

CLUSTERING ANALYSIS

- Pendahuluan
- Agglomerative Clustering

Pendahuluan

1

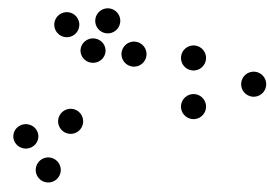
- *Cluster Analysis :*
pengelompokkan data (objek) yang didasarkan pada informasi yang ditemukan dalam data yang menggambarkan objek tersebut dan hubungan diantaranya.
- *Tujuan :*
Agar objek-objek yang bergabung dalam satu kelompok merupakan objek-objek yang mirip/berhubungan satu sama lain dan berbeda/tidak berhubungan dengan objek pada kelompok lain

Pendahuluan

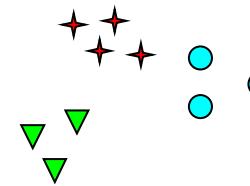
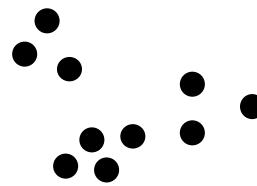
2

- Konsepnya :
 - Lebih besar kemiripan(homogenitas) dalam kelompok
 - Lebih besar perbedaan antara kelompok yang lain
- Clustering = unsupervised classification
 - Label kelas data training tidak diketahui
 - Pembentukan adanya kelas atau kelompok dalam data
- Klasifikasi = supervised classification
 - Ada kelas dalam data training
 - Data baru diklasifikasikan didasarkan pada data training

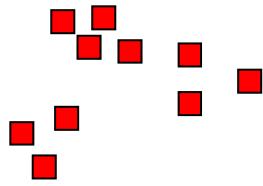
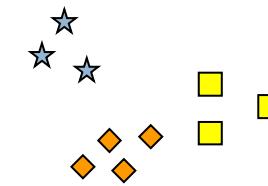
Pendahuluan



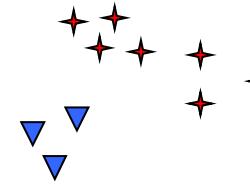
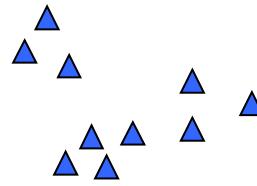
How many clusters?



Six Clusters



Two Clusters



Four Clusters

Penentuan *cluster* terbaik tergantung dari kondisi data serta hasil yang diinginkan

Jenis-jenis Clustering

1. Struktur Pengelompokan :

□ Hierarkhi : metode untuk mengklasterkan / menggerombolkan / mengelompokkan pengamatan secara terstruktur berdasarkan kemiripan sifatnya dan klaster yang diinginkan belum diketahui banyaknya

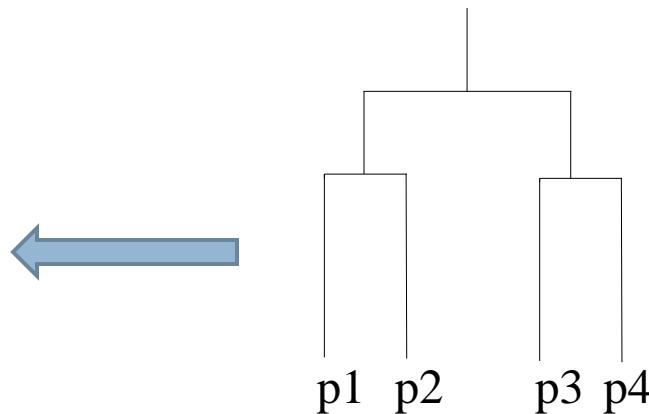
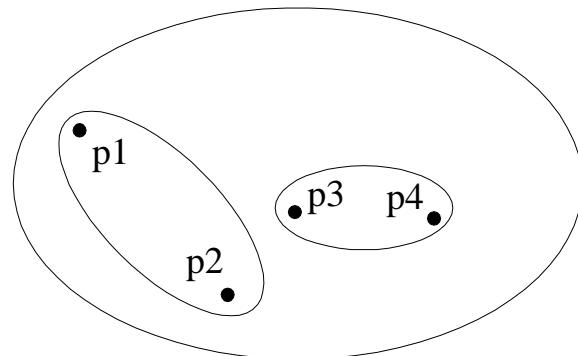
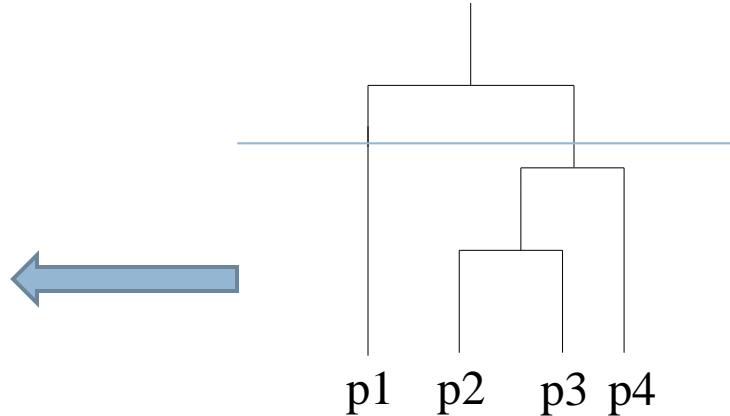
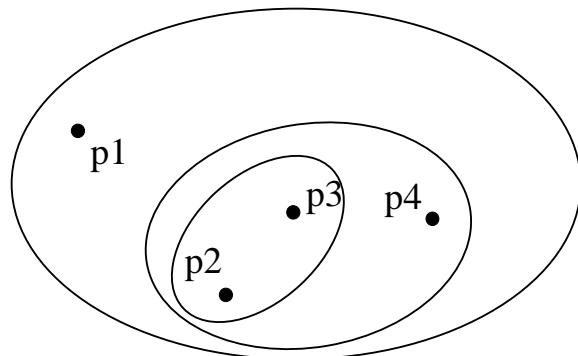
■ Agglomerative

- melakukan pengklasteran dari kecil ke besar
- Setiap data adalah satu klaster sampai semua data menjadi satu klaster
- Dinyatakan dalam bentu dendogram (mirip tree)
- Pilih akan dibuat berapa klaster

■ Divisive

- melakukan pengklasteran dari besar ke kecil
- Semua data adalah satu klaster sampai satu data menjadi satu klaster
- Dinyatakan dalam bentu dendogram (mirip tree)
- Pilih akan dibuat berapa klaster

Hierarchical Clustering



Hierarchical Clustering

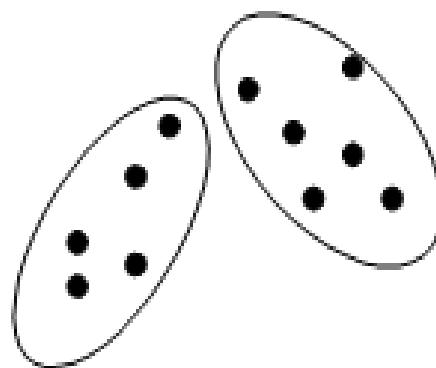
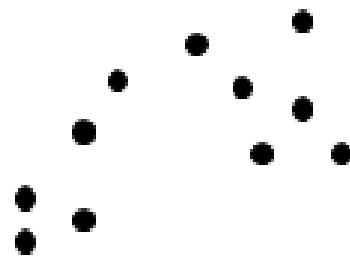
Dendrogram

Jenis-jenis Clustering

6

- Non-hierarkhi /partitioning:

- metode untuk mengklasterkan dengan terlebih dahulu menentukan jumlah klaster yang akan dibentuk



