

CN6005

# Τεχνητή Νοημοσύνη

Machine Learning  
and Data Mining



University  
of  
East London



METROPOLITAN  
COLLEGE

# Μηχανική μάθηση (Machine Learning)



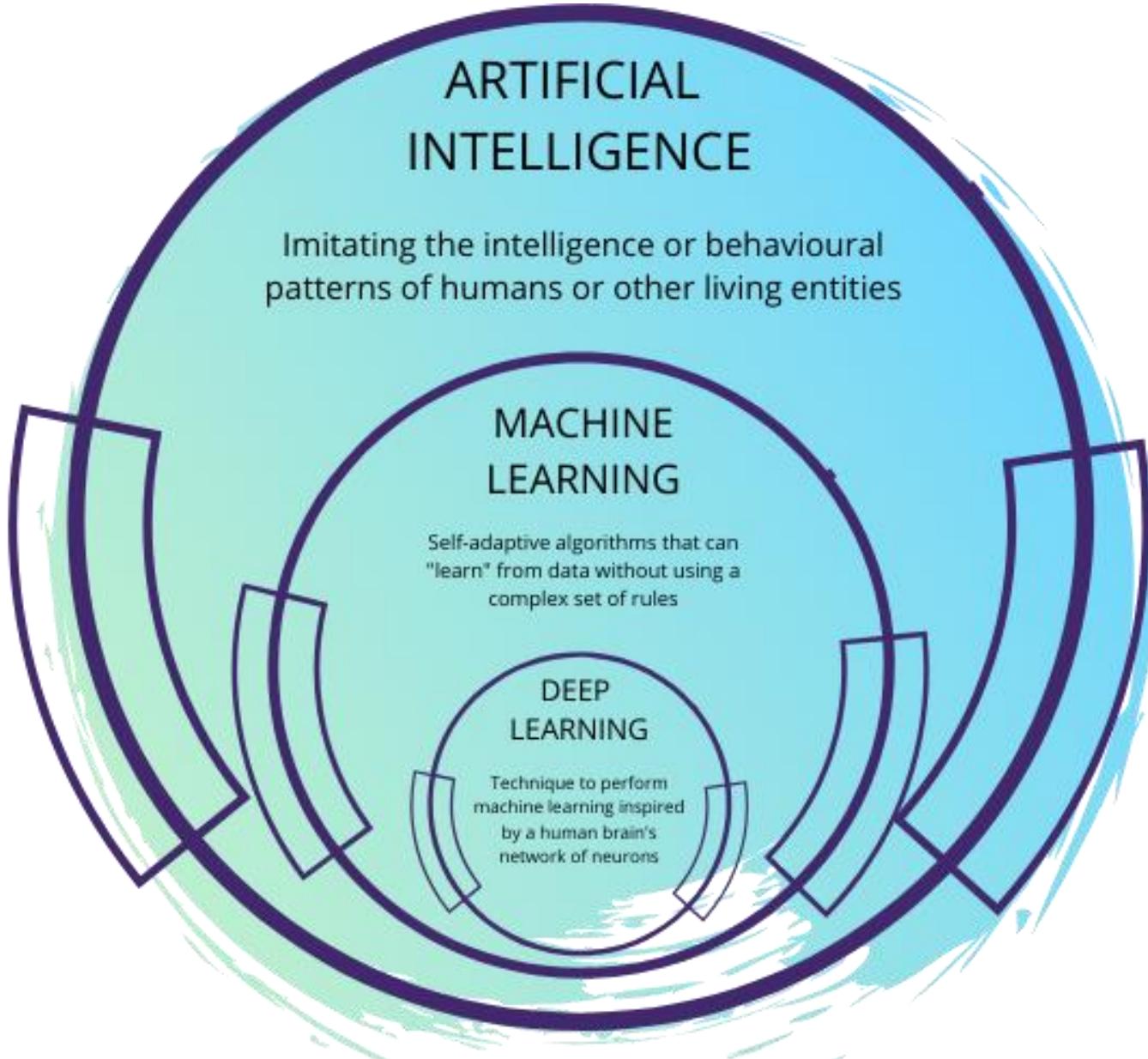
University  
of  
East London



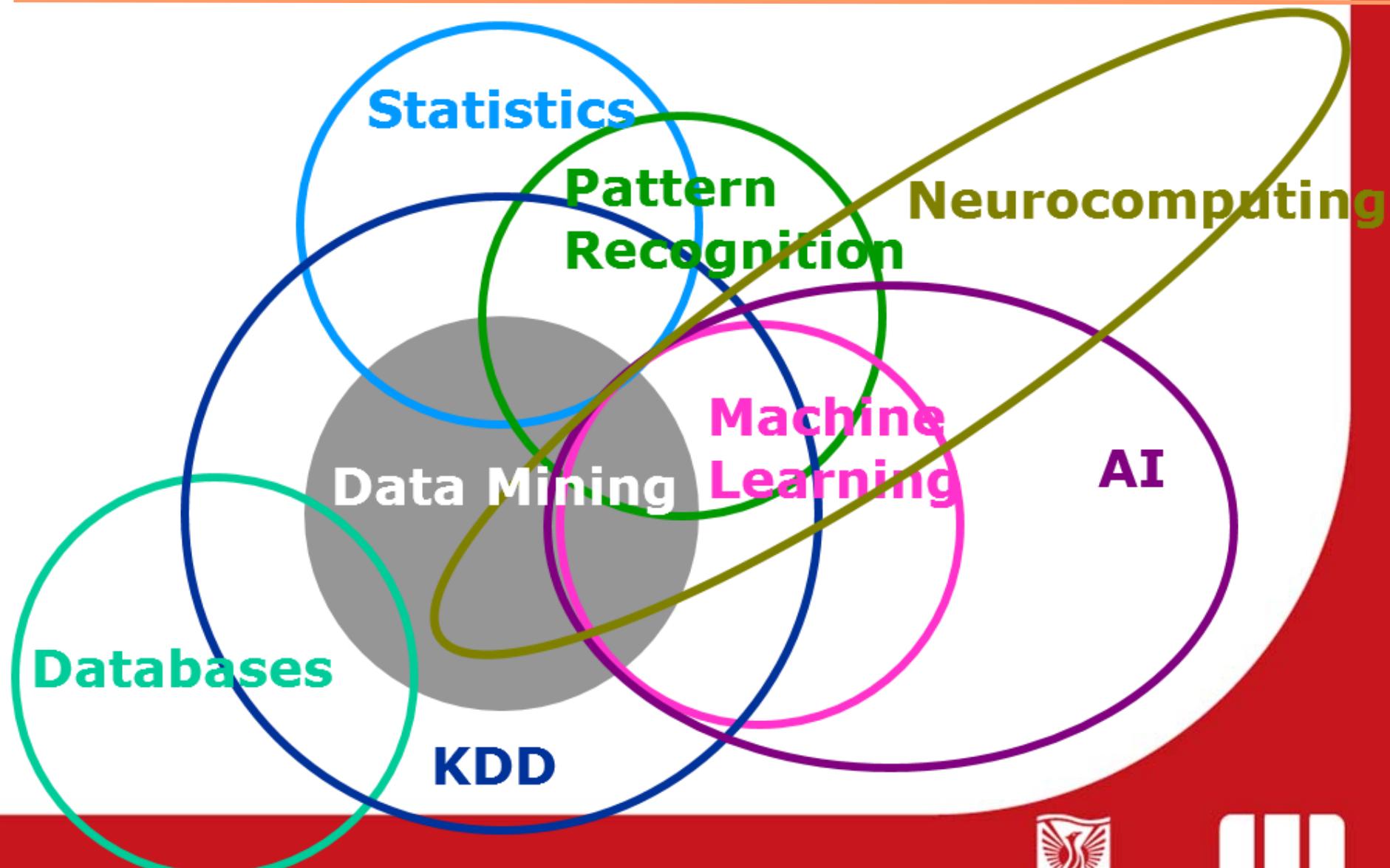
METROPOLITAN  
COLLEGE

# Μηχανική μάθηση - Εισαγωγή

---



# Μηχανική μάθηση - Εισαγωγή



University  
of  
East London



METROPOLITAN  
COLLEGE

# Μηχανική μάθηση - Εισαγωγή

- ❖ Η μάθηση σε ένα γνωστικό σύστημα, όπως γίνεται αντιληπτή στην καθημερινή ζωή, μπορεί να συνδεθεί με δύο βασικές ιδιότητες:
  - ❑ την ικανότητά στην πρόσκτηση γνώσης κατά την αλληλεπίδρασή του με το περιβάλλον,
  - ❑ την ικανότητά να βελτιώνει με την επανάληψη τον τρόπο εκτέλεσης μία ενέργειας.
- ❖ Έχουν προταθεί διάφοροι ορισμοί για τη μάθηση:
  - ❑ Simon ('83), "η μάθηση σηματοδοτεί προσαρμοστικές αλλαγές σε ένα σύστημα με την έννοια ότι αυτές του επιτρέπουν να κάνει την ίδια εργασία, ή εργασίες της ίδιας κατηγορίας, πιο αποδοτικά και αποτελεσματικά την επόμενη φορά".
  - ❑ Minsky ('85), "... είναι να κάνουμε χρήσιμες αλλαγές στο μυαλό μας".
  - ❑ Michalski ('86), "... είναι η δημιουργία ή η αλλαγή της αναπαράστασης των εμπειριών".
- ❖ Για τα συστήματα που ανήκουν στην συμβολική TN, η μάθηση προσδιορίζεται ως πρόσκτηση επιπλέον γνώσης, που επιφέρει μεταβολές στην υπάρχουσα γνώση.
  - ❑ Τα τεχνητά νευρωνικά δίκτυα (που ανήκουν στην μη συμβολική TN) έχουν δυνατότητα μάθησης μετασχηματίζοντας την εσωτερική τους δομή, παρά καταχωρώντας κατάλληλα αναπαριστάμενη γνώση.



University  
of  
East London



METROPOLITAN  
COLLEGE

# Μηχανική μάθηση - Ορισμός

- ❖ Ο άνθρωπος προσπαθεί να κατανοήσει το περιβάλλον του παρατηρώντας το και δημιουργώντας μια απλοποιημένη (αφαιρετική) εκδοχή του που ονομάζεται **μοντέλο (model)**.
  - Η δημιουργία ενός τέτοιου μοντέλου, ονομάζεται **επαγωγική μάθηση (inductive learning)** ενώ η διαδικασία γενικότερα ονομάζεται **επαγωγή (induction)**.
- ❖ Επιπλέον ο άνθρωπος έχει τη δυνατότητα να οργανώνει και να συσχετίζει τις εμπειρίες και τις παραστάσεις του δημιουργώντας νέες δομές που ονομάζονται **πρότυπα (patterns)**.

**Η δημιουργία μοντέλων ή προτύπων από ένα σύνολο δεδομένων, από ένα υπολογιστικό σύστημα, ονομάζεται μηχανική μάθηση (machine learning).**

- ❖ Διάφοροι ορισμοί:
  - Carbonell (1987), "... η μελέτη υπολογιστικών μεθόδων για την απόκτηση νέας γνώσης, νέων δεξιοτήτων και νέων τρόπων οργάνωσης της υπάρχουσας γνώσης".
  - Mitchell (1997), "Ενα πρόγραμμα υπολογιστή θεωρείται ότι μαθαίνει από την εμπειρία E σε σχέση με μια κατηγορία εργασιών T και μια μετρική απόδοσης P, αν η απόδοση του σε εργασίες της T, όπως μετριούνται από την P, βελτιώνονται με την εμπειρία E".
  - Witten & Frank (2000), "Κάτι μαθαίνει όταν αλλάζει τη συμπεριφορά του κατά τέτοιο τρόπο ώστε να αποδίδει καλύτερα στο μέλλον".



University  
of  
East London



# Μηχανική μάθηση - Είδη

- ❖ Έχουν αναπτυχθεί πολλές τεχνικές μηχανικής μάθησης που χρησιμοποιούνται ανάλογα με τη φύση του προβλήματος και εμπίπτουν σε ένα από τα παρακάτω δυο είδη:
  - μάθηση με επίβλεψη** (supervised learning) ή μάθηση με παραδείγματα (learning from examples),
  - μάθηση χωρίς επίβλεψη** (unsupervised learning) ή μάθηση από παρατήρηση (learning from observation).
- ❖ Στη μάθηση με επίβλεψη το σύστημα καλείται να "μάθει" μια έννοια ή συνάρτηση από ένα σύνολο δεδομένων, η οποία αποτελεί περιγραφή ενός μοντέλου.
- ❖ Στη μάθηση χωρίς επίβλεψη το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας **πρότυπα**, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι.

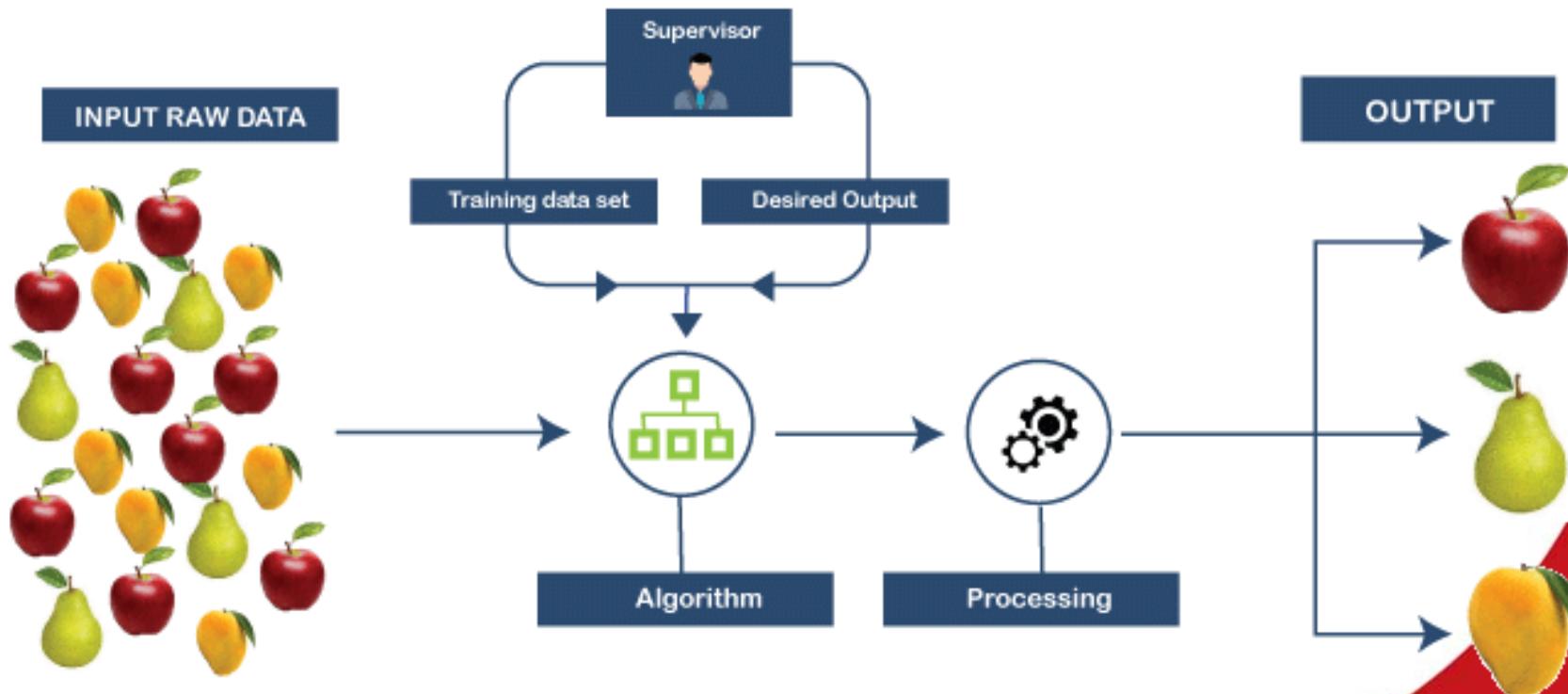


University  
of  
East London



METROPOLITAN  
COLLEGE

## SUPERVISED LEARNING

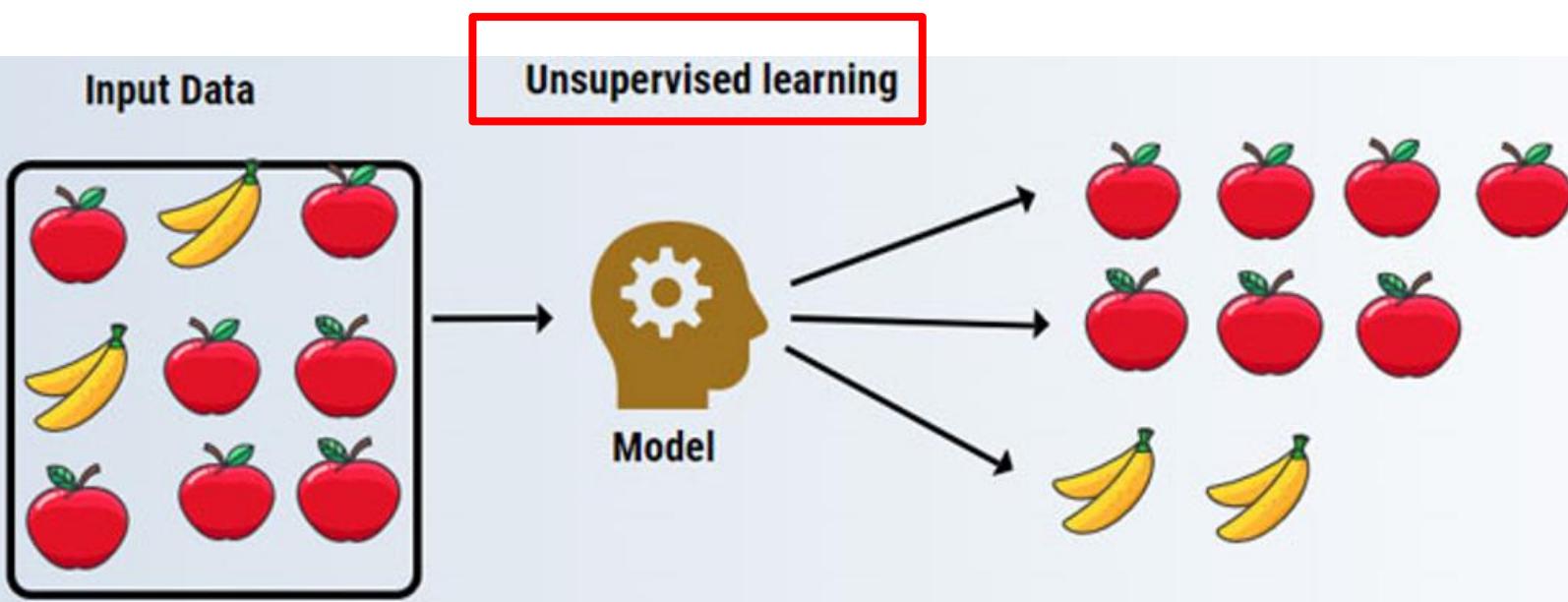


University  
of  
East London



METROPOLITAN  
COLLEGE

# Μηχανική μάθηση – Χωρίς επίβλεψη



University  
of  
East London



METROPOLITAN  
COLLEGE

# 1) Μάθηση με επίβλεψη

- ❖ Στη μάθηση με επίβλεψη το σύστημα πρέπει να "μάθει" επαγωγικά μια συνάρτηση που ονομάζεται **συνάρτηση στόχος** (**target function**) και αποτελεί έκφραση του μοντέλου που περιγράφει τα δεδομένα.
  - Η συνάρτηση στόχος χρησιμοποιείται για την πρόβλεψη της τιμής μιας μεταβλητής, που ονομάζεται **εξαρτημένη μεταβλητή** ή **μεταβλητή εξόδου**, βάσει των τιμών ενός συνόλου μεταβλητών, που ονομάζονται **ανεξάρτητες μεταβλητές** ή **μεταβλητές εισόδου** ή **χαρακτηριστικά**.

