Αναγνώριση Προτύπων

Ομαδοποίηση, Ταξινόμηση, Παλινδρόμηση

Ανδρέας Λ. Συμεωνίδης

Αν. Καθηγητής

Τμήμα Ηλεκτρολόγων Μηχ/κών

&

Μηχ/κών Υπολογιστών, Α.Π.Θ.

Email: asymeon@eng.auth.gr

Βασικές έννοιες, Είδη και μέθοδοι

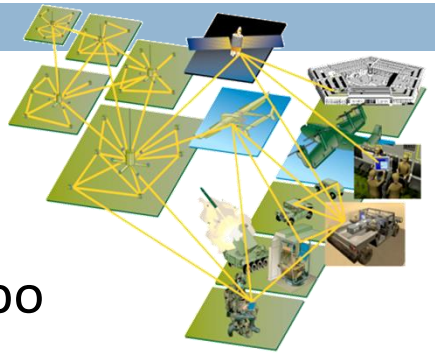
Αναγνώριση προτύπων: Η μια σκοπιά…

- Τεράστιοι όγκοι δεδομένων συγκεντρώνονται καθημερινά
  - Δεδομένα διαδικτύου, e-commerce
  - Αγορές σε πολυκαταστήματα/ supermarkets
  - Τραπεζικές συναλλαγές, συναλλαγές πιστωτικών καρτών

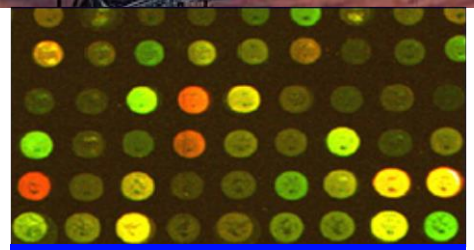
- Υπολογιστική ισχύς πλέον διαθέσιμη και φθηνή
- Μεγάλος ανταγωνισμός
  - Παροχή καλύτερης και *προσωποποιημένης* υπηρεσίας (π.χ. Customer Relationship Management)

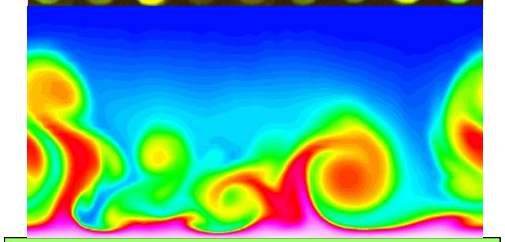
→

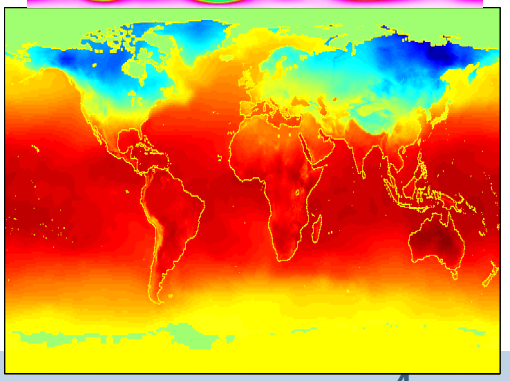
Αναγνώριση προτύπων: και η άλλη σκοπιά…



- Δεδομένα συλλέγονται και αποθηκεύονται με τεράστιους ρυθμούς (GB/hour)
  - Οι αισθητήρες σε ένα δορυφόρο
  - Τα τηλεσκόπια που σαρώνουν τον ουρανό
  - Τα microarrays που παράγουν δεδομένα έκφρασης πρωτεϊνών
  - Οι προσομοιώσεις που παράγουν terabytes δεδομένων
- Οι κλασικές τεχνικές ανάλυσης δεδομένων δεν είναι δυνατόν να εφαρμοστούν
- Η αναγνώριση προτύπων μπορεί να βοηθήσει τους ερευνητές να:
  - Ταξινομήσουν και να διαχωρίσουν τα δεδομένα
  - Σχηματίσουν υποθέσεις (Hypothesis Formation)







Στόχος της Αναγνώρισης Προτύπων

- Λήψη αποφάσεων σχετικά με τα πρότυπα.
- Οπτικό παράδειγμα – είναι το άτομο αυτό χαρούμενο ή λυπημένο;
- Φωνητικό παράδειγμα – είπε ο ομιλητής “Ναι” ή “Όχι”;
- Παράδειγμα Φυσικής – είναι αυτό άτομο ή μόριο;

Διαδικασία Ανακάλυψης Γνώσης

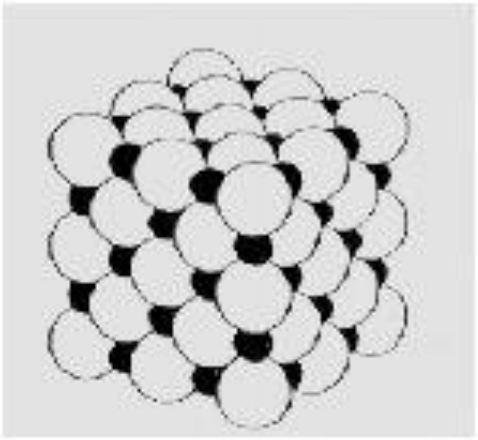
Τί είναι η ΔΑΓ;

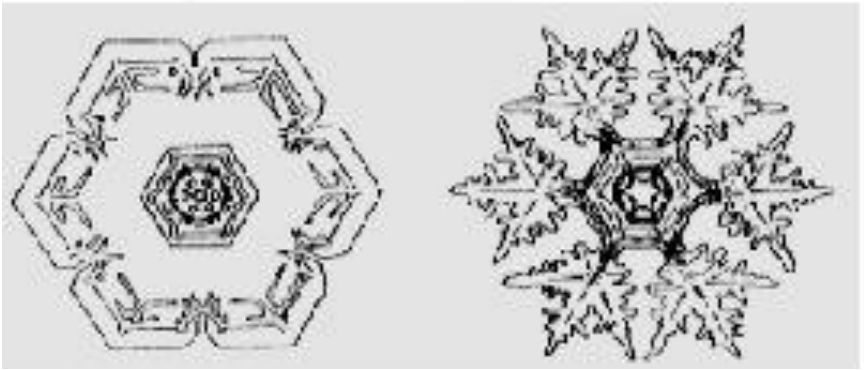
Διερεύνηση και ανάλυση των δεδομένων, με αυτοματοποιημένες και ήμι-αυτοματοποιημένες διαδικασίες, με στόχο την ανακάλυψη χρήσιμων προτύπων

δεδομένα, με στόχο την ανακάλυψη **προτύπων** μέσα σε αυτά.

