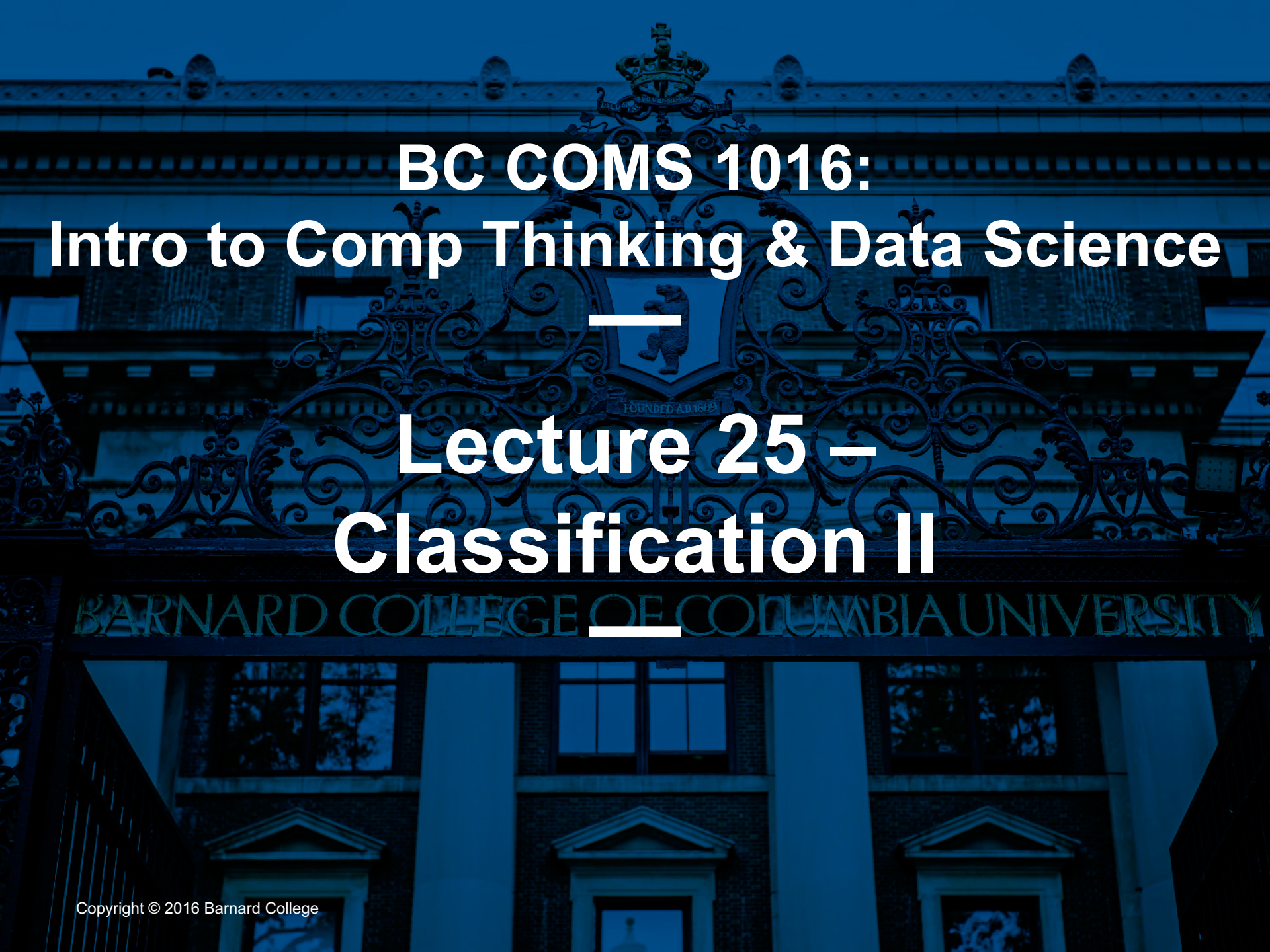


BC COMS 1016: Intro to Comp Thinking & Data Science

Lecture 25 – Classification II

Announcements

- Homework 10 – Classification
  - Due Monday 05/02
- Project 3:
  - Due Monday 05/02
- No lab this week
- Project 2:
  - Great results!
  - **Mean: 31.22, max 34**
  - **STD DEV: 2.86**
- Course Evaluations:
  - Released Monday, due two weeks after
- Thursday's content: choose your own adventure

Final Project

Final Project – Real World Data Science

- Explore a real world dataset from multiple tables
  - Choose from 8 datasets
- Ask 2 questions that the dataset can help answer
  - Hypothesis Testing
  - Prediction
- Use methods covered in in the class to answer these questions

Final Project

We will provide:

1. An overview and description of the dataset
2. A preview section with code to read in all the datasets relevant to your specific project
3. A Research Report section which contains the outline for the content of your final project.

