

ΒΑΣΕΙΣ ΚΑΙ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Databases and Data Mining

Δρ. Τιάκας Ελευθέριος



Εργαστήριο Τεχνολογίας και Επεξεργασίας Δεδομένων
Τμήμα Πληροφορικής
Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

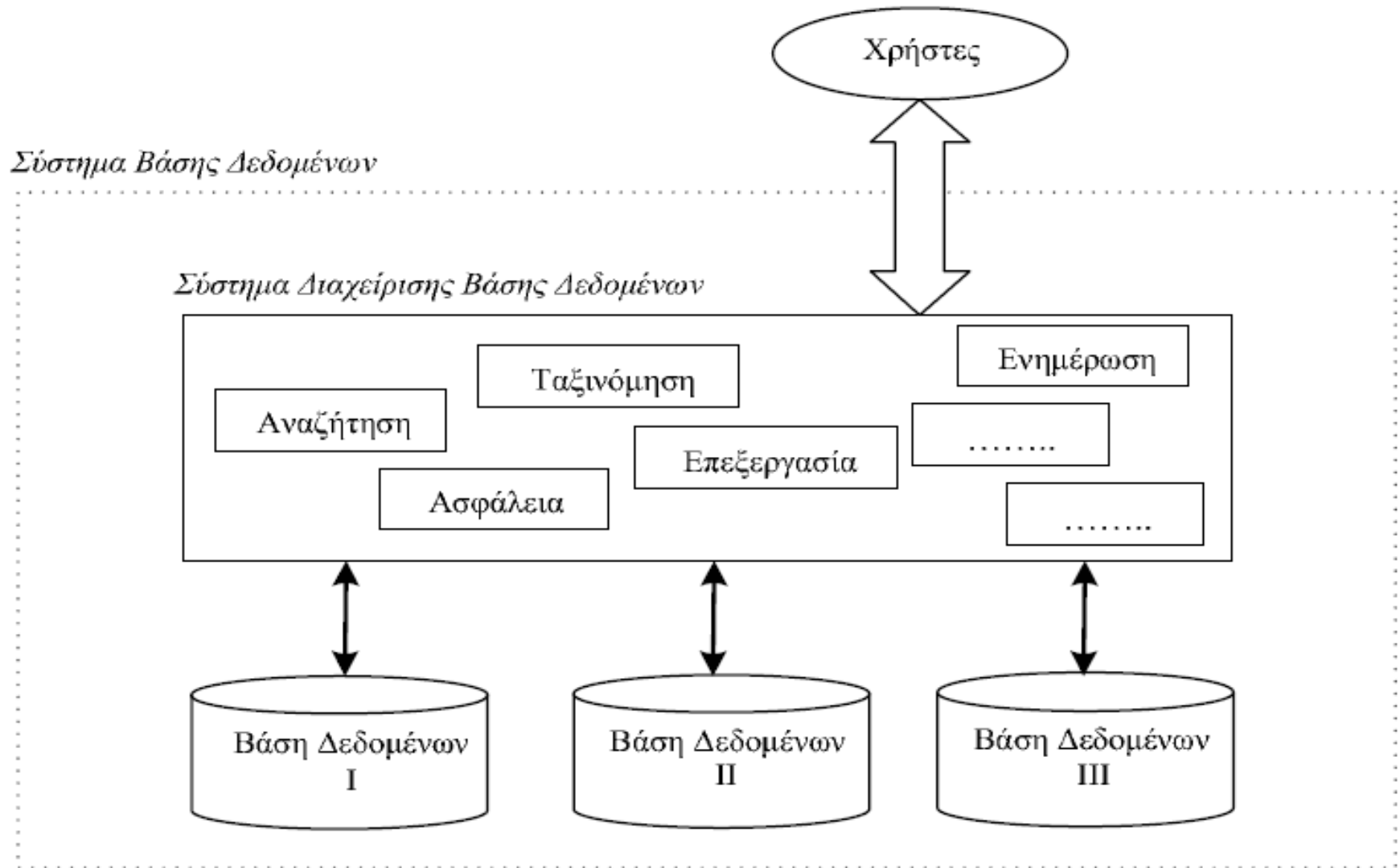
<http://delab.csd.auth.gr/~tiakas>

Ορισμός Βάσης Δεδομένων

■ Σύστημα Διαχείρισης Βάσεων Δεδομένων (database management system):

Είναι μία συλλογή πακέτων λογισμικού, τα οποία υλοποιούν όλες τις λειτουργίες που πρέπει να υποστηριχθούν, όπως αναζήτηση, εισαγωγή, διαγραφή, συγχρονισμός προσπελάσεων, προστασία και πολλές ακόμη.

Διαγραμματική απεικόνιση μιας Βάσης Δεδομένων



Πλεονεκτήματα των Βάσεων Δεδομένων

- Περιγραφή Δεδομένων (**meta-data** = δεδομένα για τα δεδομένα, κατάλογος συστήματος).
- Ανεξαρτησία Δεδομένων και Λειτουργιών (**δεν αλλάζουν τα προγράμματα εφαρμογής** [το λογισμικό] όταν αλλάζουν τα δεδομένα).
- Αποδοτικότερη Διαχείριση Δεδομένων (ο προγραμματιστής επικεντρώνεται **στην εφαρμογή και όχι στα δεδομένα και στους αλγορίθμους** επεξεργασίας τους).

