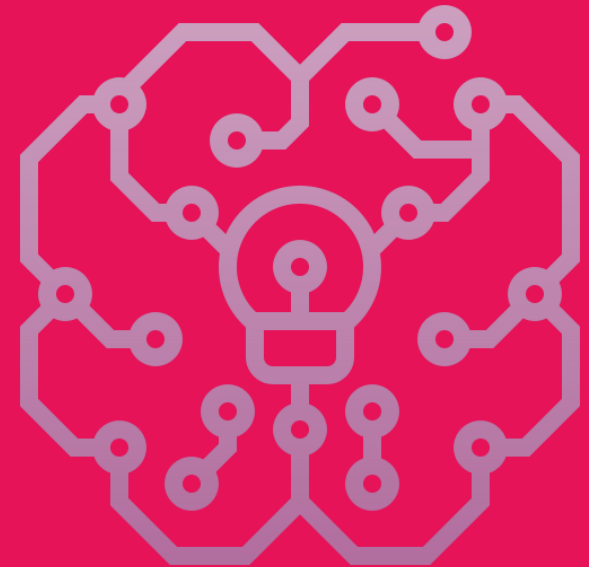


# Υπολογιστική Νοημοσύνη και Μηχανική Μάθηση



# Εισαγωγή



# Big Data

- Διάδοση χρήσης προσωπικών υπολογιστών
- Διάδοση χρήσης κινητών συσκευών
- Όλοι παράγουν δεδομένα
- Οι άνθρωποι πλέον είναι παραγωγοί και καταναλωτές δεδομένων
- Τα δεδομένα έχουν κάποια δομή (π.χ., συμπλήρωση φόρμας στο Διαδίκτυο) ή όχι
- Τα δεδομένα δεν είναι τυχαία αλλά περιγράφουν κάποιου είδους ‘συμπεριφορά’
- Επικρατεί ο όρος big data λόγω του μεγάλου όγκου των δεδομένων
- Χρειαζόμαστε τεχνολογίες για να εξάγουμε γνώση από τα δεδομένα και να προβλέψουμε καταστάσεις
  - Να καταλάβουμε τα πρότυπα και τις διαδικασίες
  - Να προβλέψουμε το μέλλον

# Μηχανική Μάθηση

- Η Μηχανική Μάθηση (MM) είναι μια ερευνητική περιοχή που περιλαμβάνει ένα σύνολο θεωριών και τεχνολογιών για να προγραμματίσουμε τους υπολογιστές να βελτιστοποιήσουν ένα κριτήριο επίδοσης χρησιμοποιώντας ιστορικά δεδομένα-παραδείγματα που περιγράφουν κάποιο φαινόμενο ή μια διαδικασία ή μια συμπεριφορά
- Παραδείγματα:
  - Να μάθει ο υπολογιστής να προβλέπει μελλοντικές θερμοκρασίες
  - Να μάθει ο υπολογιστής να υπολογίζει τις μελλοντικές ανάγκες σε μισθοδοσία για μια επιχείρηση
- Πότε χρησιμοποιούμε τη Μηχανική Μάθηση;
  - Η ανθρώπινη εμπειρία δεν υπάρχει (π.χ., πλοήγηση οχήματος στον Άρη)
  - Οι άνθρωποι δεν είναι ικανοί να εξηγήσουν την εμπειρία και τεχνογνωσία τους (π.χ., αναγνώριση ομιλίας)
  - Η λύση μπορεί να αλλάζει κατά τη διάρκεια του χρόνου (π.χ. βέλτιστη δρομολόγηση πακέτων στο δίκτυο)
  - Η λύση πρέπει να προσαρμοστεί σε συγκεκριμένα σενάρια (π.χ., επεξεργασία βιομετρικών χαρακτηριστικών)



