DOLab

live class

8 - Clustering

DQLab

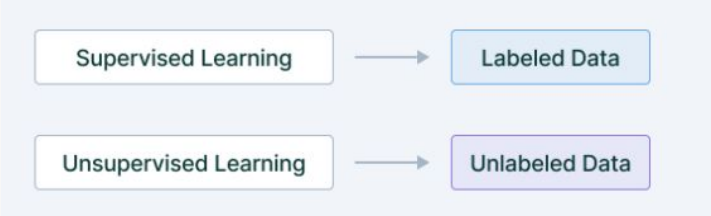
Live Class

Outline

1. Unsupervised Learning
2. Pengenalan Clustering
3. Tipe Cluster
4. Metode Clustering
5. K-means Clustering
6. Hierarchical Clustering
7. Evaluasi Kinerja Clustering
8. Penentuan Jumlah Optimal Cluster (k)
9. Hands-on

Unsupervised Learning

Unsupervised Learning

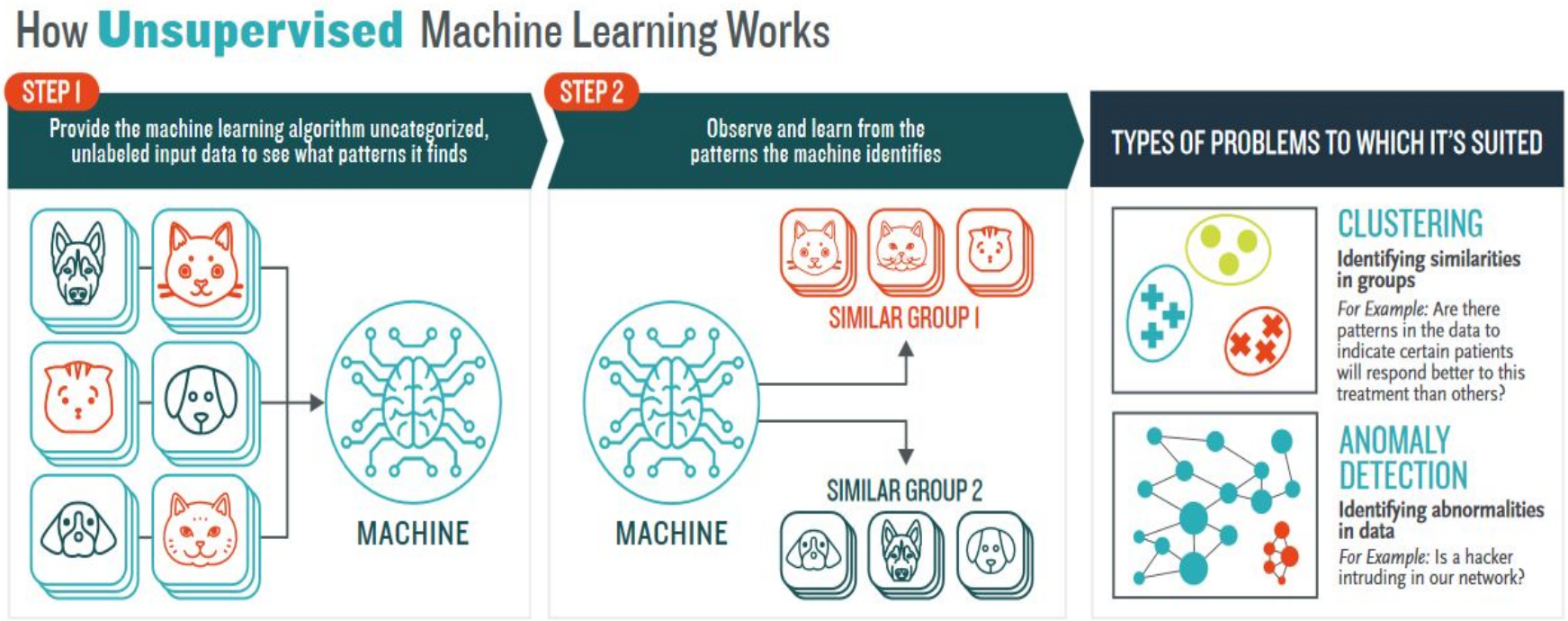


Unsupervised learning adalah pendekatan *machine learning* di mana algoritma belajar **tanpa diberikan label** atau jawaban yang diinginkan pada data yang diberikan.

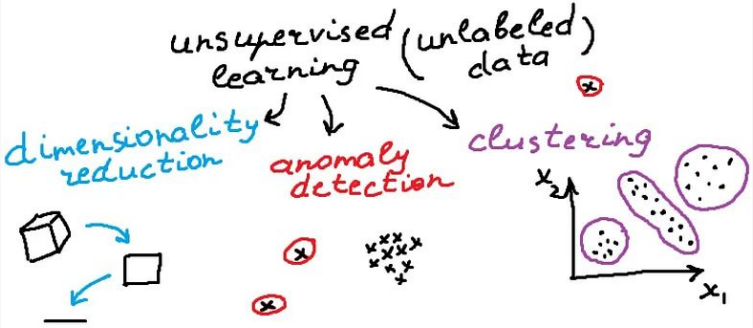
Algoritma *unsupervised learning* mencoba menemukan pola dan struktur yang tersembunyi dalam data tersebut secara mandiri, tanpa bantuan dari manusia (*unsupervised*).

Image Source

UNSUPERVISED LEARNING



Kapan *Unsupervised Learning* Digunakan?



Unsupervised learning sering digunakan dalam situasi di mana tidak ada informasi label yang tersedia untuk pelatihan model atau informasi label tidak cukup banyak atau terlalu mahal untuk diperoleh.

Beberapa contoh penggunaan unsupervised learning antara lain: **pengelompokkan data (*clustering*)**, **reduksi dimensi (*dimensionality reduction*)**, **deteksi anomali (*anomaly detection*)**, dan **pencarian asosiasi (*association rule mining*)**.

Image Source

The Advantages

Growing importance in a number of fields

- subgroups of breast cancer patients grouped by their gene expression measurements
- groups of shoppers characterised by their browsing and purchase histories
- movies grouped by the ratings assigned by movie viewers
- topic modelling of text document (NLP)

Easier to obtain unlabeled data than labelled data

The Challenges

more subjective than supervised learning

- No simple goal for analysis
- The computer have to learn how to do something that we don't tell it how to do

Have some issues

- The number of subgroups (clusters)
- The different results via K-means with different random initialisations
- How to assess the performance of the unsupervised learning methods?

The learning (or inference) procedure is hard

Clustering

Clustering adalah teknik pengelompokkan objek-objek atau data menjadi beberapa kelompok berdasarkan kesamaan karakteristik di antara objek-objek tersebut.

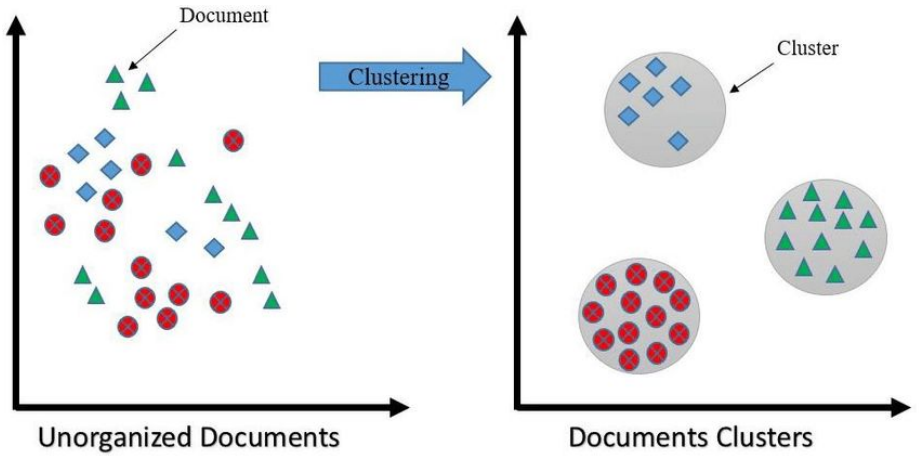


Image Source

Apa itu *Clustering*?

- *Cluster* adalah kelompok atau kumpulan dari data yang memiliki karakteristik atau atribut yang serupa atau mirip satu sama lain, dan memiliki perbedaan dengan entitas dari kelompok lain.
- Tujuan dari analisis *clustering* adalah untuk menemukan pola dan struktur dalam data dan mengidentifikasi kelompok-kelompok atau klaster-klaster yang signifikan.
- Semakin besar tingkat kemiripan di dalam satu grup (*intra-cluster*) dan semakin besar tingkat perbedaan di antara grup (*inter-cluster*), maka semakin baik *clustering* tersebut.
- Penentuan cluster terbaik tergantung dari kondisi data serta hasil yang diinginkan seperti apa.

