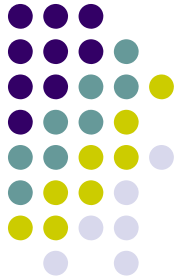
ΠΛ0823 Συστήματα Ανακάλυψης Γνώσης από Βάσεις Δεδομένων

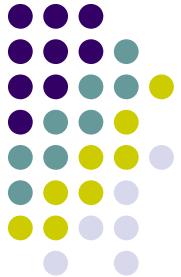
Διδάσκοντες
Γεωργία Κολωνιάρη, Γεώργιος Ευαγγελίδης

Πηγές υλικού μαθήματος

- Οι διαφάνειες της διάλεξης αυτής είναι από το site του textbook “Introduction to Data Mining” των Tan, Steinbach, Karpatne, Kumar.
- Οι μεταφρασμένες/προσαρμοσμένες στα ελληνικά διαφάνειες που ακολουθούν είναι της συναδέλφου Ε. Πιτουρά από το Πανεπιστήμιο Ιωαννίνων.



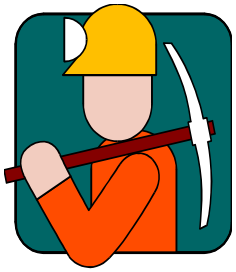
ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ



Εισαγωγή

Τεράστιος όγκος διαθέσιμων δεδομένων – χρειαζόμαστε μεθόδους για να τα αναλύσουμε

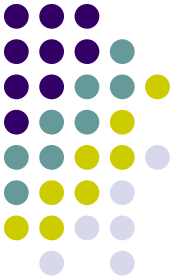
Τι είναι η Εξόρυξη Δεδομένων



(με δυο λόγια)

Αποδοτικές τεχνικές για να αναλύσουμε πολύ μεγάλες συλλογές από δεδομένα και να εξάγουμε χρήσιμες πληροφορίες από αυτά





Παραδείγματα Δεδομένων

Κυβερνητικά: IRS (εφορία), δημογραφικά δεδομένα, «ΔΙΑΦΑΝΕΙΑ», …

Αρχεία κειμένου (document data)

- Web ως συλλογή κειμένων: δισεκατομμύρια σελίδες
- Wikipedia: 4 εκατομμύρια λήμματα (που συνεχώς αυξάνονται)
- Online συλλογές επιστημονικών άρθρων

Μεγάλες εταιρίες

- WALMART: 20M συναλλαγές την ημέρα
- MOBIL: 100 TB γεωλογικά σύνολα δεδομένων
- AT&T 300 M κλήσεις την ημέρα
- Εταιρίες πιστωτικών κρατών

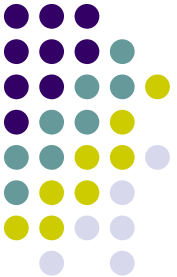
Επιστημονικά

- NASA, EOS project: 50 GB την ώρα

Δεδομένα για το περιβάλλον (για παράδειγμα)

http://www.ncdc.gov/oa/climate/ghcn-monthly/index.php

- "a database of temperature, precipitation and pressure records managed by the National Climatic Data Center, Arizona State University and the Carbon Dioxide Information Analysis Center"
- "6000 temperature stations, 7500 precipitation stations, 2000 pressure stations"



Παραδείγματα Δεδομένων

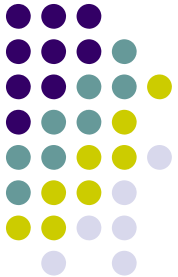
Παράδειγμα: γονιδιακές ακολουθίες genomic sequences

http://www.1000genomes.org/page.php

- Πλήρης ακολουθία για 1000 άτομα
- 310^9 nucleotides για κάθε άτομο → 310^12 nucleotides
- Στην πραγματικότητα ακόμα περισσότερα δεδομένα: ιατρικό ιστορικό ατόμων, γονίδια (gene expression data) κλο

Παράδειγμα: Διαδικτυακά δεδομένα

- Web: 50 δισεκατομμύρια σελίδες διασυνδεδεμένες
- Facebook: 400 εκατομμύρια χρήστες
- MySpace: 300 εκατομμύρια χρήστες
- Instant messenger: ~ 1 δισεκατομμύρια χρήστες
- Blogs: 250 εκατομμύρια blogs



Εξόρυξη Δεδομένων (Ορισμός)

Πολύ μεγάλα σύνολα δεδομένων (data sets)

(1) η διαδικασία ανακάλυψης (discovery) προτύπων (patterns) που πριν δεν ήταν γνωστά, ισχύουν, είναι πιθανόν χρήσιμα και είναι κατανοητά

(2) η ανάλυση τους για να βρούμε μη αναμενόμενες σχέσεις ανάμεσα στα δεδομένα καθώς και να τα συνοψίσουμε με νέους τρόπους που είναι κατανοητοί και χρήσιμοι στους χρήστες

Γιατί Εξόρυξη Δεδομένων (από εμπορική πλευρά)



- Πολλά δεδομένα συγκεντρώνονται και εισάγονται σε αποθήκες δεδομένων ή είναι διαθέσιμα στο διαδίκτυο
  - Αγορές σε πολύ-καταστήματα/αλυσίδες
  - Συναλλαγές με τράπεζες/πιστωτικές κάρτες
  - Web, web logs
  - Network traffic
  - Κοινωνικά δίκτυα (emails, συστήματα δικτύωσης)



Σχεδιασμός καλύτερων συστημάτων

- Αποφυγή spam, αποδοτικότητα

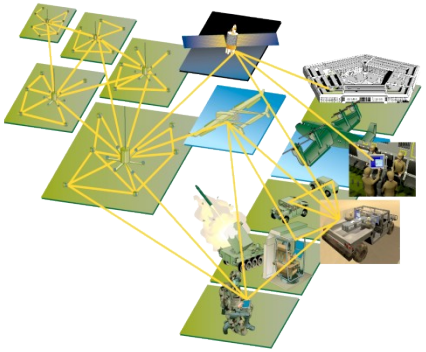
Μεγάλος ανταγωνισμός

- Παροχή καλύτερων, προσωπικών υπηρεσιών σε κάποιο πεδίο (fraud detection, target marketing)

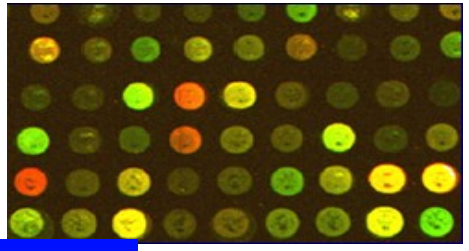


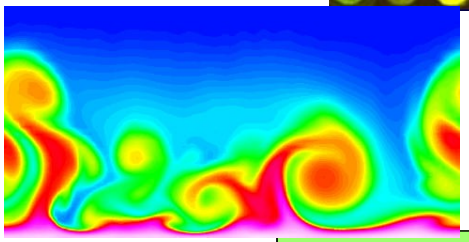
Γιατί Εξόρυξη Δεδομένων (από επιστημονική πλευρά)

