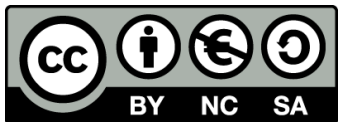
ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Αποθήκες Δεδομένων και Εξόρυξη Δεδομένων

Ενότητα 1: Εισαγωγή

Αναστάσιος Γούναρης, Επίκουρος Καθηγητής
Τμήμα Πληροφορικής ΑΠΘ



Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Εισαγωγή

Κίνητρο, ορισμοί και λειτουργίες

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Περιεχόμενα ενότητας

1. Κίνητρο.
2. Ορισμοί.
3. Ιστορικά Στοιχεία.
4. Λειτουργίες.
5. Πηγές Δεδομένων.

Σκοποί ενότητας

- Γνωριμία με το μάθημα.
- Βασικές έννοιες και ορισμοί.

Κίνητρο: Υποστήριξη Αποφάσεων

- Αυτόματη συλλογή πολύ μεγάλων συνόλων δεδομένων.
  - Πωλήσεις (bar-code scanners).
  - Παγκόσμιος Ιστός (η-εμπόριο).
  - Τράπεζες.
- Ανάγκες Χρήστη → Σωστές Αποφάσεις
  Καλύτερες Υπηρεσίες
  Προσαρμογή στις ανάγκες (CRM)

Η γνώση και η χρήσιμη πληροφορία είναι **κρυμμένη** μέσα στα δεδομένα.

Κίνητρο: Επιστήμες

- Επιστημονικά δεδομένα συλλέγονται με υπερβολικά υψηλούς ρυθμούς (GB-TB/ώρα).
  - Δεδομένα δορυφόρων (NASA, ESA).
  - Τηλεσκόπια.
  - Βιολογικά Δεδομένων (γονίδια - μικροσυστοιχίες).
  - LHC.
  - Επιστημονικές Προσομοιώσεις.
- Οι παραδοσιακές τεχνικές δεν είναι αποδοτικές.
- Η εξόρυξη δεδομένων βοηθά στην απάντηση πιο θεμελιωδών ερωτημάτων, π.χ.:
  - «ποια είναι η σχέση μεταξύ ακραίων καιρικών φαινομένων (κυκλώνων) και υπερθέρμανσης του πλανήτη;»
  - «ποια γονίδια συσχετίζονται με κάποια αρρώστια;»

Η ανάπτυξη των τεχνολογιών ΒΔ

- 1960s:
  - Συλλογή δεδομένων, δημιουργία βάσεων, ιεραρχικό – δικτυακό μοντέλο δεδομένων.
- 1970s:
  - Σχεσιακό μοντέλο, υλοποιήσεις σχεσιακών ΣΔΒΔ.
- 1980s:
  - Σχεσιακά ΣΔΒΔ, προηγμένα μοντέλα (extended-relational, OO, deductive, etc.) και ΣΔΒΔ για συγκεκριμένες εφαρμογές (π.χ., χωρικές, χωροχρονικές, επιστημονικές ΒΔ, κ.ο.κ.)
- 1990s— σήμερα:
  - **Εξόρυξη Δεδομένων (data mining) και Αποθήκες Δεδομένων (Data Warehouses)**, Πολυμεσικές ΒΔ, ΒΔ και Παγκόσμιος Ιστός

