



# DIGITAL TALENT SCHOLARSHIP 2019



Program Fresh Graduate Academy Digital Talent Scholarship 2019 | Machine Learning

# Clustering: Pengantar

M. Ramli & M. Soleh



# Clustering untuk Segmentasi

- **Contoh kasusnya**, kita harus **mengelompokkan pelanggan berdasarkan karakteristik pelanggan**. Dalam kasus ini kita dapat mengelompokkannya kedalam 2 kelas, yaitu 1 dan 0.
- Sehingga perusahaan dapat secara efektif menerapkan strategi bisnis secara spesifik kepada pelanggan atau mengalokasikan dengan optimal sumber daya untuk pemasaran.

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

# Clustering untuk Segmentasi

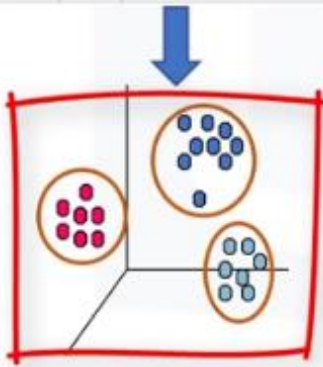
- Customer Group dapat dihasilkan berdasarkan atribut atau sering disebut sebagai fitur data
- **Proses segmentasi** adalah mencoba untuk menemukan kesamaan dari setiap pelanggan berdasarkan fitur-fitur yang ada, yaitu *Age, Edu, Years Employed, Income, Card Debt, Other Debt, Address, dan Debt Income Ratio*.

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

# Clustering untuk Segmentasi

- Salah satu metode segmentasi clustering
- **Clustering** bekerja dengan **metode tanpa pengawasan (unsupervised learning)** berdasarkan kesamaan pelanggan.

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1



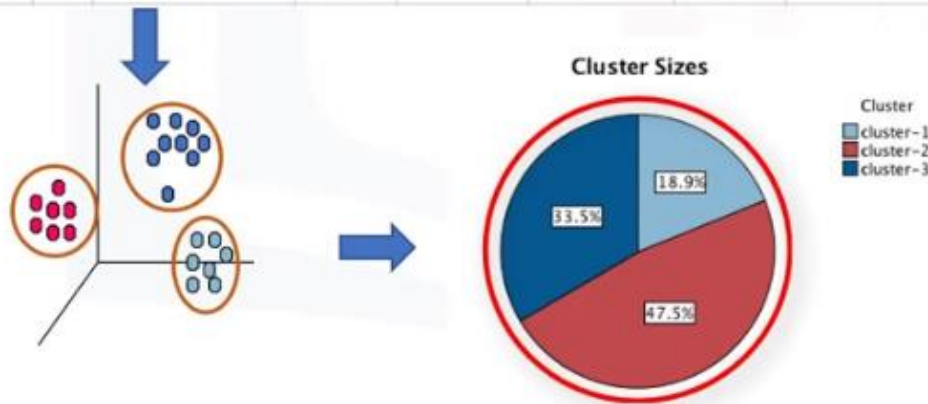




# Clustering untuk Segmentasi

- Misalnya pelanggan dikelompokkan menjadi 3 kelompok
- Setiap kelompok tersebut memiliki demografi (berdasarkan fitur) yang serupa

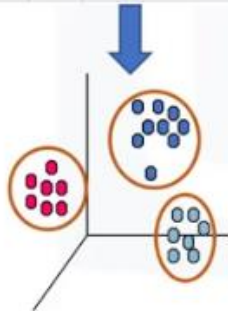
Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1



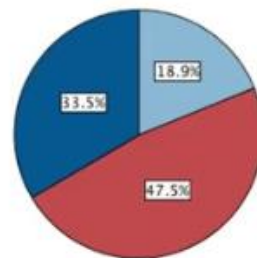
# Clustering untuk Segmentasi

- Dari hasil pengelompokan, kita dapat membuat profil untuk setiap grup

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1



Cluster Sizes



Cluster  
 cluster-1  
 cluster-2  
 cluster-3

Cluster	Segment Name
cluster-1	AFFLUENT AND MIDDLE AGED
cluster-2	YOUNG EDUCATED AND MIDDLE INCOME
cluster-3	YOUNG AND LOW INCOME

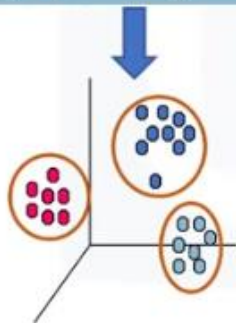
Profil

# Clustering untuk Segmentasi

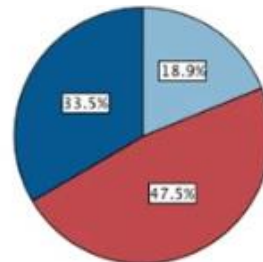
- Akhirnya kita dapat menetapkan data secara individual ke salah satu grup

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFLUENT AND MIDDLE AGED
3	AFFLUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFLUENT AND MIDDLE AGED
6	AFFLUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFLUENT AND MIDDLE AGED



## Keputusan Segmentasi Setiap Customer



■ cluster-1  
■ cluster-2  
■ cluster-3

Cluster	Segment Name
cluster-1	AFFLUENT AND MIDDLE AGED
cluster-2	YOUNG EDUCATED AND MIDDLE INCOME
cluster-3	YOUNG AND LOW INCOME





# Clustering untuk Segmentasi

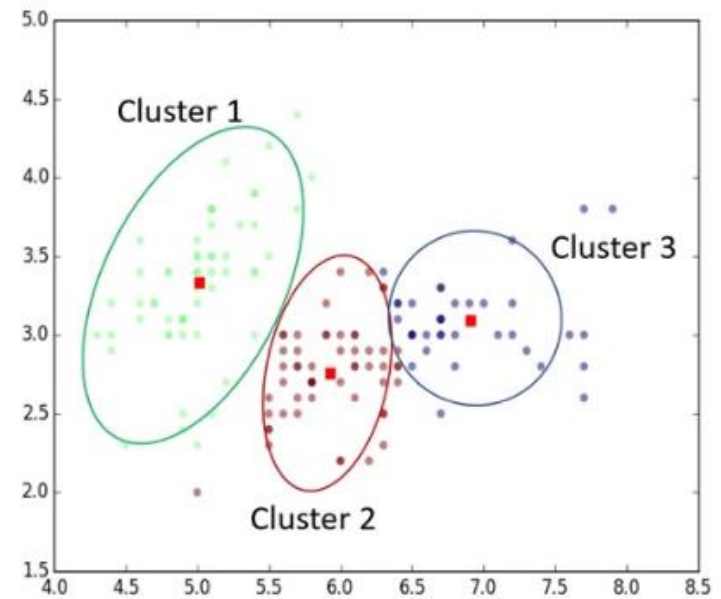
- Jadi, dari hasil clustering yang kita dapatkan adalah:
  - Preferensi pelanggan secara individu
  - Perilaku pembelian pada berbagai produk
- Kita dapat mengembangkan personal experience untuk masing-masing segmen

# Apa itu clustering?

- **Clustering** adalah menemukan cluster pada dataset tanpa pengawasan (unsupervised)

## Apa itu cluster?

Sebuah grup objek yang memiliki kesamaan (similar) diantara objek didalam cluster objek, dan memiliki ketidaksamaan (dissimilar) dengan objek di cluster lainnya.

