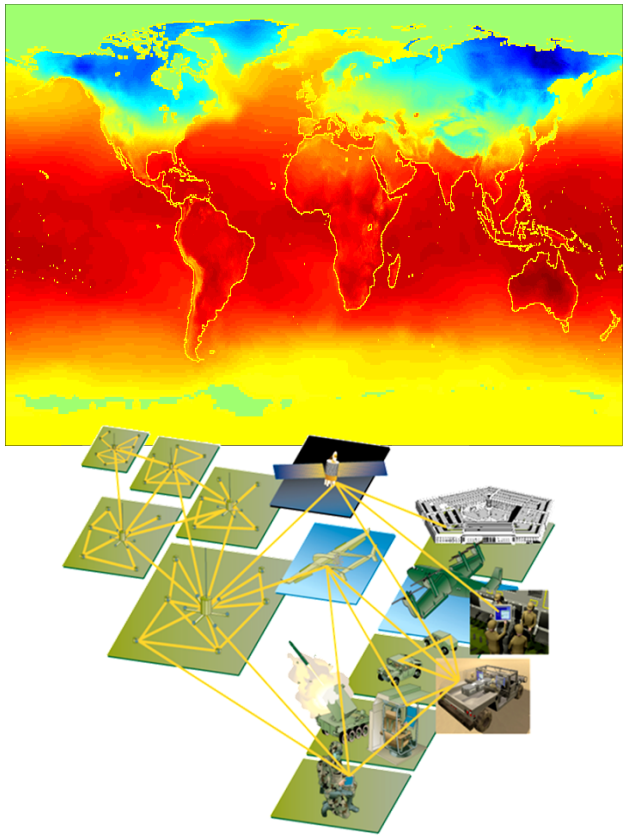
Βελτιστοποίηση Συστημάτων &
Υδροπληροφορική

Εξόρυξη Δεδομένων

Χρήστος Μακρόπουλος
Τομέας Υδατικών Πόρων και Περιβάλλοντος
Εθνικό Μετσόβιο Πολυτεχνείο

Γιατί εξόρυξη;

- Τεχνικές ανάλυσης πολύ **μεγάλων συνόλων** δεδομένων με σκοπό την **εξαγωγή νέα γνώσης**
- Δεδομένα συλλέγονται και αποθηκεύονται σε μεγάλες ταχύτητες (GB/Min)
  - Απομακρυσμένοι αισθητήρες
  - Next Generation Sequencing!
  - Προσομοιώσεις που παράγουν terabytes δεδομένων
- Η εξόρυξη δεδομένων μπορεί να βοηθήσει:
  - Στην **κατηγοριοποίηση** και την τμηματοποίηση των δεδομένων
  - Στην διατύπωση ***υποθέσεων***



Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities

BoonFei Tan[1], Charmaine Ng[2], Jean Pierre Nshimyimana[1,3,4], Lay Leng Loh[1,2], Karina Y.-H. Gin[2] and Janelle R. Thompson[1,5]*

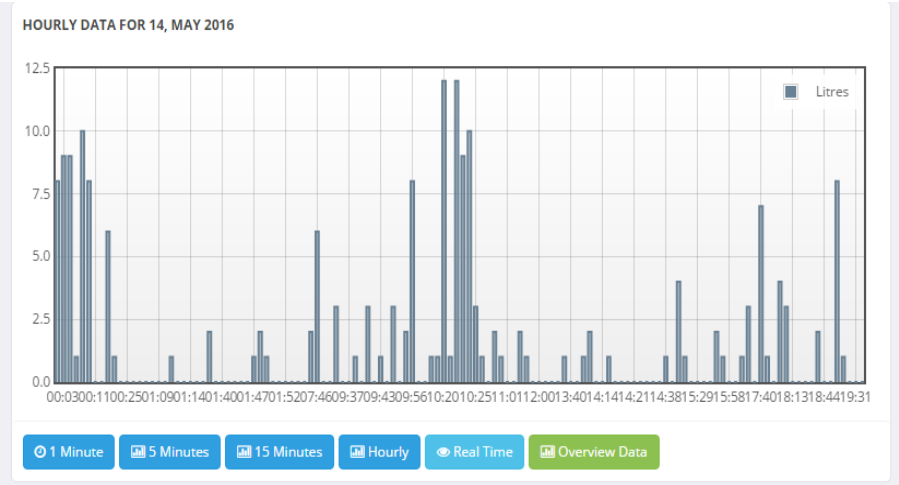
[1]Center for Environmental Sensing and Modelling, Singapore-MIT Alliance for Research and Technology Centre, Singapore, Singapore
[2]Department of Civil and Environmental Engineering, National University of Singapore, Singapore, Singapore
[3]Singapore Centre on Environmental Life Sciences Engineering, Nanyang Technological University, Singapore, Singapore
[4]School of Civil and Environmental Engineering, Nanyang Technological University, Singapore, Singapore
[5]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

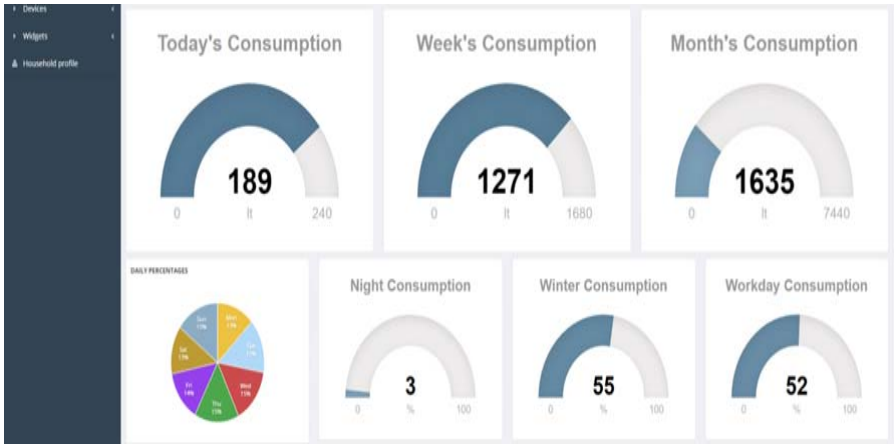
Water quality is an emergent property of a complex system comprised of interacting microbial populations and introduced microbial and chemical contaminants. Studies leveraging next-generation sequencing (NGS) technologies are providing new insights into the ecology of microbially mediated processes that influence fresh water quality such as algal blooms,

