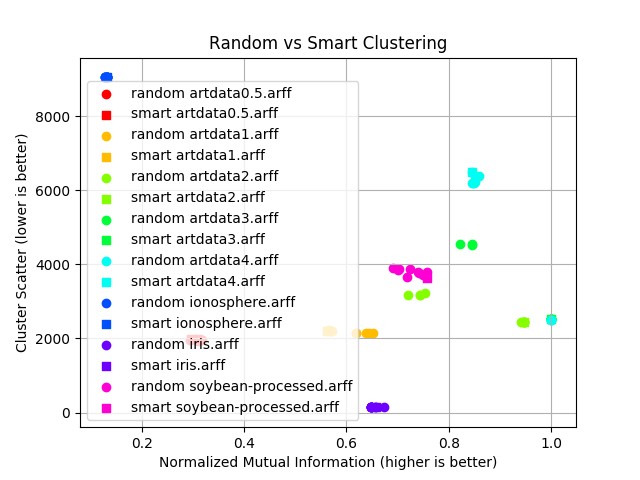
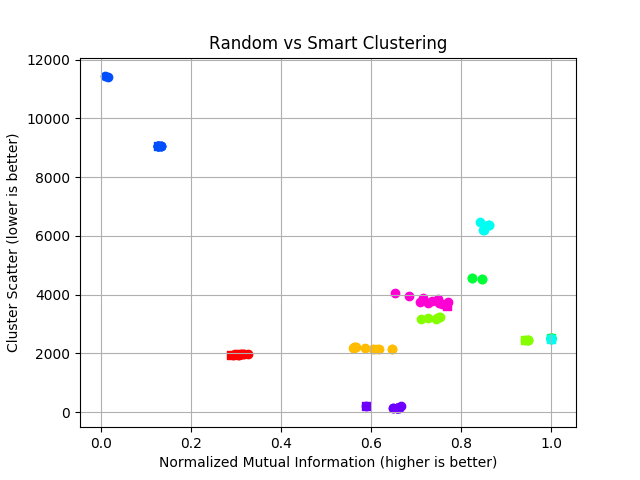
COMP135

