

K-Means Clustering

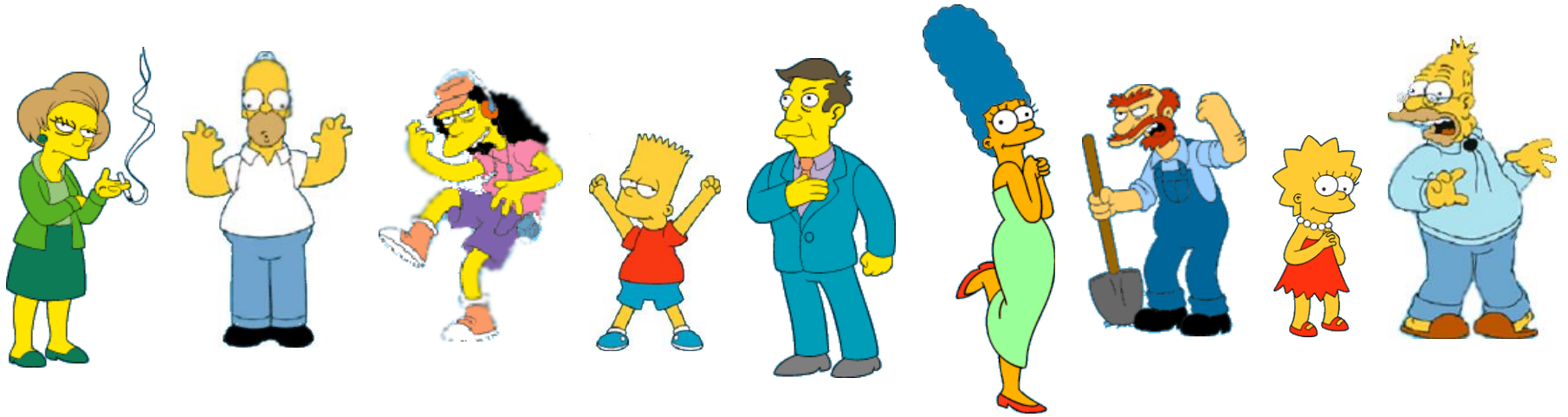
Tim Asprak Metkuan

What is Clustering?

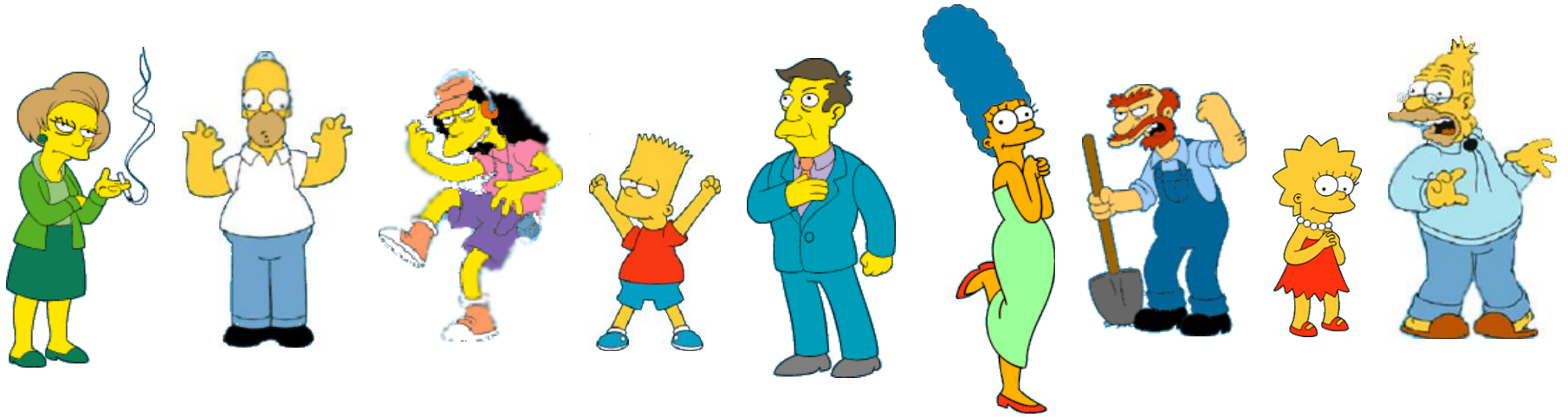
Also called *unsupervised learning*, sometimes called ***classification*** by statisticians and ***sorting*** by psychologists and ***segmentation*** by people in marketing

