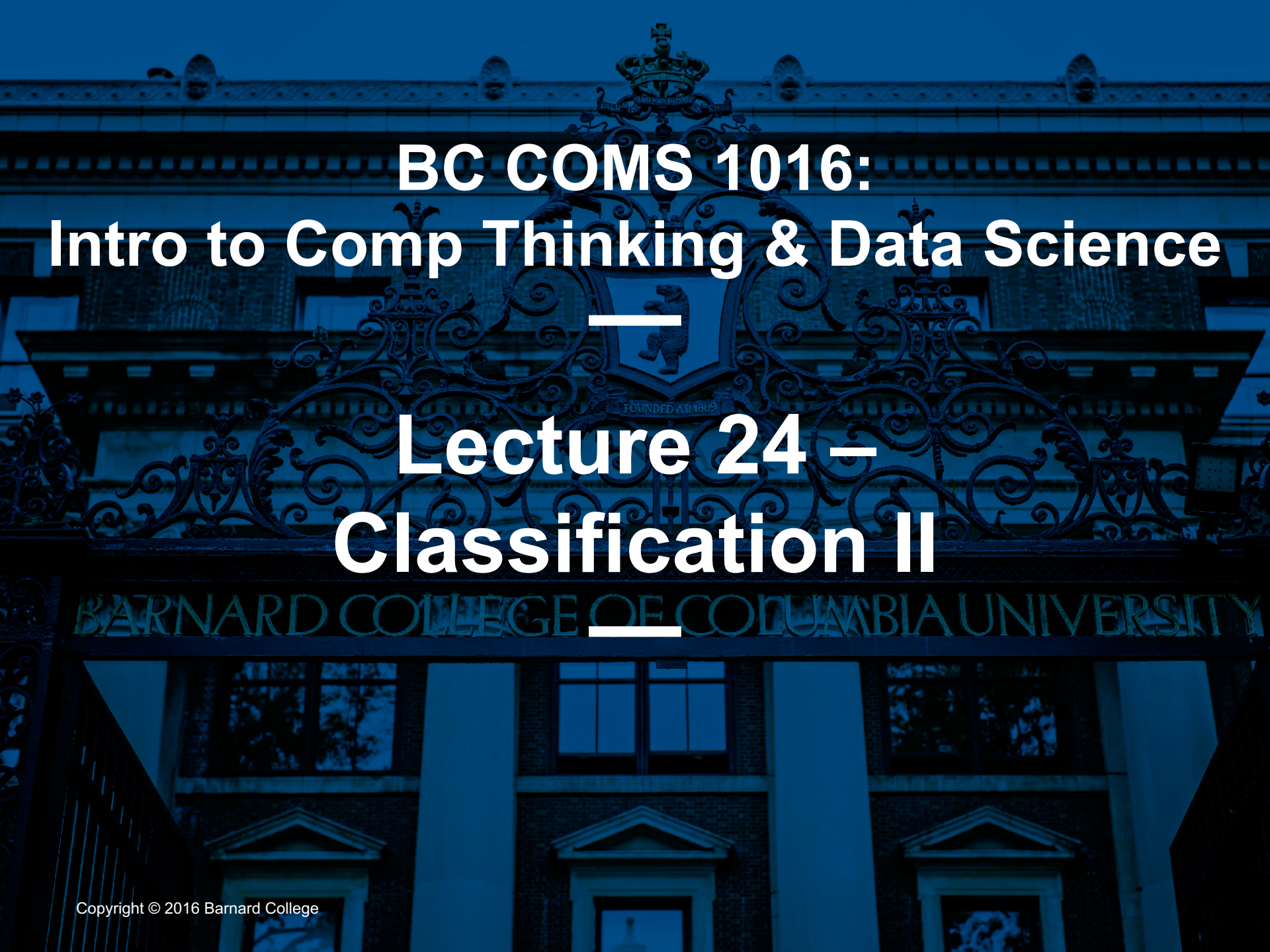


BC COMS 1016: Intro to Comp Thinking & Data Science

—

Lecture 24 – Classification II

Announcements

- Homework 9 - Regression Inference
  - Due Thursday 12/10
- Homework 10 – Classification
  - Due Monday 12/14
- Course Evaluations:
  - Due 12/16
- Project 3:
  - Due Monday 12/14

