ФИСТ
УлГТУ

Machine learning
Lesson 08. Clustering

Kirill Svyatov
Ulyanovsk State Technical University,
Faculty of Information Systems and Technologies

What is the need of segmentation?

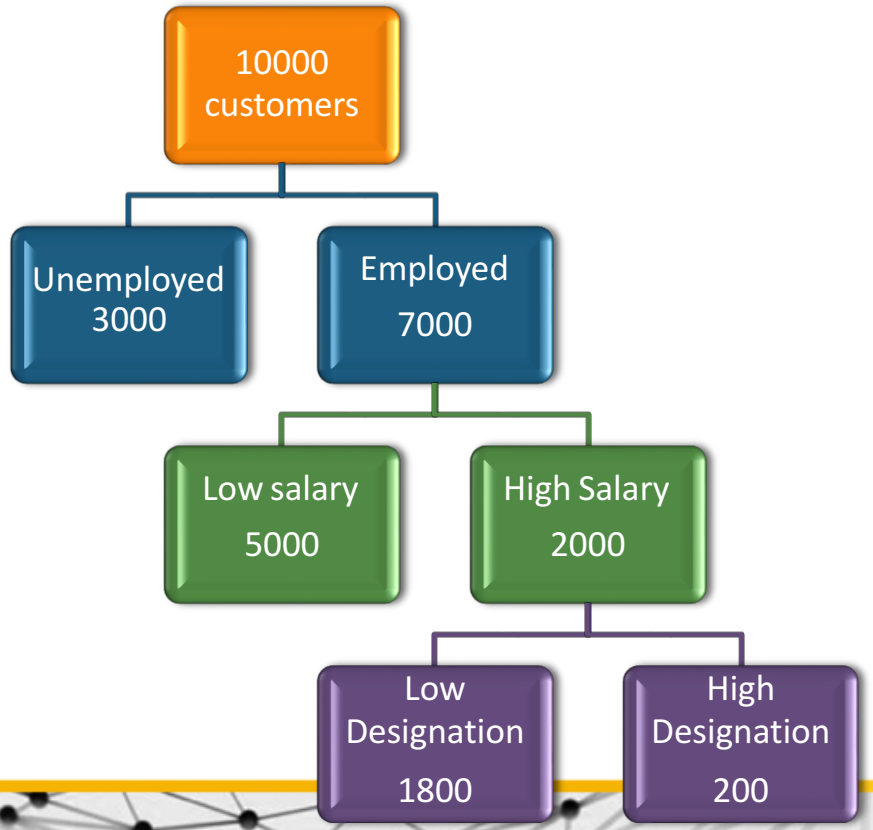
Problem:

- 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?

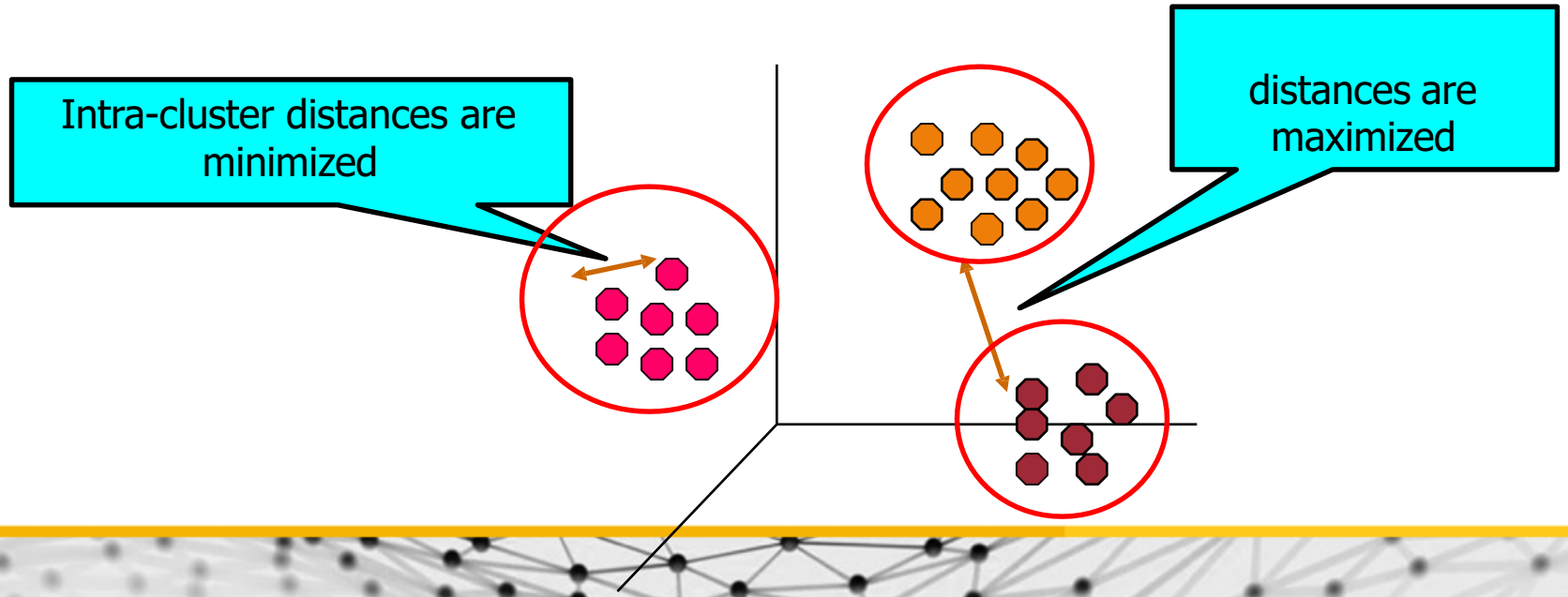
Solution

- 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



Segmentation and Cluster Analysis

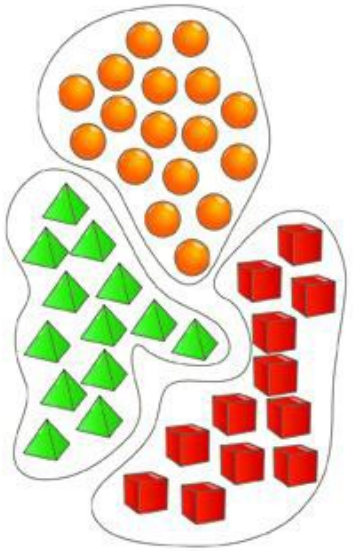
- Cluster is a group of similar objects (cases, points, observations,
- examples, members, customers, patients, locations, etc)
- Finding the groups of cases/observations/ objects in the population such that the objects are
  - Homogeneous within the group (high intra-class similarity)
  - Heterogeneous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



