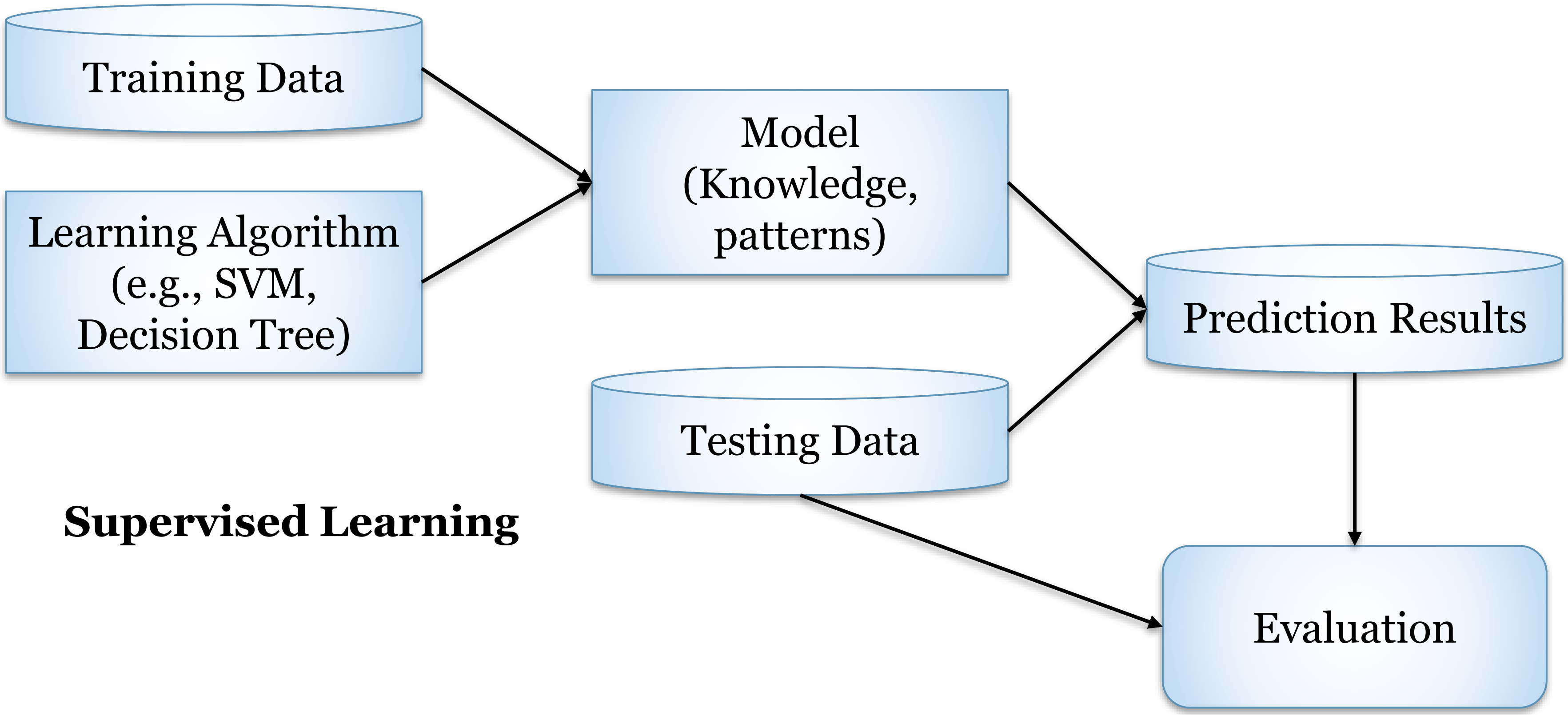
# CISA4358: Senior project and seminar
# Mohammad Abdel-Rahman

