



Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική  
Σχολή Θετικών Επιστημών  
Πανεπιστήμιο Θεσσαλίας

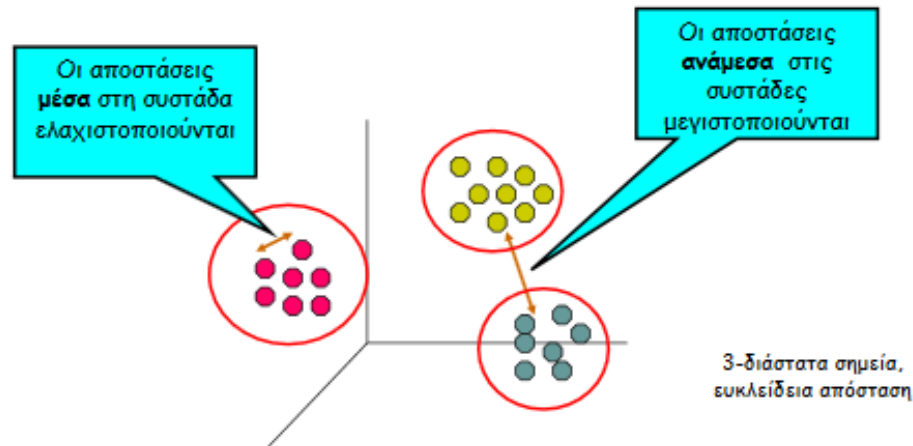
# ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

## Ομαδοποίηση

Αριστείδης Γ. Βραχάτης, Dipl-Ing, M.Sc, PhD  
Adjunct Lecturer

# Συσταδοποίηση - Clustering

- Είναι η διαδικασία της κατηγοριοποίησης των δεδομένων σε σύνολα ομοειδών αντικειμένων καλούμενα ομάδες (clusters)
- Στόχος
  - Να παράγει ένα σύνολο από ομάδες με υψηλή εντός των ομάδων ομοιότητα (intra-cluster similarity), ενώ παράλληλα να διατηρείται χαμηλή η ομοιότητα μεταξύ των διαφόρων ομάδων (inter-cluster similarity)



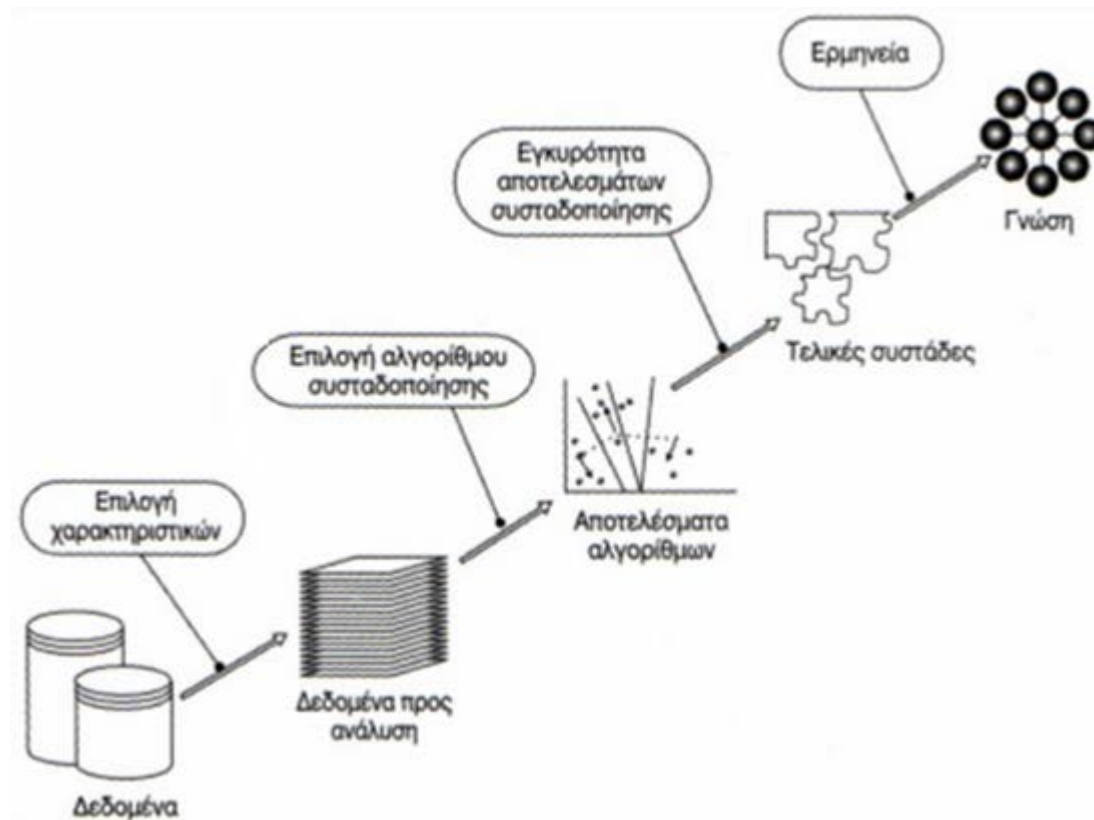
- Εφαρμογές
  - Ευρύ φάσμα εφαρμογών, από τις κοινωνικές επιστήμες, την οικονομία, την αναγνώριση προτύπων έως την βιοπληροφορική, την αστροφυσική και σεισμολογία

# Ομαδοποίηση - Clustering

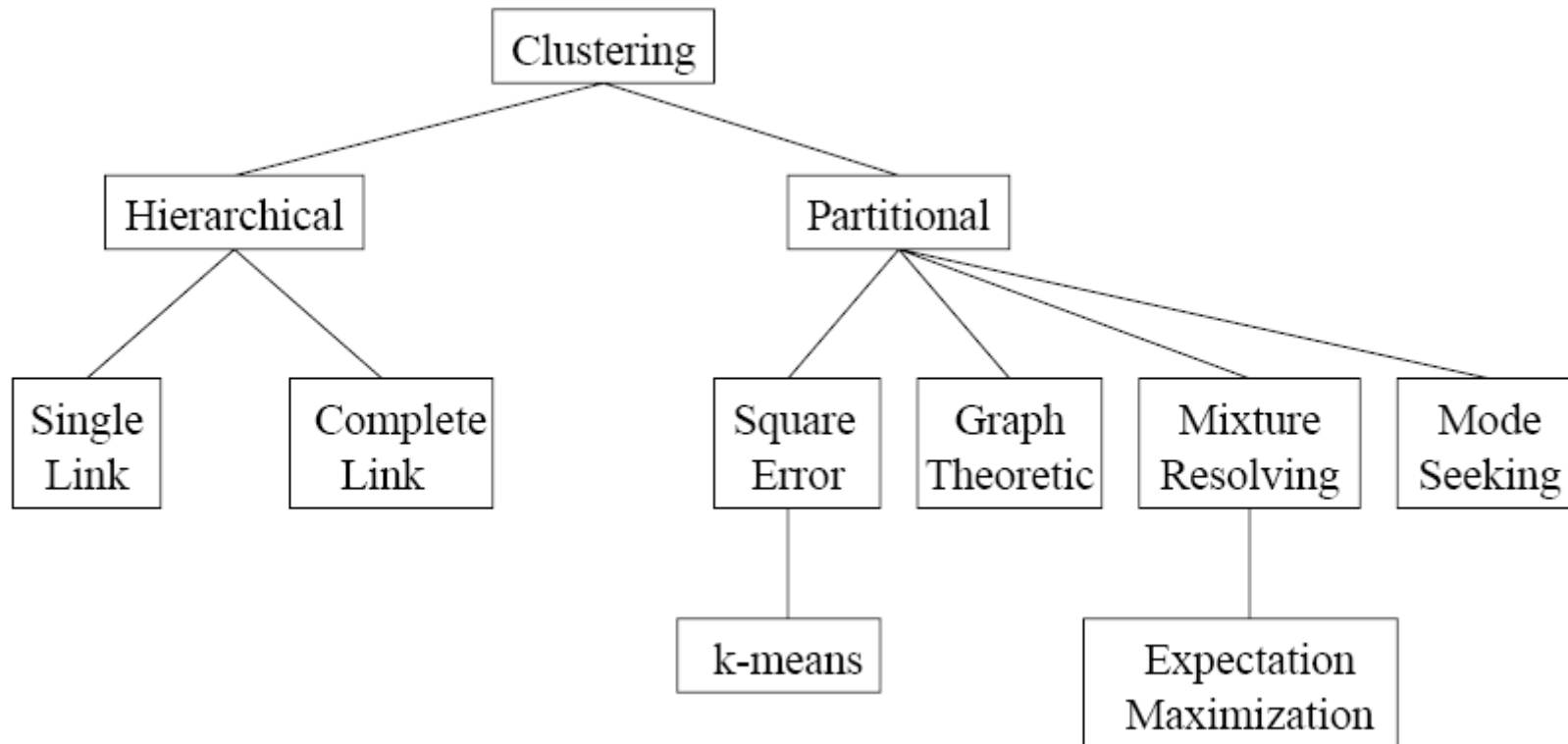
---

- Well Separated
  - μία συστάδα είναι το σύνολο των αντικειμένων όπου κάθε αντικείμενο είναι πιο κοντά σε κάθε άλλο αντικείμενο της συστάδας, από ότι σε κάποιο άλλο αντικείμενο.
- Prototype Based
  - μία συστάδα είναι τα αντικείμενα που είναι πιο κοντά σε ένα πρωτότυπο (prototype) από ότι κάποιο άλλο αντικείμενο. Συνήθως σαν πρωτότυπο επιλέγεται το μέσο των σημείων μίας συστάδας.
- Graph Based
  - μία συνεκτική συνιστώσα ή μία κλίκα του γραφήματος.
- Density Based
  - μία πυκνή περιοχή αντικειμένων που περιβάλλεται από μία αραιή
- Shared Property (conceptual clusters)
  - σύνολο αντικειμένων που μοιράζονται μία ιδιότητα – έχει εφαρμογή κυρίως σε κατηγορικά αντικείμενα

# Βήματα Διαδικασίας Συσταδοποίησης



# Κατηγοριοποίηση των Αλγορίθμων Ομαδοποίησης



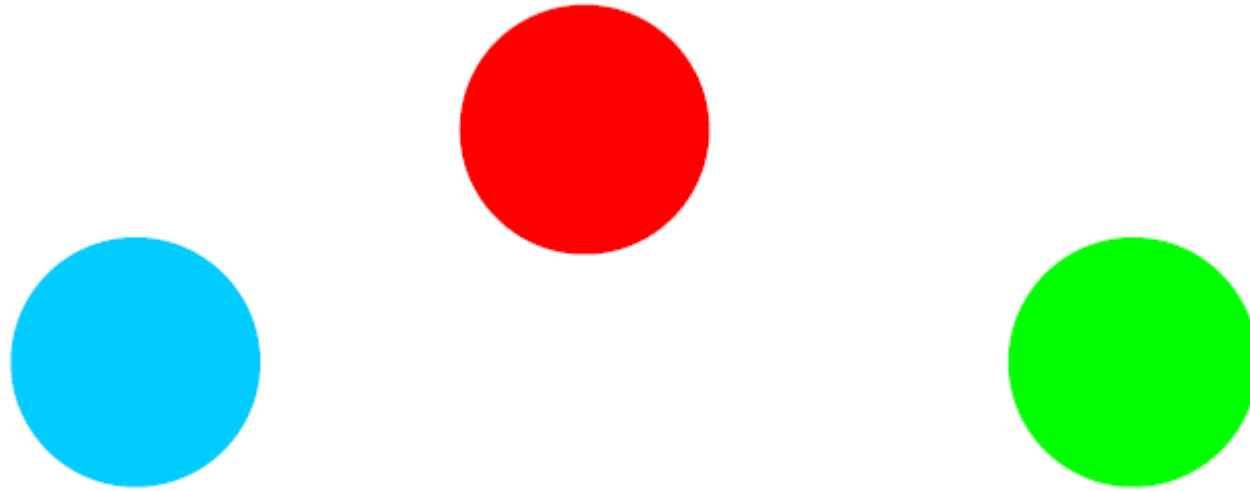
# Είδη Ομαδοποίησης

---

- Βασική διάκριση ανάμεσα στο ιεραρχικό (hierarchical) και διαχωριστικό (partitional) σύνολο από ομάδες
- Διαχωριστική Συσταδοποίηση (Partitional Clustering)
  - Ένας διαμερισμός των αντικειμένων σε μη επικαλυπτόμενα -non-overlapping - υποσύνολα (συστάδες) τέτοιος ώστε κάθε αντικείμενο ανήκει σε ακριβώς ένα υποσύνολο
- Ιεραρχική Συσταδοποίηση (Hierarchical clustering)
  - Ένα σύνολο από εμφωλευμένες (nested) ομάδες Επιτρέπουμε σε μια συστάδα να έχει υπο-συστάδες οργανωμένες σε ένα ιεραρχικό δέντρο

# Τύποι συστάδων: Καλώς Διαχωρισμένες Συστάδες

Μια συστάδα είναι ένα σύνολο από σημεία τέτοια ώστε κάθε σημείο μιας συστάδας είναι **κοντινότερο σε (ή πιο όμοιο με) όλα τα άλλα σημεία** της συστάδας από ότι σε οποιοδήποτε άλλο σημείο που δεν ανήκει στη συστάδα.



**3 καλώς-διαχωρισμένες συστάδες**

Συχνά υπάρχει η έννοια του κατωφλίου (threshold)

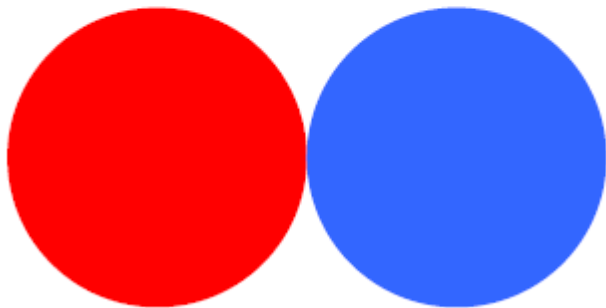
Όχι απαραίτητα κυκλικοί (οποιοδήποτε σχήμα)

# Τύποι συστάδων: Συστάδες βασισμένες σε κέντρο ή πρότυπο

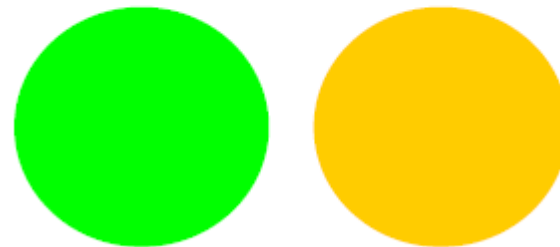
Μια συστάδα είναι ένα σύνολο από αντικείμενα τέτοιο ώστε ένα αντικείμενο στην συστάδα είναι **κοντινότερο σε (ή πιο όμοιο με) το «κέντρο»** ή **πρότυπο** της συστάδας από ό,τι από το κέντρο οποιασδήποτε άλλης συστάδας.

