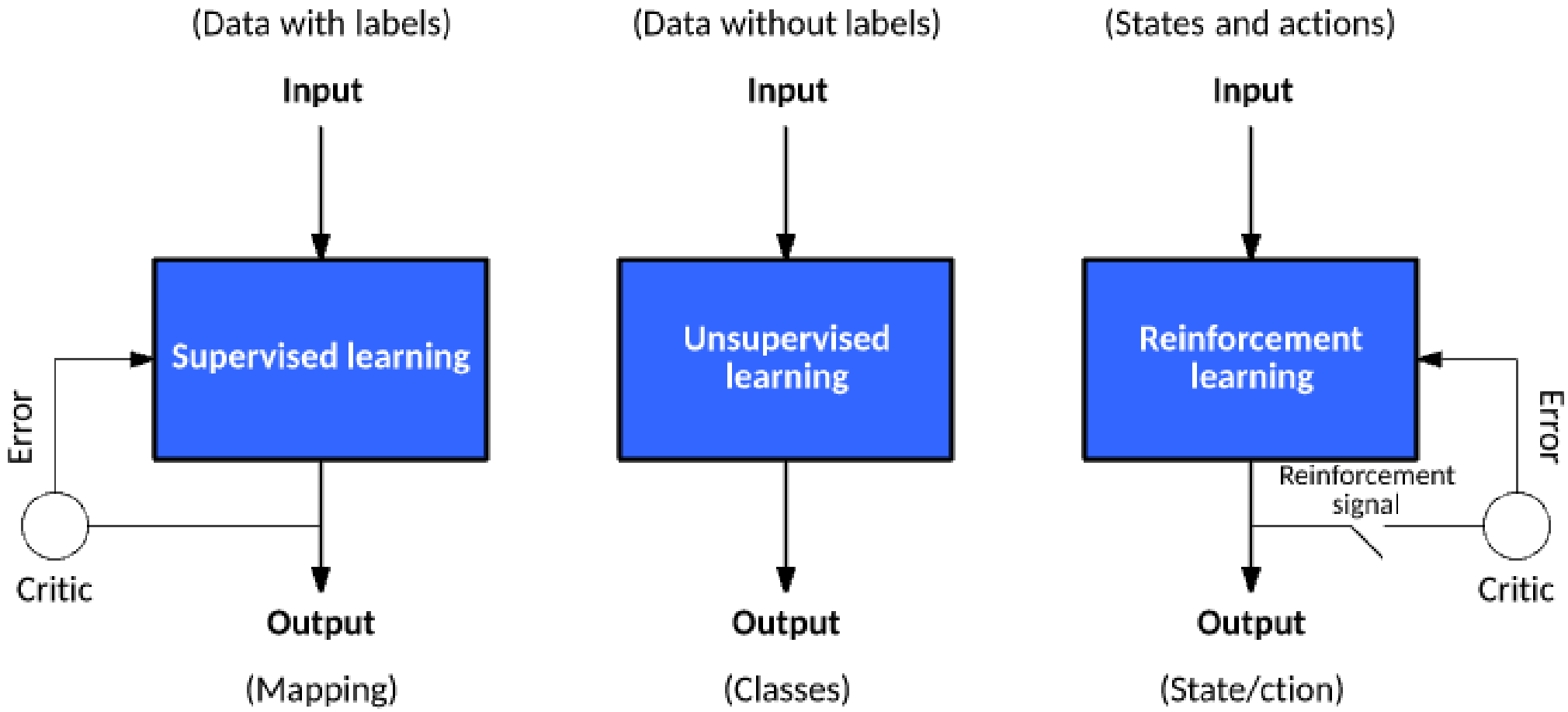


Data Leakage

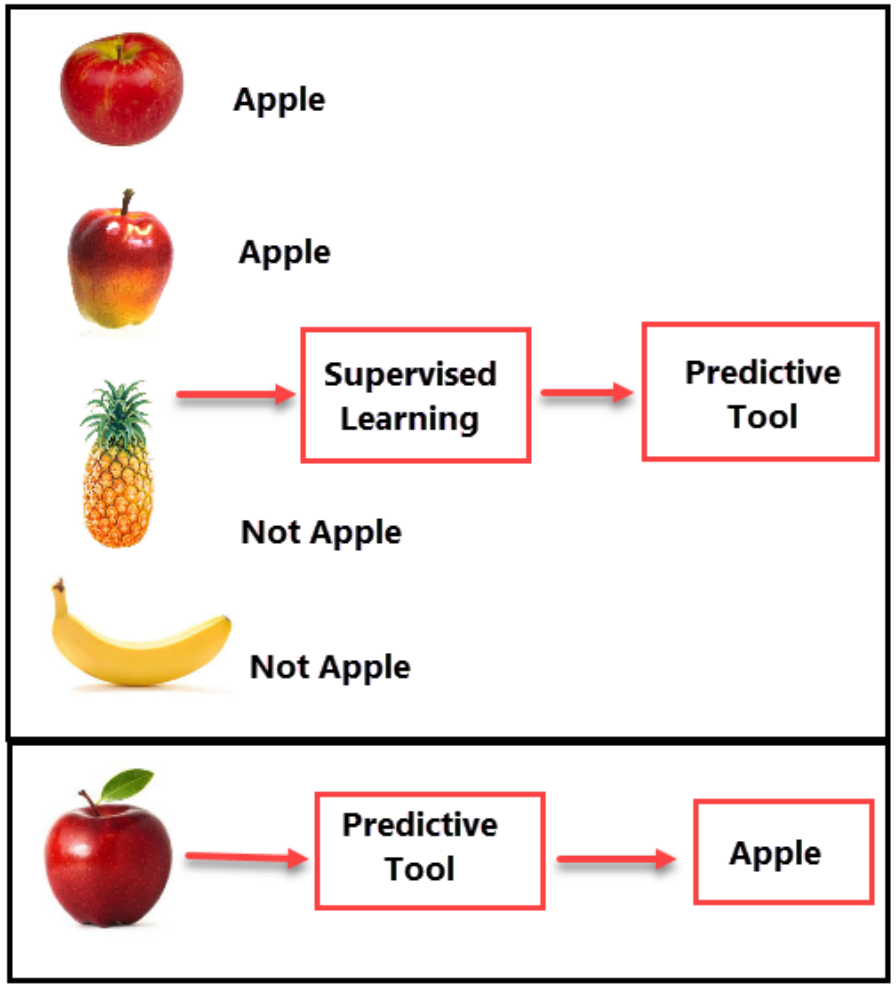
- Occurs if information from outside the training window or future data seeps into the model.
- Example: Using “final cost” as a feature when predicting cost overruns.



