

# **DATA MINING**

## **Pertemuan 8: Clustering**

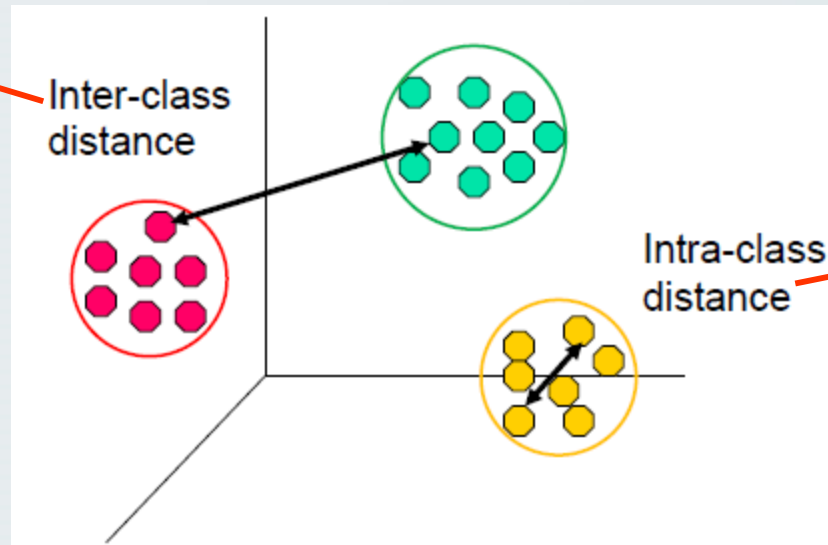
# Materi

- Pendahuluan
  - Definisi
  - Aplikasi clustering
- Model Clustering
  - Partitional Clustering
  - Hierarchical Clustering

# Clustering

- Pencarian **kelompok** dari sejumlah objek sedemikian hingga objek-objek dalam satu grup adalah **mirip** (atau berelasi) satu sama lain dan **berbeda** (atau tidak berelasi) dengan objek di kelompok lain

maksimum



minimum

# Aplikasi dari Clustering

- Perencanaan kota
  - Misal kita ingin membangun 2 kantor polisi di area berikut, dimana lokasi yang paling tepat?

Plan A



Plan B



# Aplikasi dari Clustering (lanjutan)

- Segmentasi pasar
  - Misal Anda adalah seorang manajer kantor cabang di sebuah perusahaan asuransi
  - Anda memiliki 10 tim marketing
  - Anda ingin membagi pasar ke dalam 10 segmen, sehingga tiap tim dapat berkonsentrasi pada sebuah pasar tertentu saja
  - Beberapa kriteria untuk segmentasi:
    - Jenis Kelamin
    - Usia
    - Pendapatan
    - Pengeluaran
    - Karir
    - ....

# Aplikasi dari Clustering (lanjutan)

- Mengurangi informasi yang serupa dalam penampilan hasil pencarian


Google UPN "Veteran" jatim

+Intan

Web News Images Maps Shopping More Search tools



About 347,000 results (0.66 seconds)

**UPN "Veteran" Jawa Timur**  
[www.upnjatim.ac.id/](http://www.upnjatim.ac.id/) [Translate this page](#)  
Senin 6 Oktober 2014, [Cached](#) [Persejarah bagi civitas Akademika UPN "Veteran"](#)  
Jawa Timur. Pada hari ini Susilo Bambang Yudhoyono, ...  
4.3 ★★★★★ [Similar](#) [Write a review](#) · [Google+ page](#)

 Jl. Raya Rungkut Madya, Gunung Anyar, Jawa Timur 60294, Indonesia  
+62 31 8706369  
[Siamik - Sistem Informasi ... - E-Learning](#)

**Sistem Informasi Akademik (SIAMIK) - UPN "Veteran"**  
<https://siamik.upnjatim.ac.id/> [Translate this page](#)  
UPN VETERAN JAWA TIMUR. SIAMIK (Sistem Informasi Akademik). Adalah suatu sistem informasi untuk mengelola KRS (Kartu Rencana Studi), KHS (Kartu ...

**Sistem Informasi Mahasiswa Baru 2014 - UPN "Veteran"**  
<https://simaba.upnjatim.ac.id/> [Translate this page](#)  
MABA 2015. UPN "Veteran" Jawa Timur yg berlokasi di Surabaya merupakan salah satu Perguruan Tinggi Negeri di Indonesia yang didirikan sejak 5 Juli 1959.

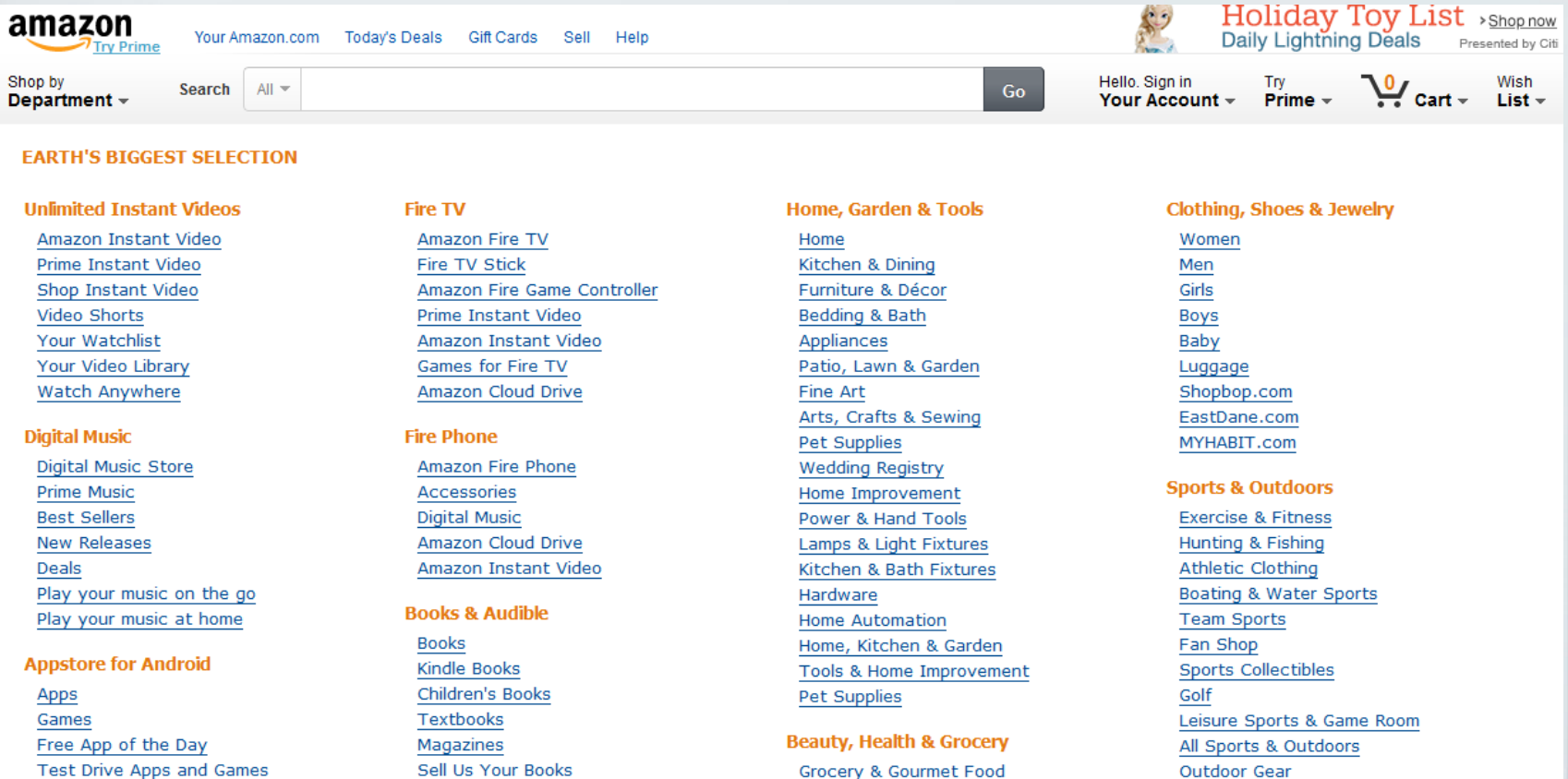
**University of Pembangunan Nasional Veteran**  
University in Surabaya, Indonesia

[Directions](#) [Write a review](#)

The University of Pembangunan Nasional Veteran East Java abbreviated as UPN Veteran is a private university in Indonesia, located at Surabaya municipality, Province of East Java. [Wikipedia](#)

# Aplikasi dari Clustering (lanjutan)

- Memudahkan pencarian



The screenshot displays the Amazon.com homepage with a navigation bar at the top. The navigation bar includes the Amazon logo, links to 'Your Amazon.com', 'Today's Deals', 'Gift Cards', 'Sell', and 'Help'. On the right side of the navigation bar, there is a 'Holiday Toy List' banner with a 'Shop now' link, and a 'Daily Lightning Deals' banner. Below the navigation bar, there is a search bar with a 'Go' button. To the left of the search bar, there is a 'Shop by Department' dropdown menu. To the right of the search bar, there are links for 'Hello. Sign in Your Account', 'Try Prime', 'Cart', and 'Wish List'.

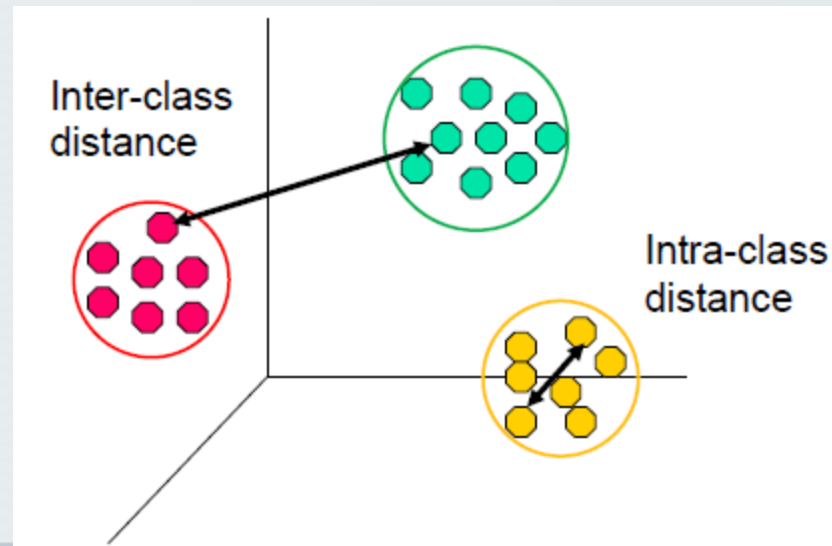
**EARTH'S BIGGEST SELECTION**

- Unlimited Instant Videos**
  - [Amazon Instant Video](#)
  - [Prime Instant Video](#)
  - [Shop Instant Video](#)
  - [Video Shorts](#)
  - [Your Watchlist](#)
  - [Your Video Library](#)
  - [Watch Anywhere](#)
- Digital Music**
  - [Digital Music Store](#)
  - [Prime Music](#)
  - [Best Sellers](#)
  - [New Releases](#)
  - [Deals](#)
  - [Play your music on the go](#)
  - [Play your music at home](#)
- Appstore for Android**
  - [Apps](#)
  - [Games](#)
  - [Free App of the Day](#)
  - [Test Drive Apps and Games](#)
- Fire TV**
  - [Amazon Fire TV](#)
  - [Fire TV Stick](#)
  - [Amazon Fire Game Controller](#)
  - [Prime Instant Video](#)
  - [Amazon Instant Video](#)
  - [Games for Fire TV](#)
  - [Amazon Cloud Drive](#)
- Fire Phone**
  - [Amazon Fire Phone](#)
  - [Accessories](#)
  - [Digital Music](#)
  - [Amazon Cloud Drive](#)
  - [Amazon Instant Video](#)
- Books & Audible**
  - [Books](#)
  - [Kindle Books](#)
  - [Children's Books](#)
  - [Textbooks](#)
  - [Magazines](#)
  - [Sell Us Your Books](#)
- Home, Garden & Tools**
  - [Home](#)
  - [Kitchen & Dining](#)
  - [Furniture & Décor](#)
  - [Bedding & Bath](#)
  - [Appliances](#)
  - [Patio, Lawn & Garden](#)
  - [Fine Art](#)
  - [Arts, Crafts & Sewing](#)
  - [Pet Supplies](#)
  - [Wedding Registry](#)
  - [Home Improvement](#)
  - [Power & Hand Tools](#)
  - [Lamps & Light Fixtures](#)
  - [Kitchen & Bath Fixtures](#)
  - [Hardware](#)
  - [Home Automation](#)
  - [Home, Kitchen & Garden](#)
  - [Tools & Home Improvement](#)
  - [Pet Supplies](#)
- Beauty, Health & Grocery**
  - [Grocery & Gourmet Food](#)
- Clothing, Shoes & Jewelry**
  - [Women](#)
  - [Men](#)
  - [Girls](#)
  - [Boys](#)
  - [Baby](#)
  - [Luggage](#)
  - [Shopbop.com](#)
  - [EastDane.com](#)
  - [MYHABIT.com](#)
- Sports & Outdoors**
  - [Exercise & Fitness](#)
  - [Hunting & Fishing](#)
  - [Athletic Clothing](#)
  - [Boating & Water Sports](#)
  - [Team Sports](#)
  - [Fan Shop](#)
  - [Sports Collectibles](#)
  - [Golf](#)
  - [Leisure Sports & Game Room](#)
  - [All Sports & Outdoors](#)
  - [Outdoor Gear](#)



# Seperti apa clustering yang baik?

- Metode clustering yang baik adalah yang menghasilkan cluster berkualitas tinggi dengan
  - Tingkat kesamaan tinggi pada item intra-class (dalam 1 kelas)
  - Tingkat kesamaan rendah pada item inter-class (dalam kelas yang berbeda)
  - Kemampuan untuk menemukan pola yang tersembunyi

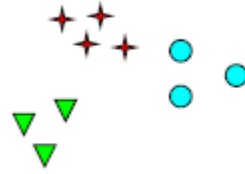




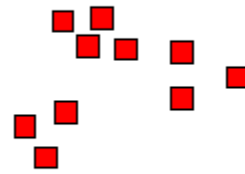
# Pertanyaan: Ada berapa cluster yang dapat terbentuk?



