Data Analysis Course

Cluster Analysis

Venkat Reddy

Contents

- What is the need of Segmentation
- Introduction to Segmentation & Cluster analysis
- Applications of Cluster Analysis
- Types of Clusters
- K-Means clustering

What is the need of segmentation?

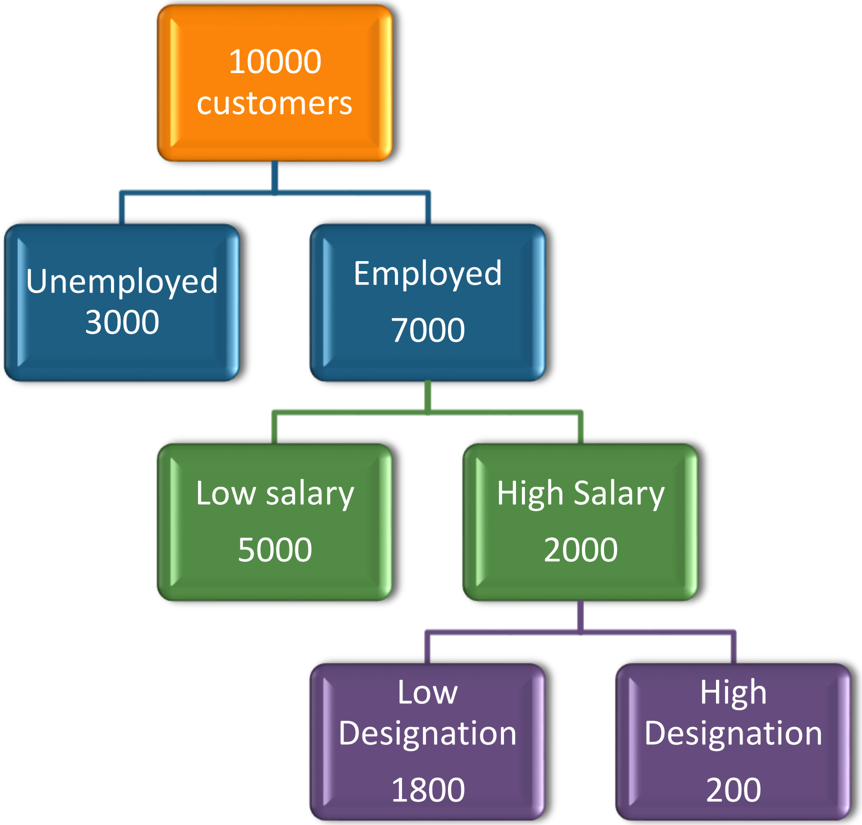
Problem:

- 10,000 Customers - we know their age, city name, income, employment status, designation
- You have to sell 100 Blackberry phones(each costs $1000) to the people in this group. You have maximum of 7 days
- If you start giving demos to each individual, 10,000 demos will take more than one year. How will you sell maximum number of phones by giving minimum number of demos?

What is the need of segmentation?

Solution

- Divide the whole population into two groups employed / unemployed
- Further divide the employed population into two groups high/low salary
- Further divide that group into high /low designation



10000 customers

Unemployed 3000

Employed 7000

Low salary 5000

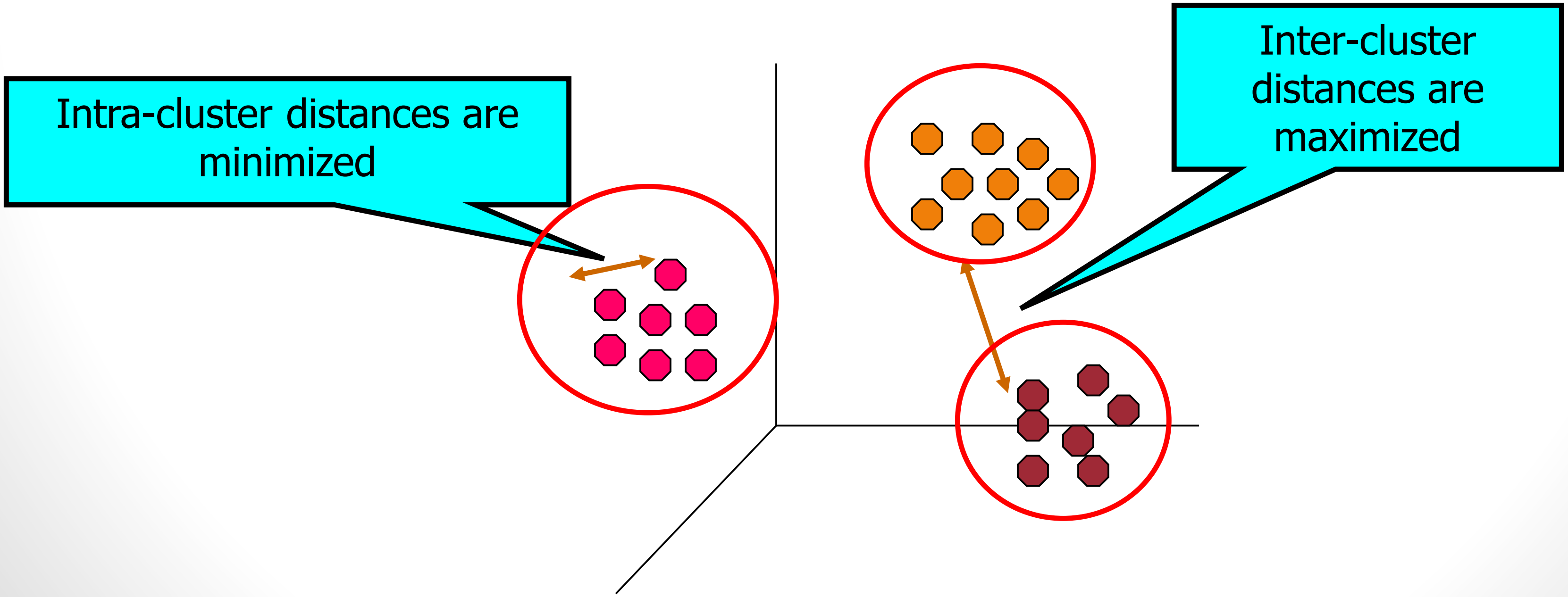
High Salary 2000

Low Designation 1800

High Designation 200

Segmentation and Cluster Analysis

- Cluster is a group of similar objects (cases, points, observations, examples, members, customers, patients, locations, etc)
- Finding the groups of cases/observations/ objects in the population such that the objects are
  - Homogeneous within the group (high intra-class similarity)
  - Heterogeneous between the groups(low inter-class similarity)



