Πτυχιακή εργασία

Εξόρυξη δεδομένων και εφαρμογές της

Σπουδαστής: Ντάϊκο Γεώργιος (ΑΜ 9273)

Επιβλέπων: Στέφανος Ι. Καρναβάς

Φεβρουάριος 2026

Εξόρυξη δεδομένων και KDD

Εξόρυξη δεδομένων (data mining): εξαγωγή ή εξόρυξη γνώσης, από μεγάλες ποσότητες δεδομένων

KDD– Knowledge Discovery in Databases: ανακάλυψη γνώσης, σε βάσεις δεδομένων

Κάποιοι θεωρούν την εξόρυξη δεδομένων ως συνώνυμο της KDD, ενώ άλλοι ως απλά ένα βήμα σε ολόκληρη τη διαδικασία της KDD

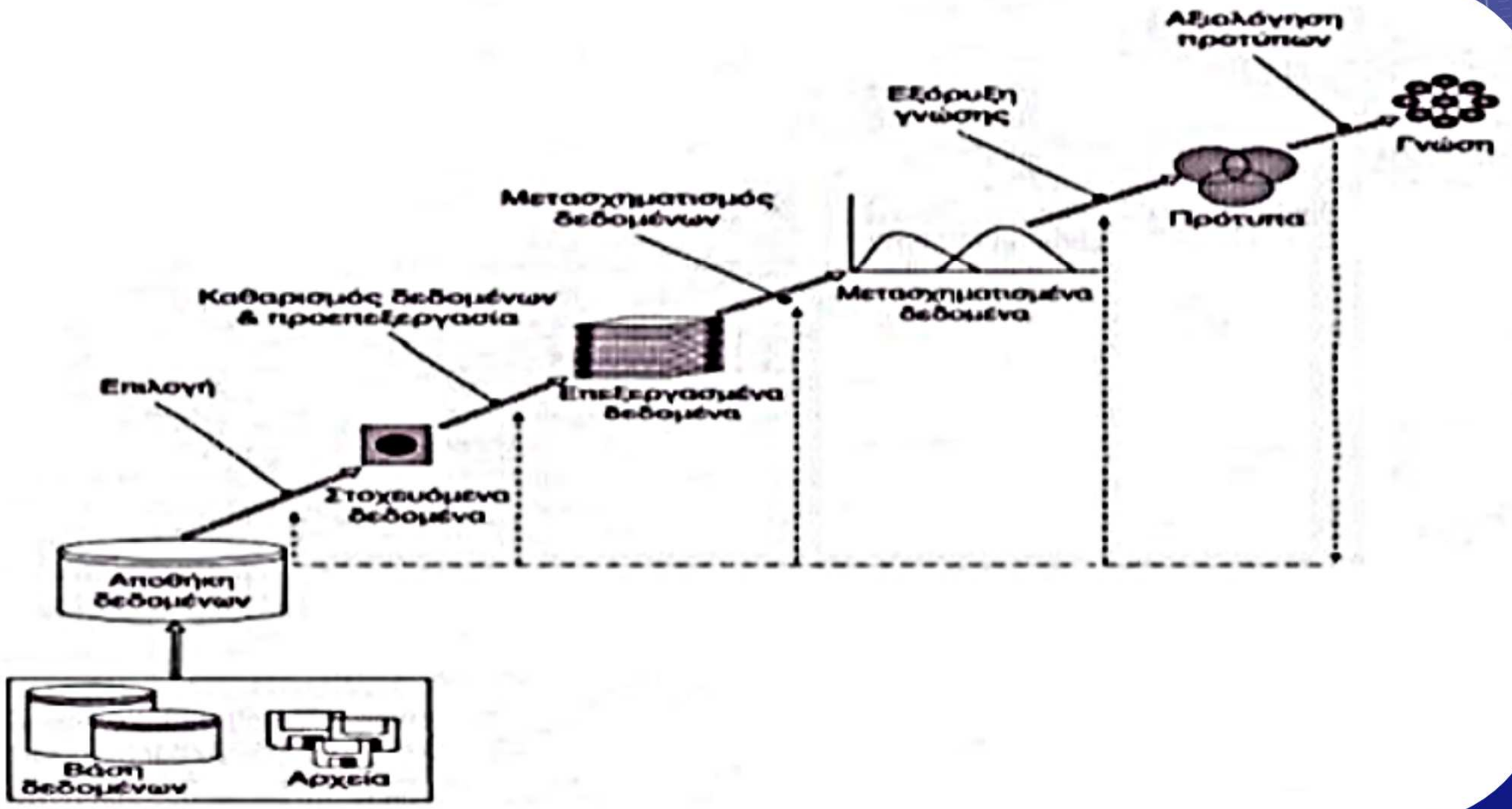
Εξόρυξη δεδομένων ⇨ Διαδικασία ανακάλυψης γνώσης

