

Unsupervised Learning

Clustering





Hafizh Adi Prasetya

Education Background



2011-2015
Bachelor Degree
Informatics



2017-2019
Master Degree
Artificial Intelligence



Hafizh Adi Prasetya
<https://id.linkedin.com/in/hafizhadi>

Hands On Required

Hands - On : Clustering

Klik disini untuk
mengakses folder Hands
On

Unsupervised Learning

- Sesi I: Intro + Dimensionality Reduction
- **Sesi II : Clustering**

Objektif



Mendapatkan **pemahaman dasar mengenai clustering, aplikasinya di industri, dan konsep jarak antar data.**

Mendapatkan **kemampuan untuk mengimplementasikan proses clustering sederhana** menggunakan Python.

Expected Output



1. Memahami konsep dan motivasi dasar dari proses clustering
2. Memahami konsep segmentasi dan aplikasi teknik clustering di dunia nyata
3. Memahami konsep jarak antara 2 titik data dan kepentingannya dalam unsupervised learning
4. Memahami teknik Agglomerative clustering dan cara mengimplementasikannya menggunakan Python
5. Memahami teknik K-means Clustering dan cara mengimplementasikannya menggunakan Python
6. Memahami dua jenis cara untuk melakukan evaluasi pada cluster yang dihasilkan.

Outline Pembelajaran

Topik Unsupervised Learning

- Sesi I: Intro + Dimensionality Reduction
- Definisi dan Jenis-jenis Unsupervised Learning
- Contoh Kasus Unsupervised Learning
- Dimensionality Reduction dan Penggunaannya
- Intuisi dan Motivasi Principal Component Analysis (PCA)
- PCA (Praktik)
- Algoritma PCA Langkah-demi-langkah

- Sesi II : Clustering**
- Intuisi dan Motivasi Clustering**
- Clustering dan Segmentasi dalam Bisnis**
- Intermezzo: Pengukuran Jarak**
- Algoritma Agglomerative Clustering dan Praktik**
- Algoritma K-means Clustering dan Praktik**
- Evaluasi Clustering**

Apa itu
Clustering?

Outline

Pembelajaran

Topik Unsupervised Learning

- Sesi I: Intro + Dimensionality Reduction
- Definisi dan Jenis-jenis Unsupervised Learning
- Contoh Kasus Unsupervised Learning
- Dimensionality Reduction dan Penggunaannya
- Intuisi dan Motivasi Principal Component Analysis (PCA)
- PCA (Praktik)
- Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering



- Intuisi dan Motivasi Clustering**
- Clustering dan Segmentasi dalam Bisnis**
- Intermezzo: Pengukuran Jarak**
- Algoritma Agglomerative Clustering dan Praktik**
- Algoritma K-means Clustering dan Praktik**
- Evaluasi Clustering**

Memisahkan/mengelompokkan data menjadi beberapa kelompok/cluster/segmen sedemikian rupa sehingga setiap titik data:

- Lebih mirip dengan anggota kelompoknya sendiri
- Dibandingkan dengan anggota kelompok lain

UNSUPERVISED



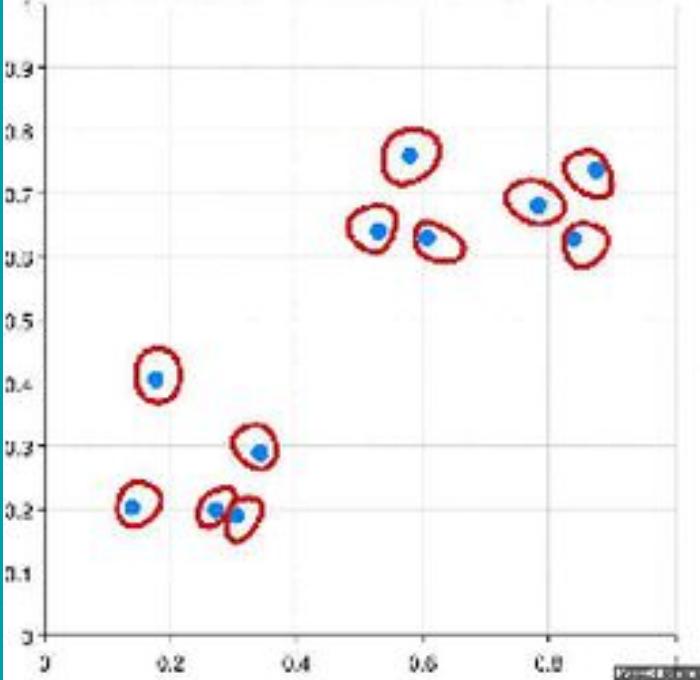
UNSUPERVISED



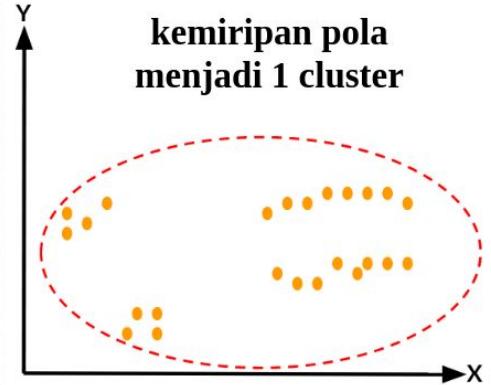
Data
tanpa
label



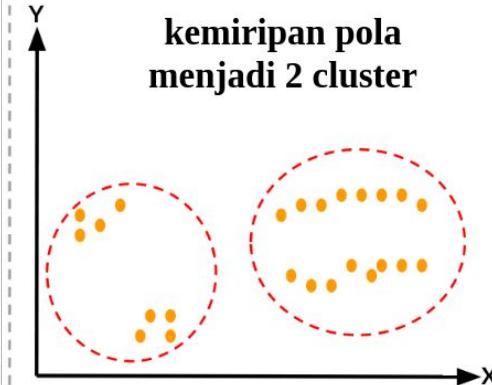
2 Pertanyaan Penting Clustering



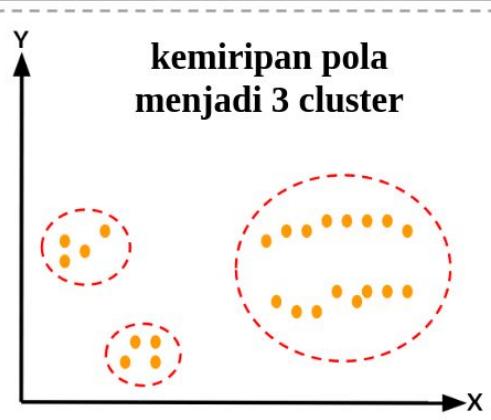
kemiripan pola menjadi 1 cluster



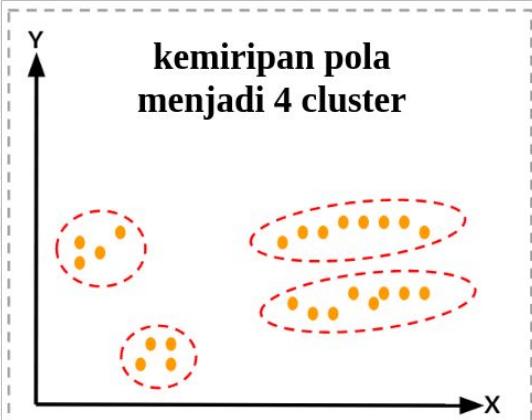
kemiripan pola menjadi 2 cluster



kemiripan pola menjadi 3 cluster



kemiripan pola menjadi 4 cluster



1. Ke dalam berapa kelompok kita ingin memisahkan data?

2. Ke dalam kelompok mana setiap baris data akan ditempatkan?

Menentukan Jumlah Cluster

Spesifikasi dari Bisnis

Sangat sering jumlah cluster ditentukan berdasarkan spesifikasi/konstrain stakeholder bisnis.

Contoh:

- “Kita ingin membagi user menjadi 3-5 kelompok berdasarkan spending”
- “Kita ingin membuat 2 versi homepage untuk personalisasi”

Evaluasi Internal

Tanpa spesifikasi khusus, kita menentukan jumlah cluster “optimal” -> memaksimalkan perbedaan antar cluster.

Contoh:

- Elbow Method
- Silhouette Score

1. Ke dalam berapa kelompok kita ingin memisahkan data?

2. Ke dalam kelompok mana setiap baris data akan ditempatkan?

Algoritma Clustering

Agglomerative Clustering

- Mengelompokkan dua data yang paling dekat
- Deterministik (pasti)

K-means Clustering

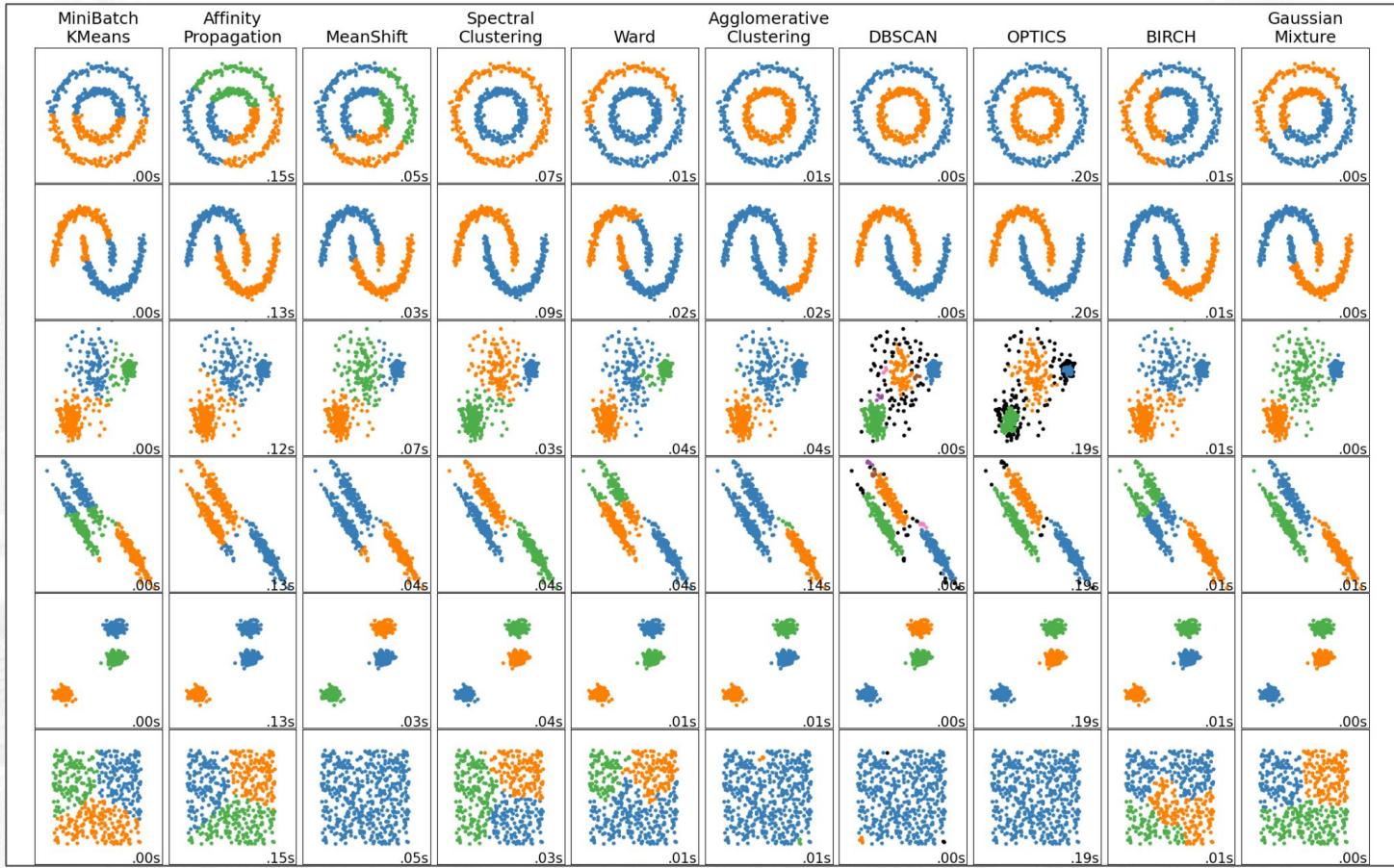
- Mencari centroid yang stabil
- Non-deterministik (acak)

Lainnya

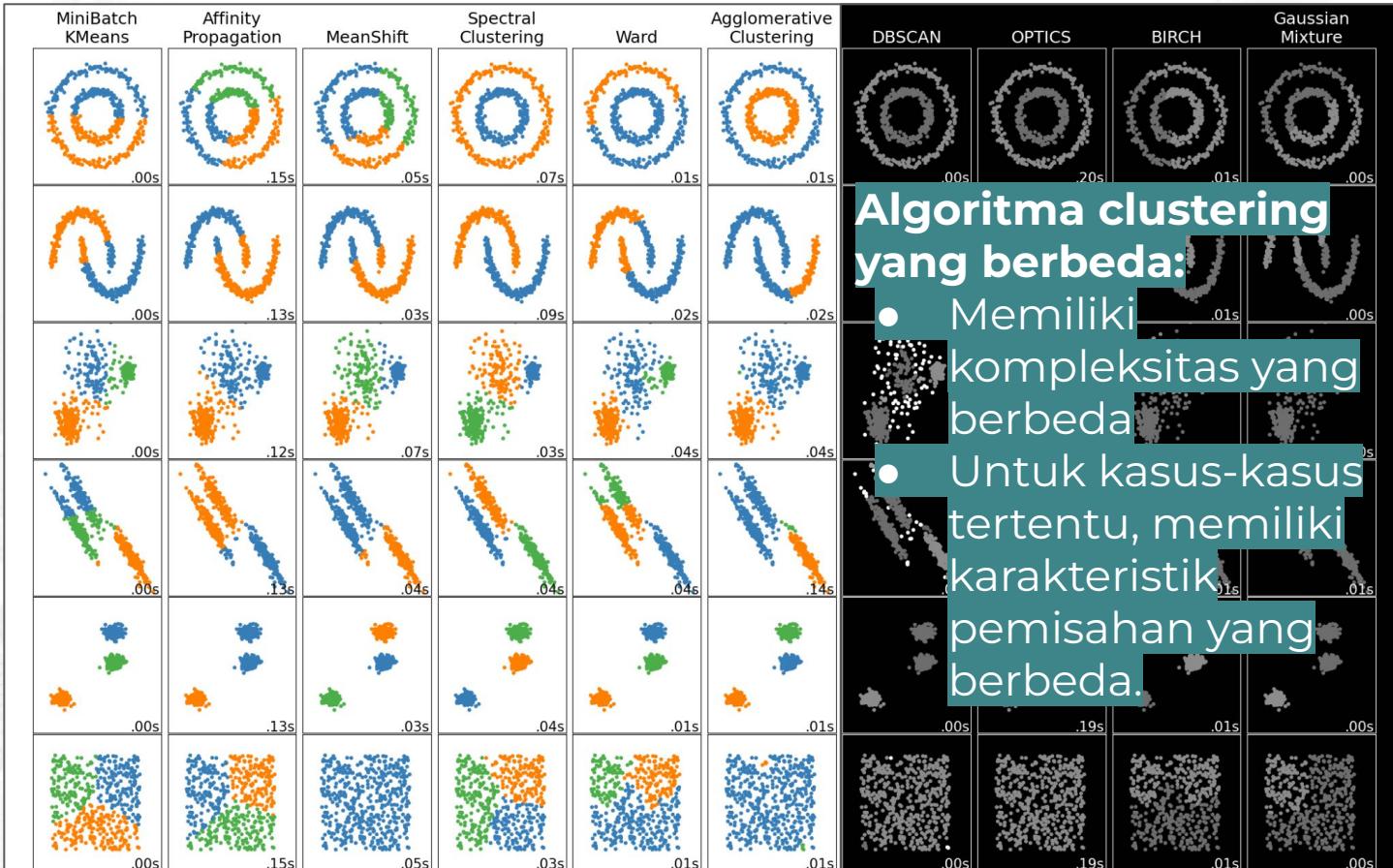
- DBScan
- Gaussian Mixture Model
- Spectral Clustering



Apa yang membedakan bermacam
Algoritma Clustering tersebut?