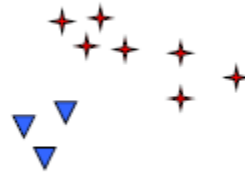
How many clusters?



Six Clusters



Two Clusters



Four Clusters

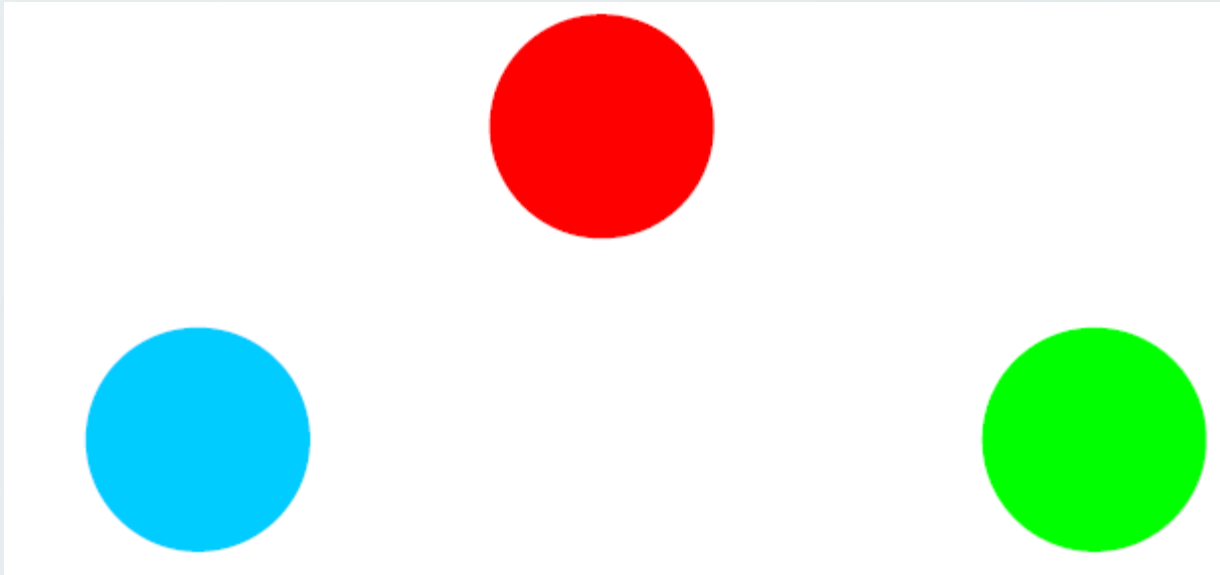
Pendefinisian sebuah cluster dapat bersifat ambigu

# Beberapa Tipe Cluster

- *Well-separated clusters* (terpisah dengan sempurna)
- *Center-based clusters* (memiliki pusat)
- *Contiguous clusters* (bersebelahan)
- *Density-based clusters* (berdasarkan kerapatan)
- Property atau Conceptual (dibagi berdasar sifat atau konsepnya)
- Dideskripsikan dengan sebuah Fungsi Objektif

# Tipe Cluster: Well-Separated

- Cluster adalah sekumpulan titik-titik sedemikian hingga sembarang titik pada sebuah cluster lebih dekat (atau lebih mirip) dengan titik pada cluster yang sama daripada dengan titik lain di luar cluster tersebut



Tiga cluster yang well-separated

# Tipe Cluster: Center-based

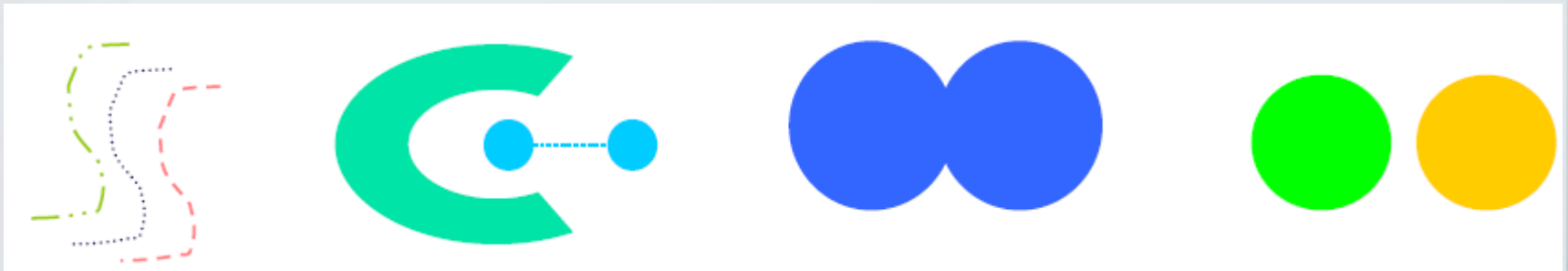
- Sebuah cluster adalah sekumpulan objek sedemikian hingga sebuah objek pada cluster tersebut lebih dekat (atau lebih mirip) dengan “pusat” cluster, daripada dengan pusat cluster lain
- Pusat dari cluster dapat berupa **centroid** (rata-rata dari semua titik pada cluster) atau **medoid** (titik yang paling representatif dari cluster tersebut)



**Empat cluster yang center-based**

# Type Cluster: Contiguity-Based

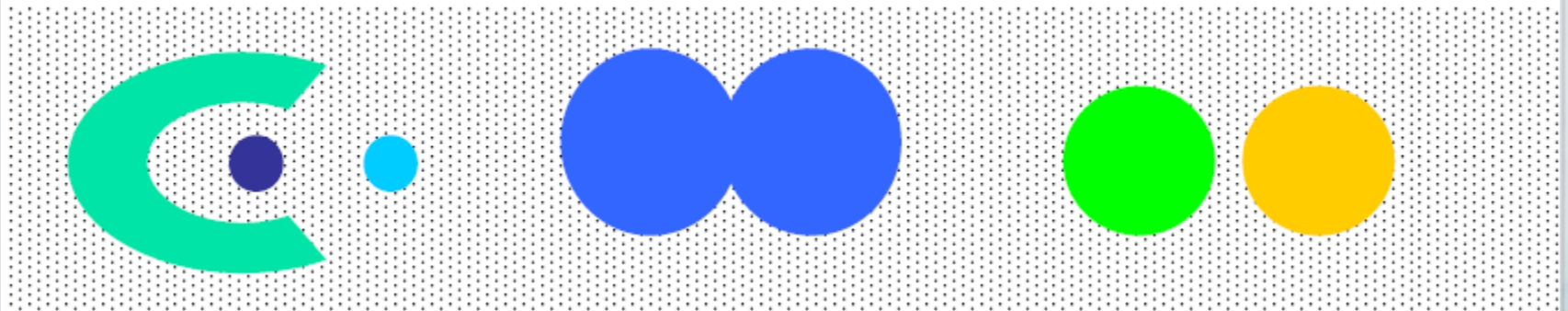
- Cluster yang contiguous (tetangga terdekat/*nearest neighbor* atau transitif):
  - Cluster adalah sekumpulan titik-titik sedemikian hingga sebuah titik pada cluster lebih dekat (atau lebih mirip) dengan satu atau lebih titik lainnya dalam cluster tersebut daripada dengan titik lainnya di luar cluster



Delapan cluster yang bersifat *contiguous*

# Tipe Cluster: Density-Based

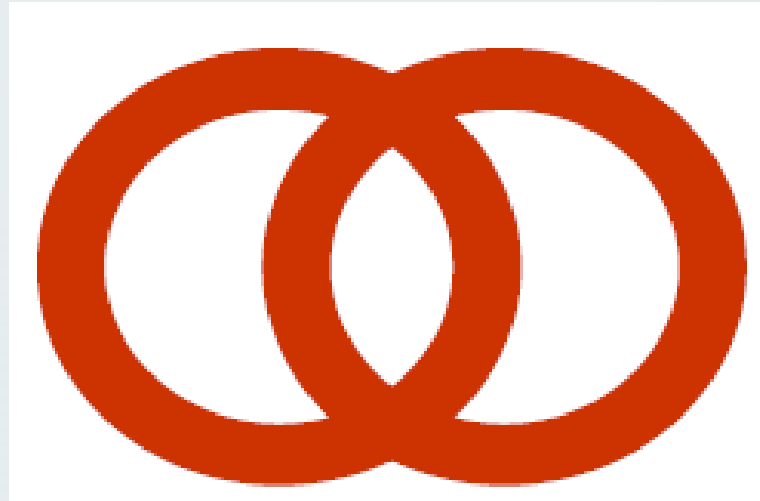
- Sebuah cluster adalah sebuah daerah yang rapat dengan titik-titik, dimana daerah dengan kerapatan rendah terpisah dengan daerah dengan kerapatan tinggi
- Tipe ini digunakan ketika cluster bersifat irreguler atau saling bertautan, dan ketika terdapat *noise* atau *outlier*



Enam cluster yang density-based

# Tipe Cluster: Conceptual

- Memiliki sifat atau konsep yang sama

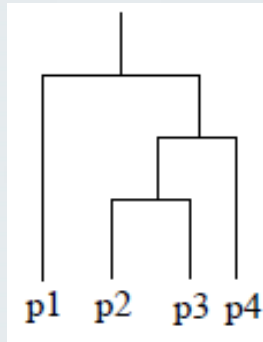


Dua cluster lingkaran yang saling overlap

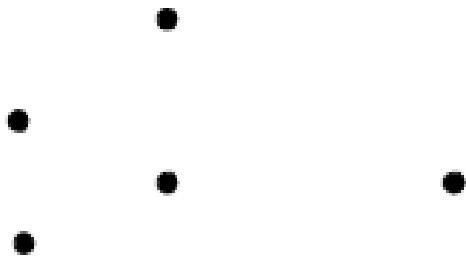
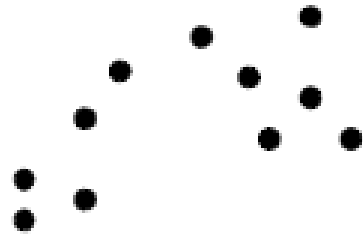


# Tipe Clustering

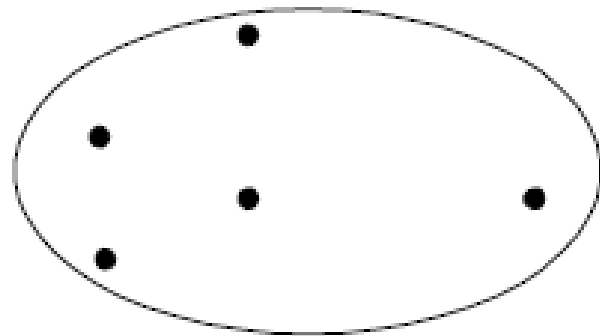
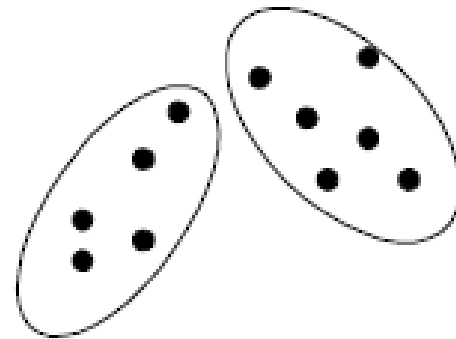
- Secara umum, ada dua tipe clustering:
  - Partitional Clustering
    - Membagi objek data ke dalam sub-himpunan (cluster) yang tidak overlap sedemikian hingga tiap objek data berada dalam 1 sub-himpunan
  - Hierarchical Clustering
    - Serangkaian cluster bersarang (nested clusters) yang teroganisir dalam bentuk pohon hirarkis → disebut **dendrogram**