Final Project - Report Section

1. Introduction:

   250-300 word background

2. Hypothesis Testing and Prediction Questions

   State the questions and how you plan to answer them

3. Exploratory Data Analysis
   1. Visualize!

4. Hypothesis Testing
5. Prediction
6. Conclusion

Final Project - Proposal Deadline – Thursday 05/05

1. Introduction:

   250-300 word background

2. Hypothesis Testing and Prediction Questions

   State the questions and how you plan to answer them

---

3. Exploratory Data Analysis
   1. Visualize!
4. Hypothesis Testing
5. Prediction
6. Conclusion

The earlier you submit the proposal the better so we can give you more feedback

Classification

Machine Learning Algorithm

- A mathematical model
- calculated based on sample data ("training data")

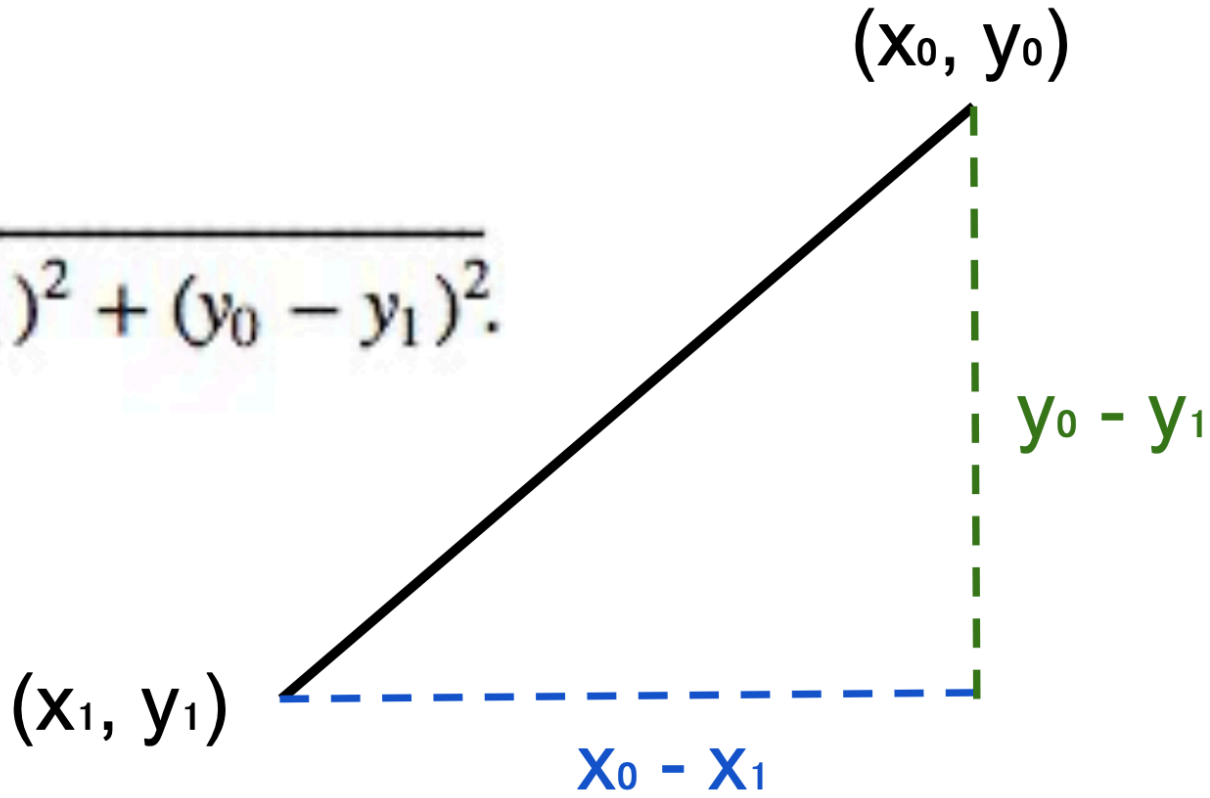
- that makes predictions or decisions without being explicitly programmed to perform the task

