CROSS-VALIDATION I

- In this method, we make *k* random split of original data set. Where *k* is specified by the user.
- The model is trained on *k-1* split of data set whereas, remaining *k* split is used for evaluation.
- The performance of the model is recorded for this step. In the second step, the role of subsets formed as training and test in step one are now swapped.
- The k^{th} split which was earlier used as test set is not included in training subset and one random split from *k-1* behaves as a test set.
- Again the performance of the model is computed. This process is repeated *k* times so that each split is used once for training and once for testing.

CROSS-VALIDATION II

Cross-validation- Example Illustration

Original data set

sepal length	sepal width	petal length	petal width	Class
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
5.0	3.6	1.4	0.2	virginica
5.5	2.3	4.0	1.3	versicolor
6.5	2.8	4.6	1.5	virginica
5.7	2.8	4.5	1.3	virginica
6.2	1.8	4.7	2.5	versicolor
5.1	2.1	3.2	1.2	versicolor
5.5	3.4	6.7	2.3	virginica

Figure 43: Original data set

$K = 5$ random split of original data set

sepal length	sepal width	petal length	petal width	Class	
5.1	2.1	3.2	1.2	versicolor	Split 1
5.7	2.8	4.5	1.3	virginica	
4.7	3.2	1.3	0.2	setosa	Split 2
5.5	3.4	6.7	2.3	virginica	
6.5	2.8	4.6	1.5	virginica	Split 3
5.5	2.3	4.0	1.3	versicolor	
4.9	3.0	1.4	0.2	setosa	Split 4
6.2	1.8	4.7	2.5	versicolor	
5.1	3.5	1.4	0.2	setosa	Split 5
5.0	3.6	1.4	0.2	virginica	

(a) Example of $k = 5$ cross-validation

$K = 2$ random split of original data set

sepal length	sepal width	petal length	petal width	Class	
5.1	2.1	3.2	1.2	versicolor	Split 1
5.7	2.8	4.5	1.3	virginica	
4.7	3.2	1.3	0.2	setosa	
5.5	3.4	6.7	2.3	virginica	
6.5	2.8	4.6	1.5	virginica	
5.5	2.3	4.0	1.3	versicolor	Split 2
4.9	3.0	1.4	0.2	setosa	
6.2	1.8	4.7	2.5	versicolor	
5.1	3.5	1.4	0.2	setosa	
5.0	3.6	1.4	0.2	virginica	

(b) Example of $k = 2$ cross-validation

CROSS-VALIDATION III

Example Illustration: 2-fold Cross-validation

With $k = 2$, we have two subsets of data sets and number of iterations in which validation will stop is 2.

In Figure 45, iteration 1 of the process of cross-validation is illustrated.