Πλεονεκτήματα των Βάσεων Δεδομένων

- Προστασία Δεδομένων και Δικαιώματα Χρηστών (για κάθε σύνολο δεδομένων **ξεχωριστά**).
- Μηχανισμοί Ταυτόχρονης Προσπέλασης (προσφέρονται από το σύστημα μηχανισμοί **προσωρινού κλειδώματος** κάθε φορά που γίνεται κάποια ενημέρωση από ένα τερματικό [**μόνο για τα δεδομένα που ενημερώνονται και μόνο για το χρονικό διάστημα που ενημερώνονται**]).

Σύγχρονες Εφαρμογές

- Πολυμεσικές Εφαρμογές (**multimedia databases**)
- Γεωγραφικά Συστήματα Πληροφοριών (**GIS**)
- Άμεση (on-line) αναλυτική επεξεργασία (**OLAP**)
- Εξόρυξη Δεδομένων (**data mining**)
- Ανάκτηση Πληροφορίας (**information retrieval**)
- Διαχείριση και Εφαρμογές Κινούμενων Αντικειμένων (**moving objects applications**)
- Εφαρμογές Φορητών Υπολογιστών (**mobile computing applications**)

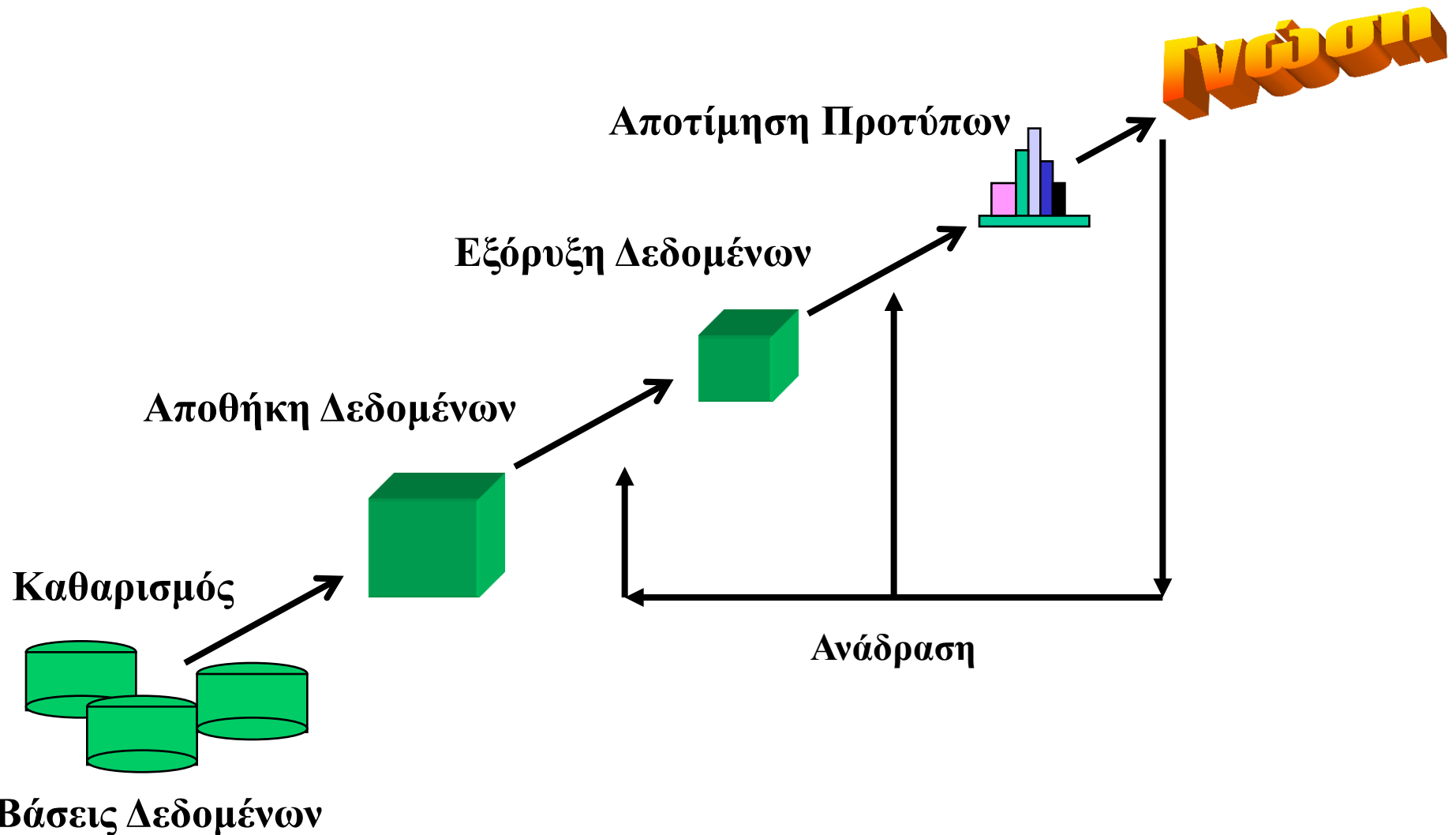
Εξόρυξη Δεδομένων - Ορισμός

- Εξόρυξη δεδομένων είναι η εξαγωγή πληροφορίας από *μεγάλες βάσεις δεδομένων*:
 - που παρουσιάζει ενδιαφέρον
 - που είναι νέα (μη γνωστής εκ των προτέρων)
 - που δεν είναι προφανής
 - που είναι χρήσιμη (αξιοποιήσιμη)