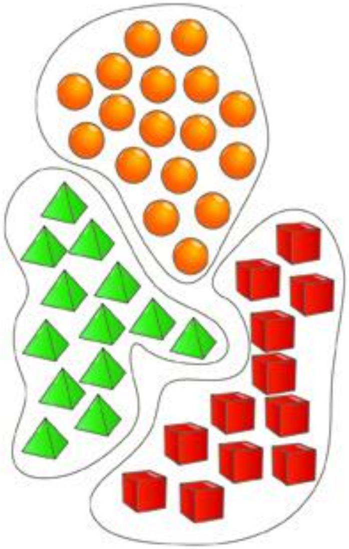
Intra-cluster distances are minimized

Inter-cluster distances are maximized

Applications of Cluster Analysis

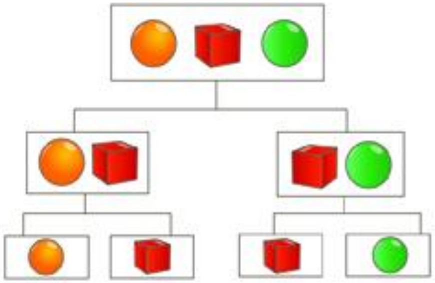
- **Market Segmentation:** Grouping people (with the willingness, purchasing power, and the authority to buy) according to their similarity in several dimensions related to a product under consideration.
- **Sales Segmentation**: Clustering can tell you what types of customers buy what products
- **Credit Risk**: Segmentation of customers based on their credit history
- **Operations**: High performer segmentation & promotions based on person's performance
- **Insurance**: Identifying groups of motor insurance policy holders with a high average claim cost.
- **City-planning:** Identifying groups of houses according to their house type, value, and geographical location
- **Geographical**: Identification of areas of similar land use in an earth observation database.

Types of Clusters



- ***Partitional* clustering or non-hierarchical :** A division of objects into non-overlapping subsets (clusters) such that each object is in exactly one cluster
- The non-hierarchical methods divide a dataset of N objects into M clusters.
- **K-means clustering**, a non-hierarchical technique, is the most commonly used one in business analytics

- ***Hierarchical* clustering:** A set of nested clusters organized as a hierarchical tree
- The hierarchical methods produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains
- **CHAID tree** is most widely used in business analytics

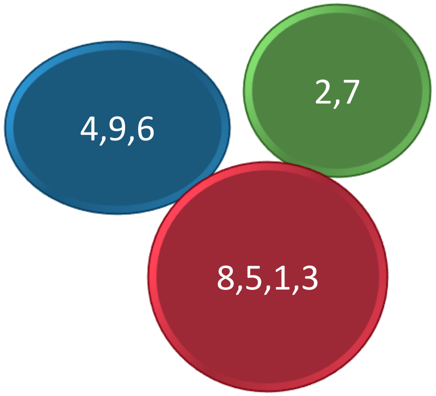


Cluster Analysis -Example

	Maths	Science	Gk	Apt
Student-1	94	82	87	89
Student-2	46	67	33	72
Student-3	98	97	93	100
Student-4	14	5	7	24
Student-5	86	97	95	95
Student-6	34	32	75	66
Student-7	69	44	59	55
Student-8	85	90	96	89
Student-9	24	26	15	22

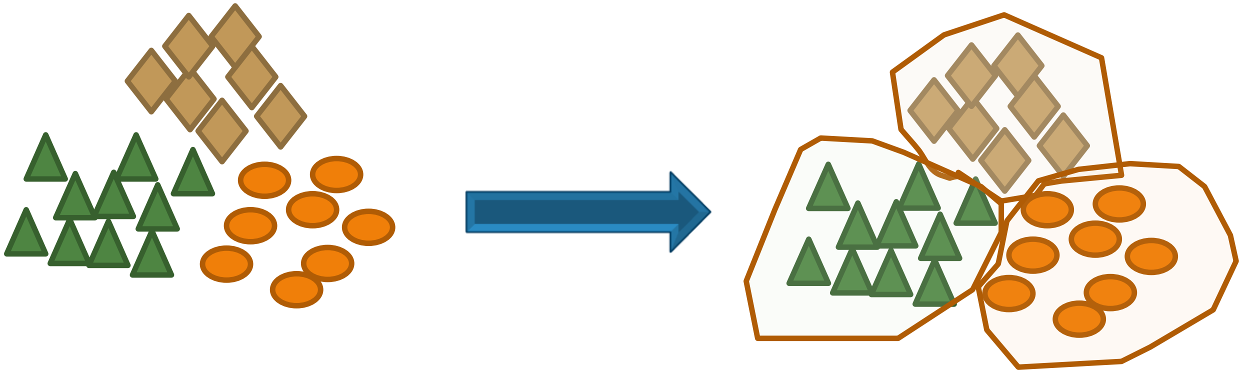
	Maths	Science	Gk	Apt
Student-1	✓ 94	✓ 82	✓ 87	✓ 89
Student-2	! 46	! 67	✗ 33	✓ 72
Student-3	✓ 98	✓ 97	✓ 93	✓ 100
Student-4	✗ 14	✗ 5	✗ 7	✗ 24
Student-5	✓ 86	✓ 97	✓ 95	✓ 95
Student-6	✗ 34	✗ 32	✓ 75	! 66
Student-7	✓ 69	! 44	! 59	! 55
Student-8	✓ 85	✓ 90	✓ 96	✓ 89
Student-9	✗ 24	✗ 26	✗ 15	✗ 22

	Maths	Science	Gk	Apt
Student-4	✗ 14	✗ 5	✗ 7	✗ 24
Student-9	✗ 24	✗ 26	✗ 15	✗ 22
Student-6	✗ 34	✗ 32	✓ 75	! 66
Student-2	! 46	! 67	✗ 33	✓ 72
Student-7	✓ 69	! 44	! 59	! 55
Student-8	✓ 85	✓ 90	✓ 96	✓ 89
Student-5	✓ 86	✓ 97	✓ 95	✓ 95
Student-1	✓ 94	✓ 82	✓ 87	✓ 89
Student-3	✓ 98	✓ 97	✓ 93	✓ 100



Building Clusters

1. Select a **distance measure**
2. Select a **clustering algorithm**
3. Define the **distance between two clusters**
4. Determine the **number of clusters**
5. **Validate** the analysis



- The aim is to build clusters i.e divide the whole population into group of similar objects
- What is similarity/dis-similarity?
- How do you define distance between two clusters

Dissimilarity & Similarity

	Weight
Cust1	68
Cust2	72
Cust3	100

Which two customers are similar?

	Weight	Age
Cust1	68	25
Cust2	72	70
Cust3	100	28

Which two customers are similar now?

	Weight	Age	Income
Cust1	68	25	60,000
Cust2	72	70	9,000
Cust3	100	28	62,000

Which two customers are similar in this case?

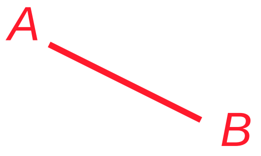
Quantify dissimilarity -Distance measures

- To measure similarity between two observations a distance measure is needed. With a single variable, similarity is straightforward
    - Example: income – two individuals are similar if their income level is similar and the level of dissimilarity increases as the income gap increases
- Multiple variables require an **aggregate distance measure**
    - Many characteristics (e.g. income, age, consumption habits, family composition, owning a car, education level, job…), it becomes more difficult to define similarity with a single value
- The most known measure of distance is the Euclidean distance, which is the concept we use in everyday life for spatial coordinates.

