



# 1. Εισαγωγή



Data Scientist is still the Sexiest Job  
of the 21st Century  
[Harvard Business Review, Jul.2022]

## Ανάλυση Δεδομένων (Data Analytics)

Χρήστος Δουλκερίδης  
2024-25

# Περιεχόμενο Σημερινής Διάλεξης

## ■ Εισαγωγικά στοιχεία μαθήματος

### ■ Δεδομένα

- Τύποι δεδομένων
- Ποιότητα δεδομένων
- Προεπεξεργασία δεδομένων

### ■ Μέτρα απόστασης και ομοιότητας

- Ομοιότητα πολυδιάστατων αριθμητικών δεδομένων
- Ομοιότητα συνόλων
- Ομοιότητα αλφαριθμητικών

# Γνωριμία

- Διδάσκων: Χρήστος Δουλκερίδης ([cdoulk@unipi.gr](mailto:cdoulk@unipi.gr))
- Όρες γραφείου (γρ.502, Γρ.Λαμπράκη 126):
  - Πέμπτη 10:00 – 11:30, Παρασκευή 11:00 – 12:00
  - Και κατόπιν συνεννόησης
- Όρες μαθήματος:
  - Δευτέρα 08:15 – 11:00, ΚΕΚΤ 002
- Όρες εργαστηρίου:
  - Τετάρτη 12:00 – 14:00 & 14:00 – 16:00 (Εργ. Ισογείου)
  - Πέμπτη 13:00 – 15:00 & 15:00 – 17:00 (Εργ. Ημιωρόφου)
  - Υπεύθυνοι εργαστηρίου:
    - Καθ. Χ. Δουλκερίδης
    - Δρ. Κ. Μούτσελος
    - Δ. Πετράτος (ΥΔ)

# Επικοινωνία

- Ιστοσελίδα μαθήματος:
  - <https://aristarchus.ds.unipi.gr/courses/DS-COURSES-SEM122/>
- Θέματα και υλικό σχετικά με το μάθημα
  - ΔΙΑΛΕΞΕΙΣ
  - ΑΝΑΚΟΙΝΩΣΕΙΣ
  - ΕΚΦΩΝΗΣΕΙΣ ΕΡΓΑΣΙΩΝ
  - ΠΑΡΑΔΟΣΗ ΕΡΓΑΣΙΩΝ
  - ΕΡΓΑΣΤΗΡΙΟ
  - ...

# Ενότητες Μαθήματος [Εαρινό 2025]

1. Εισαγωγή (Δεδομένα και Μέτρα απόστασης/ομοιότητας)
2. Μονομεταβλητή και διμεταβλητή ανάλυση
3. Ανάλυση χρονοσειρών
4. Εισαγωγή στην προβλεπτική αναλυτική
5. Προσαρμογή μοντέλου σε δεδομένα
6. Το φαινόμενο της υπερπροσαρμογής και αποτίμηση προβλέψεων
7. Εντοπισμός συστάδων
8. Ανάλυση συσχέτισης – κανόνες συσχέτισης
9. Εύρεση σημαντικών γνωρισμάτων και μείωση διάστασης
10. Ανίχνευση ανωμαλιών
11. Πιθανοτικά μοντέλα και προσομοιώσεις

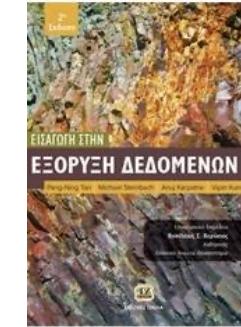
# Περιεχόμενα Εργαστηρίου

Εργαστήριο	Περιεχόμενο
#1	Εισαγωγή στην <i>Python</i> : Σύνταξη, βασικοί τύποι δεδομένων, βασικές δομές δεδομένων
#2	Η βιβλιοθήκη <i>NumPy</i> και υπολογιστικά μαθηματικά
#3	Διαχείριση αρχείων και εξερεύνηση φακέλων
#4	Εισαγωγή στην οπτικοποίηση δεδομένων – η βιβλιοθήκη <i>Matplotlib</i>
#5	Εισαγωγή στην βιβλιοθήκη <i>Pandas</i> και περιγραφική στατιστική
#6	Προεπεξεργασία δεδομένων και διερευνητική ανάλυση δεδομένων
#7	Παράδειγμα μηχανικής μάθησης (1/2): διερευνητική αναλυτική σε σύνολο δεδομένων
#8	Παράδειγμα μηχανικής μάθησης (2/2): κατασκευή προβλεπτικού μοντέλου παλινδρόμησης
#9	Συσταδοποίηση δεδομένων
#10	Εξέταση στο Εργαστήριο

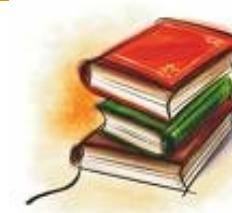
# Βιβλία Μαθήματος (1 / 2)



- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Anuj Karpatne, “*Εισαγωγή στην Εξόρυξη Δεδομένων*”, Εκδόσεις Τζιόλα, 2018.
- Foster Provost, Tom Fawcett, “*Η Επιστήμη των Δεδομένων για Επιχειρήσεις*”, Εκδόσεις Κλειδάριθμος, 2019.
- Joel Grus, “*Επιστήμη Δεδομένων: Βασικές Αρχές και Εφαρμογές με Python*”, 2<sup>η</sup> έκδοση, Εκδόσεις Παπασωτηρίου, 2021.

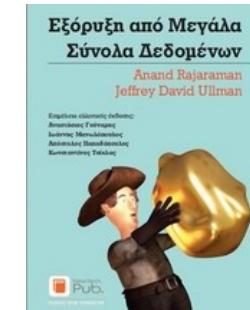


# Βιβλία Μαθήματος (2/2)



## ■ Επιπρόσθετα / Βοηθητικά Συγγράμματα:

- Philipp K. Janert, “*Data Analysis with Open Source Tools*”, O'Reilly press, 2011.
- Anand Rajaraman, Jeffrey David Ullman, “*Εξόρυξη από Μεγάλα Σύνολα Δεδομένων*”, Εκδόσεις Νέων Τεχνολογιών, 2013.
- Mohammed J. Zaki, Wagner Meira Jr, “*Εξόρυξη και Ανάλυση Δεδομένων: Βασικές Έννοιες και Αλγόριθμοι*” Εκδόσεις Κλειδάριθμος, 2017.



# Αξιολόγηση

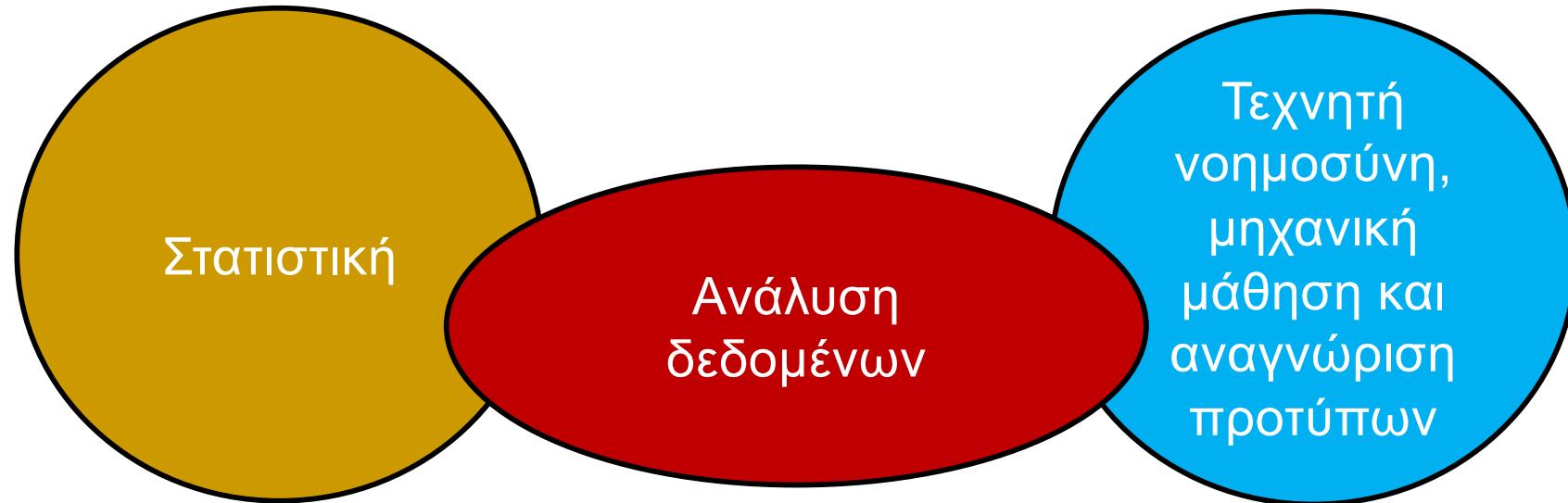
- Γραπτή Εξέταση ( $\Gamma$ )
- 2 Τεστ Εργαστηρίου ( $T1, T2$ )
  - Υπό τη μορφή προόδου στην ύλη του εργαστηρίου

$$\text{Βαθμός} = \begin{cases} (70\%) \times \Gamma + (10\%) \times T1 + (20\%) \times T2 & , \text{ αν: } \Gamma \geq 4 \\ (70\%) \times \Gamma & , \text{ αν: } \Gamma < 4 \end{cases}$$

Οι βαθμοί των Τεστ  $T1$  και  $T2$  ισχύουν **για το τρέχον ακαδημαϊκό έτος 2024-25**

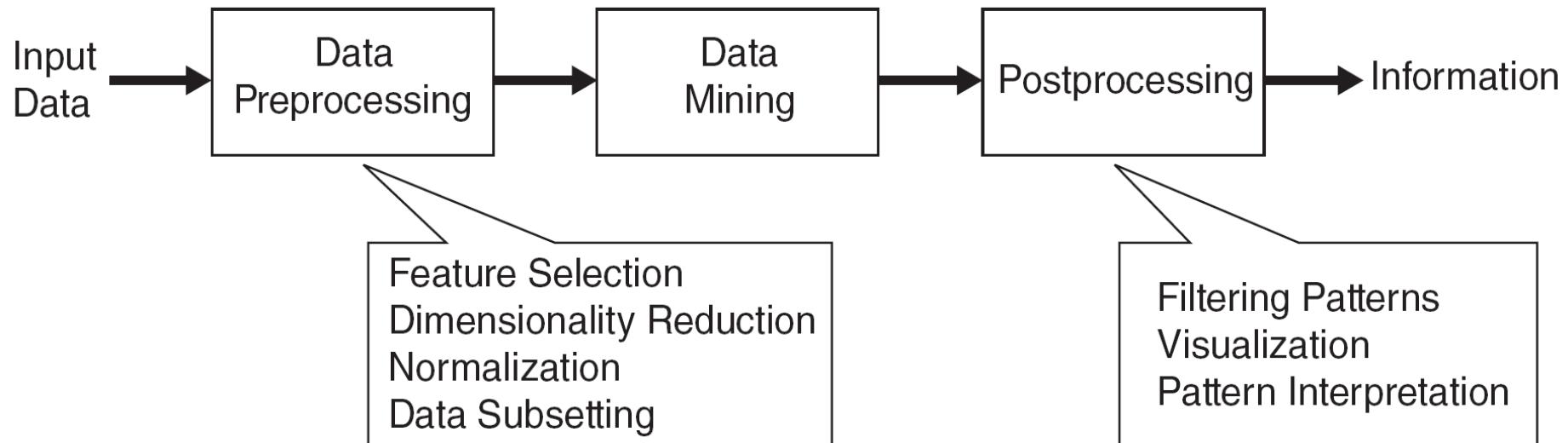
- και για την επαναληπτική εξεταστική Σεπτεμβρίου
- και για τυχόν εμβόλιμη εξεταστική

# Σχέση Ανάλυσης Δεδομένων με Άλλα Επιστημονικά Πεδία



Τεχνολογία βάσεων δεδομένων, παράλληλος υπολογισμός,  
κατανεμημένος υπολογισμός

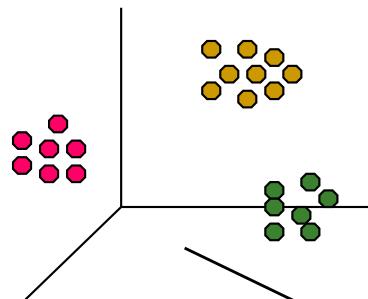
# Η Διαδικασία Εξόρυξης Γνώσης



- Ενδιαφέρουσες προκλήσεις:
  - Κλιμάκωση (scalability)
  - Πολλές διαστάσεις (high dimensionality)
  - Ετερογενή και πολύπλοκα δεδομένα (heterogeneous and complex data)
  - Κυριότητα και διανομή των δεδομένων (data ownership and distribution)
  - Μη παραδοσιακή ανάλυση (non-traditional analysis)

# Εργασίες Ανάλυσης Δεδομένων

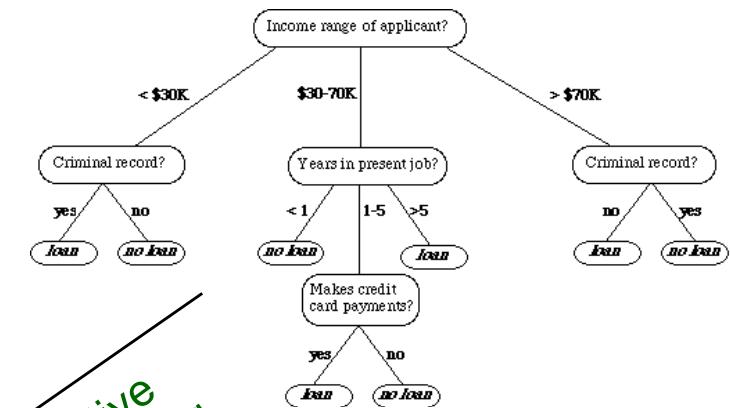
## Data



Clustering

Tid	Home Owner	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Association Rules



Predictive  
Modeling

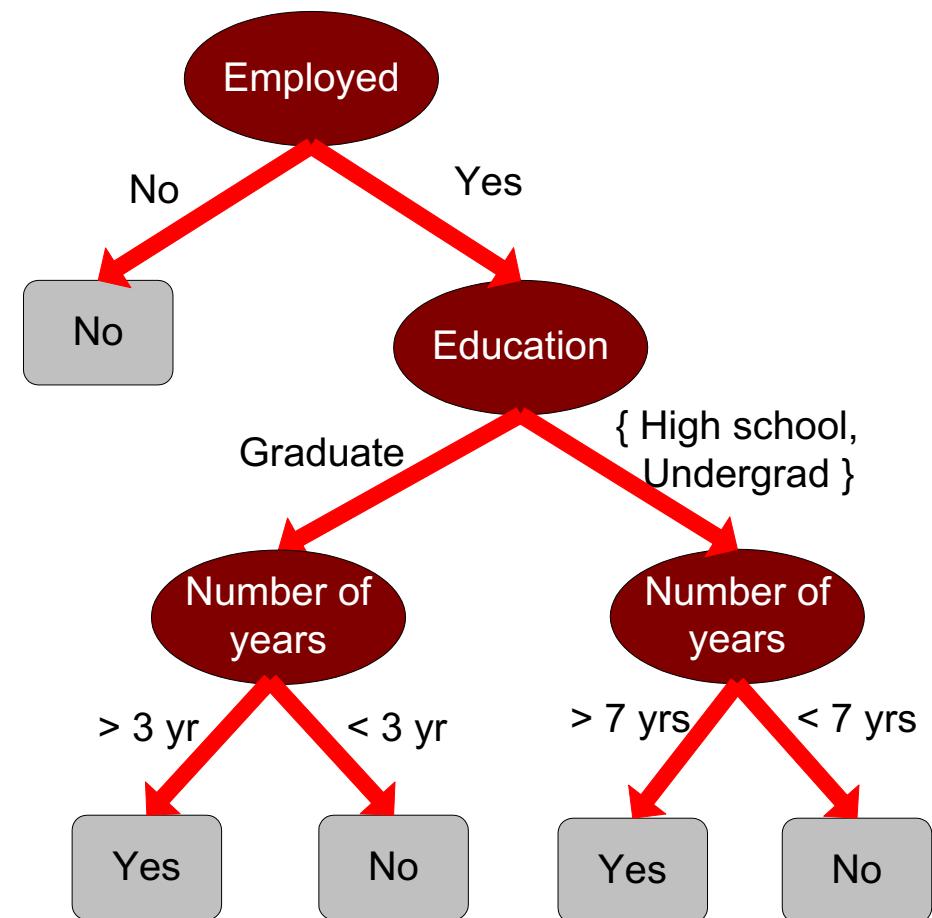
Anomaly  
Detection

# Προβλεπτική Αναλυτική/Μοντελοποίηση

Εύρεση ενός μοντέλου για το γνώρισμα class ως συνάρτηση των υπόλοιπων γνωρισμάτων

Class					
Tid	Employed	Level of Education	# years at present address	Credit Worthy	
1	Yes	Graduate	5	Yes	
2	Yes	High School	2	No	
3	No	Undergrad	1	No	
4	Yes	High School	10	Yes	
...	...	...	...	...	

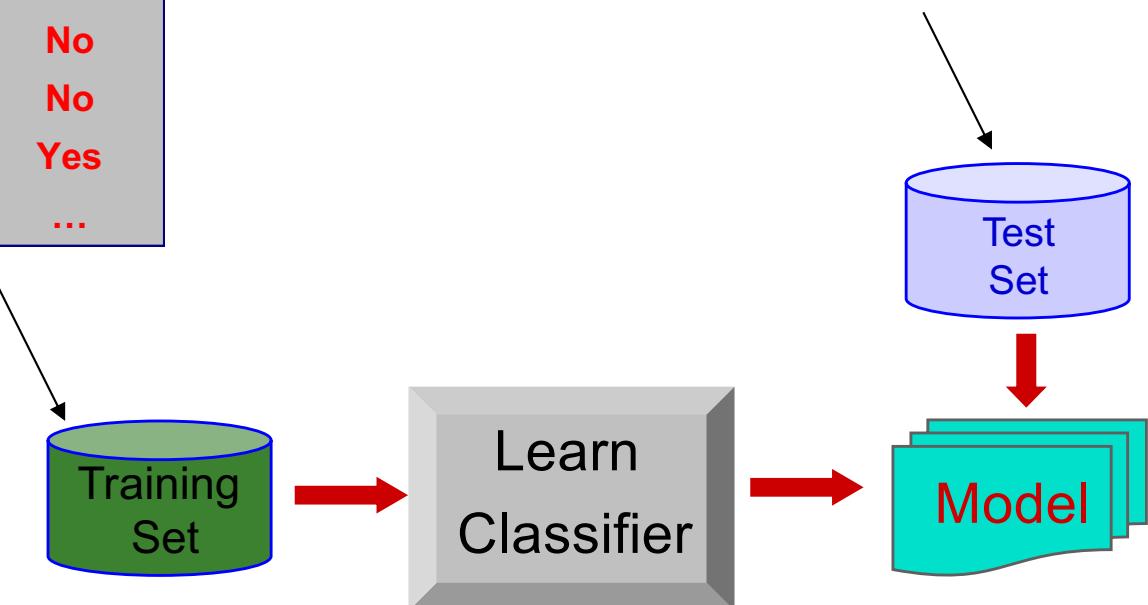
Model for predicting credit worthiness



# Παράδειγμα Κατηγοριοποίησης

class				
<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...	...	...	...	...

<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
...	...	...	...	...



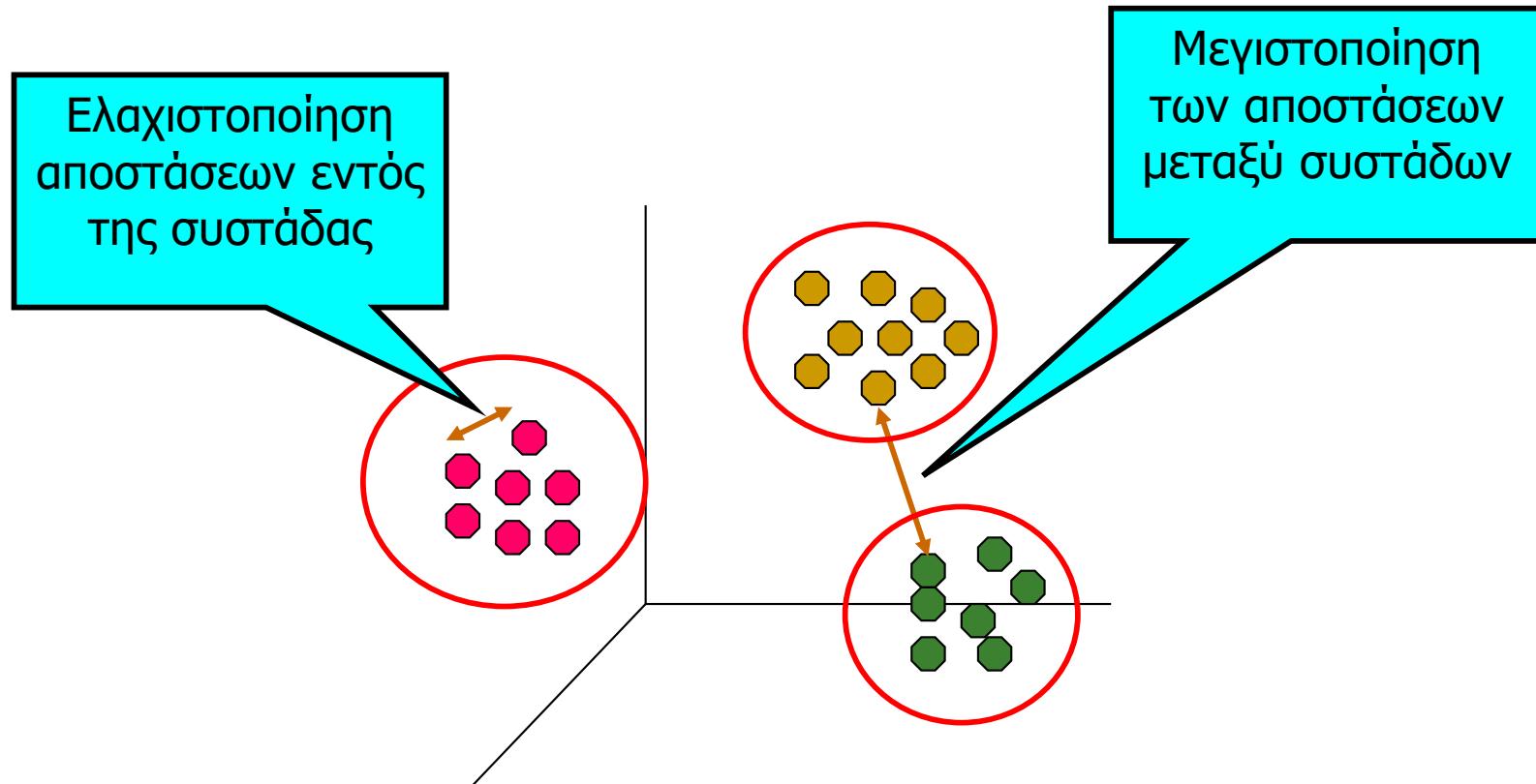
# Εφαρμογές Κατηγοριοποίησης

- Κατηγοριοποίηση κίνησης πιστωτικής κάρτας ως «σωστή» ή «απάτη»
- Κατηγοριοποίηση ειδήσεων ως οικονομικές, πολιτικές, αθλητικές, κτλ.
- Κατηγοριοποίηση e-mail ως spam
- Εντοπισμός εισβολέων στον κυβερνοχώρο
- Πρόβλεψη φύσης όγκων (καλοήθεις/κακοήθεις)
- ...

# Παράδειγμα Συσταδοποίησης

Εύρεση ομάδων αντικειμένων, έτσι ώστε:

- τα αντικείμενα σε κάθε ομάδα να είναι όμοια (ή σχετιζόμενα) μεταξύ τους,
- και ταυτόχρονα διαφορετικά από (ή άσχετα με) τα αντικείμενα άλλων ομάδων



# Εφαρμογές Συσταδοποίησης

- **Κατανόηση**
  - Κατάτμηση πελατών σε ομάδες με κοινό προφίλ
  - Δημιουργία προφίλ πελατών για προσωποποιημένο μάρκετινγκ
  - Ομαδοποίηση ιστοσελίδων για πλοήγηση (βλ. μηχανές αναζήτησης)
  - Ομαδοποίηση γονιδίων και πρωτεϊνών με όμοια λειτουργικότητα
  - Ομαδοποίηση μετοχών με παρόμοια συμπεριφορά στη μεταβολή τιμής
- **Περίληψη – συμπίεση συνόλων δεδομένων**
  - Μείωση μεγέθους μεγάλων συνόλων δεδομένων

# Περιεχόμενο Σημερινής Διάλεξης

- Εισαγωγικά στοιχεία μαθήματος
- **Δεδομένα**
  - Τύποι δεδομένων
  - Ποιότητα δεδομένων
  - Προεπεξεργασία δεδομένων
- Μέτρα απόστασης και ομοιότητας
  - Ομοιότητα πολυδιάστατων αριθμητικών δεδομένων
  - Ομοιότητα συνόλων
  - Ομοιότητα αλφαριθμητικών

# Τύποι Δεδομένων

- Ένα **σύνολο δεδομένων** (**data set**) μπορεί να θεωρηθεί ως μια συλλογή **αντικειμένων δεδομένων** (**data objects**)
- Εναλλακτικές ονομασίες για **αντικείμενο δεδομένων**:
  - Εγγραφή (record), σημείο (point), διάνυσμα (vector), πρότυπο (pattern), γεγονός (event), περίπτωση (case), δείγμα (sample), παρατήρηση (observation) ή οντότητα (entity)
- Τα αντικείμενα δεδομένων περιγράφονται από ένα πλήθος **χαρακτηριστικών** (**attributes**) που το περιγράφουν
  - Παράδειγμα: η μάζα ενός φυσικού αντικειμένου, ή ο χρόνος που συνέβη ένα γεγονός
- Εναλλακτικές ονομασίες για **χαρακτηριστικό**:
  - Μεταβλητή (variable), πεδίο (field), γνώρισμα (feature) ή διάσταση (dimension)

# Παράδειγμα

Χαρακτηριστικά

Κωδικός φοιτητή	Έτος	Μέσος όρος	...
1034262	Τελειόφοιτος	8.29	...
1052663	Δευτεροετής	6.12	...
1082246	Πρωτοετής	7.53	...

Αντικείμενα δεδομένων  
(εγγραφές)

# Τύποι Χαρακτηριστικών

Τύπος	Περιγραφή	Παραδείγματα	
Κατηγορικά (Ποιοτικά)	<b>Ονομαστικό (nominal)</b>	Απλώς διαφορετικές τιμές (=, ≠)	Ταχυδρομικοί κώδικες, αναγνωριστικά υπαλλήλων, χρώμα ματιών, φύλο
	<b>Τακτικό (ordinal)</b>	Είναι δυνατή η ταξινόμηση των αντικειμένων (<,>)	Σκληρότητα ορυκτών, βαθμοί, αριθμοί οδών
Αριθμητικά (Ποσοτικά)	<b>Διαστημάτων (interval)</b>	Οι διαφορές μεταξύ των τιμών έχουν σημασία (+,-)	Ημερομηνίες ημερολογίων, θερμοκρασία σε Κελσίου ή Φαρενάιτ
	<b>Αναλογιών (ratio)</b>	Τόσο οι διαφορές όσο και οι αναλογίες έχουν σημασία (*,/)	Θερμοκρασία Κέλβιν, νομισματικές ποσότητες, μετρήσεις, ηλικία, μάζα

**Κάθε τύπος χαρακτηριστικού έχει όλες τις ιδιότητες των τύπων που είναι πάνω από αυτόν**

# Περιγραφή Χαρακτηριστικών από ένα Πλήθος Τιμών

- **Διακριτό (discrete) χαρακτηριστικό:** έχει πεπερασμένο ή μετρήσιμα απεριόριστο σύνολο τιμών
  - **Παράδειγμα:** χρώμα αυτοκινήτου = {κόκκινο, πράσινο, μπλε, ...}
  - Τα **δυαδικά** χαρακτηριστικά (**binary** attributes) είναι μια ειδική περίπτωση των διακριτών χαρακτηριστικών και θεωρούν δεδομένη την ύπαρξη δύο μόνο τιμών
- **Συνεχές (continuous) χαρακτηριστικό:** λαμβάνει ως τιμές πραγματικούς αριθμούς
  - **Παράδειγμα:** θερμοκρασία = {37.3, 38.1, 38.4, ...}

# Γενικά Χαρακτηριστικά των Συνόλων Δεδομένων

- **Διάσταση (dimensionality)**: το πλήθος των χαρακτηριστικών που περιέχουν τα αντικείμενα του συνόλου
- **Σποραδικότητα (sparsity)**: όταν τα πιο πολλά χαρακτηριστικά έχουν μηδενικές τιμές
- **Ανάλυση (resolution)**: μπορεί να υπάρχουν δεδομένα διαφορετικών επιπέδων ανάλυσης
  - **Παράδειγμα**: η επιφάνεια της γης φαίνεται πολύ ανώμαλη σε μια ανάλυση λίγων μέτρων, αλλά είναι ιδιαίτερα ομαλή σε μια ανάλυση δεκάδων χιλιομέτρων

# Τύποι Συνόλων Δεδομένων

- Δεδομένα εγγραφών
  - Δεδομένα συναλλαγών ή καλαθιού αγοράς
  - Μήτρα δεδομένων
  - Μήτρα σποραδικών δεδομένων (document-term matrix)
- Δεδομένα γράφων
  - Δεδομένα με σχέσεις μεταξύ αντικειμένων (συνδεδεμένες ιστοσελίδες)
  - Δεδομένα αντικειμένων που είναι γράφοι (μόρια βενζίνης)
- Διατεταγμένα δεδομένα
  - Ακολουθιακά δεδομένα (υπάρχει η έννοια του χρόνου, χρονικά δεδομένα)
  - Δεδομένα ακολουθίας (γονίδια)
  - Δεδομένα χρονικών σειρών (χρονικές μετρήσεις)
  - Χωρικά δεδομένα

# Δεδομένα Εγγραφών (Records)

- Δεδομένα που αποτελούν μια συλλογή εγγραφών
- Κάθε εγγραφή περιγράφεται από σταθερό σύνολο χαρακτηριστικών
- Γνωστά και ως *tabular data*

Tid	Home Owner	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

# Μήτρα Δεδομένων

- Αν τα δεδομένα έχουν το ίδιο σταθερό σύνολο αριθμητικών χαρακτηριστικών, τότε μπορούν να αναπαρασταθούν ως **σημεία** σε ένα **πολυδιάστατο χώρο**, όπου κάθε διάσταση είναι ένα χαρακτηριστικό
- **Μ γραμμές** – μια για κάθε αντικείμενο και
- **Ν στήλες** – μια για κάθε χαρακτηριστικό

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

# Δεδομένα Εγγράφων (Documents)

- Κάθε **έγγραφο** αναπαρίσταται ως ένα **διάνυσμα** όρων
  - Κάθε όρος είναι ένα χαρακτηριστικό του διανύσματος
  - Η τιμή κάθε όρου είναι ο αριθμός των φορών που ο συγκεκριμένος όρος εμφανίζεται μέσα στο έγγραφο

	team	coach	play	ball	score	game	wi	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

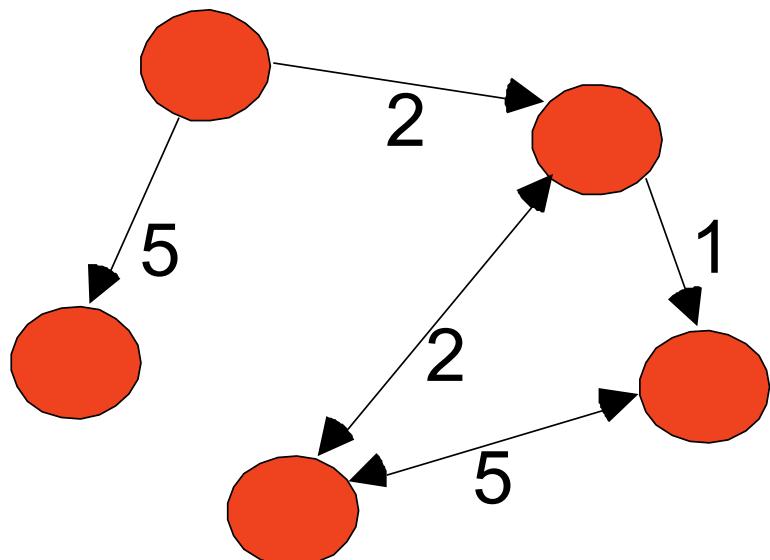
# Δεδομένα Συναλλαγών

- Ειδικός τύπος δεδομένων εγγραφών
  - Κάθε εγγραφή (συναλλαγή) εμπεριέχει ένα σύνολο στοιχείων

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

# Δεδομένα Γράφων

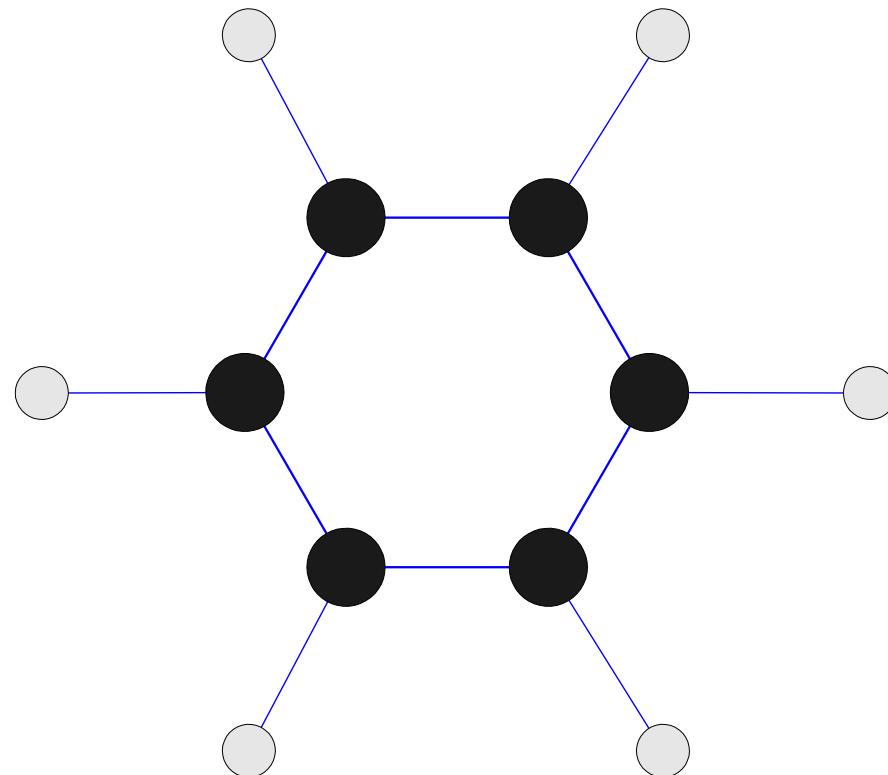
- Παραδείγματα:
  - Γενικευμένοι γράφοι και HTML υπερσύνδεσμοι



```
<a href="papers/papers.html#bbbb">  
Data Mining </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Graph Partitioning </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Parallel Solution of Sparse Linear System of Equations </a>  
<li>  
<a href="papers/papers.html#ffff">  
N-Body Computation and Dense Linear System Solvers
```

# Επιστημονικά Δεδομένα

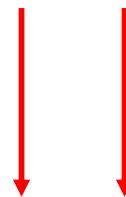
- Το μόριο της βενζίνης:  $C_6H_6$



# Διατεταγμένα Δεδομένα

## ■ Ακολουθίες συναλλαγών

Items/Events



( A B) (D) (C E)

( B D) (C) (E)

( C D) (B) (A E)



An element of  
the sequence

# Διατεταγμένα Δεδομένα

## ■ Διατάξεις νουκλεοτιδίων (γονίδια)

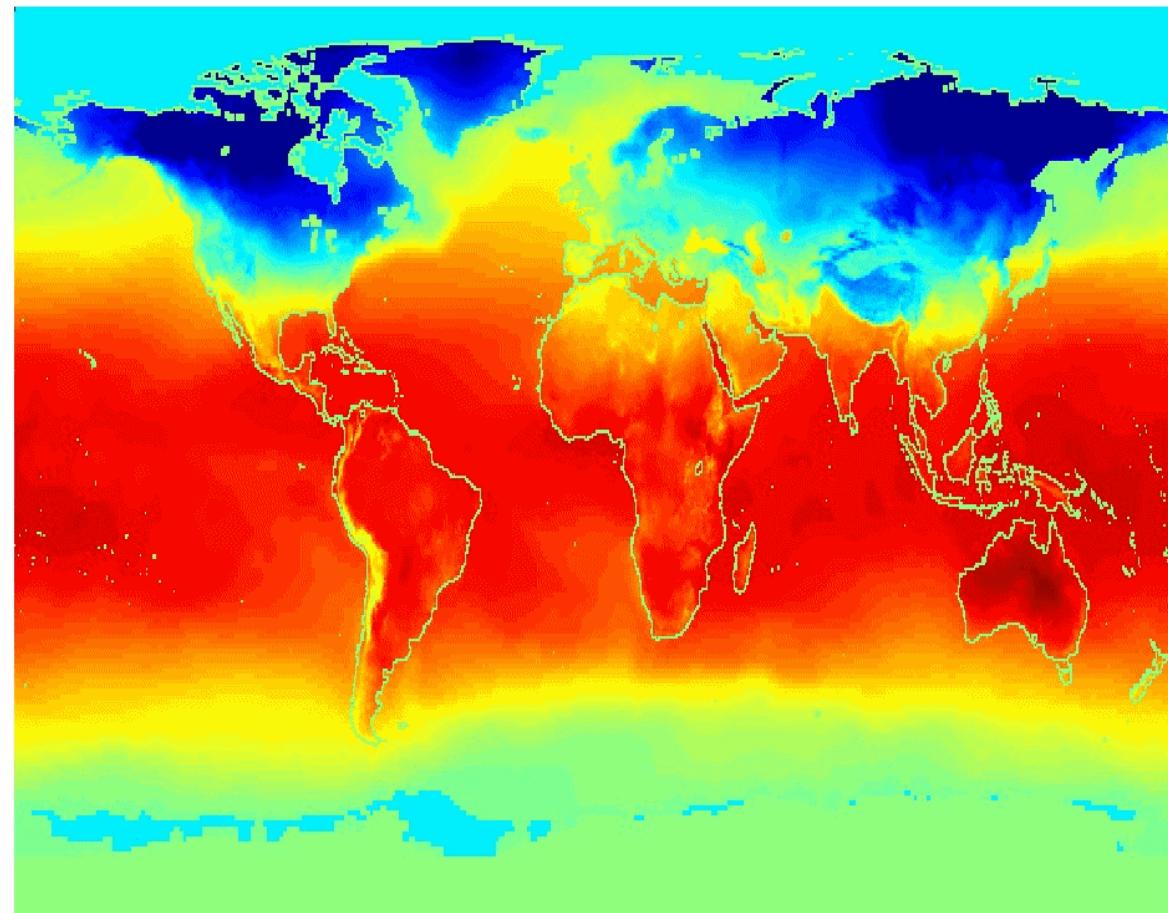
GGTTCCGCCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
**GAGAAGGGCCC**GCCTGGCGGGCG  
GGGGGAGGCAGGGCCGCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCAGCGGACAG  
GCCAAGTAGAACACCGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG

# Διατεταγμένα Δεδομένα

## ■ Χωροχρονικά δεδομένα

Μέση μηνιαία  
θερμοκρασία  
στεριάς και  
ωκεανών

Jan



# Ποιότητα Δεδομένων

- Τι είδους **προβλήματα ποιότητας δεδομένων**;
- Πώς μπορούμε να εντοπίσουμε προβλήματα στα δεδομένα;
- Τι μπορούμε να κάνουμε για αυτά τα προβλήματα?

“The most important point is that poor data quality is an unfolding disaster.

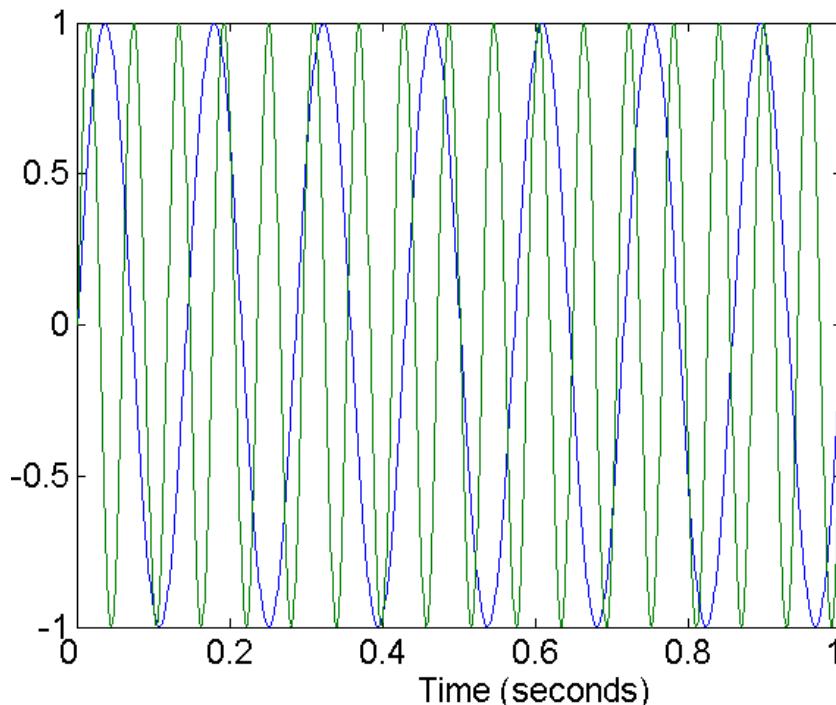
- Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate.”

Thomas C. Redman, DM Review, August 2004

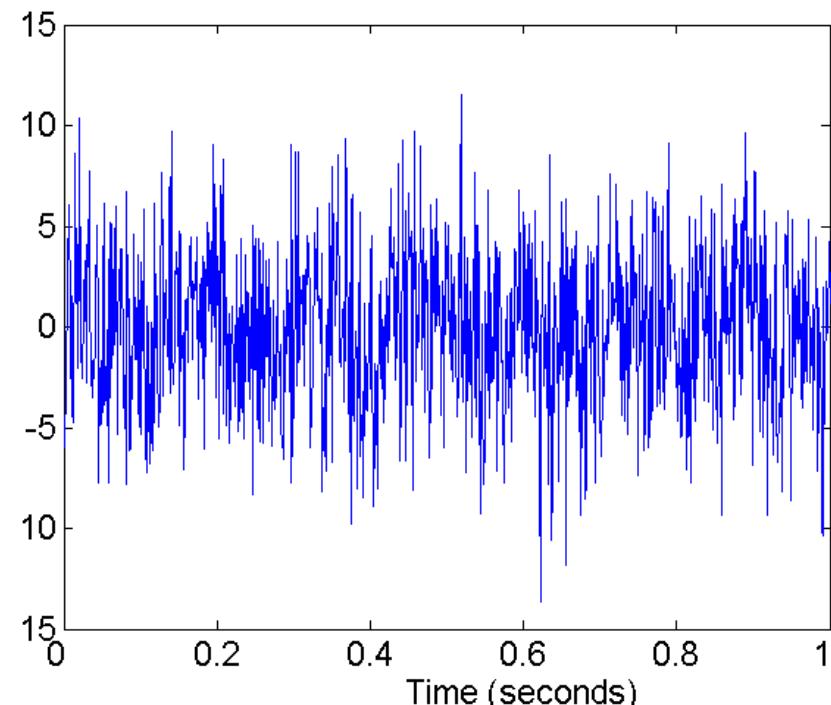
- Παραδείγματα προβλημάτων ποιότητας δεδομένων:
  - Θόρυβος και ακραίες τιμές
  - Ελλιπείς τιμές
  - Διπλότυπα

# Θόρυβος

- Ο **θόρυβος** είναι μια **τυχαία συνιστώσα** ενός **σφάλματος μέτρησης**
  - Μπορεί να αφορά τη διαστρέβλωση μιας τιμής ή την **προσθήκη** μη αληθινών αντικειμένων



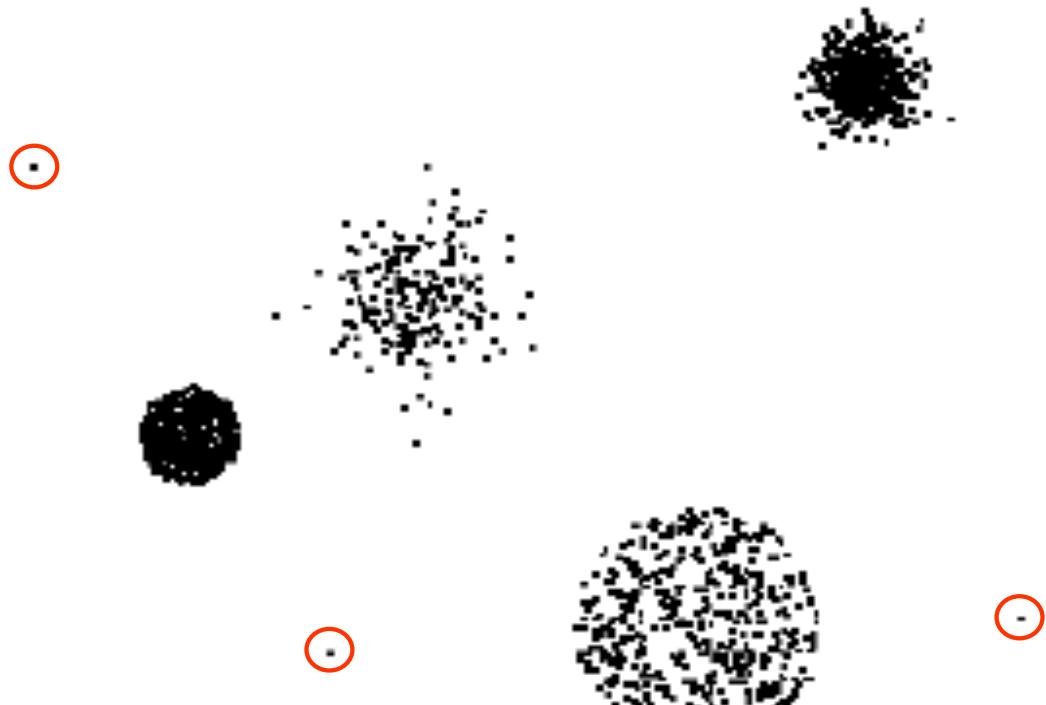
Δύο ημιτονοειδείς συναρτήσεις



Δύο ημιτονοειδείς συναρτήσεις  
με προσθήκη θορύβου

# Ακραίες (Περιθωριακές) Τιμές

- Είναι δεδομένα με τιμές χαρακτηριστικών που είναι **πολύ διαφορετικές** από τα περισσότερα σημεία στο σύνολο δεδομένων
- Σε αντίθεση με το θόρυβο, οι ακραίες τιμές (outliers) αντιστοιχούν σε **υπαρκτά δεδομένα** και η ανάλυσή τους έχει ενδιαφέρον



# Παράδειγμα Αιχαίων Τιμών

Όνομα	Ύψος	Βάρος	...	...
A	1,75	85	...	
B	1,82	94	...	...
Γ	1,59	61	...	...
Δ	2,13	122	...	...
Ε	1,77	85	...	...
Ζ	1,68	73	...	...
Η	1,72	132	...	...

# Ελλιπείς (ή Αγνοούμενες) Τιμές

- Λόγοι ύπαρξης ελλιπών τιμών (missing values)
  - Η πληροφορία δε συλλέγεται
    - **Παράδειγμα:** άνθρωποι αρνούνται να δηλώσουν το βάρος ή την ηλικία τους
  - Κάποια χαρακτηριστικά δεν είναι εφαρμόσιμα σε ορισμένες περιπτώσεις
    - **Παράδειγμα:** το ετήσιο εισόδημα δεν έχει νόημα για παιδιά
- Χειρισμός ελλιπών τιμών
  - **Εξάλειψη** των αντικειμένων
  - **Παράβλεψη** της ελλιπούς τιμής κατά την ανάλυση
  - **Εκτίμηση** των ελλιπών τιμών
  - **Αντικατάσταση** με όλες τις πιθανές τιμές (με κάποιο βάρος που προκύπτει από την πιθανότητά τους)

# Παράδειγμα Χειρισμού Ελλιπών Τιμών

Όνομα	Ύψος	Βάρος	...	...
A	1,75	-	...	
B	1,82	94	...	...
Γ	1,59	61	...	...
Δ	2,13	122	...	...
Ε	-	85	...	...
Ζ	1,68	73	...	...
Η	1,72	132	...	...

# Παράδειγμα Χειρισμού Ελλιπών Τιμών

## Εξάλειψη των αντικειμένων

Όνομα	Ύψος	Βάρος	...	...
B	1,82	94	...	...
Γ	1,59	61	...	...
Δ	2,13	122	...	...
Z	1,68	73	...	...
H	1,72	132	...	...

# Παράδειγμα Χειρισμού Ελλιπών Τιμών

## Παράβλεψη ελλιπούς τιμής κατά την ανάλυση δεδομένων

Όνομα	Ύψος	Βάρος	...	...
A	1,75	-	...	
B	1,82	94	...	...
Γ	1,59	61	...	...
Δ	2,13	122	...	...
Ε	-	85	...	...
Ζ	1,68	73	...	...
Η	1,72	132	...	...

$$\text{Μέσος όρος ύψους} = (1,75+1,82+1,59+2,13+1,68+1,72) / 6 = 1,78$$

# Παράδειγμα Χειρισμού Ελλιπών Τιμών

## Εκτίμηση ελλιπών τιμών

Όνομα	Ύψος	Βάρος	...	...
A	1,75	-	...	
B	1,82	94	...	...
Γ	1,59	61	...	...
Δ	2,13	122	...	...
Ε	1,78	85	...	...
Ζ	1,68	73	...	...
Η	1,72	132	...	...

Ύψος Ε = Μέσος όρος ύψους = 1,78

# Διπλότυπα Δεδομένα

- Το σύνολο δεδομένων μπορεί να περιέχει αντικείμενα που είναι **διπλότυπα** ή **σχεδόν διπλότυπα**
  - Σύνηθες πρόβλημα όταν συγχωνεύονται **δεδομένα από ετερογενείς πηγές**
- Παραδείγματα:
  - Το ίδιο πρόσωπο με πολλαπλούς λογαριασμούς e-mail
- **Καθαρισμός δεδομένων (data cleaning)**
  - Εντοπισμός και διόρθωση σφαλμάτων στα δεδομένα (**error detection**)

# Παράδειγμα Διπλότυπων Δεδομένων

Όνοματεπώνυμο	E-mail	...
Βασίλης σπανούλης	vs@gsp.gr	...
Βασίλης Σπανούλης	vs@gmail.com	...
Νίκος Παππάς	nikp@gmail.com	...
Δημήτρης Διαμαντίδης	ddiam@gmail.com	...
Νικόλαος Παππάς	npappas@gmail.com	...
Δημήτρης Διαμαντίδης	ddiam@gmail.com	...
Βασλης Σπανούλης	vs@osfp.gr	...

**Ακριβώς η ίδια εγγραφή πάνω από μία φορές (εύκολο)**

# Παράδειγμα Διπλότυπων Δεδομένων

Όνοματεπώνυμο	E-mail	...
Βασίλης σπανούλης	vs@gsp.gr	...
Βασίλης Σπανούλης	vs@gmail.com	...
Νίκος Παππάς	nikp@gmail.com	...
Δημήτρης Διαμαντίδης	ddiam@gmail.com	...
Νικόλαος Παππάς	npappas@gmail.com	...
Δημήτρης Διαμαντίδης	ddiam@gmail.com	...
Βασλης Σπανούλης	vs@osfp.gr	...

**Ίδια εγγραφή, αλλά χωρίς κοινές τιμές**

# Παράδειγμα Διπλότυπων Δεδομένων

Όνοματεπώνυμο	E-mail	...
Βασίλης σπανούλης	vs@gsp.gr	...
Βασίλης Σπανούλης	vs@gmail.com	...
Νίκος Παππάς	nikp@gmail.com	...
Δημήτρης Διαμαντίδης	ddiam@gmail.com	...
Νικόλαος Παππάς	npappas@gmail.com	...
Δημήτρης Διαμαντίδης	ddiam@gmail.com	...
Βασλης Σπανούλης	vs@osfp.gr	...

**Ίδια εγγραφή, αλλά χωρίς κοινές τιμές**

# Περιεχόμενο Σημερινής Διάλεξης

- Εισαγωγικά στοιχεία μαθήματος
- **Δεδομένα**
  - Τύποι δεδομένων
  - Ποιότητα δεδομένων
  - **Προεπεξεργασία δεδομένων**
- Μέτρα απόστασης και ομοιότητας
  - Ομοιότητα πολυδιάστατων αριθμητικών δεδομένων
  - Ομοιότητα συνόλων
  - Ομοιότητα αλφαριθμητικών

# Προεπεξεργασία Δεδομένων

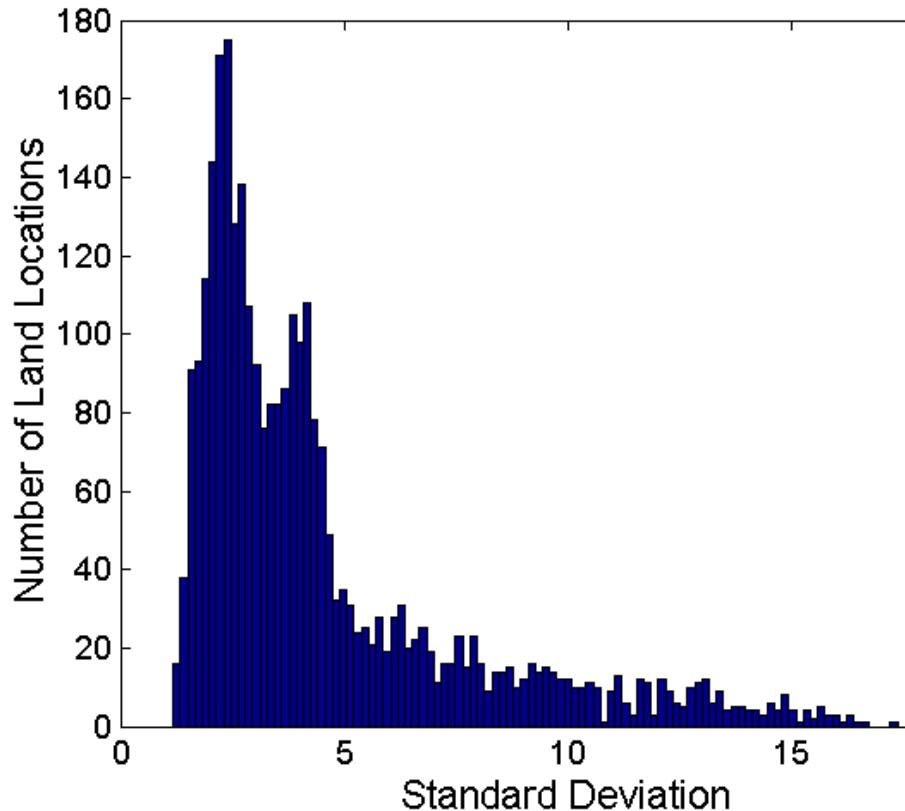
- Συνάθροιση (Aggregation)
- Δειγματοληψία (Sampling)
- Μείωση διαστάσεων (Dimensionality Reduction)
- Επιλογή υποσυνόλου γνωρισμάτων (Feature subset selection)
- Δημιουργία γνωρισμάτων (Feature creation)
- Διακριτοποίηση και δυαδικοποίηση (Discretization and Binarization)
- Μετασχηματισμοί μεταβλητών (Attribute Transformation)

# Συνάθροιση

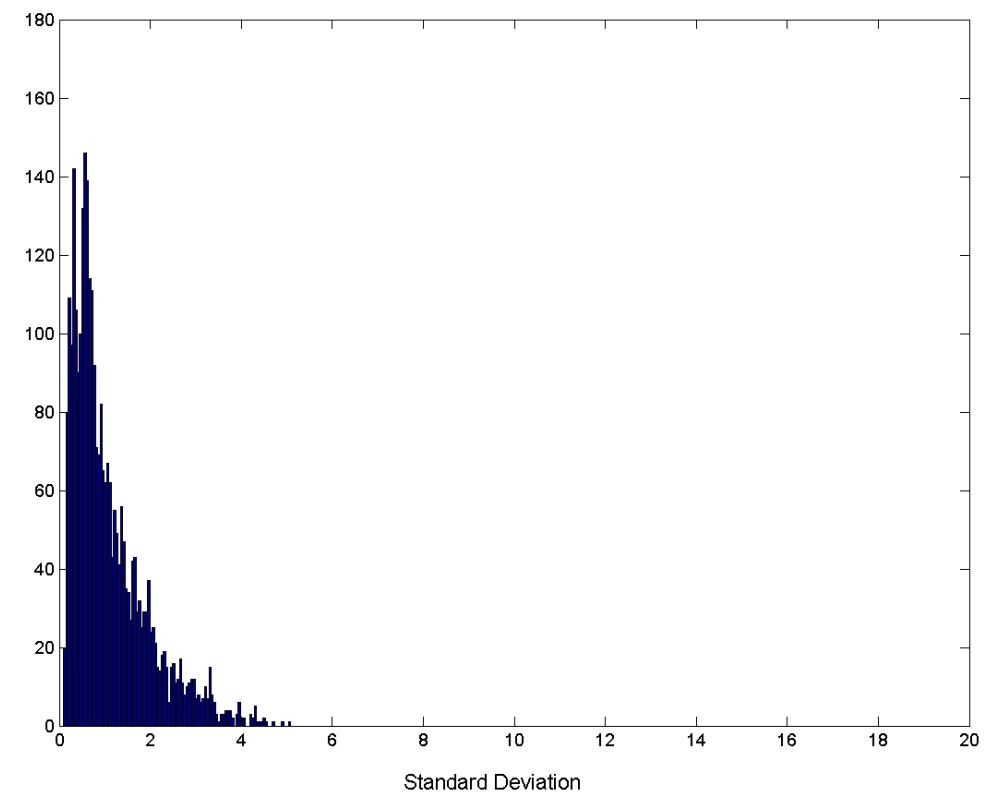
- Συνδυασμός δύο ή περισσότερων γνωρισμάτων (ή αντικειμένων) σε ένα γνώρισμα (ή αντικείμενο)
- Σκοπός
  - Μείωση δεδομένων
    - Μείωση του πλήθους γνωρισμάτων ή αντικειμένων
  - Αλλαγή κλίμακας
    - Πόλεις συναθροίζονται σε περιοχές, πολιτείες, χώρες, κτλ.
  - Πιο σταθερή συμπεριφορά δεδομένων
    - Τα συναθροισμένα δεδομένα τείνουν να έχουν μικρότερη μεταβλητότητα

# Συνάθροιση (Αντικειμένων)

- Η βροχόπτωση στην Αυστραλία σε μηνιαία και ετήσια βάση
- Τα συναθροισμένα δεδομένα τείνουν να έχουν μικρότερη μεταβλητότητα



Τυπική απόκλιση μέσης **μηνιαίας**  
βροχόπτωσης



Τυπική απόκλιση μέσης **ετήσιας**  
βροχόπτωσης

# Δειγματοληψία

- Είναι η κύρια τεχνική για την **επιλογή υποσυνόλου** των αντικειμένων δεδομένων
  - Χρησιμοποιείται εκτεταμένα τόσο για την **προκαταρκτική** έρευνα των δεδομένων όσο και για την **τελική** τους ανάλυση
- Οι στατιστικολόγοι χρησιμοποιούν τη δειγματοληψία επειδή η λήψη **ολόκληρου** του συνόλου των δεδομένων που ενδιαφέρει είναι **πολύ ακριβή** ή **πολύ χρονοβόρα**
- Στην ανάλυση δεδομένων χρησιμοποιείται επειδή η **επεξεργασία** όλων των δεδομένων είναι **πολύ ακριβή** ή **χρονοβόρα**

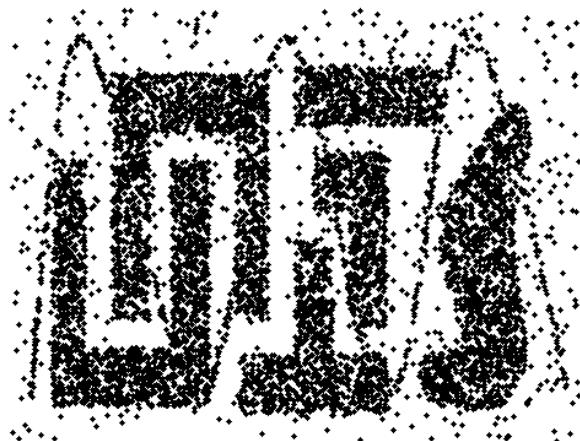
# Δειγματοληψία

- Η βασική αρχή της αποτελεσματικής δειγματοληψίας είναι η ακόλουθη:
  - Η χρήση ενός δείγματος θα λειτουργήσει σχεδόν όσο καλά όσο και η χρήση ολόκληρου του συνόλου δεδομένων, αν το δείγμα είναι **αντιπροσωπευτικό**
  - Ένα δείγμα είναι **αντιπροσωπευτικό** αν έχει κατά προσέγγιση τις **ίδιες ιδιότητες** (**που ενδιαφέρουν**) όπως το αρχικό σύνολο δεδομένων

# Τύποι Δειγματοληψίας

- Απλή **τυχαία δειγματοληψία**
  - Ίση πιθανότητα επιλογής ενός στοιχείου
- Δειγματοληψία **χωρίς αντικατάσταση**
  - Το κάθε επιλεγμένο στοιχείο αφαιρείται από το σύνολο όλων των αντικειμένων που συνολικά απαρτίζουν τον πληθυσμό
- Δειγματοληψία **με αντικατάσταση**
  - Τα αντικείμενα δεν αφαιρούνται από τον πληθυσμό όταν επιλέγονται στο δείγμα
    - Άρα το ίδιο αντικείμενο μπορεί να επιλεχθεί πάνω από μια φορά
- **Στρωματοποιημένη δειγματοληψία**
  - Χωρίζονται τα δεδομένα σε προκαθορισμένες ομάδες αντικειμένων, και μετά επιλέγονται τυχαία δείγματα από κάθε ομάδα

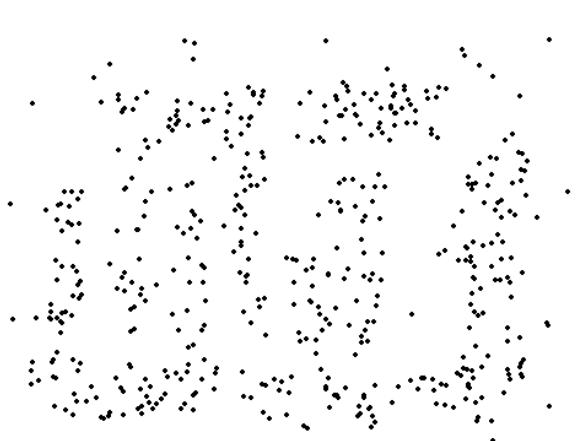
# Μέγεθος Δείγματος



8000 σημεία



2000 σημεία

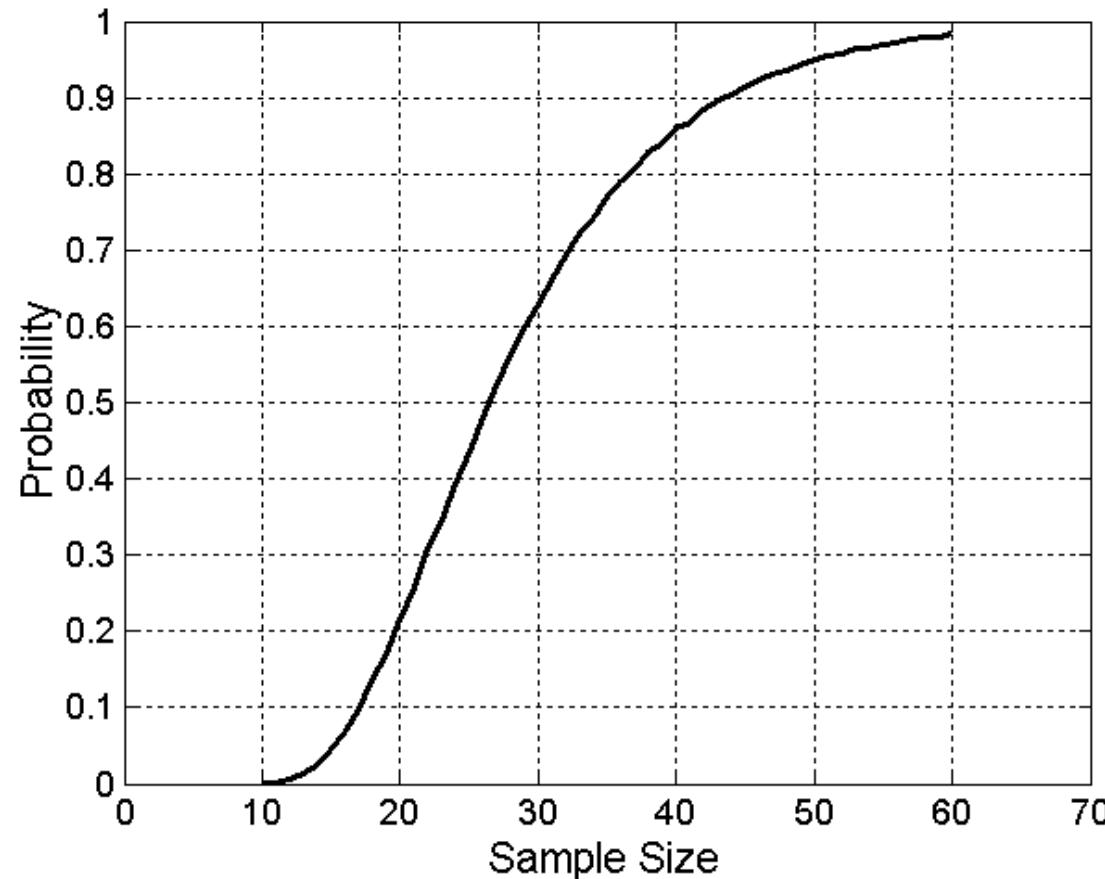
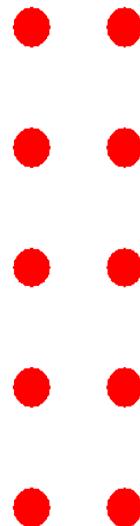


500 σημεία

- Δύο αντικρουόμενοι στόχοι:
  - **Μικρό** μέγεθος δείγματος: ώστε να μπορούμε να επεξεργαστούμε τα δεδομένα
  - Όμως, **αρκετά μεγάλο** ώστε να διατηρούνται οι ιδιότητες των δεδομένων

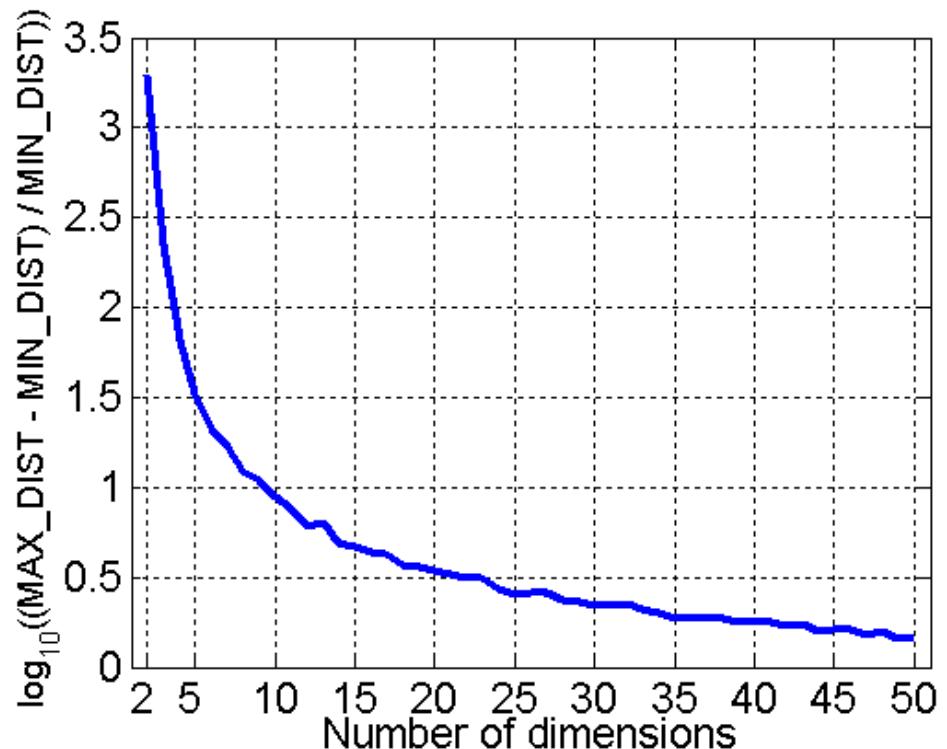
# Μέγεθος Δείγματος

- Εύρεση αντιπροσωπευτικών σημείων από 10 ομάδες
- Ποιο μέγεθος δείγματος χρειάζεται για να λάβουμε τουλάχιστον ένα αντικείμενο από κάθε μία από τις 10 ομάδες?



# Η Κατάρα των Πολλών Διαστάσεων

- Όταν αυξάνεται η διάσταση, αυξάνεται η **σποραδικότητα** των δεδομένων
- Οι ορισμοί της **πυκνότητας** και της **απόστασης** μεταξύ σημείων, που είναι κρίσιμοι για συσταδοποίηση και εντοπισμό ακραίων τιμών, δεν έχουν τόσο νόημα



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

# Μείωση Διάστασης

## ■ Σκοπός:

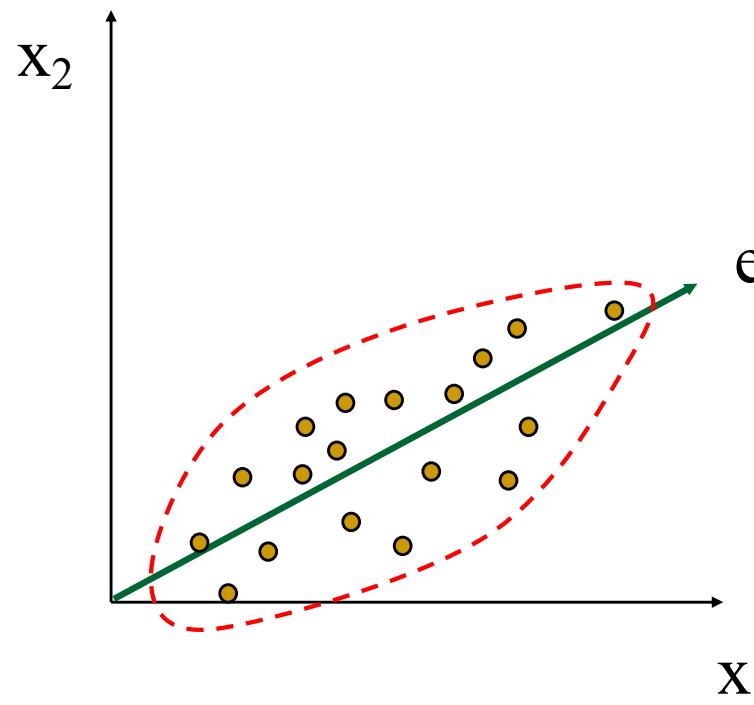
- Αποφυγή της κατάρας των πολλών διαστάσεων
- Μείωση του χρόνου και μνήμης που απαιτούν αλγόριθμοι ανάλυσης δεδομένων
- Ευκολία στην οπτικοποίηση
- Μπορεί να βοηθήσει στο να απορριφθούν άσχετα γνωρίσματα ή να μειωθεί ο θόρυβος

## ■ Τεχνικές

- Principle Component Analysis (Ανάλυση Κύριων Συνιστωσών)
- Singular Value Decomposition (Διάσπαση Μοναδιαίων Τιμών)
- Πολλές άλλες τεχνικές: εποπτευόμενες και μη-γραμμικές τεχνικές

# Μείωση Διάστασης: PCA

- Ο στόχος είναι η εύρεση μιας προβολής που πιάνει το μεγαλύτερο ποσό μεταβολής στα δεδομένα
- Εύρεση ιδιοδιανυσμάτων του πίνακα συνδιακύμανσης
- Τα ιδιοδιανύσματα ορίζουν το νέο χώρο



# Επιλογή Υποσυνόλου Γνωρισμάτων (Feature Subset Selection)

- Ένας άλλος τρόπος μείωσης διάστασης
- **Περιττά** γνωρίσματα
  - Διπλασιάζουν ένα μέρος ή ολόκληρη την πληροφορία που εμπεριέχεται σε ένα ή περισσότερα άλλα γνωρίσματα
  - **Παράδειγμα:** τιμή αγοράς προϊόντος και ποσό φόρου πωλήσεως που πληρώνεται
- **Άσχετα** γνωρίσματα
  - Δεν περιέχουν σχεδόν καθόλου χρήσιμες πληροφορίες για την εργασία ανάλυσης δεδομένων
  - **Παράδειγμα:** τα αναγνωριστικά φοιτητών είναι άσχετα με την εργασία πρόβλεψης της μέσης βαθμολογίας τους

# Επιλογή Υποσυνόλου Γνωρισμάτων (Feature Subset Selection)

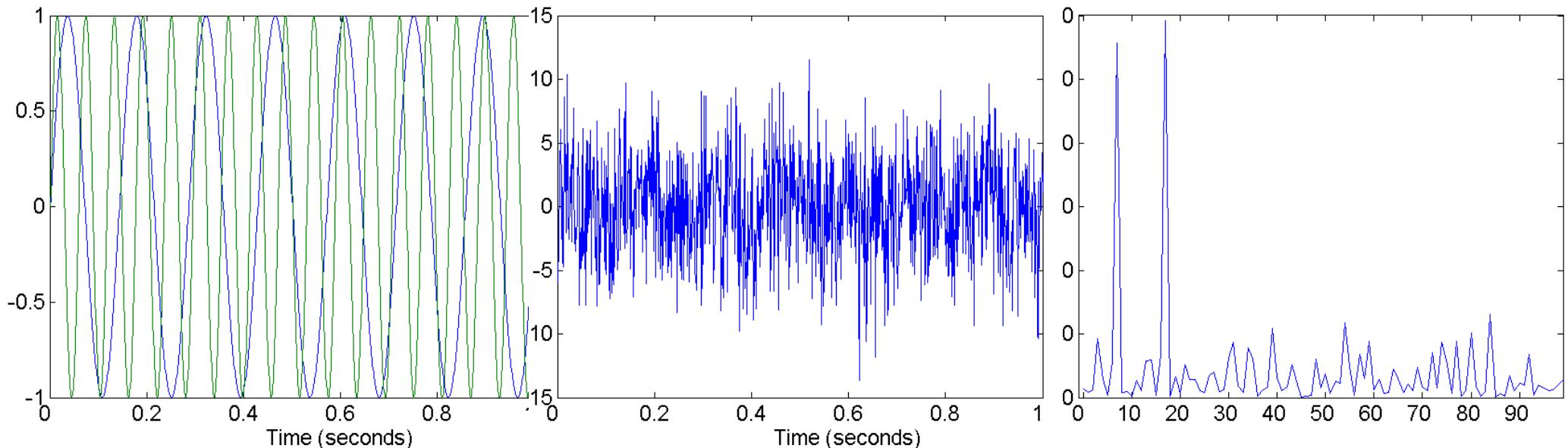
- Τεχνικές επιλογής υποσυνόλου γνωρισμάτων
  - Εξαντλητική προσέγγιση (brute-force approach)
    - Δοκιμή όλων των δυνατών υποσυνόλων γνωρισμάτων
  - Ενσωματωμένη (embedded approaches)
    - Η επιλογή των γνωρισμάτων λαμβάνει χώρα φυσιολογικά ως τμήμα του αλγόριθμου ανάλυσης δεδομένων
  - Φίλτρου (filter approaches)
    - Τα γνωρίσματα επιλέγονται πριν εκτελεστεί ο αλγόριθμος ανάλυσης δεδομένων
  - Περιτυλίγματος (wrapper approaches)
    - Χρησιμοποιούν τον αλγόριθμο ανάλυσης δεδομένων ως μαύρο κουτί για την εύρεση του καλύτερου υποσυνόλου γνωρισμάτων

# Δημιουργία Γνωρισμάτων

- Κατασκευάζονται **νέα γνωρίσματα** (από τα αρχικά) που διατηρούν τη σημαντική πληροφορία με καλύτερο τρόπο
- Τρεις μεθοδολογίες
  - Εξαγωγή γνωρισμάτων (**feature extraction**)
    - Εξαρτημένη από το πεδίο (domain-specific)
    - **Παράδειγμα:** εξαγωγή ακμών από εικόνες
  - Απεικόνιση σε **νέο χώρο**
    - **Παράδειγμα:** ανάλυση Fourier ή Wavelet
  - **Κατασκευή γνωρισμάτων (feature construction)**
    - Συνδυασμός γνωρισμάτων
    - **Παράδειγμα:** διαίρεση μάζας διά όγκου για υπολογισμό πυκνότητας

# Απεικόνιση Δεδομένων σε νέο Χώρο

- Μετασχηματισμός Fourier (Fourier transform)
- Κυματοειδής μετασχηματισμός (Wavelet transform)



Δύο ημιτονοειδείς συναρτήσεις

Δύο ημιτονοειδείς συναρτήσεις με προσθήκη θορύβου

Στο πεδίο της συχνότητας

# Διακριτοποίηση (Discretization)

- Διακριτοποίηση είναι η διαδικασία μετασχηματισμού ενός συνεχούς γνωρίσματος σε διακριτό
  - Ένα (δυνητικά) άπειρο πλήθος τιμών απεικονίζεται σε ένα περιορισμένο σύνολο κατηγοριών
  - Η διακριτοποίηση χρησιμοποιείται συχνά στην κατηγοριοποίηση
  - Πολλοί αλγόριθμοι κατηγοριοποίησης δουλεύουν καλά όταν τόσο οι ανεξάρτητες όσο και η εξαρτημένη μεταβλητή έχουν λίγες διακριτές τιμές
- Ο μετασχηματισμός συντελείται σε **δύο βήματα**:
  1. Απόφαση σχετικά με το πλήθος κατηγοριών
  2. Καθορισμός τρόπου απεικόνισης των τιμών του συνεχούς γνωρίσματος σε αυτές τις κατηγορίες

# Παράδειγμα: Το Σύνολο Δεδομένων Iris Plant

## ■ Iris Plant data set

- Can be obtained from the UCI Machine Learning Repository  
<http://www.ics.uci.edu/~mlearn/MLRepository.html>
- From the statistician Douglas Fisher
- Three flower types (classes):
  - Setosa
  - Versicolour
  - Virginica
- Four (non-class) attributes
  - Sepal width and length
  - Petal width and length

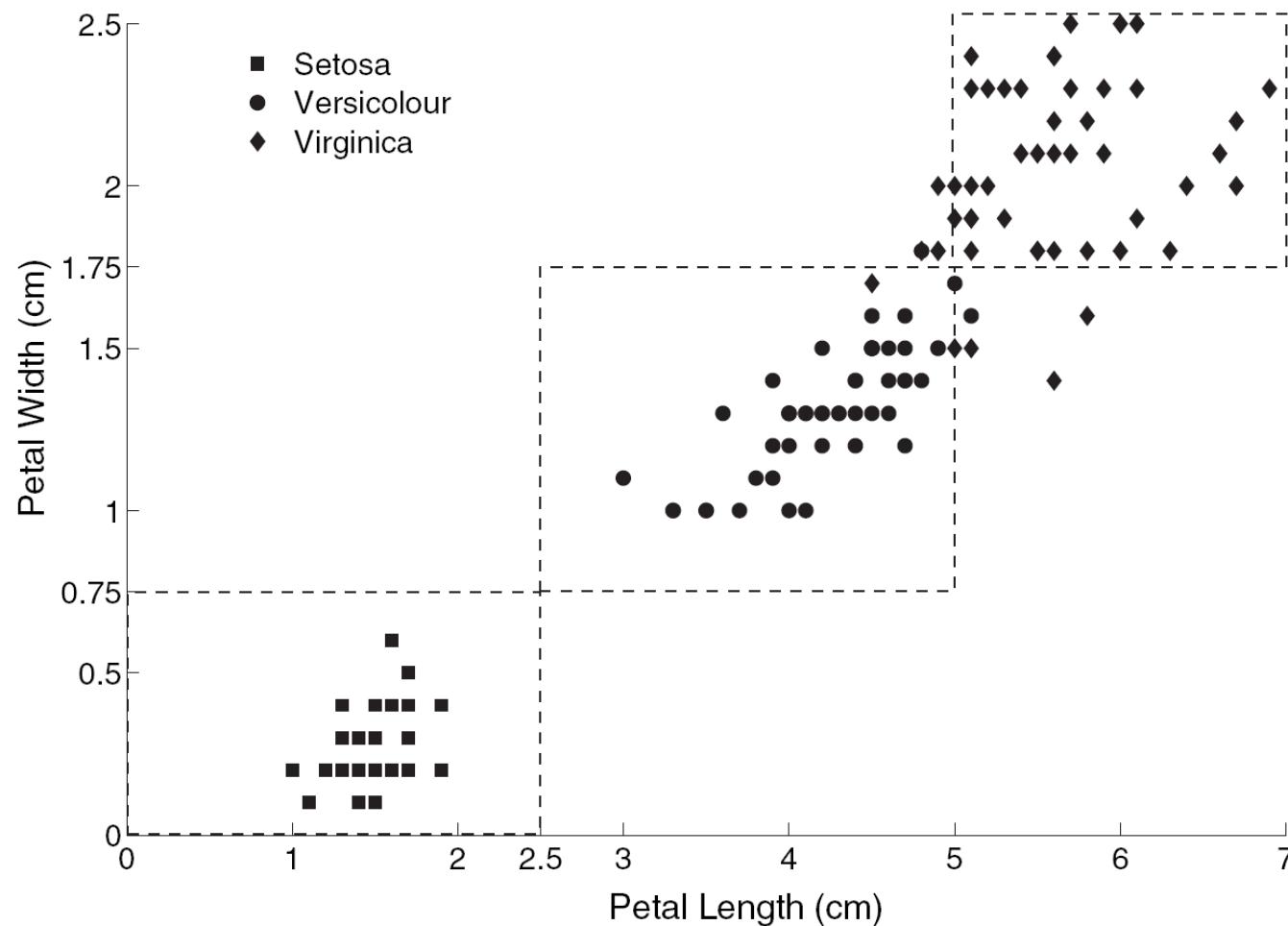


Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

# Παράδειγμα: Το Σύνολο Δεδομένων Iris Plant

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...	...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

# Παράδειγμα Διακριτοποίησης στο Iris Plant



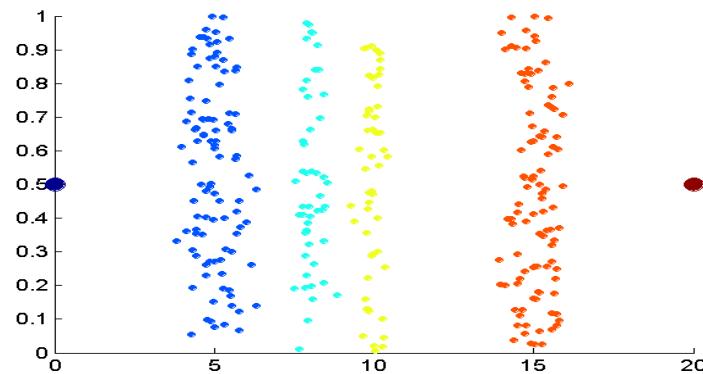
Petal width low or petal length low implies **Setosa**.

Petal width medium or petal length medium implies **Versicolour**.

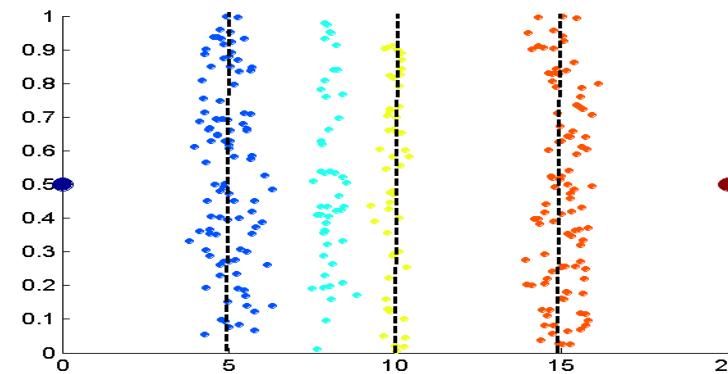
- Petal width high or petal length high implies **Virginica**.

# Διακριτοποίηση δίχως χρήση Επικετών Κλάσεων

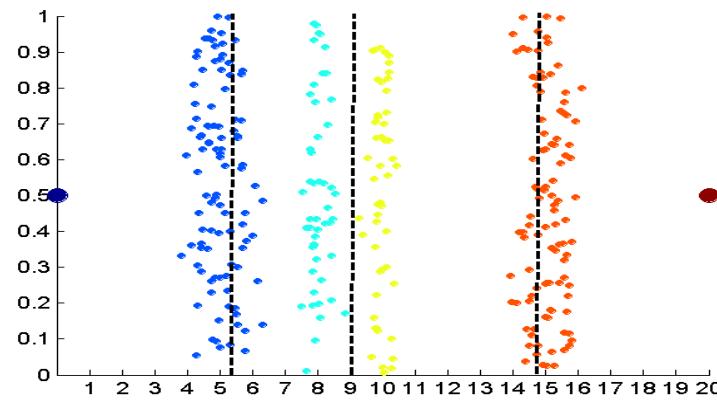
## ■ Διακριτοποίηση χωρίς επίβλεψη



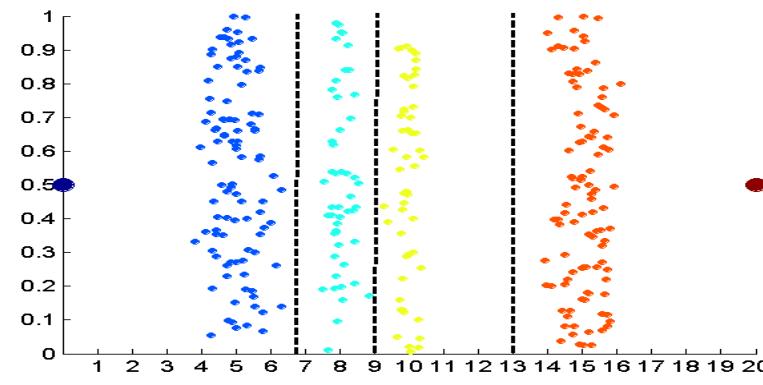
Data



Equal interval



Equal frequency



K-means

# Μετασχηματισμός Μεταβλητών

- Μια συνάρτηση που απεικονίζει το σύνολο τιμών μιας συγκεκριμένης μεταβλητής σε ένα σύνολο τιμών αντικατάστασης, έτσι ώστε κάθε παλιά τιμή μπορεί να αναγνωριστεί με μια από τις νέες τιμές
  - Απλές συναρτήσεις:  $x^k$ ,  $\log(x)$ ,  $e^x$ ,  $|x|$
- **Κανονικοποίηση (Normalization)** ή **Τυποποίηση (Standardization)**
  - **Παράδειγμα:** τυποποίηση μιας μεταβλητής (στη Στατιστική) → δημιουργία νέας μεταβλητής με μέσο 0 και τυπική απόκλιση 1

$$z = \frac{x - \mu}{\sigma}$$

$\mu$  = Mean

$\sigma$  = Standard Deviation

- **Παράδειγμα:** σύγκριση ανθρώπων βάσει ηλικίας και εισοδήματος → απαιτεί κανονικοποίηση των 2 μεταβλητών στο ίδιο διάστημα τιμών, ώστε οι διαφορές στο εύρος τιμών να είναι συγκρίσιμες

# Κανονικοποίηση (Normalization)

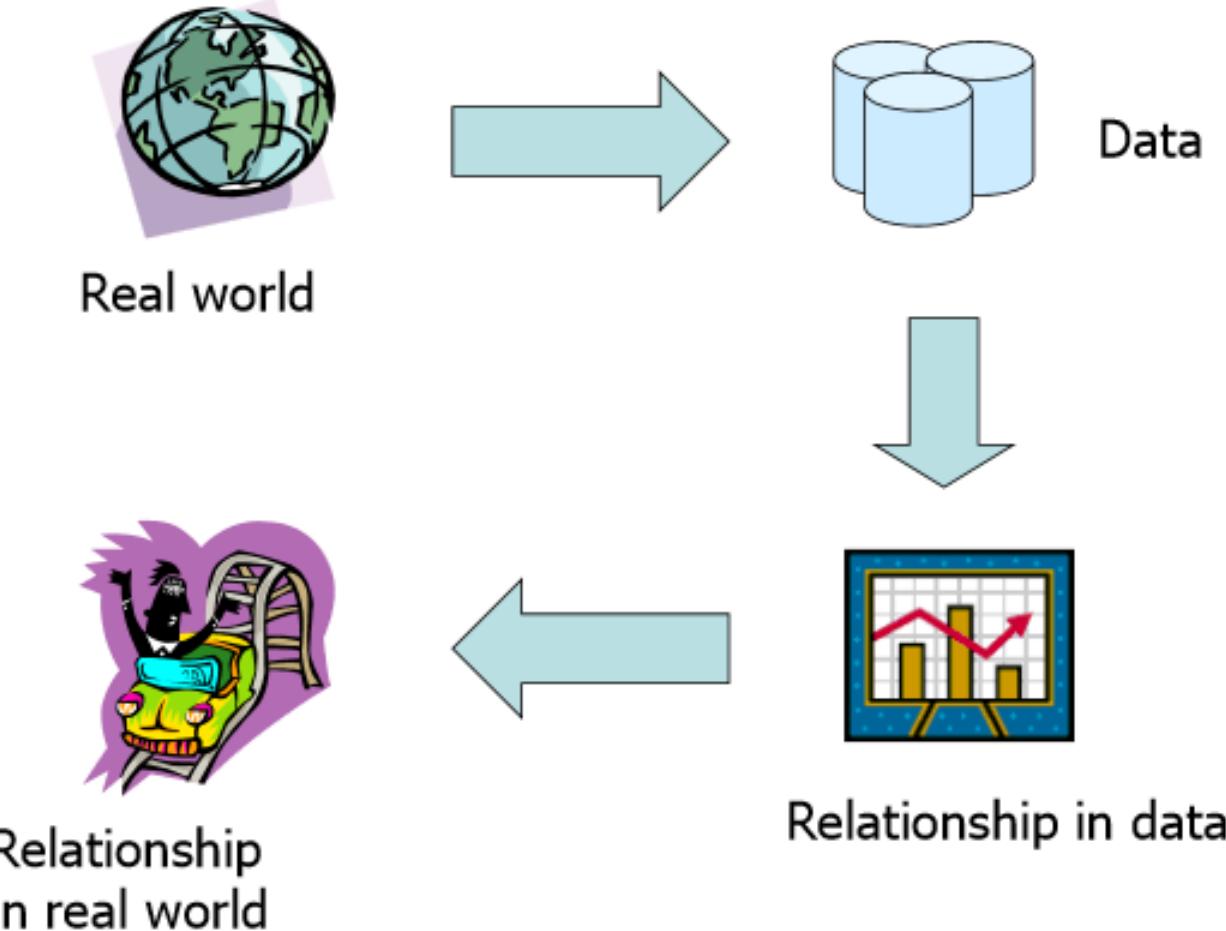
- Κανονικοποίηση μπορεί να γίνει με διάφορους τρόπους
- Άλλος βασικός τρόπος (υπάρχουν και άλλοι, και ο κατάλληλος τρόπος εξαρτάται από την εφαρμογή):
  - Feature scaling: απεικονίζει όλες τις τιμές στο διάστημα  $[0,1]$ 
$$(x - \text{min}) / (\text{max} - \text{min})$$
  - Μπορεί να προσαρμοστεί για απεικόνιση σε αυθαίρετο διάστημα  $[\alpha, \beta]$  (πώς;)

# Περιεχόμενο Σημερινής Διάλεξης

- Εισαγωγικά στοιχεία μαθήματος
- Δεδομένα
  - Τύποι δεδομένων
  - Ποιότητα δεδομένων
  - Προεπεξεργασία δεδομένων
- **Μέτρα απόστασης και ομοιότητας**
  - **Ομοιότητα πολυδιάστατων αριθμητικών δεδομένων**
  - Ομοιότητα συνόλων
  - Ομοιότητα αλφαριθμητικών

# Μετρήσεις

- Αντιστοίχιση οντοτήτων του πραγματικού κόσμου σε συμβολικές αναπαραστάσεις



# Μέτρα Ομοιότητας και Ανομοιότητας

- **Μέτρο ομοιότητας**
  - Αριθμητική μέτρηση που υποδεικνύει πόσο **μοιάζουν** δύο αντικείμενα
  - Είναι υψηλότερη όσο τα αντικείμενα μοιάζουν περισσότερο μεταξύ τους
  - Συνήθως παίρνει τιμές στο διάστημα: [0, 1]
- **Μέτρο ανομοιότητας**
  - Αριθμητική μέτρηση που υποδεικνύει πόσο **διαφέρουν** δύο αντικείμενα
  - Χαμηλή τιμή όταν τα αντικείμενα μοιάζουν περισσότερο μεταξύ τους
  - Η ελάχιστη ανομοιότητα είναι συνήθως 0
  - Η μέγιστη τιμή ποικίλει
- **Η εγγύτητα (proximity)** αναφέρεται είτε σε ομοιότητα ή ανομοιότητα

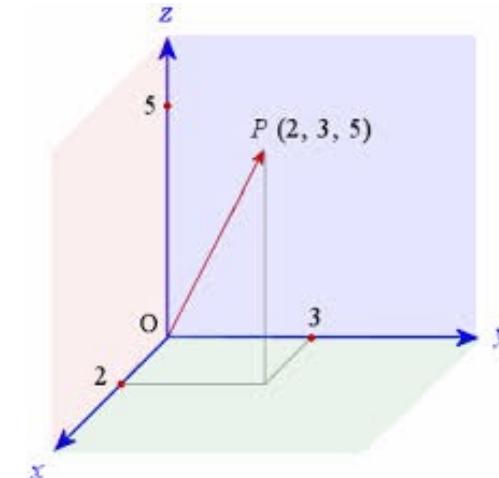
# Ομοιότητα/Ανομοιότητα για Απλά Γνωρίσματα

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d =  x - y /(n - 1)$ (values mapped to integers 0 to $n-1$ , where $n$ is the number of values)	$s = 1 - d$
Interval or Ratio	$d =  x - y $	$s = -d, s = \frac{1}{1+d}, s = e^{-d},$ $s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

# Αναπαράσταση: Διανύσματα και Πίνακες Αποστάσεων

- Τα δεδομένα δίνονται με τη μορφή **η διανυσμάτων** που αναπαρίστανται σε χώρο **d διαστάσεων**

$$\mathbf{V} = [2, 3, 5]$$



- Εναλλακτικά, τα δεδομένα μπορεί να δίνονται με τη μορφή **ηxη πίνακα ομοιοτήτων ή αποστάσεων**

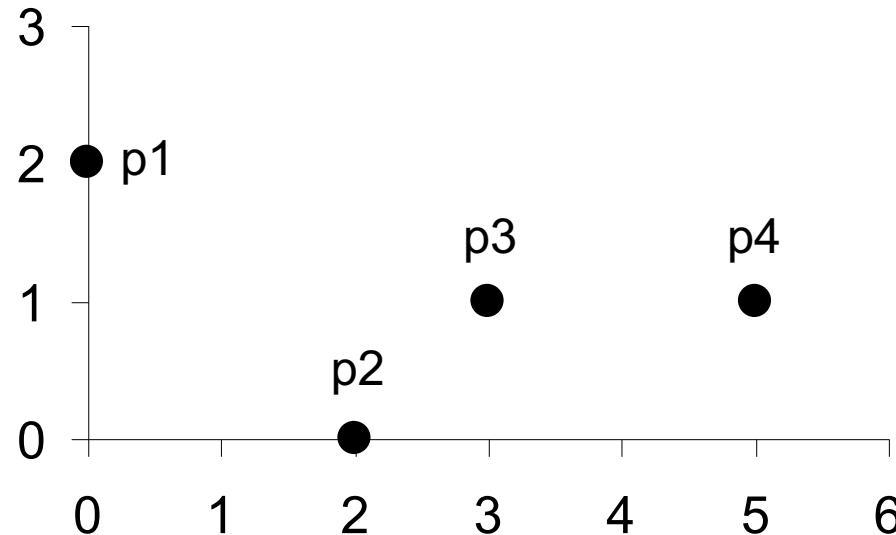
Vi	Vj
	0,43

# Ευκλείδεια Απόσταση

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

- Όπου  $n$  το πλήθος των διαστάσεων, και  $x_k$  ( $y_k$ ) η τιμή του  $k$ -οστού γνωρίσματος του αντικειμένου  $\mathbf{x}$  ( $\mathbf{y}$ )
- Έχει νόημα όταν οι μετρήσεις (γνωρίσματα) μετρούνται στις ίδιες μονάδες μέτρησης
- Εάν είναι διαφορετικές, π.χ. μήκος σε διαφορετικές μονάδες μέτρησης, η ευκλείδεια απόσταση δεν έχει νόημα

# Ευκλείδεια Απόσταση



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

## Πίνακας Αποστάσεων

# Απόσταση Minkowski

$$d(\mathbf{x}, \mathbf{y}) = \left( \sum_{k=1}^n |x_k - y_k|^r \right)^{1/r}$$

- Αποτελεί γενίκευση της Ευκλείδειας απόστασης
- Όπου  $r$  μια παράμετρος
  - $r=1$ : απόσταση Manhattan ή  $L_1$
  - $r=2$ : ευκλείδεια απόσταση ή  $L_2$
  - $r \rightarrow \infty$  : “supremum” ( $L_{\max}$  norm,  $L_\infty$  norm) απόσταση
- Προσοχή: να μην συγχέεται το  $r$  με τη διάσταση  $n$ 
  - Όλες αυτές οι αποστάσεις ορίζονται για οποιοδήποτε πλήθος διαστάσεων  $n$

# Απόσταση Minkowski

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

$L_{\infty}$	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

## Πίνακας Αποστάσεων

# Ιδιότητες Συναρτήσεων Απόστασης

- Ορισμένες συναρτήσεις απόστασης, όπως η Ευκλείδεια, έχουν γνωστές ιδιότητες:
  1.  $d(x, y) \geq 0$  για κάθε  $x$  και  $y$  και  $d(x, y) = 0$  μόνο εάν  $x = y$ . (**Positive definiteness**)
  2.  $d(x, y) = d(y, x)$  για κάθε  $x$  και  $y$ . (**Symmetry**)
  3.  $d(x, z) \leq d(x, y) + d(y, z)$  για οποιαδήποτε  $x$ ,  $y$ , και  $z$ . (**Triangle Inequality**)όπου  $d(x, y)$  είναι η απόσταση (ανομοιότητα) μεταξύ σημείων (αντικειμένων),  $x$  και  $y$ .
- Μια συνάρτηση απόστασης που ικανοποιεί αυτές τις ιδιότητες λέγεται **μετρική** (metric)

# Ιδιότητες Συναρτήσεων Ομοιότητας

- Οι συναρτήσεις ομοιότητας έχουν κι αυτές ορισμένες γνωστές ιδιότητες:
  1.  $s(x, y) = 1$  (ή μέγιστη ομοιότητα) μόνο εάν  $x = y$ .
  2.  $s(x, y) = s(y, x)$  για όλα τα  $x$  και  $y$ . (Συμμετρία)

όπου  $s(x, y)$  η ομοιότητα μεταξύ σημείων (αντικειμένων),  $x$  και  $y$ .

# Περιεχόμενο Σημερινής Διάλεξης

- Εισαγωγικά στοιχεία μαθήματος
- Δεδομένα
  - Τύποι δεδομένων
  - Ποιότητα δεδομένων
  - Προεπεξεργασία δεδομένων
- **Μέτρα απόστασης και ομοιότητας**
  - Ομοιότητα πολυδιάστατων αριθμητικών δεδομένων
  - **Ομοιότητα συνόλων**
  - Ομοιότητα αλφαριθμητικών

# Τομή Συνόλων – Προβλήματα

- Δοθέντων δύο συνόλων  $A$  και  $B$ , η τομή τους είναι το υποσύνολο των κοινών τους στοιχείων
- Συνήθως αναφερόμαστε στο **πλήθος** των κοινών τους στοιχείων
- Παράδειγμα:
  - $A=\{\alpha, \beta, \gamma, \delta, \varepsilon\}$  και  $B=\{\alpha, \beta, \zeta\}$ , τότε  $A \cap B = \{\alpha, \beta\}$  και  $|A \cap B| = 2$
- Πρόβλημα: με βαση την τομή, τα ακόλουθα ζεύγη συνόλων είναι το ίδιο όμοια
  - $A=\{\alpha, \beta, \gamma, \delta, \varepsilon, \zeta, \eta, \theta, \iota, \kappa, \dots\}$  και  $B=\{\alpha\} \rightarrow |A \cap B| = 1$
  - $A=\{\alpha, \beta\}$  και  $B=\{\alpha\} \rightarrow |A \cap B| = 1$

# Συντελεστής Jaccard (Jaccard Coefficient)

- Ένα κοινό μέτρο που δείχνει πόσο μεγάλη είναι η τομή δύο συνόλων
- Έστω  $X$  και  $Y$  δύο σύνολα, τότε ο **συντελεστής Jaccard (Jaccard Coefficient)** είναι  $|X \cap Y| / |X \cup Y|$
- Ισούται με 1, όταν τα  $X$  και  $Y$  έχουν τα ίδια στοιχεία, και με 0 όταν έχουν εντελώς διαφορετικά
- Τα  $X$  και  $Y$  **δε χρειάζεται να έχουν το ίδιο μήκος**
- Πάντα παράγει έναν αριθμό μεταξύ 0 και 1
  - Ένα κατώφλι (threshold) καθορίζει εάν πρόκειται για ταίριασμα
  - Π.χ., εάν **J.C.** > 0.8, τότε έχουμε ταίριασμα

# Δυαδικά Διανύσματα

	j=1	j=0
i=1	$n_{11}$	$n_{10}$
i=0	$n_{01}$	$n_{00}$

Πλήθος μεταβλητών όπου  
item i=1 και item j=0

- Simple matching coefficient
- Jaccard coefficient

$$\frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}}$$

$$\frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

## Παράδειγμα: SMC vs. Jaccard

$\mathbf{x} = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$

$\mathbf{y} = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$

$n_{01} = 2$  (the number of attributes where  $p$  was 0 and  $q$  was 1)

$n_{10} = 1$  (the number of attributes where  $p$  was 1 and  $q$  was 0)

$n_{00} = 7$  (the number of attributes where  $p$  was 0 and  $q$  was 0)

$n_{11} = 0$  (the number of attributes where  $p$  was 1 and  $q$  was 1)

$$\begin{aligned}\text{SMC} &= (n_{11} + n_{00}) / (n_{01} + n_{10} + n_{11} + n_{00}) \\ &= (0+7) / (2+1+0+7) = 0.7\end{aligned}$$

$$J = (n_{11}) / (n_{01} + n_{10} + n_{11}) = 0 / (2 + 1 + 0) = 0$$

# Ομοιότητα Συνημιτόνου (Cosine Similarity)

- Αν  $\mathbf{d}_1$  και  $\mathbf{d}_2$  αναπαριστούν δύο διανύσματα (εγγράφων):

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \|\mathbf{d}_2\| ,$$

όπου  $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$  το εσωτερικό γινόμενο των διανυσμάτων  $\mathbf{d}_1$  και  $\mathbf{d}_2$ , και  $\|\mathbf{d}\|$  είναι το μήκος (νόρμα) του διανύσματος  $\mathbf{d}$ .

- Παράδειγμα:

$$\mathbf{d}_1 = \begin{matrix} 3 & 2 & 0 & 5 & 0 & 0 & 0 & 2 & 0 & 0 \end{matrix}$$

$$\mathbf{d}_2 = \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2 \end{matrix}$$

$$\langle \mathbf{d}_1, \mathbf{d}_2 \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$\|\mathbf{d}_1\| = (3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481$$

$$\|\mathbf{d}_2\| = (1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)^{0.5} = (6)^{0.5} = 2.449$$

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = 0.3150$$

# Επεκταμένη Jaccard

## (Extended Jaccard Coefficient – Tanimoto)

- Παραλλαγή της Jaccard για συνεχή γνωρίσματα
  - Ανάγεται στην απλή Jaccard στην περίπτωση δυαδικών διανυσμάτων

$$EJ(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}$$

# Περιεχόμενο Σημερινής Διάλεξης

- Εισαγωγικά στοιχεία μαθήματος
- Δεδομένα
  - Τύποι δεδομένων
  - Ποιότητα δεδομένων
  - Προεπεξεργασία δεδομένων
- **Μέτρα απόστασης και ομοιότητας**
  - Ομοιότητα πολυδιάστατων αριθμητικών δεδομένων
  - Ομοιότητα συνόλων
  - **Ομοιότητα αλφαριθμητικών**

# Απόσταση Διόρθωσης (Edit Distance)

- Δοθέντων δύο αλφαριθμητικών  $S_1$ , και  $S_2$ , ο ελάχιστος αριθμός πράξεων διόρθωσης ώστε να μετασχηματιστεί το ένα στο άλλο
- Οι πράξεις είναι σε επίπεδο χαρακτήρα
  - Εισαγωγή, διαγραφή, αντικατάσταση
- Π.χ.., η edit distance του **dof** με το **dog** είναι 1
  - Του **cat** με το **act** είναι 2
  - Του **cat** με το **dog** είναι 3
- Γενικά υπολογίζεται με δυναμικό προγραμματισμό

# Edit Distance – Υπολογισμός

		f	a	s	t
	0	1	2	3	4
c	1	1	2	3	4
a	2	2	1	2	3
t	3	3	2	2	2
s	4	4	3	2	3

# Edit Distance – Αλγόριθμος

LEVENSHTEINDISTANCE( $s_1, s_2$ )

```
1  for  $i \leftarrow 0$  to  $|s_1|$ 
2  do  $m[i, 0] = i$ 
3  for  $j \leftarrow 0$  to  $|s_2|$ 
4  do  $m[0, j] = j$ 
5  for  $i \leftarrow 1$  to  $|s_1|$ 
6  do for  $j \leftarrow 1$  to  $|s_2|$ 
7    do if  $s_1[i] = s_2[j]$ 
8      then  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]\}$ 
9      else  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1\}$ 
10 return  $m[|s_1|, |s_2|]$ 
```

Operations: insert (cost 1), delete (cost 1), replace (cost 1), copy (cost 0)

# Edit Distance – Αλγόριθμος

LEVENSHTEINDISTANCE( $s_1, s_2$ )

```
1  for  $i \leftarrow 0$  to  $|s_1|$ 
2  do  $m[i, 0] = i$ 
3  for  $j \leftarrow 0$  to  $|s_2|$ 
4  do  $m[0, j] = j$ 
5  for  $i \leftarrow 1$  to  $|s_1|$ 
6  do for  $j \leftarrow 1$  to  $|s_2|$ 
7    do if  $s_1[i] = s_2[j]$ 
8      then  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]\}$ 
9      else  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1\}$ 
10 return  $m[|s_1|, |s_2|]$ 
```

Operations: **insert (cost 1)**, delete (cost 1), replace (cost 1), copy (cost 0)

# Edit Distance – Αλγόριθμος

LEVENSHTEINDISTANCE( $s_1, s_2$ )

```
1  for  $i \leftarrow 0$  to  $|s_1|$ 
2  do  $m[i, 0] = i$ 
3  for  $j \leftarrow 0$  to  $|s_2|$ 
4  do  $m[0, j] = j$ 
5  for  $i \leftarrow 1$  to  $|s_1|$ 
6  do for  $j \leftarrow 1$  to  $|s_2|$ 
7    do if  $s_1[i] = s_2[j]$ 
8      then  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]\}$ 
9      else  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1\}$ 
10 return  $m[|s_1|, |s_2|]$ 
```

Operations: insert (cost 1), delete (cost 1), replace (cost 1), copy (cost 0)

# Edit Distance – Αλγόριθμος

LEVENSHTEINDISTANCE( $s_1, s_2$ )

```
1  for  $i \leftarrow 0$  to  $|s_1|$ 
2  do  $m[i, 0] = i$ 
3  for  $j \leftarrow 0$  to  $|s_2|$ 
4  do  $m[0, j] = j$ 
5  for  $i \leftarrow 1$  to  $|s_1|$ 
6  do for  $j \leftarrow 1$  to  $|s_2|$ 
7    do if  $s_1[i] = s_2[j]$ 
8      then  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]\}$ 
9      else  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1\}$ 
10 return  $m[|s_1|, |s_2|]$ 
```

Operations: insert (cost 1), delete (cost 1), replace (cost 1), copy (cost 0)

# Edit Distance – Αλγόριθμος

LEVENSHTEINDISTANCE( $s_1, s_2$ )

```
1 for  $i \leftarrow 0$  to  $|s_1|$ 
2 do  $m[i, 0] = i$ 
3 for  $j \leftarrow 0$  to  $|s_2|$ 
4 do  $m[0, j] = j$ 
5 for  $i \leftarrow 1$  to  $|s_1|$ 
6 do for  $j \leftarrow 1$  to  $|s_2|$ 
7     do if  $s_1[i] = s_2[j]$ 
8         then  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]\}$ 
9     else  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1\}$ 
10 return  $m[|s_1|, |s_2|]$ 
```

Operations: insert (cost 1), delete (cost 1), replace (cost 1), **copy** (cost 0)

# Edit Distance – Παράδειγμα

		f	a	s	t
	0	1 1	2 2	3 3	4 4
c	1	1 2	2 3	3 4	4 5
a	1	2 1	2 2	3 3	4 4
t	2	2 2	1 3	3 4	4 5
s	2	3 2	3 1	2 2	3 3
	3	3 3	3 2	2 3	2 4
	3	4 3	4 2	3 2	3 2
	4	4 4	4 3	2 3	3 3
	4	5 4	5 3	4 2	3 3

# Κελιά του Πίνακα

cost of getting here from my upper left neighbor <b>(copy or replace)</b>	cost of getting here from my upper neighbor <b>(delete)</b>
cost of getting here from my left neighbor <b>(insert)</b>	the <b>minimum</b> of the three possible “movements”; the cheapest way of getting here

# Hamming Distance

- Δοθέντων δύο αλφαριθμητικών ίσου μεγέθους, η **Hamming Distance** ορίζεται ως ο αριθμός των θέσεων όπου τα αντίστοιχα σύμβολα διαφέρουν
- Εναλλακτικά: ο ελάχιστος αριθμός αντικαταστάσεων που απαιτούνται για το μετασχηματισμό του ενός αλφαριθμητικού στο άλλο
- Παραδείγματα [πηγή: wikipedia]
  - "karolin" and "kathrin" is 3
  - "karolin" and "kerstin" is 3
  - **1011101** and **1001001** is 2
  - **2173896** and **2233796** is 3
- Η Hamming Distance είναι μετρική

# Πηγές Αναφοράς

- “Εισαγωγή στην Εξόρυξη Δεδομένων”, κεφάλαια 1 & 2
  - P-N.Tan, M.Steinbach, V.Kumar
  - Εκδόσεις Τζιόλα
- “Foundations of Multidimensional and Metric Data Structures”
  - Hanan Samet
  - Morgan Kaufmann, 2006, ISBN-10: 0123694469



## 2. Μονομεταβλητή και Διμεταβλητή Ανάλυση



---

Ανάλυση Δεδομένων  
(*Data Analytics*)

Χρήστος Δουλκερίδης  
2024-25

# Ανάλυση μιας Μεταβλητής

# Το Πρόβλημα

- Σύνολο δεδομένων μιας μεταβλητής:
  - Ένα σύνολο από τιμές που έχουν μετρηθεί και αφορούν κάτι συγκεκριμένο
  - Παράδειγμα: μέσος όρος πόντων παίκτων NBA για τη σαιζόν 2019-20
- Ζητούμενο:
  - Να απαντηθούν ορισμένα βασικά ερωτήματα
  - Για καλύτερη κατανόηση του συνόλου δεδομένων
- Πώς;
  - Τεχνικές μονομεταβλητής ανάλυσης δεδομένων

#	PLAYER	PTS
1	James Harden	34.4
2	Bradley Beal	30.5
3	Giannis Antetokounmpo	29.6
4	Trae Young	29.6
5	Damian Lillard	28.9
6	Luka Doncic	28.7
7	Russell Westbrook	27.5
8	Kawhi Leonard	26.9
9	Anthony Davis	26.7
10	Devin Booker	26.1
11	LeBron James	25.7
12	Zach LaVine	25.5
13	Brandon Ingram	24.3
14	Donovan Mitchell	24.2
15	Pascal Siakam	23.6

# Βασικά Ερωτήματα (1/2)

- **Πού** βρίσκονται τα δεδομένα, και πόσο **εύρος** καταλαμβάνουν; Ποιες είναι οι πιο **συνηθισμένες τιμές**, καθώς και **ελάχιστη** και **μέγιστη** τιμή;
- Πώς έχουν κατανεμηθεί τα δεδομένα στο χώρο; Είναι **ομοιόμορφα** ή σχηματίζουν **συστάδες (clusters)** σε συγκεκριμένες περιοχές;
- **Πόσα** σημεία υπάρχουν; Πρόκειται για **μεγάλο** σύνολο δεδομένων ή **σχετικά μικρό**;
- Είναι η **κατανομή συμμετρική** ή όχι; Με άλλα λόγια, είναι η ουρά της κατανομής πολύ μεγαλύτερη στη μία πλευρά από τις δύο;
- Είναι οι ουρές της κατανομής σχετικά «βαριές» (δηλαδή βρίσκονται πολλά σημεία μακριά από την κεντρική περιοχή) ή βρίσκονται όλα τα σημεία – με εξαίρεση πιθανές περιθωριακές τιμές (outliers) - σε μια περιορισμένη περιοχή;

## Βασικά Ερωτήματα (2/2)

- Εάν υπάρχουν **συστάδες**, πόσες είναι; Μία ή πολλές; **Πού** βρίσκονται οι συστάδες και **πόσο μεγάλες** είναι; Τόσο ως προς το εύρος όσο και ως προς το πλήθος σημείων.
- Περιέχει το σύνολο δεδομένων **περιθωριακά στοιχεία**; Δηλαδή στοιχεία που διαφέρουν από όλα τα υπόλοιπα.
- Τέλος, υπάρχουν τίποτα **ασυνήθιστα** ή **σημαντικά χαρακτηριστικά** στο σύνολο δεδομένων; Κενά, απότομες αλλαγές, ασυνήθιστες τιμές, οτιδήποτε μπορεί να παρατηρηθεί.

# Ορισμένα Εργαλεία για Μονομεταβλητή Ανάλυση

- **Συνοπτική στατιστική (Descriptive statistics)**
- Dot και jitter διαγράμματα
- Ιστογράμματα
- Εκτιμητές πυρήνα (kernel density estimates)
- Συνάρτηση αθροιστικής κατανομής
- Θηκογράμματα

# Συνοπτική Στατιστική (Descriptive Statistics)

- Τα στατιστικά δεδομένα παρουσιάζονται συνοπτικά μέσω:
  - αριθμητικών μέτρων, όπως **μέτρων θέσης** και **διασποράς**, ή
  - κατάλληλων διαγραμμάτων
- Η **συνοπτική στατιστική** περιλαμβάνει **ποσοτικά μέτρα**, όπως
  - ο **μέσος** και
  - η **τυπική απόκλιση**

Που καταγράφουν διάφορα χαρακτηριστικά ενός συνόλου δεδομένων, με ένα μόνο αριθμό (ή με μικρό σύνολο αριθμών)

- Μέσο εισόδημα νοικοκυριών
- Ποσοστό φοιτητών που ολοκληρώνουν τις σπουδές σε 4 χρόνια

# Συνοπτική Στατιστική (Descriptive Statistics)

- Συχνότητες και επικρατούσα τιμή
- Εκατοστημόρια
- Μέτρα θέσης: μέσος και διάμεσος
- Μέτρα διασποράς: εύρος και διακύμανση

# Συχνότητες και Επικρατούσα Τιμή

- Θεωρήστε ένα σύνολο μη διατεταγμένων, κατηγορικών τιμών  $\{v_1, v_2, \dots, v_i, \dots, v_k\}$  που μπορεί να λάβει ένα γνώρισμα  $x$
- Έστω  $m$  το σύνολο των αντικειμένων
- Υπολογισμός **συχνότητας εμφάνισης (frequency)**:
  - $frequency(v_i) = N / m$
  - όπου  $N$  το πλήθος αντικειμένων με τιμή  $v_i$
- **Επικρατούσα τιμή (mode)**: ενός κατηγορικού γνωρίσματος είναι η τιμή που έχει τη μεγαλύτερη συχνότητα

# Εκατοστημόρια

- Για διατεταγμένα δεδομένα, είναι πιο χρήσιμα τα **εκατοστημόρια (percentiles)** ενός συνόλου τιμών
  - Το **p-ποσοστιαίο σημείο** ενός δείγματος με **n** παρατηρήσεις ορίζεται ως η παρατήρηση για την οποία **το πολύ p%** των παρατηρήσεων είναι μικρότερες από αυτή
  - (Γνωστά και ως **quantiles**: το 10-percentile ισούται με το 0,1 quantile)
  - Αν οι παρατηρήσεις είναι διατεταγμένες σε αύξουσα διάταξη, το p-ποσοστιαίο σημείο είναι στη θέση: **(n+1)p/100** όπου  $1 \leq p \leq 99$
  - Για  $p=25, 50, 75$  έχουμε πρώτο, δεύτερο και τρίτο τεταρτημόριο αντίστοιχα

# Εκατοστημόρια – Παράδειγμα

Σύνολο δεδομένων 20 εγγραφών:  $\{x_i\}$

15	3	8	11	4	1	5	9	6	2	12	11	13	8	6	7	3	5	11	6
----	---	---	----	---	---	---	---	---	---	----	----	----	---	---	---	---	---	----	---

Ταξινομημένες εγγραφές

1	2	3	3	4	5	5	6	6	6	7	8	8	9	11	11	11	12	13	15
---	---	---	---	---	---	---	---	---	---	---	---	---	---	----	----	----	----	----	----



10-ποσοστιαίο σημείο: 2



90-ποσοστιαίο σημείο: 13

# Μέτρα Θέσης: Μέσος και Διάμεσος

## ■ Μέσος (mean)

- Είναι ευαίσθητος σε ακραίες τιμές

$$\text{mean}(x) = \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$$

## ■ Διάμεσος (median)

- Η τιμή της μεσαίας παρατήρησης (στοιχείου), όταν οι παρατηρήσεις (στοιχεία) ταξινομηθούν με φθίνουσα ή αύξουσα διάταξη
- $\{x_{(1)}, \dots, x_{(m)}\}$  ταξινομημένα σε μη φθίνουσα σειρά, τότε:  $x_{(1)} = \min(x)$  και  $x_{(m)} = \max(x)$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r + 1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

## ■ Περικομμένος μέσος (trimmed mean)

- Ορίζεται ένα ποσοστό  $p$  μεταξύ 0 και 100, εξαιρούνται το ανώτερο και κατώτερο  $(p/2)\%$  των δεδομένων και υπολογίζεται ο μέσος
- Διάμεσος = περικομμένος μέσος με  $p=100\%$
- Μέσος = περικομμένος μέσος με  $p=0\%$

# Μέσος, Διάμεσος – Παράδειγμα

Σύνολο δεδομένων 20 εγγραφών:  $\{x_i\}$

15	3	8	11	4	1	5	9	6	2	12	11	13	8	6	7	3	5	11	6
----	---	---	----	---	---	---	---	---	---	----	----	----	---	---	---	---	---	----	---

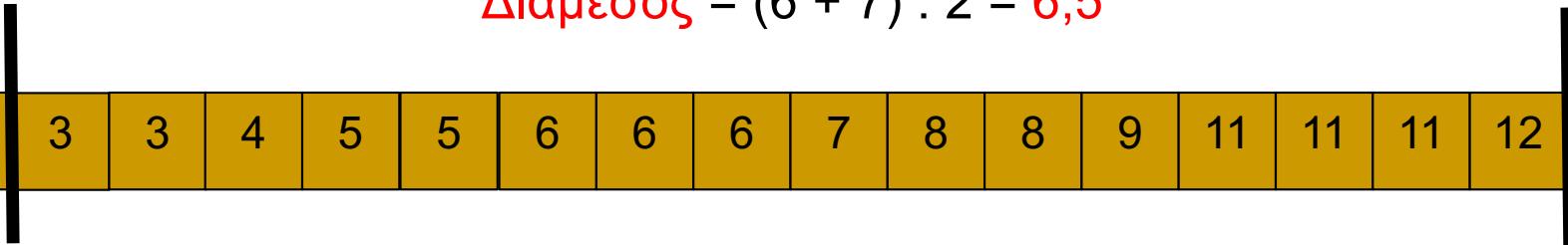
$$\text{Μέσος} = (1/20) \sum x_i = 7,3$$

1	2	3	3	4	5	5	6	6	6	7	8	8	9	11	11	11	12	13	15
---	---	---	---	---	---	---	---	---	---	---	---	---	---	----	----	----	----	----	----



$$\text{Διάμεσος} = (6 + 7) : 2 = 6,5$$

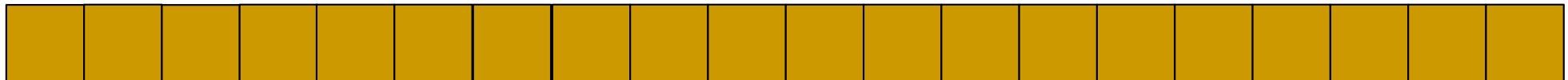
4	2	3	3	4	5	5	6	6	6	7	8	8	9	11	11	11	12	13	15
---	---	---	---	---	---	---	---	---	---	---	---	---	---	----	----	----	----	----	----



$$\text{Περικομμένος μέσος (για } p=20\%) = 7,19$$

# Μέσος ή Διάμεσος;

Σύνολο δεδομένων 20 εγγραφών:  $\{x_i\}$



Δίνεται ότι Μέσος = 10

Περίπτωση #1:

Διάμεσος = 10



Περίπτωση #2:

Διάμεσος = 10



Περίπτωση #3:

Διάμεσος = 1,5



# Μέτρα Διασποράς: Εύρος και Διακύμανση

- Εύρος (range) είναι η διαφορά μεταξύ μέγιστης και ελάχιστης τιμής
- Η διακύμανση (variance)  $s_x^2$  είναι το πιο κοινό μέτρο διασποράς, όπου  $s_x$  η τυπική απόκλιση (standard deviation)

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2$$

$i$	$x_i$	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	7	6	1	1
2	8	6	2	4
3	9	6	3	9
4	6	6	0	0
5	5	6	-1	1
6	4	6	-2	4
7	3	6	-3	9

$$s_x^2 = 28 / (7 - 1) = 4.667$$

$$s_x = 2.16$$

# Μέτρα Διασποράς: Εύρος και Διακύμανση

- Πιο εύρωστες εκτιμήσεις διασποράς (λιγότερο ευαίσθητες σε ακραίες τιμές):

Απόλυτη μέση απόκλιση

$$\text{AAD}(x) = \frac{1}{m} \sum_{i=1}^m |x_i - \bar{x}|$$

Διάμεσος απόλυτη  
απόκλιση

$$\text{MAD}(x) = \text{median}\left(\{|x_1 - \bar{x}|, \dots, |x_m - \bar{x}|\}\right)$$

Ενδοτεταρτημοριακό εύρος

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

# Παράδειγμα

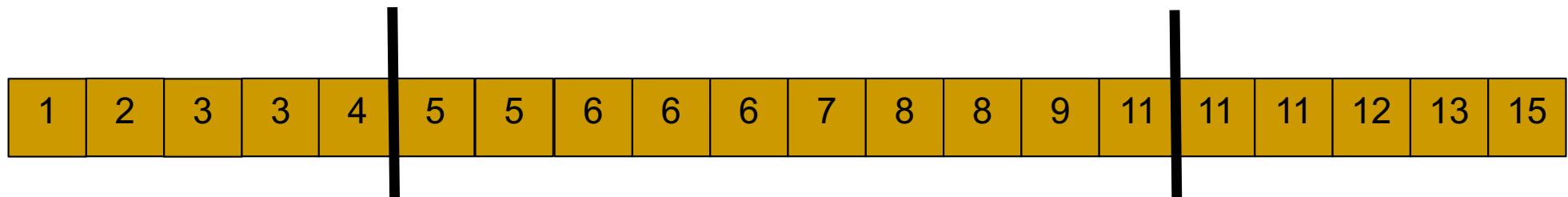
Σύνολο δεδομένων 20 εγγραφών:  $\{x_i\}$

15	3	8	11	4	1	5	9	6	2	12	11	13	8	6	7	3	5	11	6
----	---	---	----	---	---	---	---	---	---	----	----	----	---	---	---	---	---	----	---

Ελάχιστη τιμή = 1

Μέγιστη τιμή = 15

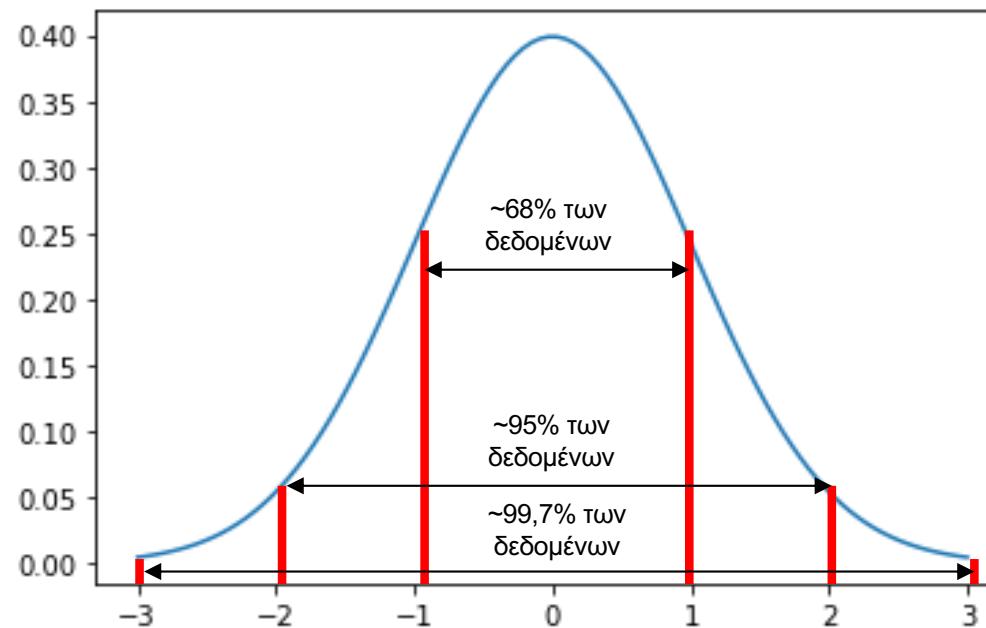
Εύρος =  $15 - 1 = 14$



Ενδοτεταρτημοριακό εύρος IQR = 6,5 (η τιμή εξαρτάται από τον τρόπο υπολογισμού)

# Προβλήματα

- Συχνά (όχι πάντα!), σε πολλά σύνολα δεδομένων αναμένεται
  - Τα **2/3** των σημείων να βρίσκονται στο διάστημα
$$[\text{mean}(x) - s_x, \text{mean}(x) + s_x]$$
  - Το **99%** των σημείων να βρίσκονται στο ευρύτερο διάστημα
$$[\text{mean}(x) - 3s_x, \text{mean}(x) + 3s_x]$$



# Προβλήματα

- Αυτές οι (εύκολα υπολογιζόμενες) ποσότητες:  $\text{mean}(x)$ ,  $s_x$  έχουν νόημα μόνο όταν τα δεδομένα είναι συμμετρικά και δεν περιέχουν περιθωριακές τιμές (outliers)
  - Παράδειγμα:
    - Καλάθι με 10 προϊόντα με 1€ και 1 προϊόν με 20€
    - $\text{mean}(x) = 2,73\text{€}$  και  $s_x = 5,46\text{€}$
    - Άρα (?) τα περισσότερα προϊόντα είναι μεταξύ:

$$[2,73\text{€} - 5,46\text{€}, 2,73\text{€} + 5,46\text{€}]$$

# Προσοχή

- Απλά αριθμητικά μέτρα, όπως:
  - μέσος (mean),
  - διάμεσος (median),
  - τυπική απόκλιση (standard deviation),
  - $p$ -ποσοστιαία σημεία (quantiles)
- έχουν νόημα μόνο υπό συγκεκριμένες προϋποθέσεις:
  - Μόνο όταν η κατανομή στην οποία εφαρμόζονται έχει ένα κεντρικό σημείο
  - Διαφορετικά, τα συμπεράσματα που βασίζονται σε τέτοια μέτρα θα είναι λάθος

# Συμπεράσματα (1/2)

- Μελετώντας ένα σύνολο τιμών
- Όταν δεν ακολουθεί **συμμετρική κατανομή** ή περιέχει **περιθωριακές τιμές**
  - Προτιμούμε τη **διάμεσο** και τα **ποσοστιαία σημεία**
  - Αντί της **μέσης τιμής** και της **τυπικής απόκλισης**
- Όταν το δείγμα ακολουθεί **συμμετρική κατανομή**
  - Η μέση τιμή και η διάμεσος θα είναι κοντά
  - Οποιοδήποτε μπορεί να χρησιμοποιηθεί
- Π.χ. για τον υπολογισμό του μέσου μισθού ενός νοικοκυριού χρησιμοποιείται η διάμεσος

# Συμπεράσματα (2/2)

- Όταν υπάρχουν **περιθωριακές τιμές**
  - Η υπόθεση ότι η τυπική απόκλιση υποδεικνύει το εύρος της κατανομής τιμών παραβιάζεται
  - **Προτιμάται η χρήση του IQR**
- Γιατί δε χρησιμοποιούμε πάντα διάμεσο και ποσοστιαία σημεία;
  - Η μέση τιμή και η τυπική απόκλιση υπολογίζονται ευκολότερα
  - Οι ποσοστιαίες τιμές απαιτούν ταξινόμηση, άρα **O(n log n)** αντί για **O(n)**

# Ορισμένα Εργαλεία για Μονομεταβλητή Ανάλυση

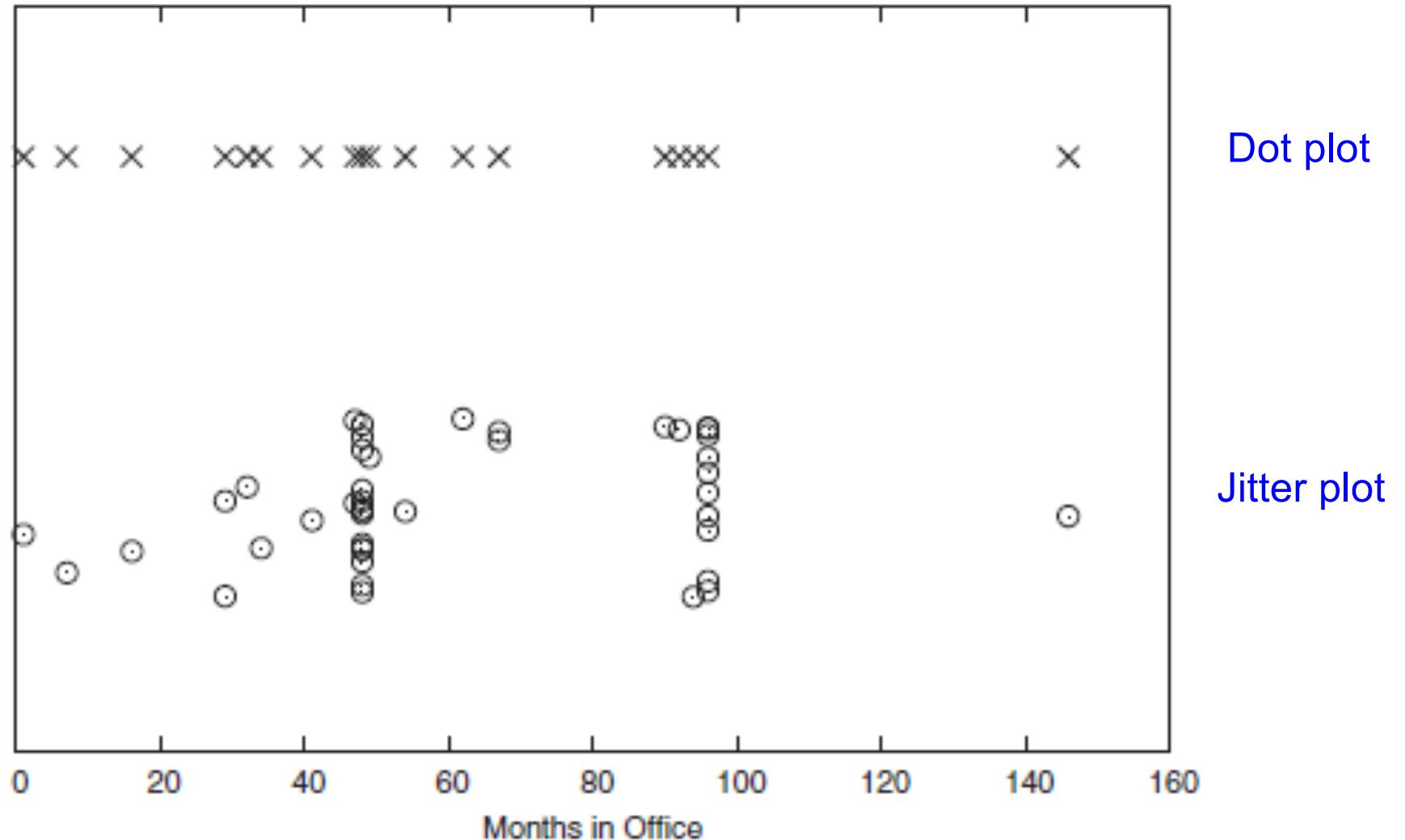
- Συνοπτική στατιστική (Descriptive statistics)
- **Dot και jitter διαγράμματα**
- Ιστογράμματα
- Εκτιμητές πυρήνα (kernel density estimates)
- Συνάρτηση αθροιστικής κατανομής
- Θηκογράμματα

# Dot και Jitter Διαγράμματα

- Το παρακάτω σύνολο δεδομένων περιέχει μια εγγραφή για κάθε πρόεδρο των Η.Π.Α. και τους μήνες θητείας του

1	Washington	94	16	Lincoln	49	30	Coolidge	67
2	Adams	48	17	Johnson	47	31	Hoover	48
3	Jefferson	96	18	Grant	96	32	Roosevelt	146
4	Madison	96	19	Hayes	48	33	Truman	92
5	Monroe	96	20	Garfield	7	34	Eisenhower	96
6	Adams	48	21	Arthur	41	35	Kennedy	34
7	Jackson	96	22	Cleveland	48	36	Johnson	62
8	Van Buren	48	23	Harrison	48	37	Nixon	67
9	Harrison	1	24	Cleveland	48	38	Ford	29
10	Tyler	47	25	McKinley	54	39	Carter	48
11	Polk	48	26	Roosevelt	90	40	Reagan	96
12	Taylor	16	27	Taft	48	41	Bush	48
13	Fillmore	32	28	Wilson	96	42	Clinton	96
14	Pierce	48	29	Harding	29	43	Bush	96
15	Buchanan	48						

# Dot και Jitter Διαγράμματα



Πλήθος μηνών θητείας (*months in office*) προέδρων των Η.Π.Α.

