Εισαγωγή στη Μηχανική Μάθηση και στην Αναλυτική Δεδομένων

Χάρης Γεωργίου (MSc, PhD)

Ένωση Πληροφορικών Ελλάδας

Στόχοι:

- Πρώτος "καθολικός" φορέας εκπροσώπησης πτυχιούχων Πληροφορικής.
- Αρμόδιος φορέας εκπροσώπησης επαγγελματιών Πληροφορικής.
- Αρμόδιος επιστημονικός "συμβουλευτικός" φορέας για το Δημόσιο.
- Αρωγός της Εθνικής Ψηφιακής Στρατηγικής & Παιδείας της χώρας.

Τομείς παρέμβασης

Ποιοι είναι οι κύριοι τομείς παρεμβάσεων της ΕΠΕ;

1. Εθνική Ψηφιακή Στρατηγική & Οικονομία
2. Εργασιακά (ΤΠΕ), Δημόσιος & ιδιωτικός τομέας
3. Παιδεία (Α′, Β′, Γ′)
4. Έρευνα & Τεχνολογία
5. Έργα & υπηρεσίες ΤΠΕ
6. Ασφάλεια συστημάτων & δεδομένων
7. Ανοικτά συστήματα & πρότυπα
8. Χρήση ΕΛ/ΛΑΚ
9. Πνευματικά δικαιώματα
10. Κώδικας Δεοντολογίας (ΤΠΕ)
11. Κοινωνική μέριμνα (ICT4D)

Harris Georgiou (MSc, PhD) – https://github.com/xgeorgio/info

- R&D: Associate post-doc researcher and lecturer with the University Athens (NKUA) and University of Piraeus (UniPi)
- Consultant in Medical Imaging, Machine Learning, Data Analytics, Signal Processing, Process Optimization, Dynamic Systems, Complexity & Emergent A.I., Game Theory
- HRTA member since 2009, LEAR / scientific advisor
- HRTA field operator (USAR, scuba diver)
- Wilderness first aid, paediatric (child/infant)
- Humanitarian aid & disaster relief in Ghana, Lesvos, Piraeus
- Support of unaccomp. minors, teacher in community schools
- Streetwork training, psychological first aid & victim support
- 2+ books, 160+ scientific papers/articles (and 5 marathons)

Επισκόπηση – Πηγές

- Περιεχόμενα:
  - Τι είναι η Μηχανική Μάθηση και η Αναλυτική Δεδομένων (ML/DA)
  - Προπαρασκευή δεδομένων (pre-processing), είδη προβλημάτων ML/DA
  - Αλγόριθμοι: κατηγοριοποίηση, συσταδοποίηση, ανακάλυψη συχνών προτύπων , ...
  - Ειδικά θέματα (π.χ. δεδομένα ήχου, εικόνας, ιατρικά, ...)

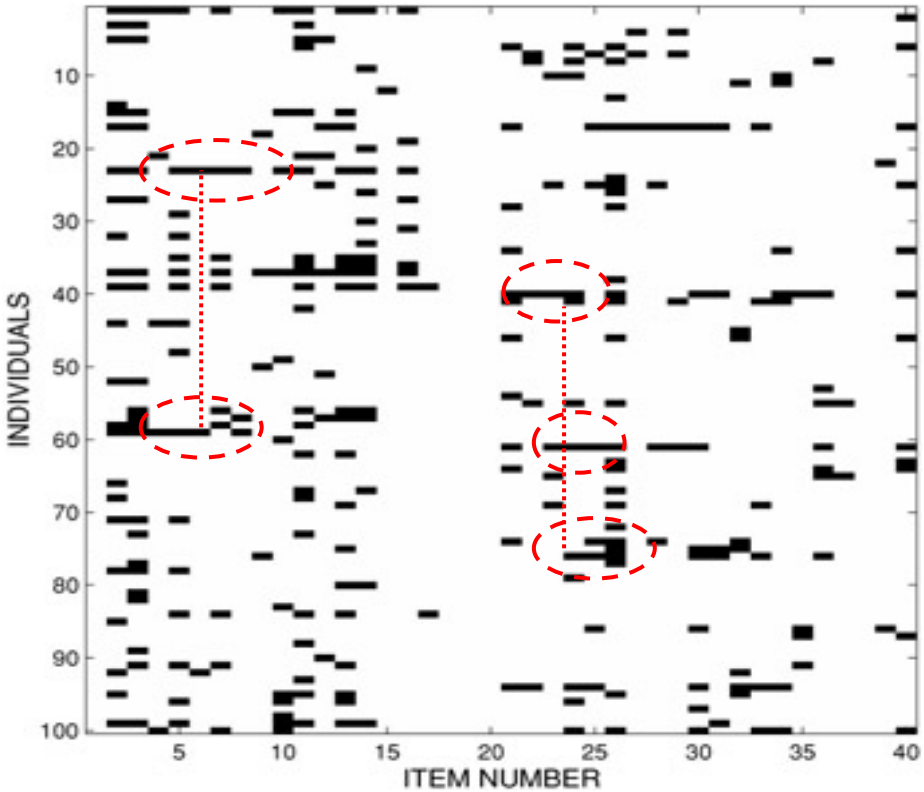
- Πηγές:
  - «Αναλυτική Δεδομένων» – μάθημα ΠΜΣ Πανεπ. Πειραιά (σημειώσεις) 2017-2021.
  - Dunham: Data Mining – Introductory and Advanced Topics. Prentice Hall, 2003.
  - Tan, Steinbach, Kumar: Introduction to Data Mining. Addison Wesley, 2006.
  - Hand, Mannila, Smyth: Principles of Data Mining. MIT Press, 2001.

Δεδομένα παντού ...

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

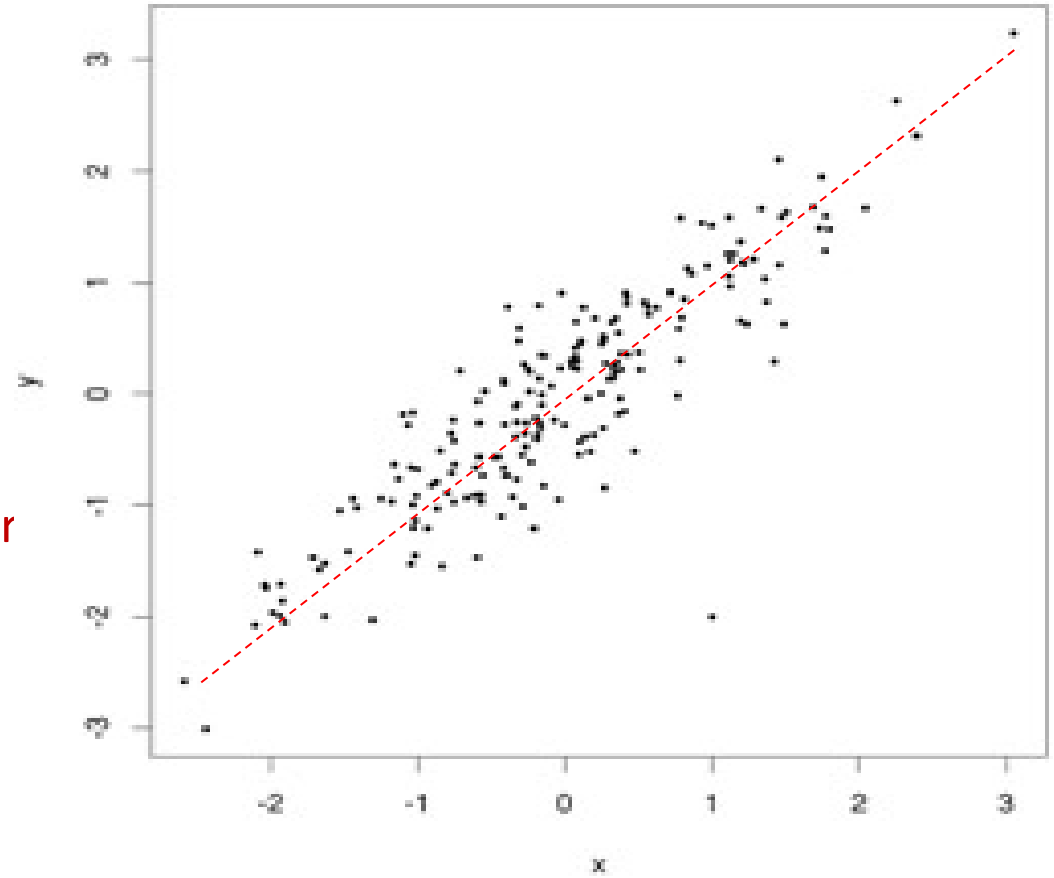
Δεδομένα…

- 2-D πίνακας: πελάτες (γραμμές) προς προϊόντα (στήλες)
- Στόχος ανάλυσης: (εύκολος;) εντοπισμός συλλογικής συμπεριφοράς
- Αποτέλεσμα: Αναγνώριση ισχυρών εξαρτήσεων μεταξύ μεταβλητών
- Στον πίνακα: «συνεμφάνιση» τιμών ή γεγονότων



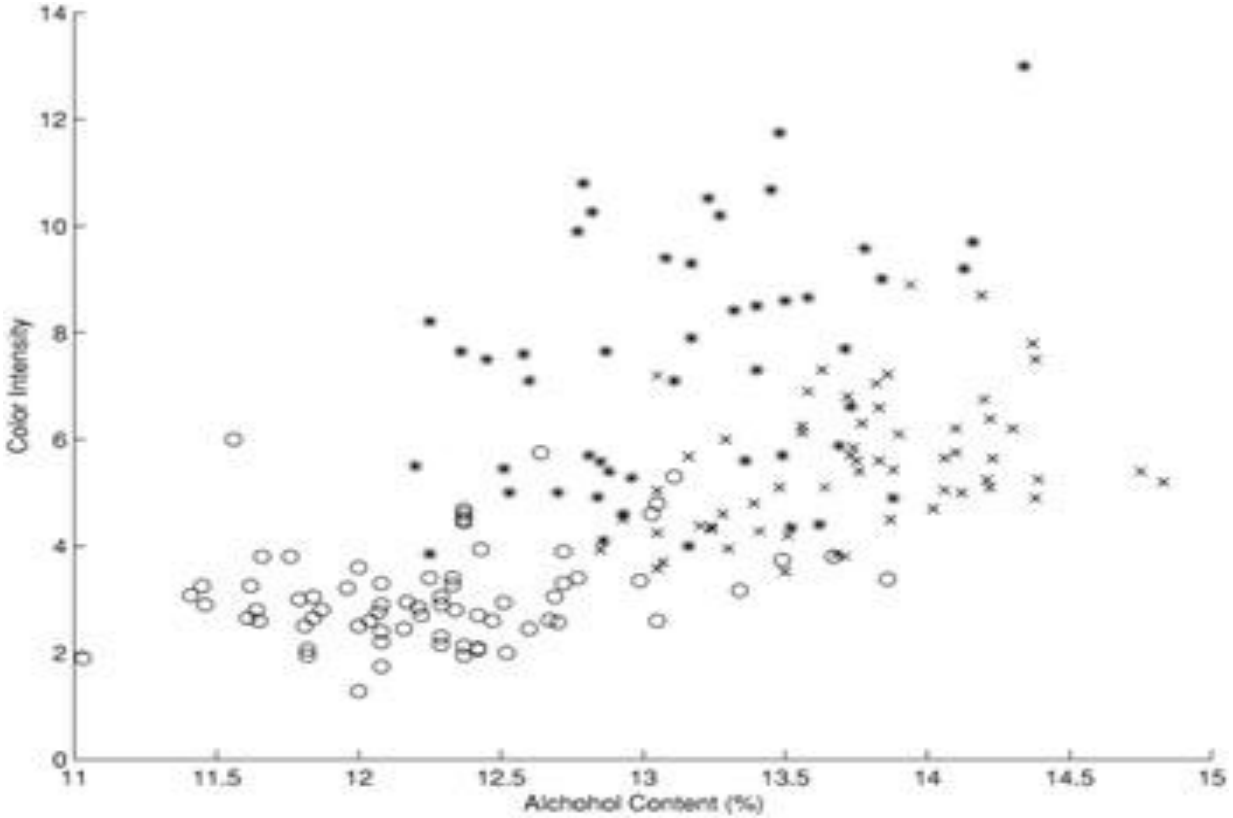
Δεδομένα…

- 2-D διάνυσμα (x,y)
- Στόχος ανάλυσης: πρόβλεψη «αναμενόμενης» εξόδου για οποιαδήποτε είσοδο
- Εναλλακτικά: εντοπισμός «ακραίας συμπεριφοράς»
- Γραμμική προσαρμογή (linear regression): η «καλύτερη» ευθεία που «εξηγεί» τα δεδομένα