Announcements - Office Hours

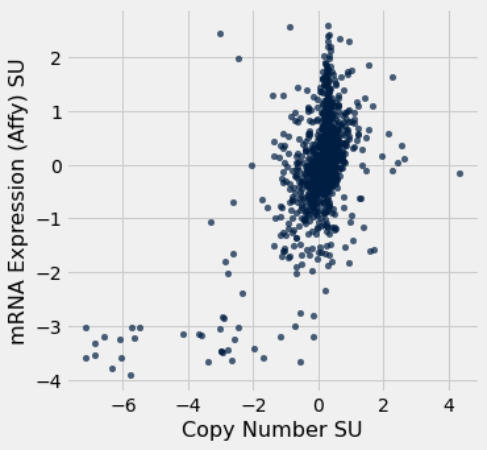
- Tonight (12/10) ~~5pm – 6pm~~ 5:30-6:30pm
- Next week (Week of 12/14):
  - Same as usual *in general*
  - Except:
    - No Tuesday/Wed 9:00 – 10:00PM
    - I'll make new times for me
- Final Project:
  - Group Project
  - Will send out detailed write up tonight or tomorrow

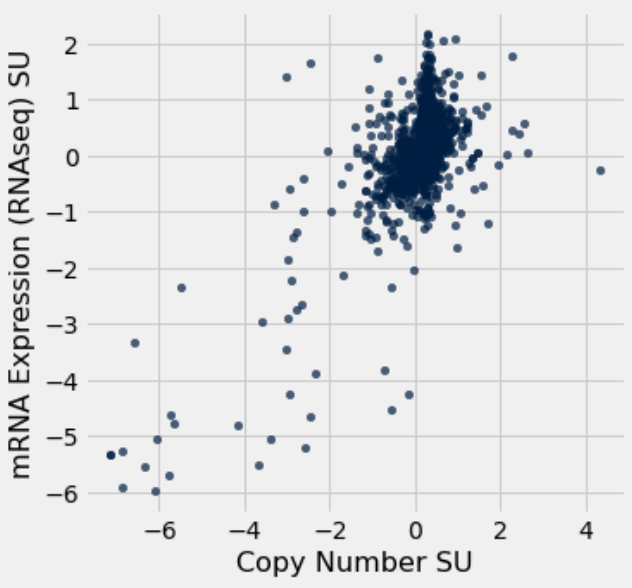
Announcements - Office Hours

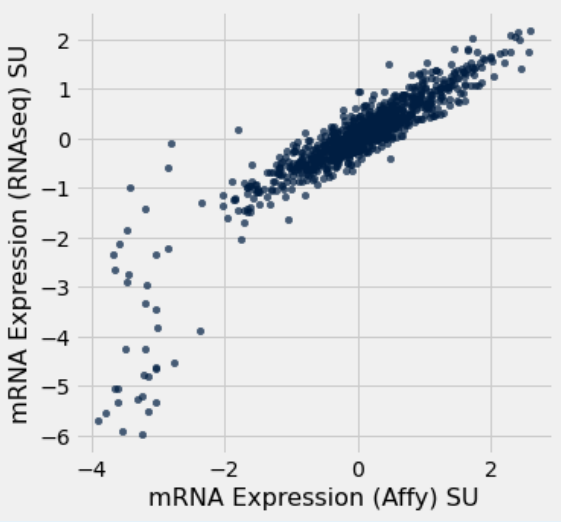
- Tonight (12/10) ~~5pm – 6pm~~ 5:30-6:30pm

- Next week (Week of 12/14):
  - Same as usual *in general*
  - Except:
    - No Tuesday/Wed 9:00 – 10:00PM
    - I'll make new times for me

What is a Pattern?







