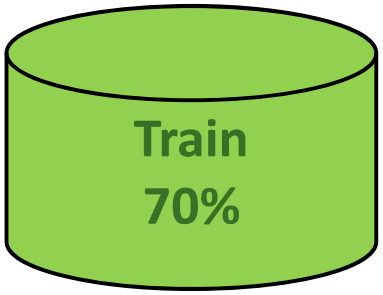
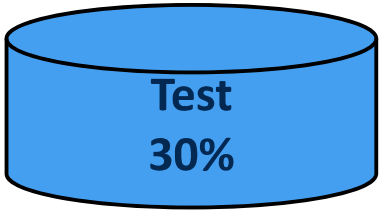
Validation of Machine Learning Models

Dr. Fani Deligianni,
fani.deligianni@glasgow.ac.uk
Lecturer (Assistant Professor)
Lead of the Computing Technologies for Healthcare Theme
https://www.gla.ac.uk/schools/computing/staff/fanideligianni

Performance Evaluation



Train Model



Test Performance

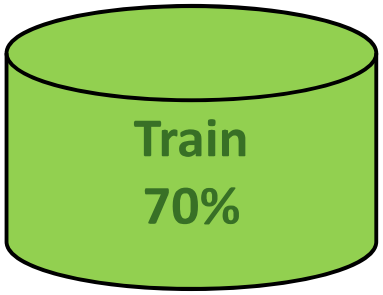
$$R_S(f) = \frac{1}{m}\sum_{i=1}^{m} L(y_i, f(\mathbf{x}_i))$$

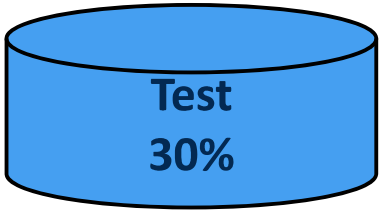
Empirical Risk

Loss function

Classifier

Performance Evaluation – Confidence Intervals





Train Model

Test Performance

$$R_S(f) = \frac{1}{m}\sum_{i=1}^{m} L(y_i, f(\mathbf{x}_i))$$

Empirical Risk

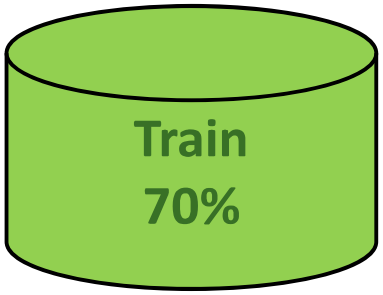
Loss function

Classifier

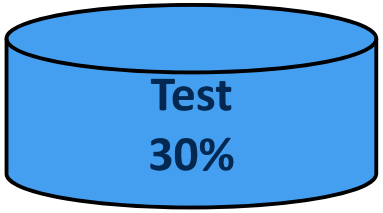
Model the error based on Bernoulli distribution

$$\left|R_T(f) - R(f)\right| \leq t_{1-\delta} = \epsilon = \sqrt{\frac{1}{2m'}\ln\left(\frac{2}{\delta}\right)}$$

