CSCI 491: Data Visualization

17- Clustering (visualizing subgroups)

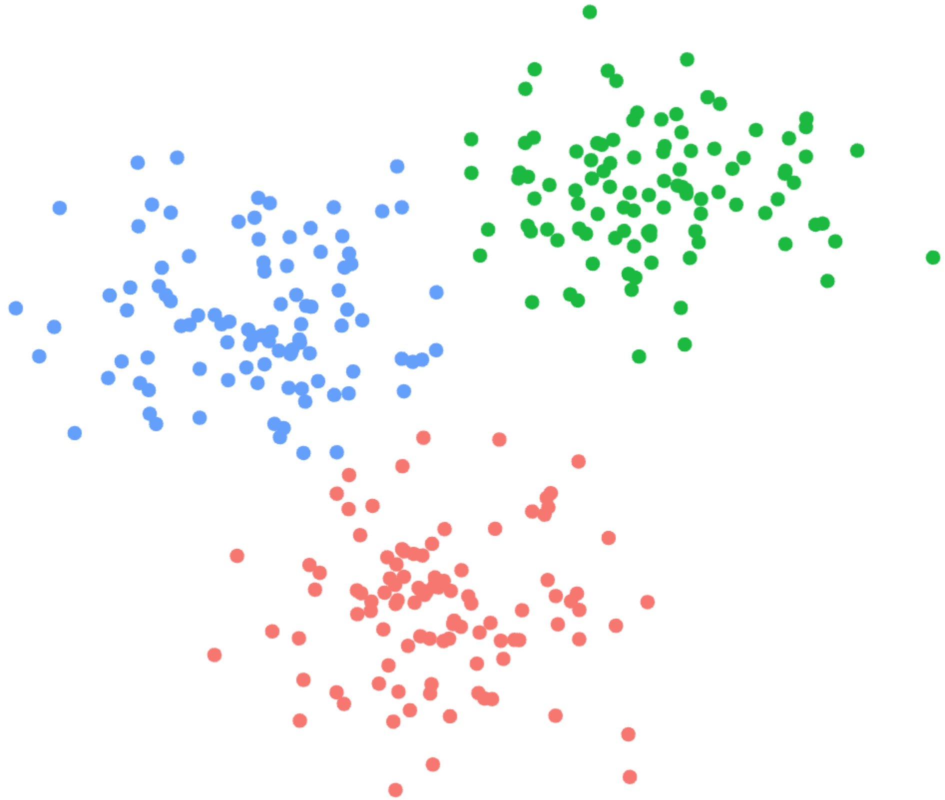
MONTANA STATE UNIVERSITY

Clustering

- These points correspond to three clusters. Can a computer find them automatically?

Clustering

- These points correspond to three clusters. Can a computer find them automatically?



Clustering

- What if we don't know their actual group assignments?
- The goal of clustering is to discover interesting things about the measurements: is there an informative way to visualize the data?
- Can we discover subgroups among the variables or among the observations?
- Clustering is a subcategory of unsupervised learning techniques

Unsupervised Learning

- In Unsupervised learning, we observe only the features because we do not have an associated response variable Y.
- We already discuss principal components analysis, as an unsupervised learning tool used for data visualization.
- Clustering, is a broad class of methods for discovering unknown subgroups in data.

The Challenge of Unsupervised Learning

- Unsupervised learning is more subjective as there is no simple goal for the analysis.
- But techniques for unsupervised learning are of growing importance in a number of fields:
  - subgroups of breast cancer patients grouped by their gene expression measurements
  - groups of shoppers characterized by their browsing and purchase histories,
  - movies grouped by the ratings assigned by movie viewers.

Advantage of Unsupervised learning

- It is often easier to obtain unlabeled data from a lab instrument or a computer than labeled data, which can require human intervention.
- For example it is difficult to automatically assess the overall sentiment of a movie review: is it favorable or not?

Clustering

- Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set.
- We seek a partition of the data into distinct groups so that the observations within each group are quite similar to each other,
- To make this concrete, we must define what it means for two or more observations to be similar or different.
- Indeed, this is often a domain-specific consideration that must be made based on knowledge of the data being studied.

Clustering vs. PCA

- PCA looks for a low-dimensional representation of the observations that explains a good fraction of the variance.
- Clustering looks for homogeneous subgroups among the observations.