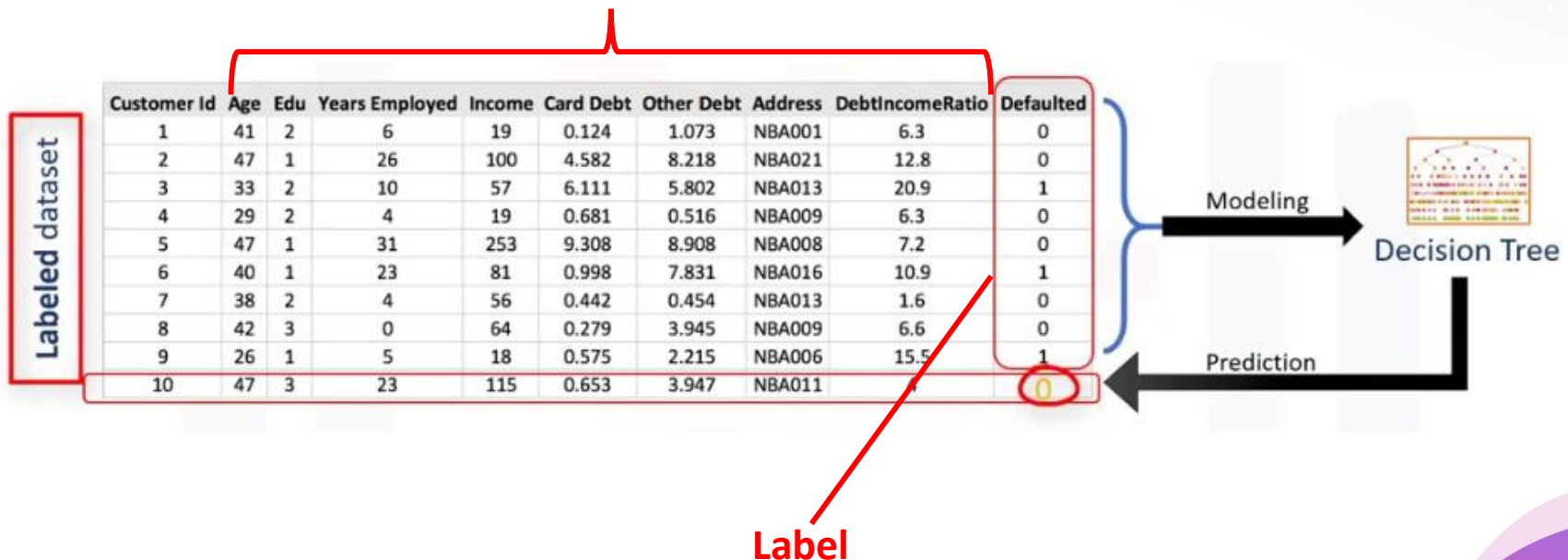


# Clustering Vs. Classification

## • Classification

- Dibimbing/diawasi menggunakan set data berlabel saat dilakukan training (proses pembelajaran)
- Training (proses pembelajaran/pembentukan model) menggunakan attributes dan label.

**Attributes**



# Clustering Vs. Classification

- **Clustering**

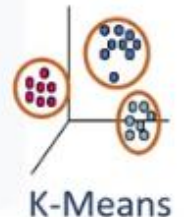
- Proses pemodelan tidak diawasi dengan menggunakan label dataset
- Label hanya digunakan untuk validasi model

Unlabeled dataset

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Modeling

Segmentation





# Clustering Vs. Classification

Labeled dataset

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1
10	47	3	23	115	0.653	3.947	NBA011	4	0

Modeling



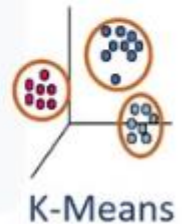
Prediction

Unlabeled dataset

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Modeling

Segmentation





# Penggunaan Clustering

- **RETAIL MARKETING**

- Mengidentifikasi pola pembelian pelanggan
- Merekomendasikan buku atau film baru kepada pelanggan baru

- **BANKING**

- Deteksi penipuan dalam penggunaan kartu kredit
- Mengidentifikasi kelompok pelanggan (misalnya: Loyal/Tidak Loyal)

- **INSURANCE**

- Fraud detection (Deteksi penipuan) dalam analisis klaim asuransi
- Risiko asuransi pelanggan



# Penggunaan Clustering

- **PUBLICATION**

- Mengelompokkan berita secara otomatis berdasarkan kontennya
- Merekomendasikan artikel berita serupa

- **MEDICINE**

- Mengkarakterisasi perilaku pasien

- **BIOLOGY**

- Mengelompokkan penanda genetik (genetic markers) untuk mengidentifikasi ikatan keluarga



# Mengapa menggunakan clustering?

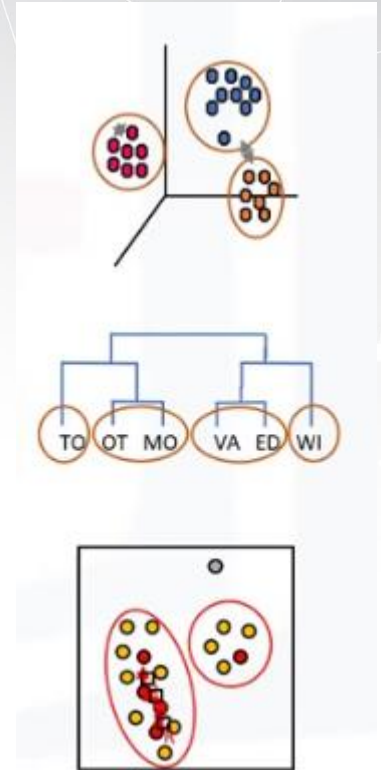
- Merupakan exploratory data analysis
- Dapat melakukan summary generation
- Dapat melakukan outlier detection
- Dapat menemukan duplikasi data
- Terdapat step pra-pemrosesan





# Algoritma Clustering

- **Partitioned-based Clustering**
  - Relatively efficient
  - E.g., k-Means, k-Median, Fuzzy c-Means
- **Hierarchical Clustering**
  - Produces trees of clusters
  - E.g. Agglomerative, Divisive
- **Density-based Clustering**
  - Produces arbitrary shaped clusters
  - E.g. DBSCAN



Program Fresh Graduate Academy Digital Talent Scholarship 2019 | Machine Learning

# Clustering: K-Means

M. Ramli & M.Soleh

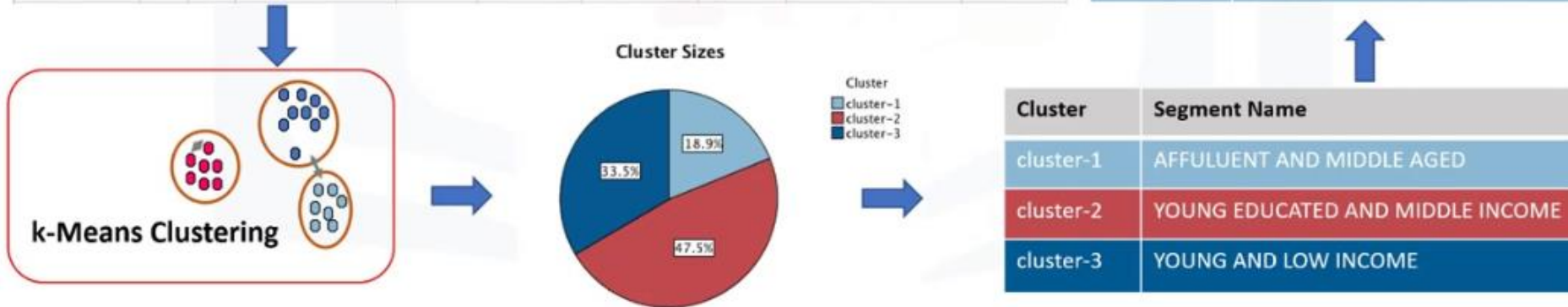


# Apa itu k-Means clustering?

- Clustering bekerja pada data yang tidak diawasi berdasarkan kesamaan setiap dataset

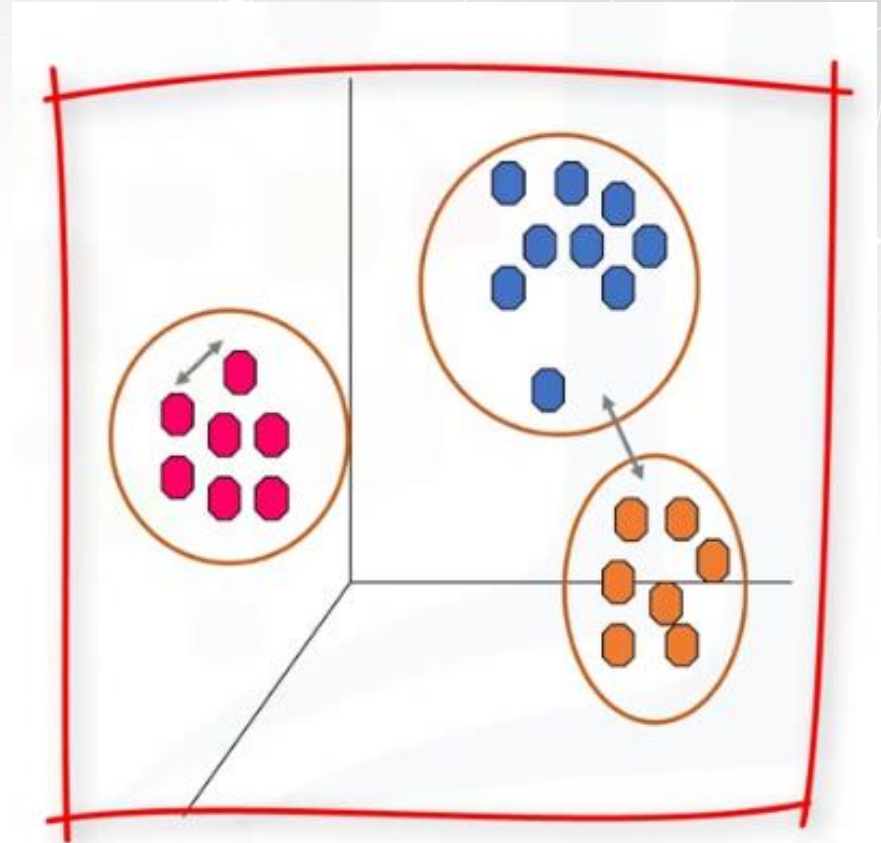
Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFLUENT AND MIDDLE AGED
3	AFFLUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFLUENT AND MIDDLE AGED
6	AFFLUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFLUENT AND MIDDLE AGED



