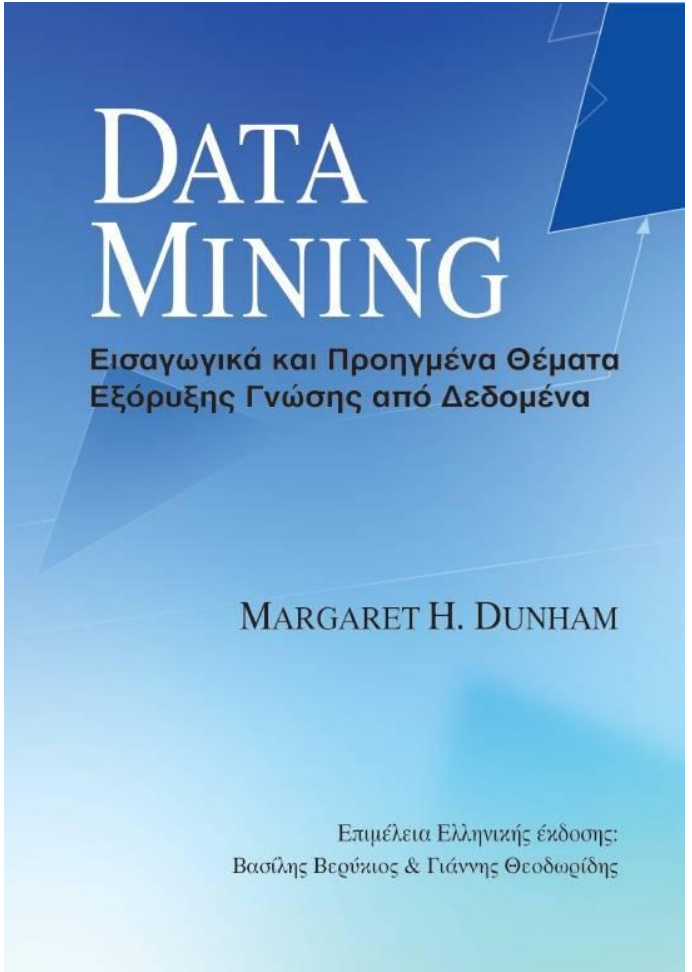
ΑΠΟΘΗΚΕΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ

Διαδικαστικά

- Διδάσκοντες
  - Θέμης Έξαρχος
  - Κάτια Κερμανίδου

Βιβλία

- Data Mining (στα Ελληνικά)
  - Margaret H.,Dunham
    - Έτος Έκδοσης: 2004
    - Εκδότης: ΕΚΔΟΣΕΙΣ ΝΕΩΝ ΤΕΧΝΟΛΟΓΙΩΝ
    - Αριθμός σελίδων: 315
    - Κωδικός ISBN: 960-8105-72-2



Βιβλία

- Εισαγωγή στην Εξόρυξη Δεδομένων και τις Αποθήκες Δεδομένων
- Αλ. Νανόπουλος - Γ. Μανωλόπουλος
  - Έτος Έκδοσης: 2008
  - Εκδότης: ΕΚΔΟΣΕΙΣ ΝΕΩΝ ΤΕΧΝΟΛΟΓΙΩΝ
  - Αριθμός σελίδων: 384
  - Κωδικός ISBN: 978-960-6759-17-8

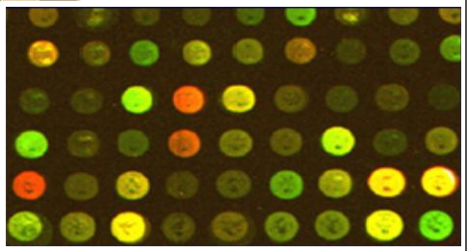
Πλημμύρα Δεδομένων

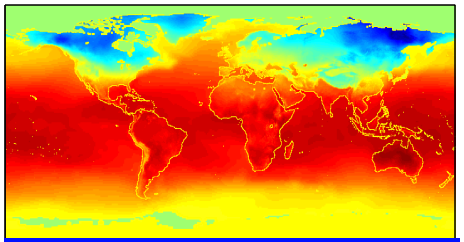
- Δεδομένα δημιουργούνται διαρκώς από:
  - Τράπεζες, τηλεπικοινωνίες, άλλες εμπορικές συναλλαγές…
  - Επιστημονικά δεδομένα: αστρονομία,βιολογία, ιατρική…
  - Διαδίκτυο, κείμενα, ηλ/κό εμπόριο

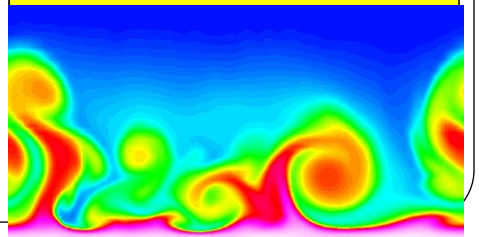
Γιατί να εξορύξουμε δεδομένα? Η εμπορική προοπτική

- Πολλά δεδομένα συλλέγονται και συσσωρεύονται σε αποθήκες δεδομένων (data warehouses)
  - Από το διαδίκτυο, το ηλ/κό εμπόριο
  - Από αγορές σε εμπορικά καταστήματα
  - Από συναλλαγές με τράπεζες, πιστωτικές κάρτες
- Οι υπολογιστές έχουν γίνει ισχυρότεροι και φθηνότεροι
- Ισχυρή πίεση από τον ανταγωνισμό
  - Παροχή καλύτερων, προσαρμοσμένων στον καταναλωτή υπηρεσιών

Γιατί να εξορύξουμε δεδομένα? Η επιστημονική προοπτική



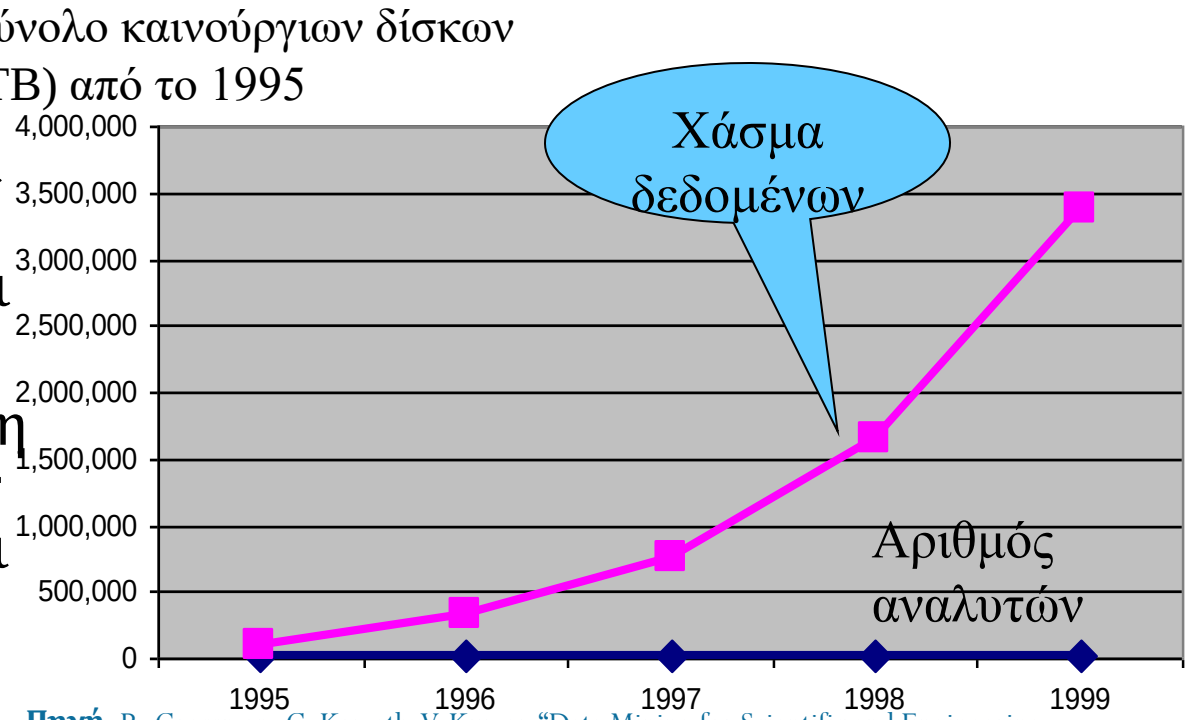




- Τα δεδομένα συλλέγονται και αποθηκεύονται με εντυπωσιακούς ρυθμούς (GB/ώρα)
  - Αισθητήρες σε δορυφόρους
  - Τηλεσκόπια ανιχνεύουν τους ουρανούς
  - Εξομοιωτές δημιουργούν τεχνητά δεδομένα
  - Ανάλυση του DNA
- Οι παραδοσιακές τεχνικές είναι μη εφικτές για αδρά δεδομένα
- Η εξόρυξη δεδομένων μπορεί να συμβάλλει
  - Στην κατηγοριοποίηση και τμηματοποίηση των δεδομένων
  - Στη διατύπωση υποθέσεων

Εξόρυξη σε μεγάλα σύνολα δεδομένων - Κίνητρο

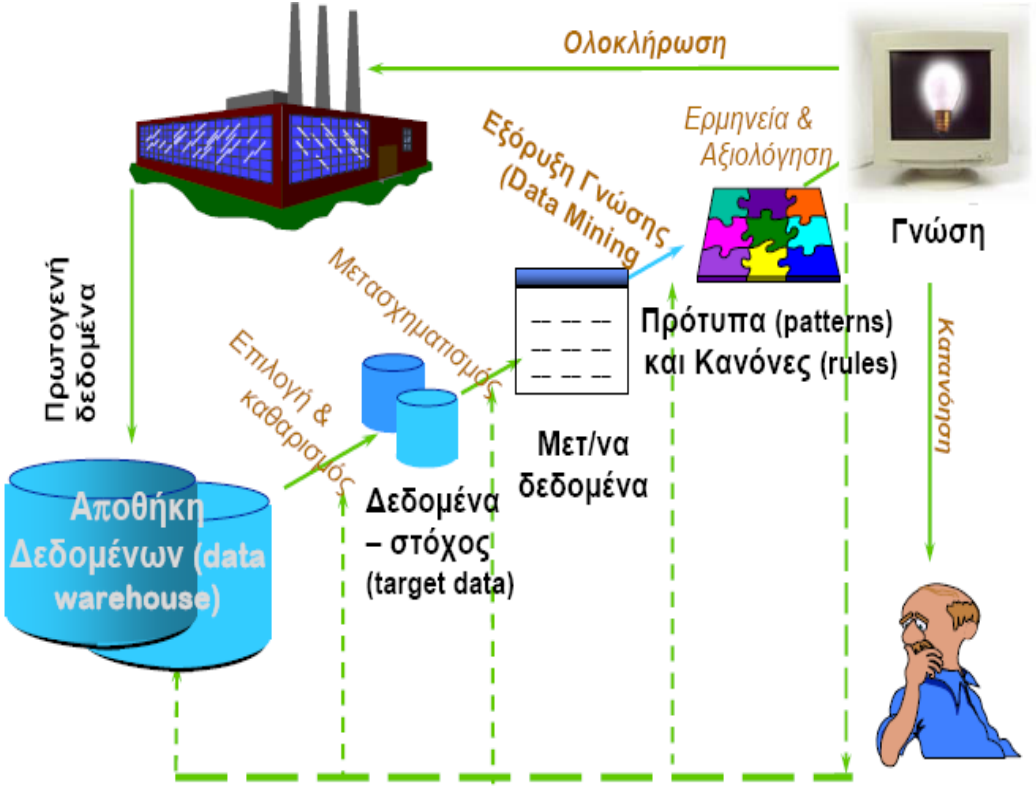
- Υπάρχει πολύ συχνά πληροφορία που είναι «κρυμμένη» μέσα στα δεδομένα και δεν είναι καθόλου προφανής
- Οι ανθρώπινη ανάλυση μπορεί να πάρει μήνες μέχρι να την εντοπίσει
- Πολλά δεδομένα δεν αναλύονται καθόλου



Πηγή: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

Τι είναι η εξόρυξη δεδομένων;

- Πληθώρα ορισμών
  - Μη τετριμμένη εξαγωγή υποκρυπτόμενης, άγνωστης και εν δυνάμει χρήσιμης πληροφορίας από τα δεδομένα
  - Εξερεύνηση και ανάλυση, με αυτόματο ή ημι-αυτόματο τρόπο, μεγάλων ποσοτήτων δεδομένων για την ανακάλυψη χρήσιμων προτύπων



Τι είναι η εξόρυξη δεδομένων;

- Η διαδικασία ημί-αυτόματης ανάλυσης μεγάλων βάσεων δεδομένων για την εύρεση προτύπων (patterns) που είναι:
  - Έγκυρα → ισχύουν για νέα δεδομένα με κάποια βεβαιότητα
  - Καινούργια → μη-προφανή για το σύστημα
  - Χρήσιμα → πιθανόν να εφαρμόζονται σε αντικείμενα
  - Κατανοήσιμα → οι άνθρωποι μπορούν να τα ερμηνεύουν
- Γνωστή και ως Ανακάλυψη Γνώσης από Δεδομένα (Knowledge Discovery in Databases) ή (Knowledge Discovery from Data) KDD

Τι (ΔΕΝ) είναι εξόρυξη δεδομένων

ΔΕΝ είναι εξόρυξη

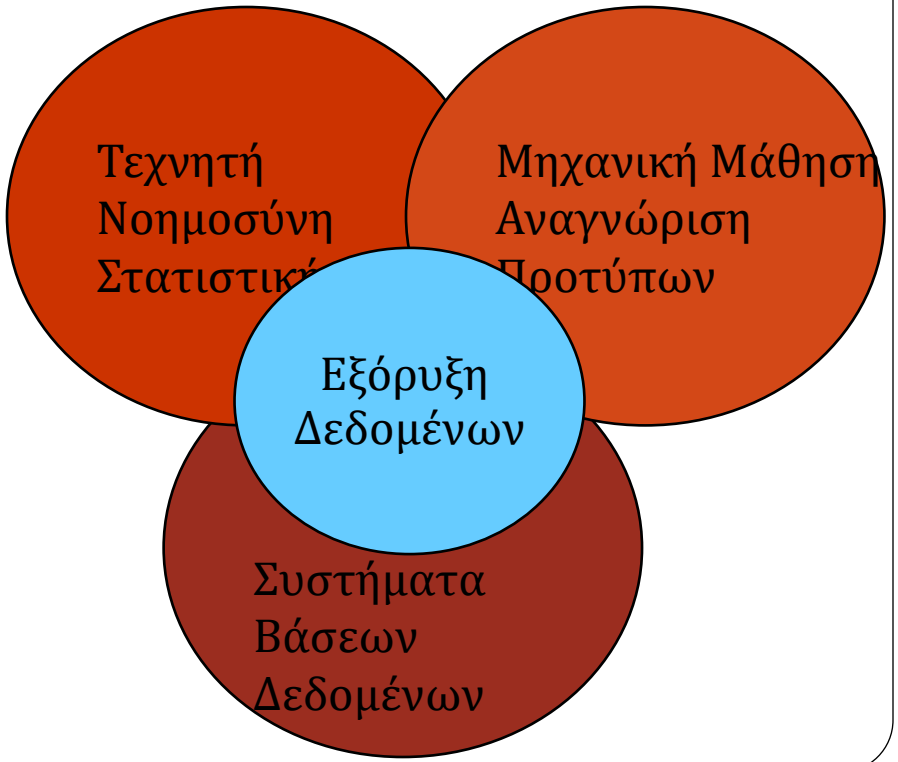
- Η εύρεση ονομάτων σε ένα τηλεφωνικό κατάλογο
- Η επερώτηση σε μια μηχανή αναζήτησης στο διαδίκτυο