Examples of distances

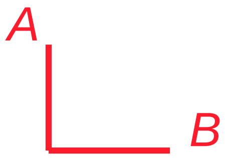
$$D_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ki} - x_{kj} \right)^2}$$

Euclidean distance



$$D_{ij} = \sum_{k=1}^{n} \left| x_{ki} - x_{kj} \right|$$

City-block (Manhattan) distance



D_{ij} distance between cases *i* and *j* x_{kj} - value of variable x_k for case *j*

Other distance measures: Chebychev, Minkowski, Mahalanobis, maximum distance, cosine similarity, simple correlation between observations etc.,

Data matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Dissimilarity matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Calculating the distance

	Weight
Cust1	68
Cust2	72
Cust3	100

- Cust1 vs Cust2 :- (68-72)= **4**
- Cust2 vs Cust3 :- (72-100) = **28**
- Cust3 vs Cust1 :- (100-68) =**32**

	Weight	Age
Cust1	68	25
Cust2	72	70
Cust3	100	28

- Cust1 vs Cust2 :- sqrt((68-72)^2 + (25-70)^2)=**44.9**
- Cust2 vs Cust3 :- **50.54**
- Cust3 vs Cust1 :- **32.14**

Demo: Calculation of distance

```
proc distance data=cust_data out=Dist method=Euclid nostd;
   var interval(Credit_score  Expenses);
run;

proc print data=Dist;
run;
```

Lab: Distance Calculation

```
proc distance data=cust_data out=Count_Dist method=Euclid
nostd;
   var interval(Area_Sq_Miles_ GDP_MM_ Unemp_rate);
run;

proc print data=Count_Dist;
run;
```

Clustering algorithms

- k-means clustering algorithm
- Fuzzy c-means clustering algorithm
- Hierarchical clustering algorithm
- Gaussian(EM) clustering algorithm
- Quality Threshold (QT) clustering algorithm
- MST based clustering algorithm
- Density based clustering algorithm
- kernel k-means clustering algorithm

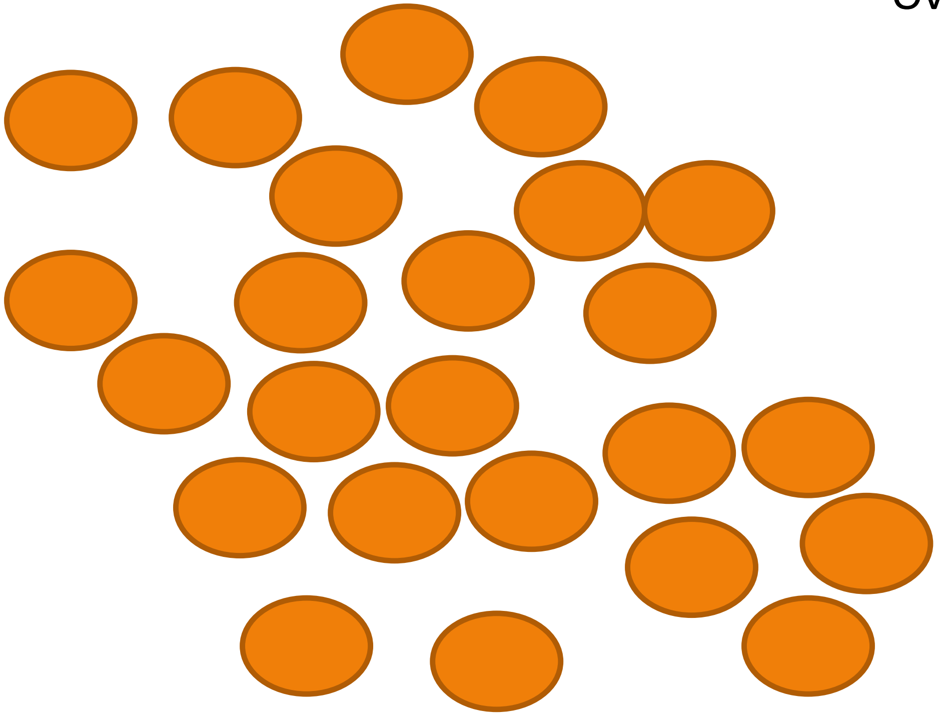
K -Means Clustering – Algorithm

1. The number *k* of clusters is fixed
2. An initial set of *k “seeds” (aggregation centres)* is provided
    1. First *k* elements
    2. Other seeds (randomly selected or explicitly defined)
3. Given a certain fixed threshold, all units are assigned to the nearest cluster seed
4. New seeds are computed
5. Go back to step 3 until no reclassification is necessary

Or simply

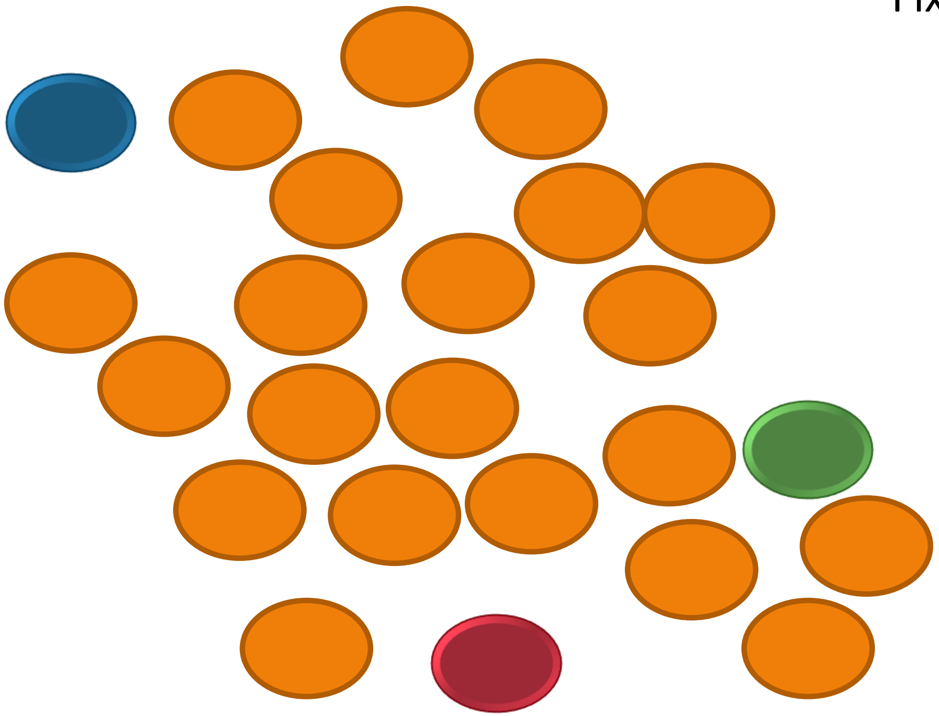
```
Initialize k cluster centers
Do
   Assignment step: Assign each data point to its closest cluster center
   Re-estimation step: Re-compute cluster centers
While (there are still changes in the cluster centers)
```