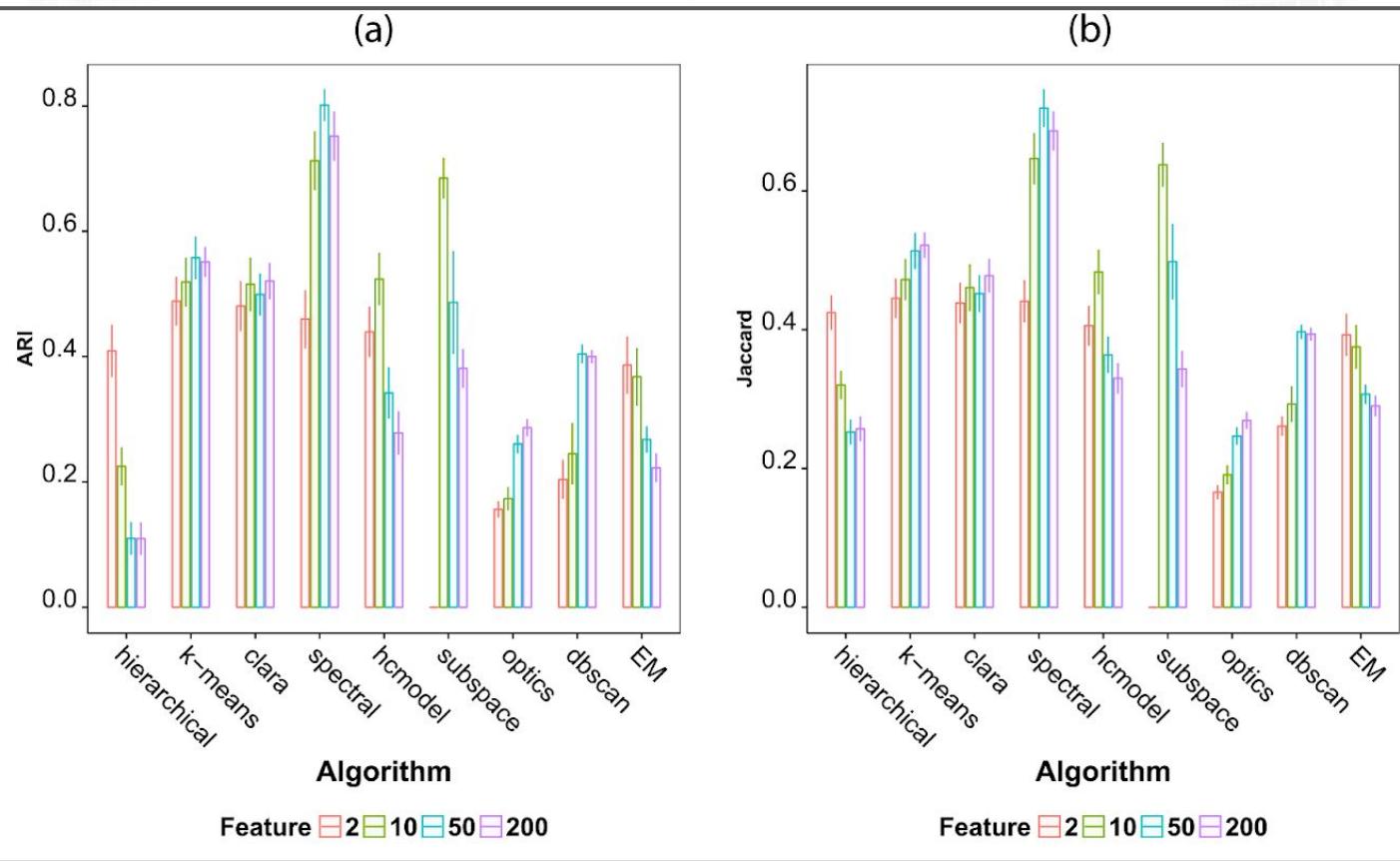






Algoritma clustering yang berbeda:

- Memiliki kompleksitas yang berbeda
- Untuk kasus-kasus tertentu, memiliki karakteristik pemisahan yang berbeda.



Outline

Pembelajaran

Topik Unsupervised Learning

- Sesi I: Intro + Dimensionality Reduction
- Definisi dan Jenis-jenis Unsupervised Learning
- Contoh Kasus Unsupervised Learning
- Dimensionality Reduction dan Penggunaannya
- Intuisi dan Motivasi Principal Component Analysis (PCA)
- PCA (Praktik)
- Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

- Intuisi dan Motivasi Clustering**
- Clustering dan Segmentasi dalam Bisnis**
- Intermezzo: Pengukuran Jarak**
- Algoritma Agglomerative Clustering dan Praktik**
- Algoritma K-means Clustering dan Praktik**
- Evaluasi Clustering**

Segmentasi: Clustering dalam **Konteks Bisnis**

Outline Pembelajaran

Topik Unsupervised Learning

- Sesi I: Intro + Dimensionality Reduction
- Definisi dan Jenis-jenis Unsupervised Learning
- Contoh Kasus Unsupervised Learning
- Dimensionality Reduction dan Penggunaannya
- Intuisi dan Motivasi Principal Component Analysis (PCA)
- PCA (Praktik)
- Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

Intuisi dan Motivasi Clustering



Clustering dan Segmentasi dalam Bisnis

Intermezzo: Pengukuran Jarak

Algoritma Agglomerative Clustering dan Praktik

Algoritma K-means Clustering dan Praktik

Evaluasi Clustering

Contoh Kasus Online Marketplace: Segmentasi

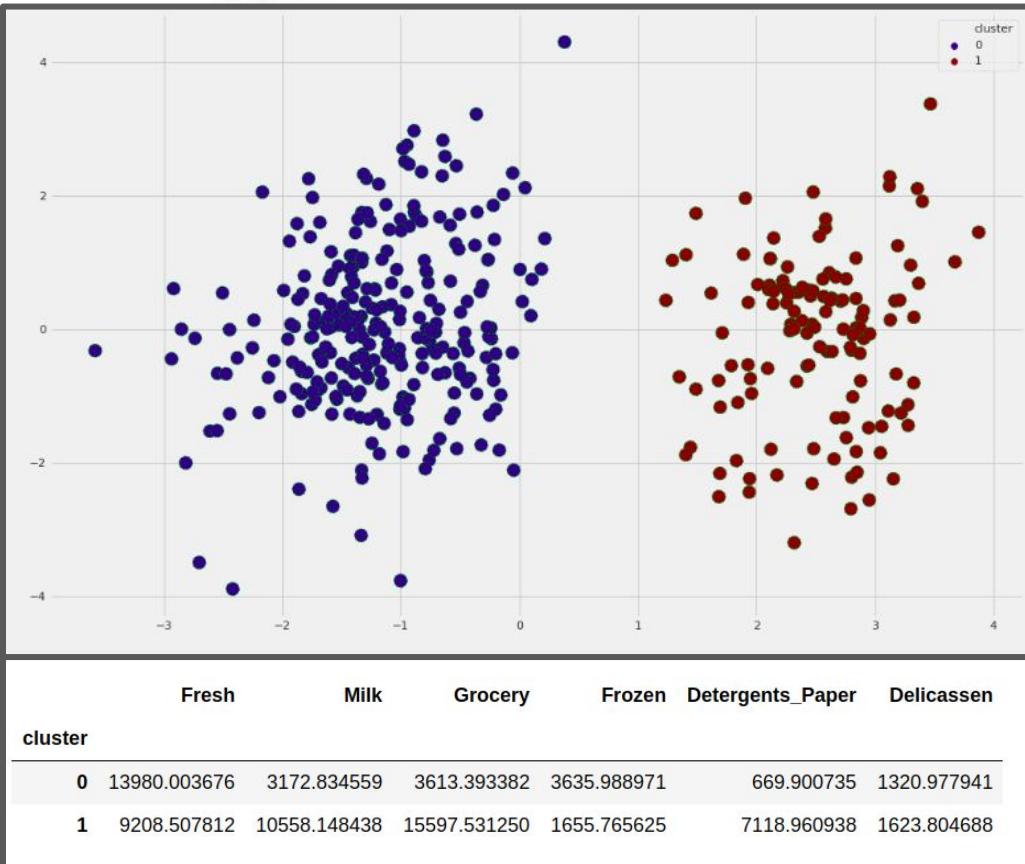
Sangat umum pada bisnis yang berurusan dengan konsumen, kita ingin melakukan **segmentasi**: memisahkan konsumen kita menjadi beberapa kelompok yang berbeda secara konteks bisnis.

Tujuan akhir -> **memberikan perlakuan yang berbeda pada segmen yang berbeda untuk mengoptimasi proses bisnis.**

Data Pelanggan

Satu row: satu pelanggan di marketplace

| ID_Customers | Fresh | Milk | Grocery | Frozen | Detergents_Paper | Delicassen |
|--------------|-------|-------|---------|--------|------------------|------------|
| 0 | 0 | 12669 | 9656 | 7561 | 214 | 2674 |
| 1 | 1 | 7057 | 9810 | 9568 | 1762 | 3293 |
| 2 | 2 | 6353 | 8808 | 7684 | 2405 | 3516 |
| 3 | 3 | 13265 | 1196 | 4221 | 6404 | 507 |
| 4 | 4 | 22615 | 5410 | 7198 | 3915 | 1777 |
| 5 | 5 | 9413 | 8259 | 5126 | 666 | 1795 |
| 6 | 6 | 12126 | 3199 | 6975 | 480 | 3140 |
| 7 | 7 | 7579 | 4956 | 9426 | 1669 | 3321 |
| 8 | 8 | 5963 | 3648 | 6192 | 425 | 1716 |
| 9 | 9 | 6006 | 11093 | 18881 | 1159 | 7425 |



Hasil Segmentasi

- Ditemukan 2 segmen pembeli dengan pola konsumsi yang berbeda
- Kita dapat mendesain strategi marketing yang berbeda untuk 2 segmen tersebut

2 Alasan Segmentasi dalam Bisnis

Memberikan perlakuan khusus

Ciri:

- Produk sudah cukup dewasa
- Pemahaman tinggi tentang customer
- Gambaran konkret tentang *action item*

Contoh use case:

- Personalisasi
- Targeted campaign

Eksplorasi tipe-tipe customer

Ciri:

- Produk/fitur baru
- Pemahaman tentang customer masih lemah
- Tidak ada gambaran mengenai *action item*

Contoh use case:

- Desain produk baru
- Analisis produk

Kapan Clustering dan Kapan Manual?

- Requirement teknis/bisnis yang detil
- Jumlah fitur sedikit dan sudah ditentukan
- Auditabilitas dan akuntabilitas segmen penting
- Segmen digunakan tim bisnis

Segmentasi Manual

Clustering

- Requirement teknis/bisnis tidak terlalu banyak
- Jumlah fitur yang banyak
- Segmen yang homogen dan terpisah penting
- Segmen digunakan untuk eksperimentasi

Outline Pembelajaran

Topik Unsupervised Learning

- Sesi I: Intro + Dimensionality Reduction
- Definisi dan Jenis-jenis Unsupervised Learning
- Contoh Kasus Unsupervised Learning
- Dimensionality Reduction dan Penggunaannya
- Intuisi dan Motivasi Principal Component Analysis (PCA)
- PCA (Praktik)
- Algoritma PCA Langkah-demi-langkah

- Sesi II : Clustering**
- Intuisi dan Motivasi Clustering**
- Clustering dan Segmentasi dalam Bisnis**
- Intermezzo: Pengukuran Jarak**
- Algoritma Agglomerative Clustering dan Praktik**
- Algoritma K-means Clustering dan Praktik**
- Evaluasi Clustering**

Ukuran Jarak antar Data

Outline Pembelajaran

Topik Unsupervised Learning

- Sesi I: Intro + Dimensionality Reduction
- Definisi dan Jenis-jenis Unsupervised Learning
- Contoh Kasus Unsupervised Learning
- Dimensionality Reduction dan Penggunaannya
- Intuisi dan Motivasi Principal Component Analysis (PCA)
- PCA (Praktik)
- Algoritma PCA Langkah-demi-langkah

- Sesi II : Clustering**
- Intuisi dan Motivasi Clustering
- Clustering dan Segmentasi dalam Bisnis
-  **Intermezzo: Pengukuran Jarak**
- Algoritma Agglomerative Clustering dan Praktik**
- Algoritma K-means Clustering dan Praktik**
- Evaluasi Clustering**

Bagaimana mengukur kedekatan? Bagaimana cara mengukur ‘jarak’ antara dua titik data?

3 contoh:

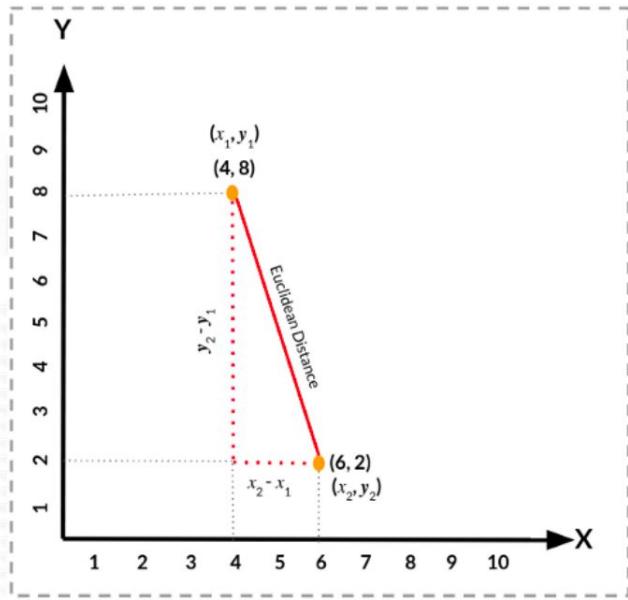
Euclidean Distance

Manhattan Distance

Levenshtein
Distance

Euclidean Distance

$$\sqrt{\sum_{i=1}^n (q_i - p_i)^2}.$$



$$\begin{array}{ccccc} \begin{matrix} 2 \\ 6 \end{matrix} & - & \begin{matrix} 8 \\ 4 \end{matrix} & = & \begin{matrix} -6 \\ 2 \end{matrix} \\ \text{q} & & \text{point 1} & & \\ & & & & \\ \begin{matrix} 36 \\ 4 \end{matrix} & & & \xrightarrow{\text{Kuadrat}} & \begin{matrix} 40 \\ 4 \end{matrix} \\ & & & & \xrightarrow{\text{Total}} \\ & & & & \begin{matrix} 6.3 \\ \text{Akar} \end{matrix} \end{array}$$

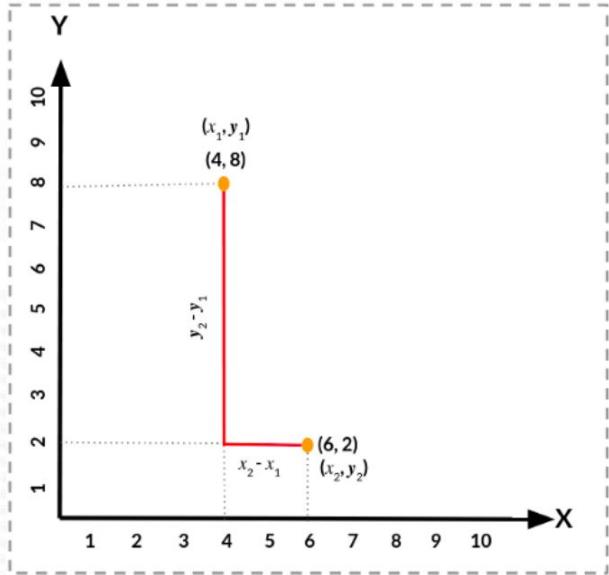
Manhattan Distance

$$\sum_{i=1}^n |p_i - q_i|$$

| | | | | | | | | |
|---|---|---|---|----|--------------|---|-------|---|
| 2 | - | 8 | = | -6 | Absolut e | 6 | Total | 8 |
| 6 | - | 4 | = | 2 | | 2 | | |

q **p** Total

point 1 point 1



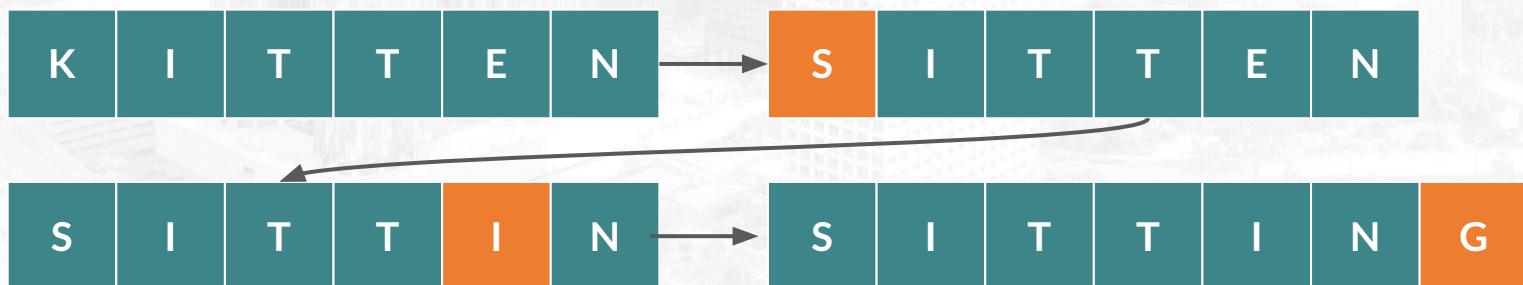
Levenshtein/Minimum Edit Distance

(jarak antar kata)

$$\text{lev}(a, b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \text{lev}(\text{tail}(a), \text{tail}(b)) & \text{if } a[0] = b[0] \\ 1 + \min \begin{cases} \text{lev}(\text{tail}(a), b) \\ \text{lev}(a, \text{tail}(b)) \\ \text{lev}(\text{tail}(a), \text{tail}(b)) \end{cases} & \text{otherwise,} \end{cases}$$

lev(kitten, smitten) = 3

- kitten → sitten ("s" diganti "k"),
- sitten → sittin ("i" diganti "e"),
- sittin → sitting
(penambahan "g" di ujung).



