Web Mining

Cluster II (berbasis density)

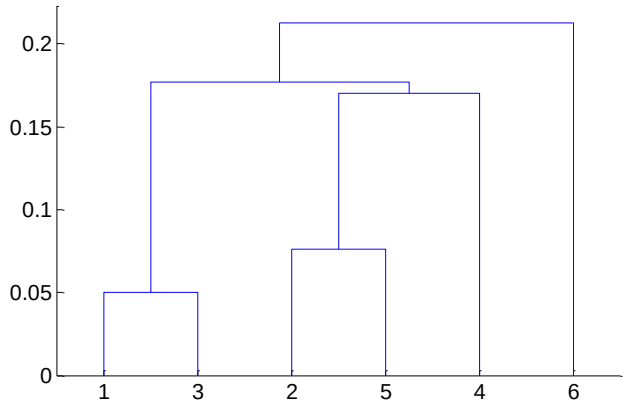
Prodi Teknik Informatika

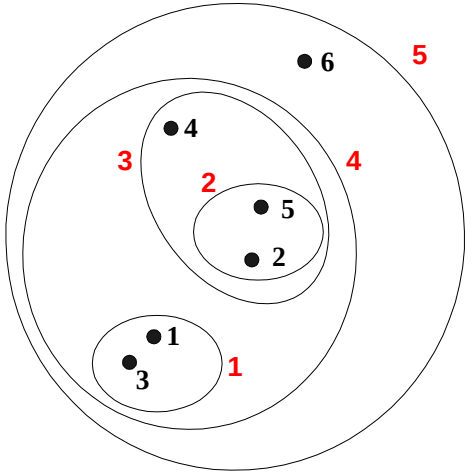
Universitas Trunojoyo Madura

2024

Hierarchical Clustering

- Menghasilkan sekumpulan kluster berlapis yang disusun sebagai pohon hierarkis
- Dapat divisualisasikan sebagai dendrogram
  - Diagram seperti pohon yang mencatat urutan penggabungan atau pemisahan





Strengths of Hierarchical Clustering

- Tidak harus mengasumsikan jumlah cluster tertentu
  - Jumlah cluster yang diinginkan dapat diperoleh dengan 'memotong' dendrogram pada tingkat yang tepat
- Mereka mungkin sesuai dengan taksonomi yang bermakna
  - Contoh dalam ilmu biologi (misalnya, kerajaan hewan, rekonstruksi filogeni, ...)

Hierarchical Clustering

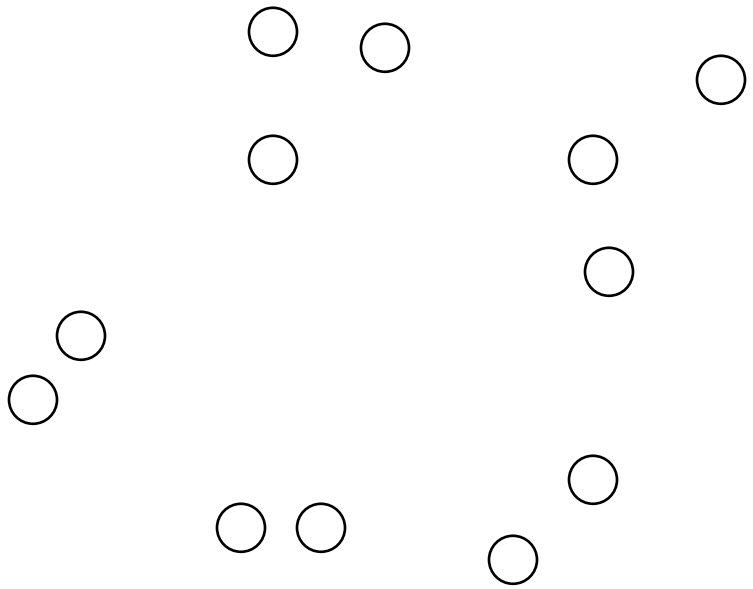
- Dua jenis utama pengelompokan hierarkis
  - Agglomerative:
    - Mulailah dengan titik sebagai kluster individual
    - Pada setiap langkah, gabungkan pasangan kluster terdekat hingga hanya satu kluster (atau k kluster) yang tersisa
  - Divisive:
    - Mulai dengan satu klaster lengkap
    - Pada setiap langkah, pisahkan kluster hingga setiap kluster berisi titik individual (atau ada k kluster)
- Algoritma hierarkis tradisional menggunakan matriks kesamaan atau jarak
  - Gabungkan atau pisahkan satu kluster pada satu waktu

Agglomerative Clustering Algorithm

- Teknik pengelompokan hierarkis paling populer
- Algoritma dasar sangat mudah
  1. Hitung matriks kedekatan
  2. Biarkan setiap titik data menjadi kluster
  3. **Repeat**
  4. Gabungkan dua kluster terdekat
  5. Memperbarui matriks kedekatan
  6. **Until** hanya satu kluster yang tersisa
- Operasi kunci adalah perhitungan kedekatan dua kluster
  - Pendekatan yang berbeda untuk menentukan jarak antar kluster membedakan algoritma yang berbeda

Starting Situation

- Mulailah dengan kelompok titik individual dan matriks kedekatan

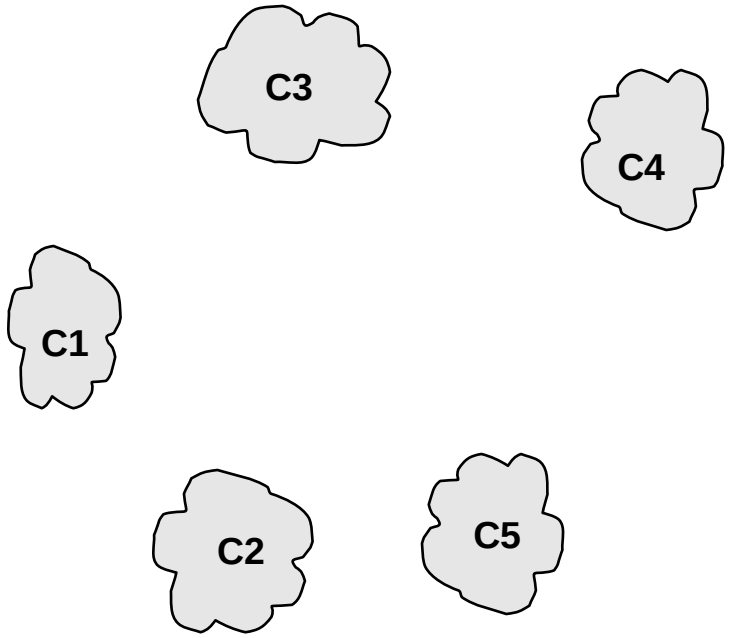


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

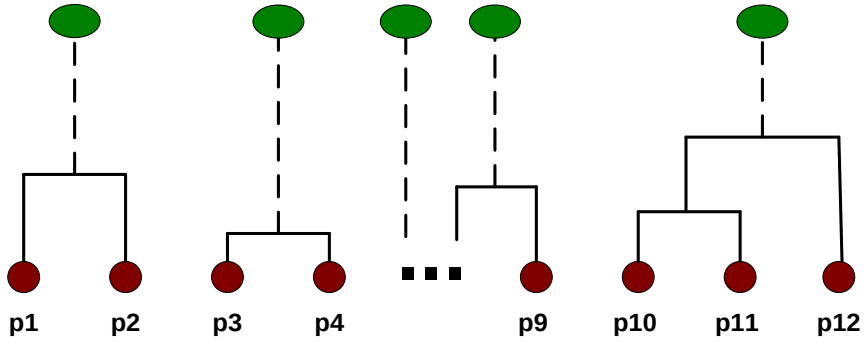
Intermediate Situation

- Setelah beberapa langkah penggabungan, kita memiliki beberapa cluster



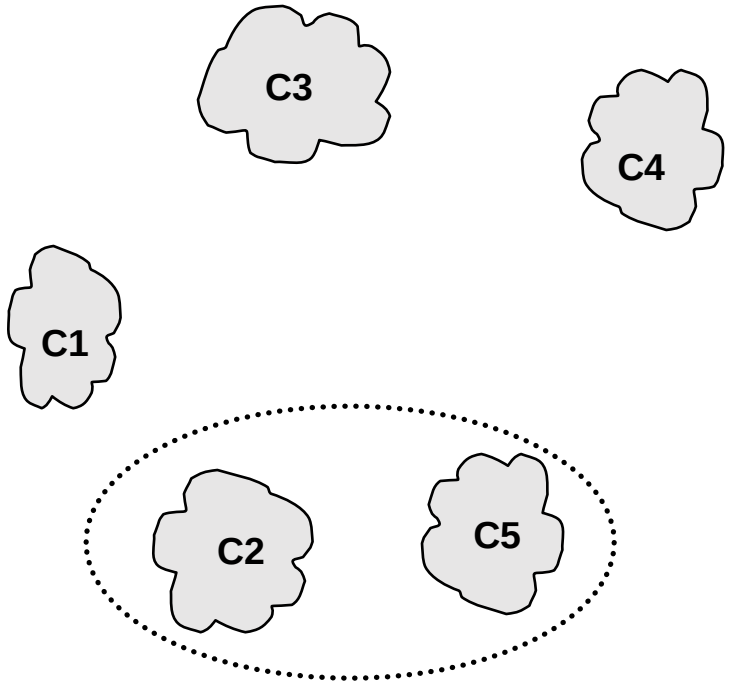
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



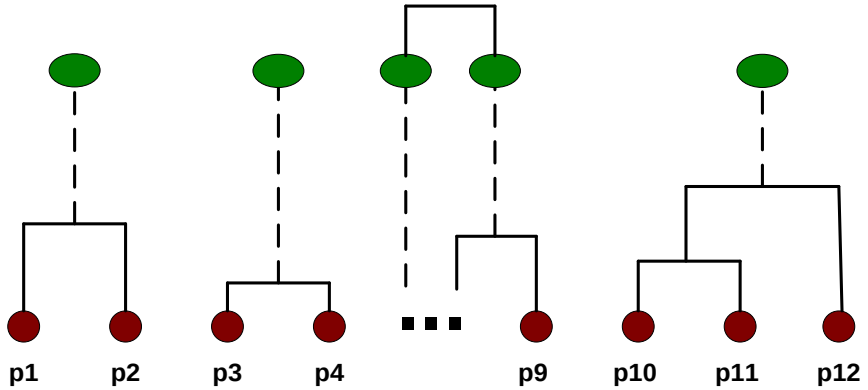
Intermediate Situation

- Kita ingin menggabungkan dua kluster terdekat (C2 dan C5) dan memperbarui matriks kedekatan.



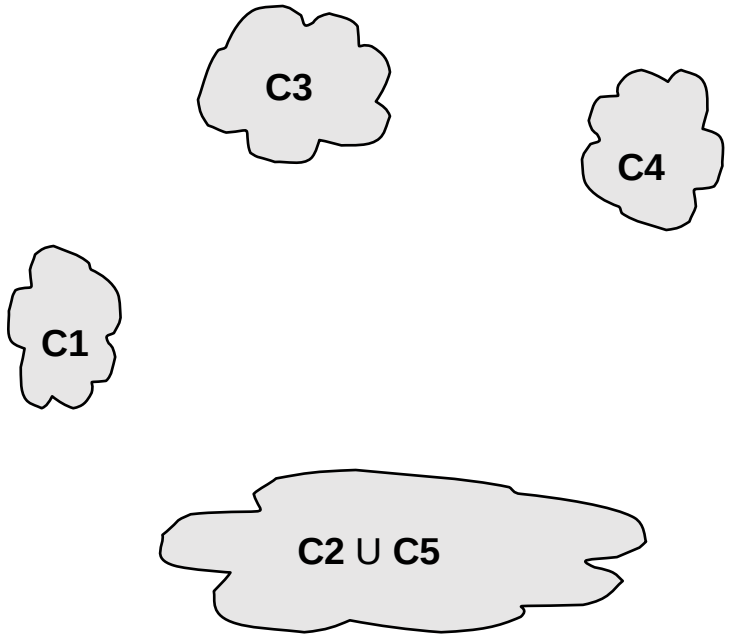
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



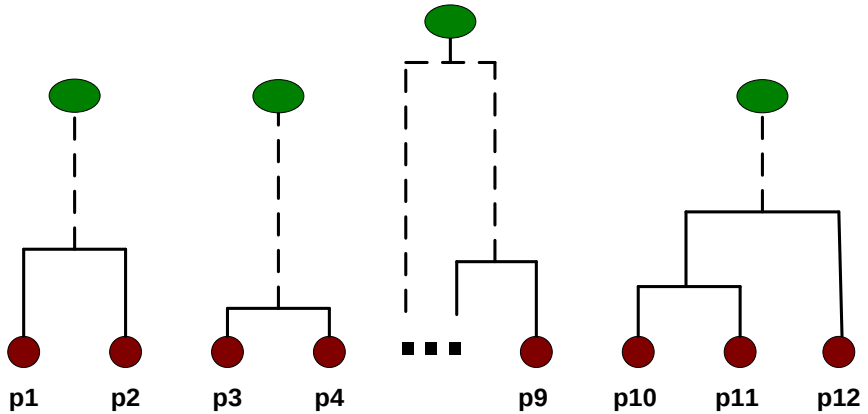
After Merging

- Pertanyaannya adalah "Bagaimana kita memperbarui matriks kedekatan?"

