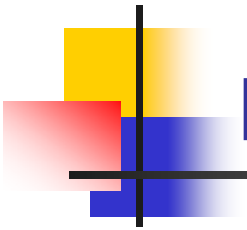
Θεμελιώδεις Αρχές Συστημάτων Βάσεων Δεδομένων

Β. Μεγαλοοικονόμου

Εισαγωγή στην Εξόρυξη Δεδομένων

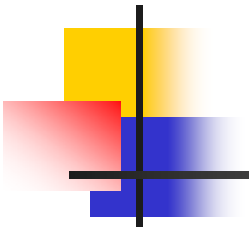


Γενική Επισκόπηση- Σχεσιακό μοντέλο

- Σχεσιακό Μοντέλο - SQL
- Συναρτησιακές εξαρτήσεις & Κανονικοποίηση
- Φυσικός σχεδιασμός, Δεικτοδότηση
- Επεξεργασία / Βελτιστοποίηση ερωτημάτων
- Επεξεργασία δοσοληψιών
- Προηγμένα θέματα
  - Κατανεμημένες βάσεις δεδομένων
  - Αντικειμενοστραφή και Αντικειμενο-σχεσιακά Συστήματα Διαχείρισης Βάσεων Δεδομένων
  - Εξόρυξη δεδομένων

Γενικά

- Κίνητρο: Πού χρειάζεται η εξόρυξη δεδομένων;
- Τι είναι η εξόρυξη δεδομένων;
- Εξόρυξη δεδομένων: Σε τι είδους δεδομένα;
- Λειτουργικότητα εξόρυξης δεδομένων;
- Είναι όλα τα πρότυπα ενδιαφέροντα;
- Ομαδοποίηση συστημάτων εξόρυξης δεδομένων;
- Σημαντικά θέματα στην εξόρυξη δεδομένων;



Κίνητρο

- Έχουμε πολλά δεδομένα αλλά μικρή πληροφορία!
- Πρόβλημα έκρηξης δεδομένων
  - Τα εργαλεία αυτοματοποιημένης συλλογής δεδομένων και η τελευταία τεχνολογία βάσεων δεδομένων οδηγεί σε ένα μεγάλο πλήθος δεδομένων το οποίο αποθηκεύεται σε βάσεις δεδομένων, data warehouses και άλλες αποθήκες δεδομένων
- Λύση: Εξόρυξη δεδομένων
  - Εξόρυξη ενδιαφέρουσας γνώσης (κανόνες, πρότυπα, περιορισμοί) από δεδομένα μεγάλων βάσεων δεδομένων

Εξέλιξη της τεχνολογίας βάσεων δεδομένων

- 1960:
  - Συλλογή δεδομένων, δημιουργία βάσης δεδομένων, συστήματα διαχείρισης πληροφορίας (IMS) και δικτυωτές βάσεις δεδομένων (network DBMS)
- 1970:
  - Σχεσιακό μοντέλο δεδομένων, υλοποίηση σχεσιακού ΣΔΒΔ
- 1980:
  - ΣΣΔΒΔ, προηγμένα μοντέλα δεδομένων (extended-relational, OO, deductive,κτλ.) και συστήματα προσανατολισμένα στην εφαρμογή (χωρικά, επιστημονικά, μηχανικά, κτλ.)
- 1990—2000:
  - Εξόρυξη δεδομένων και αποθηκών δεδομένων (Data mining, data warehousing), βάσεις δεδομένων πολυμέσων, βάσεις δεδομένων στο Παγκόσμιο Ιστό (Web databases)

Τι είναι η εξόρυξη δεδομένων;

- Εξόρυξη δεδομένων (Ανακάλυψη γνώσης από βάσεις δεδομένων):
  - Εξαγωγή ενδιαφέρουσας (μη τετριμμένης, προηγούμενα άγνωστης και πιθανά χρήσιμης) πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων
- Εναλλακτικά ονόματα:
  - Εξόρυξη (Ανακάλυψη) γνώσης σε βάσεις δεδομένων (KDD), εξαγωγή γνώσης (knowledge extraction), ανάλυση δεδομένων/προτύπων(data/pattern analysis) κτλ.
- Τι δεν είναι εξόρυξη δεδομένων;
  - (Deductive) επεξεργασία ερωτημάτων
  - Έμπειρα συστήματα ή μικρής κλίμακας στατιστικά προγράμματα

Τι είναι η εξόρυξη δεδομένων;

- Τώρα που έχουμε συλλέξει τόσα δεδομένα, τι κάνουμε με αυτά;
- Εξαγωγή **χρήσιμων προτύπων** (αυτόματα)
  - Συσχετίσεις (π.χ., ψωμί + βούτυρο --> γάλα)
  - Ακολουθίες (π.χ., χρονικά δεδομένα που σχετίζονται με το χρηματιστήριο)
  - Κανόνες που διαμοιράζουν τα δεδομένα (π.χ. πρόβλημα τοποθεσίας αποθήκευσης)
- Ποια πρότυπα έχουν "**ενδιαφέρον**";

