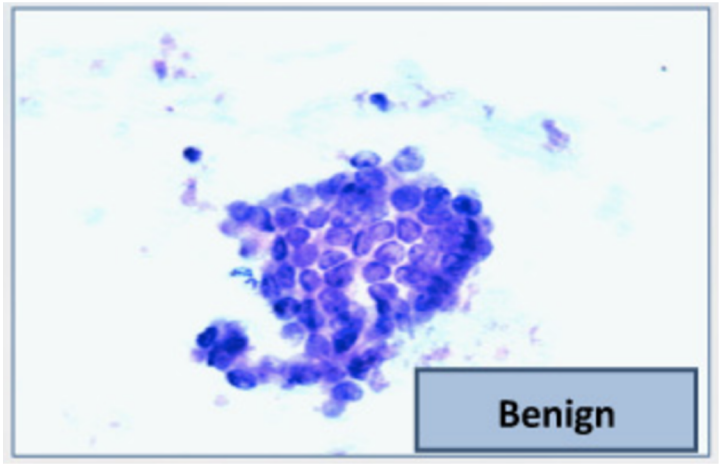
CompSci 116: Lecture 11: Prediction - Classification

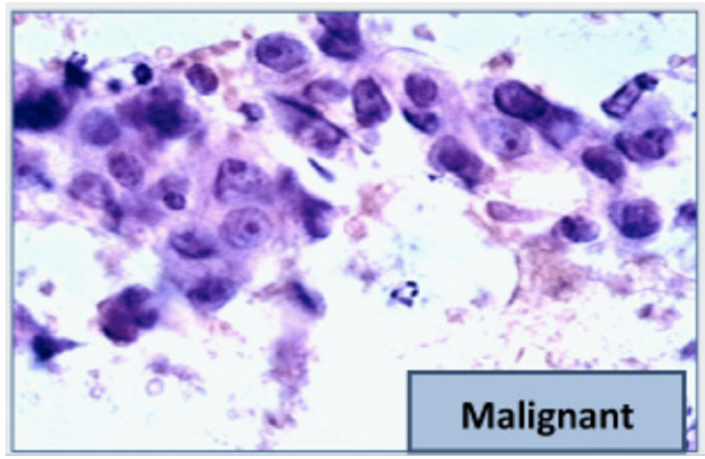
Jeff Forbes

April 4, 2019

The Google Science Fair

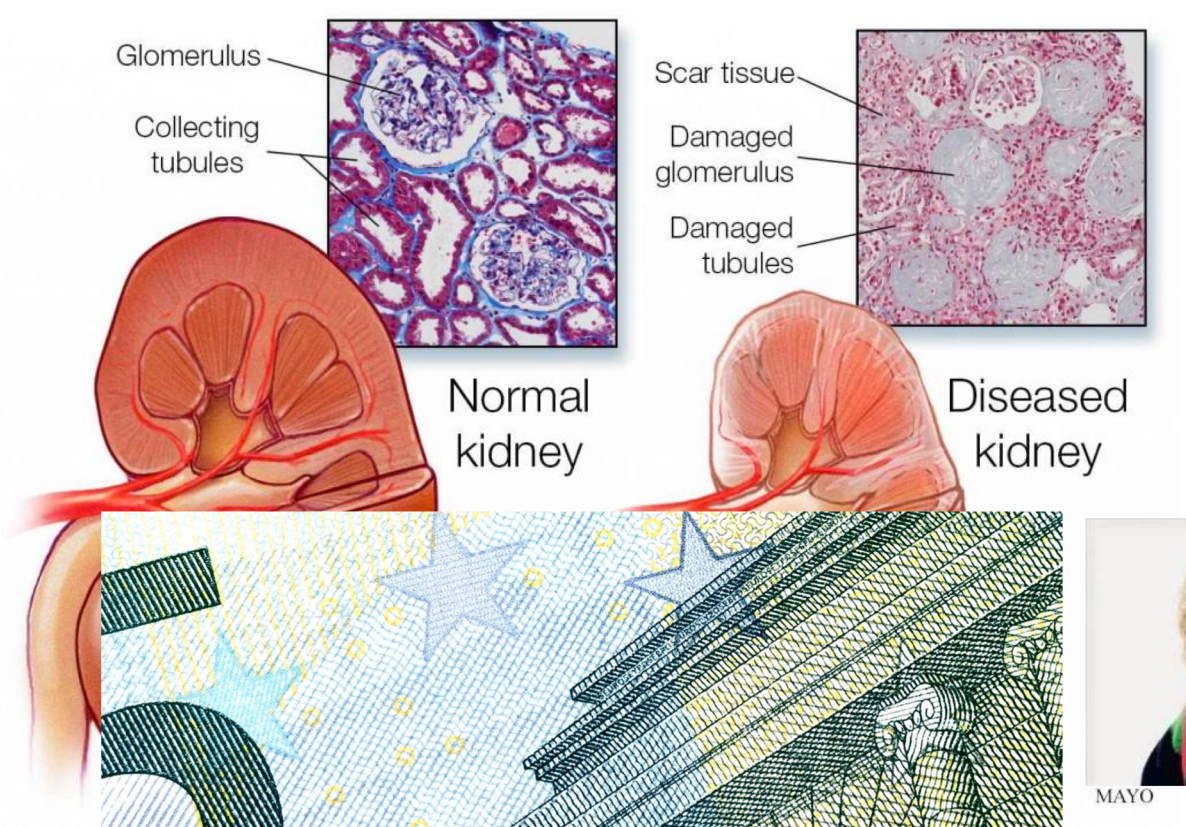
- Brittany Wenger, Trinity 2017
- Won by 2012 Science Fair building a breast cancer classifier with 99% accuracy