- ❖ Η επαγωγική μάθηση στηρίζεται στην "**υπόθεση επαγωγικής μάθησης**" (**inductive learning hypothesis**), σύμφωνα με την οποία:

Κάθε υπόθεση  $h$  που προσεγγίζει καλά τη συνάρτηση στόχο για ένα αρκετά μεγάλο σύνολο παραδειγμάτων, θα προσεγγίζει το ίδιο καλά τη συνάρτηση στόχο και για περιπτώσεις που δεν έχει εξετάσει.

- ❖ Στην μάθηση με επίβλεψη διακρίνονται δυο είδη προβλημάτων (learning tasks), τα προβλήματα ταξινόμησης και τα προβλήματα παρεμβολής.

- Η **ταξινόμηση<sup>1</sup>** (**classification**) αφορά στη δημιουργία μοντέλων πρόβλεψης διακριτών τάξεων (κλάσεων/κατηγοριών) (π.χ. ομάδα αίματος).
  - Η **παρεμβολή** (**regression**) αφορά στη δημιουργία μοντέλων πρόβλεψης αριθμητικών τιμών (π.χ. πρόβλεψη ισοτιμίας νομισμάτων ή τιμής μετοχής).

**Παρεμβολή ή Παλινδρόμηση**



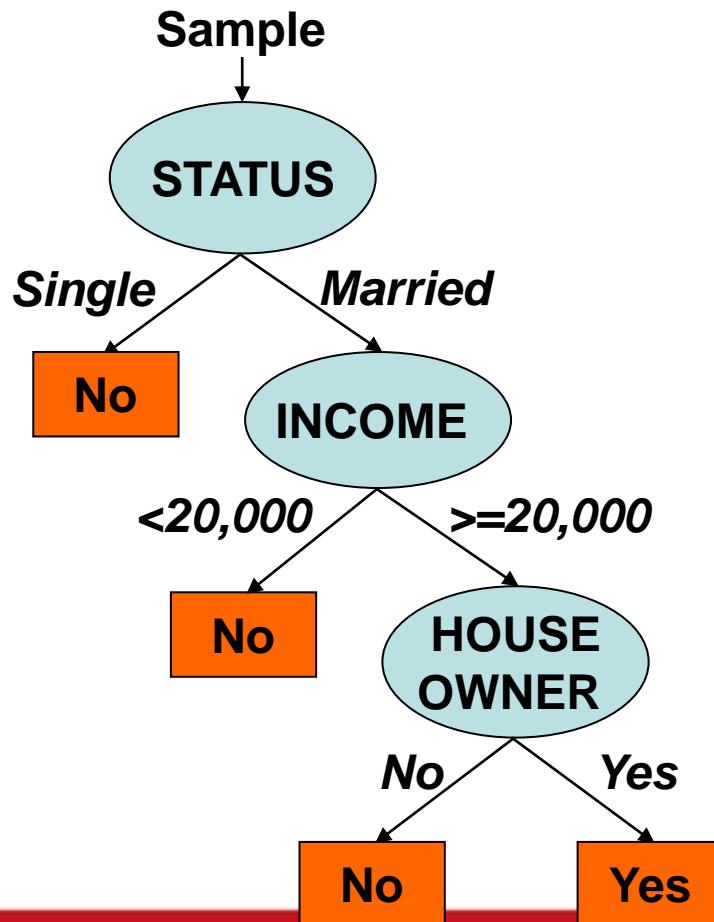
University  
of  
East London



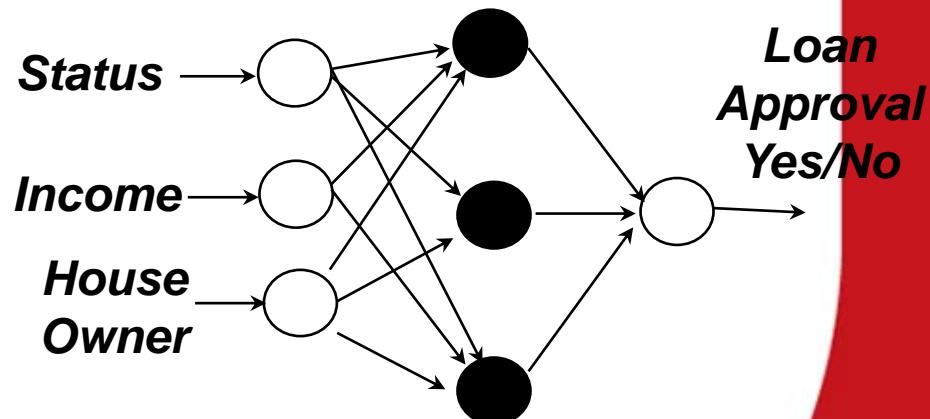
METROPOLITAN  
COLLEGE

# Παράδειγμα μάθησης με επίβλεψη

## Decision Trees



## Artificial Neural Networks



# Decision Trees

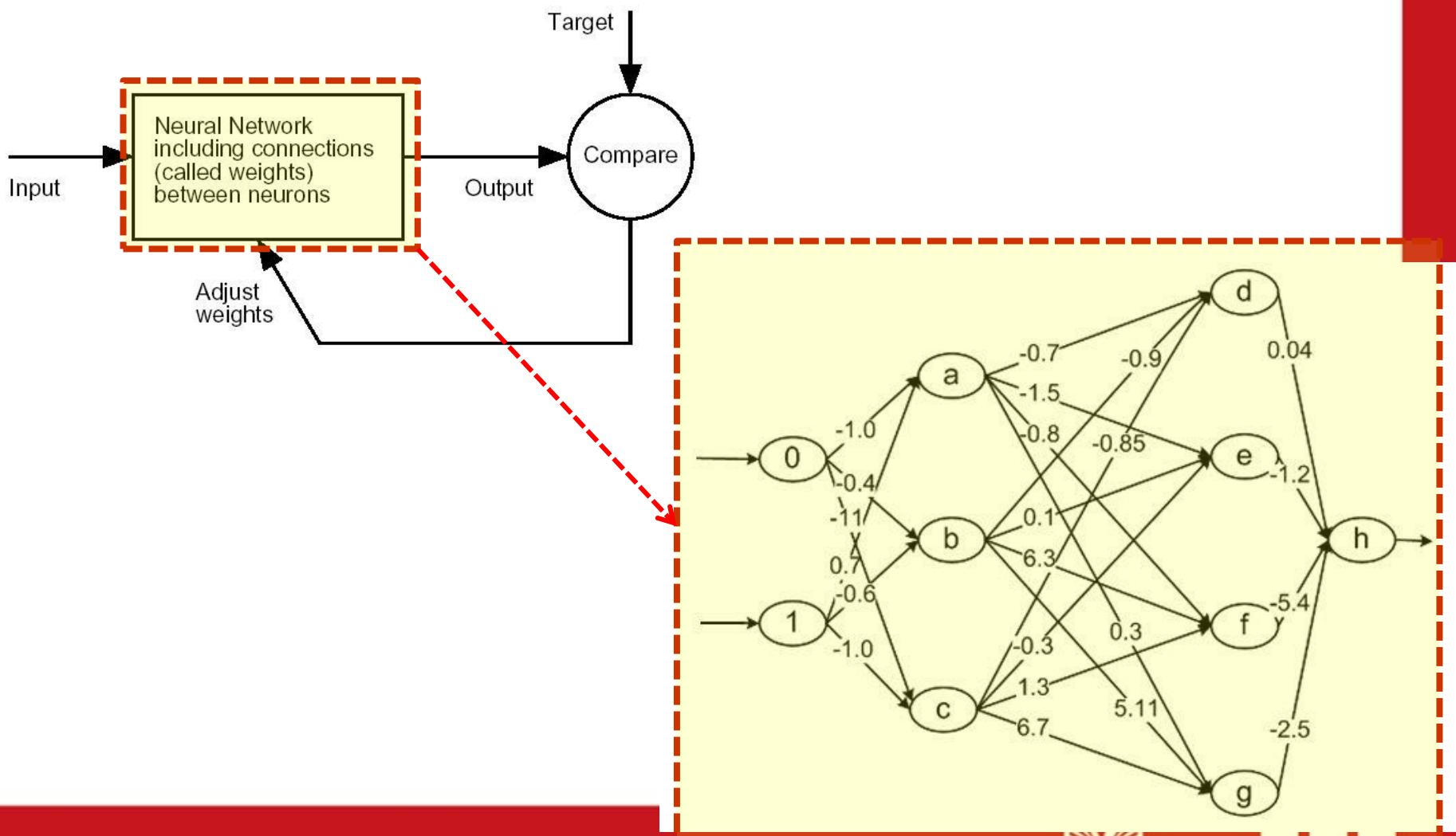


University of  
East London



METROPOLITAN  
COLLEGE

# Νευρωνικά δίκτυα: Supervised εκμάθηση



## 2) Μάθηση χωρίς επίβλεψη

- ❖ Στη μάθηση χωρίς επίβλεψη το σύστημα έχει στόχο να ανακαλύψει συσχετίσεις και ομάδες από τα δεδομένα, βασιζόμενο μόνο στις ιδιότητές τους.
- ❖ Σαν αποτέλεσμα προκύπτουν πρότυπα (περιγραφές), κάθε ένα από τα οποία περιγράφει ένα μέρος από τα δεδομένα.
- ❖ Παραδείγματα προτύπων πληροφόρησης είναι
  - ❑ οι κανόνες συσχέτισης (association rules) και
  - ❑ οι ομάδες (clusters), οι οποίες προκύπτουν από τη διαδικασία της ομαδοποίησης (clustering).



University  
of  
East London

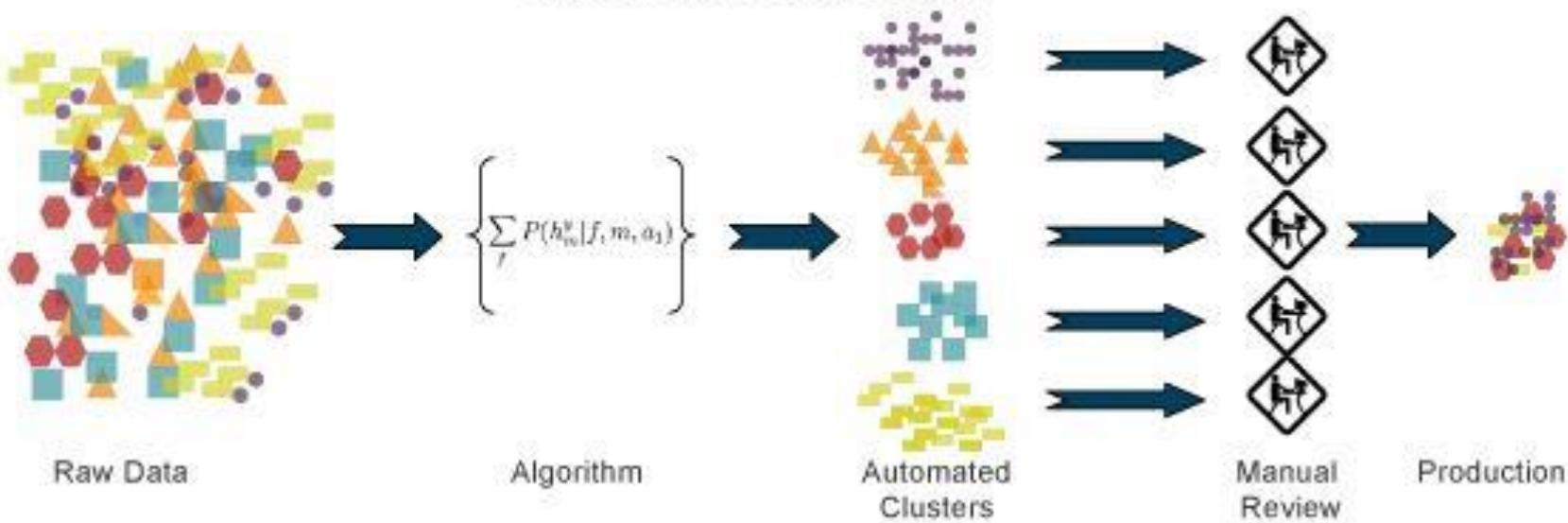


METROPOLITAN  
COLLEGE

# Μηχανική μάθηση - Εισαγωγή

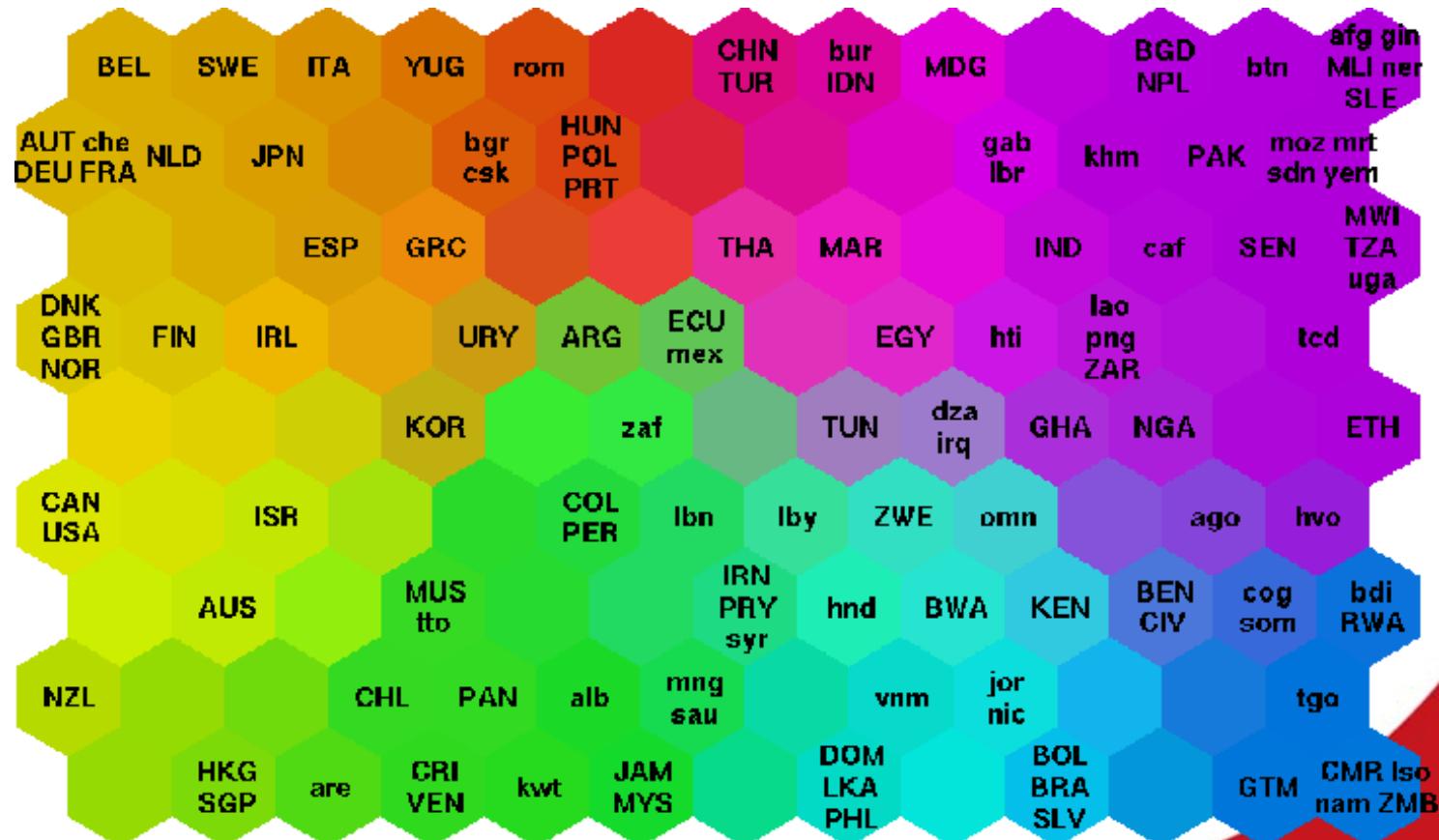
## UNSUPERVISED LEARNING

High reliance on algorithm for raw data, large expenditure on manual review for review for relevance and coding



# Νευρωνικά Δίκτυα: Unsupervised εκμάθηση

## SOM (self organizing maps)



'Χάρτης φτώχιας' βασιζόμενος σε 39 χαρακτηριστικά (από World Bank statistics (1992))



University of  
East London



METROPOLITAN  
COLLEGE

# Data Mining (and Business Intelligence)



University of  
East London



METROPOLITAN  
COLLEGE

# Επιχειρηματική Ευφυία (Business Intelligence) και Επιχειρηματική Ανάλυση (Business Analytics)

## • Επιχειρηματική Ευφυία (BI):

- Είναι ένας όρος «ομπρέλα» που περιλαμβάνει τις **αρχιτεκτονικές, διεργασίες, τεχνολογίες, εργαλεία, βάσεις δεδομένων, εφαρμογές και μεθοδολογίες που απαιτούνται για να μετασχηματίσουν τα δεδομένα σε πληροφορία, την πληροφορία σε αποφάσεις και τις αποφάσεις σε δράσεις.**
- Δίνει τη δυνατότητα σε ανθρώπους σε όλα τα επίπεδα ενός οργανισμού να έχουν αλληλεπιδραστική πρόσβαση **σε πραγματικό χρόνο** στην ανάλυση και διαχείριση δεδομένων και πληροφοριών για τη διοίκηση του οργανισμού, τη βελτίωση της απόδοσης του και την ανακάλυψη νέων ευκαιριών.

