Introduction to Data Science

(Lecture 8)

Dr. Mohammad Pourhomayoun

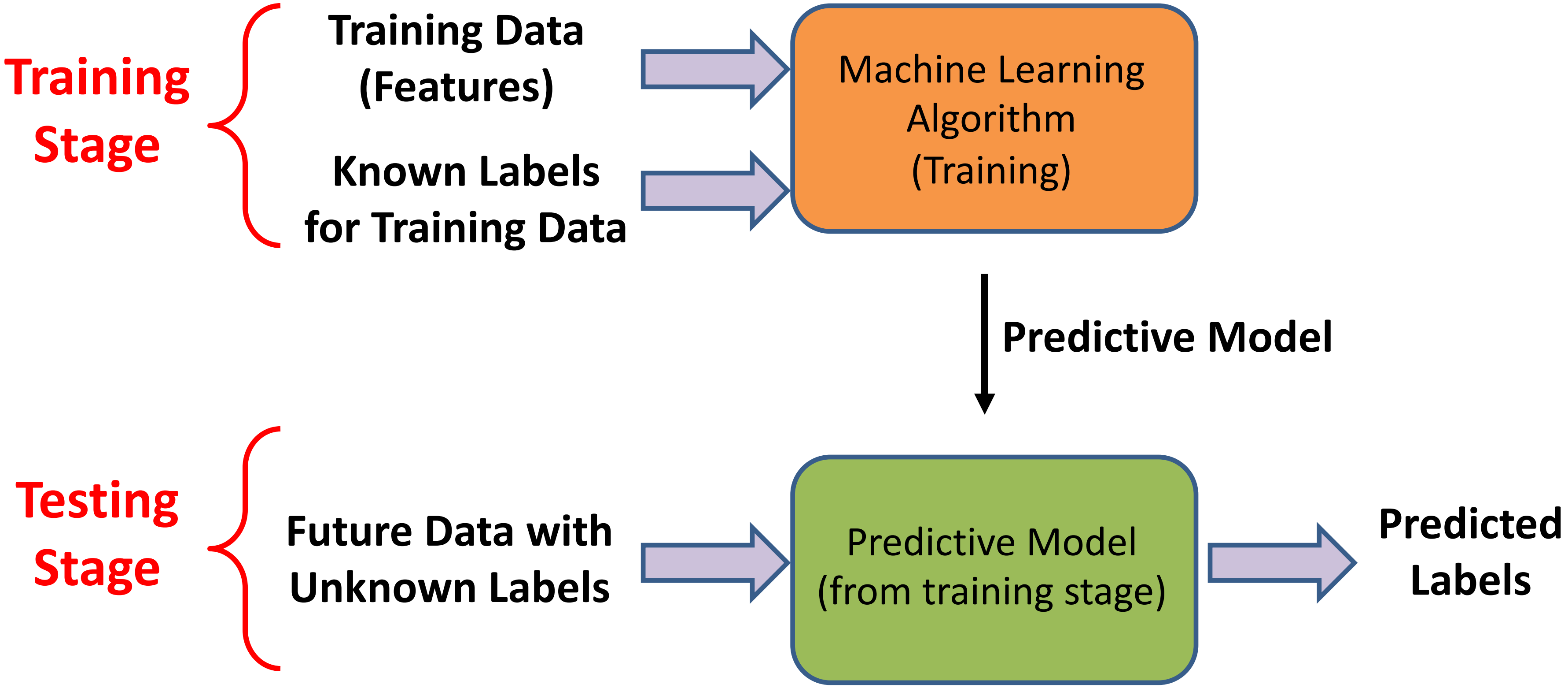
Associate Professor

Computer Science Department

California State University, Los Angeles

Evaluating The Accuracy Of Our Predictive Model

Review: Supervised Learning: Learning from labeled Data


Training Stage
Training Data (Features)
Known Labels for Training Data
Machine Learning Algorithm (Training)
Predictive Model
Testing Stage
Future Data with Unknown Labels
Predictive Model (from training stage)
Predicted Labels

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).