Το κέντρο της ομάδας είναι συχνά

- **centroid**, ο μέσος όρος των σημείων της συστάδας, ή
- a **medoid**, το πιο «αντιπροσωπευτικό» σημείο της συστάδας (πχ όταν κατηγορικά γνωρίσματα)



4 συστάδες βασισμένες σε κέντρο



Τείνουν στο να είναι κυκλικοί

# Τύποι συστάδων: Συνεχής Συστάδες

Συνεχής Συστάδες (Contiguous Cluster) (Κοντινότερος γείτονα ή μεταβατικά) – Βάσει γειτνίασης

Μια συστάδα είναι ένα σύνολο σημείων τέτοιο ώστε κάθε σημείο είναι **πιο κοντά σε ένα ή περισσότερα σημεία της συστάδας από ό,τι σε οποιοδήποτε άλλο σημείο εκτός συστάδας**

Συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα – ή όταν έχουμε γραφήματα και θέλουμε να βρούμε συνεκτικά υπογραφήματα

Πρόβλημα με θόρυβο



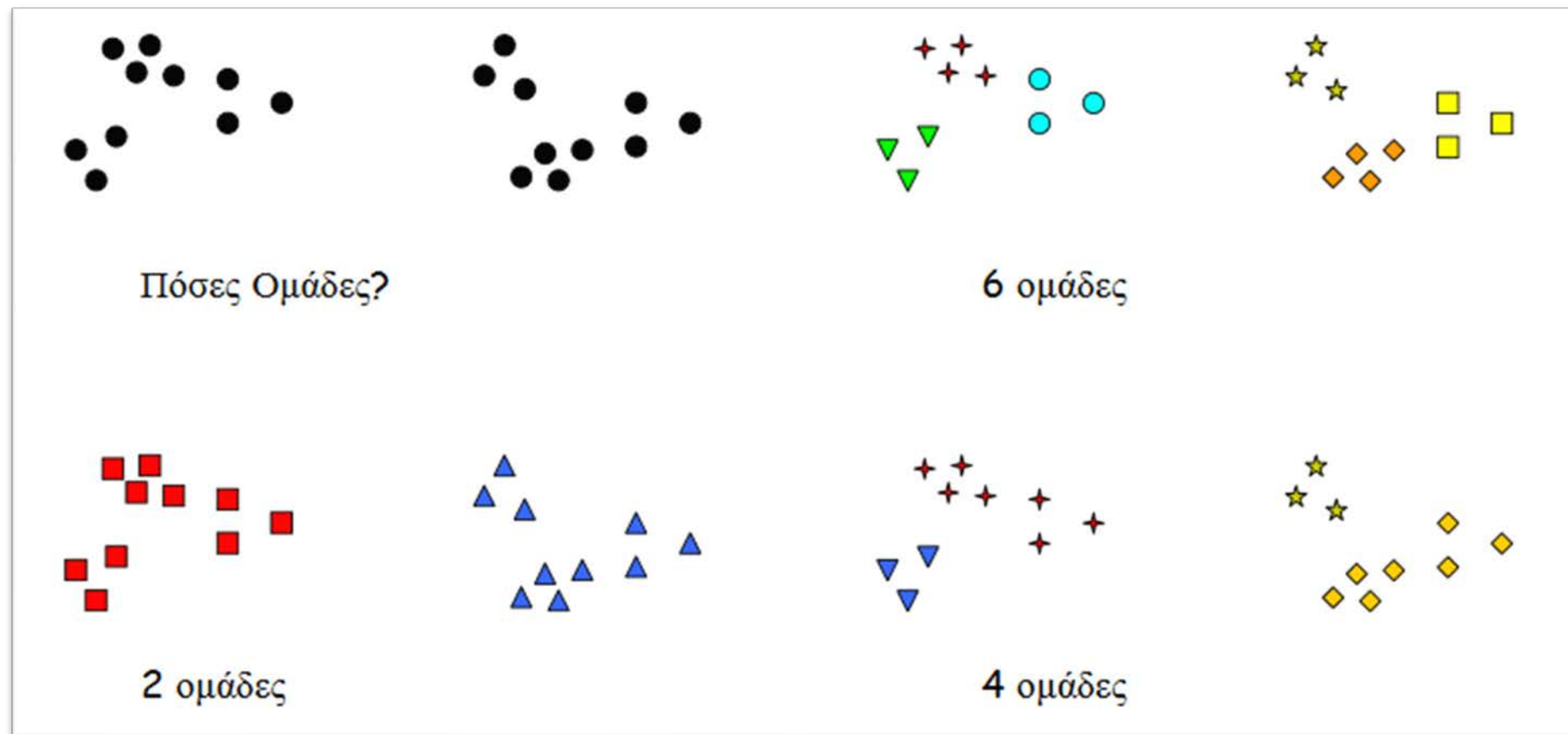
**8 συνεχείς συστάδες**

# Τύποι συστάδων: Συστάδες βασισμένες στην πυκνότητα

- Μια συστάδα είναι μια πυκνή περιοχή από σημεία την οποία χωρίζουν από άλλες περιοχές μεγάλης πυκνότητας περιοχές χαμηλής πυκνότητας
- Συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα ή όταν θόρυβος ή outliers



# Ασαφεια



# Ο αλγόριθμος K-μέσων (k-means)

- Θέλουμε να βρούμε εκείνο το σύνολο  $k$  σημείων στον  $d$ -διάστατο χώρο, το οποίο ελαχιστοποιεί την μέση απόσταση ελαχίστων τετραγώνων κάθε σημείου από το κοντινότερό του κέντρο

$$d_E(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- Η συνάρτηση βαθμολόγησης που βασίζεται στο άθροισμα των τετραγώνων των σφαλμάτων (SSE) ορίζεται ως:

$$SSE(\mathbf{C}) = \sum_{i=1}^k \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

- Ο στόχος μας είναι να βρούμε εκείνη τη συσταδοποίηση που ελαχιστοποιεί τη βαθμολογία SSE:

$$\mathbf{C}^* = \arg \min_{\mathbf{C}} \{SSE(\mathbf{C})\}$$

- Ο αλγόριθμος K μέσων χρησιμοποιεί μια άπληστη επαναληπτική τεχνική για να βρει μια συσταδοποίηση που ελαχιστοποιεί την αντικειμενική συνάρτηση SSE.

# Ο αλγόριθμος K-μέσων (k-means)

- Ο αλγόριθμος K μέσων καθορίζει τις αρχικές τιμές των μέσων για τις συστάδες παράγοντας με τυχαίο τρόπο  $k$  σημεία στον χώρο δεδομένων. Κάθε επανάληψη του αλγορίθμου K μέσων αποτελείται από δύο βήματα: (1) την αντιστοίχιση σε συστάδες και (2) την ενημέρωση των κέντρων βάρους.
- Με την προϋπόθεση ότι δίνονται οι μέσοι των  $k$  συστάδων, κάθε σημείο  $\mathbf{x}_j \in D$  αντιστοιχίζεται στον πλησιέστερο μέσο κατά τη διάρκεια του πρώτου βήματος του αλγορίθμου· αυτό προκαλεί μια συσταδοποίηση, με κάθε συστάδα  $C_i$  να περιλαμβάνει σημεία που βρίσκονται πιο κοντά στον μέσο  $\mu_i$  σε σύγκριση με τον μέσο οποιασδήποτε άλλης συστάδας. Δηλαδή, κάθε σημείο  $\mathbf{x}_j$  αντιστοιχίζεται στη συστάδα  $C_{j^*}$ , όπου

$$j^* = \arg \min_k^{i=1} \{ \| \mathbf{x}_j - \boldsymbol{\mu}_i \|^2 \}$$

- Για ένα καθορισμένο σύνολο συστάδων  $C_i$ ,  $i = 1, \dots, k$ , στο δεύτερο βήμα του αλγορίθμου (ενημέρωση των κέντρων βάρους) υπολογίζονται νέες μέσες τιμές για κάθε συστάδα από τα σημεία του συνόλου  $C_i$ .
- Τα βήματα της αντιστοίχισης σε συστάδες και της ενημέρωσης των κέντρων βάρους εκτελούνται επαναληπτικά μέχρι να καταλήξουμε σε ένα σταθερό σημείο ή σε τοπικά ελάχιστα.

# Ο αλγόριθμος K-μέσων (k-means)

- ΣΚΟΠΟΣ : Εύρεση των κέντρων των ομάδων

- ΜΕΘΟΔΟΣ : Ελαχιστοποίηση του σφάλματος,  $J$

$$J = \sum_{j=1}^k \sum_{i=1}^n (\|x_i^{(j)} - c_j\|)^2$$

- ΒΗΜΑΤΑ

I. Ορισμός  $K$  κέντρων συστάδων με τυχαίο τρόπο

II. Εισαγωγή αντικειμένου στη συστάδα με το πιο κοντινό κέντρο

III. Ανανέωση του κέντρου της συστάδας

IV. Επανάληψη των βημάτων 2,3 μέχρι τη σύγκλιση (αλλαγή στις συστάδες μικρότερη από ένα κατώφλι)

- Ουσιαστικά, ο αλγόριθμος προσπαθεί επαναληπτικά να «μειώσει» την απόσταση όλων των σημείων από ένα σημείο της συστάδας

# Ο αλγόριθμος K-μέσων (k-means)

**K-MEANS** ( $\mathbf{D}, k, \epsilon$ ):

- 1  $t \leftarrow 0$
- 2 Καθορισμός αρχικής τιμής για  $k$  κέντρα βάρους με τυχαίο τρόπο:  $\mu_1^t, \mu_2^t, \dots, \mu_k^t \in \mathbb{R}^d$
- 3 **repeat**
- 4      $t \leftarrow t + 1$
- 5      $C_j \leftarrow \emptyset$  για όλα τα  $j = 1, \dots, k$   
      // Βήμα αντιστοίχισης σε συστάδες
- 6     **foreach**  $\mathbf{x}_j \in \mathbf{D}$  **do**
- 7          $j^* \leftarrow \arg \min_i \left\{ \|\mathbf{x}_j - \mu_i^{t-1}\|^2 \right\}$  // Αντιστοίχιση του  $\mathbf{x}_j$  στο πλησιέστερο κέντρο βάρους
- 8          $C_{j^*} \leftarrow C_{j^*} \cup \{\mathbf{x}_j\}$
- 9         // Βήμα ενημέρωσης των κέντρων βάρους
- 10         **foreach**  $i = 1$  **to**  $k$  **do**
- 11              $\mu_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$
- 11 **until**  $\sum_{i=1}^k \|\mu_i^t - \mu_i^{t-1}\|^2 \leq \epsilon$

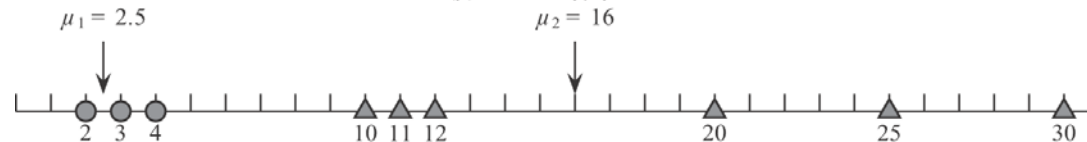
# Ο αλγόριθμος K μέσων στη μία διάσταση



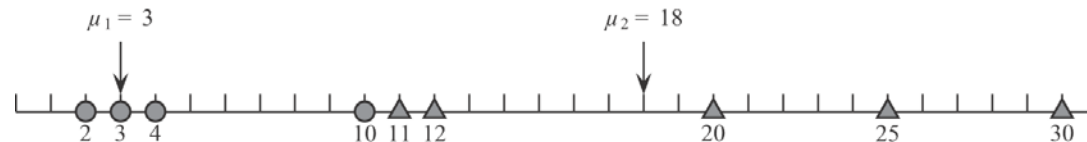
(α) Αρχικό σύνολο δεδομένων



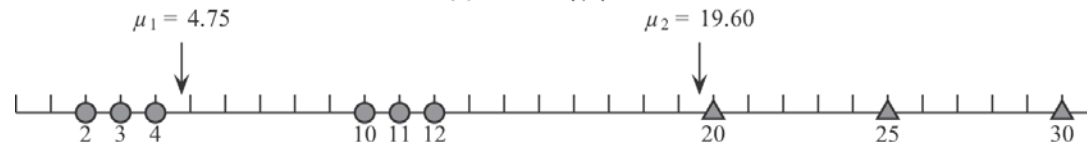
(β) Επανάληψη:  $t = 1$



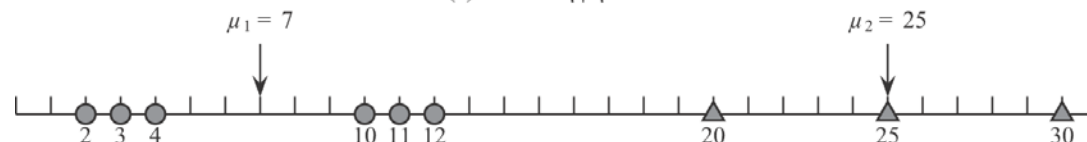
(γ) Επανάληψη:  $t = 2$



(δ) Επανάληψη:  $t = 3$



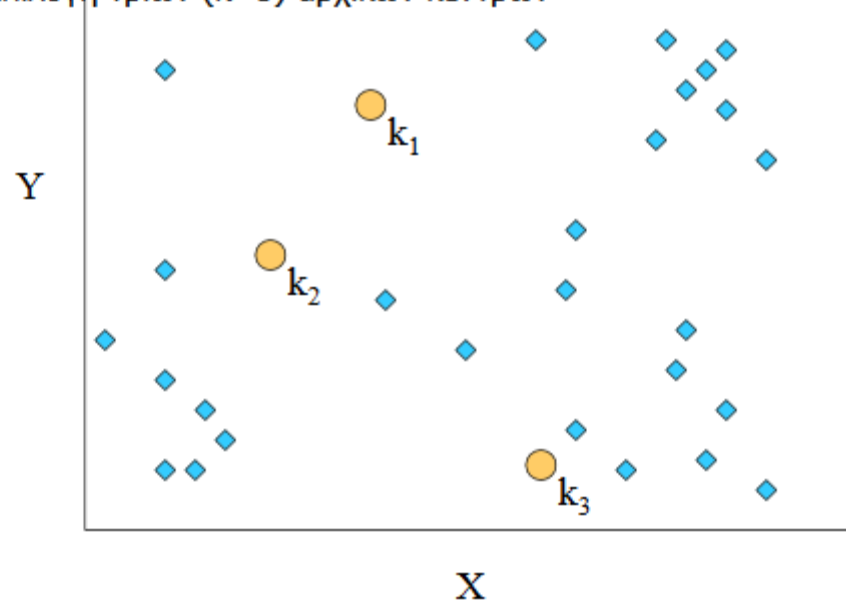
(ε) Επανάληψη:  $t = 4$



(στ) Επανάληψη:  $t = 5$  (σύγκλιση)