K-Means clustering



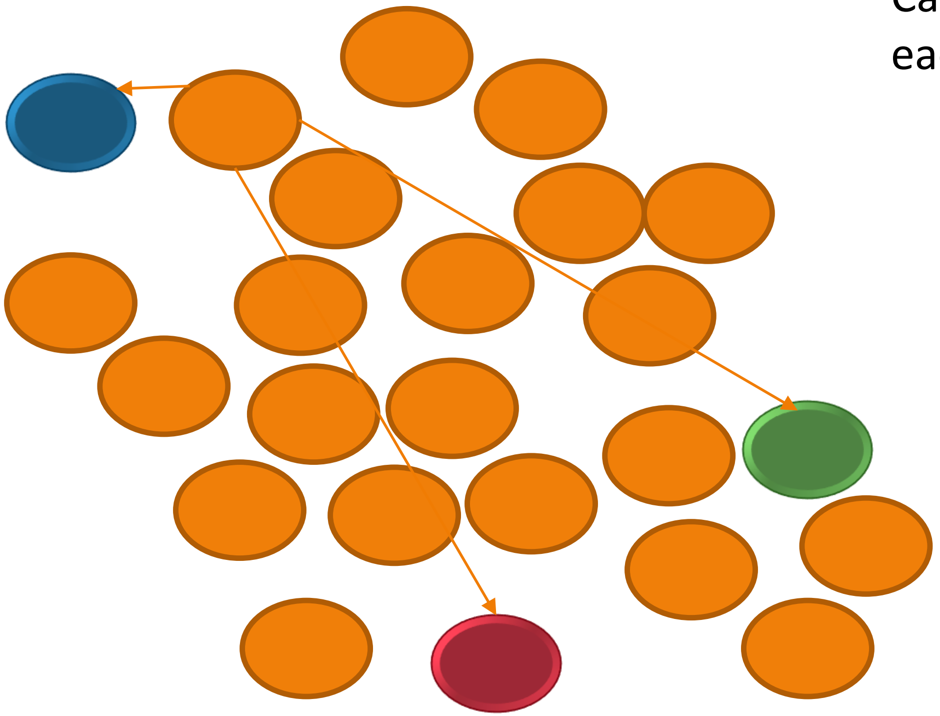
Overall population

K-Means clustering



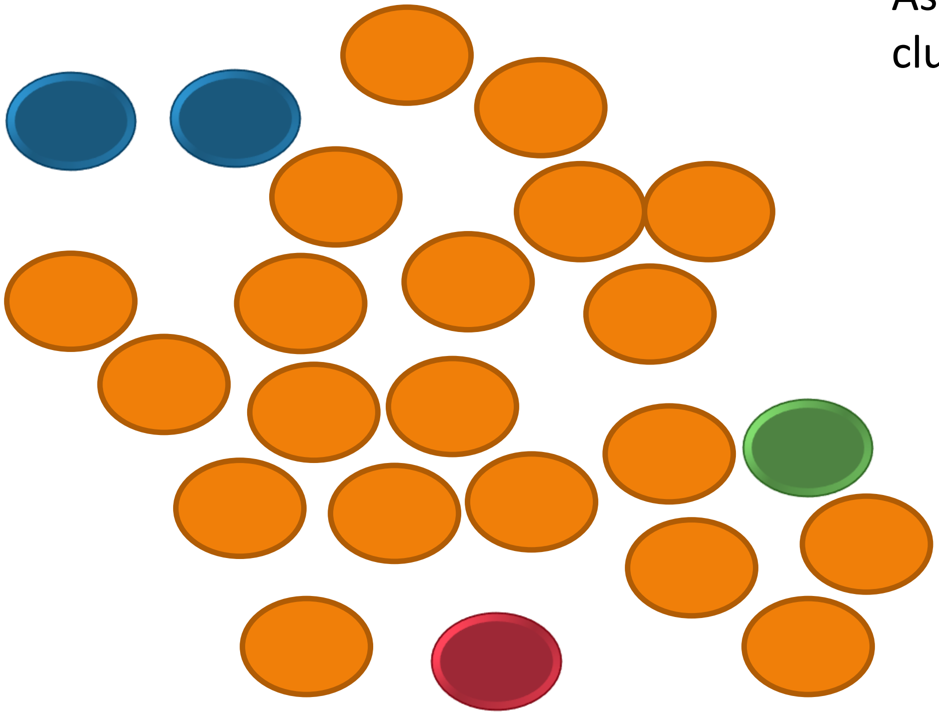
Fix the number of clusters

K-Means clustering



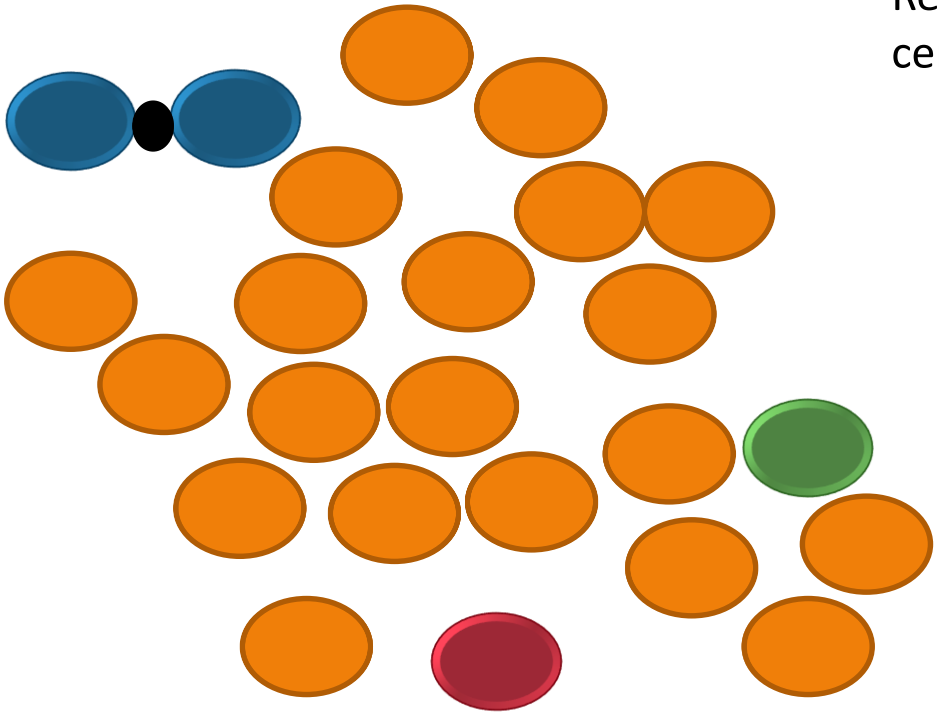
Calculate the distance of each case from all clusters