Όλο και περισσότερη σύνθεση πληροφορίας

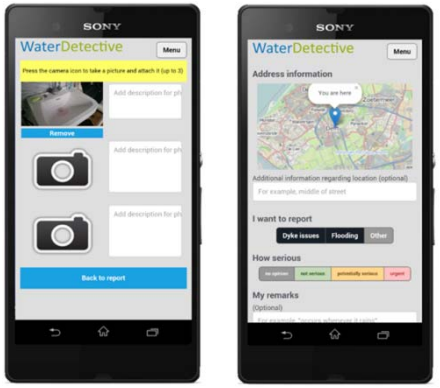
Παραδείγματα

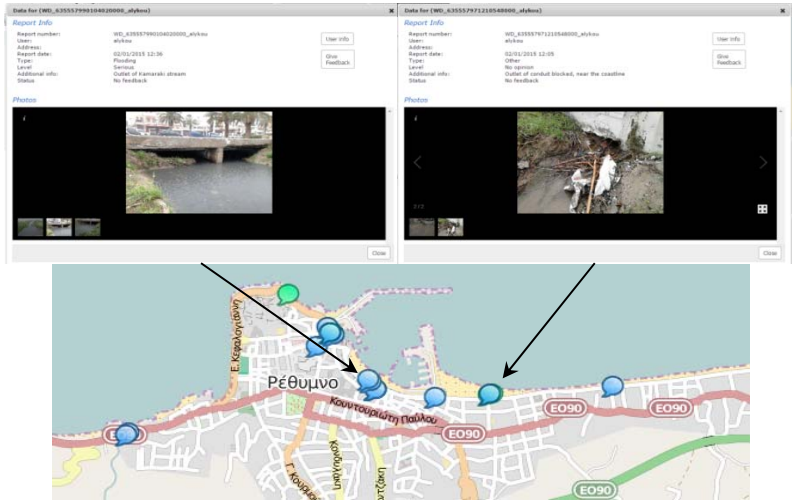
Έξυπνοι μετρητές κατανάλωσης νερού





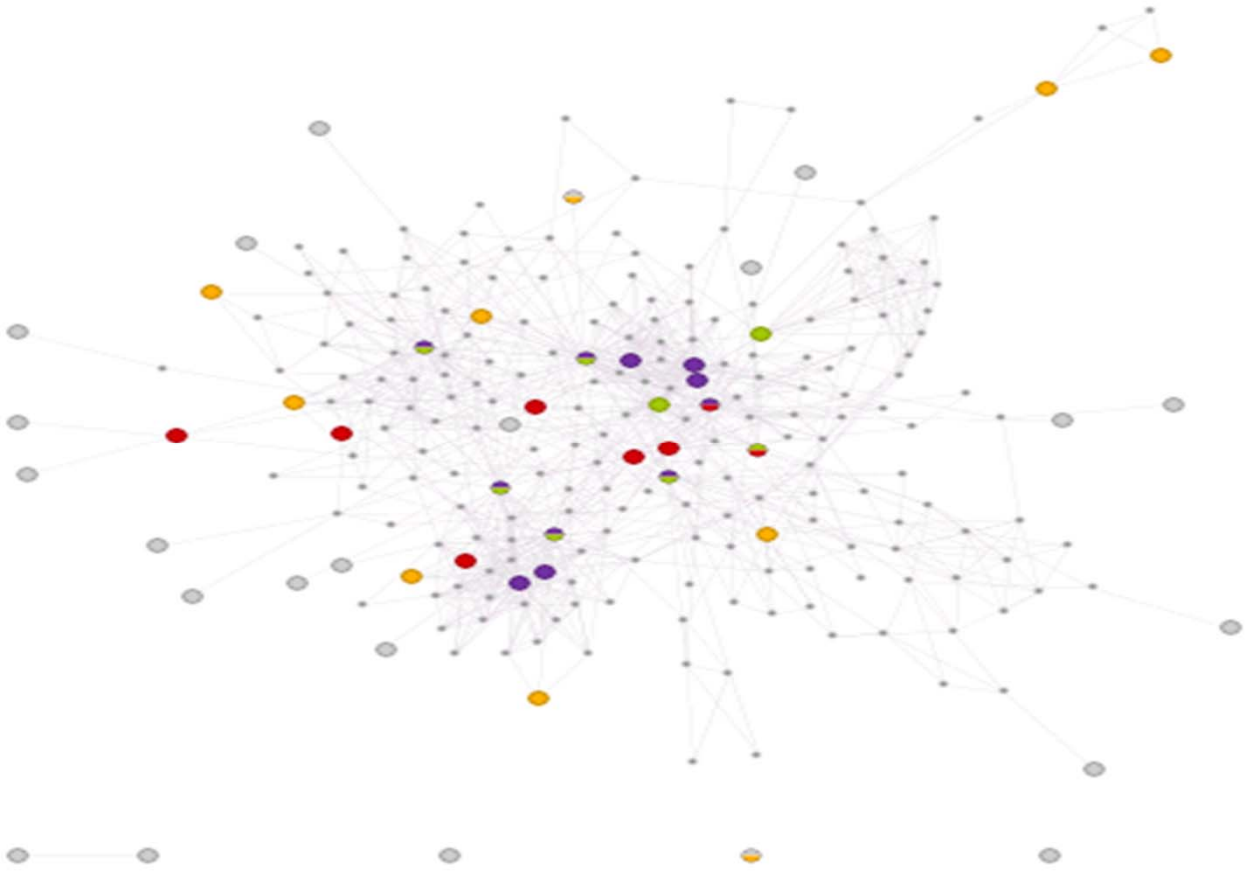
Συμμετοχικά παρατηρητήρια νερού

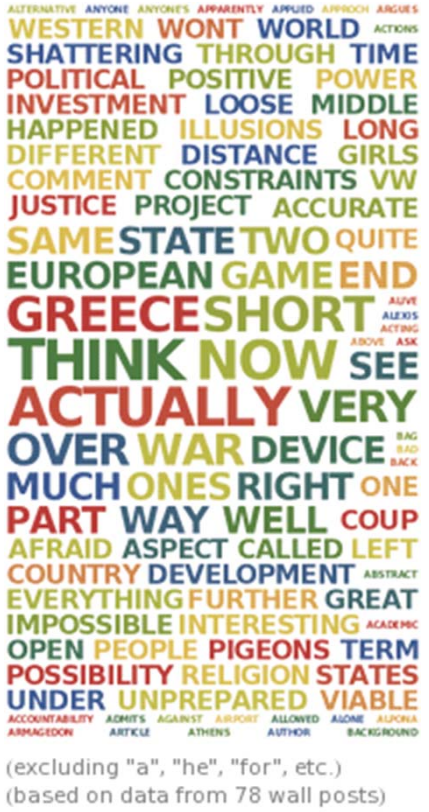




Συχνότητα μέτρησης 10 sec X 1-2 εκ. μετρητές..

