Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

DATA SCIENCE DARMAJAYA
"YOUR BEST FUTURE IN DATA"

MEETING: [14]

PARTITIONING-BASED CLUSTERING

# ALGORITMA K-MEANS

BY: HENDRA KURNIAWAN

[ 12-09-2023 ] | [KODE MK: BDG23407 | MATAKULIAH: DATA SCIENCE AND BUSINESS ANALYTICS | SKS: 2/2 | VERSI: 01

# K-MEANS ALGORITHM

1. K-Means Clustering
2. **K-Means Clustering: Similarity/Dissimilarity**
3. K-Means Algorithm
4. Strengths of K-Means
5. Weaknesses of K-Means
6. Case Study

# K-Means Clustering

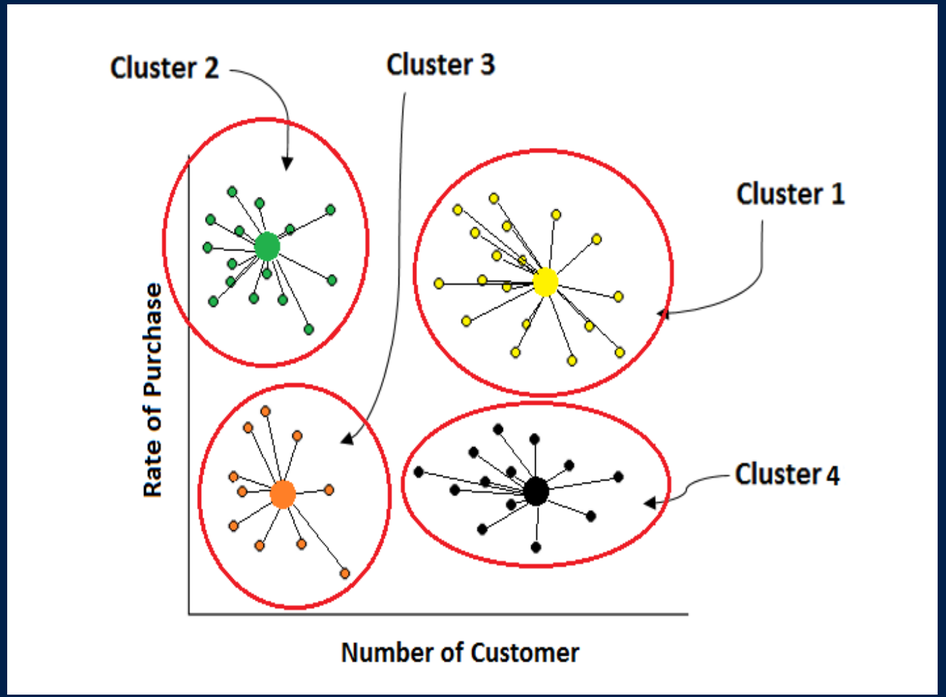
- K-means is a partitional clustering algorithm
- Let the set of data points (or instances) *D* be

  $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$,

  where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ is a vector in a real-valued space $X \subseteq R^r$, and $r$ is the number of attributes (dimensions) in the data.
- The *k*-means algorithm partitions the given data into *k* clusters.
  - Each cluster has a cluster **center**, called **centroid**.
  - *k* is specified by the user

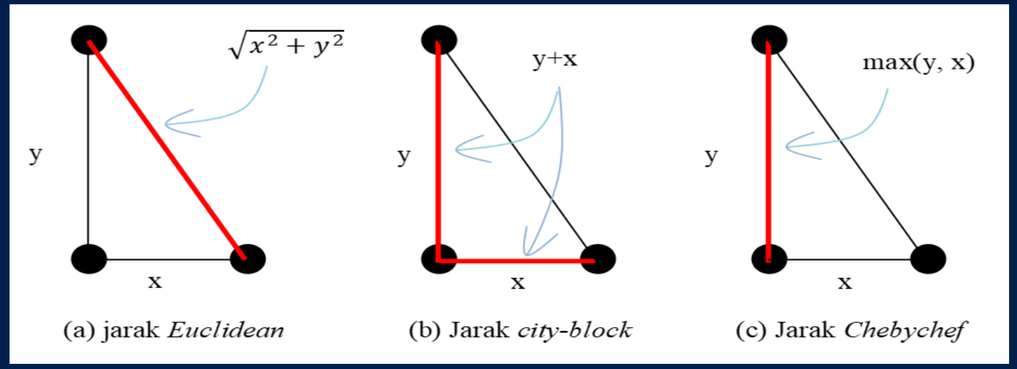
# K-Means Clustering: Similarity/Dissimilarity

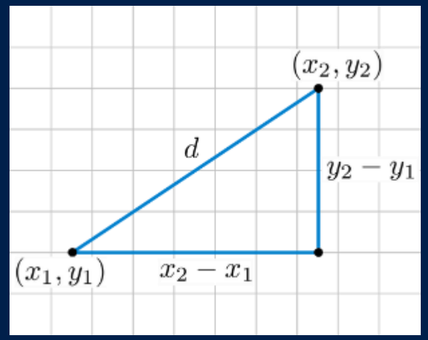
- Intra-cluster:
  - Memaksimalkan similarity (kesamaan) di dalam klaster
  - Meminimalkan dissimilarity di dalam klaster
- Inter-cluster:
  - Meminimalkan similarity antar-klaster
  - Memaksimalkan dissimilarity antar-klaster



# K-Means Clustering: Similarity/Dissimilarity

- Jarak Euclidean
- Jarak City-Block
- Jarak Kotak Catur (Chebychef)
- Jarak Minkowski
- Jarak Canberra
- Jarak Bray-Curtis (Sorensen)
- Divergensi Kullback Leibler
- Divergensi Jensen Shannon
- dll




$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

# K-Means Algorithm

- Given *k*, the *k-means* algorithm works as follows:
  1)Randomly choose *k* data points (seeds) to be the initial centroids, cluster centers
  2)Assign each data point to the closest centroid
  3)Re-compute the centroids using the current cluster memberships.
  4)If a convergence criterion is not met, go to 2).