Performance Evaluation – Confidence Intervals



Train Model



Test Performance

$$R_S(f) = \frac{1}{m}\sum_{i=1}^{m} L(y_i, f(\mathbf{x}_i))$$

Empirical Risk

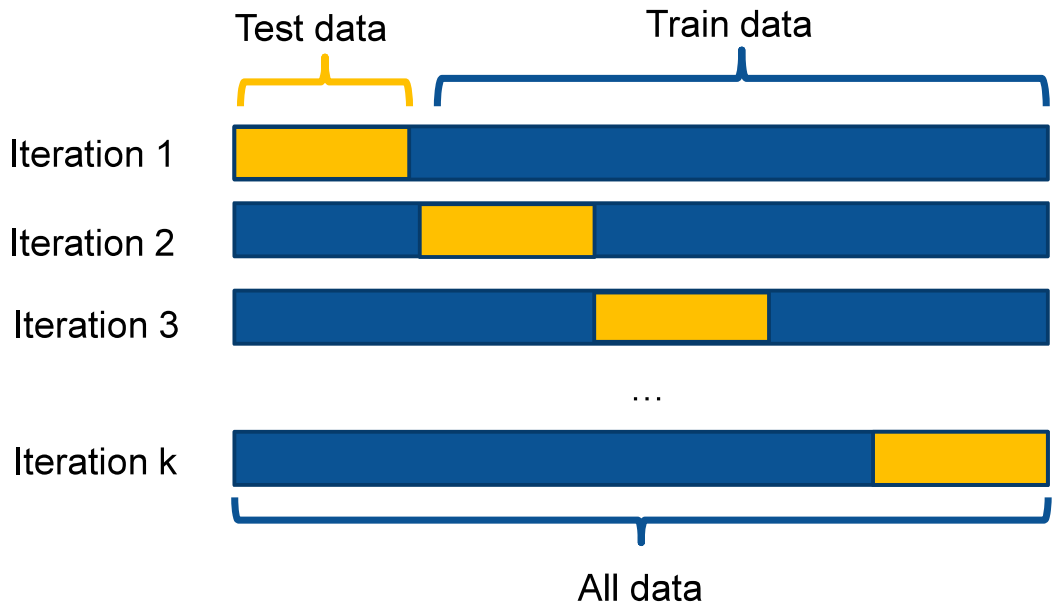
Loss function

Classifier

Model the error based on Bernoulli variable

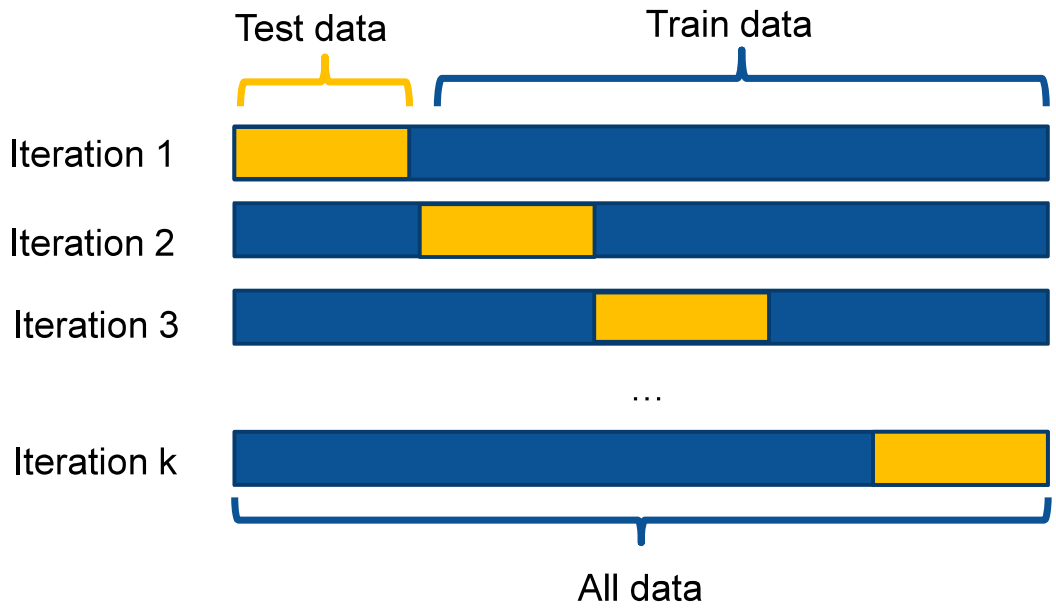
$$m' \geq \frac{1}{2\epsilon^2}\ln\left(\frac{2}{\delta}\right)$$

