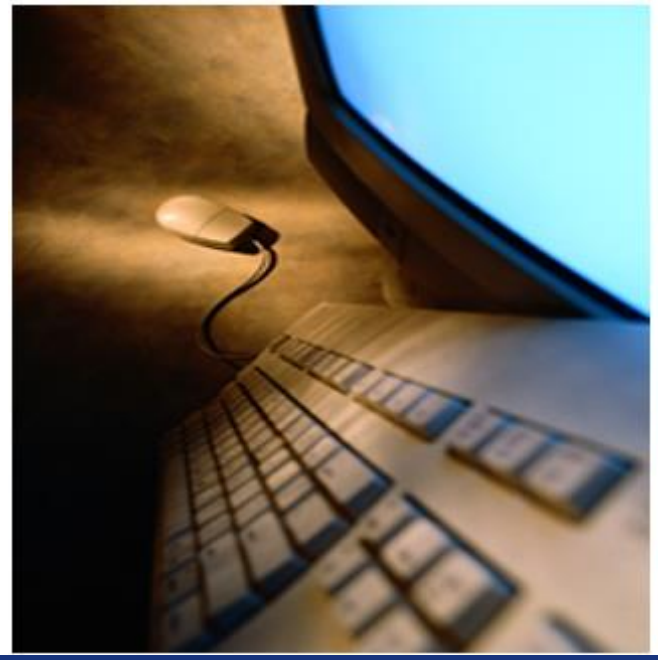
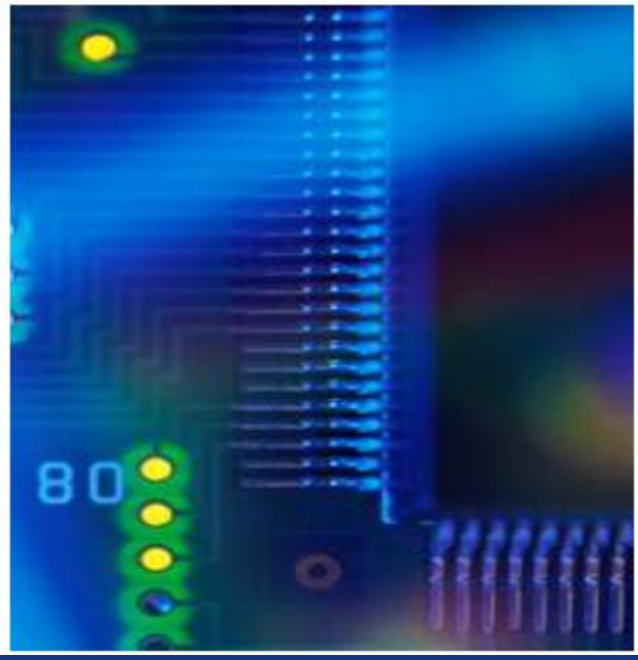
UNIVERSITAS
BUDI LUHUR




FAKULTAS TEKNOLOGI INFORMASI

# PENAMBANGAN DATA

# [ KP368 / 3 SKS ]


Pertemuan 11

# UNSUPERVISED LEARNING:
# K-MEANS CLUSTERING


# Tujuan Pembelajaran

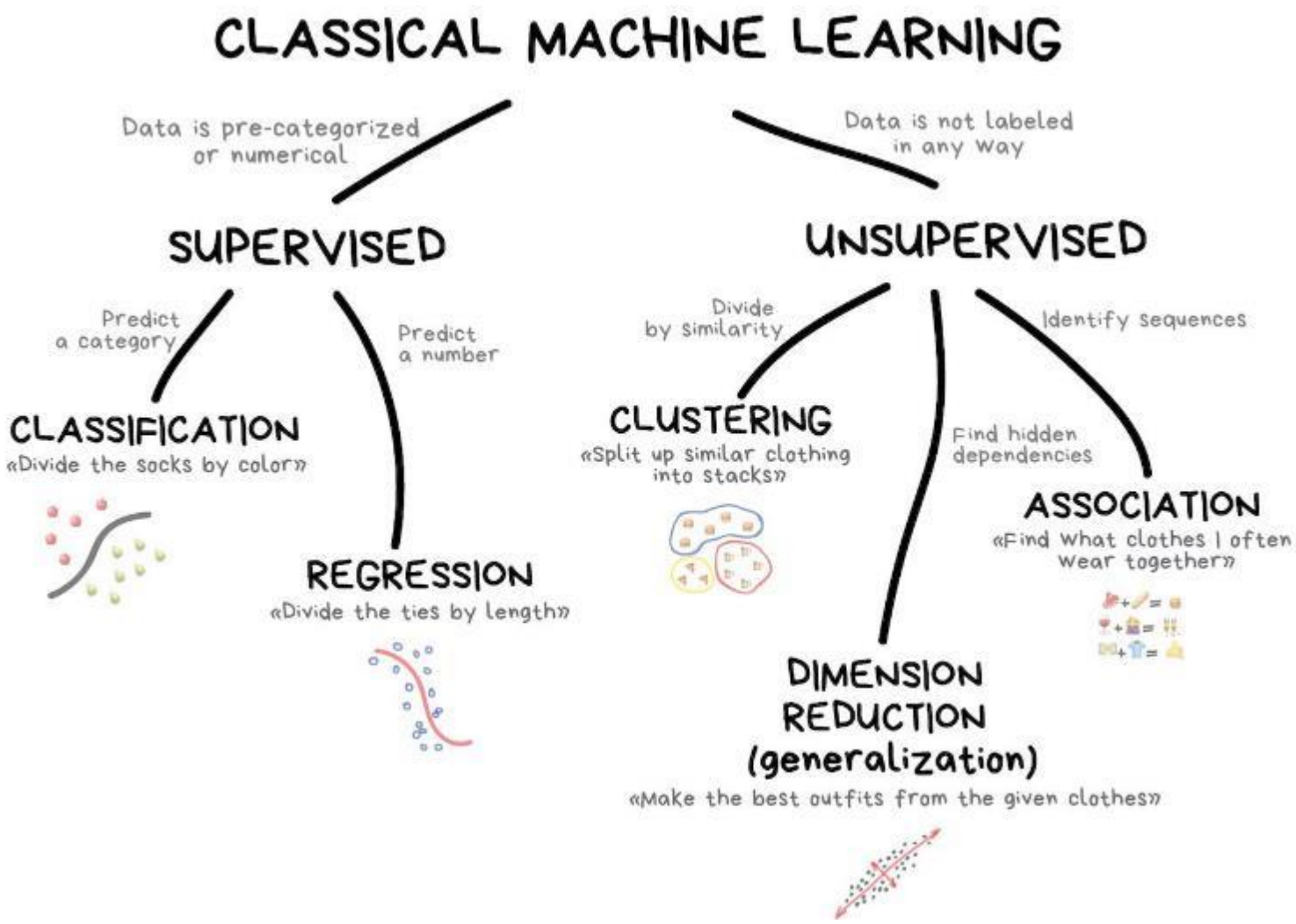
- Mahasiswa dapat memahami konsep pembelajaran tidak tersupervisi (unsupervised learning), khususnya menggunakan metode K-Means clustering
- Mahasiswa dapat menjelaskan beberapa penerapan algoritma klastering dalam menyelesaikan berbagai masalah.


# Outline

- Pengantar Clustering
- Penerapan Algoritma Clustering
- Macam-macam  Algoritma Clustering
- Langkah Algoritma K-Means Clustering
- Contoh Perhitungan Algoritma K-Means
- Optimasi Nilai k pada K-Means

# Supervised vs Unsupervised Learning

- Supervised
  - Classification
  - Regression
- Unsupervised
  - Clustering
  - Association
  - Dimension Reduction




# Pengantar

- Diberikan data profil pelanggan (customer), bagaimana memilih data pelanggan yang potensial untuk ditawarkan produk tertentu?

| | Customer Id | Age | Edu | Years Employed | Income | Card Debt | Other Debt | Defaulted | Address | DebtIncomeRatio |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 41 | 2 | 6 | 19 | 0.124 | 1.073 | 0.0 | NBA001 | 6.3 |
| **1** | 2 | 47 | 1 | 26 | 100 | 4.582 | 8.218 | 0.0 | NBA021 | 12.8 |
| **2** | 3 | 33 | 2 | 10 | 57 | 6.111 | 5.802 | 1.0 | NBA013 | 20.9 |
| **3** | 4 | 29 | 2 | 4 | 19 | 0.681 | 0.516 | 0.0 | NBA009 | 6.3 |
| **4** | 5 | 47 | 1 | 31 | 253 | 9.308 | 8.908 | 0.0 | NBA008 | 7.2 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **845** | 846 | 27 | 1 | 5 | 26 | 0.548 | 1.220 | NaN | NBA007 | 6.8 |
| **846** | 847 | 28 | 2 | 7 | 34 | 0.359 | 2.021 | 0.0 | NBA002 | 7.0 |
| **847** | 848 | 25 | 4 | 0 | 18 | 2.802 | 3.210 | 1.0 | NBA001 | 33.4 |
| **848** | 849 | 32 | 1 | 12 | 28 | 0.116 | 0.696 | 0.0 | NBA012 | 2.9 |
| **849** | 850 | 52 | 1 | 16 | 64 | 1.866 | 3.638 | 0.0 | NBA025 | 8.6 |

850 rows × 10 columns

Kita diminta untuk mengelompokkan data customer di samping, berdasarkan kesamaan profil pelanggan

↓

Customer Segmentation

↓

Clustering

# Pengantar

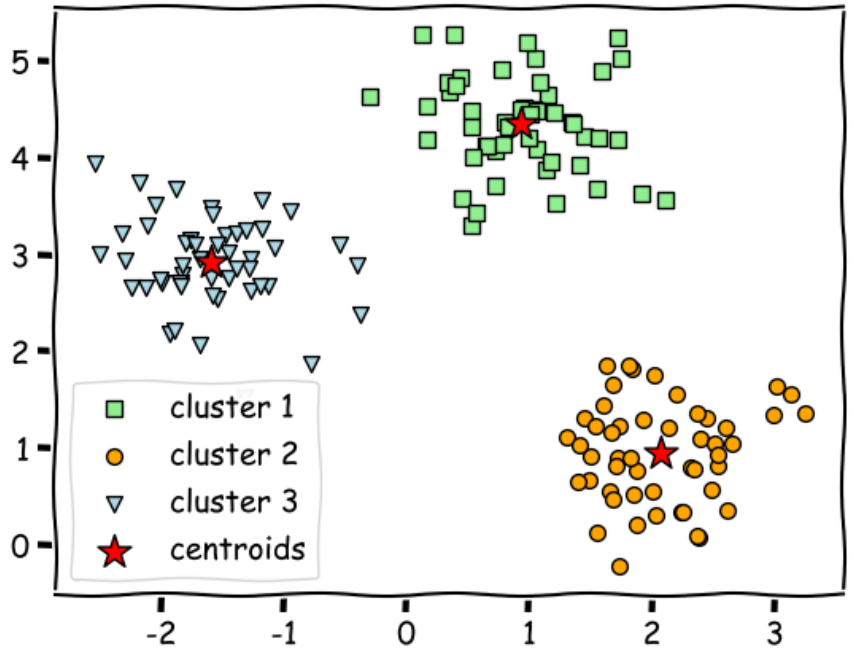
- Contoh hasil clustering / segmentasi pelanggan

| | Customer Id | Age | Edu | Years Employed | Income | Card Debt | Other Debt | Defaulted | DebtIncomeRatio | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 41 | 2 | 6 | 19 | 0.124 | 1.073 | 0.0 | 6.3 | 2 |
| **1** | 2 | 47 | 1 | 26 | 100 | 4.582 | 8.218 | 0.0 | 12.8 | 0 |
| **2** | 3 | 33 | 2 | 10 | 57 | 6.111 | 5.802 | 1.0 | 20.9 | 2 |
| **3** | 4 | 29 | 2 | 4 | 19 | 0.681 | 0.516 | 0.0 | 6.3 | 2 |
| **4** | 5 | 47 | 1 | 31 | 253 | 9.308 | 8.908 | 0.0 | 7.2 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **845** | 846 | 27 | 1 | 5 | 26 | 0.548 | 1.220 | NaN | 6.8 | 2 |
| **846** | 847 | 28 | 2 | 7 | 34 | 0.359 | 2.021 | 0.0 | 7.0 | 2 |
| **847** | 848 | 25 | 4 | 0 | 18 | 2.802 | 3.210 | 1.0 | 33.4 | 2 |
| **848** | 849 | 32 | 1 | 12 | 28 | 0.116 | 0.696 | 0.0 | 2.9 | 2 |
| **849** | 850 | 52 | 1 | 16 | 64 | 1.866 | 3.638 | 0.0 | 8.6 | 0 |

Setiap pelanggan berhasil dikelompokkan

# Apa itu Clustering?

- **Cluster** adalah sekumpulan data / object yang memiliki **kesamaan (similarity)** diantara setiap anggota klaster, atau **ketidaksamaan (dissimilarity)** dengan data pada klaster yang lain




# Contoh Penerapan Clustering

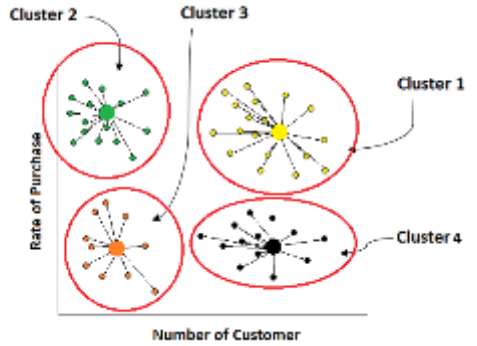
- Retail / Marketing
  - Analisis pola transaksi yang dilakukan pelanggan
  - Rekomendasi buku, film atau produk baru untuk pelanggan baru
- Perbankan
  - Deteksi fraud dalam transaksi perbankan
  - Pengelompokan nasabah (program loyalitas nasabah)
- Asuransi
  - Deteksi fraud dalam klaim asuransi
  - Analisis resiko asuransi bagi pelanggan
- Berita dan Penerbitan
  - Kategorisasi berita secara otomatis
  - Rekomendasi artikel / berita baru