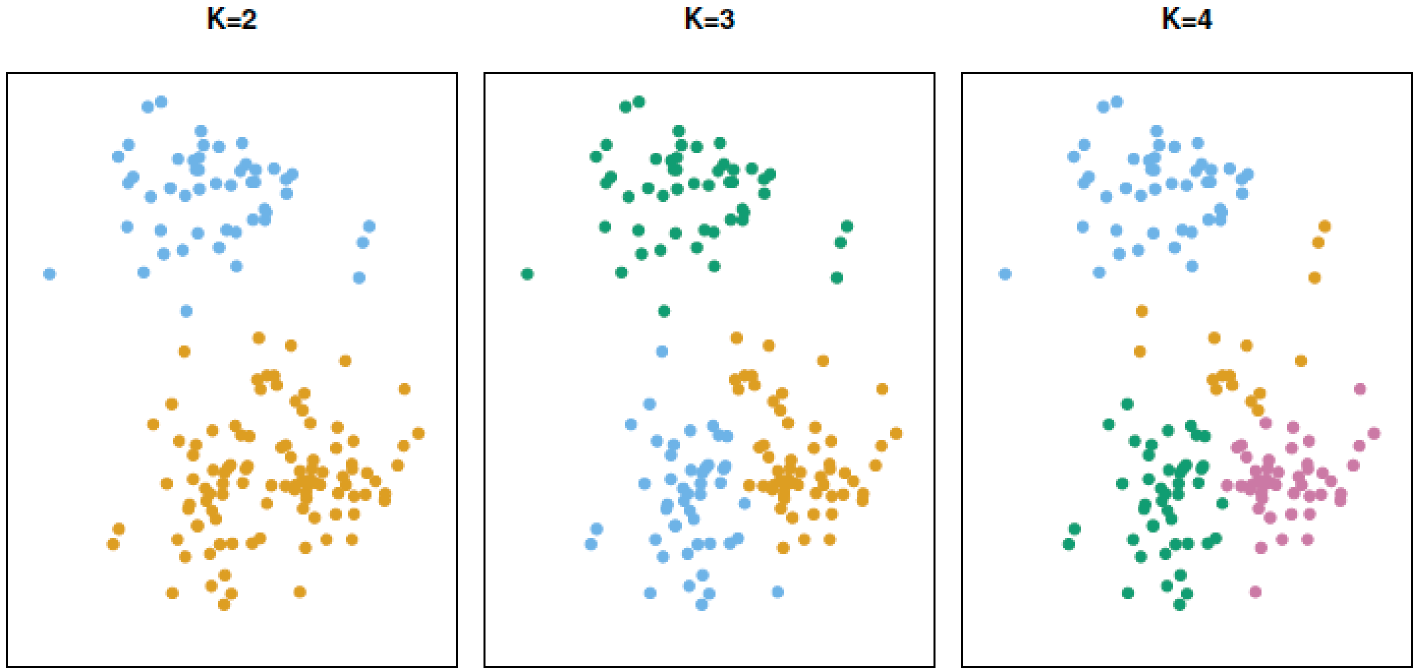
Clustering for Market Segmentation

- Suppose we have access to a large number of measurements (e.g. median household income, occupation, distance from nearest urban area, and so forth) for a large number of people.
- Our goal is to perform market segmentation by identifying subgroups of people who might be more receptive to a particular form of advertising, or more likely to purchase a particular product.
- The task of performing market segmentation amounts to clustering the people in the data set.

Two clustering methods

- In K-means clustering, we seek to partition the observations into a pre-specified number of clusters.
- In hierarchical clustering, we do not know in advance how many clusters we want; in fact, we end up with a tree-like visual representation of the observations, called a dendrogram, that allows us to view at once the clusterings obtained for each possible number of clusters, from 1 to n.

K-means clustering



A simulated data set with 150 observations in 2-dimensional space. Panels show the results of applying K-means clustering with different values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm.

Details of K-means clustering

- The idea behind K-means clustering is that a good clustering is one for which the *within-cluster variation* is as small as possible.
- Hence we want to solve the problem:

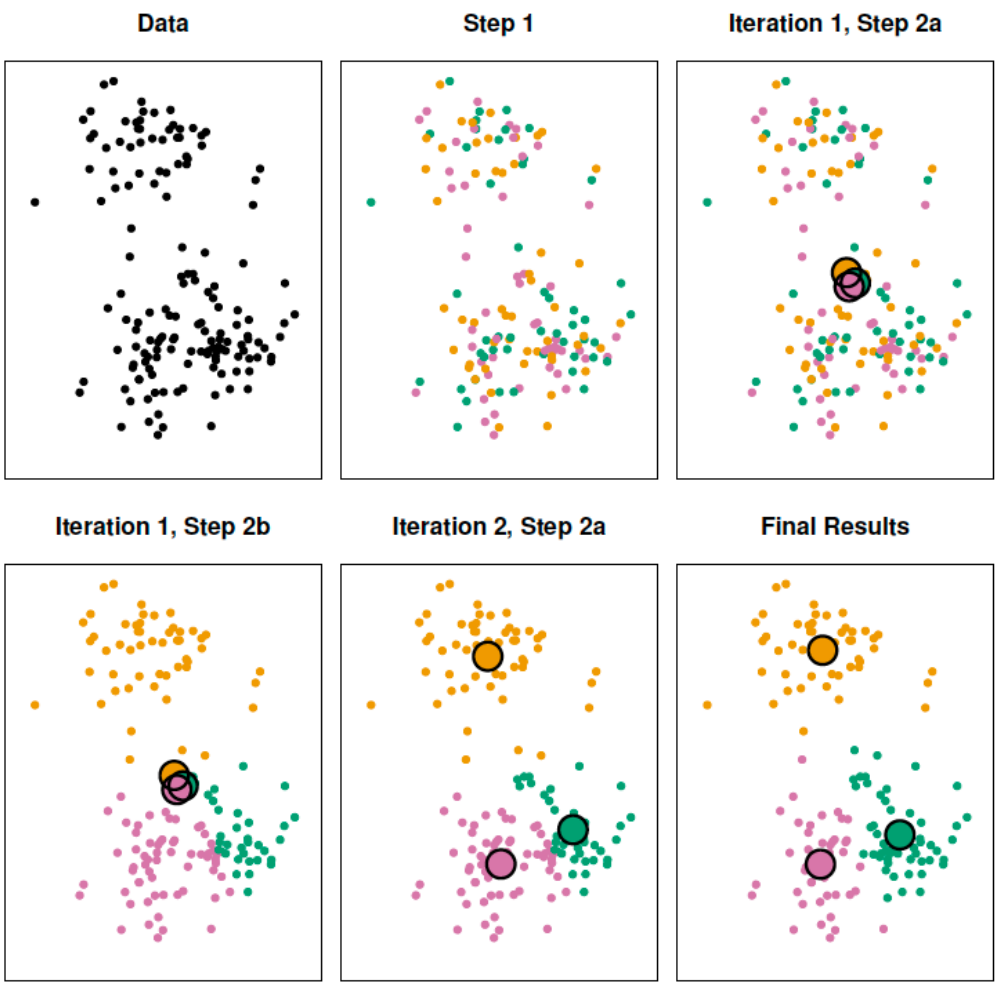
$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \text{WCV}(C_k) \right\}$$

- In words, this formula says that we want to partition the observations into K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible.