K-fold Cross Validation



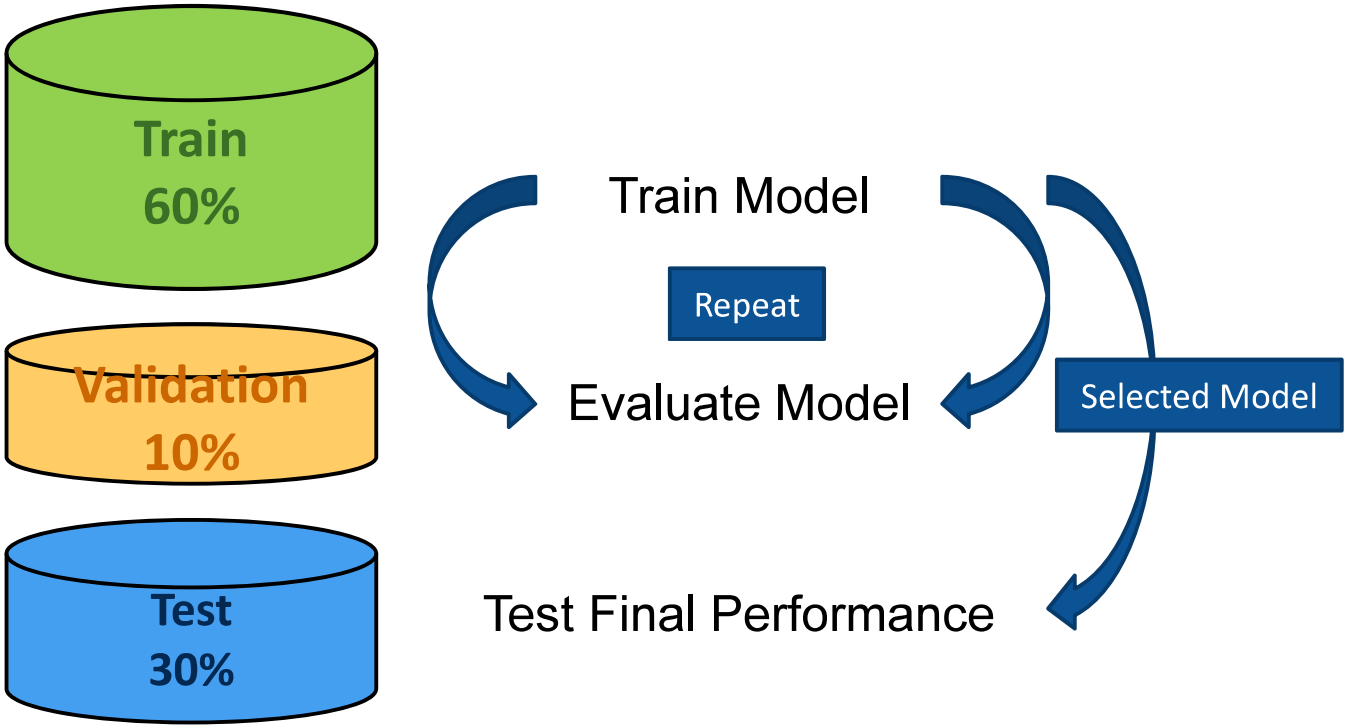
- **Cross-validation** is a resampling procedure used to evaluate machine learning models
- One key advantage is that the testing samples are **independent** between folds