Ανακαλύπτοντας γνώση κρυμμένη μέσα στα δεδομένα

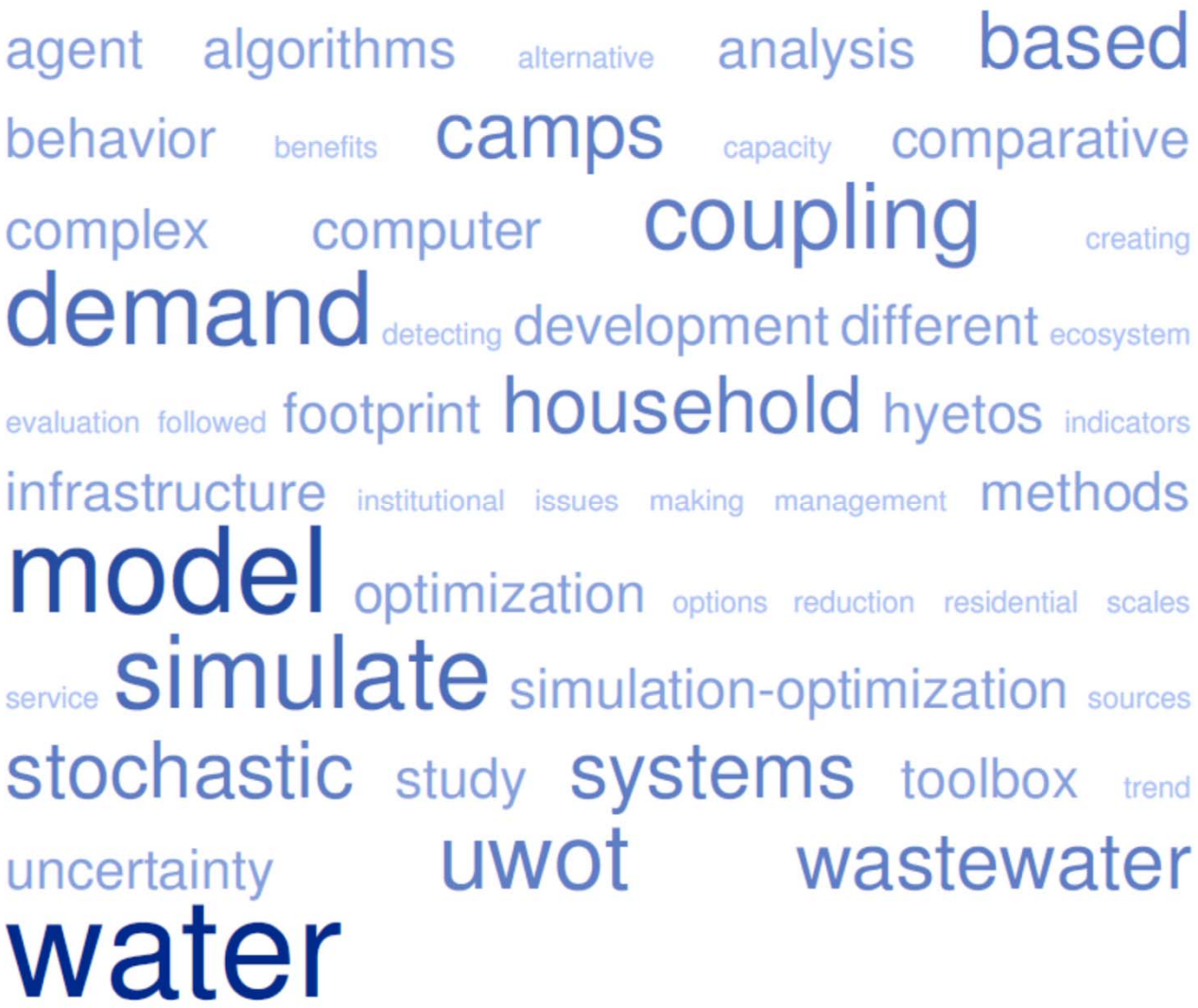




Wolfram alpha: facebook report

http://tagcrowd.com/

Το tag cloud των θεμάτων διπλωματικών στο μάθημα…

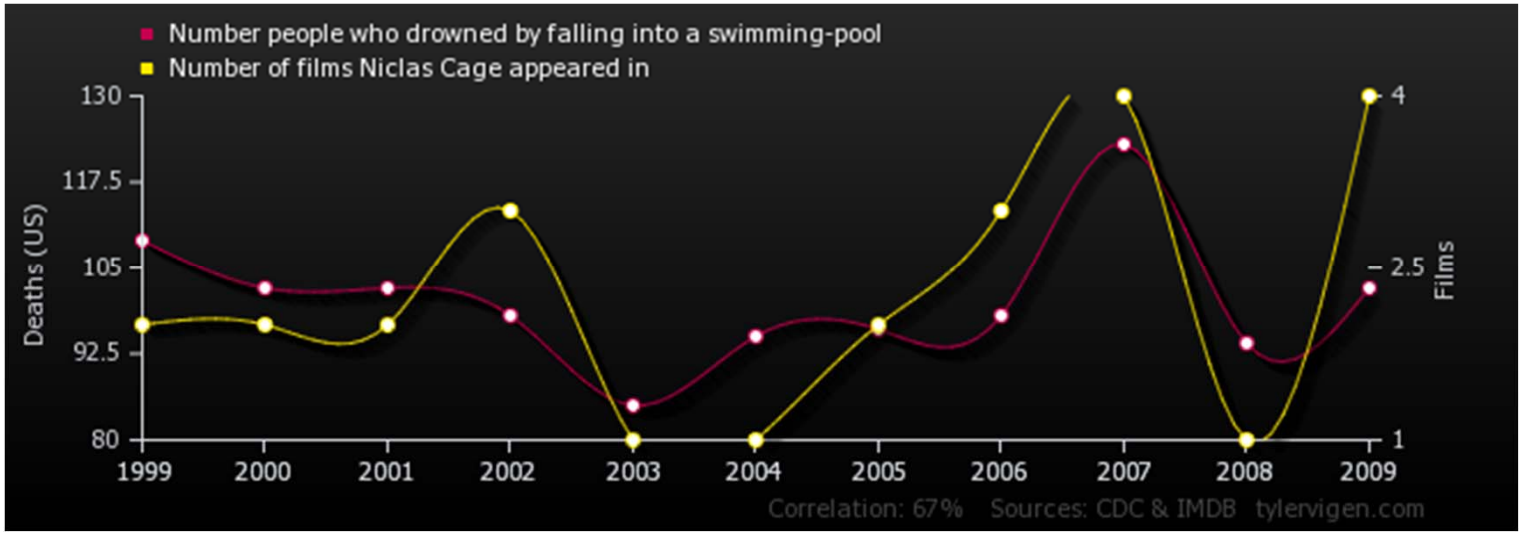


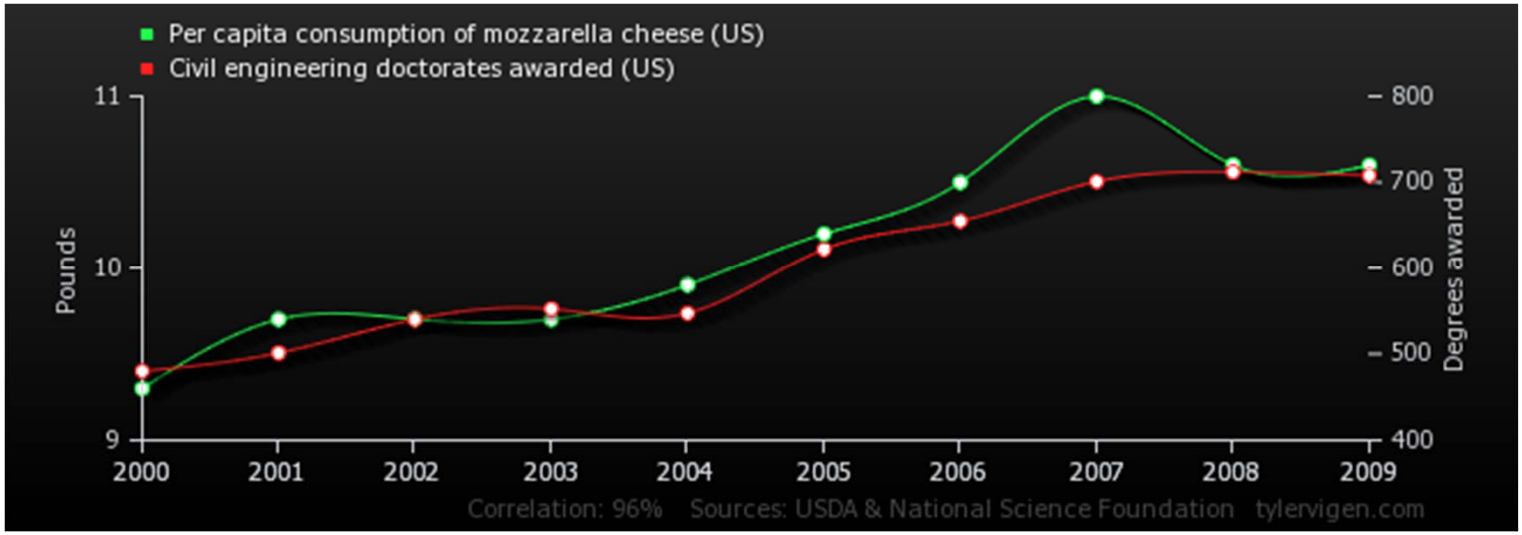
Ένας Ορισμός..

Για πολύ μεγάλα σύνολα δεδομένων :

- η διαδικασία **ανακάλυψης** (discovery) προτύπων (patterns) που πριν δεν ήταν γνωστά,
- η ανάλυση τους για να βρούμε **μη αναμενόμενες σχέσεις** ανάμεσα στα δεδομένα (causality?) και να τα συνοψίσουμε με νέους τρόπους (πχ. κανόνες) που είναι κατανοητοί και χρήσιμοι.
- Με άλλα λόγια ψάχνουμε για **συσχετίσεις** και σκεφτόμαστε αν αυτές μεταφράζονται σε **αιτιότητα.** Με προσοχή...

Προσοχή: συσχέτιση ≠ αιτιότητα



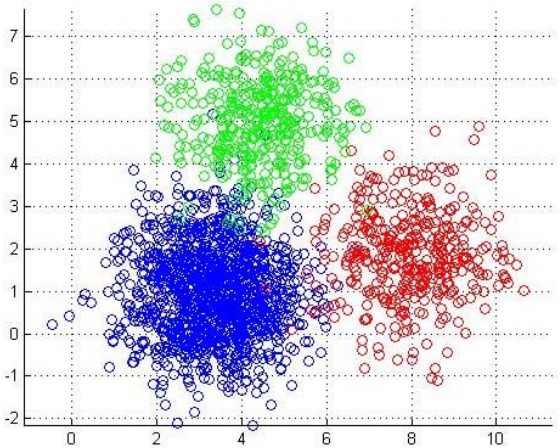


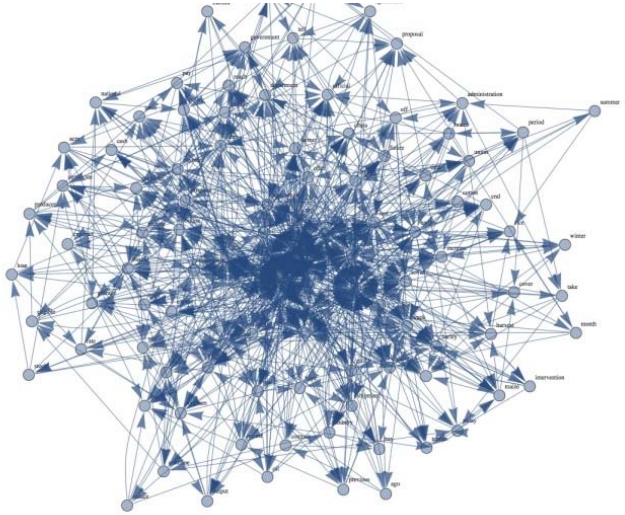
Μια διεπιφάνεια

- Όχι πραγματικά κάτι απόλυτα διαφορετικό.
- Ένας νέος τρόπος χρήσης γνωστών εργαλείων (ANN, Stats, BBF).
- Data-Driven

Εργαλεία…

Case Based Reasoning; Rule Extraction (Fuzzy Inference); Artificial Neural Networks; Support Vector Machines; Principal Component Analysis; Bayesian Inference, Bayesian Belief Networks; Statistics (incl. Stochastics); Multiple Liner Regression; Decision Trees; Factor Analysis; Survival Analysis; Graph-based mining; Symbolic Regression, Evolutionary Polynomial Regression, Wavelets but also *SBO-Optimisation*





Είδη/Μέθοδοι Εξόρυξης Δεδομένων

1. **Ταξινόμηση** - Classification: κατασκευή ενός μοντέλου που αντιστοιχεί ένα *στοιχείο* σε ένα σύνολο από (προκαθορισμένες) *κλάσεις* (mapping)
2. **Συσταδοποίηση** - Clustering: εύρεση ενός συνόλου από ομάδες με όμοια στοιχεία
3. **Εύρεση Συχνών Προτύπων**, Εξαρτήσεων και Συσχετίσεων – Dependencies and associations: εύρεση σημαντικών/συχνών εξαρτήσεων μεταξύ γνωρισμάτων
4. **Συνοψίσεις** - Summarization: εύρεση μιας συνοπτικής περιγραφής του συνόλου δεδομένων ή ενός υποσυνόλου του

Ταξινόμηση: Classification

- Δοθέντος ενός συνόλου από εγγραφές (**σύνολο εκπαίδευσης - training set**) με κάθε εγγραφή να έχει ένα σύνολο από γνωρίσματα (**attributes**), ένα από τα οποία να είναι η κλάση (**class**)
- Εύρεση ενός **μοντέλου** για το γνώρισμα της κλάσης ως συνάρτηση της τιμής των άλλων γνωρισμάτων.
- **Στόχος**: να αναθέτει το μοντέλο εγγραφές **που δεν έχουμε δει** σε μια από τις γνωστές μας **κλάσεις** με την μεγαλύτερη δυνατή ακρίβεια.
- **Διαδικασία**: Συνήθως, το σύνολο δεδομένων χωρίζεται σε ένα σύνολο *εκπαίδευσης* και σε ένα σύνολο *ελέγχου* – το πρώτο χρησιμοποιείται για την κατασκευή του μοντέλου και το δεύτερο για τον έλεγχο του.

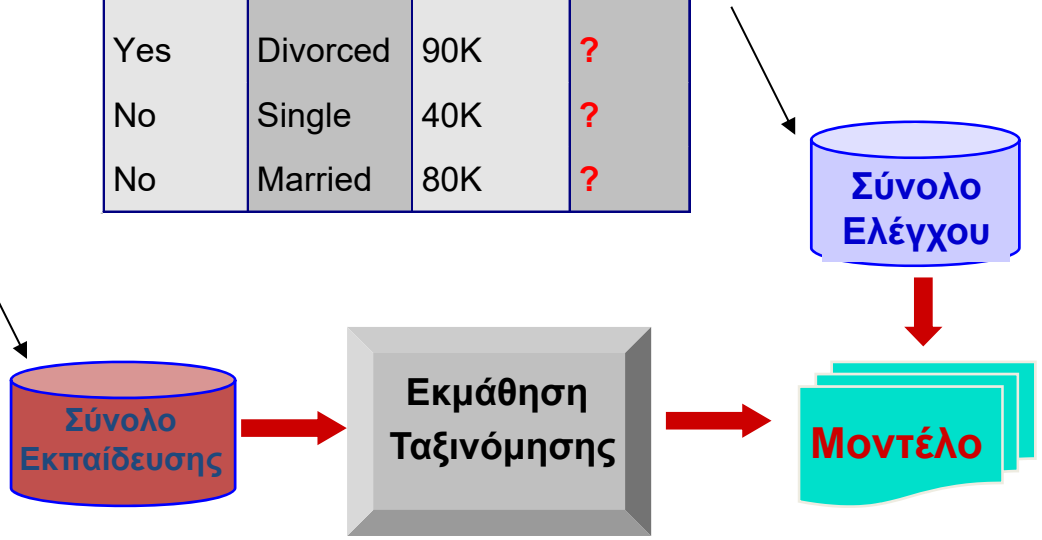
Παράδειγμα

Γνωρίσματα

ΚΛΑΣΗ

Tid	Home Owner	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

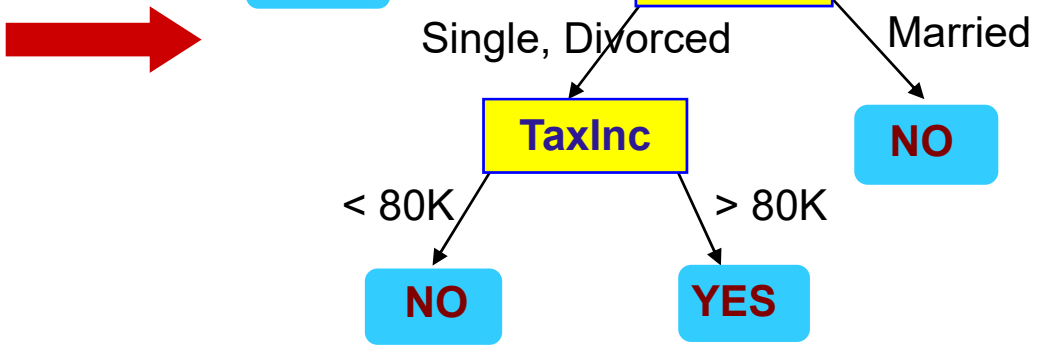
Home Owner	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Παράδειγμα Μοντέλου: **Δέντρο Απόφασης** – Decision tree