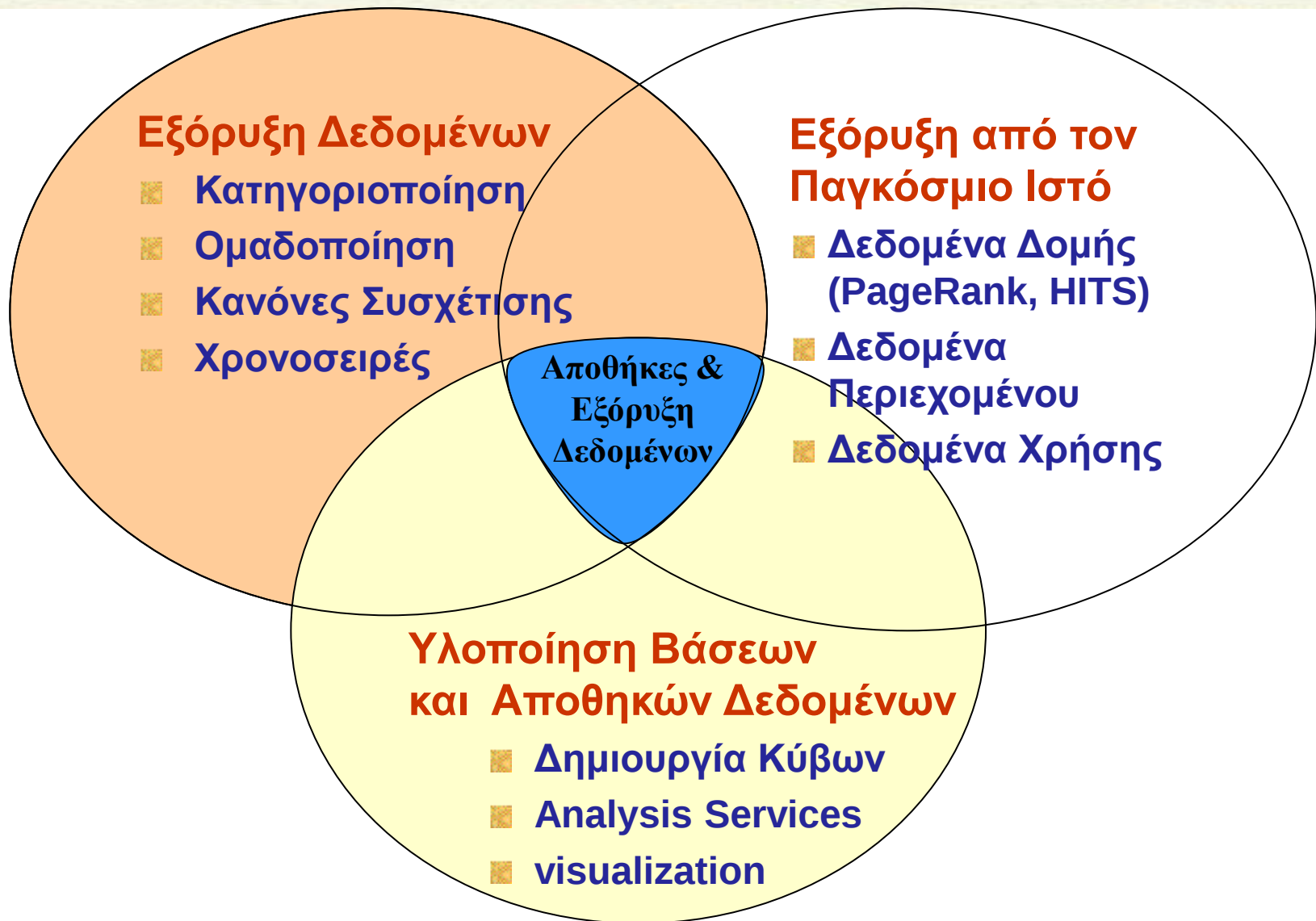
Κίνητρα για Εξόρυξη Δεδομένων

- Υπάρχουν παντού δεδομένα:
 - Σχεσιακές Βάσεις
 - Αποθήκες Δεδομένων
 - Δεδομένα Συναλλαγών (**ATM, υπερκαταστήματα**)
 - Χωρικά Δεδομένα (**GIS, δεδομένα δορυφόρων**)
 - Δεδομένα Χρονοσειρών (**δεδομένα σεισμών κλπ.**)
 - Πολυμεσικά Δεδομένα (**ήχοι, εικόνες, βίντεο**)
 - Δεδομένα Παγκοσμίου Ιστού (**logs, ιστοσελίδες**)
- Η εξόρυξη δίνει το πλεονέκτημα να κρατάμε μόνο δεδομένα χρήσιμα αντί να κρατάμε όλα τα δεδομένα που είναι πρόβλημα

Η Εξόρυξη Δεδομένων ως διαδικασία



Αποθήκες & Εξόρυξη Δεδομένων



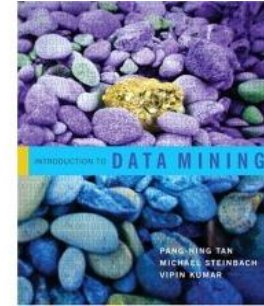
Στόχοι του Μαθήματος

Τρεις διδακτικοί στόχοι:

1. Γνώση προχωρημένων θεμάτων Βάσεων και Εξόρυξης Δεδομένων.
2. Ικανότητα για κριτική αξιολόγηση ερευνητικών εργασιών στην Εξόρυξη Δεδομένων.
3. Εφαρμογή αλγορίθμων Εξόρυξης Δεδομένων για την επίλυση προβλημάτων.