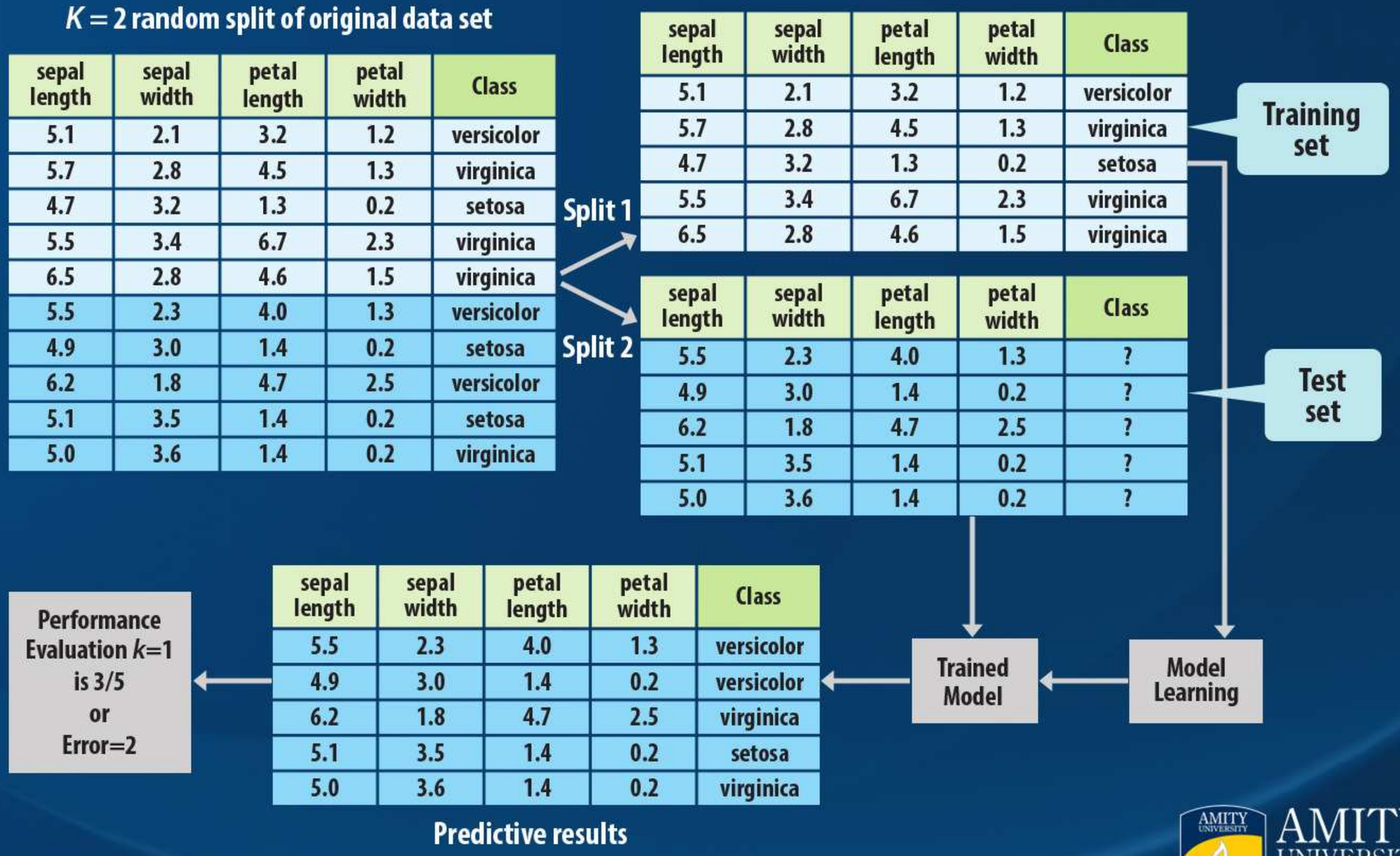

Figure 45: First iteration of 2-fold cross-validation

CROSS-VALIDATION IV

Example Illustration -2-fold Cross-validation

In Figure 46, iteration 2 of the process of Cross-validation is illustrated.