ΚΛΑΣΗ

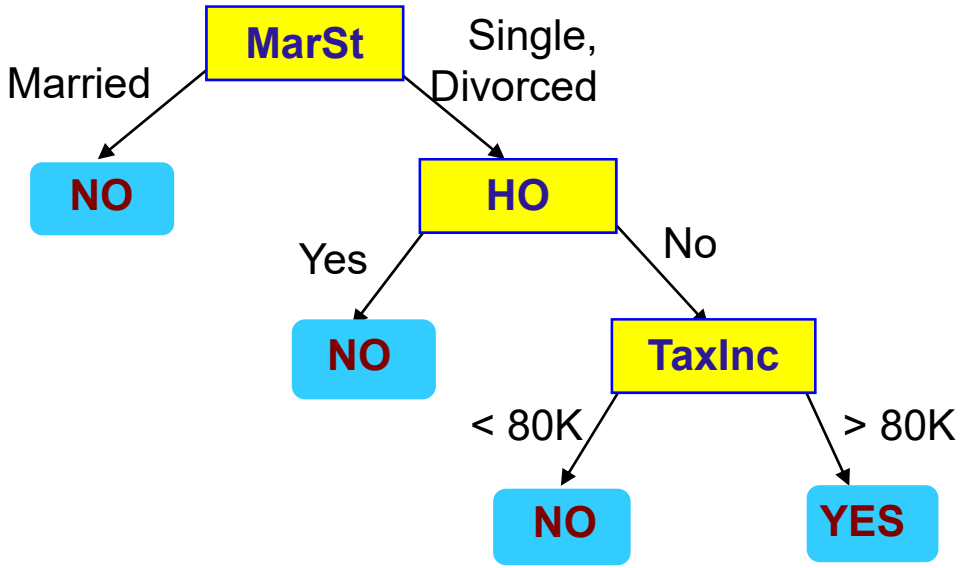
Tid	Home Owner	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Σύνολο Εκπαίδευσης

Εναλλακτικό παράδειγμα Μοντέλου: Δέντρο Απόφασης – Decision tree

Tid	Home Owner	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Για τα ίδια δεδομένα μπορεί να υπάρχουν παραπάνω από ένα δέντρα απόφασης (μοντέλα)

Παράδειγμα: Decision Trees

load *ionosphere*
t = **classregtree** (X, Y)
view(t)

Or

load *fisher**iris***

Ionosphere dataset from the UCI machine learning repository*. This **radar data** was collected by a system in Goose Bay, Labrador. A system of an array of high-frequency antennas.

"Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere.

- X is a 351x34 real-valued matrix of **predictors**.
- Y is a categorical **response**: "b" for bad radar returns and "g" for good radar returns.

*http://archive.ics.uci.edu/ml/datasets/Ionosphere

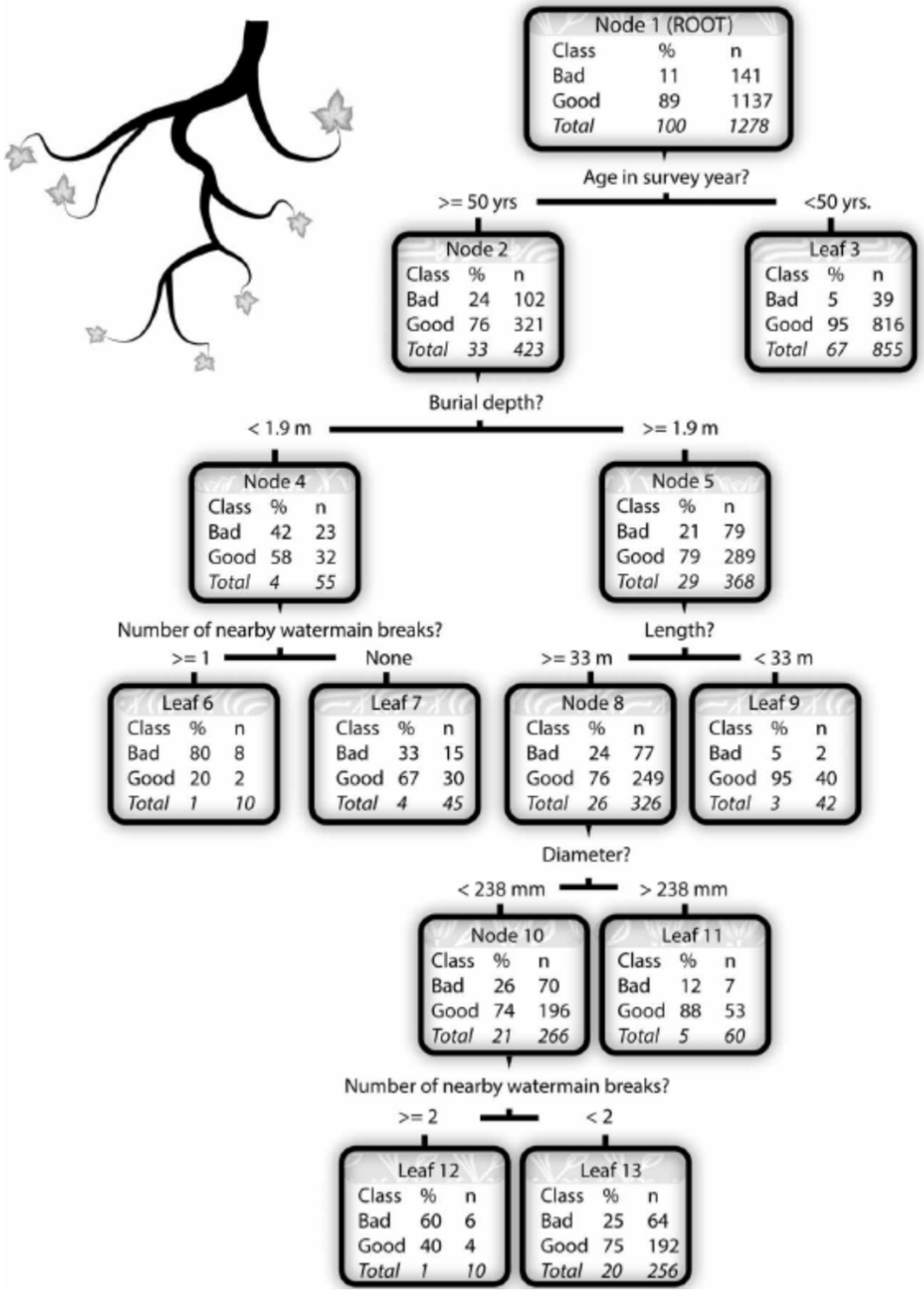
Παράδειγμα

- Φτιάξτε ένα δικό σας σετ (με δεδομένο decision tree) και βάλτε το matlab να το «ανακαλύψει»

Tip:
To generate n random numbers within a given [a b]:

x=a+ (b-a).*rand(n,1)

Harvey and McBean, 2014

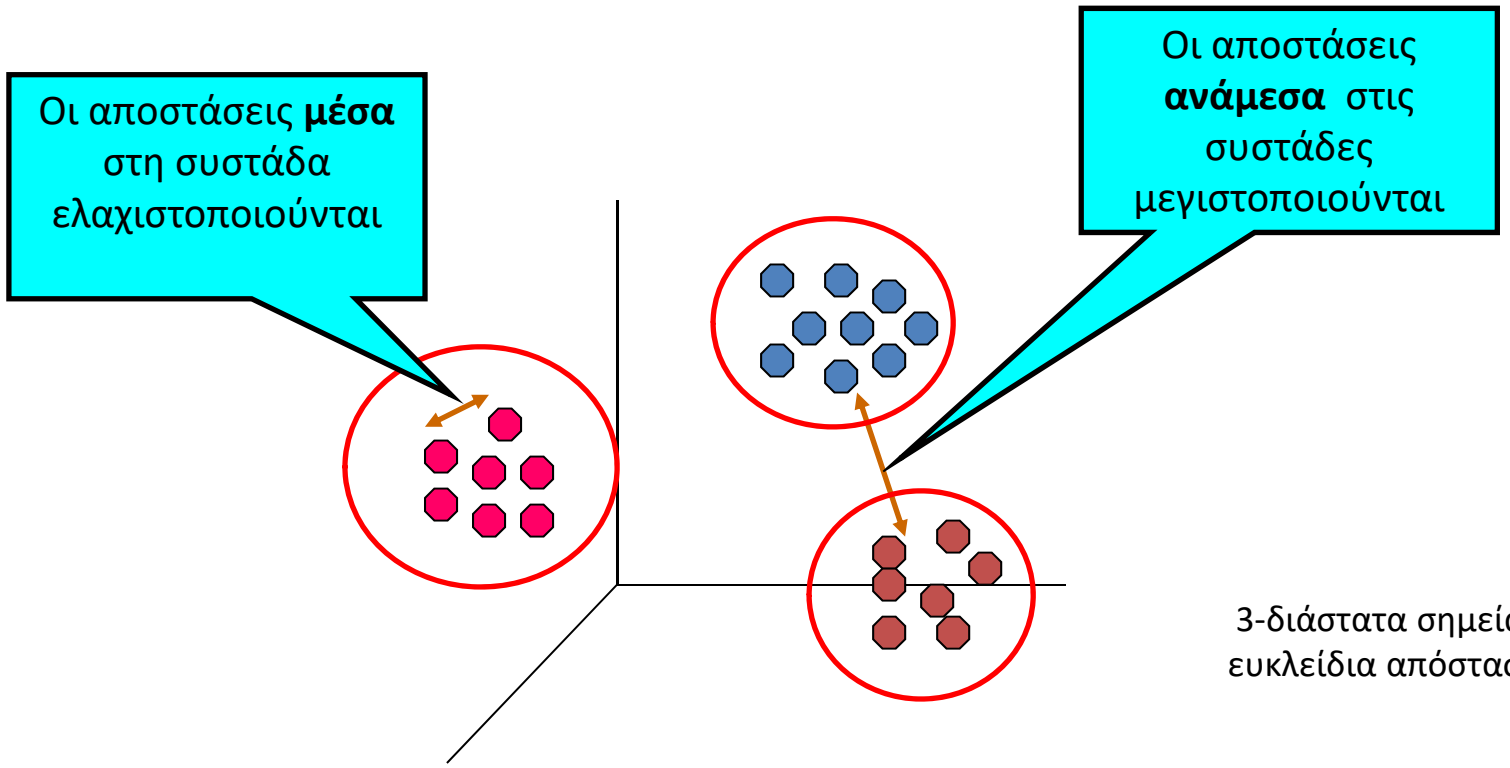


Παράδειγμα

- Ταξινόμηση με Νευρωνικά (nnstart>nprtool)
- Για τα δεδομένα *fisheriris* και πάλι
- Διαφορά*;* Έχει σημασία;