# Μηχανική Μάθηση

- Η μάθηση εστιάζει στην αναγνώριση γενικών μοντέλων πάνω από ιστορικά δεδομένα-παραδείγματα
- Τα δεδομένα λόγω της παρούσας κατάστασης είναι αρκετά και 'φθηνά', η γνώση είναι 'ακριβή' και δύσκολη να εξαχθεί ενώ εξάγεται εστιασμένη σε συγκεκριμένα σενάρια
- Παράδειγμα
  - Καταγραφή συναλλαγών και εξαγωγή της συμπεριφοράς των καταναλωτών
  - Καταγραφή ενός συνόλου παραμέτρων, π.χ., θερμοκρασίας, υγρασίας, κ.λπ. και αναγνώριση φαινομένων π.χ., φωτιάς, για λήψη αποφάσεων

## Στόχος

**Δημιουργία ενός καλού μοντέλου που αποτελεί μια χρήσιμη προσέγγιση των δεδομένων που περιγράφουν πραγματικές καταστάσεις**



# Εξόρυξη Δεδομένων

- Η εξόρυξη δεδομένων/γνώσης στοχεύει στην παραγωγή νέων δεδομένων και της προαναφερθείσας γνώσης
- Παραδείγματα
  - Ανάλυση καλαθιού αντικειμένων (Market basket analysis), ανάλυση και διαχείριση σχέσεων πελατών (Customer relationship management - CRM)
  - Αποτίμηση οικονομικού προφίλ (Credit scoring), αναγνώριση οικονομικής απάτης (fraud detection)
  - Ρομποτική, αυτόματος έλεγχος, αυτόματη αναγνώριση και διόρθωση λαθών
  - Ιατρική διάγνωση
  - Αναγνώριση ψηφιακής εισβολής, αναγνώριση και διαχείριση spam filters
  - Λειτουργία μηχανών αναζήτησης στη εξόρυξη γνώσης στο Διαδίκτυο
  - ...

# Στόχοι της MM

- Όπως περιγράφηκε, η MM στοχεύει στην κατασκευή ενός μοντέλου που βελτιστοποιεί ένα κριτήριο επίδοσης μέσω της επεξεργασίας ιστορικών δεδομένων-παραδειγμάτων
- Στατιστική → Μαθηματικά μοντέλα που βοηθούν στο συμπερασμό μέσα από πλειάδες/δείγματα-παραδείγματα (samples)
- Επιστήμη Υπολογιστών → Δημιουργία αποδοτικών αλγορίθμων για να επιλυθεί το πρόβλημα βελτιστοποίησης που προκύπτει για να καταλήξουμε στο τελικό μοντέλο MM
- Η δημιουργία του μοντέλου ακολουθείται από την αποτίμησή του ώστε να διαπιστωθεί αν η επίδοσή του είναι αποδεκτή ώστε να χρησιμοποιηθεί σε πραγματικές συνθήκες

# Κατηγορίες

- Επιβλεπόμενα μοντέλα (μάθηση υπό επίβλεψη – Supervised Learning)
- Μη επιβλεπόμενα μοντέλα (μάθηση χωρίς επίβλεψη – Unsupervised Learning)
- Υβριδικά μοντέλα
- Παραδείγματα
  - Εξαγωγή κανόνων συσχετίσεων (Association rules learning)
  - Κατηγοριοποίηση (Classification)
  - Παλινδρόμηση (Regression)
  - Ενισχυτική μάθηση (Reinforcement learning)



# Επίβλεψη

- Για να παραχθεί το μοντέλο, τα δεδομένα εκπαίδευσης περιλαμβάνουν την κλάση/αποτέλεσμα στην οποία ανήκει η κάθε πλειάδα – καθοδήγηση από τα δεδομένα για το σωστό αποτέλεσμα
- Παράδειγμα: Δένδρα αποφάσεων
- Τελικά: χρήση κανόνων για να προβλέψουμε τις μελλοντικές κλάσεις στα άγνωστα δεδομένα
- Πλεονέκτημα: Οι κανόνες που εξάγονται μπορούν να γίνουν εύκολα αντιληπτοί από τους ανθρώπους – υψηλή επεξηγησιμότητα (interpretability & explainability)