Βιβλιογραφία

[1] Introduction to Data Mining (Tan, Steinbach, Kumar)



[2] Mining of Massive Datasets (Rajaraman, Leskovec, Ullman)

Mining of Massive Datasets

Anand Rajaraman
Jure Leskovec
Jeffrey D. Ullman

[3] Εισαγωγή στην εξόρυξη και τις αποθήκες δεδομένων (Νανόπουλος, Μανωλόπουλος)



Βιβλιογραφία

[4] Βάσεις, Αποθήκες και Εξόρυξη Δεδομένων (Συμεωνίδης)



Βαθμολογία Μαθήματος

■ Παρουσίαση Ερευνητικής Εργασίας

(2 μονάδες)

■ Συνθετική Εργασία

(3 μονάδες)

■ Γραπτές Εξετάσεις

(5 μονάδες)

*Οι μονάδες αθροίζονται χωρίς προϋποθέσεις

Παρουσίαση Ερευνητικής Εργασίας

■ Δυνατότητες επιλογής θέματος

1. State-of-the-art papers from Conferences or Workshops
2. Papers from Data Engineering Lab (DELAB)
3. Data Mining Book project topics

■ Τρόπος Παρουσίασης

- Διάλεξη 15 λεπτών
- Προβολή διαφανειών σε Power Point (το πολύ 30 διαφάνειες)
- Ερωτήσεις – συζήτηση (5 λεπτά)

Συνθετική Εργασία

■ Χρήση του **MS SQL Server 2014 R2 Business Intelligence**

■ Σενάριο

- Μία επιχείρηση διαθέτει μία βάση δεδομένων με στοιχεία των πελατών της.
- Θα σας δοθούν τρία προβλήματα που αντιμετωπίζει η επιχείρηση και θα προτείνετε λύσεις.

■ Εφαρμογή αλγορίθμων Εξόρυξης δεδομένων

- Αλγόριθμοι Δέντρων Απόφασης, Ομαδοποίησης σε Συστάδες, Χρονοσειρών και Κανόνων Συσχέτισης

■ Παραδοτέα

- Προσδιορισμός πιθανών μοντέλων
 - Διαγράμματα κύβων, δέντρων, χρονοσειρών, συστάδων και κανόνων συσχέτισης
- Αξιολόγηση των πιθανών μοντέλων
- Επιλογή καλύτερου μοντέλου

Εξετάσεις

■ Ύλη μαθήματος

(5-6 κεφάλαια από τα βιβλία που προτείνονται και από το υλικό των διαφανειών)

■ Θα πρέπει να απαντηθούν πέντε θέματα

Οργάνωση του Μαθήματος

- Περιγραφή-Επισκόπηση του Μαθήματος, Βασικές έννοιες
- Βάσεις Δεδομένων: Εννοιολογικό Επίπεδο, ER-model
- Βάσεις Δεδομένων: Εκτεταμένο και Σχεσιακό Μοντέλο
- Ερωτήματα ΒΔ, Η γλώσσα SQL, Αποθήκες Δεδομένων
- Κατηγοριοποίηση: Θεωρία και Μέθοδοι
- Ομαδοποίηση: Θεωρία και Μέθοδοι
- Κανόνες Συσχέτισης: Θεωρία και Μέθοδοι
- Χρονοσειρές Θεωρία και προετοιμασία SQL Server 2014
- SQL Server 2014 ΒΔ-Ερωτήματα-Διαγράμματα-Κύβοι
- SQL Server 2014 (Κατηγοριοποίηση-Δέντρα Απόφασης)
- SQL Server 2014 (Ομαδοποίηση)
- SQL Server 2014 (Κανόνες Συσχέτισης)
- Εξόρυξη Δεδομένων Παγκόσμιου Ιστού & Ειδικού Σκοπού

Ενότητες στις Βάσεις Δεδομένων

- Εκτεταμένο μοντέλο οντοτήτων συσχετίσεων (ER-Model)
- Δημιουργία Βάσης Δεδομένων
- Ερωτήματα σε Βάση Δεδομένων

Μέθοδοι Ομαδοποίησης

- K-means
- Agglomerative Hierarchical Clustering
- Density-based algorithms (**DBSCAN**)
- Graph-based algorithms (**Two-way nCut**)
- Scalable Clustering algorithms
- Cluster Evaluation

Μέθοδοι Κατηγοριοποίησης

- **Decision Tree Classifier**
- **Model Over-fitting**
- **Naïve Bayes Classifier**
- **Nearest Neighbor Classifier**
- **Evaluating and Comparing Classifiers**
- **Ensemble Methods**

Μέθοδοι Εύρεσης Κανόνων Συσχέτισης

- Frequent Item-set Generation (**A priori**)
- Alternative Item-set Generation (**FP-Growth**)
- Sequential Patterns (**temporal information**)
- Recommendations based on Association Rules (**cross-sales**)
- Evaluation of Association Patterns

