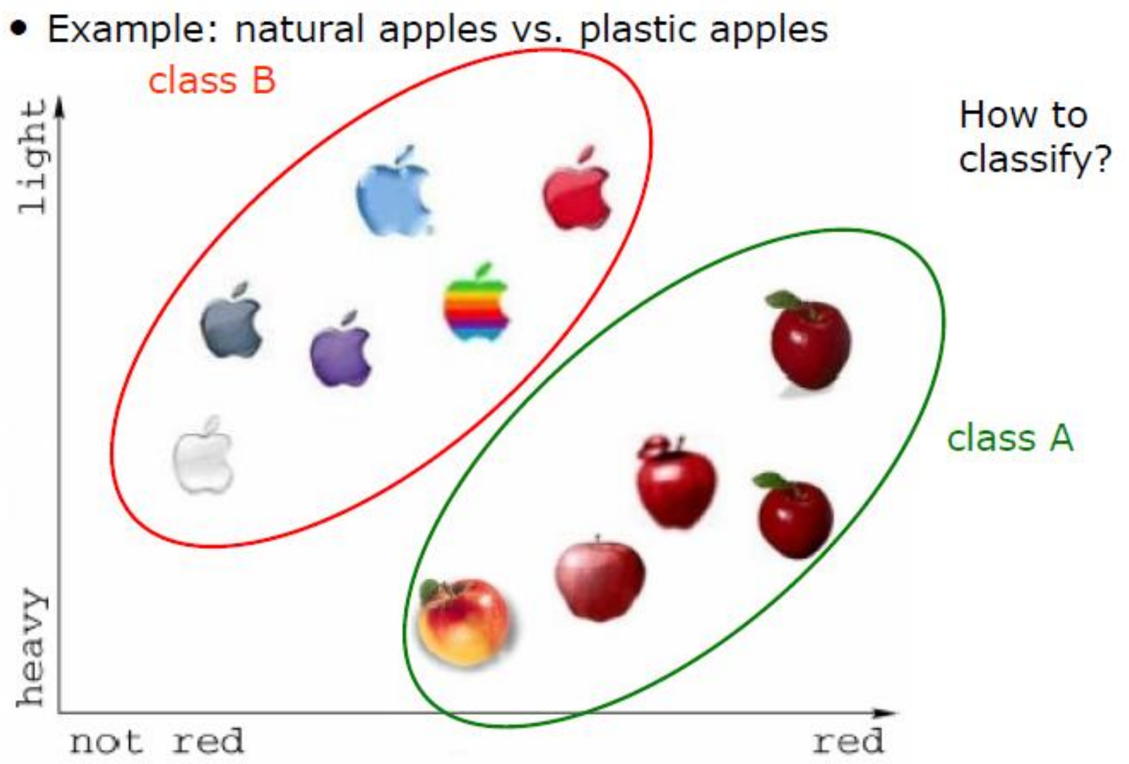
# CLUSTERING

**Clustering** is the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often according to some defined distance measure.

**Clustering** is unsupervised classification

# CLUSTERING

- There is no explicit teacher and the system forms clusters or “natural groupings” or structure in the input pattern



- Example: natural apples vs. plastic apples

How to classify?

# CLUSTERING

- Data WITHOUT classes or labels

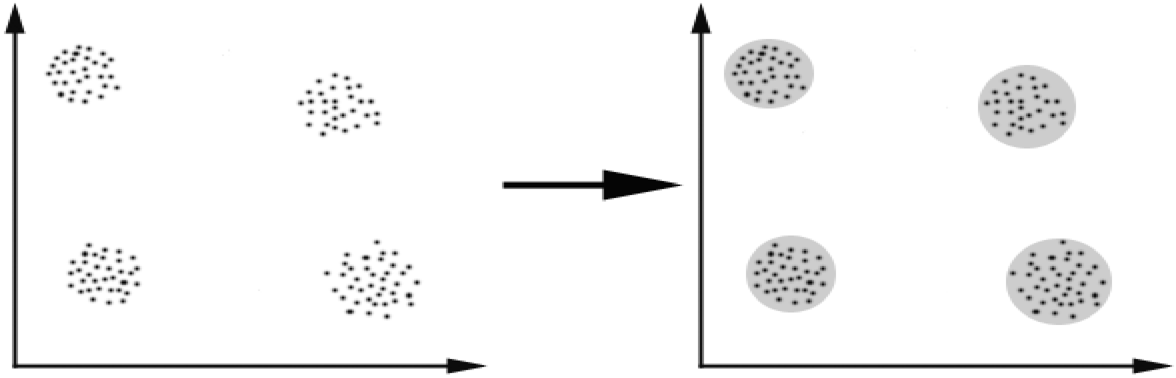
$$\left\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \cdots\cdots \mathbf{x}_n\right\}, \quad \mathbf{x} \in \mathbb{R}^d$$

- Deals with finding a *structure* in a collection of unlabeled data.

