



Εξόρυξη Γνώσης από Δεδομένα

1: Γενικά για το μάθημα

Βασικές έννοιες

Περιεχόμενα

- Πώς γίνεται το μάθημα, τι διαβάζω, πώς εξετάζομαι ;
- Τι είναι η εξόρυξη γνώσης;
- Πώς υλοποιείται;
- Ποια η αρχιτεκτονική ενός συστήματος εξόρυξης γνώσης;
- Τι πρέπει να προσέχω;
- Από που ενημερώνομαι;

Διαδικαστικά

- Ώρες μαθήματος: Παρασκευή 10:00-13:00
- Διδάσκων: Ηρακλής Βαρλάμης
- Γραφείο: 5.1
- Ώρες γραφείου: Παρασκευή 9:00-10:00
- E-mail: varlamis@hua.gr
- Web: <http://eclass.hua.gr/courses/DIT129/>
 - Σημειώσεις, ανακοινώσεις, ασκήσεις, κλπ.
- Αίθουσα: 2.3
<https://meet.google.com/rjk-ythi-wrs>

Διδακτικές μέθοδοι

- Θεωρία:
 - Κατηγορίες προβλημάτων εξόρυξης γνώσης από δεδομένα
 - Αλγόριθμοι, τεχνικές, μέθοδοι
- Εργαστήριο:
 - Εφαρμογές των πιο πάνω σε σύνολα δεδομένων
 - Εξοικείωση με εργαλεία εξόρυξης γνώσης και μηχανικής μάθησης
- Ασκήσεις:
 - Σε θέματα του εργαστηρίου αλλά και θεωρίας

Βιβλιογραφία

- Εισαγωγή στην εξόρυξη δεδομένων, 2η Έκδοση (2018), P.N. Tan, M. Steinbach, V. Kumar, (επιμ. Β. Βερύκιος)
- Εξόρυξη και ανάλυση δεδομένων: Βασικές έννοιες και αλγόριθμοι (2018), M. J. Zaki, W. Meira Jr. (επιμ. Β. Μεγαλοοικονόμου, Χ. Μακρής)
- Εξόρυξη από Μεγάλα Σύνολα Δεδομένων - 3η Έκδοση (2020), A. Rajaraman, J. D. Ullman, J. Leskovec (επιμ. Α. Γούναρης, Ι. Μανωλόπουλος κλπ)

Επιπλέον πηγές

- J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011.
 - Jiawei Han homepage: <http://web.engr.illinois.edu/~hanj/>
- Tan, P. N., Steinbach, M., & Kumar, V. (2016). Introduction to data mining. Pearson Education India.
 - Στα έγγραφα του eclass

Βαθμολογία

- Βαθμολογία (ισχύει για Φεβρουάριο & Σεπτέμβριο):
 - Τελική εξέταση: 50% [βαθμός ≥ 5]
 - Εργασία (2 milestones): 50% [βαθμός ≥ 5 , **υποχρεωτική**]
 - Bonus 10% στις 2 καλύτερες εργασίες
- Τελική εξέταση: με ανοιχτό βιβλίο και σημειώσεις μαθήματος
- Κατοχύρωση:
 - Εργασίες δίνετε **μόνο** μέσα στο εξάμηνο και ο βαθμός τους κατοχυρώνεται για το Σεπτέμβριο
 - Τίποτε δεν κατοχυρώνεται για επόμενη χρονιά!



Ας ξεκινήσουμε

Η εξέλιξη στη διαχείριση δεδομένων

- 1960: Συλλογές δεδομένων, δημιουργία βάσεων δεδομένων, συστήματα διαχείρισης πληροφορίας
- 1970: Σχεσιακό μοντέλο δεδομένων, RDBMS
- 1980: Εξελιγμένα μοντέλα δεδομένων (extended-relational, OO, active, deductive, κλπ.) και ΒΔ προσαρμοσμένες στις ανάγκες των εφαρμογών (spatial, temporal, scientific, engineering, κλπ.)
- 1990-2000: Εξόρυξη δεδομένων και αποθήκες δεδομένων (data mining and data warehousing), ΒΔ πολυμέσων, ΒΔ στον παγκόσμιο ιστό
- 2010: Web of Data, Linked Data, Ontologies

Τι είναι η Εξόρυξη Γνώσης

- Αποδοτικές τεχνικές για να **αναλύσουμε** πολύ μεγάλες συλλογές από **δεδομένα** και να **εξάγουμε** χρήσιμες **πληροφορίες** από αυτά
- Η διαδικασία ανακάλυψης **ενδιαφέρουσας** (μη τετριμμένης, κατανοητής, επικυρωμένης, προηγούμενα άγνωστης και πιθανά χρήσιμης) πληροφορίας ή **προτύπων**
- Η ανάλυση δεδομένων με στόχο να βρούμε **μη αναμενόμενες σχέσεις** ανάμεσά τους καθώς και να τα συνοψίσουμε με νέους τρόπους που είναι κατανοητοί και χρήσιμοι στους χρήστες
- Data mining □ Εξόρυξη ή εξαγωγή γνώσης, εξόρυξη δεδομένων, ανάλυση δεδομένων/προτύπων

Παράδειγμα

- Μια τράπεζα διατηρεί δεδομένα σχετικά με τα στεγαστικά δάνεια που έχει δώσει και το αν αποπληρώθηκαν ή όχι

| S e x | A g e | Time Addr | ResStat | occup | Time Emp | Time Bank | House Exp | PAID BACK |
|-------------|-------------|--------------|---------|-----------|-------------|--------------|--------------|--------------|
| M | 50 | 0.5 | owner | unemploye | 0 | 0 | 00145 | NO |
| M | 19 | 10 | rent | labourer | 0.8 | 0 | 00140 | NO |
| F | 52 | 15 | owner | creative | 5.5 | 14 | 00000 | YES |
| M | 22 | 2.5 | rent | creative | 2.6 | 0 | 00000 | YES |
| M | 29 | 13 | owner | driver | 0.5 | 0 | 00228 | NO |
| F | 16 | 0.3 | owner | unemploye | 0 | 01 | 00160 | NO |
| M | 23 | 11 | owner | professio | 0.5 | 01 | 00100 | YES |
| F | 27 | 3 | owner | manager | 2.8 | 01 | 00280 | NO |
| F | 19 | 5.4 | owner | guard_etc | 0.3 | 0 | 00080 | NO |
| F | 27 | 0.3 | owner | manager | 0.1 | 01 | 00272 | NO |
| M | 34 | 4 | rent | guard_etc | 8.5 | 07 | 00195 | YES |
| M | 20 | 1.3 | rent | labourer | 0.1 | 0 | 00140 | NO |
| M | 34 | 1.3 | owner | guard_etc | 0.1 | 0 | 00440 | NO |