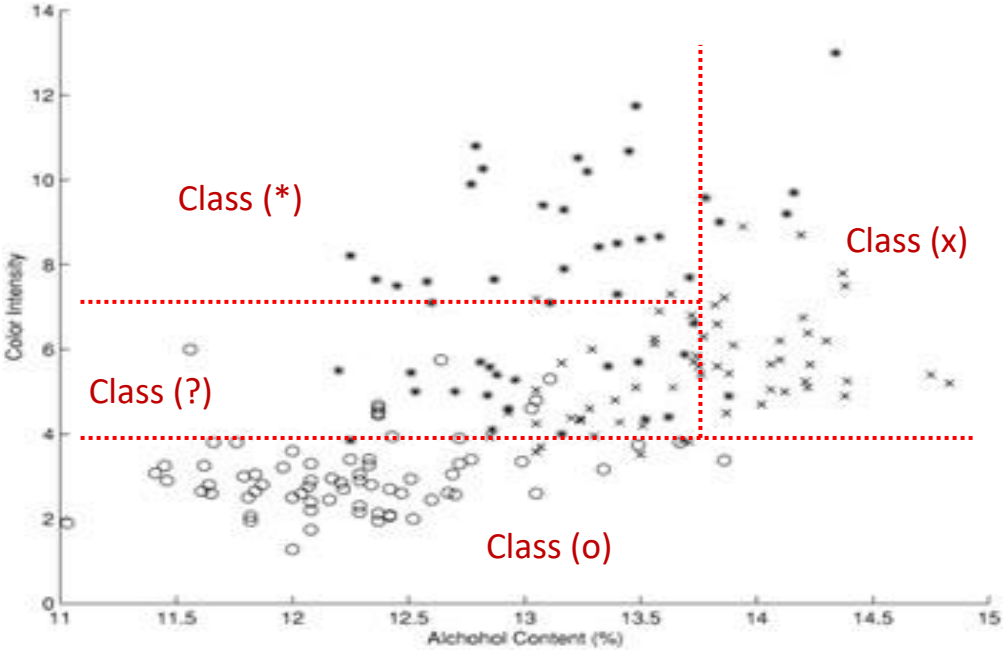


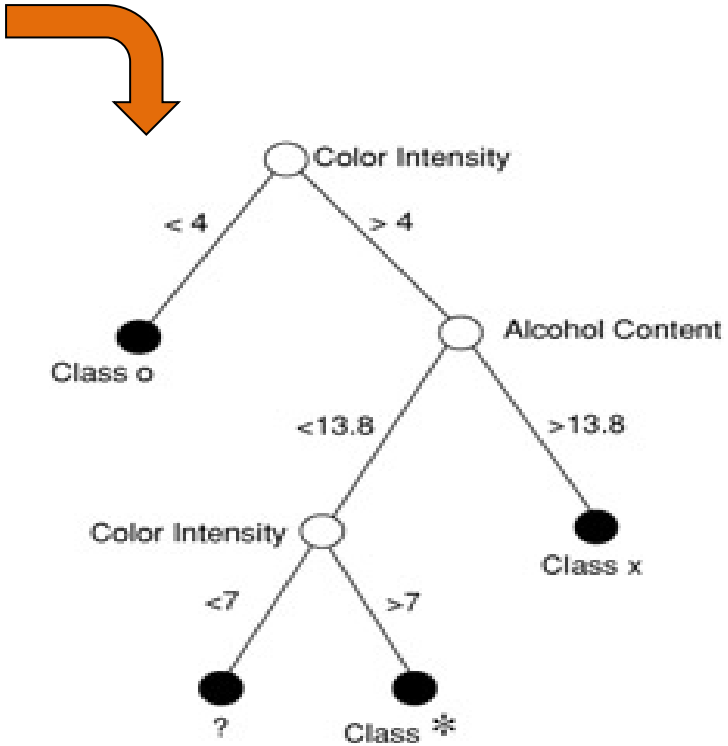
Δεδομένα…

- 2-διάστατο διάνυσμα (βαθμός αλκοόλης, ένταση χρώματος)
- Στόχος ανάλυσης: κατηγοριοποίηση κρασιών σε κλάσεις



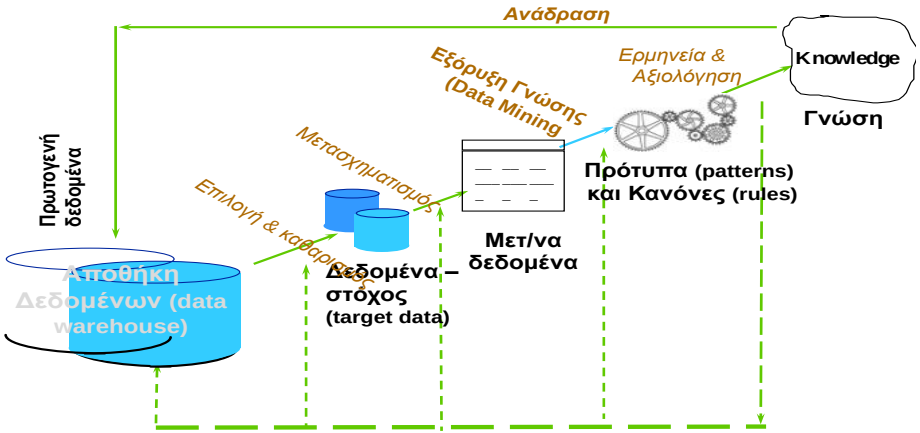
Από τα δεδομένα στη γνώση…

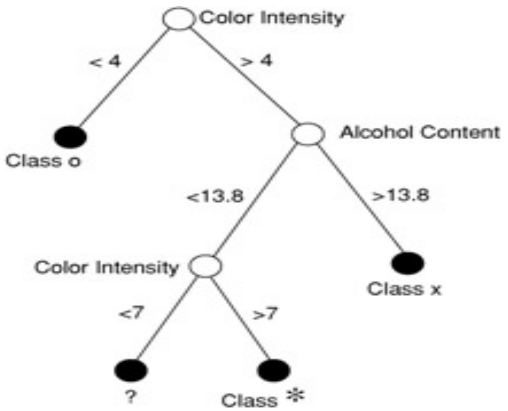


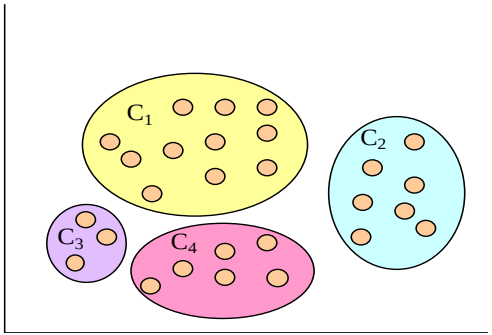


Εναλλακτικοί ορισμοί

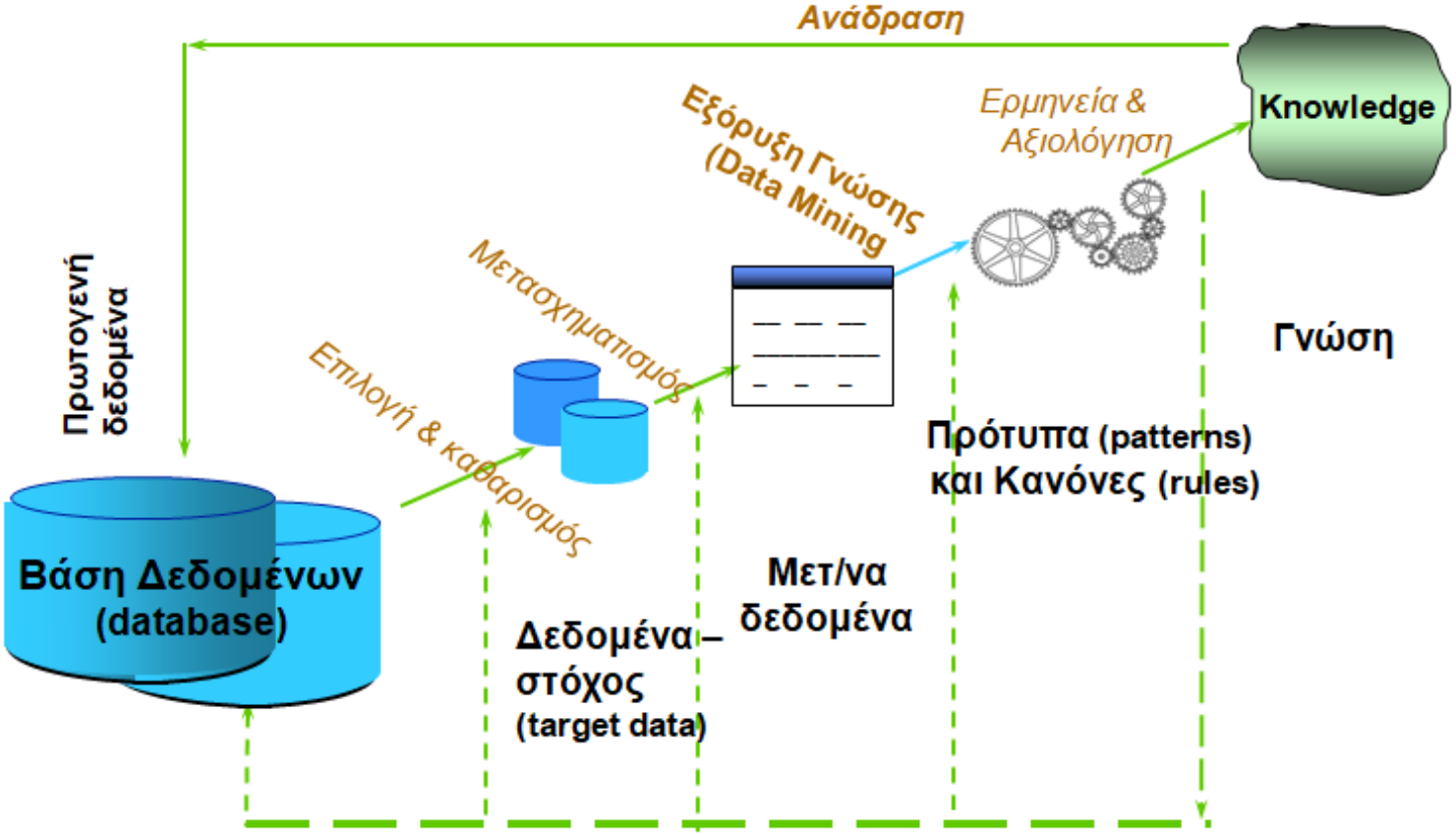
- **Knowledge Discovery in Data (KDD)** ή **Data mining**: *μη-τετριμμένη* διαδικασία εύρεσης έγκυρων, πρωτότυπων, πιθανώς χρήσιμων και, οπωσδήποτε, κατανοητών προτύπων (patterns) μέσα στα δεδομένα.