- The process of organizing objects into groups whose members are *similar* in some way

- A *cluster* is therefore a collection of objects which are "*similar*" between them and are "*dissimilar*" to the objects belonging to other clusters.

# CLUSTERING



- In this case we easily identify the 4 clusters into which the data can be divided;
- The similarity criterion is *distance*: two or more objects belong to the same cluster if they are "*close*" according to a given distance

# Distance Measures

- Each clustering problem is based on some kind of "distance" between points.
- Two major classes of distance measure:
  1. Euclidean
  2. Non-Euclidean

# Euclidean Vs. Non-Euclidean

- A Euclidean space has some number of real-valued dimensions and “dense” points.
  - There is a notion of “average” of two points.
  - A Euclidean distance is based on the locations of points in such a space.
- A Non-Euclidean distance is based on properties of points, but not their “location” in a space.

# Euclidean Distance

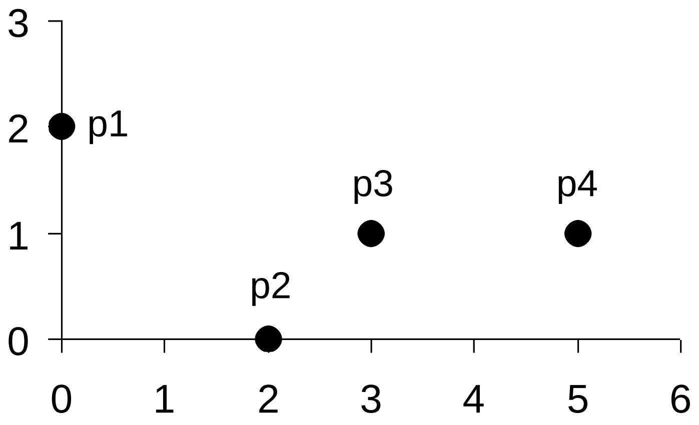
- Euclidean Distance

$$\boldsymbol{dist} = \sqrt{\sum_{\boldsymbol{k}=1}^{\boldsymbol{n}} (\boldsymbol{p_k} - \boldsymbol{q_k})^2}$$

Where $n$ is the number of dimensions (attributes) and $p_k$ and $q_k$ are, respectively, the k$^{th}$ attributes (components) or data objects $p$ and $q$.

- Standardization is necessary, if scales differ.

# Euclidean Distance



| point | x | y |
|---|---|---|
| **p1** | 0 | 2 |
| **p2** | 2 | 0 |
| **p3** | 3 | 1 |
| **p4** | 5 | 1 |

| | p1 | p2 | p3 | p4 |
|---|---|---|---|---|
| **p1** | 0 | 2.828 | 3.162 | 5.099 |
| **p2** | 2.828 | 0 | 1.414 | 3.162 |
| **p3** | 3.162 | 1.414 | 0 | 2 |
| **p4** | 5.099 | 3.162 | 2 | 0 |

**Distance Matrix**

# Types of Clustering

- **Hierarchical algorithms**
  - These find successive clusters using previously established clusters.

1. Agglomerative ("bottom-up"): Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters.

2. Divisive ("top-down"): Divisive algorithms begin with the whole set and proceed to divide it into successively into smaller clusters.