K-Means clustering



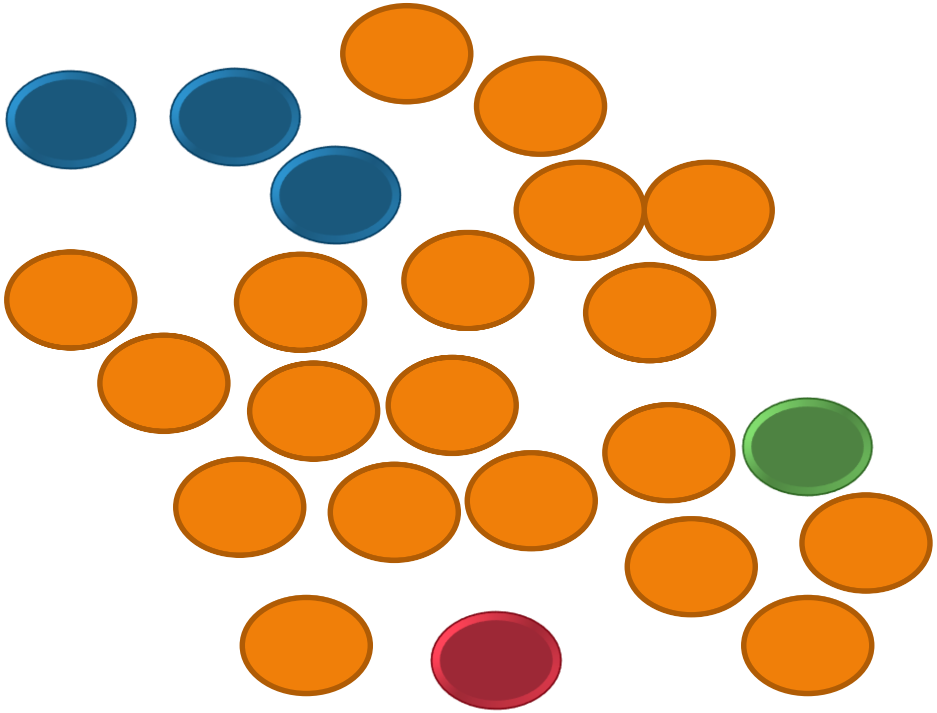
Assign each case to nearest cluster

K-Means clustering

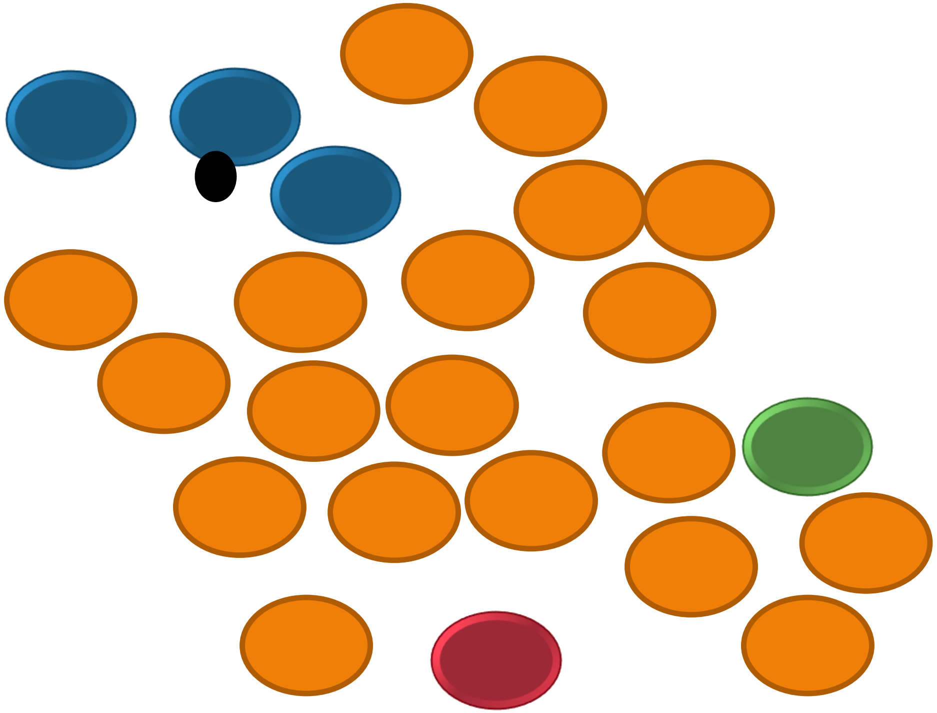


Re calculate the cluster centers

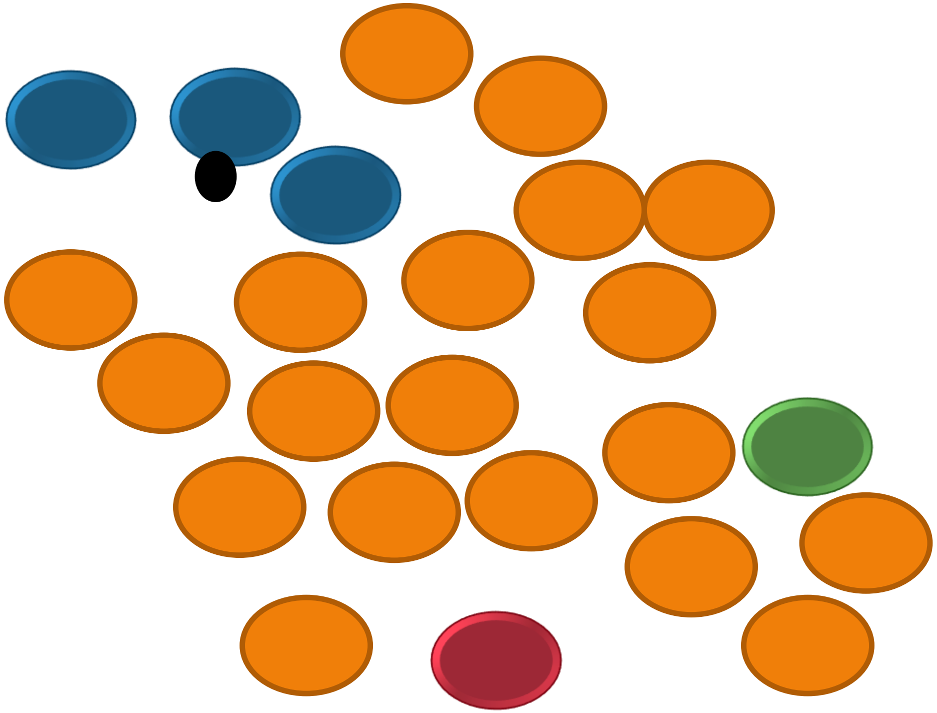
K-Means clustering



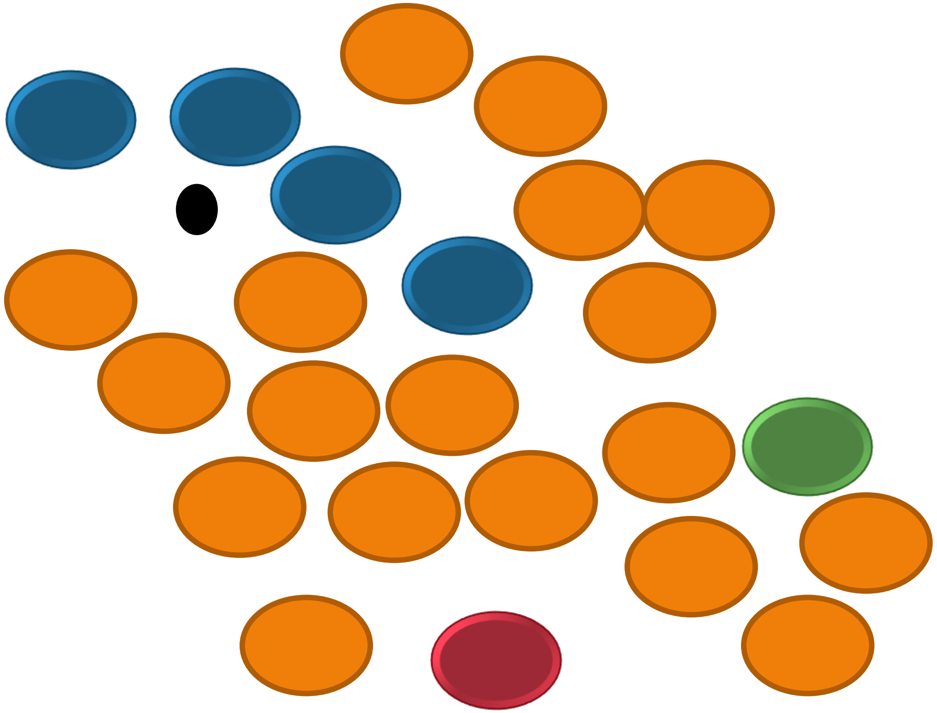
K-Means clustering



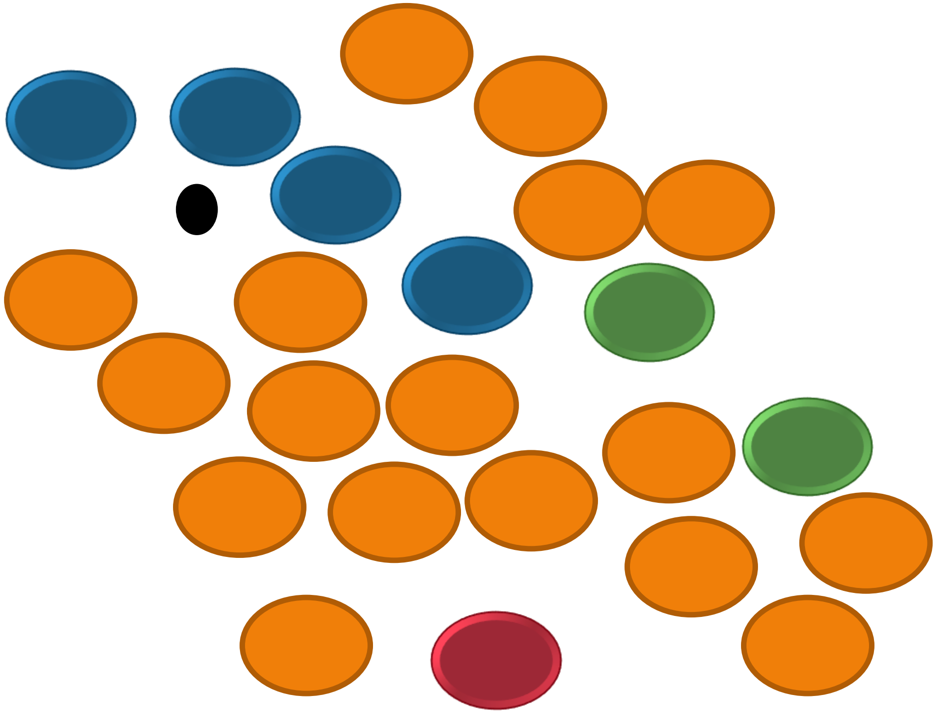
K-Means clustering



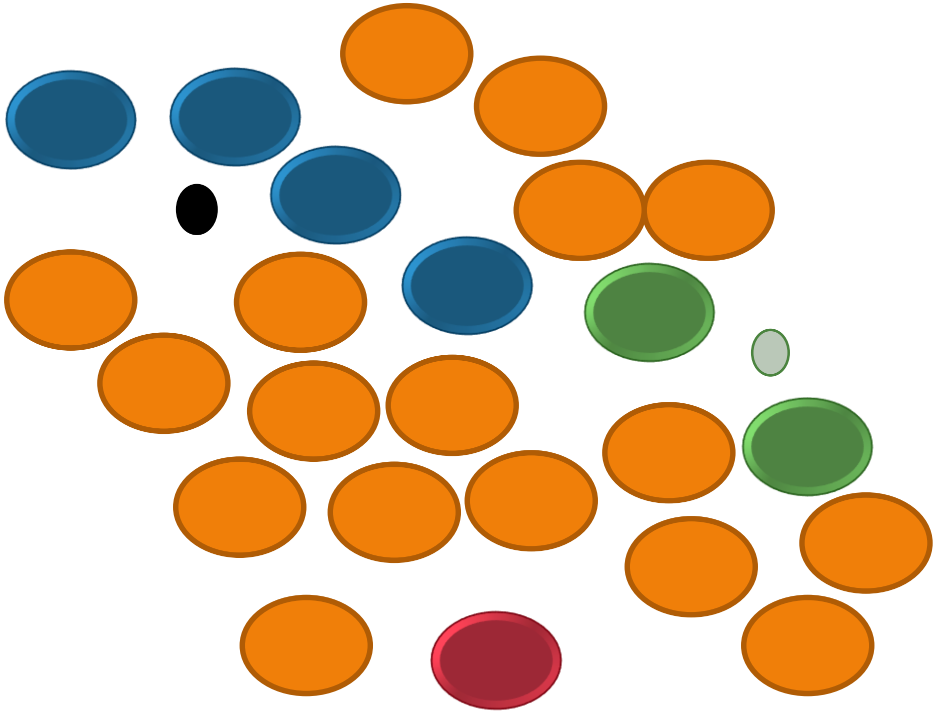
K-Means clustering



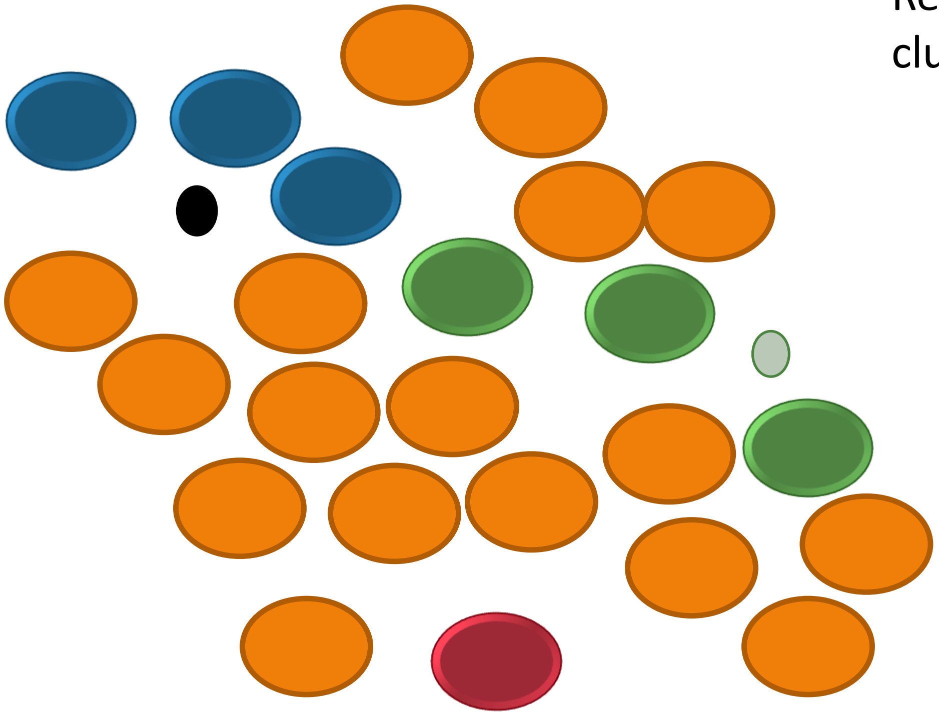
K-Means clustering



K-Means clustering

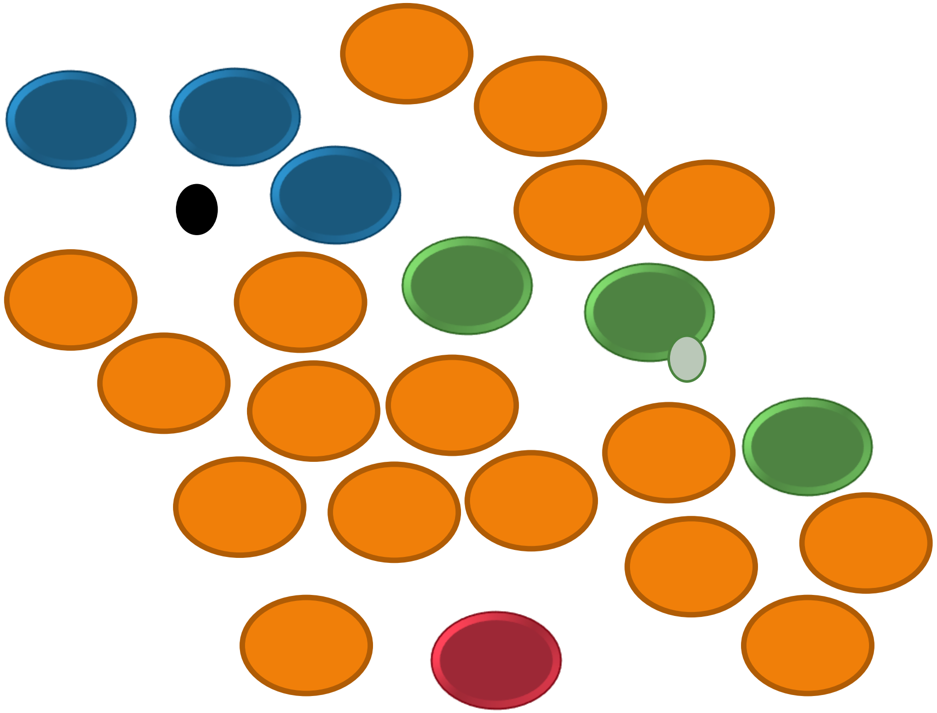


K-Means clustering

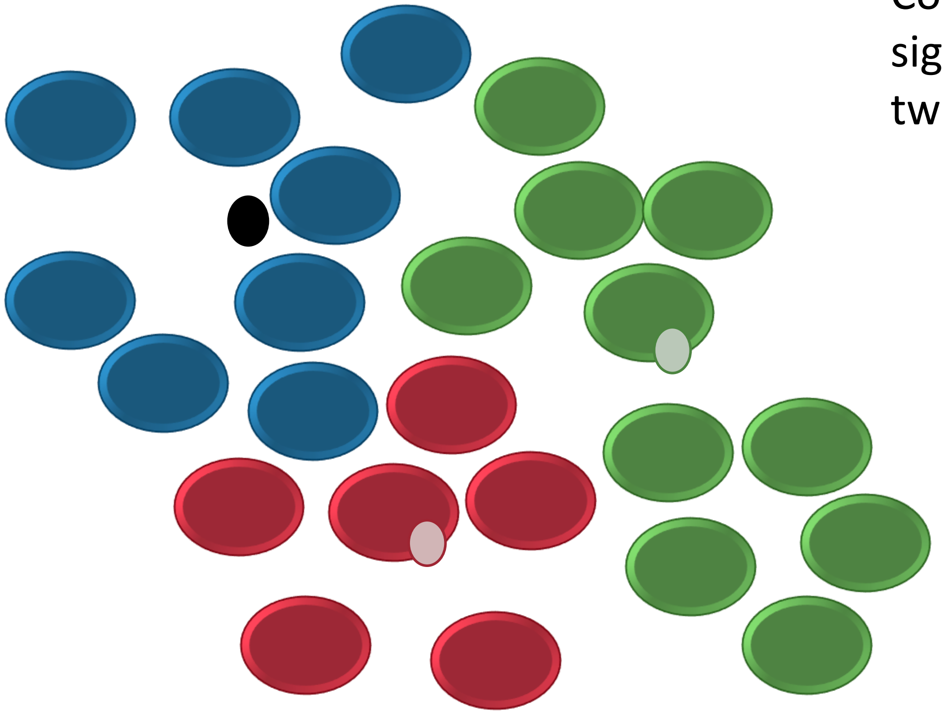


Reassign after changing the cluster centers

K-Means clustering



K-Means clustering

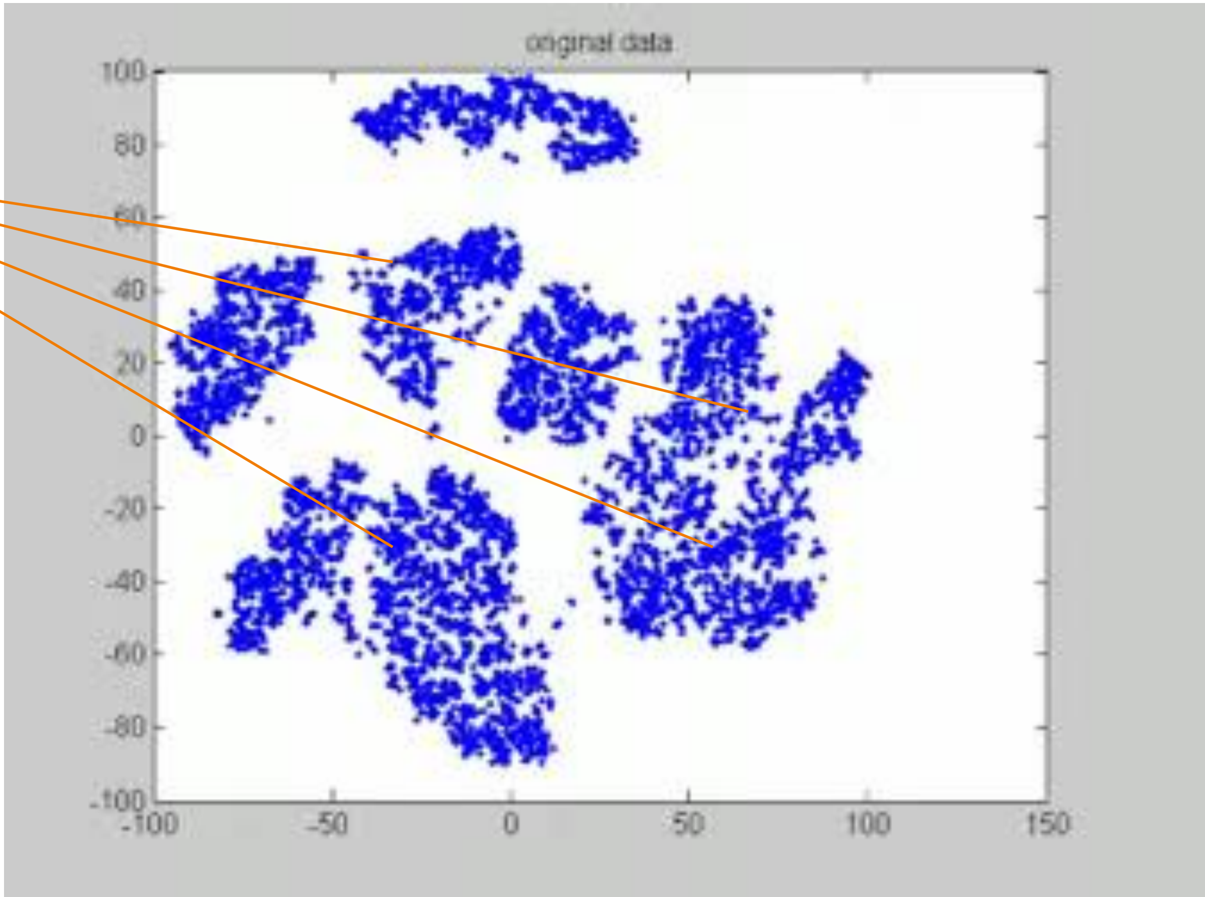


Continue till there is no significant change between two iterations

K Means clustering in action

- Dividing the data into 10 clusters using K-Means

Distance metric will decide cluster for these points



K-Means Clustering SAS Demo

- A Supermarket wanted to send some promotional coupons to 100 families
- The idea is to identify 100 customers with medium income and low recent spends

```
proc fastclus  data= sup_market radius=0 replace=full
maxclusters =5  maxiter =20 distance out=clustr_out;
id cust_id;
Var age  family_size income spend visit_Other_shops;
run;
```

Lab: K- Means Clustering