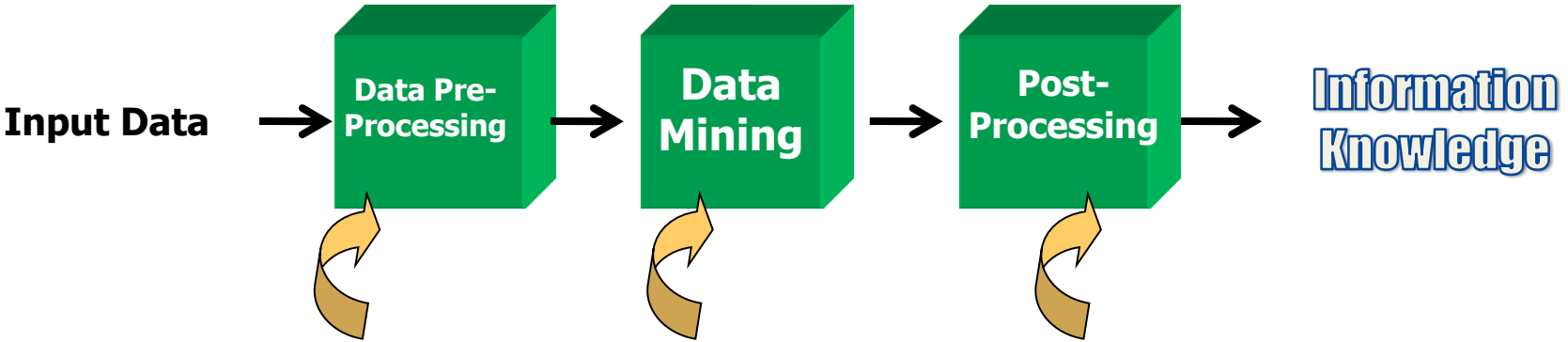




Η "σκάλα" της διαδικασίας KDD



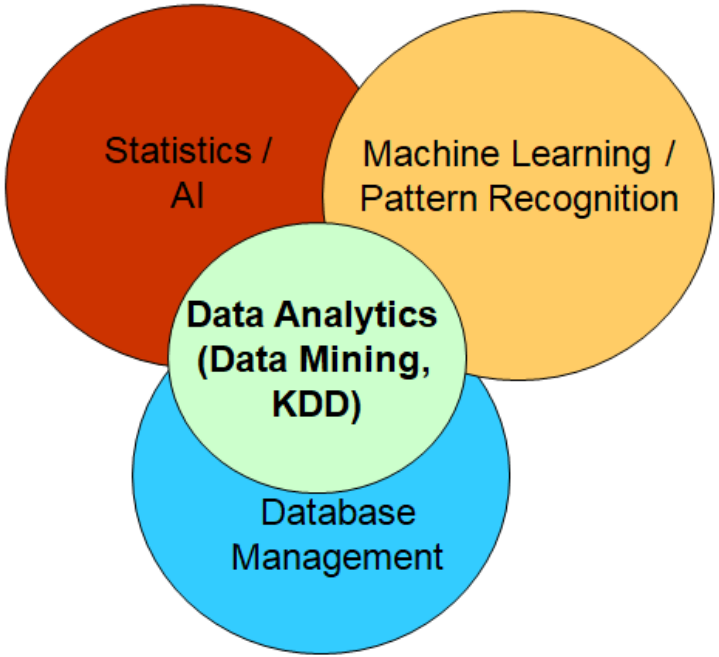
Εναλλακτικά...



- Τυπική θεώρηση από τη σκοπιά της Στατιστικής και της Μηχανικής Μάθησης

Σχετικά επιστημονικά πεδία

- Στατιστική / *«Τεχνητή Νοημοσύνη»*, Μηχανική Μάθηση / Αναγνώριση Προτύπων, Διαχείριση Βάσεων Δεδομένων
- Οι παραδοσιακές τεχνικές επεξεργασίας δεδομένων που μας προσφέρουν αυτές οι επιστημονικές περιοχές μπορεί να είναι ανεφάρμοστες λόγω:
  - του μεγάλου όγκου,
  - των πολλών διαστάσεων,
  - της ετερογένειας των δεδομένων,
  - των απαιτήσεων επεξεργασίας,
  - ...



Εφαρμογές Αναλυτικής Δεδομένων

- **Ανάλυση συμπεριφοράς**
  - στοχευμένο marketing
  - ανάλυση καλαθιού αγοράς (market basket analysis)
  - τμηματοποίηση αγοράς (customer segmentation)
  - ηλεκτρονικό εμπόριο
  - …
- **Ανάλυση κινδύνου**
  - πρόβλεψη τάσεων (συγκράτηση / διαρροή πελατών)
  - ανάλυση ανταγωνισμού
  - ανίχνευση απάτης (fraud detection)
  - …

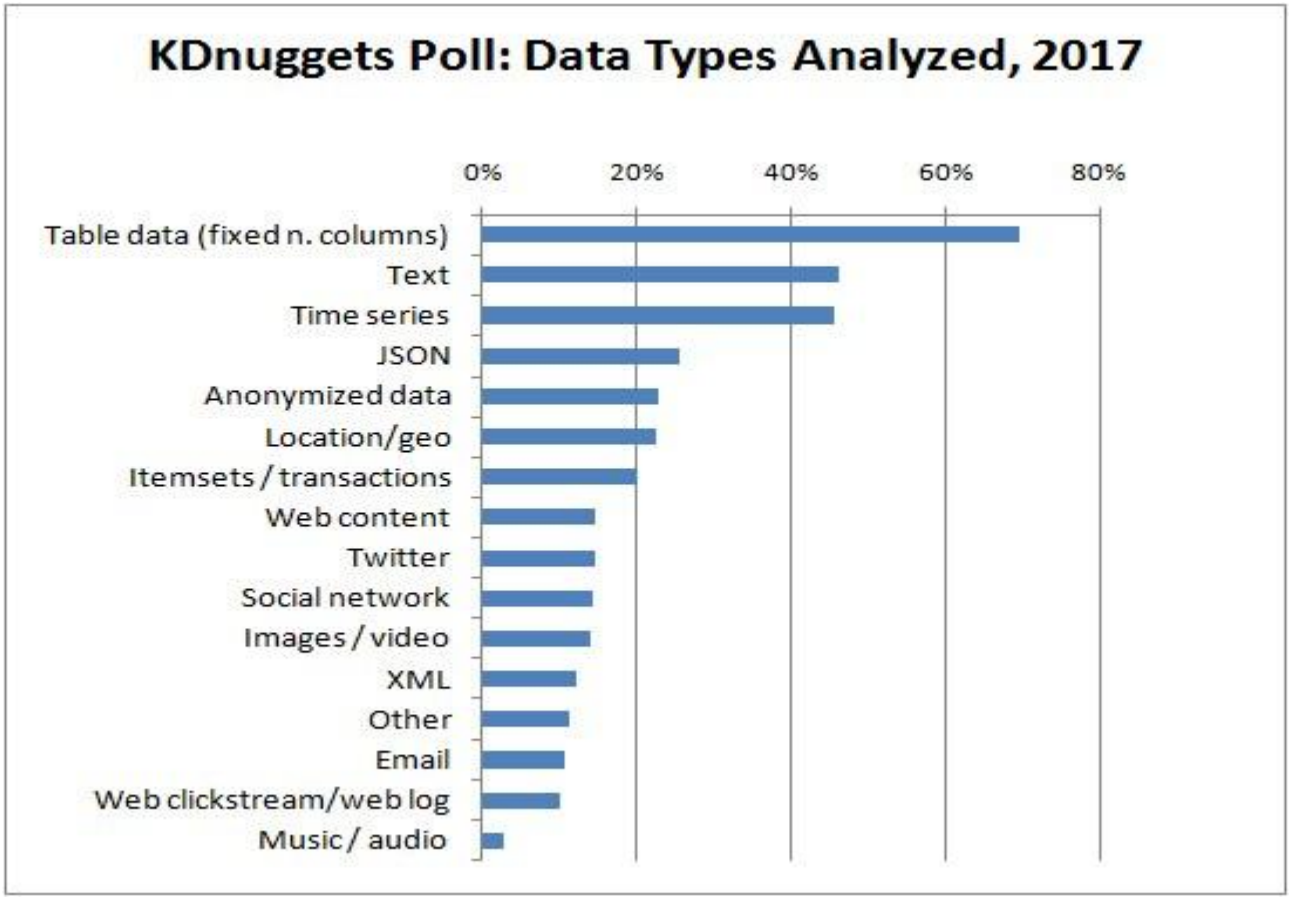
Facebook–Cambridge Analytica data scandal

From Wikipedia, the free encyclopedia

The **Facebook–Cambridge Analytica data scandal** was a major political scandal in early 2018 when it was revealed that Cambridge Analytica had harvested the personal data of millions of people's Facebook profiles without their consent and used it for political purposes. It has been described as a watershed moment in the public understanding of personal data and precipitated a massive fall in Facebook's stock price and calls for tighter regulation of tech companies' use of data.

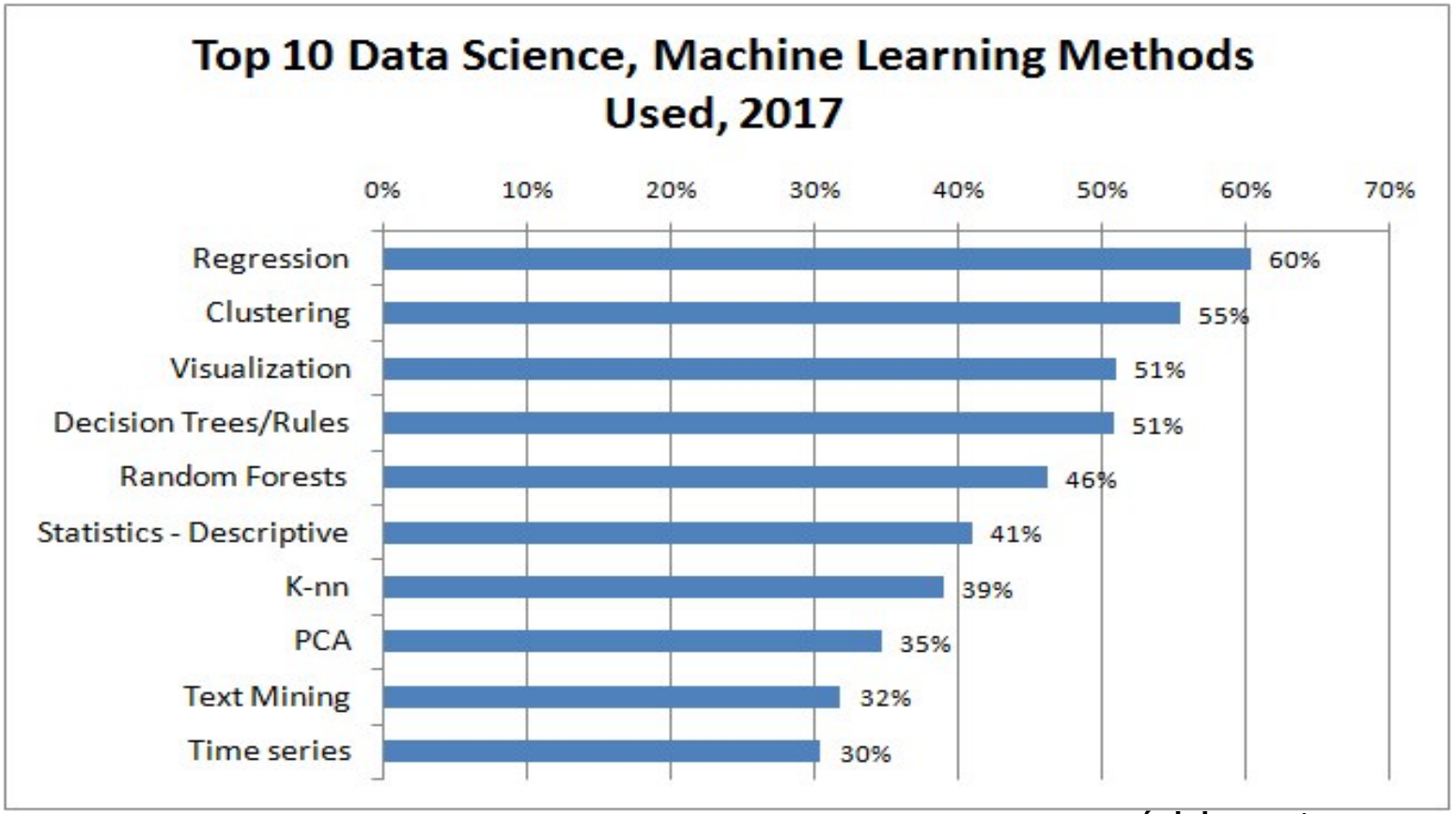
The illicit harvesting of personal data by Cambridge Analytica was first reported in December 2015 by Harry Davies, a journalist for *The Guardian*