# Μη Επίβλεψη

- Για να παραχθεί το μοντέλο, τα δεδομένα εκπαίδευσης **δεν** περιλαμβάνουν την κλάση/αποτέλεσμα στην οποία ανήκει η κάθε πλειάδα
- Το μοντέλο μαθαίνει ποια δεδομένα είναι 'φυσιολογικά' όπως αυτά αποτυπώνονται στο σύνολο δεδομένων εκπαίδευσης
- Παραδείγματα:
  - Συσταδοποίηση - Clustering: Ομαδοποίηση των δεδομένων σε ομάδες με βάση κάποια συνάρτηση ομοιότητας

# Ενισχυτική Μάθηση

- Εκμάθηση μια πολιτικής/στρατηγικής → αλληλουχία εξόδων/κινήσεων
- Δεν υπάρχει επίβλεψη στη διαδικασία μάθησης αλλά οφέλη (rewards) για κάθε σωστή κίνηση
- Παραδείγματα:
  - Κίνηση ρομπότ
  - Παιχνίδια
  - Πολλαπλοί πράκτορες λογισμικού που παρατηρούν μερικώς το περιβάλλον/χώρο τους και επιλέγουν κινήσεις
  - ...

# • Εκμάθηση Συσχετίσεων •

- Ανάλυση καλαθιού προϊόντων/αντικειμένων

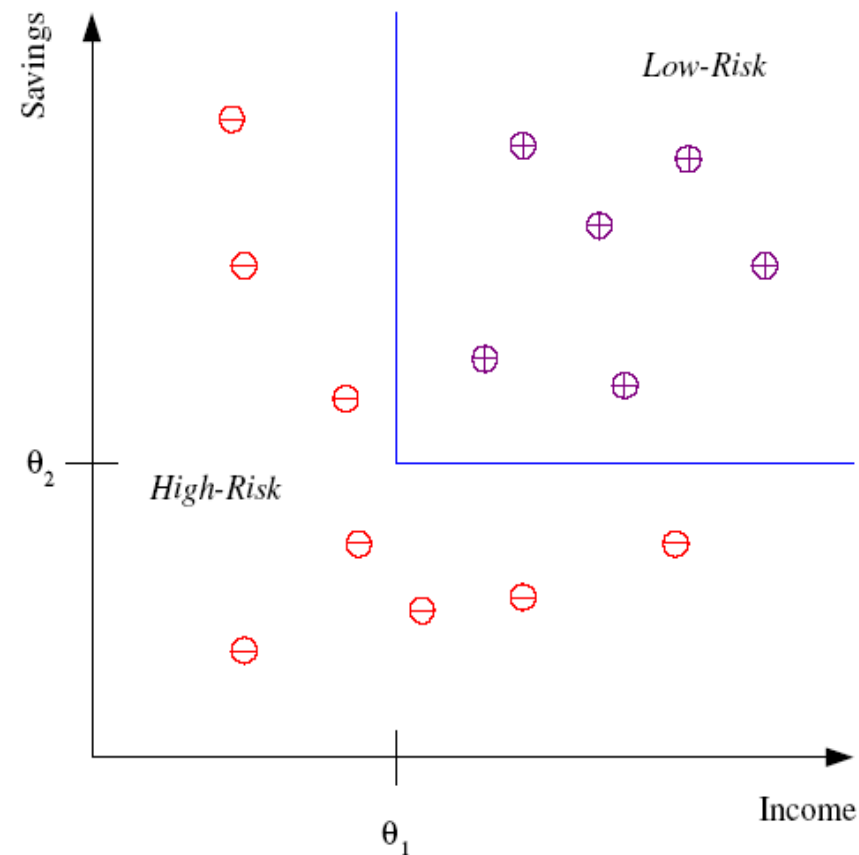
$P(Y | X)$ : πιθανότητα κάποιος που αγόρασε το  $X$  να αγοράσει και το  $Y$  όπου  $X, Y$  είναι δύο προϊόντα/αντικείμενα/υπηρεσίες