Outline Pembelajaran

Topik Unsupervised Learning

- Sesi I: Intro + Dimensionality Reduction
- Definisi dan Jenis-jenis Unsupervised Learning
- Contoh Kasus Unsupervised Learning
- Dimensionality Reduction dan Penggunaannya
- Intuisi dan Motivasi Principal Component Analysis (PCA)
- PCA (Praktik)
- Algoritma PCA Langkah-demi-langkah

- Sesi II : Clustering**
- Intuisi dan Motivasi Clustering**
- Clustering dan Segmentasi dalam Bisnis**
- Intermezzo: Pengukuran Jarak**
- Algoritma Agglomerative Clustering dan Praktik**
- Algoritma K-means Clustering dan Praktik**
- Evaluasi Clustering**

Algoritma **Agglomerative Clustering**

Outline Pembelajaran

Topik Unsupervised Learning

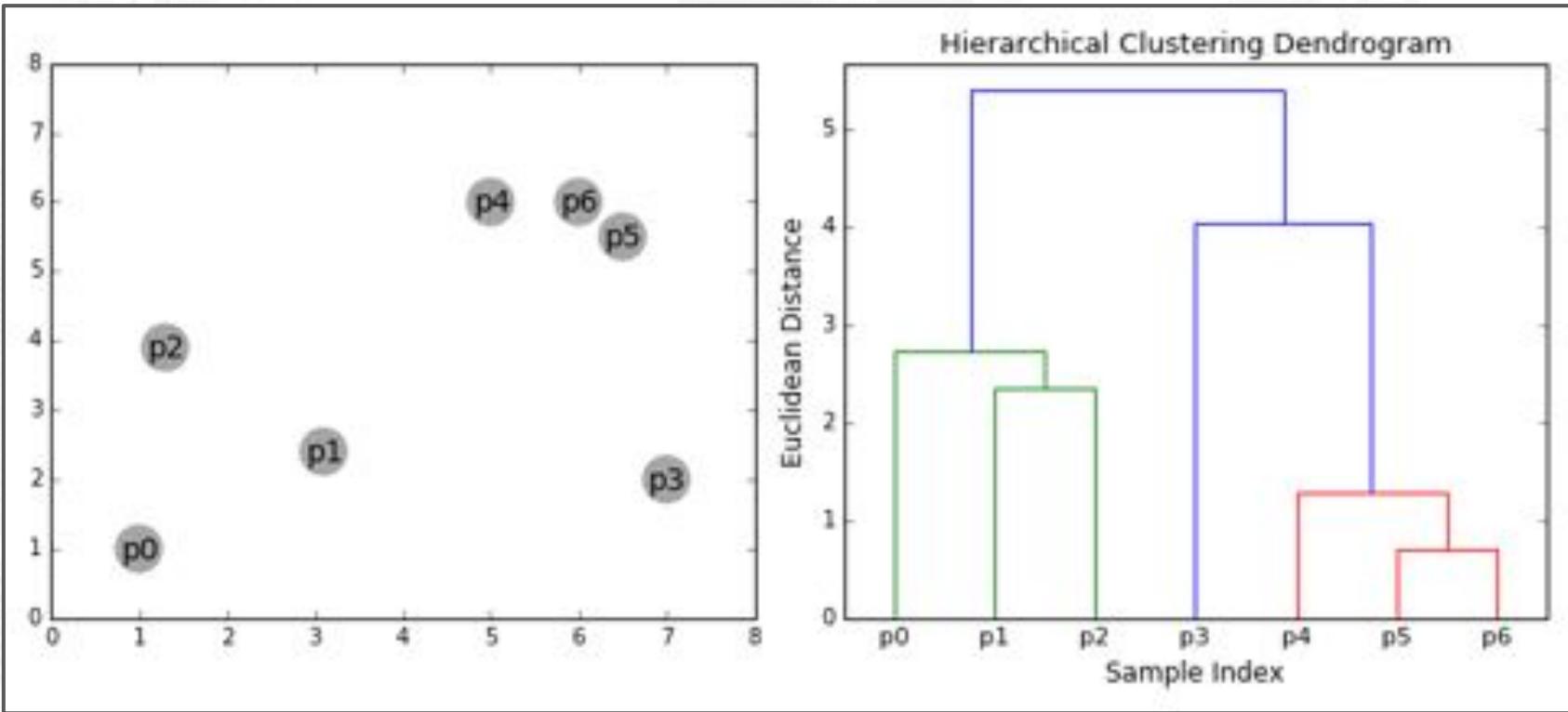
- Sesi I: Intro + Dimensionality Reduction
- Definisi dan Jenis-jenis Unsupervised Learning
- Contoh Kasus Unsupervised Learning
- Dimensionality Reduction dan Penggunaannya
- Intuisi dan Motivasi Principal Component Analysis (PCA)
- PCA (Praktik)
- Algoritma PCA Langkah-demi-langkah

- Sesi II : Clustering**
- Intuisi dan Motivasi Clustering
- Clustering dan Segmentasi dalam Bisnis
- Intermezzo: Pengukuran Jarak
-  Algoritma Agglomerative Clustering dan Praktik
- Algoritma K-means Clustering dan Praktik
- Evaluasi Clustering

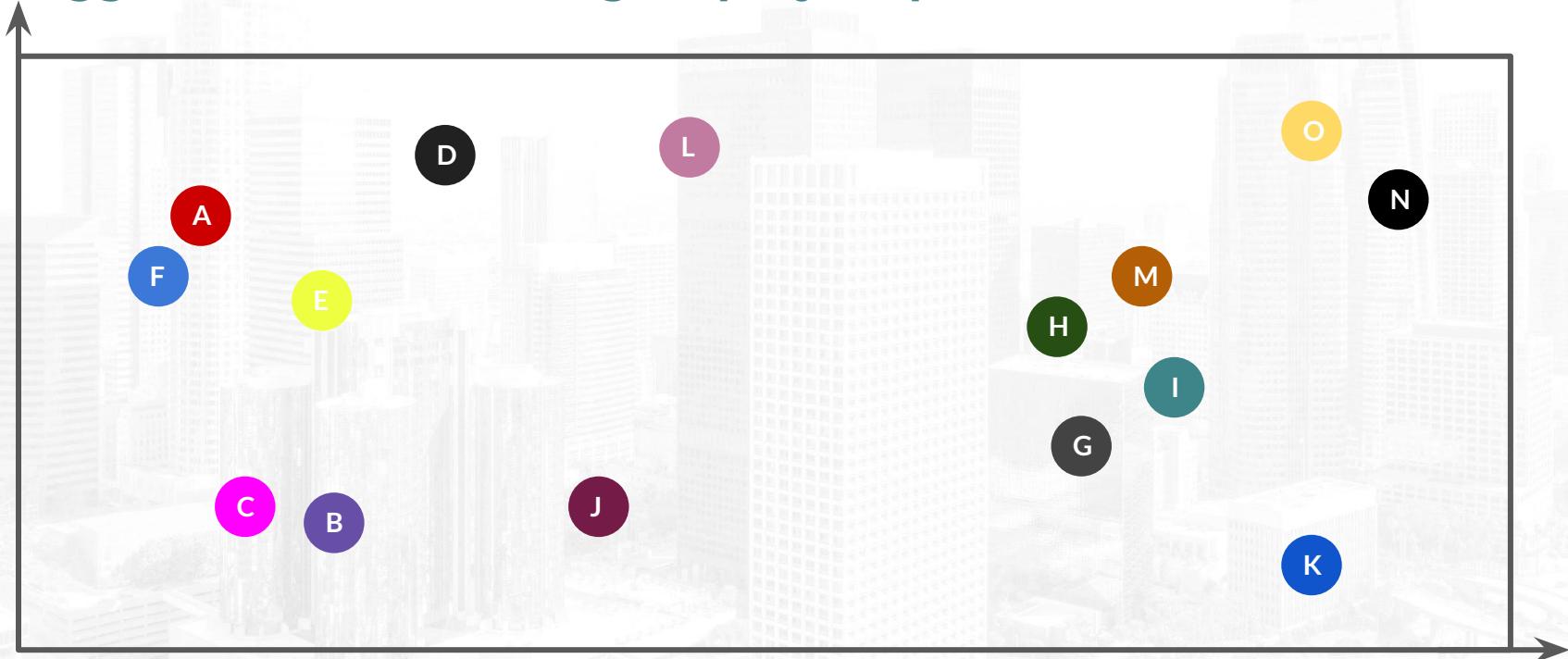
Prinsip Utama Agglomerative Clustering

“2 titik yang paling dekat berada di cluster yang sama.”

Intuisi

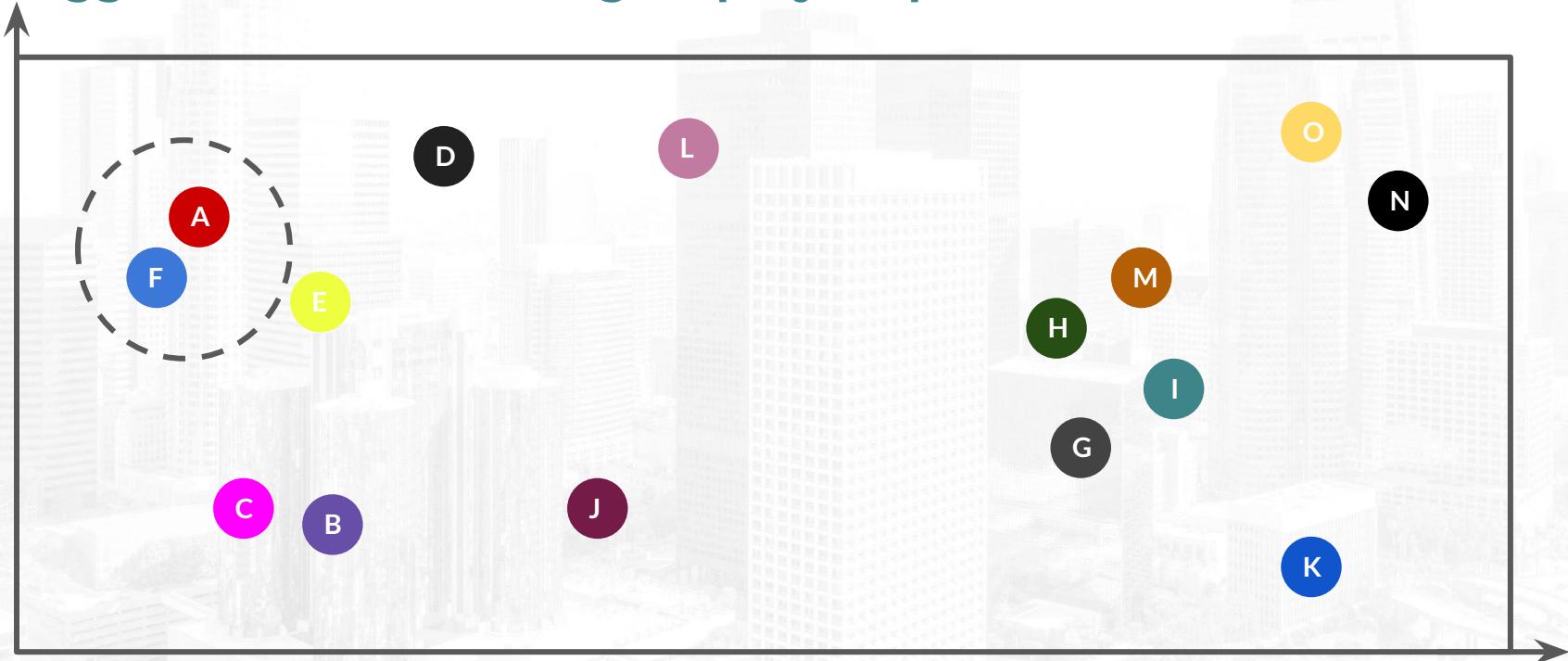


Agglomerative Clustering Step-by-step: N Cluster



- Pertama kita tentukan jumlah cluster yang ingin kita dapatkan di akhir MISAL: 2 CLUSTER
- Di awal algoritma, setiap titik data adalah clusternya masing-masing

Agglomerative Clustering Step-by-step: 2 Cluster Terdekat

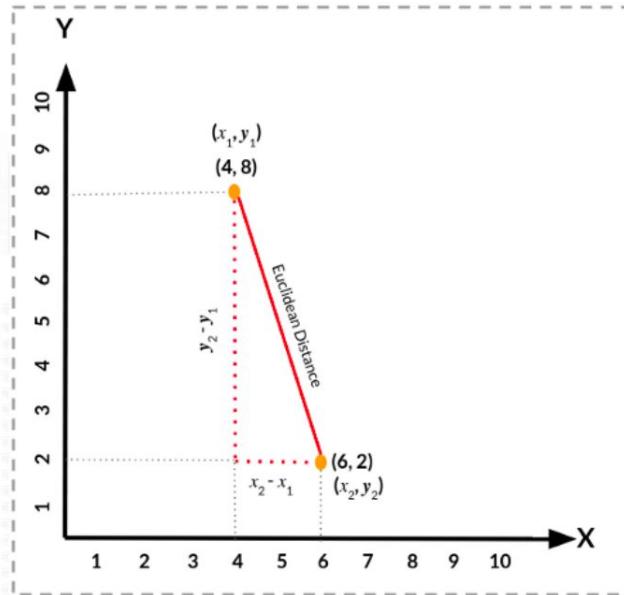


- Cari 2 cluster yang paling dekat satu sama lain
- Dalam kasus di atas, A dan F

Kilas Balik mengukur kedekatan

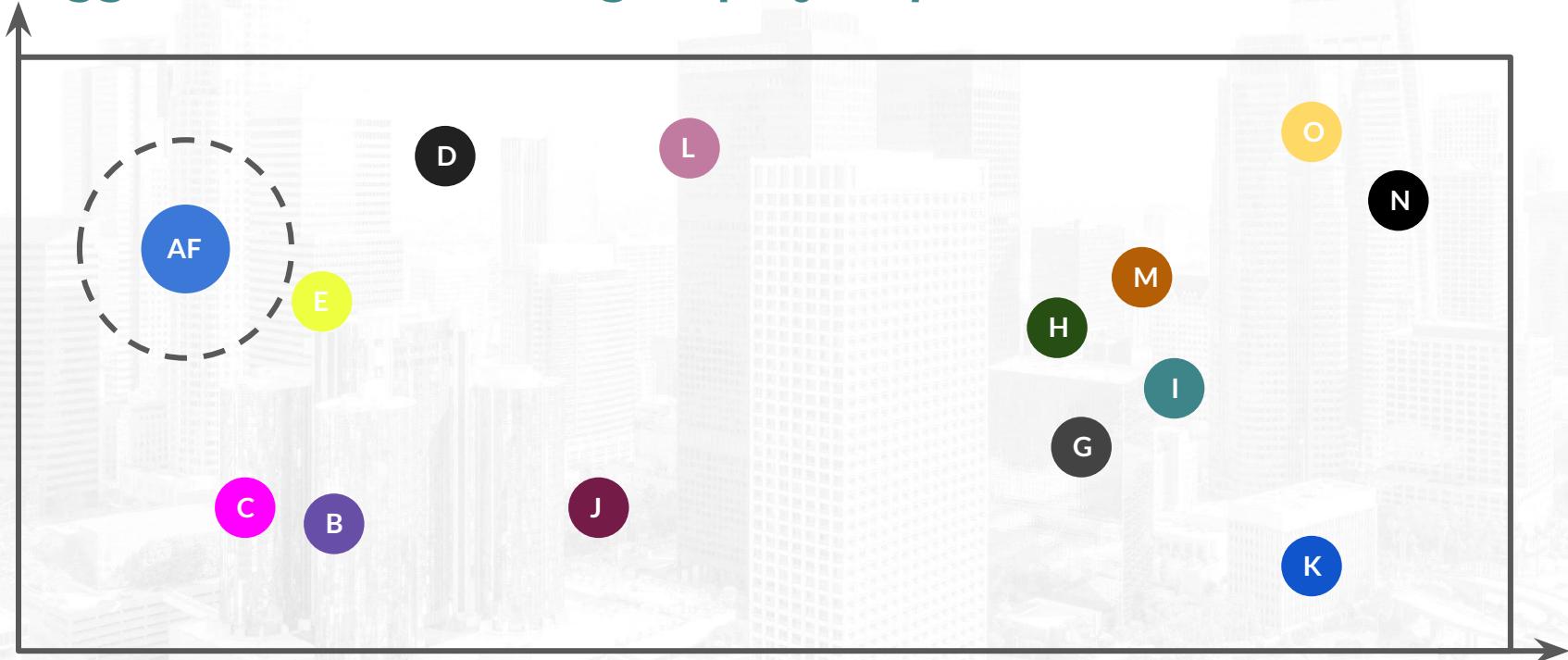
Euclidean Distance

$$\sqrt{\sum_{i=1}^n (q_i - p_i)^2}.$$



$$\begin{array}{ccccc} \begin{matrix} 2 \\ 6 \end{matrix} & - & \begin{matrix} 8 \\ 4 \end{matrix} & = & \begin{matrix} -6 \\ 2 \end{matrix} \\ \text{q} & & \text{point 1} & & \\ & & & & \\ \begin{matrix} 36 \\ 4 \end{matrix} & & & & \begin{matrix} \text{Kuadrat} \\ \longrightarrow \end{matrix} \\ & & & & \begin{matrix} 40 \\ 4 \end{matrix} \\ & & & & \begin{matrix} \text{Total} \\ \longrightarrow \end{matrix} \\ & & & & \begin{matrix} 40 \\ 4 \end{matrix} \\ & & & & \begin{matrix} \text{Akar} \\ \longrightarrow \end{matrix} \\ & & & & \begin{matrix} 6.3 \\ \boxed{6.3} \end{matrix} \end{array}$$

