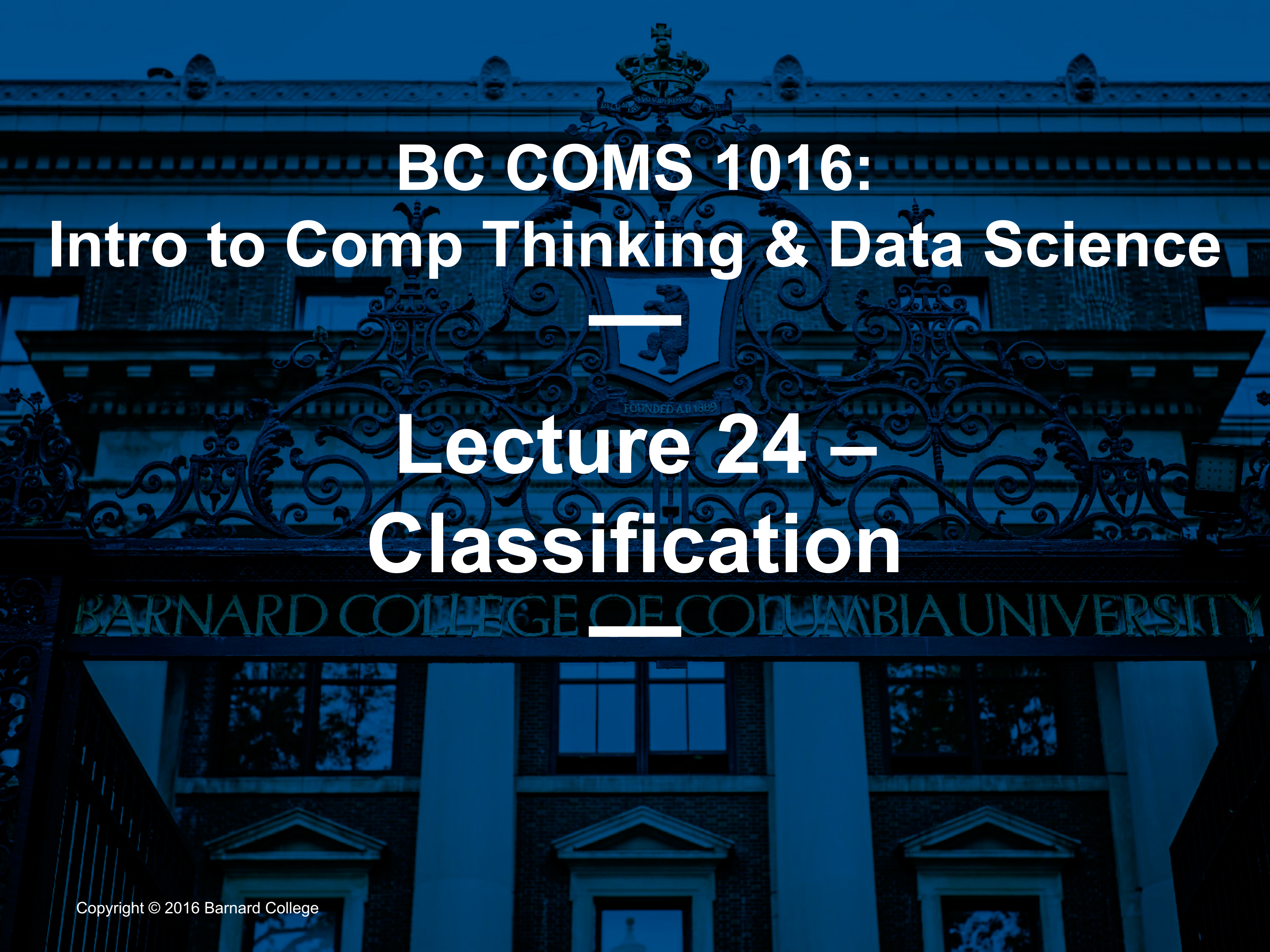


BC COMS 1016:
Intro to Comp Thinking & Data Science

—

Lecture 24 –
Classification

—

Announcements

- Homework 9 - Regression Inference
  - Due Monday 04/25
- Homework 10 – Classification
  - Due Monday 05/02
- Course Evaluations:
  - Due 1216
- Project 3:
  - Due Monday 05/02

Prediction Variability

Regression Prediction

- If the data come from the regression model,
- And if the sample is large, then:

- The regression line is close to the true line
- Given a new value of *x*, predict *y* by finding the point on the regression line at that *x*

Confidence Interval for Prediction

- **Bootstrap the scatter plot**
- **Get a prediction for *y* using the regression line that goes through the resampled plot**
- Repeat the two steps above many times
- Draw the empirical histogram of all the predictions.
- Get the “middle 95%” interval.
- That’s an approximate 95% confidence interval for the height of the true line at *y*.