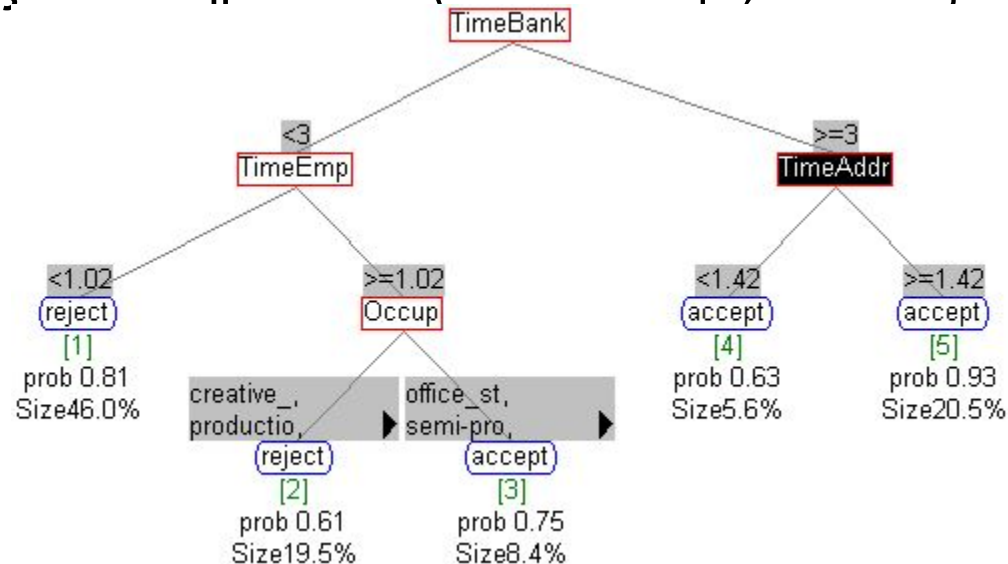
- Γνωρίσματα δανειολήπτη: Φύλο, ηλικία, μονιμότητα κατοικίας, ιδιοκτησία, επάγγελμα, χρόνια εργασίας, χρόνια πελάτης, ετήσιες δαπάνες σπιτιού, **Αποπλήρωσε ή όχι**
- Υπάρχουν κανόνες που θα με βοηθήσουν να αποφασίσω αν θα εγκρίνω μια νέα αίτηση δανείου;

Παράδειγμα

- Τετριμμένη γνώση: **Αν** είναι άνεργος και δεν έχει μόνιμη κατοικία **τότε** δεν αποπληρώνει το δάνειο
 - Το **αν** υποστηρίζεται από 2/13 περιπτώσεις αλλά και στις 2 ισχύει το **τότε**.
- Μη επικυρωμένη γνώση: **Αν** έχουν ιδιόκτητο σπίτι **τότε** αποπληρώνουν το δάνειο.
 - Υπάρχουν αρκετά παραδείγματα (9/13) αλλά δεν επιβεβαιώνονται όλα (αποπληρώνουν μόνο οι 2 στους 9)
- Μη κατανοητή γνώση: **Αν** εμφανίζουν μηδενικές ετήσιες δαπάνες **τότε** αποπληρώνουν το δάνειο.
 - Το **αν** υποστηρίζεται από 2/13 περιπτώσεις, και στις 2 ισχύει το τότε, αλλά δεν έχουμε στοιχεία για να το εξηγήσουμε.

Παράδειγμα

- Προηγούμενα άγνωστη και πιθανά χρήσιμη: Όσοι είναι πάνω από 3 χρόνια στην ίδια κατοικία και πάνω από 1,42 χρόνια πελάτες αποπληρώνουν: (YES \Leftrightarrow accept, NO \Leftrightarrow reject)



- Το δέντρο αυτό μπορεί να χρησιμοποιηθεί για να προτείνει πως θα χειριστούμε νέες αιτήσεις

Γιατί Εξόρυξη Γνώσης;

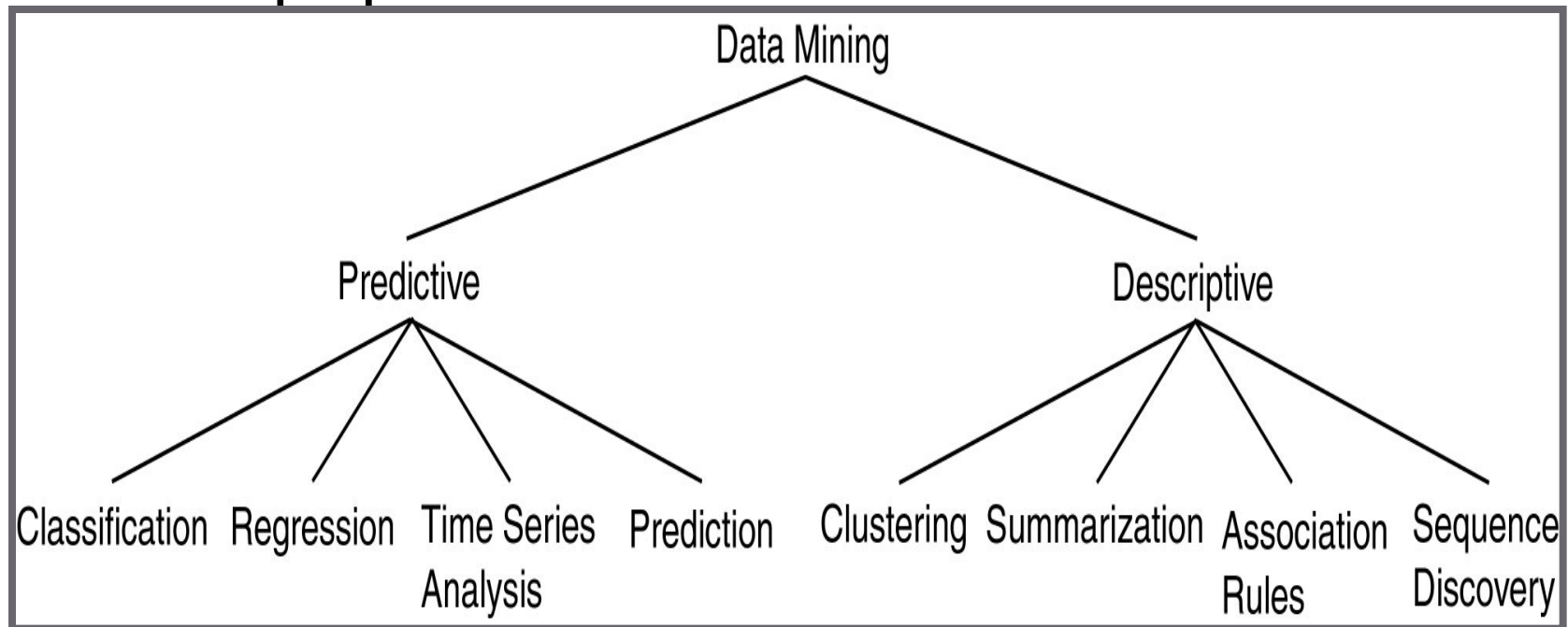
- Πολλά δεδομένα συγκεντρώνονται και εισάγονται σε αποθήκες δεδομένων
 - Web δεδομένα, ηλ. Εμπόριο, δείκτες μετοχών
 - Αγορές σε πολύ-καταστήματα/αλυσίδες
 - Συναλλαγές με τράπεζες/πιστωτικές κάρτες
 - Εφαρμογές κοινωνικής δικτύωσης
- Τα δεδομένα συλλέγονται και αποθηκεύονται με τρομερή συχνότητα (GB/hour)
 - Απομακρυσμένοι αισθητήρες (remote sensors) σε δορυφόρους
 - Τηλεσκόπια στον ουρανό
 - Microarrays που παράγουν γονιδιακά δεδομένα
 - Επιστημονικές προσομοιώσεις που παράγουν terabytes δεδομένων

Γιατί Εξόρυξη Γνώσης;

- Τα δεδομένα έχουν πολύπλοκες σχέσεις μεταξύ τους που δύσκολα ανιχνεύονται
- Αν βρούμε τις κρυμμένες σχέσεις (γνώση) τότε κερδίζουμε
 - Μεγάλος ανταγωνισμός
 - Παροχή καλύτερων υπηρεσιών (fraud detection, targeting marketing)
 - Εξατομίκευση υπηρεσιών (personalization)
- Η εξόρυξη δεδομένων μπορεί να βοηθήσει τους
 - Στην κατηγοριοποίηση και την τμηματοποίηση των δεδομένων
 - Στην Διατύπωση Υποθέσεων
 - Στην καλύτερη οπτικοποίηση της πληροφορίας

Στόχος της εξόρυξης

- Να ταιριάζει τα δεδομένα σε κάποιο μοντέλο:
 - Περιγραφικό
 - Προγνωστικό



Τι δεν είναι η εξόρυξη;