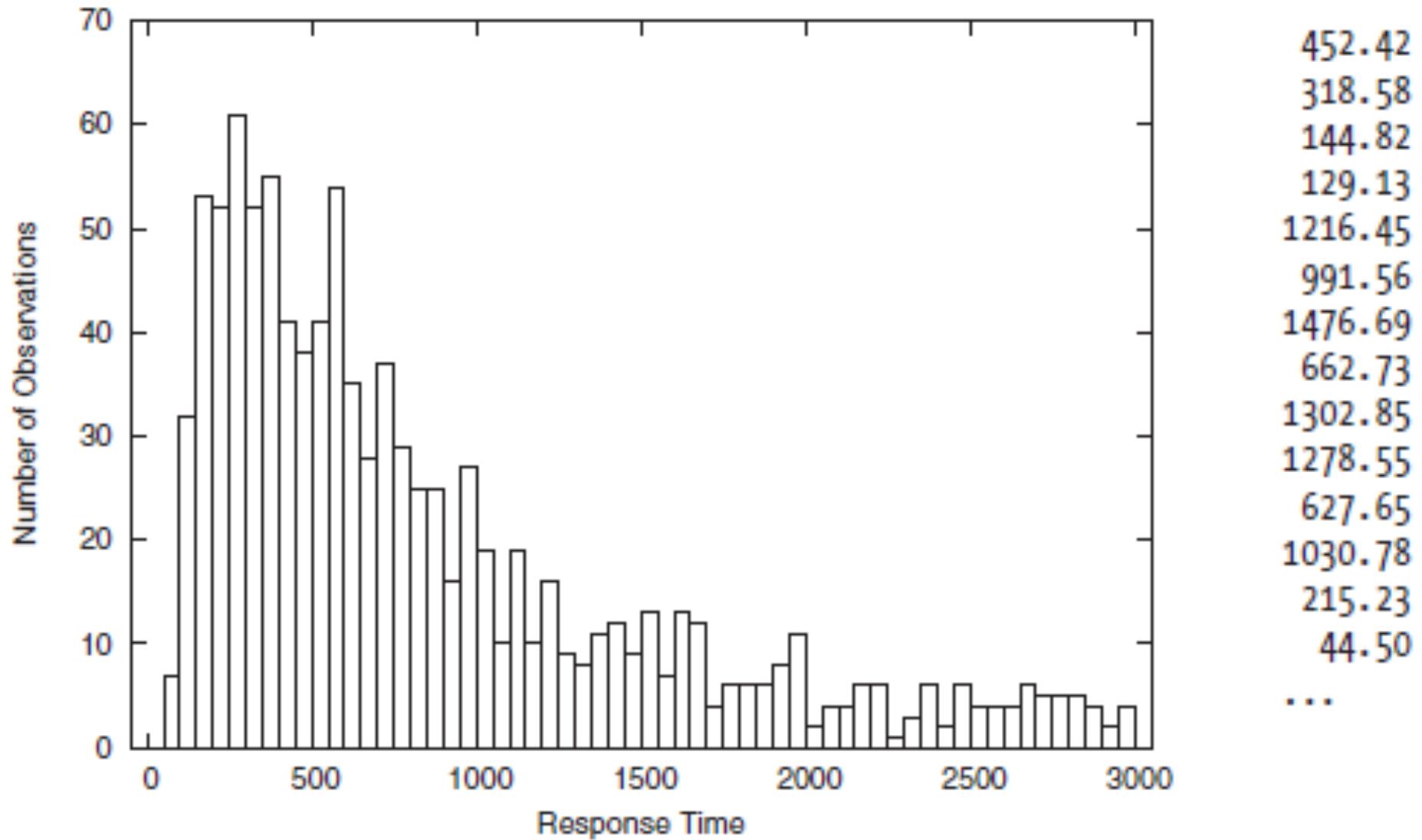
# Ορισμένα Εργαλεία για Μονομεταβλητή Ανάλυση

- Συνοπτική στατιστική (Descriptive statistics)
- Dot και jitter διαγράμματα
- **Ιστογράμματα**
- Εκτιμητές πυρήνα (kernel density estimates)
- Συνάρτηση αθροιστικής κατανομής
- Θηκογράμματα

# Ιστογράμματα (Histograms)

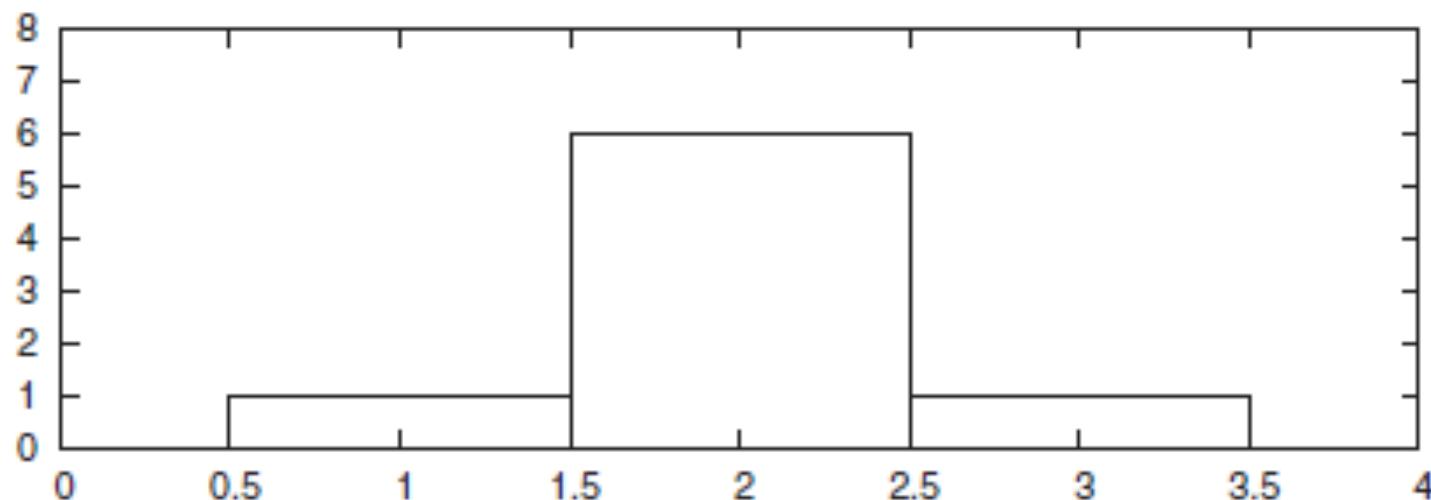
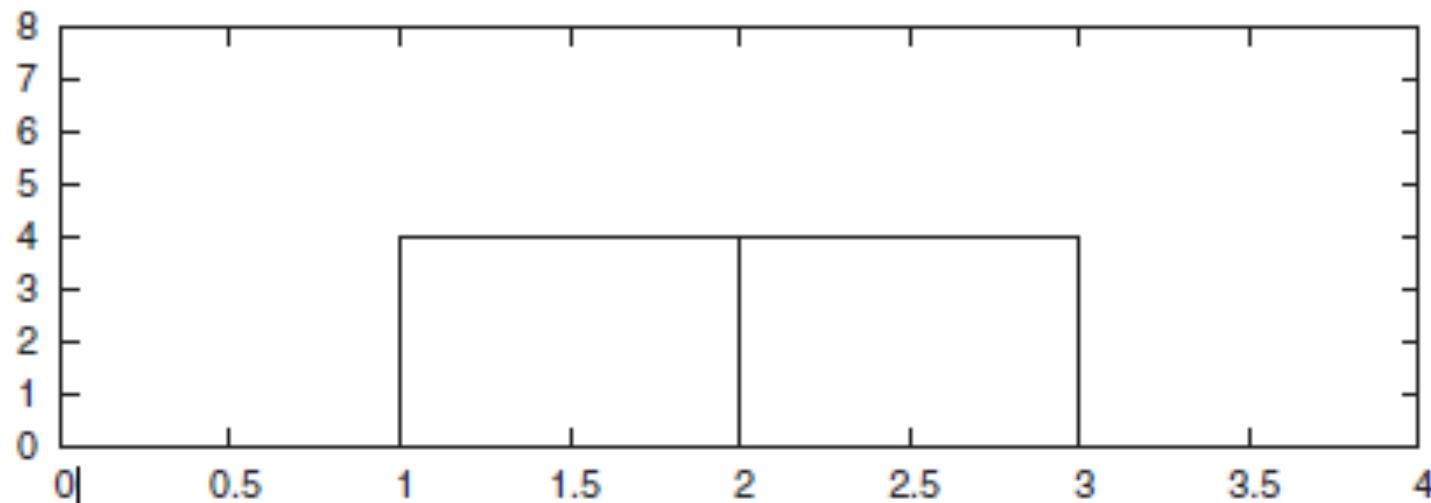
- Μια συνοπτική οπτική περιγραφή της κατανομής των δεδομένων
- Για να παράξουμε ένα **ιστόγραμμα**
  - χωρίζουμε το εύρος τιμών σε ένα σύνολο από **κάδους (bins)** και
  - **μετράμε το πλήθος των σημείων** που ανήκουν σε κάθε κάδο
- Κατόπιν, σχεδιάζουμε τις **μετρήσεις για κάθε κάδο** σαν συνάρτηση της θέσης του κάδου

# Ιστογράμματα (Histograms)



Ένα ιστόγραμμα με τους χρόνους απόκρισης ενός εξυπηρετητή

# Ιστογράμματα (Histograms)



Τα ιστογράμματα μπορεί να διαφέρουν στην εμφάνιση ανάλογα με την επιλογή του κεντρικού σημείου για το πρώτο bin

# Μειονεκτήματα Ιστογραμμάτων

- Υπάρχει **απώλεια πληροφορίας**, αφού αντικαθίσταται η θέση ενός σημείου με έναν κάδο (bin) με πεπερασμένο εύρος
- Τα ιστογράμματα **δεν ορίζονται με μοναδικό τρόπο**
- Οι τιμές σε ένα ιστόγραμμα **δε μεταβάλλονται με ομαλό τρόπο**, αλλά απότομα
- Τα ιστογράμματα **δεν μπορούν να χειριστούν καλά περιθωριακές τιμές (outliers)**
  
- Για τον καθορισμό του εύρους μπορεί να χρησιμοποιηθεί ο κανόνας του Scott (Scott's rule):
  - Κάνει υπόθεση κανονικής κατανομής  $w = 3.5\sigma / \sqrt[3]{n}$

# Ιστογράμματα – Επιπλέον Παρατηρήσεις

- Δύο παράμετροι επηρεάζουν την κατασκευή:
  - το **εύρος του κάδου** (bin) και
  - το **σημείο εκκίνησης στον άξονα**
- Κανονικοποιημένα ή όχι ιστογράμματα:
  - **Μη-κανονικοποιημένα**: εμφανίζεται το πλήθος των εγγραφών
  - **Κανονικοποιημένα**: διαιρούμε το πλήθος με το σύνολό τους
  - Προσοχή: να φαίνεται κάπως αν το ιστόγραμμα είναι κανονικοποιημένο
- Υπόθεση (ως τώρα): όλοι οι κάδοι έχουν το ίδιο εύρος
  - Δεν είναι απαραίτητο
  - Μπορούμε να έχουμε **πιο στενούς κάδους** εκεί όπου τα δεδομένα είναι περισσότερα και **πιο αραιούς κάδους** εκεί όπου είναι λίγα

# Ορισμένα Εργαλεία για Μονομεταβλητή Ανάλυση

- Συνοπτική στατιστική (Descriptive statistics)
- Dot και jitter διαγράμματα
- Ιστογράμματα
- **Εκτιμητές πυρήνα (kernel density estimates)**
- Συνάρτηση αθροιστικής κατανομής
- Θηκογράμματα

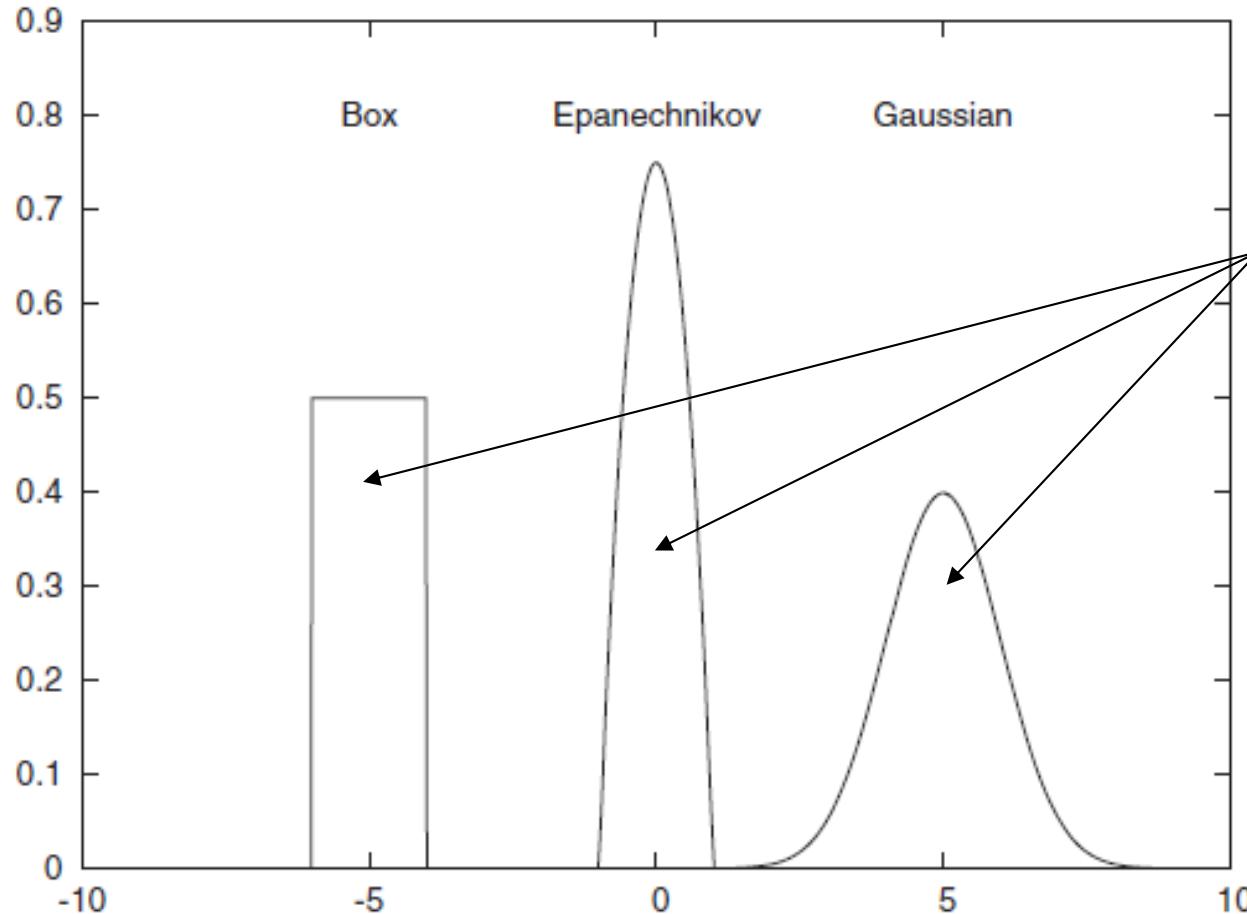
# Εκτιμητές Πυρήνα

## Kernel Density Estimates (KDE)

- Μια μη-παραμετρική μέθοδος για την εκτίμηση της συνάρτησης πυκνότητας πιθανότητας ενός συνόλου δεδομένων
- Με χρήση μιας συνάρτησης πυρήνα (*kernel*)
- Μια ομαλή συνάρτηση με μια έντονη κορυφή

$$K(x) = \begin{cases} \frac{1}{2} & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{box or boxcar kernel}$$
$$K(x) = \begin{cases} \frac{3}{4} (1 - x^2) & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{Epanechnikov kernel}$$
$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \quad \text{Gaussian kernel}$$

# Εκτιμητές Πυρήνα Kernel Density Estimates (KDE)



Το εμβαδό της περιοχής κάτω από την καμπύλη πυρήνα πρέπει να ισούται με 1

# Εκτιμητές Πυρήνα Kernel Density Estimates (KDE)

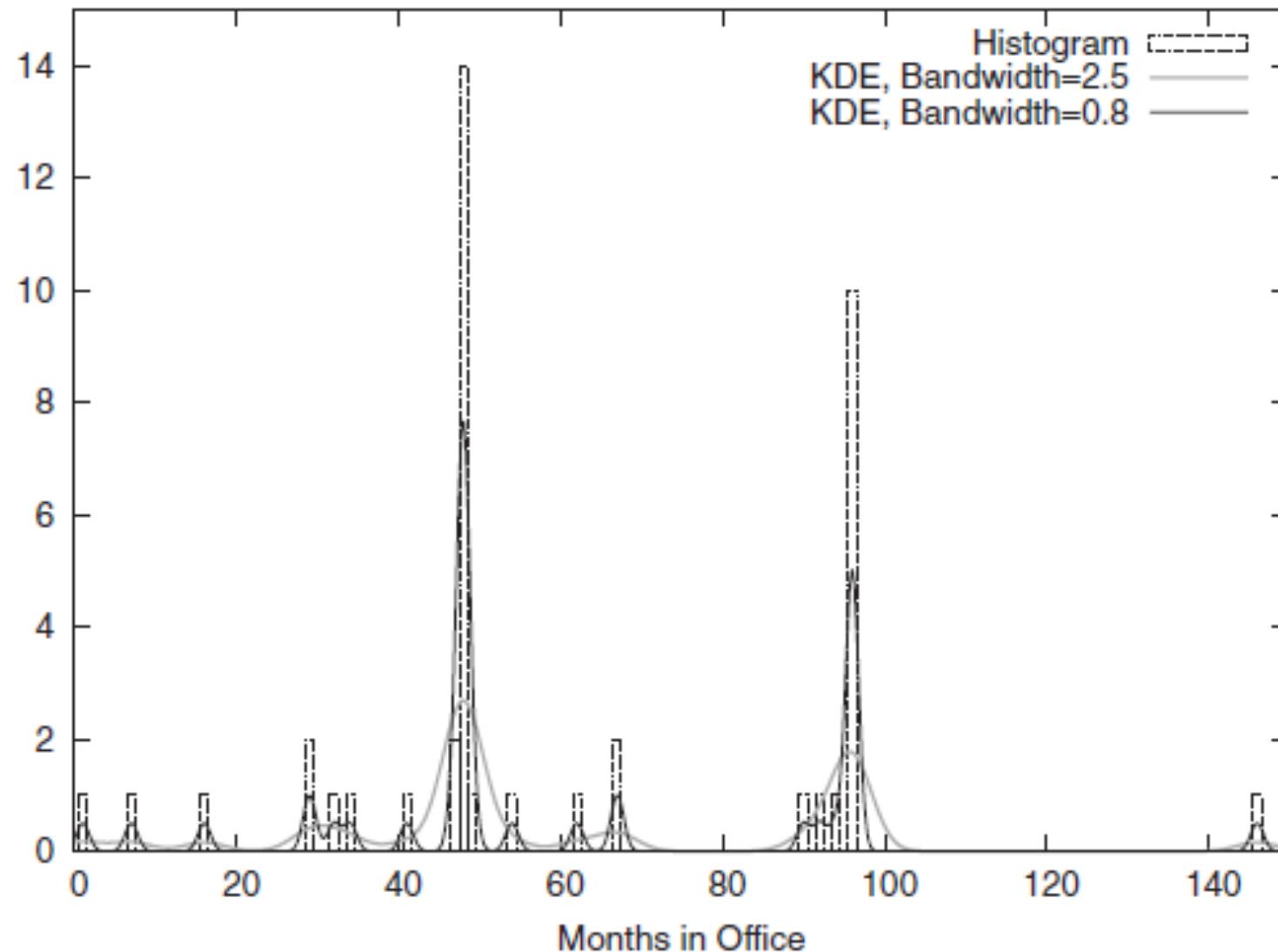
- Για να υπολογιστεί ένας εκτιμητής πυρήνα (KDE)
  - Τοποθετούμε τη συνάρτηση πυρήνα (*kernel*) στη θέση κάθε σημείου του συνόλου δεδομένων
- Στη συνέχεια, αθροίζουμε τις συνεισφορές όλων των kernels για να εξαχθεί μια ομαλή καμπύλη, την οποία μπορούμε να αποτιμήσουμε για οποιοδήποτε σημείο στο x άξονα

Για σύνολο δεδομένων:

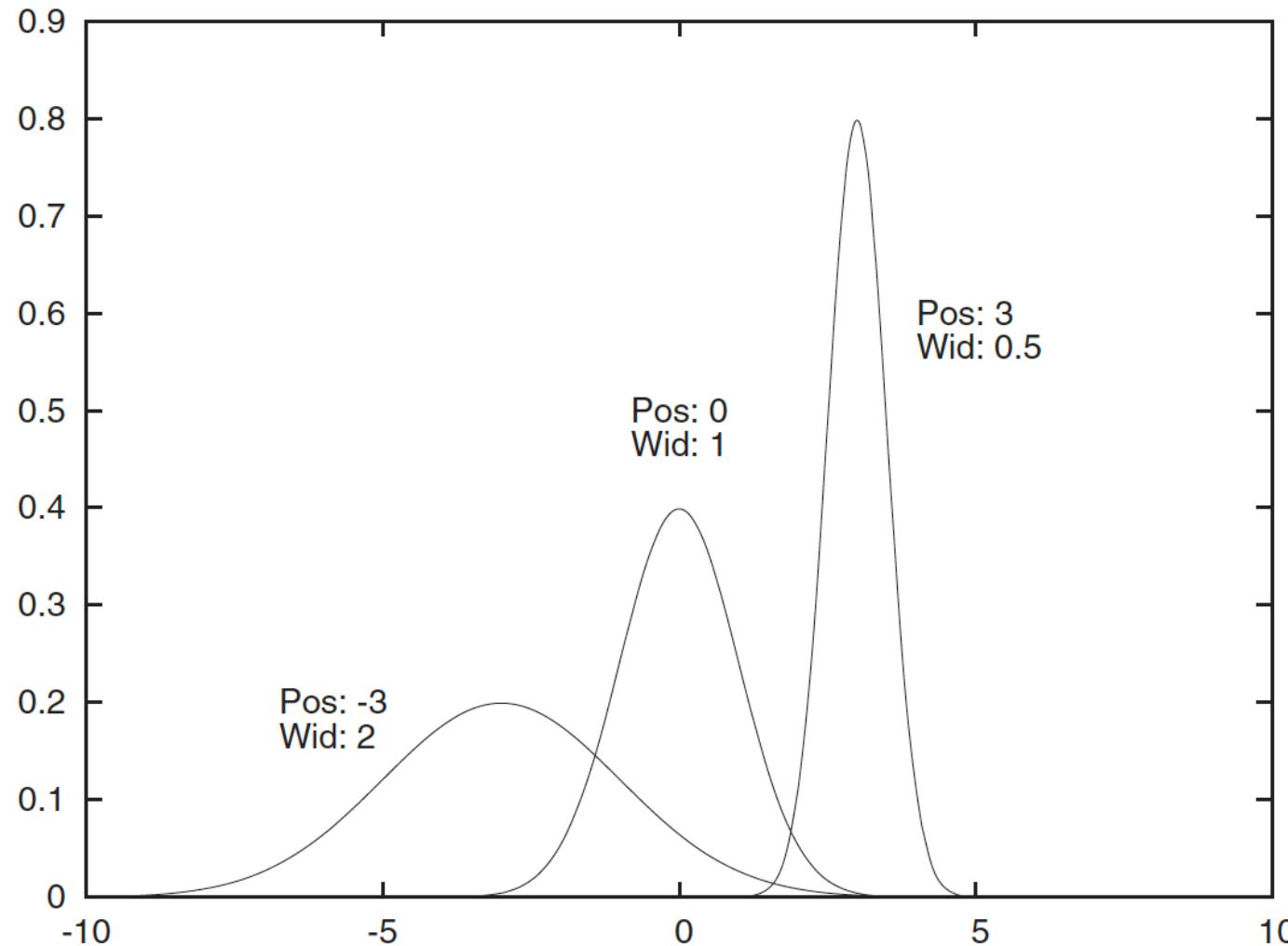
$$\{x_1, x_2, \dots, x_n\}$$

$$\sum_{i=1}^n \frac{1}{h} K \left( \frac{x - x_i}{h} \right)$$

# Kernel Density Estimates (KDE)



# O Gaussian Kernel για Διαφορετικές Τιμές Εύρους $h$

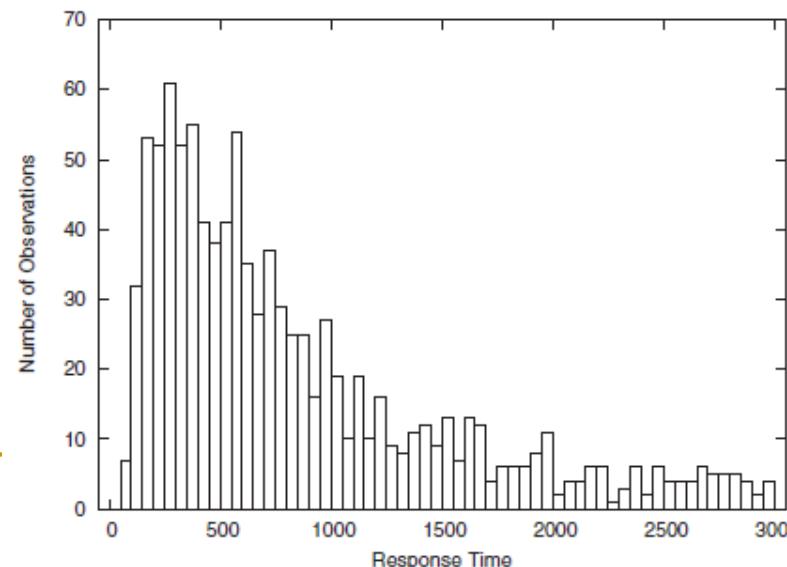


# Ορισμένα Εργαλεία για Μονομεταβλητή Ανάλυση

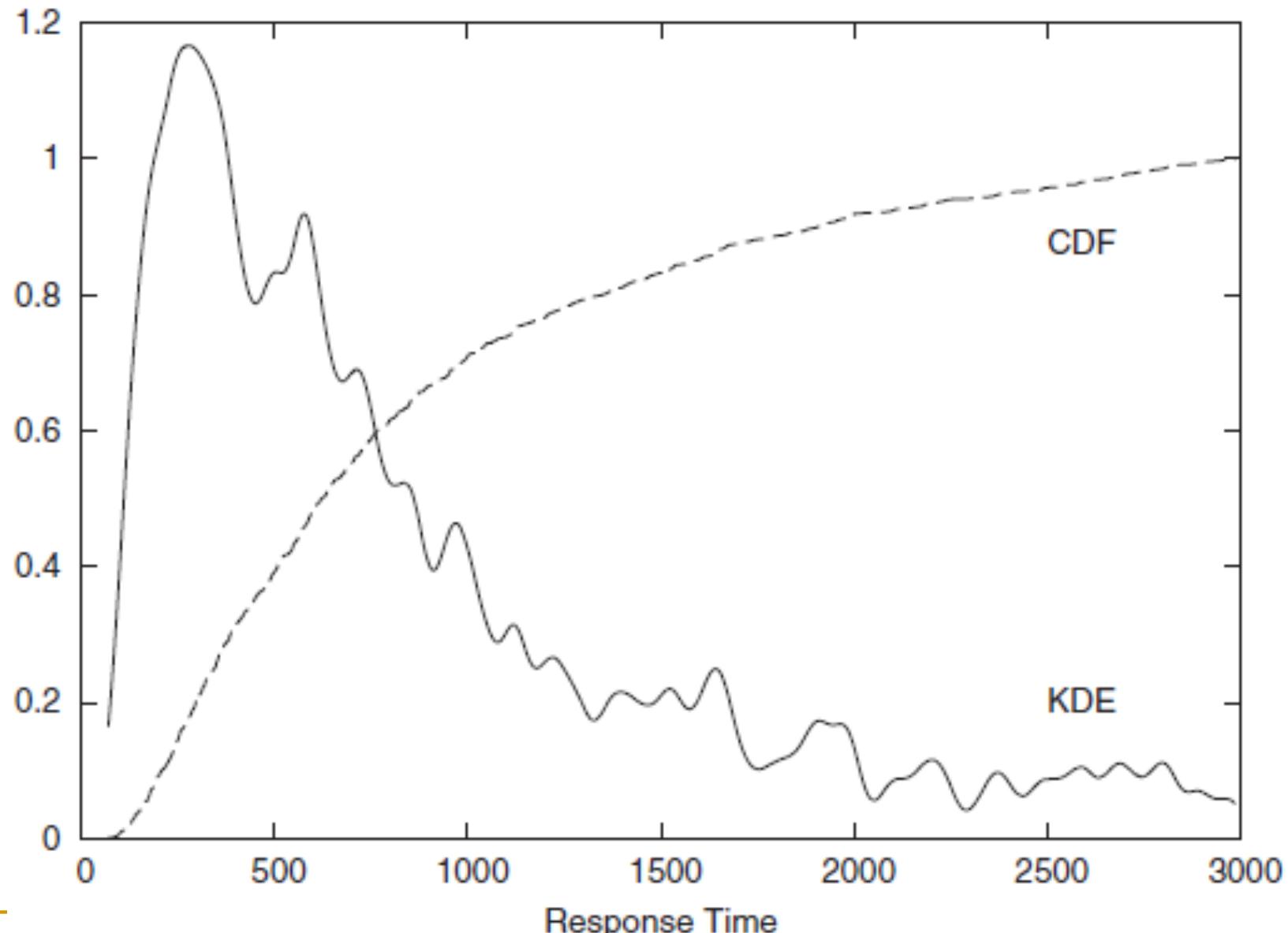
- Συνοπτική στατιστική (Descriptive statistics)
- Dot και jitter διαγράμματα
- Ιστογράμματα
- Εκτιμητές πυρήνα (kernel density estimates)
- **Συνάρτηση αθροιστικής κατανομής**
- Θηκογράμματα

# Συνάρτηση Αθροιστικής Κατανομής (Cumulative Distribution Function)

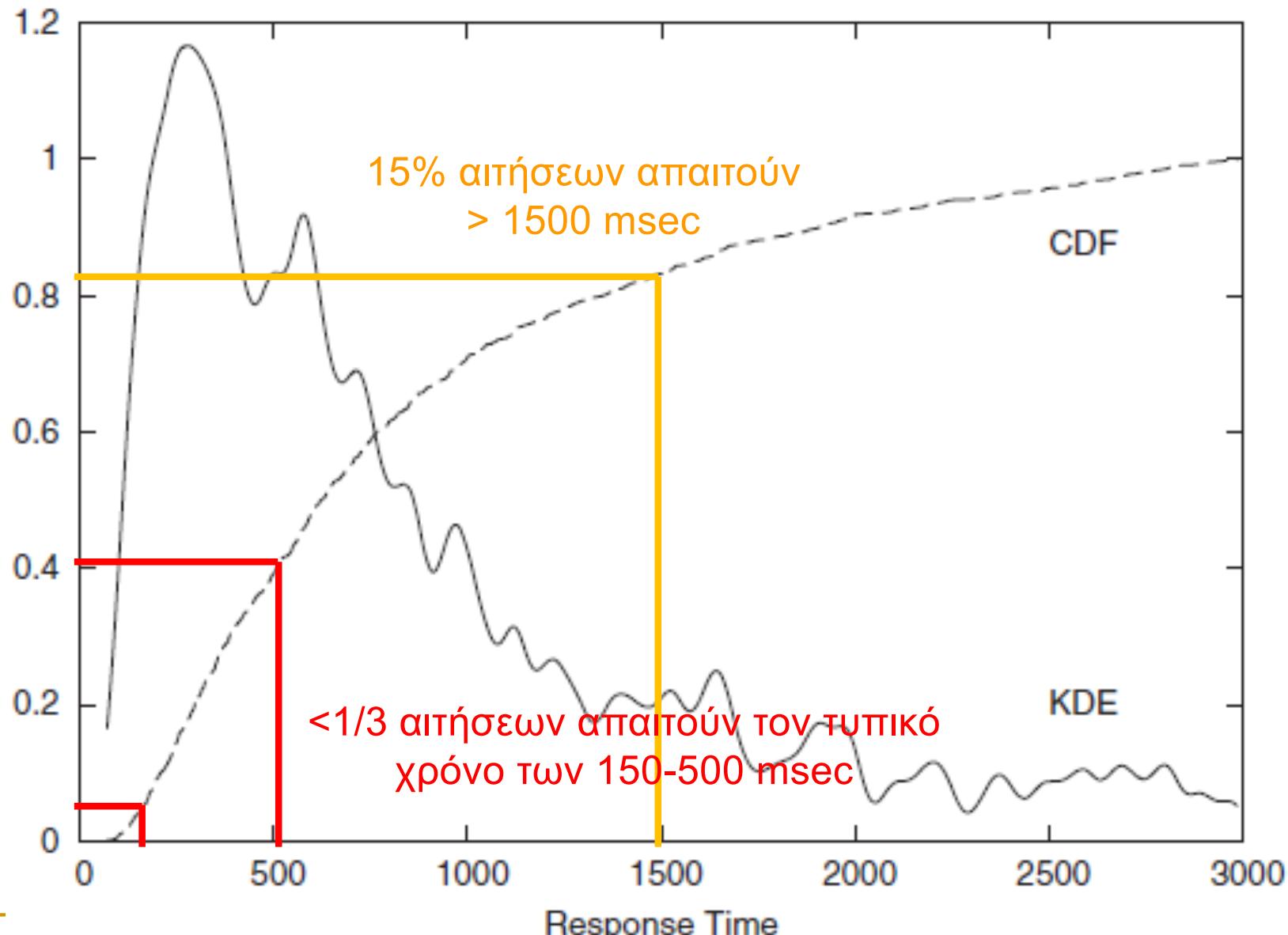
- Κύριο πλεονέκτημα των ιστογραμμάτων και KDE → διαισθητικά κατανοητά
  - Περιγράφουν πόσο πιθανό είναι να βρεθεί ένα στοιχείο του συνόλου δεδομένων με μια συγκεκριμένη τιμή
  - Όμως πόσο ακριβώς είναι αυτή η πιθανότητα;
  - Π.χ. ποιο ποσοστό αιτήσεων ολοκληρώνεται μεταξύ 150 και 350 msec;
- Η *cumulative distribution function (CDF)* πετυχαίνει αυτό το σκοπό
  - Η CDF στη θέση  $x$  δίνει το ποσοστό των σημείων  $x_i$  με  $x_i \leq x$



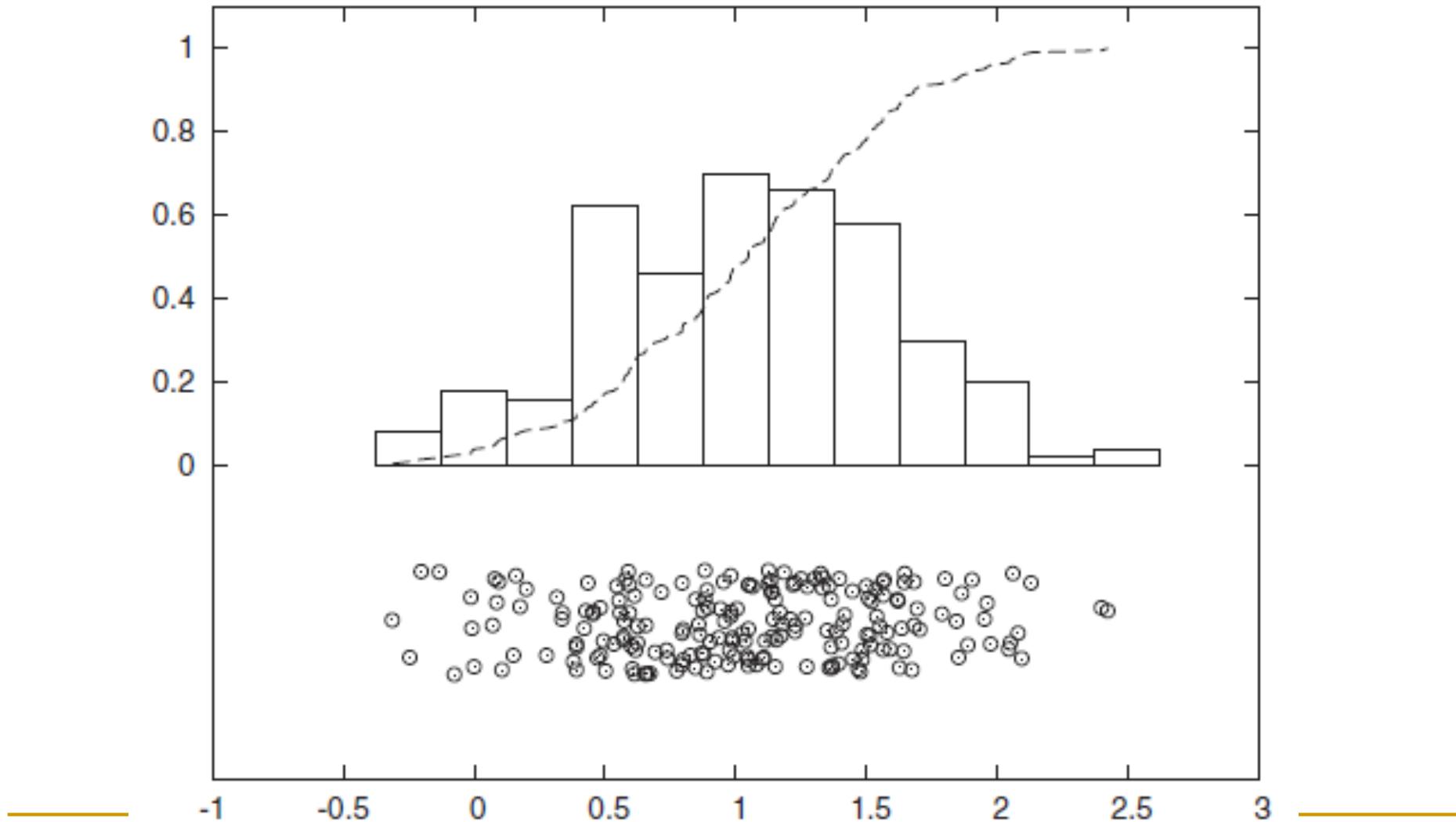
# Συνάρτηση Αθροιστικής Κατανομής



# Διαβάζοντας Ποσοτική Πληροφορία



# Jitter plot, Ιστόγραμμα, και CDF για ένα Σύνολο Δεδομένων που ακολουθεί Κανονική Κατανομή



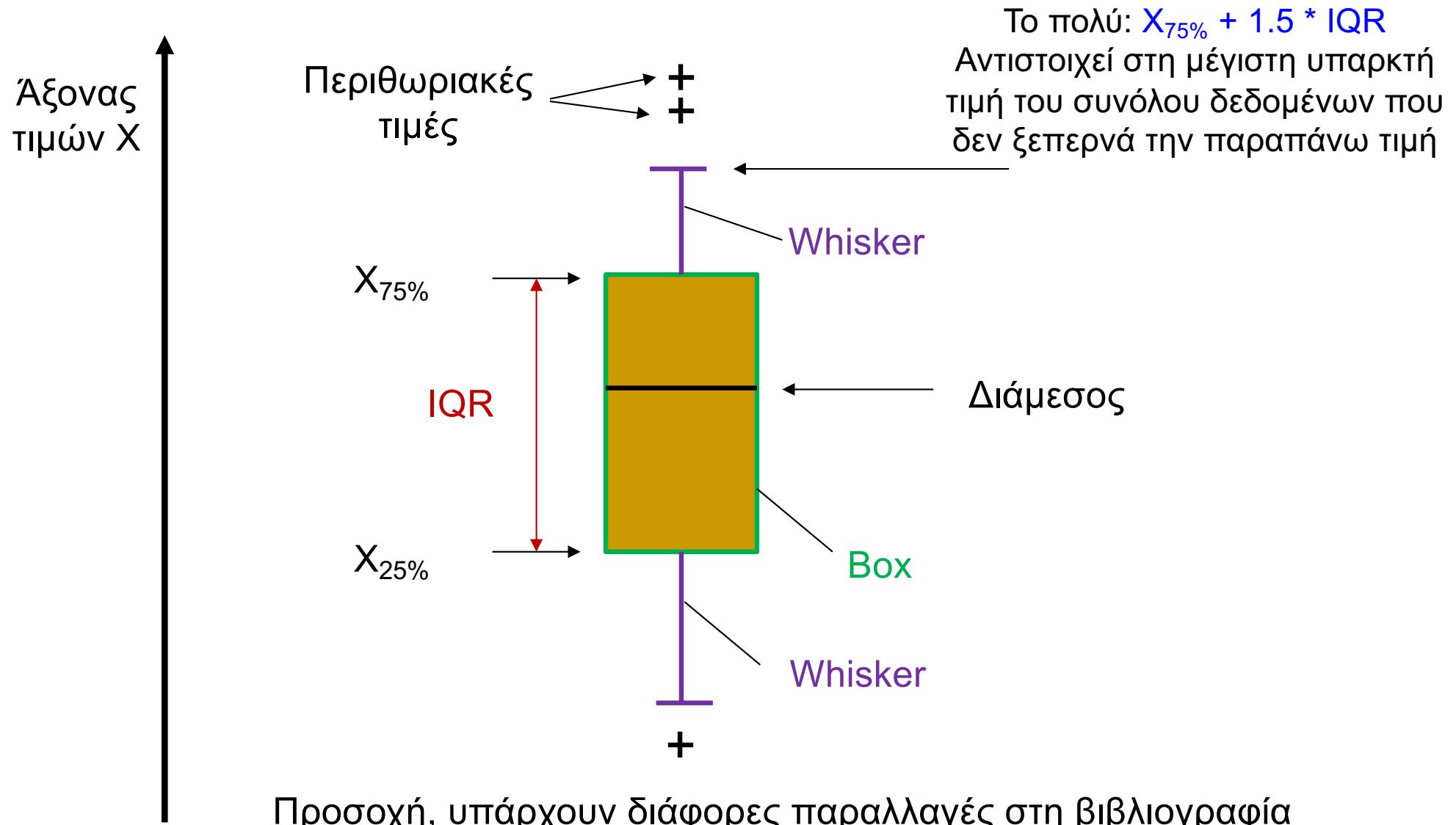
# Ιδιότητες των CDF

- Μια CDF είναι μονότονα αύξουσα με το  $x$  (επειδή η τιμή της  $CDF(x)$  στη θέση  $x$  είναι ίση με το ποσοστό των σημείων αριστερά του  $x$ )
- Μια CDF είναι πιο ομαλή από ένα ιστόγραμμα, και περιέχει την ίδια πληροφορία
- Οι CDFs δεν οδηγούν σε απώλεια πληροφορίας και αποτελούν πιο πιστές αναπαραστάσεις των δεδομένων από ότι τα ιστογράμματα
- Οι CDFs είναι συνήθως κανονικοποιημένες και προσεγγίζουν το 1 (ή 100%) όταν το  $x \rightarrow \infty$ , και το 0 όταν το  $x \rightarrow -\infty$
- Η CDF ενός συνόλου δεδομένων είναι μοναδική

# Ορισμένα Εργαλεία για Μονομεταβλητή Ανάλυση

- Συνοπτική στατιστική (Descriptive statistics)
- Dot και jitter διαγράμματα
- Ιστογράμματα
- Εκτιμητές πυρήνα (kernel density estimates)
- Συνάρτηση αθροιστικής κατανομής
- Θηκογράμματα

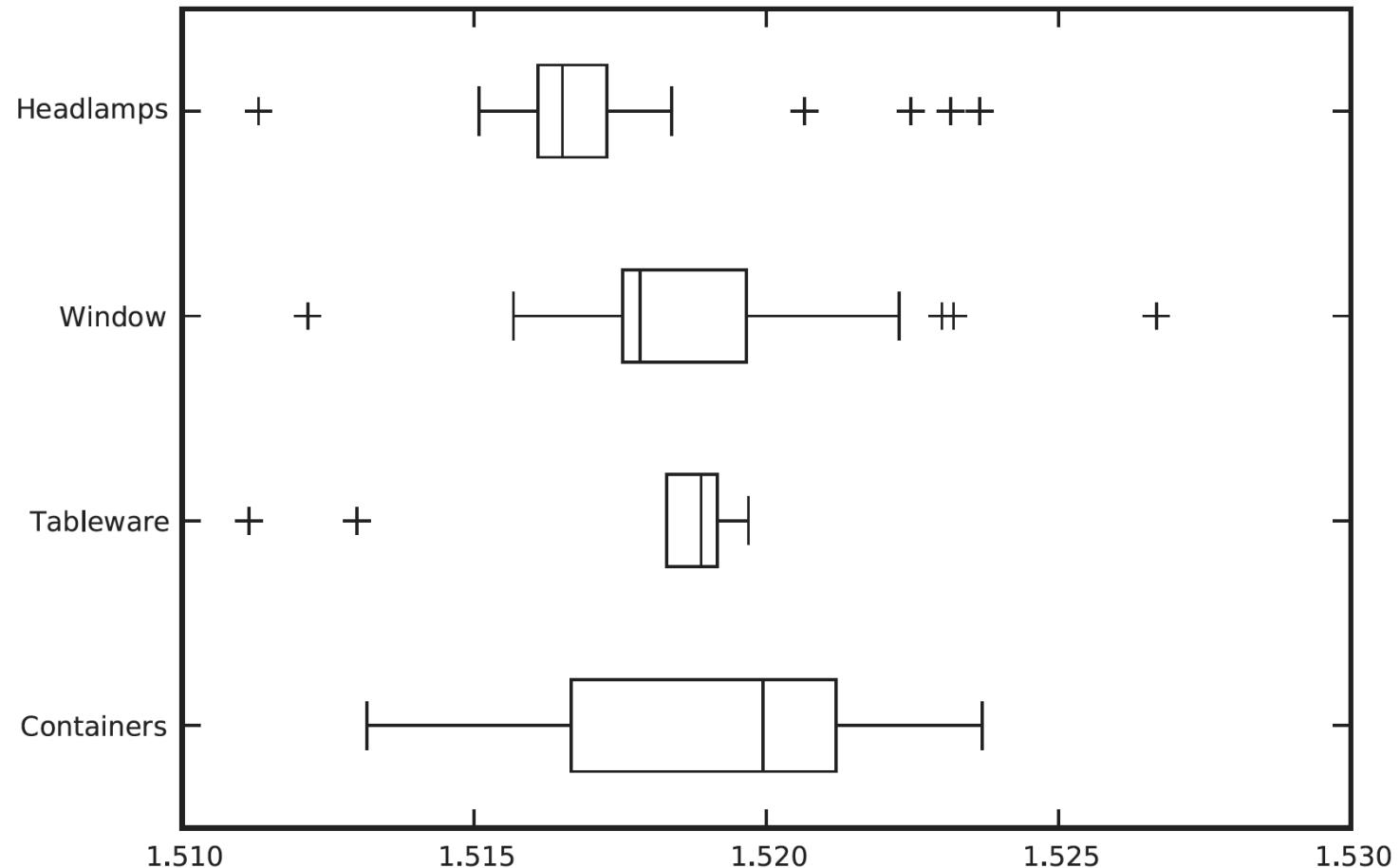
# Θηκόγραμμα (Box-and-whisker plot)



# Θηκόγραμμα (Box-and-whisker plot)

Σύνολο δεδομένων  
με 121 είδη  
γυαλιού(\*):

- Headlamps = 29
- Window = 70
- Tableware = 9
- Containers = 13

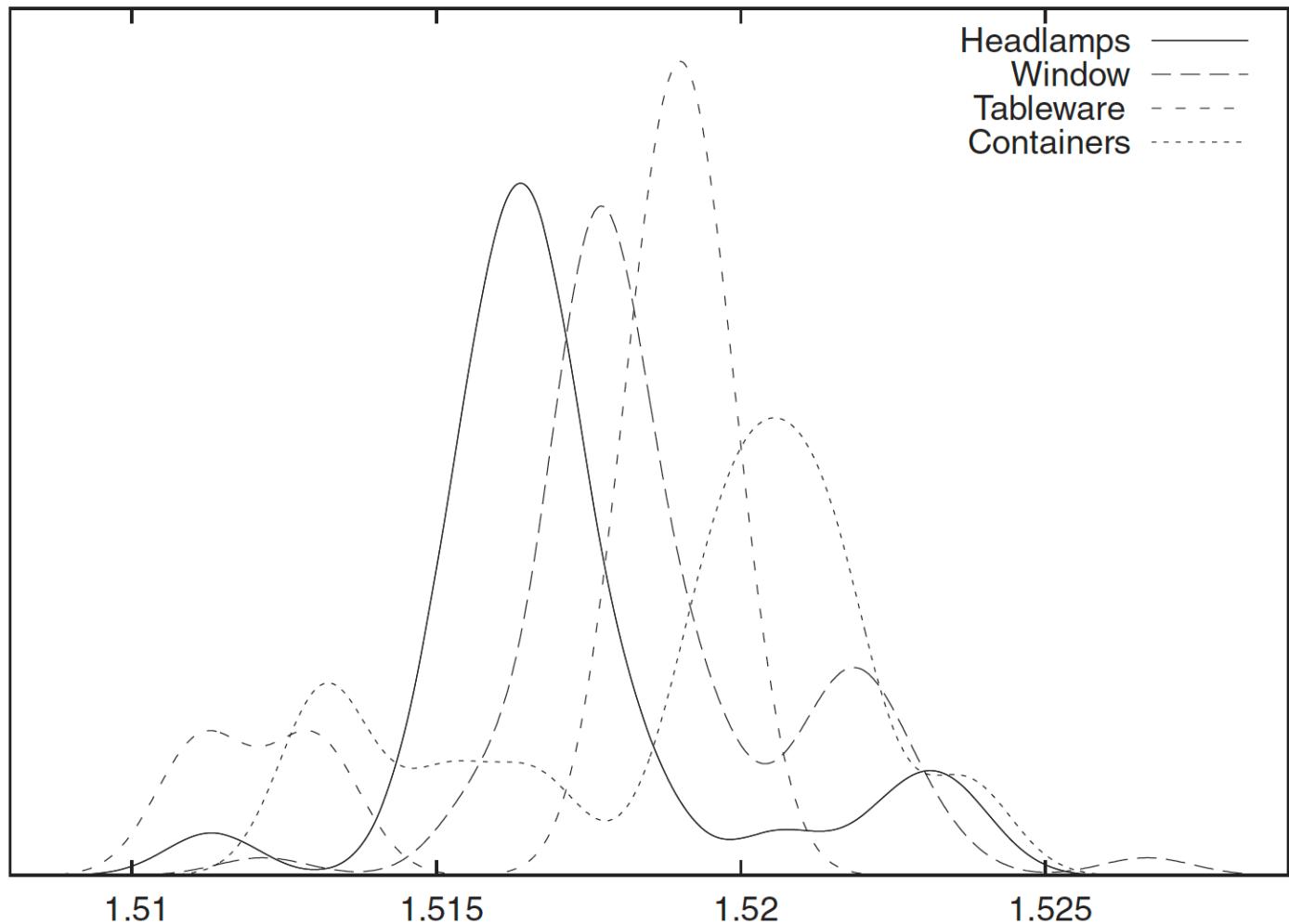


(\*) “Glass Identification Data Set” on the UCI Machine Learning Repository  
<http://archive.ics.uci.edu/ml/>

# Το ίδιο Σύνολο Δεδομένων με KDE

Σύνολο δεδομένων  
με 121 είδη  
γυαλιού(\*):

- Headlamps = 29
- Window = 70
- Tableware = 9
- Containers = 13



(\*) “Glass Identification Data Set” on the UCI Machine Learning Repository  
<http://archive.ics.uci.edu/ml/>

# Ανάλυση Δύο Μεταβλητών: Εντοπίζοντας Συσχετίσεις

—

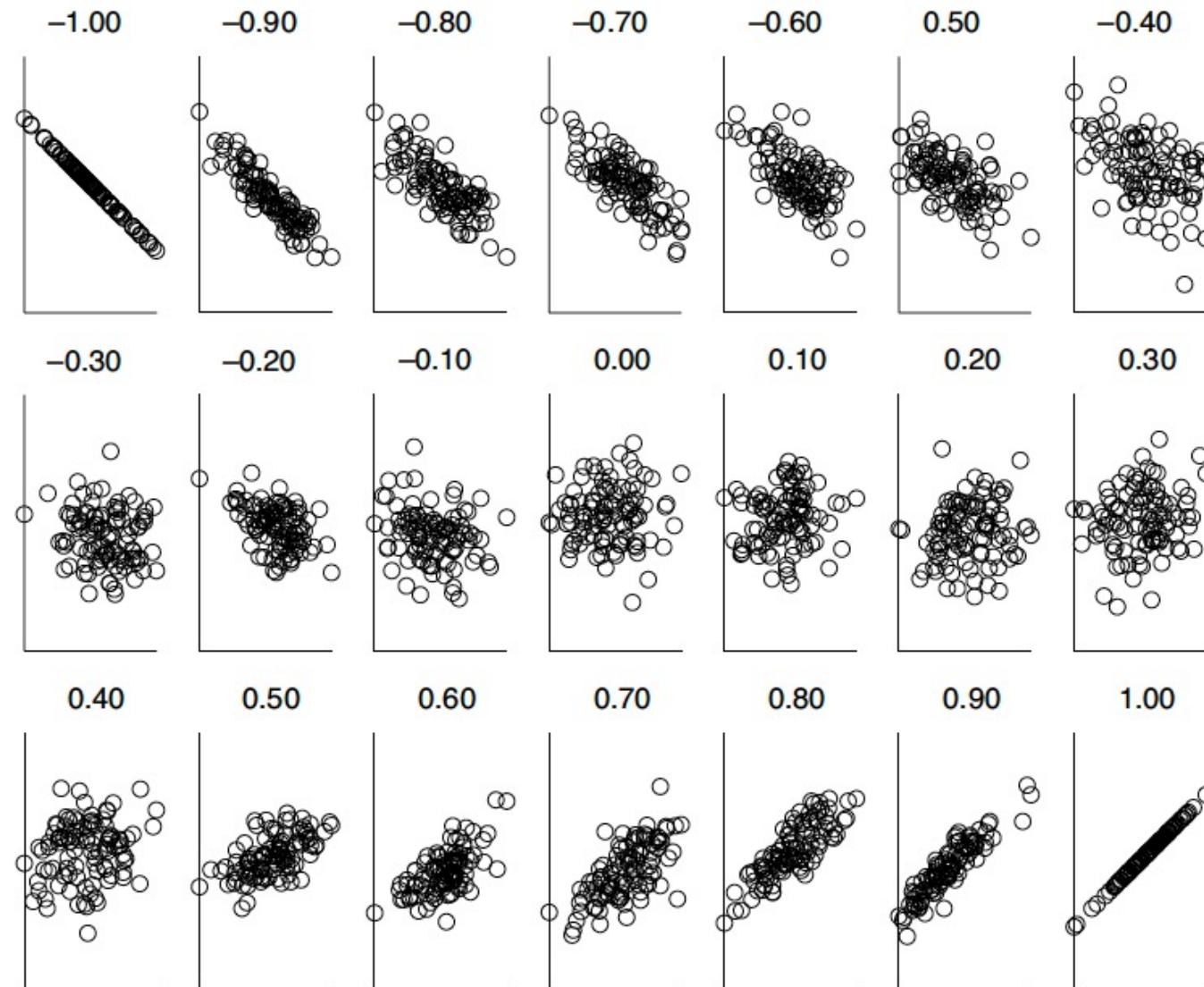
# Ορισμένα Εργαλεία για Διμεταβλητή Ανάλυση

- Χρήση δισδιάστατων διαγραμμάτων
- Συντελεστής συσχέτισης (Pearson's correlation coefficient)
- Αντιμετώπιση δεδομένων διαφορετικής κλίμακας
  - Διάγραμμα λογαριθμικής κλίμακας
- Αντιμετώπιση θορύβου στα δεδομένα
  - Εξομάλυνση (smoothing)

# Δουλεύοντας με Δύο Μεταβλητές

- Υπάρχει κάποια σχέση μεταξύ των δύο μεταβλητών;
- Εάν ναι, τι είδους σχέση;
  
- Ένας εύκολος τρόπος είναι να φτιάξουμε ένα διάγραμμα
- Ένας άλλος είναι να υπολογίσουμε το **συντελεστή συσχέτισης** (correlation coefficient) μεταξύ των δύο μεταβλητών
  
- Όμως, πώς μπορεί να αντιμετωπιστεί η περίπτωση που υπάρχει **θόρυβος** στα δεδομένα;

# Διαφορετικές Τιμές Συντελεστή Συσχέτισης για Διάφορα Σύνολα Δεδομένων



# Pearson's Correlation Coefficient

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x s_y},$$

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})$$

$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2}$$

$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2}$$

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k \text{ is the mean of } \mathbf{x}$$

$$\bar{y} = \frac{1}{n} \sum_{k=1}^n y_k \text{ is the mean of } \mathbf{y}$$

# Παραδείγματα Συντελεστή Συσχέτισης

Παράδειγμα 1 (τέλεια συσχέτιση):

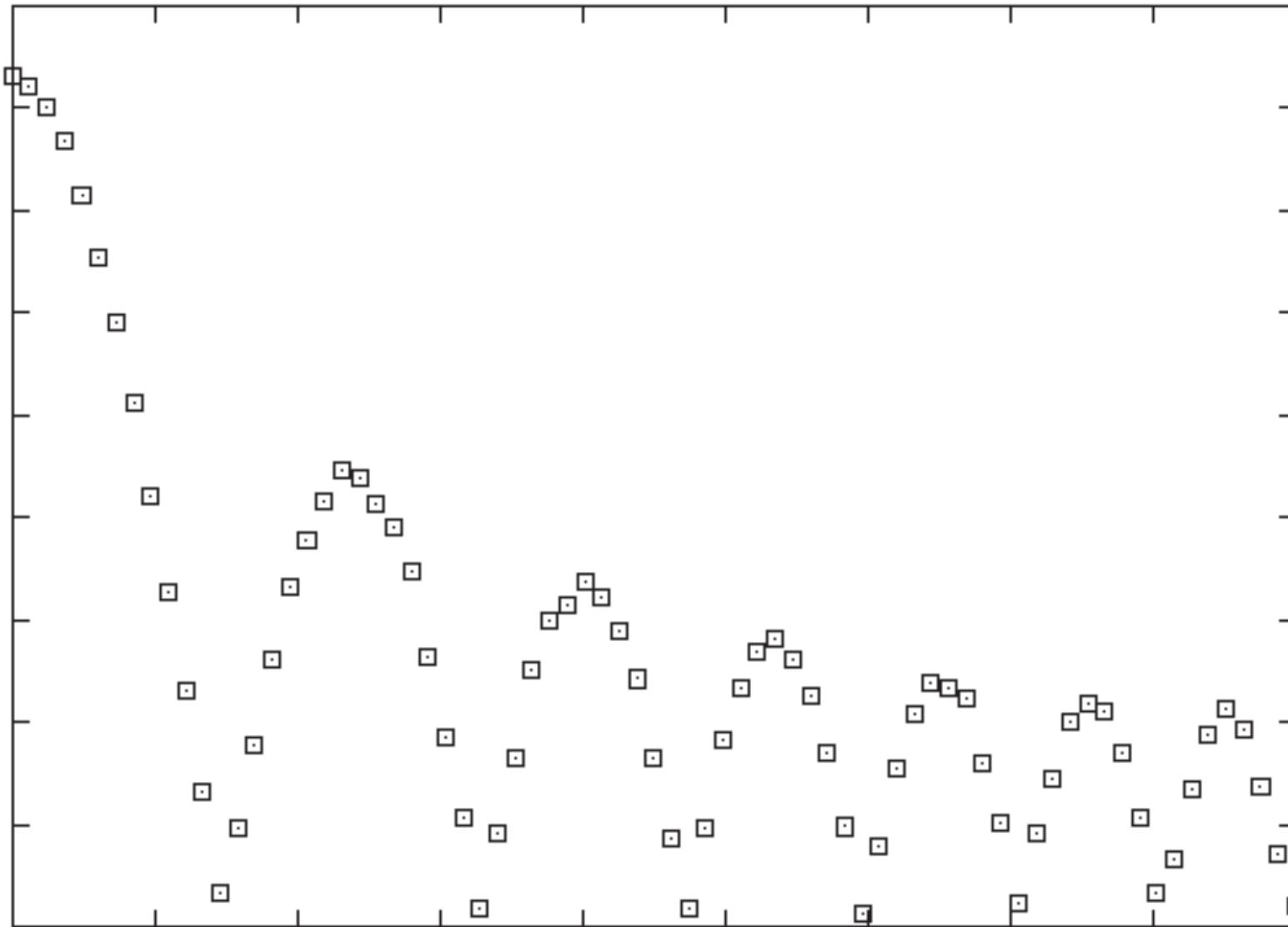
$$\begin{aligned} \mathbf{x} &= (-3, 6, 0, 3, -6) & \text{correlation} &= -1 \\ \mathbf{y} &= (-1, -2, 0, -1, 2) \end{aligned}$$

$$\begin{aligned} \mathbf{x} &= (3, 6, 0, 3, 6) & \text{correlation} &= 1 \\ \mathbf{y} &= (1, 2, 0, 1, 2) \end{aligned}$$

Παράδειγμα 2 (μη γραμμική συσχέτιση):

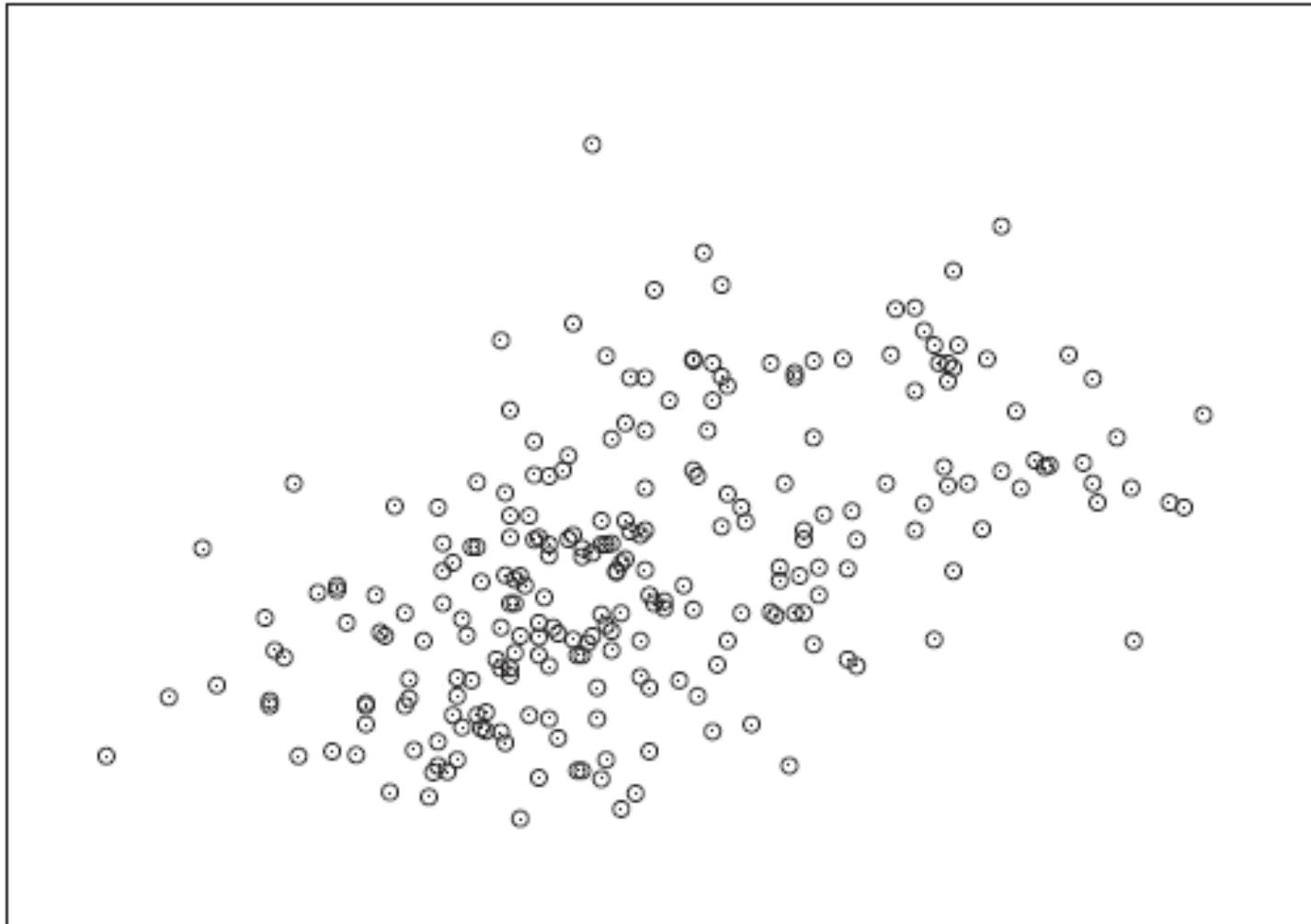
$$\begin{aligned} \mathbf{x} &= (-3, -2, -1, 0, 1, 2, 3) & \text{correlation} &= 0 \text{ (όμως } y = x^2) \\ \mathbf{y} &= (9, 4, 1, 0, 1, 4, 9) \end{aligned}$$

# Παράδειγμα



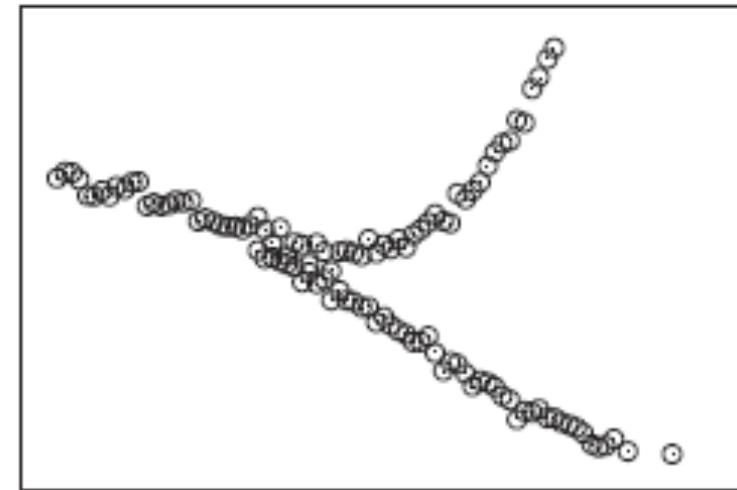
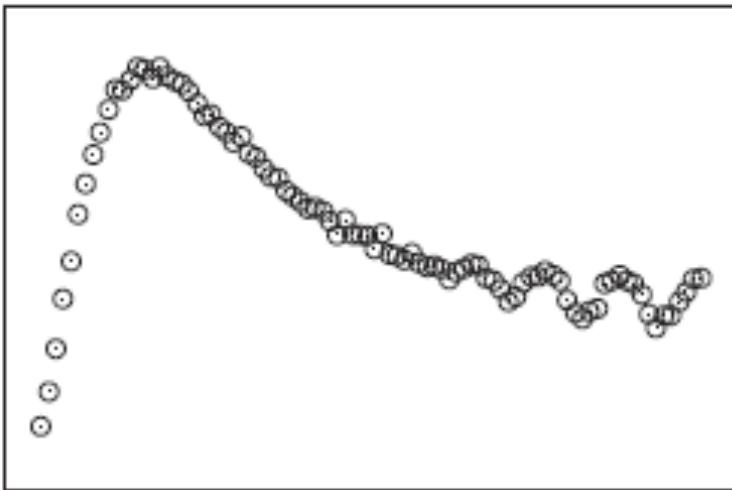
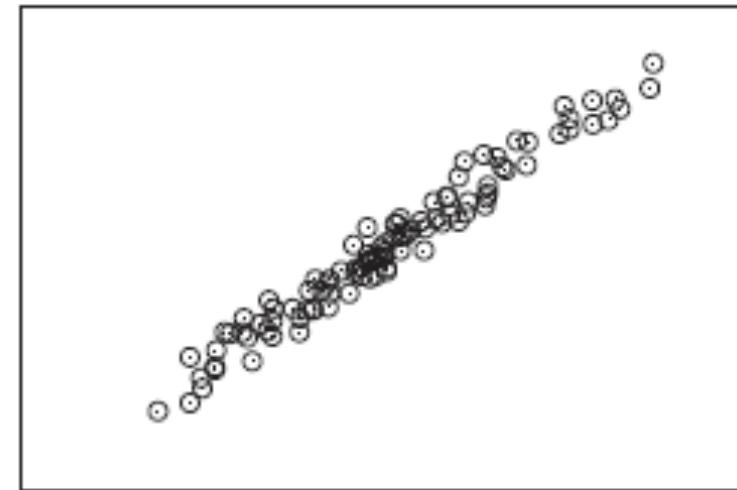
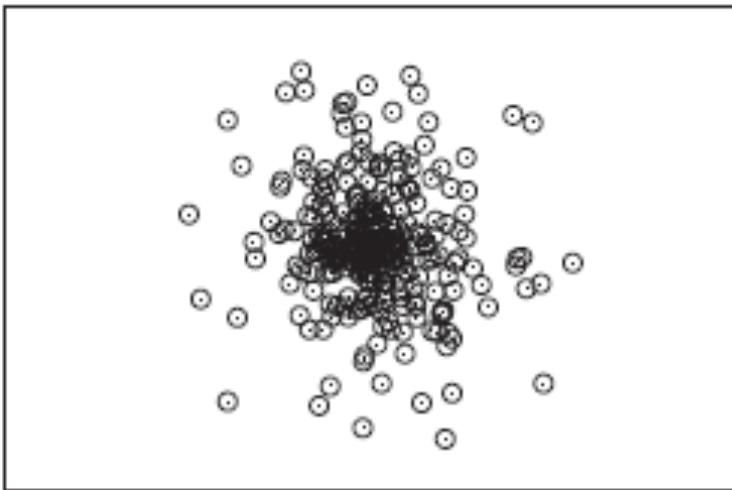
Εμφανής η ύπαρξη κάποιας (πολύπλοκης) συσχέτισης μεταξύ X και Y

# Παράδειγμα: Διάγραμμα Διασποράς (Scatter plot ή xy plot)



Ένα σύνολο δεδομένων με θόρυβο... υπάρχει κάποια συσχέτιση μεταξύ των  $X$  και  $Y$ ;

# Τι Μπορεί να Συναντήσουμε

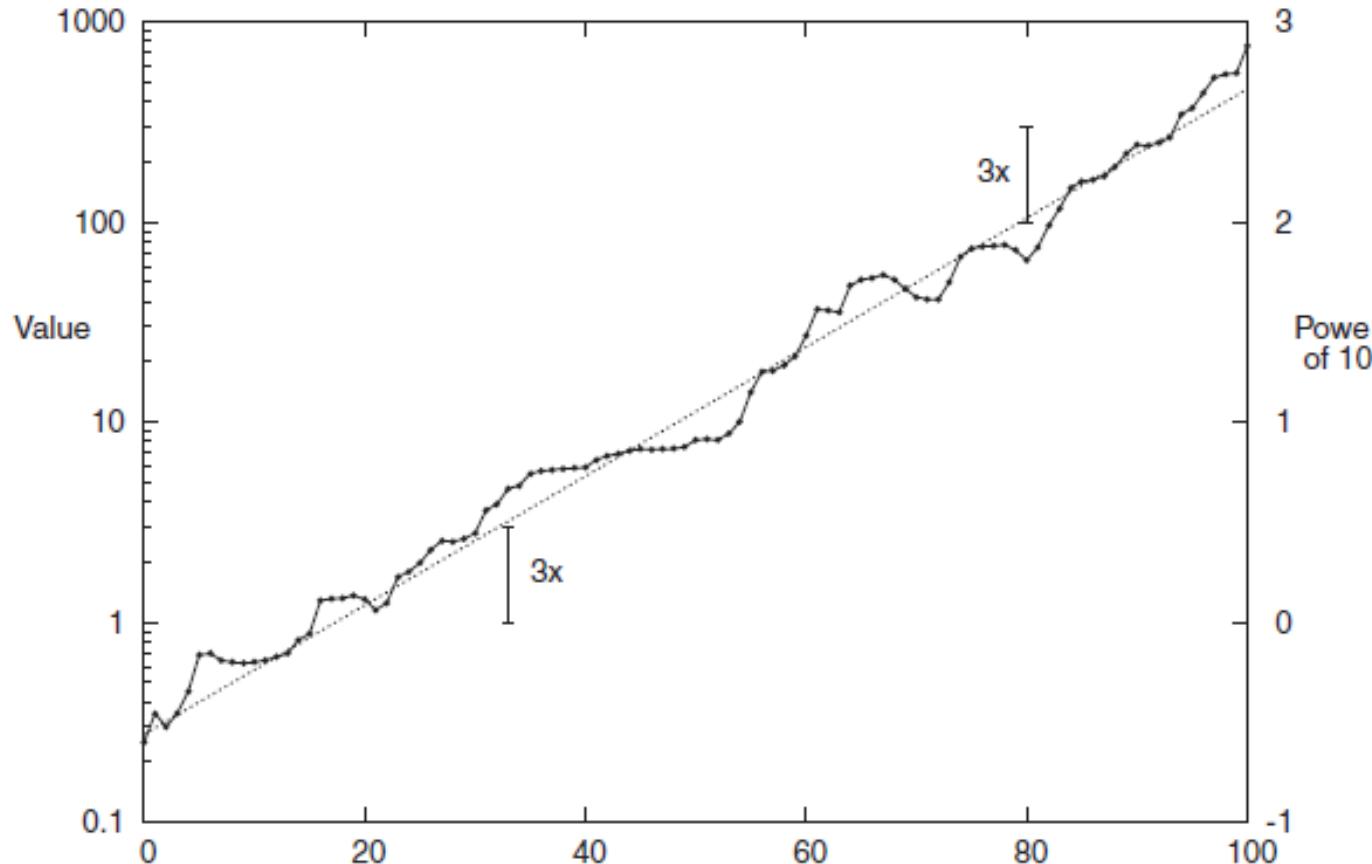


Τέσσερις διαφορετικοί τύποι συσχετίσεων (από αριστερά προς δεξιά, πάνω προς κάτω):  
(1) καμία συσχέτιση, (2) ισχυρή, απλή συσχέτιση, (3) ισχυρή, πολύπλοκη συσχέτιση,  
(4) πολυμεταβλητή συσχέτιση

# Για τον Εντοπισμό Δομής σε ένα Σύνολο Δεδομένων

- Λογαριθμική κλίμακα (**logarithmic plots**)
  - Βοηθούν όταν τα δεδομένα καταλαμβάνουν αρκετές τάξεις μεγέθους
- Μέθοδοι εξομάλυνσης (**smoothing methods**)
  - Μειώνουν το θόρυβο στα δεδομένα

# Λογαριθμική Κλίμακα (Logarithmic plots)



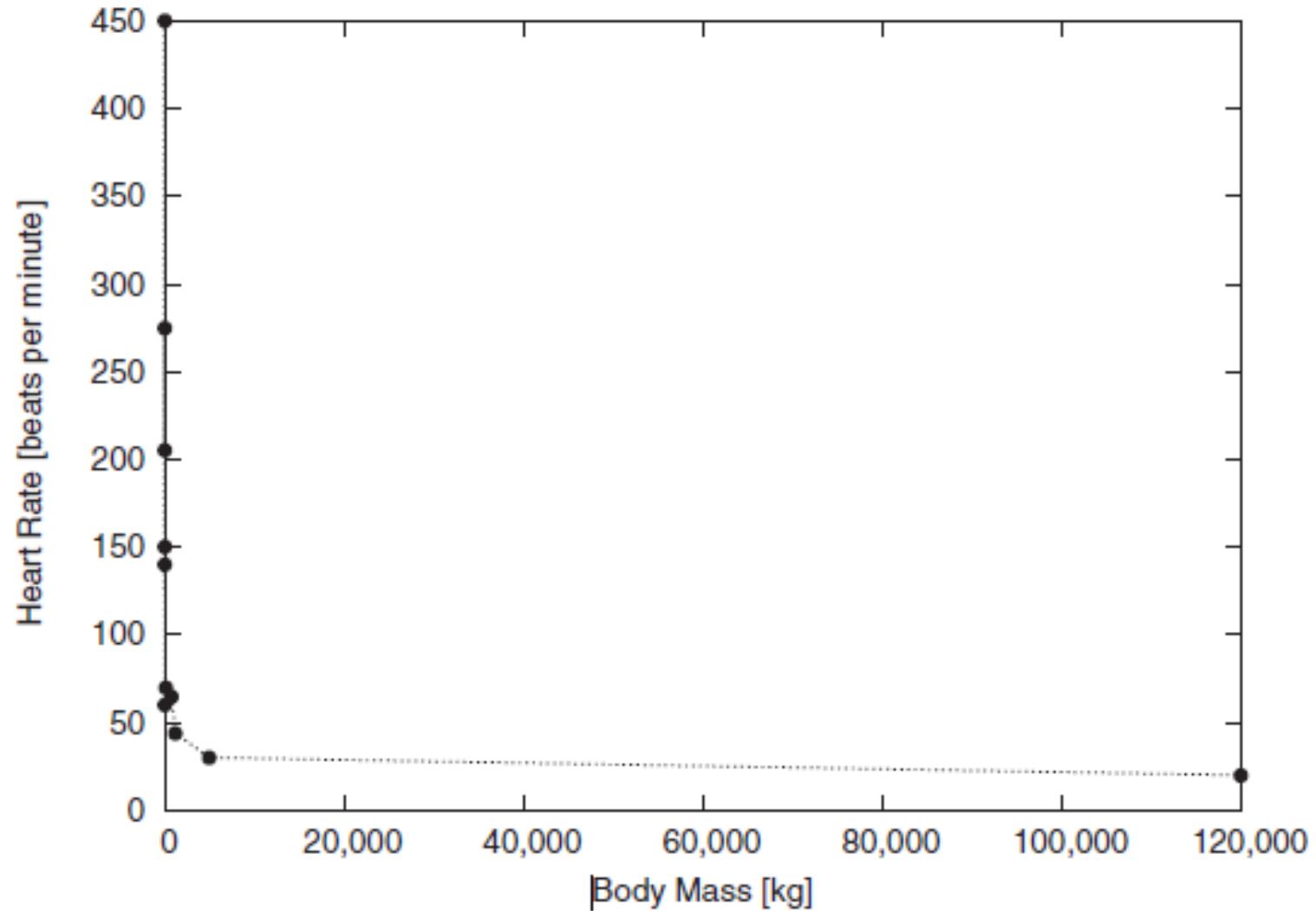
$$y = C \exp(\alpha x) \quad \text{where } C \text{ and } \alpha \text{ are constants}$$

$$\log y = \alpha x + \log C$$

Βασικά πλεονεκτήματα διαγραμμάτων λογαριθμικής κλίμακας:

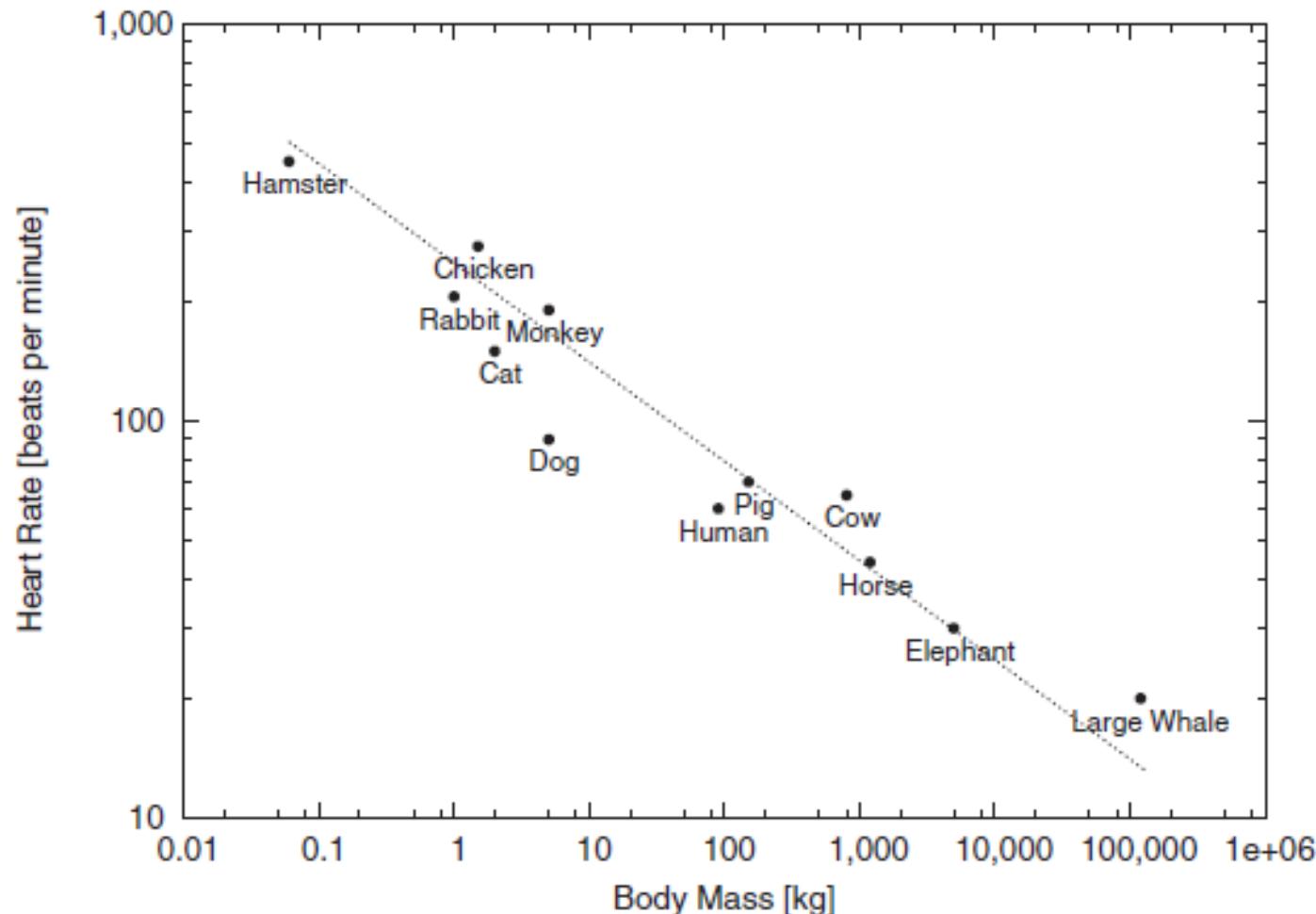
- (1) Συγκρατούν μεγάλες διακυμάνσεις στις τιμές δεδομένων
- (2) Μετατρέπουν πολλαπλασιαστικές διακυμάνσεις σε προσθετικές
- (3) Αποκαλύπτουν εκθετική συμπεριφορά

# Παράδειγμα



Χτύποι καρδιάς (Υ-άξονας) vs. Βάρος (Χ-άξονας) για θηλαστικά

## Ίδιο Παράδειγμα, σε (Διπλή) Λογαριθμική Κλίμακα



Τα δεδομένα φαίνεται ότι βρίσκονται πάνω σε μια ευθεία, γεγονός που υποδεικνύει ότι ισχύει ένας **εκθετικός νόμος (power law)** μεταξύ χτύπων καρδιάς και βάρους (βλ. Allometric scaling)

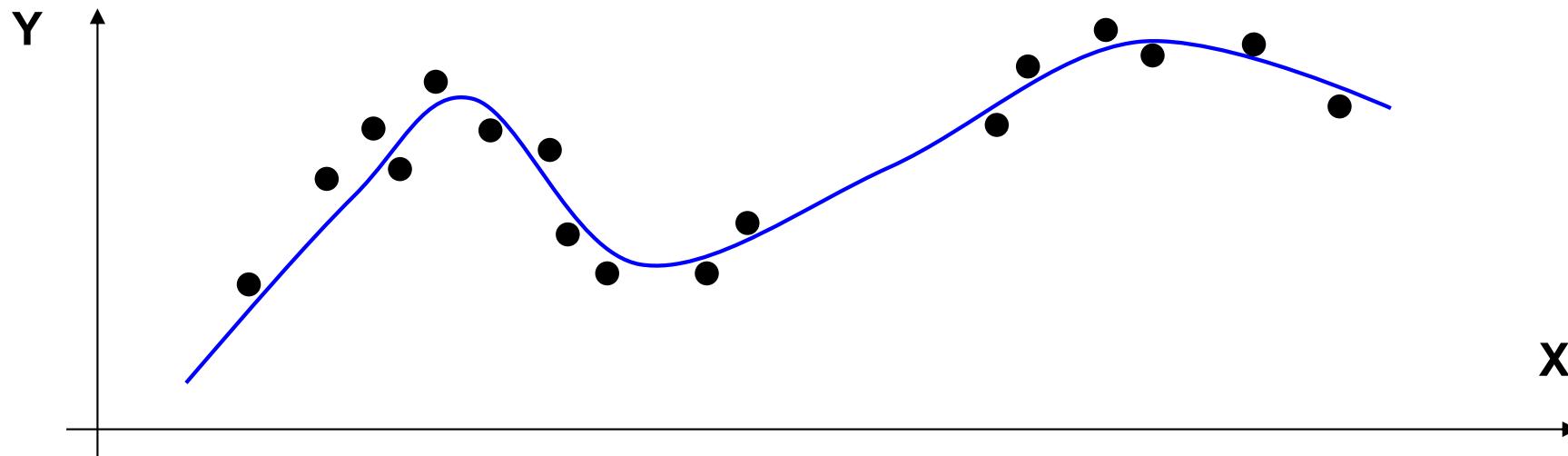
# Αντιμετώπιση Θορύβου: Εξομάλυνση (Smoothing)

- Συχνά, είναι χρήσιμο να βρούμε μια **ομαλή καμπύλη** που **αναπαριστά το σύνολο δεδομένων** (που περιέχει θόρυβο)
- Κάποια **τάση** ή **δομή** στα δεδομένα μπορεί να γίνει εμφανής ευκολότερα από μια τέτοια καμπύλη, παρά από ένα σύνολο σημείων
- Μία μέθοδος που χρησιμοποιείται συχνά για την εξαγωγή μιας **ομαλής αναπαράστασης** για σύνολα δεδομένων **που περιέχουν θόρυβο**
  - **LOESS**: LOcally weighted EStimated Smoothing ή
  - **LOWESS**: LOcally WEighted Scatterplot Smoothing

# Η Μέθοδος LOWESS

Δοθέντος ενός συνόλου δεδομένων  $(x_i, y_i)$

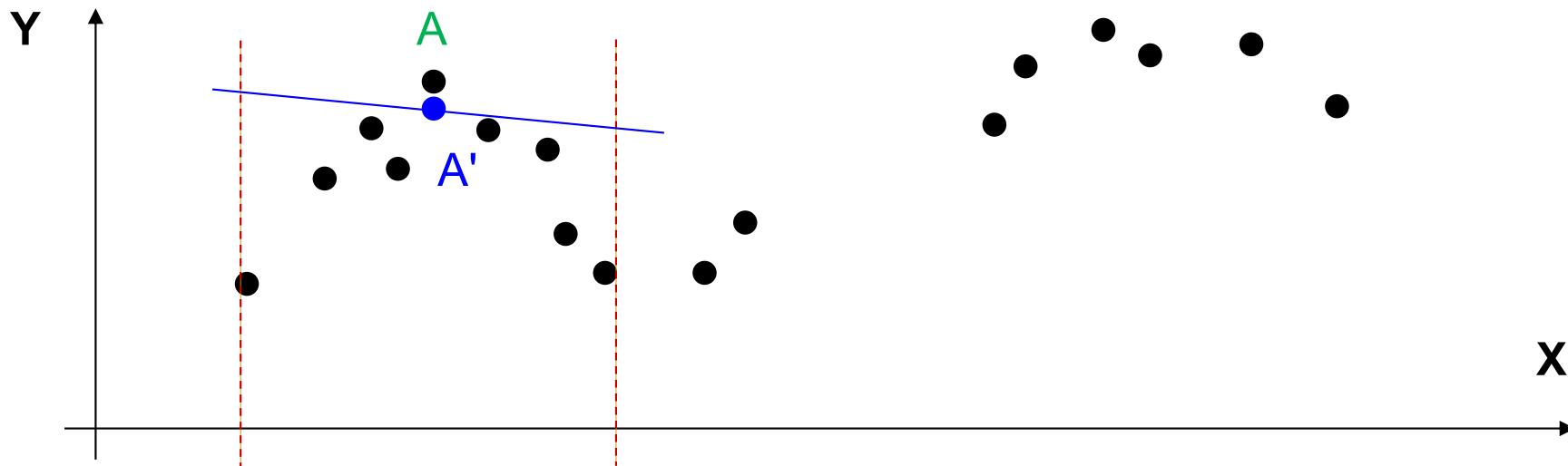
Ζητείται να βρεθεί μια καμπύλη που «προσεγγίζει» τα δεδομένα



Η LOWESS προσπαθεί να **προσεγγίσει τοπικά τα δεδομένα**, με  
ένα πολυώνυμο χαμηλού βαθμού (συνήθως πρώτου βαθμού)

# Η Μέθοδος LOWESS – Λειτουργία

- (1) Επιλέγεται μια «**γειτονιά**» από σημεία (μέγεθος γειτονιάς = *bandwidth*)
- (2) Για κάθε σημείο της «**γειτονιάς**», π.χ. για το **σημείο A**



- (3) Υπολογίζεται η **ευθεία** που ταιριάζει καλύτερα στα δεδομένα μέσω **τοπικά ζυγισμένης παλινδρόμησης** (*locally weighted regression*)
- (4) Υπολογίζεται το **σημείο A'** στη θέση που βρίσκεται το A

# Τοπικά Ζυγισμένη Παλινδρόμηση

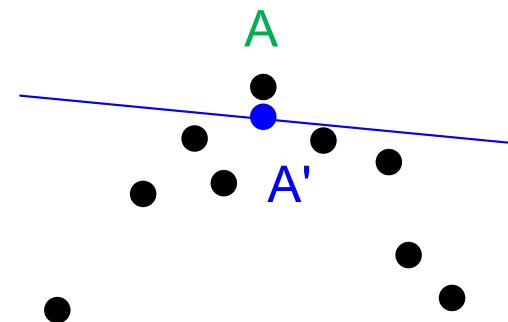
Υποθέτουμε ένα απλό γραμμικό μοντέλο:

$$y = \alpha * x + \beta$$

**Μέθοδος ελαχίστων τετραγώνων:**

Να βρεθεί εκείνη η ευθεία (ισοδύναμα τα  $\alpha$  και  $\beta$ ),  
έτσι ώστε να ελαχιστοποιηθεί το συνολικό  
**τετραγωνικό σφάλμα:**

$$\sum w(x - x_i) * (\alpha * x_i + \beta - y_i)^2$$



Συνήθως στη LOWESS χρησιμοποιείται η συνάρτηση:

$$w(x) = \begin{cases} (1 - |x|^3)^3 & \text{for } |x| < 1 \\ 0 & \text{for } |x| \geq 1 \end{cases}$$

Τα κοντινά σημεία συνεισφέρουν περισσότερο από τα μακρινά (local weighting)

# Τοπικά Ζυγισμένη Παλινδρόμηση

Μία γειτονιά σημείων (bandwidth)

	X	Y	απόσταση	Κανον. απόσταση	Βάρος w()
$x_1$	$x_1$	$y_1$	$x_1 - x_1$	$(x_1 - x_1) / d_{max} = 0$	$w((x_1 - x_1) / d_{max})$
	$x_2$	$y_2$	$x_2 - x_1$	$(x_2 - x_1) / d_{max}$	$w((x_2 - x_1) / d_{max})$
	$x_3$	$y_3$	$x_3 - x_1$	$(x_3 - x_1) / d_{max}$	$w((x_3 - x_1) / d_{max})$
	$x_4$	$y_4$	$x_4 - x_1$	$(x_4 - x_1) / d_{max}$	$w((x_4 - x_1) / d_{max})$
	$x_5$	$y_5$	$x_5 - x_1$	$(x_5 - x_1) / d_{max}$	$w((x_5 - x_1) / d_{max})$
	$x_6$	$y_6$	$x_6 - x_1$	$(x_6 - x_1) / d_{max} = 1$	$w((x_6 - x_1) / d_{max})$

$$d_{max} = x_6 - x_1$$

υπολογισμός:  $\alpha, \beta$

# Τοπικά Ζυγισμένη Παλινδρόμηση

Μία γειτονιά σημείων (bandwidth)

	X	Y	απόσταση	Κανον. απόσταση	Βάρος w()
x <sub>2</sub>	x <sub>1</sub>	y <sub>1</sub>	x <sub>1</sub> - x <sub>1</sub>	(x <sub>1</sub> - x <sub>2</sub> ) / d <sub>max</sub> = 0	w((x <sub>1</sub> - x <sub>2</sub> ) / d <sub>max</sub> )
	x <sub>2</sub>	y <sub>2</sub>	x <sub>2</sub> - x <sub>1</sub>	(x <sub>2</sub> - x <sub>2</sub> ) / d <sub>max</sub>	w((x <sub>2</sub> - x <sub>2</sub> ) / d <sub>max</sub> )
	x <sub>3</sub>	y <sub>3</sub>	x <sub>3</sub> - x <sub>1</sub>	(x <sub>3</sub> - x <sub>2</sub> ) / d <sub>max</sub>	w((x <sub>3</sub> - x <sub>2</sub> ) / d <sub>max</sub> )
	x <sub>4</sub>	y <sub>4</sub>	x <sub>4</sub> - x <sub>1</sub>	(x <sub>4</sub> - x <sub>2</sub> ) / d <sub>max</sub>	w((x <sub>4</sub> - x <sub>2</sub> ) / d <sub>max</sub> )
	x <sub>5</sub>	y <sub>5</sub>	x <sub>5</sub> - x <sub>1</sub>	(x <sub>5</sub> - x <sub>2</sub> ) / d <sub>max</sub>	w((x <sub>5</sub> - x <sub>2</sub> ) / d <sub>max</sub> )
	x <sub>6</sub>	y <sub>6</sub>	x <sub>6</sub> - x <sub>1</sub>	(x <sub>6</sub> - x <sub>2</sub> ) / d <sub>max</sub> = 1	w((x <sub>6</sub> - x <sub>2</sub> ) / d <sub>max</sub> )

$$d_{\max} = x_6 - x_1$$

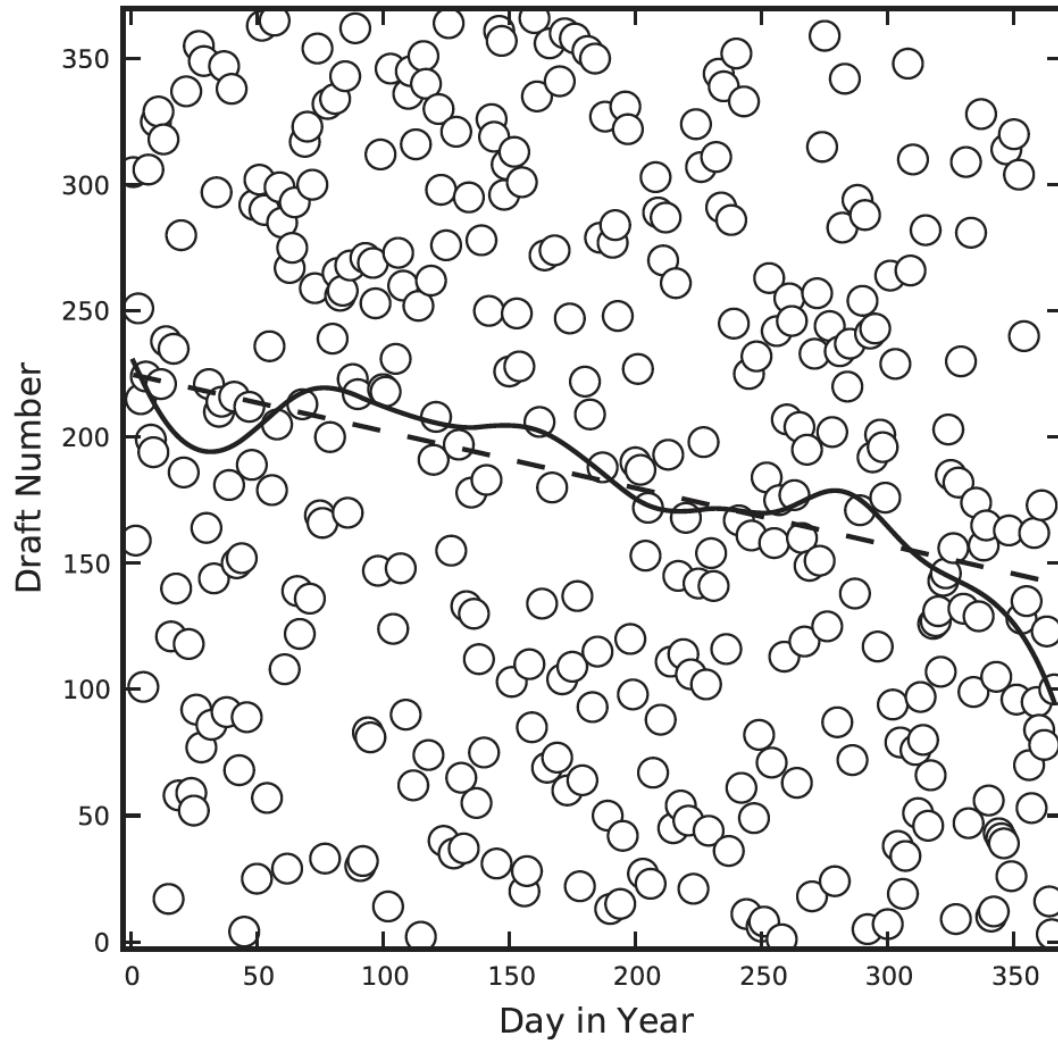
**Εκ νέου** υπολογισμός: α, β

Άσκηση: Προσπαθήστε να υλοποιήσετε μόνοι σας τη μέθοδο LOWESS

# Παραδείγματα Εξομάλυνσης

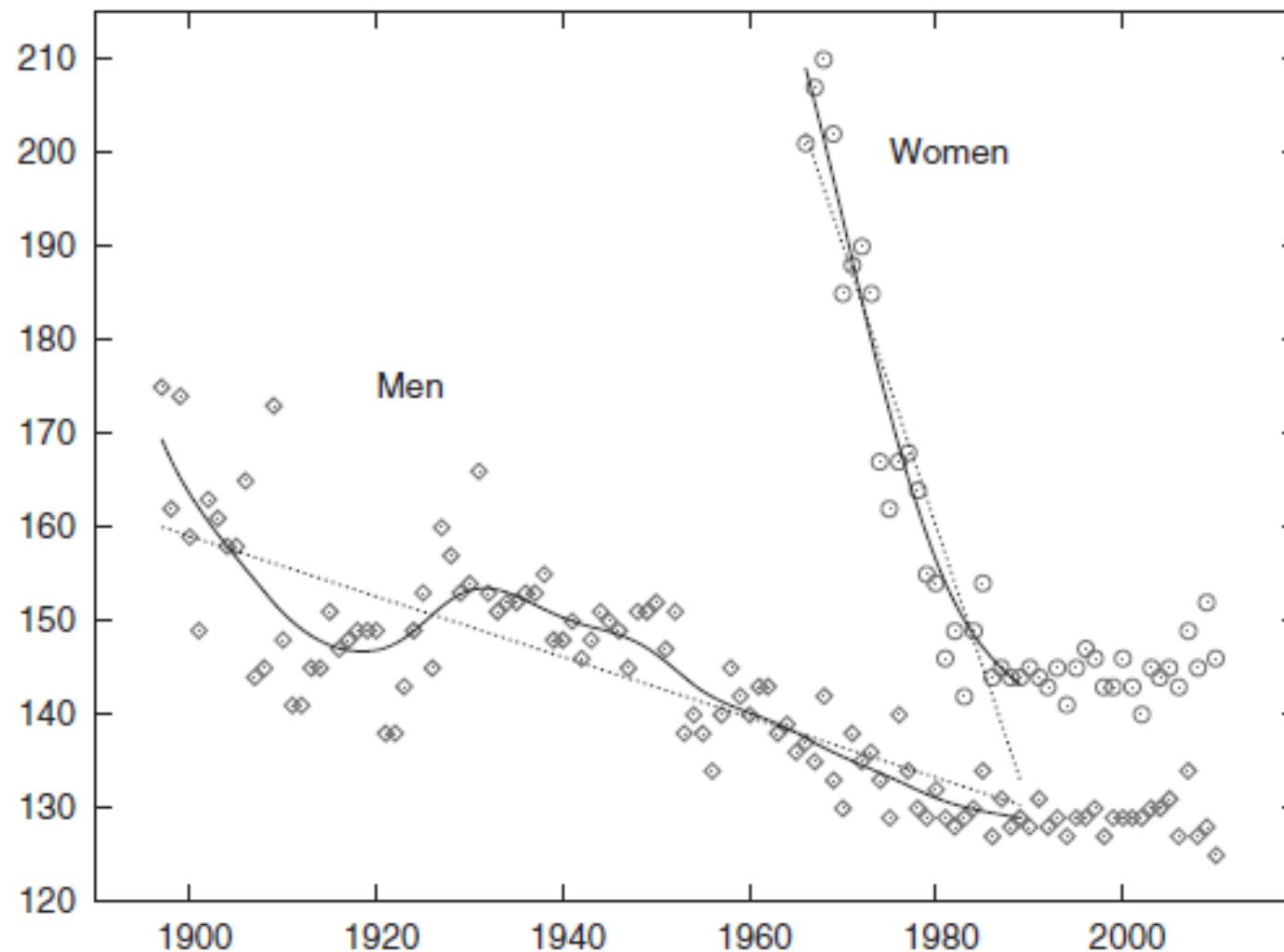
- Ακολουθούν δύο παραδείγματα
- Δείχνουν τη χρησιμότητα της εξομάλυνσης (smoothing)
- Για την ανακάλυψη κάποιου μοτίβου στα δεδομένα, που δεν είναι εμφανές

# Παράδειγμα 1



Κλήρωση 1970 (ΗΠΑ): αριθμός κλήρωσης vs. ημερομηνία γέννησης. Δύο καμπύλες LOWESS με διαφορετική τιμή για την παράμετρο *bandwidth* δείχνουν ότι όσοι άντρες ήταν γεννημένοι αργότερα μες στο έτος είχαν μικρότερους αριθμούς κλήρωσης.

## Παράδειγμα 2

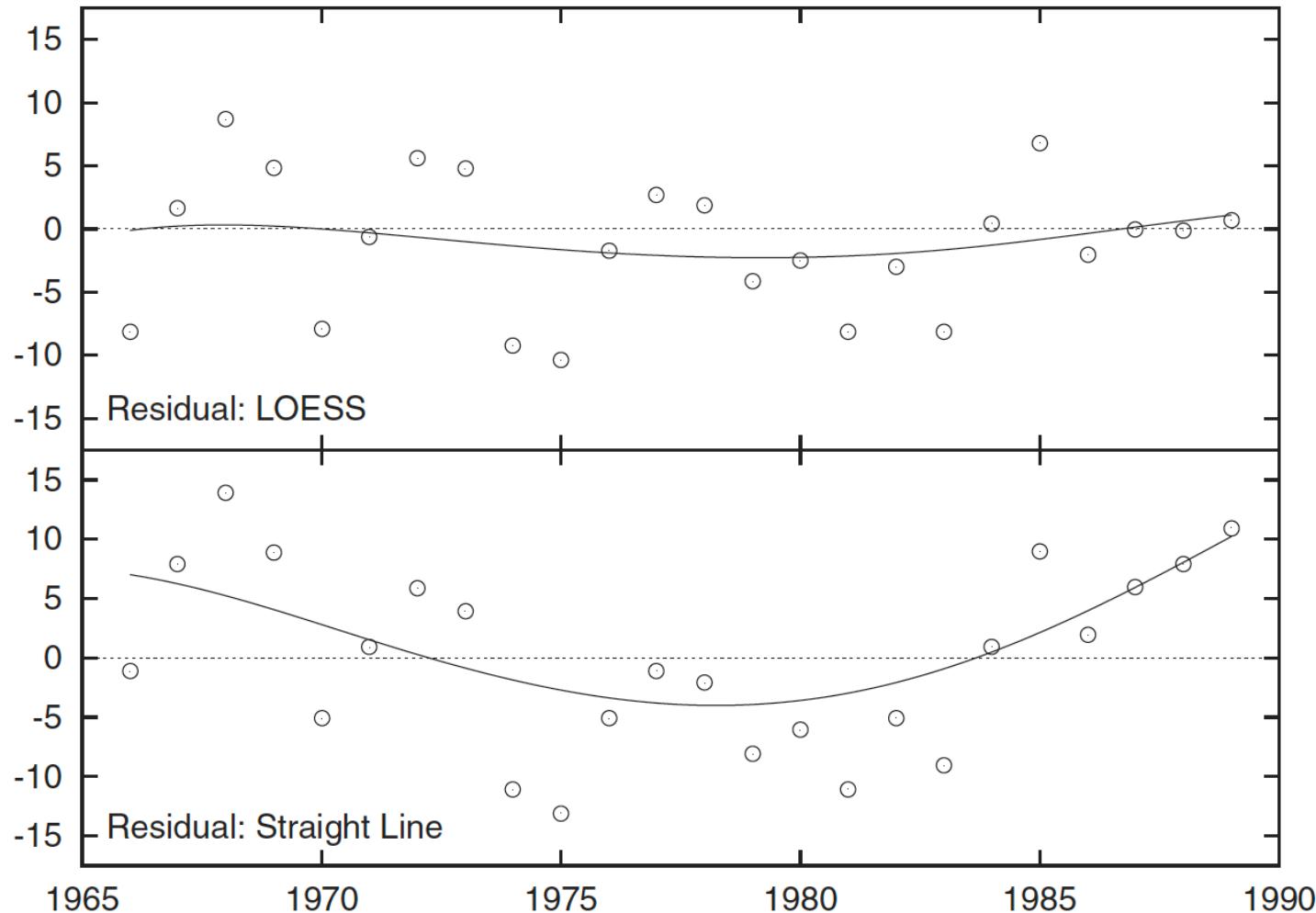


Χρόνοι τερματισμού (σε λεπτά) για ετήσιο μαραθώνιο, χωριστά για άντρες και γυναίκες. Επίσης, φαίνονται οι προσεγγίσεις με ευθεία γραμμή και με ομαλή καμπύλη. Οι προσεγγίσεις βασίζονται στα δεδομένα μέχρι το 1990 μόνο.

# Υπόλοιπα (Residuals)

- Αφού έχουμε μια εξομαλυμένη προσέγγιση των δεδομένων, ελέγχουμε το **υπόλοιπο (residual)** που είναι αυτό που μένει όταν αφαιρέσουμε την προσέγγιση από τα δεδομένα
  - Τα υπόλοιπα πρέπει να είναι **συμμετρικά κατανεμημένα** γύρω από το μηδέν
  - Τα υπόλοιπα **δεν πρέπει να περιέχουν κάποια τάση**
    - Εάν υπάρχει κάποια τάση ή άλλη συστηματική συμπεριφορά, υποδεικνύει ότι το μοντέλο δεν είναι κατάλληλο
  - Τα υπόλοιπα **παίρνουν και θετικές και αρνητικές τιμές**, και θα περνούν από το μηδέν

# Υπόλοιπα (Residuals)



Χρόνοι τερματισμού (σε λεπτά) για ετήσιο μαραθώνιο, χωριστά για άντρες και γυναίκες. Επίσης, φαίνονται οι προσεγγίσεις με ευθεία γραμμή και με ομαλή καμπύλη. Οι προσεγγίσεις βασίζονται στα δεδομένα μέχρι το 1990 μόνο.

# Εμφάνιση/Παρουσίαση των Σημαντικών Στοιχείων

## ■ #1: Βασικά στοιχεία

- Επιλογή κατάλληλων περιοχών τιμών
- Αφαίρεση κάποιας σταθερής ποσότητας
- Χρήση συμβόλων (για σημεία) ή γραμμών (για συνεχή δεδομένα) ή και των δύο

## ■ #2: Εμφάνιση

- Λογαριθμική κλίμακα
- Προσθήκη μιας ομαλής καμπύλης

# Εμφάνιση/Παρουσίαση των Σημαντικών Στοιχείων

- #3: **Κατασκευή ενός μαθηματικού μοντέλου (και εντοπισμός διαφορών με τα δεδομένα)**
  - Αφαίρεση κάποιας τάσης
  - Σχηματίζω την αναλογία σε σχέση με μια τιμή αναφοράς
  - Αλλαγή κλίμακας ώστε καμπύλες να έρθουν η μία πάνω στην άλλη
- #4: **Προσθήκη διακοσμητικών στοιχείων (για γραφήματα μόνο)**
  - Ετικέτες, βέλη, επισημειώσεις, ειδικά σύμβολα, επεξηγήσεις μπορούν να αυξήσουν την πληροφορία σε ένα γράφημα και να γίνει έτσι αυτο-περιγραφικό
  - Σημαντικό για την παρουσίαση των αποτελεσμάτων ανάλυσης σε τρίτους

# Τι Πρέπει να Προσέχουμε ηλιτά την Παρουσίαση Αποτελεσμάτων

- Οι λέξεις που χρησιμοποιούνται στο γράφημα να είναι αυτο-περιγραφικές
  - Δε βασιζόμαστε στη λεζάντα, όλη η πληροφορία πρέπει να βρίσκεται στο γράφημα
- Πρέπει να υπάρχει επεξήγηση των αξόνων: όνομα, μονάδες μέτρησης(!)
- Οι ετικέτες να είναι αυτο-περιγραφικές
- Κατάλληλη γραμματοσειρά για τις λέξεις
  - Μέγεθος 10-12pt
  - Sans-serif fonts (π.χ. Helvetica) για μικρές ετικέτες
  - Serif fonts (π.χ. Times) για κείμενο
- Αν υπάρχουν error bars, να εξηγείται τι δείχνουν
  - Τυπική απόκλιση; Inter-quartile range; ...
- Επιλογή κατάλληλων περιοχών τιμών

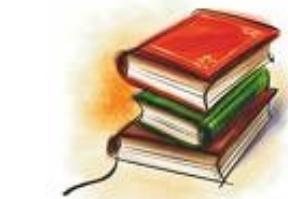
# Περίληψη – Βασικά Στοιχεία

- Διερεύνηση μονομεταβλητού συνόλου δεδομένων (πώς;)
  - Συνοπτική στατιστική
  - Οπτικοποίηση
    - Ιστογράμματα
    - Εκτιμητές πυρήνα (kernel density estimates)
    - Συνάρτηση αθροιστικής κατανομής
    - Θηκογράμματα
- Εντοπισμός δομής σε σύνολο δεδομένων δύο μεταβλητών (πώς;)
  - Διάγραμμα διασποράς
  - Συντελεστής συσχέτισης Pearson
  - Εξομάλυνση (smoothing methods: LOWESS)
  - Λογαριθμική κλίμακα

# Πηγές Αναφοράς

- *Data Analysis with Open Source Tools*, by Philipp K. Janert.

- 2011 Philipp K. Janert, 978-0-596-80235-6.
  - Κεφάλαια 2 και 3



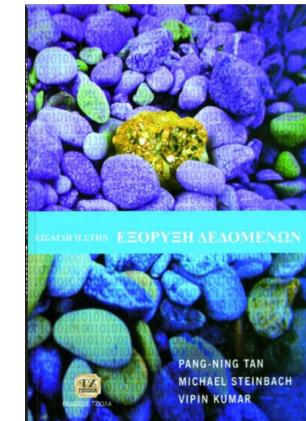
A Hands-On Guide for Programmers and Data Scientists



O'REILLY®

Philipp K. Janert

- Εισαγωγή στην Εξόρυξη Δεδομένων
  - Pang-Ning Tan, Michael Steinbach, Vipin Kumar
    - Εκδόσεις Τζιόλα
  - Κεφάλαιο 3.2 και Παράρτημα Δ





## 3. Ανάλυση Χρονοσειρών



---

Ανάλυση Δεδομένων  
(*Data Analytics*)

Χρήστος Δουλκερίδης  
2024-25

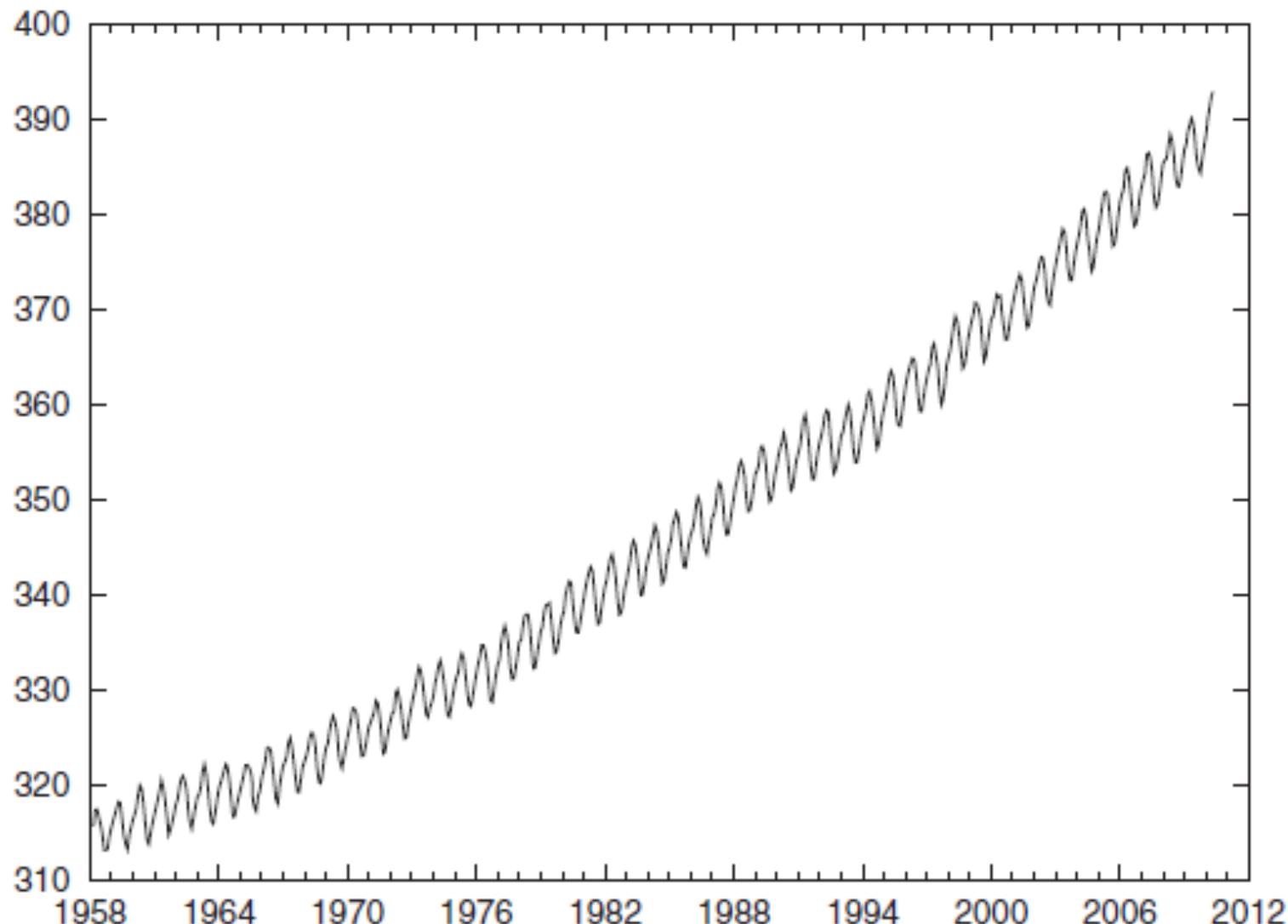
# Περιεχόμενο Διάλεξης

- **Παραδείγματα χρονοσειρών**
- Ανάλυση χρονοσειρών – βασικές έννοιες
- Εντοπισμός τάσης και εποχικότητας
- Εξομάλυνση: κινητοί μέσοι όροι
- Εξομάλυνση: εκθετική εξομάλυνση
- Συνάρτηση αυτοσυσχέτισης

# Διάγραμμα Χρόνου – Time Plot

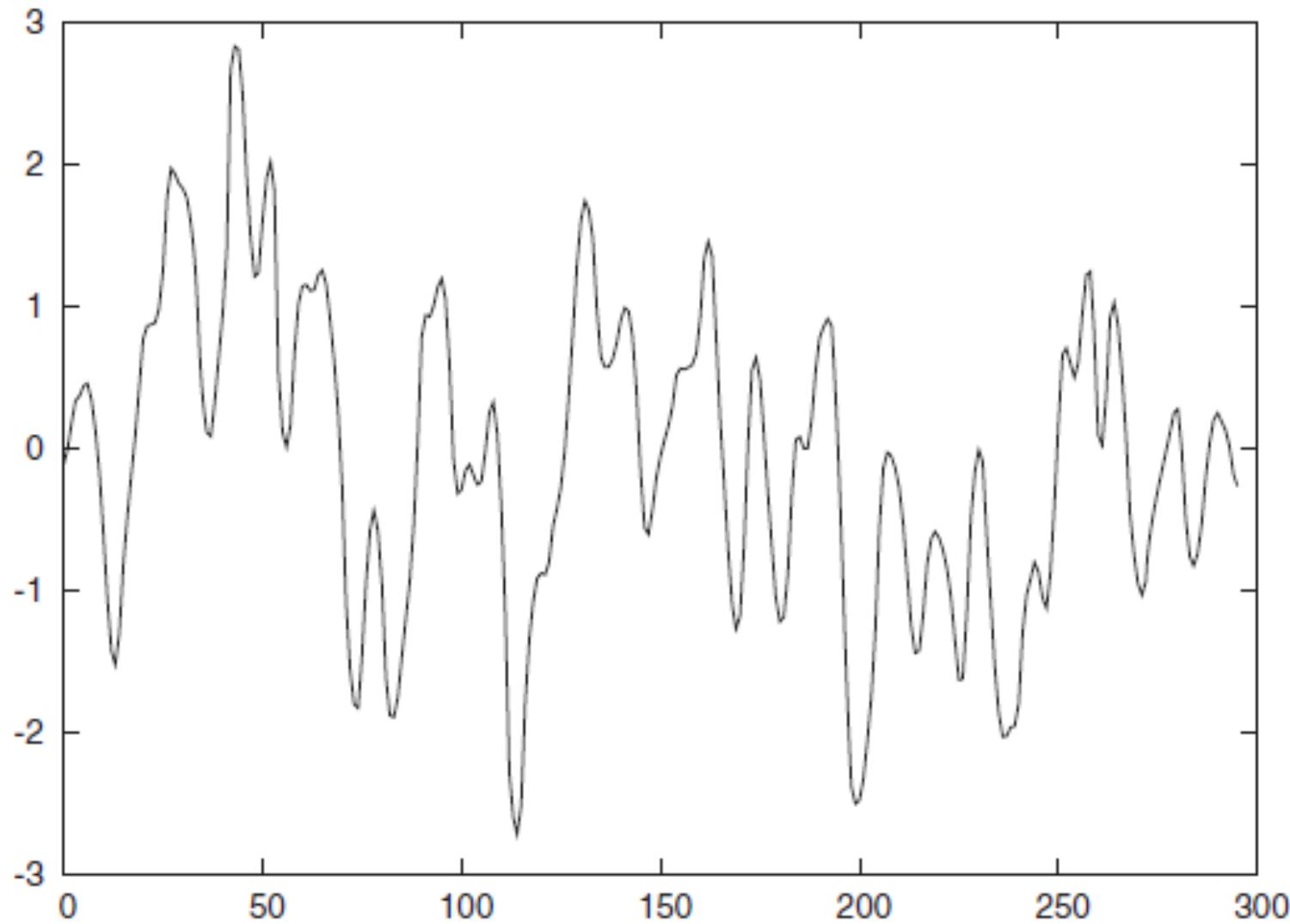
- Απεικόνιση των παρατηρήσεων (μετρήσεων)
  - καθώς μεταβάλλεται ο χρόνος
- Η κατασκευή του είναι το **πιο σημαντικό βήμα στη ανάλυση χρονοσειρών**
  - Βοηθάει στην περιγραφή των δεδομένων και
  - Στο να διατυπωθεί ένα καλό μοντέλο περιγραφής
- Δεν είναι τόσο εύκολο/απλό να φτιαχτεί σωστά:
  - Επιλογή κλίμακας
  - Γραφικής αναπαράστασης (σημεία ή συνεχής γραμμή)
  - Τίτλος
  - Μονάδες μέτρησης
  - Ετικέτες στους άξονες, κτλ.

# 1° Παράδειγμα Χρονοσειράς



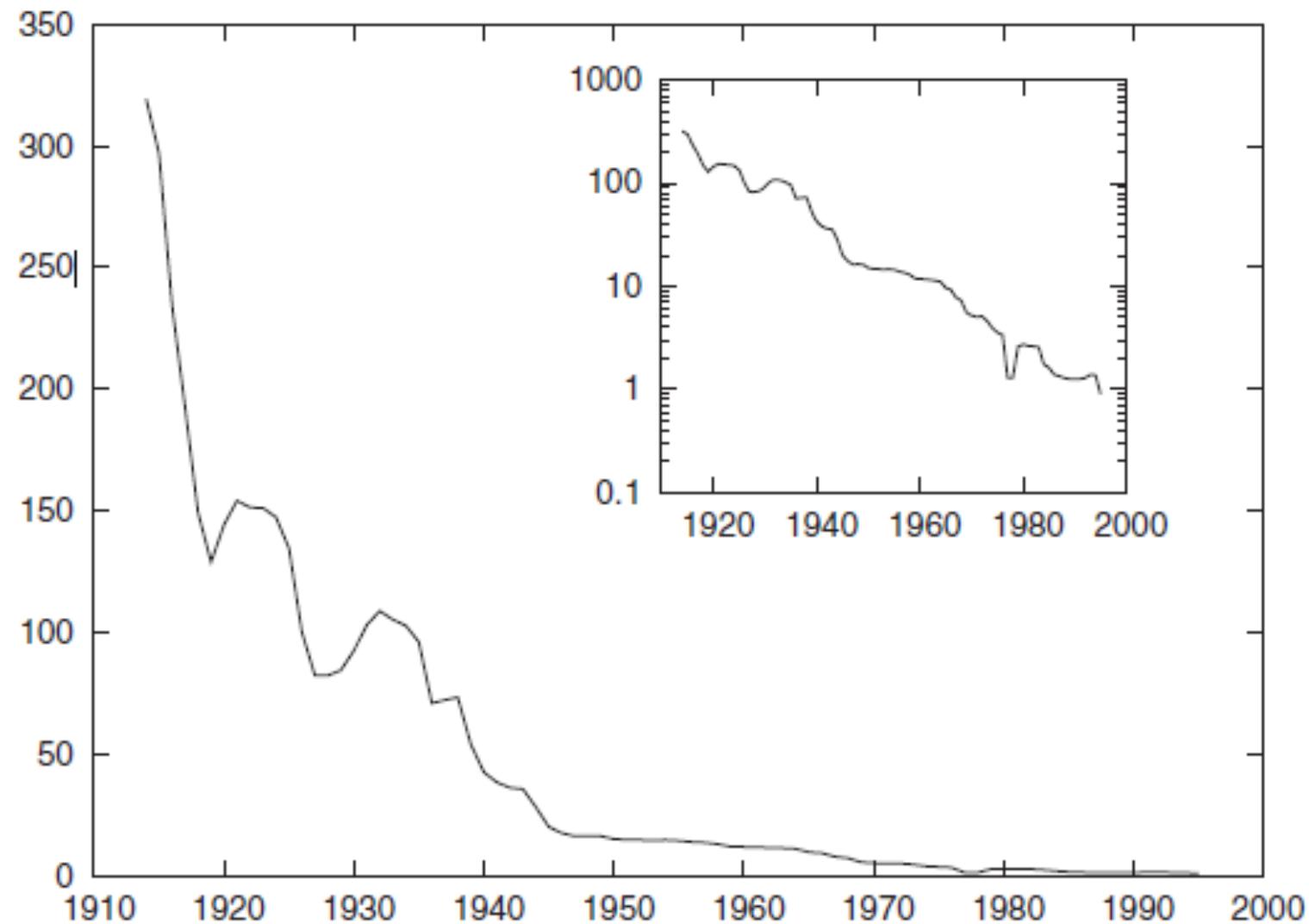
*Τάση (trend) και εποχικότητα (seasonality): η συγκέντρωση CO<sub>2</sub> (parts per million) στην ατμόσφαιρα, από μετρήσεις στο παρατηρητήριο Mauna Loa, Hawaii, κάθε μήνα.*

## 2<sup>ο</sup> Παράδειγμα Χρονοσειράς



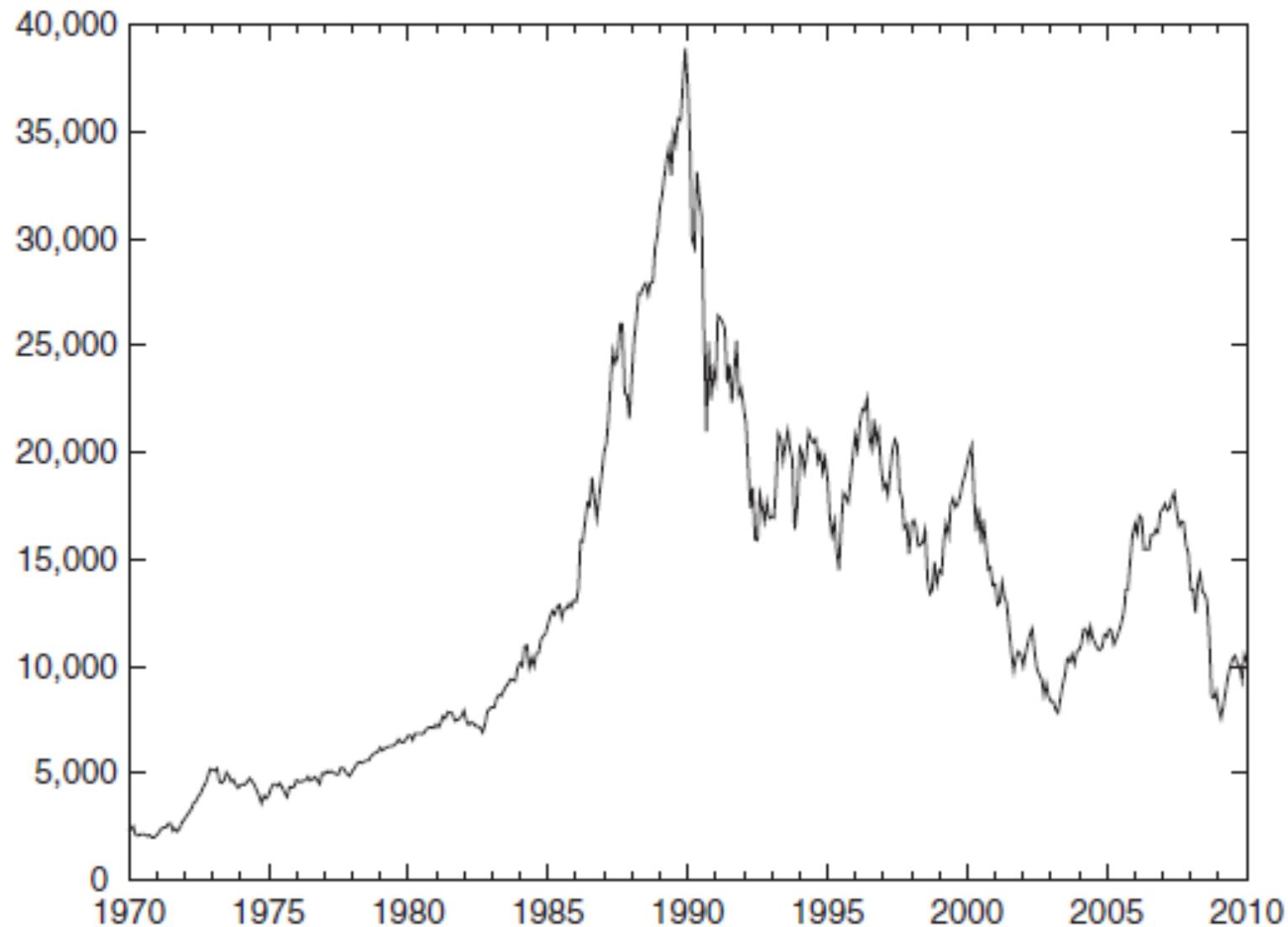
Δεν υπάρχει τάση, όμως σχετικά ομαλή μεταβολή με την πάροδο του χρόνου: συγκέντρωση ενός αερίου σε εξάτμιση φούρνου (σε αυθαίρετες μονάδες).

### 3<sup>ο</sup> Παράδειγμα Χρονοσειράς



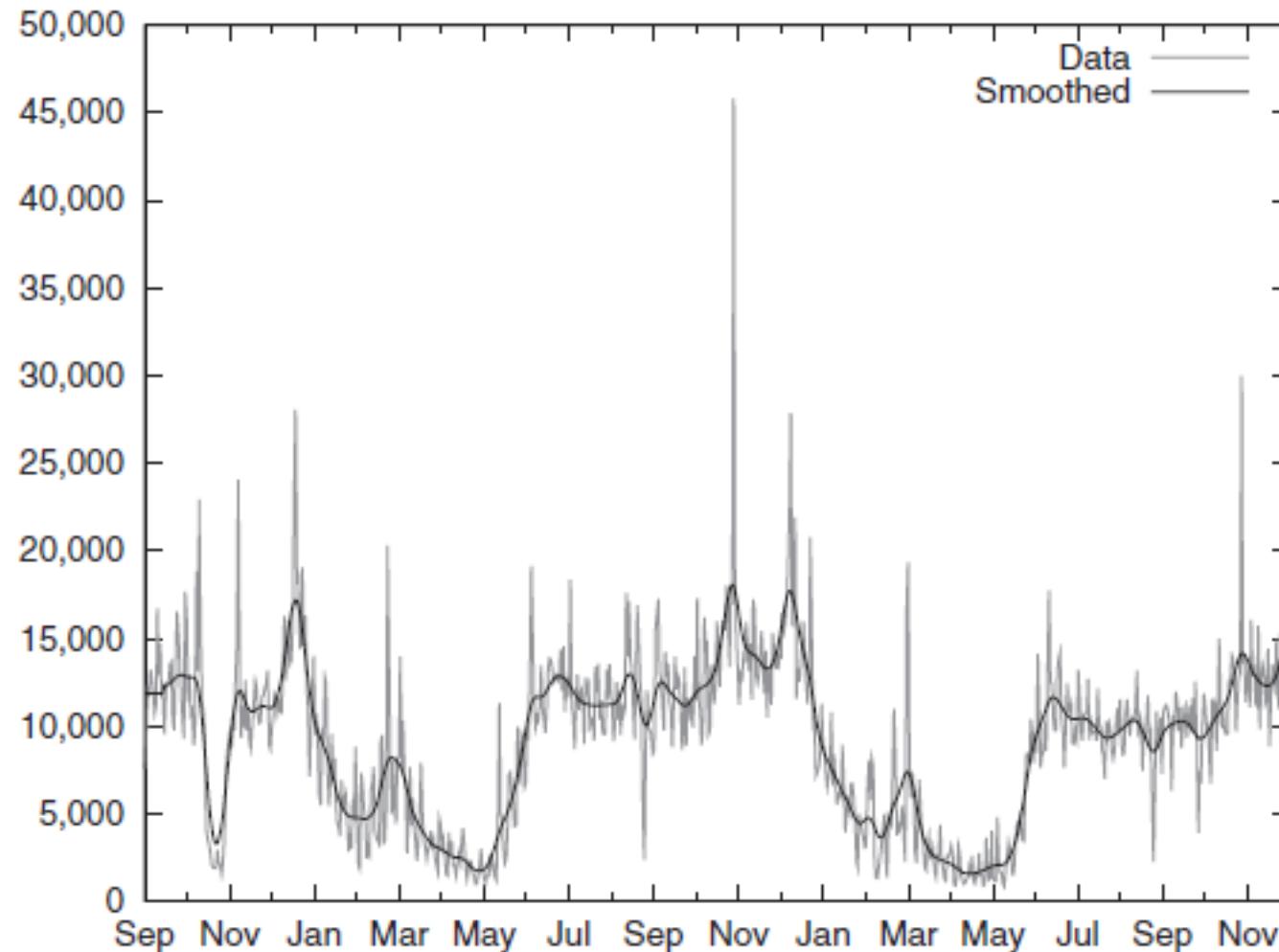
*Μη-γραμμική τάση: το κόστος ενός υπεραστικού τηλεφωνήματος στις Η.Π.Α.*

## 4<sup>ο</sup> Παράδειγμα Χρονοσειράς



Αλλαγή στη συμπεριφορά: η πορεία του χρηματιστηριακού δείκτη Nikkei για περίοδο 40 ετών

## 5° Παράδειγμα Χρονοσειράς



Το πλήθος ημερήσιων κλήσεων σε ένα τηλεφωνικό κέντρο. Τα δεδομένα εμφανίζουν βραχυπρόθεσμη και μακρυπρόθεσμη εποχικότητα, θόρυβο και πιθανώς αλλαγές στη συμπεριφορά. Εμφανίζεται και ένα γκαουσιανό φίλτρο εξομάλυνσης 31-σημείων.

# Περιεχόμενο Διάλεξης

- Παραδείγματα χρονοσειρών
- **Ανάλυση χρονοσειρών – βασικές έννοιες**
- Εντοπισμός τάσης και εποχικότητας
- Εξομάλυνση: κινητοί μέσοι όροι
- Εξομάλυνση: εκθετική εξομάλυνση
- Συνάρτηση αυτοσυσχέτισης

# Τύποι Μεταβολών

- Οι παραδοσιακές μέθοδοι ανάλυσης χρονοσειρών εστιάζουν στην **αποσύνθεση της μεταβολής** σε:
  - **τάση, εποχική μεταβολή**, άλλες **κυκλικές μεταβολές** και τις εναπομείνασες «ακανόνιστες» διακυμάνσεις
- **Εποχική μεταβολή:**
  - Πολλές χρονοσειρές παρουσιάζουν μεταβολές που έχουν περιοδικότητα μες στο έτος (π.χ. υψηλή ανεργία το χειμώνα, χαμηλή το καλοκαίρι)
  - Εύκολα κατανοητή, μπορεί να μετρηθεί και να αφαιρεθεί
- **Κυκλικές μεταβολές** (π.χ. μεταβολή θερμοκρασίας εντός ημέρας)
- **Τάση** (μακροπρόθεσμη αλλαγή του μέσου)
- **Εναπομείνασες «ακανόνιστες» διακυμάνσεις:**
  - Ό,τι απομένει (υπόλοιπα) μετά την αφαίρεση των παραπάνω

# Συστατικά Στοιχεία Χρονοσειρών (1/2)

## ■ Τάση (Trend)

- Η τάση μπορεί να είναι γραμμική ή μη-γραμμική
- Είναι επιθυμητό να διερευνήσουμε το μέγεθός της

## ■ Εποχικότητα (Seasonality)

- Το μοτίβο της εποχικότητας μπορεί να είναι προσθετικό ή πολλαπλασιαστικό
  - Προσθετικό: η εποχική αλλαγή έχει το ίδιο απόλυτο μέγεθος, ανεξάρτητα από το μέγεθος της τρέχουσας τιμής της χρονοσειράς
  - Πολλαπλασιαστικό: η εποχική αλλαγή έχει το ίδιο σχετικό μέγεθος συγκριτικά με την τρέχουσα τιμή της χρονοσειράς

# Συστατικά Στοιχεία Χρονοσειρών (2/2)

## ■ Θόρυβος (Noise)

- Ορισμένες **τυχαίες διακυμάνσεις** είναι σχεδόν πάντα μέρος μιας χρονοσειράς
- *Η εύρεση τρόπων μείωσης ή απομάκρυνσης του θορύβου* από τα δεδομένα αποτελεί σημαντικό μέρος της αναλυτικής διαδικασίας

## ■ Άλλα (Other)

- Περιλαμβάνει οτιδήποτε άλλο μπορεί να παρατηρηθεί σε μια χρονοσειρά
  - Σημαντικές αλλαγές στη συνολική συμπεριφορά
  - Ιδιαίτερες ακραίες τιμές
  - Ελλιπείς τιμές
  - ...

# Ανάλυση Χρονοσειρών

## ■ Τρεις βασικές εργασίες:

- **Περιγραφή (Description)**

Κοιτάει το παρελθόν

- Αναγνώριση τμημάτων της χρονοσειράς
- Όπως τάση, εποχικότητα, απότομες αλλαγές στη συμπεριφορά

- **Πρόβλεψη (Prediction)**

Κοιτάει το μέλλον

- Πρόβλεψη μελλοντικών τιμών

- **Έλεγχος (Control)**

Κοιτάει το παρόν

- Παρακολούθηση μιας διεργασίας κατά την πάροδο του χρόνου
- Με σκοπό να παραμένει σε ένα εύρος τιμών
- Πρόκειται για κλασική εργασία σε περιβάλλον παραγωγής

# Απαιτήσεις και ο Πραγματικός Κόσμος

- Οι περισσότερες μέθοδοι ανάλυσης χρονοσειρών κάνουν ορισμένες **υποθέσεις** για τα διαθέσιμα δεδομένα, όπως:
  - Τα δεδομένα έχουν ληφθεί σε **ίσα χρονικά διαστήματα**, και **δεν υπάρχουν ελλιπείς τιμές**
  - Η χρονοσειρά είναι **αρκετά μεγάλη** (ένα σύνολο 50 παρατηρήσεων θεωρείται συχνά ως το ελάχιστο δυνατό)
  - Η χρονοσειρά είναι **στάσιμη (stationary)**: **δεν περιέχει κάποια τάση**, **δεν έχει εποχικότητα**, και ο **χαρακτήρας θορύβου** (μέγεθος και συχνότητα) που πιθανώς υπάρχει **δεν αλλάζει με το χρόνο**
- Οι περισσότερες από αυτές τις υποθέσεις παραβιάζονται όταν μελετούμε ένα πραγματικό σύνολο δεδομένων
  - Ως εκ τούτου, απαιτείται **καθαρισμός δεδομένων (data cleaning)**

# Περιεχόμενο Διάλεξης

- Παραδείγματα χρονοσειρών
- Ανάλυση χρονοσειρών – βασικές έννοιες
- **Εντοπισμός τάσης και εποχικότητας**
- Εξομάλυνση: κινητοί μέσοι όροι
- Εξομάλυνση: εκθετική εξομάλυνση
- Συνάρτηση αυτοσυσχέτισης

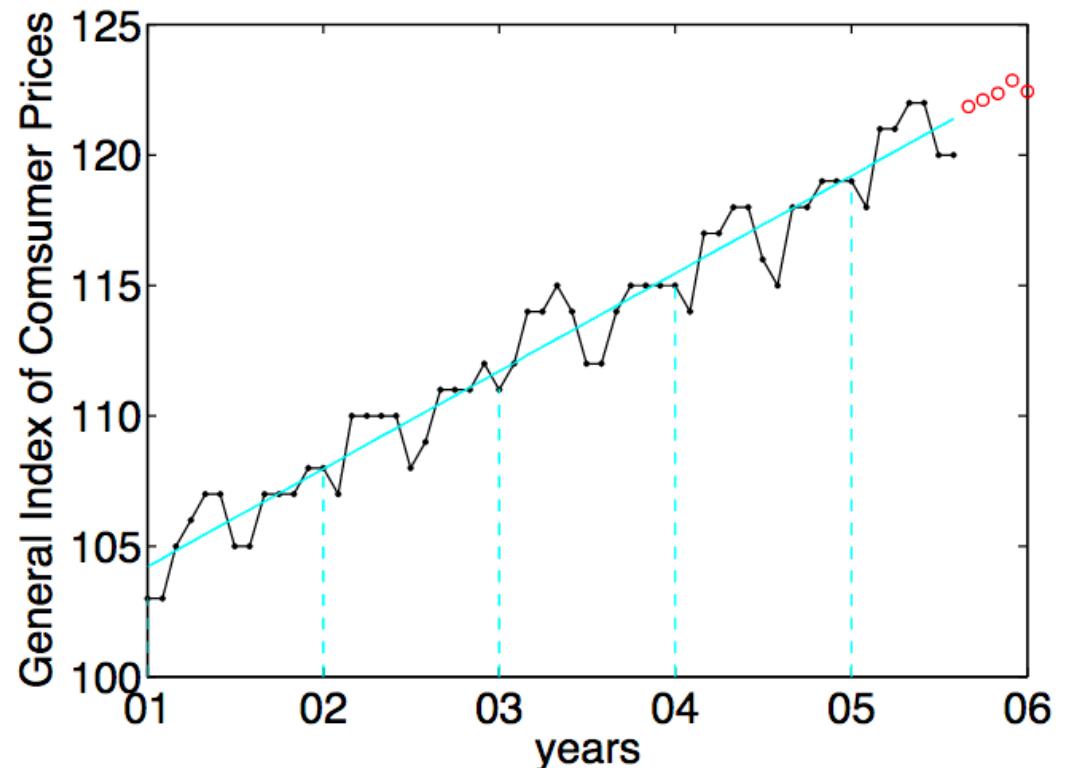
# Απαλοιφή Τάσης και Εποχικότητας

Αναπαράσταση χρονοσειράς:  $X_t = m_t + s_t + y_t$

- $m_t$  : η συνιστώσα της τάσης
- $s_t$  : η συνιστώσα της περιοδικότητας για περίοδο  $d$  ( $s_{t-d}=s_t$ )
- $y_t$  : η εναπομείνουσα χρονοσειρά (**χρονοσειρά υπολοίπων – residuals**) αν αφαιρεθεί η τάση και η εποχικότητα

# Παράδειγμα (1/5)

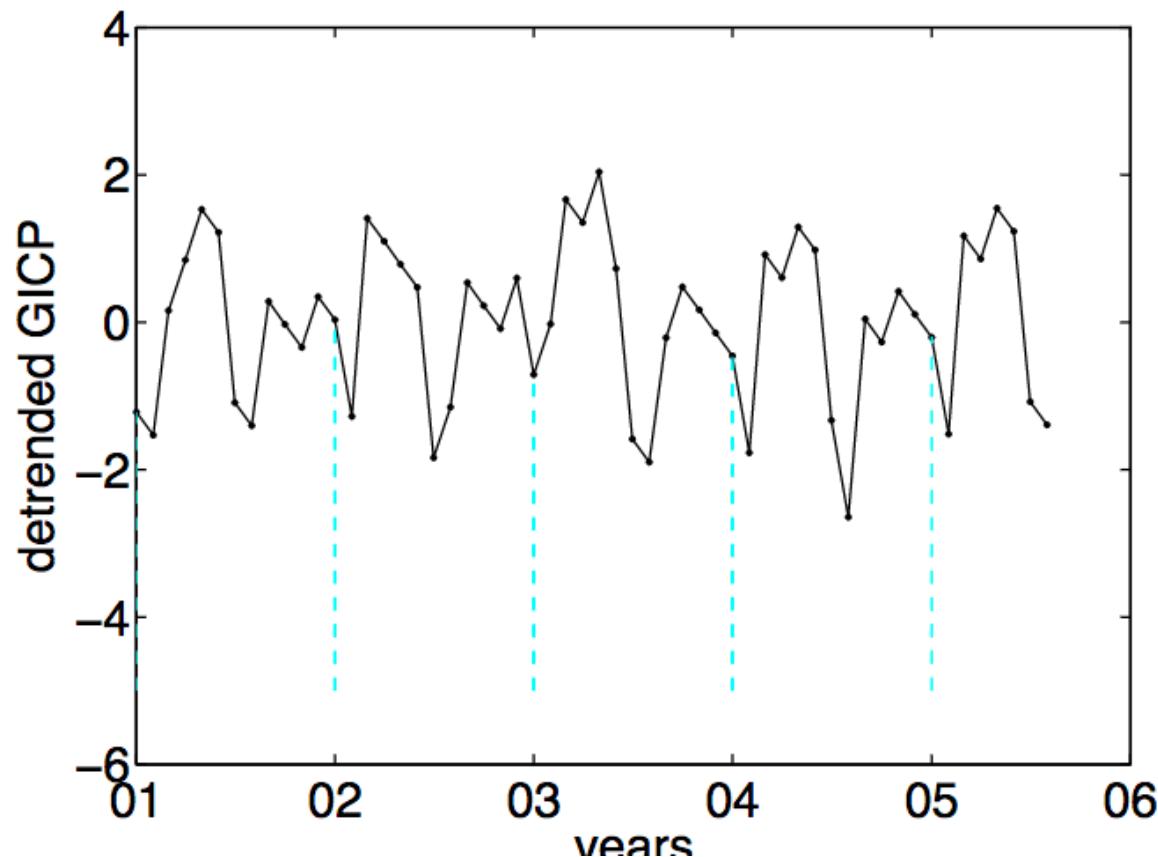
- Γενικός δείκτης τιμών καταναλωτή
  - 1/2001 – 8/2005
  - $\{x_1 \dots x_{56}\}$
- Εντοπισμός τάσης
  - Με προσαρμογή απλού μοντέλου γραμμικής παλινδρόμησης:
  - $\mu_t = 103.9 + 0.31 * t$



Πηγή: <http://users.auth.gr/dkugiu/Teach/DataAnalysis/Chp6.pdf>

## Παράδειγμα (2/5) – Απαλοιφή Τάσης

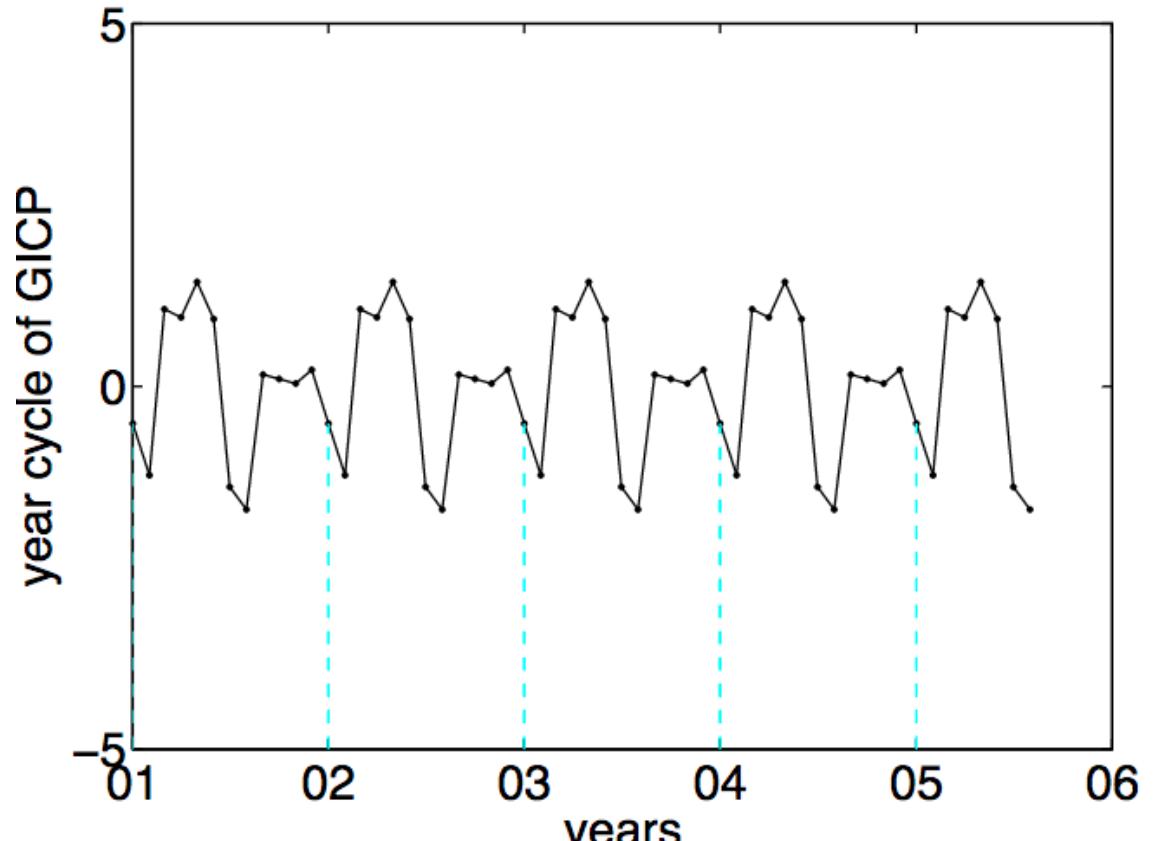
$$x'_t = x_t - \mu_t$$



Πηγή: <http://users.auth.gr/dkugiu/Teach/DataAnalysis/Chp6.pdf>

# Παράδειγμα (3/5) – Απαλοιφή Εποχιότητας

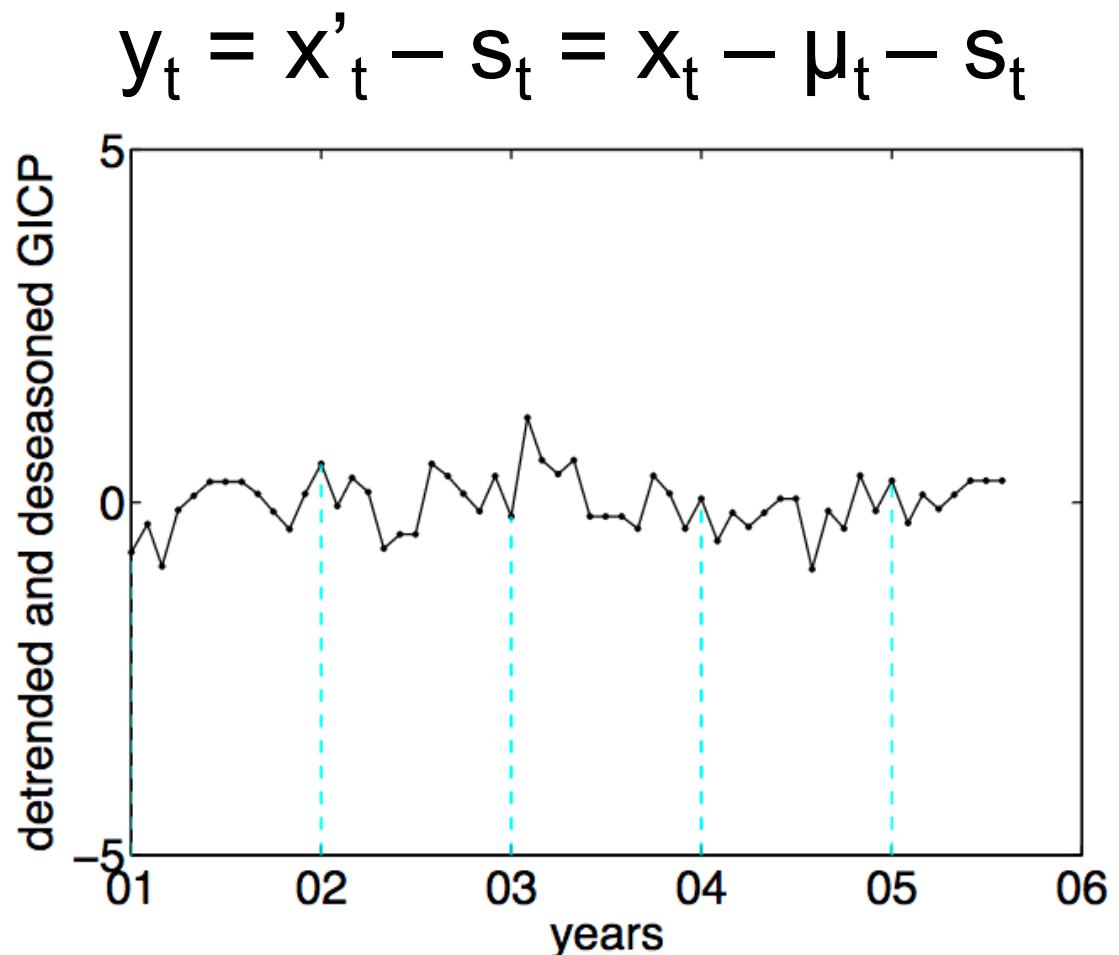
- Υπολογίζουμε για κάθε μήνα τη μέση τιμή
  - Π.χ. για Ιανουάριο, τη μέση τιμή των μηνών Ιανουαρίου για τα έτη 2001-2005
  - Αφαιρούμε αυτή την τιμή από τις τιμές κάθε Ιανουαρίου
- Αποτέλεσμα:
  - **Η χρονοσειρά απαλλαγμένη από εποχικότητα**



Πηγή: <http://users.auth.gr/dkugiu/Teach/DataAnalysis/Chp6.pdf>

# Παράδειγμα (4/5) – Υπόλοιπα

- Η χρονοσειρά των υπολοίπων
- Απαλλαγμένη από τάση και εποχικότητα
- Εάν η χρονοσειρά είναι εντελώς τυχαία, η ανάλυση σταματά εδώ
  - Δηλαδή να μη διακρίνεται κάποια περιοδικότητα ή δομή

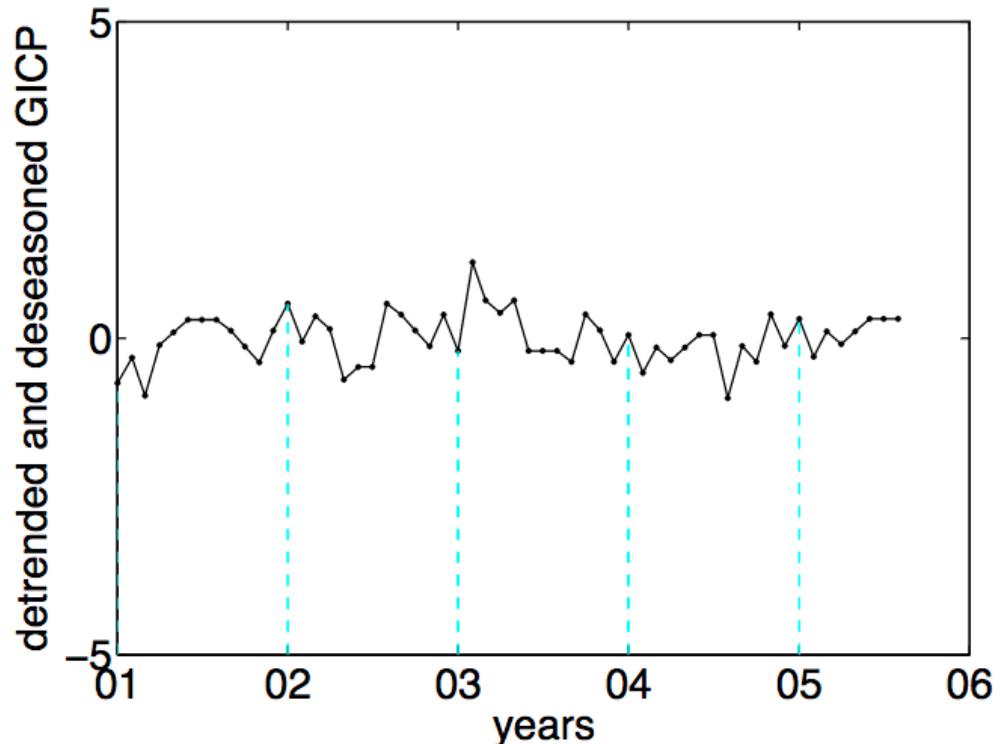


Πηγή: <http://users.auth.gr/dkugiu/Teach/DataAnalysis/Chp6.pdf>

# Παράδειγμα (5/5)

- Για την πρόβλεψη τιμής (π.χ. Σεπτεμβρίου)
  - Επεκτείνουμε (extrapolate) τη συνάρτηση τάσης για το μήνα Σεπτέμβριο και θα προσθέσουμε σε αυτήν την τιμή του ετήσιου κύκλου για το μήνα Σεπτέμβριο
- $\mu_{n+1} = 103.9 + 0.31 * (n + 1)$
- $\mu_{57} = 103.9 + 0.31 * 57 = 121.7$
- $s_9 = 0.16$
- $\bar{x}_{57} = \mu_{57} + s_9 = 121.86$

$$y_t = x'_t - s_t = x_t - \mu_t - s_t$$



Πηγή: <http://users.auth.gr/dkugiu/Teach/DataAnalysis/Chp6.pdf>

# Απαλοιφή Τάσης

- Στο προηγούμενο παράδειγμα χρησιμοποιήθηκε προσαρμογή ενός μοντέλου στα δεδομένα για τον εντοπισμό της τάσης
  - Θα μπορούσε η προσέγγιση να γενικευτεί με χρήση πολυώνυμου μεγαλύτερου βαθμού
- Όταν όμως η τάση δεν μπορεί να περιγραφεί ικανοποιητικά ως μια συνάρτηση του χρόνου, αλλά είναι τυχαία (στοχαστική τάση), τότε για την απαλοιφή της μπορούμε να πάρουμε τις διαφορές
  - $y_t = \nabla x_t = x_t - x_{t-1}$
- Υπάρχουν και άλλα φίλτρα, όπως κινητός μέσος όρος

# Εξομάλυνση (Smoothing)

- Η απομάκρυνση του θορύβου ή του λάχιστον η μείωση της επίδρασής του
  - Είναι εξαιρετικά σημαντική στην ανάλυση χρονοσειρών
- Με άλλα λόγια, αναζητούμε τρόπους για να **εξομαλύνουμε το σήμα**
- Όταν η τεχνική εξομάλυνσης εφαρμόζεται επιτυχώς, μπορεί να αναδείξει πιο ξεκάθαρα την τάση ή την εποχικότητα ή κάποια περιοδικότητα
- Δύο βασικές κατηγορίες μεθόδων εξομάλυνσης:
  - **Κινητοί μέσοι όροι**
  - **Εκθετική εξομάλυνση**

# Περιεχόμενο Διάλεξης

- Παραδείγματα χρονοσειρών
- Ανάλυση χρονοσειρών – βασικές έννοιες
- Εντοπισμός τάσης και εποχικότητας
- **Εξομάλυνση: κινητοί μέσοι όροι**
- Εξομάλυνση: εκθετική εξομάλυνση
- Συνάρτηση αυτοσυσχέτισης

# Απλός Μέσος Όρος

- Ο απλούστερος τρόπος εξομάλυνσης είναι με χρήση απλού μέσου όρου
- Ένας διευθυντής αποθήκης θέλει να υπολογίσει τι ποσό λαμβάνει από έναν τυπικό προμηθευτή (σε χιλιάδες δολάρια)
- Παίρνει δείγμα 12 πελατών
- Ο μέσος όρος στο διπλανό παράδειγμα δίνει 10 χιλ. δολάρια
- Πρόκειται για καλή ή κακή εκτίμηση;

Sample of 12 suppliers

Supplier	Amount	Supplier	Amount
1	9	7	11
2	8	8	7
3	9	9	13
4	12	10	9
5	9	11	11
6	12	12	10

# Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error)

- "error": πραγματικό ποσό μείον το εκτιμώμενο
- "error squared": το τετράγωνο του παραπάνω
- "SSE": το άθροισμα των τετραγωνικών σφαλμάτων
- "MSE": ο μέσος όρος των τετραγωνικών σφαλμάτων

$$SSE = 36 \text{ and } MSE = 36/12 = 3$$

Sample of 12 suppliers

Supplier	Amount	Supplier	Amount
1	9	7	11
2	8	8	7
3	9	9	13
4	12	10	9
5	9	11	11
6	12	12	10

Supplier	Amount	Error	Error Squared
1	9	-1	1
2	8	-2	4
3	9	-1	1
4	12	2	4
5	9	-1	1
6	12	2	4
7	11	1	1
8	7	-3	9
9	13	3	9
10	9	-1	1
11	11	1	1
12	10	0	0

# Μέσο Τετραγωνικό Σφάλμα σε σχέση με άλλες Εκτιμήσεις

- Πόσο καλή ήταν η εκτίμηση;
- Συγκρίνουμε την εκτίμηση (10) με άλλες εκτιμήσεις: 7, 9, 12
- *Η εκτίμηση που ελαχιστοποιεί την τιμή μέσου τετραγωνικού σφάλματος (MSE) είναι η καλύτερη*
  - Μπορεί να αποδειχθεί ότι για ένα τυχαίο σύνολο δεδομένων, **η εκτίμηση που ελαχιστοποιεί το MSE είναι ο μέσος όρος**

Estimator	7	9	10	12
SSE	144	48	36	84
MSE	12	4	<b>3</b>	7

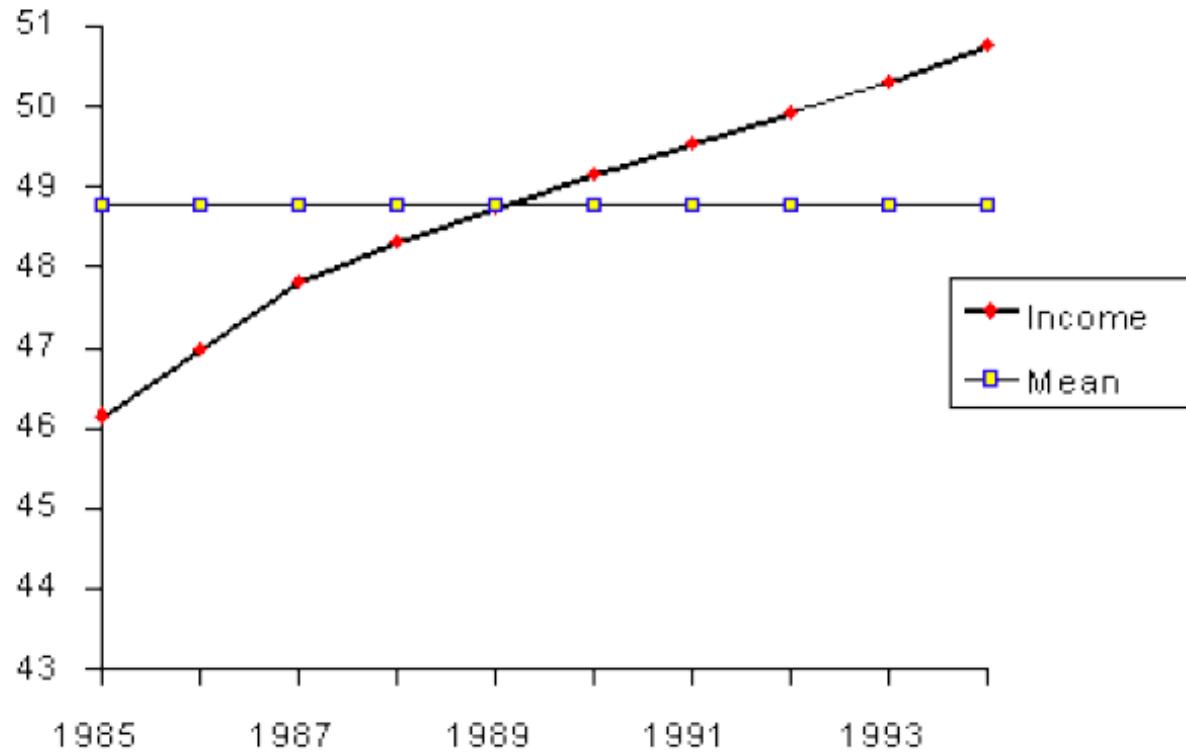
# Χρήση του Μέσου Όρου για Πρόβλεψη

MSE = 1.8129.