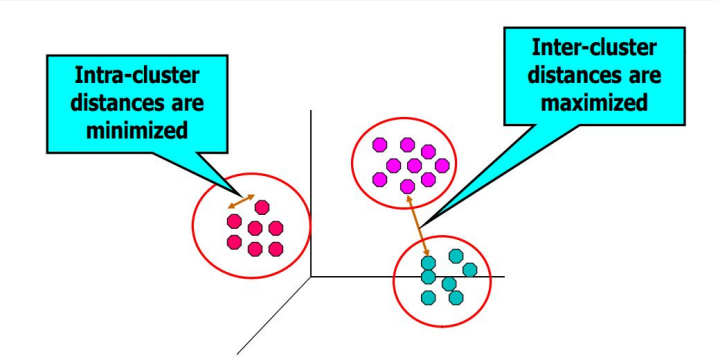


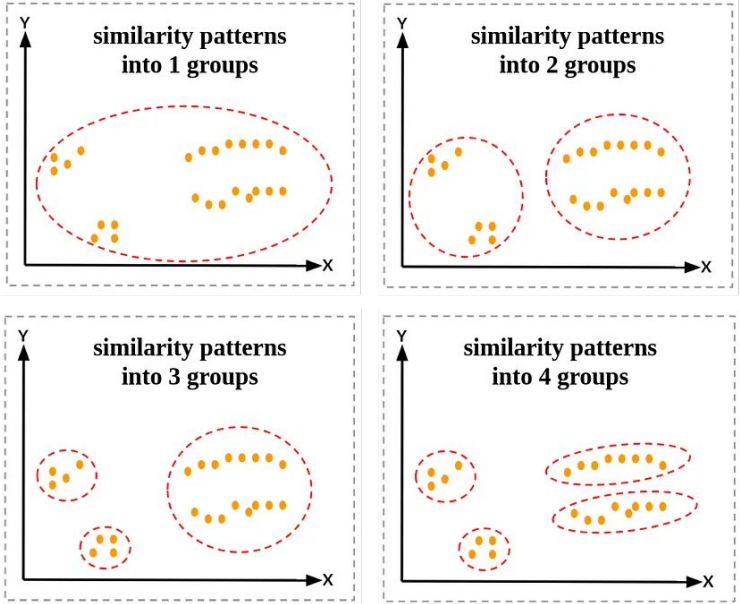
Image Source

Clustering

Contoh penggunaan *clustering*:

- **Pemasaran:** Pengelompokkan pelanggan berdasarkan preferensi mereka.
- **Segmentasi pasar:** Pengelompokkan konsumen berdasarkan karakteristik demografi, seperti usia, pendapatan, atau pendidikan.
- **Segmentasi pasien:** Pengelompokkan pasien berdasarkan profil kesehatan mereka, seperti riwayat penyakit, obat-obatan yang dikonsumsi, dan faktor risiko kesehatan.
- **Sistem rekomendasi:** Sistem yang merekomendasikan item-item yang serupa dalam kelompok yang sama kepada pengguna yang memiliki preferensi yang serupa.

Clustering Intuition



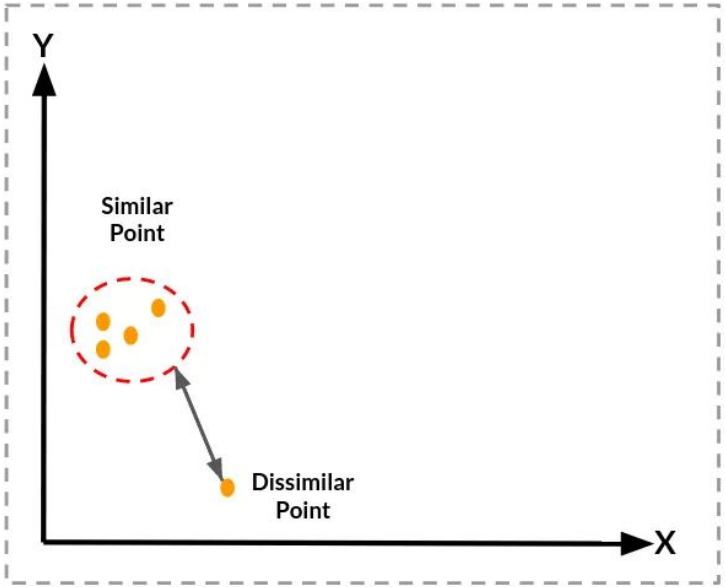
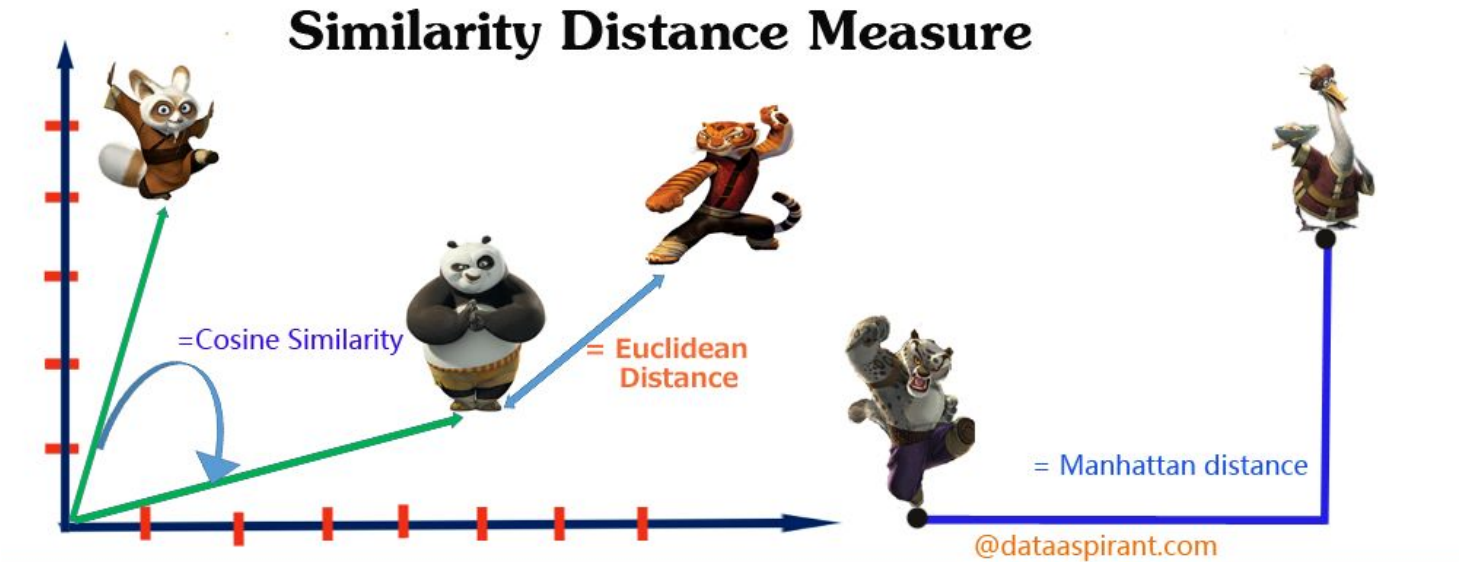


Image Source

Clustering Intuition



Kemiripan dua dengan data lainnya bisa diukur dengan ukuran *distance* (jarak) ataupun *similarity* (kemiripan). Beberapa ukuran yang dapat digunakan adalah *Euclidean distance*, *Manhattan distance*, dan *Cosine similarity*.

Image Source

Cosine Similarity

- *Cosine similarity* menghitung cosinus dari sudut antara dua vektor.
- Nilai *cosine similarity* 1 menunjukkan bahwa kedua vektor sepenuhnya identik, sedangkan nilai *cosine similarity* 0 menunjukkan bahwa kedua vektor sepenuhnya berbeda arah.
- Semakin besar nilai *cosine similarity*, semakin mirip kedua vektor tersebut.