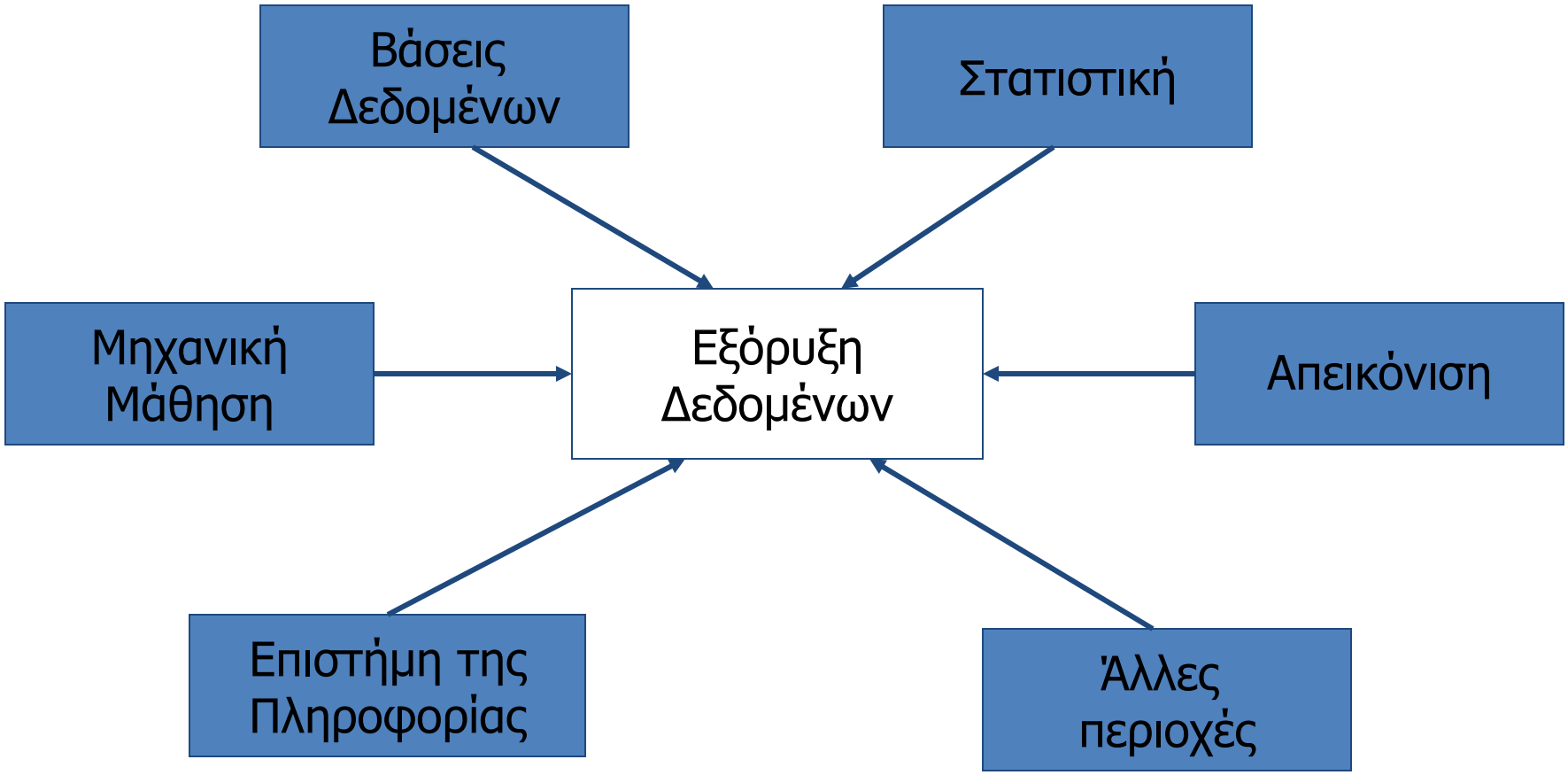
Τί είναι εξόρυξη δεδομένων;

- Εξόρυξη Δεδομένων:
  - Μέρος της διαδικασίας ανακάλυψης γνώσης σε ΒΔ.
  - Εξαγωγή ενδιαφέρουσας (μη-τετριμμένης, υποκρυπτόμενης, άγνωστης προηγουμένως και ενδεχομένως χρήσιμης), πληροφορίας ή προτύπων από δεδομένα σε μεγάλες ΒΔ.
- Εναλλακτικές ονομασίες (οι περισσότερες είναι άστοχες):
  - Ανακάλυψη γνώσης σε ΒΔ (Knowledge discovery in databases - KDD) --ευρύτερη περιοχή.
  - Business intelligence
  - ...

Τί ΔΕΝ είναι εξόρυξη δεδομένων;

- Πολλά από τα αυτοαποκαλούμενα συστήματα εξόρυξης δεδομένων της αγοράς.
- Τεχνικές.
  - (συμπερασματική - deductive) επεξεργασία ερωτημάτων.
  - Έμπειρα συστήματα.
  - Μικρά προγράμματα μηχανικής-στατιστικής μάθησης ML/statistical programs.
  - Συμπερασματική απάντηση ερωτήσεων.
- Ένα πραγματικό σύστημα Εξόρυξης Δεδομένων πρέπει να είναι σε θέση να διαχειρίζεται τεράστιο όγκο δεδομένων.
- Έμφαση σε αποδοτικότητα και κλιμάκωση – επεκτασιμότητα.
  - Χρόνος εκτέλεσης = Ο(μέγεθος ΒΔ).