- Για έναν κατασκευαστή υπολογιστών από το 1985 ως το 1994
- Μπορούμε να χρησιμοποιήσουμε το μέσο όρο για την πρόβλεψη του εισοδήματος στο μέλλον, εάν υπάρχει κάποια τάση στα δεδομένα;

Year	\$ (millions)	Mean	Error	Squared Error
1985	46.163	48.676	-2.513	6.313
1986	46.998	48.676	-1.678	2.814
1987	47.816	48.676	-0.860	0.739
1988	48.311	48.676	-0.365	0.133
1989	48.758	48.676	0.082	0.007
1990	49.164	48.676	0.488	0.239
1991	49.548	48.676	0.872	0.761
1992	48.915	48.676	0.239	0.057
1993	50.315	48.676	1.639	2.688
1994	50.768	48.676	2.092	4.378

# Χρήση του Μέσου 'Ορου για Πρόβλεψη



- Ο μέσος όρος δίνει την ίδια στάθμιση (βάρος) σε όλες τις παρατηρήσεις
- Τα βάρη είναι  $1/n$  και αθροίζουν στο 1:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \left(\frac{1}{n}\right)x_1 + \left(\frac{1}{n}\right)x_2 + \dots + \left(\frac{1}{n}\right)x_n$$

# Εξομάλυνση με Κινητό Μέσο 'Ορο

MSE = 2.42

- Ένας εναλλακτικός τρόπος για να εξάγουμε μια περίληψη των παρελθουσών τιμών
- Κοιτώντας 3 παρατηρήσεις
  - Ο κινητός μέσος όρος γίνεται:  $(9+8+9)/3=8.667$

Supplier	\$	MA	Error	Error squared
1	9			
2	8			
3	9	8.667	0.333	0.111
4	12	9.667	2.333	5.444
5	9	10.000	-1.000	1.000
6	12	11.000	1.000	1.000
7	11	10.667	0.333	0.111
8	7	10.000	-3.000	9.000
9	13	10.333	2.667	7.111
10	9	9.667	-0.667	0.444
11	11	11.000	0	0
12	10	10.000	0	0

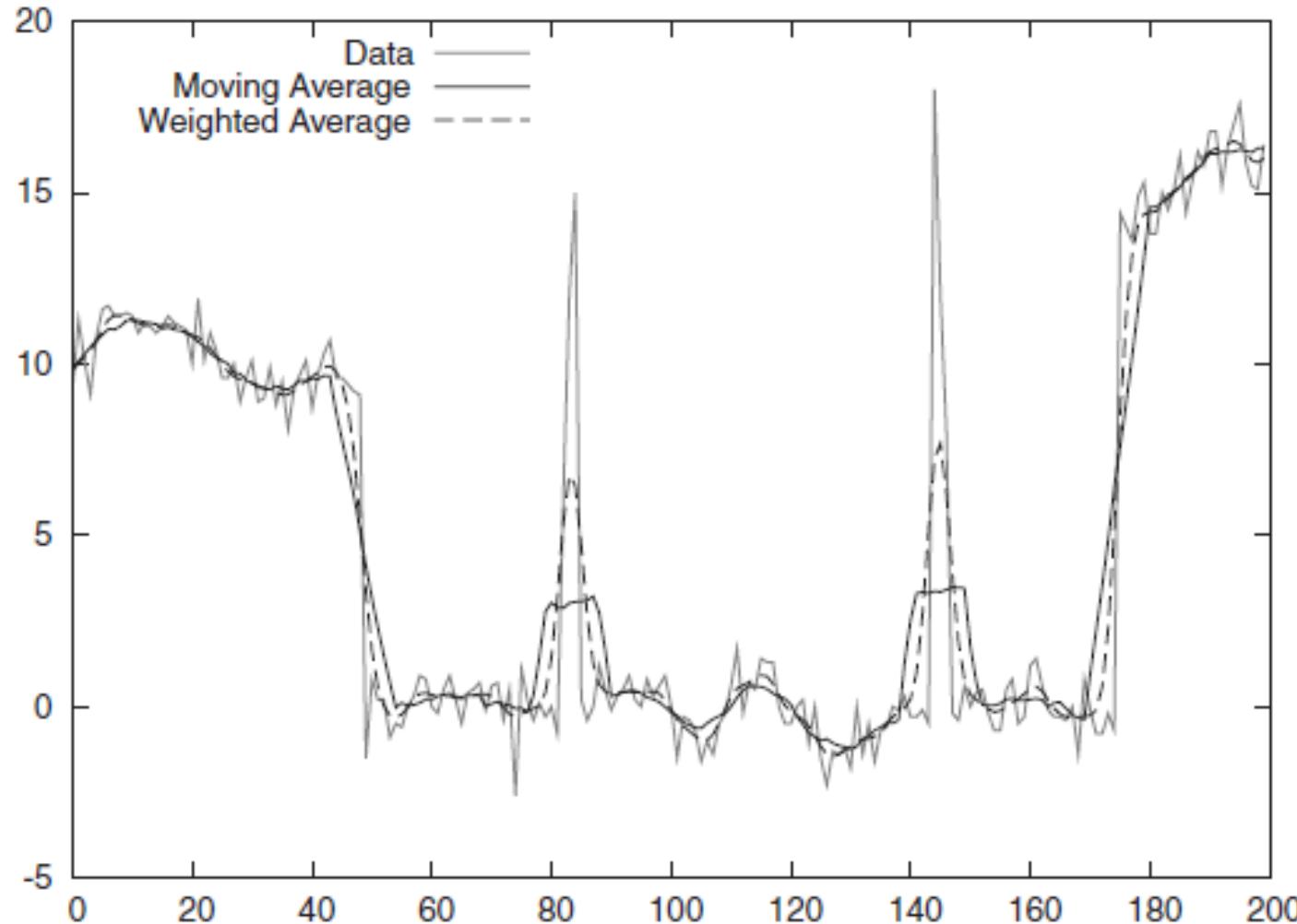
# Κινητοί Μέσοι 'Όροι (Running Averages)

- Πιθανώς ο απλούστερος αλγόριθμος εξομάλυνσης
  - Για οποιοδήποτε περιττό σύνολο  $2k+1$  συνεχόμενων σημείων, αντικαθιστούμε την **κεντρική τιμή** με το **μέσο όρο** των υπολοίπων

$$s_i = \frac{1}{2k+1} \sum_{j=-k}^k x_{i+j}$$

- $\{x_i\}$  είναι το σύνολο δεδομένων
  - $s_i$  είναι η εξομαλυμένη τιμή στη θέση  $i$

# Εξομάλυνση – Κινητοί Μέσοι 'Όροι



## Πρόβλημα:

Κάθε φορά που μια πολύ υψηλή τιμή εισέρχεται στο τρέχον «παράθυρο» 11 σημείων, ο κινητός μέσος όρος αλλάζει απότομα μέχρι να εξέλθει η τιμή αυτή από το παράθυρο

Απλός και σταθμισμένος (γκαουσιανός) κινητός μέσος όρος 11 σημείων: ο σταθμισμένος μέσος όρος επηρεάζεται λιγότερο από απότομες αλλαγές στα δεδομένα

# Σταθμισμένος Κινητός Μέσος 'Ορος (Weighted Moving Average)

- Δίνει μικρότερο βάρος στα σημεία που βρίσκονται στα άκρα του παραθύρου εξομάλυνσης
- Έτσι, κάθε νέο σημείο που εισέρχεται στο παράθυρο, συνεισφέρει σταδιακά (κι όχι απότομα) στο μέσο όρο και σταδιακά αφαιρείται
- Για παράδειγμα, για 3 σημεία τα βάρη είναι  $(1/4, 1/2, 1/4)$

$$s_i = \sum_{j=-k}^k w_j x_{i+j} \quad \text{where} \quad \sum_{j=-k}^k w_j = 1$$

# Υπολογισμός Συντελεστών Στάθμισης

- Μπορεί να χρησιμοποιηθεί η κανονική κατανομή για την κατασκευή των συντελεστών στάθμισης:

$$f(x, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{x}{\sigma}\right)^2\right) \quad (\text{επιστρέφει σχεδόν μηδενικές τιμές για } x > 3.5\sigma)$$

- $f(x, 1)$ : προκύπτουν 9 σημεία αποτιμώντας την  $f(x, 1)$  στα σημεία [-4, -3, -2, -1, 0, 1, 2, 3, 4]
  - $f(x, 2)$ : προκύπτουν 15 σημεία αποτιμώντας την  $f(x, 2)$  στα σημεία [-7...7]
- όμως γενικά, μας κάνει οποιαδήποτε συνάρτηση με μια κορυφή στο κέντρο και βάρη που αθροίζουν στη μονάδα

# Περιεχόμενο Διάλεξης

- Παραδείγματα χρονοσειρών
- Ανάλυση χρονοσειρών – βασικές έννοιες
- Εντοπισμός τάσης και εποχικότητας
- Εξομάλυνση: κινητοί μέσοι όροι
- **Εξομάλυνση: εκθετική εξομάλυνση**
- Συνάρτηση αυτοσυσχέτισης

# Περιορισμοί Μεθόδων Κινητών Μέσων 'Ορων

- Όλες οι μέθοδοι κινητών μέσων όρων έχουν κάποια **προβλήματα**
  - Είναι **ακριβοί υπολογιστικά**
    - καθώς για κάθε σημείο πρέπει να γίνει η αποτίμηση εκ νέου
    - δεν μπορεί να υπολογιστεί ενημερώνοντας απλά το προηγούμενο αποτέλεσμα
  - Δεν μπορούν να εφαρμοστούν στα **άκρα** του συνόλου δεδομένων
  - Δεν ορίζεται κινητός μέσος όρος **εκτός του εύρους** του πεδίου ορισμού
    - δε χρησιμεύει για πρόβλεψη

# Εκθετική Εξομάλυνση (Exponential Smoothing ή Μέθοδος Holt-Winters)

- Για την αντιμετώπιση αυτών των προβλημάτων
- **Εκθετική εξομάλυνση (ή μέθοδος Holt-Winters)**
  - **Απλή εκθετική εξομάλυνση** (single exponential smoothing):
    - για χρονοσειρές που δεν παρουσιάζουν τάση ούτε εποχικότητα
  - **Διπλή εκθετική εξομάλυνση** (double exponential smoothing):
    - για χρονοσειρές που παρουσιάζουν τάση, όμως όχι εποχικότητα
  - **Τριπλή εκθετική εξομάλυνση** (triple exponential smoothing):
    - για χρονοσειρές που παρουσιάζουν και τάση και εποχικότητα

# Εκθετική Εξομάλυνση

- Η αναδρομική σχέση για την απλή εκθετική εξομάλυνση  
**(single exponential smoothing)**

$$s_i = \alpha x_i + (1 - \alpha)s_{i-1} \quad \text{with } 0 \leq \alpha \leq 1$$

$$\begin{aligned} s_i &= \alpha x_i + (1 - \alpha)s_{i-1} \\ &= \alpha x_i + (1 - \alpha)[\alpha x_{i-1} + (1 - \alpha)s_{i-2}] \\ &= \alpha x_i + (1 - \alpha)[\alpha x_{i-1} + (1 - \alpha)[\alpha x_{i-2} + (1 - \alpha)s_{i-3}]] \\ &= \alpha [x_i + (1 - \alpha)x_{i-1} + (1 - \alpha)^2 x_{i-2}] + (1 - \alpha)^3 s_{i-3} \\ &= \dots \\ &= \alpha \sum_{j=0}^i (1 - \alpha)^j x_{i-j} \end{aligned}$$

Μπορεί να χρησιμοποιηθεί και για πρόβλεψη (όπου  $s_i$  η τελευταία υπολογισθείσα τιμή):

$$x_{i+h} = s_i$$

# Διπλή Εκθετική Εξομάλυνση

- Η απλή εκθετική εξομάλυνση δουλεύει καλά όταν δεν υπάρχει κάποια τάση στα δεδομένα
- Λύση: **διπλή εκθετική εξομάλυνση**
  - Διατηρεί την πληροφορία για την τάση

$$s_i = \alpha x_i + (1 - \alpha)(s_{i-1} + \underline{t_{i-1}})$$

$$t_i = \beta(s_i - s_{i-1}) + (1 - \beta)\underline{t_{i-1}}$$

← Εξομαλυμένη τάση

- Πρόβλεψη:

$$x_{i+h} = s_i + h t_i$$

# Τριπλή Εκθετική Εξομάλυνση

- Προσθέτουμε μια τρίτη ποσότητα που περιγράφει την εποχικότητα

Προσθετική

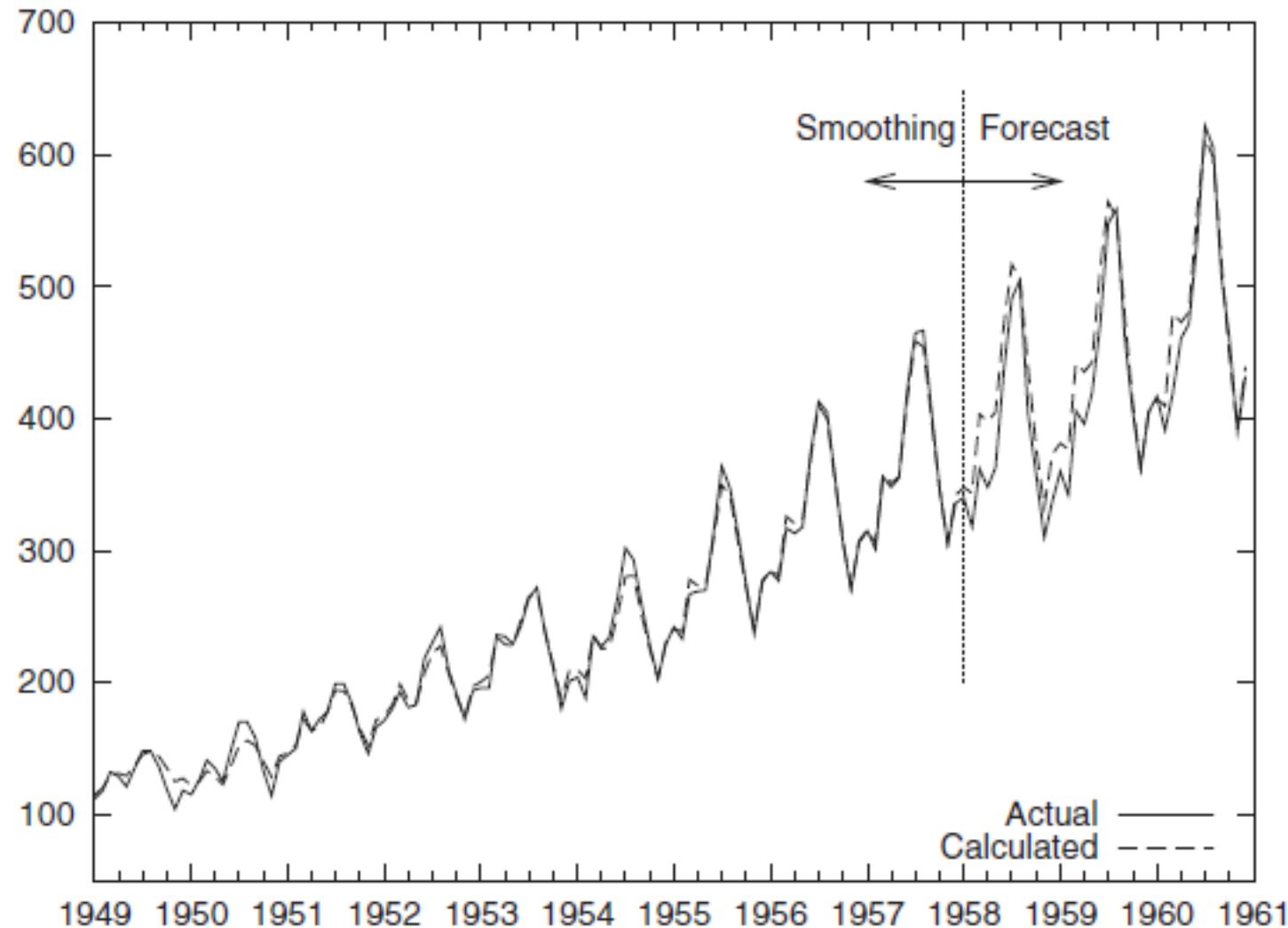
$$\begin{aligned}s_i &= \alpha(x_i - p_{i-k}) + (1 - \alpha)(s_{i-1} + t_{i-1}) \\t_i &= \beta(s_i - s_{i-1}) + (1 - \beta)t_{i-1} \\p_i &= \gamma(x_i - s_i) + (1 - \gamma)p_{i-k} \\x_{i+h} &= s_i + h t_i + p_{i-k+h}\end{aligned}$$

Πολλαπλασιαστική

$$\begin{aligned}s_i &= \alpha \frac{x_i}{p_{i-k}} + (1 - \alpha)(s_{i-1} + t_{i-1}) \\t_i &= \beta(s_i - s_{i-1}) + (1 - \beta)t_{i-1} \\p_i &= \gamma \frac{x_i}{s_i} + (1 - \gamma)p_{i-k} \\x_{i+h} &= (s_i + h t_i) p_{i-k+h}\end{aligned}$$

όπου k: η περίοδος

# Εκθετική Εξομάλυνση



Μηνιαίο πλήθος επιβατών σε πτήσεις εξωτερικού (σε χιλιάδες επιβατών)

Τριπλή εκθετική εξομάλυνση: σύγκριση μεταξύ των πρωτογενών δεδομένων (συνεχής γραμμή) και εξομαλυμένων (διακεκομμένη γραμμή)

# Περιεχόμενο Διάλεξης

- Παραδείγματα χρονοσειρών
- Ανάλυση χρονοσειρών – βασικές έννοιες
- Εντοπισμός τάσης και εποχικότητας
- Εξομάλυνση: κινητοί μέσοι όροι
- Εξομάλυνση: εκθετική εξομάλυνση
- **Συνάρτηση αυτοσυσχέτισης**

# Συνάρτηση Συσχέτισης (Correlation Function)

- Πρόκειται για το **βασικό διαγνωστικό εργαλείο** για την ανάλυση χρονοσειρών
  - Ενώ οι μέθοδοι εξομάλυνσης χρησιμοποιούν τα πρωτογενή δεδομένα, η συνάρτηση συσχέτισης παρέχει μια άλλη όψη των ίδιων δεδομένων
- **Συνάρτηση συσχέτισης – Correlation function  $r_k$  σε τιμή υστέρησης  $k$  (lag value) (όπου  $N$  είναι το πλήθος σημείων):**

$$r_k = \frac{\sum_{i=1}^{N-k} (x_i - \mu)(x_{i+k} - \mu)}{\sum_{i=1}^N (x_i - \mu)^2}$$

$r_0$

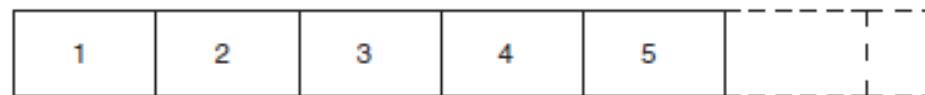
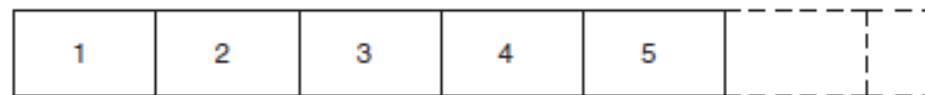
$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

# Υπολογισμός Συνάρτησης Συσχέτισης

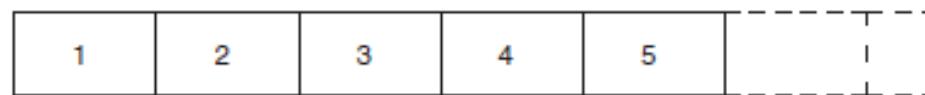
- Παίρνουμε δύο αντίγραφα του συνόλου δεδομένων και αφαιρούμε το μέσο όρο από όλες τις τιμές
- Ευθυγραμμίζουμε (χρονικά) τα δύο σύνολα δεδομένων και πολλαπλασιάζουμε τις τιμές στις αντίστοιχες χρονικές στιγμές
- Αθροίζουμε τα γινόμενα για όλες τις χρονικές στιγμές
- Το αποτέλεσμα είναι ο (μη κανονικοποιημένος) συντελεστής συσχέτισης στην τιμή υστέρησης 0
- Ολισθαίνουμε τα δύο αντίγραφα κατά μία χρονική στιγμή
  - Επαναλαμβάνουμε πολλαπλασιασμό και άθροισμα: το αποτέλεσμα είναι ο συντελεστής συσχέτισης σε τιμή υστέρησης 1
- Συνεχίζουμε με τον ίδιο τρόπο για όλο το μήκος της χρονοσειράς
  - Το σύνολο όλων των συντελεστών συσχέτισης είναι η συνάρτηση αυτοσυσχέτισης
- Τέλος, διαιρούμε όλους τους συντελεστές με το συντελεστή σε τιμή υστέρησης 0, ώστε τώρα ο συντελεστής σε τιμή υστέρησης 0 ισούται με 1

# Υπολογισμός Συνάρτησης Συσχέτισης

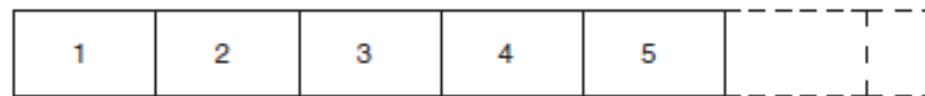
Lag: 0



Lag: 1



Lag: 2



# Επεξήγηση Συνάρτησης Συσχέτισης

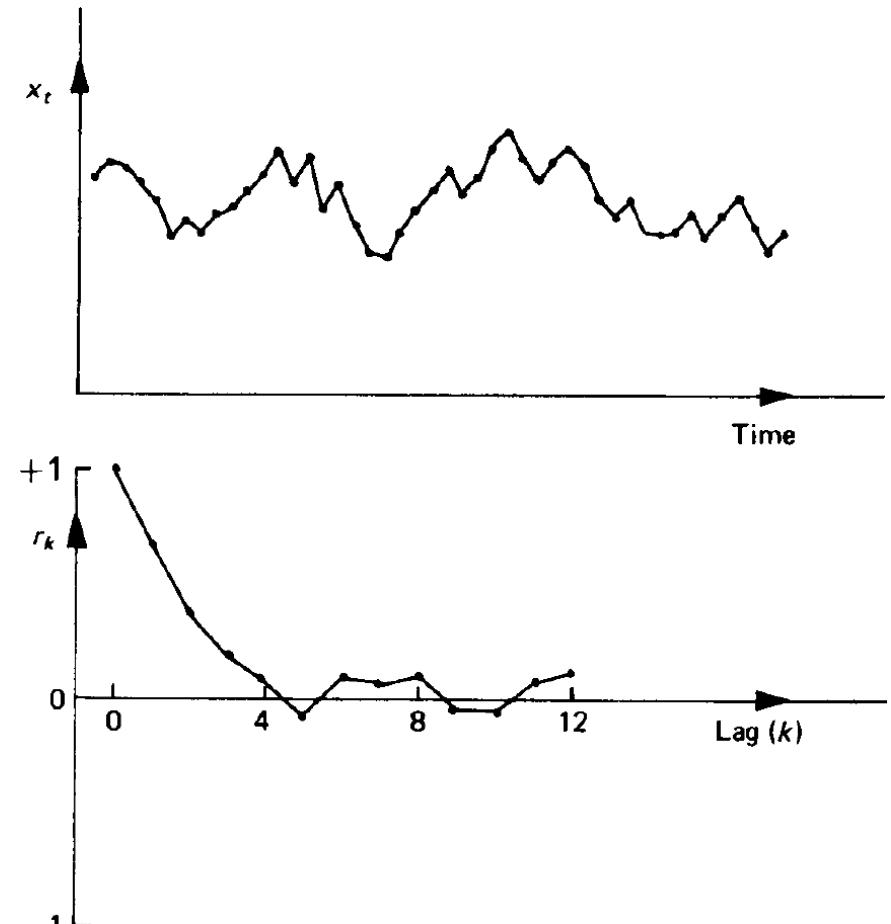
- Αρχικά, τα δύο σήματα είναι ευθυγραμμισμένα στο χρόνο και η συσχέτιση ισούται με 1
- Καθώς ολισθαίνουμε το ένα σήμα ως προς το άλλο, η συσχέτιση μειώνεται
- Το **πόσο γρήγορα μειώνεται** μας λέει πόση «μνήμη» υπάρχει στα δεδομένα
  - Εάν η συσχέτιση μειώνεται γρήγορα, γνωρίζουμε ότι **μετά από λίγα βήματα το σήμα έχει χάσει όλη τη μνήμη** του πρόσφατου παρελθόντος
  - Αν όμως η συσχέτιση μειώνεται αργά, γνωρίζουμε ότι παρατηρούμε μια χρονοσειρά που είναι **σχετικά σταθερή** σε μεγάλες χρονικές περιόδους
- Είναι επίσης **πιθανό αρχικά να μειώνεται** η συνάρτηση συσχέτισης και **μετά να αυξάνει** ξανά για να σχηματίσει δεύτερη (και ίσως τρίτη ή τέταρτη) κορυφή
  - Αυτό σημαίνει ότι τα σήματα θα ευθυγραμμιστούν ξανά αν ολισθήσουν αρκετά, ára **υπάρχει περιοδικότητα (εποχικότητα)** στα δεδομένα
  - Η θέση της δεύτερης κορυφής καθορίζει την **περιοδικότητα**

# Διάγραμμα Συσχέτισης (Correlogram)

- Το διάγραμμα συσχέτισης (correlogram) προκύπτει όταν σχεδιάζουμε τις τιμές  $r_k$  (γνωστές και ως συντελεστές συσχέτισης - correlation coefficients) ως προς την υστέρηση (lag)  $k$
- Η οπτική επισκόπηση του διαγράμματος συσχέτισης μπορεί να οδηγήσει σε πολύ ενδιαφέρουσες παρατηρήσεις, όπως:
  - **Τυχαία χρονοσειρά:** τότε για μεγάλες τιμές του  $N$  παρατηρούμε ότι  $r_k \rightarrow 0$  για μη μηδενικές τιμές του  $k$

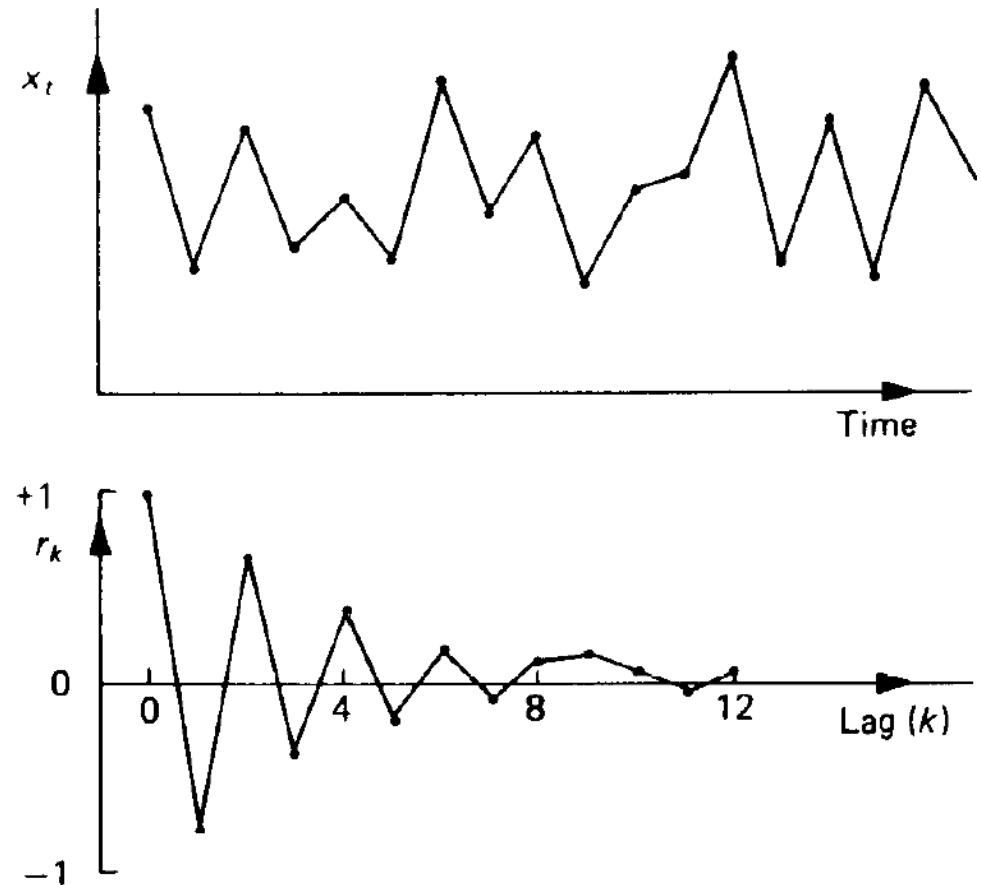
# Χρονοσειρά με Βραχεία Συσχέτιση

- Σε στάσιμες χρονοσειρές παρατηρείται μια υψηλή τιμή  $r_1$  που ακολουθείται από ορισμένους ακόμη συντελεστές που ενώ είναι μεγαλύτεροι του μηδέν, τείνουν να μικραίνουν συνεχώς
- Μια χρονοσειρά είναι **στάσιμη** όταν
  - δεν υπάρχει συστηματική μεταβολή στο μέσο (δεν περιέχει τάση),
  - δεν υπάρχει συστηματική μεταβολή στη διακύμανση, και
  - εάν περιοδικές μεταβολές έχουν αφαιρεθεί



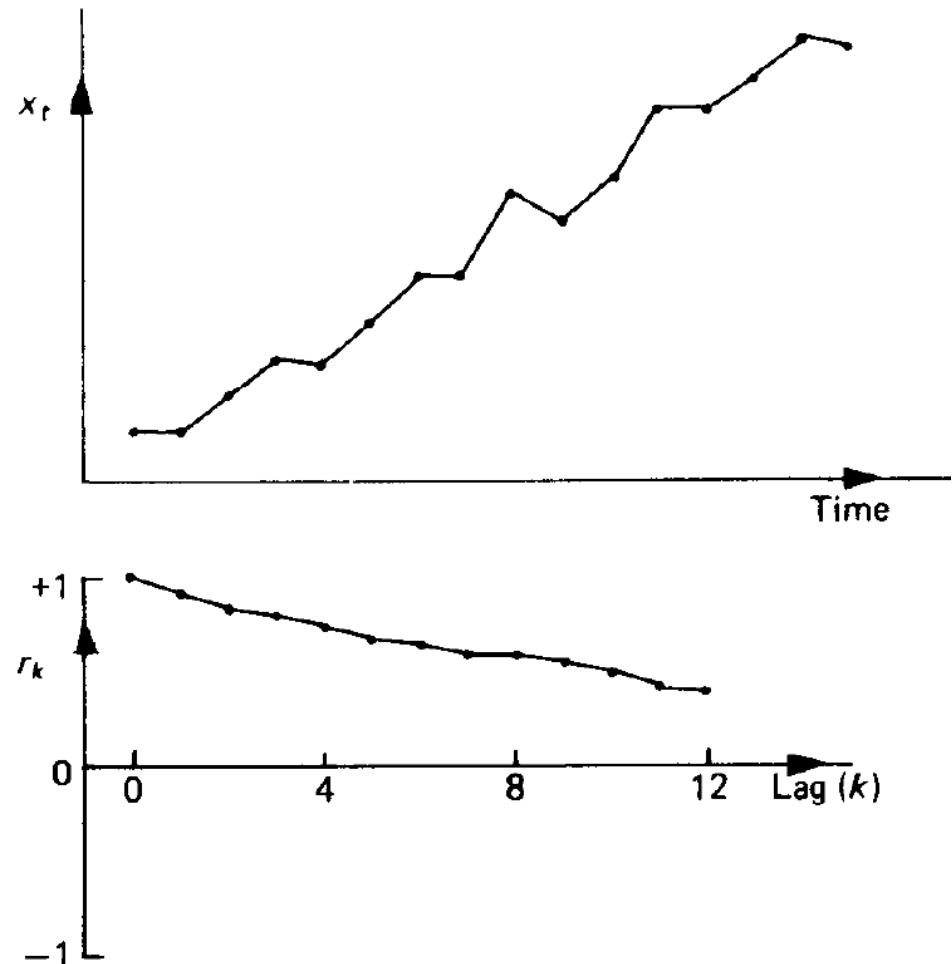
# Χρονοσειρά με Εναλλαγή Τιμών

- Αν οι τιμές μιας χρονοσειράς **εναλλάσσονται συνεχώς** σε σχέση με τη μέση τιμή, τότε **το ίδιο παρατηρείται** και στο διάγραμμα συσχέτισης



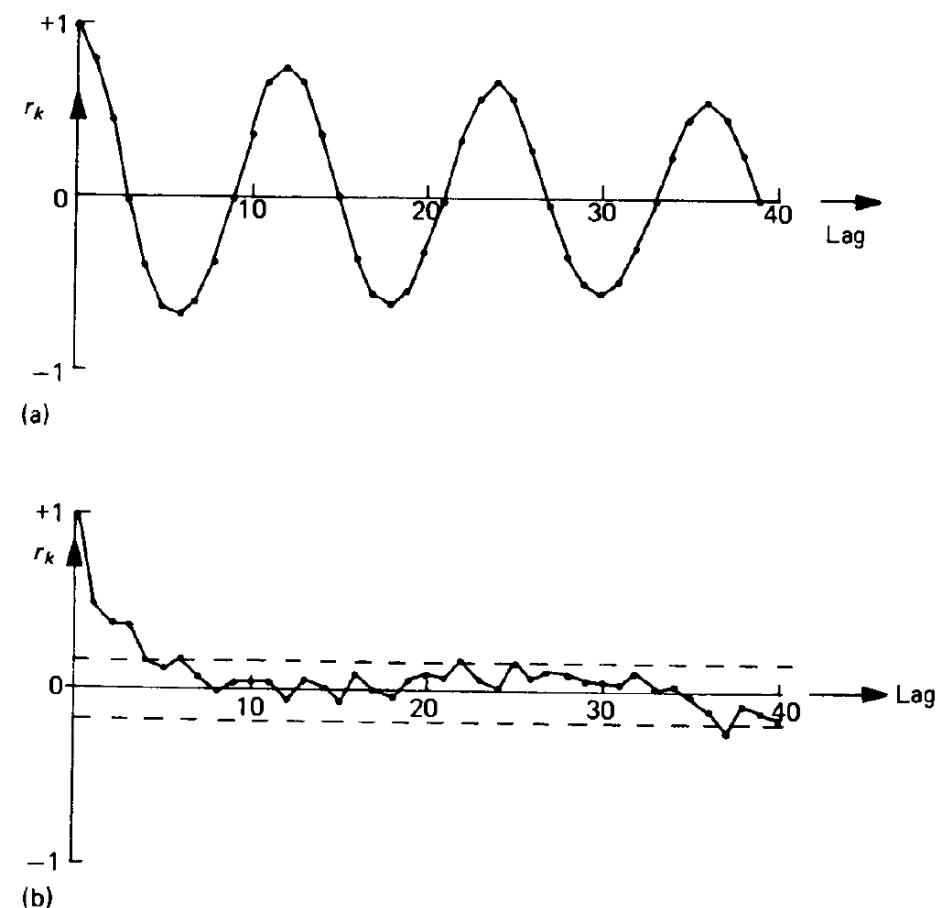
# Μη Στάσιμη Χρονοσειρά

- Εάν η χρονοσειρά περιέχει τάση, τότε οι τιμές  $r_k$  δε θα πλησιάζουν το μηδέν, παρά μόνο για υψηλές τιμές του  $k$
- Σε αυτή την περίπτωση το διάγραμμα δε βοηθάει πολύ
- Στην πραγματικότητα, η συνάρτηση αυτοσυσχέτισης έχει νόημα μόνο για στάσιμες χρονοσειρές



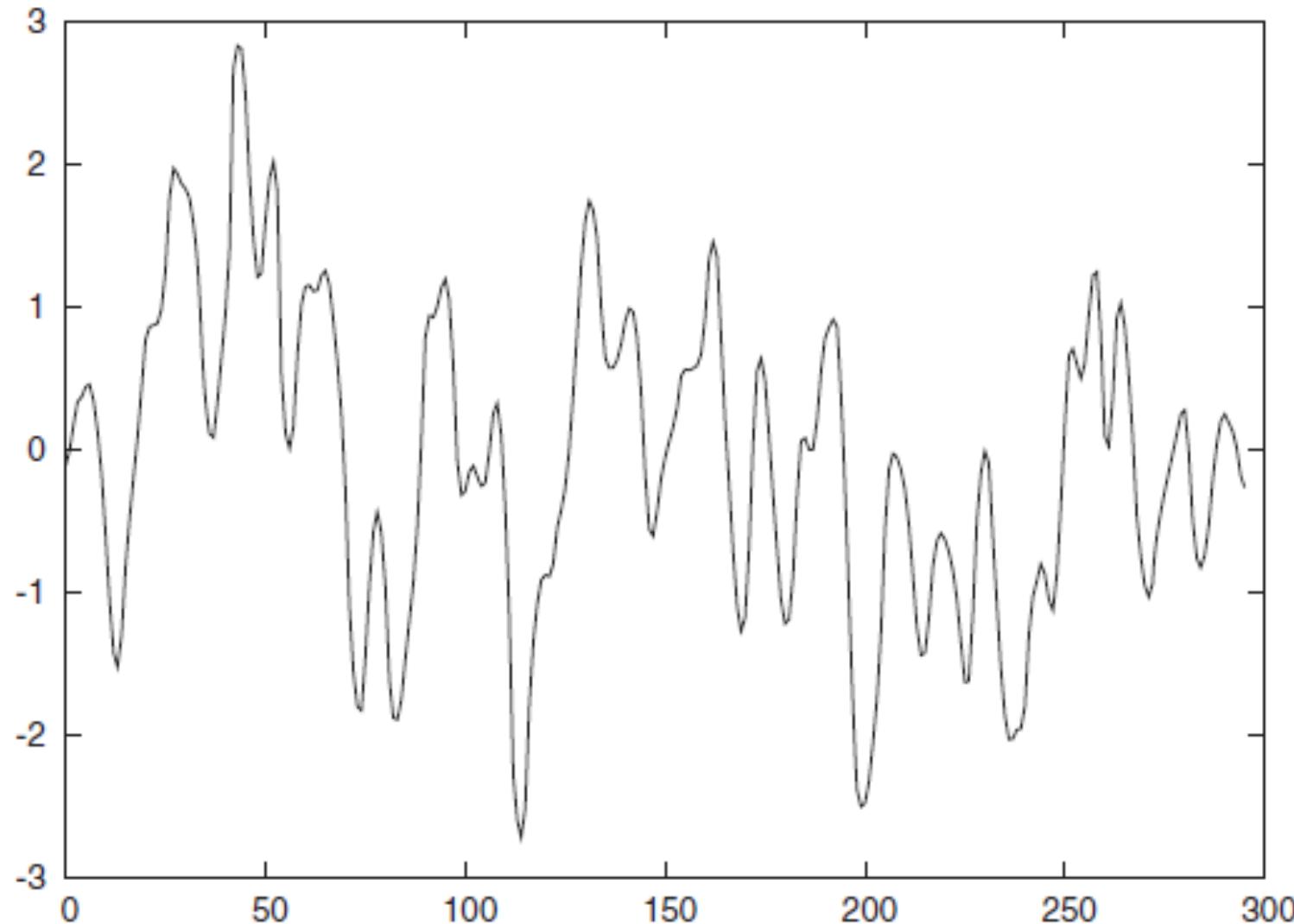
# Χρονοσειρά με Περιοδικότητα

- Εάν η χρονοσειρά περιέχει περιοδικότητα, τότε και το διάγραμμα συσχέτισης θα περιέχει μια ταλάντωση με ίδια συχνότητα
- Εφόσον αφαιρεθεί η περιοδικότητα, μπορούμε να βγάλουμε συμπεράσματα
- Π.χ. οι πρώτοι τρεις συντελεστές είναι αρκετά διαφορετικοί από το μηδέν, που υποδεικνύει **βραχεία συσχέτιση**

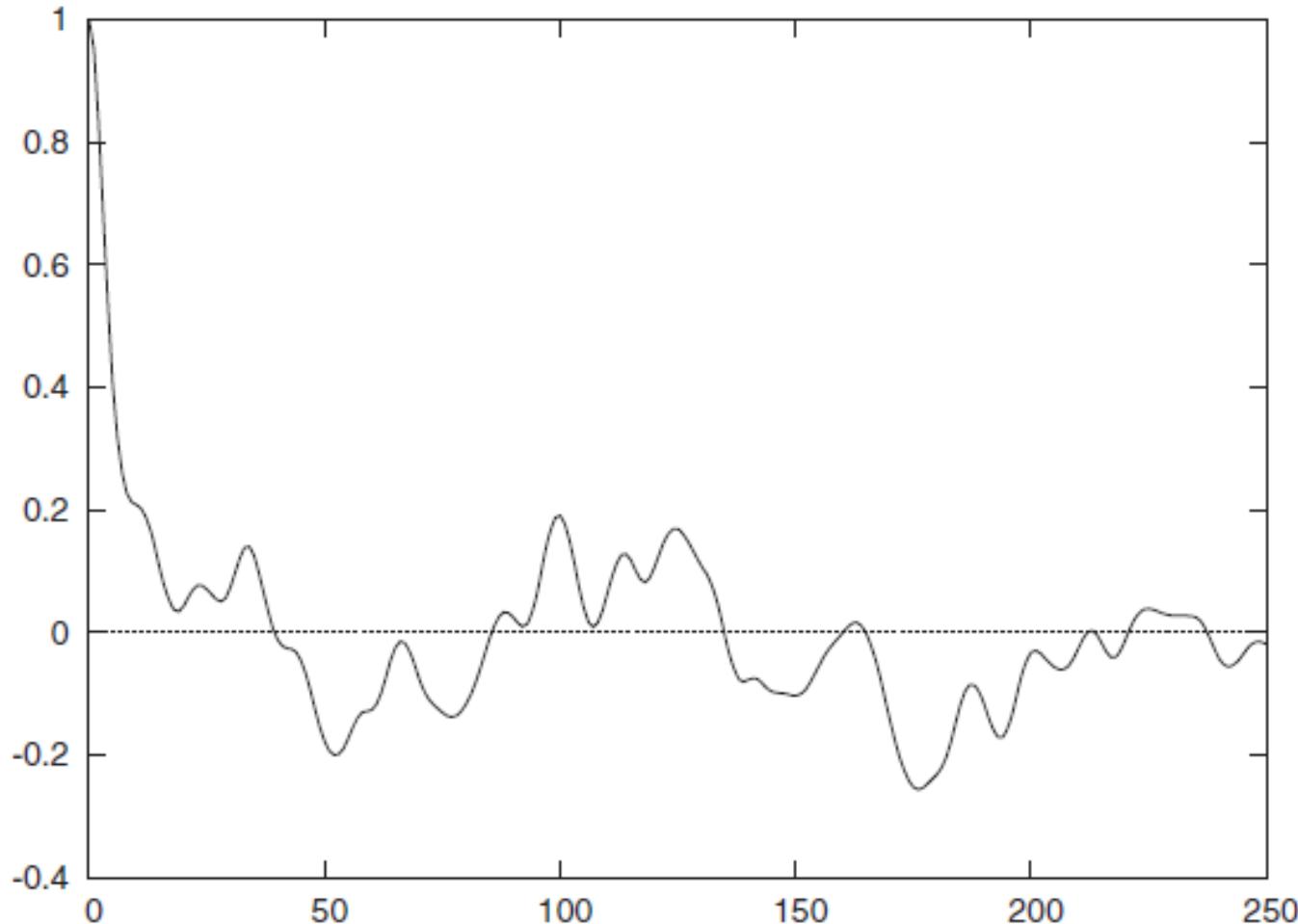


# Παράδειγμα #1: Το Σύνολο Δεδομένων

## Συγκέντρωσης Αερίων σε Εξάτμιση

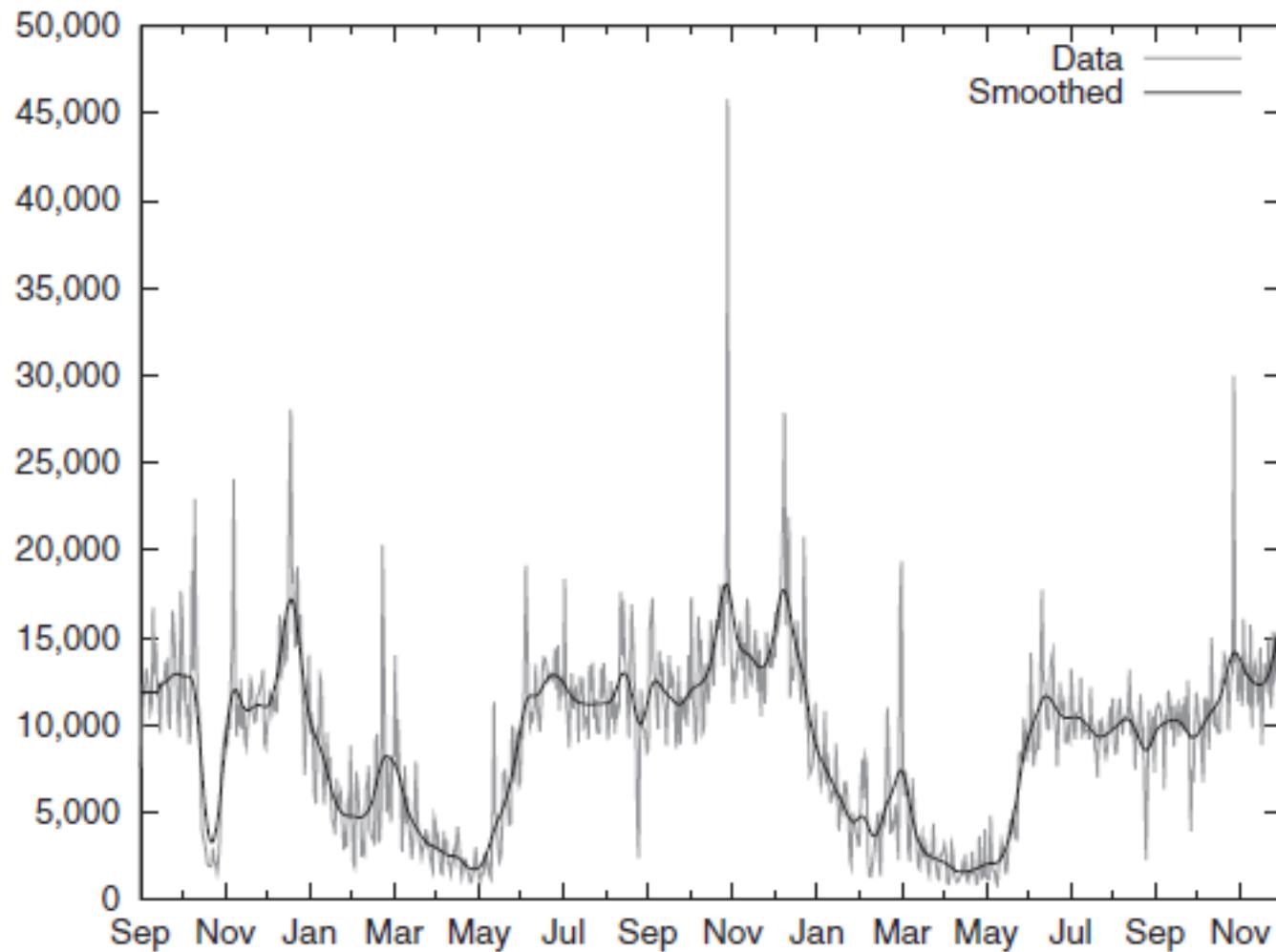


# Παράδειγμα #1: Η Συνάρτηση Συσχέτισης για το Ιδιο Σύνολο Δεδομένων

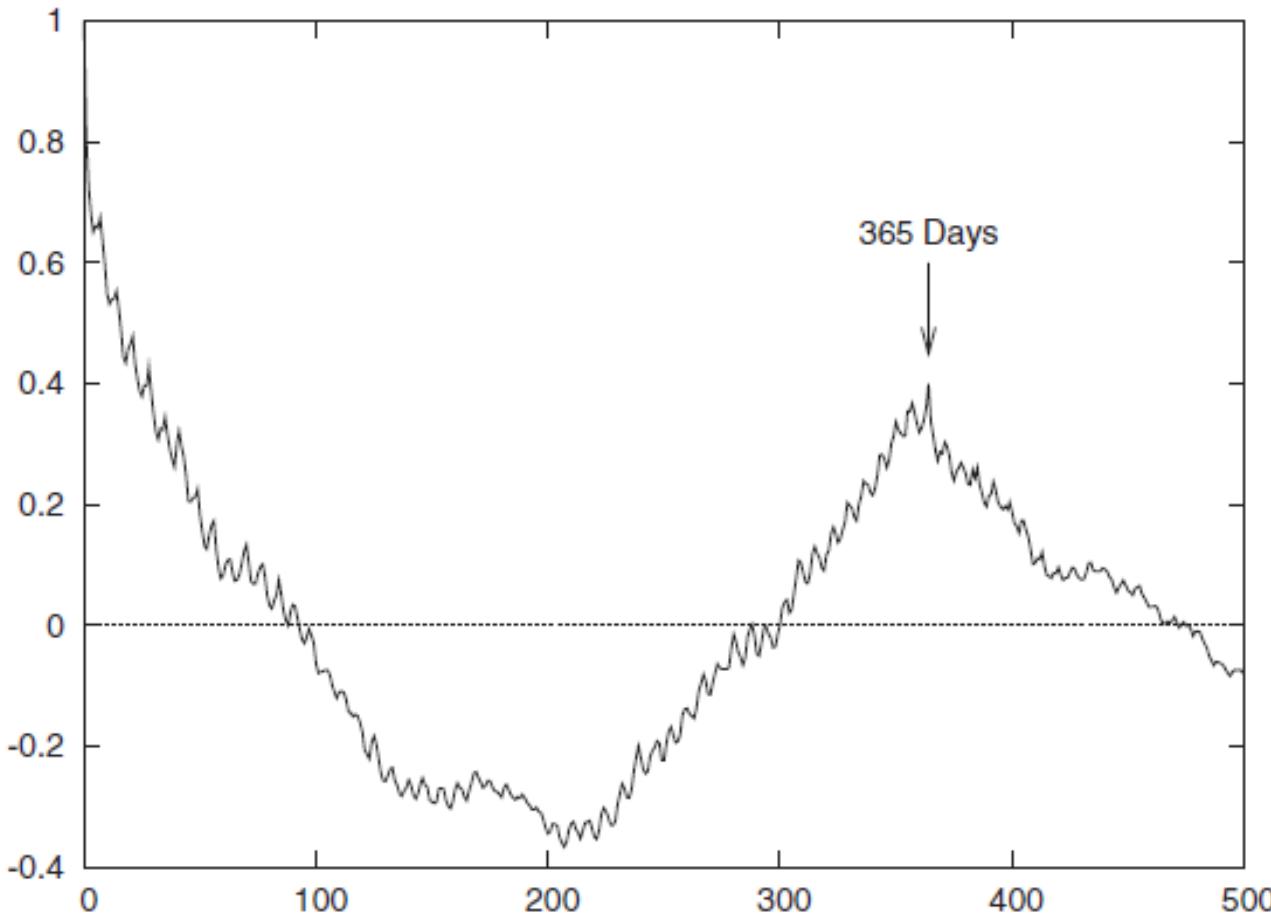


- Μια κλασική συνάρτηση συσχέτισης για χρονοσειρά που παρουσιάζει μόνο βραχεία συσχέτιση
- Η τιμή συσχέτισης πέφτει γρήγορα στο μηδέν, όχι όμως αμέσως
- Δεν υπάρχει περιοδικότητα, καθώς μετά την πρώτη μείωση δεν εμφανίζονται άλλες (ισχυρές) κορυφές

## Παράδειγμα #2: Το Σύνολο Δεδομένων του Τηλεφωνικού Κέντρου



# Παράδειγμα #2: Η Συνάρτηση Συσχέτισης για το ίδιο Σύνολο Δεδομένων



- Η χρονοσειρά έχει αρκετά μεγαλύτερη «μνήμη»
  - Χρειάζονται  $\sim 100$  μέρες για:  $r_k \rightarrow 0$
- Η δευτερεύουσα κορυφή εμφανίζεται με υστέρηση 365 ημερών
- Μικρή, αλλά επαναλαμβανόμενη, δομή (7 ημερών)
  - Δευτερεύουσα εποχικότητα με περίοδο 7 ημερών

# Άλλα Θέματα Προεπεξεργασίας Χρονοσειρών

## ■ Διαχείριση ελλιπών τιμών

- Με χρήση γραμμικής παρεμβολής
- Όστε να έχουμε συγχρονισμένες χρονοσειρές με τιμές σε ισαπέχουσες (χρονικά) θέσεις
- Τιμές  $y_i, y_j$  τις στιγμές  $t_i, t_j$ , και  $t \in (t_i, t_j)$

$$y = y_i + \left( \frac{t - t_i}{t_j - t_i} \right) \cdot (y_j - y_i)$$

## ■ Απομάκρυνση Θορύβου

- Binning: αντικατάσταση ενός συνόλου τιμών με τη μέση τιμή τους για ισομήκη διαστήματα  $[t_1, t_k], [t_{k+1}, t_{2k}], \dots$
- Γνωστή και ως Piecewise Aggregate Approximation (PAA)

# Άλλα Θέματα Προεπεξεργασίας Χρονοσειρών

## ■ Κανονικοποίηση

- Εύρους (range-based normalization)

$$y'_i = \frac{y_i - \min}{\max - \min}$$

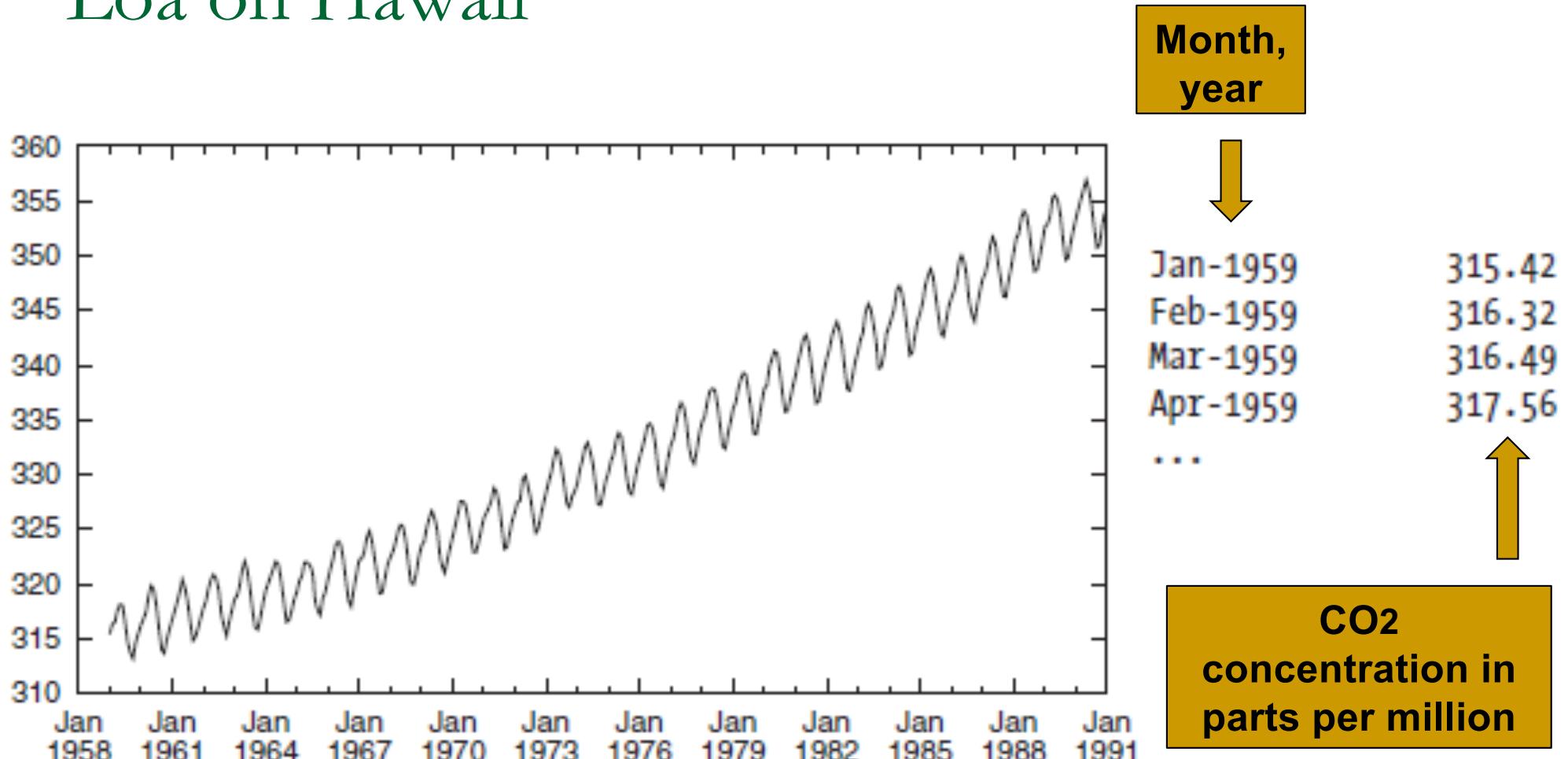
- Standardization

$$z_i = \frac{y_i - \mu}{\sigma}$$

# Παράρτημα:

## Ανάλυση Χρονοσειράς Βήμα προς Βήμα με το Gnuplot

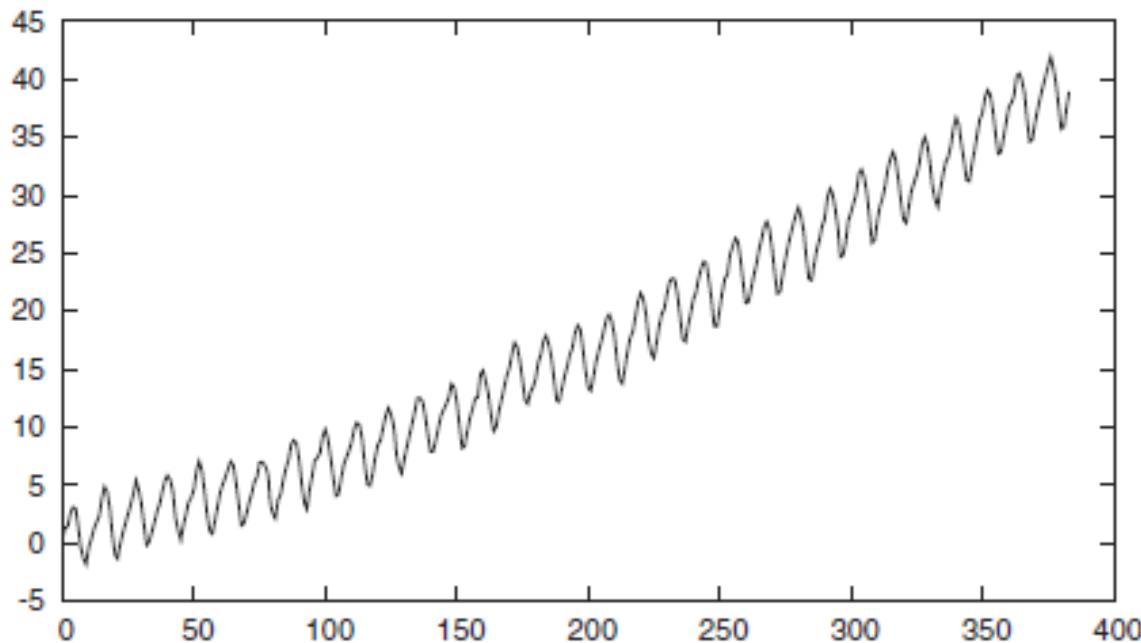
# Data set: CO<sub>2</sub> measurements above Mauna Loa on Hawaii



```
plot "data" u 2 w l
```

# Making the x values numeric and subtracting the constant vertical offset

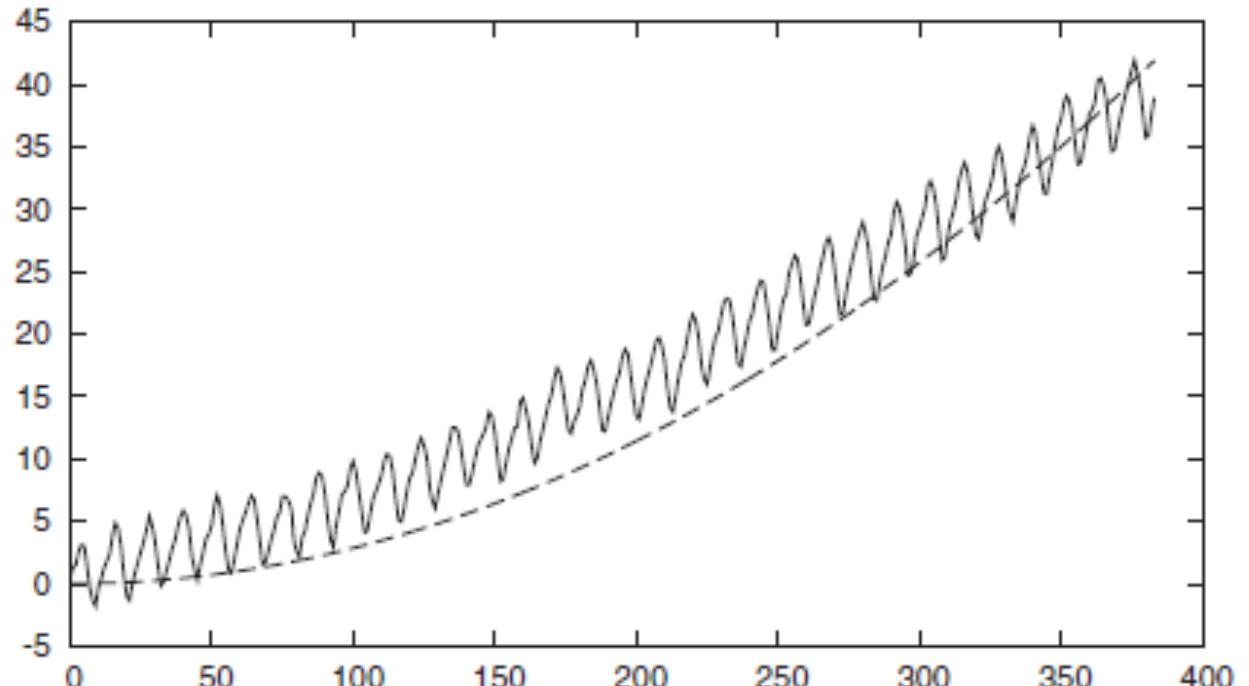
- subtracting the vertical offset from the data and expressing the horizontal position as the number of months since the first measurement



```
plot "data" u 0:(\$2-315) w l
```

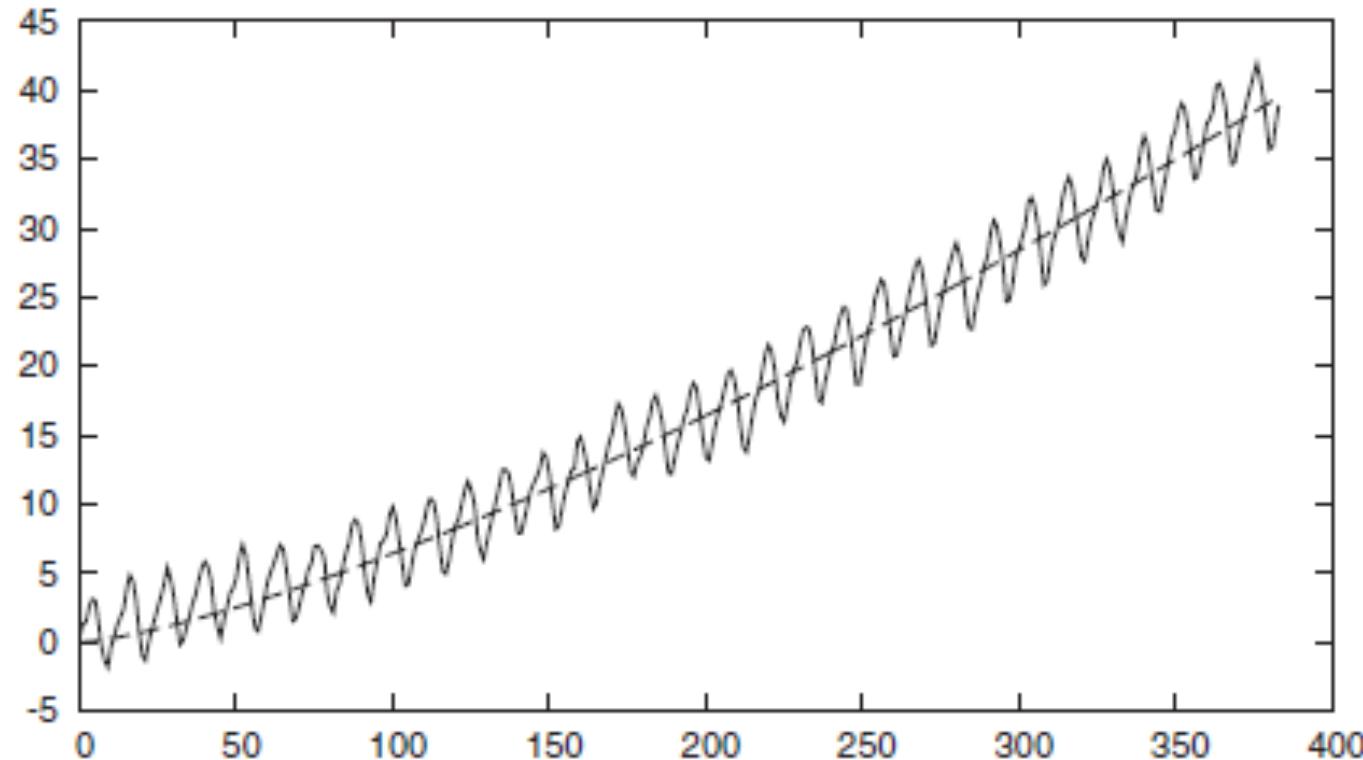
# Adding a Function

- The most predominant feature is the trend
  - First of all, the trend is nonlinear
  - If we ignore the short-term variation, the curve is convex downward
  - This suggests a power law with an as-yet-unknown exponent:  $x^k$
- To fix the upper-right corner, we need to rescale both axes: if  $x^k$  goes through  $(1, 1)$ , then  $b (x/a)^k$  goes through  $(a, b)$
- Try  $k=2$



```
plot "data" u 0:(\$2-315) w l, 35*(x/350)**2
```

# Getting the exponent right $f(x) = 35 \left(\frac{x}{350}\right)^{1.35}$



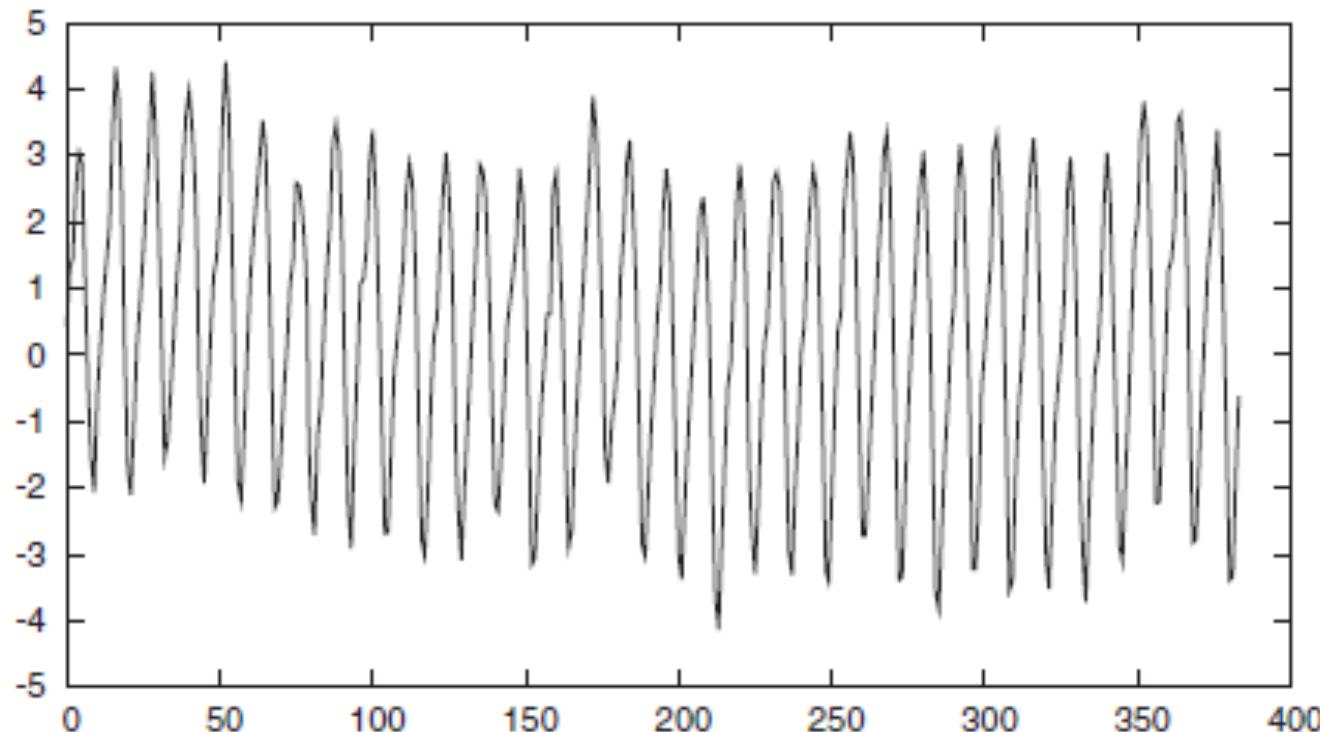
```
plot "data" u 0:(\$2-315) w l, 35*(x/350)**1.35
```

# Residuals

- The remainder when you subtract the smooth “trend” from the actual data
- Residuals should be balanced: **symmetrically distributed around zero**
- Residuals should be **free of a trend**
  - The presence of a trend or of any other large-scale systematic behavior in the residuals suggests that the model is inappropriate!
- Residuals will **necessarily straddle the zero value**; they will take on both positive and negative values

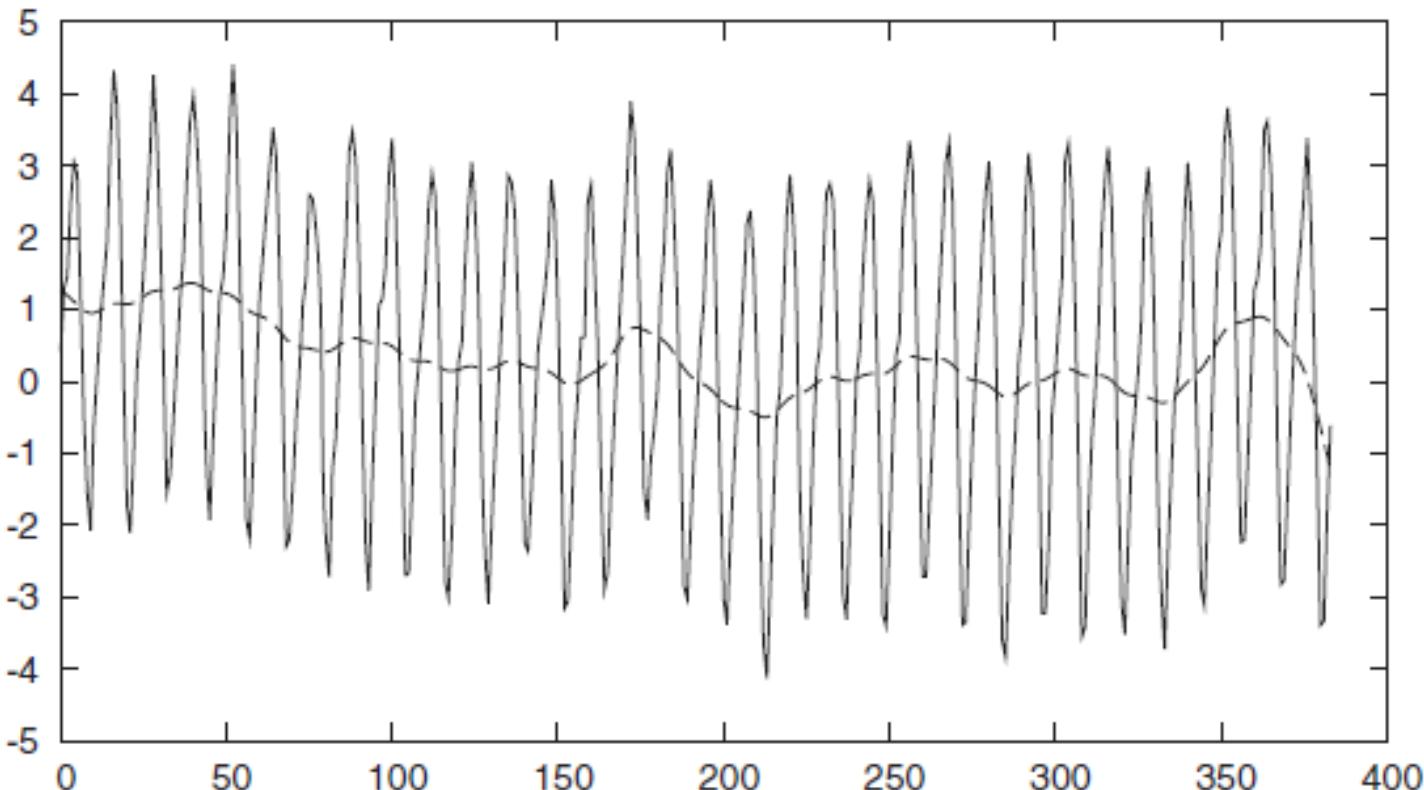
# The residual, after subtracting the function from the data.

- If our guess for the trend is correct, then the residual should not exhibit any trend itself—it should just straddle  $y = 0$  in a balanced fashion



```
plot "data" u 0:(\$2-315 - 35*(\$0/350)**1.35) w l
```

# Weighted spline approximation



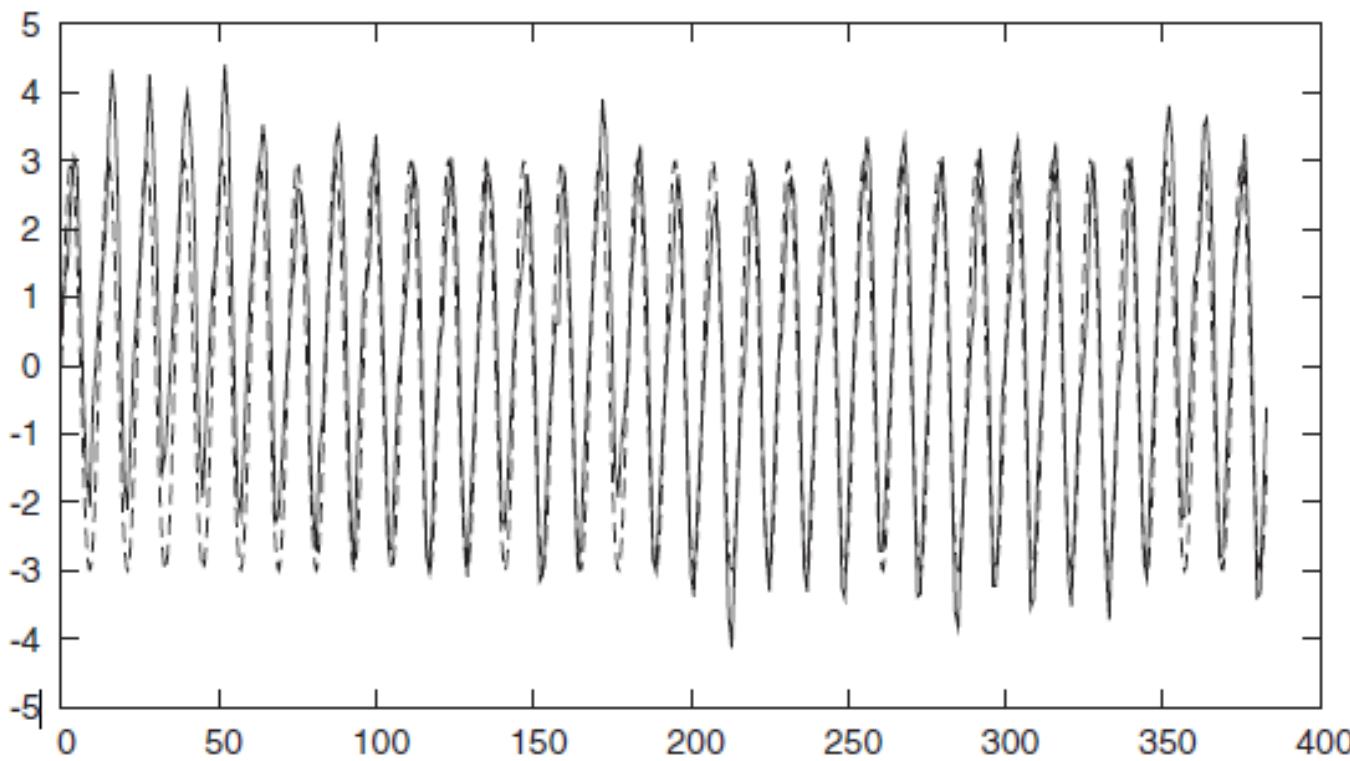
- Plotting a smoothed version of the residual together with the unsmoothed residual to test whether there is any systematic trend remaining in the residual

```
plot "data" u 0:(2 - 315 - 35 * (0/350)**1.35) w l, \
      "" u 0:($2-315 - 35*($0/350)**1.35):(0.001) s acs w l
```

$$f(x) = 315 + 35 \cdot (x/350)^{1.35}$$

```
plot "data" u 0:($2-f($0)) w l, "" u 0:($2-f($0)):(0.001) s acs w l
```

# Seasonality

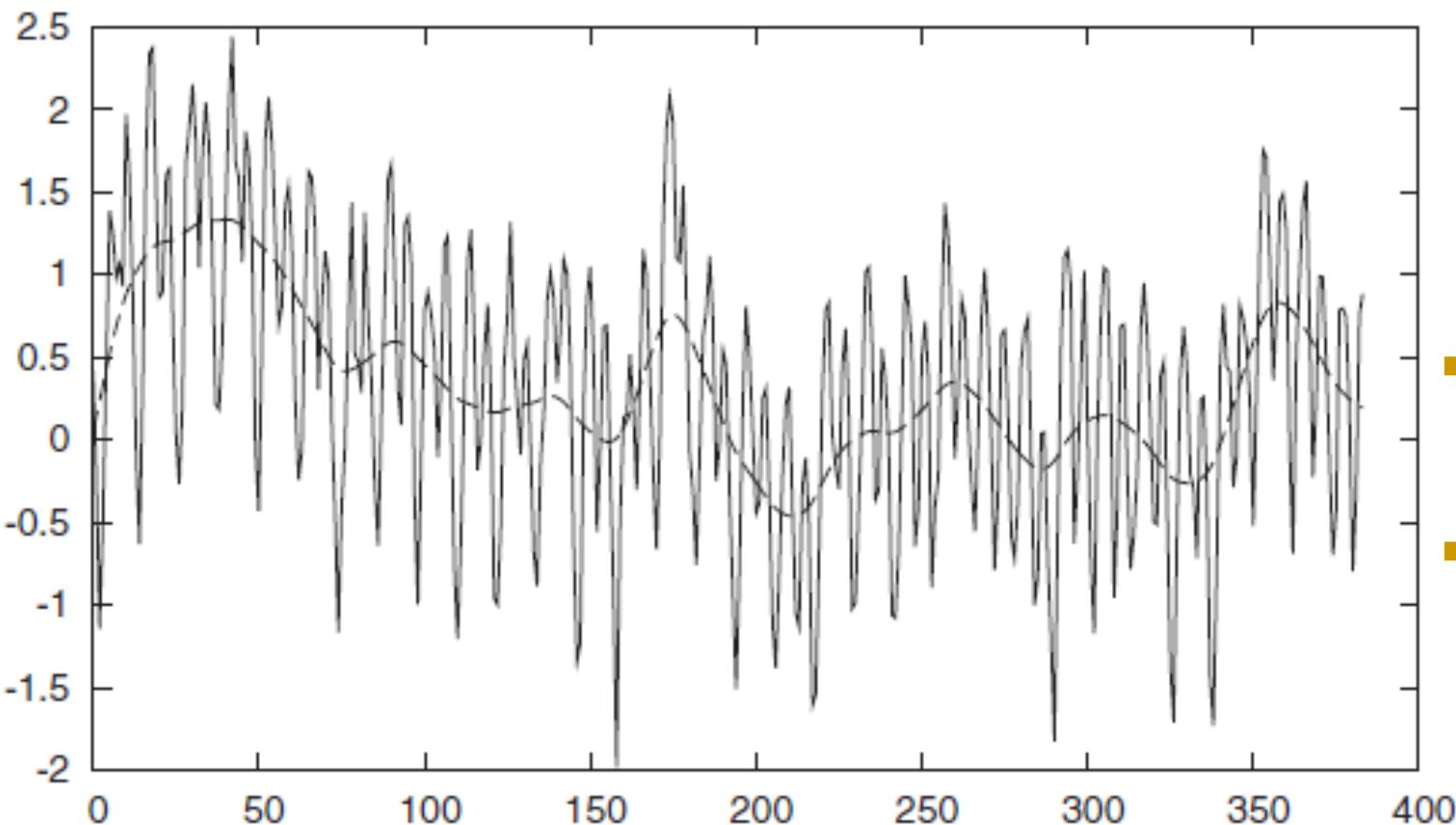


- regular oscillations  
→ sines and cosines
- data has been taken on a monthly basis—perhaps there is a year-over-year periodicity → one sine period corresponds to 12 points
- for the amplitude, the graph suggests a value close to 3

$$3 \sin\left(2\pi \frac{x}{12}\right)$$

```
plot "data" u 0:(\$2-f(\$0)) w l, 3*sin(2*pi*x/12) w l
```

# Residuals after subtracting both trend and seasonality

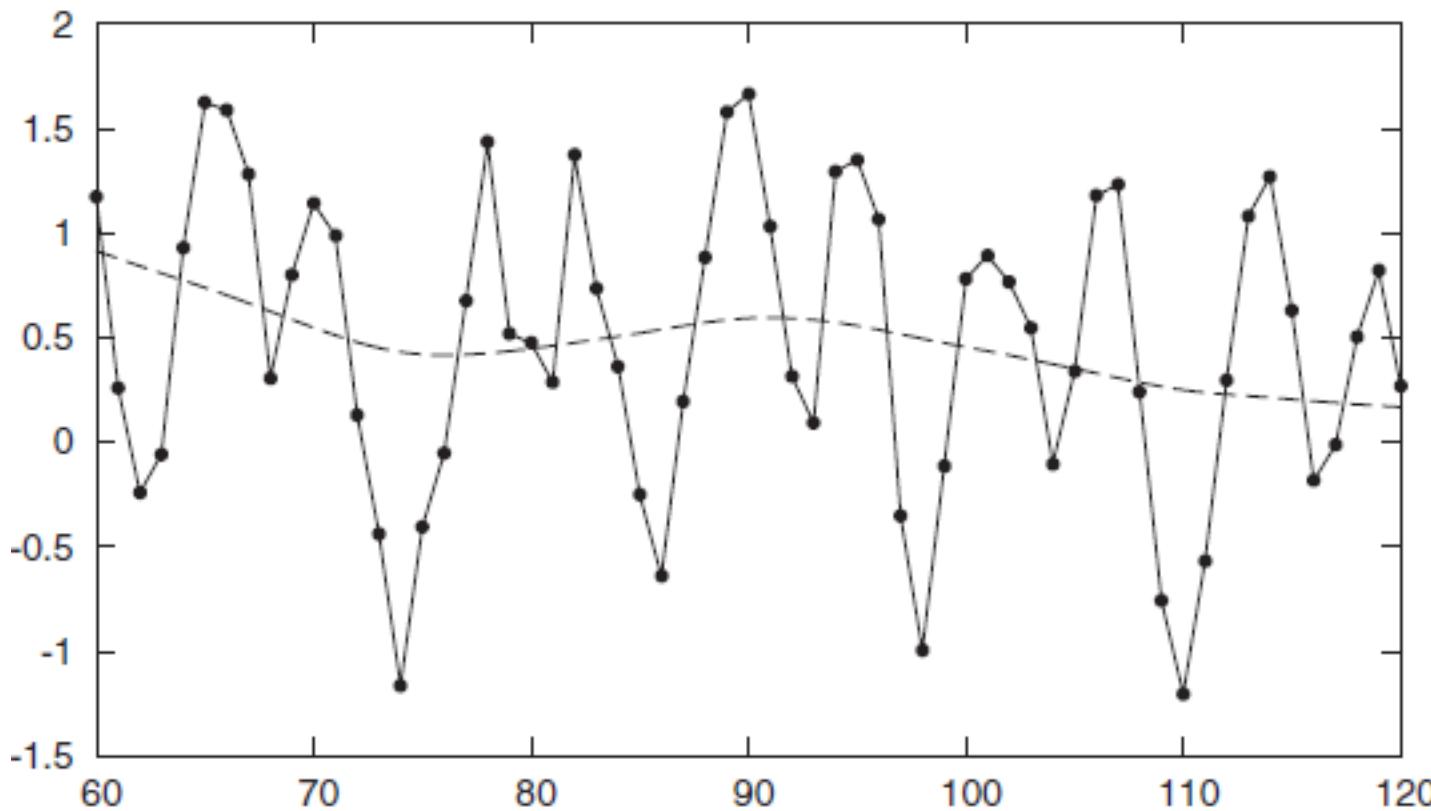


- There is clearly some regularity remaining in the data, although at a higher frequency than the main seasonality
- Let's zoom in on a smaller interval of the data
- Data in [60:120] appears particularly regular, so let's look there

```
f(x) = 315 + 35*(x/350)**1.35 + 3*sin(2*pi*x/12)
plot "data" u 0:($2-f($0)) w l, "" u 0:($2-f($0)):(0.001) s acs w l
```

# Zooming in for a closer look at [60,120]

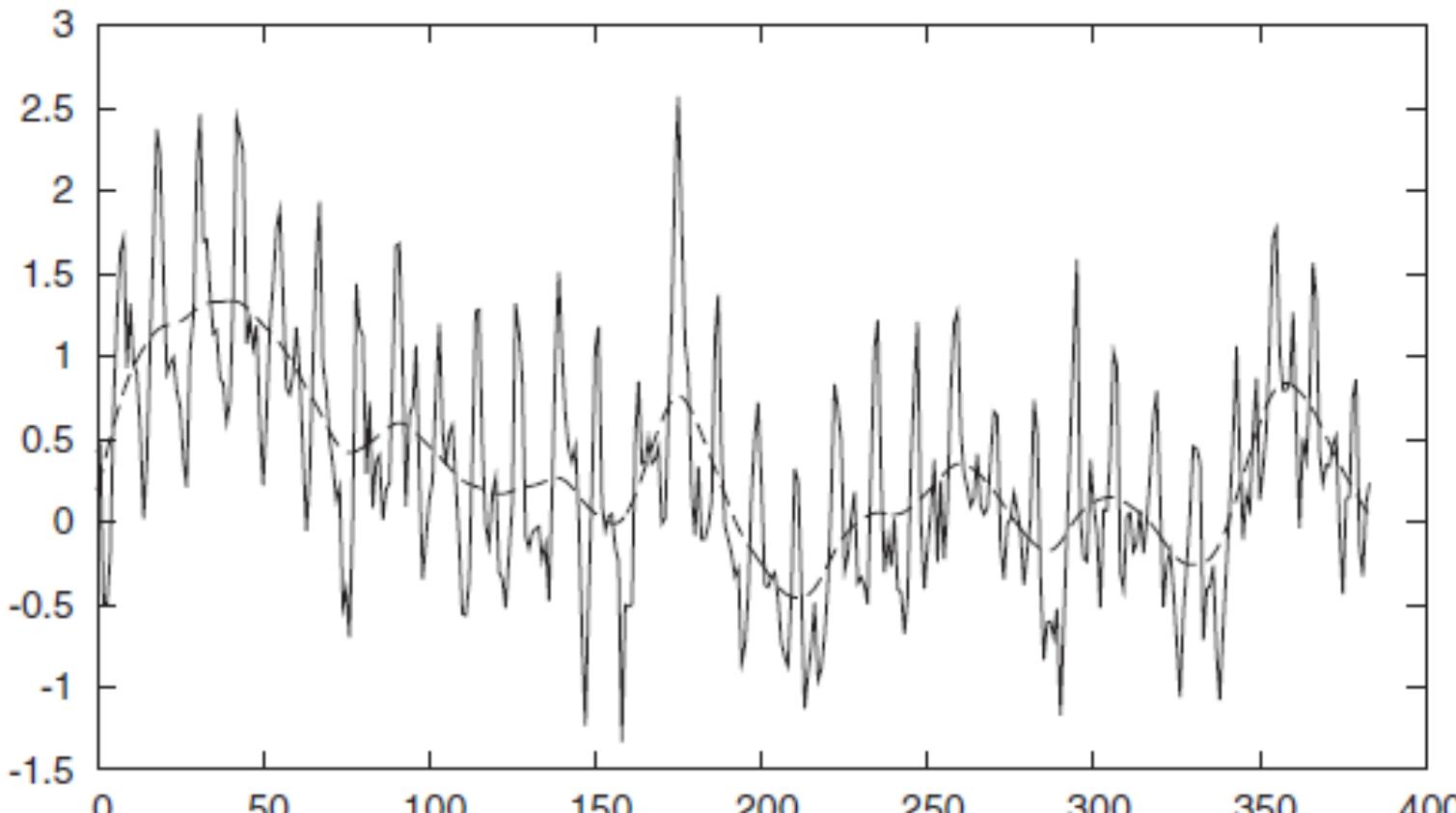
## Individual data points are marked by symbols



- But it seems that between any two primary valleys there is exactly one secondary valley
- 7 months rising, 5 months falling
- This kind of asymmetry implies that the seasonality cannot be represented by a simple sine wave
- We have to take into account higher harmonics

```
plot [60:120] "data" u 0:(\$2-f(\$0)) w lp,  
"" u 0:(\$2-f(\$0)):(0.001) s acs w l
```

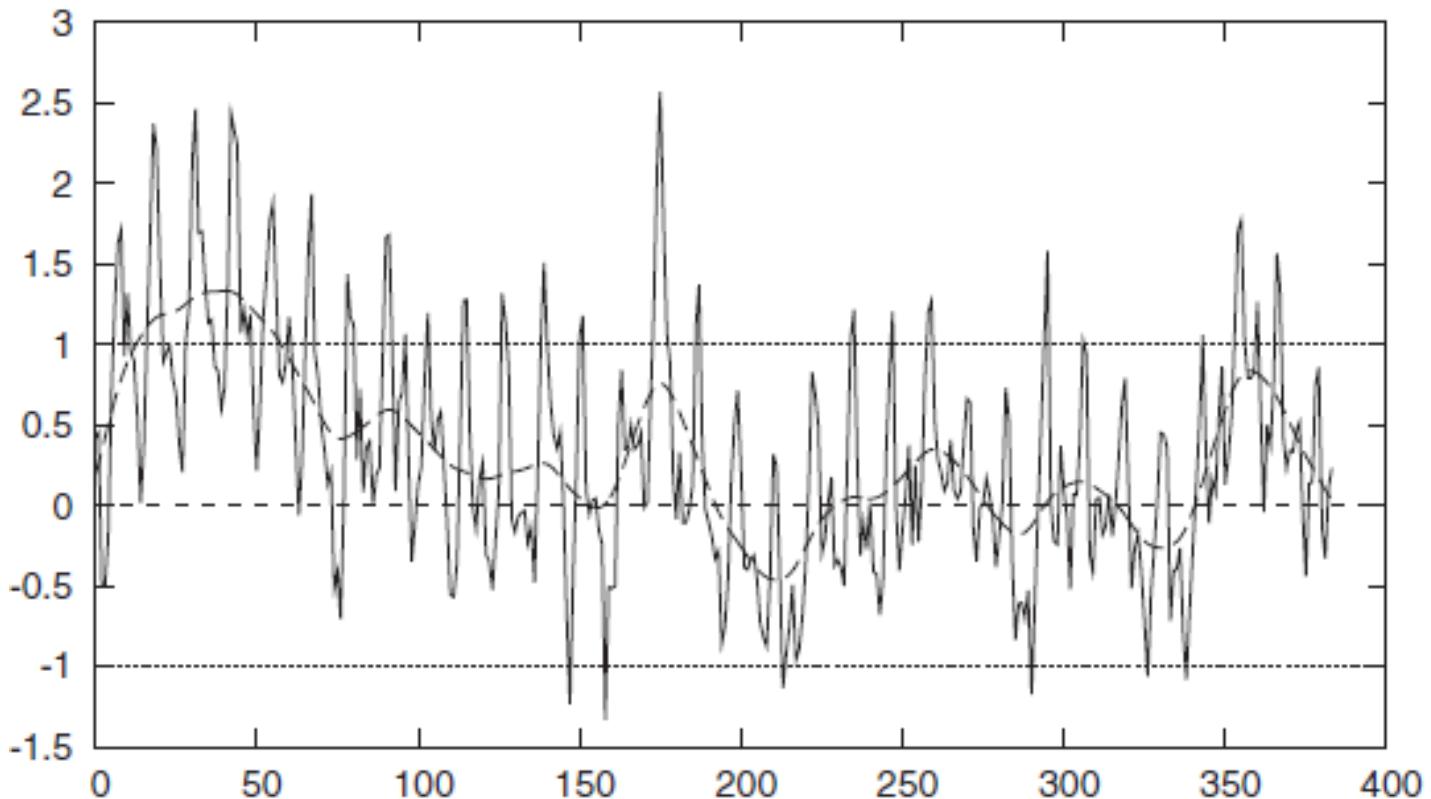
# Residual after removing trend and the first and second harmonic of the seasonality



- Now we are really pretty close
- Look at the residual—in particular, for values of  $x > 150$
- The data starts to look quite “random,” although there is some systematic behavior for  $x$  in the range [0:70] that we don’t really capture

```
f(x) = 315 + 35*(x/350)**1.35 + 3*sin(2*pi*x/12) - 0.75*sin(2*pi*$0/6)
plot "data" u 0:($2-f($0)) w l, "" u 0:($2-f($0)):(0.001) s acs w l
```

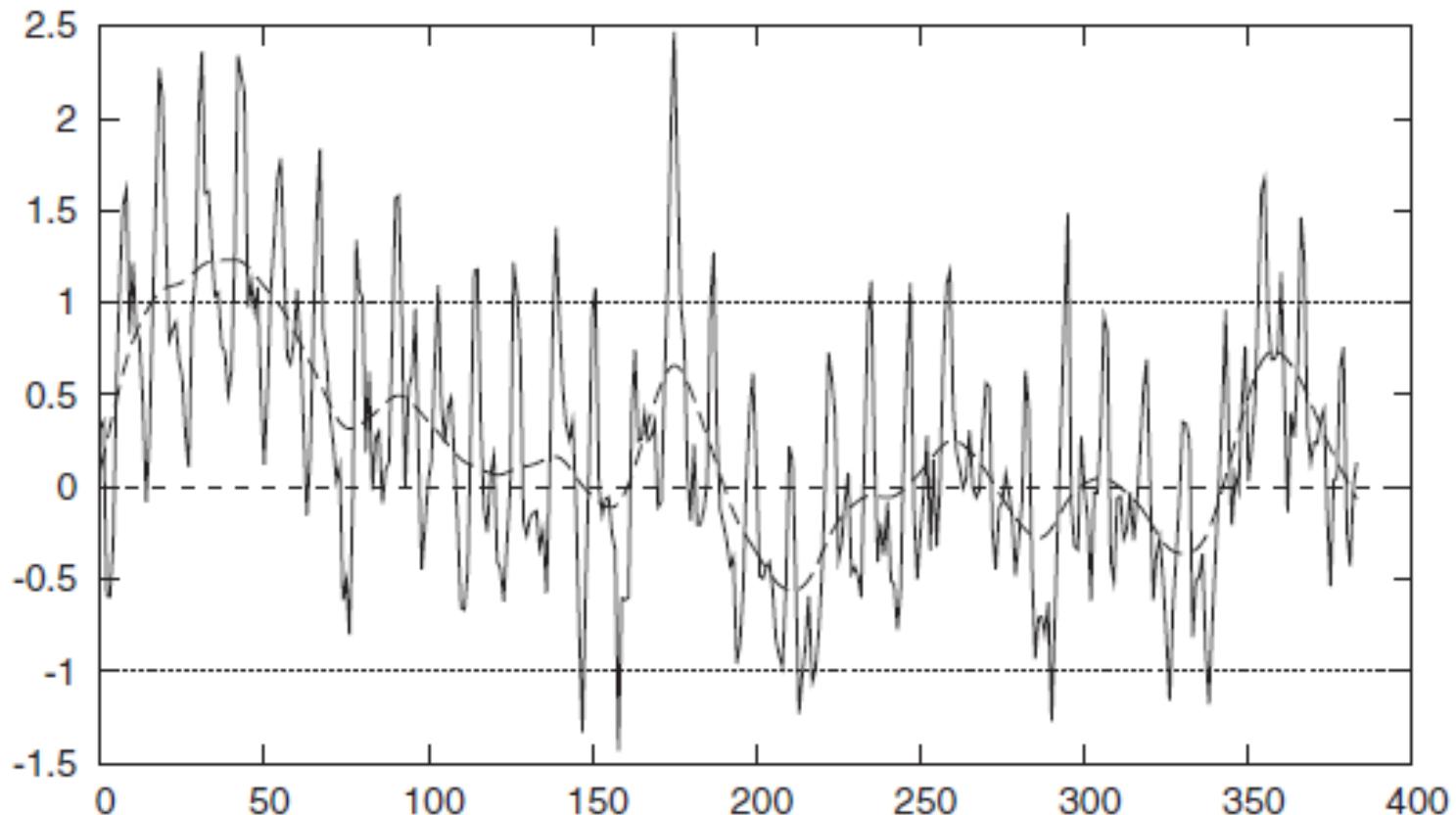
# Adding some grid lines for comparison



- It looks as if the residual is skewed toward positive values, so let's adjust the vertical offset by 0.1

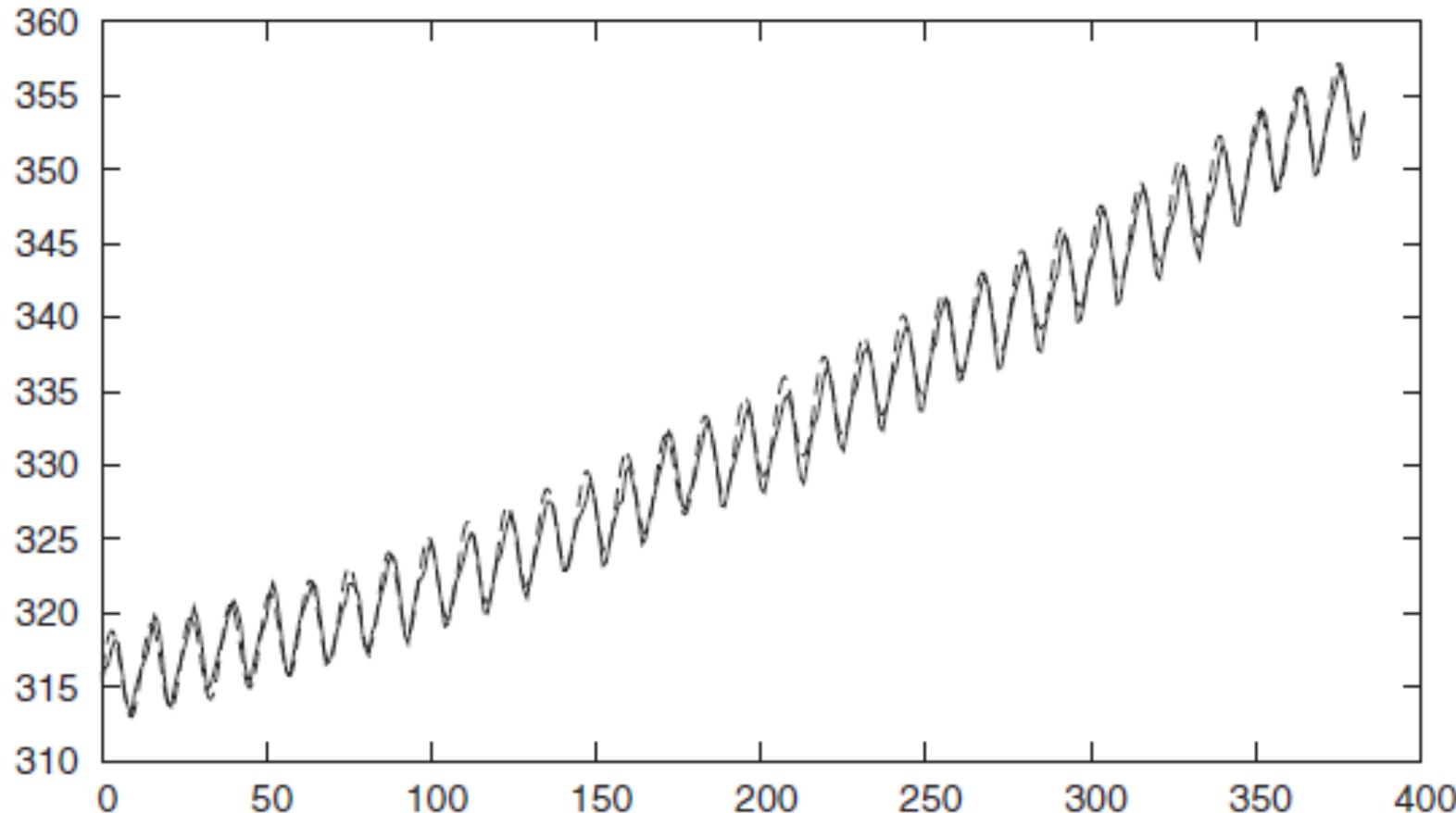
```
plot "data" u 0:(\$2-f(\$0)) w l, "" u 0:(\$2-f(\$0):(0.001) s acs w l, 0, 1, -1
```

# The final residual



```
f(x) = 315 + 35*(x/350)**1.35 + 3*sin(2*pi*x/12) - 0.75*sin(2*pi*$0/6) + 0.1  
plot "data" u 0:($2-f($0)) w l, "" u 0:($2-f($0)):(0.001) s acs w l, 0, 1, -1
```

# Raw data and final fit

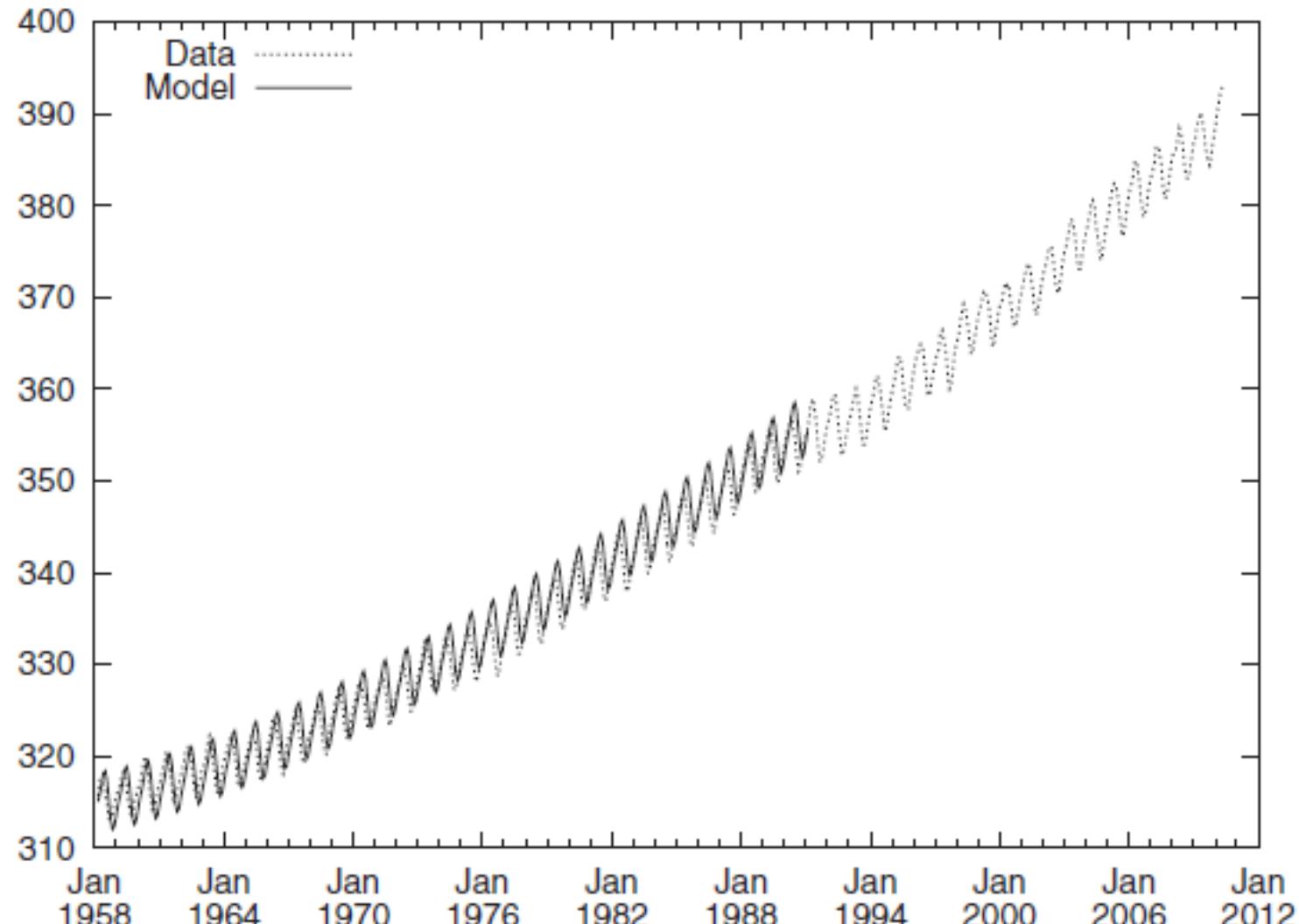


```
f(x) = 315 + 35*(x/350)**1.35 + 3*sin(2*pi*x/12) - 0.75*sin(2*pi*$0/6) + 0.1  
plot "data" u 0:2 w l, f(x)
```

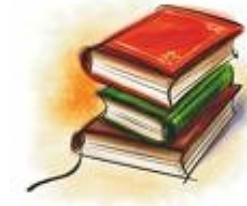
# Take-away

- We started out with nothing—no idea at all of what the data looked like
  - Then, layer by layer, we peeled off components of the data, until only random noise remained
  - We ended up with an explicit, analytical formula that describes the data remarkably well
- We did so entirely “manually”
  - No black-box fitting routine
- We learned *more* by doing this work manually than if we had used a fitting routine

# The extended data set up to early 2010 together with the model (up to 1990)



# Πηγές Αναφοράς



- *Data Analysis with Open Source Tools, by Philipp K. Janert.*  
*Copyright 2011 Philipp K. Janert, 978-0-596-80235-6.*
  - Κεφάλαια 4 και 6
- *Data Mining: The Textbook, by Charu C. Aggarwal (2015) Springer*
  - Κεφάλαιο 14
- *The Analysis of Time-Series: An Introduction (2003), by Chris Chatfield, 6<sup>th</sup> edition, Chapman & Hall, CRC*



## 4. Εισαγωγή στην Προγνωστική Μοντελοποίηση – Δέντρα Απόφασης



---

**Ανάλυση Δεδομένων  
(Data Analytics)**

Χρήστος Δουλκερίδης  
2024-25

# Περίγραμμα Μαθήματος

- **Εισαγωγή στην προγνωστική μοντελοποίηση**
- Επιλογή γνωρισμάτων (**feature selection**)
  - Εντροπία (**entropy**), κέρδος πληροφορίας (**information gain**)
- Δέντρα απόφασης (**decision trees**)
- Άλλα μέτρα επιλογής γνωρισμάτων
  - **Gini**, αναλογία κέρδους (**gain ratio**)

# Ορολογία

- **Μοντέλο**: μια απλουστευμένη αναπαράσταση της πραγματικότητας που δημιουργείται για να εξυπηρετεί έναν σκοπό
- **Πρόβλεψη**: η εκτίμηση μιας άγνωστης τιμής
- **Εποπτευόμενη μάθηση (supervised learning)**: η δημιουργία ενός **μοντέλου** που περιγράφει μια σχέση μεταξύ **επιλεγμένων μεταβλητών** (γνωρισμάτων) και μιας προκαθορισμένης **μεταβλητής** (ενός γνωρίσματος) **στόχου**
- **Επαγωγή μοντέλου (model induction)**: η δημιουργία **μοντέλου** από δεδομένα
- **Αλγόριθμος επαγωγής (induction algorithm)**: η διαδικασία με την οποία δημιουργείται το **μοντέλο** από τα δεδομένα
- **Δεδομένα εκπαίδευσης (training data)**: τα δεδομένα εισόδου για τον **αλγόριθμο επαγωγής**

# Ορολογία

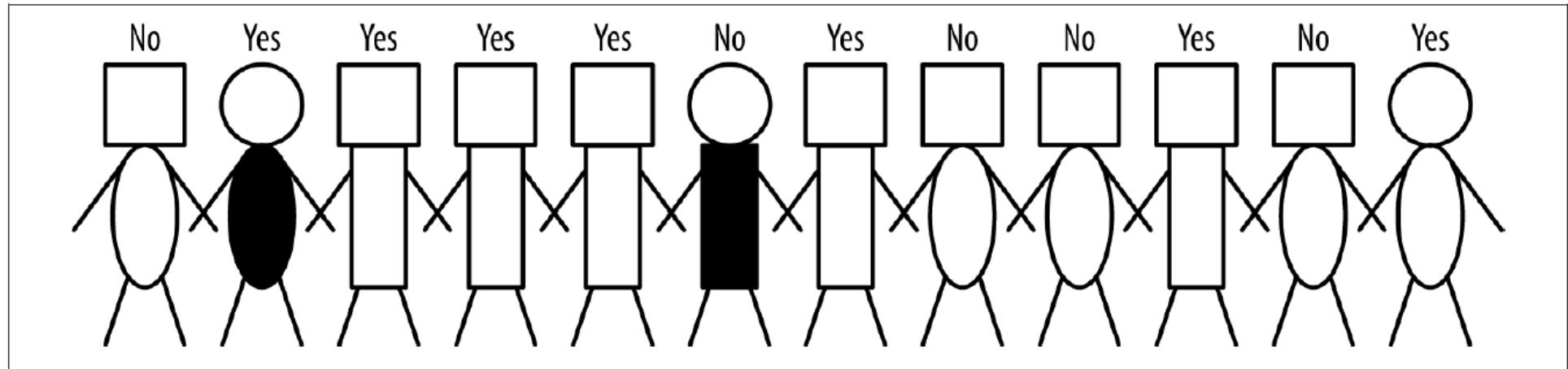
Name	Balance	Age	Employed	Write-off
Mike	\$200,000	42	no	yes
Mary	\$35,000	33	yes	no
Claudio	\$115,000	40	no	no
Robert	\$29,000	23	yes	yes
Dora	\$72,000	31	no	no

Αυτή είναι μια γραμμή (παράδειγμα)

Διάνυσμα χαρακτηριστικών: <Cladio, \$115.000, 40, no>

Ετικέτα κατηγορίας (τιμή γνωρίσματος-στόχου) είναι: no

# Παράδειγμα Κατηγοριοποίησης/Classification



## ■ Γνωρίσματα/Μεταβλητές:

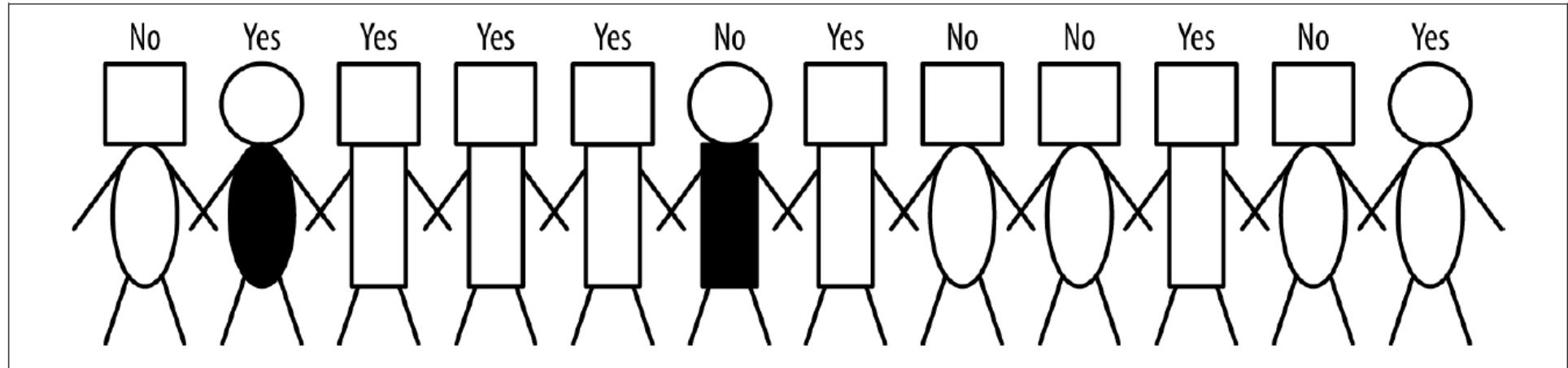
- **Σχήμα κεφαλιού:** {τετράγωνο, στρογγυλό}
- **Σχήμα σώματος:** {ορθογώνιο, ωοειδές}
- **Χρώμα σώματος:** {σκούρο, ανοικτό}

## ■ Γνώρισμα/Μεταβλητή – στόχος:

- **Παραγραφή χρέους:** {ναι, όχι}

Ποια από αυτά τα γνωρίσματα θα ήταν τα καταλληλότερα για να **χωρίσουμε** τους ανθρώπους **σε ομάδες**, με τρόπο που να επιτρέπει να διακρίνουμε ποιοι θα αθετήσουν την υποχρέωση αποπληρωμής από αυτούς που δε θα το κάνουν;

# Παράδειγμα Κατηγοριοποίησης (συνέχ.)



- Στόχος: οι ομάδες που θα προκύψουν να είναι όσο **πιο ομοιογενείς** γίνεται **σε σχέση με τη μεταβλητή-στόχο**
  - Αν **κάθε** μέλος της ομάδας έχει την **ίδια τιμή** για τη μεταβλητή-στόχο, τότε η ομάδα είναι **ομοιογενής**
- Σε πραγματικά δεδομένα, σπανίως ένα γνώρισμα θα παράγει ομοιογενή τμήματα
  - Άρα ο στόχος είναι: **(α) να μειώσουμε την ανομοιογένεια**, και **(β) να χρησιμοποιήσουμε το γνώρισμα σε ένα προγνωστικό μοντέλο**

# Περίγραμμα Μαθήματος

- Εισαγωγή στην προγνωστική μοντελοποίηση
- Επιλογή γνωρισμάτων (**feature selection**)
  - Εντροπία (**entropy**), κέρδος πληροφορίας (**information gain**)
- Δέντρα απόφασης (**decision trees**)
- Άλλα μέτρα επιλογής γνωρισμάτων
  - **Gini**, αναλογία κέρδους (**gain ratio**)

# Επιλογή Γνωρισμάτων (Feature Selection)

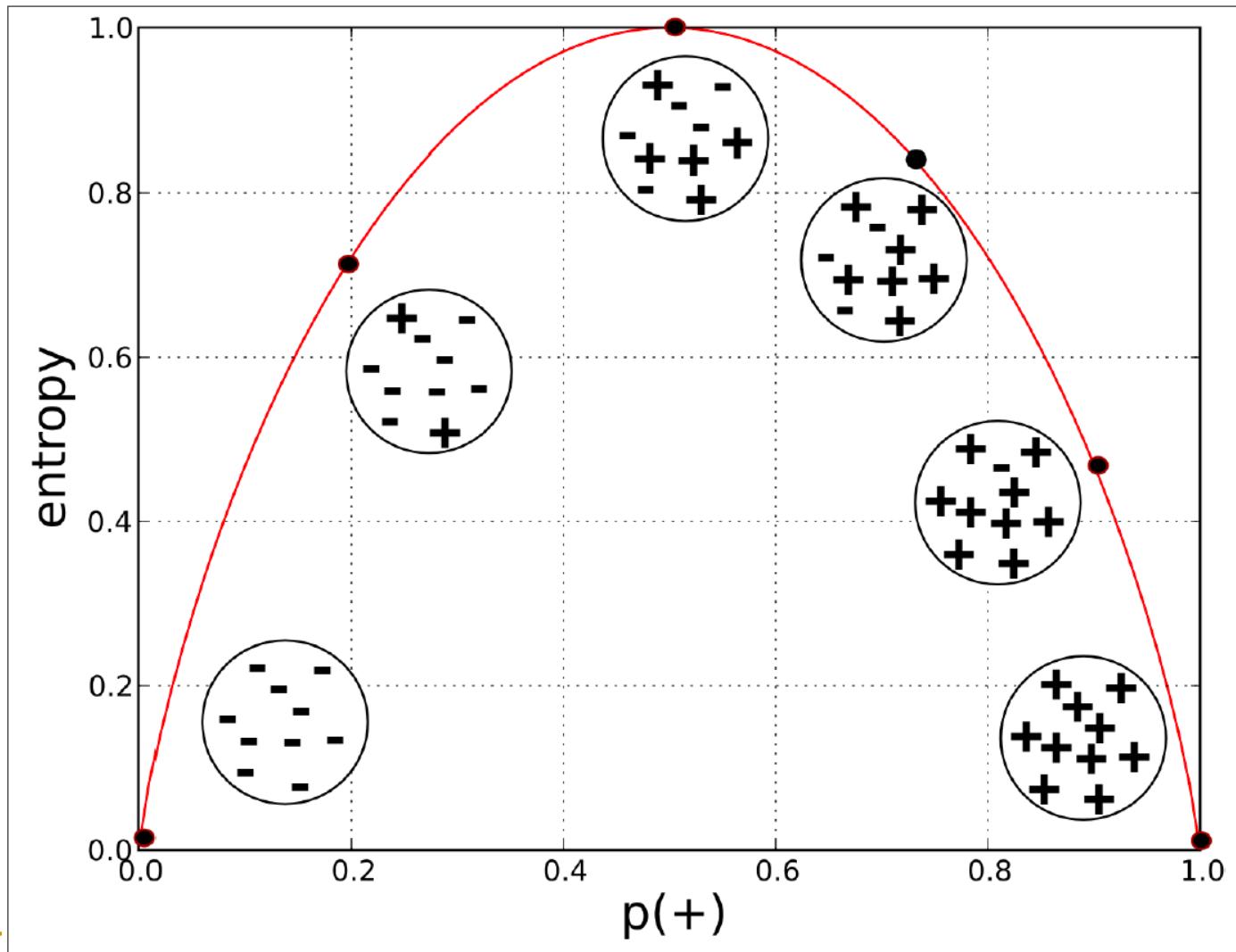
- Μέτρο ομοιογένειας (purity measure):
  - Ένας τύπος που αξιολογεί πόσο καλά κάθε γνώρισμα χωρίζει ένα σύνολο παραδειγμάτων σε τμήματα
- Κέρδος πληροφορίας (information gain):
  - Το πιο συνηθισμένο κριτήριο διαχωρισμού
  - Βασίζεται σε ένα μέτρο που ονομάζεται εντροπία (entropy) [Shannon, 1948]

$$\text{εντροπία} = - p_1 \log_2(p_1) - p_2 \log_2(p_2) - \dots$$

- $p_i$ : η πιθανότητα της ιδιότητας  $i$  μέσα στο σύνολο
- Η ιδιότητα αντιστοιχεί σε τιμή της μεταβλητής-στόχου

# Εντροπία ενός Συνόλου Κατηγοριών ως Συνάρτηση του $p(+)$

10 στιγμιότυπα δύο κλάσεων: + και -



# Παράδειγμα Εντροπίας

- Ας υποθέσουμε 10 ανθρώπους:
  - 7 στην κατηγορία **non-write-off** (μη παραγραφή χρέους)
  - 3 στην κατηγορία **write-off** (παραγραφή χρέους)

$$p(\text{non-write-off}) = 7 / 10 = 0.7$$

$$p(\text{write-off}) = 3 / 10 = 0.3$$

$$\begin{aligned} \text{entropy}(S) &= -[0.7 \times \log_2 (0.7) + 0.3 \times \log_2 (0.3)] \\ &\approx -[0.7 \times -0.51 + 0.3 \times -1.74] \\ &\approx 0.88 \end{aligned}$$

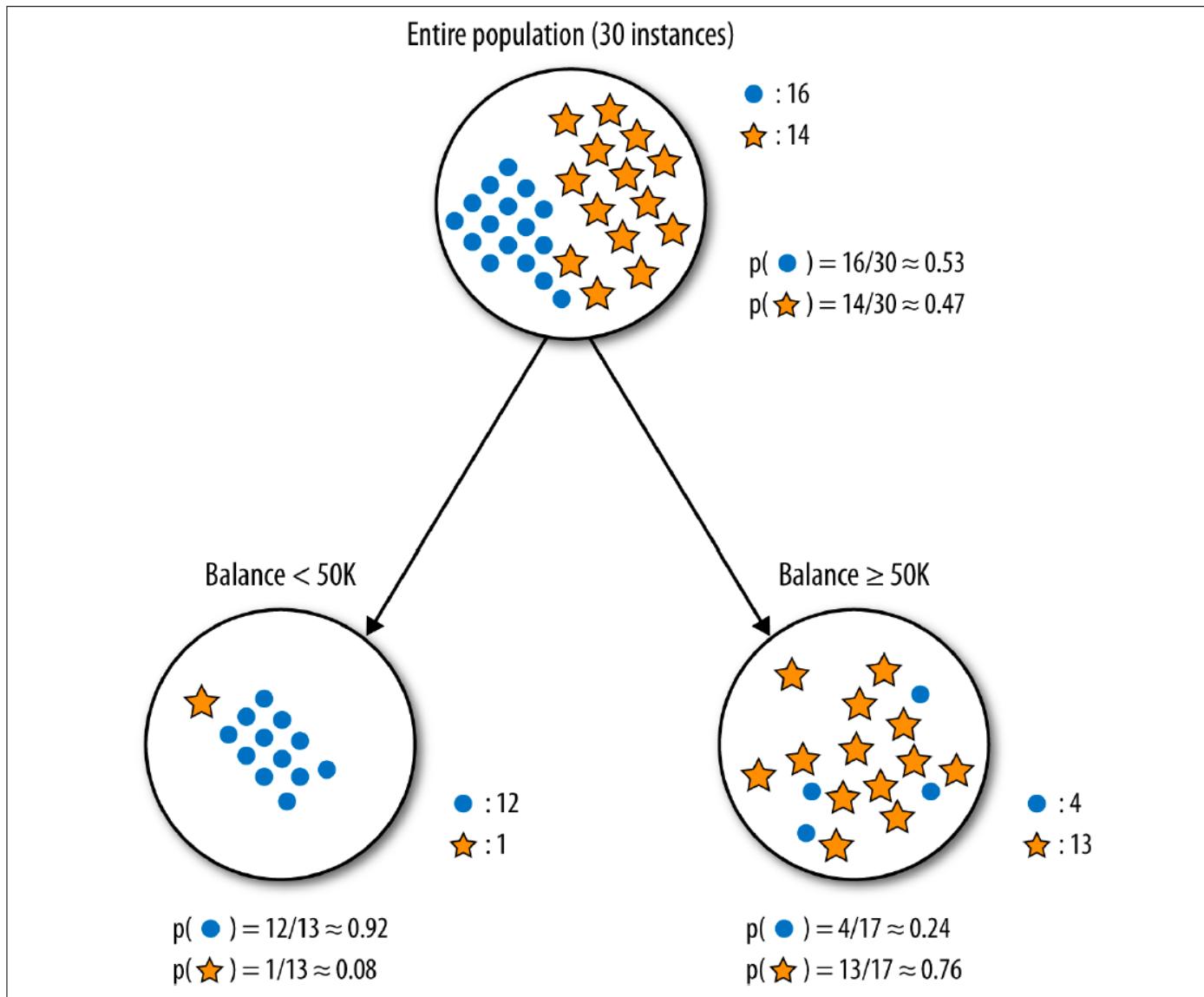
# Κέρδος Πληροφορίας (Information Gain)

- Ένα γνώρισμα χωρίζει ένα σύνολο από στιγμιότυπα σε αρκετά υποσύνολα
- Η εντροπία μας λέει **πόσο ανομοιογενές είναι ένα υποσύνολο**
- Το **κέρδος πληροφορίας (IG)** μπορεί να χρησιμοποιηθεί για να μετρήσουμε **πόσο ένα γνώρισμα βελτιώνει την εντροπία** (δλδ. τη μειώνει) μέσω της τμηματοποίησης που δημιουργεί:

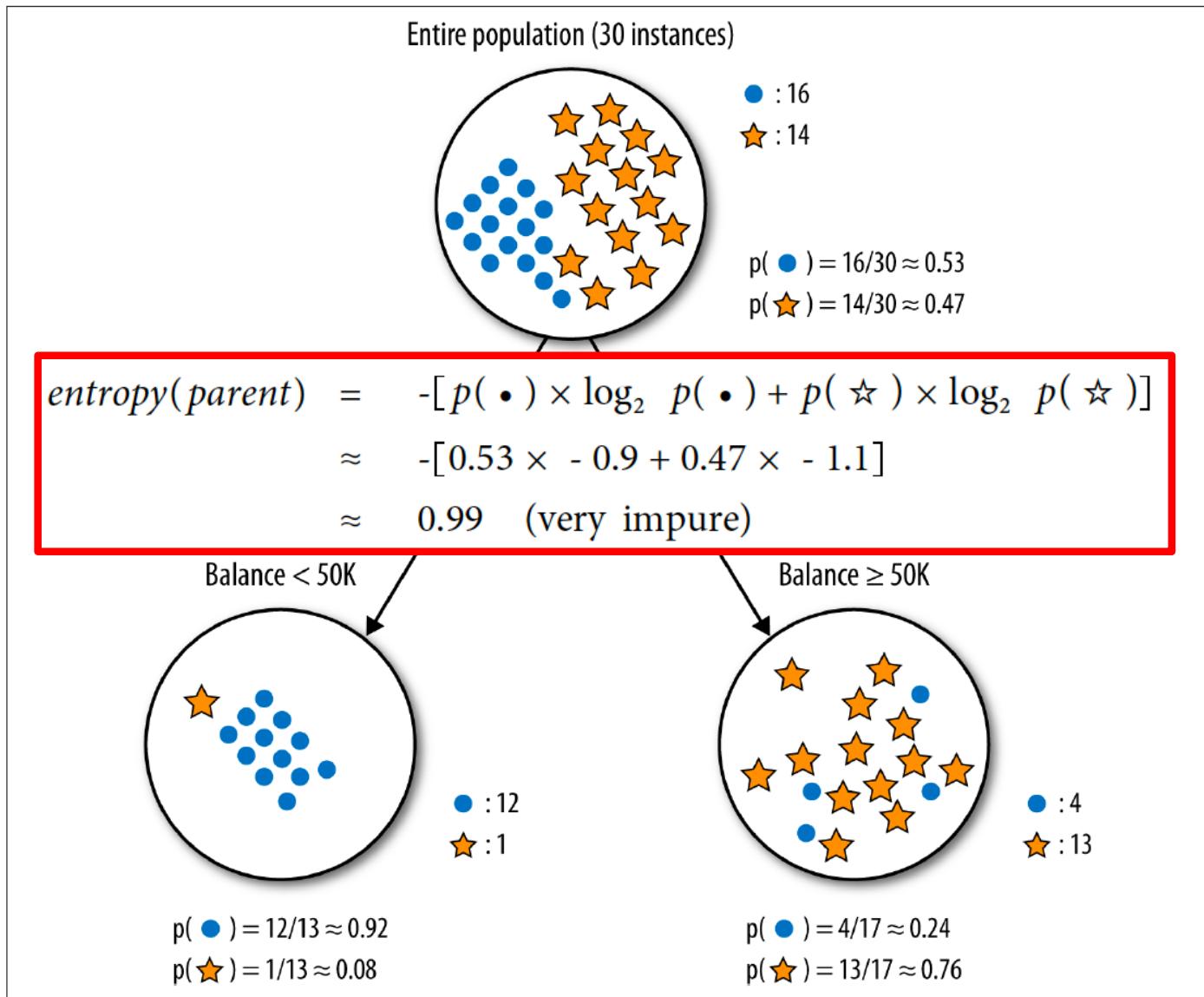
$$IG(P,\theta) = \text{entropy}(P) - [ p(\theta_1) \text{ entropy}(\theta_1) + p(\theta_2) \text{ entropy}(\theta_2) + \dots ]$$

- $P$ : αρχικό σύνολο παραδειγμάτων, σύνολο γονέας (parent set)
- $\theta_i$ : τα θυγατρικά σύνολα (children sets)
- $p(\theta_i)$ : η αναλογία των στιγμιότυπων στο θυγατρικό σύνολο  $\theta_i$

# Διαχωρισμός Δείγματος Παραγραφής Χρέους



# Διαχωρισμός Δείγματος Παραγραφής Χρέους



# Διαχωρισμός Δείγματος Παραγραφής Χρέους

Entire population (30 instances)

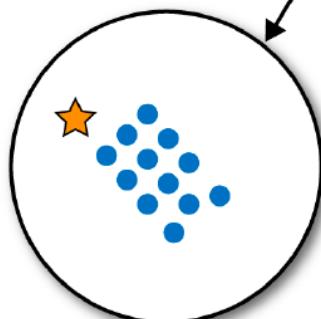


● : 16  
★ : 14

$$p(\bullet) = 16/30 \approx 0.53$$
$$p(\star) = 14/30 \approx 0.47$$

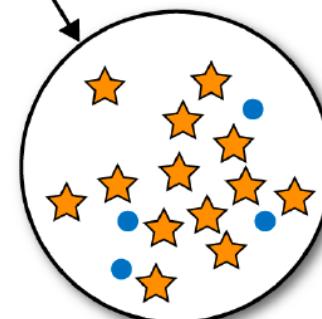
$$\begin{aligned} \text{entropy}(\text{Balance} < 50K) &= -[p(\bullet) \times \log_2 p(\bullet) + p(\star) \times \log_2 p(\star)] \\ &\approx -[0.92 \times (-0.12) + 0.08 \times (-3.7)] \\ &\approx 0.39 \end{aligned}$$

Balance < 50K



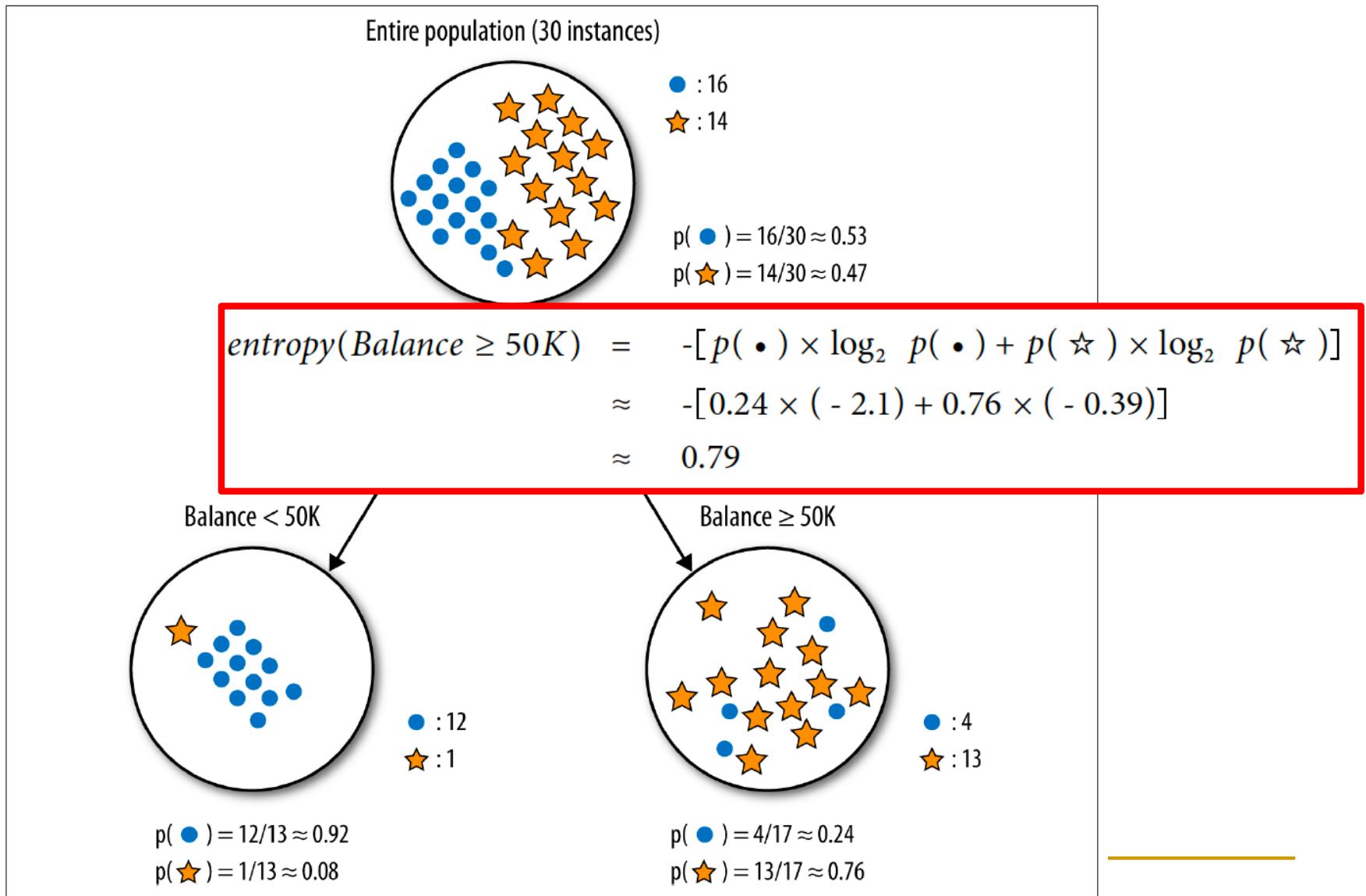
$$p(\bullet) = 12/13 \approx 0.92$$
$$p(\star) = 1/13 \approx 0.08$$

Balance ≥ 50K



$$p(\bullet) = 4/17 \approx 0.24$$
$$p(\star) = 13/17 \approx 0.76$$

# Διαχωρισμός Δείγματος Παραγραφής Χρέους



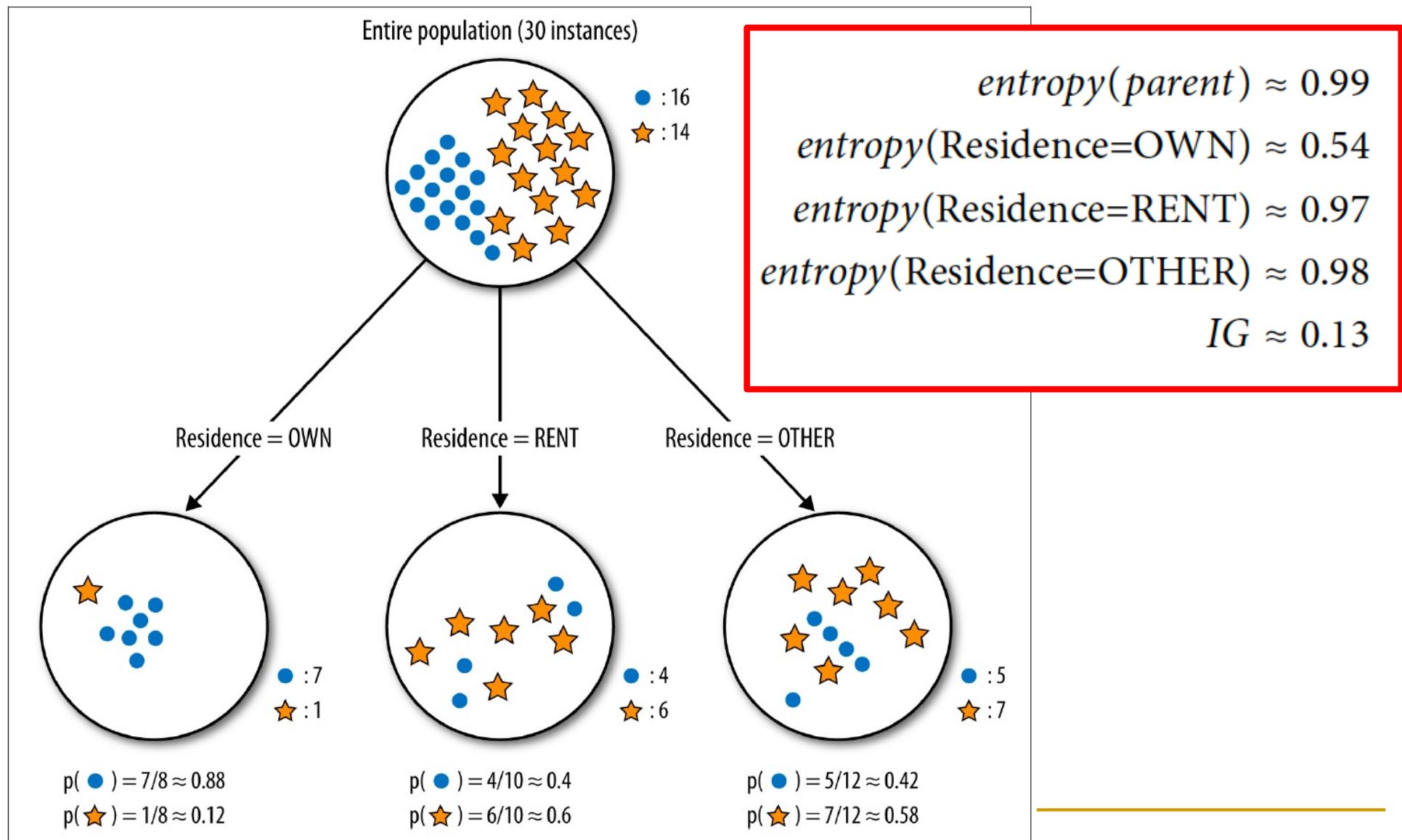
# Υπολογισμός Κέρδους Πληροφορίας

$$\begin{aligned} IG &= \text{entropy}(\text{parent}) - [p(\text{Balance} < 50K) \times \text{entropy}(\text{Balance} < 50K) \\ &\quad + p(\text{Balance} \geq 50K) \times \text{entropy}(\text{Balance} \geq 50K)] \\ &\approx 0.99 - [0.43 \times 0.39 + 0.57 \times 0.79] \\ &\approx 0.37 \end{aligned}$$

*0.43=13/30      0.57=17/30*

- Αυτός ο διαχωρισμός μειώνει σημαντικά την εντροπία
- Σε όρους προγνωστικής μοντελοποίησης:
  - Το γνώρισμα παρέχει πολλές πληροφορίες για την τιμή της μεταβλητής-στόχου

# Άλλος Πιθανός Διαχωρισμός



# 'Άλλος Πιθανός Διαχωρισμός

- Η μεταβλητή **Residence** δίνει κέρδος πληροφορίας (0,13), όμως μικρότερο από τη μεταβλητή **Balance** (0,37)
  - Οφείλεται στο ότι ενώ το θυγατρικό σύνολο **OWN** έχει αρκετά περιορισμένη εντροπία
  - Τα θυγατρικά σύνολα **RENT**, **OTHER**, δεν είναι περισσότερο ομοιογενή από το γονικό
- Άρα η μεταβλητή **Residence** είναι **λιγότερο κατατοπιστική** από τη μεταβλητή **Balance**

# Παράδειγμα Επιλογής Γνωρισμάτων στο Σύνολο Δεδομένων Μανιταριών

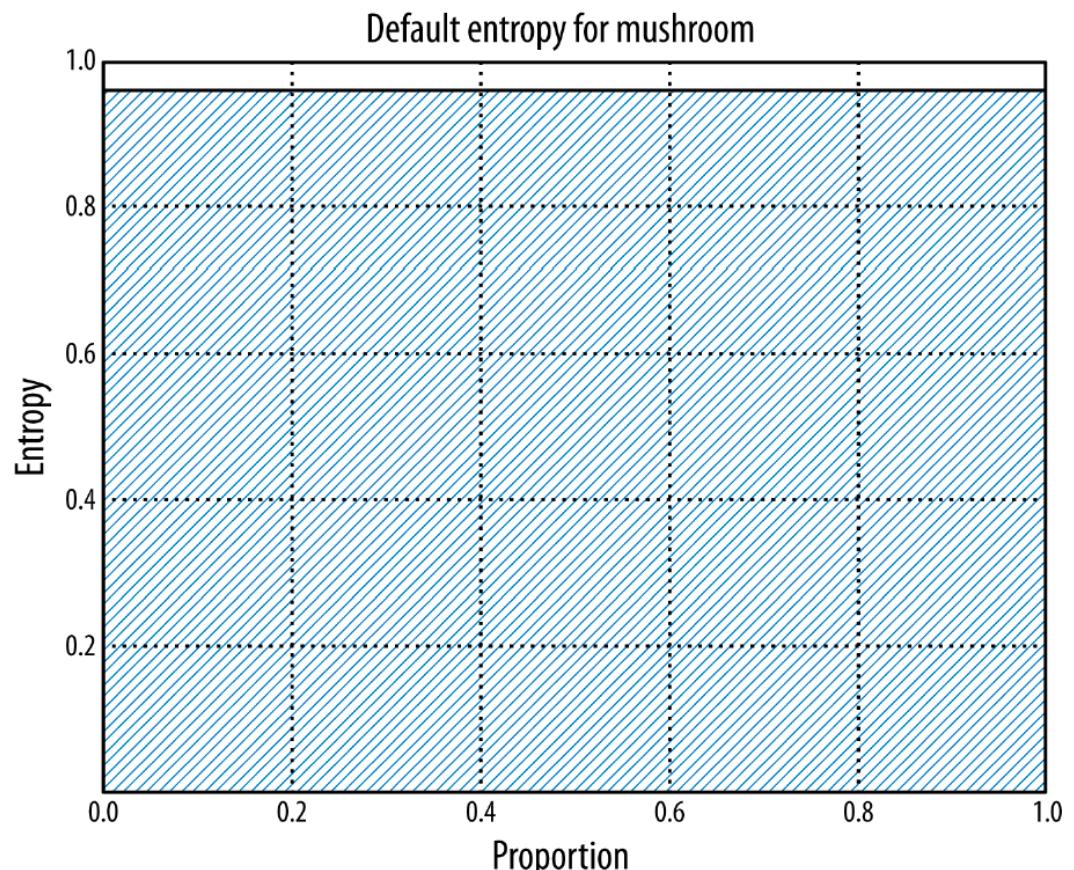
- Δίνεται σύνολο δεδομένων με 5.644 εδώδιμα και δηλητηριώδη μανιτάρια (2.156 δηλητηριώδη και 3.488 εδώδιμα) που περιγράφονται από 22 γνωρίσματα
  - Διαθέσιμο: <http://archive.ics.uci.edu/ml/datasets/Mushroom>
- Πρόβλημα: Ποιο είναι το πιο χρήσιμο γνώρισμα για να διακρίνουμε τα μανιτάρια σε εδώδιμα (edible=Yes) ή σε δηλητηριώδη (edible>No);

# Γνωρίσματα Συνόλου Δεδομένων Μανιταριών

Attribute name	Possible values
GILL-SPACING	close, crowded, distant
GILL-SIZE	broad, narrow
GILL-COLOR	black, brown, buff, chocolate, gray, green, orange, pink, purple, red, white, yellow
STALK-SHAPE	enlarging, tapering
STALK-ROOT	bulbous, club, cup, equal, rhizomorphs, rooted, missing
STALK-SURFACE-ABOVE-RING	fibrous, scaly, silky, smooth
STALK-SURFACE-BELOW-RING	fibrous, scaly, silky, smooth
STALK-COLOR-ABOVE-RING	brown, buff, cinnamon, gray, orange, pink, red, white, yellow
STALK-COLOR-BELOW-RING	brown, buff, cinnamon, gray, orange, pink, red, white, yellow
VEIL-TYPE	partial, universal
VEIL-COLOR	brown, orange, white, yellow
RING-NUMBER	none, one, two
RING-TYPE	cobwebby, evanescent, flaring, large, none, pendant, sheathing, zone
SPORE-PRINT-COLOR	black, brown, buff, chocolate, green, orange, purple, white, yellow
POPULATION	abundant, clustered, numerous, scattered, several, solitary
HABITAT	grasses, leaves, meadows, paths, urban, waste, woods
EDIBLE? (Target variable)	yes, no

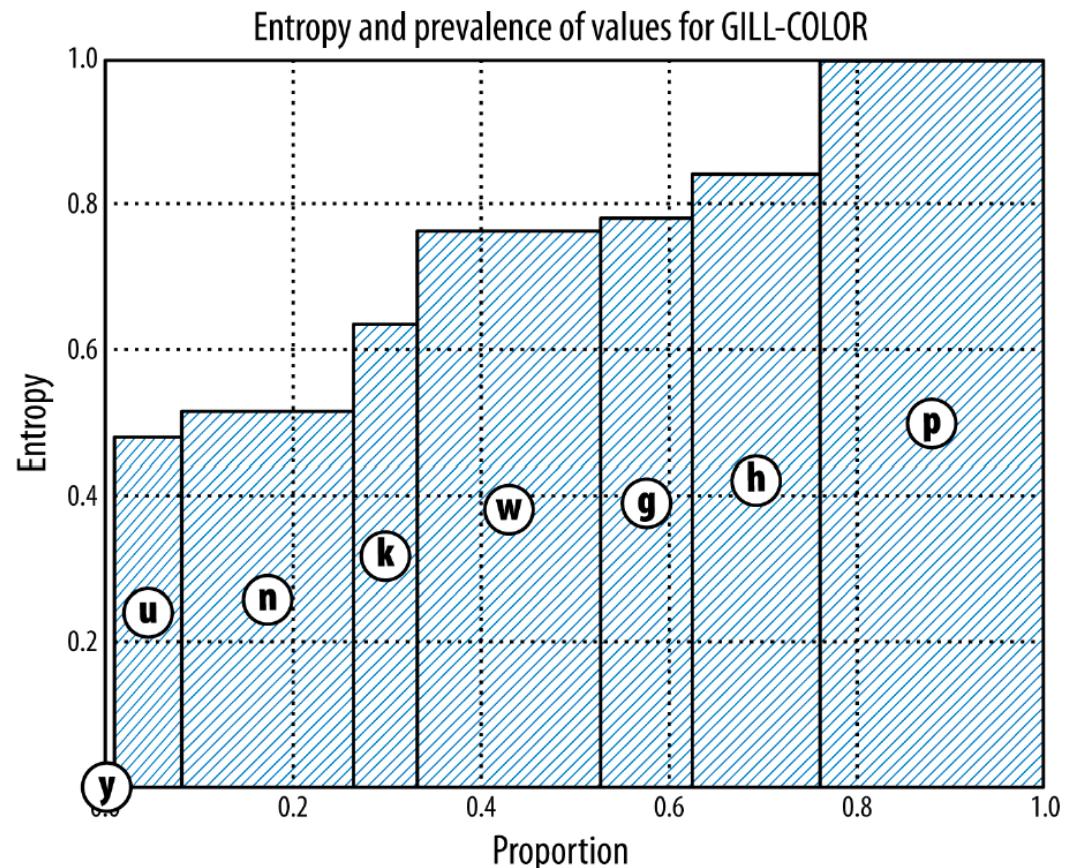
# Παράδειγμα Επιλογής Γνωρισμάτων στο Σύνολο Δεδομένων Μανιταριών

- Υπολογίζουμε το κέρδος πληροφορίας που επιτυγχάνεται από το διαχωρισμό για κάθε γνώρισμα ξεχωριστά
- Πρώτα υπολογίζεται η εντροπία όλου του συνόλου ( $=0,96$ )
- Αν οι δύο κατηγορίες ήταν απόλυτα ισοκατανεμημένες, η εντροπία θα ήταν ίση με 1
- **Γράφημα εντροπίας**
  - Η **έκταση της σκιασμένης επιφάνειας** αντιστοιχεί στην ποσότητα της εντροπίας
  - **Στόχος: χαμηλή εντροπία**, άρα όσο δυνατόν λιγότερο σκιασμένη επιφάνεια



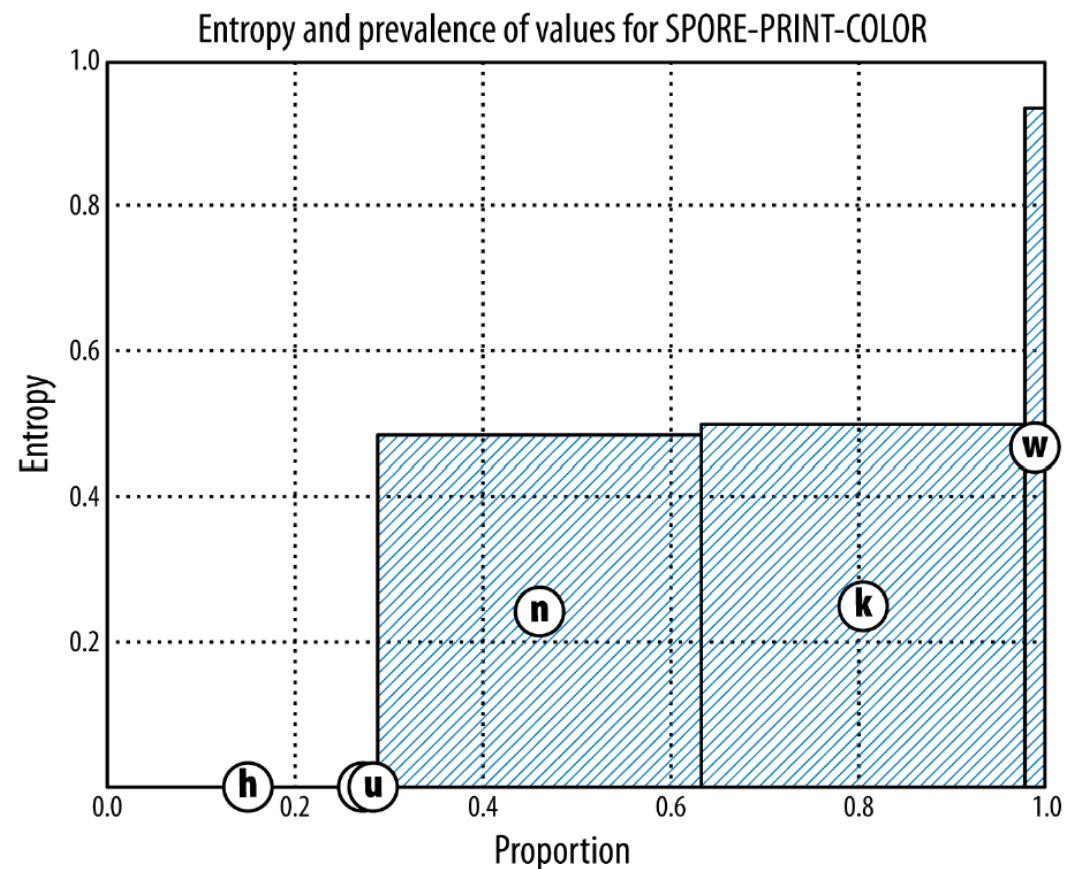
# Παράδειγμα Επιλογής Γνωρισμάτων στο Σύνολο Δεδομένων Μανιταριών

- Γράφημα εντροπίας για το γνώρισμα **GILL-COLOR**
- Το πλάτος κάθε γνωρίσματος αντιπροσωπεύει το ποσοστό του συνόλου δεδομένων με αυτή την τιμή
- Το ύψος του κάθε γνωρίσματος είναι η εντροπία του
- Το **GILL-COLOR** μειώνει κάπως την εντροπία



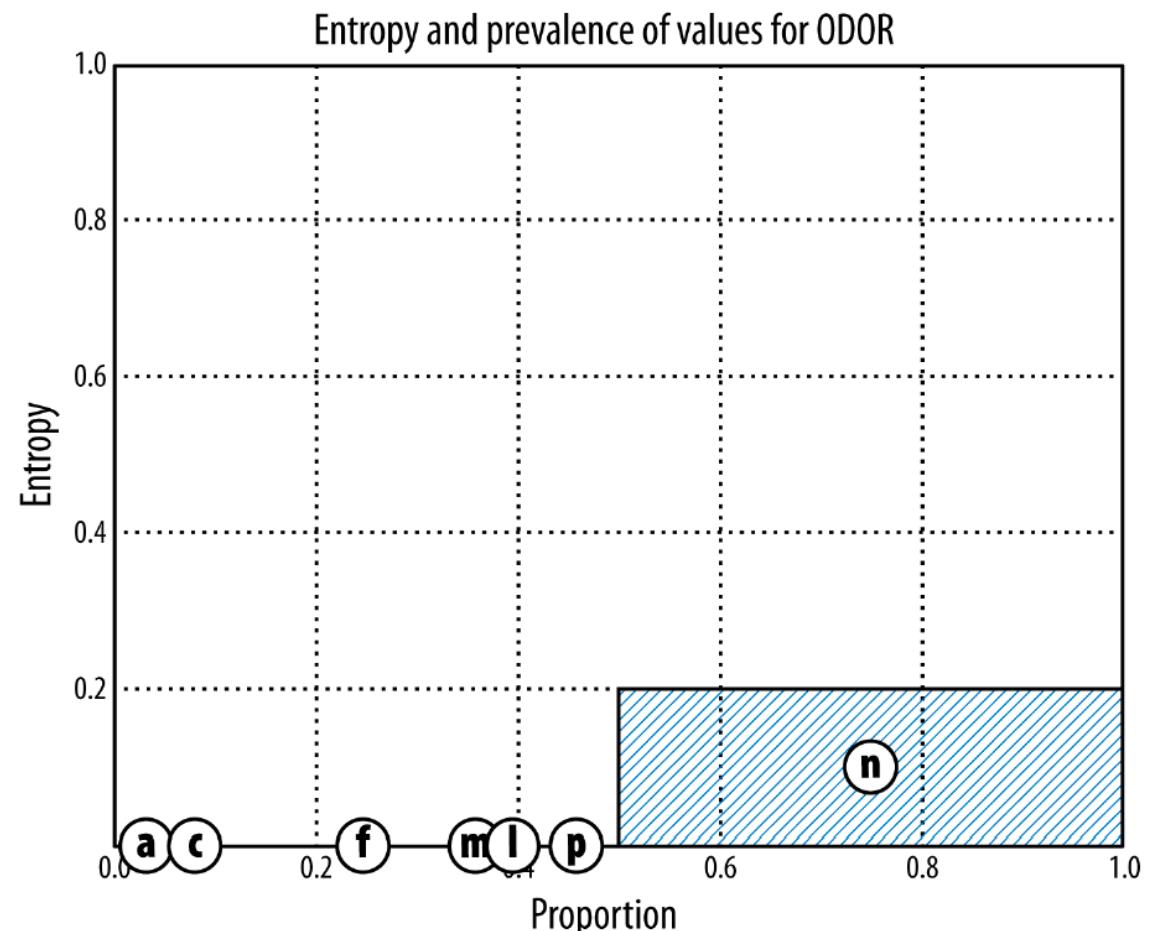
# Παράδειγμα Επιλογής Γνωρισμάτων στο Σύνολο Δεδομένων Μανιταριών

- Το γνώρισμα **SPORE-PRINT-COLOR** μειώνει την εντροπία
- Κάποιες τιμές (π.χ. **h**) καθορίζουν τέλεια την τιμή-στόχο
  - Για αυτό παράγουν ράβδο μηδενικής εντροπίας



# Παράδειγμα Επιλογής Γνωρισμάτων στο Σύνολο Δεδομένων Μανιταριών

- Το γνώρισμα **ODOR** έχει το υψηλότερο κέρδος πληροφορίας από κάθε άλλο γνώρισμα
- Μειώνει τη συνολική εντροπία του συνόλου δεδομένων στο 0,1 περίπου
- Άρα, κέρδος πληροφορίας =  $0,96 - 0,1 = 0,86$
- → *Η οσμή (ODOR) είναι απόλυτα χαρακτηριστική για την αναγνώριση της εδωδιμότητας των μανιταριών*



## Άσκηση για το σπίτι:

Κατασκευάστε ένα python notebook με την παραπάνω μελέτη επιλογής γνωρισμάτων για το σύνολο δεδομένων *Mushroom*, υλοποιώντας μόνοι σας όλους τους υπολογισμούς.