Stratified K-fold Cross Validation

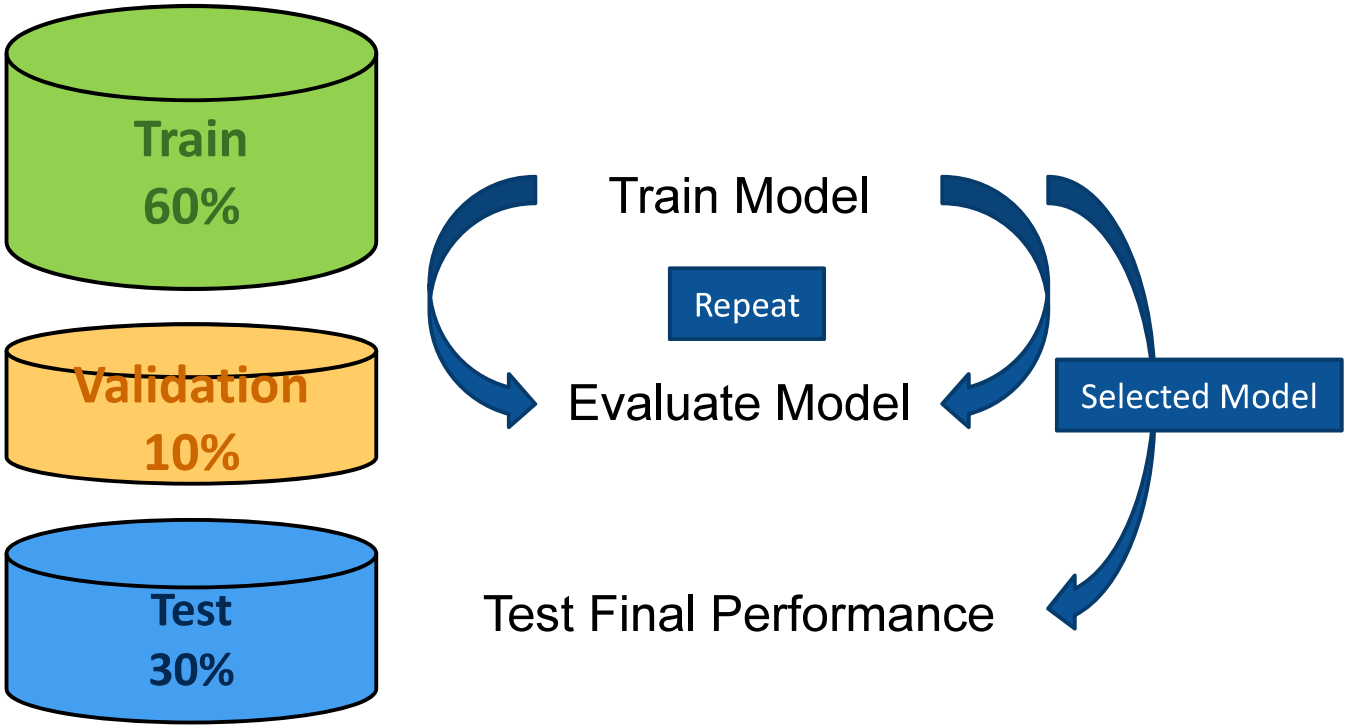


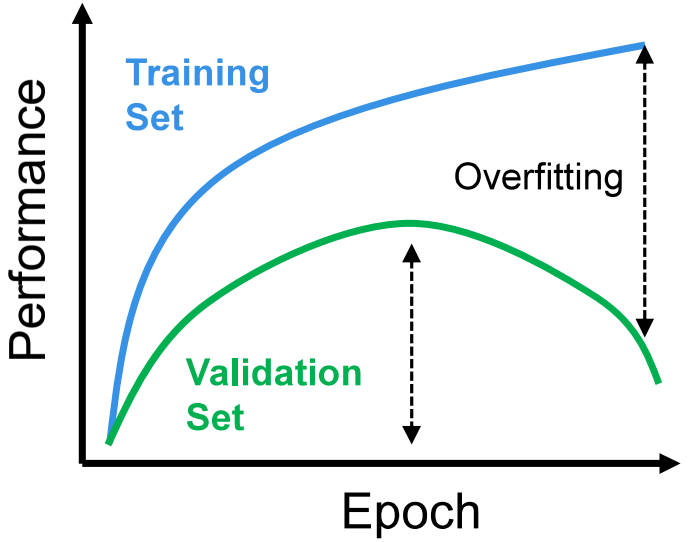
- **Cross-validation** is a resampling procedure used to evaluate machine learning models
- One key advantage is that the testing samples are **independent** between folds
- There is a need to control the class' distributions of the training/testing data

Performance Evaluation vs Model Selection



Performance Evaluation - Overfitting





Subsampling Methods

- **Hold-Out method**
- **Simple Resampling Methods**
  - K-fold Cross Validation
  - Stratified K-fold Cross Validation
  - Leave One Out Cross Validation
- **Multiple Resampling Methods**
  - Repeated k-fold cross validation

Subsampling Methods

- **Hold-Out method**
- **Simple Resampling Methods**
  - K-fold Cross Validation
  - Stratified K-fold Cross Validation
  - Leave One Out Cross Validation
- **Multiple Resampling Methods**
  - Random Sub-sampling
    - Extension of the hold-out method
    - Loss of independence of the data

Subsampling Methods

- **Hold-Out method**
- **Simple Resampling Methods**
  - K-fold Cross Validation
  - Stratified K-fold Cross Validation
  - Leave One Out Cross Validation
- **Multiple Resampling Methods**
  - Bootstrap Sampling
    - Draw with replacement
    - Useful for very small datasets