ΕΙΝΑΙ εξόρυξη

- Η εύρεση παρόμοιων βιβλίων σε ένα ηλ/κό κατάστημα με βάση το περιεχόμενο τους
- Ορισμένα ονόματα είναι πιο συχνά σε συγκεκριμένες περιοχές
- Το φιλτράρισμα ανεπιθύμητων μηνυμάτων

Προέλευση της εξόρυξης δεδομένων

- Σταχυολογεί ιδέες από
  - μηχανική μάθηση, τεχνητή νοημοσύνη, αναγνώριση προτύπων, στατιστική και βάσεις δεδομένων
- Οι παραδοσιακές τεχνικές δεν είναι κατάλληλες λόγω
  - Πληθώρας των δεδομένων
  - Μεγάλης διαστατικότητας των δεδομένων
  - Ετερογενείς, κατανεμημένες οργανώσεις δεδομένων



Πραγματικές περιπτώσεις εξόρυξης

- **Επιτυχημένες**
  - Amazon ®
  - Εργασία: να συστήσουμε στον πελάτη και άλλα συναφή προϊόντα (βιβλία) τα οποία μπορεί και να αγοράσει
    - Αυτοί που αγόρασαν το «**Data Mining**» αγόρασαν και το «**Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**»
  - Το πρόγραμμα έχει βελτιώσει τις πωλήσεις της Amazon®

- **Αποτυχημένες**
  - Gazelle.com
  - Εργασία: KDD-CUP 2000
    - Στόχος: αναγνώριση πελατών που μπορεί να ξοδέψουν πάνω από 12$ /παραγγελία
    - Βάση:3465 αγορές, 1831 πελάτες
    - Έξοδα: Χ.000.000$
    - Έσοδα: Υ.000$
    - Επικήδειος: Έκλεισε τον 08/2000

Εργασίες Εξόρυξης Δεδομένων

- Προβλεπτικές Μέθοδοι
  - Χρησιμοποίηση ορισμένων μεταβλητών για να προβλεφθεί η τιμή (άγνωστη ή μελλοντική) τιμή άλλων μεταβλητών
    - Π.χ. η μέτρηση του όγκου συναλλαγών μιας μετοχής για την πρόβλεψη της πορείας του ΓΔ την επόμενη ημέρα
- Περιγραφικές Μέθοδοι
  - Εύρεση προτύπων που μπορεί να ερμηνεύσει ο άνθρωπος για την περιγραφή των δεδομένων
    - Π.χ. η εύρεση ενός κανόνα με βάση τον οποίο ένα σύστημα θα κατατάσσει ένα μήνυμα ως ανεπιθύμητο

Εργασίες Εξόρυξης Δεδομένων…

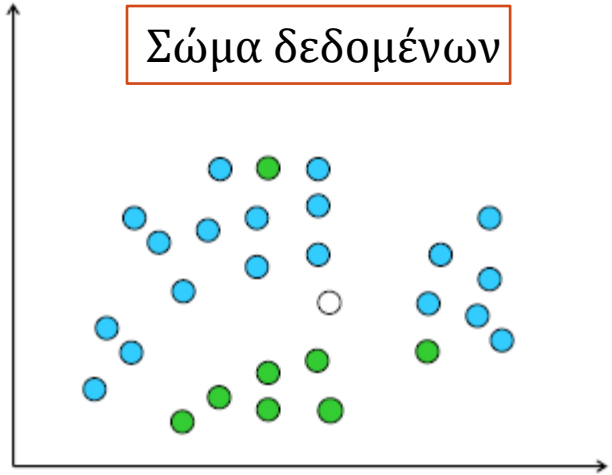
- Κατηγοριοποίηση ή Ταξινόμηση-Classification [Προβλεπτική]
- Συσταδοποίηση ή Ομαδοποίηση-Clustering [Περιγραφική]
- Ανακάλυψη Κανόνων Συσχετίσεων-Association Rule Discovery [Περιγραφική]
- Ανακάλυψη Χρονολογικών Προτύπων-Sequential Pattern Discovery [Περιγραφική]
- Παλινδρόμηση-Regression [Προβλεπτική]
- Ανίχνευση Εξαιρέσεων-Deviation Detection [Προβλεπτική]

Κατηγοριοποίηση: Ορισμός

- Δοσμένης μιας συλλογής εγγραφών (σώμα εκπαίδευσης)
  - Κάθε εγγραφή περιέχει ένα σύνολο ιδιοτήτων, μια εκ των οποίων είναι η κλάση (ή κατηγορία)
- Βρες ένα μοντέλο για την κλάση ως συνάρτηση των τιμών των άλλων μεταβλητών
- Στόχος: η ακριβής ανάθεση κλάσης σε νέες εγγραφές
  - Ένα σύνολο αξιολόγησης χρησιμοποιείται για την αξιολόγηση του μοντέλου. Συνήθως το σώμα δεδομένων χωρίζεται σε σώμα εκπαίδευσης και αξιολόγησης

Κατηγοριοποίηση: Παράδειγμα 1

Σώμα δεδομένων



```
if X > 5 then blue
  else if Y > 3 then blue
    else if X > 2 then green
      else blue
```

Μοντέλο κατηγοριοποίησης με δέντρα αποφάσεων

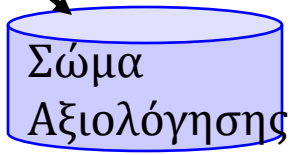
Κατηγοριοποίηση: Παράδειγμα 2

Διακριτή | Διακριτή | Συνεχής | Κλάση

Tid	Επιστροφή	Οικογεν. Κατάσταση	Εισόδημα	Απάτη;
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

Επιστροφή	Οικογεν. Κατάσταση	Εισόδημα	Απάτη;
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

Εκμάθηση
Μοντέλου



Κατηγοριοποίηση: Εφαρμογή 1

- Μάρκετινγκ
  - Στόχος: Η μείωση του κόστους με την ταχυδρόμηση ενός συγκεκριμένου συνόλου καταναλωτών που είναι πιθανόν να αγοράσουν ένα νέο κινητό τηλέφωνο

Πηγή: [Berry & Linoff] Data Mining Techniques, 1997

- Προσέγγιση:
  - Χρήση δεδομένων από ένα προγενέστερο, παρόμοιο τηλέφωνο
  - Γνωρίζουμε ποιοι πελάτες αποφάσισαν να αγοράσουν και ποιο όχι. Αυτή είναι και η κλάση του προβλήματος
  - Συλλογή διαφόρων δημογραφικών στοιχείων και εταιρικών στοιχείων
    - Φύλο, ηλικία, τόπος διαμονής, εισόδημα, λόγοι χρήσης, κτλ.
  - Χρήση των δεδομένων για την εκμάθησης ενός μοντέλου κατηγοριοποίησης

Κατηγοριοποίηση: Εφαρμογή 2

- Ανίχνευση Απάτης
  - Στόχος: Η πρόβλεψη χρήσης κλεμμένων πιστωτικών καρτών

- Προσέγγιση:
  - Χρησιμοποίησε τις συναλλαγές με πιστωτικές κάρτες και πληροφορίες για τους κατόχους τους ως ιδιότητες.
    - Πότε αγοράσθηκε κάτι, τι αγόρασε, πόσο συχνά πληρώνει τη δόση της κάρτας, κτλ.
  - Επισημείωσε τις παλαιότερες συναλλαγές σαν κανονικές ή όχι. Αυτή είναι η κλάση του προβλήματος
  - Εκμάθηση ενός μοντέλου της κλάσης
  - Χρησιμοποίησε το μοντέλο για την ανίχνευση πιθανής χρήσης κλεμμένης κάρτας.

Κατηγοριοποίηση: Εφαρμογή 3

- Διαρροή πελατών
  - Στόχος: Να προβλέψουμε εάν ένας πελάτης θα πάει σε κάποιον ανταγωνιστή

Πηγή: [Berry & Linoff] Data Mining Techniques, 1997

- Προσέγγιση:
  - Χρήση λεπτομερών εγγραφών με τις συναλλαγές με τον πελάτη
    - Πόσο συχνά καλεί την εταιρία, τι ώρα καλεί, την οικονομική κατάσταση, κτλ.
  - Επισημείωσε τον κάθε πελάτη σαν πιστό ή όχι
  - Εύρεση μοντέλου για την πιστότητα

Κατηγοριοποίηση: Εφαρμογή 4

- Ουράνια αντικείμενα
  - Στόχος: η πρόβλεψη της κλάσης (άστρο ή γαλαξίας) από ουράνια αντικείμενα, με βάση τηλεσκοπικές εικόνες του Αστεροσκοπείου Palomar
    - 3000 εικόνες με 23,040 x 23,040 pixel ανά εικόνα

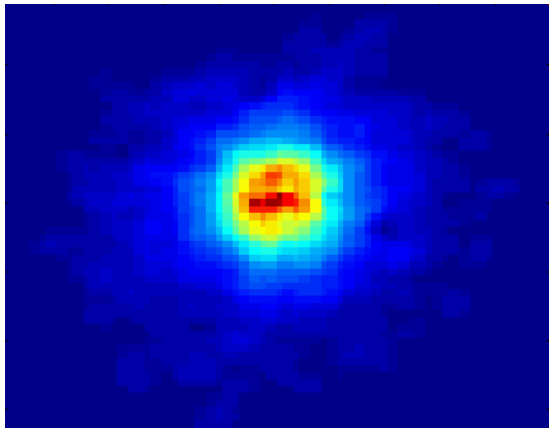
- Προσέγγιση
  - Τμηματοποίηση της εικόνας
  - Μέτρηση ιδιοτήτων εικόνας (40 ανά αντικείμενο)
  - Μοντελοποίηση της κλάσης με βάση τις παραπάνω ιδιότητες
  - Ιστορικό επιτυχίας: Εύρεση 16 νέων κόκκινων κβάζαρ, ένα από τα δυσκολότερα και πιο απομακρυσμένων αντικειμένων στο σύμπαν

Πηγή: [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Πηγή: http://aps.umn.edu

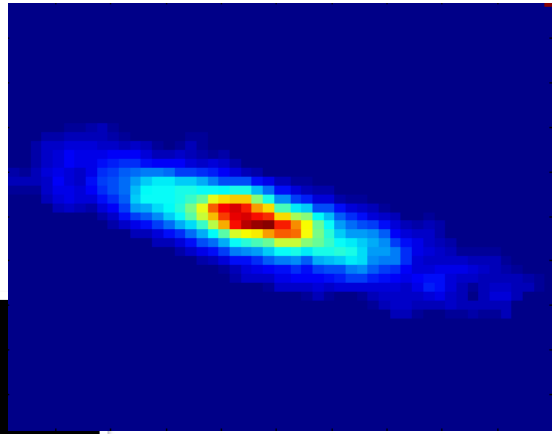
Κατηγοριοποίηση Γαλαξιών

Αρχικό



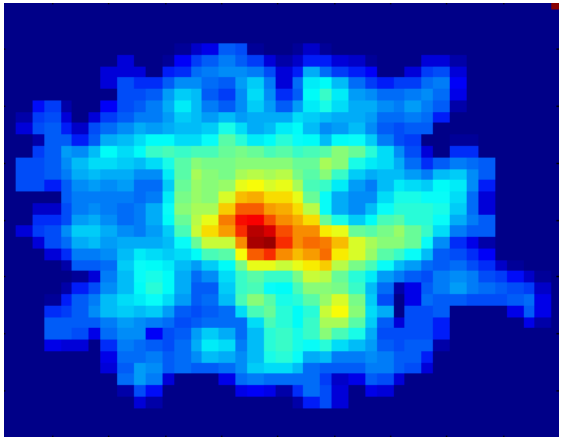
Κλάση:
- Στάδια της Ανάπτυξης

Ενδιάμεσο



Ιδιότητες:
- Χαρακτηριστικά Εικόνας,
- Χαρακτηριστικά των κυμάτων φωτός, κτλ.

Τελικό



Μέγεθος δεδομένων:
- 72M άστρα, 20M γαλαξίες
- Κατάλογος Αντικειμένων: 9 GB
- Βάση δεδομένων εικόνων: 150 GB

Ομαδοποίηση (ή Συσταδοποίηση): Ορισμός

- Δεδομένου ενός συνόλου δεδομένων, αποτελούμενο από ένα σύνολο ιδιοτήτων και ένα μέτρο ομοιότητας μεταξύ τους, βρες ομάδες τέτοιες ώστε:
  - Τα δεδομένα μιας ομάδας να είναι πιο όμοια μεταξύ τους
  - Τα δεδομένα ξεχωριστών ομάδων να είναι λιγότερο όμοια μεταξύ τους
- Μέτρα ομοιότητας
  - Ευκλείδεια απόσταση, αν οι ιδιότητες είναι συνεχείς
  - Άλλα μέτρα, εξαρτώμενα της εφαρμογής

Ομαδοποίηση Εγγράφων

- Δεδομένα ομαδοποίησης
  - 3204 άρθρα των Los Angeles Times.
  - Μέτρο ομοιότητας:
    - Πόσες λέξεις είναι κοινές στα έγγραφα αυτά (μετά από ένα στάδιο φιλτραρίσματος)

Κατηγορία	*Άρθρα*	*Σωστή Ομάδα*
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

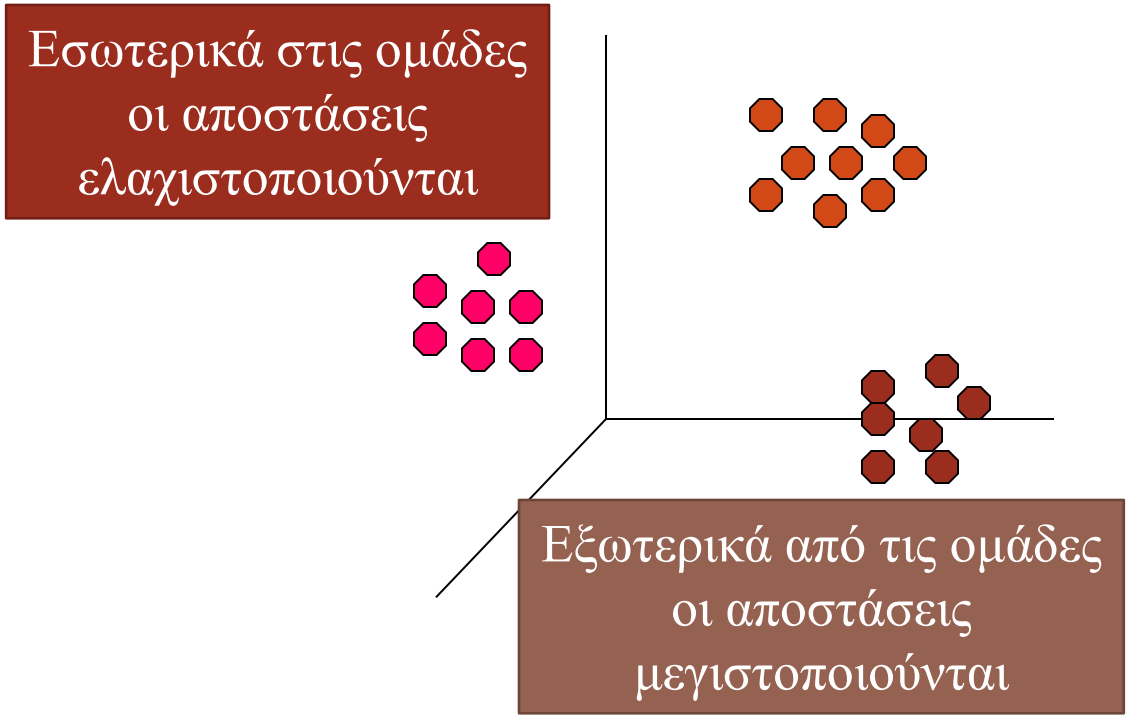
Ομαδοποίηση μετοχών

- Παρατήρηση μετοχών σε ημερήσια βάση.
- Ομάδες: Μετοχή{UP/DOWN}
- Μέτρο ομοιότητας
  - Δυο μετοχές είναι πιο όμοιες αν τα γεγονότα που τα περιγράφουν συμβαίνουν συχνά την ίδια ημέρα

	Εξαγόμενες Ομάδες	***Κατηγορία Μετοχής***
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Απεικόνιση της ομαδοποίησης

- Ομαδοποίηση στον 3Δ χώρο με Ευκλείδεια απόσταση



Ομαδοποίηση: Εφαρμογή 1

- Τμηματοποίηση Αγοράς:
  - Στόχος: διαχωρισμός της αγοράς σε υποσύνολα πελατών, με κάθε υποσύνολο να αποτελεί ξεχωριστό target group
- Προσέγγιση:
  - Συλλογή διαφορετικών ιδιοτήτων των πελατών, βασισμένων στη γεωγραφική και κοινωνική τους κατανομή
  - Εύρεση ομάδων παρόμοιων πελατών
  - Μέτρηση της ποιότητας των ομάδων με παρατήρηση αγοραστικών προτύπων των πελατών της ίδιας ομάδας

Ομαδοποίηση: Εφαρμογή 2

- Ομαδοποίηση Εγγράφων
  - Στόχος: η εύρεση ομάδων κοινών εγγράφων τους όρους
- Προσέγγιση:
  - Αναγνώριση συχνών όρων σε κάθε έγγραφο. Σχηματισμός μέτρου ομοιότητας με βάση τις συχνότητες. Ομαδοποίηση με βάση το μέτρο αυτό
  - Κέρδος: η ανάκτηση πληροφορίας μπορεί να αξιοποιήσει τις ομάδες για νε συσχετίσει ένα έγγραφο ή λέξη κλειδί με έγγραφα από τις ομάδες.