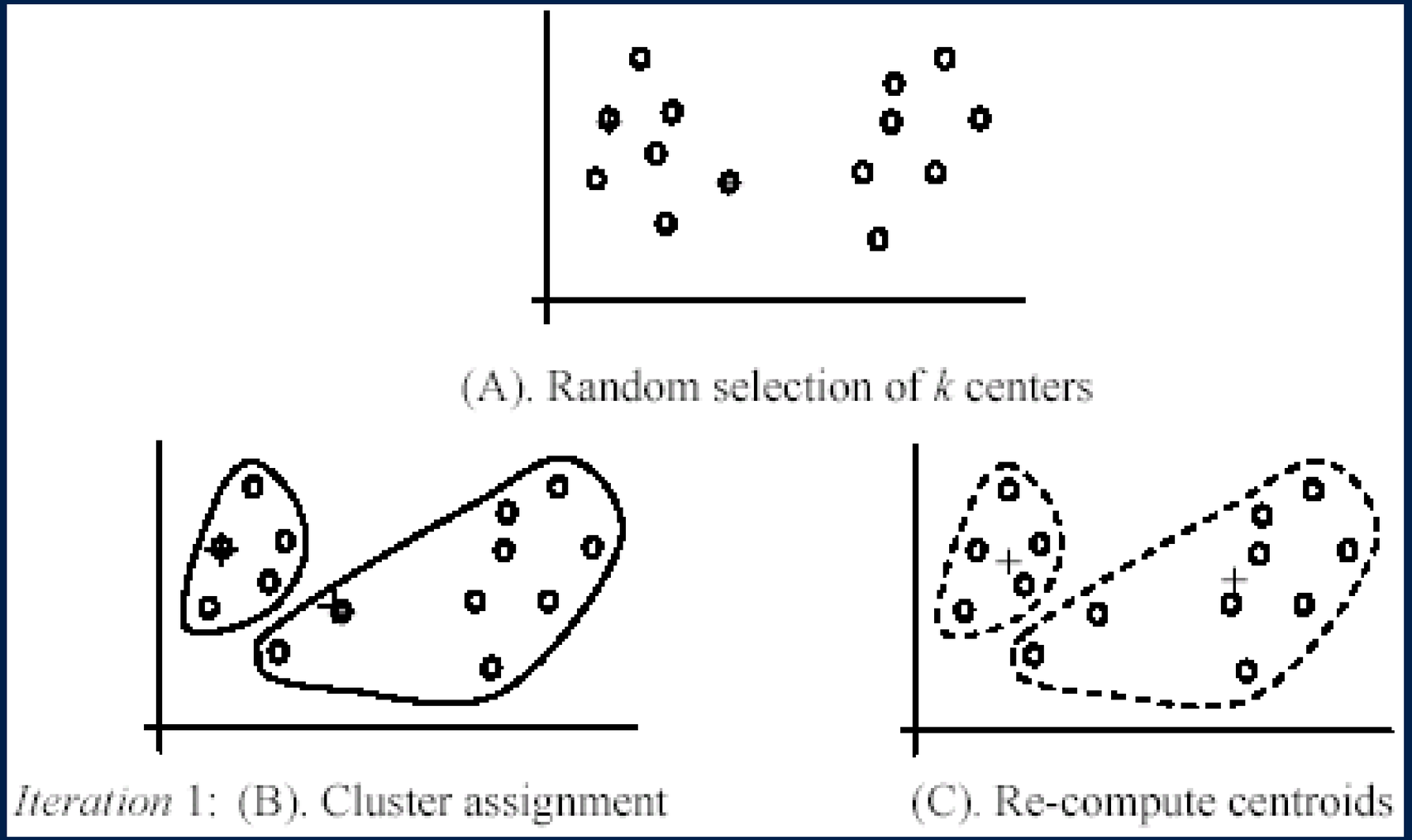
# K-Means Algorithm

**Algorithm** *k*-means(*k*, *D*)

1 Choose *k* data points as the initial centroids (cluster centers)
2 **repeat**
3 **for** each data point $\mathbf{x} \in D$ do
4 compute the distance from $\mathbf{x}$ to each centroid;
5 assign $\mathbf{x}$ to the closest centroid // a centroid represents a cluster
6 **endfor**
7 re-compute the centroids using the current cluster memberships
8 **until** the stopping criterion is met

# K-Means Algorithm

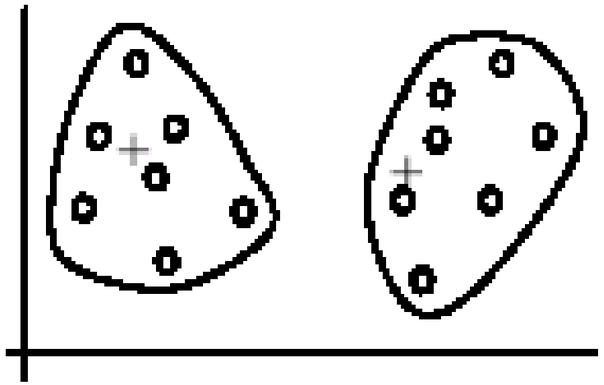
1. Tentukan jumlah klaster (nilai k)
2. Inisialisasi nilai centroid awal setiap klaster secara acak
3. Hitung jarak setiap titik data dengan setiap centroid
4. Masukkan setiap titik data ke dalam klaster berdasarkan jarak terdekat dengan pusat klaster
5. Untuk setiap klaster, tentukan nilai centroid baru berdasarkan rerata (means) dari setiap data di dalam klaster
6. Ulangi langkah 3-5 sedemikian hingga tidak ada perubahan anggota klaster.