Jenis-jenis Clustering

7

2. Keanggotaan Data

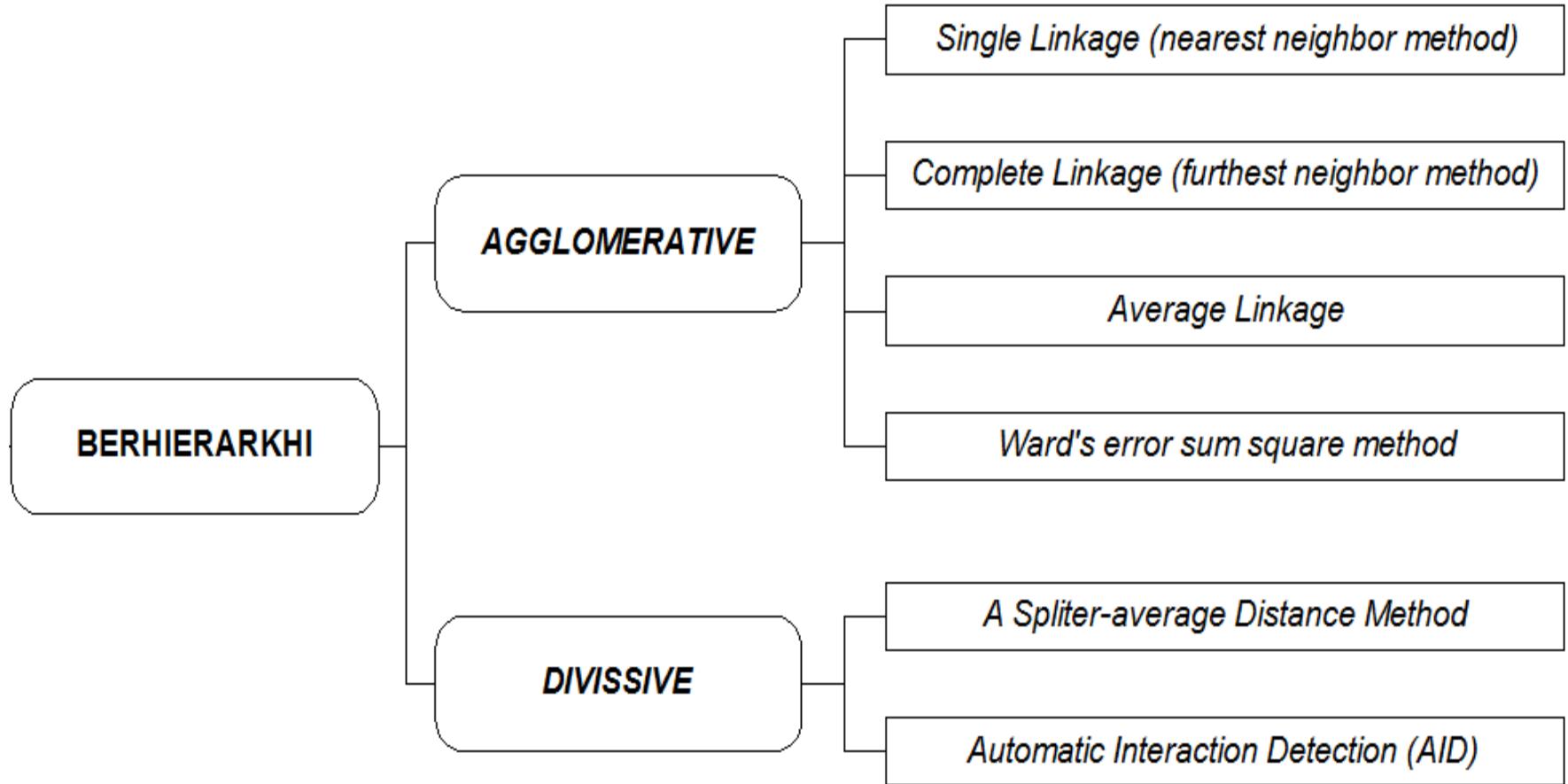
- ❑ Ekslusif : sebuah data bisa dipastikan hanya menjadi anggota satu kelompok saja
- ❑ Non-eksklusif : memperbolehkan suatu data menjadi anggota di lebih dari satu kelompok (fuzzy)

3. Kekompakan Data

- ❑ Komplet : setiap data menjadi anggota klaster
- ❑ Parsial : ada data yang tidak masuk anggota klaster, biasanya data yg outlier/anomali

Agglomerative Clustering

8



Agglomerative Clustering

9

- Metode pengklasteran hierarkhi dengan pendekatan bottom-up
 - Pengklasteran dimulai dari masing-masing data sebagai satu klaster
 - Secara rekursif mencari klaster terdekat sebagai pasangan untuk bergabung menjadi satu kelompok yg lebih besar
 - Proses diulang sehingga membentuk jenjang (hierarkhi)
- Membutuhkan parameter kedekatan klaster (metode single link, complete link, average link, centroid)

Agglomerative Clustering

10

- Metode *Single Linkage (Nearest Neighbor)*
 - Metode pautan tunggal
 - Jarak antara dua klaster dapat ditentukan dari dua obyek berpasangan yang memiliki kesamaan atau memiliki jarak terdekat (*Nearest Neighbor*) dalam klaster berbeda
 - Ukuran kedekatan yang digunakan (MIN):

$$d_{UV} = \min\{d_{U_iV_j}\} \quad d_{U_iV_j} \in D$$

- U, V = klaster
- $d_{U_iV_j}$ = jarak data ke-i di U dengan data ke-j di V

Agglomerative Clustering

11

- **Complete Linkage (Furthest Neighbor)**
 - Metode pautan lengkap,
 - Jarak antara dua klaster ditentukan dari jarak terjauh antara dua obyek pada klaster yang berbeda (*Furthest Neighbor*)
 - Ukuran kedekatan yang digunakan (MAX):

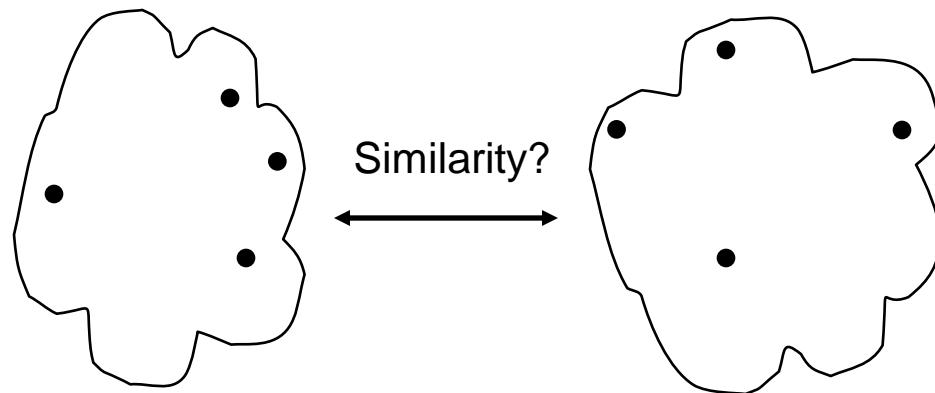
$$d_{UV} = \max\{d_{U_iV_j}\} \quad d_{U_iV_j} \in D$$

Agglomerative Clustering

12

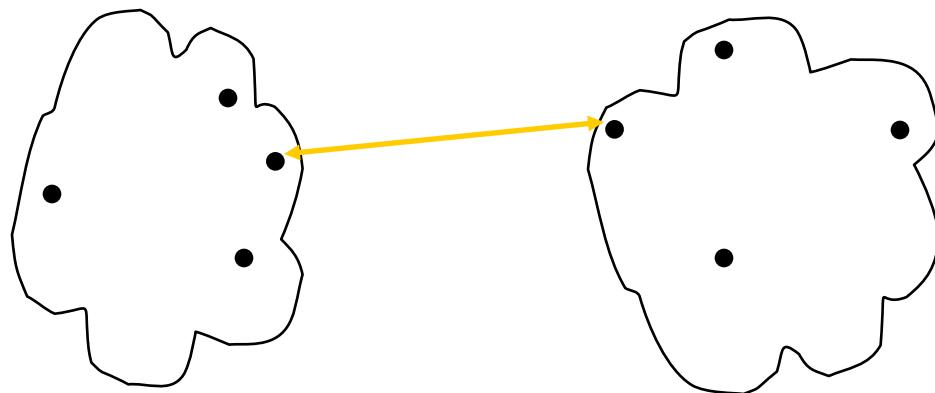
- *Average Linkage*
 - Metode pautan rataan
 - Jarak antar dua klaster ditentukan dari nilai rata-rata jarak seluruh obyek suatu klaster terhadap seluruh obyek pada klaster lainnya
 - Pada berbagai keadaan, metode ini dianggap lebih stabil dibandingkan dengan kedua metode di atas
- *Metode centroid*

Agglomerative Clustering



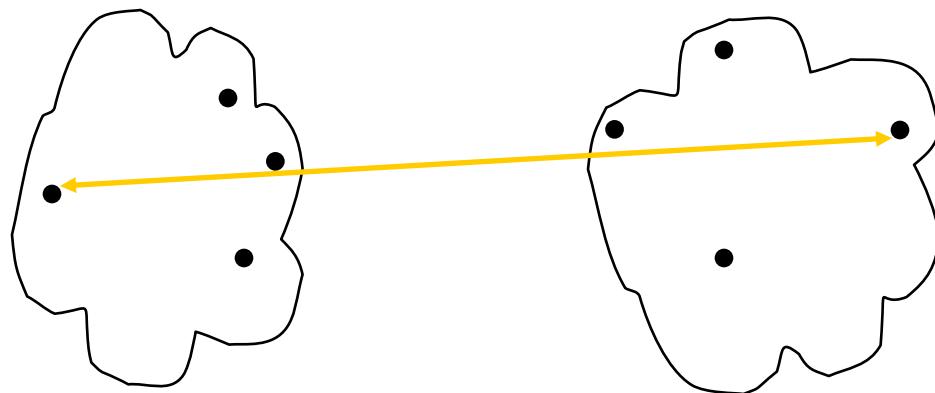
- Single Linkage (MIN)
- Complete Linkage (MAX)
- Average Linkage
- Centroid