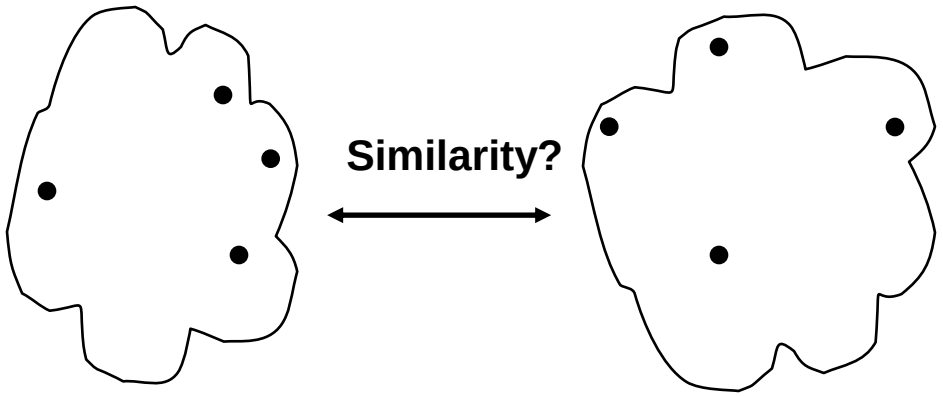


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



How to Define Inter-Cluster Distance

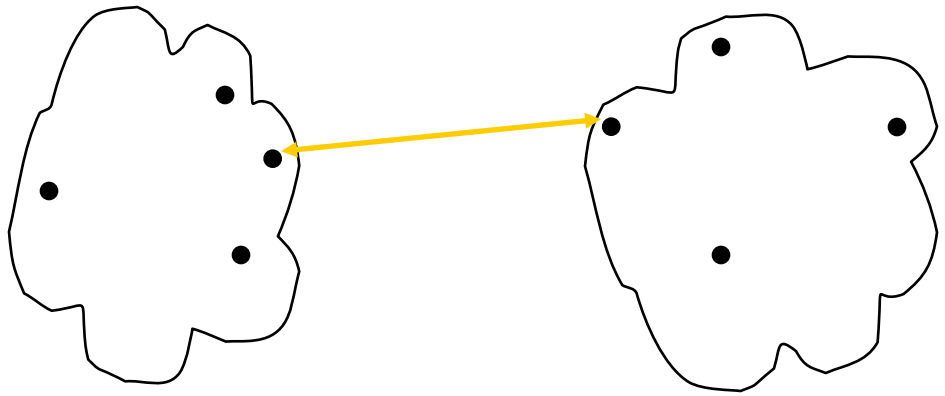


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

How to Define Inter-Cluster Similarity

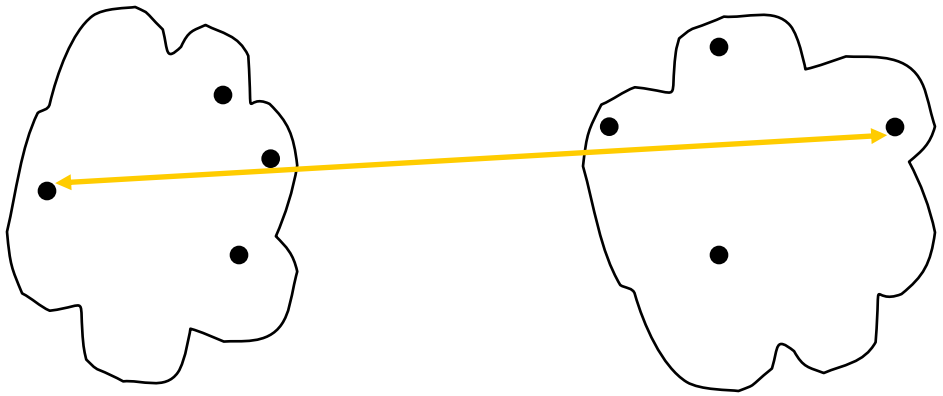


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

How to Define Inter-Cluster Similarity

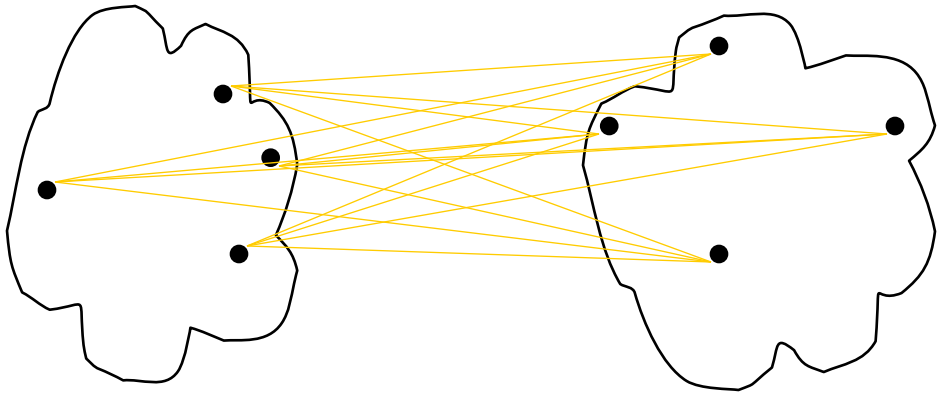


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

How to Define Inter-Cluster Similarity

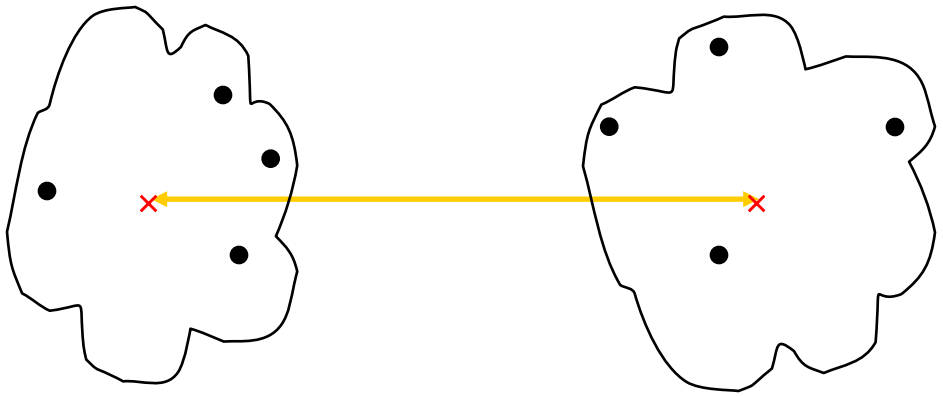


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

How to Define Inter-Cluster Similarity



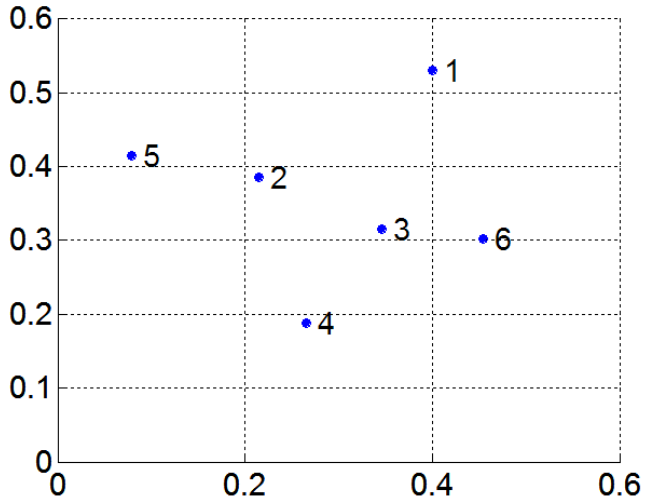
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

MIN or Single Link

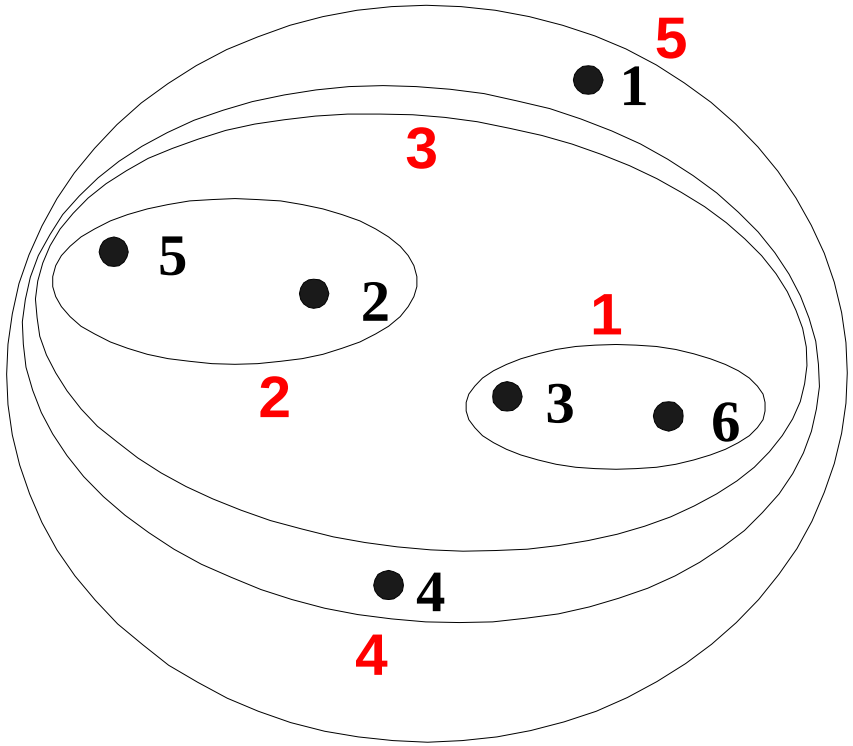
- Kedekatan dua kluster didasarkan pada dua titik terdekat dalam kluster yang berbeda
  - Ditentukan oleh satu pasang titik, yaitu oleh satu tautan dalam grafik kedekatan
- Example:



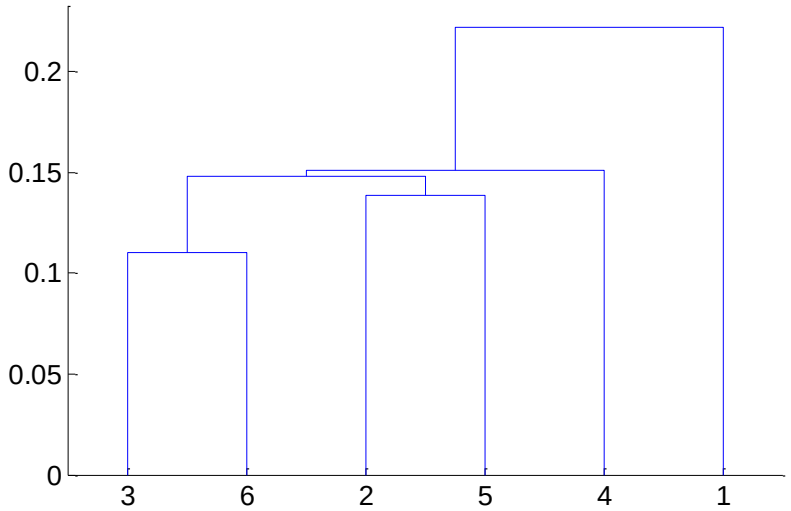
Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Hierarchical Clustering: MIN