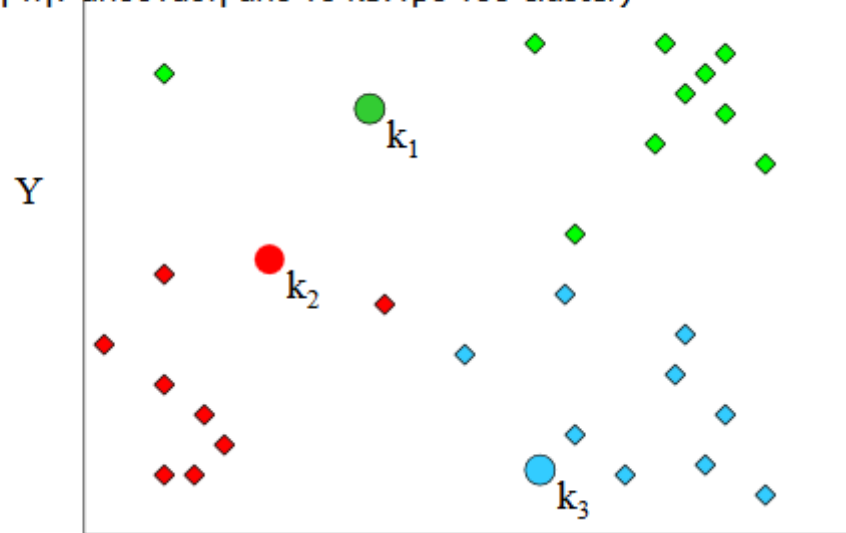
# K-means σε 2 διαστάσεις

- Τυχαία επιλογή τριών ( $k=3$ ) αρχικών κέντρων



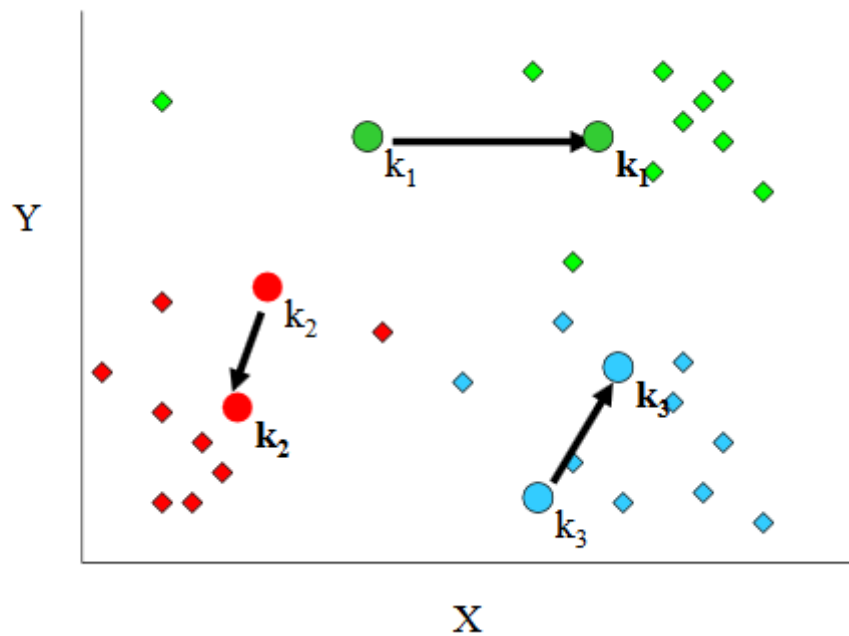
# K-means σε 2 διαστάσεις

- Εκχώρηση κάθε στοιχείου στο πλησιέστερό του cluster (με βάση την απόσταση από το κέντρο του cluster)



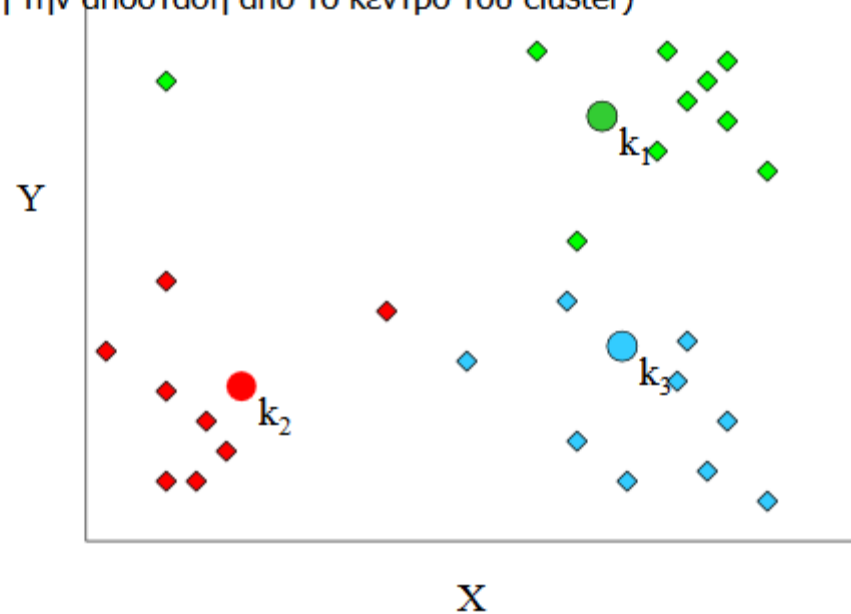
# K-means σε 2 διαστάσεις

- Επανυπολογισμός του νέου κέντρου βάρους του κάθε cluster

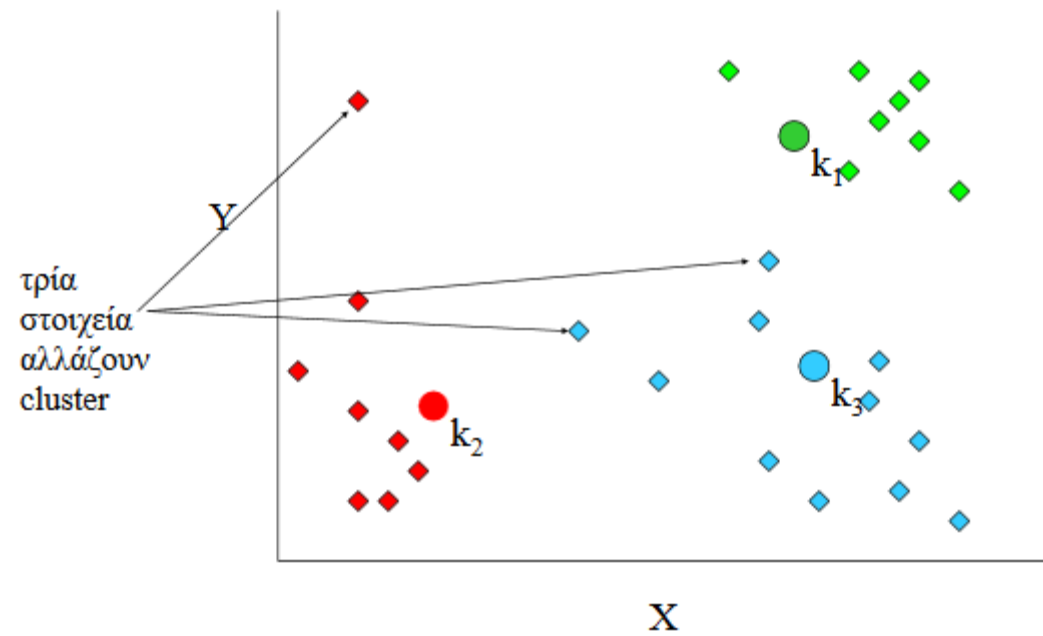


# K-means σε 2 διαστάσεις

- Εκχώρηση κάθε στοιχείου στο πλησιέστερό του cluster (με βάση την απόσταση από το κέντρο του cluster)

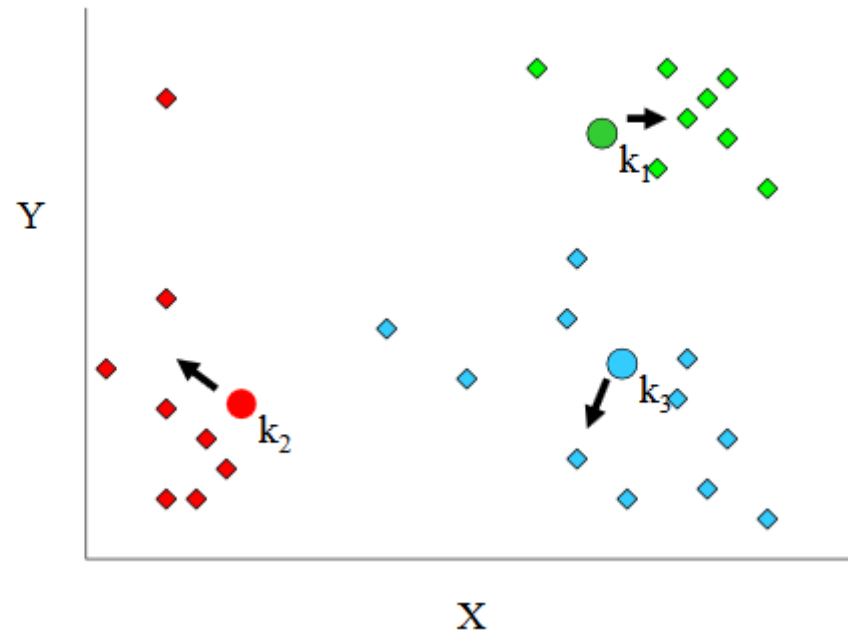


# K-means σε 2 διαστάσεις

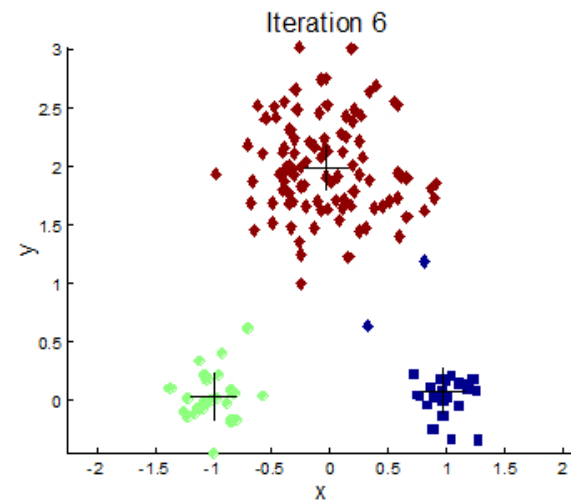
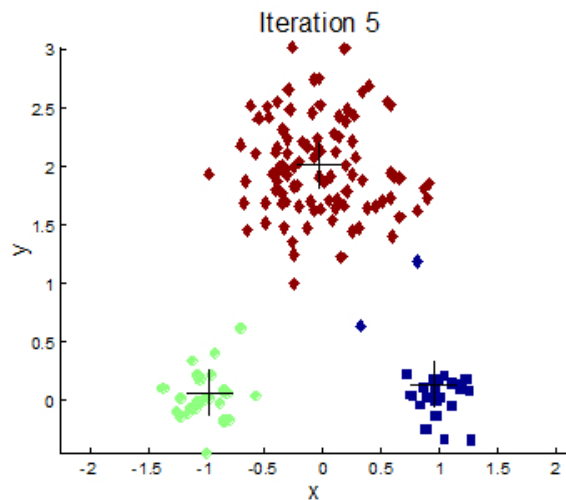
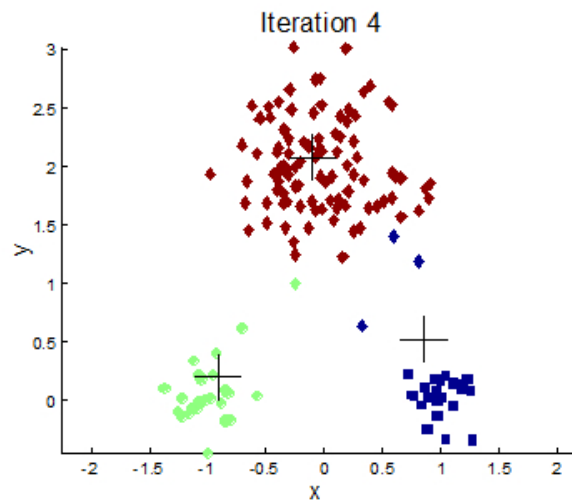
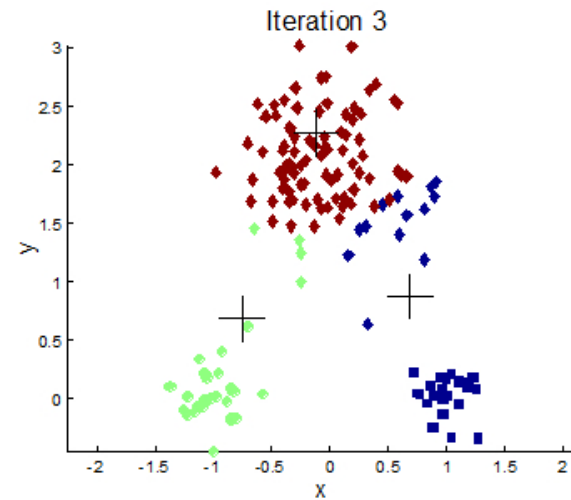
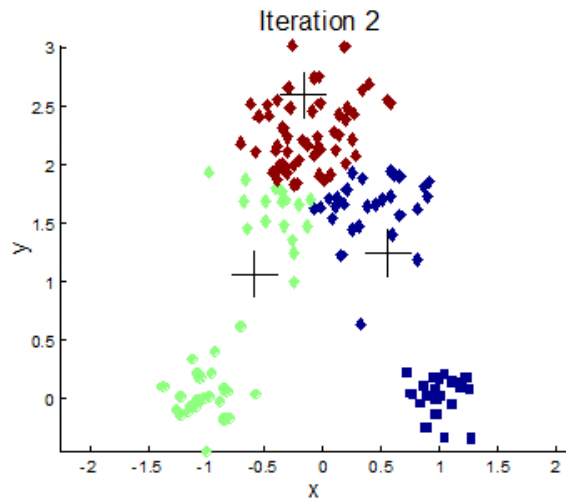
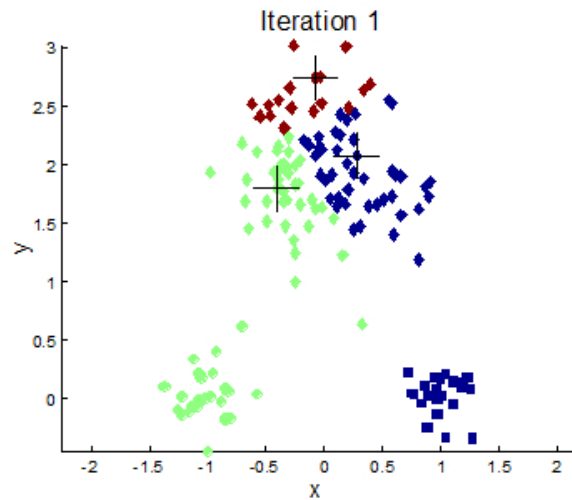