Agglomerative Clustering



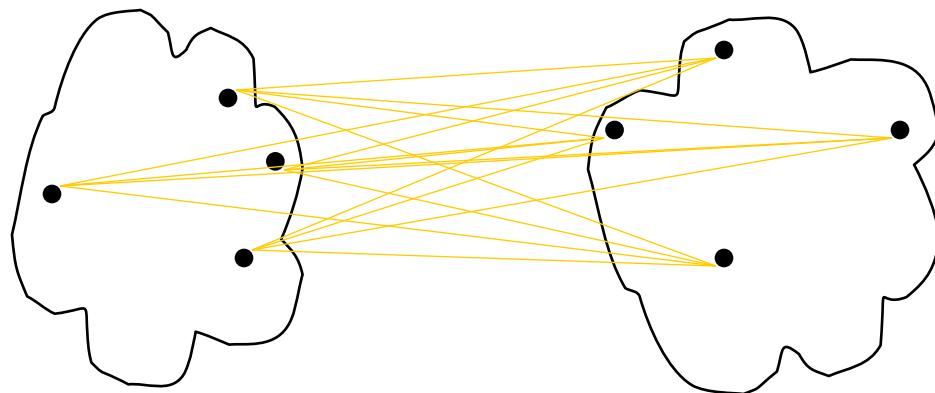
- Single Linkage (MIN)
- Complete Linkage (MAX)
- Average Linkage
- Centroid

Agglomerative Clustering



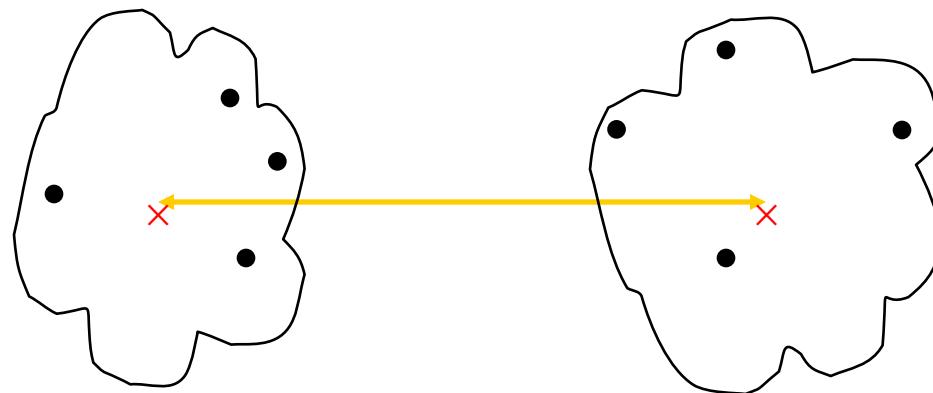
- Single Linkage (MIN)
- Complete Linkage (MAX)
- Average Linkage
- Centroid

Agglomerative Clustering



- Single Linkage (MIN)
- Complete Linkage (MAX)
- Average Linkage
- Centroid

Agglomerative Clustering



- Single Linkage (MIN)
- Complete Linkage (MAX)
- Average Linkage
- Centroid

Agglomerative Clustering

18

□ Algoritma Agglomerative Clustering

1. Hitung matriks kedekatan antar data berdasarkan jarak yang digunakan
2. Gabungkan 2 data/klaster berdasarkan jarak terdekat
3. Buat matriks kedekatan baru setelah beberapa data/klaster digabung, berdasarkan parameter kedekatan yang digunakan
4. Ulangi langkah 2 dan 3, sampai menjadi satu klaster

Agglomerative Clustering

19

□ Contoh :

Terdapat matriks jarak antara 5 buah obyek/data

	A	B	C	D	E
A	0.0	1.0	5.0	6.0	8.0
B	1.0	0.0	3.0	8.0	7.0
C	5.0	3.0	0.0	4.0	6.0
D	6.0	8.0	4.0	0.0	2.0
E	8.0	7.0	6.0	2.0	0.0

Agglomerative Clustering

20

Langkah penyelesaiannya :

1. Ok.
2. Gabung 2 data/klaster terdekat

- $d(A,B)=1$ paling dekat
- A dan B menjadi satu klaster AB

3. Buat matriks kedekatan baru, hitung jarak baru antara klaster/data (**SINGLE LINKAGE**)

- $d_{(AB)C} = \min \{d_{AC}, d_{BC}\} = \min (5,3) = d_{BC} = 3$
- $d_{(AB)D} = \min \{d_{AD}, d_{BD}\} = \min (6,8) = d_{AD} = 6$
- $d_{(AB)E} = \min \{d_{AE}, d_{BE}\} = \min (8,7) = d_{BE} = 7$

	A	B	C	D	E
A	0.0	1.0	5.0	6.0	8.0
B	1.0	0.0	3.0	8.0	7.0
C	5.0	3.0	0.0	4.0	6.0
D	6.0	8.0	4.0	0.0	2.0
E	8.0	7.0	6.0	2.0	0.0

Agglomerative Clustering

	A	B	C	D	E
A	0.0	1.0	5.0	6.0	8.0
B	1.0	0.0	3.0	8.0	7.0
C	5.0	3.0	0.0	4.0	6.0
D	6.0	8.0	4.0	0.0	2.0
E	8.0	7.0	6.0	2.0	0.0

	AB	C	D	E
AB	0.0	3.0	6.0	7.0
C	3.0	0.0	4.0	6.0
D	6.0	4.0	0.0	2.0
E	7.0	6.0	2.0	0.0

- Gabung 2 klaster terdekat
 - $d(D,E) = 2$ paling dekat, D dan E menjadi satu klaster DE
- Buat matriks kedekatan baru, hitung jarak baru antara klaster
 - $d_{(AB)DE} = \min \{d_{AD}, d_{AE}, d_{BD}, d_{BE}\} = d_{AD} = 6$
 - $d_{C(DE)} = \min \{d_{CD}, d_{CE}\} = d_{CD} = 4$

Agglomerative Clustering

2

	A	B	C	D	E
A	0.0	1.0	5.0	6.0	8.0
B	1.0	0.0	3.0	8.0	7.0
C	5.0	3.0	0.0	4.0	6.0
D	6.0	8.0	4.0	0.0	2.0
E	8.0	7.0	6.0	2.0	0.0

	AB	C	DE
AB	0	3	6
C	3	0	4
DE	6	4	0

- Gabung 2 klaster terdekat
 - $d((AB),C) = 3$ paling dekat, AB dan C menjadi satu klaster ABC
- Buat matriks kedekatan baru, hitung jarak baru antara klaster
 - $d_{(ABC)(DE)} = \min \{d_{AD}, d_{AE}, d_{BD}, d_{BE}, d_{CD}, d_{CE}\}$
 $= d_{CD} = 4$

Agglomerative Clustering

23

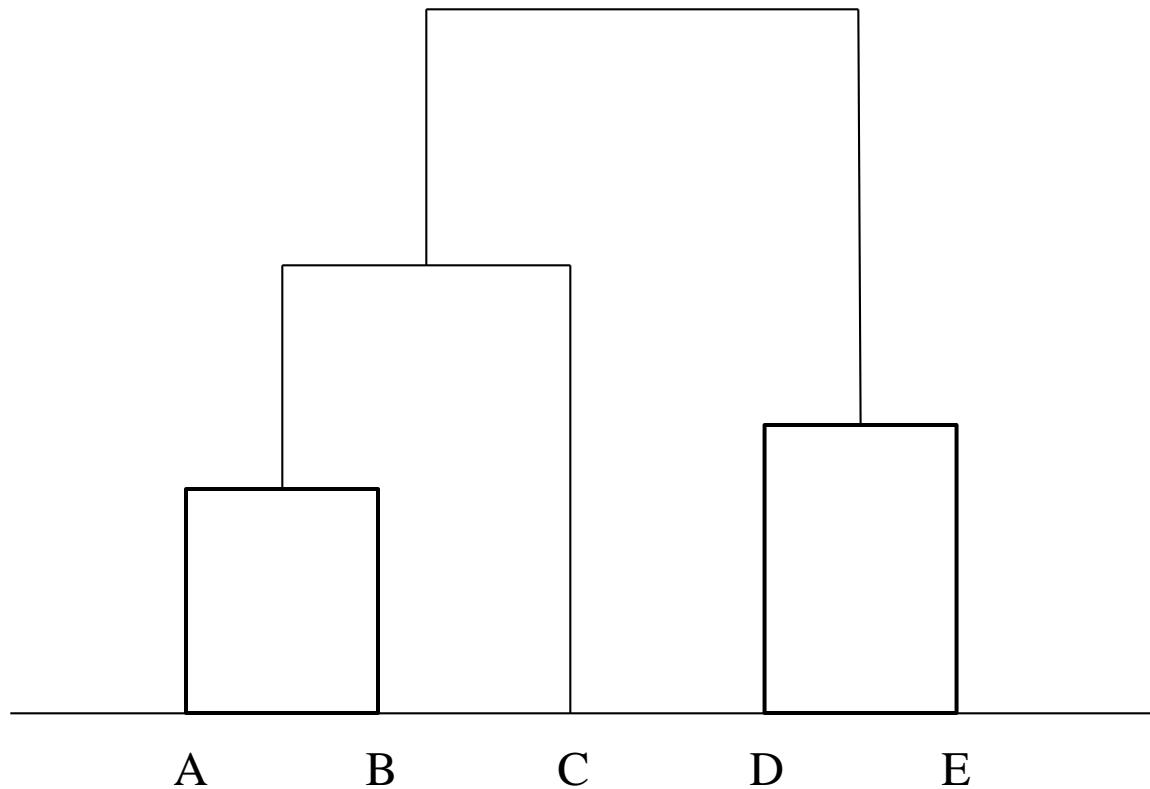
	ABC	DE
ABC	0	4
DE	4	0

- Gabung 2 klaster terakhir (ABC) dengan (DE)

