ANALISIS GEROMBOL (CLUSTER ANALYSIS)

Bahan Kuliah Secara Daring
Mahasiswa Departemen Statistika-FMIPA-IPB
Oleh: Dr. Ir. Budi Susetyo

Latar Belakang

- Dalam beberapa hal, tujuan suatu penelitian adalah untuk mengelompokkan individu-individu berdasarkan banyak peubah penciri
- Individu-individu dalam satu kelompok memiliki kemiripan atau keragaman yang kecil dibandingkan individu-individu di kelompok yang berbeda
- Analisis gerombol merupakan metode yang dapat menggabungkan beberapa individu ke dalam kelompok-kelompok berdasarkan sifat kemiripan atau sifat ketidakmiripan antar objek, sehingga objek dalam kelompok lebih mirip dibandingkan dengan objek antar kelompok
- Kemiripan/ketakmiripan antar objek dalam analisis gerombol menggunakan konsep jarak
- Sebagai contoh ingin menggerombolkan 34 provinsi di Indonesia berdasarkan indicator kesejahteraan rakyat
- Terdapat beberapa metode penggerombolan

Konsep Jarak Antar Objek

- Objek yang berada pada gerombol yang sama memiliki kemiripan yang lebih besar dibandingkan objek yg ada dalam gerombol lainnya.
- Kemiripan antar objek diukur dengan konsep jarak
- Jarak antar 2 objek *a* dan *b,* dinotasikan dengan d*(a,b),* dimana :

- $d(a, b) \geq 0$
- $d(a, a) = 0$
- $d(a, b) = d(b, a)$
- d(*a*, *b*) meningkat seiring semakin tidak mirip kedua objek a dan b
- $d(a,c) \leq d(a,b) + d(b,c)$