# Partitional Clustering

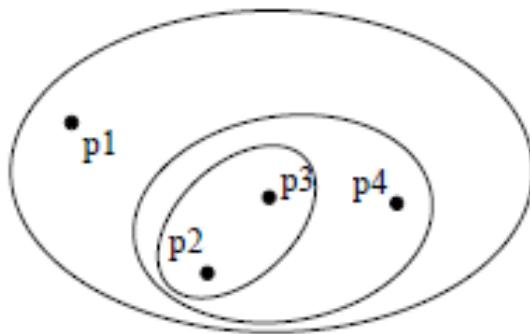


Original Points

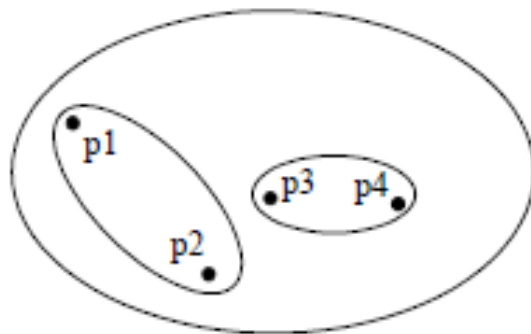


A Partitional Clustering

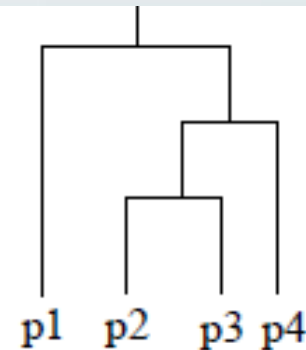
# Hierarchical Clustering



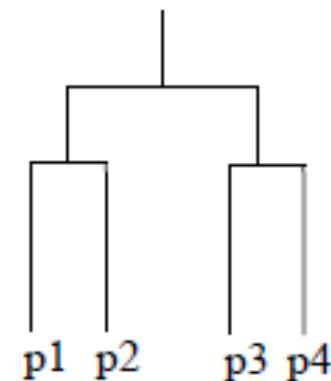
Traditional Hierarchical Clustering



Non-traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Dendrogram

Model Clustering:

**PARTITIONAL CLUSTERING**

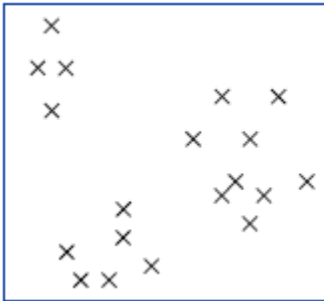
# 1. K-Means

- Langkah:
  - Pilih  $k$  titik sebagai centroid awal
  - Ulangi
    - Memberikan semua titik kepada centroid terdekat
    - Hitung ulang nilai centroid
  - Sampai semua centroid tidak berubah

# 1. K-Means (keterangan)

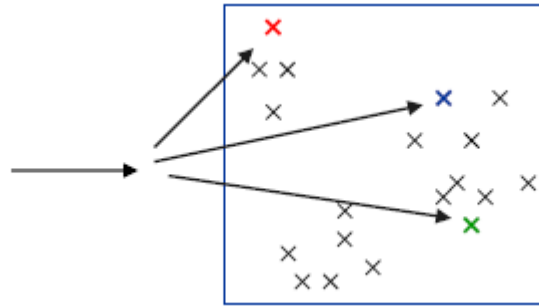
- Centroid awal biasanya dipilih secara random, sehingga cluster yang terbentuk dapat bervariasi setiap kali algoritma dijalankan
- Centroid biasanya merupakan rata-rata dari keseluruhan titik pada cluster
- Tingkat “kedekatan” diukur dengan Euclidean distance, cosine similarity, korelasi, dsb
- K-means akan mengerucut akibat dari pengukuran jarak di atas
- Sebagian besar pengerucutan terjadi pada beberapa iterasi pertama

# 1. K-Means (Contoh)

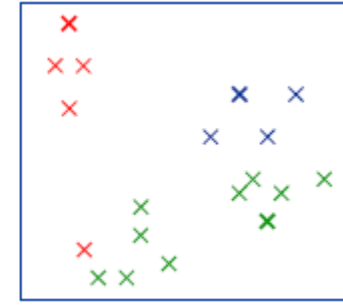


Data point awal

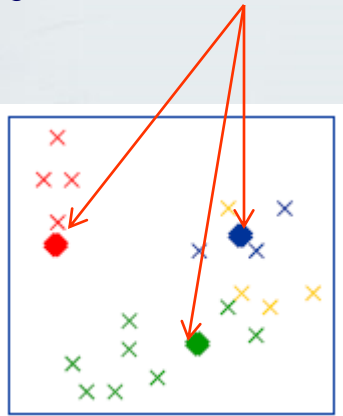
3 centroid baru  
yg bukan berasal dari data point



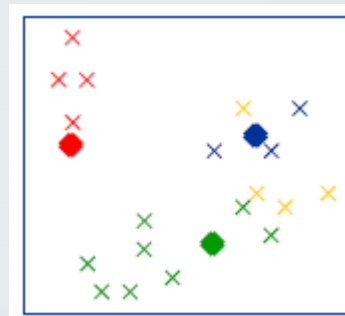
Secara acak memilih  
3 titik sebagai centroid



Berikan semua titik ke  
centroid terdekatnya



Hitung ulang centroid  
untuk setiap cluster



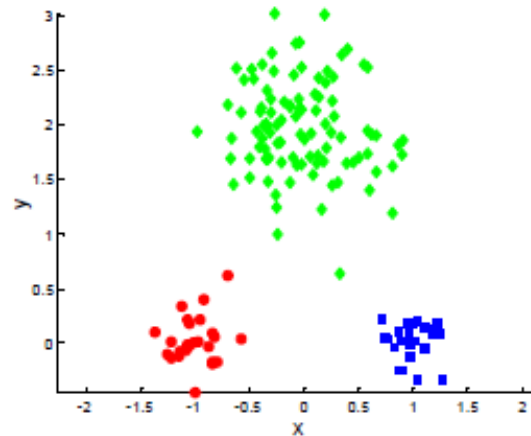
Berikan semua titik ke  
centroid terdekatnya

Ulangi:

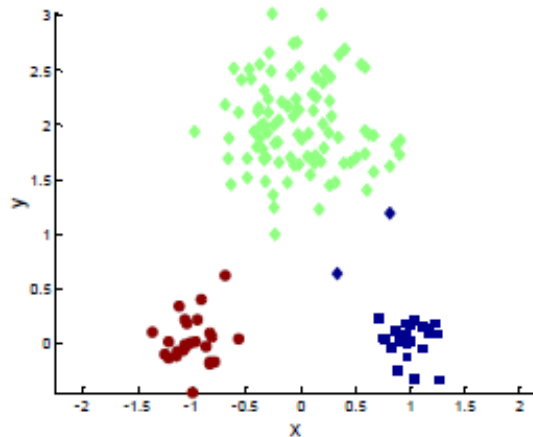
- Penghitungan ulang nilai centroid untuk tiap cluster
  - Berikan semua titik ke centroid terdekatnya
- Sampai nilai centroid tidak berubah



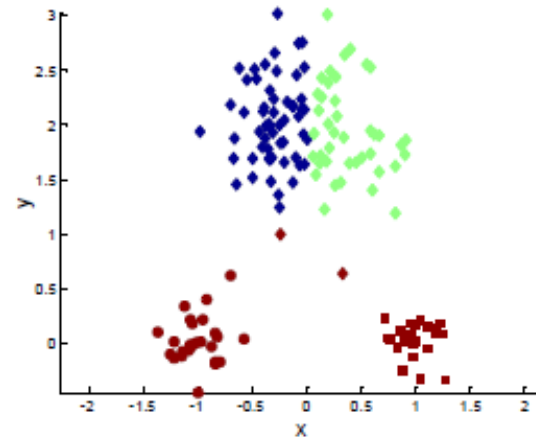
# Contoh Dua K-Means Clustering yang berbeda