# K-means σε 2 διαστάσεις

- Επανυπολογισμός του νέου κέντρου βάρους του κάθε cluster



# Αλγόριθμος k-means - ΒΗΜΑΤΑ



# Παραδείγματα

## Άσκηση 1

- Δίνεται:  $\{2, 4, 10, 12, 3, 20, 30, 11, 25\}$ ,  $k=2$  και Τυχαία επιλέγουμε, έστω κέντρα  $m_1=3$ ,  $m_2=4$

## Άσκηση 2

- Δίνεται:

x1	x2
1	1
2	1
4	4
1	2
4	5
5	4
1	3
8	3

και τυχαία κέντρα  $k_1 = \{1, 0\}$ ,  $k_2 = \{1, 3\}$

# Επιλογή $k$

---

- Χρήση άλλης μεθόδου ομαδοποίησης
- Εφαρμογή του αλγορίθμου για διάφορες τιμές του  $k$
- Χρήση πρότερης γνώσης για το είδος των δεδομένων
  - Κακοήθης - Καλοήθης

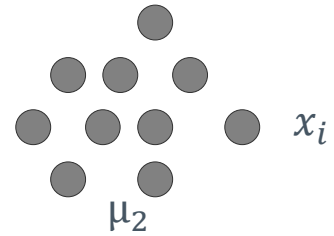
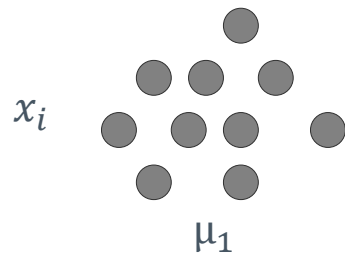
# K-means - Συμπέρασμα

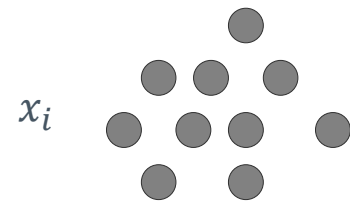
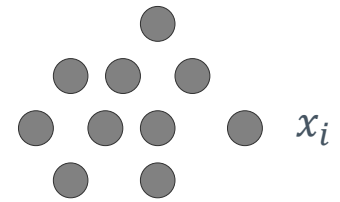
---

- Πλεονεκτήματα
  - Απλός, κατανοητός
  - Τα αντικείμενα ανατίθενται αυτόματα σε κάποιο cluster
  - Ταχύτητα σύγκλισης
- Μειονεκτήματα
  - Πρέπει να οριστεί ο αριθμός των clusters
  - Όλα τα αντικείμενα πρέπει υποχρεωτικά να ανήκουν σε κάποιο cluster
  - Δε δουλεύει για μη αριθμητικά δεδομένα
  - Μη-ντετερμινιστικός

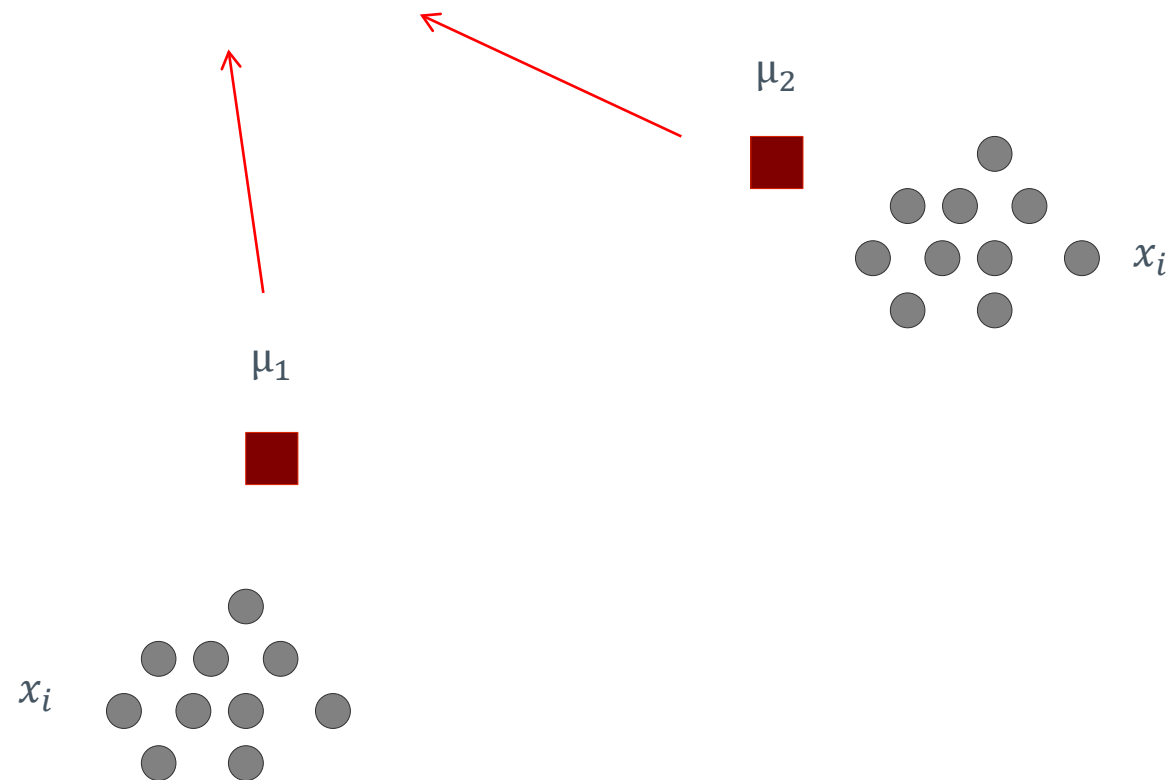
# Ερώτηση

- Ποια αρχικοποίηση θα μπέρδευε τον K-means ? (εστω  $k = 2$ )

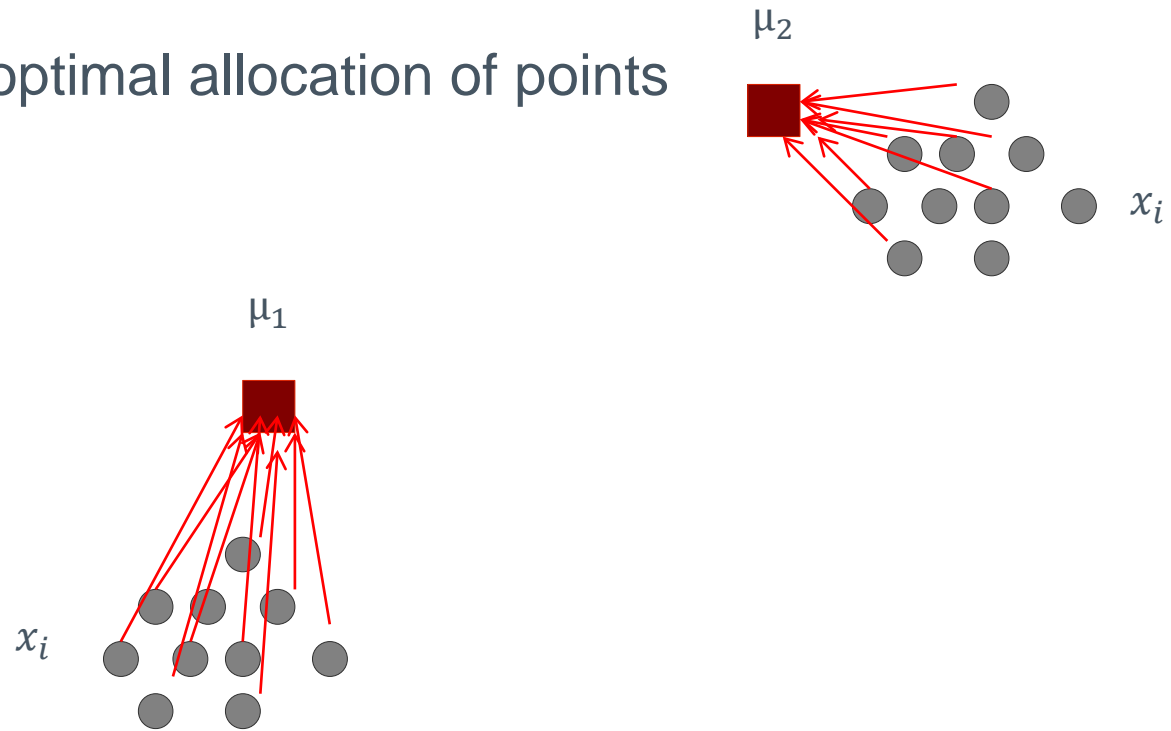




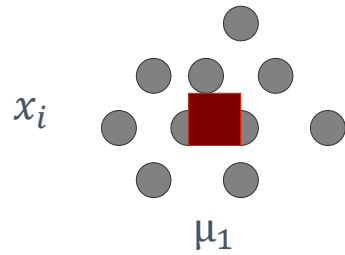
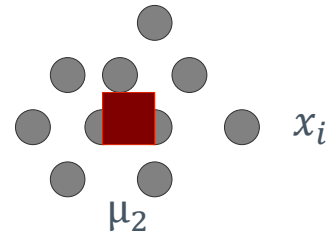
Randomly placed



Find optimal allocation of points

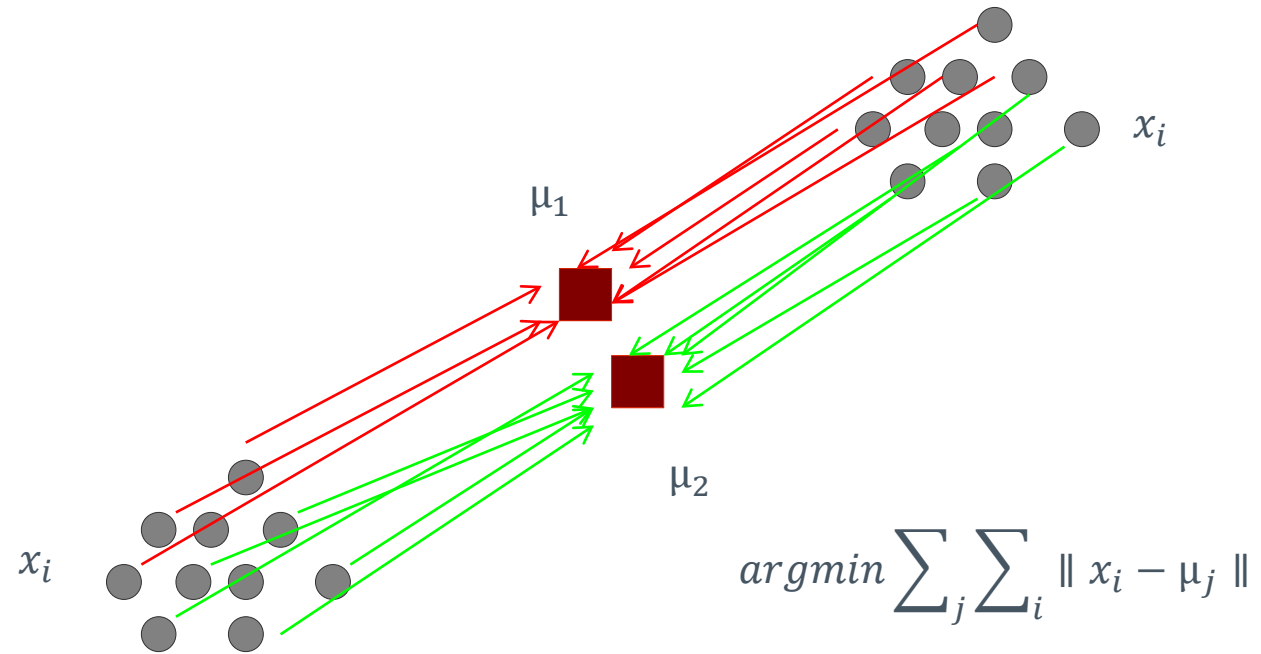


# Reassign and repeat



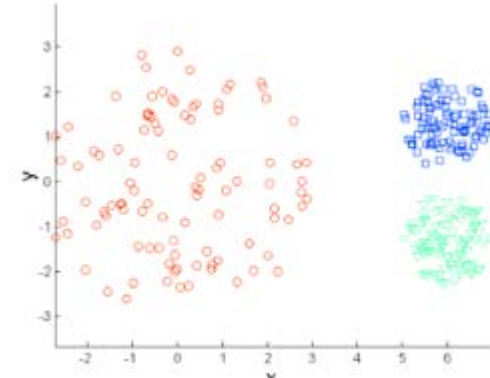
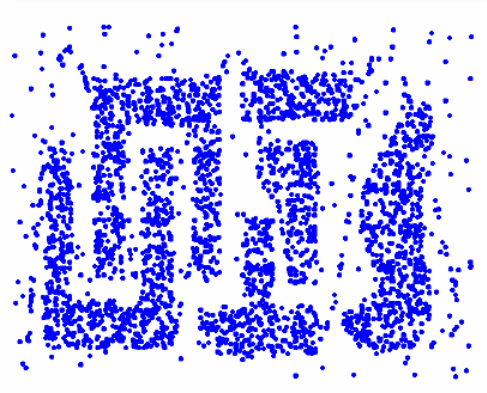
$$\operatorname{argmin} \sum_j \sum_i \|x_i - \mu_j\|$$

# Problems with local optimum of the optimization function

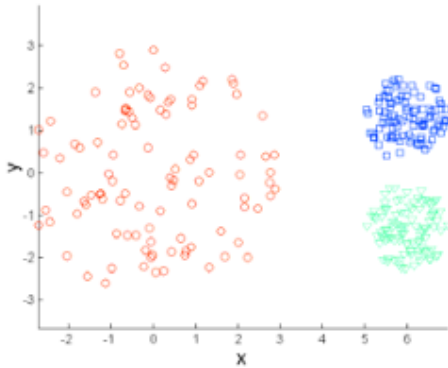


# Ερώτηση

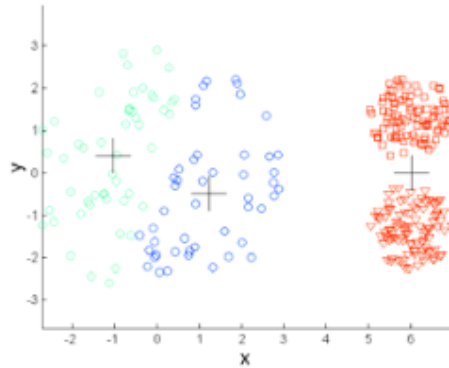
- Θα λειτουργούσε καλά ο k-means (εστω ότι δίνουμε σωστό  $K$ )