Συμβολή πολλών περιοχών



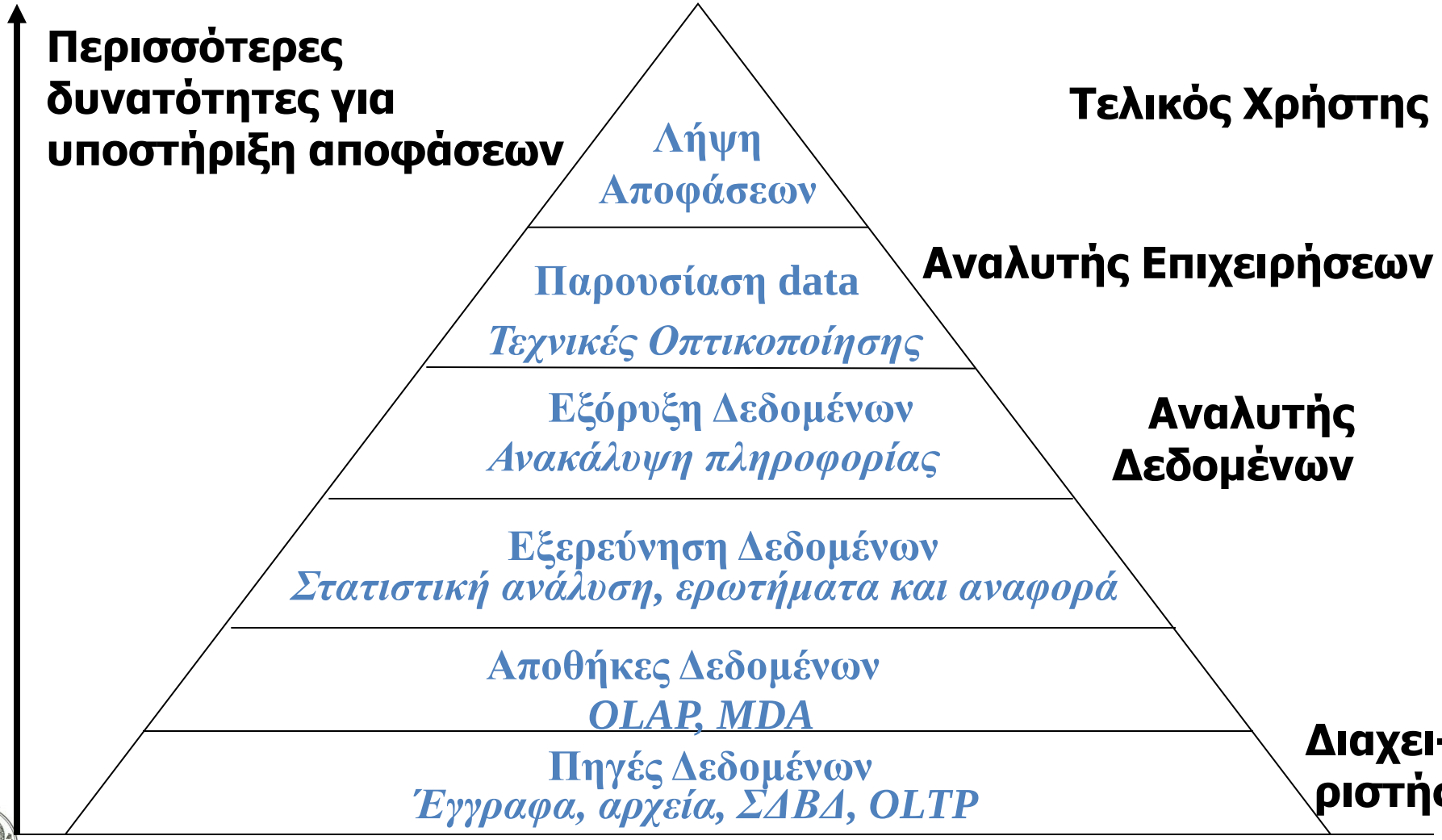
Αποθήκες Δεδομένων – OLAP (1)

