Quiz 9 Solution

1. EDA, curse of dimensionality, computational cost, noise reduction.
2. SVD is a factorization of matrix that generalizes the eigendecomposition of square normal matrix to any m*n matrix.
3. NMF factorized matrix into two matrix: L & R while SVD factorized matrix into three; NMF only contains non-negative elements in L and R while SVD may have negative elements.
4. Biclustering is a data mining technique which allows simultaneous clustering of the rows and columns of a matrix to generate a subset of rows which exhibit similar behavior across a subset of columns, or vice versa.
5. Both are methods of reconstructing the probability density function using a set of given data points.
6. a) k=3

   b) The columns of U tell us something about the people and the columns of V tell us something about the movies. (Similar to what professor covered during the lecture using newsgroup data as an example: columns of U give insights about the documents and columns of V give insights about the words)

   The first column of U is more like a 'scientific movie liker' predictor or tells us how much a person would prefer scientific movie (the first four person are scientific movie likers and it seems the fourth person really likes it). The second column of U is a 'romance movie liker' predictor or how much a person would prefer romance movies (negative values represent higher likeness, for example, the sixth person really likes romance movies).

   The first column of V represents a 'scientific movie selector' which shows the first three movies are more related to 'science'. The second column of V represents a 'romance movie selector' which shows the last two movies are more related to 'romance'.

   In conclusion, using the first two components, we are able to extract almost all the useful information in this data set.

   c) Not very important since the singular value is 1.3