Agglomerative Clustering Step-by-step: 1 Cluster Baru



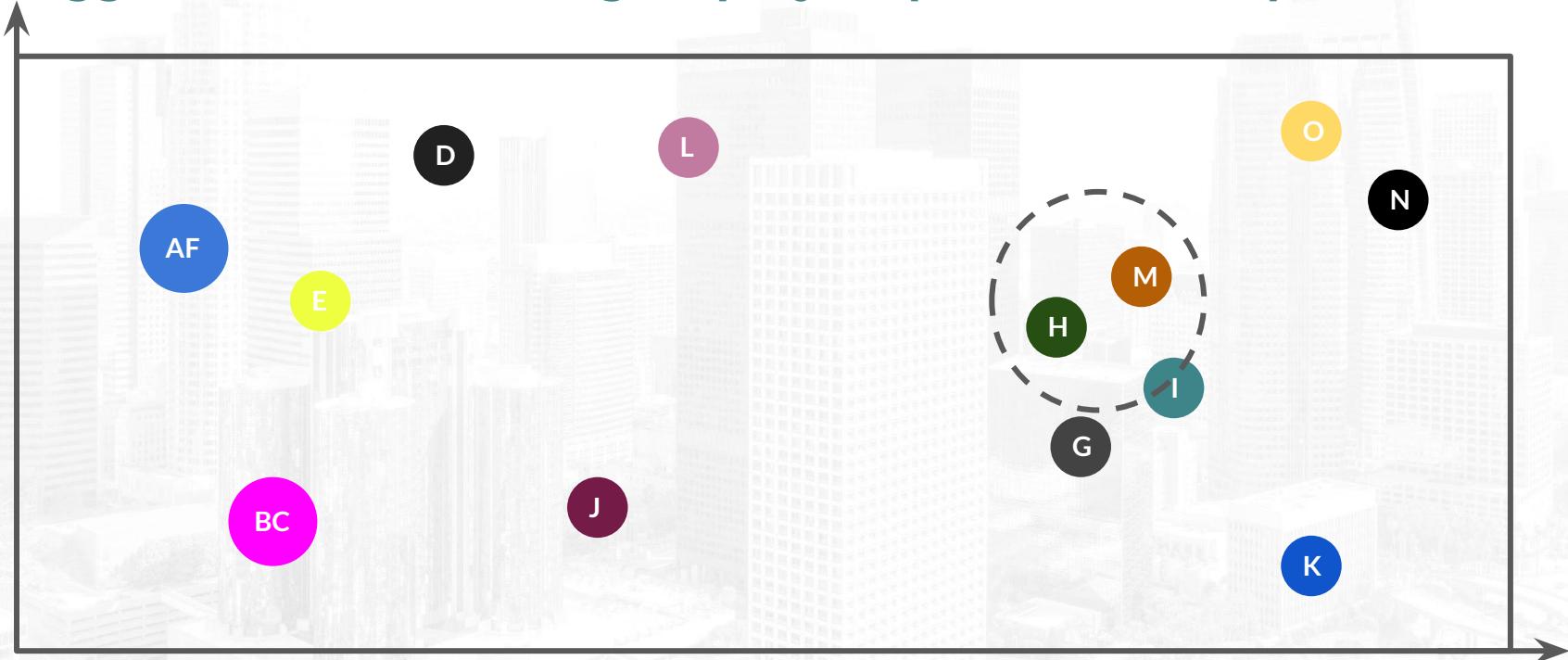
- Kita anggap kedua cluster tersebut adalah satu. Sekarang jumlah cluster berkurang 1.
- Kemudian kita ambil rata2 koordinat kedua cluster tersebut; anggap itu sebagai koordinat cluster baru, AF

Agglomerative Clustering Step-by-step: Rinse and Repeat



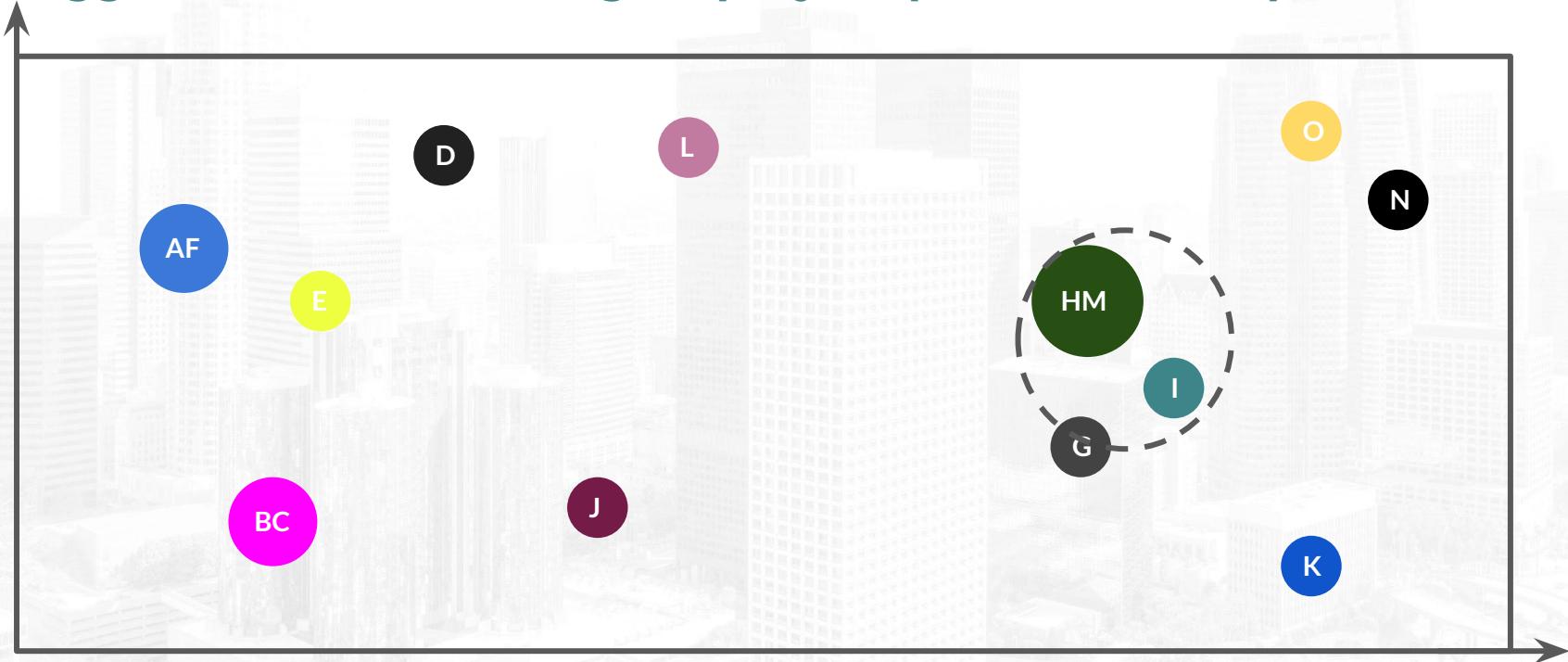
- Apakah kita sudah mendapatkan 2 cluster? **BELUM**
- Maka kita ulang kembali langkah-langkah tadi - sekarang C dan B adalah 2 cluster yang paling dekat

Agglomerative Clustering Step-by-step: Rinse and Repeat



- Cluster B dan C bergabung menjadi cluster BC. Apakah sudah tersisa 2 cluster?
BELUM
- Kita satukan 2 cluster lagi - sekarang H dan M adalah 2 cluster yang paling dekat

Agglomerative Clustering Step-by-step: Rinse and Repeat



- Cluster H dan M bergabung menjadi cluster HM. Apakah sudah tersisa 2 cluster?
BELUM
- Kita satukan 2 cluster lagi - sekarang HM dan I adalah 2 cluster yang paling dekat

Agglomerative Clustering Step-by-step: Rinse and Repeat



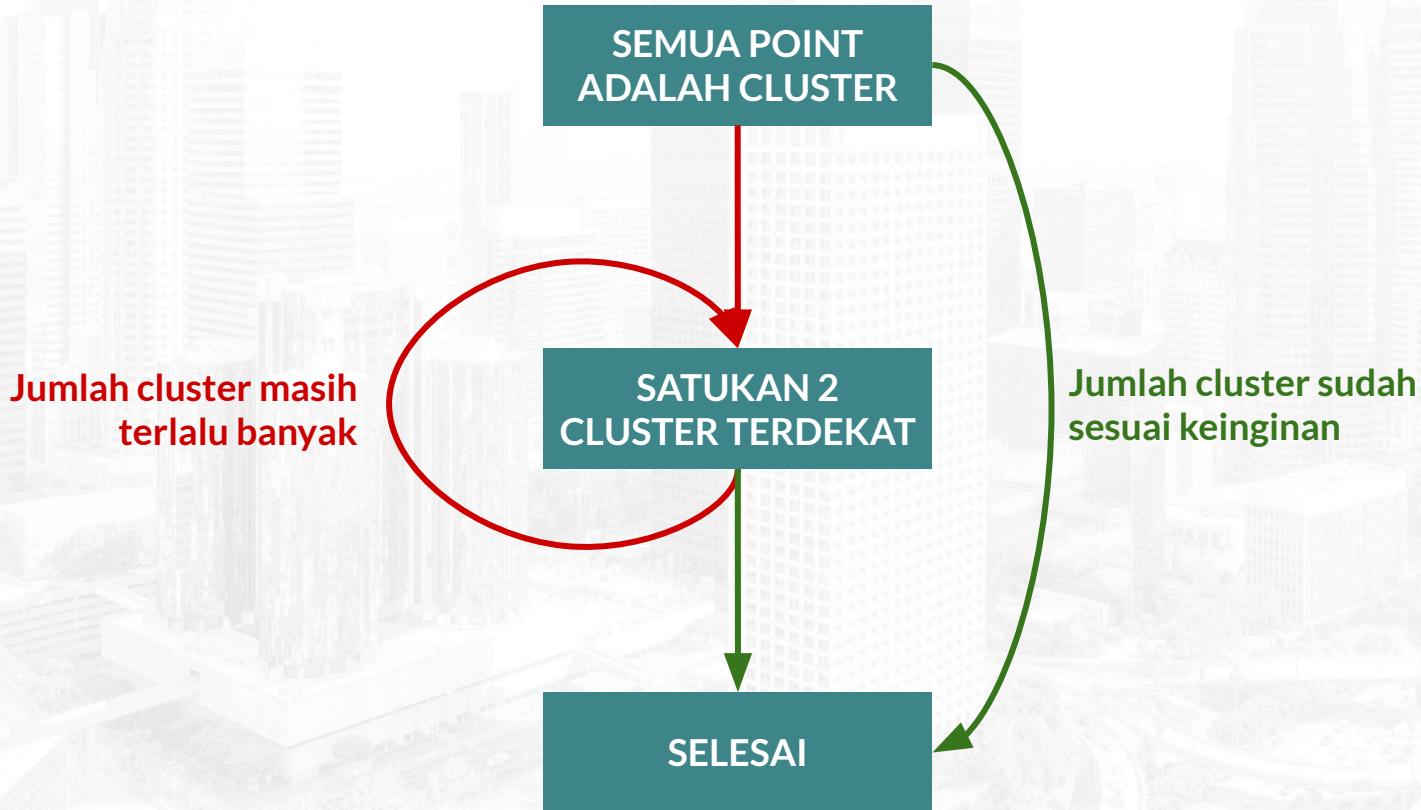
- Demikian proses akan terus diulangi hingga tersisa 2 cluster
- Berdasarkan gambar di atas, cluster selanjutnya yang akan terbentuk adalah cluster ON, HMIG, dan AFE

Agglomerative Clustering Step-by-step: Final

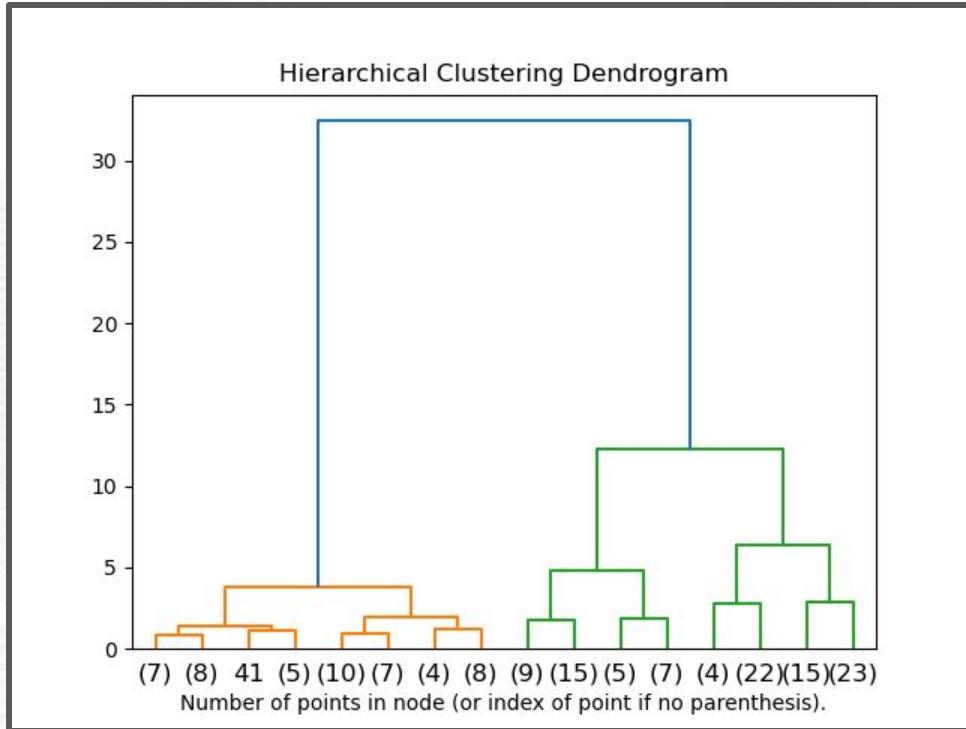


- Pada akhirnya kita akan sampai di keadaan dimana hanya 2 cluster yang tersisa
- Algoritma SELESAI!

Agglomerative Clustering Overview



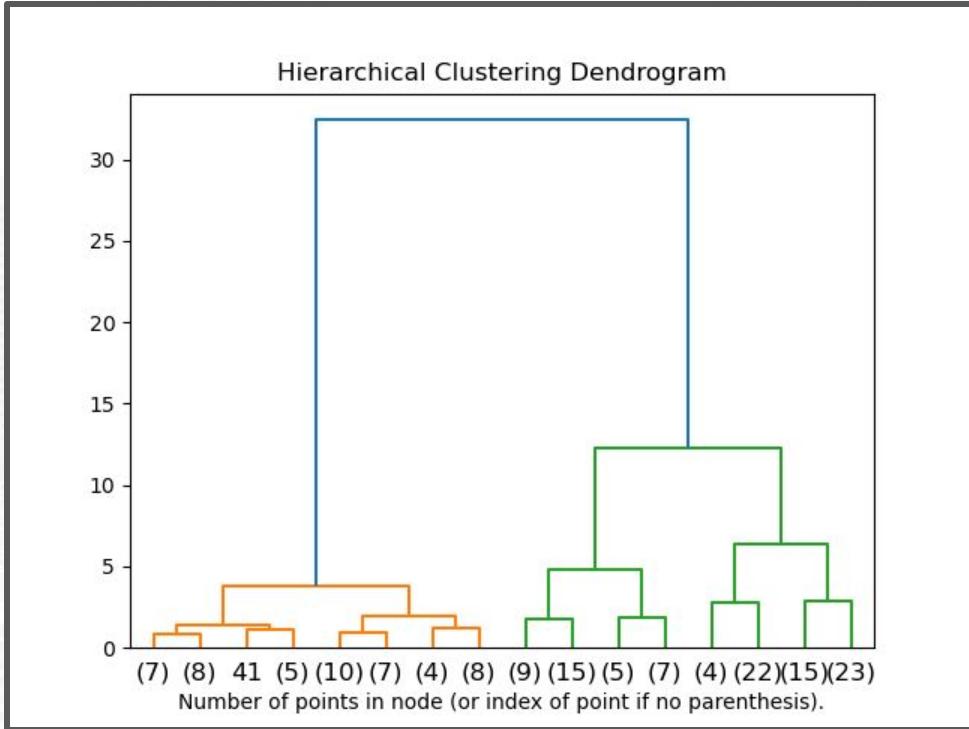
Evaluasi Internal Agglomerative Clustering: Dendrogram



Dendrogram menunjukkan proses agglomerative clustering

- Sumbu X menunjukkan ujung cluster yang tersedia di awal-awal proses (atau sampai batas kedalaman tertentu)
- Sumbu Y menunjukkan jarak antar cluster

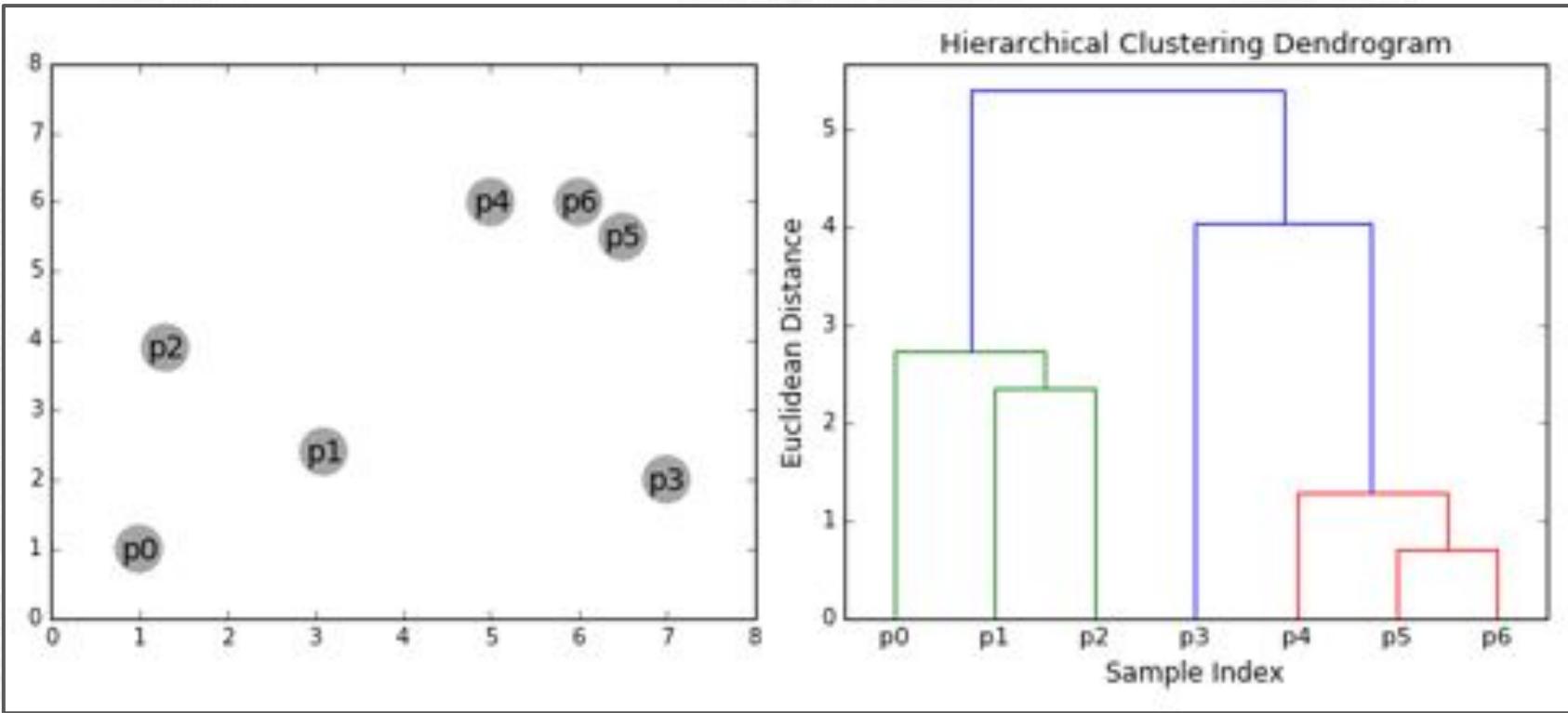
Evaluasi Internal Agglomerative Clustering: Cluster optimal?