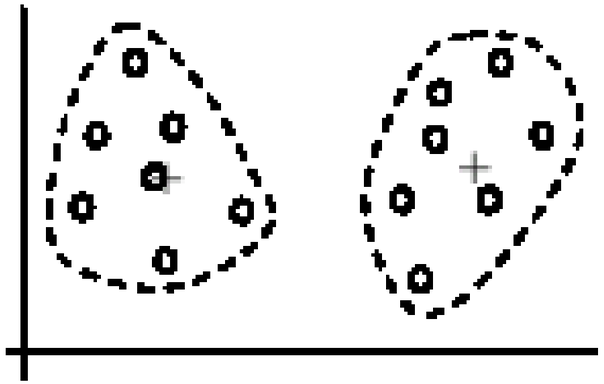
(A). Random selection of $k$ centers

*Iteration* 1: (B). Cluster assignment

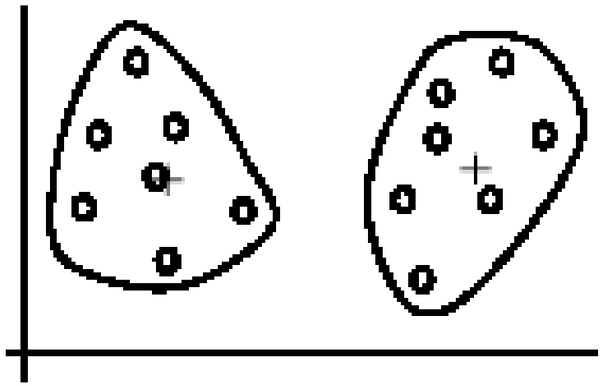
(C). Re-compute centroids



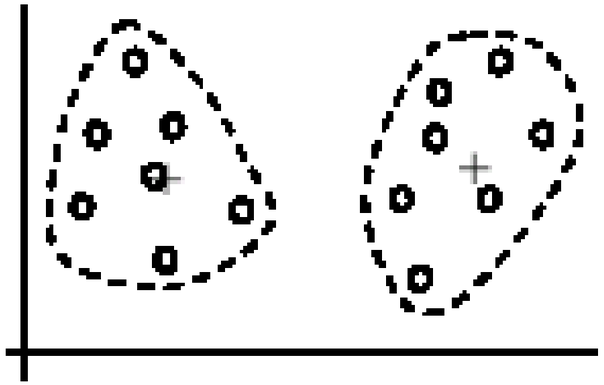
*Iteration* 2: (D). Cluster assignment



(E). Re-compute centroids



*Iteration* 3: (F). Cluster assignment



(G). Re-compute centroids

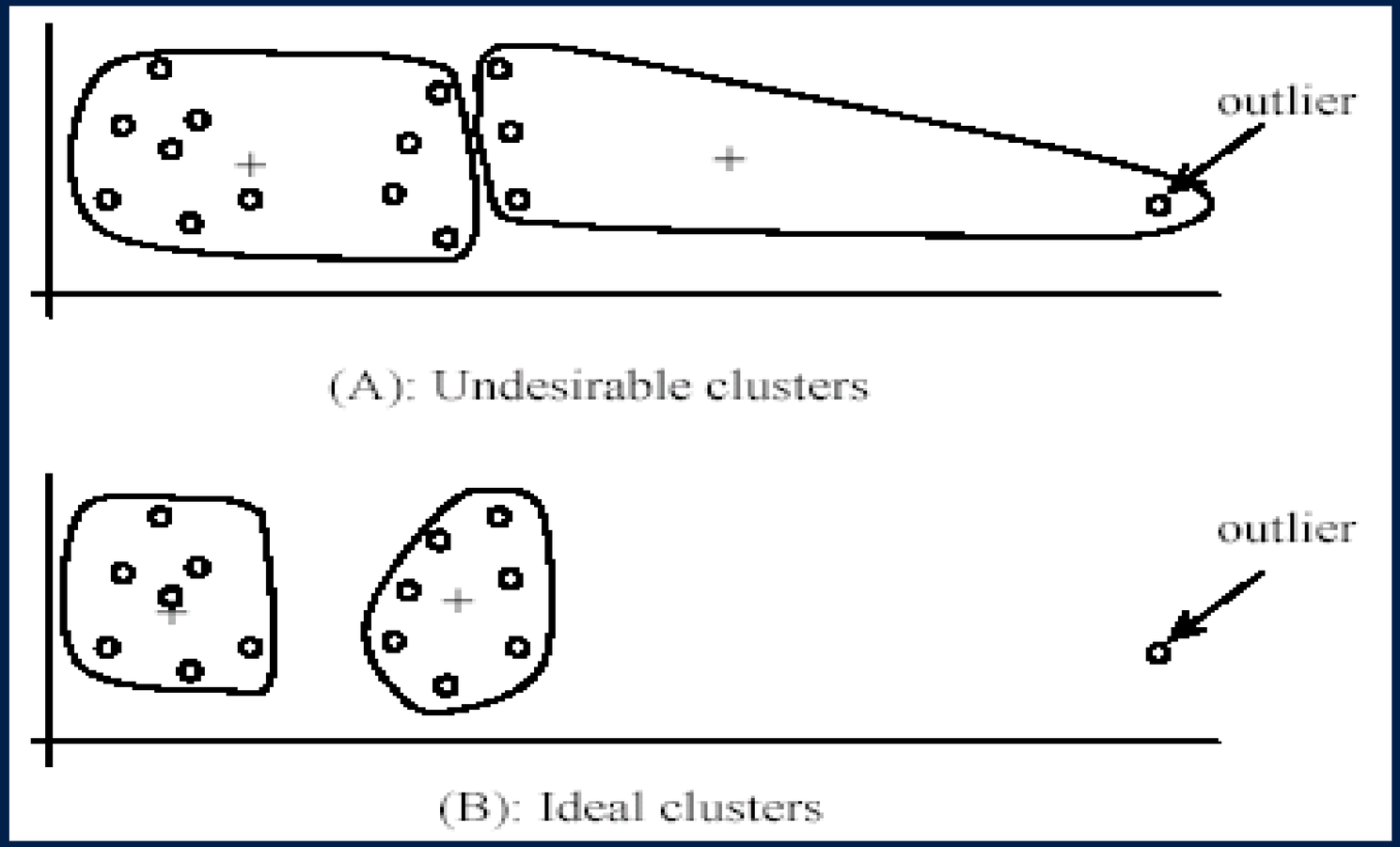
# Strengths of K-Means

- Strengths:
  - Simple: easy to understand and to implement
  - Efficient: Time complexity: $O(tkn)$,
    where $n$ is the number of data points,
    $k$ is the number of clusters, and
    $t$ is the number of iterations.
  - Since both $k$ and $t$ are small. $k$-means is considered a linear algorithm.
- K-means is the most popular clustering algorithm.
- Note that: it terminates at a local optimum if SSE is used. The global optimum is hard to find due to complexity.

# Weaknesses of k-means

- The algorithm is only applicable if the mean is defined.
  - For categorical data, *k*-mode - the centroid is represented by most frequent values.
- The user needs to specify *k*.
- The algorithm is sensitive to **outliers**
  - Outliers are data points that are very far away from other data points.
  - Outliers could be errors in the data recording or some special data points with very different values.

# Weaknesses of k-means: Problems with outliers

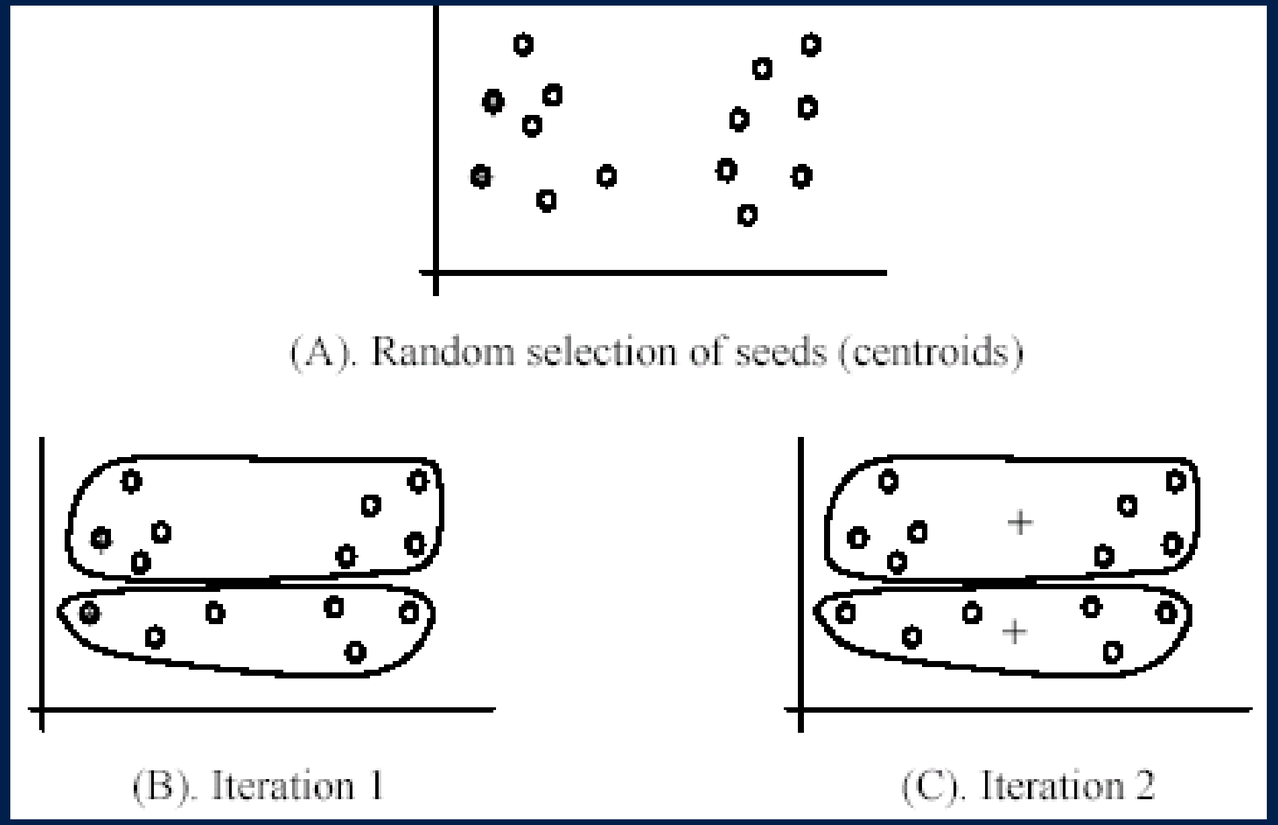


# Weaknesses of k-means: To deal with outliers

- One method is to remove some data points in the clustering process that are much further away from the centroids than other data points.
  - To be safe, we may want to monitor these possible outliers over a few iterations and then decide to remove them.
- Another method is to perform random sampling. Since in sampling we only choose a small subset of the data points, the chance of selecting an outlier is very small.
  - Assign the rest of the data points to the clusters by distance or similarity comparison, or classification