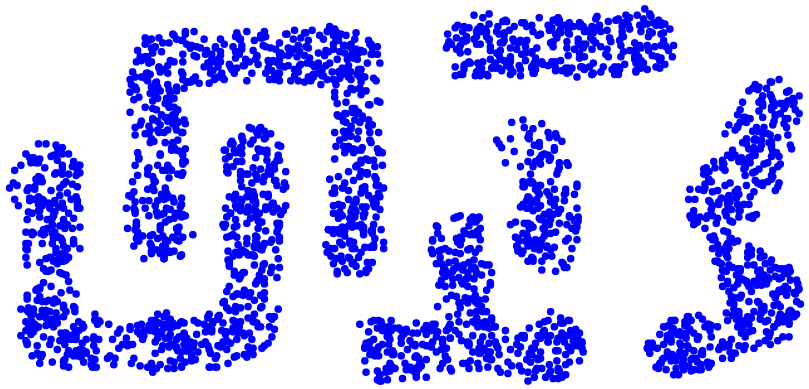


Nested Clusters

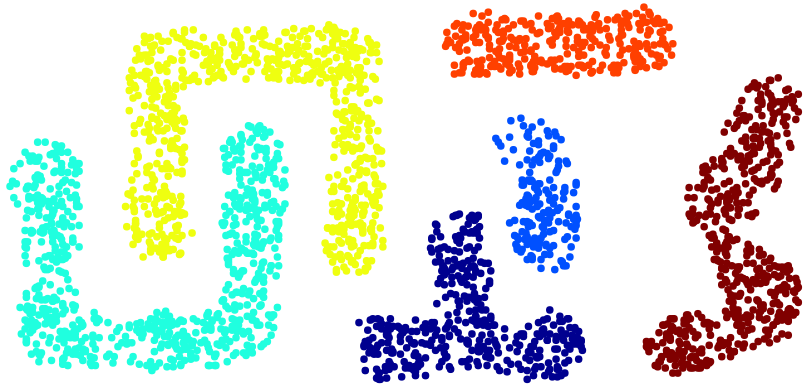


Dendrogram

Strength of MIN



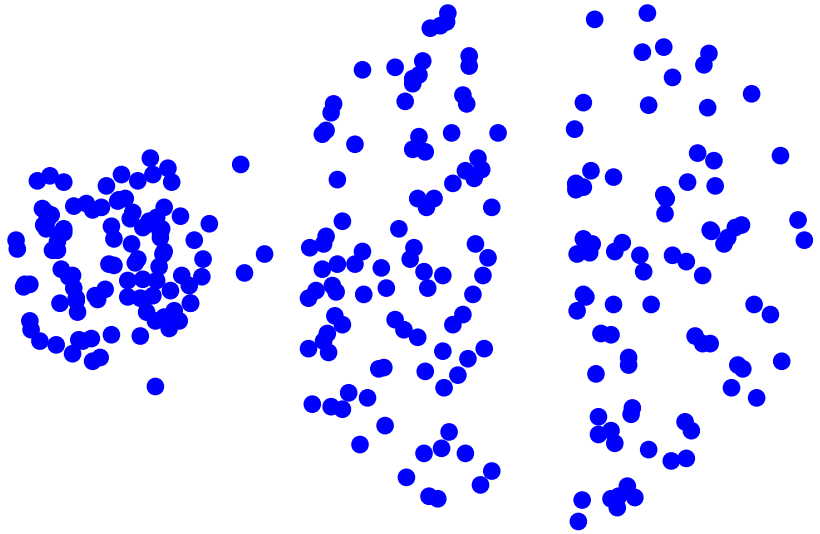
Original Points



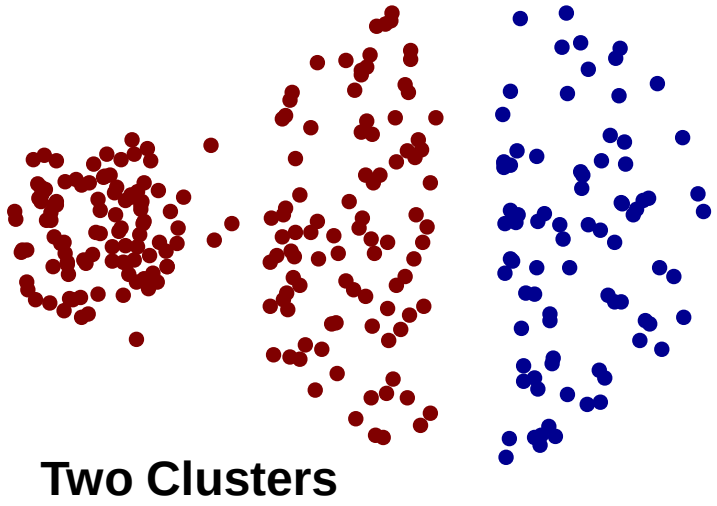
Six Clusters

- **Can handle non-elliptical shapes**

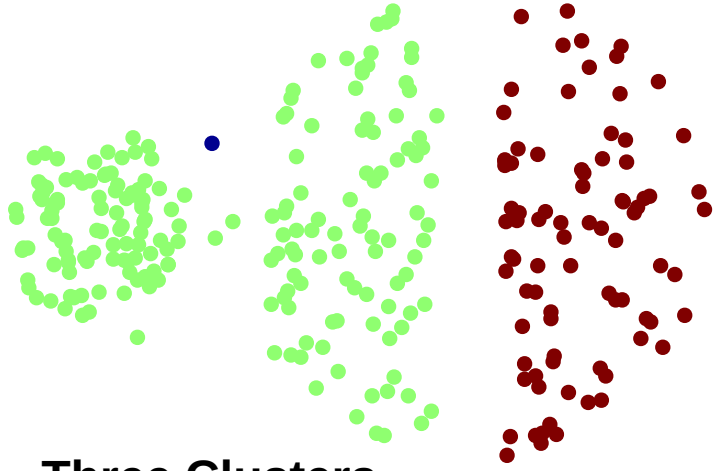
Limitations of MIN



Original Points



Two Clusters

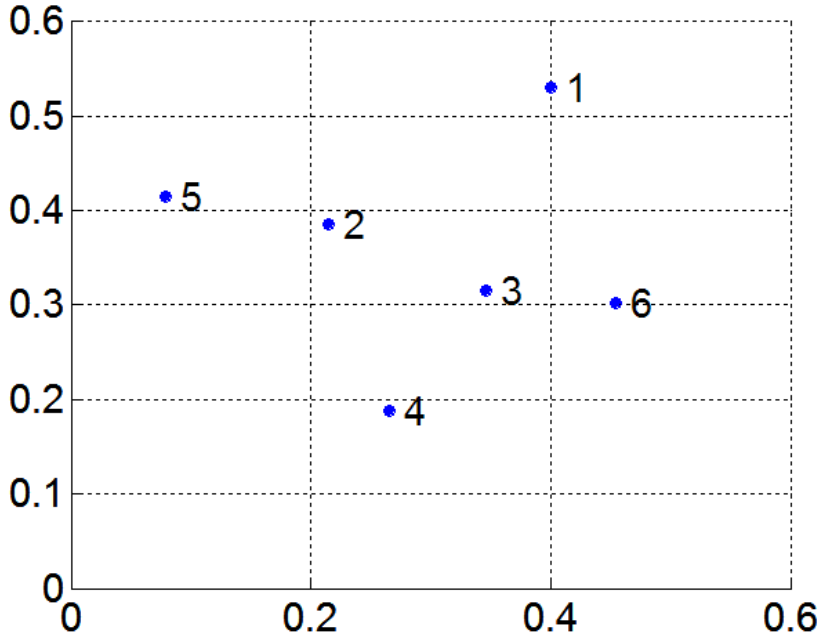


Three Clusters

- Sensitive to noise and outliers

MAX or Complete Linkage

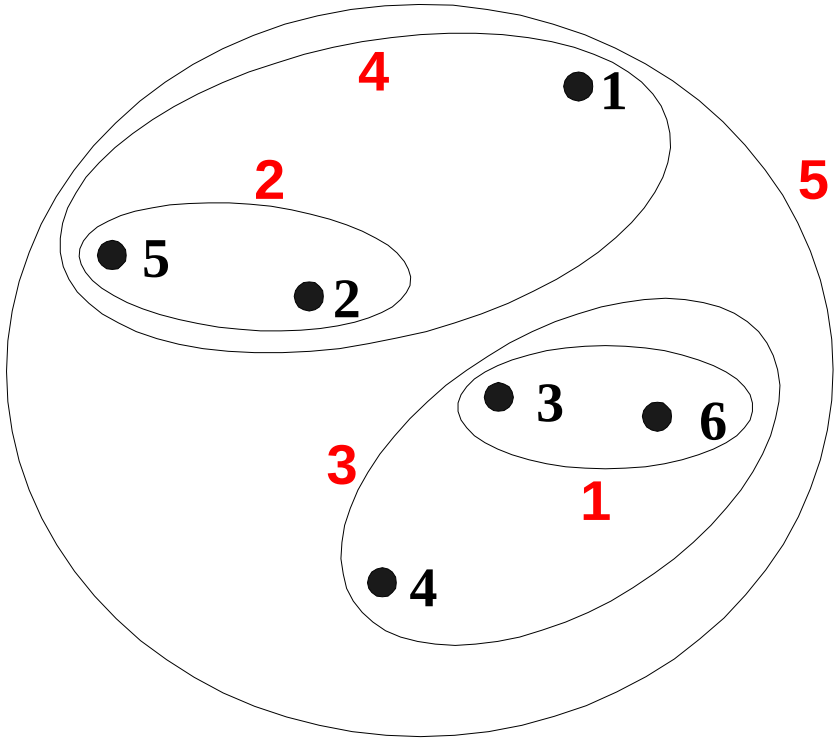
- Kedekatan dua gugus didasarkan pada dua titik terjauh dalam kluster yang berbeda
  - Ditentukan oleh semua pasangan titik dalam dua gugus



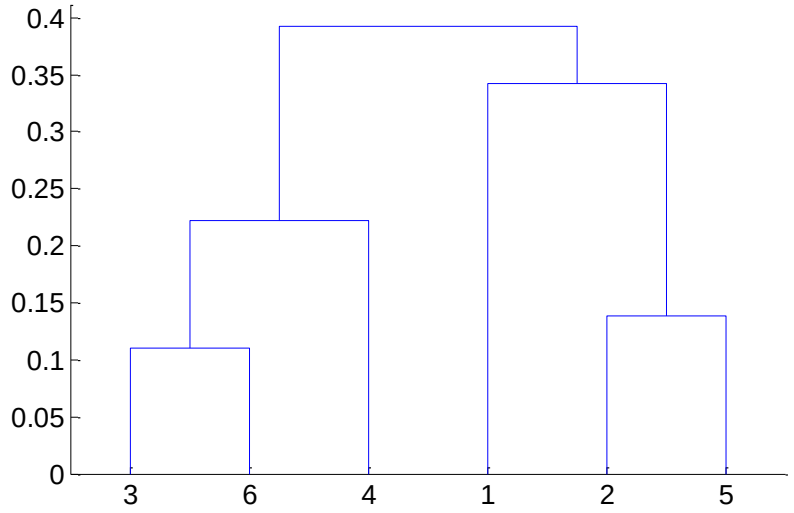
Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Hierarchical Clustering: MAX

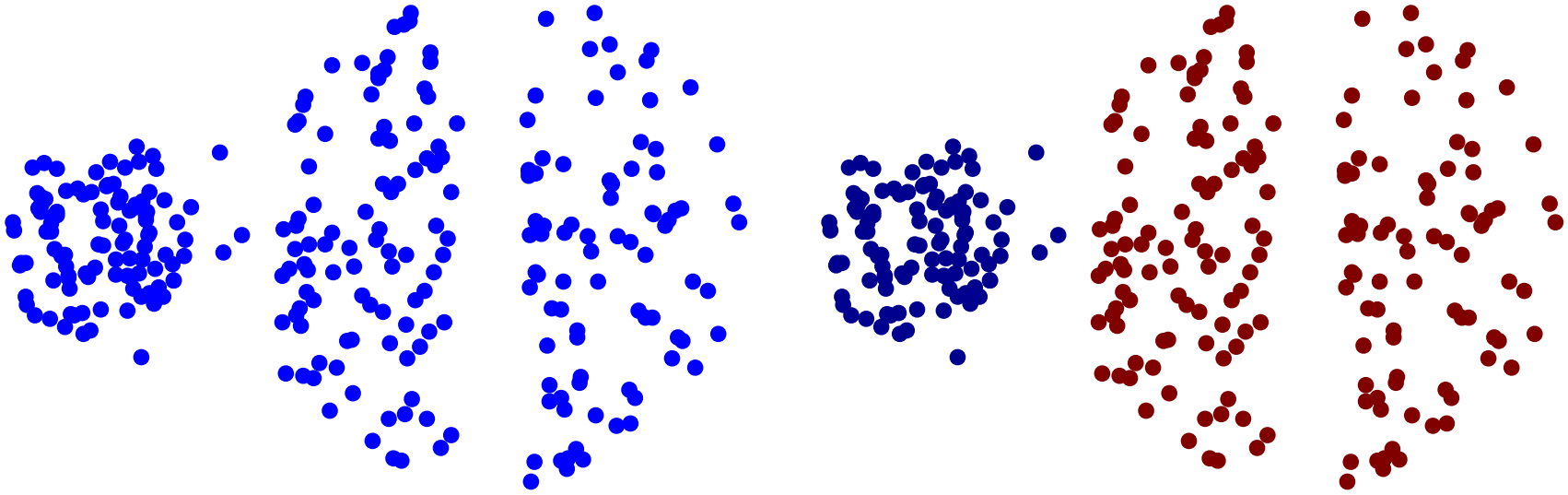


Nested Clusters



Dendrogram

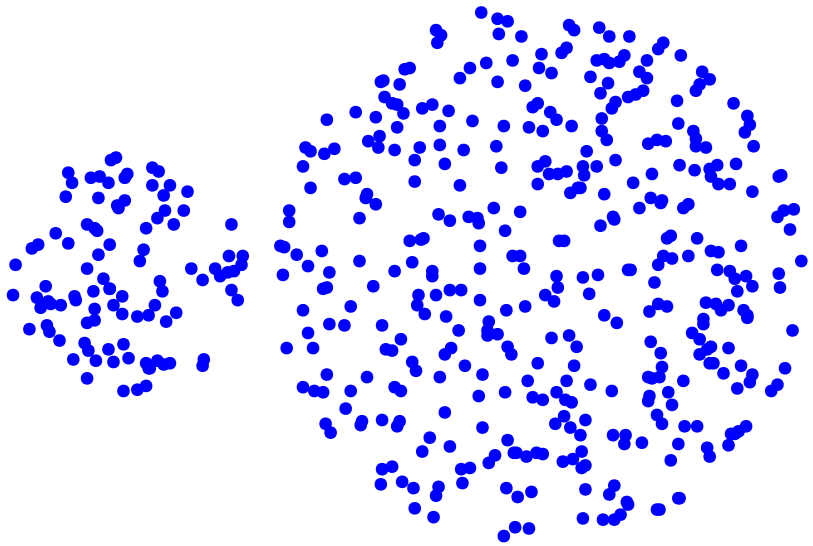
Strength of MAX



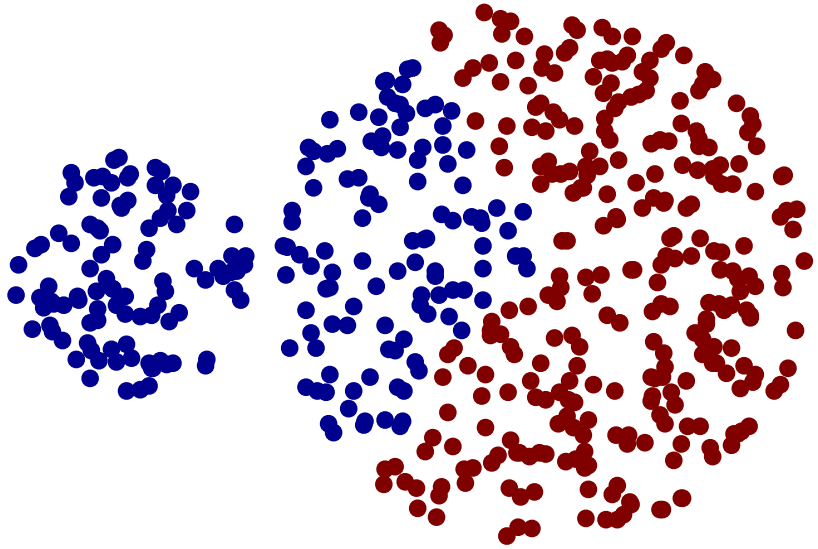
Original Points **Two Clusters**

- **Kurang rentan terhadap kebisingan dan outlier**

Limitations of MAX



Original Points



Two Clusters

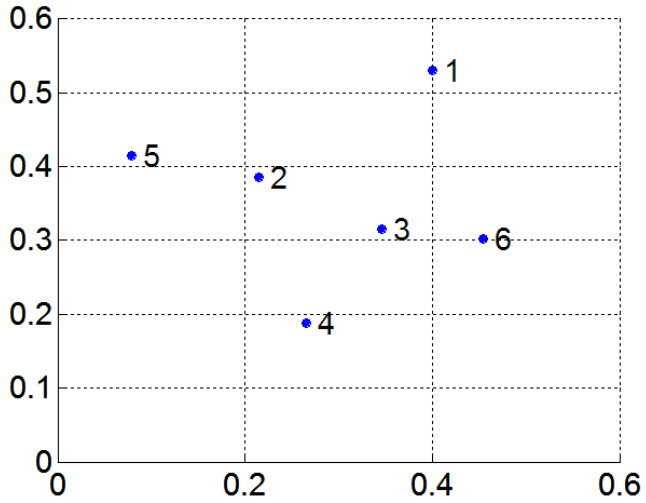
- **Cenderung memecah cluster besar**
- **Bias terhadap gugus bola**

Group Average

- Kedekatan dua cluster adalah rata-rata kedekatan berpasangan antara titik-titik dalam dua kluster.