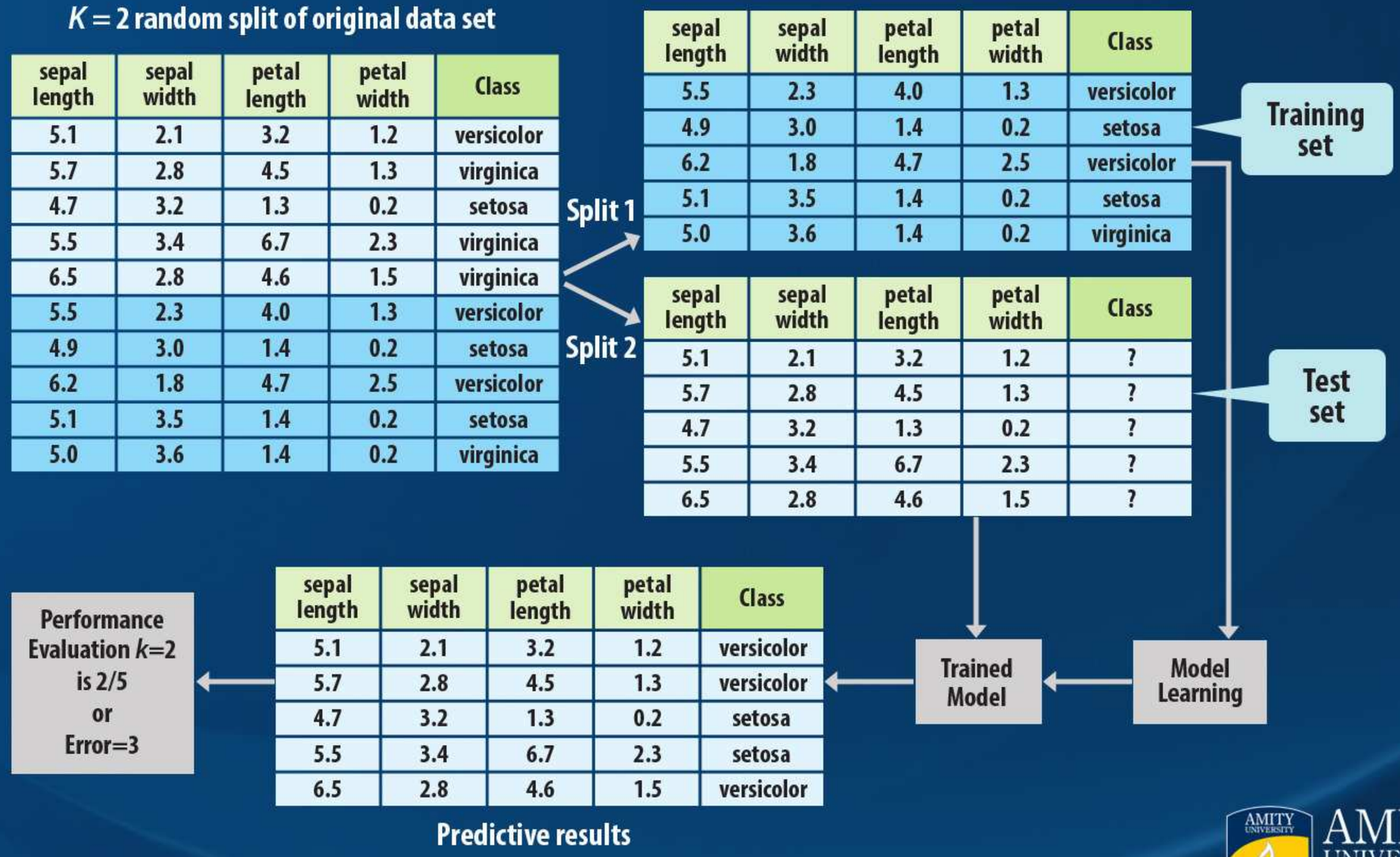

Figure 46: Second iteration of 2-fold Cross-validation

CROSS-VALIDATION V

Performance Evaluation-2-fold Cross-validation

In order to calculate the performance of the model, the performances at $k = 1$ and $k = 2$ are averaged.

Performance Evaluation K=1 Is 3/5 or Error=2	+	Performance Evaluation K=2 Is 2/5 or Error=3	=	Performance = 50% Average Error - 0.5

Figure 47: Performance evaluation of 2-fold Cross-validation

CROSS-VALIDATION VI

- **The key characteristics of Cross-validation method are:**

1. Cross-validation is an iterative process that involves *k* steps, where *k* is specified by the user.
2. The division of data records in *k* splits are random.
3. In each run of method, *k-1* split is used for training and k^{th} split is used for test.
4. The split used for testing or model validation is always unsupervised data, i.e., class labels are not provided or hidden from to the model.
5. The overall performance is an average of performances at each run of Cross-validation.
6. No test case is common between each split.

CROSS-VALIDATION VII

- **Strengths and Weakness of Cross-validation**
  - **Strengths**
    1. It is the best model evaluation to be used particularly for small data set.
    2. In this method since every data point once behave as a training sample and once as test sample, the overall performance estimate is less sensitive to the partitioning of the data.
  - **Weakness**
    1. Computationally expensive especially for the cases where number of splits are low than observations in the data set.

BOOTSTRAP I

- **Bootstrap**
  - In Hold-out and Cross-validations methods training and test sets are mutually exclusive in nature. That is the two sets prepared for learning and testing contains no overlapping observations.
  - However, in Bootstrap, the training dataset is randomly selected with replacement. The remaining examples that were not selected for training are used for testing.
  - Refer Figure 48 for illustration. The overall performance of the model is the average of the performance of each iteration.