## Model Evaluation

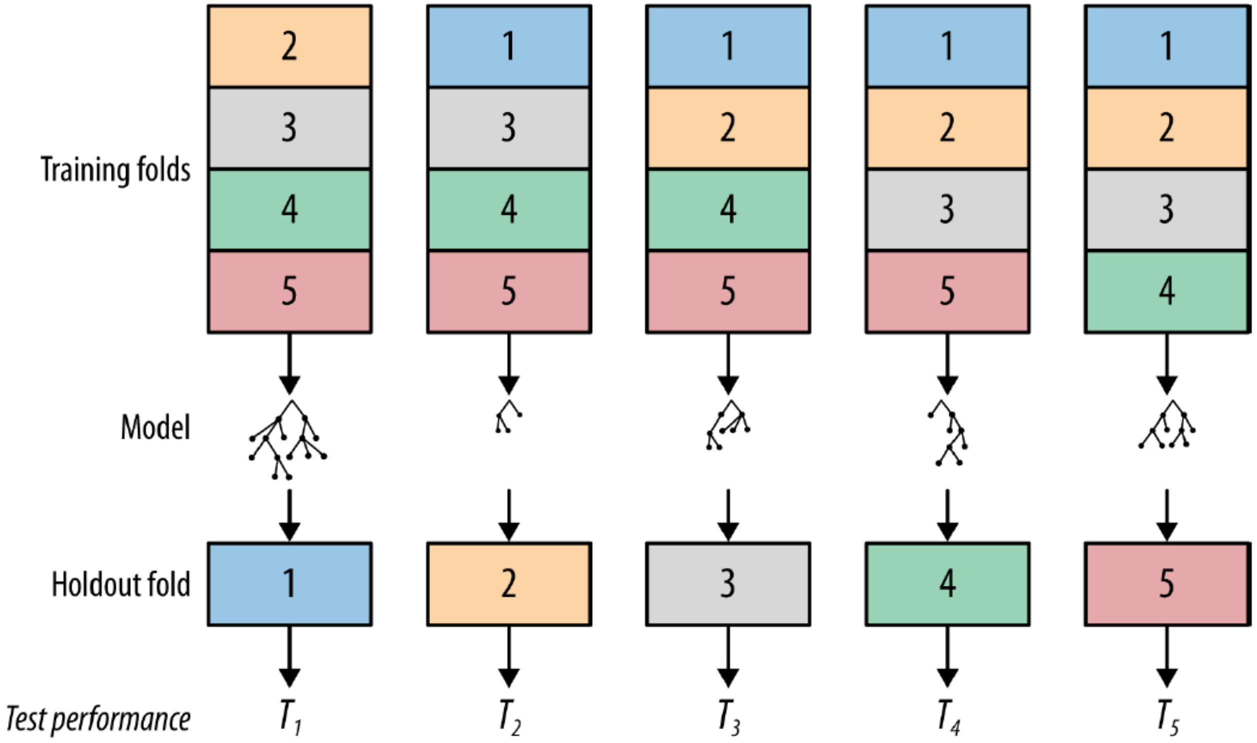
# Evaluating a Predictive Model


Training Data
Learning Algorithm (e.g., SVM, Decision Tree)
Model (Knowledge, patterns)
Testing Data
Prediction Results
Evaluation

**Supervised Learning**

# K-Fold Cross Validation

- Split a data set into *k* partitions (i.e. folds).
- In each iteration, use *1* partition as the testing set and other *k-1* partitions as the training set.
- Aggregate the performance from the *k* tests (e.g., average)
- Variation
  - Stratified
  - Leave-one-out
- Typically considered sufficient:
  - 10 times 10-fold cross validation



Picture from "Data Science for Business"

# Confusion Matrix

| | | Predicted Class | |
|---|---|---|---|
| | | Positive | Negative |
| Actual Class | Positive | True Positive | False Negative |
| | Negative | False Positive | True Negative |

$$\text{Precison} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\text{Number of correctly classified instances}}{\text{Total number of instances classified}}$$

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

# The Problem with Accuracy

- An insurance claim data set contains 100 claim, 10 of which are fraud.
- Model 1: Predicting all claims as non-fraud
- Model 2: Using other information in the claim data

Predicting all claims as non-fraud

| | Predicted Class | | |
|---|---|---|---|
| | | Fraud | Non-Fraud |
| Actual Class | Fraud | 0 | 10 |
| | Non-Fraud | 0 | 90 |

Model 2

| | Predicted Class | | |
|---|---|---|---|
| | | Fraud | Non-Fraud |
| Actual Class | Fraud | 6 | 4 |
| | Non-Fraud | 8 | 82 |

- Accuracy: 90% vs 88%
- Precision: 50% vs 42.86%
- Recall: 0% vs 60%
- F-measure: 0% vs 50%

# Cost-Sensitive Classification

- Unequal cost among classes
    - Example:
        - It costs nothing if a correct prediction is made (TP and TN)
        - It costs $1,000 if a non-fraud case is misclassified as fraud.
        - It costs $10,000 if a fraud case is misclassified as non-fraud.
        - Total loss for using Model 1: 10*10,000 = 100,000
        - Total loss for using Model 2: 4*10,000+8*1,000 = 48,000

Prediction cost in thousand dollars

| | | Predicted Class | |
|---|---|---|---|
| | | Fraud | Non-Fraud |
| Actual Class | Fraud | 0 | 10 |
| | Non-Fraud | 1 | 0 |