# Types of Clustering

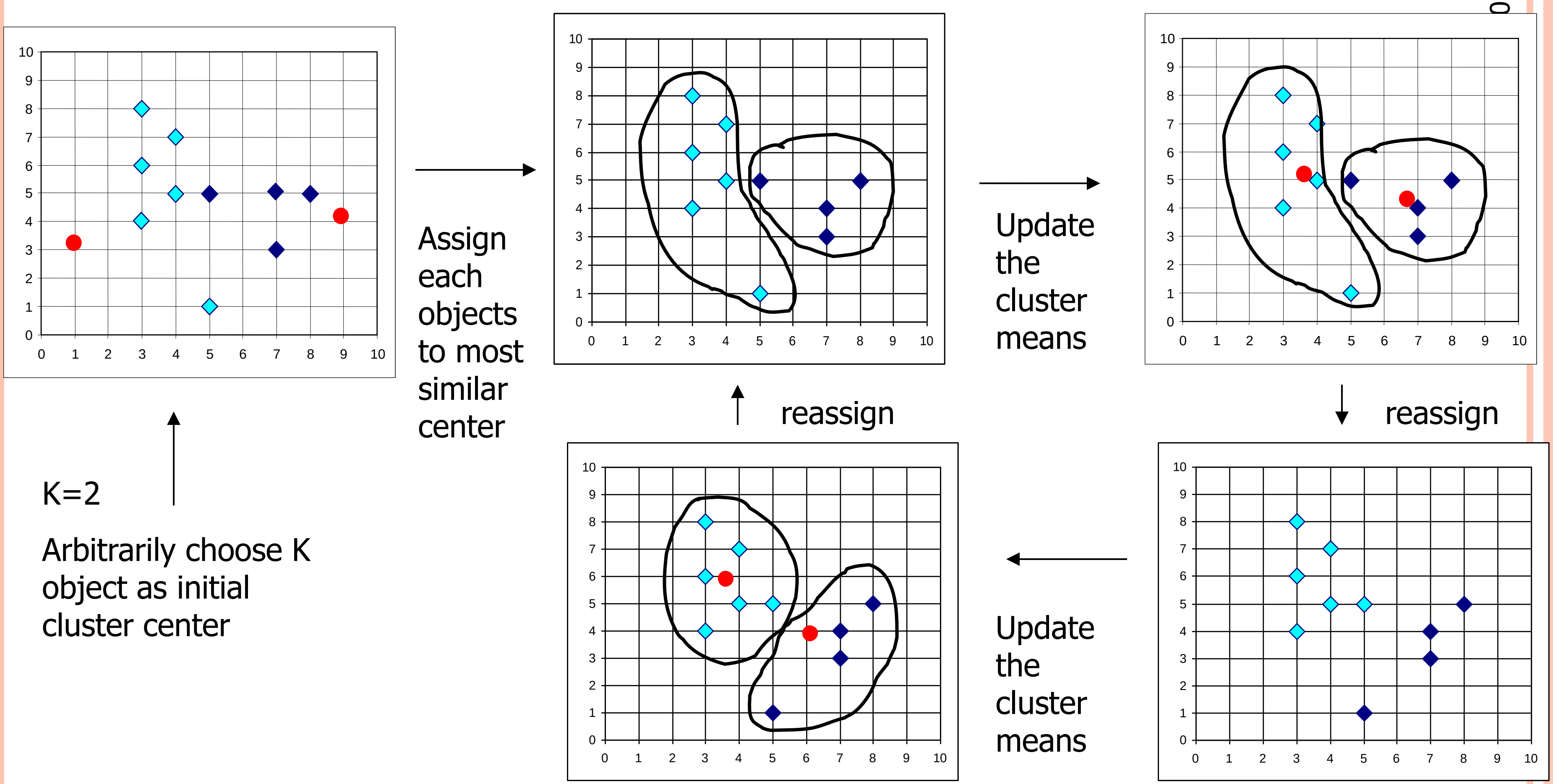
- **Partitional clustering**
  - Construct a partition of a data set to produce several clusters – At once
  - The process is repeated iteratively – Termination condition
  - Examples
    - *K-means clustering*
    - *Fuzzy c-means clustering*

# K MEANS CLUSTERING

1. Chose the number (K) of clusters and randomly select the centroids of each cluster.
2. For each data point:
   I. Calculate the distance from the data point to each cluster.
   II. Assign the data point to the closest cluster.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until there is no further change in the assignment of data points (or in the centroids).

# THE *K-MEANS* CLUSTERING METHOD

- Example



K=2

Arbitrarily choose K object as initial cluster center

Assign each objects to most similar center

Update the cluster means

reassign

reassign

Update the cluster means

# EXAMPLE

- Cluster the following eight points (with (x, y) representing locations) into three clusters
  - A1(2, 10) A2(2, 5) A3(8, 4) A4(5, 8) A5(7, 5) A6(6, 4) A7(1, 2) A8(4, 9)
- Initial cluster centers are:

| | |
|---|---|
| C1 | A1(2, 10), |
| C2 | A4(5, 8) and |
| C3 | A7(1, 2) |

- The distance function between two points $a=(x1, y1)$ and $b=(x2, y2)$ is defined as:
  - $\rho(a, b) = |x2 - x1| + |y2 - y1|$
- Use k-means algorithm to find the three cluster centers after the second iteration

# EXAMPLE

| | | (2, 10) | (5, 8) | (1, 2) | |
|---|---|---|---|---|---|
| | Point | Dist Mean 1 | Dist Mean 2 | Dist Mean 3 | Cluster |
| A1 | (2, 10) | | | | |
| A2 | (2, 5) | | | | |
| A3 | (8, 4) | | | | |
| A4 | (5, 8) | | | | |
| A5 | (7, 5) | | | | |
| A6 | (6, 4) | | | | |
| A7 | (1, 2) | | | | |
| A8 | (4, 9) | | | | |