K-Means Clustering Algorithm

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   i. For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   ii. Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

Example



Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

Details of Previous Figure

The progress of the K-means algorithm with K=3.

- Top left: The observations are shown.
- Top center: In Step 1 of the algorithm, each observation is randomly assigned to a cluster.
- Top right: In Step 2(a), the cluster centroids are computed. These are shown as large colored disks. Initially the centroids are almost completely overlapping because the initial cluster assignments were chosen at random.
- Bottom left: In Step 2(b), each observation is assigned to the nearest centroid.
- Bottom center: Step 2(a) is once again performed, leading to new cluster centroids.
- Bottom right: The results obtained after 10 iterations.

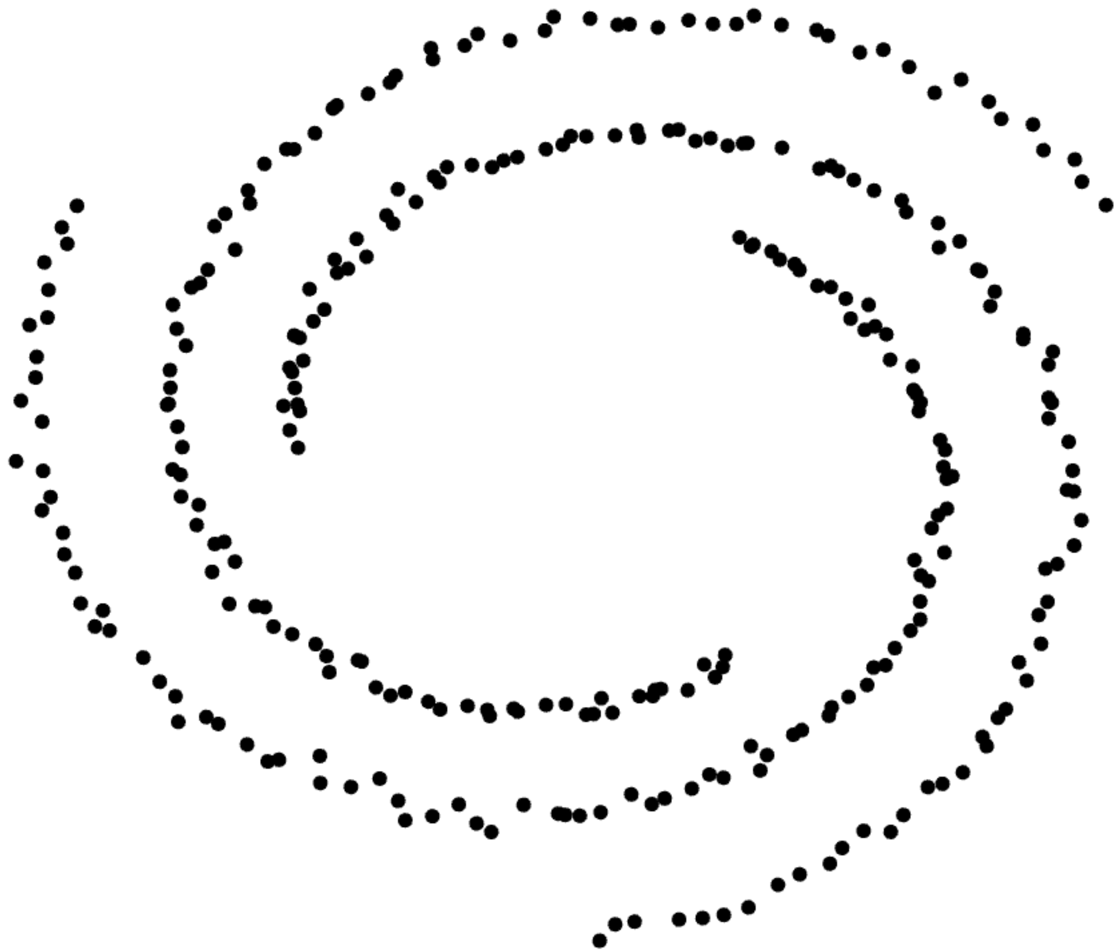
Example: different starting values

Because the K-means algorithm finds a local rather than a global optimum,
the results obtained will depend on the initial (random) cluster assignment of each observation in Step 1

