

Πανεπιστήμιο Πειραιώς - Τμήμα Πληροφορικής



Αποθήκες Δεδομένων και Εξόρυξη Γνώσης (Data Warehousing & Data Mining)

Εισαγωγή

Γιάννης Θεοδορίδης, Νίκος Πελέκης

Ομάδα Διαχείρισης Δεδομένων
Εργαστήριο Πληροφοριακών Συστημάτων

<http://isl.cs.unipi.gr/db>

“πλημμύρα” από δεδομένα



- Παράγονται όλο και περισσότερα δεδομένα:
 - Τραπεζικά, τηλεπικοινωνιακά, ...
 - Επιστημονικά δεδομένα: αστρονομικά, βιολογικά κλπ.
 - Κείμενα στο web κ.α.
- Αποθηκεύονται όλο και περισσότερα δεδομένα:
 - Γρήγορη και φθηνή τεχνολογία αποθήκευσης
 - Ικανά ΣΔΒΔ για μεγάλες ΒΔ



Παραδείγματα



- Το ευρωπαϊκό Very Long Baseline Interferometry (VLBI) διαθέτει 16 τηλεσκόπια, καθένα από τα οποία παράγει **1 Gigabit/second** αστρονομικά δεδομένα σε συνόδους παρατήρησης των 25 ημερών
 - η αποθήκευση και ανάλυση τέτοιου όγκου δεδομένων είναι πρόβλημα ☹
- Ο τηλεπικοινωνιακός κολοσός AT&T χειρίζεται δισεκατομμύρια κλήσεις / μέρα
 - τόσο μεγάλος είναι ο όγκος των δεδομένων που αυτά **δεν** αποθηκεύονται – η ανάλυση γίνεται «στον αέρα» (on the fly)
- Η ΒΔ της επιχείρησης λιανεμπορίου Wal-Mart είναι της τάξης των 24 Tbytes
- Το UC Berkeley έκανε την εκτίμηση ότι μέσα στο 2002 παρήχθησαν 5 Exa-bytes (5 εκατομμύρια TBytes) δεδομένων !!!

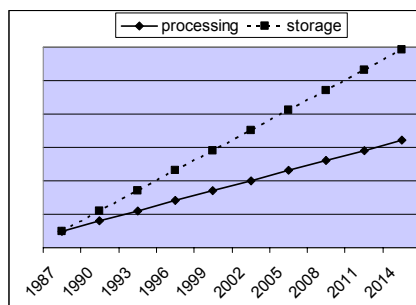
3

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Τάσεις ανάπτυξης



- Ο νόμος του Moore
 - Η ταχύτητα των υπολογιστών διπλασιάζεται κάθε 18 μήνες
- Ο νόμος της αποθήκευσης
 - Τα δεδομένα που αποθηκεύονται διπλασιάζονται κάθε 9 μήνες
- Κατά συνέπεια ...
 - πολύ λίγα από αυτά τα δεδομένα μπορεί να κοιτάξει (και να αναλύσει) ο άνθρωπος
- **Άρα** χρειάζεται η ανακάλυψη γνώσης μέσα από τα δεδομένα (Knowledge Discovery in Data - KDD) για να δώσει νόημα και χρήση στα δεδομένα



4

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Ορισμός της διαδικασίας KDD



Η ανακάλυψη γνώσης από δεδομένα (Knowledge Discovery in Data – KDD) είναι η μη τετριμμένη διαδικασία εύρεσης έγκυρων, πρωτότυπων, πιθανώς χρήσιμων και οπωσδήποτε κατανοητών προτύπων (patterns) μέσα στα δεδομένα.

