1. What is the “Curse of Dimensionality” and why is it a major problem in data mining?
2. “Randomly sampling data from a large data set to create a model will always work well.” Do you agree with this statement? Why or why not?
3. Describe how stratified sampling differs from random sampling.
4. A movie reviews website provides ratings on a scale of 1 to 100, where 1 is an unfathomably poor film and 100 is a classic work of art. Provide a discretization scheme that would allow users to quickly gauge whether or not a given movie is worth watching.
5. Why are variance and standard deviation sensitive to outliers? **Hint: think of their relationship to the mean**.
6. What is the mean, median, and IQR of the data set “11 3 1 6 7 5 4 5”? (Show your work to get full credit)

1. Why is it important to provide axis labels when visualizing data?
2. We saw that no single attribute, or pair of attributes, allow us to visually separate the three classes of irises. What could we ask a domain expert to do to help us with this problem?
3. Open iris.arff in Weka. For each attribute in the data set, note if the attribute is continuous or discrete. If the attribute is continuous, list its min, max, mean, and standard deviation. If the attribute is discrete, list each attribute value, its frequency, and if there is a single mode among the attribute values.