Introduction to Data Science

(Lecture 13)

Dr. Mohammad Pourhomayoun

Assistant Professor

Computer Science Department

California State University, Los Angeles

Review:

Evaluating the Accuracy of a Predictive Model Using a Random Training and Testing Sets

Evaluating The Accuracy Of Our Predictive Model

Here is a simple way to evaluate the accuracy of our predictive model:

1- Let's split the dataset **RANDOMLY** into two new datasets: **Training Set** (e.g. 70% of the data samples) and **Testing Set** (30% of the data).

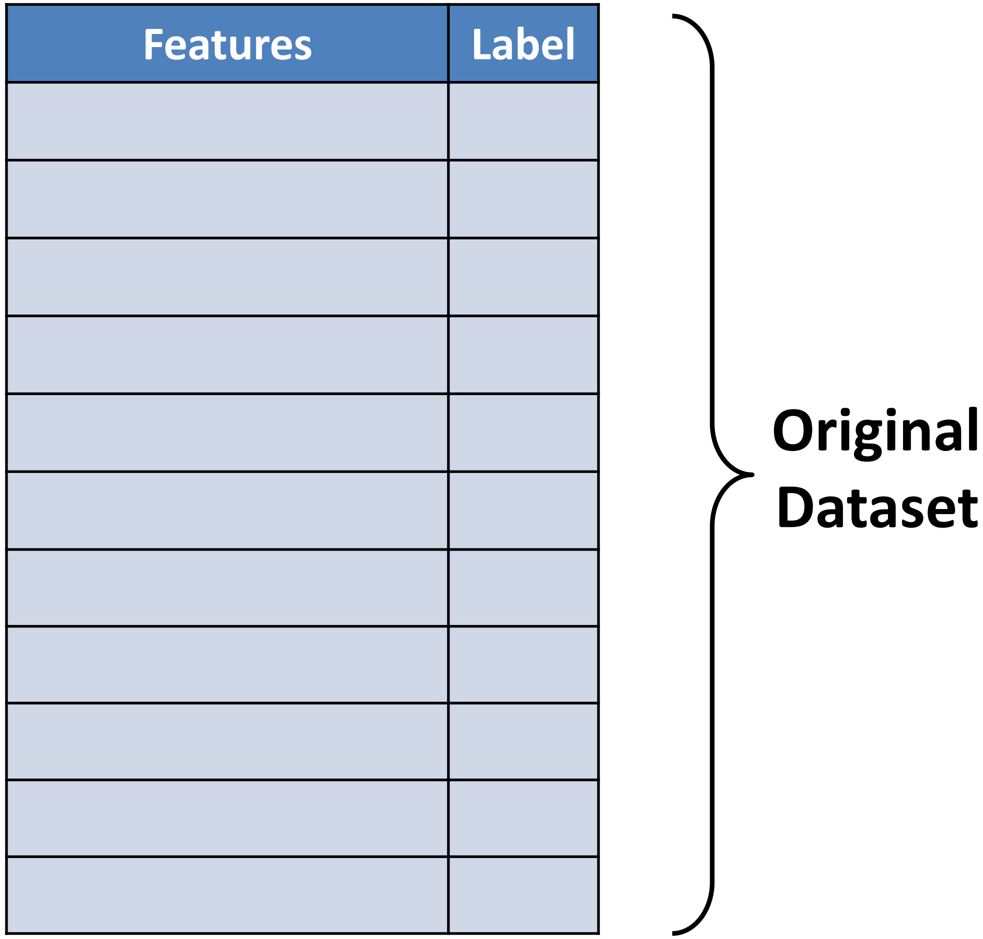
2- Let's **pretend** that we do **NOT** know the label of the Testing Set!

3- Let's Train the model **ONLY on Training Set**, and then Predict on the Testing Set!

4- After prediction, we can compare the **predicted labels** for the Testing Set with the **actual labels** of it to evaluate the accuracy of our prediction!

We will learn more techniques for model evaluation (e.g. **Cross Validation** method) later in this class!