Τι δεδομένα αναλύουμε συνήθως …



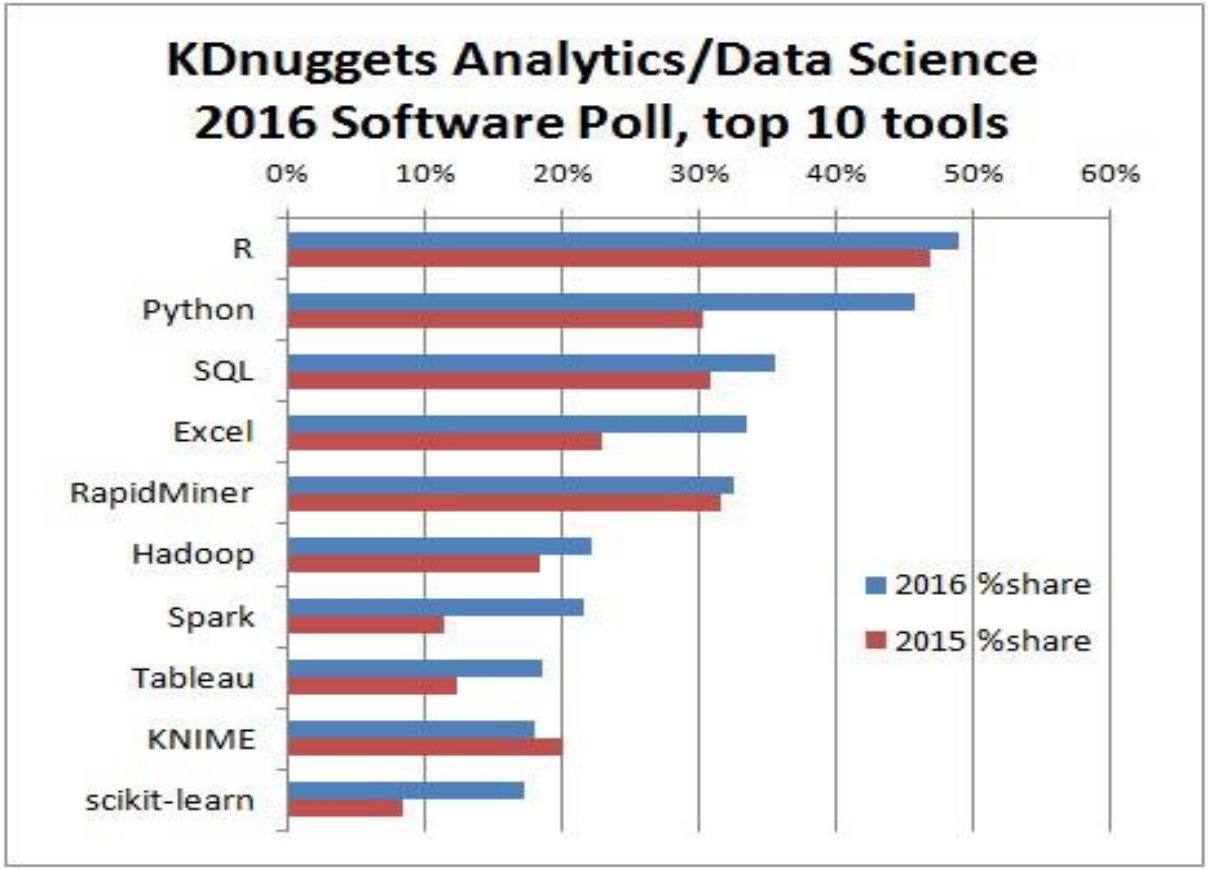
πηγή: kdnuggets.com

Με ποιες τεχνικές …



πηγή: kdnuggets.com

Με τι λογισμικό …



πηγή: kdnuggets.com

Βασικές έννοιες & γνώσεις

- **Γνωριμία με τα δεδομένα**
  - Προπαρασκευή δεδομένων
- **Τεχνικές και αλγόριθμοι**
  - Κατηγοριοποίηση / ταξινόμηση (classification)
  - Ανάλυση συστάδων (cluster analysis)
  - Μοντέλα προσαρμογής (regression)
  - Εξόρυξη συχνών προτύπων (frequent pattern mining)
- **Παραγόμενα αποτελέσματα**
  - Προτυποποίηση κώδικα
  - Βελτιστοποίηση διαδικασιών (process optimization)
  - Ενσωματωμένα συστήματα (IoT, edge, ...)

Μηχανική Μάθηση & Αναλυτική Δεδομένων

1. **Γνωριμία με τα δεδομένα**
   - Προπαρασκευή δεδομένων για αναλυτικές εργασίες
   - Αποθήκες δεδομένων και πολυδιάστατη ανάλυση
   - Οπτικοποίηση δεδομένων και εποπτική ανάλυση
2. **Τεχνικές και αλγόριθμοι**
   - Στατιστική ανάλυση δεδομένων
   - Τεχνικές Μηχανικής Μάθησης
   - Ειδικά θέματα
   - Ιδιωτικότητα δεδομένων

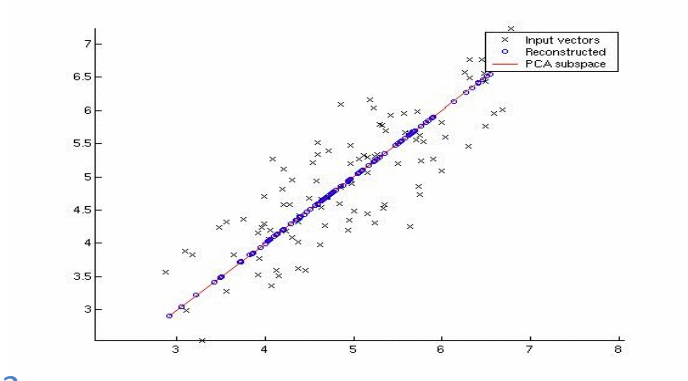
Τεχνικές Μηχανικής Μάθησης

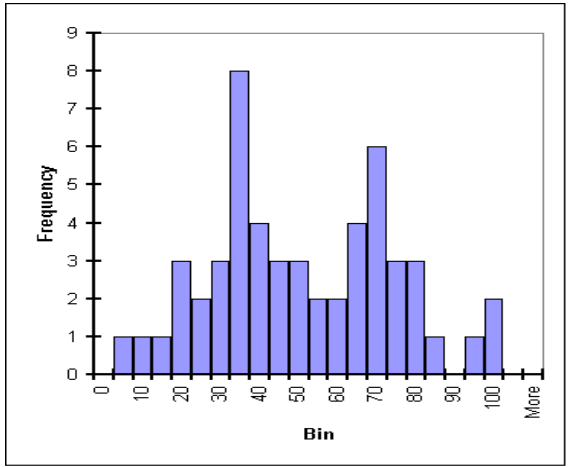
- **Επιβλεπόμενη μάθηση (classification, regression, ...)**
  - Επίβλεψη: τα δεδομένα εκπαίδευσης (παρατηρήσεις, μετρήσεις κλπ.) συνοδεύονται από *ετικέτες* (**labels**) που δηλώνουν την κλάση/κατηγορία στην οποία ανήκει η κάθε παρατήρηση
  - Τα νέα δεδομένα ταξινομούνται/κατηγοριοποιούνται βάσει του συνόλου εκπαίδευσης
- **Μη-επιβλεπόμενη μάθηση (clustering, association rules, ...)**
  - Δεν γνωρίζουμε ετικέτες κλάσεων/κατηγοριών
  - Με βάση κάποια δεδομένα (παρατηρήσεις, μετρήσεις κλπ.), στοχεύουμε στην *ανακάλυψη* κλάσεων ή ομάδων (συστάδων) μέσα σε αυτά

Προπαρασκευή δεδομένων

- **Καθαρισμός δεδομένων** (data cleansing / restoration)
  - Συμπλήρωση ελλιπών, εξομάλυνση θορύβου, εντοπισμός outliers, κανονικοποίηση
- **Μείωση όγκου δεδομένων** (data reduction)
  - Μείωση χαρακτηριστικών ή εγγραφών
  - Διακριτοποίηση δεδομένων

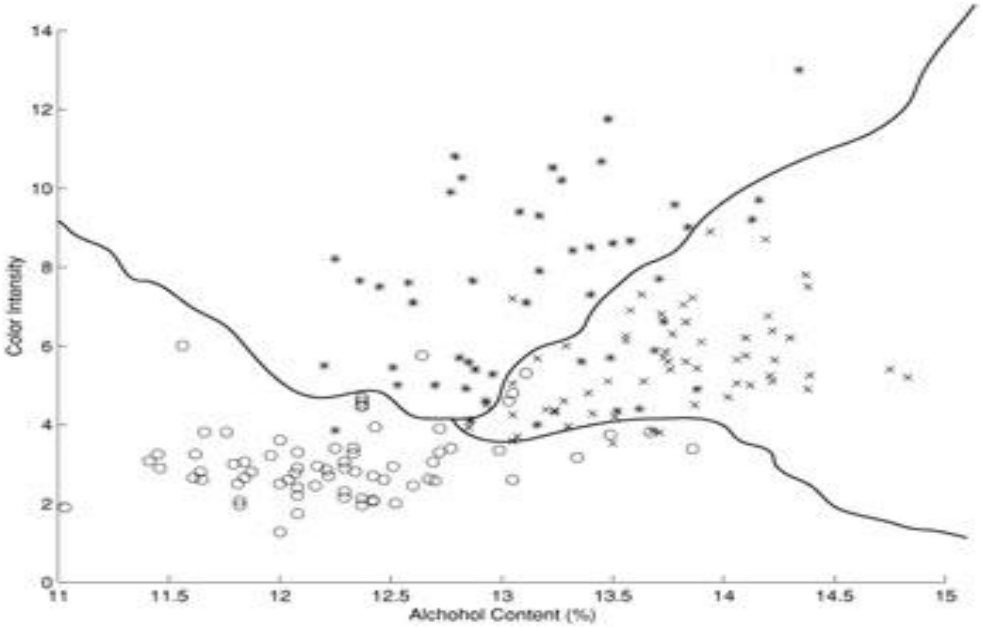
Humidity	Windy
70%	false
68%	true
80%	false
?	false
50%	false
45%	true
58%	?
65%	false
40%	false
0%	false
?	true

