Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

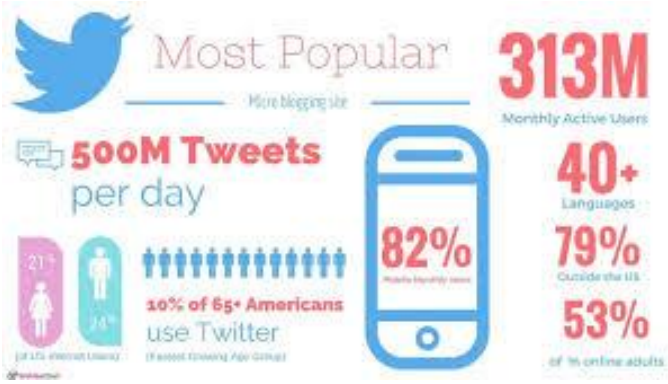
Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών

Βασικές Αρχές Εξόρυξης Γνώσης

Τι είναι η εξόρυξη δεδομένων (Data Mining)

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Wikipedia

Δεδομένα υπάρχουν παντού .....



amazon

Top 5 shopping days for Amazon.com during the last year

Date	Visits
Cyber Monday 2015	95,347,279
Black Friday 2015	87,057,257
Amazon Prime Day 2015	86,374,093
Amazon Prime Day 2016	81,574,924
Sunday before Cyber Monday 2015	80,225,862

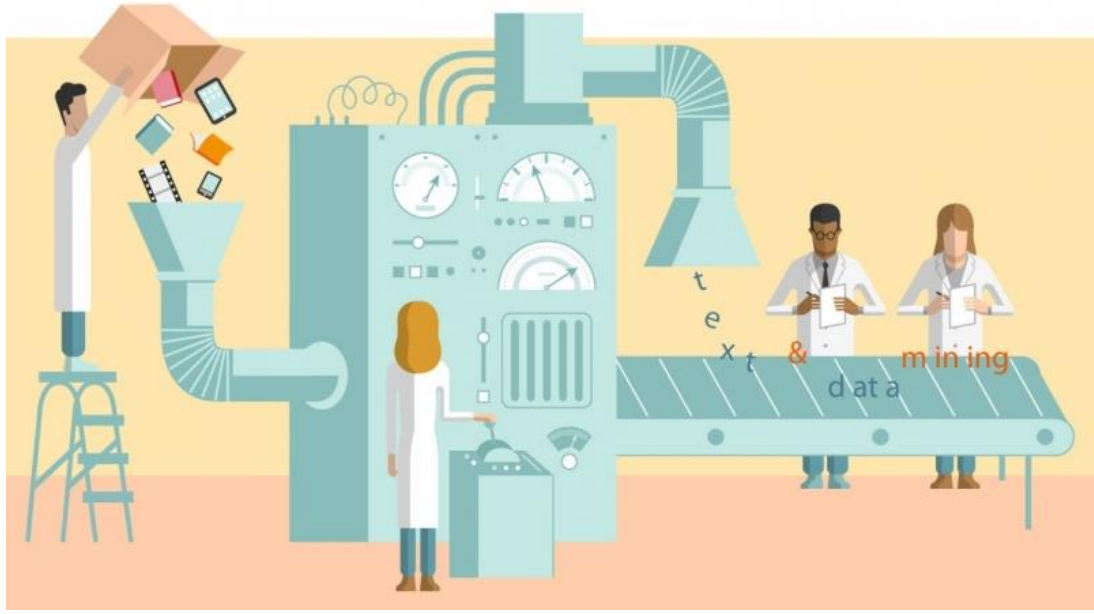
Source: Hitwise, a division of Connexity

