

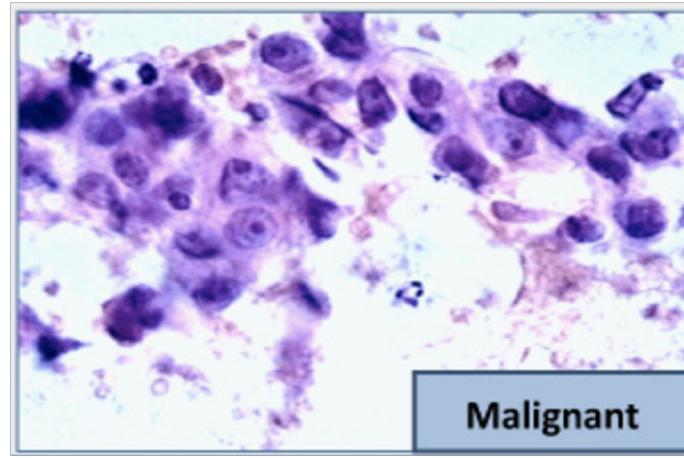
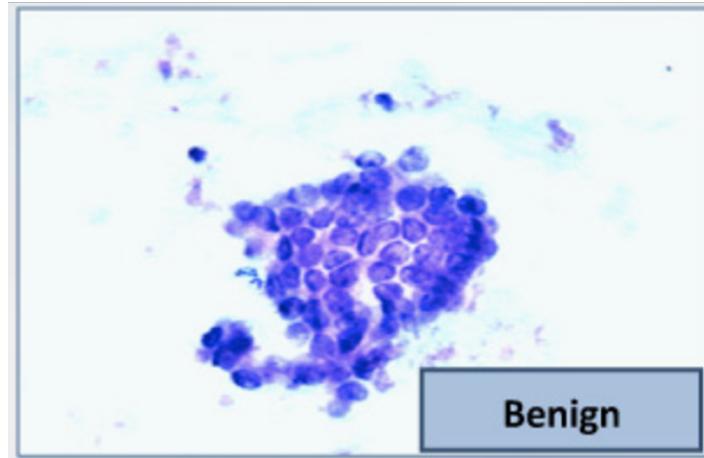
# CompSci 190: Lecture 14: Classification

