YaleNUSCollege

YSC2239 Lecture 12

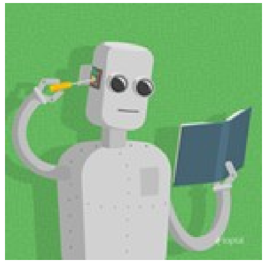
Today’s class

- Classification

- Reading: Chapter 16, 17

Prediction

Review: Taxonomy of Machine Learning



Labeled Data

Unlabeled Data

Supervised Learning

Unsupervised Learning

Quantitative Response

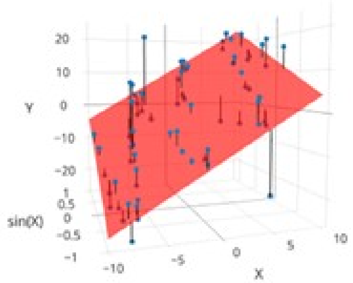
Categorical Response

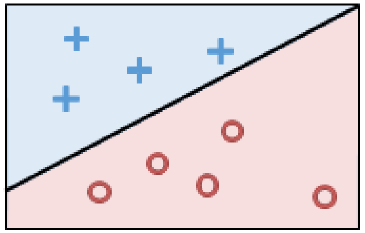
Regression

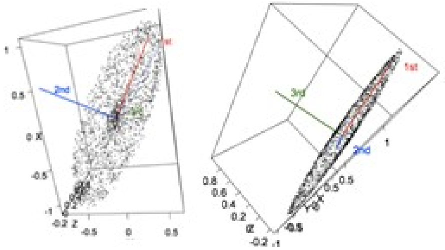
Classification

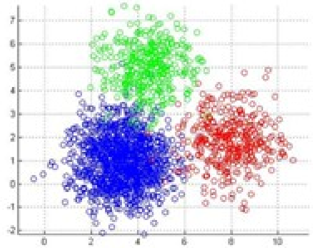
Dimensionality Reduction

Clustering







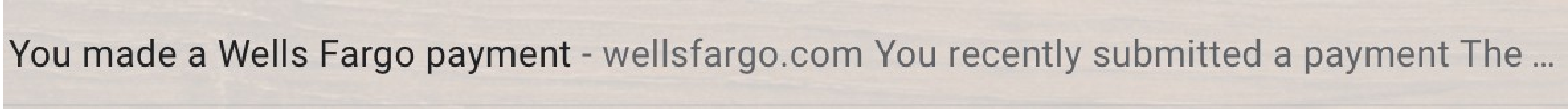


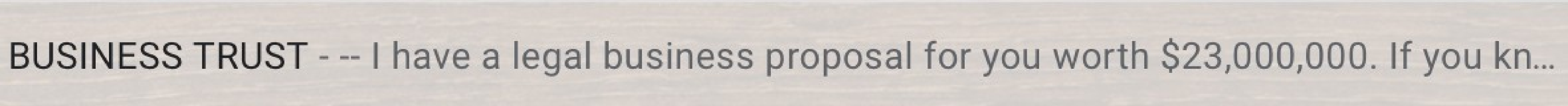
Guessing the Value of an Attribute

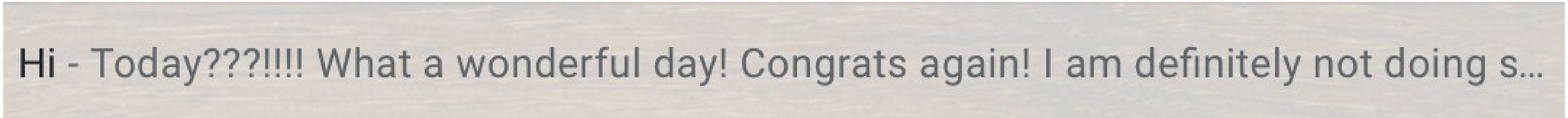
- Based on incomplete information
- One way of making predictions:
    - To predict an outcome for an individual,
    - find others who are like that individual
    - and whose outcomes you know.
    - Use those outcomes as the basis of your prediction.

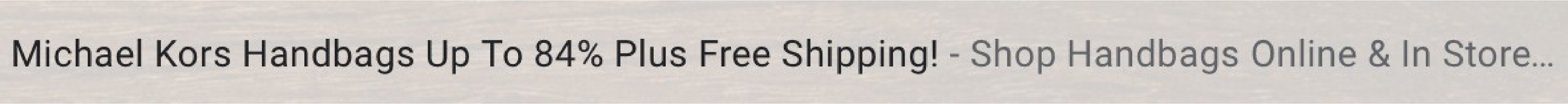
- Two Types of Prediction
    - Classification = Categorical; Regression = Numeric

Prediction Example: Spam or Not?