Training and Testing Sets

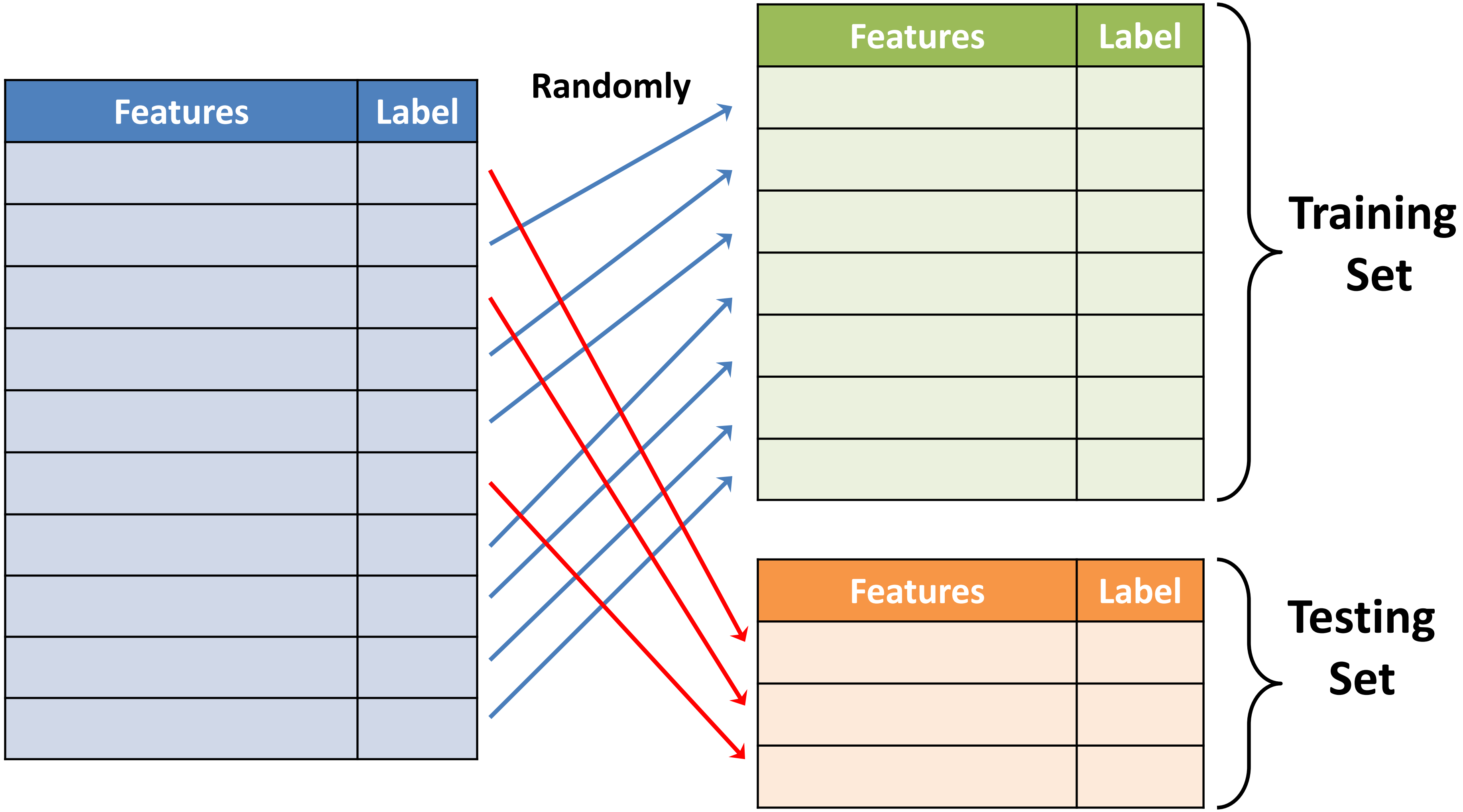


Features

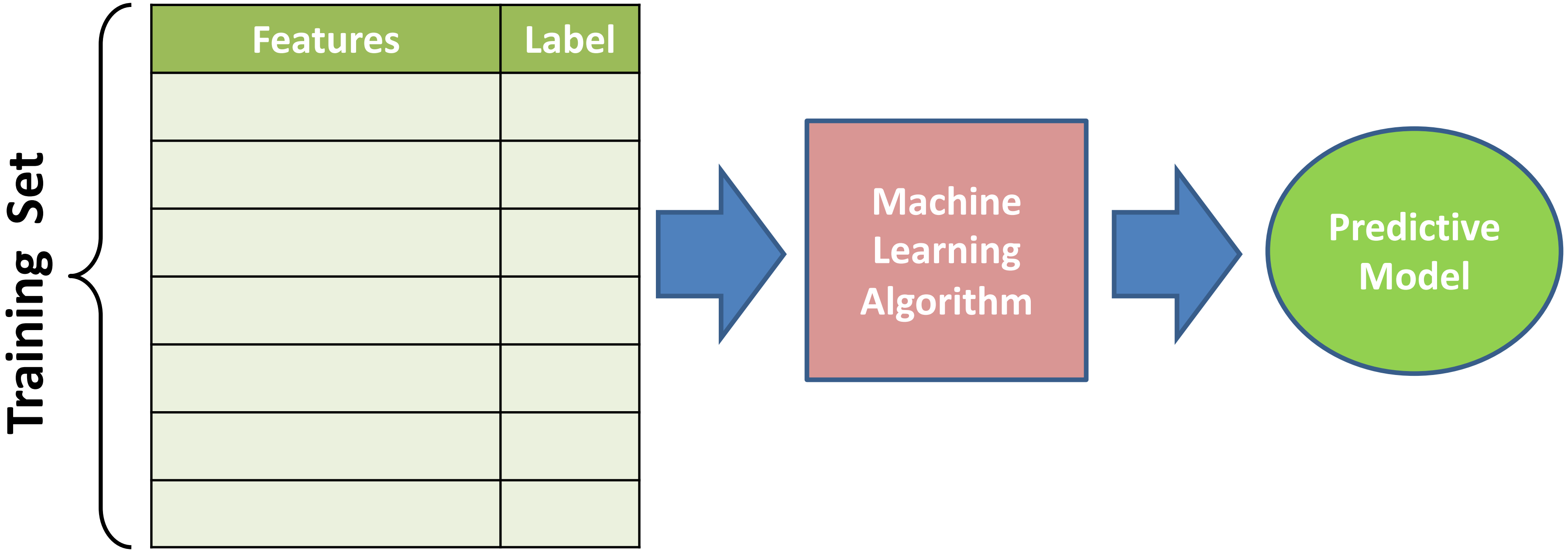
Label

Original Dataset

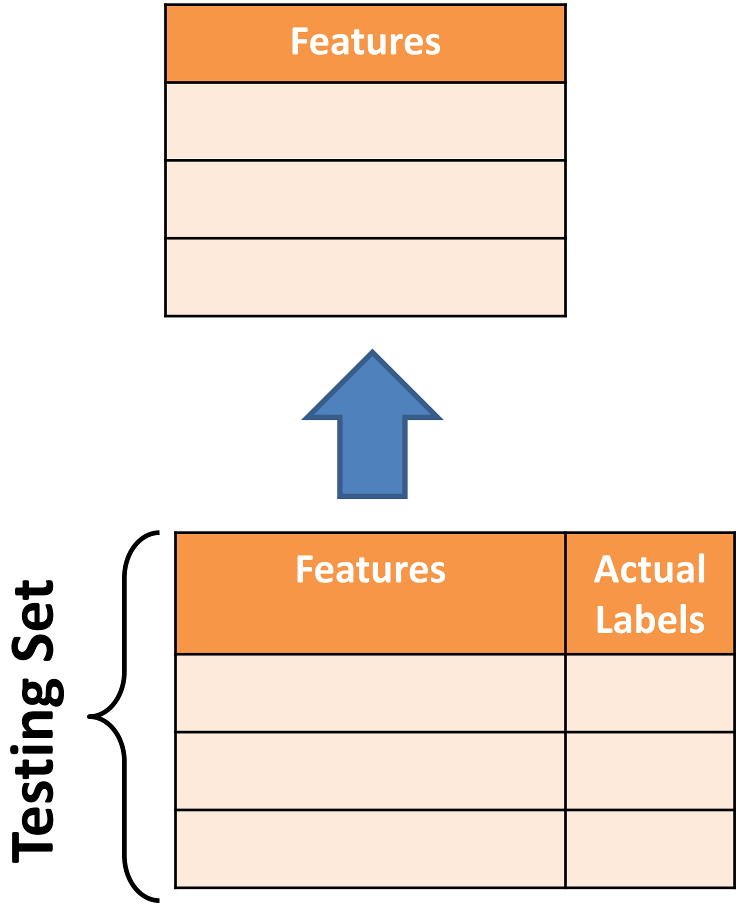
Training and Testing Sets



Training Stage



Testing Stage

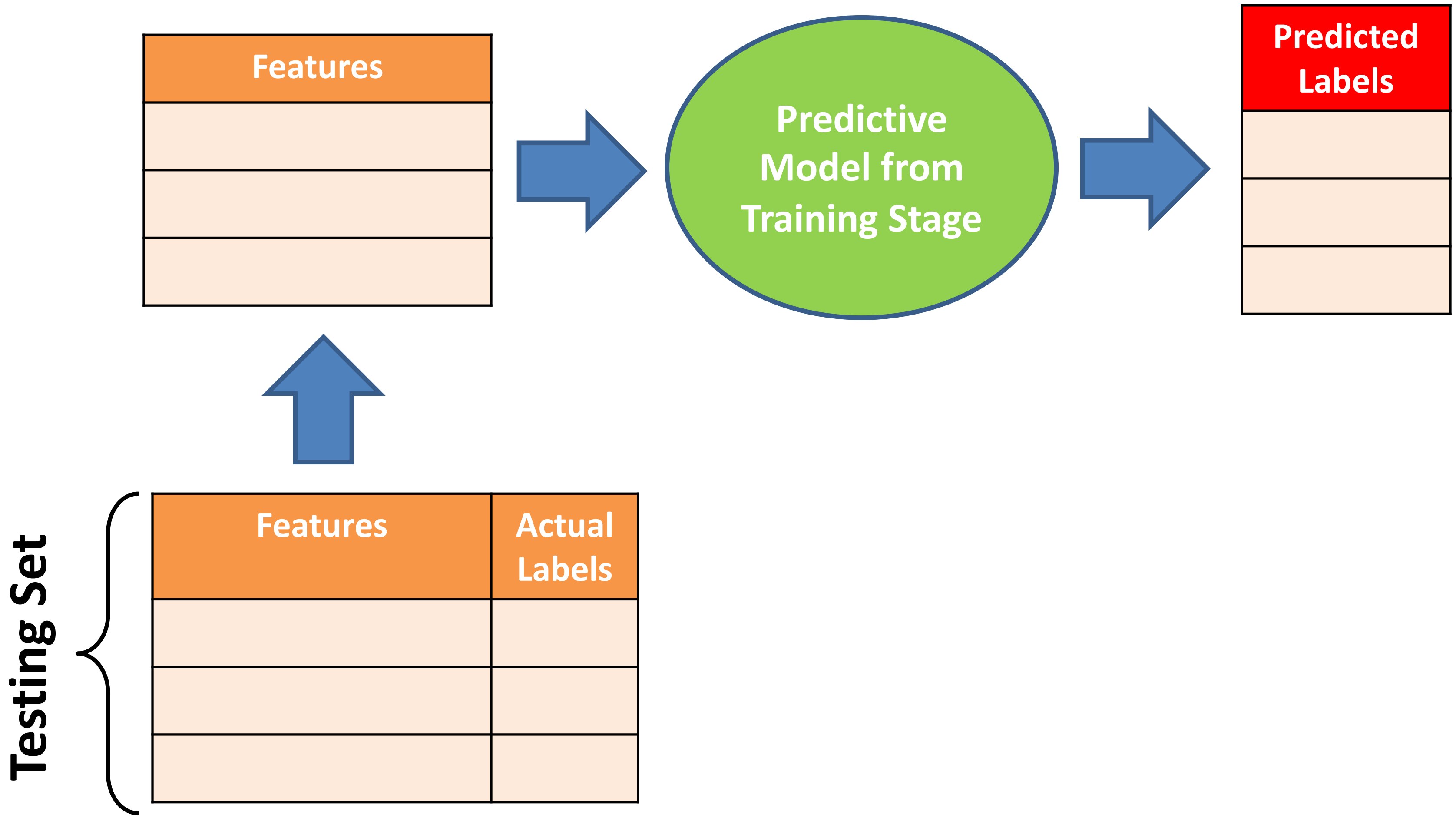


Features

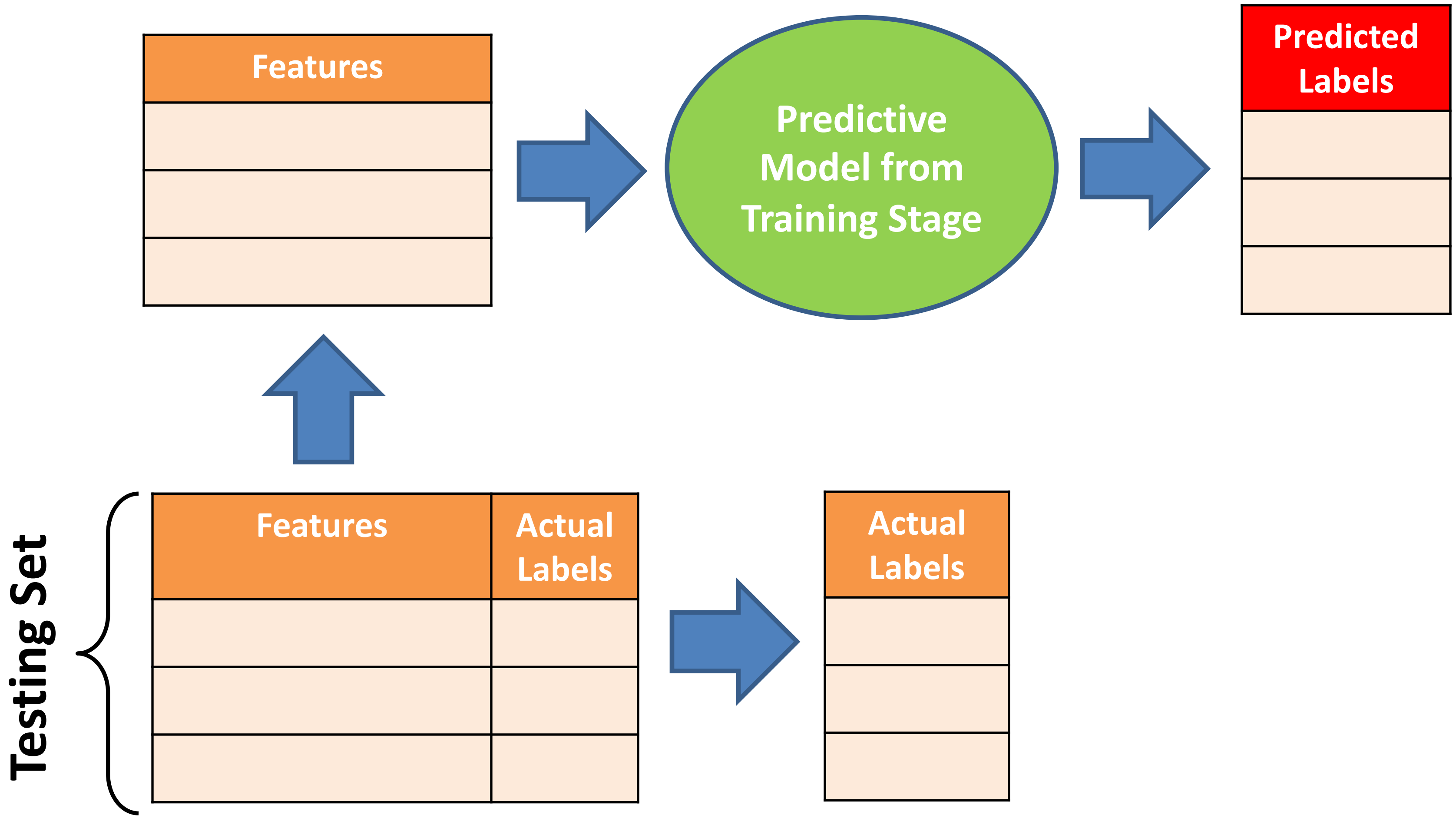
Testing Set

Features	Actual Labels

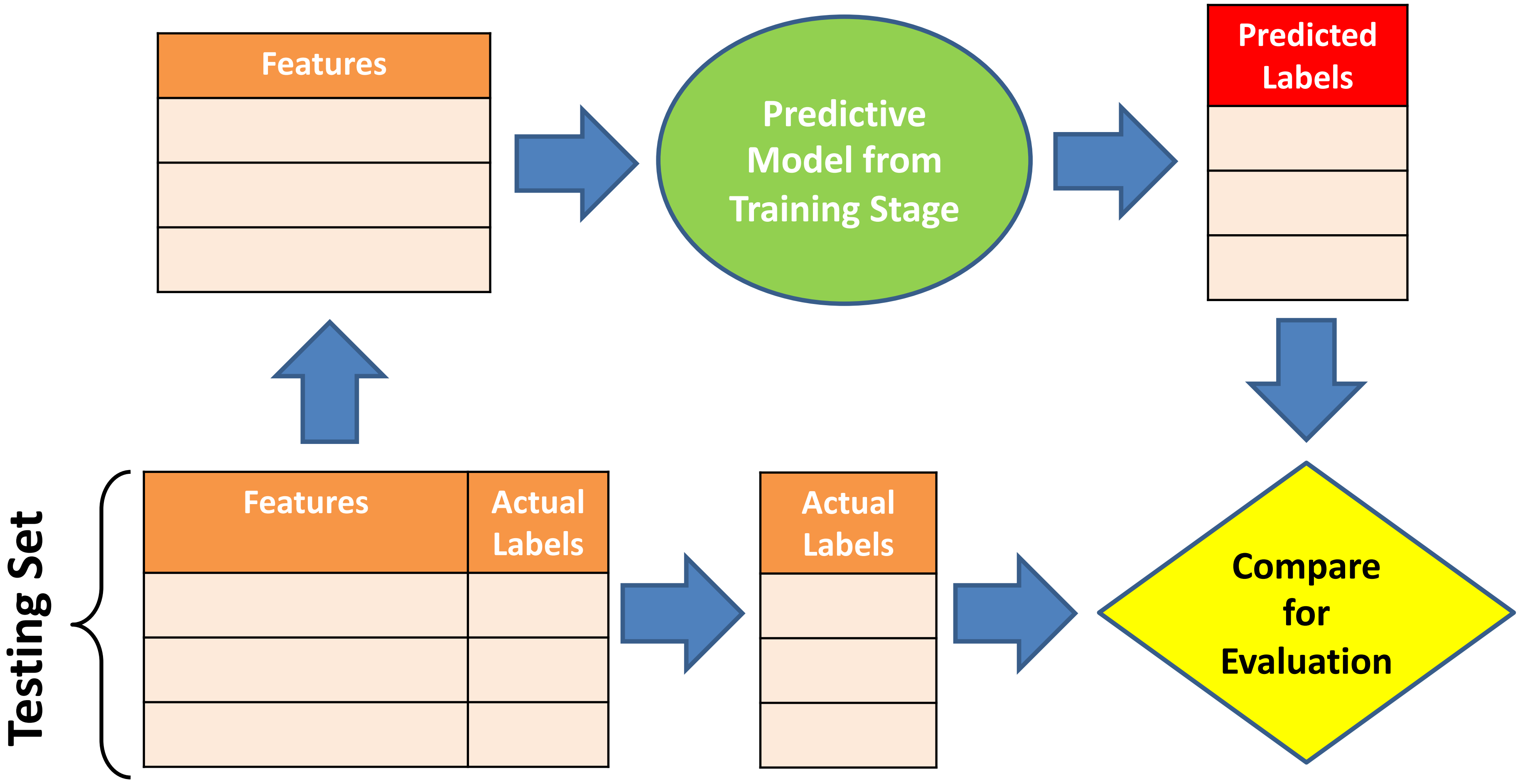