**Λήψη αποφάσεων (Decision-making): Η παροχή της κατάλληλης πληροφορίας, στην κατάλληλη χρονική στιγμή, στον κατάλληλο άνθρωπο που χρειάζεται την πληροφορία για να ολοκληρώσει μια εργασία ή για να πάρει μια απόφαση.**

Η επιχειρηματική ανάλυση είναι ένα βασικό δομικό στοιχείο της επιχειρηματικής ευφυίας.



University of  
East London



METROPOLITAN  
COLLEGE

# Επιχειρηματική Ευφυία

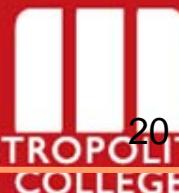
- Η λήψη αποφάσεων σχετίζεται με οποιαδήποτε λειτουργία μιας επιχείρησης όπου απαιτούνται **τεκμηριωμένες** αποφάσεις για τη **βελτιστοποίηση** της απόδοσης σε κάποιον τομέα.
- Απαντά σε κρίσιμες ερωτήσεις που αφορούν έναν οργανισμό, όπως, π.χ. :
  1. Ποιοι είναι οι **καλύτεροι** και οι **χειρότεροι** προμηθευτές?
  2. Ποιοι είναι οι **καλύτεροι** και οι **χειρότεροι** πελάτες?
  3. Ποιοι πελάτες είναι πιθανόν να πάνε σε ανταγωνιστή?
  4. Ποια προϊόντα **συνεισφέρουν** περισσότερο στο κέρδος?
  5. Πως μπορεί η επιχείρηση να γίνει πιο **επικερδής**?
  6. Γιατί κάποια **τμήματα** φέρνουν περισσότερο κέρδος από κάποια άλλα?
  7. Πως μπορεί να βελτιωθεί η παραγωγικότητα?
  8. .....

## Πληροφορία υψηλότερης προστιθέμενης αξίας

- “For example, somewhere in the more than 15+ terabytes of data in the Verizon customer database is evidence that
  - some customers are about to change cell phone companies
  - a different pricing scheme would generate more revenue
  - the district manager in, say, Chicago, is doing a better job than any other district manager.
- All of that information is in there. The question is: How can Verizon get it out?”



University of  
East London



20

# Επιχειρηματική Ευφυία – περιοχές συχνής εφαρμογής

## Κλάδος

Χρηματοοικονομικός

Ασφαλιστικός

Τηλεπικοινωνίες

Μεταφορές

Καταναλωτικά Αγαθά

Ηλεκτροπαραγωγικός

Μάρκετινγκ

## Εφαρμογή

Ανάλυση Αγορών με Πιστωτική Κάρτα

Ανάλυση απάτης (Fraud Analysis)

Ανάλυση κλήσεων (Call record analysis)

Διαχείριση Logistics

Ανάλυση προωθήσεων

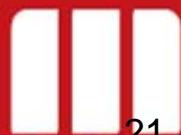
Ανάλυση χρήσης ρεύματος

Ανάλυση MBA (market basket analysis)

Ανάλυση RFM (recency, frequency, money)



University of  
East London



METROPOLITAN  
COLLEGE

Οι γενικές κατηγορίες εφαρμογών είναι:

- ❖ **Spreadsheets**
- ❖ **Reporting and querying software**: εφαρμογές που εξορύσσουν, ταξινομούν, κάνουν περίληψη και παρουσιάζουν επιλεγμένα δεδομενα.
- ❖ **Online analytical processing (OLAP)**
- ❖ **Digital dashboards**
- ❖ **Data mining**
- ❖ **Process visualization**
- ❖ **Data warehousing**

Εκτός από τα spreadsheets, αυτά τα εργαλεία δίνονται ως αυτόνομες εφαρμογές, σουίτες εφαρμογών, **συστατικά ERP συστημάτων**, ή ως συστατικά λογισμικού που στοχεύει σε εξειδικευμένους κλάδους. Πολλές φορές τα εργαλεία είναι ενσωματωμένα σε data warehouses.