Ποιοι είναι οι λόγοι που οδήγησαν στην ανάπτυξη της εξόρυξης των δεδομένων;

- Η τεράστια αύξηση των δεδομένων
- Η μείωση του κόστους επεξεργασίας και η αύξηση της χωρητικότητας των αποθηκευμένων δεδομένων
- Η διαθεσιμότητα λογισμικού για data mining
- Η τεράστια και ταχέως αναπτυσσόμενη ποσότητα των δεδομένων, έχει ξεπεράσει κατά πολύ την ανθρώπινη ικανότητά για κατανόηση, χωρίς ισχυρά εργαλεία. Έτσι, τα δεδομένα που συλλέγονται σε μεγάλες βάσεις δεδομένων γίνονται «τάφοι δεδομένων».

Στόχος της εξόρυξης δεδομένων: μετατροπή των «τάφων δεδομένων» σε «χρυσά ψήγματα» γνώσης

Η βασική ροή των βημάτων της διαδικασίας KDD



Εξόρυξη δεδομένων: συμβολή πολλαπλών επιστημονικών κλάδων

Ρίζες της εξόρυξης δεδομένων:
- Στατιστική
- Μηχανική μάθηση
- Βάσεις δεδομένων

Σχεδιασμός και υλοποίηση ενός έργου εξόρυξης δεδομένων

Υπάρχουν διαφορετικές προσεγγίσεις

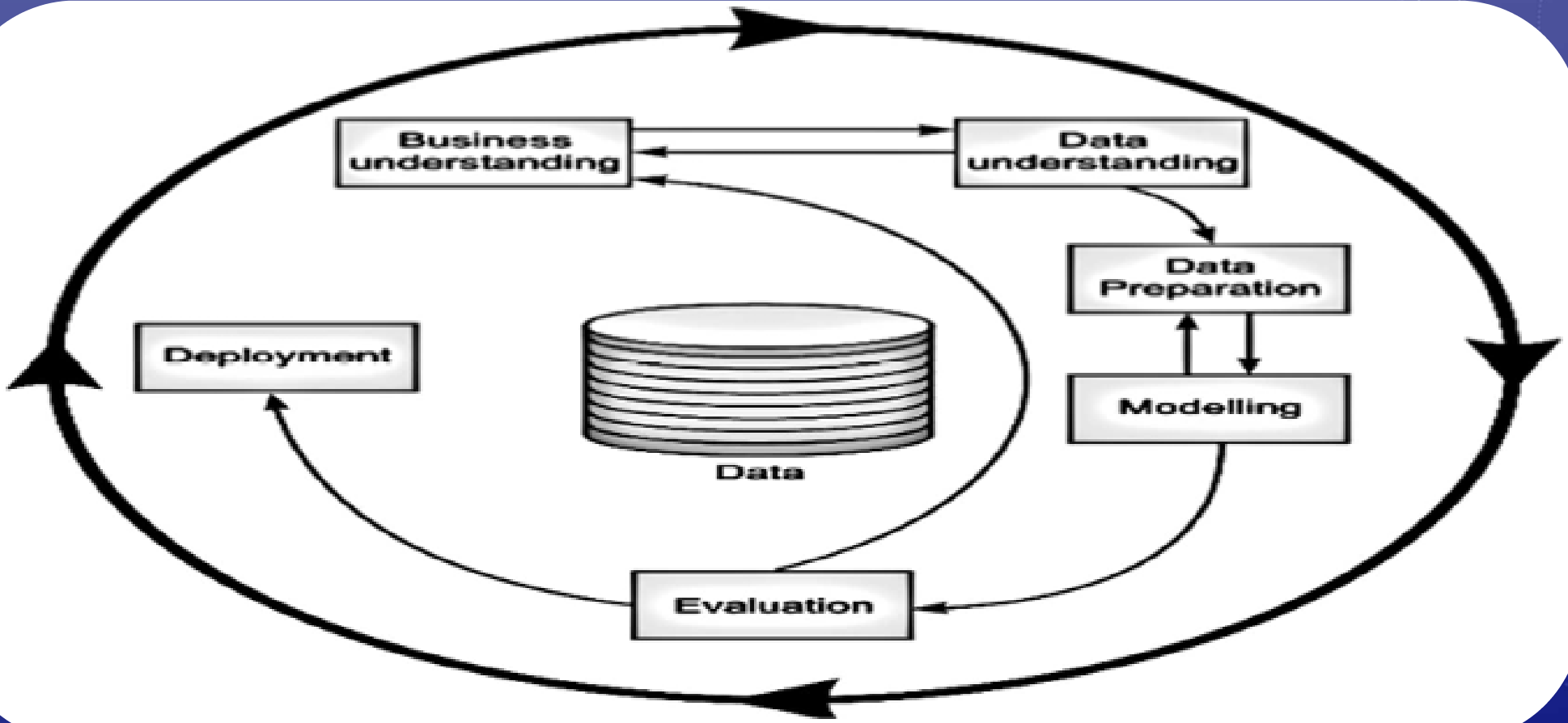
Μία προσέγγιση που αποτελεί προσαρμογή της γνωστής διαδικασίας ανάπτυξης λογισμικού, περιλαμβάνει τα ακόλουθα βήματα:

1) Ανάλυση των απαιτήσεων
2) Επιλογή και συλλογή των δεδομένων
3) Καθαρισμός και προετοιμασία των δεδομένων
4) Εξερεύνηση και επικύρωση της εξόρυξης των δεδομένων
5) Υλοποίηση, αξιολόγηση και παρακολούθηση
6) Οπτικοποίηση των αποτελεσμάτων

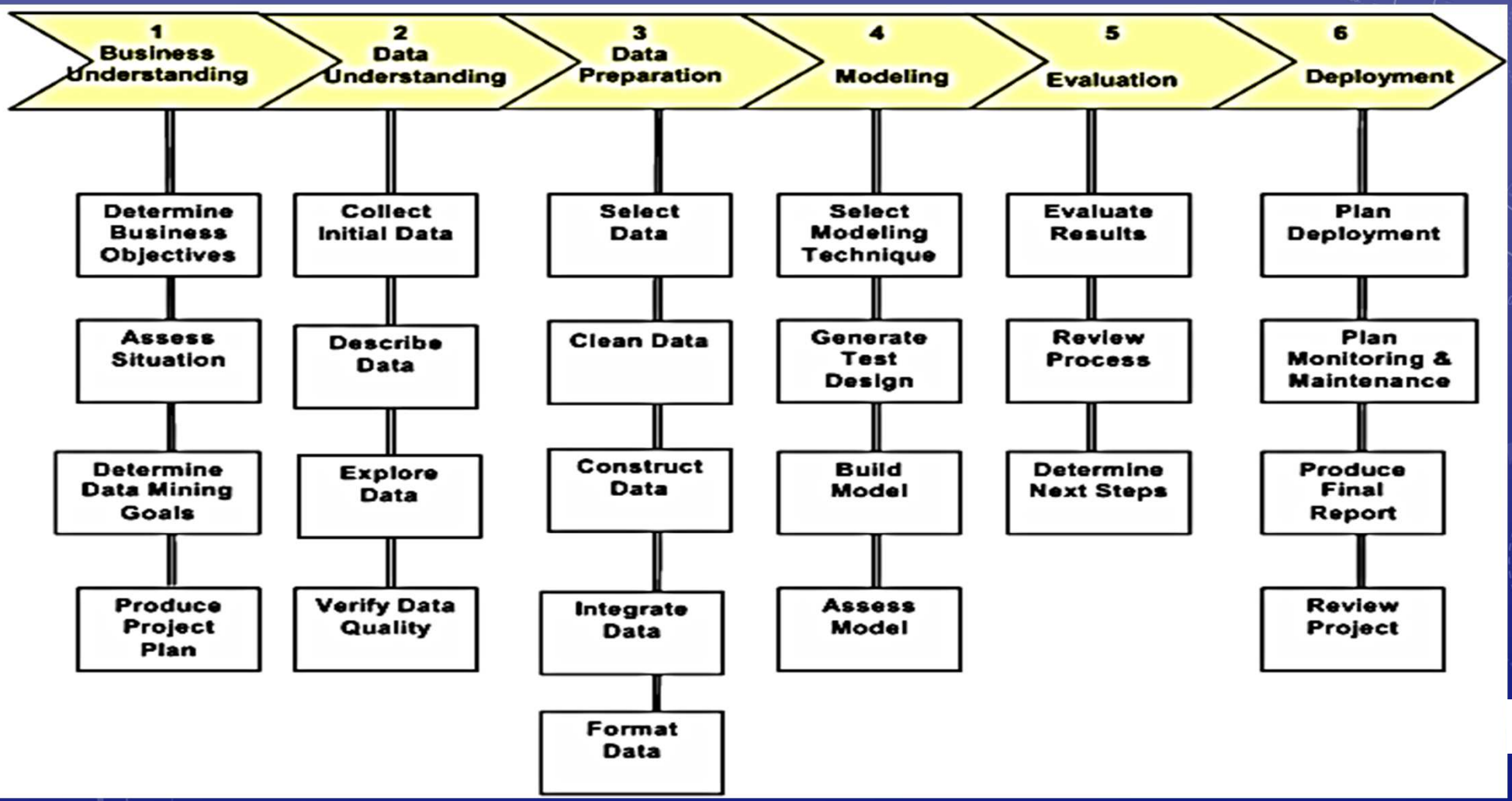
Άλλες προσεγγίσεις αποτελούν οι εξής:

- SPSS (Statistical Package for the Social Sciences)
- SAS (Statistical Analysis System)
- CRISP DM (Cross–Industry Standard Process for Data Mining)

Το μοντέλο εξόρυξης δεδομένων CRISP



Μεθοδολογία της διαδικασίας εξόρυξης δεδομένων CRISP–DM



Δεδομένα στα οποία πραγματοποιείται εξόρυξη δεδομένων

Η εξόρυξη δεδομένων, πρέπει να εφαρμόζεται σε κάθε είδους αποθετήριο πληροφοριών

Αυτό περιλαμβάνει:

- σχεσιακές βάσεις δεδομένων,
- αποθήκες δεδομένων,
- βάσεις δεδομένων συναλλαγών,
- προηγμένα συστήματα βάσεων δεδομένων και προηγμένες εφαρμογές βάσεων δεδομένων

Βασικές εργασίες της εξόρυξης γνώσης από δεδομένα 1/2

Κατηγοριοποίηση (Classification): Απεικονίζει τα δεδομένα, σε προκαθορισμένες ομάδες–κλάσεις (classes)