Παράδειγμα

$$P(\text{chips} | \text{beer}) = 0.7$$

# Κατηγοριοποίηση

- Παράδειγμα
  - Αποτίμηση οικονομικού/πιστωτικού προφίλ και έγκριση δανείου
  - Διαφοροποίηση ανάμεσα σε χαμηλού και υψηλού κινδύνου πελάτες με βάση το εισόδημά τους και τις αποταμιεύσεις τους
  - Βάση γνώσης με τη μορφή κανόνων:
    - IF income >  $\theta_1$  AND savings >  $\theta_2$   
THEN low-risk ELSE high-risk



# Κατηγοριοποίηση

- Εφαρμογές:
  - Αναγνώριση προτύπων - Pattern recognition
  - Αναγνώριση προσώπου - Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
  - Αναγνώριση χαρακτήρων σε εικόνες - Character recognition: Different handwriting styles.
  - Αναγνώριση ομιλίας - Speech recognition
  - Ιατρική διάγνωση - Medical diagnosis
  - Αναγνώριση βιομετρικών χαρακτηριστικών - Biometrics: Recognition/authentication using physical and/or behavioral characteristics: Face, iris, signature, etc
  - Αναγνώριση ανωμαλιών στα δεδομένα - Outlier/novelty detection

# • Αναγνώριση Προσώπου

- Δεδομένα εκπαίδευσης



- Δεδομένα αποτίμησης



# Παλινδρόμηση

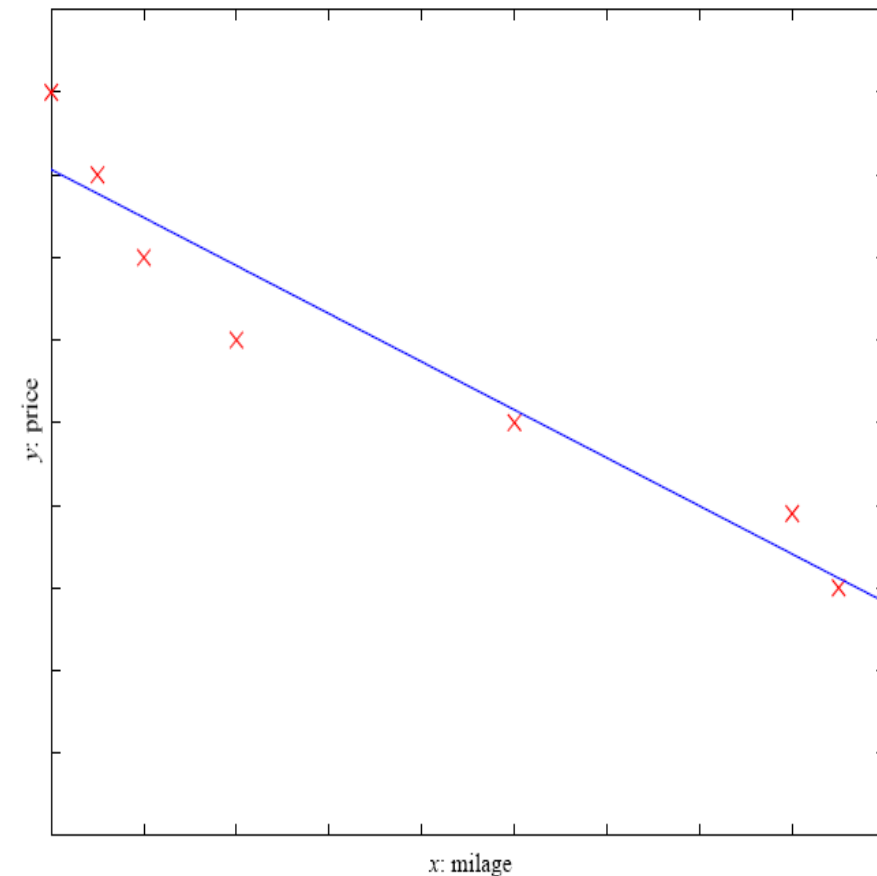
- Παράδειγμα

- Η τιμή ενός μεταχειρισμένου αυτοκινήτου
- $x$  : οι ιδιότητες/χαρακτηριστικά/στοιχεία του αυτοκινήτου
- $y$  : η τιμή του αυτοκινήτου