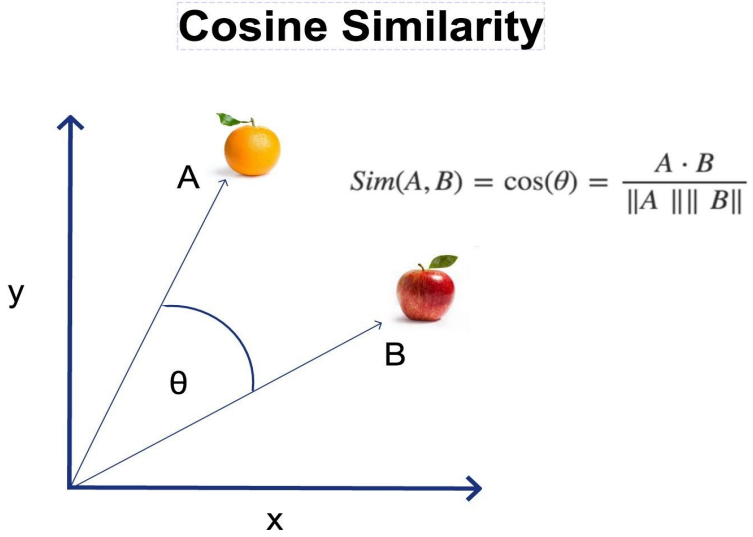
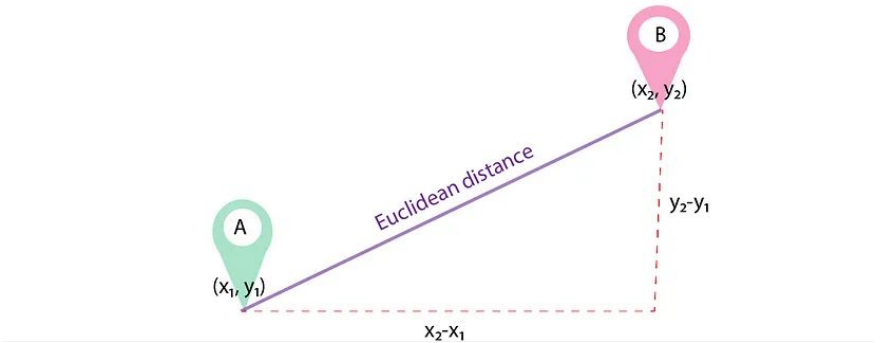


Image Source

Euclidean Distance

$$\text{dist}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

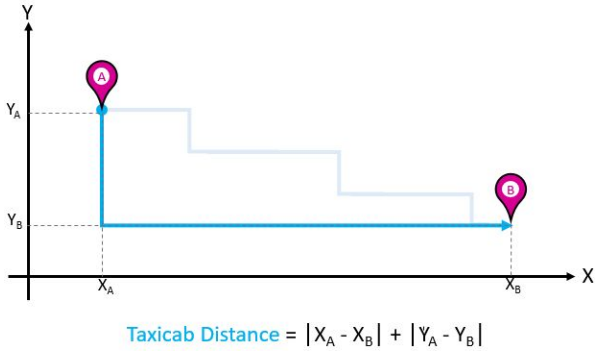


Euclidean distance adalah perhitungan jarak dari 2 buah titik dalam *Euclidean space*. *Euclidean distance* mewakili jarak terpendek antara dua titik.

Image Source

Manhattan Distance

$$d(x, y) = \sum_{i=1}^{k} |x_i - y_i|$$



Jarak Manhattan (*Manhattan distance*) antara dua titik adalah jumlah dari panjang ruas garis kedua titik tersebut terhadap tiap sumbu dalam koordinat Kartesius.

Image Source

Tipe Cluster

Well-separated Clusters

- Tipe *cluster* di mana setiap *cluster* terpisah dengan jelas dan tidak tumpang tindih.
- Setiap titik data hanya termasuk dalam satu *cluster*, dan tidak ada titik data yang berada di antara dua atau lebih *cluster*.
- Tipe *cluster* ini biasanya digunakan dalam data yang sangat terstruktur dan mudah dibedakan, seperti dalam data geospasial.

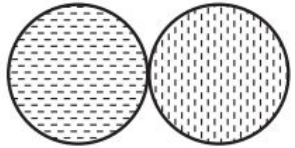
(a) Well-separated clusters. Each point is closer to all of the points in its cluster than to any point in another cluster.

Image Source

DQLab

Center-based Clusters

- Tipe *cluster* di mana setiap *cluster* memiliki pusat yang mewakili kelompok itu.
- Titik-titik data dalam suatu *cluster* lebih dekat dengan pusat *cluster*-nya dibandingkan dengan pusat *cluster* lainnya.
- Jenis pusat *cluster*: ***centroid*** (rata-rata dari semua titik dalam *cluster*) dan ***medoid*** (titik yang paling dekat dengan semua titik dalam *cluster*).



(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.

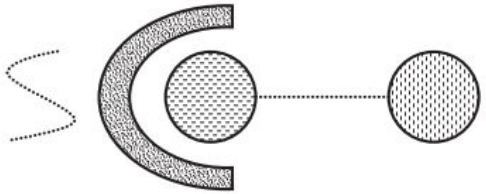
Image Source

DQLab

Live Class

Contiguity-based Clusters

- Tipe *cluster* di mana setiap *cluster* dibentuk berdasarkan kedekatan atau keterhubungan antara titik-titik data.
- Setiap titik data berada lebih dekat dengan beberapa objek lain dalam *cluster*-nya daripada ke titik mana pun dalam *cluster* lain (*nearest neighbor*).
- Tipe *cluster* ini biasanya digunakan untuk data yang tidak memiliki struktur spasial atau struktur yang tidak teratur, e.g. data sosial.

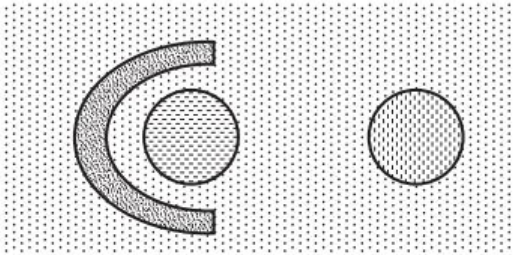


(c) Contiguity-based clusters. Each point is closer to at least one point in its cluster than to any point in another cluster.

Image Source

Density-based Clusters

- Tipe *cluster* di mana setiap *cluster* dibentuk berdasarkan kepadatan titik-titik data.
- *Cluster* dibentuk berdasarkan titik-titik data yang memiliki kepadatan yang tinggi dan diapit oleh area dengan kepadatan yang lebih rendah.
- Penentuan *cluster* dilakukan dengan mempertimbangkan jarak antara titik-titik data dan nilai ambang batas (*threshold*) kepadatan.



(d) Density-based clusters. Clusters are regions of high density separated by regions of low density.

Image Source

Metode Clustering

Clustering Approaches

Partitional Clustering

- Perlu menentukan jumlah kluster
- Iterasi untuk menempatkan data ke dalam cluster
- K-Means

Hierarchical Clustering

- Pembentukan cluster dilakukan secara hirarki
  - **Agglomerative :** Bottom-up
    - Menggabungkan dua titik yang memiliki kemiripan ke dalam sebuah cluster
  - **Divisive :** Top-down
    - Mulai dari sebuah cluser besar kemudian dibagi

Density Based Clustering

- Pembentukan cluster dilakukan berdasar kepadatan titik data pada suatu area;
- Antara cluster dipisahkan oleh area dengan kepadatan titik data yang rendah;
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

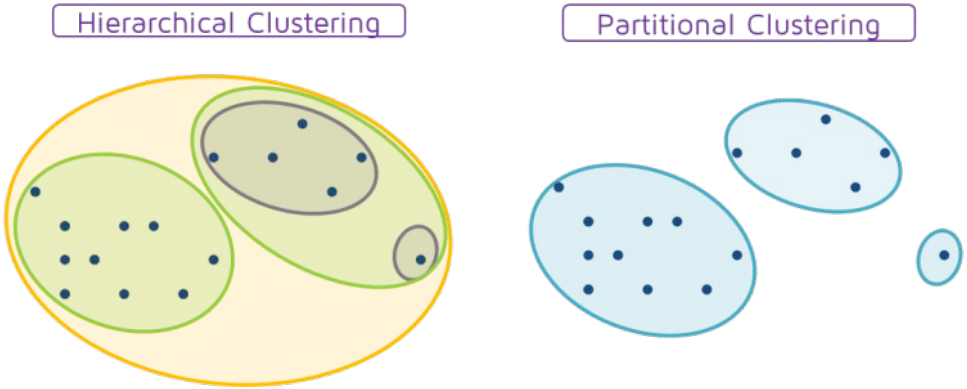
Partitional Clustering

- *Partitional clustering* adalah metode yang membagi seluruh *instance* (objek) ke dalam beberapa *cluster* secara eksklusif (tidak *overlap*), sehingga setiap objek data berada dalam tepat satu *cluster*.
- Setiap *cluster* memiliki pusat (*centroid*) yang telah ditentukan terlebih dahulu secara random, *instance* akan ditentukan masuk ke *cluster* mana berdasarkan *distance* terpendek dari *instance* ke tiap *centroid* pada tiap *cluster*.
- *Partitional clustering* biasanya dilakukan dengan teknik seperti ***k-means clustering*** atau ***Fuzzy C-Means clustering***.
- **Kelebihan:** sederhana dan efisien dalam pemrosesan data dengan jumlah titik data yang besar.
- **Kekurangan:** sensitif terhadap kondisi awal atau *seed* yang digunakan dalam algoritma, dan rentan terhadap data yang mengandung *noise* atau *outlier*.