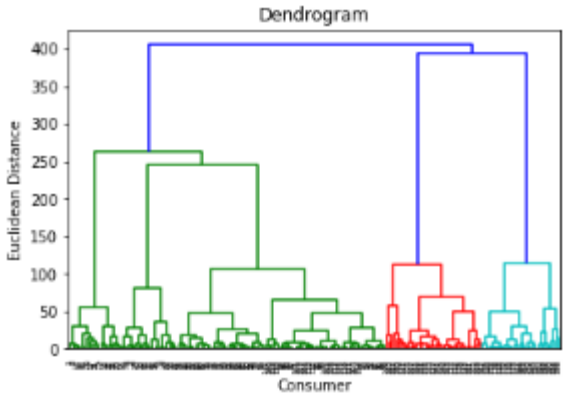

# Penggunaan Algoritma Clustering

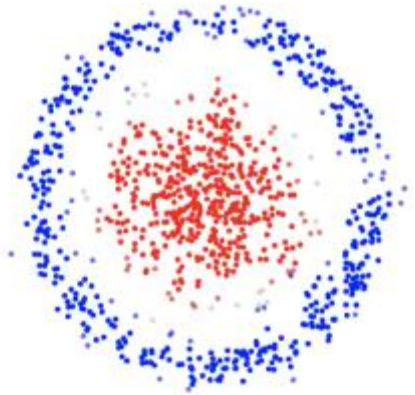
- Exploratory Data Analysis (EDA)
- Generate Rangkuman (summary generation)
- Deteksi pencilan (outlier detection)
- Mencari duplikat (finding duplicates)
- Tahap pra-pemrosesan data (Data pre-processing)
- Kompresi data / image
- Optimasi algoritma k-NN
- dll

# Algoritma Clustering

- Partitioning-based clustering
  - K-Means,
  - K-Medoid,
  - K-Medians,
  - Fuzzy C-Means, dll
- Hierarchical Clustering
  - Agglomerative
  - Divisive, dll
- Density-based Clustering
  - DBSCAN, dll




## BAGIAN 1: PARTITIONING-BASED CLUSTERING

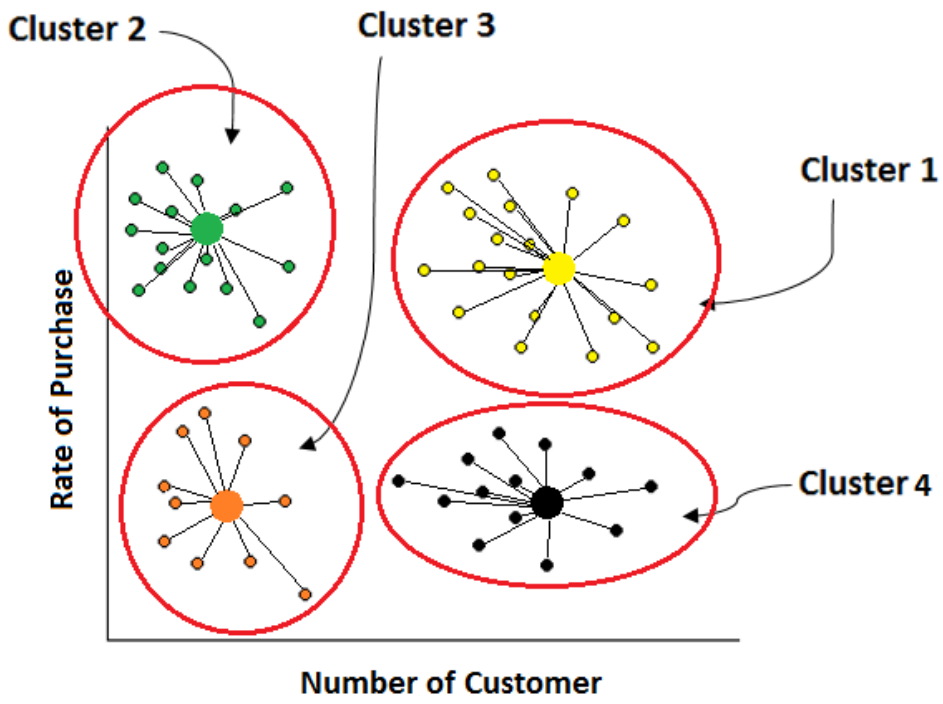
# ALGORITMA K-MEANS


# K-Means Clustering

- Algoritma K-Means adalah salah satu algoritma clustering yang bersifat iteratif yang mencoba untuk mempartisi dataset menjadi subkelompok non-overlapping berbeda yang ditentukan oleh K (cluster) yang mana setiap titik data hanya dimiliki oleh satu kelompok.
- K-Means mencoba membuat titik data intra-cluster semirip mungkin dengan titik data yang lain pada satu cluster.
- K-Means menetapkan poin data ke cluster sedemikian rupa sehingga jumlah jarak kuadrat antara titik data dan pusat massa cluster (centroid) adalah minimal.
- Semakin sedikit variasi dalam sebuah cluster, semakin homogen (serupa) titik data dalam cluster yang sama.

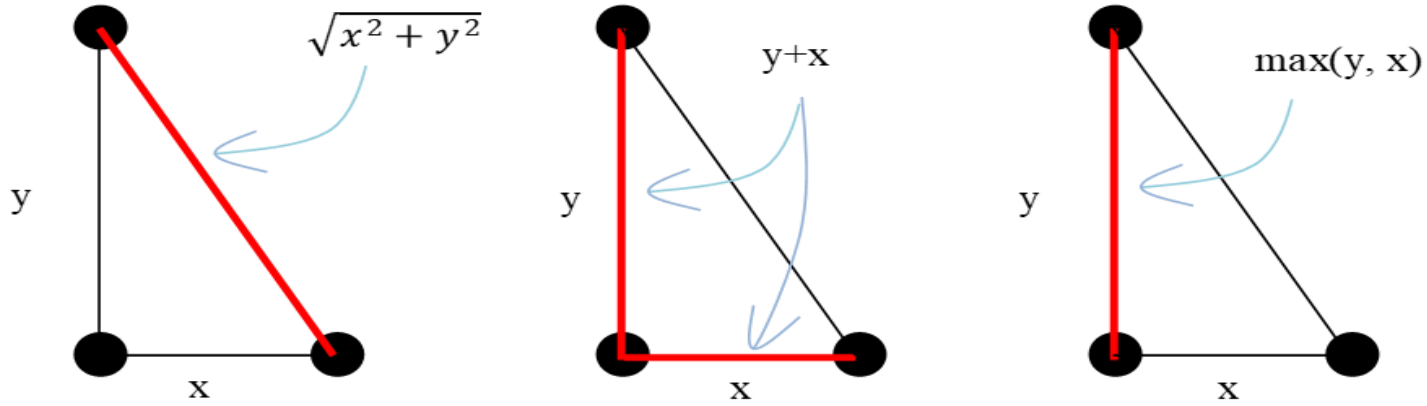
# K-Means Clustering: Similarity / Dissimilarity

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

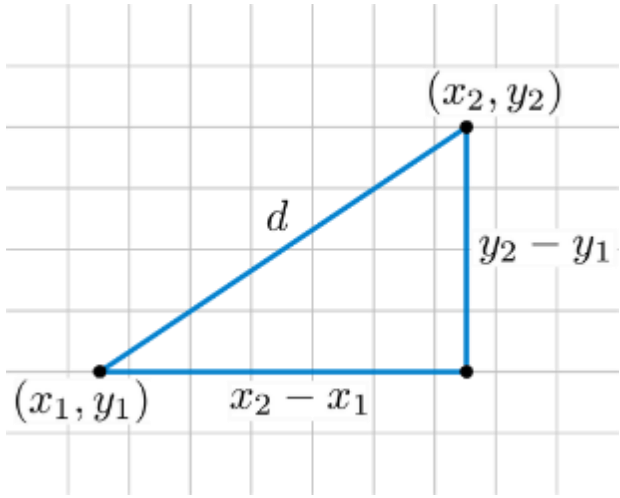


# K-Means Clustering: Metode Perhitungan Similarity

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




(a) jarak *Euclidean* (b) Jarak *city-block* (c) Jarak *Chebychef*




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

# Langkah / Algoritma K-Means Clustering

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.

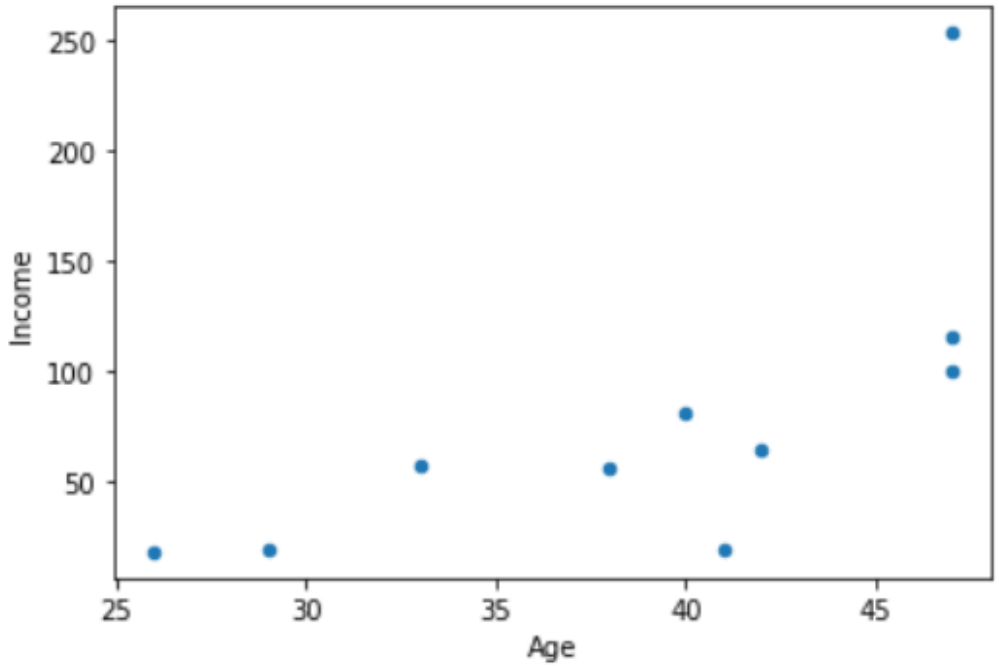
# Ilustrasi Cara Kerja Algoritma K-Means

# Contoh Kasus: Klasterisasi Pelanggan

## Data Pelanggan

| 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 |

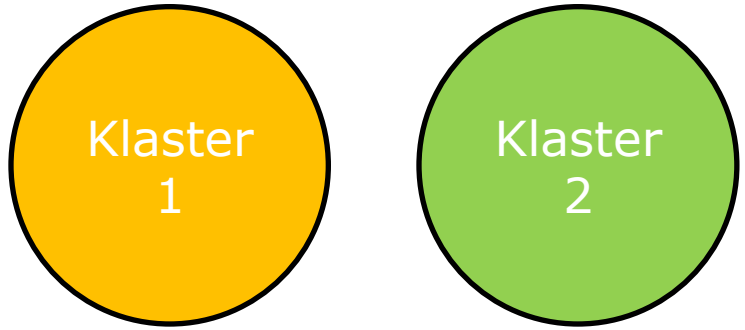
Diketahui data pelanggan sebagaimana tabel di samping, kita diminta mengelompokkan data pelanggan menjadi 2 (dua) kelompok.



# Contoh Kasus: Klasterisasi Pelanggan

**1. Tentukan jumlah klaster. Dalam contoh kasus ini kita gunakan nilai k=2**

| 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 |



# Contoh Kasus: Klasterisasi Pelanggan

## 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 |

**Cara penentuan centroid awal:**

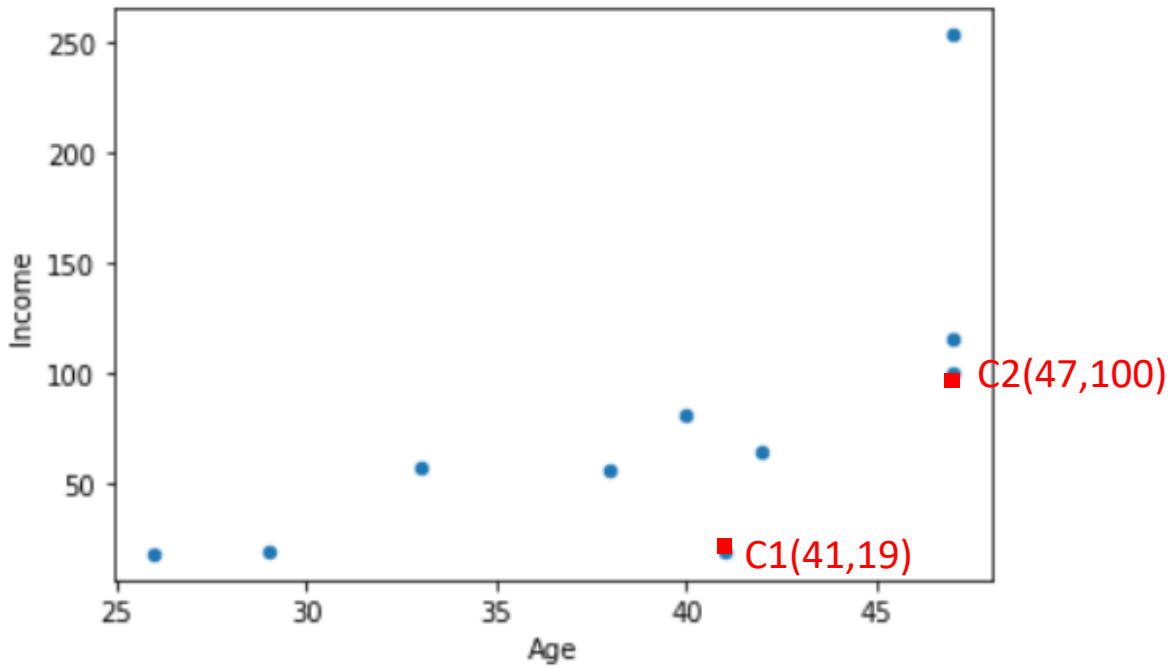
1. Memilih salah satu data untuk atribut "Age" dan "Income" secara acak
2. Membangkitkan bilangan acak sesuai rentang nilai "Age" dan "Income"

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

# Contoh Kasus: Klasterisasi Pelanggan

## 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 |




# Contoh Kasus: Klasterisasi Pelanggan

**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 | $\sqrt{(47-41)^2+(115-19)^2}=96{,}19$ | $\sqrt{(47-47)^2+(115-100)^2}=15{,}0$ |