You made a Wells Fargo payment - wellsfargo.com You recently submitted a payment The ...


BUSINESS TRUST - -- I have a legal business proposal for you worth $23,000,000. If you kn...


Hi - Today???!!!! What a wonderful day! Congrats again! I am definitely not doing s...

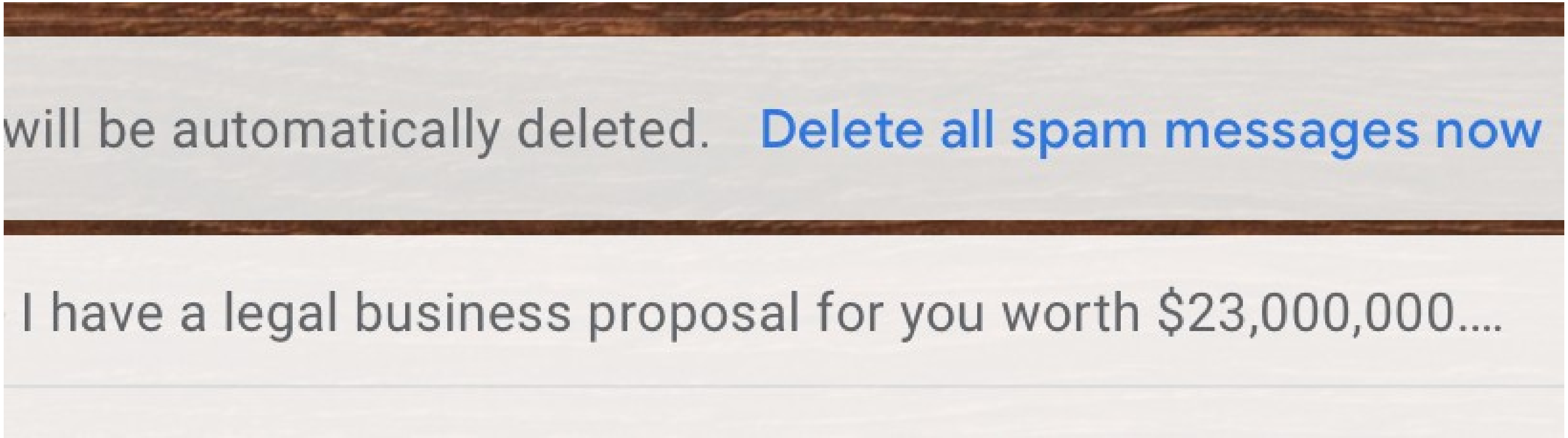

Michael Kors Handbags Up To 84% Plus Free Shipping! - Shop Handbags Online & In Store...