Jeff Forbes

November 28, 2018

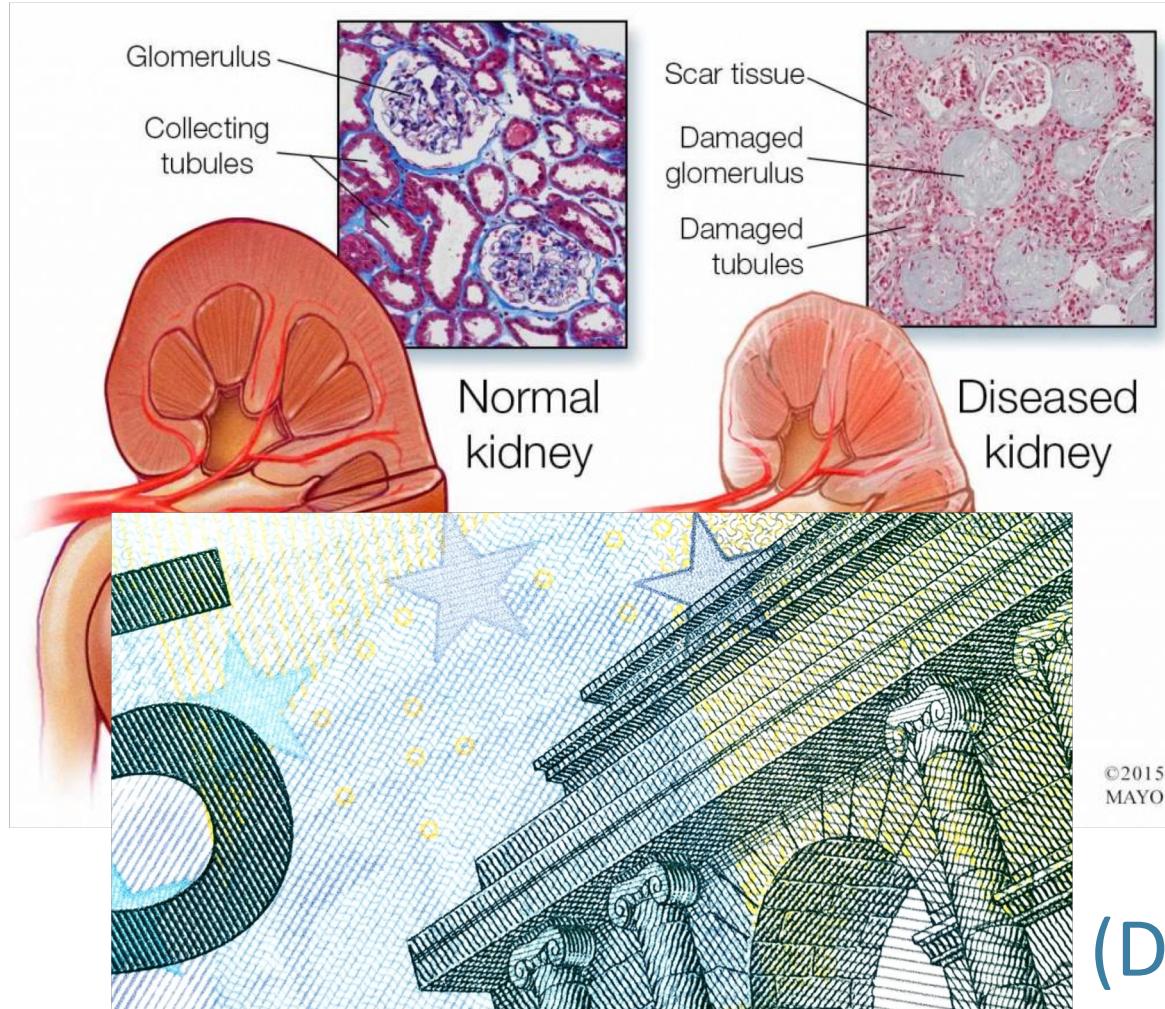
# The Google Science Fair

- [Brittany Wenger, Trinity 2017](#)
- Won by 2012 Science Fair  
building a breast cancer classifier  
with 99% accuracy

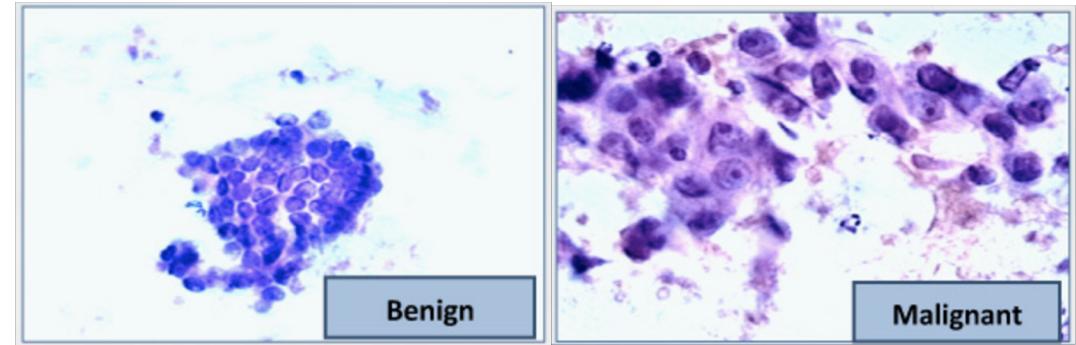


[April 22, 2013](#)

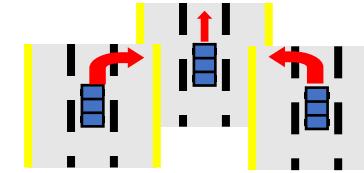
# Classification Examples



(Demo)