  Περιεχόμενο πληροφορίας, εμπιστοσύνη και υποστήριξη, Απρόσμενο, χρησιμότητα στην λήψη απόφασης, κτλ

Γιατί εξόρυξη δεδομένων; — Πιθανές εφαρμογές

- Ανάλυση βάσης δεδομένων και υποστήριξη απόφασης
  - Ανάλυση και διαχείριση αγοράς
    - Μάρκετινγκ στόχου, διαχείριση σε σχέση με τον πελάτη, ανάλυση καλαθιού αγορών, cross selling, τμηματοποίηση αγοράς
  - Ανάλυση και διαχείριση κινδύνου
    - Πρόβλεψη, απομνημόνευση πελατών, έλεγχος ποιότητας, ανάλυση ανταγωνισμού
  - Εντοπισμός και διαχείριση λάθους
- Άλλες εφαρμογές
  - Εξόρυξη κειμένου (news group, email, documents) και ανάλυση Ιστού.
  - Χωρική εξόρυξη δεδομένων
  - Έξυπνη απάντηση ερωτημάτων

Ανάλυση και διαχείριση αγοράς

- Που είναι οι πηγές των δεδομένων για την ανάλυση; (Δοσοληψίες με κάρτες, εκπτωτικά κουπόνια, κλήσεις παραπόνων από τους πελάτες, κτλ.)
- Μάρκετινγκ στόχου (Εύρεση ομάδων "μοντέλων" πελατών που έχουν κοινά χαρακτηριστικά: εισόδημα, συνήθης αγορές, κτλ.)
- Καθορισμός προτύπων συναλλαγών των πελατών με το χρόνο (Μετατροπή ενός λογαριασμού ενός μόνο ατόμου σε κοινό : γάμος, κτλ.)
- «Cross-market» ανάλυση (Συσχετίσεις μεταξύ προϊόντων πωλήσεων και προβλέψεις με βάσει τις πωλήσεις)
- Προφίλ πελάτη (Ποια προϊόντα αγοράζει ο πελάτης)
- Προσδιορισμός απαιτήσεων πελάτη (Τα καλύτερα προϊόντα για διαφορετικούς πελάτες)
- Παροχή συνοπτικής πληροφορίας (πολυδιάστατες αναφορές)

Ανάλυση και διαχείριση ρίσκου/κινδύνου

- Οικονομικός προγραμματισμός
  - Ανάλυση και πρόβλεψη κίνησης μετρητών
  - Ανάλυση χρονοσειρών και cross-sectional (ανάλυση τάσης, κτλ.)
- Προγραμματισμός πηγών:
  - Σύνοψη και σύγκριση των πηγών και των εξόδων
- Ανταγωνισμός:
  - Έλεγχος ανταγωνιστών και τάσεων της αγοράς
  - Ομαδοποίηση πελατών και απόδοση τιμών με βάσει τις ομάδες αυτές
  - Ορισμός στρατηγικής απόδοσης τιμών σε μία ιδιαίτερα ανταγωνιστική αγορά

Εντοπισμός και διαχείριση λάθους/κλοπών

- Εφαρμογές
  - Φροντίδα υγείας, υπηρεσίες πιστωτικών καρτών, τηλεπικοινωνίες κτλ.
- Προσέγγιση
  - Χρήση στοιχείων ιστορικού για την δόμηση μοντέλων κανονικής και λανθασμένης συμπεριφοράς και χρήση εξόρυξης δεδομένων για τον προσδιορισμό λανθασμένων στιγμιοτύπων
- Παραδείγματα
  - Αυτόματη ασφάλεια: εντοπισμός ομάδων που προκαλούν ατυχήματα για την είσπραξη της ασφάλειας
  - Ξέπλυμα χρήματος: εντοπισμός ύποπτων δοσοληψιών χρημάτων
  - Ιατρική ασφάλεια: εντοπισμός επαγγελματιών ασθενών, ακατάλληλων ιατρικών θεραπειών
  - Εντοπισμός τηλεφωνικού σφάλματος: Μοντέλο τηλεφωνικής κλήσης: προορισμός κλήσης, διάρκεια, χρονική στιγμή ημέρας/εβδομάδας. Ανάλυση προτύπων που αποκλίνουν από τα αναμενόμενα