ML Foundations Refresher

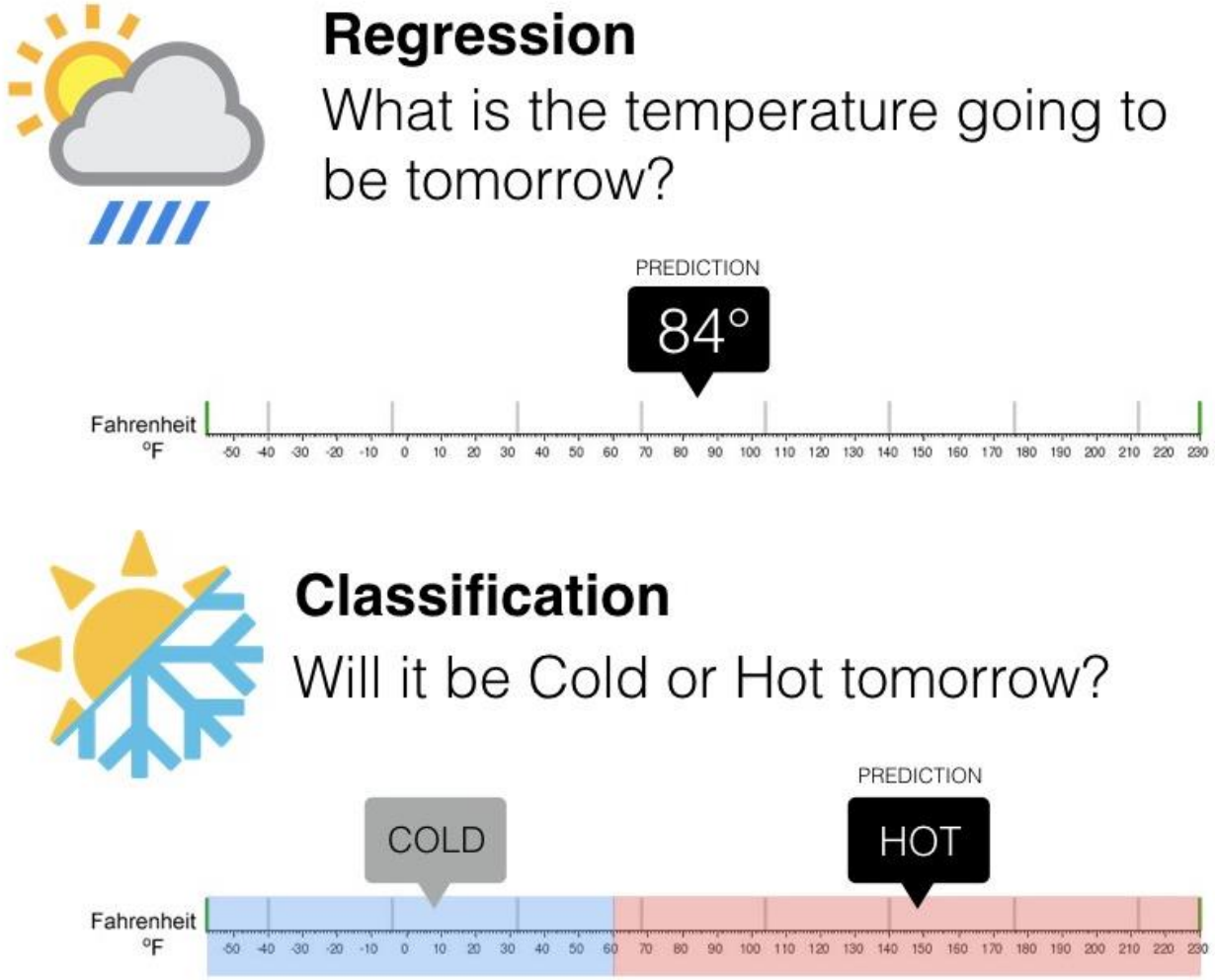


Supervised Learning



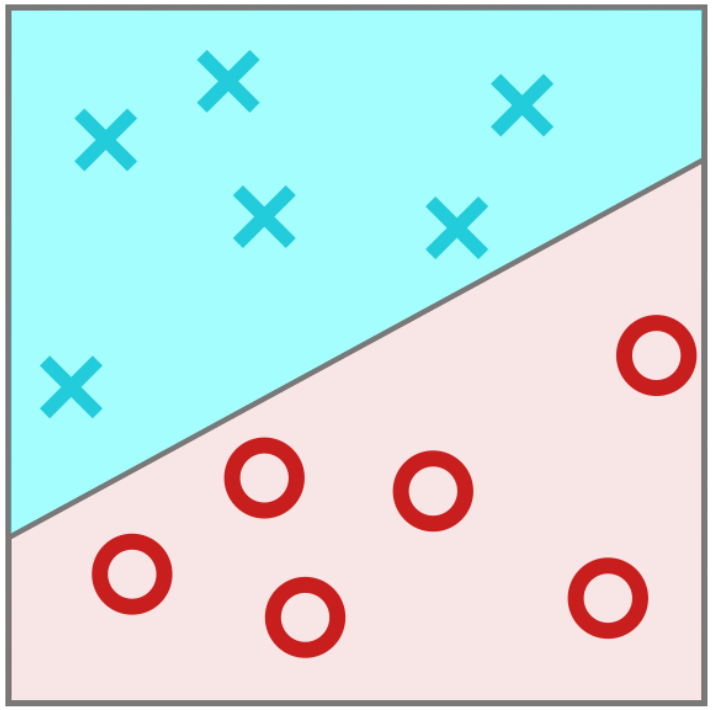
Could you summarize what is the supervised learning?

Supervised Learning: Classification vs. Regression



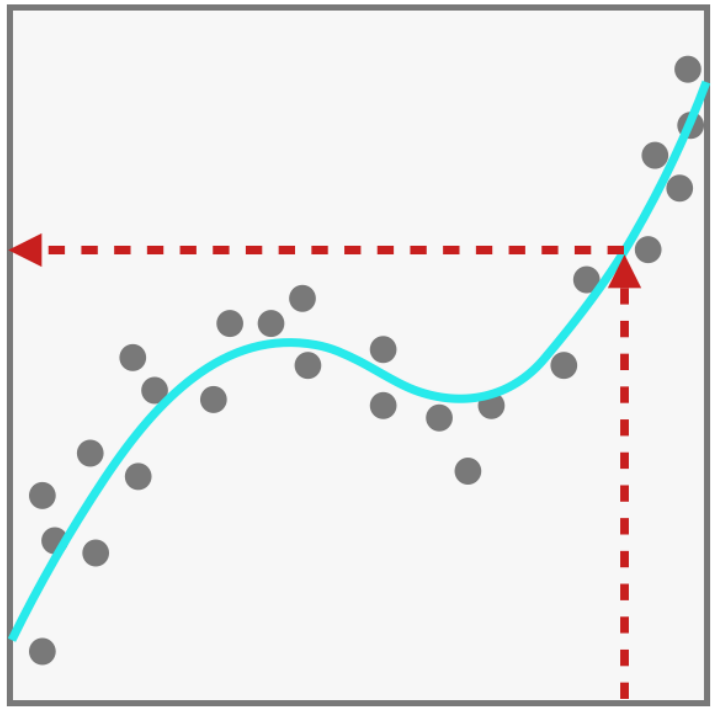
Supervised Learning: Classification vs. Regression

Classification Groups observations into "classes"