Applications of Cluster Analysis

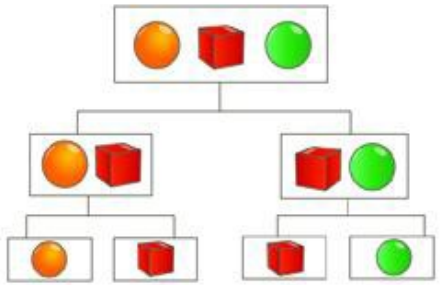
- **Market Segmentation:** Grouping people (with the willingness, purchasing power, and the authority to buy) according to their similarity in several dimensions related to a product under consideration.
- **Sales Segmentation**: Clustering can tell you what types of customers buy what products
- **Credit 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



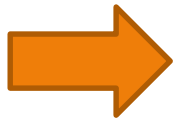
- ***Partitional*** **clustering or non-hierarchical :** A division of objects into non-overlapping subsets (clusters) such that each object is in exactly one cluster
- The non-hierarchical methods divide a dataset of N objects into M clusters.
- **K-means clustering**, a non-hierarchical technique,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



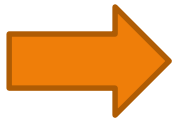
Cluster Analysis-Example

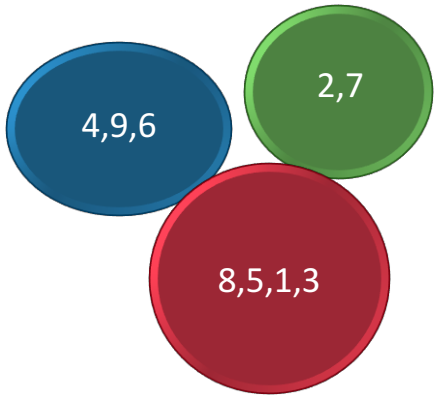
	Maths	Science	Gk	Apt
Student-1	94	82	87	89
Student-2	46	67	33	72
Student-3	98	97	93	100
Student-4	14	5	7	24
Student-5	86	97	95	95
Student-6	34	32	75	66
Student-7	69	44	59	55
Student-8	85	90	96	89
Student-9	24	26	15	22



	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

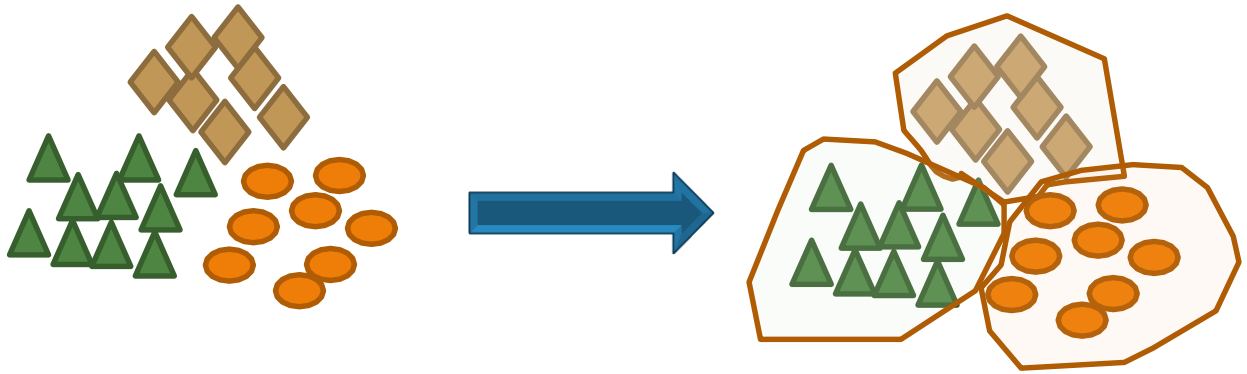
	Maths	Science	Gk	Apt
Student-4	✗ 14	✗ 5	✗ 7	✗ 24
Student-9	✗ 24	✗ 26	✗ 15	✗ 22
Student-6	✗ 34	✗ 32	✓ 75	! 66
Student-2	! 46	! 67	✗ 33	✓ 72
Student-7	✓ 69	! 44	! 59	! 55
Student-8	✓ 85	✓ 90	✓ 96	✓ 89
Student-5	✓ 86	✓ 97	✓ 95	✓ 95
Student-1	✓ 94	✓ 82	✓ 87	✓ 89
Student-3	✓ 98	✓ 97	✓ 93	✓ 100





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

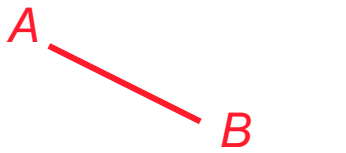
Which two customers are similar in this case?