# Περίγραμμα Μαθήματος

- Εισαγωγή στην προγνωστική μοντελοποίηση
- Επιλογή γνωρισμάτων (**feature selection**)
  - Εντροπία (**entropy**), κέρδος πληροφορίας (**information gain**)

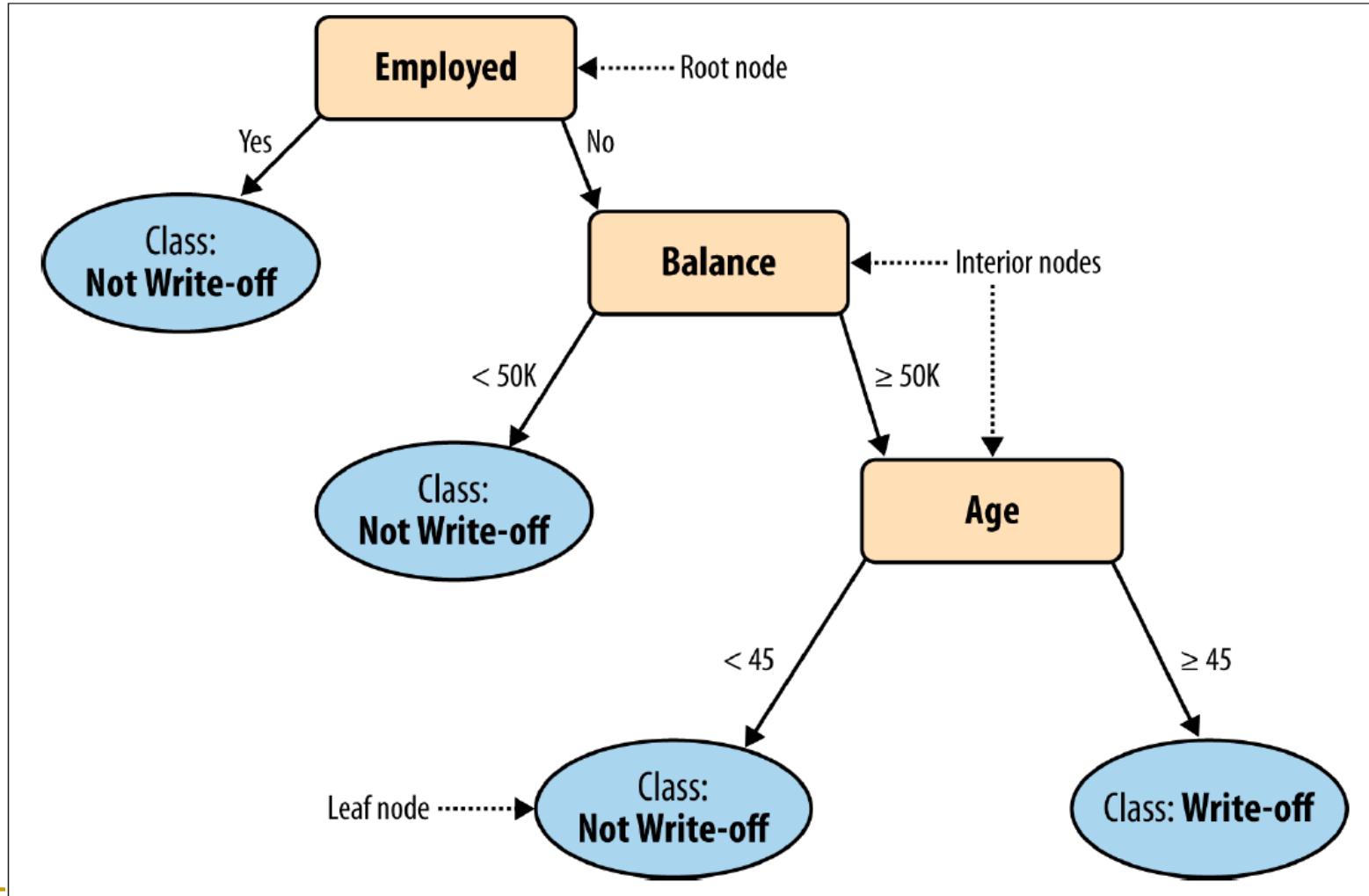
## ■ Δέντρα απόφασης (**decision trees**)

- Άλλα μέτρα επιλογής γνωρισμάτων
  - **Gini**, αναλογία κέρδους (**gain ratio**)

# Εισαγωγή στα Δέντρα Απόφασης (Decision Trees)

- Τα δέντρα απόφασης (decision trees) χρησιμοποιούνται συχνά ως μοντέλα πρόβλεψης
- Κάθε εσωτερικός κόμβος περιέχει έναν έλεγχο ενός γνωρίσματος
- Το δέντρο δημιουργεί μια τμηματοποίηση (διαμέριση) των δεδομένων
- Κάθε εγγραφή δεδομένων αντιστοιχεί σε μία και μόνο μία διαδρομή, άρα σε ένα μόνο φύλλο
- Κάθε φύλλο περιλαμβάνει μια τιμή για τη μεταβλητή-στόχο
- Επομένως το δέντρο αποτελεί εποπτευόμενη τμηματοποίηση (supervised segmentation)

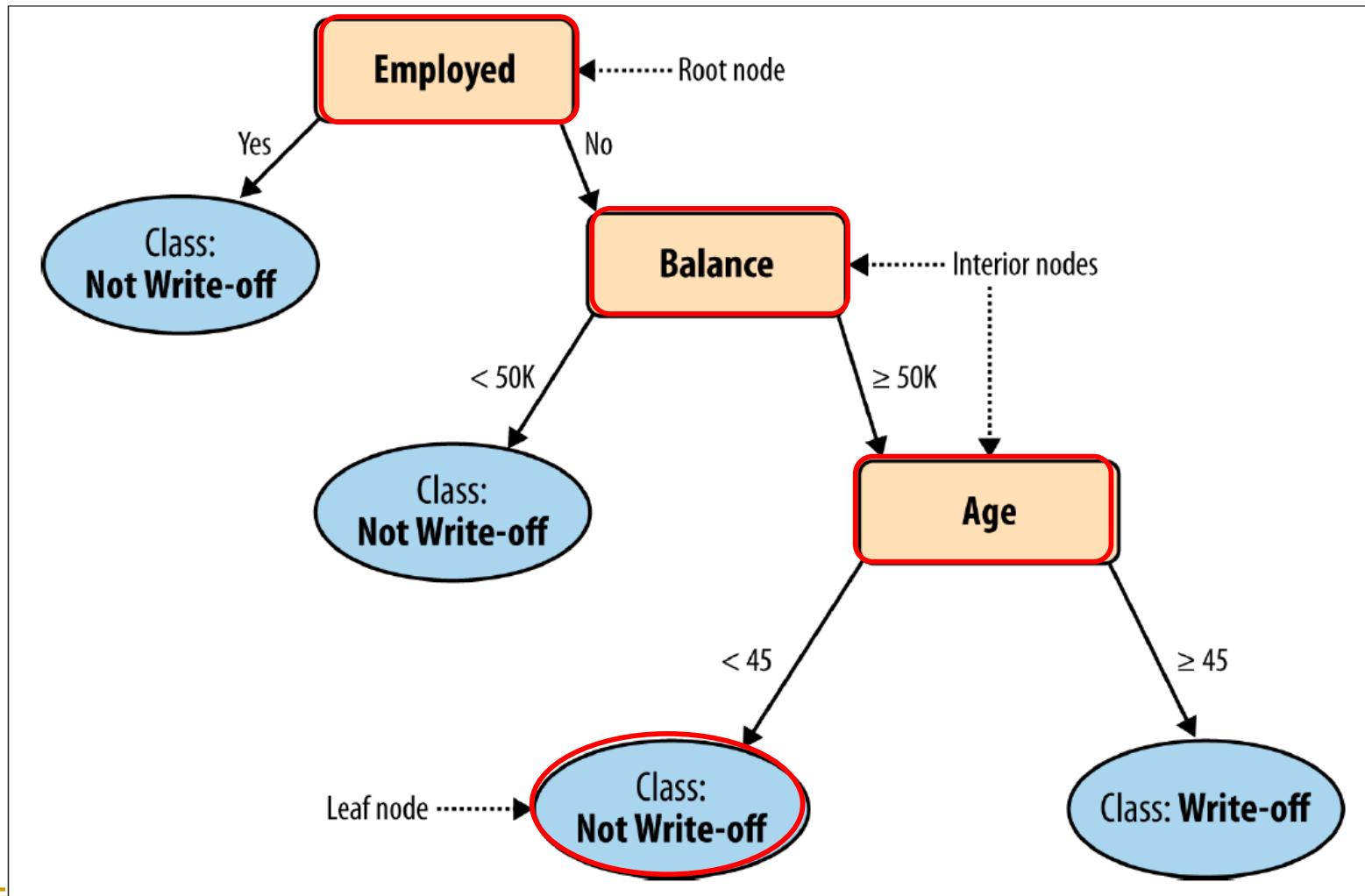
# Εισαγωγή στα Δέντρα Απόφασης – Παράδειγμα



# Εισαγωγή στα Δέντρα Απόφασης – Πρόβλεψη για νέα Εγγραφή

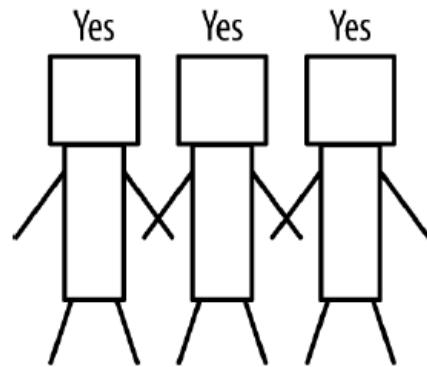
Claudio:

<Balance=115K, Employed=No, Age=40>

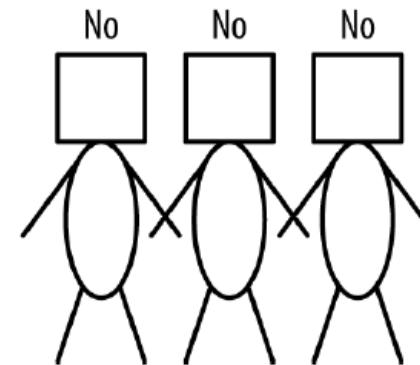
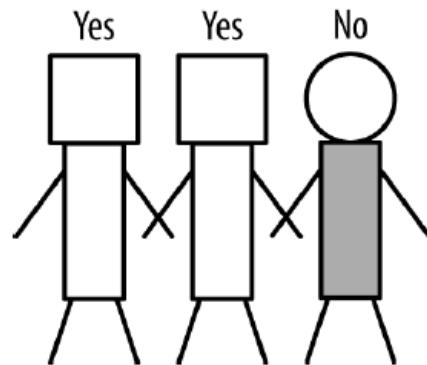
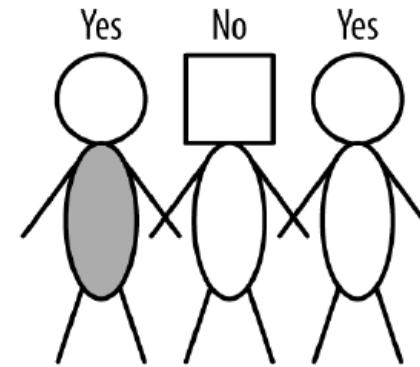


# Επαγωγή Δέντρου (Tree Induction)

Rectangular Bodies

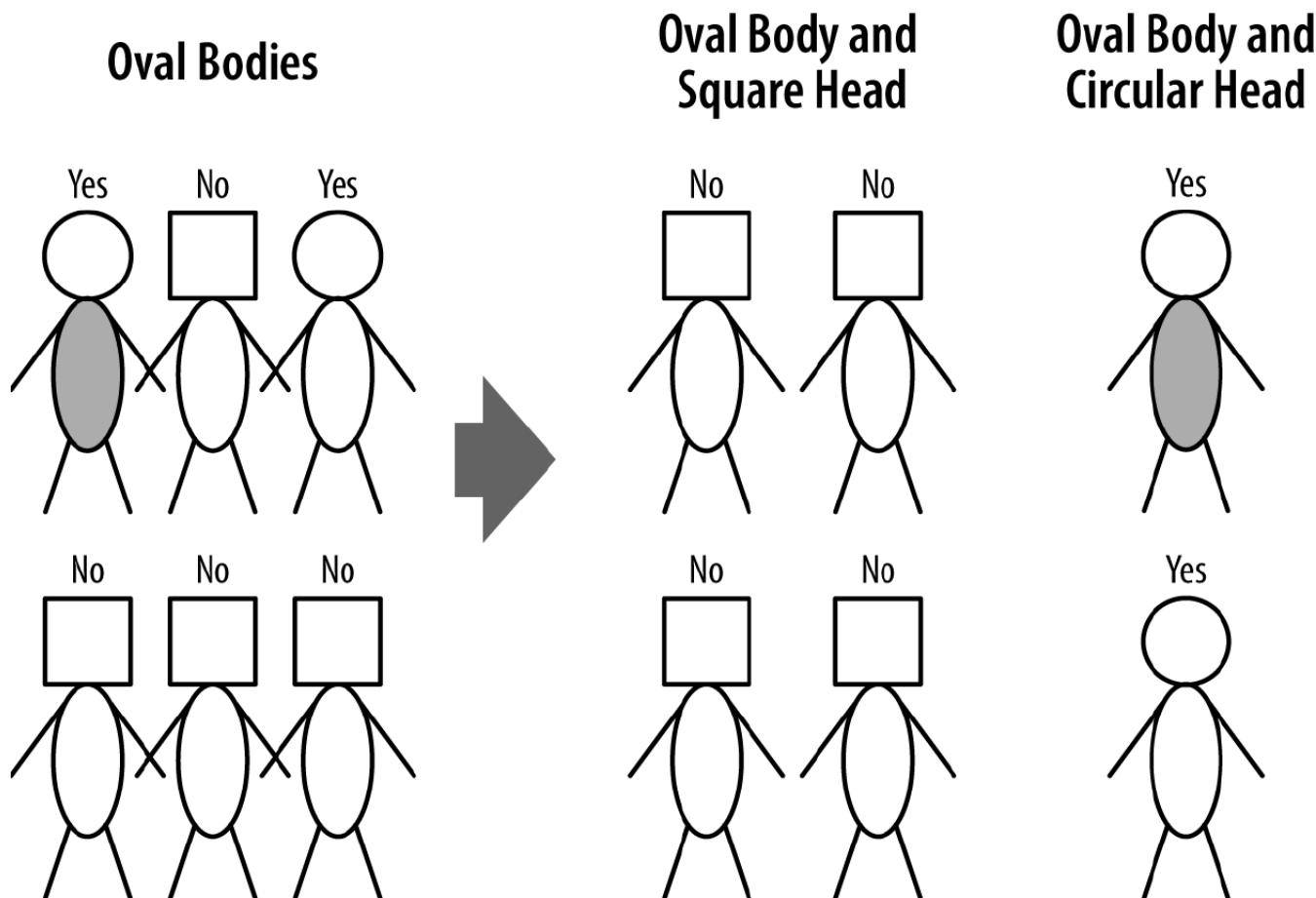


Oval Bodies



**Πρώτη διαμέριση:** βάσει **σχήματος σώματος** (ορθογώνιο – ωοειδές)

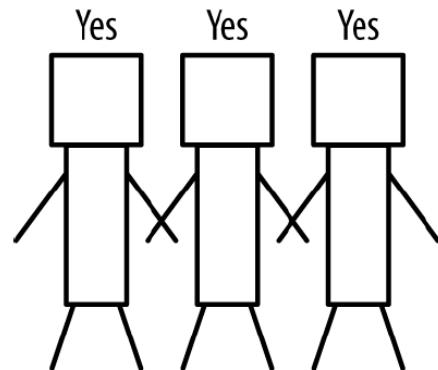
# Επαγωγή Δέντρου (Tree Induction)



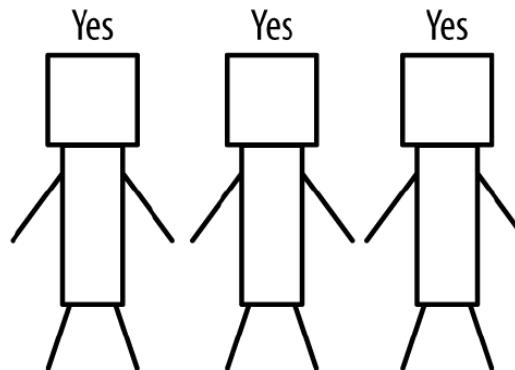
**Δεύτερη διαμέριση:** βάσει τύπου κεφαλιού (ωοειδές – στρογγυλό)

# Επαγωγή Δέντρου (Tree Induction)

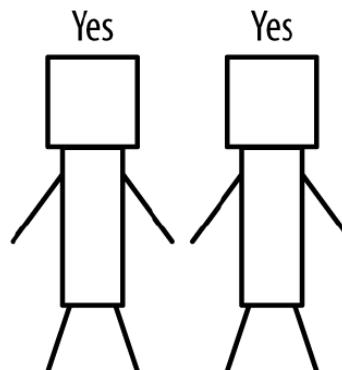
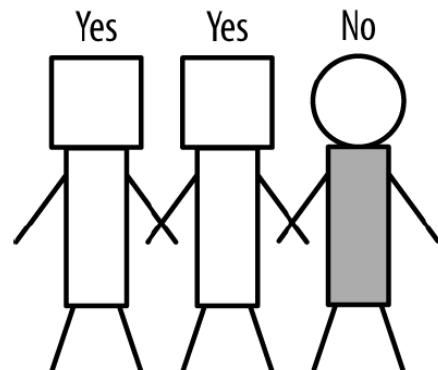
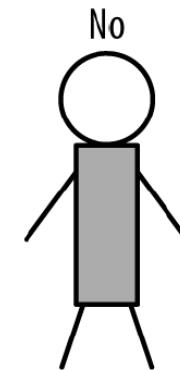
Rectangular Bodies



Rectangular Body  
and White

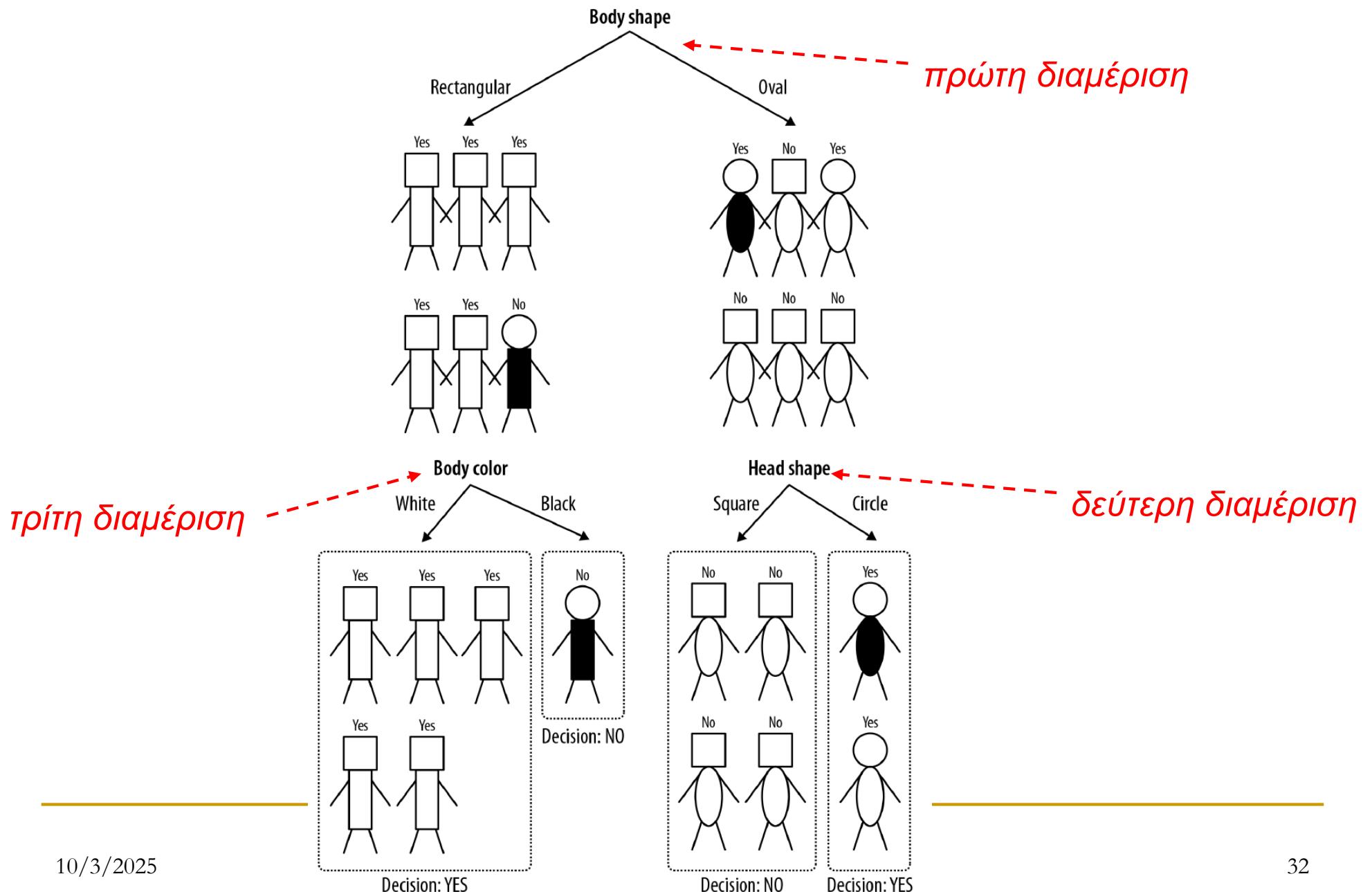


Rectangular Body  
and Gray



**Τρίτη διαμέριση:** βάσει χρώματος σώματος (ανοικτό – σκούρο)

# Επαγωγή Δέντρου (Tree Induction)



# Επαγωγή Δέντρου (Tree Induction)

- Μια **επαναληπτική** διαδικασία **διαίρει-και-βασίλευε** (divide-and-conquer)
- Στόχος σε κάθε βήμα να επιλέξουμε ένα γνώρισμα για να **χωρίσουμε την τρέχουσα ομάδα σε όσο το δυνατόν πιο ομοιογενείς υποομάδες**, σε σχέση με τη μεταβλητή-στόχο
- Ο διαχωρισμός εκτελείται αναδρομικά
- Δοκιμάζουμε όλα τα γνωρίσματα, επιλέγουμε εκείνα που οδηγούν σε πιο ομοιογενείς υποομάδες
- Πότε ολοκληρώνεται αυτή η διαδικασία;
  - Όταν οι κόμβοι είναι ομοιογενείς ή
  - Όταν εξαντλήσουμε τα γνωρίσματα που χρησιμοποιούμε για το διαχωρισμό
  - **Στην πράξη μπορεί να θέλουμε να σταματήσουμε νωρίτερα**

# Ο Αλγόριθμος του Hunt

- Αποτελεί τη **βάση πολλών αλγορίθμων επαγωγής δέντρου**
  - Το δέντρο απόφασης μεγαλώνει αναδρομικά
  - Διαιρώντας τις εγγραφές σε διαδοχικά πιο αμιγή σύνολα
- Ακολουθεί μια **άπληστη (greedy)** στρατηγική
  - Διότι η εύρεση του **βέλτιστου δέντρου** είναι **υπολογιστικά ανέφικτη**
  - Λόγω του εκθετικά αυξανόμενου μεγέθους του χώρου αναζήτησης

Έστω  $D_t$  το σύνολο εγγραφών σε κόμβο  $t$ , και  $y=\{y_1, y_2, \dots\}$  οι ετικέτες κατηγορίας (=τιμές μεταβλητής-στόχου)

- **BHMA 1:** Αν όλες οι εγγραφές στο  $D_t$  ανήκουν στην κατηγορία  $y_t$ , τότε το  $t$  είναι κόμβος φύλλο με ετικέτα  $y_t$
- **BHMA 2:** Αν το  $D_t$  περιέχει εγγραφές που ανήκουν σε πολλές κατηγορίες, τότε επιλέγεται μια **συνθήκη ελέγχου χαρακτηριστικού** για να διαιρέσει τις εγγραφές σε μικρότερα υποσύνολα. Για κάθε αποτέλεσμα της συνθήκης ελέγχου, δημιουργείται ένας κόμβος παιδί και οι εγγραφές κατανέμονται στα παιδιά.

Ο αλγόριθμος εφαρμόζεται αναδρομικά σε κάθε παιδί.

# Πρόβλεψη Δανειοληπτών που αθετούν την Πληρωμή Δανείου

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

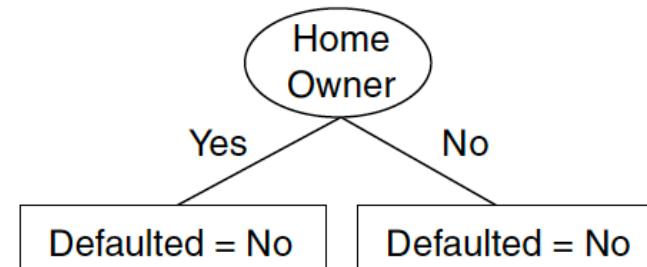
Σύνολο δεδομένων  
εκπαίδευσης

# Πρόβλεψη Δανειοληπτών που αθετούν την Πληρωμή Δανείου

Ξεκινάει με κόμβο που  
αντιστοιχεί στην ετικέτα  
των περισσότερων  
εγγραφών

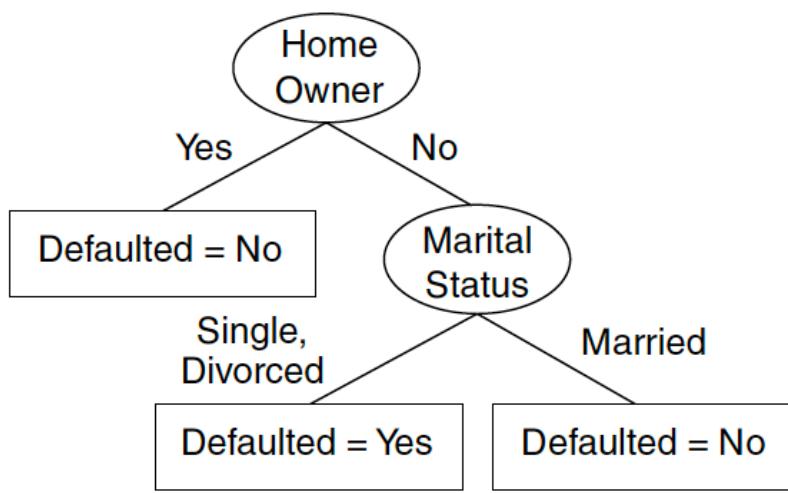
Defaulted = No

(a)

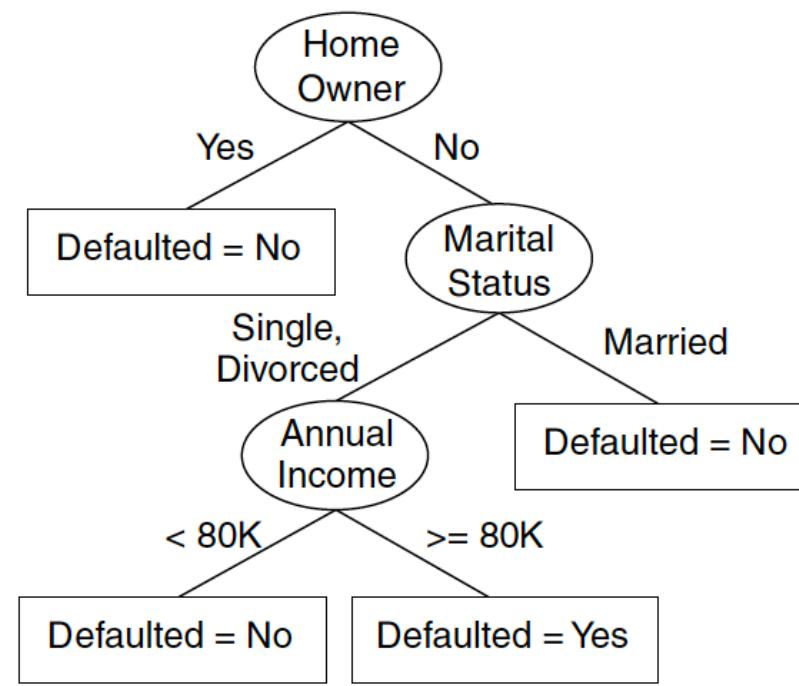


Κόμβος φύλλο (όλες οι  
εγγραφές έχουν την ίδια  
ετικέτα)

(b)



(c)

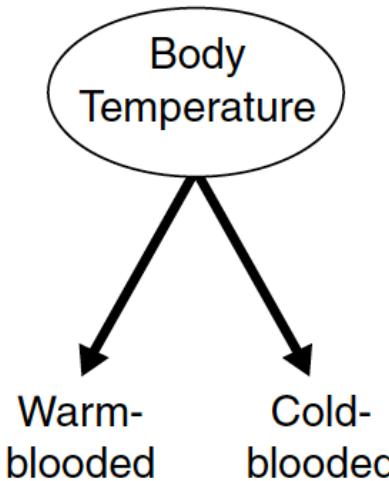


(d)

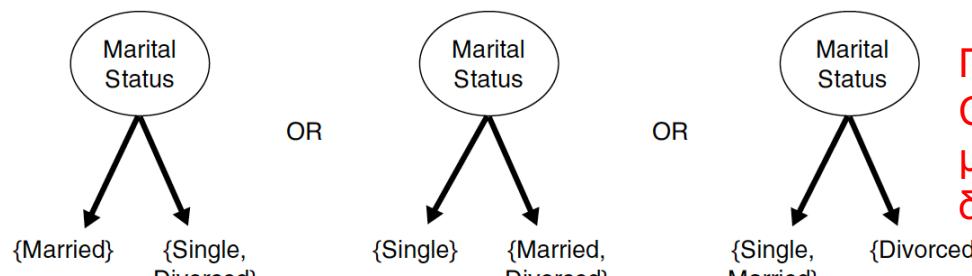
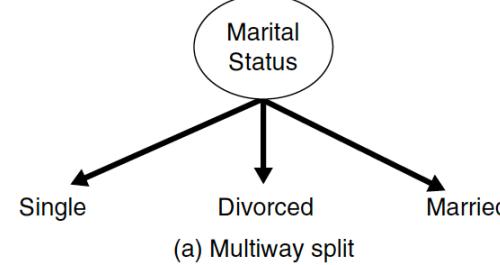
# Παρατηρήσεις στον Αλγόριθμο του Hunt

- Λειτουργεί καλά όταν κάθε συνδυασμός των τιμών των γνωρισμάτων
  - είναι παρών στο σύνολο δεδομένων, και
  - έχει μοναδική ετικέτα κατηγορίας
- Επειδή αυτές οι υποθέσεις είναι πολύ αυστηρές, απαιτούνται επιπλέον συνθήκες
  - Αν παράγονται **κόμβοι παιδιά χωρίς εγγραφές**, τότε αυτοί **δηλώνονται** ως **φύλλο** με **ετικέτα** αυτήν που έχει η **πλειοψηφία** του κόμβου γονέα
  - Αν στο Βήμα 2, **όλες** οι εγγραφές έχουν **ίδιες τιμές** γνωρισμάτων (εκτός της ετικέτας), **τότε ο κόμβος δηλώνεται** ως **φύλλο** με **ετικέτα** αυτήν που έχει **η πλειοψηφία** των εγγραφών

# Τρόποι Έκφρασης Συνθηκών Ελέγχου

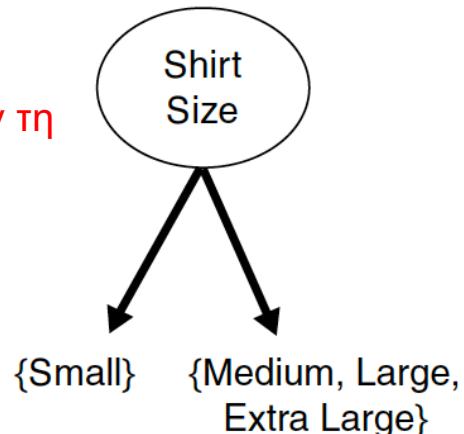
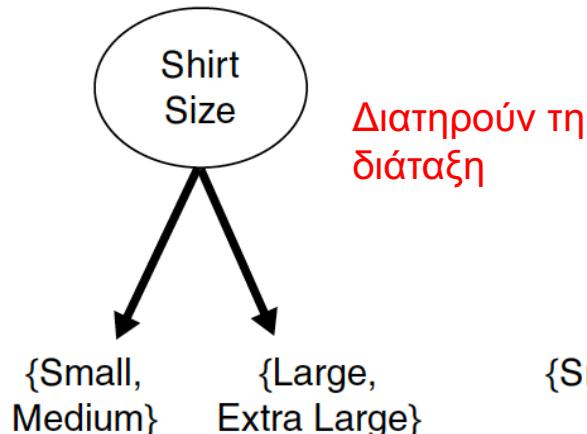


Δυαδικά γνωρίσματα (binary)



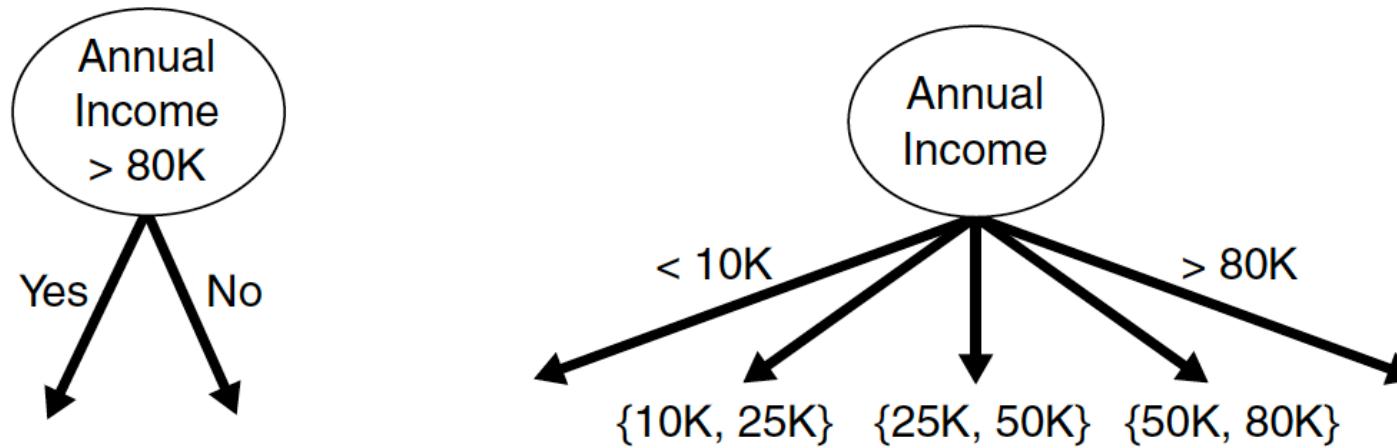
Π.χ. ο αλγόριθμος  
CART παράγει  
μόνο δυαδικούς  
διαχωρισμούς

Ονομαστικά γνωρίσματα (nominal)



Τακτικά γνωρίσματα (ordinal)

# Τρόποι Έκφρασης Συνθηκών Ελέγχου

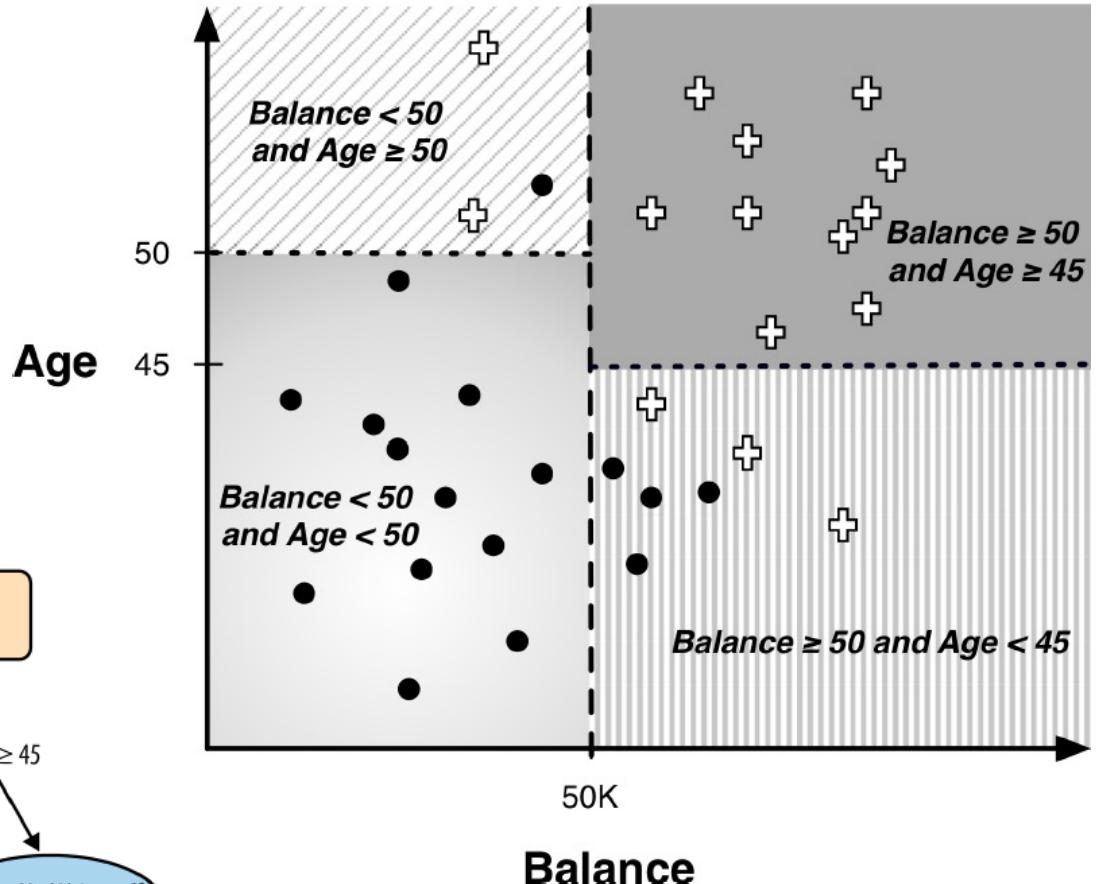
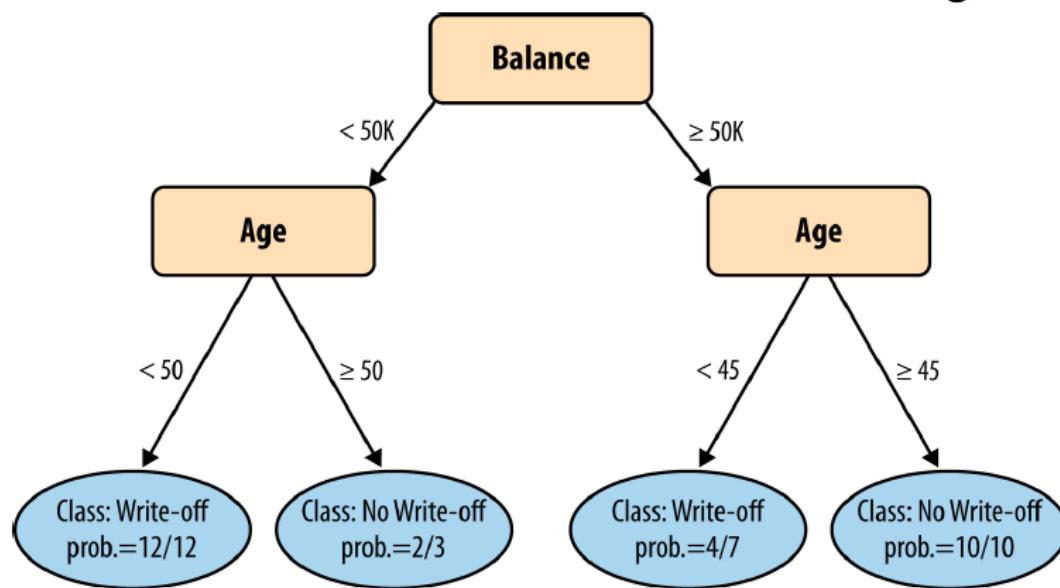


Συνεχή γνωρίσματα (continuous)

Η συνθήκη ελέγχου μπορεί να εκφραστεί ως:

- Ένα κριτήριο σύγκρισης ( $A < u$ ) ή ( $A \geq u$ ) με δυαδικά αποτελέσματα ή
- Ένα ερώτημα κάλυψης (range query) με αποτέλεσμα ( $u_i \leq A < u_{i+1}$ )

# Οπτικοποίηση Τυχαιτοποιήσεων



# Ευθείες Απόφασης και Υπερεπίπεδα

- Οι ευθείες που χωρίζουν τις περιοχές είναι γνωστές ως
  - Στις  $2\Delta$ : ευθείες απόφασης (decision lines) ή
  - Σε  $>2\Delta$ : επιφάνειες απόστασης (decision surfaces) ή όρια απόφασης (decision boundaries)
- Σε κάθε κόμβο γίνεται έλεγχος μιας μεταβλητής, άρα *πάντα το όριο απόφασης είναι κάθετο στον άξονα που αντιστοιχεί σε αυτή τη μεταβλητή*

# Δένδρα ως Σύνολα Κανόνων

- Αν ακολουθήσουμε μια διαδρομή του δέντρου συκεντρώνοντας τις συνθήκες ελέγχου, δημιουργούμε έναν **κανόνα**
- Το **δέντρο κατηγοριοποίησης** είναι **ισοδύναμο** με ένα **σύνολο κανόνων**:

IF (Balance < 50K) AND (Age < 50) THEN Class=Write-off

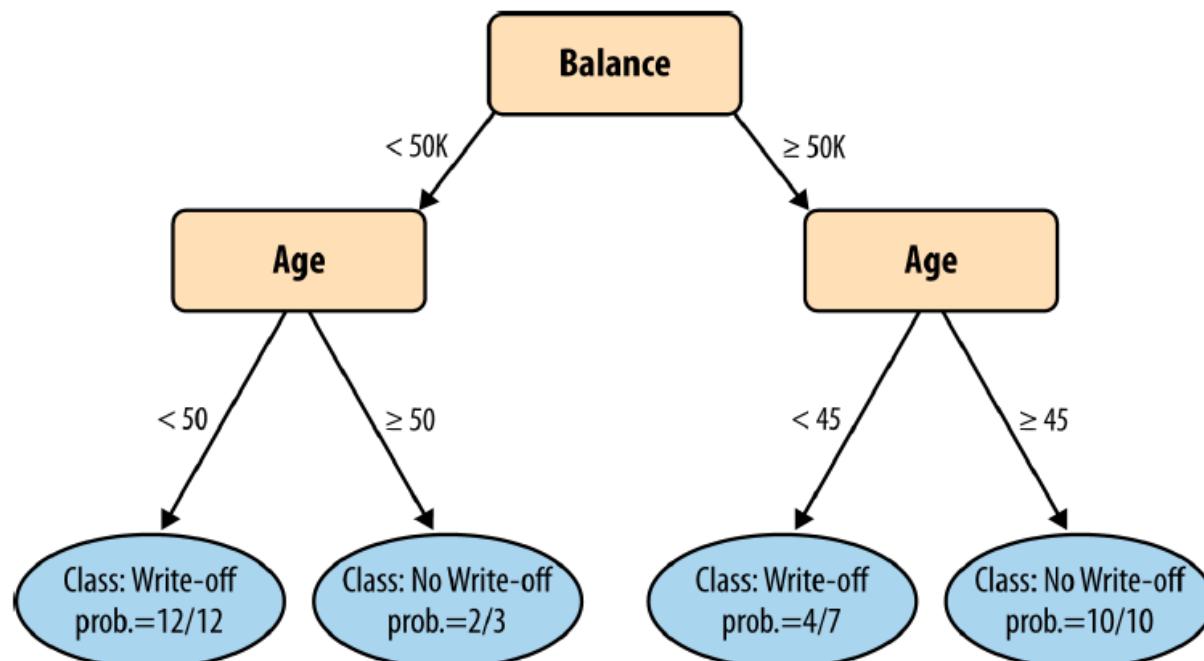
IF (Balance < 50K) AND (Age ≥ 50) THEN Class=No Write-off

IF (Balance ≥ 50K) AND (Age < 45) THEN Class=Write-off

IF (Balance ≥ 50K) AND (Age < 45) THEN Class=No Write-off

# Εκτίμηση Πιθανότητας

- Ορισμένες φορές απαιτείται ένας πιο κατατοπιστικός τύπος πρόβλεψης από μια απλή κατηγοριοποίηση
  - Παράδειγμα (πρόβλημα απώλειας πελατών):
    - Αντί για την πρόβλεψη αν θα αποχωρήσει κάποιος πελάτης
    - Ποια η πιθανότητα να αποχωρήσει κάποιος πελάτης
- Δενδρικό μοντέλο εκτίμησης πιθανότητας



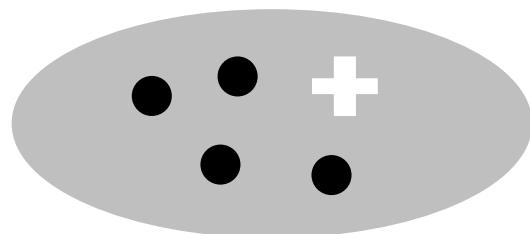
# Εκτίμηση Πιθανότητας

## ■ Εκτίμηση πιθανότητας κατηγορίας

- Αν ένα φύλλο περιέχει **n** θετικά στιγμιότυπα και **m** αρνητικά στιγμιότυπα
- Η πιθανότητα ενός νέου στιγμιότυπου να είναι θετικό: **n / (n+m)**
- Αυτό ονομάζεται **εκτίμηση βάσει συχνότητας** (frequency-based estimate) της πιθανότητας συμμετοχής στην κατηγορία
- Όμως, τι γίνεται αν ο αριθμός στιγμιότυπων σε ένα τμήμα είναι πολύ μικρός;
  - Ακραία περίπτωση: ένα μόνο στιγμιότυπο σε κάποιο φύλλο σημαίνει 100% πιθανότητα;

# Εκτίμηση Πιθανότητας – Διόρθωση Laplace

- Για την αντιμετώπιση του προβλήματος των μικρών δειγμάτων
  - Χρησιμοποιούμε μια «εξομαλυμένη» εκδοχή της εκτίμησης βάσει συχνότητας, γνωστή ως **Διόρθωση Laplace**
    - $p(c) = (n+1) / (n+m+2)$  για δυαδικές κλάσεις, όπου
    - **n** ο αριθμός παραδειγμάτων της κατηγορίας **c**, και
    - **m** ο αριθμός παραδειγμάτων που δεν ανήκουν στην κατηγορία **c**
  - Σκοπός της διόρθωσης Laplace είναι να **μετριάσει την επιρροή των φύλλων με λίγα στιγμιότυπα**

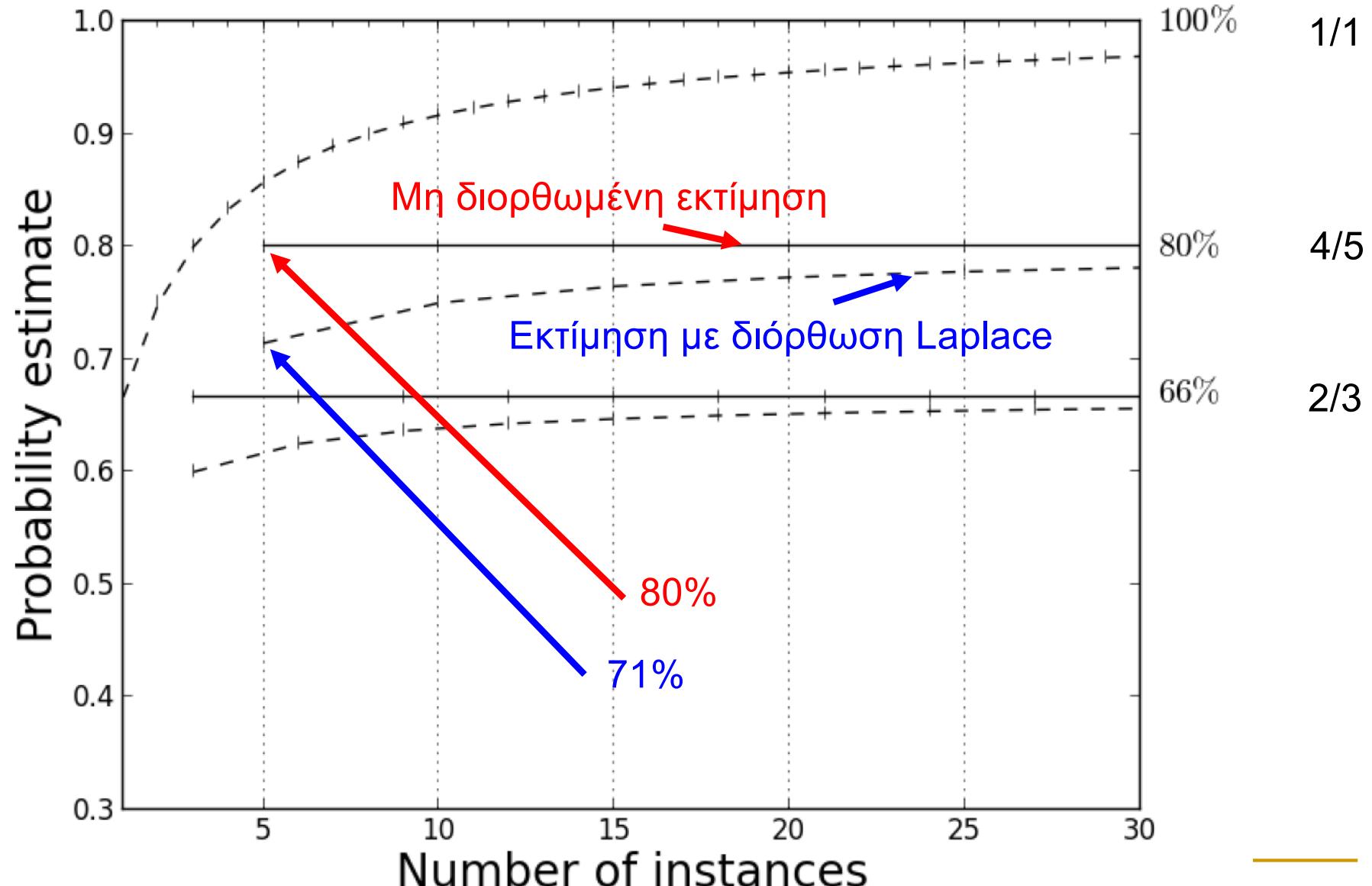


$$p(c=\text{μαύρο}) = n / (n + m) = 4 / 5 = 80\%$$

*Με διόρθωση Laplace:*

$$p(c=\text{μαύρο}) = (n + 1) / (n + m + 2) = 5 / 7 \approx 71\%$$

# Εκτίμηση Πιθανότητας – Διόρθωση Laplace



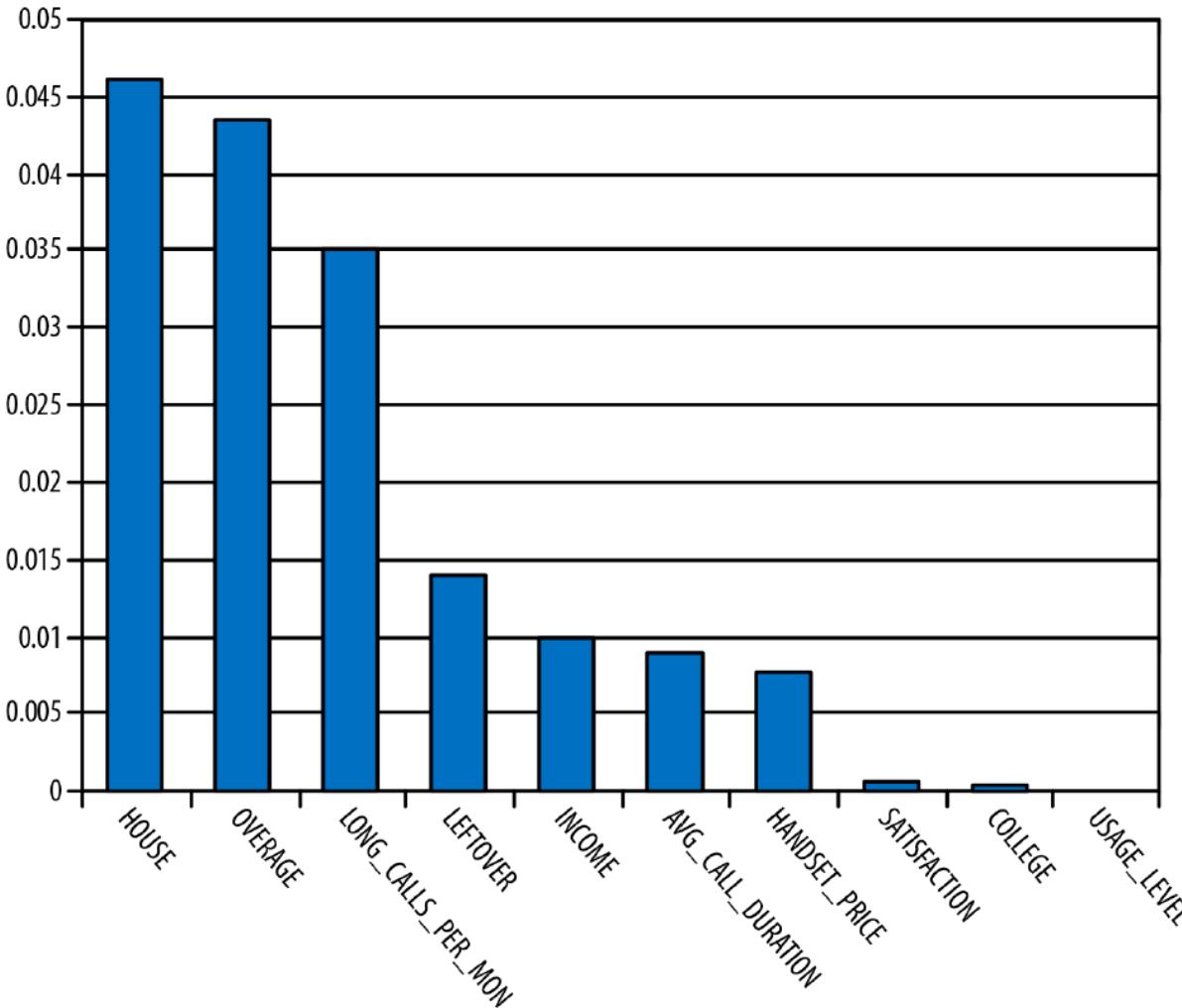
# Παράδειγμα: Πρόβλημα Απώλειας Πελατών με Επαγωγή Δέντρου

- Δίνεται σύνολο δεδομένων 20.000 πελατών εταιρίας κινητής τηλεφωνίας
- Περιγράφονται από τα γνωρίσματα στα δεξιά
- Θέλουμε να **προβλέψουμε ποιοι νέοι πελάτες θα αποχωρήσουν** με την τεχνική επαγωγής δέντρου

Variable	Explanation
COLLEGE	Is the customer college educated?
INCOME	Annual income
OVERAGE	Average overcharges per month
LEFTOVER	Average number of leftover minutes per month
HOUSE	Estimated value of dwelling (from census tract)
HANDSET_PRICE	Cost of phone
LONG_CALLS_PER_MONTH	Average number of long calls (15 mins or over) per month
AVERAGE_CALL_DURATION	Average duration of a call
REPORTED_SATISFACTION	Reported level of satisfaction
REPORTED_USAGE_LEVEL	Self-reported usage level
LEAVE ( <i>Target variable</i> )	Did the customer stay or leave (churn)?

# Παράδειγμα (συνέχ.)

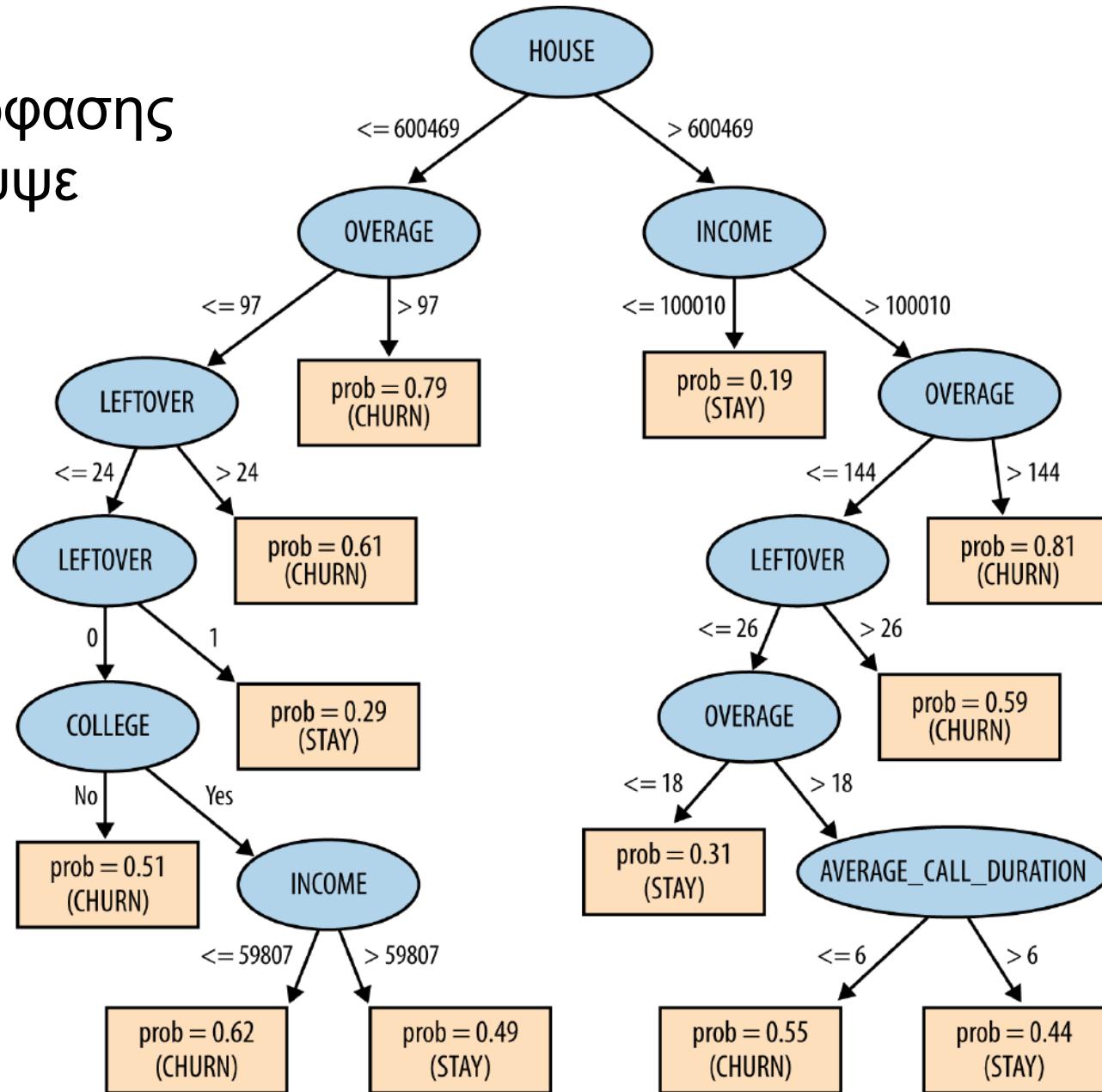
## Κατάταξη γνωρισμάτων βάσει κέρδους πληροφορίας



Rank	Info. gain	Attribute name
1	0.0461	HOUSE
2	0.0436	OVERAGE
3	0.0350	LONG_CALLS_PER_MON
4	0.0136	LEFTOVER
5	0.0101	INCOME
6	0.0089	AVG_CALL_DURATION
7	0.0076	HANDSET_PRICE
8	0.0003	SATISFACTION
9	0.0000	COLLEGE
10	0.0000	USAGE_LEVEL

# Παράδειγμα (συνέχ.)

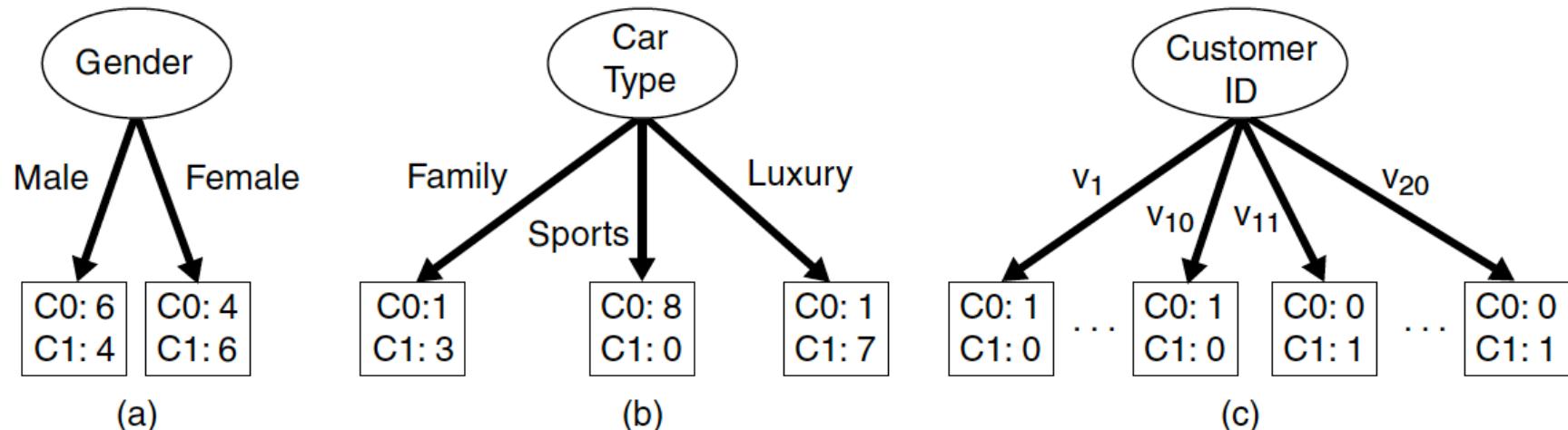
Δέντρο απόφασης  
που προέκυψε



# Περίγραμμα Μαθήματος

- Εισαγωγή στην προγνωστική μοντελοποίηση
  - Επιλογή γνωρισμάτων (**feature selection**)
    - Εντροπία (**entropy**), κέρδος πληροφορίας (**information gain**)
  - Δέντρα απόφασης (**decision trees**)
- 
- **Άλλα μέτρα επιλογής γνωρισμάτων**
    - **Gini**, αναλογία κέρδους (**gain ratio**)

# Επιλογή Γνωρισμάτων (πιο λεπτομερώς)



- Έστω  $p(i | t)$  (ή  $p_i$ ) το ποσοστό των εγγραφών που ανήκουν στην κατηγορία  $i$  στον κόμβο  $t$
- Ο διαχωρισμός (b) παράγει πιο αιμιγή χωρίσματα από τον (a)
- Άρα θέλουμε να επιλέξουμε τον καλύτερο διαχωρισμό βάσει του βαθμού ανομοιογένειας των κόμβων παιδιών

# Μέτρα Ανομοιογένειας

$$\text{Entropy}(t) = - \sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t),$$

Χρησιμοποιείται στους:  
ID3, C4.5, C5.0

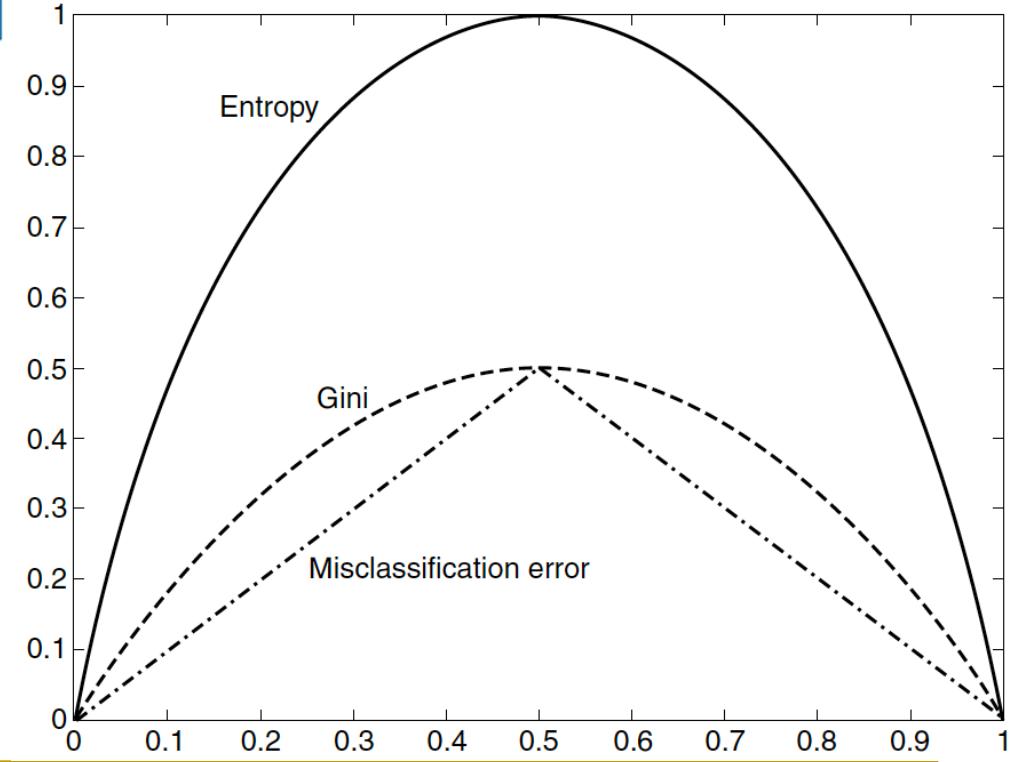
$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2,$$

Χρησιμοποιείται στον:  
CART

$$\text{Classification error}(t) = 1 - \max_i [p(i|t)]$$

Για προβλήματα δυαδικής κατηγοριοποίησης:

- ❖ Όταν  $p=0.5$ , τα μέτρα λαμβάνουν τη μέγιστη τιμή τους
- ❖ Όταν  $p=1$  ή  $p=0$ , παίρνουν τη μικρότερη τιμή τους (=όλες οι εγγραφές στην ίδια κατηγορία)



# Μέτρα Ανομοιογένειας

Node $N_1$	Count
Class=0	0
Class=1	6

$$\begin{aligned} \text{Gini} &= 1 - (0/6)^2 - (6/6)^2 = 0 \\ \text{Entropy} &= -(0/6) \log_2(0/6) - (6/6) \log_2(6/6) = 0 \\ \text{Error} &= 1 - \max[0/6, 6/6] = 0 \end{aligned}$$

Node $N_2$	Count
Class=0	1
Class=1	5

$$\begin{aligned} \text{Gini} &= 1 - (1/6)^2 - (5/6)^2 = 0.278 \\ \text{Entropy} &= -(1/6) \log_2(1/6) - (5/6) \log_2(5/6) = 0.650 \\ \text{Error} &= 1 - \max[1/6, 5/6] = 0.167 \end{aligned}$$

Node $N_3$	Count
Class=0	3
Class=1	3

$$\begin{aligned} \text{Gini} &= 1 - (3/6)^2 - (3/6)^2 = 0.5 \\ \text{Entropy} &= -(3/6) \log_2(3/6) - (3/6) \log_2(3/6) = 1 \\ \text{Error} &= 1 - \max[3/6, 3/6] = 0.5 \end{aligned}$$

Ο κόμβος  $N_1$  διαθέτει τη μικρότερη τιμή ανομοιογένειας

# Συνθήκη Ελέγχου

- Για να καθοριστεί πόσο καλά αποδίδει μια συνθήκη ελέγχου
  - Πρέπει να συγκριθεί ο **βαθμός ανομοιογένειας** του **κόμβου γονέα** με αυτόν των **κόμβων παιδιών**
- Χρησιμοποιείται η έννοια του **κέρδους Δ**:

$$\Delta = I(\text{parent}) - \sum_{j=1}^k \frac{N(v_j)}{N} I(v_j)$$

όπου:

- $I()$  το μέτρο ανομοιογένειας ενός κόμβου
- $N$  το συνολικό πλήθος εγγραφών του κόμβου γονέα
- $k$  το πλήθος τιμών ενός γνωρίσματος
- $N(v_j)$  το πλήθος εγγραφών του κόμβου  $j$

# Συνθήκη Ελέγχου

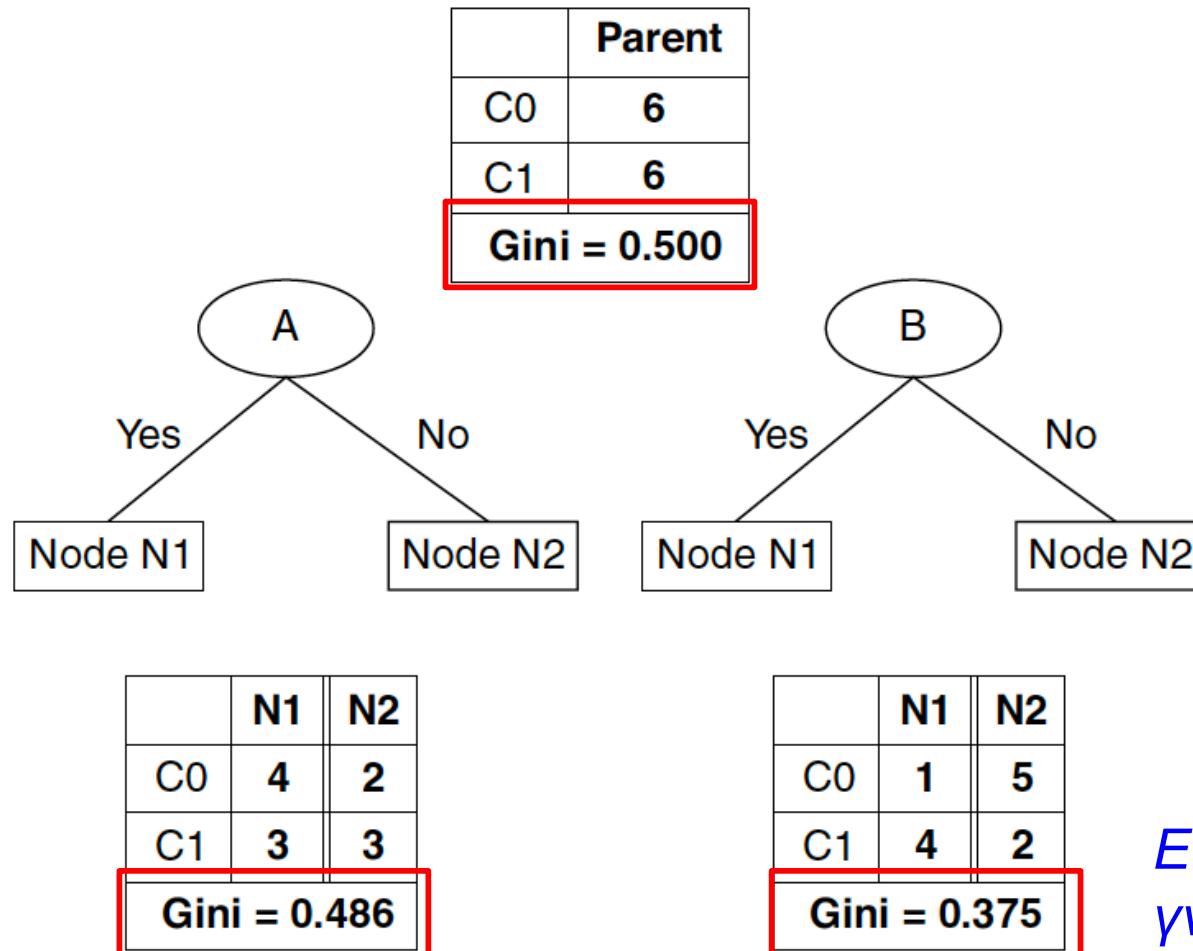
- Για να καθοριστεί πόσο καλά αποδίδει μια συνθήκη ελέγχου
  - Πρέπει να συγκριθεί ο **βαθμός ανομοιογένειας** του **κόμβου γονέα** με αυτόν των **κόμβων παιδιών**
- Χρησιμοποιείται η έννοια του **κέρδους  $\Delta$** :

$$\Delta = I(\text{parent}) - \sum_{j=1}^k \frac{N(v_j)}{N} I(v_j)$$

Δεδομένου ότι το  $I(\text{parent})$  είναι το ίδιο για όλες τις συνθήκες ελέγχου, η **μεγιστοποίηση του κέρδους** είναι ισοδύναμη με την **ελαχιστοποίηση των σταθμισμένων μέσων μέτρων ανομοιογένειας** των κόμβων παιδιών

Όταν χρησιμοποιείται η **εντροπία** ως το μέτρο ανομοιογένειας, η διαφορά στην εντροπία είναι γνωστή ως **κέρδος πληροφορίας (information gain)**

# Διαχωρισμός Δυαδικών Γνωρισμάτων

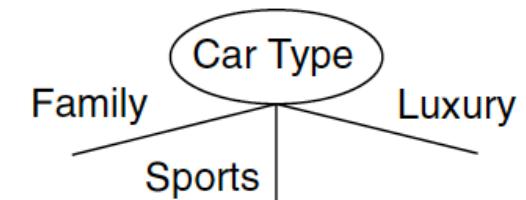
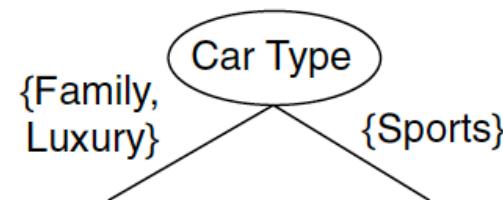
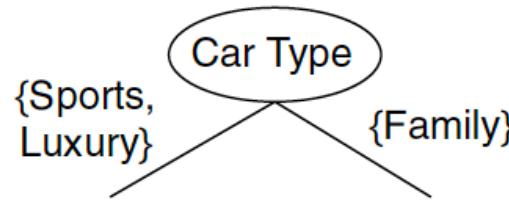


$$(7/12) \times 0.4898 + (5/12) \times 0.480 = 0.486$$

Όπου Gini(N1)=0.4898, Gini(N2)=0.480

Επιλέγεται το γνώρισμα B που παράγει μικρότερο δείκτη Gini

# Διαχωρισμός Ονομαστικών Γνωρισμάτων



Car Type		
	{Sports, Luxury}	{Family}
C0	9	1
C1	7	3
Gini	0.468	

Car Type		
	{Sports}	{Family, Luxury}
C0	8	2
C1	0	10
Gini	0.167	

(a) Binary split

Car Type			
	Family	Sports	Luxury
C0	1	8	1
C1	3	0	7
Gini	0.163		

(b) Multiway split

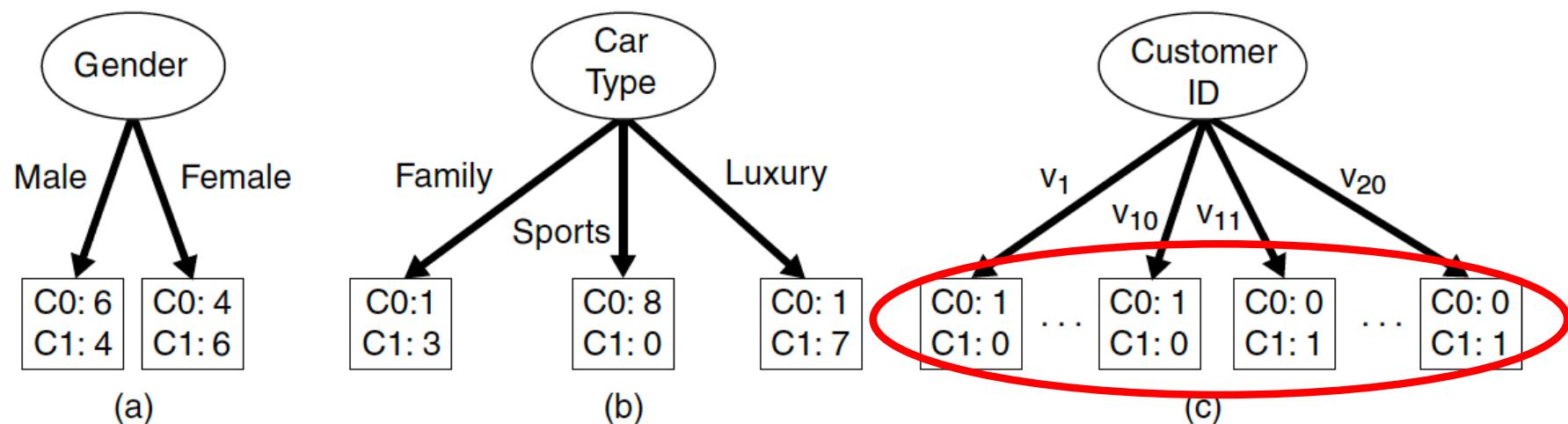
Ο διαχωρισμός πολλών κατευθύνσεων έχει μικρότερο δείκτη Gini, συγκρινόμενος και με τους δύο δυαδικούς διαχωρισμούς

Αυτό δεν πρέπει να μας εκπλήσσει, διότι:

Ο δυαδικός διαχωρισμός συγχωνεύει τα αποτελέσματα του διαχωρισμού πολλών κατευθύνσεων, άρα παράγει λιγότερο αμιγή υποσύνολα

# Αναλογία Κέρδους (Gain Ratio)

Τα μέτρα ανομοιογένειας, όπως η εντροπία και ο δείκτης Gini, τείνουν να ευνοούν τα γνωρίσματα που έχουν μεγάλο αριθμό διακριτών τιμών



- Όμως ο κωδικός πελάτη δεν είναι προγνωστικό γνώρισμα (βλ. **παράδειγμα**)
- Γενικότερα: όταν μια συνθήκη ελέγχου παράγει πάρα πολλούς κόμβους παιδιά, αυτό δεν είναι επιθυμητό, διότι πολύ λίγες εγγραφές σχετίζονται με κάθε διαμέριση, άρα οι προβλέψεις δεν είναι αξιόπιστες

# Αναλογία Κέρδους (Gain Ratio)

- Δύο στρατηγικές για την επίλυση του προβλήματος
  - Χρήση μόνο δυαδικών διαχωρισμών (π.χ. αλγόριθμος CART)
  - Τροποποίηση κριτηρίου διαχωρισμού ώστε να λαμβάνει υπόψιν το πλήθος των αποτελεσμάτων
- Ο αλγόριθμος C4.5 χρησιμοποιεί την αναλογία κέρδους ως κριτήριο διαχωρισμού

$$\text{Gain ratio} = \frac{\Delta_{\text{info}}}{\text{Split Info}}$$

$$\text{όπου } \text{Split Info} = - \sum_{i=1}^k P(v_i) \log_2 P(v_i)$$

Αν κάθε κατηγορία (από τις  $k$ ) έχει το ίδιο πλήθος εγγραφών:  $P(v_i)=1/k$   
και τότε η πληροφορία διαχωρισμού:  $\text{Split Info} = \log_2 k$

Αν ένα γνώρισμα παράγει μεγάλο πλήθος διαχωρισμών, τότε μεγάλη Split Info, και άρα μικρό Gain Ratio

# Σύνοψη για τα Δέντρα Απόφασης (1/4)

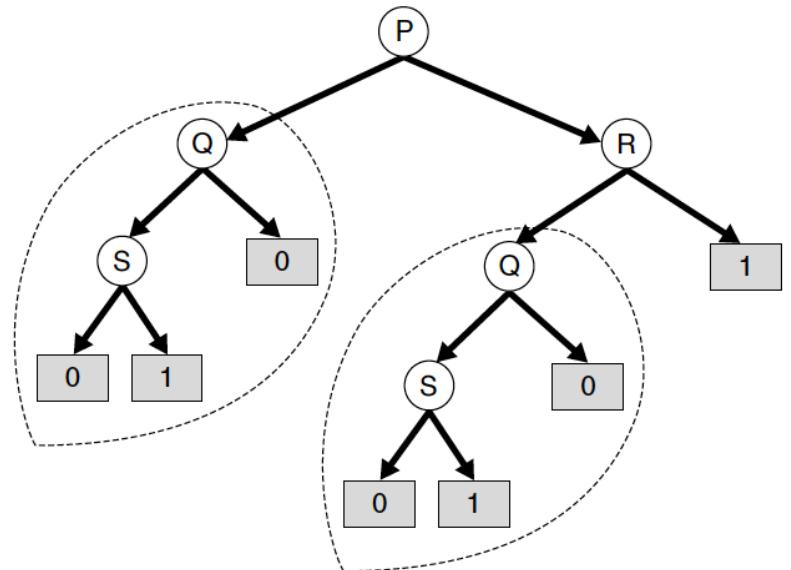
- Η επαγωγή δέντρων απόφασης είναι μια **μη παραμετρική προσέγγιση** (**non-parametric**) για τη δημιουργία μοντέλων κατηγοριοποίησης
  - Δεν απαιτούνται εκ των προτέρων υποθέσεις σχετικά με τον τύπο των κατανομών πιθανοτήτων της κατηγορίας
- Η εύρεση του **βέλτιστου** δέντρου απόφασης είναι **NP-πλήρες** πρόβλημα
  - Για αυτό το λόγο χρησιμοποιούνται **ευριστικές προσεγγίσεις** (**heuristics**), όπως άπληστοι αλγόριθμοι
  - Έτσι, η κατασκευή του δέντρου απόφασης γίνεται **φθηνή υπολογιστικά** εργασία, άρα εφαρμόσιμη σε πολύ μεγάλα σύνολα δεδομένων
  - Επίσης, αφού χτιστεί το δέντρο, η κατηγοριοποίηση μιας εγγραφής είναι πολύ γρήγορη (πολυπλοκότητα  $O(L)$ , όπου  $L$  το μέγιστο βάθος του δέντρου)

## Σύνοψη για τα Δέντρα Απόφασης (2/4)

- Τα δέντρα απόφασης είναι σχετικά **εύκολο να ερμηνευτούν**
  - Ειδικά τα μικρά σε μέγεθος δέντρα
- Η **ακρίβειά τους** είναι **συγκρίσιμη** με άλλες τεχνικές κατηγοριοποίησης
- Οι αλγόριθμοι δέντρων απόφασης είναι «**ανθεκτικοί**» στην **παρουσία θορύβου**, ειδικά όταν χρησιμοποιούνται μέθοδοι για την αποφυγή υπερπροσαρμογής
- Η **παρουσία περιττών γνωρισμάτων δεν επηρεάζει** αρνητικά την ακρίβεια των δέντρων απόφασης
  - Όμως, υπάρχει η πιθανότητα να επιλεγούν άσχετα γνωρίσματα, άρα να παραχθούν μεγαλύτερα από ότι χρειάζεται δέντρα απόφασης

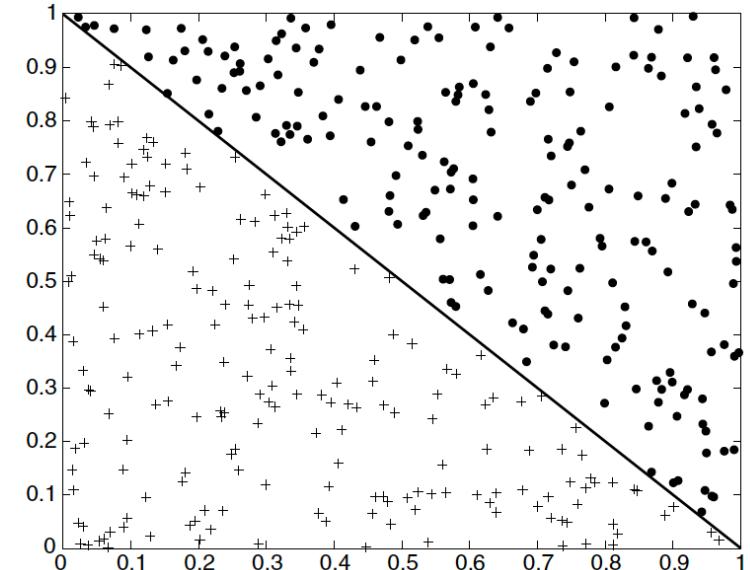
## Σύνοψη για τα Δέντρα Απόφασης (3/4)

- Μπορεί να προκύψουν κόμβοι φύλλα με μικρό πλήθος εγγραφών, που να μην επιτρέπουν μια στατιστικά σημαντική απόφαση για την αναπαράσταση κατηγοριών
  - Γνωστό ως data fragmentation
  - Πιθανή λύση: να μη συνεχίζεται ο διαχωρισμός όταν το πλήθος εγγραφών πέσει κάτω από ένα κατώφλι
- Ένα υπόδεντρο μπορεί να εμφανιστεί πολλές φορές σε ένα δέντρο απόφασης



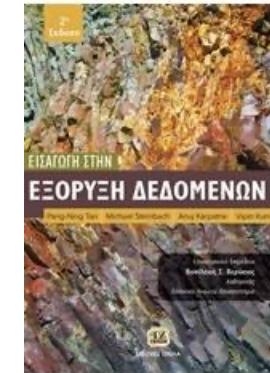
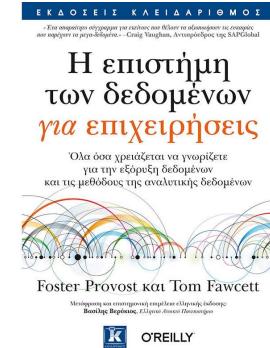
# Σύνοψη για τα Δέντρα Απόφασης (4/4)

- Τα σύνορα απόφασης (decision boundaries) είναι παράλληλα στους άξονες, γεγονός που περιορίζει την εκφραστικότητα αναπαράστασης των δέντρων απόφασης
- Η επιλογή του μέτρου ανομοιογένειας έχει μικρή επίδραση στην απόδοση των αλγορίθμων
  - Καθώς πολλά μέτρα είναι συμβατά μεταξύ τους
  - Ο τρόπος με τον οποίο γίνεται το κλάδεμα (pruning) έχει μεγαλύτερη επιρροή



# Πηγές Αναφοράς

- F. Provost, T. Faucett. “Η Επιστήμη των Δεδομένων για Επιχειρήσεις”. Εκδόσεις Κλειδάριθμος.
  - *Κεφ. 3: Εισαγωγή στην Προγνωστική Μοντελοποίηση*
- P. Tan, M. Steinbach, V. Kumar. “Εισαγωγή στην Εξόρυξη Δεδομένων”. Εκδόσεις Τζιόλα.
  - *Κεφ. 4: Κατηγοριοποίηση: Βασικές έννοιες, Δένδρα απόφασης, και Εκτίμηση μοντέλων*





## 5. Προσαρμογή Μοντέλου στα Δεδομένα

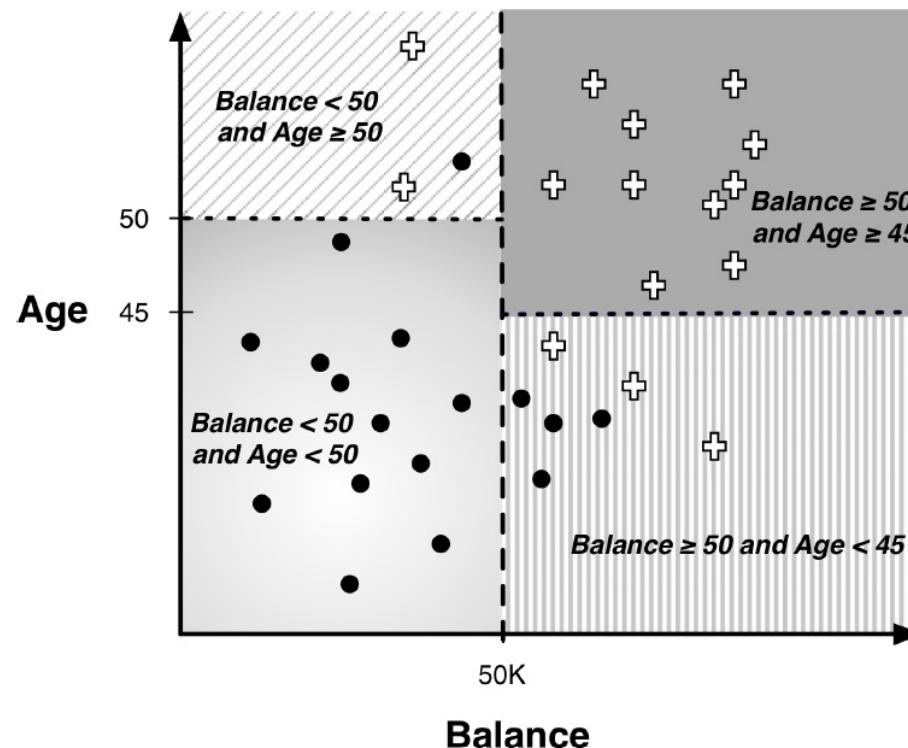


Ανάλυση Δεδομένων  
(*Data Analytics*)

Χρήστος Δουλκερίδης  
2024-25

# Δέντρα Απόφασης (προηγ. μάθημα)

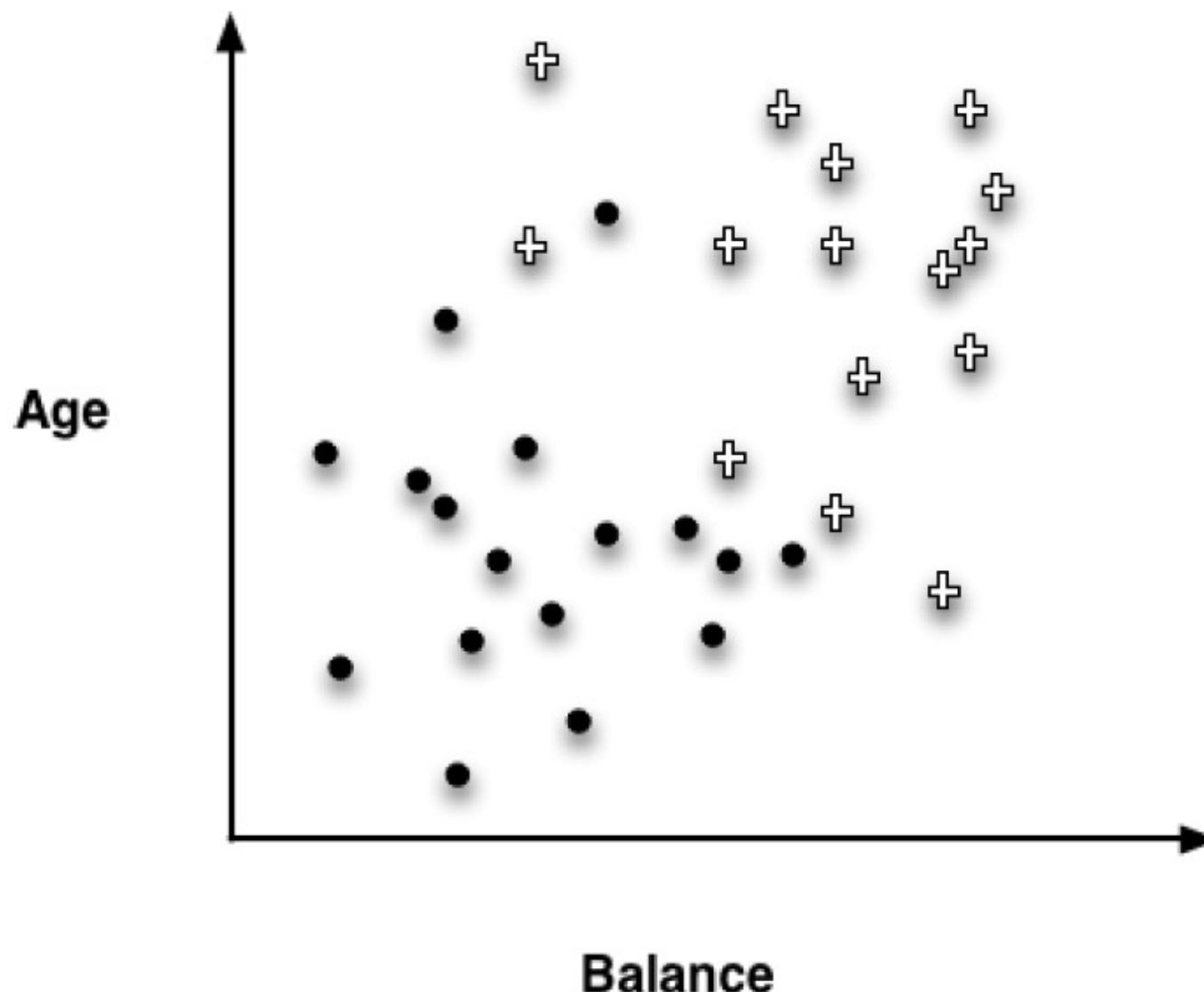
- Στα δέντρα απόφασης ιόσο η δομή του μοντέλου (δέντρο) όσο και οι αριθμητικές παράμετροι του μοντέλου (εκτιμήσεις πιθανότητας στα φύλλα) προέκυψαν από τα δεδομένα



# Εισαγωγή στην Παραμετρική Μοντελοποίηση

- Μια **άλλη μέθοδος** για την εκμάθηση ενός προγνωστικού μοντέλου από ένα σύνολο δεδομένων είναι
  - να ξεκινήσουμε **καθορίζοντας τη δομή του μοντέλου** με αριθμητικές παραμέτρους που δεν έχουν καθοριστεί
  - να **υπολογίσουμε τις βέλτιστες τιμές** των παραμέτρων, εξετάζοντας το σύνολο δεδομένων
- Παράδειγμα: η δομή του μοντέλου είναι μια παραμετροποιημένη συνάρτηση, π.χ.,  $y = a x + b$
- Αυτή η προσέγγιση ονομάζεται **μάθηση παραμέτρων (parameter learning)** ή **παραμετρική μοντελοποίηση (parametric modeling)**

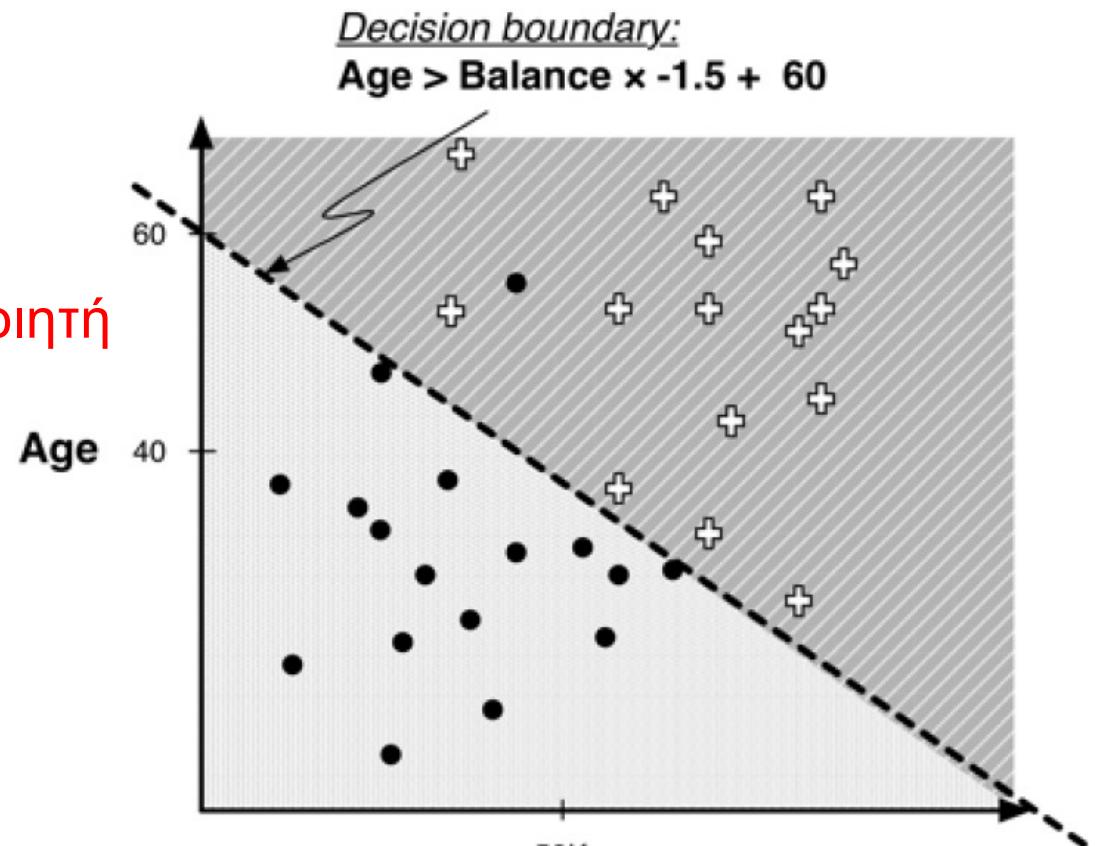
# To 'Ιδιο Σύνολο Δεδομένων



# Γραμμικός Διαχωρισμός

Παράδειγμα γραμμικού κατηγοριοποιητή  
(linear classifier)

$$Age = (-1.5) \times Balance + 60$$



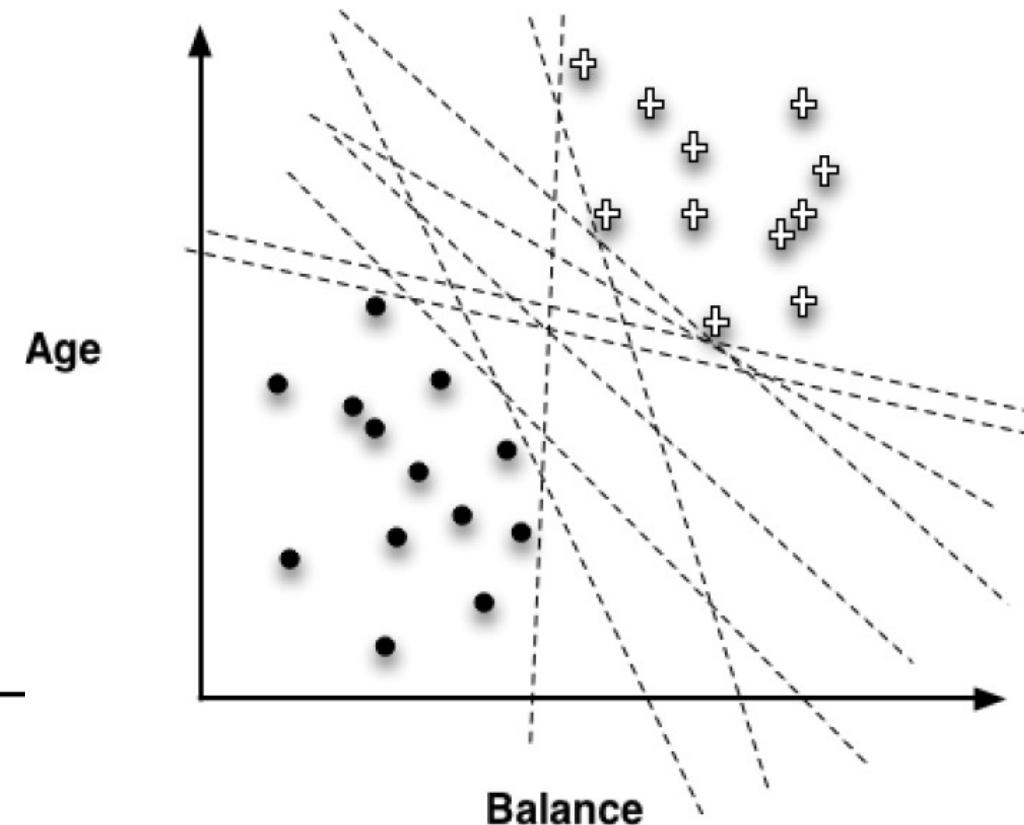
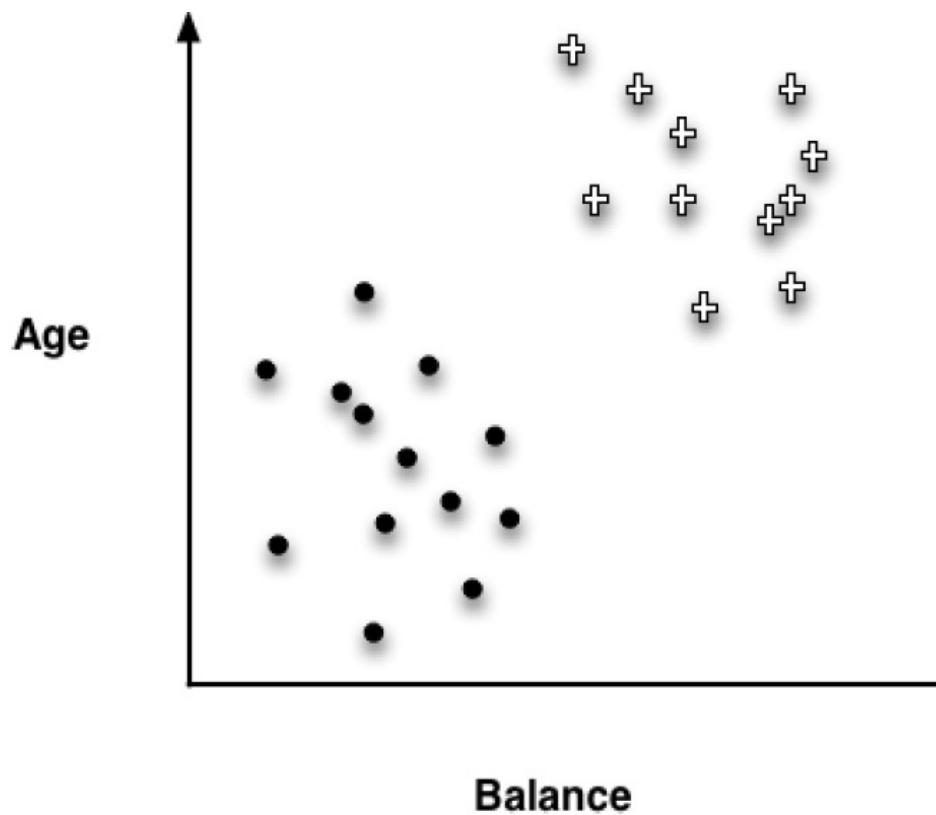
$$\text{class}(\mathbf{x}) = \begin{cases} + \text{ if } 1.0 \times Age - 1.5 \times Balance + 60 > 0 \\ • \text{ if } 1.0 \times Age - 1.5 \times Balance + 60 \leq 0 \end{cases}$$

# Παραμετροποιημένα Μοντέλα

- Γενική μορφή **γραμμικού μοντέλου**:
  - $f(x) = w_0 + w_1 x_1 + w_2 x_2 + \dots$
- Στόχος να προσαρμόσουμε το μοντέλο στα δεδομένα
  - Ήστε να διαχωρίζονται καλά τα δεδομένα εκπαίδευσης
  - Και να προβλέπεται όσο το δυνατόν ακριβέστερα η τιμή της μεταβλητής-στόχου
- Οι **συντελεστές στάθμισης  $w_i$**  είναι οι παράμετροι
  - Υποδεικνύουν τη σπουδαιότητα των γνωρισμάτων  $x_i$

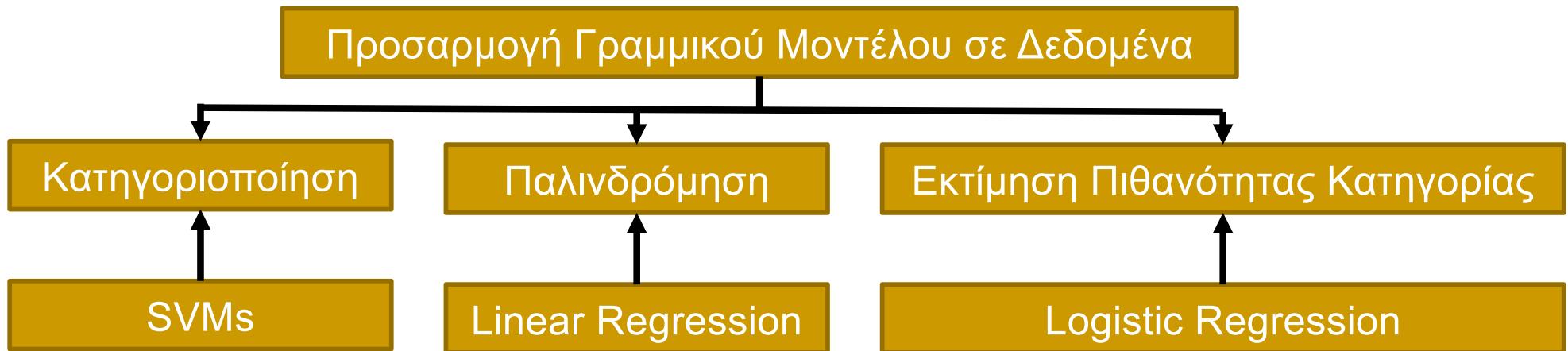
# Πολλοί Πιθανοί Γραμμικοί Διαχωριστές

Κάθε γραμμικός διαχωριστής αντιπροσωπεύει ένα **διαφορετικό μοντέλο** των δεδομένων  
**Πώς θα επιλέξουμε τους συντελεστές στάθμισης;**



# Περίγραμμα Σημερινού Μαθήματος

- Προσαρμογή μοντέλου στα δεδομένα (model fitting)
  - Γραμμική παλινδρόμηση (linear regression)
  - Μηχανές διανυσμάτων υποστήριξης (support vector machines)
  - Λογιστική παλινδρόμηση (logistic regression)



# Γραμμική Παλινδρόμηση (Linear Regression)

# Γραμμική Παλινδρόμηση (Linear Regression)

- Πρόκειται για μια τεχνική προγνωστικής μοντελοποίησης, για την **πρόβλεψη** της τιμής μιας **συνεχούς μεταβλητής**
  - Παράδειγμα: πρόβλεψη δείκτη χρηματιστηρίου
- **Παλινδρόμηση** είναι η διαδικασία εκμάθησης μιας στοχευμένης συνάρτησης **f** η οποία απεικονίζει κάθε χαρακτηριστικό **x** σε μια έξοδο συνεχών τιμών **y**
- **Στόχος**: η εύρεση μιας συνάρτησης **f** που να μπορεί να **προσαρμοστεί** στα δεδομένα εισόδου με **ελάχιστο σφάλμα**

# Το Πρόβλημα Γραμμικής Παλινδρόμησης

Δίνεται σύνολο εγγραφών:

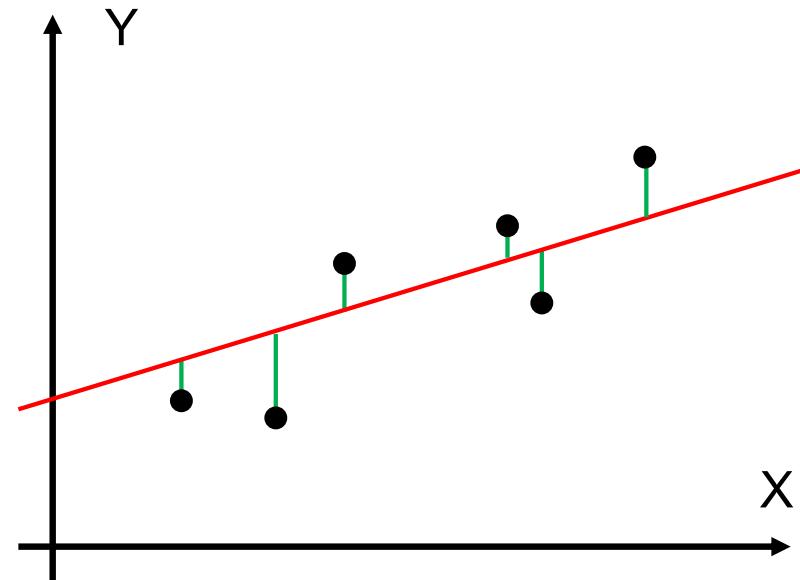
$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

Το πρόβλημα γραμμικής παλινδρόμησης αφορά στην εύρεση μια γραμμικής συνάρτησης:

$$\hat{y} = f(x) = a \cdot x + b$$

που μπορεί να χρησιμοποιηθεί για την πρόβλεψη τιμών, π.χ.

για δοθέν  $x_{n+1}$  ποια είναι η τιμή  $y_{n+1} =$ ;



# Η Μέθοδος Ελαχίστων Τετραγώνων

Ζητείται συνάρτηση ευθείας:

$$\hat{y} = f(x) = a \cdot x + b$$

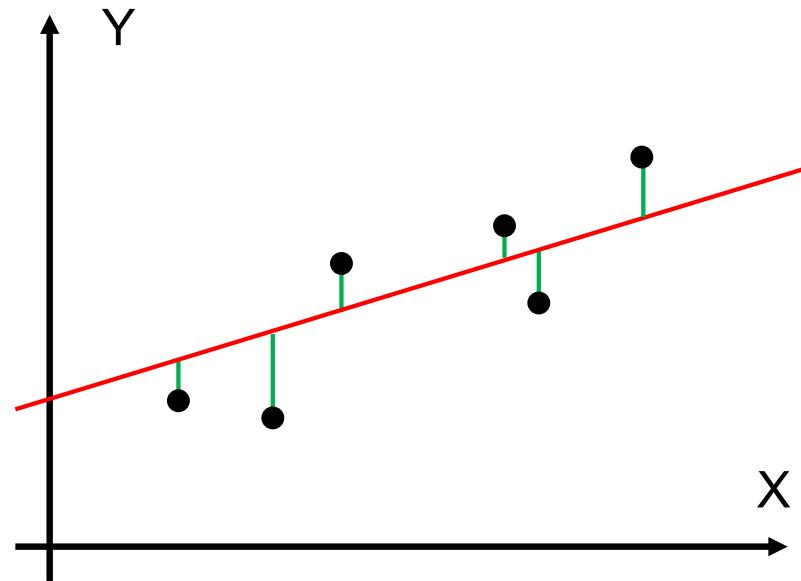
τέτοια ώστε να ελαχιστοποιηθεί το σφάλμα:

$$\hat{y} = y + \epsilon$$

και συγκεκριμένα να ελαχιστοποιηθεί το συνολικό σφάλμα  $E$ :

$$E = \sum_{i=1}^N \epsilon_i^2 = \sum_{i=1}^N (\hat{y}_i - y_i)^2$$

$$E = \sum_{i=1}^N (a \cdot x_i + b - y_i)^2$$



# Επίλυση (1/3)

Αναζητούμε τα  $a$ ,  $b$  που **ελαχιστοποιούν** τη συνάρτηση:

$$E = \sum_{i=1}^N (a \cdot x_i + b - y_i)^2$$

Απαιτούμε οι μερικές παράγωγοι ως προς  $a$  και  $b$  να είναι ίσες με μηδέν:

$$\frac{\partial E}{\partial a} = 0 \quad \frac{\partial E}{\partial b} = 0$$

και έχουμε:

$$\frac{\partial E}{\partial a} = \sum_{i=1}^N 2 \cdot x_i \cdot (a \cdot x_i + b - y_i)$$

$$\frac{\partial E}{\partial b} = \sum_{i=1}^N 2 \cdot (a \cdot x_i + b - y_i)$$

## Επίλυση (2/3)

$$\frac{\partial E}{\partial b} = \sum_{i=1}^N 2 \cdot (a \cdot x_i + b - y_i)$$

$$\sum_{i=1}^N a \cdot x_i + \sum_{i=1}^N b - \sum_{i=1}^N y_i = 0$$

$$a \cdot \sum_{i=1}^N x_i + N \cdot b - \sum_{i=1}^N y_i = 0$$

$$b = \frac{\sum_{i=1}^N y_i - a \cdot \sum_{i=1}^N x_i}{N}$$

$$b = \bar{y} - a \cdot \bar{x}$$

## Επίλυση (3/3)

$$\frac{\partial E}{\partial a} = 0$$

$$\frac{\partial E}{\partial a} = \sum_{i=1}^N 2 \cdot x_i \cdot (a \cdot x_i + b - y_i)$$

$$\sum_{i=1}^N x_i \cdot (a \cdot x_i + b - y_i) = 0$$

$$\sum_{i=1}^N (a \cdot x_i^2 + b \cdot x_i - x_i \cdot y_i) = 0$$

replace  $b$

$$\sum_{i=1}^N (a \cdot x_i^2 + (\bar{y} - a \cdot \bar{x}) \cdot x_i - x_i \cdot y_i) = 0$$

$$\sum_{i=1}^N a \cdot x_i^2 + \sum_{i=1}^N \bar{y} \cdot x_i - \sum_{i=1}^N a \cdot \bar{x} \cdot x_i - \sum_{i=1}^N x_i \cdot y_i = 0$$

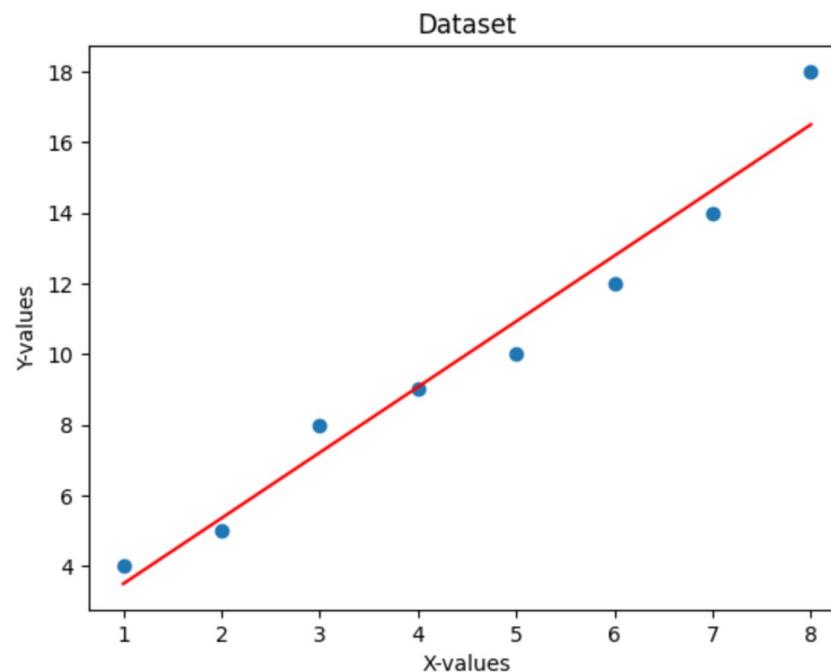
$$a = \frac{\sum_{i=1}^N x_i \cdot y_i - \bar{y} \cdot \sum_{i=1}^N x_i}{\sum_{i=1}^N x_i^2 - \bar{x} \cdot \sum_{i=1}^N x_i}$$

# 'Ασκηση

x	1	2	3	4	5	6	7	8
y	4	5	8	9	10	12	14	18

Ζητείται να εφαρμόσετε γραμμική παλινδρόμηση με τη μέθοδο ελαχίστων τετραγώνων:

$$\hat{y} = f(x) = a \cdot x + b$$



# Παρατηρήσεις στη Μέθοδο Ελαχίστων Τετραγώνων (1/3)

- Μπορεί εύκολα να δειχθεί ότι *ισοδύναμα* ισχύει:

$$a = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} = \frac{\sigma_{xy}}{\sigma_{xx}}$$

# Παρατηρήσεις στη Μέθοδο Ελαχίστων Τετραγώνων (2/3)

- Αν κανονικοποιήσουμε το  $\mathbf{x}$  διαιρώντας όλες τις τιμές του με τη διακύμανση  $\sigma_{xx}$  του  $\mathbf{x}$ :

$$x'_i = \frac{x_i}{\sigma_{xx}} \quad \text{και} \quad \bar{x}' = \frac{\bar{x}}{\sigma_{xx}}$$

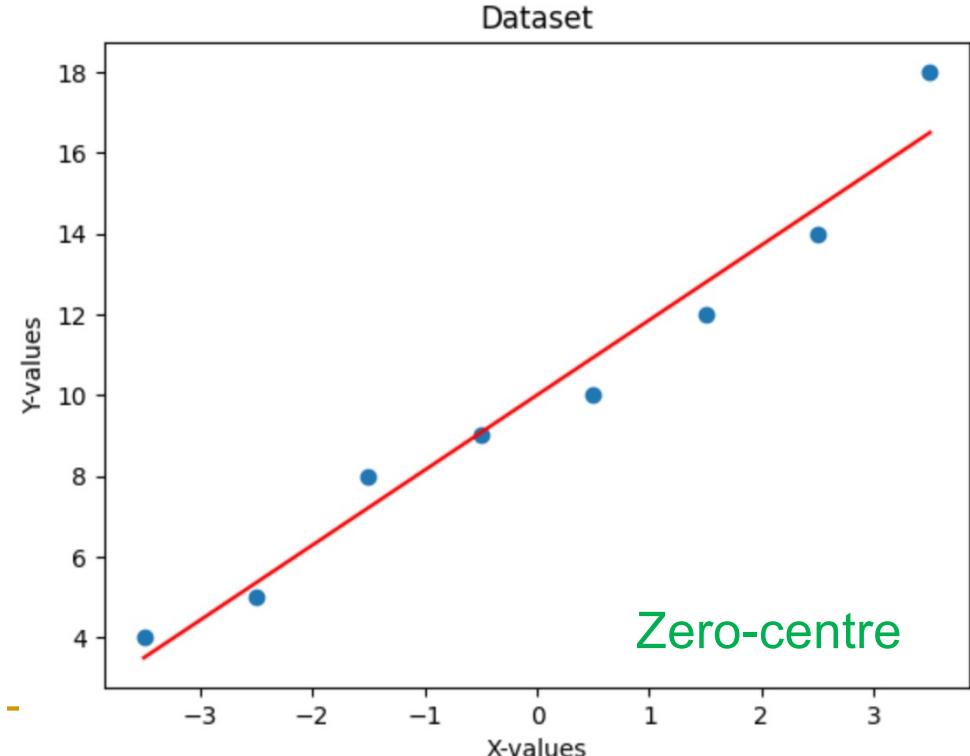
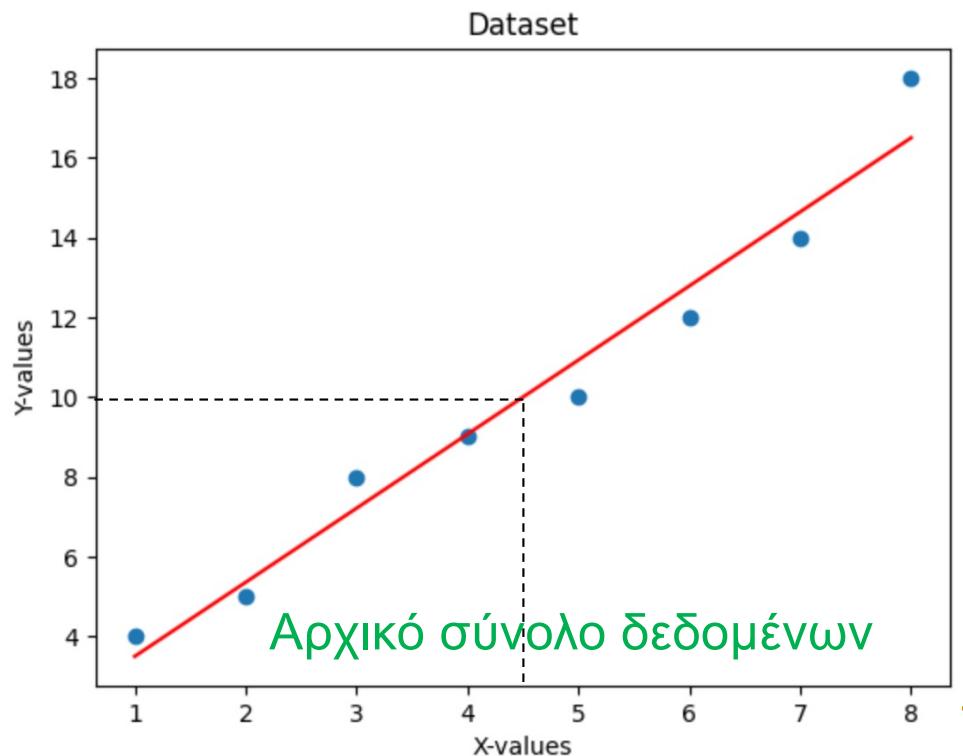
- Βρίσκουμε:

$$a = \frac{1}{N} \sum_{i=1}^N (x'_i - \bar{x}') (y_i - \bar{y}) = \sigma_{x'y}$$

- Δηλαδή ο συντελεστής παλινδρόμησης  $\alpha$  ισούται με τη συνδιακύμανση μεταξύ του **κανονικοποιημένου  $x$**  και του  **$y$**

# Παρατηρήσεις στη Μέθοδο Ελαχίστων Τετραγώνων (3/3)

- Η τιμή **b** είναι τέτοια ώστε η ευθεία να περνάει από το:  $(\bar{x}, \bar{y})$
- Μπορούμε να αφαιρέσουμε μια σταθερή ποσότητα από όλα τα  $x_i$  χωρίς να επηρεάσουμε την τιμή του συντελεστή **a**
- (zero-centre) ára μπορούμε να αφαιρέσουμε το:  $\bar{x}$

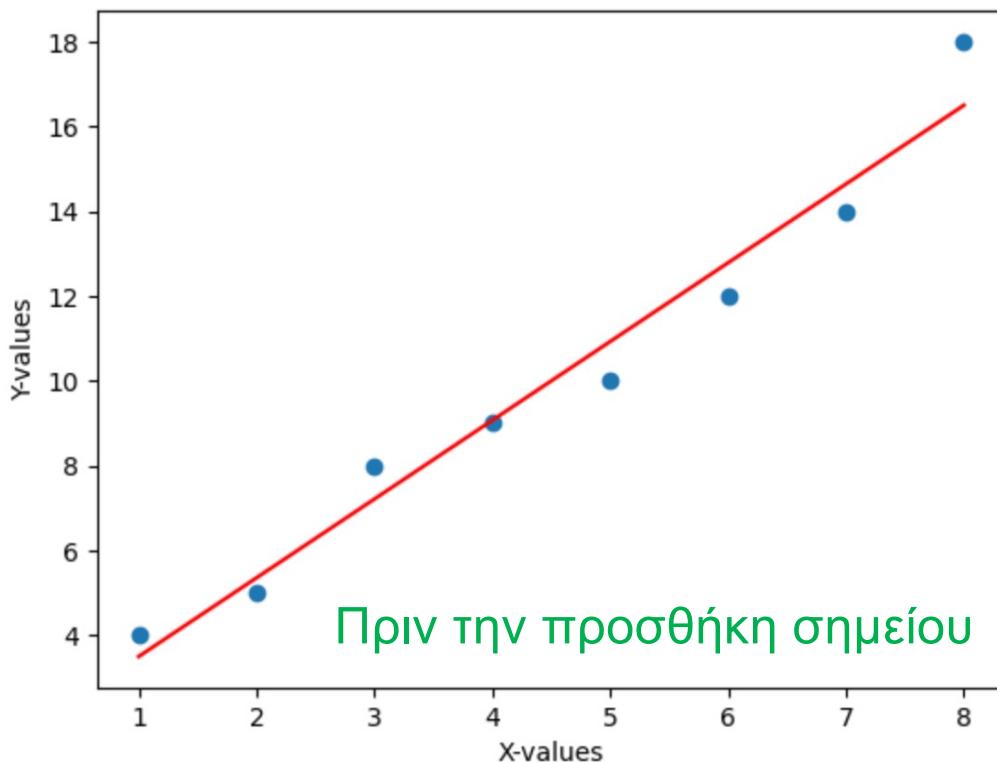


# Επιρροή Αναραίων Τιμών

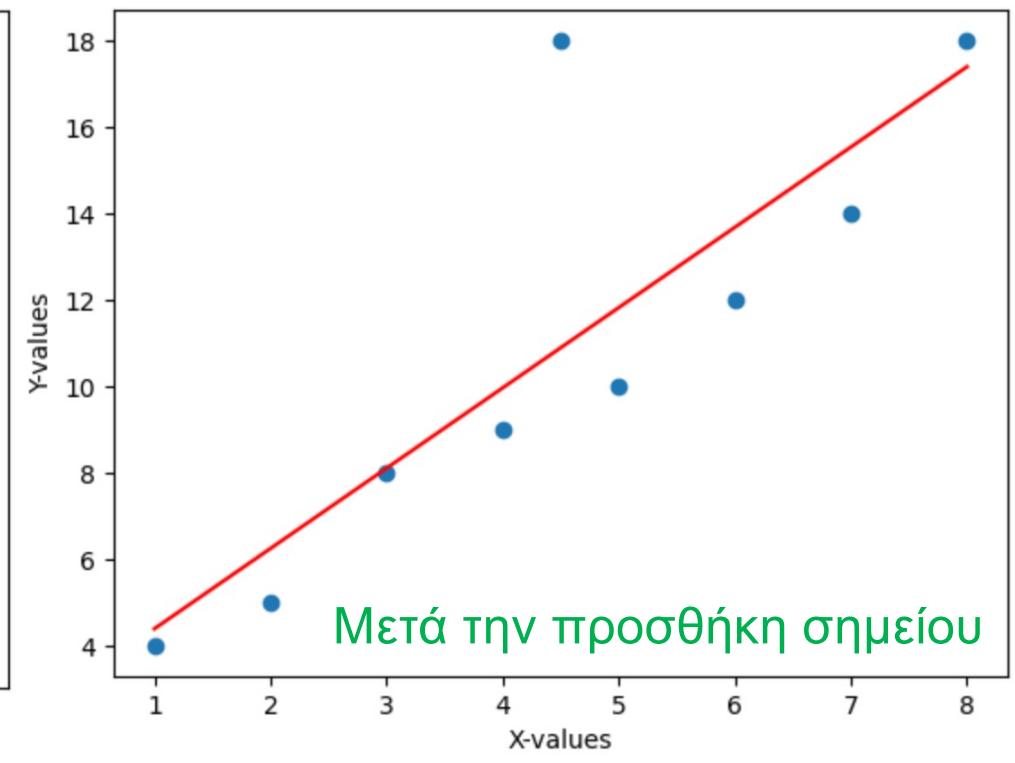
- Το άθροισμα των υπολοίπων ισούται με το μηδέν  
Αυτό κάνει τη γραμμική παλινδρόμηση ευάλωτη σε ακραίες τιμές

x	1	2	3	4	4.5	5	6	7	8
y	4	5	8	9	18	10	12	14	18

Dataset



Dataset



# Σχόλια στη Γραμμική Παλινδρόμηση

- Το πρόβλημα της παλινδρόμησης εξετάζει τη σχέση μεταξύ δύο μεταβλητών ( $x, y$ ), με σκοπό την πρόβλεψη των τιμών της μίας ( $y$ )
- Στη μέθοδος γραμμικής παλινδρόμησης χρησιμοποιείται γραμμικό μοντέλο
- Η μέθοδος γενικεύεται για πολυμεταβλητή ανάλυση:
  - $y = f(x_1, x_2, \dots, x_n) = w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n$
  - αντί  $y = f(x) = w_0 + w_1x_1$
- Η ιδέα μπορεί να γενικευτεί για μη γραμμικά μοντέλα
  - Με χρήση πολυώνυμου με βαθμό  $> 1$
  - Μη γραμμική παλινδρόμηση

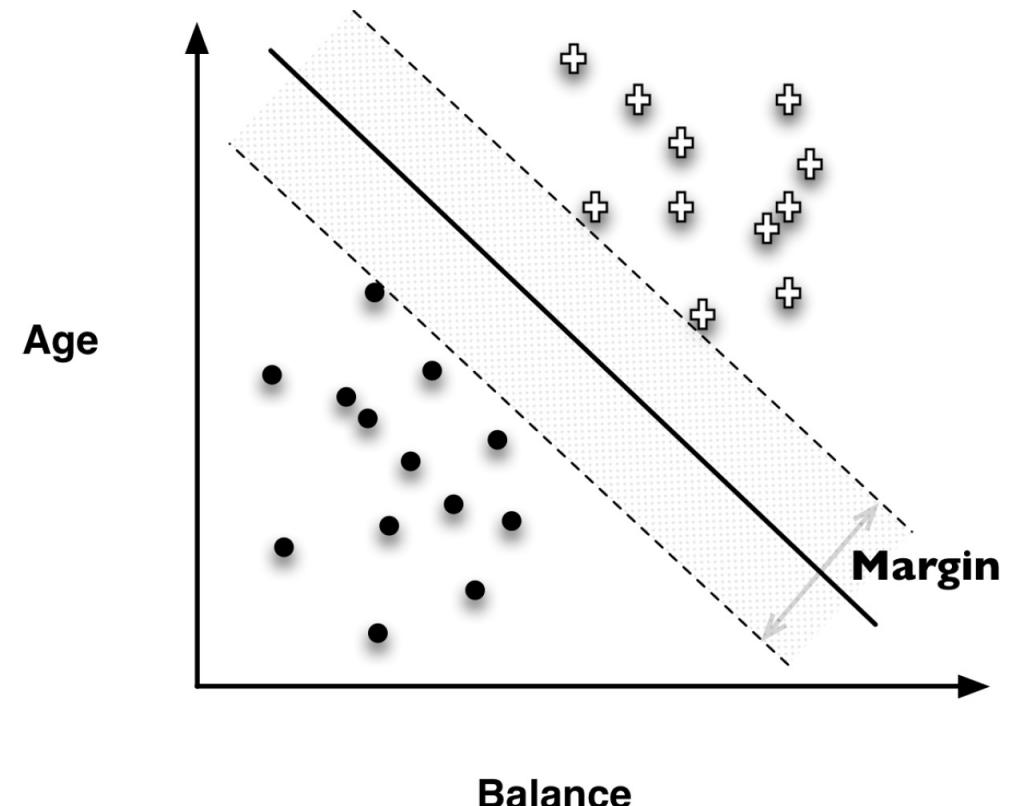
# Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

# Βελτιστοποίηση Αντικειμενικής Συνάρτησης

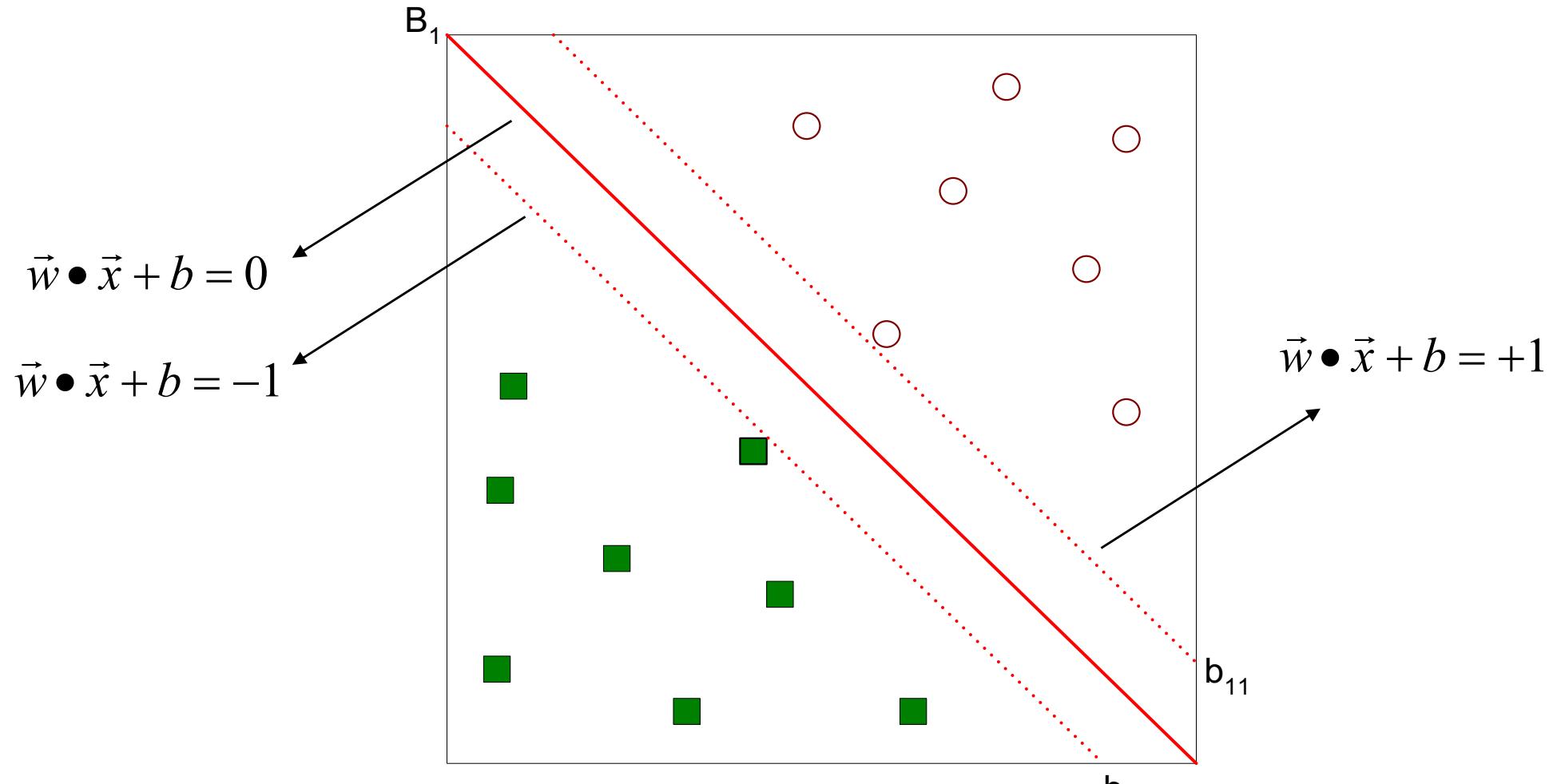
- Ορίζουμε μια **αντικειμενική συνάρτηση (objective function)** η οποία αντιπροσωπεύει το στόχο μας
  - Μπορεί να υπολογιστεί για ένα συγκεκριμένο σύνολο δεδομένων και για συγκεκριμένους συντελεστές στάθμισης
- Εύρεση της **βέλτιστης τιμής** των συντελεστών **μεγιστοποιώντας** ή **ελαχιστοποιώντας** την τιμή της αντικειμενική συνάρτησης
- Υπόθεση: οι συντελεστές θα είναι «βέλτιστοι» μόνο εφόσον η αντικειμενική συνάρτηση αντιπροσωπεύει πραγματικά τον στόχο

# Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

- Βήμα 1: εύρεση της **πιο παχιάς ράβδου** που μπορεί να σχηματιστεί μεταξύ των δειγμάτων των κατηγοριών
- Βήμα 2: ο γραμμικός διαχωριστής είναι η **κεντρική ευθεία** που περνά μέσα από τη ράβδο
- Η απόσταση μεταξύ των διακεκομένων ευθειών λέγεται **περιθώριο (margin)** και ο στόχος είναι η **μεγιστοποίησή του**



# Υπολογισμός



$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x} + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x} + b \leq -1 \end{cases}$$

$$\text{Margin} = \frac{2}{\|\vec{w}\|}$$

# Γραμμικό SVM

- Γραμμικό μοντέλο

$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x} + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x} + b \leq -1 \end{cases}$$

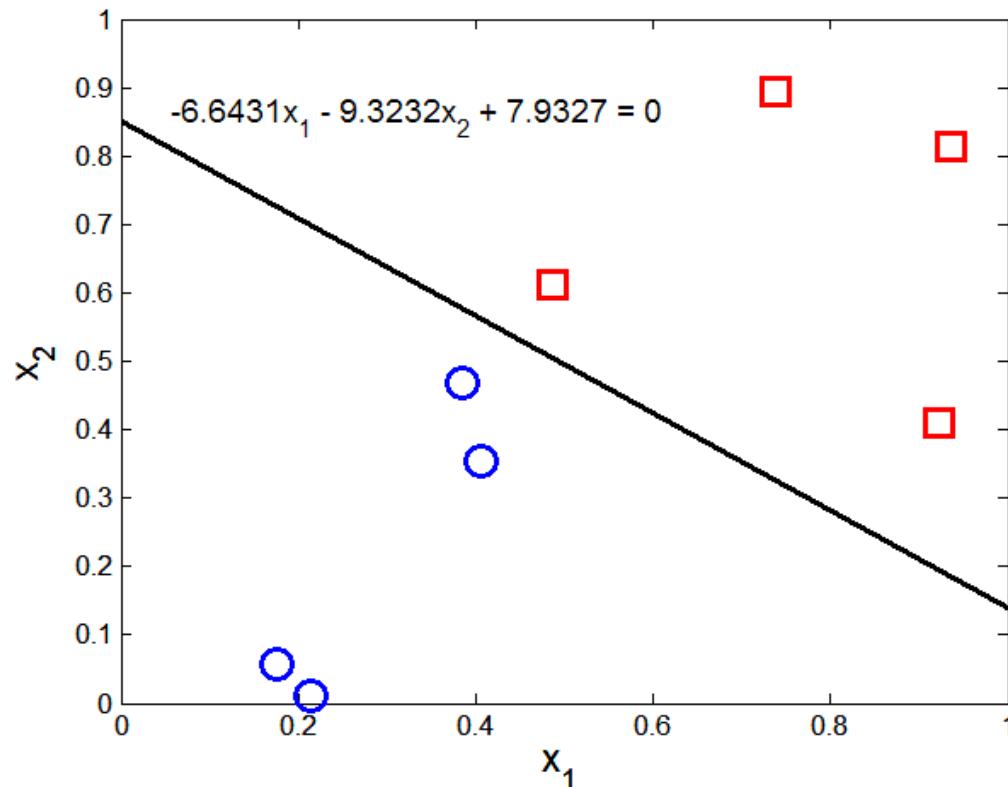
- Η εκμάθηση του μοντέλου ισοδυναμεί με την εκμάθηση των παραμέτρων **w** και **b**
- Πώς μπορούμε να μάθουμε τις παραμέτρους αυτές από τα δεδομένα;

# Εκπαίδευση Γραμμικού SVM

- Στόχος η **μεγιστοποίηση** της έκφρασης:  $\text{Margin} = \frac{2}{\|\vec{w}\|}$
- Ισοδύναμα, η ελαχιστοποίηση της έκφρασης:  $L(\vec{w}) = \frac{\|\vec{w}\|^2}{2}$   
με τους περιορισμούς:  
 $y_i = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x}_i + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x}_i + b \leq -1 \end{cases}$   
ή:  
 $y_i(\mathbf{w} \bullet \mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \dots, N$

- Κυρτό (convex) πρόβλημα βελτιστοποίησης
- Επιλύεται με τη μέθοδο του **πολλαπλασιαστή του Lagrange**(\*)  
(\*) λεπτομέρειες στο βιβλίο

# Παράδειγμα Γραμμικού SVM



Support vectors

<b>x1</b>	<b>x2</b>	<b>y</b>	<b><math>\lambda</math></b>
0.3858	0.4687	1	65.5261
0.4871	0.611	-1	65.5261
0.9218	0.4103	-1	0
0.7382	0.8936	-1	0
0.1763	0.0579	1	0
0.4057	0.3529	1	0
0.9355	0.8132	-1	0
0.2146	0.0099	1	0

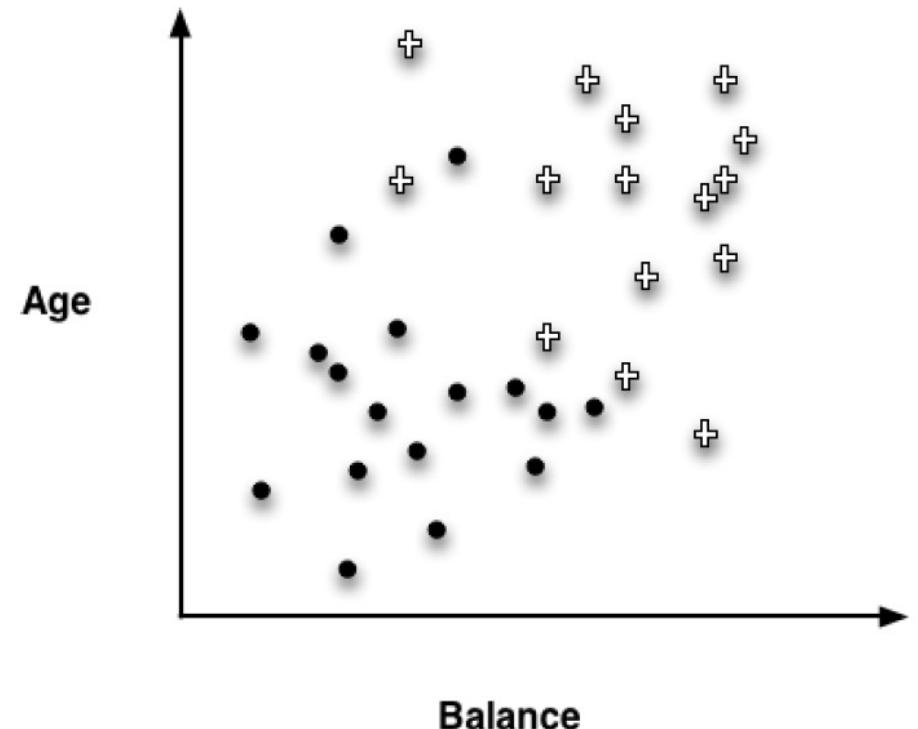
# Εκπαίδευση Γραμμικού SVM

- Το όριο απόφασης εξαρτάται **μόνο** από τα **διανύσματα υποστήριξης (support vectors)**
  - Εάν ένα σύνολο δεδομένων έχει τα ίδια διανύσματα υποστήριξης, το όριο απόφασης δεν αλλάζει
  - Πώς κατηγοριοποιείται μια νέα εγγραφή  $\vec{x}_i$  αφού έχουν υπολογιστεί τα **w** και **b**;

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x}_i + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x}_i + b \leq -1 \end{cases}$$

# Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

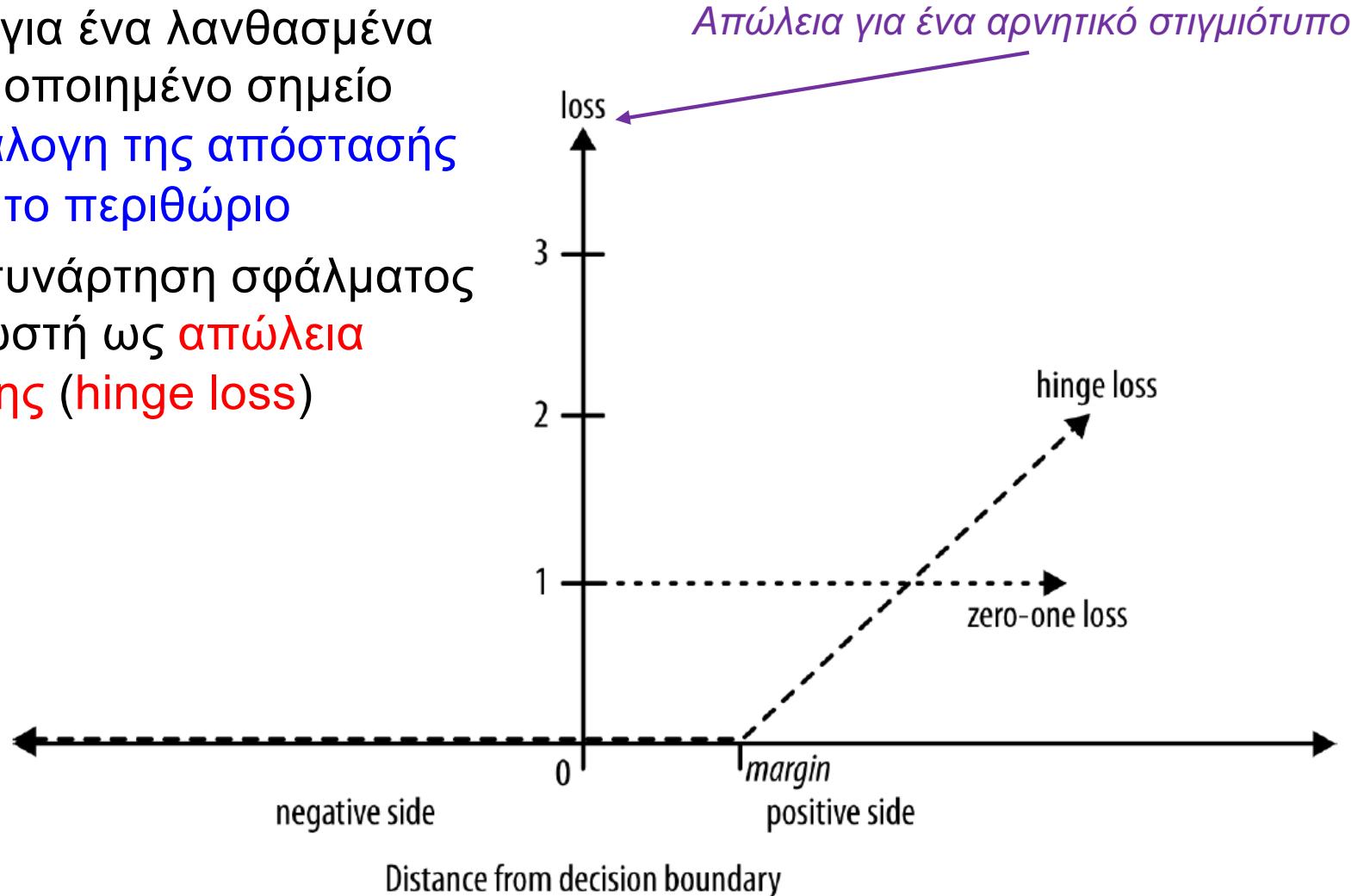
- Πώς χειρίζονται τα SVMs τα σημεία που βρίσκονται στη λάθος πλευρά του διαχωριστικού ορίου;
- Ορίζοντας στην αντικειμενική συνάρτηση μια **ποινή (penalty)** για κάθε σημείο του συνόλου εκπαίδευσης που βρίσκεται στη λάθος πλευρά
- Όταν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, η καλύτερη προσαρμογή είναι κάποιος **συμβιβασμός** μεταξύ **μεγάλου περιθωρίου** και **χαμηλής ποινής συνολικού σφάλματος**



Παράδειγμα μη-γραμμικά διαχωρίσιμου συνόλου δεδομένων

# Υπολογισμός Ποινής

- Η ποινή για ένα λανθασμένα κατηγοριοποιημένο σημείο είναι ανάλογη της απόστασής του από το περιθώριο
- Αυτή η συνάρτηση σφάλματος είναι γνωστή ως **απώλεια άρθρωσης** (hinge loss)

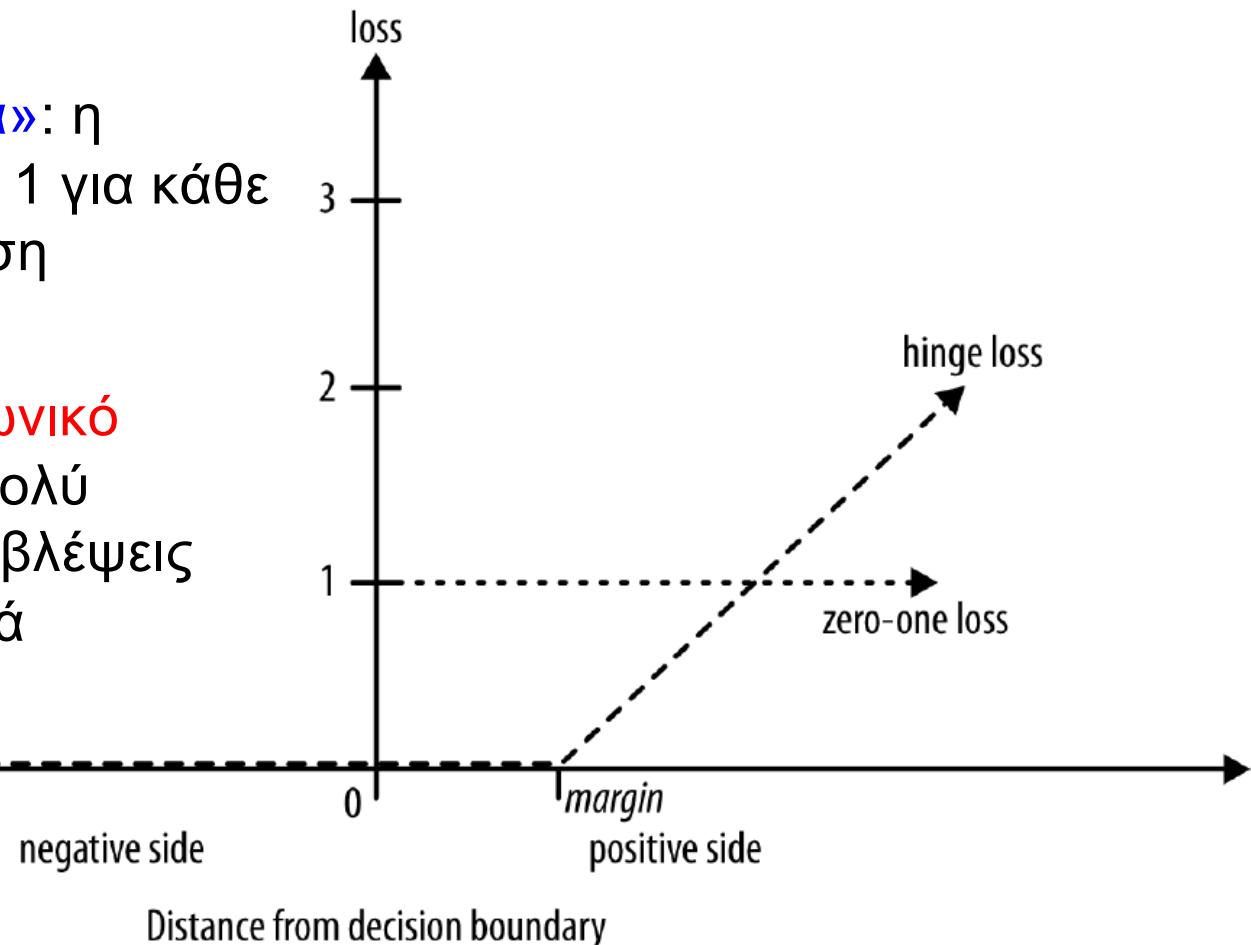


# Συναρτήσεις Απώλειας (Loss Functions)

- **Απώλεια áρθρωσης**: η απώλεια αυξάνεται γραμμικά με την απόσταση
- **Απώλεια «μηδέν-ένα»**: η απώλεια είναι ίση με 1 για κάθε λανθασμένη απόφαση
- Σύγκριση με **τετραγωνικό σφάλμα**, που δίνει πολύ υψηλές τιμές σε προβλέψεις που είναι υπερβολικά λανθασμένες

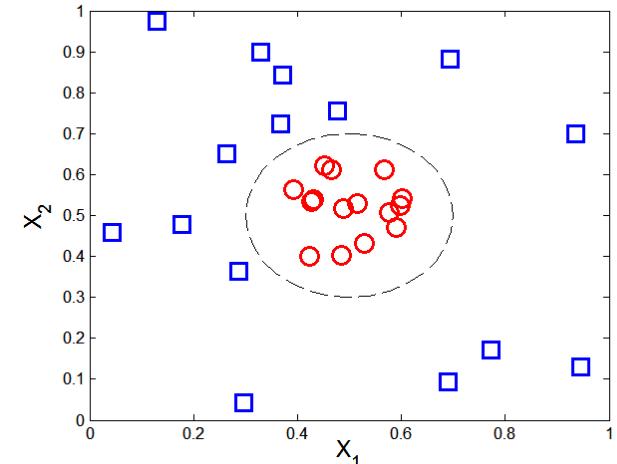


Συνήθως χρησιμοποιείται στην παλινδρόμηση, όχι στην κατηγοριοποίηση



# Γενικά για SVM

- Υπάρχει δυνατότητα να εφαρμοστούν σε σύνολα δεδομένων που έχουν μη γραμμικά όρια απόφασης (non-linear SVM) – βλ. βιβλίο
- Εντοπίζουν το **καθολικά ελάχιστο** της αντικειμενικής συνάρτησης
  - Ενώ τα δέντρα απόφασης χρησιμοποιούν μια άπληστη στρατηγική, που συχνά βρίσκει μόνον **τοπικά βέλτιστη λύση**
- Δυσκολία στην αντιμετώπιση ελλιπών τιμών
- Εύρωστη στο θόρυβο
- Υψηλή υπολογιστική πολυπλοκότητα για την κατασκευή του μοντέλου
- Για μη-γραμμικό SVM:
  - Πρέπει ο χρήστης να καθορίσει την παράμετρο  $C$  και τη συνάρτηση πυρήνα



# Λογιστική Παλινδρόμηση (Logistic Regression)

# Λογιστική Παλινδρόμηση (Logistic Regression)

- Τι κάνουμε όταν μας ενδιαφέρει η **πιθανότητα** να ανήκει ένα νέο στιγμιότυπο σε μια **κατηγορία**;
- Μπορούμε να χρησιμοποιήσουμε την ίδια **μεθοδολογία προσαρμογής γραμμικών μοντέλων** σε δεδομένα
  - Επιλέγοντας διαφορετική αντικειμενική συνάρτηση
- Η πιο συνηθισμένη διαδικασία για να το κάνουμε αυτό είναι η **λογιστική παλινδρόμηση** (logistic regression)

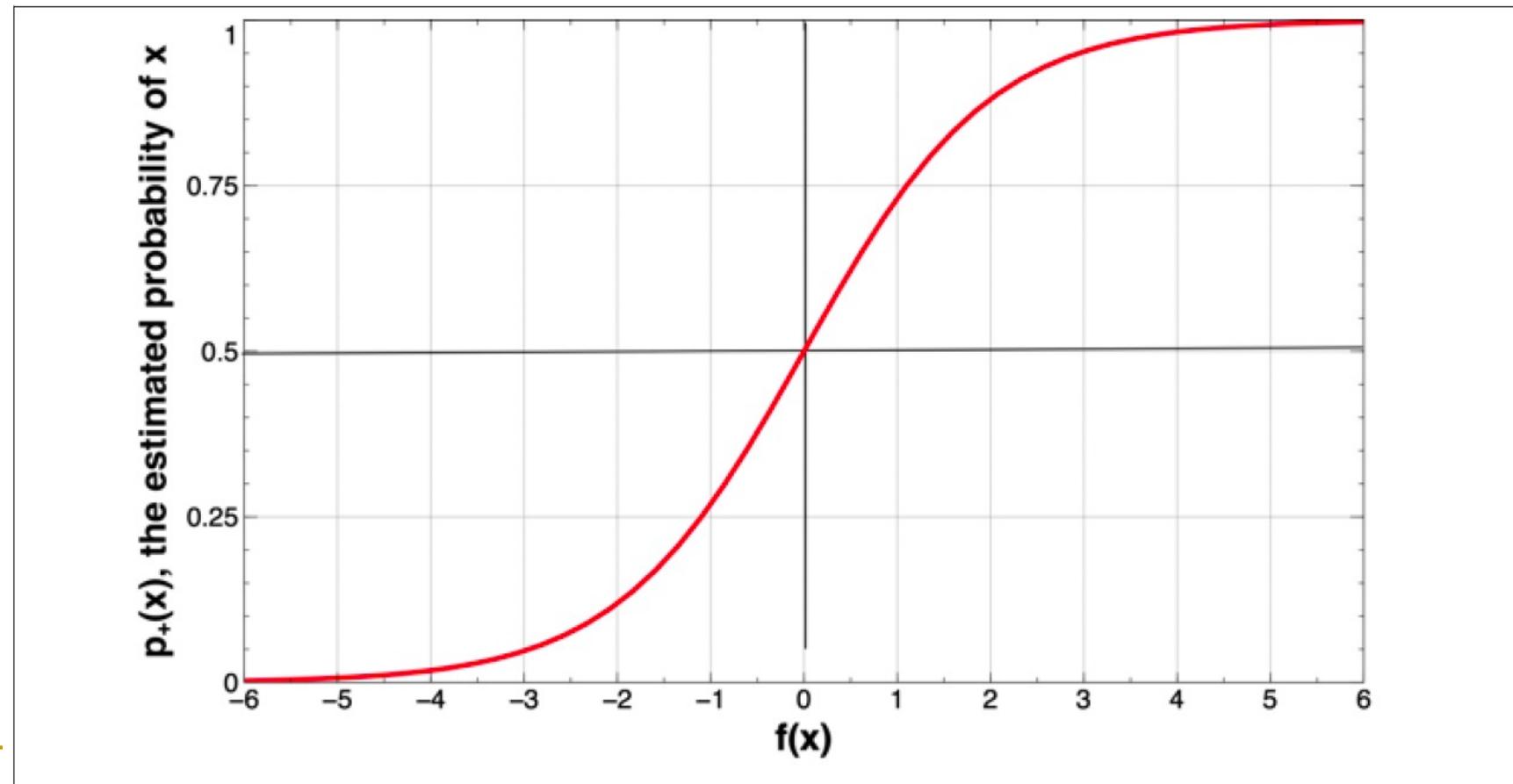
# Η Έννοια της Απόδοσης (Odd)

- Η απόδοση (odd) ενός ενδεχομένου είναι ο λόγος της πιθανότητας να λάβει χώρα το ενδεχόμενο προς την πιθανότητα να μη λάβει χώρα
  - Παράδειγμα:  $p=70\%$  σημαίνει  $odd=70/30=7/3$
- Εύρεση ενός γραμμικού διαχωριστή που δίνει την απόδοση
  - Επιλέγουμε το λογάριθμο των αποδόσεων (log-odds)
  - Η γραμμική συνάρτηση  $f(x)$  είναι η εκτίμηση του μοντέλου για το λογάριθμο της απόδοσης ότι το στιγμιότυπο  $x$  ανήκει στην κατηγορία

Probability	Odds	Log-odds
0.5	50:50 or 1	0
0.9	90:10 or 9	2.19
0.999	999:1 or 999	6.9
0.01	1:99 or 0.0101	-4.6
0.001	1:999 or 0.001001	-6.9

# Λογιστική Παλινδρόμηση

$$\log \left( \frac{p_+(\mathbf{x})}{1 - p_+(\mathbf{x})} \right) = f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \dots \rightarrow p_+(\mathbf{x}) = \frac{1}{1 + e^{-f(\mathbf{x})}}$$



# Λογιστική Παλινδρόμηση

- Η **αντικειμενική συνάρτηση** υπολογίζει την πιθανοφάνεια ενός παραδείγματος να ανήκει στη σωστή κατηγορία

$$g(\mathbf{x}, \mathbf{w}) = \begin{cases} p_+(\mathbf{x}) & \text{if } \mathbf{x} \text{ is a +} \\ 1 - p_+(\mathbf{x}) & \text{if } \mathbf{x} \text{ is a •} \end{cases}$$

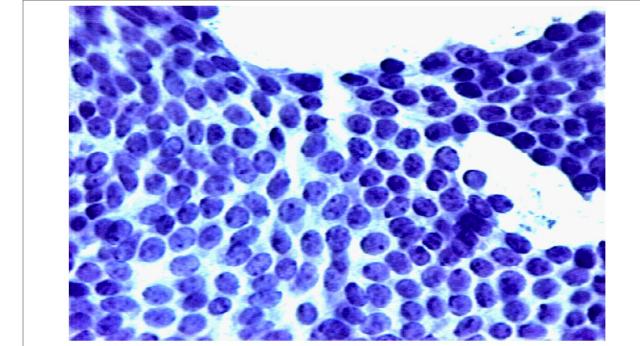
- Το μοντέλο (δλδ. το σύνολο συντελεστών στάθμισης  $\mathbf{w}$ ) που δίνει το **υψηλότερο άθροισμα** των τιμών της  $g()$  για όλα τα παραδείγματα είναι το **μοντέλο μέγιστης πιθανοφάνειας**

# Λογιστική Παλινδρόμηση – Βασικά Στοιχεία

- Για την εκτίμηση πιθανότητας, στη λογιστική παλινδρόμηση χρησιμοποιείται το ίδιο γραμμικό μοντέλο όπως στους γραμμικούς διαχωριστές (για την κατηγοριοποίηση) και στη γραμμική παλινδρόμηση (για την εκτίμηση αριθμητικών τιμών)
- Το αποτέλεσμα του μοντέλου λογιστικής παλινδρόμησης ερμηνεύεται ως ο λογάριθμος της απόδοσης για το ενδεχόμενο συμμετοχής στην κατηγορία
- Αυτές οι λογαριθμικές αποδόσεις μπορούν να μεταφραστούν απευθείας στην πιθανότητα συμμετοχής στην κατηγορία

# Παράδειγμα: Wisconsin Breast Cancer Dataset

- Περιγραφή των χαρακτηριστικών καρκινικών κυττάρων σε περιπτώσεις καρκίνου του μαστού στην πολιτεία του Wisconsin
- #Στιγμιοτύπων: 357 καλοήθη, 212 κακοήθη
- Εξαγωγή 10 χαρακτηριστικών από κάθε εικόνα
  - Παραγωγή 30 χαρακτηριστικών (από τα 10):
    - μέση τιμή ([mean](#))
    - τυπικό σφάλμα ([se](#))
    - μέση τιμή τριών μεγαλύτερων ([worst](#))



Attribute name	Description
RADIUS	<i>Mean of distances from center to points on the perimeter</i>
TEXTURE	<i>Standard deviation of grayscale values</i>
PERIMETER	<i>Perimeter of the mass</i>
AREA	<i>Area of the mass</i>
SMOOTHNESS	<i>Local variation in radius lengths</i>
COMPACTNESS	<i>Computed as: perimeter<sup>2</sup>/area – 1.0</i>
CONCAVITY	<i>Severity of concave portions of the contour</i>
CONCAVE POINTS	<i>Number of concave portions of the contour</i>
SYMMETRY	<i>A measure of the nuclei's symmetry</i>
FRACTAL DIMENSION	<i>'Coastline approximation' – 1.0</i>
DIAGNOSIS (Target)	<i>Diagnosis of cell sample: malignant or benign</i>

[http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(diagnostic\)](http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic))

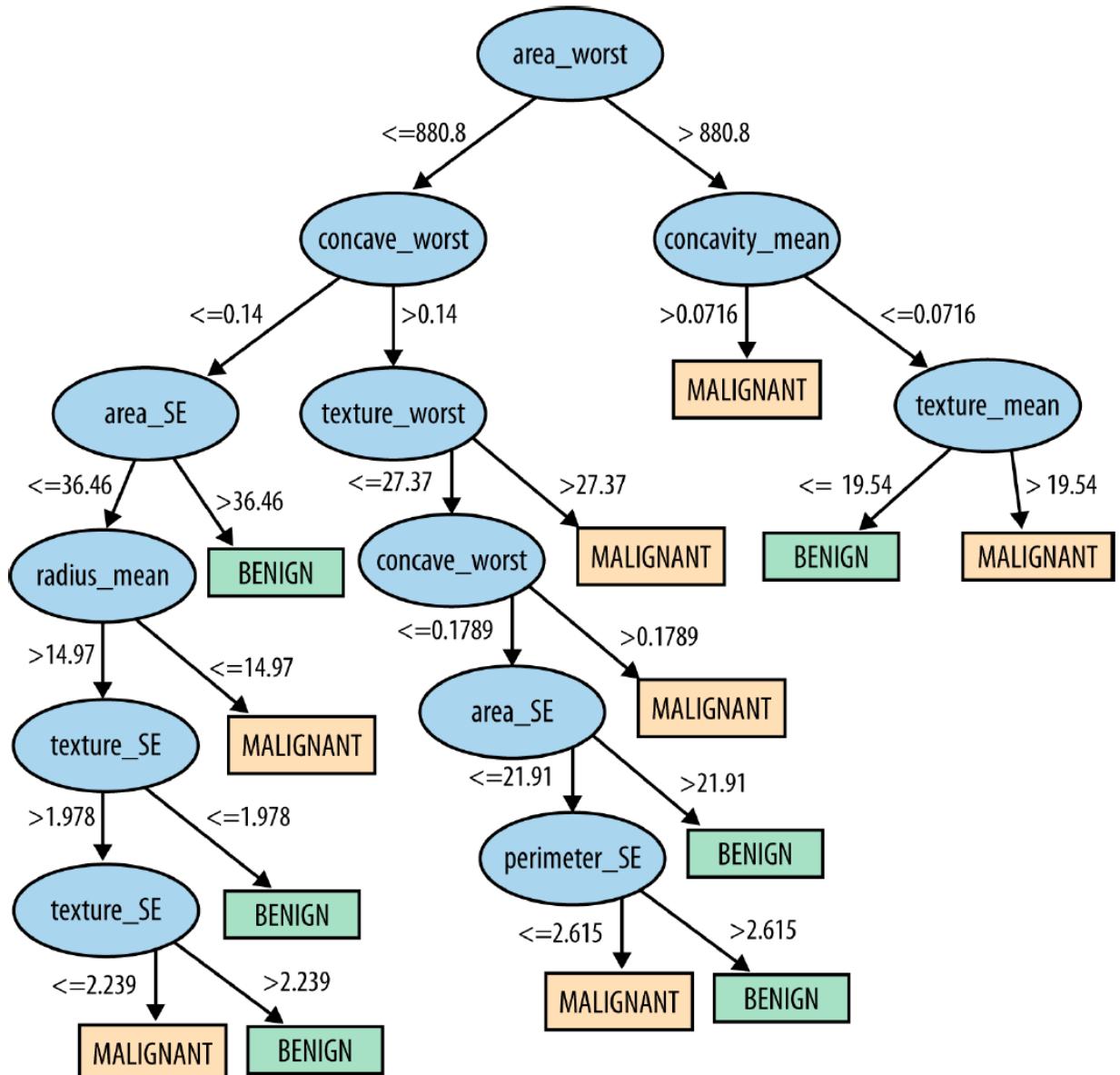
# Γραμμική Εξίσωση (Λογιστική Παλινδρόμηση)

- Μη μηδενικοί συντελεστές στάθμισης γραμμικού μοντέλου, ταξινομημένοι από τον υψηλότερο στον χαμηλότερο
- Ακρίβεια = 98.9% (μόνο 6 λάθη)

Attribute	Weight (learned parameter)
SMOOTHNESS_worst	22.3
CONCAVE_mean	19.47
CONCAVE_worst	11.68
SYMMETRY_worst	4.99
CONCAVITY_worst	2.86
CONCAVITY_mean	2.34
RADIUS_worst	0.25
TEXTURE_worst	0.13
AREA_SE	0.06
TEXTURE_mean	0.03
TEXTURE_SE	-0.29
COMPACTNESS_mean	-7.1
COMPACTNESS_SE	-27.87
$w_0$ (intercept)	-17.7

# Δέντρο Κατηγοριοποίησης

25 κόμβοι συνολικά  
13 κόμβοι-φύλλα  
Ακρίβεια = 99.1%

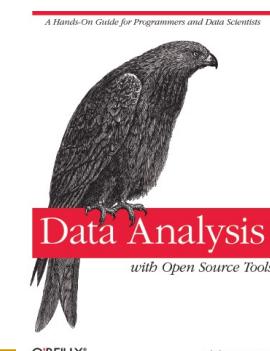
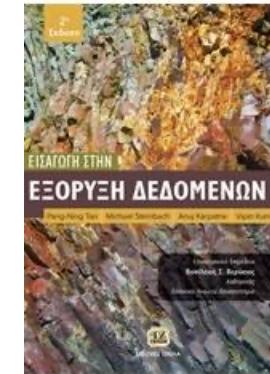
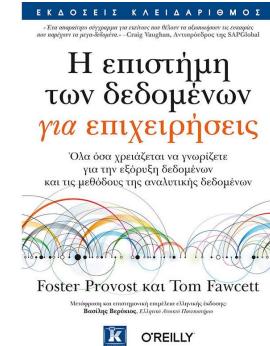


# Σύνοψη

- Η γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση και οι μηχανές διανυσμάτων υποστήριξης είναι πολύ παρόμοιες περιπτώσεις μιας θεμελιώδους τεχνικής
  - **Της προσαρμογής ενός (γραμμικού) μοντέλου στα δεδομένα**
- Η βασική τους διαφορά είναι η χρήση διαφορετικής αντικειμενικής συνάρτησης

# Πηγές Αναφοράς

- F. Provost, T. Fawcett. “Η Επιστήμη των Δεδομένων για Επιχειρήσεις”. Εκδόσεις Κλειδάριθμος.
  - *Κεφ. 4: Η προσαρμογή ενός μοντέλου στα δεδομένα*
- P. Tan, M. Steinbach, V. Kumar. “Εισαγωγή στην Εξόρυξη Δεδομένων”. Εκδόσεις Τζιόλα.
  - *Κεφ. 4: Κατηγοριοποίηση: Εναλλακτικές Τεχνικές*
- P.K. Janert. “Data Analysis with Open Source Tools”. O'Reilly, 2011.
  - *Κεφ. 18: Predictive Analytics*





## 6. Κατηγοριοποίηση, Αξιολόγηση Μοντέλων και Υπερπροσαρμογή



---

**Ανάλυση Δεδομένων  
(Data Analytics)**

Χρήστος Δουλκερίδης  
2024-25

# Περίγραμμα Μαθήματος

- Άλλες μέθοδοι κατηγοριοποίησης
  - Κατηγοριοποίηση βάσει k-κοντινότερων γειτόνων (k-nearest neighbors)
  - Κατηγοριοποιητής Bayes (Bayes classifier)
- Αξιολόγηση μοντέλων
  - Μέτρα αξιολόγησης κατηγοριοποιητών
    - Accuracy, confusion matrix, precision, recall, F-measure
  - Οπτικοποιήσεις της απόδοσης κατηγοριοποιητών
    - Καμπύλη ROC, Area Under Curve (AUC), Lift curve
- Η έννοια της υπερπροσαρμογής (overfitting)
  - Αναγνώριση υπερπροσαρμογής
  - Γράφημα προσαρμογής (fitting graph)
  - Παρακράτηση δεδομένων (hold out data)
  - Διασταυρωτική επικύρωση (cross-validation)
  - Καμπύλη μάθησης (learning curve)

# Υπενθύμιση: Το Πρόβλημα Κατηγοριοποίησης

Δεδομένου ενός συνόλου δεδομένων (εκπαίδευσης) με γνωστή τιμή-στόχο:

A.M.	Γνώση Μαθηματικών	Γνώση Προγραμματισμού	Βαθμός Εισαγωγής	Ολοκλήρωση σε 4 έτη
19001	Μέτρια	Άριστη	17.5	N
18002	Καλή	Καλή	16.9	O
19003	Πολύ καλή	Καλή	18.0	N
18004	Καλή	Μέτρια	16.8	O
17005	Άριστη	Άριστη	17.7	N
19006	Πολύ καλή	Μέτρια	17.2	N
19007	Μέτρια	Μέτρια	17.5	O

Ζητείται η κατασκευή ενός μοντέλου που θα επιτρέπει την **πρόβλεψη τιμών-στόχου**:

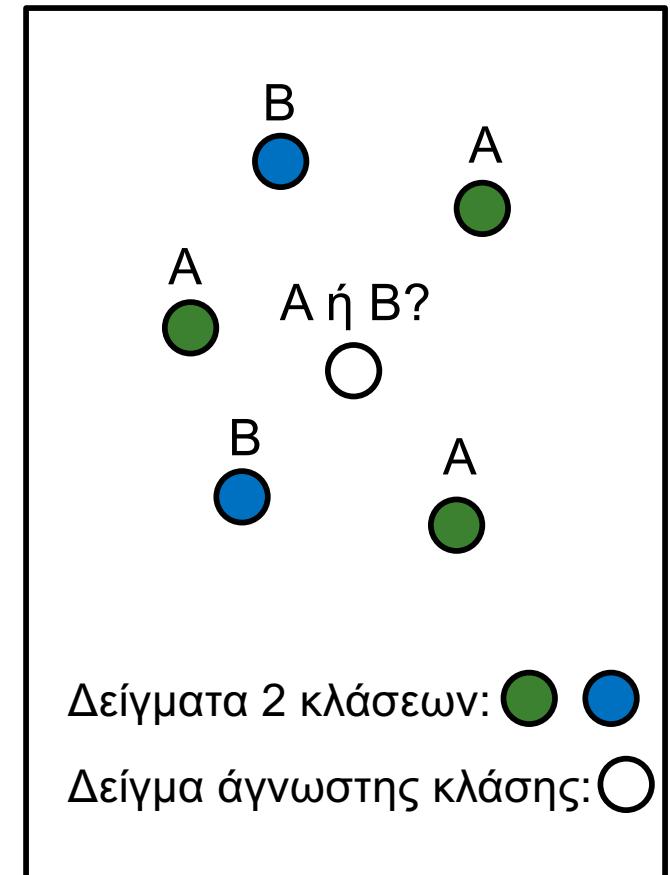
22001	Άριστη	Πολύ καλή	17.8	
-------	--------	-----------	------	--

# Κατηγοριοποίηση k-κοντινότερων γειτόνων (k-nearest neighbors)

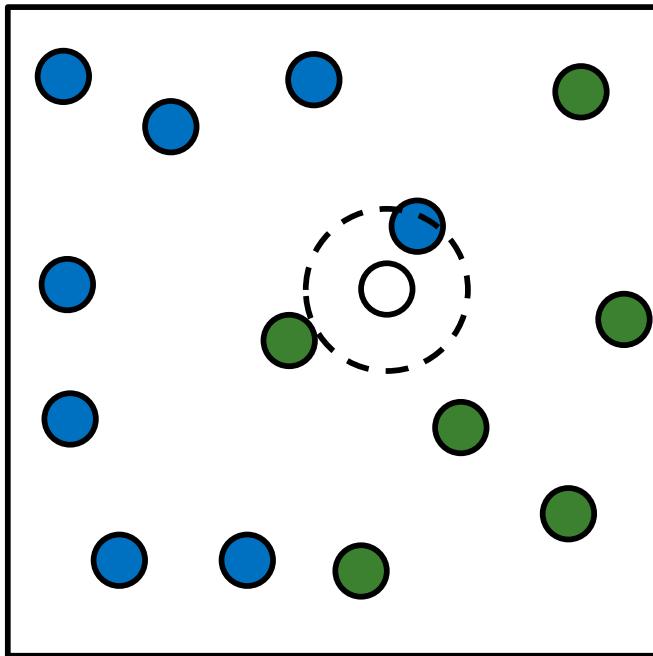
---

# Κατηγοριοποιητές Κοντινότερου Γείτονα (Nearest-neighbor Classifiers)

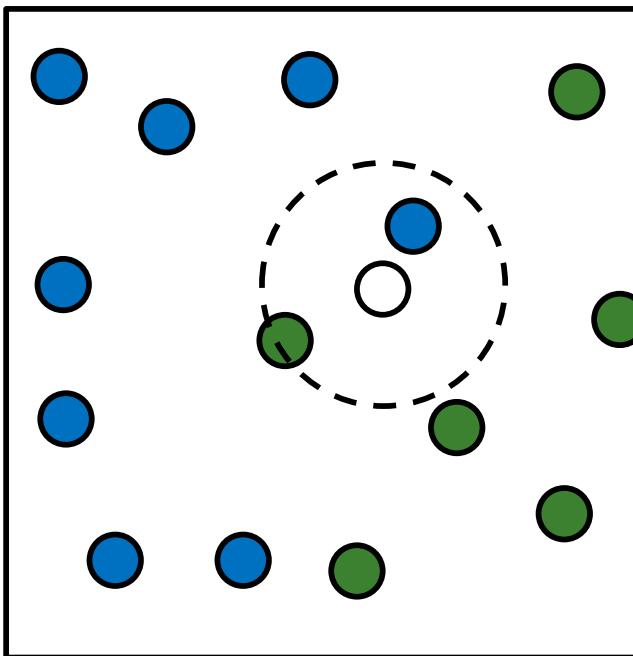
- Βασική ιδέα: για να κατηγοριοποιήσουμε ένα δείγμα άγνωστης κλάσης:
  - βρίσκουμε ένα υπάρχον δείγμα που είναι **το πιο όμοιο** (*ισοδύναμα: πιο κοντινό*) με το νέο δείγμα,
  - και αναθέτουμε την ετικέτα του
- Η βασική ιδέα μπορεί να γενικευτεί:
  - Με χρήση των **K πιο όμοιων στιγμιότυπων** και ανάθεση της ετικέτας της πλειοψηφίας
  - Με χρήση κατάλληλου **μέτρου απόστασης**



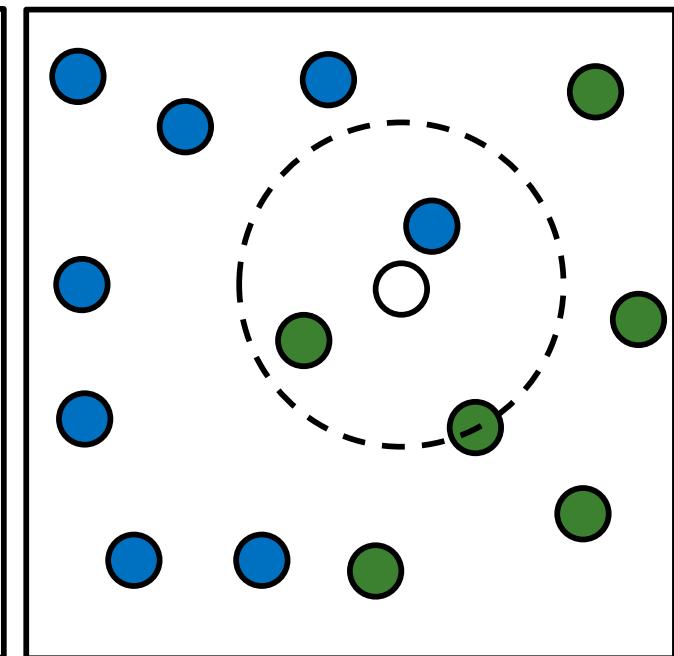
# Παράδειγμα k-NN



1-NN



2-NN

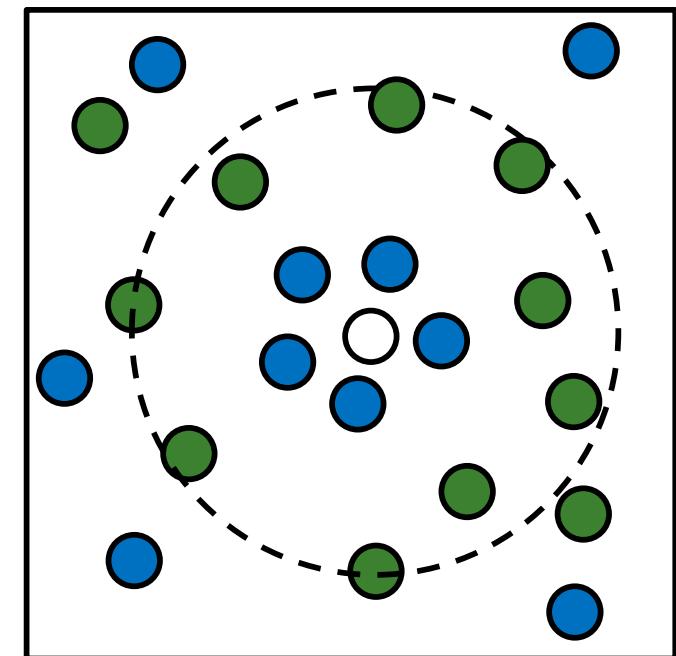


3-NN

- Το δείγμα κατηγοριοποιείται με βάση την πλειοψηφία των γειτόνων του
- Αν υπάρχει ισοψηφία, μπορεί να επιλεχθεί τυχαία ένας από τους γείτονες για την κατηγοριοποίηση (υπάρχουν και άλλες παραλλαγές)

# Επίδραση της Τιμής: k

- Από το παράδειγμα φαίνεται η σημασία της σωστής επιλογής της τιμής **k**
  - **Αν το k είναι μικρό**, ο κατηγοριοποιητής k-NN πιθανόν να γίνει επιρρεπής σε υπερπροσαρμογή, λόγω **Θορύβου** στα δεδομένα
  - **Αν το k είναι πολύ μεγάλο**, μπορεί να κατηγοριοποιήσει **λάθος** ένα δείγμα, διότι ίσως η λίστα κοντινότερων γειτόνων περιέχει δείγματα που βρίσκονται πολύ μακριά από τη γειτονιά του



# Αλγόριθμος

---

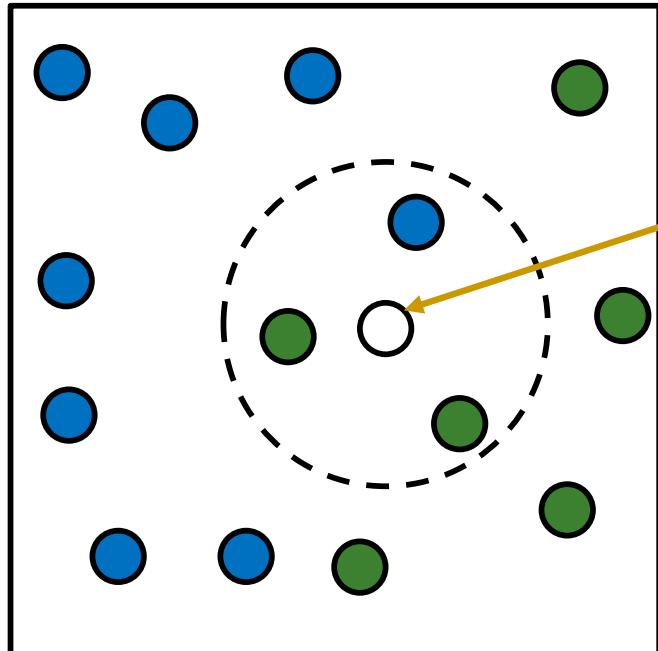
**Algorithm 5.2** The  $k$ -nearest neighbor classification algorithm.