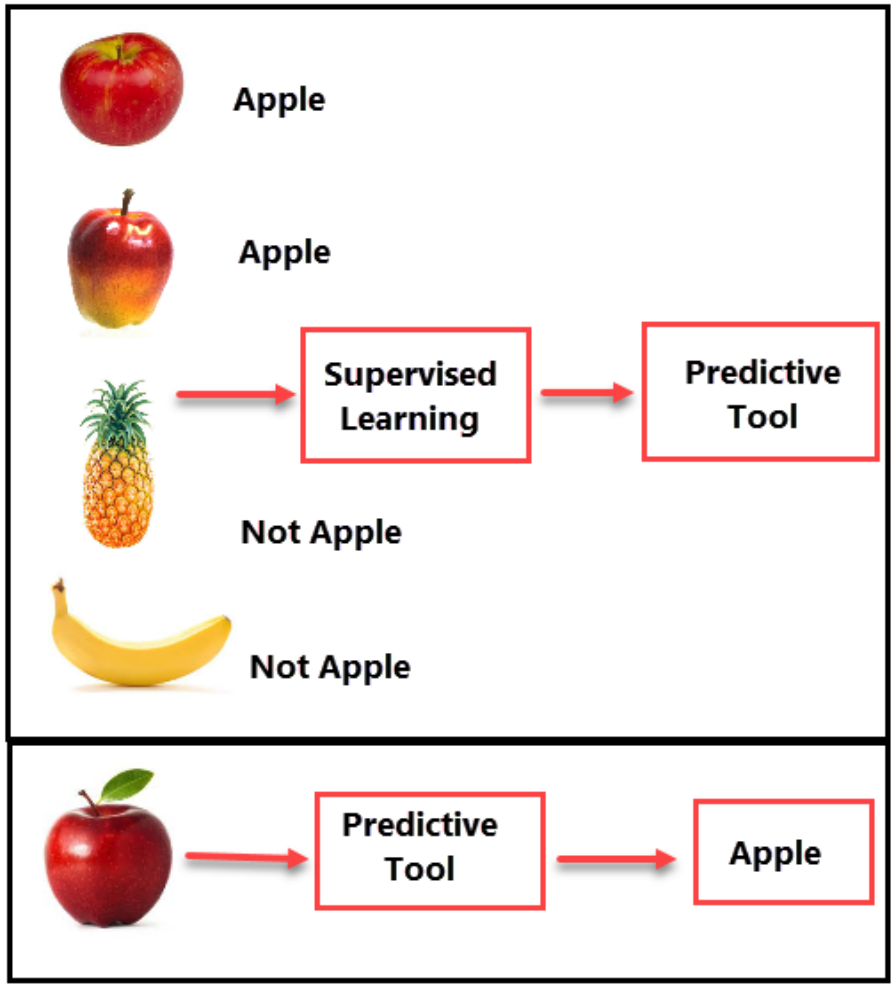
Here, the line classifies the observations into X's and O's

Regression predicts a numeric value



Here, the fitted line provides a predicted output, if we give it an input

Supervised Learning



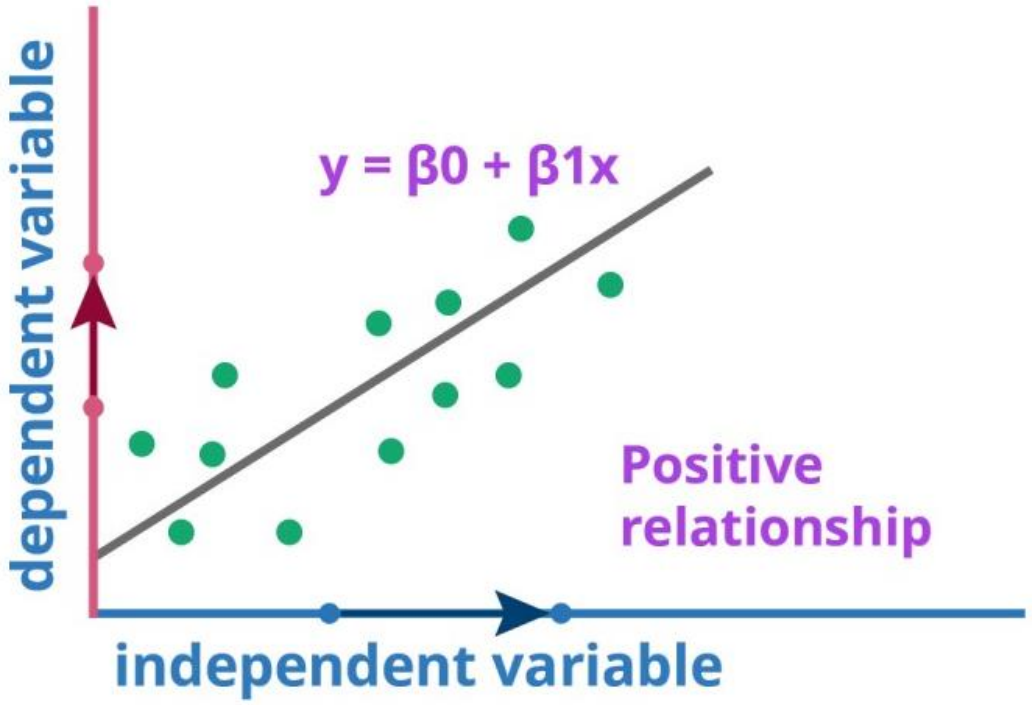
Classification or Regression?

Supervised Learning

- **Linear & Logistic Regression**
  - Predict cost (regression) or binary classification (e.g., on-schedule vs. off-schedule).
- **Decision Trees & Random Forests**
  - Interpretable rules for classification (e.g., “High risk site” vs. “Low risk site”).
- **Gradient Boosting (XGBoost, LightGBM)**
  - High performance in many construction-related datasets with structured data.
- **Neural Networks**
  - Time-series forecasting (progress, site conditions), complex relationships.

Supervised Learning: Linear Regression

- Model the relationship between one or more independent variables (**features**) and a continuous dependent variable (**target**)



Supervised Learning: Linear Regression

- Predict **total project cost (y)** from features like floor area ($x1$), structural complexity ($x2$), location cost index ($x3$), and project duration ($x4$)

- Cost=β0 +β1(Floor Area-$x1$)+β2 (Complexity-$x2$)+β3 (Location Index-$x3$)+β4 (Duration-$x4$) + ϵ

Dataset:

1. X Project A: size, location, types, areas, stories, story height, etc

Y Project A: cost

2. X Project B: size, location, types, areas, stories, story height, etc

Y Project A: cost

3. X Project C: size, location, types, areas, stories, story height, etc

Y Project A: cost

…

Supervised Learning: Linear Regression

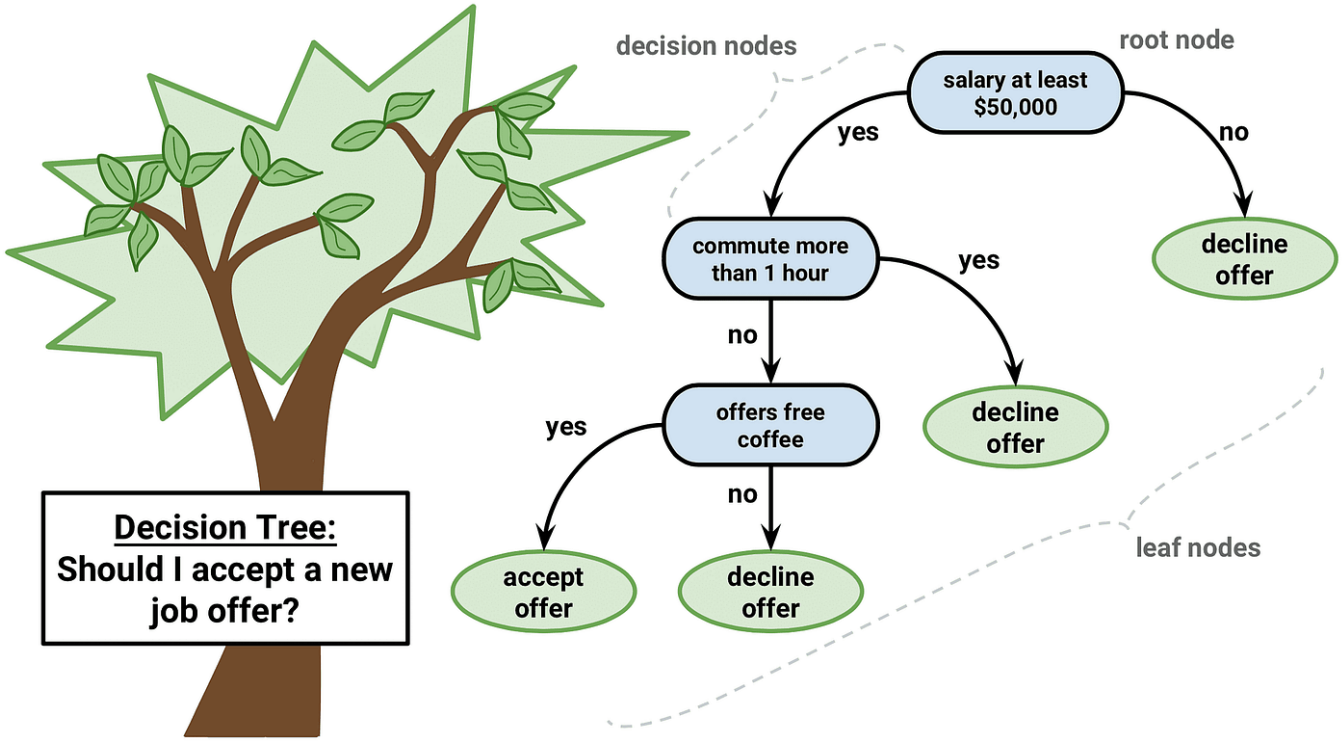
- Predict **total project cost (y)** from features like floor area ($x1$), structural complexity ($x2$), location cost index ($x3$), and project duration ($x4$)

- Cost=β0 +β1(Floor Area-$x1$)+β2 (Complexity-$x2$)+β3 (Location Index-$x3$)+β4 (Duration-$x4$) + ϵ

- Any other features missed in this model?

Supervised Learning: Decision Tree

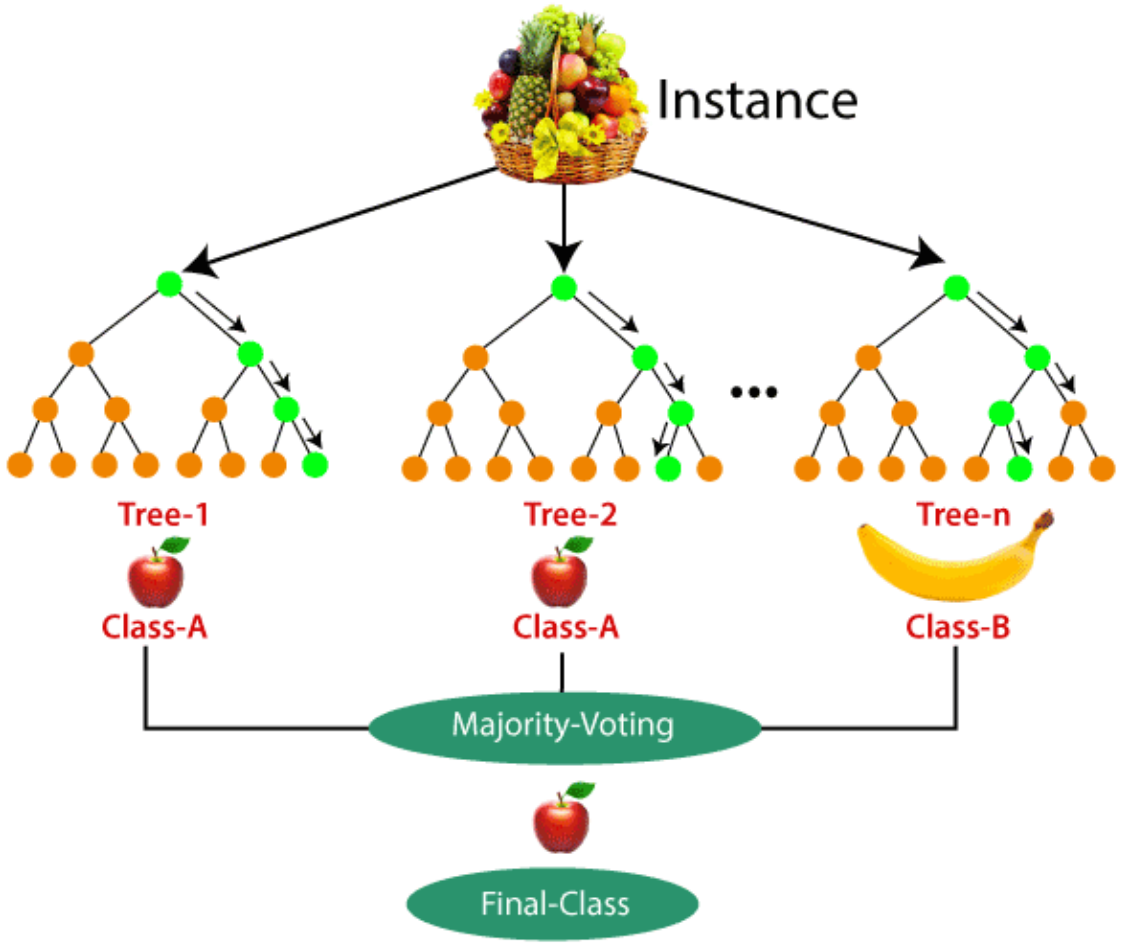
- **Tree-structured model**
- **Each node represents a feature (or attribute)** in the dataset
- **Each branch represents a decision rule**
- **Each leaf node (terminal node) represents an outcome or value** (for regression, a numeric prediction; for classification, a class label)