Συσταδοποίηση (clustering)

Εύρεση “συστάδων” αντικειμένων έτσι ώστε τα αντικείμενα σε κάθε ομάδα να είναι όμοια (ή να σχετίζονται) και διαφορετικά (ή μη σχετιζόμενα) από τα αντικείμενα των άλλων ομάδων



3-διάστατα σημεία, ευκλείδια απόσταση

Είδη Συστάδων

- Καλώς διαχωρισμένες συστάδες
- Συστάδες βασισμένες σε κέντρο
- Συνεχής (contiguous) συστάδες
- Συστάδες βασισμένες σε πυκνότητα
- Βασισμένα σε ιδιότητες ή έννοιες
- Περιγράφονται από μια αντικειμενική συνάρτηση (Objective Function)

Πότε μια συσταδοποίηση είναι “καλή”

- **Μεγάλη** *ομοιότητα* **εντός** της συστάδας και
- **Μικρή** *ομοιότητα* **ανάμεσα** στις συστάδες
- … τι είναι όμως ομοιότητα;

«Ομοιότητα»

- Η εγγύτητα των σημείων υπολογίζεται με βάση κάποια ***απόσταση*** που εξαρτάται από το είδος των σημείων (πχ. *Ευκλείδεια απόσταση)*
- Επειδή η απόσταση υπολογίζεται συχνά ο υπολογισμός της πρέπει να είναι σχετικά *απλός*
- Το κεντρικό σημείο είναι (συνήθως) το μέσο των σημείων της συστάδας (το οποίο μπορεί να *μην είναι ένα από τα δεδομένα εισόδου*)

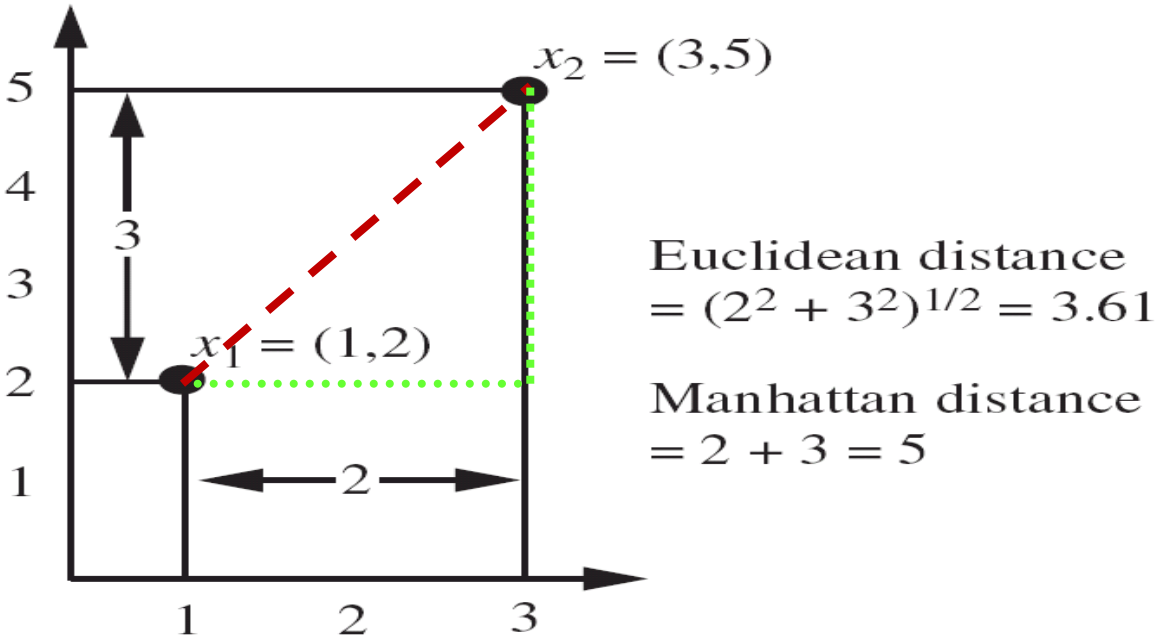
Κριτήρια Ομοιότητας: Απόσταση

Ευκλείδεια απόσταση

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{in}-x_{jn}|^2)}$$

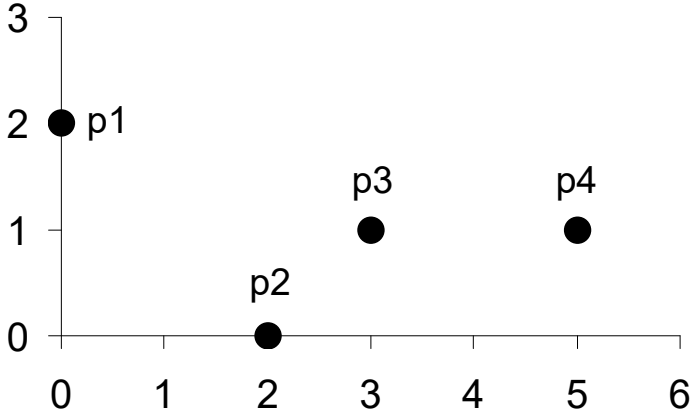
Manhattan

$$L_1(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+...+|x_{in}-x_{jn}|$$



Ορισμός Απόστασης

Παράδειγμα



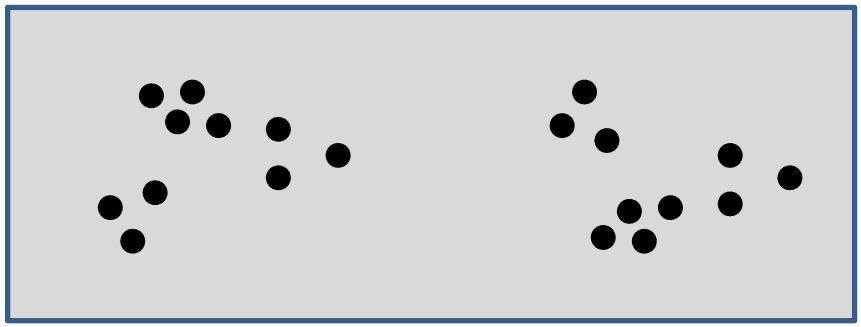
Πίνακας Δεδομένων

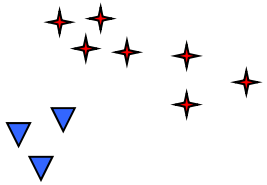
point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

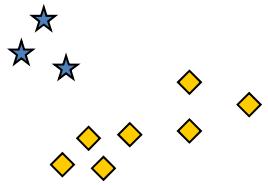
	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Πίνακας Απόστασης

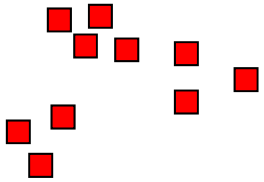
Πόσες Ομάδες?

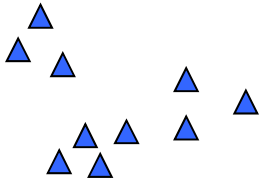




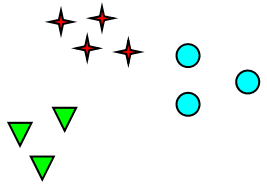


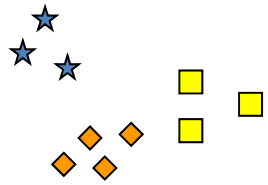
4 ομάδες





2 ομάδες

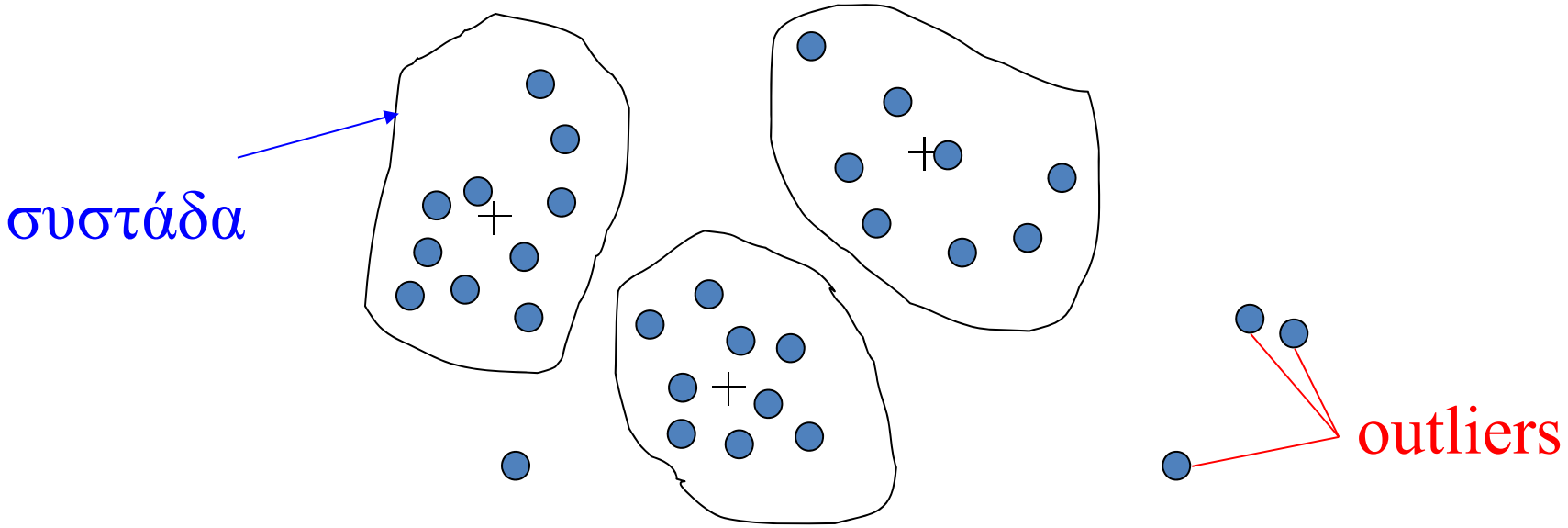




6 ομάδες

Ακραία Σημεία & Θόρυβος

Outlier (ακραίο σημείο) τιμές που είναι εξαιρέσεις ως προς τα συνηθισμένες ή αναμενόμενες τιμές (domain knowledge!)