Agglomerative Clustering

24

Dendogram



Latihan

25

Data	Atribut P	Atribut Q
A	1	1
B	4	1
C	1	2
D	3	4
E	5	4

- Buat pengklasteran dari data diatas dengan metode
 - Single Linkage
 - Complete Linkage
- (Gunakan konsep jarak Manhattan)

Tugas

26

- Buat klaster untuk data (contoh) dan (latihan) menggunakan metode average linkage dengan konsep jarak euclid dan manhattan

Terima Kasih

CLUSTERING ANALYSIS

K-MEANS

Pendahuluan

1

- K-Means, termasuk pada *partitioning clustering*
- Memisahkan data ke k daerah bagian yang terpisah (kelompok/klaster)
- Setiap data harus masuk ke klaster tertentu
- Jumlah klaster (k), ditentukan terlebih dahulu
- Perlu penentuan pusat (sentroid) dari klaster yang berdasarkan rataan

Pendahuluan

2

- Memungkinkan bagi setiap data yang termasuk di klaster tertentu pada suatu tahapan proses, pada tahap berikutnya berpindah ke klaster yang lain

- Kemudahan dan kemampuan untuk meng-klaster data besar dan data outlier dengan sangat cepat

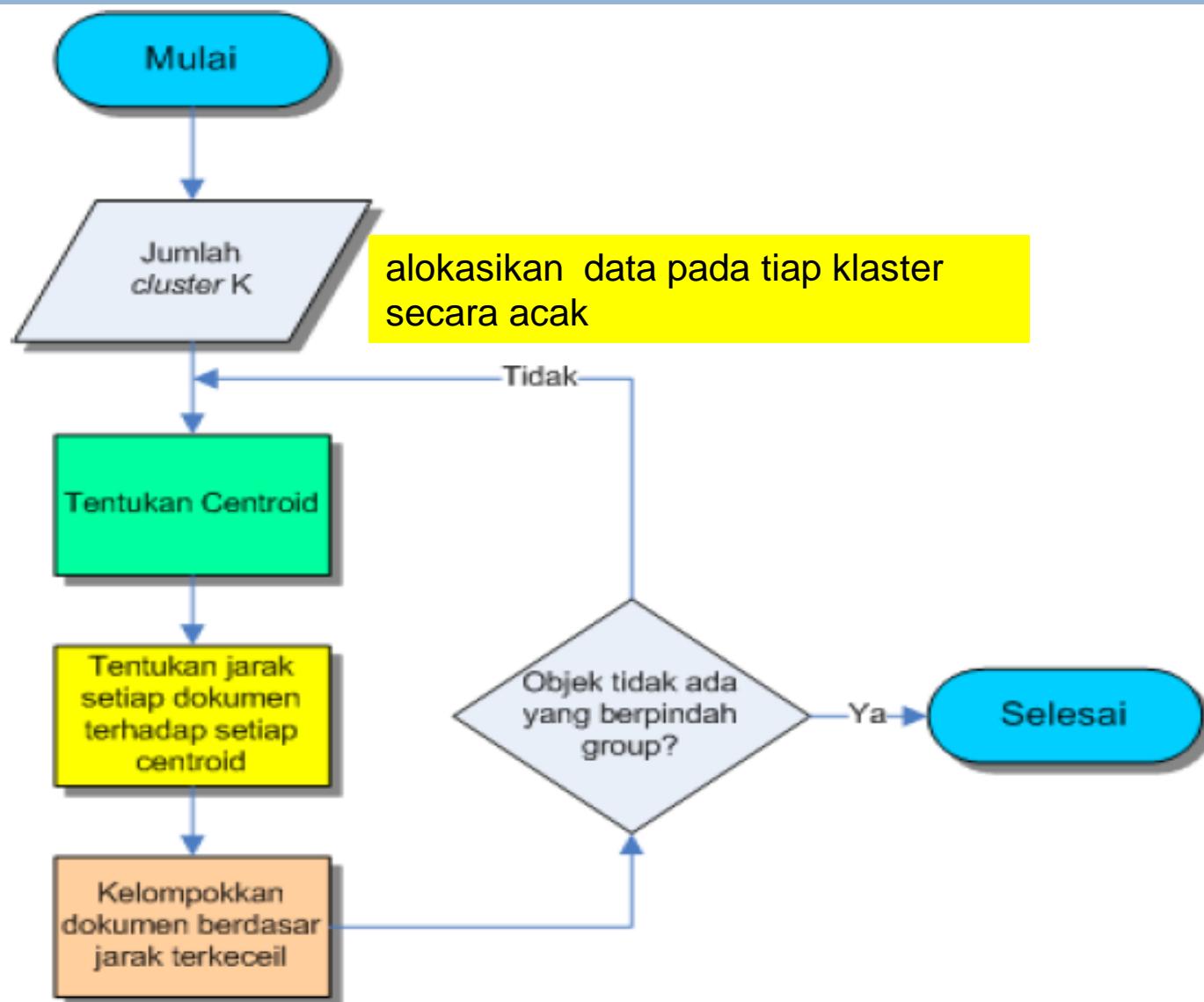
Algoritma K-Means Clustering

3

1. Tentukan jumlah klaster (K)
2. Alokasikan data ke dalam klaster secara acak
3. Hitung pusat (sentroid) klaster dari data yang ada di masing-masing klaster
4. Hitung jarak tiap data ke sentroid tiap klaster
5. Alokasikan masing-masing data ke klaster yang jaraknya terdekat
6. Jika tidak ada data yg pindah ke klaster yg lain, maka clustering selesai. Jika ada yg pindah klaster maka embali ke langkah 3.

Algoritma K-Means Clustering

4



Contoh Kasus

5

- Jika terdapat 10 data dengan 2 atribut/fitur. Buatlah data ini menjadi 3 klaster

Data ke - i	Atribut x	Atribut y
1	1	1
2	4	1
3	6	1
4	1	2
5	2	3
6	5	3
7	2	5
8	3	5
9	2	6
10	3	8

Contoh Kasus

6

- Alokasikan ke-10 data secara random pada 3 klaster

Data ke - i	Atribut x	Atribut y	C1	C2	C3
1	1	1	*		
2	4	1			*
3	6	1		*	
4	1	2		*	
5	2	3			*
6	5	3		*	
7	2	5		*	
8	3	5			*
9	2	6			*
10	3	8		*	

Contoh Kasus

7

- **ITERASI 1**
- Hitung sentroid masing-masing klaster
- Sentroid C1 :

	Data ke - i	Atribut x	Atribut y
	1	1	1
Jumlah data/ Total nilai data	1	1	1
Sentroid (rataan)		1	1

Contoh Kasus

8

□ Sentroid C2 :

	Data ke - i	Atribut x	Atribut y
	3	6	1
	4	1	2
	6	5	3
	7	2	5
	10	3	8
Jumlah data/ Total nilai data	5	17	19
Sentroid (rataan)		3.4	3.8

Contoh Kasus

9

□ Sentroid C3 :

	Data ke - i	Atribut x	Atribut y
	2	4	1
	5	2	3
	8	3	5
	9	2	6
Jumlah data/ Total nilai data	4	11	15
Sentroid (rataan)		2.75	3.75

Contoh Kasus

10

- Hitung jarak tiap data pada sentroid tiap klaster
- Alokasikan tiap data pada klaster baru (berdasar jarak terdekat)

Data ke - i	Atribut x	Atribut y	c1	c2	c3	Min	Klaster
1	1	1	0.0000	3.6878	3.2596	0.0000	1
2	4	1	3.0000	2.8636	3.0208	2.8636	2
3	6	1	5.0000	3.8210	4.2573	3.8210	2
4	1	2	1.0000	3.0000	2.4749	1.0000	1
5	2	3	2.2361	1.6125	1.0607	1.0607	3
6	5	3	4.4721	1.7889	2.3717	1.7889	2
7	2	5	4.1231	1.8439	1.4577	1.4577	3
8	3	5	4.4721	1.2649	1.2748	1.2649	2
9	2	6	5.0990	2.6077	2.3717	2.3717	3
10	3	8	7.2801	4.2190	4.2573	4.2190	2

Contoh Kasus

11

Data ke - i	Atribut x	Atribut y	Klaster Lama	Klaster Baru
1	1	1	1	1
2	4	1	3	2
3	6	1	2	2
4	1	2	2	1
5	2	3	3	3
6	5	3	2	2
7	2	5	2	3
8	3	5	3	2
9	2	6	3	3
10	3	8	2	2

- Ada data yg pindah klaster, lakukan iterasi selanjutnya

Contoh Kasus

12

- **ITERASI 2**
- Hitung sentroid masing-masing klaster
- Sentroid C1 :

	Data ke - i	Atribut x	Atribut y
	1	1	1
	4	1	2
Jumlah data/ Total nilai data	2	2	3
Sentroid (rataan)		1	1.5

Contoh Kasus

13

□ Sentroid C2 :

	Data ke - i	Atribut x	Atribut y
	2	4	1
	3	6	1
	6	5	3
	8	3	5
	10	3	8
Jumlah data/ Total nilai data	5	21	18
Sentroid (rataan)		3.4	3.6

Contoh Kasus

14

□ Sentroid C3 :