Quantify dissimilarity -Distancemeasures

- 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

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

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**

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

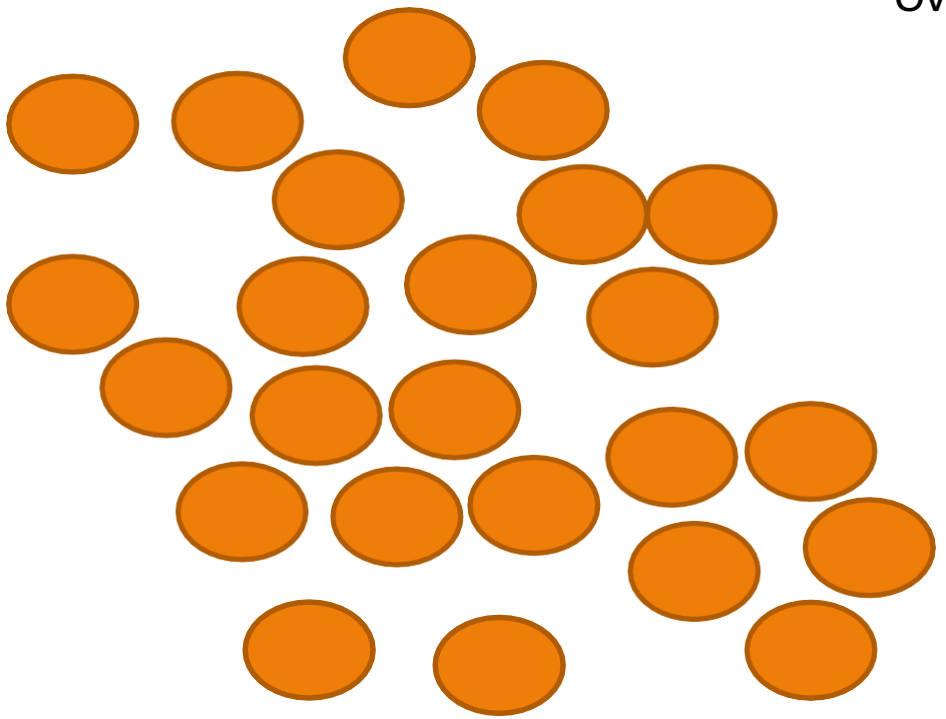
Assignment step: Assign each data point to its closest cluster center

Re-estimation step: Re-compute cluster centers

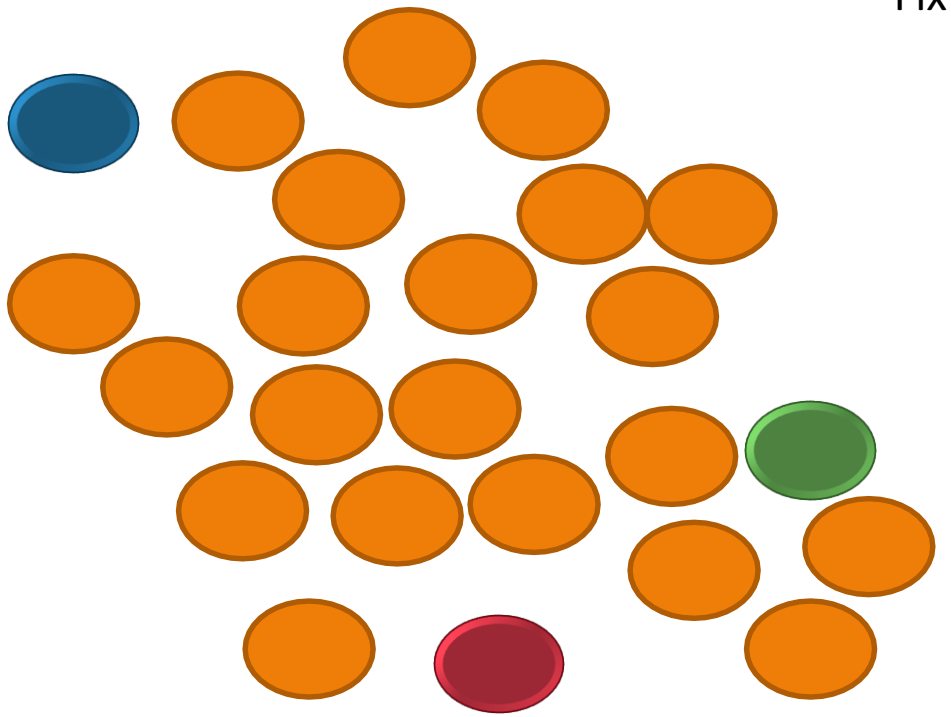
While (there are still changes in the cluster centers)