- Δεν είναι ΒΔ και SQL

- Οι ΒΔ οργανώνουν τα δεδομένα και η SQL χρησιμοποιείται για να ελέγξει αν τα δεδομένα επικυρώνουν (ή όχι) τη γνώση που ήδη έχουμε
- Η ΒΔ περιέχει λειτουργικά δεδομένα
- Με την SQL προσδιορίζουμε σαφείς πληροφοριακές ανάγκες και παίρνουμε ως απάντηση υποσύνολα της ΒΔ

- Παράδειγμα

- Βρες όσους πήραν δάνειο και διαμένουν σε ενοίκιο
- Χώρισε αυτούς που αποπλήρωσαν το δάνειο σε 4 ομάδες
- Ποιος είναι ο μέσος χρόνος καθυστέρησης στην αποπληρωμή του δανείου
- Να προχωρήσω την αίτηση για το συγκεκριμένο δάνειο; Πρόκειται να αποπληρωθεί;

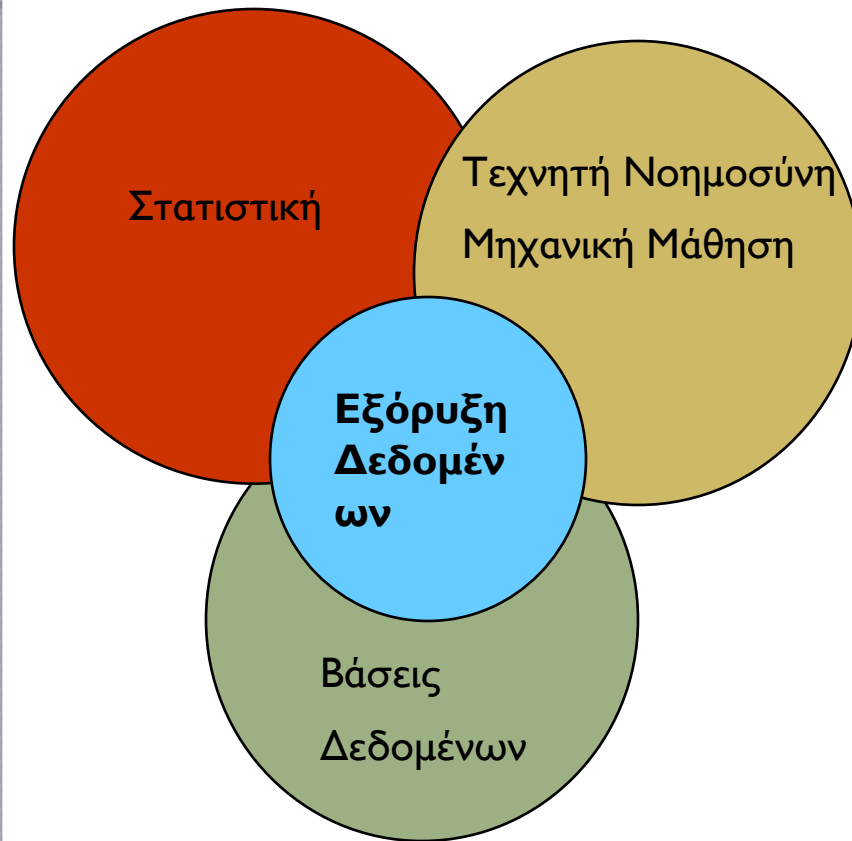
Τι δεν είναι η εξόρυξη;

- Έμπειρο σύστημα (Expert system)
 - ένα έμπειρο σύστημα διαθέτει πρότερη γνώση του πεδίου εφαρμογής, διατυπωμένη με συστηματικό τρόπο, και κανόνες για την εξαγωγή συμπερασμάτων
 - π.χ. Ένα σύστημα που θα υπολογίζει αμέσως τις δόσεις στις οποίες θα πρέπει να αποπληρωθεί ένα νέο δάνειο, ώστε να μεγιστοποιήσει την πιθανότητα αποπληρωμής του
- Στατιστικό πρόγραμμα
 - ένα στατιστικό πρόγραμμα έχει πρότερη γνώση των δεδομένων και προσαρμόζει την επεξεργασία σε αυτά, αλλά και στη γνώση που θέλουμε να εξάγουμε
 - Με βάση τα ιστορικά στοιχεία της τράπεζας υπάρχει συσχέτιση μεταξύ της αποπληρωμής του δανείου και των ετήσιων δαπανών και πόσο ισχυρή είναι;

Ιδιαιτερότητες της εξόρυξης γνώσης

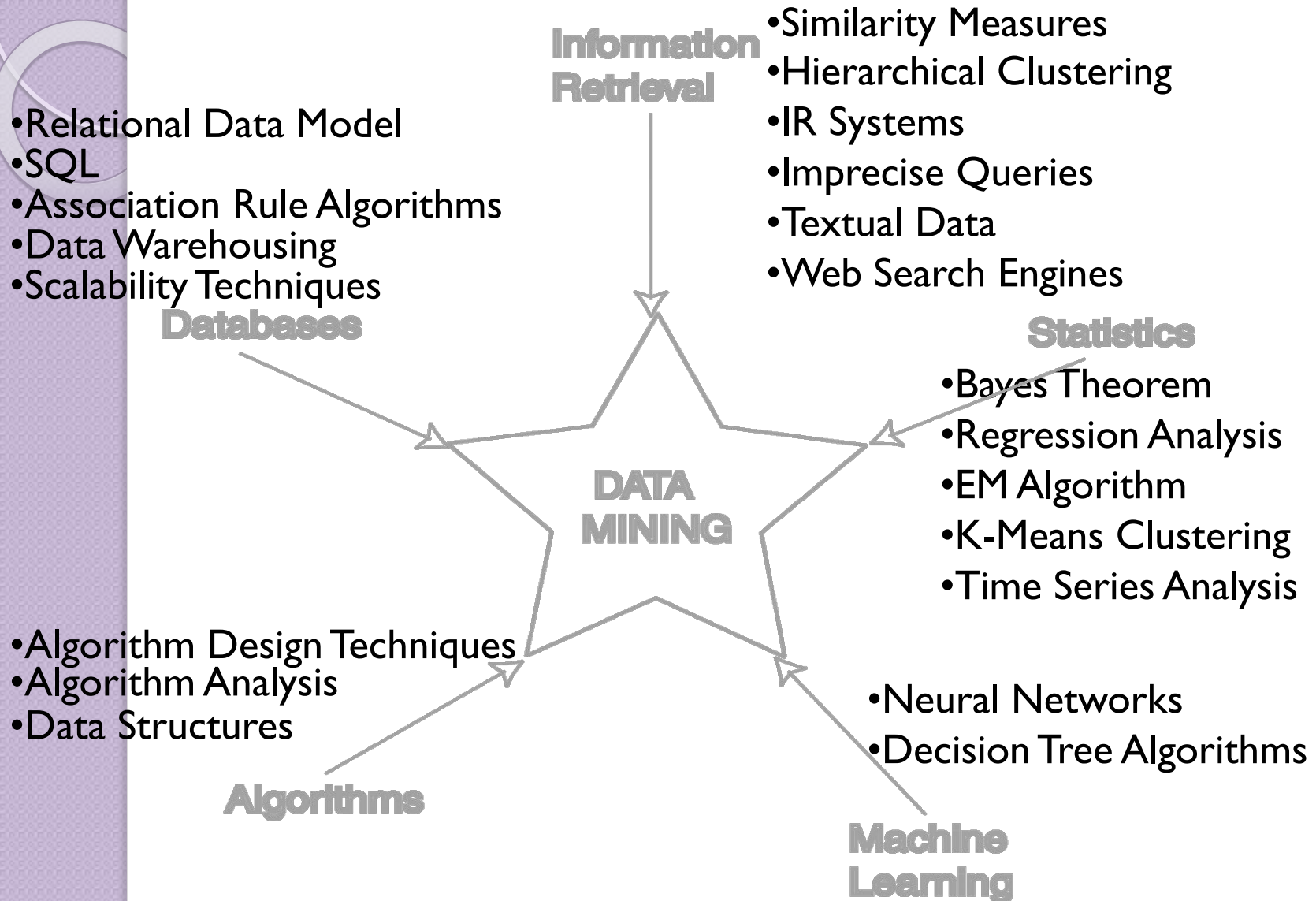
- Συχνά υπάρχει πληροφορία «κρυμμένη» στα δεδομένα, η οποία δεν είναι προφανής και αναμενόμενη
 - Χρειάζονται πολλές και προσεκτικά σχεδιασμένες ενέργειες για να την ανακαλύψουμε
 - Ενδέχεται να μην την ανακαλύψουμε αν δεν ελέγχουμε όλα τα δεδομένα για όλα τα πιθανά ενδεχόμενα
- Συχνά η εξόρυξη γνώσης χρησιμοποιεί στατιστικές μεθόδους, παράγει και διατυπώνει γνώση με συστηματικό τρόπο, εφαρμόζεται στα δεδομένα μιας ΒΔ