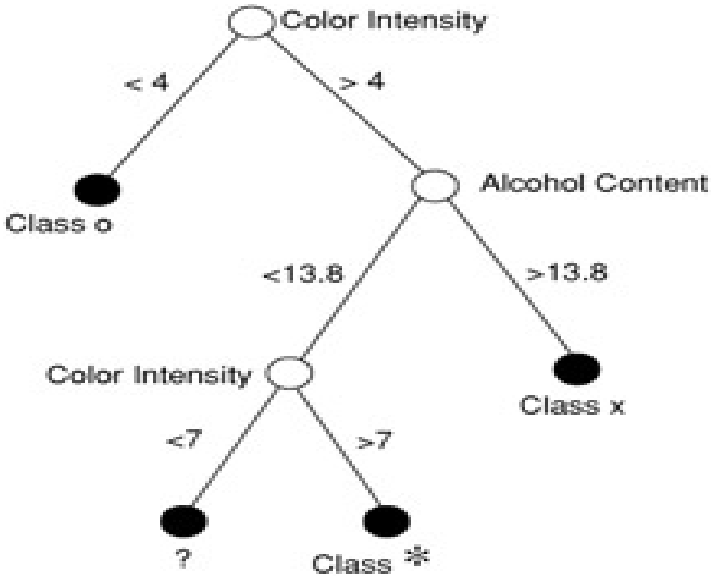




Κατηγοριοποίηση / Ταξινόμηση

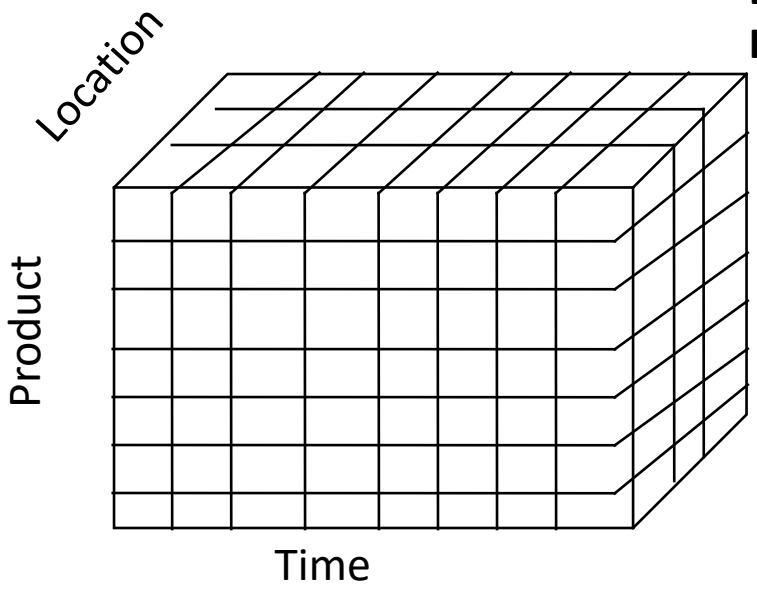
- Εκμάθηση μιας τεχνικής (δέντρο απόφασης, νευρωνικό δίκτυο κλπ.) να **προβλέπει** την κλάση ενός στοιχείου επιλέγοντας από προκαθορισμένες τιμές



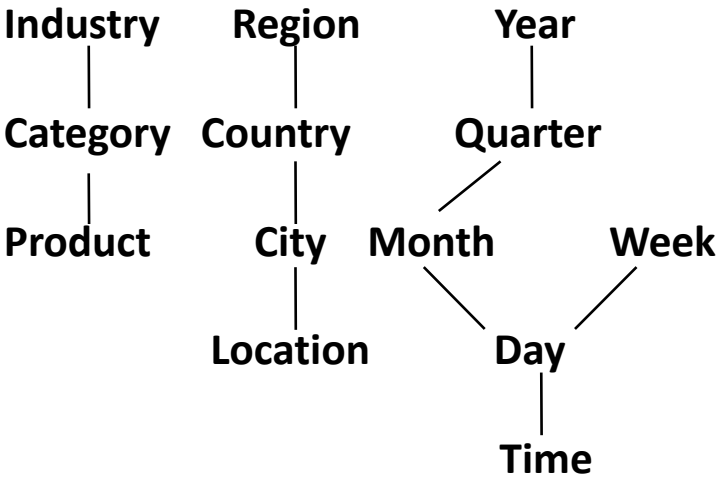


Πολυδιάστατη ανάλυση δεδομένων

- π.χ. πωλήσεις ως συνάρτηση προϊόντος / τοποθεσίας / χρόνου

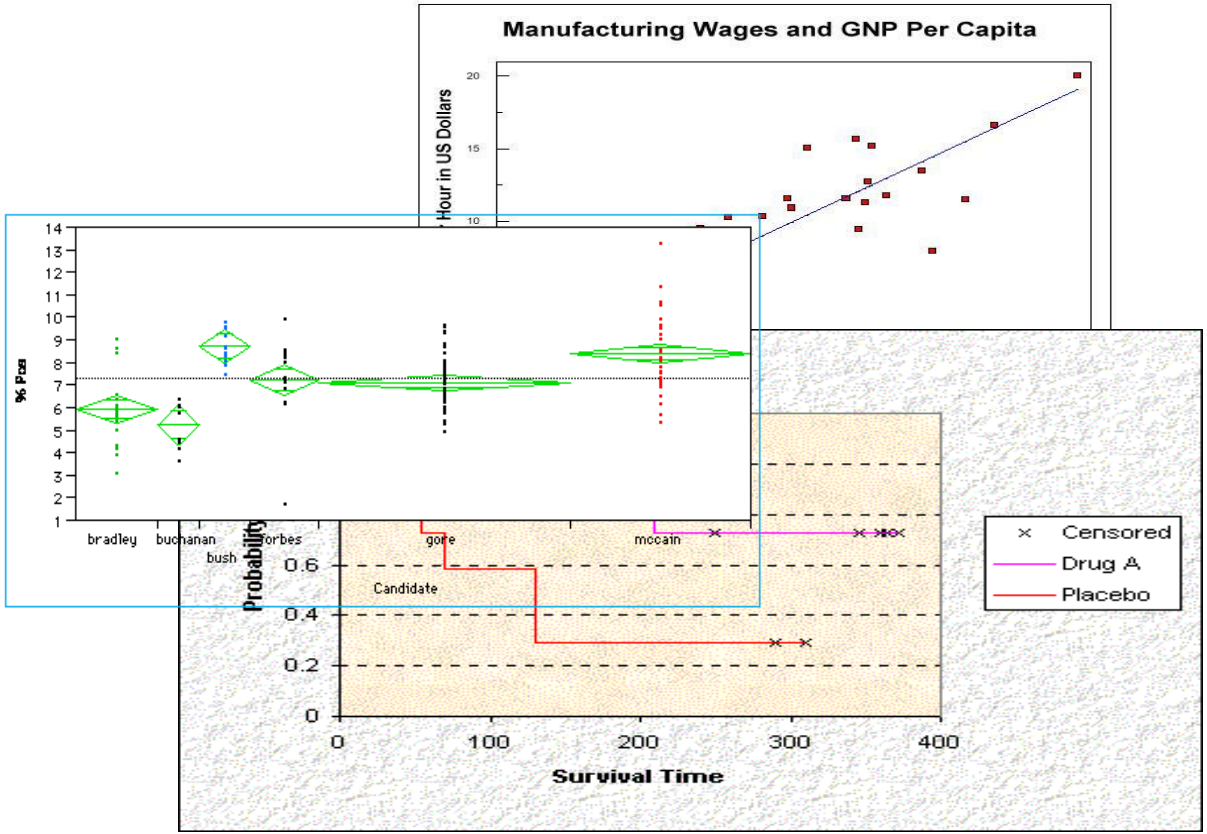


Dimensions: *Product, Location, Time*
Hierarchical summarization paths

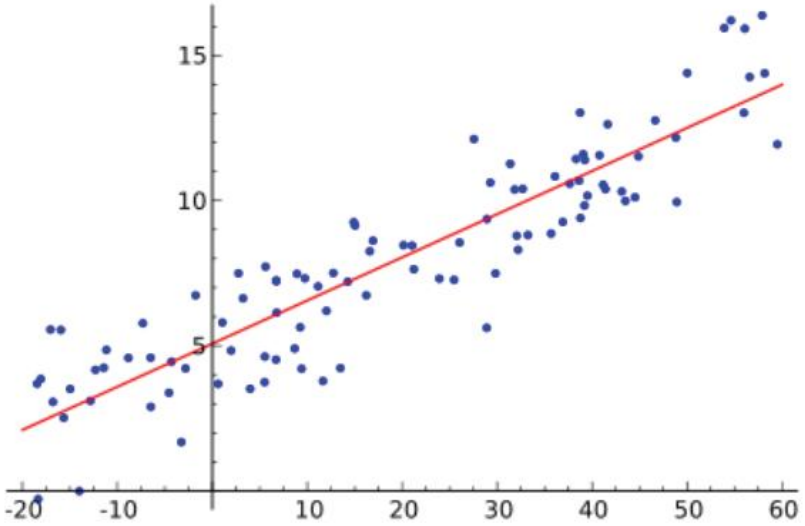


Ανάλυση & μοντελοποίηση δεδομένων

- **Regression**
- Generalized Linear Model
- Analysis of Variance
- Mixed-Effect Models
- Factor Analysis
- Discriminant Analysis
- …