Τι είναι τα πρότυπα;

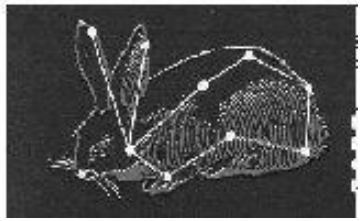
- Οι νόμοι της Φυσικής και της Χημείας δημιουργούν πρότυπα.





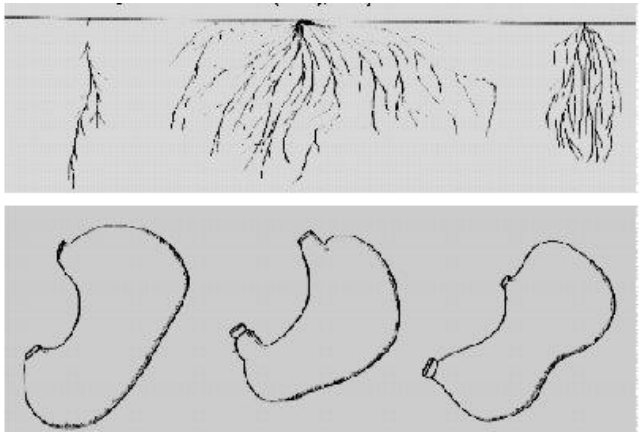
Πρότυπα στην Αστρονομία

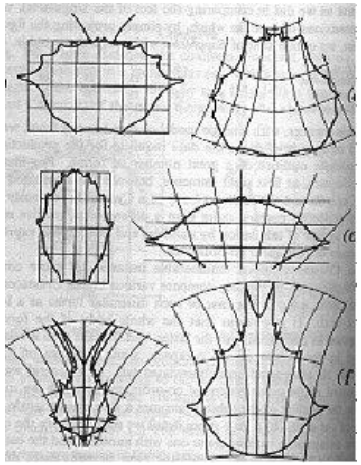
- Οι άνθρωποι συνηθίζουν να βλέπουν πρότυπα παντού.

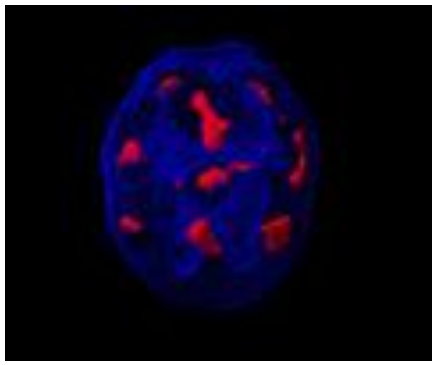


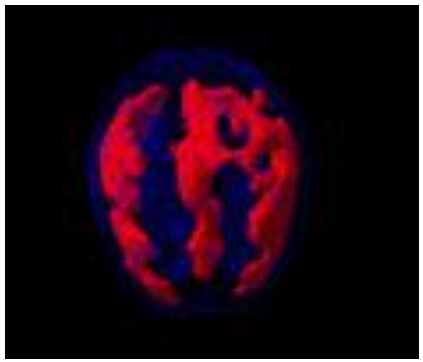
Πρότυπα στη Βιολογία

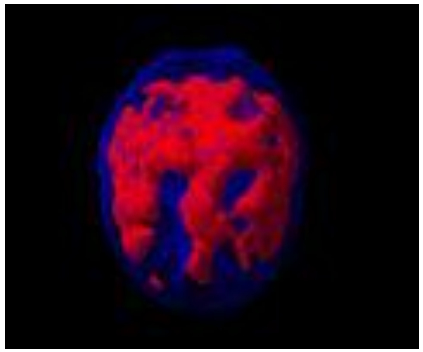
- Εφαρμογές: Βιομετρικά, Υπολογιστική Ανατομία, Εγκεφαλική δραστηριότητα.