Mengelompokkan data-data menjadi beberapa cluster berdasarkan kesamaannya

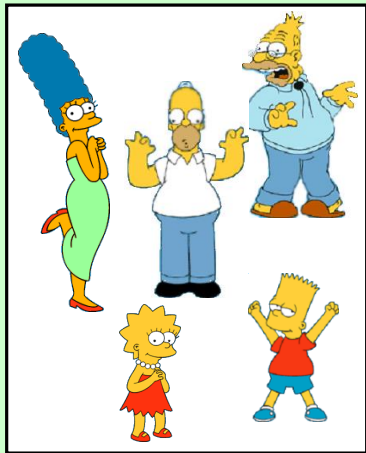
What is a natural grouping among these objects?



What is a natural grouping among these objects?



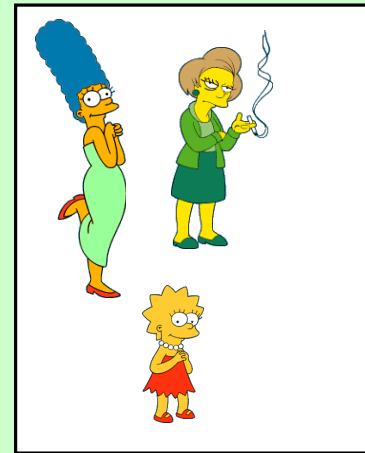
Clustering is subjective



Simpson's Family



School Employees



Females

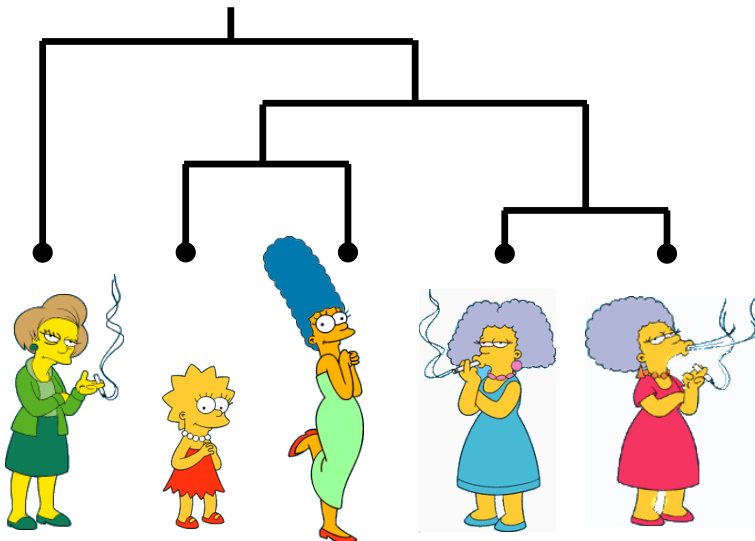


Males

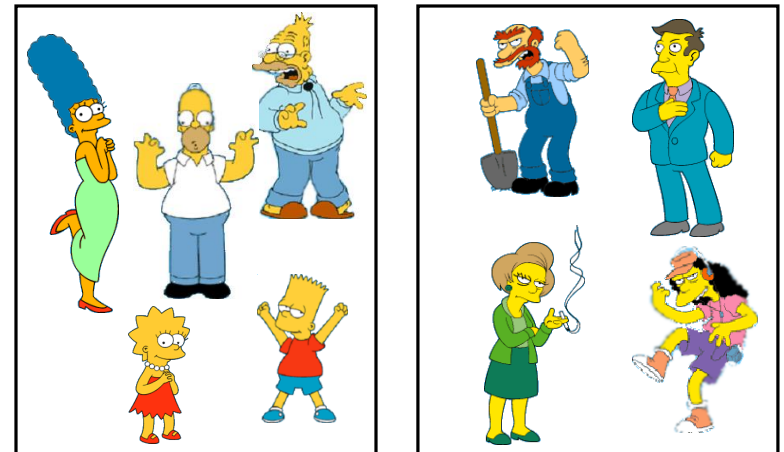
Two Types of Clustering

- **Partitional algorithms:** Membuat beberapa partisi dan mengelompokkan objek berdasarkan kriteria tertentu
- **Hierarchical algorithms:** Membuat dekomposisi pengelompokan objek berdasarkan kriteria tertentu. Misal= tua-muda, tua-muda(merokok-tidak merokok)

Hierarchical



Partitional



What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.