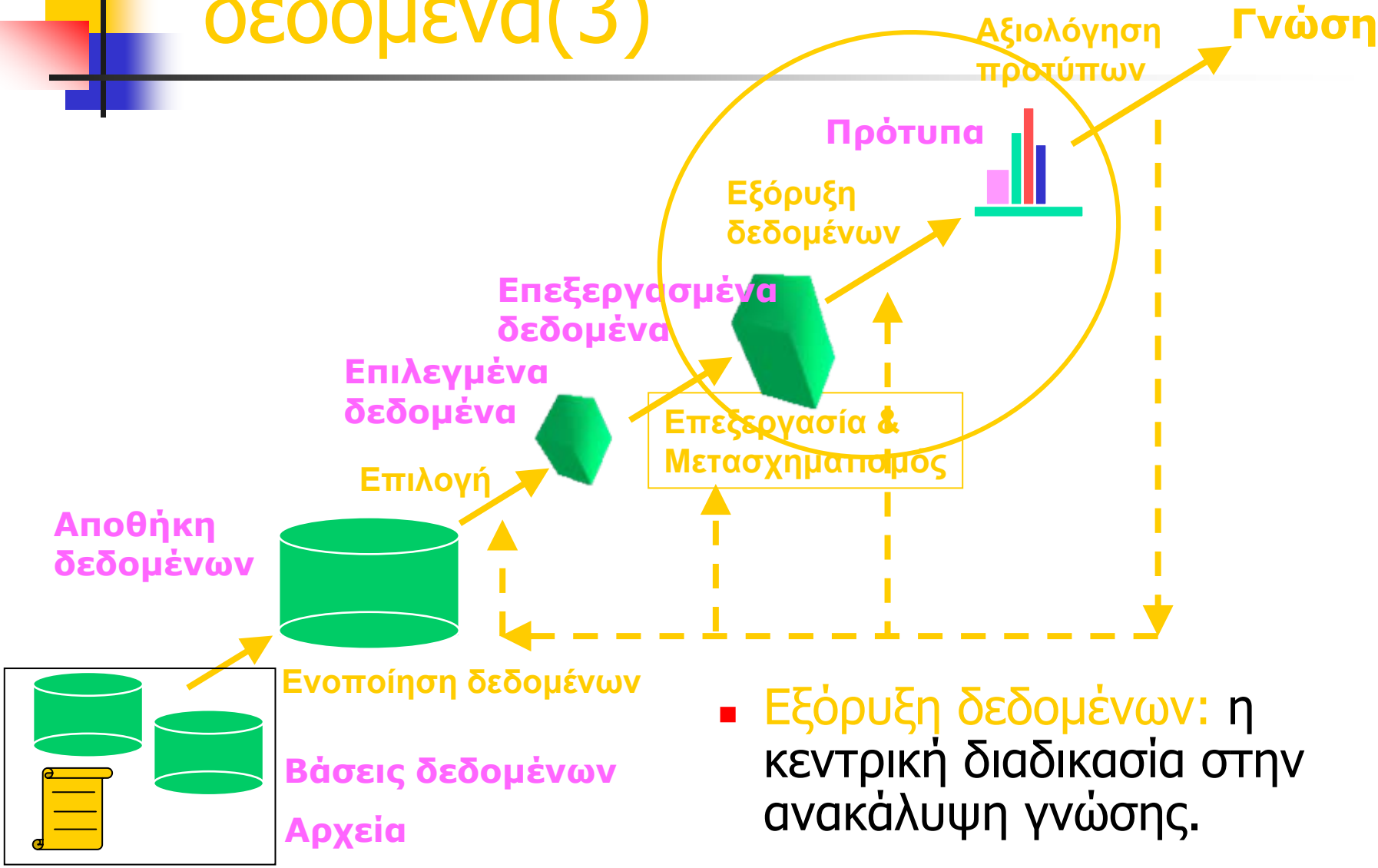
Ανακάλυψη Ιατρικής/ Βιολογικής γνώσης

- Ανακάλυψη συσχετίσεων δομής-λειτουργίας
  - Χαρτογράφηση ανθρώπινου εγκεφάλου (lesion-deficit, task-activation associations)
  - Δομή κυττάρου (cytoskeleton) και λειτουργικότητα ή παθολογία
  - Δομή πρωτεϊνών και λειτουργικότητα
- Ανακάλυψη αιτιακών σχέσεων
  - Συμπτώματα και ιατρικοί όροι
- Ανάλυση ακολουθίας DNA
  - Βιοπληροφορική (microarrays, κτλ.)

Άλλες εφαρμογές

- Αθλήματα
  - Η IBM Advanced Scout ανέλυσε στατιστικά των NBA παιχνιδιών (shots blocked, assists, fouls) για να κερδίσει ένα ανταγωνιστικό πλεονέκτημα υπέρ των New York Knicks και Miami Heat
- Αστρονομία
  - Οι JPL και Palomar Observatory ανακάλυψαν 22 αστέρια με την βοήθεια τεχνικών εξόρυξης δεδομένων
- Internet Web Surf-Aid
  - Η IBM Surf-Aid εφαρμόζει αλγορίθμους εξόρυξης δεδομένων σε αρχεία καταγραφής πρόσβασης στο διαδίκτυο για ιστοσελίδες που σχετίζονται με αγορές προκειμένου να ανακαλυφθούν οι σελίδες προτίμησης και συμπεριφοράς των πελατών, αναλύοντας την αποτελεσματικότητα του Web marketing, βελτιώνοντας την οργάνωση ιστότοπων κτλ.