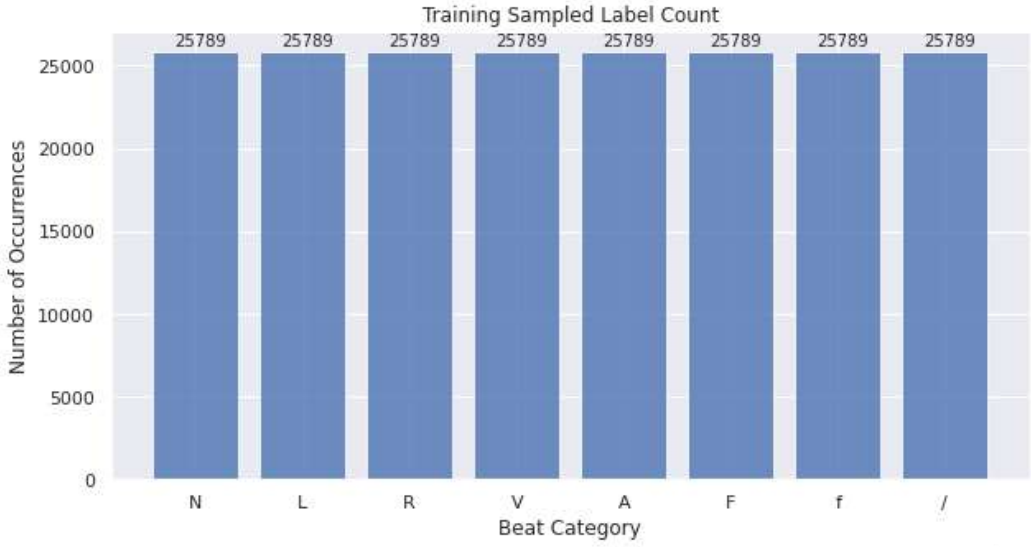
Subsampling Methods

- **Hold-Out method**
- **Simple Resampling Methods**
  - K-fold Cross Validation
  - Stratified K-fold Cross Validation
  - Leave One Out Cross Validation
- **Multiple Resampling Methods**
  - Repeated k-fold cross validation
  - Random Sub-sampling
  - Bootstrap Sampling
  - Permutation Testing

Subsampling in Healthcare Applications

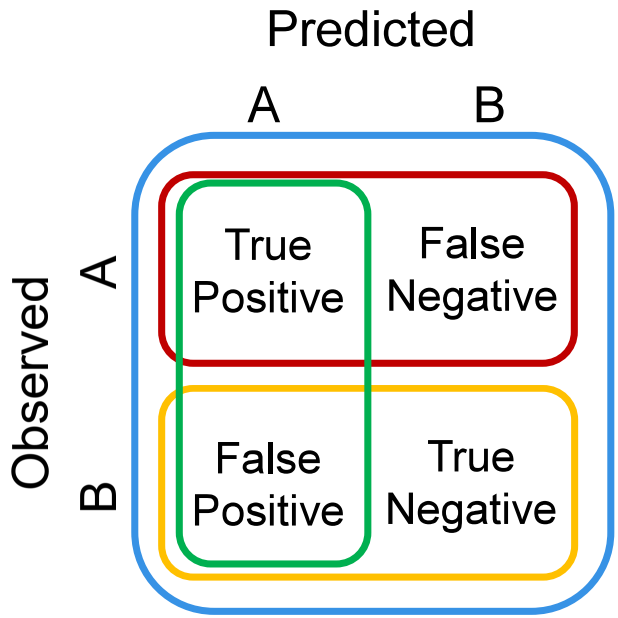
- Intra-subject evaluation
- Inter-subject evaluation