Dari dendrogram, berapa cluster yang optimal?

- Kita tidak dapat tahu berapa cluster yang optimal hanya dari dendrogram
- Yang dapat kita lakukan adalah melihat jarak antar 2 cluster
- Kita bisa menentukan ‘maksimal’ jarak yang kita inginkan
- Apabila jarak antara 2 cluster lebih besar dari batas maksimal berarti saatnya berhenti

Kilas balik:



Implementasi **Agglomerative Clustering**

Dataset

Berat_tinggi (sintetis)

- **Deskripsi:**

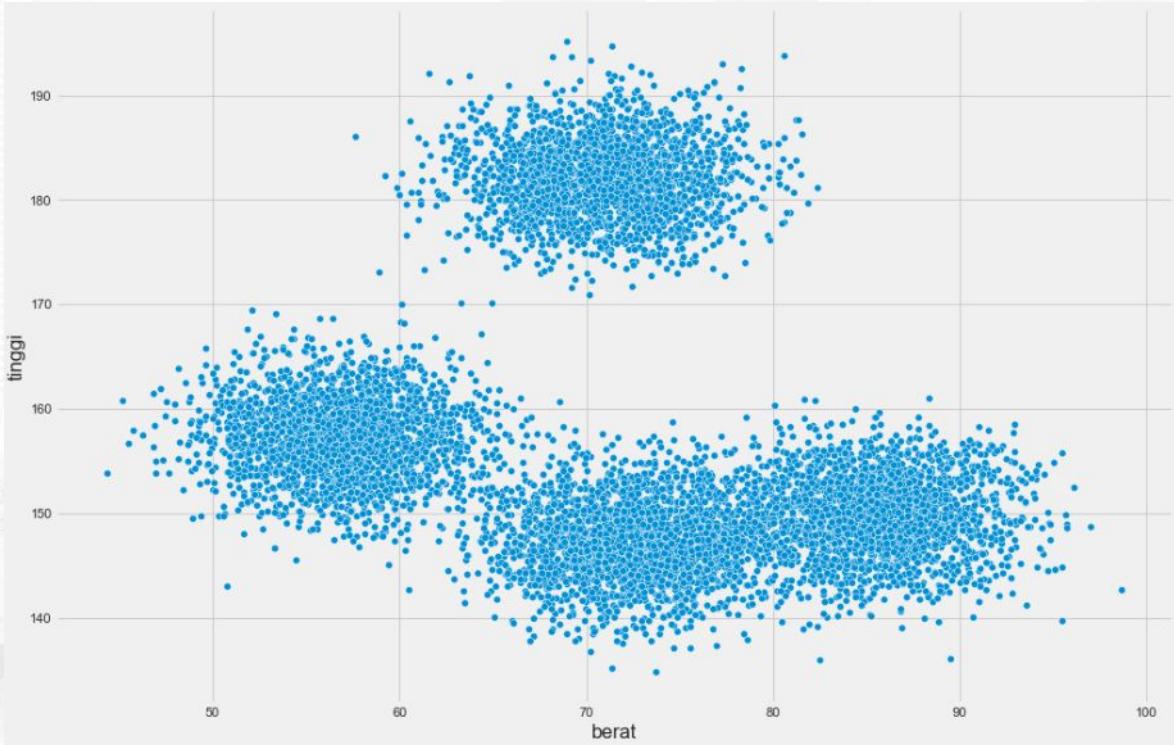
Data berisi pengukuran berat dan tinggi dari ribuan orang. Asumsikan data ini dihasilkan secara manual oleh petugas kesehatan.

- **Data:**

Setiap baris mewakili satu pengukuran, terdapat 3 kolom: 'berat', 'tinggi', dan deskripsi

- **Task:** Lakukan segmentasi dan analisis hasil segmentasi tersebut

Seperi apa datanya?
Kira-kira terdapat berapa segmen/cluster?



DATA PREPROCESSING UNTUK CLUSTERING: 2 HAL WAJIB

1. **OUTLIER TREATMENT:** Clustering adalah proses pemodelan yang cukup sensitif terhadap outlier
2. **STANDARDIZATION:** Fitur dengan skala yang lebih besar akan lebih dominan, maka skala harus sama

Standardisasi Data

```
1 feats = ['berat', 'tinggi']
2 X = df[feats].values
3
4 from sklearn.preprocessing import StandardScaler
5 X_std = StandardScaler().fit_transform(X)
6 new_df = pd.DataFrame(data = X_std, columns = feats)
7 new_df.describe()
```

| | berat | tinggi |
|-------|---------------|---------------|
| count | 8.888000e+03 | 8.888000e+03 |
| mean | -9.445390e-16 | -1.882882e-15 |
| std | 1.000056e+00 | 1.000056e+00 |

PENTING! Sebelum menggunakan algoritma apapun yang berhubungan dengan 'jarak', data harus distandardisasi.

- Bila data tidak distandardisasi, maka dimensi dengan skala data besar akan menjadi dominan
- Standardisasi dilakukan dengan menggunakan StandardScaler()

Agglomerative Clustering

```
1 from sklearn.cluster import AgglomerativeClustering  
2 ac = AgglomerativeClustering(n_clusters=4)  
3 ac.fit(new_df.values)  
  
AgglomerativeClustering(affinity='euclidean', compute_full_tree='auto',  
                      connectivity=None, distance_threshold=None,  
                      linkage='ward', memory=None, n_clusters=4)
```

Lakukan Agglomerative Clustering dengan menggunakan sklearn

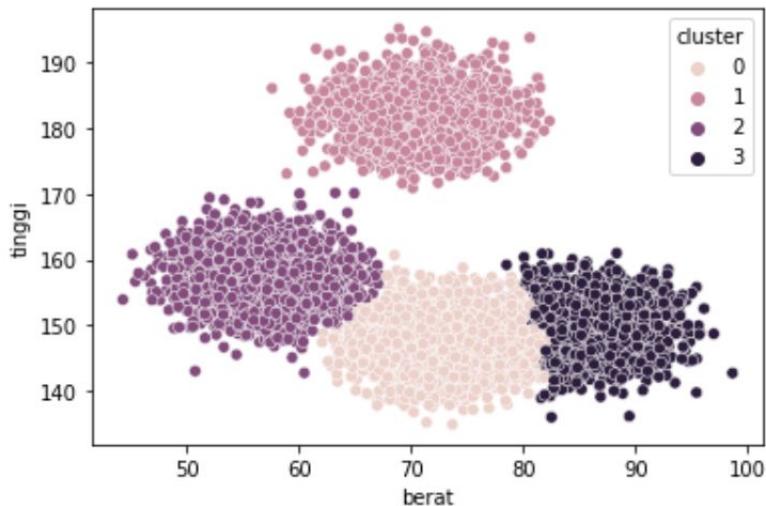
sklearn.cluster berisi banyak algoritma clustering, termasuk agglomerative clustering dengan objek AgglomerativeClustering

- n_clusters mengatur berapa cluster yang ingin kita hasilkan
- fit mengeksekusi algoritma clustering

Analisis Agglomerative Clustering

```
1 sns.scatterplot(data=df, x='berat', y='tinggi',  
2                  hue='cluster')
```

```
<AxesSubplot:xlabel='berat', ylabel='tinggi'>
```



Analisis

cluster 0 : Segment Fat

cluster 1 : Segment Slim

cluster 2 : Segment Normal

cluster 3 : Segment Obese

Statistical Summary

```
1 df['fit'] = ac.labels_
2 df.groupby('fit').agg(['mean', 'median', 'std'])
```

| | tinggi | | | berat | | |
|-----|------------|------------|----------|-----------|-----------|----------|
| | mean | median | std | mean | median | std |
| fit | | | | | | |
| 0 | 147.513296 | 147.515258 | 3.712023 | 73.180222 | 73.033571 | 4.078332 |
| 1 | 182.095791 | 182.100077 | 3.660836 | 70.845000 | 70.918470 | 3.784643 |
| 2 | 157.254633 | 157.304927 | 3.858061 | 56.902492 | 56.893684 | 3.647512 |
| 3 | 149.945901 | 150.043124 | 3.805872 | 86.212358 | 85.910828 | 3.144928 |

- Tempel cluster ke dataframe asli
- Cek statistik setelah grouping berdasarkan cluster
- Tentukan strategi bisnis berdasarkan statistik

Outline Pembelajaran

Topik Unsupervised Learning

- Sesi I: Intro + Dimensionality Reduction
- Definisi dan Jenis-jenis Unsupervised Learning
- Contoh Kasus Unsupervised Learning
- Dimensionality Reduction dan Penggunaannya
- Intuisi dan Motivasi Principal Component Analysis (PCA)
- PCA (Praktik)
- Algoritma PCA Langkah-demi-langkah

- Sesi II : Clustering**
- Intuisi dan Motivasi Clustering**
- Clustering dan Segmentasi dalam Bisnis**
- Intermezzo: Pengukuran Jarak**
- Algoritma Agglomerative Clustering dan Praktik**
- Algoritma K-means Clustering dan Praktik**
- Evaluasi Clustering**

Algoritma **K-means Clustering**

Outline Pembelajaran

Topik Unsupervised Learning

- Sesi I: Intro + Dimensionality Reduction
- Definisi dan Jenis-jenis Unsupervised Learning
- Contoh Kasus Unsupervised Learning
- Dimensionality Reduction dan Penggunaannya
- Intuisi dan Motivasi Principal Component Analysis (PCA)
- PCA (Praktik)
- Algoritma PCA Langkah-demi-langkah

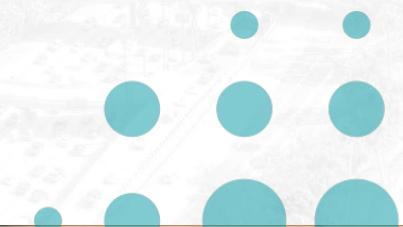
Sesi II : Clustering

- Intuisi dan Motivasi Clustering**
- Clustering dan Segmentasi dalam Bisnis**
- Intermezzo: Pengukuran Jarak**
- Algoritma Agglomerative Clustering dan Praktik**
-  **Algoritma K-means Clustering dan Praktik**
- Evaluasi Clustering**

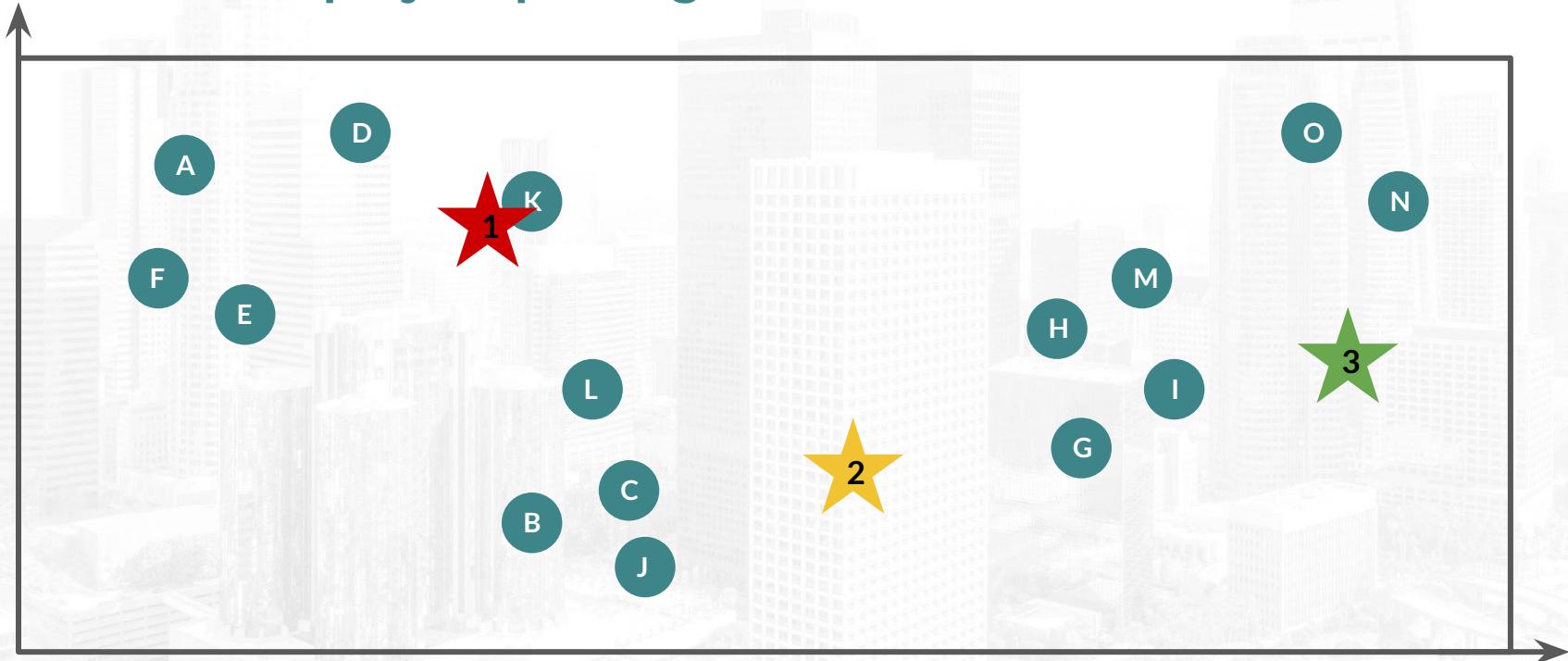
Prinsip Utama K-Means

“Temukan pusat cluster yang meminimalisir total jarak setiap titik ke pusatnya.”

K-means Step-by-step: Berapa cluster?