Testing Stage



Testing Stage



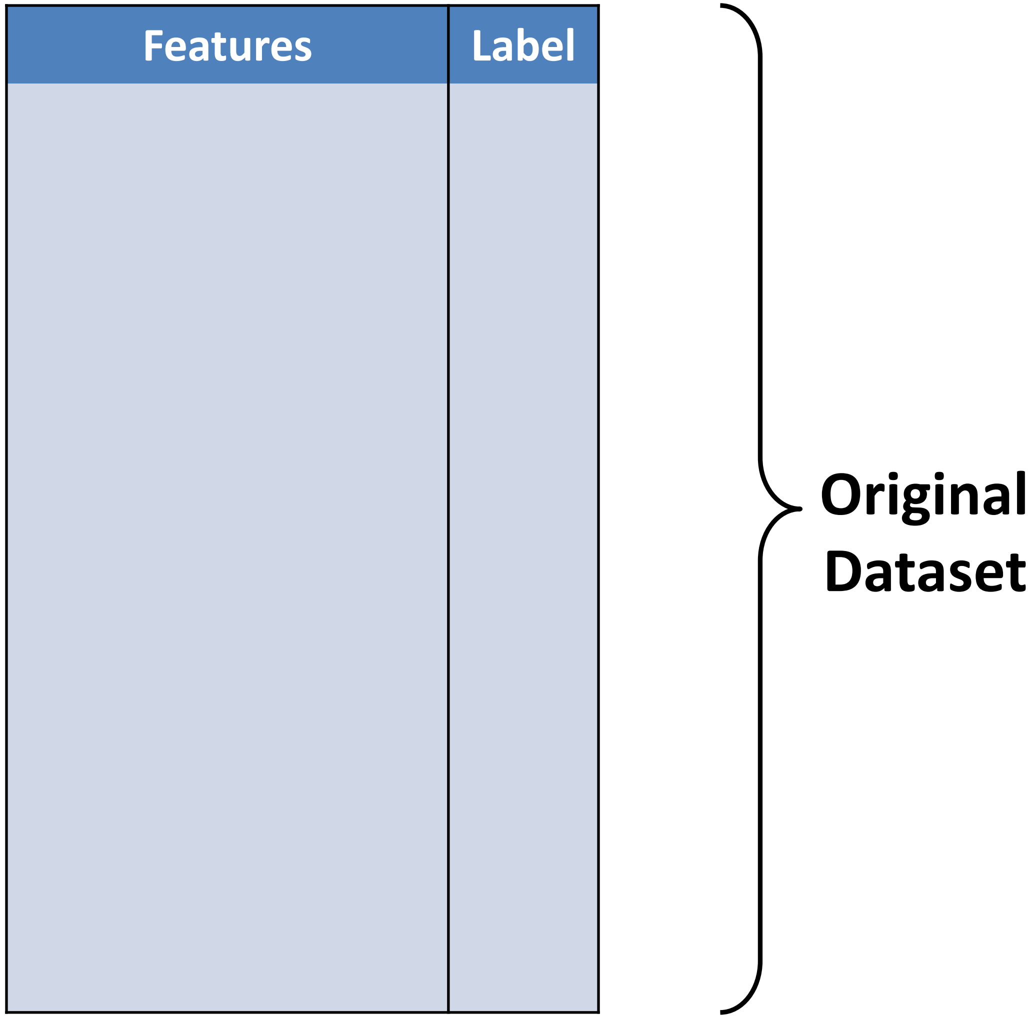
Testing Stage



Evaluating the Accuracy of a Predictive Model Using

Cross Validation

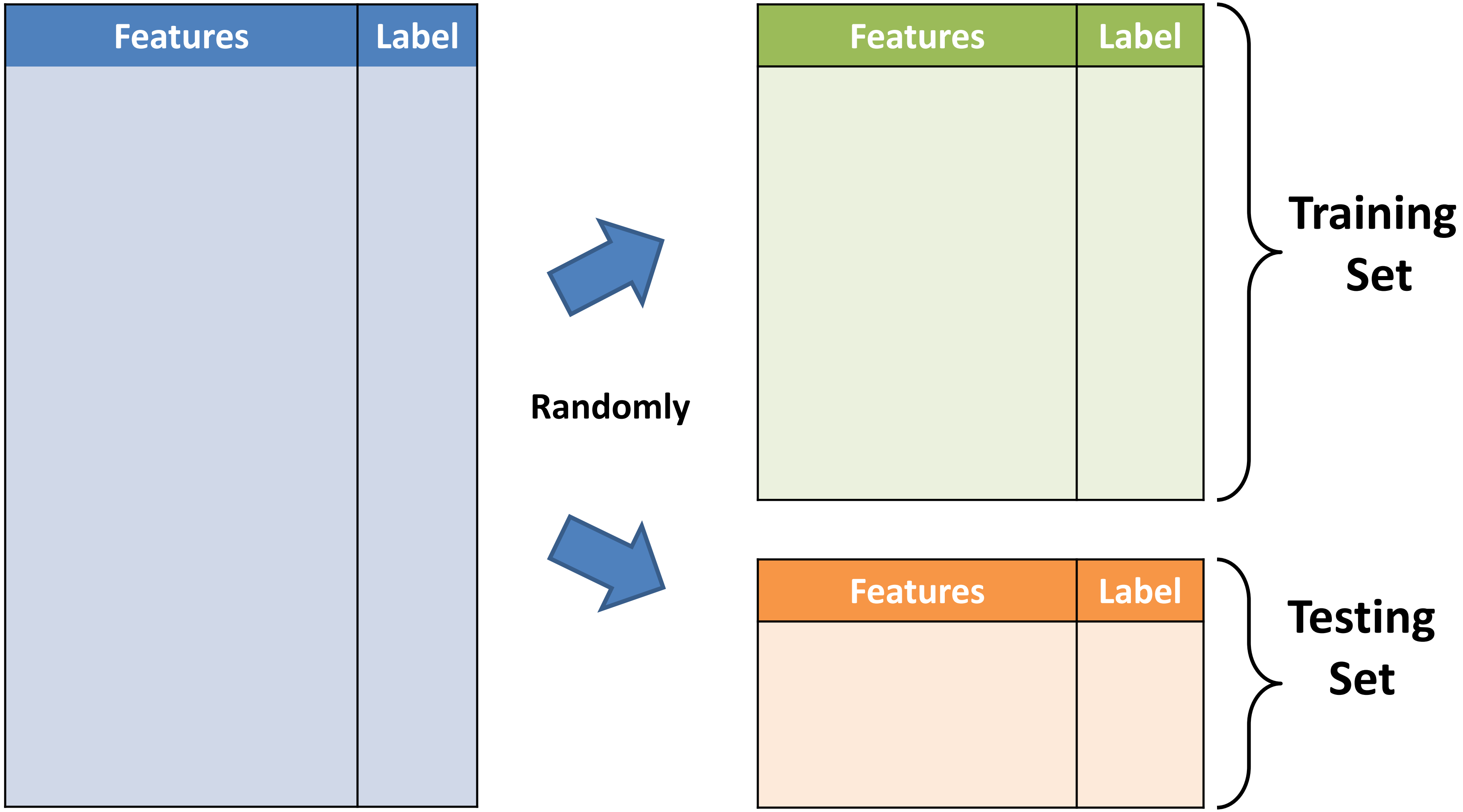
Training and Testing Sets



Features	Label

Original Dataset

Training and Testing Sets



Cross Validation

- We saw how to split the dataset into Training and Testing sets, Fit the model on "training set", and then predict on "testing set" to evaluate the accuracy.