Predictions at Different Values of *x*

- Since *y* is correlated with *x*, the predicted values of *y* depend on the value of *x*.

- The width of the prediction's CI also depends on *x*.
  - Typically, intervals are wider for values of *x* that are further away from the mean of *x*.

Inference about the True Slope

Confidence Interval for True Slope

- **Bootstrap the scatter plot.**
- **Find the slope of the regression line through the bootstrapped plot.**
- Repeat.
- Draw the empirical histogram of all the generated slopes.
- Get the “middle 95%” interval.
- That’s an approximate 95% confidence interval for the slope of the true line.

Test Whether There Really is a Slope

- **Null hypothesis:** The slope of the true line is 0.
- **Alternative hypothesis:** No, it’s not.
- Method:
    - Construct a bootstrap confidence interval for the true slope.
    - If the interval doesn’t contain 0, the data are more consistent with the alternative
    - If the interval does contain 0, the data are more consistent with the null

Classification

Guessing the Value of an Attribute

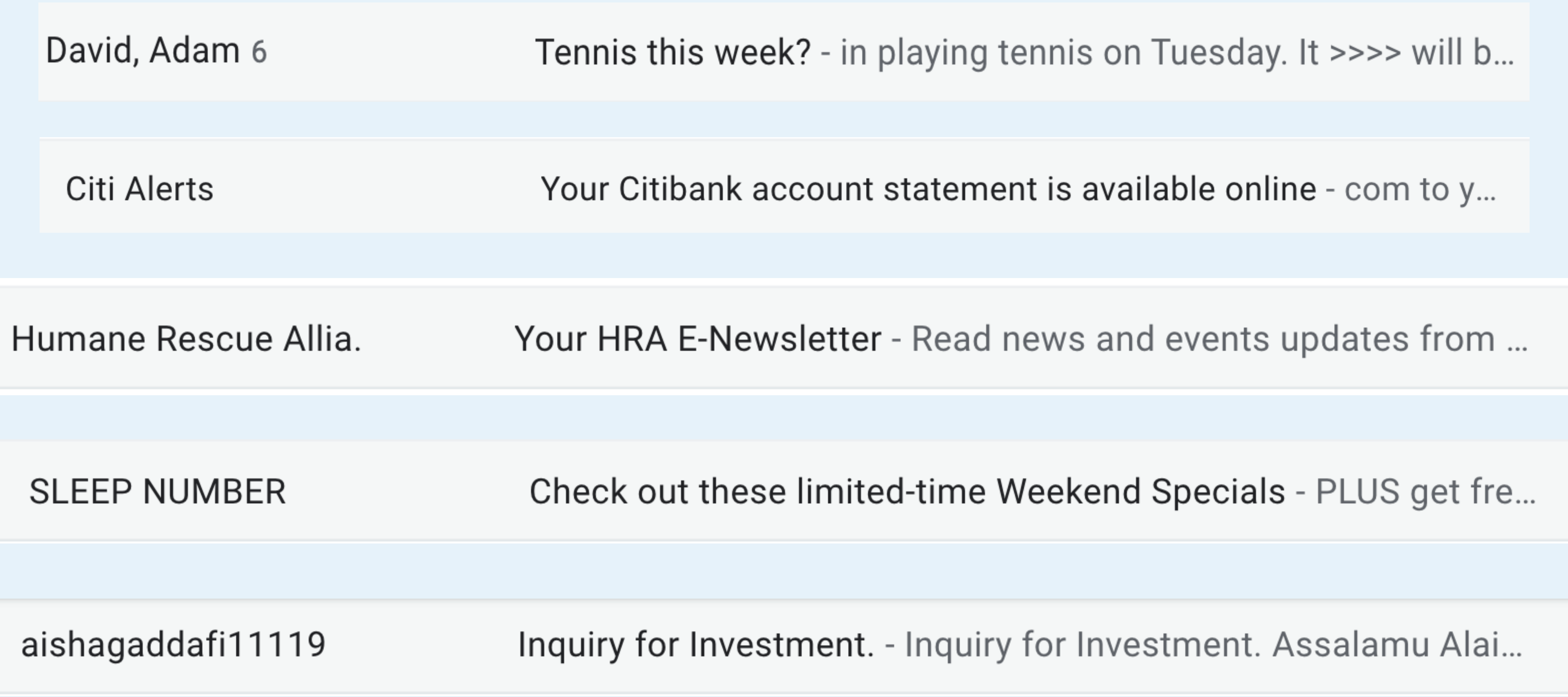
- Based on incomplete information

- One way of making predictions:
  - To predict an outcome for an individual,
  - find others who are like that individual
  - and whose outcomes you know.
  - Use those outcomes as the basis of your prediction.