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

Guessing the Value of an Attribute

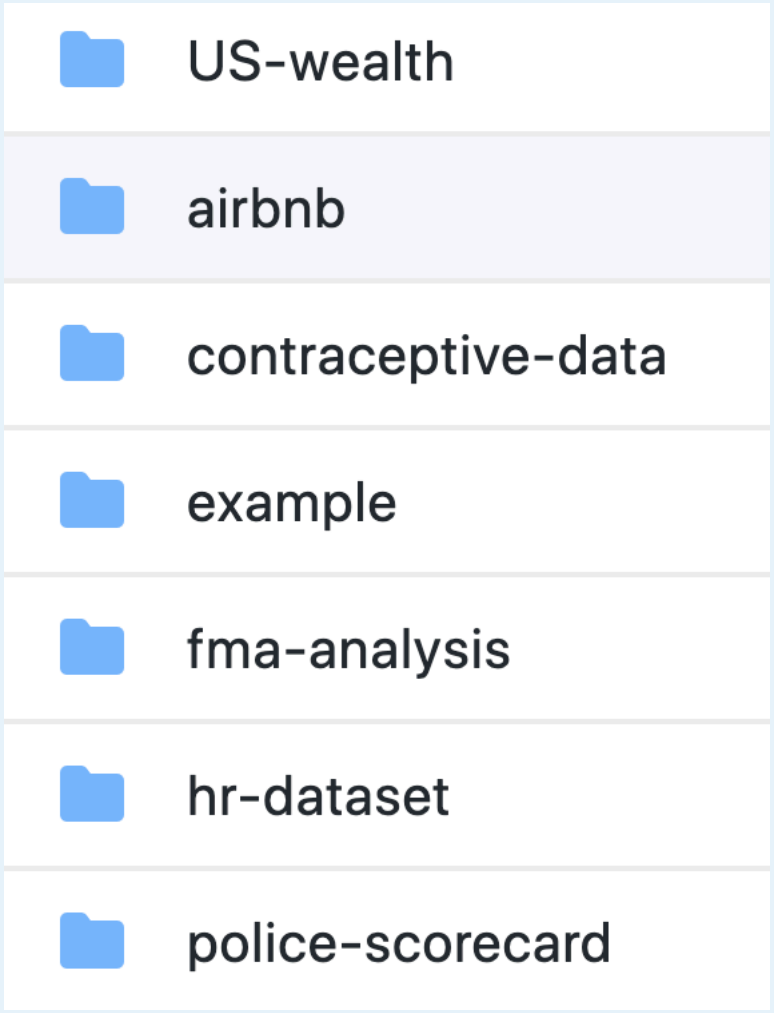
- Based on incomplete information
- One way of making predictions:
  - To predict an outcome for an individual,
  - find others who are like that individual
  - and whose outcomes you know.
  - Use those outcomes as the basis of your prediction.
- Two Types of Prediction
  - Classification = Categorical; Regression = Numeric

Final Project

Final Project – Real World Data Science

- Explore a real world dataset from multiple tables
  - Choose from 6 datasets
- Ask 2 questions that the dataset can help answer
  - Hypothesis Testing
  - Prediction
- Use methods covered in in the class to answer these questions

6 Datasets to choose from



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 12/17

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

Classifiers

A Classifier

Attributes (features) of an example

Predicted label of the example

Rows

(Zoom Poll)