Asumsi : semua pengukuran bersifat numerik

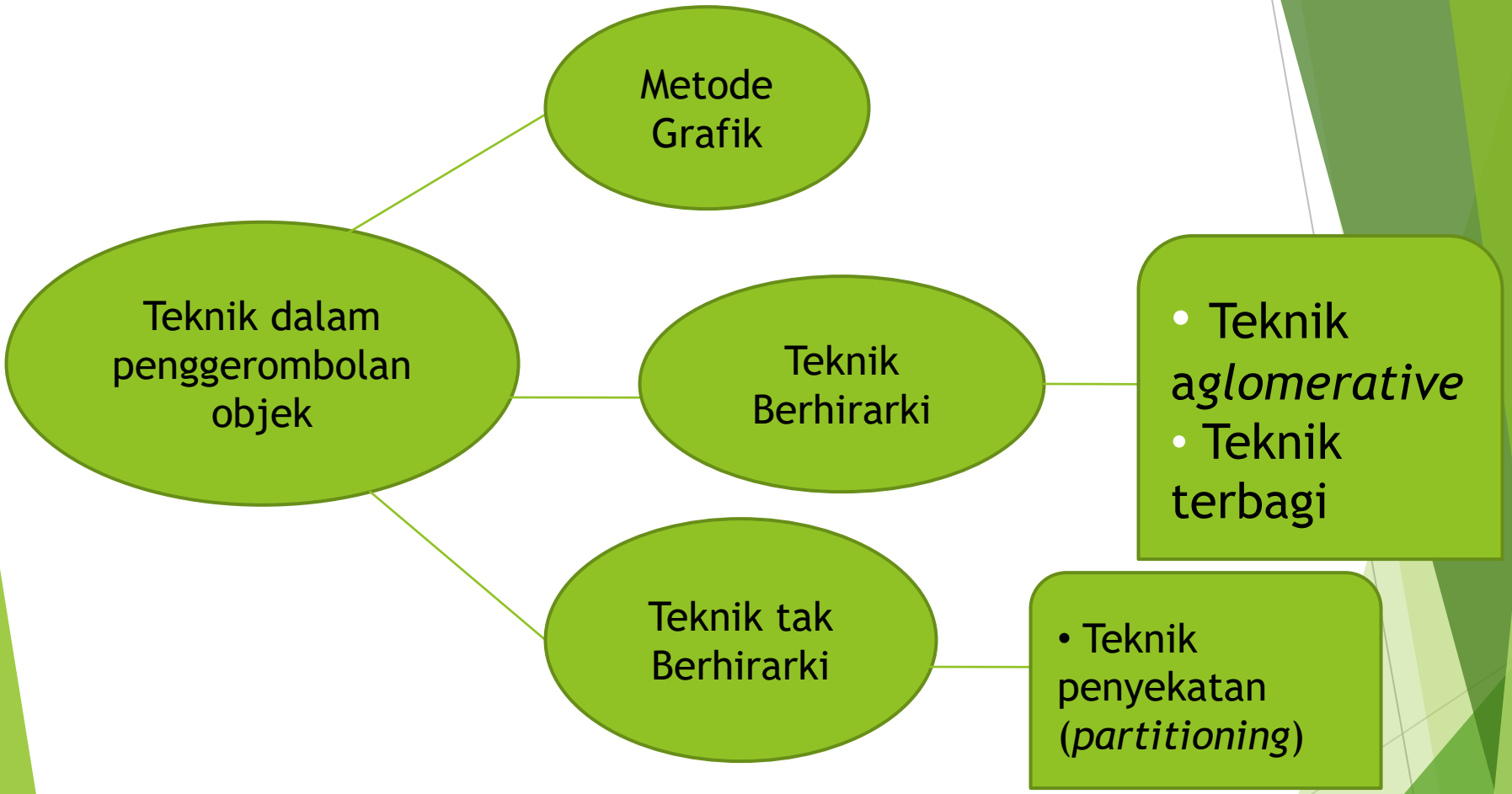
Struktur Data Amatan

Individu	Peubah				
	X1	X2	X3	...	Xp
1	x11	x12	x13		x1p
2	x21	x22	x23		x2p
3	x31	x32	x33		x3p
4	x41	x42	x43		x4p
5	x51	x52	x53		x5p
...	...	...	...	...	...
...	...	...	...	...	...
n	xn1	xn2	xn3		xnp

Beberapa konsep jarak

Jarak	Formula
Jarak Euclidean	$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}$ $= \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$
Jarak Minkowski / Jarak city-block / Jarak Manhattan	$d(x, \mathbf{y}) = \left[\sum_{i=1}^{p}\lvert x_i - y_i\rvert^k\right]^{\frac{1}{k}}$
Jarak Mahalanobis	$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'\mathbf{S}^{-1}(\mathbf{x}-\mathbf{y})}$

Beberapa Teknik Penggerombolan



Metode grafik sangat subjetif untuk menarik kesimpulan

Perbedaan antara teknik berhirarki dengan teknik tak berhirarki

Teknik berhirarki	• banyaknya gerombol yang akan dihasilkan belum diketahui • hasil penggerombolan ditampilkan dalam bentuk dendrogram
Teknik tak berhirarki	• banyaknya gerombol sudah ditentukan dulu • Beberapa metode: K-rataan Macqueen, metode Chernoff dan kurva Andrews

Metode berhirarkhi lebih popular digunakan

Beberapa metode penggerombolan berhirarkhi:

- Pautan Tunggal
- Pautan Lengkap
- Pautan Centroid
- Pautan Median
- Pautan Rataan

- **Pautan Tunggal (Single Linkage)**

  Jarak antar dua gerombol diukur dengan jarak terdekat antara sebuah objek dalam gerombol yang satu dengan sebuah objek dalam gerombol yang lain.

  $$h(Br, Bs) = \min \{ d(\mathbf{x}i, \mathbf{x}j); \mathbf{x}i \text{ anggota } Br, \text{ dan } \mathbf{x}j \text{ anggota } Bs \}$$

- **Pautan Lengkap (Complete Linkage)**

  Jarak antar dua gerombol diukur dengan jarak terjauh antara sebuah objek dalam gerombol yang satu dengan sebuah objek dalam gerombol yang lain.

$h(Br, Bs) = \max \{ d(\mathbf{x}i, \mathbf{x}j); \mathbf{x}i \text{ anggota } Br, \text{ dan } \mathbf{x}j \text{ anggota } Bs \}$

Pautan Centroid (Centroid Linkage)

Jarak antara dua buah gerombol diukur sebagai jarak Euclidean antara kedua rataan (centroid) gerombol.

Jika $\bar{\mathbf{x}}_r$ dan $\bar{\mathbf{x}}_s$ adalah vektor rataan (centroid) dari gerombol *Br* dan *Bs*, maka jarak kedua gerombol tersebut didefinisikan sebagai :

$$h(B_r, B_s) = d(\bar{\mathbf{x}}_r, \bar{\mathbf{x}}_s)$$

Centroid cluster yang baru didefinisikan sebagai :

$$\frac{n_r \bar{\mathbf{x}}_r + n_s \bar{\mathbf{x}}_s}{n_r + n_s}$$

Pautan Median (Median Linkage)

Jarak antar gerombol didefinisikan sebagai jarak antar median, dan gerombol-gerombol dengan jarak terkecil akan digabungkan.