Οι «ρίζες» της Εξόρυξης Δεδομένων



- Πρέπει να αντιμετωπίσει:
 - Το τεράστιο μέγεθος των δεδομένων
 - Το μεγάλο αριθμό διαστάσεων (παραγόντων που μπορούμε να λάβουμε υπόψη)
 - Την μη ομοιογενή και την κατανεμημένη φύση των δεδομένων

Data Mining Development



Εφαρμογές

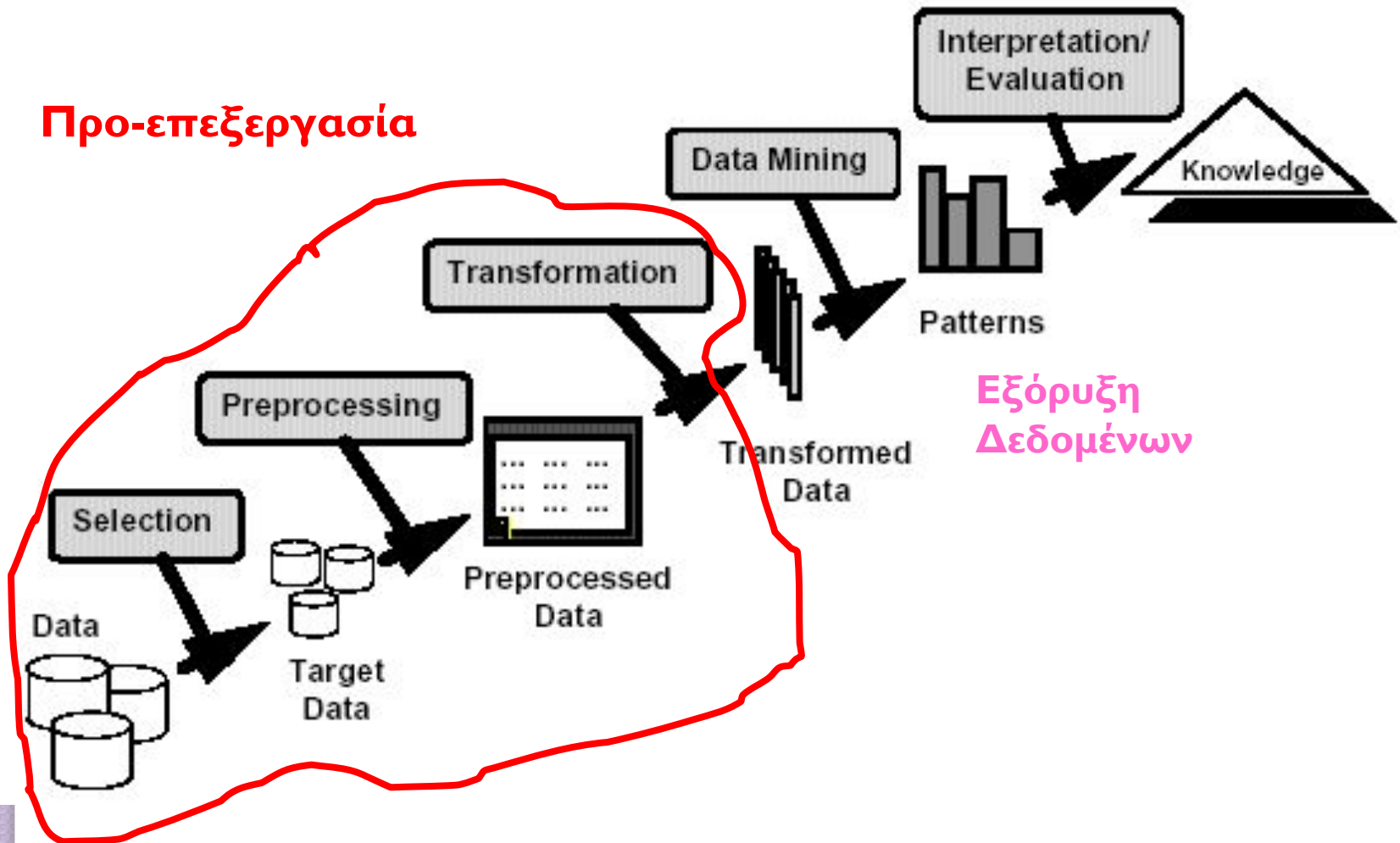
- Εξόρυξη στο διαδίκτυο
- Μηχανές αναζήτησης με βάση τις ερωτήσεις που υποβάλλονται και τις επιλογές στα αποτελέσματα (clickstream analysis)
- Εξατομίκευση περιεχομένου και υπηρεσιών με βάση τα δεδομένα χρήσης (προσπέλασης)
- Πακέτα προσφορών σε πολυκαταστήματα με βάση τις αγορές
- Και άλλα: αποτελέσματα επιστημονικών πειραμάτων, κίνηση μετοχών, βιολογικά δεδομένα κλπ



Η διαδικασία της εξόρυξης γνώσης

Η διαδικασία εξόρυξης γνώσης

Προ-επεξεργασία



Προ-επεξεργασία

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

Καθαρισμός δεδομένων

Τα πραγματικά δεδομένα είναι

- Ελλιπή - incomplete:
 - Λείπουν τιμές γνωρισμάτων (δεν καταγράφηκαν, καταγράφηκαν λανθασμένα),
 - Λείπουν ενδιαφέροντα γνωρίσματα (δε θεωρήθηκαν σημαντικά ή δεν ήταν διαθέσιμα),
 - 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