---

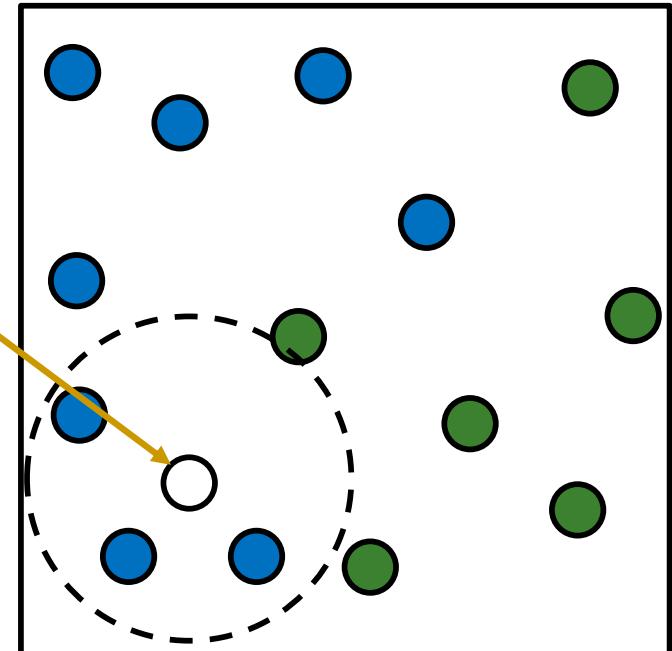
- 1: Let  $k$  be the number of nearest neighbors and  $D$  be the set of training examples.
  - 2: **for** each test example  $z = (\mathbf{x}', y')$  **do**
  - 3:   Compute  $d(\mathbf{x}', \mathbf{x})$ , the distance between  $z$  and every example,  $(\mathbf{x}, y) \in D$ .
  - 4:   Select  $D_z \subseteq D$ , the set of  $k$  closest training examples to  $z$ .
  - 5:    $y' = \underset{v}{\operatorname{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$
  - 6: **end for**
- 

- Υπολογίζεται η απόσταση του δείγματος ελέγχου με **κάθε δείγμα** του συνόλου εκπαίδευσης
  - Ακριβή υπολογιστικά διαδικασία
- **Πλειοψηφική ψηφοφορία**
  - $I()$ : συνάρτηση που επιστρέφει **1** αν το όρισμά της είναι αληθές, αλλιώς **0**

# Παράδειγμα (για $k=3$ )



νέο δείγμα

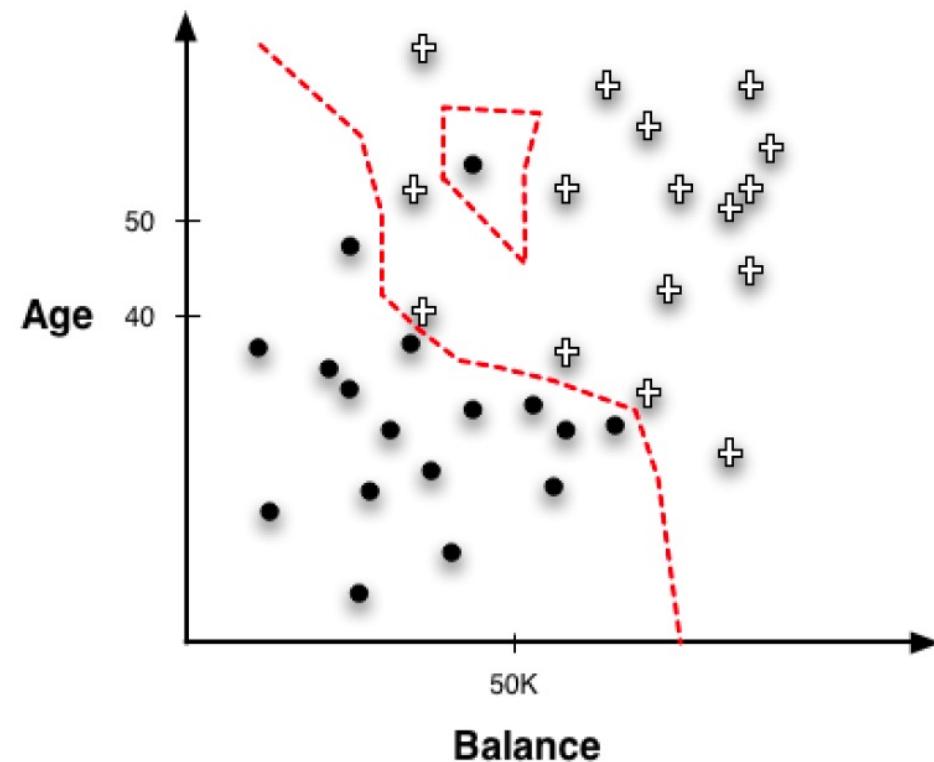
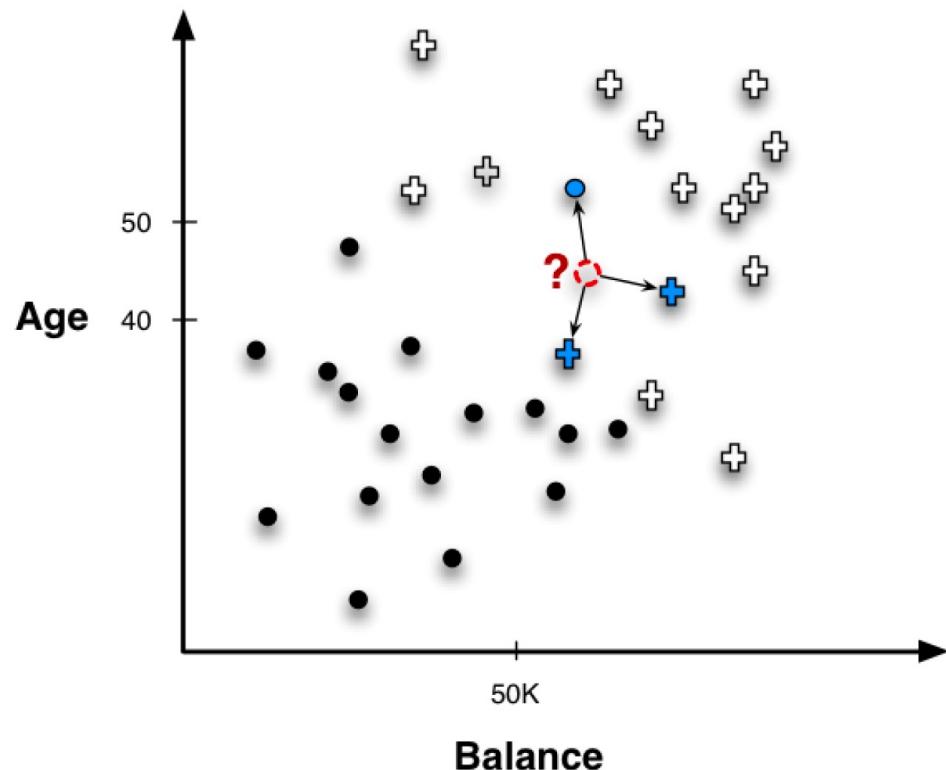


- Υπολογισμός 14 αποστάσεων
- Εύρεση των  $k=3$  κοντινότερων γειτόνων
- Κατηγοριοποίηση ως: πράσινο
- Υπολογισμός 14 αποστάσεων (**εκ νέου**)
- Εύρεση των  $k=3$  κοντινότερων γειτόνων
- Κατηγοριοποίηση ως: μπλε

# Στάθμιση Κοντινότερων Γειτόνων

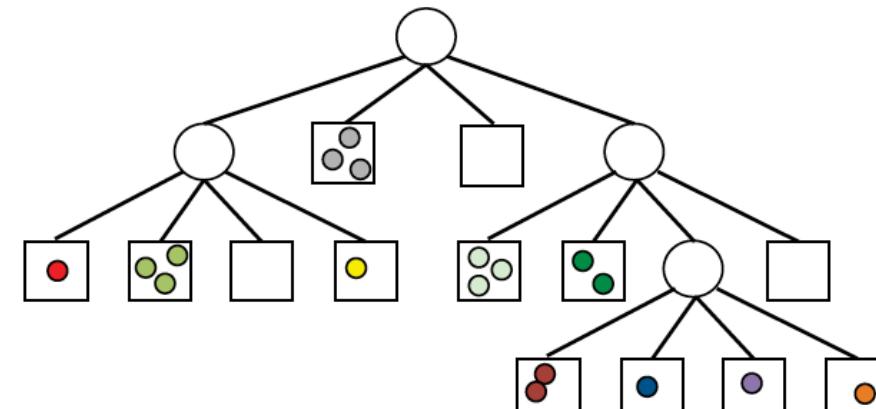
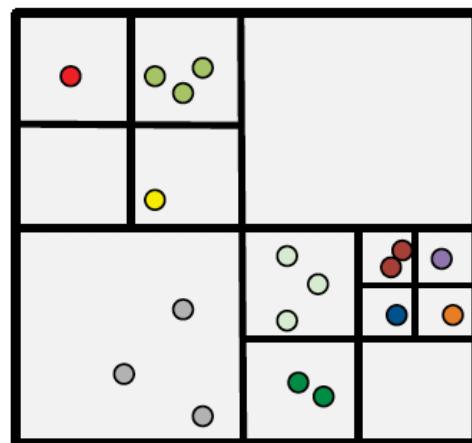
- Στην προσέγγιση πλειοψηφικής ψηφοφορίας κάθε γείτονας  $x_i$  έχει την **ίδια** επίδραση στην κατηγοριοποίηση
- Αυτό καθιστά τον αλγόριθμο **ευαίσθητο** στην επιλογή του  $k$
- **Τρόπος μείωσης της επίδρασης του  $k$** 
  - Σταθμίζοντας τη συνεισφορά κάθε κοντινότερου γείτονα  $x_i$  με βάση την απόστασή του  $d(x', x_i)$  από το δείγμα  $x'$
  - Στάθμιση (βάρος):  $w_i = 1 / d^2(x', x_i)$
  - Τα μακρινά δείγματα εκπαίδευσης έχουν μικρότερη επίδραση στην κατηγοριοποίηση

# 'Ορια Απόφασης για k-NN Κατηγοριοποιητές



# Θέματα Απόδοσης

- Για να μειωθεί το κόστος υπολογισμού χρησιμοποιείται ένα κατάλληλο ευρετήριο, όπως Quadtree (βλ.εικόνα), R-tree, ..., και ένας αποδοτικός αλγόριθμος υπολογισμού k-NN
- Επιτρέπει τον υπολογισμό των k κοντινότερων γειτόνων, χωρίς να υπολογιστούν οι αποστάσεις από όλα τα δείγματα



# Χαρακτηριστικά Κατηγοριοποιητή k-NN (1/2)

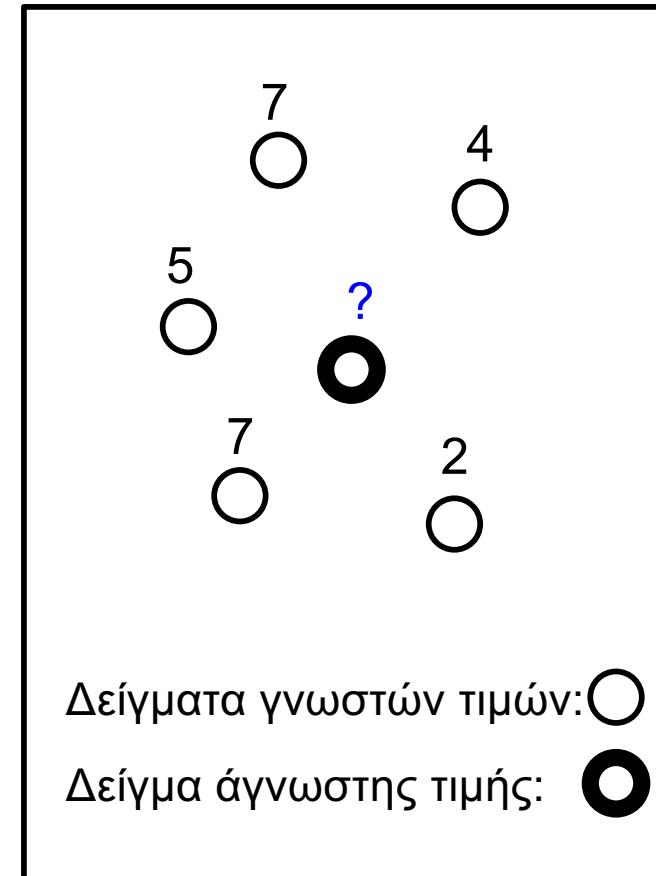
- Αποτελεί παράδειγμα εκπαίδευσης βάσει στιγμιοτύπων (**instance-based classifier**), ára **δεν κατασκευάζει κάποιο γενικό μοντέλο**
  - Απαιτείται όμως η χρήση κάποιου μέτρου απόστασης
  - Αποκαλείται και **lazy learner** (**μέθοδος οκνηρής μάθησης**)
- Η κατηγοριοποίηση νέου αντικειμένου μπορεί να είναι αρκετά **ακριβή**
  - Απαιτείται ο υπολογισμός απόστασης από τα δείγματα εκπαίδευσης
- Κάνει τις προβλέψεις με βάση **τοπικές πληροφορίες**
  - Σε αντίθεση με τα δέντρα απόφασης που κατασκευάζουν ένα γενικό μοντέλο που προσαρμόζει ολόκληρο τον χώρο εισόδου
- Μπορεί να παράγει όρια απόφασης αυθαίρετου σχήματος

## Χαρακτηριστικά Κατηγοριοποιητή k-NN (2/2)

- Δυσκολεύεται να διαχειριστεί **ελλιπείς** τιμές
- Η παρουσία **άσχετων χαρακτηριστικών** μπορεί να αλλοιώσει τα συνηθισμένα μέτρα απόστασης, ειδικά όταν το πλήθος τους είναι μεγάλο
- Απαιτείται χρήση **κατάλληλου μέτρου απόστασης** και βημάτων **προεπεξεργασίας**
  - **Κανονικοποίηση** χαρακτηριστικών
  - Παράδειγμα: ύψος (1.5-1.85), βάρος (60κ-120κ), οπότε οι διαφορές βάρους κυριαρχούν στον υπολογισμό απόστασης

# Παλινδρόμηση με χρήση Μεθόδου k-Κοντινότερων Γειτόνων

- Η βασική ιδέα παραμένει ίδια
- Εύρεση των k-κοντινότερων γειτόνων (όπως και πριν), βάσει
  - κάποιας τιμής για το k
  - κάποιας συνάρτησης απόστασης
- Πρόβλεψη τιμής βάσει των τιμών των k-κοντινότερων γειτόνων
- Παράδειγμα:
  - Απλός μέσος όρος
  - $(7 + 4 + 5 + 7 + 2) / 5 = 5$



# Κατηγοριοποιητής Bayes (Bayes classifier)

# Bayesian Classifiers

- Ακολουθούν μια **πιθανοτική προσέγγιση** στο θέμα της κατηγοριοποίησης με χρήση θεωρίας πιθανοτήτων
- Τα μοντέλα κατηγοριοποίησης που χρησιμοποιούν τη θεωρία πιθανοτήτων για να αναπαριστούν τη σχέση ανάμεσα στα χαρακτηριστικά και τις ετικέτες κατηγοριών είναι γνωστά ως **πιθανοφανή μοντέλα κατηγοριοποίησης**
- Δοθέντος ενός συνόλου γνωρισμάτων, υπολογίζουν την **πιθανότητα** ένα στιγμιότυπο να ανήκει στη μία ή στην άλλη κλάση
- Το στιγμιότυπο ανατίθεται στην κλάση με τη μέγιστη πιθανότητα
- Ουσιαστικά, πρέπει να υπολογιστεί η υπό συνθήκη πιθανότητα:

$$P(\text{class } C | \{x_1, x_2, x_3, \dots, x_n\})$$

# Θεώρημα Bayes (Υπενθύμιση)

X, Y: τυχαίες μεταβλητές

Από κοινού πιθανότητα:  $P(X = x, Y = y)$

Υπό συνθήκη πιθανότητα:  $P(Y = y|X = x)$

Οι πιθανότητες αυτές σχετίζονται ως εξής:

$$P(X, Y) = P(Y|X) \cdot P(X) = P(X|Y) \cdot P(Y)$$

ή ισοδύναμα (*Θεώρημα Bayes*):

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

# Bayesian Classifiers (συνέχ.)

## ■ Από Θεώρημα Bayes:

$$P(\text{class } C | \{x_i\}) = \frac{P(\{x_i\} | \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

Συχνότητα εμφάνισης

Εκ των προτέρων  
πιθανότητα

Μπορεί (θεωρητικά) να προσεγγιστεί με χρήση ιστογραμμάτων, όμως είναι εξαιρετικά πολλοί οι συνδυασμοί των διαφορετικών τιμών για τα γνωρίσματα – στην πράξη μπορεί να είναι αδύνατο

*Χρησιμοποιούνται ορισμένες απλοποιήσεις*

# Bayesian Classifiers – Απλοποίηση

## ■ Απλοϊκός Bayes κατηγοριοποιητής (Naïve Bayes Classifier)

- Υποθέτει ότι όλα τα γνωρίσματα είναι **υπό συνθήκη ανεξάρτητα μεταξύ τους**

$$P(\{x_1, x_2, x_3, \dots, x_n\} | C) = P(x_1|C)P(x_2|C)P(x_3|C) \cdots P(x_n|C)$$

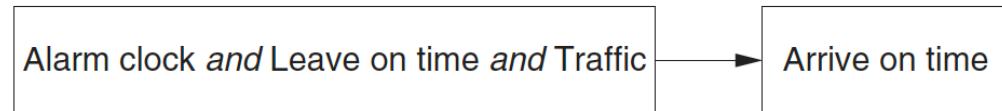
- Απλοποιεί πολύ το πρόβλημα: η κάθε πιθανότητα μπορεί να υπολογιστεί με ένα ιστόγραμμα ενός γνωρίσματος

## ■ Δίκτυα πεποίθησης Bayes (Bayesian Networks)

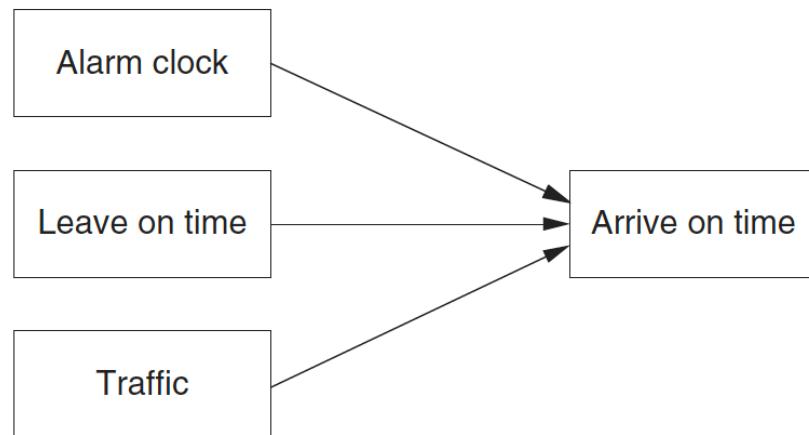
- Επιτρέπει να καθοριστεί ποιο ζεύγος γνωρισμάτων είναι υπό συνθήκη ανεξάρτητο (αντί να απαιτείται από όλα τα γνωρίσματα να είναι υπό συνθήκη ανεξάρτητα)

# Bayesian Classifiers – Απλοποίηση

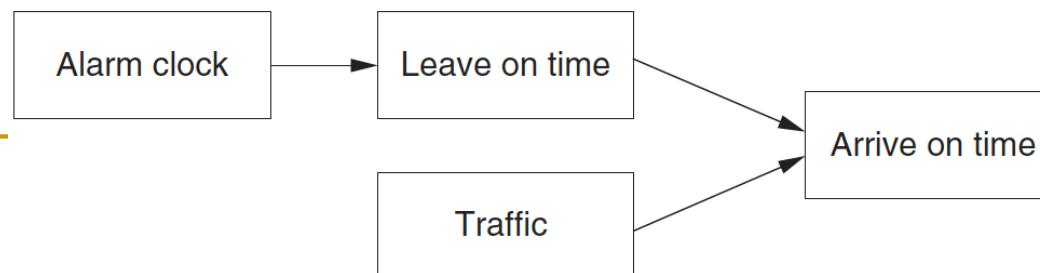
## All Combinations



## Naive Bayesian



## Bayesian Network



# Παράδειγμα Απλού Bayes

Δίνεται δείγμα ελέγχου:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Μπορούμε να εκτιμήσουμε τις πιθανότητες:  $P(\text{Evade} = \text{Yes} | X)$  και  $P(\text{Evade} = \text{No} | X)$ ;

Using Bayes Theorem:

$$\square P(\text{Yes} | X) = \frac{P(X | \text{Yes})P(\text{Yes})}{P(X)}$$

$$\square P(\text{No} | X) = \frac{P(X | \text{No})P(\text{No})}{P(X)}$$

□ How to estimate  $P(X | \text{Yes})$  and  $P(X | \text{No})$ ?

# Παράδειγμα Απλού Bayes

Δίνεται δείγμα ελέγχου:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- $P(X | \text{Yes}) =$   
 $P(\text{Refund} = \text{No} | \text{Yes}) \times$   
 $P(\text{Divorced} | \text{Yes}) \times$   
 $P(\text{Income} = 120\text{K} | \text{Yes})$
- $P(X | \text{No}) =$   
 $P(\text{Refund} = \text{No} | \text{No}) \times$   
 $P(\text{Divorced} | \text{No}) \times$   
 $P(\text{Income} = 120\text{K} | \text{No})$

# Παράδειγμα Απλού Bayes

Δίνεται δείγμα ελέγχου:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable Income	Evide
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Εκτίμηση πιθανοτήτων από δεδομένα
- Κλάση:  $P(Y) = N_c/N$ 
  - $P(\text{No}) = 7/10, P(\text{Yes}) = 3/10$
- Για **κατηγορικά γνωρίσματα**:
- $P(X_i | Y_k) = |X_{ik}| / N_c$ 
  - όπου  $|X_{ik}|$  είναι ο αριθμός δειγμάτων που έχουν τιμή γνωρίσματος  $X_i$  και ανήκουν στην κλάση  $Y_k$
  - Παραδείγματα:  
 $P(\text{Status}=\text{Married} | \text{No}) = 4/7$   
 $P(\text{Refund}=\text{Yes} | \text{Yes})=0$

# Παράδειγμα Απλού Bayes

Δίνεται δείγμα ελέγχου:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Εκτίμηση πιθανοτήτων από δεδομένα
- Κλάση:  $P(Y) = N_c/N$ 
  - $P(\text{No}) = 7/10, P(\text{Yes}) = 3/10$
- Για **συνεχή γνωρίσματα (2 εναλλακτικές)**:
  - Διακριτοποίηση (άρα όπως πριν)
  - Εκτίμηση συνάρτησης πυκνότητας πιθανότητας
    - ◆ Υποθέτουμε κανονική κατανομή για το γνώρισμα
    - ◆ Εκτιμούμε τις παραμέτρους της κατανομής από τα δεδομένα (το μέσο και την τυπική απόκλιση)
    - ◆ Αφού μάθουμε την κατανομή, τη χρησιμοποιούμε για την εκτίμηση της υπό συνθήκη πιθανότητας  $P(X_i | Y)$

# Παράδειγμα Απλού Bayes

Δίνεται δείγμα ελέγχου:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Εκτίμηση πιθανοτήτων από δεδομένα
- Κλάση:  $P(Y) = N_c/N$ 
  - $P(\text{No}) = 7/10, P(\text{Yes}) = 3/10$
- Για **συνεχή γνωρίσματα** (2 εναλλακτικές):
  - Εκτίμηση συνάρτησης πυκνότητας πιθανότητας

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

$$\bar{x} = \frac{125 + 100 + 70 + \dots + 75}{7} = 110$$

$$s^2 = \frac{(125 - 110)^2 + (100 - 110)^2 + \dots + (75 - 110)^2}{7(6)} = 2975$$

$$s = \sqrt{2975} = 54.54.$$

# Παράδειγμα Απλού Bayes

Δίνεται δείγμα ελέγχου:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$P(\text{Refund} = \text{Yes} | \text{No}) = 3/7$   
 $P(\text{Refund} = \text{No} | \text{No}) = 4/7$   
 $P(\text{Refund} = \text{Yes} | \text{Yes}) = 0$   
 $P(\text{Refund} = \text{No} | \text{Yes}) = 1$   
 $P(\text{Marital Status} = \text{Single} | \text{No}) = 2/7$   
 $P(\text{Marital Status} = \text{Divorced} | \text{No}) = 1/7$   
 $P(\text{Marital Status} = \text{Married} | \text{No}) = 4/7$   
 $P(\text{Marital Status} = \text{Single} | \text{Yes}) = 2/3$   
 $P(\text{Marital Status} = \text{Divorced} | \text{Yes}) = 1/3$   
 $P(\text{Marital Status} = \text{Married} | \text{Yes}) = 0$

For Taxable Income:  
If class = No: sample mean = 110  
sample variance = 2975  
If class = Yes: sample mean = 90  
sample variance = 25

# Παράδειγμα Απλού Bayes

Δίνεται δείγμα ελέγχου:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- $P(X | \text{No}) = P(\text{Refund}=\text{No} | \text{No}) \times P(\text{Divorced} | \text{No}) \times P(\text{Income}=120\text{K} | \text{No}) = 4/7 \times 1/7 \times 0.0072 = 0.0006$
- $P(X | \text{Yes}) = P(\text{Refund}=\text{No} | \text{Yes}) \times P(\text{Divorced} | \text{Yes}) \times P(\text{Income}=120\text{K} | \text{Yes}) = 1 \times 1/3 \times 1.2 \times 10^{-9} = 4 \times 10^{-10}$

Αφού:  $P(X | \text{No}) P(\text{No}) > P(X | \text{Yes}) P(\text{Yes})$   
Άρα:  $P(\text{No} | X) > P(\text{Yes} | X)$   
 $\Rightarrow \text{Class} = \text{No}$

# Χαρακτηριστικά Απλοϊκών Bayes Κατηγοριοποιητών

- Είναι μοντέλα κατηγοριοποίησης τα οποία μπορούν να ποσοτικοποιούν την αβεβαιότητα των προβλέψεων
- Λειτουργούν ακόμη και σε **πολυδιάστατα** προβλήματα, υπό την προϋπόθεση ότι τα χαρακτηριστικά είναι υπό συνθήκη ανεξάρτητα μεταξύ τους
- Είναι **εύρωστοι** σε μεμονωμένα **σημεία θορύβου**, επειδή αυτά δεν μπορούν να επηρεάσουν σημαντικά τις εκτιμήσεις των υπό συνθήκη πιθανοτήτων

# Χαρακτηριστικά Απλοϊκών Bayes Κατηγοριοποιητών

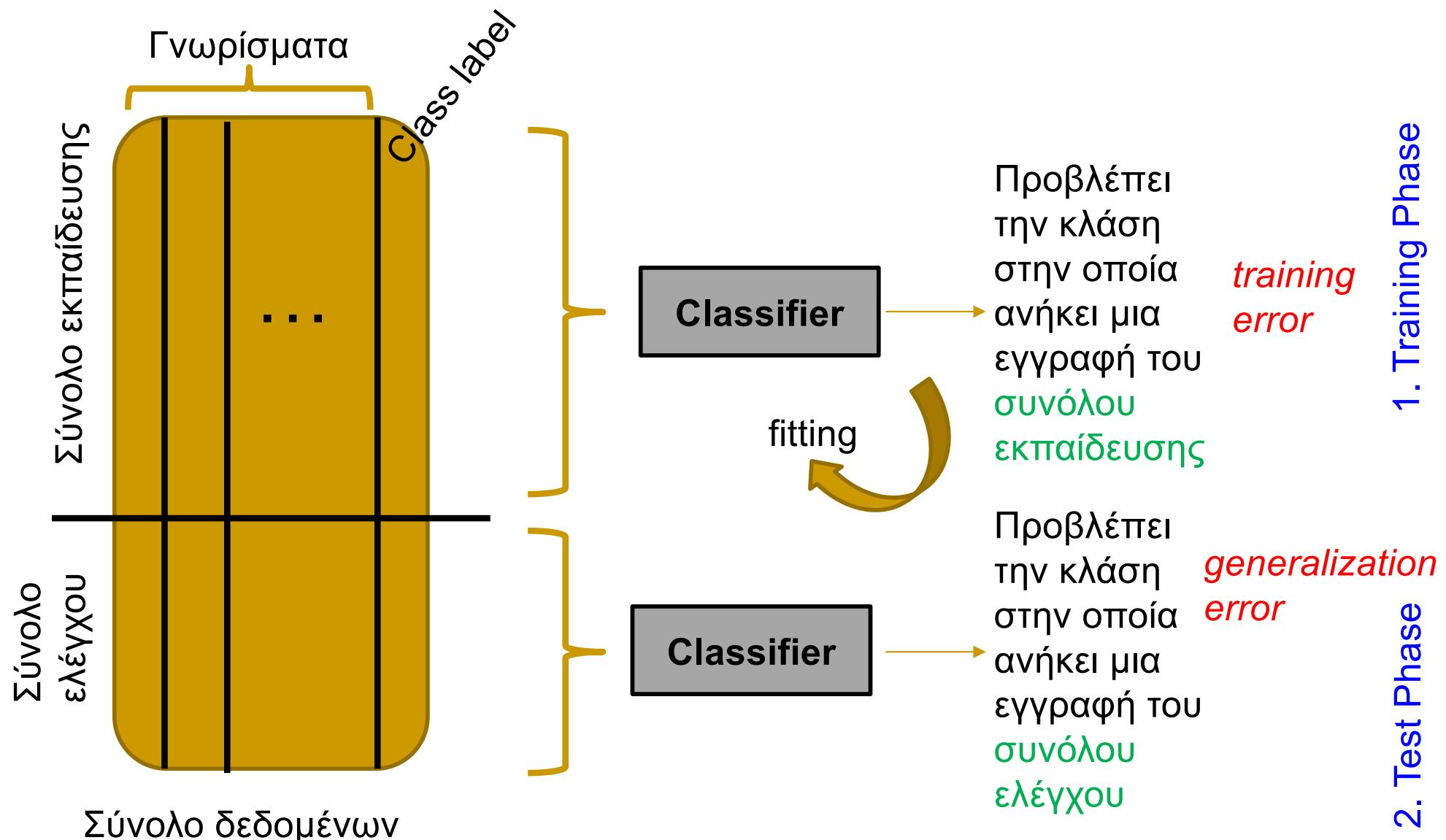
- Μπορούν να διαχειρίζονται **ελλιπείς τιμές**, αγνοώντας τες κατά τον υπολογισμό των πιθανοτήτων
- Δεν είναι κατάλληλη μέθοδος για τη διαχείριση **περιττών** ή **συσχετιζόμενων** χαρακτηριστικών

# Αξιολόγηση Μοντέλων

# Κατηγοριοποίηση – Τρόπος Λειτουργίας

- Ένας κατηγοριοποιητής ή ταξινομητής (classifier)
  - παίρνει ως είσοδο μια εγγραφή
  - και παράγει μια ετικέτα κλάσης για αυτή
  - ουσιαστικά αναθέτει την εγγραφή σε μια κλάση
- Κατασκευή και χρήση ενός κατηγοριοποιητή (3 βήματα):
  - Εκπαίδευση (Training)
  - Δοκιμή – έλεγχος (Testing)
  - Εφαρμογή
- Αρχικά χωρίζουμε τα δεδομένα σε
  - σύνολο εκπαίδευσης (training set) και
  - σύνολο ελέγχου (test set)

# Η Διαδικασία Κατηγοριοποίησης Οπτικά



# Αξιολόγηση Κατηγοριοποιητών

- Θα υποθέσουμε δυαδική κατηγοριοποίηση καταρχήν για **απλότητα**
  - Δύο κατηγορίες: «θετική» και «αρνητική»
- ***Η αξιολόγηση θα γίνει στα δεδομένα που παρακρατήθηκαν***
- ***Ωστε να αποτιμήσουμε την απόδοση του μοντέλου στη γενίκευση***
- Πώς υπολογίζεται η απόδοση του μοντέλου στη γενίκευση;

# Ακρίβεια (Accuracy)

- Ακρίβεια (accuracy) είναι το ποσοστό σωστών αποφάσεων ενός classifier:

$$\text{accuracy} = \frac{\text{Number of correct decisions made}}{\text{Total number of decisions made}}$$

- Είναι ένα απλοϊκό μέτρο αξιολόγησης που παρουσιάζει ορισμένα προβλήματα (βλ. παρακάτω)

# Μήτρα Σύγχυσης (Confusion Matrix)

- Πολλά προβλήματα κατηγοριοποίησης είναι δυαδικά (binary):  
έγκυρη/απάτη ή spam/όχι spam
  - Όμως (γενικά) μπορεί να έχουμε περισσότερες κλάσεις

**Πραγματικές κατηγορίες: **p** και **n****

<b>p</b>	<b>n</b>
Y	True positives
N	False positives
N	False negatives
	True negatives

**Προβλέψεις (από το μοντέλο): **Y** και **N****

Προσοχή: συχνά αναπαρίσταται με τις προβλέψεις ως στήλες

*Μήτρα σύγχυσης (Confusion matrix) για δυαδική κατηγοριοποίηση*

- Η διαγώνιος περιλαμβάνει τις σωστές αποφάσεις
- Τα σφάλματα του classifier είναι τα **ψευδώς θετικά (false positives)** και τα **ψευδώς αρνητικά (false negatives)**

# Παράδειγμα

- Πρόβλημα κατηγοριοποίησης:
    - *αν ένας φοιτητής θα περάσει την εξέταση*
  - Ένας κατηγοριοποιητής εφαρμόζεται σε 100 εγγραφές (φοιτητές)
    - 50 φορές προβλέπει ΝΑΙ και συμβαίνει ΝΑΙ (*true positives*)
    - 30 φορές προβλέπει ΟΧΙ και συμβαίνει ΟΧΙ (*true negatives*)
    - 15 φορές προβλέπει ΝΑΙ και συμβαίνει ΟΧΙ (*false positives*)
    - 5 φορές προβλέπει ΟΧΙ και συμβαίνει ΝΑΙ (*false negatives*)
- (1) Να υπολογιστεί η **ακρίβεια (accuracy)** του κατηγοριοποιητή
- (2) Να φτιαχτεί η **μήτρα σύγχυσης**

# Παράδειγμα – Επίλυση

- Πρόβλημα κατηγοριοποίησης:
  - *αν ένας φοιτητής θα περάσει την εξέταση*
- Ένας κατηγοριοποιητής εφαρμόζεται σε 100 εγγραφές (φοιτητές)
  - 50 φορές προβλέπει ΝΑΙ και συμβαίνει ΝΑΙ (*true positives*)
  - 30 φορές προβλέπει ΟΧΙ και συμβαίνει ΟΧΙ (*true negatives*)
  - 15 φορές προβλέπει ΝΑΙ και συμβαίνει ΟΧΙ (*false positives*)
  - 5 φορές προβλέπει ΟΧΙ και συμβαίνει ΝΑΙ (*false negatives*)
- (1) Να υπολογιστεί η **ακρίβεια (accuracy)** του κατηγοριοποιητή

$$\begin{aligned} \text{accuracy} &= \text{σωστές αποφάσεις} / \text{συνολικές αποφάσεις} \\ &= (50 + 30) / (50 + 30 + 15 + 5) \\ &= 80\% \end{aligned}$$

# Παράδειγμα - Επίλυση

- Πρόβλημα κατηγοριοποίησης:
  - *αν ένας φοιτητής θα περάσει την εξέταση*
- Ένας κατηγοριοποιητής εφαρμόζεται σε 100 εγγραφές (φοιτητές)
  - 50 φορές προβλέπει ΝΑΙ και συμβαίνει ΝΑΙ (*true positives*)
  - 30 φορές προβλέπει ΟΧΙ και συμβαίνει ΟΧΙ (*true negatives*)
  - 15 φορές προβλέπει ΝΑΙ και συμβαίνει ΟΧΙ (*false positives*)
  - 5 φορές προβλέπει ΟΧΙ και συμβαίνει ΝΑΙ (*false negatives*)

## (2) Να φτιαχτεί η μήτρα σύγχυσης

Προβλέψεις

Πραγματικές κατηγορίες

	ΝΑΙ	ΟΧΙ
ΝΑΙ	50	15
ΟΧΙ	5	30

# 'Άλλα Μέτρα Αξιολόγησης

	p	n
Y	True positives	False positives
N	False negatives	True negatives

- **Precision = TP / (TP + FP)**
- Μετράει την **ακρίβεια των θετικών προβλέψεων**
- Μετράει το ποσοστό των αληθινά θετικών σε σχέση με όλα όσα προβλέφθηκαν ως θετικά
  - Υψηλό precision σημαίνει υψηλή ακρίβεια στις θετικές προβλέψεις
  - Χαμηλό precision σημαίνει πολλές εσφαλμένες θετικές προβλέψεις

# 'Άλλα Μέτρα Αξιολόγησης

	p	n
Y	True positives	False positives
N	False negatives	True negatives

- **Recall = TP / (TP + FN)**
- Μετράει το ποσοστό των αληθινά θετικών προβλέψεων σε σχέση με όλα όσα ήταν πραγματικά θετικά
  - Υψηλό recall σημαίνει υψηλή ακρίβεια στη θετική κατηγορία
  - Χαμηλό recall σημαίνει χαμηλή ακρίβεια στη θετική κατηγορία
- Γνωστό και ως: **True Positive Rate (TPR)** και **Sensitivity**

# 'Άλλα Μέτρα Αξιολόγησης

p	n
Y True positives	False positives
N False negatives	True negatives

- **F1-score = 2 x Precision x Recall / (Precision + Recall)**
- Συνδυαστικό μέτρο, συνδυάζει precision και recall
  - Υψηλό F1-score σημαίνει ότι ο κατηγοριοποιητής τα πάει καλά και ως προς το precision και ως προς το recall
  - Χαμηλό F1-score σημαίνει σημαίνει ότι ο κατηγοριοποιητής δεν τα πάει καλά ούτε ως προς το precision ούτε ως προς το recall
- Είναι ο **αρμονικός μέσος** των precision και recall

# Άνισα Κατανεμημένες Κατηγορίες

- Πρόβλημα απώλειας πελατών
  - *Churn*: πελάτες που φεύγουν
  - *Not churn*: πελάτες που μένουν
- Και τα δύο μοντέλα κατηγοριοποιούν **σωστά το 80%** του πληθυσμού σε ένα σύνολο δεδομένων εκπαίδευσης **1.000 στιγμιότυπων**, όμως
  - Ο classifier A προβλέπει λανθασμένα ότι κάποιος θα αποχωρήσει
  - Ο classifier B κάνει σφάλματα προβλέποντας ότι κάποιος θα παραμείνει στην εταιρεία, ενώ δε συμβαίνει αυτό
- Όμως όταν εφαρμόζονται σε **άνισα κατανεμημένο πληθυσμό**, η ακρίβεια του A είναι 64% και του B 96%

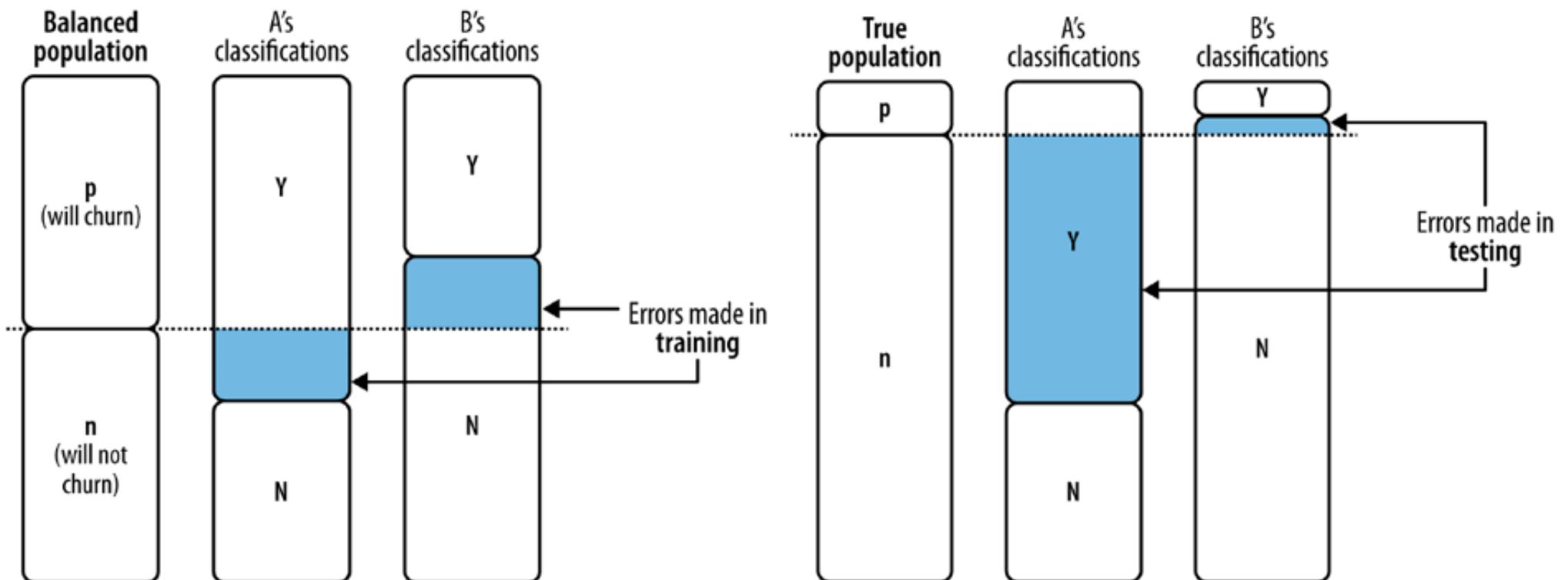
Μοντέλο A

	churn	not churn
Y	500	200
N	0	300

Μοντέλο B

	churn	not churn
Y	300	0
N	200	500

# Άνισα Κατανεμημένες Κατηγορίες

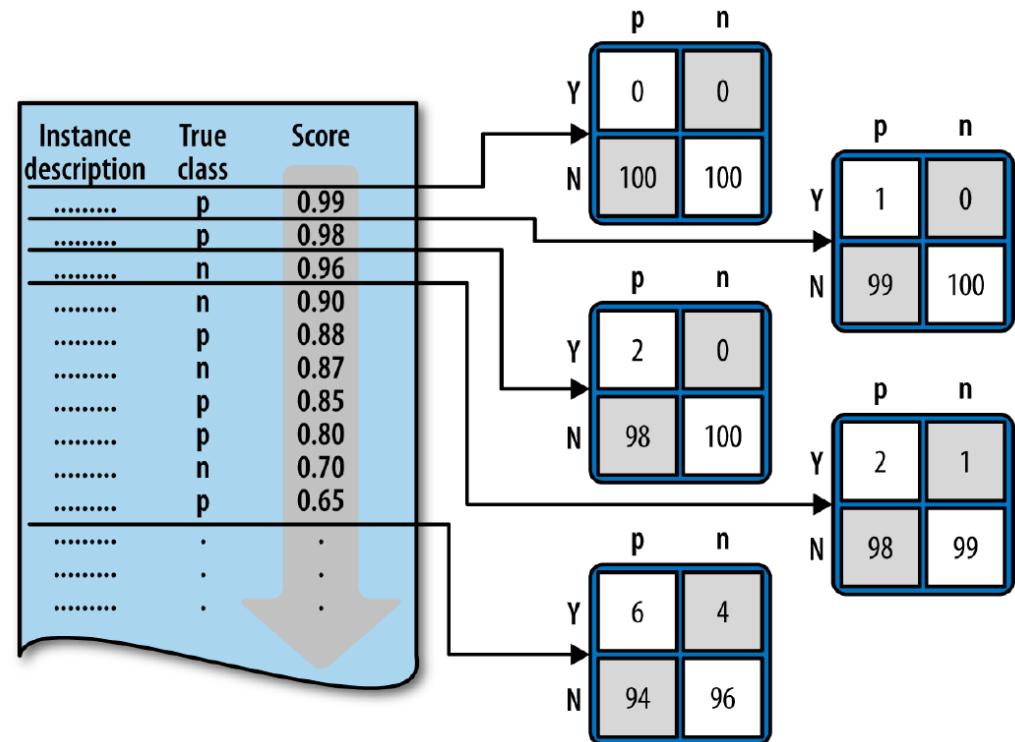


# Απόδοση Βάσης Αναφοράς

- Συχνά, είναι χρήσιμο να χρησιμοποιείται μια εύλογη **βάση αναφοράς** (**baseline**) με την οποία θα συγκρίνουμε την απόδοση των μοντέλων μας
  - Εξαρτάται από την εφαρμογή
  - Η εύρεση κατάλληλης βάσης αναφοράς σχετίζεται με την κατανόηση του προβλήματος
- **Γενικές αρχές** για **βάση αναφοράς**:
  - **Τυχαία** κατηγοριοποίηση (**random classification**)
  - **Πλειοψηφικός** κατηγοριοποιητής (**majority classifier**)
    - Επιλέγει πάντα την επικρατούσα κατηγορία του συνόλου δεδομένων εκπαίδευσης
  - Για παλινδρόμηση: πρόβλεψη της **μέσης τιμής** του πληθυσμού
  - Προγνωστικά μοντέλα που στηρίζονται σε **ένα μόνο χαρακτηριστικό**
    - Επαγωγή δέντρου για δημιουργία «άκρου απόφασης» (decision stump)

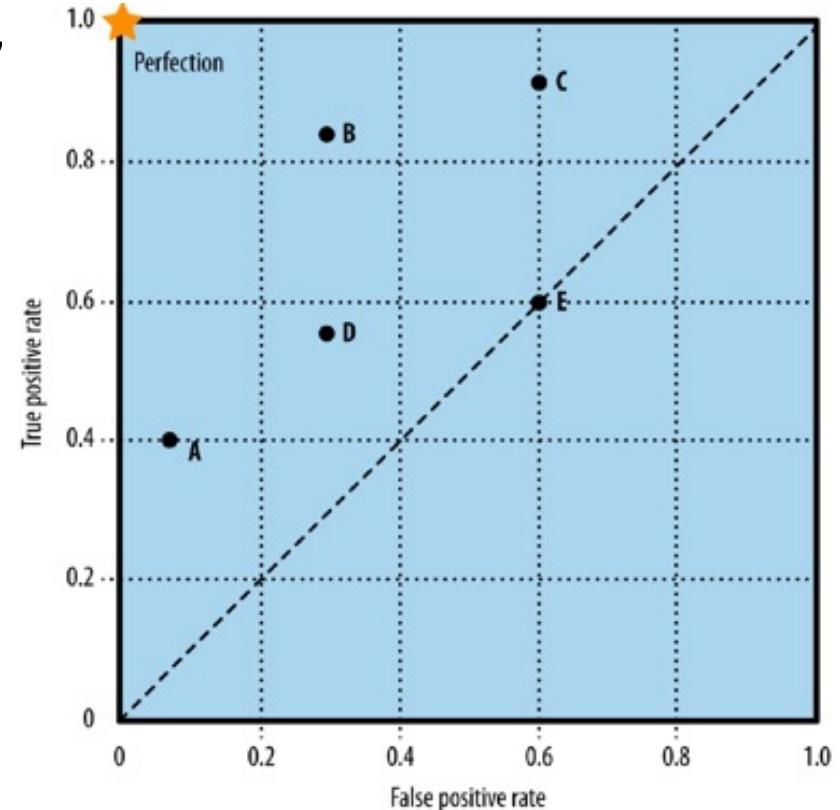
# Κατάταξη αντί Κατηγοριοποίησης

- Όταν το μοντέλο δίνει μια βαθμολογία που κατατάσσει τα στιγμιότυπα με βάση την πιθανοφάνειά τους
- Ορίζουμε μια **τιμή κατώφλι**
- Θετικά: όσα στιγμιότυπα είναι πάνω από αυτή την τιμή
- Αρνητικά: όσα στιγμιότυπα είναι κάτω από αυτή την τιμή
- Όμως, πώς να καθορίσουμε μια κατάλληλη τιμή κατωφλιού;



# Καμπύλη ROC

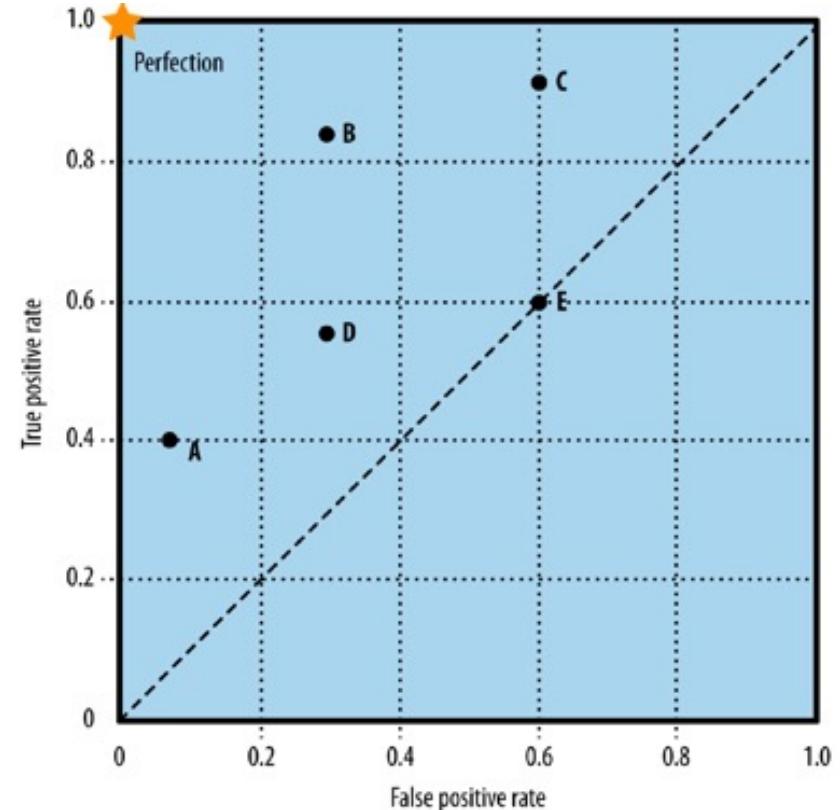
- Γράφημα λειτουργικών χαρακτηριστικών δέκτη ή γράφημα ROC (Receiver Operating Characteristics graph)
- Δισδιάστατη γραφική παράσταση ενός κατηγοριοποιητή στην οποία απεικονίζεται
  - το ποσοστό ψευδών θετικών (false positives) στον **άξονα X** και
  - το ποσοστό των αληθώς θετικών (true positives) στον **άξονα Y**



Χώρος ROC και η απόδοση πέντε κατηγοριοποιητών A-E

# Καμπύλη ROC

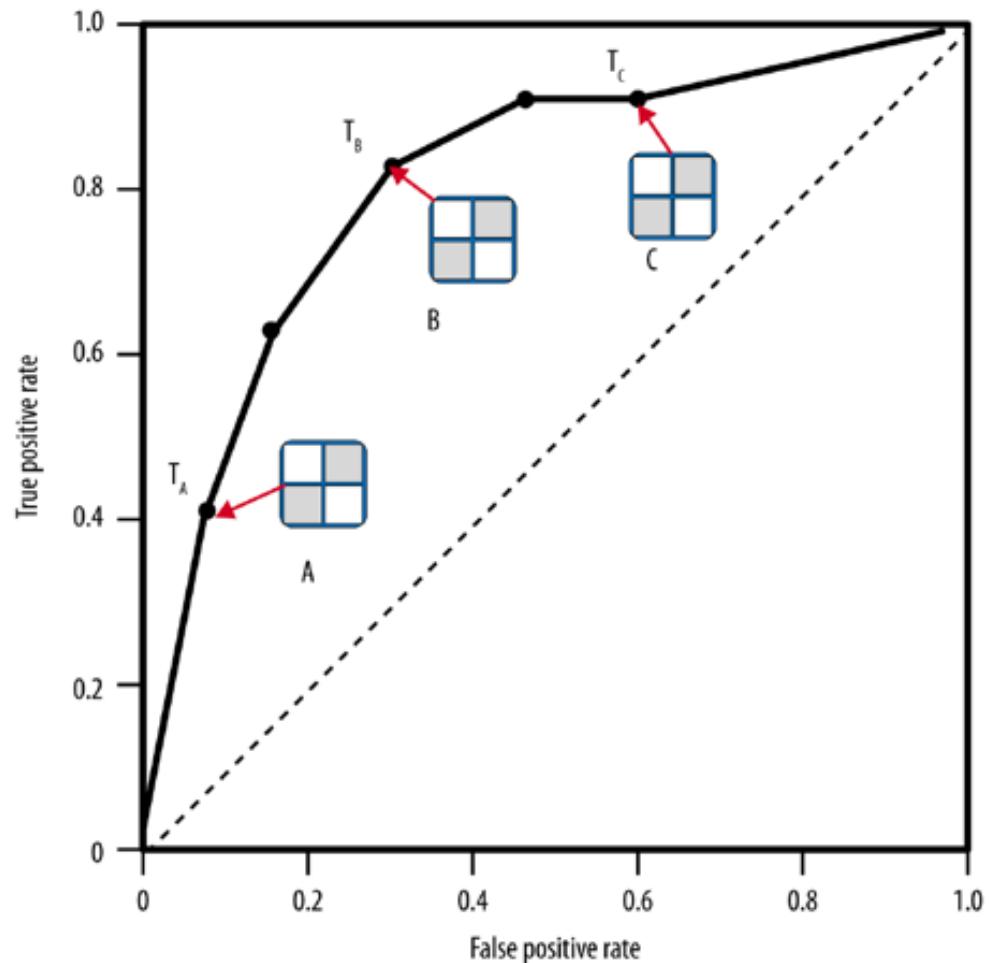
- Το γράφημα ROC απεικονίζει τις σχετικές αντισταθμίσεις (tradeoffs) που κάνει ένας κατηγοριοποιητής ανάμεσα στα οφέλη (αληθινά θετικά) και τα κόστη (ψευδώς θετικά)
- Ένας **τυχαίος κατηγοριοποιητής** παράγει ένα σημείο που κινείται στη διαγώνιο ανάλογα με τη συχνότητα με την οποία μαντεύει τη θετική κατηγορία



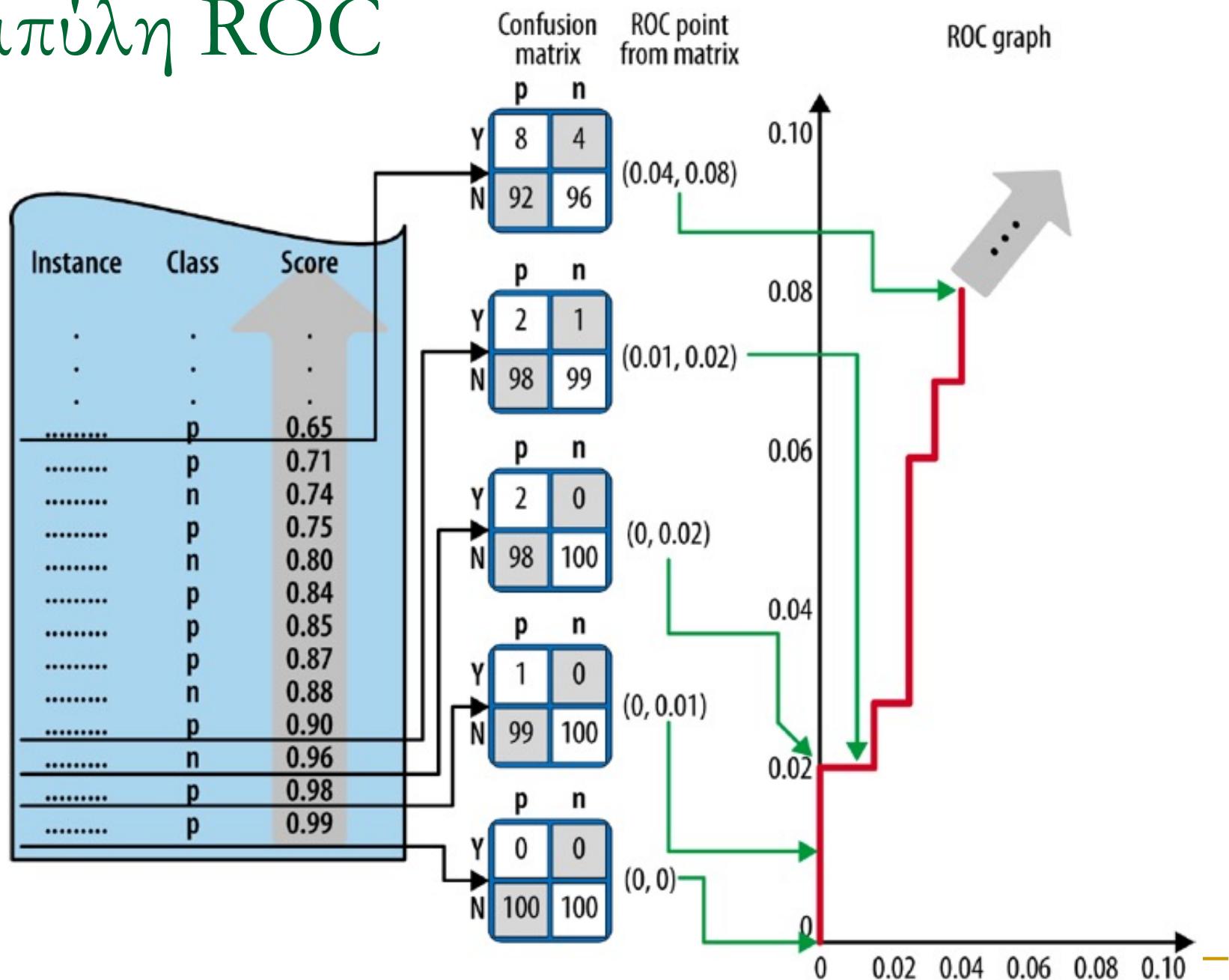
Χώρος ROC και η απόδοση πέντε κατηγοριοποιητών A-E

# Καμπύλη ROC

- Οι κατηγοριοποιητές που είναι στην αριστερή πλευρά (κοντά στον άξονα x) θεωρούνται «συντηρητικοί»
  - Προβλέπουν θετική κατηγορία, μόνο όταν υπάρχουν ισχυρά στοιχεία
  - Α πιο συντηρητικός από Β, και αυτός πιο συντηρητικός από Ζ
- Αντίθετα, στην πάνω δεξιά πλευρά μπορούν να θεωρηθούν «ανεκτικοί»
  - Κάνουν πολύ συχνά θετική κατηγοριοποίηση, με αποτέλεσμα να έχουν υψηλά ποσοστά false positives



# Καμπύλη ROC

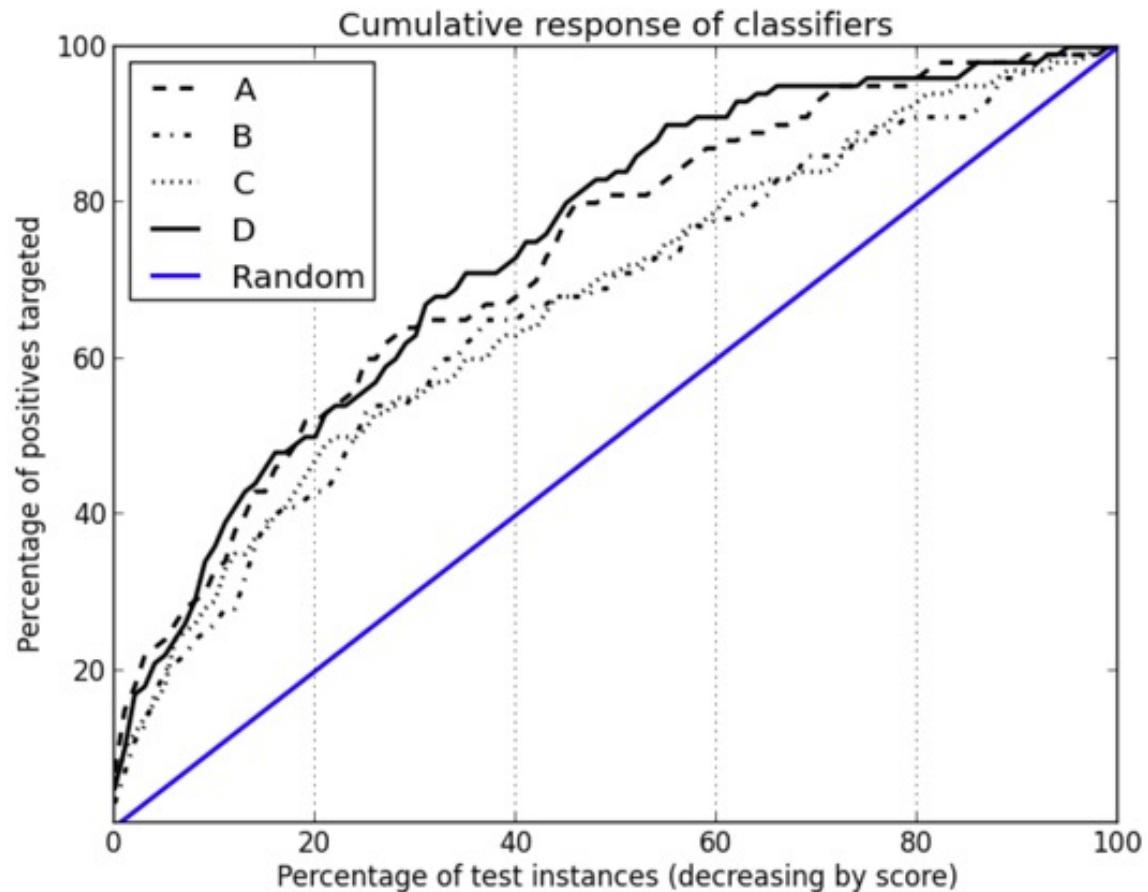


# Area Under Curve (AUC)

- *Η περιοχή κάτω από την καμπύλη ROC*
  - Είναι μια τιμή που κυμαίνεται από 0 ως 1
  - Εκφράζεται ως κλάσμα της τετραγωνικής μονάδας
- Πρόκειται για ένα στατιστικό στοιχείο σύνοψης της προγνωστικής ικανότητας ενός κατηγοριοποιητή
  - Παρόλο που η καμπύλη ROC παρέχει περισσότερες πληροφορίες
- Η περιοχή AUC είναι ισοδύναμη με
  - την πιθανότητα ότι ένα τυχαία επιλεγμένο θετικό στιγμιότυπο θα καταταγεί μπροστά από ένα τυχαία επιλεγμένο αρνητικό στιγμιότυπο

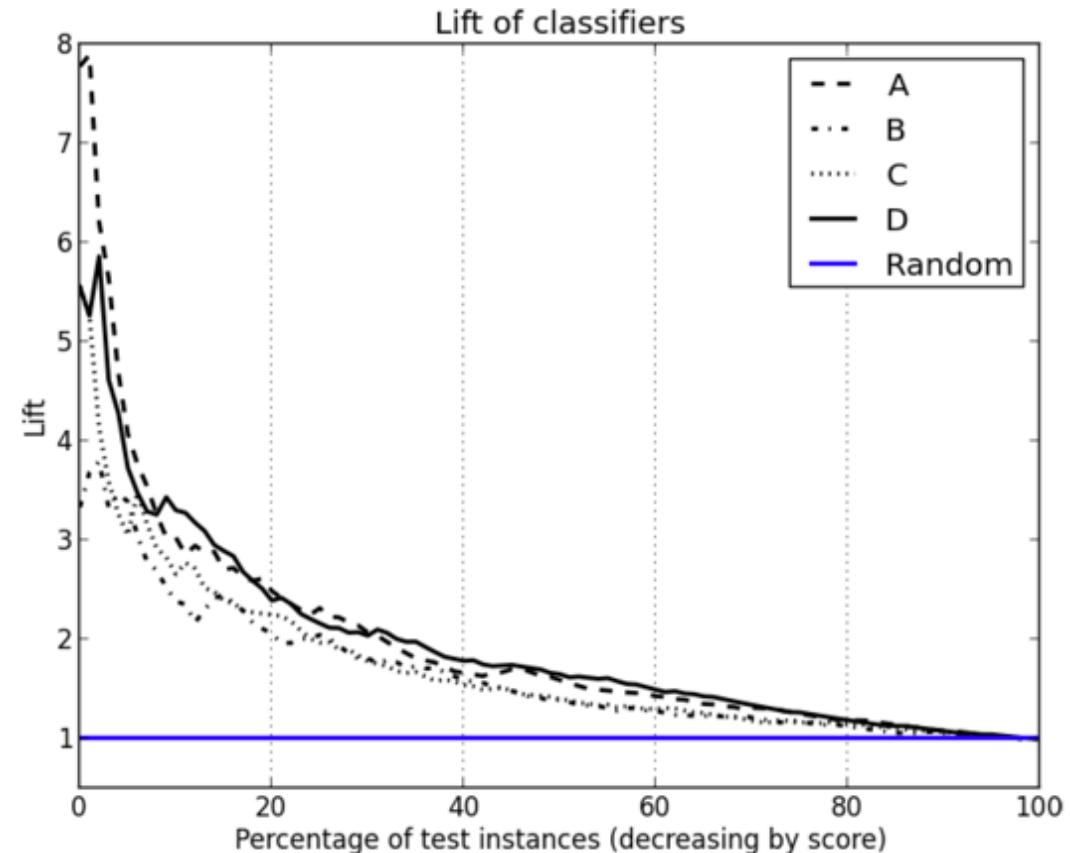
# Αθροιστική Καμπύλη Ανταπόκρισης

- Αθροιστική καμπύλη ανταπόκρισης (cumulative response curve)
  - Άξονας Y: true positives σε ποσοστό
  - Άξονας X: ποσοστό του στοχευμένου πληθυσμού
- Η διαγώνιος αντιπροσωπεύει την τυχαία απόδοση
  - Εάν στοχεύσουμε με εντελώς τυχαίο τρόπο στο 20% του συνόλου στιγμιοτύπων, τότε θα στοχεύσουμε και στο 20% των θετικών στιγμιοτύπων



# Καμπύλη Ανύψωσης (Lift Curve)

- Είναι η τιμή της αθροιστικής καμπύλης ανταπόκρισης σε ένα σημείο  $x$ , διαιρεμένη με την τιμή της διαγώνιας ευθείας ( $y=x$ ) στο δεδομένο σημείο
- Η καμπύλη ανύψωσης δείχνει την **αριθμητική ανύψωση (lift)** στον άξονα  $Y$  σε σχέση με το ποσοστό του πληθυσμού που στοχεύουμε



Σε αντίθεση με τις καμπύλες ROC, σε αυτές τις καμπύλες γίνεται η παραδοχή ότι το σύνολο δεδομένων δοκιμής έχει ακριβώς τις ίδιες εκ των προτέρων πιθανότητες της κατηγορίας-στόχου με τον πληθυσμό στον οποίο θα εφαρμοστεί το μοντέλο

# Η Έννοια της Υπερπροσαρμογής

# Γενίκευση και Υπερπροσαρμογή

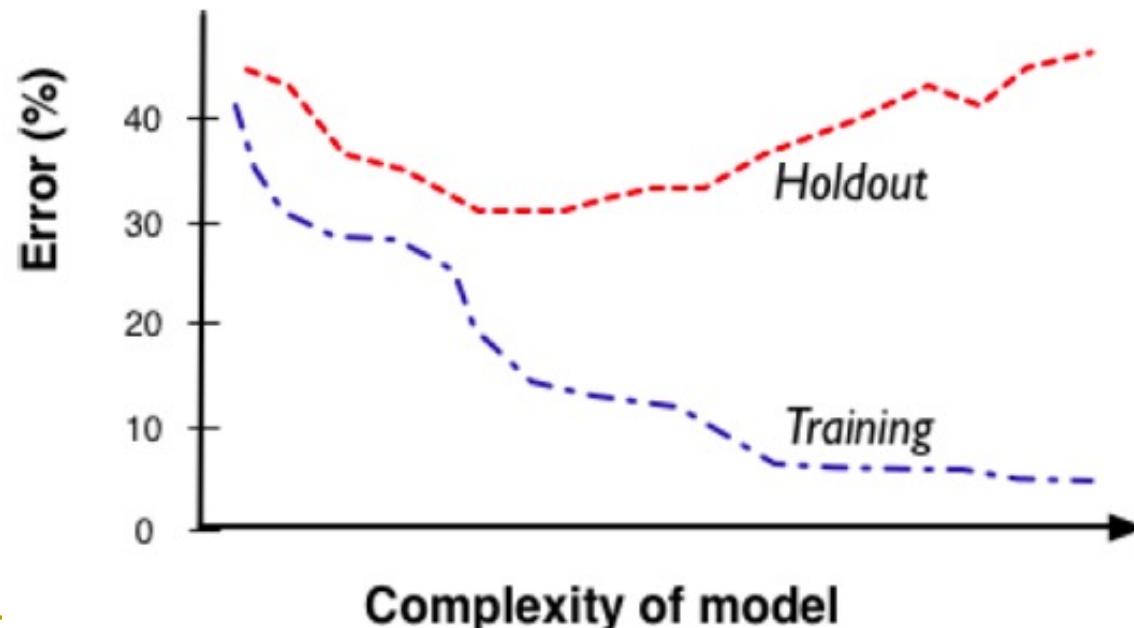
- Η γενίκευση (**generalization**) είναι η ιδιότητα ενός μοντέλου ή διαδικασίας μοντελοποίησης, στο πλαίσιο της οποίας **το μοντέλο εφαρμόζεται σε δεδομένα που δεν χρησιμοποιήθηκαν για την κατασκευή του μοντέλου**
- **Υπερπροσαρμογή (overfitting)** είναι η **τάση** των διαδικασιών ανάλυσης δεδομένων **να προσαρμόζουν τα μοντέλα στα δεδομένα εκπαίδευσης**, σε βάρος της γενίκευσης σε νέα δεδομένα

# Υπερπροσαρμογή – Σημειώσεις

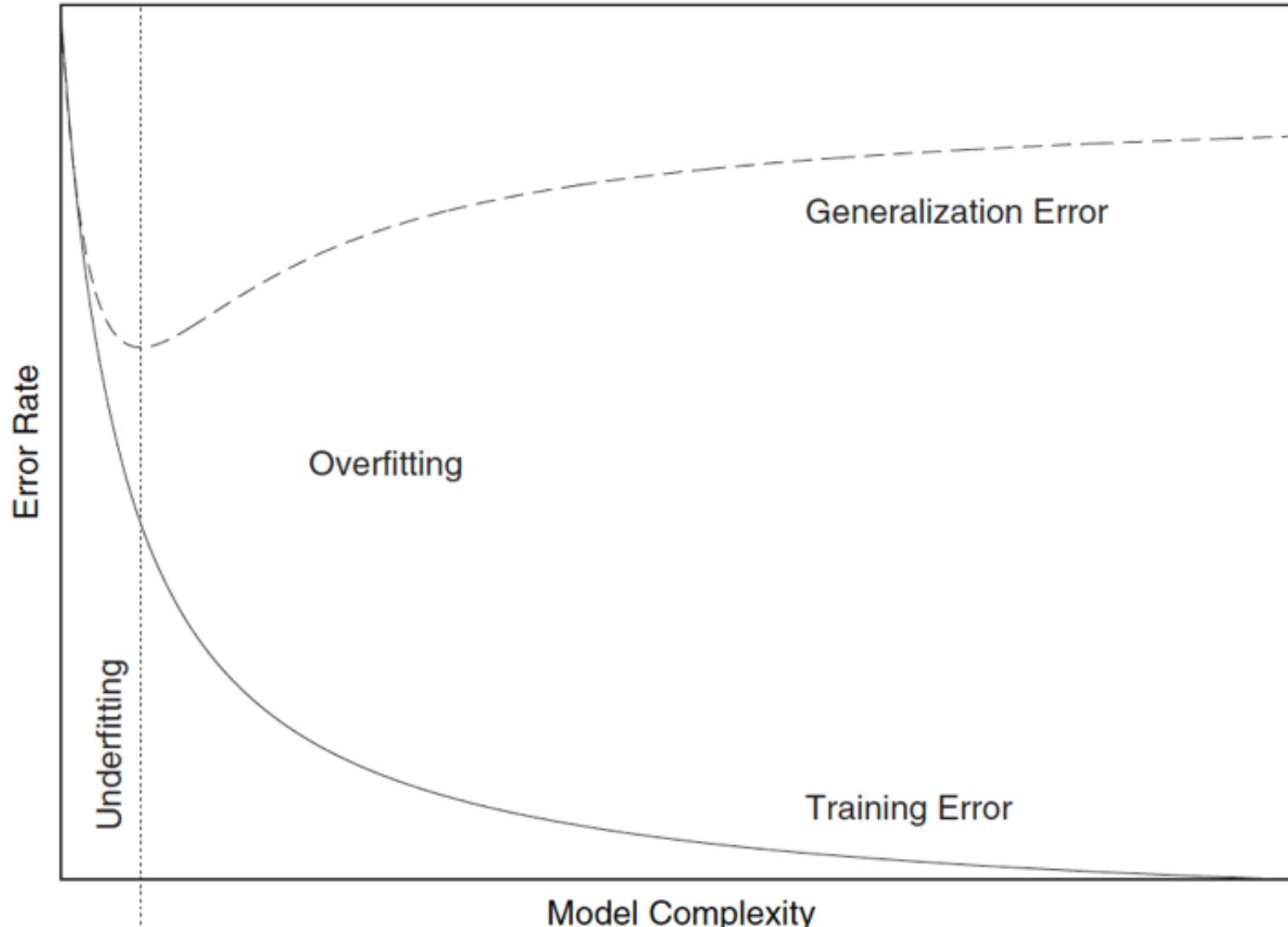
- Όλες οι διαδικασίες ανάλυσης/εξόρυξης δεδομένων έχουν την τάση να «υπερπροσαρμόζουν» σε κάποιο βαθμό
- Υπάρχει ένας **συμβιβασμός (tradeoff)** μεταξύ της ικανότητας ενός μοντέλου να **αντιπροσωπεύει πολύπλοκες σχέσεις** και της **πιθανότητας υπερπροσαρμογής**
  - Μερικές φορές θέλουμε πιο σύνθετα μοντέλα επειδή θα καταγράψουν καλύτερα την πραγματική πολυπλοκότητα της εφαρμογής και έτσι θα είναι πιο ακριβή
- Δεν υπάρχει κάποια μοναδική επιλογή ή διαδικασία που θα εξαλείψει την υπερπροσαρμογή
- Η καλύτερη στρατηγική είναι η **αναγνώριση της υπερπροσαρμογής** και η **διαχείριση της πολυπλοκότητας του μοντέλου** βάσει ορισμένων αρχών

# Παρακράτηση Δεδομένων και Γράφημα Προσαρμογής

- Το γράφημα προσαρμογής (fitting graph) δείχνει την ακρίβεια ενός μοντέλου ως συνάρτηση της πολυπλοκότητας
- Παρακράτηση δεδομένων (hold out data): δεδομένα που δε χρησιμοποιούνται για την κατασκευή του μοντέλου, των οποίων γνωρίζουμε την ετικέτα



# Υπερπλοσαρμογή (Overfitting)

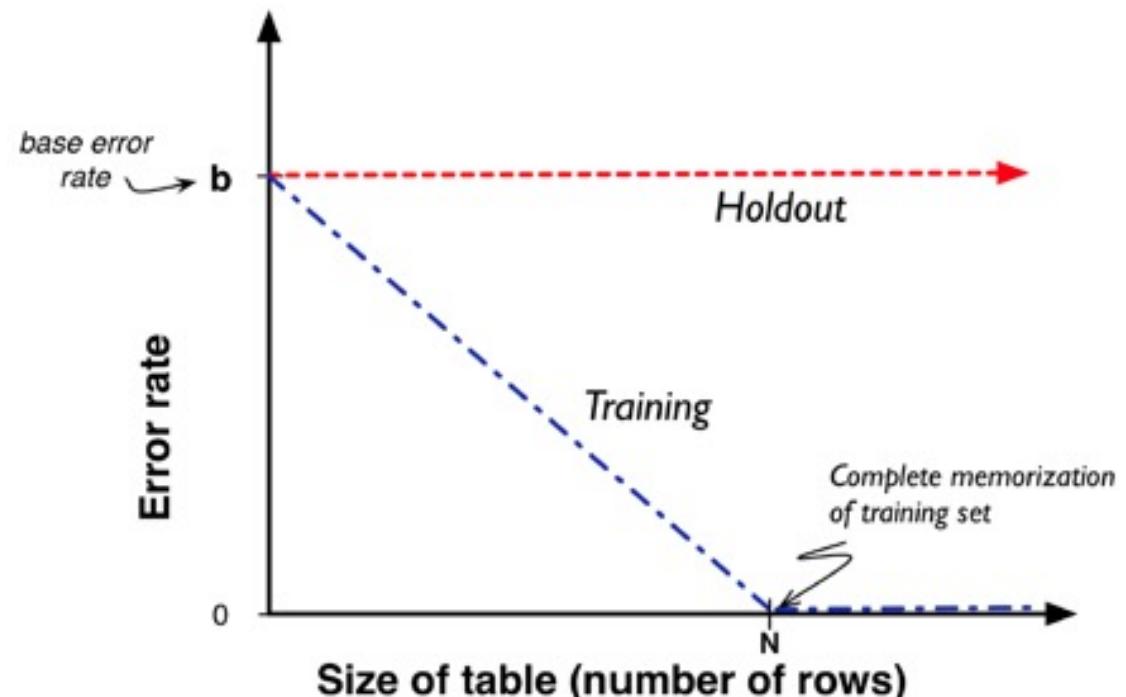


# Παράδειγμα «Μοντέλου Πίνακα»: Απομνημόνευση 'Όλων των Δεδομένων

- Έστω ότι θέλουμε να κατασκευάσουμε ένα μοντέλο που να διακρίνει τους πελάτες μιας εταιρείας που θα αποχωρήσουν
- Ιστορικά δεδομένα
  - Πελατών που παρέμειναν στην εταιρεία
  - Πελατών που διέκοψαν τη συνεργασία εντός 6 μηνών από τη λήξη του συμβολαίου
- Μοντέλο (*χάριν παραδείγματος μόνο!*)
  - Διατηρεί σε έναν πίνακα όλα τα χαρακτηριστικά κάθε πελάτη που αποχώρησε
  - Πρόβλεψη: αν ένας πελάτης βρίσκεται στον πίνακα → 100% πιθανότητα απώλειας, αλλιώς → 0%
  - Δίνει τέλειες προβλέψεις, όμως στην πράξη είναι τελείως άχρηστο
  - Όταν του δώσουμε έναν νέο πελάτη, δε θα τον βρει στον πίνακα, και áρα θα δίνει πάντα πιθανότητα 0%(!)

# Παράδειγμα «Μοντέλου Πίνακα»: Γράφημα Προσαρμογής

- Ποσοστό σφάλματος ως προς των αριθμών γραμμών που διατηρεί ο πίνακας
- Όταν ο πίνακας διατηρεί όλες τις  $N$  εγγραφές, το ποσοστό σφάλματος στο σύνολο εκπαίδευσης είναι μηδέν
- **Βασικό ποσοστό σφάλματος (base error rate):** το ποσοστό των περιπτώσεων απώλειας στον πληθυσμό
- **Κατηγοριοποιητής βασικού ποσοστού (base rate classifier):** επιλέγει πάντα την πλειοψηφική κατηγορία



# Υπερποσαρμογή στην Επαγωγή Δέντρου



# Υπερπροσαρμογή σε Μαθηματικές Συναρτήσεις

- Ένας τρόπος για να γίνουν οι μαθηματικές συναρτήσεις περισσότερο περίπλοκες είναι η **προσθήκη περισσότερων μεταβλητών (γνωρισμάτων)**

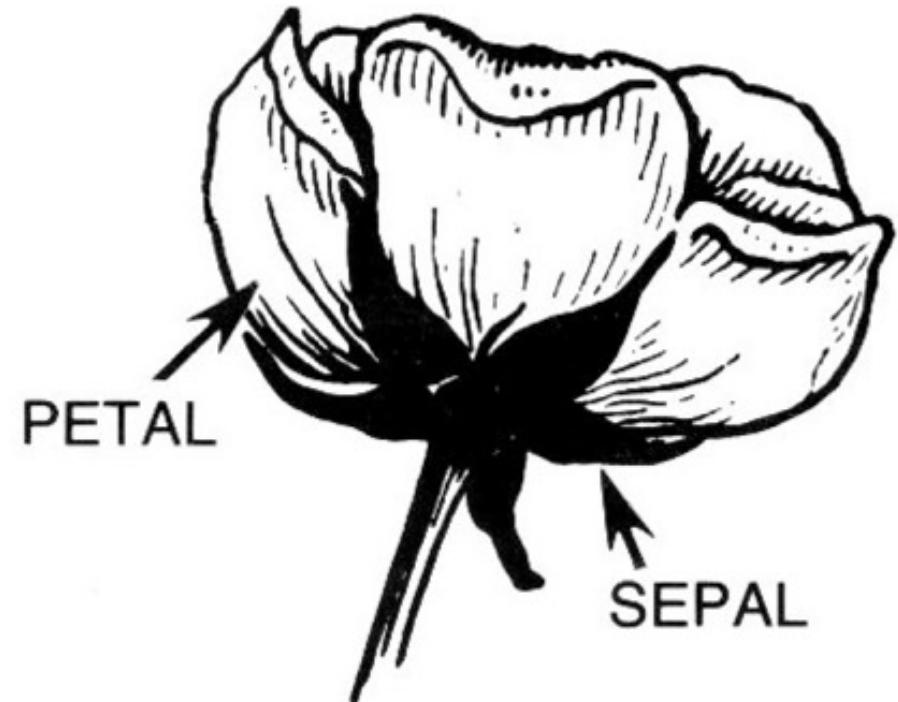
$$f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + w_3x_3 \rightarrow f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5$$

με προσθήκη των  $x_4=x_1^2$  και  $x_5=x_2/x_3$  καθώς μπορεί να είναι σημαντικές

- Όμως καθώς αυξάνει η διαστασιμότητα, μπορούμε να προσαρμόσουμε τέλεια ολοένα και μεγαλύτερα σύνολα σημείων
  - Για  $N$  σημεία μπορούμε να προσαρμόσουμε τέλεια ένα υπερεπίπεδο στις  $N$  διαστάσεις (γενίκευση της ευθείας για 2 σημεία στις 2 διαστάσεις)

# Παράδειγμα: Υπερπροσαρμογή Γραμμικών Συναρτήσεων

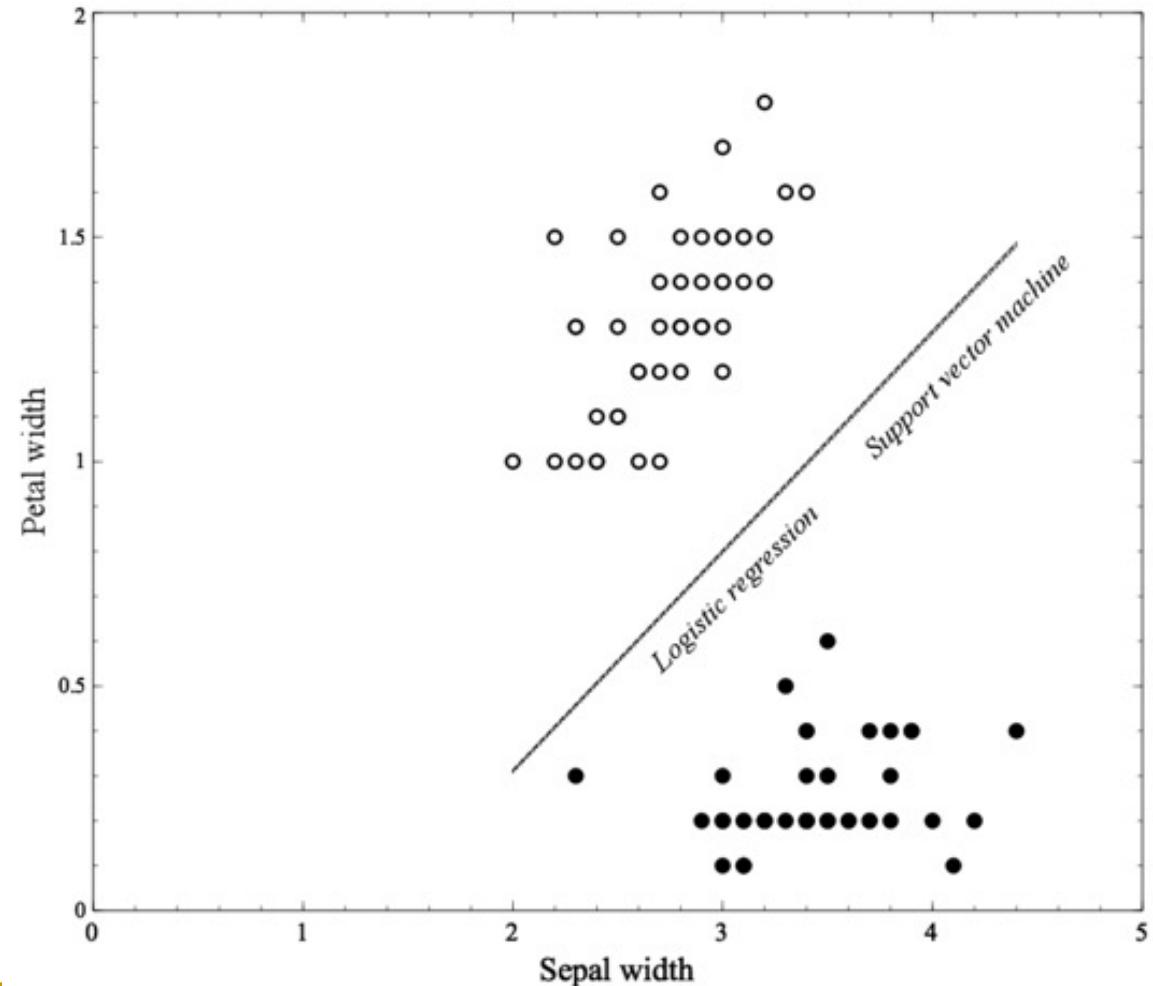
- Το σύνολο δεδομένων Iris (ένα γένος ανθοφόρου φυτού)
- Τέσσερα χαρακτηριστικά
- Τρεις κατηγορίες



<http://archive.ics.uci.edu/ml/datasets/Iris>

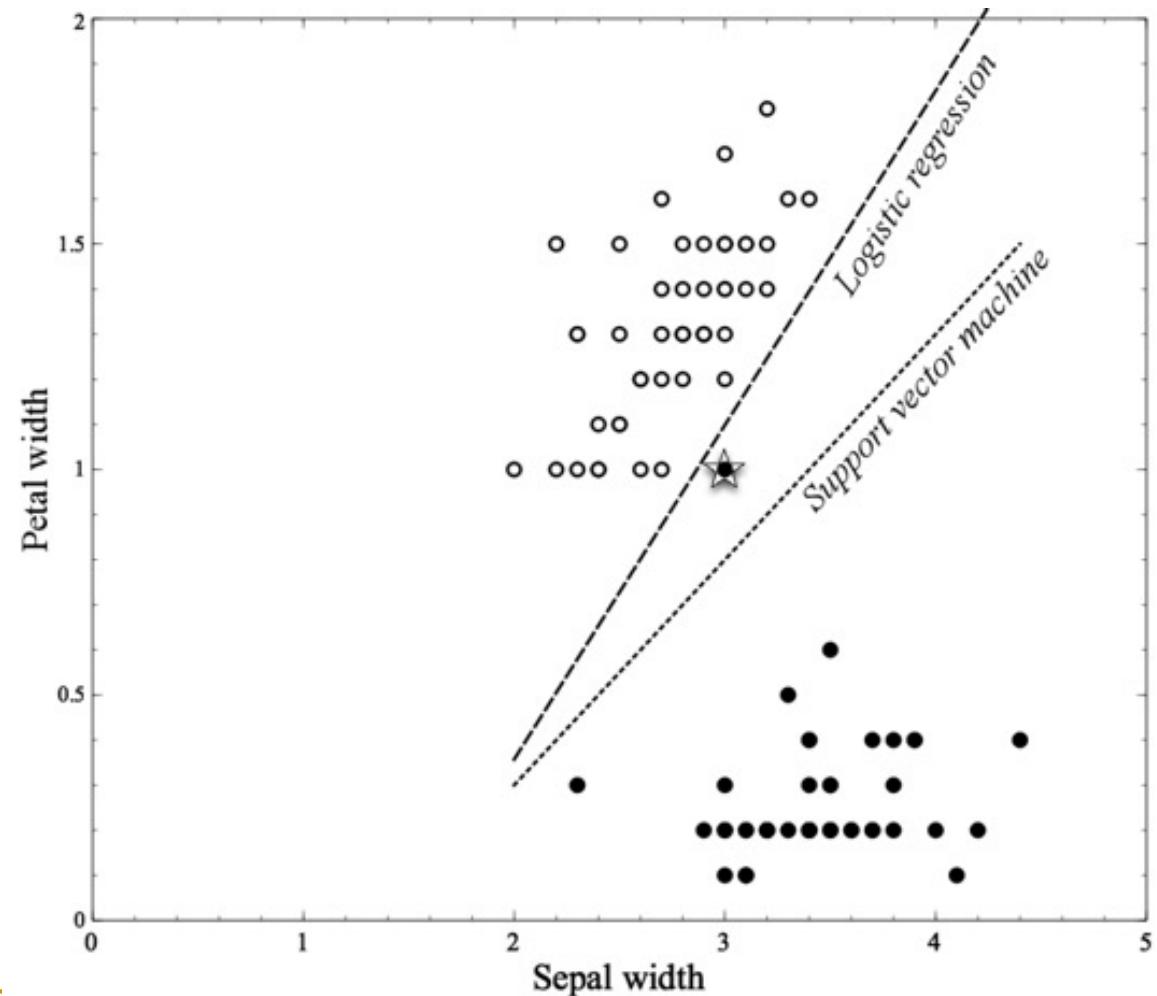
# Παράδειγμα: Υπερπροσαρμογή Γραμμικών Συναρτήσεων

- Οι δύο κατηγορίες (κουκίδες άσπρες και μαύρες) είναι πολύ διακριτές και διαχωρίσιμες
- Τόσο η λογιστική παλινδρόμηση όσο και το SVM καταλήγουν σε παρόμοιο μοντέλο ευθείας



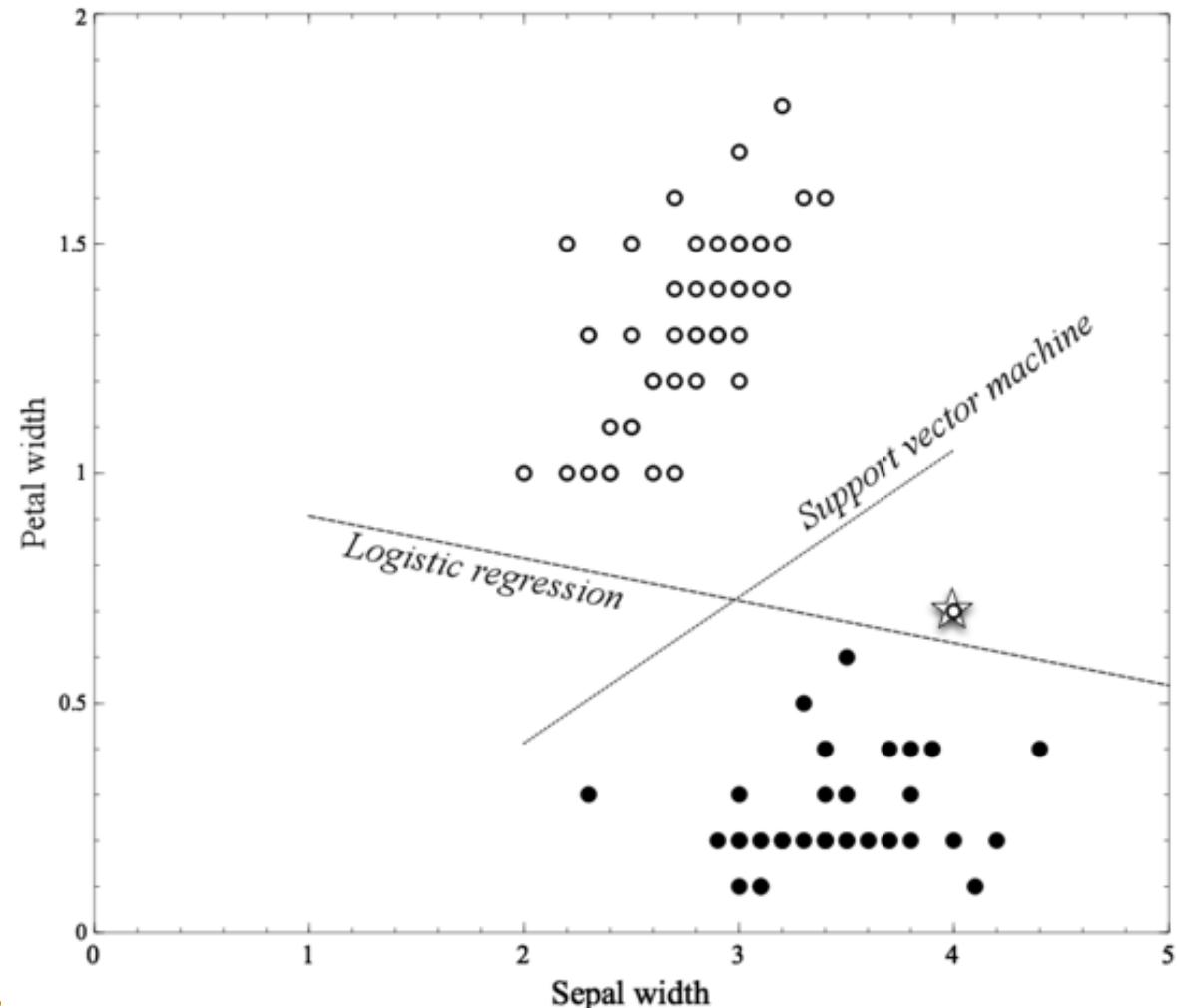
# Παράδειγμα: Υπερπροσαρμογή Γραμμικών Συναρτήσεων

- Προσθέτουμε ένα νέο σημείο (3,1) (που μοιάζει με outlier)
- Το μοντέλο λογιστικής παλινδρόμησης έχει αλλάξει σημαντικά
- Η λογιστική παλινδρόμηση φαίνεται να κάνει υπερπροσαρμογή
- Περισσότερο ευαίσθητη τεχνική σε outliers



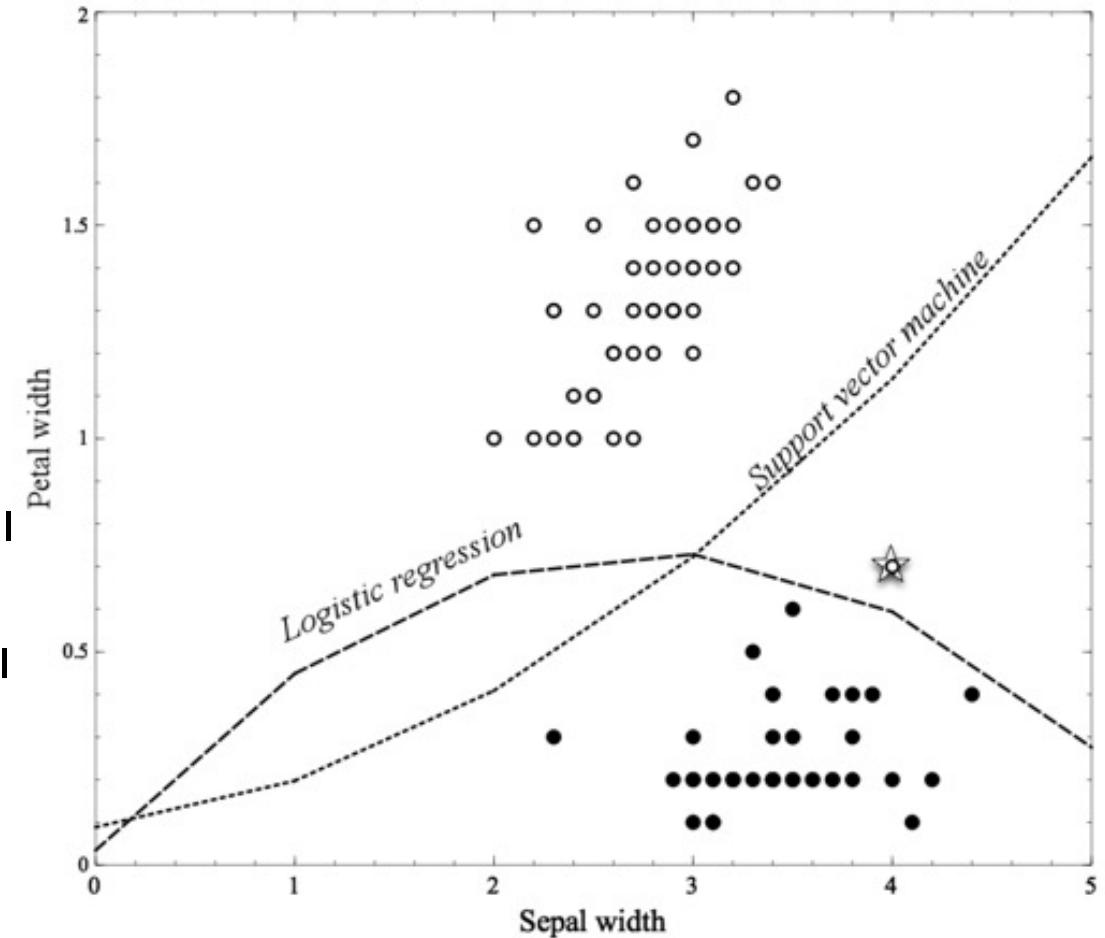
# Παράδειγμα: Υπερπροσαρμογή Γραμμικών Συναρτήσεων

- Προσθέτουμε άλλο σημείο  $(4, 0.7)$
- Πάλι η ευθεία του SVM μετατοπίζεται ελάχιστα, ενώ η αντίστοιχη της λογιστικής παλινδρόμησης αλλάζει σημαντικά



# Παράδειγμα: Υπερπροσαρμογή Γραμμικών Συναρτήσεων

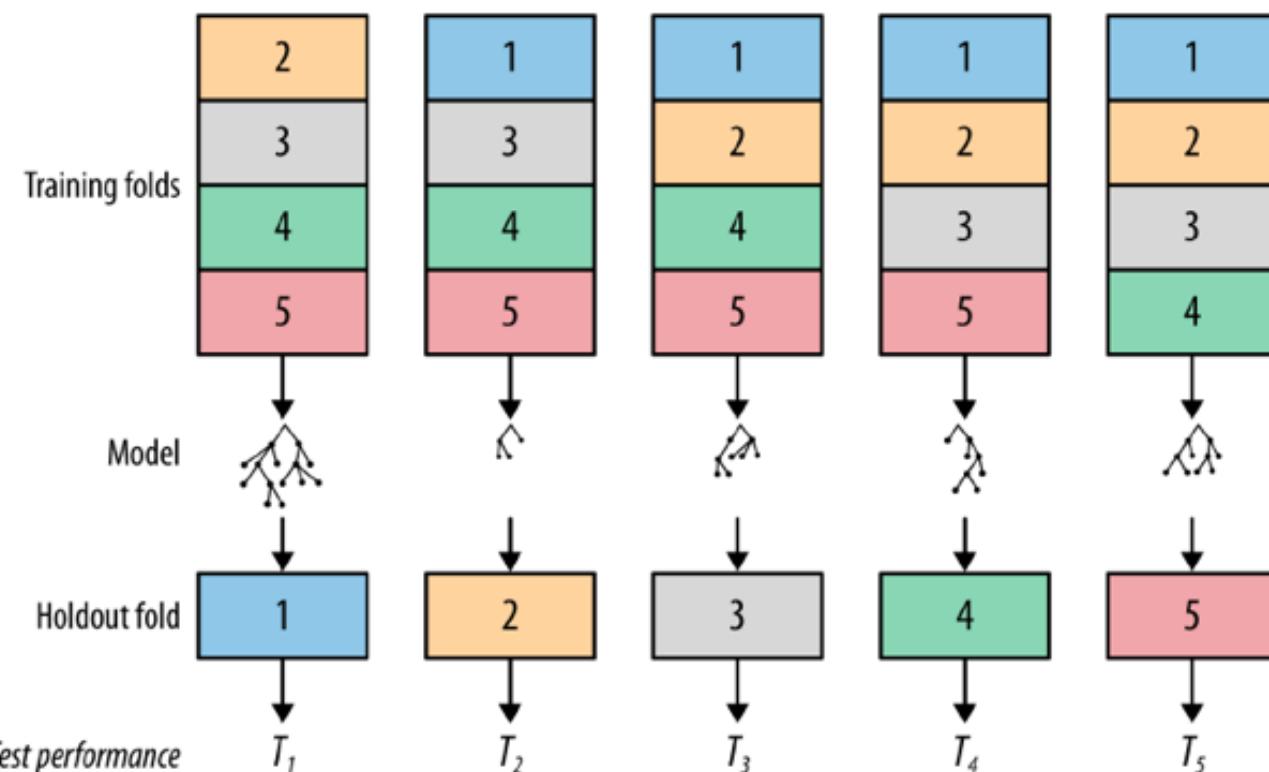
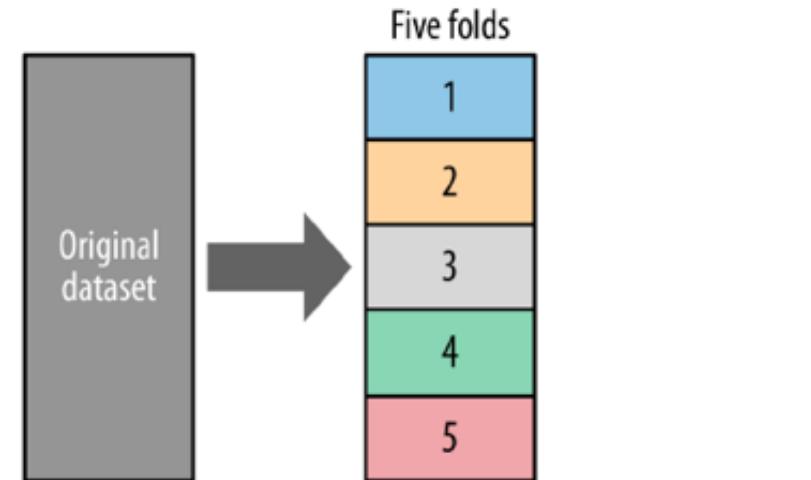
- Προσθήκη γνωρίσματος:  $(\text{Sepal width})^2$
- Ως αποτέλεσμα και οι δύο μέθοδοι παράγουν τιο σύνθετα, μη γραμμικά μοντέλα
- Γεωμετρικά, το όριο διαχωρισμού μπορεί να είναι όχι μόνο ευθεία αλλά και παραβολή (προσαρμόζονται καλύτερα στα δεδομένα)
- Αυτό δημιουργεί όμως περισσότερες ευκαιρίες για υπερπροσαρμογή



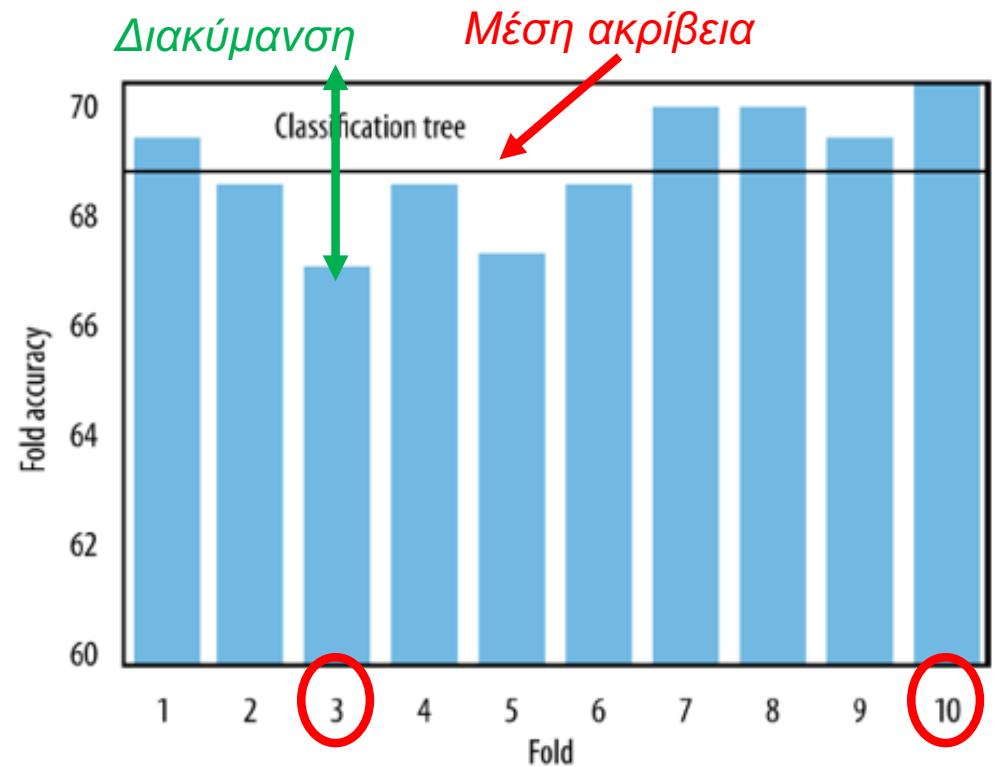
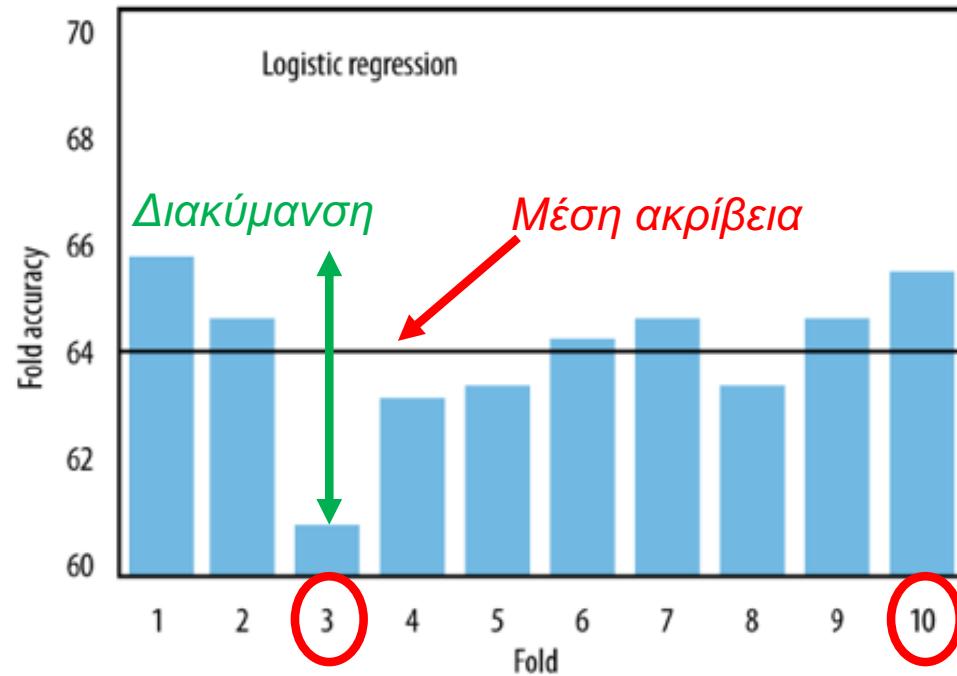
# Διασταυρωτική Επικύρωση

- Η διασταυρωτική επικύρωση (cross-validation) είναι μια πιο εξελιγμένη διαδικασία εκπαίδευσης και δοκιμής με παρακράτηση δεδομένων
  - Διαχωρίζουμε το σύνολο δεδομένων σε **k** διαμερίσεις που ονομάζονται **πτυχές** (folds)
    - Συνήθως  $k=5$  ή  $10$
  - Επαναλαμβάνουμε **k** φορές την εκπαίδευση και δοκιμή
  - Σε κάθε επανάληψη
    - **k-1** πτυχές χρησιμοποιούνται για **εκπαίδευση**
    - **1** (διαφορετική) πτυχή χρησιμοποιείται για **δοκιμή**
  - Τελικά υπολογίζουμε μέση τιμή και τυπική απόκλιση για τη μετρούμενη ποσότητα

## Παράδειγμα διασταυρωτικής επικύρωσης (cross-validation)



# Παράδειγμα: Πρόβλημα Απώλειας Πελατών



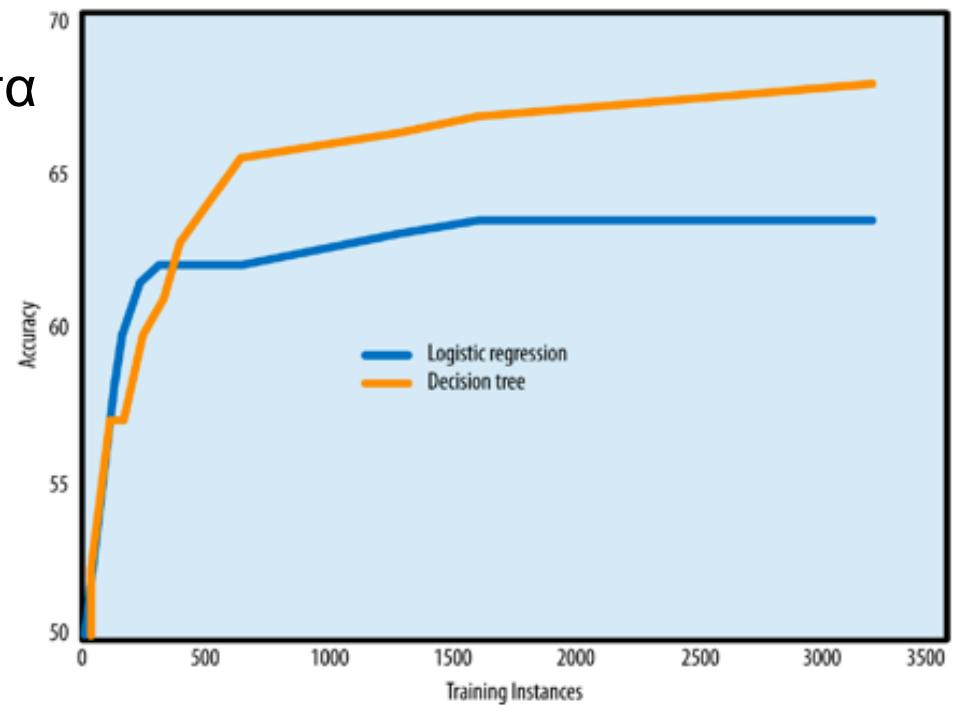
- Επαγωγή δέντρου πετυχαίνει 73% (χωρίς διασταυρωτική επικύρωση) (βλ. προηγούμενη διάλεξη για δέντρα απόφασης)
- Άρα υπήρξε κάποια υπερπροσαρμογή

# Καμπύλες Μάθησης

- Οι καμπύλες μάθησης (learning curves) αποτελούν το διάγραμμα που δείχνει την απόδοση στη γενίκευση ως συνάρτηση της ποσότητας των δεδομένων εκπαίδευσης
  - Σχηματικά, μια καμπύλη μάθησης είναι αρχικά απότομη (καθώς βρίσκει τους πιο εμφανείς κανόνες)
  - Σταδιακά έχει μια ελαφριά αυξητική τάση (βρίσκει πιο ακριβή μοντέλα όταν εκπαιδεύεται σε μεγαλύτερα σύνολα δεδομένων)
  - Όσπου μερικές φορές γίνεται εντελώς ευθεία (δεν μπορεί να βελτιώσει την ακρίβεια ακόμα και με περισσότερα δεδομένα)

# Καμπύλες Μάθησης

- Διαφορετική βελτίωση για τις δύο τεχνικές επαγωγής με την ποσότητα των δεδομένων εκπαίδευσης
- Η λογιστική παλινδρόμηση
  - Λιγότερη ευελιξία και λιγότερη υπερπροσαρμογή σε λίγα δεδομένα
  - Μακροπρόθεσμα δεν μπορεί όμως να μοντελοποιήσει πλήρως την πολυπλοκότητα των δεδομένων
- Η επαγωγή δέντρου
  - Είναι πολύ πιο ευέλικτη, οπότε κάνει περισσότερη υπερπροσαρμογή σε λίγα δεδομένα
  - Όμως έτσι μπορεί να μοντελοποιήσει καλύτερα σε μεγαλύτερα σύνολα δεδομένων εκπαίδευσης



# Καμπύλες Μάθησης

- **Καμπύλη μάθησης vs. Γράφημα προσαρμογής**
  - Και τα δύο δείχνουν την απόδοση στη **γενίκευση**
  - Η **καμπύλη μάθησης** συναρτήσει της **ποσότητας δεδομένων εκπαίδευσης**
  - Το **γράφημα προσαρμογής** συναρτήσει της **πολυπλοκότητας του μοντέλου** (άρα για σταθερή ποσότητα δεδομένων εκπαίδευσης)
- Η καμπύλη μάθησης μπορεί να δείξει ότι οι επιδόσεις στη γενίκευση έχουν σταθεροποιηθεί, και άρα δε χρειάζεται να συλλέξουμε κι άλλα δεδομένα εκπαίδευσης

# Αποφυγή Υπερπροσαρμογής και Έλεγχος Πολυπλοκότητας

- Ως παράδειγμα παίρνουμε την επαγωγή δέντρου, όπου το δέντρο συνεχίζει να αυξάνεται ώσπου να δημιουργήσει ομοιογενείς κόμβους-φύλλα (υπερπροσαρμόζεται στα δεδομένα)
- Δύο τεχνικές αντιμετώπισης
  - **Σταματάμε να επεκτείνουμε το δέντρο**
    - Βάσει τιμής κατωφλιού (=ελάχιστος αριθμός στιγμιότυπων σε ένα φύλλο) ή
    - Στατιστικό έλεγχο υποθέσεων για το αν η παρατηρούμενη διαφορά στο κέρδος πληροφορίας είναι τυχαία
  - **Εκ των υστέρων, «κλαδεύουμε» το δέντρο**
    - Π.χ. εκτιμώντας αν η αντικατάσταση ενός κλάδου από ένα φύλλο θα μείωνε την ακρίβεια
- (**γενική ιδέα**): Αν κατασκευάζαμε δέντρα με διάφορα επίπεδα πολυπλοκότητας και εκτιμούσαμε την ακρίβειά τους στη γενίκευση, θα μπορούσαμε να κρατήσουμε το καλύτερο

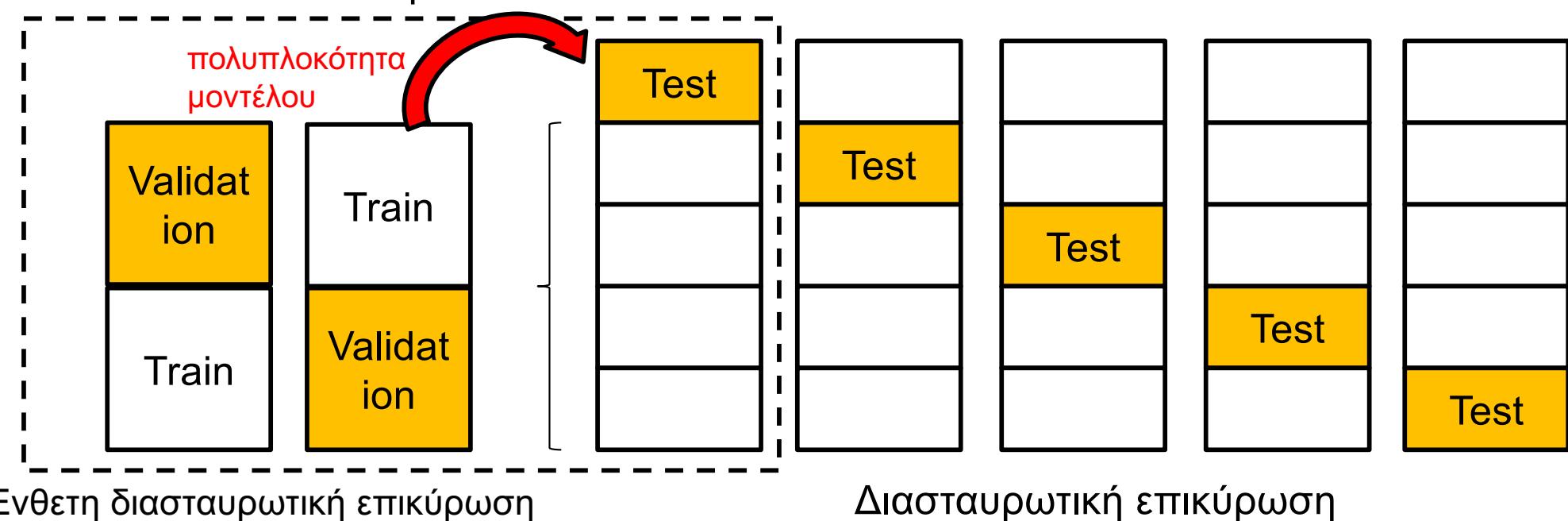
# Γενική Μέθοδος Αποφυγής Υπερπροσαρμογής

- Δοκιμή με ένθετη παρακράτηση δεδομένων (**nested holdout test**):
  - Χωρίζουμε το σύνολο δεδομένων εκπαίδευσης σε: **σύνολο υποεκπαίδευσης (subtraining set)** και **σύνολο επικύρωσης (validation set)**
  - Με αυτά τα δύο σύνολα βρίσκουμε τη **βέλτιστη πολυπλοκότητα** του μοντέλου (π.χ. το βέλτιστο μέγεθος δέντρου για επαγωγή δέντρου)
  - Κατόπιν, **εκπαιδεύουμε** το μοντέλο **με όλο το σύνολο εκπαίδευσης**
  - Εκτιμούμε την απόδοση του μοντέλου στα δεδομένα δοκιμής, που είναι ανεξάρτητα από τη δημιουργία του μοντέλου



# Γενική Μέθοδος Αποφυγής Υπερπροσαρμογής

- Ένθετη διασταυρωτική επικύρωση (**nested cross-validation**)
  - Εκτελούμε διασταυρωτική επικύρωση (cross-validation)
  - Πριν από την κατασκευή μοντέλου από κάθε πτυχή, εκτελούμε μια μικρότερη διασταυρωτική επικύρωση για να βρούμε τη βέλτιστη πολυπλοκότητα



# Ολοκληρωμένο Παράδειγμα Σύγκρισης Κατηγοριοποιητών

# 'Ενα Ολοκληρωμένο Παράδειγμα

- Μοντελοποίηση της απώλειας πελατών από εταιρεία τηλεφωνίας
- Σύνολο δεδομένων από το διαγωνισμό KDD Cup 2009
- Συνολικά **47.000 στιγμιότυπα** από τα οποία
  - 7% πελάτες που έχουν **αποχωρήσει** (θετικά παραδείγματα) και
  - 93% αφορά πελάτες που έχουν **παραμείνει** (αρνητικά παραδείγματα)
  - Βασικό ποσοστό = 93%
- Στόχος: συγκριτική αξιολόγηση διάφορων κατηγοριοποιητών
  - decision tree
  - logistic regression
  - k-nearest neighbor
  - naïve Bayes

# 'Ενα Ολοκληρωμένο Παράδειγμα

Αποτελέσματα **ακρίβειας** με εκπαίδευση και δοκιμή **σε όλο το σύνολο δεδομένων** (*αφελής προσέγγιση*)

Model	Accuracy
Classification tree	95%
Logistic regression	93%
<i>k</i> -Nearest Neighbor	100%
Naive Bayes	76%

Παρατηρήσεις:

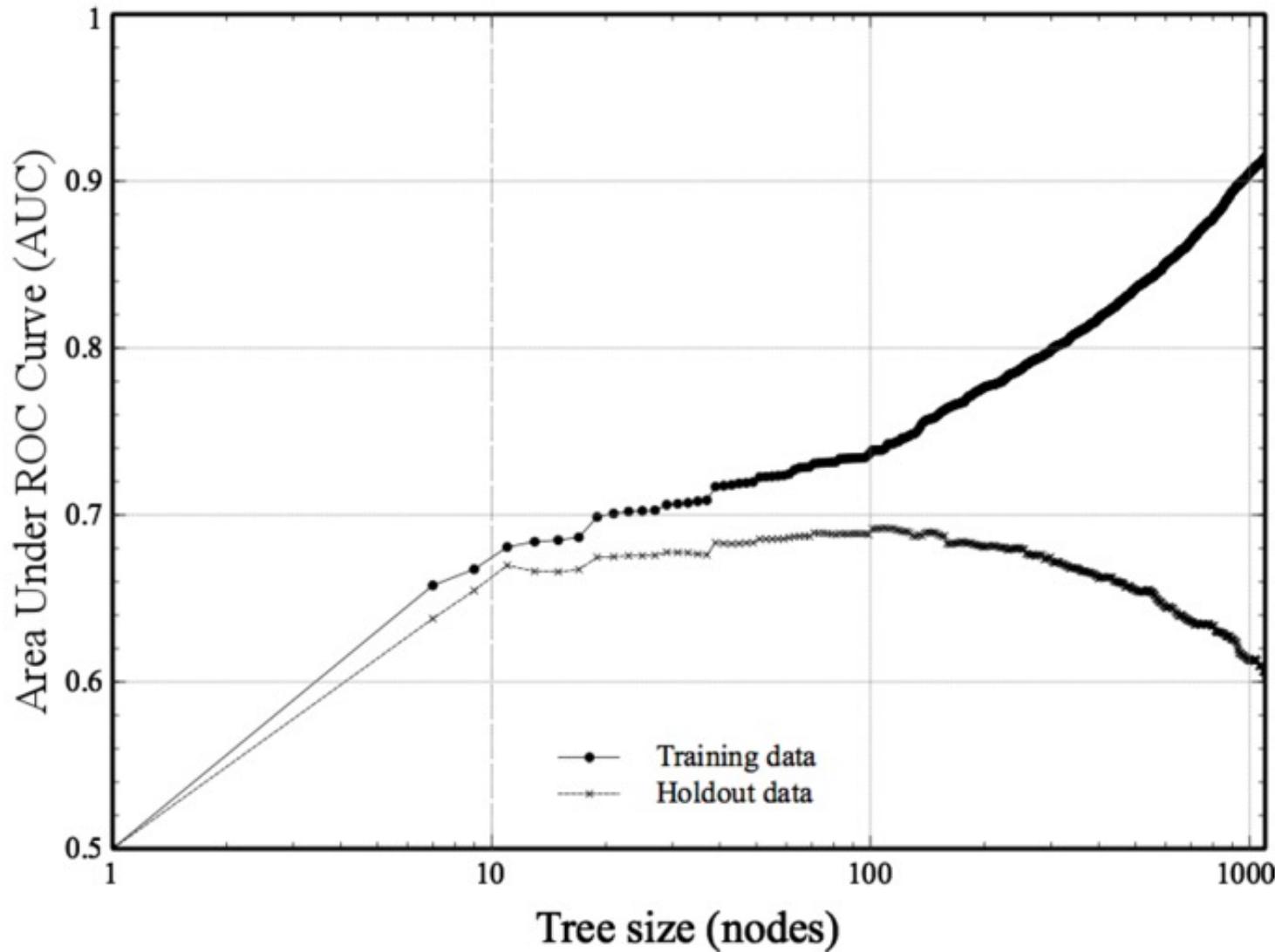
- (1) Εύρος τιμών: 76% - 100%
- (2) NB (76%) χειρότερο από το βασικό ποσοστό (93%)
- (3) *k*-NN (100%) !?!