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

Classification

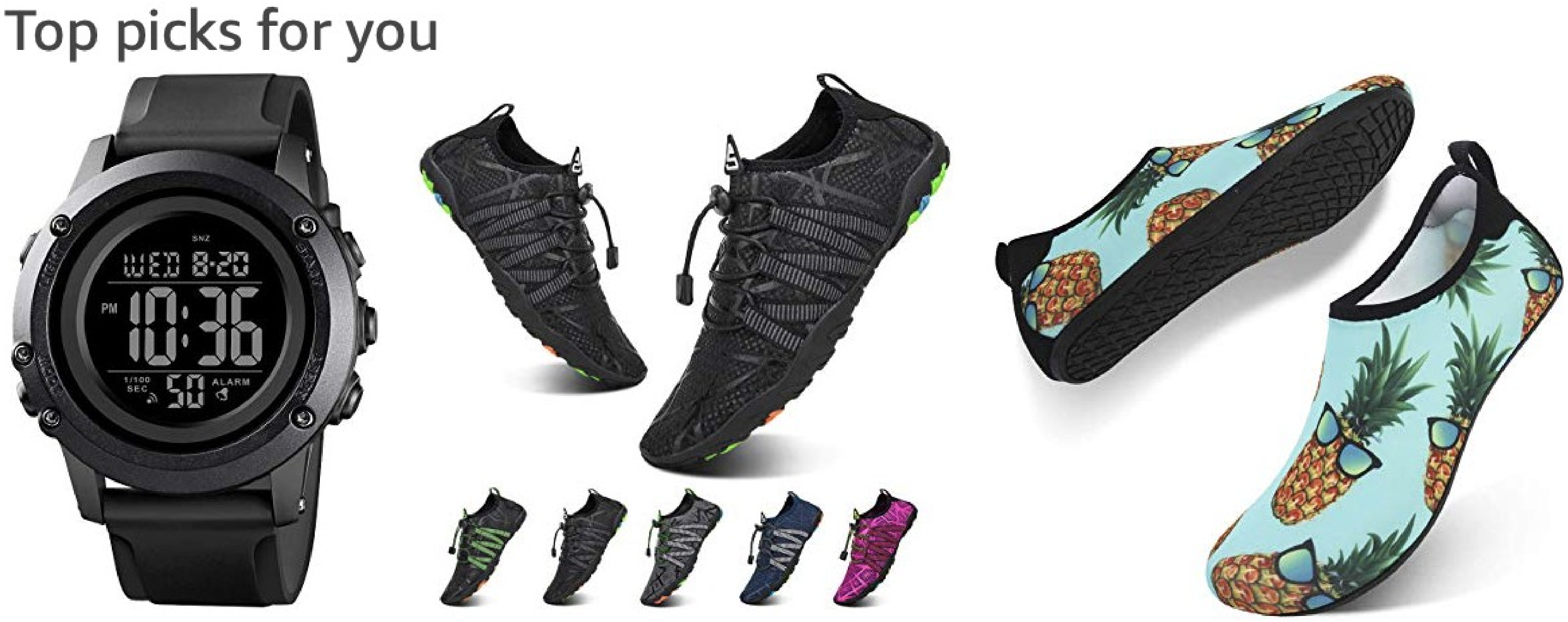
Classification Examples



will be automatically deleted. Delete all spam messages now

I have a legal business proposal for you worth $23,000,000....