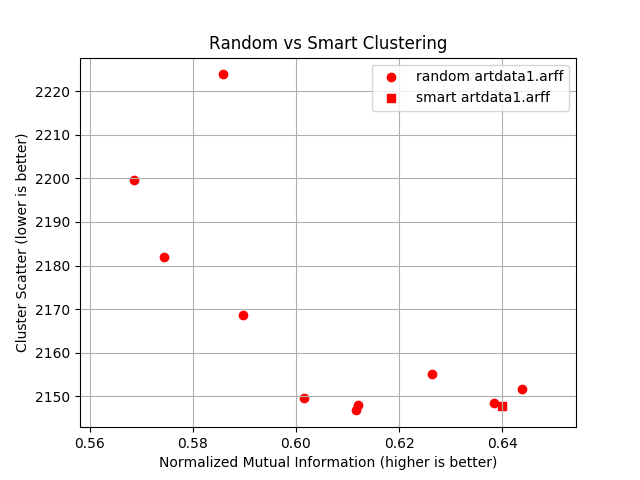
Empirical/Programming Assignment 3

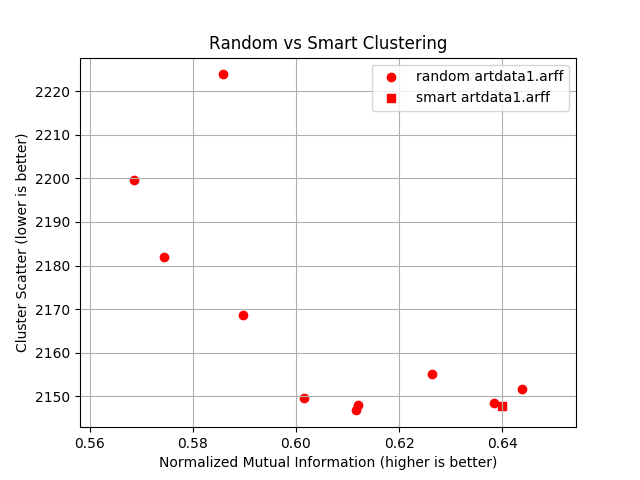
Behnam Heydarshahi

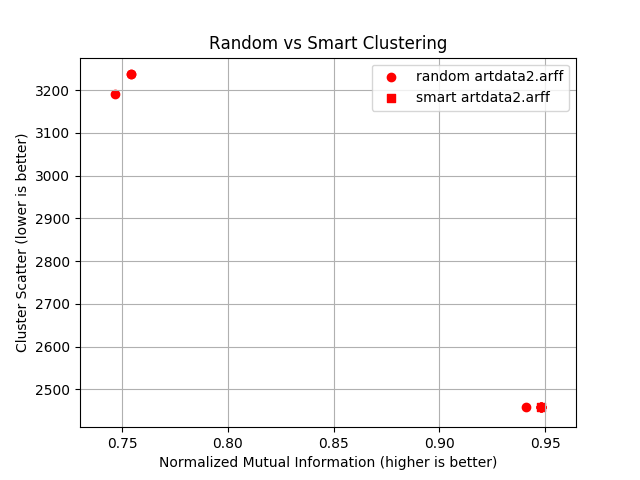
**1. Investigating the Effect of Initialization on k-means**