- The problem with this method is that **the results may depend on the choice of split**. For example, if you are lucky, some easily predictable samples may happen to be located in the testing set (or vice versa!).
- In order to get fair results, we can repeat the splitting process several times, compute the prediction accuracy for each split, and then average the results.
- **Cross Validation tries to repeat the splitting procedure K times in a smart way such that all data samples will be used in "testing set" one time and in "Training Set" (K-1) times!**

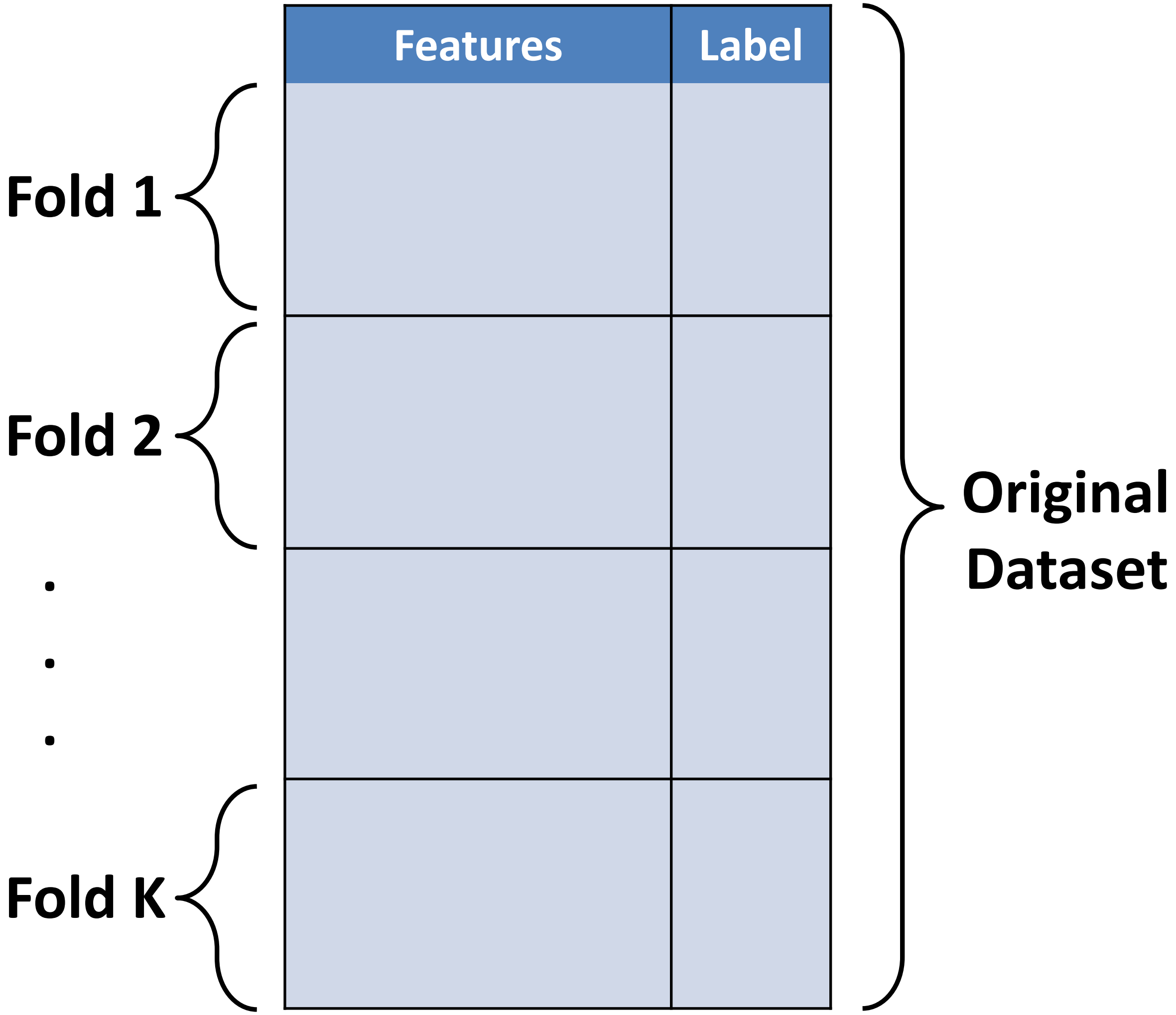
Cross Validation

Three main steps for K-fold cross-validation:

1. Partition the dataset Randomly into K equal, non-overlapping sections (called Fold).
2. Use one of the sections as **testing set** at a time and the union of the other (K-1) sections as the **training set**. Perform training stage, testing stage, and compute the accuracy based on the split each time. Repeat this procedure K times, so that each one of the K sections is used as **testing set** one time, and as a part of **training set** (K-1) times.
3. Calculate the average of the accuracies as final result.