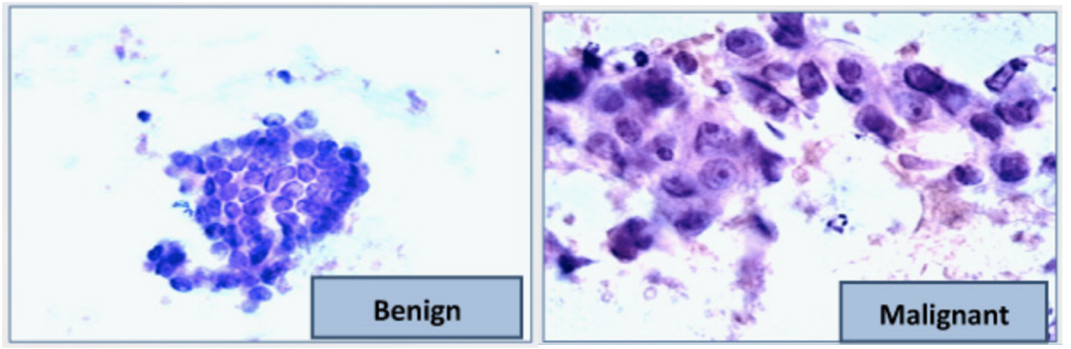




April 22, 2013

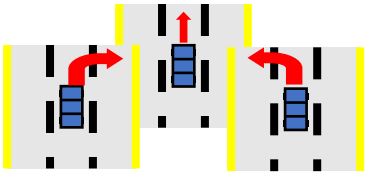
Classification Examples





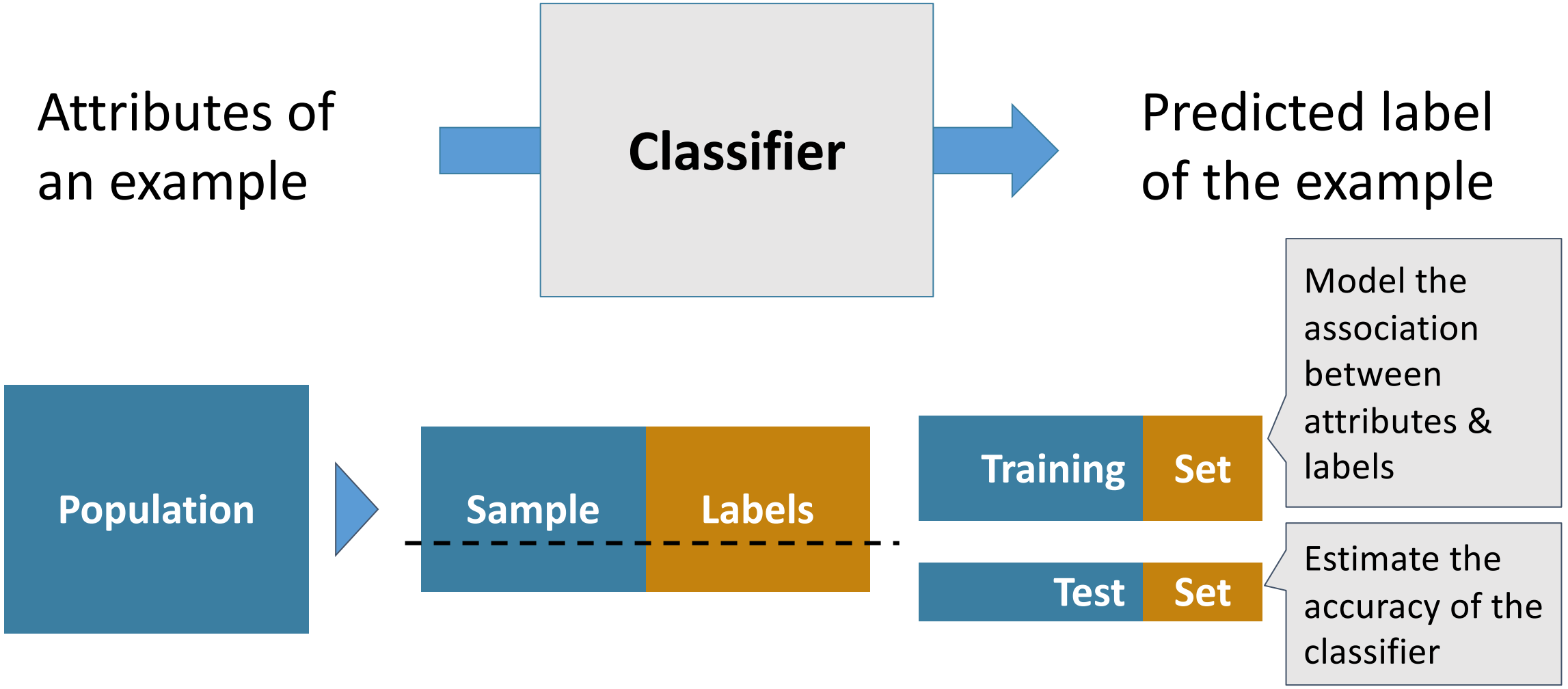
Wenger 2012

(Demo)