K-Means clustering

Overall population

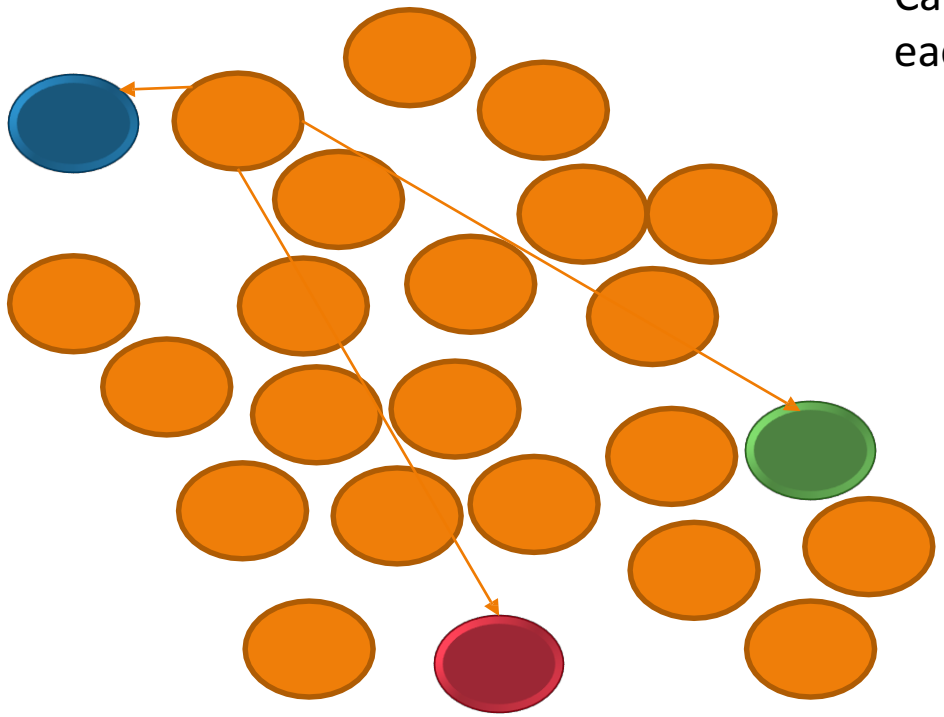


K-Means clustering



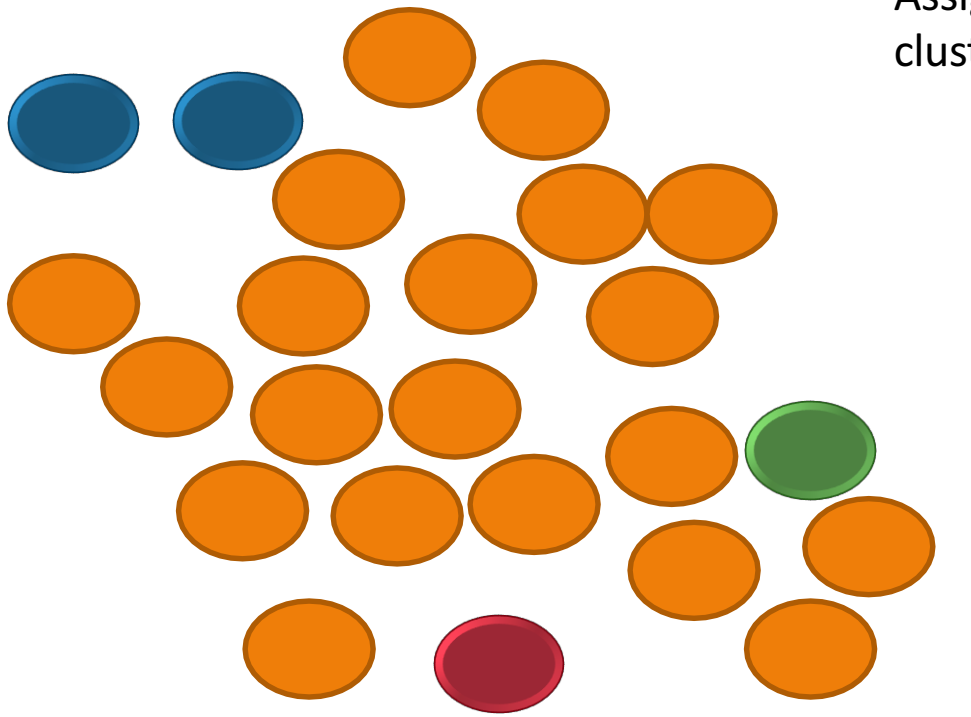
Fix the number of clusters

K-Means clustering



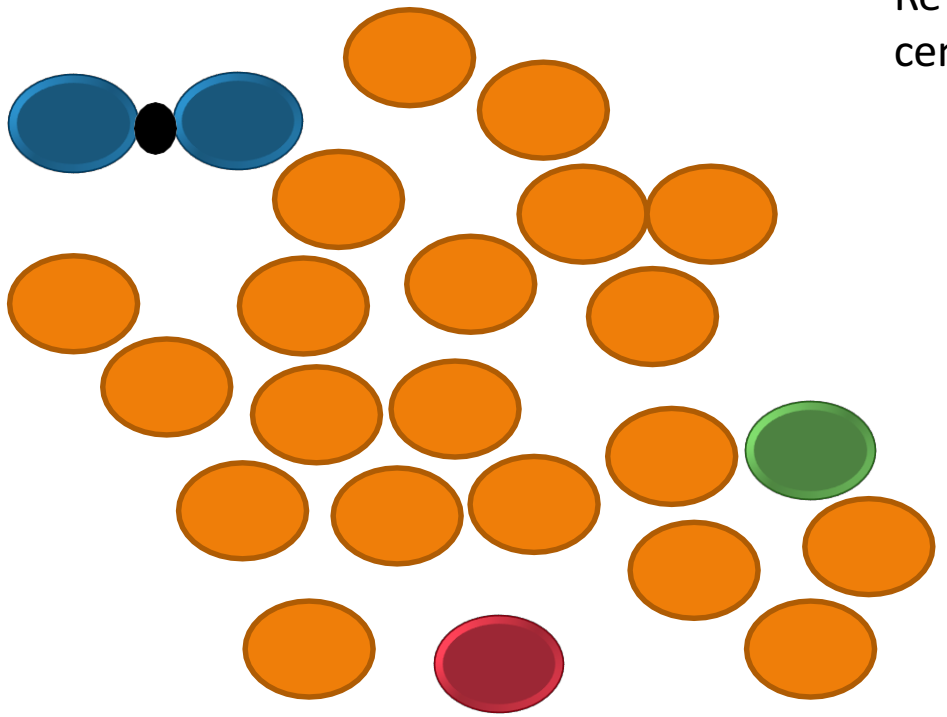
Calculate the distance of each case from all clusters