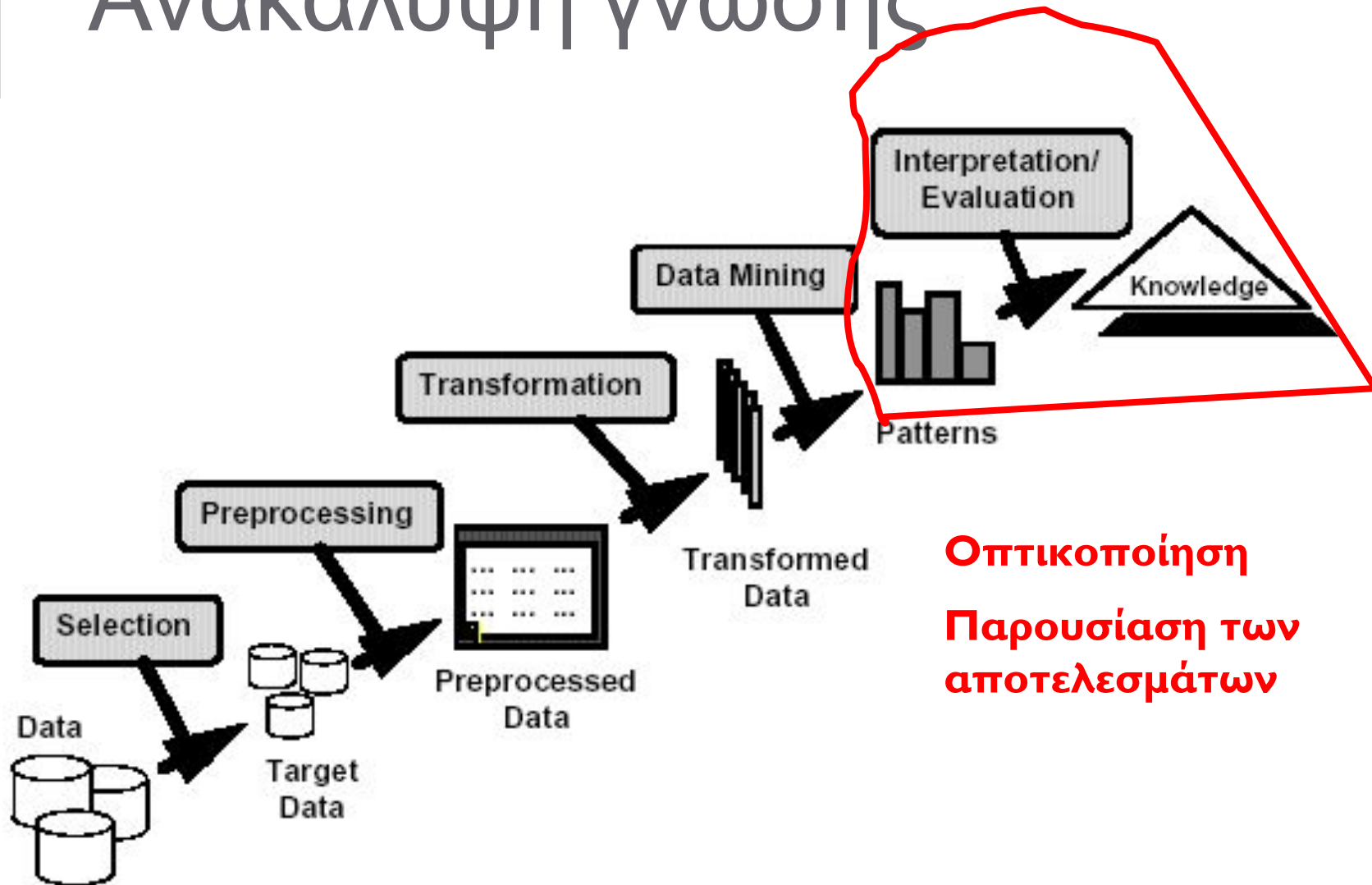
Ενοποίηση δεδομένων

- 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

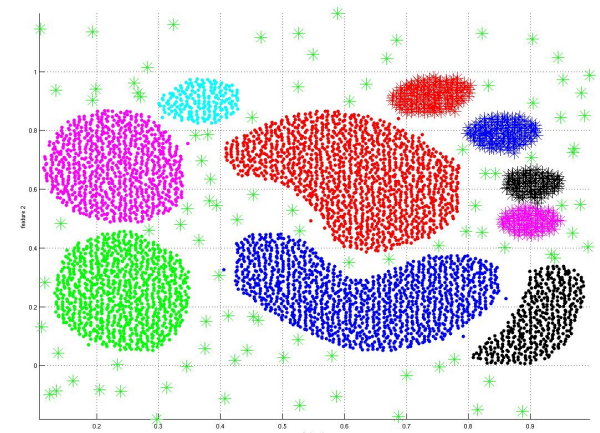
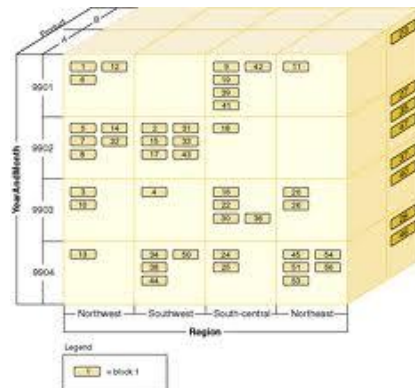
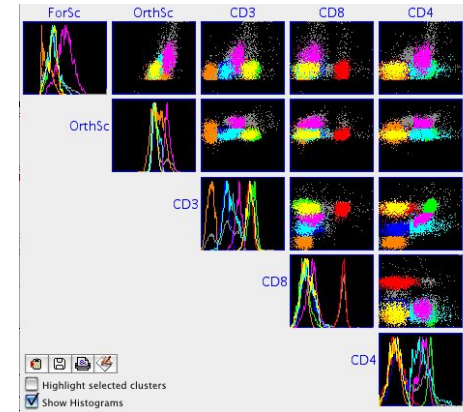
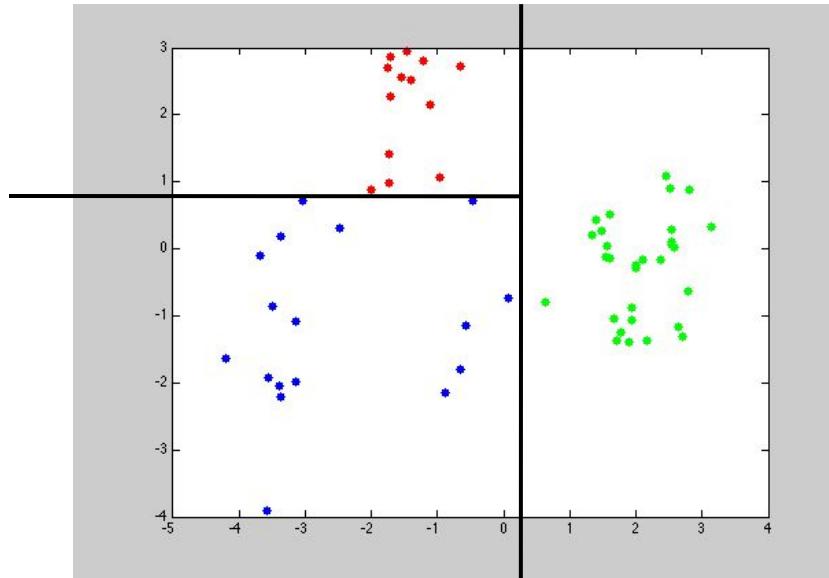
Μετασχηματισμός δεδομένων

- Ο συνδυασμός δεδομένων από διαφορετικές πηγές συχνά απαιτεί προσαρμογή των δεδομένων
- Discretization (μετασχηματισμός σε μια διακριτή τιμή) ή binarization (μετασχηματισμός σε δυαδική τιμή)
- Variable transformation – μετασχηματισμοί των τιμών των μεταβλητών
 - π.χ. Κανονικοποίηση

