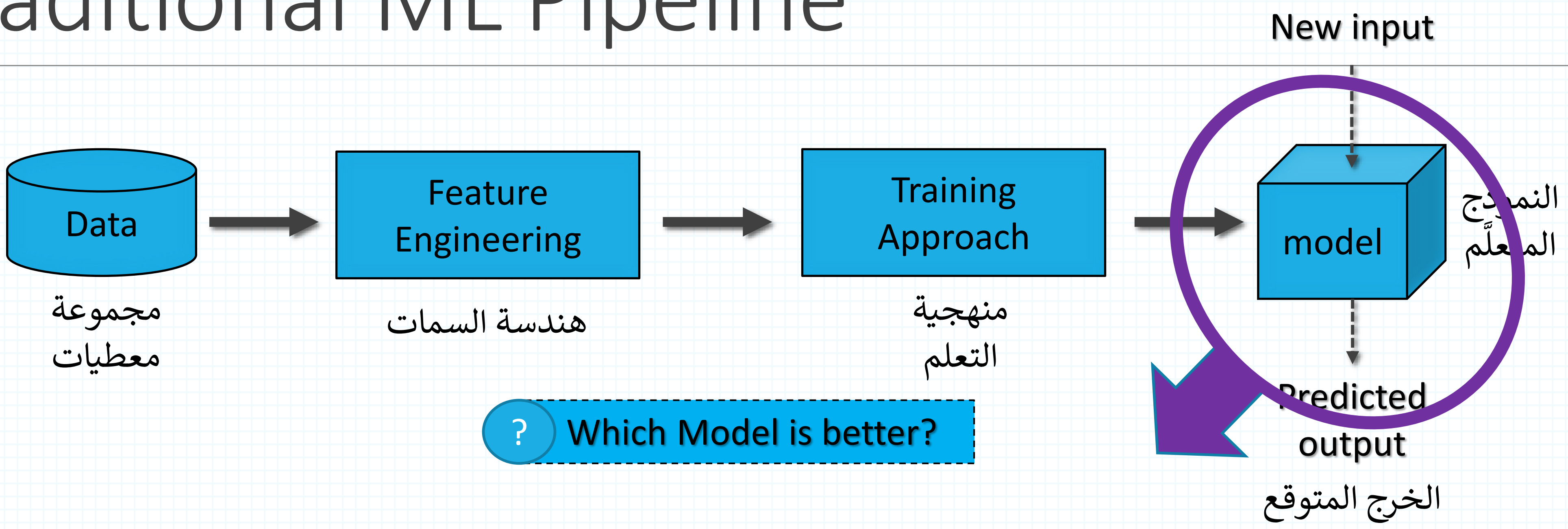
الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