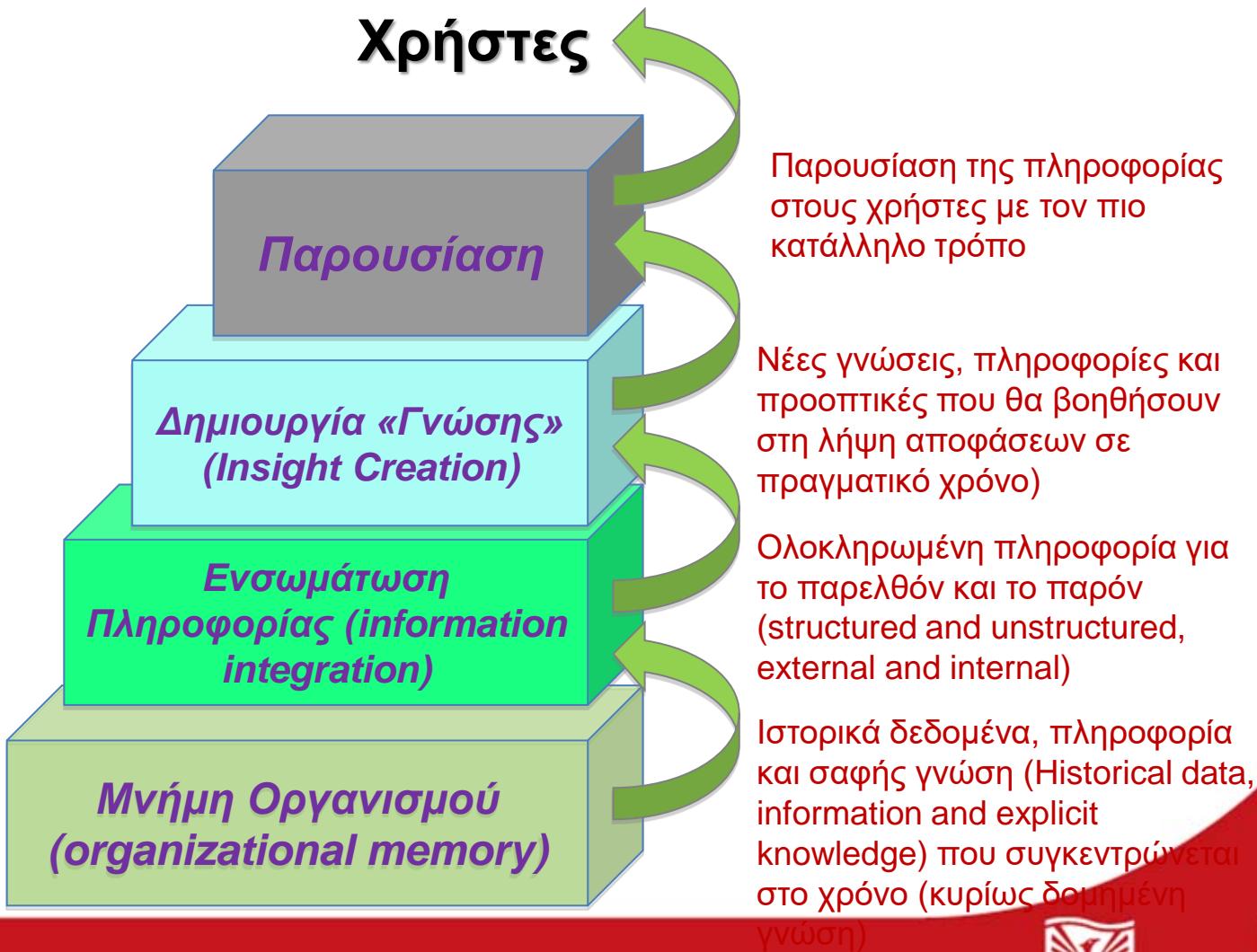


University  
of  
East London

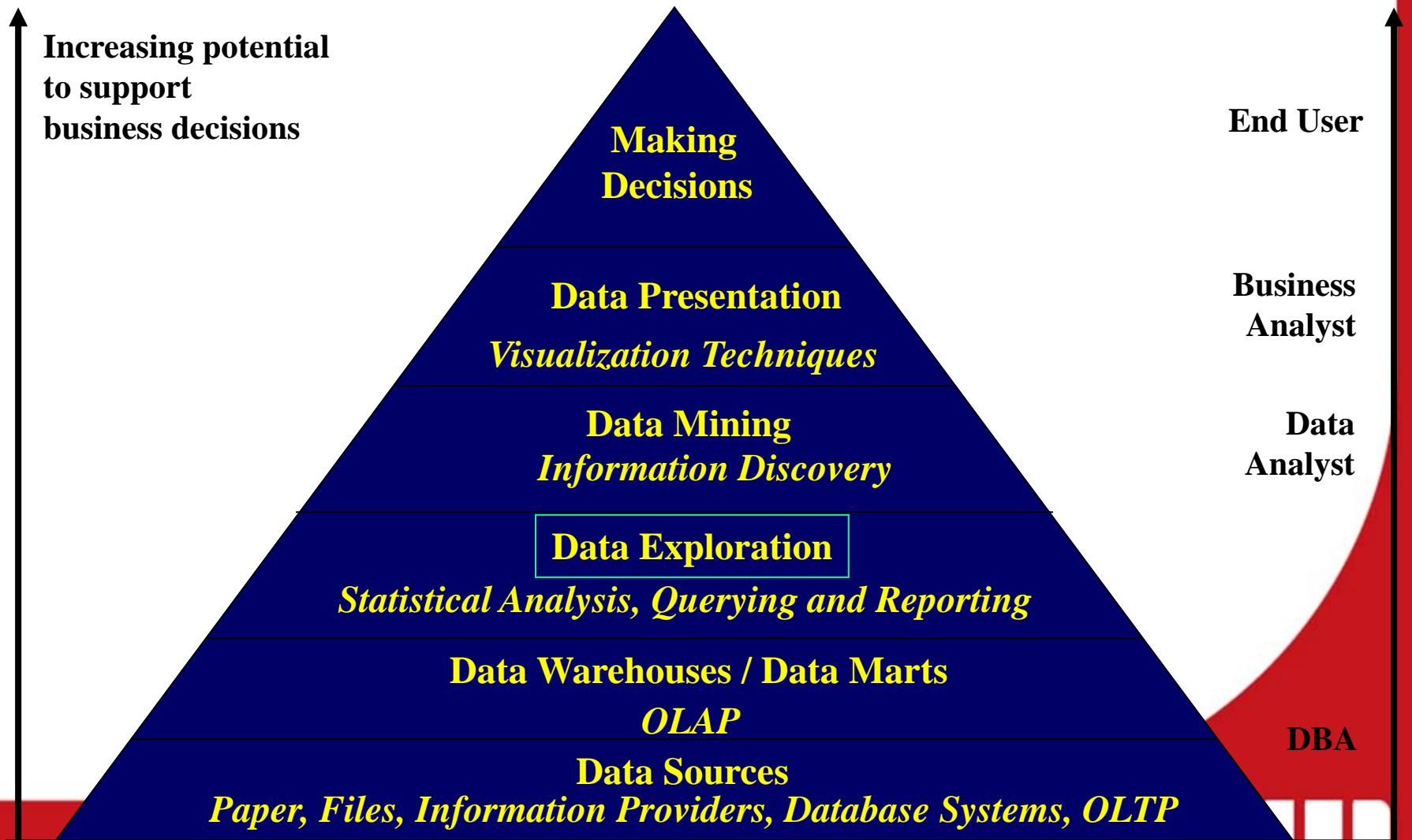


METROPOLITAN  
COLLEGE

# 4 κύρια χαρακτηριστικά της Επιχειρηματικής Ευφυΐας



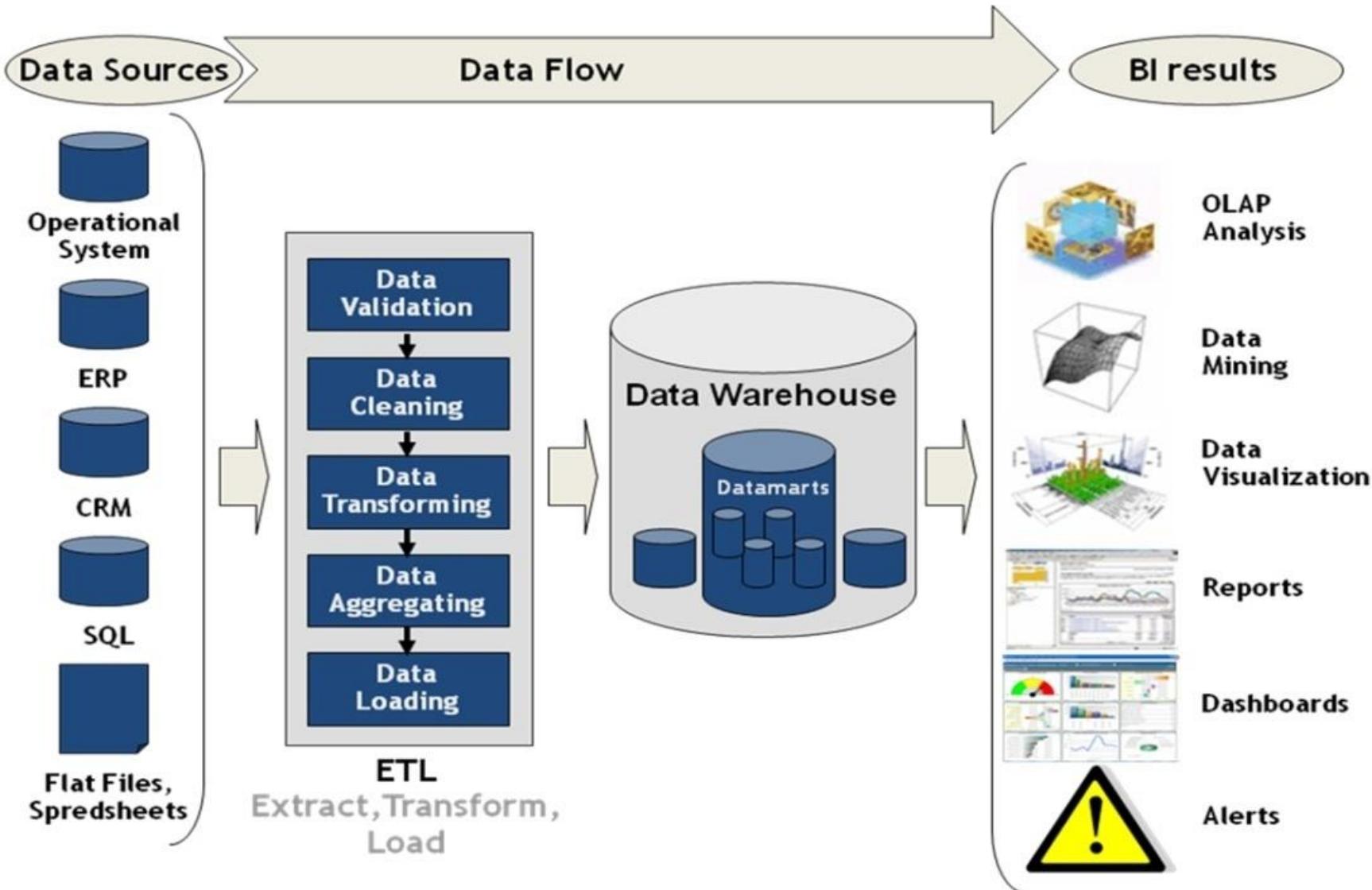
# Η πυραμίδα του Business Intelligence



University of  
East London

METROPOLITAN  
COLLEGE

# Συστήματα Επιχειρηματικής Ευφνίας



# Visualisation



University  
of  
East London



26

METROPOLITAN  
COLLEGE



Figure 1: Types of Data Visualizations

Streamgraph



Force-directed graphs



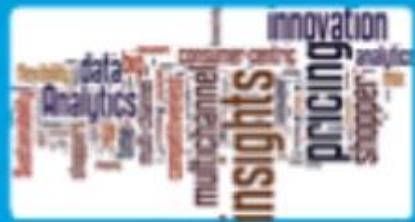
Tree maps



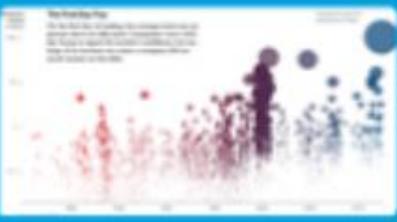
Sunburst



Word Tag Cloud



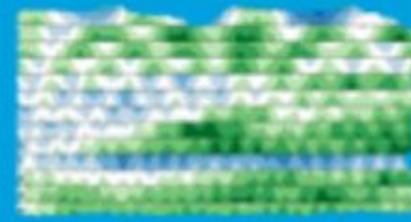
Bubble Chart



Many Eye Bubble Chart



Time Series Analysis



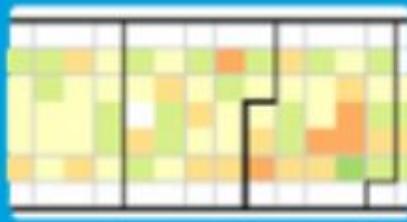
Geospatial



Parallel Chord



Calendar View



Heat Maps



Source: D3js.org



University  
of  
East London

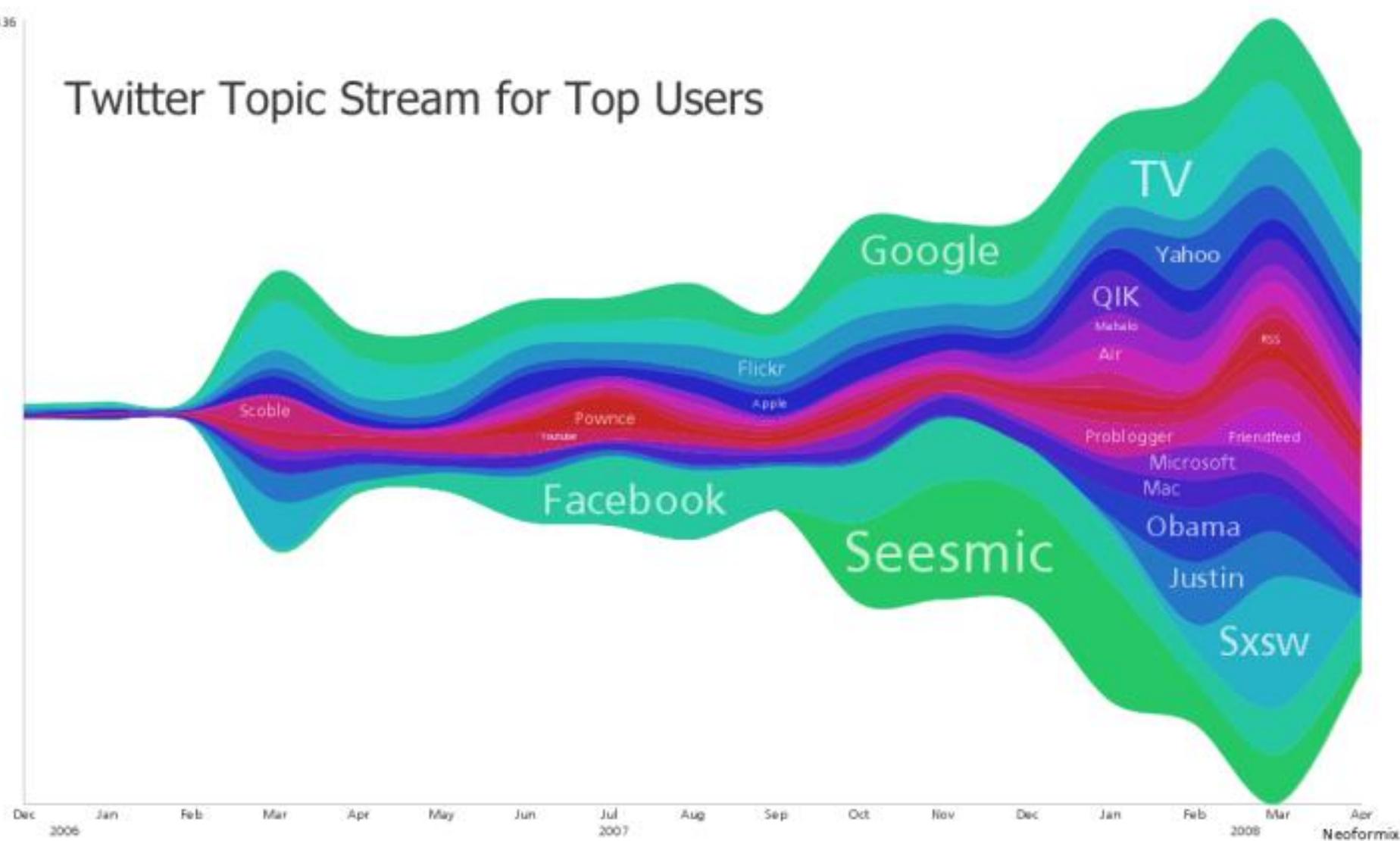
27

METROPOLITAN  
COLLEGE

# Visualisation – Οπτικοποίηση, Streamgraph



Twitter Topic Stream for Top Users



Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr  
2006 2007 2008

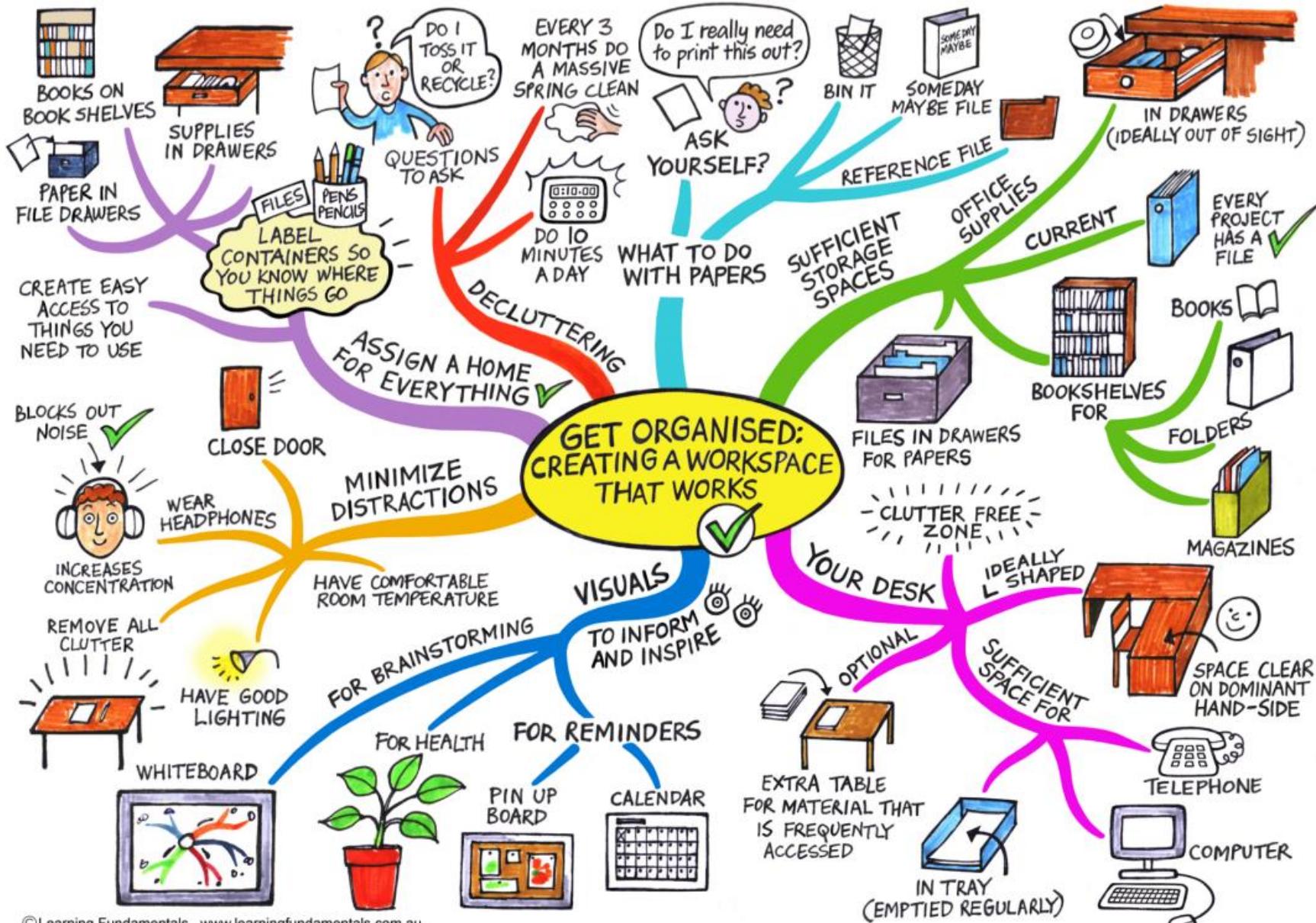


University of  
East London

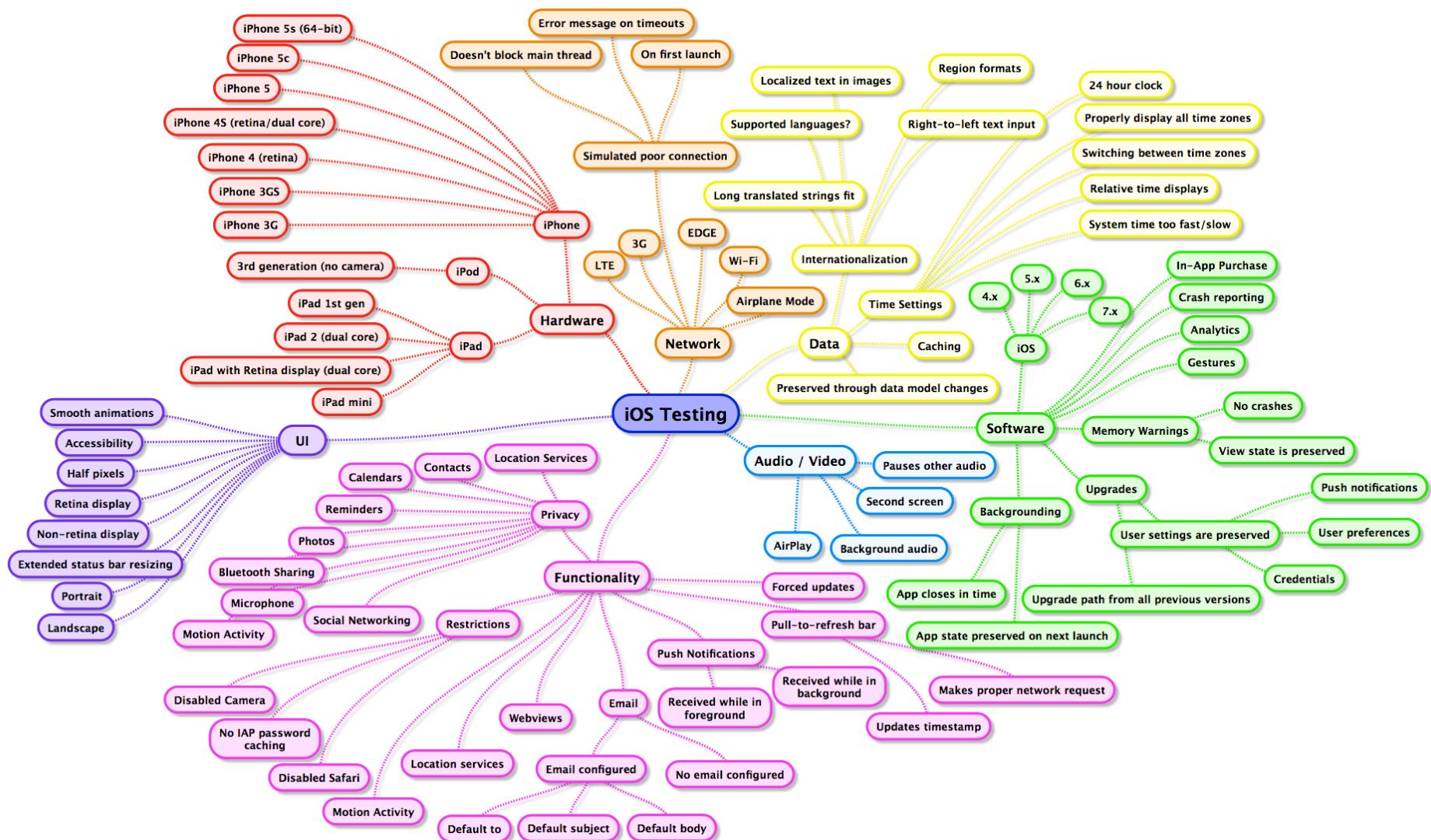
28

METROPOLITAN  
COLLEGE

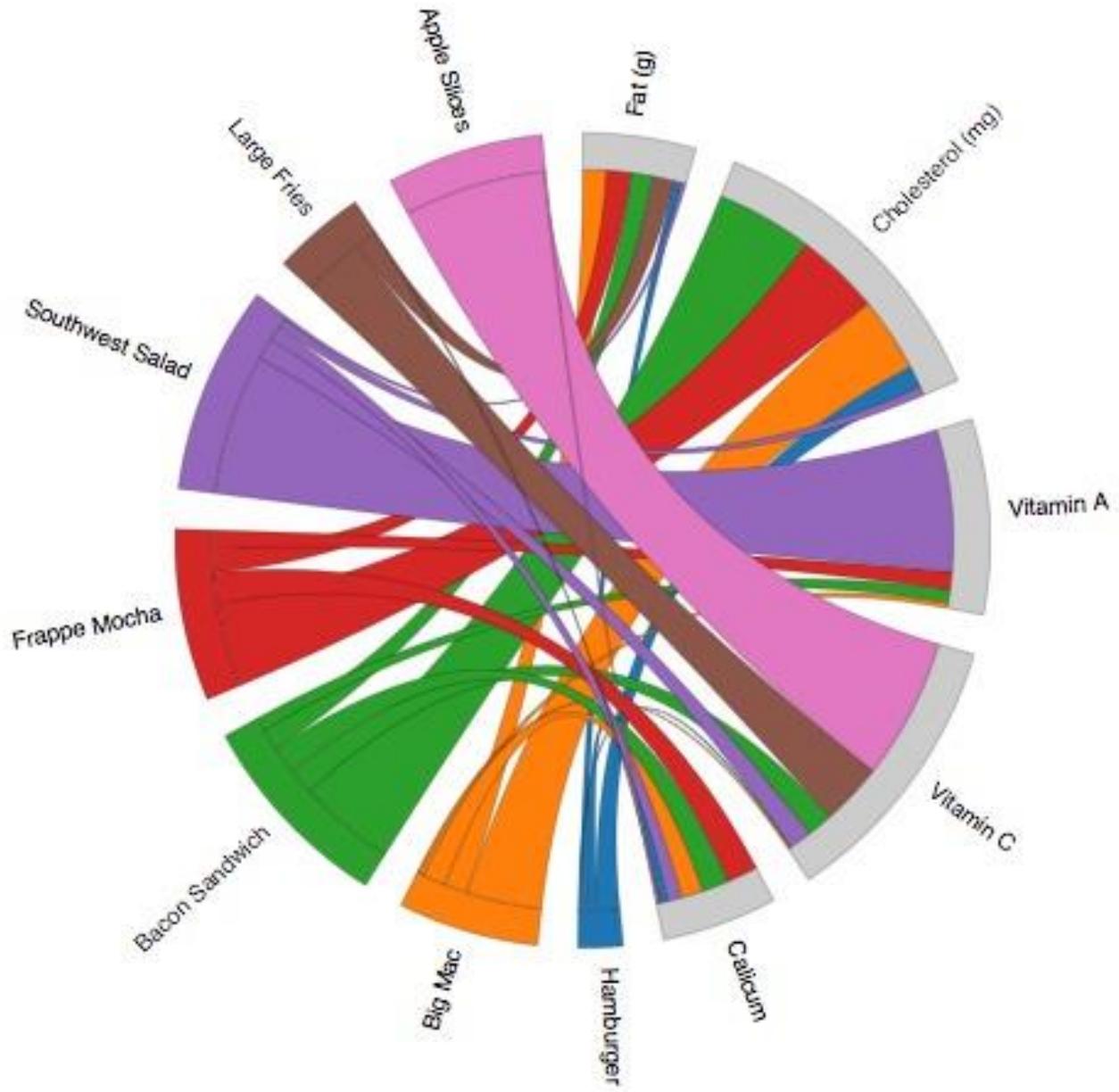
# Visualisation – Οπτικοποίηση, Mindmaps