Hierarchical Clustering

- *Hierarchical clustering* adalah teknik *clustering* yang membagi titik-titik data ke dalam *cluster* secara bertahap dengan mempertimbangkan jarak antara titik-titik data.
- Dalam hal ini, *hierarchical clustering* membentuk *cluster* dengan cara membuat pohon *cluster* (dendrogram) yang memetakan hubungan antara titik-titik data dan klaster.
- Ada dua jenis *hierarchical clustering*:
    - ***Agglomerative clustering*** memulai dengan setiap titik data sebagai *cluster* terpisah, kemudian menggabungkan *cluster* berdasarkan jarak antara *cluster* tersebut.
    - ***Divisive clustering*** memulai dengan seluruh titik data dalam satu *cluster*, kemudian membaginya menjadi *cluster* yang lebih kecil berdasarkan jarak antara titik-titik data.
- **Kelebihan:** memiliki fleksibilitas dalam pemilihan jumlah *cluster*, interpretasi hasil yang mudah, dan dapat menunjukkan hubungan antara *cluster* dan titik-titik data yang tidak terlihat dalam teknik *clustering* lainnya.
- **Kekurangan:** rentan terhadap *overfitting* dan waktu pemrosesan yang lama.

Partitional Clustering vs Hierarchical Clustering



- Perbedaan utama antara *partitional clustering* dan *hierarchical clustering* adalah cara *cluster* dibentuk.
- *Partitional clustering* membagi data ke dalam sejumlah *cluster* yang eksklusif, sedangkan *hierarchical clustering* membentuk kelompok atau *cluster* secara bertahap dengan mempertimbangkan jarak antara titik-titik data.

Image Source

K-means Clustering

K-means Clustering

- Pada *k-means clustering*, data yang ada akan dikelompokkan ke dalam *k cluster*, di mana setiap *cluster* memiliki *centroid* atau pusat data tertentu (rata-rata seluruh data cluster).
- Proses pembentukan *cluster* dilakukan dengan menghitung jarak antara setiap data dengan *centroid*, dan data tersebut akan ditempatkan ke dalam *cluster* yang memiliki *centroid* terdekat.
- Tujuannya adalah untuk meminimalkan jarak antara setiap data dengan centroid di dalam cluster yang sama, dan memaksimalkan jarak antara centroid dari setiap cluster yang berbeda.
- Algoritma dasarnya sangat sederhana namun Jumlah cluster, K, harus ditentukan diawal termasuk pemilihan Centroid awal secara acak.

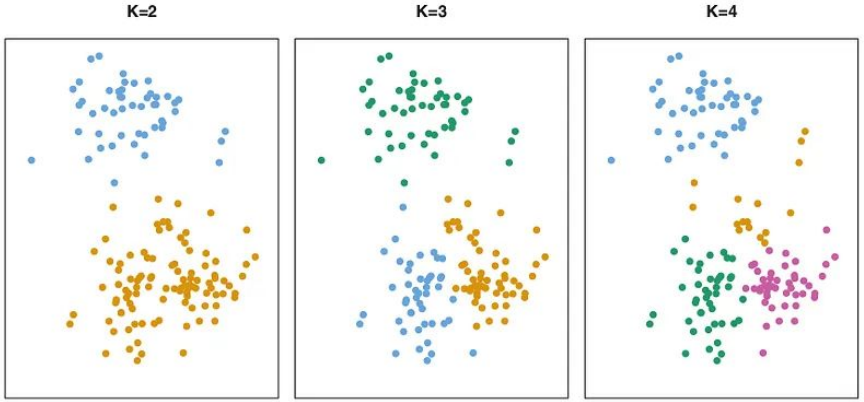
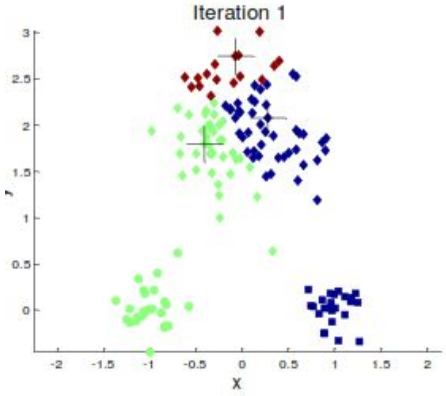
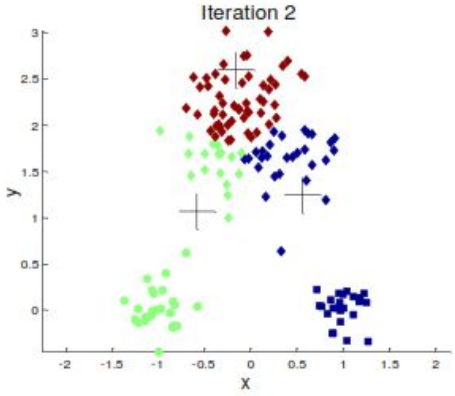
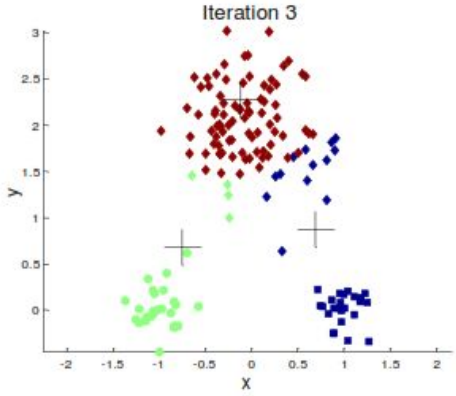


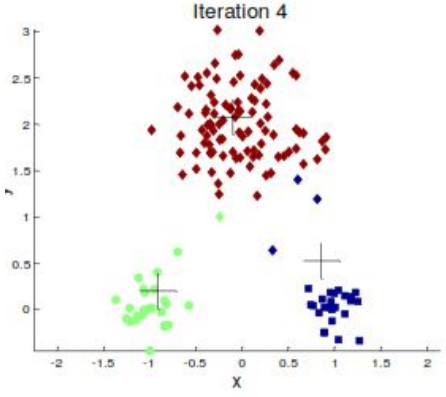
Image Source

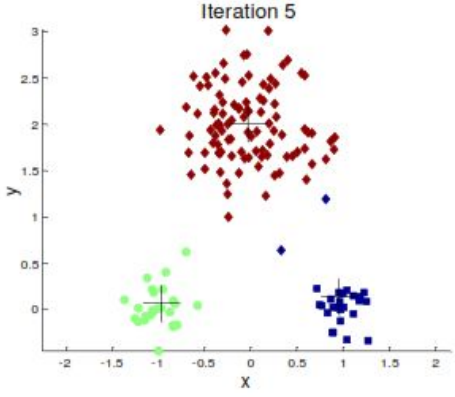
K-Means Algorithm









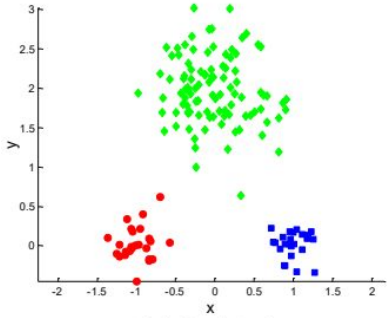


Algorithm 1 Basic K-means Algorithm.