- Οι αποθήκες δεδομένων είναι εξειδικευμένες ΒΔ που μπορούν να αποθηκεύουν δεδομένα από πολλαπλές, ετερογενείς βάσεις, οργανωμένες με κοινό σχήμα και διευκολύνουν τη λήψη αποφάσεων.
- OLAP (On-Line Analytical Processing) είναι μία τεχνική ανάλυσης που εκτελείται στις ΑΔ.
  - Είναι διερευνητικής φύσης.
  - Πολύ χρήσιμη, αλλά περισσότερο κατάλληλη για επαλήθευση υποθέσεων.
  - ΣΥΜΠΛΗΡΩΜΑΤΙΚΗ διαδικασία της εξόρυξης.

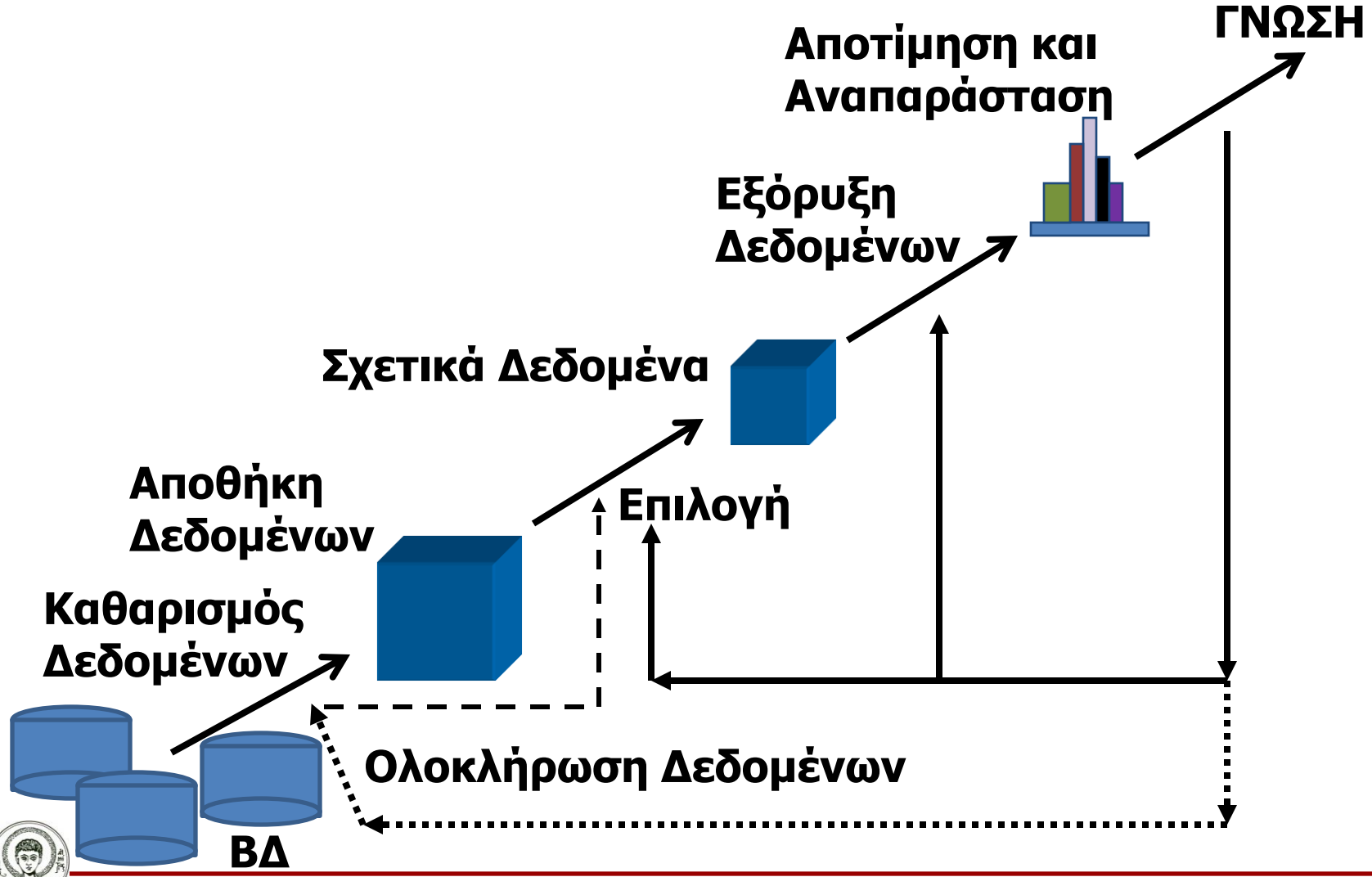
Αποθήκες Δεδομένων – OLAP (2)