$$y = g(x | q)$$

$g()$  : το μοντέλο παλινδρόμησης

$q$  : παράμετροι





# Αποθετήρια

- Υπάρχουν ανοιχτά αποθετήρια στα οποία μπορεί κανείς να βρει σύνολα δεδομένων
- UCI Repository: <http://www.ics.uci.edu/~mlearn/MLRepository.html>
- Kaggle: <https://www.kaggle.com/>
- Statlib: <http://lib.stat.cmu.edu/>
- Zenodo
- ...

Δ ε δ ο μ έ ν α



# Τύποι

- **Εγγραφές (Records):** Relational records, Data matrix, e.g., numerical matrix, crosstabs, Document data: text documents: term-frequency vector, Transaction data
- **Γράφοι και Δίκτυα (Graph and network):** World Wide Web, Social or information networks, Molecular Structures
- **Ακολουθιακά (Ordered):** Video data: sequence of images, Temporal data: time-series, Sequential Data: transaction sequences, Genetic sequence data
- **Χωρικά, εικόνες και πολυμεσικά δεδομένα (Spatial, image and multimedia):** Spatial data: maps, Image data, Video data:

|            | team | coach | play | ball | score | game | win | 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      |

| <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        |

# Χαρακτηριστικά

- Διαστάσεις (Dimensionality): πόσες διαστάσεις/ιδιότητες/χαρακτηριστικά έχουν τα δεδομένα
  - Κατάρα των διαστάσεων (Curse of dimensionality): όσο περισσότερες διαστάσεις έχουν τόσο πιο πολύπλοκη γίνεται η διαδικασία της εκπαίδευσης
- Έλλειψη (Sparsity)
  - Μόνο τα δεδομένα που υπάρχουν είναι αυτά που μετρούν
- Ανάλυση (Resolution)
  - Τα πρότυπα εξαρτώνται από το scale
- Κατανομή (Distribution)
  - Κεντρική τάση, διασπορά, στατιστικές τεχνικές

# • Αντικείμενα Δεδομένων •

- Τα σύνολα δεδομένων αποτελούνται από ένα αριθμό αντικειμένων δεδομένων
- Ένα αντικείμενο δεδομένων αναπαριστά μια οντότητα της πραγματικότητας
- Παράδειγμα: Φανταστείτε μια γραμμή σε ένα πίνακα μιας βάσης δεδομένων
  - Βάση δεδομένων πωλήσεων: πελάτες, προϊόντα, πωλήσεις
  - Βάση ιατρικών δεδομένων: ασθενείς, θεραπείες, φάρμακα
  - Πανεπιστημιακή βάση δεδομένων: φοιτητές, καθηγητές, τμήματα, μαθήματα
- Ορολογία: Τα αντικείμενα δεδομένων θα τα συναντήσουμε με μια από τις ακόλουθες αναφορές: δείγματα (samples), παραδείγματα (examples), στιγμιότυπα (instances), σημεία δεδομένων (data points), αντικείμενα (objects), πλειάδες (tuples)
- Τα αντικείμενα δεδομένων περιγράφονται από τις διαστάσεις (dimensions), ιδιότητες (attributes), στοιχεία (features), στήλες (columns), μεταβλητές (variables)
- Αναγωγή
  - Γραμμές σε ένα πίνακα βάσης δεδομένων → αντικείμενο δεδομένων, Στήλες → διαστάσεις/ιδιότητες