Ανακάλυψη γνώσης από δεδομένα(3)



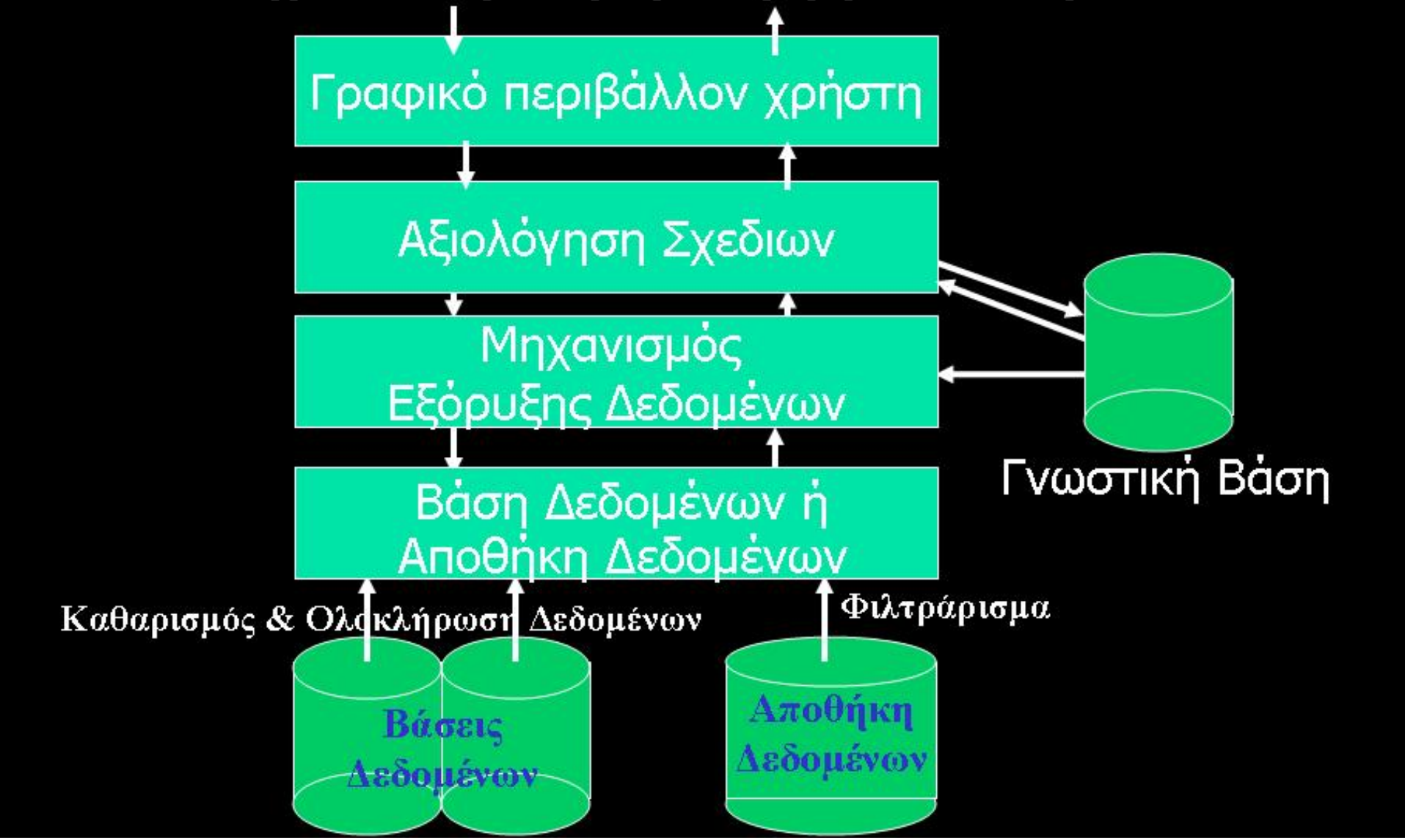
- Εξόρυξη δεδομένων: η κεντρική διαδικασία στην ανακάλυψη γνώσης.

Βήματα της KDD Διαδικασίας

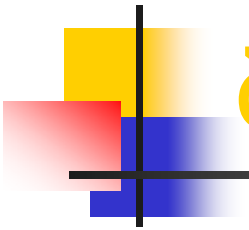
- Εκμάθηση της περιοχής της εφαρμογής:
  - Σχετική πρότερη γνώση και στόχοι της εφαρμογής
  - Δημιουργία ενός στόχου συνόλου δεδομένων: επιλογή δεδομένων
- Καθαρισμός δεδομένων και προεπεξεργασία: (μπορεί να απαιτήσει το 60% της προσπάθειας!)
- Μείωση και μετασχηματισμός δεδομένων:
  - Εύρεση χρήσιμων χαρακτηριστικών, μείωση διαστατικότητας/μεταβλητής, αμετάβλητη αναπαράσταση.
- Επιλογή λειτουργιών της εξόρυξης δεδομένων
  - σύνοψη, ταξινόμηση, οπισθοδρόμηση, συσχέτιση, ομαδοποίηση.
- Επιλογή των πηγών εξόρυξης (s)
- Εξόρυξη δεδομένων: αναζήτησης ενδιαφερόντων προτύπων
- Αξιολόγηση προτύπων και αναπαράσταση γνώσης:
  - οπτικοποίηση, μετασχηματισμός, αφαίρεση περιττών προτύπων, κτλ.
- Χρήση της γνώσης που ανακαλύπτεται