- Αντίθετα, η εξόρυξη δεδομένων δουλεύει με ένα ημι-αυτόματο τρόπο.
  - Και στοχεύει στην εύρεση «κρυφής» γνώσης.
- Οι αποθήκες δεδομένων αποτελούν ένα κατάλληλο πλαίσιο για την εφαρμογή τεχνικών εξόρυξης.
  - Εκτελούν τη συλλογή, ολοκλήρωση, καθαρισμό και μετασχηματισμό των δεδομένων.

Εξόρυξη Δεδομένων και Business Intelligence



Εξ. Δεδ.: Το κέντρο της διαδικα-σίας ανακάλυψης γνώσης σε ΒΔ



Λειτουργίες Εξόρυξης Δεδομένων

- Περιγραφικές Λειτουργίες: χαρακτηρισμός των βασικών ιδιοτήτων των δεδομένων στη βάση.
- Προγνωστικές Λειτουργίες: εφαρμογή συμπερασματισμού (inference) στα υπάρχοντα δεδομένα για την παραγωγή προβλέψεων.
- Βασικές Λειτουργίες που θα ασχοληθούμε είναι η **κατηγοριοποίηση** (classification), η **ομαδοποίηση** (clustering) και η **εύρεση κανόνων συσχέτισης** (association rules).
- Γιατί χρειαζόμαστε πολλές διαφορετικές λειτουργίες;
  - Οι χρήστες συχνά δεν έχουν εκ των προτέρων καθαρή εικόνα για το ποια πληροφορία είναι ενδιαφέρουσα.
  - Ή ενδιαφέρονται για διαφορετικά τύπου πρότυπα –μοτίβα (patterns) παράλληλα.