Ανακάλυψη κανόνων συσχετίσεων: Ορισμός

- Δεδομένου ενός συνόλου από εγγραφών, κάθε μια εκ των οποίων περιέχει ένα αριθμό αντικειμένων από μια δεδομένη συλλογή
  - Παραγωγή κανόνων εξάρτησης οι οποίοι προβλέπουν την εμφάνιση ενός αντικειμένου με βάση την εμφάνιση άλλων αντικειμένων

ID	*Αντικείμενα*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Εξαγωγή Κανόνων:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Ανακάλυψη κανόνων συσχετίσεων: Εφαρμογή 1

- Προώθηση προϊόντων
  - Έστω ότι ο κανόνας που εξήχθηκε λέει ότι:
    - {παξιμάδια,…}→πατατάκια

- Τα πατατάκια ως συνεπαγωγή → χρησιμοποιούνται για να εξακριβωθεί πως θα προωθήσουν την πώληση τους
- Τα παξιμάδια ως προγενέστερο αντικείμενο της συνεπαγωγής → χρησιμοποιούνται για να εξακριβωθούν οι συνέπειες εάν σταματήσει η πώληση τους
- Ο συνολικός κανόνας → μπορεί να χρησιμοποιηθεί για να εξακριβωθεί ποια προϊόντα πρέπει να πωλούνται μαζί με τα παξιμάδια για να πωλούνται και πατατάκια μαζί!

Ανακάλυψη κανόνων συσχετίσεων: Εφαρμογή 2

- Αυτοδιοίκηση σε πολυκατάστημα
  - Στόχος: να αναγνωρίσουν ποια προϊόντα αγοράζονται μαζί από πολλούς πελάτες
- Προσέγγιση:
  - Επεξεργασία των κωδικών κάθε προϊόντος όπως αυτά περνάνε από το barcode του ταμείου
- Ένας κλασικός κανόνας
  - Εάν ο πελάτης αγοράσει γάλα και πάνες, τότε πιθανό να αγοράσει και μπύρα
  - Επομένως, δεν είναι απίθανο να τοποθετηθούν οι μπύρες δίπλα στις πάνες!

Ανακάλυψη κανόνων συσχετίσεων: Εφαρμογή 3

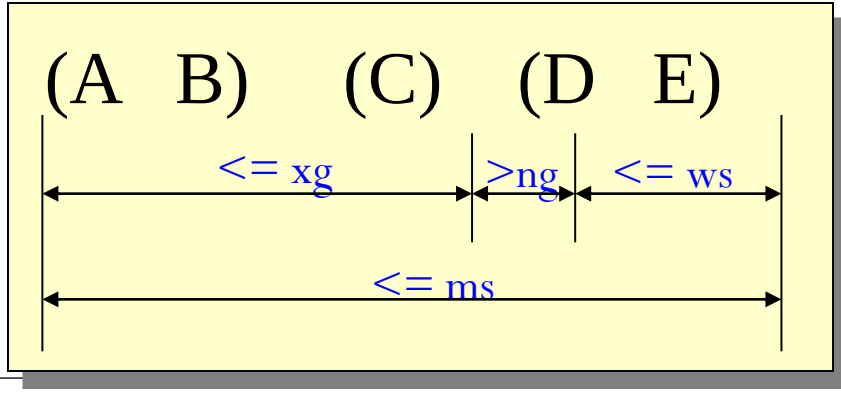
- Διαχείριση εργαλειοθήκης:
  - Στόχος: μια εταιρία επισκευής ηλεκτρικών συσκευών θέλει να προβλέπει τη φύση της βλάβης και να έχει στα οχήματα τα κατάλληλα εργαλεία
- Προσέγγιση:
  - Επεξεργασία των δεδομένων για τα εργαλεία και τις βλάβες σε προγενέστερες περιπτώσεις βλαβών και ανακάλυψη προτύπων συνεμφάνισης

Ανακάλυψη χρονολογικών προτύπων: Ορισμός

- Δεδομένου ενός συνόλου αντικειμένων, με κάθε αντικείμενο να έχει τη δική του χρονολογική σειρά γεγονότων, βρες κανόνες που προβλέπουν ισχυρή χρονολογική αλληλεξάρτηση μεταξύ γεγονότων

- Οι κανόνες σχηματίζονται από ανακάλυψη προτύπων. Τα συμβάντα των γεγονότων διέπονται από χρονικούς περιορισμούς

(A B) (C) (D E)



Ανακάλυψη χρονολογικών προτύπων: Παραδείγματα

- Στις τηλεπικοινωνίες
  - (Inverter_Problem Excessive_Line_Current)
    (Rectifier_Alarm) → (Fire_Alarm)
- Στα βιβλιοπωλεία
  - (Intro_To_Visual_C) (C++_Primer) → (Perl_for_dummies,Tcl_Tk)
- Στα καταστήματα αθλητικών
  - (Shoes) (Racket, Racketball) --> (Sports_Jacket)

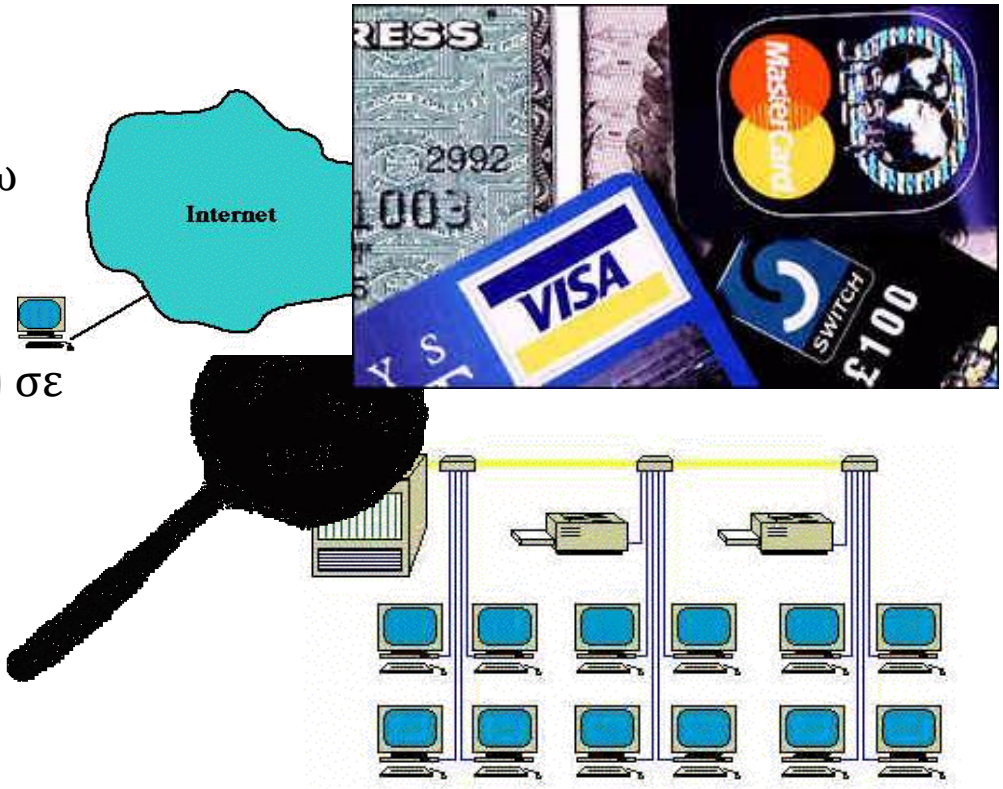
Παλινδρόμηση

- Η πρόβλεψη της τιμής μιας δοσμένης συνεχούς μεταβλητής, με βάση τις τιμές άλλων μεταβλητών, υποθέτοντας μια γραμμική ή όχι εξάρτηση του μοντέλου
- Μεγάλη εξέλιξη στη στατιστική, τα νευρωνικά δίκτυα κτλ.

- Παραδείγματα
  - Πρόβλεψη των πωλήσεων ενός νέου προϊόντος με βάση τα έξοδα διαφήμισης.
  - Πρόβλεψη των ταχυτήτων των ανέμων ως συνάρτηση της υγρασίας, θερμοκρασίας και ατμοσφαιρικής πίεσης.
  - Πρόβλεψη χρονοσειρών των δεικτών του χρηματιστηρίου

Ανίχνευση Εξαιρέσεων

- Ορισμός: Εξαίρεση
  - Ανίχνευση σημαντικών παρακάμψεων από κανονικές συμπεριφορές
- Εφαρμογές
  - Ανίχνευση απάτης (Fraud Detection) σε αγορές με πιστωτικές κάρτες
  - Αναγνώριση εισβολών σε δίκτυα
  - Αναγνώριση ξεπλύματος χρήματος
  - Τρομοκρατία



Εξόρυξη γνώσης και ιδιωτικό απόρρητο

- TIA: Terrorism (πρώην Total) Information Awareness Program
  - Επιστημονικό πρόγραμμα χρηματοδοτούμενο από το Υπουργείο Άμυνας των ΗΠΑ
  - Διακόπηκε από το Κογκρέσο
  - Μερικές από τις λειτουργίες του μεταφέρθηκαν στις υπηρεσίες πληροφοριών
- CAPPS II – φωτογράφηση όλων των επιβατών των αερομεταφορών
  - έχει προκαλέσει αντιδράσεις (από επιβάτες, εταιρείες, κυβερνήσεις)

- Οι τεχνικές εξόρυξης δεδομένων αναζητούν πρότυπα, όχι ανθρώπους!
- Υπάρχουν τεχνικές λύσεις που μπορούν να περιορίσουν την πρόσβαση σε προσωπικά δεδομένα
  - Αντικατάσταση ευαίσθητων δεδομένων με ανώνυμους κωδικούς (data anonymization)
  - Κατανεμημένα δεδομένα – κατανεμημένος υπολογισμός (distributed data mining)

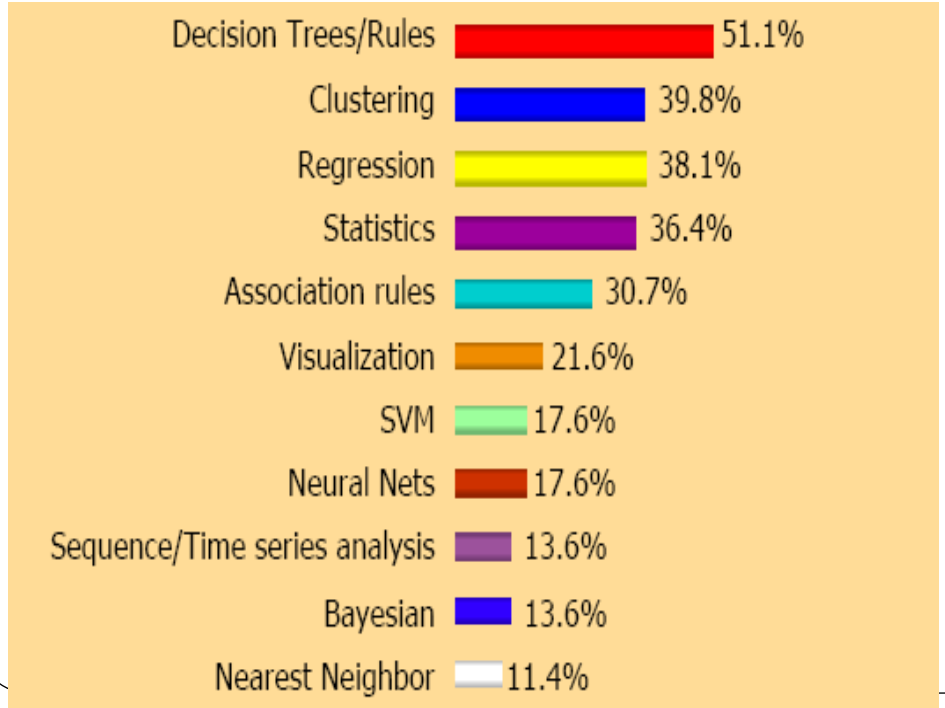
Εξόρυξη γνώσης και ιδιωτικό απόρρητο….