▪ Τι **δεν είναι** data mining

- επεξεργασία ερωτήσεων βάσεων δεδομένων
- χρήση προγραμμάτων μηχανικής μάθησης ή στατιστικής

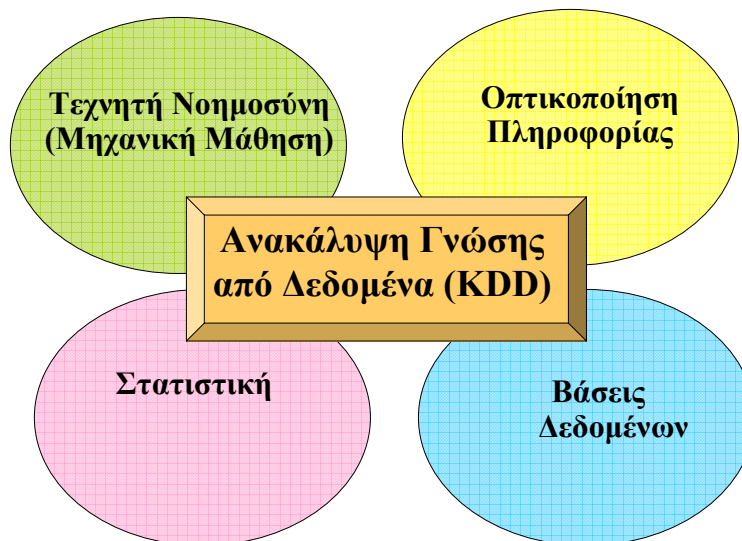
Εναλλακτικές μεταφράσεις του όρου data mining:

- «εξόρυξη δεδομένων»
- «εξόρυξη γνώσης από δεδομένα»
- «όρυξη δεδομένων»

5

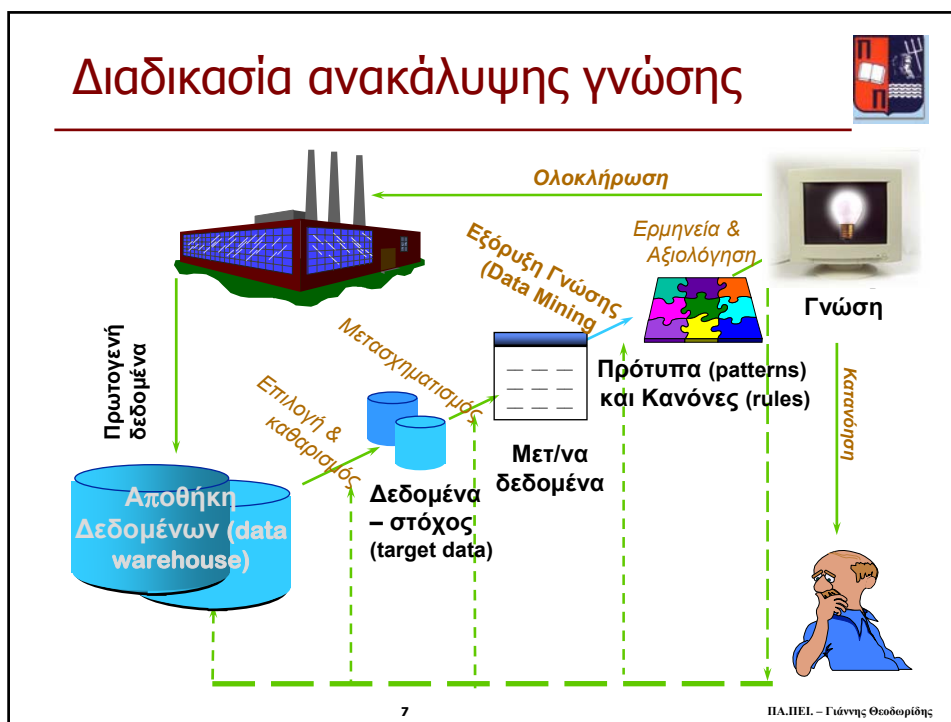
ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Σχετικά επιστημονικά πεδία