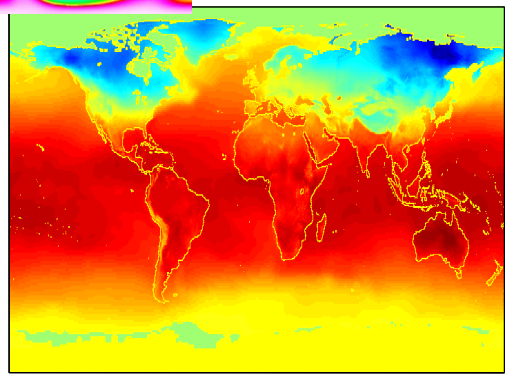
- Τα δεδομένα συλλέγονται και αποθηκεύονται σε τρομερές ταχύτητες (GB/hour)
  - Απομακρυσμένοι αισθητήρες (remote sensors) σε δορυφόρους
  - Τηλεσκόπια στον ουρανό
  - Microarrays που παράγουν γονιδιακά δεδομένα
  - Επιστημονικές προσομοιώσεις που παράγουν terabytes δεδομένων
- Η εξόρυξη δεδομένων μπορεί να βοηθήσει τους επιστήμονες
  - Στην κατηγοριοποίηση και την τμηματοποίηση των δεδομένων
  - Στη διατύπωση υποθέσεων



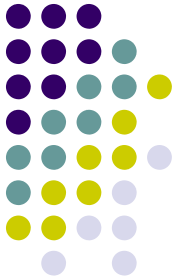








Ποια γονίδια σχετίζονται με κάποια αρρώστια, ποια είναι η συσχέτιση μεταξύ ακραίων καιρικών συνθηκών και της υπερθέρμανσης του πλανήτη



Είδη/Τεχνικές Εξόρυξης Δεδομένων (συνοπτικά)

Κατηγοριοποίηση (Classification)

κατηγοριοποιούμε τα δεδομένα τοποθετώντας τα σε μια (ή περισσότερες) από έναν αριθμό από δοσμένες κατηγορίες

Ομαδοποίηση (συσταδοποίηση) – clustering

χωρίζουμε τα δεδομένα σε ομάδες από «όμοια» σύνολα

Κανόνες συσχέτισης (Association rule mining)

βρίσκουμε συσχετίσεις ανάμεσα στα δεδομένα, πχ ποια δεδομένα εμφανίζονται συχνά μαζί σε συναλλαγές

Είδη με βάση τα δεδομένα στα οποία γίνεται η εξόρυξη

Οι «ρίζες» της Εξόρυξης Δεδομένων

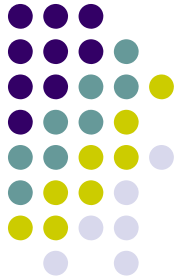


Πρέπει να αντιμετωπίσει:

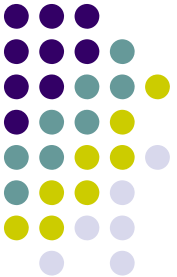
- Το τεράστιο μέγεθος των δεδομένων
- Το μεγάλο αριθμό διαστάσεων
- Την μη ομοιογενή και την κατανεμημένη φύση των δεδομένων

Η προσέγγιση στο μάθημα θα είναι σε αλγορίθμους/δομές και μεγάλα σύνολα δεδομένων – από την πλευρά των συστημάτων λογισμικού

Φιλοσοφίες



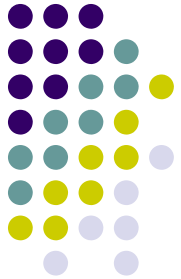
- Βάσεις δεδομένων: έμφαση σε πολύ μεγάλης κλίμακας (εκτός κύριας μνήμης) δεδομένων
- ΤΝ (μηχανική μάθηση): έμφαση σε περίπλοκες μεθόδους, λίγα δεδομένα
- Στατιστική: έμφαση σε μοντέλα



Μοντέλα vs. Αναλυτική Επεξεργασία

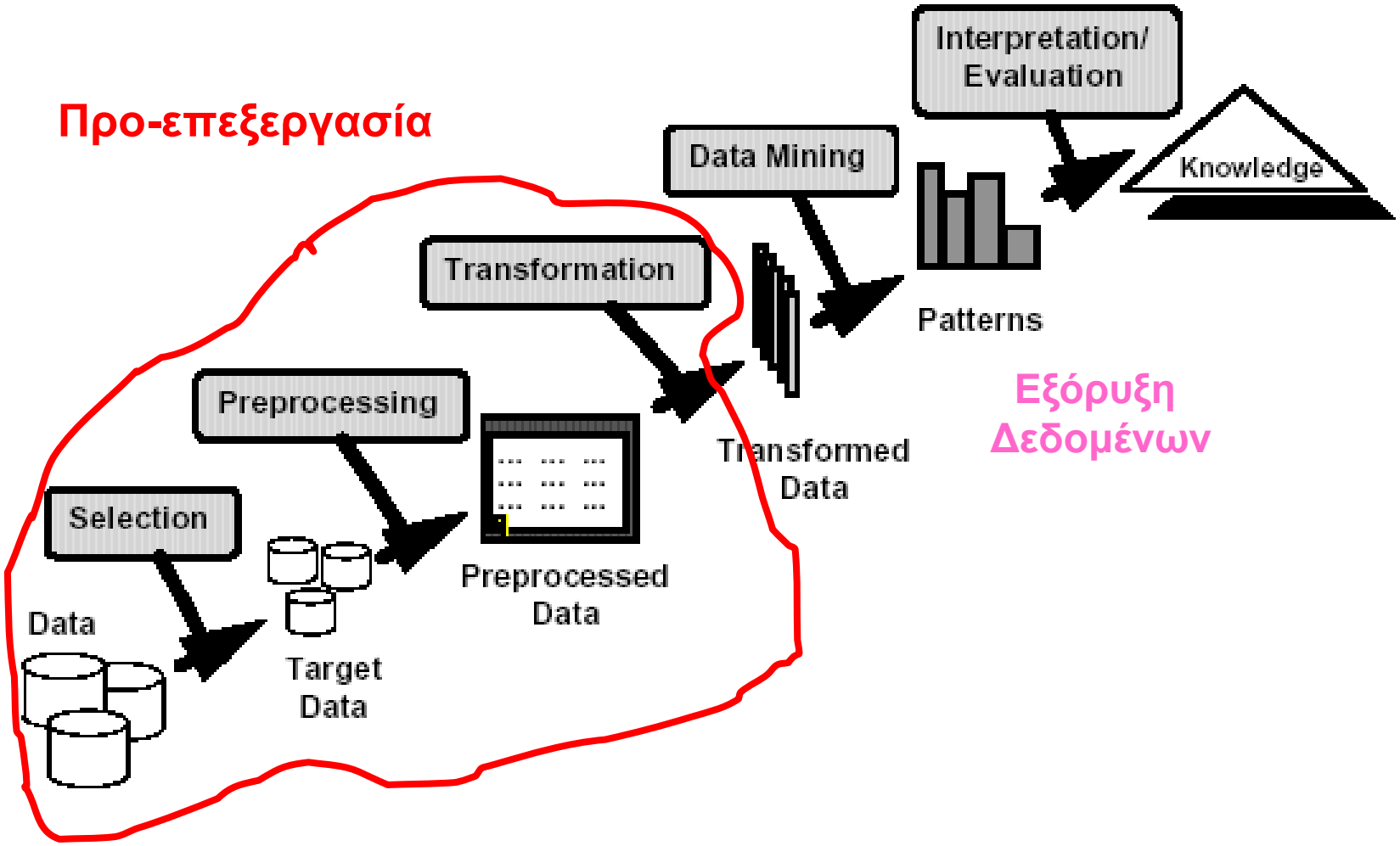
- Από τη πλευρά των βάσεων δεδομένων, η εξόρυξη δεδομένων είναι μια ακραία μορφή αναλυτικής επεξεργασίας – ερωτήσεις (queries) που εξετάζουν πολύ μεγάλο όγκο δεδομένων.
  - Το αποτέλεσμα είναι η απάντηση της ερώτησης.
- Από τη πλευρά της στατιστικής, η εξόρυξη δεδομένων είναι η επαγωγή (inference) ενός μοντέλου.
  - Το αποτέλεσμα είναι οι παράμετροι του μοντέλου.