Παλινδρόμηση (Regression): Χρησιμοποιείται, προκειμένου να απεικονιστεί ένα στοιχειώδες δεδομένο, σε μία πραγματική μεταβλητή πρόβλεψης.

Ανάλυση χρονοσειρών (Time series analysis): Εξετάζεται η τιμή ενός γνωρίσματος, καθώς μεταβάλλεται με τον χρόνο, με τιμές που λαμβάνονται σε ίσα χρονικά διαστήματα (ωριαία, ημερήσια, εβδομαδιαία)

Πρόβλεψη (prediction): Θεωρείται το να δίνεται τιμή σε μία μελλοντική κατάσταση, παρά σε μία τρέχουσα

Συσταδοποίηση (clustering): Αναφέρεται και ως μη εποπτευόμενη μάθηση ή τμηματοποίηση. Δύναται να θεωρηθεί ως μία διαμέριση των δεδομένων σε ομάδες, που μπορεί να είναι (ή να μην είναι) διακριτές μεταξύ τους

Βασικές εργασίες της εξόρυξης γνώσης από δεδομένα 2/2

Παρουσίαση συνόψεων (summarization) **ή χαρακτηρισμός (characterization) ή γενίκευση** (generalization): Απεικονίζει τα δεδομένα σε υποσύνολα τους, με συνοδευτικές απλές αντιπροσωπευτικές περιγραφές ή συνοπτικές πληροφορίες (μέσος όρος ενός χαρακτηριστικού) σχετικά με τις βάσεις δεδομένων.