Original Points

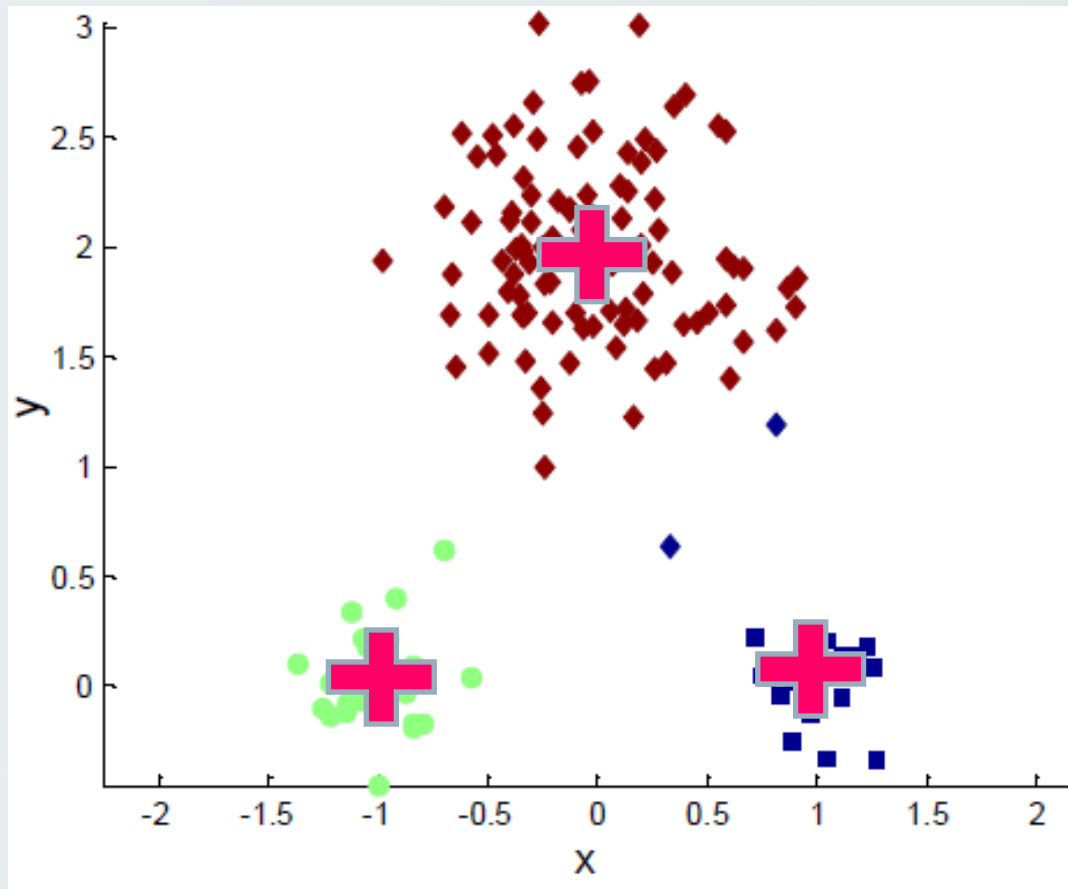


Optimal Clustering

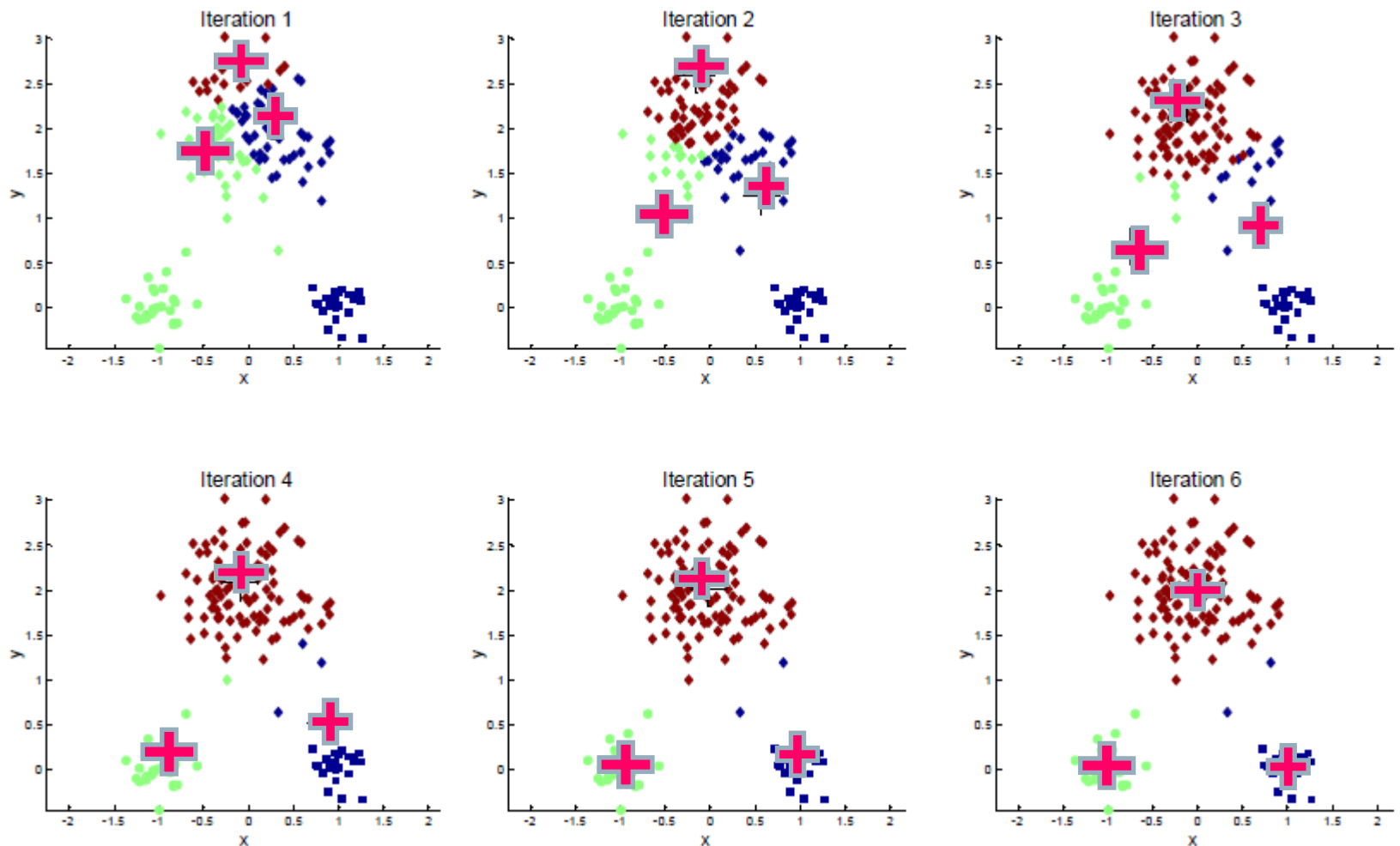


Sub-optimal Clustering

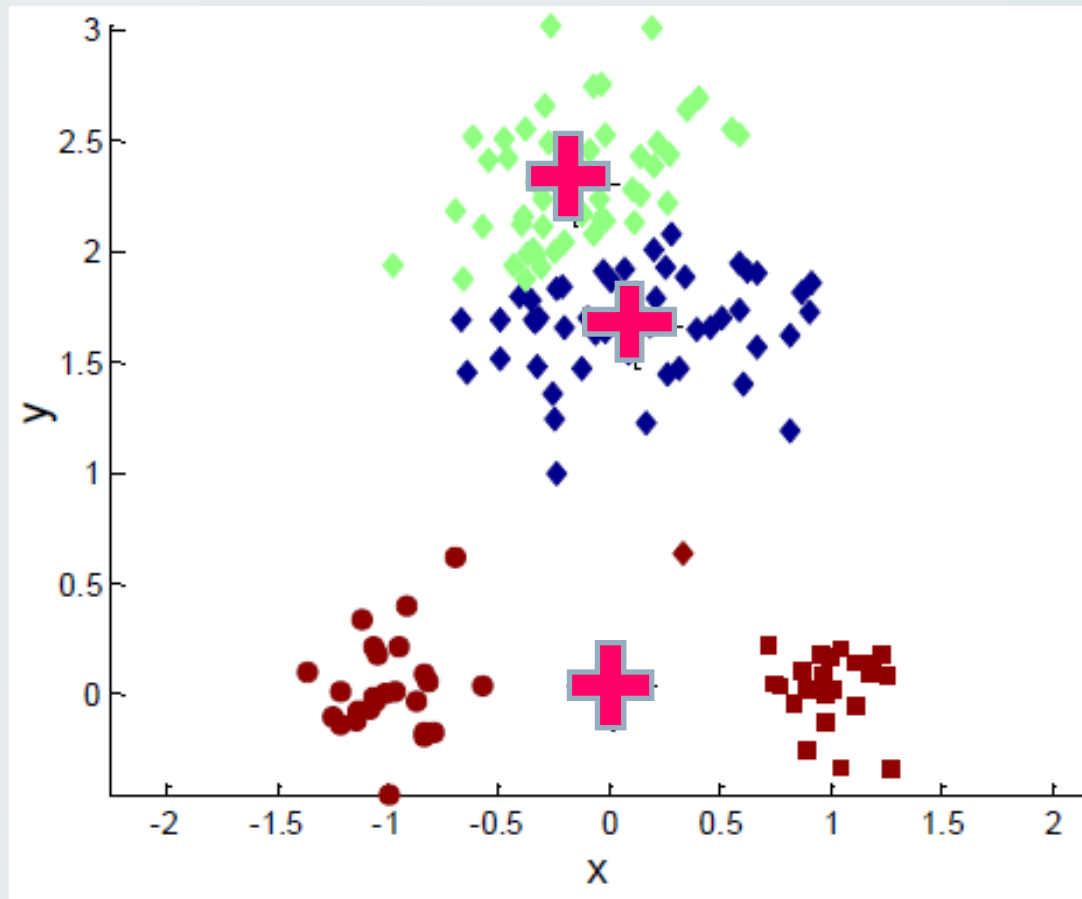
# Pentingnya Memilih Centroid Awal



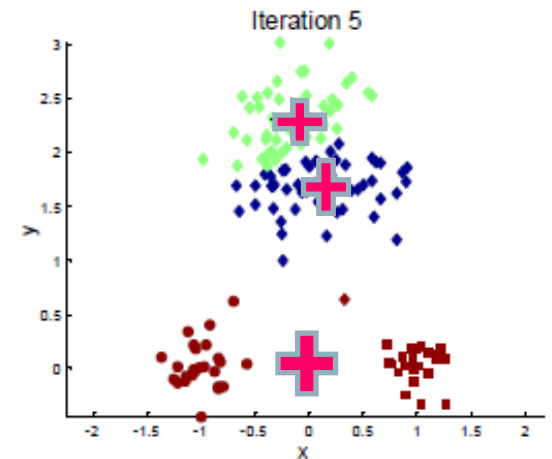
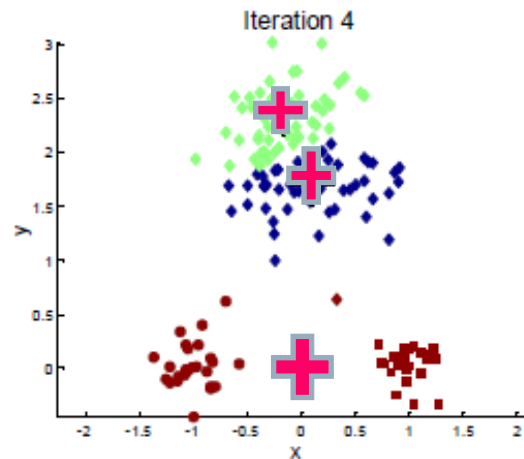
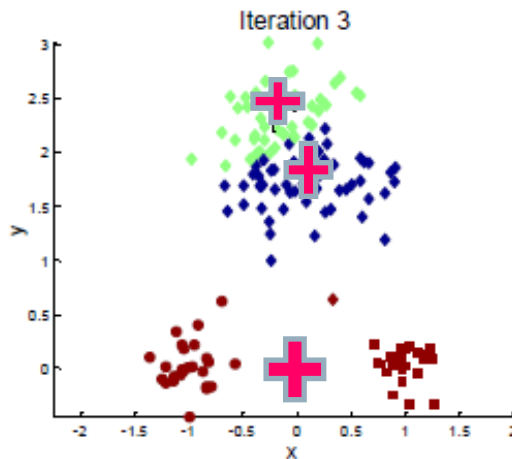
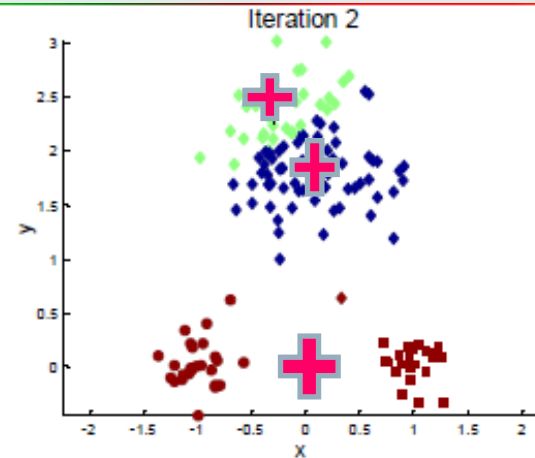
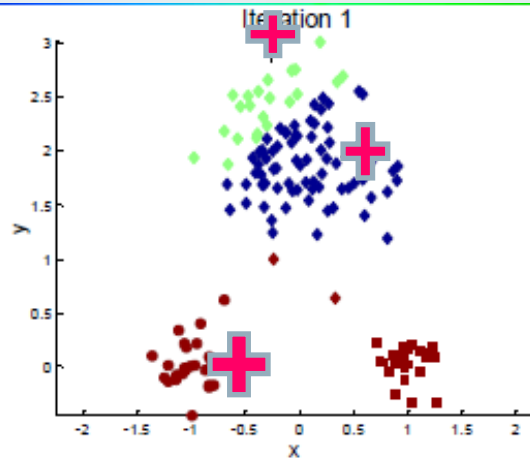
# Pentingnya Memilih Centroid Awal



# Pentingnya Memilih Centroid Awal



# Pentingnya Memilih Centroid Awal



# Mengevaluasi Cluster K-Means

- Metode pengukuran yang paling umum adalah dengan Sum of Squared Error (SSE)

- Untuk setiap titik, errornya adalah jarak ke cluster terdekat
- Untuk mendapatkan SSE, kita mengkuadratkan error dan menjumlahkannya

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

- $x$  adalah titik data pada cluster  $C_i$  dan  $m_i$  adalah titik yang mewakili cluster  $C_i$ . Dapat dikatakan  $m_i$  adalah pusat dari  $C_i$
- Jika ada 2 cluster, kita dapat memilih yang memiliki error terkecil
- Satu cara untuk mengurangi SSE adalah dengan menaikkan  $K$  (jumlah cluster)
  - Tapi ingat, clustering yang baik dengan nilai  $K$  yang lebih kecil dapat memiliki nilai SSE yang lebih rendah dari clustering yang buruk dengan nilai  $K$  yang lebih besar

# Latihan

- Misal Anda diminta untuk mengelompokkan variabel *usia* berikut ke dalam 3 kelompok: 18, 22, 25, 42, 27, 43, 33, 35, 56, 28
1. Gunakan K-Means untuk menunjukkan prosedur clustering langkah-per-langkah dengan centroid awal: 22, 35, dan 43
  2. Hitung nilai SSE-nya

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



# Solusi Terhadap Permasalahan Centroid Awal

- Menjalankan algoritma beberapa kali
- Mengambil sampel dan menggunakan hierarchical clustering untuk menentukan centroid awal
- Pilih lebih dari  $k$  centroid awal lalu pilih di antara centroid awal tersebut → Pilih yang terpisah paling jauh antar centroid

# Untuk menangani Cluster Kosong

- Algoritma basic K-Means dapat menghasilkan cluster kosong
- Beberapa strategi untuk menanganinya:
  - Pilih titik yang berkontribusi paling besar terhadap nilai SSE
  - Pilih titik pada cluster yang memiliki SSE terbesar
  - Jika ada beberapa cluster kosong, ulangi langkah di atas beberapa kali