Εξόρυξη Δεδομένων και Ευφυΐα Εταιριών

Αρχιτεκτονική ενός τυπικού συστήματος Εξόρυξης Δεδομένων



Εξόρυξη Δεδομένων: Σε τι είδους δεδομένα;



- Σχεσιακές βάσεις δεδομένων
- Data warehouses
- Βάσεις δεδομένων δοσοληψιών
- Προηγμένες ΒΔ και αποθήκες πληροφορίας
  - Object-oriented (OO)και object-relational (OR) βάσεις δεδομένων
  - Χωρικές βάσεις δεδομένων (ιατρικές, ΒΔ εικόνων δορυφόρου, GIS)
  - Δεδομένα χρονοσειρών και δεδομένα χρόνου
  - Βάσεις δεδομένων κειμένου
  - Βάσεις δεδομένων πολυμέσων (Εικόνα, Βίντεο, κτλ.)
  - Ετερογενείς βάσεις δεδομένων
  - WWW

Λειτουργικότητες εξόρυξης δεδομένων – Πρότυπα που μπορούν να εξαχθούν

- Περιγραφή κεντρικής ιδέας: Χαρακτηρισμός και διάκριση
  - Γενίκευση, σύνοψη, και αντιπαράθεση χαρακτηριστικών δεδομένων, π.χ., ξηρές vs. υγρές περιοχές
- Συσχέτιση (συσχετισμός και αιτιότητα)
  - Πολυδιάστατη και μονοδιάστατη συσχέτιση
  - ηλικία, "20..29") ^ εισόδημα, "20..29K") → αγορές, "PC") [υποστήριξη = 2%, εμπιστοσύνη = 60%]
  - περιέχει, "υπολογιστή") → περιέχει, "λογισμικό") [1%, 75%]
  - Εμπιστοσύνη → y) = P(y|x): βαθμός βεβαιότητας της συσχέτισης
  - Υποστήριξη → y) = P(x ∪y): % των δοσοληψιών που ικανοποιεί ο κανόνας

Λειτουργικότητες εξόρυξης δεδομένων– Πρότυπα που μπορούν να εξαχθούν

- Ομαδοποίηση και πρόβλεψη
  - Εύρεση μοντέλων (if-then κανόνες, δένδρα απόφασης, μαθηματικοί τύποι, νευρωνικά δίκτυα, κανόνες ταξινόμησης) που περιγράφουν και διαχωρίζουν τις κλάσεις ή τις ιδέες για την μελλοντική πρόβλεψη
  - Π.χ., ταξινόμηση χωρών με βάση το κλίμα, ή ταξινόμηση των αυτοκινήτων με βάσει την κατανάλωση
  - Πρόβλεψη: Πρόβλεψη κάποιων αγνώστων τιμών ή τιμών που δεν υπάρχουν
- Ανάλυση ομάδας
  - Η ετικέτα της κλάσης είναι άγνωστη: Ομαδοποίηση δεδομένων για την δημιουργία νέας κλάσης
  - Ομαδοποίηση βασισμένη στην αρχή: μεγιστοποίηση της «intra-class» ομοιότητας και ελαχιστοποίηση της «interclass» ομοιότητας

Λειτουργικότητες εξόρυξης δεδομένων– Πρότυπα που μπορούν να εξαχθούν

- Ανάλυση Outlier
  - Outlier: ένα αντικείμενο δεδομένων το οποίο δεν ακολουθεί την γενική συμπεριφορά των δεδομένων (μπορεί να εντοπισθεί με την χρήση στατιστικών ελέγχων που υιοθετούν ένα πιθανοτικό μοντέλο)
  - Συχνά θεωρείται θόρυβος αλλά είναι χρήσιμο στον εντοπισμό σφάλματος, την ανάλυση σπάνιων γεγονότων
- Ανάλυση τάσης και εξέλιξης
  - Μελετά την κανονικότητα των αντικειμένων των οποίων η συμπεριφορά μεταβάλλεται με τον χρόνο
  - Τάση και απόκλιση: ανάλυση οπισθοχώρησης
  - Εξόρυξη ακολουθιακού προτύπου, ανάλυση περιοδικότητας
  - Ανάλυση με βάση την ομοιότητα