Supervised Learning: Random Forest

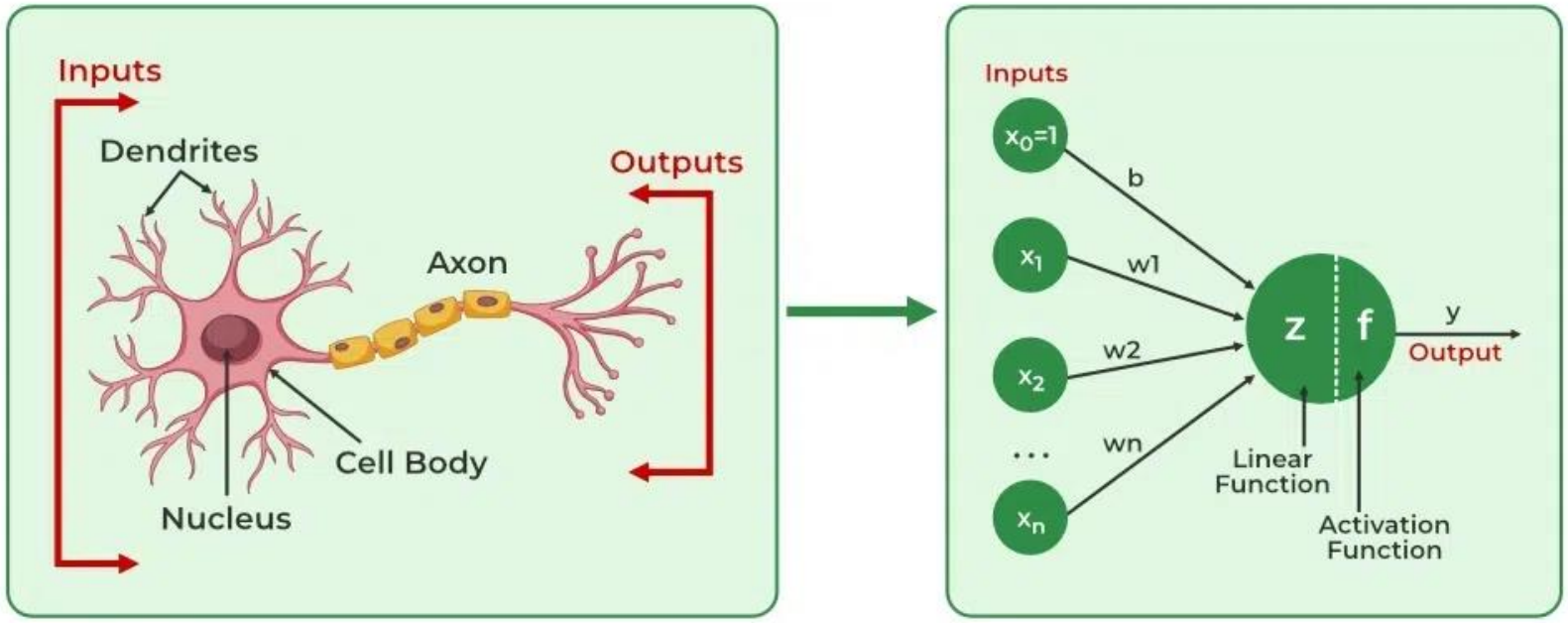
- Build **multiple decision trees** and aggregates (averages for regression, majority vote for classification) their predictions

- Different samples of the data, also randomly sampling features at each split



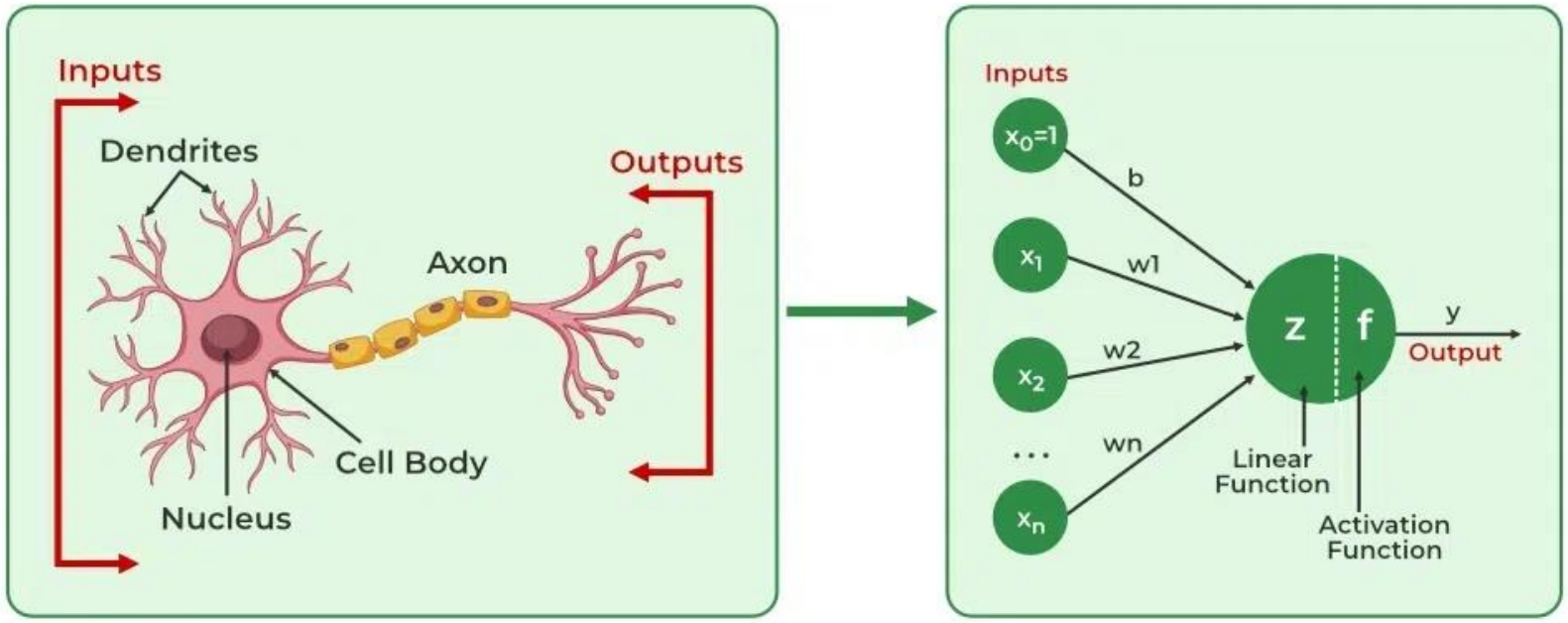
Supervised Learning: Neural Networks

- An artificial mathematical model used to approximate nonlinear functions



Supervised Learning: Neural Networks

- An artificial mathematical model used to approximate nonlinear functions



Will go deep into NN in the following lectures!

Supervised Learning Workflow:

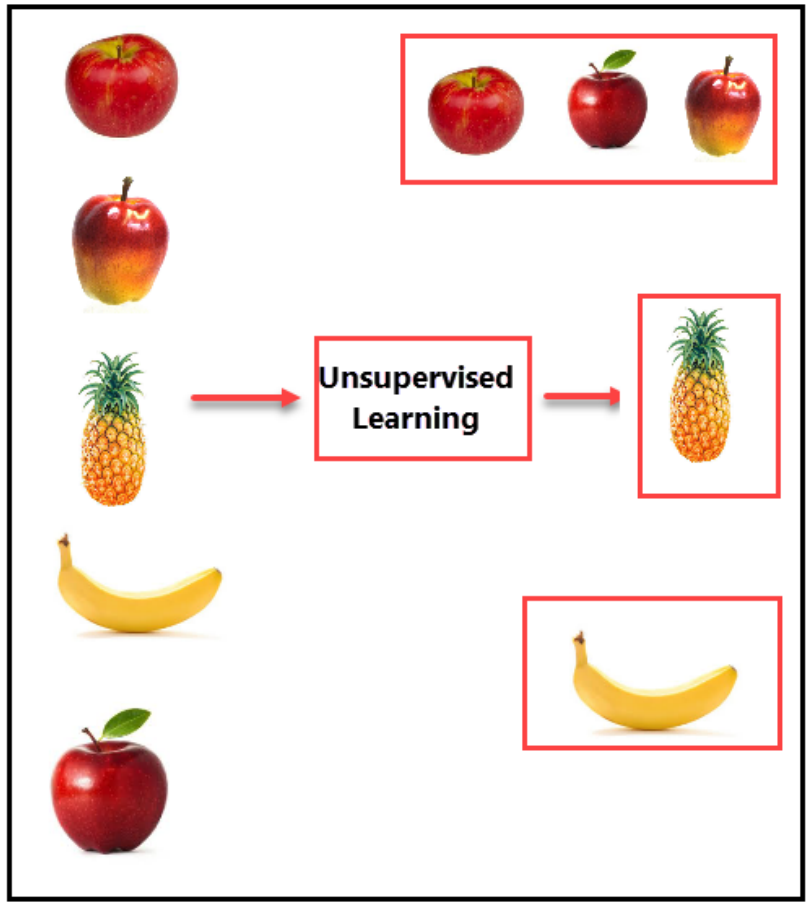
- **Initialization**: Hyperparameters, model structure
- **Forward Pass**: predict y-hat, calculate loss
- **Backpropagation** (for neural networks) or iterative improvement (for tree-based methods)
- **Hyperparameter Tuning**: grid search

Could you name some problems that maybe solved using the Supervised Learning Methods?