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

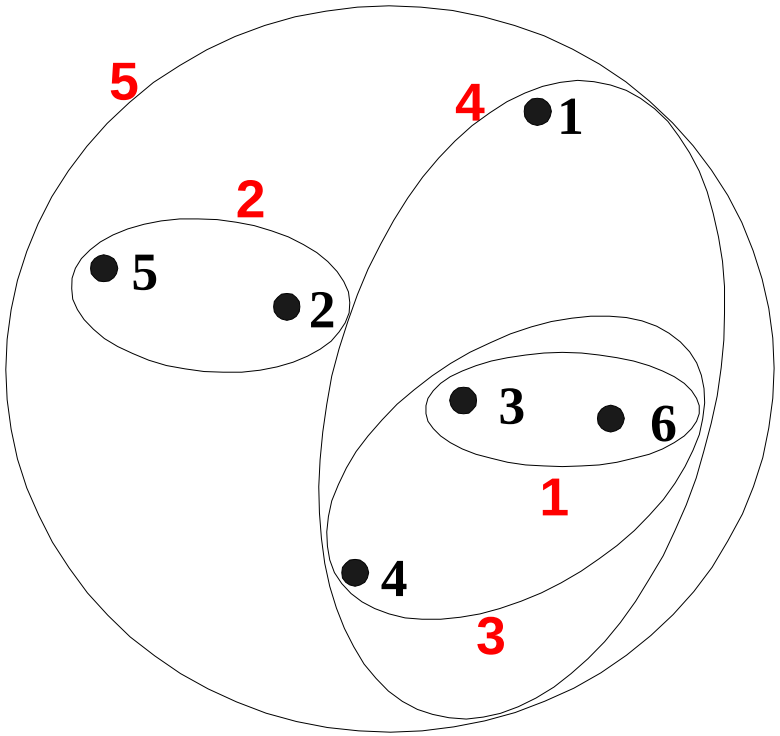
- Perlu menggunakan konektivitas rata-rata untuk skalabilitas karena kedekatan total mendukung kluster besar



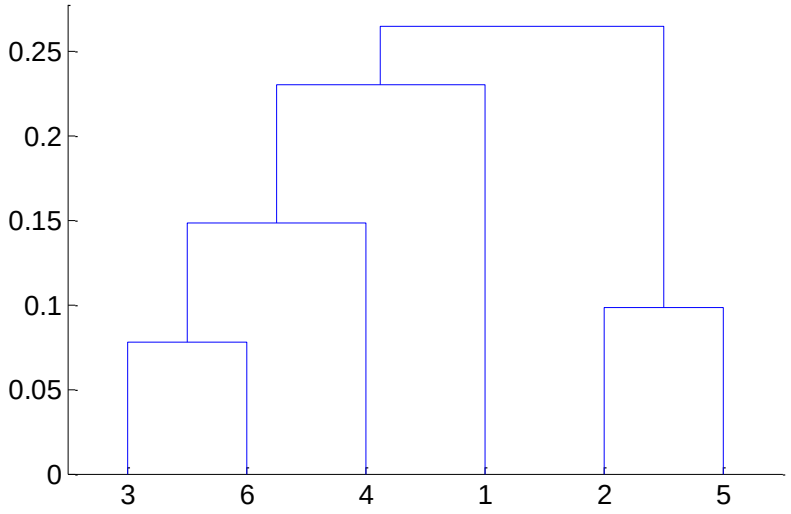
Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Hierarchical Clustering: Group Average



Nested Clusters



Dendrogram

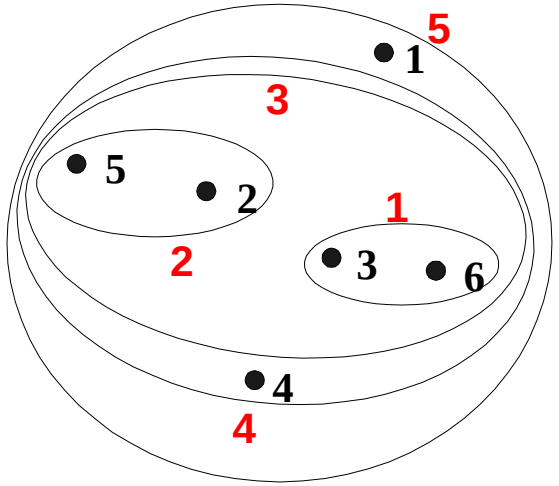
Hierarchical Clustering: Group Average

- Kompromi antara Single and Complete Link
- Strengths
  - Kurang rentan terhadap kebisingan dan outlier
- Limitations
  - Bias terhadap gugus bola

Cluster Similarity: Ward's Method

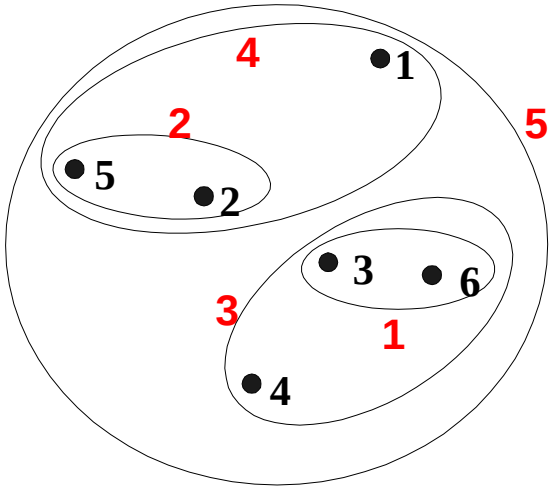
- Kesamaan dua kluster didasarkan pada peningkatan kesalahan kuadrat ketika dua kluster digabungkan
  - Mirip dengan rata-rata grup jika jarak antar titik adalah jarak kuadrat
- Kurang rentan terhadap kebisingan dan outlier
- Bias terhadap gugus bola
- Analog hierarkis dari K-means
  - Dapat digunakan untuk menginisialisasi K-means

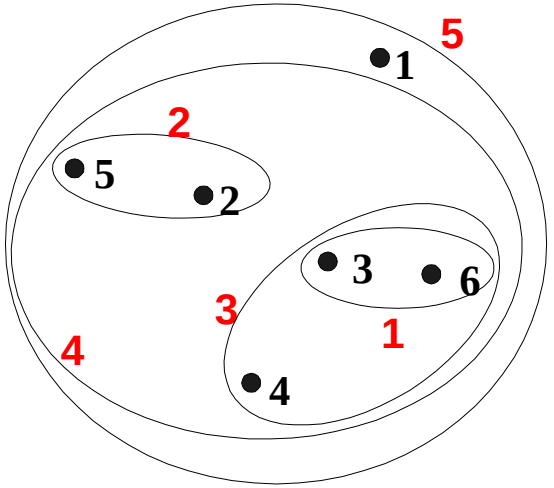
Hierarchical Clustering: Comparison



MIN

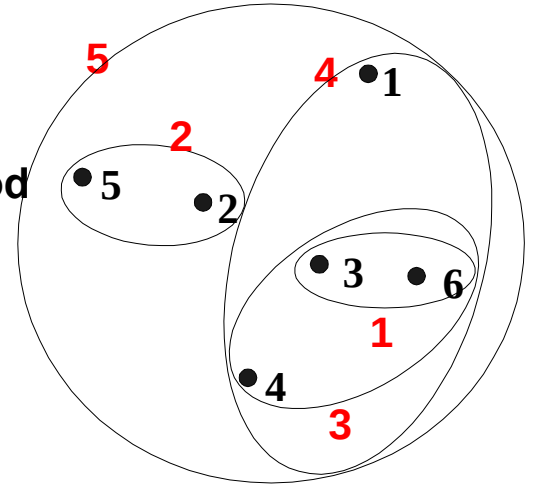
MAX





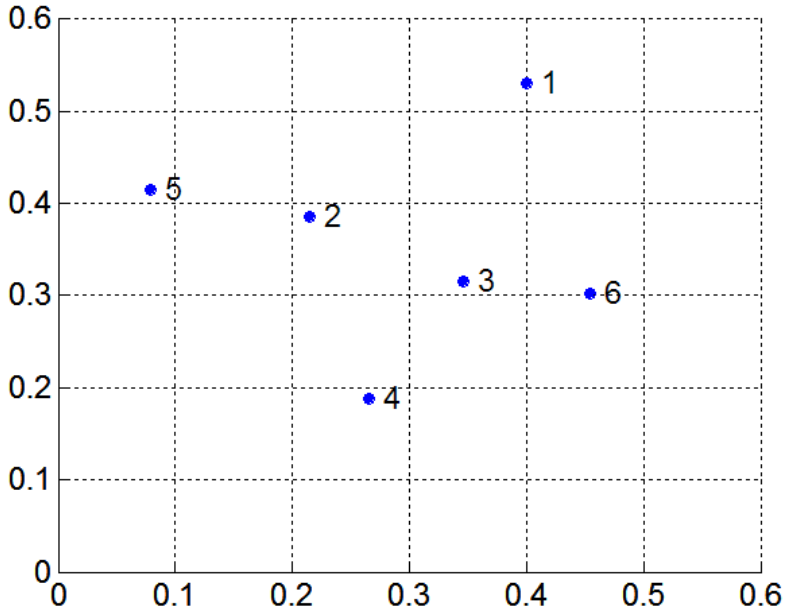
Group Average

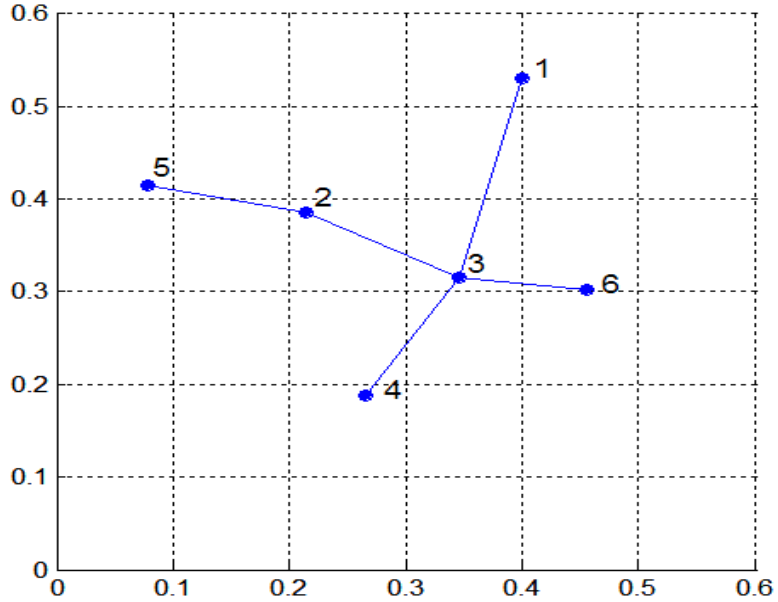
Ward's Method



MST: Divisive Hierarchical Clustering

- Membangun MST (Minimum Spanning Tree)
  - Mulailah dengan pohon yang terdiri dari titik apa pun
  - Dalam langkah-langkah berturut-turut, cari pasangan titik terdekat (p, q) sedemikian rupa sehingga satu titik (p) berada di pohon saat ini tetapi yang lain (q) tidak
  - Tambahkan q ke pohon dan letakkan tepi antara p dan q





MST: Divisive Hierarchical Clustering

- Gunakan MST untuk membangun hierarki kluster

Algorithm 7.5 MST Divisive Hierarchical Clustering Algorithm

1: Compute a minimum spanning tree for the proximity graph.
2: **repeat**
3: Create a new cluster by breaking the link corresponding to the largest distance (smallest similarity).
4: **until** Only singleton clusters remain

Hierarchical Clustering: Time and Space requirements

- O(N2) ruang karena menggunakan proximity matrix.
  - N is the number of points.