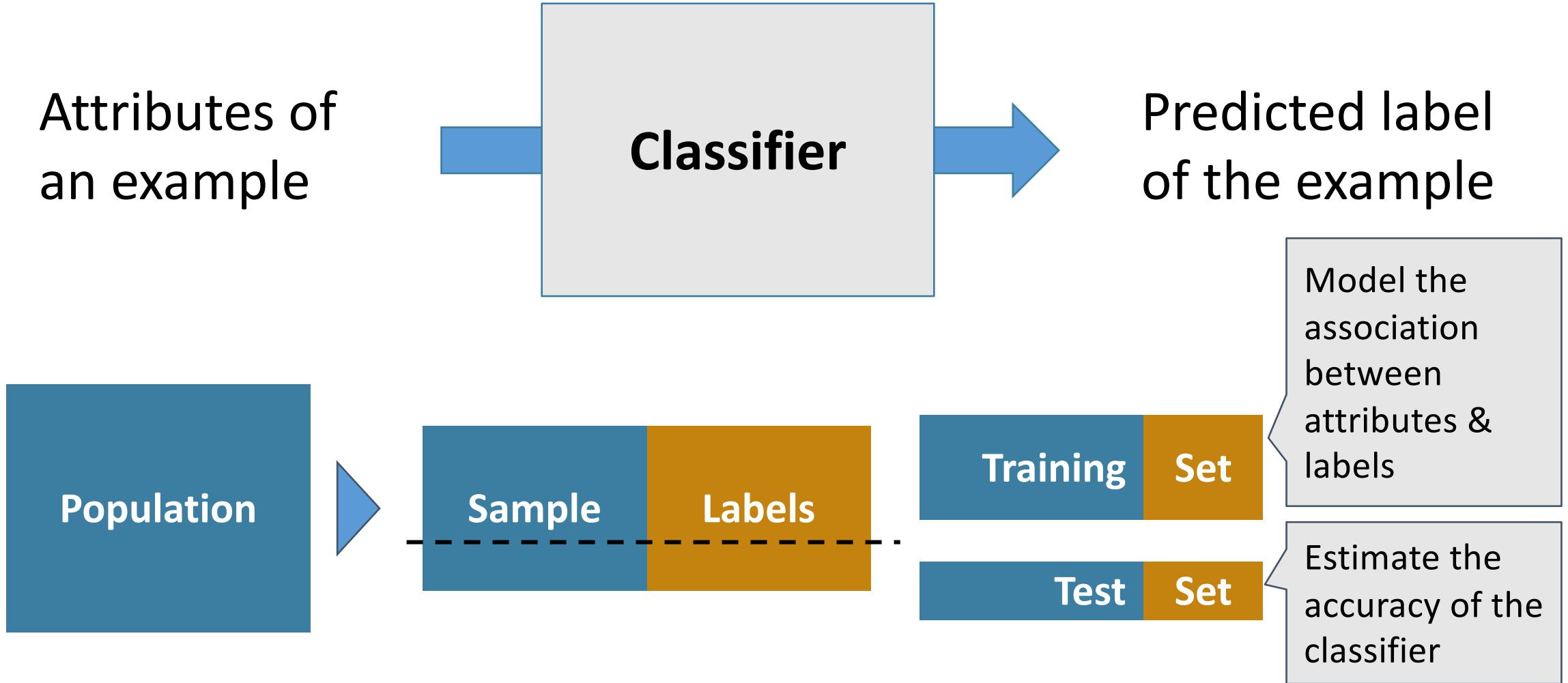


Wenger 2012

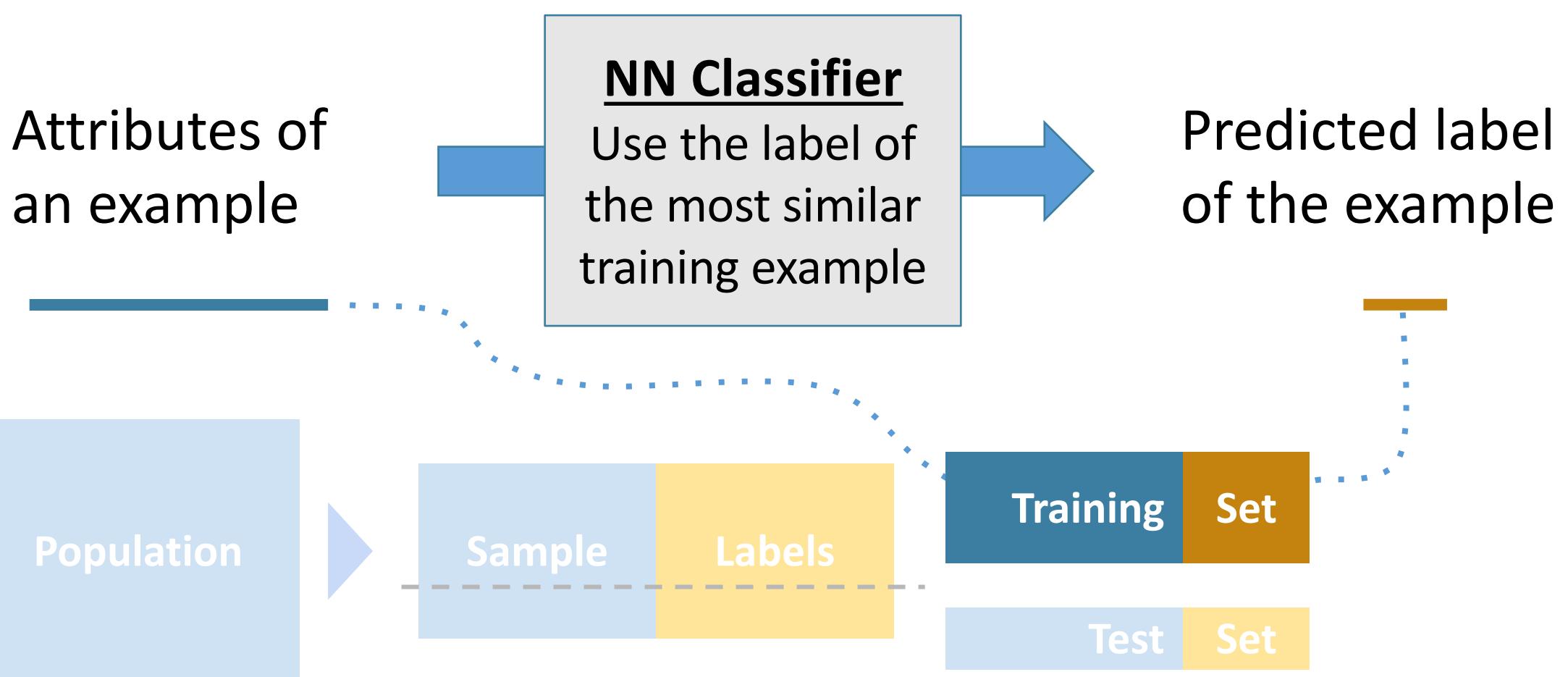


Forbes, 1998

# Training a Classifier



# Nearest Neighbor Classifier



# 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  $z$ :

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

- and so on ... <http://bit.ly/FoDS-f18-1128>
-

# 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) ...`

(Demo)

# 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)

---

# 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)