Could you name some problems that maybe solved using the Supervised Learning Methods?

- Progress forecasting (time-to-completion)
- Cost estimation from historical data
- Quality inspections (image classification for defect detection)

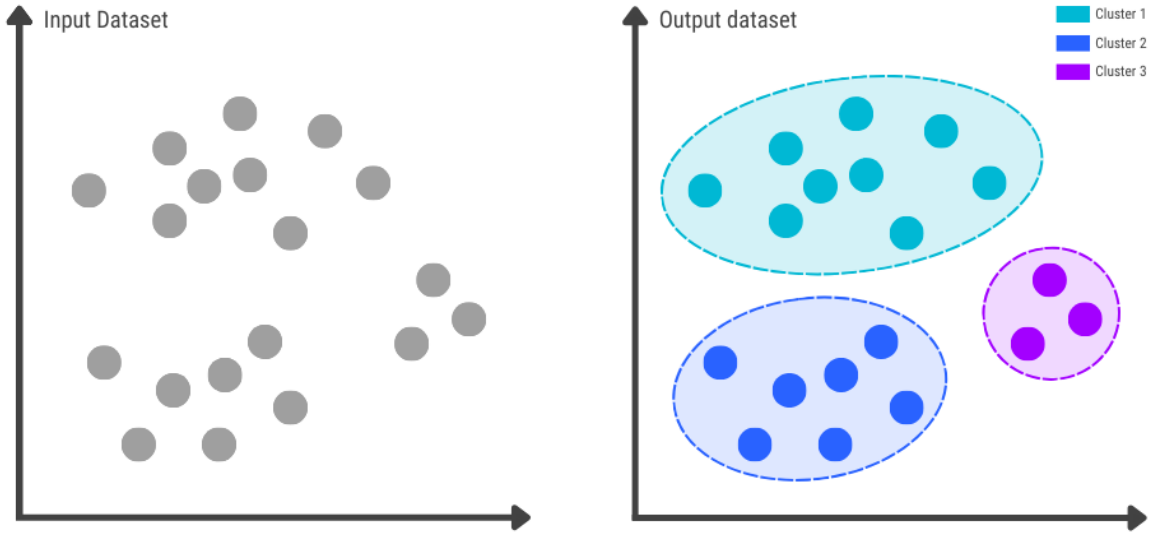
Unsupervised Learning



Could you summarize what is the unsupervised learning?

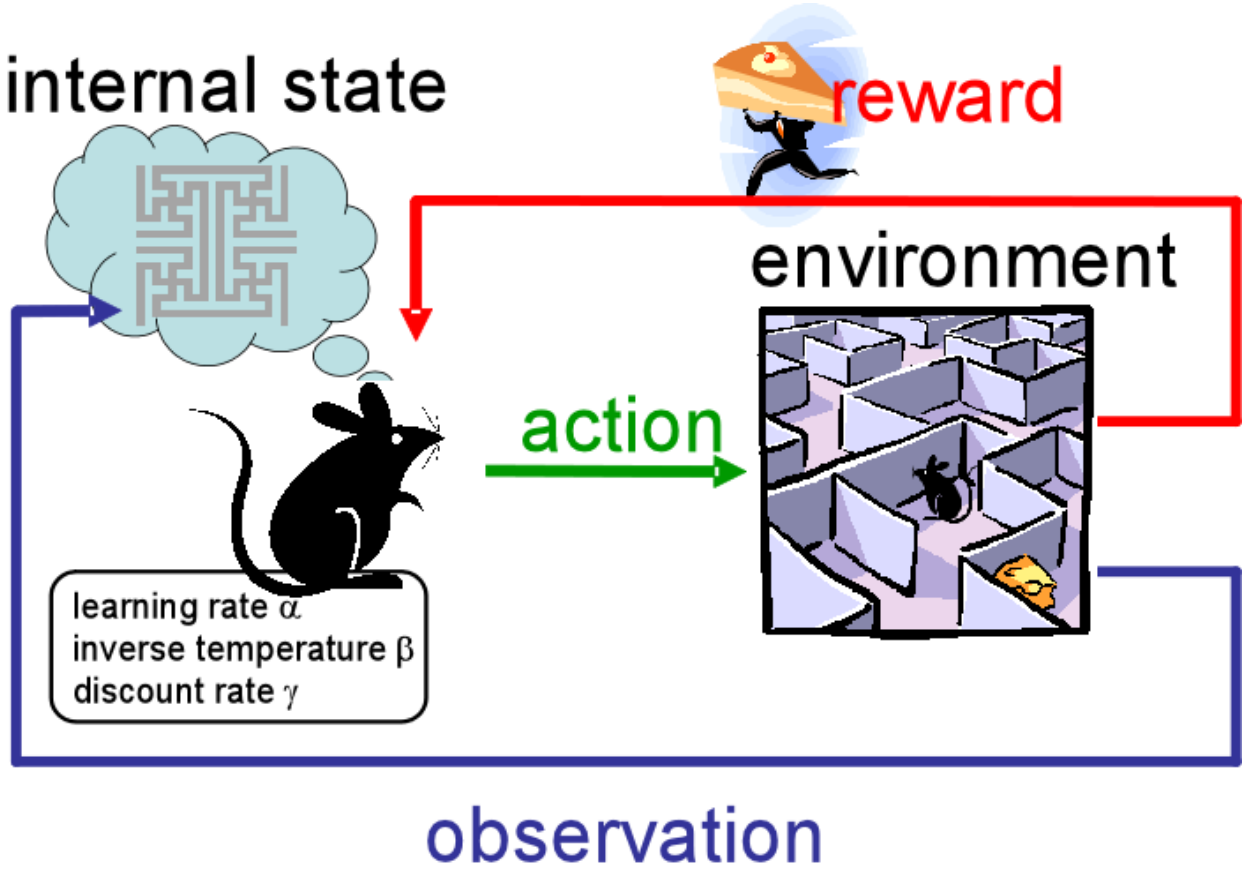
Unsupervised Learning & Clustering

- **Clustering**: Group similar data points (e.g., grouping of safety incident patterns, grouping project tasks by resource demand).
- **Outlier Detection**: Identify abnormal sensor readings or unusual cost inflations.



Reinforcement Learning

- Learning by rewards and penalties in a dynamic setting



Reinforcement Learning

Reinforcement Learning

Could you name some problems that maybe solved using the Unsupervised Learning or RL Methods?

Could you name some problems that maybe solved using the Unsupervised Learning or RL Methods?

- **Unsupervised Learning:**
  - Clustering safety incidents
  - Anomaly detection in sensor data

- **Reinforcement Learning:**
  - Robotics, autonomous equipment operation
  - Scheduling optimization (multi-agent, dynamic environments)

Evaluation Metrics

- **Regression Metrics:**
  - MAE (Mean Absolute Error), RMSE (Root Mean Square Error), R^2
- **Classification Metrics:**
  - Accuracy, Precision, Recall, F1-score, ROC AUC
- **Domain-Specific Considerations:**
  - Over/underestimation cost → which is worse financially?
  - Safety classification → is a false negative (missing a high-risk situation) more dangerous?