مقرر تعلم الآلة | كلية الهندسة المعلوماتية | المحاضرة الرابعة

استراتيجيات التدريب والاختبار

د. رياض سنبل

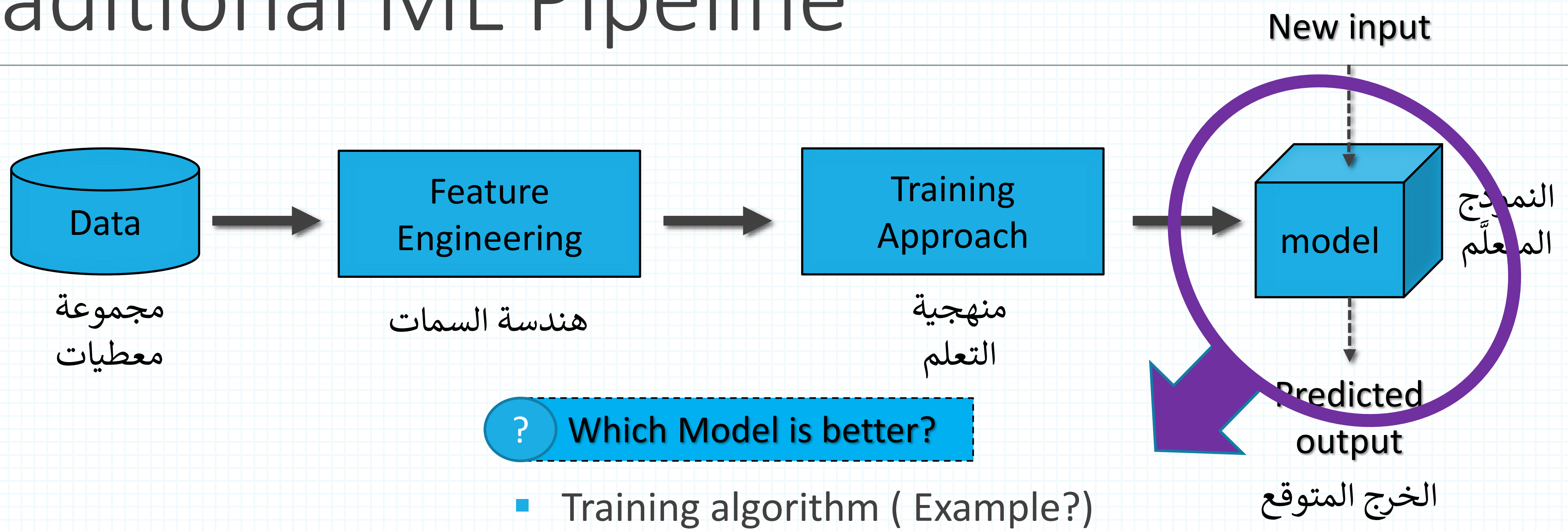
Traditional ML Pipeline



"All models are wrong; some are useful."

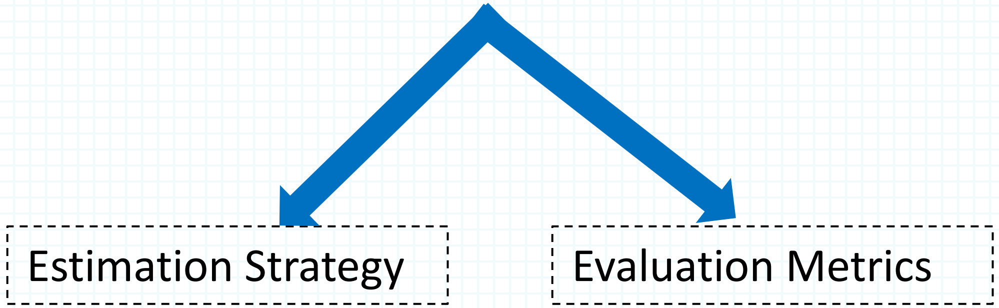
—George E. P. Box

Traditional ML Pipeline



? Which Model is better?

- Training algorithm (Example?)
- Parameters (Hyperparameters)?
- Handle unseen cases



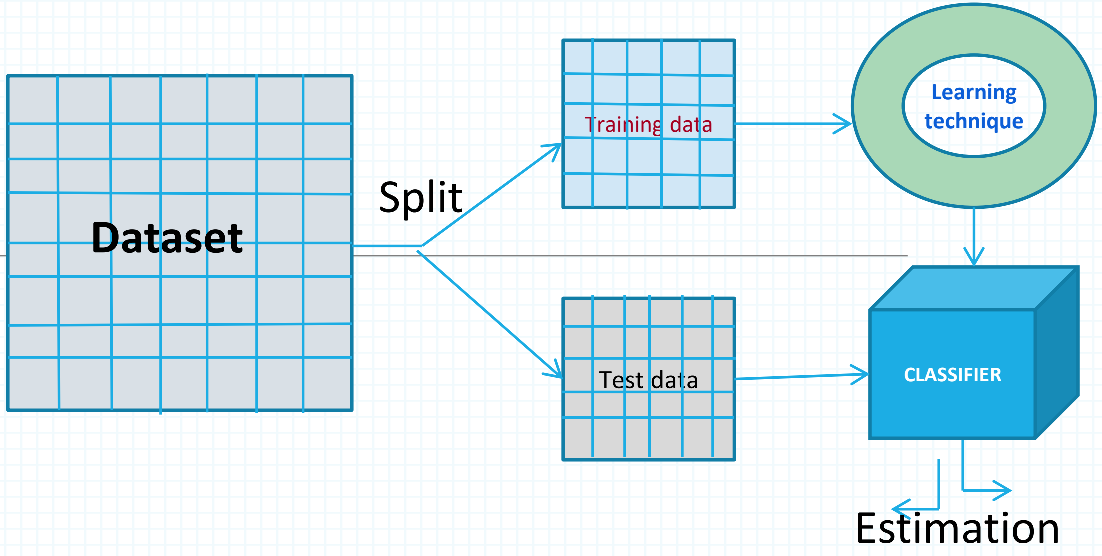
Hyperparameters

- Hyperparameters are the explicitly specified parameters that control the training process.
- What are the hyperparameters for decision trees?