# K-means: Περιορισμοί – Διαφορετικές Πυκνότητες



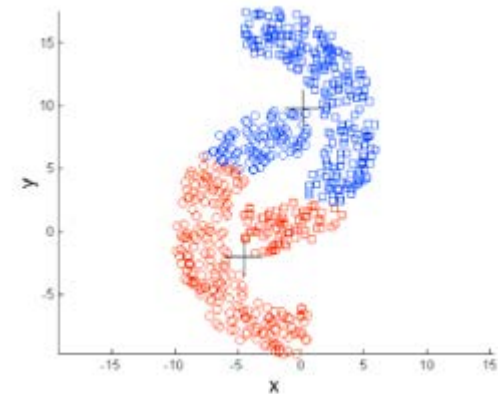
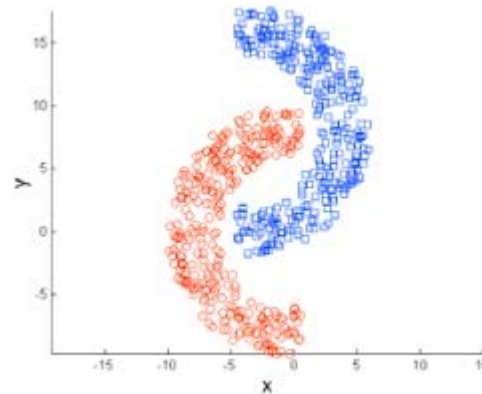
Αρχικά σημεία



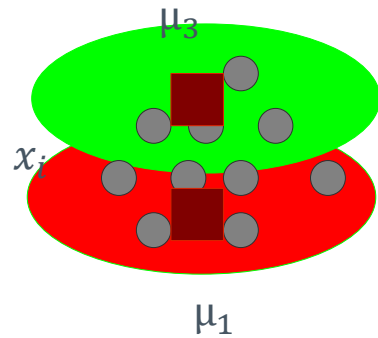
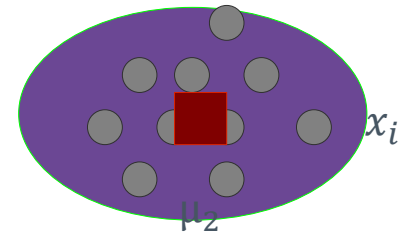
K-means (3 συστάδες)

Δεν μπορεί να διαχωρίσει τους δυο μικρούς γιατί είναι πολύ πυκνοί σε σχέση με τον ένα μεγάλο

Δεν μπορεί να βρει τις δύο συστάδες γιατί έχουν μη κυκλικά σχήματα



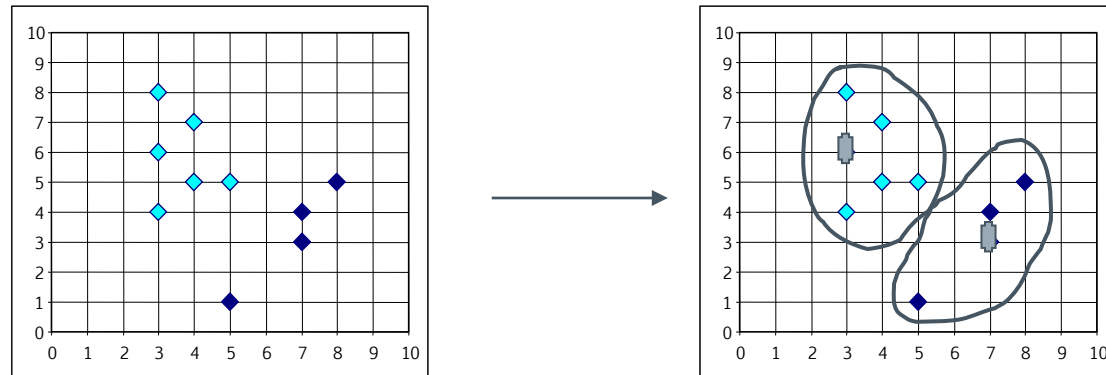
## Problems with wrong number of clusters



$$\operatorname{argmin} \sum_j \sum_i \|x_i - \mu_j\|$$

# K-medoid

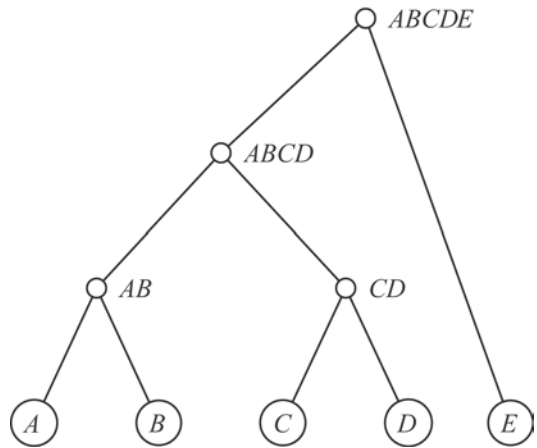
- Συνήθως συνεχή d-διάστατο χώρο
- Διαλέγει ένα αντιπροσωπευτικό σημείο από τα δεδομένα και ελαχιστοποιεί την απόσταση από αυτό – Medoid: το πιο κεντρικό σημείο της συστάδας (αντί να χρησιμοποιεί το mean)
- Μειώνει την ευαισθησία σε outliers
- Μπορεί να εφαρμοστεί σε δεδομένα οποιουδήποτε τύπου (πχ και για κατηγορικά δεδομένα)



# Ιεραρχική συσταδοποίηση: Ένθετες διαμερίσεις

- Αν δίνεται ένα σύνολο δεδομένων  $\mathbf{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ , όπου  $\mathbf{x}_i \in \mathbb{R}^d$ , μια συσταδοποίηση  $C = \{C_1, \dots, C_k\}$  αποτελεί διαμέριση του  $\mathbf{D}$ .
- Λέμε ότι μια συσταδοποίηση  $A = \{A_1, \dots, A_r\}$  είναι ένθετη σε μια άλλη συσταδοποίηση  $B = \{B_1, \dots, B_s\}$  αν και μόνο αν ισχύει  $r > s$ , και για κάθε συστάδα  $A_i \in A$  υπάρχει μια συστάδα  $B_j \in B$  τέτοια ώστε να ισχύει  $A_i \subseteq B_j$ .
- Η ιεραρχική συσταδοποίηση παράγει μια ακολουθία  $n$  ένθετων διαμερίσεων  $C_1, \dots, C_N$ . Η συσταδοποίηση είναι ένθετη στη συσταδοποίηση  $C_t$ .
- Το δενδρόγραμμα συστάδων είναι ένα δυαδικό δένδρο με ρίζα που αποτυπώνει την ένθετη δομή, και στο οποίο υπάρχουν ακμές μεταξύ της συστάδας  $C_i \in C_{t-1}$  και της συστάδας  $C_j \in C_t$  αν η  $C_i$  είναι ένθετη στη  $C_j$ , δηλαδή αν ισχύει  $C_i \subset C_j$ .

# Ιεραρχική συσταδοποίηση: Ένθετες διαμερίσεις



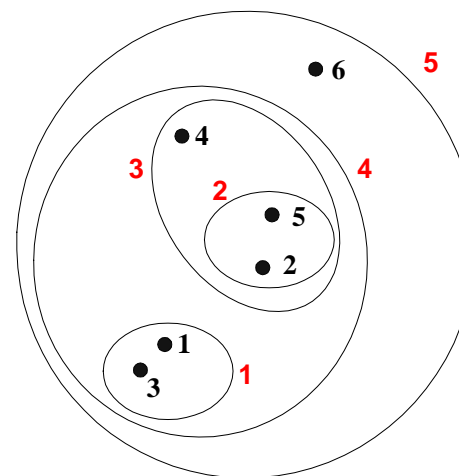
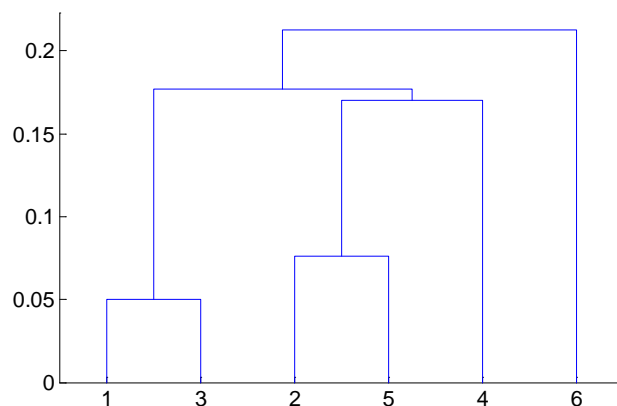
Συσταδοποίηση	Συστάδες
	$\{A\}, \{B\}, \{C\}, \{D\}, \{E\}$
	$\{AB\}, \{C\}, \{D\}, \{E\}$
	$\{AB\}, \{CD\}, \{E\}$
	$\{ABCD\}, \{E\}$
	$\{ABCDE\}$

Το δενδρόγραμμα αναπαριστά την παρακάτω ακολουθία ένθετων διαμερίσεων

- με  $C_{t-1} \in C_t$  για  $t = 2, \dots, 5$ .
- Υποθέτουμε ότι οι συστάδες  $A$  και  $B$  συγχωνεύονται πριν από τις συστάδες  $C$  και  $D$ .

## Ιεραρχική Συσταδοποίηση: Βασικά

- Παράγει ένα σύνολο από εμφωλευμένες συστάδες οργανωμένες σε ένα ιεραρχικό δέντρο
- Μπορεί να παρασταθεί με ένα δένδρο-γγραμμα
  - Ένα διάγραμμα που μοιάζει με δένδρο και καταγράφει τις ακολουθίες από συγχωνεύσεις (merges) και διαχωρισμούς (splits)

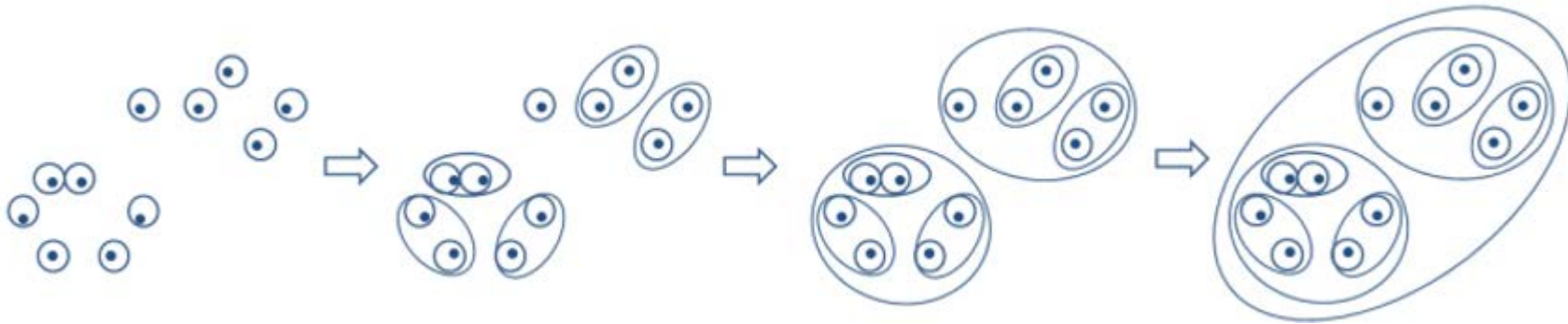


# Ιεραρχική συσταδοποίηση

Δυο βασικοί τύποι ιεραρχικής συσταδοποίησης

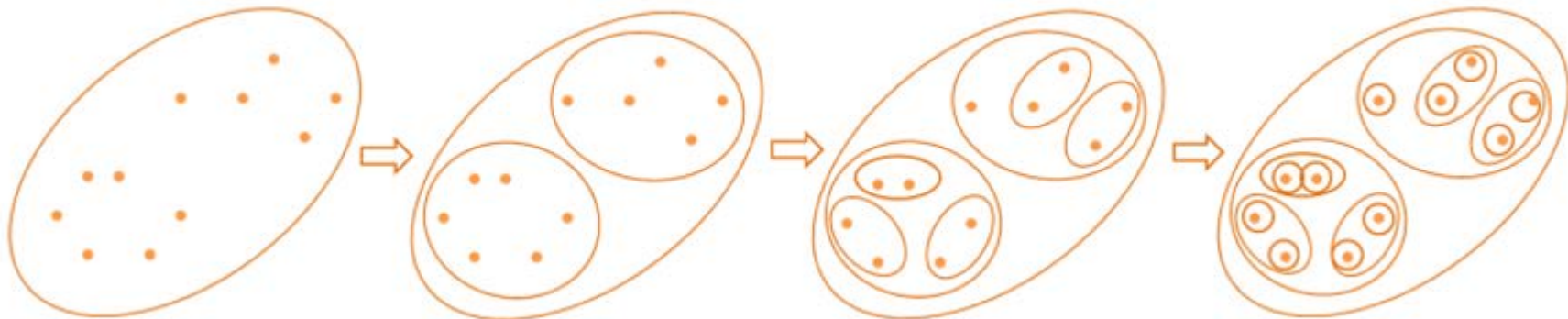
- **Συσσωρευτικός (Agglomerative):**

- Αρχίζει με τα σημεία ως ξεχωριστές συστάδες
- Σε κάθε βήμα, συγχωνεύει το πιο κοντινό ζευγάρι συστάδων μέχρι να μείνει μόνο μία (ή  $k$ ) συστάδες

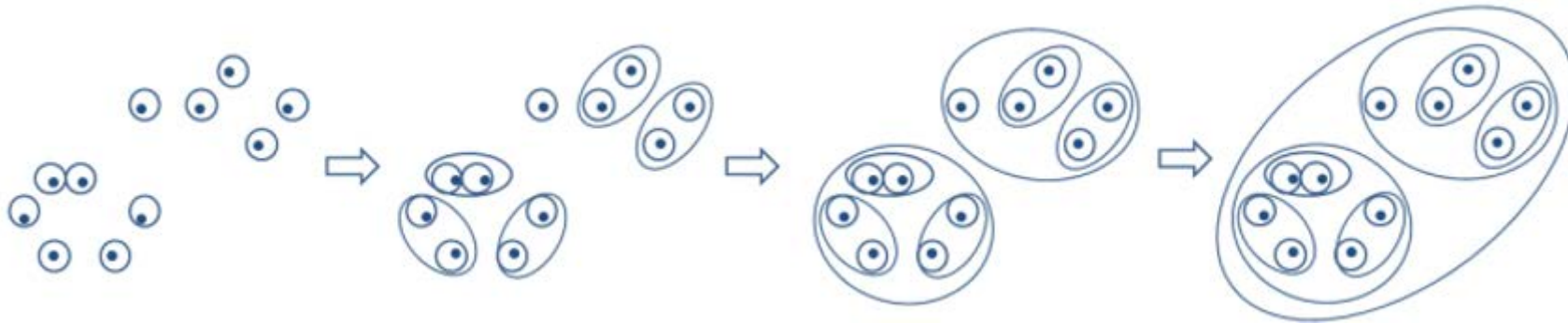


- **Διαιρετικός (Divisive):**

- Αρχίζει με μία συστάδα που περιέχει όλα τα σημεία
- Σε κάθε βήμα, διαχωρίζει μία συστάδα, έως κάθε συστάδα να περιέχει μόνο ένα σημείο (ή να δημιουργηθούν  $k$  συστάδες)