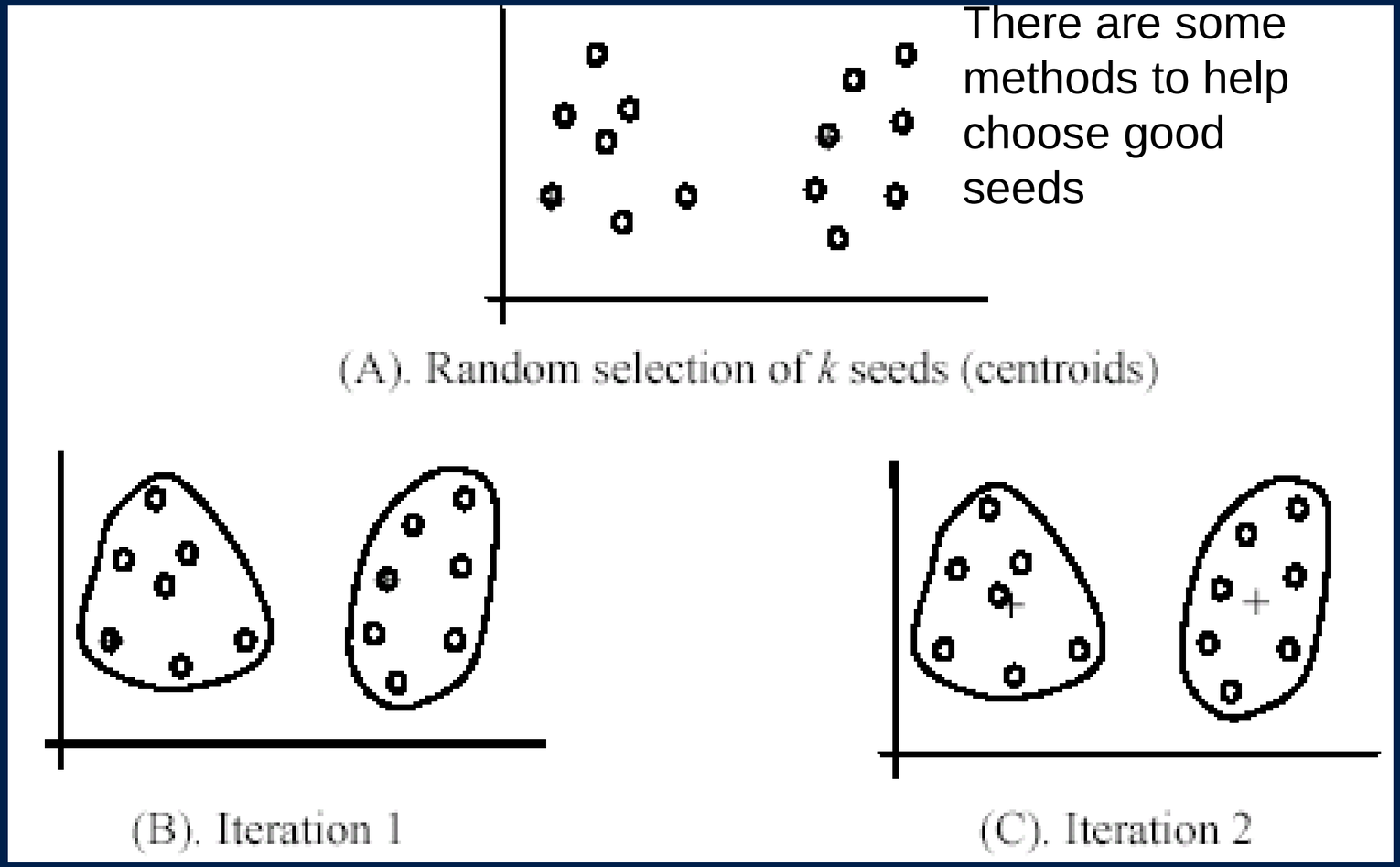
# Weaknesses of k-means (cont …)

- The algorithm is sensitive to initial seeds.



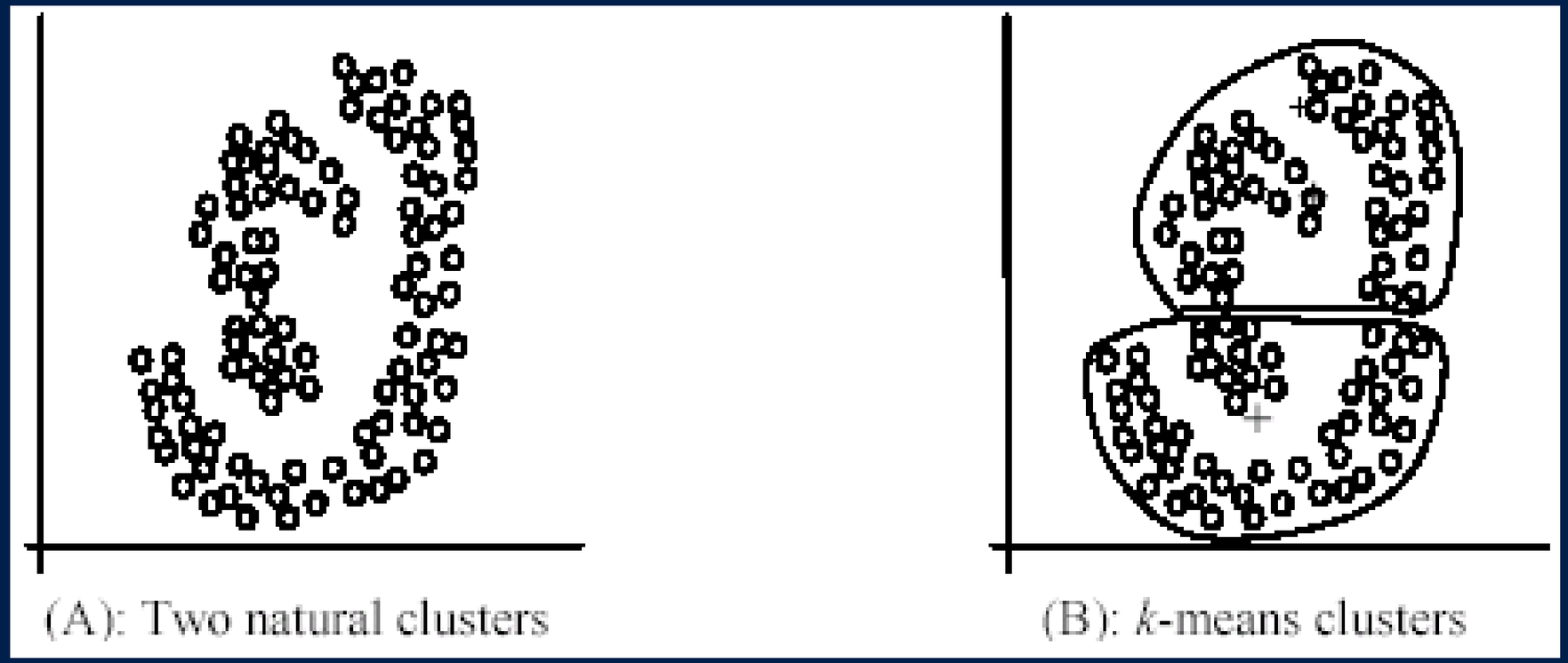
# Weaknesses of k-means (cont …)

- If we use different seeds: good results



# Weaknesses of k-means (cont …)

- The *k*-means algorithm is not suitable for discovering clusters that are not hyper-ellipsoids (or hyper-spheres).