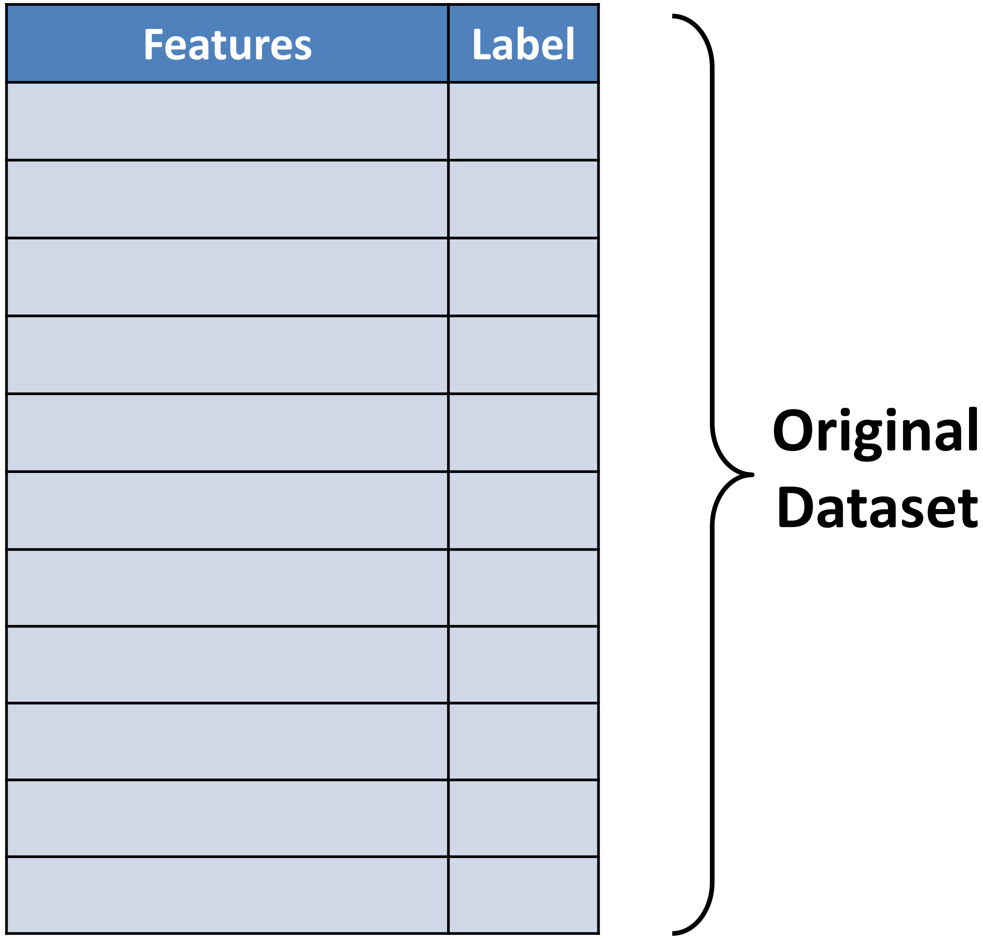
2- Let's **pretend** that we do **NOT** know the label of the Testing Set!

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!