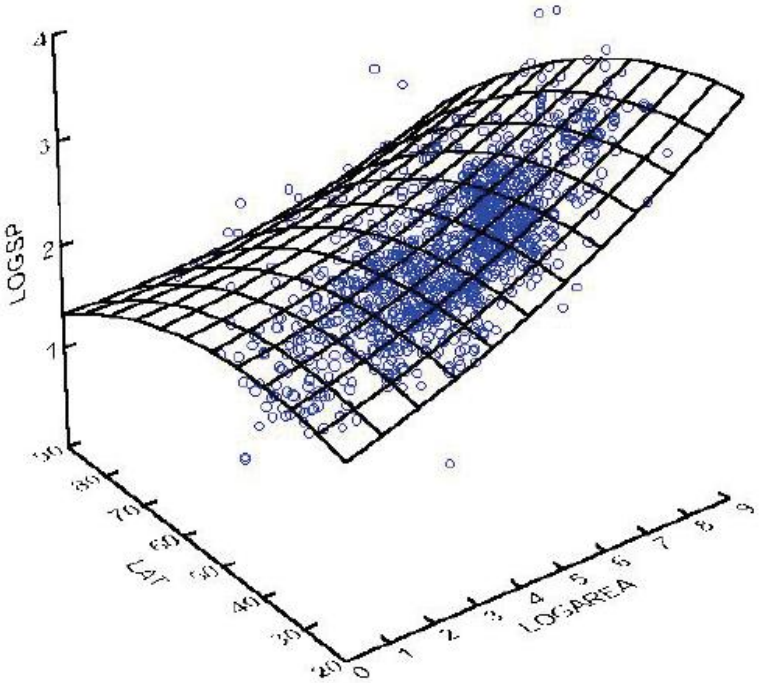


Regression models



simple linear X-Y
regression model

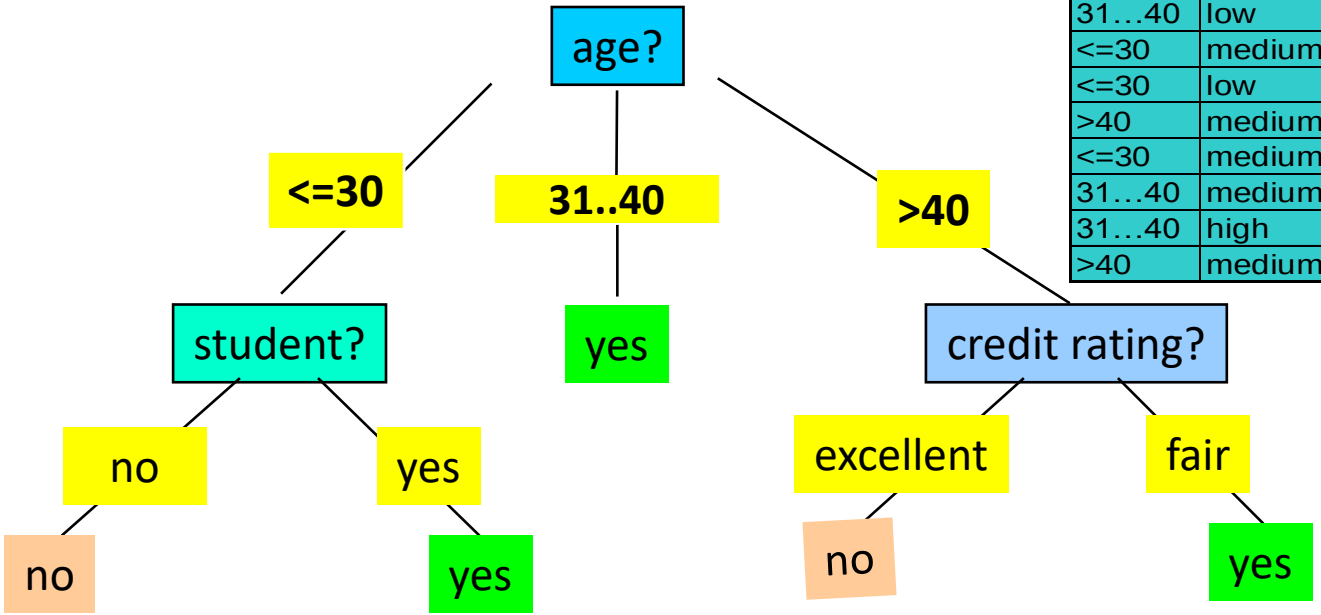
polynomial X-Y-Z
regression model



Παράδειγμα 1: Classification

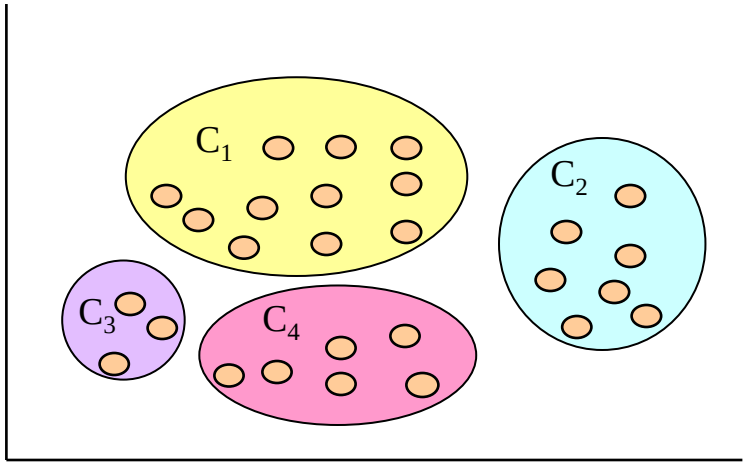
- ετικέτες: “Buys_computer” (yes/no)
- μοντέλο: δέντρο απόφασης (decision tree)

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



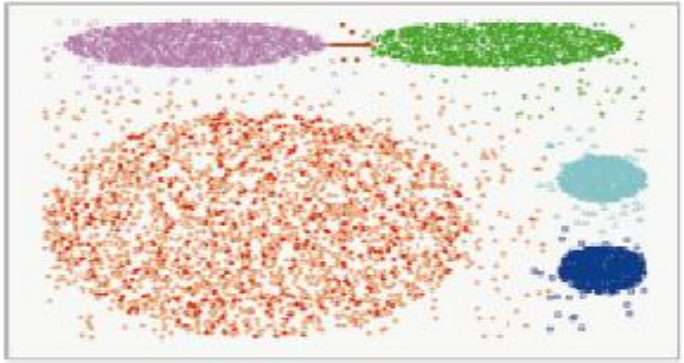
Ανάλυση Συστάδων / Ομαδοποίηση

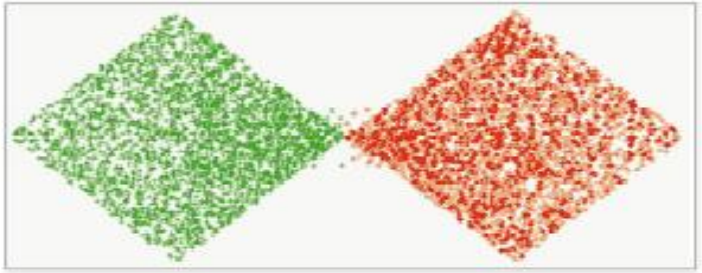
- Εύρεση μιας “φυσικής” ομαδοποίησης των δεδομένων, χωρίς προκαθορισμό των ομάδων

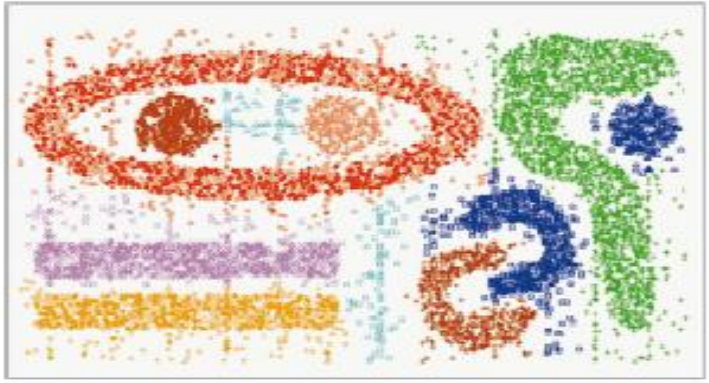


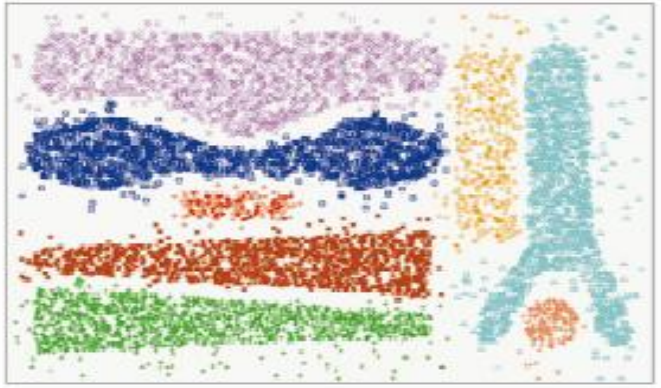
- με βάση κάποιο μέτρο (αν-)ομοιότητας
  - Πόσες / ποιες συστάδες;
  - Γιατί 4 και όχι π.χ. 3;

Ανάλυση Συστάδων / Ομαδοποίηση









Ανάλυση συχνών προτύπων

- **Εξόρυξη συχνών προτύπων (Frequent pattern mining)**: 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%

Major Tasks in Data Preprocessing

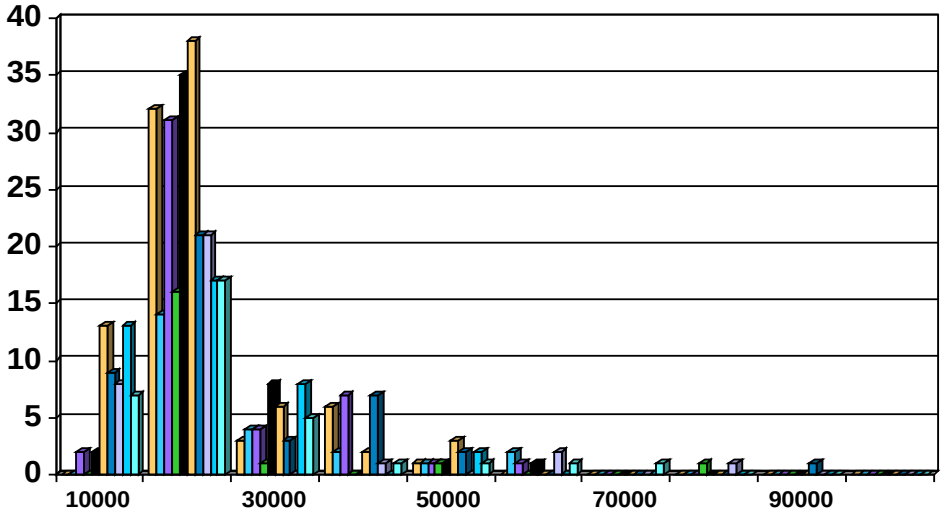
- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data fusion (multiple sensors)
- **Data reduction**
  - Dimensionality reduction (feature selection)
  - Numerosity reduction (resampling)
  - Data compression
- **Data transformation and data discretization**
  - Normalization, rescaling, domain change (e.g. freq.)
  - Concept hierarchy generation

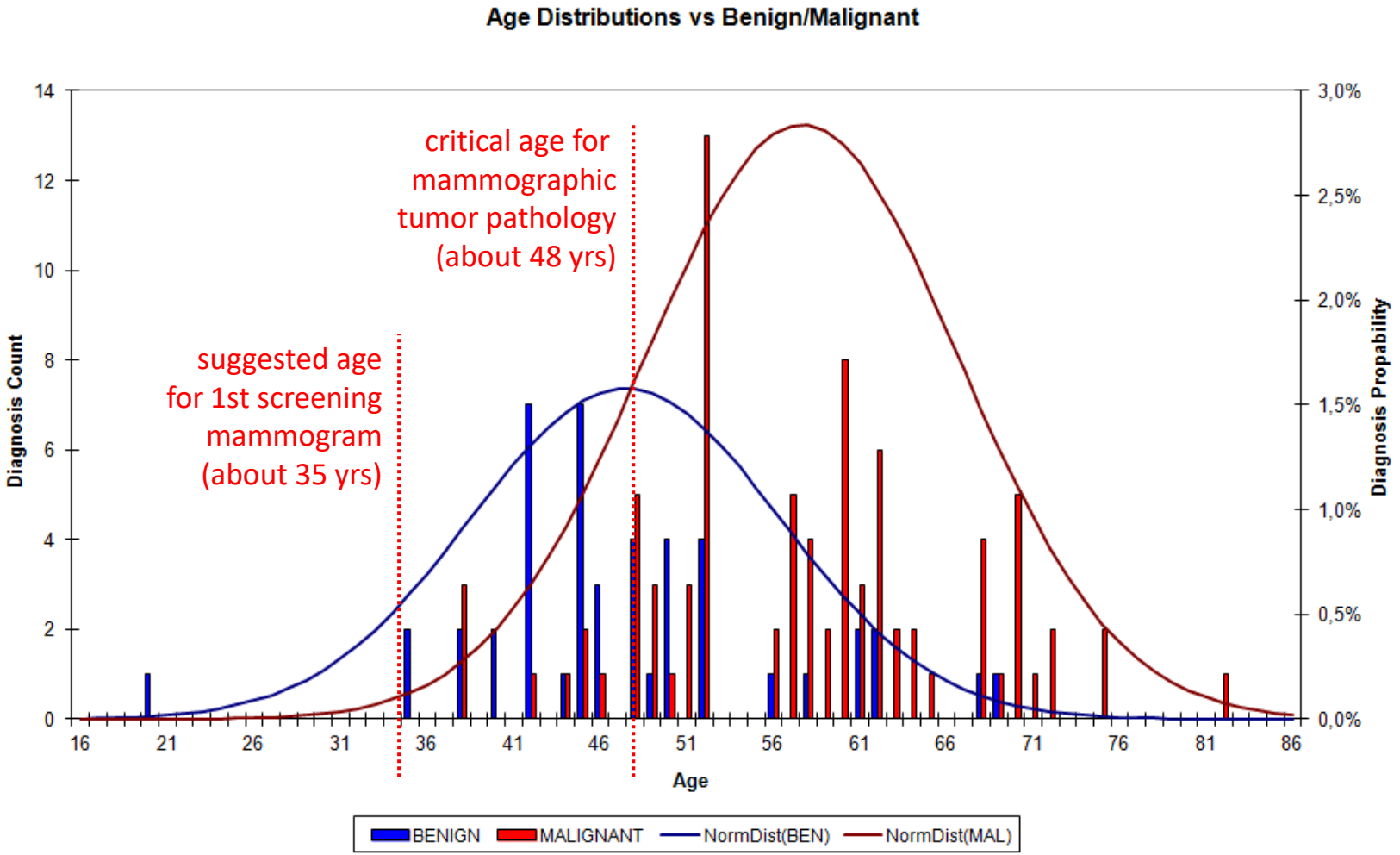
How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers / extremes
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible exceptional conditions that are valid but not for typical training)