# 'Ενα Ολοκληρωμένο Παράδειγμα

Model	Accuracy (%)	AUC
Classification Tree	$91.8 \pm 0.0$	$0.614 \pm 0.014$
Logistic Regression	$93.0 \pm 0.1$	$0.574 \pm 0.023$
k-Nearest Neighbor	$93.0 \pm 0.0$	$0.537 \pm 0.015$
Naive Bayes	$76.5 \pm 0.6$	$0.632 \pm 0.019$

- Με διασταυρωτική επικύρωση ( $k=10$ )
- Μέση τιμή ακρίβειας (των  $k$ )
  - Οι τυπικές αποκλίσεις είναι μικρές σε σχέση με τους μέσους (καλό, δεν υπάρχει μεγάλη διαφοροποίηση στην απόδοση στις διάφορες πτυχές)
- Τιμές AUC
  - Μέτριες τιμές, υποδηλώνουν ότι το πρόβλημα είναι «δύσκολο»
  - Αντίστροφη εικόνα για τον Naïve Bayes

# 'Ενα Ολοκληρωμένο Παράδειγμα



# 'Ενα Ολοκληρωμένο Παράδειγμα

	p	n
Y	127 (3%)	848 (18%)
N	200 (4%)	3518 (75%)

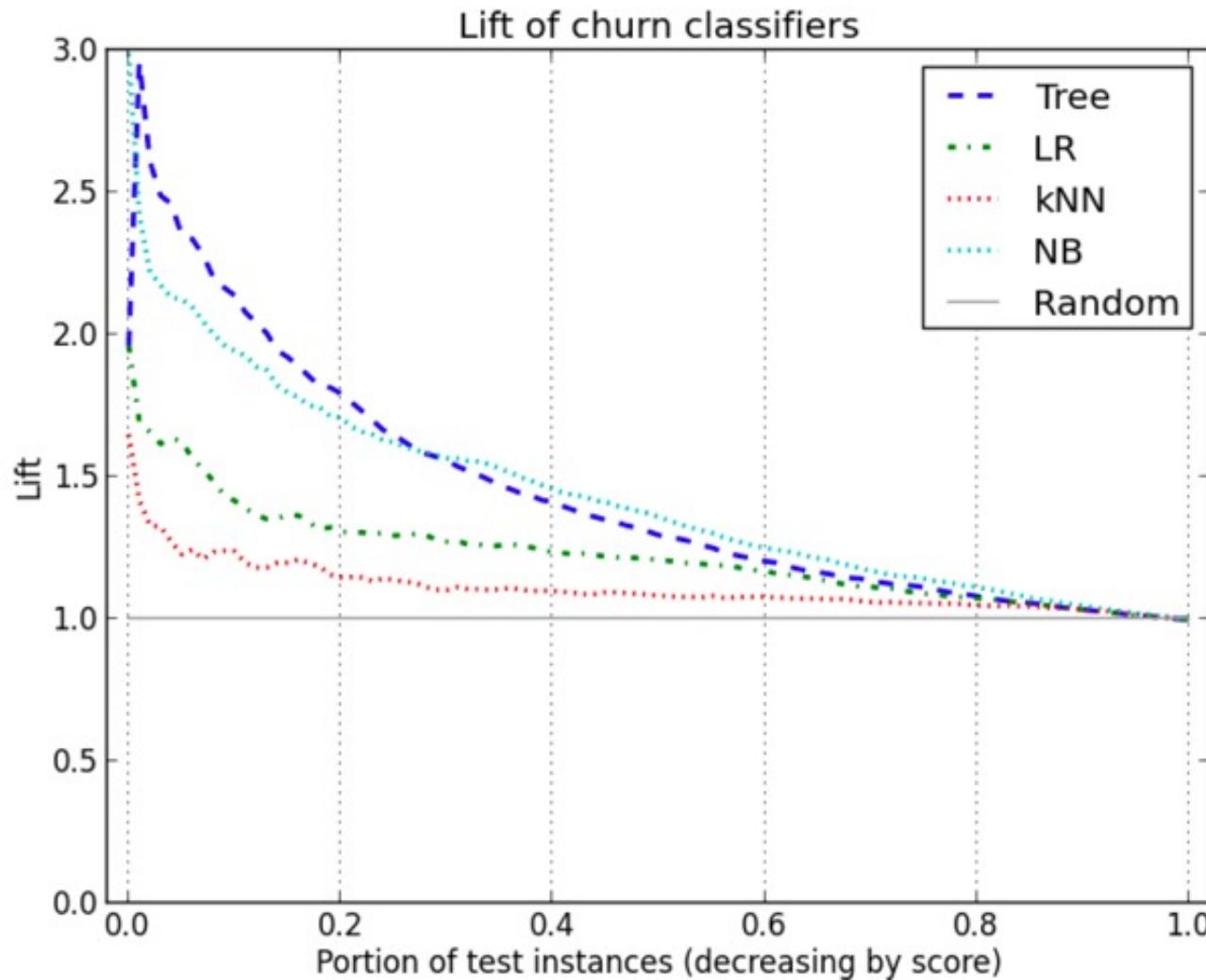
Naïve Bayes Classifier

	p	n
Y	3 (0%)	15 (0%)
N	324 (7%)	4351 (93%)

K-NN Classifier

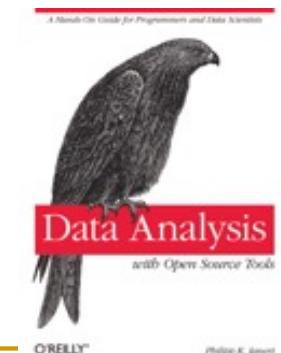
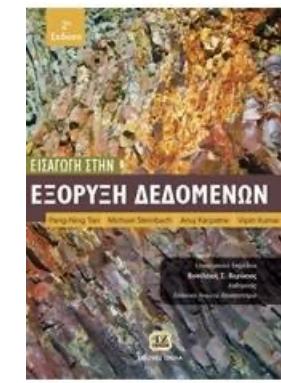
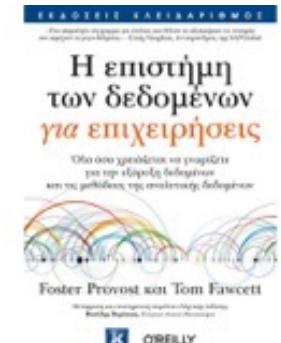
- Ο k-NN classifier σπάνια προβλέπει απώλεια πελατών
  - Λειτουργεί σαν baseline κατηγοριοποιητής
- Ο Naive Bayes classifier κάνει περισσότερα λάθη και για αυτό έχει χαμηλότερη ακρίβεια
  - Όμως εντοπίζει πολύ περισσότερους από αυτούς που πρόκειται να αποχωρήσουν

# 'Ενα Ολοκληρωμένο Παράδειγμα



# Πηγές Αναφοράς

- F. Provost, T. Fawcett. “**Η Επιστήμη των Δεδομένων για Επιχειρήσεις**”. Εκδόσεις Κλειδάριθμος.
  - Κεφ. 5: Η υπερπροσαρμογή και η αποφυγή της
  - Κεφ. 7: Αναλυτικός τρόπος σκέψης για τη λήψη αποφάσεων I
  - Κεφ. 8: Οπτικοποίηση της απόδοσης των μοντέλων
- P. Tan, M. Steinbach, V. Kumar. “**Εισαγωγή στην Εξόρυξη Δεδομένων**”. Εκδόσεις Τζιόλα.
  - Κεφ. 4: Κατηγοριοποίηση: Βασικές Έννοιες
  - Κεφ. 5: Κατηγοριοποίηση: Εναλλακτικές Τεχνικές
- P.K. Janert. “**Data Analysis with Open Source Tools**”. O'Reilly, 2011.
  - Κεφ. 18: Predictive Analytics





## 7. Εύρεση Συστάδων (Clusters)



Ανάλυση Δεδομένων  
(*Data Analytics*)

Χρήστος Δουλκερίδης  
2024-25

# Ορισμός Συσταδοποίησης

- Η **συσταδοποίηση (clustering)** είναι η
  - διαδικασία εύρεσης ομάδων
  - μεταξύ των αντικειμένων (σημείων) ενός συνόλου δεδομένων
- Μέθοδος **μη εποπτευόμενης** μάθησης (**unsupervised learning**)
  - Δε γνωρίζουμε εκ των προτέρων
  - πού βρίσκονται οι συστάδες ή πώς μοιάζουν
- Σε αντίθεση με την εποπτευόμενη μάθηση (**supervised learning**)
  - Όπου τα αντικείμενα ανατίθενται σε προϋπάρχουσες ομάδες
  - Ομάδες: τιμές γνωρίσματος – ετικέτα

# Περίγραμμα Διάλεξης

- Τι ορίζεται ως μια συστάδα (cluster);
- Μέτρα απόστασης και ομοιότητας
- Μέθοδοι συσταδοποίησης
  - Αλγόριθμοι που αναζητούν κέντρα (k-means)
  - Αλγόριθμοι κατασκευής δέντρων (HAC)
  - Αλγόριθμοι μεγέθυνσης γειτονιών (DBSCAN)
- Προεπεξεργασία και μετεπεξεργασία

# Τι Ορίζεται ως μια Συστάδα (cluster);

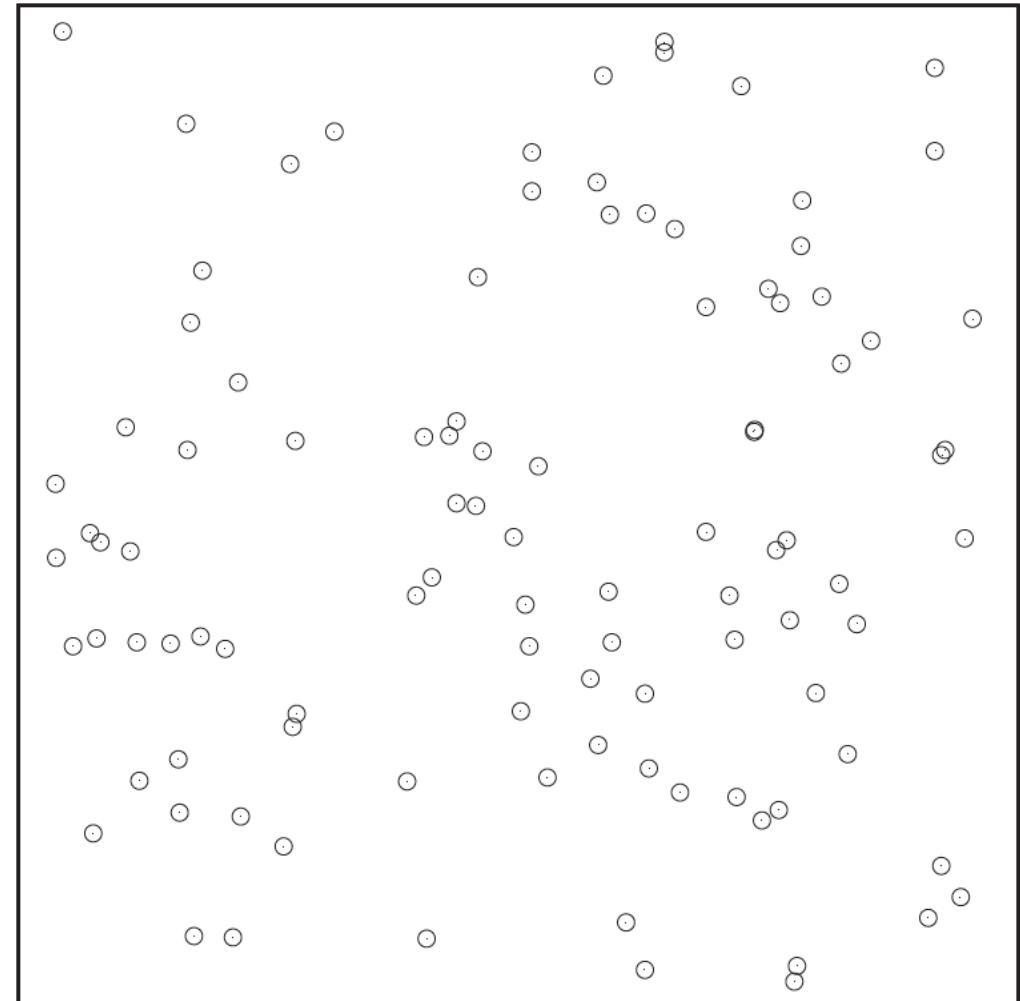
- Η έννοια της συστάδας δεν είναι καλά ορισμένη
- Πιθανές περιγραφές του τι αποτελεί μια συστάδα:
  - Ομάδες αντικειμένων που **μοιάζουν** ή
  - Ομάδες σημείων που **βρίσκονται κοντά** στο χώρο
- Επίσης, οι συστάδες πρέπει να είναι **καλά διαχωρισμένες** μεταξύ τους

# Απεικόνιση Ομοιόμορφα Κατανεμημένων Δεδομένων

Δεν υπάρχουν ευδιάκριτες συστάδες

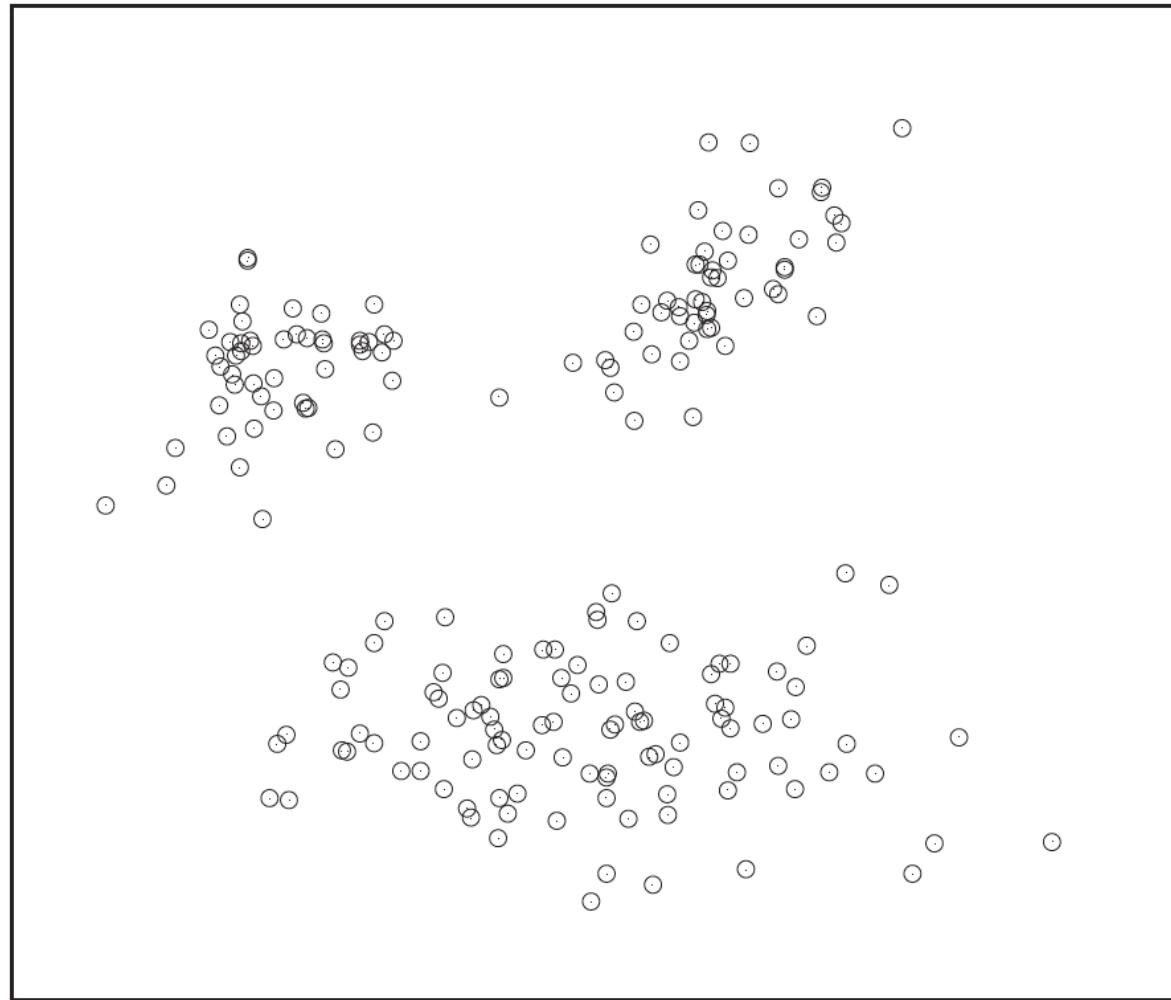
## Πιθανός ορισμός:

συνεχείς περιοχές υψηλής πυκνότητας σημείων,  
διαχωρισμένες από περιοχές χαμηλότερης πυκνότητας σημείων



# Τρεις Καλά Διαχωρισμένες, Σφαιροειδείς Συστάδες

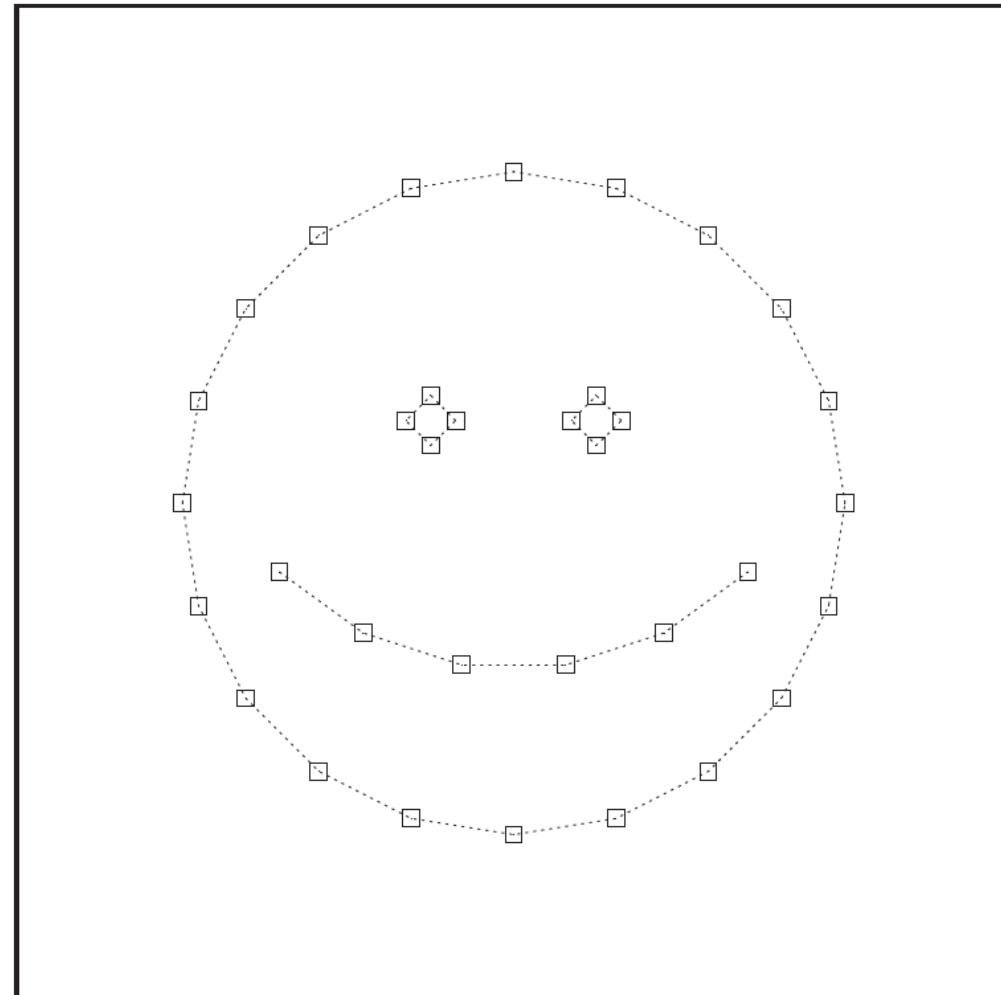
Συμβατές με τον προηγούμενο ορισμό



# Παράδειγμα Μη-σφαιροειδών Συστάδων

Επίσης συμβατές με τον  
προηγούμενο ορισμό

Μερικές συστάδες είναι  
εμφωλευμένες, δηλαδή  
περιέχονται πλήρως μέσα σε  
άλλες



# Συσταδοποίηση Άλλων Τύπων Δεδομένων

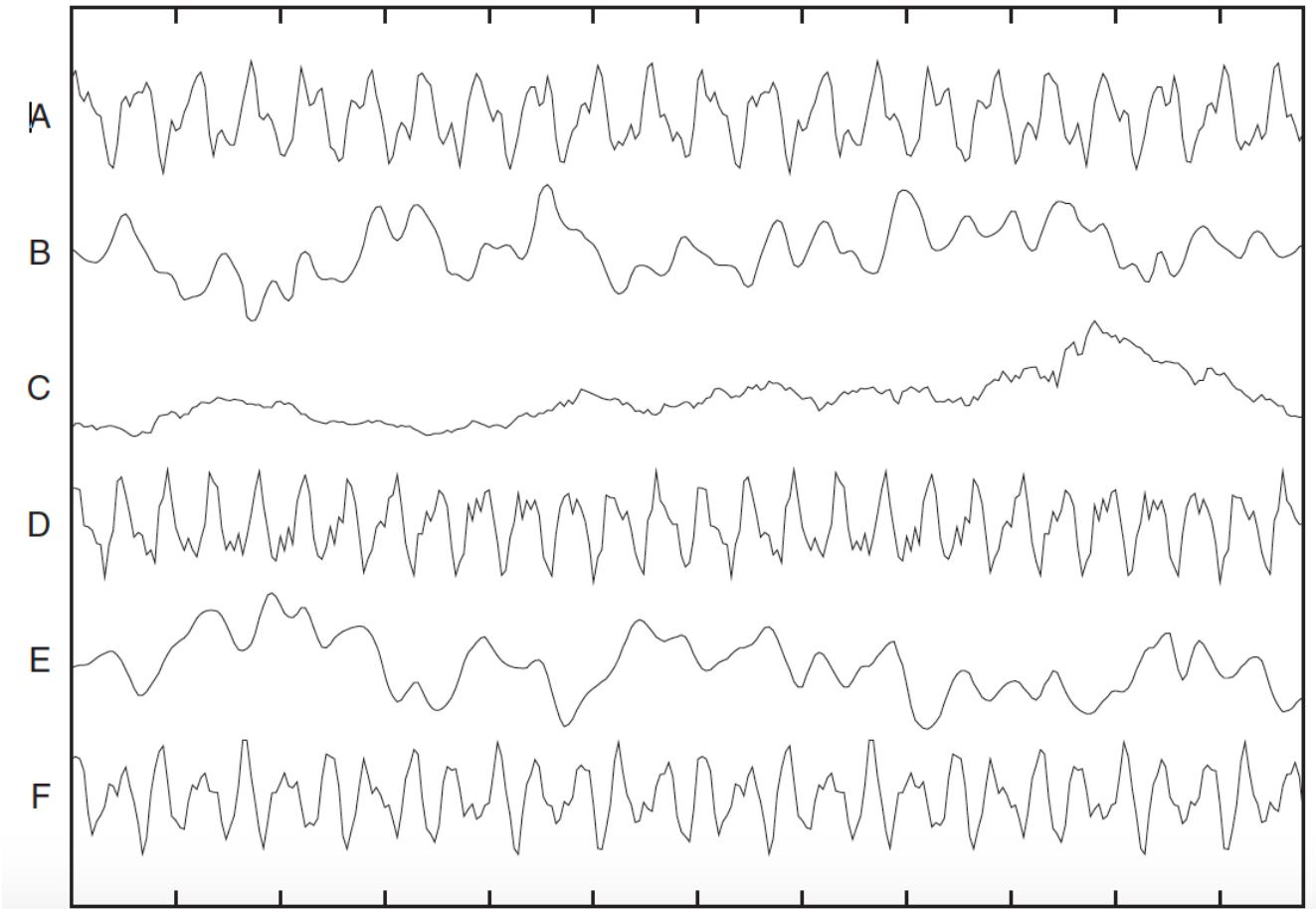
First Avenue 35	48 Second Street	Main Boulevard 9
First Avenue 53	Main Blvd 19	Mn Boulevard 11
45 Second Street E	45 Second St	First Ave 35
Furst Avenue 33	44 second street	Main Boulevrd 1
1st Avenue 53	Second Street, 48	Main Bulevard 19

Συσταδοποίηση αλφαριθμητικών:

- παρόλο που δεν υπάρχουν ίδια αλφαριθμητικά
- μπορούμε να διακρίνουμε αρκετές ομάδες όμοιων αλφαριθμητικών

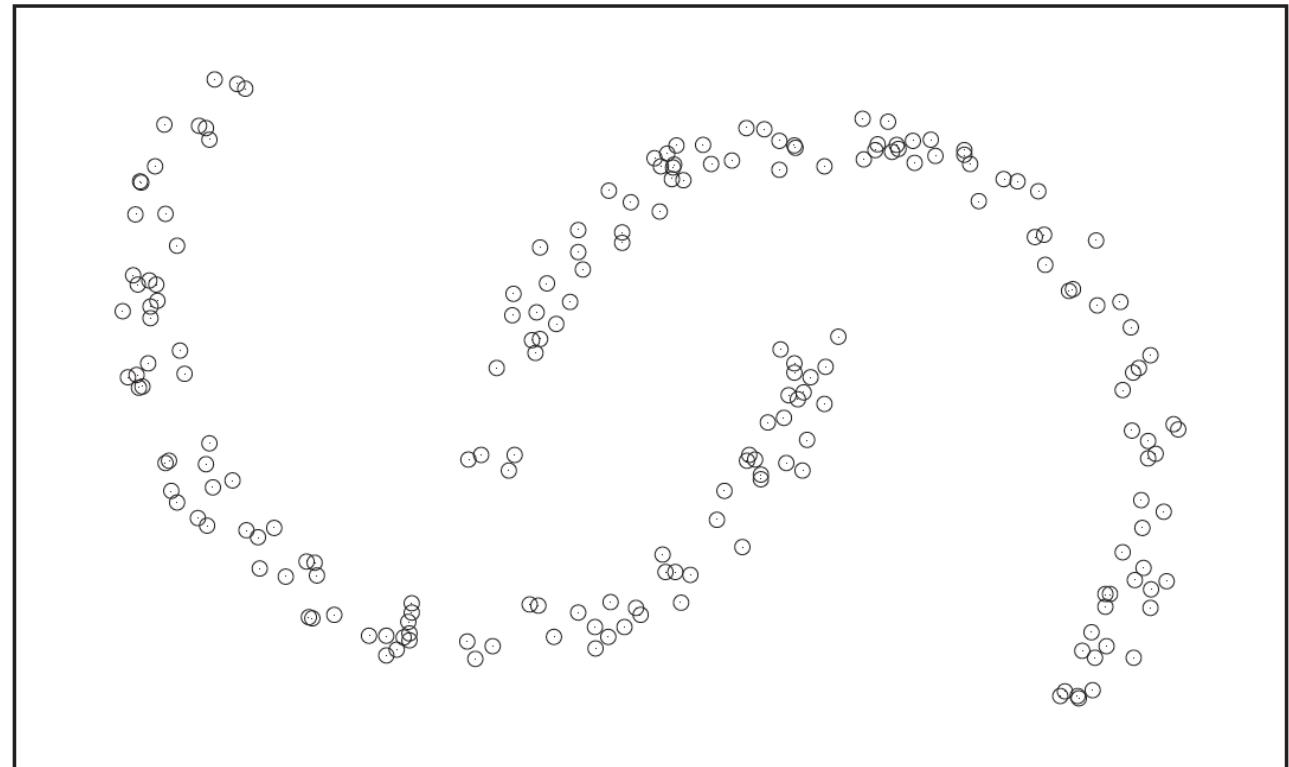
# Συσταδοποίηση Άλλων Τύπων Δεδομένων

- Έξι χρονοσειρές
- μπορούμε να διακρίνουμε χρονοσειρές που είναι περισσότερο όμοιες μεταξύ τους, παρά με άλλες



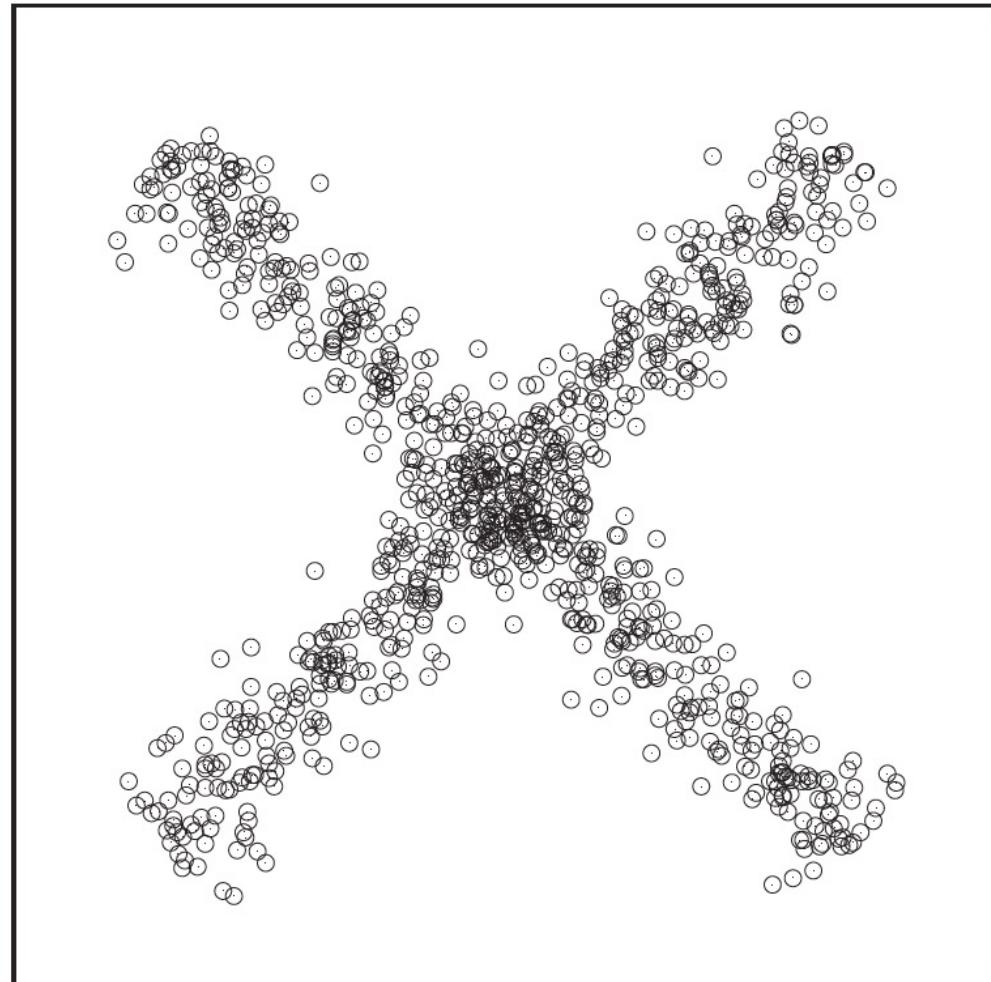
# Συστάδες με Παράξενα Σχήματα

- Καλά διαχωρισμένες συστάδες
- Όμως όχι σφαιροειδούς σχήματος
- Ορισμένοι αλγόριθμοι (π.χ. k-means) δεν μπορούν να αναγνωρίσουν τέτοιες συστάδες



# Μια Διαφορετική Οπτική

- Διαισθητικά, αναγνωρίζουμε δύο επικαλυπτόμενες συστάδες
- Όμως σχεδόν κανείς αλγόριθμος δεν μπορεί να τις διαχωρίσει



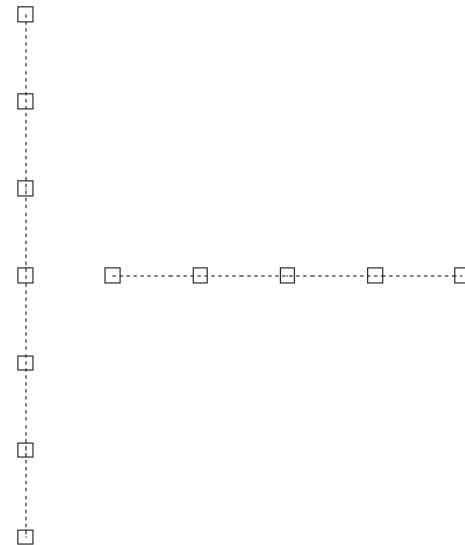
# Μια Διαφορετική Οπτική

Η απόσταση μεταξύ οποιοδήποτε δύο διαδοχικών σημείων είναι η ίδια, όμως αναγνωρίζουμε την «κάθετη» και την «οριζόντια» συστάδα

Αυτή η έννοια της συστάδας  
στηρίζεται στην ομοιότητα μεταξύ  
σημείου και **μιας ιδιότητας μιας  
συστάδας** (όχι ομοιότητα σημείου με  
σημείο)

Οδηγούμαστε σε ένα πρόβλημα:

- για να κάνουμε συσταδοποίηση,  
πρέπει να γνωρίζουμε τις ιδιότητες  
των συστάδων
- αλλά για να γνωρίζουμε αυτές τις  
ιδιότητες πρέπει να μπορούμε να  
αναθέτουμε σημεία σε συστάδες



# Μέτρα Απόστασης και Ομοιότητας

---



Υπενθύμιση:  
1η Διάλεξη

# Μέτρα Απόστασης και Ομοιότητας

- Για τη συσταδοποίηση
  - δεν είναι αναγκαίο να έχουμε σημεία σε κάποιο γεωμετρικό χώρο
- Αρκεί να υπάρχει ορισμός της **απόστασης** (ή ισοδύναμα της **ομοιότητας**) ανάμεσα σε **κάθε ζεύγος αντικειμένων**
  - Για σύνολο δεδομένων **n** αντικειμένων
  - Μήτρα (**nxn**) απόστασης ή ομοιότητας
- Έτσι, είναι δυνατό να εφαρμόσουμε συσταδοποίηση σε ένα σύνολο αλφαριθμητικών

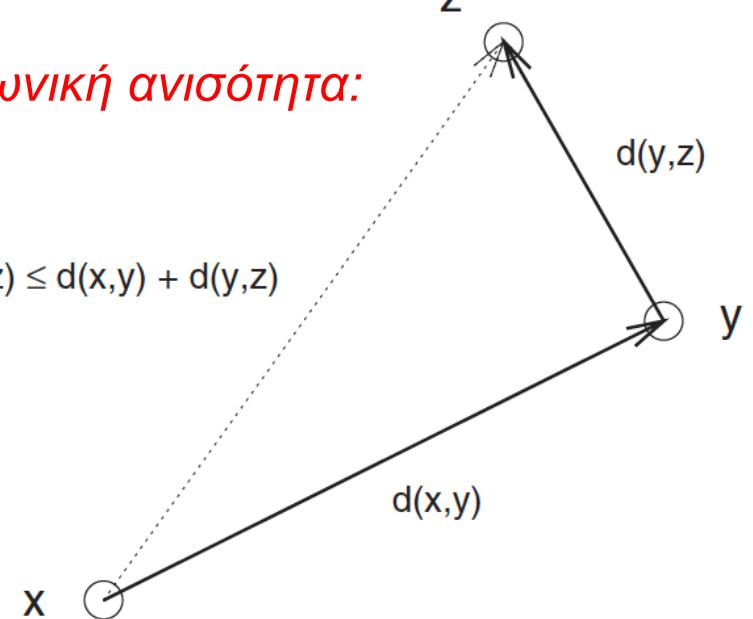
# Μέτρα Απόστασης και Ομοιότητας

- **Απόσταση  $d(x,y)$ :** οποιαδήποτε συνάρτηση που
  - παίρνει ως είσοδο δύο σημεία  $x$  και  $y$  και
  - επιστρέφει μια αριθμητική ποσότητα που δείχνει πόσο **διαφέρουν** αυτά
  - όσο μεγαλύτερη η διαφορά, τόσο μεγαλύτερη η απόσταση
- **Ομοιότητα  $s(x,y)$ :** οποιαδήποτε συνάρτηση που
  - παίρνει ως είσοδο δύο σημεία  $x$  και  $y$  και
  - επιστρέφει μια αριθμητική ποσότητα που δείχνει πόσο **μοιάζουν** αυτά
  - όσο πιο διαφορετικά είναι, τόσο χαμηλότερη η τιμή της ομοιότητας
- Οποιαδήποτε **απόσταση** μπορεί να μετατραπεί σε **ομοιότητα** και το αντίστροφο
- Για παράδειγμα, αν η ομοιότητα  $s$  παίρνει τιμές στο  $[0,1]$ , μπορούμε να ορίσουμε μια ισοδύναμη απόσταση ως  $d = 1 - s$
- Εναλλακτικά  $d = 1/s$  ή  $s = e^{-d}$

# Μετρική Απόσταση

*Η τριγωνική ανισότητα:*

$$d(x,z) \leq d(x,y) + d(y,z)$$



## ■ Ιδιότητες:

$$d(x, y) \geq 0$$

$$d(x, y) = 0 \quad \text{if and only if } x = y$$

$$d(x, y) = d(y, x)$$

$$d(x, y) + d(y, z) \geq d(x, z)$$

- Παράδειγμα μη συμμετρικής απόστασης:
  - Σε ένα σύνολο ανθρώπων, πόσο συμπαθεί ο ένας των άλλον
- Είναι χρήσιμο μια απόσταση να είναι συμμετρική
- Μπορούμε πάντα να κατασκευάσουμε μια συμμετρική απόσταση από μια μη συμμετρική:

$$d_S(x, y) = \frac{d(x, y) + d(y, x)}{2}$$

# Κοινά Μέτρα Απόστασης

Υπενθύμιση:  
1η Διάλεξη

Name	Definition
Manhattan	$d(x, y) = \sum_i^d  x_i - y_i $
Euclidean	$d(x, y) = \sqrt{\sum_i^d (x_i - y_i)^2}$
Maximum	$d(x, y) = \max_i  x_i - y_i $
Minkowski	$d(x, y) = \left( \sum_i^d  x_i - y_i ^p \right)^{1/p}$
Dot product	$x \cdot y = \frac{\sum_i^d x_i y_i}{\sqrt{\sum_i^d x_i^2} \sqrt{\sum_i^d y_i^2}}$
Correlation coefficient	$\text{corr}(x, y) = \frac{\sum_i^d (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^d (x_i - \bar{x})^2} \sqrt{\sum_i^d (y_i - \bar{y})^2}}$
	$\bar{x} = \frac{1}{d} \sum_i^d x_i \quad \bar{y} = \frac{1}{d} \sum_i^d y_i$

# Μέθοδοι Συσταδοποίησης

# Τύποι Αλγορίθμων Συσταδοποίησης

- Αλγόριθμοι που αναζητούν κέντρα
  - K-means
- Αλγόριθμοι κατασκευής δέντρων
  - Hierarchical agglomerative clustering
- Αλγόριθμοι μεγέθυνσης γειτονιών
  - DBSCAN

# 1. Αλγόριθμοι που Αναζητούν Κέντρα k-means

- Ο **k-means**<sup>1</sup> είναι ένας από τους δημοφιλέστερους αλγόριθμους συσταδοποίησης
  - Απαιτεί τον **αριθμό** των αναμενόμενων **συστάδων** ως είσοδο
  - Είναι **επαναληπτικός** (**iterative**) αλγόριθμος
- Τρόπος λειτουργίας
  - Υπολογίζει τη θέση του **κέντρου** (**centroid**) κάθε συστάδας από τις θέσεις των σημείων που ανήκουν στη συστάδα
  - Στη συνέχεια, **αναθέτει** κάθε σημείο στο **κοντινότερο κέντρο**
  - Η διαδικασία **επαναλαμβάνεται** μέχρι να επιτευχθεί κάποιο **κριτήριο σύγκλισης**

<sup>1</sup>Stuart P. Lloyd: Least squares quantization in PCM. *IEEE Trans. Inf. Theory* 28(2): 129-136 (1982).  
Algorithm first proposed in Bell Telephone Laboratories Paper (1957).

# k-means – Ψευδοιώδης

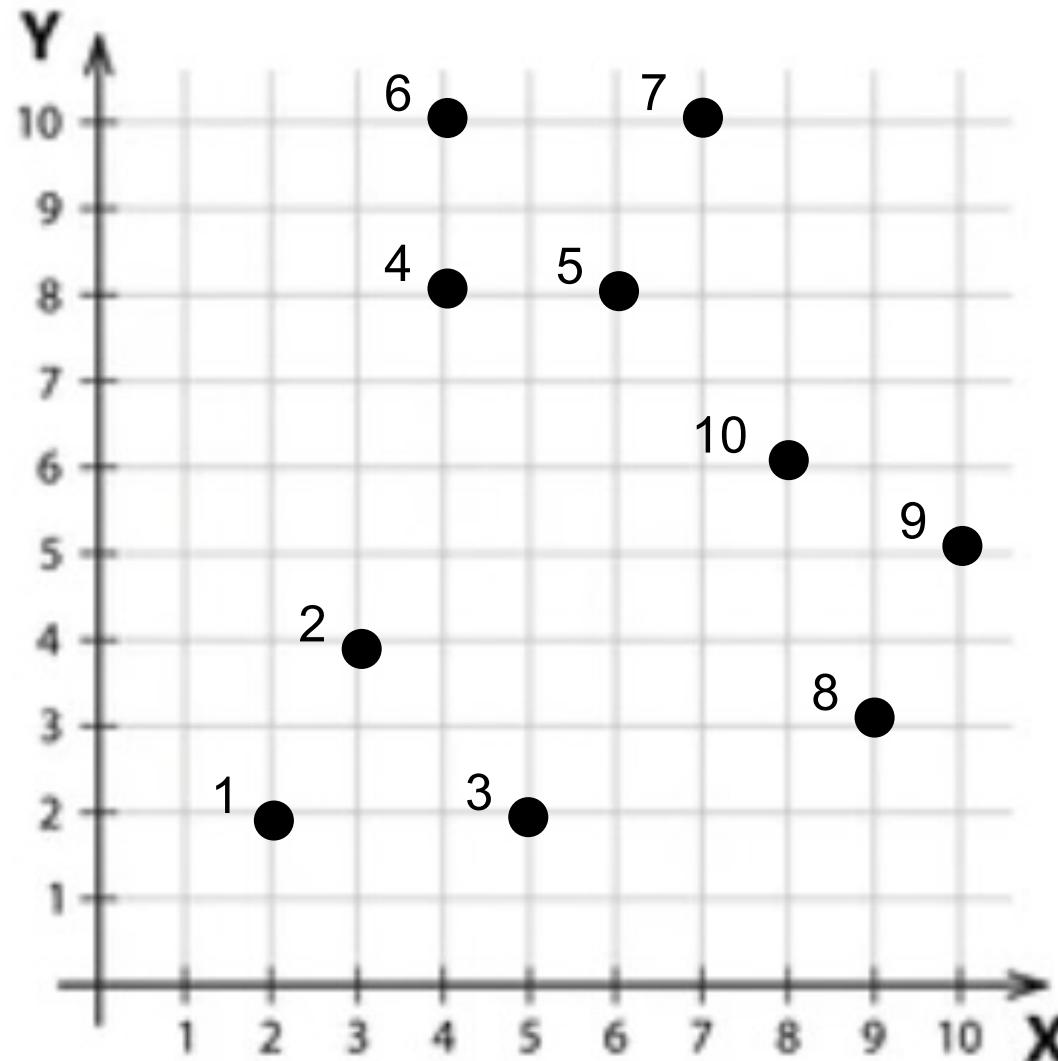
```
choose initial positions for the cluster centroids  
  
repeat:  
    for each point:  
        calculate its distance from each cluster centroid  
        assign the point to the nearest cluster  
  
    recalculate the positions of the cluster centroids
```

Τα νέα κέντρα υπολογίζονται χρησιμοποιώντας το κέντρο μάζας μιας συστάδας:

$$x_c = \frac{1}{n} \sum_i^n x_i \quad y_c = \frac{1}{n} \sum_i^n y_i$$

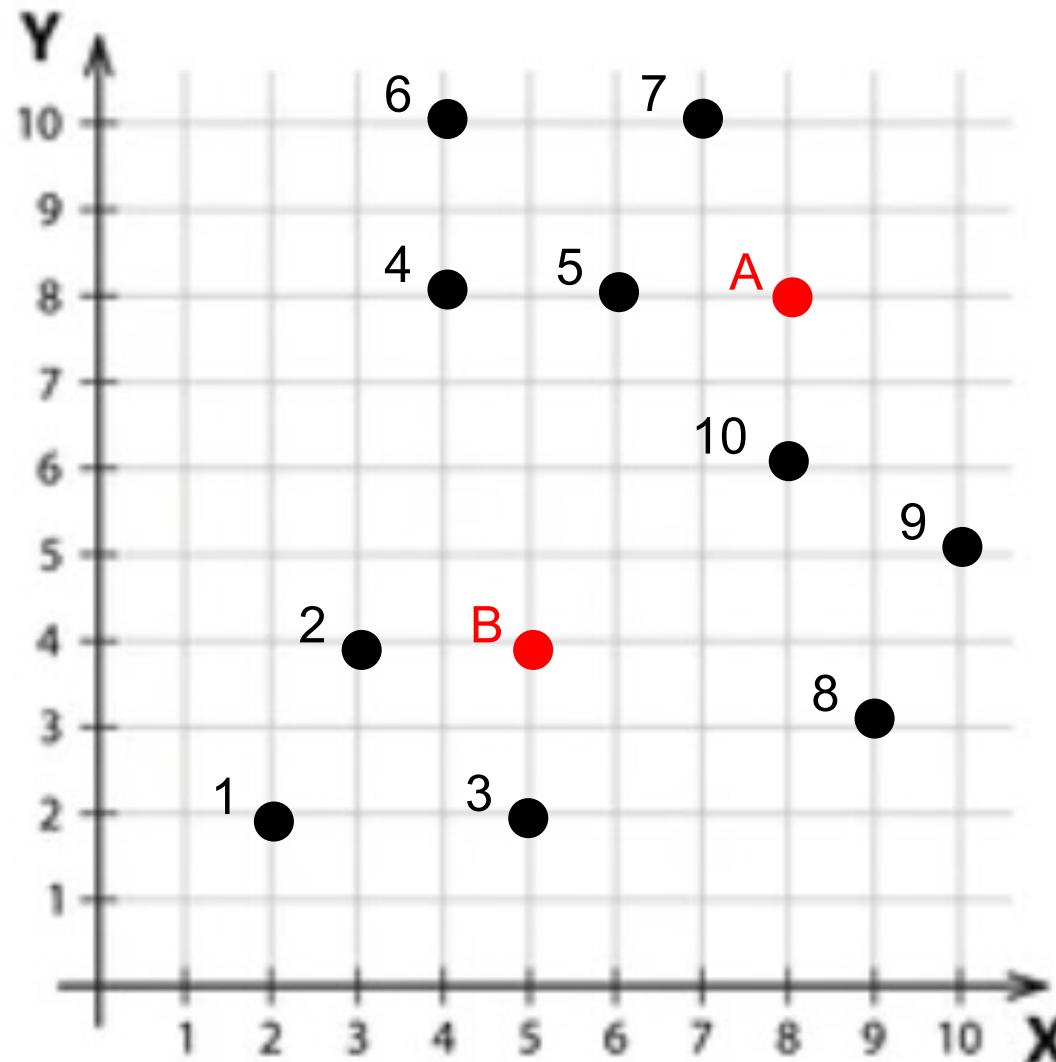
# Παράδειγμα: Σύνολο 10 Σημείων

	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6



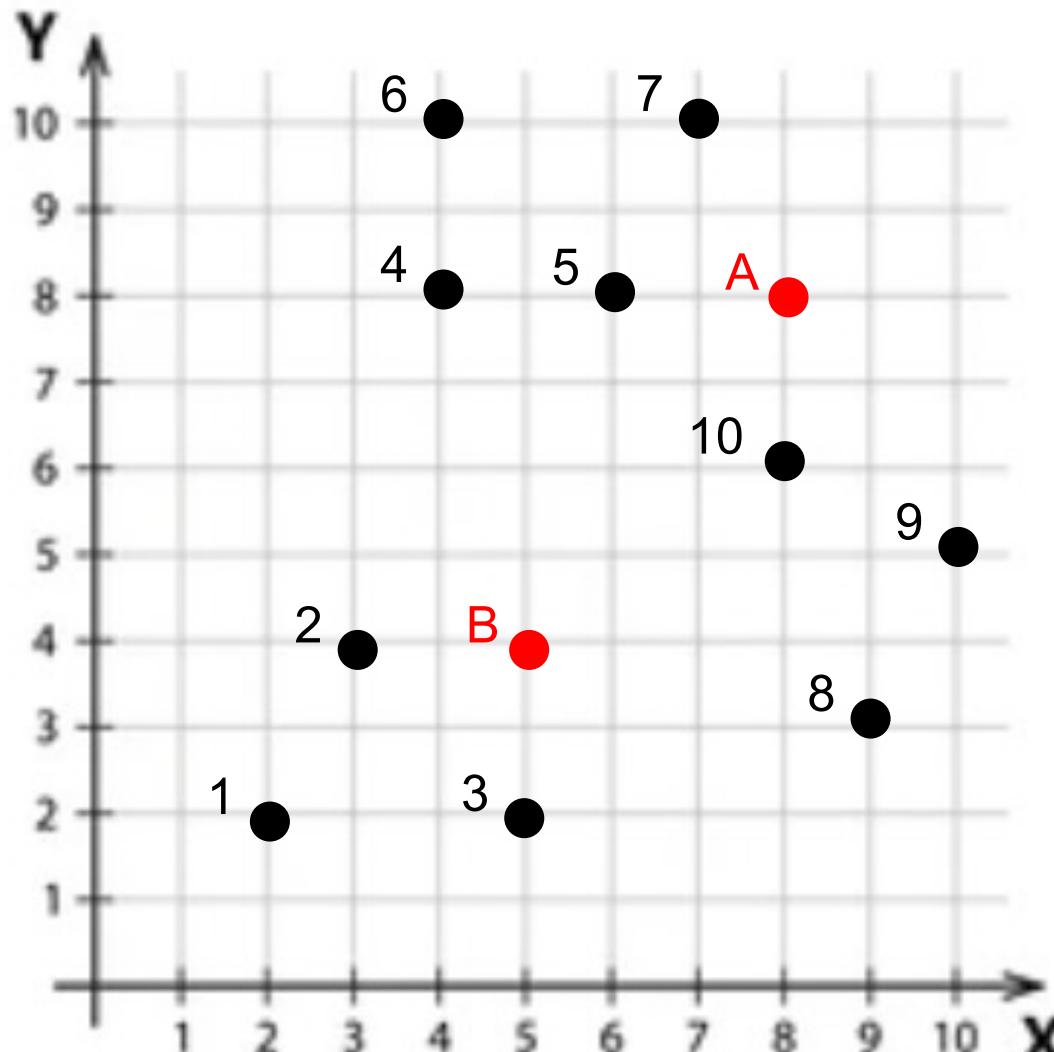
# Παράδειγμα: $k=2$ Τυχαία Κέντρα

	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6



# Παράδειγμα: Υπολογισμός Αποστάσεων

	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6

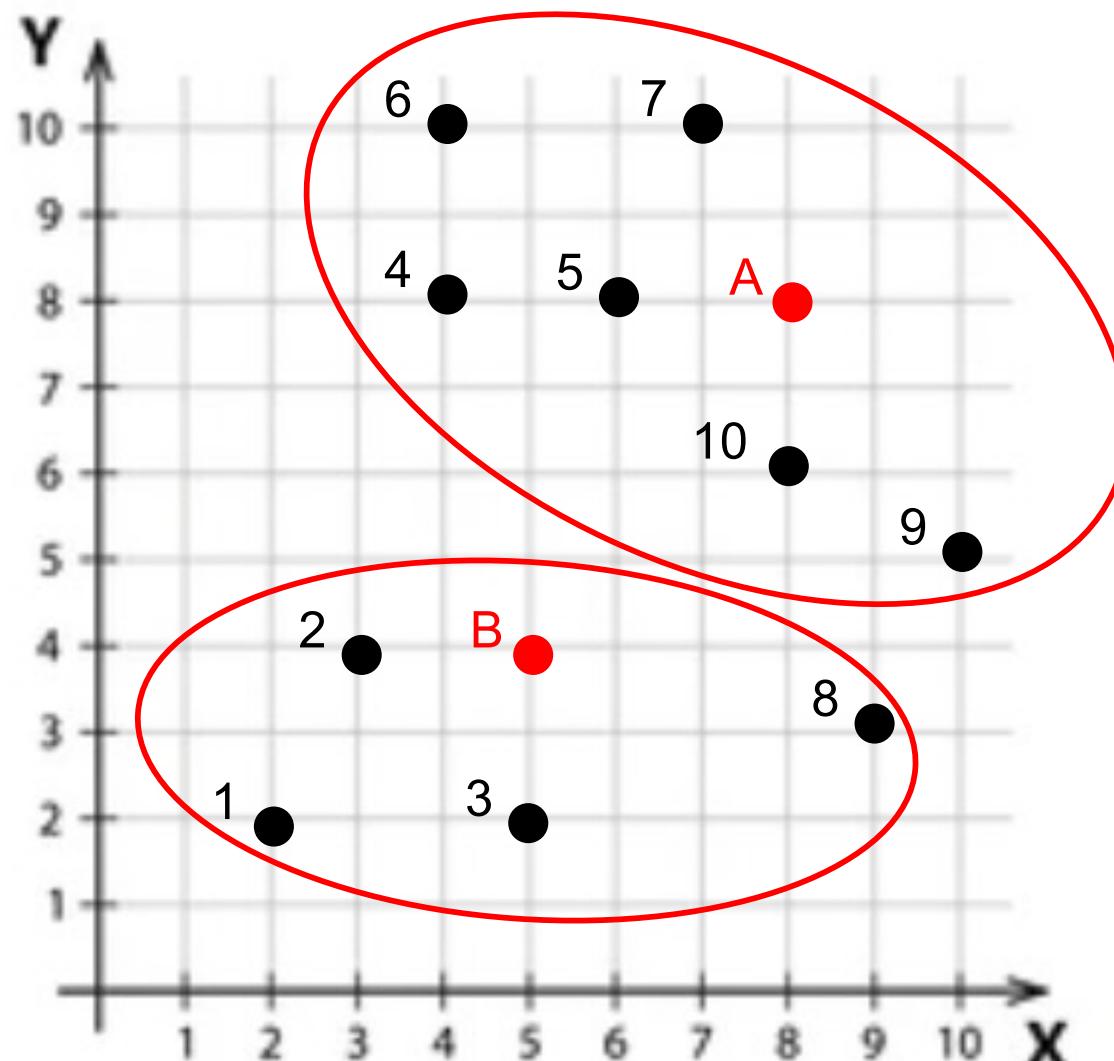


$$\begin{aligned}
 d(A,1) &= 6\sqrt{2} \\
 d(A,2) &= \sqrt{41} \\
 d(A,3) &= 3\sqrt{5} \\
 d(A,4) &= 4 \\
 d(A,5) &= 2 \\
 d(A,6) &= 2\sqrt{5} \\
 d(A,7) &= \sqrt{5} \\
 d(A,8) &= \sqrt{26} \\
 d(A,9) &= \sqrt{13} \\
 d(A,10) &= 2 \\
 d(B,1) &= \sqrt{13} \\
 d(B,2) &= 2 \\
 d(B,3) &= 2 \\
 d(B,4) &= \sqrt{17} \\
 d(B,5) &= \sqrt{17} \\
 d(B,6) &= \sqrt{37} \\
 d(B,7) &= 2\sqrt{10} \\
 d(B,8) &= \sqrt{17} \\
 d(B,9) &= \sqrt{26} \\
 d(B,10) &= \sqrt{13}
 \end{aligned}$$

# Παράδειγμα: Ανάθεση σε Πλησιέστερα

## Κέντρα

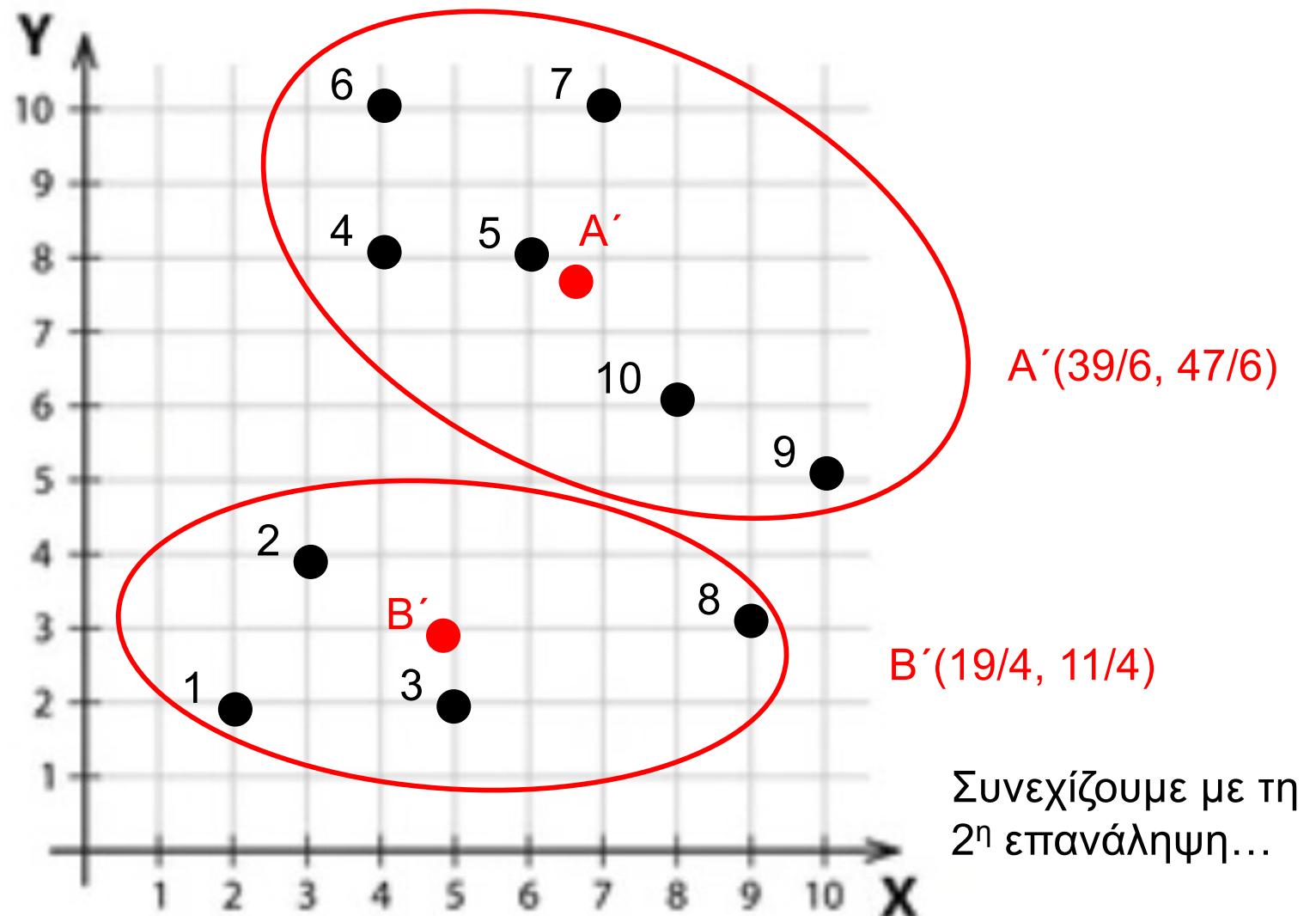
	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6



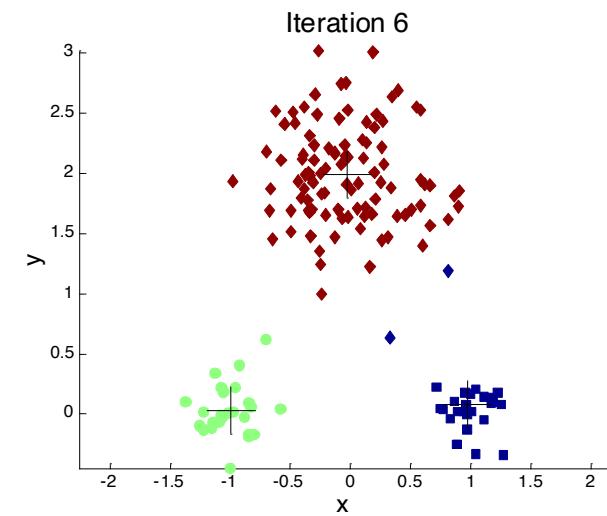
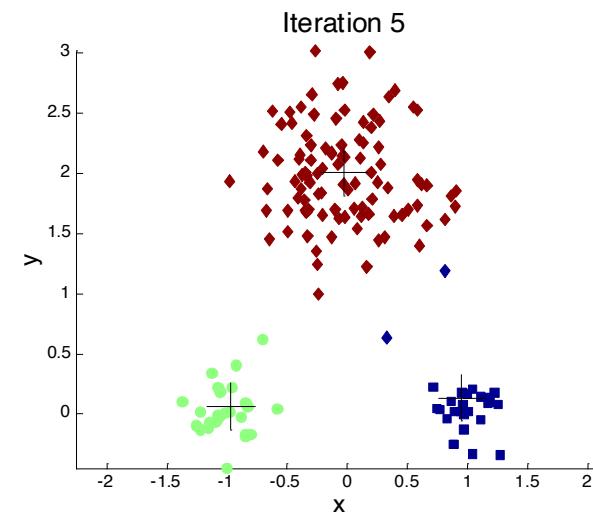
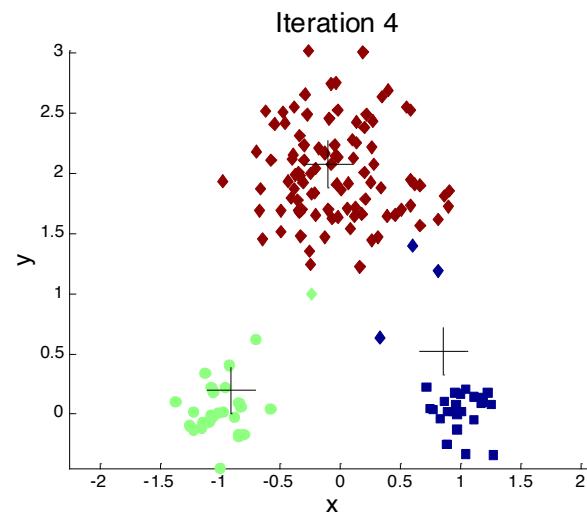
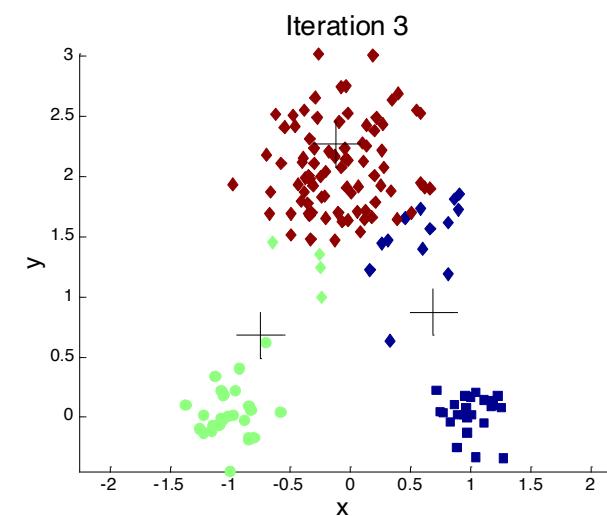
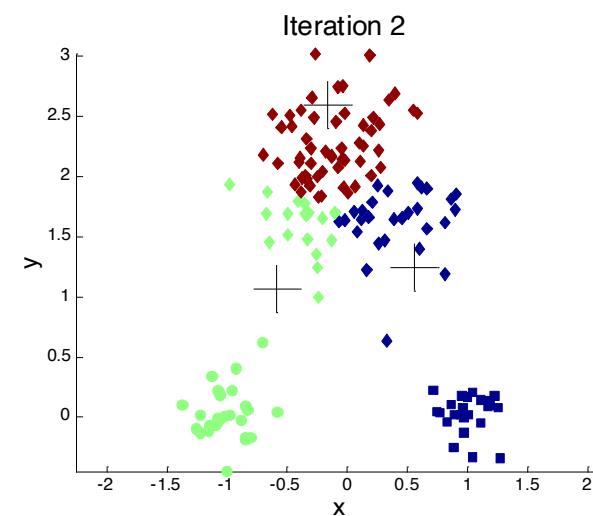
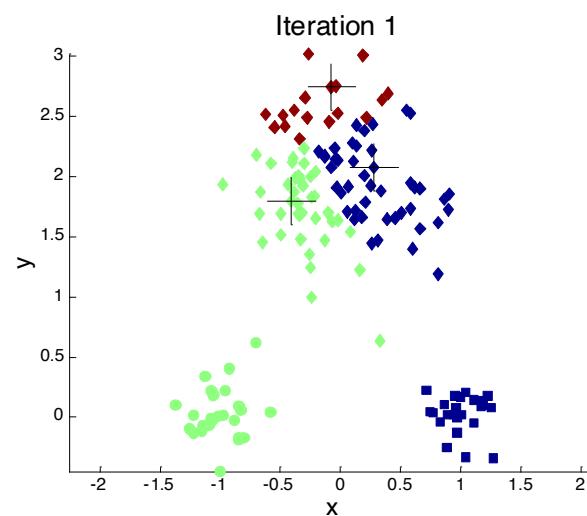
$$\begin{aligned}
 d(A,1) &= 6\sqrt{2} \\
 d(A,2) &= \sqrt{41} \\
 d(A,3) &= 3\sqrt{5} \\
 d(A,4) &= 4 \\
 d(A,5) &= 2 \\
 d(A,6) &= 2\sqrt{5} \\
 d(A,7) &= \sqrt{5} \\
 d(A,8) &= \sqrt{26} \\
 d(A,9) &= \sqrt{13} \\
 d(A,10) &= 2 \\
 d(B,1) &= \sqrt{13} \\
 d(B,2) &= 2 \\
 d(B,3) &= 2 \\
 d(B,4) &= \sqrt{17} \\
 d(B,5) &= \sqrt{17} \\
 d(B,6) &= \sqrt{37} \\
 d(B,7) &= 2\sqrt{10} \\
 d(B,8) &= \sqrt{17} \\
 d(B,9) &= \sqrt{26} \\
 d(B,10) &= \sqrt{13}
 \end{aligned}$$

# Παράδειγμα: Υπολογισμός Νέων Κέντρων

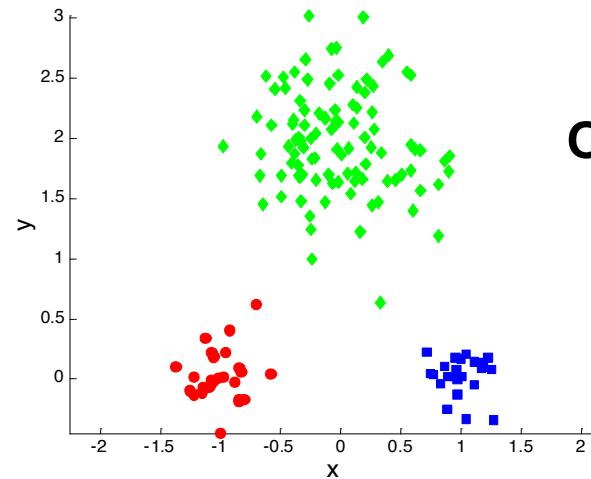
	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6



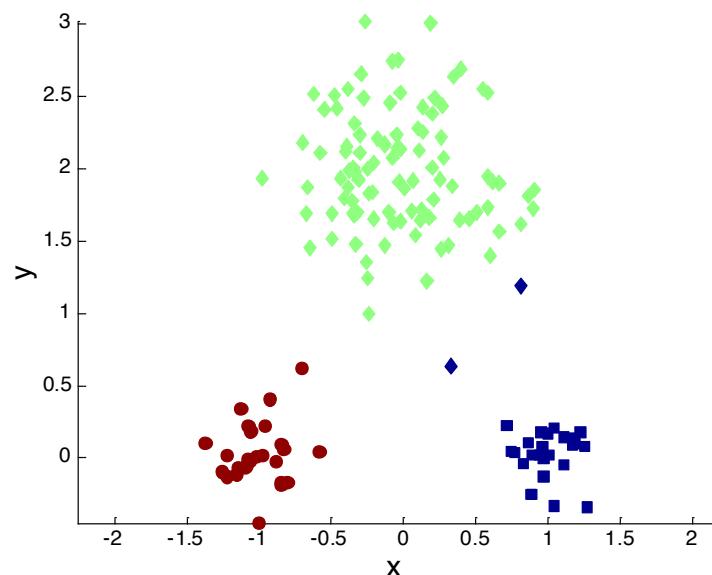
# k-means – Παράδειγμα



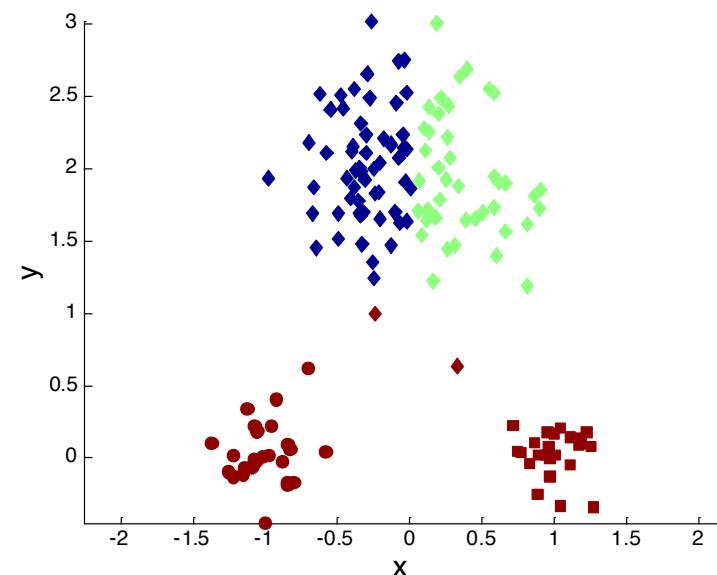
# Δύο Διαφορετικά Αποτελέσματα



Original Points

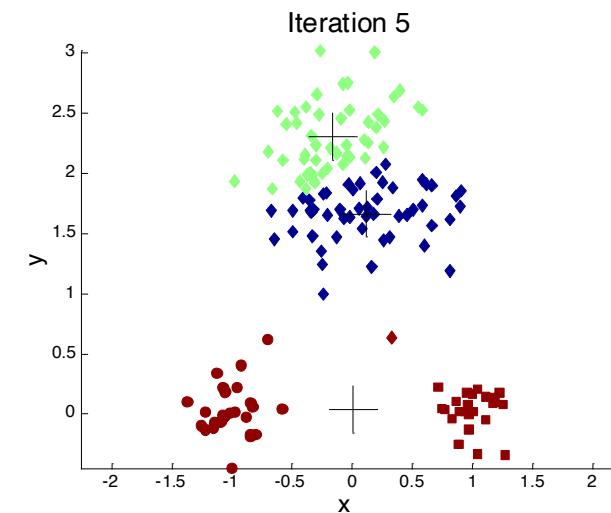
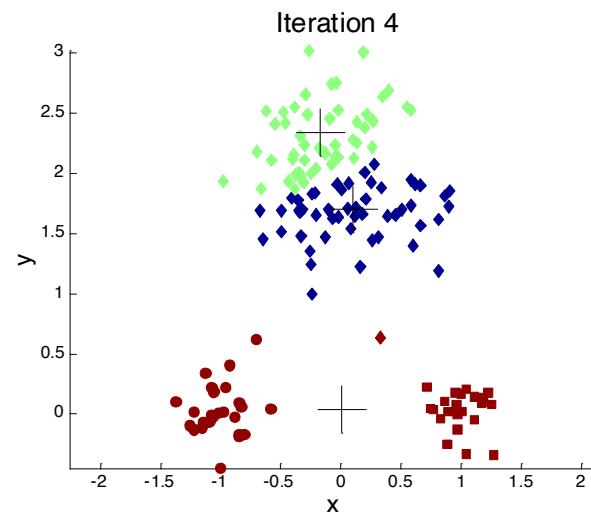
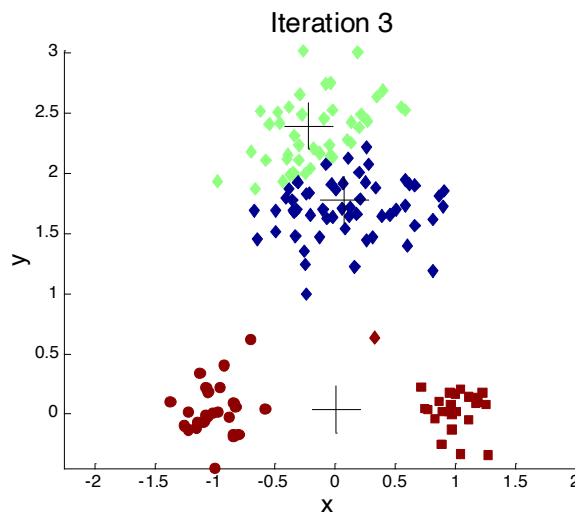
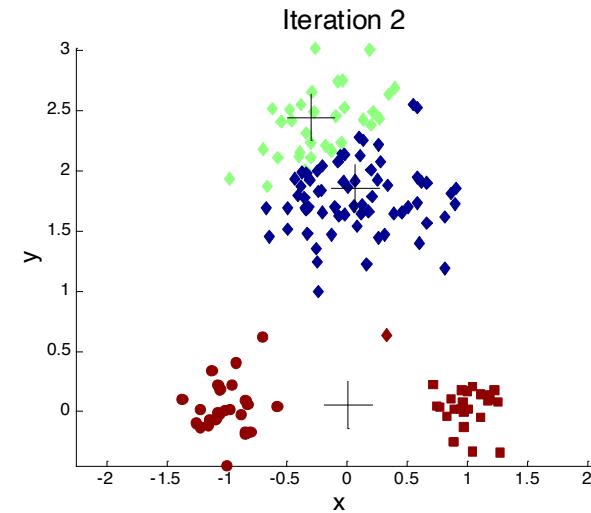
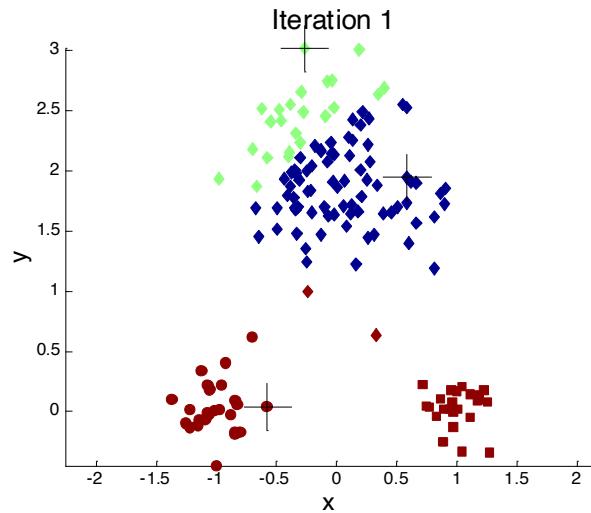


Optimal Clustering



Sub-optimal Clustering

# Επιλογή Αρχικών Κέντρων

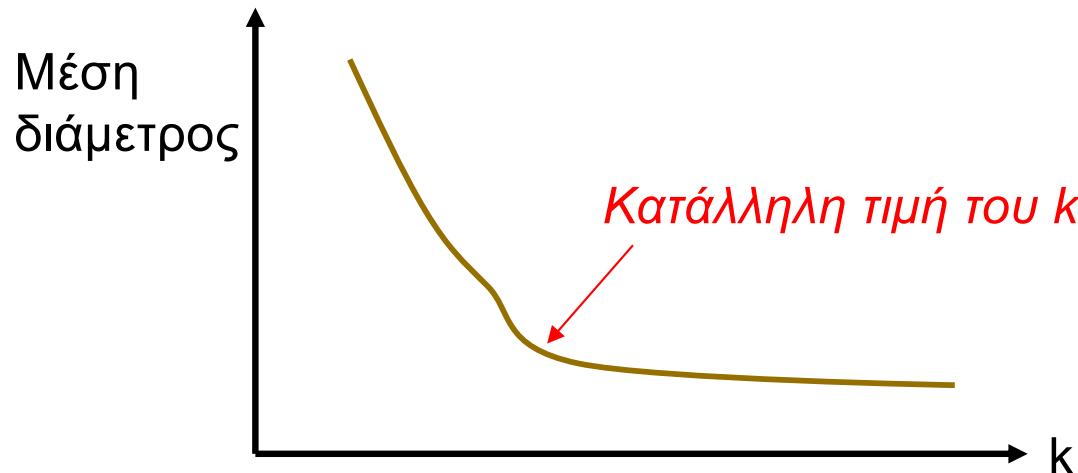


# Τρόποι Επιλογής Αρχικών Κέντρων

- **Τυχαία αρχικοποίηση**
  - Εκτελούμε τον αλγόριθμο για πολλές τυχαίες αρχικοποιήσεις
  - Μεγάλη πιθανότητα να μην επιτευχθεί η βέλτιστη συσταδοποίηση
- **Άλλες στρατηγικές επιλογής**
  - Αλγόριθμος **k-means++**
    - Το πρώτο κέντρο επιλέγεται τυχαία
    - Το κάθε επόμενο κέντρο επιλέγεται με πιθανότητα ανάλογη του τετραγώνου της απόστασής του από το πλησιέστερο κέντρο (μεταξύ των υπαρχόντων)
- Διχοτομικός αλγόριθμος k-means (**bisecting k-means**)
  - Ιεραρχική προσέγγιση, top-down
  - Διαχωρίζεται το σύνολο δεδομένων σε 2 συστάδες
  - Επιλέγεται μία από αυτές για διαχωρισμό, κ.ο.κ.

<sup>1</sup>David Arthur, Sergei Vassilvitskii: k-means++: the advantages of careful seeding. SODA 2007: 1027-1035

# Επιλογή της «Κατάλληλης» Τιμής του $k$



- Μια τεχνική γνωστή ως: “*Elbow rule*”
- Εκτελούμε συσταδοποίηση για διάφορες τιμές του  $k$
- Υπολογίζουμε κάποιο μέτρο διασποράς, όπως τη μέση διάμετρο
  - Μόλις το  $k$  πέσει κάτω από τον πραγματικό αριθμό των συστάδων στα δεδομένα, το μέτρο διασποράς αυξάνεται απότομα

# k-means – Χαρακτηριστικά

- Ο k-means είναι **μη ντετερμινιστικός** αλγόριθμος
  - *Διαφορετική επιλογή των αρχικών τιμών οδηγεί σε διαφορετική ανάθεση σημείων σε κέντρα*
  - Για αυτό συνήθως εκτελούμε τον αλγόριθμο αρκετές φορές και συγκρίνουμε τα αποτελέσματα
- Όμως εάν γνωρίζουμε περίπου τη θέση των κέντρων των συστάδων
  - Μπορούμε να αρχικοποιήσουμε τον αλγόριθμο με αυτά
- Ο αλγόριθμος είναι αποδοτικός
  - Διότι **δε χρειάζεται να ψάχνει στα δεδομένα για να ανακαλύψει «καλά» κέντρα**
  - Αντίθετα, απλά υπολογίζει το κέντρο μάζας μιας συστάδας

# k-means – Χαρακτηριστικά

- Για **κατηγορικά δεδομένα** → **k-medoids**
  - Αντί για υπολογισμό του κέντρου μιας συστάδας
  - Εύρεση του **σημείου μιας συστάδας** (που ονομάζεται **medoid**)
  - Που έχει την ελάχιστη μέση απόσταση από όλα τα σημεία της συστάδας
- Ο **k-means** είναι αποδοτικός αλγόριθμος
  - Κάθε επανάληψη έχει υπολογιστική πολυπλοκότητα: **O(k n)**
    - καθώς υπολογίζεται η απόσταση μεταξύ κάθε σημείου και κέντρου
  - Συνήθως 10-50 επαναλήψεις αρκούν για να οδηγήσουν σε σύγκλιση
- Αντίθετα, ο **k-medoids** απαιτεί για κάθε επανάληψη: **O(n<sup>2</sup>)**

# Πότε Μπορεί να Αποτύχει ο k-means

- Συνήθως ο k-means λειτουργεί καλά και γρήγορα
- Όμως, μπορεί να αποτύχει ακόμη και σε περιπτώσεις που υπάρχει ισχυρή δομή συσταδοποίησης
- Λόγω της επαναληπτικής φύσης του, λειτουργεί καλά σε περιπτώσεις που η **πυκνότητα των δεδομένων αλλάζει σταδιακά**
- Αντίθετα, αν το σύνολο δεδομένων περιέχει πολύ πυκνές και καλά διαχωρισμένες συστάδες, ο k-means μπορεί να «κολλήσει» αν αρχικά δύο κέντρα ανατεθούν στην ίδια συστάδα
  - Η μετακίνηση του ενός κέντρου σε άλλη συστάδα θα απαιτούσε μια μεγάλη κίνηση, που δεν είναι πιθανή λόγω των τοπικών κυρίως μεταβολών (βημάτων) του k-means

# Ο k-means Περιληπτικά

- Ο k-means και οι παραλλαγές του **λειτουργούν καλά** για **σφαιροειδείς συστάδες**
- Μπορεί να αποτύχει σε περίπτωση:
  - Πολύπλοκων γεωμετρικών σχημάτων και εμφωλευμένων συστάδων
  - Συστάδων πολύ διαφορετικού μεγέθους ή διαφορετικής πυκνότητας
- Ο αναμενόμενος αριθμός συστάδων ( $k$ ) απαιτείται ως είσοδος
- Ο αλγόριθμος είναι **επαναληπτικός** και **μη ντετερμινιστικός**
- Απαιτεί διανυσματικά δεδομένα
  - Ενώ για κατηγορικά δεδομένα χρησιμοποιείται ο k-medoids
- Είναι γραμμικής πολυπλοκότητας ως προς το σύνολο σημείων
- Ενώ ο k-medoids είναι τετραγωνικής πολυπλοκότητας

# Τύποι Αλγορίθμων Συσταδοποίησης

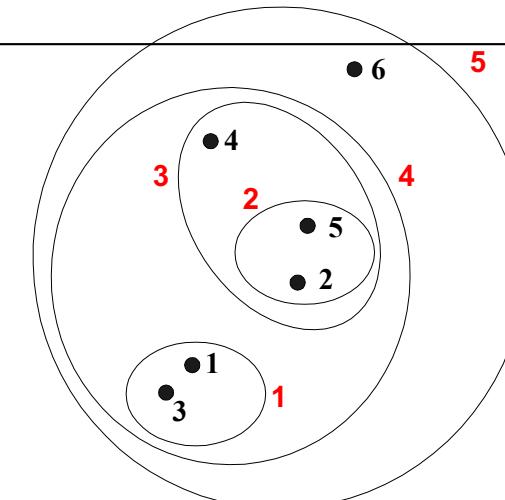
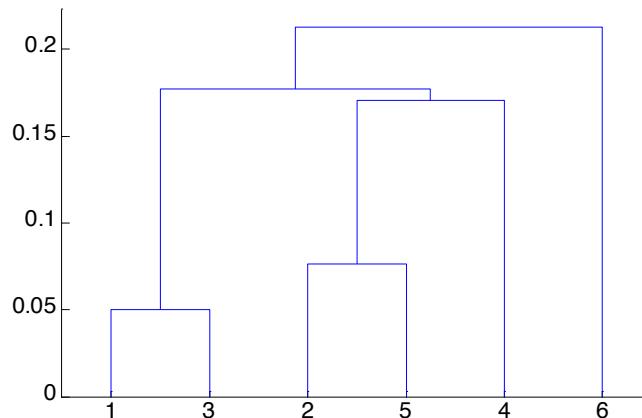
- Αλγόριθμοι που αναζητούν κέντρα
  - K-means
- Αλγόριθμοι κατασκευής δέντρων
  - Hierarchical agglomerative clustering
- Αλγόριθμοι μεγέθυνσης γειτονιών
  - DBSCAN

## 2. Αλγόριθμοι Κατασκευής Δέντρων

# Agglomerative Hierarchical Clustering

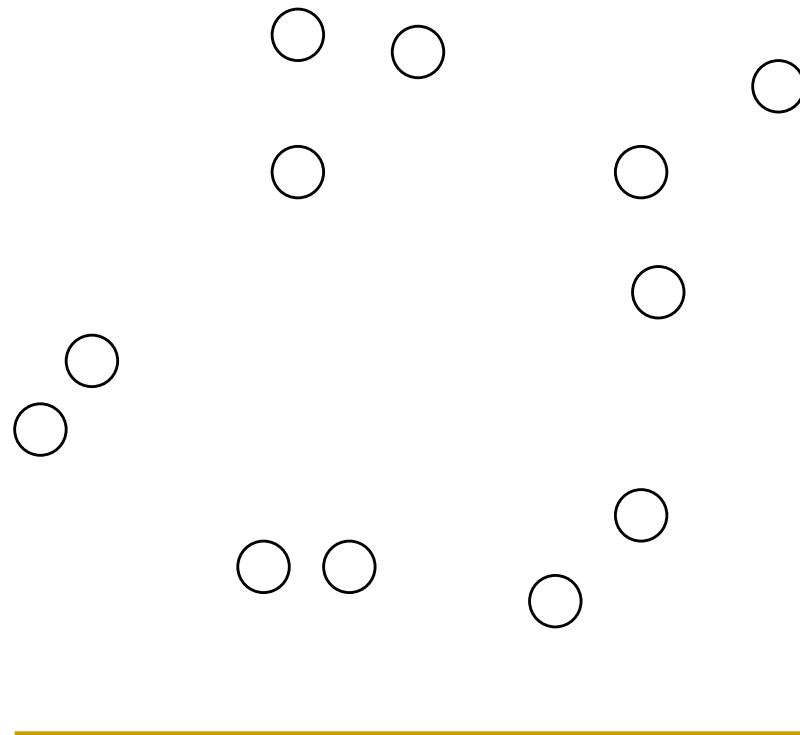
- Ένας εναλλακτικός τρόπος σχηματισμού συστάδων είναι η διαδοχική συνένωση συστάδων που βρίσκονται σε κοντινή απόσταση σε μεγαλύτερες συστάδες, μέχρι να απομείνει μια μόνο συστάδα
- Καταλήγει σε μια δενδροειδή ιεραρχία από συστάδες

1. Examine all pairs of clusters.
2. Combine the two clusters that are closest to each other into a single cluster.
3. Repeat.



# Παράδειγμα – Εκπίνηση

Ξεκινούμε με clusters που αποτελούνται από ένα σημείο το καθένα, και ένα πίνακα εγγύτητας (ή πίνακα αποστάσεων)

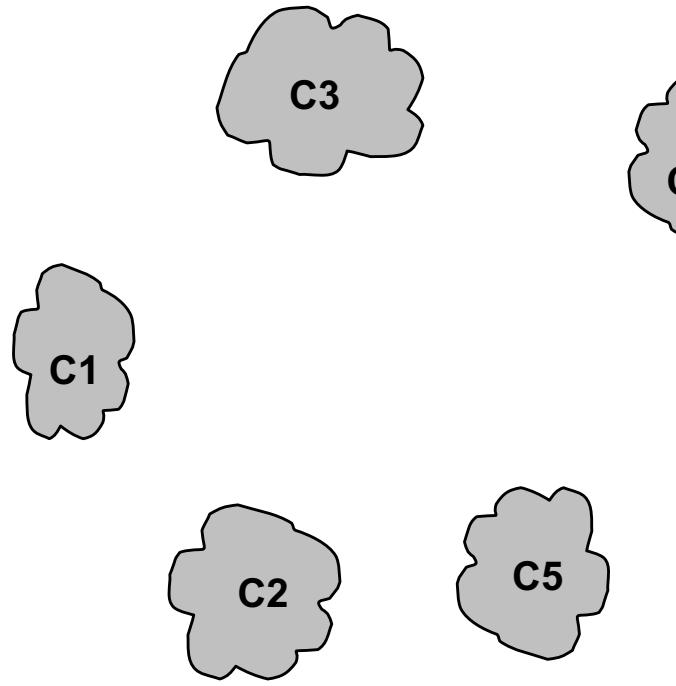


	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						

**Proximity Matrix**

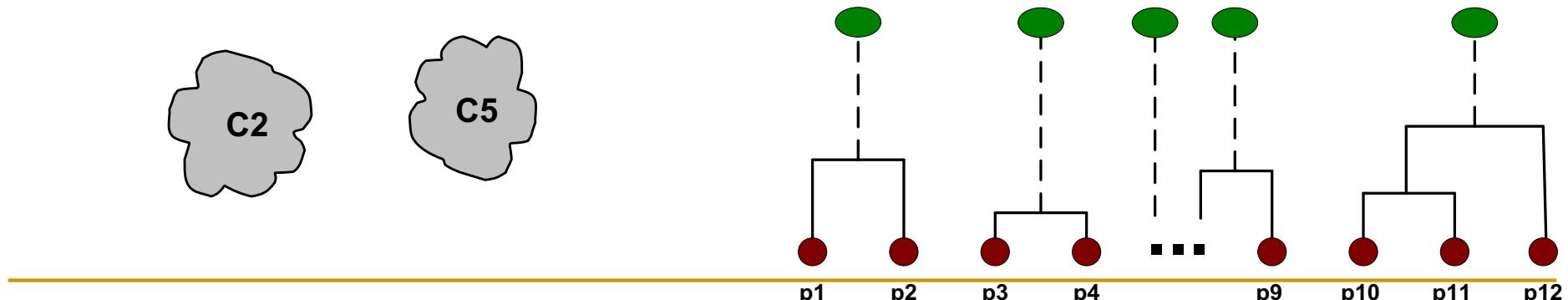
# Παράδειγμα – Ενδιάμεση Κατάσταση

Μετά από μερικά βήματα έχουμε βρει ορισμένα clusters



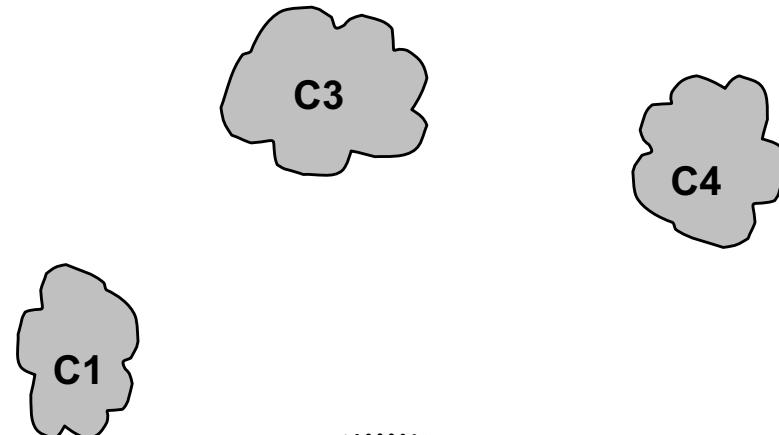
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



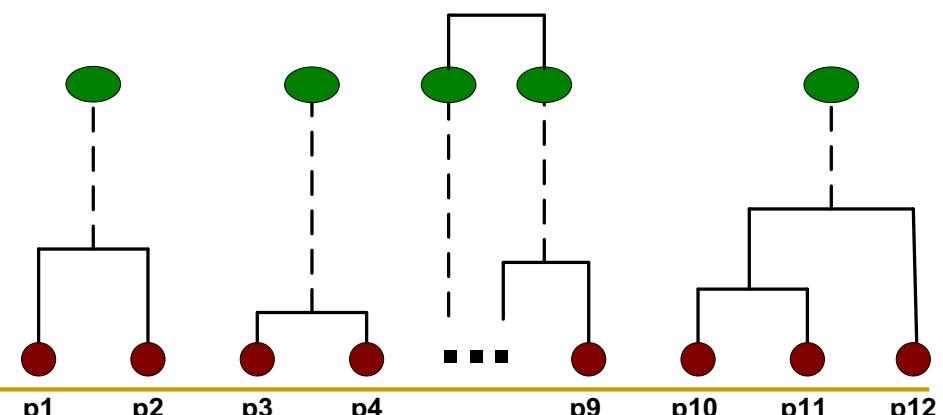
# Παράδειγμα – Ενδιάμεση Κατάσταση

Θέλουμε να συγχωνεύσουμε τα δυο κοντινότερα clusters και να ενημερώσουμε τον πίνακα



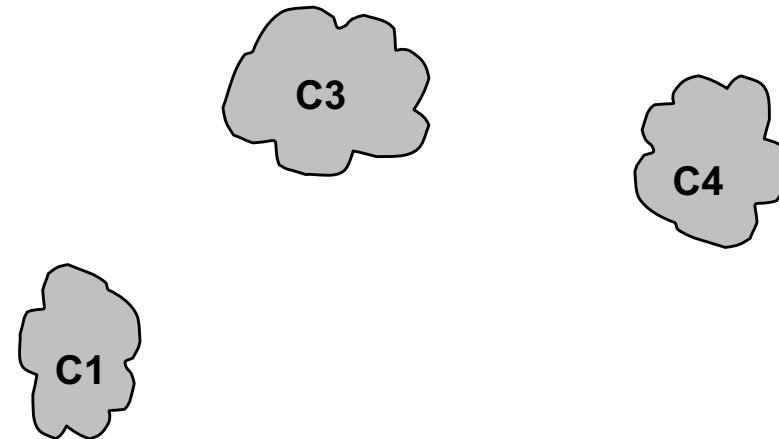
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



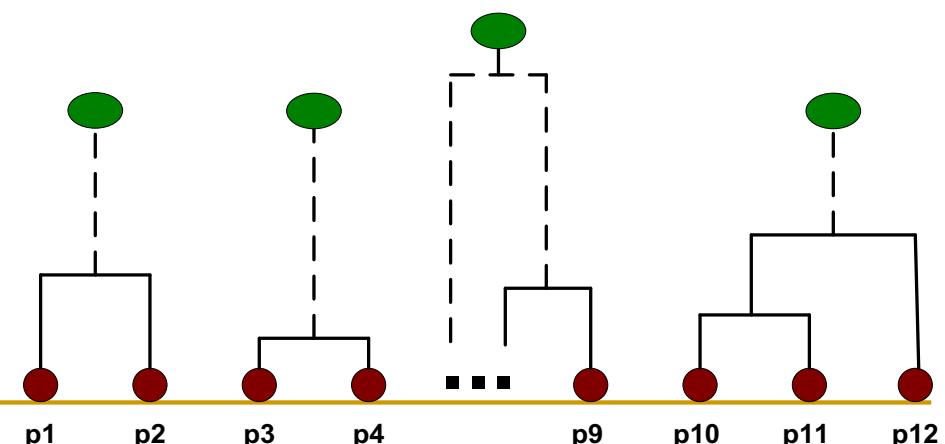
# Παράδειγμα – Μετά τη Συγχώνευση

Πώς ενημερώνουμε τον πίνακα;



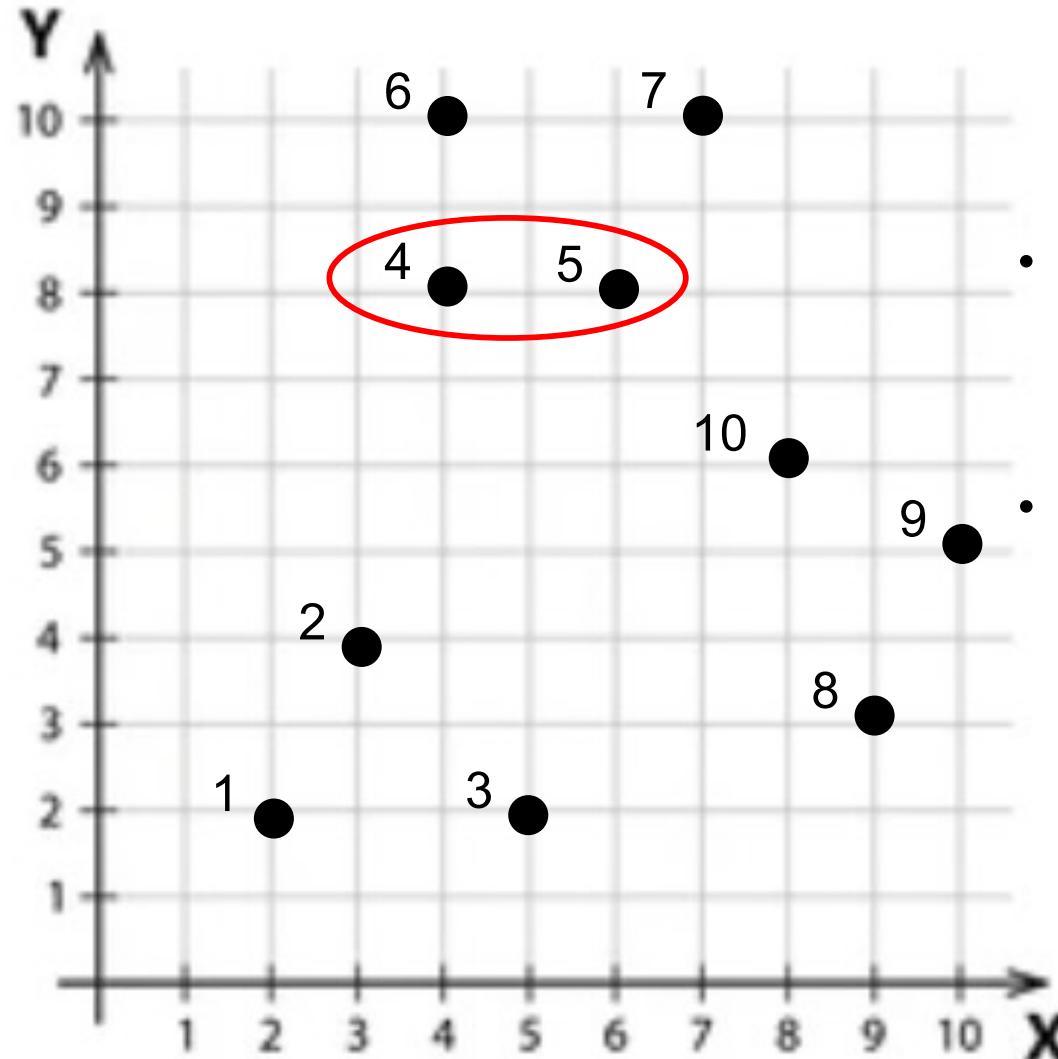
	C1		C3	C4
C1		?		
C2 ∪ C5	?	?	?	?
C3		?		
C4		?		

**Proximity Matrix**



# Παράδειγμα: Σύνολο 10 Σημείων

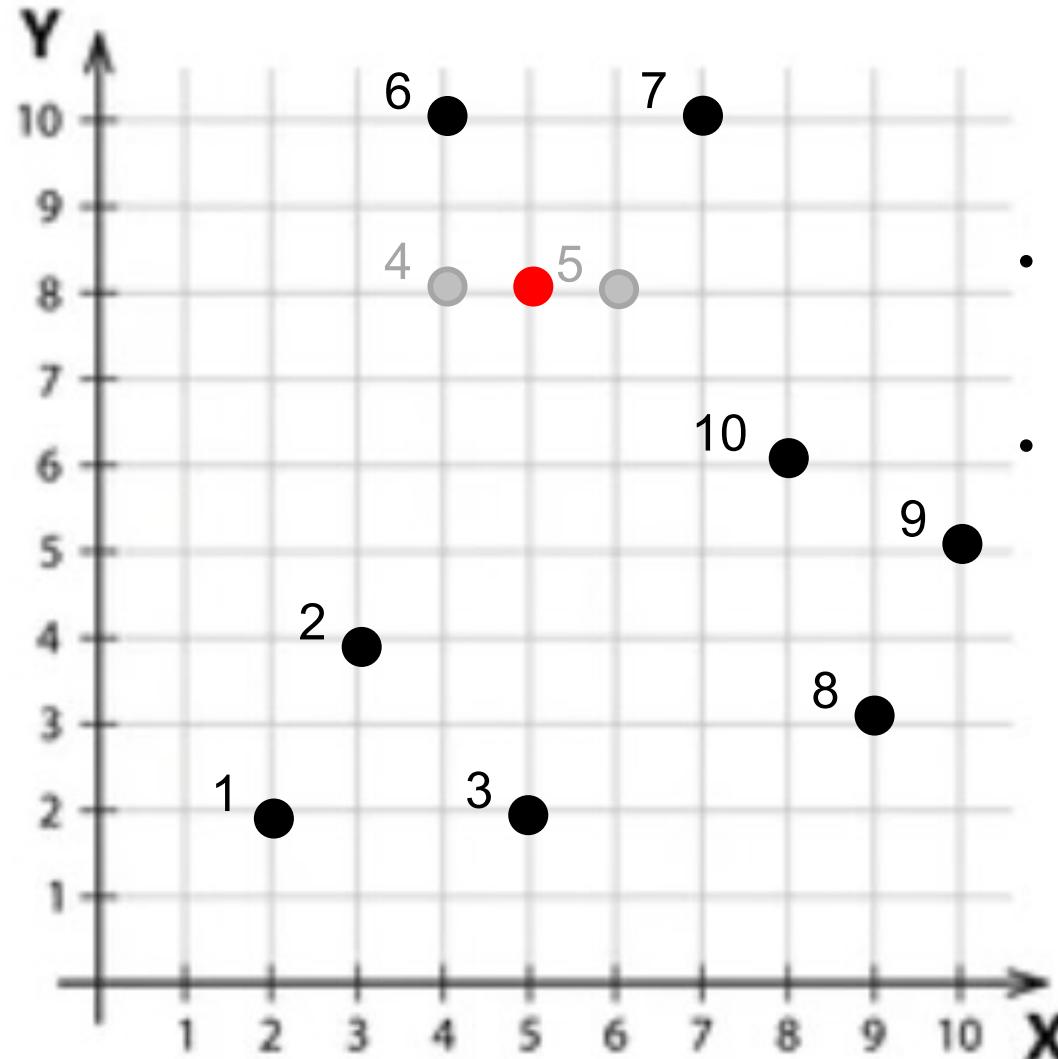
	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6



- Υπολογισμός όλων των αποστάσεων (πόσες;)
- Εύρεση κοντινότερου ζεύγους σημείων

# Παράδειγμα: Σύνολο 10 Σημείων

	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6

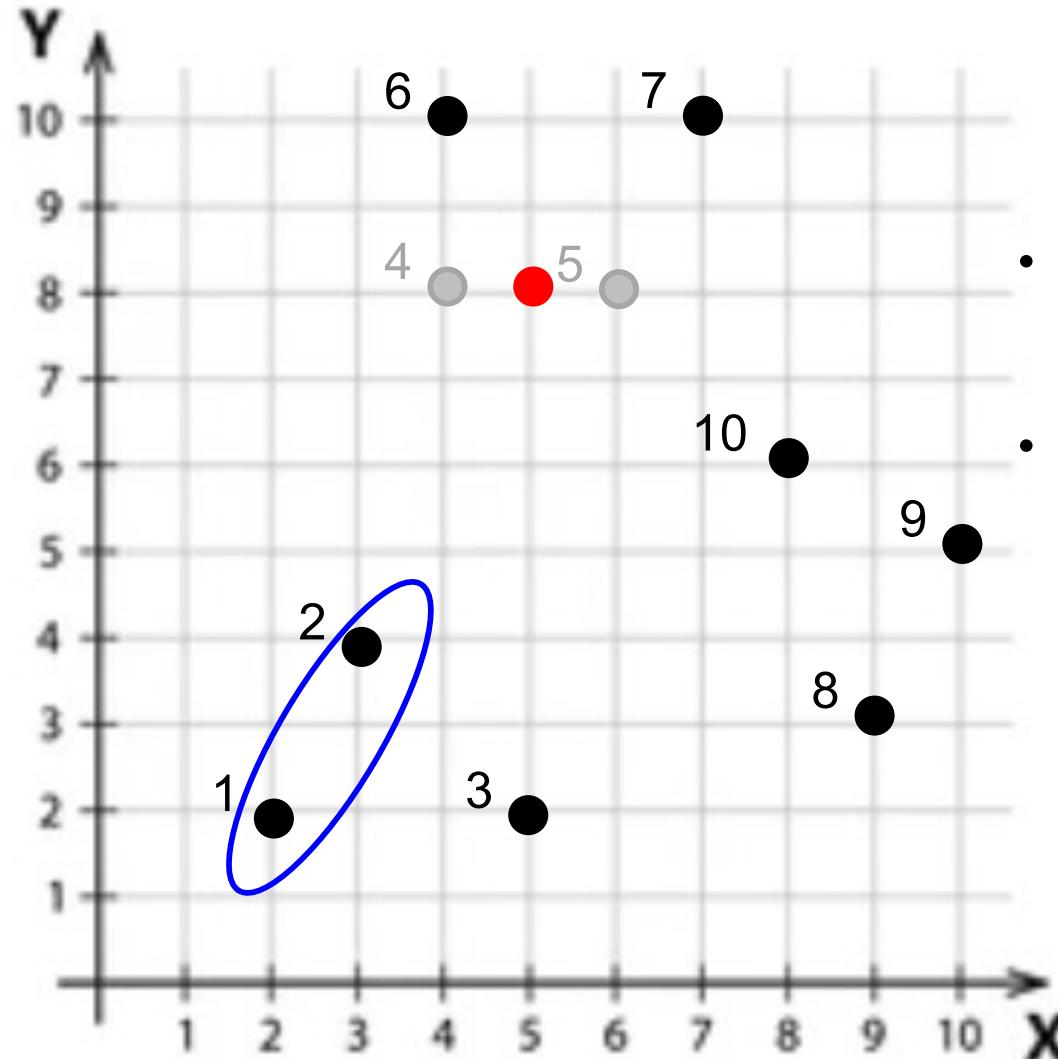


- Συγχώνευση σημείων σε μια συστάδα
- Αναπαράσταση με το **κέντρο**<sup>(\*)</sup>

<sup>(\*)</sup> Υπάρχουν και εναλλακτικοί τρόποι

# Παράδειγμα: Σύνολο 10 Σημείων

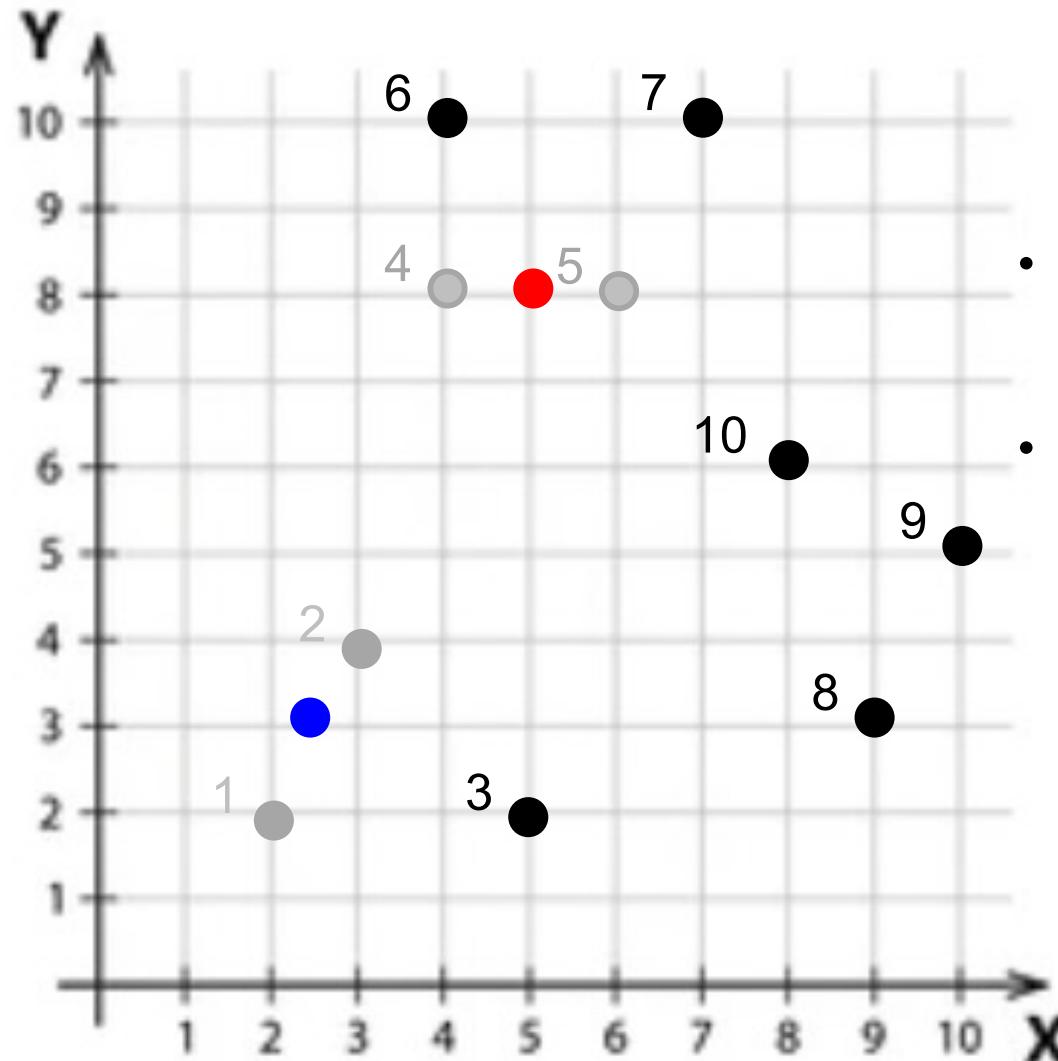
	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6



- Ενημέρωση των αποστάσεων
- Εύρεση κοντινότερου ζεύγους σημείων

# Παράδειγμα: Σύνολο 10 Σημείων

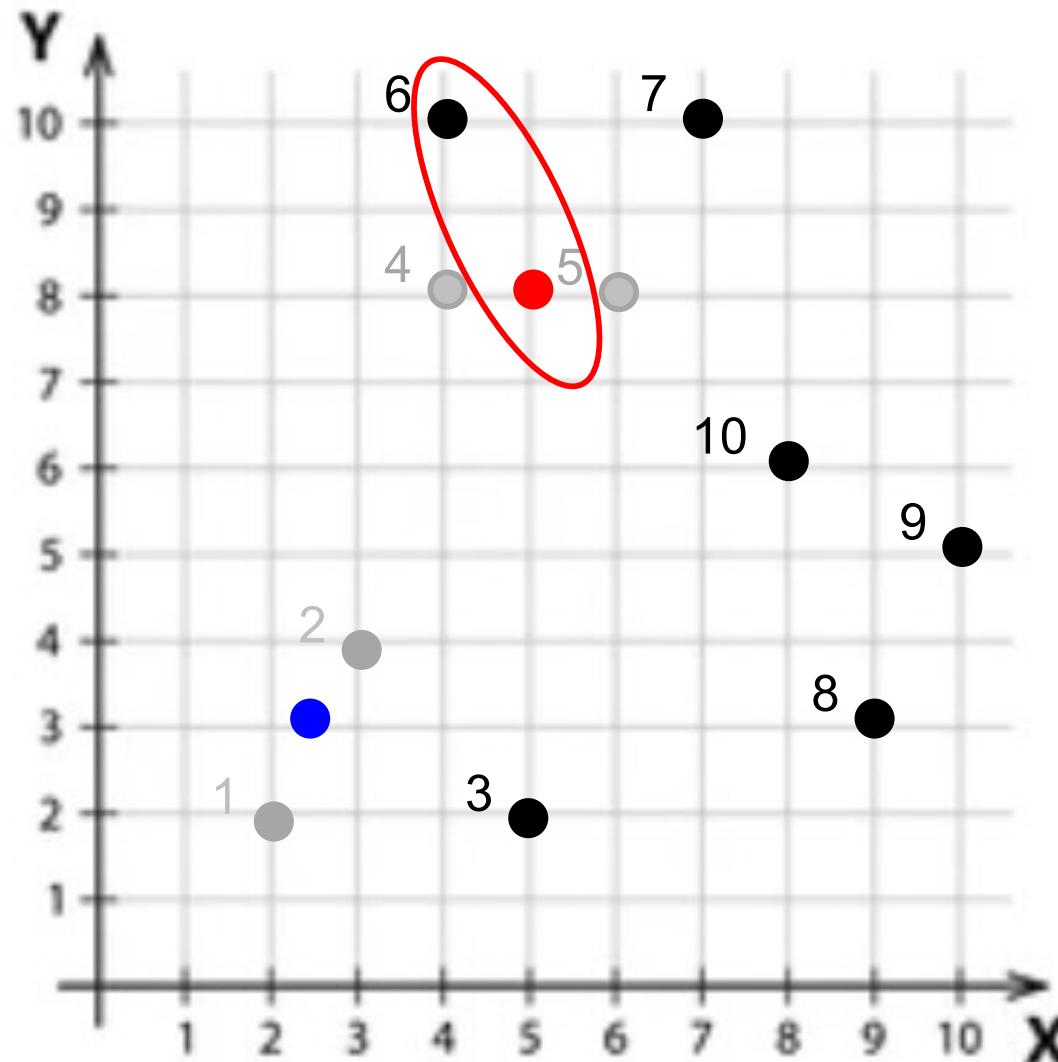
	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6



- Συγχώνευση σημείων σε μια συστάδα
- Αναπαράσταση με το **κέντρο**

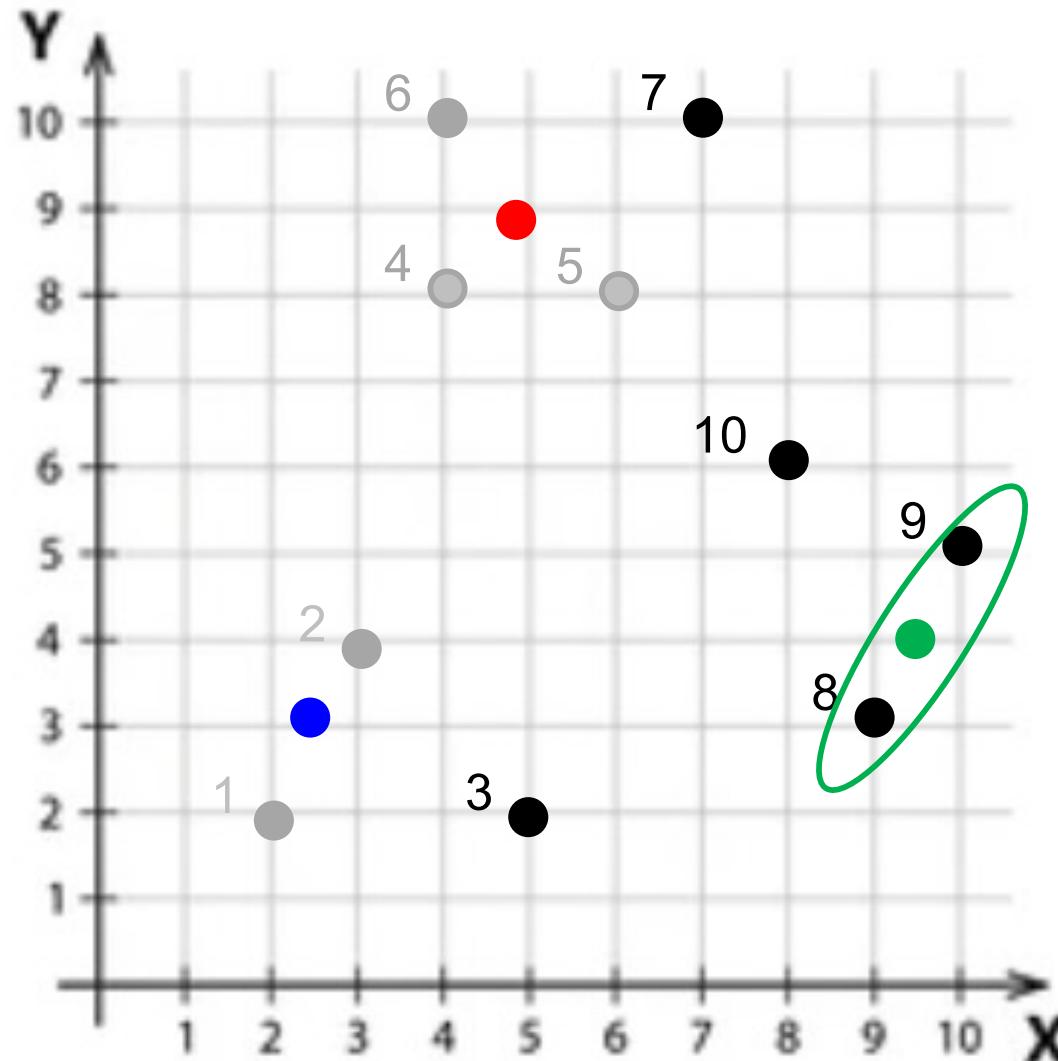
# Παράδειγμα: Σύνολο 10 Σημείων

	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6



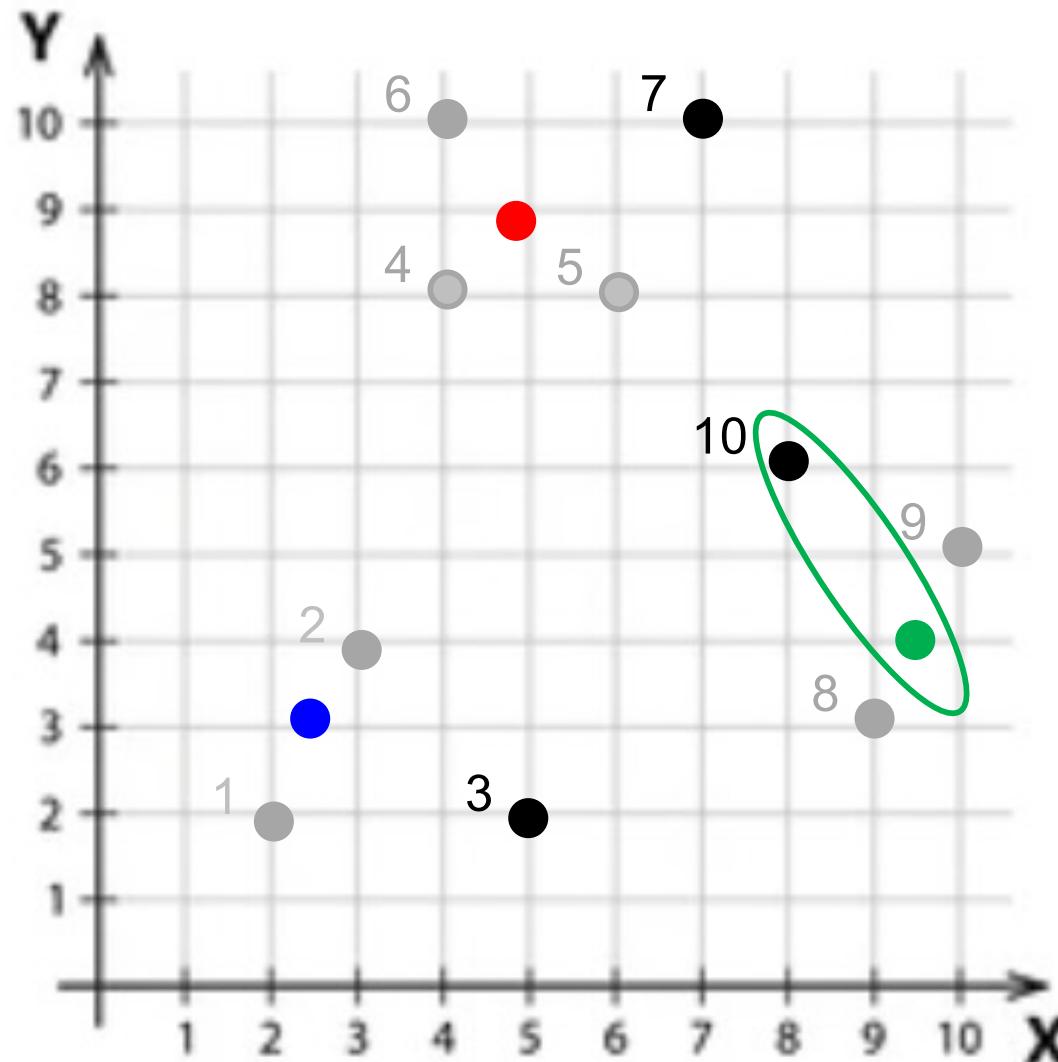
# Παράδειγμα: Σύνολο 10 Σημείων

	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6



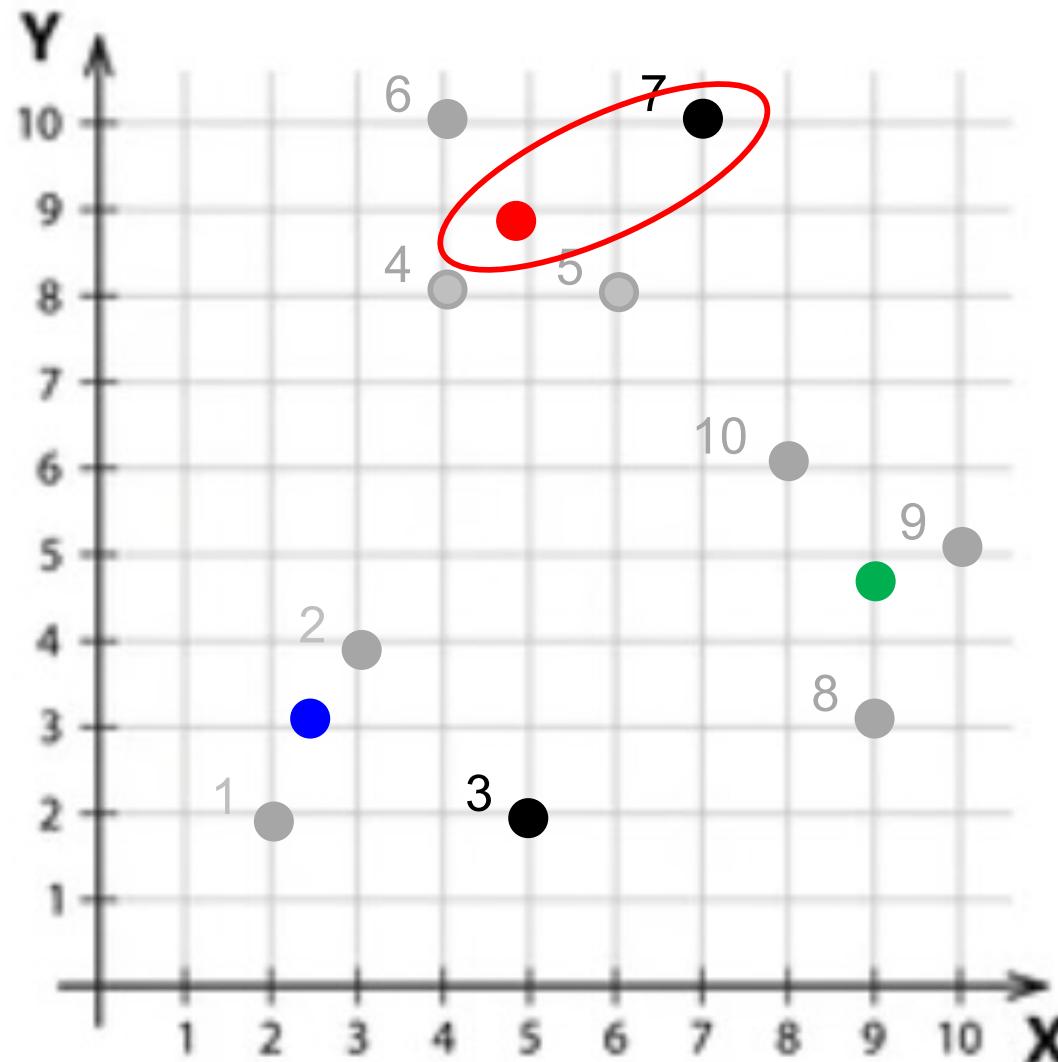
# Παράδειγμα: Σύνολο 10 Σημείων

	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6



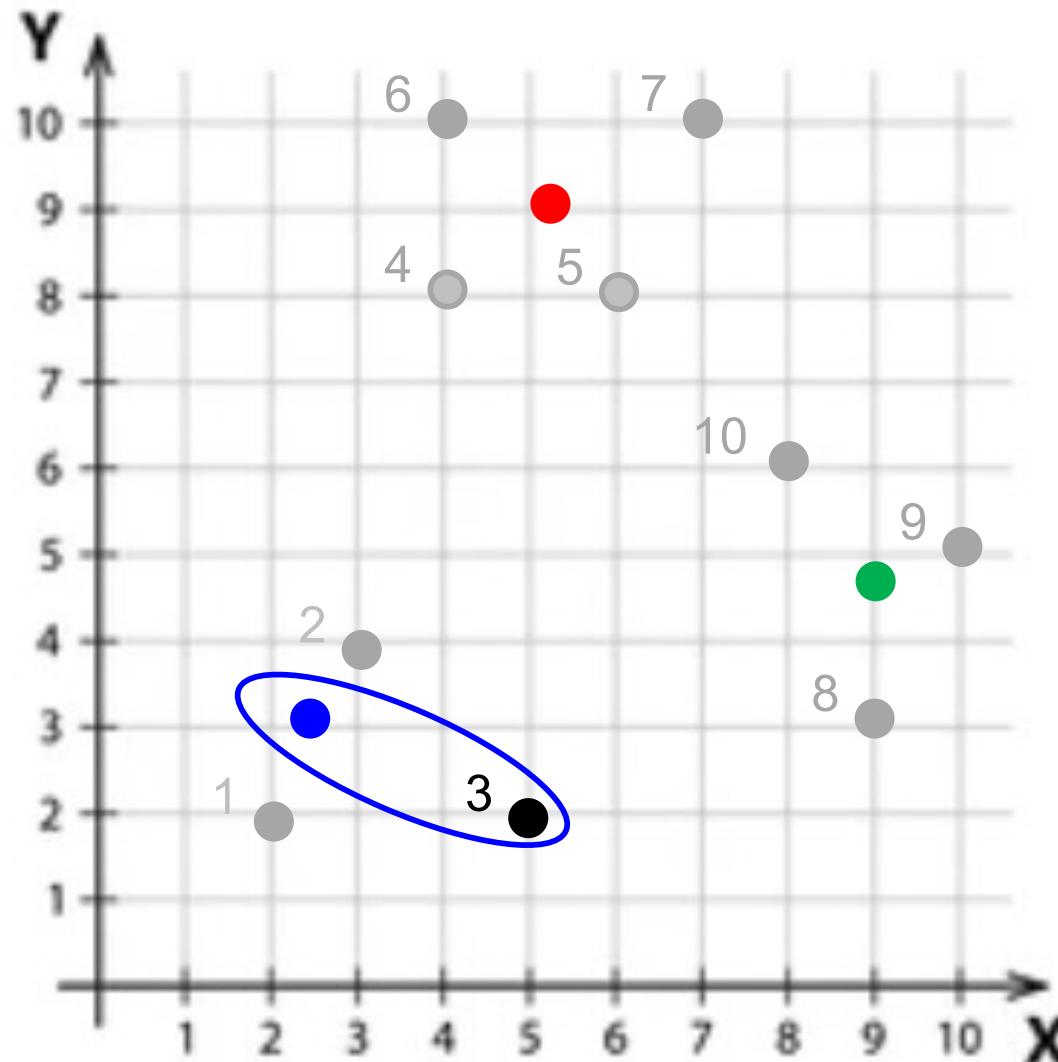
# Παράδειγμα: Σύνολο 10 Σημείων

	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6



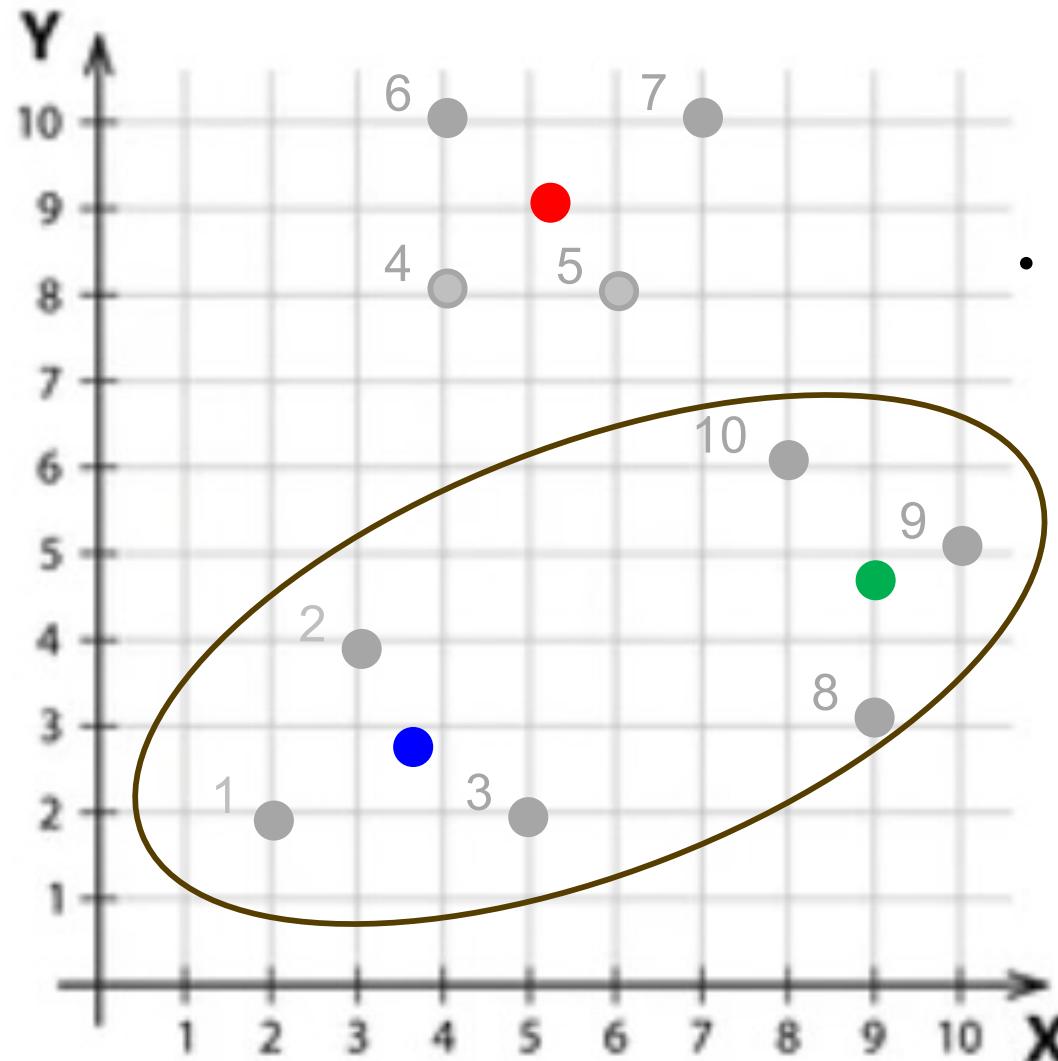
# Παράδειγμα: Σύνολο 10 Σημείων

	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6



# Παράδειγμα: Σύνολο 10 Σημείων

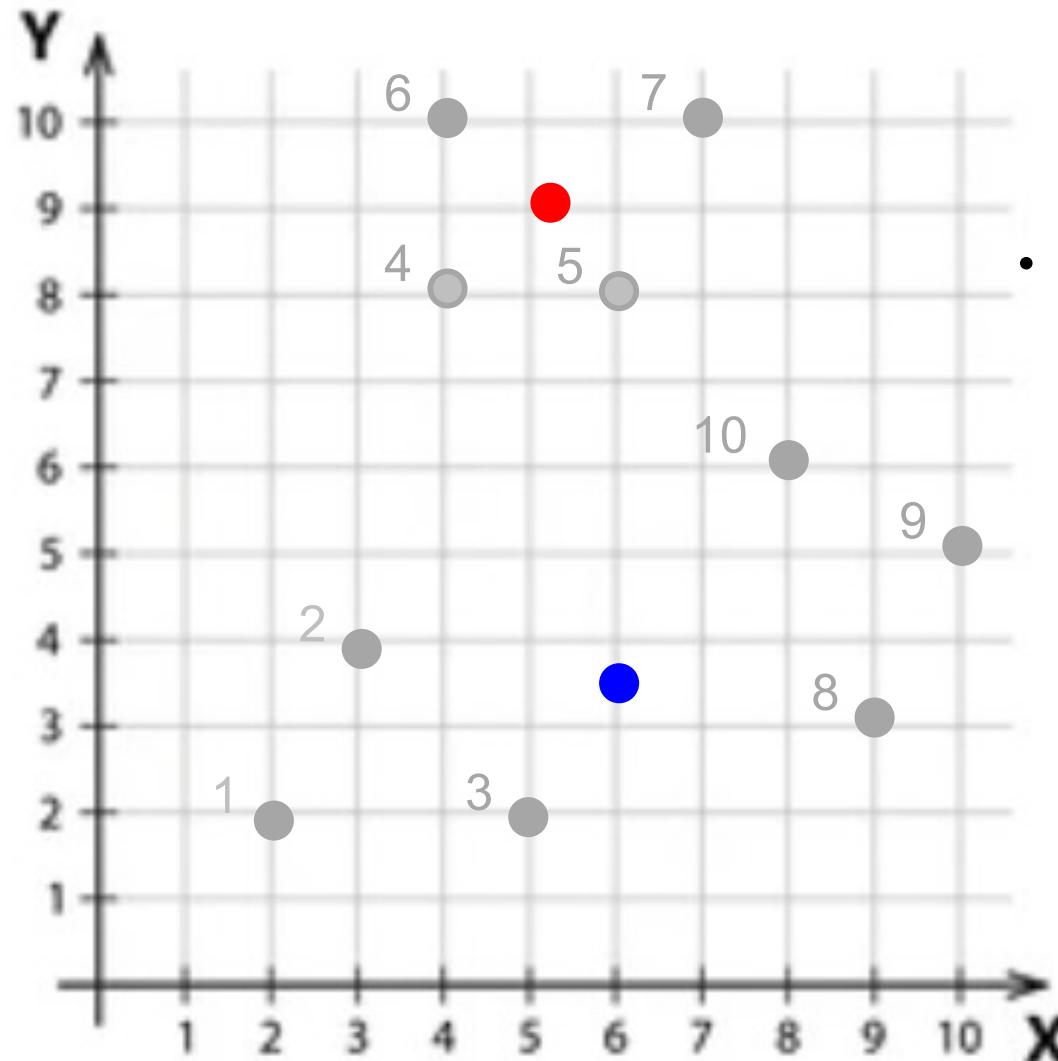
	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6



- Ως τώρα, δεν έτυχε συγχώνευση δύο συστάδων

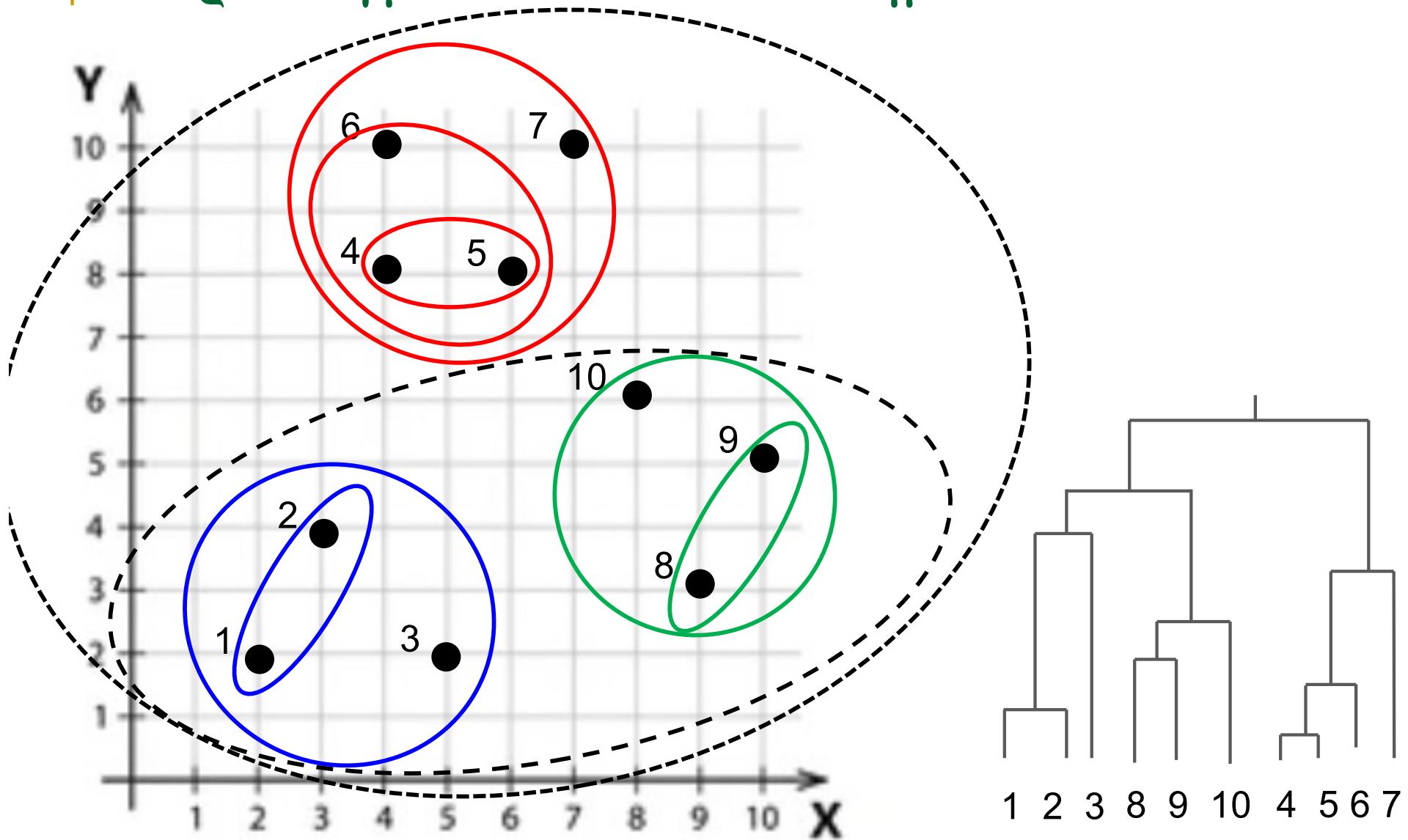
# Παράδειγμα: Σύνολο 10 Σημείων

	X	Y
1	2	2
2	3	4
3	5	2
4	4	8
5	6	8
6	4	10
7	7	10
8	9	3
9	10	5
10	8	6



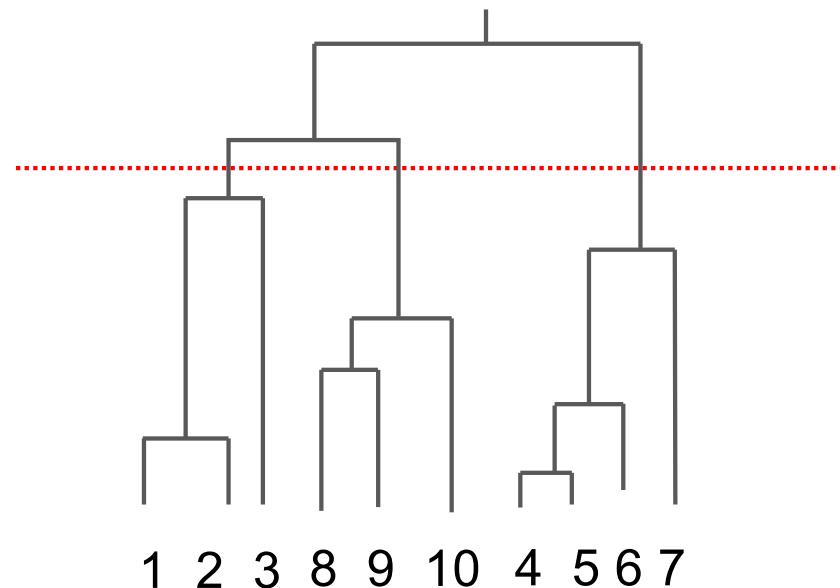
- Τελευταία συγχώνευση συστάδων

# Παράδειγμα: Σύνολο 10 Σημείων



# Αποτέλεσμα Ιεραρχικής Συσταδοποίησης – Δενδρόγραμμα

- Για να βρούμε τις τελικές συστάδες πρέπει:
  - να διασχίσουμε το δέντρο,
  - να αποτιμήσουμε τη συσταδοποίηση για κάθε υπόδεντρο, και
  - να «**κόψουμε**» το δέντρο για να καταλήξουμε σε συστάδες



# Απόσταση Μεταξύ Συστάδων – Εναλλακτικοί Τρόποι (1/2)

## ■ **Minimum ή single link**

- Η απόσταση μεταξύ των δύο κοντινότερων σημείων (ένα από κάθε συστάδα)
- Οδηγεί σε **επεκταμένες** και **αδύναμα συνδεδεμένες** συστάδες
- Μπορεί να **χειριστεί παράξενα σχήματα**, αλλά είναι **ευαίσθητη σε ακραίες τιμές**

## ■ **Maximum ή complete link**

- Η απόσταση μεταξύ των δύο μακρινότερων σημείων (ένα από κάθε συστάδα)
- Δύο συστάδες θα συνενωθούν όταν όλα τα σημεία καθεμιάς συστάδας συνδέονται με τα άλλα
- Οδηγεί σε **συνεκτικές** και **σφαιροειδείς** συστάδες

# Απόσταση Μεταξύ Συστάδων – Εναλλακτικοί Τρόποι (2/2)

## ■ Average

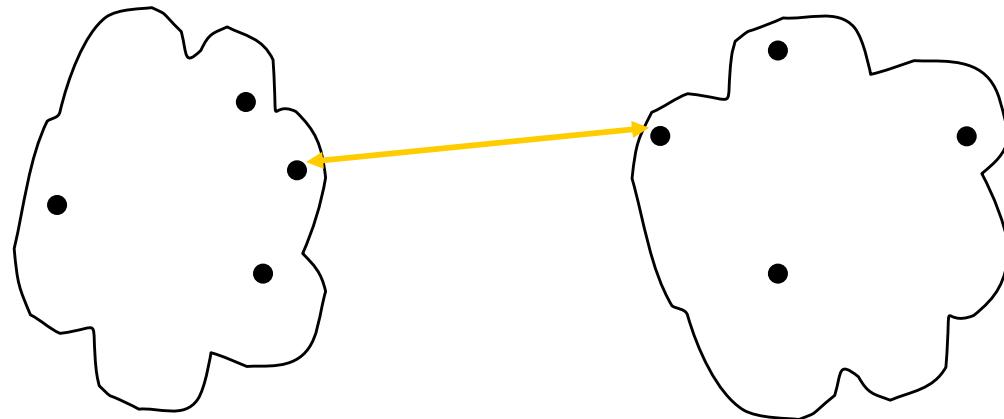
- Παίρνουμε το μέσο όρο όλων των αποστάσεων μεταξύ ζευγών σημείων (ένα από κάθε συστάδα)
- Συνδυάζει χαρακτηριστικά των δύο προηγούμενων προσεγγίσεων

## ■ Centroid

- Για κάθε συστάδα υπολογίζουμε ένα κέντρο, και ορίζουμε την απόσταση μεταξύ συστάδων ως την απόσταση μεταξύ κέντρων

## ■ Μέθοδος Ward

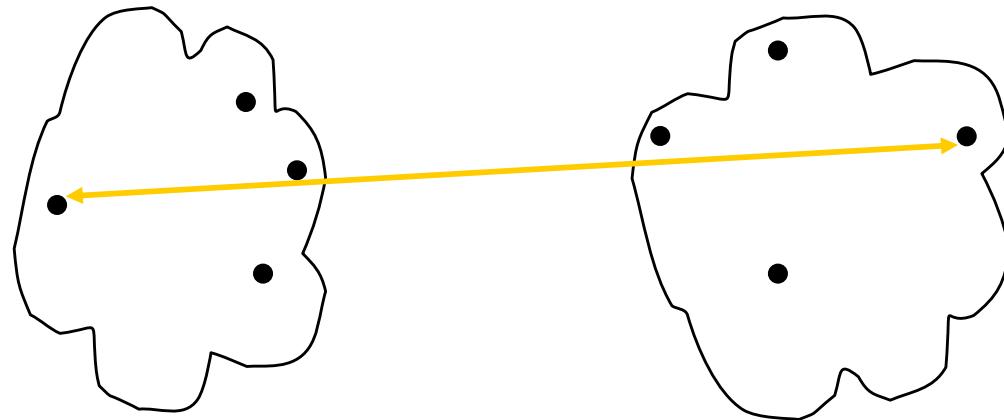
- Μετράει την απόσταση μεταξύ δύο συστάδων ως τη μείωση της συνεκτικότητας (coherence) που παρατηρείται όταν συνενώνονται δύο συστάδες



- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.	.	.	.	.	.	.

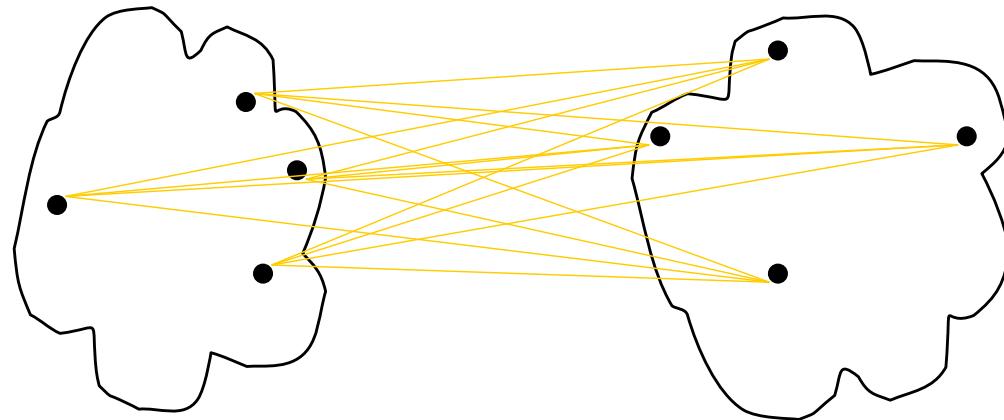
**Proximity Matrix**



- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.	.	.	.	.	.	.

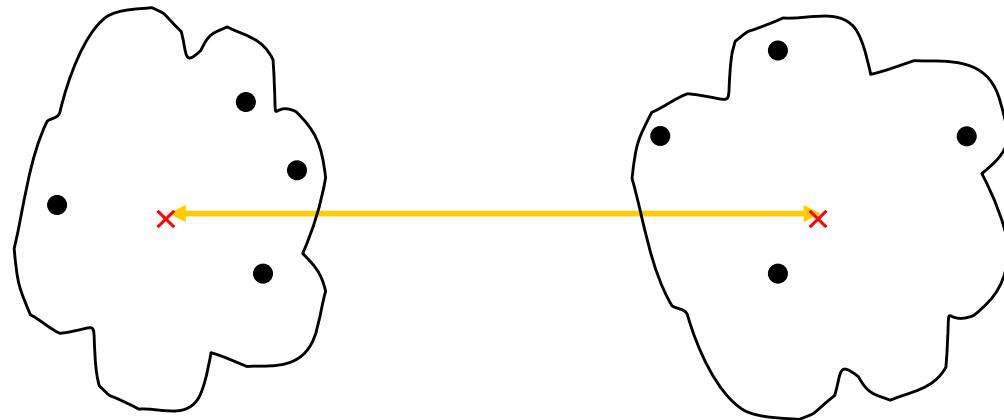
**Proximity Matrix**



- MIN
- MAX
- **Group Average**
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						

**Proximity Matrix**



- MIN
- MAX
- Group Average
- **Distance Between Centroids**
- Other methods driven by an objective function
  - Ward's Method uses squared error

	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						

### Proximity Matrix

# Απόδοση

- Οι αλγόριθμοι κατασκευής δέντρων είναι «**ακριβοί**» υπολογιστικά
  - Απαιτούν τον πίνακα αποστάσεων μεταξύ όλων των σημείων:  $O(n^2)$
  - Η κατασκευή ολόκληρου του δέντρου απαιτεί  $O(n)$  επαναλήψεις
  - Συνολικό κόστος:  $O(n^3)$
  - Μειώνεται σε  $O(n^2 \log n)$  με χρήση μεθόδων ευρετηρίασης (indexing)
- Οι αλγόριθμοι κατασκευής δέντρων έχουν ένα χαρακτηριστικό
  - Δεν παράγουν απλά μια επίπεδη ομαδοποίηση των σημείων σε συστάδες
  - Άλλα αναδεικνύουν τις συσχετίσεις τους με ρητό τρόπο
  - Πρέπει ο χρήστης να αποφασίσει εάν αυτό του είναι χρήσιμο για την εκάστοτε εφαρμογή
  - Η επιλογή του μέτρου απόστασης συστάδων μπορεί να επηρεάσει σημαντικά την εμφάνιση της δεντρικής δομής

# Τύποι Αλγορίθμων Συσταδοποίησης

- Αλγόριθμοι που αναζητούν κέντρα
  - K-means
- Αλγόριθμοι κατασκευής δέντρων
  - Hierarchical agglomerative clustering
- Αλγόριθμοι μεγέθυνσης γειτονιών
  - DBSCAN

# 3. Αλγόριθμοι Μεγέθυνσης Γειτονιών DBSCAN

- Οι αλγόριθμοι αυτής της κατηγορίας συνδέουν μαζί σημεία που βρίσκονται σε κοντινή απόσταση ώστε να ανήκουν στην ίδια συστάδα
  - Συνεχίζουν με αυτό τον τρόπο ώσπου να αναθέσουν όλα τα σημεία
- ***Συσταδοποίηση βάσει πυκνότητας***
- Πρόκειται για την πιο άμεση εφαρμογή του ορισμού συστάδας ως μια **περιοχή υψηλής πυκνότητας, χωρίς να κάνει υποθέσεις για το σχήμα** μιας συστάδας
- Τέτοιοι αλγόριθμοι αποτελούν ένα εξειδικευμένο εργαλείο καθώς εφαρμόζονται
  - Είτε όταν άλλοι αλγόριθμοι αποτυγχάνουν ή
  - Για να εκλεπτύνουν τα αποτελέσματα συσταδοποίησης που έχουν παραχθεί από έναν πιο «κλασικό» αλγόριθμο, όπως ο k-means

# Ο Αλγόριθμος DBSCAN<sup>1</sup>

- Απαιτεί δύο παραμέτρους εισόδου
  - Την **ελάχιστη πυκνότητα** (**minimum density**) που αναμένουμε εντός μιας συστάδας
    - Σημεία που βρίσκονται σε περιοχές μικρότερης πυκνότητας δε θα σχηματίσουν συστάδες
  - Το **μέγεθος περιοχής** (**size of a region**) όπου αναμένεται να εντοπίσουμε την παραπάνω τιμή πυκνότητας
    - Θα πρέπει να είναι μεγαλύτερη από τη μέση απόσταση μεταξύ γειτονικών σημείων, αλλά μικρότερη από ολόκληρη τη συστάδα
- Στην πράξη χρησιμοποιούμε δύο ελαφρώς διαφορετικές παραμέτρους για λόγους ευκολίας
  - Την ακτίνα **r** γειτονιάς (απαντάται και ως **Eps**)
  - Τον ελάχιστο αριθμό σημείων **n** (απαντάται και ως **MinPts**) που αναμένεται να βρούμε στη γειτονιά κάθε σημείου εντός μιας συστάδας

<sup>1</sup>Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD 1996: 226-231

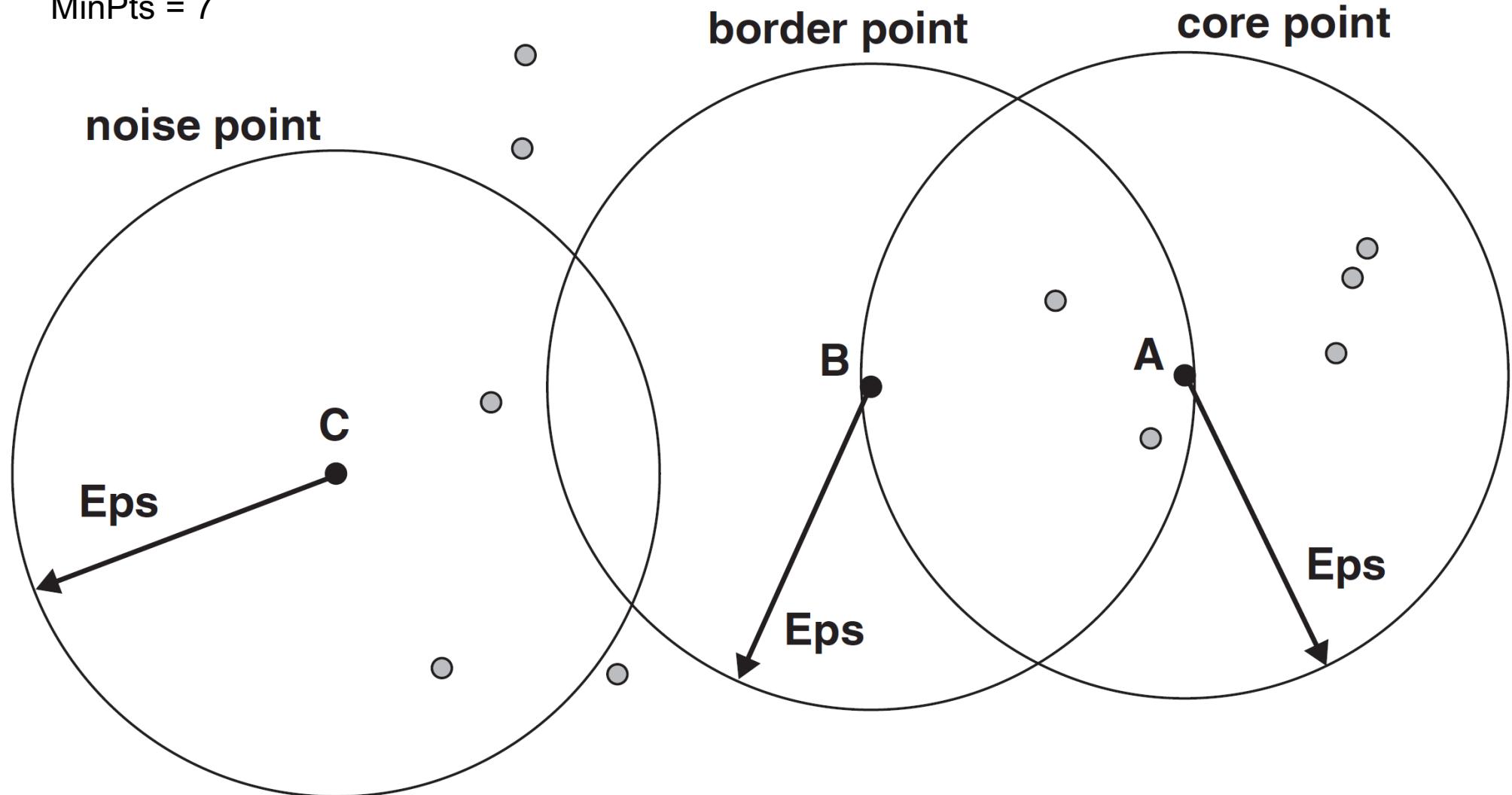
# Ο Αλγόριθμος DBSCAN

## ■ Διακρίνονται τρεις τύποι σημείων

- **Noise point**: ένα σημείο που έχει λιγότερα από **n** σημεία στη γειτονιά του (ακτίνας **r**). Τέτοια σημεία δεν ανατίθενται σε συστάδες.
- **Core point**: περιέχει τουλάχιστον **n** γείτονες<sup>(\*)</sup>
- **Edge/border point**: έχει λιγότερους γείτονες από ότι απαιτείται για ένα core point, όμως το ίδιο είναι στη γειτονιά ενός core point
- (\*) γείτονας ενός σημείου είναι οποιοδήποτε άλλον σημείο βρίσκεται εντός απόστασης **r**

# Παράδειγμα

MinPts = 7

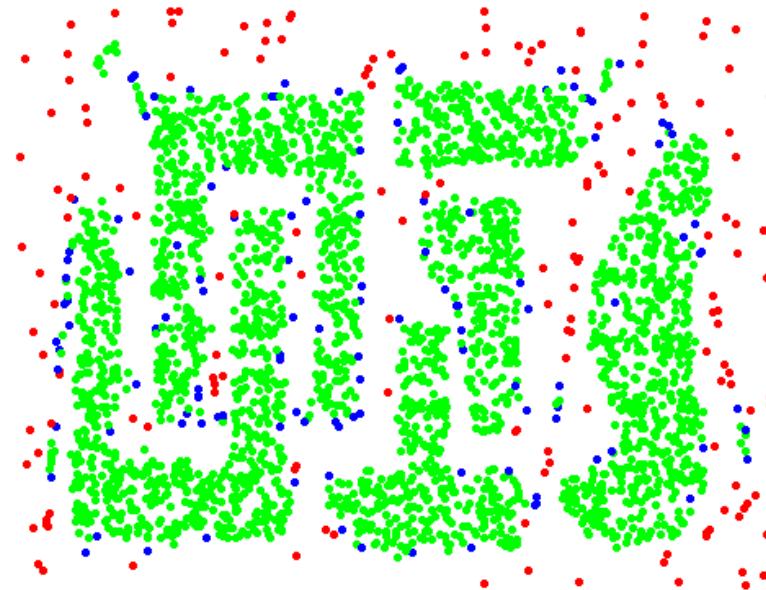


# DBSCAN: Core, Border και Noise Points



Original Points

Eps = 10, MinPts = 4



Point types: **core**,  
**border** and **noise**

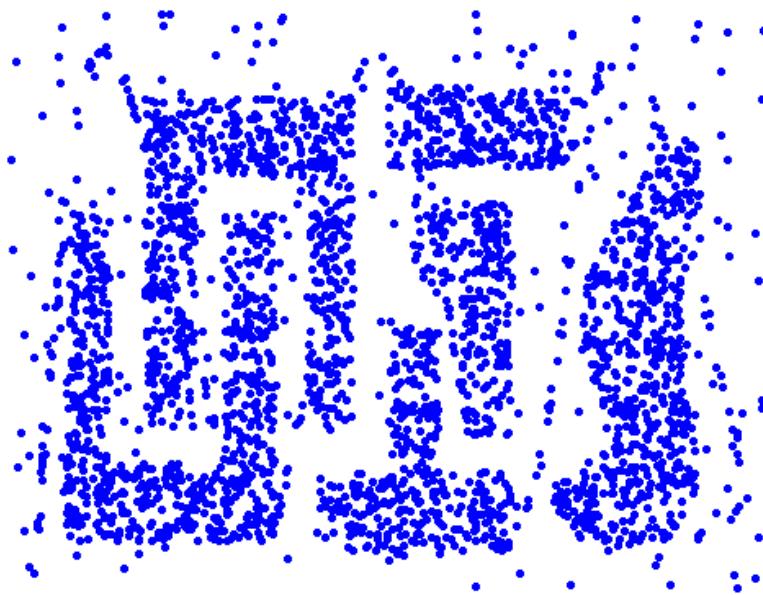
# Αλγόριθμος

## Algorithm 7.5 DBSCAN algorithm.

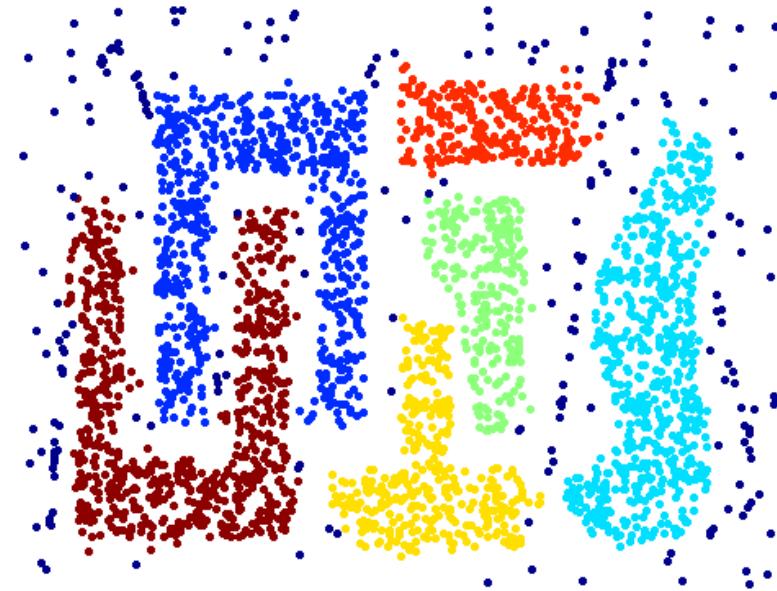
- 1: Label all points as core, border, or noise points.
- 2: Eliminate noise points.
- 3: Put an edge between all core points within a distance  $Eps$  of each other.
- 4: Make each group of connected core points into a separate cluster.
- 5: Assign each border point to one of the clusters of its associated core points.