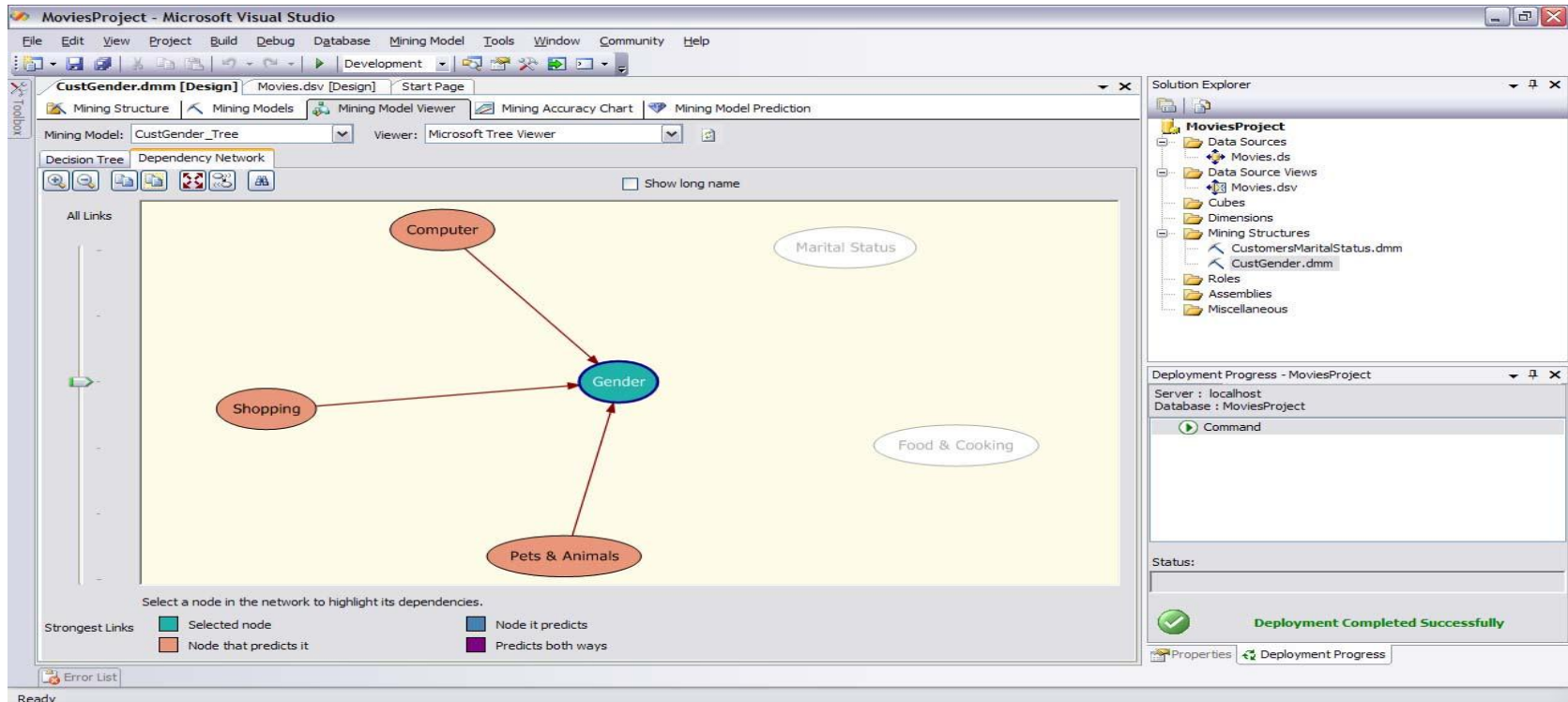
MS SQL Server 2014 (Δημιουργία Κύβου)

- Δημιουργία Κύβου Δεδομένων
- Προβολή του Κύβου Δεδομένων
- Αξιολόγηση του Κύβου Δεδομένων

Drop Filter Fields Here					
		The Year ▼ Quarter The Month			
		[-] 1997			
		[-] Q1			
		[+] February	[+] January	[+] March	Total
Store Country ▼	Store State	Store Sales	Store Sales	Store Sales	Store Sales
[-] USA	[+] CA	12654.65	11737.29	11783.26	36175.2
	[+] OR	9675.340000000001	14302.7	16192.25	40170.28999999999
	[+] WA	21728.8	19499.7	22054.36	63282.85999999998
	Total	44058.78999999999	45539.69	50029.86999999999	139628.35
Grand Total		44058.78999999999	45539.69	50029.86999999999	139628.35000000001