# Visualisation – Οπτικοποίηση, Mindmaps



# Visualisation – Οπτικοποίηση, Sankey diagrams



# Visualisation – Οπτικοποίηση, Animation



## Out of Sight, Out of Mind.

ATTACKS VICTIMS NEWS INFO SHARE

CHILDREN CIVILIAN  
**190** 5.7% **534** 16%

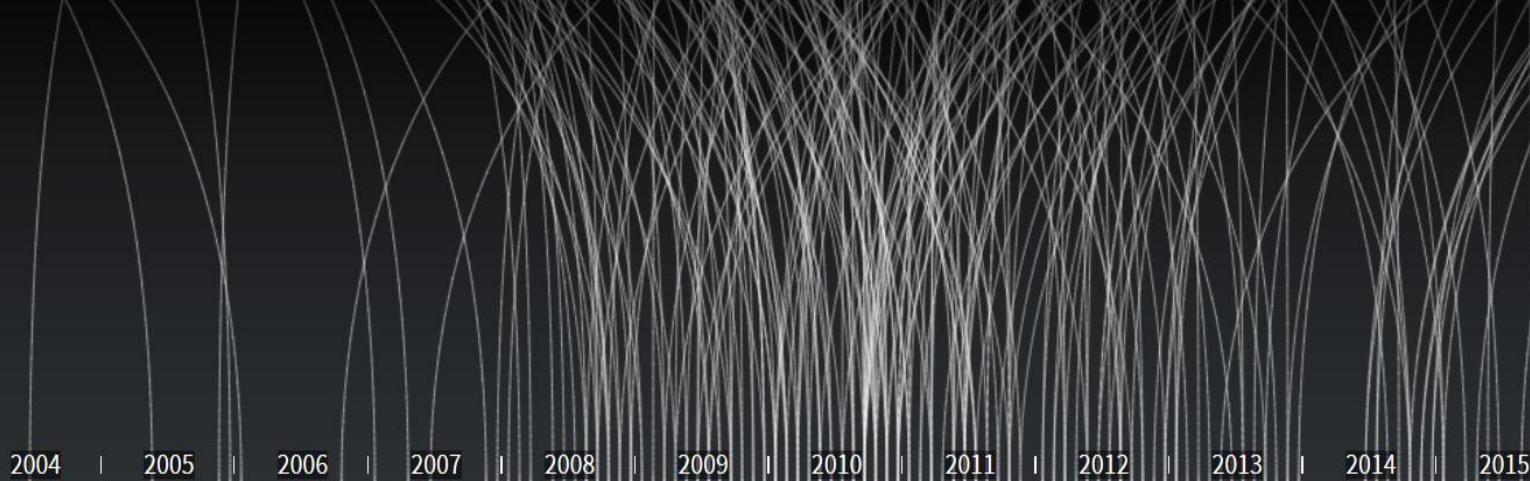
HIGH PROFILE  
**52** 1.6%

ESTIMATED TOTAL FATALITIES **3341**

EN FR

OTHER  
76.8% **2565**

PAKISTAN



2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015

RECENT NEWS ABOUT DRONES

<http://blog.visme.co/examples-data-visualizations/>

# Visualisation - Οπτικοποίηση



SLEEP    CREATIVE WORK    DAY JOB/ADMIN    FOOD/LEISURE    EXERCISE    OTHER

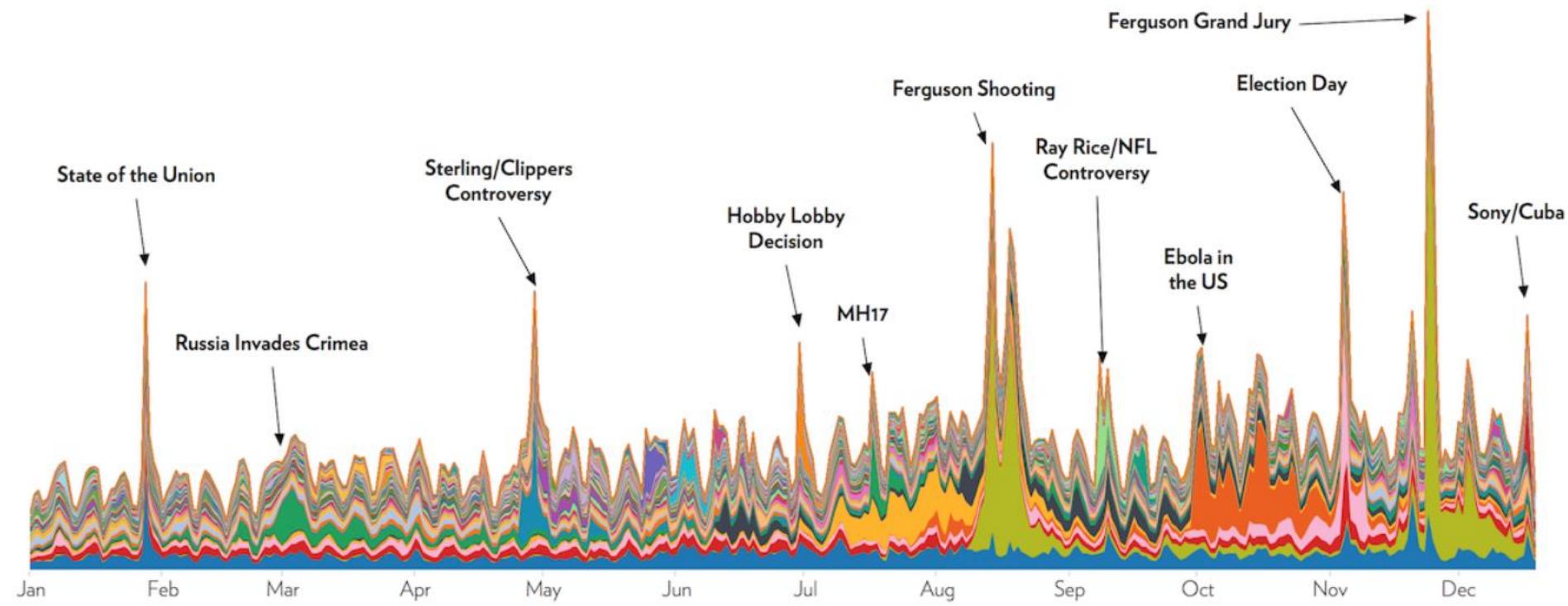
12 1 2 3 4 5 6 7 8 9 10 11 12  
AM → PM →



# Visualisation – Οπτικοποίηση, Stream Graph



What America talked about in 2014, as viewed through 184.5 million Twitter mentions.



<http://blog.visme.co/examples-data-visualizations/>



University  
of  
East London



METROPOLITAN  
COLLEGE

# Visualisation – Οπτικοποίηση, Tree Maps



Total: \$589M

Raw Cotton

20.52%

Coconuts, Brazil Nuts,  
and Cashews

12.12%

Rice

Hot-  
rolled  
Iron  
Bars

Other  
Iron  
Bars

Scrap  
Iron

Scrap  
Copper

Raw  
Iron  
Bars

Refined  
Petroleum

7.51%

Raw Sugar

Rolled Tobacco

1.51%

Other Vegetable  
Residues

Poultry  
Meat

11.38%

Other  
Vegetable  
Oils

Palm  
Oil

Rough  
Wood

Sawn  
Wood

2.73%

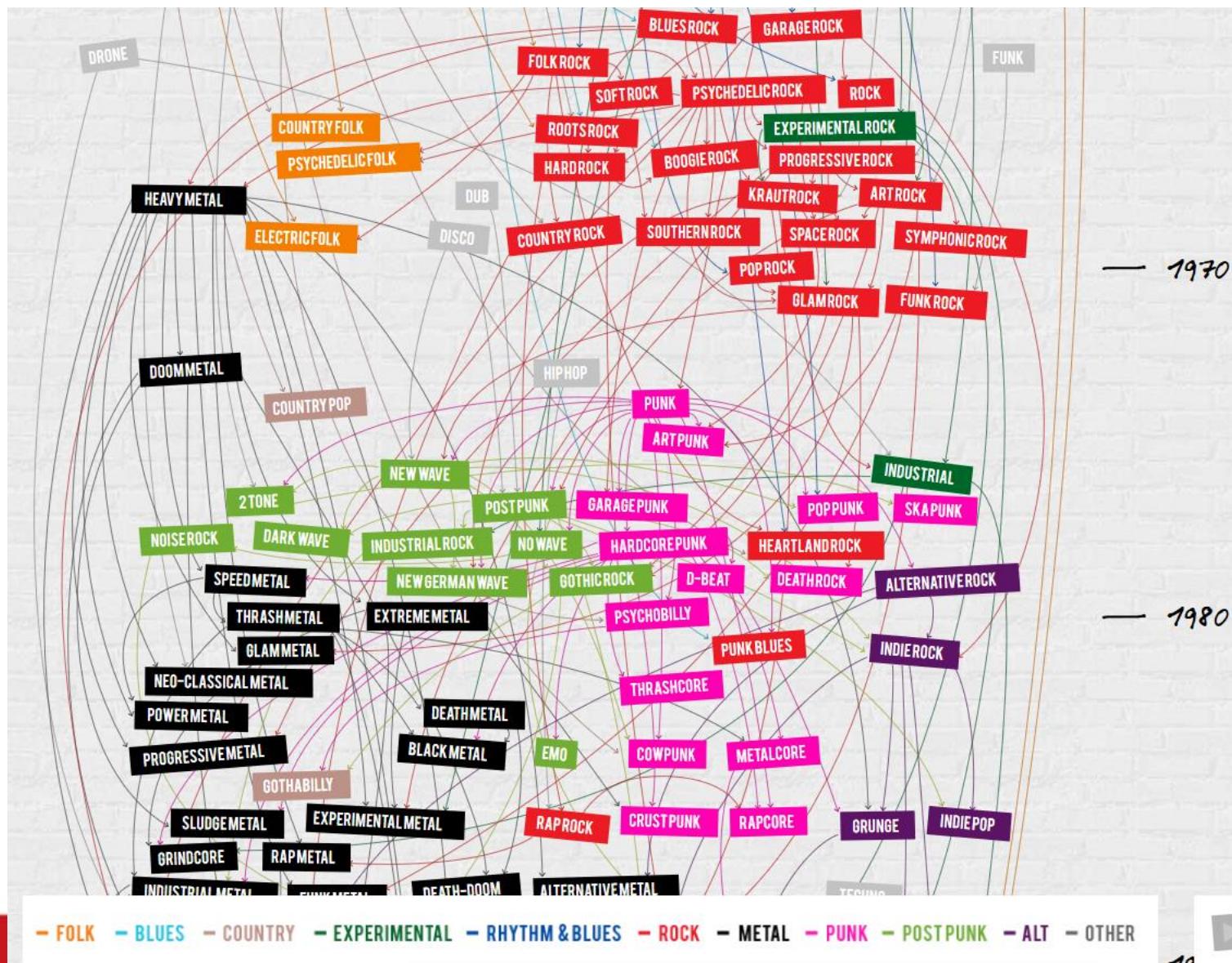
2.7%

2.18%

Gold



# Visualisation – Οπτικοποίηση, Semantic Map



— FOLK — BLUES — COUNTRY — EXPERIMENTAL — RHYTHM & BLUES — ROCK — METAL — PUNK — POST PUNK — ALT — OTHER



**SUSTAINABLE DEVELOPMENT GOALS**  
17 GOALS TO TRANSFORM OUR WORLD

1 NO POVERTY	2 ZERO HUNGER	3 GOOD HEALTH AND WELL-BEING	4 QUALITY EDUCATION	5 GENDER EQUALITY	6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY	8 DECENT WORK AND ECONOMIC GROWTH	9 INDUSTRY, INNOVATION AND INFRASTRUCTURE	10 REDUCED INEQUALITIES	11 SUSTAINABLE CITIES AND COMMUNITIES	12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION	14 LIFE BELOW WATER	15 LIFE ON LAND	16 PEACE, JUSTICE AND STRONG INSTITUTIONS	17 PARTNERSHIPS FOR THE GOALS	SUSTAINABLE DEVELOPMENT GOAL 17



# Data Mining

## *Εξόρυξη Πληροφοριών*

## *Εξόρυξη Γνώσης*



University  
of  
East London



METROPOLITAN  
COLLEGE



- “Η τράπεζα ABC” θέλει να αποφασίσει ποια από τα πολλά τραπεζικά προϊόντα της θα προωθήσει και σε ποιους συγκεκριμένα από τους πελάτες της».
- “Η XYZ Internet Portal, Inc.” θέλει να βρει ποιοι από τους πελάτες της θα είναι πιο δεκτικοί σε στοχευμένες προσπάθειες να αυξηθεί η επισκεψιμότητα του portal.
- Το σουπερμάρκετ XY θέλει να βρει σε ποιους πελάτες του να κάνει συγκεκριμένες στοχευμένες προσφορές.
- Ποια είναι τα χαρακτηριστικά των αναγνωστών ενός περιοδικού αυτοκινήτων.
- .....



University  
of  
East London

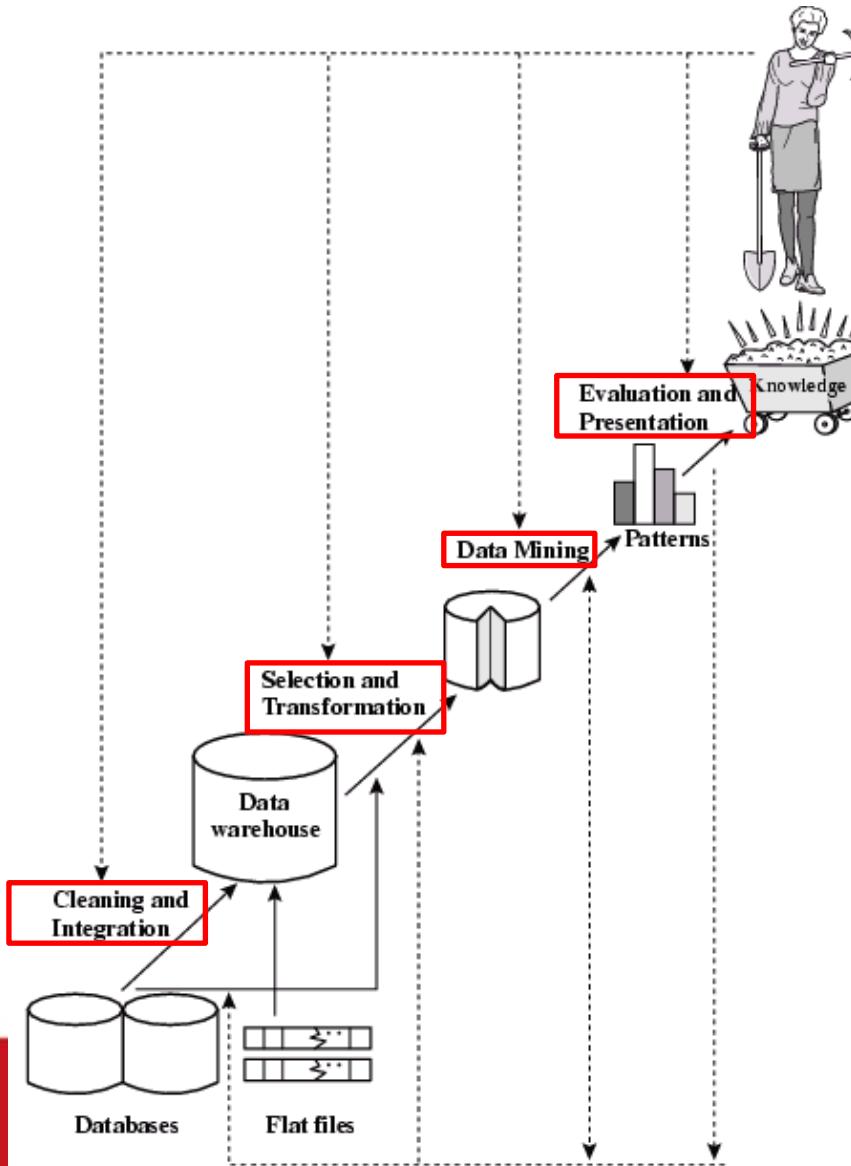


39

# Η πορεία προς την εξόρυξη γνώσης (Knowledge Discovery)



**Data mining:** Η κύρια δραστηριότητα για την εξόρυξη γνώσης





## 1. Προσδιορισμός Στόχου:

- Καθορισμός προβλήματος
- Προηγούμενη σχετική γνώση στην περιοχή και στόχοι της εφαρμογής

## 2. Δημιουργία του σετ δεδομένων (dataset):

Επιλογή των δεδομένων που θα χρησιμοποιηθούν

## 3. Προεπεξεργασία δεδομένων (Data preprocessing):

(μπορεί να χρειαστεί το 60%-80% της συνολικής προσπάθειας!)

- αφαίρεση θορύβου ή/και outliers
- στρατηγική διαχείρισης δεδομένων που λείπουν

## 4. Συμπύκνωση και μετασχηματισμός δεδομένων:

- Εύρεση χρήσιμων χαρακτηριστικών, μείωση της διάστασης των δεδομένων





## 5. Data Mining:

- Επιλογή λειτουργιών του data mining:
  - summarization, classification, regression, association, clustering.
- Επιλογή αλγορίθμων εξόρυξης:
  - μοντέλα και παράμετροι
- Αναζήτηση ενδιαφέροντων μοτίβων (patterns)

## 6. Παρουσίαση και Αξιολόγηση:

- visualisation, transformation, removing redundant patterns, etc.

## 7. Δράση (χρήση της ανακαλυφθήσας γνώσης):

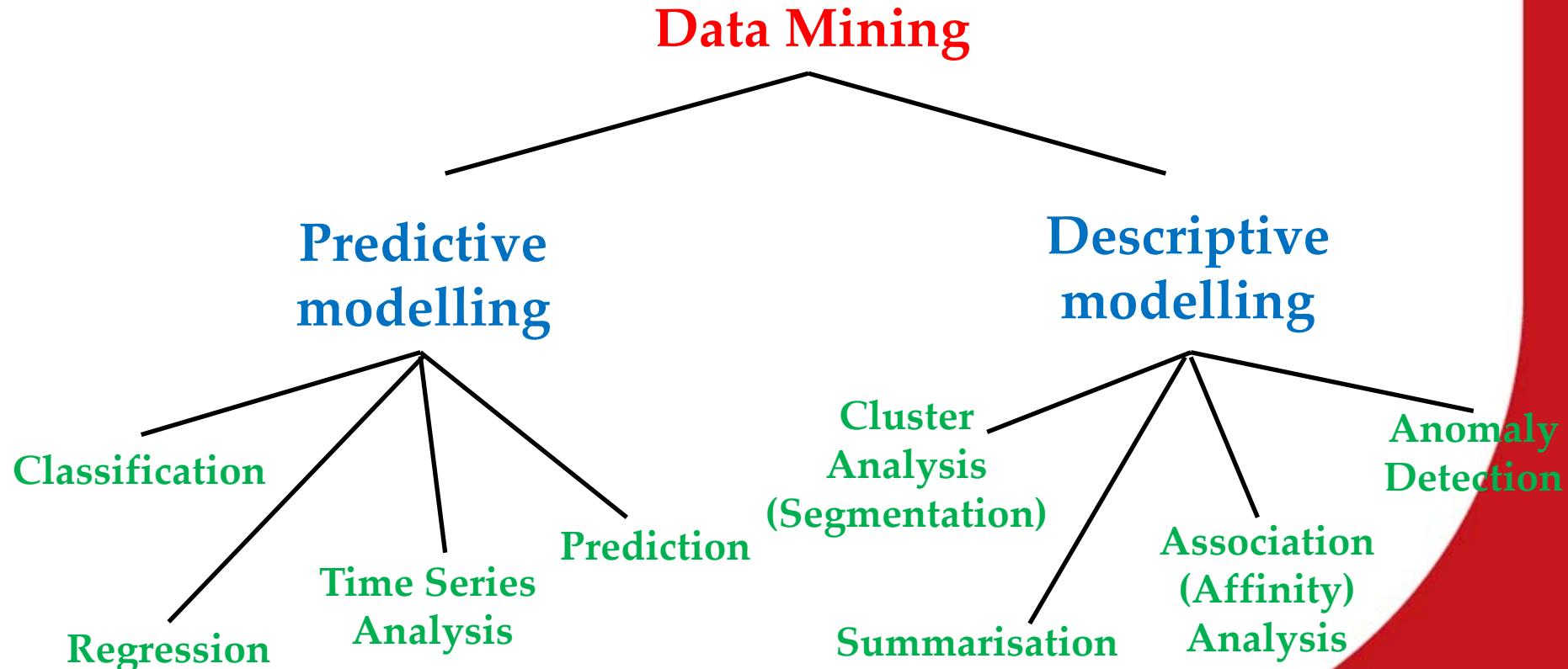
- ενσωμάτωση στο σύστημα
- τεκμηρίωση
- αναφορές στα ενδιαφερόμενα μέρη



