

# ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

---

Διαδικαστικά

# Συστάσεις I

- Ποιός είμαι εγώ:
  - Email: tsap@cs.uoi.gr
  - Γραφείο: B.3
  - Προτιμώμενες ώρες γραφείου: 10:00-14:00
- Ενδιαφέροντα
  - Social networks, User Generated Content
  - Algorithmic Fairness
  - Συνδυασμός θεωρίας και πράξης

# Συστάσεις II

- Ποιοί είσαστε εσείς:
  - Στο τέλος του μαθήματος θα συμπληρώσετε ένα quiz με κάποιες πληροφορίες.

# Γενικές πληροφορίες για το μάθημα

- Διαλέξεις: Δευτέρα 12:00 – 3:00 μ.μ.
  - Οι διαφάνειες θα είναι στα **αγγλικά**, αλλά η διάλεξη θα γίνεται στα ελληνικά.
  - Θα έχουμε και κάποια επιπλέον φροντιστήρια/αναπληρώσεις κάποιες εβδομάδες.
    - Πότε σας βολεύει?
- Web: <http://www.cs.uoi.gr/~tsap/teaching/cse012/>
  - Ανακοινώσεις, ασκήσεις, υλικό για διάβασμα διαφάνειες από τις διαλέξεις
  - Γραφτείτε και στη σελίδα του μαθήματος στο **ecourse**.
- Βαθμολογία:
  - Η αξιολόγηση του μαθήματος στηρίζεται σε **εργασίες**. Θα υπάρχει προσωπική εξέταση για τις εργασίες. Οι εξετάσεις που γίνανε το 2020-2021 ήταν επίσης με την μορφή εργασίας. Φέτος μπορεί να έχουμε κάποια **εξέταση σε μορφή quiz**
- Πολιτική για καθυστερημένες εργασίες:
  - Μία μέρα καθυστέρηση -10%, δύο μέρες -20%, τρεις μέρες -40%, τέσσερις μέρες -70%, πέντε μέρες -100%.
  - **Free pass policy**: Έχετε 4 free passes τα οποία μπορείτε να χρησιμοποιήσετε όποτε θέλετε για να καθυστερήσετε την παράδοση μιας εργασίας. Το κάθε pass σας δίνει μία μέρα επιπλέον.

# Ασκήσεις

- Οι ασκήσεις θα έχουν (συνήθως) τριών ειδών ερωτήσεις:
  - **Θεωρητικές**: Θα σας ζητηθεί να σχεδιάσετε ένα αλγόριθμο, ή να αποδείξετε κάποια ιδιότητα
  - **Αλγοριθμικές**: Θα σας ζητηθεί να σχεδιάσετε ένα μια λύση για ένα πρόβλημα.
  - **Προγραμματιστικές**: Θα σας ζητηθεί να υλοποιήσετε ένα αλγόριθμο, ή να χρησιμοποιήσετε κάποιο έτοιμο εργαλείο σε κάποια δεδομένα.
- **Αναφορά**: Στις κάποιες ερωτήσεις θα πρέπει να παραδώσετε μία αναφορά. Η αναφορά αυτή μετράει ένα **σημαντικό ποσοστό** του βαθμού της ερώτησης και πρέπει να γίνεται προσεκτικά. Τις περισσότερες φορές σας ζητείται να εξηγήσετε τα αποτελέσματα κάποιου πειράματος.
- **Προγραμματισμός**: Η επεξεργασία μεγάλων ποσοτήτων δεδομένων απαιτεί έξυπνο και αποτελεσματικό προγραμματισμό.
  - Πρέπει να αποφεύγετε δαπανηρές λειτουργίες.
  - Πρέπει να χρησιμοποιείτε τις κατάλληλες δομές.
  - Πρέπει να προσπαθείτε να χρησιμοποιείτε λίγη μνήμη.
  - Κάποιες φορές το πρόγραμμα σας μπορεί να πάρει μερικές ώρες να τελειώσει.