- Οποιαδήποτε δύο σημεία πυρήνα (core points) είναι εντός απόστασης  $Eps$ , τοποθετούνται στην ίδια συστάδα
- Κάθε σημείο ορίου (border point) που είναι αρκετά κοντά σε σημείο πυρήνα, τοποθετείται στην ίδια συστάδα με το σημείο πυρήνα

# Παράδειγμα Επιτυχούς Λειτουργίας



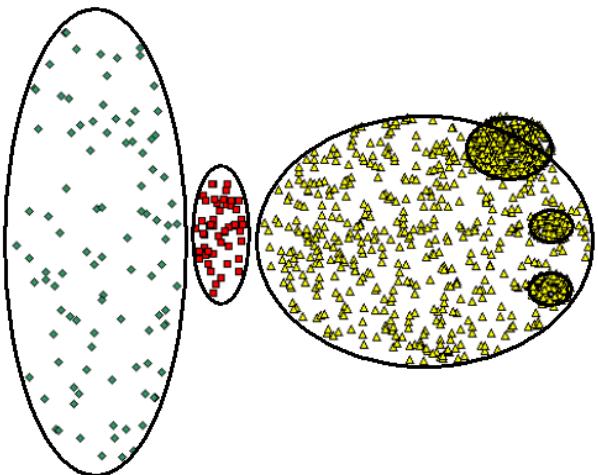
Αρχικό σύνολο δεδομένων



Συστάδες

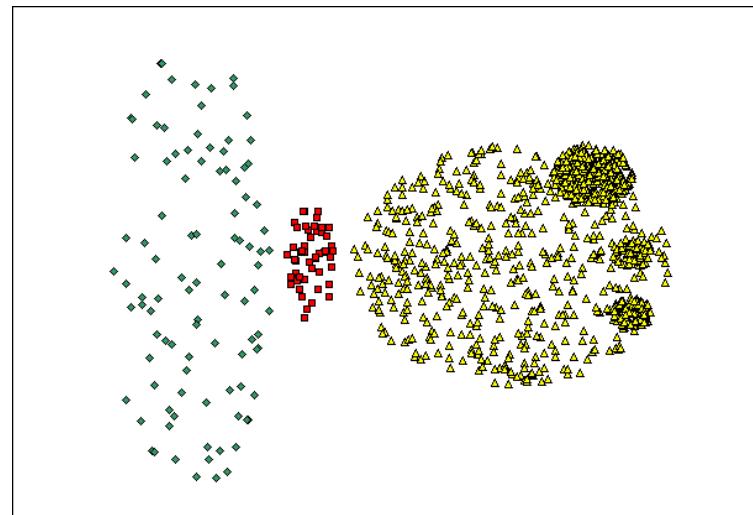
- Ανθεκτικός αλγόριθμος σε θόρυβο
- Μπορεί να χειριστεί συστάδες διαφορετικού σχήματος και μεγέθους

# Παράδειγμα Μη Επιτυχούς Λειτουργίας

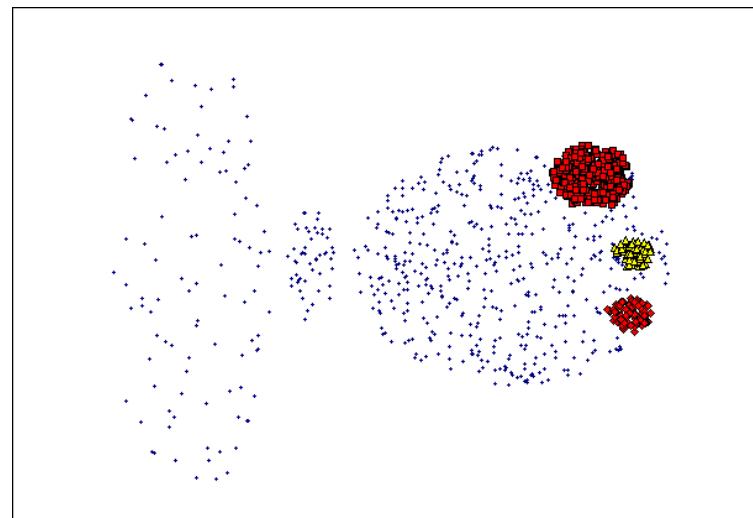


Αρχικό σύνολο δεδομένων

- Μεταβλητή πυκνότητα
- Δεδομένα υψηλής διάστασης



MinPts=4  
Eps=9.75



MinPts=4  
Eps=9.92

# Αποτίμηση του DBSCAN

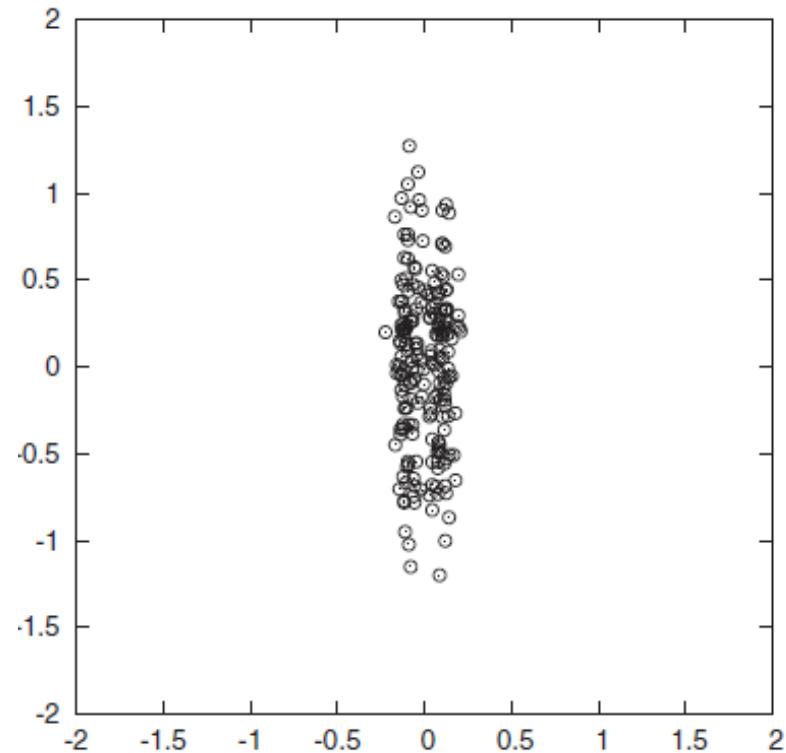
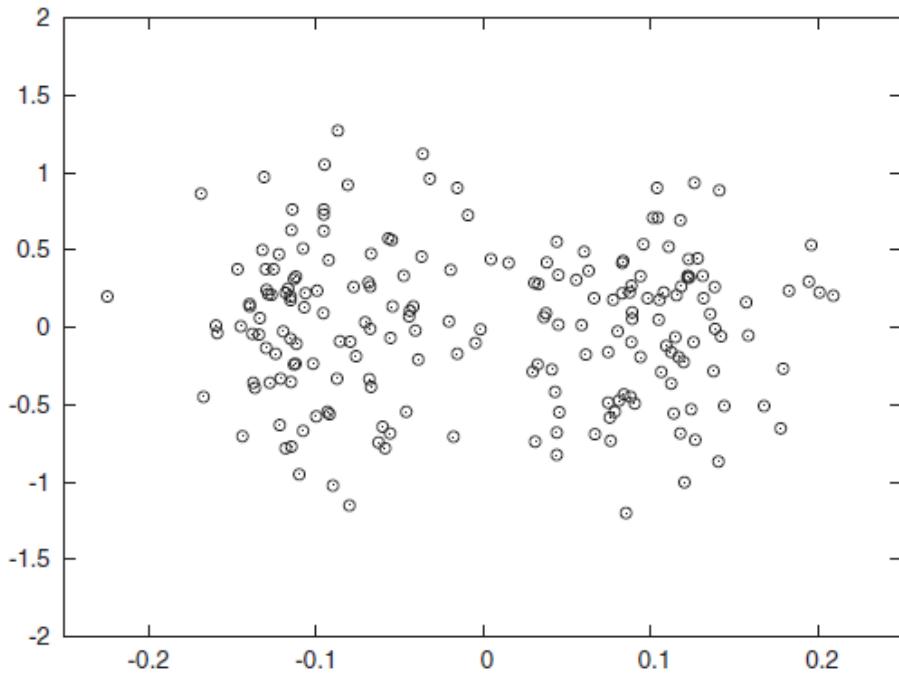
- Βασικό πλεονέκτημα
  - μπορεί να ανακαλύψει συστάδες παράξενων γεωμετρικών σχημάτων
  - καθώς και εμφωλευμένες συστάδες
- Είναι ευαίσθητος όμως στις τιμές των παραμέτρων εισόδου του
  - και ο καθορισμός κατάλληλων τιμών είναι επίπονη διαδικασία
- Μερικές φορές μπορεί να συνδυαστεί με τον k-means
  - Σε πρώτο πέρασμα χρησιμοποιείται ο k-means για τον εντοπισμό υποψήφιων συστάδων
  - Στατιστικά αυτών (όπως ακτίνα και πυκνότητα) μπορεί να δοθούν στον DBSCAN ως είσοδος
- Υπολογιστικό κόστος: καθορίζεται από τις αναζητήσεις γειτονικών σημείων
  - Για κάθε σημείο, ελέγχονται όλα τα υπόλοιπα  $\rightarrow O(n^2)$
  - Μπορεί να βελτιωθεί με χρήση κατάλληλων χωρικών ευρετηρίων

# Προεπεξεργασία & Μετεπεξεργασία

---

# Προεπεξεργασία

*Μπορείτε να αναγνωρίσετε συστάδες;*



Όταν τα δεδομένα βρίσκονται σε πολύ διαφορετικά διαστήματα τιμών, πρέπει να γίνει **κανονικοποίηση** πριν εκτελεστεί ο αλγόριθμος συσταδοποίησης

# Κανονικοποίηση

- Υπάρχουν διαφορετικοί τρόποι κανονικοποίησης
  - #1: διαίρεση κάθε τιμής κάθε διάστασης με το εύρος τιμών των δεδομένων στην αντίστοιχη διάσταση
  - #2: διαίρεση κάθε τιμής κάθε διάστασης με την τιμή της τυπικής απόκλισης της αντίστοιχης διάστασης

# Ποιότητα Συσταδοποίησης

- Οι βασικές έννοιες είναι η συνεκτικότητα (*cohesion*) σε μια συστάδα και ο διαχωρισμός (*separation*) ανάμεσα σε συστάδες
  - Η μέση απόσταση όλων των σημείων εντός μιας συστάδας είναι ένα μέτρο της συνεκτικότητας
  - Η μέση απόσταση όλων των σημείων μιας συστάδας από όλα τα σημεία μιας άλλης συστάδας είναι ένα μέτρο του διαχωρισμού μεταξύ των δύο συστάδων
- Εάν το σύνολο δεδομένων μπορεί να χωριστεί ξεκάθαρα σε ομάδες, τότε αναμένουμε να έχουμε μεγάλη απόσταση μεταξύ συστάδων συγκριτικά με την ακτίνα των συστάδων

separation  
cohesion

Θέλουμε ο λόγος αυτός να  
είναι μεγάλος

Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis: On Clustering Validation Techniques. J. Intell. Inf. Syst. 17(2-3): 107-145 (2001).

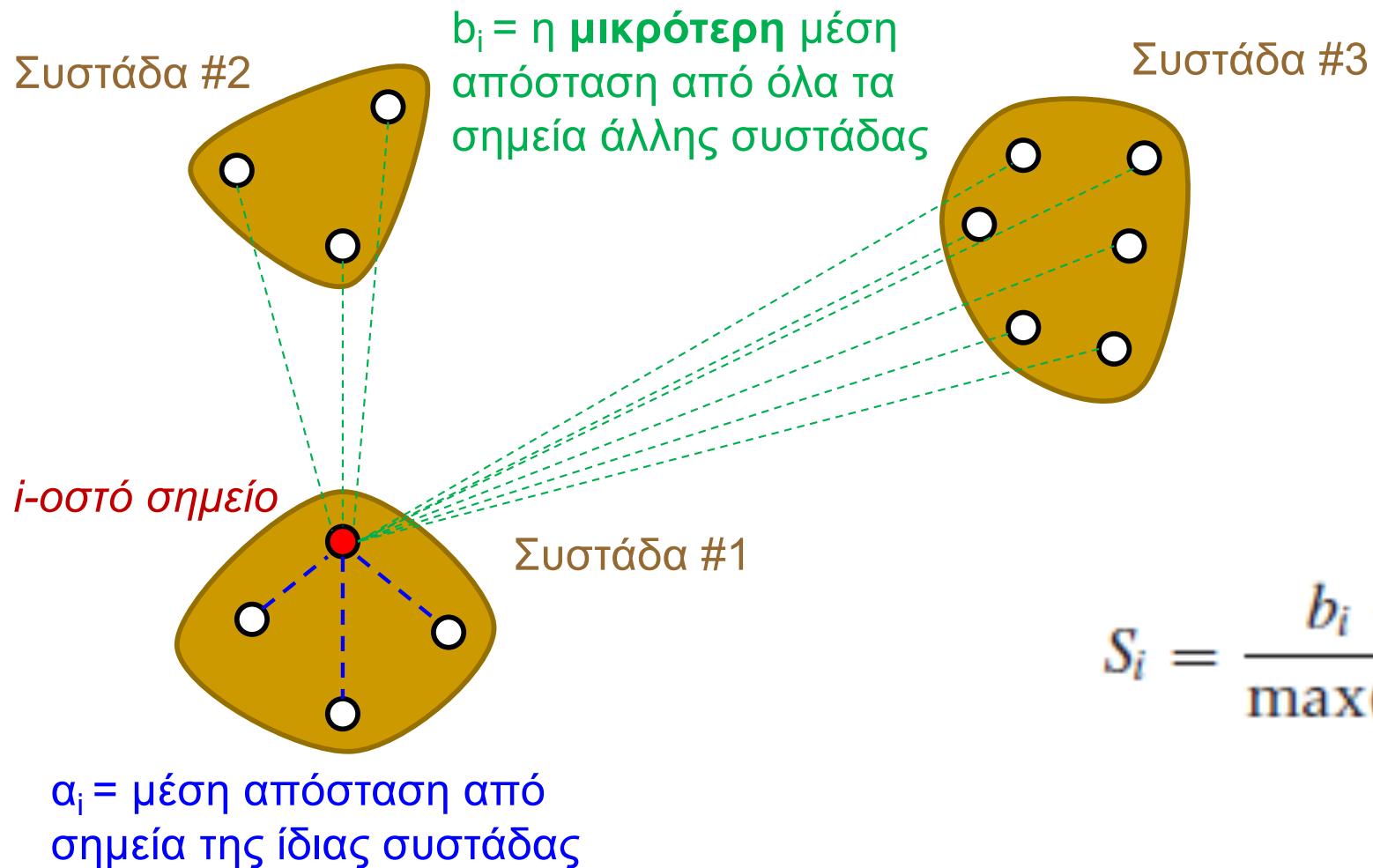
# Silhouette coefficient

- Ορίζεται για σημεία:

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

- Έστω  $a_i$  η μέση απόσταση που έχει το σημείο  $i$  από όλα τα άλλα σημεία **εντός της συστάδας** στην οποία ανήκει (**συνεκτικότητα**)
- Μετράμε τη μέση απόσταση του σημείου  $i$  από όλα τα σημεία οποιασδήποτε **άλλης συστάδας** που δεν ανήκει, και έστω  $b_i$  η μικρότερη τέτοια τιμή (το  $b_i$  είναι ο **διαχωρισμός** από την κοντινότερη συστάδα)
- Ο αριθμητής είναι δείκτης του “κενού χώρου” μεταξύ συστάδων
- Ο παρονομαστής είναι η μεγαλύτερη τιμή από δύο μήκη: την ακτίνα των συστάδων και την απόσταση μεταξύ συστάδων

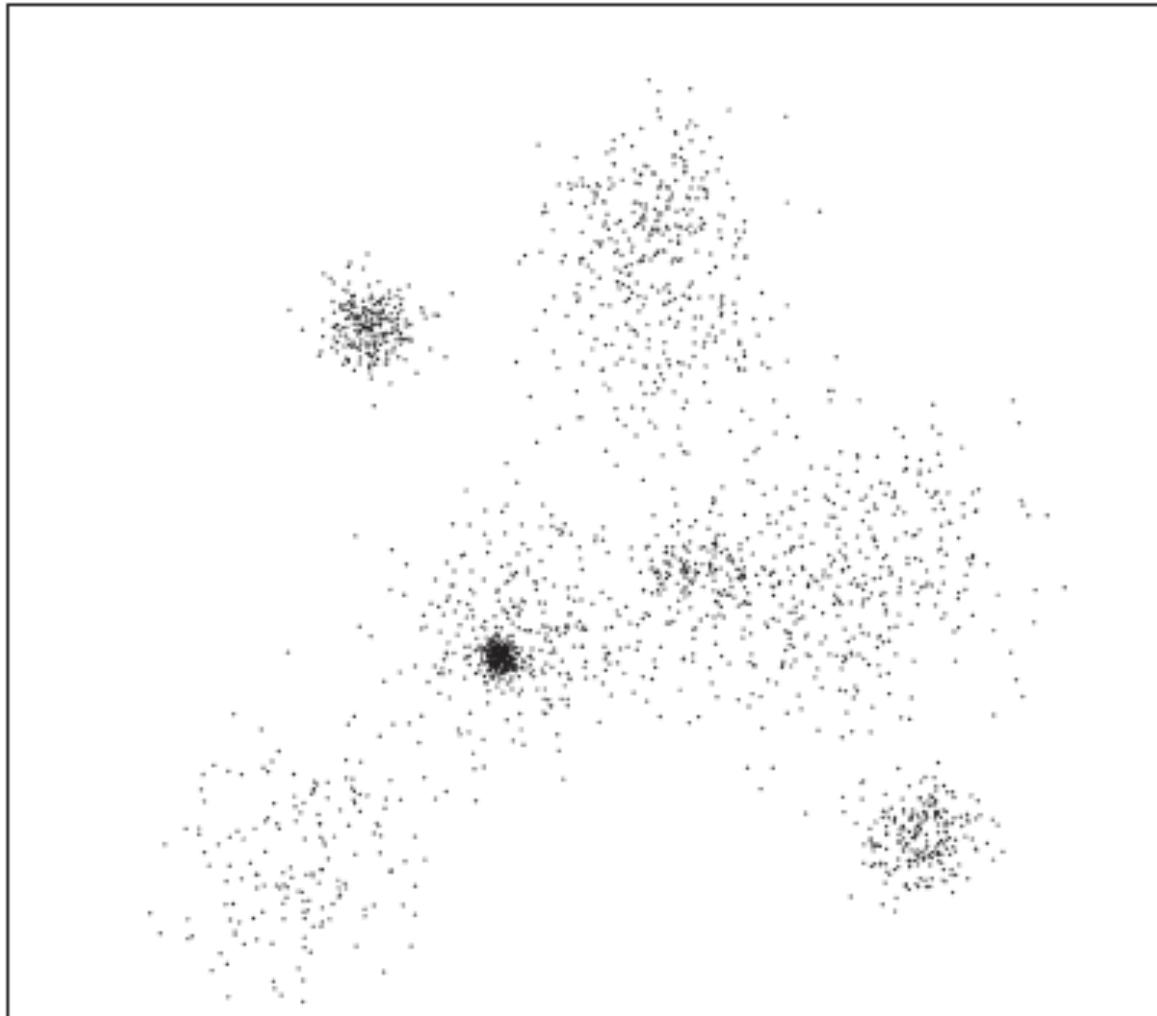
# Silhouette coefficient (Παράδειγμα)



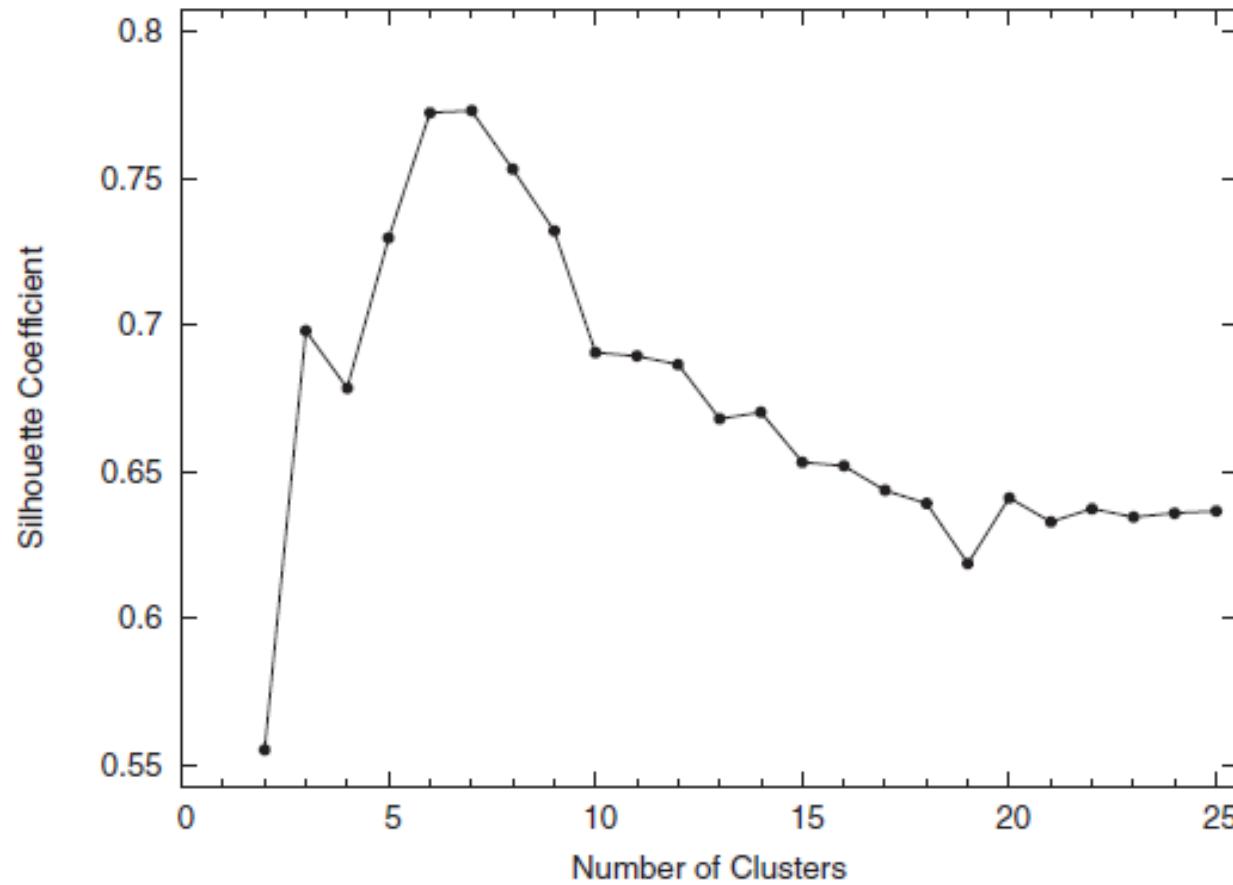
# Silhouette coefficient (συνέχ.)

- To silhouette coefficient παίρνει τιμές από -1 ως 1
  - Αρνητικές τιμές σημαίνουν ότι η ακτίνα των συστάδων είναι μεγαλύτερη από την απόσταση μεταξύ συστάδων → ένδειξη **κακής συσταδοποίησης**
  - Υψηλές θετικές τιμές αποτελούν ένδειξη **καλής συσταδοποίησης**
- Μπορούμε να ορίσουμε το μέσο όρο όλων των τιμών του silhouette coefficient για όλα τα σημεία, ως το **συνολικό silhouette coefficient** για όλο το σύνολο δεδομένων

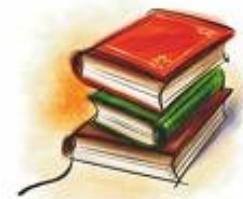
# Πόσες Συστάδες Υπάρχουν σε αυτό το Σύνολο Δεδομένων;



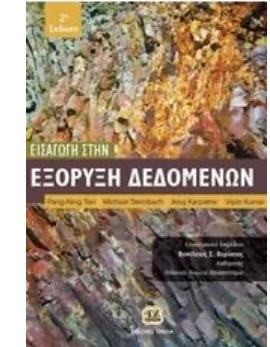
# Μετρώντας το Silhouette coefficient



# Βιβλιογραφικές Πηγές

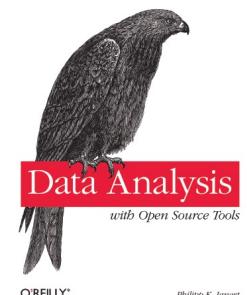


- P. Tan, M. Steinbach, A. Karpatne, V. Kumar. “Εισαγωγή στην Εξόρυξη Δεδομένων”, 2<sup>η</sup> Έκδοση, Εκδόσεις Τζιόλα.
  - *Κεφάλαιο 7*



A Hands-On Guide for Programmers and Data Scientists

- Philipp K. Janert. “Data Analysis with Open Source Tools” 2011, O'Reilly, 978-0-596-80235-6.
  - *Κεφάλαιο 13*





## 8. Ανάλυση Συσχέτισης (Association Analysis)



---

**Ανάλυση Δεδομένων**  
**(Data Analytics)**

Χρήστος Δουλκερίδης  
2024-25

# Παράδειγμα Συναλλαγών Καλαθιού Αγοράς

<i>TID</i>	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

- **Συναλλαγές καλαθιού αγοράς (market basket transactions)**
  - Εμφανίζονται σε πολλές σύγχρονες επιχειρήσεις
  - Η ανάλυσή τους αναδεικνύει την **αγοραστική συμπεριφορά** πελατών
  - Υποβοηθά εφαρμογές όπως:
    - **προώθηση προϊόντων**
    - **διαχείριση αποθήκης**
    - **διαχείριση σχέσεων με πελάτες**

# Ανάλυση Συσχέτισης

- Η ανάλυση συσχέτισης (**association analysis**) είναι μια μεθοδολογία χρήσιμη για την ανακάλυψη ενδιαφερουσών **σχέσεων** που είναι κρυμμένες σε μεγάλα σύνολα δεδομένων
- Οι **σχέσεις** μπορούν να αναπαρασταθούν στη μορφή στοιχείων συνόλων που είναι παρόντα σε πολλές συναλλαγές, γνωστά και ως **σύνολα συχνών στοιχείων (frequent itemsets)**
- Παράδειγμα ενός εξαγόμενου κανόνα:
  - {Diapers → Beer}
  - «πελάτες που αγοράζουν πάνες αγοράζουν και μπύρα»
- Άλλες εφαρμογές ανάλυσης συσχέτισης:
  - Βιοπληροφορική, ιατρική διάγνωση, εξόρυξη Ιστού, ανάλυση επιστημονικών δεδομένων, ...

# Παράδειγμα Κανόνα Συσχέτισης

{Diapers → Beer} [support=2%, confidence=60%]

- Σε 2% των συναλλαγών αγοράζονται μαζί diapers και beer
- 60% των πελατών που αγόρασαν diapers, αγόρασαν και beer

Συνήθως ενδιαφέροντες κανόνες είναι αυτοί που ικανοποιούν

- ένα ελάχιστο κατώφλι υποστήριξης (**support**) και
- ένα ελάχιστο κατώφλι εμπιστοσύνης (**confidence**)

# Δύο Βασικά Ζητήματα

- Υπάρχουν δύο βασικά ζητήματα προς αντιμετώπιση όταν εφαρμόζεται ανάλυση συσχετίσεων:
  1. Η ανακάλυψη υποδειγμάτων από ένα μεγάλο σύνολο δεδομένων συναλλαγών μπορεί να είναι **υπολογιστικά ακριβή**
  2. Μερικά από τα υποδείγματα που ανακαλύπτονται είναι **πιθανόν ψευδή**, επειδή μπορεί απλά να είναι **τυχαία**

Επομένως θέλουμε:

- **Αποδοτικούς αλγόριθμους επεξεργασίας**
- **Εύρεση ισχυρών κανόνων (ή συσχετίσεων)**

# Περίγραμμα Μαθήματος

- Βασικές έννοιες
- Παραγωγή συχνών στοιχειοσυνόλων
  - Αλγόριθμος Apriori
- Παραγωγή κανόνων
- Σύντομη αναπαράσταση συχνών στοιχειοσυνόλων

# Αναπαράσταση Δεδομένων Καλαθιού Αγοράς

$$I = \{i_1, i_2, \dots, i_d\}$$

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

$T = \{t_1, t_2, \dots, t_N\}$

- Σύνολο αντικειμένων (ή στοιχείων):  $I = \{i_1, i_2, \dots, i_d\}$
- Σύνολο συναλλαγών:  $T = \{t_1, t_2, \dots, t_N\}$
- Κάθε συναλλαγή  $t_i$  περιέχει ένα υποσύνολο στοιχείων
- Μια συλλογή από 0 ή περισσότερα αντικείμενα λέγεται στοιχειοσύνολο (**itemset**)
- Εάν ένα στοιχειοσύνολο περιέχει  $k$  αντικείμενα ονομάζεται στοιχειοσύνολο- $k$

# Μέτρηση Υποστήριξης

- Μέτρηση υποστήριξης  $\sigma(X)$  ενός στοιχειοσυνόλου  $X$ :
  - Σε πόσες συναλλαγές περιέχεται το  $X$ :

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}|$$

- Π.χ. για  $X = \{\text{Beer}, \text{Diapers}, \text{Milk}\}$  έχουμε  $\sigma(X) = 2$

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

# Υποστήριξη (Support)

- Συχνά, μας ενδιαφέρει η υποστήριξη (support):
  - Το ποσοστό συναλλαγών στις οποίες εμφανίζεται ένα στοιχειοσύνολο:
- Ένα στοιχειοσύνολο **X** ονομάζεται **συχνό**
  - αν η τιμή **s(X)** είναι μεγαλύτερη από μια τιμή κατωφλιού **minsup**
  - η τιμή είναι καθοριζόμενη από το χρήστη

# Κανόνας Συσχέτισης (Association Rule)

- Ένας **κανόνας συσχέτισης** είναι μια πρόταση συνεπαγώγης της μορφής  $X \rightarrow Y$ , όπου  $X \cap Y = \emptyset$
- Η ισχύς ενός κανόνα συσχέτισης μπορεί να μετρηθεί με βάση την **υποστήριξη (support)** και την **εμπιστοσύνη (confidence)**
  - **Υποστήριξη**: πόσο συχνά είναι εφαρμόσιμος ένας κανόνας σε ένα σύνολο δεδομένων
  - **Εμπιστοσύνη**: πόσο συχνά τα αντικείμενα του  $Y$  εμφανίζονται σε συναλλαγές που περιέχουν το  $X$

$$\text{Support, } s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

$$\text{Confidence, } c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

# Παράδειγμα

<i>TID</i>	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Κανόνας: {Milk, Diapers} → {Beer}

- Support = ???
- Confidence = ???

# Παράδειγμα

<i>TID</i>	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{ <u>Milk</u> , Diapers, Beer, Cola}
4	{Bread, <u>Milk</u> , <u>Diapers</u> , <u>Beer</u> }
5	{Bread, Milk, Diapers, Cola}

Κανόνας: {Milk, Diapers} → {Beer}

- Support =  $2/5 = 0.4$
- Confidence = ???

# Παράδειγμα

<i>TID</i>	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Κανόνας: {Milk, Diapers} → {Beer}

- Support =  $2/5 = 0.4$
- Confidence =  $2/3 = 0.67$

# Σχόλια – Παρατηρήσεις

- Ένας κανόνας που έχει πολύ χαμηλή υποστήριξη ίσως να εμφανίζεται **τυχαία**
  - Η υποστήριξη χρησιμοποιείται για να εξαλείψει αδιάφορους κανόνες
- Η **εμπιστοσύνη** μετράει την **αξιοπιστία** του συμπεράσματος
  - όσο πιο μεγάλη είναι η εμπιστοσύνη ενός κανόνα  $X \rightarrow Y$ ,
  - τόσο πιο πιθανό είναι για το στοιχειοσύνολο  $Y$  να είναι παρόν σε συναλλαγές που περιέχουν το στοιχειοσύνολο  $X$
- Το συμπέρασμα που προκύπτει από έναν κανόνα συσχέτισης δεν υπονοεί **υποχρεωτικά αιτιότητα**
  - Αντίθετα, υποδεικνύει **μια ισχυρή σχέση συνύπαρξης** των  $X, Y$

# Ενδιαφέρον (Interest) ενός Κανόνα

- Το ενδιαφέρον (interest) ενός κανόνα  $X \rightarrow Y$  ορίζεται ως η διαφορά της εμπιστοσύνης του και του κλάσματος των συναλλαγών που περιέχουν το  $Y$ :
  - $\text{Interest}(X \rightarrow Y) = c(X \rightarrow Y) - \sigma(Y) / N$
- Αν το  $X$  δεν επηρεάζει το  $Y$ , τότε θα αναμέναμε ότι το ποσοστό των συναλλαγών που περιλαμβάνουν το  $X$  και περιέχουν επίσης το  $Y$  θα ήταν ακριβώς το ίδιο με το ποσοστό των συναλλαγών που περιέχουν το  $Y$
- Ένας τέτοιος κανόνας συσχέτισης έχει  $\text{interest}=0$
- $\{\text{Diapers}\} \rightarrow \{\text{Beer}\}$  έχει υψηλό interest:
  - Γιατί το ποσοστό των ανθρώπων που ενώ αγοράζουν πάνες αγοράζουν και μπύρα είναι σημαντικά μεγαλύτερο από εκείνων που αγοράζουν μπύρα
- Παράδειγμα κανόνα με αρνητικό interest:  $\{\text{Coke}\} \rightarrow \{\text{Pepsi}\}$

# Παράδειγμα

<i>TID</i>	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

- Κανόνας: {Milk, Diapers} → {Beer}
  - $\text{Interest}(X \rightarrow Y) = c(X \rightarrow Y) - \sigma(Y)/N = ???$

# Παράδειγμα

<i>TID</i>	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, <u>Beer</u> , Cola}
4	{Bread, Milk, Diapers, <u>Beer</u> }
5	{Bread, Milk, <u>Diapers</u> , <u>Cola</u> }

- Κανόνας: {Milk, Diapers} → {Beer}
  - $\text{Interest}(X \rightarrow Y) = 2/3 - \sigma(Y)/N = ???$

# Παράδειγμα

<i>TID</i>	Items
1	{Bread, Milk}
2	{Bread, Diapers, <u>Beer</u> , Eggs}
3	{Milk, Diapers, <u>Beer</u> , Cola}
4	{Bread, Milk, Diapers, <u>Beer</u> }
5	{Bread, Milk, Diapers, Cola}

- Κανόνας: {Milk, Diapers} → {Beer}
  - $\text{Interest}(X \rightarrow Y) = 2/3 - 3/5 = 1/15$

# Το Πρόβλημα Εξόρυξης Κανόνων Συσχέτισης

Δοθέντος ενός συνόλου συναλλαγών  $T$ , να βρεθούν **όλοι οι κανόνες** με **υποστήριξη**  $\geq \text{minsup}$  και **εμπιστοσύνη**  $\geq \text{minconf}$ , όπου  $\text{minsup}$  και  $\text{minconf}$  είναι οι αντίστοιχες τιμές κατωφλιού της υποστήριξης και της εμπιστοσύνης

# Το Πρόβλημα Εξόρυξης Κανόνων Συσχέτισης

Δοθέντος ενός συνόλου συναλλαγών  $T$ , να βρεθούν **όλοι οι κανόνες** με **υποστήριξη**  $\geq \text{minsup}$  και **εμπιστοσύνη**  $\geq \text{minconf}$ , όπου  $\text{minsup}$  και  $\text{minconf}$  είναι οι αντίστοιχες τιμές κατωφλιού της υποστήριξης και της εμπιστοσύνης

- Προσέγγιση ωμής βίας (brute-force approach):
  - Υπολογίζει υποστήριξη και εμπιστοσύνη για **όλους** τους πιθανούς κανόνες
  - Όμως, το πλήθος κανόνων **R** είναι εκθετικά μεγάλο(!):
  - **$R = 3^d - 2^{d+1} + 1$**  όπου **d** το πλήθος αντικειμένων

# Το Πρόβλημα Εξόρυξης Κανόνων Συσχέτισης

Δοθέντος ενός συνόλου συναλλαγών  $T$ , να βρεθούν όλοι οι κανόνες με υποστήριξη  $\geq \text{minsup}$  και εμπιστοσύνη  $\geq \text{minconf}$ , όπου  $\text{minsup}$  και  $\text{minconf}$  είναι οι αντίστοιχες τιμές κατωφλιού της υποστήριξης και της εμπιστοσύνης

- Προσέγγιση ωμής βίας (brute-force approach):
  - Υπολογίζει υποστήριξη και εμπιστοσύνη για όλους τους πιθανούς κανόνες
  - Όμως, το πλήθος κανόνων  $R$  είναι εκθετικά μεγάλο(!):
  - $R = 3^d - 2^{d+1} + 1$  όπου  $d$  το πλήθος αντικειμένων

<i>TID</i>	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

$$R = 3^6 - 2^7 + 1 = 602 \text{ πιθανοί κανόνες}$$

# Απόδοση Εξόρυξης Κανόνων Συσχέτισης

- Για τη **βελτίωση της απόδοσης**, ένα βήμα είναι η **αποσύνδεση των απαιτήσεων υποστήριξης και εμπιστοσύνης**
- Η υποστήριξη ενός κανόνα  $X \rightarrow Y$  είναι **ίδια** με την υποστήριξη του αντίστοιχου στοιχειοσυνόλου  $X \cup Y$
- Για παράδειγμα, οι ακόλουθοι κανόνες έχουν την **ίδια υποστήριξη** αφού εμπεριέχουν αντικείμενα από το ίδιο στοιχειοσύνολο:

$\{\text{Beer, Diapers}\} \rightarrow \{\text{Milk}\}$ ,     $\{\text{Beer, Milk}\} \rightarrow \{\text{Diapers}\}$ ,  
 $\{\text{Diapers, Milk}\} \rightarrow \{\text{Beer}\}$ ,     $\{\text{Beer}\} \rightarrow \{\text{Diapers, Milk}\}$ ,  
 $\{\text{Milk}\} \rightarrow \{\text{Beer, Diapers}\}$ ,     $\{\text{Diapers}\} \rightarrow \{\text{Beer, Milk}\}$ .

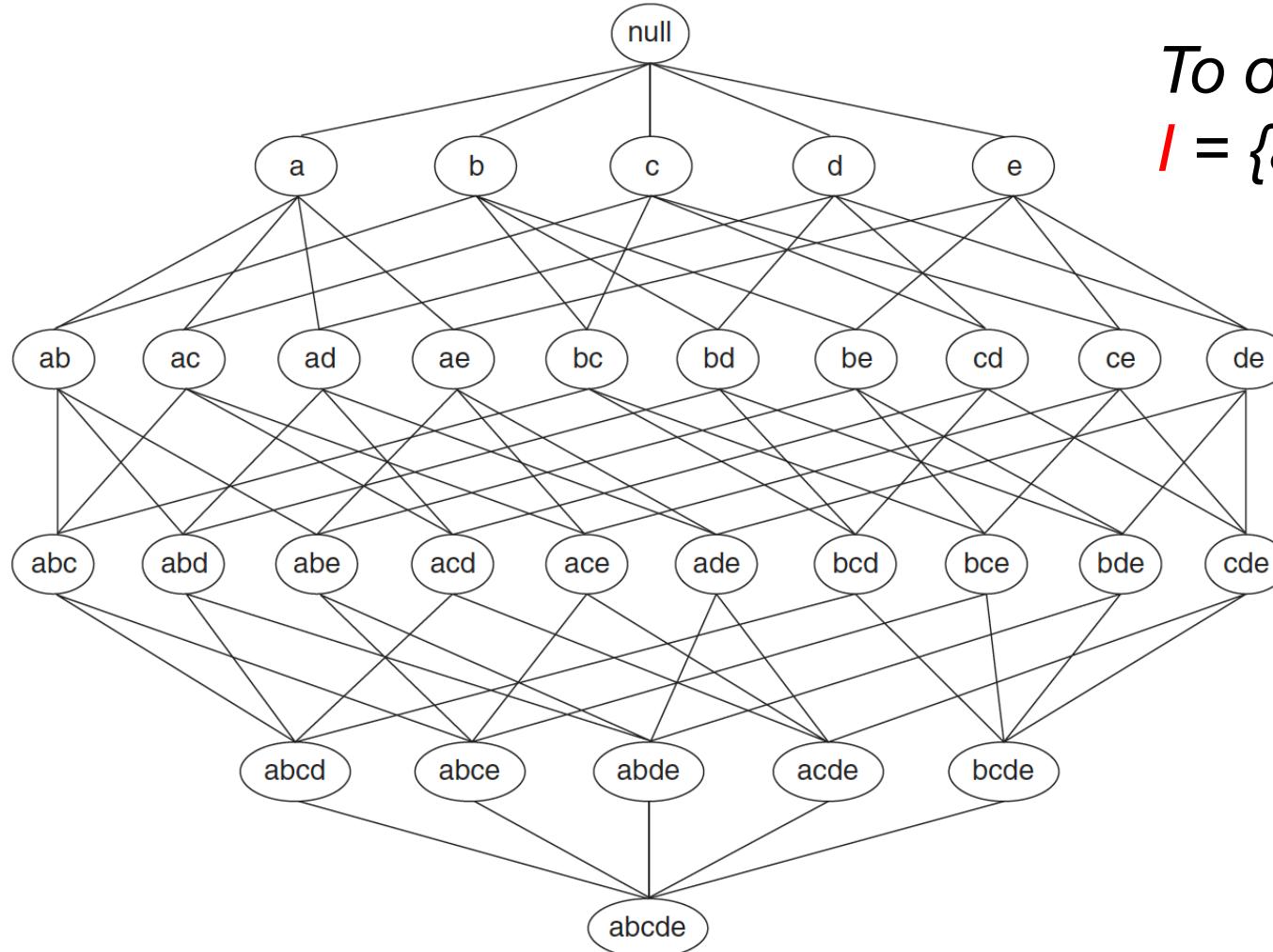
# Απόδοση Εξόρυξης Κανόνων Συσχέτισης

- Μια κοινή στρατηγική πολλών αλγορίθμων εξόρυξης κανόνων συσχέτισης είναι ο διαχωρισμός του προβλήματος σε δύο βασικές επιμέρους εργασίες:
  - **Παραγωγή συχνών στοιχειοσυνόλων**
    - Να βρεθούν όλα τα στοιχεισύνολα που ικανοποιούν το κατώφλι *minsup*
  - **Παραγωγή κανόνων**
    - Να εξαχθούν όλοι οι κανόνες **υψηλής εμπιστοσύνης** από τα συχνά στοιχειοσύνολα

# Περίγραμμα Μαθήματος

- Βασικές έννοιες
- Παραγωγή συχνών στοιχειοσυνόλων
  - Αλγόριθμος Apriori
- Παραγωγή κανόνων
- Σύντομη αναπαράσταση συχνών στοιχειοσυνόλων

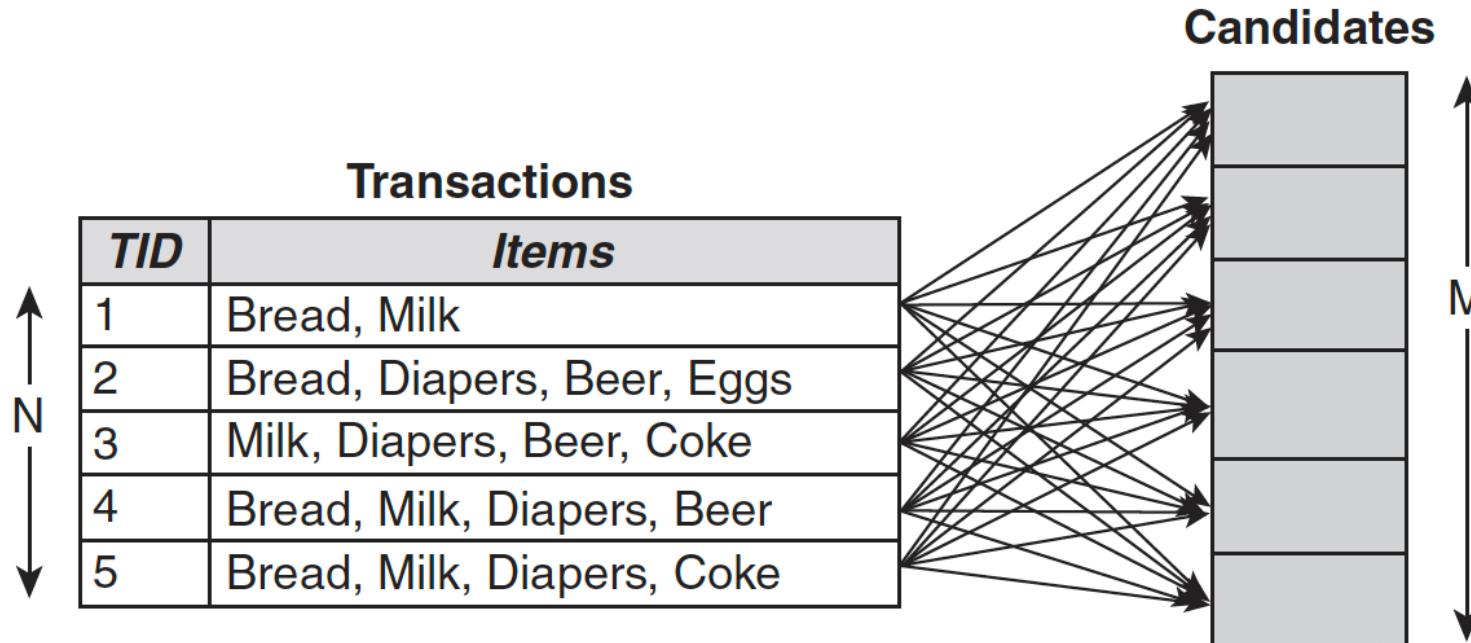
# Δομή Πλέγματος για Απαρίθμηση Στοιχειοσυνόλων



Το σύνολο στοιχείων:  
 $I = \{a, b, c, d, e\}$

Για σύνολο  $k$  στοιχείων,  
μπορούν να παραχθούν  
ως και  $2^k - 1$  συχνά  
στοιχειοσύνολα  
εξαιρουμένου του κενού

# Προσέγγιση Ωμής Βίας (Brute-force Approach)



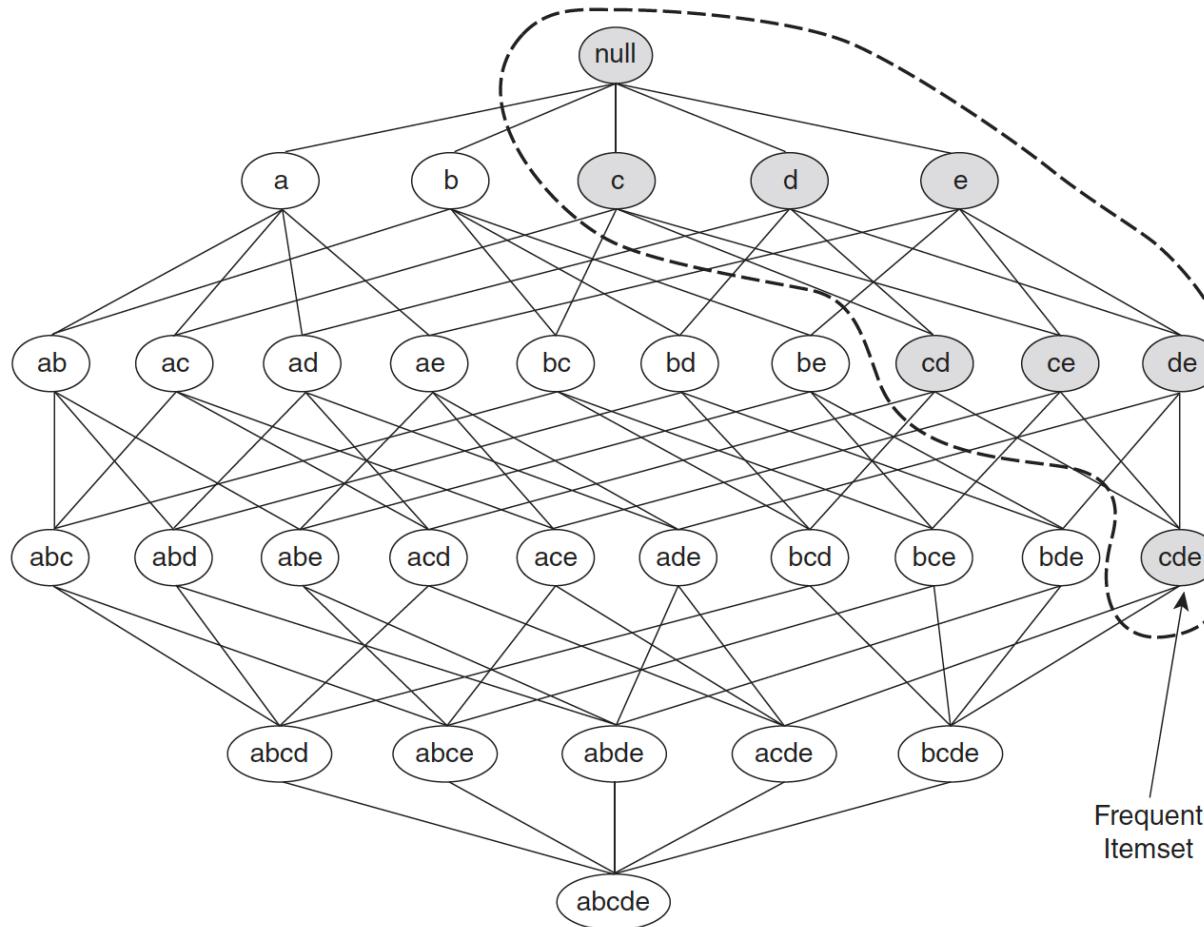
- Μέτρηση υποστήριξης για κάθε υποψήφιο στοιχειοσύνολο
- Συγκρίνεται κάθε υποψήφιο στοιχειοσύνολο με κάθε συναλλαγή, και αν περιλαμβάνεται σε αυτή, η μέτρηση υποστήριξης θα αυξηθεί
- Παράδειγμα: {Bread, Milk} αυξάνεται 3 φορές (συναλλαγές: 1, 4, 5)
- Πολυπλοκότητα:  $O(N M w)$ , όπου  $M=2^k - 1$  και  $w$  το μέγιστο πλάτος συναλλαγής

# Τρόποι Μείωσης Υπολογιστικής Πολυπλοκότητας

- **Μείωση πλήθους υποψήφιων στοιχειοσυνόλων (M)**
  - Η **εκ των προτέρων (Apriori) αρχή** εξαλείφει ορισμένα υποψήφια στοιχειοσύνολα χωρίς να υπολογίσει την τιμή υποστήριξής τους
- **Μείωση του πλήθους των συγκρίσεων**
  - Αξιοποιώντας πιο προηγμένες δομές δεδομένων είτε για την αποθήκευση των στοιχειοσυνόλων είτε για τη συμπίεση του συνόλου δεδομένων
- **Μείωση του πλήθους των συναλλαγών (N)**
  - Για παράδειγμα, αφαίρεση των συναλλαγών πλάτους 2, πριν αναζητήσουμε συχνά στοιχειοσύνολα μεγέθους 3 ή μεγαλύτερα

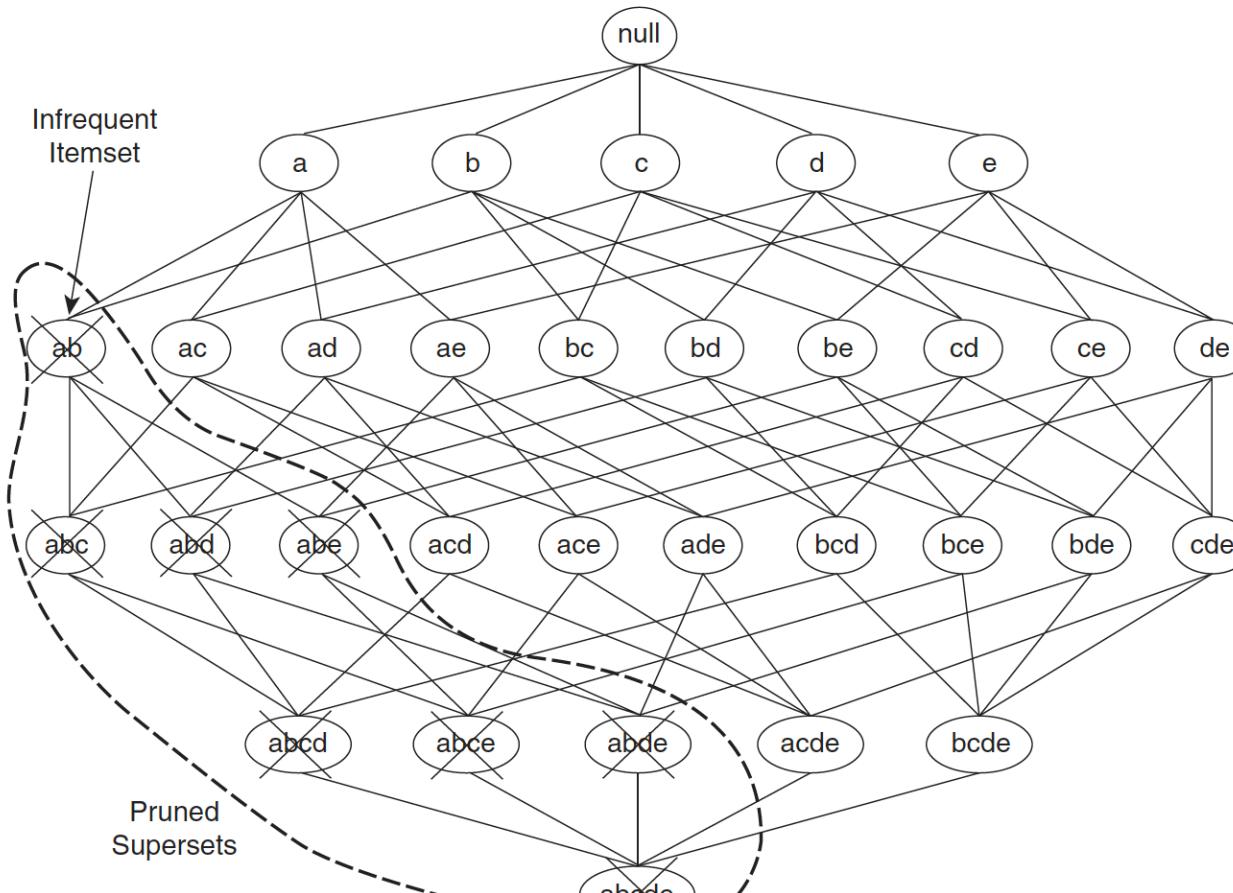
# Η Εκ των Προτέρων (Apriori) Αλγή

Αν ένα στοιχειοσύνολο είναι συχνό, τότε όλα τα υποσύνολά του πρέπει επίσης να είναι συχνά.



# Η Εκ των Προτέρων (Apriori) Αρχή

Κλάδεμα βάσει υποστήριξης (support-based pruning):  
επιτυγχάνει περικοπή του εκθετικού χώρου αναζήτησης



# Ο Αλγόριθμος Apriori

Candidate  
1-Itemsets

Item	Count
Beer	3
Bread	4
Cola	2
Diapers	4
Milk	4
Eggs	1

Minimum support count = 3

Candidate  
2-Itemsets

Itemset	Count
{Beer, Bread}	2
{Beer, Diapers}	3
{Beer, Milk}	2
{Bread, Diapers}	3
{Bread, Milk}	3
{Diapers, Milk}	3

Itemsets removed  
because of low  
support

Candidate  
3-Itemsets

Itemset	Count
{Bread, Diapers, Milk}	2

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

# Ο Αλγόριθμος Apriori

Candidate  
1-Itemsets

Item	Count
Beer	3
Bread	4
Cola	2
Diapers	4
Milk	4
Eggs	1

Minimum support count = 3

Candidate  
2-Itemsets

Itemset	Count
{Beer, Bread}	2
{Beer, Diapers}	3
{Beer, Milk}	2
{Bread, Diapers}	3
{Bread, Milk}	3
{Diapers, Milk}	3

Itemsets removed  
because of low  
support

Προσέγγιση ωμής βίας:

$${6 \choose 1} + {6 \choose 2} + {6 \choose 3} = 6 + 15 + 20 = 41$$

Αλγόριθμος Apriori:

$${6 \choose 1} + {4 \choose 2} + 1 = 6 + 6 + 1 = 13$$

Candidate  
3-Itemsets

Itemset	Count
{Bread, Diapers, Milk}	2

# O Αλγόριθμος Apriori

---

**Algorithm 5.1** Frequent itemset generation of the *Apriori* algorithm.

---

```
1:  $k = 1$ .
2:  $F_k = \{ i \mid i \in I \wedge \sigma(\{i\}) \geq N \times \text{minsup} \}$ . {Find all frequent 1-itemsets}
3: repeat
4:    $k = k + 1$ .
5:    $C_k = \text{candidate-gen}(F_{k-1})$ . {Generate candidate itemsets.}
6:    $C_k = \text{candidate-prune}(C_k, F_{k-1})$ . {Prune candidate itemsets.}
7:   for each transaction  $t \in T$  do
8:      $C_t = \text{subset}(C_k, t)$ . {Identify all candidates that belong to  $t$ .}
9:     for each candidate itemset  $c \in C_t$  do
10:        $\sigma(c) = \sigma(c) + 1$ . {Increment support count.}
11:     end for
12:   end for
13:    $F_k = \{ c \mid c \in C_k \wedge \sigma(c) \geq N \times \text{minsup} \}$ . {Extract the frequent  $k$ -itemsets.}
14: until  $F_k = \emptyset$ 
15: Result =  $\bigcup F_k$ .
```

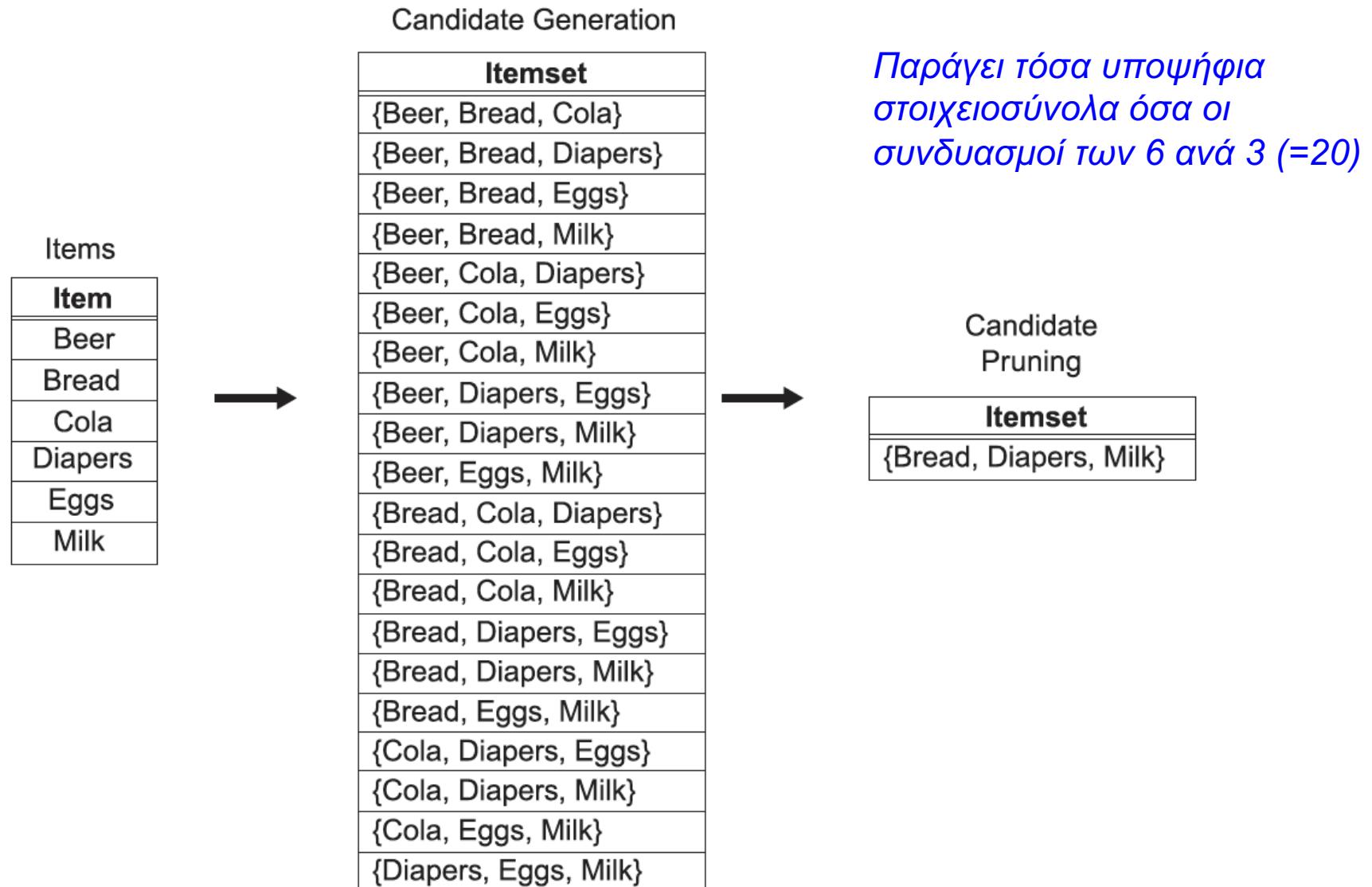
---

Rakesh Agrawal, Ramakrishnan Srikant: Fast Algorithms for Mining Association Rules in Large Databases. VLDB 1994: 487-499

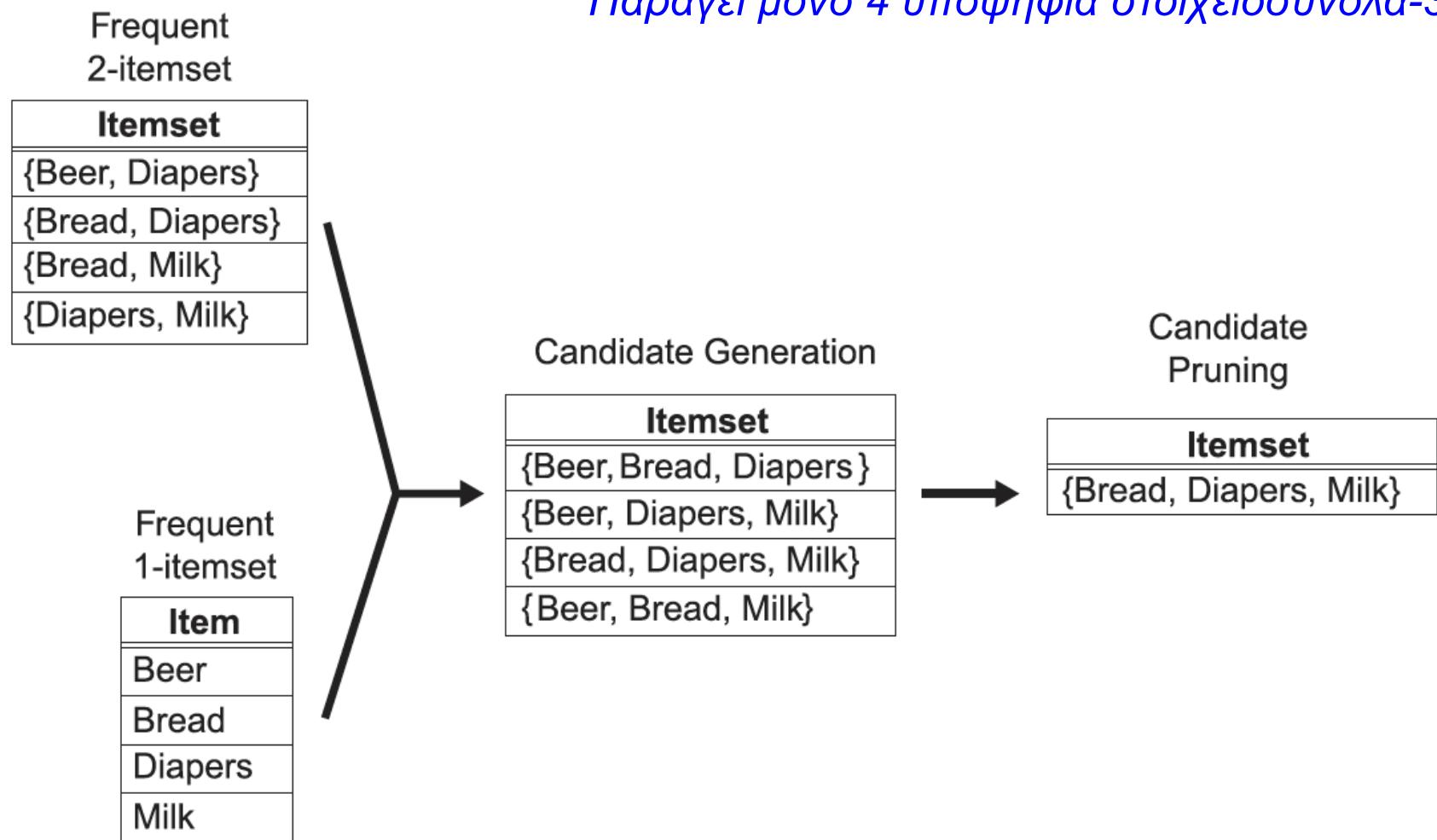
# Παραγωγήσεις

- Το τμήμα παραγωγής συχνών στοιχειοσυνόλων του αλγορίθμου Apriori έχει δύο σημαντικά χαρακτηριστικά
  - Διασχίζει το πλέγμα **επίπεδο-επίπεδο**, από τα συχνά στοιχειοσύνολα-1 προς το μέγιστο μέγεθος συχνών στοιχειοσυνόλων
  - Χρησιμοποιεί μια στρατηγική **παραγωγής και ελέγχου** για να βρει τα συχνά στοιχειοσύνολα
    - Σε κάθε επανάληψη **παράγονται νέα υποψήφια στοιχειοσύνολα** από αυτά της προηγούμενης επανάληψης
    - Στη συνέχεια, **μετριέται η υποστήριξη** ως προς το κατώφλι

# Παραγωγή Υποψηφίων – Μέθοδος Ωμής Βίας

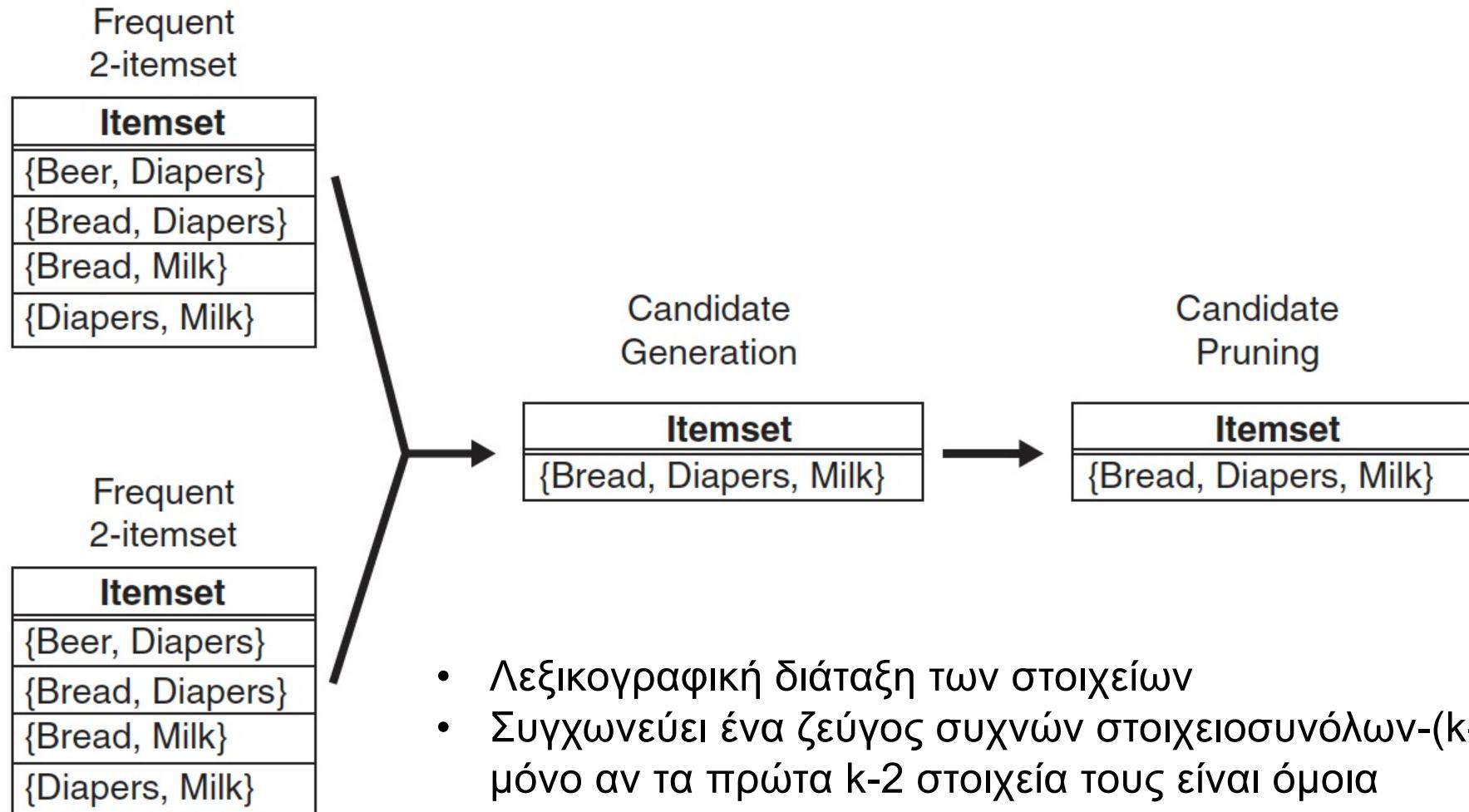


# Παραγωγή Υποψηφίων – Μέθοδος F<sub>k-1</sub>xF<sub>1</sub>



# Παραγωγή Υποψηφίων – Μέθοδος $F_{k-1} \times F_{k-1}$

Χρησιμοποιείται από τον *Apriori*



# Κλάδεμα Υποψηφίων (Prune Candidates)

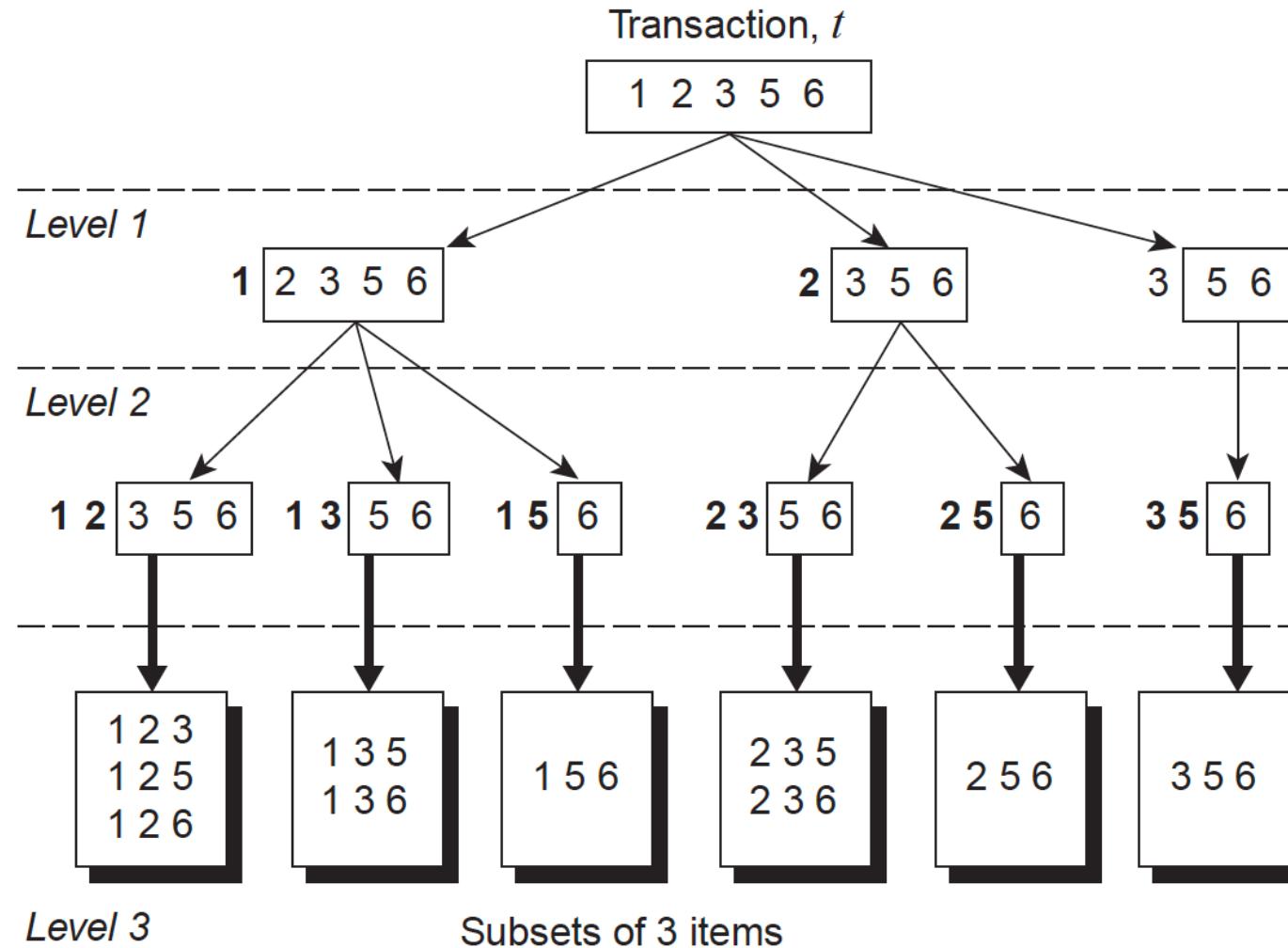
- Έστω  $X = \{i_1, i_2, \dots, i_k\}$  και ας θεωρήσουμε τα  $k$  γνήσια υποσύνολά του  $X - \{i_j\}$ , για  $j=1\dots k$
- *Εάν έστω ένα από αυτά είναι σπάνιο, τότε το  $X$  κλαδεύεται βάσει της αρχής Apriori*
- Για κάθε υποψήφιο στοιχειοσύνολο- $k$ :
  - Μέθοδος ωμής βίας: ελέγχει  $k$  υποσύνολα μεγέθους  $k-1$
  - Μέθοδος  $F_{k-1} \times F_1$ : ελέγχει  $k-1$  υποσύνολα μεγέθους  $k-1$
  - Μέθοδος  $F_{k-1} \times F_{k-1}$ : ελέγχει  $k-2$  υποσύνολα μεγέθους  $k-1$

# Μέτρηση Υποστήριξης

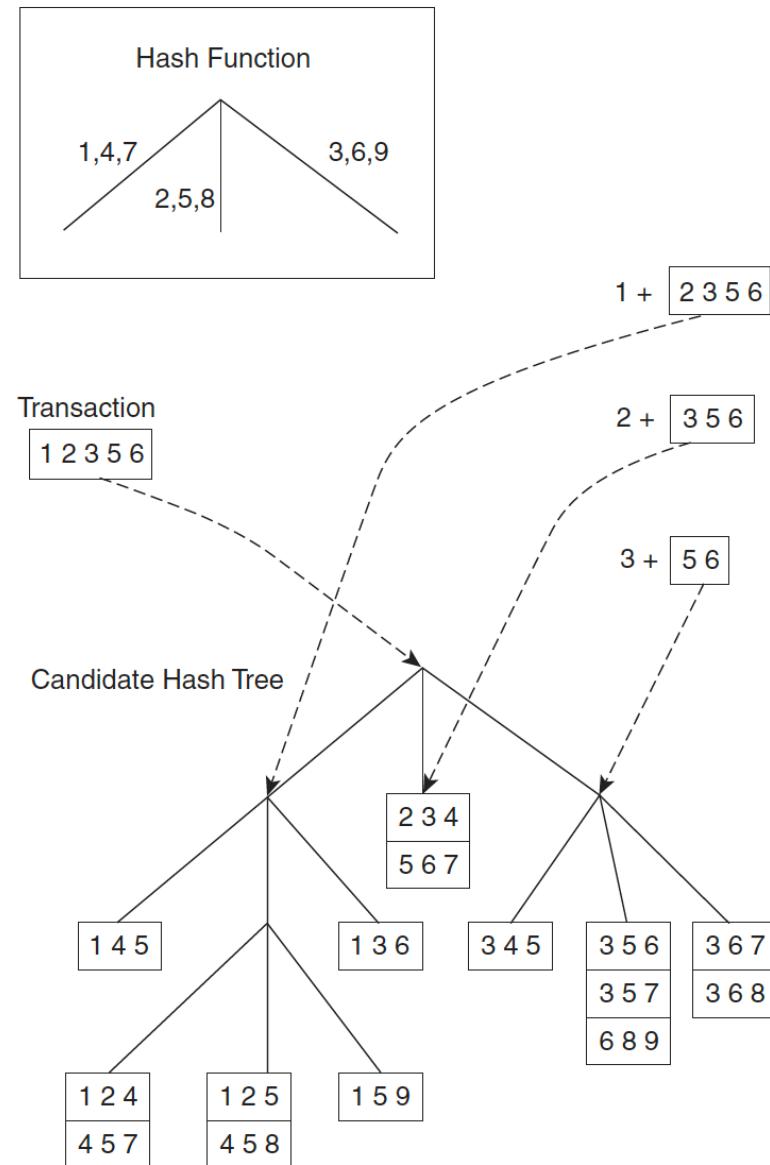
**Algorithm 5.1** Frequent itemset generation of the *Apriori* algorithm.

```
1:  $k = 1$ .
2:  $F_k = \{ i \mid i \in I \wedge \sigma(\{i\}) \geq N \times \text{minsup} \}$ . {Find all frequent 1-itemsets}
3: repeat
4:    $k = k + 1$ .
5:    $C_k = \text{candidate-gen}(F_{k-1})$ . {Generate candidate itemsets.}
6:    $C_k = \text{candidate-prune}(C_k, F_{k-1})$ . {Prune candidate itemsets.}
7:   for each transaction  $t \in T$  do
8:      $C_t = \text{subset}(C_k, t)$ . {Identify all candidates that belong to  $t$ .}
9:     for each candidate itemset  $c \in C_t$  do
10:       $\sigma(c) = \sigma(c) + 1$ . {Increment support count.}
11:    end for
12:  end for
13:   $F_k = \{ c \mid c \in C_k \wedge \sigma(c) \geq N \times \text{minsup} \}$ . {Extract the frequent  $k$ -itemsets.}
14: until  $F_k = \emptyset$ 
15: Result =  $\bigcup F_k$ .
```

# Μέτρηση Υποστήριξης – Απαρίθμηση Στοιχειοσυνόλων 3 Στοιχείων



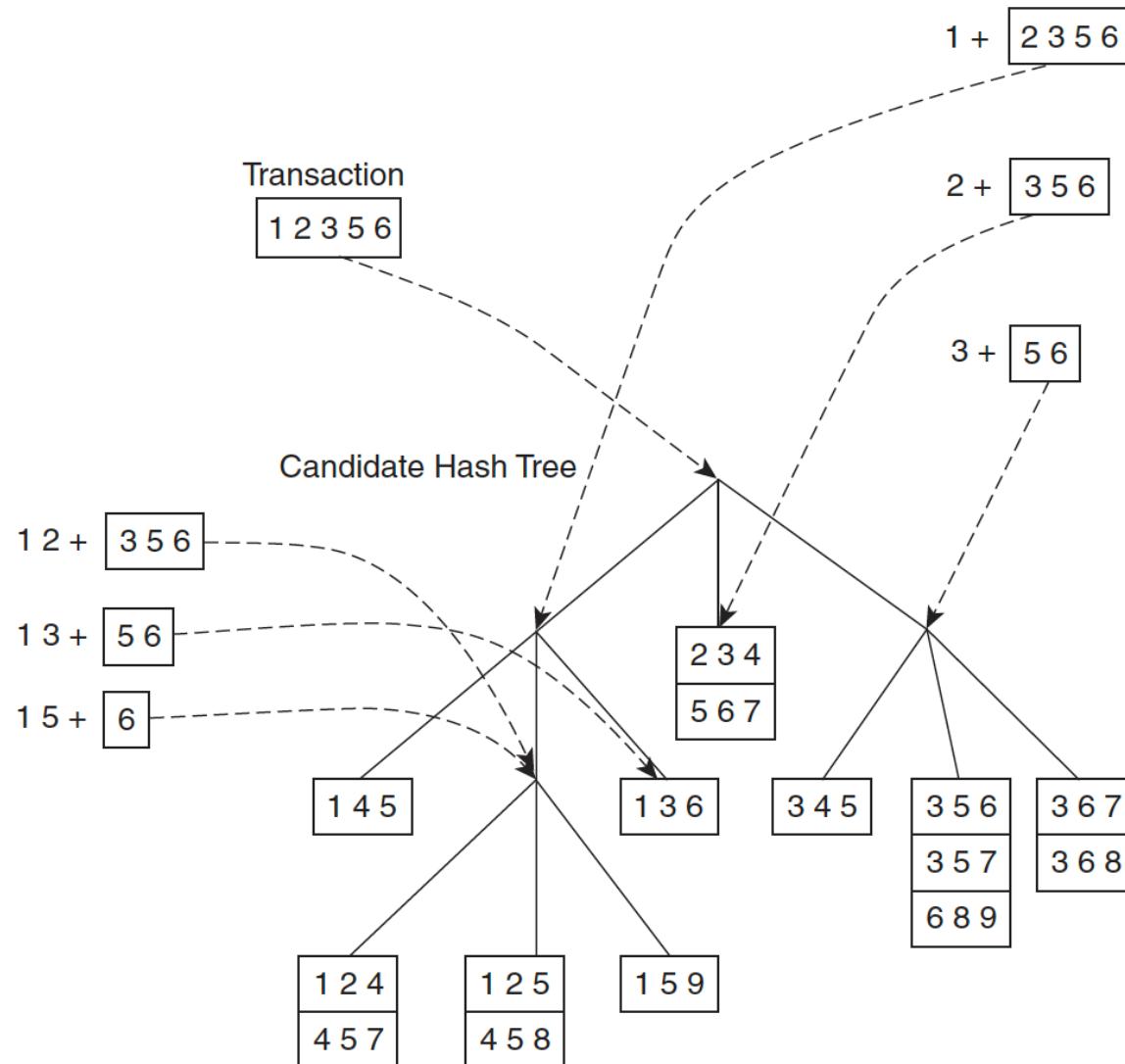
# Μέτρηση Υποστήριξης – Δένδρα Κατακερματισμού



Στον *Apriori*, τα **υποψήφια στοιχειοσύνολα** διαιρούνται σε διαφορετικά τμήματα και αποθηκεύονται σε δέντρα κατακερματισμού

Συνάρτηση κατακερματισμού:  
 $h(p) = p \bmod 3$

# Μέτρηση Υποστήριξης – Δένδρα Κατακερματισμού



- 9 κόμβοι φύλλα
- 15 στοιχειοσύνολα

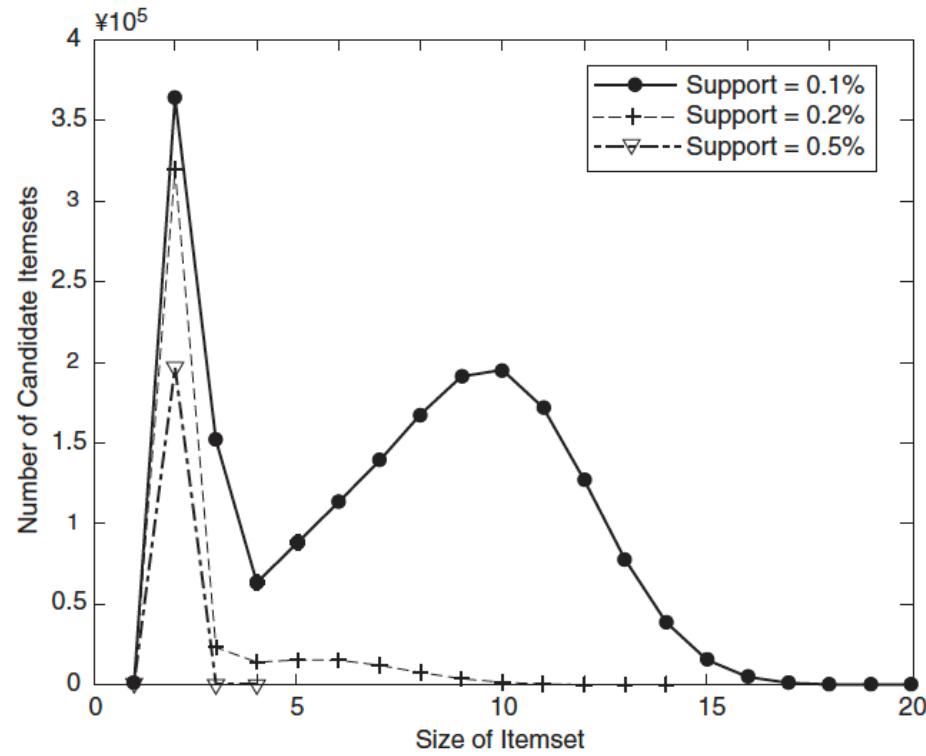
5 από τους 9 κόμβους-φύλλα δέχονται επίσκεψη

9 από τα 15 στοιχειοσύνολα συγκρίνονται με τη συναλλαγή  $t$

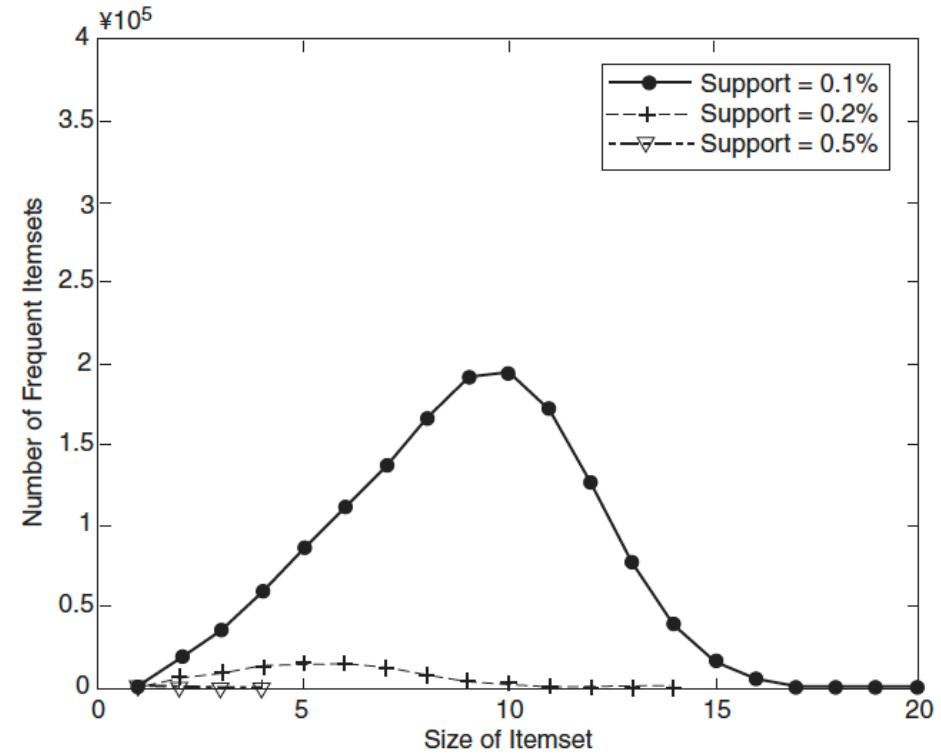
# Πολυπλοκότητα Υπολογισμών

- Η υπολογιστική πολυπλοκότητα του αλγόριθμου Apriori επηρεάζεται από τους ακόλουθους παράγοντες
  - Κατώφλι υποστήριξης  $\text{minsup}$
  - Πλήθος των στοιχείων (διαστάσεις)  $d$
  - Πλήθος των συναλλαγών  $N$
  - Μέσο πλάτος συναλλαγής  $w$

# Επίδραση Κατωφλιού Υποστήριξης

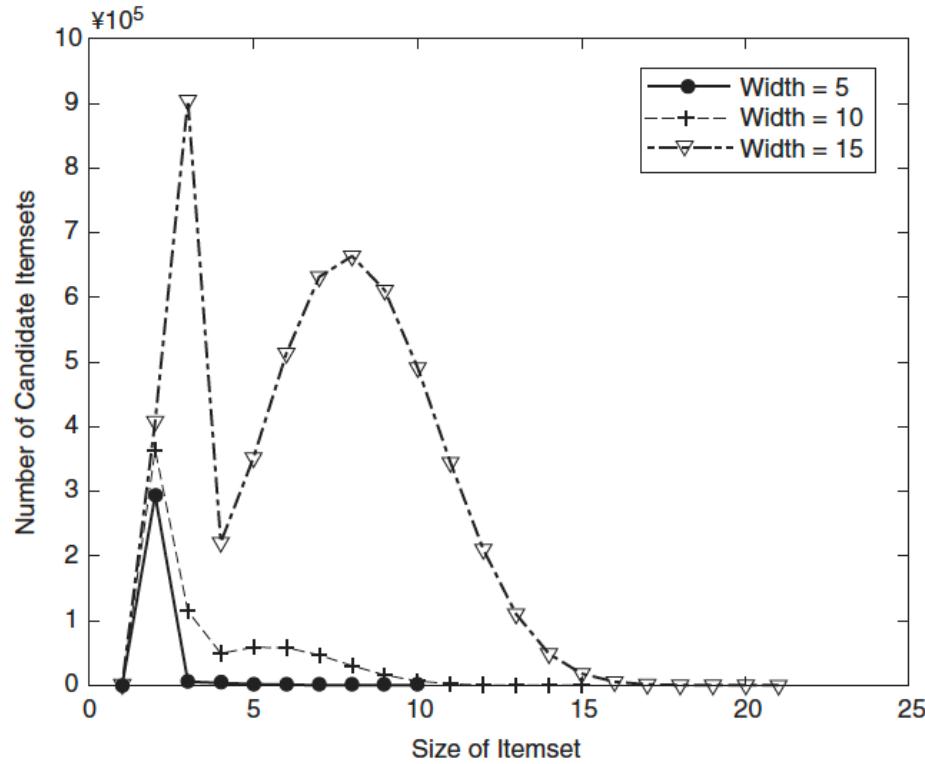


(a) Number of candidate itemsets.

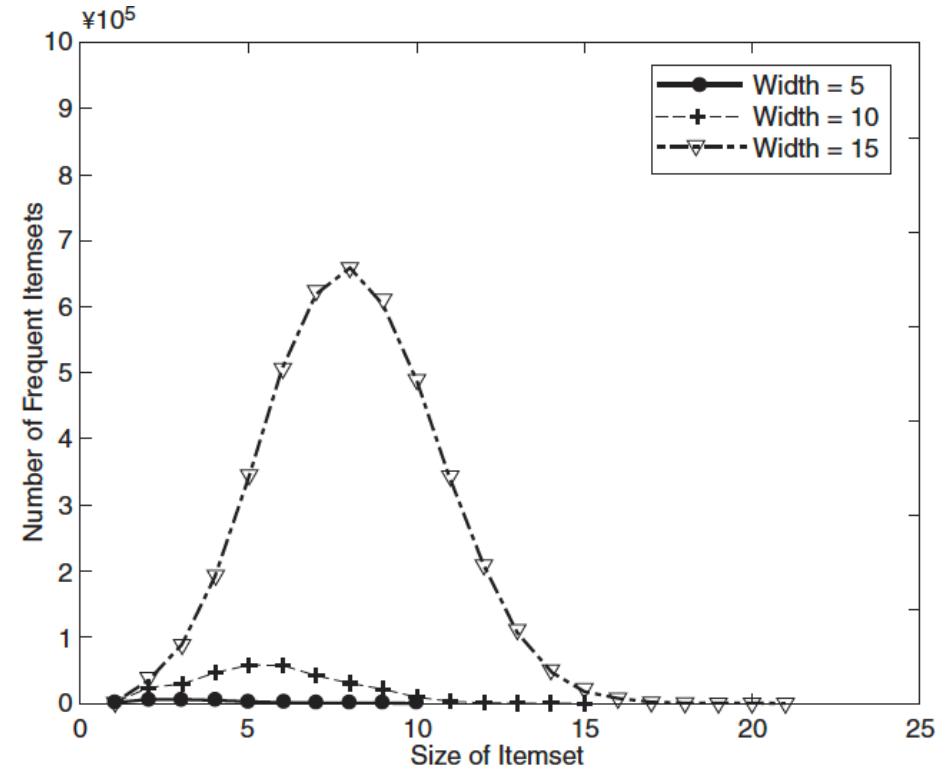


(b) Number of frequent itemsets.

# Επίδραση Μέσου Πλάτους Συναλλαγής



(a) Number of candidate itemsets.



(b) Number of Frequent Itemsets.

# Περίγραμμα Μαθήματος

- Βασικές έννοιες
- Παραγωγή συχνών στοιχειοσυνόλων
  - Αλγόριθμος Apriori
- Παραγωγή κανόνων
- Σύντομη αναπαράσταση συχνών στοιχειοσυνόλων

# Παραγωγή Κανόνων

- *Πώς μπορούμε να παράγουμε κανόνες συσχέτισης με αποτελεσματικό τρόπο από ένα δοθέν συχνό στοιχειοσύνολο Y;*
- Ένας κανόνας συσχέτισης προκύπτει διαχωρίζοντας το Y σε δύο μη κενά υποσύνολα: X και Y – X
- Όλοι οι παραγόμενοι κανόνες με αυτόν τον τρόπο ικανοποιούν το κατώφλι υποστήριξης (αφού το Y είναι συχνό στοιχειοσύνολο)
- Παράδειγμα:
  - $Y = \{a, b, c\}$
  - Υποψήφιοι κανόνες συσχέτισης από το Y:

$\{a, b\} \rightarrow \{c\}$	$\{a, c\} \rightarrow \{b\}$	$\{b, c\} \rightarrow \{a\}$
$\{a\} \rightarrow \{b, c\}$	$\{b\} \rightarrow \{a, c\}$	$\{c\} \rightarrow \{a, b\}$

# Κλάδεμα βάσει Εμπιστοσύνης

$$\text{Confidence, } c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

- Η εμπιστοσύνη ενός κανόνα  $X \rightarrow Y$  μπορεί να είναι μεγαλύτερη, μικρότερη ή ίση με εκείνη του κανόνα  $X' \rightarrow Y'$ , όπου  $X' \subseteq X$  και  $Y' \subseteq Y$
- Όμως ισχύει το ακόλουθο θεώρημα για το μέτρο εμπιστοσύνης:

Έστω  $Y$  ένα στοιχειοσύνολο και  $X$  ένα υποσύνολο του  $Y$  ( $X \subseteq Y$ ). Αν ένας κανόνας  $X \rightarrow Y - X$  δεν ικανοποιεί το κατώφλι εμπιστοσύνης, τότε κάθε κανόνας  $X' \rightarrow Y - X'$ , όπου  $X' \subseteq X$ , πρέπει επίσης να μην ικανοποιεί το κατώφλι εμπιστοσύνης.

# Κλάδεμα βάσει Εμπιστοσύνης

$$\text{Confidence, } c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

- Η εμπιστοσύνη ενός κανόνα  $X \rightarrow Y$  μπορεί να είναι μεγαλύτερη, μικρότερη ή ίση με εκείνη του κανόνα  $X' \rightarrow Y'$ , όπου  $X' \subseteq X$  και  $Y' \subseteq Y$
- Όμως ισχύει το ακόλουθο θεώρημα για το μέτρο εμπιστοσύνης:

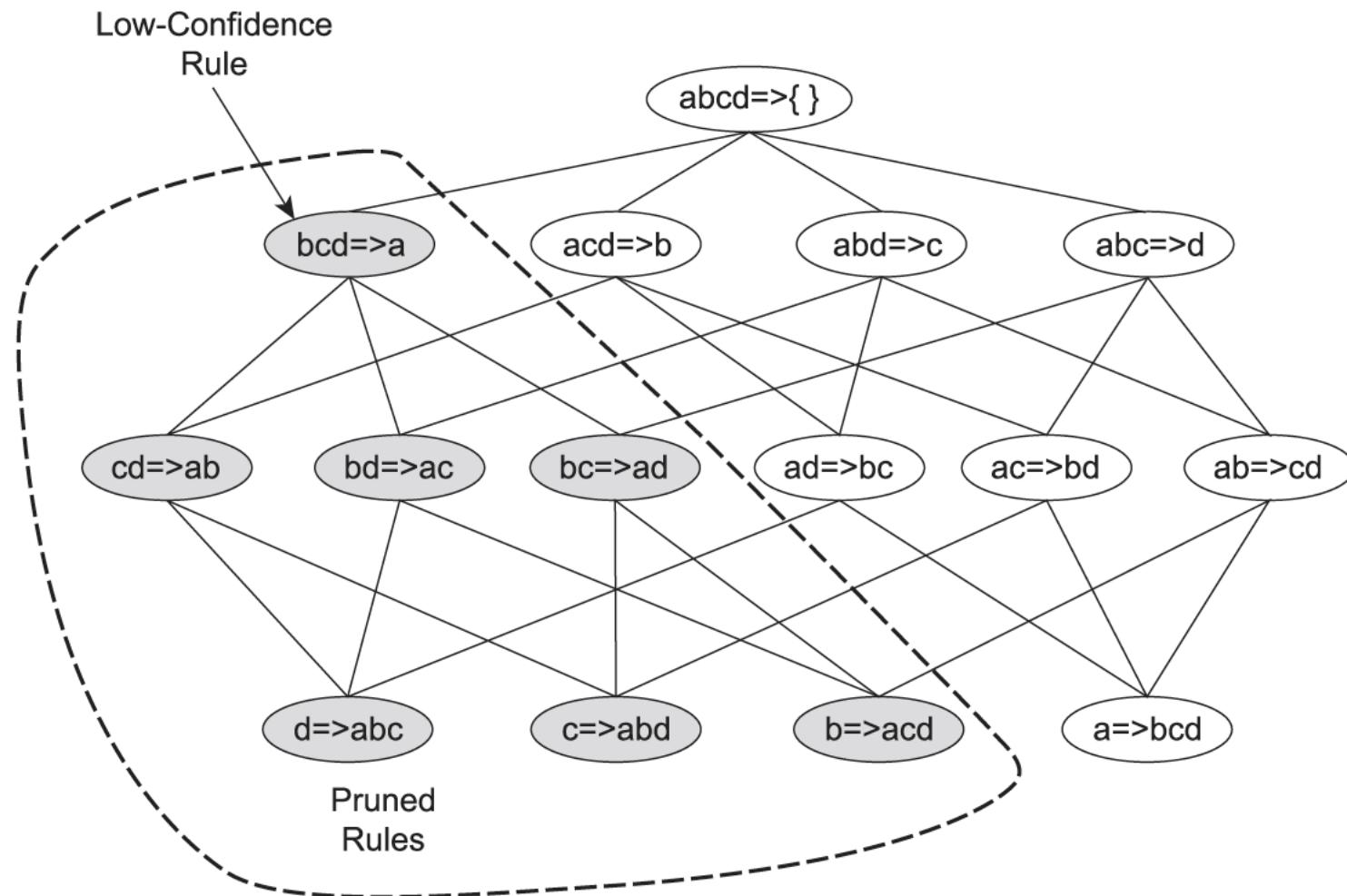
Έστω  $Y$  ένα στοιχειοσύνολο και  $X$  ένα υποσύνολο του  $Y$  ( $X \subseteq Y$ ). Αν ένας κανόνας  $X \rightarrow Y - X$  δεν ικανοποιεί το κατώφλι εμπιστοσύνης, τότε κάθε κανόνας  $X' \rightarrow Y - X'$ , όπου  $X' \subseteq X$ , πρέπει επίσης να μην ικανοποιεί το κατώφλι εμπιστοσύνης.

## Απόδειξη:

- Έστω  $X' \rightarrow Y - X'$  και  $X \rightarrow Y - X$  με  $X' \subseteq X$
- Η εμπιστοσύνη:  $c(X' \rightarrow Y - X') = \sigma(Y)/\sigma(X')$  και  $c(X \rightarrow Y - X) = \sigma(Y)/\sigma(X)$
- Δεδομένου ότι:  $X' \subseteq X$ , άρα  $\sigma(X') \geq \sigma(X)$ , άρα ο κανόνας  $X' \rightarrow Y - X'$  δεν μπορεί να έχει μεγαλύτερη εμπιστοσύνη από τον  $X \rightarrow Y - X$

# Κλάδεμα βάσει Εμπιστοσύνης

$$\text{Confidence, } c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$



# Κλάδεμα βάσει Εμπιστοσύνης

Ομοιότητα του *ap-genrules()* με τη διαδικασία παραγωγής συχνών στοιχειοσυνόλων (Apriori)

---

**Algorithm 5.2** Rule generation of the *Apriori* algorithm.

---

```
1: for each frequent  $k$ -itemset  $f_k$ ,  $k \geq 2$  do
2:    $H_1 = \{i \mid i \in f_k\}$  {1-item consequents of the rule.}
3:   call ap-genrules( $f_k, H_1$ .)
4: end for
```

---

**Algorithm 5.3** Procedure *ap-genrules*( $f_k, H_m$ ).

---

```
1:  $k = |f_k|$  {size of frequent itemset.}
2:  $m = |H_m|$  {size of rule consequent.}
3: if  $k > m + 1$  then
4:    $H_{m+1} = \text{candidate-gen}(H_m)$ .
5:    $H_{m+1} = \text{candidate-prune}(H_{m+1}, H_m)$ .
6:   for each  $h_{m+1} \in H_{m+1}$  do
7:      $\text{conf} = \sigma(f_k)/\sigma(f_k - h_{m+1})$ .
8:     if  $\text{conf} \geq \text{minconf}$  then
9:       output the rule  $(f_k - h_{m+1}) \rightarrow h_{m+1}$ .
10:    else
11:      delete  $h_{m+1}$  from  $H_{m+1}$ .
12:    end if
13:   end for
14:   call ap-genrules( $f_k, H_{m+1}$ .)
15: end if
```

---

# Παράδειγμα: Εγγραφές Ψηφοφορίας του Κογκρέσου

**Table 5.3.** List of binary attributes from the 1984 United States Congressional Voting Records. Source: The UCI machine learning repository.

1. Republican	18. aid to Nicaragua = no	435 συναλλαγές
2. Democrat	19. MX-missile = yes	34 γνωρίσματα
3. handicapped-infants = yes	20. MX-missile = no	
4. handicapped-infants = no	21. immigration = yes	
5. water project cost sharing = yes	22. immigration = no	
6. water project cost sharing = no	23. synfuel corporation cutback = yes	
7. budget-resolution = yes	24. synfuel corporation cutback = no	
8. budget-resolution = no	25. education spending = yes	
9. physician fee freeze = yes	26. education spending = no	
10. physician fee freeze = no	27. right-to-sue = yes	
11. aid to El Salvador = yes	28. right-to-sue = no	
12. aid to El Salvador = no	29. crime = yes	
13. religious groups in schools = yes	30. crime = no	
14. religious groups in schools = no	31. duty-free-exports = yes	
15. anti-satellite test ban = yes	32. duty-free-exports = no	
16. anti-satellite test ban = no	33. export administration act = yes	
17. aid to Nicaragua = yes	34. export administration act = no	

<https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records>

# Παράδειγμα: Εγγραφές Ψηφοφορίας του Κογκρέσου

- Μερικοί κανόνες συσχέτισης που παράγονται από τον Apriori για  $\text{minsup} = 30\%$  και  $\text{minconf} = 90\%:$

Association Rule	Confidence
{budget resolution = no, MX-missile=no, aid to El Salvador = yes } → {Republican}	91.0%
{budget resolution = yes, MX-missile=yes, aid to El Salvador = no } → {Democrat}	97.5%
{crime = yes, right-to-sue = yes, physician fee freeze = yes} → {Republican}	93.5%
{crime = no, right-to-sue = no, physician fee freeze = no} → {Democrat}	100%

# Περίγραμμα Μαθήματος

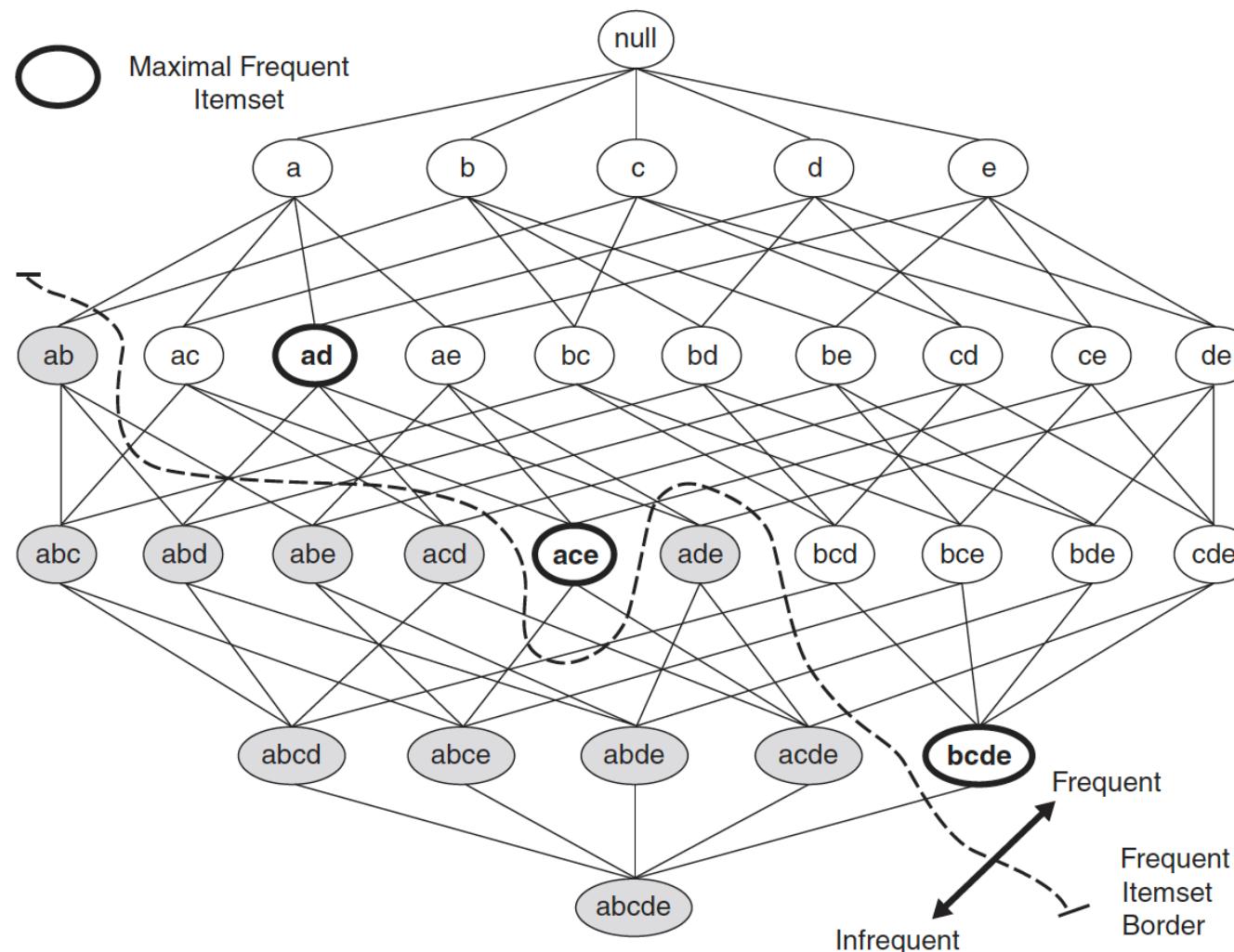
- Βασικές έννοιες
- Παραγωγή συχνών στοιχειοσυνόλων
  - Αλγόριθμος Apriori
- Παραγωγή κανόνων
- Σύντομη αναπαράσταση συχνών στοιχειοσυνόλων

# Σύντομες Αναπαραστάσεις

- Το πλήθος συχνών στοιχειοσυνόλων μπορεί να είναι πολύ μεγάλο
- Είναι χρήσιμο να προσδιοριστεί **ένα μικρό, αντιπροσωπευτικό σύνολο στοιχειοσυνόλων**, από τα οποία μπορούν να προκύψουν όλα τα άλλα συχνά στοιχειοσύνολα
- Δύο αναπαραστάσεις:
  - Μέγιστα συχνά στοιχειοσύνολα
  - Κλειστά συχνά στοιχειοσύνολα

# Μέγιστα Συχνά Στοιχειοσύνολα

Ένα **μέγιστο συχνό στοιχειοσύνολο (maximal frequent itemset)** ορίζεται ως ένα συχνό στοιχειοσύνολο για το οποίο κανένα από τα άμεσα υπερσύνολά του δεν είναι συχνό



Τα  $\{a,d\}$ ,  $\{a,c,e\}$  και  $\{b,c,d,e\}$  είναι **μέγιστα συχνά στοιχειοσύνολα**, διότι τα άμεσα υπερσύνολά τους είναι σπάνια

Αντίθετα, το  $\{a,c\}$  είναι μη μέγιστο, επειδή ένα από τα άμεσα υπερσύνολά του, το  $\{a,c,e\}$  είναι συχνό

# Κλειστά Στοιχειοσύνολα

- Παρότι τα μέγιστα συχνά στοιχειοσύνολα παρέχουν μια σύντομη αναπαράσταση, **δεν περιέχουν πληροφορίες υποστήριξης των υποσυνόλων τους**
  - Εκτός από το ότι ικανοποιείται μία τιμή κατωφλιού υποστήριξης
- Άρα απαιτείται ένα επιπλέον πέρασμα στο σύνολο δεδομένων για να καθοριστούν οι μετρήσιες υποστήριξης των μη-μέγιστων συχνών στοιχειοσυνόλων
- Επομένως, είναι χρήσιμη μια ελάχιστη αναπαράσταση που **διατηρεί τις πληροφορίες υποστήριξης**

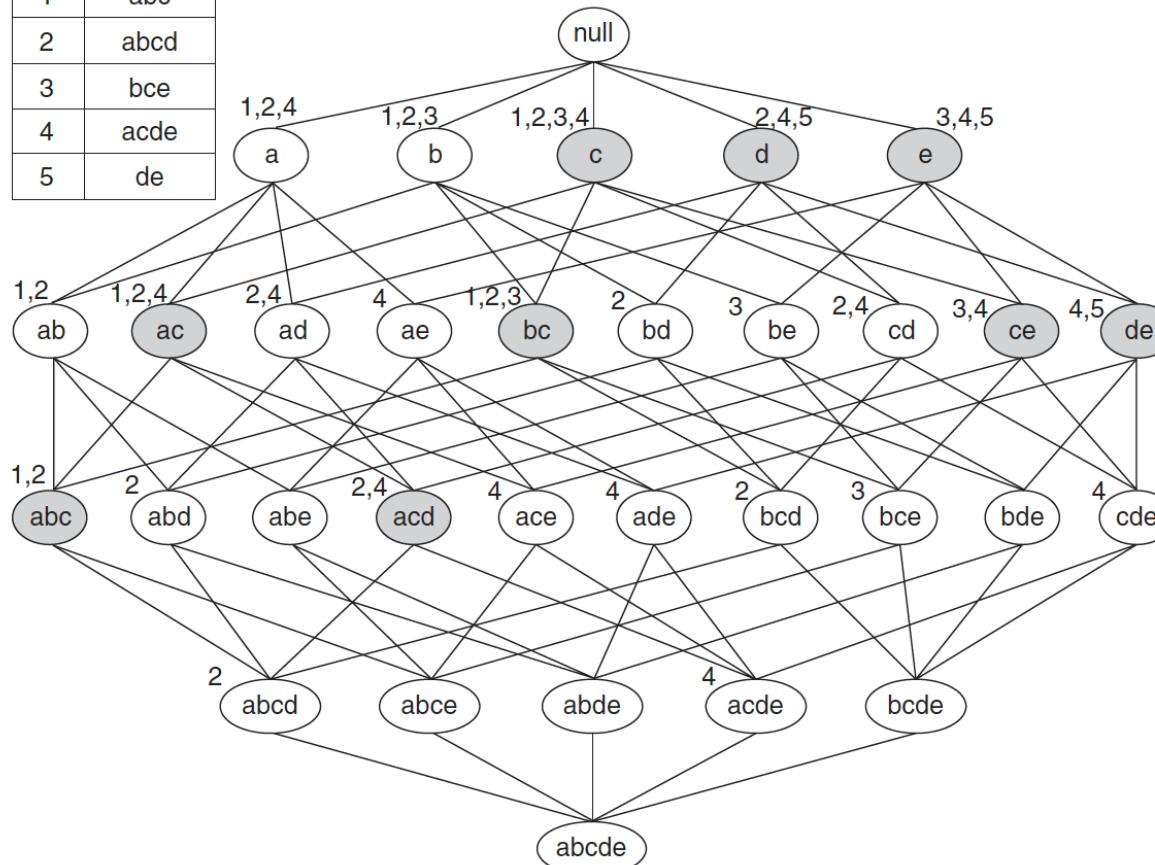
Ένα στοιχειοσύνολο  $X$  είναι **κλειστό** (*closed itemset*) αν κανένα από τα άμεσα υπερσύνολά του δεν έχει την ίδια μέτρηση υποστήριξης με το  $X$ .

Ισοδύναμα, το  $X$  **δεν είναι κλειστό** αν τουλάχιστον ένα από τα άμεσα υπερσύνολά του έχει την ίδια μέτρηση υποστήριξης με το  $X$ .

# Κλειστά Στοιχειοσύνολα

Ένα στοιχειοσύνολο  $X$  είναι **κλειστό (closed itemset)** αν κανένα από τα άμεσα υπερσύνολά του δεν έχει την ίδια μέτρηση υποστήριξης με το  $X$ .

TID	Items
1	abc
2	abcd
3	bce
4	acde
5	de



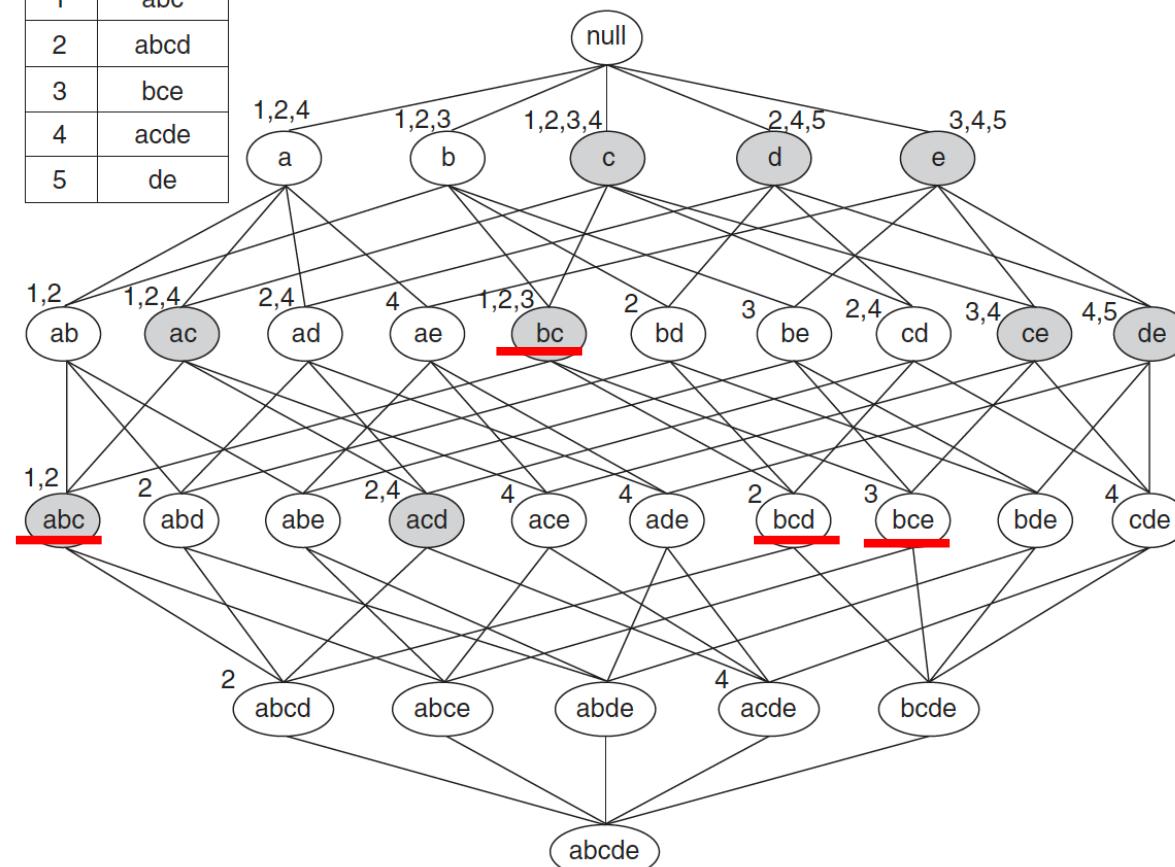
{b,c}: κλειστό  
γιατί;

{a,d}: μη κλειστό  
γιατί;

# Κλειστά Στοιχειοσύνολα

Ένα στοιχειοσύνολο  $X$  είναι **κλειστό (closed itemset)** αν κανένα από τα άμεσα υπερσύνολά του δεν έχει την ίδια μέτρηση υποστήριξης με το  $X$ .

TID	Items
1	abc
2	abcd
3	bce
4	acde
5	de



{b,c}: κλειστό  
διότι κανένα από τα  
{a,b,c}, {b,c,d}, {b,c,e}  
δεν έχει την ίδια  
μέτρηση υποστήριξης  
με αυτό

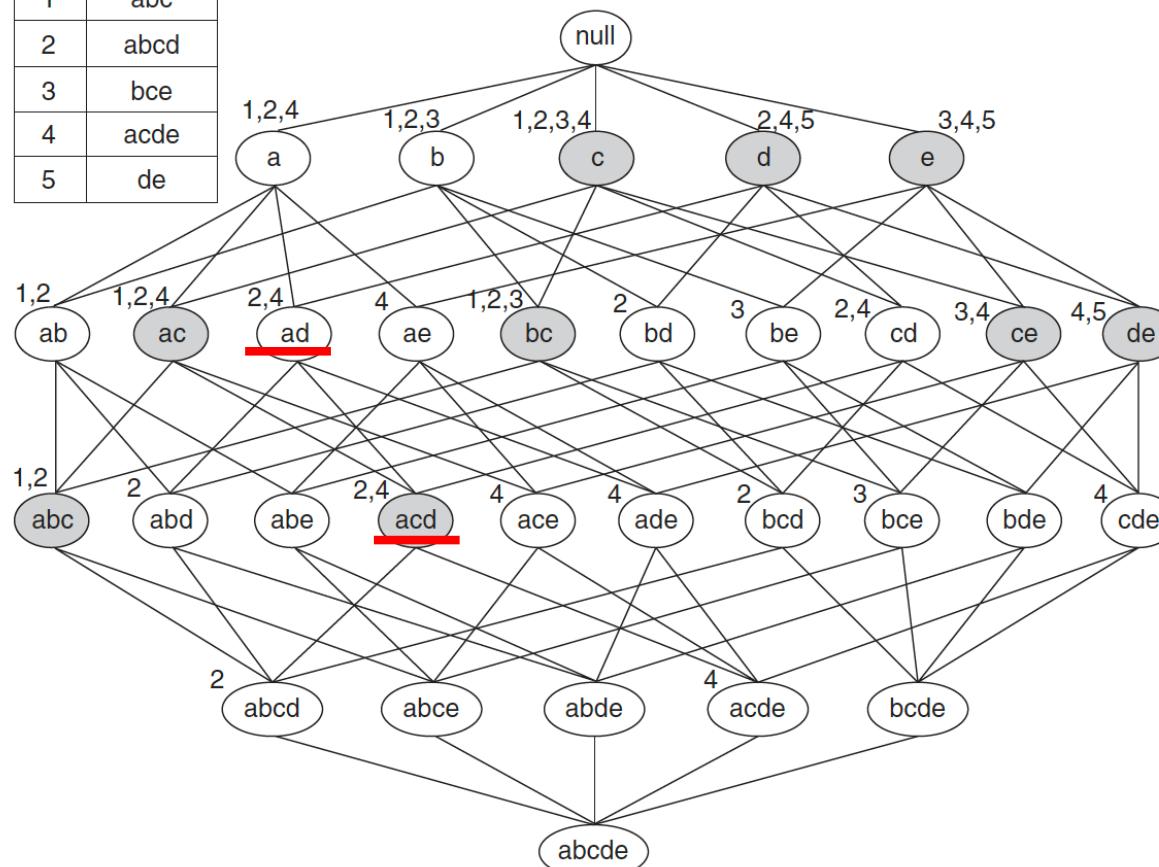
{a,d}: μη κλειστό  
γιατί:

# Κλειστά Στοιχειοσύνολα

Ένα στοιχειοσύνολο  $X$  είναι **κλειστό (closed itemset)** αν κανένα από τα άμεσα υπερσύνολά του δεν έχει την ίδια μέτρηση υποστήριξης με το  $X$ .

TID	Items
1	abc
2	abcd
3	bce
4	acde
5	de

minsup = 40%



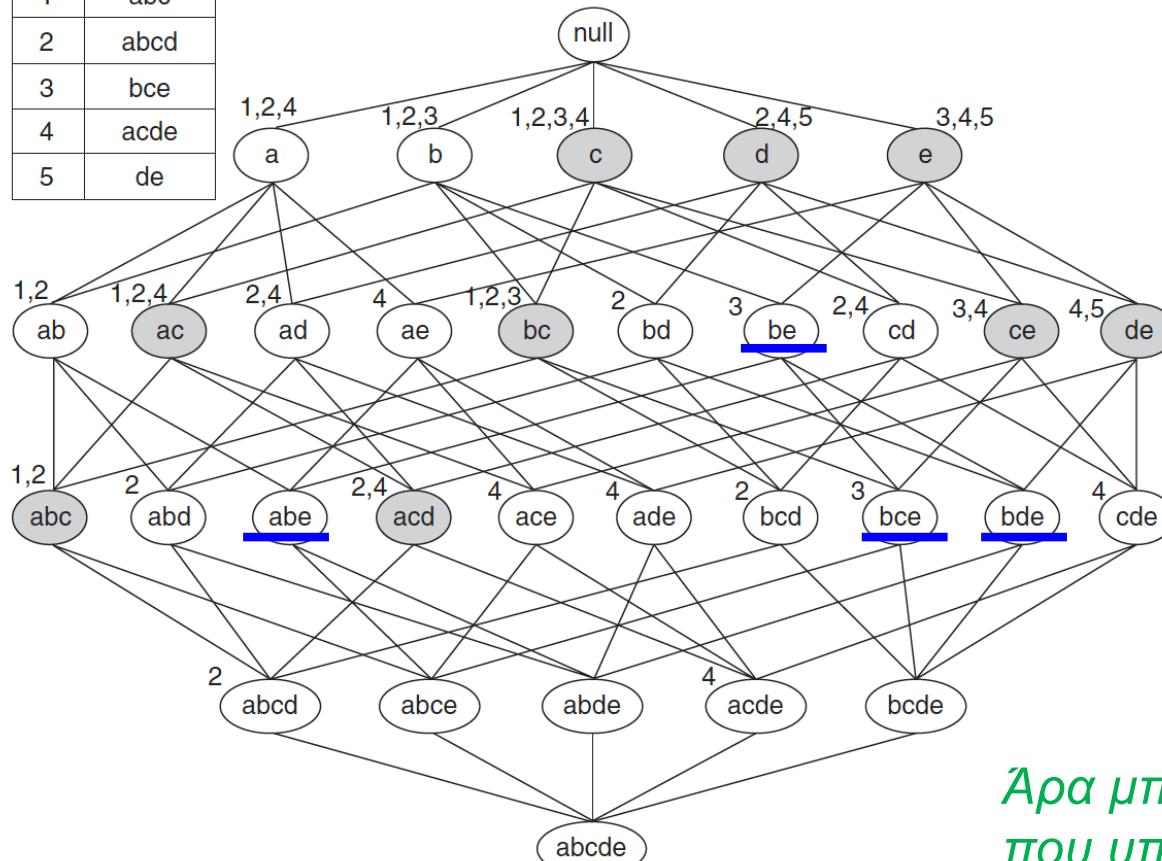
{b,c}: κλειστό

{a,d}: μη κλειστό  
διότι έχει την ίδια  
υποστήριξη με το  
υπερσύνολό του  
{a,c,d}

# Κλειστά Στοιχειοσύνολα

**Ιδιότητα κλειστών στοιχειοσυνόλων:** Αν γνωρίζουμε τις μετρήσεις υποστήριξής τους, μπορούμε να υπολογίσουμε τη μέτρηση υποστήριξης για οποιοδήποτε στοιχειοσύνολο

TID	Items
1	abc
2	abcd
3	bce
4	acde
5	de



To  $\{b, e\}$  δεν είναι κλειστό  
 → ένα από τα υπερσύνολά του έχει ίδια τιμή, βλ.  $\{b, c, e\}$   
 → κανένα από αυτά δεν μπορεί να έχει μεγαλύτερη τιμή  
**ΣΥΝΕΠΩΣ:** η υποστήριξή του είναι ίση με την μέγιστη υποστήριξη των:  $\{a, b, e\}$ ,  $\{b, c, e\}$  και  $\{b, d, e\}$

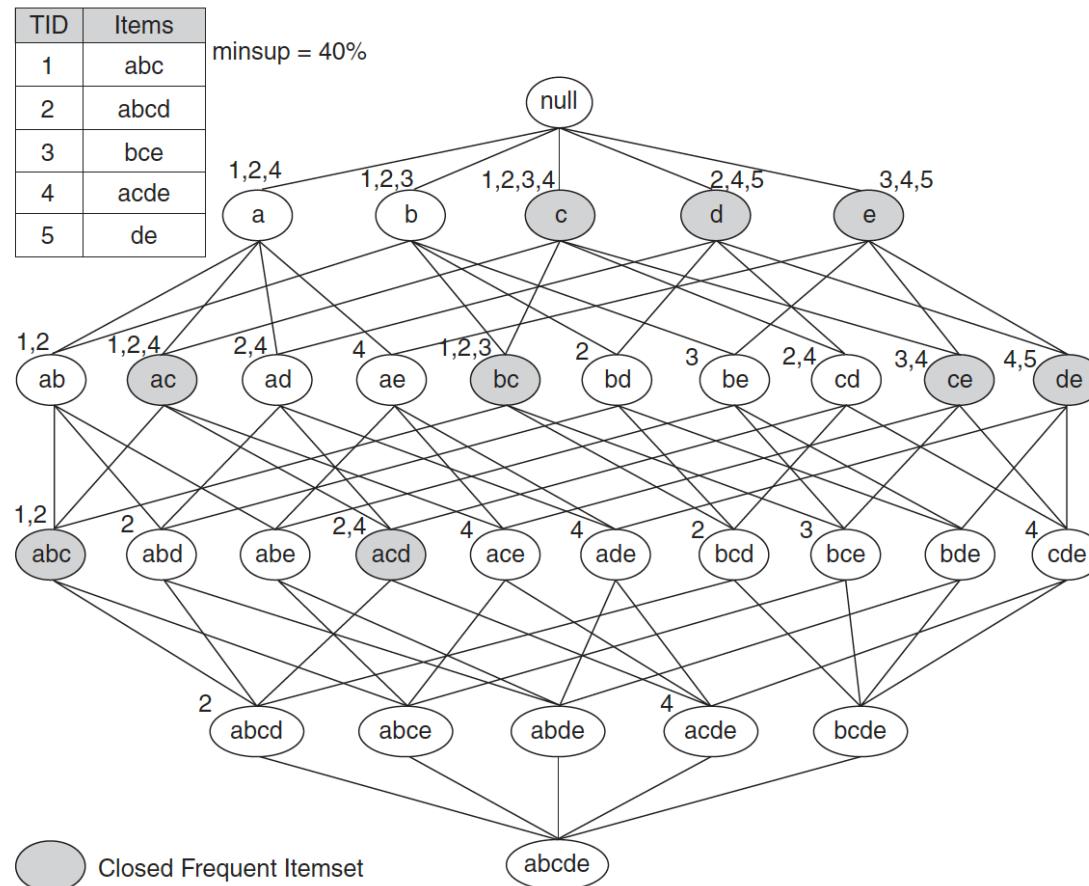
Άρα μπορεί να φτιαχτεί αλγόριθμος που υπολογίζει τις τιμές στο επίπεδο  $k-1$  βάσει των τιμών στο επίπεδο  $k$

# Γιατί τα Κλειστά Στοιχειοσύνολα δε μάς ικανοποιούν;

- Παρότι παρέχουν μια σύντομη αναπαράσταση των μετρήσεων υποστήριξης όλων των στοιχειοσυνόλων
  - Είναι εκθετικά πολλά σε πλήθος
  - Συνήθως, μας ενδιαφέρουν μόνο οι τιμές των συχνών στοιχειοσυνόλων (όχι όλων των στοιχειοσυνόλων)
- Για αυτούς τους λόγους χρησιμοποιούμε την έννοια των **κλειστών συχνών στοιχειοσυνόλων**

# Κλειστά Συχνά Στοιχειοσύνολα

Ένα στοιχειοσύνολο είναι **κλειστό συχνό στοιχειοσύνολο (closed frequent itemset)**, αν είναι **κλειστό** και η **υποστήριξή του** είναι **μεγαλύτερη** ή **ίση** με την τιμή κατωφλιού **υποστήριξης minsup**



# Παράδειγμα

TID	Items	
1	$\{a_1, a_2, \dots, a_{100}\}$	$\text{minsup} = 50\%$
2	$\{a_1, a_2, \dots, a_{50}\}$	

Κλειστά συχνά στοιχειοσύνολα:

$\{a_1, a_2, \dots, a_{100}\}$  με support = 1

$\{a_1, a_2, \dots, a_{50}\}$  με support = 2

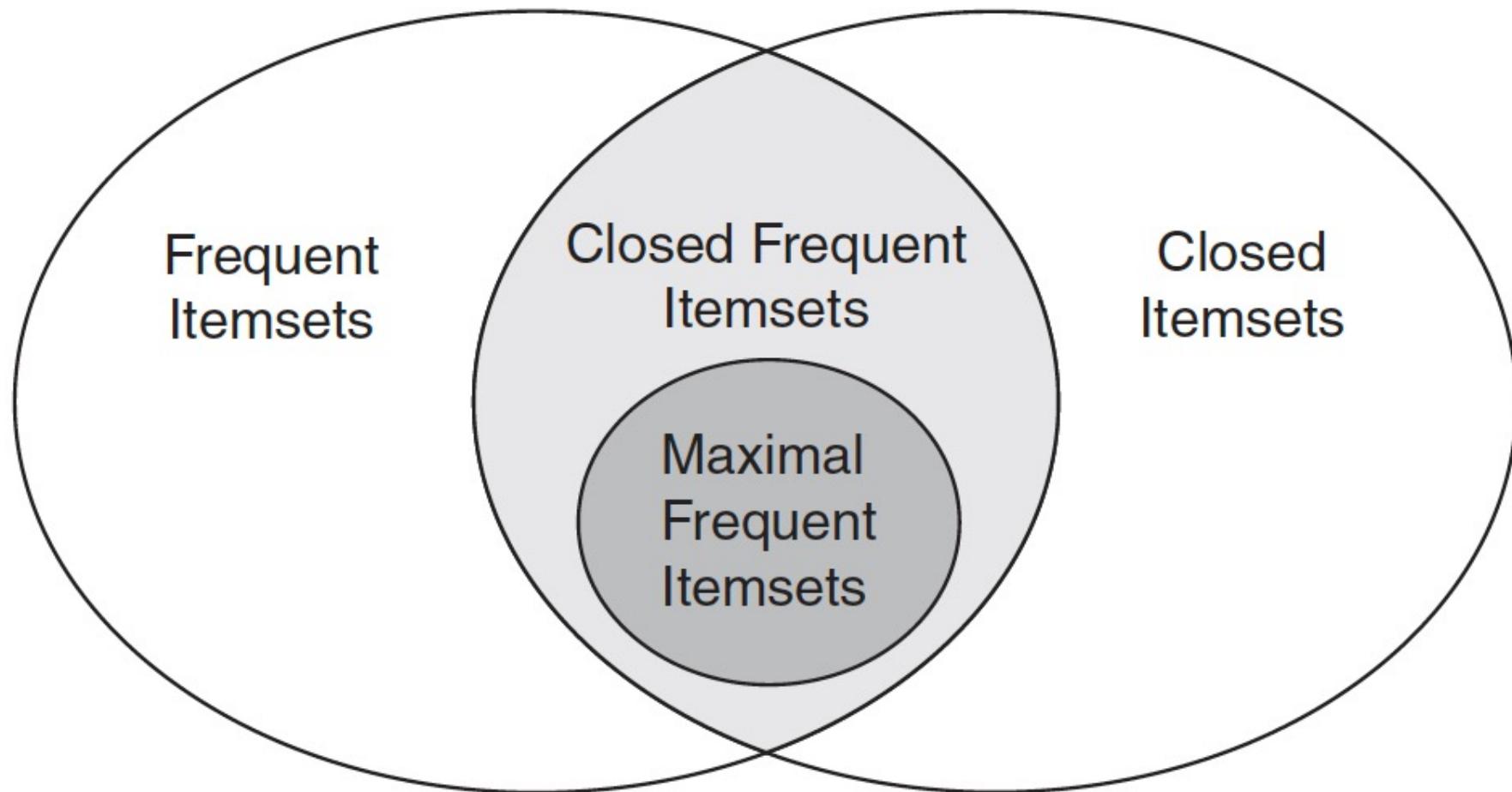
Μέγιστα συχνά στοιχειοσύνολα:

$\{a_1, a_2, \dots, a_{100}\}$  με support = 1

Απαρίθμηση των συχνών στοιχειοσυνόλων:

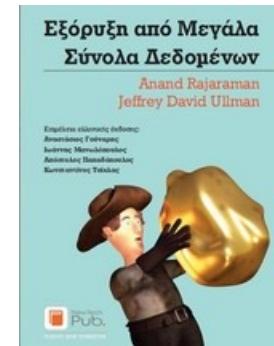
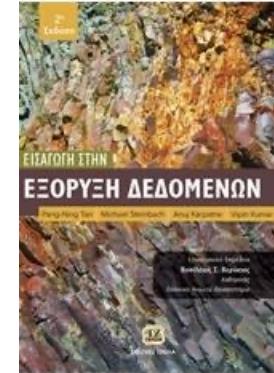
$2^{100} - 1$  στοιχειοσύνολα

# Σχέσεις που Ισχύουν



# Πηγές Αναφοράς

- ❑ P. Tan, M. Steinbach, A. Karpatne, V. Kumar. “Εισαγωγή στην Εξόρυξη Δεδομένων”, 2<sup>η</sup> Έκδοση, Εκδόσεις Τζιόλα.
  - *Κεφ. 5: Ανάλυση Συσχέτισης: Βασικές Έννοιες και Αλγόριθμοι*
- Anand Rajaraman, Jeffrey David Ullman. “Εξόρυξη από Μεγάλα Σύνολα Δεδομένων”, Εκδόσεις Νέων Τεχνολογιών.
  - ❑ *Κεφ.6: Συχνά Στοιχειοσύνολα*





## 9. Εύρεση Σημαντικών Γνωρισμάτων



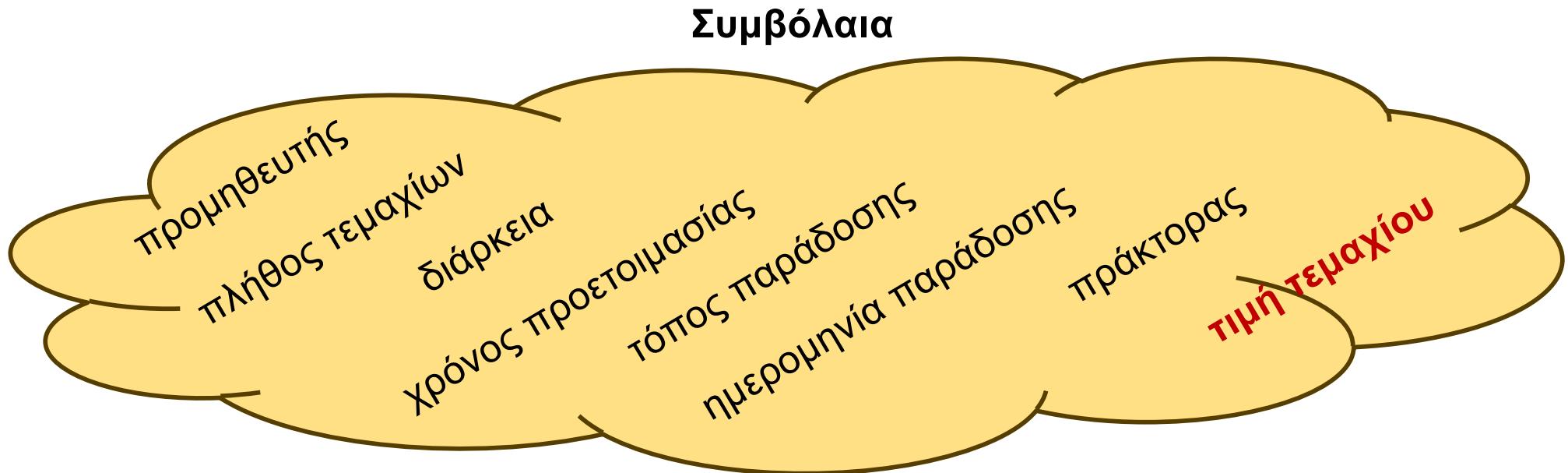
Ανάλυση Δεδομένων  
(*Data Analytics*)

Χρήστος Δουλκερίδης  
2024-25

# Το Πρόβλημα

- Δίνεται ένα πολυδιάστατο σύνολο δεδομένων
- Περιγράφεται από πολλά διαφορετικά γνωρίσματα
- Θέλουμε να απαντήσουμε ερωτήματα όπως:
  - *Από ποια γνωρίσματα εξαρτάται ένα συγκεκριμένο γνώρισμα;*
  - *Ποια από αυτά τα γνωρίσματα εξαρτώνται μεταξύ τους;*

# Παράδειγμα



- **Συμβόλαια που περιγράφονται από:**
  - προμηθευτή, πλήθος τεμαχίων, διάρκεια, χρόνο προετοιμασίας, τόπο παράδοσης, ημερομηνία παράδοσης, πράκτορα, **τιμή τεμαχίου**
- Ποια από αυτά τις γνωρίσματα επηρεάζουν περισσότερο την **τιμή τεμαχίου**;

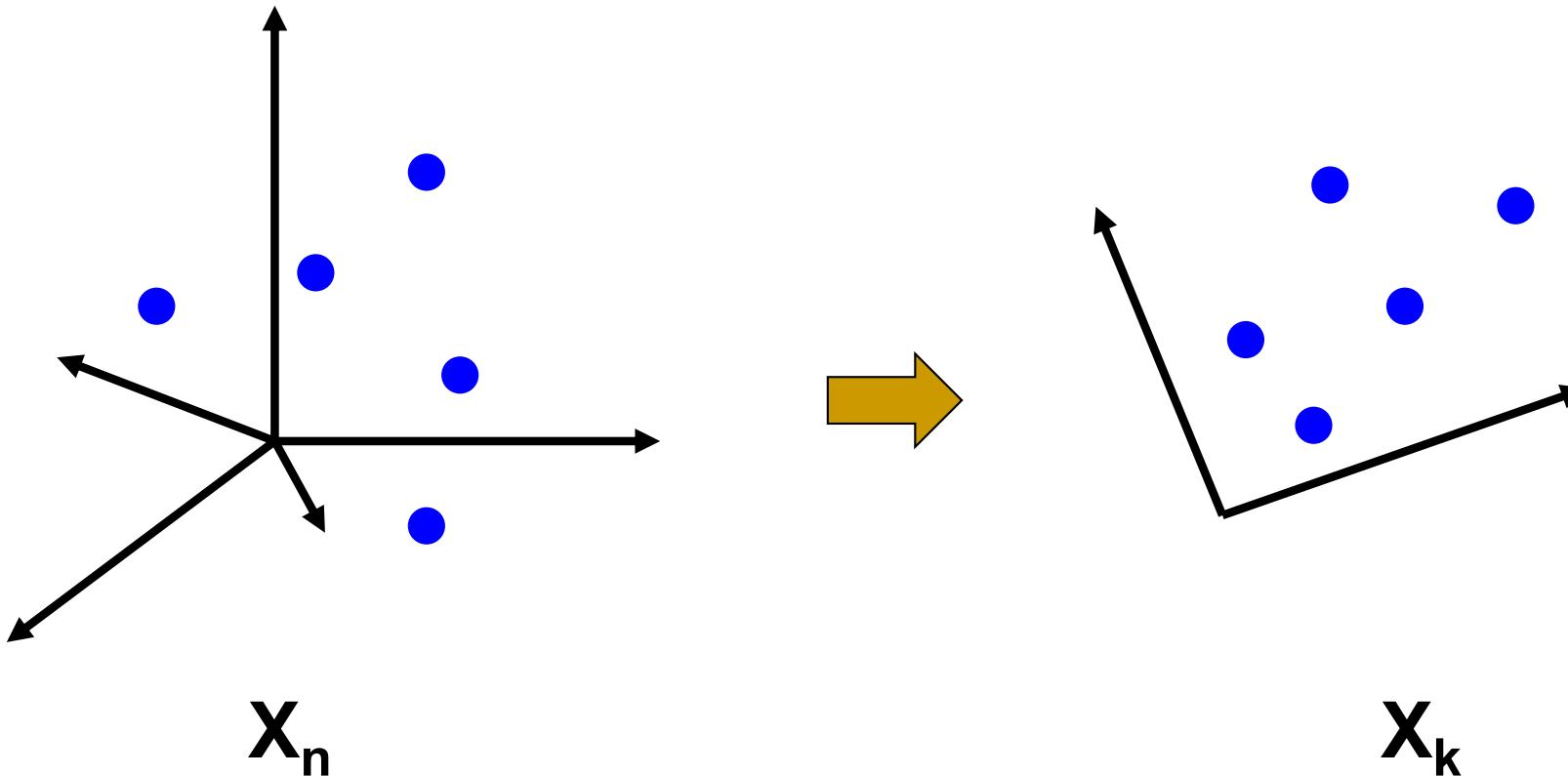
# Το Πρόβλημα

- Δύσκολο να αντιμετωπιστεί: πολλές μεταβλητές, καμιά δε φαίνεται να ξεχωρίζει, κάποιες μπορεί να είναι εξαρτώμενες
  - Για παράδειγμα: προμηθευτής και τόπος παράδοσης μπορεί να είναι συσχετιζόμενες μεταβλητές
    - Επιλογή προμηθευτή που είναι κοντά στον τόπο παράδοσης
- Πιθανή λύση
  - Με οπτικοποίηση όλων των ζευγών μεταβλητών
  - Σε πίνακα διαγραμμάτων διασποράς (scatter plot matrix)
  - Όμως, τι κάνουμε αν το πλήθος μεταβλητών είναι πολύ μεγάλο;
- Χρειάζεται μια υπολογιστική τεχνική

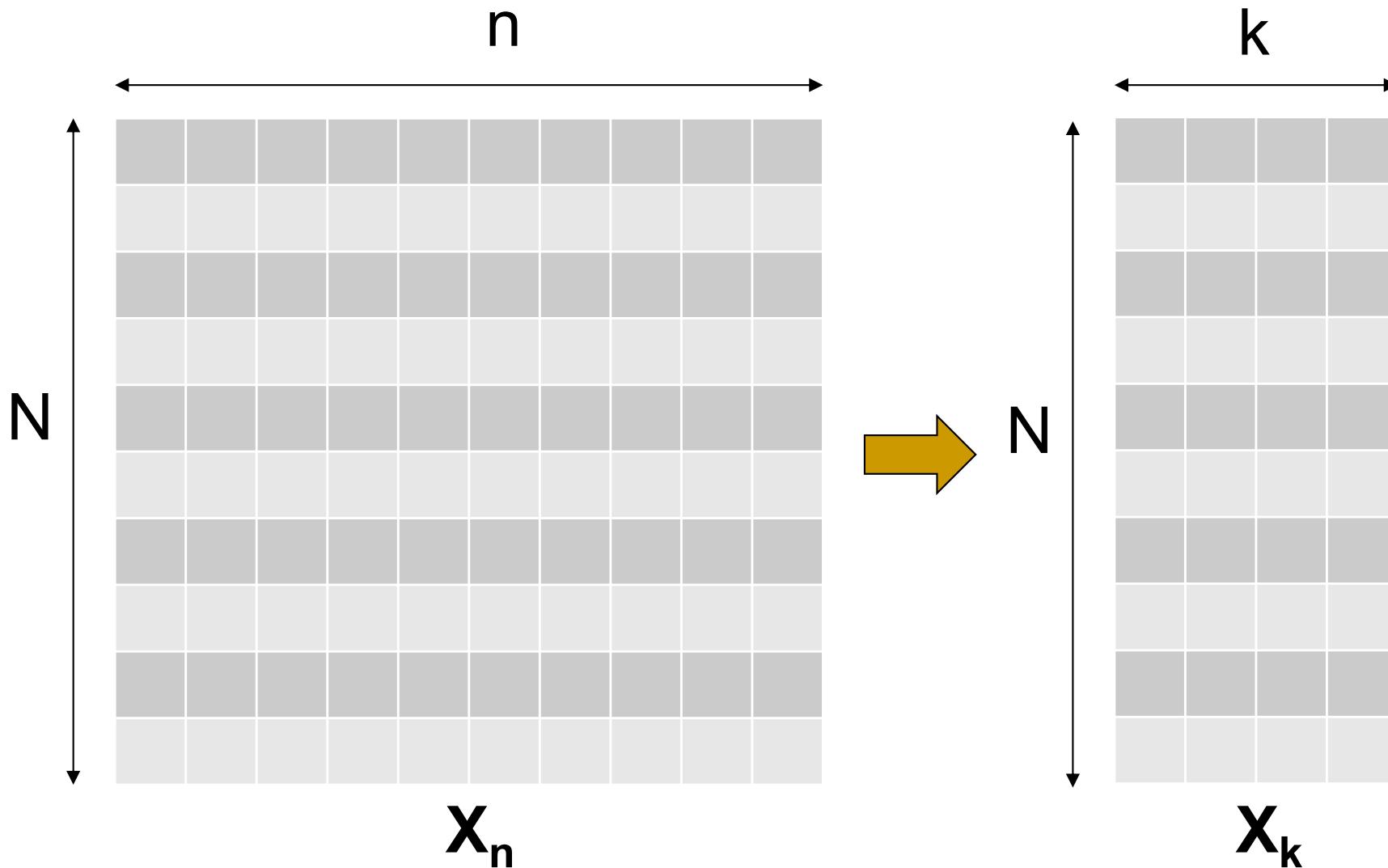
# Λύση

- *Επιλογή των πιο σημαντικών μεταβλητών (variables) ή γνωρισμάτων (features) από ένα πολυδιάστατο σύνολο δεδομένων*
- *Μετασχηματίζοντας το σύνολο από έναν αρχικό χώρο  $R^n$  σε έναν άλλο χώρο μικρότερης διάστασης  $R^k$  ( $k < n$ )*
  - Με την ελπίδα ότι στο χώρο μικρότερης διάστασης θα αναδεικνύεται κάποια ενδιαφέρουσα συμπεριφορά/ιδιότητα του συνόλου
- Τέτοιες μέθοδοι ή τεχνικές είναι γνωστές με τον όρο
  - επιλογή γνωρισμάτων (feature selection) ή
  - μείωση διάστασης (dimensionality reduction)

# Βασική Ιδέα (Γεωμετρικά)



# Βασική Ιδέα (Αλγεβρικά)



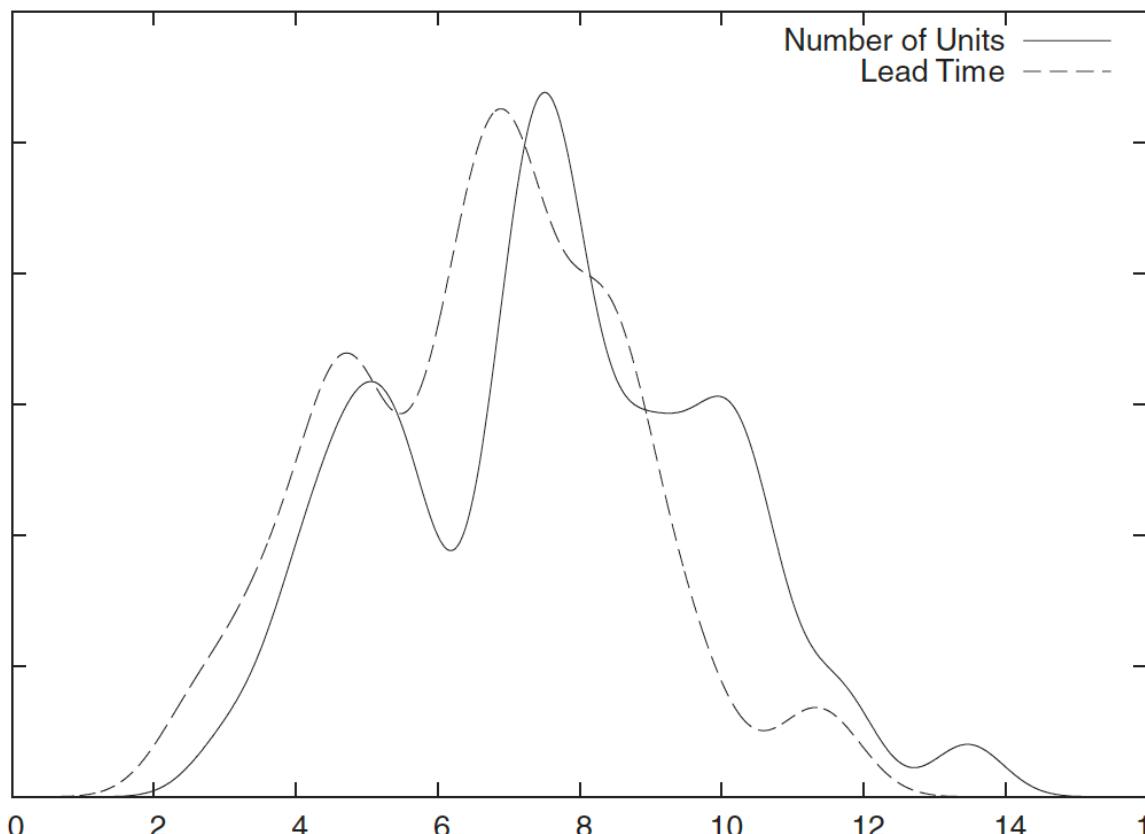
# Ανάλυση Κύριων Συνιστωσών (ΑΚΣ) Principal Component Analysis (PCA)

# Κινητρό (1/3)

Συμβόλαια που εξαρτώνται από δύο μόνο μεταβλητές:

- πλήθος τεμαχίων (Number of Units) και
- χρόνος προετοιμασίας (Lead Time)

Απεικόνιση κατανομής τιμών με Εκτιμητές Πυρήνα (*Kernel Density Estimates*)

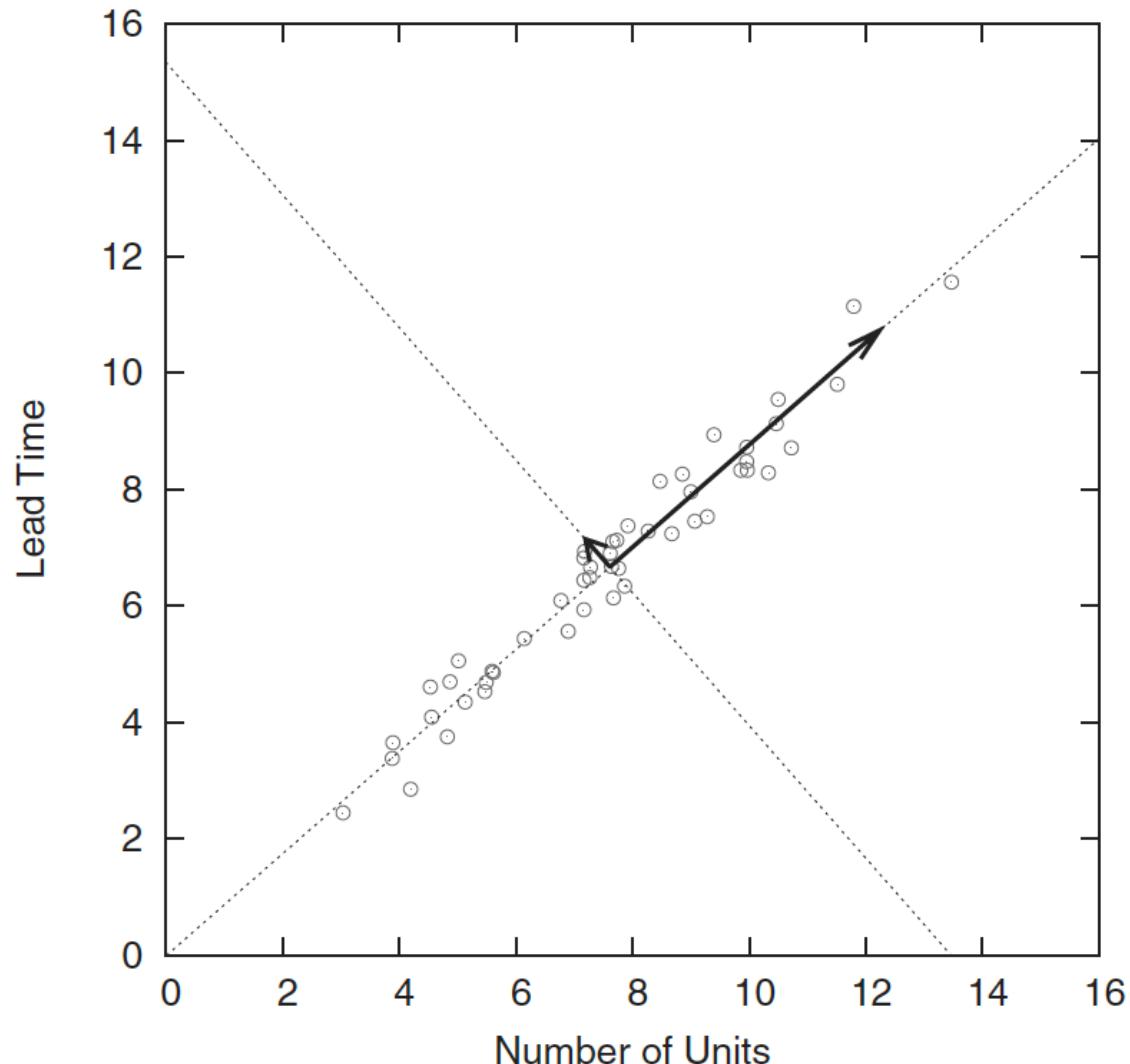


# Κινητρό (2/3)

Τα ίδια δεδομένα σε  
διάγραμμα διασποράς  
(scatter plot)

Φαίνεται ότι ο χρόνος  
προετοιμασίας αυξάνεται με  
το μέγεθος της παραγγελίας  
(πλήθος τεμαχίων)

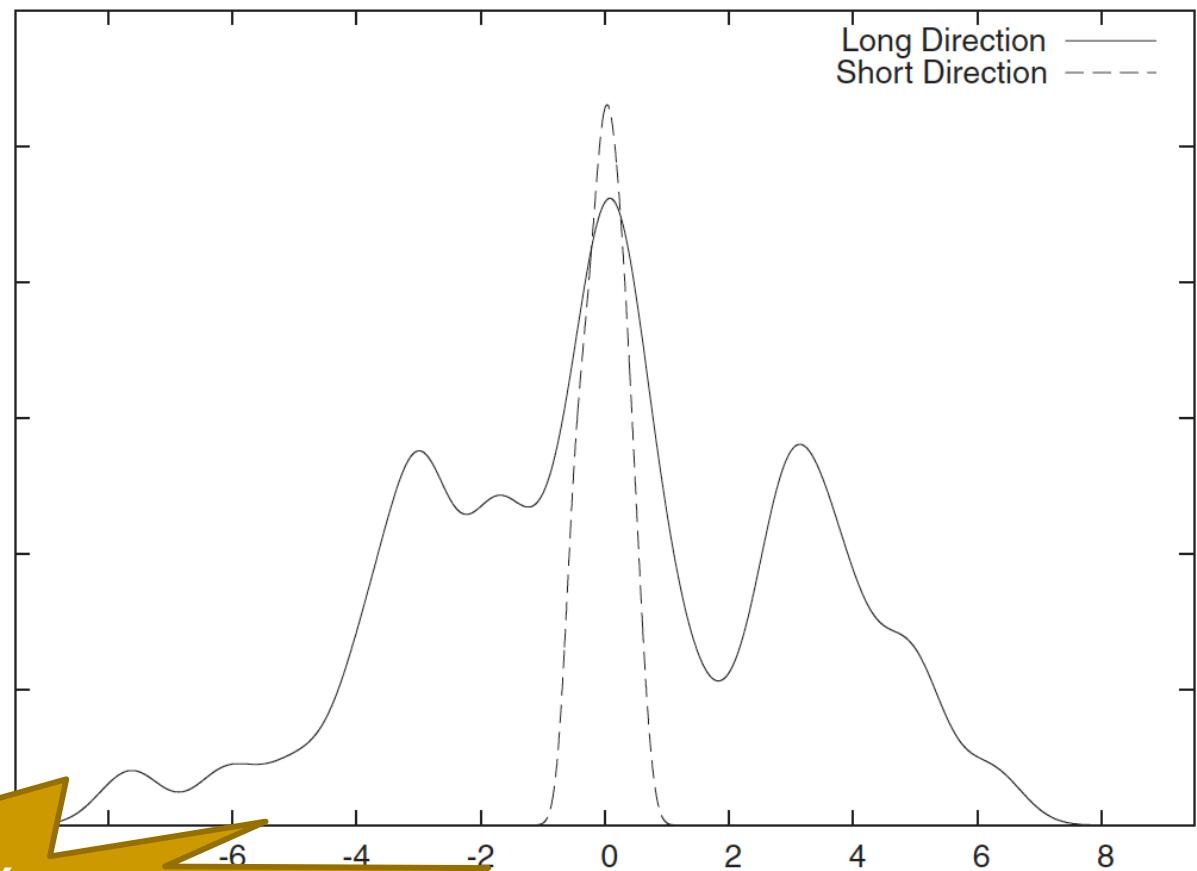
Επίσης, φαίνεται ότι όλα τα  
σημεία είναι σχεδόν σε μια  
ευθεία → *απαιτείται μόνο μία  
μεταβλητή* για να περιγράψει  
τη θέση κάθε σημείου: η  
απόσταση πάνω στην ευθεία



# Κίνητρο (3/3)

- Τα δεδομένα εξαρτώνται από δύο μεταβλητές
- Όμως, η περισσότερη διακύμανση (**variance**) υπάρχει σε **μια μόνο κατεύθυνση**
  - Εάν μετρούσαμε τα δεδομένα μόνο πάνω σε αυτή την κατεύθυνση, πάλι θα μπορούσαμε να καταγράψουμε οτιδήποτε «ενδιαφέρον» στα δεδομένα

Πάλι σχεδίαση εκτιμητών πυρήνα, αλλά χρησιμοποιώντας τις «**νέες** **κατευθύνσεις** που δείχνουν τα δύο βέλη στο προηγούμενο σχήμα



Τι κάνουμε όμως όταν  
έχουμε περισσότερες  
διαστάσεις?