K-Means clustering



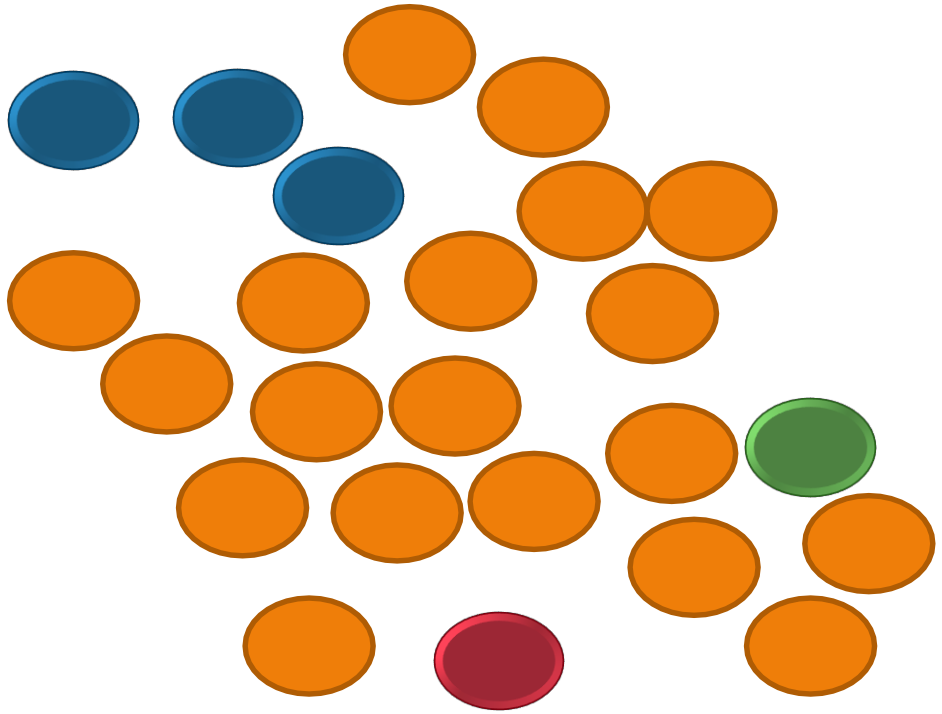
Assign each case to nearest cluster

K-Means clustering

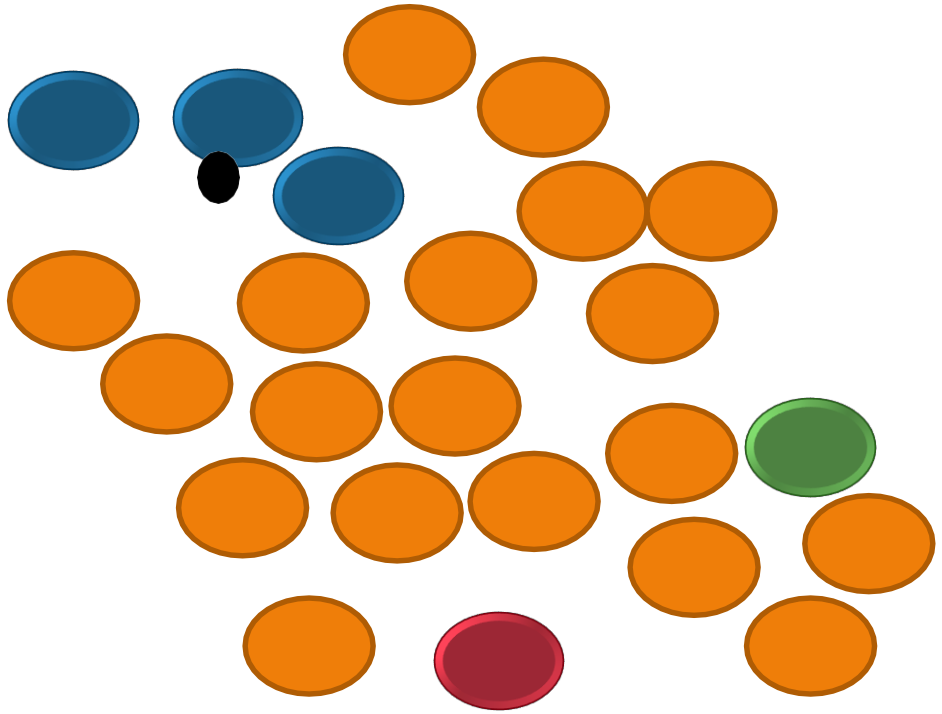


Re calculate the cluster centers

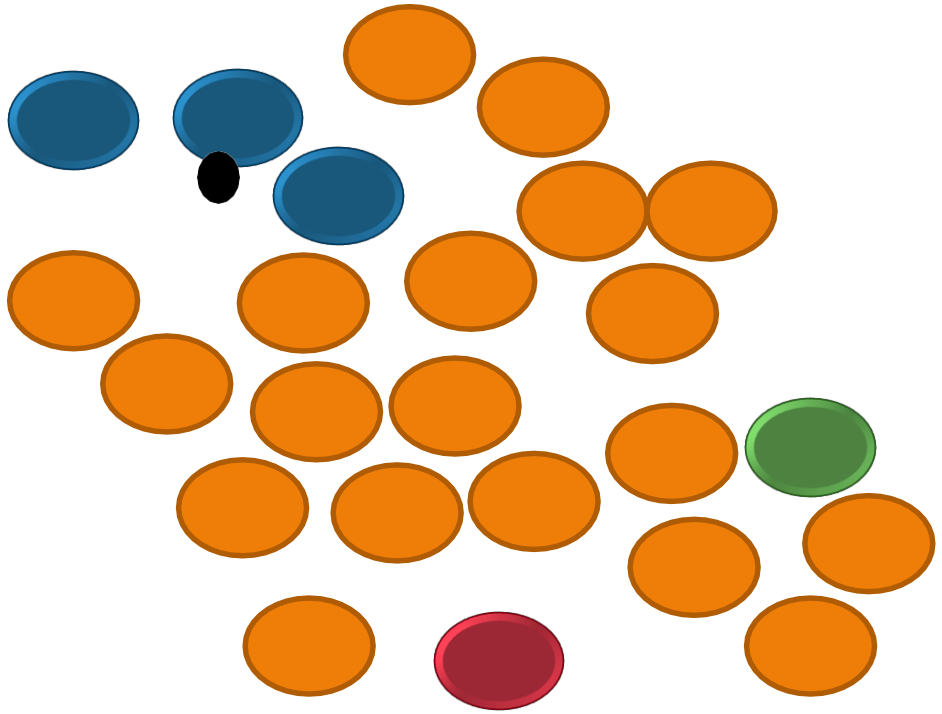
K-Means clustering



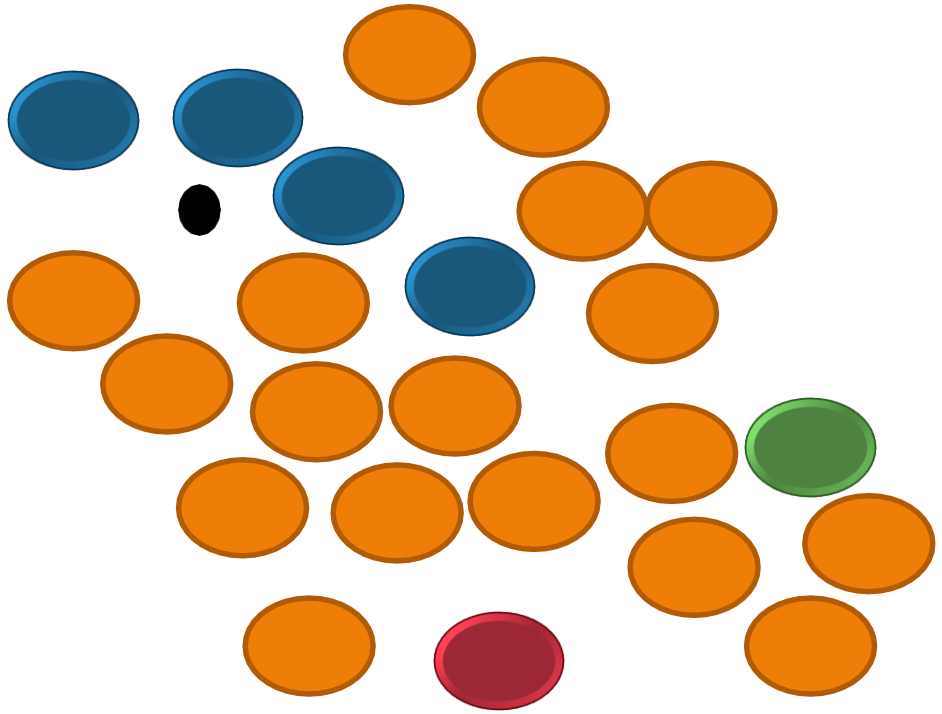