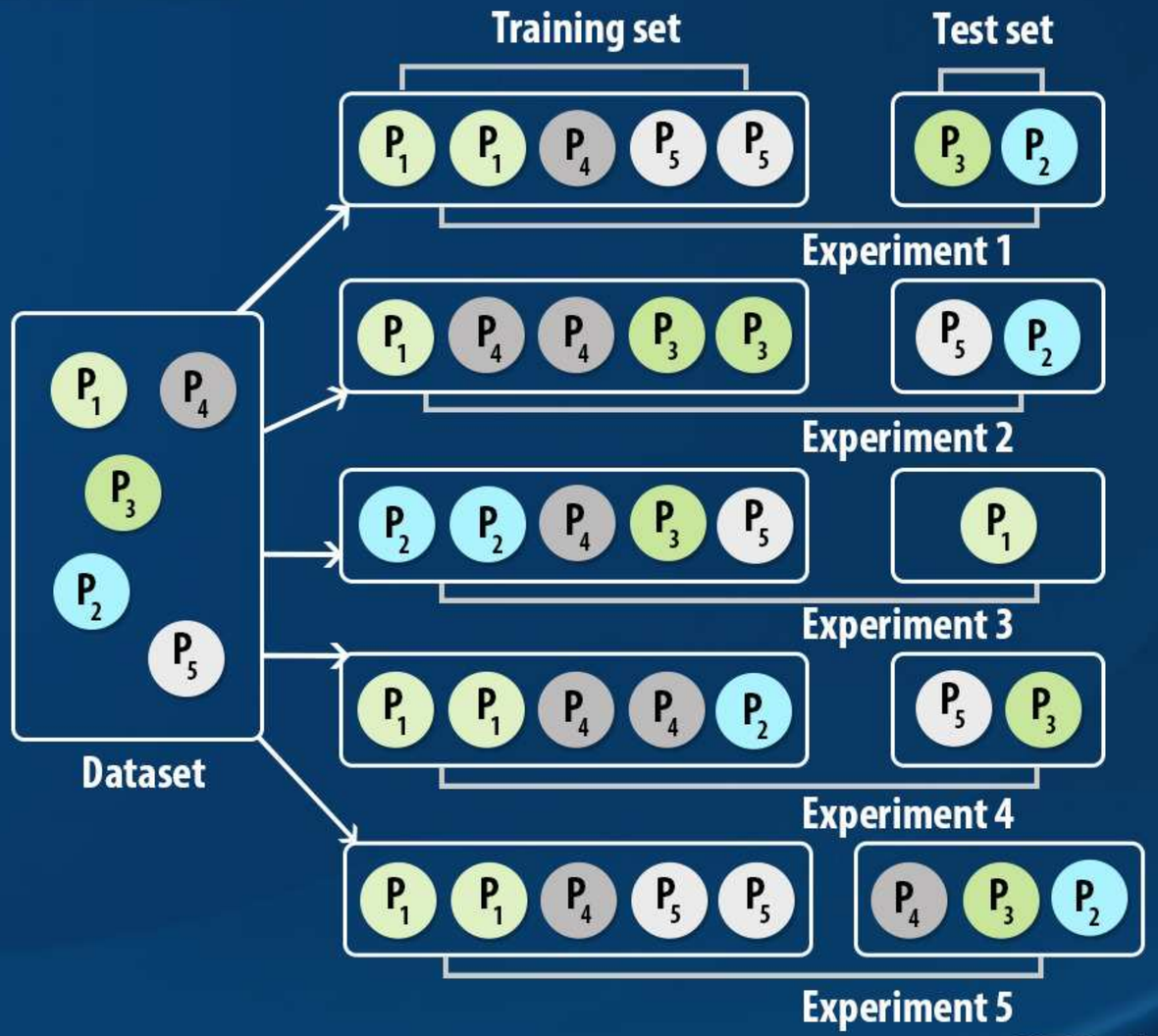

Figure 48: Illustration of Bootstrapping on hypothetical data set with 5 observations

BOOTSTRAP II

- **Strengths and Weaknesses of Bootstrap**
  - **Strengths**
    1. It can develop more robust, stable and consistent models.
    2. It is very useful in case of small data sets.
    3. Bootstrapping can be used to overcome "overfitting".
  - **Weakness**
    1. Sensitive to noisy data present in the data set.