Δεδομένα Δημιουργούνται Συνέχεια

Πηγή https://www.domo.com/learn/data-never-sleeps-7

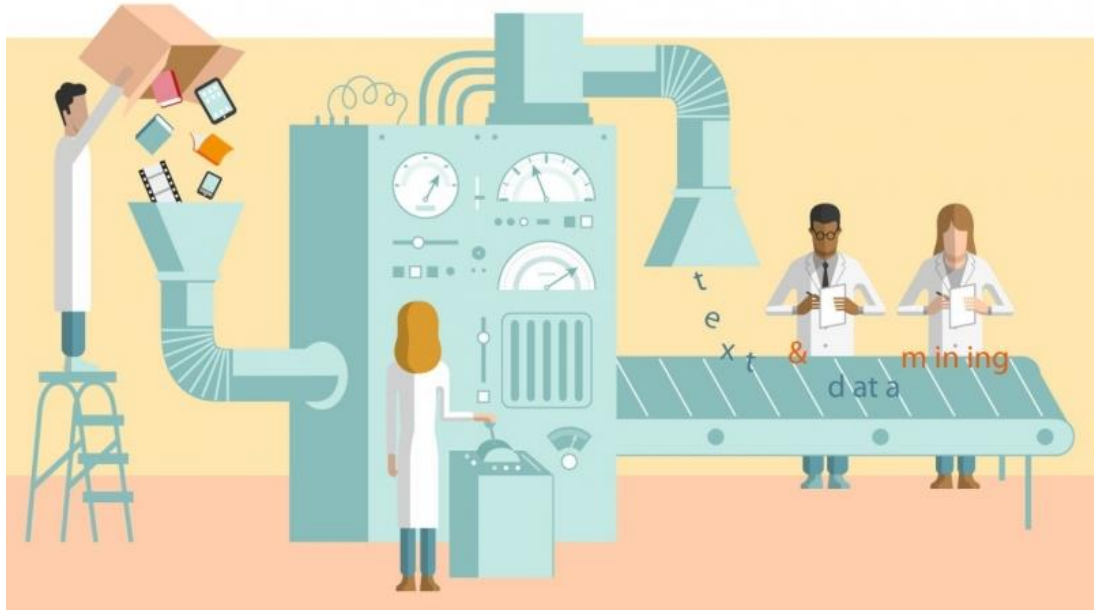
Γιατί Data Mining

- Μπορούμε να κάνουμε ανάλυση μεγάλου όγκου δεδομένων, όπου με τους παραδοσιακούς τρόπους θα ήταν δύσκολο έως αδύνατο.
- Οι επιστημονικές ομάδες βοηθούνται μέσω της ανάλυσης των δεδομένων στην μελέτη και έγκυρη διάγνωση φαινομένων.
- Μια εταιρία μπορεί να μελετήσει τους πελάτες της και να σχεδιάσει πιο αποδοτικές στρατηγικές διαφήμισης με σκοπό να αυξήσει τα έσοδα και να μειώσει τα κόστη.



Γιατί Data Mining

- Μπορούν να αναλυθούν οι τάσεις που παρατηρούνται.
- Να εντοπιστούν ακραίες συμπεριφορές οι οποίες ενδεχομένως να οδηγούν σε φαινόμενα απάτης.
- Να εφαρμοστούν προγνωστικές διαδικασίες σχετικά με την κατάσταση που μπορεί να περιέλθει ένα αντικείμενο ή μια οντότητα.
- Να πραγματοποιηθεί έλεγχος Υποθέσεων
- Να εντοπιστούν πιθανές ομάδες εντός ενός συνόλου πραγμάτων
- Να εντοπιστούν συσχετίσεις μεταξύ πραγμάτων