# Algoritma K-Means

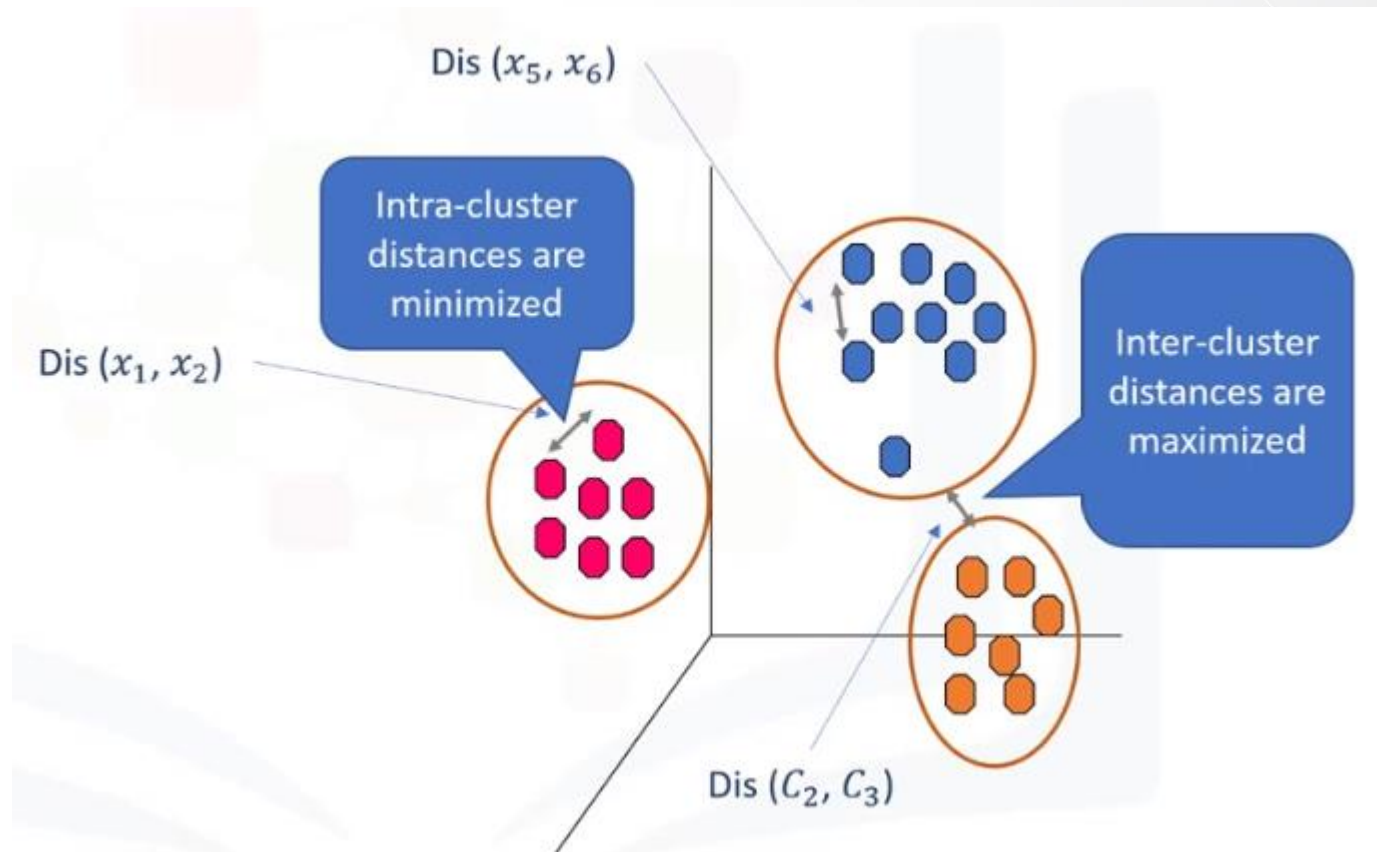
- K-Means termasuk dalam Partitioning Clustering
- K-Means membagi data menjadi subset (cluster) yang tidak tumpang tindih
- Data dalam sebuah cluster sangat mirip
- Data antar kelompok sangat berbeda





# Menentukan similarity atau dissimilarity

- Similarity digunakan untuk dataset dalam internal satu cluster
- Dissimilarity digunakan untuk dataset antar cluster





# 1-dimentional similarity/distance

- 1-dimentional similarity dapat digunakan untuk mengukur jarak dua titik menggunakan 1 nilai.
- Rumus *Euclidean Distance* dapat digunakan untuk mengukur similarity



**Customer 1**

Age

54



**Customer 2**

Age

50

$$\text{Dis}(x_1, x_2) = \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2}$$

$$\text{Dis}(x_1, x_2) = \sqrt{(34 - 30)^2} = 4$$

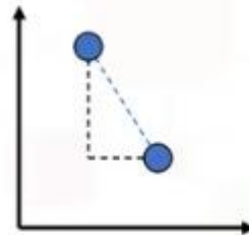
# 2-dimentional similarity/distance

- 2-dimentional similarity dapat digunakan untuk mengukur jarak dua titik menggunakan 2 nilai atau 2D *matrix space*.



Customer 1

Age	Income
54	190



Customer 2

Age	Income
50	200

$$\begin{aligned} \text{Dis}(x_1, x_2) &= \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2} \\ &= \sqrt{(54 - 50)^2 + (190 - 200)^2} = 10.77 \end{aligned}$$

# Multi-dimentional similarity/distance

- Dengan demikian metode *Euclidean Distance* dapat digunakan untuk multidimensi dengan menambahkan titik pada rumus.



**Customer 1**

Age	Income	education
54	190	3



**Customer 2**

Age	Income	education
50	200	8

$$\text{Dis } (x_1, x_2) = \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2}$$

$$= \sqrt{(54 - 50)^2 + (190 - 200)^2 + (3 - 8)^2} = 11.87$$



# Bagaimana k-Means Clustering bekerja?

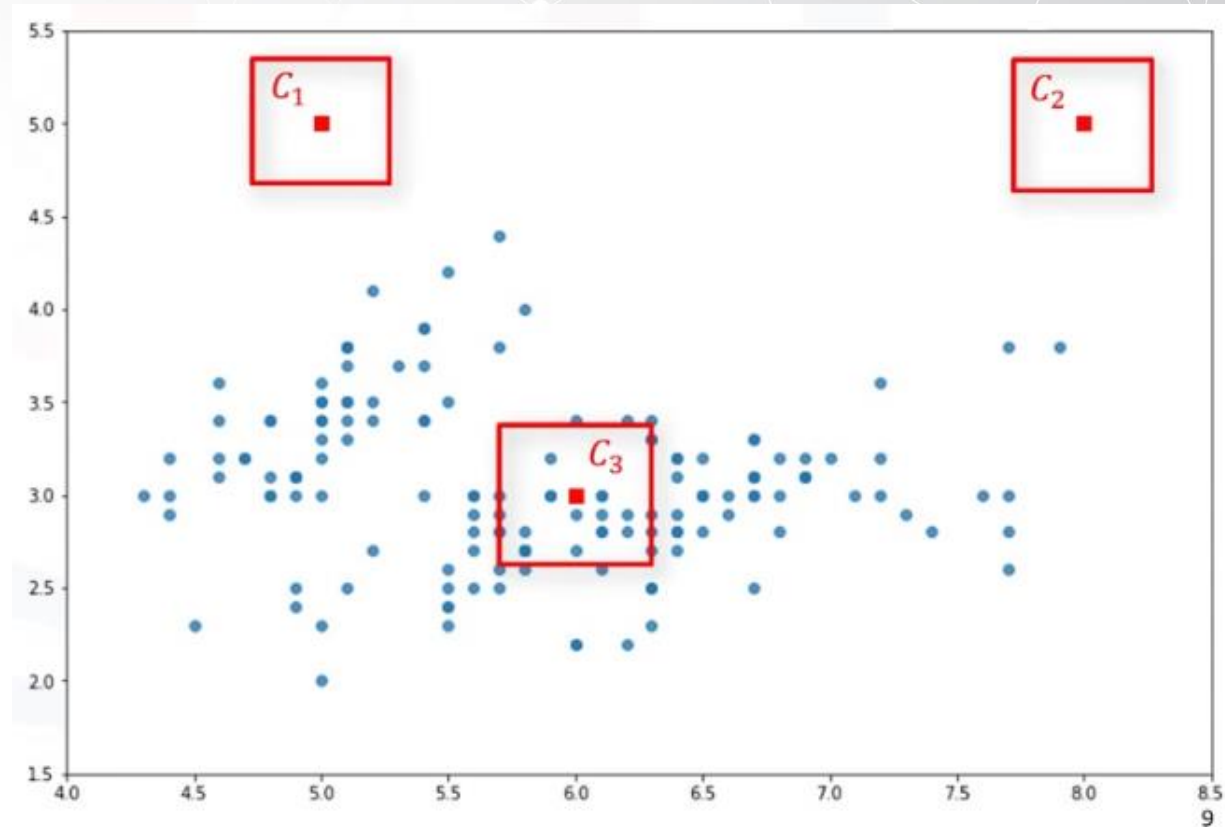
- Misalkan terdapat data yang memiliki attributes **age** dan **income**, dan tersebar dalam *matrix 2D space* (dapat digambarkan dalam diagram Cartesian)

Customer ID	Age	Income
1	3	4
2	2	6
3	3.5	2
...	...	..



