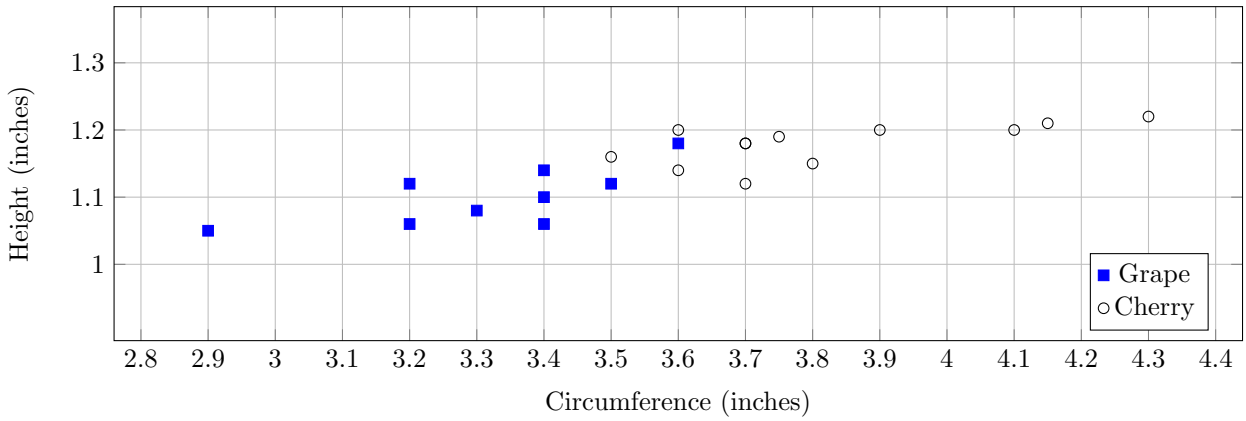
Won't you be my neighbor?

There are a lot of different algorithms which can be used to classify data. The first one that we'll cover is called the *k-nearest neighbors* algorithm. For this algorithm,

1. We pick a number k.
2. In order to classify a point P, we look at the k data points which are closest to P.
3. These k data points "vote" on what they think P should be, and the majority wins.

We'll practice this with an example.



For our example, we'll work with $k = 3$. So, for a new point P, we'll use the 3 closest data point to classify that point.

1. Plot the point P for a fruit with circumference 4 inches and height 1.2 inches. Which 3 data points are closest to P? Are they grapes or cherries?

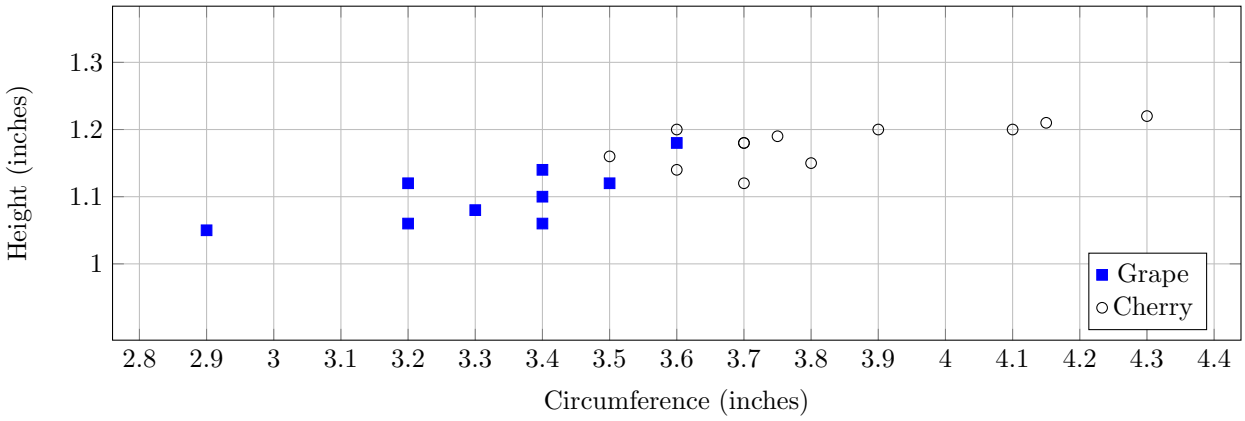
   Based on those three neighbors, would we classify P as a grape or a cherry?

2. Plot the point Q for a fruit with circumference 3.6 inches and height 1.16 inches. Which 3 data points are closest to Q? Are they grapes or cherries?

   Based on those three neighbors, would we classify Q as a grape or a cherry?

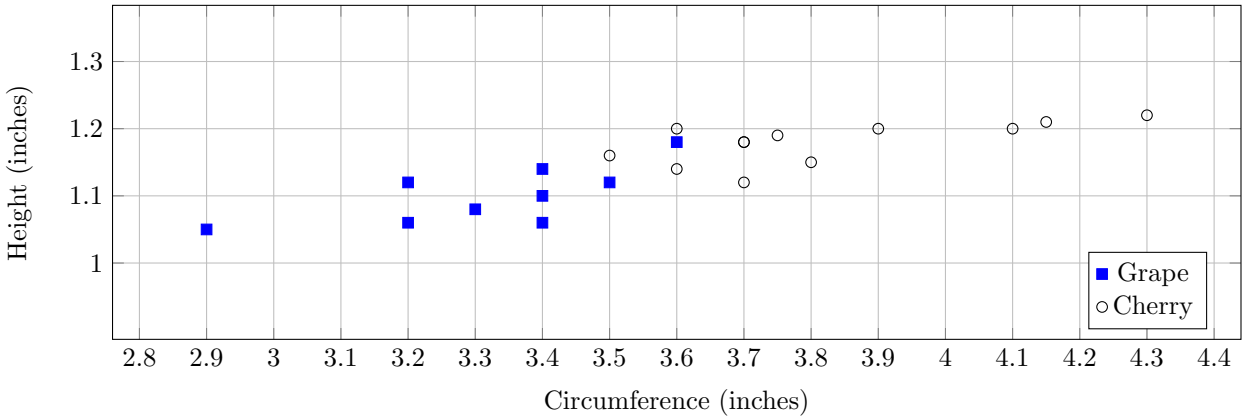
3. Shade in the region of your graph that would be classified as cherries using the 3-nearest neighbors algorithm.

4. Suppose instead of using the 3 nearest neighbors, we took $k = 1$ and only used the single closest data point. Shade in the region of the graph that would be classified as cherries using the 1-nearest neighbors algorithm.



Do you think the 1-nearest neighbor classification is better or worse than the 3-nearest neighbor classification? Why?

5. Suppose instead of using the 3 nearest neighbors, we took $k = 20$ and used the 20 closest data points. Shade in the region of the graph that would be classified as cherries using the 20-nearest neighbors algorithm.



Do you think the 20-nearest neighbor classification is better or worse than the 3-nearest neighbor classification? Why?

How can you decide what value of k works well for k-nearest neighbor classification?