- Το 2006, η NSA (National Security Agency) κατηγορήθηκε ότι παρακολουθεί όλες τις εσωτερικές τηλεφωνικές συνομιλίες (~200M πελάτες) για να εντοπίσει τρομοκράτες
- Παραβίαση της ιδιωτικότητας
  - Θα σας πείραζε η παρακολούθηση των συνομιλιών σας από την κυβέρνηση;
    - (βλ. ταινία: *Οι ζωές των άλλων*)
- Τι θα συνέβαινε αν η NSA εύρισκε 1 πραγματικό ύποπτο σε 1000 εσφαλμένους υπόπτους;
  - Από 1.000.000;

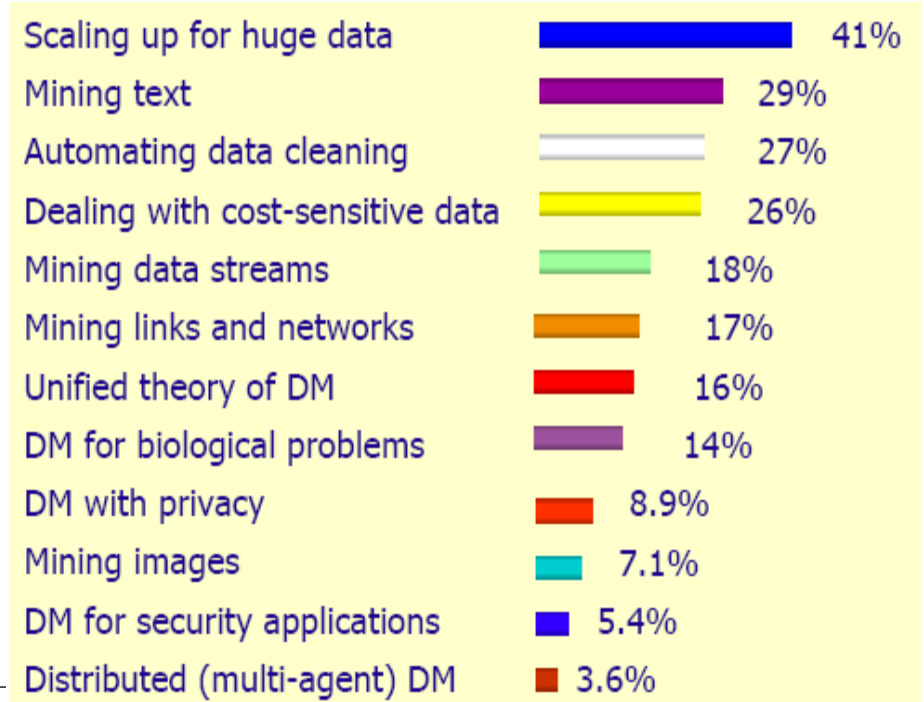
Μερικά στατιστικά

Πηγή: kdnuggets.com

- Οι πιο δημοφιλείς

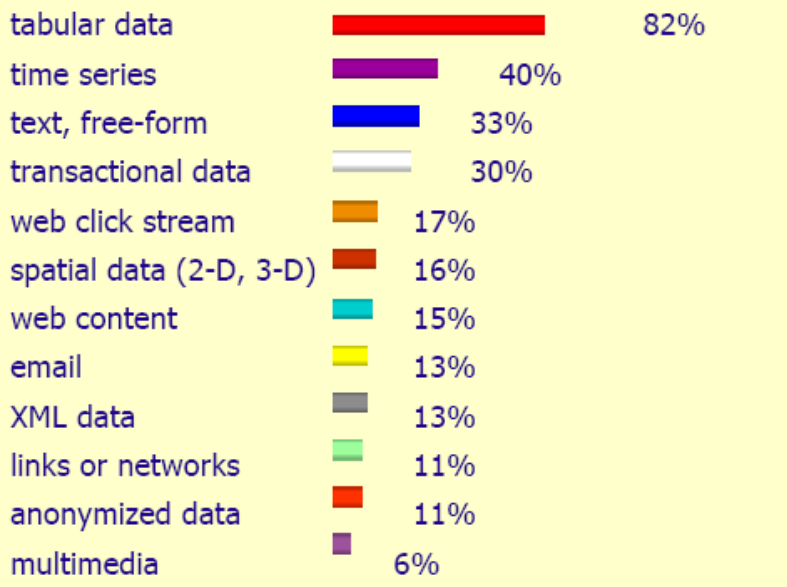


- Ερευνητικά προβλήματα



Ακόμη μερικά στατιστικά

- Τι δεδομένα αναλύουμε



Πηγή: kdnuggets.com

Προκλήσεις στην εξόρυξη δεδομένων

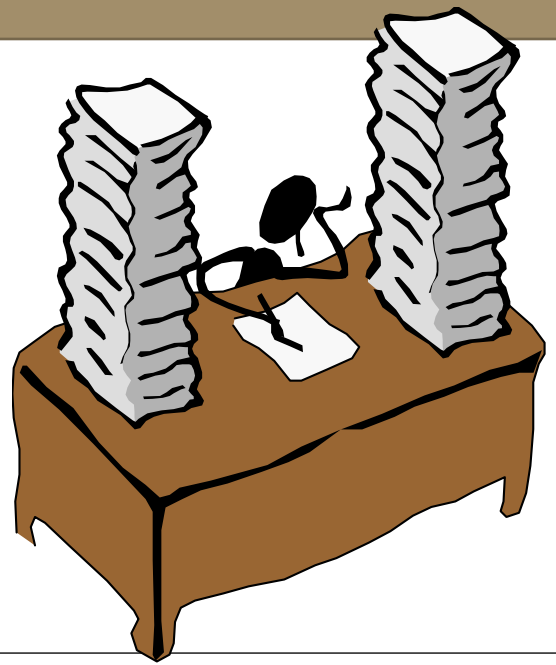
- Προσαρμοστικότητα
- Διαστατικότητα
- Πολύπλοκα και ετερογενή δεδομένα
- Ποιότητα δεδομένων
- Ιδιοκτησία και κατανομή δεδομένων
- Διατήρηση ανωνυμίας
- Συνεχής ροή δεδομένων

Γιατί εξόρυξη δεδομένων τώρα;

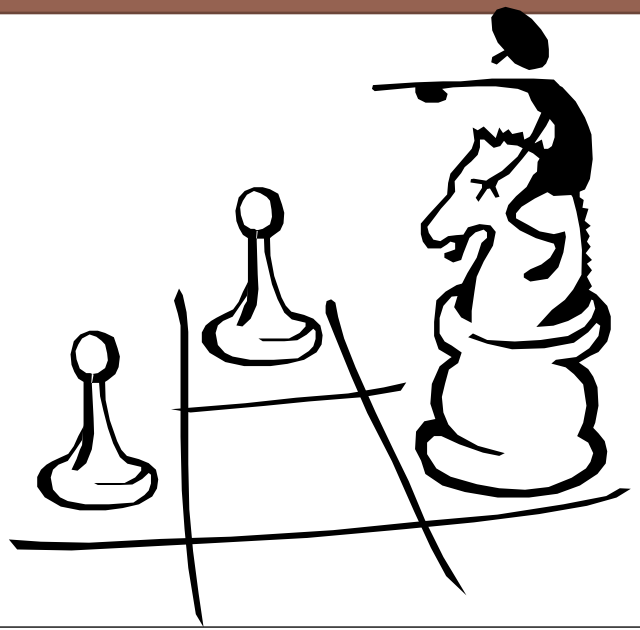
- Τα δεδομένα παράγονται και οργανώνονται
- Η υπολογιστική ισχύς
  - Είναι διαθέσιμη
  - Είναι προσιτή
- Υψηλός ανταγωνισμός
- Διάθεση εμπορικών εφαρμογών

Η εξόρυξη δουλεύει με αποθήκες δεδομένων

Οι αποθήκες δεδομένων παρέχουν στην επιχείρηση τη **μνήμη**



Η εξόρυξη δεδομένων παρέχει στην επιχείρηση την **ευφυΐα**



Σενάρια χρήσης

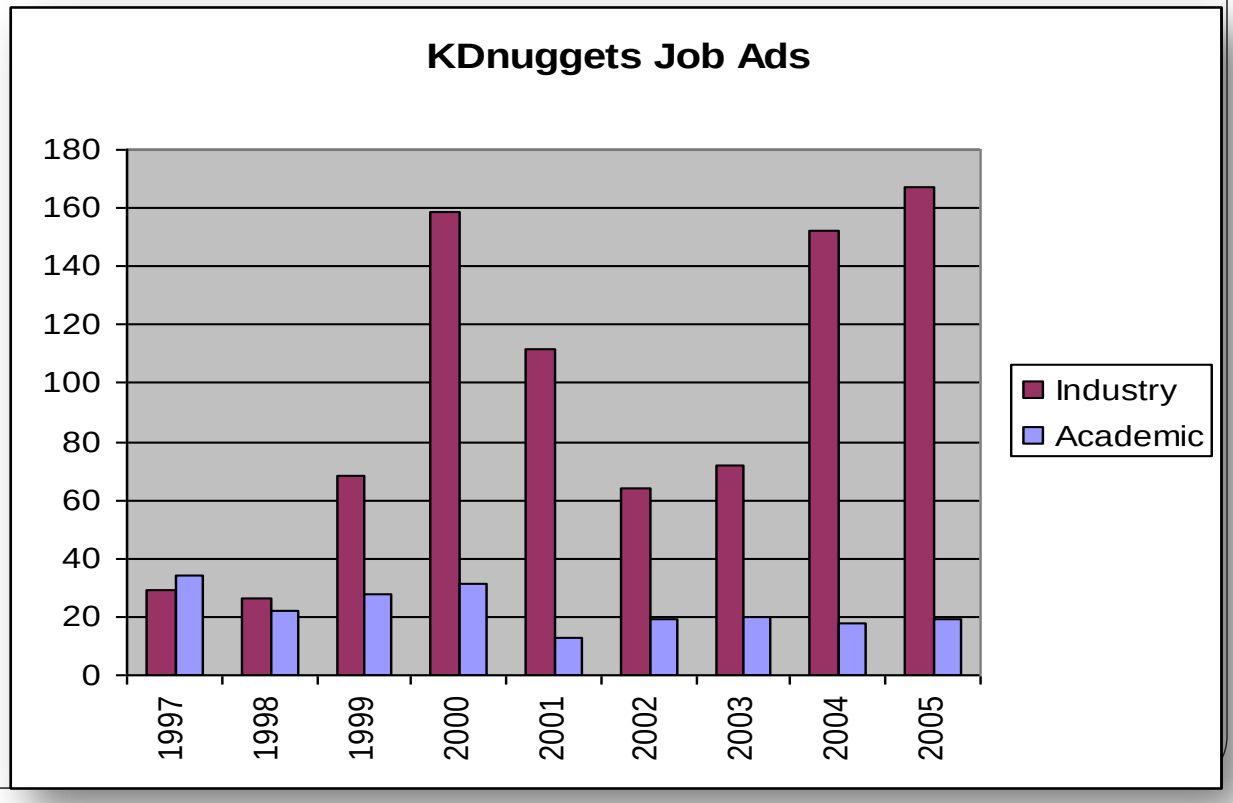
- Εξόρυξη δεδομένων από αποθήκες δεδομένων
  - Συγχώνευση δεδομένων από λειτουργικές πηγές
  - Εξόρυξη στατικών δεδομένων
- Εξόρυξη αρχείων καταγραφής χρήσης (Log data)
- Συνεχής εξόρυξη: παράδειγμα στον έλεγχο διαδικασιών (process control)
- Στάδια εξόρυξης:
  - **Επιλογή δεδομένων → προ-επεξεργασία → μετασχηματισμός → εξόρυξη → αξιολόγηση → οπτικοποίηση**

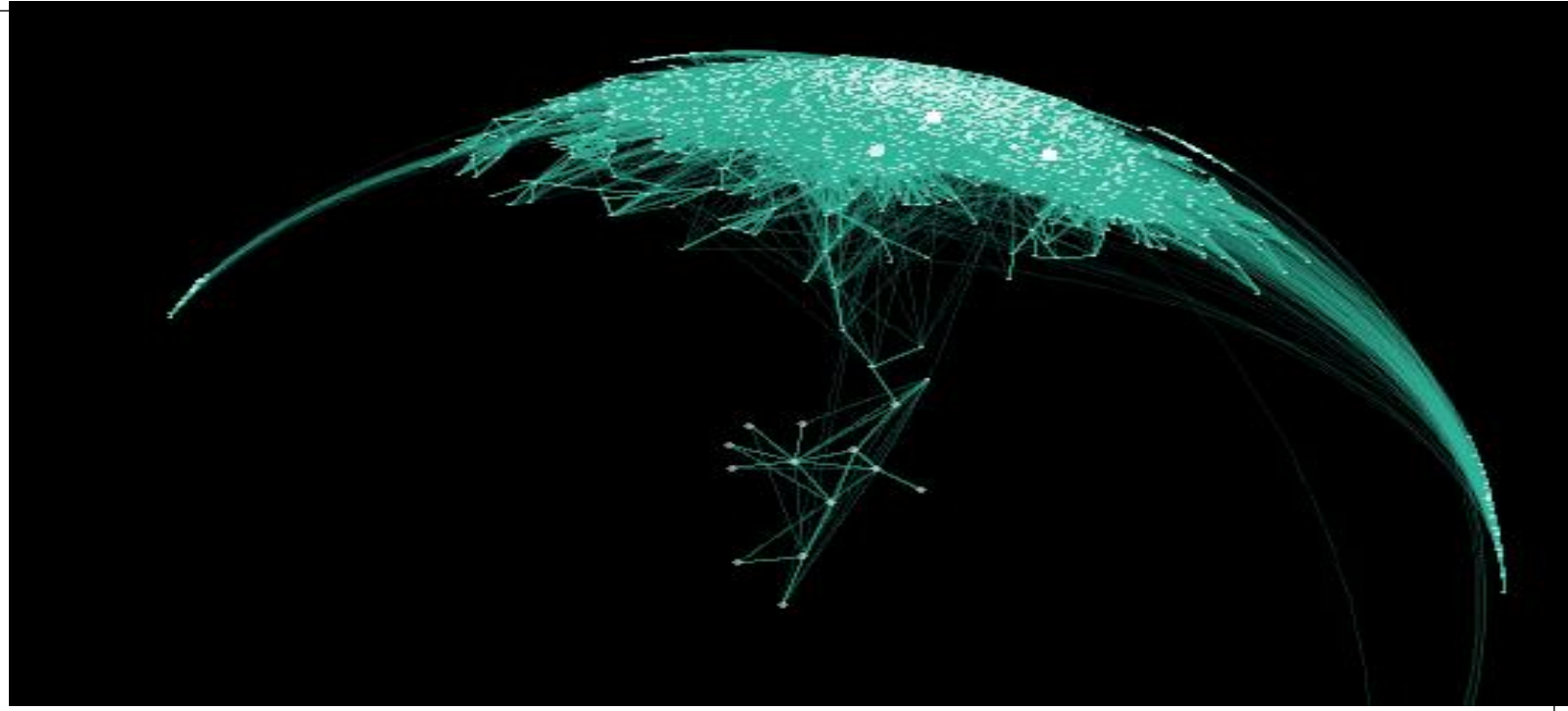
OLAP mining

- OLAP (On Line Analytical Processing)
  - Γρήγορη διαδραστική εξερεύνηση πολυδιάστατων συγκεντρωτικών δεδομένων
  - Βασίζεται σε μεγάλο βαθμό στη χειροκίνητη ανάλυση έμπειρων ανθρώπων
  - Επιρρεπής σε λάθη για μεγάλα δεδομένα υψηλής διαστατικότητας
- State of the art
  - Δέντρα απόφασης [Information discovery, Cognos]
  - Συσταδοποίηση [Pilot software]
  - Ανάλυση χρονοσειρών: [Seagate's Holos]
  - Πολύ επίπεδοι κανόνες αντιστοίχησης [Han et al.]
  - Sarawagi [VLDB2000

Εργασία και Εξόρυξη Δεδομένων

- Πηγή:
  - KDnuggets.com
  - Αφορά αγγελίες για εύρεση εργασίας στις ΗΠΑ
    - Μερικές από Ευρώπη, Ασία και Αυστραλία





Ενότητα 2: Δεδομένα

Τι ονομάζουμε δεδομένα;

- Συλλογή αντικειμένων και των ιδιοτήτων τους
- Μια ιδιότητα είναι ένα χαρακτηριστικό ενός αντικειμένου
  - Παραδείγματα: χρώμα ματιών, θερμοκρασία, εισόδημα, φύλο, κτλ.
  - Συνώνυμα: μεταβλητή, πεδίο, χαρακτηριστικό, γνώρισμα
- Μια συλλογή ιδιοτήτων περιγράφει ένα αντικείμενο
  - Συνώνυμα: εγγραφή, σημείο, περίπτωση, δείγμα, οντότητα ή στιγμιότυπο

Ιδιότητες

Αντικείμενα

Tid	Αποζημίωση	Οικ Κατάστ.	Εισόδημα	Απάτη
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

Τιμές ιδιοτήτων

- Οι τιμές είναι αριθμοί ή σύμβολα που ανατίθενται σε μια ιδιότητα
- Διαφορά τιμών από ιδιότητες
  - Μια ιδιότητα μπορεί να έχει διαφορετικές τιμές
    - Το ύψος μπορεί να εκφραστεί σε μέτρα ή πόδια
  - Διαφορετικές ιδιότητες μπορεί να έχουν κοινές τιμές
    - Η ηλικία και το ID είναι ακέραιοι
      - Έχουν όμως άλλα γνωρίσματα
        - Το ID δεν έχει όριο ενώ η ηλικία έχει μέγιστο και ελάχιστο όριο

Τύποι ιδιοτήτων

- Υπάρχουν διαφορετικοί τύποι ιδιοτήτων
  - Ονομαστικές (Nominal)
    - ΑΔΤ, ταχ. κώδικες, χρώμα ματιών, κτλ
  - Τακτικές (Ordinal)
    - Κλίμακες (π.χ. θετική γνώμη για ένα πολιτικό σε κλίμακα 1-10), βαθμοί, ύψος (ψηλός, κανονικός, κοντός)
  - Διαστήματα (Interval)
    - Ημερομηνίες, θερμοκρασίες Κελσίου…
  - Αναλογικές (Ratio)
    - Θερμοκρασίες Κέλβιν, μήκος, μετρήσεις

Χαρακτηριστικά των τιμών

- Ο τύπος μιας ιδιότητας εξαρτάται στα χαρακτηριστικά τα οποία τη διέπουν
  - Ατομικότητα = ≠
  - Τάξη < >
  - Πρόσθεση + -
  - Πολλαπλασιαστικότητα * /
- Nominal: Ατομικότητα
- Ordinal: Ατομικότητα & τάξη
- Interval: Ατομικότητα , τάξη & πρόσθεση
- Ratio : και τα 4 χαρακτηριστικά

Είδος Ιδιότητας	Μετασχηματισμός	Σχόλια
Nominal	Οποιοσδήποτε συνδυασμός	Εάν όλοι οι κωδικοί των υπαλλήλων ξανα-ανατεθούν, θα υπάρχει διαφορά;
Ordinal	Μια μονοτονική συνάρτηση f, τέτοια ώστε *νέα_τιμή= f(παλαιά_τιμής)*	Μια ιδιότητα που περιγράφει την έννοια καλό και κακό αναπαριστώνται εξίσου από τις τιμές{1,-1}.
Interval	*νέα_τιμή=a * παλαιά_τιμή+ b* όπου a και b σταθερές	Μετατροπή από Κελσίου σε Φαρενάιτ
Ratio	*νέα_τιμή = a * παλαιά_τιμή*	Το μήκος εκφράζεται σε πόδια ή μέτρα.