Webster's Dictionary




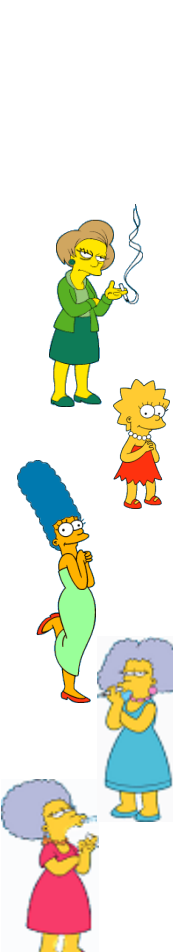
Similarity is hard to define, but...
"We know it when we see it".











Distance :

Adalah ukuran kesamaan antar objek yang dihitung berdasarkan rumusan tertentu


$$D(\text{Marge}, \text{Lisa}) = 8$$

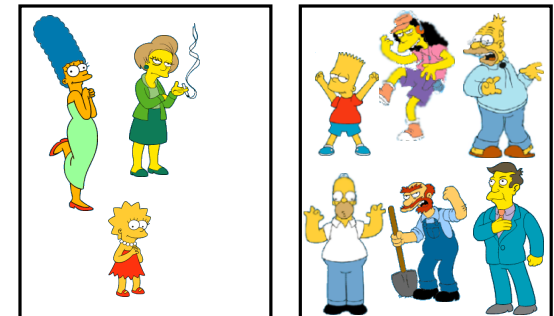
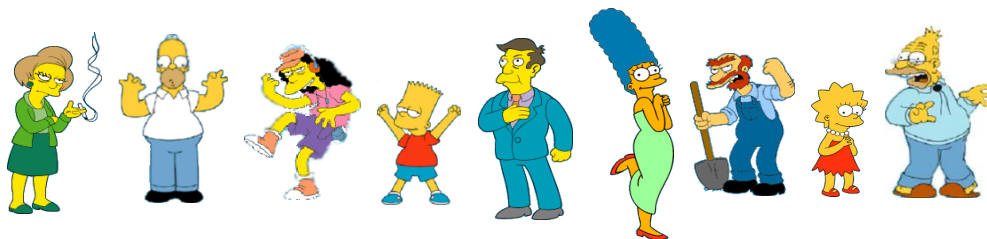

$$D(\text{Barbara}, \text{Edna}) = 1$$



				
0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0
				

Partitional Clustering

- Nonhierarchical, setiap objek ditempatkan di salah satu cluster
- Nonoverlapping cluster
- Jumlah kluster yang akan dibentuk ditentukan sejak awal