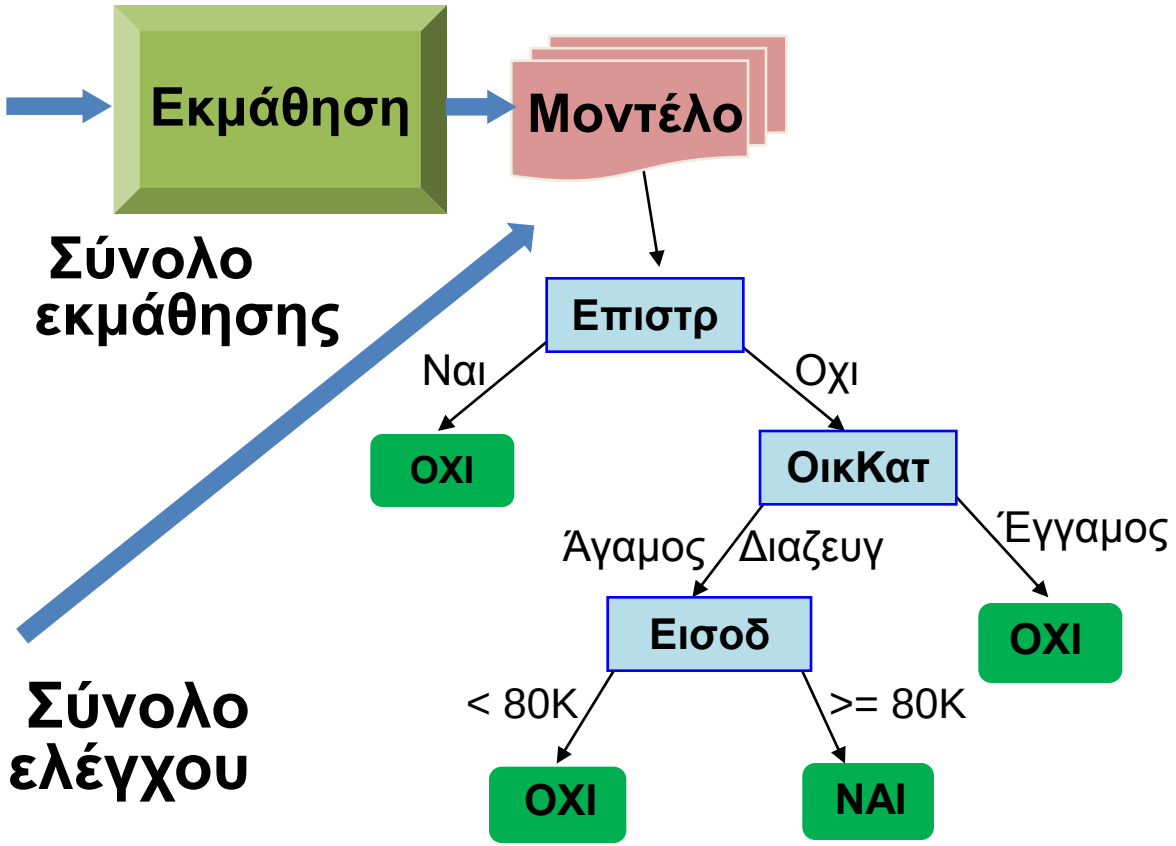
Παράδειγμα Κατηγοριοποίησης

Ιδιότητες **Κλάση**

A/A	Επιστροφή	Οικογ. κατάσταση	Εισόδημα	Απάτη
1	Ναι	Άγαμος	125K	Όχι
2	Όχι	Διαζευγμ.	95K	Ναι
3	Όχι	Άγαμος	70K	Όχι
4	Όχι	Έγγαμος	60K	Όχι
5	Όχι	Έγγαμος	100K	Όχι
6	Ναι	Έγγαμος	120K	Όχι

Επιστρ	ΟικΚατ	Εισοδ	Απάτη
Όχι	Άγαμος	75K	? (ΟΧΙ)
Όχι	Έγγαμος	150K	? (ΝΑΙ)
Ναι	Διαζευγμ.	90K	? (ΟΧΙ)
Ναι	Έγγαμος	50K	? (ΟΧΙ)

Ανάθεση τιμής στη κλάση «Απάτη»



Παραδείγματα Εφαρμογής Κατηγοριοποίησης

- Στόχευση πελατών (Marketing).
  - Δεδομένα από πελάτες (δημογραφικά, ερωτηματολόγια).
  - 2 κατηγορίες {buyer, non-buyer}.
  - Δημιουργία Μοντέλου.
  - Ταξινόμηση πελατών.
- Ταξινόμηση ουρανίων σωμάτων.
  - (αστέρας ή γαλαξίας σε κάποια φάση).
  - Εξαγωγή Δεδομένων από Εικόνα (π.χ., χροιά χρώματος hue, ιστογράμματα, κ.ο.κ.).
  - επιτυχία: με αυτή τη μέθοδο έχουν ανακαλυφθεί πολλά quasars!

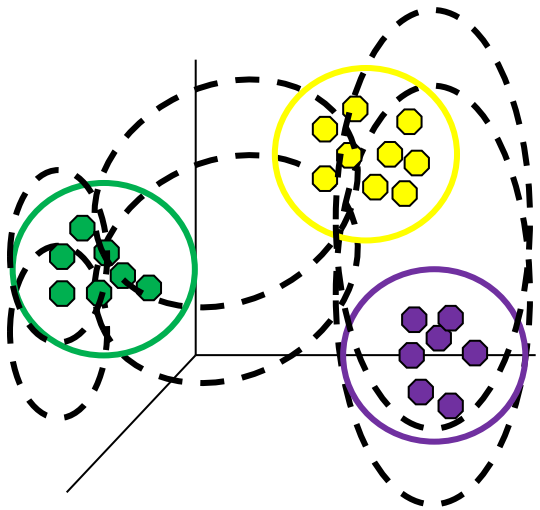
Ομαδοποίηση

- Δεν υπάρχει το χαρακτηριστικό κλάσης – κατηγορίας.
- Διαχωρισμός σε ομάδες με τον καλύτερο τρόπο.