Διακριτές και συνεχείς ιδιότητες

- Διακριτές ιδιότητες
  - Έχει ένα πεπερασμένο σύνολο τιμών
    - Παραδείγματα: λέξεις, ταχυδρομικοί κωδικοί, χώρες, ονόματα, κτλ.
  - Συχνά αναπαρίστανται ως ακέραιοι
    - Σημείωση: οι δυαδικές ιδιότητες αποτελούν ειδική περίπτωση διακριτών ιδιοτήτων
- Συνεχείς Ιδιότητες
  - Έχουν πραγματικούς αριθμούς ως τιμές
    - Παραδείγματα: θερμοκρασία, ύψος, χρόνος, κτλ.
  - Πρακτικά, οι πραγματικοί αριθμοί αναπαρίστανται από πεπερασμένο αριθμό ψηφίων.
  - Οι συνεχείς ιδιότητες αναπαρίστανται ως μεταβλητές κινητής υποδιαστολής.

Σώματα δεδομένων

- **Εγγραφές**
  - **Πίνακας δεδομένων**
  - **Έγγραφα**
  - **Δεδομένα συναλλαγών**
- **Γράφοι**
  - **World Wide Web**
  - **Μοριακές δομές**
- **Ταξινομημένα**
  - **Χωρικά δεδομένα**
  - **Χρονικά δεδομένα**
  - **Σειριακά δεδομένα**
  - **Γενετικές σειρές**

- **Χαρακτηριστικά**
  - **Διαστατικότητα**
    - **Η κατάρα της διαστατικότητας**
  - **Αραιότητα**
    - **Μετράει μόνο η παρουσία**
  - **Ανάλυση**
    - **Τα πρότυπα εξαρτώνται από την κλίμακα**

Εγγραφές

- Δεδομένα που αποτελούνται από συλλογές αντικειμένων που περιέχουν ένα κλειστό σύνολο ιδιοτήτων

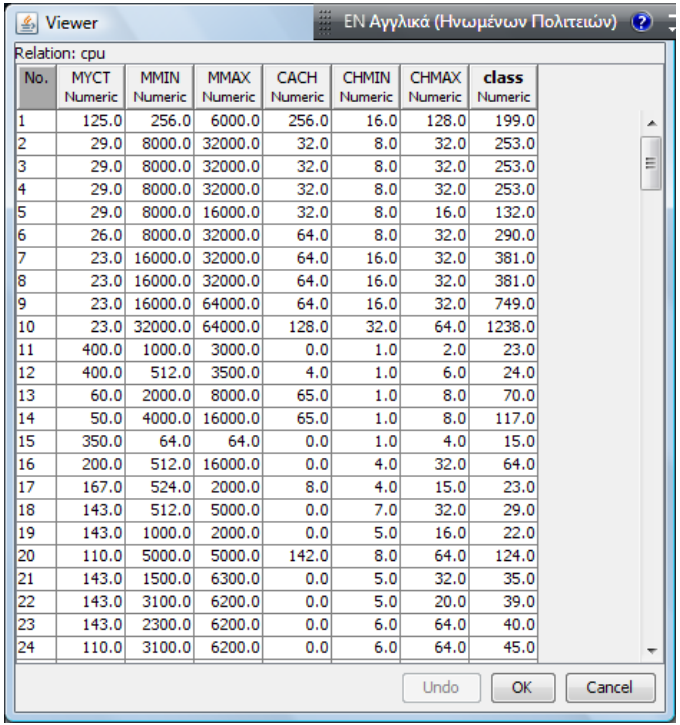
Πηγή: WEKA, University of Waikato

Relation: contact-lenses

No.	age Nominal	spectacle-prescrip Nominal	astigmatism Nominal	tear-prod-rate Nominal	**contact-lenses** Nominal
1	young	myope	no	reduced	none
2	young	myope	no	normal	soft
3	young	myope	yes	reduced	none
4	young	myope	yes	normal	hard
5	young	hypermetrope	no	reduced	none
6	young	hypermetrope	no	normal	soft
7	young	hypermetrope	yes	reduced	none
8	young	hypermetrope	yes	normal	hard
9	pre-pr...	myope	no	reduced	none
10	pre-pr...	myope	no	normal	soft
11	pre-pr...	myope	yes	reduced	none
12	pre-pr...	myope	yes	normal	hard
13	pre-pr...	hypermetrope	no	reduced	none
14	pre-pr...	hypermetrope	no	normal	soft
15	pre-pr...	hypermetrope	yes	reduced	none
16	pre-pr...	hypermetrope	yes	normal	none
17	presb...	myope	no	reduced	none
18	presb...	myope	no	normal	none
19	presb...	myope	yes	reduced	none
20	presb...	myope	yes	normal	hard
21	presb...	hypermetrope	no	reduced	none
22	presb...	hypermetrope	no	normal	soft
23	presb...	hypermetrope	yes	reduced	none
24	presb...	hypermetrope	yes	normal	none

Πίνακας δεδομένων

- Όταν τα δεδομένα έχουν ένα κοινό κλειστό σύνολο αριθμητικών ιδιοτήτων, μπορούν να θεωρηθούν ως σημεία σε ένα πολύ-διάστατο χώρο, όπου κάθε διάσταση αναπαριστά μια συγκεκριμένη μεταβλητή
- Τέτοια δεδομένα αναπαρίστανται ως ένας m x n πίνακας, όπου m η κάθε σειρά (αντικείμενο) και n η κάθε στήλη (ιδιότητα)



No.	MYCT Numeric	MMIN Numeric	MMAX Numeric	CACH Numeric	CHMIN Numeric	CHMAX Numeric	class Numeric
1	125.0	256.0	6000.0	256.0	16.0	128.0	199.0
2	29.0	8000.0	32000.0	32.0	8.0	32.0	253.0
3	29.0	8000.0	32000.0	32.0	8.0	32.0	253.0
4	29.0	8000.0	32000.0	32.0	8.0	32.0	253.0
5	29.0	8000.0	16000.0	32.0	8.0	16.0	132.0
6	26.0	8000.0	32000.0	64.0	8.0	32.0	290.0
7	23.0	16000.0	32000.0	64.0	16.0	32.0	381.0
8	23.0	16000.0	32000.0	64.0	16.0	32.0	381.0
9	23.0	16000.0	64000.0	64.0	16.0	32.0	749.0
10	23.0	32000.0	64000.0	128.0	32.0	64.0	1238.0
11	400.0	1000.0	3000.0	0.0	1.0	2.0	23.0
12	400.0	512.0	3500.0	4.0	1.0	6.0	24.0
13	60.0	2000.0	8000.0	65.0	1.0	8.0	70.0
14	50.0	4000.0	16000.0	65.0	1.0	8.0	117.0
15	350.0	64.0	64.0	0.0	1.0	4.0	15.0
16	200.0	512.0	16000.0	0.0	4.0	32.0	64.0
17	167.0	524.0	2000.0	8.0	4.0	15.0	23.0
18	143.0	512.0	5000.0	0.0	7.0	32.0	29.0
19	143.0	1000.0	2000.0	0.0	5.0	16.0	22.0
20	110.0	5000.0	5000.0	142.0	8.0	64.0	124.0
21	143.0	1500.0	6300.0	0.0	5.0	32.0	35.0
22	143.0	3100.0	6200.0	0.0	5.0	20.0	39.0
23	143.0	2300.0	6200.0	0.0	6.0	64.0	40.0
24	110.0	3100.0	6200.0	0.0	6.0	64.0	45.0

Δεδομένα Εγγράφων

- Κάθε έγγραφο μετατρέπεται σε ένα διάνυσμα όρων
  - Κάθε όρος είναι ένα συστατικό (ιδιότητα) του διανύσματος
  - Η τιμή κάθε συστατικού είναι ο αριθμός εμφανίσεων του όρου στο έγγραφο

ΠΑΡΑΔΕΙΓΜΑ
D1: "Shipment of gold damaged in a fire"
D2: "Delivery of silver arrived in a silver truck"
D3: "Shipment of gold arrived in a truck

Doc ID	a	arrived	damaged	delivery	fire	gold	in	of	shipment	silver	truck
D1	1	0	1	0	1	1	1	1	1	0	0
D2	1	0	0	1	0	0	1	1	0	2	1
D3	1	1	0	0	0	1	1	1	1	0	1

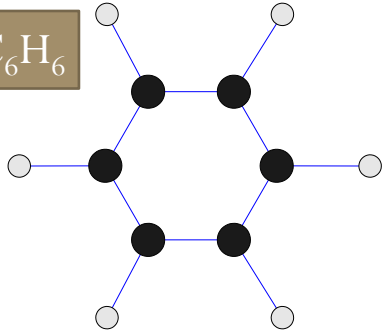
Δεδομένα συναλλαγών

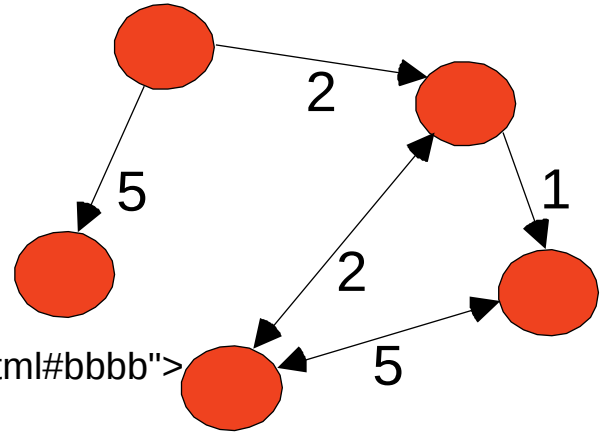
- Μια ειδική περίπτωση εγγραφών όπου
  - Κάθε εγγραφή (ή συναλλαγή) έχει ένα σύνολο αντικειμένων
  - Για παράδειγμα, θεωρήστε ένα σούπερ μάρκετ. Το σύνολο των προϊόντων αγοράς από ένα πελάτη αποτελούν δεδομένα συναλλαγών