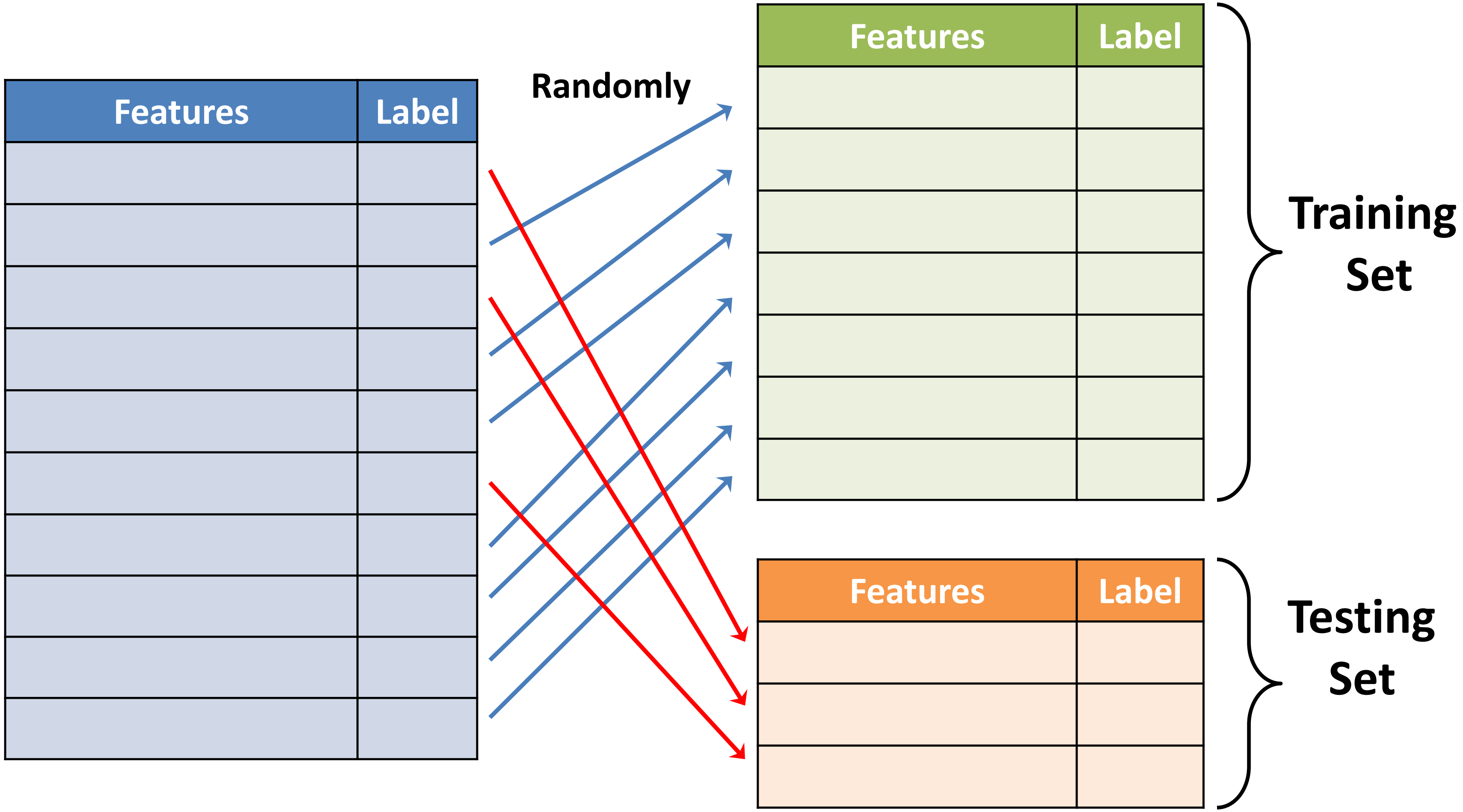
We will learn more techniques for model evaluation (e.g. **Cross Validation** method) later in this class!