- O(N3) waktu dalam banyak kasus
  - Ada N langkah dan pada setiap langkah ukuran, N2, matriks kedekatan harus diperbarui dan dicari
  - Kompleksitas dapat dikurangi menjadi waktu O(N2 log(N)) dengan beberapa kepintaran

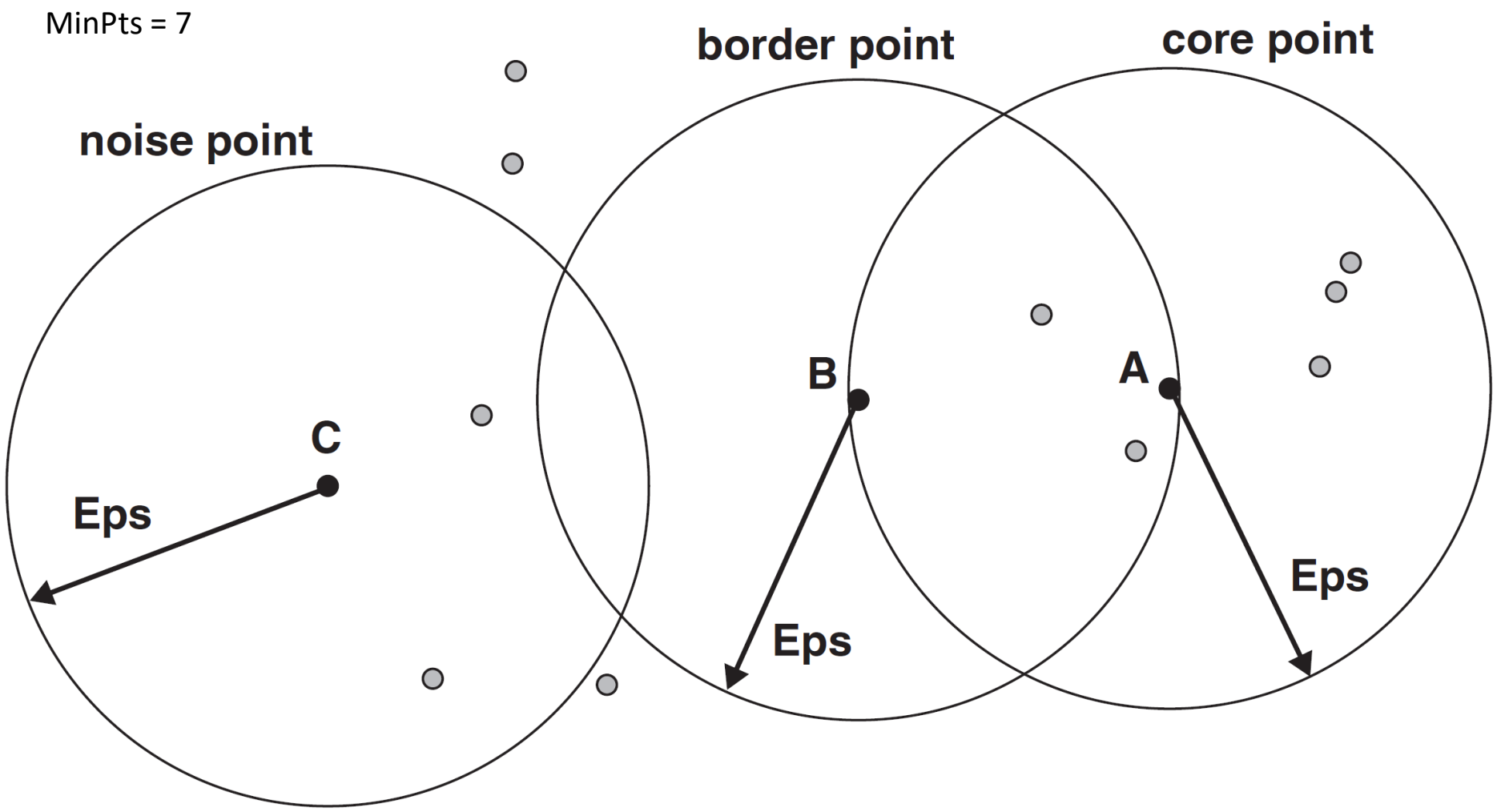
Hierarchical Clustering: Problems and Limitations

- Setelah keputusan dibuat untuk menggabungkan dua klaster, proses tidak dapat dibatalkan
- Tidak ada fungsi objektif global yang diminimalkan secara langsung
- Skema yang berbeda memiliki masalah dengan satu atau beberapa hal berikut:
  - Sensitivitas terhadap kebisingan dan outlier
  - Kesulitan menangani kelompok dengan berbagai ukuran dan bentuk non-bola
  - Memecah cluster besar

DBSCAN

- DBSCAN is a density-based algorithm.
  - Density = jumlah titik dalam radius tertentu (Eps)
  - Poin adalah poin inti jika memiliki setidaknya jumlah poin tertentu (MinPts) dalam Eps
    - Ini adalah titik-titik yang berada di bagian dalam cluster
    - Menghitung poin itu sendiri
  - Titik border bukanlah titik inti, tetapi berada di lingkungan titik inti
  - Titik noise adalah titik apa pun yang bukan titik inti atau titik border

DBSCAN: Core, Border, and Noise Points



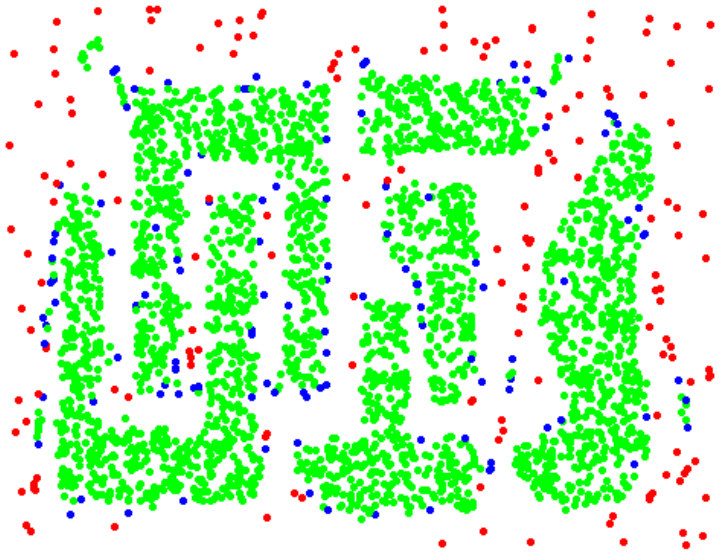
DBSCAN Algorithm

- Eliminate noise points
- Perform clustering on the remaining points

```
current_cluster_label ← 1
for all core points do
  if the core point has no cluster label then
    current_cluster_label ← current_cluster_label + 1
    Label the current core point with cluster label current_cluster_label
  end if
  for all points in the Eps-neighborhood, except i^th the point itself do
    if the point does not have a cluster label then
      Label the point with cluster label current_cluster_label
    end if
  end for
end for
```

DBSCAN: Core, Border and Noise Points

Original Points

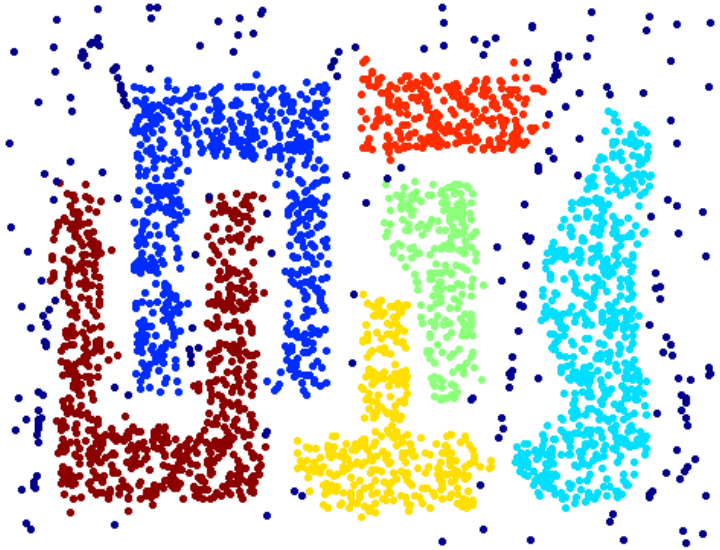


Point types: core, border and noise

Eps = 10, MinPts = 4

When DBSCAN Works Well

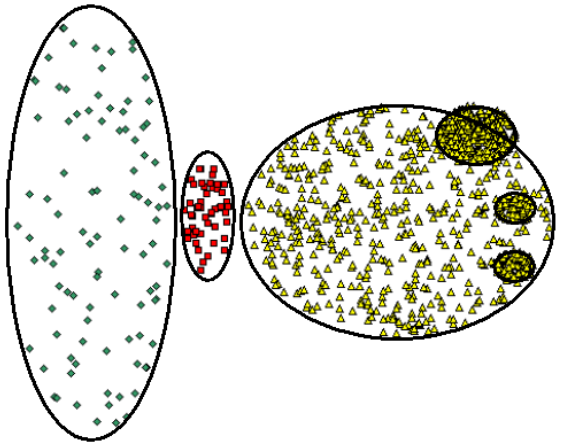
Original Points



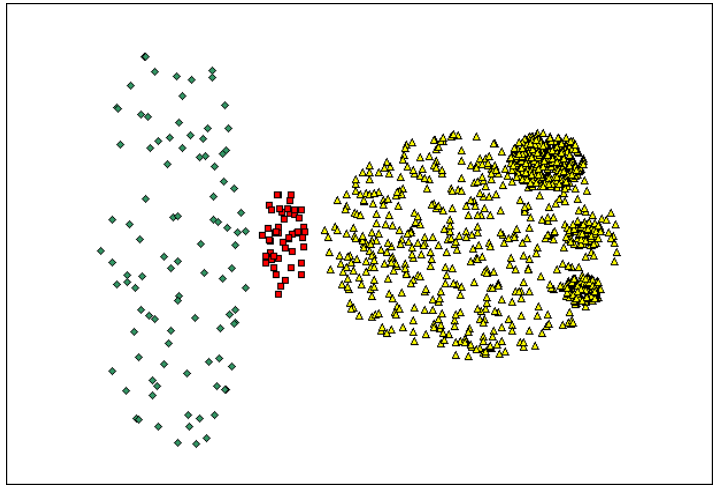
Clusters

- **Resistant to Noise**
- **Can handle clusters of different shapes and sizes**

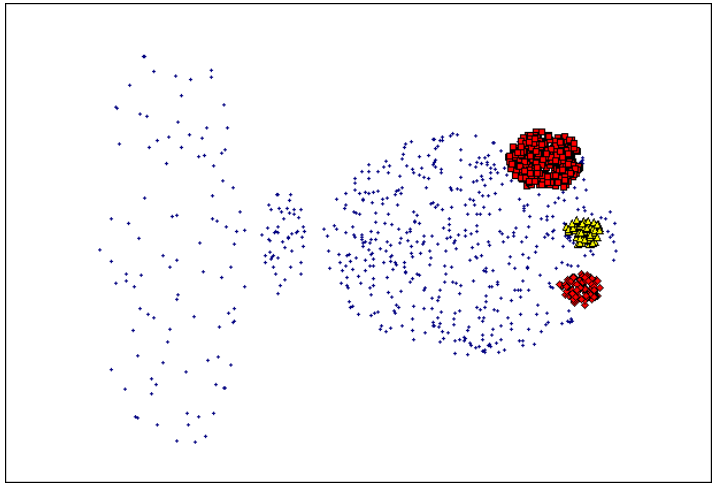
When DBSCAN Does NOT Work Well



Original Points



(MinPts=4, Eps=9.75).

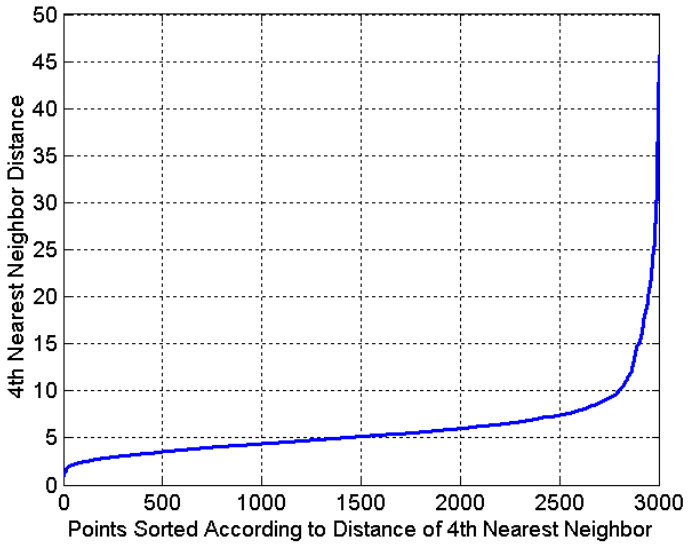


(MinPts=4, Eps=9.92)

- **Varying densities**
- **High-dimensional data**

DBSCAN: Determining EPS and MinPts

- Idenya adalah bahwa untuk titik-titik dalam sebuah gugus, tetangga terdekat ke-k mereka berada pada jarak yang kira-kira sama
- Titik kebisingan memiliki tetangga terdekat ke-k pada jarak yang lebih jauh
- Jadi, plot jarak yang diurutkan dari setiap titik ke tetangga terdekat ke-knya