Median untuk gerombol yang baru adalah

$$\text{Mbaru} = \frac{\mathbf{m}_r + \mathbf{m}_s}{2}$$

Pautan Rataan (Average Linkage)

Jarak antara dua buah gerombol, *Br* dan *Bs* didefinisikan sebagai rataan dari $n_r n_s$ jarak yang dihitung antara $\mathbf{x}_i$ anggota *Br* dan $\mathbf{x}_j$ anggota *Bs*

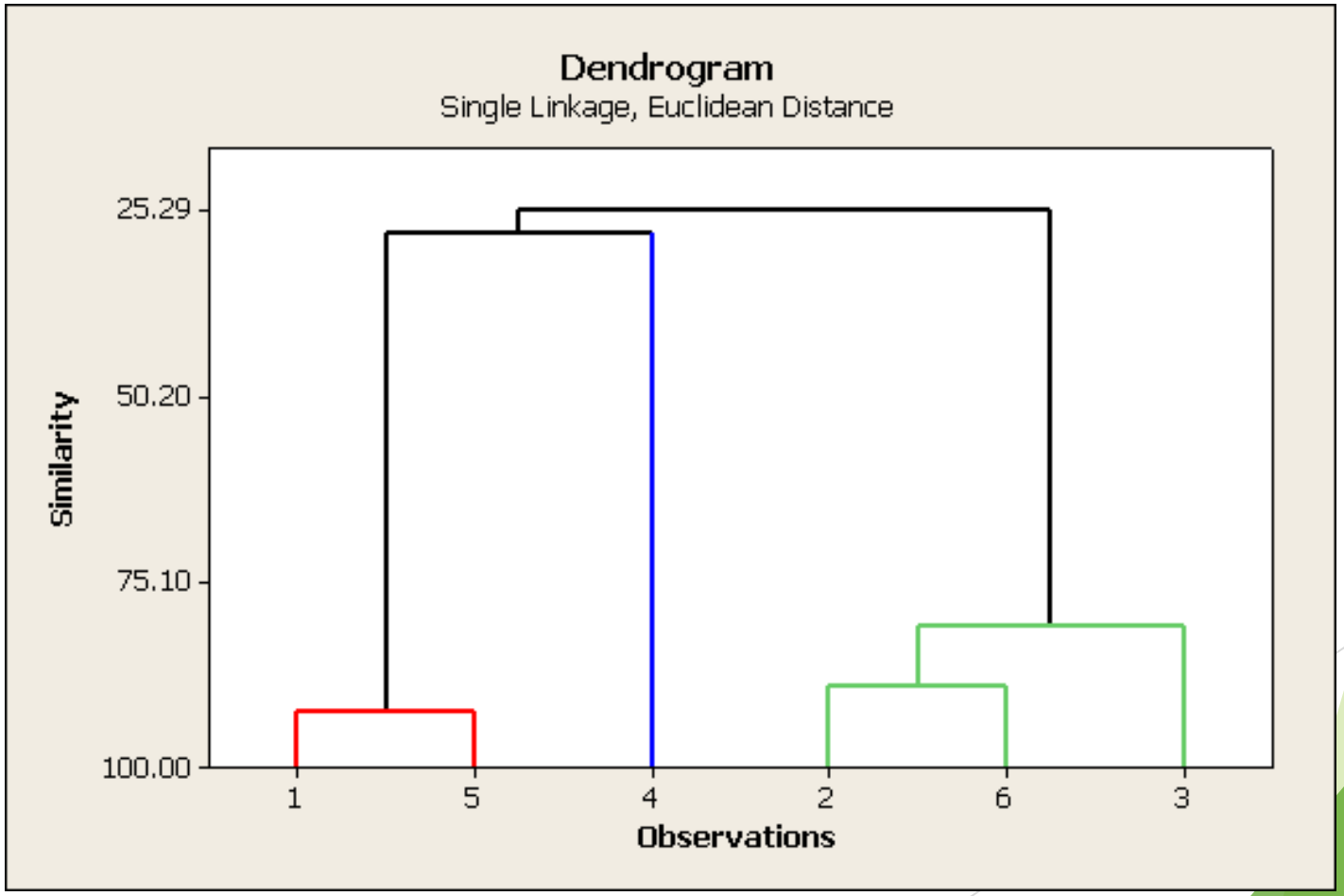
$$h(B_r, B_s) = \frac{1}{n_r n_s} \sum_{\mathbf{x}_i \in B_r} \sum_{\mathbf{x}_j \in B_s} d(\mathbf{x}_i, \mathbf{x}_j)$$