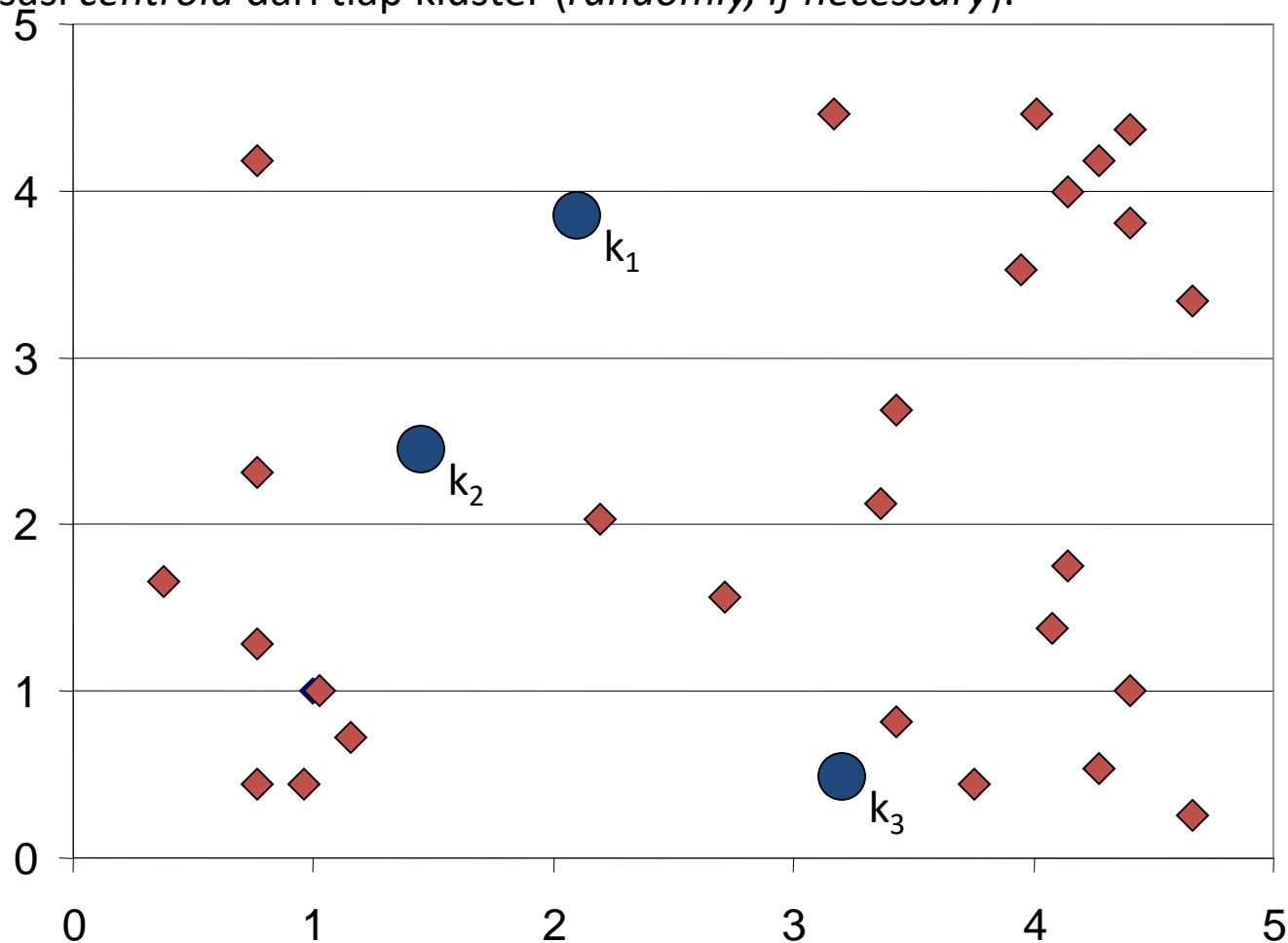
Algorithm *k-means*

1. Tentukan jumlah kluster k .
2. Inisialisasi *centroid* dari tiap kluster (*randomly, if necessary*).
3. Hitung jarak objek ke *centroid* tiap kluster. Anggota kluster adalah objek dengan jarak minimum
4. Setelah kluster dan anggotanya terbentuk, hitung *mean* tiap cluster dan jadikan sebagai *centroid* baru
5. Jika *centroid* baru tidak sama dengan *centroid* lama, maka kembali ke langkah 3. Tetapi jika *centroid* tidak berubah, maka iterasi selesai.