TID	*Αντικείμενα*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Γράφοι

Μόριο βενζίνης C_6H_6





- Παραδείγματα:
  - Γενικευμένοι γράφοι (κυκλοφορίας, φορτίου σε δίκτυο, κτλ) όπως επίσης και έγγραφα σε HTML ή XML ή άλλης markup μορφής

```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

Ταξινομημένα

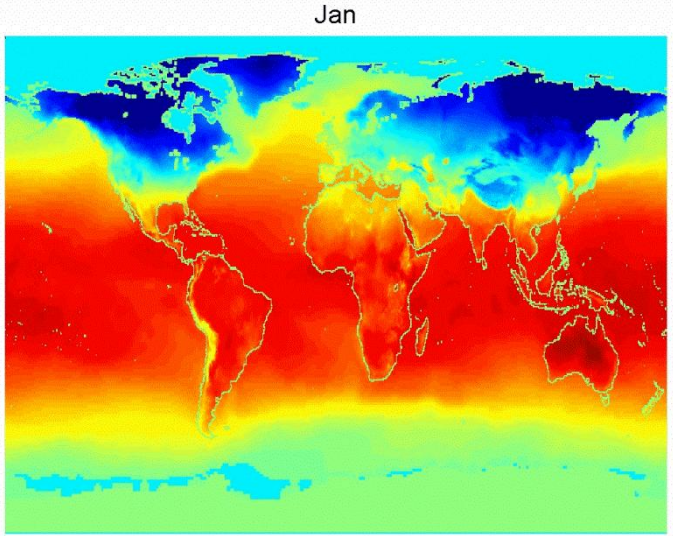
- Αλληλουχίες συμβάντων
- Χώρο-χρονικά δεδομένα

Αντικείμενα/Συμβάντα

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

Ένα συμβάν της ακολουθίας

Μέσες μηνιαίες θερμοκρασίες στην ξηρά και τον ωκεανό



Ποιότητα δεδομένων

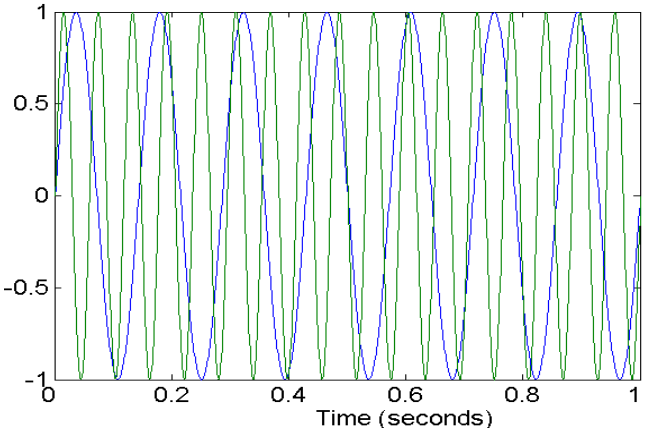
- Τι είδους προβλήματα ποιότητας υπάρχουν;
- Πώς ανιχνεύουμε προβλήματα στα δεδομένα
- Πως αντιμετωπίζουμε τα προβλήματα αυτά

- Παραδείγματα προβλημάτων ποιότητας δεδομένων
  - Θόρυβος και εξαιρέσεις
  - Ανύπαρκτες τιμές
  - Διπλότυπα

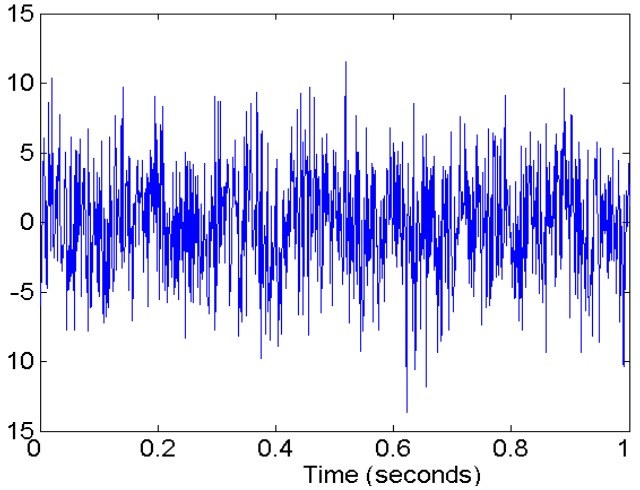
Θόρυβος

- Ο θόρυβος αναφέρεται σε αλλοιώσεις των αρχικών τιμών μιας ιδιότητας
  - Παράδειγμα: παραμόρφωση στο σήμα ομιλίας λόγω κακής ποιότητας ενός μικροφώνου, χιόνια σε σήμα τηλεόρασης λόγω κακής λήψης, κτλ

2 ημιτονοειδή σήματα

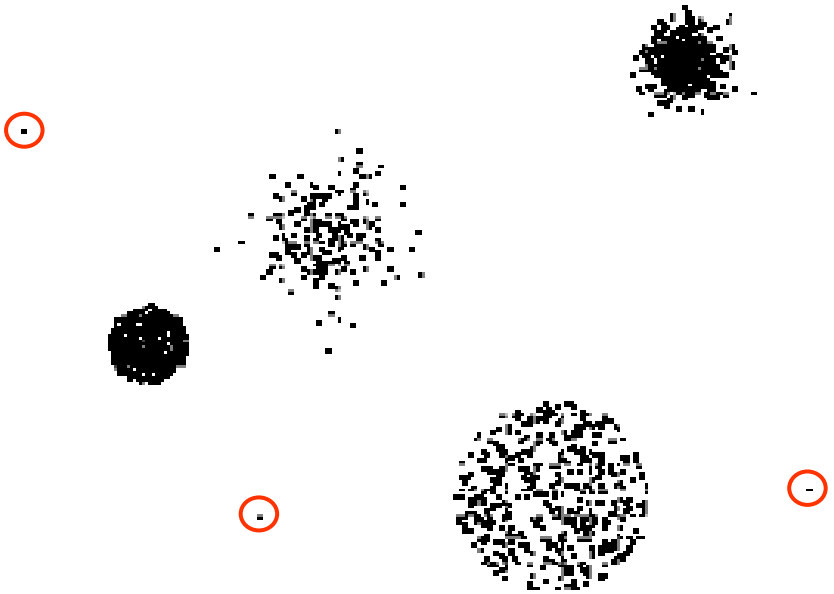


2 ημιτονοειδή σήματα **με θόρυβο**



Εξαιρέσεις

- Οι εξαιρέσεις (outliers) είναι δεδομένα με χαρακτηριστικά που διαφέρουν σημαντικά από τα περισσότερα δεδομένα του συνόλου



Ανύπαρκτες τιμές

- Λόγοι ανύπαρκτων τιμών
  - Δεν συλλέγεται η πληροφορία
    - Π.χ. ορισμένοι αρνούνται να αποκαλύψουν την ηλικία ή το βάρος τους
  - Οι ιδιότητες δεν χαρακτηρίζουν όλα τα δεδομένα
    - Π.χ. το εισόδημα δεν εφαρμόζεται στα παιδιά
- Πως χειριζόμαστε τις ανύπαρκτες τιμές
  - Αφαίρεση δεδομένων που τις περιέχουν
  - Εκτίμηση των τιμών
  - Να τις αγνοήσουμε
  - Αντικατάσταση με όλες τις πιθανές τιμές (με στάθμιση των πιθανοτήτων τους)

Διπλότυπα

- Το σύνολο δεδομένων μπορεί να περιλαμβάνει δεδομένα με διπλές εγγραφές ή σχεδόν διπλότυπα
  - Μεγάλο ζήτημα όταν συνενώνονται δεδομένα από ετερογενείς πηγές
- Παραδείγματα:
  - Ένα άτομο με πολλές διευθύνσεις ηλ. ταχυδρομείου
- Καθαρισμός δεδομένων
  - Η διαδικασία εξομάλυνσης του προβλήματος των διπλότυπων

Προ-επεξεργασία δεδομένων

- Συνάθροιση
- Δειγματοληψία
- Μείωση Διαστατικότητας
- Επιλογή υποσυνόλου ιδιοτήτων
- Δημιουργία ιδιοτήτων
- Διακριτοποίηση και Δυαδοποίηση
- Μετασχηματισμός Ιδιοτήτων

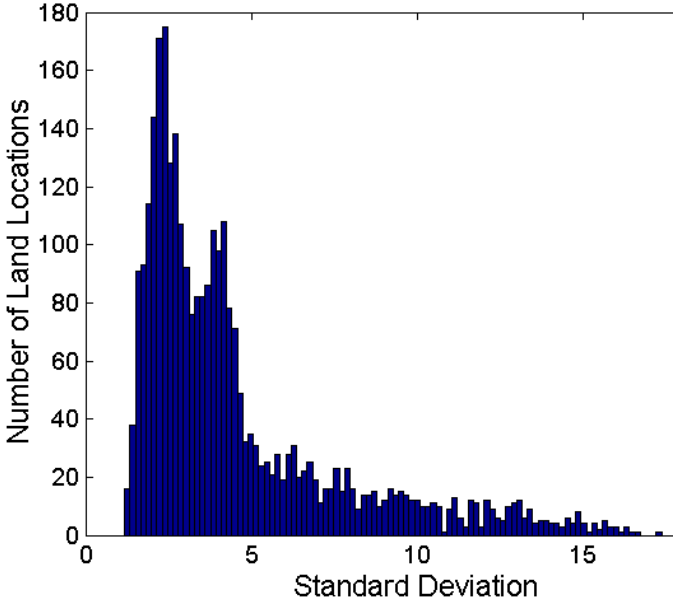
Συνάθροιση

- Ο συνδυασμός 2 ή περισσοτέρων ιδιοτήτων (ή αντικειμένων) σε μια μοναδική ιδιότητα (ή αντικείμενο)

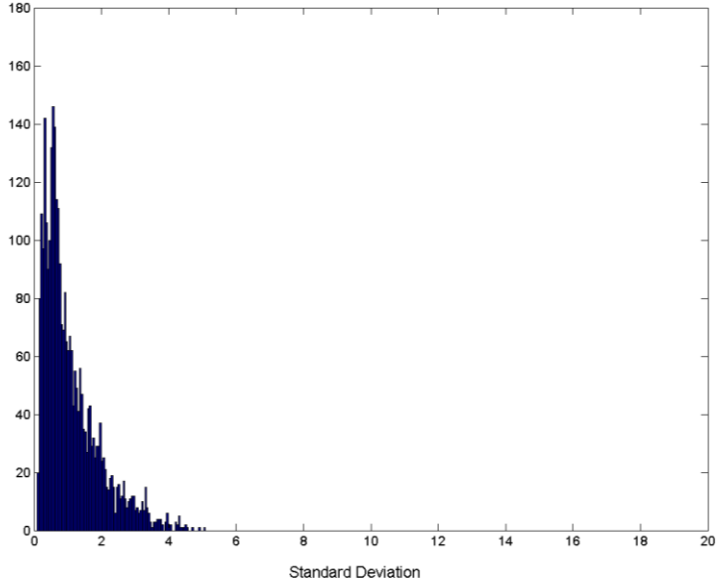
- Σκοπός:
  - Μείωση δεδομένων
    - Μείωση του αριθμού των ιδιοτήτων ή αντικειμένων
  - Αλλαγή κλίμακας
    - Οι πόλεις συναθροίζονται σε νομούς, περιφέρειες, χώρες, κτλ
  - Πιο «σταθερά» δεδομένα
    - Τα συναθροισμένα δεδομένα τείνουν να έχουν λιγότερη μεταβλητότητα

Συνάθροιση

Μεταβολή ατμοσφαιρικής πίεσης στην Αυστραλία



Μηνιαία



Ετήσια

Δειγματοληψία

- Είναι η κύρια τεχνική για επιλογή δεδομένων
  - Χρησιμοποιείται τόσο σε προκαταρκτικές έρευνες όσο και στην τελική ανάλυση δεδομένων
- Οι στατιστικολόγοι κάνουν δειγματοληψία επειδή η απόκτηση όλων των δεδομένων είναι ακριβή ή χρονοβόρα
- Η δειγματοληψία χρησιμοποιείται στην εξόρυξη δεδομένων επειδή η επεξεργασία όλων των δεδομένων μπορεί να είναι υπολογιστικά απαγορευτική

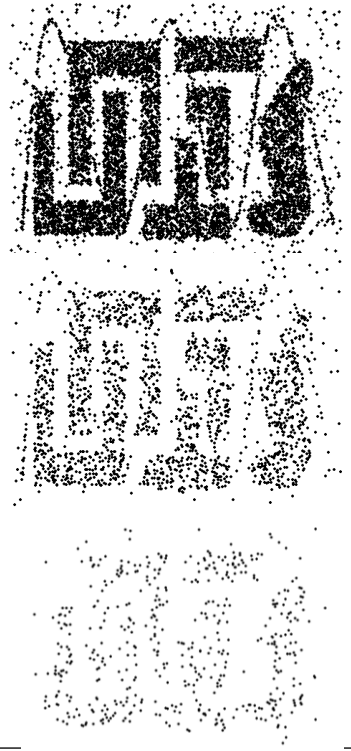
- Αρχή αποτελεσματικής δειγματοληψίας
  - Αν το δείγμα είναι αντιπροσωπευτικό, θα δουλέψει το ίδιο καλά με το συνολικό σώμα δεδομένων
  - Ένα δείγμα είναι αντιπροσωπευτικό εάν έχει προσεγγιστικά την ίδια ιδιότητα με το αρχικό σύνολο δεδομένων (π.χ. κατανομή)

Τύποι δειγματοληψίας

- Απλή τυχαία δειγματοληψία
  - Υπάρχει ίδια πιθανότητα επιλογής ενός αντικειμένου
- Δειγματοληψία χωρίς αντικατάσταση
  - Όταν επιλέγεται ένα αντικείμενο, αφαιρείται από τον πληθυσμό
    - Επιλέγεται μόνο μια φορά
- Δειγματοληψία με αντικατάσταση
  - Το ίδιο αντικείμενο μπορεί να επιλεχθεί περισσότερο από μια φορά
- Διαστρωματική δειγματοληψία
  - Διαχωρισμός των δεδομένων σε τομείς, και επιλογή τυχαίων αντικειμένων από κάθε τομέα

Μέγεθος Δείγματος

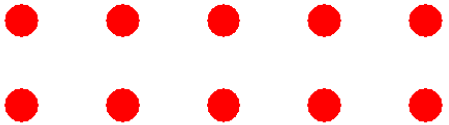
Τι μέγεθος δείγματος χρειάζεται για να επιλεχθεί τουλάχιστον ένα αντικείμενο από κάθε ένα από τα 10 σύνολα;

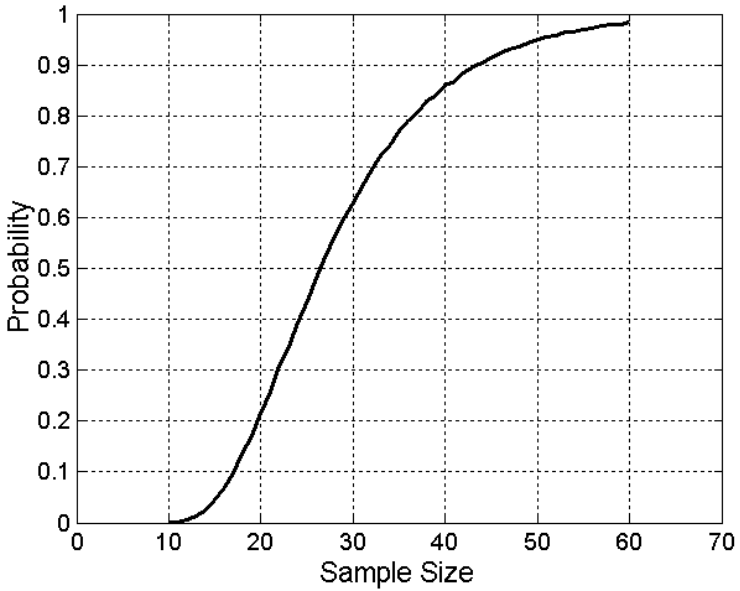


8000 σημεία

2000 σημεία

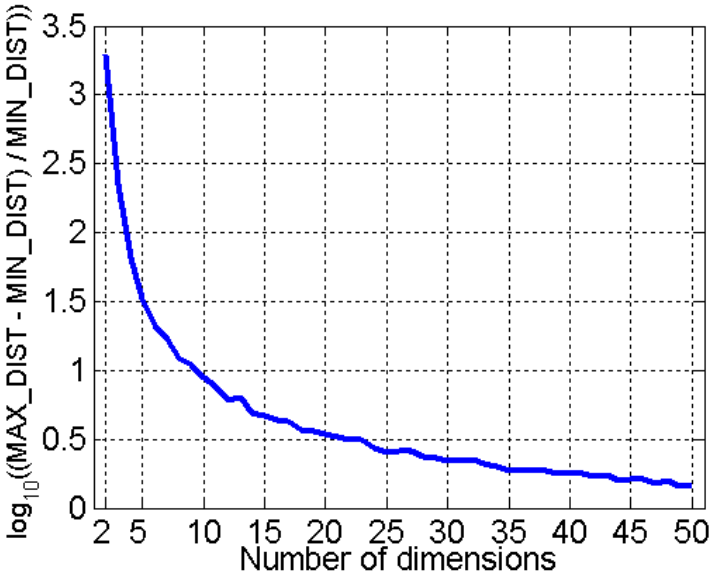
500 σημεία





Η κατάρα της διαστατικότητας

- Όταν αυξάνουν οι διαστάσεις (ιδιότητες), τα δεδομένα γίνονται ολοένα και πιο αραιά στον πολυδιάστατο χώρο
- Η απόσταση και η πυκνότητα μεταξύ των σημείων που είναι κρίσιμες για την ομαδοποίηση ή την ανίχνευση εξαιρέσεων χάνουν τη σημασία τους



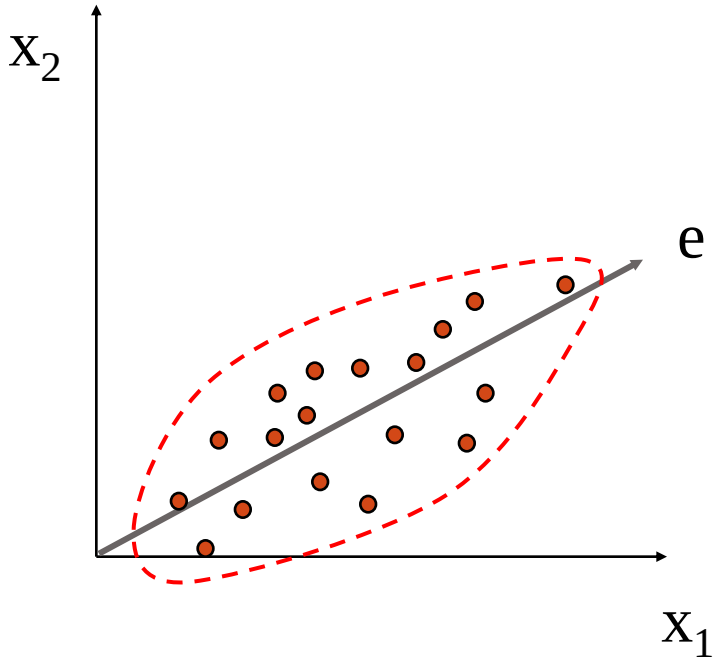
- **Τυχαία δημιουργία 500 σημείων**
- **Υπολογισμός της διαφοράς μεταξύ μέγιστης και ελάχιστης απόστασης μεταξύ όλων των ζευγών σημείων**