Beat Holdout method



Patient Holdout method

Performance Metrics



Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

Specificity = $\frac{TN}{TN+FP}$

Precision = $\frac{TP}{TP+FP}$

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

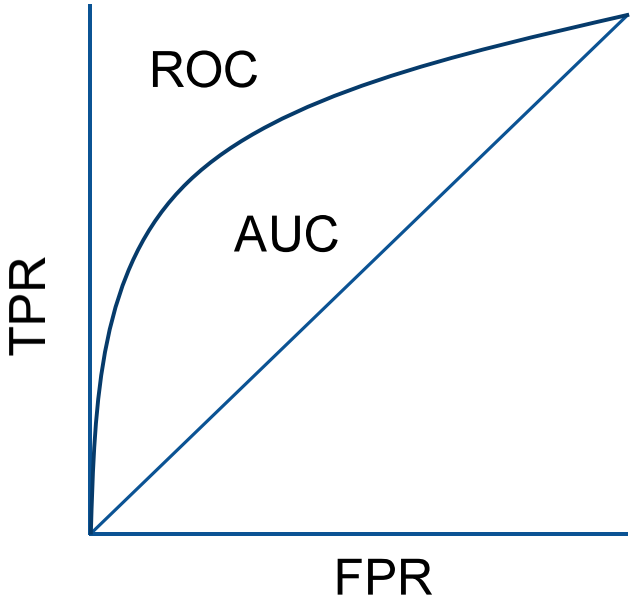
F1-Score

$$F1 = \frac{2}{\frac{1}{recall} * \frac{1}{precision}} = 2 * \frac{precision * recall}{precision + recall}$$

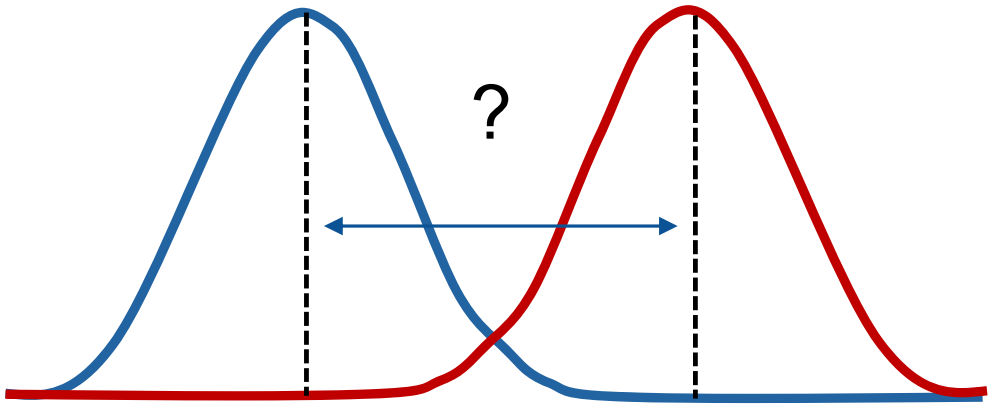
- The F1-score is a way of **combining the precision and recall of the model**, and it is defined as the harmonic mean of the model's precision and recall.

Receiver Operating Characteristic

- **ROC** curve is a performance measurement for the classification problems at various threshold settings
- ROC is a **probability curve** and AUC represents the degree or **measure of separability**
- The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.