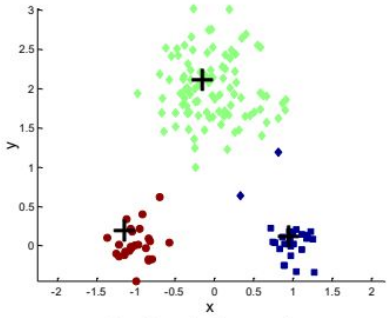
1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

DOLab

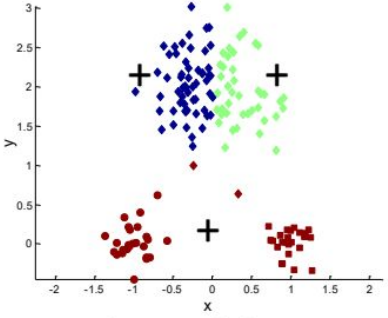
K-Means Limitation



Original Points



Optimal Clustering



Sub-optimal Clustering

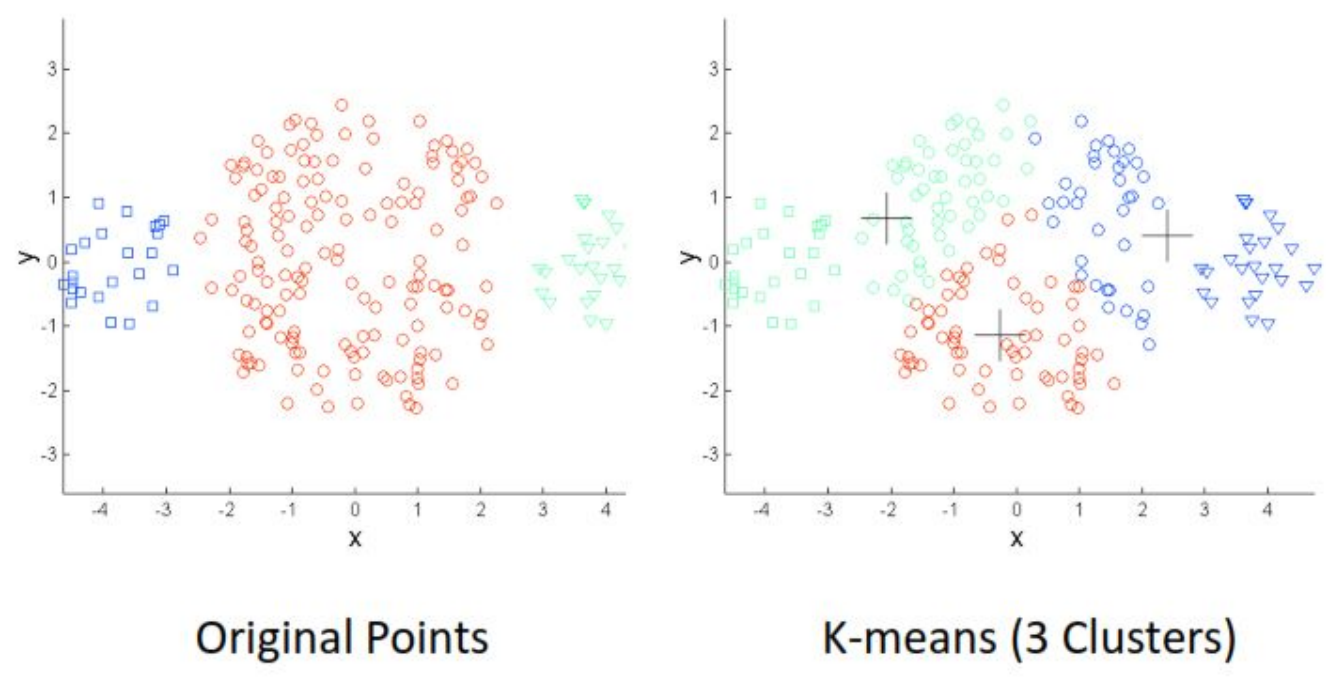
Permasalahan *Initial Centroid*

- *Initial centroid* pada *k-means clustering* adalah titik awal yang digunakan untuk menghitung jarak antara data dengan *centroid* dan menentukan *cluster* mana yang menjadi tempat data tersebut ditempatkan.
- Pemilihan *initial centroid* dapat mempengaruhi performa clustering yang dihasilkan.
- Beberapa masalah yang sering terjadi pada *initial centroid* diantaranya:
    - **Randomness:** Initial centroid dipilih secara acak, sehingga kemungkinan ditemukan centroid yang buruk, yang menyebabkan clustering yang dihasilkan tidak optimal.
    - **Ketergantungan pada pengamat:** Initial centroid dapat dipilih secara manual oleh pengamat, sehingga tergantung pada pengetahuan dan pengalaman pengamat dalam menentukan centroid yang baik.
    - **Data berbentuk tidak simetris:** Jika data memiliki bentuk yang tidak simetris, seperti data dengan outliers atau data yang memiliki kelompok yang kecil, maka initial centroid dapat dipilih secara tidak optimal dan clustering yang dihasilkan menjadi tidak baik.

Image Source

K-Means Limitation

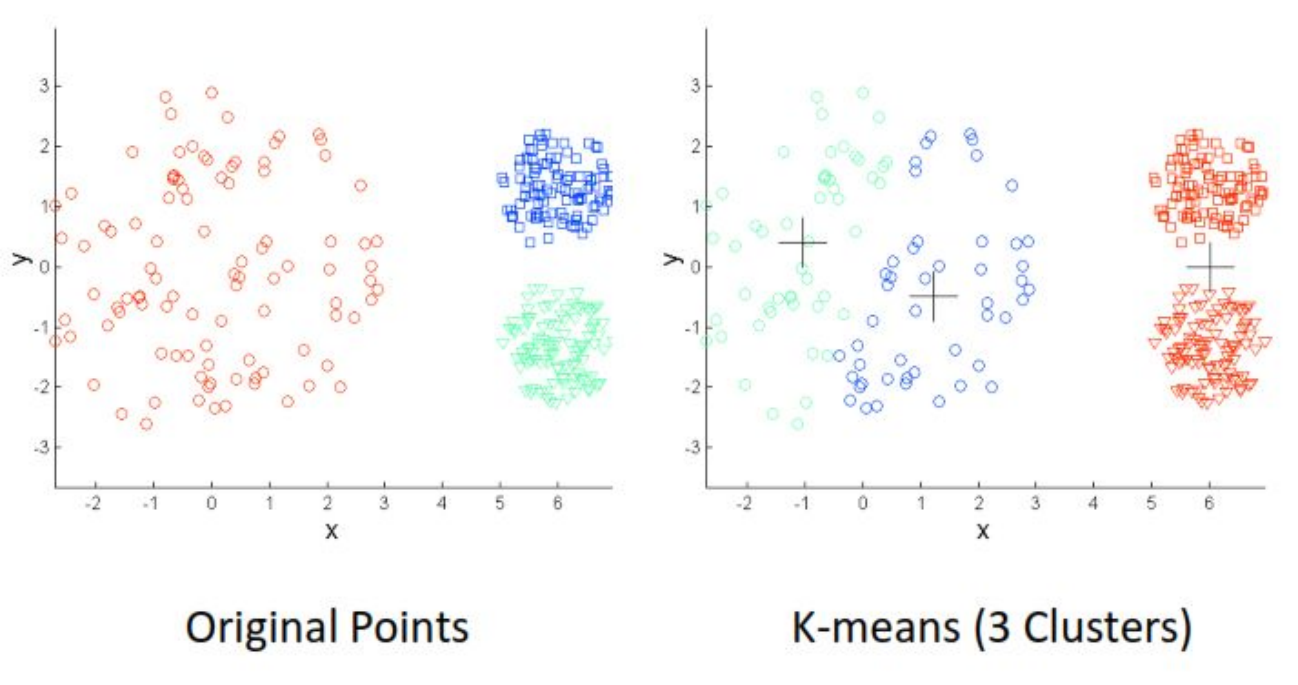
Problem: Cluster Size difference



Hasil Clustering berbeda dengan klasifikasi data original (clustering dicoba pada data dengan label)

K-Means Limitation

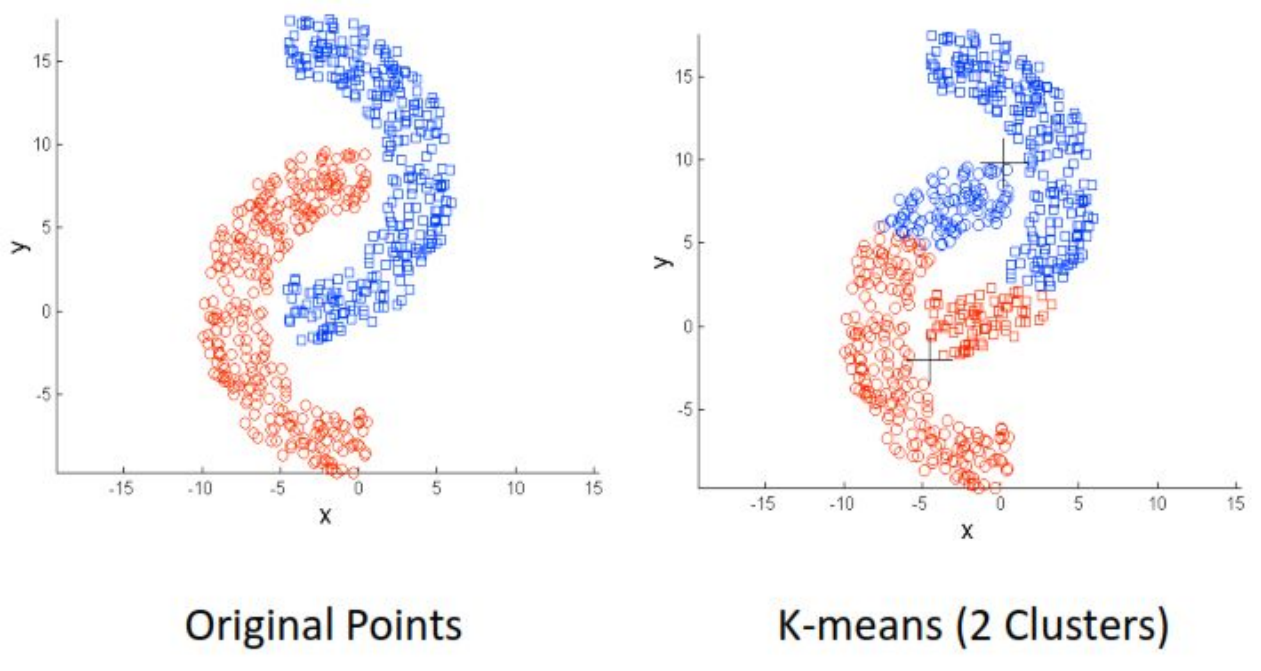
Problem: Cluster Density difference



Hasil Clustering berbeda dengan klasifikasi data original (clustering dicoba pada data dengan label)