K-Means clustering



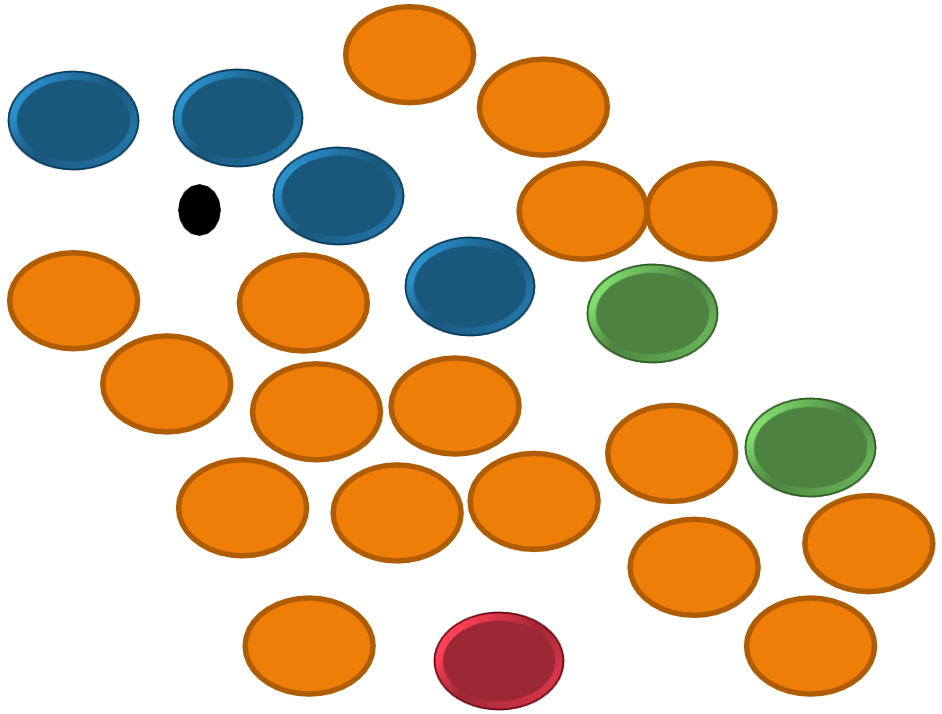
K-Means clustering



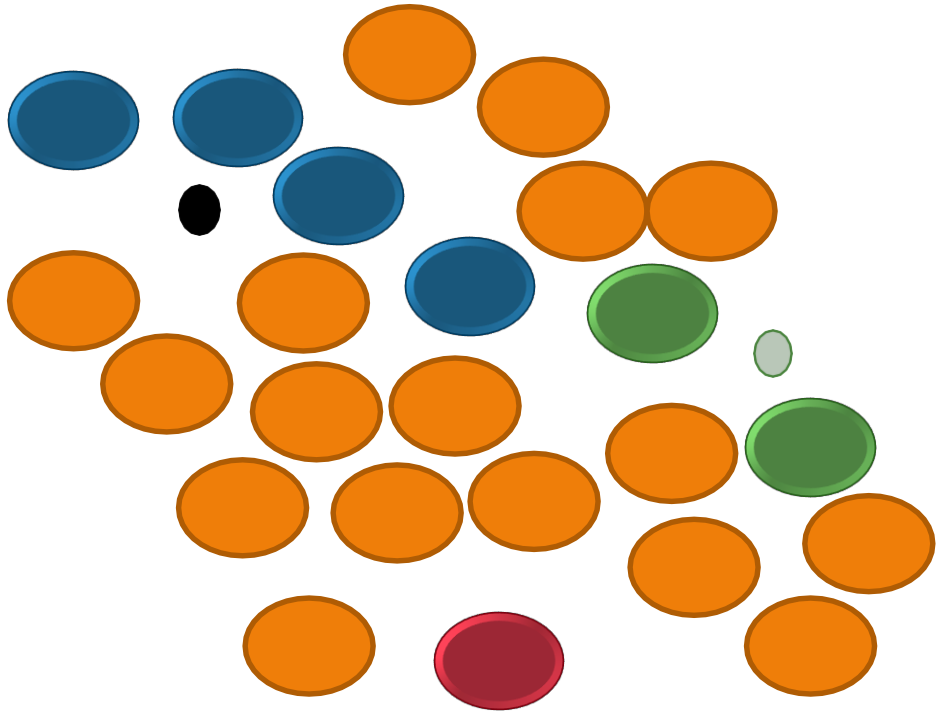
K-Means clustering



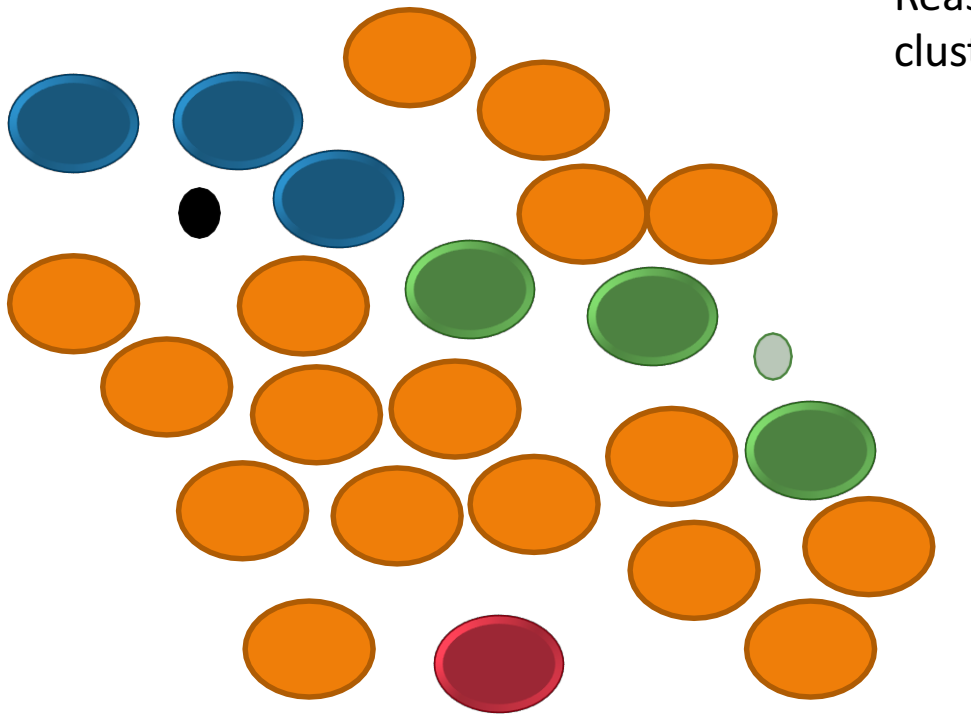
K-Means clustering



K-Means clustering

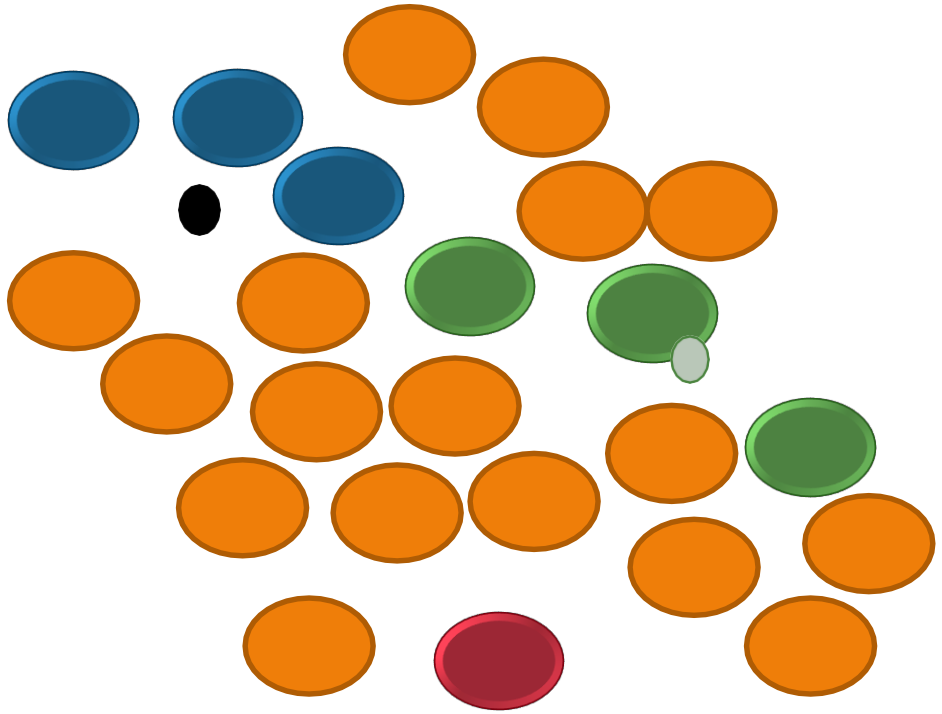