Berikut adalah data nilai 7 mata pelajaran dari 6 siswa. Berdasarkan data tersebut ingin diketahui kemiripan prestasi 6 siswa tersebut

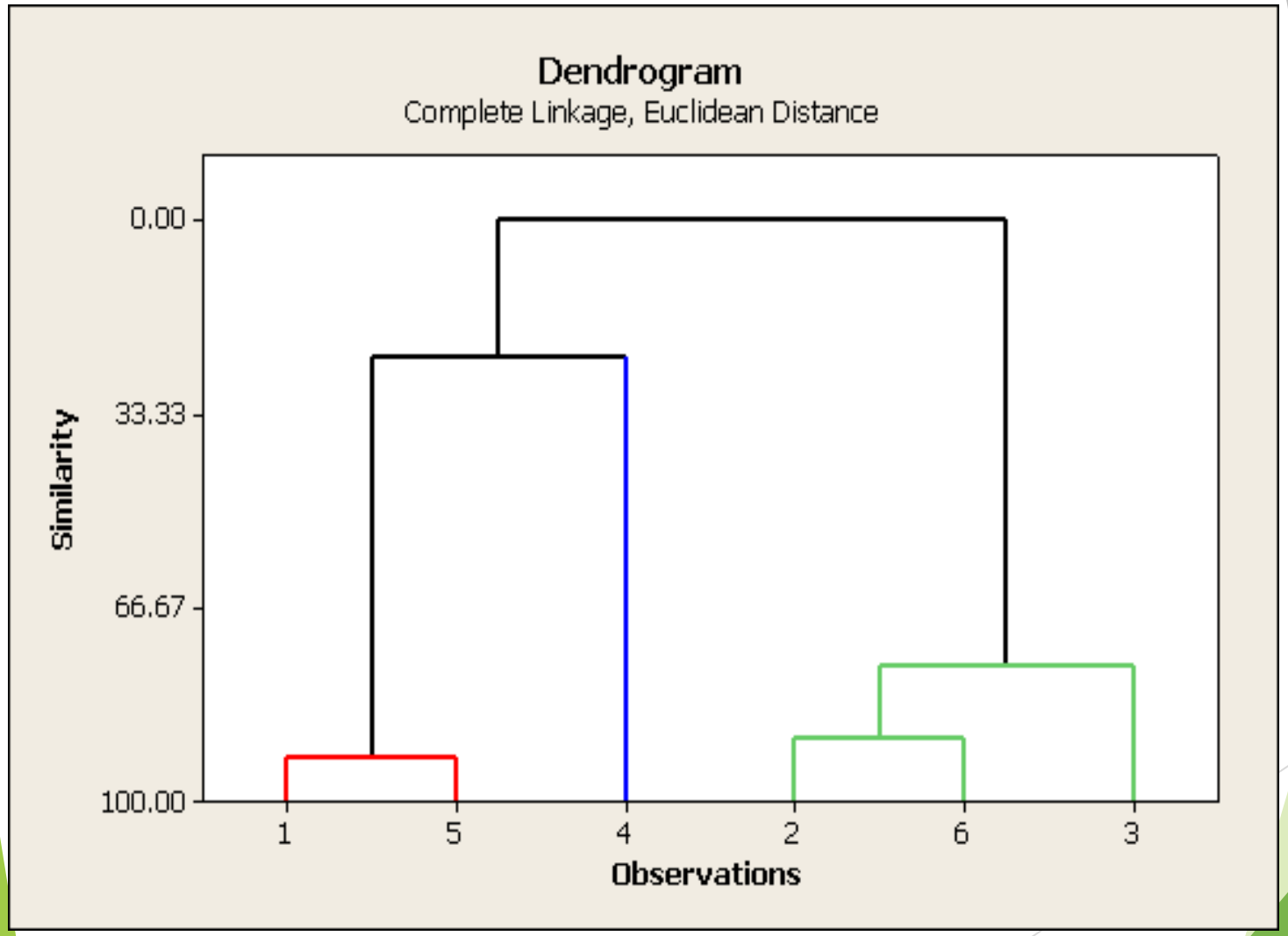
	Mat	Fis	Bio	Sej	Kew	Sos	Seni
1.Andi	8.1	8.3	7.6	6.2	5.8	5.4	6.0
2. Benny	5.6	6.3	6.1	7.3	7.4	7.6	6.0
3.Budi	5.2	5.8	5.7	7.0	6.8	7.2	5.7
4. Ika	6.7	6.8	5.6	7.4	5.3	5.4	7.9
5. Maya	8.2	8.2	7.4	6.4	5.7	5.5	6.1
6. Ana	5.7	6.4	5.9	7.1	7.2	7.3	5.8