# k-Means clustering – inisialisasi k

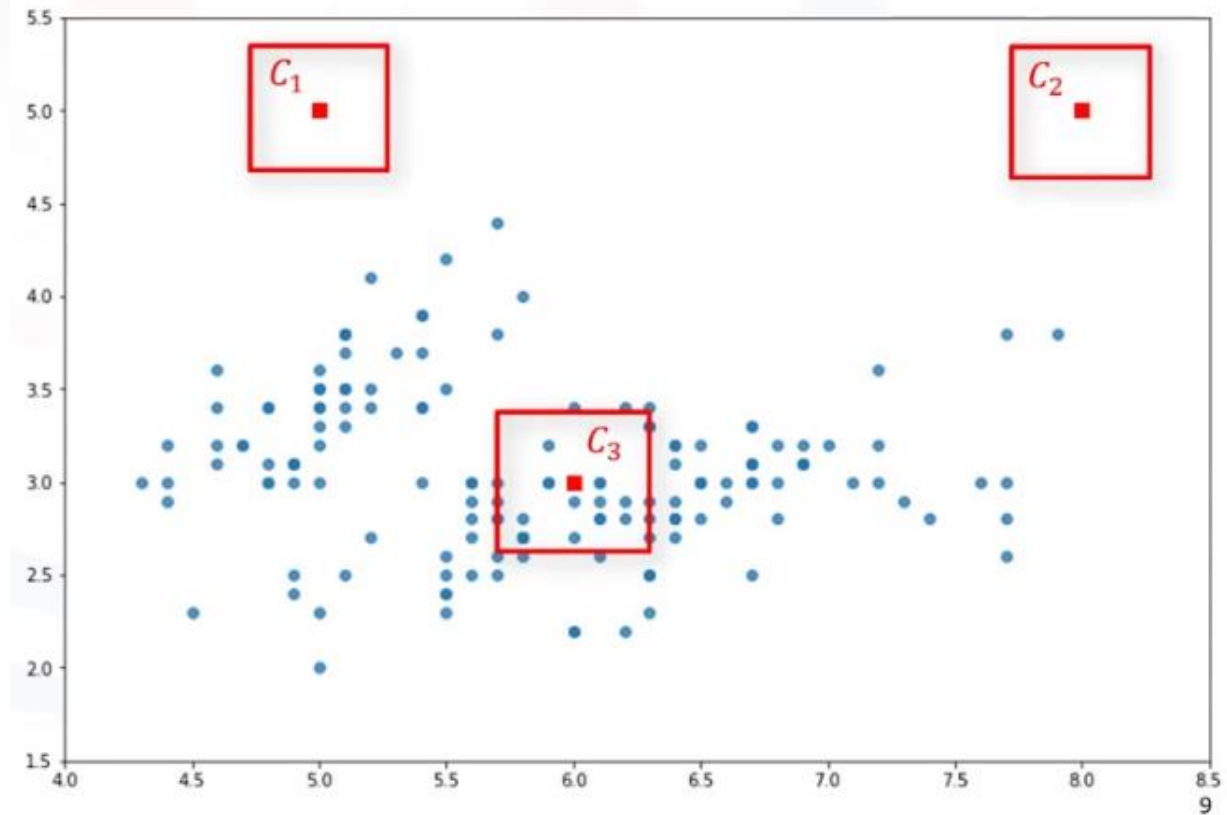
1. Misal kita inisialisasi  $k=3$ ,  $k$  sebagai centroids yang dipilih secara random



# k-Means clustering – inisialisasi k

1. Misal kita inisialisasi  $k=3$ ,  $k$  sebagai centroids yang dipilih secara random

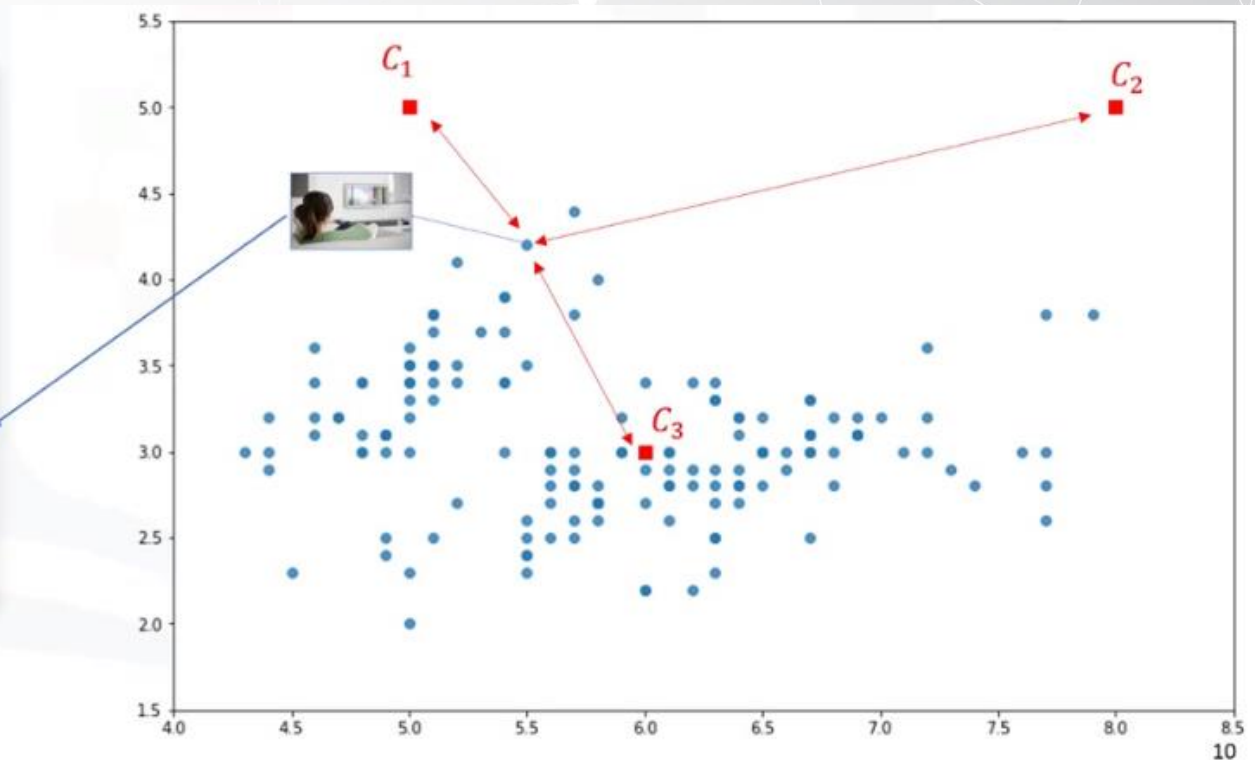
$C_1 = [5., 5.]$   
 $C_2 = [8., 5.]$   
 $C_3 = [6., 3.]$