Note: K is arbitrary, but Using K=10 (10-fold cross-validation) is very common and recommended in machine learning.

Cross Validation



Features

Label

Fold 1

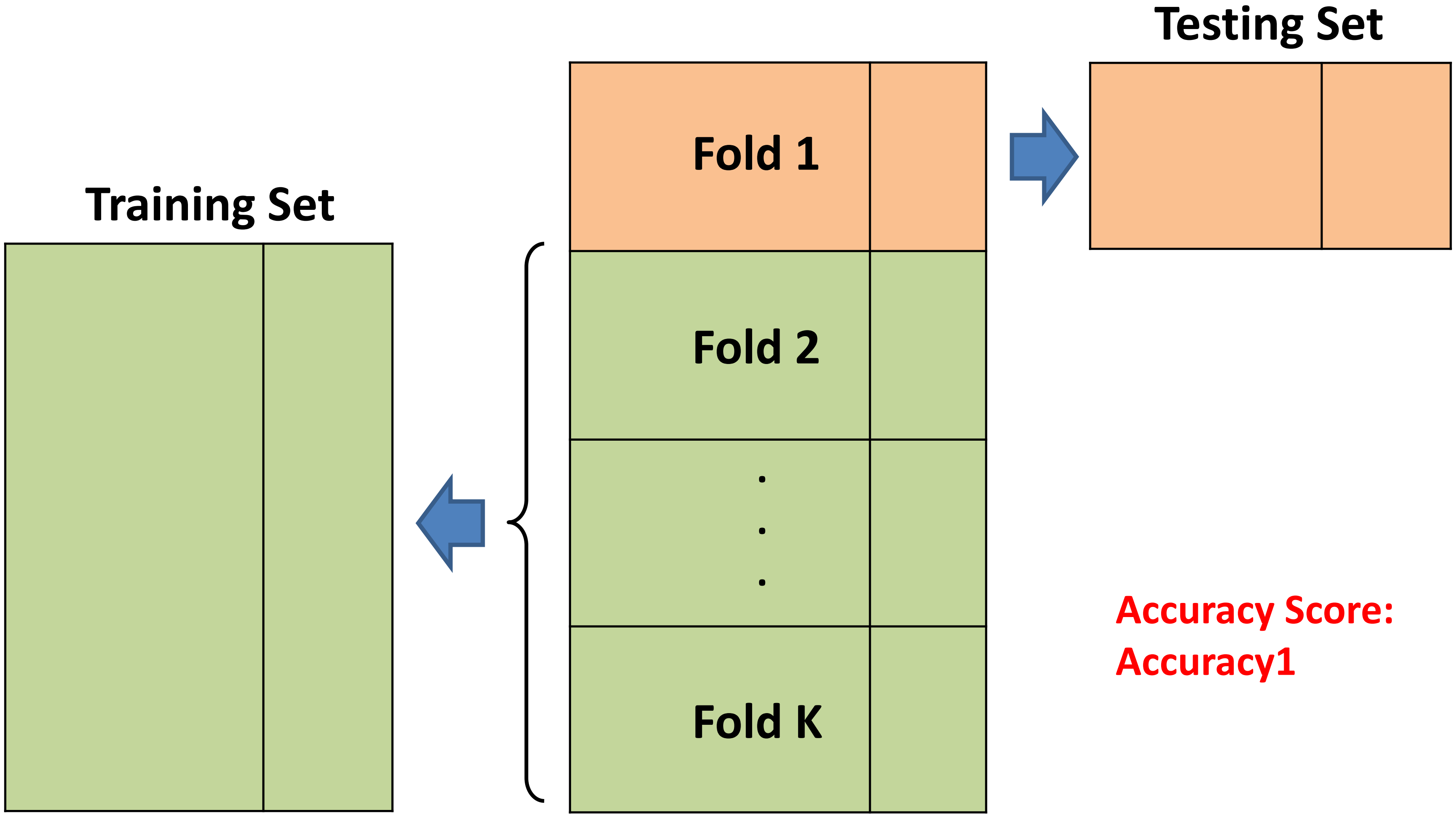
Fold 2

.
.
.