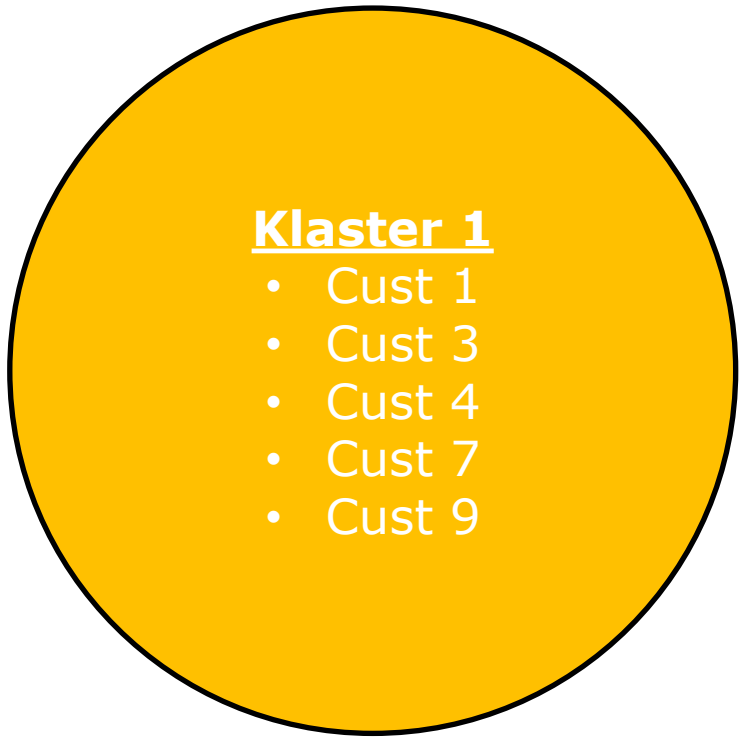
# Contoh Kasus: Klasterisasi Pelanggan

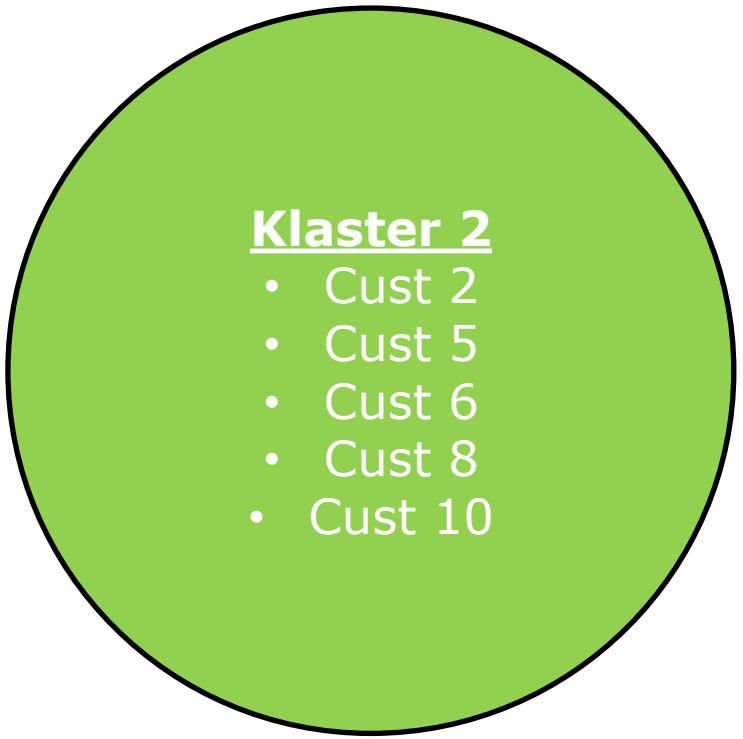
**4. Masukkan setiap titik data ke dalam klaster berdasarkan jarak terdekat dengan centroid**

| 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 |

# Contoh Kasus: Klasterisasi Pelanggan

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



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

# Contoh Kasus: Klasterisasi Pelanggan

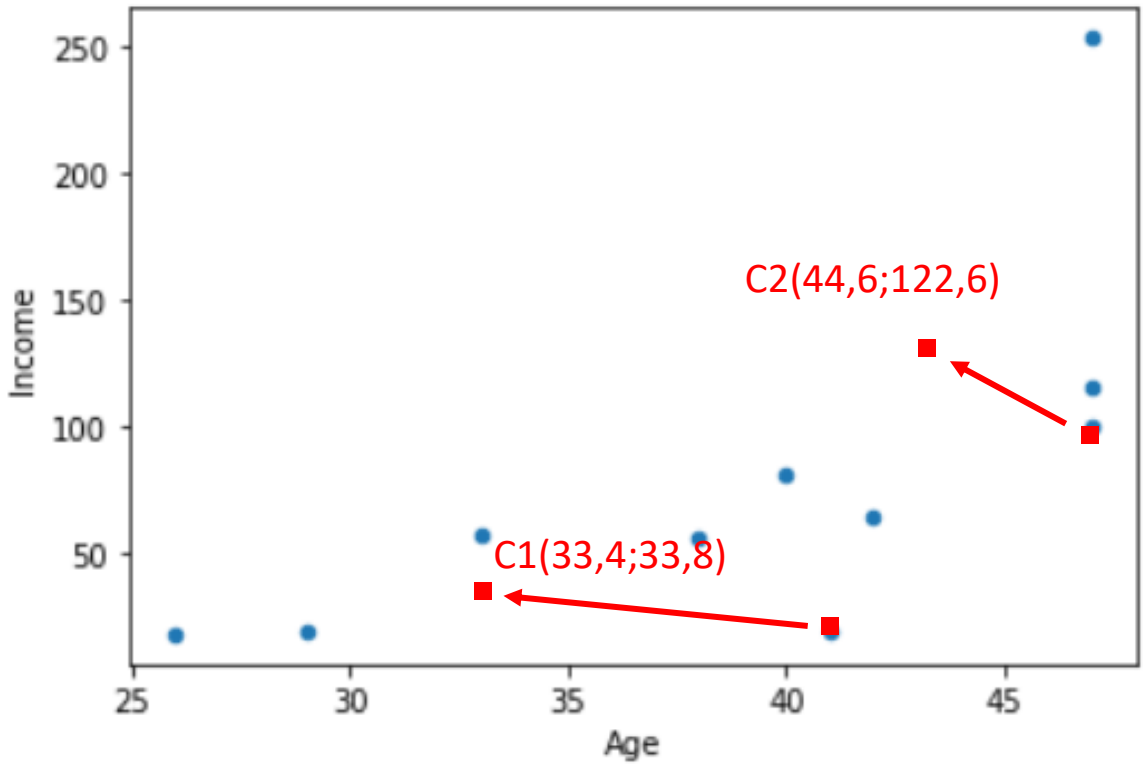
**5. Untuk setiap klaster, hitung nilai centroid baru berdasarkan rerata (means) dari setiap data di dalam klaster**

| 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)**

# Contoh Kasus: Klasterisasi Pelanggan

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




# Contoh Kasus: Klasterisasi Pelanggan

## 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 |

# Contoh Kasus: Klasterisasi Pelanggan

## 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

# Contoh Kasus: Klasterisasi Pelanggan

| 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

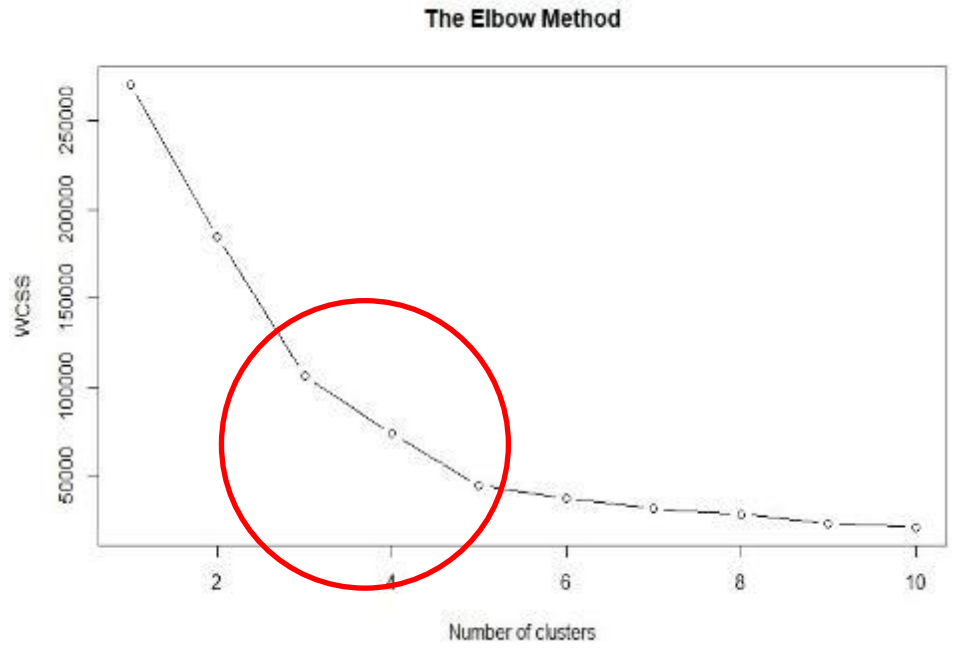
- Salah satu faktor krusial baik tidaknya metode K-Means adalah **jumlah klusternya (nilai K)**. Hasil pengelompokan akan menghasilkan analisa yang berbeda untuk jumlah klaster yang berbeda juga.
- **Semakin kecil nilai K** (misal 2), maka pembagian kluster menjadi cepat, namun mungkin ada informasi 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)**.



# Kesimpulan Bagian 1

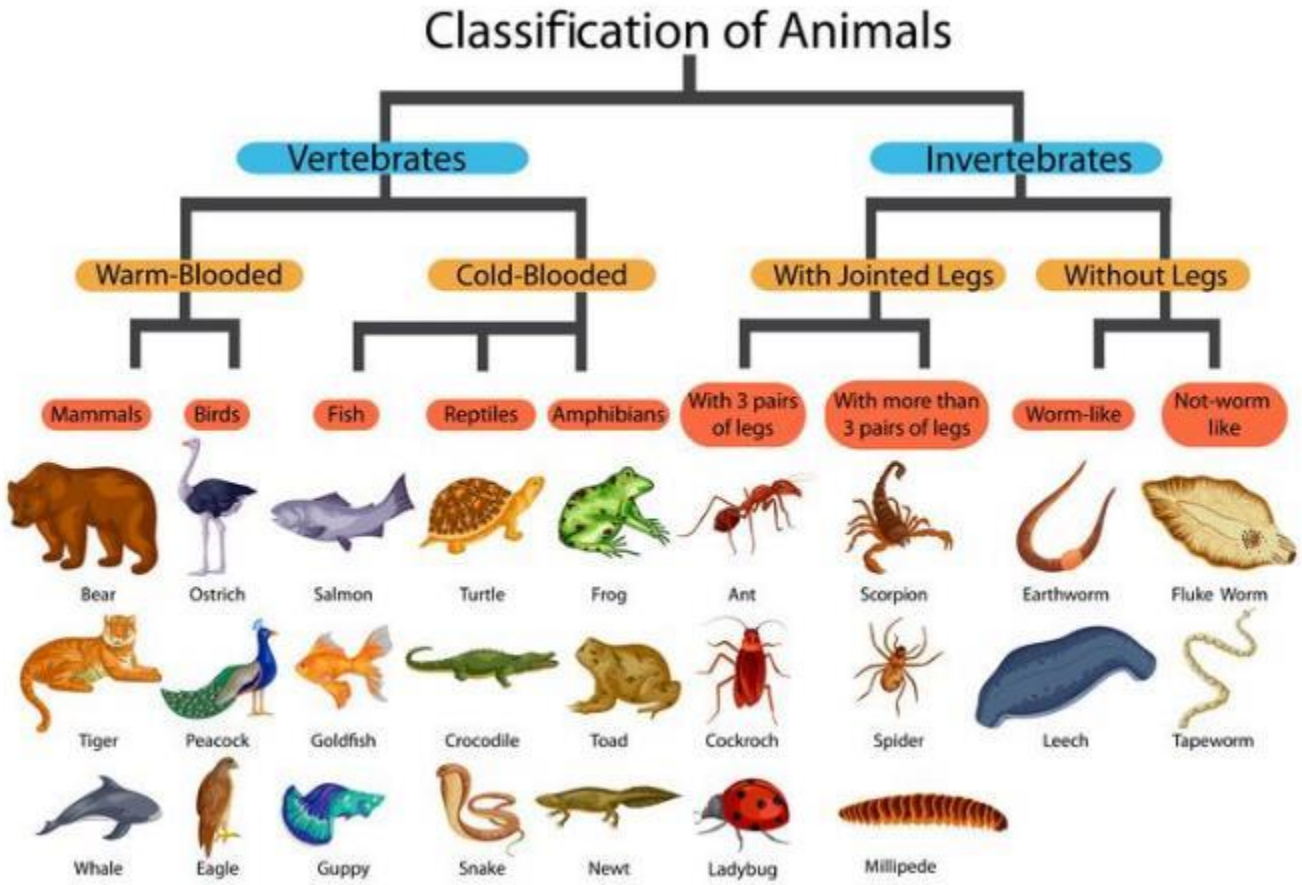
- Clustering merupakan salah satu metode pembelajaran tidak tersupervisi
- Metode clustering dibagi menjadi 3 jenis: partitioning-based, hierarchical, dan density-based.
- Algoritma K-means adalah salah satu algoritma clustering yang bersifat iteratif yang mempartisi dataset menjadi subkelompok non-overlapping berbeda yang ditentukan oleh K (cluster)
- Algoritma K-Means:
  - Relatif efisien untuk data kecil hingga besar
  - Menghasilkan kelompok klaster
  - Memerlukan inisialisasi nilai k


Bagian 2

# HIERARCHICAL CLUSTERING

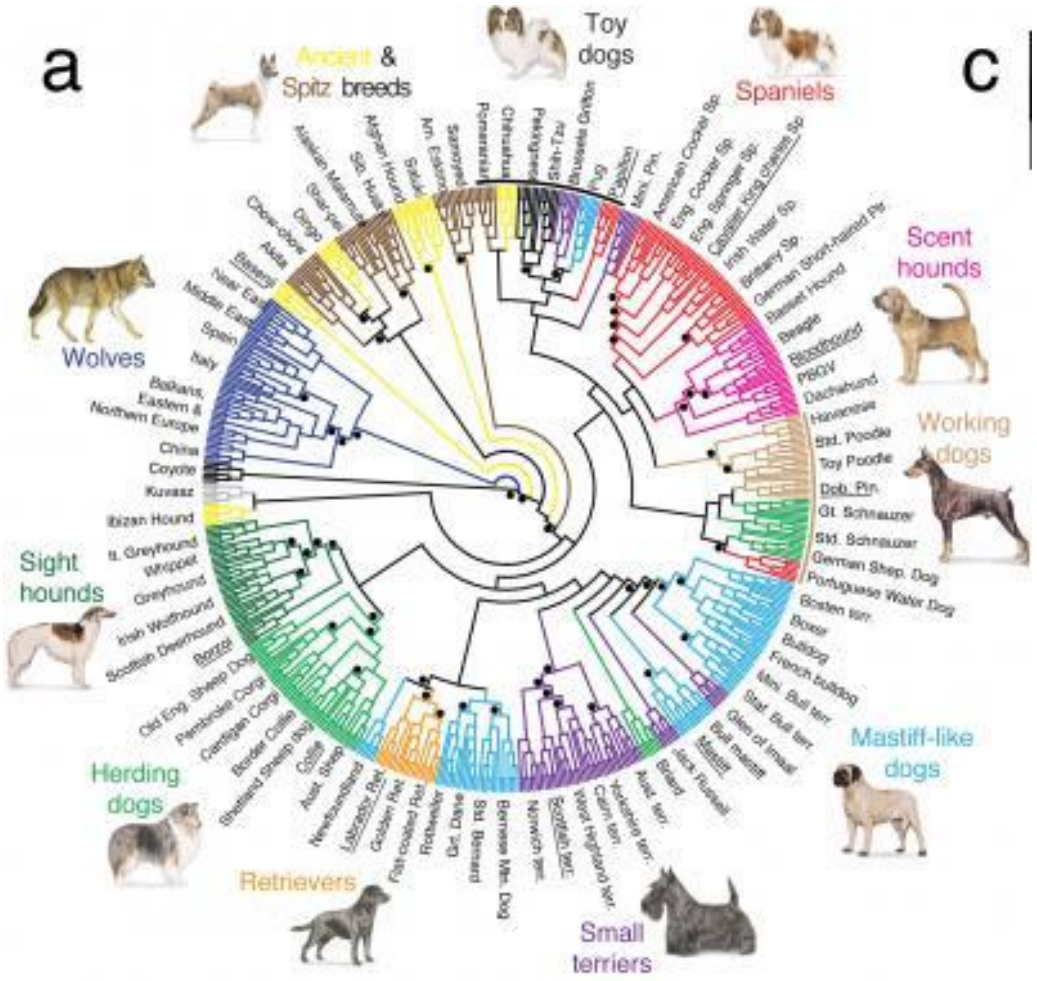
# Animal Kingdom: Klasifikasi Hewan

- Para ahli biologi mengklasifikasi makhluk hidup secara hierarkis sesuai tingkatan taksonominya: **kingdom**, **divisi**, **class**, **ordo**, **familia**, **genus**, hingga **species**.
- Klasifikasi makhluk hidup (contohnya hewan), didasarkan pada **kesamaan ciri** dari makhluk hidup tersebut. Contoh: **beruang** dan **burung elang** memiliki kesamaan ciri: hewan bertulang belakang, dan berdarah panas.