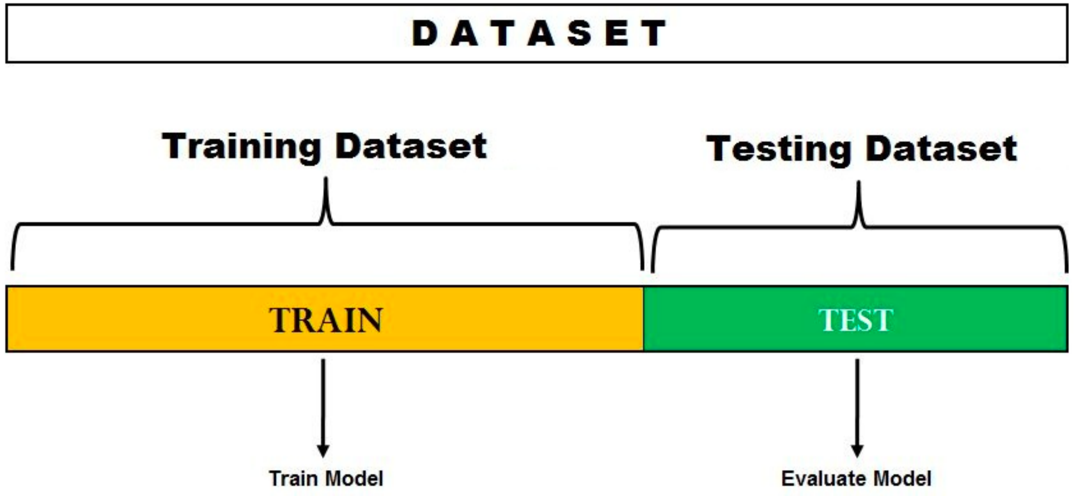
class `sklearn.tree.`**`DecisionTreeClassifier`**(**, criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, class_weight=None, ccp_alpha=0.0*) [source]