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

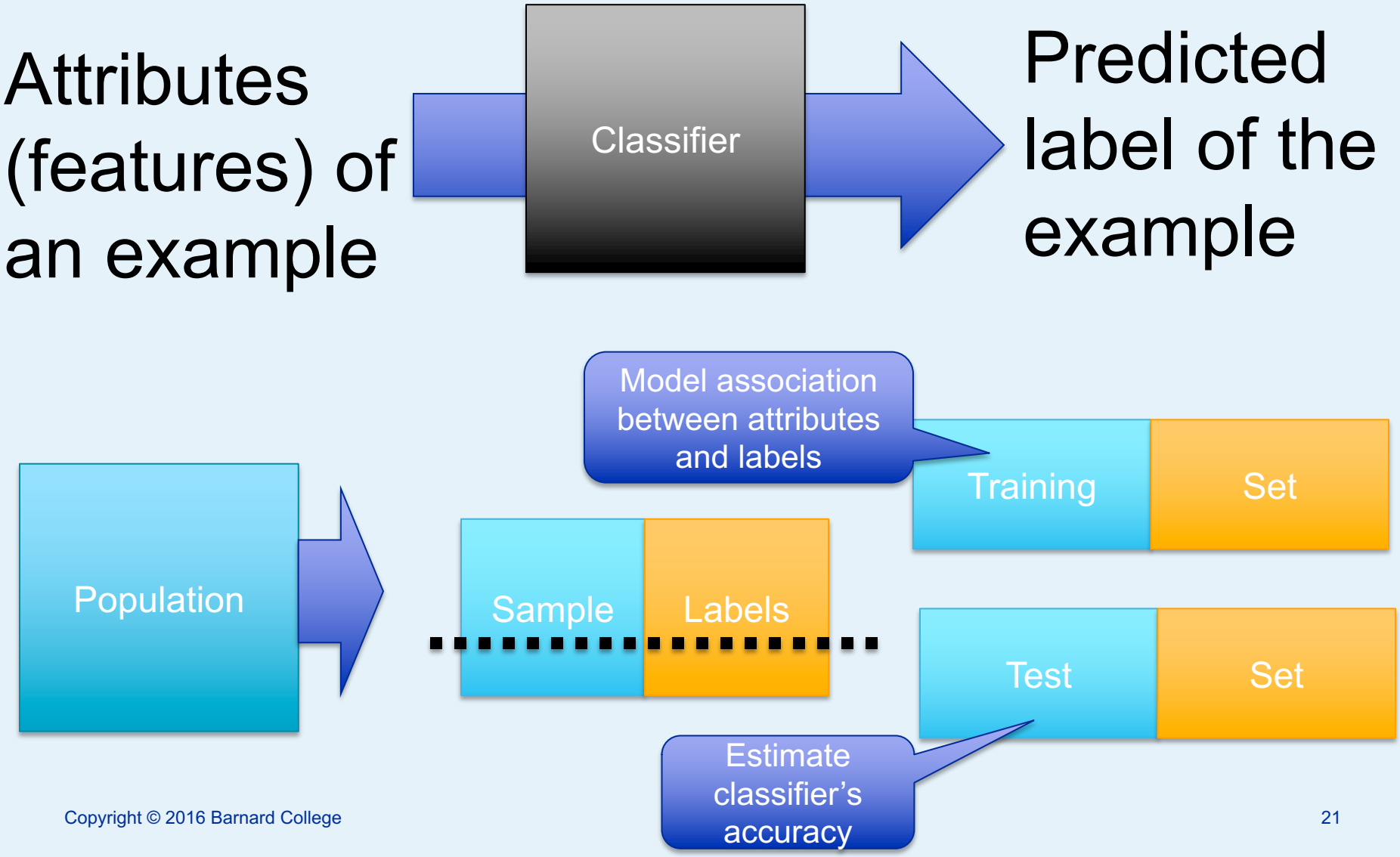
Evaluation

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

Training and Evaluating a Classifier

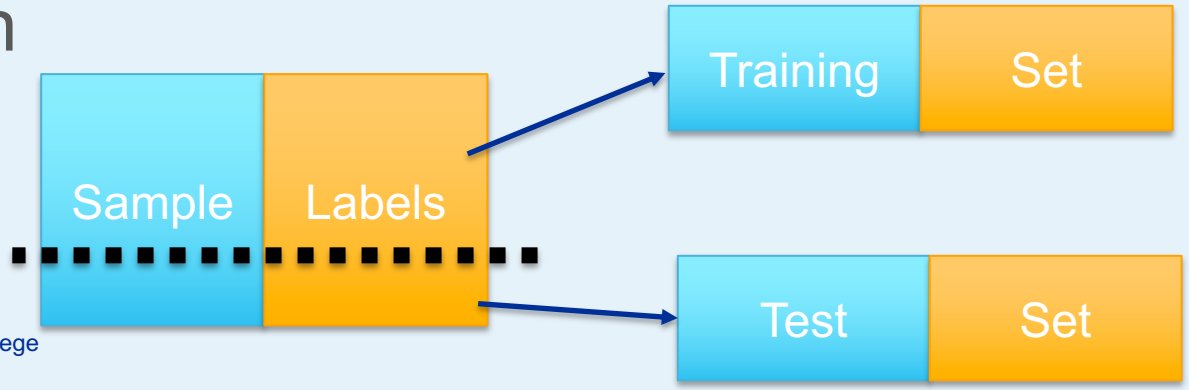


Accuracy of a Classifier

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