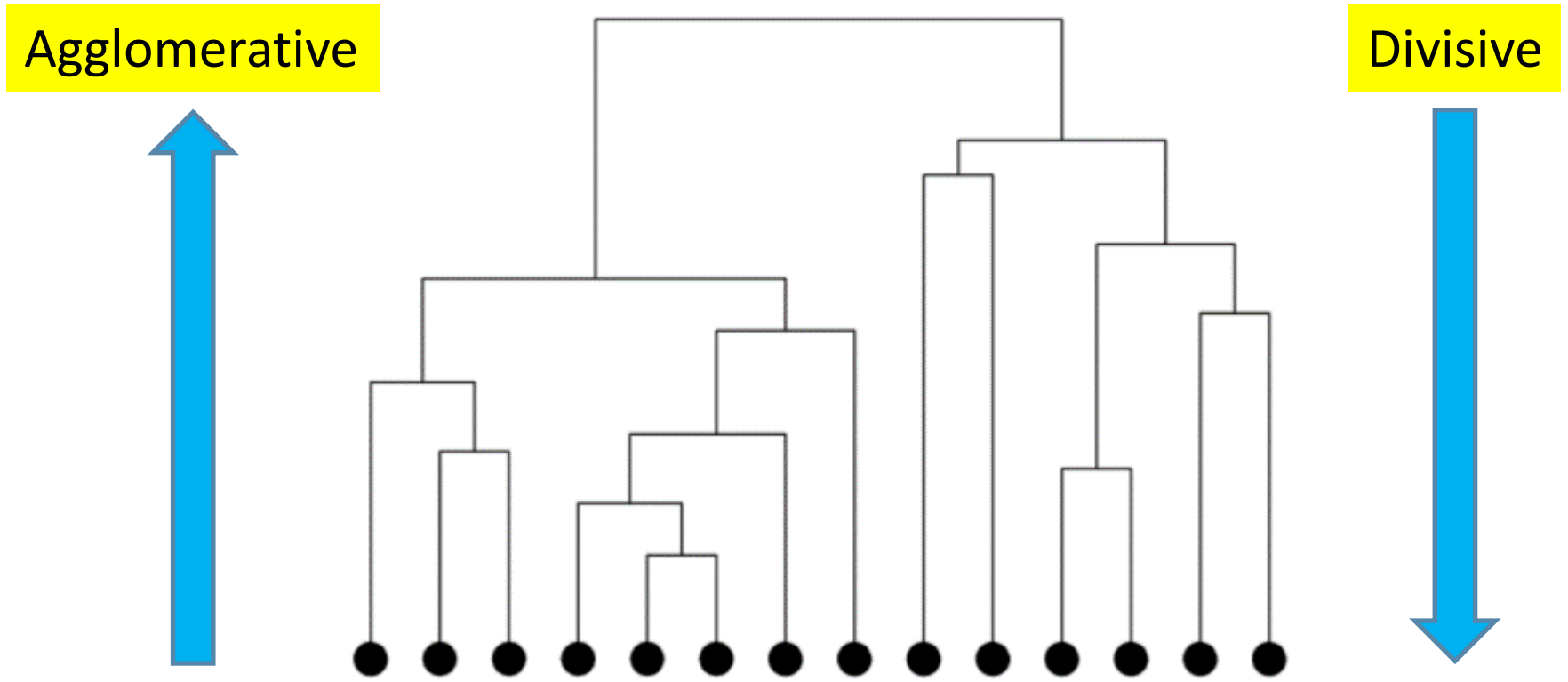
# Klasifikasi Peranakan Anjing

- Pada tahun 2010, sejumlah ilmuwan internasional yang dikoordinir oleh UCLA mempublikasikan sebuah **dendogram** (lihat gambar di samping) yang merupakan hasil pengelompokan lebih dari **900 spesies anjing dari 85+ peranakan** di seluruh dunia.
- Pengelompokan dilakukan dengan menganalisis **kesamaan ciri** pada sekitar **48.000 data genetic** spesies anjing.

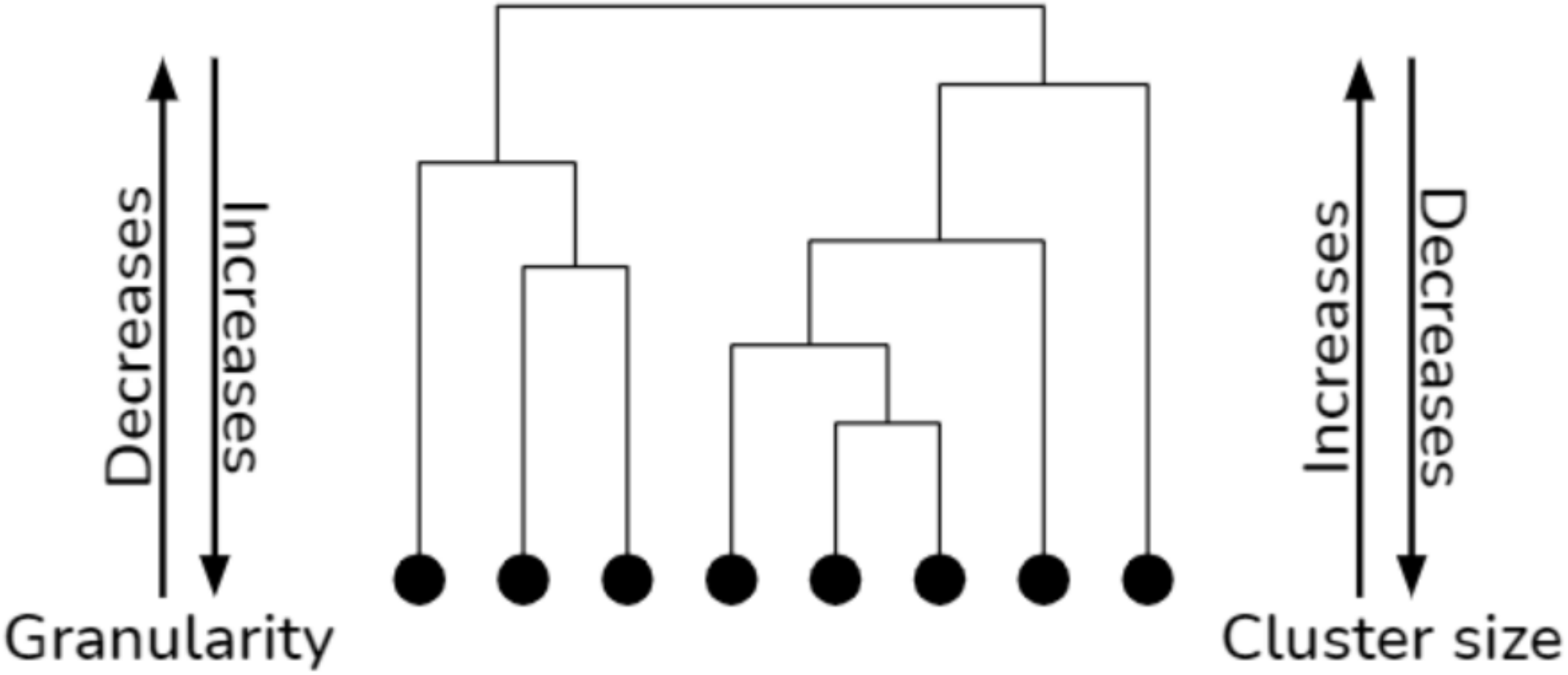


# Hierarchical Clustering

Metode Hierarchical Clustering ada **2 (dua) pendekatan**:

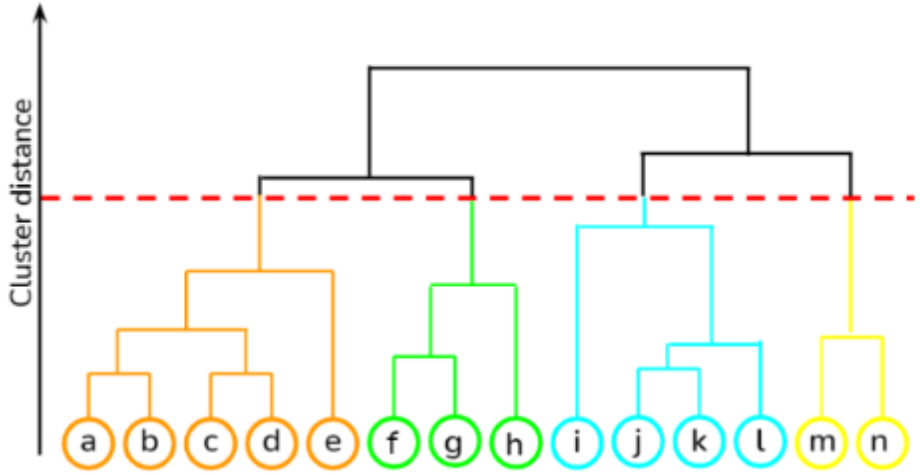


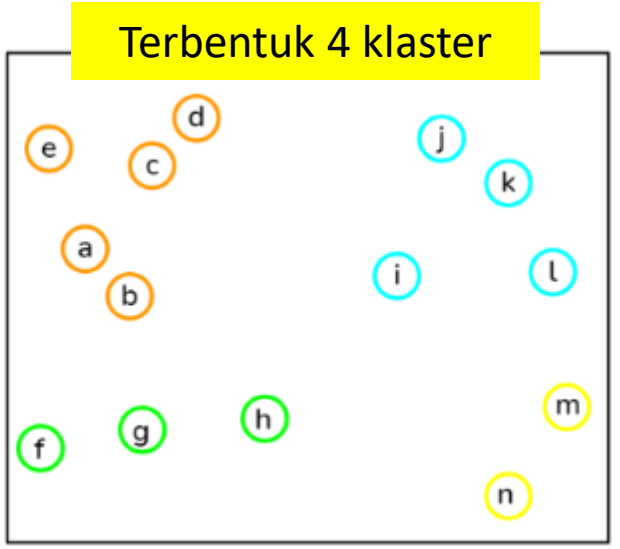
# Hierarchical Clustering: Granularity vs Cluster Size



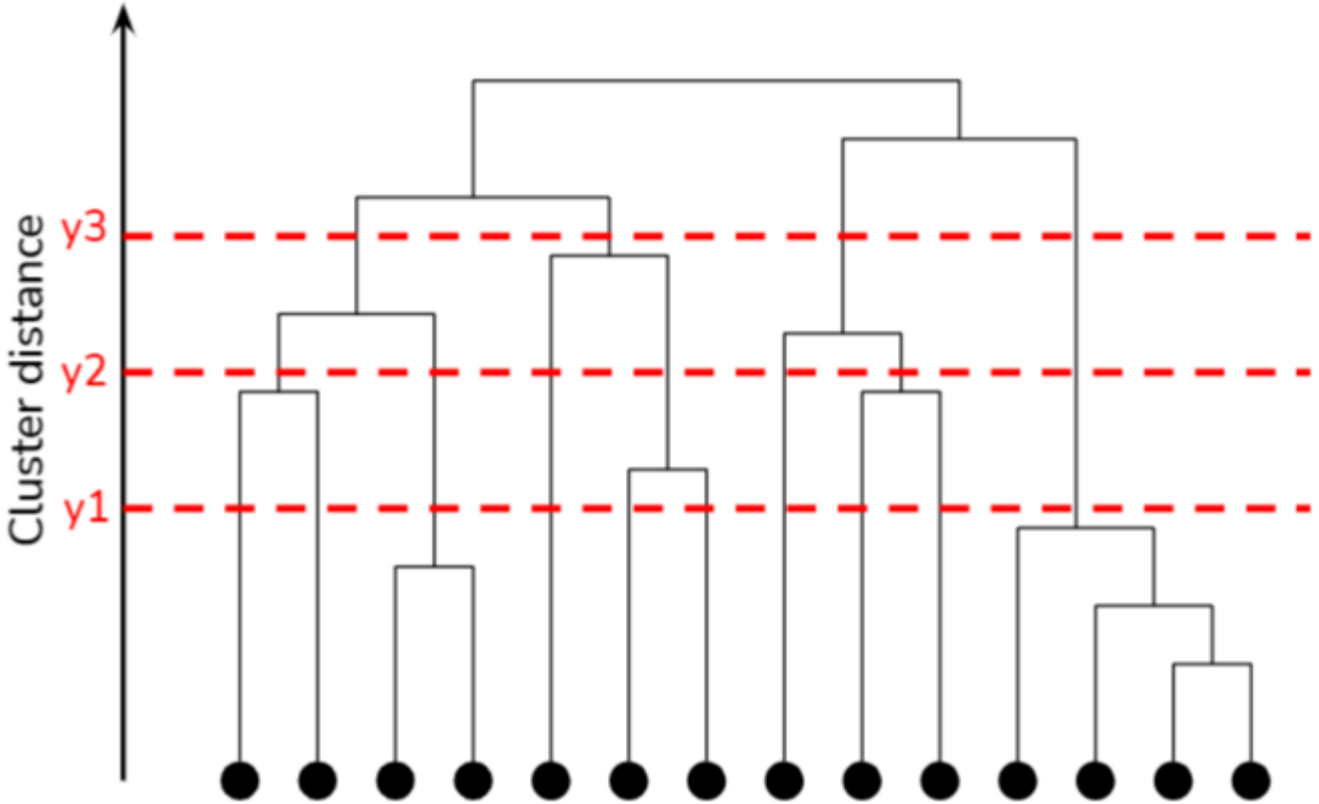
# Hierarchical Clustering: Jumlah Klaster

- Berbeda dengan metode K-Means, pada metode Hierarchical Clustering, **jumlah klaster tidak ditentukan di awal** proses klasterisasi.
- Jumlah klaster ditentukan berdasarkan **kebutuhan pengguna** setelah dendogram terbentuk dengan melakukan pemotongan pada level tertentu.