Συστάδες βασισμένες σε κέντρο ή πρότυπο

Μια συστάδα είναι ένα σύνολο από στοιχεία τέτοιο ώστε κάθε στοιχείο της ομάδας είναι πιο κοντά σε (ή πιο όμοιο με) το «κέντρο» ή πρότυπο της ομάδας από ότι με το κέντρο οποιασδήποτε άλλης ομάδας.

Το κέντρο της ομάδας είναι συχνά

- centroid, ο μέσος όρος των σημείων της συστάδας, ή
- medoid, το πιο «αντιπροσωπευτικό» σημείο της συστάδας

Η βασική διαδικασία clustering: K-means

- Διαχωριστικός αλγόριθμος
- Κάθε συστάδα **συσχετίζεται** με ένα κεντρικό σημείο (centroid)
- Κάθε σημείο **ανατίθεται** στη συστάδα με το κοντινότερο κεντρικό σημείο
- Ο αριθμός των ομάδων, Κ, είναι **είσοδος** στον αλγόριθμο

K-means: Βασικός Αλγόριθμος

1: Επιλογή *K* σημείων ως τα αρχικά κεντρικά σημεία (*συνήθως τυχαία*)

2: **Repeat**

3: Ανάθεση όλων των αρχικών σημείων στο *κοντινότερο* τους από τα *K* κεντρικά σημεία

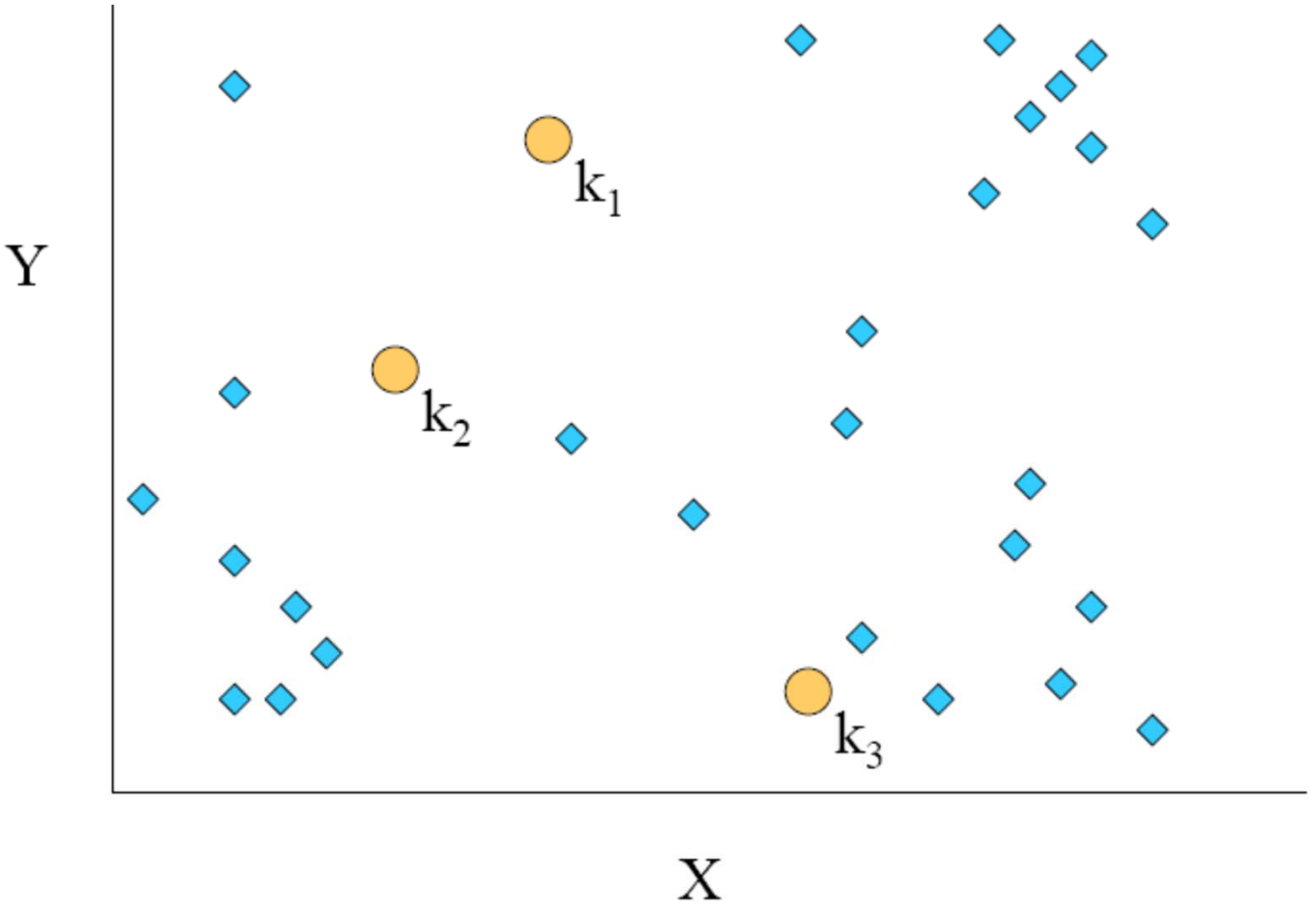
4: Επαν-υπολογισμός του *κεντρικού σημείου* κάθε συστάδας

