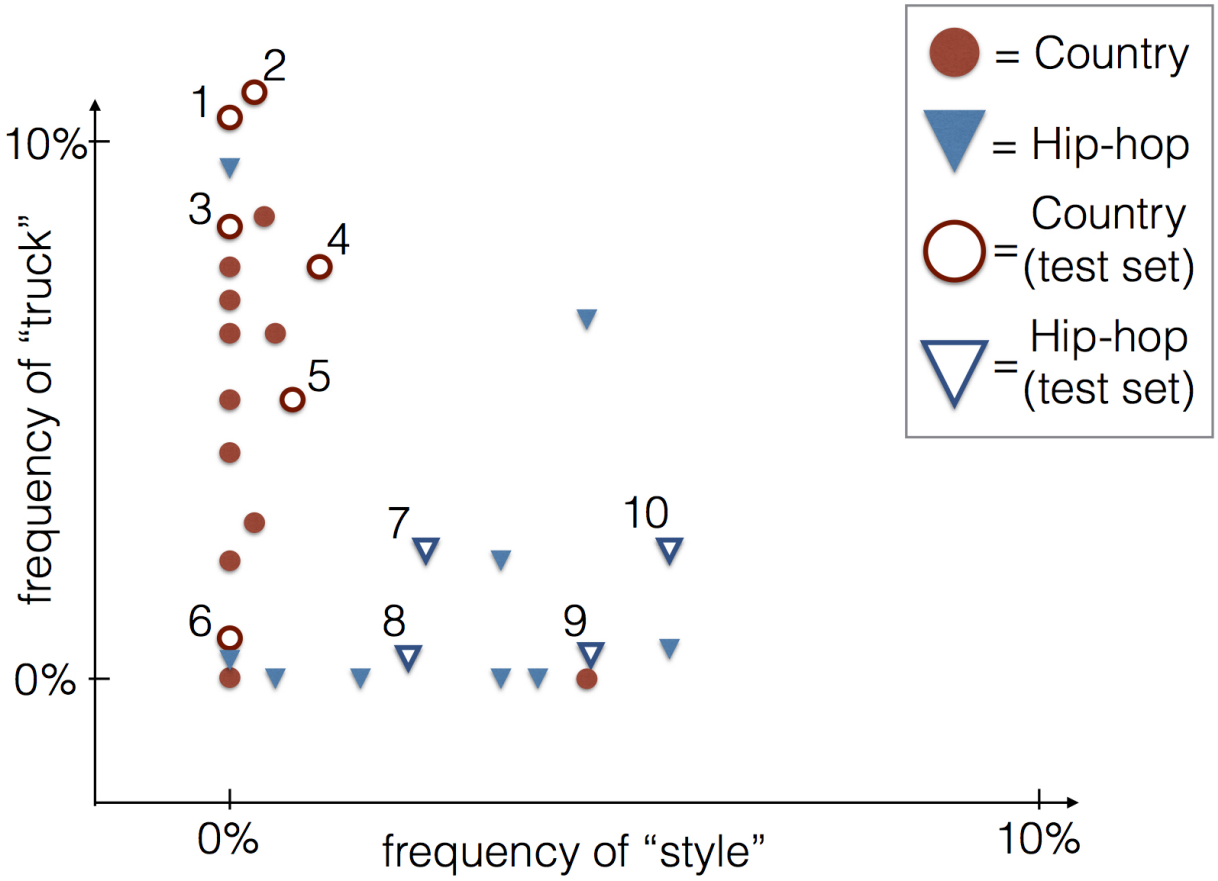
Discussion: Classifying Songs by Lyrics

Today's discussion is designed to prepare you for Project 3. In this project, you'll create a classifier that predicts a song's genre based on its lyrics (the words in the song). Each observation is a song. Every song's genre is either hip-hop or country. There is one attribute for every word that's in any song. A song's value for an attribute is the *proportion* of the song's lyrics that are that word.

In this discussion, we'll limit ourselves to just two attributes: the frequency of the word "truck" and the frequency of the word "style". It's easiest to see this with a picture.



The hollow blue triangle labeled 7 represents one hip-hop song. The graph says that song contains a few instances of the word "truck" - around 2% of its lyrics - and a few more instances of the world "style" - around 3%.

Whenever you build a classifier, it's important to have a separate dataset to test it on. This is called a *test set*. The hollow shapes represent songs in the test set, and the filled-in shapes represent songs in the training set. The test set songs have been labeled with numbers.

Question 1. Suppose you train a 1-nearest neighbor classifier on the training set. For each song in the test set, find its nearest neighbor and draw a line between them. How will the classifier classify each song in the test set? Which labels are correct? (For convenience, we've written in the actual

label of each song.)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
True label	c	c	c	c	c	c	h	h	h	h
Label										
Correct?										

Question 2. What percentage of the test set does the 1-nearest neighbor classifier get right?

Question 3. Draw the *decision boundaries* of the 1-nearest neighbor classifier in the top-left region of the graph (the region to the left and above song 4) and in the bottom-right region of the graph (around song 9).

Question 4. Suppose you train a **3-nearest neighbor** classifier on the training set. For each song in the test set, find its next 2 nearest neighbors and draw a line to them. How will the classifier classify each song in the test set? Which labels are correct? What proportion did we get right this time?

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
True label	c	c	c	c	c	c	h	h	h	h
Label										
Correct?										

Question 5. Why would it be a bad idea to use a 19-nearest neighbor classifier with this dataset?