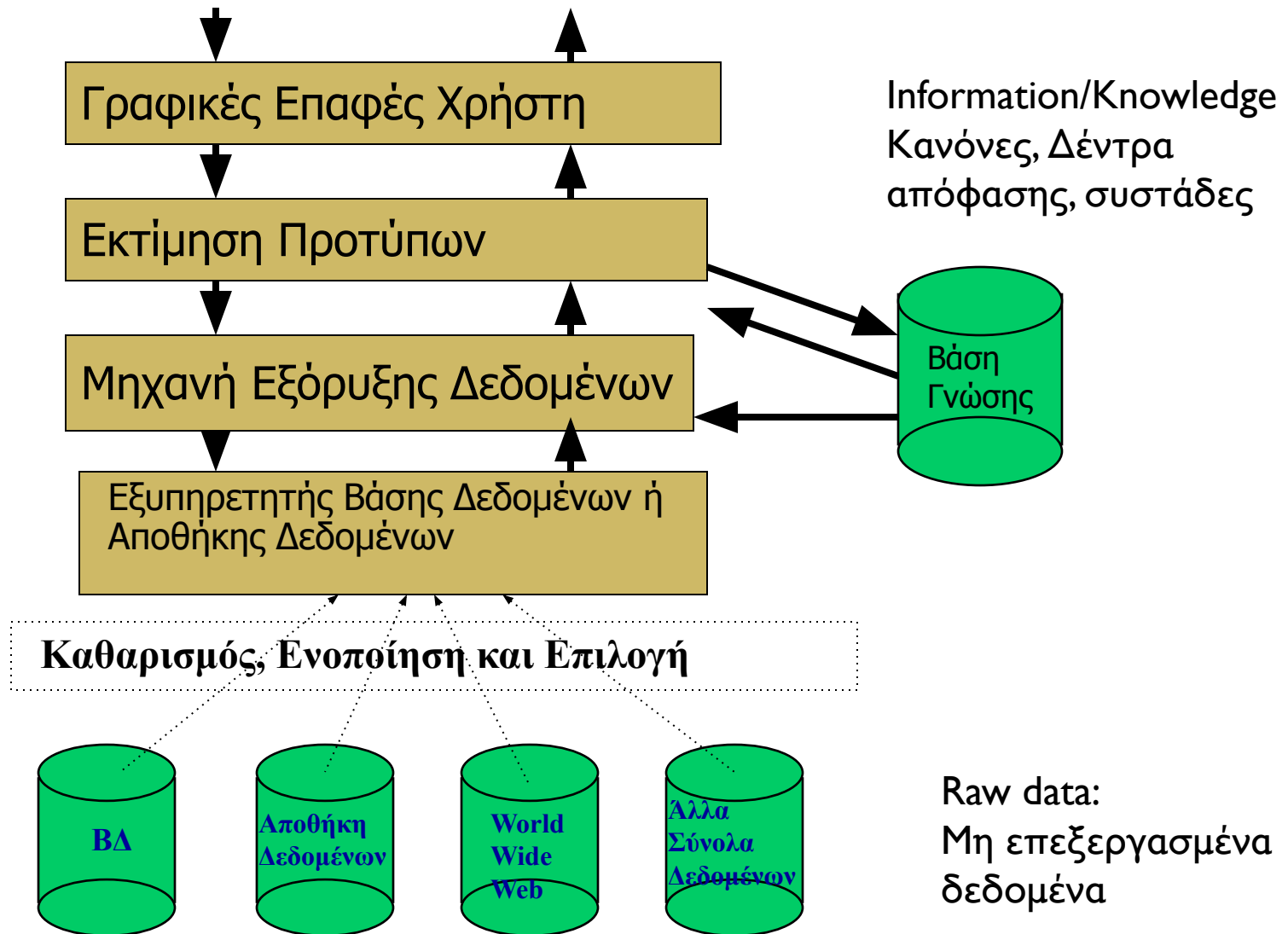
Ανακάλυψη γνώσης



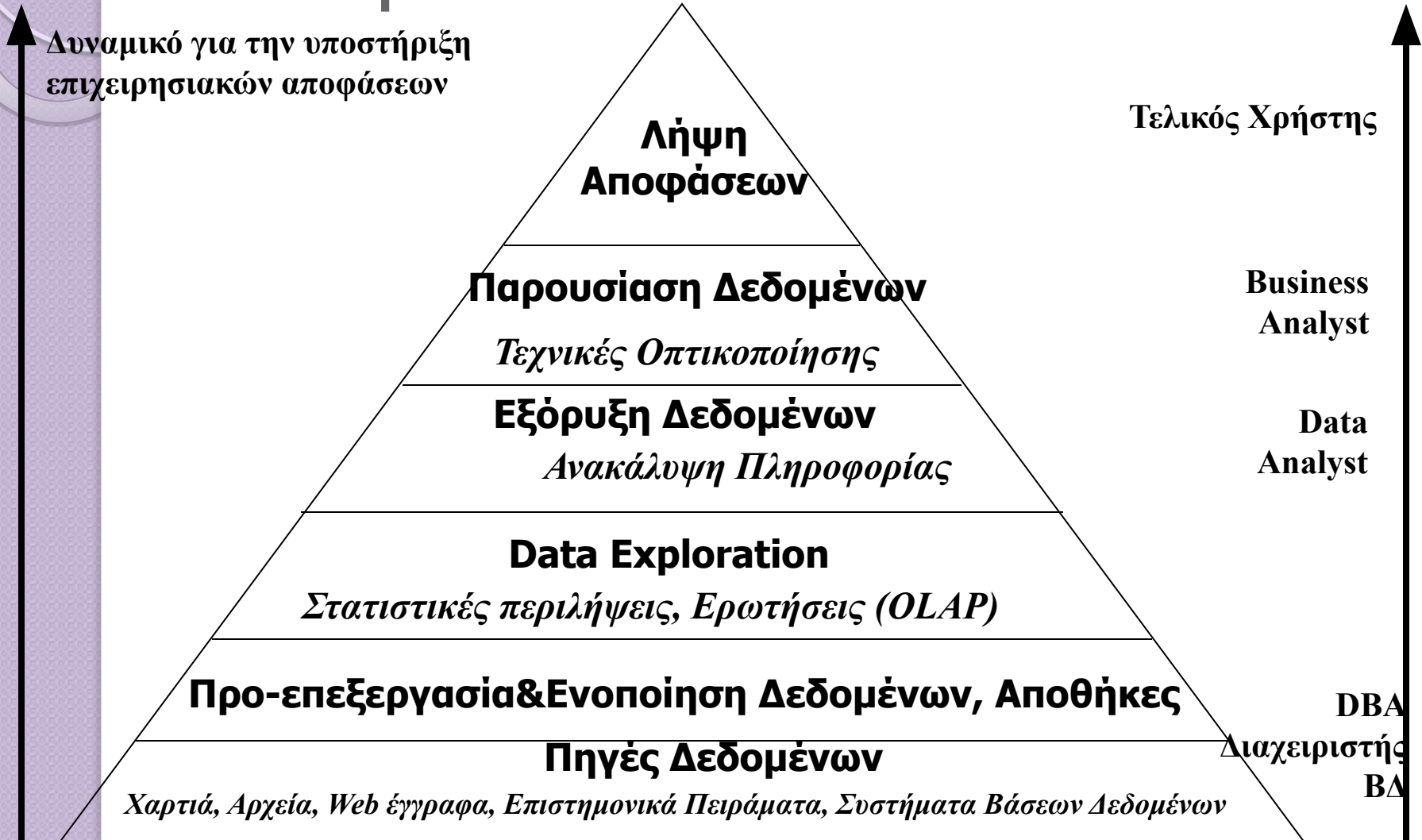
Παράδειγμα



Αρχιτεκτονική ενός συστήματος DM



Υποστήριξη λήψης αποφάσεων



Δυσκολίες

- Μεγάλος όγκος δεδομένων
 - Οι αλγόριθμοι πρέπει να έχουν μικρή πολυπλοκότητα και υψηλή κλιμάκωση ώστε να διαχειρίζονται πολλά δεδομένα
- Πολυδιάστατα και Πολύπλοκα δεδομένα
 - Ροές δεδομένων (data streams) π.χ. δεδομένα αισθητήρων, ροές ειδήσεων
 - Χρονολογικές σειρές (time-series), χρονικά και ακολουθιακά δεδομένα
 - Ημι-δομημένα δεδομένα, γραφήματα, δίκτυα
 - Ετερογενείς πηγές δεδομένων
 - Χωρικά και χωροχρονικά δεδομένα, πολυμέσα, δεδομένα στο web, πολυγλωσσικά δεδομένα

Χρειαζόμαστε νέες και εξειδικευμένες εφαρμογές

Παράδειγμα: Μηχανές αναζήτησης

- Οι μηχανές αναζήτησης αρχικά χρησιμοποιούσαν τις λέξεις που περιέχονταν στις σελίδες – με αποτέλεσμα να παραπλανούνται εύκολα
- Η Google κατάφερε με τον αλγόριθμό της (PageRank) να πετύχει πιο αξιόπιστα αποτελέσματα γιατί βασίστηκε στις πληροφορίες που μεταφέρουν τα links προς μια σελίδα
- Ο Sergey Brin και ο Larry Page ήταν σπουδαστές στο Stanford σε θέματα ΒΔ και εξόρυξης δεδομένων το 1998.