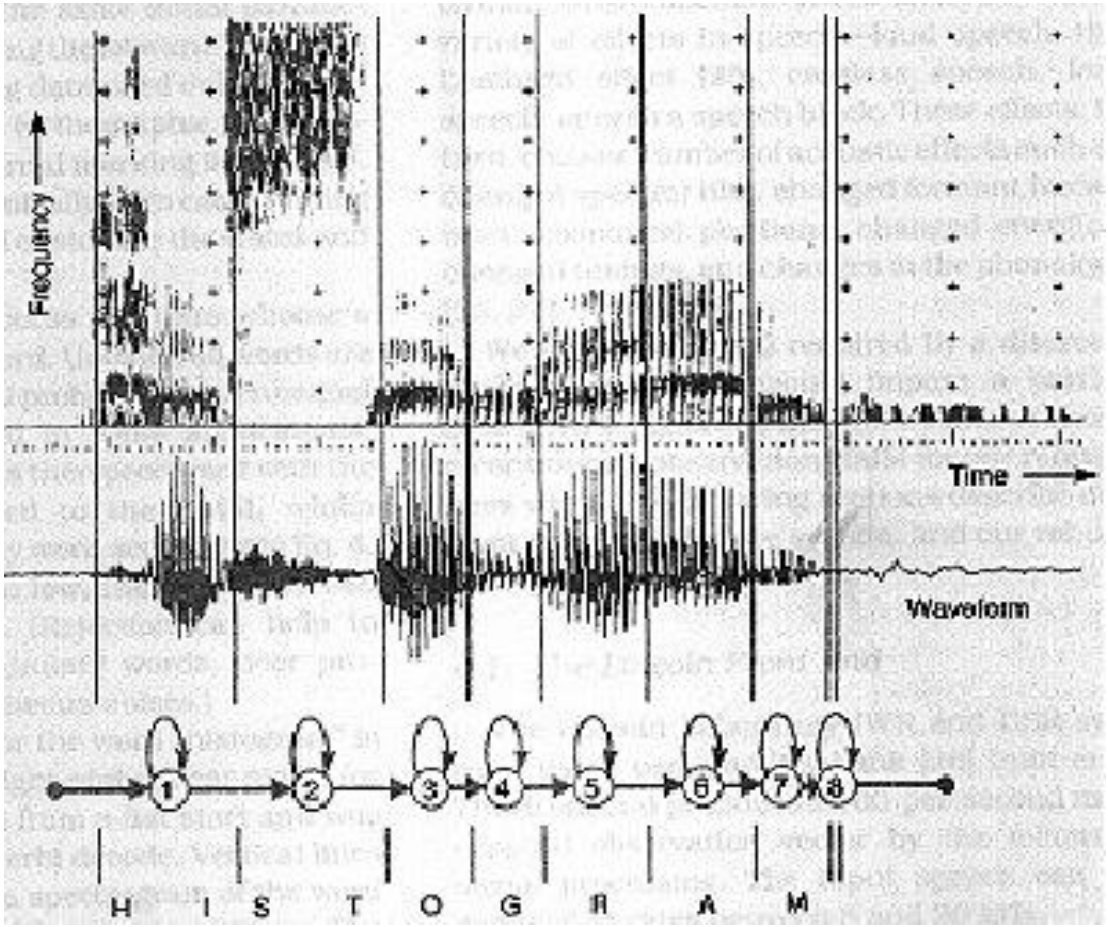






Πρότυπα φωνής

- Ακουστικά σήματα

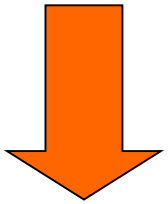


Διαφοροποίηση των προτύπων

- Τα πρότυπα διαφοροποιούνται με την έκφραση, τον φωτισμό, κτλ.

Είναι όλα τα πρότυπα ενδιαφέροντα;

Ένα σύστημα αναγνώρισης προτύπων μπορεί να γεννήσει χιλιάδες πρότυπα. Δεν είναι όλα ενδιαφέροντα!!!



Σημασία

Ένα πρότυπο είναι **ενδιαφέρον** είναι εύκολα κατανοητό, ισχύον σε ένα νέο σετ δεδομένων, πιθανόν χρήσιμο, καινούριο, ή επικυρώνει μια υπόθεση που επιθυμεί να ελέγξει ο χρήστης.

Ένα ενδιαφέρον πρότυπο αναπαριστά **γνώση**

Μέθοδοι Αναγνώρισης Προτύπων

- **Μέθοδοι Πρόβλεψης**
  - Χρήση κάποιων μεταβλητών για την πρόβλεψη άγνωστων ή μελλοντικών τιμών από άλλες μεταβλητές.
- **Μέθοδοι Περιγραφής**
  - Εύρεση προτύπων που περιγράφουν τα δεδομένα.

Εργασίες Αναγνώρισης Προτύπων

- Ταξινόμηση
- Ομαδοποίηση
- Παλινδρόμηση
- Αναγνώριση Αποκλίσεων

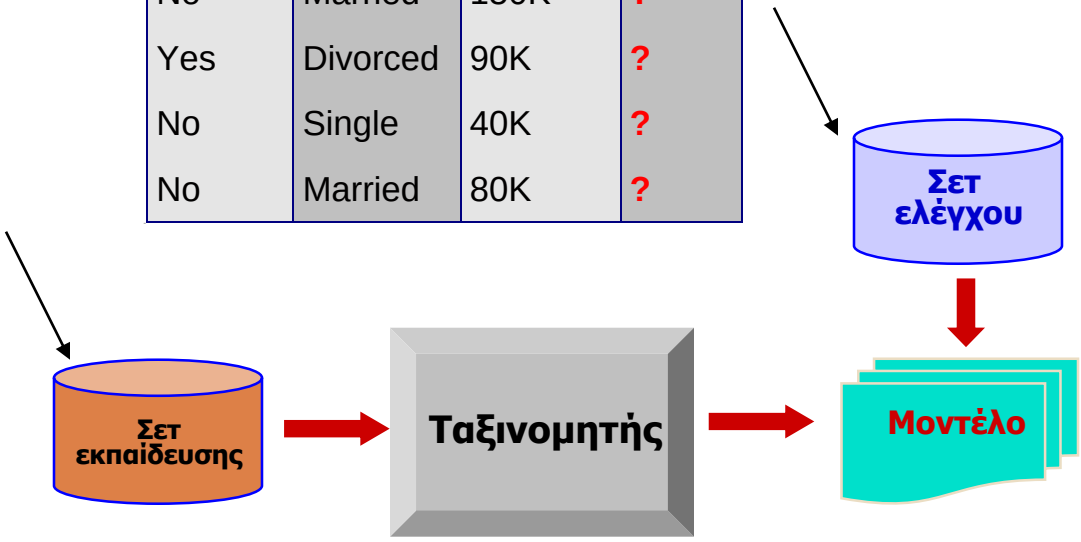
Ταξινόμηση: Ορισμός

- Δεδομένης μιας συλλογής εγγραφών (*σετ εκπαίδευσης*)
  - Κάθε εγγραφή περιέχει ένα σετ από *χαρακτηριστικά*, όπου το ένα είναι η *κλάση*.
- Εύρεση ενός ***μοντέλου*** για το χαρακτηριστικό κλάση ως μια συνάρτηση των τιμών των άλλων χαρακτηριστικών.
- Στόχος: **πρότερα μη γνωστές** εγγραφές να μπορούν να ταξινομηθούν σε μια κλάση με όσο μεγαλύτερη ακρίβεια.
  - Ένα *σετ ελέγχου* χρησιμοποιείται η ακρίβεια του μοντέλου. Συνήθως, το αρχικό σετ χωρίζεται σε εκπαίδευσης και ελέγχου, με το σετ εκπαίδευσης να χρησιμοποιείται για τη δημιουργία του μοντέλου και με το σετ ελέγχου να χρησιμοποιείται για την επικύρωσή του.