(A): Two natural clusters

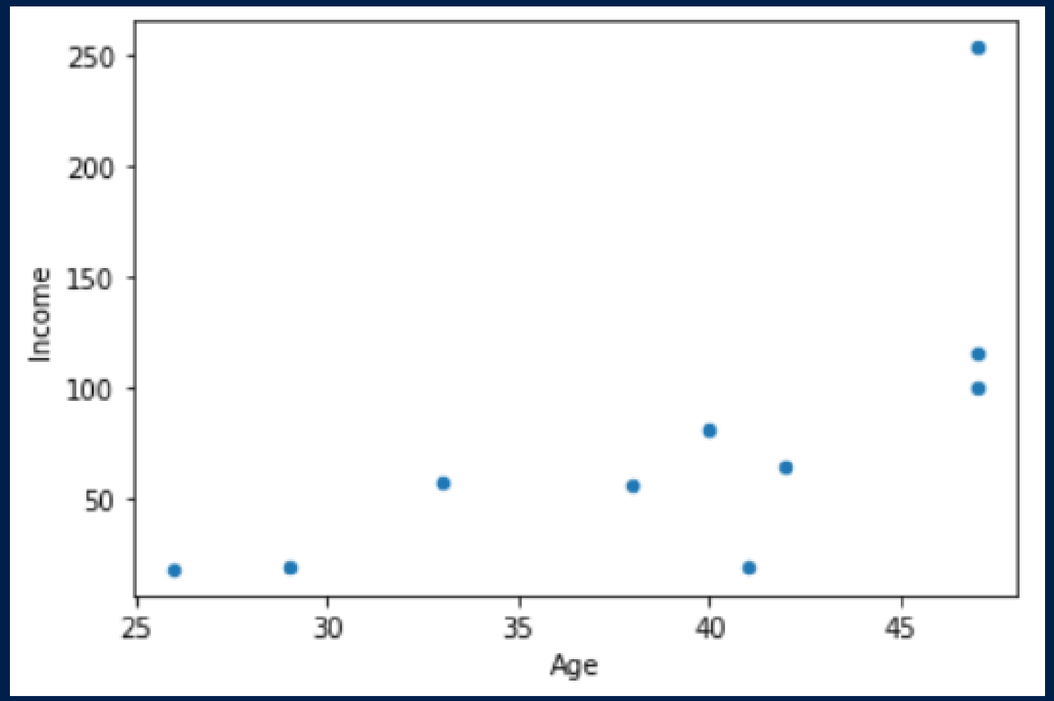
(B): *k*-means clusters

# Example: Customer Clustering

## Customer

| CustID | Age | Income |
|---|---|---|
| 1 | 41 | 19 |
| 2 | 47 | 100 |
| 3 | 33 | 57 |
| 4 | 29 | 19 |
| 5 | 47 | 253 |
| 6 | 40 | 81 |
| 7 | 38 | 56 |
| 8 | 42 | 64 |
| 9 | 26 | 18 |
| 10 | 47 | 115 |

## Clustering data into 2 cluster



# Example: Customer Clustering

| CustID | Age | Income |
|---|---|---|
| 1 | 41 | 19 |
| 2 | 47 | 100 |
| 3 | 33 | 57 |
| 4 | 29 | 19 |
| 5 | 47 | 253 |
| 6 | 40 | 81 |
| 7 | 38 | 56 |
| 8 | 42 | 64 |
| 9 | 26 | 18 |
| 10 | 47 | 115 |

1. **Determine the number of clusters. we use the value k=2**

## 2. initialize the initial centroid value of each cluster randomly

| CustID | Age | Income |
|---|---|---|
| 1 | 41 | 19 |
| 2 | 47 | 100 |
| 3 | 33 | 57 |
| 4 | 29 | 19 |
| 5 | 47 | 253 |
| 6 | 40 | 81 |
| 7 | 38 | 56 |
| 8 | 42 | 64 |
| 9 | 26 | 18 |
| 10 | 47 | 115 |

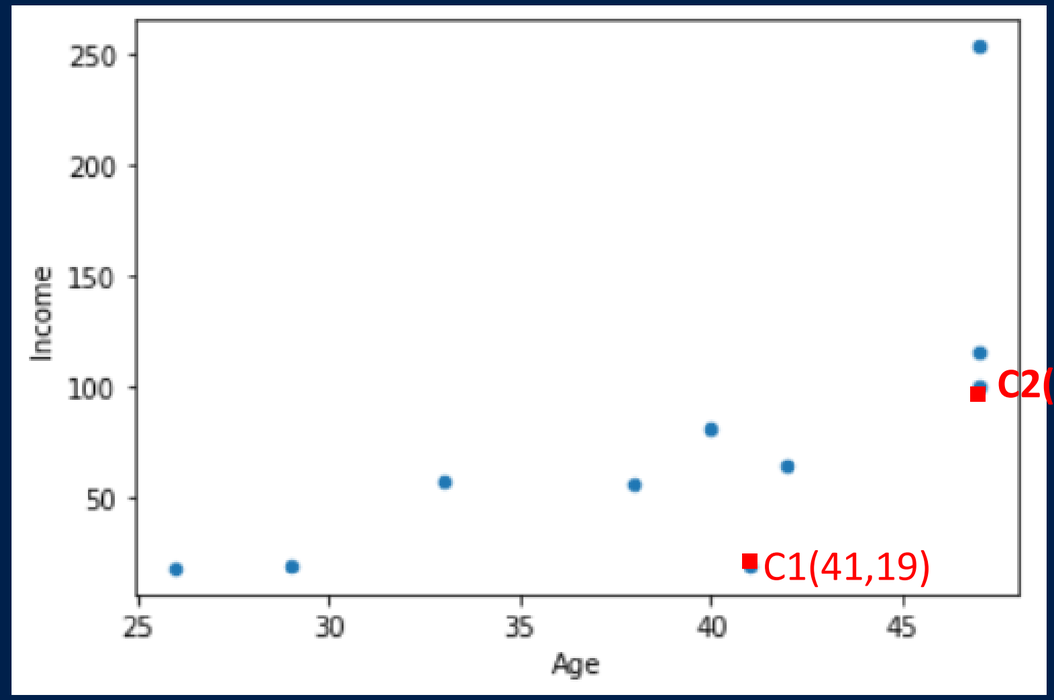
**Cara penentuan centroid awal:**

1. Memilih salah satu data untuk atribut “Age” dan “Income” secara acak
2. Membangkitkan bilangan acak sesuai rentang nilai “Age” dan “Income”