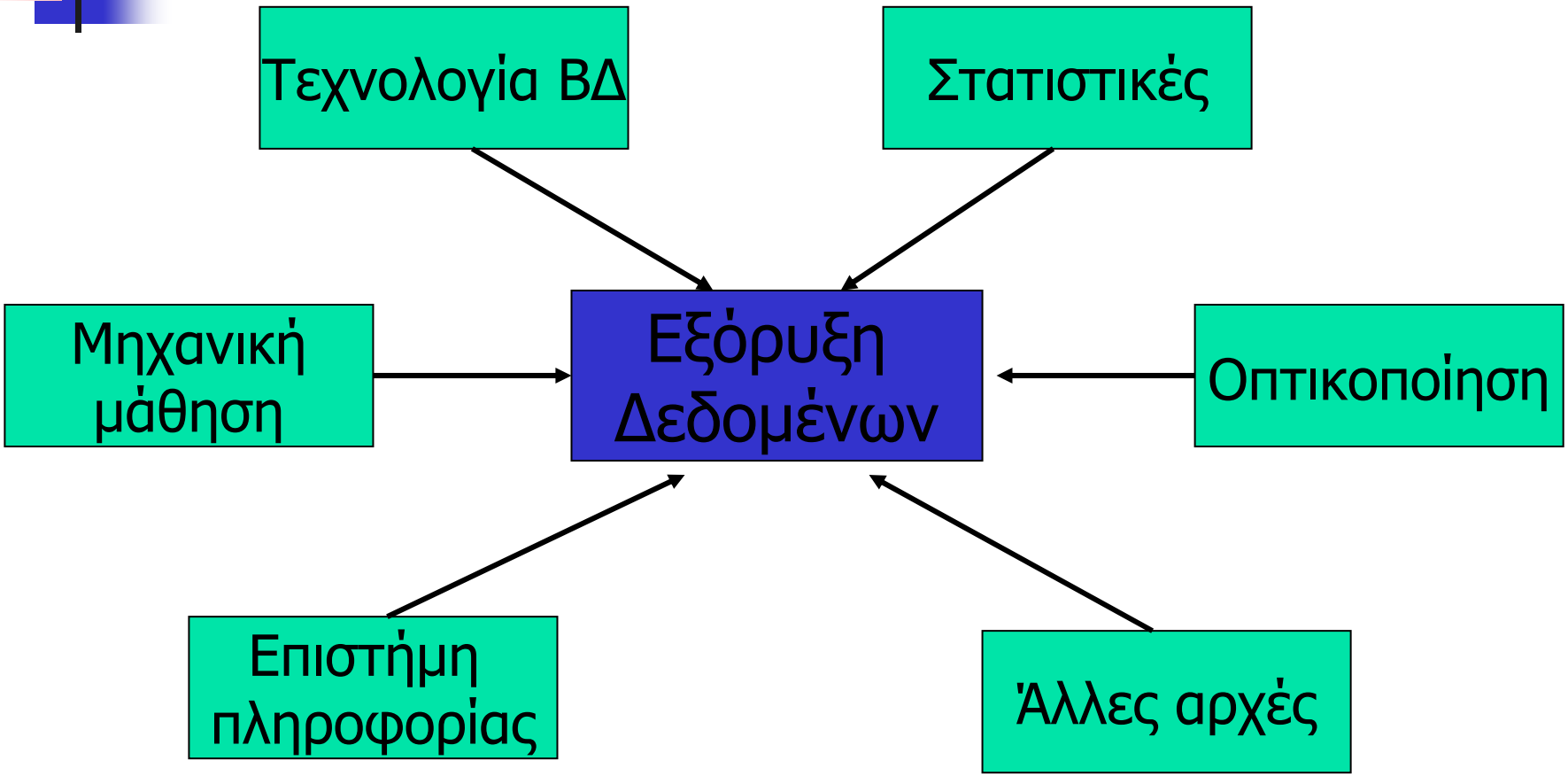
Πότε είναι ενδιαφέρον ένα "Ανακαλυφθέν" πρότυπο;

- Ένα σύστημα/ερώτημα δεδομένων μπορεί να παράγει χιλιάδες προτύπων, τα οποία δεν είναι όλα ενδιαφέροντα.
  - Προτεινόμενη προσέγγιση: Ανθρωποκεντρική, με βάση το ερώτημα, με βάση την εξόρυξη
- **Μετρικές ενδιαφέροντος**: Ένα πρότυπο είναι **ενδιαφέρον** εάν είναι εύκολα κατανοητό από τους ανθρώπους, έγκυρο σε νέα ή πειραματικά δεδομένα με κάποιο βαθμό βεβαιότητας, πιθανά χρήσιμο, καινούριο, ή επικυρώνει κάποια υπόθεση που ο χρήστης επιδιώκει να επαληθεύσει
- **Αντικειμενικές vs. υποκειμενικές μετρικές ενδιαφέροντος:**
  - Αντικειμενικές: βασίζονται σε στατιστικές και δομές προτύπων
  - Υποκειμενικές: βασίζονται στην γνώμη του χρήστη για τα δεδομένα

Μπορούμε να βρούμε όλα τα ενδιαφέροντα πρότυπα;

- Εύρεση όλων των ενδιαφερόντων προτύπων: **Πληρότητα**
  - Μπορεί ένα σύστημα εξόρυξης δεδομένων να βρει όλα τα ενδιαφέροντα πρότυπα;
  - Συσχέτιση vs. Ταξινόμηση vs. Ομαδοποίηση
- Αναζήτηση μόνο των ενδιαφερόντων προτύπων: **Βελτιστοποίηση**
  - Μπορεί ένα σύστημα εξόρυξης δεδομένων να βρει μόνο τα ενδιαφέροντα πρότυπα;
  - Προσεγγίσεις
    - Πρώτα παρήγαγε όλα τα πρότυπα και στην συνέχεια φιλτράρισε αυτά που δεν ενδιαφέρουν
    - Παρήγαγε μόνο τα ενδιαφέροντα πρότυπα - βελτιστοποίηση ερωτήματος