K-means Clustering: Step 1-2

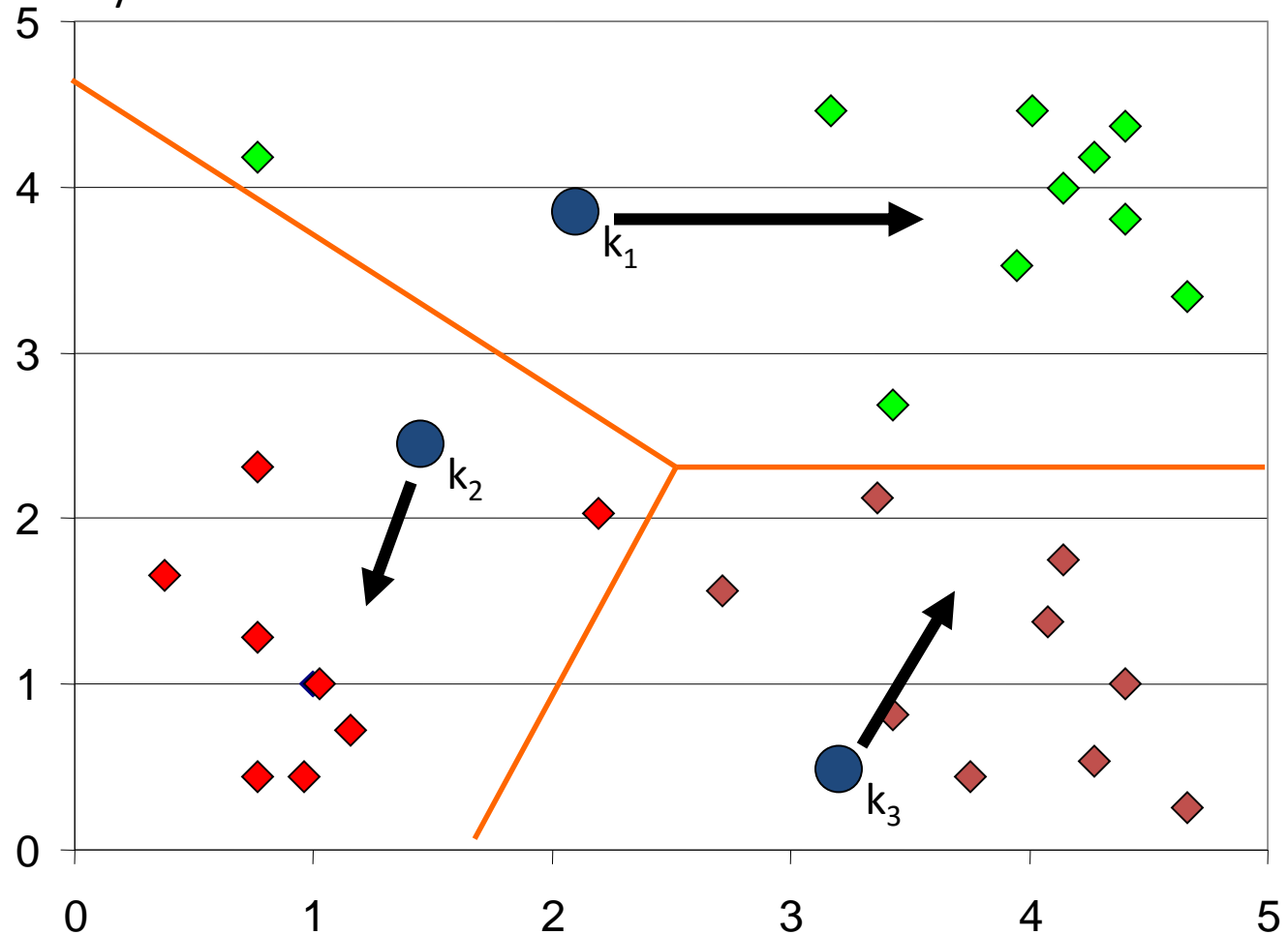
Tentukan jumlah k .

Inisialisasi *centroid* dari tiap kluster (*randomly, if necessary*).