# «Προαπαιτούμενα»

- Δεν υπάρχουν προαπαιτούμενα αλλά καλό θα είναι να έχετε κάποια άνεση με:
  - **Πιθανότητες**: Άνεση με βασικές γνώσεις πιθανοτήτων.
  - **Προγραμματισμός + Python**: Ευκολία στην εκμάθηση νέων εργαλείων. Θα χρησιμοποιήσουμε κάποιες νέες βιβλιοθήκες python.
  - **Γραμμική άλγεβρα**: πίνακες, διανύσματα, ιδιοδιανύσματα.
  - **Αλγόριθμους**: γνώση βασικών αλγορίθμων (π.χ., sorting), και σχεδίασης αλγορίθμων (greedy algorithms, dynamic programming).
  - **Βάσεις δεδομένων**: Χρήση βασικών SQL λειτουργιών
  - **Δομές δεδομένων**: χρήση βασικών δομών δεδομένων.
  - **Γραφήματα**: βασικές έννοιες γραφημάτων

# Στόχοι του μαθήματος

- Να μάθετε **βασικές έννοιες** του data mining, που καλύπτουν και το θεωρητικό υπόβαθρο, και την εφαρμογή στην πράξη.
- Να καταλάβετε το **είδος των προβλημάτων** που αντιμετωπίζουμε στην εξόρυξη δεδομένων.
- Να καταλάβετε τη **θεωρία** και τα **μαθηματικά** πίσω από τους αλγόριθμους και τις τεχνικές
- Να αποκτήσετε ένα σύνολο από **εργαλεία (toolbox)** για εξόρυξη δεδομένων.
- Να παίξετε με **πραγματικά δεδομένα** και να δείτε κάποια ενδιαφέροντα **πραγματικά προβλήματα** (ελπίζω).
- Να το **διασκεδάσετε**.

# Μάθημα

- Η παρακολούθηση και συμμετοχή είναι απαραίτητες
  - Κάνετε ερωτήσεις. Κάποια πράγματα δεν θα είναι ξεκάθαρα και θα πρέπει να τα επαναλάβω.
  - Αν κάτι στηρίζεται σε παλαιότερη γνώση που δεν θυμάστε ζητήστε να κάνουμε μια (σύντομη) επισκόπηση.
  - Αν υπάρχει πρόβλημα με αγγλική ορολογία και τις διαφάνειες μπορούμε να κάνουμε κάποιες ρυθμίσεις.
- Για τα εργαλεία που θα χρησιμοποιήσουμε θα προσπαθήσω να κάνουμε ένα ξεχωριστό φροντιστήριο.

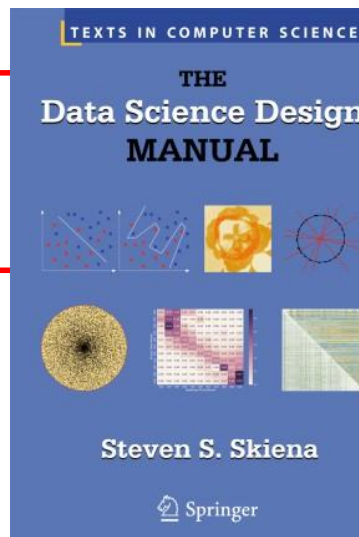
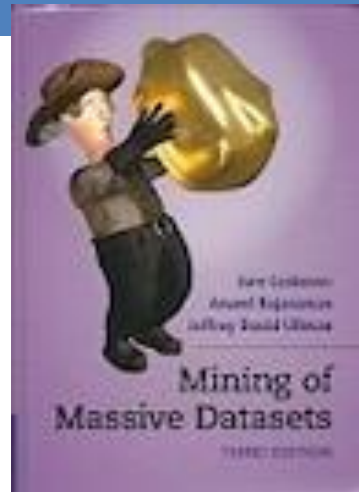
# Θέματα που θα καλύψουμε

- Κάποιο υποσύνολο από τα παρακάτω
  - Frequent itemsets and association rules (συσχετισμοί)
  - Definitions and Computation of Similarity
  - Στατιστική
  - Clustering (συσταδιοποίηση), co-clustering, compression
  - Regression και Classification (κατηγοριοποίηση)
  - Dimensionality Reduction
  - Ranking (ιεραρχηση/ταξινόμηση)
  - Recommendation systems
  - Graph Analysis
  - Covering problems
  - Map-Reduce tools
  - Time-series analysis
  - Aggregation
  - Privacy preserving data mining