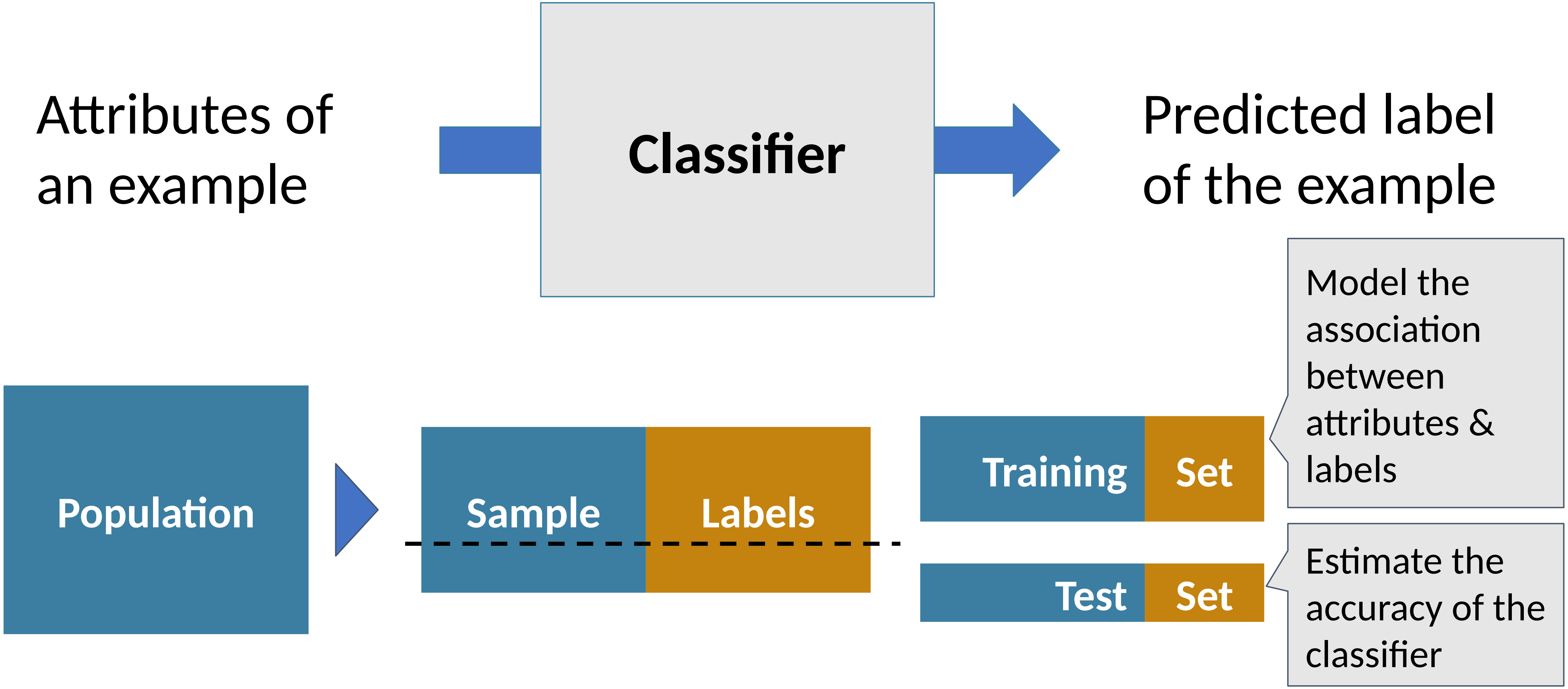
Classification Examples



Top picks for you

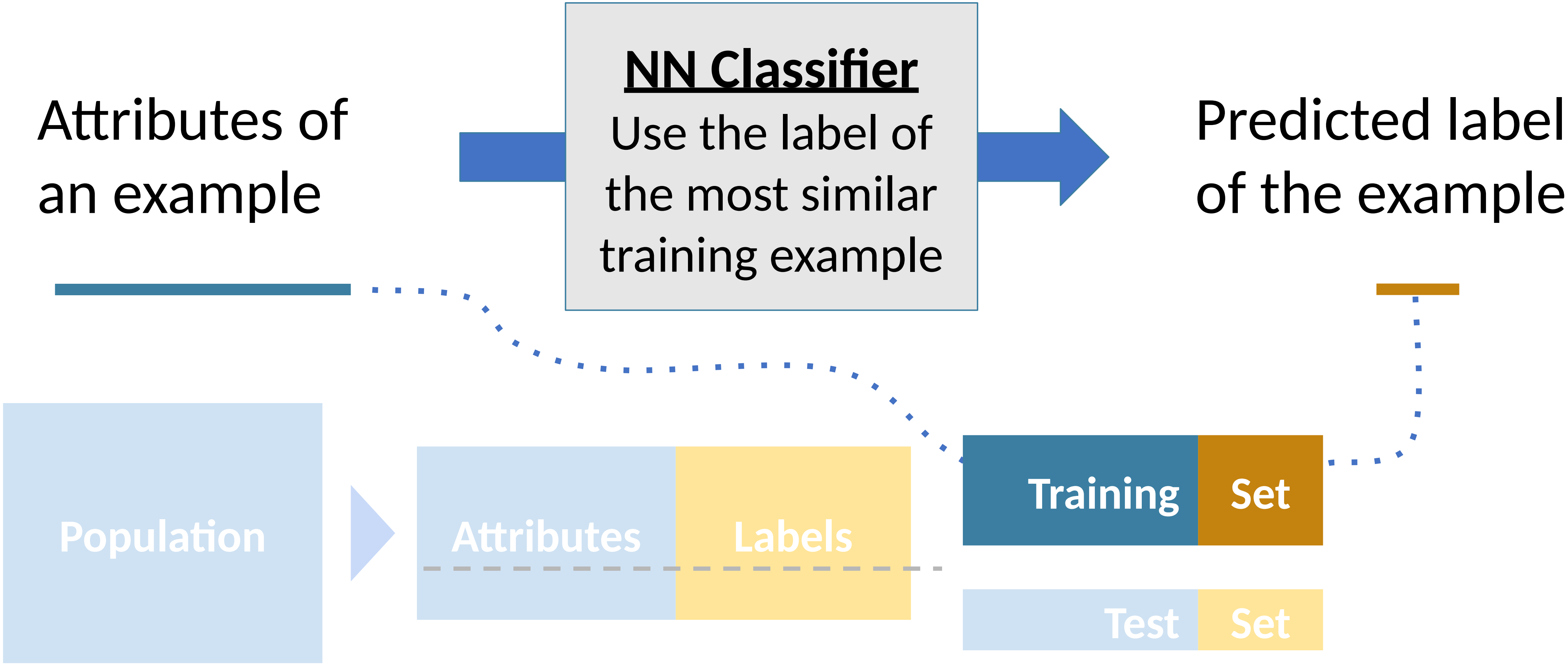
Classifiers

Training a Classifier



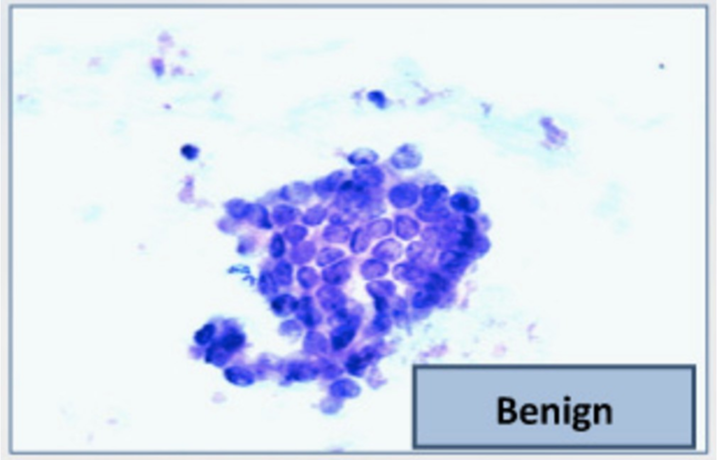
Attributes of an example
Classifier
Predicted label of the example
Population
Sample
Labels
Training
Set
Test
Set
Model the association between attributes & labels
Estimate the accuracy of the classifier

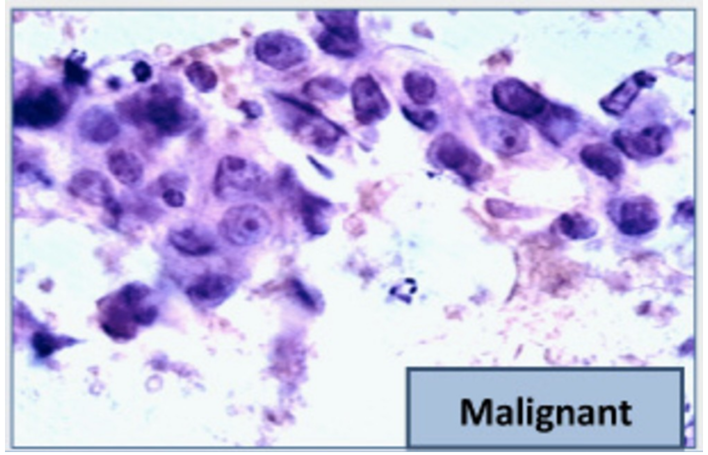
Nearest Neighbor Classifier



The Google Science Fair

- Brittany Wenger, a 17-year-old high school student in 2012
- Won by building a breast cancer classifier with 99% accuracy





(Demo)

Distance

Rows of Tables