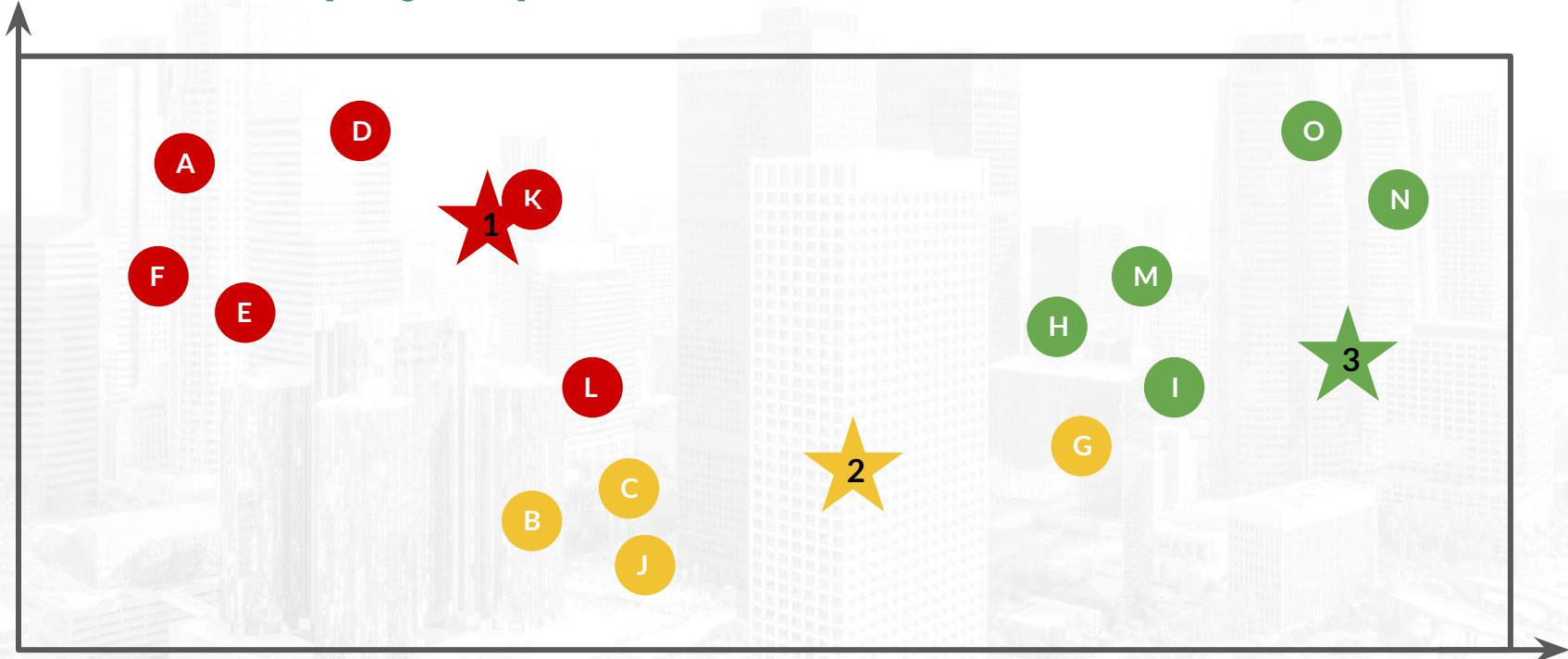


K-means Step-by-step: Mengenal Centroid



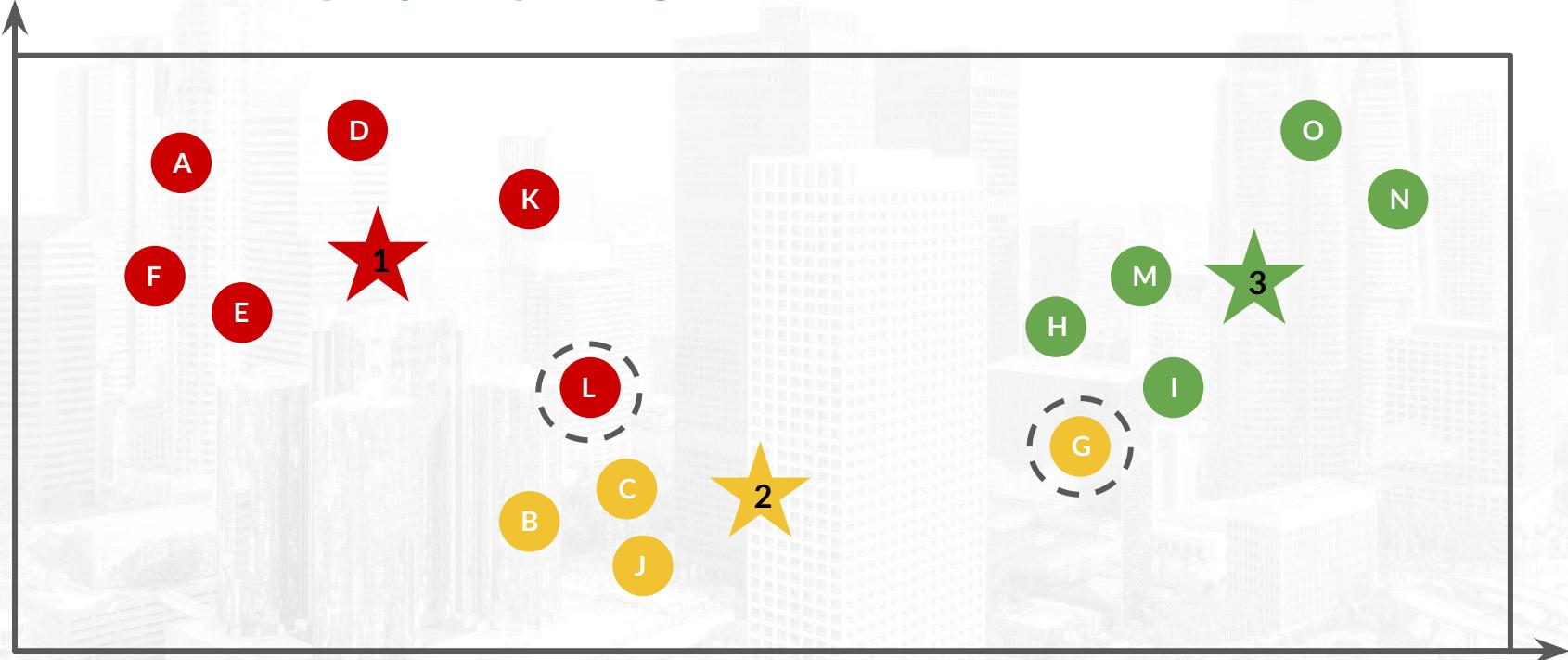
- Secara acak, tempatkan CENTROID sejumlah banyak cluster yang diinginkan
- Centroid ini akan menjadi titik pusat setiap cluster yang dihasilkan

K-means Step-by-step: Cluster Sementara



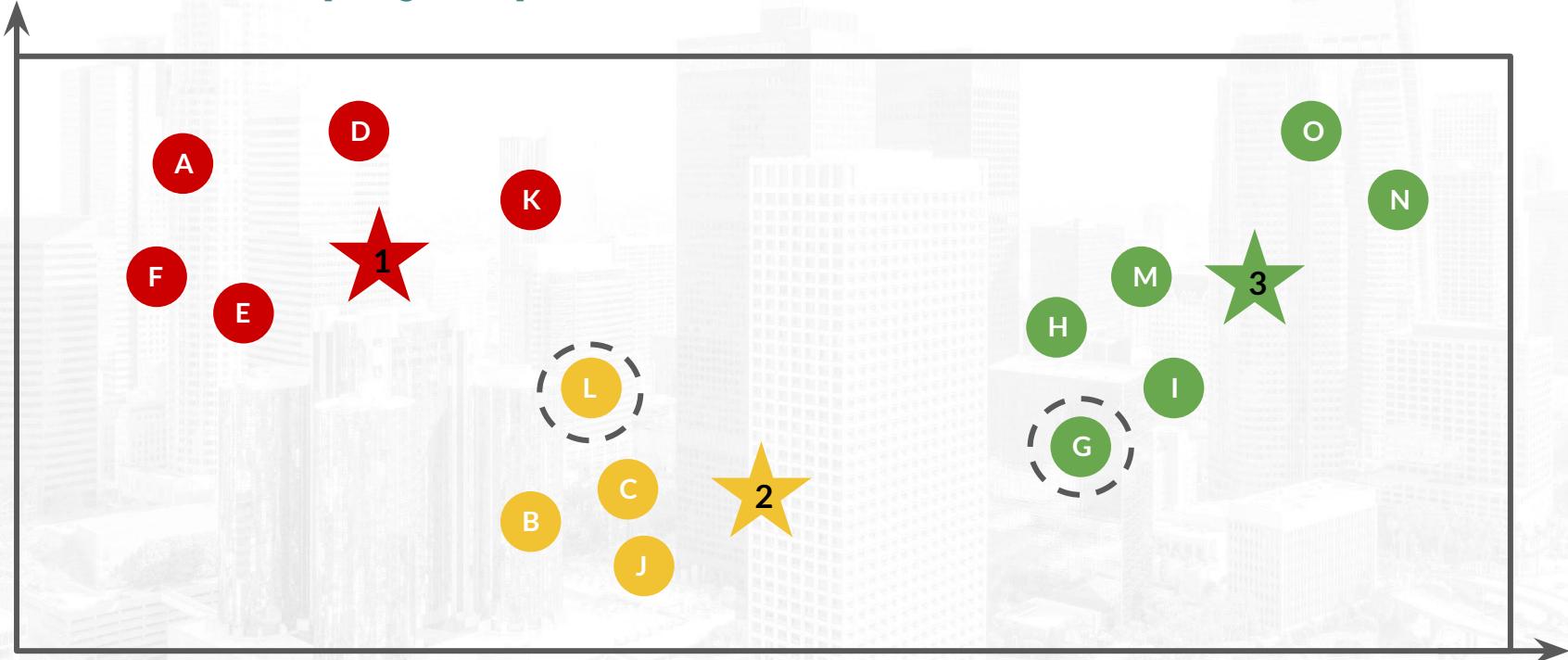
- Pilihkan cluster untuk setiap data yang kita miliki
- Cluster satu data point = centroid nya paling dekat dibanding centroid lain

K-means Step-by-step: Pergeseran Centroid



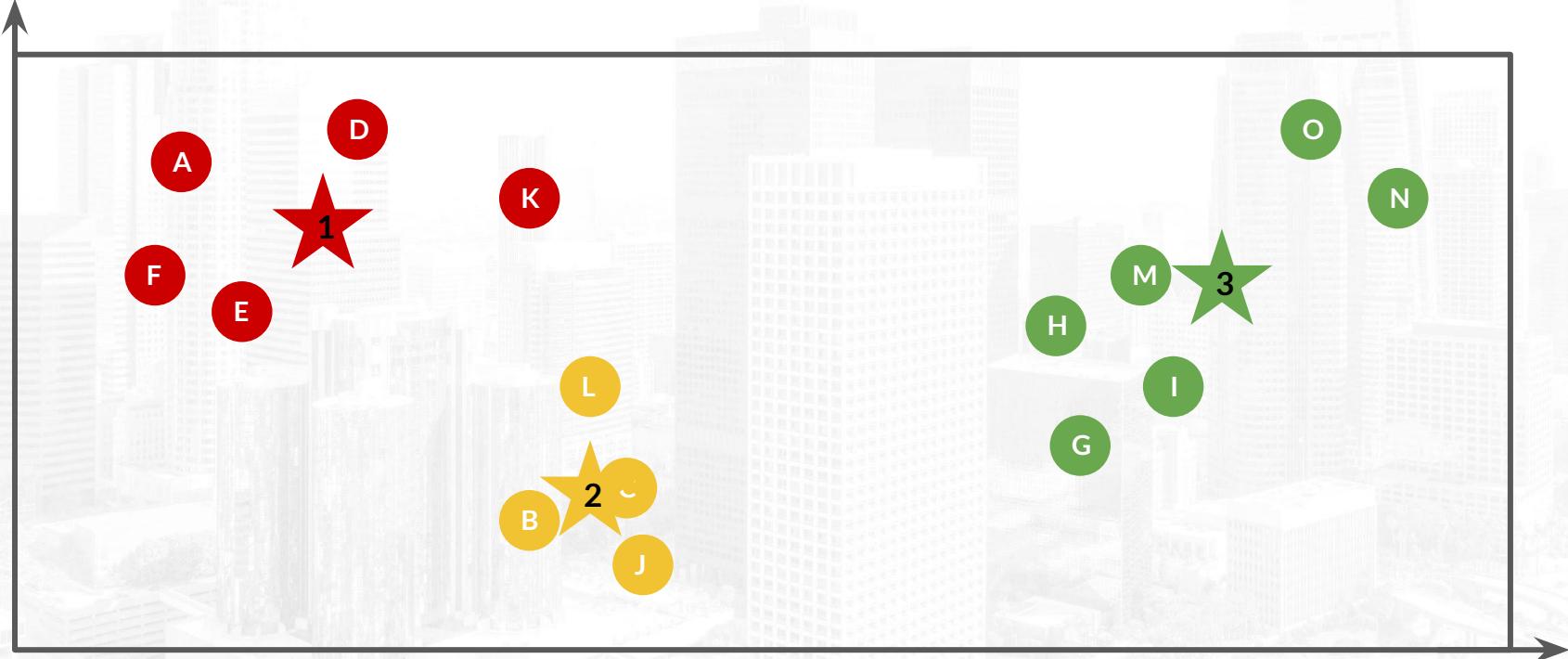
- Setelah kita mendapatkan cluster untuk setiap data, kita geser centroid
- Posisi baru centroid = rata-rata posisi anggota clusternya
- Setelah ini prosedur tadi kita ulangi lagi

K-means Step-by-step: Cluster Baru



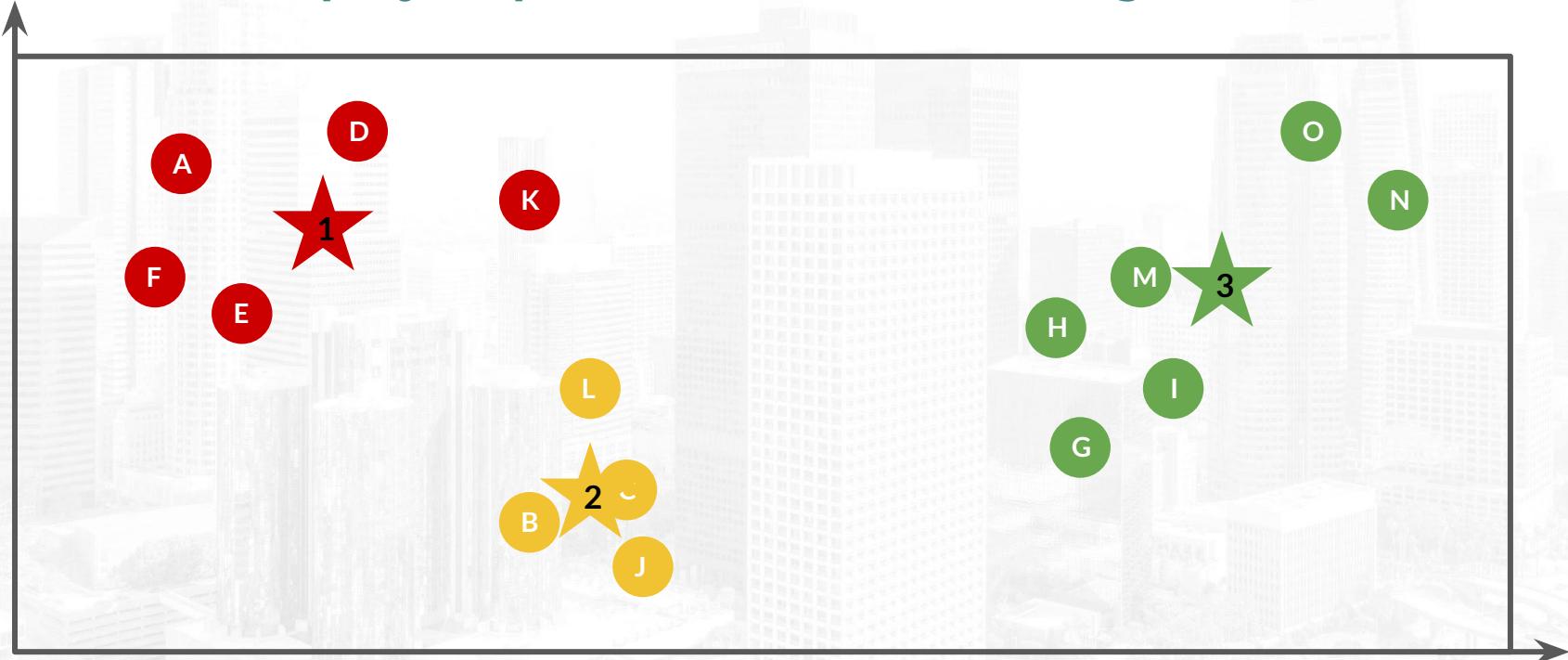
- Pilihkan cluster untuk setiap data yang kita miliki
- Cluster satu data point = centroid nya paling dekat dibanding centroid lain
- Perhatikan ada data point yang clusternya berubah setelah digeser

K-means Step-by-step: Pergerakan Centroid (2)



- Setelah kita mendapatkan cluster BARU untuk setiap data, kita geser centroid
- Posisi baru centroid = rata-rata posisi anggota clusternya
- Setelah ini prosedur tadi kita ulangi lagi

K-means Step-by-step: Cluster Baru??? Konvergensi



- Setelah beberapa kali melakukan proses pergeseran ini, akan tiba kondisi dimana tidak ada data point yang berganti cluster lagi
- Algoritma K-means telah selesai!