# Βιβλιογραφία (Εύδοξος)

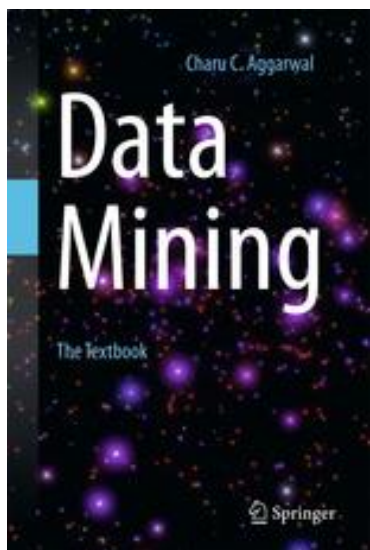
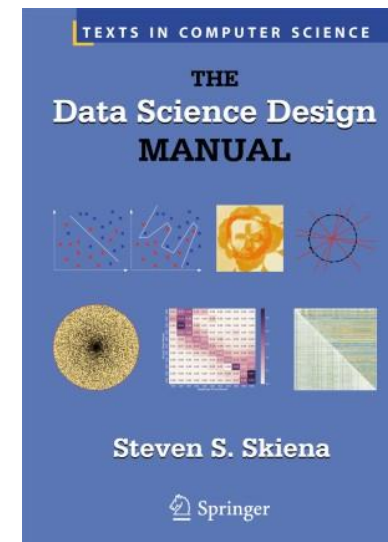
- Βιβλίο [94700707]: Εξόρυξη από Μεγάλα Σύνολα Δεδομένων - 3η Έκδοση, Anand Rajaraman, Jeffrey David Ullman, Jure Leskovec
- Βιβλίο [77107675]: Εισαγωγή στην εξόρυξη δεδομένων, 2η Έκδοση, Tan Pang - Ning, Steinbach Michael, Kumar Vipin, Βερύκιος Βασίλειος
- Βιβλίο [68386089]: ΕΞΟΡΥΞΗ ΚΑΙ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ: ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ, MOHAMMED J. ZAKI, WAGNER MEIRA JR.
- Βιβλίο [122074432]: Επιστήμη των Δεδομένων-Εγχειρίδιο Σχεδιασμού, Skiena S.S. Λεπτομέρειες

Το πρώτο βιβλίο υπάρχει και [online](#) στα Αγγλικά



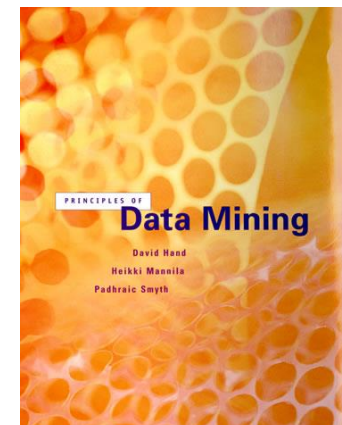
# Επιπλέον Βιβλιογραφία (αγγλικά)

Steven S. Skiena, [The Data Science Design Manual](#) Springer



Charu Aggarwal, [Data Mining, The Textbook](#), Springer, 2015

Hand, Mannila, Smyth. [Principles of Data Mining](#)



# Υλικό

- Εκτός από βιβλία θα χρησιμοποιήσουμε υλικό και από δημοσιευμένα άρθρα
- Για τις διαφάνειες θα δανειστούμε από πολλές πηγές
  - Anand Rajaraman and Jeff Ullman [Mining Massive Datasets](#).
  - P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Addison Wesley, 2006
  - Data Mining, E. Terzi
  - Data Mining, Aris Anagnostopoulos

# Quiz

- Σύντομο ερωτηματολόγιο για να δω τι ξέρετε
  - Χρησιμεύει για να πάρω μια ιδέα του τι κενά μπορεί να χρειαστεί να καλύψουμε.
  - Δεν επηρεάζει βαθμό ή κάτι άλλο.