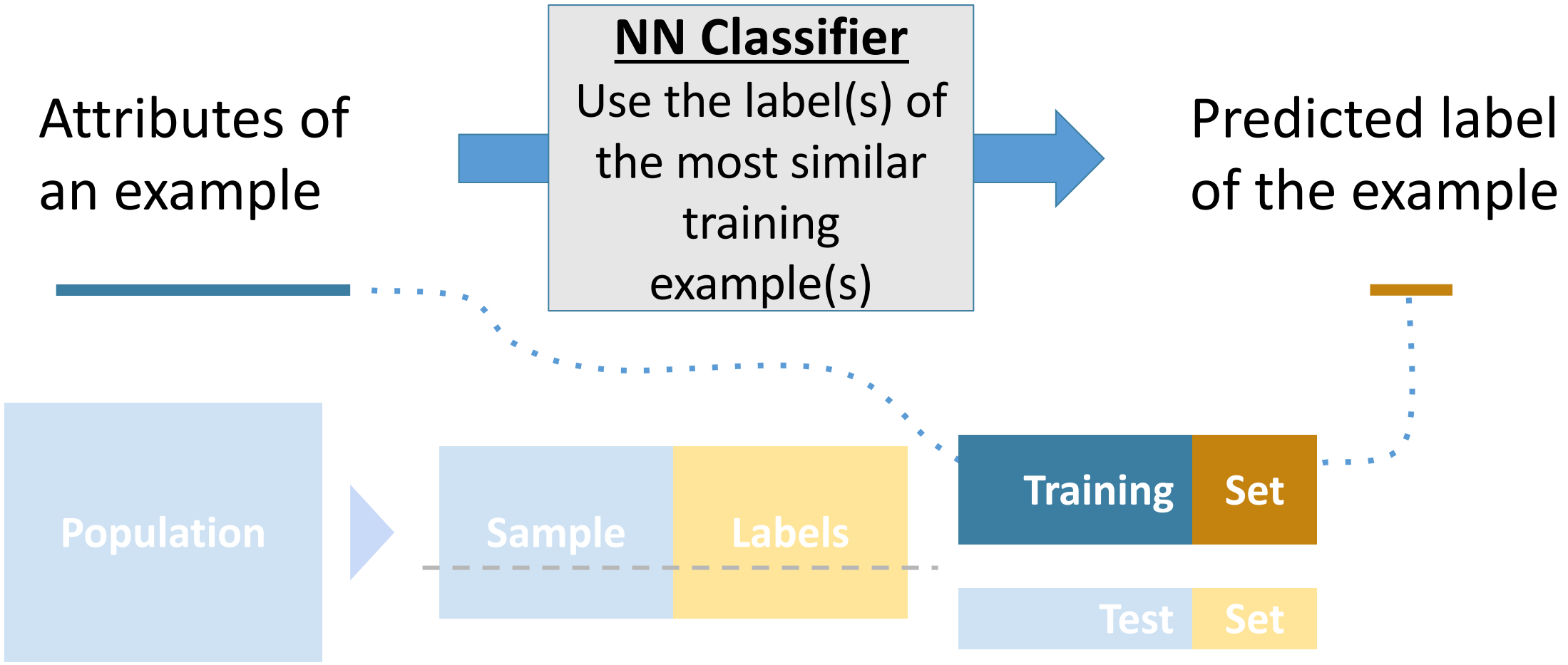


Forbes, 1998

Training a Classifier



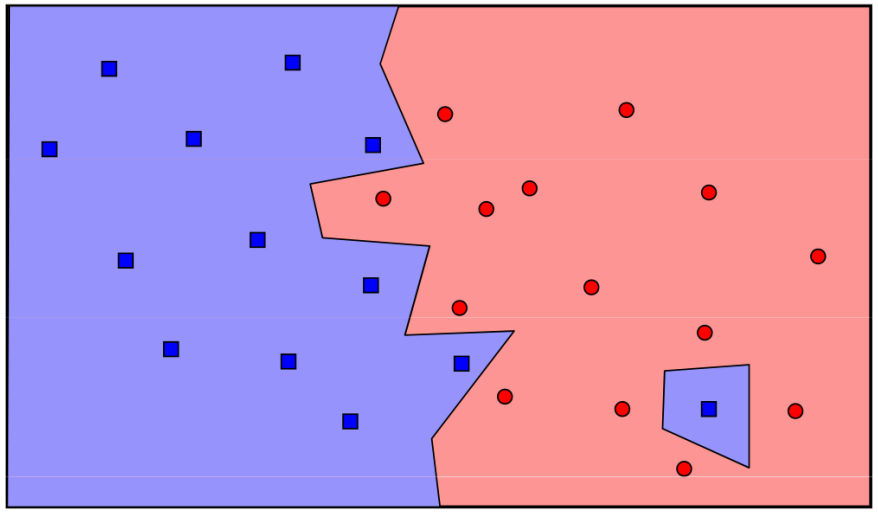
Nearest Neighbor Classifier



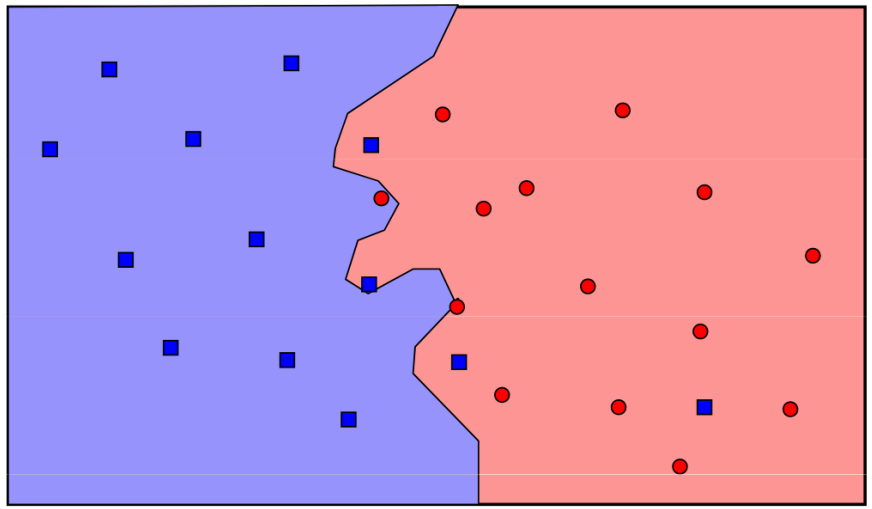
The Classifier