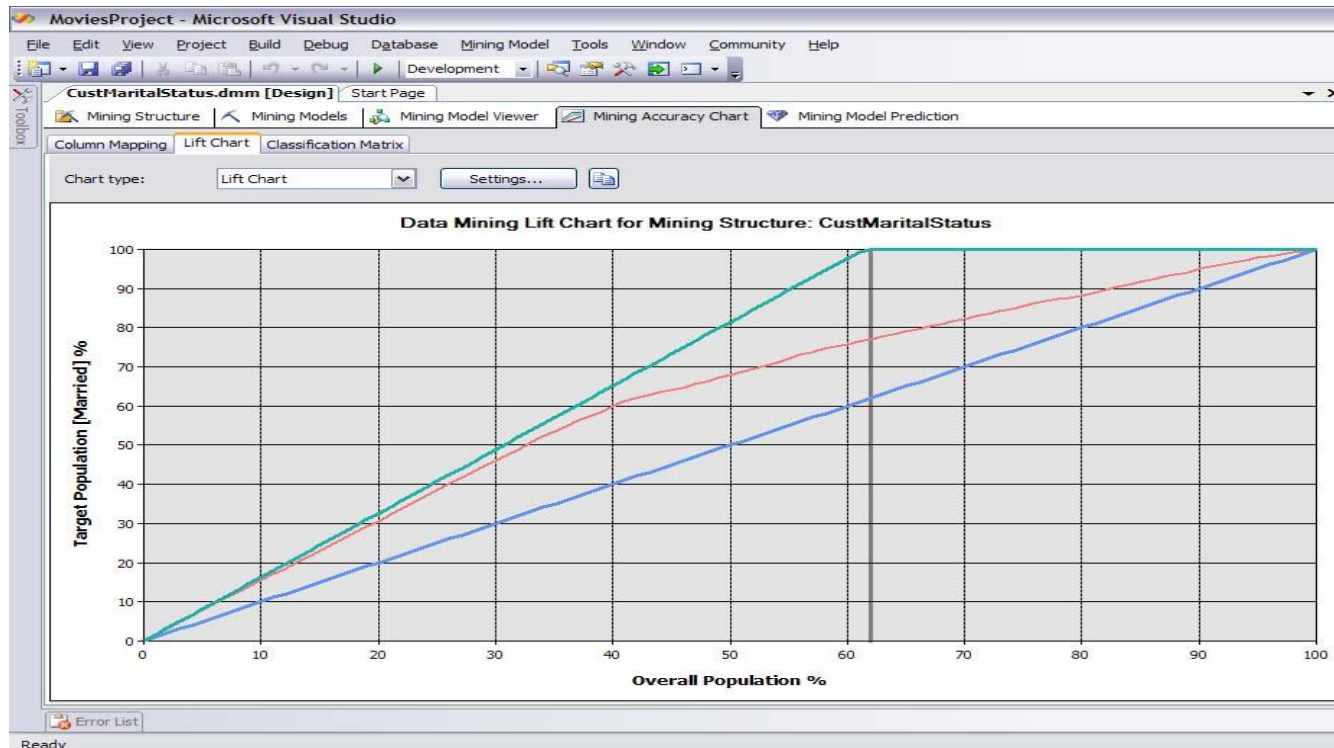
MS SQL Server 2014 (Κατηγοριοποίηση)

- Εφαρμογή Αλγορίθμου Δέντρου Απόφασης
- Παράμετροι Αλγορίθμου
- Προβολή του Δέντρου Απόφασης
- Αξιολόγηση του Δέντρου Απόφασης



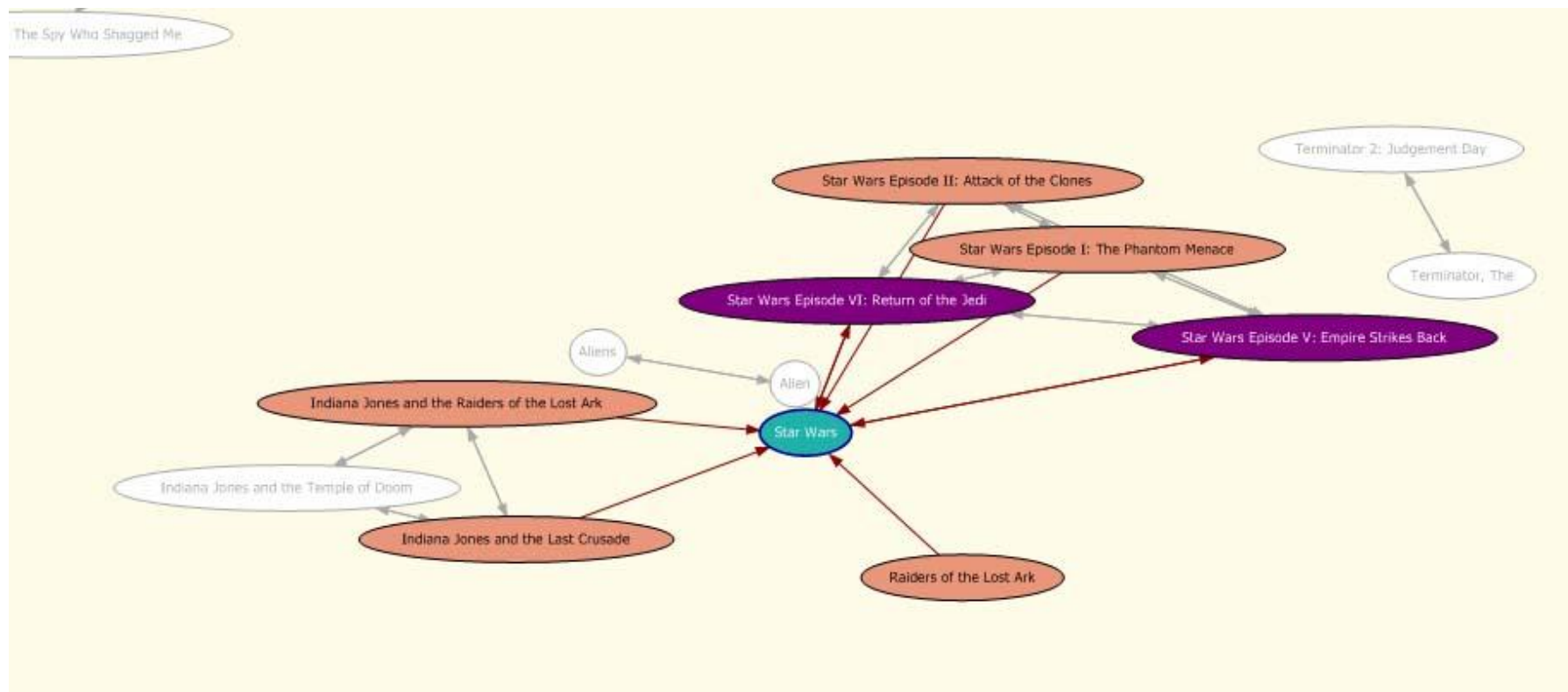
MS SQL Server 2014 (Ομαδοποίηση)