# EXAMPLE

- Calculate the distance from the first point (2, 10) to each of the three means, by using the distance function

| point | mean1 |
|---|---|
| *x1, y1* | *x2, y2* |
| (2, 10) | (2, 10) |

$$\rho(point, mean1) = |x2 - x1| + |y2 - y1|$$
$$= |2 - 2| + |10 - 10|$$
$$= 0$$

| point | mean2 |
|---|---|
| *x1, y1* | *x2, y2* |
| (2, 10) | (5, 8) |

$$\rho(point, mean2) = |x2 - x1| + |y2 - y1|$$
$$= |5 - 2| + |8 - 10|$$
$$= 5$$

| point | mean3 |
|---|---|
| *x1, y1* | *x2, y2* |
| (2, 10) | (1, 2) |

$$\rho(point, mean2) = |x2 - x1| + |y2 - y1|$$
$$= |1 - 2| + |2 - 10|$$
$$= 9$$

# Example

| | | (2, 10) | (5, 8) | (1, 2) | |
|---|---|---|---|---|---|
| | Point | Dist Mean 1 | Dist Mean 2 | Dist Mean 3 | Cluster |
| A1 | (2, 10) | 0 | 5 | 9 | 1 |
| A2 | (2, 5) | | | | |
| A3 | (8, 4) | | | | |
| A4 | (5, 8) | | | | |
| A5 | (7, 5) | | | | |
| A6 | (6, 4) | | | | |
| A7 | (1, 2) | | | | |
| A8 | (4, 9) | | | | |

| Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|
| (2, 10) | | |

# Example

- Calculate the distance of second point (2, 5) to each of the three means, by using the distance function:

| point | mean1 |
|---|---|
| *x1*, *y1* | *x2*, *y2* |
| (2, 5) | (2, 10) |

$$\begin{aligned} \rho(point, mean1) &= |x2 - x1| + |y2 - y1| \\ &= |2 - 2| + |10 - 5| \\ &= 5 \end{aligned}$$

| point | mean2 |
|---|---|
| *x1*, *y1* | *x2*, *y2* |
| (2, 5) | (5, 8) |

$$\begin{aligned} \rho(point, mean2) &= |x2 - x1| + |y2 - y1| \\ &= |5 - 2| + |8 - 5| \\ &= 6 \end{aligned}$$

| point | mean3 |
|---|---|
| *x1*, *y1* | *x2*, *y2* |
| (2, 5) | (1, 2) |

$$\begin{aligned} \rho(point, mean2) &= |x2 - x1| + |y2 - y1| \\ &= |1 - 2| + |2 - 5| \\ &= 4 \end{aligned}$$

# Example

| | | (2, 10) | (5, 8) | (1, 2) | |
|---|---|---|---|---|---|
| | Point | Dist Mean 1 | Dist Mean 2 | Dist Mean 3 | Cluster |
| A1 | (2, 10) | 0 | 5 | 9 | 1 |
| A2 | (2, 5) | 5 | 6 | 4 | 3 |
| A3 | (8, 4) | | | | |
| A4 | (5, 8) | | | | |
| A5 | (7, 5) | | | | |
| A6 | (6, 4) | | | | |
| A7 | (1, 2) | | | | |
| A8 | (4, 9) | | | | |

| Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|
| (2, 10) | | (2, 5) |

# EXAMPLE

- Iteration I
  - Analogically, we fill in the rest of the table, and place

| | | (2, 10) | (5, 8) | (1, 2) | |
|---|---|---|---|---|---|
| | Point | Dist Mean 1 | Dist Mean 2 | Dist Mean 3 | **Cluster** |
| **A1** | (2, 10) | 0 | 5 | 9 | **1** |
| **A2** | (2, 5) | 5 | 6 | 4 | **3** |
| **A3** | (8, 4) | 12 | 7 | 9 | **2** |
| **A4** | (5, 8) | 5 | 0 | 10 | **2** |
| **A5** | (7, 5) | 10 | 5 | 9 | **2** |
| **A6** | (6, 4) | 10 | 5 | 7 | **2** |
| **A7** | (1, 2) | 9 | 10 | 0 | **3** |
| **A8** | (4, 9) | 3 | 2 | 10 | **2** |