K-Means clustering

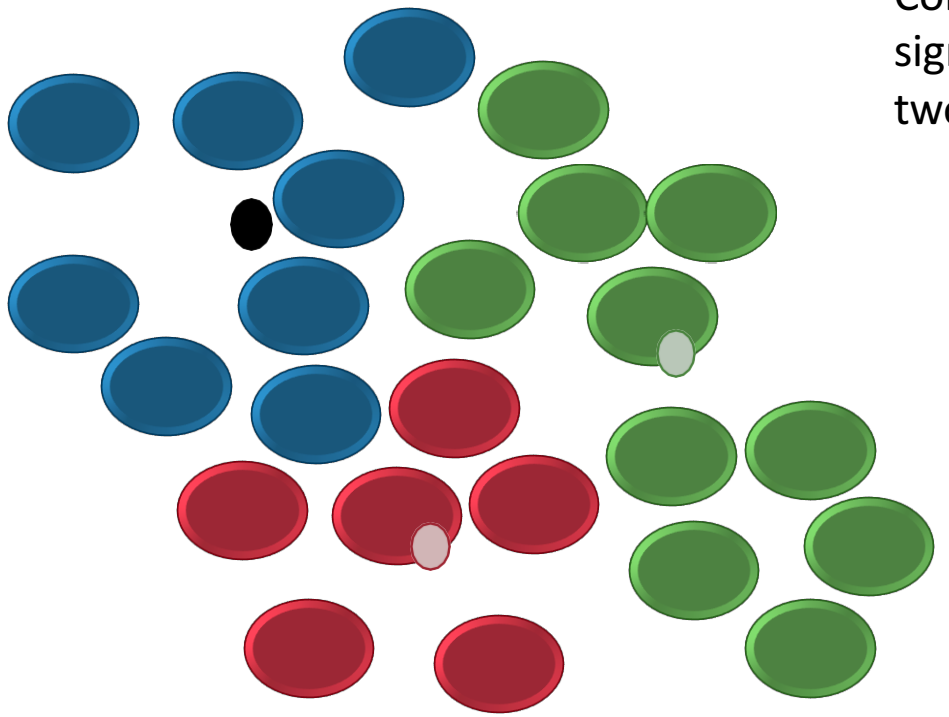


Reassign after changing the cluster centers

K-Means clustering



K-Means clustering

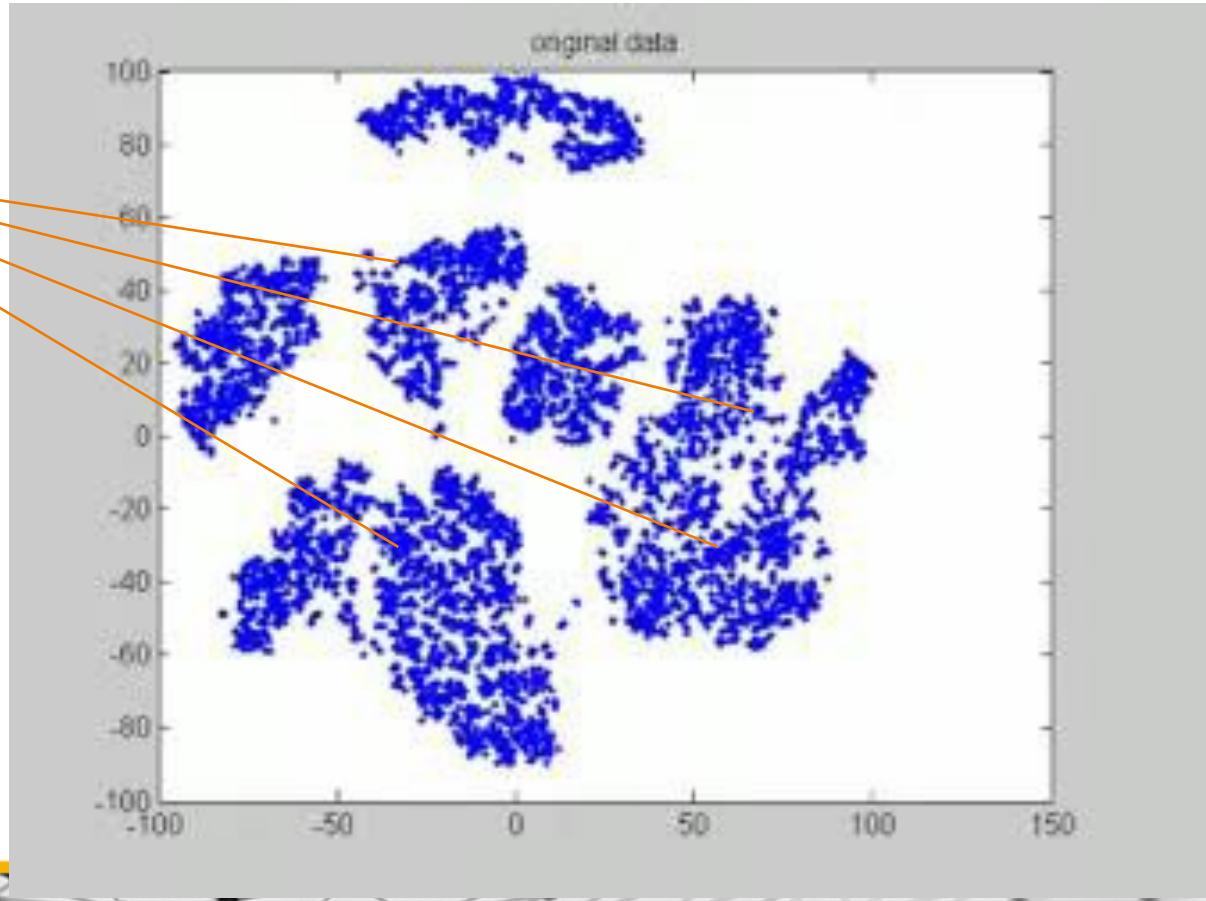


Continue till there is no significant change between two iterations

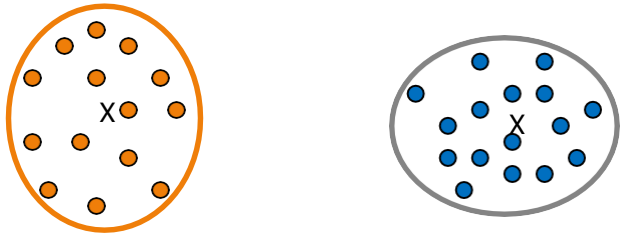
K Means clustering in action

- Dividing the data into 10 clusters using K-Means

Distance metric will decide cluster for these points



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

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