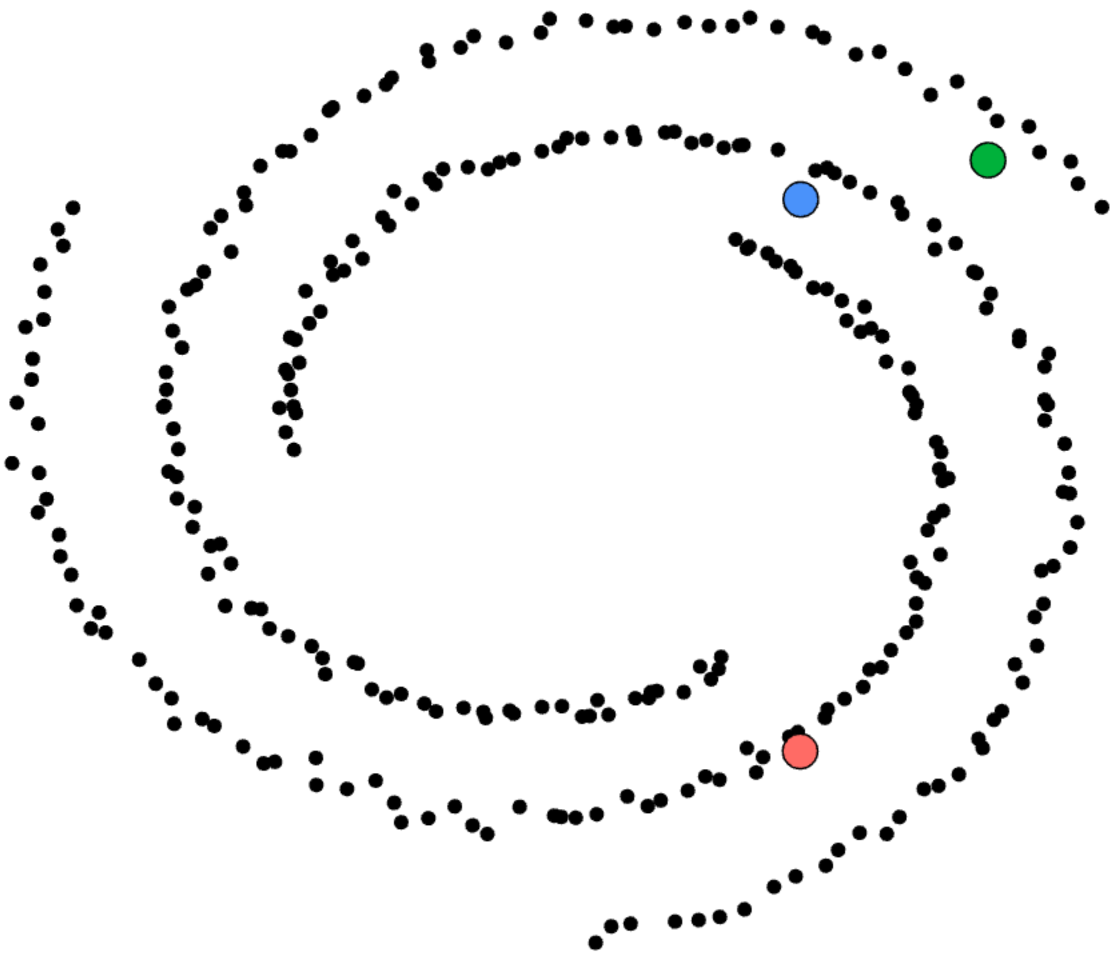
Details of Previous Figure

- K-means clustering performed six times on the data from previous figure with K = 3, each time with a different random assignment of the observations in Step 1 of the K-means algorithm.
- Above each plot is the value of the objective function
- Three different local optima were obtained, one of which resulted in a smaller value of the objective and provides better separation between the clusters.
- Those labeled in red all achieved the same best solution, with an objective value of 235.8