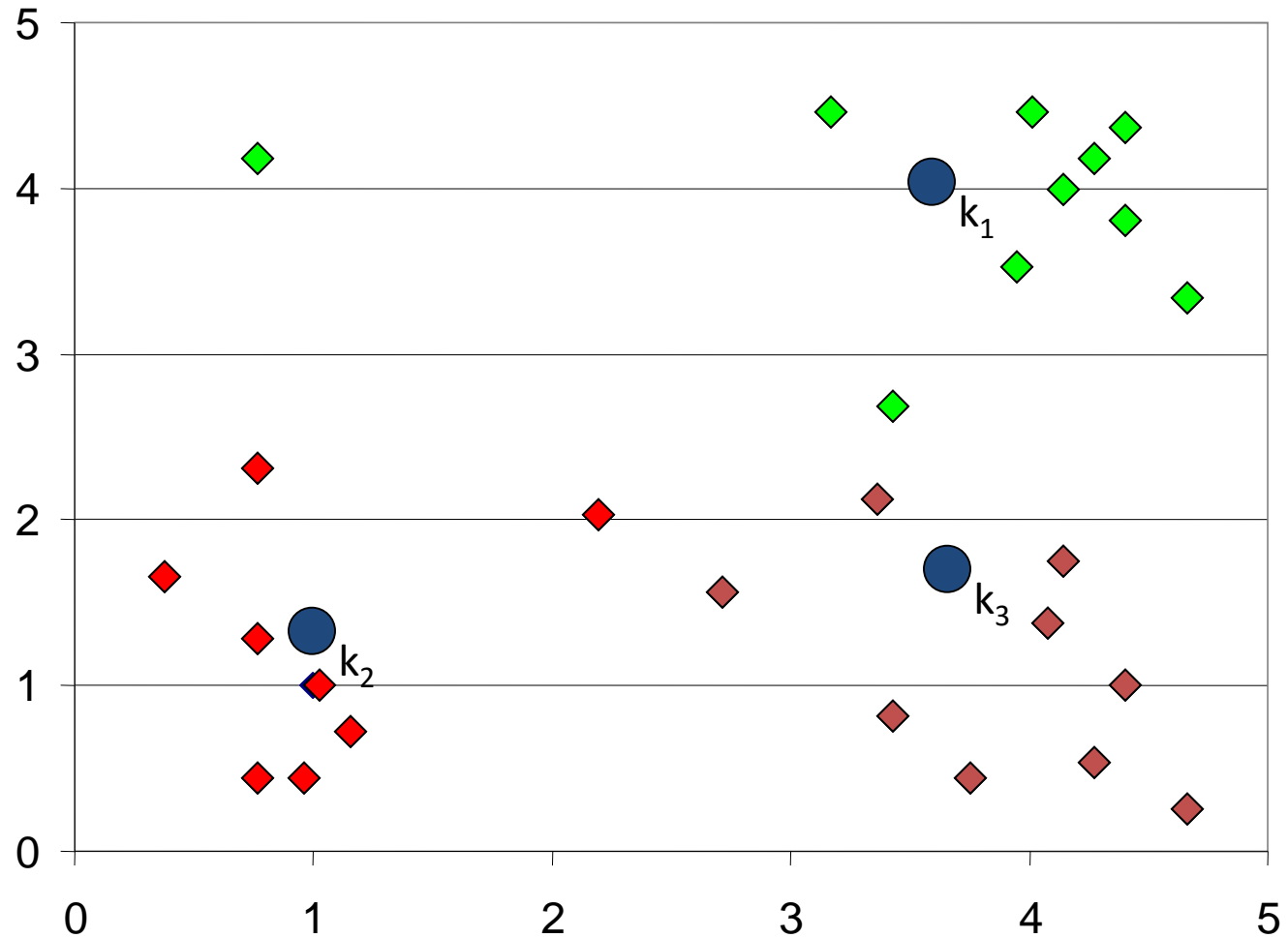
K-means Clustering: Step 3

Tentukan keanggotaan objek-objek yang lain dengan mengklasifikasikannya sesuai *centroid* terdekat