K-Means Limitation

Problem: Non Globular Shape of Clusters



Hasil Clustering berbeda dengan klasifikasi data original (clustering dicoba pada data dengan label)

DQLab

Live Class

Handling Limitations and Problems

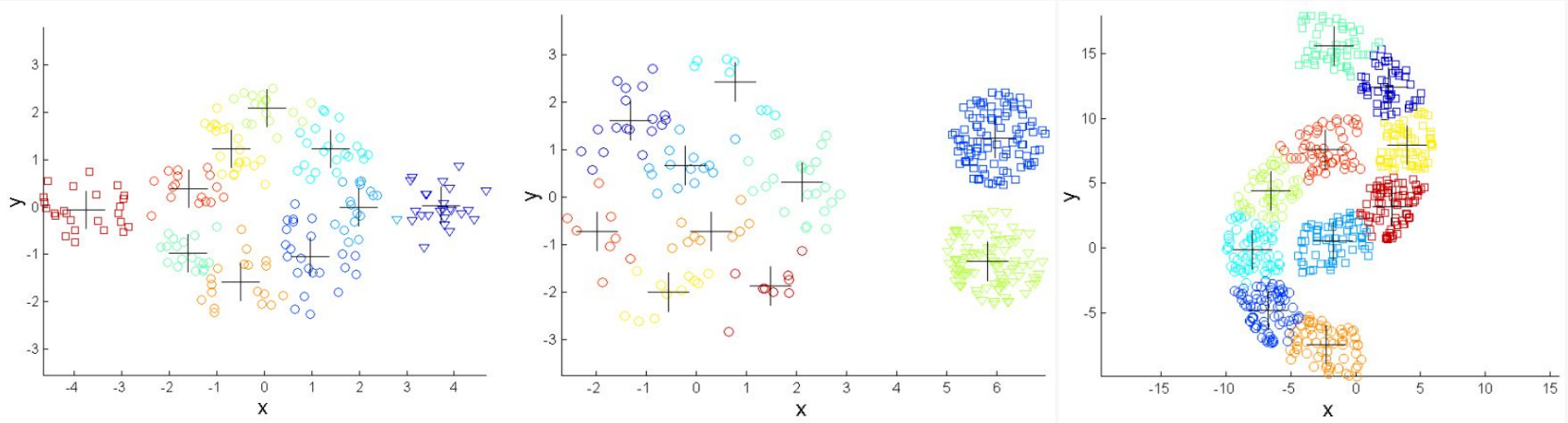
What to do?

Beberapa cara untuk mengatasi masalah initial centroid pada k-means clustering antara lain:

- Pengulangan: Melakukan pengulangan k-means clustering dengan centroid awal yang berbeda-beda untuk mendapatkan hasil clustering yang lebih baik.
- PCA (*Principal Component Analysis*): Melakukan PCA untuk mengurangi dimensi data dan memilih initial centroid berdasarkan hasil PCA.
- Pre-processing:
  - Normalisasi data
  - Hapus outliers
- Post-processing:
  - Hapus *cluster* yang kemungkinan merepresentasikan *outliers*
  - Split cluster dengan skor SSE yang tinggi
  - Merge cluster dengan skor SSE yang rendah

Image Source

Overcoming K-Means Limitation



Salah satu solusi dari masalah dalam hasil Clustering

Buat cluster sebanyak mungkin cluster, kemudian analisis hasilnya dan gabungkan cluster-cluser yang berdekatan

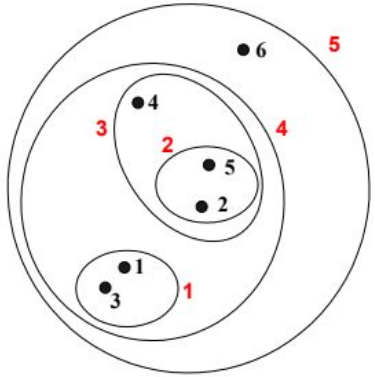
Image Source

Hierarchical Clustering

Hierarchical Clustering

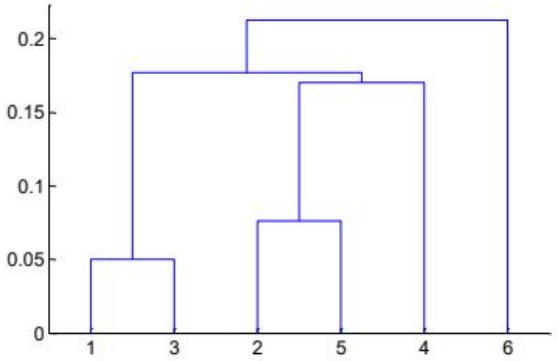
Menghasilkan satu set nested cluster yang terorganisir sebagai pohon hierarki

- Dapat divisualisasikan sebagai dendogram
- Sebuah pohon seperti diagram yang merekam urutan penggabungan atau pemisahan



Tidak harus mengasumsikan sejumlah cluster tertentu

- Setiap jumlah klaster yang diinginkan dapat diperoleh dengan 'memotong' dendogram pada tingkat yang tepat



Hasil Ccluster mungkin sesuai dengan taksonomi yang bermakna

- Contoh dalam ilmu biologi (misalnya, kerajaan hewan, rekonstruksi filogeni, …)

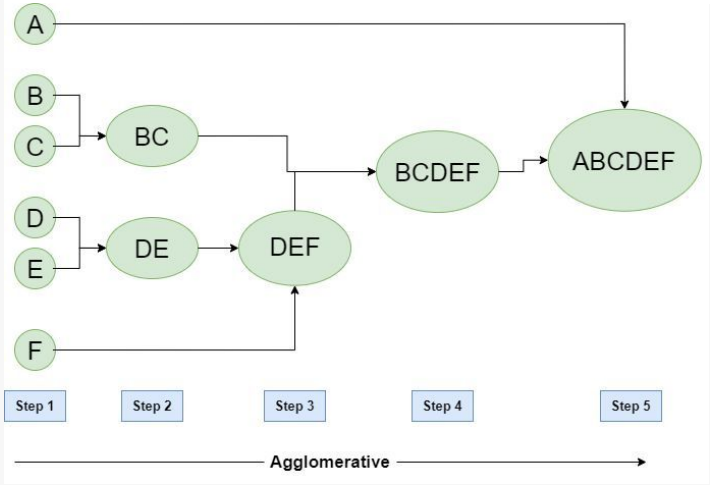
Hierarchical Clustering

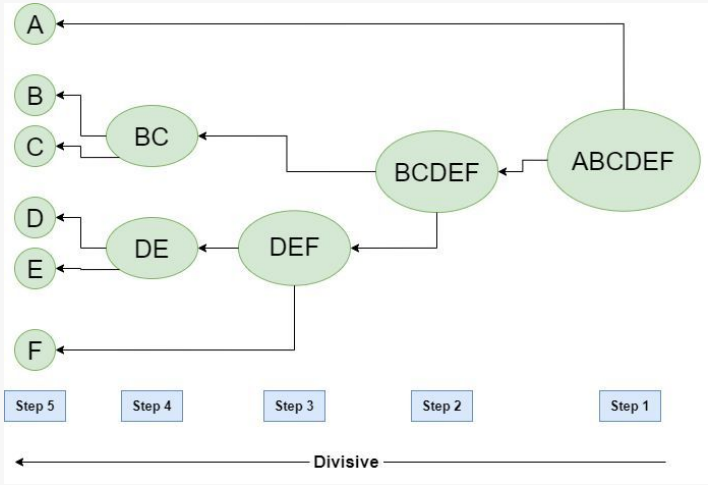
Two main types of hierarchical clustering

- Agglomerative:
  - Start with the points as individual clusters
  - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
- Divisive:
  - Start with one, all-inclusive cluster
  - At each step, split a cluster until each cluster contains a point (or there are k clusters)

Agglomerative
Bottom-up

Divisive
Top-Down





Similarity?
Kesamaan?