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) ...
```

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*:

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

(Demo)

  - and so on …

Nearest Neighbors

Finding the *k* Nearest Neighbors

To find the *k* nearest neighbors of an example:

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

(Demo)

The Classifier

To classify a point:

- Find its *k* nearest neighbors
- Take a majority vote of the *k* nearest neighbors to see which of the two classes appears more often
- Assign the point the class that wins the majority vote

(Demo)

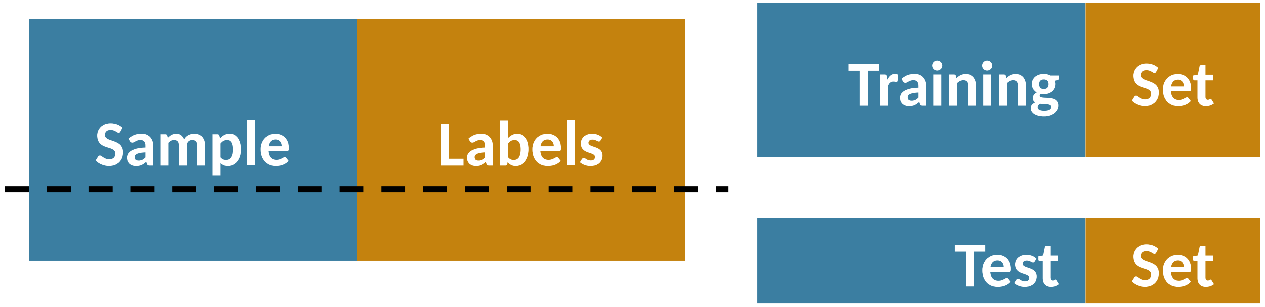
Evaluation

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



(Demo)