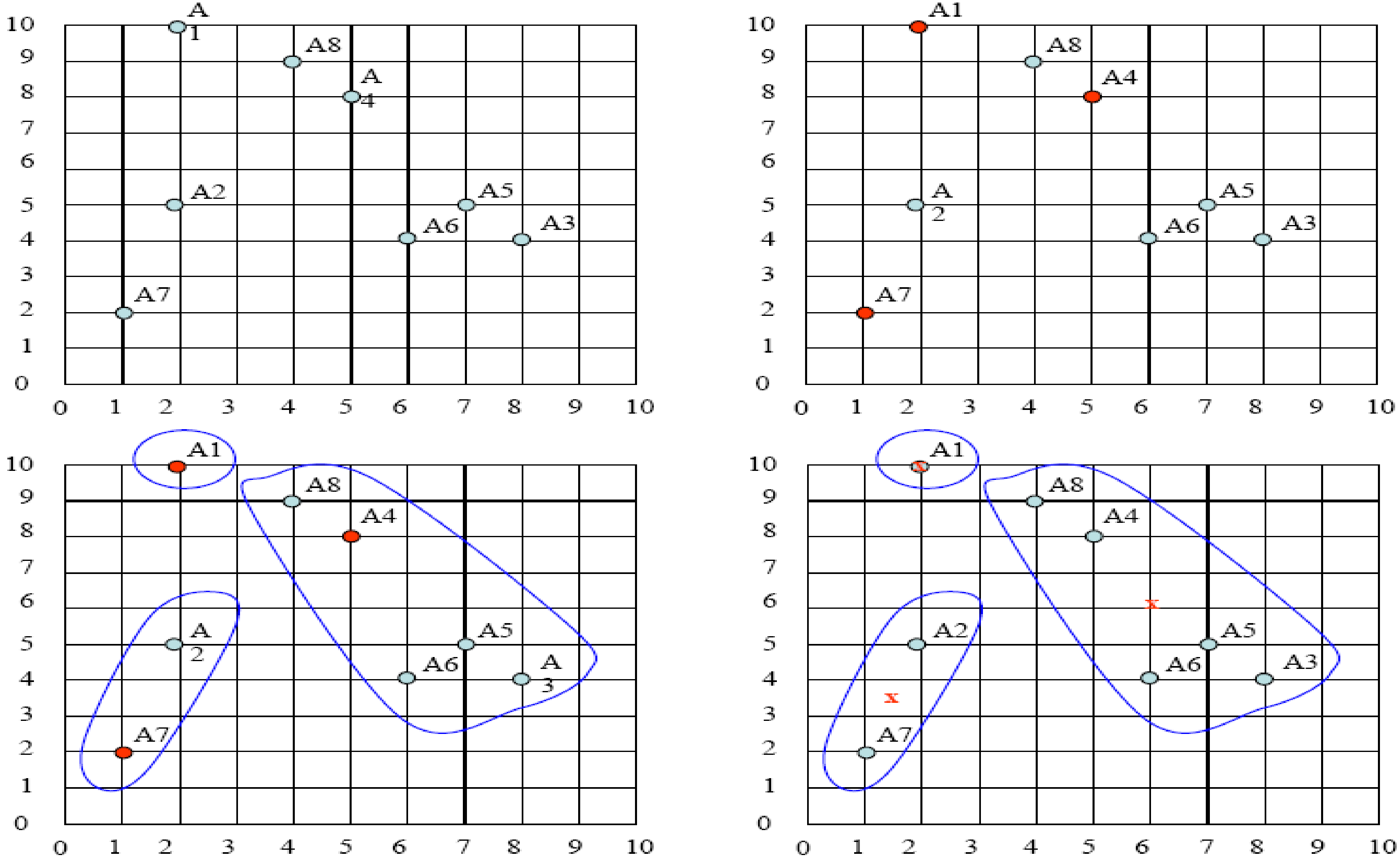
# EXAMPLE

- Clusters after Iteration I

| Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|
| (2, 10) | (8, 4) | (2, 5) |
| | (5, 8) | (1, 2) |
| | (7, 5) | |
| | (6, 4) | |
| | (4, 9) | |

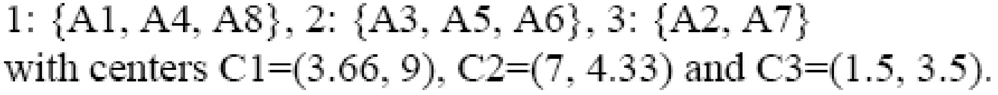
b) centers of the new clusters:
C1= (2, 10), C2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6), C3= ((2+1)/2, (5+2)/2) = (1.5, 3.5)

c)