Ελαχιστοποίηση των αποστάσεων μέσα στην ομάδα

Μεγιστοποίηση των αποστάσεων μεταξύ των ομάδων

Ομαδοποίηση στον 3-Δ Ευκλείδειο χώρο (με διακεκομμένες γραμμές)



Εφαρμογές Ομαδοποίησης

- **Κατάτμηση Αγοράς**: Διαχωρισμός πελατών σε ομάδες με διαφορετική αντιμετώπιση.
  - Δεδομένα: γεωγραφικά, δημογραφικά, κλπ.
  - Αξιολόγηση κάθε ομάδας με σύγκριση με τις υπόλοιπες.
- **Οικολογία**: Εντοπισμός φωλιών πουλιών.
  - Δεδομένα: χωρικά.
  - Κάθε ομάδα (συστάδα) αξιολογείται αναφορικά με χαρακτηριστικά όπως απόσταση από νερό, κ.ο.κ.

Κανόνες συσχέτισης - Παράδειγμα

- Είσοδος: δεδομένα συναλλαγών – αγορασμένα αντικείμενα.
- Εύρεση κανόνων που εκφράζουν τις συσχετίσεις μεταξύ της ύπαρξης αντικειμένων κατά τις συναλλαγές.

A/A	*Αντικείμενα*
1	Ψωμί, Αλεύρι, Γάλα
2	Μπύρα, Πάνες, Γάλα, Ψωμί
3	Μπύρα, Ψωμί
4	Μπύρα, Ψωμί, Πάνες, Γάλα
5	Αλεύρι, Πάνες, Γάλα

Κανόνες:

{Αλεύρι} --> {Γάλα} p(Γάλα|Αλεύρι)=1
{Γάλα} --> {Αλεύρι} p(Αλεύρι |Γάλα)=0.5

{Μπύρα, Ψωμί} --> {Πάνες}
p(Πάνες|Μπύρα, Ψωμί)=0.66

- Άλλο παράδειγμα στη Βιολογία – Δεδομένα μικροσυστοιχιών: Γονίδια που εμφανίζονται ταυτόχρονα => αλληλεπίδραση.
  - {disease} => {gene A ↑ gene B↓ gene C↑}

Επιπλέον Λειτουργίες Εξόρυξης Δεδομένων

- Ανάλυση ακραίων δεδομένων (Outlier analysis).
  - Ακραία δεδομένα - Outliers: δεδομένα που δεν ακολουθούν τη γενική κατανομή και δεν συμφωνούν με τη γενική συμπεριφορά των υπολοίπων δεδομένων.
  - Μπορούν να θεωρηθούν ως θόρυβος ή εξαιρέσεις.
  - ΑΛΛΑ είναι πολύ χρήσιμα στον εντοπισμό απατών και στην ανάλυση σπανίων φαινομένων.
- Ανάλυση τάσεων και εξέλιξης.
  - Τάσεις και αποκλίσεις: ανάλυση παλινδρόμησης (regression).
  - Ανάλυση βάσει ομοιότητας.
  - Ανάλυση περιοδικότητας, Εξόρυξη Ακολουθιακών Προτύπων.

Σε τι δεδομένα εφαρμόζεται η εξόρυξη δεδομένων

- Σχεσιακές ΒΔ - ΒΔ συναλλαγών.
- Αποθήκες δεδομένων.
- Προηγμένες ΒΔ.
  - Αντικειμενοστραφείς – αντικειμενοσχεσιακές ΒΔ.
  - Χωρικές ΒΔ.
  - Χρονοσειρές και χρονικά δεδομένα.
  - Πολυμεσικές ΒΔ και ΒΔ κειμένου.
  - Ετερογενείς και κληρονομημένες ΒΔ.
  - Παγκόσμιος Ιστός.