Πεδία Εφαρμογής

Οικονομικές απάτες

Πρόβλεψη του καιρού

Στο χώρο της διαφήμισης

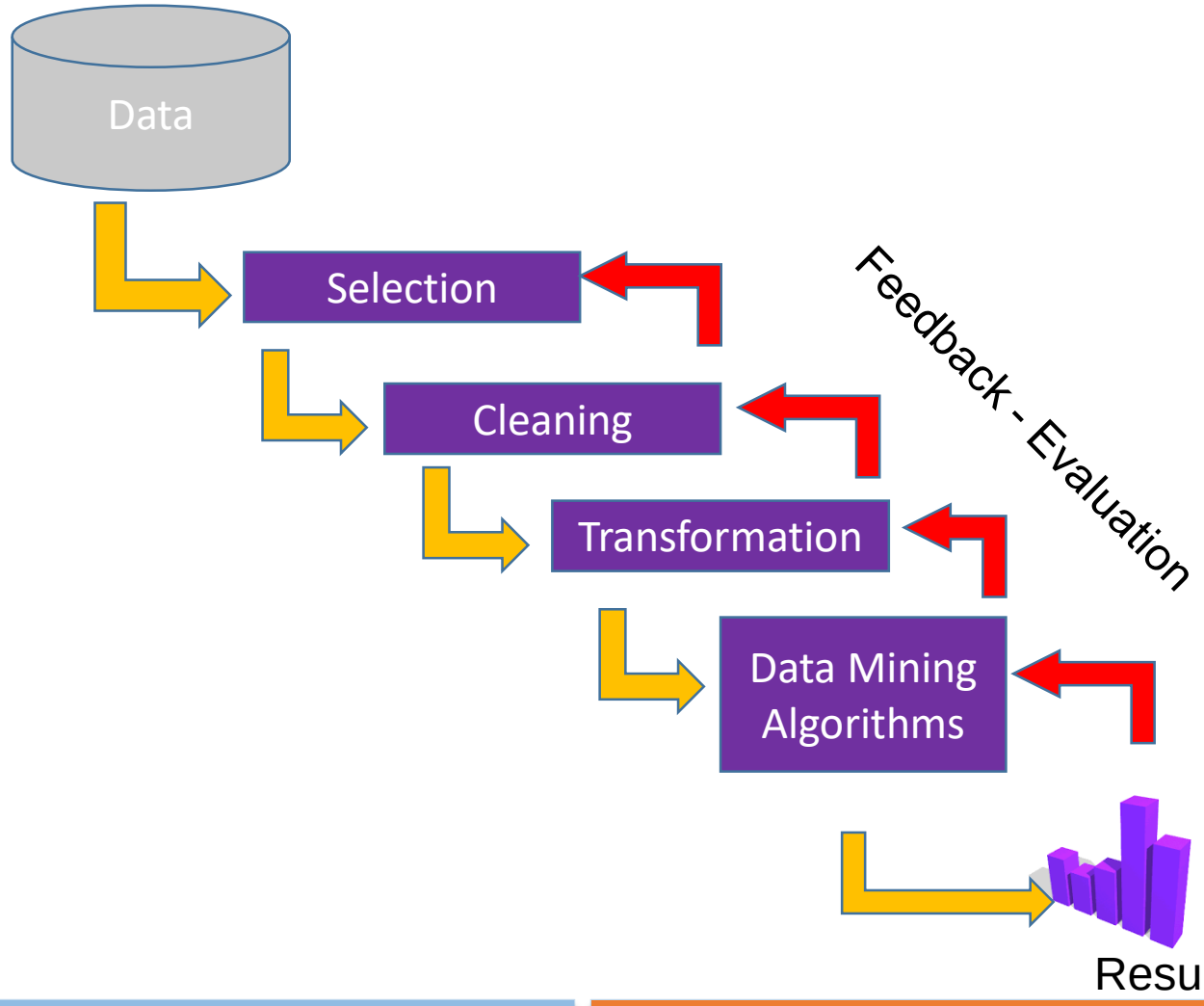
Βελτίωση του τρόπου ζωής

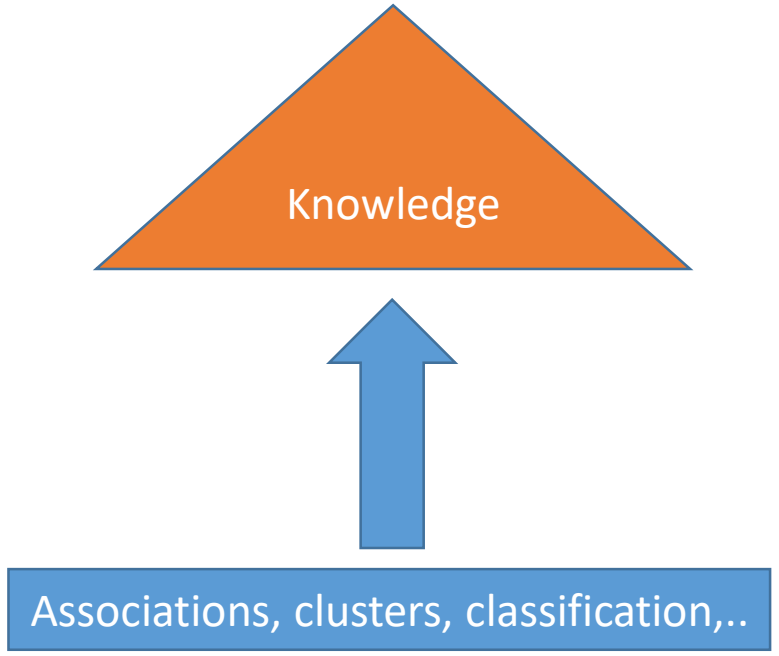
Διαδικασία Εξόρυξης Γνώσης

Η διαδικασία της Εξόρυξης Γνώσης είναι μια αμφίδρομη διαδικασία με πολλά επαναλαμβανόμενα στάδια δεδομένου ότι