(Πολύ Απλουστευμένο) Παράδειγμα

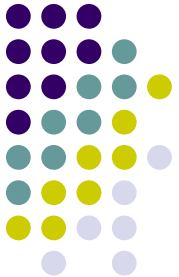


- Δοθέντος ενός δισεκατομμυρίου αριθμών, από τη πλευρά των βάσεων δεδομένων θα υπολογίζαμε πχ τη μέση τιμή και την τυπική απόκλιση.
- Από τη πλευρά της στατιστικής θα προσπαθούσε να ταιριάξει (fit) τα δισεκατομμύρια σημεία στην καλύτερη Gaussian κατανομή και να καταγράψει τη μέση τιμή και την τυπική απόκλιση *αυτής της κατανομής.*

Ανακάλυψη Γνώσης (Knowledge Discovery)



Ανακάλυψη Γνώσης (Knowledge Discovery)



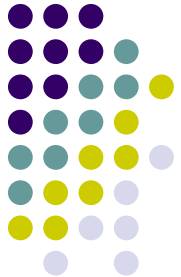
ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ

- Data Cleaning – Καθαρισμός Δεδομένων
- Data Integration – Ενοποίηση Δεδομένων
- Data Transformation – Μετασχηματισμοί Δεδομένων

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

ΑΝΑΠΑΡΑΣΤΑΣΗ

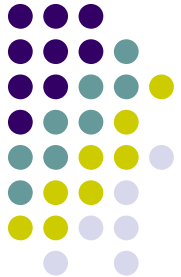
Προ-επεξεργασία δεδομένων - Καθαρισμός



Τα δεδομένα στο πραγματικό κόσμο είναι «βρώμικα»

- Ελλειπή - incomplete: μπορεί να *λείπουν* κάποιες τιμές γνωρισμάτων (να μην καταγράφηκαν, να *καταγράφηκαν λανθασμένα* λόγω μη συνεννόησης ή λανθασμένης λειτουργίας), να λείπουν κάποια *ενδιαφέροντα γνωρίσματα* (που να μην θεωρήθηκαν σημαντικά ή να μην ήταν διαθέσιμα), ή να περιέχουν μόνο συναθροιστικά (aggregate) δεδομένα
  - Συμπλήρωση των γνωρισμάτων και τιμών που λείπουν
- Με θόρυβο - noisy: περιέχουν λάθη ή outliers (περιθωριακές τιμές - τιμές που διαφέρουν πολύ από την πλειοψηφία)
  - Εύρεση των περιθωριακών τιμών και απομάκρυνση θορύβου
- Ασυνεπή - inconsistent: περιέχουν ασυνέπειες, διπλότιμα
  - Διόρθωση ασυνεπών τιμών

Προ-επεξεργασία δεδομένων



Επιλογή Δεδομένων και Γνωρισμάτων και εφαρμογή κατάλληλων Μετασχηματισμών

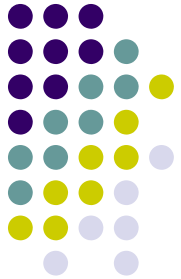
- Συνάθροιση – Aggregation: συνδυασμούς δεδομένων από πολλές πηγές
- Sampling – δειγματοληψία: χρήση αντιπροσωπευτικού δείγματος των δεδομένων για βελτίωση της απόδοσης
- Dimensionality reduction – μείωση διαστάσεων - Κατάρα της διάστασης (curse of dimensionality)

Πολλές τεχνικές για την ανάλυση δεδομένων γίνονται δυσκολότερες με την αύξηση της διάστασης των δεδομένων (αυξάνει εκθετικά η πολυπλοκότητα ή τα δεδομένα γίνονται πολύ αραιά)

Τεχνικές της γραμμικής άλγεβρας (SVD, PCA)