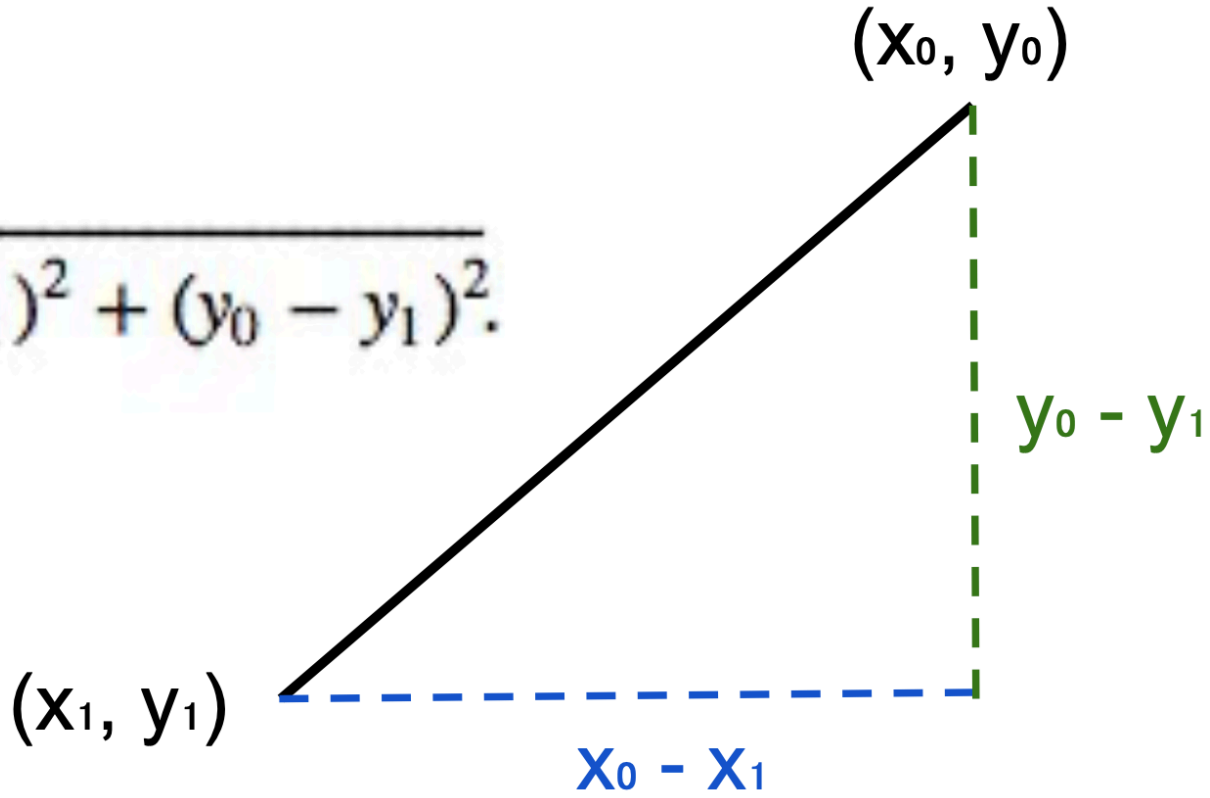
If the labeled data set is sampled at random from a population, then we can infer accuracy on that population



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

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

Nearest Neighbor Classifier