# Συσσωρευτικός (Agglomerative)



## Βασικός Αλγόριθμος

- 1: Υπολογισμός του Πίνακα Γειτνίασης
- 2: Έστω κάθε σημείο αποτελεί και μια συστάδα
- 3: **Repeat**
- 4:     Συγχώνευση των δύο κοντινότερων συστάδων
- 5:     Ενημέρωση του Πίνακα Γειτνίασης
- 6: **Until** να μείνει μία μόνο συστάδα

- Βασική λειτουργία είναι ο υπολογισμός της γειτνίασης δυο συστάδων
- Διαφορετικοί αλγόριθμοι με βάση το πως ορίζεται η απόσταση ανάμεσα σε δύο συστάδες

---

## Απόσταση μεταξύ συστάδων: Μοναδικός σύνδεσμος, πλήρης σύνδεσμος και μέσος όρος ομάδας

---

- Η απόσταση δύο σημείων υπολογίζεται συνήθως με χρήση της Ευκλείδειας απόστασης ή  $L_2$ -νόρμας, που ορίζεται ως

$$\delta(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2 = \left( \sum_{i=1}^d (x_i - y_i)^2 \right)^{1/2}$$

- Οι «διασυσταδικές» αποστάσεις υπολογίζονται ως εξής.
- **Μοναδικός σύνδεσμος:** Η ελάχιστη απόσταση ενός σημείου της συστάδας  $C_i$  από ένα σημείο της συστάδας  $C_j$

$$\delta(C_i, C_j) = \min \{ \delta(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C_i, \mathbf{y} \in C_j \}$$

- **Πλήρης σύνδεσμος:** Η μέγιστη απόσταση μεταξύ ενός σημείου της συστάδας  $C_i$  και ενός σημείου της συστάδας  $C_j$

$$\delta(C_i, C_j) = \max \{ \delta(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C_i, \mathbf{y} \in C_j \}$$

- **Μέσος όρος ομάδας :** Ο μέσος όρος της απόστασης ανά ζεύγη μεταξύ σημείων της συστάδας  $C_i$  και της συστάδας  $C_j$

$$\delta(C_i, C_j) = \frac{\sum_{\mathbf{x} \in C_i} \sum_{\mathbf{y} \in C_j} \delta(\mathbf{x}, \mathbf{y})}{n_i \cdot n_j}$$

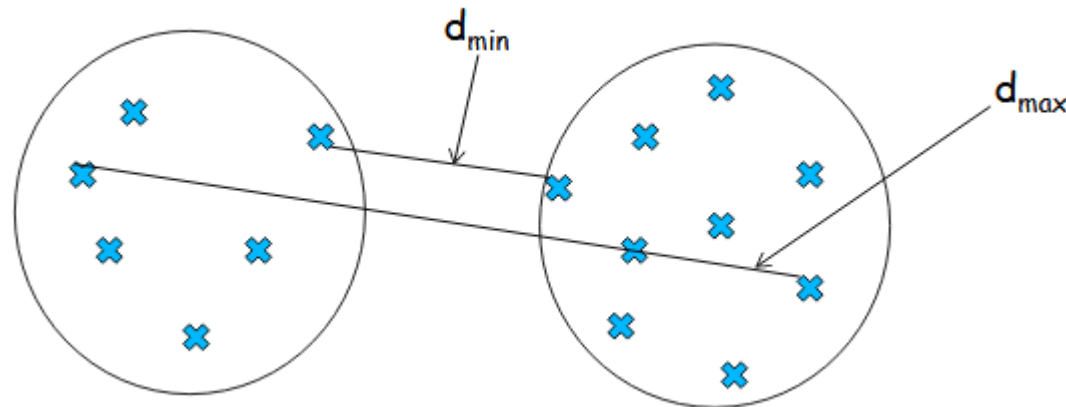
## Απόσταση μεταξύ συστάδων: Μοναδικός σύνδεσμος, πλήρης σύνδεσμος και μέσος όρος ομάδας

- Οι «διασυσταδικές» αποστάσεις υπολογίζονται ως εξής.
- **Μοναδικός σύνδεσμος:** Η ελάχιστη απόσταση ενός σημείου της συστάδας  $C_i$  από ένα σημείο της συστάδας  $C_j$

$$— \delta(C_i, C_j) = \min \{ \delta(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C_i, \mathbf{y} \in C_j \}$$

- **Πλήρης σύνδεσμος:** Η μέγιστη απόσταση μεταξύ ενός σημείου της συστάδας  $C_i$  και ενός σημείου της συστάδας  $C_j$

$$— \delta(C_i, C_j) = \max \{ \delta(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C_i, \mathbf{y} \in C_j \}$$



# Ιεραρχική Ομαδοποίηση

Στον πίνακα δίνονται οι τιμές από δύο μετρήσεις ( $X_1$  και  $X_2$ ).

1. Να ομαδοποιήσετε τα δεδομένα με τον αλγόριθμο «Συσσωρευτικής Ιεραρχικής Ομαδοποίησης» (Agglomerative Hierarchical Clustering), χρησιμοποιώντας ως κριτήριο της απόστασης μεταξύ ομάδων, την περίπτωση του απλού συνδέσμου (MIN ή single link) και πλήρους συνδέσμου (MAX ή complete linkage).
2. Να κατασκευαστεί το δενδρόγραμμα της κάθε ομαδοποίησης και να τα συγκρίνετε.

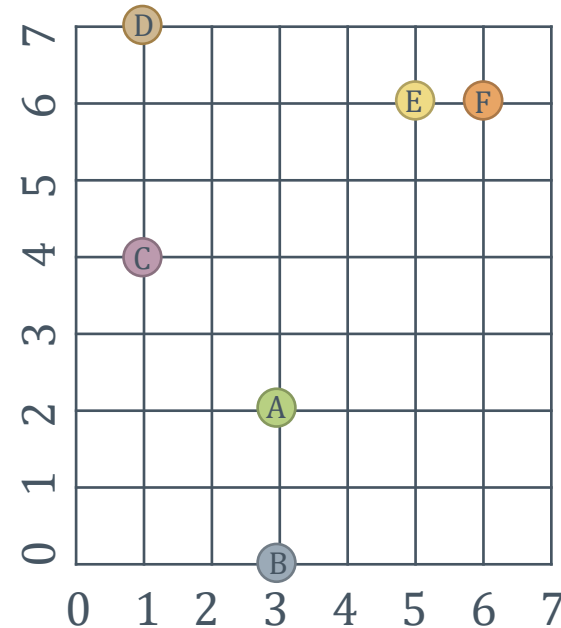
	$X_1$	$X_2$
D	1	7
B	3	0
A	3	2
E	5	6
F	6	6
C	1	4

# Ιεραρχική Ομαδοποίηση

Στον πίνακα δίνονται οι τιμές από δύο μετρήσεις ( $X_1$  και  $X_2$ ).

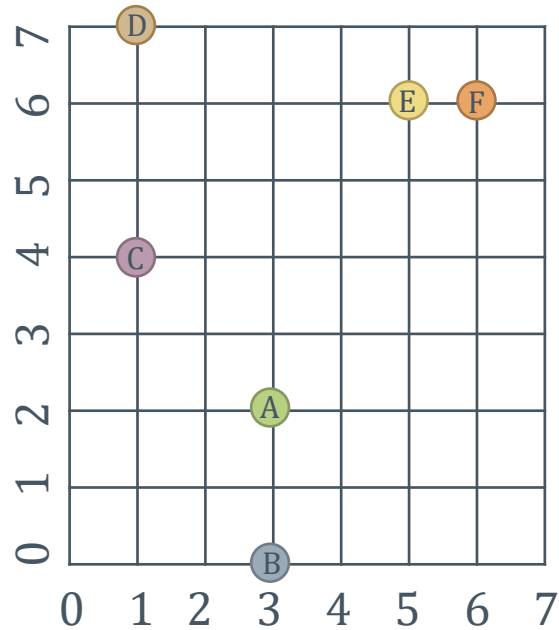
1. Να ομαδοποιήσετε τα δεδομένα με τον αλγόριθμο «Συσσωρευτικής Ιεραρχικής Ομαδοποίησης» (Agglomerative Hierarchical Clustering), χρησιμοποιώντας ως κριτήριο της απόστασης μεταξύ ομάδων, την περίπτωση του απλού συνδέσμου (MIN ή single link) και πλήρους συνδέσμου (MAX ή complete linkage).
2. Να κατασκευαστεί το δενδρόγραμμα της κάθε ομαδοποίησης και να τα συγκρίνετε.

	$X_1$	$X_2$
D	1	7
B	3	0
A	3	2
E	5	6
F	6	6
C	1	4



# Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)

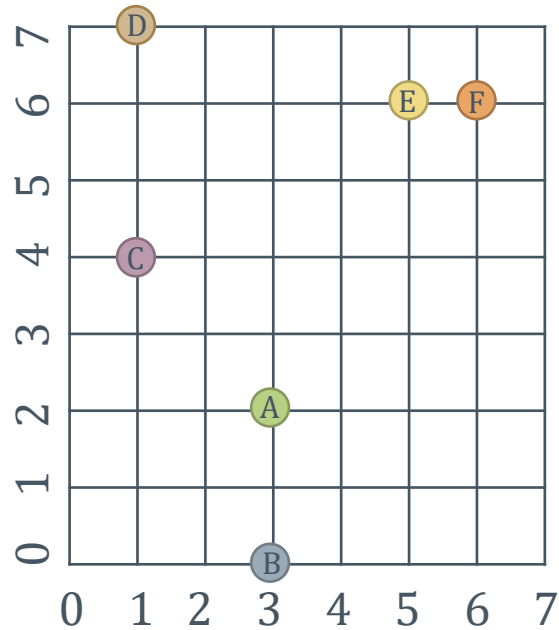


Πίνακας Αποστάσεων

	A	B	C	D	E	F
A	0	2	4	7	6	7
B	2	0	6	9	8	9
C	4	6	0	3	6	7
D	7	9	3	0	5	6
E	6	8	6	5	0	1
F	7	9	7	6	1	0

# Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



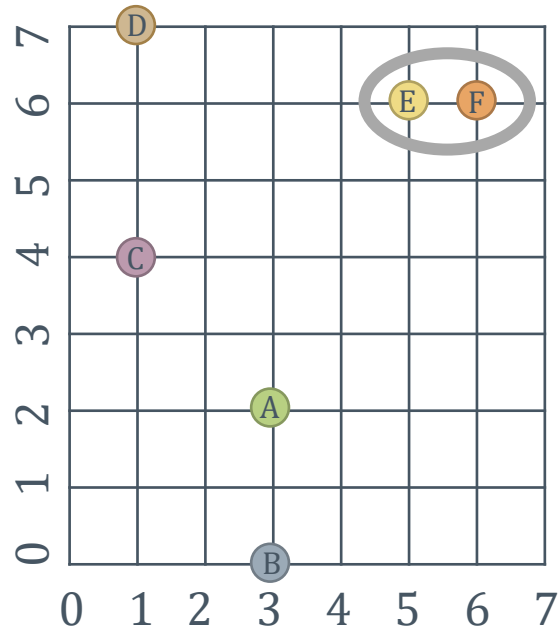
Πίνακας Αποστάσεων

	A	B	C	D	E	F
A	0	2	4	7	6	7
B	2	0	6	9	8	9
C	4	6	0	3	6	7
D	7	9	3	0	5	6
E	6	8	6	5	0	1
F	7	9	7	6	1	0



# Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



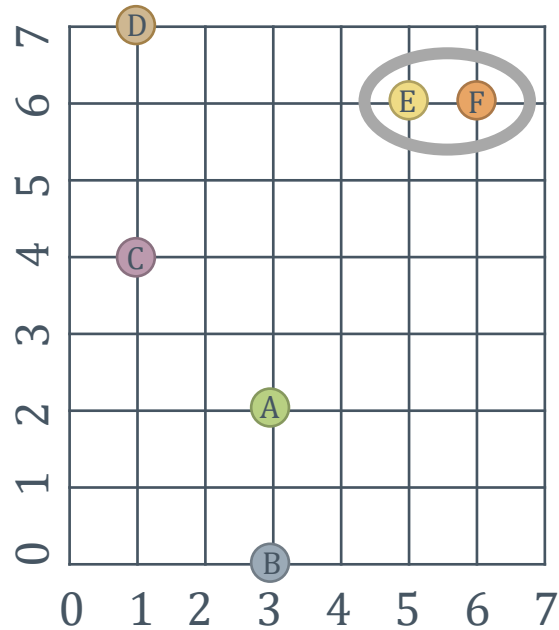
Πίνακας Αποστάσεων

	A	B	C	D	E	F
A	0	2	4	7	6	7
B	2	0	6	9	8	9
C	4	6	0	3	6	7
D	7	9	3	0	5	6
E	6	8	6	5	0	1
F	7	9	7	6	1	0



# Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



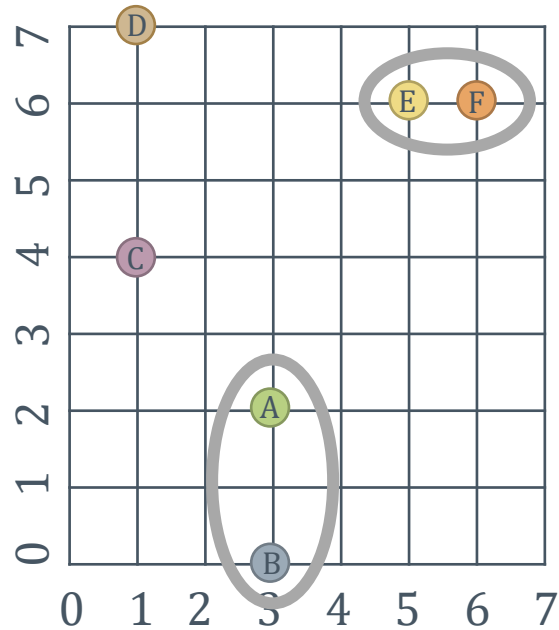
Πίνακας Αποστάσεων

	A	B	C	D	E/F
A	0	2	4	7	6
B	2	0	6	9	8
C	4	6	0	3	6
D	7	9	3	0	5
E/F	6	8	6	5	0



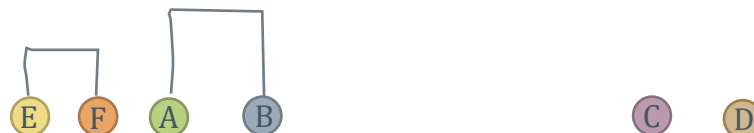
# Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