Μείωση διαστατικότητας

- Σκοπός:
  - Αποφυγή της «κατάρας» της διαστατικότητας
  - Μείωση χρόνου και χώρου για τους αλγόριθμους εξόρυξης δεδομένων
  - Επιτρέπει την οπτικοποίηση των δεδομένων
  - Βοηθάει στην αφαίρεση άσχετων ιδιοτήτων ή στη μείωση του θορύβου
- Τεχνικές
  - Principle Component Analysis (PCA)
  - Singular Value Decomposition (SVD)
  - Άλλες: τεχνικές επίβλεψης και μη-γραμμικές τεχνικές

Μείωση διαστατικότητας: PCA

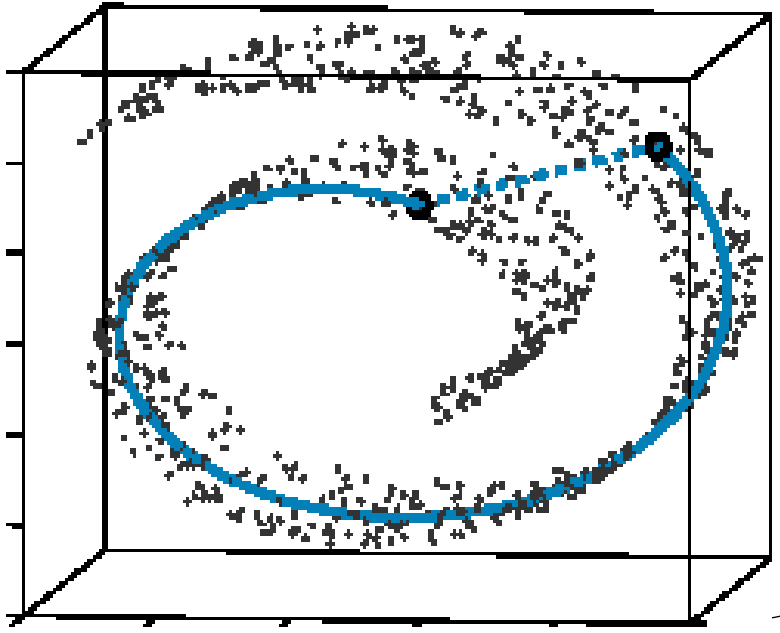
- Ο στόχος είναι η εύρεση μιας προβολής που αιχμαλωτίζει τη μεγαλύτερη ποσότητα διασποράς των δεδομένων



Μείωση διαστατικότητας: ISOMAP

- Κατασκευή ενός γράφου γειτόνων
- Για κάθε ζεύγος σημείων του γράφου, υπολόγισε το συντομότερο μονοπάτι αποστάσεων-γεωδαιτικών αποστάσεων

Από: Tenenbaum, de Silva, Langford (2000)



Επιλογή υποσυνόλου ιδιοτήτων

- Εναλλακτικός τρόπος μείωσης της διαστατικότητας
- Πλεονάζοντα χαρακτηριστικά
  - Περιγράφουν πολύ πληροφορία που ήδη υπάρχει σε άλλες ιδιότητες
    - Παράδειγμα: η τιμή ενός προϊόντος και η τιμή του φόρου που του αναλογεί
- Μη σχετικά χαρακτηριστικά
  - Δεν περιέχουν πληροφορία που μπορεί να αξιοποιηθεί για την εφαρμογή που ενδιαφέρει
    - Παράδειγμα: το ΑΦΜ ενός πολίτη είναι άσχετο για την εφαρμογή πρόβλεψης της πιστοληπτικής ικανότητας του

- Τεχνικές
  - Εξαντλητική (Brute-force):
    - Δοκιμή όλων των πιθανών υποσυνόλων
  - Ενσωματωμένες:
    - Η επιλογή γίνεται εσωτερικά από τη φύση του αλγόριθμου εξόρυξης δεδομένων
  - Φίλτρα:
    - Οι ιδιότητες επιλέγονται πριν εκτελεστεί ο αλγόριθμος εξόρυξης δεδομένων

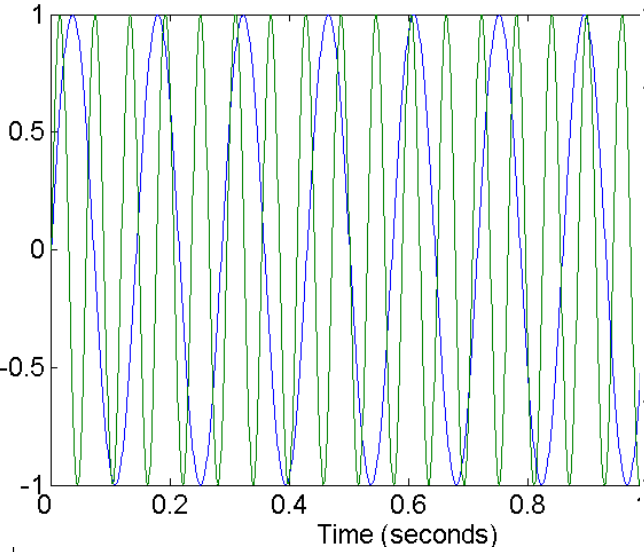
Δημιουργία ιδιοτήτων

- Δημιουργία νέων ιδιοτήτων που μπορούν να αιχμαλωτίζουν σημαντική πληροφορία ενός συνόλου δεδομένων πιο αποτελεσματικά από τις αρχικές ιδιότητες

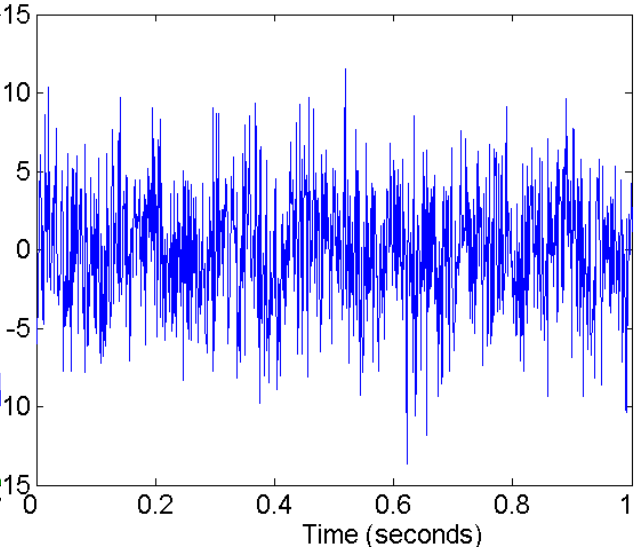
- 3 κύριες προσεγγίσεις
  - Εξαγωγή Ιδιοτήτων από ειδήμονες
    - Π.χ. στην ιατρική
  - Αντιστοίχιση δεδομένων σε νέο χώρο
  - Συνδυασμός ιδιοτήτων
    - Γραμμικός, με χρήση πιθανοτήτων, κτλ.

Αντιστοίχηση σε νέο χώρο

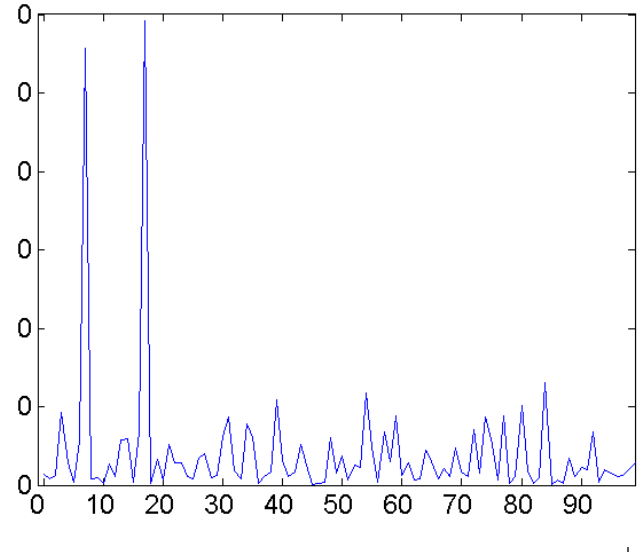
- Μετασχηματισμός Fourier
- Wavelets



Δυο ημιτονοειδή κύματα



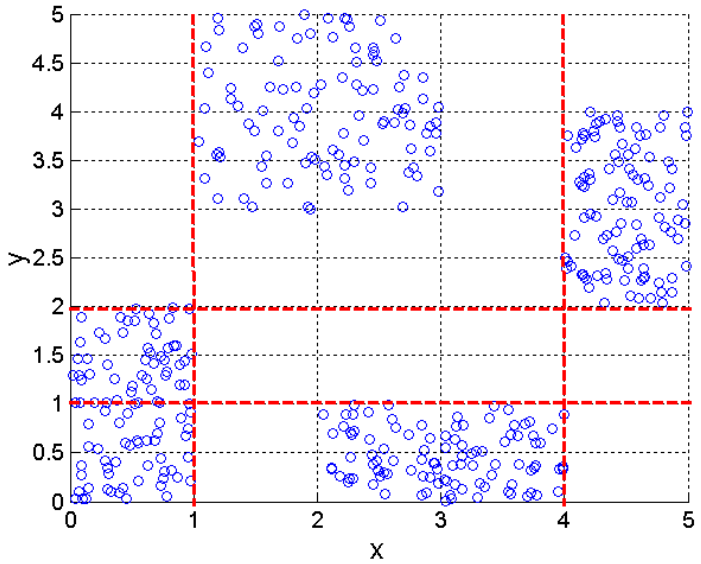
Δυο ημιτονοειδή με θόρυβο



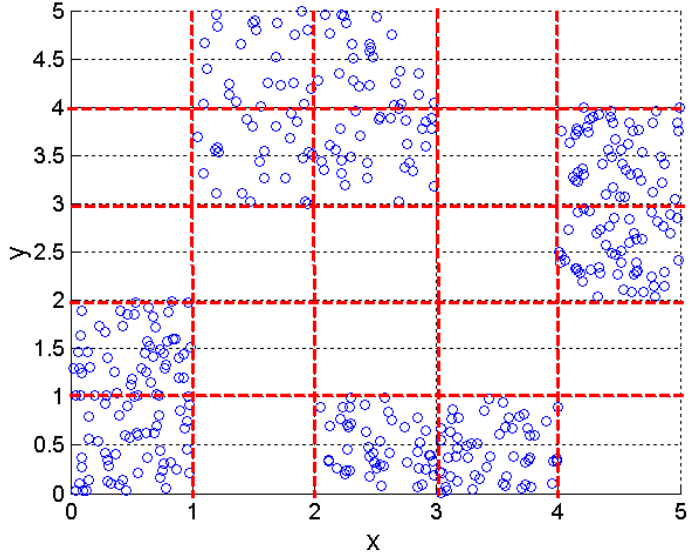
Συχνότητα

Διακριτοποίηση με τιμές κλάσης

- Με βάση την εντροπία

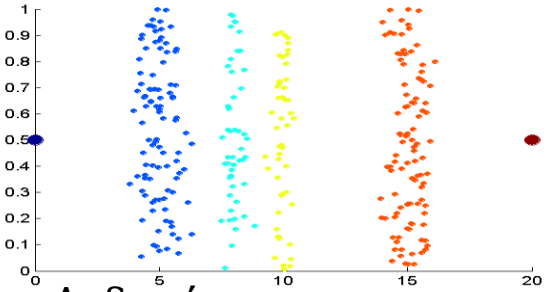


3 κατηγορίες ανά x και y

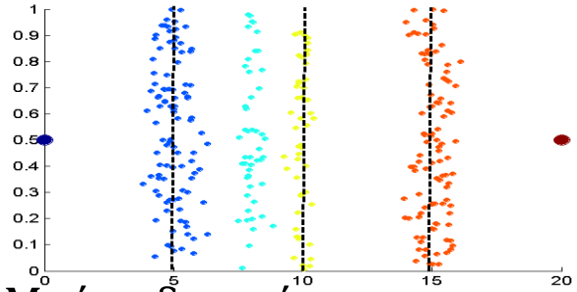


5 κατηγορίες ανά x και y

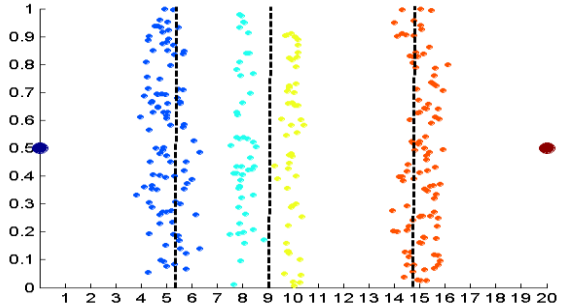
Διακριτοποίηση χωρίς τιμές κλάσης



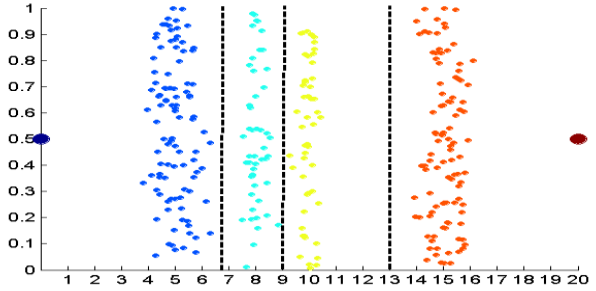
Δεδομένα



Με ίσα διαστήματα



Ίσες συχνότητες



K-means

Μετασχηματισμός ιδιοτήτων

- Μια συνάρτηση που αντιστοιχίζει το σύνολο των τιμών μιας ιδιότητας σε ένα νέο σύνολο τιμών, τέτοιων ώστε κάθε παλαιά τιμή να ταυτοποιείται από μια νέα
  - Απλές συναρτήσεις
    - x^k, $\log(x)$, e^x, $|x|$
  - Συναρτήσεις κανονικοποίησης

Ομοιότητες και ανομοιότητες

- Ομοιότητα
  - Αριθμητική μέτρηση του πόσο μοιάζουν δυο αντικείμενα
  - Μεγαλώνει όταν τα αντικείμενα μοιάζουν
  - Συχνά εμπίπτει στο όριο [0,1]
- Ανομοιότητα
  - Αριθμητική μέτρηση του πόσο διαφορετικά είναι δυο αντικείμενα
  - Μειώνεται όταν τα αντικείμενα μοιάζουν
  - Η ελάχιστη ανομοιότητα είναι συχνά 0
  - Το ανώτατο όριο διαφέρει από περίπτωση σε περίπτωση

Η εγγύτητα αναφέρεται στην ομοιότητα ή την ανομοιότητα

Ομοιότητες και μη για απλές ιδιότητες

- Έστω p,q δυο ιδιότητες:

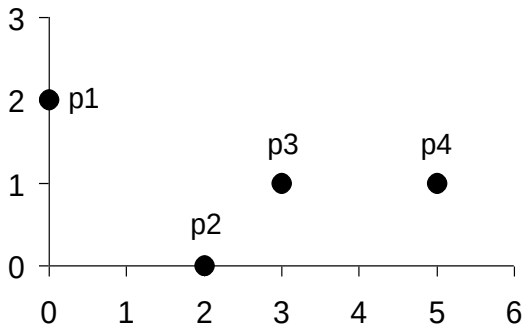
Τύπος ιδιότητας	Ανομοιότητα (d)	Ομοιότητα (s)
Nominal	$d = \begin{cases} 0 \text{ αν } p=q \\ 1 \text{ αν } p \neq q \end{cases}$	$s = \begin{cases} 1 \text{ αν } p=q \\ 0 \text{ αν } p \neq q \end{cases}$
Ordinal	Οι τιμές κυμαίνονται από 0 έως n-1, όπου n ο αριθμός των τιμών $d = \frac{\lvert p-q \rvert}{n-1}$	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval ή Ratio	$d = \lvert p-q \rvert$	$s = -d,\ s = \frac{1}{1+d}$, ή $s = 1 - \frac{\text{d-min}_d}{\max_d - \min_d}$

Ευκλείδεια απόσταση

- Τύπος:

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

- Όπου n ο αριθμός των διαστάσεων (ιδιότητες) και p_k, q_k η k-οστή ιδιότητα των αντικειμένων p, q
- Η κανονικοποίηση επιβάλλεται εάν διαφέρουν οι κλίμακες



σημείο	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

Απόσταση Minkowski

- Τύπος

$$d = \left(\sum_{k=1}^{n} | p_k - q_k |^r\right)^{\frac{1}{r}}$$

- Όπου r μια παράμετρος, n ο αριθμός των διαστάσεων, και p_k, q_k η k-οστή ιδιότητα των αντικειμένων p, q

$r = 1$. απόσταση City block (Manhattan, taxicab, L_1 norm).

"Ενα κοινό παράδειγμα είναι η απόσταση Hamming που είναι απλά ο αριθμός των bits που διαφέρουν μεταξύ δυο δυαδικών διανυσμάτων

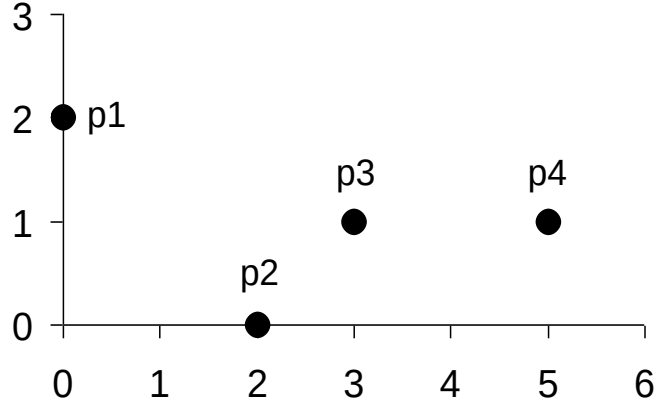
$r = 2$. Ευκλείδεια απόσταση