Παράδειγμα ταξινόμησης

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
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

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



Ταξινόμηση: Παράδειγμα Εφαρμογής

Καμπάνια Μάρκετινγκ

- **Στόχος:** Μείωση του ταχυδρομικού κόστους με τη *στόχευση* σε ένα τμήμα από πελάτες που είναι πιθανοί να αγοράσουν ένα καινούριο προϊόν.
- **Προσέγγιση:**
  - Χρήση δεδομένων από την προώθηση ενός παρόμοιου προϊόντος.
  - Γνώση του ποίοι πελάτες αγόρασαν το προϊόν. Με τον τρόπο αυτό ορίζουμε το χαρακτηριστικό *{αγοράζω, δεν αγοράζω}*.
  - Συλλογή διάφορων δημογραφικών στοιχείων και στοιχείων για τους πελάτες (εργασία, περιοχή διαμονής, οικογενειακή κατάσταση κλπ)
  - Χρήση της πληροφορίας αυτής για να χτιστεί ένας ταξινομητής.

[Berry & Linoff, 1997]

Ομαδοποίηση: Ορισμός

- Δεδομένου ενός σετ από πολυδιάστατα σημεία, και μιας μετρικής ομοιότητας μεταξύ τους, βρες **ομάδες** σημείων, τέτοιες ώστε:
  - Τα σημεία σε μια ομάδα έχουν μεγάλη ομοιότητα μεταξύ τους.
  - Τα σημεία διαφορετικών ομάδων είναι ανόμοια μεταξύ τους.
- Μετρικές ομοιότητας:
  - Ευκλείδεια απόσταση, αν τα χαρακτηριστικά είναι συνεχείς τιμές.
  - Σχετικές με το πρόβλημα.

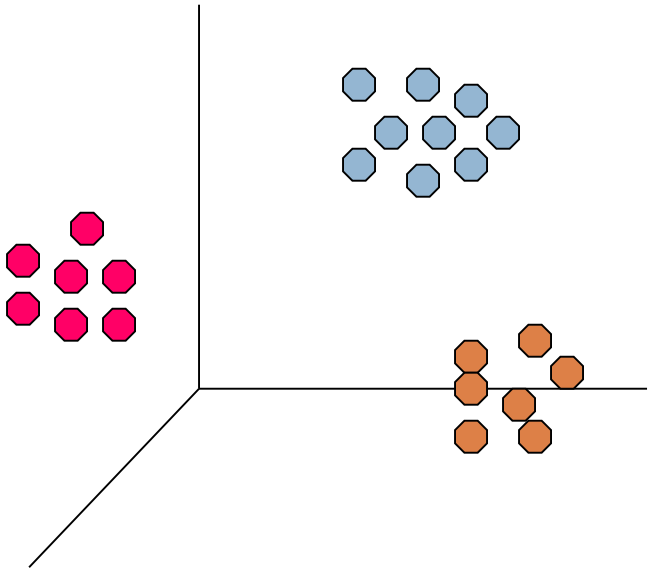
Βασικές έννοιες ομαδοποίησης

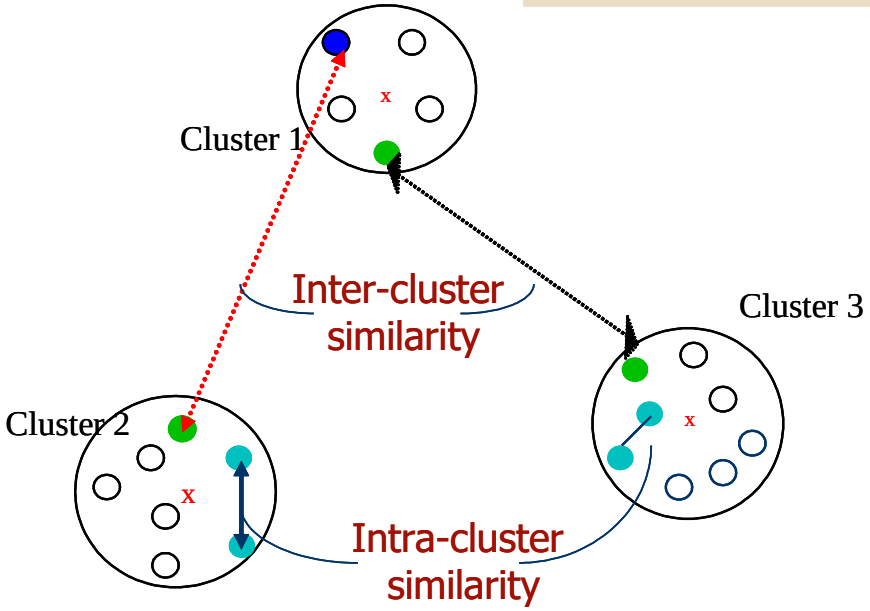
Ομαδοποίηση στον 3-D χώρο, με βάση την Ευκλείδεια απόσταση

Οι ενδο-ομαδικές αποστάσεις ελαχιστοποιούνται

$$E_K = \sum_k \left\| x_k - m_{c(x_k)} \right\|^2$$

Οι δια-ομαδικές αποστάσεις μεγιστοποιούνται





Ομαδοποίηση: Παράδειγμα Εφαρμογής

Κατάτμηση Αγοράς

- **Στόχος:** κατάτμηση της αγοράς σε διακριτές υποομάδες πελατών, όπου κάθε υποομάδα μπορεί δυνητικά να αποτελέσει συγκεκριμένη κατηγορία σε καμπάνια μάρκετινγκ.
- **Προσέγγιση:**
  - Συλλογή διαφόρων χαρακτηριστικών των πελατών σχετικά με τη δημογραφία τους και τον τρόπο ζωής τους.
  - Δημιουργία ομάδων παρόμοιων πελατών.
  - Αποτίμηση της ποιότητας των ομάδων και χαρακτηρισμός τους, παρατηρώντας τις συνήθειες αγοράς των πελατών στην ίδια ομάδα vs. αυτών από άλλες ομάδες.