# k-Means clustering – hitung jarak (distance)

2. Hitung jarak setiap titik dataset dengan 3 centroid yang telah ditentukan secara random

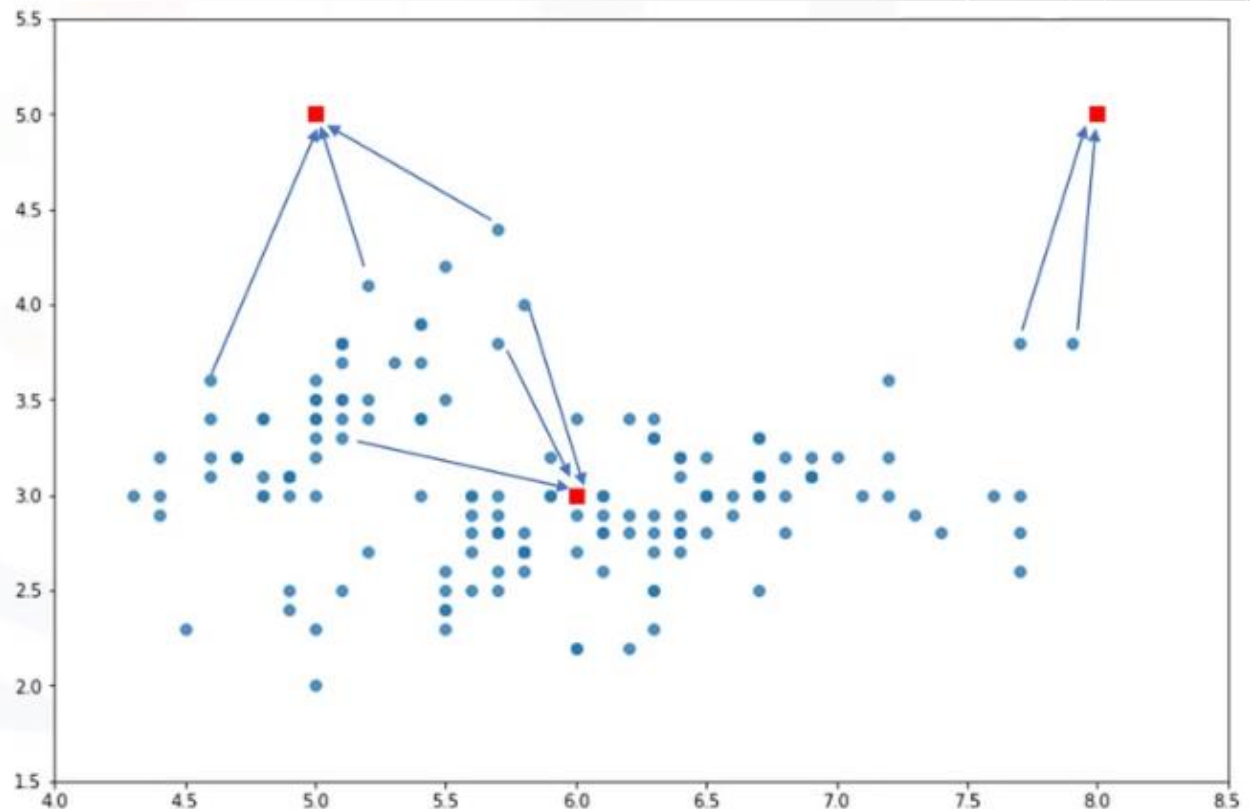
$C_1$	$C_2$	$C_3$
$d(p1, c1)$	$d(p1, c2)$	$d(p1, c3)$
$d(p2, c1)$	$d(p2, c2)$	$d(p2, c3)$
$d(p3, c1)$	$d(p3, c2)$	$d(p3, c3)$
$d(p4, c1)$	$d(p4, c2)$	$d(p4, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$



# k-Means clustering – tetapkan ke centroid

## 3. Tetapkan setiap titik dataset ke centroid terdekat

$C_1$	$C_2$	$C_3$
$d(p1, c1)$	$d(p1, c2)$	$d(p1, c3)$
$d(p2, c1)$	$d(p2, c2)$	$d(p2, c3)$
$d(p3, c1)$	$d(p3, c2)$	$d(p3, c3)$
$d(p4, c1)$	$d(p4, c2)$	$d(p4, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$

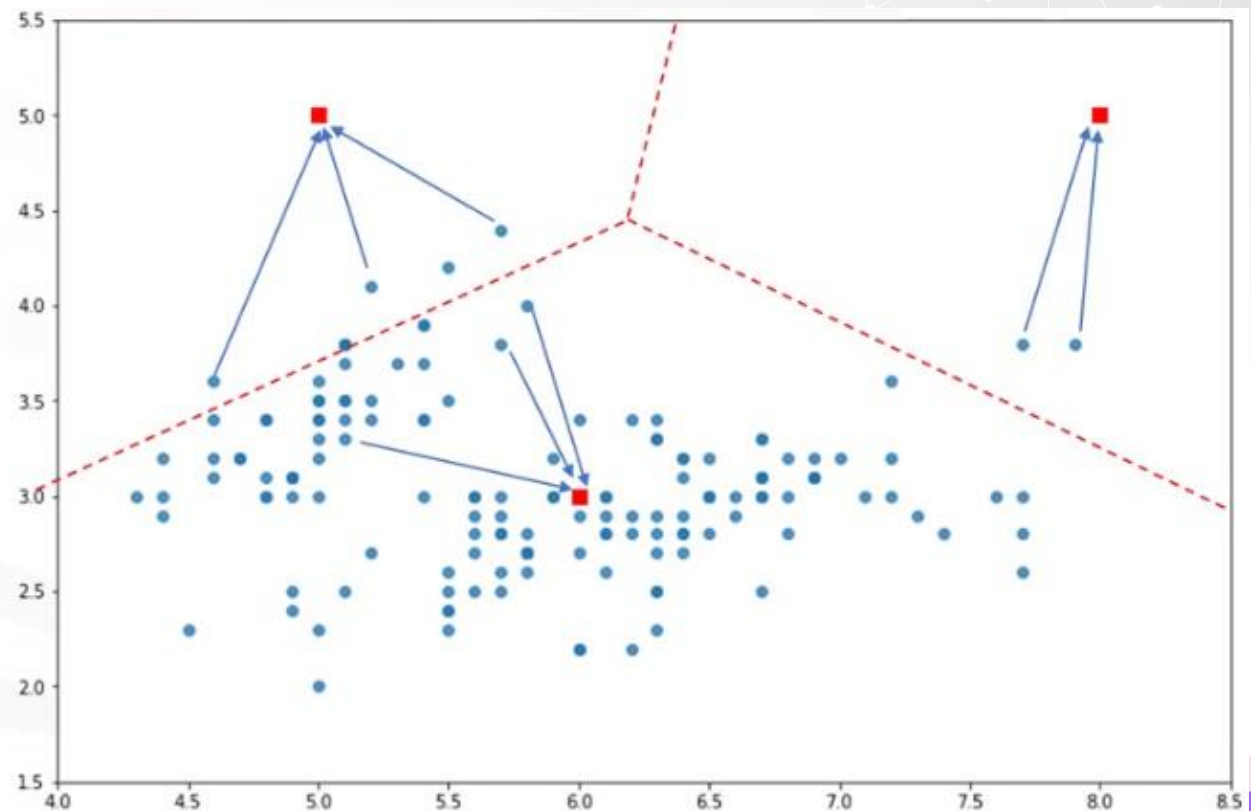




# k-Means clustering – tetapkan titik ke satu centroid

3. Tetapkan setiap titik dataset ke centroid terdekat, sehingga terbentuk voronoi diagram yang menunjukkan pembatas antar cluster.

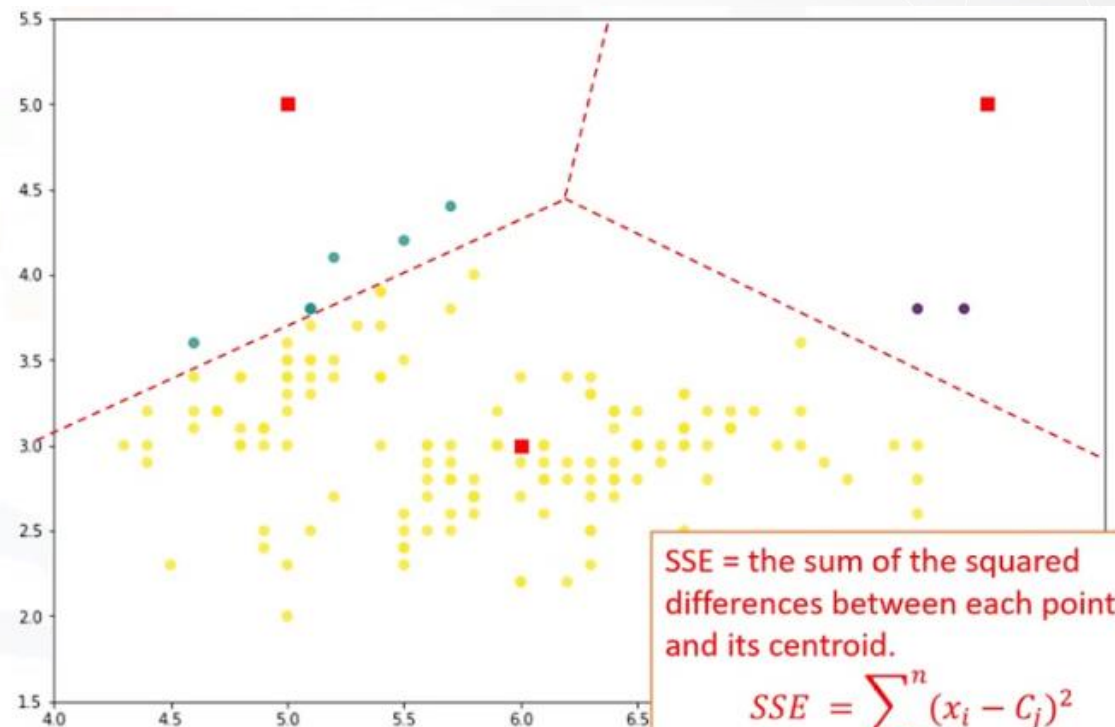
$C_1$	$C_2$	$C_3$
$d(p1, c1)$	$d(p1, c2)$	$d(p1, c3)$
$d(p2, c1)$	$d(p2, c2)$	$d(p2, c3)$
$d(p3, c1)$	$d(p3, c2)$	$d(p3, c3)$
$d(p4, c1)$	$d(p4, c2)$	$d(p4, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$



# k-Means clustering – tetapkan titik ke satu centroid

3. Tetapkan setiap titik dataset ke centroid terdekat, dan hitung SSE. SSE merupakan jumlah error yang terjadi antara titik dataset dengan centroid. Tugas k-Means adalah mengoptimalkan (memperkecil) nilai SSE disetiap iterasi.