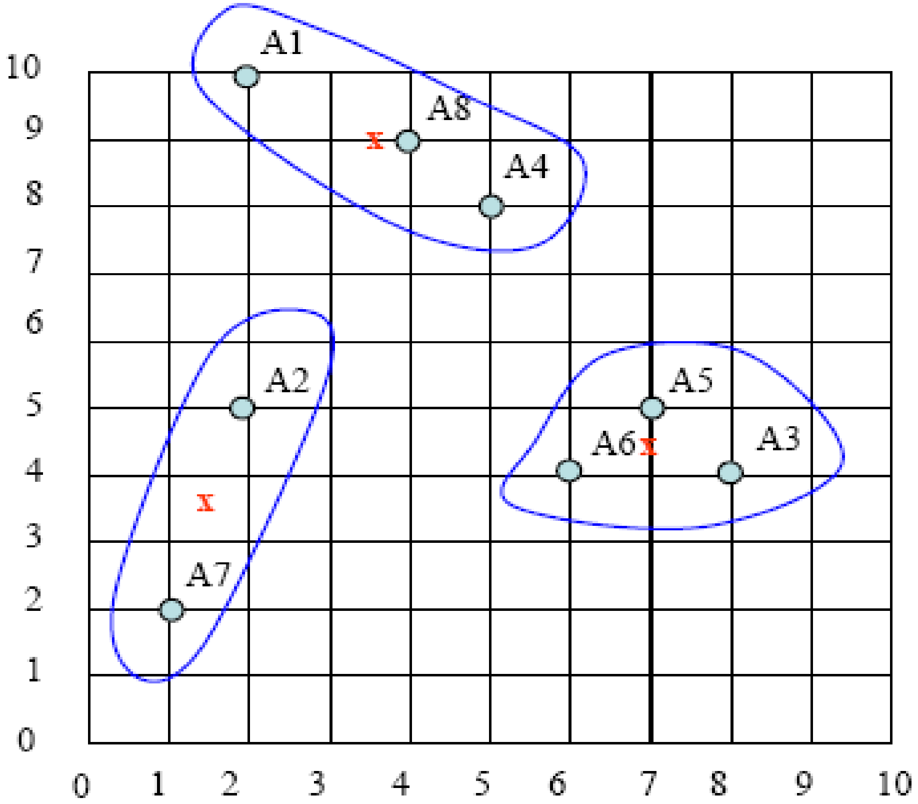


# EXAMPLE

- Calculate 2nd , 3rd iteration and so on till the new means do not change anymore
- Result ample



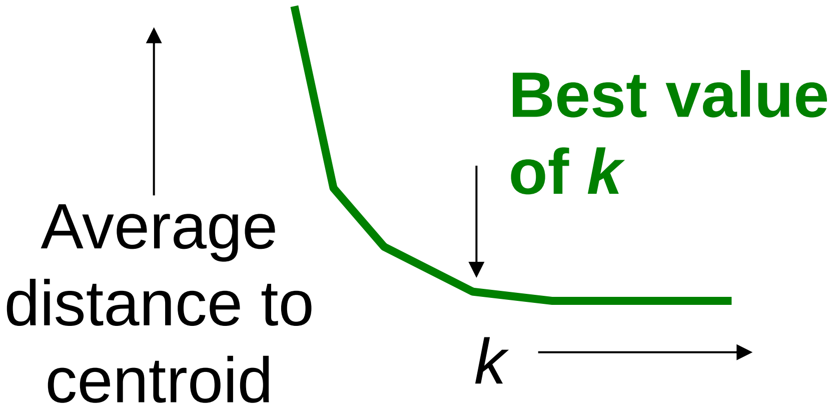
1: {A1, A4, A8}, 2: {A3, A5, A6}, 3: {A2, A7} with centers C1=(3.66, 9), C2=(7, 4.33) and C3=(1.5, 3.5).