# • Αντικείμενα Δεδομένων •

- Οι διαστάσεις/ιδιότητες/χαρακτηριστικά αποτυπώνουν ένα πεδίο δεδομένων που αναπαριστά ένα συγκεκριμένο χαρακτηριστικό ενός αντικειμένου δεδομένων
- Παραδείγματα: customer\_ID, name, address
- Τύποι:
  - Nominal
  - Ordinal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

# Τύποι

- Nominal: κατηγορίες (categories), καταστάσεις (states), ή ονόματα (“names of things”)
  - Hair\_color = {auburn, black, blond, brown, grey, red, white}, marital status, occupation, ID numbers, zip codes
- Binary (δυναδικά δεδομένα)
  - Ιδιότητες nominal με μόνο δύο τιμές/καταστάσεις (0, 1 ή οποιοδήποτε άλλο ζεύγος τιμών)
  - Συμμετρικά δεδομένα: και τα δύο στοιχεία του ζεύγους έχουν την ίδια βαρύτητα/σημασία, π.χ. φύλλο γέννησης
  - Μη συμμετρικά δεδομένα: οι τιμές δεν έχουν την ίδια βαρύτητα/σημασία, π.χ. ένα ιατρικό τεστ (θετικό, αρνητικό – συνήθως δίνουμε μεγαλύτερη σημασία στο θετικό αποτέλεσμα)
- Ordinal
  - Οι τιμές έχουν μια σειρά με νόημα αλλά η διαφοροποίηση τους δεν είναι γνωστή, π.χ. Size = {small, medium, large}, τάξεις σχολείου

# • Αριθμητικά Δεδομένα •

- Ποσότητα: Ακέραιοι ή Πραγματικοί αριθμοί
- Interval
  - Μετρώνται σε μια κλίμακα ίσου μεγέθους μονάδες
  - Οι τιμές έχουν μια σειρά
  - Παραδείγματα: θερμοκρασία σε  $C^{\circ}$  ή  $F^{\circ}$ , ημερομηνίες
  - Δεν υπάρχει κάποιο σημείο αρχής (true zero-point)
- Ratio
  - Υπάρχει ένα σημείο αρχής και όλες οι τιμές καθορίζονται σε σχέση με αυτό (zero-point)
  - Οι τιμές καθορίζονται σε σχέση με το πόσο μεγαλύτερες είναι συγκριτικά με το σημείο αρχής (10  $K^{\circ}$  είναι διπλάσιοι σε σχέση με τους 5  $K^{\circ}$ ).
  - Παραδείγματα: θερμοκρασία σε βαθμούς Kelvin, μήκος (length), χρηματικές ποσότητες



# Τύποι

- Διακριτές (Discrete Attributes)
  - Έχουν ένα πεπερασμένο σύνολο τιμών ή άπειρο σύνολο διακριτών τιμών
  - Παραδείγματα: ταχυδρομικοί κώδικες, επάγγελμα, σύνολο λέξεων μιας γλώσσας
  - Πολλές φορές αναπαριστώνται με ακεραίους αριθμούς
  - Σημείωση: οι δυαδικές (binary) ιδιότητες είναι μια ειδική περίπτωση αυτής της κατηγορίας
- Συνεχείς (Continuous Attributes)
  - Παίρνουν πραγματικές τιμές
  - Παραδείγματα: Θερμοκρασία, ύψος, βάρος
  - Πρακτικά οι πραγματικές τιμές μπορούν να μετρηθούν και να αναπαρασταθούν με χρήση ενός αριθμού ψηφίων