University  
of  
East London



METROPOLITAN  
COLLEGE



# Predictive Data Mining



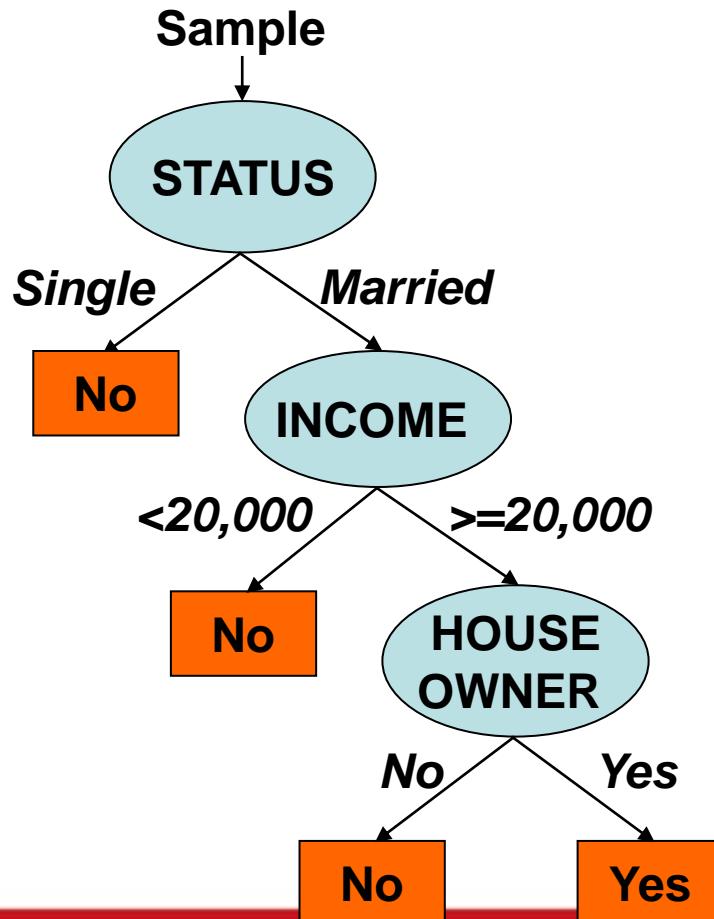
- ... εύρεση **αναλυτικών (ή μη) μοντέλων** που περιγράφουν ένα πρόβλημα και η χρήση αυτών των μοντέλων για πρόβλεψη της τιμής ενός συγκεκριμένου χαρακτηριστικού (target attribute) με βάση τις τιμές που έχουν άλλα χαρακτηριστικά (independent variables).
- **ΠΡΟΣΟΧΗ: Ο χρήστης/ερευνητής γνωρίζει τι θέλει να προβλεφθεί.**
- Χωρίζουμε τα δεδομένα που έχουμε στη διάθεσή μας σε **δύο είδη**:
  - τα δεδομένα με τα οποία βρίσκουμε το μοντέλο που περιγράφει τη σχέση τους (**training data**) και
  - αυτά με τα οποία ελέγχουμε αν το μοντέλο που βρήκαμε είναι σωστό (**testing data**)



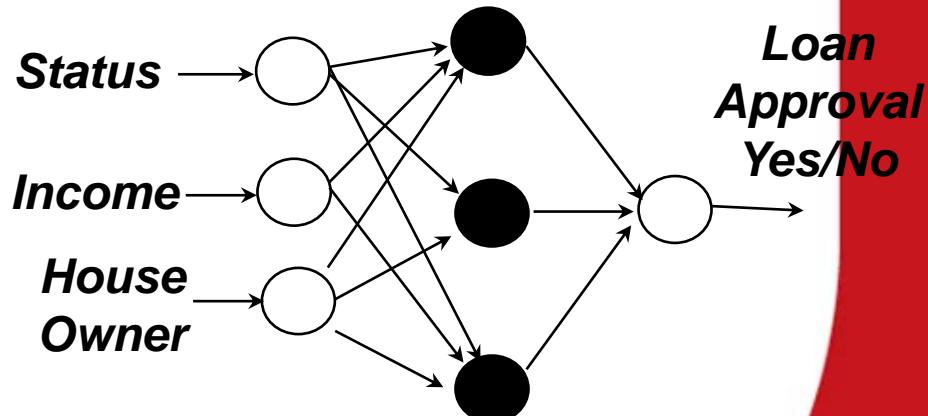
# Predictive Modelling: ένα παράδειγμα



## Decision Trees



## Artificial Neural Networks



University of  
East London



METROPOLITAN  
COLLEGE

# Data Mining συνέχεια, Descriptive DM...



- ... εύρεση **μοτίβων (patterns)** που χαρακτηρίζουν και περιγράφουν τις σχέσεις που υπάρχουν μεταξύ των δεδομένων (**τάσεις** (trends), **συσχετίσεις** (correlations), **σχηματισμοί** (clusters) και **ανωμαλίες** οπως outliers).
- **ΠΡΟΣΟΧΗ:** Ο χρήστης/ερευνητής καθορίζει ποιά από τα παραπάνω χαρακτηριστικά είναι σημαντικά

Βασικοί εκπρόσωποι του descriptive data mining:

- **Cluster analysis (Segmentation):** ομαδοποίηση παρόμοιων αντικειμένων, γεγονότων, ανθρώπων κλπ.
- **Association analysis (Affinity):** εύρεση πόσο συχνά συμβαίνουν πράγματα ταυτόχρονα
- **Anomaly (outlier) detection:** ανεύρεση μοτίβων σε ένα δεδομένο dataset που δεν ταιριάζουν με μια καθορισμένη φυσιολογική συμπεριφορά που ισχύει για τα υπόλοιπα δεδομένα.





τα είδη των πληροφοριών που μπορούν από παραχθούν από data-mining είναι

## 1. κατηγοριοποίηση (classification)

- ένταξη σε κατηγορίες
- π.χ. πελάτες με πιθανότητα να παύσουν να είναι πελάτες της εταιρείας

## 2. ομαδοποιήσεις (clustering)

- ομαδοποιήσεις χωρίς την εξαρχής ύπαρξη κατηγοριών
- π.χ. ομοειδείς πελάτες

## 3. προβλέψεις (forecasts)

- χρήση υπαρχουσών τιμών για την πρόβλεψη άλλων τιμών
- π.χ. πρόβλεψη πωλήσεων

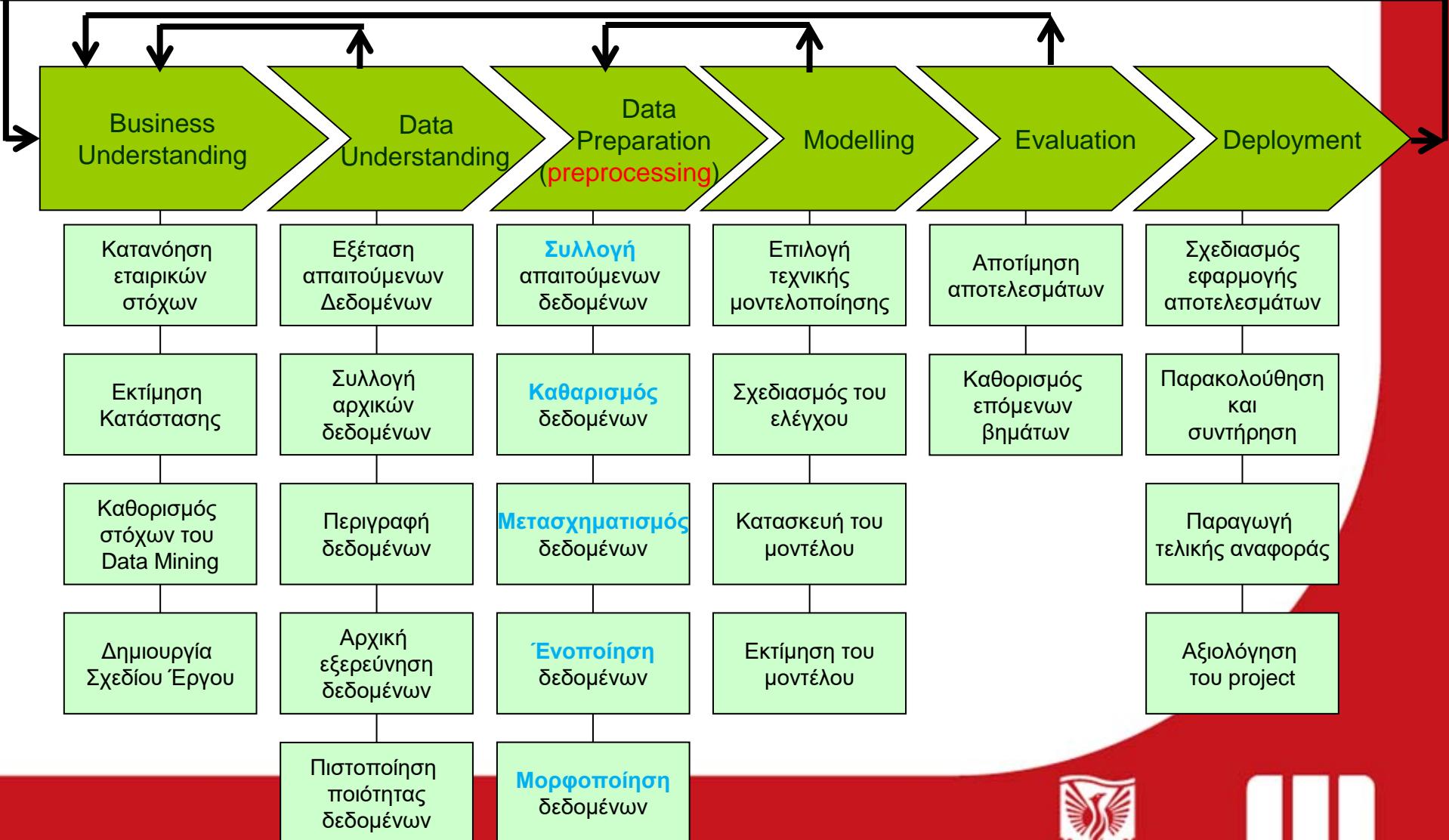
## 4. συσχετισμοί (associations)

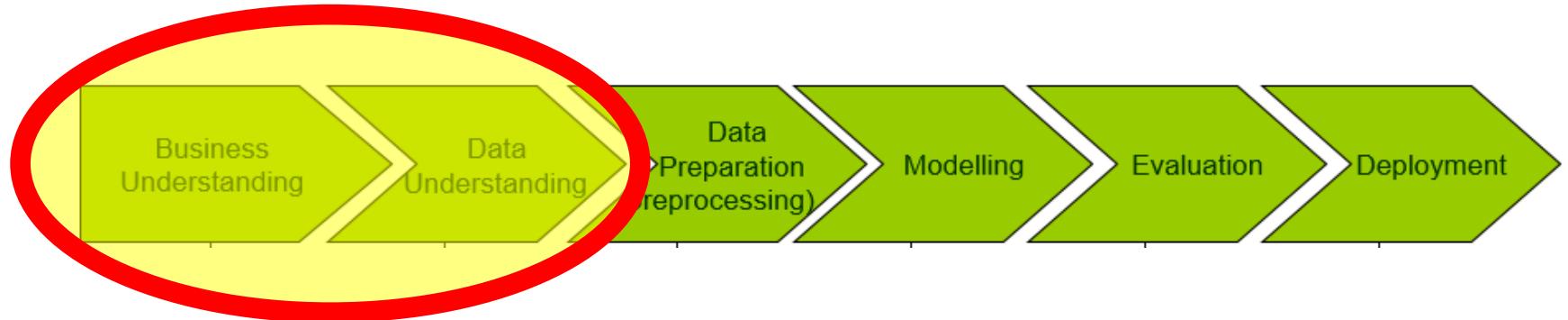
- συμβάντα που συνδέονται με το ίδιο γεγονός
- π.χ. προϊόντα αγοραζόμενα από κοινού την ίδια χρονική στιγμή

## 5. ακολουθίες (sequences)

- σύνδεση γεγονότων συνδεμένων χρονικά
- π.χ. προϊόντα αγοραζόμενα από κοινού σε άλλες χρονικές στιγμές

# Μεθοδολογία Data Mining





## Business και Data Understanding στην Εξόρυξη Δεδομένων



University of  
East London



49

METROPOLITAN  
COLLEGE

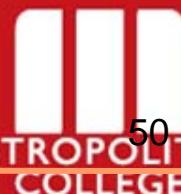
# Φάση 1: Business Understanding



- Κατανόηση των στόχων του έργου και των απαιτήσεων από μια επιχειρηματική οπτική.
- **Μετάφραση των παραπάνω στόχων και περιορισμών σε ένα πρόβλημα data mining – Πολύ σημαντικό!**
  - Ένας επιχειρηματικός στόχος χρησιμοποιεί **επιχειρηματική ορολογία**
  - Ένας data mining στόχος χρησιμοποιεί **τεχνικούς όρους**
- **Παράδειγμα:**
  - Επιχειρηματικός στόχος: **“Να αυξηθούν οι πωλήσεις των προϊόντων του καταλόγου στους υπάρχοντες πελάτες.”**
  - Data mining στόχος: **“Να προβλεφθούν οι πωλήσεις σε κάποιον πελάτη με δεδομένα τις αγορές που έκανε τα τελευταία τρία χρόνια, δημογραφικά χαρακτηριστικά (ηλικία, μισθό, πόλη, κλπ.) και την τιμή του αντικειμένου.”**
- Κατάστρωση ενός αρχικού σχεδίου για την επίτευξη του επιχειρηματικού στόχου και του data mining στόχου



University  
of  
East London



50

METROPOLITAN  
COLLEGE

## Φάση 2: Data Understanding



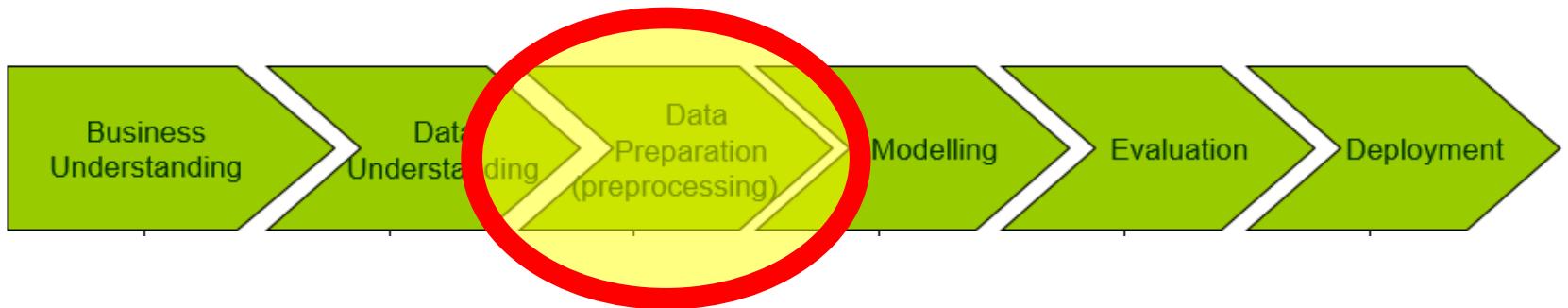
- Συλλογή των αρχικών δεδομένων
- Εξερεύνηση και εξοικείωση με τα δεδομένα
- Αξιολόγηση της ποιότητας των δεδομένων (π.χ. πληρότητα, εγκυρότητα)



University  
of  
East London



METROPOLITAN  
COLLEGE



## Προεπεξεργασία Δεδομένων (Data Preprocessing) στην Εξόρυξη Δεδομένων



University of  
East London



52

METROPOLITAN  
COLLEGE



**Μεταβλητή** είναι κάθε ιδιότητα ενός αντικειμένου ή μια κατάσταση που παίρνει διαφορετικές τιμές. Οι τιμές αυτές δεν είναι απαραίτητο να είναι αριθμητικές.

**Παραδείγματα:** Το βάρος, η νοημοσύνη, η στάση απέναντι στο ρατσισμό, τα πολιτικά κόμματα, η τιμή πώλησης ενός προιόντος.

**Εξαρτημένη Μεταβλητή** είναι η μεταβλητή που μας ενδιαφέρει και την οποία μετράμε με σκοπό π.χ. να κάνουμε πρόβλεψη της. **Παράδειγμα:** Τιμή πώλησης, υποψήφιοι πελάτες για παροχή δανείου.

**Ανεξάρτητη Μεταβλητή** είναι η μεταβλητή που χειριζόμαστε για να διαπιστώσουμε τι είδος επιρροή ασκεί στην εξαρτημένη μεταβλητή. **Παράδειγμα:** φύλο, οικογενειακή κατάσταση, ηλικία.



University  
of  
East London



METROPOLITAN  
COLLEGE

# Είδη (κλίμακες) χαρακτηριστικών (attributes) / μεταβλητών



Είδος Χαρακτηριστικού	Περιγραφή	Παράδειγμα	Πράξεις
Κατηγορικές (Categorical) ή Ποιοτικές (Qualitative)	Ονομαστικό (Nominal)	Μόνο για απαρίθμηση με βάση κάποια κριτήρια, χωρίς συγκεκριμένη διάταξη	Χρώμα ματιών, φύλο, οικογενειακή κατάσταση, ομάδα αίματος = ≠
	Τακτικό (Ordinal)	Όπως το ονομαστικό μόνο που έχει σημασία η σειρά	Φυσική κατάσταση («κακή», «μέτρια», «καλή», «άριστη»), Συχνότητα παρακολούθησης TV («ποτέ», σπάνια», «μερικές φορές», «συχνά») = ≠ < >
Αριθμητικές ή Ποσοτικές (Numerical) (Quantitative)	Ισοδιαστημικό (Interval)	Είναι αριθμοί και η διαφορά μεταξύ τους έχει σημασία (δεν ορίζεται όμως η αναλογία τους). Δεν υπάρχει η έννοια του «μηδενός»	Η μέτρηση του IQ: έχει σημασία η διαφορά μεταξύ των IQ (δείχνει διαφορά στην ευφυία, αλλά δεν μπορεί κανείς να πει ότι κάποιος με IQ = 100 είναι διπλάσια έξυπνος από έναν με IQ = 50. Δεν ορίζεται το IQ = 0 = ≠ < > + -
	Αναλογικό (Ratio)	Είναι αριθμοί και έχει σημασία και η διαφορά μεταξύ τους αλλά και η αναλογία τους). Ορίζεται το «μηδέν» που δείχνει την παντελή απουσία του μετρούμενου μεγέθους	Βάρος, ταχύτητα, μήκος, ηλικία = ≠ < > + - * /





Χωρίς ποιοτικά δεδομένα δεν μπορούμε να πετύχουμε ποιοτικό data mining (**GIGO**: Garbage In Garbage Out)

Η προεπεξεργασία είναι δυνάμει πολύπλοκη και χρονοβόρα και είναι δυνατόν να καταλάβει ακόμα και το **80%** της συνολικής προσπάθειας σε μια εργασία data mining.