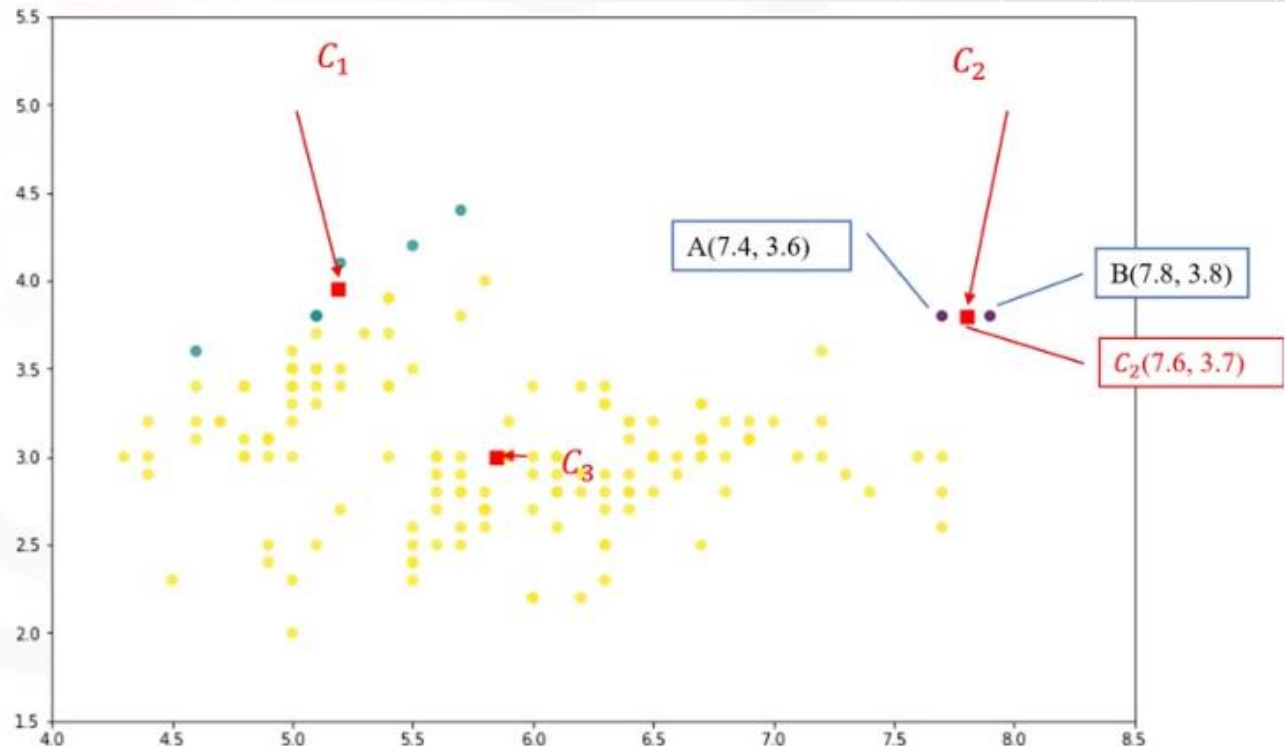
$C_1$	$C_2$	$C_3$
$d(p1, c1)$	$d(p1, c2)$	$d(p1, c3)$
$d(p2, c1)$	$d(p2, c2)$	$d(p2, c3)$
$d(p3, c1)$	$d(p3, c2)$	$d(p3, c3)$
$d(p4, c1)$	$d(p4, c2)$	$d(p4, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(p \dots, c1)$	$d(p \dots, c2)$	$d(p \dots, c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$



# k-Means clustering – compute new centroids

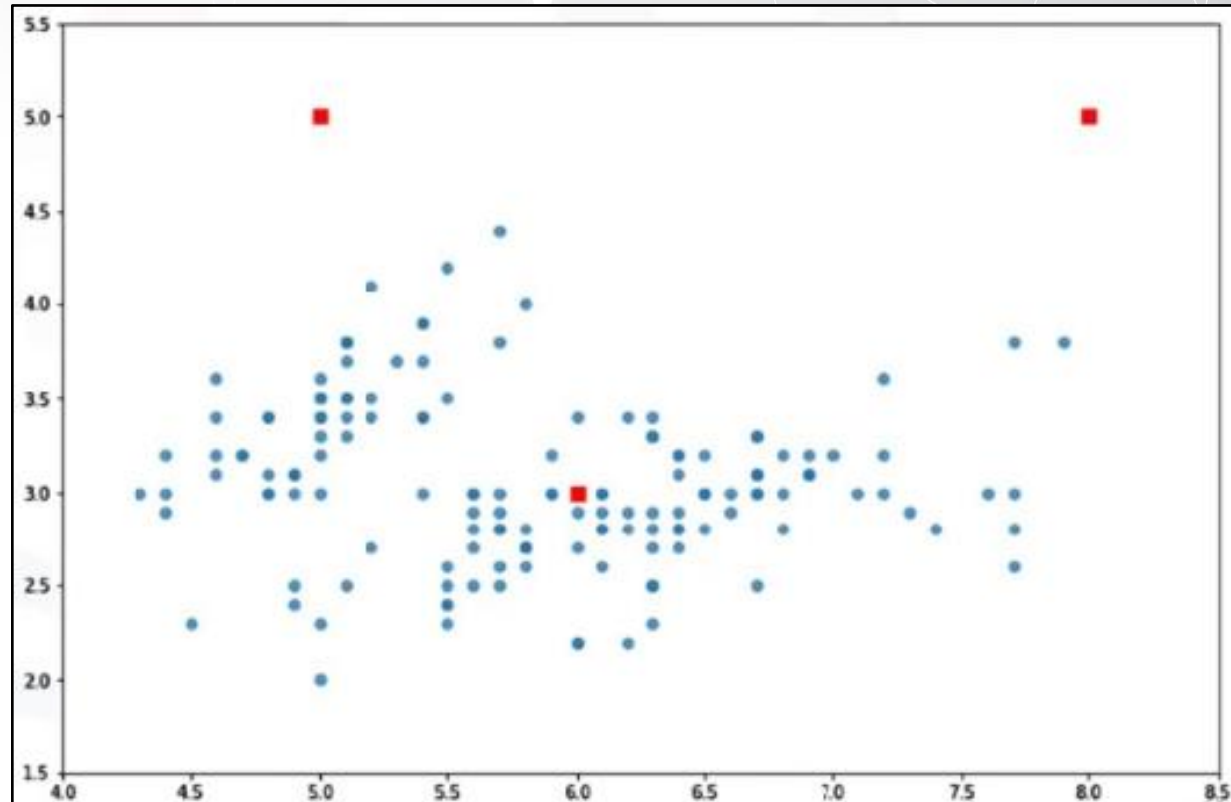
4. Hitung centroid baru dari setiap cluster ( $C_1$ ,  $C_2$ , dan  $C_3$ ).