- Two Types of Prediction
  - Classification = Categorical; Regression = Numeric

Prediction Example: Hot dog or not Hot dog?

Prediction Example: Spam or Not Spam?



David, Adam 6 Tennis this week? - in playing tennis on Tuesday. It >>>> will b...

Citi Alerts Your Citibank account statement is available online - com to y...

Humane Rescue Allia. Your HRA E-Newsletter - Read news and events updates from ...

SLEEP NUMBER Check out these limited-time Weekend Specials - PLUS get fre...

aishagaddafi11119 Inquiry for Investment. - Inquiry for Investment. Assalamu Alai...

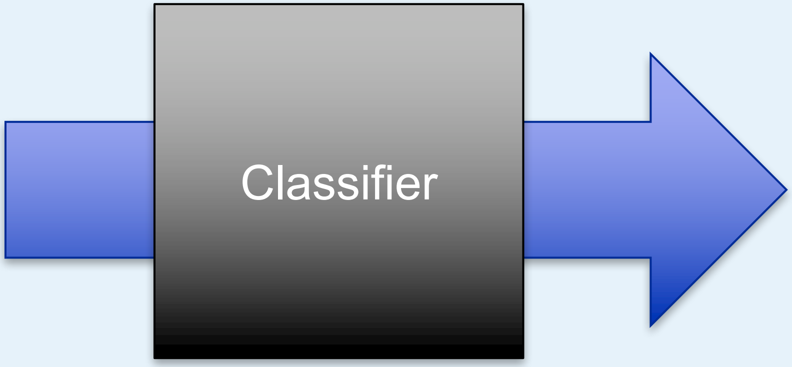
Machine Learning Algorithm

- A mathematical model
- calculated based on sample data ("training data")
- that makes predictions or decisions without being explicitly programmed to perform the task

Classifiers

A Classifier

Attributes (features) of an example



Classifier

Predicted label of the example

Rows

Rows of a Table

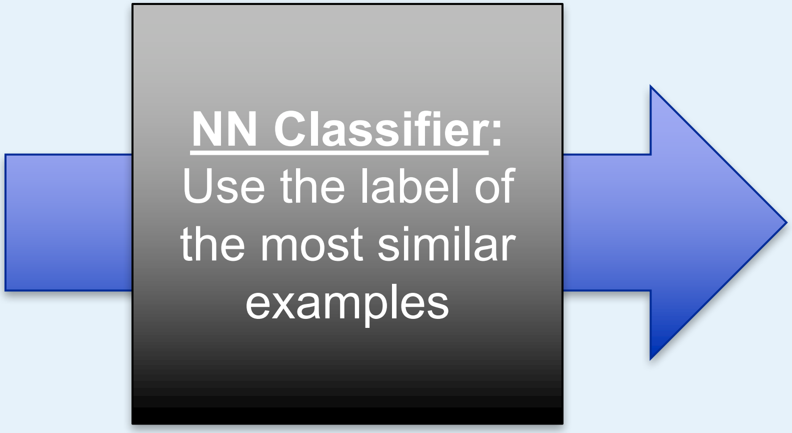
Each row contains all the data for one individual

- **t.row(i)** evaluates to **i**th row of table **t**
- **t.row(i).item(j)** is the value of column **j** in row **i**
- If all values are numbers, then **np.array(t.row(i))** evaluates to an array of all the numbers in the row.
- To consider each row individually, use

```
for row in t.rows:
    ... row.item(j) ...
```