# • Στατιστική Περιγραφή •

- Η στατιστική περιγραφή είναι σημαντική για να δούμε την κατανομή και διασπορά των δεδομένων
- Τεχνικές-Στατιστικά Χαρακτηριστικά
  - Διάμεσος (median), μέγιστος (max), ελάχιστος (min), quantiles, ανωμαλίες (outliers), διακύμανση (variance), κ.λπ.
- Ανάλυση αριθμητικών διαστάσεων σε ταξινομημένα αριθμητικά τμήματα (sorted intervals)
  - Διασπορά: αναλύεται με πολλαπλές αναλύσεις
  - Ανάλυση με χρήση boxplots ή ανάλυση των quantiles
- Ανάλυση αριθμητικών διαστάσεων με στατιστικές τεχνικές σε υπολογισμένες μετρικές
  - Χρήση τεχνικών folding σε αριθμητικές διαστάσεις
  - Ανάλυση με χρήση boxplots ή ανάλυση των quantiles

# Κεντρική Τάση

- Μέση τιμή (sample vs. population):
  - Σημείωση:  $n$  είναι το μέγεθος του δείγματος και  $N$  το σύνολο του πληθυσμού
  - Weighted arithmetic mean (βάρος στα δείγματα)
  - Trimmed mean: απορρίπτουμε τις υπερβολικές τιμές
- Διάμεσος
  - Η μεσαία τιμή σε περιττό αριθμό δειγμάτων ή ο μέσος όρος των δύο μεσαίων τιμών διαφορετικά
  - Εκτίμηση:
$$median = L_1 + \left( \frac{N/2 - (\sum freq)_1}{freq_{median}} \right) width.$$
- Επικρατούσα τιμή (Mode)
- Η πιο συχνή τιμή στα δεδομένα
- Εμπειρικός μαθηματικός τύπος:  $mean - mode = 3 \times (mean - median)$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \mu = \frac{\sum x}{N}$$

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

| <i>age</i> | <i>frequency</i> |
|------------|------------------|
| 1–5        | 200              |
| 6–15       | 450              |
| 16–20      | 300              |
| 21–50      | 1500             |
| 51–80      | 700              |
| 81–110     | 44               |

# Κεντρική Τάση

- Ο υπολογισμός της διαμέσου είναι ακριβή διαδικασία ιδιαίτερα να πραγματοποιείται συνεχώς πάνω σε μεγάλους όγκους δεδομένων
- Για αριθμητικές τιμές, μπορούμε να προσεγγίσουμε τη διάμεσο με τη χρήση του ακόλουθου μαθηματικού τύπου (χωρίζουμε τα δεδομένα σε διαστήματα):

$$median = L_1 + \left( \frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width.$$

- $L_1$ : το κατώτερο όριο του ενδιάμεσου διαστήματος
- $N$ : το πλήθος των τιμών
- $(\sum freq)_l$ : το άθροισμα των συχνοτήτων όλων των διαστημάτων που είναι κάτω από το ενδιάμεσο
- $freq_{median}$ : είναι η συχνότητα των τιμών στο ενδιάμεσο διάστημα
- $width$ : είναι το πλάτος του ενδιάμεσου διαστήματος

# Κεντρική Τάση

- Παράδειγμα

- Τιμές: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110
- Διαστήματα: [10,50], [51,90], [91,130]
- $L_1 = 51$
- $N = 12$
- $(\sum freq)_l = 4$
- $freq_{median} = 7$
- $width = 40$

$$median = L_1 + \left( \frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width.$$

- Εκτιμώμενη διάμεσος =  $51 + ((12/2 - 4) / 7) * 40 = 51 + 2/7 * 40 = 51 + 11.43 = 62.43$
- Θεωρητική διάμεσος =  $(52+56) / 2 = 54$

# Κεντρική Τάση

- Παράδειγμα

- Τιμές: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 71, 72, 85, 86, 88, 90, 110, 140
- Διαστήματα: [11,40], [41,70], [71,100], [101,130], [131,160]
- $L_1 = 71$
- $N = 19$
- $(\sum freq)_l = 11$
- $freq_{median} = 6$
- $width = 30$

$$median = L_1 + \left( \frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$$

- Εκτιμώμενη διάμεσος =  $71 + ((19/2 - 11) / 6) * 30 = 71 - 1.5/6 * 30 = 71 - 7.5 = 63.5$
- Θεωρητική διάμεσος = 70