Estimation Strategies

The holdout method

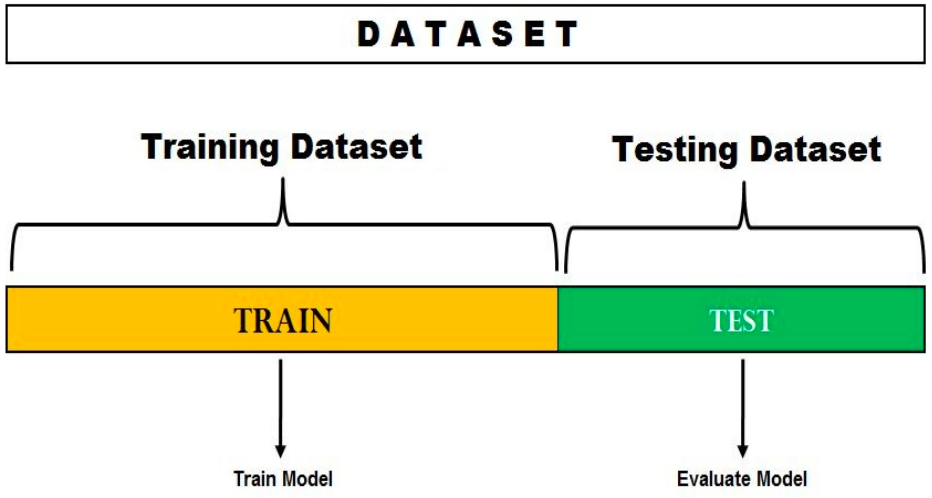


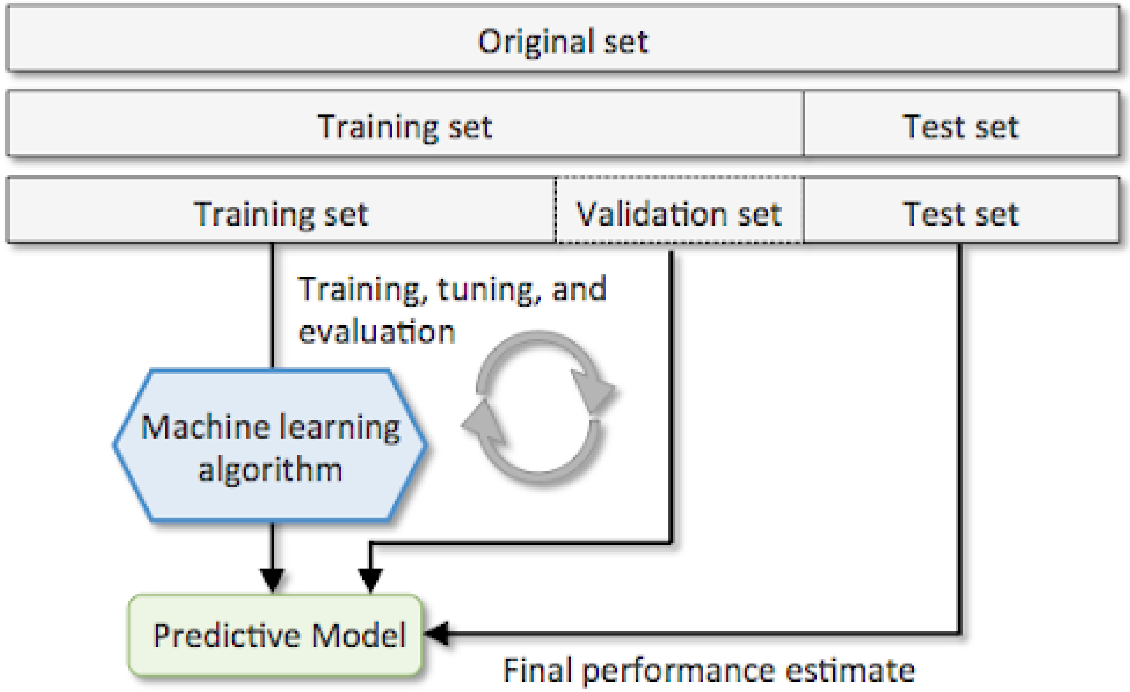
- Split dataset into two groups
  - <u>Training set</u>: used to train the classifier
  - <u>Test set</u>: used to estimate the error rate of the trained classifier.
- Ratio of training and testing sets is at the discretion of analyst;
  - Typically **1:1 or 2:1**, and there is a trade-off between these sizes of these two sets.
  - If the training set is too large, then model may be good enough, but estimation may be less reliable due to small testing set and vice-versa.



Holdout method

But how can we tune hyper-parameters?



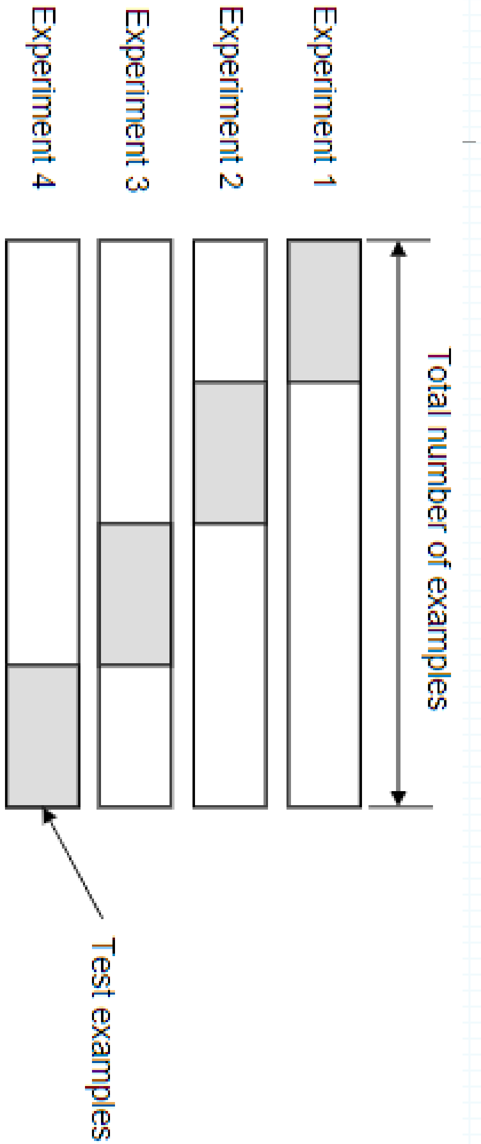


Holdout method - Drawbacks

- The holdout method has **two basic drawbacks**
  - In problems where we have a sparse dataset we may not be able to afford **the "luxury" of setting aside a portion** of the dataset for testing
  - Since it is **a single train-and-test experiment**, the holdout estimate of error rate will be misleading if we happen to get an "unfortunate" split
- The limitations of the holdout can be overcome with a family of re-sampling methods at the expense of higher computational cost
  - Cross Validation
    - Random Subsampling
    - K-Fold Cross-Validation
    - Leave-one-out Cross-Validation
  - Bootstrap

K-Fold Cross-validation

- Create a K-fold partition of the dataset
- For each of K experiments, use K-1 folds for training and a different fold for testing
- The advantage of K-Fold Cross validation is that **all the examples** in the dataset are eventually <u>used for both training and testing</u>
- The true error is estimated as the average error rate on test examples



Leave-one-out Cross-Validation

- Leave-one-out is the degenerate case of K-Fold Cross Validation, where K is chosen as the total number of examples
  - For a dataset with N examples, perform N experiments
  - For each experiment use N-1 examples for training and the remaining example for testing