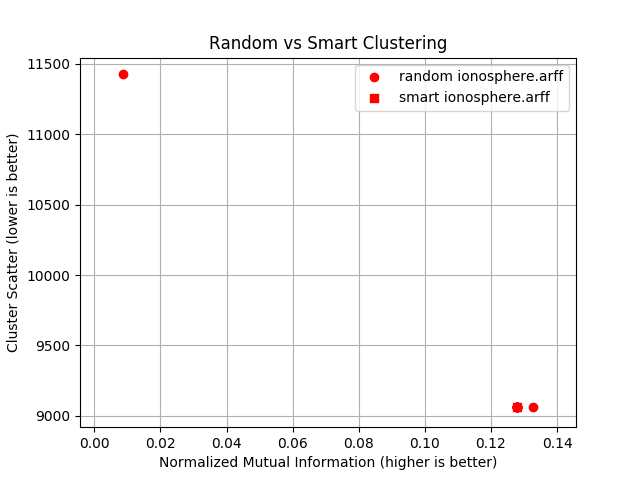


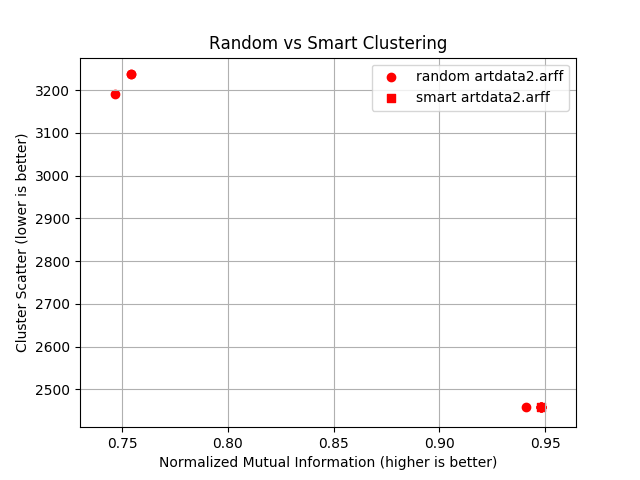


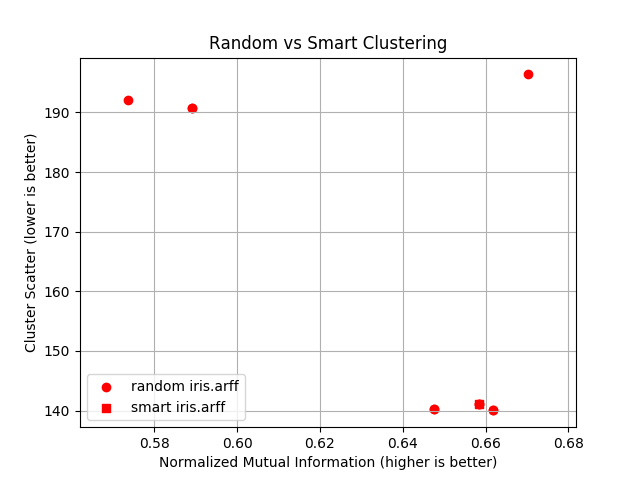
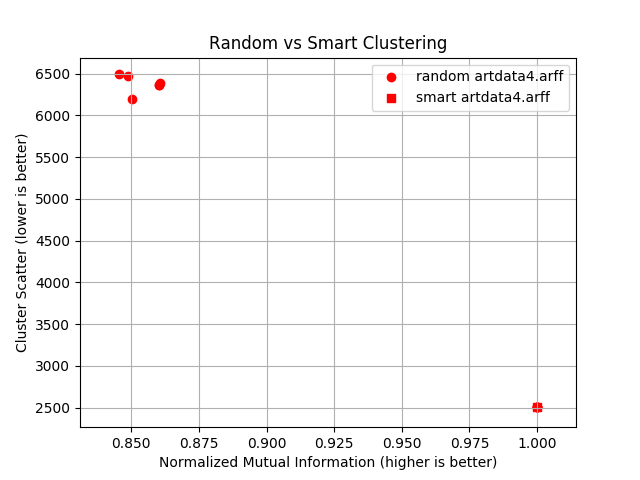
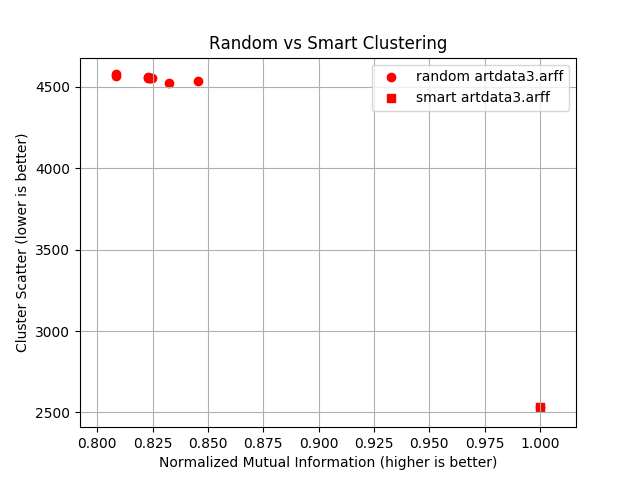
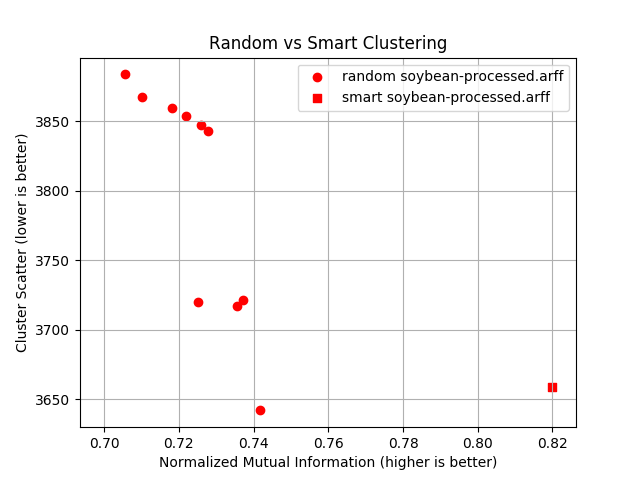
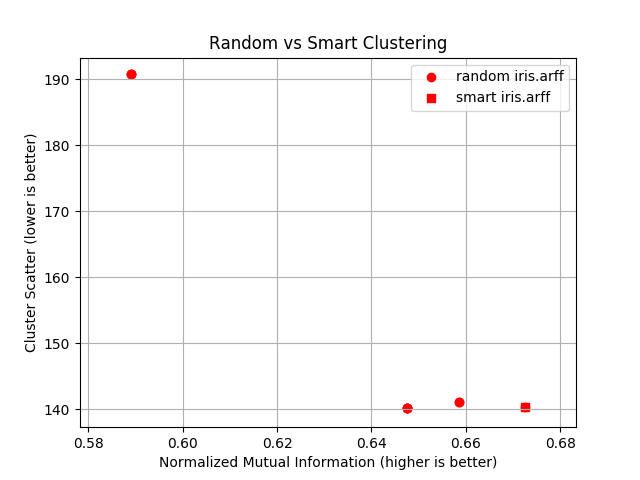
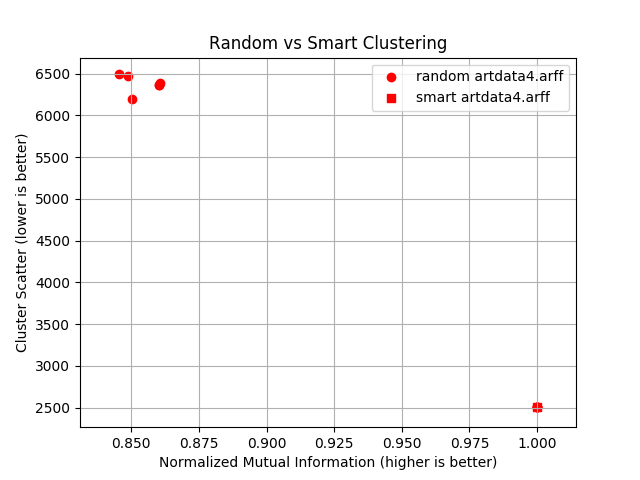
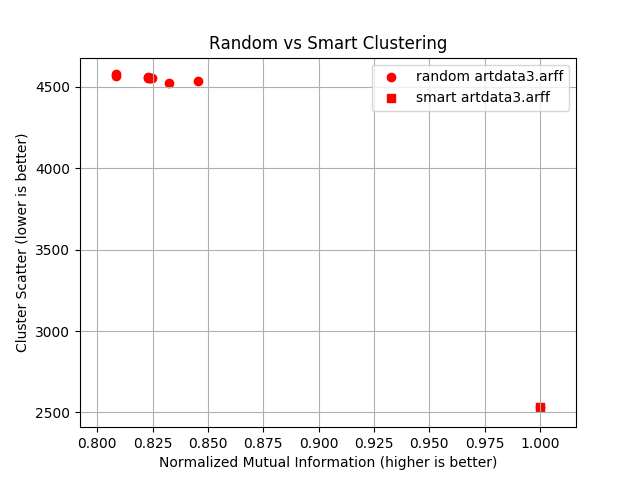
****