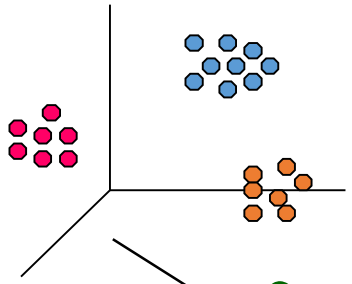
- Οι χρήστες συχνά δεν έχουν εκ των προτέρων καθαρή εικόνα για το ποια πληροφορία είναι ενδιαφέρουσα.
- Μέσω της παραγωγής των πρώτων συμπερασμάτων προκύπτουν νέα ερωτήματα.
- Τα αποτελέσματα της ανάλυσης των δεδομένων δεν οδήγησαν σε χρήσιμα συμπεράσματα, με σκοπό τον επανασχεδιασμό της διαδικασίας.

Knowledge Discovery



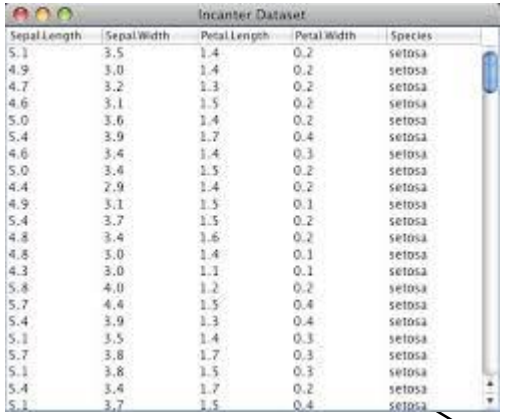


Εξόρυξη Γνώσης...

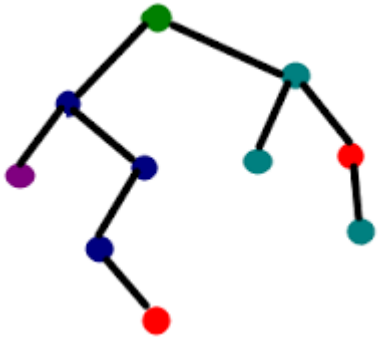


Clustering

Δεδομένα

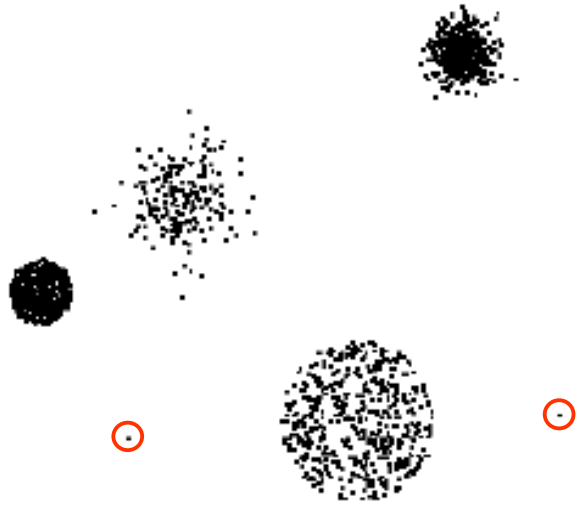


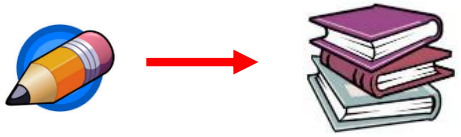
Prediction



Association Rules

Outlier Detection





Τα βασικά στάδια της εξόρυξης γνώσης

- Γνώση του αντικειμένου και του χώρου που καλούμαστε να αναλύσουμε.
- Καλή κατανόηση του προβλήματος.
- Επιλογή των δεδομένων.
- Προετοιμασία και μετασχηματισμός των δεδομένων (data cleaning, reduction & transformation).
- Επιλογή του κατάλληλου αλγορίθμου.
- Αξιολόγηση των αποτελεσμάτων.
- Εξαγωγή συμπερασμάτων.
- Προβολή των αποτελεσμάτων.