Histogram Analysis

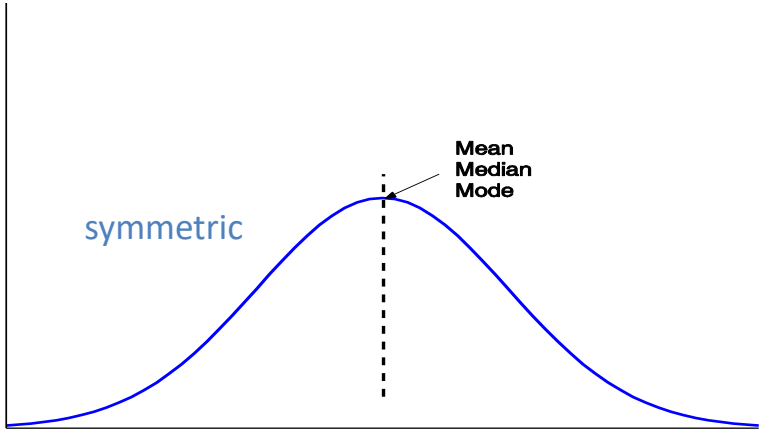
- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent

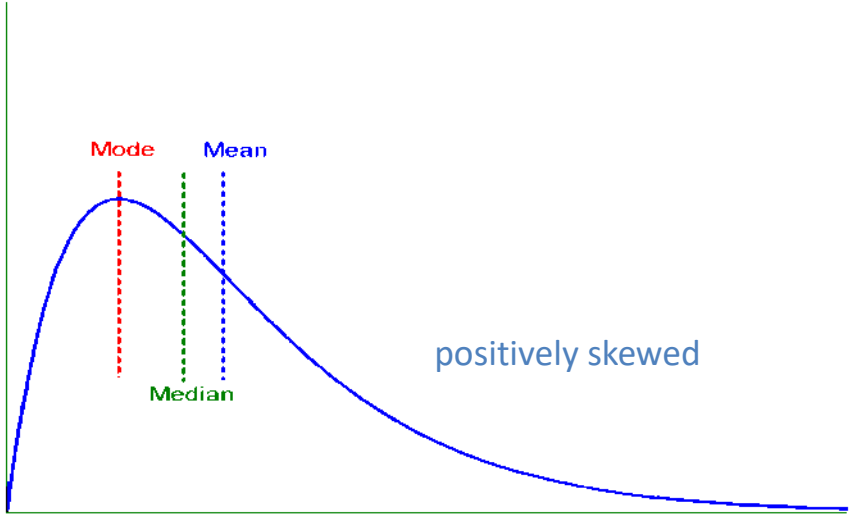


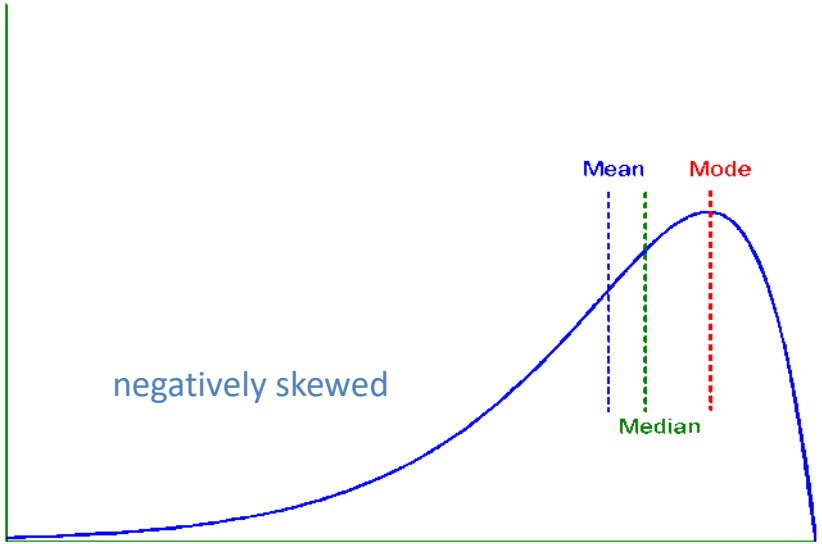

Age Distributions vs Benign/Malignant
critical age for mammographic tumor pathology (about 48 yrs)
suggested age for 1st screening mammogram (about 35 yrs)
Diagnosis Count
Diagnosis Propability
Age
BENIGN
MALIGNANT
NormDist(BEN)
NormDist(MAL)

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data

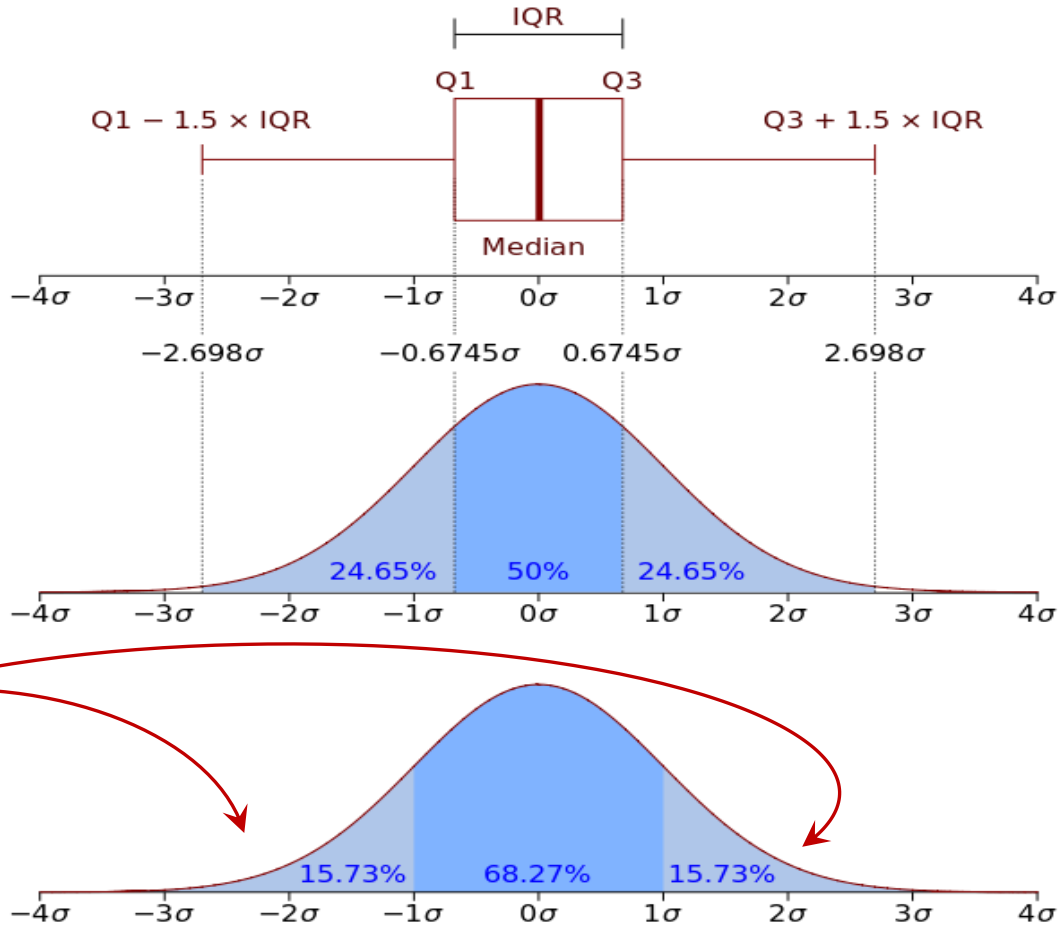






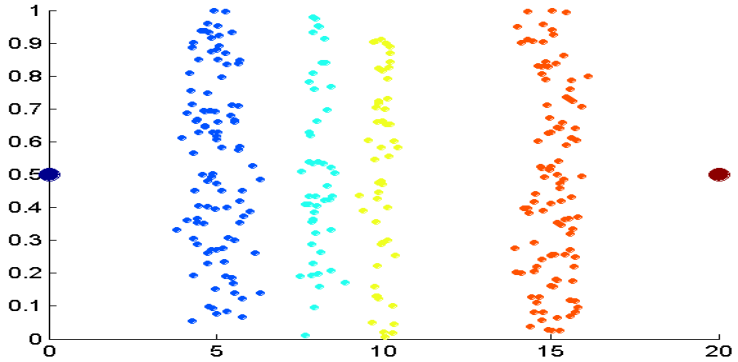
Properties of Normal Distribution Curve

- The normal (Gaussian) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it

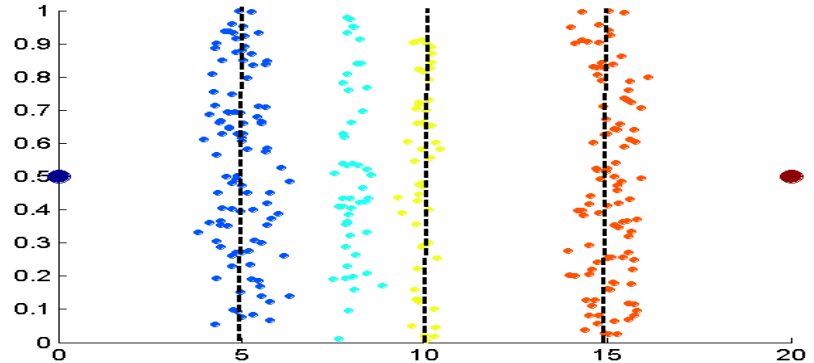


“outliers” (?)

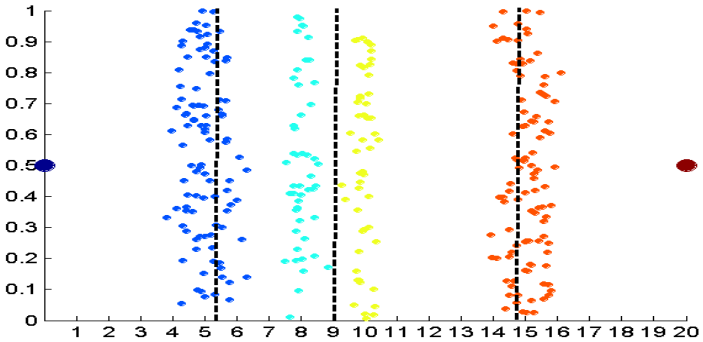
Discretization Without Using Class Labels (Binning vs. Clustering)



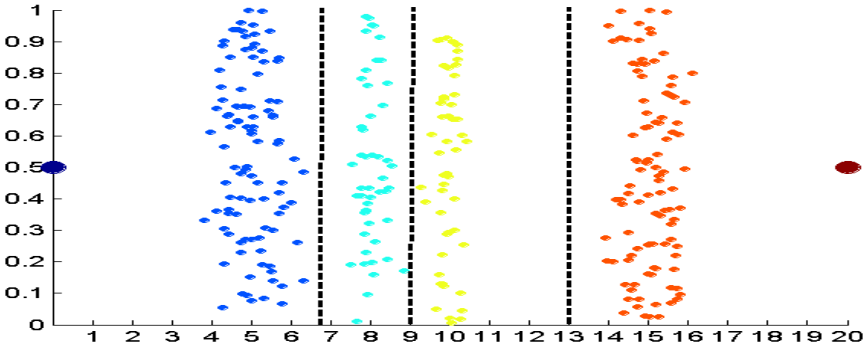
Original data



Equal depth (binning)