# Hierarchical Clustering: Jumlah Klaster

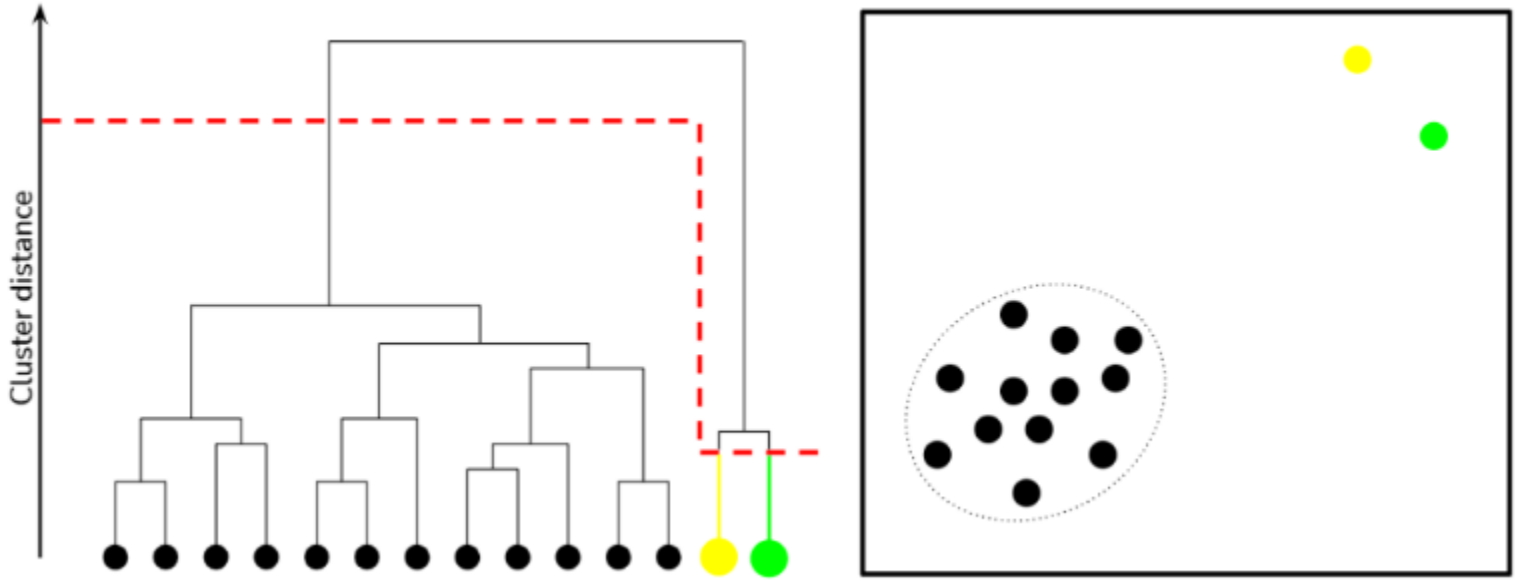



- Pemotongan y1 membentuk 10 klaster
- Pemotongan y2 membentuk 7 klaster
- Pemotongan y3 membentuk 4 klaster

Semakin rendah granularity-nya, semakin sedikit jumlah klaster yang dihasilkan

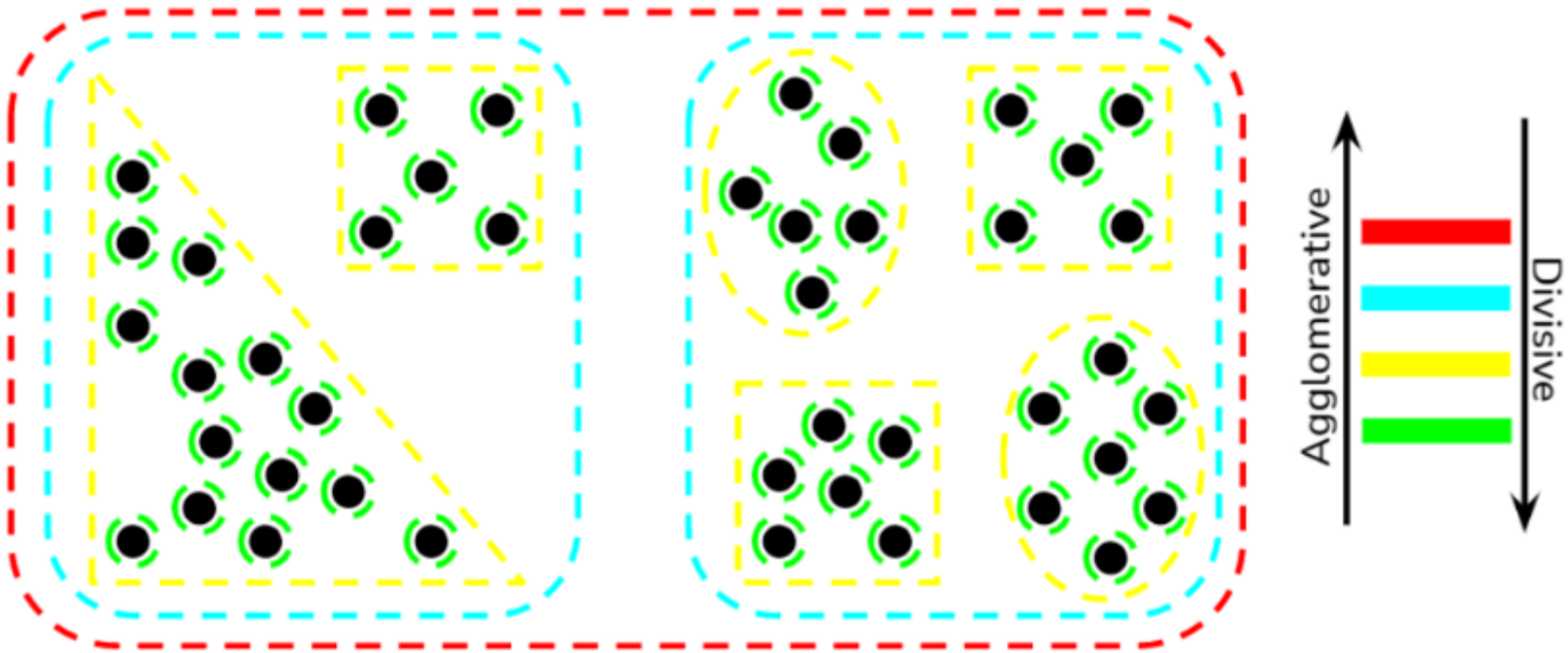
# Hierarchical Clustering: Jumlah Klaster

- Pemotongan dendrogram sangat bergantung pada tujuan, kebutuhan dan analisis yang diinginkan.
- Contoh berikut ini pemotongan dendogram untuk memisahkan data outlier (**outlier removal**)



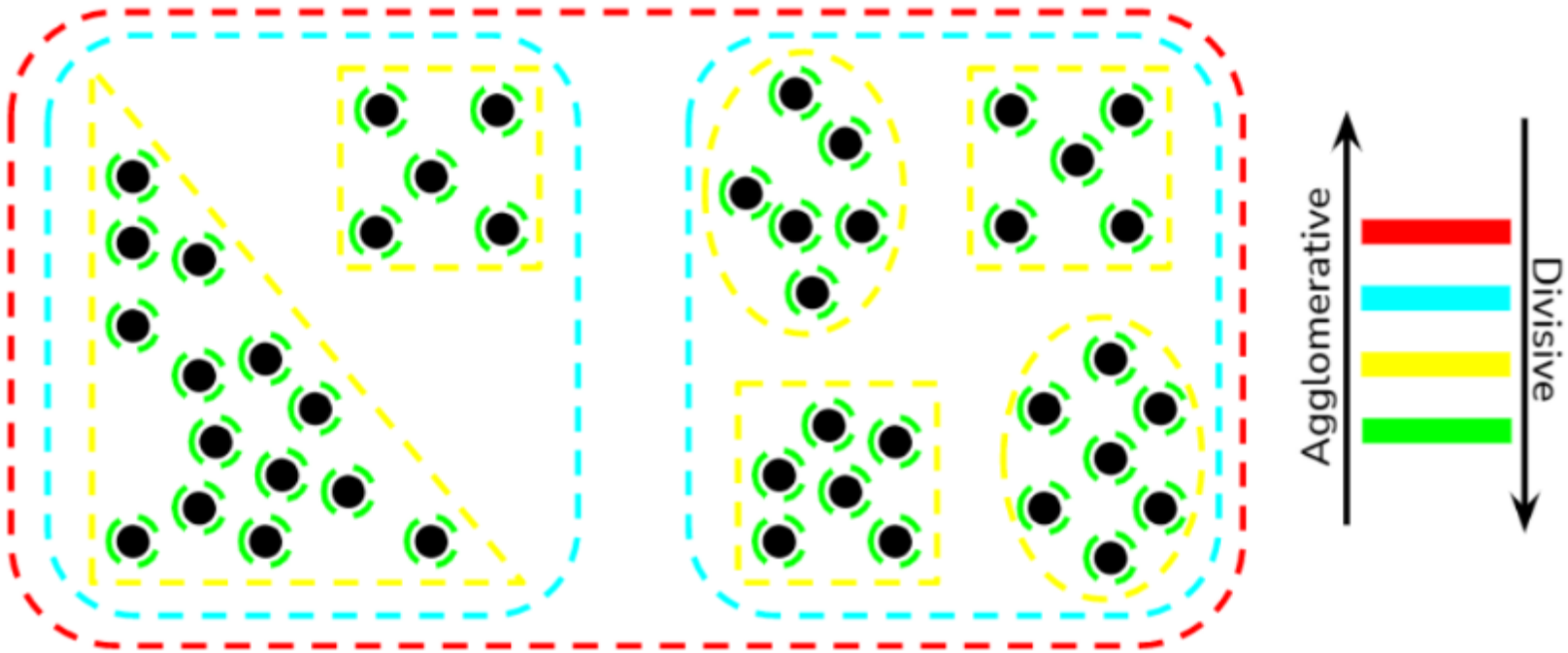
# Agglomerative Clustering

- Metode Agglomerative Clustering menyusun hirarki berdasarkan kedekatan (kesamaan) setiap individu dengan penggabungan klaster secara progresif (berjenjang).



# Divisive Clustering

- Metode Divisive Clustering menyusun hirarki dengan memecah (split) data ke dalam klaster secara progresif (berjenjang).




# Hierarchical Clustering: Kelebihan dan Kekurangan

| Kelebihan | Kekurangan |
|---|---|
| Tidak perlu menentukan jumlah klaster di awal proses | Dalam pembentukan klaster, algoritma tidak dapat kembali ke tahap sebelumnya |
| Mudah diterapkan | Secara umum pembentukan klaster (dendogram) membutuhkan waktu lama |
| Menghasilkan dendogram yang dapat mempermudah pemahaman data | Terkadang sulit menentukan jumlah klaster dari dendogram yang terbentuk. |


# Hierarchical Clustering vs K-Means

| K-Means | Hierarchical Clustering |
|---|---|
| Lebih efisien | Relative lambat untuk jumlah data yang besar |
| Jumlah klaster ditentukan di awal proses | Tidak perlu menentukan jumlah klaster di awal proses |
| Hanya terdapat satu hasil klaster untuk setiap kali proses (sesuai jumlah klaster) | Dapat menghasilkan beberapa hasil klaster (sesuai kebutuhan) |
| Berpotensi menghasilkan klaster yang berbeda untuk setiap proses, karena perbedaan centroid awal | Selalu menghasilkan klaster (dendogram) yang sama |

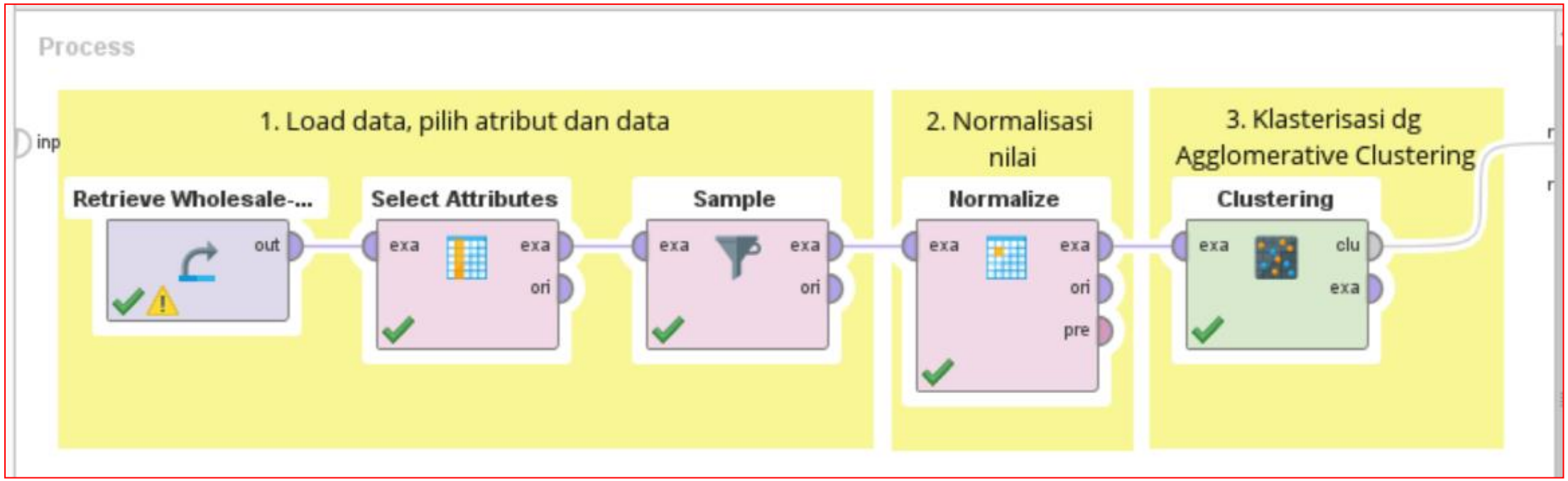
# Latihan: Agglomerative Clustering

- Dataset: **Wholesale-customers-data.csv**
- Data jumlah penjualan kebutuhan sehari-hari dari sebuah toko waralaba yang disajikan per wilayah

| | Channel | Region | Fresh | Milk | Grocery | Frozen | Detergents_Paper | Delicassen |
|---|---|---|---|---|---|---|---|---|
| 0 | 2 | 3 | 12669 | 9656 | 7561 | 214 | 2674 | 1338 |
| 1 | 2 | 3 | 7057 | 9810 | 9568 | 1762 | 3293 | 1776 |
| 2 | 2 | 3 | 6353 | 8808 | 7684 | 2405 | 3516 | 7844 |
| 3 | 1 | 3 | 13265 | 1196 | 4221 | 6404 | 507 | 1788 |
| 4 | 2 | 3 | 22615 | 5410 | 7198 | 3915 | 1777 | 5185 |