Εξόρυξη δεδομένων: Συμβολή πολλαπλών αρχών



Εξόρυξη δεδομένων: Σχήματα Ταξινόμησης

- Γενική λειτουργικότητα
  - Περιγραφική εξόρυξη δεδομένων
  - Προβλεπτική εξόρυξη δεδομένων
- Διαφορετικές όψεις, διαφορετικές ταξινομήσεις
  - Είδη βάσεων δεδομένων που εξάγονται
  - Είδη γνώσεων που θα ανακαλυφθούν
  - Είδη τεχνικών που θα χρησιμοποιηθούν
  - Είδη εφαρμογών που προσαρμόζονται

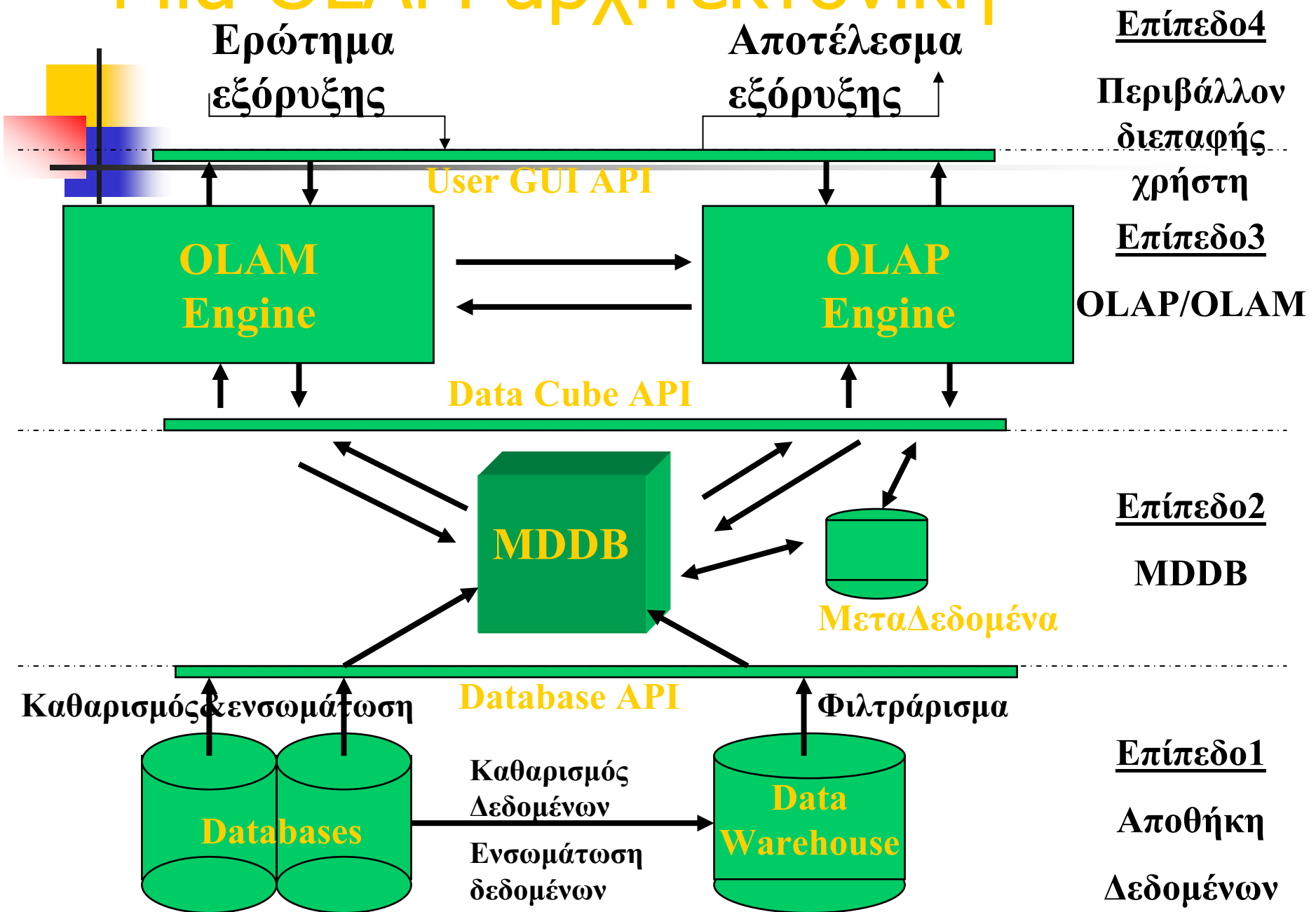
Μία πολυδιάστατη οπτική της ταξινόμησης της εξόρυξης δεδομένων

- **Βάσεις δεδομένων που θα εξαχθούν**
  - Σχεσιακές, δοσοληψιών, αντικειμενοστραφείς, χωρικές, χρονοσειρών, κειμένου, πολυμέσων, ετερογενείς, object-relational, legacy, active, WWW, κτλ.
- **Γνώση που θα εξαχθεί**
  - Χαρακτηρισμός, διάκριση, συσχέτιση, ταξινόμηση, ομαδοποίηση, τάση, απόκλιση και ανάλυση outlier κτλ.
- **Τεχνικές που χρησιμοποιούνται**
  - Στατιστικές, οπτικοποίηση, νευρωνικό δίκτυο, μηχανική μάθηση, Database-oriented, data warehouse (OLAP), κτλ.
- **Εφαρμογές που προσαρμόζονται**
  - Τηλεπικοινωνίες, τραπεζικές εργασίες, ανάλυση σφάλματος, εξόρυξη DNA, εξόρυξη Ιστού κτλ.