# GETTING THE *K* RIGHT

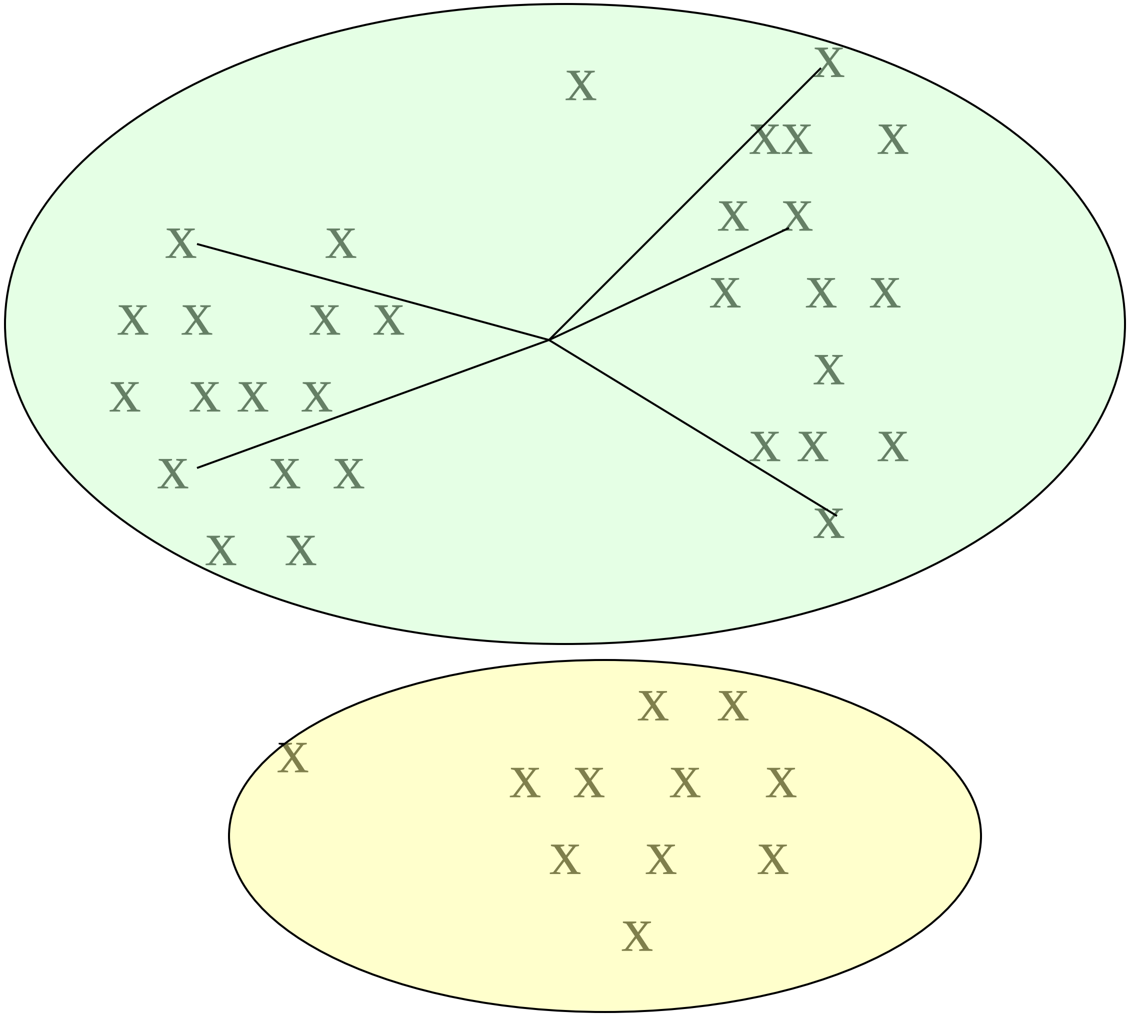
**How to select *k*?**

- Try different **k**, looking at the change in the average distance to centroid as **k** increases
- Average falls rapidly until right **k**, then changes little