- Download contact center agents data
- The performance data contains
  - Average handling time
  - Average number of calls
  - CSAT
  - Resolution score
- Identify top 10 agents for promotion based on below criteria
  - High C_SAT
  - High Resolution
  - Low Average handling time
  - High number of calls

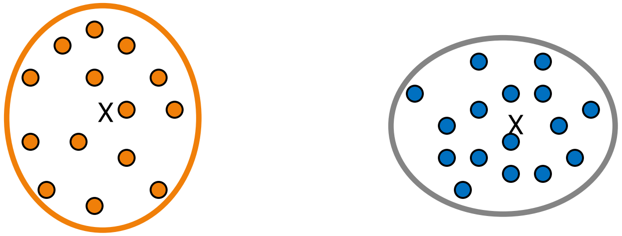
SAS Code Options

- The **RADIUS**= option establishes the minimum distance criterion for selecting new seeds. No observation is considered as a new seed unless its minimum distance to previous seeds exceeds the value given by the RADIUS= option. The default value is 0.
- The **MAXCLUSTERS**= option specifies the maximum number of clusters allowed. If you omit the MAXCLUSTERS= option, a value of 100 is assumed.
- The REPLACE= option specifies how seed replacement is performed.
  - FULL :requests default seed replacement.
  - PART :requests seed replacement only when the distance between the observation and the closest seed is greater than the minimum distance between seeds.
  - NONE : suppresses seed replacement.
  - RANDOM :Selects a simple pseudo-random sample of complete observations as initial cluster seeds.

SAS Code & Options

- The **MAXITER**= option specifies the maximum number of iterations for re computing cluster seeds. When the value of the MAXITER= option is greater than 0, each observation is assigned to the nearest seed, and the seeds are recomputed as the means of the clusters.