OLAP Εξόρυξη: Συγχώνευση Εξόρυξης Δεδομένων και Data Warehousing

- **Συστήματα εξόρυξης δεδομένων, ΣΔΒΔ, Data warehouse συστήματα ένωσης**
  - Χωρίς ένωση, χαλαρή-ένωση, ημι-σφιχτή-ένωση, σφιχτή-ένωση
- **On-line αναλυτική εξόρυξη δεδομένων**
  - Συγχώνευση της εξόρυξης και των OLAP τεχνολογιών
- **Διαδραστική εξόρυξη γνώσης σε πολλά επίπεδα**
  - Ανάγκη εξόρυξης γνώσης και προτύπων σε διαφορετικά επίπεδα γενίκευσης (drilling/rolling, pivoting, slicing/dicing, κτλ.)
- **Συγχώνευση πολλών λειτουργιών εξόρυξης**
  - Χαρακτηρισμένη ταξινόμηση, πρώτα ομαδοποίηση και μετά συσχέτιση

Μία OLAM αρχιτεκτονική

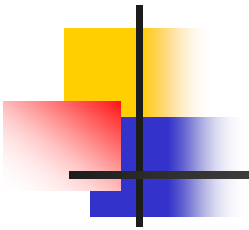


Βασικά ζητήματα στην εξόρυξη δεδομένων

- Μεθοδολογία εξορυξης και αλληλεπίδραση χρηστών
  - Εξόρυξη διαφορετικών ειδών γνώσεων στις βάσεις δεδομένων
  - Διαδραστική εξόρυξη γνώσης σε πολλαπλά επίπεδα αφαίρεσης
  - Ενσωμάτωση της γνώσης που υπάρχει για την καθοδήγηση της διαδικασίας ανακάλυψης
  - Γλώσσες διατύπωσης ερωτημάτων εξόρυξης δεδομένων και ad-hoc εξόρυξη δεδομένων
  - Έκφραση και οπτικοποίηση των αποτελεσμάτων της εξόρυξης δεδομένων
  - Χειρισμός θορύβου και ανολοκλήρωτων δεδομένων
  - Αξιολόγηση προτύπου: το πρόβλημα ενδιαφέροντος
- Απόδοση και εξελιξιμότητα
  - Απόδοση και εξελιξιμότητα των αλγορίθμων
  - Παράλληλες, κατανεμημένες και επαυξητικές μέθοδοι εξόρυξης

Σημαντικά θέματα στην εξόρυξη δεδομένων

- Ζητήματα σχετικά με την ποικιλομορφία των τύπων δεδομένων
  - Χειρισμός σχεσιακών και σύνθετων τύπων δεδομένων
  - Εξόρυξη πληροφορίας από ετερογενείς βάσεις δεδομένων και σφαιρικά πληροφοριακά συστήματα (WWW)
- Ζητήματα σχετικά με τις εφαρμογές και τις κοινωνικές επιδράσεις
  - Εφαρμογή της ανακαλυφθείσας γνώσης
    - Εργαλεία εξόρυξης δεδομένων με βάση τον πεδίο
    - Έξυπνη απάντηση ερωτήματος
    - Έλεγχος διαδικασίας και λήψη απόφασης
  - Ένωση της ανακαλυφθείσας γνώσης με την ήδη υπάρχουσα: πρόβλημα συγχώνευσης της γνώσης
  - Προστασία της ασφάλειας των δεδομένων, της ακαιρεότητας και της μυστικότητας



Σύνοψη

- Εξόρυξη δεδομένων: ανακάλυψη ενδιαφερόντων προτύπων από μεγάλους όγκους δεδομένων
- Μία φυσική εξέλιξη της τεχνολογίας βάσεων δεδομένων, με μεγάλη ζήτηση, με μεγάλο εύρος εφαρμογών
- Μία KDD διαδικασία περιλαμβάνει τον καθαρισμό, την συγχώνευση, και την επιλογή δεδομένων, μετασχηματισμό, εξόρυξη δεδομένων, αξιολόγηση προτύπου και παρουσίαση της γνώσης
- Η εξόρυξη μπορεί να γίνει σε διάφορες αποθήκες πληροφοριών
- Λειτουργικότητες της εξόρυξης δεδομένων: χαρακτηρισμός, επιλογή, συσχέτιση, ταξινόμηση, ομαδοποίηση, ανάλυση τάσης και outlier, κτλ.
- Ταξινόμηση των συστημάτων εξορυξης δεδομένων
- Βασικά ζητήματα στην εξόρυξη δεδομένων