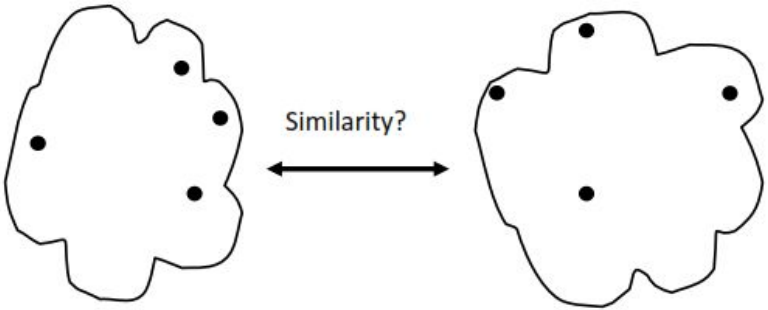
Inter-Cluster Similarity

How to decide closest pair??

- Inter-cluster similarity atau kemiripan antar-klaster adalah ukuran seberapa mirip atau berbedanya dua atau lebih klaster dalam sebuah analisis clustering

- Linkage method adalah metode yang digunakan untuk menghitung jarak antara dua klaster. Metode ini menggabungkan konsep jarak antar-objek data dalam sebuah klaster menjadi jarak antar-klaster dengan menggunakan berbagai teknik penghitungan yang berbeda.

- Semakin kecil nilai jarak antar-klaster, semakin mirip kedua klaster tersebut dan semakin tinggi inter-cluster similarity. Sebaliknya, semakin besar jarak antar-klaster, semakin berbeda kedua klaster tersebut dan semakin rendah inter-cluster similarity.

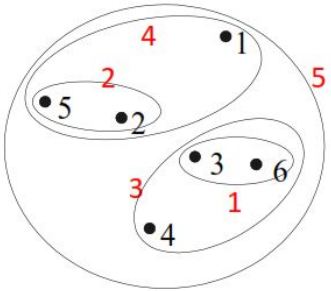


Linkage Method

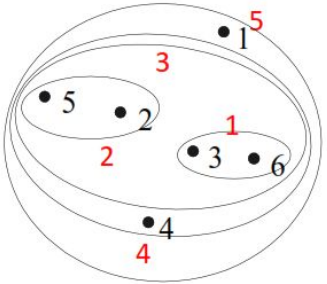
- Linkage method yang umum digunakan dalam hierarchical clustering, yaitu:
  - Single linkage: Jarak minimum antar klaster.
  - Complete linkage: Jarak maksimum antar klaster.
  - Average linkage: Jarak rata-rata antar klaster.
  - Centroid linkage: Jarak centroid antar klaster.

- Setiap jenis linkage method akan menghasilkan nilai jarak antar-klaster yang berbeda-beda. Pemilihan linkage method yang tepat sangat tergantung pada karakteristik data dan tujuan analisis clustering.

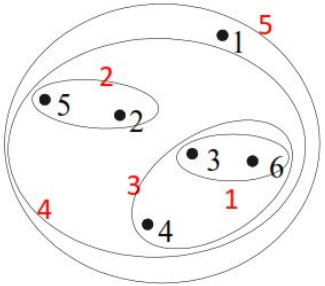
- Misalnya, jika data memiliki varians yang besar dan heterogen, maka complete linkage atau average linkage mungkin lebih tepat. Namun, jika data memiliki variasi yang kecil dan homogen, maka single linkage dapat dipertimbangkan.



MAX (Complete Linkage)

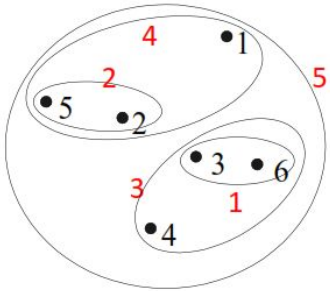


MIN (Single Linkage)



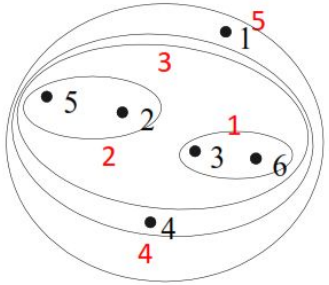
Group Average Distance

Linkage Method



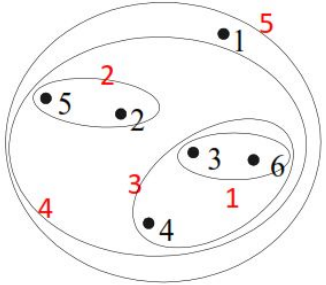
MAX (Complete Linkage)
STRENGTH:
- Less susceptible to noise and outliers

LIMITATION:
- Tends to break large clusters
- Biased towards globular clusters



MIN (Single Linkage)
STRENGTH:
- Can handle non-elliptical shapes

LIMITATION :
- Sensitive to noise and outliers



Group AVERAGE Distance
STRENGTH:
- Less susceptible to noise and outliers

LIMITATION:
- Biased towards globular clusters

DQLab

Kendala Hierarchical Clustering

Live Class

Catatan atas Hierachical Clustering

Meskipun hierarchical clustering memiliki beberapa kelebihan dalam analisis clustering, seperti kemampuan untuk menghasilkan struktur hierarkis yang lebih mudah dipahami dan diinterpretasikan, namun terdapat beberapa hal yang perlu dipertimbangkan ketika menggunakannya, antara lain:

- Ketergantungan pada jumlah objek data: Hierarchical clustering membutuhkan jumlah objek data yang cukup besar dan kurang efektif untuk dataset dengan jumlah objek data yang terbatas.
- Ketergantungan pada linkage method: Berbagai jenis linkage method dapat menghasilkan hasil clustering yang berbeda-beda, sehingga pemilihan jenis linkage method yang tepat sangat penting.
- Interpretasi yang sulit: Struktur hierarkis yang dihasilkan oleh hierarchical clustering dapat menjadi sangat kompleks dan sulit untuk diinterpretasikan, terutama jika jumlah objek data dan klaster yang besar.
- Kompleksitas waktu yang tinggi: Hierarchical clustering membutuhkan waktu komputasi yang cukup lama untuk menganalisis dataset yang besar dan kompleks.

Evaluasi Kinerja Clustering

Sum of Squared Errors (SSE)

To get SSE, we square these errors and sum them:

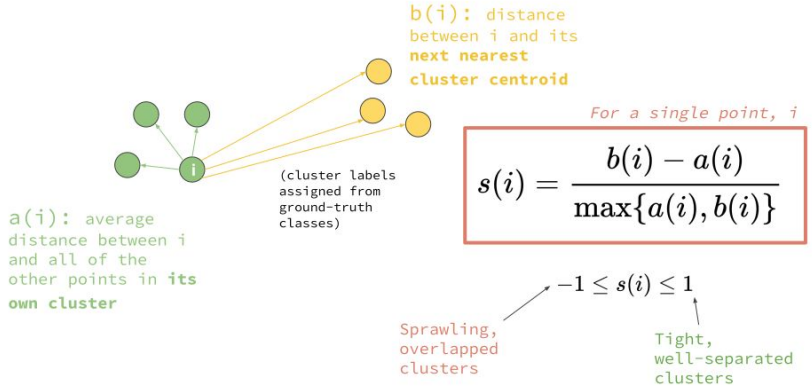
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

where x is a data point in cluster C_i and m_i is the centroid of C_i.

Sum of Squared Errors (SSE) menghitung jarak kuadrat antara setiap data dengan centroid dalam cluster, kemudian menjumlahkan semua nilai jarak tersebut untuk setiap cluster.
Semakin kecil nilai SSE, semakin baik performa k-means clustering.

Image Source

Silhouette Score



Metode ini mengukur seberapa baik setiap data cocok dengan cluster tempat ia berada, dibandingkan dengan cluster lain.
Semakin tinggi nilai silhouette score, semakin baik performa k-means clustering.

Image Source

Dunn's Index

$$\text{Dunn index}(U) = \min_{1 \le i \le c} \left\{ \min_{1 \le j \le c,\, j \ne i} \left\{ \frac{\delta(X_i, X_j)}{\max_{1 \le k \le c} \{\Delta(X_k)\}} \right\} \right\}$$

Dimana:

$\delta(X_i, X_j)$ is the intercluster distance i.e. the distance between cluster Xi and Xj

$\Delta(X_k)$ is the intracluster distance of cluster Xk i.e. distance within the cluster Xk

Mengukur kualitas hasil clustering dengan membandingkan jarak antara dua klaster (inter-cluster distance) dengan jarak antara objek data dalam satu klaster yang sama (intra-cluster distance).
Semakin besar nilai Dunn index, semakin baik kualitas hasil clustering.