- The **LIST** option lists all observations, giving the value of the ID variable (if any), the number of the cluster to which the observation is assigned, and the distance between the observation and the final cluster seed.
- The **DISTANCE** option computes distances between the cluster means.
- The **ID** variable, which can be character or numeric, identifies observations on the output when you specify the LIST option.
- The **VAR** statement lists the numeric variables to be used in the cluster analysis. If you omit the VAR statement, all numeric variables not listed in other statements are used.

Distance between Clusters



- **Single link:** smallest distance between an element in one cluster and an element in the other, i.e., dist(K_i, K_j) = min(t_{ip}, t_{jq})
- **Complete link:** largest distance between an element in one cluster and an element in the other, i.e., dist(K_i, K_j) = max(t_{ip}, t_{jq})
- **Average:** avg distance between an element in one cluster and an element in the other, i.e., dist(K_i, K_j) = avg(t_{ip}, t_{jq})
- **Centroid:** distance between the centroids of two clusters, i.e., dist(K_i, K_j) = dist(C_i, C_j)
- **Medoid:** distance between the medoids of two clusters, i.e., dist(K_i, K_j) = dist(M_i, M_j) Medoid: a chosen, centrally located object in the cluster

SAS output interpretation

- **RMSSTD** - Pooled standard deviation of all the variables forming the cluster.(Variance within a cluster) Since the objective of cluster analysis is to form homogeneous groups, the
  - RMSSTD of a cluster should be as small as possible

- **SPRSQ** -Semipartial R-squared is a measure of the homogeneity of merged clusters, so SPRSQ is the loss of homogeneity due to combining two groups or clusters to form a new group or cluster. (error incurred by combining two groups)
  - Thus, the SPRSQ value should be small to imply that we are merging two homogeneous groups

SAS output interpretation

- **RSQ** (R-squared) measures the extent to which groups or clusters are different from each other. (Variance between the clusters)
  - So, when you have just one cluster RSQ value is, intuitively, zero). Thus, the RSQ value should be high.
- **Centroid Distanc**e is simply the Euclidian distance between the centroid of the two clusters that are to be joined or merged.
  - So, Centroid Distance is a measure of the homogeneity of merged clusters and the value should be small.

Distance Calculation on standardized data

	Weight	Income
Cust1	68	60,000
Cust2	72	9,000
Cust3	100	62,000

Average	80	43667
Stdev	14	24527

	Weight	Income
Cust1	-0.84	0.67
Cust2	-0.56	-1.41
Cust3	1.40	0.75