Classifiers

Nearest Neighbor Classification

Pythagoras' Formula

$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}.$$



Distance Between Two Points

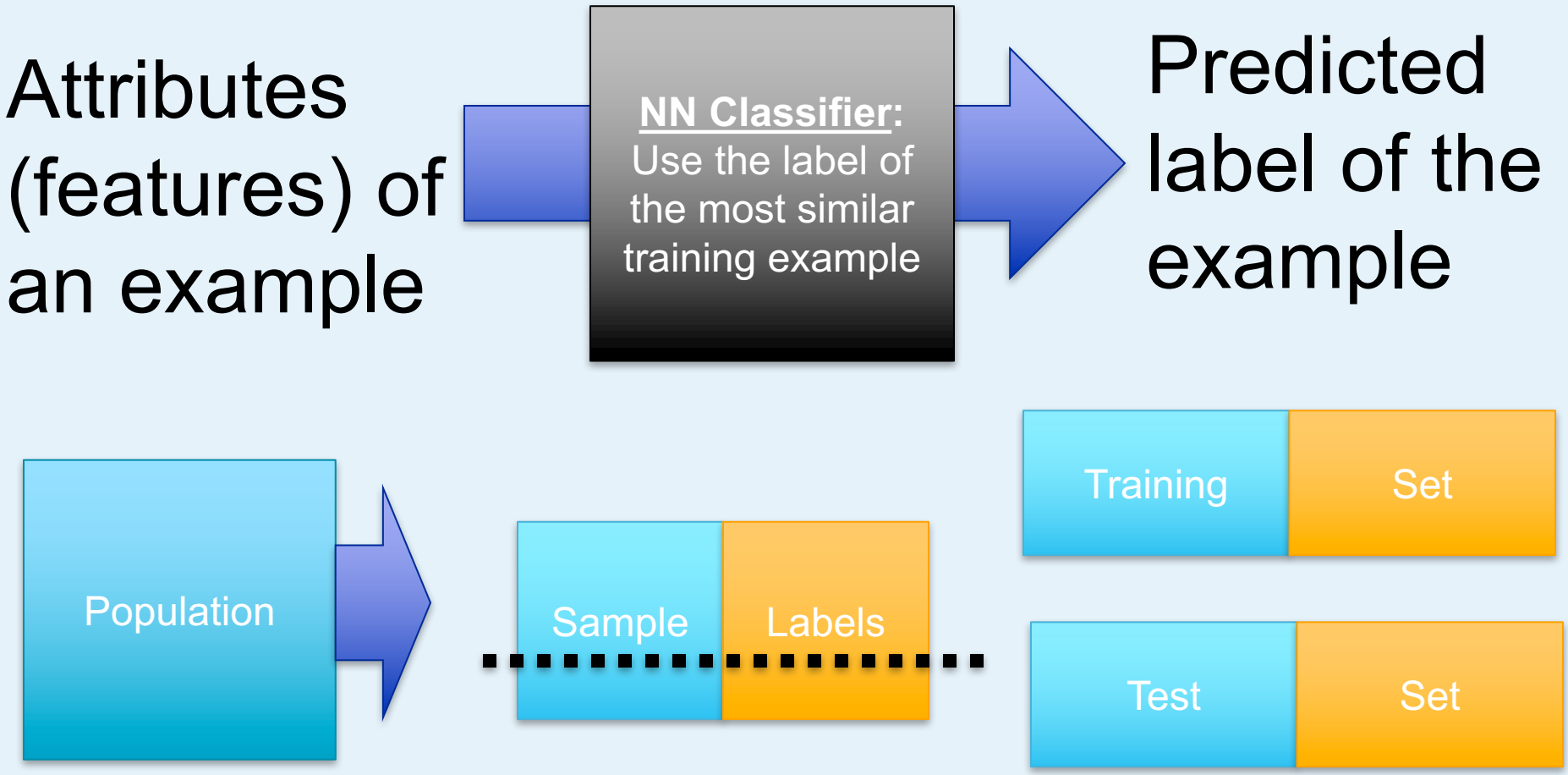
- Two attributes x and y:

$$D = \sqrt{((x_0 - x_1)^2 + (y_0 - y_1)^2)}$$

- Three attributes x, y, and z:
  - $D = \sqrt{((x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2)}$