Πίνακας Αποστάσεων

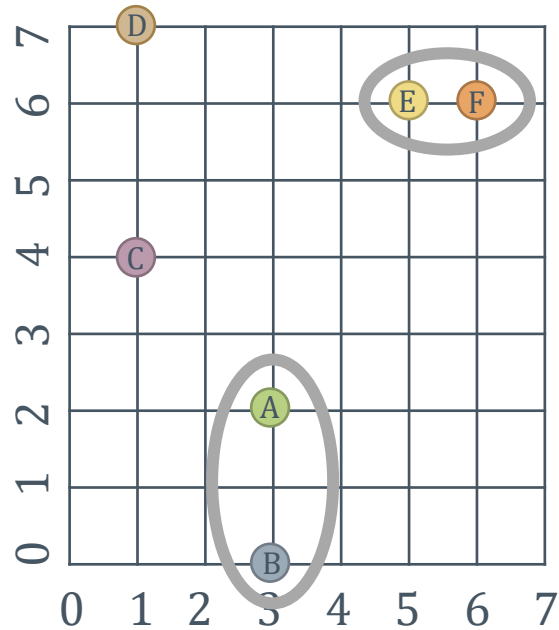
	A	B	C	D	E/F
A	0	2	4	7	6
B	2	0	6	9	8
C	4	6	0	3	6
D	7	9	3	0	5
E/F	6	8	6	5	0



# Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)

Πίνακας Αποστάσεων

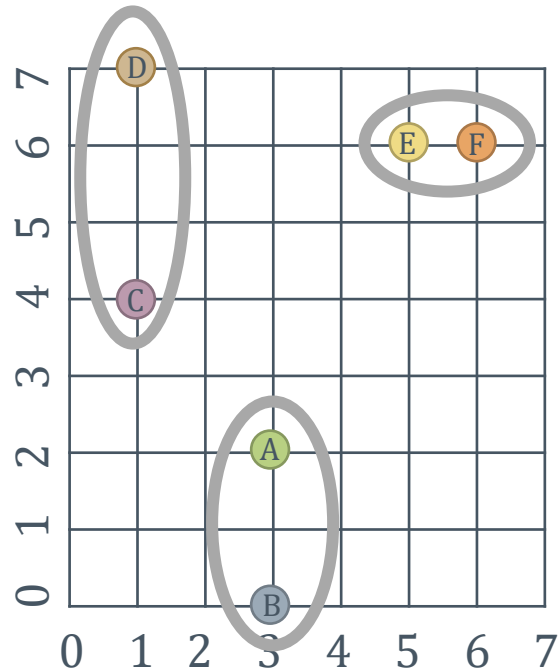


	A/B	C	D	E/F
A/B	0	4	7	6
C	4	0	3	6
D	7	3	0	5
E/F	6	6	5	0



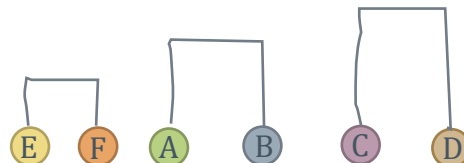
# Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



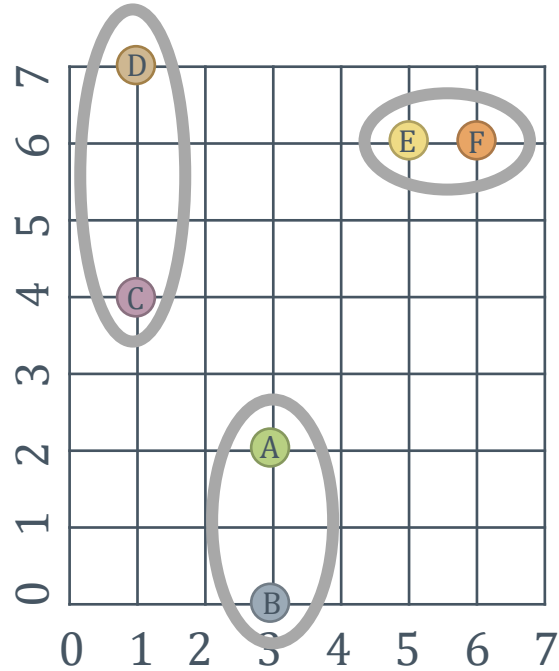
Πίνακας Αποστάσεων

	A/B	C	D	E/F
A/B	0	4	7	6
C	4	0	3	6
D	7	3	0	5
E/F	6	6	5	0



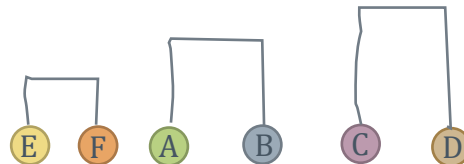
# Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



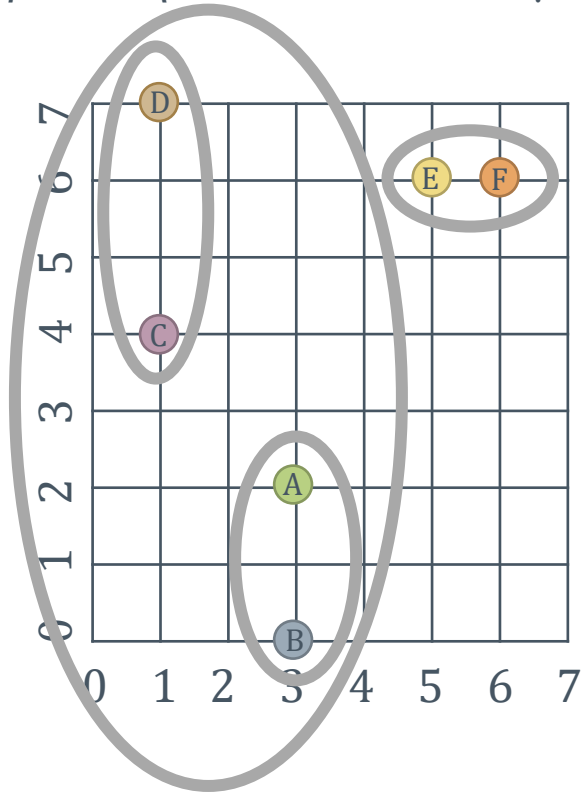
Πίνακας Αποστάσεων

	A/B	C/D	E/F
A/B	0	4	6
C/D	4	0	5
E/F	6	5	0



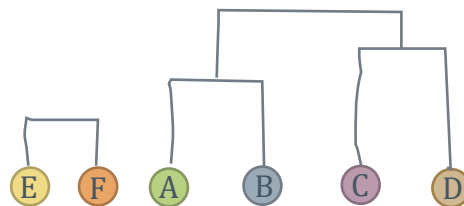
# Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



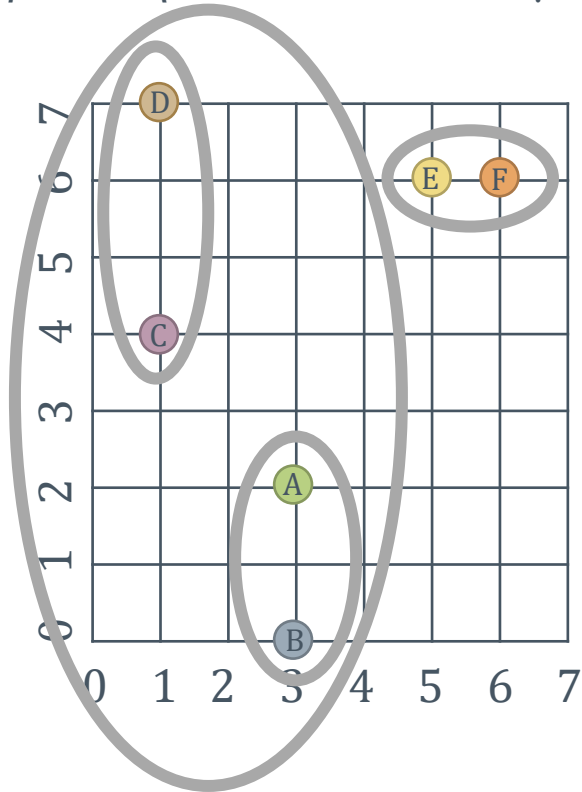
Πίνακας Αποστάσεων

	A/B	C/D	E/F
A/B	0	4	6
C/D	4	0	5
E/F	6	5	0



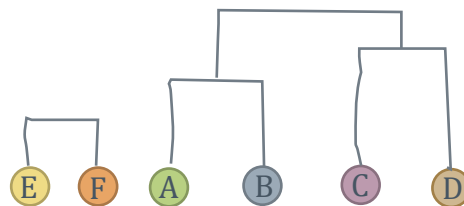
# Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



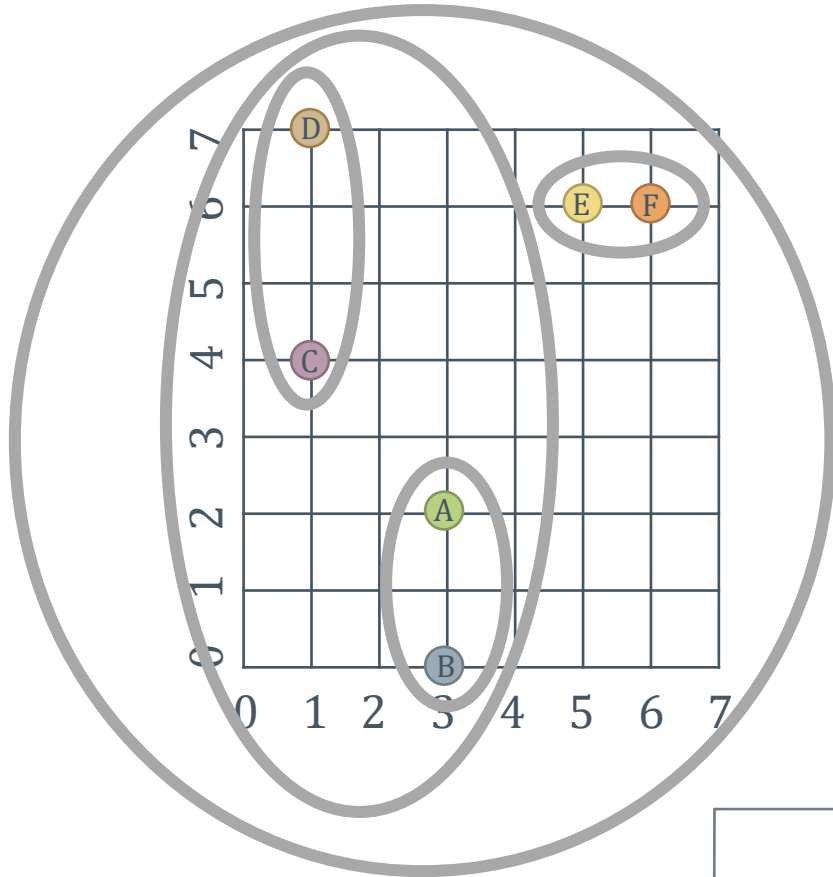
Πίνακας Αποστάσεων

	A/B/C/D	E/F
A/B/C/D	0	5
E/F	5	0



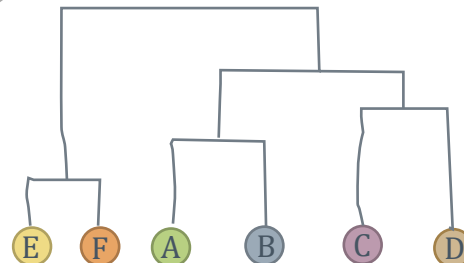
# Ιεραρχική Ομαδοποίηση

- Περίπτωση του απλού συνδέσμου (MIN ή single link)



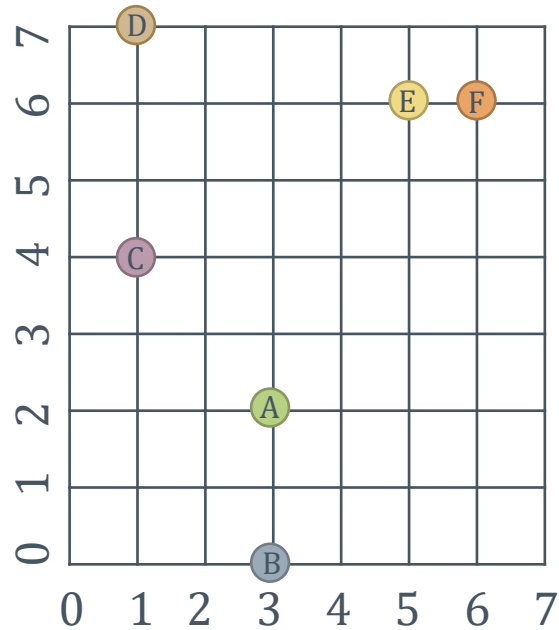
Πίνακας Αποστάσεων

	A/B/C/DE/F
A/B/C/D/E/F	0



# Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)



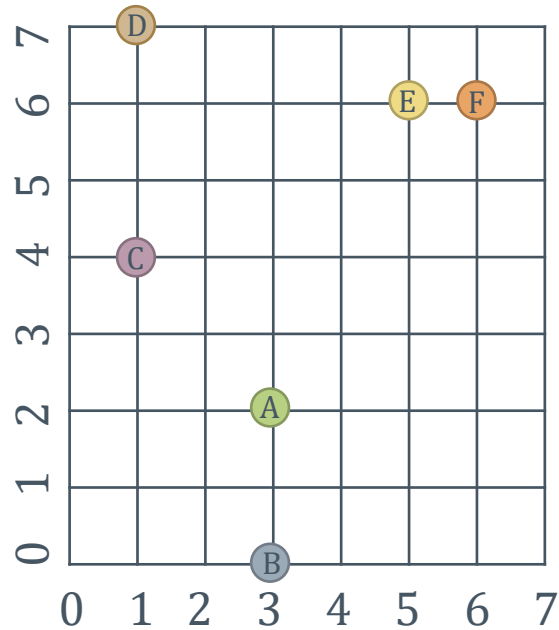
Πίνακας Αποστάσεων

	A	B	C	D	E	F
A	0	2	4	7	6	7
B	2	0	6	9	8	9
C	4	6	0	3	6	7
D	7	9	3	0	5	6
E	6	8	6	5	0	1
F	7	9	7	6	1	0

# Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)