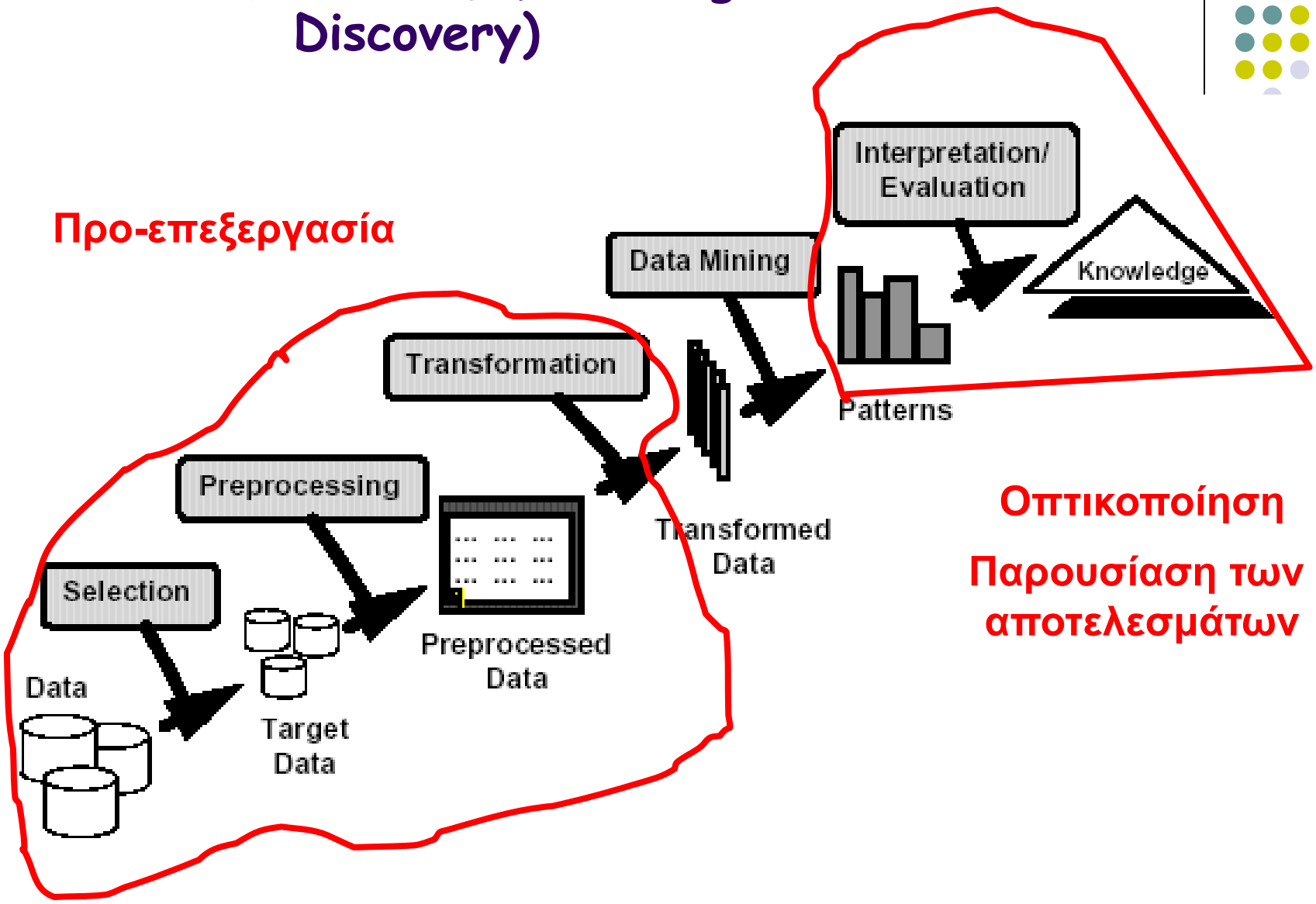
Απεικόνιση σε άλλο χώρο με μικρότερο αριθμό διαστάσεων

Προ-επεξεργασία δεδομένων

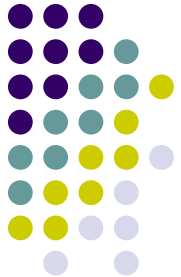


- Discretization (μετασχηματιμός σε μια διακριτή τιμή) ή binarization (μετασχηματισμός σε δυαδική τιμή)
- Variable transformation – μετασχηματισμοί των τιμών των μεταβλητών
  - Πχ Κανονικοποίηση

Ανακάλυψη Γνώσης (Knowledge Discovery)



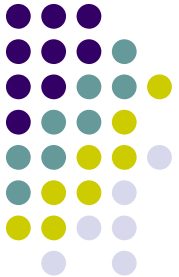

Δυναμικό για την υποστήριξη επιχειρησιακών αποφάσεων
Τελικός Χρήστης
Λήψη Αποφάσεων
Παρουσίαση Δεδομένων
Τεχνικές Οπτικοποίησης
Business Analyst
Εξόρυξη Δεδομένων
Ανακάλυψη Πληροφορίας
Data Analyst
Data Exploration
Στατιστικές περιλήψεις, Ερωτήσεις (OLAP)
Προ-επεξεργασία&Ενοποίηση Δεδομένων, Αποθήκες
DBA Διαχειριστής ΒΔ
Πηγές Δεδομένων
Χαρτιά, Αρχεία, Web έγγραφα, Επιστημονικά Πειράματα, Συστήματα Βάσεων Δεδομένων



Πως χρησιμοποιείται

1. Κατανόηση του προβλήματος
2. Χρήση τεχνικών εξόρυξης δεδομένων για να πάρουμε πληροφορία από τα δεδομένα
3. Χρήση αυτής της πληροφορίας
4. Μέτρηση των αποτελεσμάτων

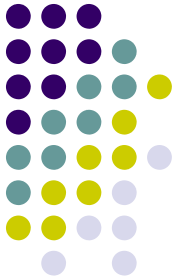
Είδη/Μέθοδοι για Εξόρυξη Δεδομένων



(συνοπτικά)

1. Κατηγοριοποίηση - Classification: εκμάθηση μια συνάρτησης – κατασκευή ενός μοντέλου που απεικονίζει ένα στοιχείο σε μια από ένα σύνολο από προκαθορισμένες κλάσεις
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
5. Συνοψίσεις - Summarization: εύρεση μιας συνοπτικής περιγραφής του συνόλου δεδομένων ή ενός υποσυνόλου του
6. Αλλα

Κατηγορίες Εξόρυξης Δεδομένων



Descriptive Methods - Περιγραφικές Μέθοδοι

Στόχος να βρεθούν κατανοητά πρότυπα που περιγράφουν τα δεδομένα – τις ιδιότητες τους

Predictive Methods – Μέθοδοι πρόβλεψης

Χρήση κάποιων μεταβλητών για να προβλέψουν άγνωστες ή μελλοντικές τιμές κάποιων άλλων μεταβλητών

Είδη/Μέθοδοι για Εξόρυξη Δεδομένων

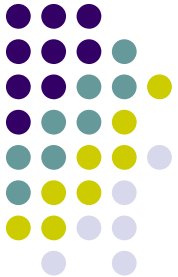


- Κατηγοριοποίηση [Predictive]
- Συσταδοποίηση [Descriptive]
- Εύρεση Κανόνων Συσχέτισης [Descriptive]

- Sequential Pattern Discovery [Descriptive]
- Regression – Συνοψίσεις [Predictive]

  ένα συνοπτικό μοντέλο για τα δεδομένα (πχ μια συνάρτηση)
- Deviation/Anomaly Detection [Predictive]

  outlier analysis (στατιστικοί έλεγχοι για σπάνια σημεία),
  evolution analysis (πχ ανάλυση χρονοσειρών – πχ μετοχές) κλπ



Ορισμός

- Δοθέντος ενός συνόλου από εγγραφές (**σύνολο εκπαίδευσης - training set**)
  - Κάθε εγγραφή έχει ένα σύνολο από γνωρίσματα, ένα από αυτά είναι η κλάση (ή κατηγορία)
- Εύρεση ενός **μοντέλου** για το γνώρισμα της κλάσης ως συνάρτηση της τιμής των άλλων γνωρισμάτων.
- Στόχος: να αναθέτει σε εγγραφές που δεν έχουμε δει μια κλάση με την μεγαλύτερη δυνατή ακρίβεια
  - Για να χαρακτηρίσουμε την ακρίβεια του μοντέλου χρησιμοποιούμε ένα **σύνολο ελέγχου** (**test set**)
  - Συνήθως, το δοθέν σύνολο δεδομένο 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