# Pre-processing dan Post-processing

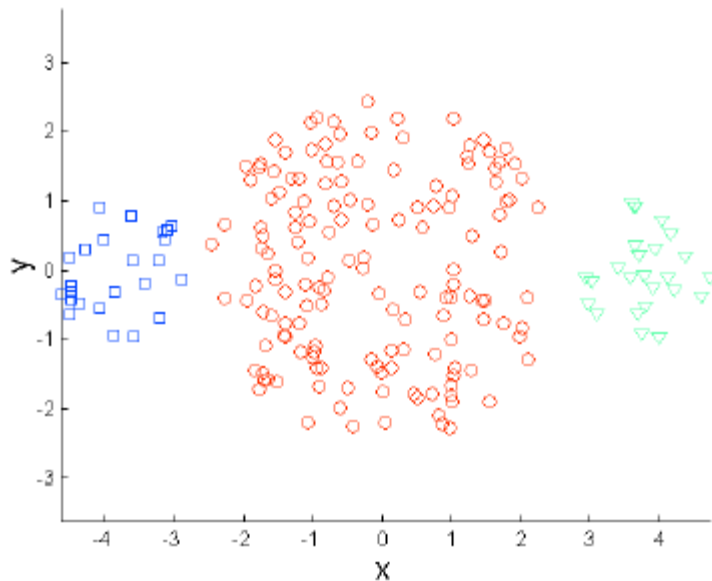
- Pre-processing
  - Normalisasi data
  - Buang outlier
- Post-processing
  - Buang cluster kecil yang mungkin mewakili outlier
  - Bagi/split cluster yang “renggang”, yakni cluster yang memiliki nilai SSE tinggi
  - Gabung cluster yang “berdekatan” dan memiliki nilai SSE rendah

# Keterbatasan K-Means

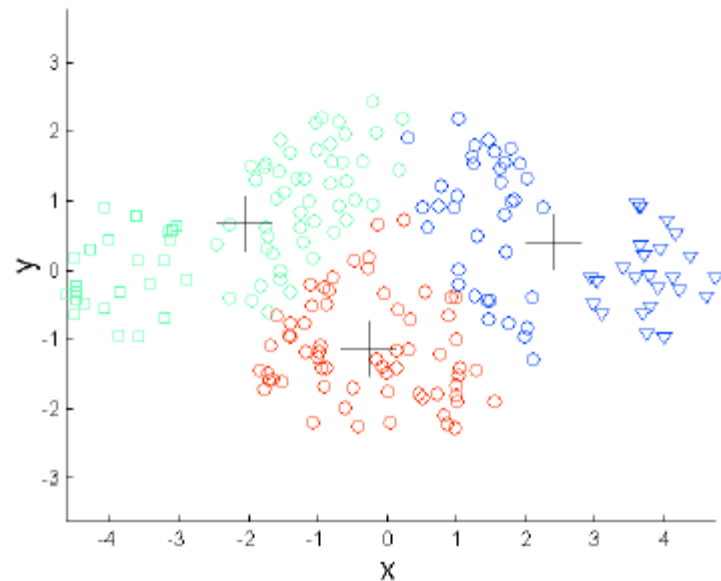
- K-Means memiliki masalah ketika cluster memiliki perbedaan:
  - Ukuran
  - Kerapatan
  - Bentuk non-globular
- K-Means memiliki masalah pada data yang memiliki outlier

# Keterbatasan K-Means

- Ketika ada perbedaan ukuran



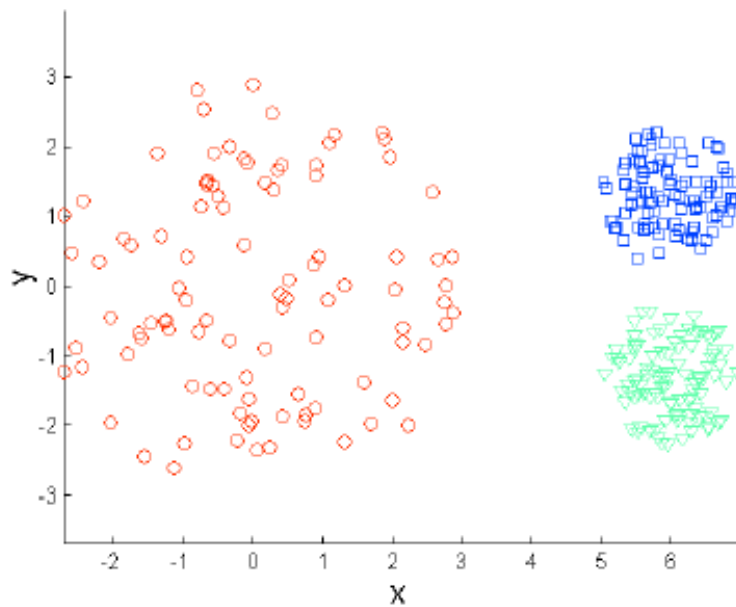
Original Points



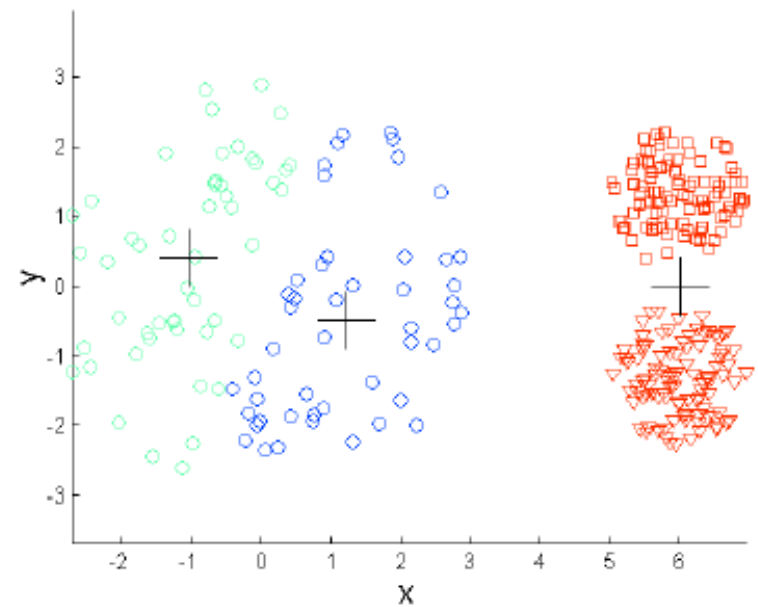
K-means (3 Clusters)

# Keterbatasan K-Means

- Ketika ada perbedaan kerapatan



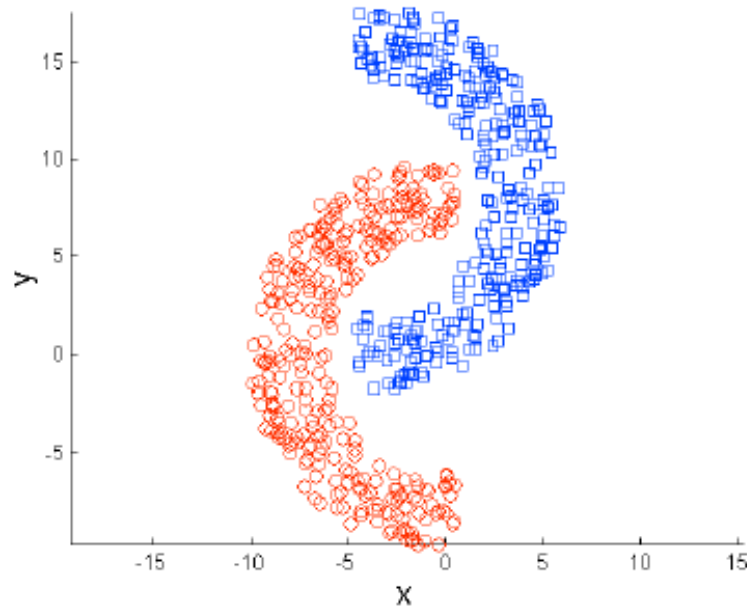
Original Points



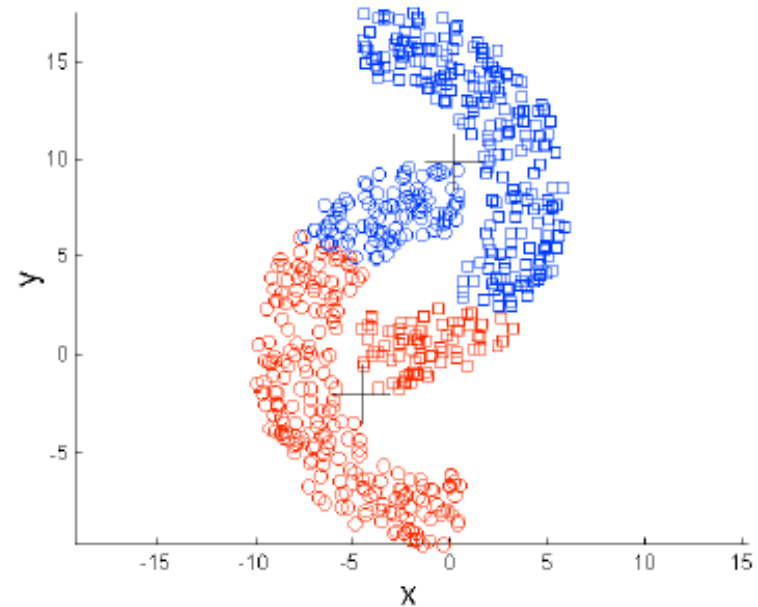
K-means (3 Clusters)

# Keterbatasan K-Means

- Ketika ada bentuk non-globular



Original Points

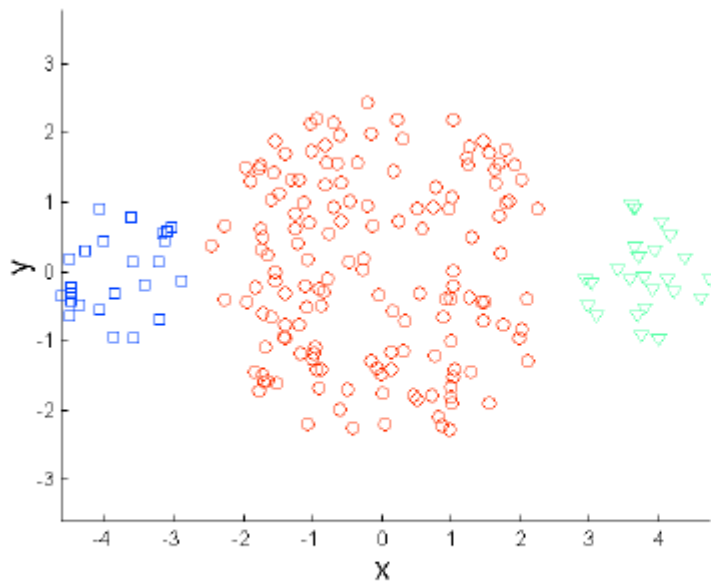


K-means (2 Clusters)

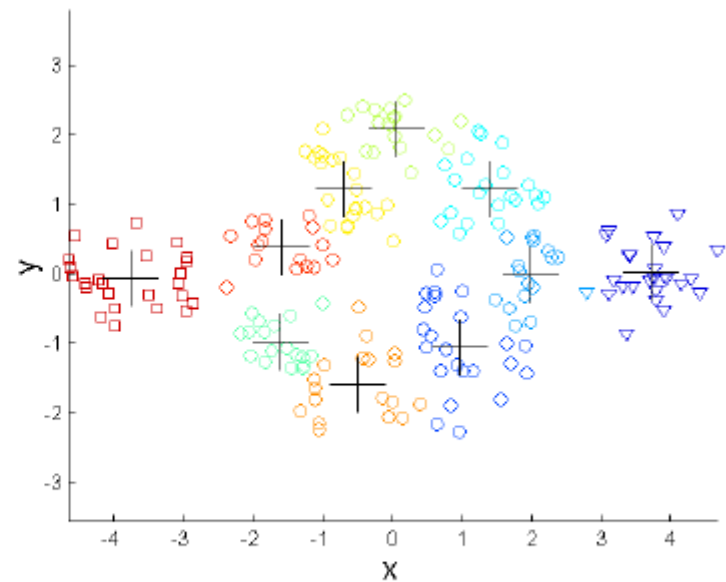


# Cara untuk menangani keterbatasan K-Means

- Menggunakan banyak cluster
  - Dapat menemukan “potongan” cluster yang perlu digabungkan