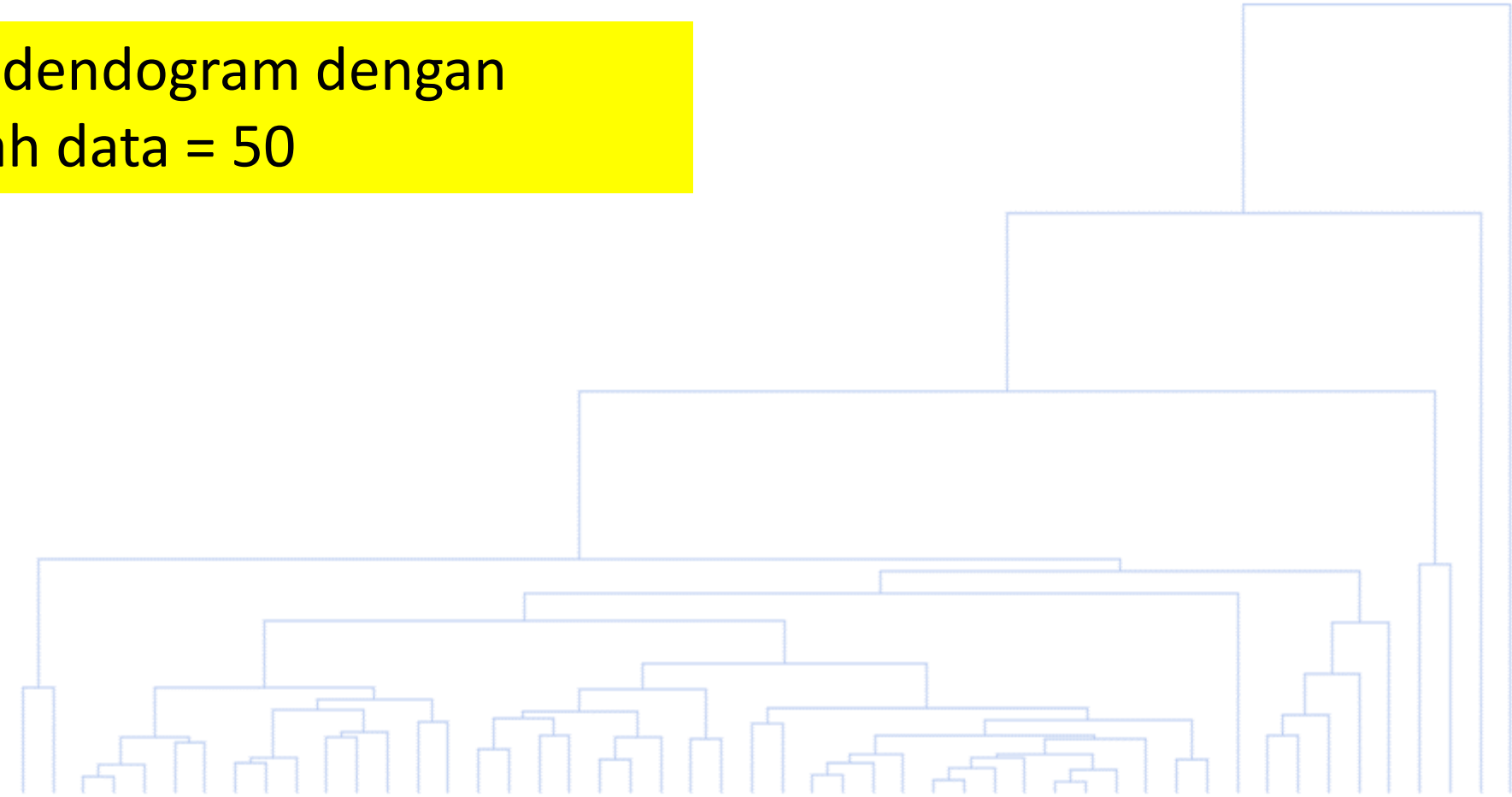
Lakukan klasterisasi dengan Agglomerative Clustering berdasarkan data jumlah penjualan tiap produk

# Latihan: Agglomerative Clustering: Pemodelan di Rapidminer



# Latihan: Agglomerative Clustering

Hasil dendogram dengan jumlah data = 50

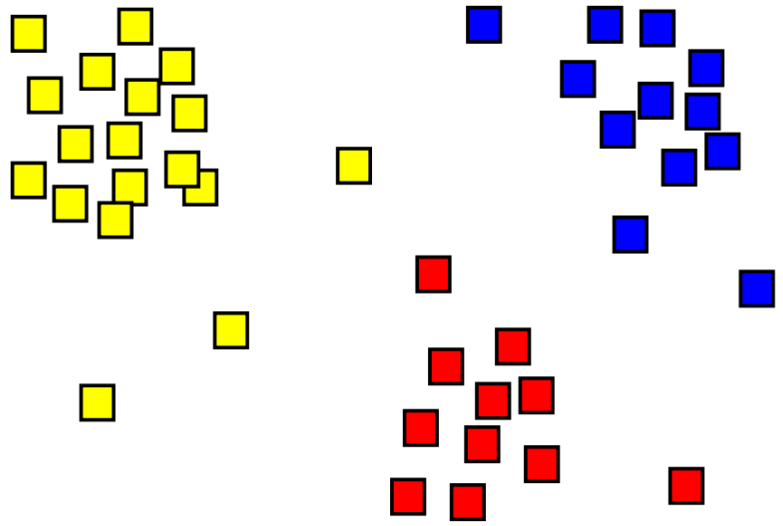



Bagian 3

# DENSITY-BASED CLUSTERING: DBSCAN

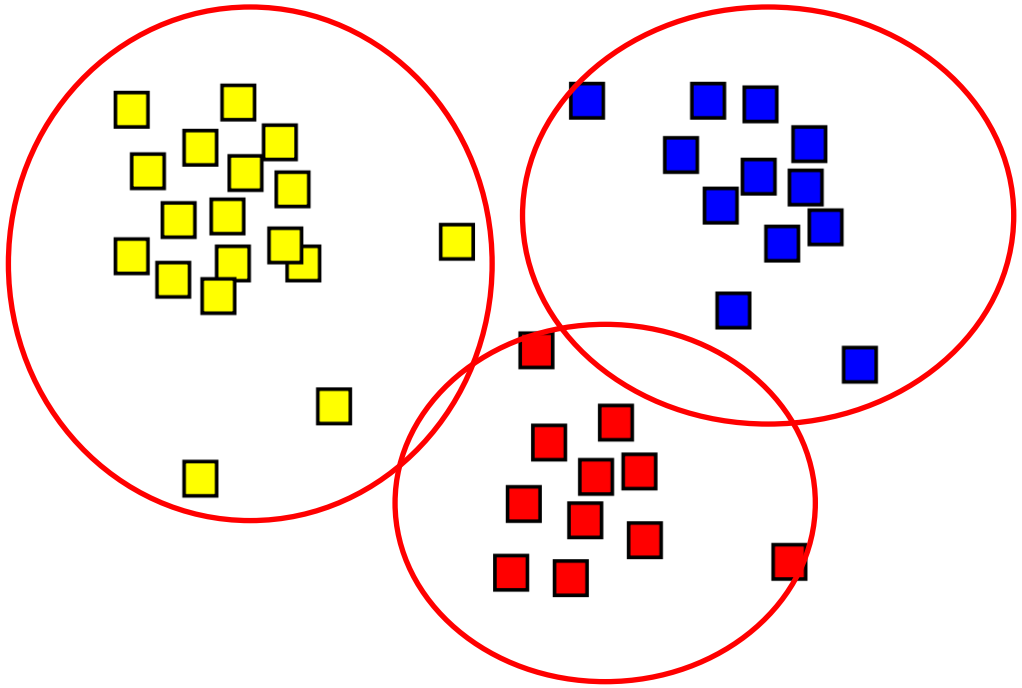
# K-Means vs Density-based Clustering

- **K-Means** akan **memasukkan setiap data** ke dalam suatu klaster, walaupun data tersebut terpisah cukup jauh dari pusat data (centroid).
- **Density-based** akan mengelompokkan data yang memiliki **densitas (kerapatan) tinggi** dan mengabaikan data yang berada di luar kelompoknya.



# K-Means vs Density-based Clustering

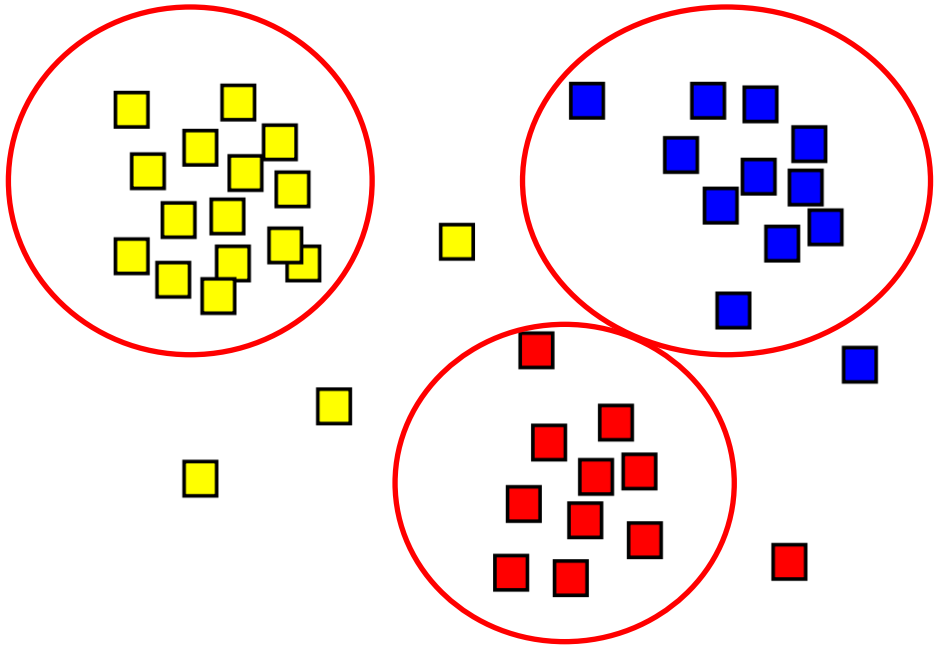
- **K-Means** akan **memasukkan setiap data** ke dalam suatu klaster, walaupun data tersebut terpisah cukup jauh dari pusat data (centroid).
- **Density-based** akan mengelompokkan data yang memiliki **densitas (kerapatan) tinggi** dan mengabaikan data yang berada di luar kelompoknya.



Dengan **K-Means**, seluruh data selalu dimasukkan dalam klaster terdekat, walaupun terpisah.

# K-Means vs Density-based Clustering

- **K-Means** akan **memasukkan setiap data** ke dalam suatu klaster, walaupun data tersebut terpisah cukup jauh dari pusat data (centroid).
- **Density-based** akan mengelompokkan data yang memiliki **densitas (kerapatan) tinggi** dan mengabaikan data yang berada di luar kelompoknya.



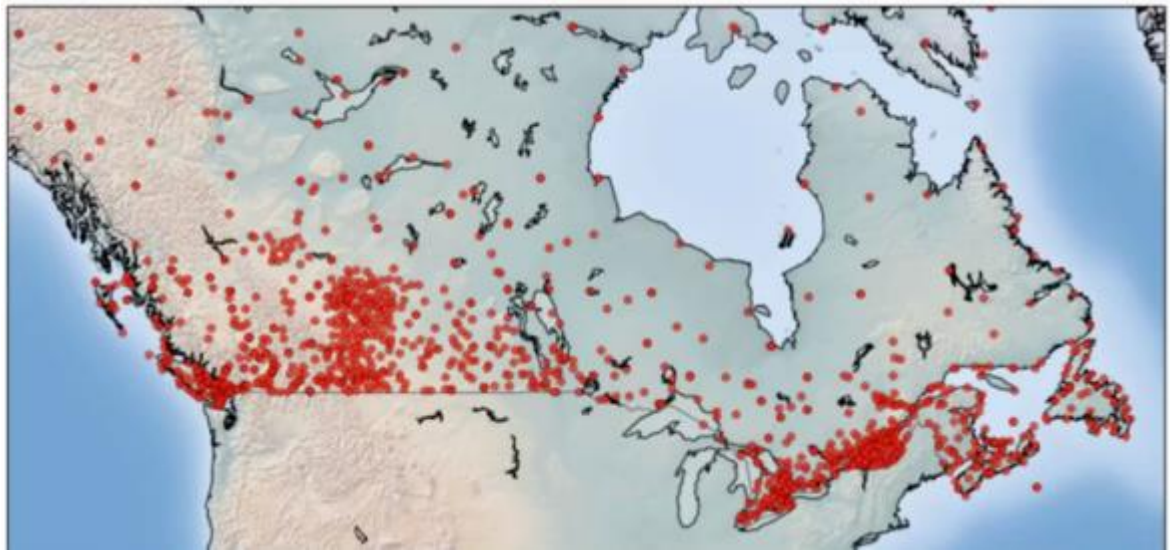
Dengan **Density-based** Clustering, data yang terpisah dari kelompoknya akan menjadi data pencilan (outlier)


# Density-based Clustering

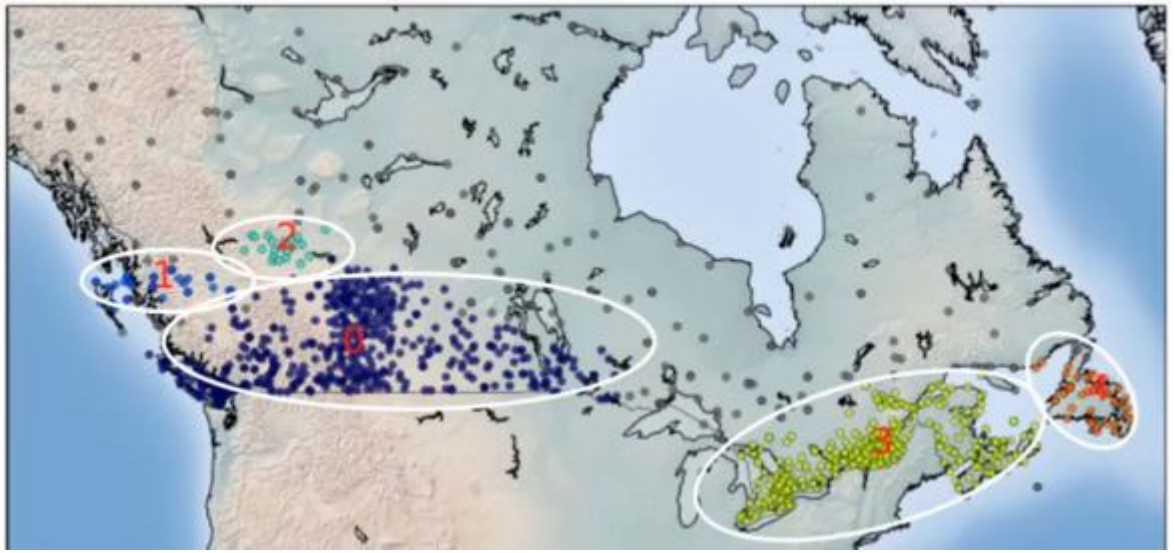
- **Density** dapat didefinisikan sebagai **jumlah poin data** yang berada pada radius tertentu.
- Salah satu metode yang menggunakan pendekatan density untuk melakukan klasterisasi adalah **DBSCAN**.
- **DBSCAN** = **D**ensity **B**ased **S**patial **C**lustering of **A**pplications with **N**oise.