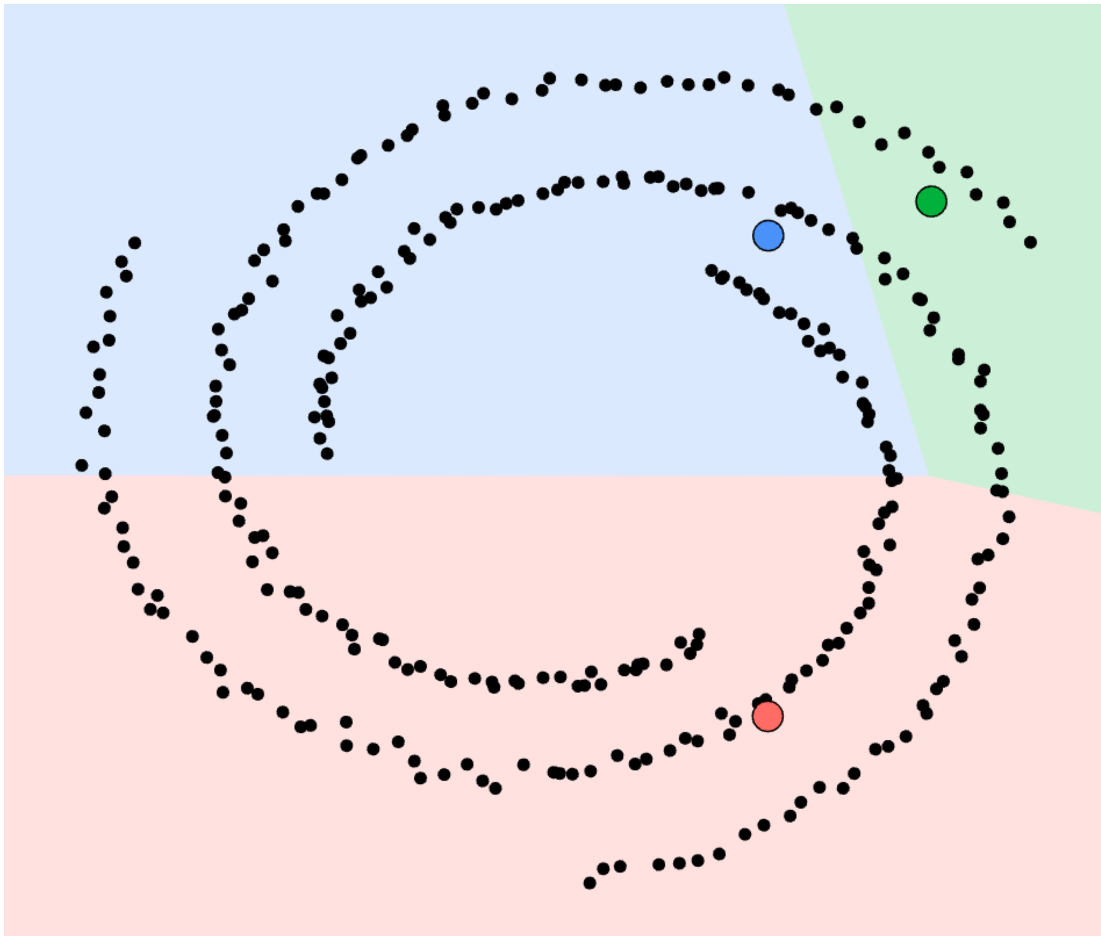
Spiral example



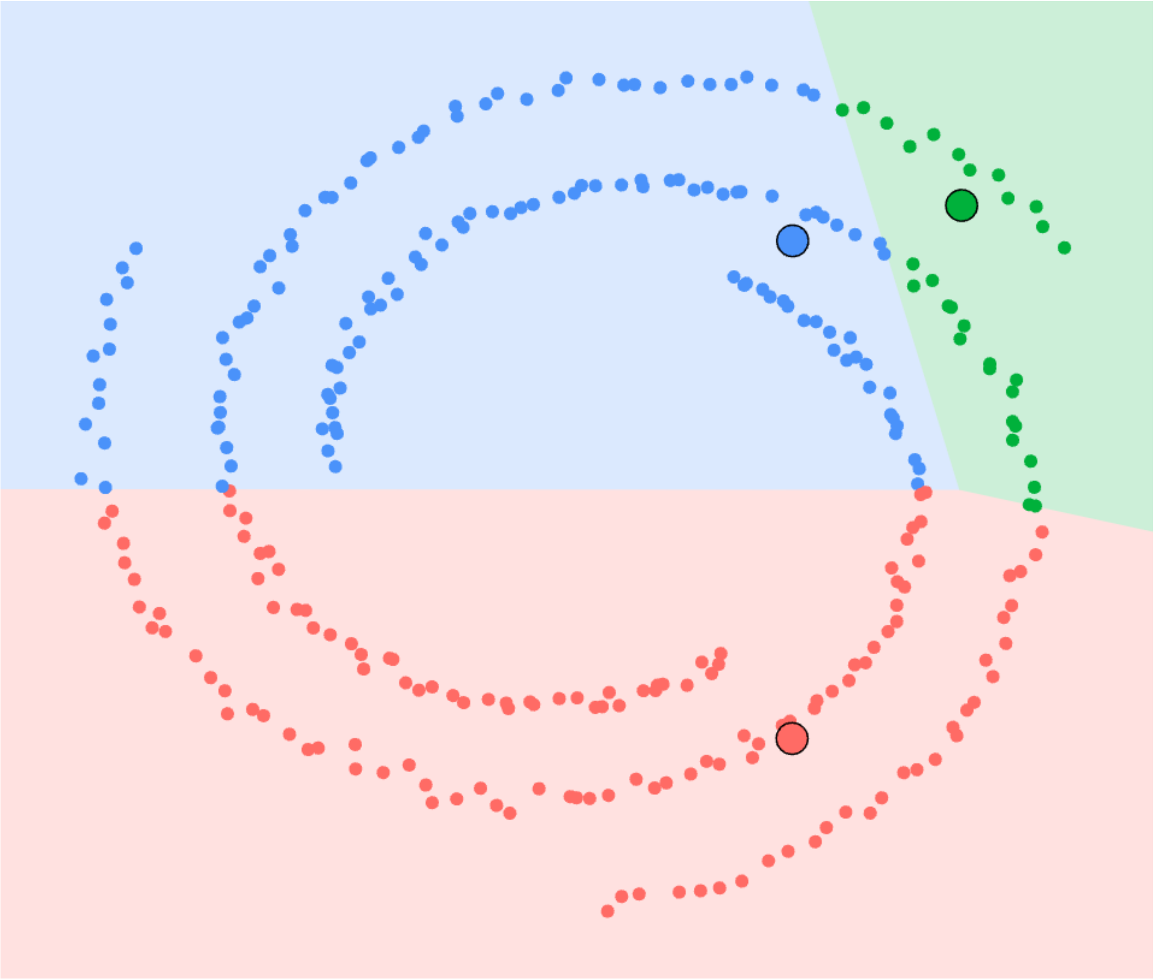
Add means at arbitrary locations



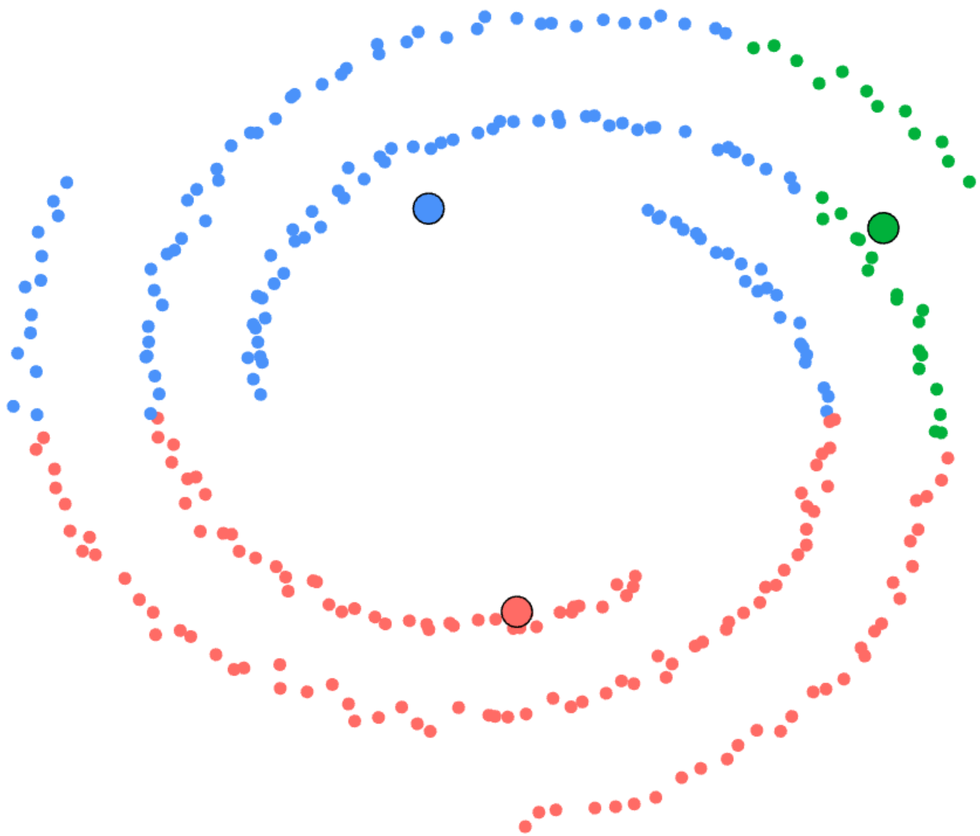