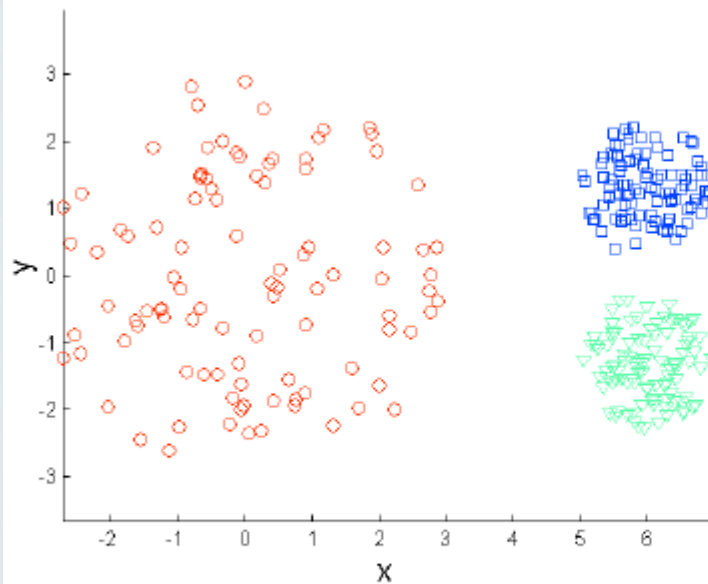
Original Points



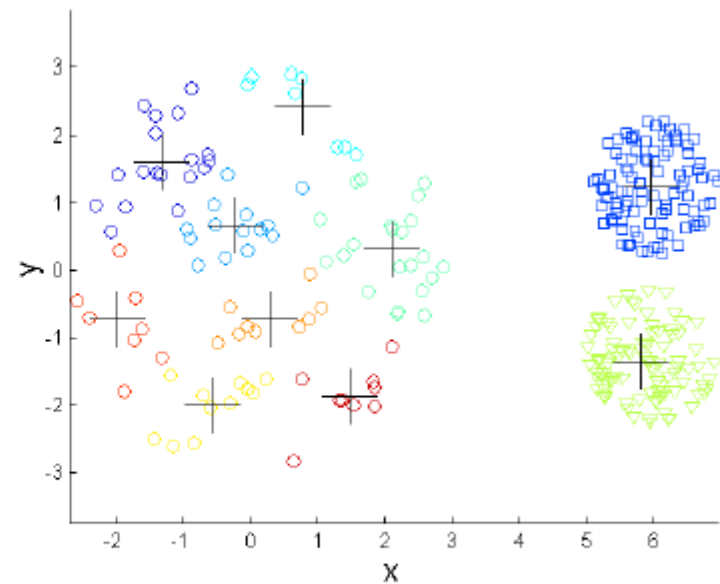
K-means Clusters

# Cara untuk menangani keterbatasan K-Means

- Menggunakan banyak cluster
  - Dapat menemukan “potongan” cluster yang perlu digabungkan



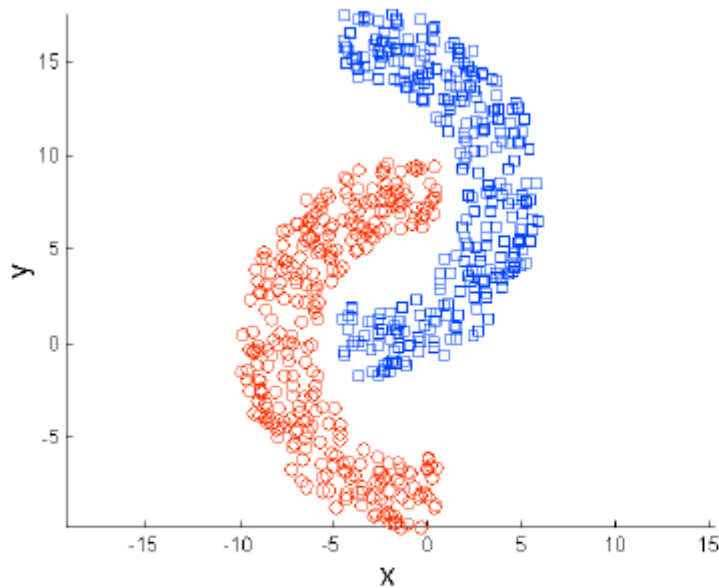
Original Points



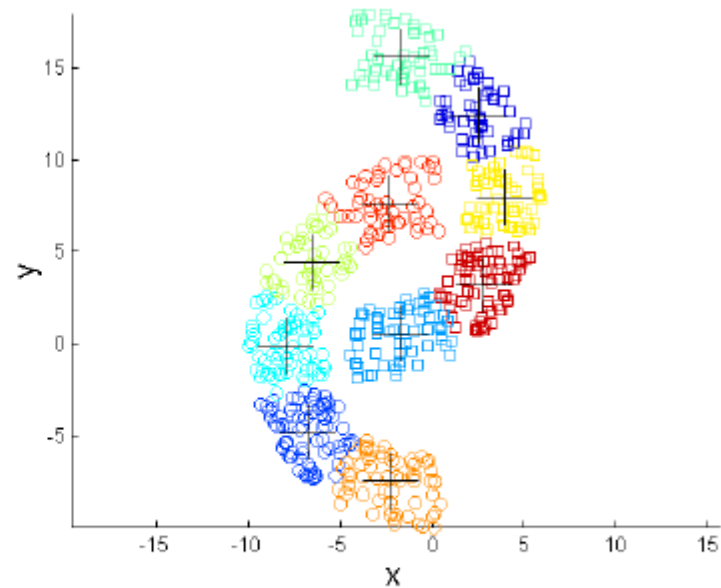
K-means Clusters

# Cara untuk menangani keterbatasan K-Means

- Menggunakan banyak cluster
  - Dapat menemukan “potongan” cluster yang perlu digabungkan



Original Points



K-means Clusters

Model Clustering:

**AGGLOMERATIVE**

**HIERARCHICAL**

**CLUSTERING**

# Kelebihan Hierarchical Clustering

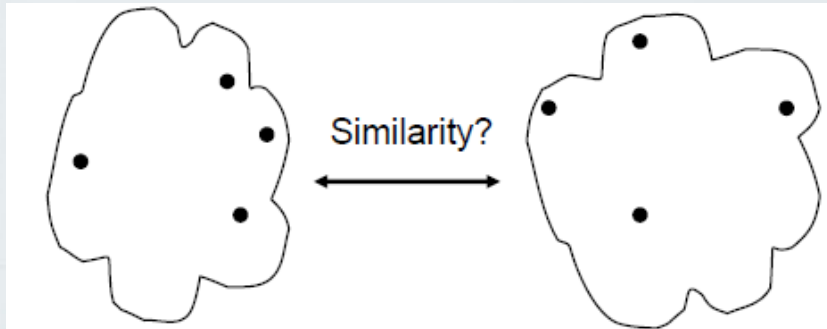
- Tidak perlu mengasumsikan jumlah cluster terlebih dahulu
  - Jumlah cluster yang diinginkan dapat diperoleh dengan “memotong” dendrogram pada level yang diinginkan
- Bentuk hierarchical clustering dapat menyerupai taksonomi (sistem klasifikasi).  
Misal: Klasifikasi hewan

## 2. Hierarchical Clustering

- Dua tipe utama hierarchical clustering:
  - Agglomerative (penggabungan)
    - Mulai dari titik-titik sebagai cluster individu
    - Pada tiap tahap, gabungkan (*merge*) pasangan cluster terdekat sampai terbentuk 1 (atau  $k$ ) cluster
  - Divisive (pembagian)
    - Mulai dari 1 cluster besar
    - Pada tiap tahap, pisahkan (*split*) sebuah cluster hingga tiap cluster memiliki 1 titik (atau sampai terdapat  $k$  cluster)
- Algoritma hierarchical clustering tradisional menggunakan matriks jarak atau similaritas

# Agglomerative Hierarchical Clustering

- Menggunakan matriks jarak sebagai pengukur kriteria pengelompokan



- Ada empat metode:
  - Min (single linkage)
  - Max (complete linkage)
  - Rata-rata grup
  - Jarak antar-centroid

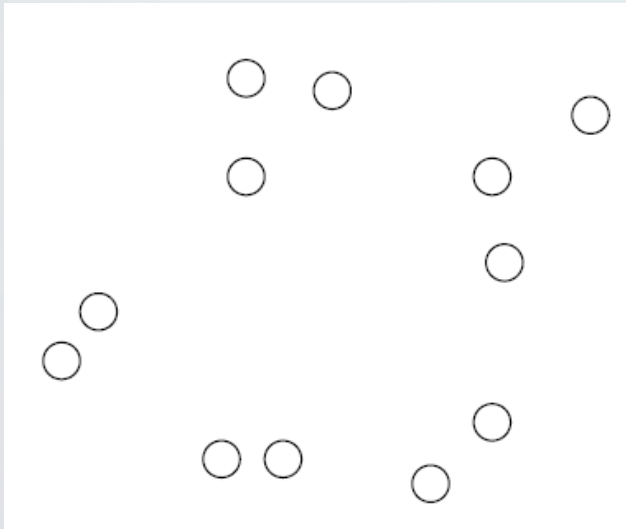
# Algoritma Agglomerative Clustering

- Algoritma dasar:
  - Hitung matriks jarak
  - Misal tiap titik data adalah sebuah cluster
  - Ulangi
    - Gabungkan 2 cluster terdekat
    - Update matriks jarak
  - Sampai tersisa 1 cluster
- Operasi dasar: perhitungan jarak antara 2 cluster



# Keadaan Awal

- Mulai dari cluster dari titik individu dan matriks jarak



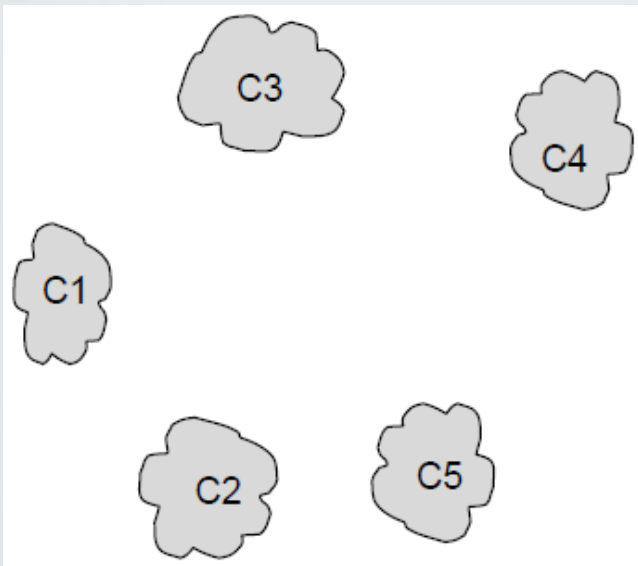
	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix



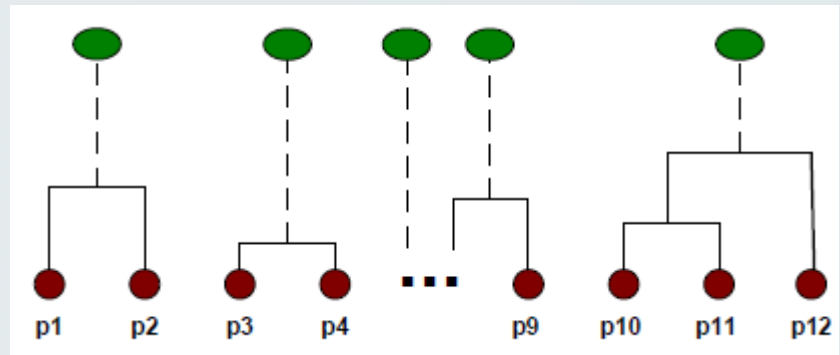
# Keadaan Intermediate

- Setelah beberapa tahap merging, terdapat sejumlah cluster



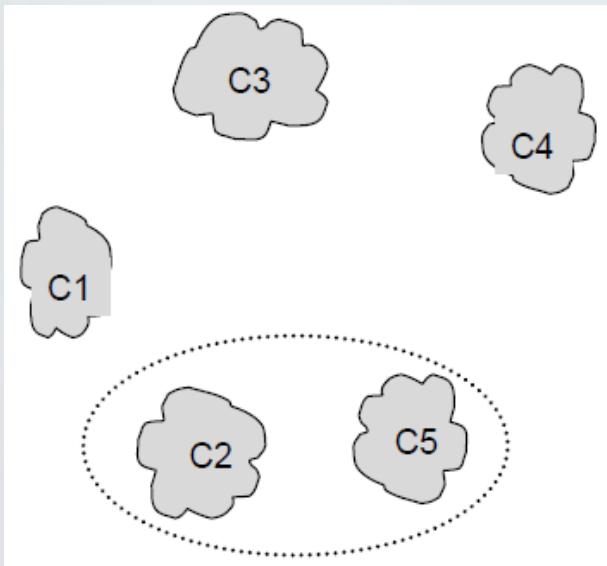
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



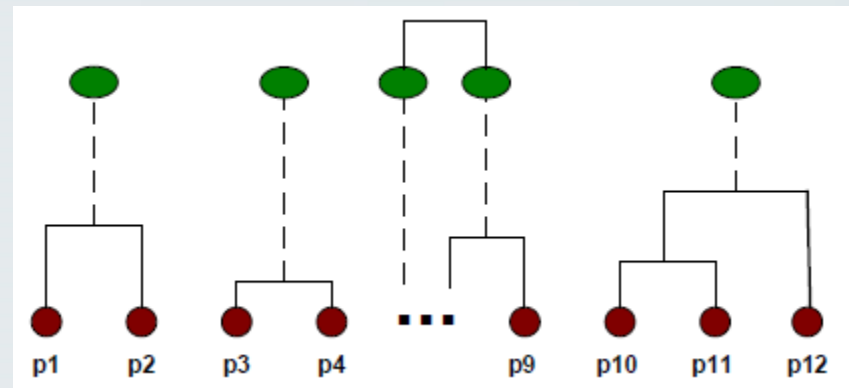
# Keadaan Intermediate

- Kita ingin menggabungkan 2 cluster terdekat (C2 dan C5) dan mengupdate matriks jarak