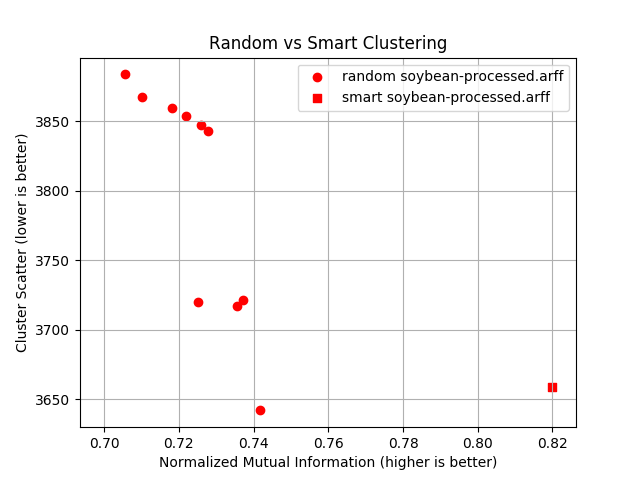












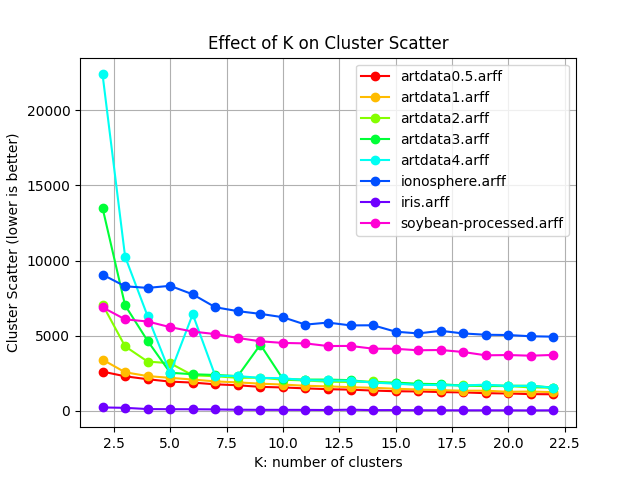
**2. Selecting k**

Is there visual evidence suggesting CS can be used as a criterion for selecting k?

Yes! The derivative of the plots drastically becomes less steep at certain k values. Usually the largest derivative change happens on k = number of different class labels.

What can we conclude from the relative performance on the artificial databases?

Two things:

(1) For the five artificial datasets, 'argdata 0.4' through 'artdata 0.5' performance relatively improves. Therefore we conclude that relative to the spheres' width, the spheres are closer in the 'argdata 4' than 'argdata 3' and so on and so forth.

(2) The variability in performance is much much higher, and datasets with close spheres perform much worse, when the cluster number is smaller than the knee value, which is 5.

Are the results consistent across datasets? What can you observe?

Generally yes. That said, I made these observations:

There are two irregularities in Artdata 4 and Artdata3 graphs, where performance suddenly bumps up for k=6 and k = 9, respectively.

The variability of results on real data is much much smaller, especially when cluster number is small.