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



1-NN



3-NN

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-s19-0404**

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)

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)

Accuracy of a Classifier

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

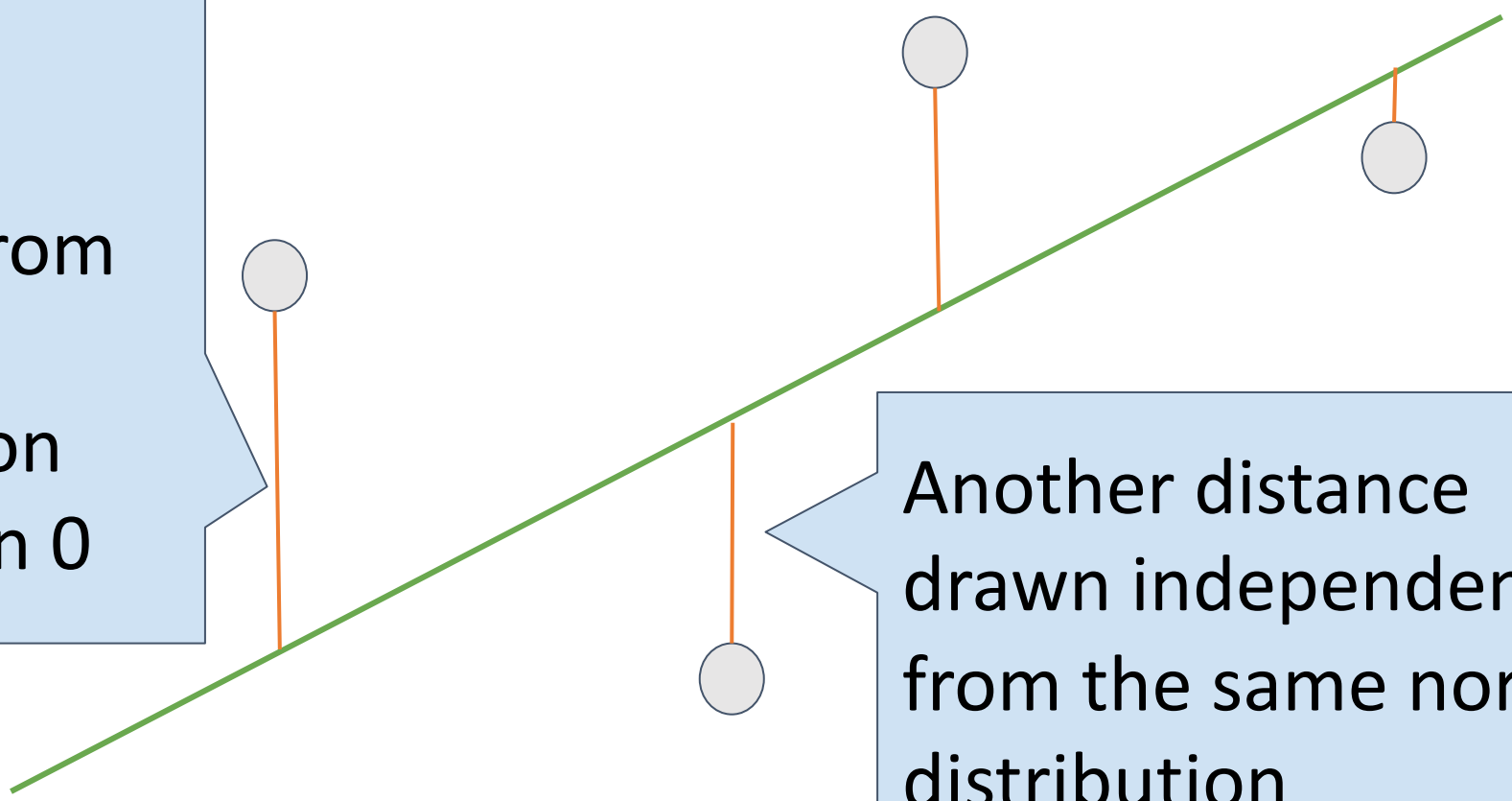
Need to compare classifier predictions to true labels