**Dalam contoh ini kita memilih centroid awal dengan cara 1, kita tentukan C1 = (41,19) dan C2 = (47,100)**

## 2. Inisialisasi nilai centroid awal setiap klaster secara acak

| CustID | Age | Income |
|---|---|---|
| 1 | 41 | 19 |
| 2 | 47 | 100 |
| 3 | 33 | 57 |
| 4 | 29 | 19 |
| 5 | 47 | 253 |
| 6 | 40 | 81 |
| 7 | 38 | 56 |
| 8 | 42 | 64 |
| 9 | 26 | 18 |
| 10 | 47 | 115 |



# 3. Hitung jarak setiap titik data dengan setiap centroid. Contoh: Euclidean Distance

| CustID | Age | Income | Jarak ke C1(41,19) | Jarak ke C2(47,100) |
|---|---|---|---|---|
| 1 | 41 | 19 | $\sqrt{(41-41)^2+(19-19)^2}=0$ | $\sqrt{(41-47)^2+(19-100)^2}=81{,}22$ |
| 2 | 47 | 100 | $\sqrt{(47-41)^2+(100-19)^2}=81{,}22$ | $\sqrt{(47-47)^2+(100-100)^2}=0$ |
| 3 | 33 | 57 | $\sqrt{(33-41)^2+(57-19)^2}=38{,}83$ | $\sqrt{(33-47)^2+(57-100)^2}=45{,}22$ |
| 4 | 29 | 19 | $\sqrt{(29-41)^2+(19-19)^2}=12{,}0$ | $\sqrt{(29-47)^2+(19-100)^2}=82{,}98$ |
| 5 | 47 | 253 | $\sqrt{(47-41)^2+(253-19)^2}=234{,}08$ | $\sqrt{(47-47)^2+(253-100)^2}=153{,}0$ |
| 6 | 40 | 81 | $\sqrt{(40-41)^2+(81-19)^2}=62{,}01$ | $\sqrt{(40-47)^2+(81-100)^2}=20{,}25$ |
| 7 | 38 | 56 | $\sqrt{(38-41)^2+(56-19)^2}=37{,}12$ | $\sqrt{(38-47)^2+(56-100)^2}=44{,}91$ |
| 8 | 42 | 64 | $\sqrt{(42-41)^2+(64-19)^2}=45{,}01$ | $\sqrt{(42-47)^2+(64-100)^2}=36{,}35$ |
| 9 | 26 | 18 | $\sqrt{(26-41)^2+(18-19)^2}=15{,}03$ | $\sqrt{(26-47)^2+(18-100)^2}=84{,}65$ |
| 10 | 47 | 115 | $\sqrt{(47-41)^2+(115-19)^2}=96{,}19$ | $\sqrt{(47-47)^2+(115-100)^2}=15{,}0$ |

# 4. Masukkan setiap titik data ke dalam klaster berdasarkan jarak terdekat dengan centroid

| CustID | Age | Income | Jarak ke C1(41,19) | Jarak ke C2(47,100) | Klaster |
|---|---|---|---|---|---|
| 1 | 41 | 19 | $\sqrt{(41-41)^2+(19-19)^2}=0$ | $\sqrt{(41-47)^2+(19-100)^2}=81{,}22$ | 1 |
| 2 | 47 | 100 | $\sqrt{(47-41)^2+(100-19)^2}=81{,}22$ | $\sqrt{(47-47)^2+(100-100)^2}=0$ | 2 |
| 3 | 33 | 57 | $\sqrt{(33-41)^2+(57-19)^2}=38{,}83$ | $\sqrt{(33-47)^2+(57-100)^2}=45{,}22$ | 1 |
| 4 | 29 | 19 | $\sqrt{(29-41)^2+(19-19)^2}=12{,}0$ | $\sqrt{(29-47)^2+(19-100)^2}=82{,}98$ | 1 |
| 5 | 47 | 253 | $\sqrt{(47-41)^2+(253-19)^2}=234{,}08$ | $\sqrt{(47-47)^2+(253-100)^2}=153{,}0$ | 2 |
| 6 | 40 | 81 | $\sqrt{(40-41)^2+(81-19)^2}=62{,}01$ | $\sqrt{(40-47)^2+(81-100)^2}=20{,}25$ | 2 |
| 7 | 38 | 56 | $\sqrt{(38-41)^2+(56-19)^2}=37{,}12$ | $\sqrt{(38-47)^2+(56-100)^2}=44{,}91$ | 1 |
| 8 | 42 | 64 | $\sqrt{(42-41)^2+(64-19)^2}=45{,}01$ | $\sqrt{(42-47)^2+(64-100)^2}=36{,}35$ | 2 |
| 9 | 26 | 18 | $\sqrt{(26-41)^2+(18-19)^2}=15{,}03$ | $\sqrt{(26-47)^2+(18-100)^2}=84{,}65$ | 1 |
| 10 | 47 | 115 | $\sqrt{(47-41)^2+(115-19)^2}=96{,}19$ | $\sqrt{(47-47)^2+(115-100)^2}=15{,}0$ | 2 |

## 4. Masukkan setiap titik data ke dalam klaster berdasarkan jarak terdekat dengan centroid

**Klaster 1**
- Cust 1
- Cust 3
- Cust 4
- Cust 7
- Cust 9

**Klaster 2**
- Cust 2
- Cust 5
- Cust 6
- Cust 8
- Cust 10

## 5. Untuk setiap klaster, hitung nilai centroid baru berdasarkan rerata (means) dari setiap data di dalam klaster

| CustID | Age | Income | Jarak ke C1(41,19) | Jarak ke C2(47,100) | Klaster |
|---|---|---|---|---|---|
| 1 | 41 | 19 | $\sqrt{(41-41)^2+(19-19)^2}=0$ | $\sqrt{(41-47)^2+(19-100)^2}=81{,}22$ | 1 |
| 2 | 47 | 100 | $\sqrt{(47-41)^2+(100-19)^2}=81{,}22$ | $\sqrt{(47-47)^2+(100-100)^2}=0$ | 2 |
| 3 | 33 | 57 | $\sqrt{(33-41)^2+(57-19)^2}=38{,}83$ | $\sqrt{(33-47)^2+(57-100)^2}=45{,}22$ | 1 |
| 4 | 29 | 19 | $\sqrt{(29-41)^2+(19-19)^2}=12{,}0$ | $\sqrt{(29-47)^2+(19-100)^2}=82{,}98$ | 1 |
| 5 | 47 | 253 | $\sqrt{(47-41)^2+(253-19)^2}=234{,}08$ | $\sqrt{(47-47)^2+(253-100)^2}=153{,}0$ | 2 |
| 6 | 40 | 81 | $\sqrt{(40-41)^2+(81-19)^2}=62{,}01$ | $\sqrt{(40-47)^2+(81-100)^2}=20{,}25$ | 2 |
| 7 | 38 | 56 | $\sqrt{(38-41)^2+(56-19)^2}=37{,}12$ | $\sqrt{(38-47)^2+(56-100)^2}=44{,}91$ | 1 |
| 8 | 42 | 64 | $\sqrt{(42-41)^2+(64-19)^2}=45{,}01$ | $\sqrt{(42-47)^2+(64-100)^2}=36{,}35$ | 2 |
| 9 | 26 | 18 | $\sqrt{(26-41)^2+(18-19)^2}=15{,}03$ | $\sqrt{(26-47)^2+(18-100)^2}=84{,}65$ | 1 |
| 10 | 47 | 115 | $\sqrt{(47-41)^2+(115-19)^2}=96{,}19$ | $\sqrt{(47-47)^2+(115-100)^2}=15{,}0$ | 2 |