Η βαθμονόμηση και η παραγωγή συστάσεων είναι ένα πεδίο εξόρυξης γνώσης με μεγάλο πρακτικό και ερευνητικό ενδιαφέρον

Παράδειγμα: Προώθηση προϊόντων

- Οι περισσότερες εταιρίες που κάνουν αποκλειστική προώθηση προϊόντων χρησιμοποιούν εργαλεία data mining
- Οι περισσότερες οικονομικές εταιρίες χρησιμοποιούν μοντέλα πελατών
- Η μοντελοποίηση είναι πιο εύκολη από την αλλαγή της συμπεριφοράς των πελατών
- Η Verizon Wireless μείωσε την απώλεια

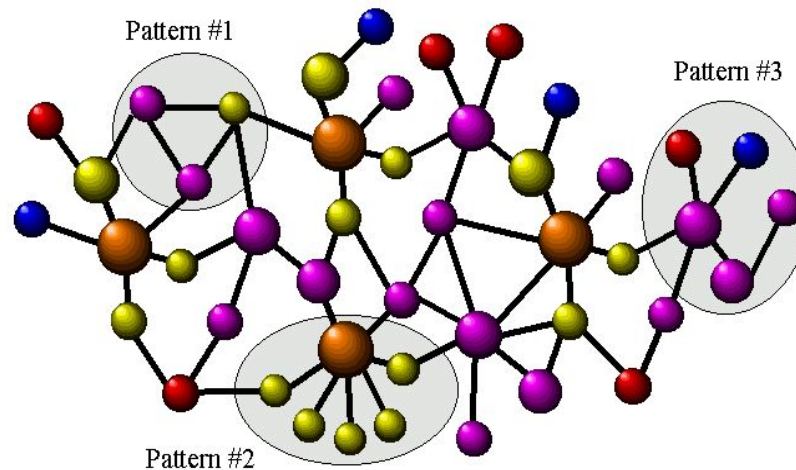
Ο διαχωρισμός των πελατών σε ομάδες μπορεί να οδηγήσει σε καλύτερα στοχευμένες υπηρεσίες

Παράδειγμα: Ασφάλεια και προστασία από απάτη

- Ανίχνευση απάτης σε πιστωτικές κάρτες
- Ξέπλυμα χρήματος
- Παραβιάσεις ασφάλειας
 - NASDAQ Sonar system
- Τηλεφωνική απάτη
 - AT&T, Bell Atlantic, British Telecom/MCI
- Ανίχνευση βιοτρομοκρατίας με ανάλυση δεδομένων αισθητήρων στους Ολυμπιακούς

Η πρόβλεψη ή έγκαιρη διάγνωση αρνητικών περιστατικών μπορεί να μειώσει τις συνέπειες

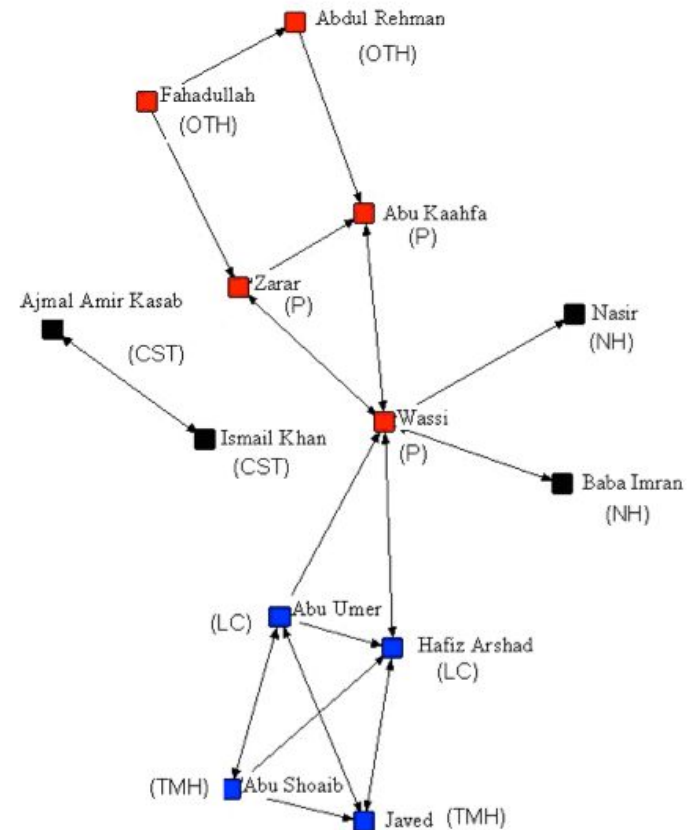
Παράδειγμα: Ανάλυση συνδέσμων



- Μπορούμε να βρούμε χρήσιμη πληροφορία, περίεργα πρότυπα κλπ.

Παράδειγμα: Τρομοκρατία

- TIA: Terrorism (formerly Total) Information Awareness Program –
 - DARPA program (σταμάτησε)
 - some functions transferred to intelligence agencies
- CAPPs II – παρακολούθηση των στοιχείων πτήσης όλων των επιβατών αεροπορικών πτήσεων
- Η εξόρυξη δεδομένων μπορεί
 - Να παραβιάσει την ιδιωτικότητα
 - Να δημιουργήσει λάθος συναγερμούς (false positives)



Terrorist network for 26/11/2008 Mumbai attack based on intercepted phone calls.

Κριτική

- Αν οι ΒΔ έχουν 5% εσφαλμένα δεδομένα, η ανάλυση 10 εκατ. υπόπτων θα δημιουργήσει 500 χιλ. false positives
- Στην πραγματικότητα τα αναλυτικά μοντέλα συσχετίζουν πολλά δεδομένα για να μειώσουν τα false positives
- Παράδειγμα: Θέλουμε να βρούμε ένα πλαστό κέρμα στα 1000.
 - Με μία ρίψη κάθε νομίσματος δεν μπορούμε
 - Με 30 ρίψεις, ένα πλαστό κέρμα θα ξεχωρίσει με μεγάλη πιθανότητα
 - Αντίστοιχα μπορούμε να βρούμε 19 πλαστά κέρματα σε 100 εκατομ. ρίψεις

Πολυδιάστατο πρόβλημα

- **Δεδομένα προς ανάλυση:** σχεσιακά δεδομένα, συναλλαγές, ροές, αντικειμενόστραφή, αντικειμενοκεντρικά, χωρικά, χωροχρονικά, χρονολογικές σειρές, πολυμέσα, γράφοι
- **Γνώση που εξάγεται:** Διάκριση (discrimination), συσχετίσεις (association), κατηγοριοποίηση (classification), συσταδοποίηση (clustering), τάση/απόκλιση (trend/deviation), εξαιρέσεις (outliers)
- **Τεχνικές που χρησιμοποιούνται:** database oriented, data warehouse (analytical processing OLAP), στατιστική, οπτικοποίηση
- **Προσαρμογή στις εφαρμογές:** Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining κλπ

Λειτουργίες της εξόρυξης γνώσης

- Περιγραφή πολυδιάστατων εννοιών
 - Γενίκευση, σύνοψη και αντιπαραβολή χαρακτηριστικών που έχουν τα δεδομένα, π.χ. «Καλοί» και «Κακοί» δανειολήπτες
 - Συχνά εμφανιζόμενα πρότυπα, π.χ. Γάλα □ Δημητριακά [0.5%, 75%] (confidence, support)
- Κατηγοριοποίηση και πρόβλεψη
 - Κατασκευή μοντέλων (συναρτήσεων) που περιγράφουν και διαχωρίζουν κατηγορίες ή έννοιες για μελλοντική πρόβλεψη π.χ. Ταξινόμησε τις χώρες με βάση το κλίμα τους αν γνωρίζεις τα δεδομένα βροχόπτωσης, ηλιοφάνειας και θερμοκρασίας
 - Πρόβλεψη τιμών που λείπουν αναλύοντας ένα δείγμα ή όλες τις προηγούμενες τιμές