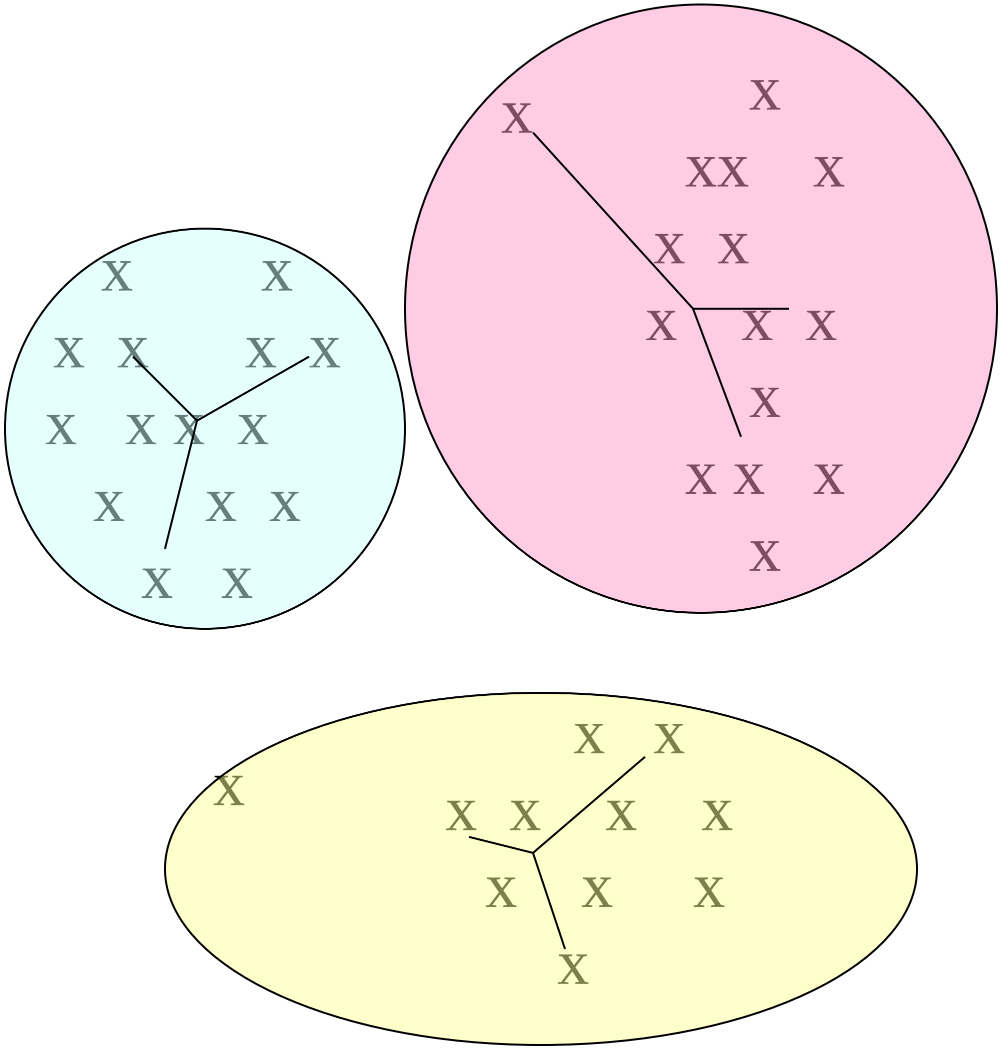
# EXAMPLE: PICKING *K*

Too few;
many long
distances
to centroid.



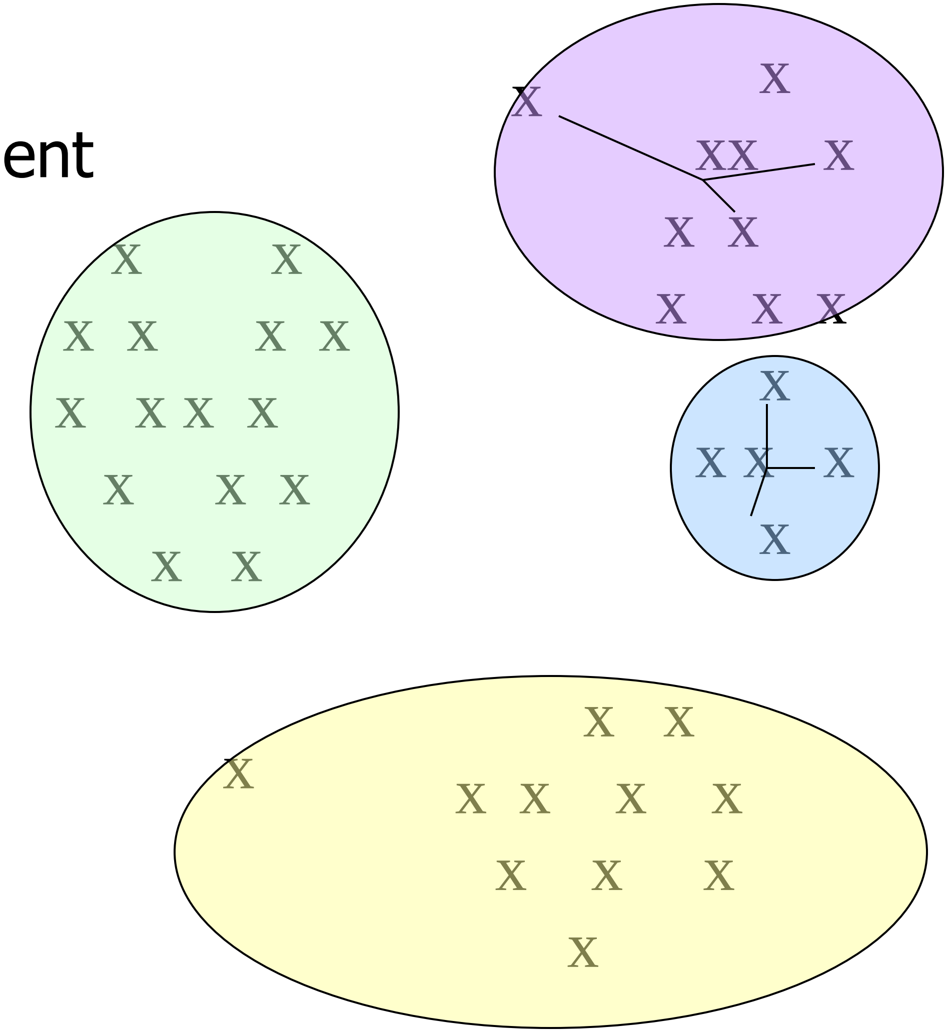
# Example: Picking *k*

Just right; distances rather short.



# EXAMPLE: PICKING *K*

Too many;
little improvement
in average
distance.


x x x x x x x x x x x x x x x x
x x x x x x x x x x x x x x x x x
x x x x x
x x x x x x x x x x x x x

# Comments on K mean Clustering

- Strength: *Relatively efficient*: *O*(*tkn*), where *n* is # objects, *k* is # clusters, and *t* is # iterations. Normally, *k*, *t* << *n*.
- Comment: Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *deterministic annealing* and *genetic algorithms*
- Weakness
  - Applicable only when *mean* is defined, then what about categorical data?
  - Need to specify *k,* the *number* of clusters, in advance
  - Unable to handle noisy data and *outliers*
  - Not suitable to discover clusters with *non-convex shapes*