**Centroid Baru** **C1** = (mean(41;33;29;38;26), mean(19;57;19;56;18)) = **(33,4; 33,8)**

# Contoh Kasus: Klasterisasi Pelanggan

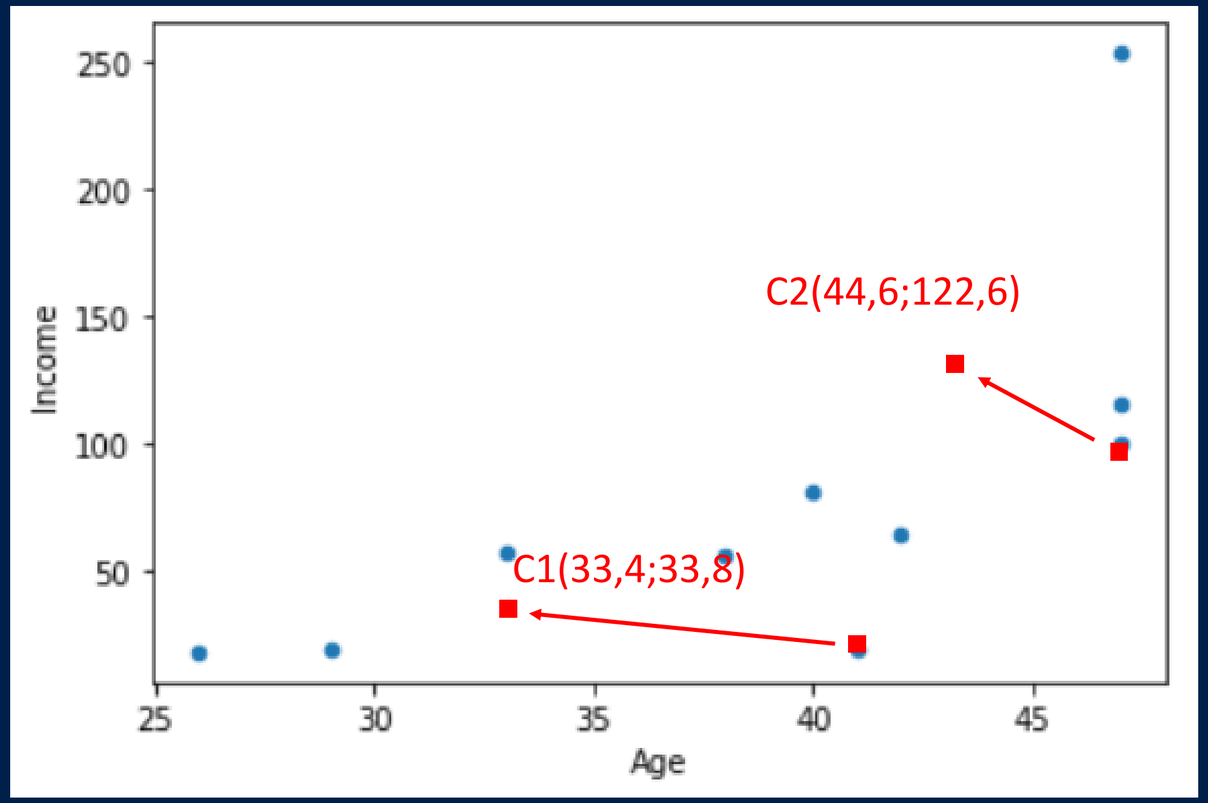
**5. Untuk setiap klaster, hitung nilai centroid baru berdasarkan rerata (means) dari setiap data di dalam klaster**

| CustID | Age | Income | Jarak ke C1(41,19) | Jarak ke C2(47,100) | Klaster |
|---|---|---|---|---|---|
| 1 | 41 | 19 | $\sqrt{(41-41)^2+(19-19)^2}=0$ | $\sqrt{(41-47)^2+(19-100)^2}=81{,}22$ | 1 |
| 2 | 47 | 100 | $\sqrt{(47-41)^2+(100-19)^2}=81{,}22$ | $\sqrt{(47-47)^2+(100-100)^2}=0$ | 2 |
| 3 | 33 | 57 | $\sqrt{(33-41)^2+(57-19)^2}=38{,}83$ | $\sqrt{(33-47)^2+(57-100)^2}=45{,}22$ | 1 |
| 4 | 29 | 19 | $\sqrt{(29-41)^2+(19-19)^2}=12{,}0$ | $\sqrt{(29-47)^2+(19-100)^2}=82{,}98$ | 1 |
| 5 | 47 | 253 | $\sqrt{(47-41)^2+(253-19)^2}=234{,}08$ | $\sqrt{(47-47)^2+(253-100)^2}=153{,}0$ | 2 |
| 6 | 40 | 81 | $\sqrt{(40-41)^2+(81-19)^2}=62{,}01$ | $\sqrt{(40-47)^2+(81-100)^2}=20{,}25$ | 2 |
| 7 | 38 | 56 | $\sqrt{(38-41)^2+(56-19)^2}=37{,}12$ | $\sqrt{(38-47)^2+(56-100)^2}=44{,}91$ | 1 |
| 8 | 42 | 64 | $\sqrt{(42-41)^2+(64-19)^2}=45{,}01$ | $\sqrt{(42-47)^2+(64-100)^2}=36{,}35$ | 2 |
| 9 | 26 | 18 | $\sqrt{(26-41)^2+(18-19)^2}=15{,}03$ | $\sqrt{(26-47)^2+(18-100)^2}=84{,}65$ | 1 |
| 10 | 47 | 115 | $\sqrt{(47-41)^2+(115-19)^2}=96{,}19$ | $\sqrt{(47-47)^2+(115-100)^2}=15{,}0$ | 2 |

**Centroid Baru**

**C2** = (mean(47;47;40;42;47), mean(100;253;81;64;115)) = **(44,6; 122,6)**

# Pergeseran Centroid setiap klaster. C1 =(33,4; 33,8) dan C2 = (44,6; 122,6)



# 6. Ulangi langkah 3-5 menggunakan centroid baru

| CustID | Age | Income | Jarak ke C1(33,4; 33,8) | Jarak ke C2(44,6; 122,6) | Klaster |
|---|---|---|---|---|---|
| 1 | 41 | 19 | 16,64 | 103,66 | 1 |
| 2 | 47 | 100 | 67,58 | 22,73 | 2 |
| 3 | 33 | 57 | 23,20 | 66,62 | 1 |
| 4 | 29 | 19 | 15,44 | 104,77 | 1 |
| 5 | 47 | 253 | 219,62 | 130,42 | 2 |
| 6 | 40 | 81 | 47,66 | 41,85 | 2 |
| 7 | 38 | 56 | 22,67 | 66,93 | 1 |
| 8 | 42 | 64 | 31,40 | 58,66 | 1 |
| 9 | 26 | 18 | 17,45 | 106,24 | 1 |
| 10 | 47 | 115 | 82,33 | 7,97 | 2 |

# 6. Ulangi langkah 3-5 menggunakan centroid baru

| CustID | Age | Income | Jarak ke C1(33,4; 33,8) | Jarak ke C2(44,6; 122,6) | Klaster |
|---|---|---|---|---|---|
| 1 | 41 | 19 | 16,64 | 103,66 | 1 |
| 2 | 47 | 100 | 67,58 | 22,73 | 2 |
| 3 | 33 | 57 | 23,20 | 66,62 | 1 |
| 4 | 29 | 19 | 15,44 | 104,77 | 1 |
| 5 | 47 | 253 | 219,62 | 130,42 | 2 |
| 6 | 40 | 81 | 47,66 | 41,85 | 2 |
| 7 | 38 | 56 | 22,67 | 66,93 | 1 |
| 8 | 42 | 64 | 31,40 | 58,66 | 1 |
| 9 | 26 | 18 | 17,45 | 106,24 | 1 |
| 10 | 47 | 115 | 82,33 | 7,97 | 2 |

Apakah hasil klasterisasinya sama dengan tahap sebelumnya?