# Density-based Clustering: DBSCAN

Lokasi stasiun cuaca di Canada

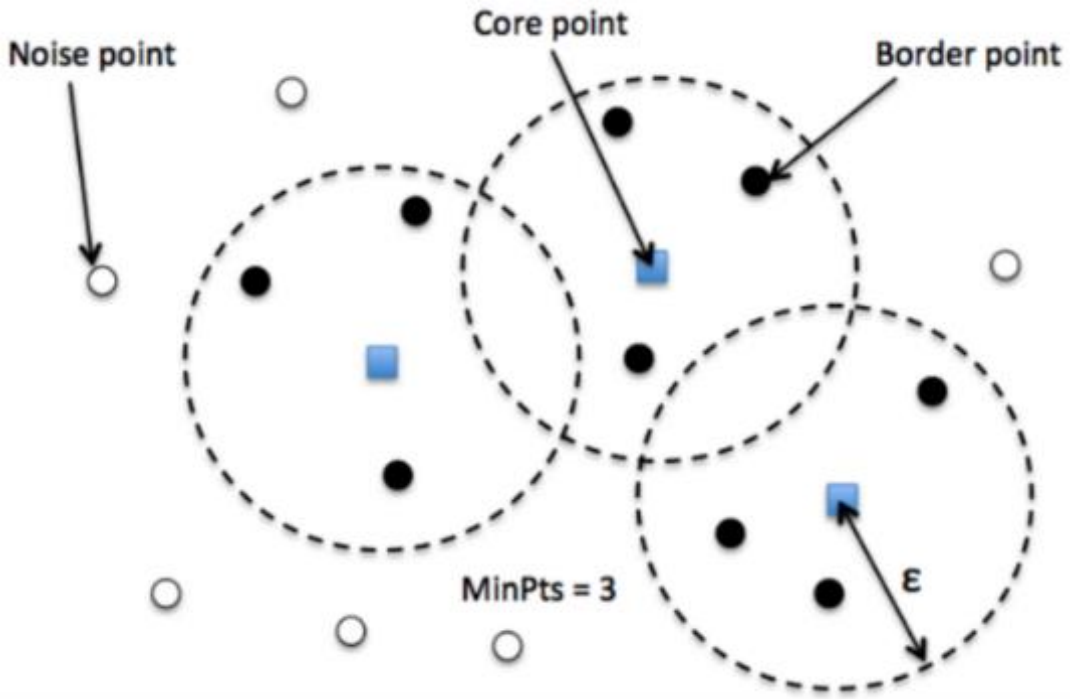


Hasil klasterisasi dengan DBSCAN



# Density-based Clustering: DBSCAN

- DBSCAN:
  - Salah satu algoritma klasterisasi berbasis densitas data yang paling popular.
- DBSCAN melakukan klasterisasi berdasarkan 2 (dua) parameter:
  - Radius of neighbourhood (**R**), yaitu jarak antara titik pusat data dengan titik terjauh dalam sebuah area (klaster).
  - Min number of neighbour (**M**), yaitu jumlah minimal data di dalam suatu area (klaster)

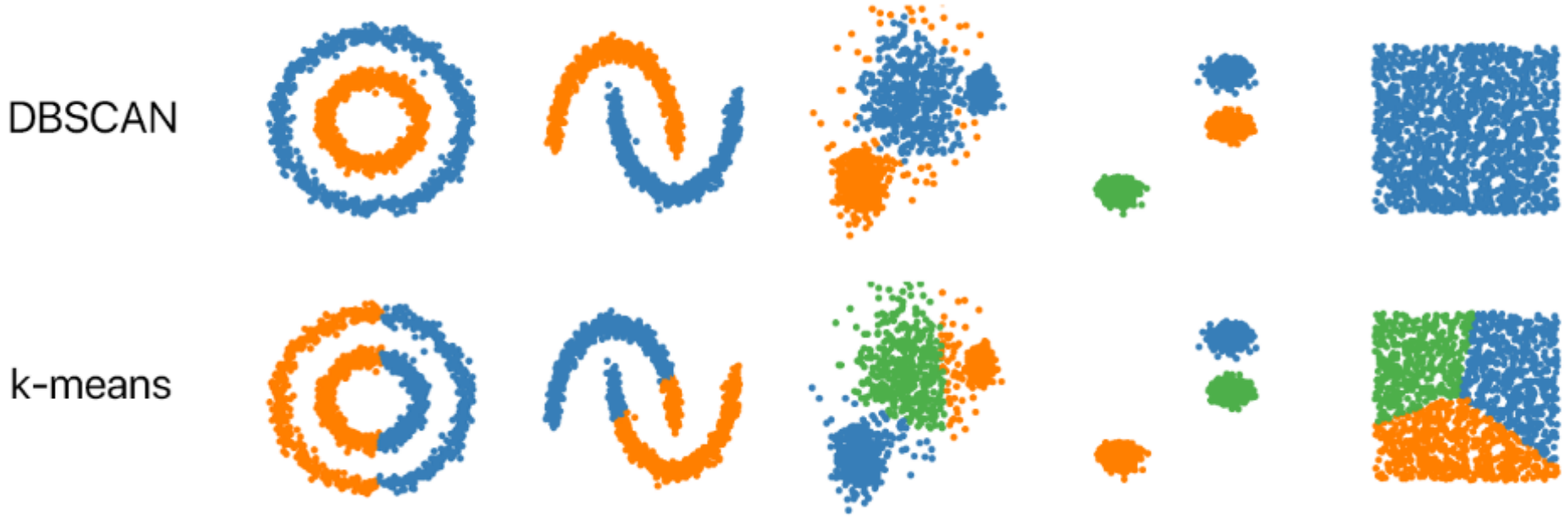


# Algoritma DBSCAN

1. Tentukan nilai minimum poin (M) dan radius (R) yang akan digunakan.
2. Pilih data awal “p” secara acak.
3. Hitung jarak antara data “p” terhadap semua data menggunakan metode perhitungan jarak (contoh: Euclidian distance).
4. Ambil semua amatan yang density-reachable dengan amatan “p”.
5. Jika amatan yang memenuhi nilai R lebih dari jumlah minimal amatan dalam satu kelompok maka amatan “p” dikategorikan sebagai core points dan kelompok terbentuk.
6. Jika amatan “p” adalah border points dan tidak ada amatan yang density-reachable dengan amatan “p”, maka lanjutkan pada amatan lainnya.
7. Ulangi langkah 3 sampai 6 hingga semua amatan diproses.

# Kapan DBSCAN digunakan?

- Distribusi data non-linear
- Klasterisasi data spasial umumnya cocok dengan DBSCAN
- Data mengandung banyak outlier



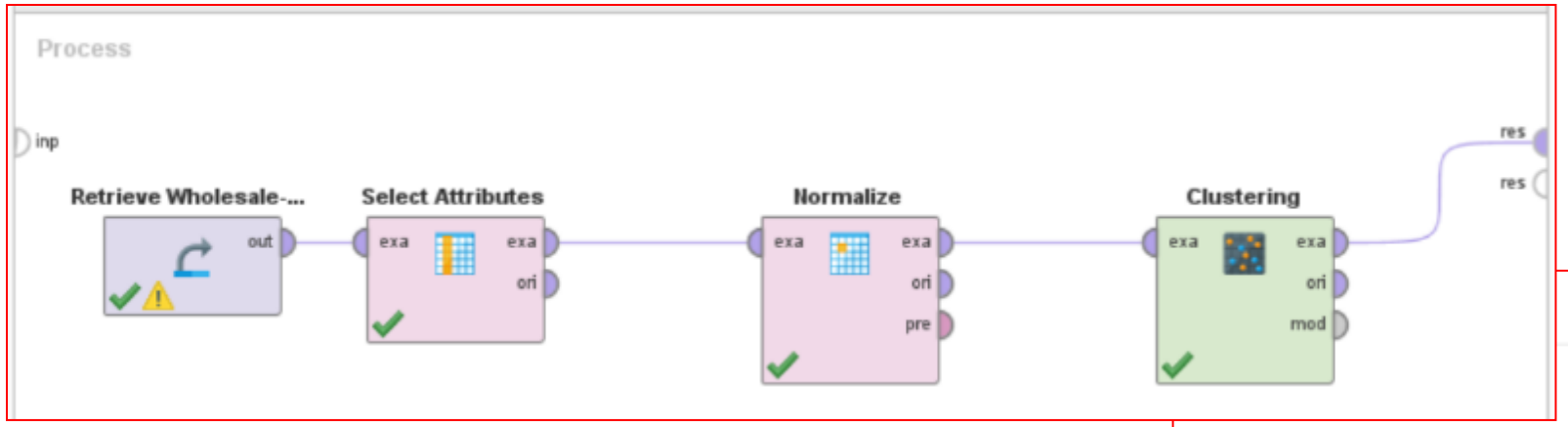
# Latihan: DBSCAN

- Dataset: **Wholesale-customers-data.csv**
- Data jumlah penjualan kebutuhan sehari-hari dari sebuah toko waralaba yang disajikan per wilayah

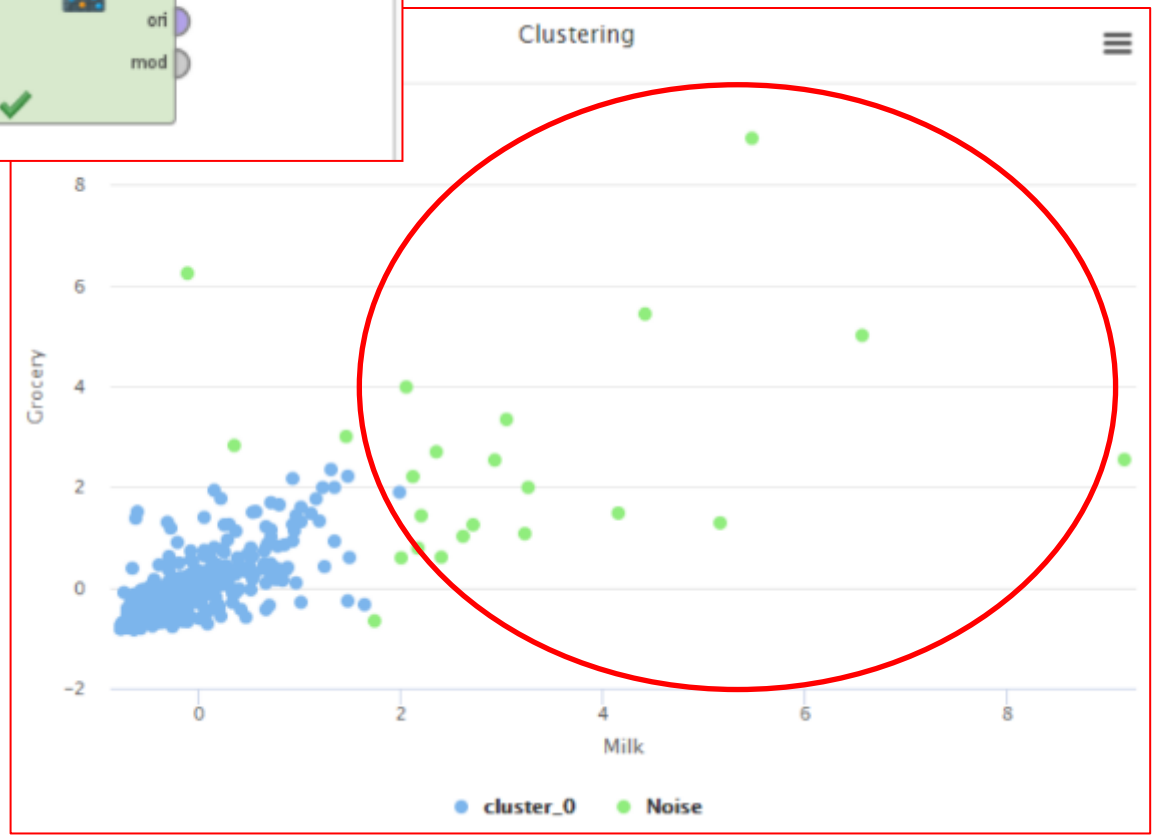
| | Channel | Region | Fresh | Milk | Grocery | Frozen | Detergents_Paper | Delicassen |
|---|---|---|---|---|---|---|---|---|
| 0 | 2 | 3 | 12669 | 9656 | 7561 | 214 | 2674 | 1338 |
| 1 | 2 | 3 | 7057 | 9810 | 9568 | 1762 | 3293 | 1776 |
| 2 | 2 | 3 | 6353 | 8808 | 7684 | 2405 | 3516 | 7844 |
| 3 | 1 | 3 | 13265 | 1196 | 4221 | 6404 | 507 | 1788 |
| 4 | 2 | 3 | 22615 | 5410 | 7198 | 3915 | 1777 | 5185 |

Lakukan klasterisasi dengan **DBSCAN** berdasarkan jumlah penjualan produk "**Milk**" dan "**Grocery**" untuk mengidentifikasi **data outlier**.

# Latihan: DBSCAN




Metode **DBSCAN** dapat mendeteksi keberadaan data **outlier**.