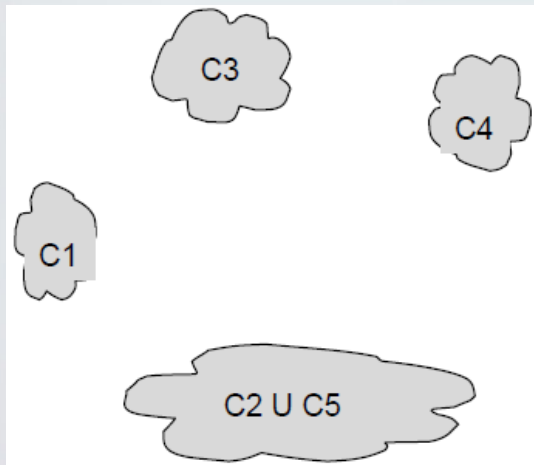
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



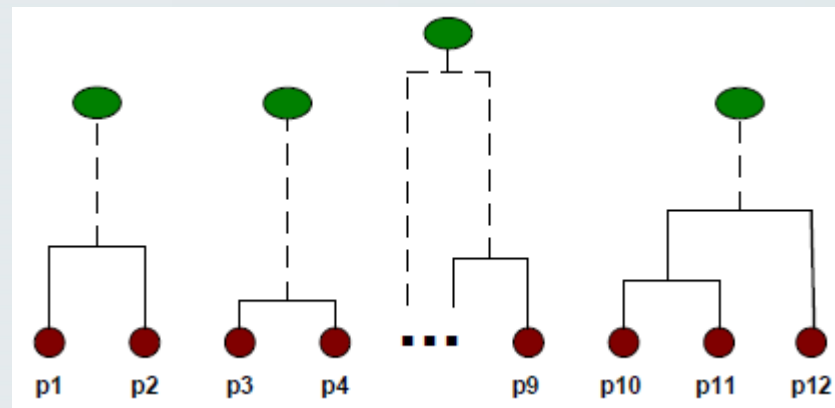
# Setelah Penggabungan

- Bagaimana cara mengupdate matriks jarak?

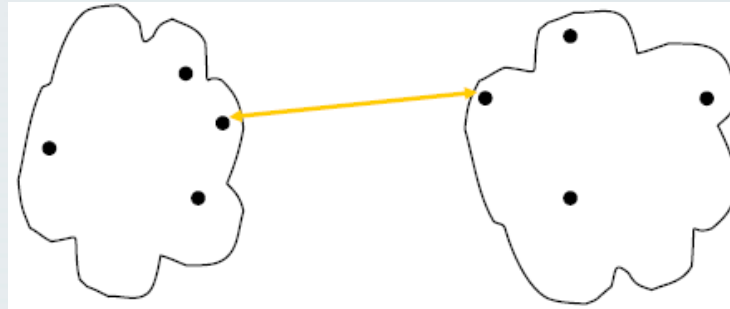


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix

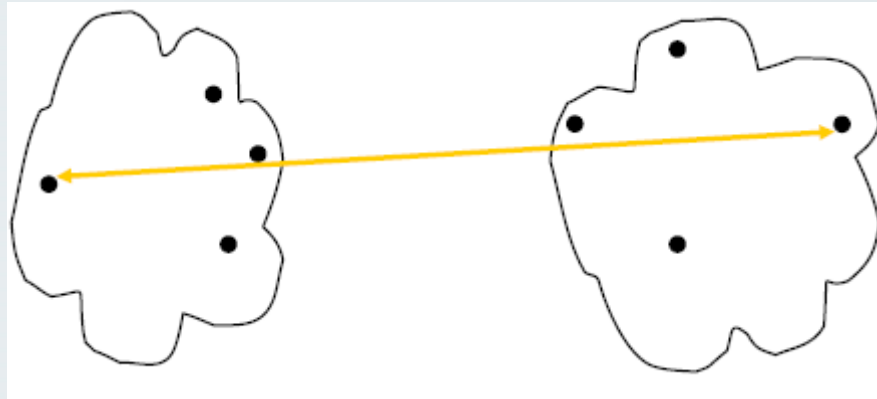


# Bagaimana Mendefinisikan Kesamaan Antar-Cluster



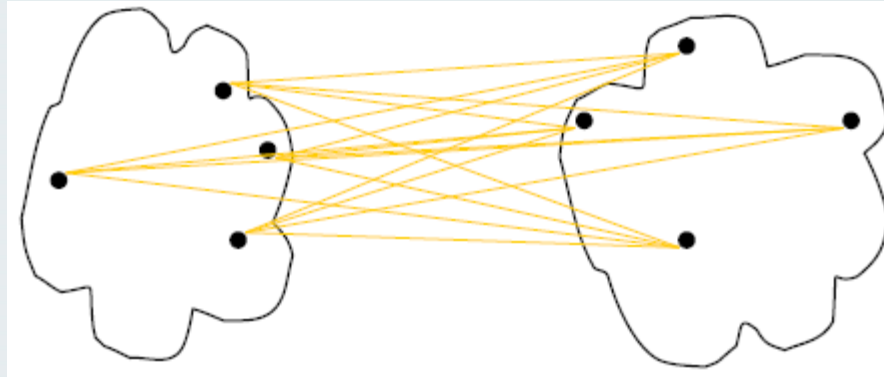
- **Min (single linkage)**
  - Dua titik terdekat (termirip) pada cluster yang berbeda
  - Ditentukan oleh 1 pasang titik, yakni 1 link dalam graph jarak

# Bagaimana Mendefinisikan Kesamaan Antar-Cluster



- **Max (complete linkage)**
  - Dua titik yang paling tidak mirip (paling jauh) pada cluster yang berbeda
  - Ditentukan oleh semua pasangan titik dari kedua cluster

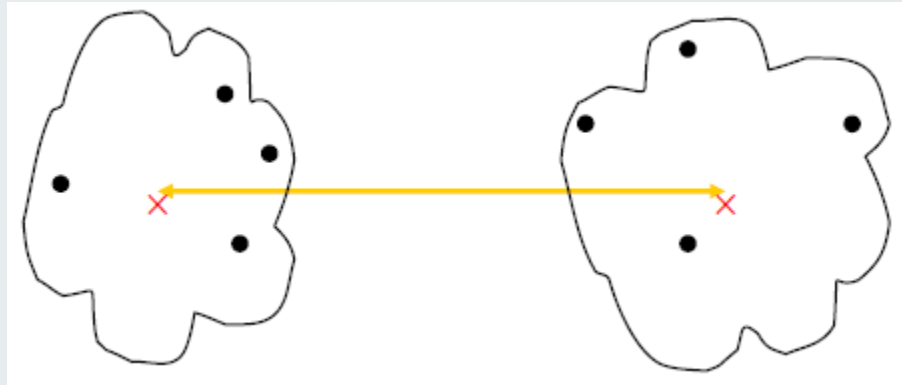
# Bagaimana Mendefinisikan Kesamaan Antar-Cluster



- Rata-rata grup
  - Rata-rata jarak dari setiap pasang titik di antara kedua cluster

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

# Bagaimana Mendefinisikan Kesamaan Antar-Cluster

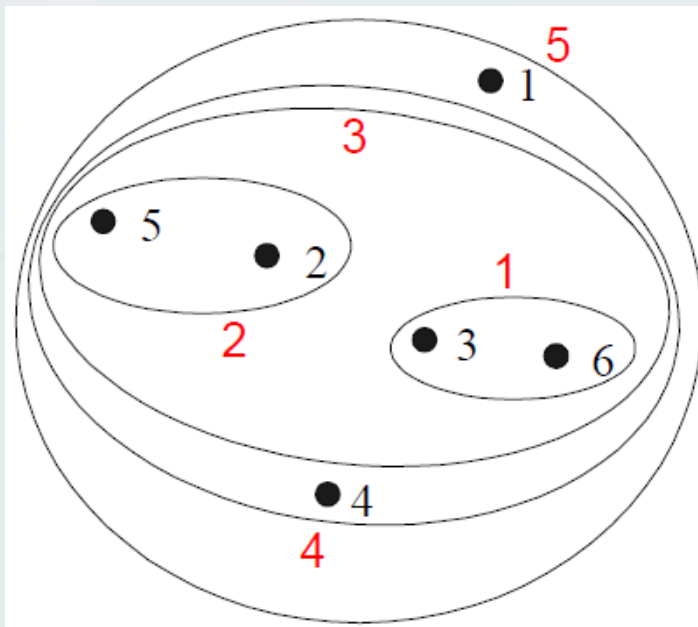


- Jarak antara centroid

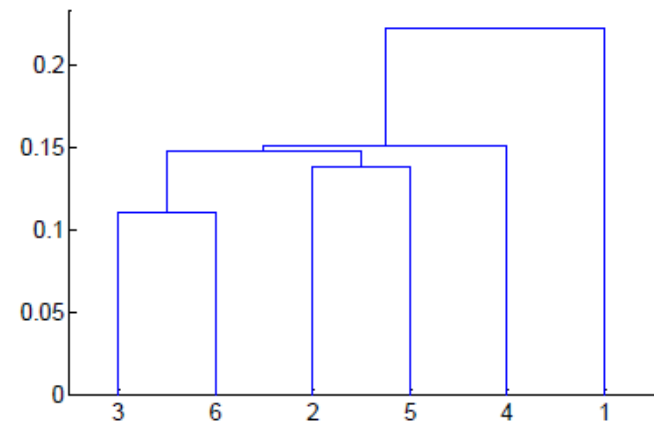


# Min (Single Linkage)

- Kesamaan antara dua cluster didasarkan atas dua titik yang paling dekat (mirip) pada cluster yang berbeda



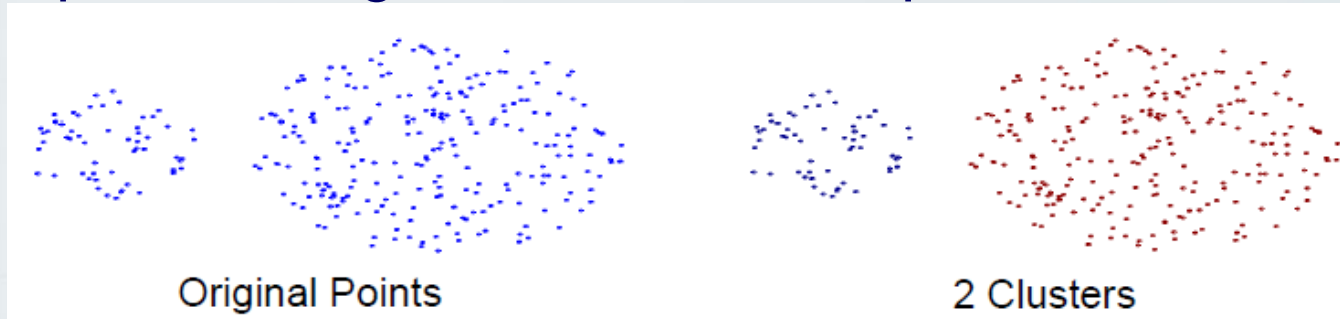
Nested Clusters



Dendrogram

# Kelebihan dan Kekurangan Min

- Kelebihan
  - Dapat menangani bentuk non-elips

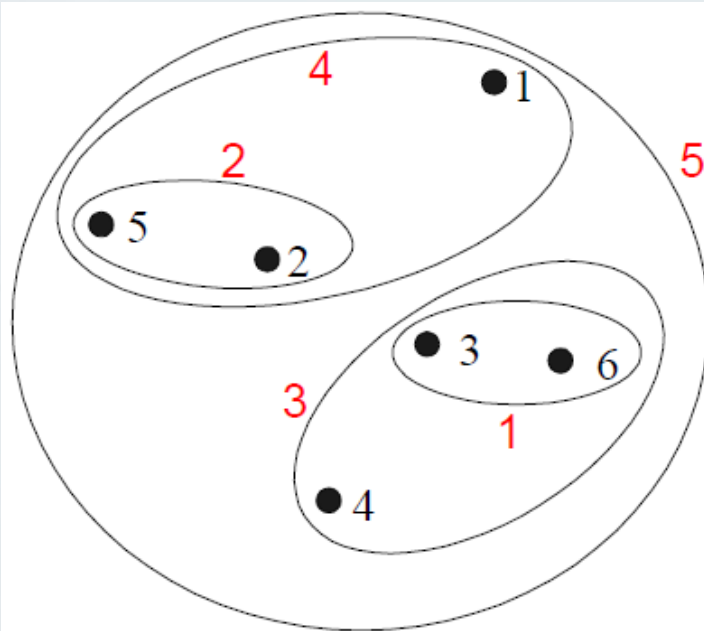


- Kekurangan
  - Sensitif terhadap noise

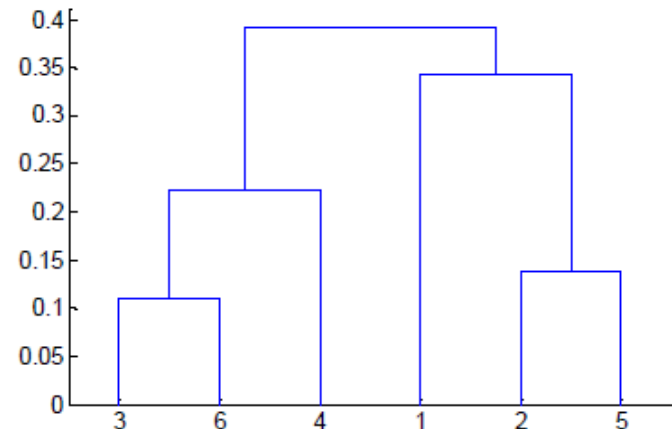


# Max (Complete Linkage)

- Kesamaan antara 2 cluster didasarkan pada dua titik yang paling tidak mirip (paling jauh) pada cluster yang berbeda