Training and Testing Sets



Features	Label

Training and Testing Sets

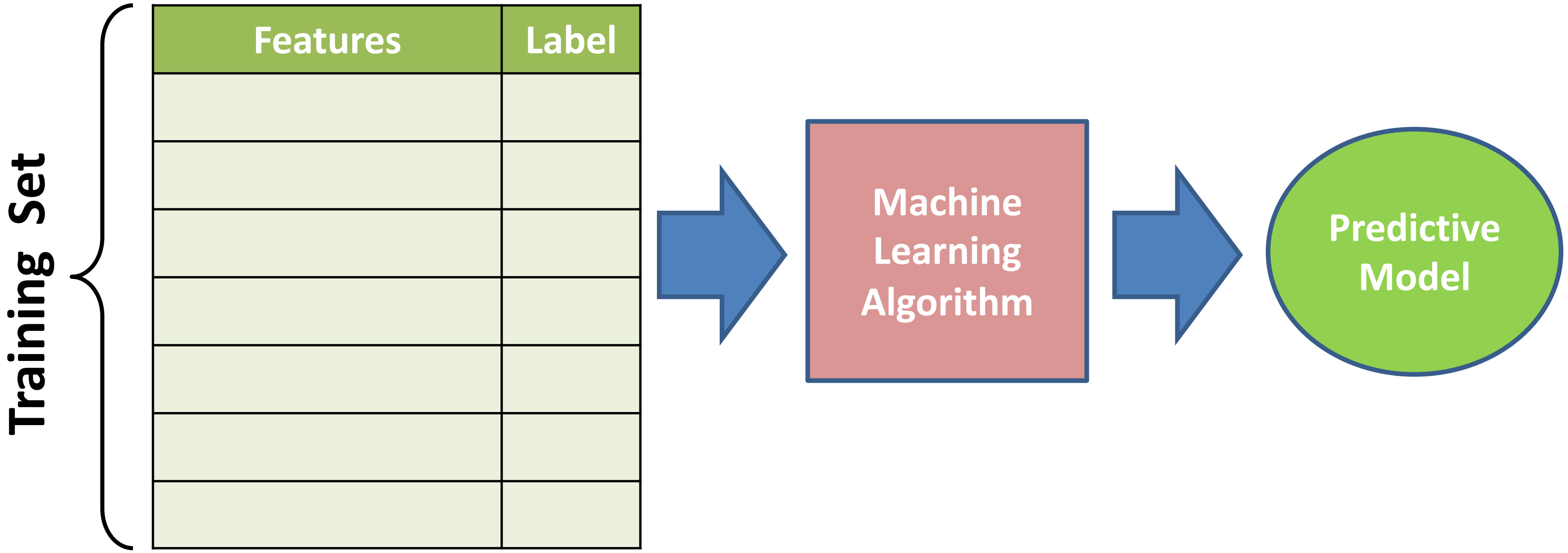


Training Stage

Training Set

Features	Label

Training Stage



Training Set

Features

Label

Machine Learning Algorithm

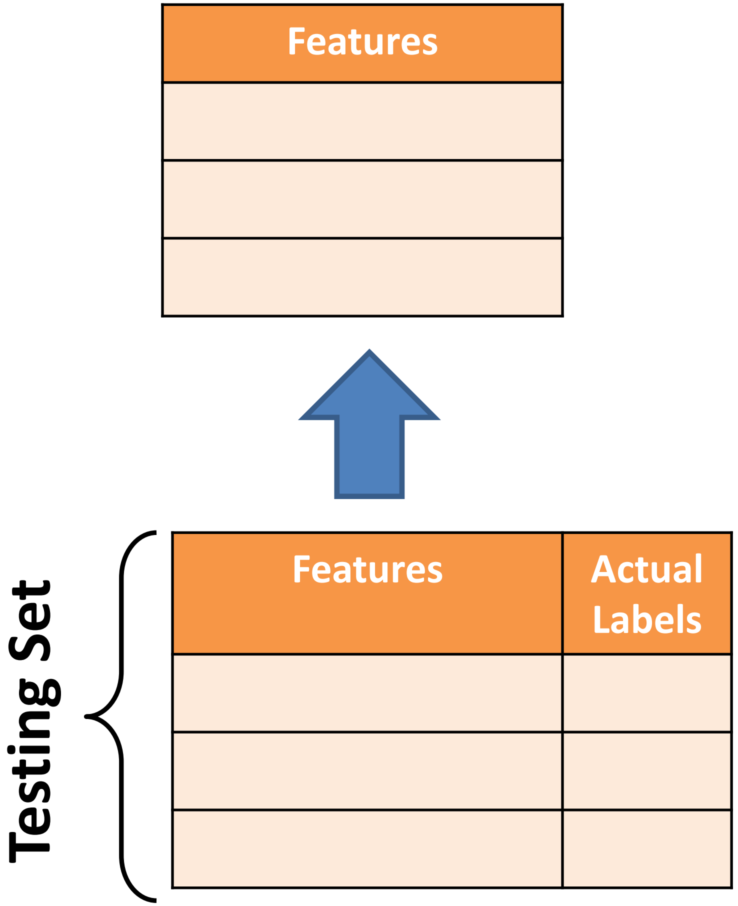
Predictive Model

Testing Stage

Testing Set

Features	Actual Labels

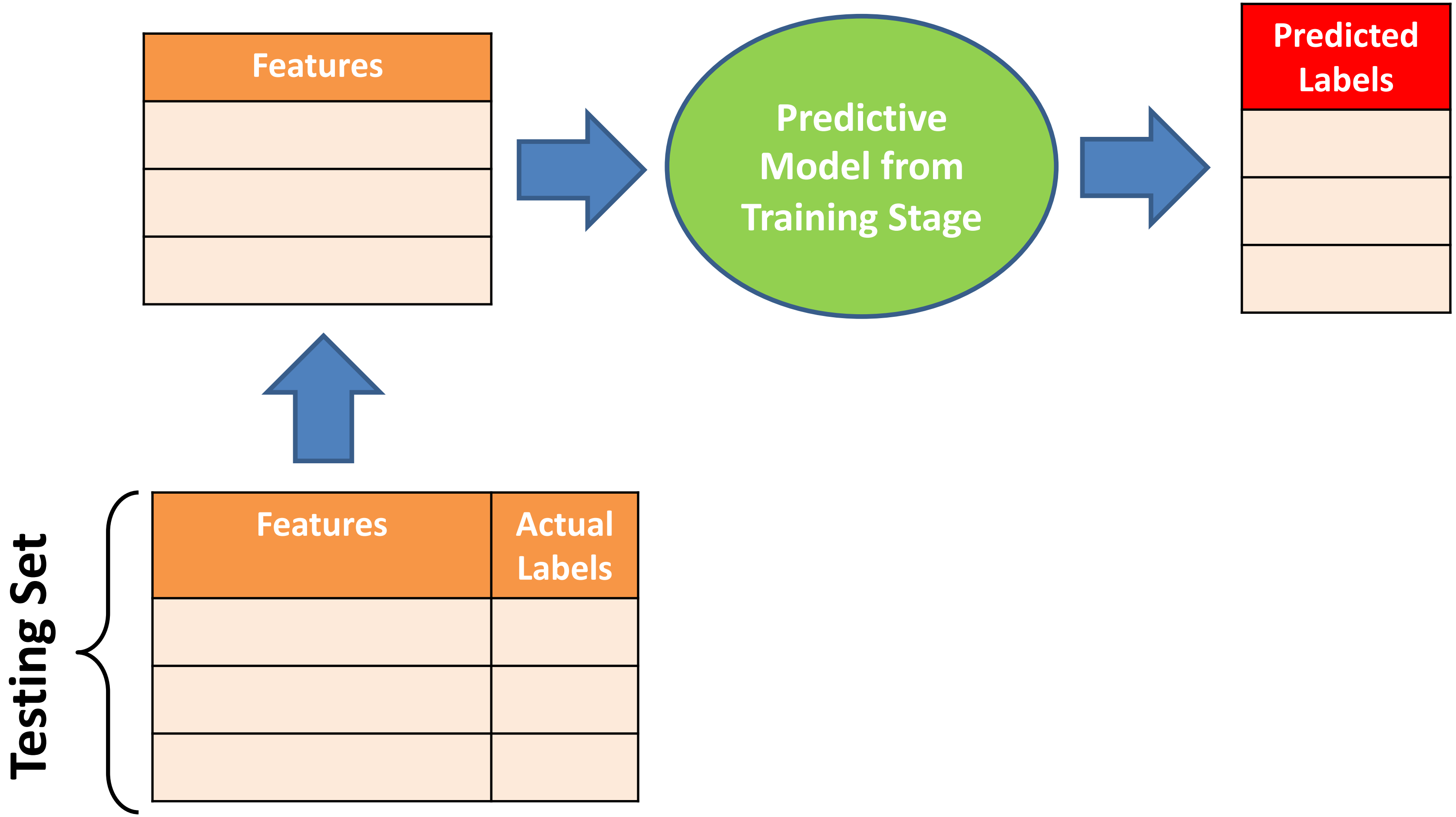
Testing Stage



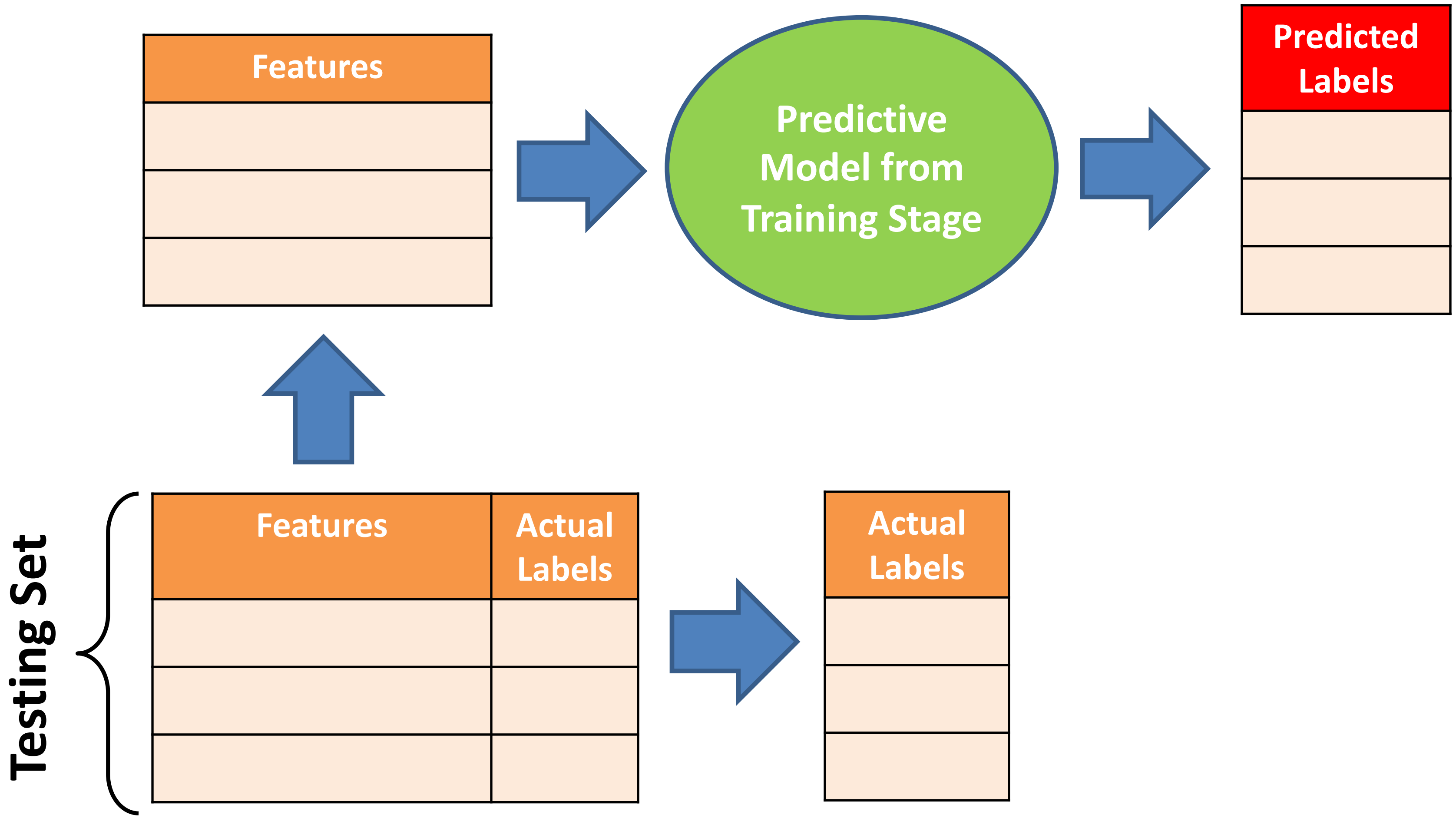
Features

Features	Actual Labels

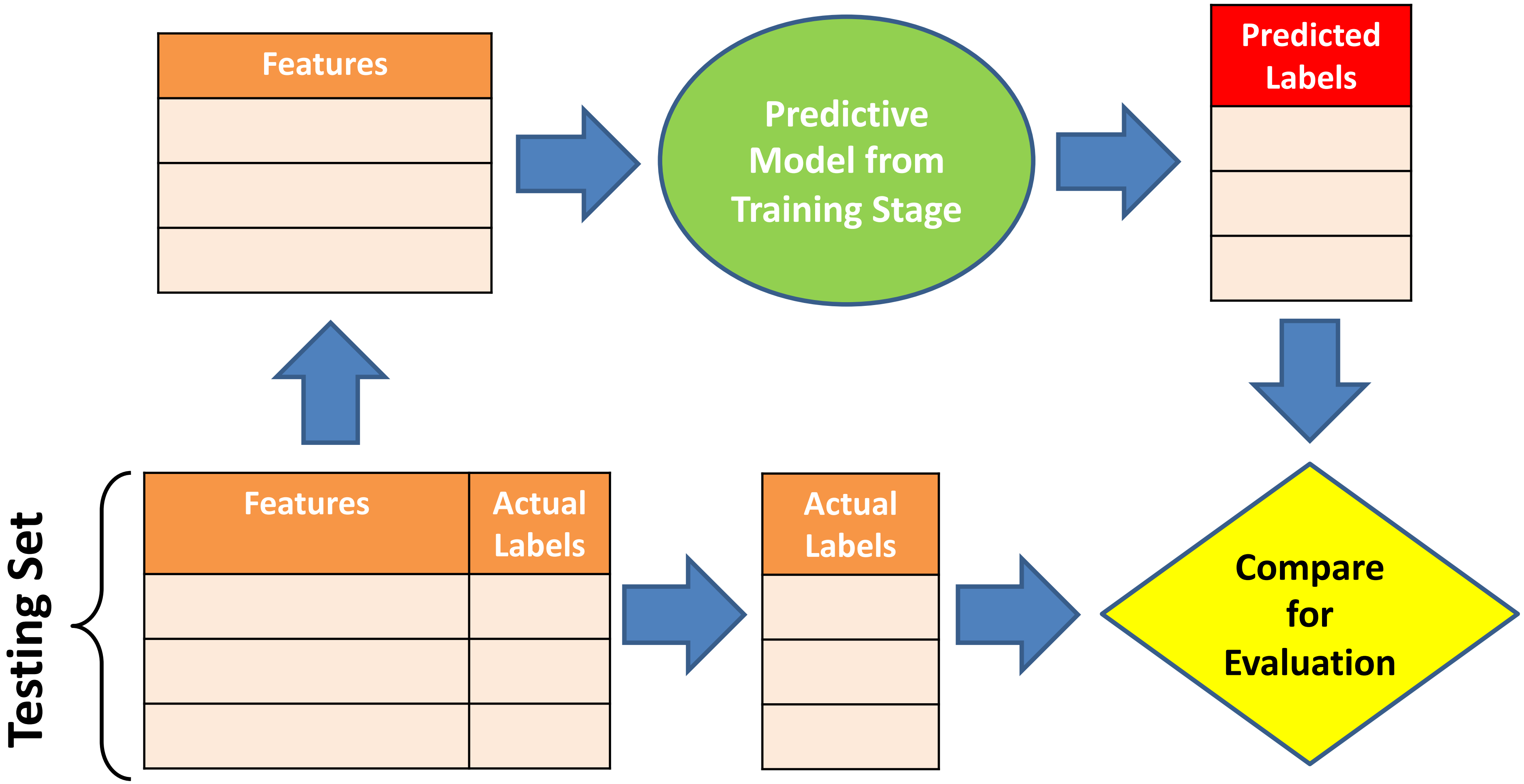
Testing Stage



Testing Stage



Testing Stage



Evaluating The Accuracy Of Our Predictive Model

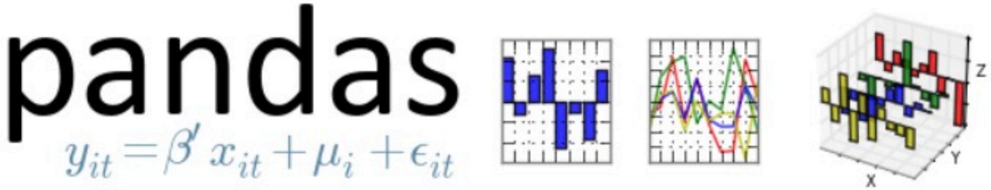