- **t.exclude(i)** evaluates to the table **t** without its **i**th row

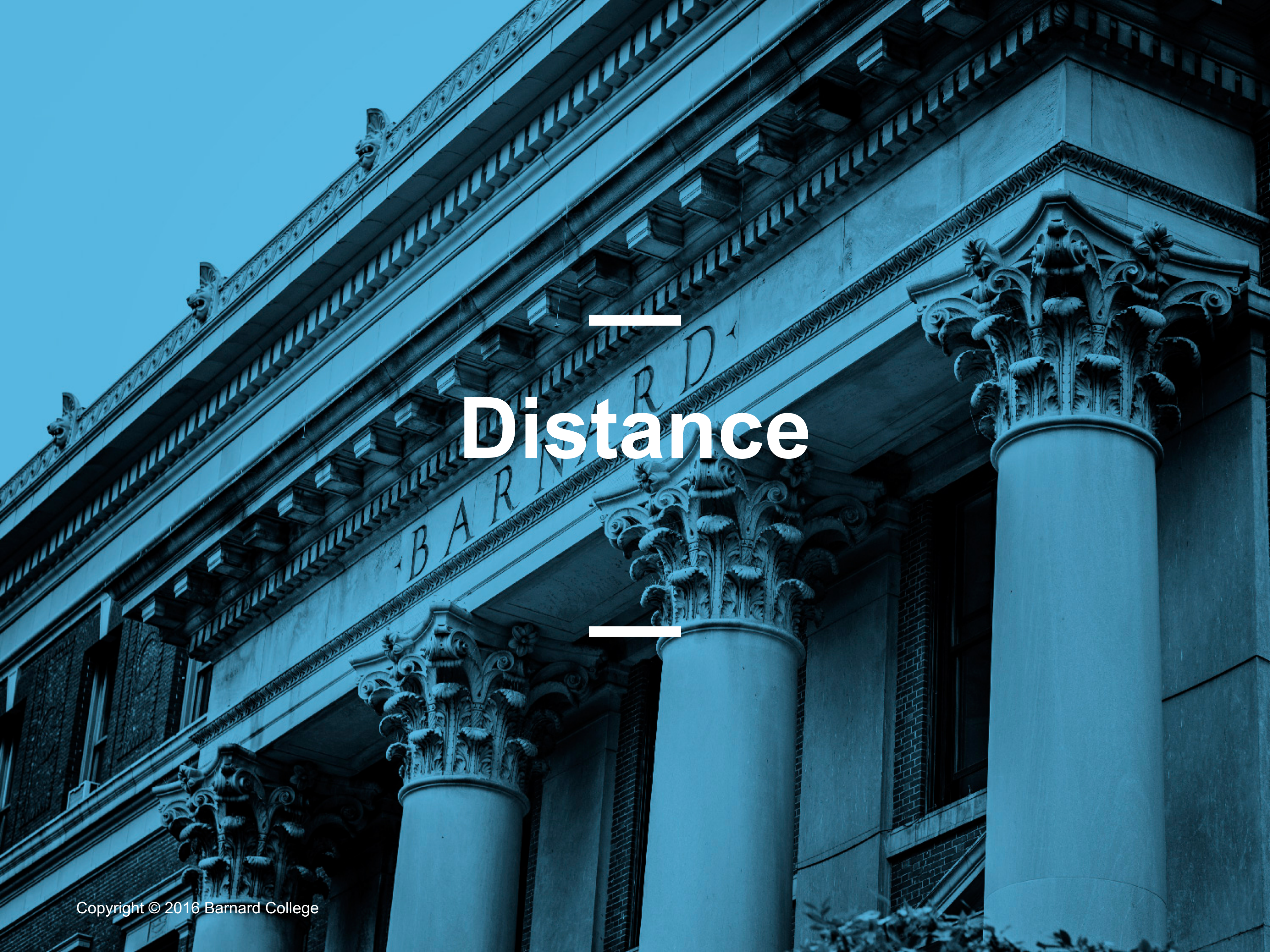
A Classifier



Attributes (features) of an example

NN Classifier: Use the label of the most similar examples

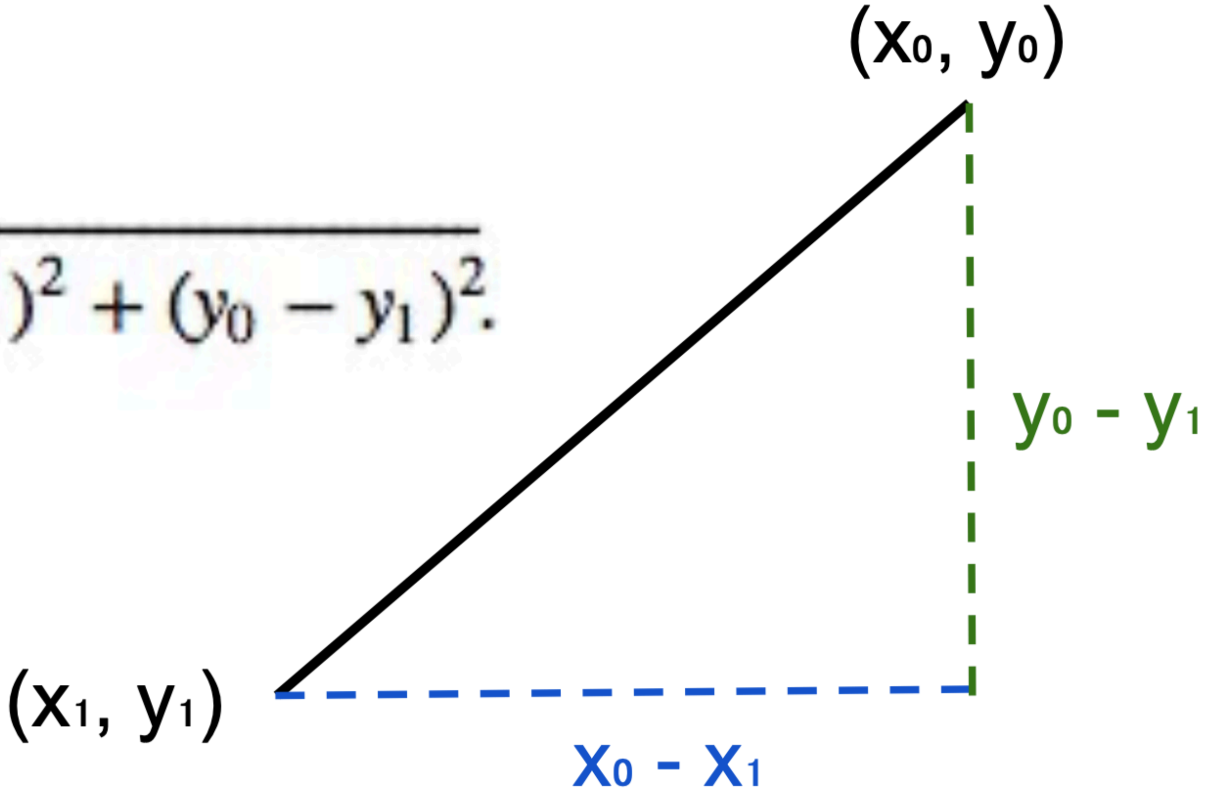
Predicted label of the example