- Jika sama, hentikan proses klasterisasi
- Jika belum sama, ulangi langkah 3-5

| Data | | | ITERASI 1 | |
|---|---|---|---|---|
| **CustID** | **Age** | **Income** | C1(41,19) | C2(47,100) |
| 1 | 41 | 19 | 0,00 | 81,22 |
| 2 | 47 | 100 | 81,22 | 0,00 |
| 3 | 33 | 57 | 38,83 | 45,22 |
| 4 | 29 | 19 | 12,00 | 82,98 |
| 5 | 47 | 253 | 234,08 | 153,00 |
| 6 | 40 | 81 | 62,01 | 20,25 |
| 7 | 38 | 56 | 37,12 | 44,91 |
| 8 | 42 | 64 | 45,01 | 36,35 |
| 9 | 26 | 18 | 15,03 | 84,65 |
| 10 | 47 | 115 | 96,19 | 15,00 |
| | | | | |
| | Centroid Baru | | 33,4 | 44,6 |
| | | | 33,8 | 122,6 |

| ITERASI 2 | |
|---|---|
| C1 | C2 |
| 16,64 | 103,66 |
| 67,58 | 22,73 |
| 23,20 | 66,62 |
| 15,44 | 104,77 |
| 219,62 | 130,42 |
| 47,66 | 41,85 |
| 22,67 | 66,93 |
| 31,40 | 58,66 |
| 17,45 | 106,24 |
| 82,33 | 7,97 |
| | |
| 34,83 | 45,25 |
| 38,83 | 137,25 |

| ITERASI 3 | |
|---|---|
| C1 | C2 |
| 20,77 | 118,33 |
| 62,36 | 37,29 |
| 18,26 | 81,18 |
| 20,67 | 119,36 |
| 214,51 | 115,76 |
| 42,48 | 56,49 |
| 17,46 | 81,57 |
| 26,17 | 73,32 |
| 22,63 | 120,79 |
| 77,13 | 22,32 |
| | |
| 35,57 | 47,00 |
| 44,86 | 156 |

| ITERASI 4 | |
|---|---|
| C1 | C2 |
| 26,42 | 137,13 |
| 56,31 | 56,00 |
| 12,41 | 99,98 |
| 26,68 | 138,18 |
| 208,46 | 97,00 |
| 36,41 | 75,33 |
| 11,40 | 100,40 |
| 20,19 | 92,14 |
| 28,51 | 139,59 |
| 71,07 | 41,00 |
| | |
| SELESAI | |

# Optimasi Nilai k pada K-Means

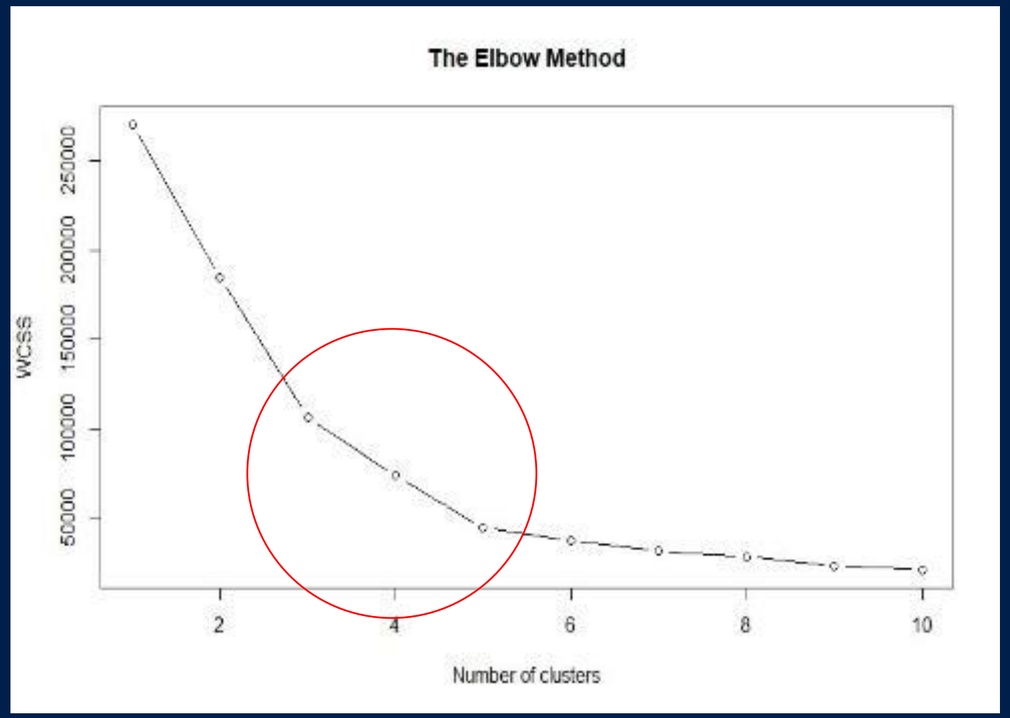
- Salah satu faktor krusial baik tidaknya metode K-Means adalah **jumlah klusternya (nilai K)**. Hasil pengelompokan akan menghasilkan analisa yang berbeda untuk jumlah klaster yang berbeda juga.
- **Semakin kecil nilai K** (misal 2), maka pembagian kluster menjadi cepat, namun mungkin ada informasi tersembunyi yang tidak terungkap.
- **Semakin besar nilai K** (misal K=10), maka terlalu banyak kluster. Mungkin akan terlalu sulit untuk membuat analisa atau memilih dukungan keputusan dari hasil cluster.

# Optimasi Nilai k pada K-Means

- Penentuan nilai k terbaik dapat dilakukan berdasarkan ukuran kualitas hasil klasterisasi.
- Beberapa ukuran kualitas klaster:
  - Sum Square Error (SSE)
  - Davies-Bouldin Index (DBI)
  - Silhouette Coefficient
  - Rand Index
  - Mutual Information
  - Calinski-Harabasz Index (C-H Index)
  - Dunn Index

# Penentuan Nilai k Terbaik dengan Metode Elbow

- Untuk mengetahui jumlah klaster yang paling baik adalah dengan cara melihat perbandingan kualitas klaster untuk setiap pilihan nilai k (missal k=2, 3, 4, 5, …).
- Nilai k yang dipilih adalah nilai k yang memiliki perubahan kualitas signifikan, seperti sebuah **siku (elbow)**.



Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

# THANK YOU!!

DATA SCIENCE DARMAJAYA “YOUR BEST FUTURE IN DATA”