Προβλήματα με τα δεδομένα

- Κακός σχεδιασμός της βάσης ή του τρόπου συλλογής των δεδομένων, με αποτέλεσμα να έχουμε ελλιπείς τιμές
- Λάθη κατά την εισαγωγή των δεδομένων
- Ασυνεπείς τιμές οι οποίες μπορεί πάλι να οφείλονται σε λάθος κατά την εισαγωγή των τιμών ή την ομογενοποίηση των δεδομένων από διαφορετικές πηγές
- Τιμές εκτός κάποιων λογικών ορίων

Όλα τα παραπάνω αποτελούν μόνο μερικά από τα προβλήματα που έχει να αντιμετωπίσει ένας αναλυτής και τα οποία αν δεν διορθωθούν θα οδηγήσουν σε λάθος συμπεράσματα.

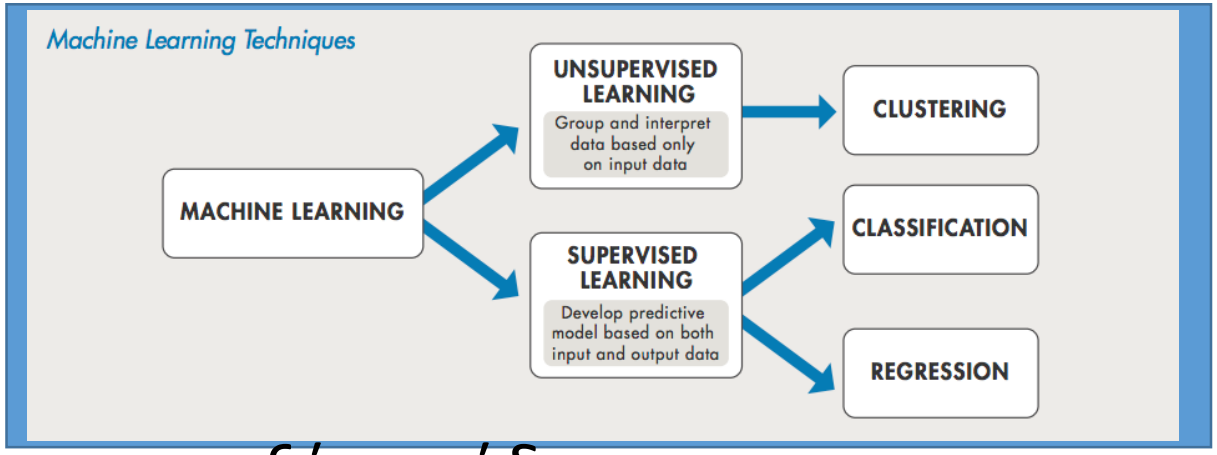
Μέθοδοι Προετοιμασίας των Δεδομένων

- Απαλοιφή θορύβου.
- Εντοπισμός ανωμαλιών.
- Διόρθωση Ελλιπών τιμών.
- Μετασχηματισμός των δεδομένων.
- Μείωση μεγέθους των δεδομένων.
  - Σε επίπεδο εγγραφών, μέσω της δειγματοληψίας.
  - Σε επίπεδο χαρακτηριστικών, μέσω αλγορίθμων όπου μπορεί να δημιουργήσουμε μια νέα μεταβλητή μέσω της συνένωσης δυο υπαρχουσών ή να παραλείψουμε κάποιες καθώς είναι ισοδύναμες με κάποιες άλλες.