Κανόνες συσχέτισης (association rules) **ή ανάλυση συνδέσμων** (link analysis) **ή ανάλυση συγγένειας** (affinity analysis): Πρόκειται για μοντέλα που αναγνωρίζουν ειδικούς τύπους συσχέτισης μεταξύ των δεδομένων

Ανακάλυψη ακολουθιών (sequence discovery**) ή ακολουθιακή ανάλυση** (sequential analysis): Χρησιμοποιείται προκειμένου να καθορισθούν στα δεδομένα, σειριακά πρότυπα, που η συσχέτιση τους βασίζεται σε μία χρονική ακολουθία ενεργειών.

Κατηγορίες των μεθόδων εξόρυξης δεδομένων

Υπάρχουν δυο βασικοί στόχοι: η πρόβλεψη και η περιγραφή.

Η **πρόβλεψη**, στοχεύει στην εκτίμηση της συμπεριφοράς κάποιων μεταβλητών που παρουσιάζουν ενδιαφέρον και βασίζονται – επηρεάζονται από τη συμπεριφορά άλλων μεταβλητών.

Η **περιγραφή**, αναπαριστά τα δεδομένα μίας πολύπλοκης βάσης δεδομένων, με κατανοητό και αξιοποιήσιμο στόχο.

Σε ότι αφορά την εξόρυξη γνώσης, η περιγραφή τείνει να είναι σημαντικότερη από την πρόβλεψη. Αντιθέτως, σε ότι αφορά την αναγνώριση προτύπων και την εφαρμογή της μηχανικής μάθησης, η πρόβλεψη είναι σημαντικότερη.

Προβλεπτικά μοντέλα	**Περιγραφικά οντέλα**
Κατηγοριοποίηση (Classification)	Συσταδοποίηση (Clustering)
Παλινδρόμηση (Regression)	Παρουσίαση συνόψεων (Summarization)
Ανάλυση χρονοσειρών (Time series analysis)	Κανόνες συσχέτισης (Association rules)
Πρόβλεψη (Prediction)	Ανακάλυψη ακολουθιών (Sequence discovery)

Απαιτήσεις της εξόρυξης δεδομένων

Χειρισμός των διαφορετικών τύπων δεδομένων

Απόδοση και εξελιξιμότητα, των αλγορίθμων εξόρυξης δεδομένων.

Χρησιμότητα, βεβαιότητα και εκφραστικότητα των αποτελεσμάτων της εξόρυξης δεδομένων

Διαλογική ανακάλυψη γνώσης, στα πολυεννοιολογικά επίπεδα

Εξόρυξη γνώσης, από διαφορετικές πηγές δεδομένων.

Λογισμικά εξόρυξης δεδομένων

Μερικά από τα σημαντικότερων προγράμματα που εφαρμόζονται στις τεχνικές εξόρυξης δεδομένων, είναι τα:

WEKA, Tanagra, Rattle, Carrot 2, XL–MINER, R,
SAS, Orange, Rapid Miner, Clementine

Πλεονεκτήματα της εξόρυξης δεδομένων

Παροχή καλύτερων πληροφοριών, για την επίτευξη ανταγωνιστικού πλεονεκτήματος

Προσθήκη αξίας, σε μια αποθήκη δεδομένων

Επίλυση προβλημάτων έρευνας

Αύξηση της λειτουργικής αποδοτικότητας

Παροχή ευελιξίας, στη χρήση των δεδομένων

Μείωση, του λειτουργικού κόστους.