Color data points by the shortest distance to any mean



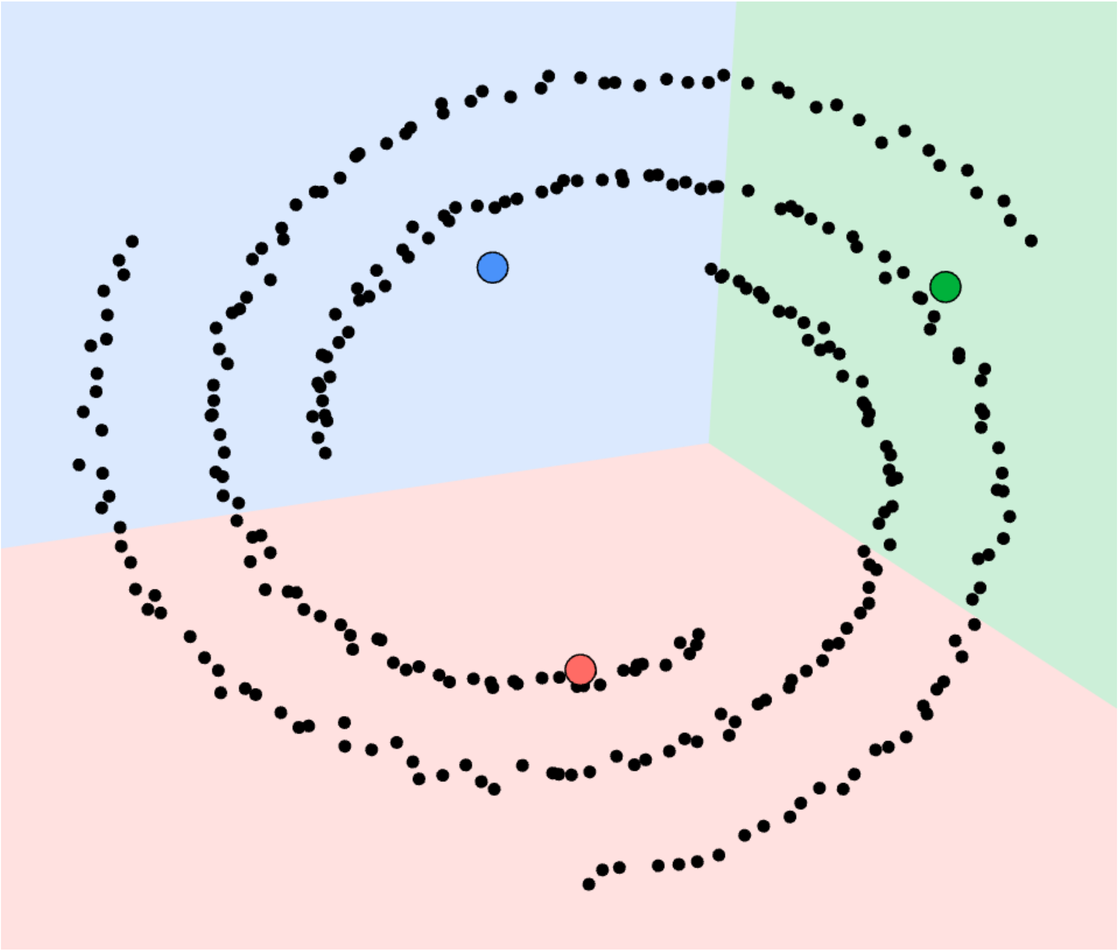
Color data points by the shortest distance to any mean