Perbandingan hasil dendogram kelima metode penggerombolan (menghasilkan hasil yang berbeda)

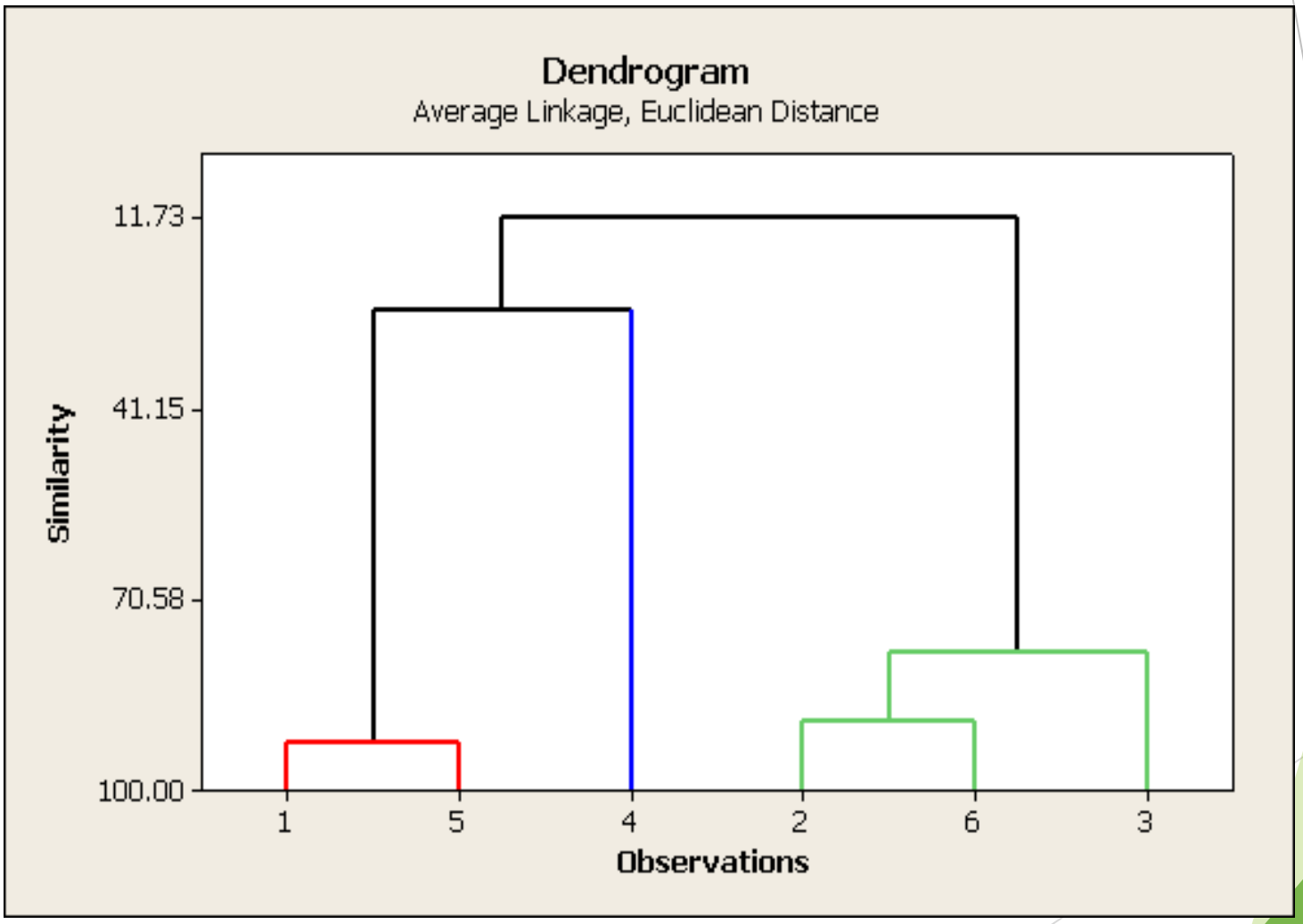
Single Linkage



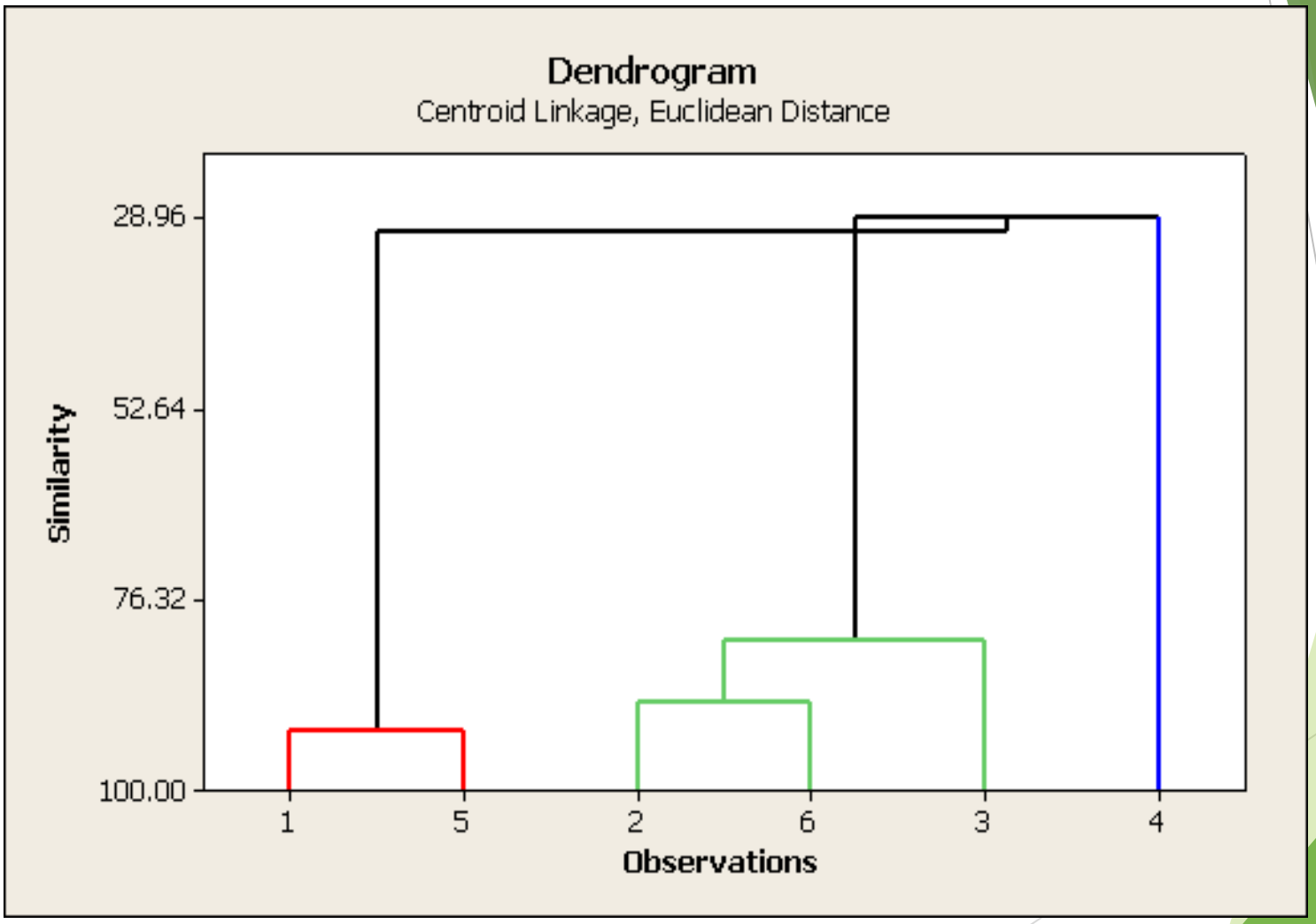
Complete Linkage



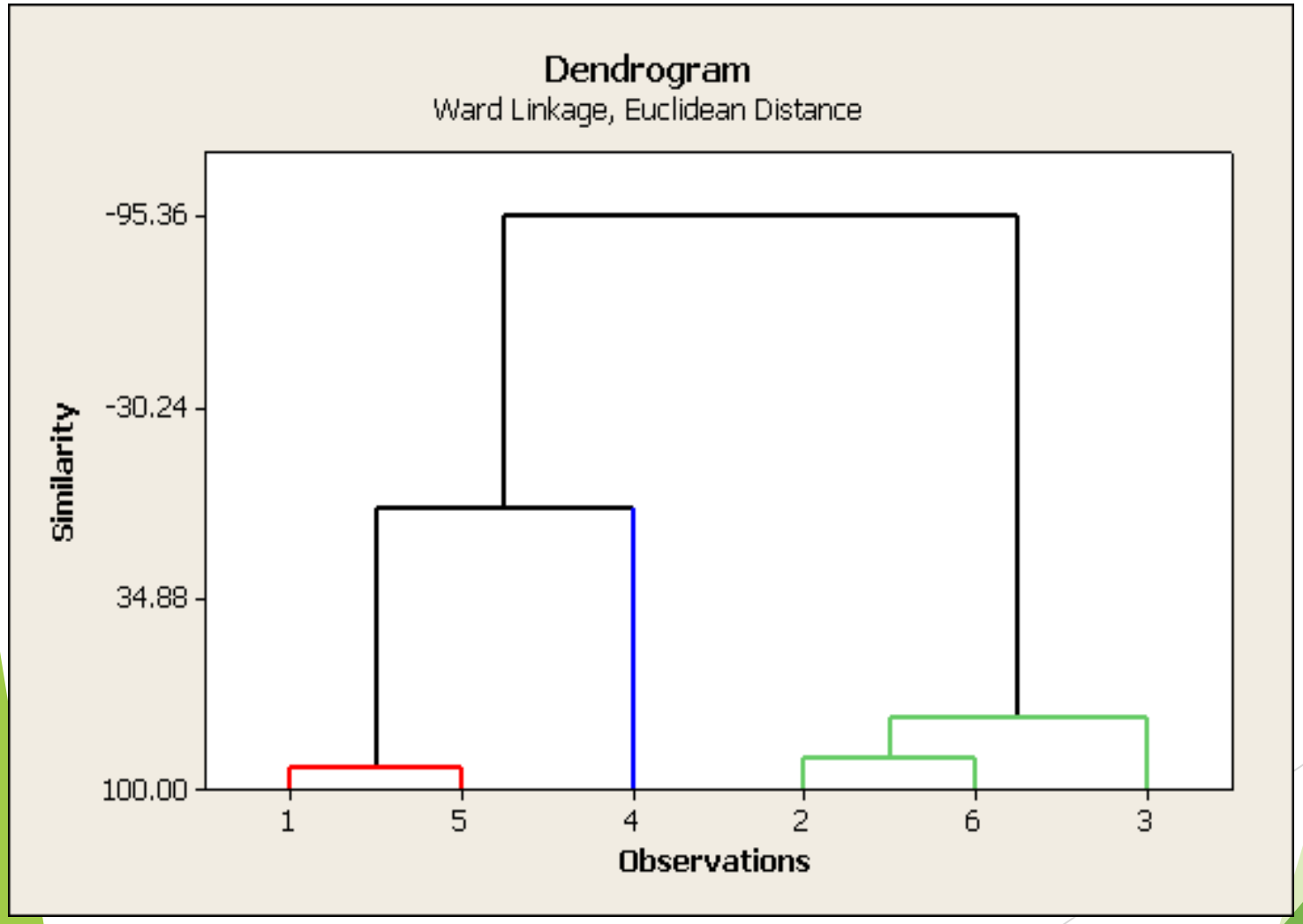
Average Linkage



Centroit Linkage



Ward Linkage



Metode Penggerombolan tak berhirarki

- Metode K rataan (*k-means*)

Algoritmanya sbb :

1. Tentukan besarnya *k*, yaitu banyaknya gerombol, dan tentukan juga centroid di tiap gerombol.
2. Hitung jarak antara setiap objek dengan setiap centroid.
3. Hitung kembali rataan (centroid) untuk gerombol yang baru terbentuk.
4. Ulangi langkah 2 sampai tidak ada lagi pemindahan objek antar gerombol.

Terimakasih