VERY IMPORTANT: **There must be NO OVERLAP between Training Set and Testing Set!**

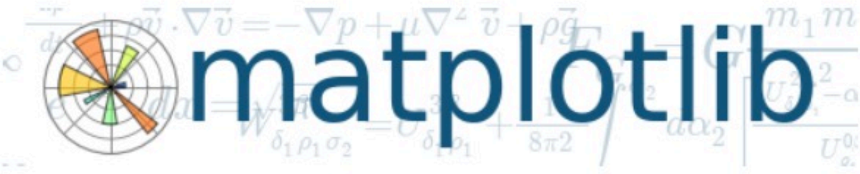
Evaluating The Accuracy Of Our Predictive Model

- **Note1:** Later, we will see that we can split the original dataset into 3 sets: **Training Set, Validation Set ,** and **Testing Set**. In this case, We can use Validation set for adjusting the classifier parameters, and then use Testing Set for final evaluation.

- **Note2:** Later, we will also talk about **Cross-Validation** approach. In Cross-Validation, several rounds of partitioning will be applied to assure that all data samples are used both in training set and testing set but not simultaneously (NO OVERLAP!)

Data Science with Python





Scikit-Learn:
A Library for Data Science and Machine Learning

Scikit-Learn (sklearn)

- Scikit-learn (aka sklearn) is the Python Machine Learning Library.
- It includes optimal implementation of various **classification, regression** and **clustering** algorithms.
- It also includes hundreds of commands and functions for data preprocessing and processing along with a number of **default datasets** to work with.
- It is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
- Scikit-learn has an exceptional documentation.

IRIS Dataset

- **Recognizing flowers**
  - 150 sample flowers in three species (50 each).
  - Species of Iris (Labels): setosa, versicolor, virginica
  - Features: sepal length, sepal width, petal length, petal width

Important Hint about sklearn

- Sklearn only accept **NUMERICAL** **features**. Thus, we have to convert non-numerical (categorical) features into numerical values.
- **Note**: In converting features (and sometimes labels), we have to be cautious to avoid defining a confusing "ordering" between categorical values (we will talk about it later in this course).
- Depending on the classification algorithm, We usually use **LabelEncoding** to convert labels, and/or **OneHotCoding** to convert features.

6 Steps To Make Prediction In sklearn

- **Step1:** Importing the sklearn class (the machine learning algorithm) that you would like to use for prediction FROM sklearn library.
- **Step2:** Set up the Feature Matrix and Label Vector.
- **Step3:** Defining (instantiating) an "object" (instance) of the sklearn class as an initial predictive object.
- **Step4:** Training Stage: Train the above predictive model using the training dataset.
- **Step5:** Testing (Prediction) Stage: Making prediction on new observations (Testing Data) using the trained model.
- **Step6:** Evaluating the machine learning model and results

Data Science Practical Tutorial

- Let’s open file ***CS4661-PythonDataScienceTutorial-Lab3.ipynb***
  in Jupyter notebook to continue the tutorial.