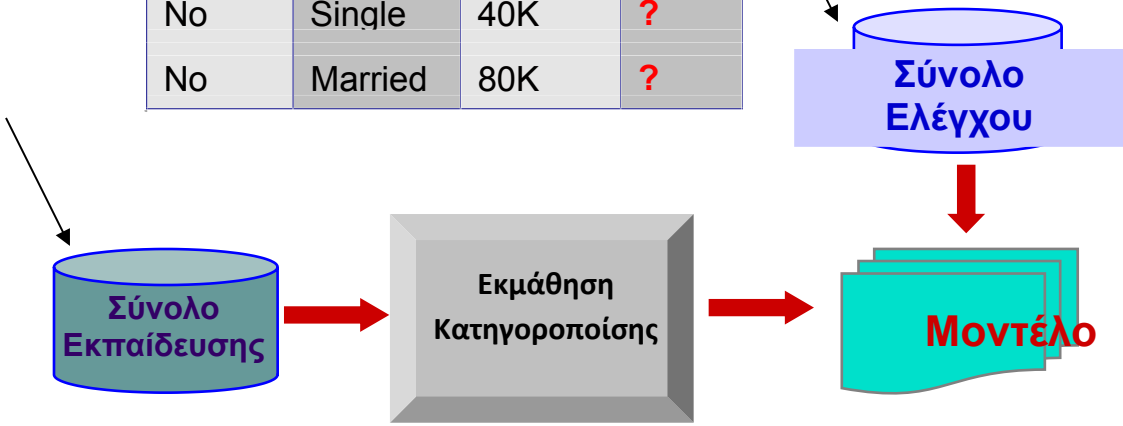
Παράδειγμα

Γνωρίσματα

ΚΛΑΣΗ

Tid	HomeOwner	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

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



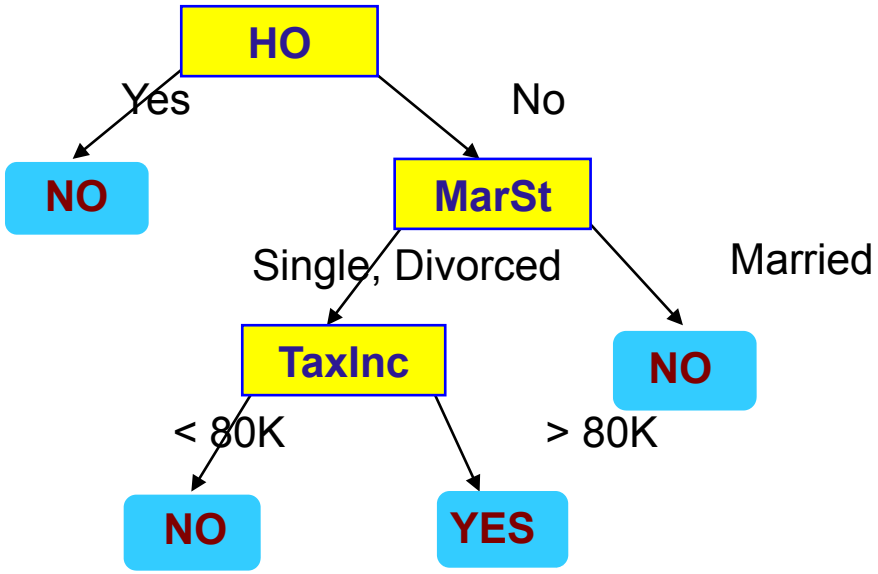
Πως μοιάζει το μοντέλο;

Παράδειγμα Μοντέλου: Δέντρο Απόφασης – Decision tree

ΚΛΑΣΗ

Tid	Home Owner	Marital Status	Taxable Income	Default
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

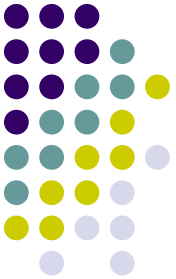




Σύνολο Εκπαίδευσης

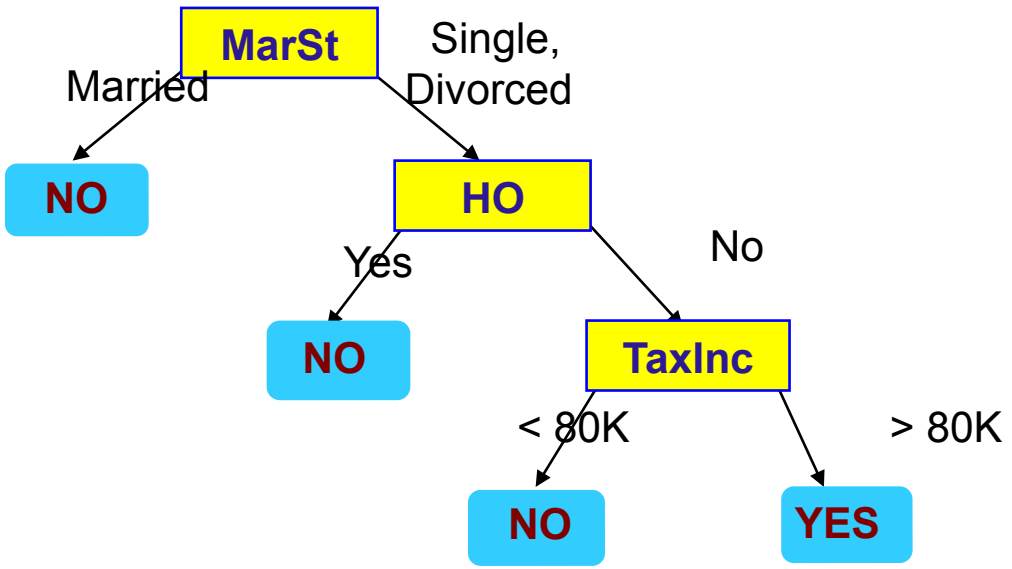
Μοντέλο: Δέντρο Απόφασης

Κατηγοριοποίηση IV

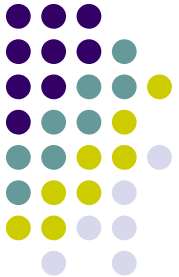


ΚΛΑΣΗ

Tid	Home Owner	Marital Status	Taxable Income	Default
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)



Regression analysis – ανάλυση παλινδρόμισης: στατιστική εκμάθηση μια συνάρτησης που απεικονίζει ένα στοιχείο σε μια πραγματική τιμή, χρήση για αριθμητικές προβλέψεις

Ανάλυση σχετικότητας (relevance analysis): ποια γνωρίσματα επηρεάζουν την ταξινόμηση

Άλλα είδη μοντέλων πλην των Δέντρων Απόφασης: νευρωνικά δίκτυα, κ-εγγύτεροι γείτονες, naive bayes, support vector machines κλπ