5: **Until** τα κεντρικά σημεία να μην αλλάζουν

K-means: Βασικός Αλγόριθμος

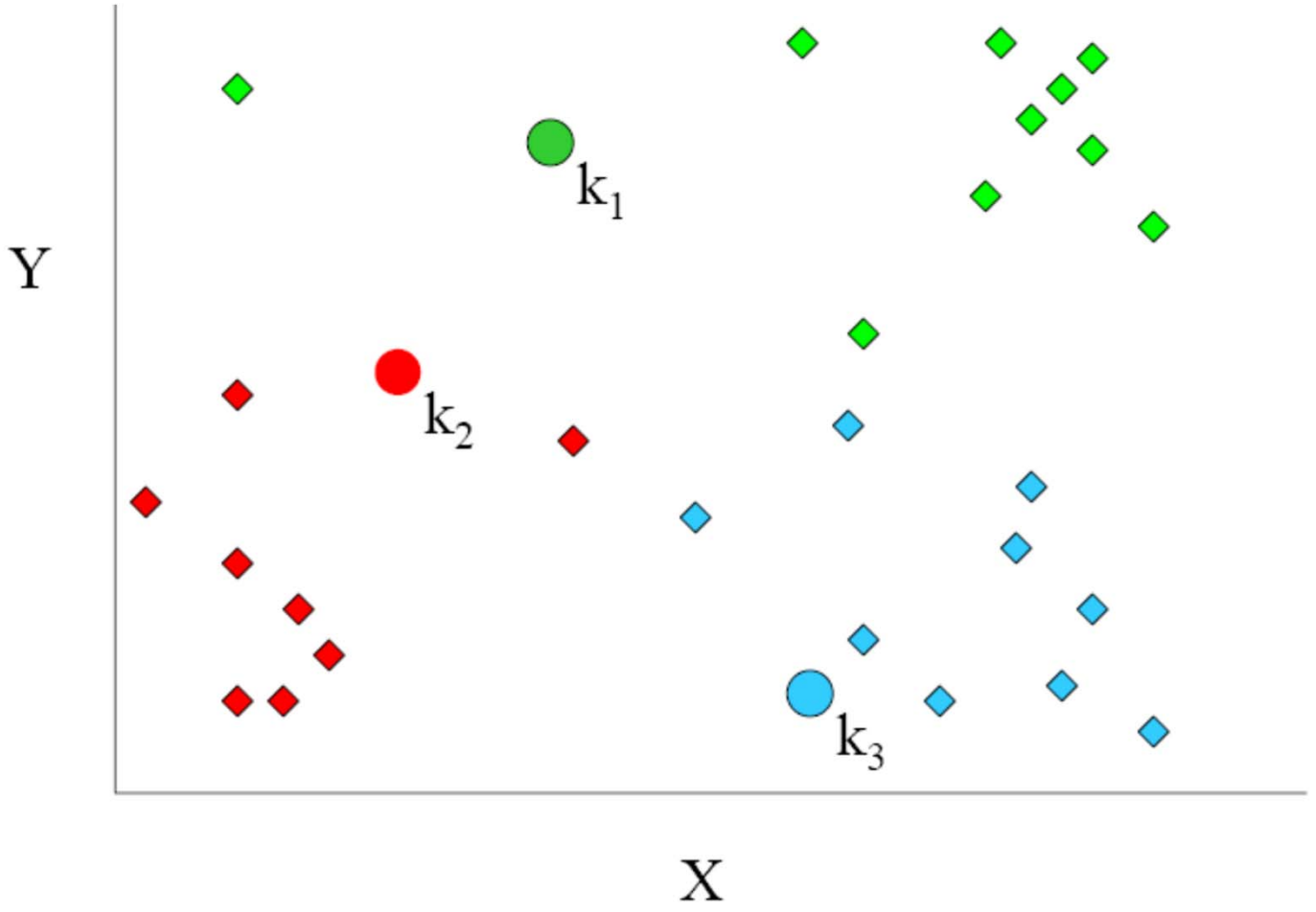
K = 3 συστάδες

Αρχικά σημεία k_1, k_2, k_3



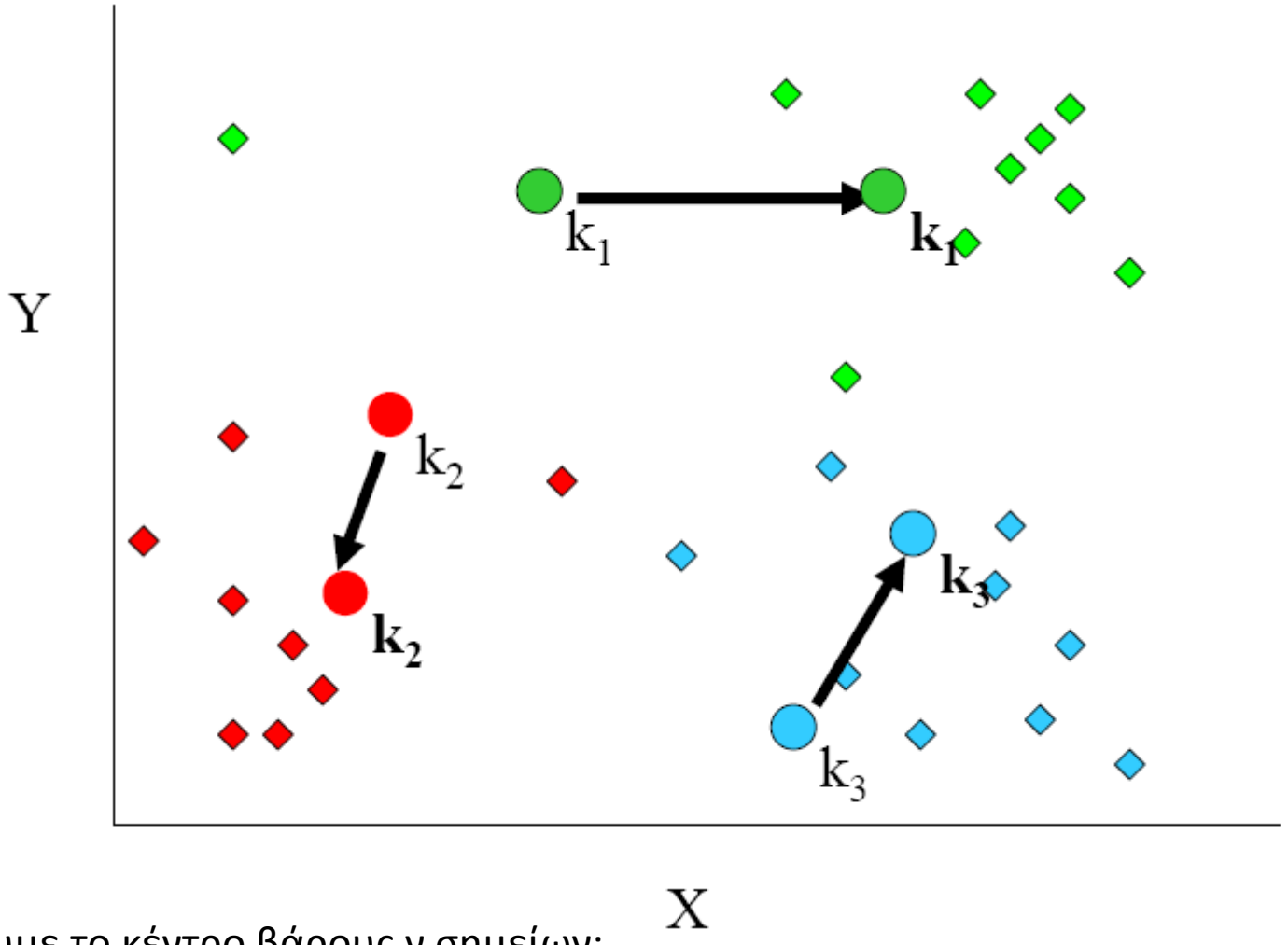
K-means: Βασικός Αλγόριθμος

Τα σημεία ανατίθενται στο πιο γειτονικό
από τα 3 αρχικά σημεία



K-means: Βασικός Αλγόριθμος

Επανυπολογισμός του κέντρου (κέντρου βάρους) κάθε συστάδας (cluster)

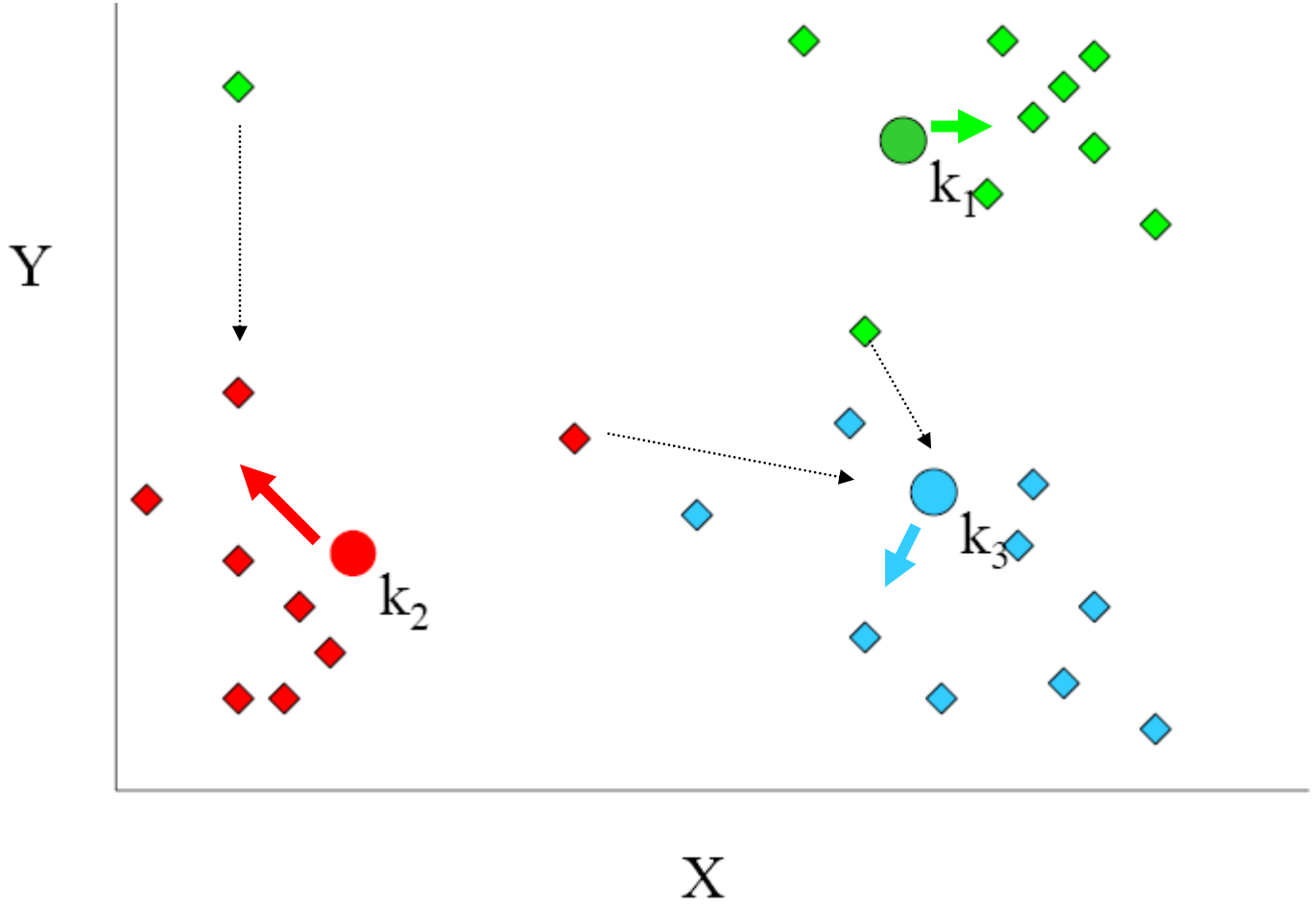


Πως υπολογίζουμε το κέντρο βάρους ν σημείων;