6

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης



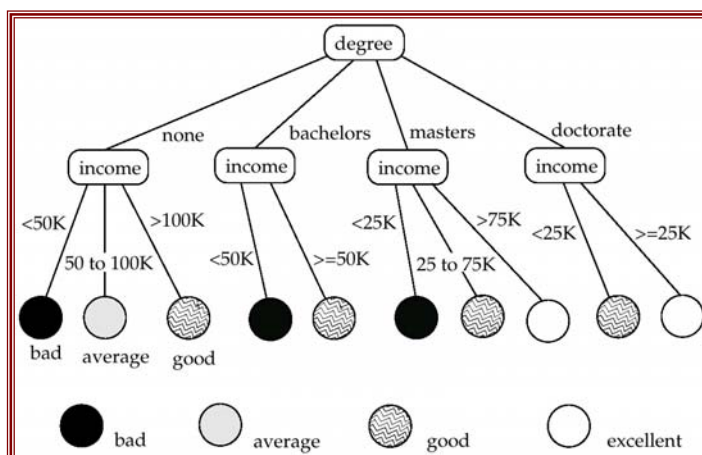
Εξόρυξη γνώσης από δεδομένα

- Data mining (ένας ευρύς ορισμός): Η **διαδικασία ημι-αυτόματης ανάλυσης μεγάλων ΒΔ με στόχο την εύρεση χρήσιμης πληροφορίας – «γνώσης»**
 - π.χ.: «καλύτεροι πελάτες είναι αυτοί με σπουδές μεταπτυχιακού επιπέδου (ανεξαρτήτως ύψους εισοδήματος) ή αυτοί με υψηλά εισοδήματα (ανεξαρτήτως επιπέδου σπουδών)
- Παράδειγμα αναπαράστασης: δέντρο απόφασης (decision tree)
- Πώς θα προκύψει ένα τέτοιο πρότυπο (pattern) από μια μεγάλη ΒΔ;

8

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Παράδειγμα Decision Tree



9

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Εφαρμογές Data Mining

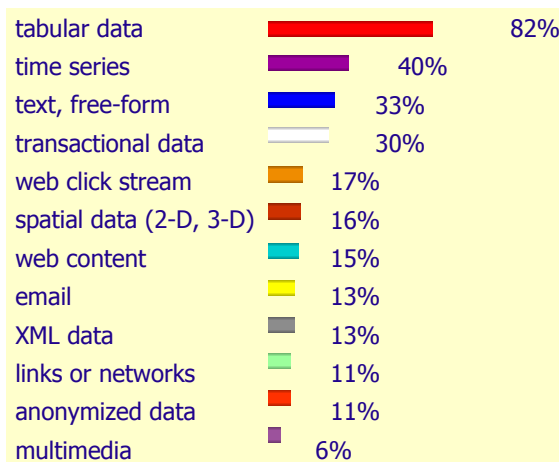


- **Market analysis and management**
 - target marketing, customer relation management, market basket analysis, cross selling, market segmentation
- **Risk analysis and management**
 - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
- **Fraud detection and management**
- **αλλά και ...**
 - Intelligent query answering
 - Text / Web mining (news group, email, documents)

10

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Τι δεδομένα αναλύουμε ...



Source: www.kdnuggets.com (Sep. 2005)

11

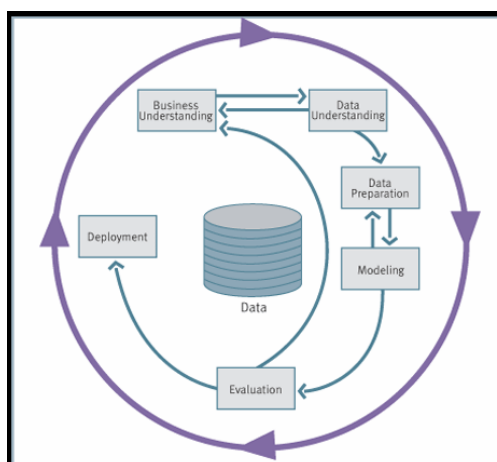
ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Βήματα εξόρυξης γνώσης



■ μεθοδολογία CRISP-DM

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment

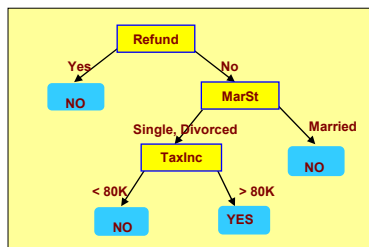


12

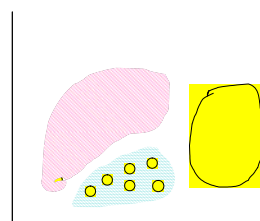
ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Τεχνικές DM

- Supervised learning (Classification / Prediction)
 - Decision Trees, Neural Nets, Bayesian classification
- Unsupervised learning (Clustering)
 - Hierarchical, partitional, density-based
- Association rule mining
- Κ.Α.