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

(Demo)

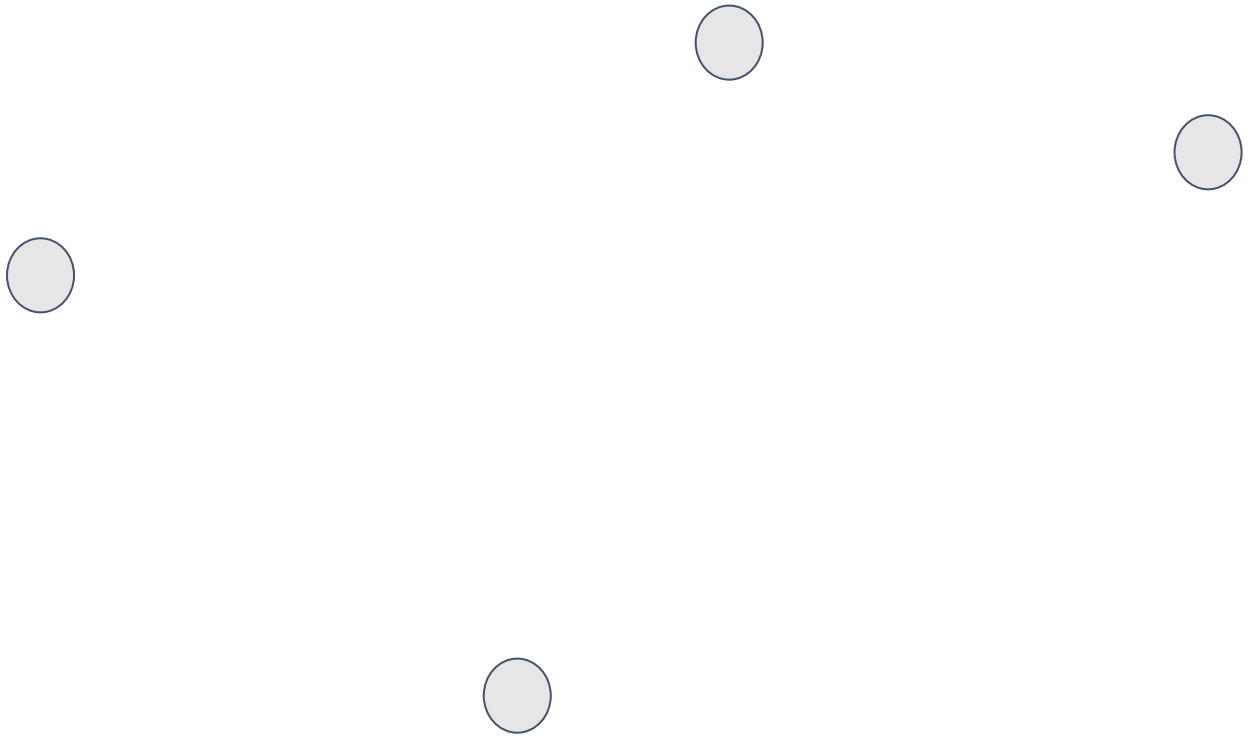
Regression "Model": Signal + Noise

Distance drawn at random from normal distribution with mean 0



Another distance drawn independently from the same normal distribution

What We Get to See



Regression Prediction

If the data come from the regression model,

- The regression line is close to 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 predicted value of *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 interval also depends on *x*.
  - Typically, intervals are wider for values of *x* that are further away from the mean of *x*.

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.

Rain on the Regression Parade

We observed a slope based on our sample of points.

But what if the sample scatter plot got its slope just by chance?

What if the true line is actually FLAT?

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, reject the null hypothesis.
  - If the interval does contain 0, there isn't enough evidence to reject the null hypothesis.