Πίνακας Αποστάσεων



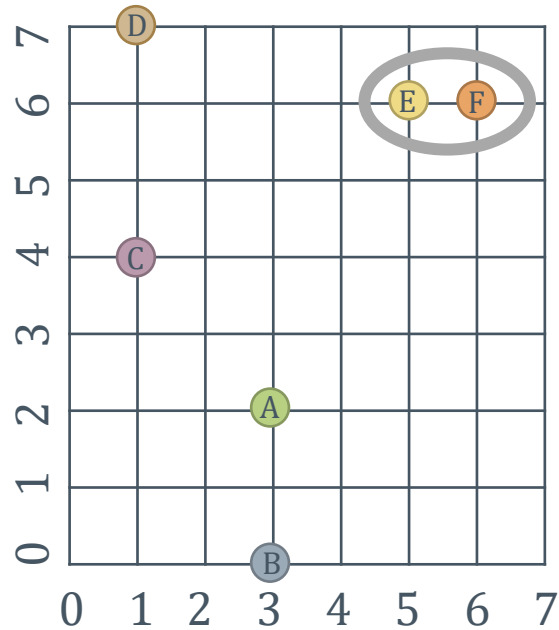
	A	B	C	D	E	F
A	0	2	4	7	6	7
B	2	0	6	9	8	9
C	4	6	0	3	6	7
D	7	9	3	0	5	6
E	6	8	6	5	0	1
F	7	9	7	6	1	0



# Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)

Πίνακας Αποστάσεων



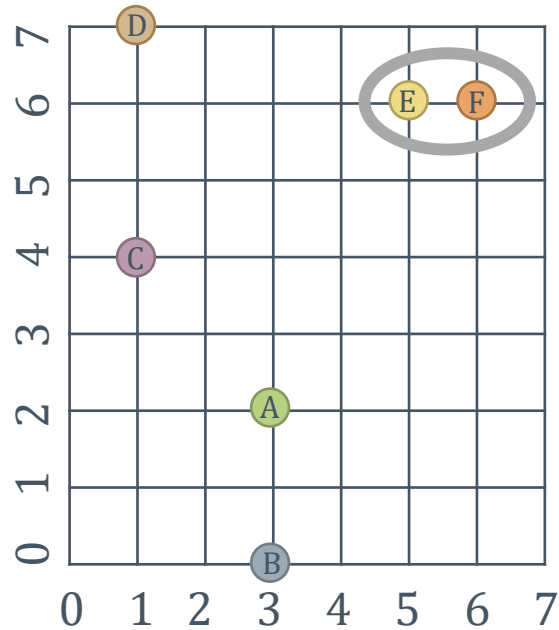
	A	B	C	D	E	F
A	0	2	4	7	6	7
B	2	0	6	9	8	9
C	4	6	0	3	6	7
D	7	9	3	0	5	6
E	6	8	6	5	0	1
F	7	9	7	6	1	0



# Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)

Πίνακας Αποστάσεων

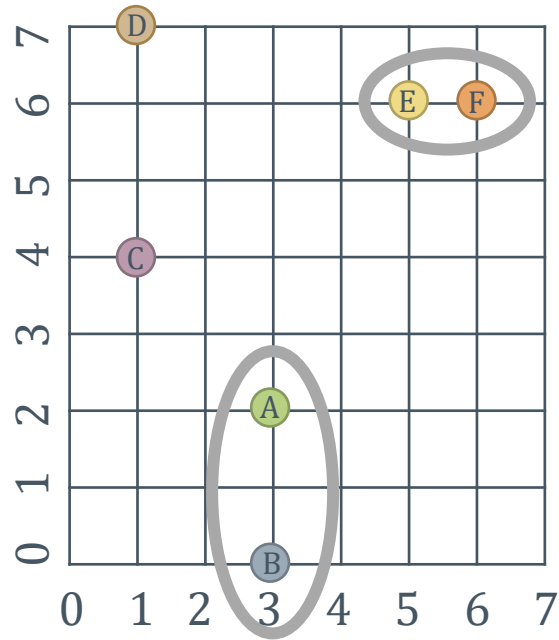


	A	B	C	D	E/F
A	0	2	4	7	7
B	2	0	6	9	9
C	4	6	0	3	7
D	7	9	3	0	6
E/F	7	9	7	6	0



# Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)



Πίνακας Αποστάσεων

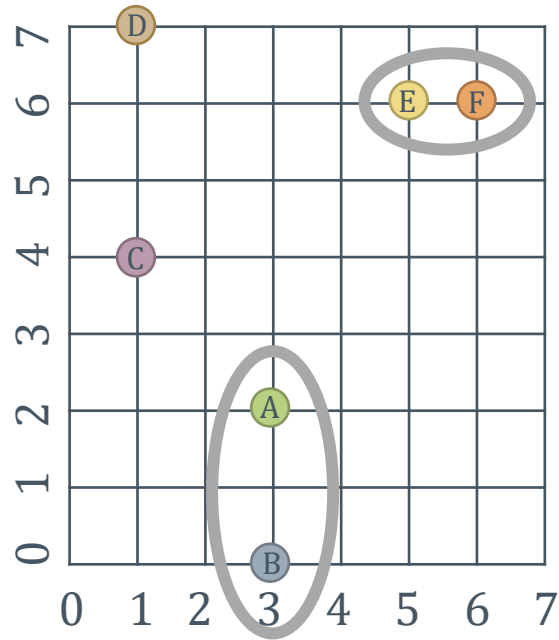
	A	B	C	D	E/F
A	0	2	4	7	7
B	2	0	6	9	9
C	4	6	0	3	7
D	7	9	3	0	6
E/F	7	9	7	6	0



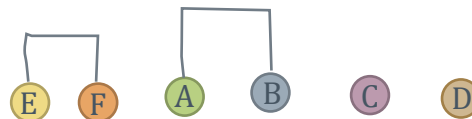
# Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)

Πίνακας Αποστάσεων

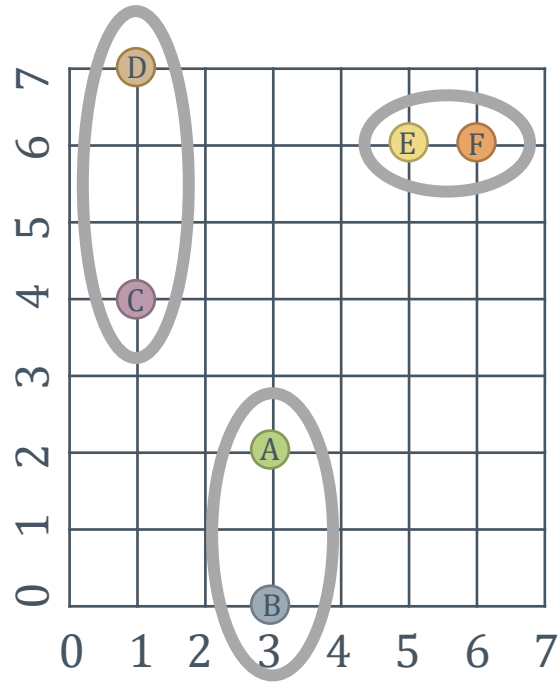


	A/B	C	D	E/F
A/B	0	6	9	9
C	6	0	3	7
D	9	3	0	6
E/F	9	7	6	0



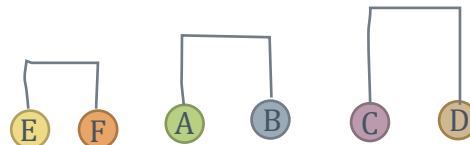
# Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)



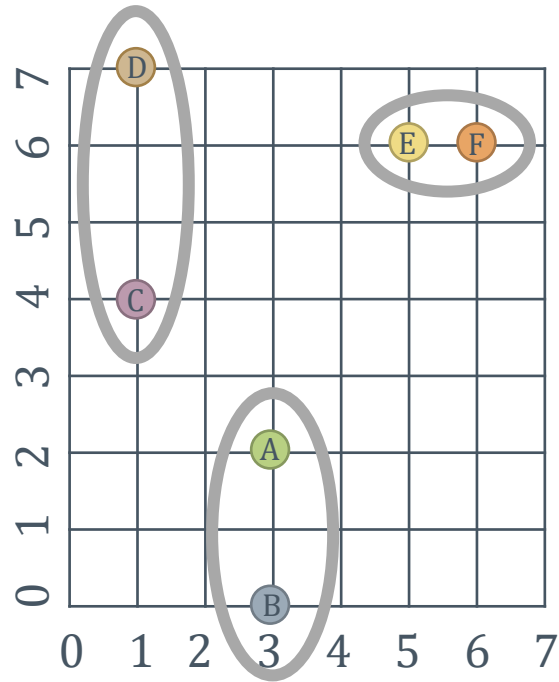
Πίνακας Αποστάσεων

	A/B	C	D	E/F
A/B	0	6	9	9
C	6	0	3	7
D	9	3	0	6
E/F	9	7	6	0



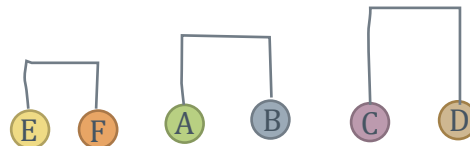
# Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)



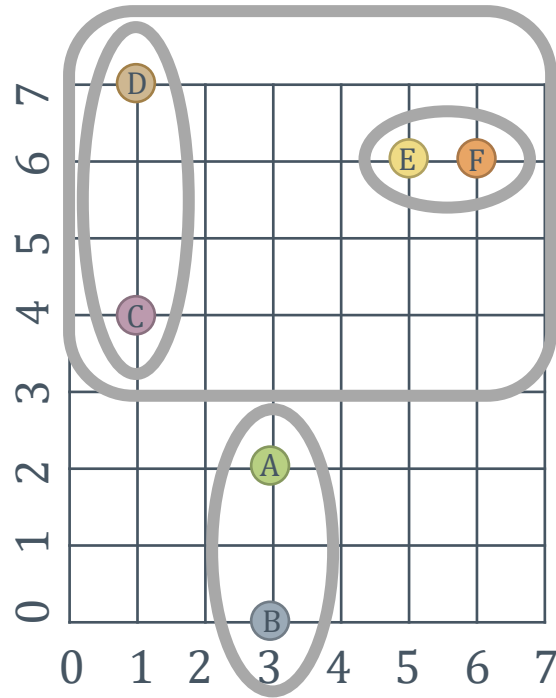
Πίνακας Αποστάσεων

	A/B	C/D	E/F
A/B	0	9	9
C/D	9	0	7
E/F	9	7	0



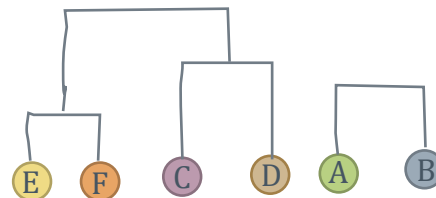
# Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)



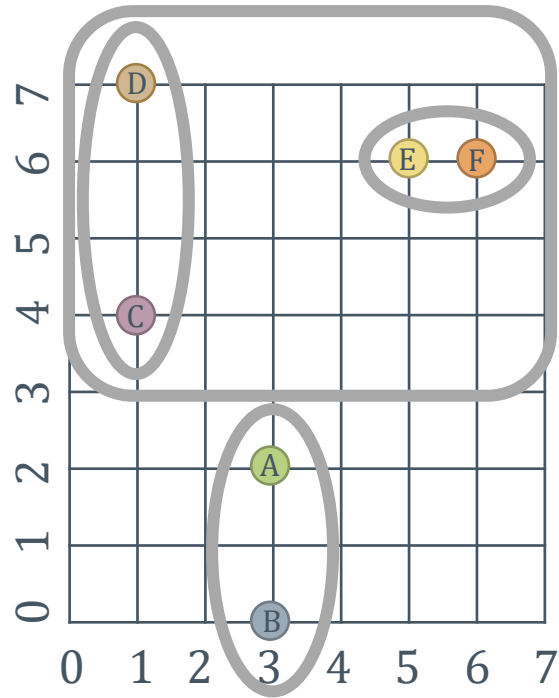
Πίνακας Αποστάσεων

	A/B	C/D	E/F
A/B	0	9	9
C/D	9	0	7
E/F	9	7	0



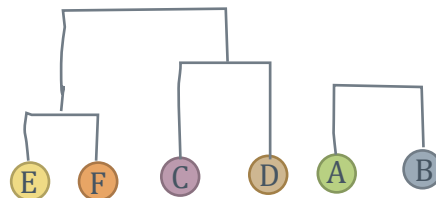
# Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)



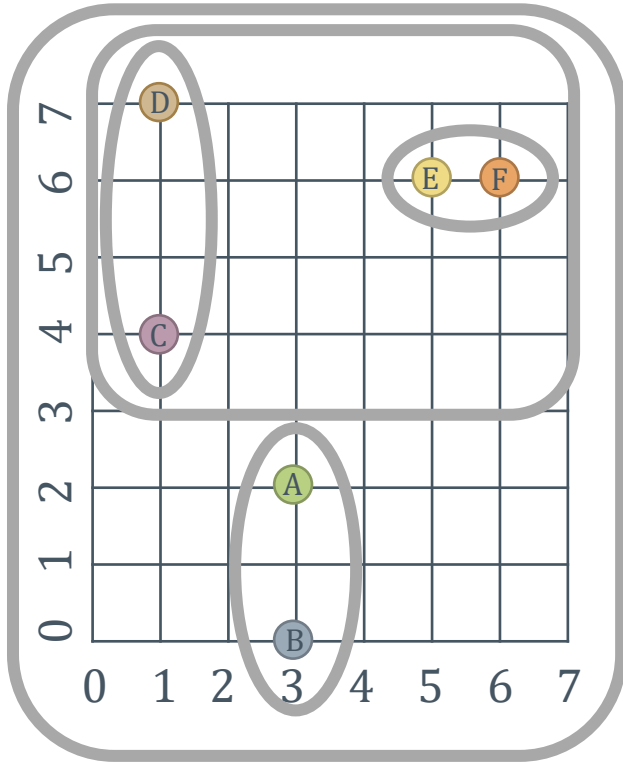
Πίνακας Αποστάσεων

	A/B	C/D/E/F
A/B	0	9
C/D/E/F	9	0



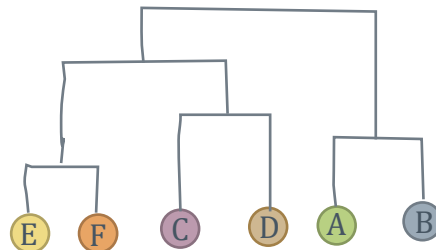
# Ιεραρχική Ομαδοποίηση

- Περίπτωση του πλήρους συνδέσμου (MAX ή complete linkage)



Πίνακας Αποστάσεων

	A/BC/D/E/F
A/BC/D/E/F	0



# Ιεραρχική Ομαδοποίηση