Ανάλυση Δεδομένων (X.Δουλερίδης)

# Ανάλυση Κύριων Συνιστωσών

- Εάν αποτυπώσουμε την πληροφορία που αφορά ένα **πολυδιάστατο σύνολο δεδομένων** σε έναν **πίνακα**
- Τότε μπορούμε να εφαρμόσουμε τεχνικές από τη γραμμική άλγεβρα
- Για να **μετασχηματίσουμε** τον πίνακα σε μια μορφή που αποκαλύπτει την υποκείμενη (κρυφή) δομή

Αυτό επιτυγχάνεται με την **Ανάλυση Κύριων Συνιστωσών (ΑΚΣ) – Principal Components Analysis (PCA)**

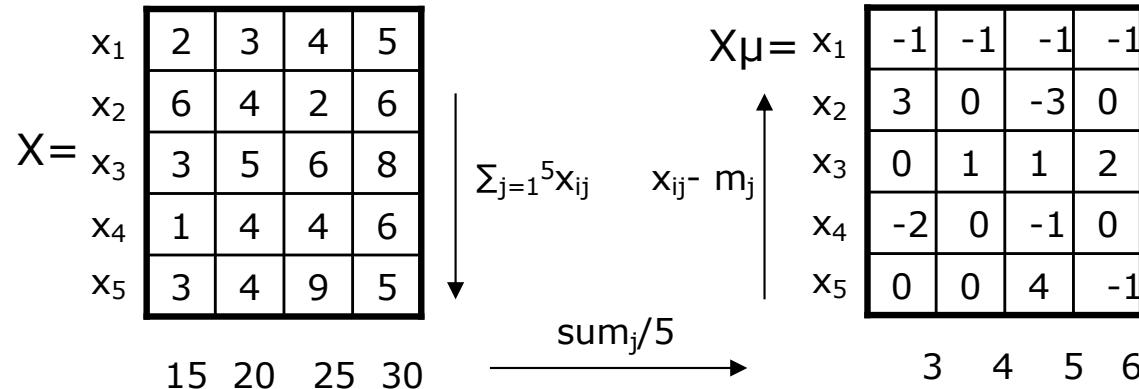
# Ανάλυση Κύριων Συνιστωσών

## ΒΗΜΑΤΑ ΕΡΓΑΣΙΑΣ

1. Κανονικοποίηση των τιμών των γνωρισμάτων
2. Υπολογισμός πίνακα συνδιακύμανσης (covariance matrix)
3. Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων
4. Επιλογή των Κ σημαντικότερων ιδιοδιανυσμάτων
5. Προβολή του συνόλου δεδομένων στις Κ διαστάσεις

# Ανάλυση Κύριων Συνιστωσών

## 1. Κανονικοποίηση των τιμών των γνωρισμάτων



1. Υπολογισμός μέσης τιμής ανά στήλη
  2. Αφαίρεση μέσης τιμής στήλης από κάθε τιμή στήλης
- Γνωστό και ως “center the data”

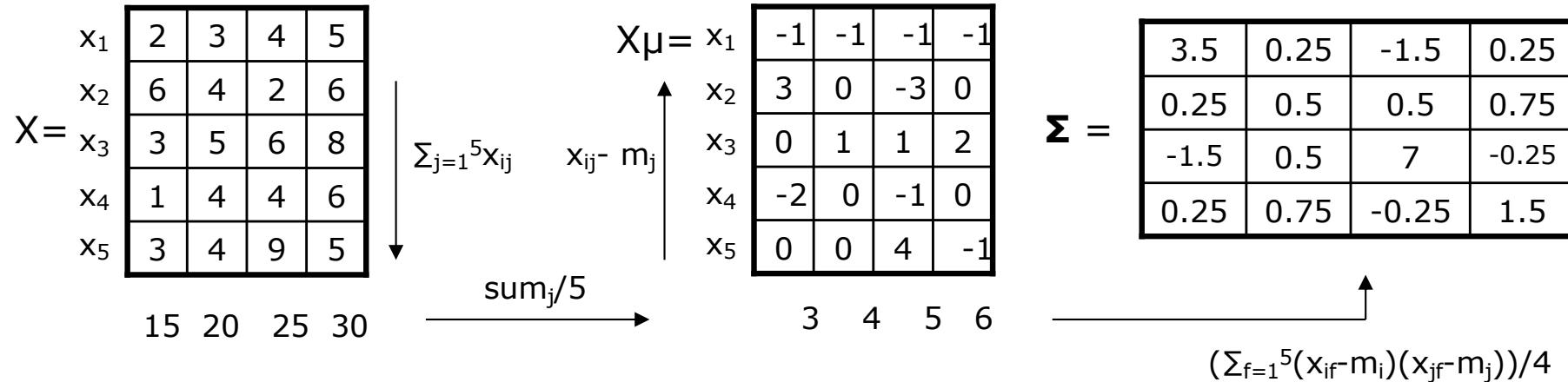
# Ανάλυση Κύριων Συνιστωσών

## ΒΗΜΑΤΑ ΕΡΓΑΣΙΑΣ

1. Κανονικοποίηση των τιμών των γνωρισμάτων
2. Υπολογισμός πίνακα συνδιακύμανσης (covariance matrix)
3. Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων
4. Επιλογή των  $K$  σημαντικότερων ιδιοδιανυσμάτων
5. Προβολή του συνόλου δεδομένων στις  $K$  διαστάσεις

# Ανάλυση Κύριων Συνιστωσών

## 2. Υπολογισμός πίνακα συνδιακύμανσης



$$\text{cov}(x, y) = \frac{1}{N} \sum_i^N (x_i - \bar{x})(y_i - \bar{y})$$

Σημείωση: τιμές στρογγυλοποιημένες στα 2 δεκαδικά

Διόρθωση Bessel

# Ανάλυση Κύριων Συνιστωσών – Θεωρία

- **Συντελεστής συσχέτισης (correlation coefficient):**
  - Μπορεί να εκφράσει την *ομοιότητα ανάμεσα σε δύο διαστάσεις x και y* ενός πολυδιάστατου συνόλου δεδομένων

$$\text{corr}(x, y) = \frac{1}{N} \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sigma(x)\sigma(y)}$$

- Ο παρονομαστής επανακαθορίζει την κλίμακα των δεδομένων σε ένα σταθερό διάστημα τιμών
- Εναλλακτικά, εάν αυτό δεν είναι επιθυμητό, χρησιμοποιούμε τη **συνδιακύμανση (covariance)** μεταξύ **x** και **y**:

$$\text{cov}(x, y) = \frac{1}{N} \sum_i^N (x_i - \bar{x})(y_i - \bar{y})$$

# Ανάλυση Κύριων Συνιστωσών – Θεωρία

- Πίνακας συνδιακύμανσης (covariance matrix)  $\Sigma$ :

$$\Sigma = \begin{pmatrix} \text{cov}(x, x) & \text{cov}(x, y) & \dots \\ \text{cov}(y, x) & \text{cov}(y, y) & \\ \vdots & & \ddots \end{pmatrix}$$

- Επειδή ισχύει ότι:  $\text{cov}(x,y)=\text{cov}(y,x)$ , ο πίνακας είναι συμμετρικός (άρα ισούται με τον ανάστροφό του:  $\Sigma=\Sigma^T$ )

# Ανάλυση Κύριων Συνιστωσών

## ΒΗΜΑΤΑ ΕΡΓΑΣΙΑΣ

1. Κανονικοποίηση των τιμών των γνωρισμάτων
2. Υπολογισμός πίνακα συνδιακύμανσης (covariance matrix)
3. Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων
4. Επιλογή των Κ σημαντικότερων ιδιοδιανυσμάτων
5. Προβολή του συνόλου δεδομένων στις Κ διαστάσεις

# Ανάλυση Κύριων Συνιστωσών

## 3. Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων

Να υπολογιστούν για τον παρακάτω πίνακα:

$$\begin{bmatrix} -2 & -4 & 2 \\ -2 & 1 & 2 \\ 4 & 2 & 5 \end{bmatrix}$$

# Στόχος Ανάλυσης Κύριων Συνιστωσών

Να βρεθεί ένα νέο σύνολο διαστάσεων (χαρακτηριστικών) που λαμβάνει καλύτερα τη μεταβλητότητα των δεδομένων

- Ο μετασχηματισμός ικανοποιεί τις ακόλουθες **Ιδιότητες**:
  - Κάθε ζεύγος νέων διαστάσεων έχει συνδιακύμανση 0 (για ξεχωριστές διαστάσεις) (**ορθογωνικότητα**)
  - Οι διαστάσεις **ταξινομούνται** σε σχέση με το **ποσοστό της διακύμανσης** των δεδομένων που λαμβάνουν
  - Η **πρώτη διάσταση** λαμβάνει όσο το δυνατόν **περισσότερη** μεταβλητότητα των δεδομένων
  - Κάθε **επόμενη διάσταση**, υπό την απαίτηση της ορθογωνικότητας, λαμβάνει όσο το δυνατόν **περισσότερη** από την **υπολειπόμενη διακύμανση**

# Ανάλυση Κύριων Συνιστωσών – Θεωρία

- Θεώρημα φασματικής ανάλυσης (spectral decomposition theorem)

Για κάθε πραγματικό, συμμετρικό NxN πίνακα  $A$  υπάρχει ένας ορθογώνιος πίνακας  $U$  τέτοιος ώστε ο  $B$  να είναι διαγώνιος:

$$B = \begin{pmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & \lambda_N & \\ & & & \ddots \end{pmatrix} = U^{-1} A U$$

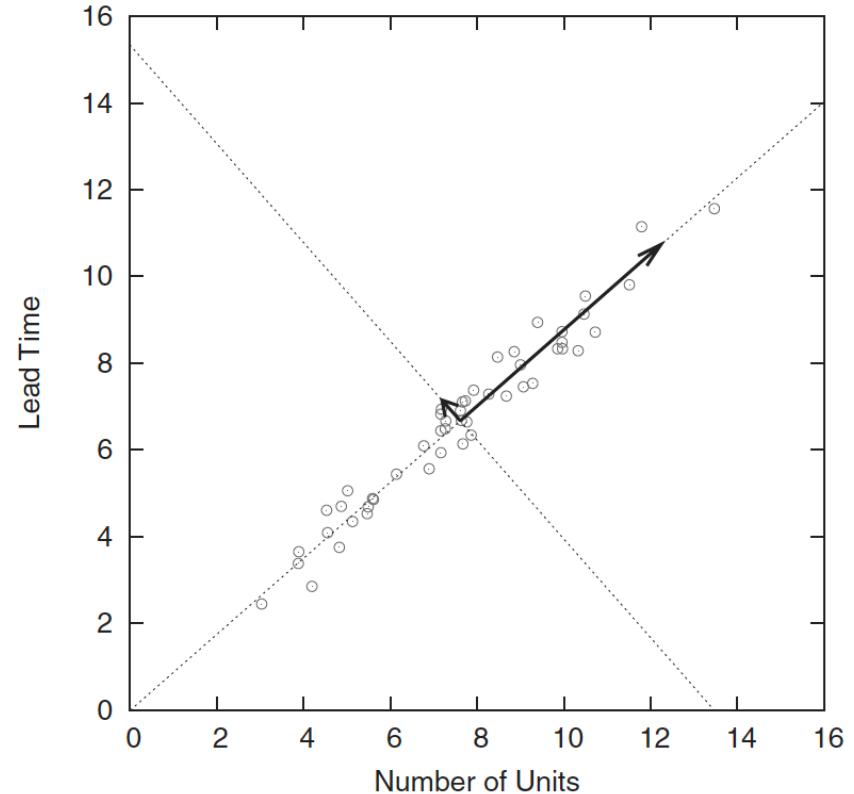
Ιδιοτιμές  $\lambda_1, \lambda_2, \dots, \lambda_N$  είναι οι Ιδιοδιανύσματα του πίνακα  $A$ .

Υπενθύμιση:

- 1) Διαγώνιος πίνακας: όταν τα μόνα μη-μηδενικά στοιχεία του βρίσκονται σημειώσεις στη διαγώνιο
- 2) Ορθογώνιος πίνακας: όταν ο ανάστροφος ισούται με τον αντίστροφό του:  $U^T = U^{-1}$  και  $UU^T = U^T U = I$

# Ερμηνεία ΑΚΣ (1/3)

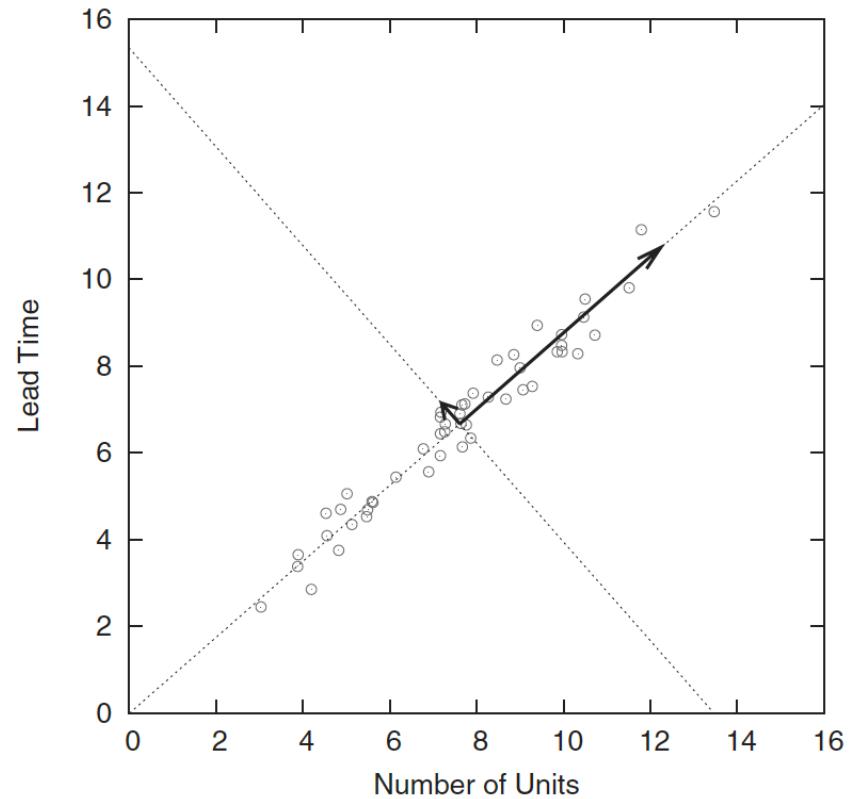
- **Ερμηνεία θεωρήματος φασματικής ανάλυσης**
  - Μπορούμε με **αλλαγές μεταβλητών** να μετασχηματίσουμε έναν οποιοδήποτε **συμμετρικό πίνακα A** σε ένα **διαγώνιο πίνακα B**
  - Ο **B** περιέχει την ίδια πληροφορία με τον **A** (απλά με διαφορετικό τρόπο)
  - Η αλλαγή μεταβλητής συνίσταται σε μια **περιστροφή** του αρχικού συστήματος συντεταγμένων σε ένα νέο σύστημα συντεταγμένων, στο οποίο ο πίνακας συσχέτισης έχει ένα βολικό σχήμα (είναι διαγώνιος)



Οι άξονες του νέου συστήματος συντεταγμένων μπορούν να εκφραστούν ως **γραμμικός συνδυασμός** των αρχικών αξόνων,  
π.χ.:  $X = (x + y) / \sqrt{2}$

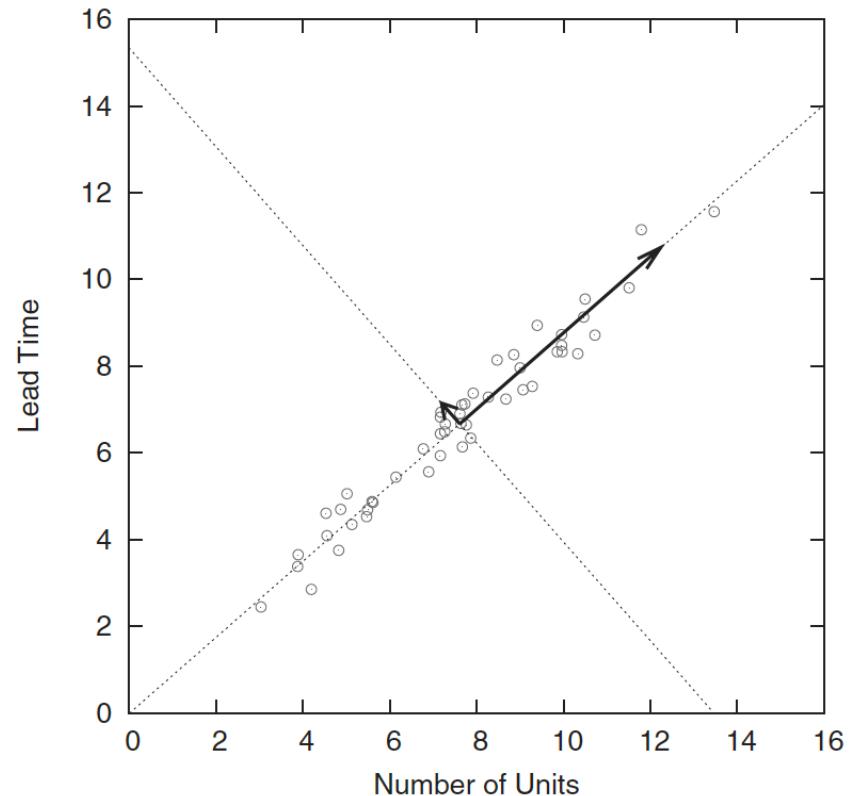
# Ερμηνεία ΑΚΣ (2/3)

- Το θεώρημα εφαρμόζεται σε οποιοδήποτε **συμμετρικό πίνακα**
- Εφαρμόζοντας το θεώρημα στον πίνακα συνδιακύμανσης, το αποτέλεσμα της παραγοντοποίησης μάς δίνει τους **κύριους αξόνες της κατανομής των σημείων**
- Στο διπλανό σχήμα:
  - Τα δεδομένα βρίσκονται σε μια έλλειψη που έχει "τεντωθεί"
  - Τα **ιδιοδιανύσματα** βρίσκονται στην κατεύθυνση των κύριων αξόνων της έλλειψης
  - Οι **ιδιοτιμές** δίνουν το σχετικό μέγεθος αντίστοιχων κύριων αξόνων

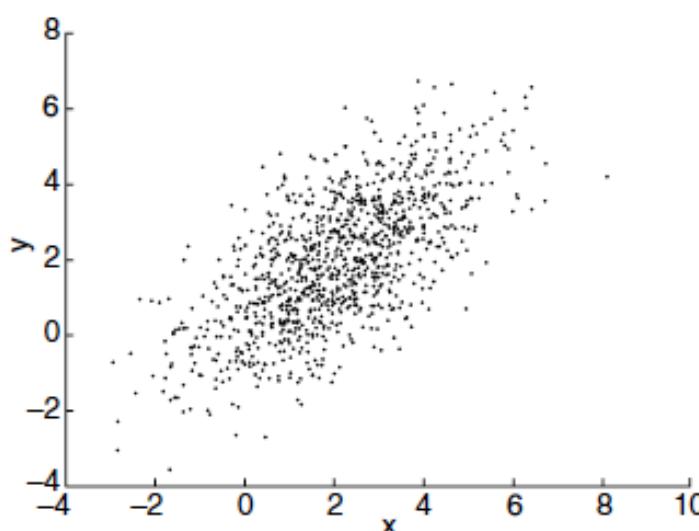


# Ερμηνεία ΑΚΣ (3/3)

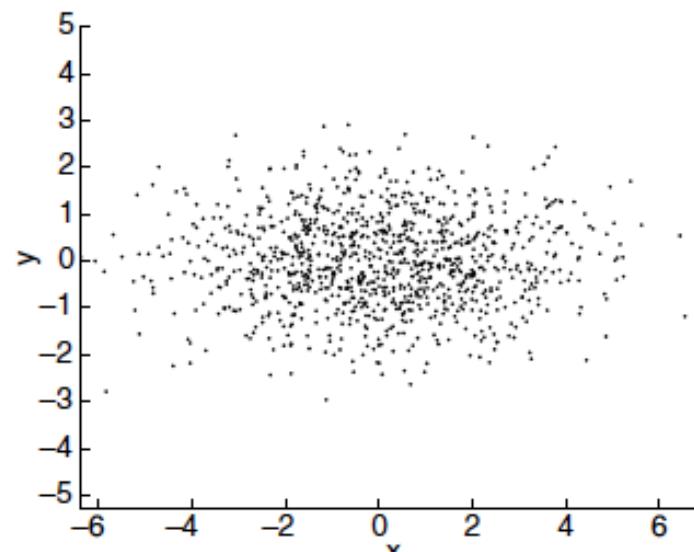
- Τα ιδιοδιανύσματα δείχνουν στις κατευθύνσεις με τη μεγαλύτερη διακύμανση: τα δεδομένα εκτείνονται περισσότερο στο χώρο εάν τα μετρήσουμε σε αυτές τις (κύριες) διαστάσεις
- Οι ιδιοτιμές αποτελούν ένα μέτρο του εύρους των δεδομένων στην αντίστοιχη διάσταση
- Στην πράξη, η κάθε ιδιοτιμή ισούται με το τετράγωνο της τυπικής απόκλισης (=διακύμανση) στην αντίστοιχη διάσταση
- Διαγώνιες τιμές του πίνακα  $\Sigma$ :  $\sigma^2(x) = \sum_i (x_i - \bar{x})^2$



# Παράδειγμα: Σύνολο Δεδομένων 2Δ



(a) Original points.



(b) Points after transformation.

Figure B.1. Using PCA to transform the data.

- Σύνολο δεδομένων 1000 σημείων στις 2Δ, πριν και μετά την εφαρμογή της ΑΚΣ
  - Άθροισμα διακύμανσης (πριν) =  $2.84 + 2.95 = 5.79$
  - Άθροισμα διακύμανσης (μετά) =  $4.81 + 0.98 = 5.79$

# Ανάλυση Κύριων Συνιστωσών

## ΒΗΜΑΤΑ ΕΡΓΑΣΙΑΣ

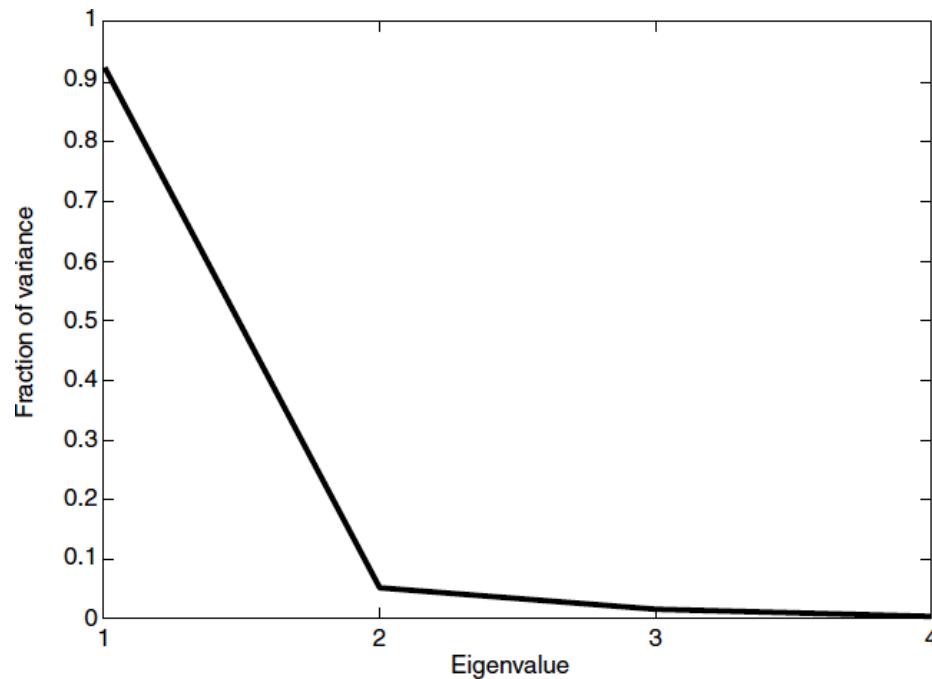
1. Κανονικοποίηση των τιμών των γνωρισμάτων
2. Υπολογισμός πίνακα συνδιακύμανσης (covariance matrix)
3. Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων
4. Επιλογή των  $K$  σημαντικότερων ιδιοδιανυσμάτων
5. Προβολή του συνόλου δεδομένων στις  $K$  διαστάσεις

# Ανάλυση Κύριων Συνιστωσών

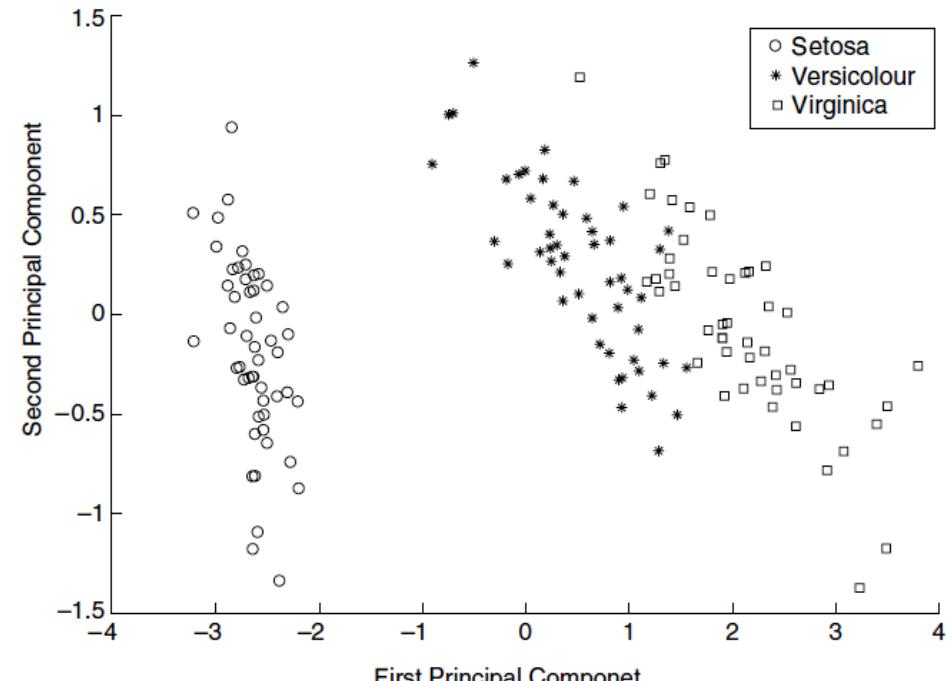
## 4. Επιλογή των K σημαντικότερων ιδιοδιανυσμάτων

- **1<sup>ος</sup> τρόπος:** Διατήρηση επιθυμητού ποσοστού μεταβλητότητας
  - Π.χ. επιλογή των K ιδιοτιμών που διατηρούν το 90% της μεταβλητότητας
- **2<sup>ος</sup> τρόπος:** Γράφημα πλαγιάς (scree plot)
  - Y-άξονας: ποσοστό μεταβλητότητας
  - X-άξονας: ιδιοτιμές
  - Επιλογή των K ιδιοτιμών
- **3<sup>ος</sup> τρόπος:** Βάσει διαφοράς μεταξύ διαδοχικών ιδιοτιμών
  - Ορίζοντας μια τιμή κατωφλιού T
  - $\sum_{j=k+1}^p \lambda_j / \sum_{j=1}^p \lambda_j > T$

# Παράδειγμα: Το Σύνολο Δεδομένων Iris



(a) Fraction of variance accounted for by each principal component.



(b) Plot of first two principal components of Iris data.

- (α) Γράφημα πλαγιάς (scree plot)
  - Χρήσιμο για τον καθορισμό του πλήθους των κύριων συνιστωσών που θα κρατήσουμε
  - Iris: 1<sup>η</sup> κύρια συνιστώσα (92.5%), 2<sup>η</sup> κύρια συνιστώσα (5.3%), 3-4<sup>η</sup> (2.2%)
- (β) Διάγραμμα διασποράς (scatter plot)
  - Παρατηρήστε το διαχωρισμό μεταξύ των κατηγοριών

# Ανάλυση Κύριων Συνιστωσών

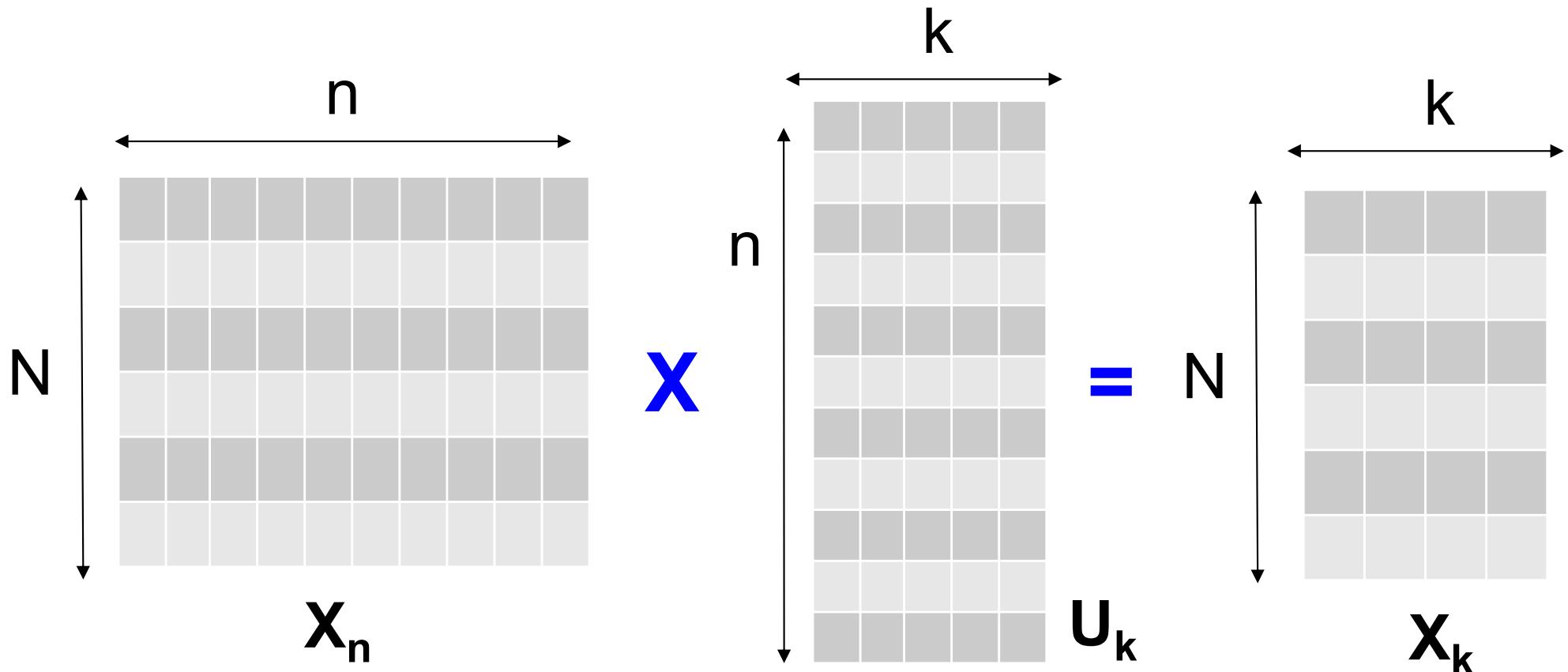
## ΒΗΜΑΤΑ ΕΡΓΑΣΙΑΣ

1. Κανονικοποίηση των τιμών των γνωρισμάτων
2. Υπολογισμός πίνακα συνδιακύμανσης (covariance matrix)
3. Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων
4. Επιλογή των  $K$  σημαντικότερων ιδιοδιανυσμάτων
5. Προβολή του συνόλου δεδομένων στις  $K$  διαστάσεις

# Ανάλυση Κύριων Συνιστωσών

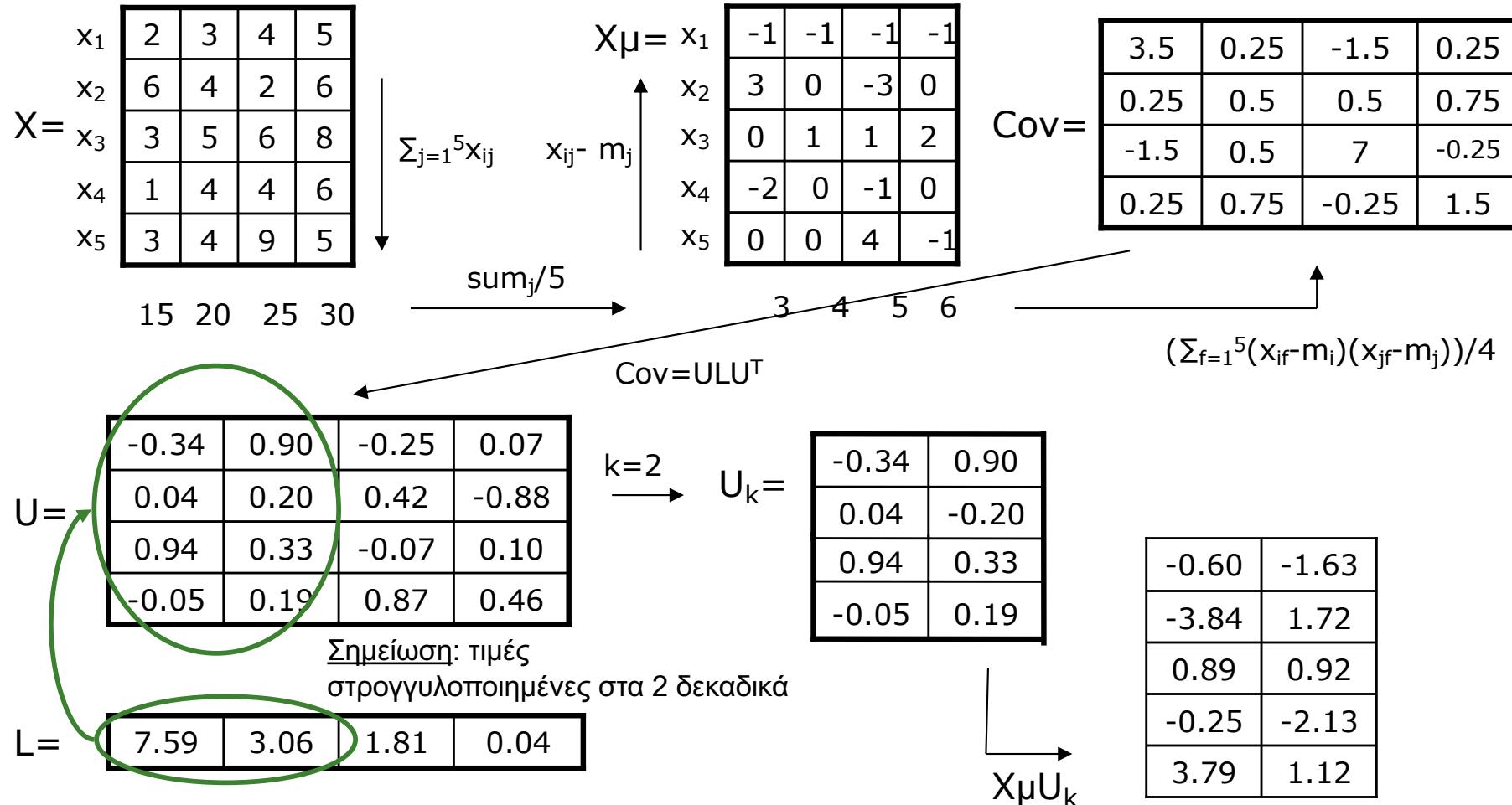
## 5. Προβολή του συνόλου δεδομένων στις $K$ διαστάσεις

- Πολλαπλασιάζοντας τον πίνακα δεδομένων (centered data) με τον πίνακα που αποτελείται από τα  $K$  πρώτα ιδιοδιανύσματα



# Ανάλυση Κύριων Συνιστωσών

## 5. Προβολή του συνόλου δεδομένων στις K διαστάσεις



# Ερμηνεία – Παρατηρήσεις

# Ανάλυση Κύριων Συνιστωσών – Ερμηνεία

- Εάν τα σημεία στον αρχικό χώρο έχουν μια σφαιροειδή κατανομή
  - Η ανάλυση κύριων συνιστωσών (PCA) δίνει τις **κατευθύνσεις** των κυρίων **αξόνων** της ελλειψοειδούς κατανομής των δεδομένων, ενώ
  - Οι **ιδιοτιμές** δίνουν το **εύρος** της κατανομής της έλλειψης σε κάθε έναν από τους κύριους άξονες
- Επιπλέον, ορισμένες από τις αρχικές μεταβλητές μπορεί να είναι **περιττές** (εκφράζουν την ίδια πληροφορία) ή **άσχετες** (περιέχουν λίγη πληροφορία)
- Μια ένδειξη ότι μεταβλητές είναι περιττές είναι όταν είναι **συσχετιζόμενες** (correlated)
  - Ο PCA χρησιμοποιεί την πληροφορία αμοιβαίας συσχέτισης μεταξύ μεταβλητών, για να αναγνωρίσει περιττές μεταβλητές

# Ανάλυση Κύριων Συνιστωσών – Ερμηνεία

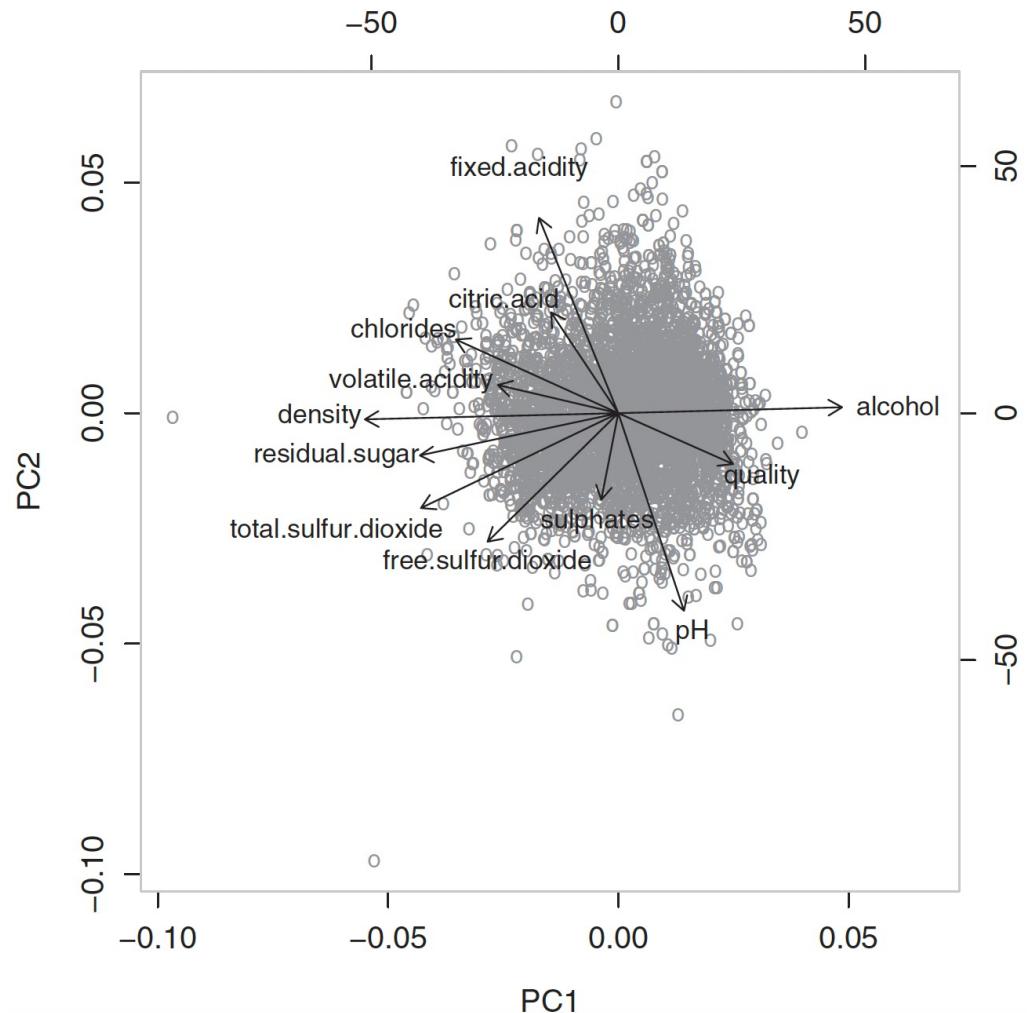
- Οι **περιττές μεταβλητές** είναι αυτές που αντιστοιχούν σε **μικρές ιδιοτιμές**
  - Σε εκείνες τις διαστάσεις το εύρος της κατανομής είναι μικρό, και μπορεί να θεωρηθούν ως σταθερές
- Οι νέες κατευθύνσεις αντιστοιχούν σε συνδυασμούς των αρχικών μεταβλητών
- Ο PCA λειτουργεί καλά όταν υπάρχει συσχέτιση μεταξύ μεταβλητών
  - Εάν δεν υπάρχει τέτοια συσχέτιση, δεν έχει νόημα η εφαρμογή του PCA

# Ανάλυση Κύριων Συνιστωσών - Παρατηρήσεις

- Ο PCA είναι είτε **διερευνητική** (exploratory) ή **προπαρασκευαστική** (preparatory) τεχνική στη διαδικασία ανάλυσης δεδομένων
- Διερευνητική:
  - Μπορεί ο αναλυτής να μελετήσει τα αποτελέσματά του (ιδιοδιανύσματα και ιδιοτιμές) για την κατανόηση του συνόλου δεδομένων
    - Δεν υπάρχουν όμως εγγυήσεις ότι θα προκύψει κάτι χρήσιμο
- Προπαρασκευαστική:
  - Μπορεί ο αναλυτής να μετασχηματίσει ένα πολυδιάστατο σύνολο δεδομένων σε μια ισοδύναμη μορφή, όπου όλες οι μεταβλητές είναι αμοιβαία ανεξάρτητες, πριν οποιαδήποτε περαιτέρω ανάλυση
    - Πάλι δεν είναι πάντα χρήσιμη
    - π.χ. αν οι αρχικές μεταβλητές δεν παρουσιάζουν κάποια συσχέτιση
    - ή αν καμιά από τις ιδιοτιμές δεν είναι σημαντικά μικρότερες ώστε να αφαιρεθούν

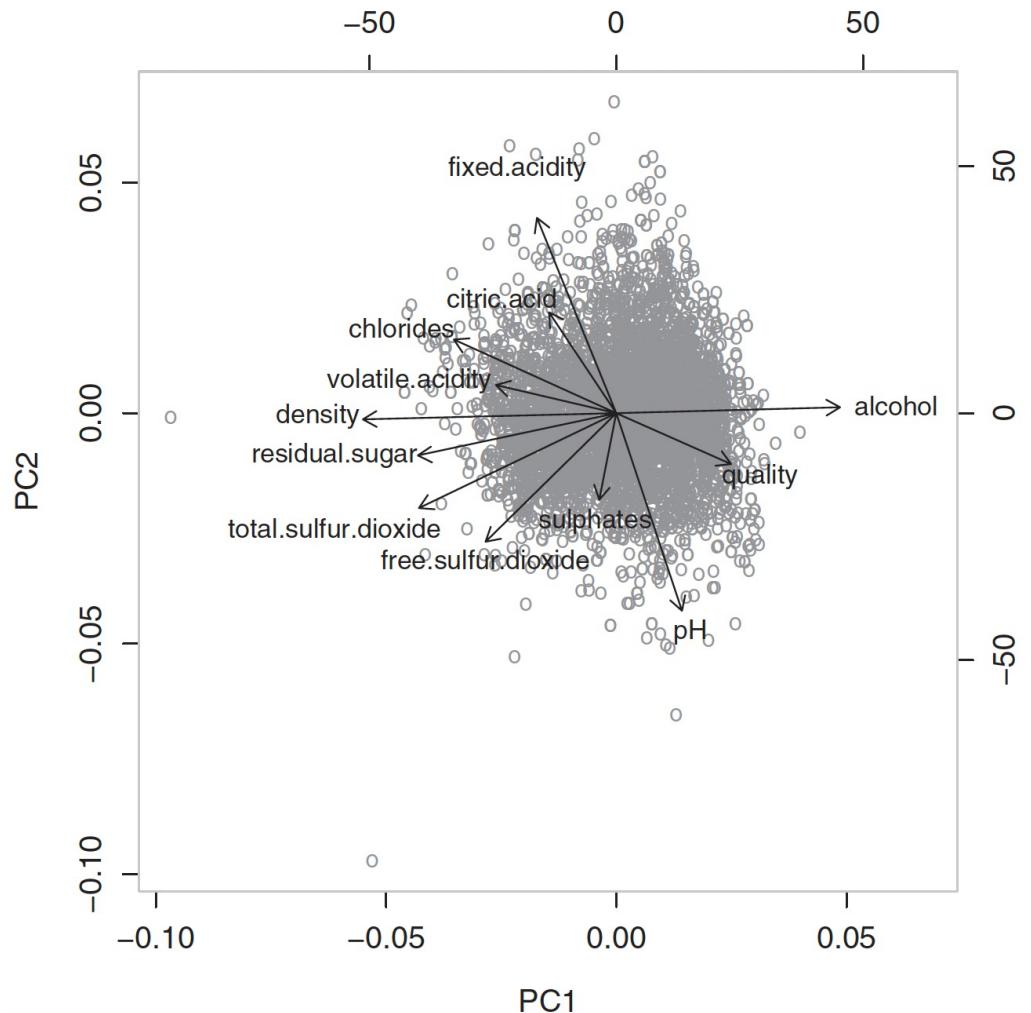
# Διγράφημα - Biplot

- Ένας τρόπος οπτικοποίησης αποτελεσμάτων του PCA
- Απεικόνιση σε σύστημα συντεταγμένων με χρήση των δύο πρώτων ιδιοδιανυσμάτων
- Σε αυτό το χώρο προβάλλονται και τα δεδομένα
- Οι κατευθύνσεις των **αρχικών μεταβλητών** αναπαρίστανται με βέλη (είναι οι προβολές τους στο χώρο των δύο πρώτων ιδιοδιανυσμάτων)



# Ερμηνεία

- Οριζόντια/κάθετα βέλη:
  - Αντιστοιχούν σε μεταβλητές που ευθυγραμμίζονται με τα δύο ιδιοδιανύσματα
  - Άρα **συνεισφέρουν πολύ στα ιδιοδιανύσματα**
- Παράλληλα βέλη:
  - **Περιπτές** μεταβλητές



# Εφαρμογή του PCA στην Πράξη

# Εφαρμογή του PCA στην Πράξη (1/8)

- Εφαρμόζεται με χρήση εξειδικευμένων εργαλείων
- Παράδειγμα χρήσης της **R**, ενός δημοφιλούς πακέτου ανοικτού κώδικα για στατιστικούς υπολογισμούς
- **Άσκηση:** επαναλάβετε με *Python* (*πακέτο scikit-learn*)

```
from sklearn.decomposition import PCA
```

# Εφαρμογή του PCA στην Πράξη (2/8)

- Εφαρμογή στο σύνολο δεδομένων **wine**
  - Περιέχει 5.000 κρασιά για τα οποία έχουν καταγραφεί 12 μεταβλητές (γνωρίσματα)
  - Μία από αυτές είναι η **ποιότητα (quality)** που είναι υποκειμενικό κριτήριο
- Ερώτημα: ποιες από αυτές συσχετίζονται και ποιες είναι οι σημαντικότερες που καθορίζουν την ποιότητα?

- 1 - fixed acidity
- 2 - volatile acidity
- 3 - citric acid
- 4 - residual sugar
- 5 - chlorides
- 6 - free sulfur dioxide
- 7 - total sulfur dioxide
- 8 - density
- 9 - pH
- 10 - sulphates
- 11 - alcohol
- 12 - quality (score between 0 and 10)

Σύνολο δεδομένων «Wine Quality» διαθέσιμο στο UCI Machine Learning Repository: <http://archive.ics.uci.edu/ml/>

## Εφαρμογή του PCA στην Πράξη (3/8)

```
wine <- read.csv( "winequality-white.csv", sep=';', header=TRUE )  
pc <- prcomp( wine )  
plot( pc )
```



- Η πρώτη ιδιοτιμή κυριαρχεί μεταξύ των άλλων, που μπορεί να υπονοεί ότι αυτή η νέα μεταβλητή καθορίζει τα πάντα
- Ας δούμε τα αντίστοιχα ιδιοδιανύσματα

# Εφαρμογή του PCA στην Πράξη (4/8)

```
print( pc )
```

```
(some output omitted...)
```

	PC1	PC2	PC3
fixed.acidity	-1.544402e-03	-9.163498e-03	-1.290026e-02
volatile.acidity	-1.690037e-04	-1.545470e-03	-9.288874e-04
citric.acid	-3.386506e-04	1.403069e-04	-1.258444e-03
residual.sugar	-4.732753e-02	1.494318e-02	-9.951917e-01
chlorides	-9.757405e-05	-7.182998e-05	-7.849881e-05
free.sulfur.dioxide	-2.618770e-01	9.646854e-01	2.639318e-02
total.sulfur.dioxide	-9.638576e-01	-2.627369e-01	4.278881e-02
density	-3.596983e-05	-1.836319e-05	-4.468979e-04
pH	-3.384655e-06	-4.169856e-05	7.017342e-03
sulphates	-3.409028e-04	-3.611112e-04	2.142053e-03
alcohol	1.250375e-02	6.455196e-03	8.272268e-02

```
(some output omitted...)
```

Τα δύο πρώτα  
ιδιοδιανύσματα έχουν  
πολύ υψηλές τιμές  
στη μεταβλητή που  
σχετίζεται με διοξείδιο  
του Θείου (sulfur  
dioxide)

Παράξενο αποτέλεσμα!

# Εφαρμογή του PCA στην Πράξη (5/8)

```
summary(wine)
fixed.acidity      volatile.acidity    citric.acid      residual.sugar
Min. : 3.800        Min. :0.0800       Min. :0.0000       Min. : 0.600
1st Qu.: 6.300       1st Qu.:0.2100      1st Qu.:0.2700      1st Qu.: 1.700
Median : 6.800       Median :0.2600      Median :0.3200      Median : 5.200
Mean   : 6.855       Mean   :0.2782      Mean   :0.3342      Mean   : 6.391
3rd Qu.: 7.300       3rd Qu.:0.3200      3rd Qu.:0.3900      3rd Qu.: 9.900
Max.   :14.200       Max.   :1.1000      Max.   :1.6600      Max.   :65.800
chlorides          free.sulfur.dioxide total.sulfur.dioxide density
Min. :0.00900       Min. : 2.00          Min. : 9.0          Min. :0.9871
1st Qu.:0.03600     1st Qu.: 23.00        1st Qu.:108.0        1st Qu.:0.9917
Median :0.04300     Median : 34.00        Median :134.0        Median :0.9937
Mean   :0.04577     Mean   : 35.31        Mean   :138.4         Mean   :0.9940
3rd Qu.:0.05000     3rd Qu.: 46.00        3rd Qu.:167.0         3rd Qu.:0.9961
Max.   :0.34600     Max.   :289.00        Max.   :440.0         Max.   :1.0390
pH                 sulphates          alcohol           quality
Min. :2.720          Min. :0.2200       Min. : 8.00        Min. :3.000
1st Qu.:3.090          1st Qu.:0.4100      1st Qu.: 9.50        1st Qu.:5.000
Median :3.180          Median :0.4700      Median :10.40        Median :6.000
Mean   :3.188          Mean   :0.4898      Mean   :10.51        Mean   :5.878
3rd Qu.:3.280          3rd Qu.:0.5500      3rd Qu.:11.40        3rd Qu.:6.000
Max.   :3.820          Max.   :1.0800      Max.   :14.20        Max.   :9.000
```

Οι δύο μεταβλητές  
μετρώνται σε πολύ<sup>μεγαλύτερες</sup>  
μονάδες από όλες<sup>τις υπόλοιπες</sup>

pcx <- prcomp(wine, scale=TRUE)

# Εφαρμογή του PCA στην Πράξη (6/8)

Έλεγχος κατανομής τιμών για τη μεταβλητή: quality

```
table( wine$quality )
```

3	4	5	6	7	8	9
20	163	1457	2198	880	175	5

Έλεγχος του φάσματος ιδιοτιμών για το σύνολο δεδομένων (μετά το scaling):

```
summary( pcx )
```

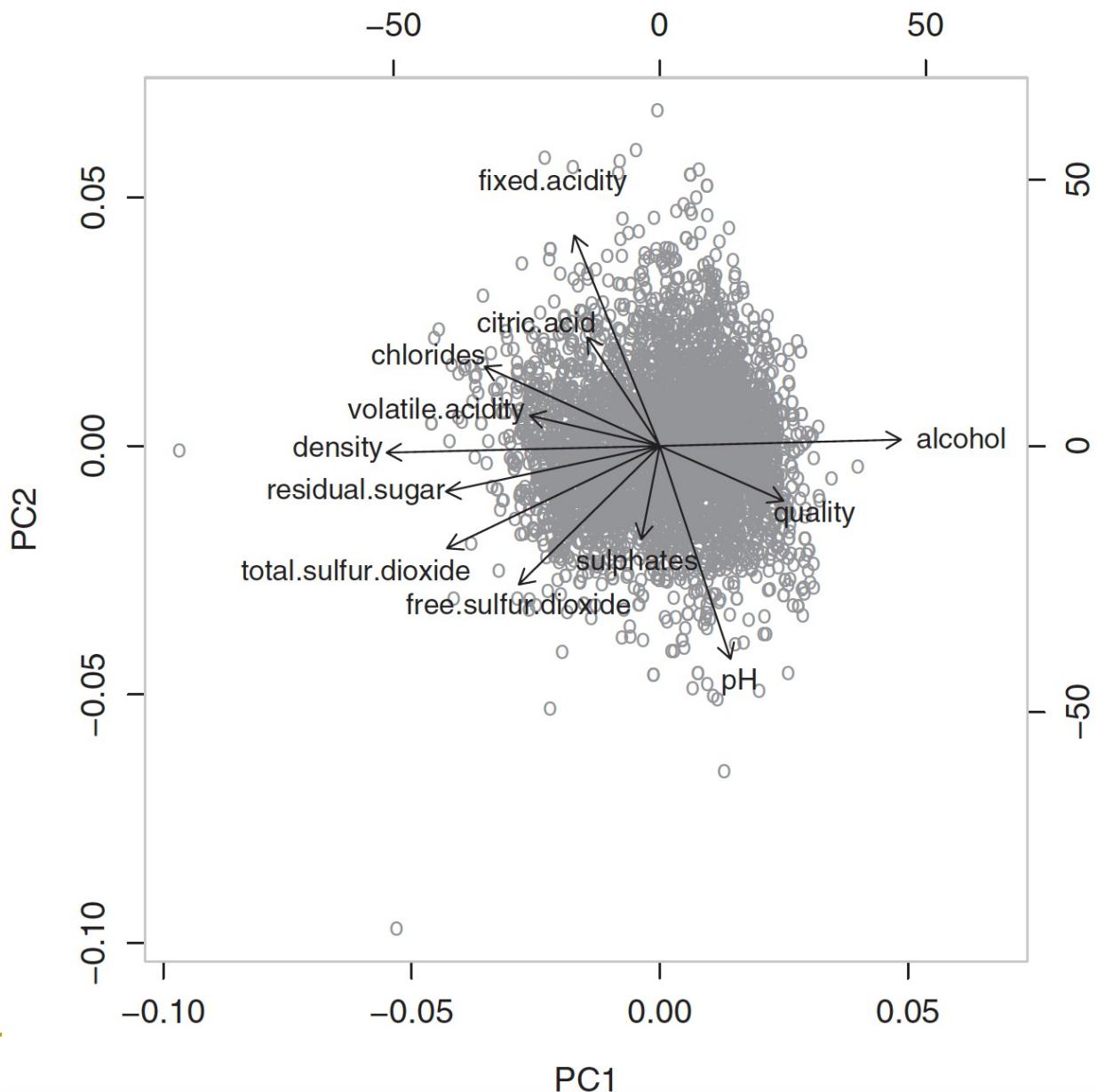
Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.829	1.259	1.171	1.0416	0.9876	0.9689
Proportion of Variance	0.279	0.132	0.114	0.0904	0.0813	0.0782
Cumulative Proportion	0.279	0.411	0.525	0.6157	0.6970	0.7752
	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	0.8771	0.8508	0.7460	0.5856	0.5330	0.14307
Proportion of Variance	0.0641	0.0603	0.0464	0.0286	0.0237	0.00171
Cumulative Proportion	0.8393	0.8997	0.9460	0.9746	0.9983	1.00000

# Εφαρμογή του PCA στην Πράξη (7/8)

`biplot( pcx )`

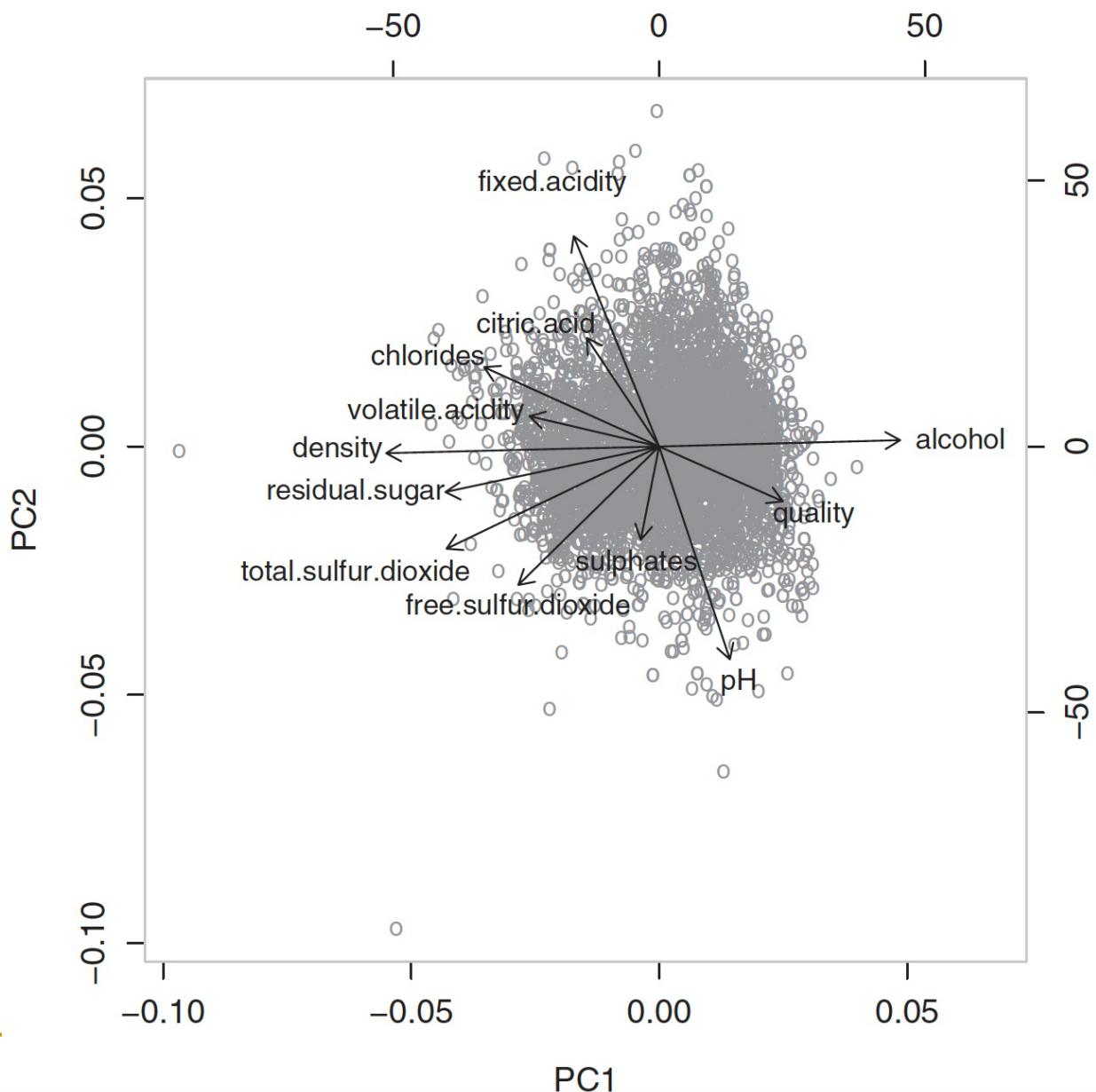
- **Alcohol, sugar, density:** παράλληλες στο πρώτο ιδιοδιάνυσμα (δείτε και τις κατευθύνσεις)
- Η αλκοόλη έχει χαμηλότερη πυκνότητα από το νερό, ενώ τα σάκχαρα υψηλότερη
- Επομένως, ο PCA μάς υπενθυμίζει ότι αυτές οι 3 μεταβλητές δεν είναι ανεξάρτητες
- Ομοίως για: **pH, acidity, citric acid** (λογικό καθώς το pH μετρά την οξύτητα)



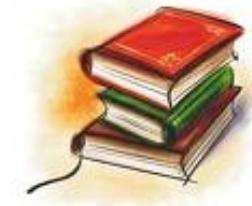
# Εφαρμογή του PCA στην Πράξη (8/8)

biplot( pcx )

- Τελικά, η ποιότητα (quality) εξαρτάται κυρίως από το περιεχόμενο σε αλκοόλ (alcohol) και από την οξύτητα (acidity)
- Όσο μεγαλύτερη η περιεκτικότητα σε αλκοόλ και όσο λιγότερη οξύτητα σε ένα κρασί, τόσο υψηλότερη η αξιολόγησή του (ποιότητα)
- Ο PCA βοηθά στην *ερμηνεία και κατανόηση του συνόλου δεδομένων*



# Περιεχόμενα



- *Data Analysis with Open Source Tools, by Philipp K. Janert. 2011, ISBN: 978-0-596-80235-6*
  - Κεφάλαιο 14
- “Εισαγωγή στην Εξόρυξη Δεδομένων”, 1<sup>η</sup> έκδοση, *Παράρτημα B*
  - P-N.Tan, M.Steinbach, V.Kumar
    - Εκδόσεις Τζιόλα
- Smith, L. I. (2002). A tutorial on Principal Components Analysis. Department of Computer Science, University of Otago.  
<https://hdl.handle.net/10523/7534>



Πανεπιστήμιο Πειραιώς  
Σχολή Τεχνολογιών Πληροφορικής και Επικοινωνιών  
Τμήμα Ψηφιακών Συστημάτων

# 10. Ανίχνευση Ανωμαλιών



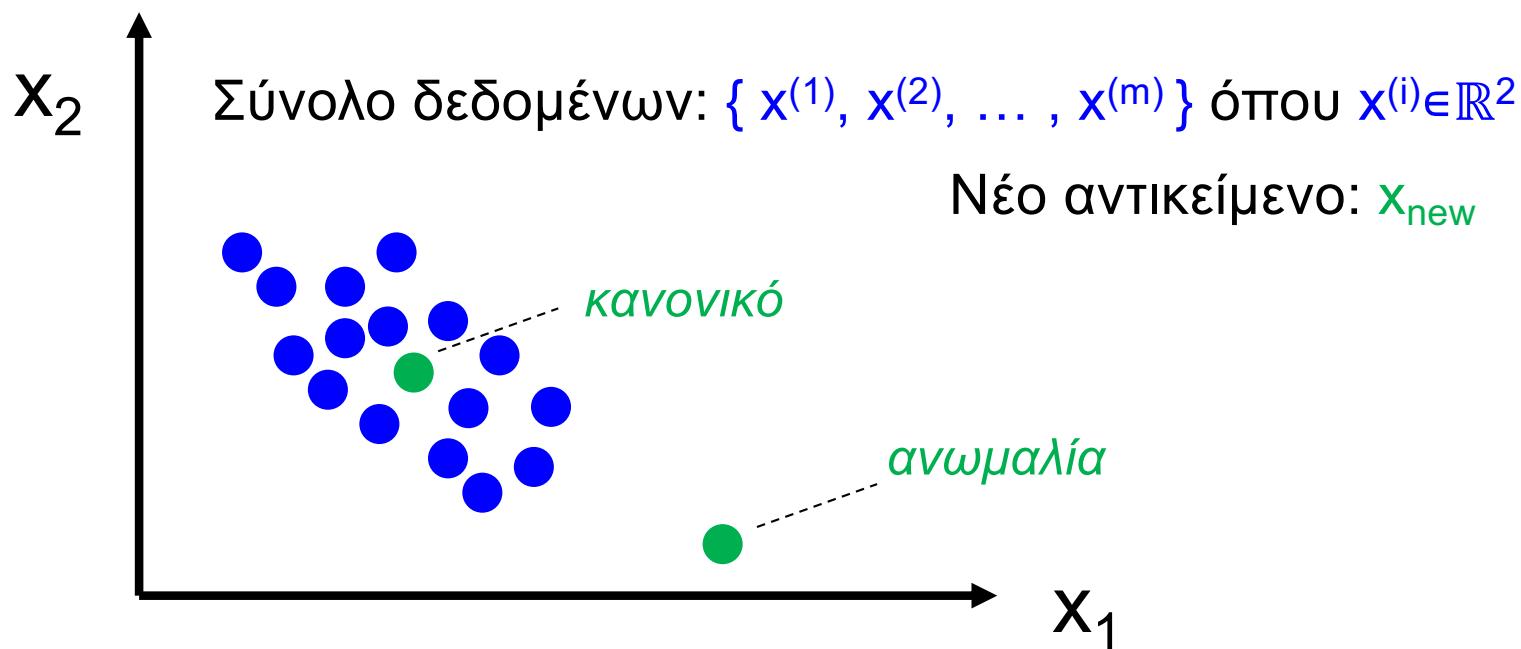
---

Ανάλυση Δεδομένων  
(*Data Analytics*)

Χρήστος Δουλκερίδης  
2024-25

# Εισαγωγή

- Ο στόχος της **ανίχνευσης ανωμαλιών** (**anomaly detection**) είναι η εύρεση αντικειμένων που **δεν ακολουθούν συνηθισμένα υποδείγματα και συμπεριφορά**
- Τα **μη ομαλά αντικείμενα** είναι γνωστά και ως **ακραίες τιμές** (**outliers**)



# Εφαρμογές

- **Ανίχνευση απάτης**
  - Υποδείγματα αγορών που αποκλίνουν από την τυπική συμπεριφορά
- **Ανίχνευση εισβολών**
  - Ασυνήθιστη συμπεριφορά υπολογιστών ή δικτύων
- **Διαταραχές οικοσυστήματος**
  - Προσδιορισμός ακραίων φαινομένων μέσω αισθητήρων
- **Δημόσια υγεία**
  - Ασυνήθιστα συμπτώματα ή αποτελέσματα ιατρικών εξετάσεων
- **Ασφάλεια πτήσεων**
  - Προσδιορισμός μιας σειράς μη ομαλών γεγονότων που καταγράφονται στη διάρκεια μιας πτήσης

# Περιεχόμενα Διάλεξης



- Χαρακτηριστικά προβλημάτων ανίχνευσης ανωμαλιών
- Χαρακτηριστικά μεθόδων ανίχνευσης ανωμαλιών
- Στατιστικές προσεγγίσεις
- Προσεγγίσεις βάσει εγγύτητας
- Προσεγγίσεις βάσει πυκνότητας
- Προσεγγίσεις βάσει συσταδοποίησης
- Αξιολόγηση ανίχνευσης ανωμαλιών

# Ορισμός και Παρατηρήσεις

**Ορισμός:** Μια **ανωμαλία** είναι μία παρατήρηση, η οποία δε συμφωνεί με την κατανομή των δεδομένων για συνηθισμένα στιγμιότυπα.

- Ο ορισμός **δεν** υποθέτει ότι η κατανομή μπορεί να εκφραστεί με όρους γνωστών στατιστικών κατανομών
- **Βαθμός ανωμαλίας:** βάσει πιθανότητας να βλέπουμε ένα αντικείμενο ή κάτι ακραίο
- Ενδέχεται να υπάρχουν **διάφορες περιπτώσεις** ανωμαλίας
  - **Θόρυβος**, αντικείμενο από **άλλη κατανομή**, ή απλά μια **σπάνια τιμή**
  - Δε μας ενδιαφέρουν οι ανωμαλίες που οφείλονται σε θόρυβο

# Φύση των Δεδομένων

- Η φύση των δεδομένων παίζει σημαντικό ρόλο στην **επιλογή κατάλληλης τεχνικής** ανίχνευσης ανωμαλιών
- Συνηθισμένα χαρακτηριστικά των δεδομένων:
  - Πλήθος χαρακτηριστικών
  - Τύπος χαρακτηριστικών
  - Αναπαράσταση για περιγραφή κάθε στιγμιότυπου

# Φύση των Δεδομένων

Μονομεταβλητά ή πολυμεταβλητά δεδομένα

## ■ Μονομεταβλητά δεδομένα

- Εύκολη περίπτωση, εξετάζεται αν η τιμή είναι μη ομαλή

## ■ Πολυμεταβλητά δεδομένα

- Πιο δύσκολο πρόβλημα – ένα αντικείμενο μπορεί να έχει ανώμαλες τιμές σε ορισμένα χαρακτηριστικά, και συνηθισμένες σε άλλα
- Επίσης, μπορεί καθεμιά από τις επιμέρους τιμές χαρακτηριστικών να είναι ομαλή, π.χ.:
  - Άνθρωπος (παιδί) ύψους 1 μέτρο – **συνηθισμένο**
  - Άνθρωπος βάρους 100 κιλά – **συνηθισμένο**
  - Άνθρωπος ύψους 1 μέτρο και βάρους 100 κιλά – **ασυνήθιστο**

# Φύση των Δεδομένων

## Δεδομένα εγγραφών ή μήτρα εγγύτητας

- Για τους σκοπούς της ανίχνευσης ανωμαλιών αρκεί να γνωρίζουμε πόσο διαφορετικό είναι ένα στιγμιότυπο σε σχέση με άλλα
- Άλλες μέθοδοι λειτουργούν απευθείας στο **σύνολο εγγραφών**
- Άλλες μέθοδοι προϋποθέτουν μια **μήτρα εγγύτητας**
  - Κάθε καταχώρηση δείχνει την εγγύτητα δύο αντικειμένων

# Φύση των Δεδομένων

## Διαθεσιμότητα ετικετών

- Αν υπάρχουν διαθέσιμες ετικέτες (ομαλό/μη ομαλό)
  - Τότε το πρόβλημα μετατρέπεται σε **πρόβλημα κατηγοριοποίησης** (με **επίβλεψη**)
  - Κυρίως ενδιαφέρουν τεχνικές που διαχειρίζονται **το πρόβλημα των σπάνιων κατηγοριών** (καθώς οι ανωμαλίες είναι σπάνιες)
- Στην πράξη μπορεί να **μην υπάρχουν ετικέτες διαθέσιμες**
  - Άρα απαιτούνται **μέθοδοι χωρίς επίβλεψη** (**σε αυτές εστιάζουμε**)
  - Με δεδομένη την απουσία ετικετών, είναι δύσκολο να διακρίνουμε ομαλά από ανώμαλα αντικείμενα
  - Χρησιμοποιούνται οι εξής **ιδιότητες** των ανωμαλιών:
    - Είναι λίγες σε πλήθος
    - Κατανέμονται αραιά στο χώρο

# Περιεχόμενα Διάλεξης



- Χαρακτηριστικά προβλημάτων ανίχνευσης ανωμαλιών
- Χαρακτηριστικά μεθόδων ανίχνευσης ανωμαλιών
- Στατιστικές προσεγγίσεις
- Προσεγγίσεις βάσει εγγύτητας
- Προσεγγίσεις βάσει πυκνότητας
- Προσεγγίσεις βάσει συσταδοποίησης
- Αξιολόγηση ανίχνευσης ανωμαλιών

# Χαρακτηριστικά Μεθόδων

- Σε υψηλό επίπεδο, τα συνηθέστερα **χαρακτηριστικά** μεθόδων ανίχνευσης ανωμαλιών:
  1. Προσεγγίσεις βάσει μοντέλου ή χωρίς μοντέλο
  2. Γενικές εναντίον τοπικών προσεγγίσεων
  3. Ετικέτα εναντίον βαθμού ανωμαλίας
- Χρησιμεύουν στην κατανόηση των κοινών τους στοιχείων και των διαφορών

# Προσεγγίσεις με Χρήση Μοντέλου vs. Προσεγγίσεις Χωρίς Μοντέλο

- **Προσεγγίσεις με χρήση μοντέλου**
  - Δημιουργούν **μοντέλο** της **συνηθισμένης κατηγορίας**
  - **Ανωμαλία είναι ό,τι δεν προσαρμόζεται στο μοντέλο**
  - Μια άλλη μορφή τεχνικών, εκπαιδεύει ένα **μοντέλο ομαλών** και ένα **ανώμαλων κατηγοριών**, και προσδιορίζει ένα αντικείμενο ως ανωμαλία αν είναι πιθανότερο να ανήκει στη δεύτερη κατηγορία
- **Προσεγγίσεις χωρίς μοντέλο**
  - Δε χαρακτηρίζουν ρητά την κατανομή ομαλών/ανώμαλων κατηγοριών
  - Λειτουργούν χωρίς ετικέτες στα δεδομένα (χωρίς δεδομένα εκπαίδευσης)
  - Για παράδειγμα, ένα αντικείμενο χαρακτηρίζεται ως ανώμαλο αν είναι διαφορετικό από άλλα αντικείμενα στη γειτονιά του

# Γενικές vs. Τοπικές Προσεγγίσεις

## ■ Γενική προσέγγιση

- Εξετάζει το γενικό περιβάλλον

## ■ Τοπική προσέγγιση

- Η έξοδός της για δοθέν στιγμιότυπο δεν αλλάζει αν αφαιρεθούν ή αλλάξουν τα στιγμιότυπα που βρίσκονται έξω από τη γειτονιά αυτού του στιγμιότυπου

## ■ Έχουν σημαντική διαφορά

- Ένα αντικείμενο μπορεί να φαίνεται μη συνηθισμένο ως προς το γενικό σύνολο αντικειμένων, αλλά όχι προς τα αντικείμενα μιας γειτονιάς
- Παράδειγμα: άτομο ύψους 1.70 είναι συνηθισμένο γενικά, αλλά όχι για ομάδα καλαθοσφαίρισης

# Ετικέτα vs. Βαθμός Ανωμαλίας

- Διαφορετικές προσεγγίσεις ανίχνευσης ανωμαλιών  
**παράγουν την έξοδο σε διαφορετική μορφή**
  - Δυαδική ετικέτα ανωμαλίας
    - Δεν παρέχουν πληροφορίες για το βαθμό ανωμαλίας
    - Κάποιες ανωμαλίες είναι πιο ακραίες από άλλες
  - Βαθμός ανωμαλίας
    - Επιτρέπει την κατάταξη των αντικειμένων
    - Μπορεί να εφαρμοστεί ένα κατώφλι περικοπής

# Περιεχόμενα Διάλεξης



- Χαρακτηριστικά προβλημάτων ανίχνευσης ανωμαλιών
- Χαρακτηριστικά μεθόδων ανίχνευσης ανωμαλιών
- Στατιστικές προσεγγίσεις
- Προσεγγίσεις βάσει εγγύτητας
- Προσεγγίσεις βάσει πυκνότητας
- Προσεγγίσεις βάσει συσταδοποίησης
- Αξιολόγηση ανίχνευσης ανωμαλιών

# Στατιστικές Προσεγγίσεις

- Οι στατιστικές προσεγγίσεις χρησιμοποιούν **κατανομές πιθανοτήτων** (π.χ. κανονική) για να **μοντελοποιούν** την ομαλή κατηγορία
- Έτσι, κάθε στιγμιότυπο δεδομένων σχετίζεται με μια **τιμή πιθανότητας**

**Ορισμός:** Οι ανωμαλίες ορίζονται ως στιγμιότυπα τα οποία είναι απίθανο να έχουν παραχθεί από την κατανομή πιθανότητας της ομαλής κατηγορίας

στιγμιότυπα  $x$  με:  $p(x) < \varepsilon \rightarrow x$ : ανωμαλία

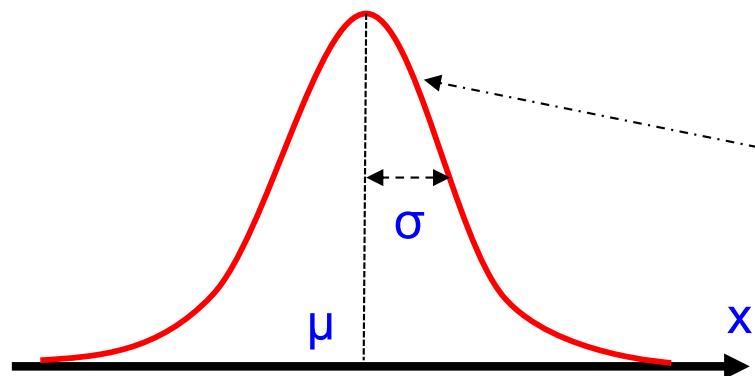
- Δύο κατηγορίες/τύποι μοντέλων:
  - **Παραμετρικά μοντέλα:** χρησιμοποιούν γνωστές στατιστικές κατανομές, εκτιμούν τις παραμέτρους από τα δεδομένα
  - **Μη παραμετρικά μοντέλα:** μαθαίνουν την κατανομή της ομαλής κατηγορίας άμεσα από τα διαθέσιμα δεδομένα

# Παραμετρικά Μοντέλα

- Είναι ιδιαίτερα αποτελεσματικά, **εφόσον** η ομαλή κατηγορία ακολουθεί μια συγκεκριμένη κατανομή
  - Όπως: κατανομή Gauss, Poisson, διωνυμική κατανομή
- Σε αυτή την περίπτωση, οι βαθμοί ανωμαλίας που υπολογίζονται έχουν ισχυρές θεωρητικές ιδιότητες
  - Οι οποίες μπορούν να αναλυθούν για την αξιολόγηση της στατιστικής σημαντικότητας
- Ας εξετάσουμε τη χρήση της κατανομής Gauss για τη μοντελοποίηση της κανονικής κατηγορίας σε μονομεταβλητά και πολυμεταβλητά δεδομένα

# Κατανομή Gauss / Κανονική Κατανομή

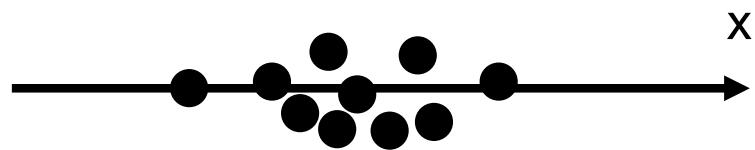
Έστω  $x \in \mathbb{R}$  ακολουθεί κανονική κατανομή με μέσο  $\mu$  και διακύμανση  $\sigma^2$   
 $x \sim N(\mu, \sigma)$



Συνάρτηση πυκνότητας πιθανότητας

$$p(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-(x-\mu)^2/2\sigma^2}$$

Για σύνολο δεδομένων:  $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$  όπου  $x^{(i)} \in \mathbb{R}$



$$\mu = \frac{1}{m} \sum_{i=1}^m x^{(i)}$$

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu)^2$$

# Κατανομή Gauss / Κανονική Κατανομή

Για σύνολο δεδομένων:  $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$  όπου  $x^{(i)} \in \mathbb{R}^n$

Έστω  $x_i \in \mathbb{R}$  ακολουθεί κανονική κατανομή με μέσο  $\mu_i$  και διακύμανση  $\sigma_i^2$

$$x_1 \sim N(\mu_1, \sigma_1)$$

$$x_2 \sim N(\mu_2, \sigma_2)$$

...

$$x_n \sim N(\mu_n, \sigma_n)$$

$$p(x) = p(x_1)p(x_2)\dots p(x_n) = \prod_{i=1}^n p(x_i)$$

# Παραμετρικά Μοντέλα

## Αλγόριθμος Ανίχνευσης Ανωμαλιών

### ■ Δοθέντος ενός συνόλου αντικειμένων

- Επιλογή ή κατασκευή των χαρακτηριστικών  $x_j$  που θα χρησιμοποιηθούν για ανίχνευση ανωμαλιών
- Υπολογισμός παραμέτρων:  $\mu_1, \dots, \mu_n, \sigma^2_1, \dots, \sigma^2_n$

$$\mu_j = \frac{1}{m} \sum_{i=1}^m x_j^{(i)}$$

$$\sigma_j^2 = \frac{1}{m} \sum_{i=1}^m (x_j^{(i)} - \mu_j)^2$$

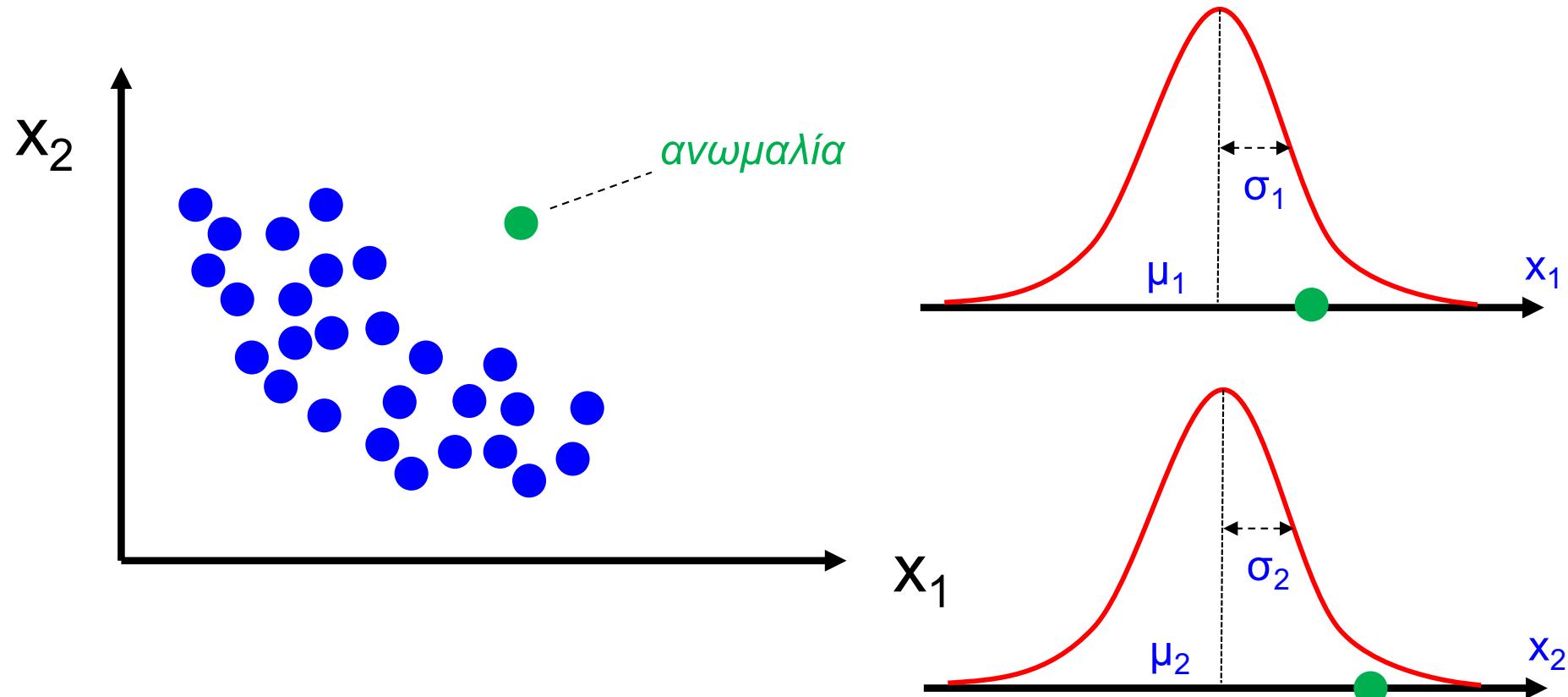
### ■ Για ένα νέο αντικείμενο: $x$

- Υπολογισμός πιθανότητας  $p(x)$

$$p(x) = \prod_{i=j}^n \frac{1}{\sqrt{2\pi} \cdot \sigma_j} e^{-(x_j - \mu_j)^2 / 2\sigma_j^2}$$

- Αν  $p(x) < \varepsilon \rightarrow$  το  $x$  χαρακτηρίζεται ως ανωμαλία

# Πολυμεταβλητή Κατανομή Gauss



# Πολυμεταβλητή Κατανομή Gauss

Για σύνολο δεδομένων:  $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$  όπου  $x^{(i)} \in \mathbb{R}^n$

Δε θα μοντελοποιήσουμε χωριστά τις:  $p(x_1), p(x_2), \dots$ , αλλά ταυτόχρονα

Παράμετροι:  $\mu \in \mathbb{R}^n$  και  $\Sigma \in \mathbb{R}^{n \times n}$  (πίνακας συνδιακύμανσης)

$$p(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

# Μη Παραμετρικά Μοντέλα

- Εκτίμηση πυκνότητας βάσει πυρήνα (kernel density estimate)
  - Προσεγγίζει την κατανομή της κανονικής κατηγορίας από τα διαθέσιμα δεδομένα
  - Ο **βαθμός ανωμαλίας** ενός στιγμιότυπου υπολογίζεται ως το **αντίστροφο της πιθανότητας**
- Κατασκευή ιστογράμματος
  - Με τεχνική διακριτοποίησης ίσου πλάτους
  - Αν ένα στιγμιότυπο δε βρίσκεται σε κάποιο τμήμα του ιστογράμματος → ανωμαλία
  - Αν βρίσκεται, τότε ο **βαθμός ανωμαλίας** είναι η **αντίστροφη τιμή του ύψους** (συχνότητα) του τμήματος όπου ανήκει το στιγμιότυπο (**ανίχνευση ανωμαλιών βάσει συχνότητας ή μέτρησης**)

# Μοντελοποίηση Ομαλών και Ανώμαλων Κατηγοριών

- Όταν υπάρχουν ακραίες τιμές στα δεδομένα εκπαίδευσης, μπορεί να προκύψει μια παραμορφωμένη κατανομή για την κανονική κατηγορία
- Συνδυαστική τεχνική μοντελοποίησης για να μάθει την κατανομή της κανονικής **και** της ανώμαλης κατηγορίας
  - Ανέχεται μια αναλογία **λ** ακραίων τιμών στο σύνολο εκπαίδευσης
  - Υπό την προϋπόθεση ότι αυτές είναι κανονικά κατανεμημένες (όχι συσταδοποιημένες) στο χώρο των χαρακτηριστικών

# Μοντελοποίηση Ομαλών και Ανώμαλων Κατηγοριών

- Βασική ιδέα/υπόθεση: το σύνολο δεδομένων **D** αποτελείται από ένα μείγμα δύο κατανομών
  - **M** (κατανομή πλειοψηφίας - majority)
  - **A** (κατανομή ανώμαλων στιγμιότυπων - anomalous)
- Αλγόριθμος
  - Αρχικά, υποθέτουμε ότι όλα τα σημεία ανήκουν στο **M**
  - Έστω  $L_t(D)$  η πιθανότητα του συνόλου δεδομένων **D** τη χρον. στιγμή **t**
  - Για κάθε τιμή  $x_t$  που ανήκει στο **M**, μετακίνησέ το στο **A**
    - Έστω  $L_{t+1}(D)$  η νέα πιθανότητα
    - Υπολογίζουμε τη διαφορά:  $\Delta = L_t(D) - L_{t+1}(D)$
    - Εάν το  $\Delta > c$  (κατώφλι), τότε το  $x_t$  είναι ανωμαλία και μετακινείται μόνιμα από το **M** στο **A**

# Μοντελοποίηση Ομαλών και Ανώμαλων Κατηγοριών

- Κατανομή δεδομένων:  $\mathbf{D} = (1 - \lambda) \mathbf{M} + \lambda \mathbf{A}$
- Η κατανομή πιθανοτήτων  $\mathbf{M}$  εκτιμάται από τα δεδομένα
- Η κατανομή  $\mathbf{A}$  υποθέτουμε στην αρχή ότι είναι η ομοιόμορφη κατανομή
- Η πιθανότητα τη χρονική στιγμή  $t$ :

$$L_t(D) = \prod_{i=1}^N P_D(x_i) = \left( (1 - \lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left( \lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1 - \lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

# Δυνατά Σημεία/Αδυναμίες Στατιστικών Προσεγγίσεων

- Σημαντικό θεωρητικό υπόβαθρο
- Σε πολλές περιπτώσεις είναι πολύ αποτελεσματικές
- Έχουν καλά αποτελέσματα, όταν η κατανομή είναι γνωστή
- Όμως συχνά τα δεδομένα μπορεί να προέρχονται από μια άγνωστη κατανομή
- Για πολυδιάστατα δεδομένα είναι δύσκολο να εκτιμηθεί η κατανομή που ακολουθούν

# Περιεχόμενα Διάλεξης



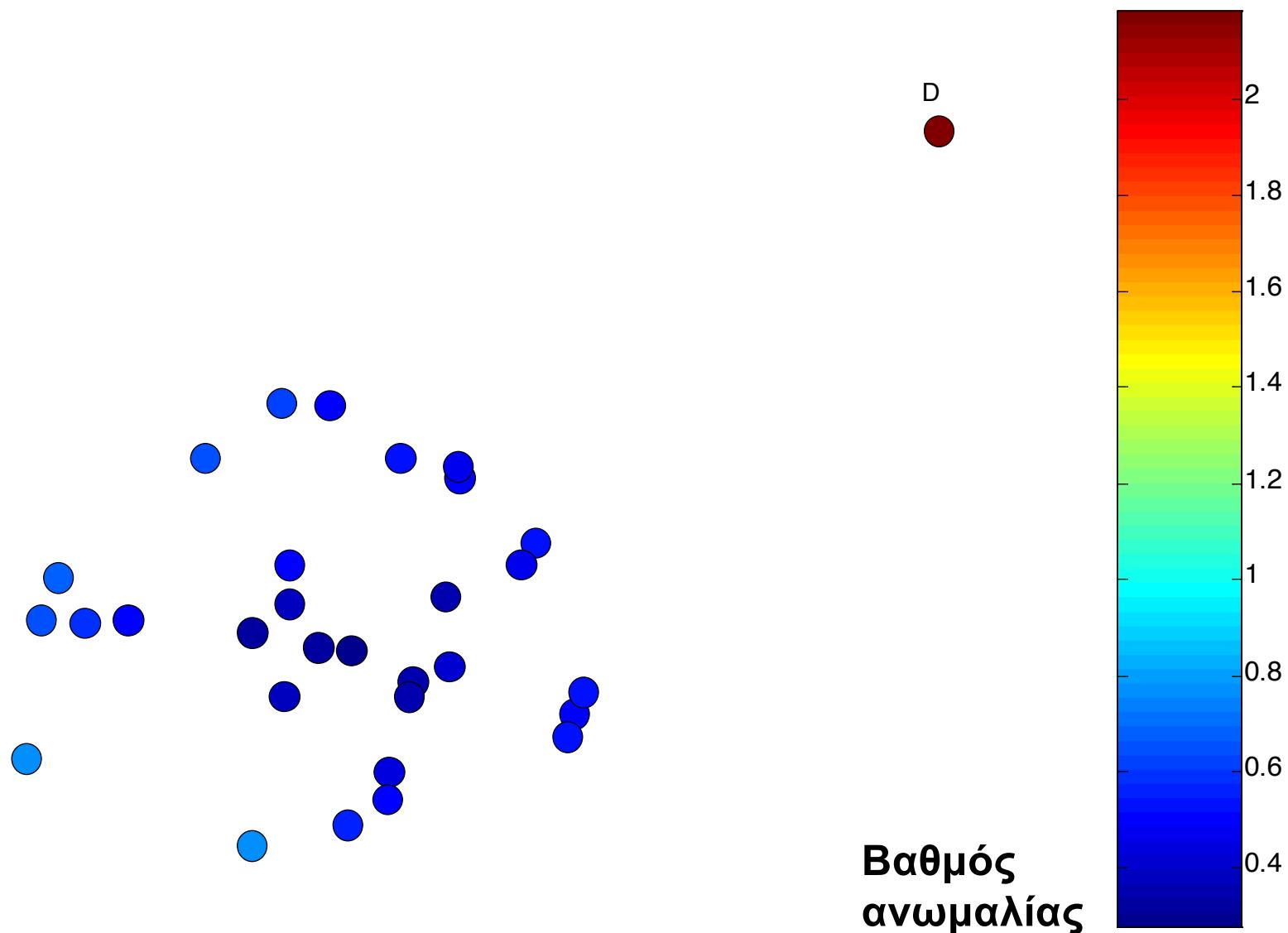
- Χαρακτηριστικά προβλημάτων ανίχνευσης ανωμαλιών
- Χαρακτηριστικά μεθόδων ανίχνευσης ανωμαλιών
- Στατιστικές προσεγγίσεις
- Προσεγγίσεις βάσει εγγύτητας
- Προσεγγίσεις βάσει πυκνότητας
- Προσεγγίσεις βάσει συσταδοποίησης
- Αξιολόγηση ανίχνευσης ανωμαλιών

# Απλός Ορισμός βάσει Εγγύτητας

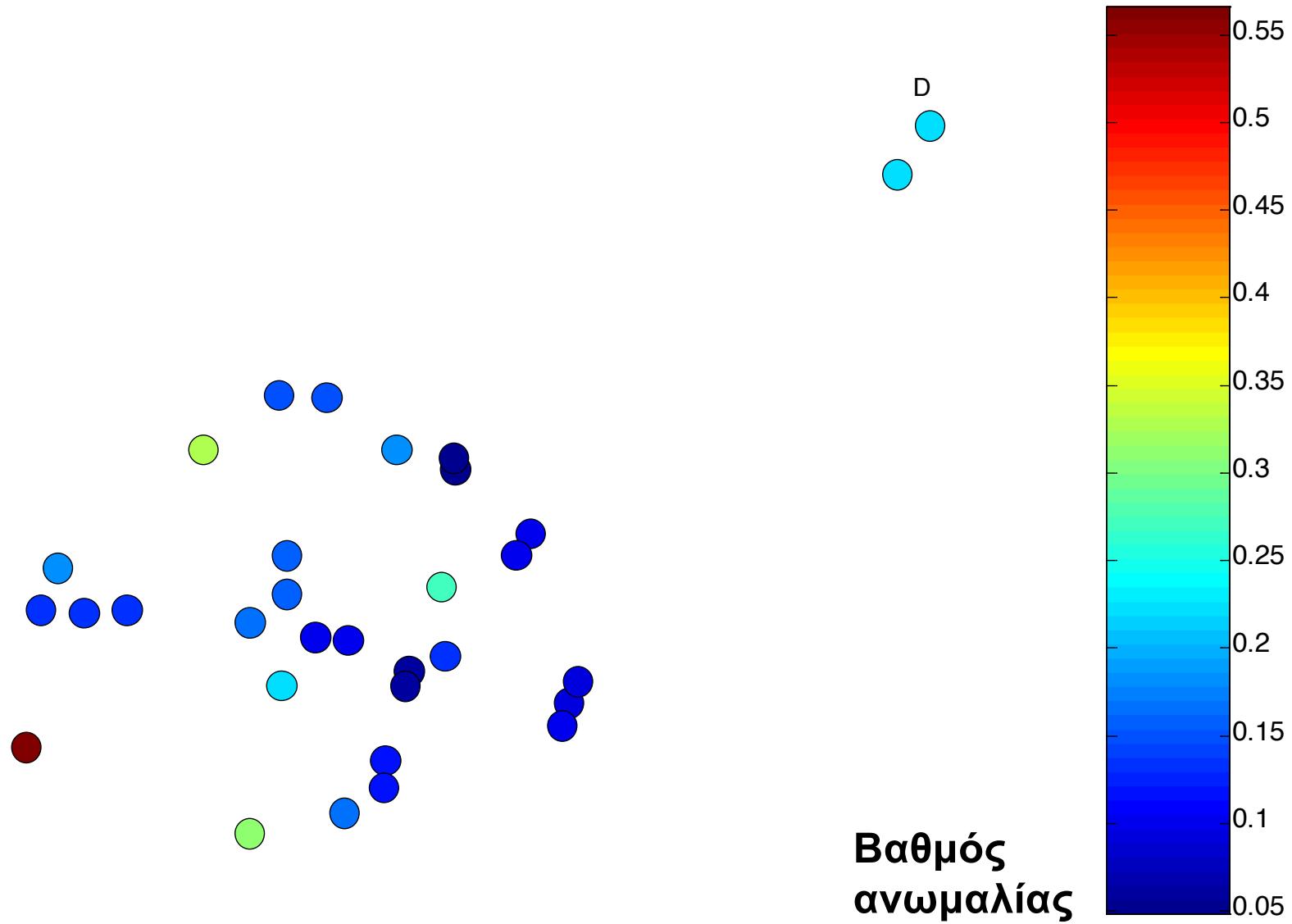
**Ορισμός:** Ο βαθμός ανωμαλίας ενός αντικειμένου μπορεί να οριστεί βάσει της **απόστασης από τον k-οστό κοντινότερο** γείτονά του

- Οι προσεγγίσεις βάσει εγγύτητας χαρακτηρίζουν ως **ανωμαλίες** εκείνα τα αντικείμενα που **βρίσκονται πιο μακριά από** άλλα αντικείμενα
- Πρόκειται για προσεγγίσεις χωρίς μοντέλο
- Απλά χρησιμοποιούν την **τοπικότητα** κάθε αντικειμένου για να υπολογίσουν το βαθμό ανωμαλίας
  - (1) Απόσταση από τον k-οστό κοντινότερο γείτονα: **dist( $\mathbf{x}$ , k)**
  - (2) Μέση απόσταση από τους k-κοντινότερους γείτονες: **avg.dist( $\mathbf{x}$ , k)**  
(πιο εύρωστη προσέγγιση από την επιλογή του k)

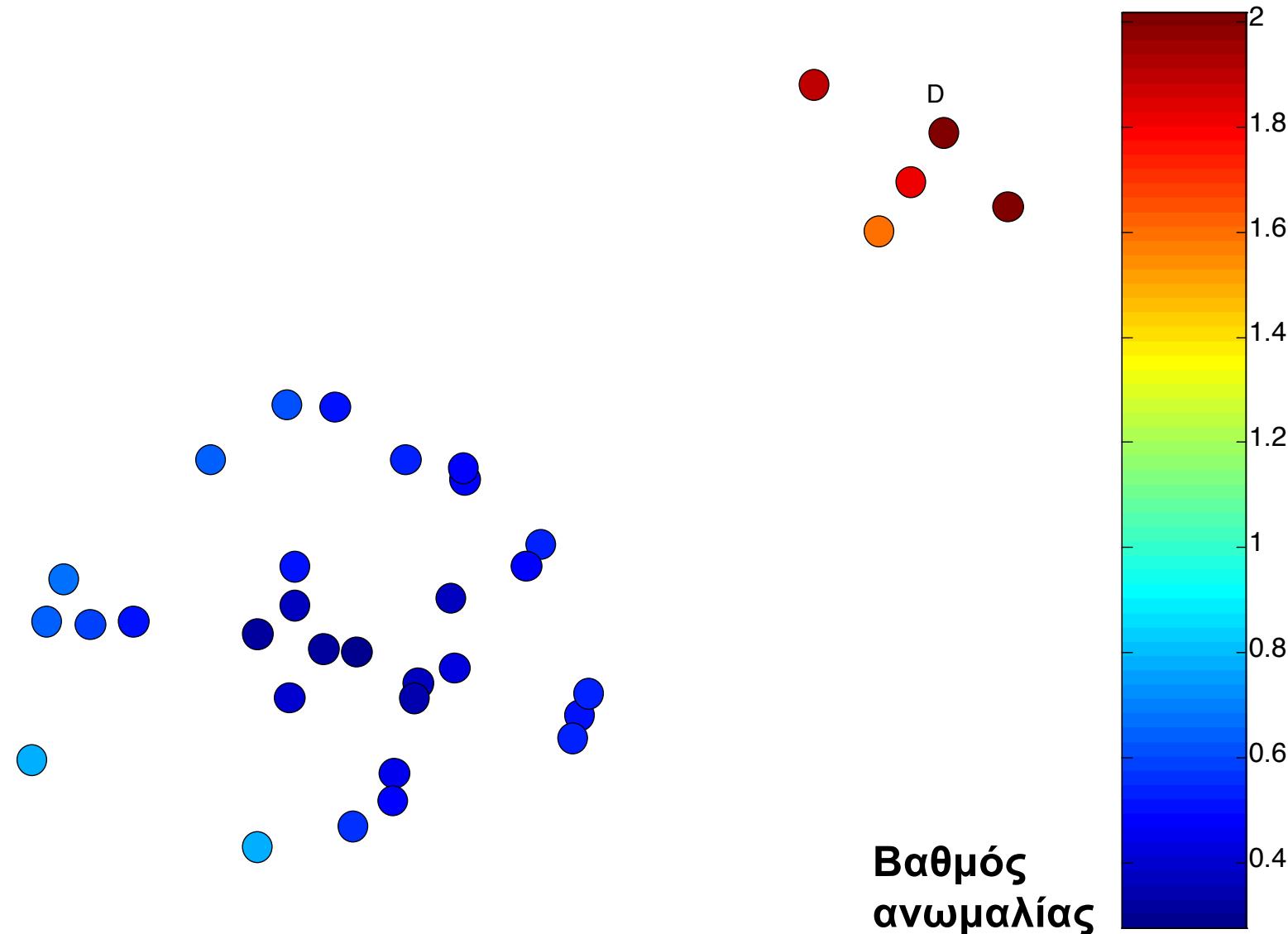
# 1-Κοντινότερος Γείτονας / 1 Outlier



# 1-Kοντινότερος Γείτονας / 2 Outliers

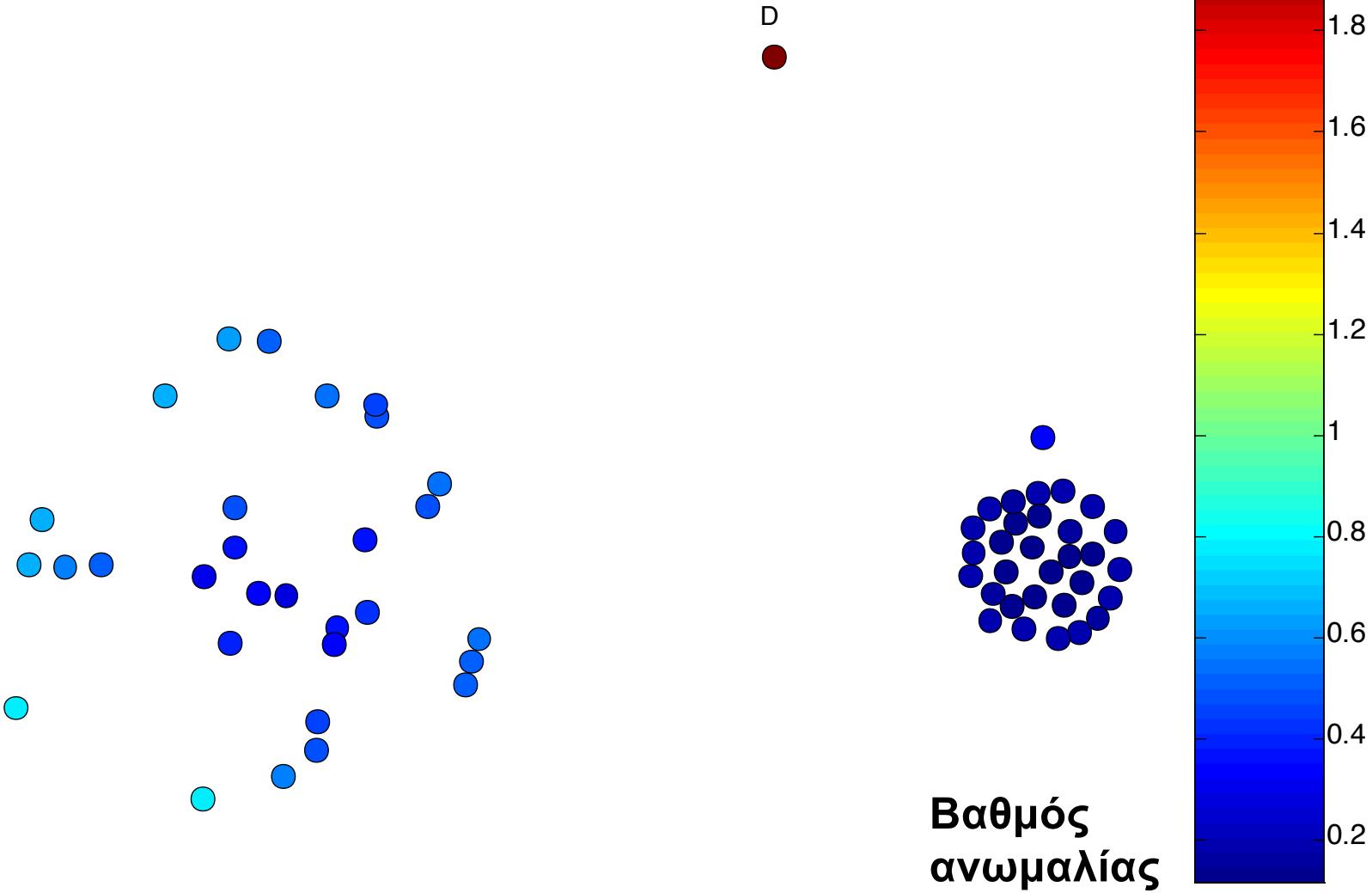


# 5-Κοντινότεροι Γείτονες / Μικρή Συστάδα



# 5-Κοντινότεροι Γείτονες / Συστάδες

## Διαφορετικής Πυκνότητας



# Δυνατά Σημεία/Αδυναμίες

## Προσεγγίσεων βάσει Απόστασης

- Είναι απλές
- Είναι υπολογιστικά ακριβές:  $O(n^2)$
- Ευαίσθητες στις παραμέτρους εισόδου ( $k$ )
- Ευαίσθητες σε διαφοροποιήσεις στην πυκνότητα
- Οι αποστάσεις δε βοηθούν πολύ όταν έχουμε πολυδιάστατα δεδομένα

# Περιεχόμενα Διάλεξης



- Χαρακτηριστικά προβλημάτων ανίχνευσης ανωμαλιών
- Χαρακτηριστικά μεθόδων ανίχνευσης ανωμαλιών
- Στατιστικές προσεγγίσεις
- Προσεγγίσεις βάσει εγγύτητας
- Προσεγγίσεις βάσει πυκνότητας
- Προσεγγίσεις βάσει συσταδοποίησης
- Αξιολόγηση ανίχνευσης ανωμαλιών

# Προσεγγίσεις βάσει Πυκνότητας

**Ορισμός:** Ο βαθμός ανωμαλίας ενός στιγμιότυπου είναι ανáλογος του αντίστροφου της πυκνότητας της γειτονιάς του

- Διαφορετικοί ορισμοί της πυκνότητας
  - Βάσει k-κοντινότερων γειτόνων:
    - Αντίστροφο απόστασης από k-κοντινότερο γείτονα
      - $\text{density}(\mathbf{x}, k) = 1 / \text{dist}(\mathbf{x}, k)$
    - Αντίστροφο της μέσης απόστασης από τους k-κοντινότερους γείτονες
      - $\text{avg.density}(\mathbf{x}, k) = 1 / \text{avg.dist}(\mathbf{x}, k)$
  - Βάσει ορισμού DBSCAN

# Σχετική Πυκνότητα

- Έστω πολυδιάστατο σημείο  $\mathbf{x}$  και έστω οι  $k$  κοντινότεροι γείτονες του:  $y_1, \dots, y_k$
- Η **σχετική πυκνότητα** βοηθά όταν υπάρχουν **περιοχές διαφορετικής πυκνότητας** στα δεδομένα

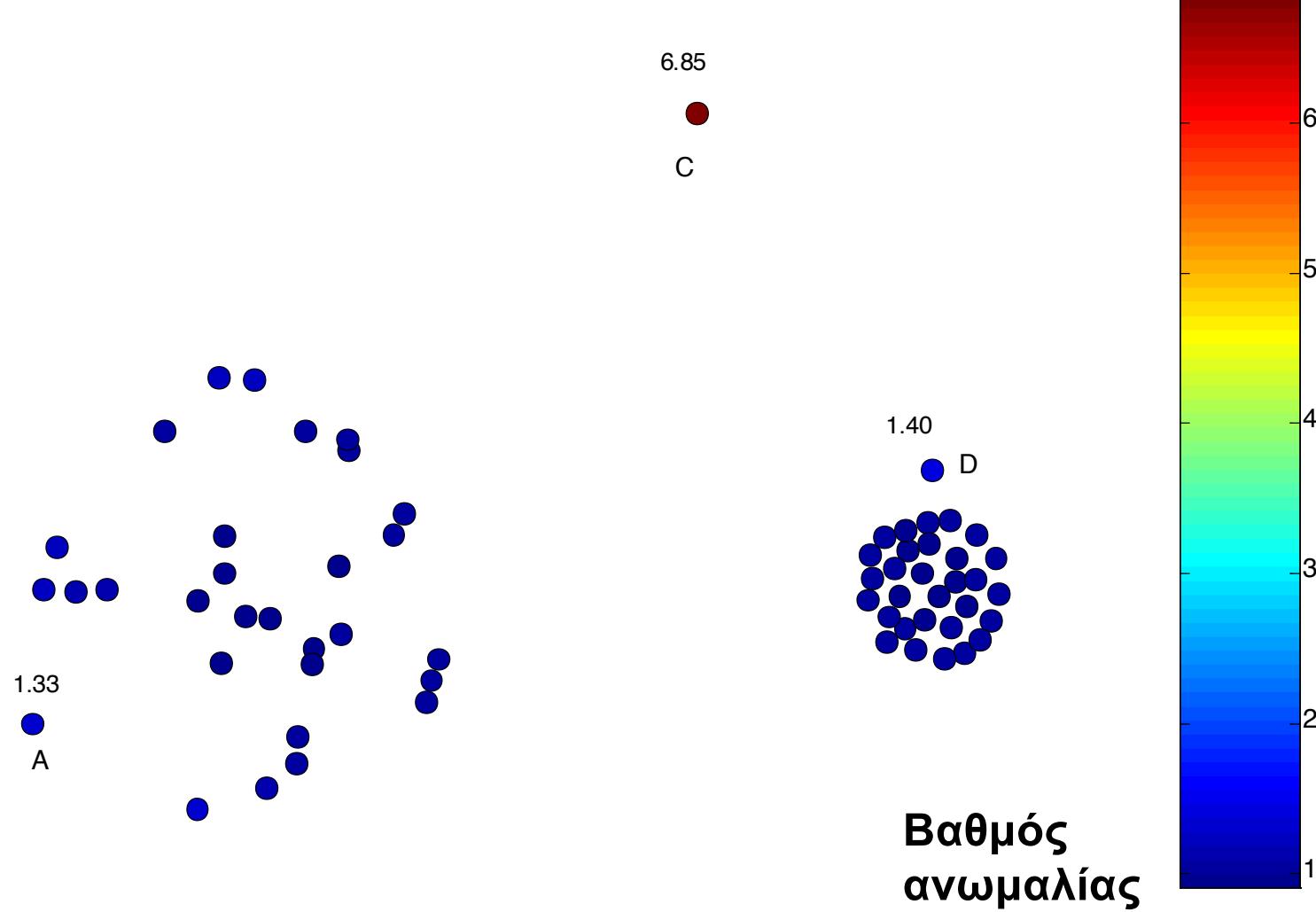
$$density(\mathbf{x}, k) = \frac{1}{dist(\mathbf{x}, k)} = \frac{1}{dist(\mathbf{x}, \mathbf{y}_k)}$$

$$\begin{aligned}relative\ density(\mathbf{x}, k) &= \frac{\sum_{i=1}^k density(\mathbf{y}_i, k)/k}{density(\mathbf{x}, k)} \\&= \frac{dist(\mathbf{x}, k)}{\sum_{i=1}^k dist(\mathbf{y}_i, k)/k} = \frac{dist(\mathbf{x}, \mathbf{y})}{\sum_{i=1}^k dist(\mathbf{y}_i, k)/k}\end{aligned}$$

- Εναλλακτικά, χρησιμοποιείται η μέση απόσταση

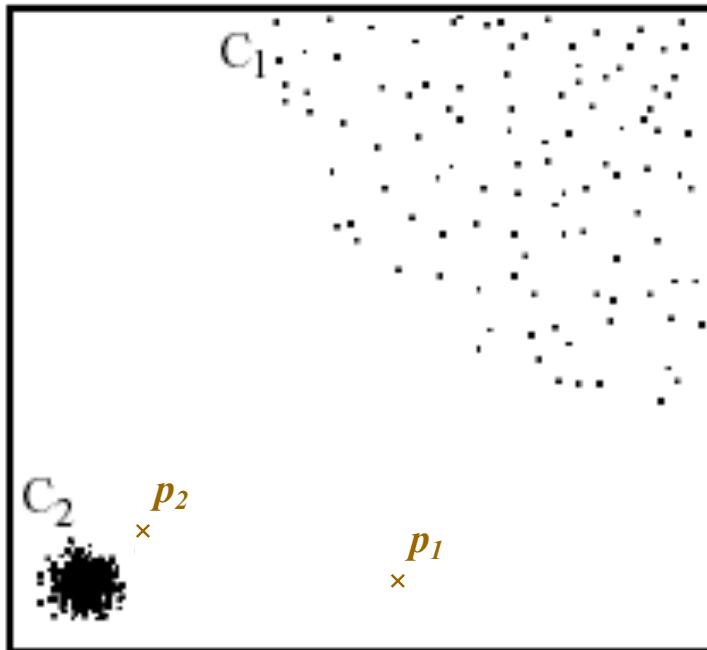
# Σχετική Πυκνότητα – Βαθμός Ανωμαλίας

Βάσει σχετικής πυκνότητας με  $k=10$



# Προσέγγιση LOF

- Για κάθε αντικείμενο υπολογίζεται η πυκνότητα της γειτονιάς του
- Υπολογισμός του Local Outlier Factor (LOF) ενός στιγμιότυπου  $p$  ως
  - Μέσος όρος πυκνότητας  $p$  και γειτόνων του
- Ανωμαλίες θεωρούνται τα σημεία με υψηλότερη τιμή LOF



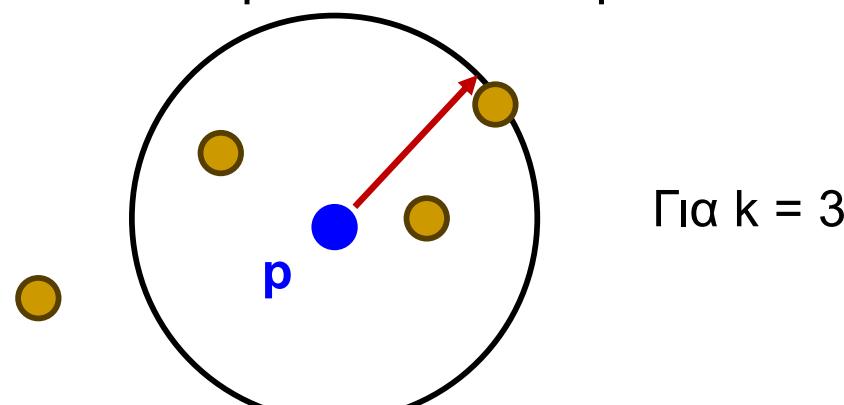
- Βάσει κοντινότερων γειτόνων:
  - το  $p_2$  δεν είναι ανωμαλία
- Βάσει LOF:
  - και το  $p_1$  και το  $p_2$  θεωρούνται ανωμαλίες

# LOF – Local Outlier Factor

- Μέσος όρος του λόγου **local reachability density** του σημείου **p** και των σημείων στην **γειτονιά** του **p**

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrд_k(o)}{lrд_k(p)}}{|N_k(p)|}$$

- $N_k(p)$ : οι **k** κοντινότεροι γείτονες του **p**
- **k-distance(p)**: η απόσταση του k-οστού γείτονα του **p**



# LOF – Local Outlier Factor

- **reachability distance** του  $p$  ως προς το  $o$ :
  - Ίση είτε με την **ακτίνα της γειτονιάς** του  $o$ , αν το  $p$  ανήκει σε αυτή ή
  - Ίση με την **απόσταση** του  $p$  από το  $o$
  - $\text{reach-dist}_k(p, o) = \max\{ \text{k-distance}(o), d(p,o) \}$
- **local reachability density**
  - Το αντίστροφο της μέσης reachability distance από τους  $k$ -κοντινότερους γείτονες του  $p$

$$lrd_k(p) = \frac{1}{\left[ \sum_{o \in N_k(p)} \text{reach-dist}_k(p, o) \right] / |N_k(p)|}$$

# Δυνατά Σημεία/Αδυναμίες

## Προσεγγίσεων βάσει Πυκνότητας

- Απλή προσέγγιση
- Υψηλό υπολογιστικό κόστος:  $O(n^2)$
- Ευαισθησία στις τιμές των παραμέτρων εισόδου
- Πρόβλημα σε πολυδιάστατα προβλήματα, όπου η πυκνότητα δεν έχει πολύ νόημα

# Περιεχόμενα Διάλεξης

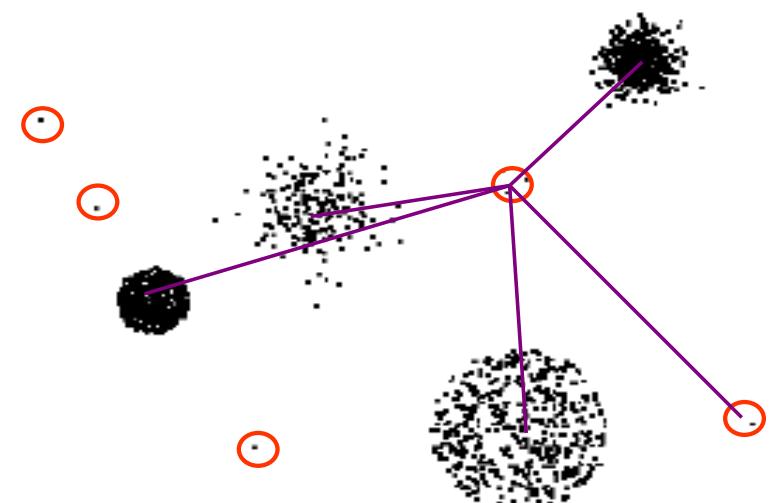


- Χαρακτηριστικά προβλημάτων ανίχνευσης ανωμαλιών
- Χαρακτηριστικά μεθόδων ανίχνευσης ανωμαλιών
- Στατιστικές προσεγγίσεις
- Προσεγγίσεις βάσει εγγύτητας
- Προσεγγίσεις βάσει πυκνότητας
- Προσεγγίσεις βάσει συσταδοποίησης
- Αξιολόγηση ανίχνευσης ανωμαλιών

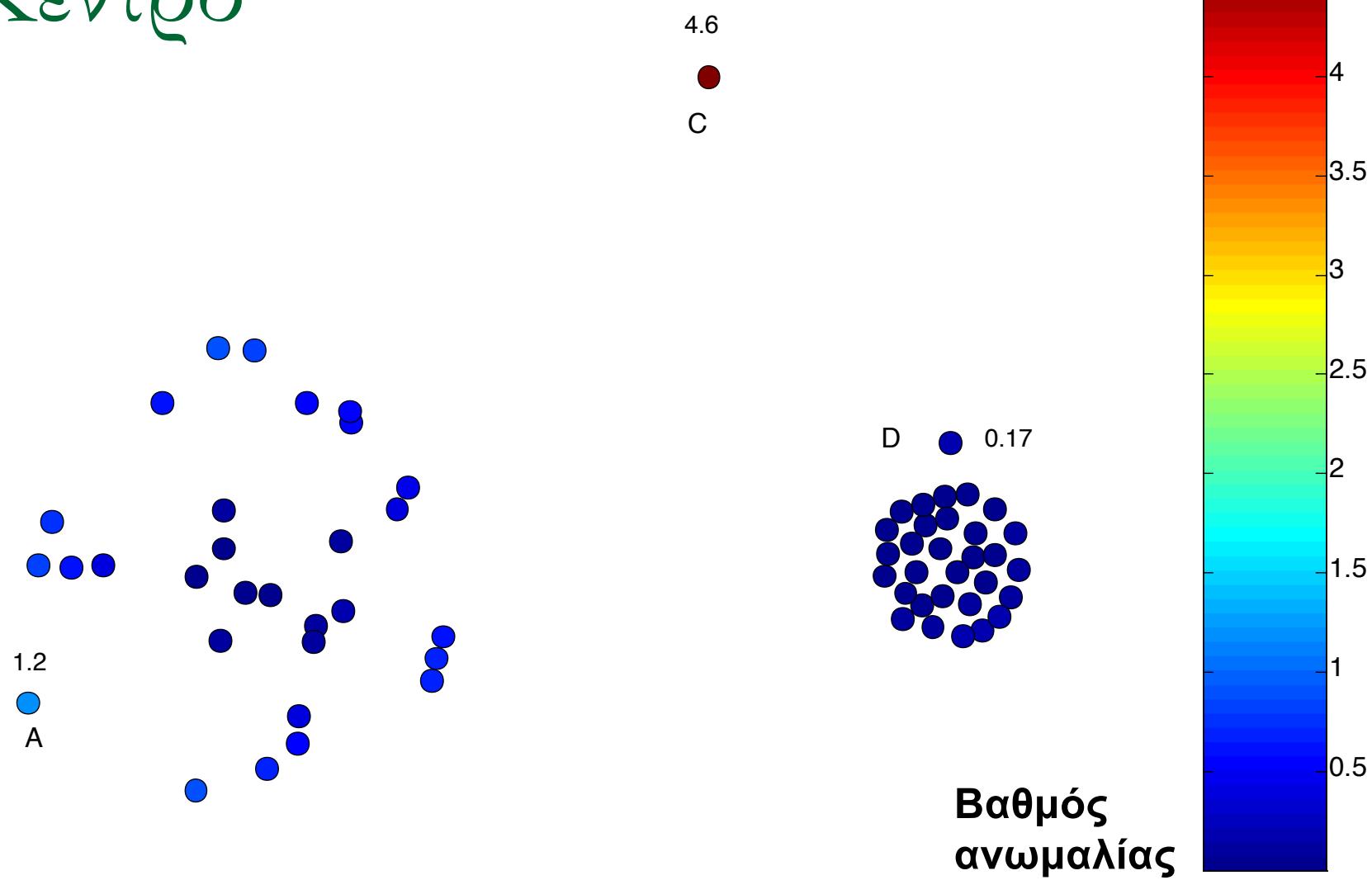
# Ορισμός βάσει Συσταδοποίησης

**Ορισμός:** Ένα στιγμιότυπο θεωρείται ανώμαλο αν δεν ανήκει σε κάποια συστάδα της ομαλής κατηγορίας

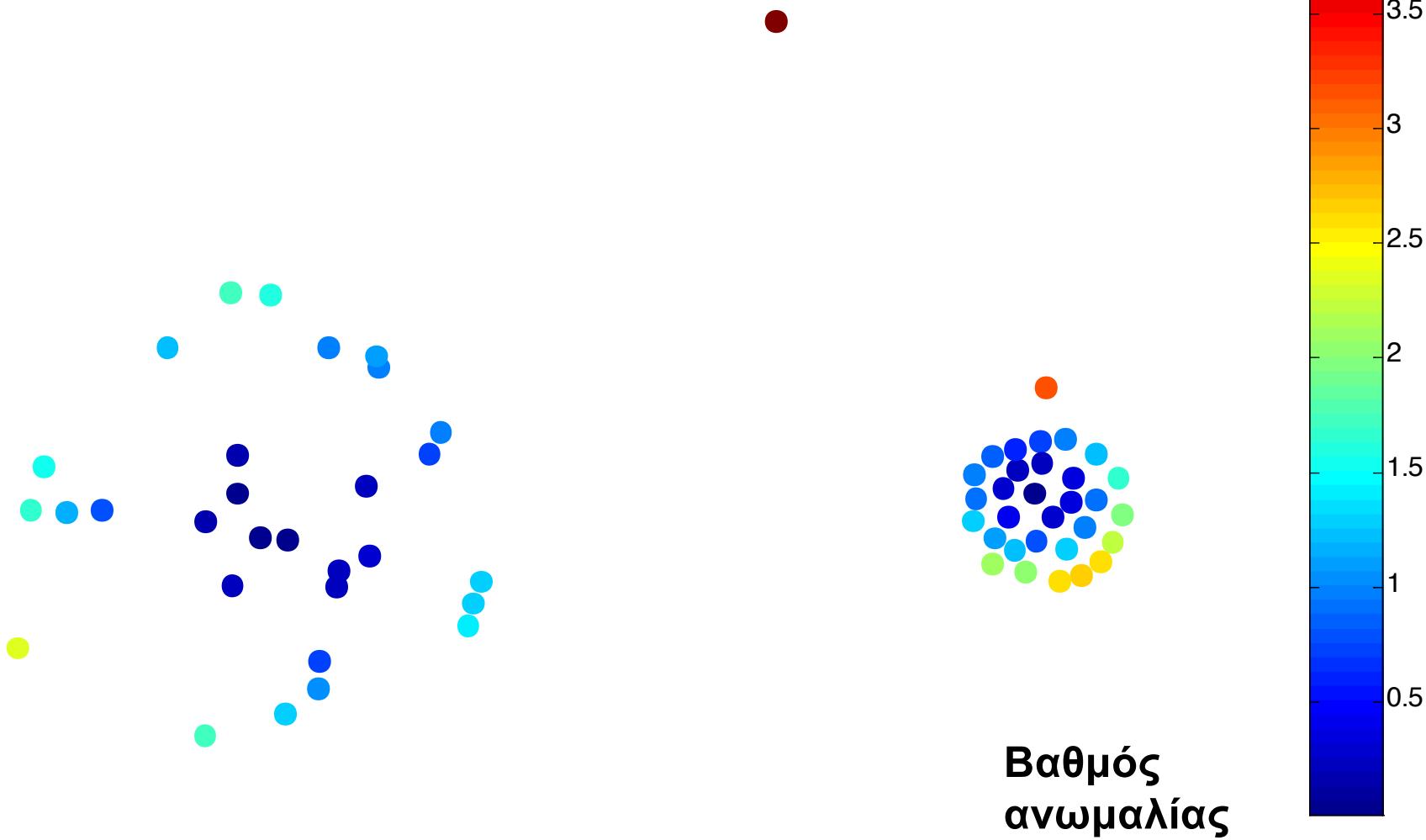
- Χρησιμοποιούνται οι συστάδες για την αναπαράσταση των κανονικών κατηγοριών
- Δύο κατηγορίες μεθόδων
  - Μέθοδοι που θεωρούν μικρές συστάδες ως ανωμαλίες
  - Μέθοδοι που ορίζουν ένα σημείο ως ανώμαλο όταν δεν προσαρμόζεται στη συσταδοποίηση



# Βάσει Απόστασης από το Κοντινότερο Κέντρο



# Βάσει Σχετικής Απόστασης από το Κοντινότερο Κέντρο



## Δυνατά Σημεία/Αδυναμίες

### Προσεγγίσεων βάσει Συσταδοποίησης

- Απλή προσέγγιση και εύκολη στη χρήση
- Πολλές τεχνικές συσταδοποίησης μπορούν να χρησιμοποιηθούν
- Δεν είναι εύκολο
  - να αποφασίσουμε ποια τεχνική συσταδοποίησης θα χρησιμοποιηθεί
  - να αποφασιστεί ο αριθμός των συστάδων
- Η ύπαρξη ακραίων τιμών μπορεί να επηρεάσει τις συστάδες που θα προκύψουν

# Περιεχόμενα Διάλεξης



- Χαρακτηριστικά προβλημάτων ανίχνευσης ανωμαλιών
- Χαρακτηριστικά μεθόδων ανίχνευσης ανωμαλιών
- Στατιστικές προσεγγίσεις
- Προσεγγίσεις βάσει εγγύτητας
- Προσεγγίσεις βάσει πυκνότητας
- Προσεγγίσεις βάσει συσταδοποίησης
- Αξιολόγηση ανίχνευσης ανωμαλιών

# Τρόποι Αξιολόγησης Ανίχνευσης Ανωμαλιών

- Διακρίνονται οι εξής τρόποι αξιολόγησης
  - Όταν υπάρχουν ετικέτες κατηγοριών
    - Χρησιμοποιούνται μέτρα αξιολόγησης κατηγοριοποίησης
    - Λόγω μικρού μεγέθους ανώμαλης κατηγορίας χρησιμοποιούνται: **ακρίβεια, ανάκληση, ρυθμός ψευδών προειδοποίησεων (false positive rate)**
  - Όταν δεν υπάρχουν ετικέτες κατηγοριών
    - Πιο δύσκολο
    - Βάσει βελτίωσης της προσαρμογής μοντέλου, όταν εξαλειφθούν οι ανωμαλίες
    - Με χρήση «εσωτερικών» μέτρων, όπως στη συσταδοποίηση

# Κατανομή Βαθμών Ανωμαλιών

- Ένας γενικότερος τρόπος εκτίμησης των αποτελεσμάτων ανίχνευσης ανωμαλιών
  - Ιδανικά, μικρό ποσοστό των δεδομένων είναι ανώμαλα
  - Άρα η πλειοψηφία των βαθμών ανωμαλίας πρέπει να είναι σχετικά χαμηλοί, ενώ ένα μικρό ποσοστό τους θα είναι υψηλοί
  - Εξετάζουμε την **κατανομή των βαθμών ανωμαλίας** με ένα **ιστόγραμμα** ή με ένα διάγραμμα πυκνότητας

# Παράδειγμα

- Δύο συστάδες 100 σημείων η καθεμία, διαφορετικής πυκνότητας
- Μέση απόσταση από k-NN
- Μέτρο LOF
- Η κατανομή των βαθμών πρέπει να μοιάζει με εκείνη των βαθμών LOF
- Μπορεί να υπάρχουν ακραίες τιμές, καθώς κινούμαστε προς τα δεξιά, αλλά αυτές πρέπει να περιέχουν μικρό ποσοστό των σημείων

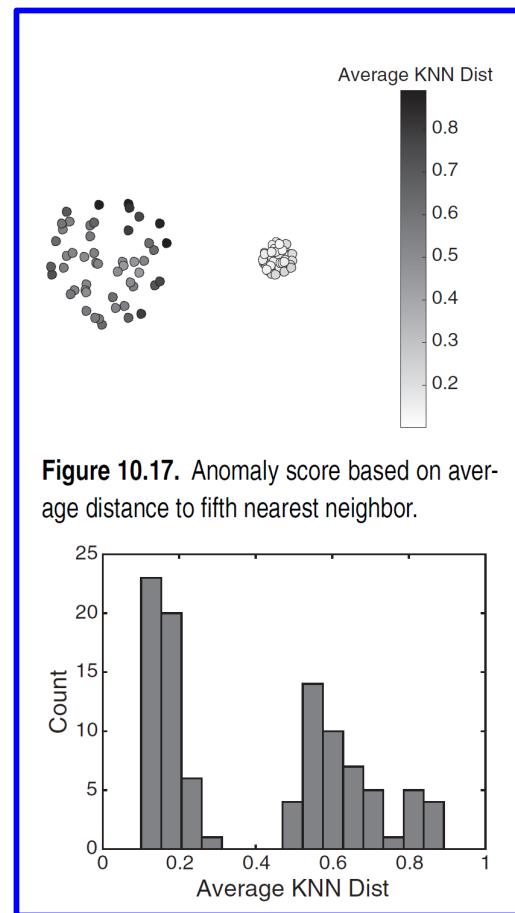


Figure 10.17. Anomaly score based on average distance to fifth nearest neighbor.

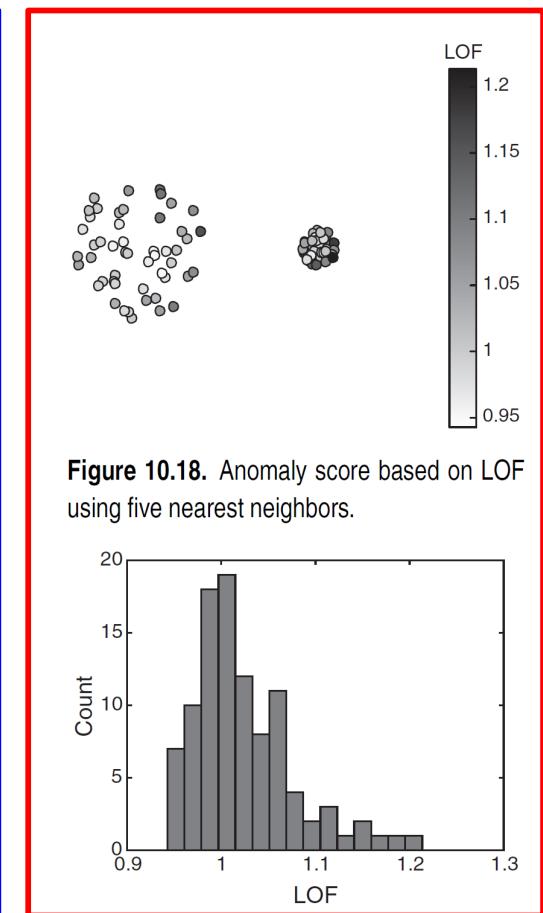
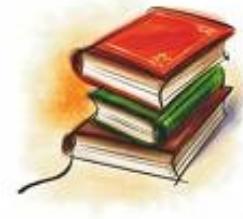
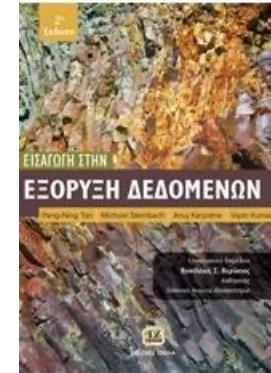


Figure 10.18. Anomaly score based on LOF using five nearest neighbors.

# Βιβλιογραφία



- P. Tan, M. Steinbach, A. Karpatne, V. Kumar. “Εισαγωγή στην Εξόρυξη Δεδομένων”, 2<sup>η</sup> Έκδοση, Εκδόσεις Τζιόλα.
  - *Κεφ. 9: Ανίχνευση Ανωμαλιών*





# 11. Επαναληπτικές Ασκήσεις



---

Ανάλυση Δεδομένων  
(*Data Analytics*)

Χρήστος Δουλκερίδης  
2024-25

# Map

1. Box plot
2. Sampling
3. Histograms
4. Normalization
5. Log plots
6. K-means
7. Association Analysis
8. K-NN
9. Naïve Bayes
10. Decision Tree

# Exercise 1A: Box Plot

- Να περιγράψετε πώς ένα θηκόγραμμα παρέχει πληροφορία για το εάν οι τιμές ενός γνωρίσματος έχουν συμμετρική κατανομή.
- Τι μπορείτε να πείτε για τη συμμετρία των κατανομών των γνωρισμάτων της διπλανής εικόνας;

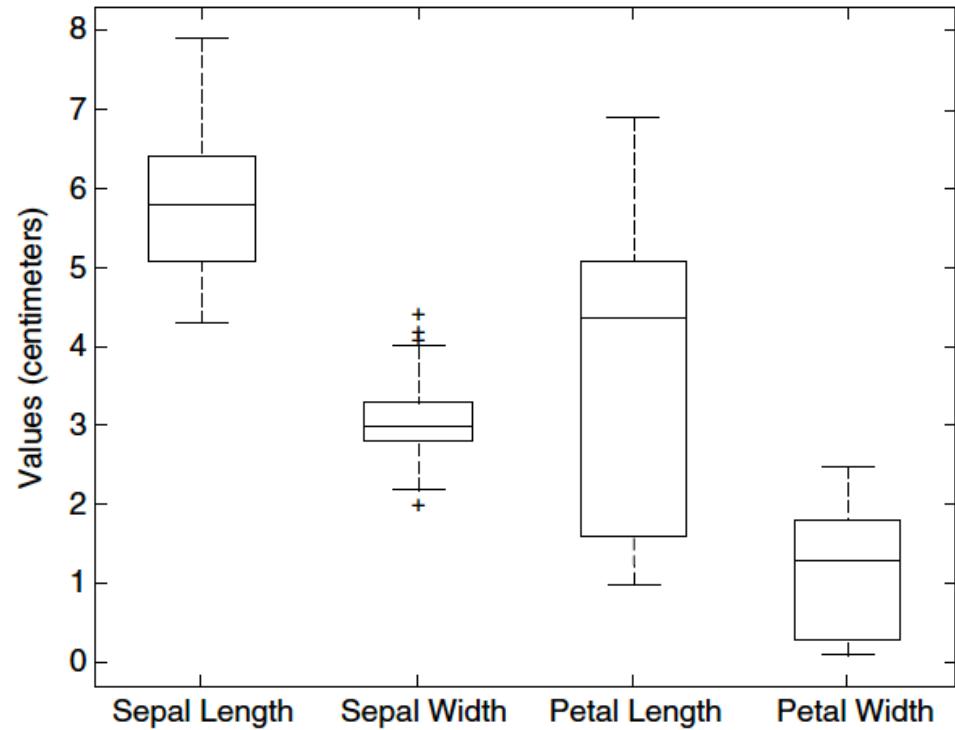
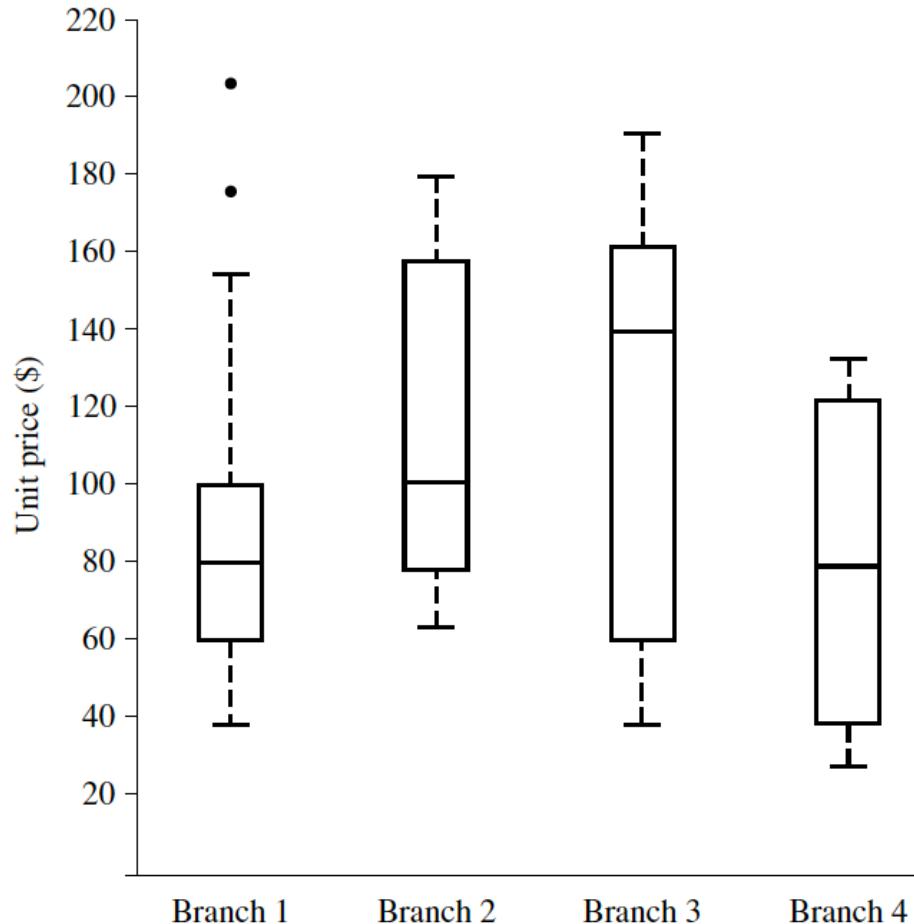


Figure 3.11. Box plot for Iris attributes.

# Exercise 1B: Box Plot



Η εικόνα δείχνει θηκογράμματα για την τιμή πώλησης (unit price) προϊόντων σε 4 υποκαταστήματα (branches) μιας εταιρείας για κάποια χρονική περίοδο.

Τι ποσοτική πληροφορία μπορείτε να πάρετε για τις τιμές του branch 1;

Boxplot for the unit price data for items sold at four branches of *AllElectronics* during a given time period.

# Exercise 2: Sampling

- Θέλουμε να οπτικοποιήσουμε ένα σύνολο δεδομένων που αποτελείται από πάρα πολλές εγγραφές. Μια συνηθισμένη πρακτική είναι να πάρουμε ένα αντιπροσωπευτικό δείγμα του συνόλου δεδομένων. Να συζητηθούν τα πλεονεκτήματα και τα μειονεκτήματα της χρήσης δειγματοληψίας για τη μείωση των αντικειμένων που θα οπτικοποιηθούν
- Είναι η απλή τυχαία δειγματοληψία (χωρίς αντικατάσταση) κατάλληλη προσέγγιση;
- Γιατί ή γιατί όχι;

# Exercise 3: Histograms

- Η ακόλουθη λίστα αντιστοιχεί σε τιμές (52) προϊόντων που έχουν πωληθεί (σε \$)
- Οι τιμές έχουν ταξινομηθεί (για ευκολία):
  - 1, 1, 5, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.
- Σχεδιάστε ένα ιστόγραμμα όπου κάθε κάδος αντιστοιχεί σε \$1
- Σχεδιάστε ένα ιστόγραμμα με κάδους των \$10

# Exercise 4A: Normalization

- Έστω ένα σύνολο δεδομένων που περιέχει γνώρισμα «ετήσια έσοδα»
- Το γνώρισμα έχει:
  - ελάχιστη τιμή \$12,000 και
  - μέγιστη τιμή \$98,000
- Ζητείται να κανονικοποιήσετε (με **min-max κανονικοποίηση**) τις τιμές στο πεδίο τιμών [0.0, 1.0]
- Ποια είναι η κανονικοποιημένη τιμή του \$73,600;

# Exercise 4B: Normalization

- Κανονικοποίηση z-score
- Υποθέστε ότι ο μέσος είναι  $\mu=\$54,000$  και η τυπική απόκλιση  $\sigma=\$16,000$
- Ποια η κανονικοποιημένη τιμή του \$73,600 κατά z-score;

# Exercise 5: Λογαριθμικά Διαγράμματα

- Για τα ακόλουθα δεδομένα, δείξτε ότι ακολουθούν νόμο δυνάμεων με χρήση διαγράμματος λογαριθμικής κλίμακας και βρείτε ακριβώς την έκφραση του νόμου

x	5	15	30	50	95
y	10	90	360	1000	3610

# Exercise 6: K-means

- Δίνονται τα ακόλουθα σημεία:  
**2, 4, 10, 12, 3, 20, 30, 11, 25**
- Υποθέτοντας  $k = 3$  και τυχαία επιλογή των αρχικών μέσων:  $\mu_1 = 2$ ,  $\mu_2 = 4$  και  $\mu_3 = 6$
- Να δείξετε τις συστάδες που προκύπτουν από την εφαρμογή του K-means μετά από 1 επανάληψη, καθώς και τους νέους μέσους
- Σε περίπτωση ισοπαλιών, ένα σημείο ανατίθεται στη συστάδα που έχει τον μικρότερο δείκτη

# Exercise 7: Association Analysis

- (1) Να υπολογιστεί η **υποστήριξη** (**support**) για τα  $\{e\}$ ,  $\{b,d\}$  και  $\{b,d,e\}$  θεωρώντας κάθε συναλλαγή ως καλάθι
- (2) Βάσει των παραπάνω, να υπολογιστεί η **εμπιστοσύνη** (**confidence**) για τους κανόνες
  - $\{b,d\} \rightarrow \{e\}$
  - $\{e\} \rightarrow \{b,d\}$
- (3) Να επαναληφθεί το πρώτο ερώτημα θεωρώντας κάθε πελάτη ως καλάθι
- (4) Βάσει των παραπάνω, να επαναληφθεί το δεύτερο ερώτημα

Πελάτης	Συναλλαγή	Προϊόντα
1	01	{a,d,e}
1	24	{a,b,c,e}
2	12	{a,b,d,e}
2	31	{a,c,d,e}
3	15	{b,c,e}
3	22	{b,d,e}
4	29	{c,d}
4	40	{a,b,c}
5	33	{a,d,e}
5	38	{a,b,e}

# Exercise 8: K-NN

- Δίνεται το διπλανό σύνολο δεδομένων
- Να βρεθεί η κλάση για το σημείο  $G(2.0, 3.5)$  αν χρησιμοποιηθεί 1-NN με ευκλείδεια απόσταση
- Να απαντηθεί το ίδιο ερώτημα για 3-NN με ευκλείδεια απόσταση και με ψηφοφορία πλειοψηφίας (majority voting)

ID	$x_1$	$x_2$	Class
A	1.0	1.0	1
B	2.0	0.5	1
C	1.0	2.5	1
D	3.0	3.5	2
E	5.5	3.5	2
F	5.5	2.5	2

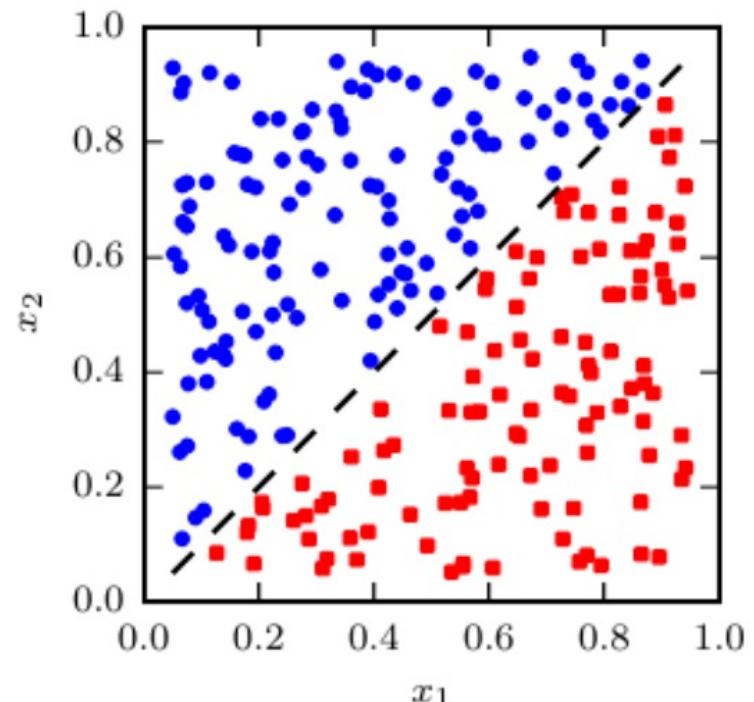
# Exercise 9: Naïve Bayes

- Θεωρήστε το παρακάτω σύνολο δεδομένων.
- Να ταξινομήσετε μια νέα εγγραφή (`Age=23, Car=truck`) με χρήση του **Naïve Bayes**.
- Να υποθέσετε ότι το πεδίο τιμών του γνωρίσματος `Car` είναι `{sports, vintage, suv, truck}`

$\mathbf{x}_i$	Age	Car	Class
$\mathbf{x}_1$	25	sports	$L$
$\mathbf{x}_2$	20	vintage	$H$
$\mathbf{x}_3$	25	sports	$L$
$\mathbf{x}_4$	45	suv	$H$
$\mathbf{x}_5$	20	sports	$H$
$\mathbf{x}_6$	25	suv	$H$

# Exercise 10A: Decision Tree

- Το διάγραμμα δείχνει ένα σύνολο δεδομένων με δύο κλάσεις που είναι γραμμικά διαχωρίσιμο
- Υπάρχει δέντρο απόφασης βάθους 1 που να πετυχαίνει 100% ακρίβεια κατηγοριοποίησης;



# Exercise 10B: Decision Tree

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

- Να παραχθεί το πρώτο επίπεδο ενός δέντρου απόφασης χρησιμοποιώντας το κέρδος πληροφορίας (information gain)

# Πηγές Αναφοράς

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Anuj Karpatne, “*Εισαγωγή στην Εξόρυξη Δεδομένων*”, Εκδόσεις Τζιόλα, 2018.