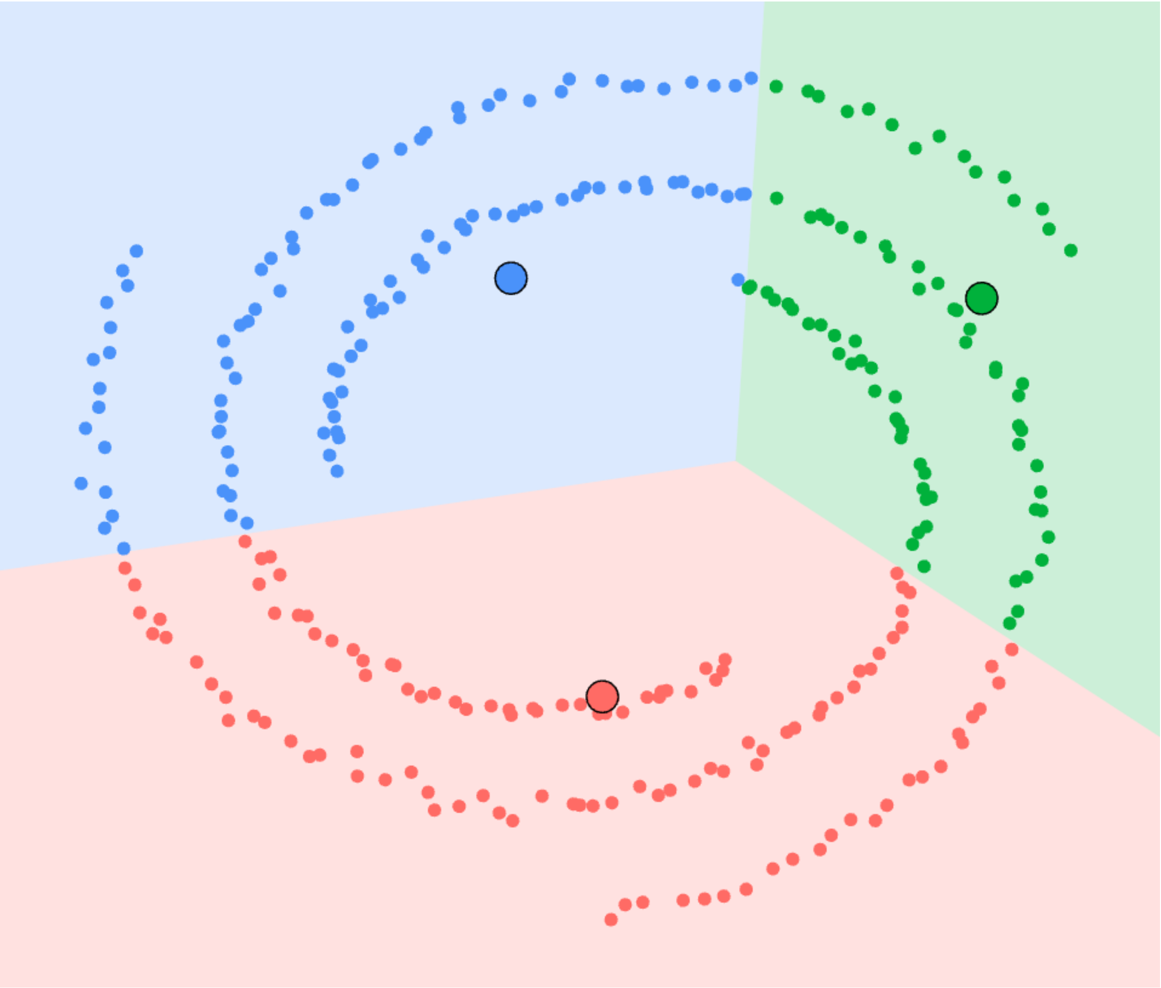
Move means to centroid position of each group of points



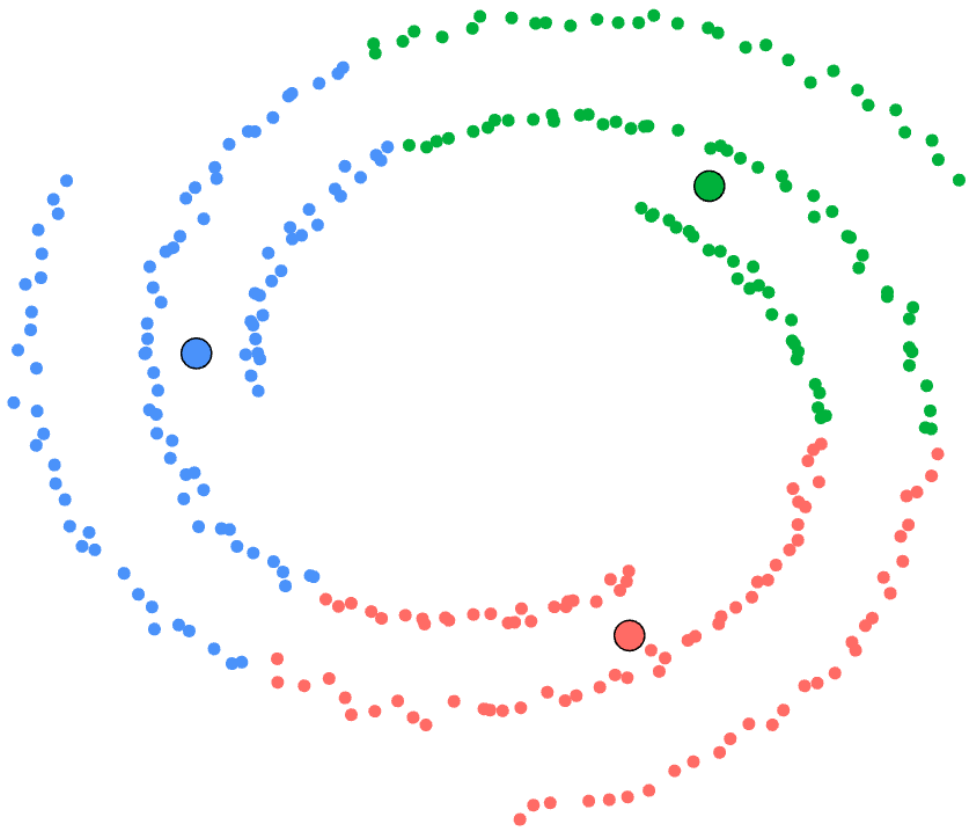
Color data points by the shortest distance to any mean