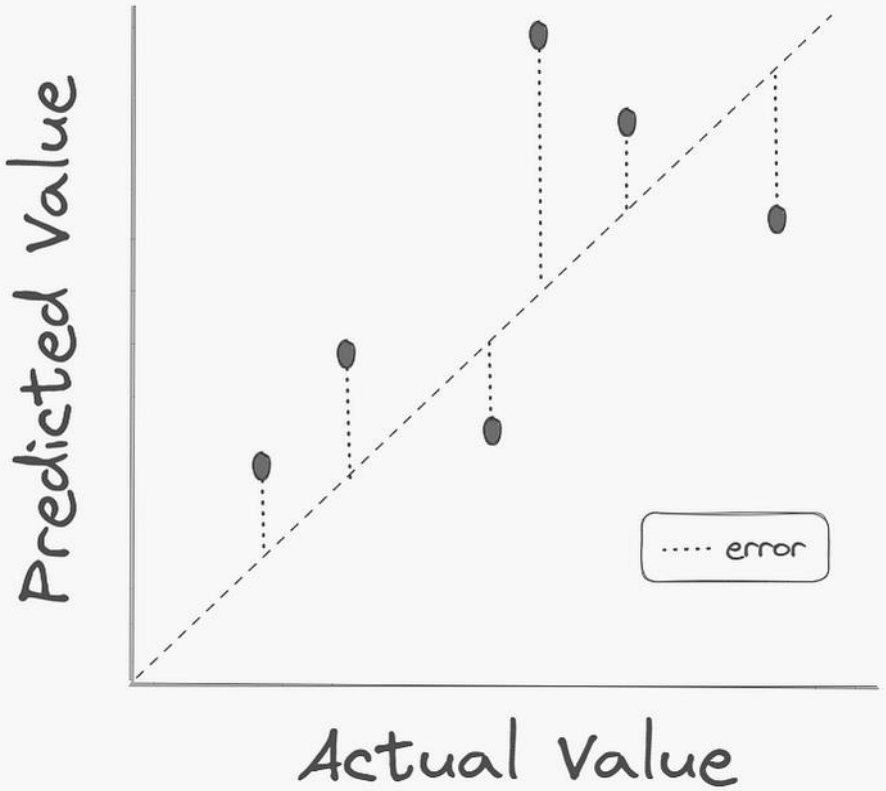
Evaluation Metrics: MAE (Mean Absolute Error), RMSE (Root Mean Square Error), R^2

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

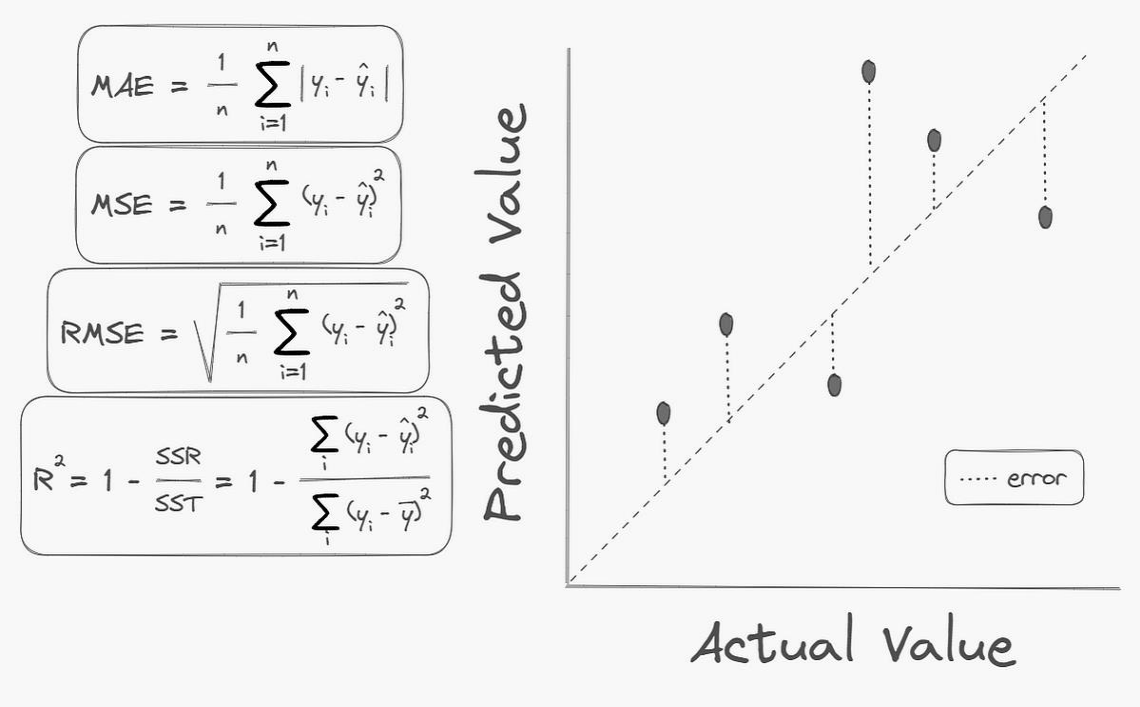
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

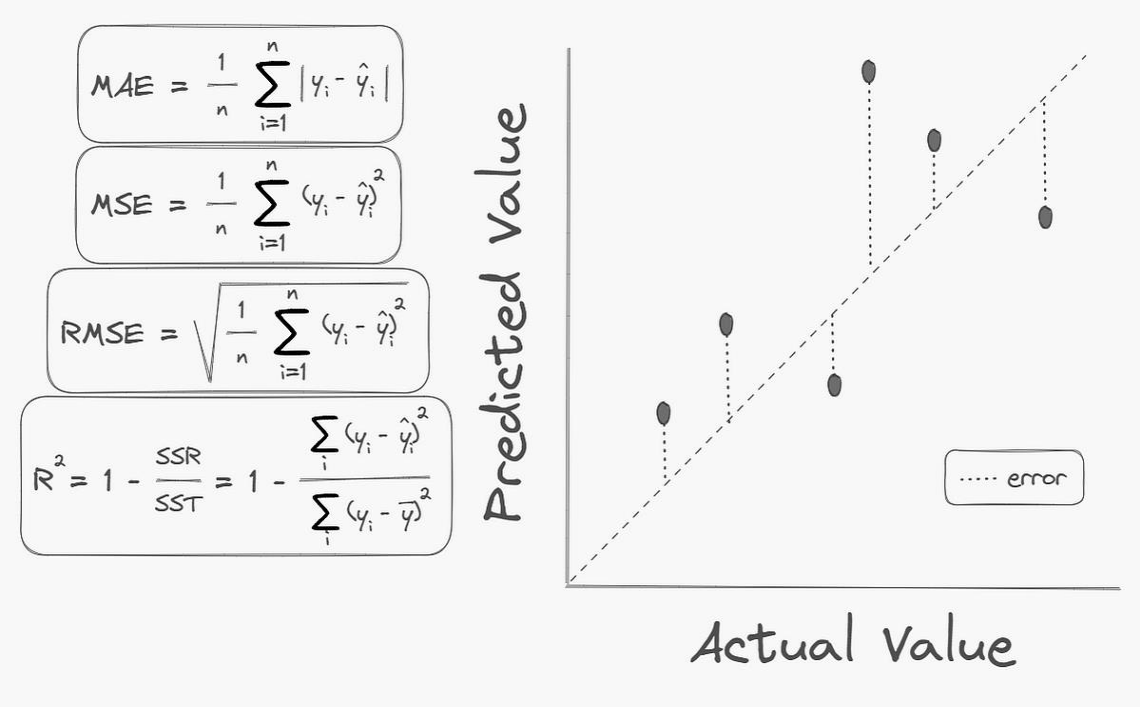


How these metrics indicate model performance?