Κατηγοριοποίηση: Εφαρμογή 1

Direct Marketing

Στόχος: Μείωση των ταχυδρομικών εξόδων για την αποστολή διαφημιστικών με τη στοχοποίηση *targeting* του συνόλου των πελατών που είναι πιο πιθανόν να αγοράσουν ένα κινητό τηλέφωνο

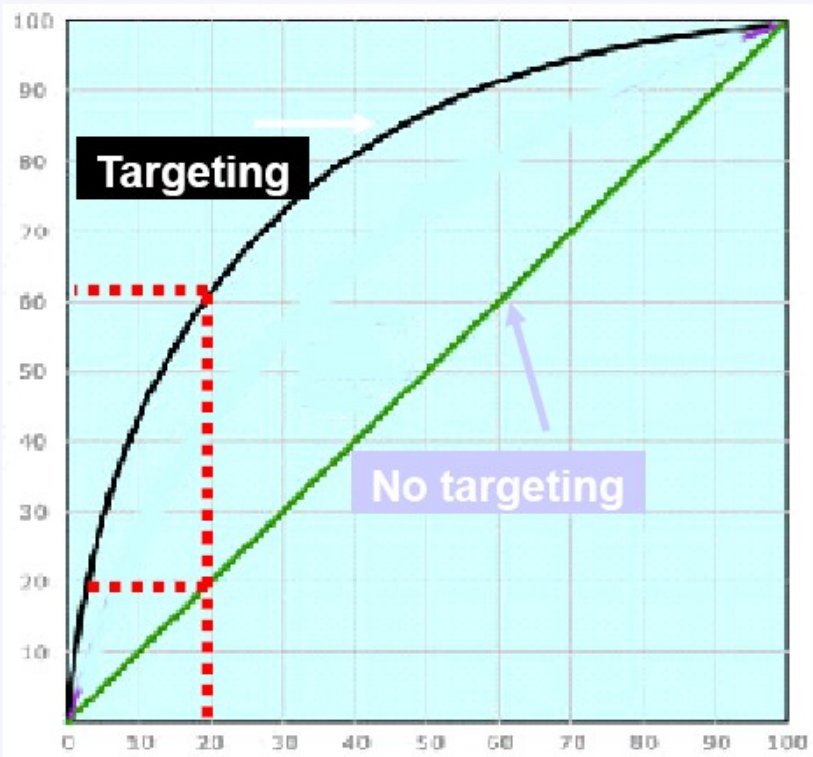
Προσέγγιση:

Χρησιμοποίηση των δεδομένων από ένα *παρόμοιο προϊόν* που βγήκε στην αγορά πρόσφατα

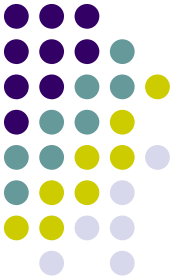
Για αυτό το προϊόν ξέρουμε ποιοι αποφάσισαν να το αγοράσουν και ποιοι όχι -> γνώρισμα της κλάσης {buy, don't buy}.

Συλλογή ποικίλων δημογραφικών δεδομένων κλπ για αυτούς τους πελάτες

Χρήση αυτής της πληροφορίας ως τα γνωρίσματα για την εκμάθηση ενός μοντέλου ταξινόμησης.



Κατηγοριοποίηση: Εφαρμογή 2



Fraud Detection – Αναγνώριση Απάτης σε Πιστωτικές Κάρτες

Στόχος: Να βρούμε ποιες συναλλαγές μιας πιστωτικής κάρτας δεν είναι από τον ιδιοκτήτη της

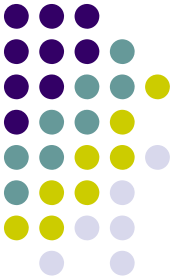
Προσέγγιση:

Χρησιμοποίηση των *δεδομένων από προηγούμενες συναλλαγές* με αυτήν την κάρτα και πληροφορίες για τον κάτοχο της (τι αγοράζει, πότε, από πού, πόσο συχνά πληρώνει)

Χαρακτηρισμός κάθε προηγούμενης συναλλαγής ως απάτη ή όχι -> γνώρισμα της κλάσης {fraud, fair}.

Χρήση αυτής της πληροφορίας ως τα γνωρίσματα για την εκμάθηση ενός μοντέλου ταξινόμησης.

Χρήση του μοντέλου για τον χαρακτηρισμό μελλοντικών συναλλαγών



Customer Attrition

Στόχος: Να εκτιμήσουμε να ένας πελάτης θα προτιμήσει μια ανταγωνιστική εταιρεία

Προσέγγιση:

Χρησιμοποίηση των δεδομένων από παλιές και νέες συναλλαγές πελατών (πόσο συχνά τηλεφωνούν, που πότε, την οικονομική του κατάσταση, την οικογενειακή του κατάσταση κλπ)

Χαρακτηρισμός κάθε πελάτη ως πιστού ή όχι -> γνώρισμα της κλάσης {loyal, disloyal}.

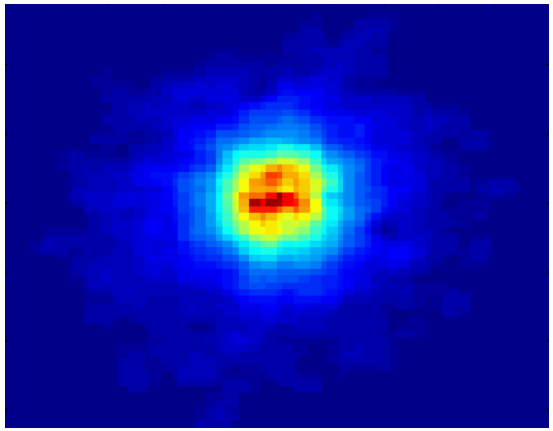
Χρήση αυτής της πληροφορίας ως τα γνωρίσματα για την εκμάθηση ενός μοντέλου ταξινόμησης.