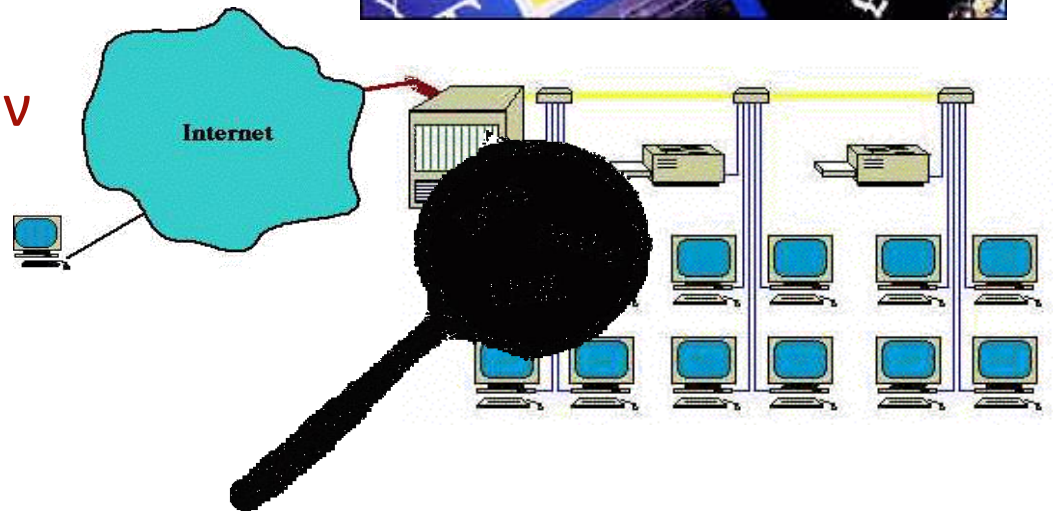
Παλινδρόμηση

- Πρόβλεψη της τιμής μιας συνεχούς μεταβλητής με βάση τις τιμές άλλων μεταβλητών, θεωρώντας ένα γραμμικό ή μη-γραμμικό μοντέλο συσχέτισης.
- Χρησιμοποιείται πάρα πολύ στη στατιστική, οικονομική θεωρία, νευρωνικά δίκτυα.
- **Παραδείγματα:**
  - Πρόβλεψη των πωλήσεων ενός προϊόντος με βάση το κόστος διαφήμισης.
  - Πρόβλεψη της ταχύτητας του ανέμου ως συνάρτηση της θερμοκρασίας, της υγρασίας, της ατμοσφαιρικής πίεσης, κτλ.
  - Πρόβλεψη της τιμής μετοχών (χρονοσειρές…)

Ανίχνευση απόκλισης/ανωμαλιών

- Ανίχνευση βασικών αποκλίσεων από την κανονική
- Εφαρμογές:
  - Ανίχνευση απάτης πιστωτικών καρτών

  - Ανίχνευση Εισβολής σε Δίκτυα Υπολογιστών



Εξερεύνηση δεδομένων

Τι είναι εξερεύνηση δεδομένων

- **Η προκαταρκτική εξερεύνηση των δεδομένων βοηθά στην καλύτερη κατανόηση των χαρακτηριστικών τους**
- Βασικοί λόγοι για εξερεύνηση:
  - Να επιλεγεί το κατάλληλο εργαλείο για προ-επεξεργασία και ανάλυση των δεδομένων
  - Να εκμεταλλευτεί την ικανότητα των ανθρώπων να αναγνωρίζουν πρότυπα
    - Οι άνθρωποι πολλές φορές αναγνωρίζουν πρότυπα που δεν συλλαμβάνονται από τα εργαλεία
- Συνδέεται άμεσα με τον χώρο της Εξερευνητικής Ανάλυσης Δεδομένων (Exploratory Data Analysis-EDA)
  - Ορίστηκε από τον στατιστικολόγο John Tukey
  - Πολύ καλό βιβλίο: *Exploratory Data Analysis,* J. Tukey
  - Καλό υλικό: http://www.itl.nist.gov/div898/handbook/index.htm

Συμβάσεις της EDA

- Πρωτεύων στόχος η οπτικοποίηση
- Η ομαδοποίηση και η αναγνώριση ανωμαλιών χρησιμοποιούνται ως τεχνικές EDA
- Στα πλαίσια του μαθήματος, η ομαδοποίηση και η αναγνώριση ανωμαλιών θεωρούνται βασικής σημασίας έννοιες.

Σετ δεδομένων Iris

- Συχνά χρησιμοποιείται ως υπόδειγμα για την εφαρμογή μεθόδων Μηχανικής Εκμάθησης
  - Διαθέσιμο: http://www.ics.uci.edu/~mlearn/MLRepository.html
  - Τρεις τύποι άνθεων (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Τέσσερις (non-class) μεταβλητές
    - Πλάτος και μήκος κάλυκα
    - Πλάτος και μήκος πέταλου

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

Στατιστικά σύνοψης

- Τα στατιστικά σύνοψης είναι μετρικές που συνοψίζουν τις ιδιότητες των δεδομένων
- Ιδιότητες σύνοψης είναι η συχνότητα, η θέση και το εύρος
  - Παραδείγματα: Θέση - *mean*
    Εύρος - *standard deviation*
- Συνήθως, τα στατιστικά σύνοψης μπορούν να υπολογιστούν με μια σάρωση των δεδομένων

Συχνότητα και Υπερισχύουσα τιμή

- Η συχνότητα (frequency) της τιμής μιας μεταβλητής είναι το ποσοστό των περιπτώσεων που η συγκεκριμένη τιμή εμφανίζεται στο σετ δεδομένων
  - Για παράδειγμα, για τη μεταβλητή ‘φύλο’ και έναν αντιπροσωπευτικό πληθυσμό, η συχνότητα της τιμής ‘γυναίκα’ εμφανίζεται περίπου 50% των περιπτώσεων.
- Η υπερισχύουσα τιμή (mode) της μεταβλητής είναι η πιο συχνή τιμή του.
- Οι έννοιες frequency και mode χρησιμοποιούνται σε κατηγορικά δεδομένα.

Εκατοστημόρια (Percentiles)

- Για συνεχή δεδομένα, χρησιμότερη είναι η έννοια του εκατοστημορίου.