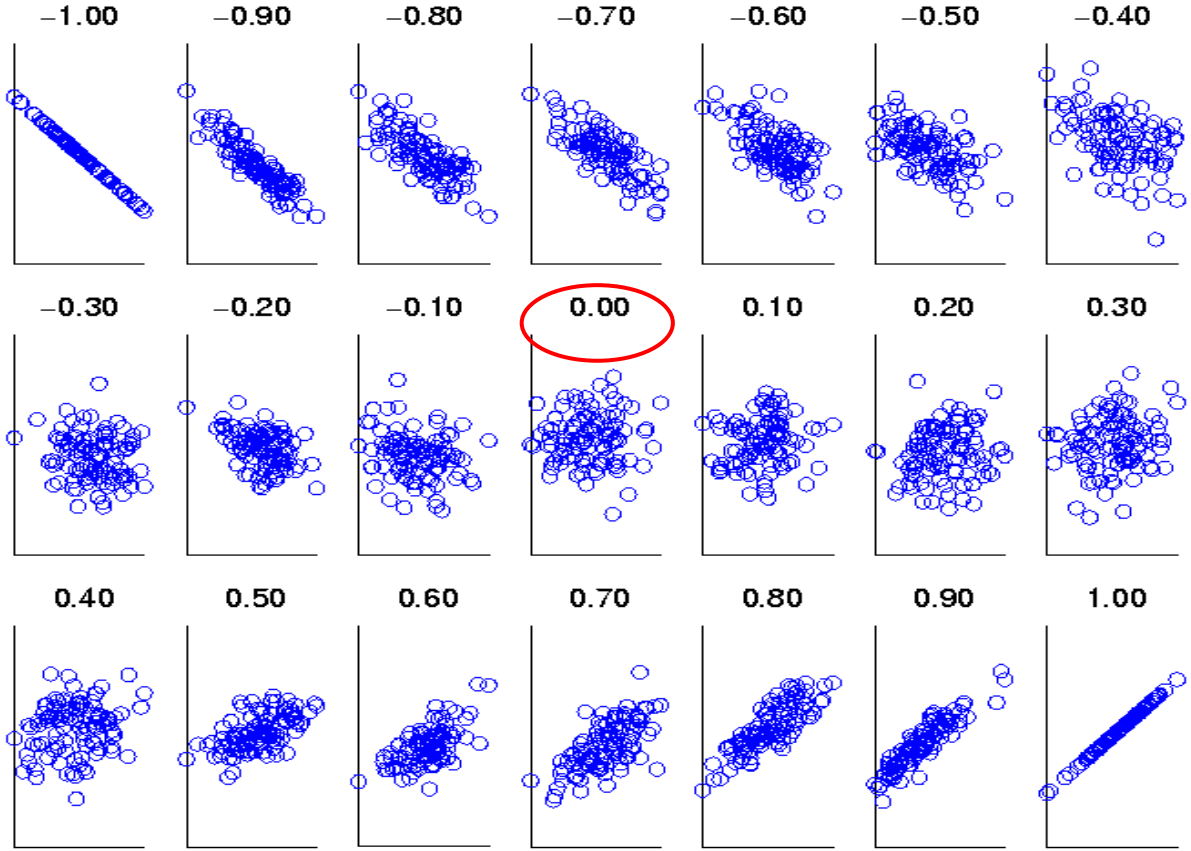


Equal frequency (binning)



K-means clustering leads to better results

Visually Evaluating Correlation



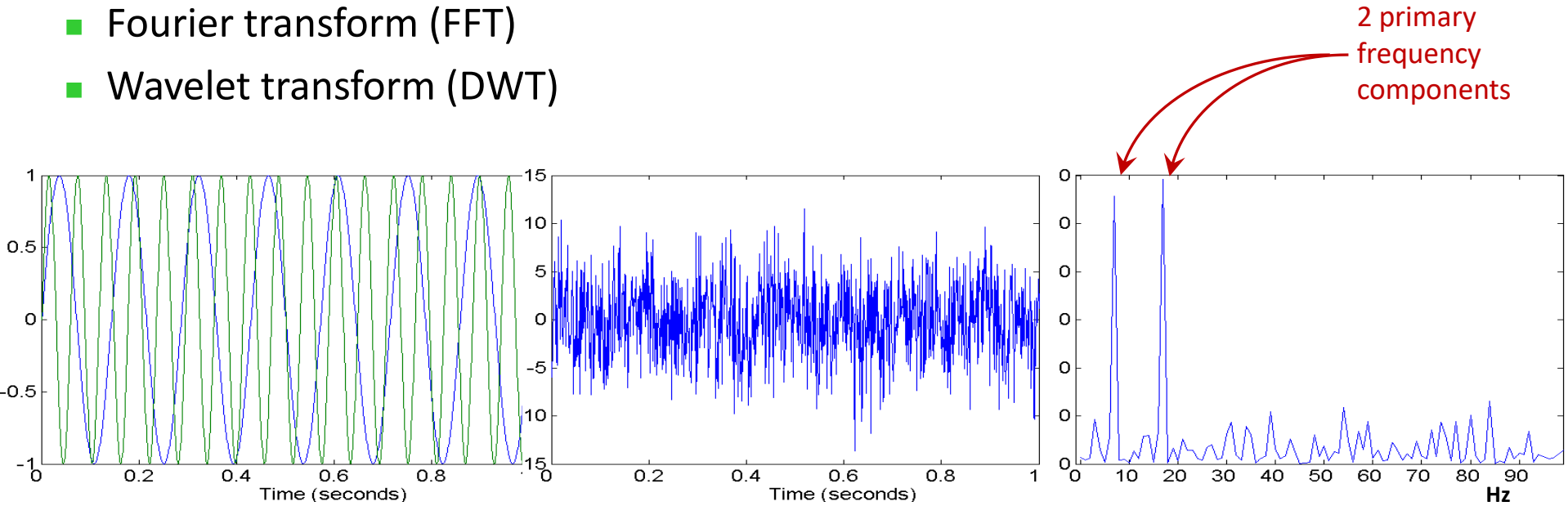
negative correlation

positive correlation

Scatter plots showing the similarity from –1 to 1

Mapping Data to a New Space

- Fourier transform (FFT)
- Wavelet transform (DWT)



Two Sine Waves

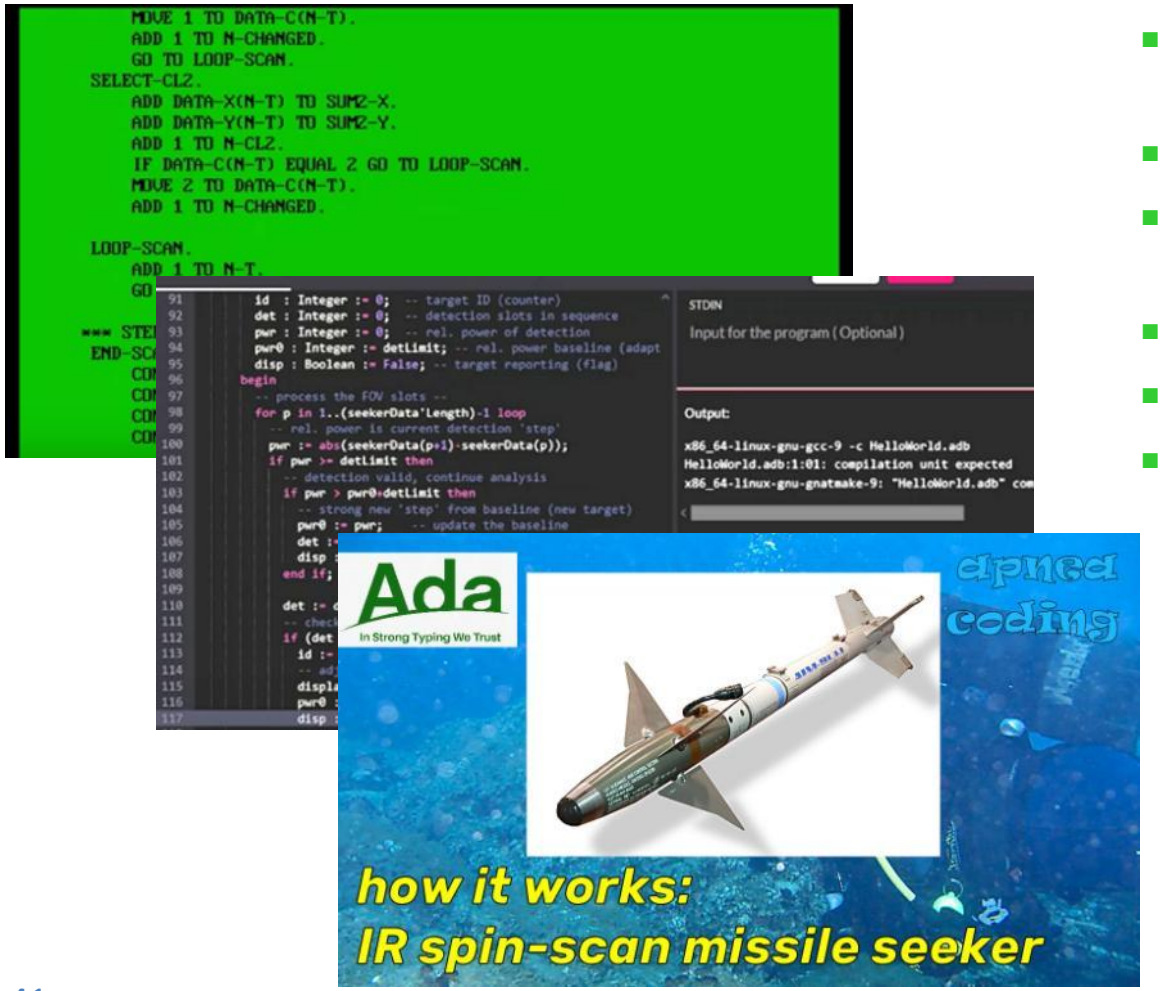
Two Sine Waves + Noise

Freq. Spectrum (FFT)

Σύνοψη

- Περιεχόμενα:
  - Τι είναι η Μηχανική Μάθηση και η Αναλυτική Δεδομένων (ML/DA)
  - Προπαρασκευή δεδομένων (pre-processing), είδη προβλημάτων ML/DA
  - Αλγόριθμοι: κατηγοριοποίηση, συσταδοποίηση, ανακάλυψη συχνών προτύπων , ...
  - Ειδικά θέματα (π.χ. δεδομένα ήχου, εικόνας, ιατρικά, ...)

- Πηγές:
  - «Αναλυτική Δεδομένων» – μάθημα ΠΜΣ Πανεπ. Πειραιά (σημειώσεις) 2017-2021.
  - Dunham: Data Mining – Introductory and Advanced Topics. Prentice Hall, 2003.
  - Tan, Steinbach, Kumar: Introduction to Data Mining. Addison Wesley, 2006.
  - Hand, Mannila, Smyth: Principles of Data Mining. MIT Press, 2001.



- Minimum Distance Classifier (MDC) in Ada
- Kmeans clustering in COBOL
- Bi-directional Associative Memory (BAM) in Arduino/C
- Linear Regression in SQL, Matlab
- k-nearest-neighbor Classifier in SQL
- ...

YouTube:

@ApneaCoding

https://www.youtube.com/@apneacoding

Github:

@xgeorgio

https://github.com/xgeorgio

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Comm. of ACM, 42:73-78, 1999
- A. Bruce, D. Donoho, and H.-Y. Gao. Wavelet analysis. *IEEE Spectrum*, Oct 1996
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- J. Devore and R. Peck. *Statistics: The Exploration and Analysis of Data*. Duxbury Press, 1997.
- H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita. Declarative data cleaning: Language, model, and algorithms. *VLDB'01*
- M. Hua and J. Pei. Cleaning disguised missing data: A heuristic approach. *KDD'07*
- H. V. Jagadish, et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), Dec. 1997
- H. Liu and H. Motoda (eds.). *Feature Extraction, Construction, and Selection: A Data Mining Perspective*. Kluwer Academic, 1998
- J. E. Olson. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, 2003
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB’2001
- T. Redman. *Data Quality: The Field Guide*. Digital Press (Elsevier), 2001
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering, 7:623-640, 1995

References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish, et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain,” Similarity measures”, IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu , et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009

Ερωτήσεις

Χάρης Γεωργίου (MSc,PhD)
https://www.linkedin.com/in/xgeorgio/
https://twitter.com/xgeorgio_gr