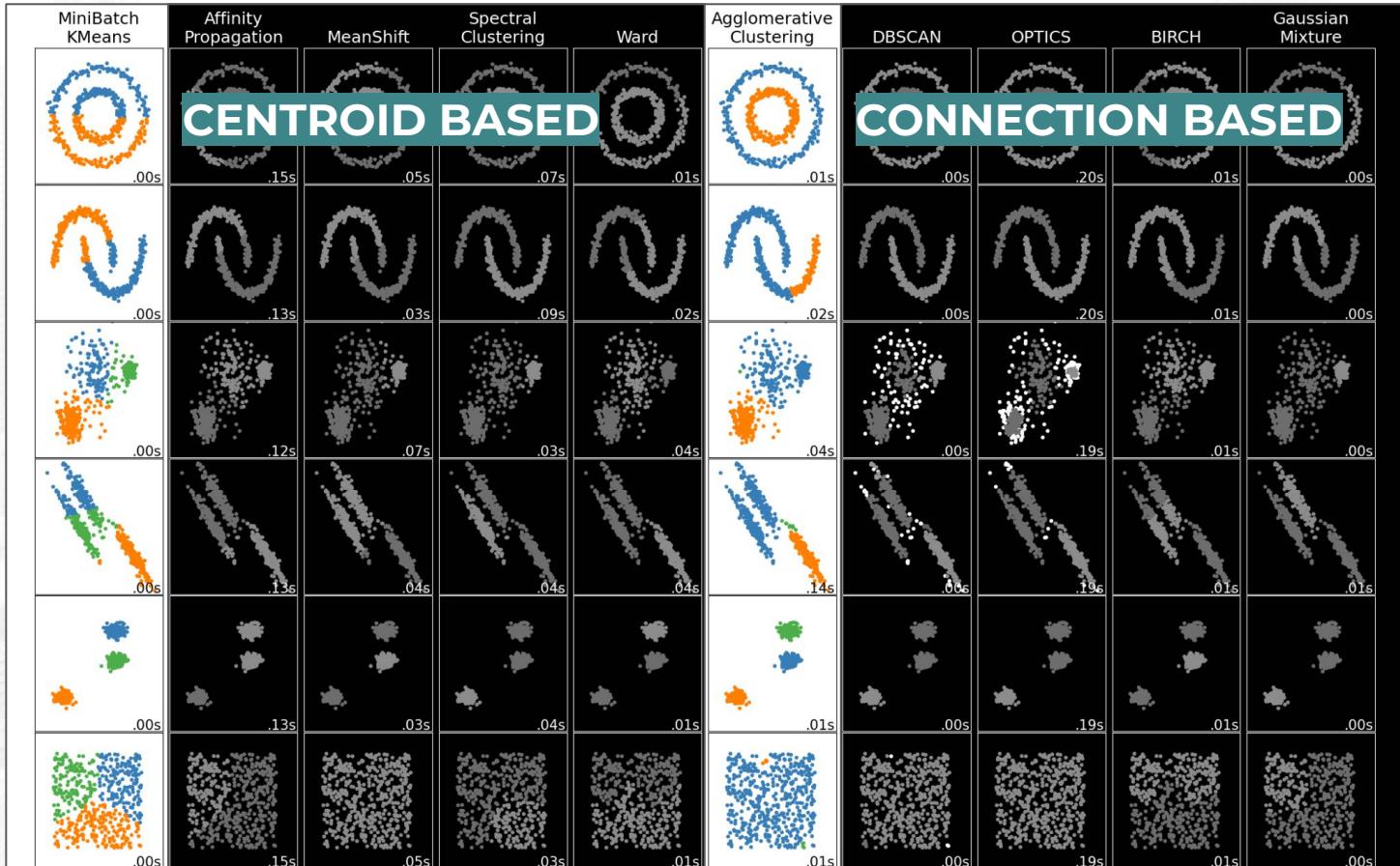
Algoritma Clustering

Agglomerative Clustering

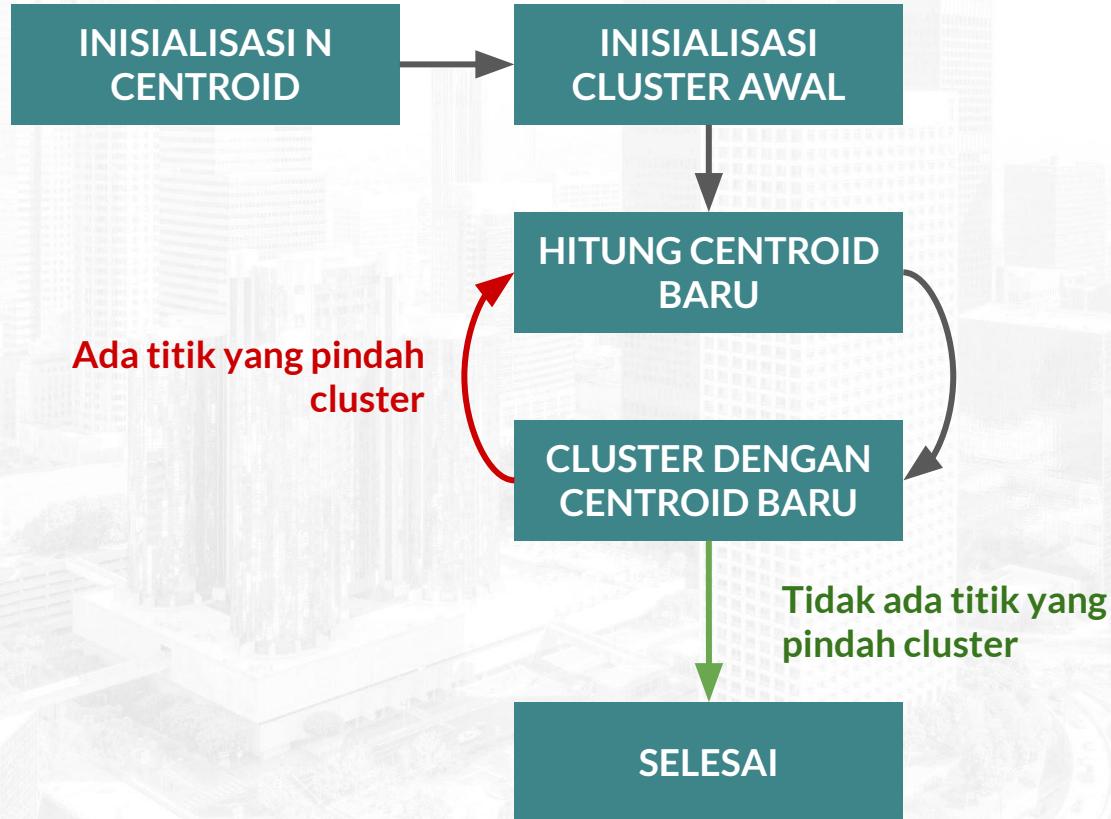
- Mengelompokkan dua data yang paling dekat
- Deterministik (pasti)

K-means Clustering

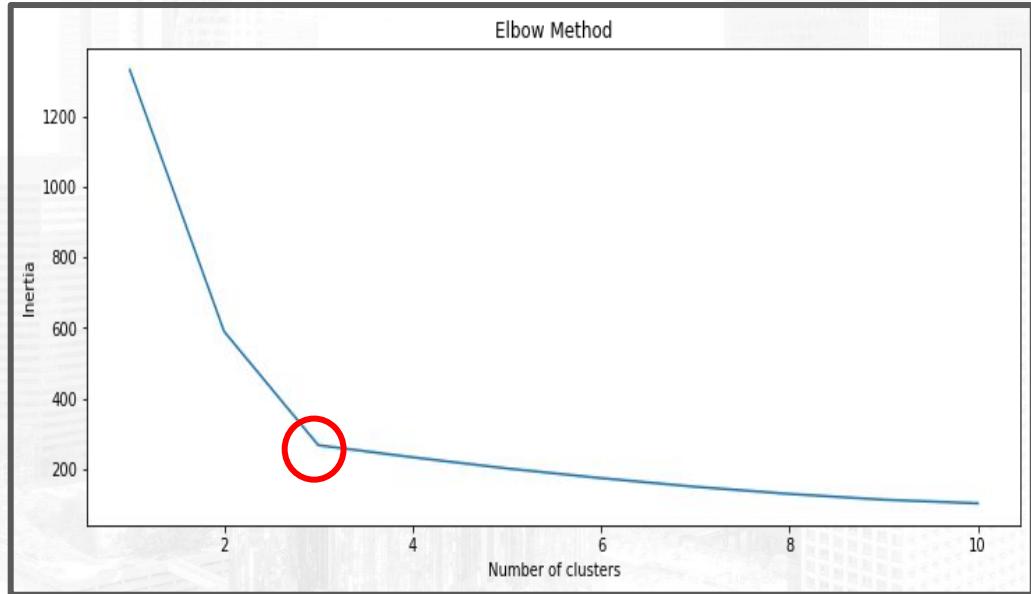
- Mencari centroid yang stabil
- Non-deterministik (acak)



K-means Clustering Overview



Evaluasi Internal K-means: Elbow Method untuk mencari jumlah cluster optimal



INERTIA:

Total jarak setiap titik
ke pusatnya

- Lakukan K-means untuk 1-N cluster lalu plot inertia akhir
- Jumlah cluster optimal -> N dimana perbedaan inertia dengan menambah cluster baru berkurang drastis

Implementasi **K-means Clustering**

K-Means

```
1 from sklearn.cluster import KMeans  
2 kmeans = KMeans(n_clusters=4, init='k-means++', max_iter=300, n_init=10, random_state=0)  
3 kmeans.fit(new_df.values)  
  
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,  
       n_clusters=4, n_init=10, n_jobs=None, precompute_distances='auto',  
       random_state=0, tol=0.0001, verbose=0)
```

Lakukan K-means dengan menggunakan sklearn

`sklearn.cluster` berisi banyak algoritma clustering, termasuk agglomerative clustering dengan objek `KMeans`

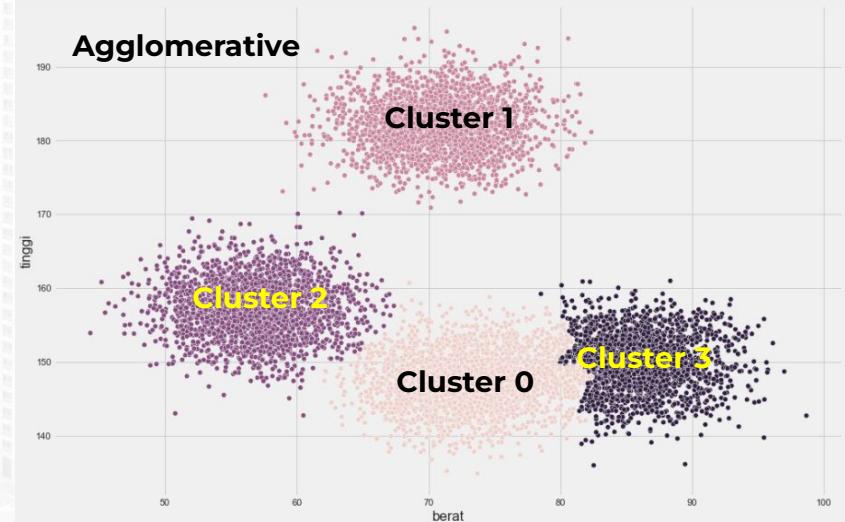
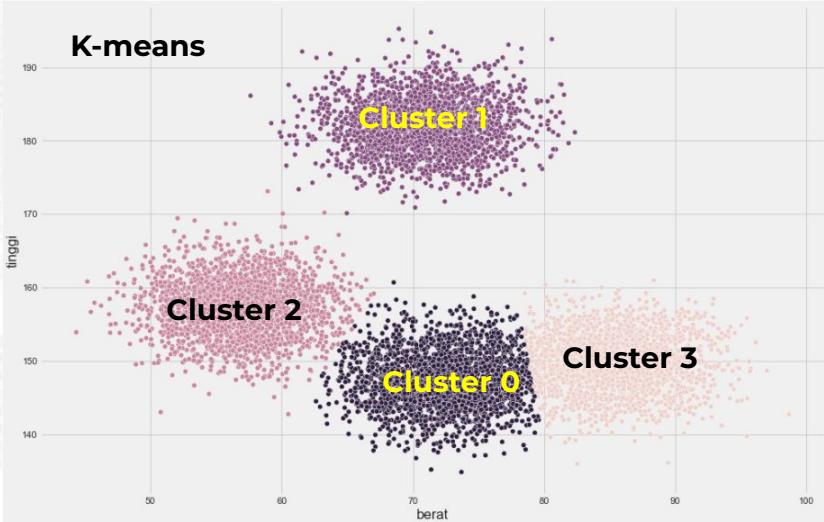
- `n_clusters` mengatur berapa cluster yang ingin kita hasilkan
- `init` mengatur strategi penempatan awal centroid, secara default valuenya adalah '`k-means++`' dan disarankan menggunakan '`k-means++`' karena cara cerdas untuk mempercepat konvergensi
- `max_iter` mengatur jumlah pengulangan prosedur
- `fit` mengeksekusi algoritma clustering

<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

<https://scikit-learn.org/stable/modules/clustering.html>

Membandingkan hasil K-means dengan Agglomerative

Terdapat hasil yang hampir mirip antara pembentukan cluster menggunakan K-means dengan pembentukan cluster menggunakan Agglomerative



Membandingkan Statistical Summary

| | tinggi | | | berat | | |
|------------|------------|------------|----------|-----------|-----------|----------|
| | mean | median | std | mean | median | std |
| fit | | | | | | |
| 0 | 147.513296 | 147.515258 | 3.712023 | 73.180222 | 73.033571 | 4.078332 |
| 1 | 182.095791 | 182.100077 | 3.660836 | 70.845000 | 70.918470 | 3.784643 |
| 2 | 157.254633 | 157.304927 | 3.858061 | 56.902492 | 56.893684 | 3.647512 |
| 3 | 149.945901 | 150.043124 | 3.805872 | 86.212358 | 85.910828 | 3.144928 |

Agglomerative Clustering

| | tinggi | | | berat | | |
|------------|------------|------------|----------|-----------|-----------|----------|
| | mean | median | std | mean | median | std |
| fit | | | | | | |
| 0 | 157.257620 | 157.313528 | 3.882061 | 56.853552 | 56.868609 | 3.589815 |
| 1 | 147.362581 | 147.326452 | 3.752352 | 72.351242 | 72.513017 | 3.494274 |
| 2 | 182.094451 | 182.100077 | 3.664672 | 70.847712 | 70.918470 | 3.778257 |
| 3 | 149.856603 | 149.891137 | 3.748861 | 85.499783 | 85.301109 | 3.553060 |

K-means Clustering

Elbow Method

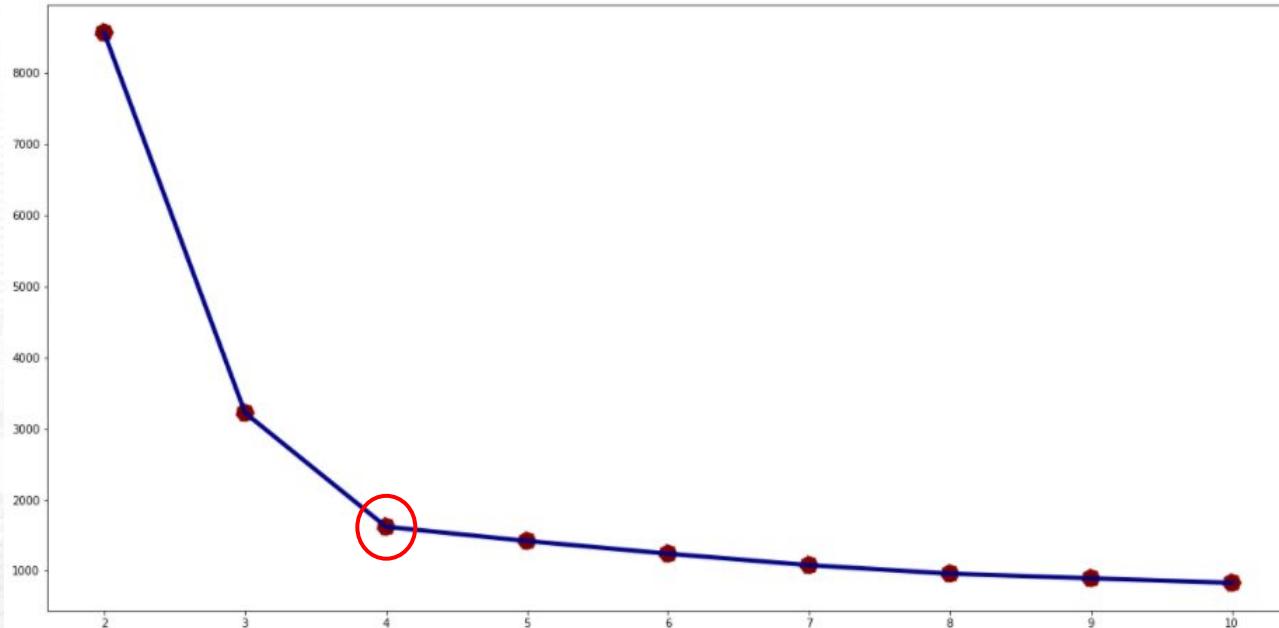
```
1 from sklearn.cluster import KMeans
2 inertia = []
3
4 for i in range(1, 11):
5     kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=0)
6     kmeans.fit(new_df.values)
7     inertia.append(kmeans.inertia_)
8
9 plt.figure(figsize=(12, 6))
10 plt.plot(inertia)
```

Inersia akhir model K-means dapat diambil dari objek `KMeans`

Untuk melihat elbow plot kita cukup mengulangi K-means untuk jumlah cluster 1-10

- Ambil inersia dengan `.inertia_`

Elbow Method



Elbow Method



Outline Pembelajaran

Topik Unsupervised Learning

- Sesi I: Intro + Dimensionality Reduction
- Definisi dan Jenis-jenis Unsupervised Learning
- Contoh Kasus Unsupervised Learning
- Dimensionality Reduction dan Penggunaannya
- Intuisi dan Motivasi Principal Component Analysis (PCA)
- PCA (Praktik)
- Algoritma PCA Langkah-demi-langkah

- Sesi II : Clustering**
- Intuisi dan Motivasi Clustering**
- Clustering dan Segmentasi dalam Bisnis**
- Intermezzo: Pengukuran Jarak**
- Algoritma Agglomerative Clustering dan Praktik**
- Algoritma K-means Clustering dan Praktik**
- Evaluasi Clustering**

Evaluasi Hasil Clustering

Outline Pembelajaran

Topik Unsupervised Learning

- Sesi I: Intro + Dimensionality Reduction
- Definisi dan Jenis-jenis Unsupervised Learning
- Contoh Kasus Unsupervised Learning
- Dimensionality Reduction dan Penggunaannya
- Intuisi dan Motivasi Principal Component Analysis (PCA)
- PCA (Praktik)
- Algoritma PCA Langkah-demi-langkah