**Προβλήματα** που μπορεί να έχουν τα δεδομένα:

- **Όχι πλήρη (incomplete):** είναι δυνατόν να λείπουν τιμές από γνωρίσματα που μας ενδιαφέρουν αλλά και να λείπουν και τα ίδια τα γνωρίσματα.
- **Με Θόρυβο (noise):** μπορεί να περιέχουν σφάλματα ή ανεπιθύμητα outliers, π.χ. «Μισθός = -1000» ή «Ηλικία = 107»
- **Ασυνεπή ή αντιφατικά (inconsistent):** π.χ. μετά τη συνένωση αρχείων από διαφορετικούς πίνακες, μπορεί η «αξιολόγηση» να έχει τιμές «Α, Β, Γ, ...» αλλά και «1, 2, 3, ...»





## 1. Καθαρισμός δεδομένων (Data cleaning)

Συμπλήρωση των χαμένων τιμών, εξομάλυνση δεδομένων με θόρυβο, αναγνώριση ή απομάκρυνση των outliers, διόρθωση ασυνεπειών στα δεδομένα

## 2. Ενοποίηση δεδομένων (Data integration)

Ενοποίηση πολλαπλών βάσεων δεδομένων

## 3. Μετασχηματισμός δεδομένων (Data transformation)

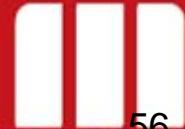
Γίνεται κανονικοποίηση (normalization), συνάθροιση (aggregation) και διακριτοποίηση (discretization) δεδομένων

## 4. Μείωση δεδομένων (Data reduction)

Διατηρούνται μειωμένες αναπαραστάσεις δεδομένων σε χωρητικότητα αλλά δημιουργούνται ίδια ή παρόμοια αποτελέσματα ανάλυσης, Μείωση διαστατικότητας

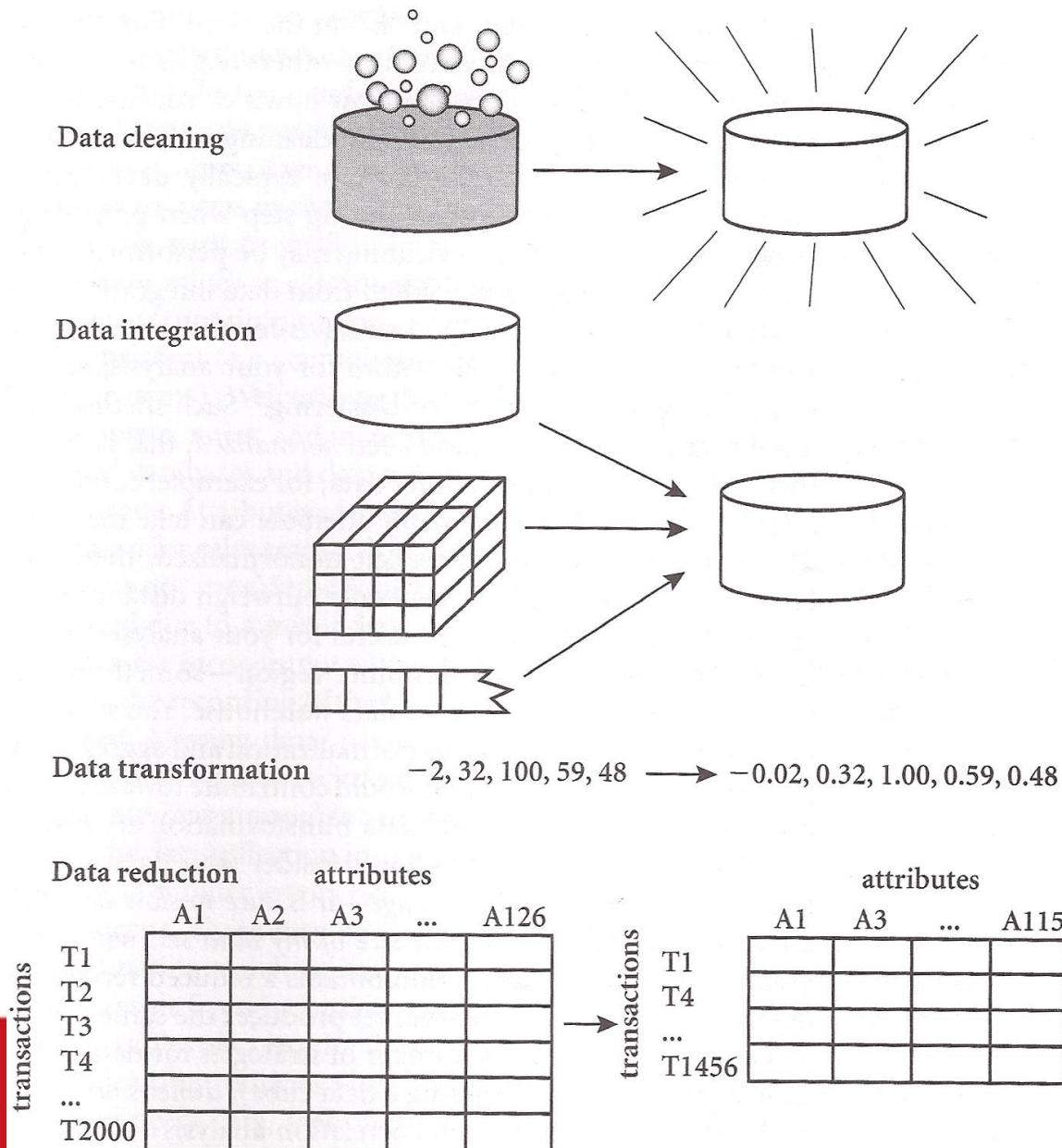


University  
of  
East London



METROPOLITAN  
COLLEGE

# Στάδια στην προ-επεξεργασία δεδομένων





## α. Συμπλήρωση των χαμένων τιμών

Τα δεδομένα δεν είναι πάντα διαθέσιμα για διάφορους λόγους (π.χ. πολλές πλειάδες (γραμμές) δεν έχουν τιμές για κάποια χαρακτηριστικά, όπως το εισόδημα του πελάτη στα δεδομένα πωλήσεων).

Κάποιες φορές είναι δυνατή η **χειρωνακτική** προσθήκη, άλλες φορές «**γεμίζουμε**» (impute) τις κενές τιμές με κάποιες γενικές σταθερές όπως «**unknowm**». Είναι δυνατόν να γεμίσουμε τις κενές τιμές **υπολογίζοντας την πιό πιθανή τιμή** βάσει των υπολοίπων με τη χρήση κάποιας στατιστικής μεθόδου (Bayesian) ή κάποιας μεθόδου machine learning όπως τα decision trees. Είναι βέβαια δυνατόν να **αγνοήσουμε** και εντελώς τις πλειάδες (γραμμές) που δεν υπάρχουν τιμές για κάποιο χαρακτηριστικό) ή και να αγνοήσουμε πλήρως ένα χαρακτηριστικό για το οποίο υπάρχουν πολλές κενές τιμές (π.χ. **>20%**)

## β. Μετατροπή των **nominal** τιμών σε **numerical**

Κάποιες μέθοδοι data mining (όπως το regression ή το nearest neighbor) απαιτούν μόνο αριθμητικές τιμές συνεπώς πρέπει να δημιουργήσουμε μια αντιστοιχία, π.χ. «άνδρας» = 1, «γυναίκα» = 2. Πολλά εργαλεία κάνουν τις αντιστοιχίσεις αυτόματα.



# 1. Εργασίες στον καθαρισμό δεδομένων (Data Cleaning)



## γ. Αναγνώριση των outliers (= δεδομένα μη ταιριαστά με την πλειοψηφία των δεδομένων) και εξομάλυνση δεδομένων με Θόρυβο

**Θόρυβος:** τυχαίο σφάλμα ή ασυμφωνία σε μετρημένες μεταβλητές. Μπορεί να οφείλεται σε λάθη στον τρόπο συλλογής δεδομένων ή σε λανθασμένα όργανα μέτρησης, σε λανθασμένη εισαγωγή δεδομένων, σε προβλήματα στη μετάδοση δεδομένων.

### Τρόποι αντιμετώπισης: Μέθοδος Binning.

- Αρχικά ταξινόμηση δεδομένων και διαχωρισμός τους σε (equi-depth, ίδιου εύρους) "bins".
- Εξομάλυνση (smoothing) μέσω bin means, εξομάλυνση μέσω bin median, εξομάλυνση μέσω bin boundaries, κλπ.

Παράδειγμα: ταξινομημένες τιμές:

**4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**

Partition the value range into three (equi-depth) bins (i.e., each bin contains 4 samples)

Bin 1: 4, 8, 9, 15

Bin 2: 21, 21, 24, 25

Bin 3: 26, 28, 29, 34      Smoothing by bin means

Bin 1: 9, 9, 9, 9  
Bin 2: 23, 23, 23, 23  
Bin 3: 29, 29, 29, 29

Smoothing by bin boundaries

Bin 1: 4, 4, 4, 15  
Bin 2: 21, 21, 25, 25  
Bin 3: 26, 26, 26, 34

In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value

## δ. Διόρθωση ασυνεπειών στα δεδομένα και απαλοιφή πλεονασμού

## 2. Ενοποίηση Δεδομένων (Data Integration)



- Ενοποίηση δεδομένων (Data integration)
  - Ενώνει δεδομένα από πολλαπλές πηγές
- Ενοποίηση σχήματος (Schema integration)
  - Ενοποίηση μετα-δεδομένων από διαφορετικές πηγές
- Ανίχνευση και επίλυση συγκρούσεων σε τιμές δεδομένων
  - Για την ίδια οντότητα οι τιμές από διαφορετικές πηγές είναι διαφορετικές (π.χ. διαφορετικές μονάδες μέτρησης των ίδιων μεγεθών)
- Είναι δυνατόν να υπάρχουν συχνά **πλεονάζοντα** (redundant) data.
  - Το ίδιο χαρακτηριστικό μπορεί να έχει διαφορετικό όνομα σε διαφορετικές βάσεις δεδομένων
  - Ένα χαρακτηριστικό μπορεί να συνεπάγεται από ένα άλλο
  - Πλεονάζοντα δεδομένα μπορούν να βρεθούν με προσεκτική ανάλυση συσχετίσεων (correlation analysis)
  - Προσεκτική ενοποίηση δεδομένων από πολλαπλές πηγές μπορεί να βοηθήσει στη μείωση των πλεοναζόντων δεδομένων

### 3. Μετασχηματισμός Δεδομένων (Data Transformation)



#### α. Smoothing

απομάκρυνση θορύβου από τα δεδομένα

#### β. Aggregation

συνάθροιση, data cube construction

#### γ. Normalization

αλλαγή κλίμακας μεγέθους για να «πέφτουν» μέσα σε ένα μικρό προκαθορισμένο εύρος τιμών

- min-max normalization

Min-max normalization is one of the most common ways to normalize data. For every feature, the minimum value of that feature gets transformed into a 0, the maximum value gets transformed into a 1, and every other value gets transformed into a decimal between 0 and 1.

- z-score normalization

Z-score normalization: This technique scales the values of a feature to have a mean of 0 and a standard deviation of 1.

- normalization by decimal scaling

In decimal scaling normalization, the decimal point of values of the attributes is moved. The movement of the decimal points in decimal scaling normalization is dependent upon the maximum values amongst all values of the attribute.

#### δ. Δημιουργία νέων χαρακτηριστικών

Χρησιμοποιούνται για να βελτιώσουν τη διαδικασία εξόρυξης γνώσης



University  
of  
East London



METROPOLITAN  
COLLEGE

### 3. Μετασχηματισμός Δεδομένων (Data Transformation)



#### Κανονικοποίηση - παράδειγμα

- Σκοπός της κανονικοποίησης:

η αντιστοίχιση των τιμών των δεδομένων από το διάστημα

$$[\min A, \max A] \rightarrow [\text{new\_min}A, \text{new\_max}A]$$

- Min-max normalization:

$$s' = \frac{s - \min\{s_k\}}{\max\{s_k\} - \min\{s_k\}}$$

Επίσης, υπάρχουν παραλλαγές της min max κανονικοποίησης ώστε το διάστημα [new\_min, new\_max] να μην είναι κατ' ανάγκη το [0,1]

- Θεωρούμε τα δεδομένα από 30-50 και έστω ότι θέλουμε να τα μετασχηματίσουμε ώστε να κυμαίνονται από 0-1.
- Θα χρησιμοποιήσουμε Min-max normalization
- Το στοιχείο 30 αντιστοιχίζεται ως εξής:
  - $s' = (30-30)/(50-30) = 0$
- Το στοιχείο 50 αντιστοιχίζεται ως εξής:
  - $s' = (50-30)/(50-30) = 1$
- Το ενδιάμεσο στοιχείο 35 αντιστοιχίζεται ως εξής:
  - $s' = (35-30)/(50-30) = 5/20 = 0.25$



University of  
East London



62

METROPOLITAN  
COLLEGE



## Πρόβλημα:

Μεγάλες αποθήκες δεδομένων μπορούν να έχουν terabytes δεδομένων,

Πολύπλοκη ανάλυση δεδομένων και εξόρυξη γνώσης μπορεί να απαιτήσει πολύ χρόνο

## Λύση:

Μείωση δεδομένων (Διατηρούνται μειωμένες αναπαραστάσεις δεδομένων σε χωρητικότητα αλλά πρέπει να διατηρούνται ίδια ή παρόμοια αποτελέσματα ανάλυσης)

## Μερικές Στρατηγικές:

- Data cube aggregation (δες στο OLAP)
- Μείωση διαστάσεων (Dimension reduction)
- ...



University  
of  
East London



METROPOLITAN  
COLLEGE



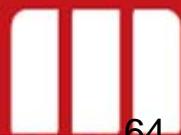
### Μείωση διαστάσεων

Μπορεί να επιτευχθεί με δύο μεθόδους:

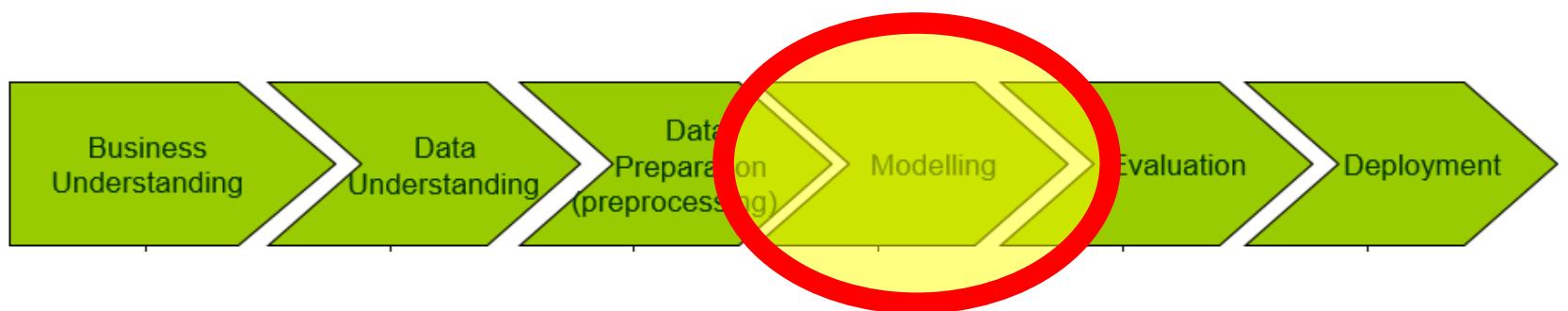
- Επιλογή χαρακτηριστικών: Επιλογή ενός ελάχιστου πλήθους ( $m$ ) χαρακτηριστικών με τα οποία είναι δυνατή η εξαγωγή ισοδύναμων ή κοντινών αποτελεσμάτων με αυτά που θα είχαμε αν είχαμε κρατήσει όλα τα χαρακτηριστικά για ανάλυση ( $n$ ). Ιδανικά  $m << n$ .
- Μετασχηματισμός χαρακτηριστικών: Είναι γνωστός ως **Principle Component Analysis (PCA)**. Ο μετασχηματισμός των χαρακτηριστικών δημιουργεί ένα νέο σύνολο χαρακτηριστικών, λιγότερων διαστάσεων από το αρχικό, αλλά χωρίς μείωση των βασικών διαστάσεων. Επίσης, συχνά χρησιμοποιείται για την οπτικοποίηση των δεδομένων.



University  
of  
East London



METROPOLITAN  
COLLEGE



## Data Mining Modeling - μερικές ενδεικτικές ΤΕΧΝΙΚΕΣ



University of  
East London



65

METROPOLITAN  
COLLEGE

# Data Mining – Association Analysis



- Η ανάλυση Καλαθιού Αγοράς (Market-Basket analysis) είναι ένα παράδειγμα ανάλυσης data mining που
  - προσδιορίζει μοτίβα αγορών (sales patterns).
  - Συσχετισμός από κοινού αγοραζόμενα προϊόντα
- Έτσι βοηθάει τις επιχειρήσεις να ανακαλύψουν ευκαιρίες cross-selling.

Market Basket Data (TID: Transaction ID)	
TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Coffee, Sugar, Eggs
6	Bread, Milk, Diapers, Coke

## Παράδειγμα Association Rules

{Diapers} → {Beer},  
{Milk, Bread} → {Eggs, Coke},  
{Beer, Bread} → {Milk},

....

→ σημαίνει συνύπαρξη

Σκεφτείτε ένα πανεπιστήμιο: σε ποιά modules οι φοιτητές κάθε course έχουν την τάση να αποτυγχάνουν ταυτόχρονα? Επίσης σε εγκληματολογική ανάλυση.