…από σχεσιακά ΣΔΒΔ

- Τα σχεσιακά ΣΔΒΔ παρέχουν τη δυνατότητα για SQL ερωτήματα.
  - Επιλογή χαρακτηριστικών, καθορισμός περιορισμών, κλπ.
- Δυνατά ερωτήματα: Ποια αντικείμενα πωλήθηκαν τον τελευταίο μήνα; Ή ένα βήμα πιο μακριά: εφαρμογή συναθροιστικών συναρτήσεων max, avg, κλπ. (π.χ., συνολικές πωλήσεις του έτους ανά κατάστημα).
- Η $1^{η}$ περίπτωση είναι απλή ανάκτηση δεδομένων ενώ η $2^{η}$ είναι διερευνητική ερώτηση.
- Η εξόρυξη δεδομένων σε σχεσιακά ΣΔΒΔ προχωρά ακόμη περισσότερο και εντοπίζει τάσεις, πρότυπα, κλπ.
  - Ανάλυση δεδομένων πελατών για πρόβλεψη ποιοι δεν μπορούν να ανταποκριθούν στις οικον. Υποχρεώσεις.
  - Εντοπισμός ισχυρών αποκλίσεων στα δεδομένα από μήνα σε μήνα.

…από Αποθήκες Δεδομένων

- Έστω ότι σε μία επιχείρηση θέλουμε ανάλυση οινομικών μεγεθών ανά κομμάτι – υποκατάστημα –μήνα.
- Δύσκολο εγχείρημα.
  - Κάθε υποκατάστημα έχει τη δική του ΒΔ πιθανώς και με διαφορετικό σχήμα.
- Η ΑΔ ξεπερνά αυτά τα προβλήματα.
- Η εξόρυξη χρησιμοποιείται συμπληρωματικά με τις τεχνικές OLAP.
  - OLAP: εργαλεία για συνοπτική πληροφορία (summarization), roll-up, drill-down.
  - Η εξόρυξη πραγματοποιεί πιο αυτοματοποιημένες αναλύσεις.
  - Αλλά και οι 2 είναι χρήσιμες.

…από ΒΔ συναλλαγών

- Κάθε εγγραφή αντιστοιχεί σε μία συναλλαγή.
  - TID: CID: item1, item2, …
  - όχι 1NF
- Τυπικά ερωτήματα:
  - «Βρες τι αγόρασε ο Χ».
  - « Σε πόσες συναλλαγές αγοράστηκε το προϊόν Υ».
  - Απλή ανάκτηση.
- Με την εξόρυξη δεδομένων μπορούμε να βρούμε
  - Ποια προϊόντα πωλούνται μαζί.

…από προηγμένα ΣΔΒΔ (1)

- Object oriented/Object relational ΒΔ:
  - Η εξόρυξη παρέχει εργαλεία για την ανάλυση περίπλοκων δομών και ιεραρχιών.
- Χωρικές ΒΔ:
  - Πρότυπα για χαρακτηριστικά σπιτιών κοντά σε συγκεκριμένες περιοχές.
  - Εισοδήματα ως συνάρτηση της απόστασης της κατοικίας από ΕΟΔ.
- Χρονοσειρές:
  - Πρότυπα εξέλιξης και αλλαγή τάσεων.

…από προηγμένα ΣΔΒΔ (2)

- Πολυμεσικές ΒΔ και ΒΔ κειμένου:
  - Συσχετισμός των λέξεων κλειδιών, ομαδοποίηση κειμένου (συνδυασμός με IR).
  - Εξαγωγή στοιχείων από πολυμεσικά δεδομένα, συνδυασμοί βάσει ομοιότητας.
- Ετερογενείς και κληρονομημένες ΒΔ:
  - Τα κληρονομημένα συστήματα έχουν ετερογενή δεδομένα σε πολλές ΒΔ.
  - Οι τεχνικές αποθήκευσης και εξόρυξης παρέχουν λύσεις στην ανταλλαγή πληροφορίας παράγοντας υψηλότερου επιπέδου, πιο γενικευμένη πληροφορία.

…από τον Παγκόσμιο Ιστό

- WWW:
  - Προσέλκυσε πολύ ενδιαφέρον =>
    - Ξεχωριστό πεδίο: Web Mining.
  - Εξόρυξη περιεχομένου (συναφές με την εξόρυξη από κείμενο).
  - Εξόρυξη συνδέσεων (εντοπισμός δομών).
  - Εξόρυξη χρήσης (εύρεση προτύπων επισκέψεων).

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Τέλος ενότητας

Επεξεργασία: Ανδρέας Κοσματόπουλος

Θεσσαλονίκη, Χειμερινό Εξάμηνο 2013-2014

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημειώματα

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.