- Εφαρμογή Αλγορίθμου Ομαδοποίησης
- Παράμετροι Αλγορίθμου
- Προβολή Συστάδων
- Αξιολόγηση των Συστάδων



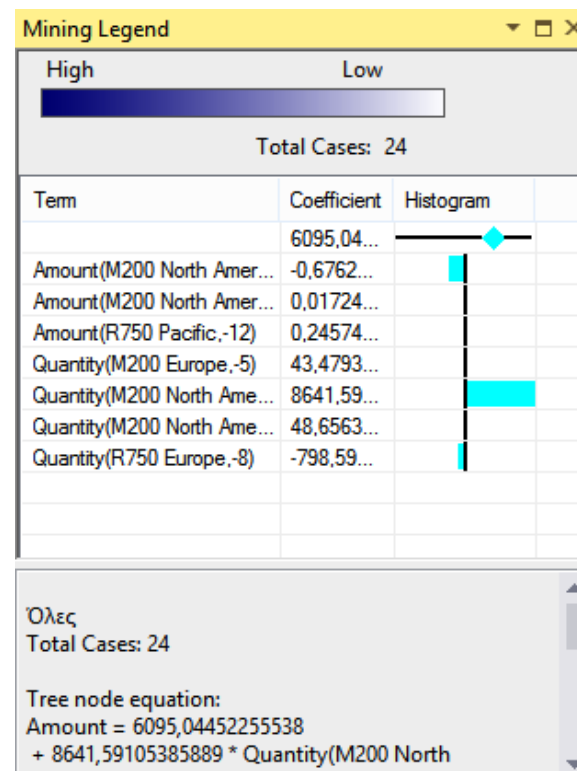
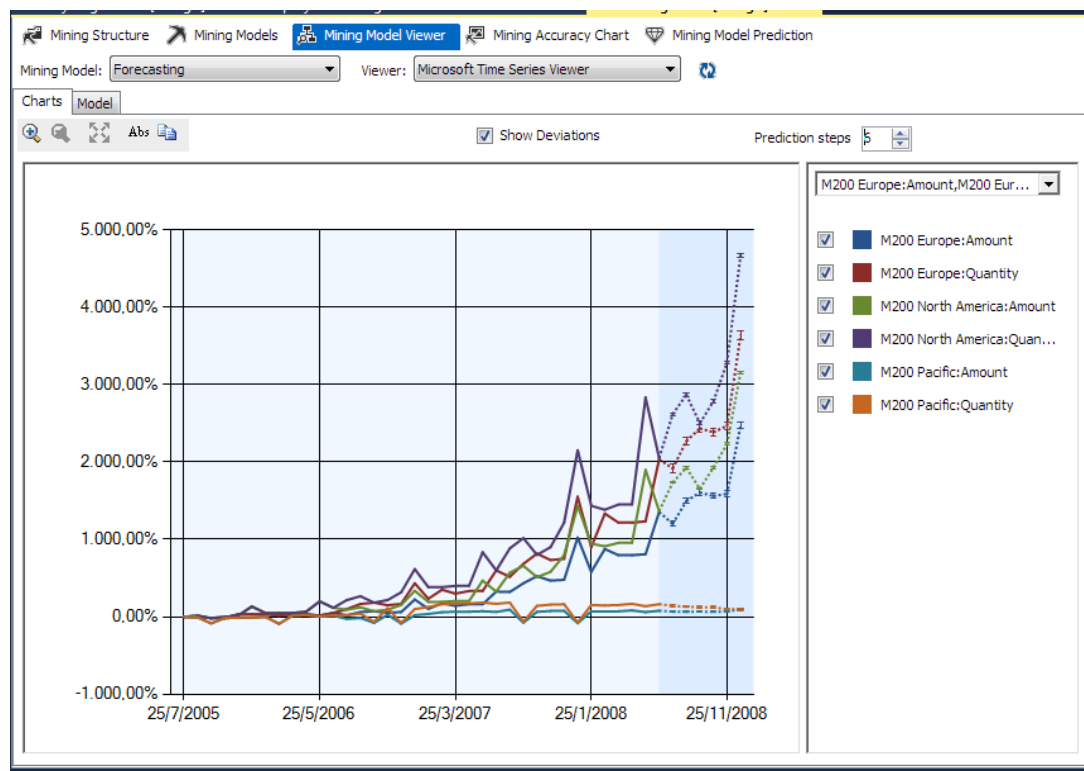
MS SQL Server 2014 (Κανόνες Συσχέτισης)