Λειτουργίες της εξόρυξης γνώσης

- Συσταδοποίηση
 - Οι κατηγορίες είναι άγνωστες. Ομαδοποίηση των δεδομένων και δημιουργία νέων κατηγοριών
 - Μεγιστοποίηση της ομοιότητας των αντικειμένων μέσα στις συστάδες και ελαχιστοποίηση της ομοιότητας μεταξύ διαφορετικών συστάδων
- Ανάλυση εξαιρέσεων
 - Outlier: αντικείμενα που δεν είναι συμβατά με τη γενική συμπεριφορά των δεδομένων
 - Διάκριση μεταξύ θορύβου και εξαίρεσης. Ανίχνευση και ειδοποίηση (alert)
- Ανάλυση τάσεων και μεταβολών
 - Τάση και απόκλιση (regression analysis)
 - Εξόρυξη ακολουθιακών προτύπων
 - Ανάλυση περιοδικότητας
 - Ανάλυση με βάση την ομοιότητα

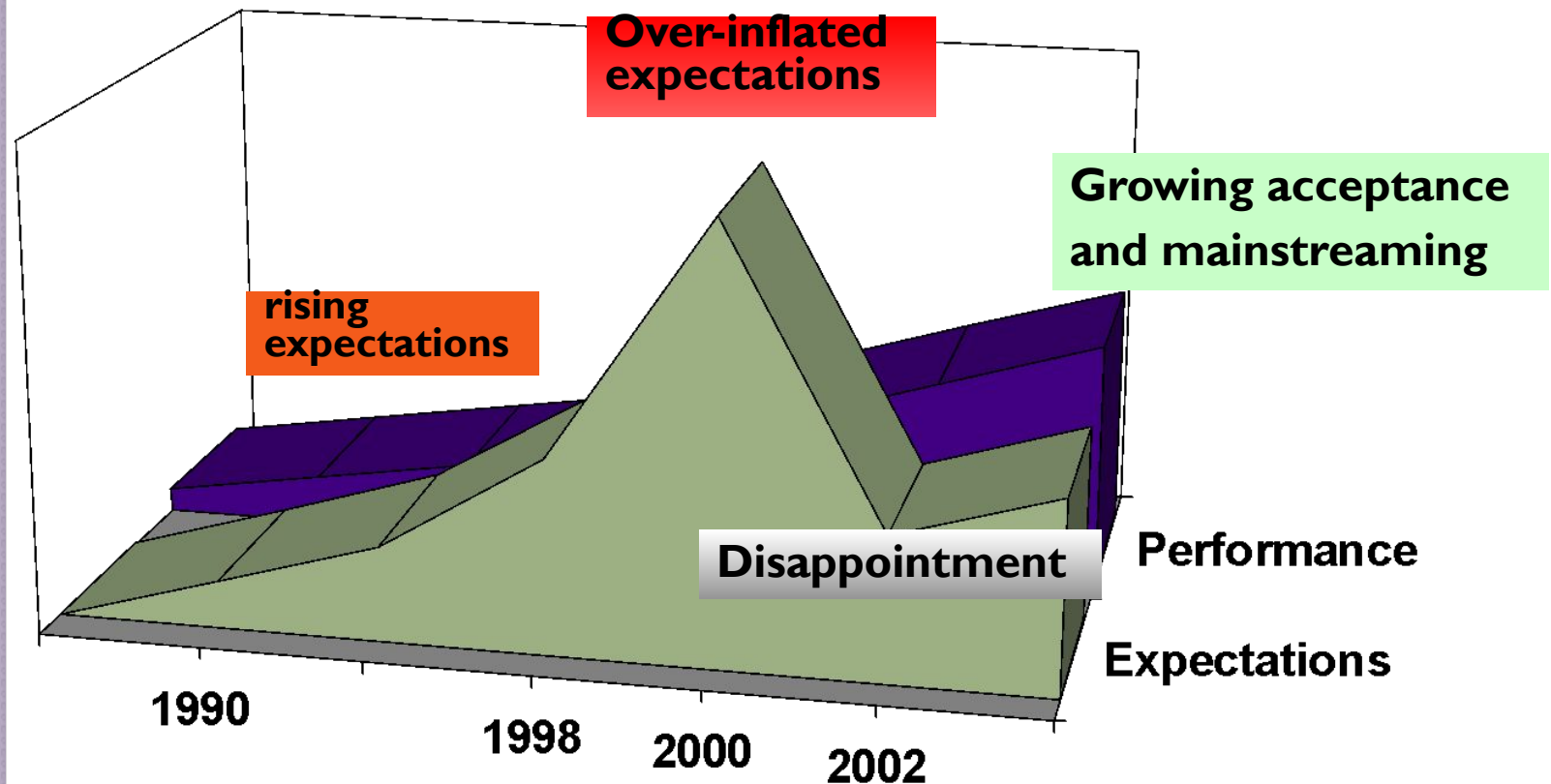
Άλλα θέματα

- Μεθοδολογία εξόρυξης
 - Εξόρυξη διαφορετικών τύπων γνώσης από διαφορετικού τύπου δεδομένα (βιολογικά, ρεύματα, ιστός κλπ)
 - Απόδοση: αποτελεσματικότητα, αποδοτικότητα, κλιμάκωση
 - Αξιολόγηση παραγόμενων προτύπων/γνώσης
 - Ενσωμάτωση προηγούμενης γνώσης
 - Ολοκλήρωση της εξαγόμενης γνώσης
 - Διαχείριση θορύβου και ελλιπών δεδομένων
 - Κατανεμημένα και αυξητική εξόρυξη γνώσης
- Αλληλεπίδραση με το χρήστη
 - Γλώσσες ερωτήσεων εξόρυξης δεδομένων
 - Έκφραση και οπτικοποίηση αποτελεσμάτων
 - Αλληλεπιδραστική (διαλογική) εξόρυξη γνώσης (π.χ. στις μηχανές αναζήτησης)
- Εφαρμογές και Κοινωνική επίδραση
 - Εξόρυξη με βάση το πεδίο αναφοράς
 - Προστασία δεδομένων, ευαίσθητα δεδομένα

Ανακεφαλαίωση

- Εξόρυξη δεδομένων: **ανακάλυψη προτύπων** γνώσης **που έχουν ενδιαφέρον** από τα δεδομένα
- Διαδικασίες που περιλαμβάνει:
 - Καθαρισμό δεδομένων, ολοκλήρωση, επιλογή, μετασχηματισμό, εξόρυξη, αξιολόγηση προτύπων, παρουσίαση της παραγόμενης γνώσης
- Μπορεί να εφαρμοστεί σε ποικίλες αποθήκες δεδομένων
- Λειτουργίες: χαρακτηρισμός, διάκριση, συσχέτιση, κατηγοριοποίηση, συσταδοποίηση, ανάλυση τάσεων, εύρεση εξαιρέσεων κλπ.

Η καμπύλη ενδιαφέροντος για την εξόρυξη



Πηγές

- **Data mining and KDD**

- Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, ECML/PKDD, PAKDD, κλπ.
- Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD

- **Database systems**

- Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
- Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., κλπ.

- **AI & Machine Learning**

- Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, κλπ.
- Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, κλπ.

- **Web and IR**

- Conferences: SIGIR, WWW, CIKM, κλπ.
- Journals: WWW: Internet and Web Information Systems,

- **Statistics**

- Conferences: Joint Stat. Meeting, κλπ.
- Journals: Annals of statistics, κλπ.