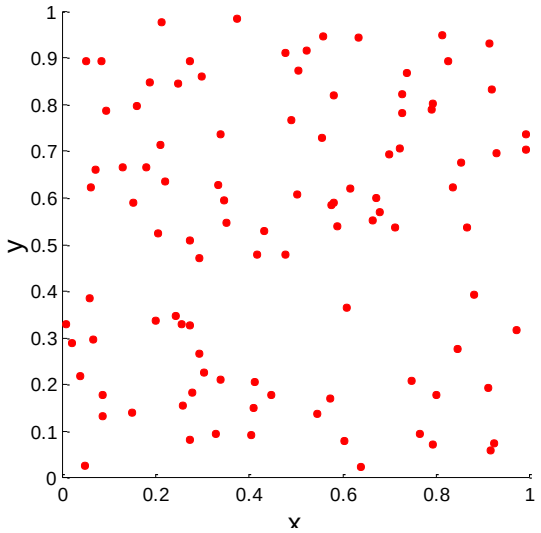


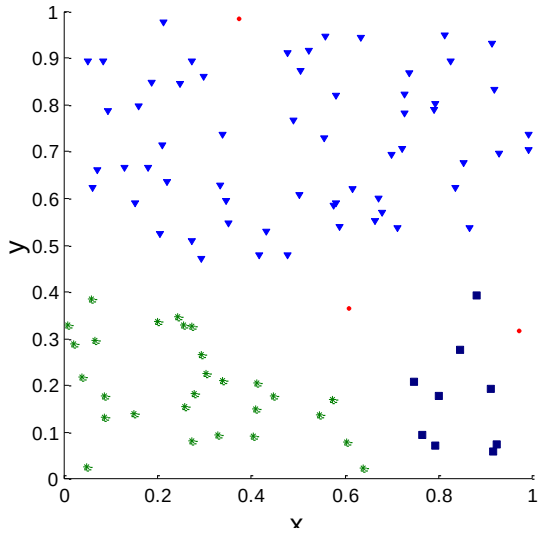
Cluster Validity

- Untuk klasifikasi yang diawasi, bisa menggunakan berbagai pengukuran untuk mengevaluasi seberapa baik model yang dihasilkan
  - Accuracy, precision, recall
- Untuk analisis cluster, pertanyaan analognya adalah bagaimana mengevaluasi "kebaikan" dari cluster yang dihasilkan?
- But “Cluster ada di mata yang melihatnya”!
- Lalu mengapa kita ingin mengevaluasi mereka?
  - Untuk menghindari menemukan pola dalam kebisingan
  - Untuk membandingkan algoritma pengelompokan
  - Untuk membandingkan dua set kluster
  - Untuk membandingkan dua kluster

Clusters found in Random Data

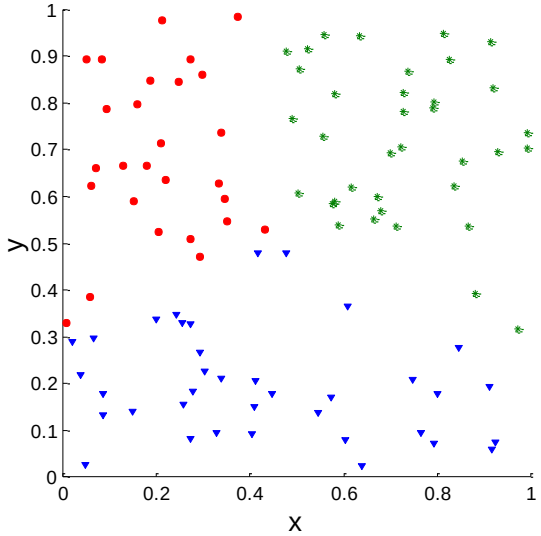
Random Points

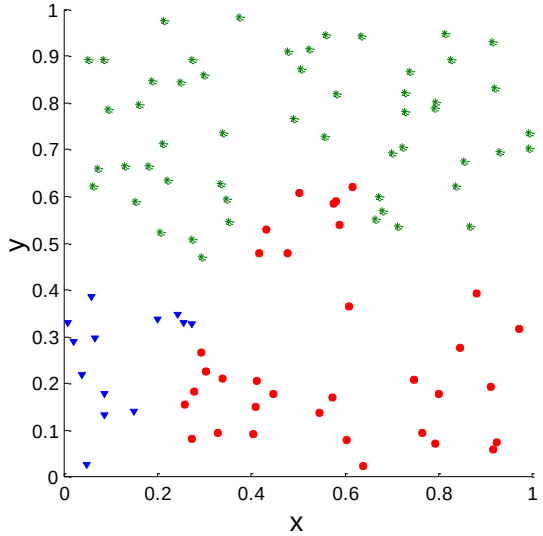




DBSCAN

K-means





Complete Link

Different Aspects of Cluster Validation

1. Menentukan kecenderungan pengelompokan sekumpulan data, yaitu, membedakan apakah struktur non-acak benar-benar ada dalam data.
2. Membandingkan hasil analisis cluster dengan hasil yang diketahui secara eksternal, misalnya, dengan label kelas yang diberikan secara eksternal.
3. Mengevaluasi seberapa baik hasil analisis klaster sesuai dengan data tanpa mengacu pada informasi eksternal.
   - Gunakan hanya data
4. Membandingkan hasil dari dua set analisis klaster yang berbeda untuk menentukan mana yang lebih baik.
5. Menentukan jumlah kluster yang 'benar'.

   Untuk 2, 3, dan 4, kita dapat membedakan lebih lanjut apakah kita ingin mengevaluasi seluruh pengelompokan atau hanya kelompok individual.

Measures of Cluster Validity

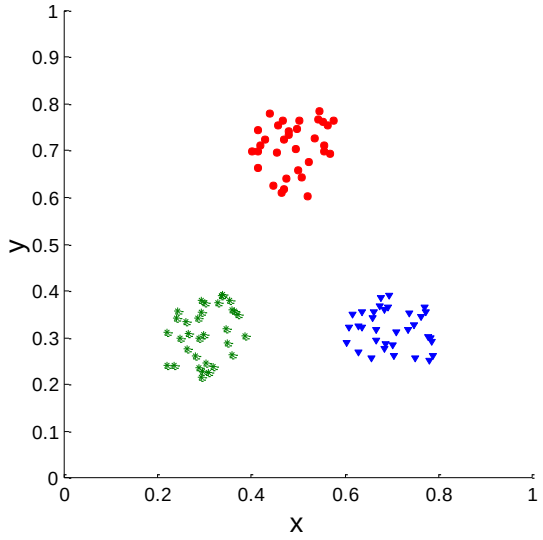
- Ukuran numerik yang diterapkan untuk menilai berbagai aspek validitas klaster, diklasifikasikan ke dalam tiga jenis berikut.
  - Indeks Eksternal: Digunakan untuk mengukur sejauh mana label kluster cocok dengan label kelas yang disediakan secara eksternal.
    - Entropy
  - Indeks Internal: Digunakan untuk mengukur kebaikan struktur pengelompokan tanpa memperhatikan informasi eksternal.
    - Sum of Squared Error (SSE)
  - Indeks Relatif: Digunakan untuk membandingkan dua pengelompokan atau kluster yang berbeda.
    - Seringkali indeks eksternal atau internal digunakan untuk fungsi ini, misalnya, SSE atau entropi
- Kadang-kadang ini disebut sebagai kriteria, bukan indeks
  - Namun, terkadang kriteria adalah strategi umum dan indeks adalah ukuran numerik yang mengimplementasikan kriteria.

Measuring Cluster Validity Via Correlation

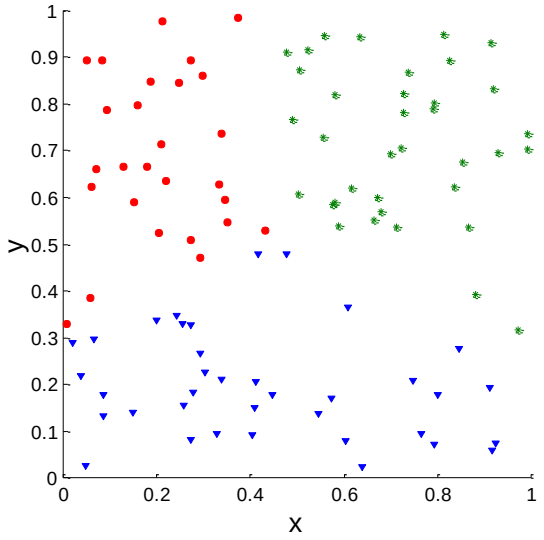
- Dua matriks
  - Proximity Matrix
  - Ideal Similarity Matrix
    - Satu baris dan satu kolom untuk setiap titik data
    - Entri adalah 1 jika pasangan titik terkait milik kluster yang sama
    - Entri adalah 0 jika pasangan titik terkait milik kluster yang berbeda
- Hitung korelasi antara dua matriks
  - Karena matriks simetris, hanya korelasi antara n(n-1) / 2 entri yang perlu dihitung.
- Korelasi tinggi menunjukkan bahwa titik-titik yang termasuk dalam kluster yang sama berdekatan satu sama lain.
- Bukan ukuran yang baik untuk beberapa kelompok berbasis kepadatan atau kedekatan.

Measuring Cluster Validity Via Correlation

- Korelasi kesamaan ideal dan matriks kedekatan untuk pengelompokan K-means dari dua kumpulan data berikut.



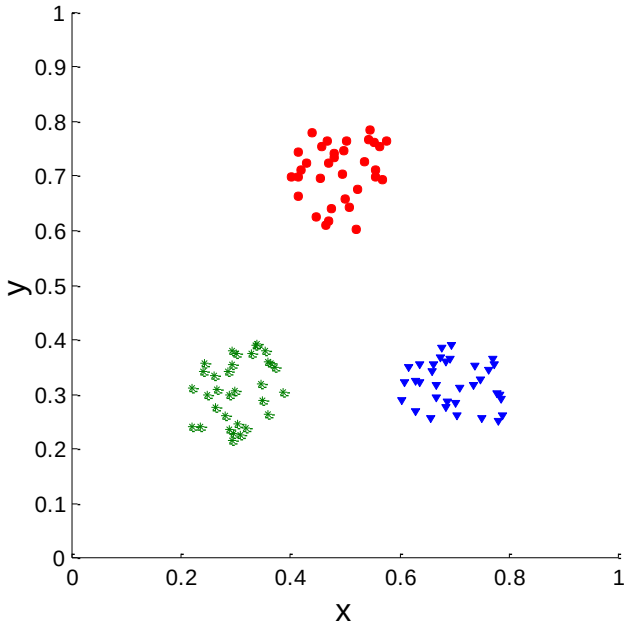
Corr = -0.9235

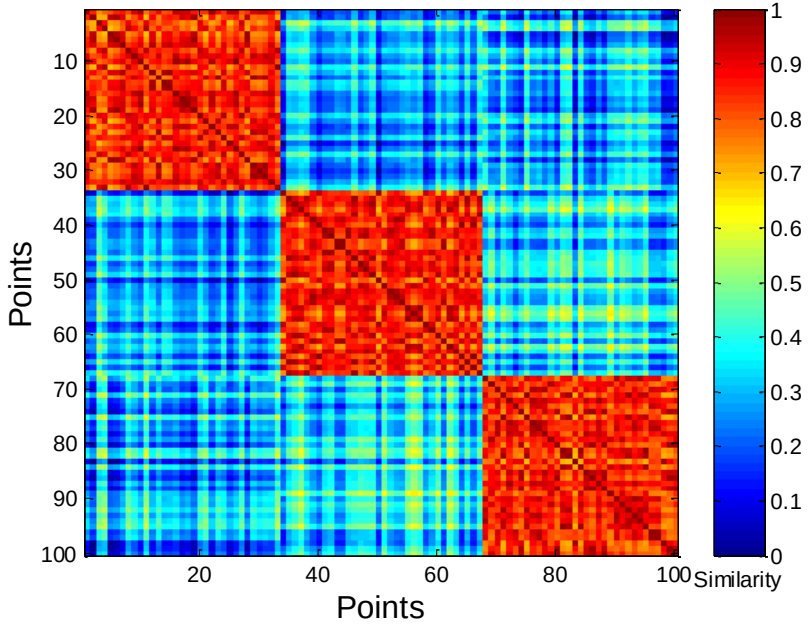


Corr = -0.5810

Using Similarity Matrix for Cluster Validation

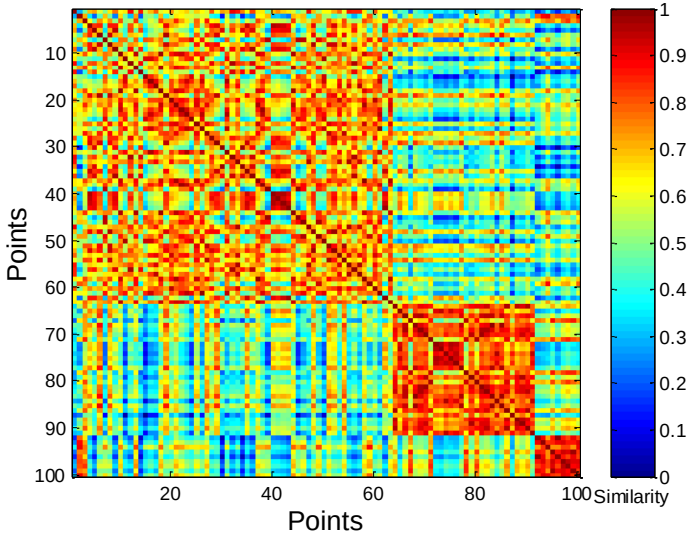
- Urutkan matriks kesamaan sehubungan dengan label kluster dan periksa secara visual.