	Data ke - i	Atribut x	Atribut y
	5	2	3
	7	2	5
	9	2	6
Jumlah data/ Total nilai data	3	6	14
Sentroid (rataan)		2	4.667

Contoh Kasus

15

- Hitung jarak tiap data pada sentroid tiap klaster
- Alokasikan tiap data pada klaster baru (berdasar jarak terdekat)

Data ke - i	Atribut x	Atribut y	c1	c2	c3	Min	Klaster
1	1	1	0.5000	4.1231	3.8006	0.5000	1
2	4	1	3.0414	2.6077	4.1767	2.6077	2
3	6	1	5.0249	3.1623	5.4263	3.1623	2
4	1	2	0.5000	3.5777	2.8480	0.5000	1
5	2	3	1.8028	2.2804	1.6667	1.6667	3
6	5	3	4.2720	1.0000	3.4319	1.0000	2
7	2	5	3.6401	2.6077	0.3333	0.3333	3
8	3	5	4.0311	1.8439	1.0541	1.0541	3
9	2	6	4.6098	3.2558	1.3333	1.3333	3
10	3	8	6.8007	4.5607	3.4801	3.4801	3

Contoh Kasus

16

Data ke - i	Atribut x	Atribut y	Klaster Lama	Klaster Baru
1	1	1	1	1
2	4	1	2	2
3	6	1	2	2
4	1	2	1	1
5	2	3	3	3
6	5	3	2	2
7	2	5	3	3
8	3	5	2	3
9	2	6	3	3
10	3	8	2	3

- Ada data yg pindah klaster, lakukan iterasi selanjutnya

Contoh Kasus

17

- **ITERASI 3**
- Hitung sentroid masing-masing klaster
- Sentroid C1 :

	Data ke - i	Atribut x	Atribut y
	1	1	1
	4	1	2
Jumlah data/ Total nilai data	2	2	3
Sentroid (rataan)		1	1.5

Contoh Kasus

18

□ Sentroid C2 :

	Data ke - i	Atribut x	Atribut y
	2	4	1
	3	6	1
	6	5	3
Jumlah data/ Total nilai data	3	15	4
Sentroid (rataan)		5	1.6667

Contoh Kasus

19

□ Sentroid C3 :

	Data ke - i	Atribut x	Atribut y
	5	2	3
	7	2	5
	8	3	5
	9	2	6
	10	3	8
Jumlah data/ Total nilai data	5	12	27
Sentroid (rataan)		2.4	5.6

Contoh Kasus

20

- Hitung jarak tiap data pada sentroid tiap klaster
- Alokasikan tiap data pada klaster baru (berdasar jarak terdekat)

Data ke - i	Atribut x	Atribut y	c1	c2	c3	Min	Klaster
1	1	1	0.5000	4.0552	4.6174	0.5000	1
2	4	1	3.0414	1.2019	4.6819	1.2019	2
3	6	1	5.0249	1.2019	5.6851	1.2019	2
4	1	2	0.5000	4.0139	3.6770	0.5000	1
5	2	3	1.8028	3.2829	2.4331	1.8028	1
6	5	3	4.2720	1.3333	3.5384	1.3333	2
7	2	5	3.6401	4.4845	0.5657	0.5657	3
8	3	5	4.0311	3.8873	0.7211	0.7211	3
9	2	6	4.6098	5.2704	0.7211	0.7211	3
10	3	8	6.8007	6.6416	2.6683	2.6683	3

Contoh Kasus

21

Data ke - i	Atribut x	Atribut y	Klaster Lama	Klaster Baru
1	1	1	1	1
2	4	1	2	2
3	6	1	2	2
4	1	2	1	1
5	2	3	3	1
6	5	3	2	2
7	2	5	3	3
8	3	5	3	3
9	2	6	3	3
10	3	8	3	3

- Ada data yg pindah klaster, lakukan iterasi selanjutnya

Contoh Kasus

22

- **ITERASI 4**
- Hitung sentroid masing-masing klaster
- Sentroid C1 :

	Data ke - i	Atribut x	Atribut y
	1	1	1
	4	1	2
	5	2	3
Jumlah data/ Total nilai data	3	4	6
Sentroid (rataan)		1.3333	2

Contoh Kasus

23

□ Sentroid C2 :

	Data ke - i	Atribut x	Atribut y
	2	4	1
	3	6	1
	6	5	3
Jumlah data/ Total nilai data	3	15	4
Sentroid (rataan)		5	1.6667

Contoh Kasus

24

□ Sentroid C3 :

	Data ke - i	Atribut x	Atribut y
	7	2	5
	8	3	5
	9	2	6
	10	3	8
Jumlah data/ Total nilai data	4	10	24
Sentroid (rataan)		2.5	6

Contoh Kasus

25

- Hitung jarak tiap data pada sentroid tiap klaster
- Alokasikan tiap data pada klaster baru (berdasar jarak terdekat)

Data ke - i	Atribut x	Atribut y	c1	c2	c3	Min	Klaster
1	1	1	1.0541	4.0552	5.2202	1.0541	1
2	4	1	2.8480	1.2019	5.2202	1.2019	2
3	6	1	4.7726	1.2019	6.1033	1.2019	2
4	1	2	0.3333	4.0139	4.2720	0.3333	1
5	2	3	1.2019	3.2829	3.0414	1.2019	1
6	5	3	3.8006	1.3333	3.9051	1.3333	2
7	2	5	3.0732	4.4845	1.1180	1.1180	3
8	3	5	3.4319	3.8873	1.1180	1.1180	3
9	2	6	4.0552	5.2704	0.5000	0.5000	3
10	3	8	6.2272	6.6416	2.0616	2.0616	3

Contoh Kasus

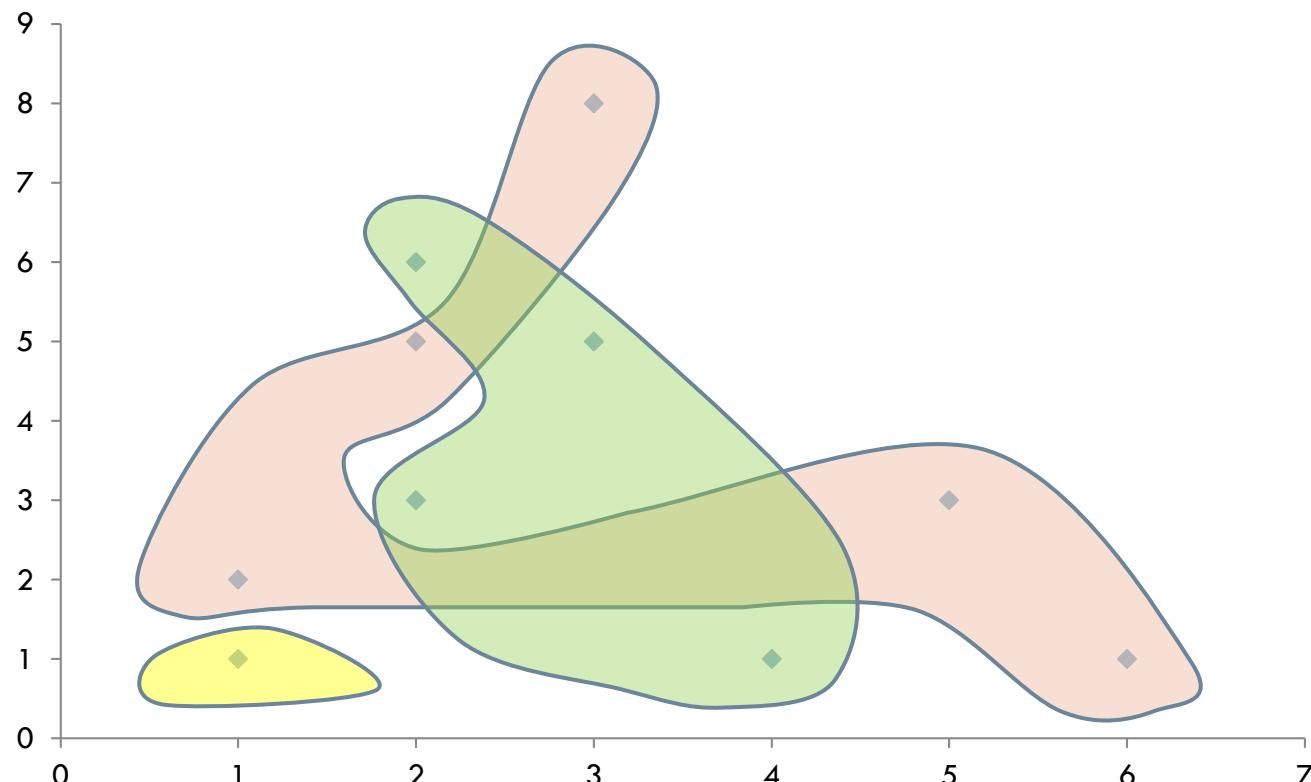
26

Data ke - i	Atribut x	Atribut y	Klaster Lama	Klaster Baru
1	1	1	1	1
2	4	1	2	2
3	6	1	2	2
4	1	2	1	1
5	2	3	1	1
6	5	3	2	2
7	2	5	3	3
8	3	5	3	3
9	2	6	3	3
10	3	8	3	3