$r \to \infty$. "supremum" (L_{max} norm, L_∞ norm) απόσταση.

Η μέγιστη διαφορά μεταξύ οποιασδήποτε διάστασης των διανυσμάτων

Μη συγχέετε το r με το n.
Όλες οι παραπάνω αποστάσεις ορίζονται για όλες τις διαστάσεις

Απόσταση Minkowski



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

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	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

L∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Απόσταση Mahalanobis

- Η απόσταση Mahalanobis μεταξύ των διανυσμάτων χ και y είναι

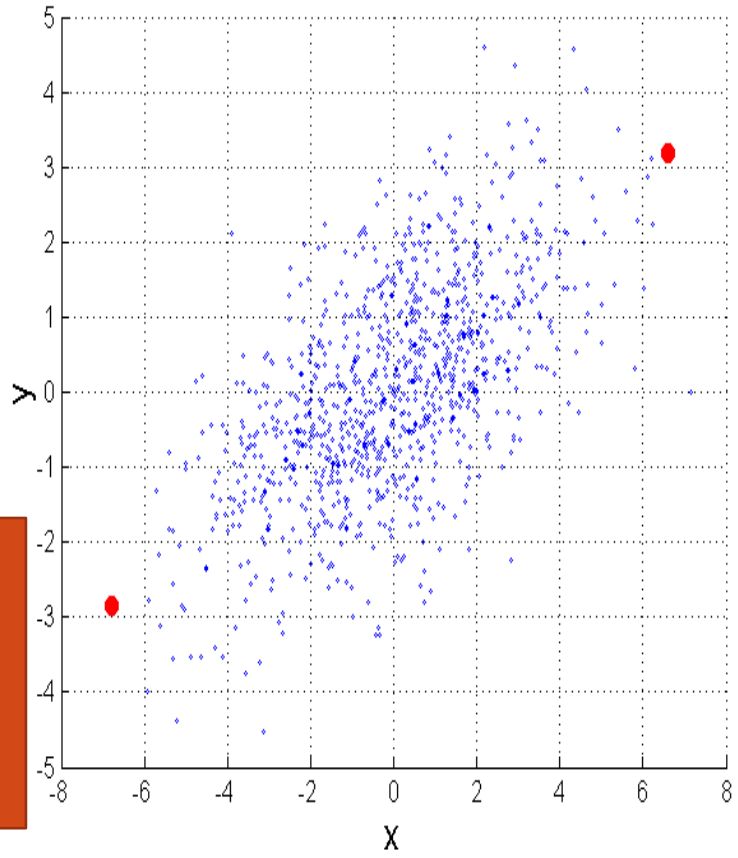
$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})}.$$

- Όπου S ο πίνακας συν-διασποράς των δεδομένων X

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})^T,$$

Μέση τιμή των τιμών του χαρακτηριστικού χ

- *Όπου x_i είναι η i-οστή εμφάνιση του χαρακτηριστικού χ, (στο i-οστό παράδειγμα), n είναι το πλήθος των παραδειγμάτων*



Για τα κόκκινα σημεία, η Ευκλείδεια απόσταση είναι 14.7, η απόσταση Mahalanobis είναι 6.

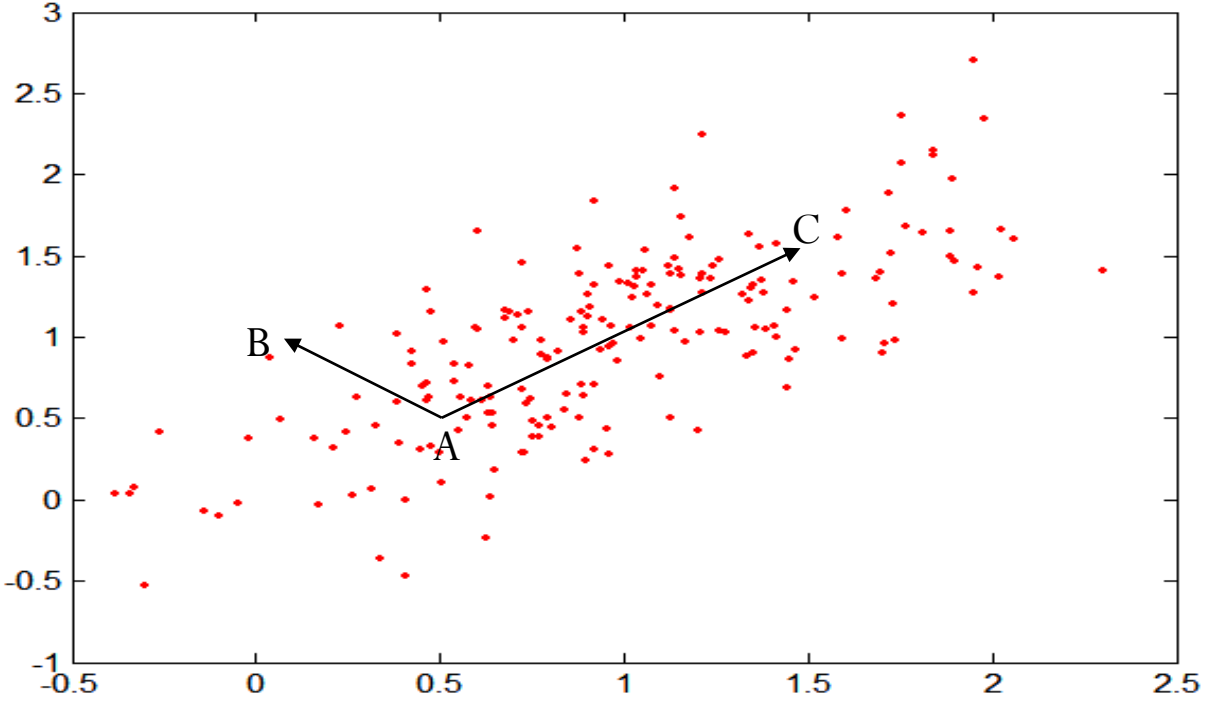
Ο πίνακας συνδιασποράς (covariance matrix) ενός τυχαίου διανύσματος **X** ορίζεται ως:

$$Cov(\mathbf{X}) = \mathbf{\Sigma} = E\{(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\}$$

$$\Sigma_{ij} = \operatorname{cov}(X_i, X_j) = \mathrm{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathrm{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathrm{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathrm{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}.$$

Απόσταση Mahalanobis



Πίνακας συν-διασποράς:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Ιδιότητες της απόστασης

- Οι αποστάσεις, όπως η Ευκλείδεια, έχουν ορισμένες γνωστές ιδιότητες
  1. $d(p, q) \geq 0$ για όλα τα p και q και $d(p, q) = 0$ μόνο αν $p = q$.
  2. $d(p, q) = d(q, p)$ για όλα τα p και q. (Συμμετρία)
  3. $d(p, r) \leq d(p, q) + d(q, r)$ για όλα τα p, q, και r. (Τριγωνική ανισότητα)

- Όποια απόσταση ικανοποιεί τις ιδιότητες αυτές ονομάζεται μετρική

Ομοιότητα μεταξύ δυαδικών διανυσμάτων

- Μια κοινή περίπτωση είναι αυτή όπου δυο αντικείμενα p, q έχουν μόνο δυαδικές ιδιότητες
- Υπολογίστε τις ιδιότητες με βάση τα παρακάτω:
  - M_{01} = # ιδιοτήτων όπου το p ήταν 0 και το q 1
  - M_{10} = # ιδιοτήτων όπου το p ήταν 1 και το q 0
  - M_{00} = # ιδιοτήτων όπου το p ήταν 0 και το q 0
  - M_{11} = # ιδιοτήτων όπου το p ήταν 1 και το q 1

- Απλό ταίριασμα (simple matching) και συντελεστές Jaccard
- SM = αριθμών συμφωνιών/ αριθμός ιδιοτήτων

$= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

- Jaccard = αριθμός των ταιριασμάτων με άσσους / αριθμός όλων των μη μηδενικών τιμών

$= (M_{11}) / (M_{01} + M_{10} + M_{11})$

Παράδειγμα SM και Jaccard

$p = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$

$q = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$

$M_{01} = 2$
$M_{10} = 1$
$M_{00} = 7$
$M_{11} = 0$

$SM = (M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$

$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$

Ομοιότητα συνημίτονου

- Αν d_1 και d_2 είναι δυο διανύσματα εγγράφων τότε:

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2|| ,$$

όπου • σημαίνει εσωτερικό γινόμενο και $||d||$ το μήκος του διανύσματος d.

- Παράδειγμα:

$d_1 =$ **3 2 0 5 0 0 0 2 0 0**

$d_2 =$ **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$\cos(d_1, d_2) = .3150$

Συντελεστής Tanimoto (Extended Jaccard Coefficient)

- Καλός για δεδομένα εγγράφων, αντιμετωπίζει καλά το sparsity

$$EJ(d_1, d_2) = (d_1 \bullet d_2) / (||d_1||^2 + ||d_2||^2 - d_1 \bullet d_2)$$

Συσχέτιση (Correlation)

- Η συσχέτιση μετρά τη γραμμική σχέση μεταξύ αντικειμένων
- Πάντα μεταξύ -1 και 1
  - X=a*Y+b
- Συχέτιση Pearson

$$corr(x,y) = \frac{covariance(x,y)}{stddev(x) * stddev(y)}$$

$$covariance(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x_m)(y_i - y_m)$$

$$stddev(x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x_m)^2}$$

Συσχέτιση (Correlation)

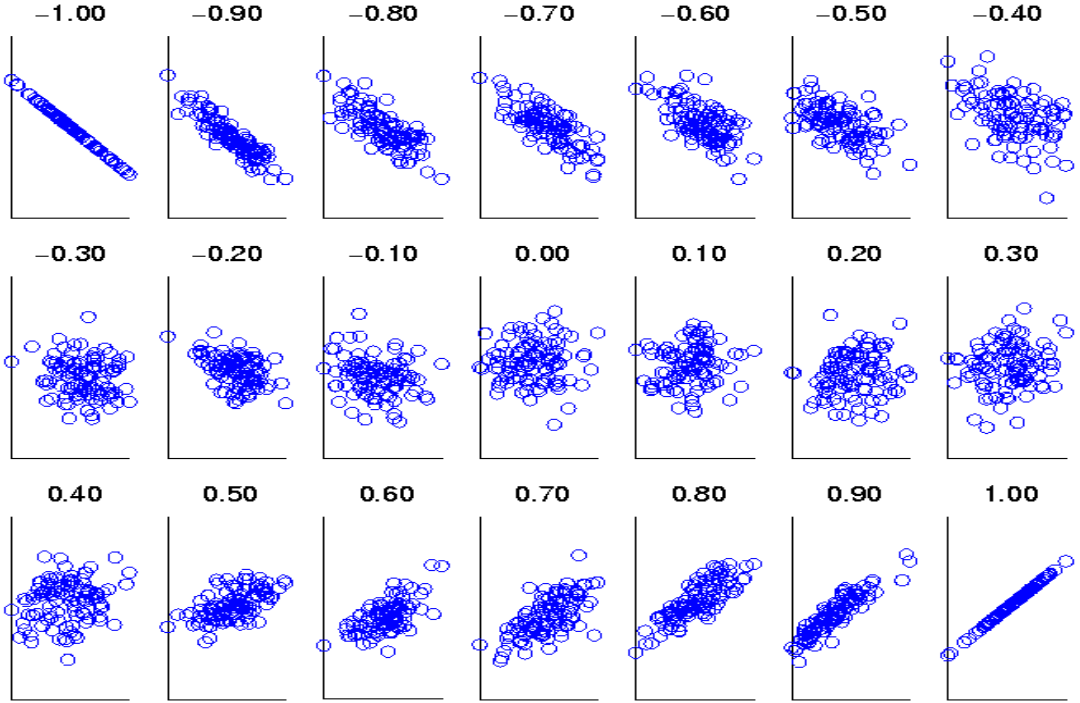
- Η συσχέτιση μετρά τη γραμμική σχέση μεταξύ αντικειμένων
- Για τον υπολογισμό της, κανονικοποιούμε τα δεδομένα και παίρνουμε το εσωτερικό τους γινόμενο

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

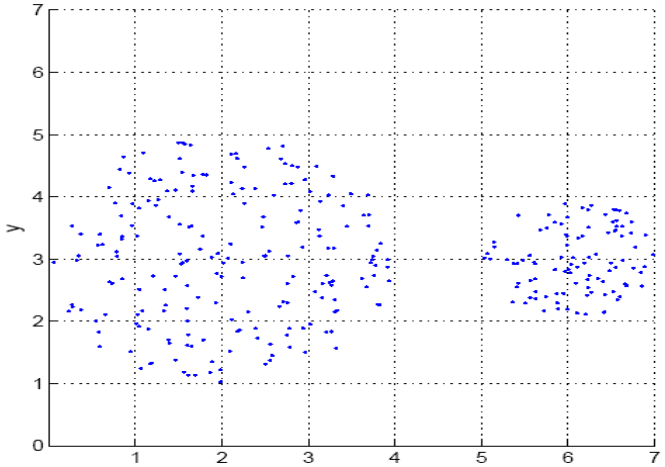
Αξιολογώντας τη συσχέτιση με εικόνες



Διαγράμματα διασποράς που δείχνουν τη συσχέτιση από το -1 έως το 1.

Πυκνότητα

- Ευκλείδεια πυκνότητα, βασισμένη σε κελιά
  - Διαιρούμε την περιοχή σε κελιά και κατασκευάζουμε ένα πίνακα με τον αριθμό των αντικειμένων κάθε περιοχής



0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Πυκνότητα

- Ευκλείδεια πυκνότητα, βασισμένη στο κέντρο σημείου
  - Διαγράφουμε ένα κύκλο με συγκεκριμένη ακτίνα γύρω από ένα σημείο και επιστρέφουμε τον αριθμό των δεδομένων που περικλείει