Bagian 4

# CONTOH: SEGMENTASI PELANGGAN

# Contoh Aplikasi Segmentasi Pelanggan Dengan K-Means

## Customer Segmentation Based on Recency Frequency Monetary (RFM) and User Event Tracking (UET) Using K-Means Algorithm

Achmad Solichin
*Computer Science, Faculty of Information Technology*
*Universitas Budi Luhur*
Jakarta, Indonesia
achmad.solichin@budiluhur.ac.id

Gunadi Wibowo
*Computer Science, Faculty of Information Technology*
*Universitas Budi Luhur*
Jakarta, Indonesia
2011600265@student.budiluhur.ac.id

***Abstract*— Business competition in e-commerce today is very tight, so every company is competing to increase sales. One of them is by providing the best service for its customers, so customer loyalty is formed to continue transacting again. Therefore, companies are required to be able to recognize and establish closeness with their customers. Customer segmentation is an effort to group customers based on customer characteristics and behavior in conducting transactions. With customer segmentation, companies can provide more targeted treatment and services. However, producing the proper customer segmentation requires a complicated data analysis process. In this study, customer segmentation was carried out using the K-Means clustering method based on the Recency, Frequency, and monetary (RFM) model combined with the User Event Tracking (UET) parameter. Based on the results of tests that have been carried out on 1,447,984 transaction data and 932,021 user tracking data, the resulting customer segmentation is divided into 3 (three) groups, namely Platinum (43.9%), Gold (9.5%), and Silver (46.6%). Companies can run different marketing strategies for each of these customer groups.**

identifying the level of customer profit (customer value) to the company is carried out using all data owned by the company. Customer value analysis is an analytical method to explore and find the character of all customers, which is then used for further analysis of specific customers for proper new knowledge from huge data [3].

Customer segmentation is done by dividing all customers into several groups according to the similarity of customer shopping behavior. If the number of customers and transactions is getting bigger, then this segmentation process becomes complicated when done manually. Therefore, the application of data mining methods will be very helpful for analyzing data to obtain patterns which will then be used as new knowledge [1], [4], [5]. Data Mining can more easily recognize certain patterns in all processed data, resulting in customer segmentation that is in accordance with each customer's behavior. This will assist the company in making business decisions that are more suited to the spending potential of each group [6].

Makalah ini merupakan hasil penelitian tesis, telah diterima dan dipresentasikan pada Seminar Internasional ITIS 2022 (The 8th Information Technology International Seminar) yang diselenggarakan di UPN Veteran Jawa Timur.


# Parameter Segmentasi Pelanggan

| Parameter | Keterangan | Sumber Data |
|---|---|---|
| Recency (R) | Kapan terakhir pelanggan bertransaksi | Transaksi Penjualan |
| Frequency (F) | Seberapa sering pelanggan bertransaksi | Transaksi Penjualan |
| Monetary (M) | Seberapa banyak pelanggan berbelanja (Rp) | Transaksi Penjualan |
| Recency (R) | Kapan terakhir pelanggan meng-klik banner promosi | User Event Tracking |
| Frequency (F) | Seberapa sering pelanggan meng-klik banner promosi | User Event Tracking |

# Dataset

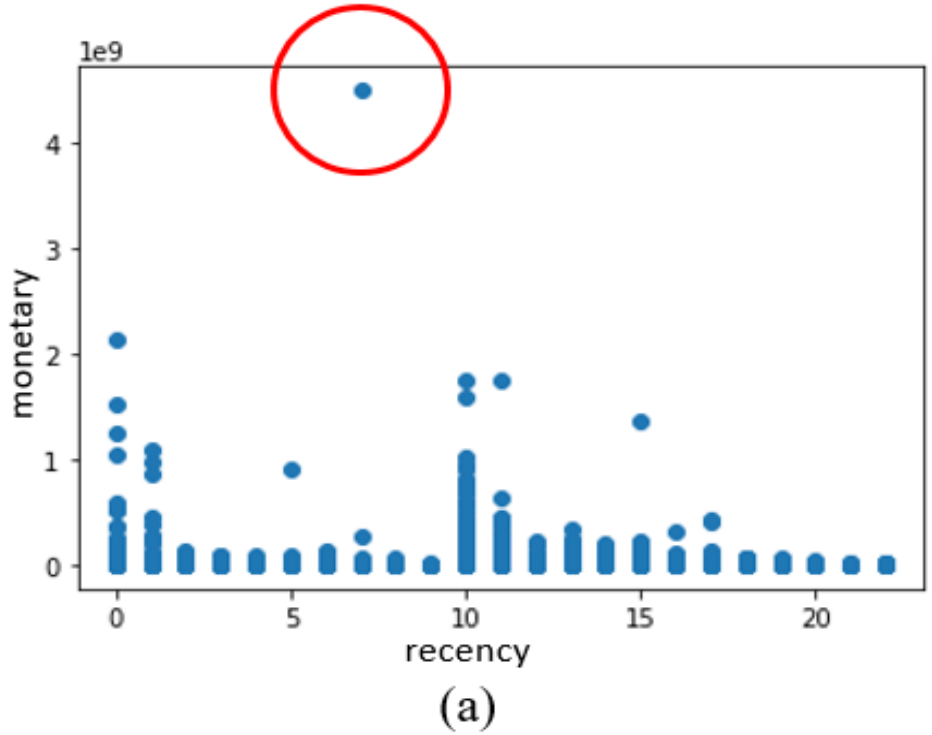
TABLE I. THE READY SNIPPET OF THE DATASET

| UUID | trx_recency | trx_freq | trx_monetery | banner_freq | banner_recency |
|---|---|---|---|---|---|
| 916f7b | 0 | 74 | 23716400 | 0 | 0 |
| 3c511d | 0 | 4 | 817300 | 2 | 2 |
| 0d7c04 | 0 | 3 | 489100 | 22 | 2 |
| a90bd6 | 0 | 49 | 18621311 | 2 | 2 |
| c9df2c | 0 | 21 | 5612200 | 0 | 0 |
| ... | ... | ... | ... | ... | ... |
| 699d17 | 22 | 1 | 62200 | 0 | 0 |

**Dataset :**

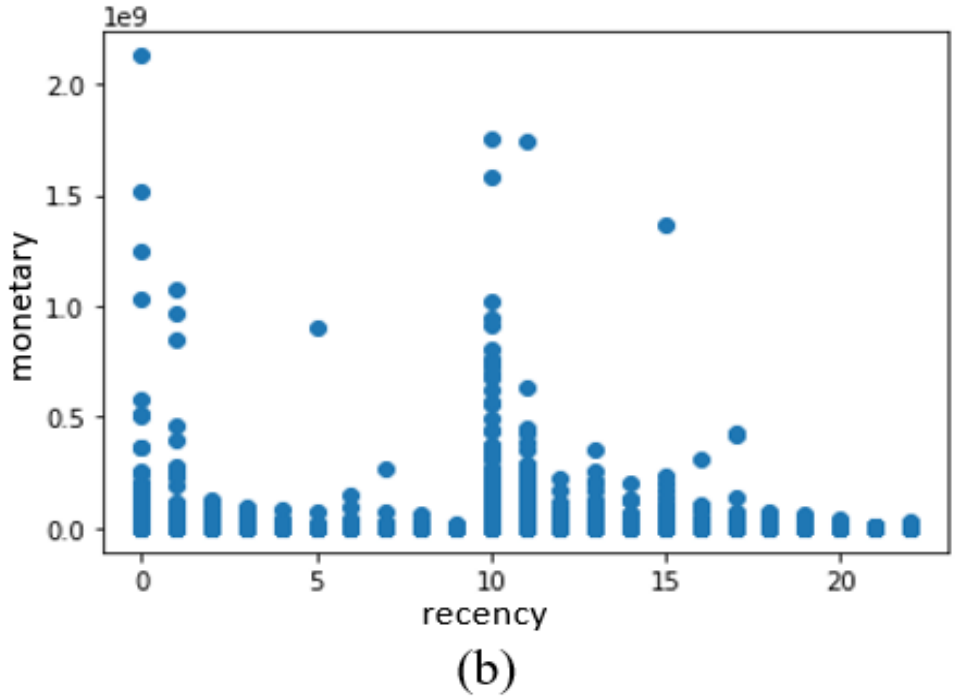
- Data selama 2 tahun (2020 – 2021)
- Jenis produk ritel
- Hanya melibatkan transaksi yang sukses
- Total data = 1,4+ juta data dari transaksi dan 900+ ribu data log UET

# Preprocessing Data: Menghilangkan Outlier




(a)

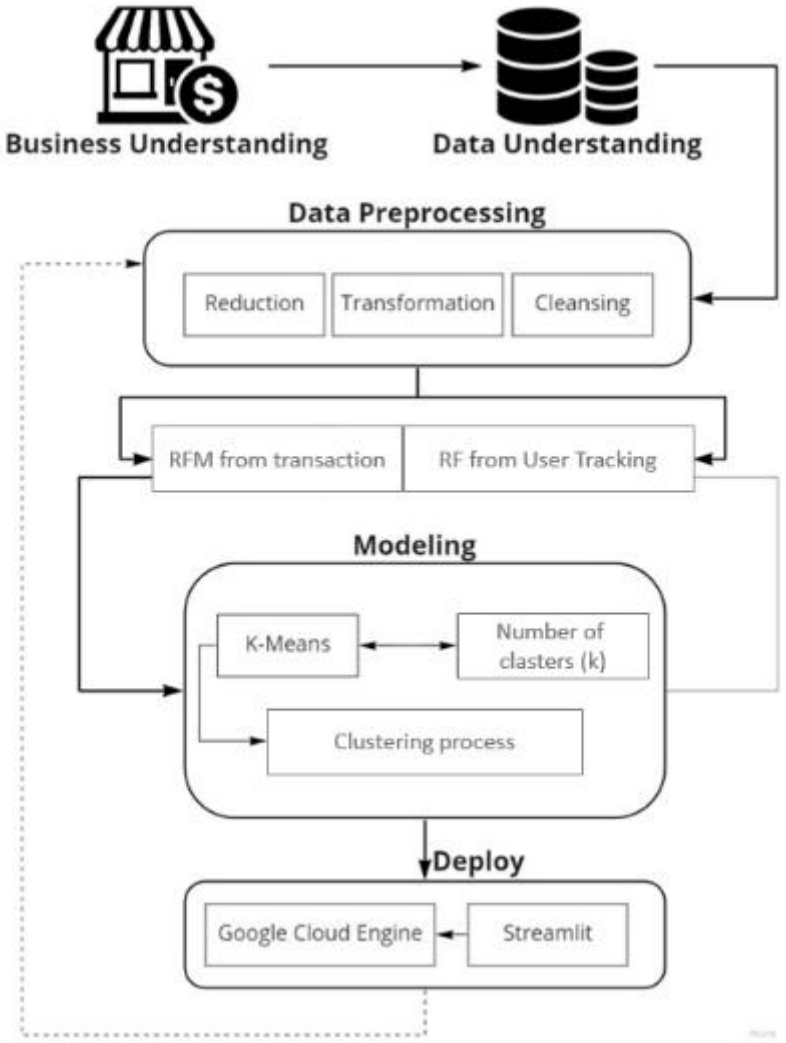
Sebelum



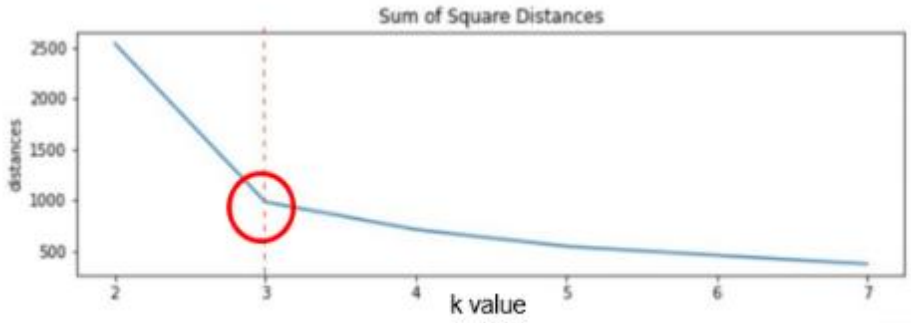

(b)

Sesudah

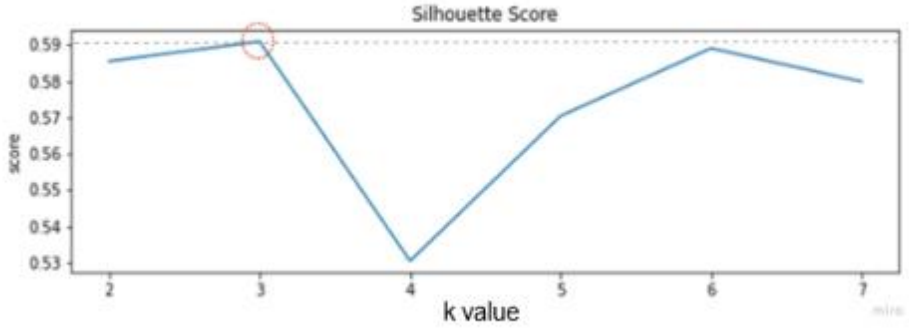
# Metode Klasterisasi




- Metode K-Means Clustering
- Jumlah klaster (k) = 3




Metode elbow



# Hasil Klasterisasi

## Distribusi data

TABLE IV. CLUSTERING RESULTS

| Cluster | Total | Percentage |
|---|---|---|
| 0 | 3807 | 9.59 % |
| 1 | 16527 | 41.63 % |
| 2 | 19362 | 48.78 % |

## Rekomendasi strategi pemasaran

TABLE VIII. CUSTOMER CHARACTERISTICS AND MARKETING STRATEGY RECOMMENDATIONS

| Cluster | Customer type | Customer characters | Marketing Strategies |
|---|---|---|---|
| Silver | First-time customer | Customers who only shop once | Product introduction email, the bundling promo offer |
| Gold | Client | The customers who buy regularly, the relationship with the product is strong, and it is not easy to switch to other company's products | Up-selling, offering products that have a higher value |
| Platinum | Advocate | Loyal customers who will not hesitate to recommend products to those closest to them | Referral Program |

# Referensi

1. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques Third Edition, *Elsevier*, 2012
2. Ian H. Witten, Frank Eibe, Mark A. Hall, Data mining: Practical Machine Learning Tools and Techniques 3rd Edition, *Elsevier*, 2011
3. Markus Hofmann and Ralf Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, *CRC Press Taylor & Francis Group*, 2014
4. Daniel T. Larose, Discovering Knowledge in Data: an Introduction to Data Mining, *John Wiley & Sons*, 2005
5. Ethem Alpaydin, Introduction to Machine Learning, 3rd ed., *MIT Press*, 2014
6. Materi "Thematic Academy: AI dan DS untuk Dosen dan Instruktur", 2021.
7. Achmad Solichin, Channel Youtube, https://youtube.com/@AchmadSolichin

# SELESAI