  Δεδομένης μιας συνεχούς μεταβλητής *x* και ενός αριθμού *p* ανάμεσα στο 0 και 100, το p^{th} percentile είναι η τιμή x_p του x, τέτοια ώστε το *p*% των παρατηρημένων τιμών να είναι μικρότερες του x_p

- Για παράδειγμα, το 50^{th} percentile είναι η τιμή $x_{50\%}$ ώστε το 50% από όλες τις τιμές που παίρνει το *x* είναι λιγότερο από $x_{50\%}$

Μετρικές θέσεις: Mean και Median

- Ο μέσος όρος (mean) είναι η πιο διαδεδομένη μετρική της θέσης ενός συνόλου από σημεία.
- Ο mean είναι πολύ ευαίσθητος σε εξωκείμενες τιμές.
- Γι αυτό και ο μέσος (median) ή ο trimmed mean χρησιμοποιούνται εξίσου:

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Μετρικές εύρους: Range και Variance

- Range: max – min
- Variance ή standard deviation:

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

- Παρόλα αυτά, και η τυπική απόκλιση επηρεάζεται από εξωκείμενες τιμές. Γι αυτό και ορίζονται επιπλέον μετρικές:

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

$$\text{MAD}(x) = median\Big(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\Big)$$

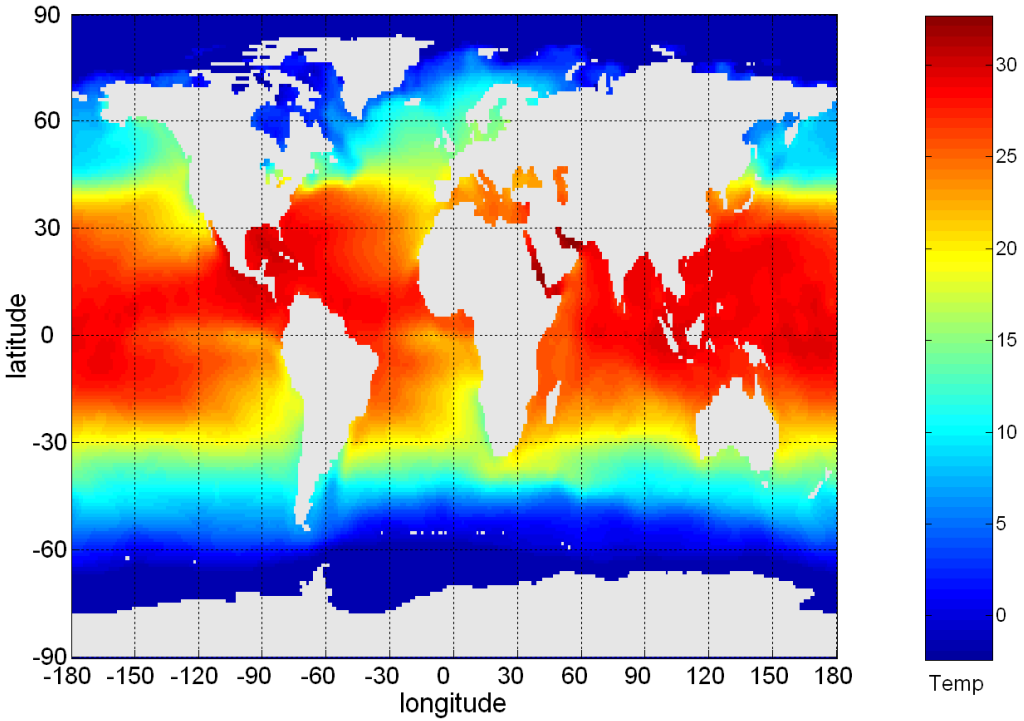
$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

Οπτικοποίηση

- Η οπτικοποίηση είναι ο μετασχηματισμός των δεδομένων σε οπτικό αντίστοιχο ή σε μορφή πίνακα, ώστε να είναι δυνατή παρατήρηση και ανάλυση στα δεδομένα.
- Η οπτικοποίηση είναι μια από τις πιο ισχυρές, αλλά και περιοριστικές τεχνικές για εξερεύνηση δεδομένων.
  - Εύκολα κατανοητή στους χρήστες.
  - Μπορεί να αναγνωρίσει γενικά πρότυπα και τάσεις.
  - Μπορεί να αναγνωρίσει εξωκείμενες τιμές και ασυνήθιστα πρότυπα.

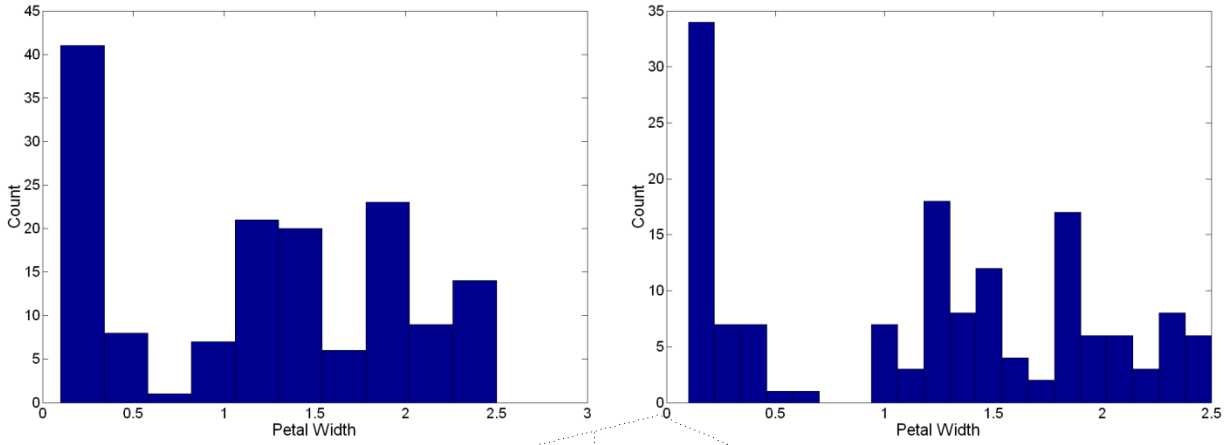
Παράδειγμα: Η θερμοκρασία της επιφάνειας της θάλασσας

- Η θερμοκρασία της επιφάνειας της θάλασσας, 13 Ιουλίου 2009.
- Δεκάδες χιλιάδες σημείων μπορούν να απεικονιστούν σε ένα διάγραμμα.