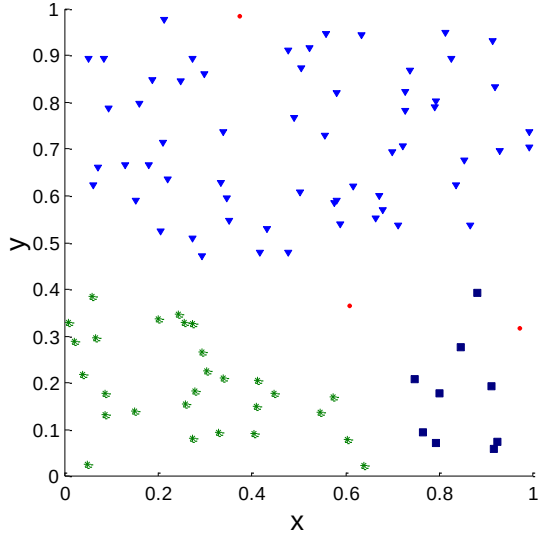




Using Similarity Matrix for Cluster Validation

- Cluster dalam data acak tidak begitu tajam

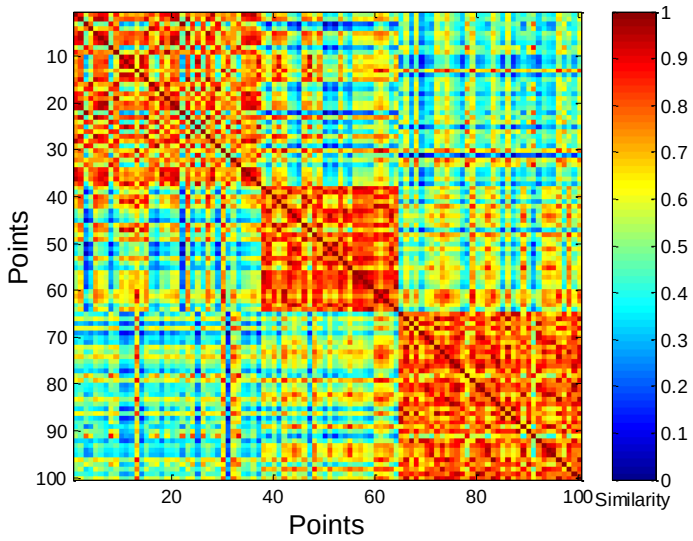


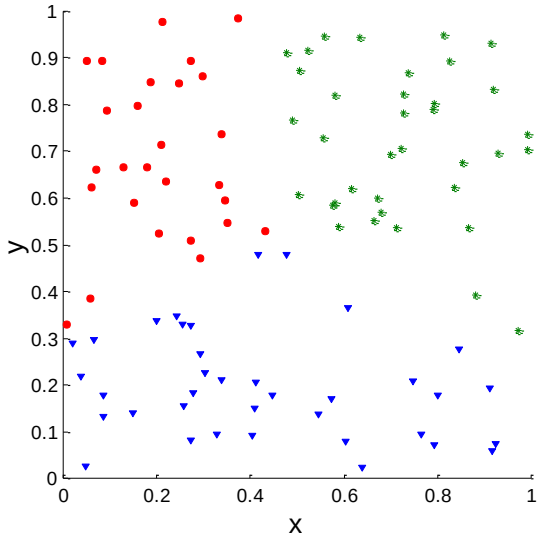


DBSCAN

Using Similarity Matrix for Cluster Validation

- Cluster dalam data acak tidak begitu tajam

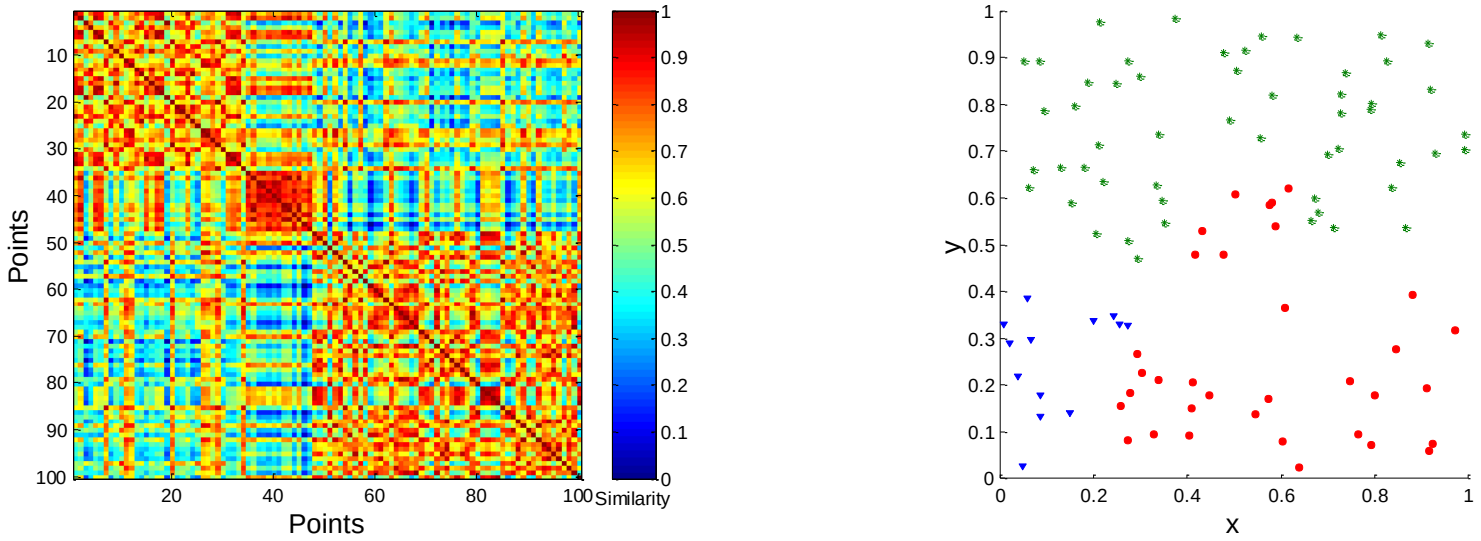




K-means

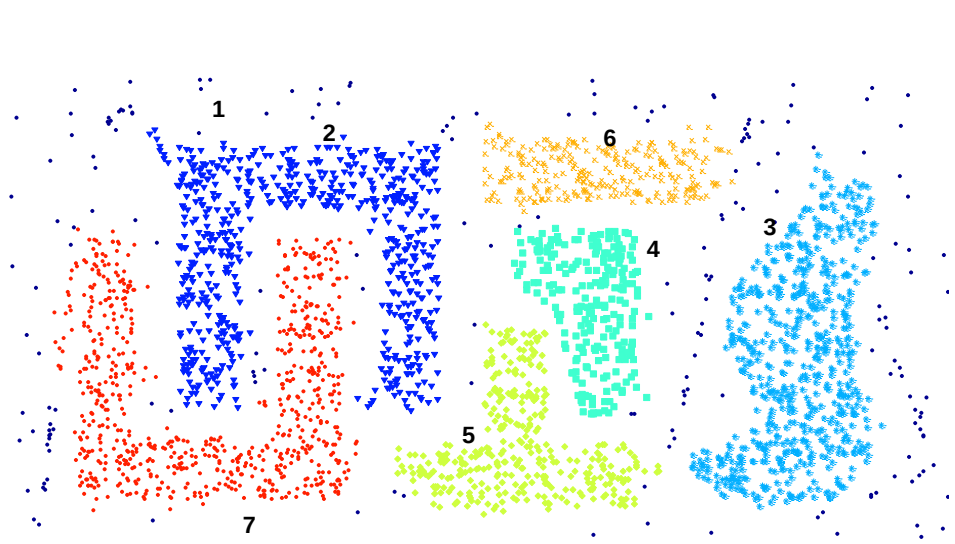
Using Similarity Matrix for Cluster Validation

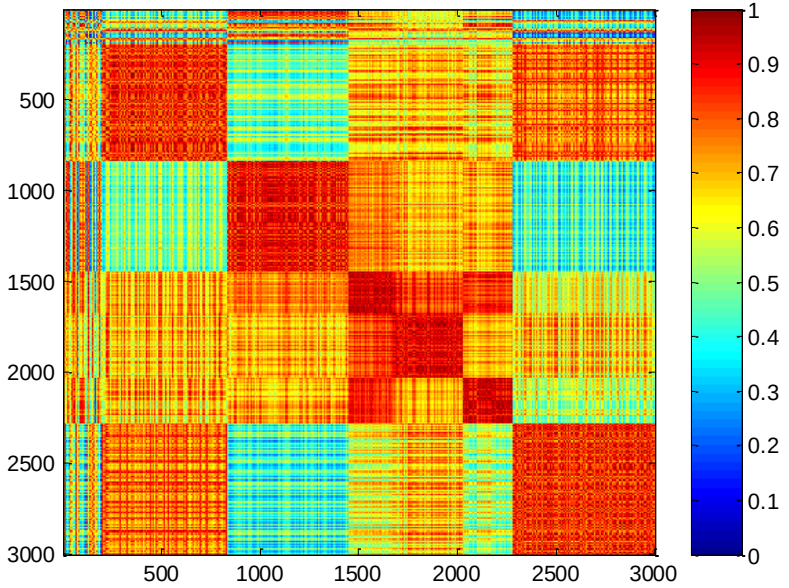
- Cluster dalam data acak tidak begitu tajam



Complete Link

Using Similarity Matrix for Cluster Validation

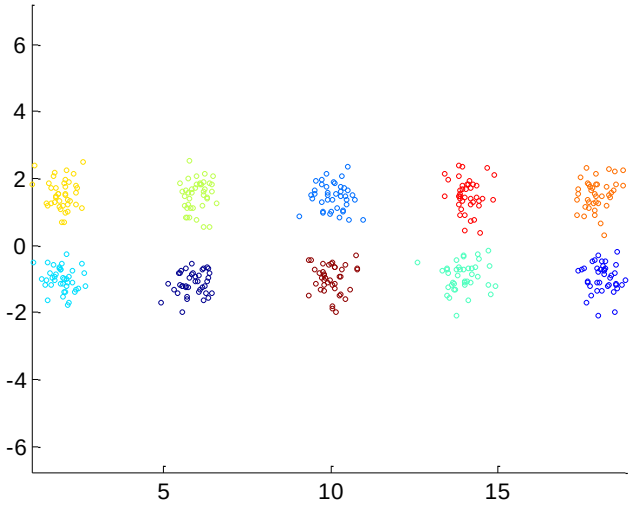


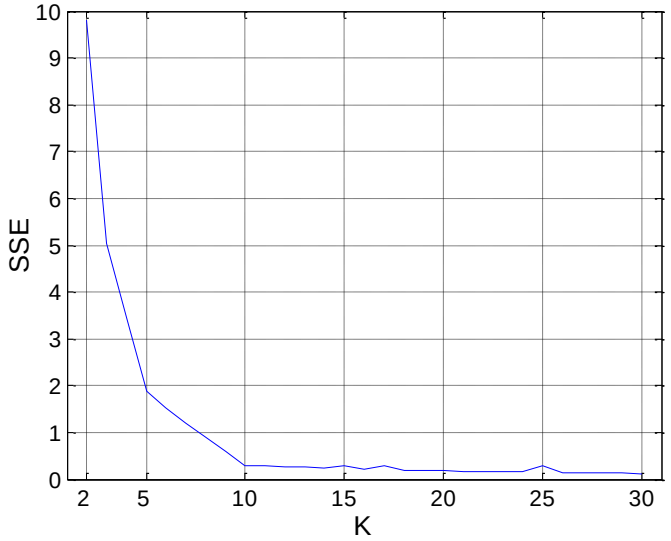


DBSCAN

Internal Measures: SSE

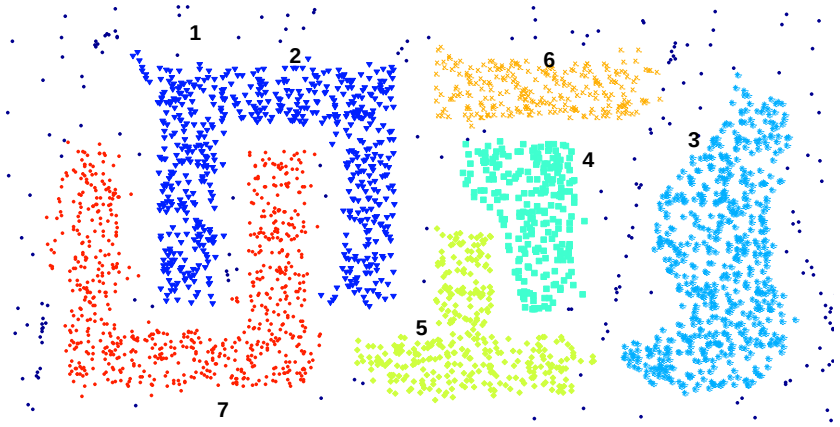
- Kelompok dalam angka yang lebih rumit tidak dipisahkan dengan baik
- Indeks Internal: Digunakan untuk mengukur kebaikan struktur pengelompokan tanpa memperhatikan informasi eksternal
  - SSE
- SSE bagus untuk membandingkan dua klaster atau dua klaster (SSE rata-rata).
- Dapat juga digunakan untuk memperkirakan jumlah cluster