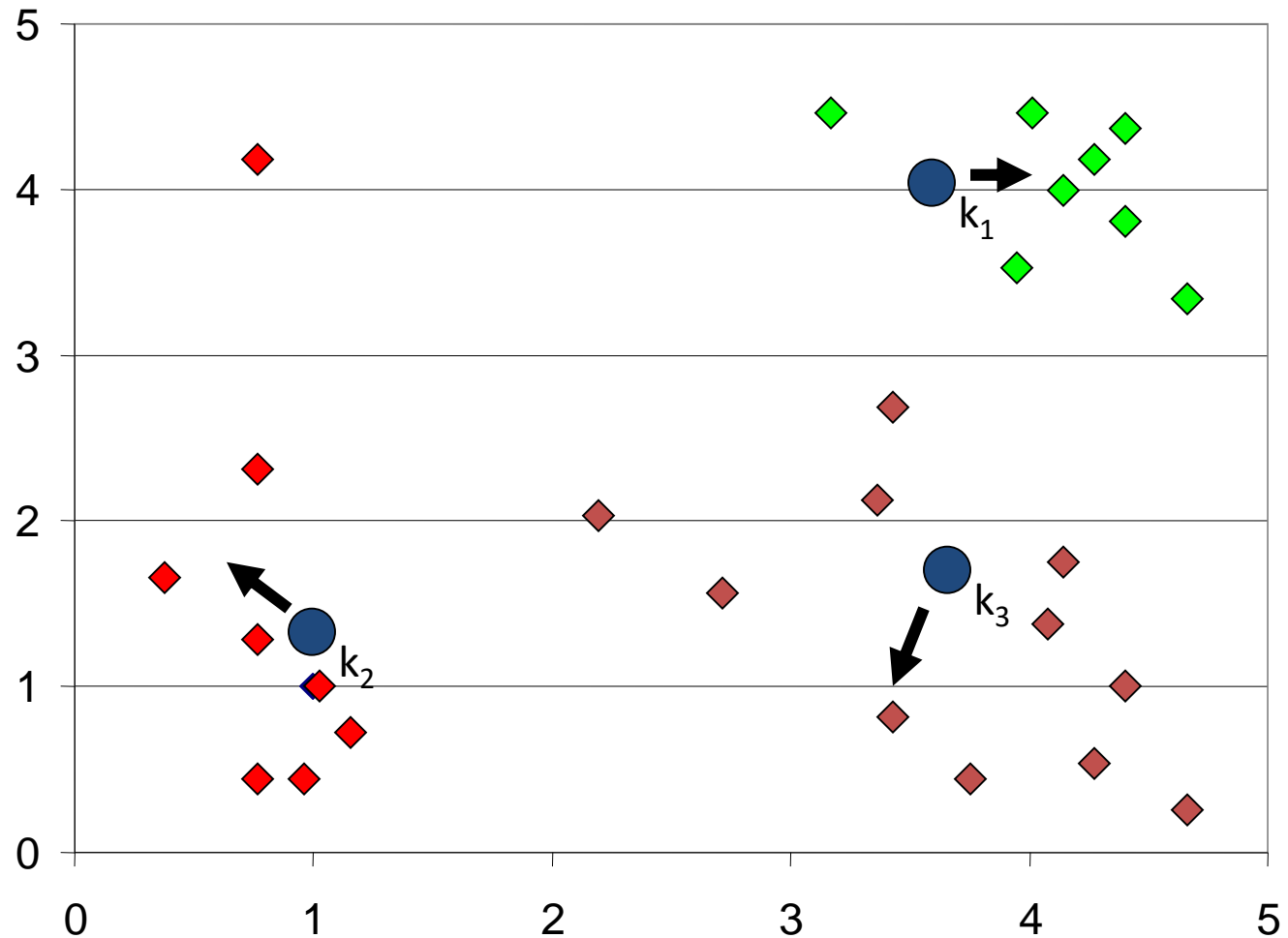
K-means Clustering: Step 4

Setelah kluster dan anggotanya terbentuk, hitung *mean* tiap kluster dan jadikan sebagai *centroid* baru



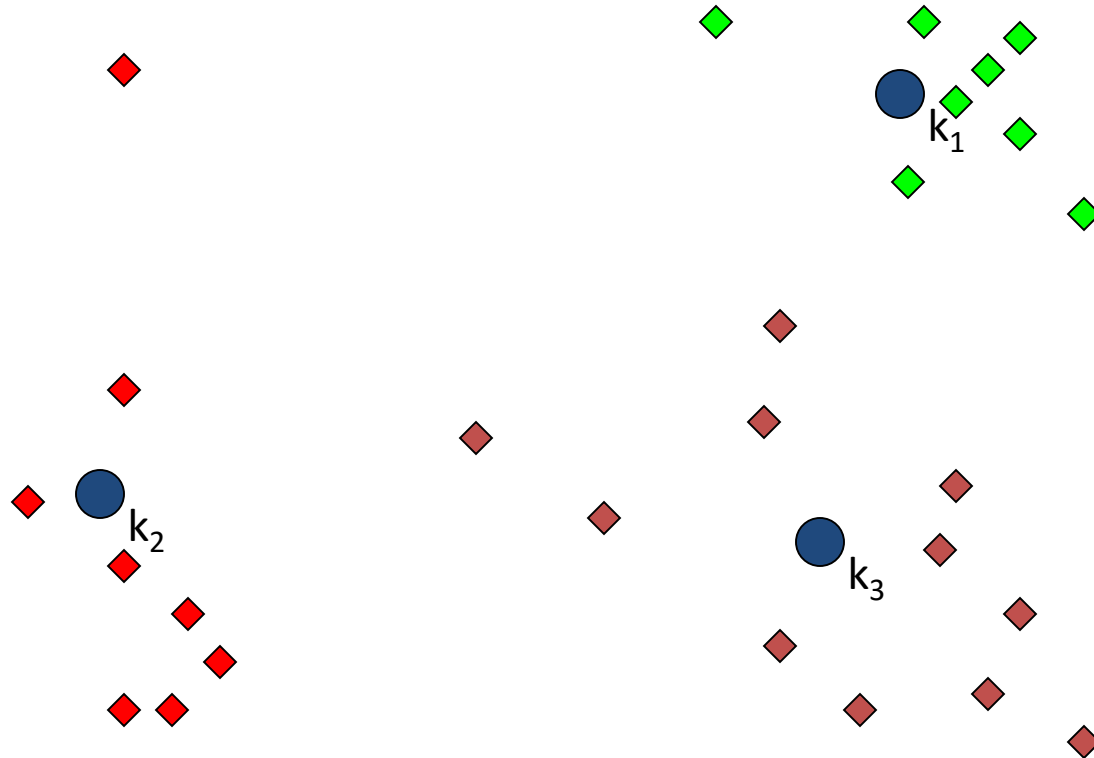
K-means Clustering: Step 5

Jika *centroid* baru tidak sama dengan *centroid* lama, maka perlu di-*update* lagi keanggotaan objek-objeknya