Ετοιμότητα προς χρήση

Μειονεκτήματα της εξόρυξης δεδομένων

Υψηλό κόστος

Πολύπλοκο και χρονοβόρο έργο

Απόρρητο

Υψηλές απαιτήσεις γνώσης από τον χρήστη

Μη διαχειρίσιμη βάση δεδομένων

Λανθασμένες πληροφορίες από σφάλματα στα δεδομένα

Τομείς που βρίσκει εφαρμογή η εξόρυξη δεδομένων 1/2

Τραπεζικές και χρηματοοικονομικές υπηρεσίες

Βιολογία, ιατρική επιστήμη, υγειονομική περίθαλψη

Ανίχνευση απάτης, πρόληψη εγκλήματος

Επιχειρηματική ευφυΐα

Μηχανές αναζήτησης ιστού

Κοινωνικά μέσα, κοινωνικά δίκτυα

Λιανικό εμπόριο, υπηρεσίες

Τηλεμάρκετινγκ και άμεσο μάρκετινγκ

Ηλεκτρονικό εμπόριο

Τηλεπικοινωνίες

Εκπαίδευση

Τομείς που βρίσκει εφαρμογή η εξόρυξη δεδομένων 2/2

Κατασκευαστές

Ασφαλιστικές εταιρείες

Αθλητισμός

Εφορία

Ψηφιακή βιβλιοθήκη

Φαρμακευτική βιομηχανία

Συστήματα συστάσεων

Εξόρυξη δεδομένων και ηθική

Η χρήση δεδομένων (ιδίως σχετικά με άτομα) για την εξόρυξη γνώσης, έχει σοβαρές ηθικές επιπτώσεις. Οι επαγγελματίες των τεχνικών εξόρυξης δεδομένων, πρέπει να ενεργούν υπεύθυνα, γνωρίζοντας τα ηθικά ζητήματα που περιβάλλουν τη συγκεκριμένη εφαρμογή. Όταν παρουσιάζονται δεδομένα, πρέπει να αναρωτηθούμε:

- ποιος επιτρέπεται να έχει πρόσβαση σε αυτά,
- για ποιο σκοπό συλλέχθηκαν
- τι είδους συμπεράσματα είναι θεμιτό να εξαχθούν από αυτά.

Η ηθική διάσταση, εγείρει δύσκολα ερωτήματα για όσους ασχολούνται με την πρακτική εξόρυξη δεδομένων η οποία αποτελεί απλά ένα εργαλείο σε όλη τη διαδικασία. Οι άνθρωποι είναι αυτοί που λαμβάνουν τα αποτελέσματα, μαζί με άλλες γνώσεις και αποφασίζουν ποια ενέργεια θα εφαρμόσουν.

Η ανάγκη για ανθρώπινη κατεύθυνση στην εξόρυξη δεδομένων

Πολλοί προμηθευτές λογισμικού, το προωθούν ως μια εφαρμογή plug and play (έτοιμη προς χρήση), που παρέχει λύσεις σε δυσεπίλυτα προβλήματα, χωρίς την ανάγκη ανθρώπινης επίβλεψης ή αλληλεπίδρασης.

Μερικοί πρώιμοι ορισμοί της εξόρυξης δεδομένων, ακολούθησαν την εστίαση στον αυτοματισμό. Αυτό έχει παραπλανήσει πολλούς στο να πιστεύουν ότι η εξόρυξη δεδομένων είναι ένα προϊόν που μπορεί να αγοραστεί και όχι ένας κλάδος που πρέπει να κατακτηθεί.

Ο αυτοματισμός, δεν υποκαθιστά τους ανθρώπους που πρέπει να συμμετέχουν ενεργά σε κάθε φάση της διαδικασίας εξόρυξης δεδομένων. Αντί να ρωτάμε πού εντάσσονται οι άνθρωποι στην εξόρυξη δεδομένων, πρέπει να αναρωτηθούμε πώς μπορούμε να την εντάξουμε στην ανθρώπινη διαδικασία επίλυσης προβλημάτων.

Η δύναμη των τρομερών αλγορίθμων εξόρυξης δεδομένων, καθιστά την κακή χρήση τους αναλογικά πιο επικίνδυνη.

Σας ευχαριστώ πολύ για την προσοχή σας ☺