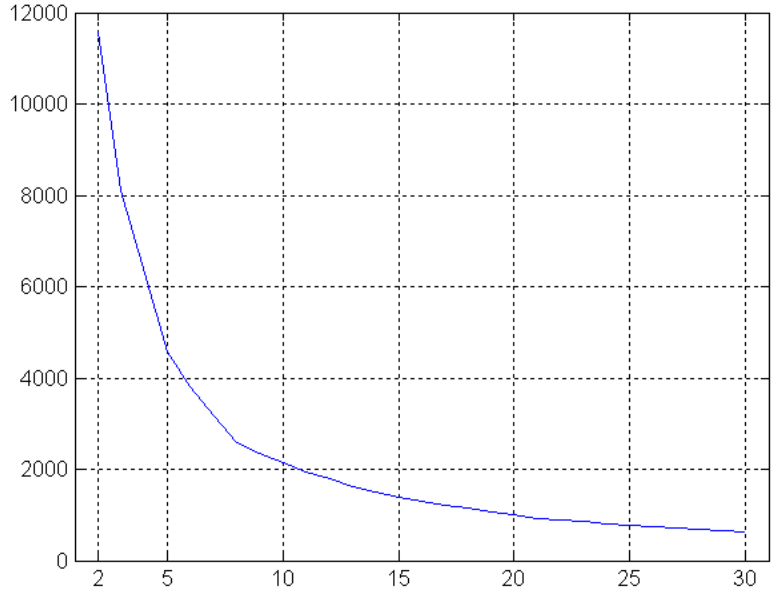




Internal Measures: SSE

- Kurva SSE untuk kumpulan data yang lebih rumit





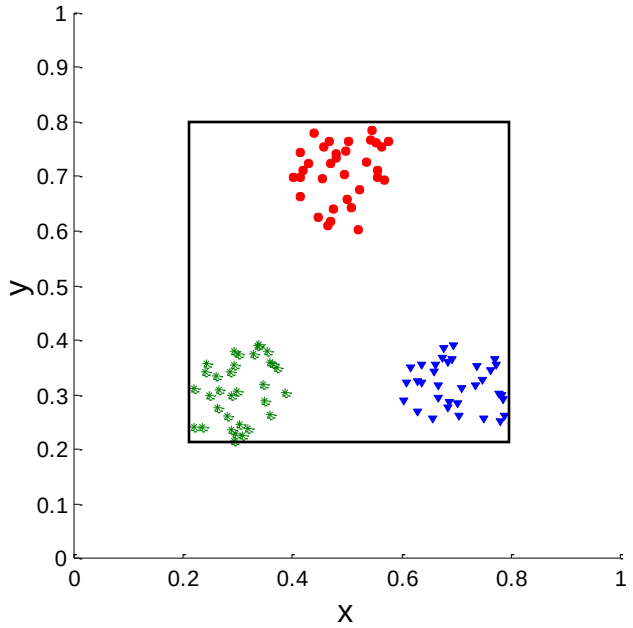
SSE of clusters found using K-means

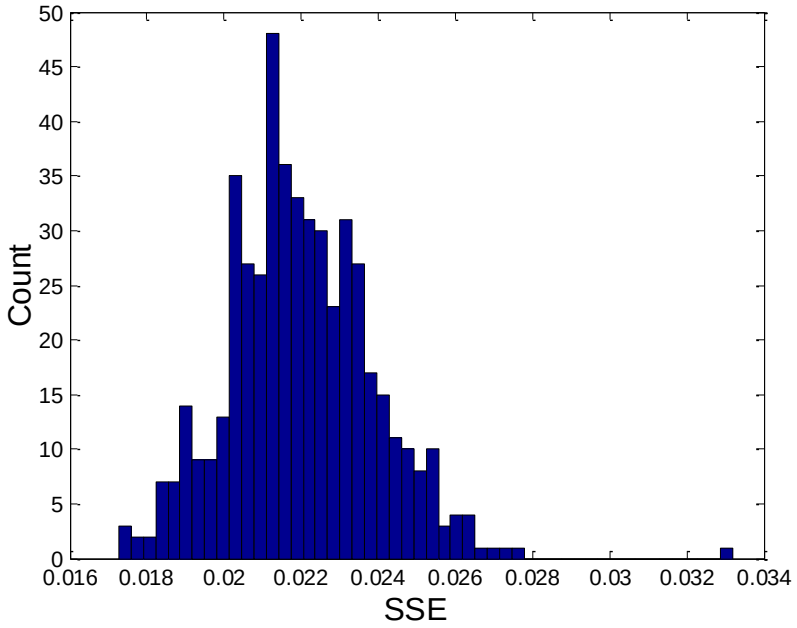
Framework for Cluster Validity

- Butuh kerangka kerja untuk menafsirkan ukuran apa pun.
  - Misalnya, jika ukuran evaluasi kita memiliki nilai, 10, apakah itu baik, adil, atau buruk?
- Statistik menyediakan kerangka kerja untuk validitas klaster
  - Semakin "atypical" hasil pengelompokan, semakin besar kemungkinan itu mewakili struktur yang valid dalam data
  - Dapat membandingkan nilai indeks yang dihasilkan dari data atau pengelompokan acak dengan nilai hasil pengelompokan.
    - Jika nilai indeks tidak sama, maka hasil klaster valid
  - Pendekatan ini lebih rumit dan lebih sulit dipahami.
- Untuk membandingkan hasil dari dua set analisis klaster yang berbeda, kerangka kerja kurang diperlukan.
  - Namun, ada pertanyaan apakah perbedaan antara dua nilai indeks signifikan

Statistical Framework for SSE

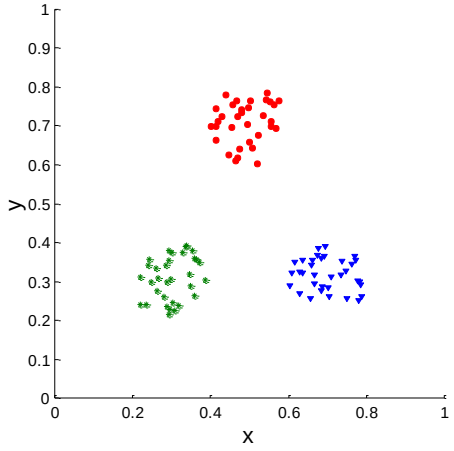
- Example
  - Bandingkan SSE 0,005 dengan tiga kluster dalam data acak
  - Histogram menunjukkan SSE dari tiga kluster dalam 500 set titik data acak berukuran 100 yang didistribusikan pada rentang 0,2 – 0,8 untuk nilai x dan y

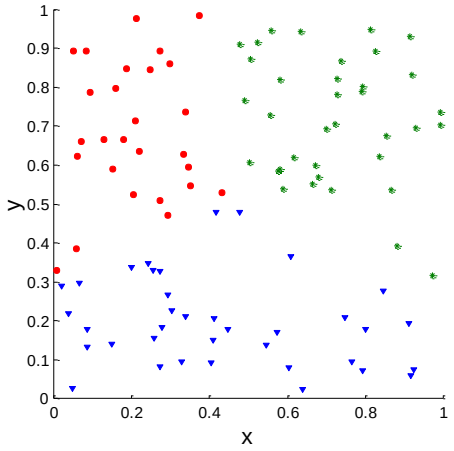


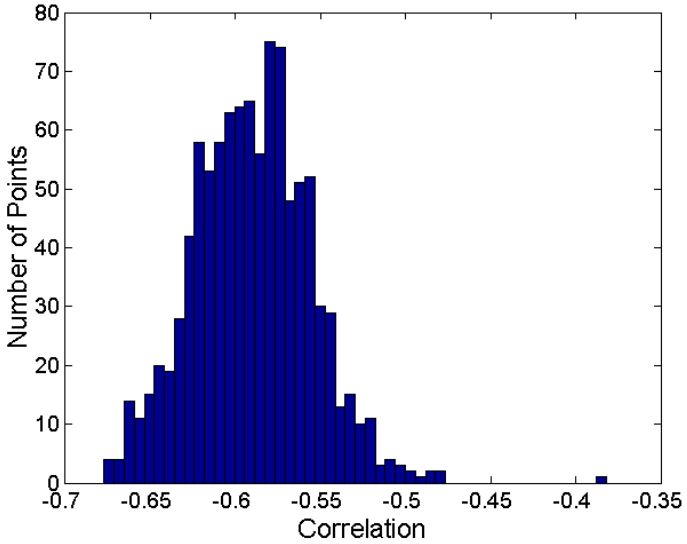


Statistical Framework for Correlation

- Korelasi kesamaan ideal dan matriks kedekatan untuk pengelompokan K-means dari dua kumpulan data berikut.







Corr = -0.9235

Corr = -0.5810

Internal Measures: Cohesion and Separation

- Cluster Cohesion: Mengukur seberapa dekat terkait objek dalam kluster
  - Example: SSE
- Cluster Separation:Mengukur seberapa berbeda atau terpisah dengan baik suatu kluster dari kluster lain
- Example: Squared Error
  - Cohesion diukur didalam cluster menggunakan sum of squares (SSE)

$$SSE = WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

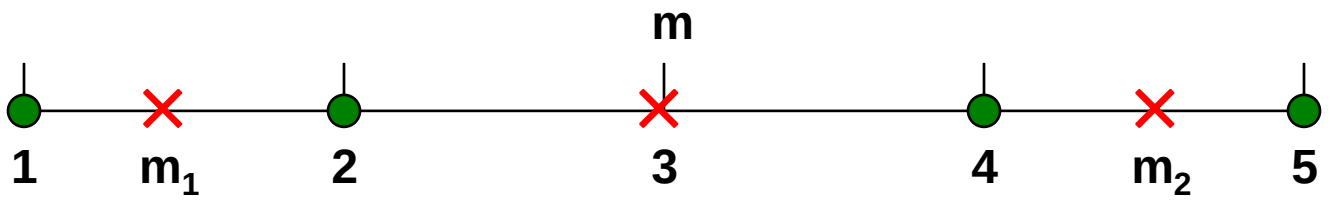
  - Separation diukur diantara cluster dengan menggunakan sum of squares

$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

- Where |Ci| is the size of cluster i

Internal Measures: Cohesion and Separation

- Example: SSE
  - BSS + WSS = constant



K=1 cluster:

$$SSE = WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$BSS = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

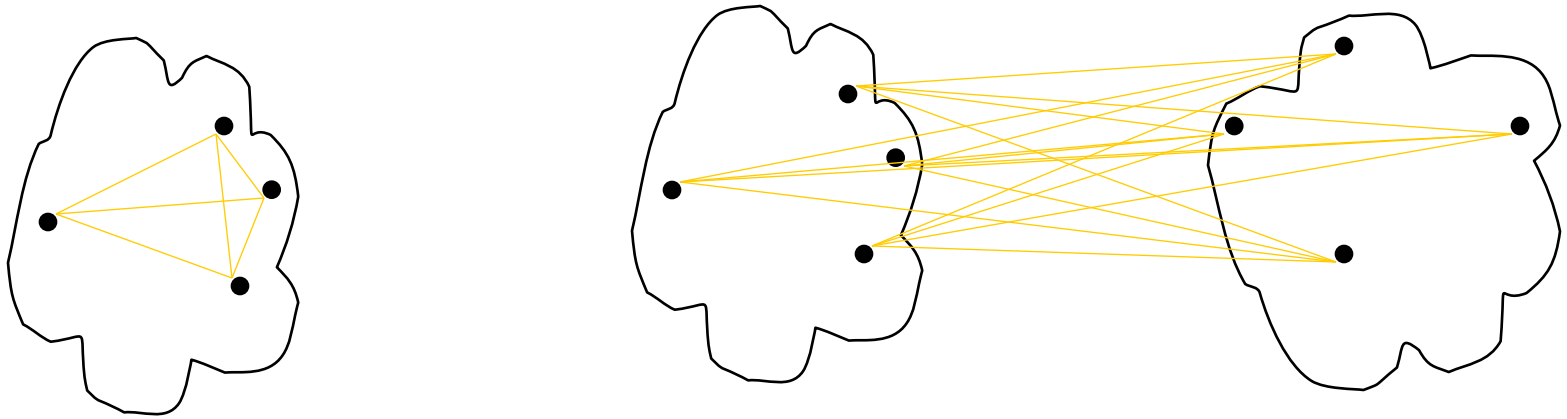
$$SSE = WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

Internal Measures: Cohesion and Separation

- Pendekatan berbasis grafik kedekatan juga dapat digunakan untuk kohesi dan pemisahan.
  - Kohesi kluster adalah jumlah bobot semua tautan dalam klaster.
  - Pemisahan kluster adalah jumlah bobot antara simpul dalam kluster dan simpul di luar kluster.



cohesion

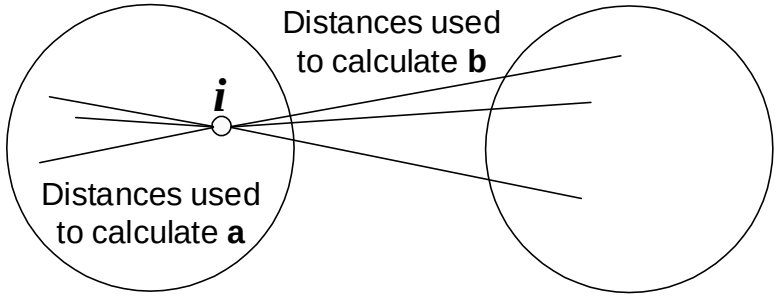
separation

Internal Measures: Silhouette Coefficient

- Silhouette coefficient menggabungkan gagasan kohesi dan pemisahan, tetapi untuk titik-titik individu, serta klaster dan klasterisasi
- For an individual point, i
  - Calculate a = Jarak rata-rata i ke titik-titik dalam kelompoknya
  - Calculate b = min (jarak rata-rata i ke titik di gugus lain)
  - Koefisien siluet untuk suatu titik kemudian diberikan oleh

    s = (b – a) / max(a,b)

  - Biasanya antara 0 dan 1.
  - Semakin dekat dengan 1 semakin baik.



- Dapat menghitung koefisien siluet rata-rata untuk cluster atau clustering

External Measures of Cluster Validity: Entropy and Purity

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

Final Comment on Cluster Validity

"Validasi struktur pengelompokan adalah bagian yang paling sulit dan membuat frustrasi dari analisis klaster.

Tanpa upaya yang kuat ke arah ini, analisis cluster akan tetap menjadi seni hitam yang hanya dapat diakses oleh orang-orang percaya sejati yang memiliki pengalaman dan keberanian besar."

Algorithms for Clustering Data, Jain and Dubes