Cost Estimation: If the **MAE is $5,000** for a house construction cost prediction, on average, your model is off by **$5,000 per project**.

How these metrics indicate model performance?



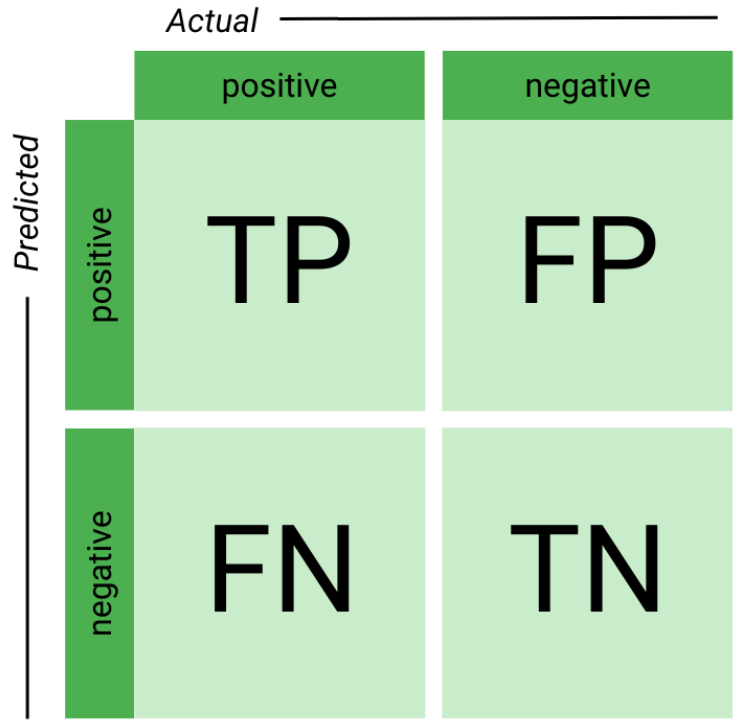
MAE (Mean Absolute Error): Lower is better

- A higher MAE means your model's average prediction error is larger—indicating it performs worse on average.

R^2 (Coefficient of Determination): Higher is better.

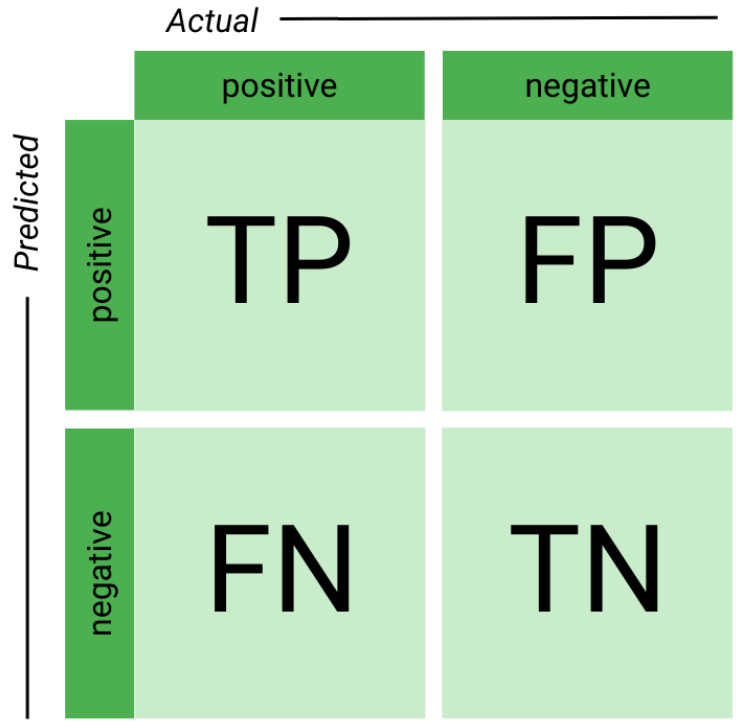
- explains more variance in the target variable

Evaluation Metrics: Accuracy, Precision, Recall, F1-score, ROC AUC



Accuracy	Predictions/ Classifications	$\frac{\text{Correct}}{\text{Correct + Incorrect}}$
Precision	Predictions/ Classifications	$\frac{\text{True Positive}}{\text{True Positive + False Positive}}$
Recall	Predictions/ Classifications	$\frac{\text{True Positive}}{\text{True Positive + False Negative}}$
F1	Predictions/ Classifications	$\frac{\text{2 * True Positive}}{\text{True Positive + 0.5 (False Positive + False Negative)}}$

How these metrics indicate model performance?



The higher, the better!

Accuracy	Predictions/ Classifications	$\frac{\text{Correct}}{\text{Correct + Incorrect}}$
Precision	Predictions/ Classifications	$\frac{\text{True Positive}}{\text{True Positive + False Positive}}$
Recall	Predictions/ Classifications	$\frac{\text{True Positive}}{\text{True Positive + False Negative}}$
F1	Predictions/ Classifications	$\frac{2 * \text{True Positive}}{\text{True Positive + 0.5 (False Positive + False Negative)}}$

How these metrics indicate model performance?

Scenario: On large construction projects, serious safety incidents might comprise only 2% of total daily incident logs; the rest are minor or no-incident days.

Dataset: 5,000 daily logs → 100 serious incidents, 4,900 minor/no incidents.

Naive Model: Predicts every day is minor/no incident.

What is the model accuracy?

How these metrics indicate model performance?

Scenario: On large construction projects, serious safety incidents might comprise only 2% of total daily incident logs; the rest are minor or no-incident days.
Dataset: 5,000 daily logs → 100 serious incidents, 4,900 minor/no incidents.
Naive Model: Predicts every day is minor/no incident.

Accuracy:4,900 correct predictions (minor/no incidents), 0 correct serious incidents
Accuracy = 4,900 / 5,000 = 98%

Please calculate the precision, recall, and F1 by yourself!

	Predicted: No Incident	Predicted: Serious Incident
Actual: Serious Incident	TP = 0	FN = 100
Actual: Minor/No	TN = 4,900	FP = 0

Challenges & Considerations

- **Data Quality & Quantity**: Often insufficient or inconsistent data in construction.
- **Change Management & Culture**: Adopting ML methods requires buy-in from stakeholders who are used to traditional methods.
- **ROI & Scalability**: Justifying investment in ML solutions, especially for smaller firms.
- **Ethical & Legal Aspects**: Privacy of workers' data (wearables, cameras), regulatory constraints.
- **Model Interpretability**: Need for clear explanations to engineers, project managers.

Summary & Next Steps

- ML can significantly improve **cost, schedule, safety, and quality outcomes**
- Both supervised and unsupervised methods have viable use cases in construction
- **Data quality and domain expertise** are critical to successful ML implementations
- Future trends: **deeper integration with IoT, robotics, and digital twins**

First course project presentation

- 10-15 mins for presentations, and 3 mins for Q&A
- Present your team works following the report workflows
- You can use your experiment report for the presentation or prepare some slides for visual aids

Thank You!