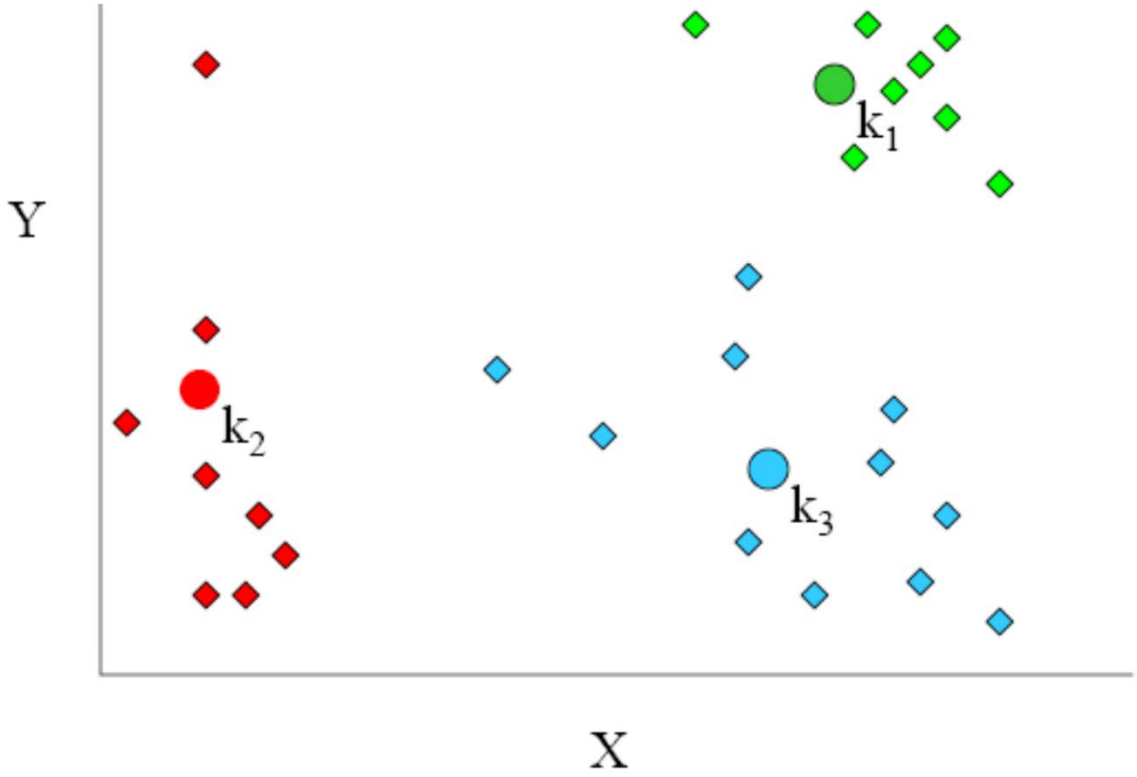
K-means: Βασικός Αλγόριθμος

Νέα ανάθεση των σημείων

Νέα κέντρα βάρους



K-means: Βασικός Αλγόριθμος



Αν δεν αλλάζει τίποτα -> ΤΕΛΟΣ

K-means: Επιλογή αρχικών σημείων

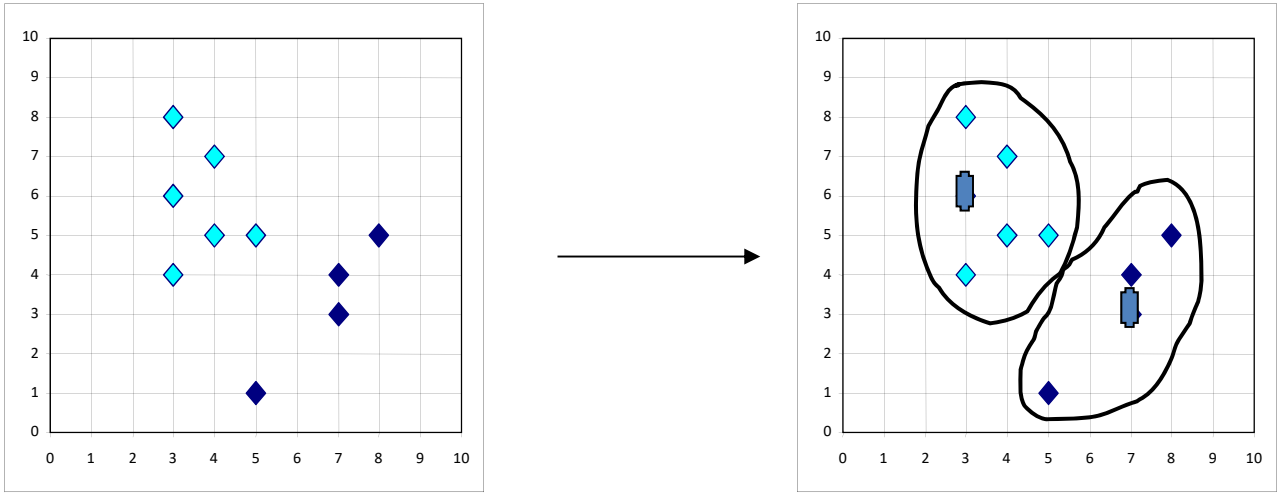
Αν υπάρχουν Κ «πραγματικές συστάδες» η πιθανότητα να επιλέξουμε ένα κέντρο από κάθε συστάδα είναι μικρή, συγκεκριμένα αν όλες οι συστάδες έχουν το ίδιο μέγεθος n, τότε:

$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K! n^K}{(Kn)^K} = \frac{K!}{K^K}$$

Για παράδειγμα, αν Κ = 10, η πιθανότητα είναι = 10!/1010 = 0.00036

K-medoid

- 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)
- Μπορεί να εφαρμοστεί σε δεδομένα οποιουδήποτε τύπου (πχ και για κατηγορικά δεδομένα)



K-means: Βασικός Αλγόριθμος

Παράδειγμα

2 4 10 12 3 20 30 11 15

Έστω **k = 2**, και αρχικά επιλέγουμε το 3 και το 4

K-Means Clustering στο Matlab

Syntax:

IDX = kmeans(X,k) ή

[IDX,C] = kmeans(X,k)

kmeans uses a two-phase iterative algorithm to minimize the sum of point-to-centroid distances, summed over all k clusters:

- **The first phase** uses *batch updates*, where each iteration consists of reassigning points to their nearest cluster centroid, all at once, followed by recalculation of cluster centroids.
- This phase occasionally does not converge to solution that is a local minimum, that is, a partition of the data where moving any single point to a different cluster increases the total sum of distances. This is more likely for small data sets.
- The batch phase is fast, but potentially only approximates a solution as a starting point for the second phase.

K-Means Clustering στο Matlab

- **The second phase** uses *online updates*, where points are individually reassigned if doing so will reduce the sum of distances, and cluster centroids are recomputed after each reassignment. Each iteration during the second phase consists of one pass though all the points.
- The second phase will converge to a local minimum, although there may be other local minima with lower total sum of distances.
- The problem of finding the global minimum can only be solved in general by an exhaustive (or clever, or lucky) choice of starting points, but using several replicates with random starting points typically results in a solution that is a global minimum.

Παράδειγμα

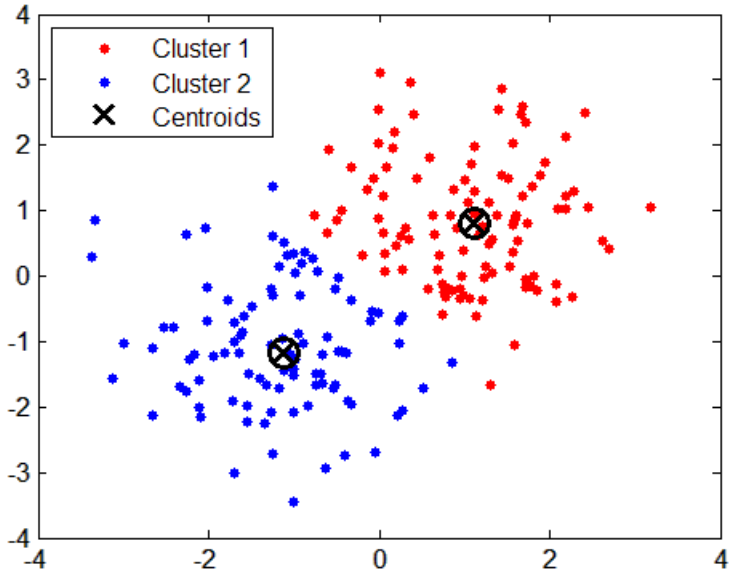
- x=[1,2,3,4,10,11,12,13]’
- [idx,ctrs]=kmeans(x,2)

Παράδειγμα

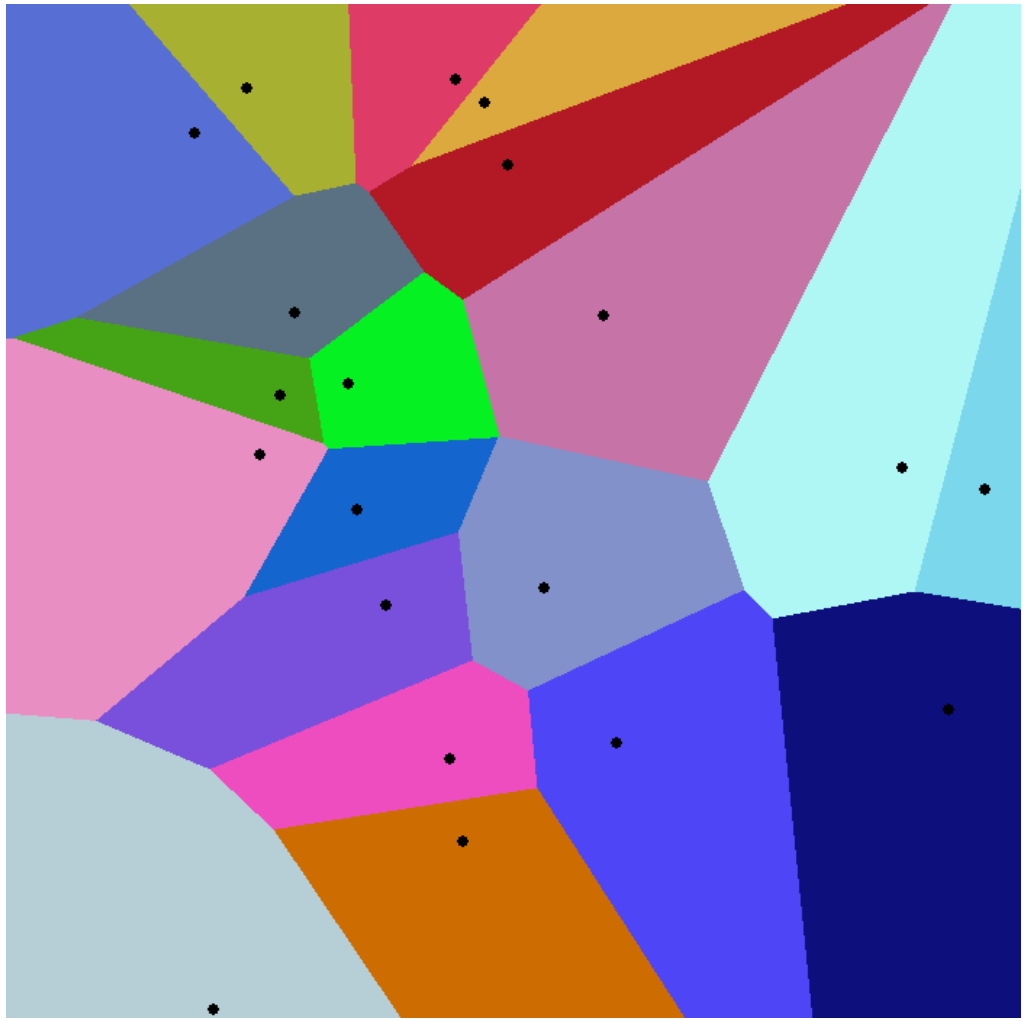
```
X = [randn(100,2)+ones(100,2);...
     randn(100,2)-ones(100,2)];
opts = statset('Display','final');

[idx,ctrs] = kmeans(X,2,...
                    'Distance','city',...
                    'Replicates',5,...
                    'Options',opts);
5 iterations, total sum of distances = 284.671
4 iterations, total sum of distances = 284.671
4 iterations, total sum of distances = 284.671
3 iterations, total sum of distances = 284.671
3 iterations, total sum of distances = 284.671

plot(X(idx==1,1),X(idx==1,2),'r.','MarkerSize',12)
hold on
plot(X(idx==2,1),X(idx==2,2),'b.','MarkerSize',12)
plot(ctrs(:,1),ctrs(:,2),'kx',...
     'MarkerSize',12,'LineWidth',2)
plot(ctrs(:,1),ctrs(:,2),'ko',...
     'MarkerSize',12,'LineWidth',2)
legend('Cluster 1','Cluster 2','Centroids',...
       'Location','NW')
```



Διαγράμματα Voronoi



Υδρολογία – σταθμησμένες μετρήσεις σταθμών

Βιβλιογραφία

Οι διαφάνειες είναι βασισμένες στους:

- P.-N. Tan, M.Steinbach, V. Kumar, «Introduction to Data Mining», Addison Wesley, 2006
- Rogkakos, D., Department of CS, University of Epirus, Lecture Notes
- Matlab 2011a Online Documentation, www.mathworks.com