Ταξινόμηση Γαλαξιών

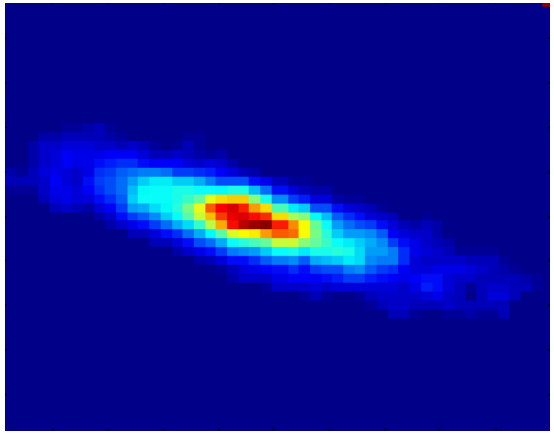
Courtesy: http://aps.umn.edu

Αρχικό



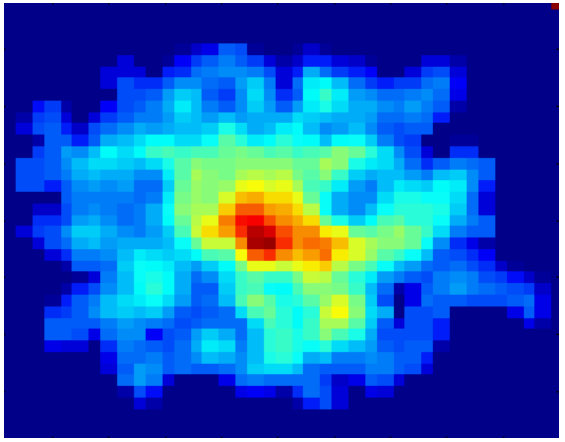
Κλάση:
- **Στάδιο δημιουργίας**

Ενδιάμεσο

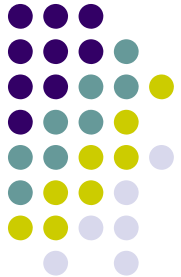


Γνωρίσματα:
- **Χαρακτηριστικά της εικόνας,**
- **Χαρακτηριστικά του κυμάτων φωτός που ελήφθησαν, κλπ.**

Προχωρημένο



Μέγεθος Δεδομένων:
- **72 εκατ. άστρα, 20 εκατ. γαλαξίες**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**



Ορισμός

- Δοθέντων
  - Ενός συνόλου από σημεία που το καθένα έχει κάποια γνωρίσματα
  - Μιας μέτρηση **ομοιότητας** μεταξύ τους

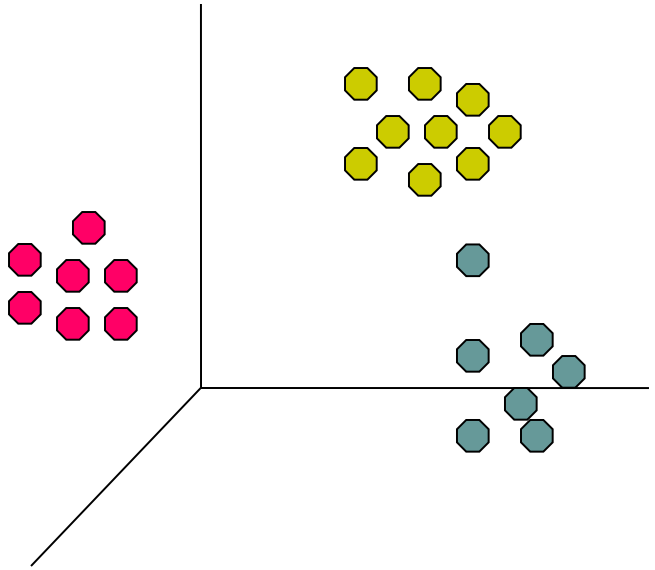
- Εύρεση **συστάδων** (**clusters**) τέτοιων ώστε:
  - Τα σημεία σε μία συστάδα είναι πιο όμοια μεταξύ τους
  - Τα σημεία σε διαφορετικές συστάδες είναι λιγότερα όμοια μεταξύ τους

Σε αντίθεση με την ταξινόμηση, οι συστάδες δεν είναι γνωστές από πριν

Παράδειγμα

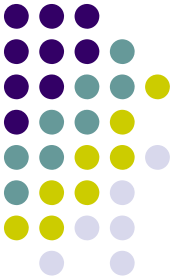
Οι αποστάσεις μέσα στη συστάδα ελαχιστοποιούνται

Οι αποστάσεις ανάμεσα στις συστάδες μεγιστοποιούνται



- 3-διάστατα σημεία, ευκλείδεια απόσταση

Συσταδοποίηση: Εφαρμογή 1



Market Segmentation

Στόχος: Χωρισμός των καταναλωτών σε ομάδες έτσι ώστε τα μέλη κάθε ομάδας να είναι ο στόχος για μια συγκεκριμένη πολιτική marketing

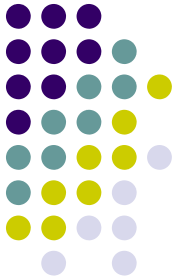
Προσέγγιση:

Συγκέντρωση διαφορετικών γνωρισμάτων για τους καταναλωτές

Ορισμός *«ομοιότητας»* ανάμεσα στους πελάτες

Δημιουργία ομάδων με όμοιους πελάτες

Μέτρηση της ποιότητας της ομαδοποίησης (πχ παρατηρώντας τις αγοραστικές συνήθειες στην ίδια ομάδα και ανάμεσα σε διαφορετικές ομάδες)

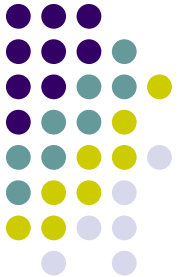


Συσταδοποίηση Εγγράφων

Στόχος: Εύρεση ομάδων από έγγραφα που είναι όμοια μεταξύ τους με βάση τους σημαντικούς όρους που εμφανίζονται σε αυτά

Προσέγγιση: Εύρεση για κάθε έγγραφο των όρων που εμφανίζονται συχνά σε αυτό.
Μέτρηση ομοιότητας με βάση τη συχνότητα των διαφορετικών όρων, Χρήση της για τη δημιουργία συστάδων

Οφελος: Μέθοδοι Ανάκτησης Πληροφορία (Information Retrieval) μπορεί να χρησιμοποιήσουν τις συστάδες για να συσχετίσουν έναν καινούργιο έγγραφο ή έναν όρο αναζήτησης με τα έγγραφα κάθε συστάδας



Στο μάθημα

Θα δούμε ενδιαφέροντες τρόπους να ορίσουμε ομοιότητα/απόσταση και τους θεμελιώδεις αλγορίθμους συσταδοποίησης (k-Means, Hierarchical, DBScan)



Ορισμός (συχνών στοιχειοσυνόλων)

- Δοθέντος
  - Ενός συνόλου από εγγραφές που η κάθε μία έχει έναν αριθμό από στοιχεία από κάποιο δοσμένο σύνολο
- Εύρεση **κανόνων εξάρτησης** που προβλέπουν την παρουσία ενός στοιχείου με βάση την παρουσία άλλων στοιχείων

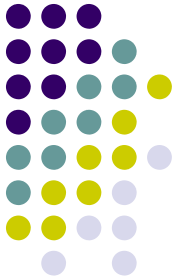
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Κανόνες που βρέθηκαν:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Κανόνες Συσχέτισης: Εφαρμογή 1



Για marketing και προώθηση πωλήσεων:

Εστω ότι ο κανόνας που ανακαλύφθηκε είναι ο:

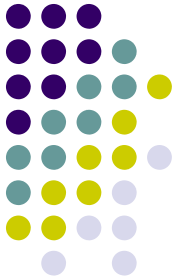
{Bagels, … } --> {Potato Chips}

Potato Chips στα δεξιά του κανόνα => Τι πρέπει να γίνει για να αυξηθούν οι πωλήσεις.