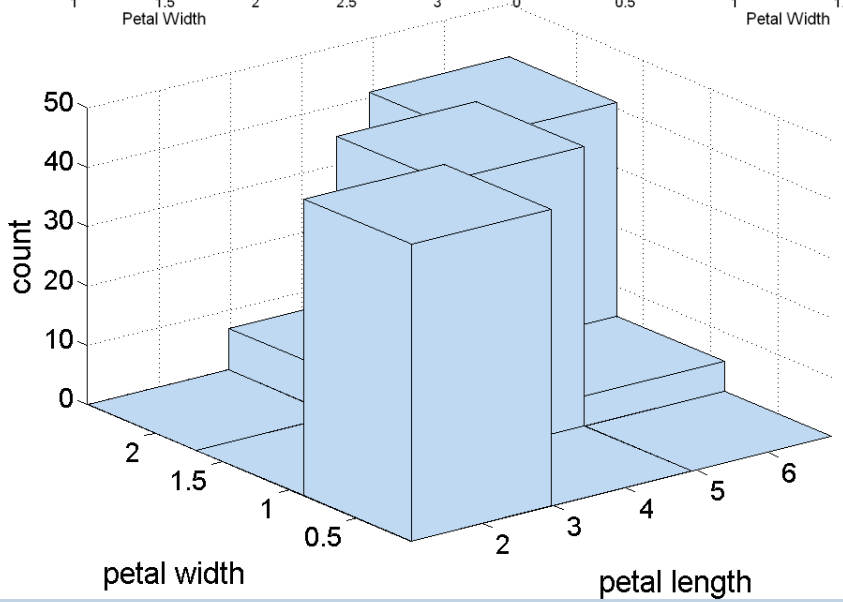


Τεχνικές οπτικοποίησης: Ιστογράμματα

- **1-D**

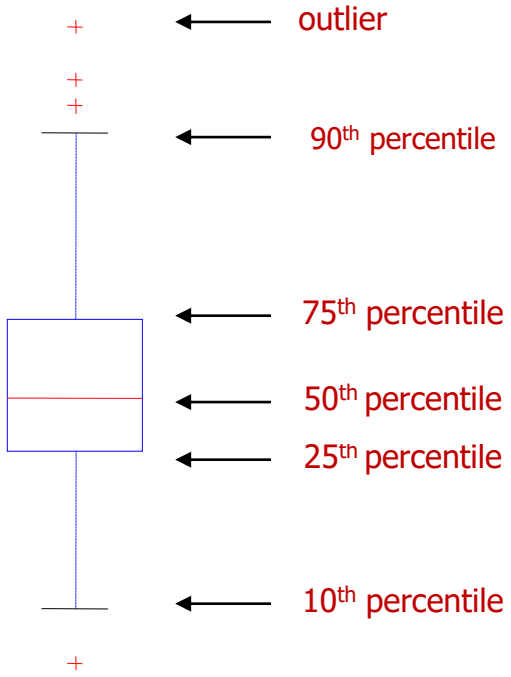


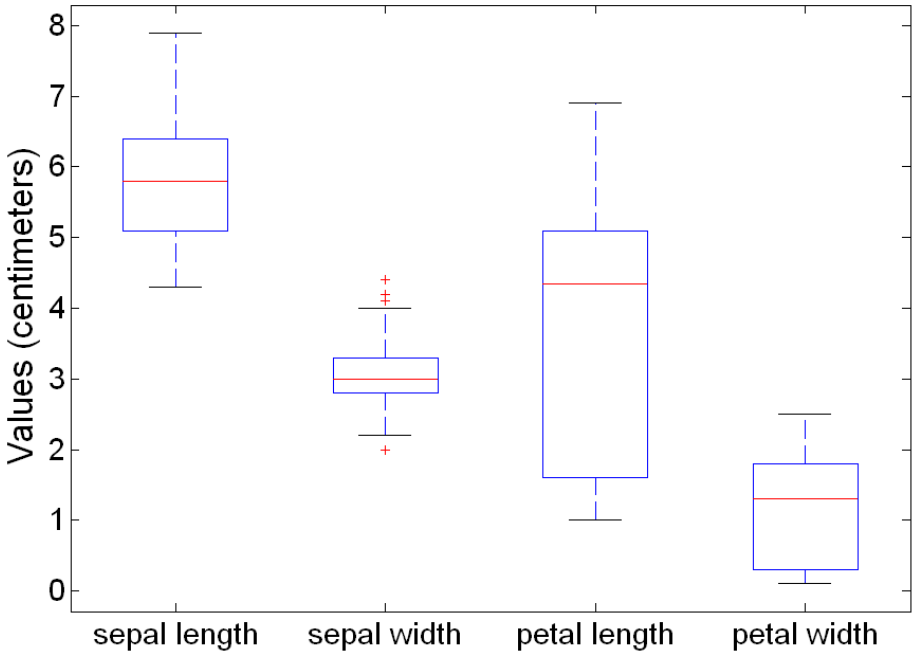
- **2-D**



Τεχνικές οπτικοποίησης: Box Plots

- Απεικόνιση κατανομών των δεδομένων

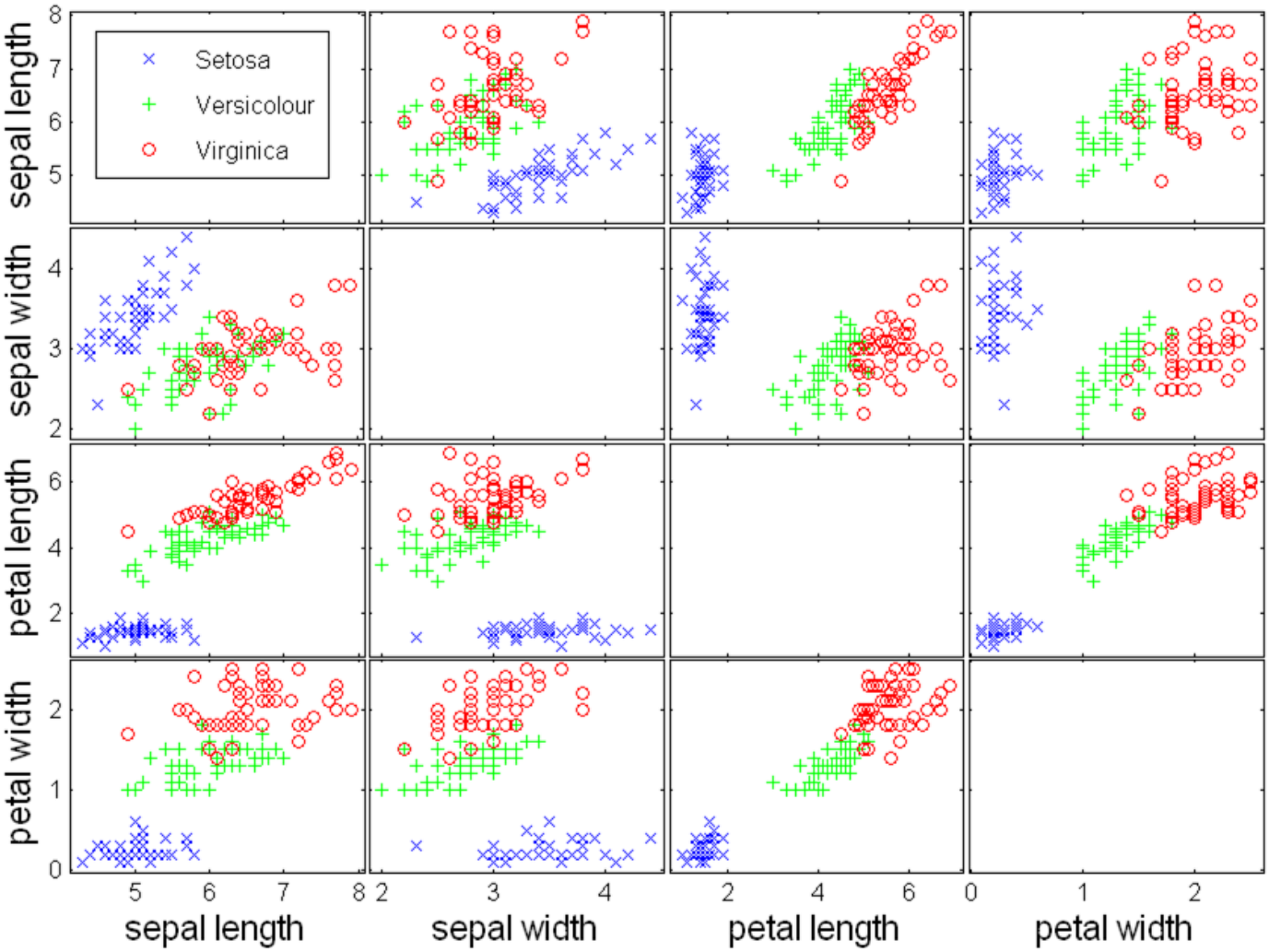




Τεχνικές οπτικοποίησης: Scatter Plots

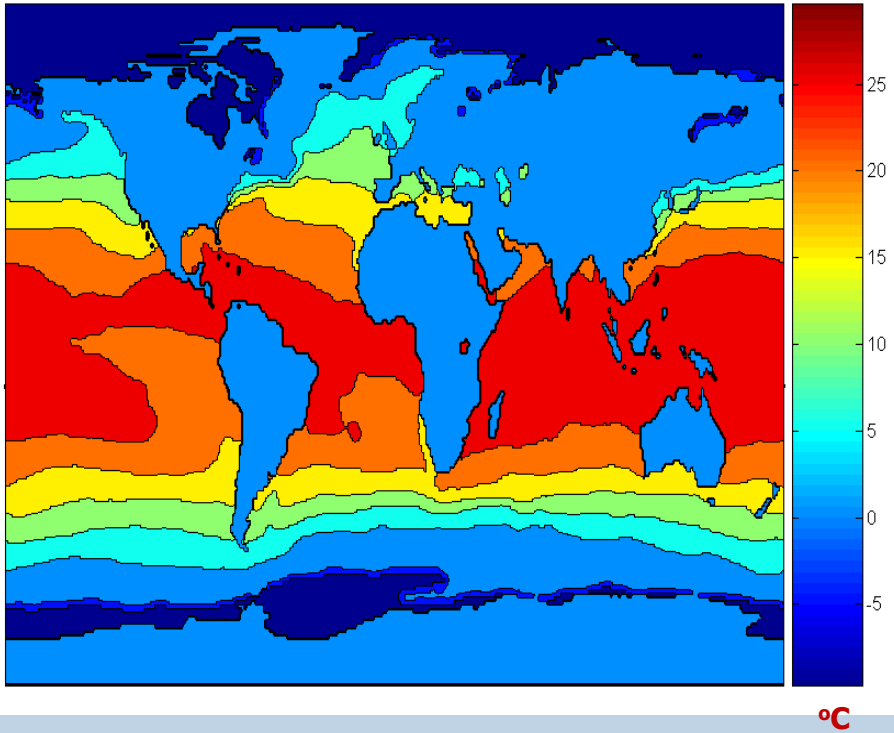
- Οι τιμές των μεταβλητών καθορίζουν τη θέση στο διάγραμμα
- Συνήθως χρησιμοποιούνται 2-Δ scatter plots (ή και 3-Δ σε κάποιες περιπτώσεις
- Επιπλέον χαρακτηριστικά μπορούν να προστεθούν με χρήση διαφορετικών χρωμάτων, μεγεθών και σχημάτων των δεικτών που αναπαριστούν τα δεδομένα.
- Είναι συχνά χρήσιμος ο συμπαγής ορισμός πινάκων από scatter plots, ώστε να συνοψίζεται η σχέση διαφόρων ζευγών χαρακτηριστικών

Scatter Plot: Παράδειγμα



Τεχνικές οπτικοποίησης: Contour Plots

- Χρήσιμα όταν μια συνεχής μεταβλητή μετριέται σε χωρικό καμβά
- Χωρίζουν τον καμβά σε περιοχές από συνεχόμενες τιμές.
- Τα περιγράμματα που ορίζουν τις περιοχές συνδέουν περιοχές με ίσες τιμές.



Τεχνικές οπτικοποίησης: Matrix Plots

- Χρήσιμο όταν τα δεδομένα ταξινομούνται σε σχέση με την ομάδα τους
- Οι μεταβλητές κανονικοποιούνται ώστε να είναι δυνατή η σύγκριση
- Μια μετρική (ομοιότητα, απόσταση, τυπική απόκλιση) καθορίζει τα χρώματα του γράφου