Οι βασικές κατηγορίες Αλγορίθμων είναι

Οι βασικές κατηγορίες Αλγορίθμων είναι

- Classification
- Regression
- Clustering
- Association



Επίσης οι αλγόριθμοι διακρίνονται στις εξής ομάδες

- Επιβλεπόμενης Μάθησης
- Μη Επιβλεπόμενης Μάθησης

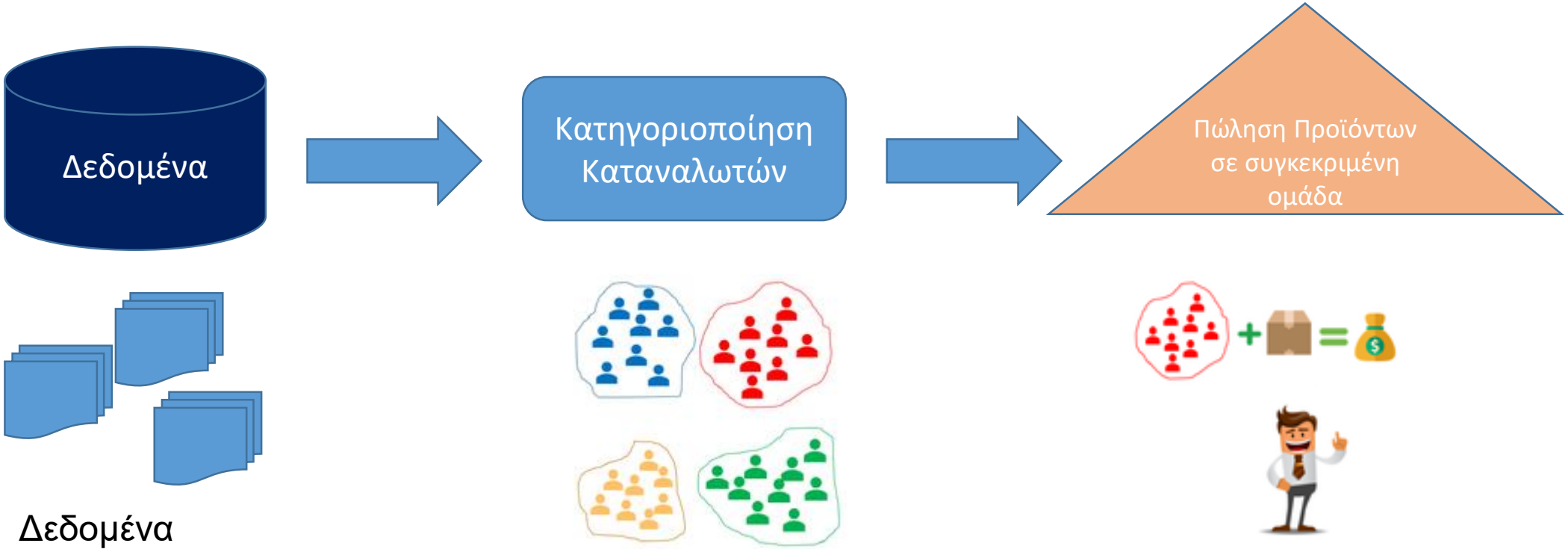
Supervised vs Unsupervised

- Στην επιβλεπόμενη μάθηση (Supervised Learning) μας δίνεται ένα σύνολο δεδομένων με τις αντίστοιχες ομάδες (κλάσεις) κάθε εγγραφής. Στόχος είναι η δημιουργία ενός μοντέλου, το οποίο όταν θα δέχεται νέα δεδομένα να μπορεί να τα κατηγοριοποιεί σε κάποια από τις προϋπάρχουσες κλάσεις.
- Στη μη επιβλεπομένη μάθηση (Unsupervised Learning) μας δίνεται ένα σύνολο δεδομένων, χωρίς όμως τις αντίστοιχες κλάσεις κάθε εγγραφής. Οπότε έχουμε δεδομένα χωρίς να γνωρίσουμε σε ποια κλάση ανήκουν. Στόχος είναι η ανάλυση αυτών των δεδομένων προκειμένου να ανακαλύψουμε κάποια ενδεχομένως ενδιαφέροντα στοιχεία στα δεδομένα.

Παράδειγμα Clustering για προώθηση προϊόντων.

- **Στόχος:** Διαχωρισμός του πληθυσμού σε ομάδες όπου θα πρέπει να επιλέγεται η κατάλληλη ομάδα κάθε φορά προκειμένου να κάνουμε στοχευμένη προώθηση ενός προϊόντος.
- **Προσέγγιση:**
  - Συγκεντρώνουμε όσες πιο πολλές πληροφορίες μπορούμε σχετικά με τους καταναλωτές του δείγματός μας.
  - Σχηματισμός ομάδων με κοινά χαρακτηριστικά
  - Εντοπισμός συμπεριφορών της κάθε ομάδας.