$X \Rightarrow Y$	s	α
Bread $\Rightarrow$ PeanutButter	60%	75%
PeanutButter $\Rightarrow$ Bread	60%	100%
Beer $\Rightarrow$ Bread	20%	50%
PeanutButter $\Rightarrow$ Jelly	20%	33.3%
Jelly $\Rightarrow$ PeanutButter	20%	100%
Jelly $\Rightarrow$ Milk	0%	0%



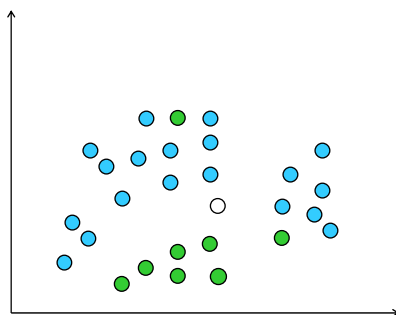
13

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Κατηγοριοποίηση (Classification)

Ταξινόμηση, Πρόβλεψη

Εκμάθηση μιας τεχνικής να προβλέπει την κλάση ενός στοιχείου επιλέγοντας από προκαθορισμένες τιμές



Προσεγγίσεις:

- στατιστικές μέθοδοι
- δένδρα αποφάσεων
- νευρωνικά δίκτυα
- ...

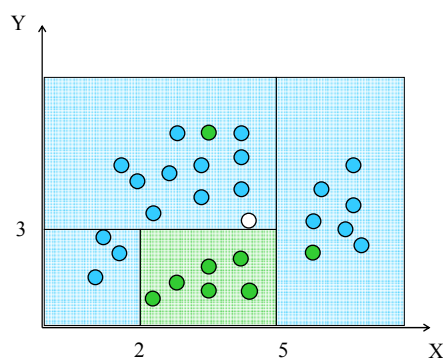
14

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Κατηγοριοποίηση (Classification)



Δέντρα Αποφάσεων (decision trees)



if $X > 5$ then **blue**
 else if $Y > 3$ then **blue**
 else if $X > 2$ then **green**
 else **blue**

15

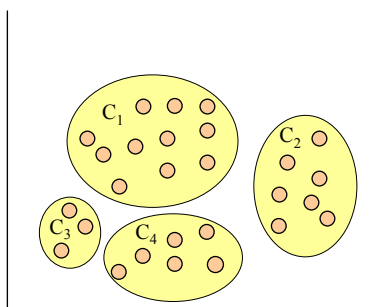
ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Συσταδοποίηση (Clustering)



Ομαδοποίηση

Εύρεση μιας “φυσικής” ομαδοποίησης των δεδομένων, χωρίς προκαθορισμό των ομάδων



Ομαδοποίηση βάσει
 • απόστασης (Εκλείδειας ή άλλης)
 • στατιστικής κατανομής
 • ...

16

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Εξόρυξη κανόνων συσχετίσεων



Εύρεση ταυτόχρονων εμφανίσεων δεδομένων
(άρα, πιθανής συσχέτισης ή εξάρτησης) μέσα σε
ένα «καλάθι» δεδομένων

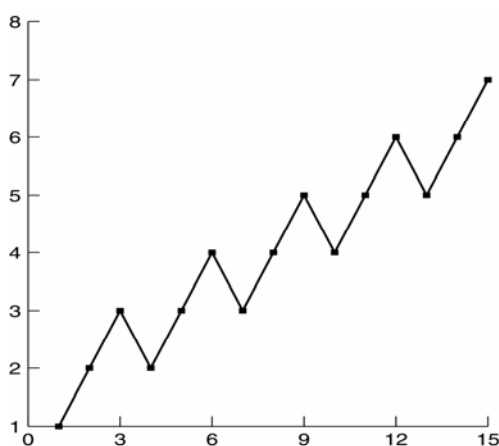
Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

$X \Rightarrow Y$	s	α
Bread $\Rightarrow$ PeanutButter	60%	75%
PeanutButter $\Rightarrow$ Bread	60%	100%
Beer $\Rightarrow$ Bread	20%	50%
PeanutButter $\Rightarrow$ Jelly	20%	33.3%
Jelly $\Rightarrow$ PeanutButter	20%	100%
Jelly $\Rightarrow$ Milk	0%	0%