$C_1$	$C_2$	$C_3$
$d(p1, c1)$	$d(p1, c2)$	$d(p1, c3)$
$d(p2, c1)$	$d(p2, c2)$	$d(p2, c3)$
$d(p3, c1)$	$d(p3, c2)$	$d(p3, c3)$
$d(p4, c1)$	$d(p4, c2)$	$d(p4, c3)$
$d(p..., c1)$	$d(p..., c2)$	$d(p..., c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p..., c1)$	$d(p..., c2)$	$d(p..., c3)$
$d(p..., c1)$	$d(p..., c2)$	$d(p..., c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p..., c1)$	$d(p..., c2)$	$d(p..., c3)$
$d(p..., c1)$	$d(p..., c2)$	$d(p..., c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$
$d(p..., c1)$	$d(p..., c2)$	$d(p..., c3)$
$d(p..., c1)$	$d(p..., c2)$	$d(p..., c3)$
$d(pn, c1)$	$d(pn, c2)$	$d(pn, c3)$



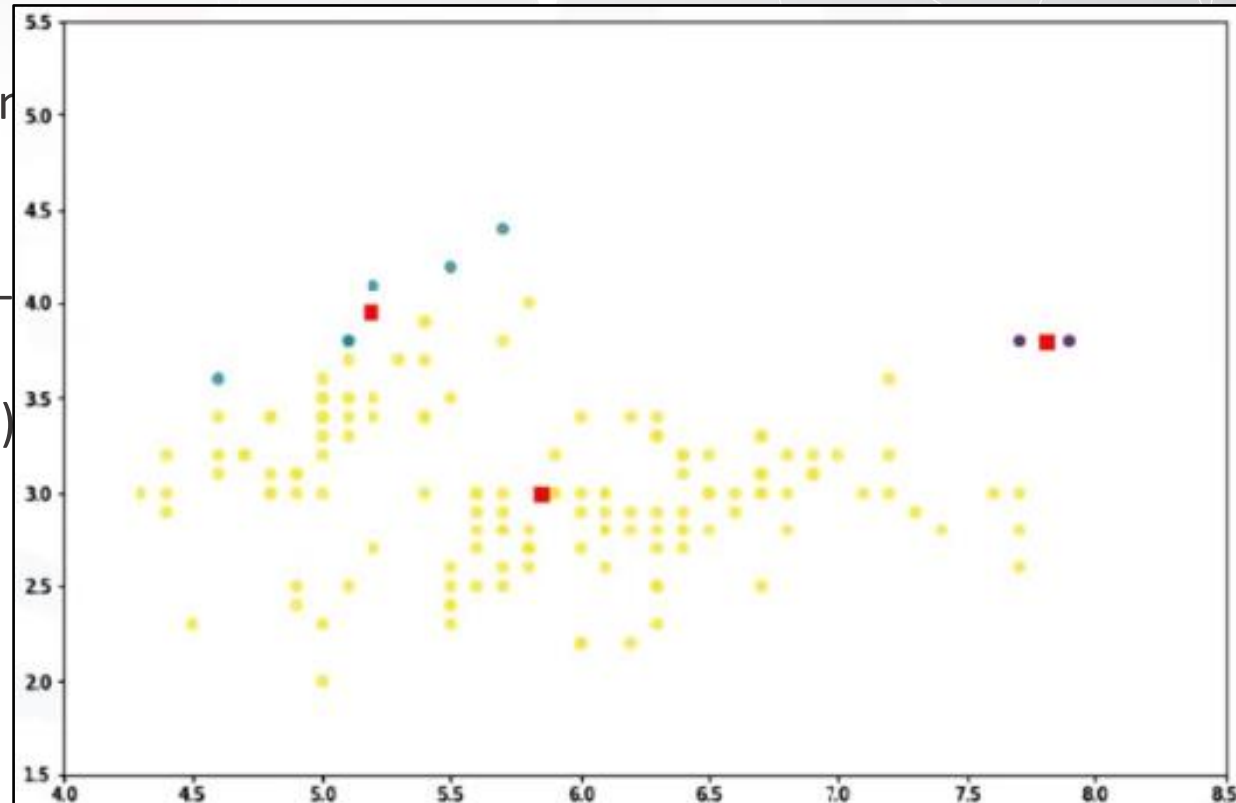
# k-Means clustering – repeat

5. Ulangi proses 2 sampai dengan 4 sampai dengan nilai SSE tertentu atau tidak ada perubahan label kelas pada masing-masing titik (tidak ada perubahan pada cluster)



# k-Means clustering – repeat

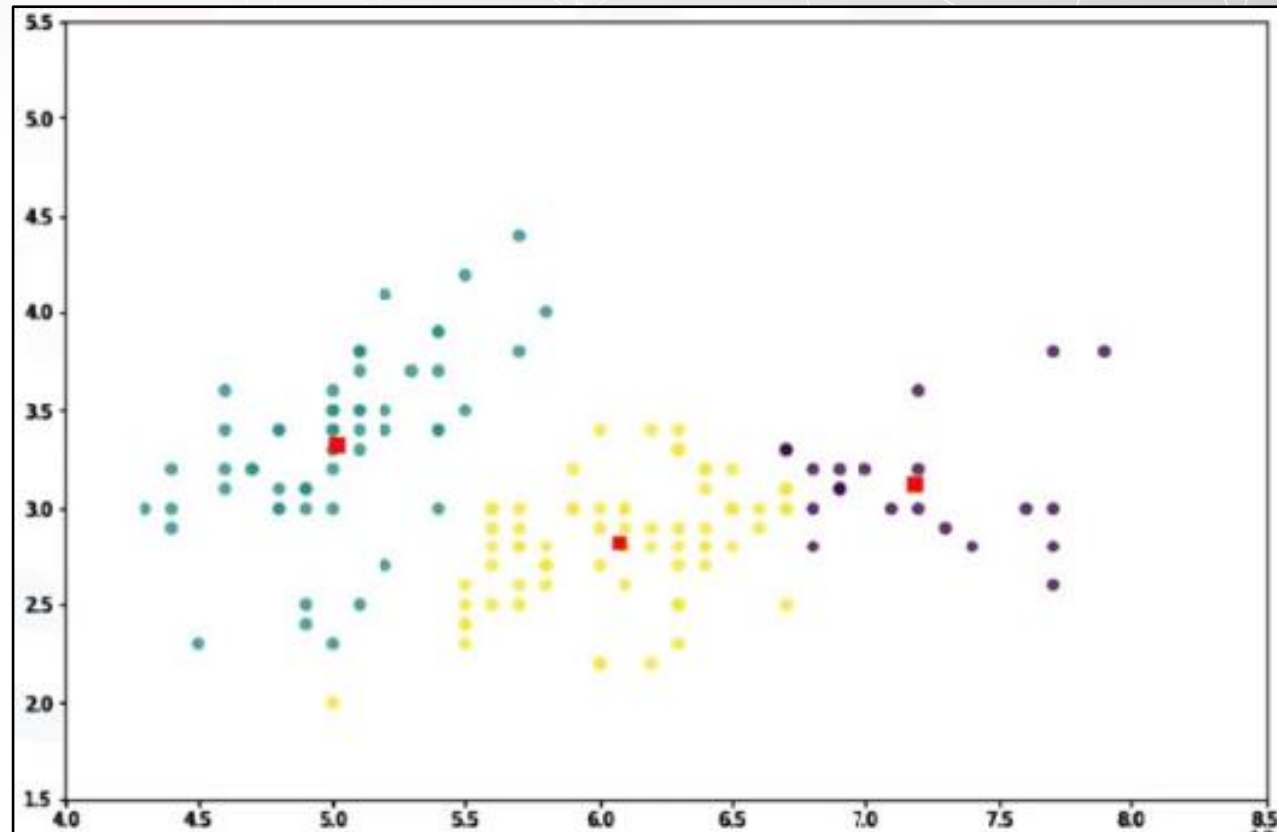
5. Ulangi proses 2 sampai dengan 4 sampai dengan nilai SSE tertentu atau tidak ada perubahan label kelas pada masing-masing titik (tidak ada perubahan pada cluster)