Fold K

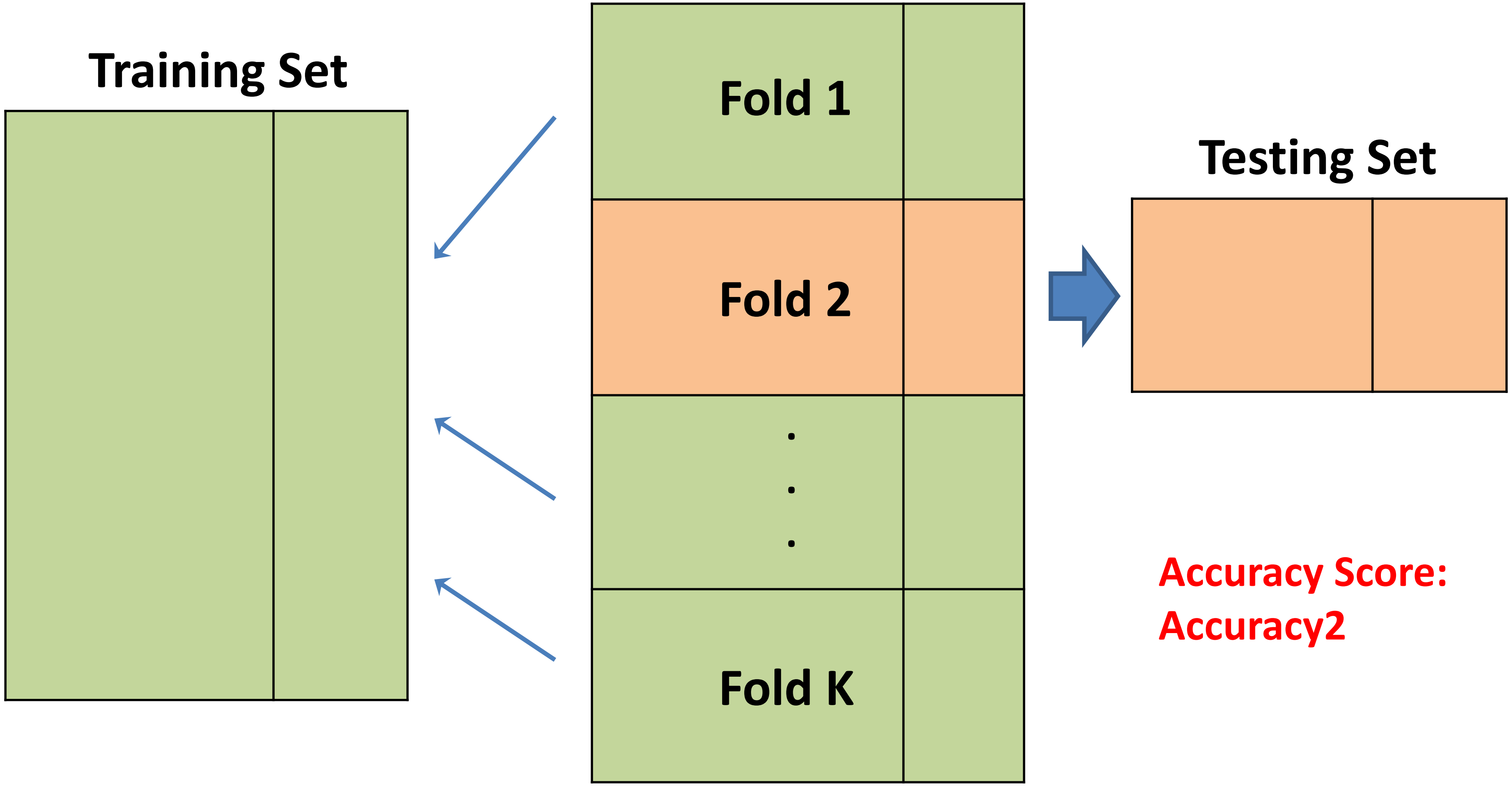
Original Dataset

Cross Validation – Round 1



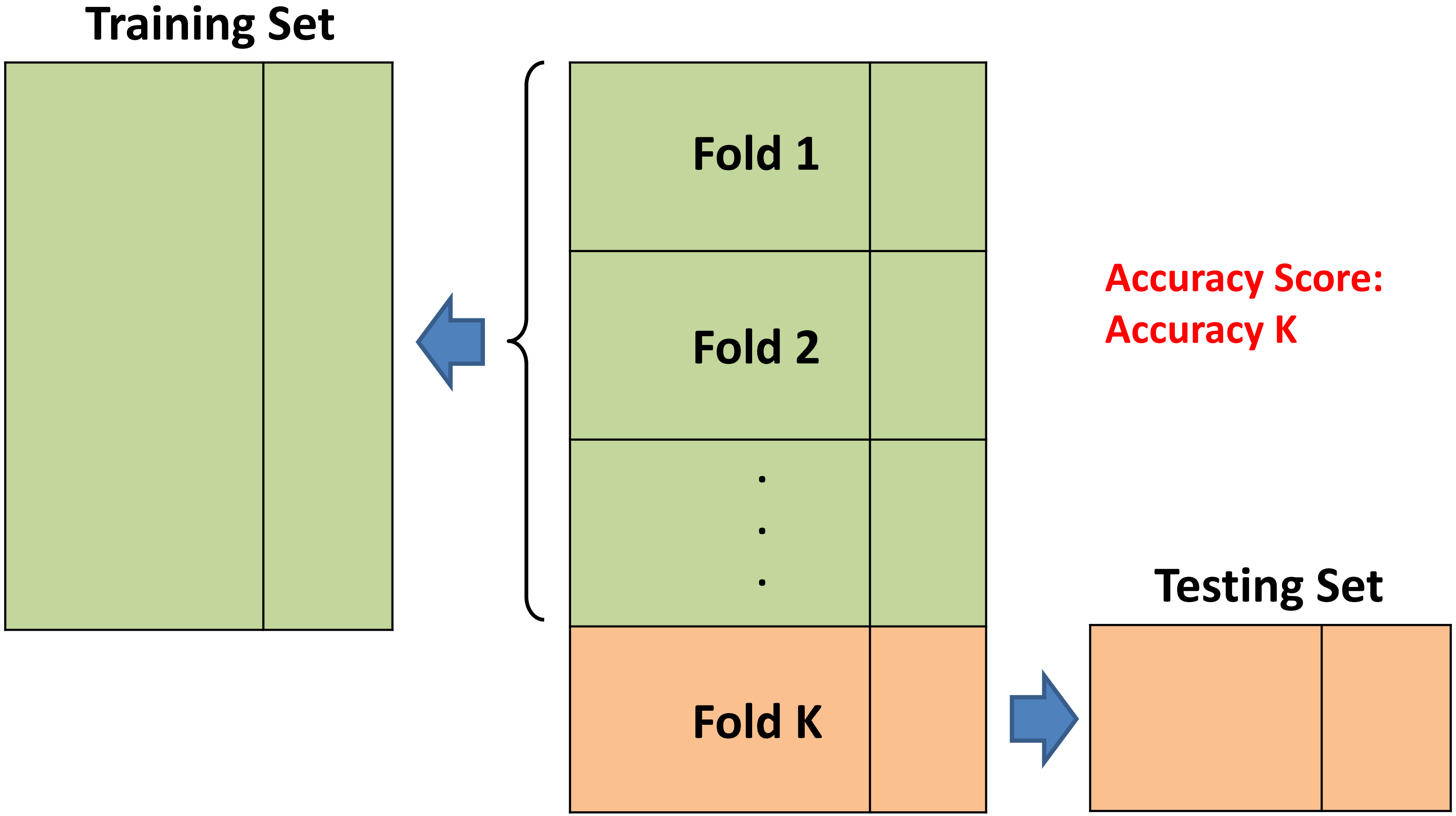
Accuracy Score:
Accuracy1

Cross Validation – Round 2



Accuracy Score:
Accuracy2

Cross Validation – Round K



Training Set

Fold 1

Fold 2

.
.
.

Fold K

Testing Set

Accuracy Score:
Accuracy K

Cross Validation

- Accuracy_Score_Total =

  (Accuracy 1 + Accuracy 2 + Accuracy 3 + … + Accuracy K) / K

Thank You!

Questions?