Distance

Pythagoras’ Formula

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



Distance Between Two Points

- Two attributes x and y:

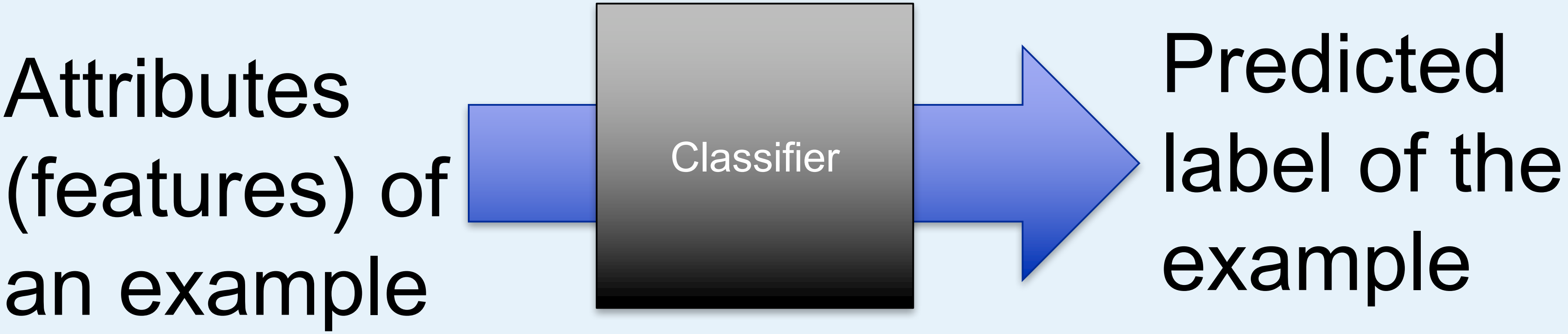
$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}$$

- Three attributes x, y, and z:

  - $D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2}$

Evaluation

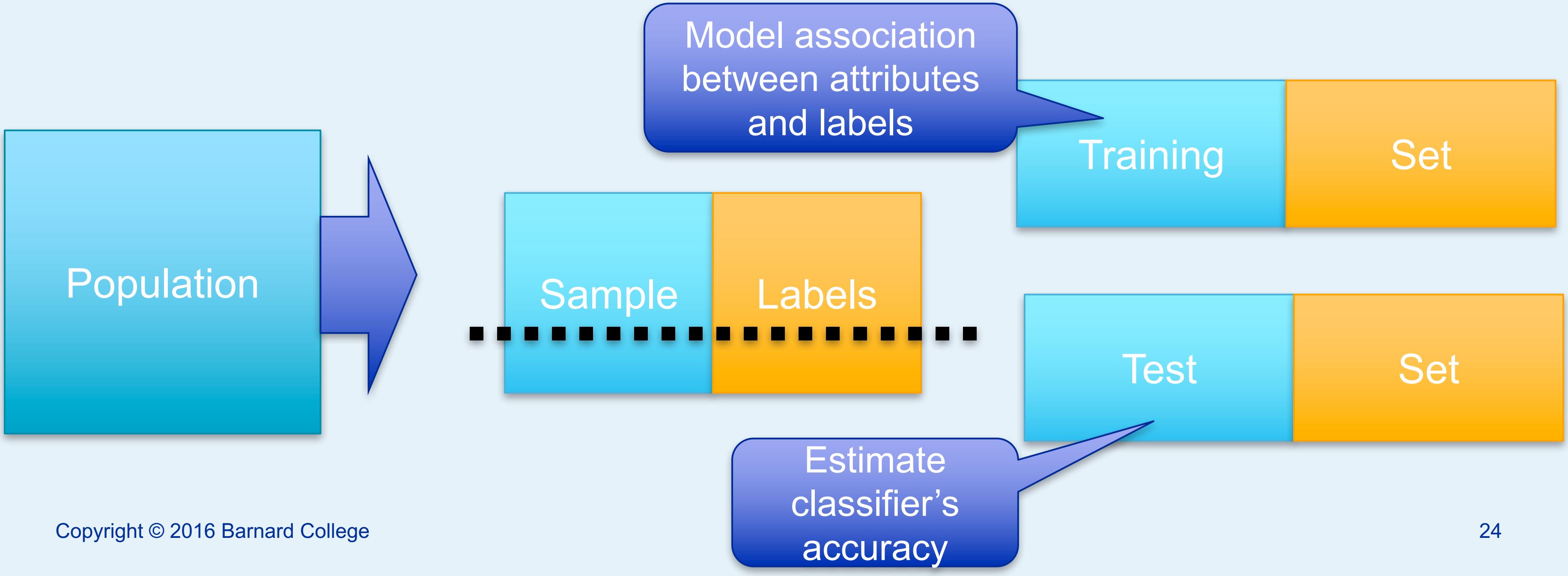
Training a Classifier



Attributes (features) of an example

Classifier

Predicted label of the example



Model association between attributes and labels

Training Set

Population

Sample Labels

Test Set

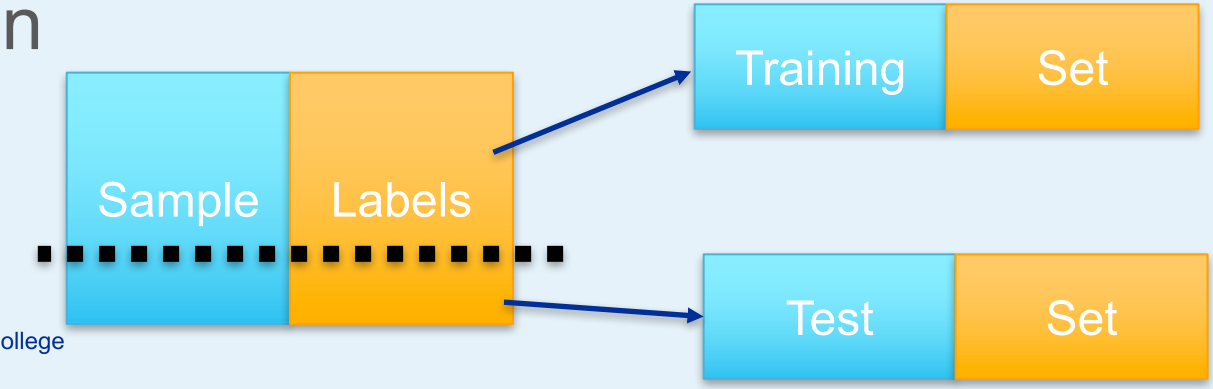
Estimate classifier's accuracy

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


Training
Set
Sample
Labels
Test
Set