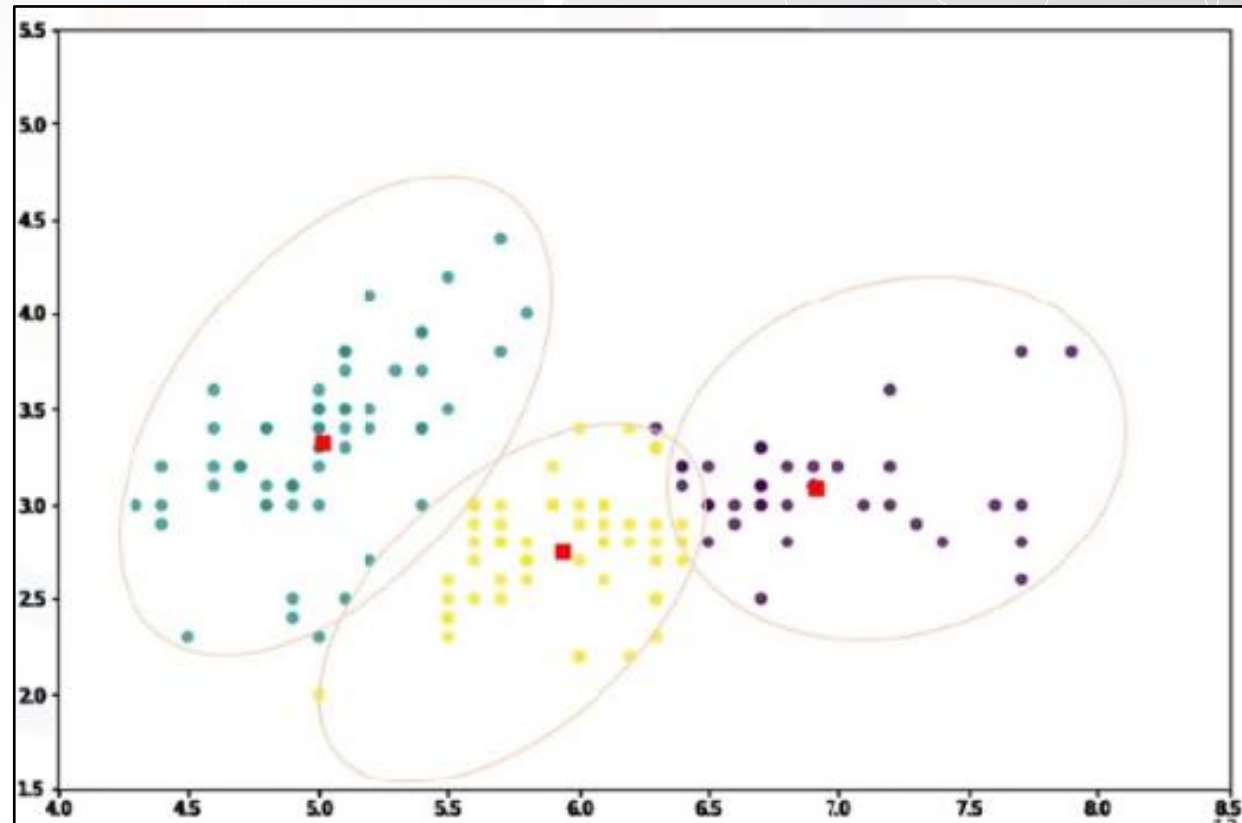
# k-Means clustering – repeat

5. Ulangi proses 2 sampai dengan 4 sampai dengan nilai SSE tertentu atau tidak ada perubahan label kelas pada masing-masing titik (tidak ada perubahan pada cluster)



# k-Means clustering – repeat

5. Ulangi proses 2 sampai dengan 4 sampai dengan nilai SSE tertentu atau tidak ada perubahan label kelas pada masing-masing titik (tidak ada perubahan pada cluster)

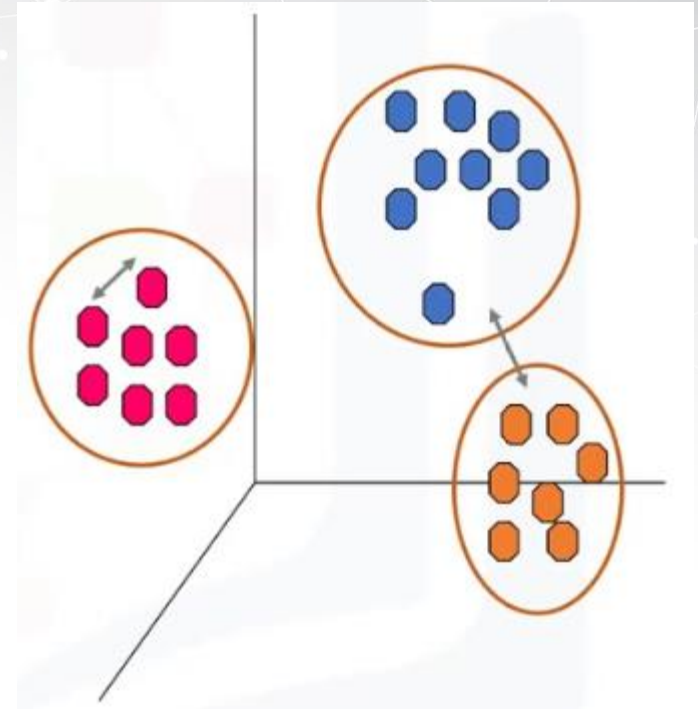


# Algoritma k-Means clustering

1. Tempatkan secara acak  $k$  centroid, satu untuk setiap cluster
2. Hitung jarak setiap titik dari setiap centroid
3. Tetapkan setiap titik data (objek) ke pusat centroid terdekatnya (membuat cluster)
4. Hitung ulang posisi centroid  $k$
5. Ulangi langkah 2-4, hingga centroid tidak lagi bergerak

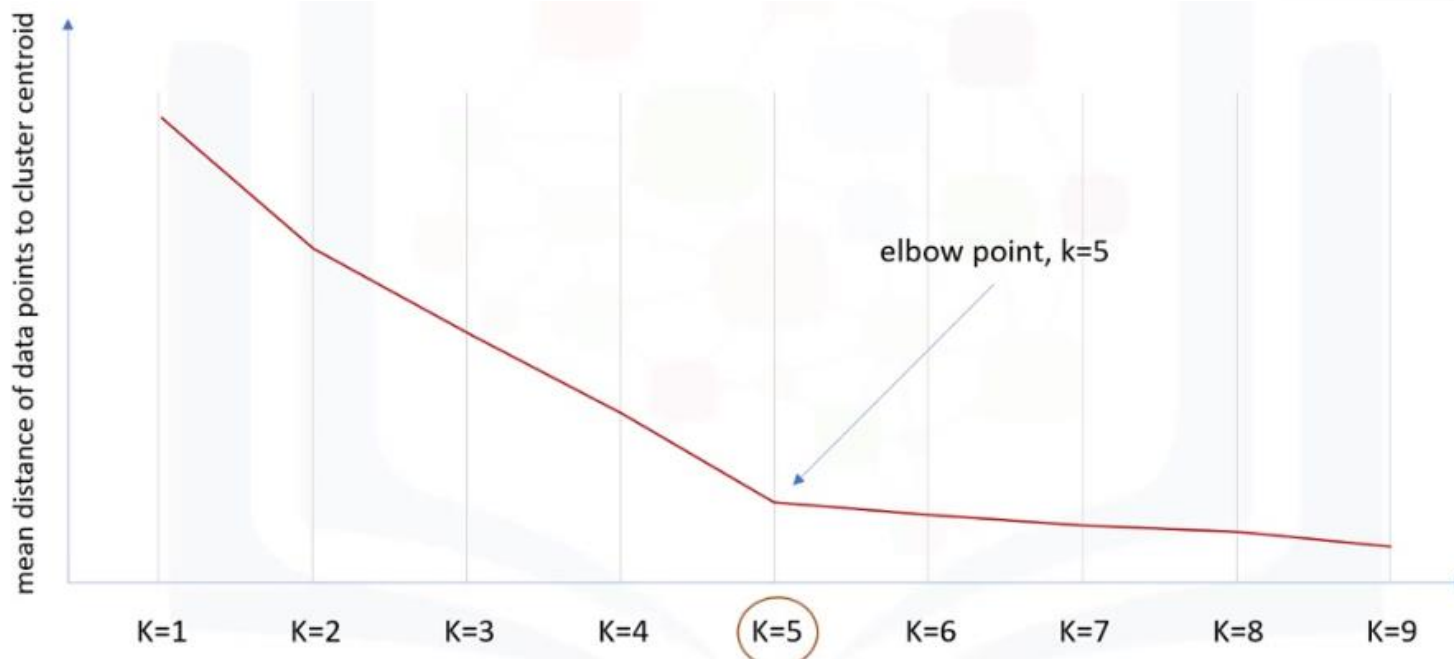
# Akurasi k-Means dengan Distance

- **Pendekatan Eksternal**
  - Bandingkan dengan prediksi cluster dengan Ground Truth (jika ada)
- **Pendekatan Internal**
  - Rata-rata jarak antara titik data dalam sebuah cluster



# Memilih k

- Dengan menghitung jarak rata-rata antara titik data ke centroid, maka dapat ditentukan jumlah centroid  $k$  yang optimal.
- Jumlah  $k$  yang optimal terletak pada *elbow point*. Hal tersebut dikarenakan setelah adanya penurunan rata-rata jarak yang signifikan berubah menjadi sedikit perubahan (landai).





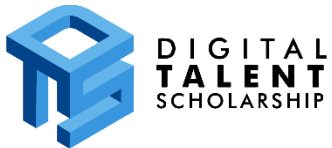






Bagian 2

# Praktikum Lab

ML0101EN-Clus-K-Means-Customer-Seg-py-v1.ipynb

IKUTI KAMI



-  digitalent.kominfo
-  digitalent.kominfo
-  DTS\_kominfo
-  Digital Talent Scholarship 2019

Pusat Pengembangan Profesi dan Sertifikasi  
Badan Penelitian dan Pengembangan SDM  
Kementerian Komunikasi dan Informatika  
Jl. Medan Merdeka Barat No. 9  
(Gd. Belakang Lt. 4 - 5)  
Jakarta Pusat, 10110