Nearest Neighbors Classification

Nearest Neighbor Classifier

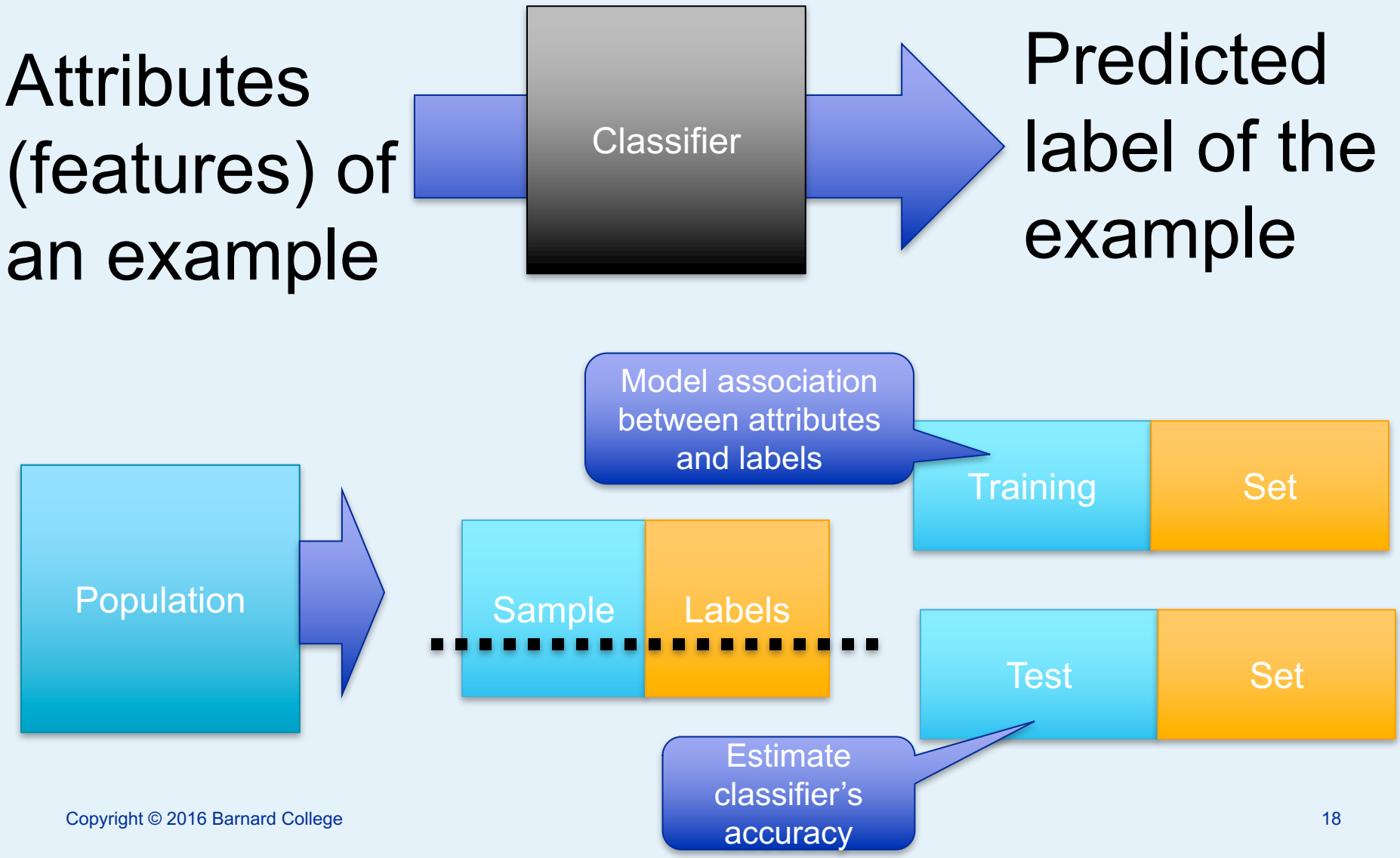


Finding the *k* nearest neighbors

1. Find the distance between the example and each example in the training set
2. Augment the training data table with a column containing all the distances
3. Sort the augmented table in increasing order of the distances
4. Take the top *k* rows of the sorted table

Evaluation

Training a Classifier



Accuracy of a Classifier

The accuracy of a classifier on a labeled data set is the proportion of examples that are labeled correctly

Need to compare classifier predictions to true labels

If the labeled data set is sampled at random from a population, then we can infer accuracy on that population