- Tidak ada data yg pindah klaster, iterasi selesai

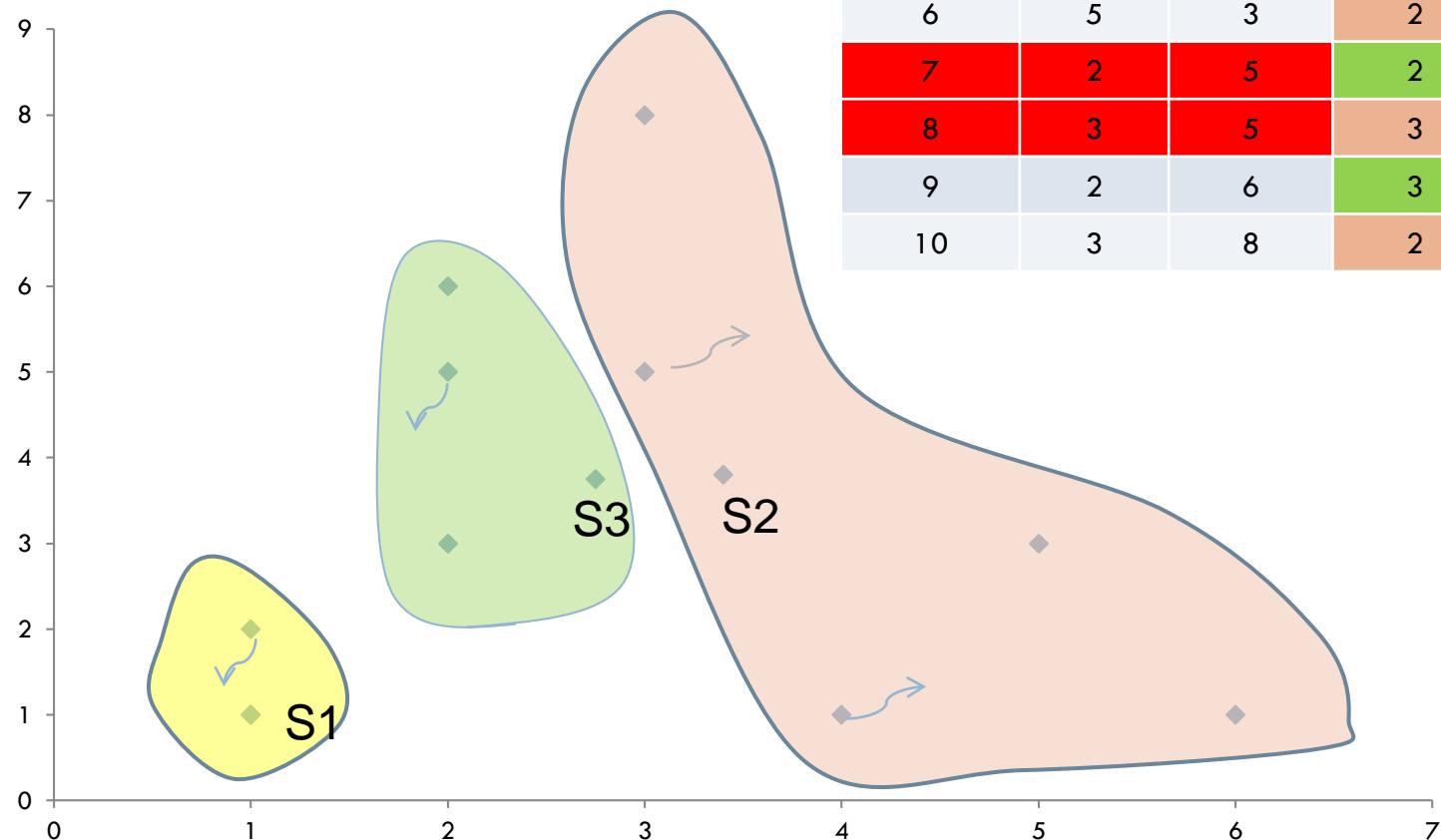
Contoh Kasus

27



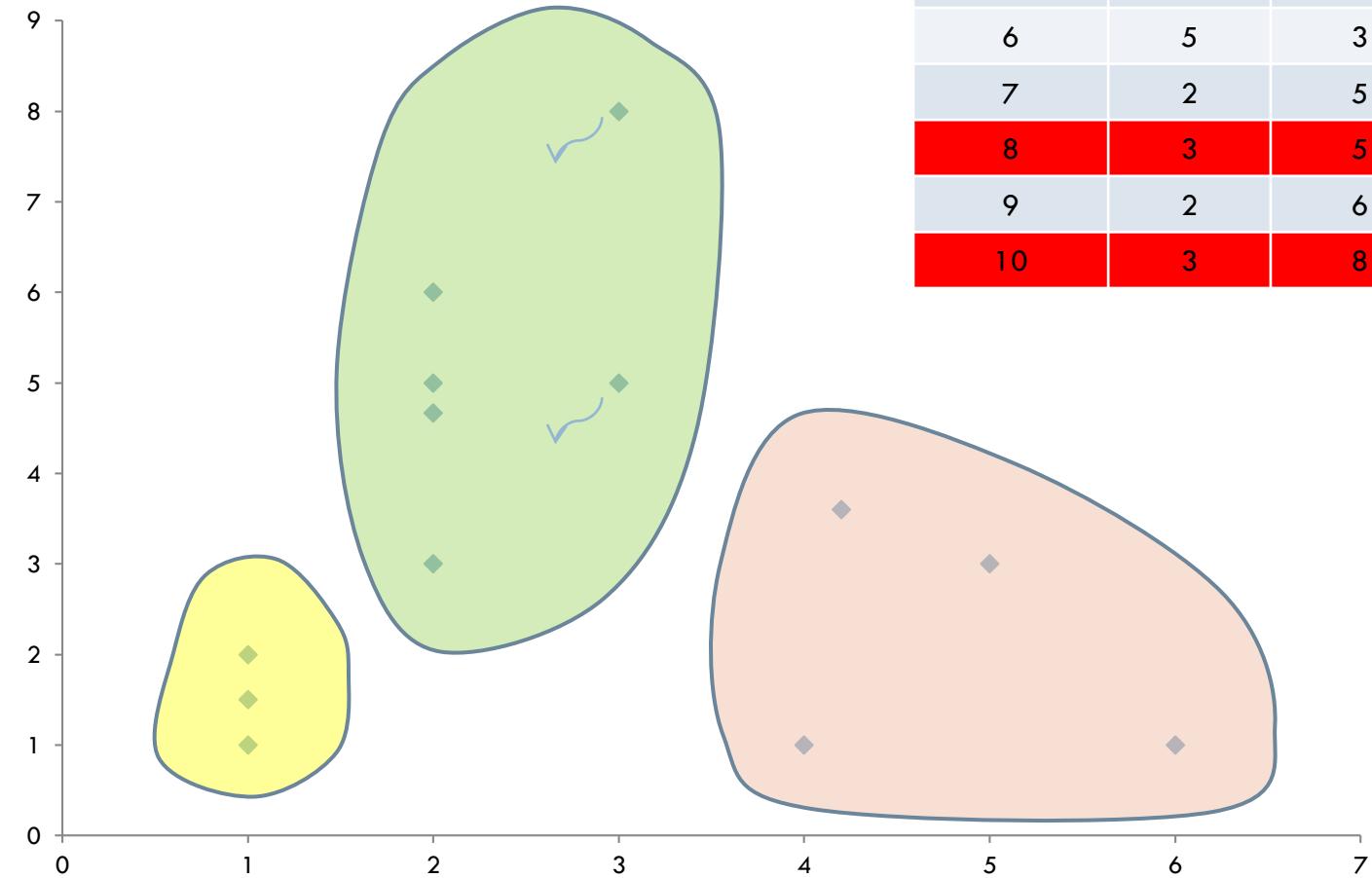
Contoh Kasus

28



Contoh Kasus

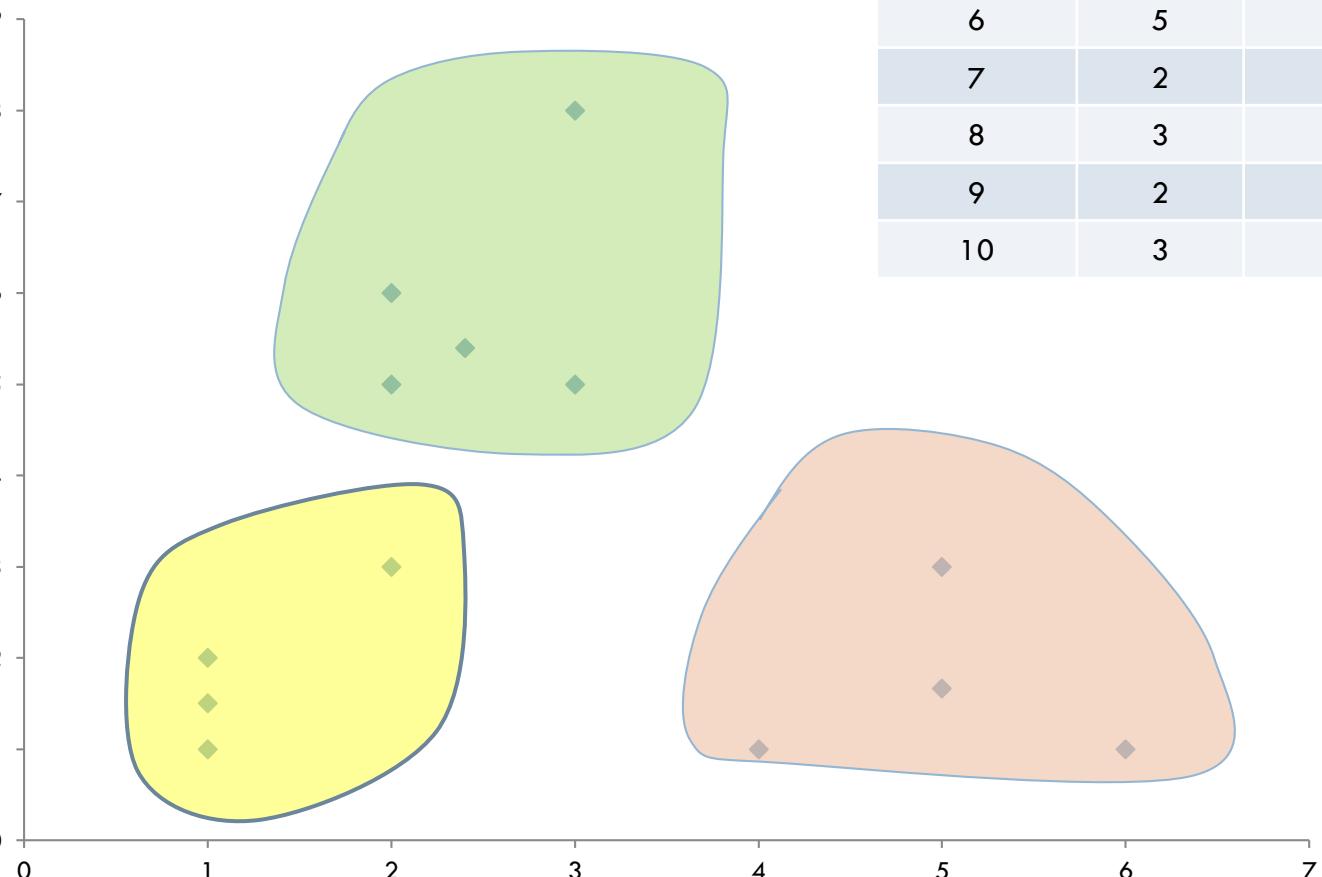
29



Data ke - i	Atribut x	Atribut y	Klaster Lama	Klaster Baru
1	1	1	1	1
2	4	1	2	2
3	6	1	2	2
4	1	2	1	1
5	2	3	3	3
6	5	3	2	2
7	2	5	3	3
8	3	5	2	3
9	2	6	3	3
10	3	8	2	3

Contoh Kasus

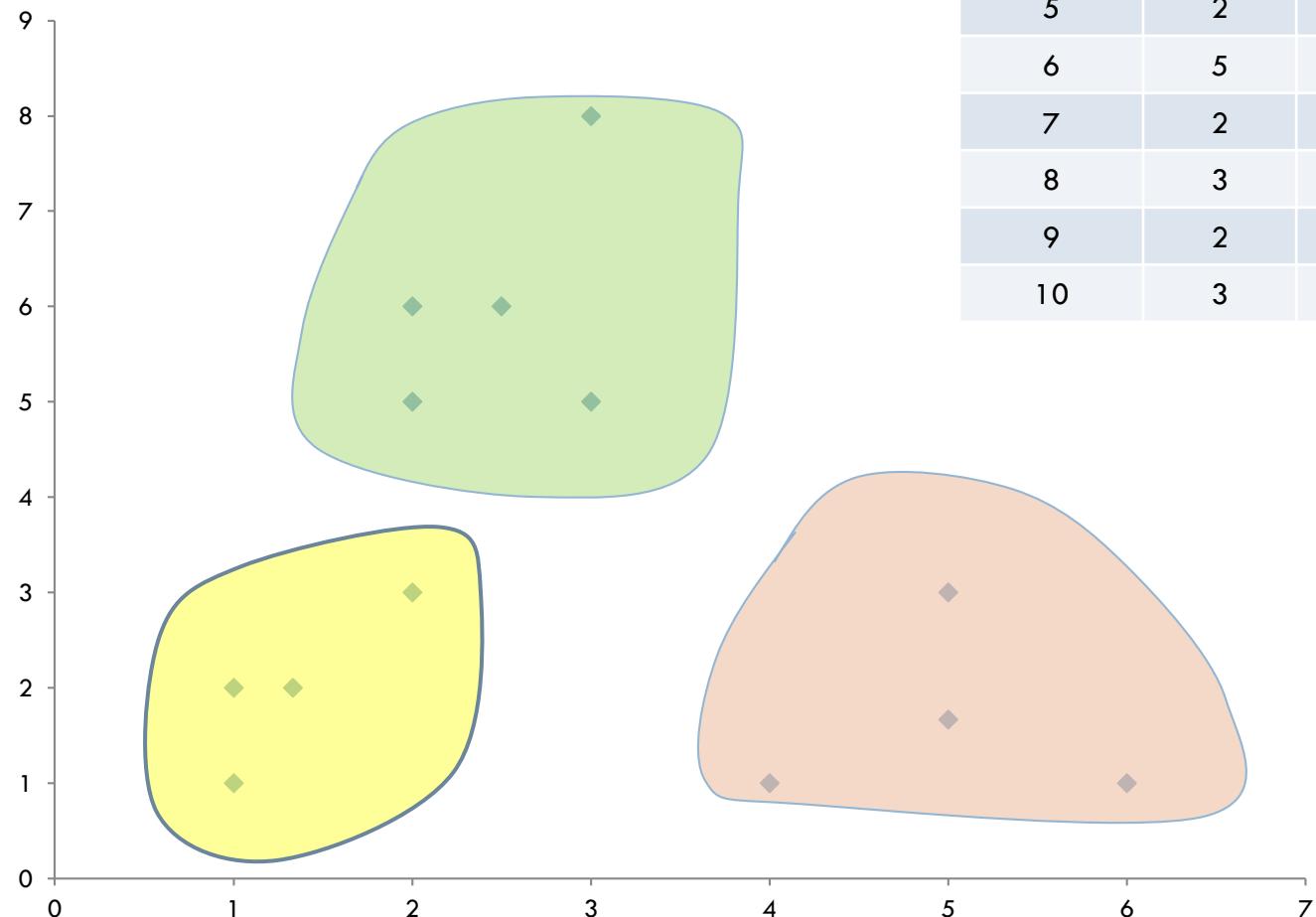
30



Data ke - i	Atribut x	Atribut y	Klaster Lama	Klaster Baru
1	1	1	1	1
2	4	1	2	2
3	6	1	2	2
4	1	2	1	1
5	2	3	3	1
6	5	3	2	2
7	2	5	3	3
8	3	5	3	3
9	2	6	3	3
10	3	8	3	3

Contoh Kasus

31



Data ke - i	Atribut x	Atribut y	Klaster Lama	Klaster Baru
1	1	1	1	1
2	4	1	2	2
3	6	1	2	2
4	1	2	1	1
5	2	3	1	1
6	5	3	2	2
7	2	5	3	3
8	3	5	3	3
9	2	6	3	3
10	3	8	3	3

Contoh Kasus

32

□ Kesimpulan :

- Klaster 1 dengan sentroid (1.3333, 2):
terdiri dari Data ke-1, Data ke-4, dan Data ke-5
- Klaster 2 dengan sentroid (5, 1.6667):
terdiri dari Data ke-2, Data ke-3, dan Data ke-6
- Klaster 3 dengan sentroid (2.5, 6):
terdiri dari Data ke-7, Data ke-8, Data ke-9, dan
Data ke-10