Color data points by the shortest distance to any mean



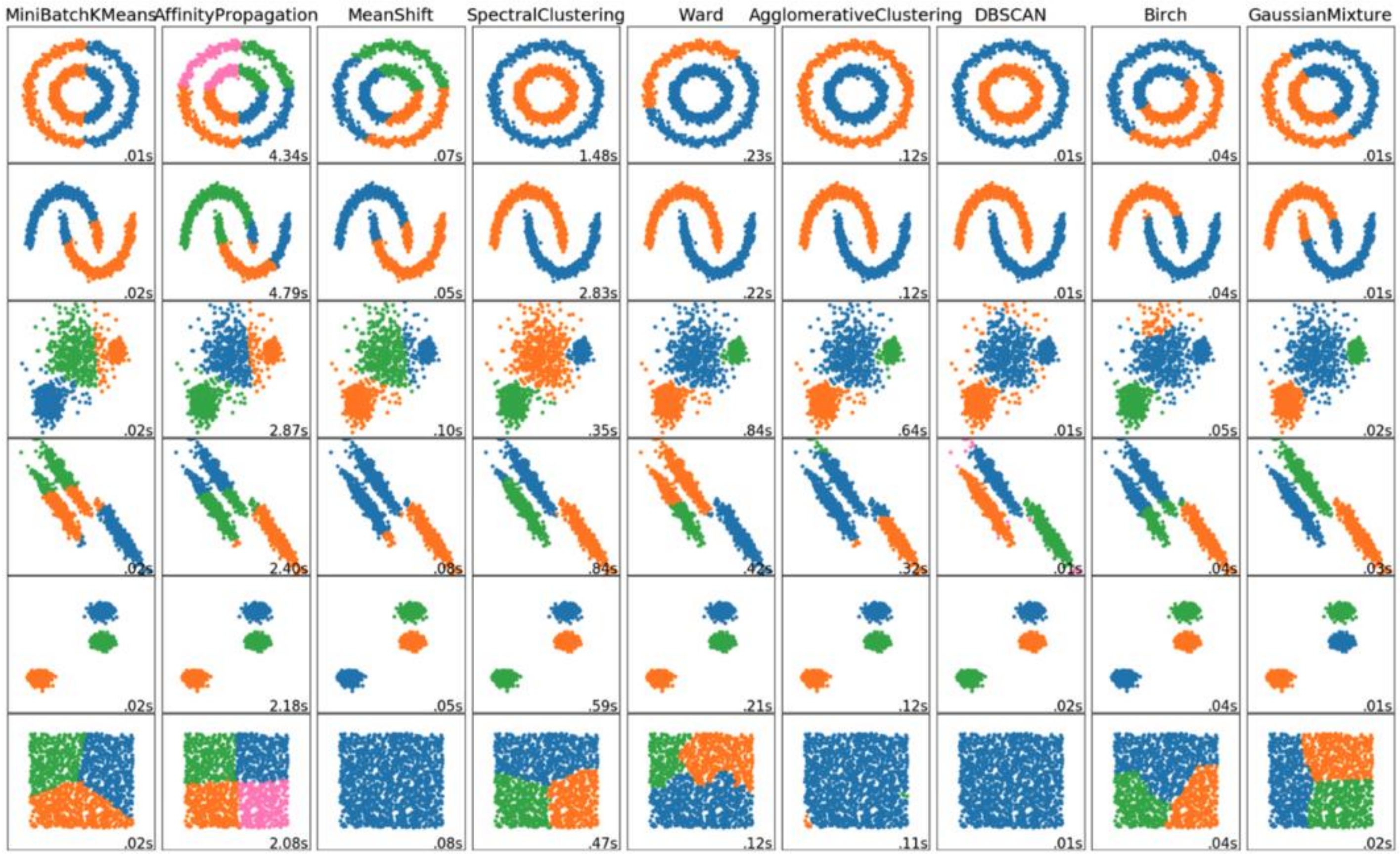
Final Result after Many Iterations



K-means Method Limitations

- We need to specify the number of clusters in advance
- k-means clustering works best when data forms distinct, compact, convex shaped clusters

Alternative clustering algorithms



From George Seif (2018) [The 5 Clustering Algorithms Data Scientists Need to Know]