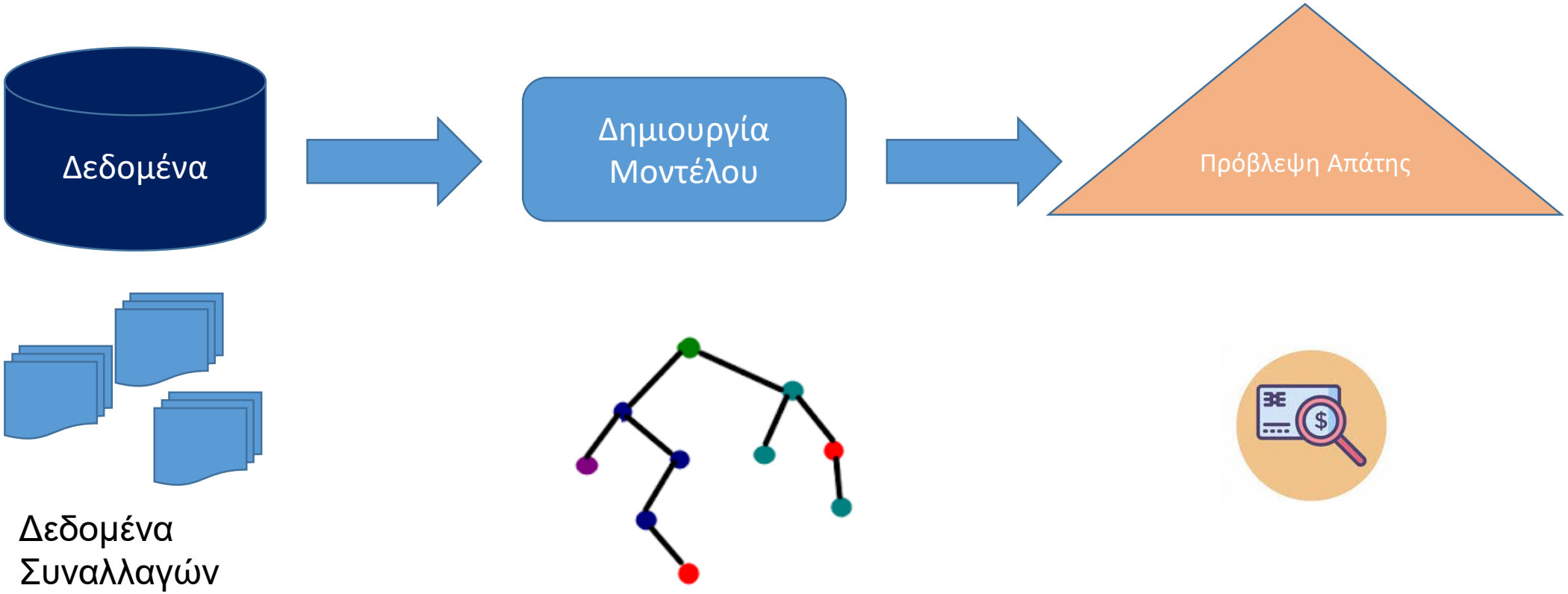
Κατηγοριοποίηση



Παράδειγμα Classification για εντοπισμό απάτης

- **Στόχος:** Πρόβλεψη περιπτώσεων απάτης σε τραπεζικές συναλλαγές.
- **Προσέγγιση:**
  - Καταγράφουμε όλες τις συναλλαγές μέσω κάρτας. Τι αγόρασε , πότε το αγόρασε, πότε πλήρωσε,…
  - Σε όλες αυτές τις συναλλαγές βάζουμε και μια τιμή η οποία δείχνει αν έχει γίνει απάτη ή όχι.
  - Χτίζουμε ένα μοντέλο το οποίο διαβάζει όλο το ιστορικό των συναλλαγών γνωρίζοντας αν υπάρχει απάτη ή όχι.
  - Στην συνέχεια όταν μας έρχεται μια νέα συναλλαγή ψάχνουμε μέσα στο μοντέλο να δούμε αν υπάρχει παρόμοια συναλλαγή και αν αυτή οδήγησε σε απάτη.