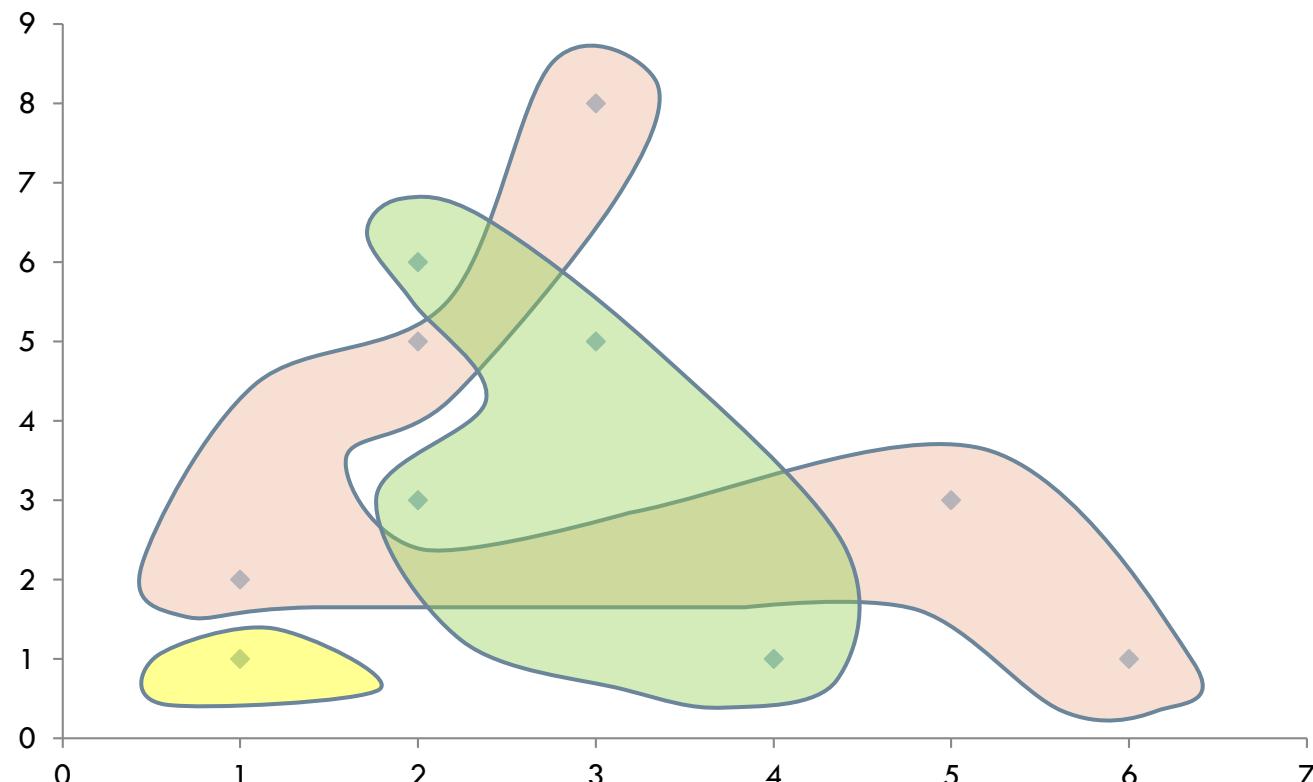
Soal

33

- Gunakan jarak Manhattan

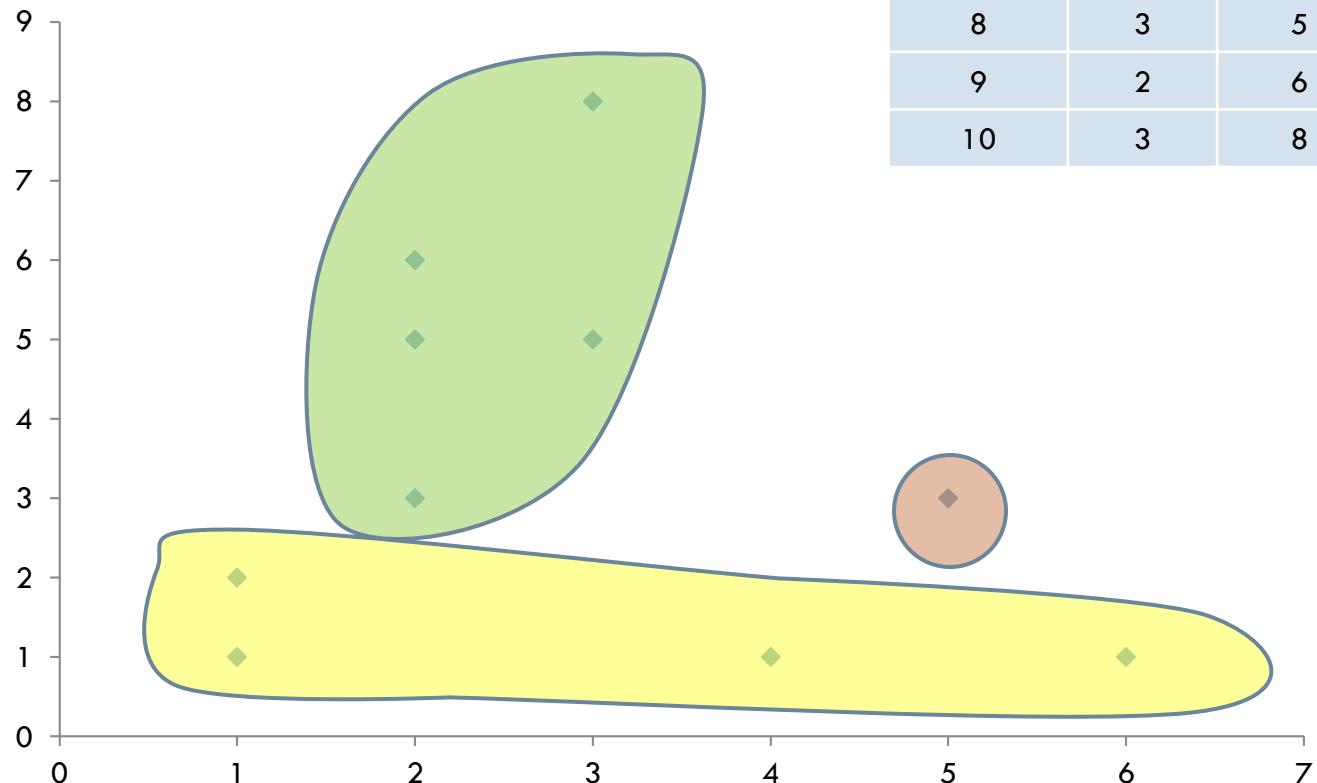
Contoh Kasus

34



Contoh Kasus

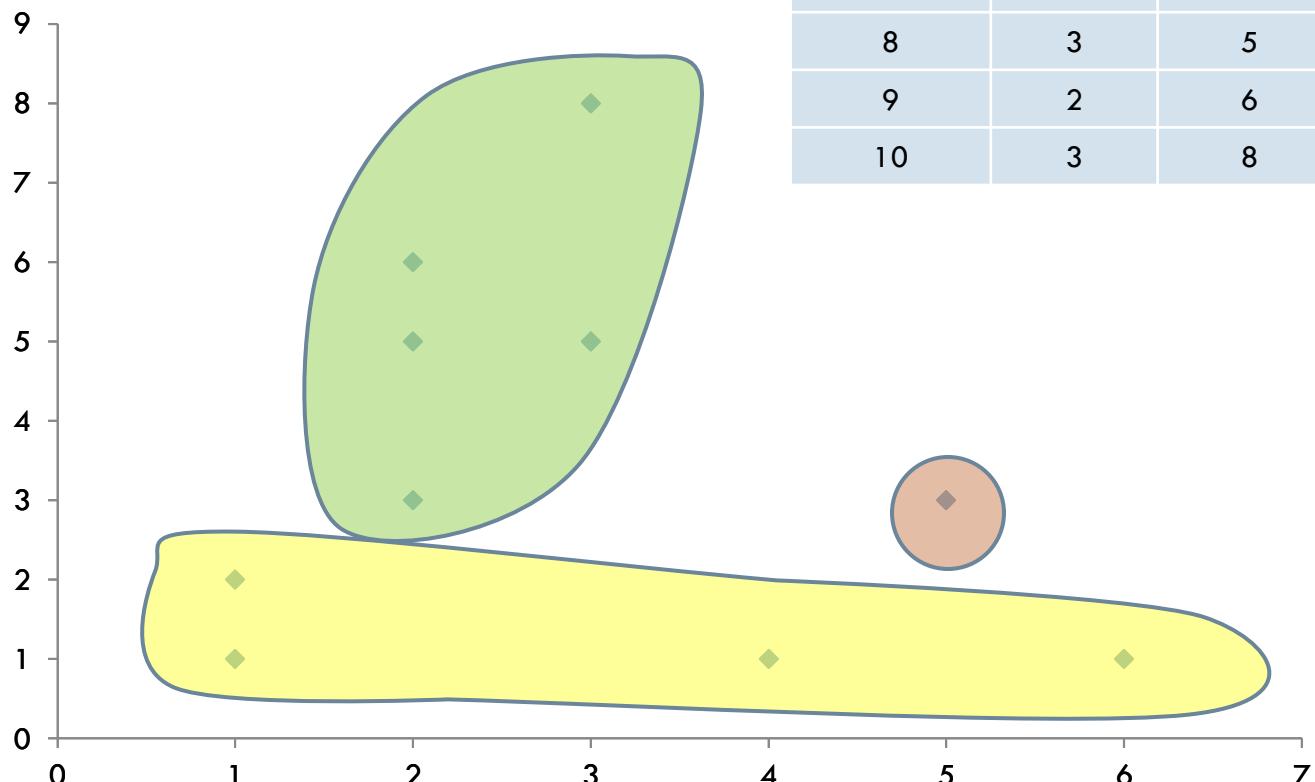
35



Data ke - i	Atribut x	Atribut y	Klaster Lama	Klaster Baru
1	1	1	1	1
2	4	1	3	1
3	6	1	2	1
4	1	2	2	1
5	2	3	3	3
6	5	3	2	2
7	2	5	2	3
8	3	5	3	3
9	2	6	3	3
10	3	8	2	3

Contoh Kasus

36



Data ke - i	Atribut x	Atribut y	Klaster Lama	Klaster Baru
1	1	1	1	1
2	4	1	2	2
3	6	1	2	2
4	1	2	1	1
5	2	3	3	1
6	5	3	2	2
7	2	5	3	3
8	3	5	3	3
9	2	6	3	3
10	3	8	3	3

Terima Kasih