17

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Ανάλυση χρονολογικών σειρών

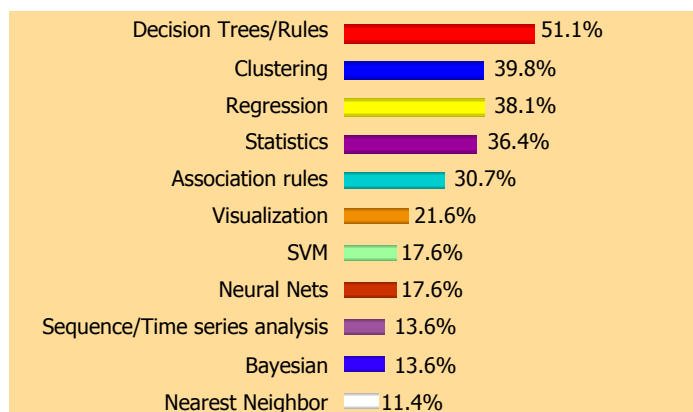


Μπορούμε να βρούμε τάσεις (κύκλους, εποχιακές συμπεριφορές) και
να κάνουμε εκτίμηση μελλοντικών τιμών

18

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Οι πιο δημοφιλείς τεχνικές DM ...



Source: www.kdnuggets.com (Apr. 2006)

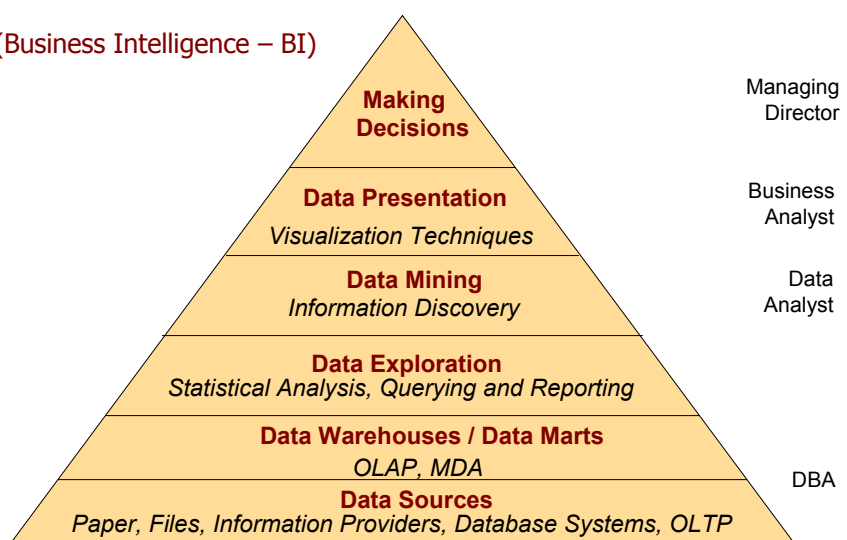
19

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

"Πυραμίδα" Επιχειρηματικής Ευφυίας



(Business Intelligence – BI)



20

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Case Study -1



- **απευθείας προώθηση προϊόντων & CRM**
- Οι εταιρείες που κάνουν απευθείας προώθηση προϊόντων (direct marketing) χρησιμοποιούν μοντέλα και τεχνικές εξόρυξης γνώσης
 - Η μοντελοποίηση είναι πιο εφικτή από την αλλαγή συμπεριφοράς του πελάτη
- Πετυχημένη ιστορία (success story)
 - Ο τηλεπικοινωνιακός παροχέας Verizon Wireless μείωσε τη «διαρροή πελατών» (churn) από 2% σε 1.5%

21

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Case Study -2



- **Ασφάλεια και ανίχνευση απάτης**
- Ανίχνευση απάτης (Fraud Detection) σε αγορές με πιστωτικές κάρτες
- Ξέπλυμα χρήματος
- Απάτη στην τηλεφωνία
 - AT&T, Bell Atlantic, British Telecom/MCI
- Καταπολέμηση τρομοκρατίας ...