Image Source

Davies–Bouldin Index (DBI)

$$DB\ index(U) = 1/k \sum_{i=1}^{k} \max_{i \neq j} \left\{ \frac{\Delta(Xi) + \Delta(Xj)}{\delta(Xi, Xj)} \right\}$$

Dimana:

δ (Xi, Xj) is the intercluster distance i.e. the distance between cluster Xi and Xj

Δ(Xk) is the intracluster distance of cluster Xk i.e. distance within the cluster Xk

Mengukur kualitas hasil clustering dengan mempertimbangkan jarak antara klaster dan jarak antara objek-objek data dalam satu klaster yang sama. Semakin kecil nilai DBI, semakin baik kualitas hasil clustering.

Image Source

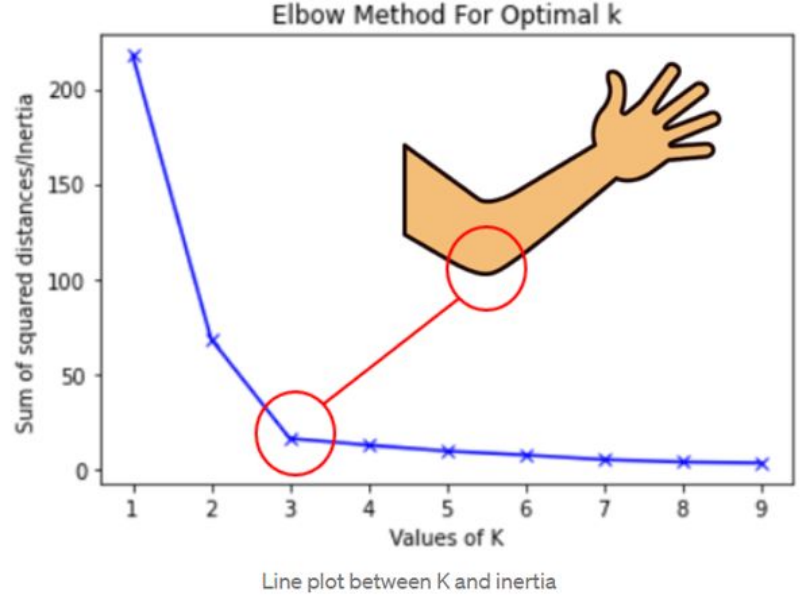
DQLab

Penentuan Jumlah Optimal Cluster (*k*)

liveClass

Elbow Method

- Metode ini melibatkan plot nilai SSE (*Sum of Squared Errors*) untuk setiap nilai k yang diuji.
- SSE adalah jumlah jarak kuadrat dari setiap data ke centroid dalam cluster.
- Pada plot SSE vs k, akan terbentuk kurva yang menyerupai lengkungan tangan.
- Kita kemudian memilih k di titik di mana penurunan SSE mulai melambat atau berbentuk seperti "siku" pada kurva, yang menunjukkan peningkatan penurunan SSE mulai kurang signifikan.
- Titik ini disebut elbow point, dan k yang sesuai dengan elbow point tersebut dianggap sebagai k yang optimal.

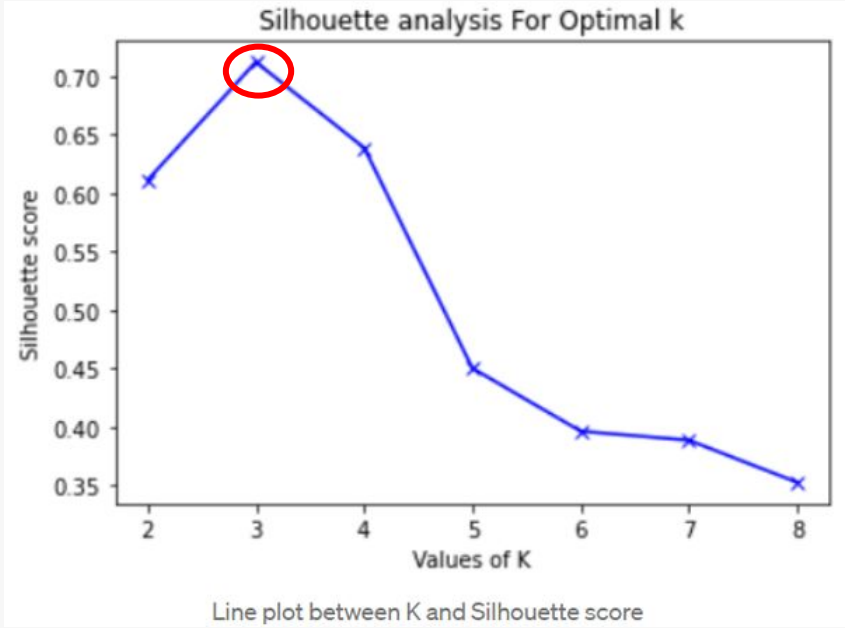


Line plot between K and inertia

Image Source

Silhouette Method

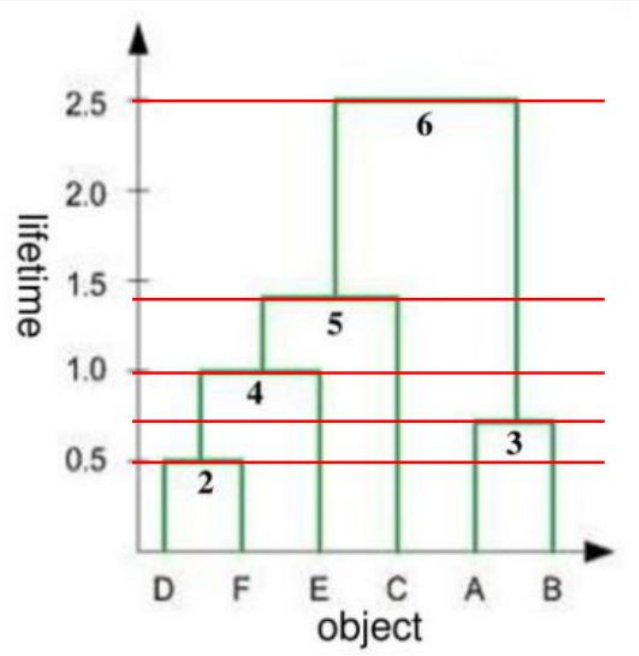
- Metode ini melibatkan plot koefisien siluet (*silhouette coefficient*) untuk setiap nilai k yang diuji.
- Silhouette coefficient adalah ukuran seberapa baik setiap data cocok dengan cluster tempat ia berada, dibandingkan dengan cluster lain.
- Pada plot siluet vs k, kita mencari nilai k yang menghasilkan siluet tertinggi, karena itu menunjukkan seberapa baik data cocok dengan cluster tempat ia berada.



Line plot between K and Silhouette score

Image Source

Lifetime dan K-Cluster Lifetime



Lifetime:
Lifetime mengacu pada jarak antara klaster sejak klaster itu terbentuk hingga klaster tersebut bergabung dengan klaster lain selama proses clustering.
Misalnya, lifetime untuk A, B, C, D, E, dan F adalah 0,71, 0,71, 1,41, 0,50, 1,00, dan 0,50, masing-masing. Lifetime dari (AB) adalah 2,50 - 0,71 = 1,79.

K-Cluster Lifetime:
K-Cluster Lifetime mengukur jarak antara klaster sejak k klaster terbentuk hingga k klaster bergabung menjadi k-1 klaster selama proses k-means clustering.
Misalnya:

- Lifetime 5-klaster adalah 0,71 - 0,50 = 0,21
- Lifetime 4-klaster adalah 1,00 - 0,71 = 0,29
- Lifetime 3-klaster adalah 1,41 - 1,00 = 0,41
- Lifetime 2-klaster adalah 2,50 - 1,41 = 1,09

Image Source

Catatan

- Jika jumlah klaster diketahui, maka kondisi penghentian telah ditentukan.
- Menggunakan k-cluster cifetime sebagai rentang nilai ambang pada pohon dendrogram yang mengarah pada identifikasi k klaster. Kita dapat memotong pohon dendrogram pada titik maksimum k-Cluster Lifetime untuk menemukan k yang "tepat".

- Dalam beberapa kasus, terkadang tidak jelas adanya elbow point pada grafik Elbow Method, atau terdapat lebih dari satu nilai k yang menghasilkan koefisien siluet yang tinggi pada grafik Silhouette Method.
- Oleh karena itu, pada akhirnya, pemilihan nilai k terbaik tetaplah bersifat subjektif dan harus dipertimbangkan dengan hati-hati berdasarkan pengalaman dan pengetahuan yang ada mengenai data yang dianalisis.

Image Source

Hands-on

Terima Kasih