K-means Clustering: Finish

Lakukan iterasi step 3-5 sampai tak ada lagi perubahan *centroid* dan tak ada lagi objek yang berpindah kelas



Comments on the *K-Means* Method

- Strength
 - Selalu konvergen atau mampu melakukan *clustering*
 - Operasi matematika sederhana
 - Beban komputasi relatif lebih ringan sehingga *clustering* bisa dilakukan dengan cepat walaupun relatif tergantung pada banyak jumlah data dan jumlah kluster yg ingin dicapai
- Weakness
 - Jumlah kluster (k) harus ditentukan
 - Nilai *centroid* awal dapat mempengaruhi hasil kluster
 - Tergantung pada *mean*
 - Tidak mampu mengatasi *noisy data* and *outliers*
 - Kluster yang dihasilkan bersifat optimum lokal

Algoritma pengukuran distance

- SqEuclidean
- Cityblock
- Cosine
- Correlation
- Hamming

MATLAB

- `[IDX,C] = kmeans(X,k)` returns the k cluster centroid locations in the k -by- p matrix C

- [...] = kmeans(...,'param1',val1,'param2',val2,...) enables you to specify optional parameter name-value pairs to control the iterative algorithm used by kmeans.
- The parameters are :
 - 'distance'
 - 'start'
 - 'replicates'
 - 'maxiter'
 - 'emptyaction'
 - 'display'

'distance'

- Distance measure, in p-dimensional space, that kmeans minimizes with respect to. kmeans computes centroid clusters differently for the different supported distance measures:

'sqEuclidean'	Squared Euclidean distance (default). Each centroid is the mean of the points in that cluster.
'cityblock'	Sum of absolute differences, i.e., L1. Each centroid is the component-wise median of the points in that cluster.
'cosine'	One minus the cosine of the included angle between points (treated as vectors). Each centroid is the mean of the points in that cluster, after normalizing those points to unit Euclidean length.
'correlation'	One minus the sample correlation between points (treated as sequences of values). Each centroid is the component-wise mean of the points in that cluster, after centering and normalizing those points to zero mean and unit standard deviation.
'Hamming'	Percentage of bits that differ (only suitable for binary data). Each centroid is the component-wise median of points in that cluster.

'start'

- Method used to choose the initial cluster centroid positions, sometimes known as "seeds". Valid starting values are:

'sample'	Select k observations from X at random (default).
'uniform'	Select k points uniformly at random from the range of X . Not valid with Hamming distance.
'cluster'	Perform a preliminary clustering phase on a random 10% subsample of X . This preliminary phase is itself initialized using 'sample'.
Matrix	k -by- p matrix of centroid starting locations. In this case, you can pass in <code>[]</code> for k , and <code>kmeans</code> infers k from the first dimension of the matrix. You can also supply a 3-dimensional array, implying a value for the 'replicates' parameter from the array's third dimension.

'replicates'

- Number of times to repeat the clustering, each with a new set of initial cluster centroid positions.
- kmeans returns the solution with the lowest value for sumd.
- You can supply 'replicates' implicitly by supplying a 3-dimensional array as the value for the 'start' parameter.

'maxiter'

- Maximum number of iterations. Default is 100.

'emptyaction'

- Action to take if a cluster loses all its member observations. Can be one of:

'error'	Treat an empty cluster as an error. (default)
'drop'	Remove any clusters that become empty. kmeans sets the corresponding return values in C and D to NaN.
'singleton'	Create a new cluster consisting of the one point furthest from its centroid.

'display'

- Controls display of output.
- **'off'** : Display no output.
- **'iter'** : Display information about each iteration during minimization, including the iteration number, the optimization phase, the number of points moved, and the total sum of distances.
- **'final'** : Display a summary of each replication.
- **'notify'** : Display only warning and error messages. (default)

Example

Dengan menggunakan data iris (gunakan nilai fitur saja), buat kluster data dengan $k = 3$

```
>> iris = [5,1  3,5  1,4  0,2;  
            4,9  3    1,4  0,2;  
            4,7  3,2  1,3  0,2;  
            ...];  
>> [idx,ctrs] = kmeans(iris,3);
```