Sesi II : Clustering

- Intuisi dan Motivasi Clustering**
- Clustering dan Segmentasi dalam Bisnis**
- Intermezzo: Pengukuran Jarak**
- Algoritma Agglomerative Clustering dan Praktik**
- Algoritma K-means Clustering dan Praktik**
-  **Evaluasi Clustering**

External vs Internal Evaluation

External Evaluation

Evaluasi atas cluster yang dihasilkan dilakukan dengan menggunakan *ground truth*, label/cluster sesungguhnya yang kita punya. Contoh teknik external evaluation:

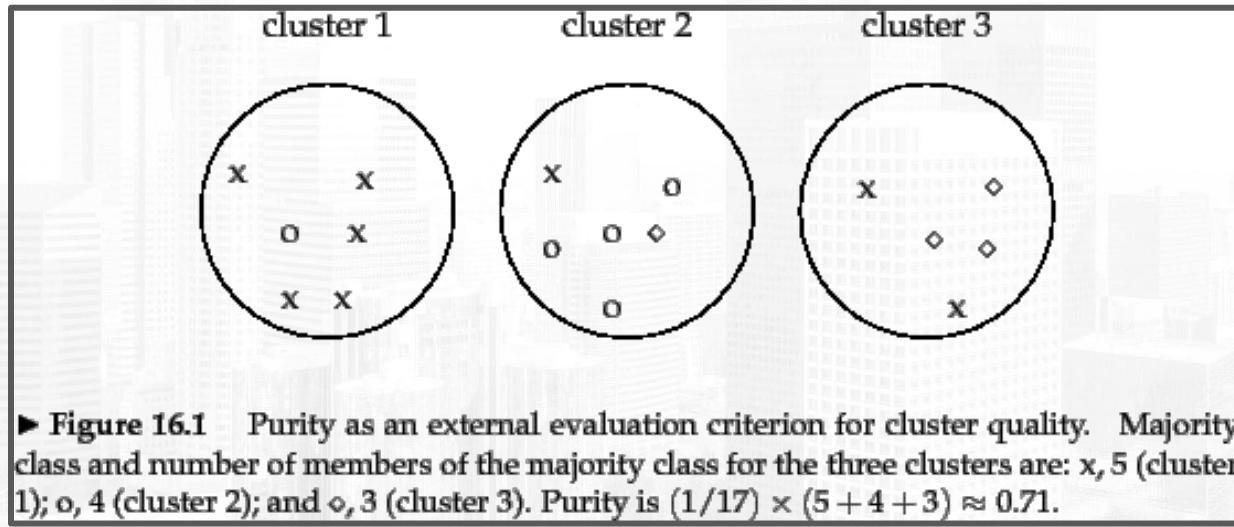
- Rand Index
- Normalized Mutual Information

Internal Evaluation

Karena tidak ada *ground truth*, kita hanya mengevaluasi seberapa mirip anggota masing-masing cluster dan apakah anggota setiap cluster paling dekat dengan clusternya dibanding cluster lain. Contoh teknik internal evaluation:

- Inertia
- Silhouette Score

External Evaluation: Purity



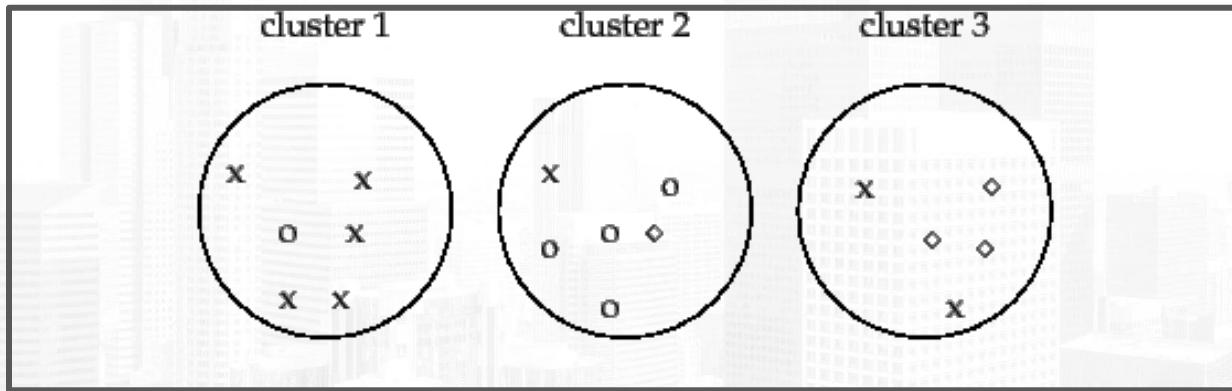
$$\text{purity}(\Omega, C) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

Purity memberikan gambaran seberapa ‘jernih’ isi setiap cluster

Purity:

- Asumsikan kelas terbanyak di setiap cluster adalah ‘benar’
- Hitung jumlah ‘benar’ dibagi jumlah baris data

External Evaluation: Rand Index



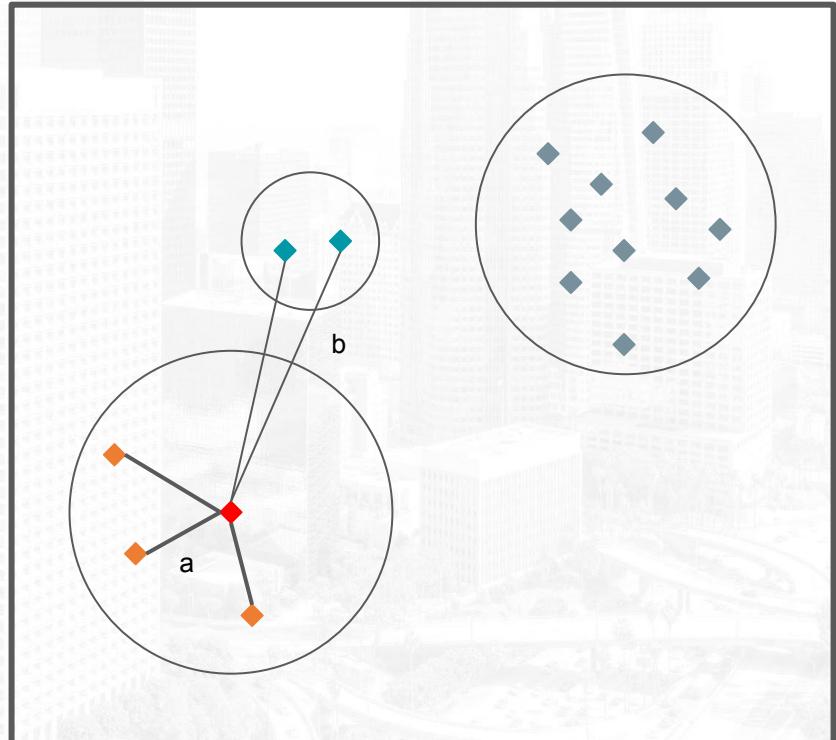
$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

Rand Index mendeskripsikan akurasi clustering berdasarkan seluruh kombinasi pasangan titik di cluster yang dihasilkan

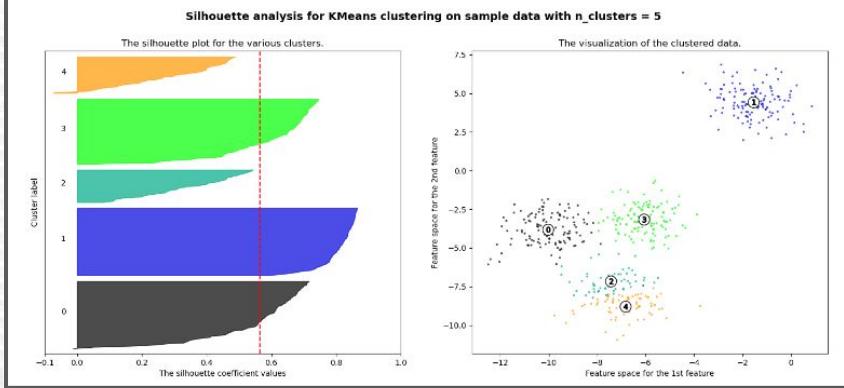
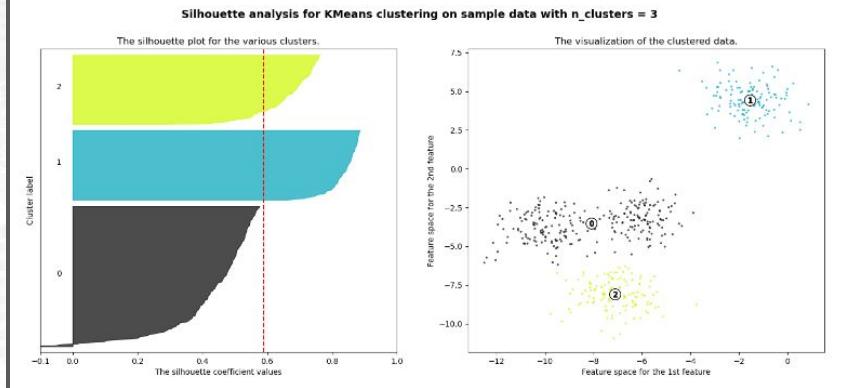
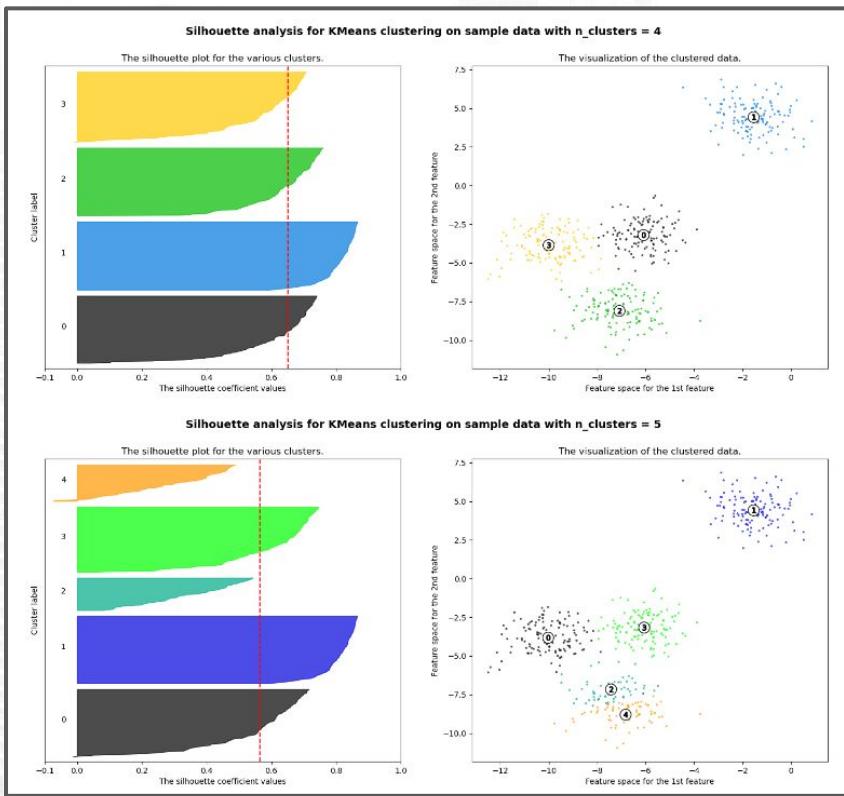
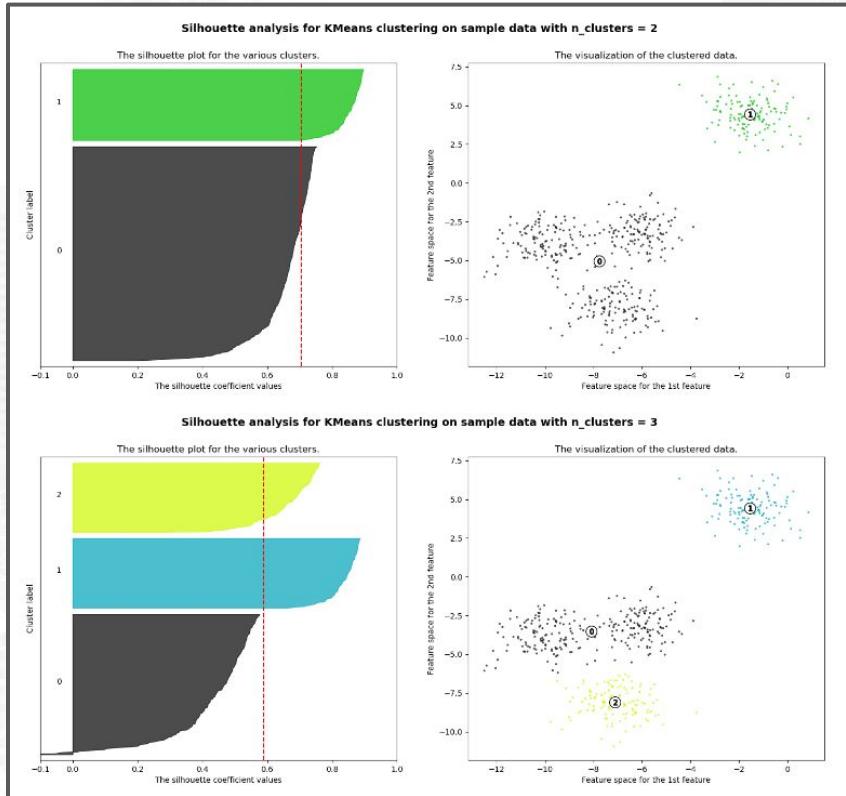
- TP: 2 titik yang satu cluster di *ground truth* dan satu cluster di cluster hasil
- TN: 2 titik yang tidak satu cluster di *ground truth* dan tidak satu cluster di cluster hasil
- FP: 2 titik yang tidak satu cluster di *ground truth* namun satu cluster di cluster hasil
- FN: 2 titik yang satu cluster di *ground truth* namun tidak satu cluster di cluster hasil

Internal Evaluation: Silhouette Score

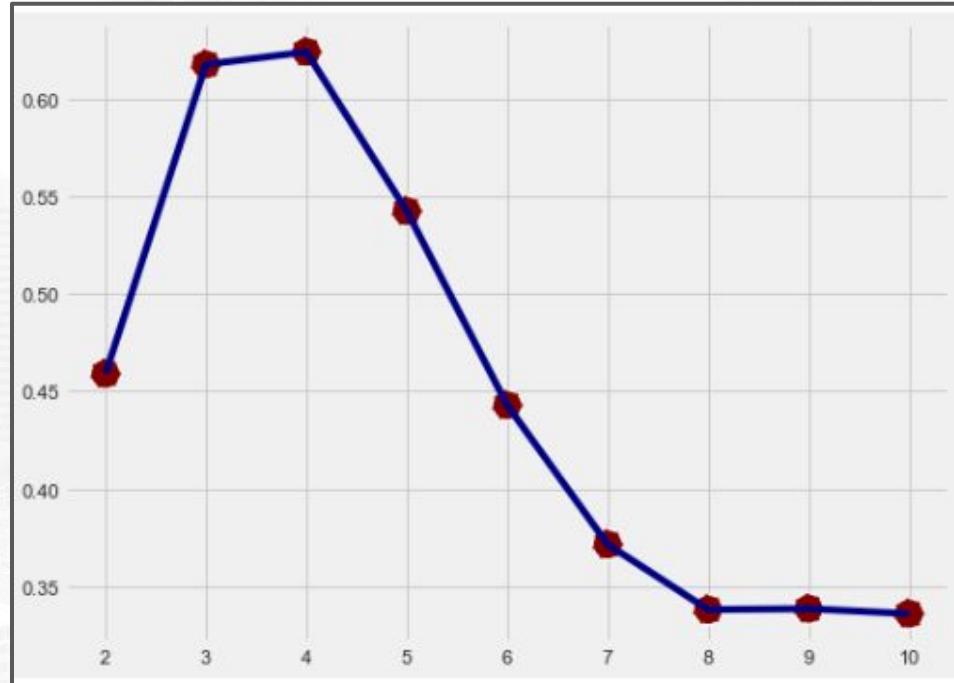
- Rasio antara
 - Perbedaan jarak rata-rata **intra-cluster (a)** dan jarak rata-rata **ekstra-cluster (b)** (ke cluster terdekat)
 - Dan nilai maksimal kedua jarak rata-rata tersebut
- Silhouette Score:
 - **Range antara -1 dan 1**
 - 1 berarti jarak rata-rata intra cluster = 0 dan ekstra-cluster > 0



Internal Evaluation: Silhouette Analysis



Internal Evaluation: Silhouette Analysis



Outline Pembelajaran

Topik Unsupervised Learning

- Sesi I: Intro + Dimensionality Reduction
- Definisi dan Jenis-jenis Unsupervised Learning
- Contoh Kasus Unsupervised Learning
- Dimensionality Reduction dan Penggunaannya
- Intuisi dan Motivasi Principal Component Analysis (PCA)
- PCA (Praktik)
- Algoritma PCA Langkah-demi-langkah

- Sesi II : Clustering**
- Intuisi dan Motivasi Clustering**
- Clustering dan Segmentasi dalam Bisnis**
- Intermezzo: Pengukuran Jarak**
- Algoritma Agglomerative Clustering dan Praktik**
- Algoritma K-means Clustering dan Praktik**
- Evaluasi Clustering**



Hafizh Adi Prasetya

Data Scientist

Bukalapak

Terima Kasih!



Hafizh Adi Prasetya

<https://id.linkedin.com/in/hafizhadi>