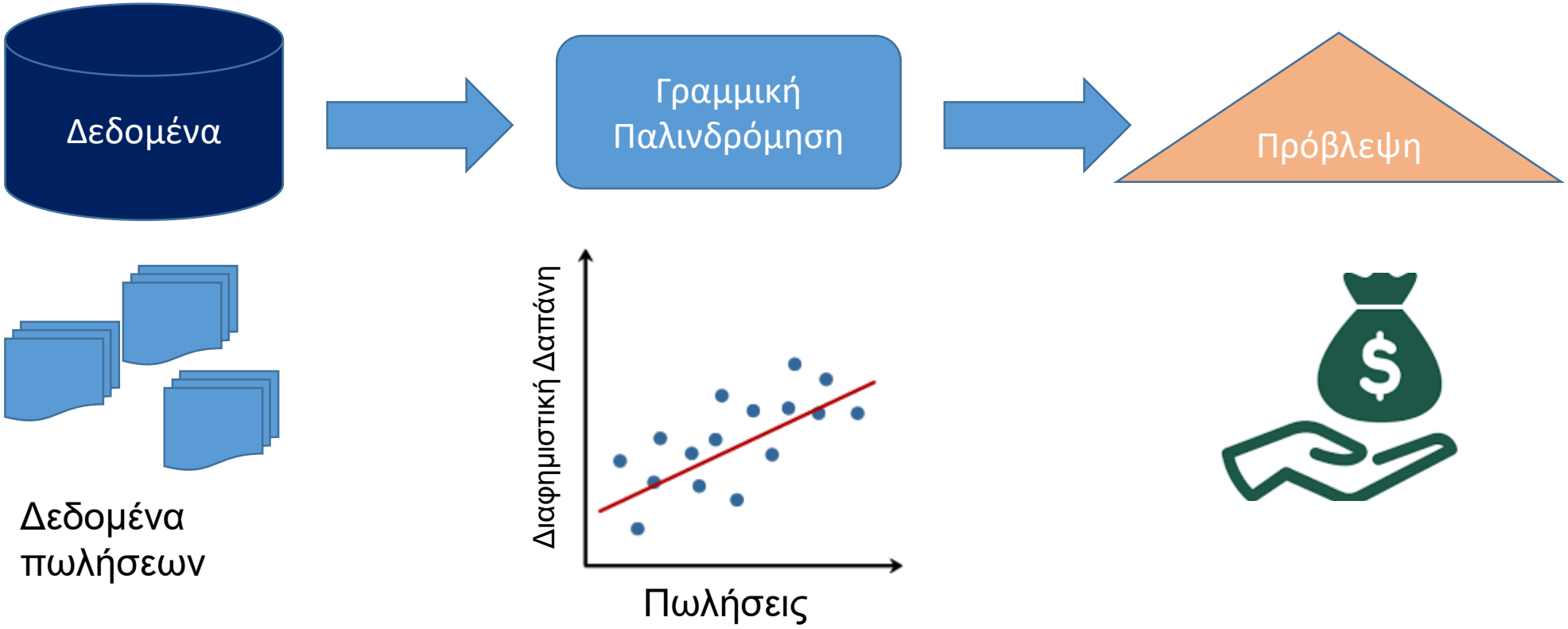
Πρόβλεψη


Δεδομένα
Δημιουργία Μοντέλου
Πρόβλεψη Απάτης
Δεδομένα Συναλλαγών

Παράδειγμα Regression για πρόβλεψη τιμών

- **Στόχος:** Πρόβλεψη του αριθμού των πωλήσεων ενός προϊόντος βάση της δαπάνης για διαφήμιση.
- **Προσέγγιση:**
  - Συγκεντρώνουμε όλα τα στοιχεία που σχετίζονται με τις πωλήσεις των προϊόντων της εταιρίας.
  - Για όλα τα παραπάνω προϊόντα καταγράφουμε το κόστος της διαφημιστικής δαπάνης.
  - Στη συνέχεια εφαρμόζουμε το μοντέλο της παλινδρόμησης το οποίο χτίζει την συνάρτηση βάση της οποίας μπορεί να γίνει η πρόβλεψη του αριθμού πωλήσεων σε σχέση με τη διαφημιστική δαπάνη.

Πρόβλεψη


Δεδομένα
Γραμμική Παλινδρόμηση
Πρόβλεψη
Δεδομένα πωλήσεων
Διαφημιστική Δαπάνη
Πωλήσεις

Παράδειγμα Κανόνων Συσχετίσεων

- **Στόχος:** Εντοπισμός συσχετίσεων στις αγοραστικές συνήθειες των καταναλωτών

- **Προσέγγιση:**
  - Συγκεντρώνουμε όλα τα στοιχεία που σχετίζονται με τις αγορές των πελατών της εταιρίας.
  - Για κάθε προϊόν συγκεντρώνουμε και την πληροφορία σχετικά με αυτό που αγοράστηκε κατά την ίδια συναλλαγή.
  - Στην συνέχει μαζεύουμε όλες τα συσχετιζόμενα προϊόντα.
  - Για κάθε συσχέτιση υπολογίζουμε το πόσες φορές εμφανίζονται σε σχέση με το συνολικό πλήθος των πωλήσεων
  - Τέλος όταν πάει ένας πελάτης να αγοράσει ένα προϊόν τότε το σύστημα του προτείνει αυτόματα τα υπόλοιπα προϊόντα με τα οποία εμφανίζει υψηλό ποσοστό συσχέτισης

Κανόνες Συσχετίσεων