- Εφαρμογή αλγορίθμου Κανόνων Συσχέτισης
- Παράμετροι Αλγορίθμου
- Προβολή Κανόνων Συσχέτισης
- Αξιολόγηση των Κανόνων Συσχέτισης



MS SQL Server 2014 (Χρονοσειρές)

- Εφαρμογή αλγορίθμου Χρονοσειρών (ARIMA, ARTXP, mixed)
- Παράμετροι Αλγορίθμου (Periodicity, Forecast method, missing values substitution, etc.)
- Προβολή του μοντέλου time series
- Αξιολόγηση των μοντέλου time series



Εξόρυξη Δεδομένων Χρήσης Παγκοσμίου Ιστού

- Χρησιμοποιεί δεδομένα από αρχεία καταγραφής των ιστοσελίδων (**Log files**) και από τα προφίλ των χρηστών (π.χ. **βαθμολογίες χρηστών** σε προϊόντα)

	Item ₁	Item ₂	Item ₃	Item ₄
User ₁	-	4	2	-
User ₂	-	3	4	-
User ₃	4	-	-	4
User ₄	5	-		5

Βαθμός
1 έως 5

- Εφαρμογές: Συστήματα Συστάσεων Συνεργατικής Διήθησης. (Recommender Systems)

Εξόρυξη Δεδομένων Χρήσης Παγκοσμίου Ιστού

- **Data Collection**
- **Session Analysis**
- **Collaborative Filtering**
- **Recommender Systems**
- **Social Tagging**
- **Singular Value Decomposition**
- **Tensor Dimensionality Reduction**

Εξόρυξη Δεδομένων Περιεχομένου Παγκοσμίου Ιστού

- Χρησιμοποιεί το περιεχόμενο των ιστοσελίδων (κείμενο, λέξεις κτλ.) προκειμένου να βρει ομοιότητα μεταξύ τους.

	Term ₁	Term ₂	Term ₃	Term ₄
Web page ₁	12	10	-	-
Web page ₂	8	9	-	-
Web page ₃	-	5	5	3
Web page ₄	-	4	6	8

Συχνότητα

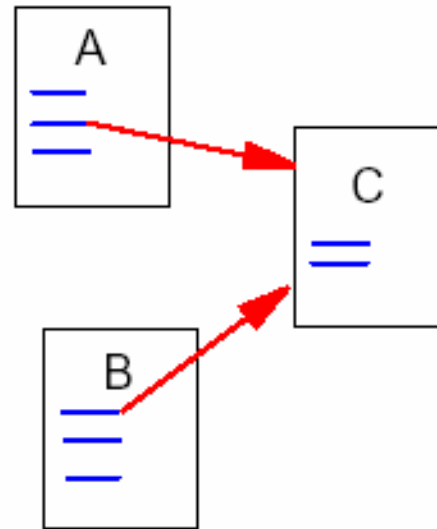
- Εφαρμογές: Συστήματα Προτάσεων Βάσει περιεχομένου, Ανάκτηση Πληροφοριών.

Εξόρυξη Δεδομένων Περιεχομένου Παγκοσμίου Ιστού

- **Information Retrieval Models**
- **Web Page Pre-processing**
- **Latent Semantic Indexing**
- **Web Spamming**
- **Content-based Collaborative Filtering**
- **Explanations in Recommender Systems**

Εξόρυξη Δεδομένων Δομής Παγκοσμίου Ιστού

- Ανακαλύπτει ενδιαφέρουσα γνώση από υπερσυνδέσμους μεταξύ ιστοσελίδων του Παγκοσμίου Ιστού.



- Εφαρμογές: Μηχανές Αναζήτησης, Ανακάλυψη Κοινοτήτων κλπ.

Εξόρυξη Δεδομένων Δομής Παγκοσμίου Ιστού

- **Social Network Analysis (friendship network)**
- **Web Search and Search Engines**
- **PageRank & Random walk with restart**
- **HITS and SimRank**
- **Community Discovery**

Εξόρυξη Δεδομένων Ειδικού Σκοπού

- Data Mining and Audience Intelligence for Advertising (**ADKDD**)
- Multimedia Data Mining (**MDM/KDD**)
- Knowledge Discovery on the Web (**WebKDD**)
- Knowledge Discovery from Sensor Data (**Sensor-KDD**)

Εξόρυξη Δεδομένων Ειδικού Σκοπού

- Data Mining in Bioinformatics (**BIOKDD**)
- Data Mining using Matrices and Tensors (**KDD**)
- Large-Scale Recommender Systems (**KDD**)
- Social Network Mining and Analysis (**SNA-KDD**)
- Advances in Social Network Analysis and Mining (**ASONAM**)
- Recommender Systems (**RecSys**)
- Principles and Practice of Knowledge Discovery (**PKDD**)

ΒΑΣΕΙΣ ΚΑΙ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Databases and Data Mining

Δρ. Τιάκας Ελευθέριος



Εργαστήριο Τεχνολογίας και Επεξεργασίας Δεδομένων
Τμήμα Πληροφορικής
Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

<http://delab.csd.auth.gr/~tiakas>