# Data Mining – Δέντρο Αποφάσεων (Decision Trees)



- Μια τεχνική data-mining που επιλέγει τα πιο χρήσιμα γνωρίσματα για την κατηγοριοποίηση στοιχείων με βάση κάποιο κριτήριο
- Οι αλγόριθμοι της τεχνικής αυτής εξετάζουν τα γνωρίσματα των δεδομένων και ανευρίσκουν αυτά για τα οποία τα δεδομένα διαφέρουν περισσότερο σε σχέση με το αρχικό κριτήριο (διακριτική ικανότητα)
- Είναι μια ιεραρχική διάταξη συνθηκών με την οποία μπορούμε να προβλέψουμε την κατηγοριοποίηση κάποιου στοιχείου

Predictors				Target
Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

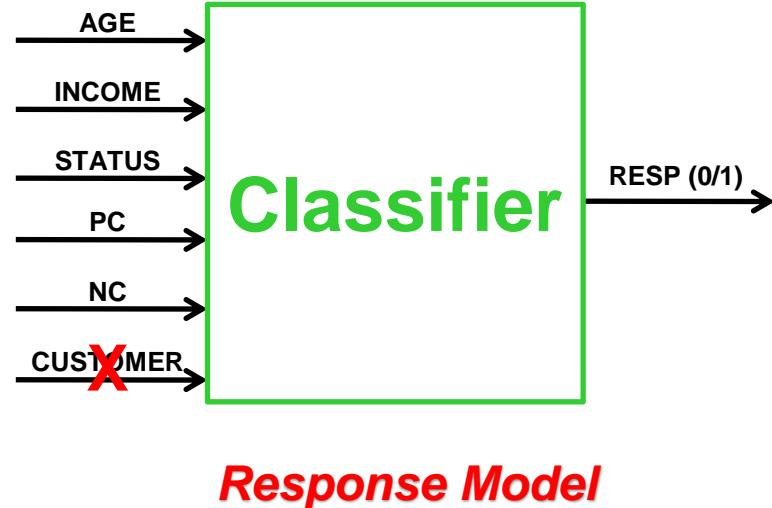


# Data Mining – Δέντρο Αποφάσεων (Decision Trees)



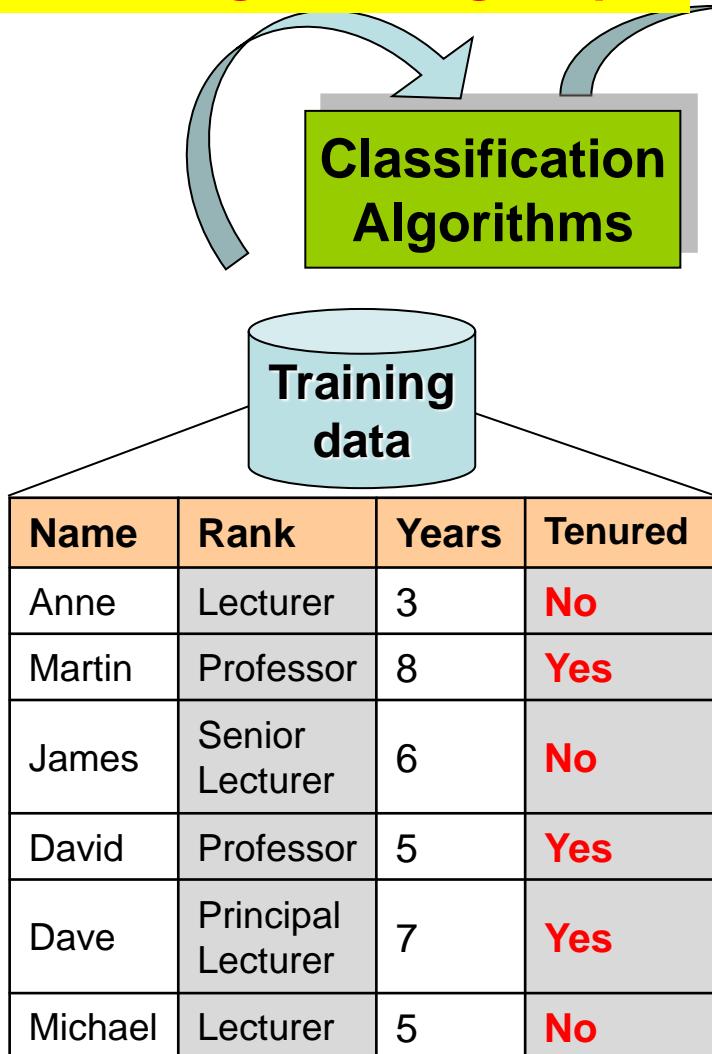
Τα δέντρα αποφάσεων είναι μια τεχνική μοντελοποίησης με την οποία επιτυγχάνεται **ταξινόμηση**.

CUSTOMER	AGE	INCOME	STATUS	PC	NC	RESP
1	25	£45,000	S	1	1	0
2	45	£65,000	MC	1	2	1
3	54	£25,000	MC	1	3	0
4	32	£31,000	MNC	0	4	0
5	43	£380,000	S	0	5	0
6	56	£145,000	MC	1	5	1
7	78	£445,000	W	0	7	1
8	6	£64,000	S	1	1	0
10	26	£40,000	S	1	2	1

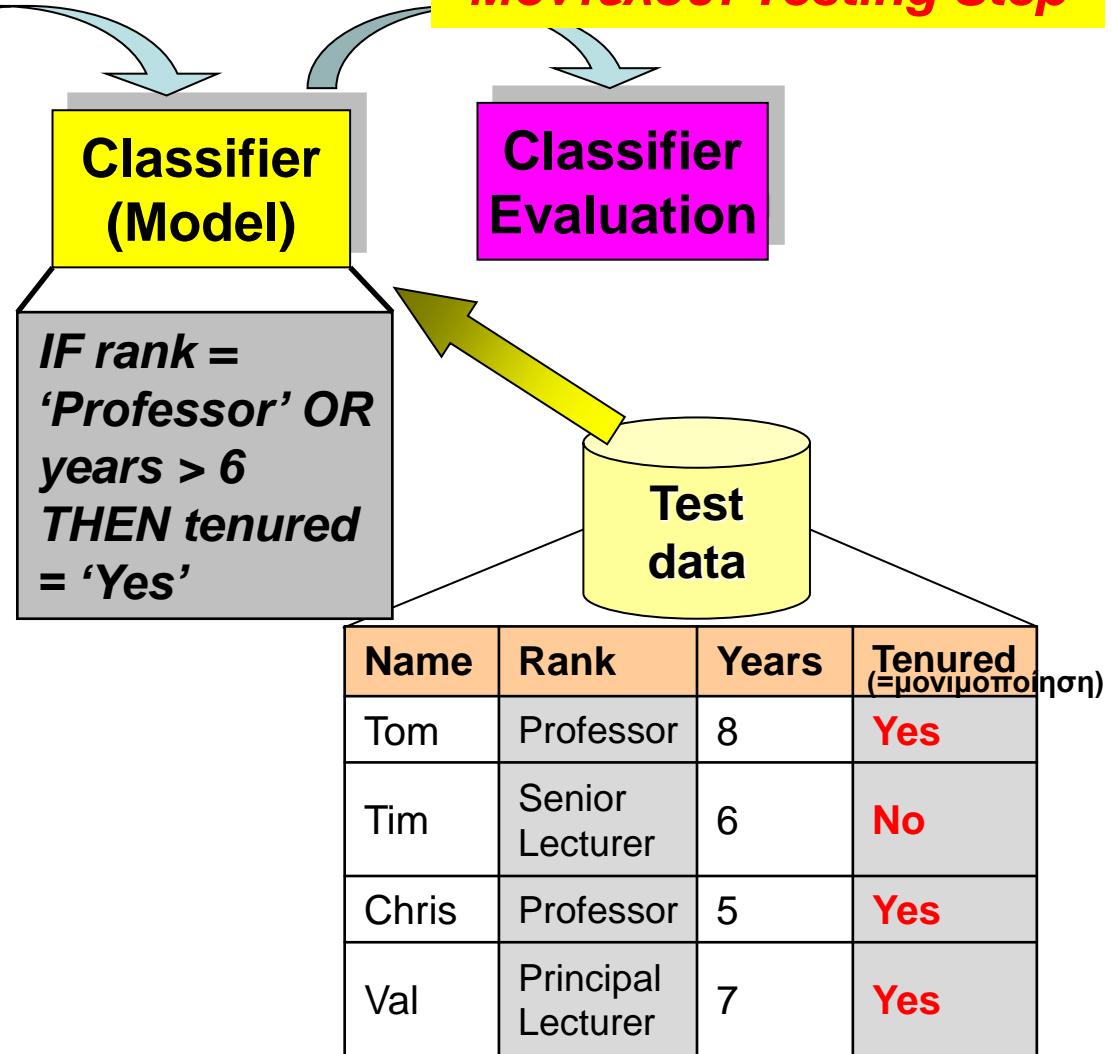


# Τι είναι η ταξινόμηση: Διαδικασία 2 βημάτων

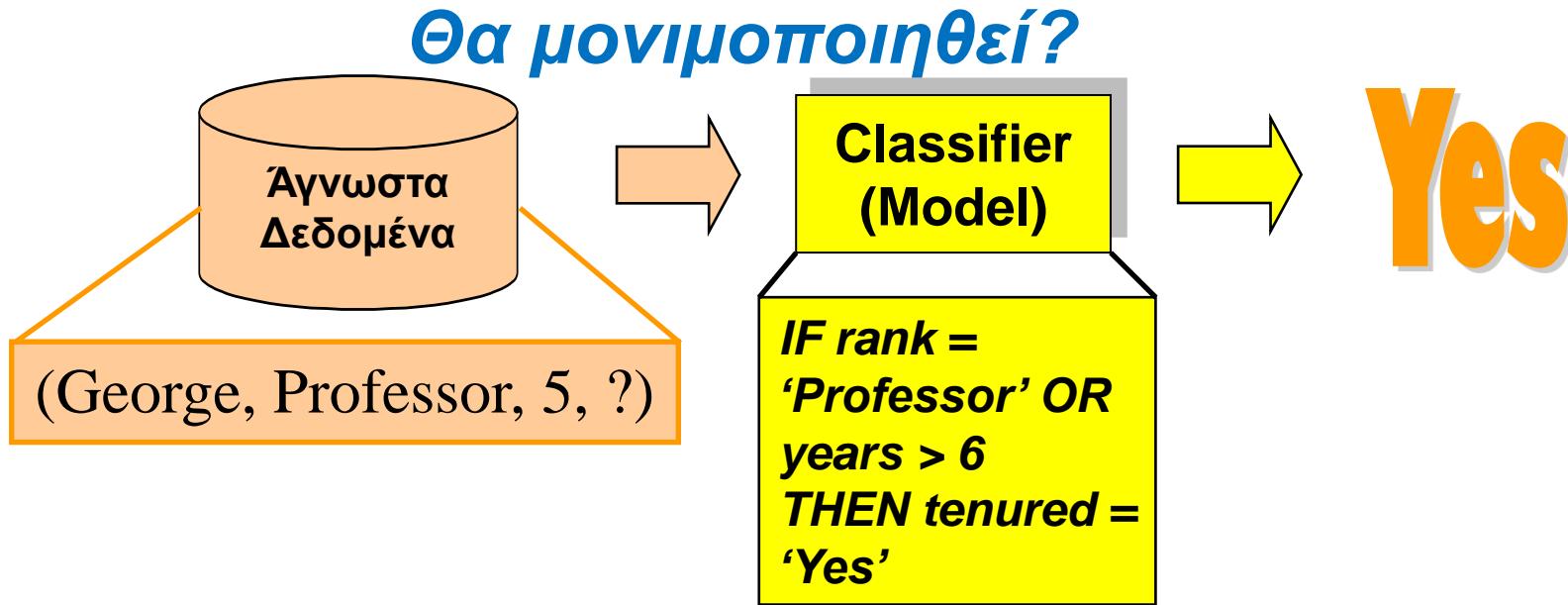
## 1. Κατασκευή μοντέλου: Learning /Training Step



## 2. Αξιολόγηση Μοντέλου: Testing Step



## (3) Model Usage: Predicting future or unknown objects



*'Learning or training' means determining the parameters in a classifier model based on samples and certain criteria*



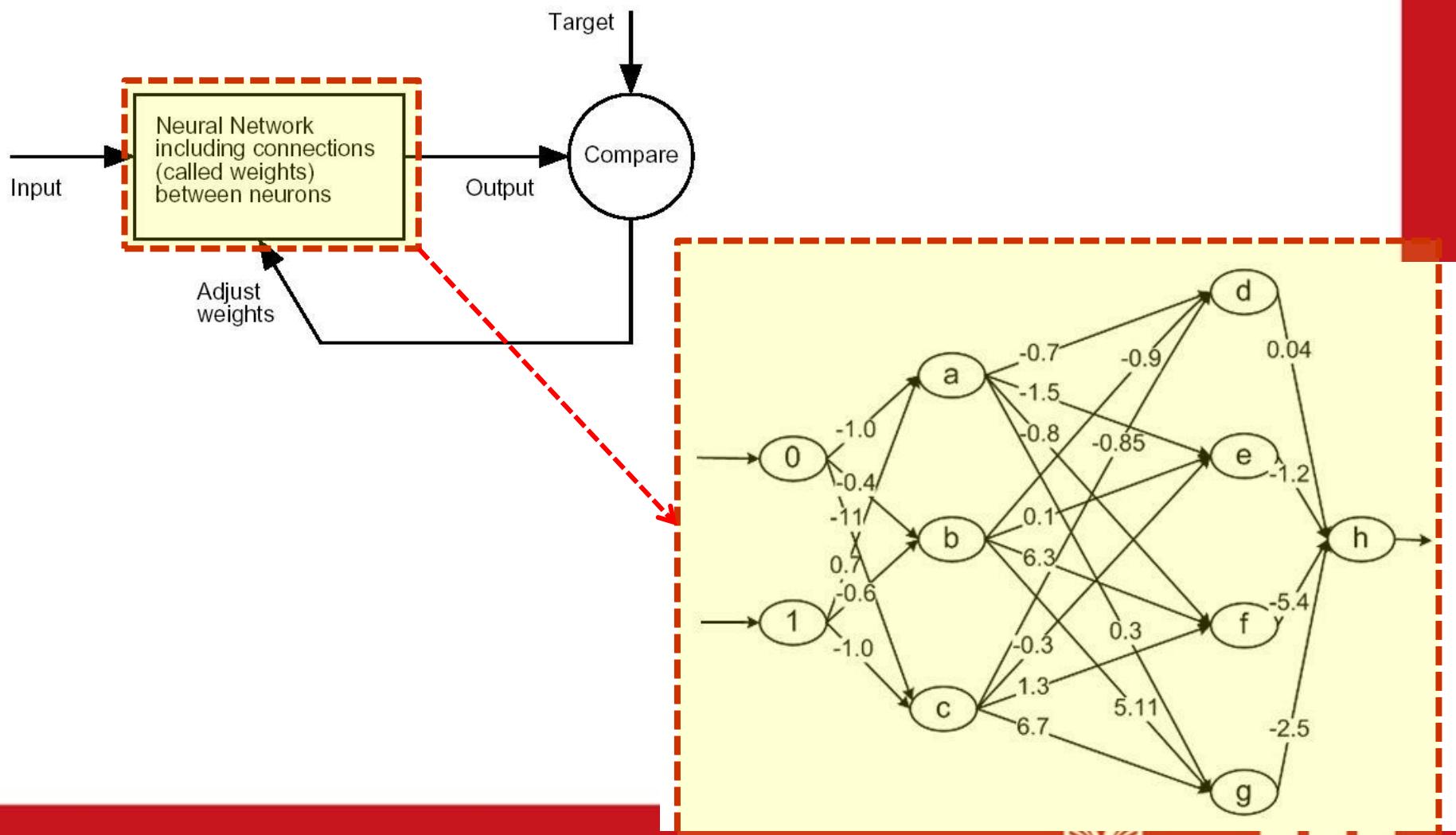
- ◆ Τα νευρωνικά δίκτυα **μαθαίνουν σχέσεις** μεταξύ εισερχόμενων και εξερχόμενων δεδομένων (αιτίων και αποτελεσμάτων) ή **κατηγοριοποιούν και Βρίσκουν μοτίβα (patterns)** σε **μεγάλες ποσότητες δεδομένων large quantities of data.**
- ◆ Τα νευρωνικά δίκτυα είναι μοντέλα των νευρώνων στο ανθρώπινο μυαλό.
- ◆ **Εκπαιδεύουμε τα νευρωνικά δίκτυα** να επιτελούν μια συγκεκριμένη λειτουργία αλλάζοντας τα βάρη (ισχύ) των συνδέσεων στους νευρώνες που τα αποτελούν.
- ◆ Δε χρειάζονται προγραμματισμό γιατί μαθαίνουν με **παραδείγματα**.
- ◆ Τα εκπαιδευμένα νευρωνικά δίκτυα έχουν την ικανότητα (αν εκπαιδευθούν σωστά) να **γενικεύουν** τη γνώση τους σε παραδείγματα που δεν έχουν «δει» ποτέ.
- ◆ Είναι εξαιρετικά καλά σε προβλήματα που είναι δύσκολο να βρεθεί ένα αναλυτικό μοντέλο που να τα περιγράφει.



Οι δύο βασικότερες κατηγορίες νευρωνικών δικτύων αίναι τα supervised και unsupervised.

- Στη **supervised εκμάθηση**, οι σωστές τιμές (*target values, επιθυμητές έξοδοι*) είναι γνωστές και δίνονται στο νευρωνικό δίκτυο κατά την εκπαίδευση ούτως ώστε το νευρωνικό δίκτυο να προσαρμόσει τα βάρη των συνδέσεών του για να πλησιάζουν οι έξοδοι του τις επιθυμητές, γνωστές τιμές. Μετά την εκπαίδευση δοκιμάζουμε το νευρωνικό δίκτυο δίνοντας του νέα inputs και ελέγχοντας πόσο τα outputs πλησιάζουν τις επιθυμητές τιμές.
- Στην **unsupervised εκπαίδευση**, οι σωστές τιμές δεν είναι γνωστές εκ των προτέρων. Το νευρωνικό δίκτυο ομαδοποιεί τα δεδομένα εισόδου και οι ομάδες (clusters) που δημιουργούνται πρέπει να αναλυθούν και εξηγηθούν.

# Νευρωνικά δίκτυα: Supervised εκμάθηση



# Πηγές

- **“Τεχνητή Νοημοσύνη – Β έκδοση»: Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας, Η. Σακελλαρίου**



University  
of  
East London



METROPOLITAN  
COLLEGE