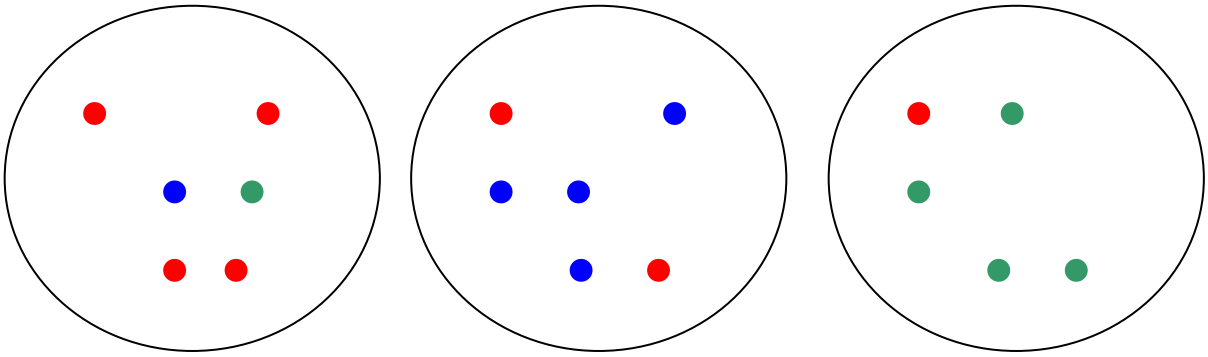
CS 620–Introduction to Data Science, HW5

1) **(30 points)** Consider the following 3 clusters.
   a. Calculate the Purity
   b. Create the contingency table (confusion matrix) and using the contingency table,
      i. Calculate the Rand index
      ii. Calculate the Balanced F measure



2) **(50 points)** Your assignment is to write a program called knn-lastname.py that behaves as follows:
   a. Dataset: iris.csv is a data set describing observed measurements of different flowers belonging to three species of Iris. The four attributes are each continuous measurements, and the label is the species of flower. This data set comes from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
   b. Start with the given sample code segment that implements standard knn classifier using train/test percentage split. Your task is to modify the code to infer the optimal K value using 10-fold cross validation given following conditions,
      i. Use a list of numbers in the range of 1-100, and filter to generate a list called "neighbors" which include only odd numbers in that range.
      ii. Use "cross_val_score" function and specify scoring='accuracy' to generate accuracy from each 10-fold cross validation for the list of "neighbors". Look at sklearn.model_selection.cross_val_score to learn more about the required parameters.
      iii. Perform 10-fold cross validation and generate a list called "MSE" (misclassification error) by using the equation, *MSE = (1- accuracy)*. Note: Here the accuracy is a list that contains the average accuracy of each 10-fold cross validation (per each neighbors).
      iv. Generate a Plot "neighbors" vs "MSE" and also find and print the optimal K using the "MSE" list. Include the plot as a figure in your pdf.

3) **(20 points)** Describe your thoughts about what you think it means to work as a data scientist. You may therefore – if you like – be very personal and describe your own plans and fears for your future career, criticism (or appreciation) for your education, skills you need to develop further, and soon. This question is intended to encourage you to reflect about yourself and your future career, and will therefore be graded generously!

What to turn in:
Each file must exactly follow the naming convention: **Lastname-hw5.zip** should contain following 2 files.

HW5-lastname.pdf
knn-lastname.py