Bagels στα αριστερά => Μπορεί να χρησιμοποιηθεί για να εκτιμηθεί ποια προϊόντα θα επηρεαστούν αν πχ ένα μαγαζί σταματήσει να τα πουλάει.

Bagels στα αριστερά and Potato chips στα δεξιά => Ποια προϊόντα πρέπει να πουληθούν μαζί με Bagels για την προώθηση των Potato chips!

Κανόνες Συσχέτισης: Εφαρμογή 2



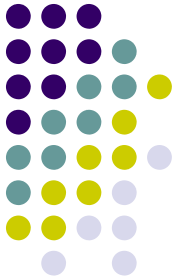
Πως θα φτιάξουμε τα ράφια στα super-markets!

«γνωστός» κανόνας --

Αν ο καταναλωτής αγοράσει πάνες, πολύ πιθανών να αγοράσει και μπύρα!

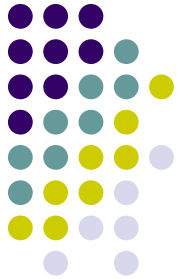
Στις ΗΠΑ, Πέμπτη και Σάββατο, άντρες που αγοράζουν πάνες αγοράζουν και μπύρα

Εκτίμηση σημασίας



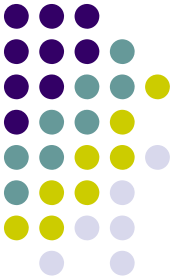
- Ενας μεγάλος κίνδυνος είναι η «εύρεση» προτύπων που δεν έχουν νόημα
- Στη στατιστική καλείται Bonferroni's principle
  - Αν ψάξεις κάτι πολύ γενικό σε πάρα πολλά δεδομένα θα το βρεις!

Παραδείγματα της αρχής του Bonferroni's



1. αν κοιτάξουμε πολύ ασαφείς συνδέσεις θα καταλήξουμε σε λάθος αποτελέσματα – ένα παράδειγμα αναζήτησης τρομοκρατών!
2. The Rhine Paradox (το παράδοξο του Rhine)

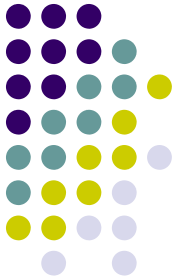
Rhine Paradox I



- Ο Joseph Rhine ήταν ένας παρα-ψυχολόγος στη δεκαετία του 1950 που ανάπτυξε τη θεωρία ότι κάποια άτομα έχουν υπερφυσικές ικανότητες - Extra-Sensory Perception (ESP).

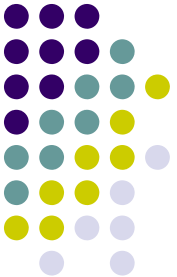
- Σχεδίασε ένα είδος πειράματος στο οποίο ζήτησε από τους ανθρώπους που συμμετείχαν σε αυτό να μαντέψουν το χρώμα μιας κρυμμένης κάρτας – κόκκινη ή μπλε.

- Ανακάλυψε ότι περίπου 1 στους 1000 είχαν ESP – βρήκαν σωστά τα χρώματα και από τις 10 κάρτες!



Rhine Paradox II

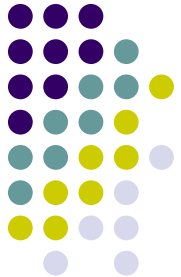
- Είπε σε αυτούς τους ανθρώπους ότι είχαν ESP και τους κάλεσε για ένα ακόμα ίδιο πείραμα
- … ανακάλυψε ότι όλοι έχασαν το ESP.
- Ποιο ήταν το συμπέρασμά του;



Rhine Paradox III

- Συμπέρανε ότι δεν πρέπει να λες στους ανθρώπους ότι έχουν ESP, γιατί αυτό έχει ως αποτέλεσμα να χάνουν αυτήν την ικανότητα

Εκτίμηση ενδιαφέροντος



Χαρακτηρισμό του «ενδιαφέροντος» ενός προτύπου:

(1) Εύκολα κατανοητό

(2) Να ισχύει σε δεδομένα ελέγχου ή σε νέα δεδομένα με κάποιο βαθμό βεβαιότητας

(3) Πιθανόν χρήσιμο

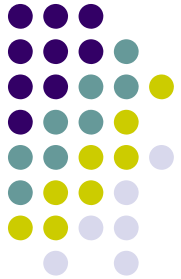
(4) Νέα πληροφορία

Υπάρχουν υποκειμενικά (αναμενόμενα και μη αναμενόμενα) και αντικειμενικά κριτήρια – Κάποιες τιμές κατωφλίου

Πληρότητα (όλα τα ενδιαφέροντα πρότυπα)

Βελτιστοποίηση (μόνο τα ενδιαφέροντα πρότυπα)

Η γενική εικόνα

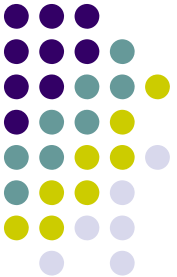


- Εκμάθηση του πεδίου εφαρμογής
  - Σχετική προηγούμενη γνώση και τους στόχους της εφαρμογής
- Δημιουργία του συνόλου δεδομένων: data selection
- Καθαρισμός και προ-επεξεργασία των δεδομένων: (έως και 60% της συνολικής προσπάθειας)
- Ελάττωση δεδομένων και μετασχηματισμοί
  - Χρήσιμα χαρακτηριστικά, ελάττωση διαστάσεων κλπ

- Επιλογή **λειτουργίας** εξόρυξης δεδομένων
  - πχ, συσταδοποίηση, ταξινόμηση, κλπ
- Επιλογή του **αλγορίθμου** εξόρυξης δεδομένων
- Εξόρυξη Δεδομένων: αναζήτηση προτύπων ενδιαφέροντος

- Εκτίμηση προτύπων και αναπαράσταση γνώσης
  - οπτικοποίηση, μετασχηματισμοί, απομάκρυνση περιττών προτύπων, κλπ

- Χρήση της γνώσης



Οι 10 καλύτεροι αλγόριθμοι ΕΔ (ICDM 2006)

#1: C4.5 (61 votes) – ταξινόμηση (δέντρο απόφασης)

#2: K-Means (60 votes) - συσταδοποίηση

#3: SVM (58 votes) – ταξινόμηση (support vector machine)

#4: Apriori (52 votes) – κανόνες συσχέτισης

#5: EM (48 votes) – στατιστική, συσταδοποίηση (expectation maximization)

#6: PageRank (46 votes) – ιστοσελίδες

#7: AdaBoost (45 votes) – μετα-ταξινομητής

#7: kNN (45 votes) – συσταδοποίηση (κοντινότερος γείτονας)

#7: Naive Bayes (45 votes) – στατιστική, ταξινόμηση

#10: CART (34 votes) – ταξινόμηση (δέντρο απόφασης)