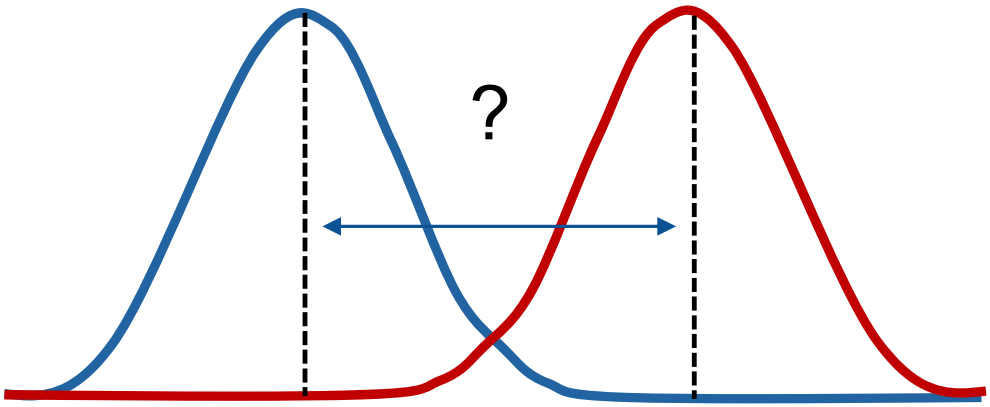


Comparing the Performance of Algorithms



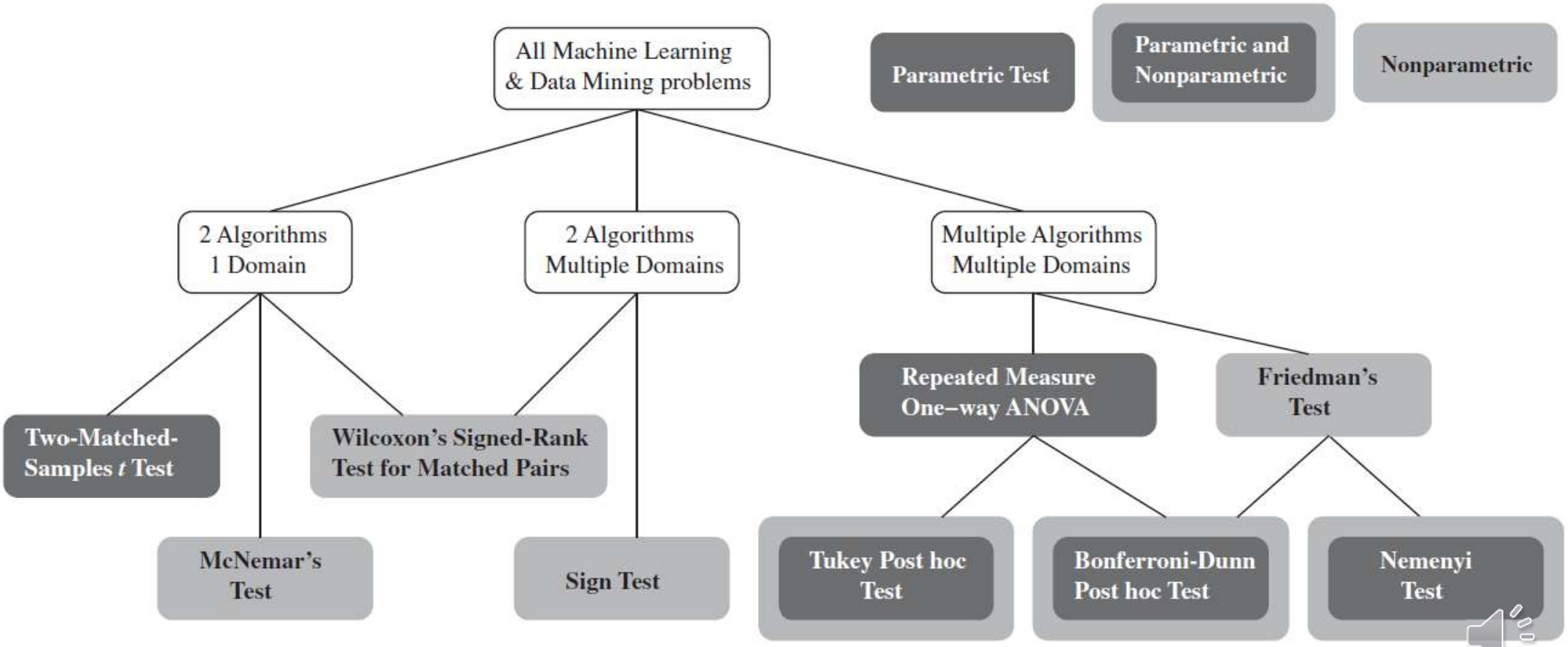
- Null-Hypothesis Statistical Testing for algorithm evaluation
- The **t-test is a statistical hypothesis** test in which the test statistic follows a Student's t-distribution under the null hypothesis.
- The **common assumptions:** independence of measurements, normality, adequate sample size, and equality of variance in standard deviation

Non parametric tests



- The **Wilcoxon signed-rank Test** is a general test to compare distributions in paired samples.
- This test is usually the preferred **alternative to the Paired t-test** when the assumptions are not satisfied.
- It determines if the two populations seem to be the same or different based upon the ranks of the absolute differences.
- Ranking procedures are commonly used in **non-parametric methods** as this moderates the effect of any outliers.

Overview of Hypothesis Testing



Japkowicz et al. Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, 2011

Summary

- Subsampling frameworks
- Performance metrics
- Significance testing

References

- Japkowicz et al. Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, 2011