Nested Clusters



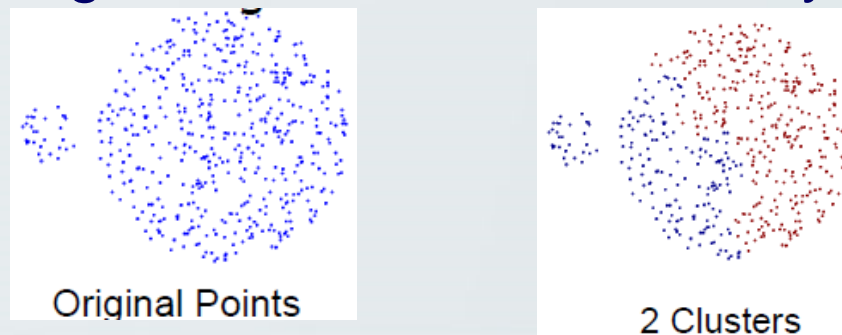
Dendrogram

# Kelebihan dan Kekurangan Max

- Kelebihan
  - Lebih tahan terhadap noise

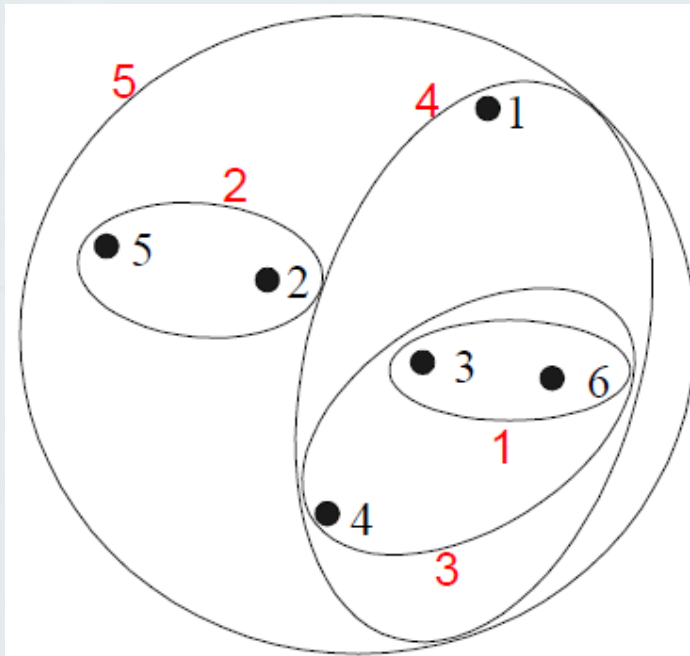


- Kekurangan
  - Cenderung untuk memecah cluster yang besar

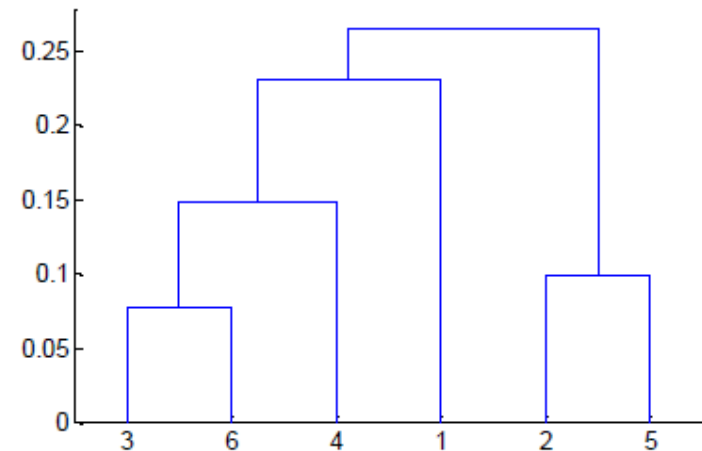


# Rata-rata Grup

- Paduan antara MIN dan MAX
- Menggunakan rata-rata jarak dari pasangan titik pada 2 cluster

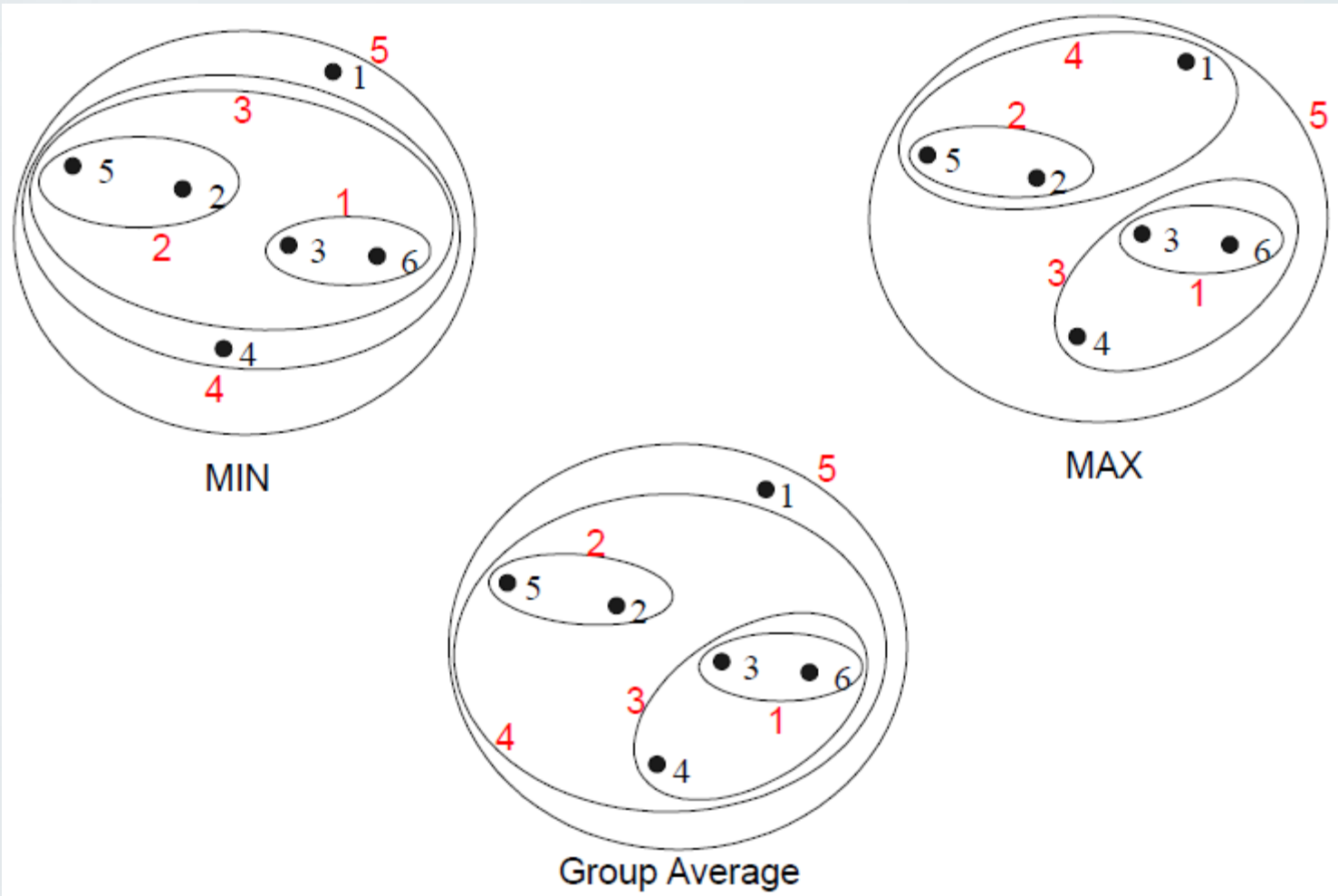


Nested Clusters



Dendrogram

# Secara Keseluruhan



# Latihan

- Misal ada sejumlah angka:  
18, 22, 25, 42, 27, dan 43,
- Gunakan algoritma Agglomerative Hierarchical Clustering untuk mengelompokkan angka tersebut
- Gunakan metode **min** untuk menggabungkan dua cluster terdekat dan meng-update matriks jarak

# Permasalahan dan Keterbatasan pada Hierarchical Clustering

- Sekali telah diputuskan untuk menggabung 2 cluster, tidak dapat dibatalkan
- Tidak ada minimalisasi langsung dari fungsi objektif
- Tergantung dari metode yang digunakan untuk menggabungkan 2 cluster, memiliki kekurangan sebagai berikut:
  - Sensitif terhadap noise dan outlier
  - Kesulitan dalam menangani ukuran cluster yang berbeda dan memiliki bentuk cekung
  - Memecah cluster besar



# Rumus Matematika

- Euclidean Distance

$$d(P, Q) = \sqrt{\sum (p_i - q_i)^2}$$

- Cosine Similarity

$$s(P, Q) = \frac{\sum p_i \cdot q_i}{\sqrt{\sum p_i^2 \times \sum q_i^2}}$$

- Nilai Cosine Similarity berkisar antara -1 (bertolak belakang) sampai 1 (sama persis)

# The Big Picture:

Classification vs Clustering vs Association vs Regression

- Studi Kasus:

- Sebuah dealer mobil Toyota ingin meningkatkan penjualan. Dealer tersebut telah memiliki catatan histori penjualan dan informasi tentang setiap orang yang membeli Toyota, melihat Toyota, dan mengunjungi showroom Toyota

# The Big Picture:

## Classification vs Clustering vs Association vs Regression

- Regression (digunakan untuk mencari sebuah perkiraan/estimasi sebuah nilai numerik)
  - Berapa harga yang sesuai untuk Toyota New Fortuner?



- Berdasarkan data penjualan lampau, harga terkini dapat dihitung dengan mempertimbangkan beberapa variabel, seperti nilai dari tiap fitur, tingkat inflasi, dsb

# The Big Picture:

## Classification vs Clustering vs Association vs Regression

- Classification

- Berapa kemungkinan seorang calon customer X membeli New Fortuner?
  - Berdasarkan atribut usia, pendapatan, jumlah mobil yang dimiliki saat ini, status pernikahan, jumlah anak, dll

# The Big Picture:

Classification vs Clustering vs Association vs Regression

- Clustering
  - Grup usia berapa yang menyukai New Fortuner warna silver metalik?



# The Big Picture:

## Classification vs Clustering vs Association vs Regression

- Association

- Jika seorang customer membeli New Fortuner, opsi lain apalagi yang mereka cenderung beli pada saat yang bersamaan?
  - Misal didapat data sarung stir+persneling, tambahan bagasi atap

# Next Generation of Data Mining

- Graph Mining

- Social Network Analysis

- Link Prediction → Friend suggestion pada Facebook → Edge apa yang perlu ditambahkan pada graph?

- Spatial Data Mining

- Spatial association

$is\_a(X, "school") \wedge close\_to(X, "sports\_center") \Rightarrow close\_to(X, "park") \quad [0.5\%, 80\%].$



# Next Generation of Data Mining

- Multimedia Data Mining
  - Image, Audio, and Video Data Mining
    - Klasifikasi genre musik
    - Deteksi similaritas gambar dan video
- Web Mining
  - Personalized Web agents
    - Automatic news filtering and retrieval