22

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Data Mining και ιδιωτικό απόρρητο

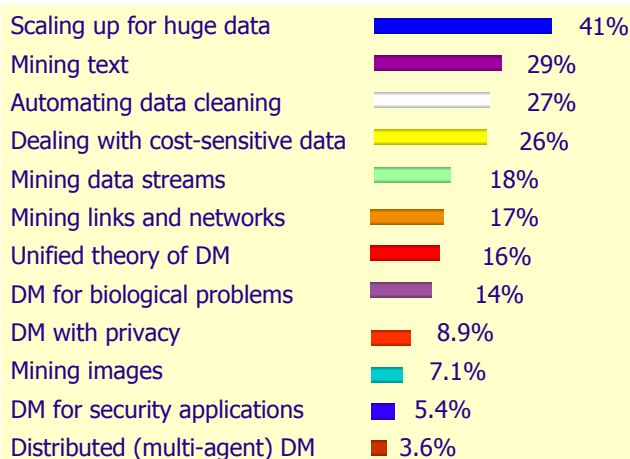


- TIA: Terrorism (πρώην Total) Information Awareness Program –
 - Επιστημονικό πρόγραμμα χρηματοδοτούμενο από το Υπουργείο Άμυνας των ΗΠΑ – διακόπηκε από το Κογκρέσο
 - μερικές από τις λειτουργίες του μεταφέρθηκαν στις υπηρεσίες πληροφοριών
- CAPPs II – φωτογράφιση όλων των επιβατών των αερομεταφορών
 - έχει προκαλέσει αντιδράσεις (από επιβάτες, εταιρείες, κυβερνήσεις)
- Οι τεχνικές Data Mining 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)

23

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Ερευνητικά προβλήματα ...

Source: www.kdnuggets.com (Nov. 2005)

24

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης

Στοιχεία μαθήματος



- **Διδάσκοντες**
 - Γιάννης Θεοδορίδης (κεντρικό κτήριο, γρ. 501)
 - Νίκος Πελέκης (κτήριο Ζέας, εργαστήριο 3^{ου} ορόφου)
- **Βοηθοί**
 - Βαγγέλης Κοτσιφάκος, Γεράσιμος Μαρκέτος, Ειρήνη Ντούτση
- **Διαδακτική ύλη μαθήματος**
 - Βιβλίο «Data Mining: Εισαγωγικά και Προηγμένα Θέματα Εξόρυξης Γνώσης από Δεδομένα» της Μ.Η. Dunham (<http://isl.cs.unipi.gr/dmbook>)
 - Διαφάνειες διδασκοντα (<http://isl.cs.unipi.gr/db/courses/dwdm>)

Εργαλεία KDD



- **Microsoft SQL Server 2005 BI Solutions**
 - *"SQL Server 2005 provides many new and enhanced business intelligence (BI) features designed to give you a competitive advantage. These advantages include integrating multiple data sources with Integration Services; enriching data and building complex business analytics with Analysis Services; and writing, managing, and delivering rich reports with Reporting Services ..."*
 - <http://www.microsoft.com/sql/solutions/bi>
- **WEKA Data Mining software**
 - *"Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization ..."*
 - <http://www.cs.waikato.ac.nz/ml/weka>

Εργασίες



- **1^η εργασία:** ανάπτυξη ενός DW και εφαρμογή τεχνικών DM πάνω σε πραγματικά δεδομένα
 - σε ομάδες των 2 ατόμων, υποχρεωτική
 - συμμετοχή 40% στον τελικό βαθμό
- **2^η εργασία:** σχεδίαση και υλοποίηση εφαρμογής BI
 - ατομική, προαιρετική, απαλλακτική της γραπτής εξέτασης
 - συμμετοχή 60% στον τελικό βαθμό
- **(και πιθανά κάποιες) βιβλιογραφικές εργασίες:** μελέτη ερευνητικών άρθρων και συγγραφή αναφοράς
 - ατομικές, υποχρεωτικές
 - bonus στον τελικό βαθμό

27

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης



Καλό εξάμηνο !

28

ΠΑ.ΠΕΙ. – Γιάννης Θεοδορίδης