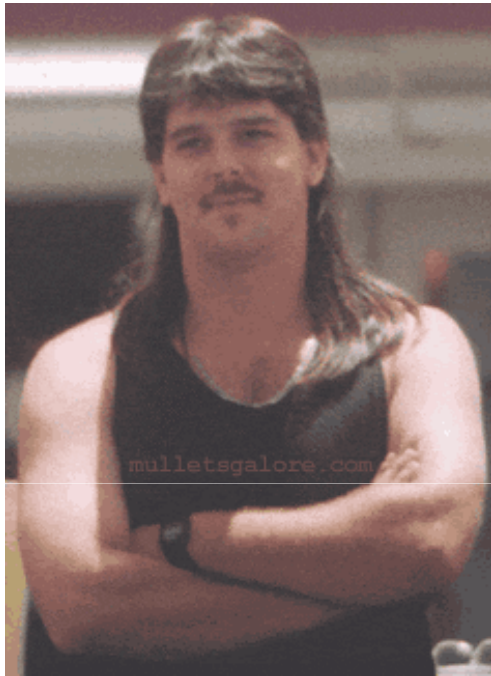


# Συστήματα Συστάσεων – Recommender Systems

Προσαρμογή διαφανειών από:

1. D. Jannach, M. Zanker, A. Felfernig, Gerhard Friedrich. Recommender Systems - An Introduction. Cambridge University Press, 2011.
2. A. Rajaraman, J. D. Ullman. Mining of Massive Datasets. Cambridge University Press, 2011.

# Example: Recommender Systems

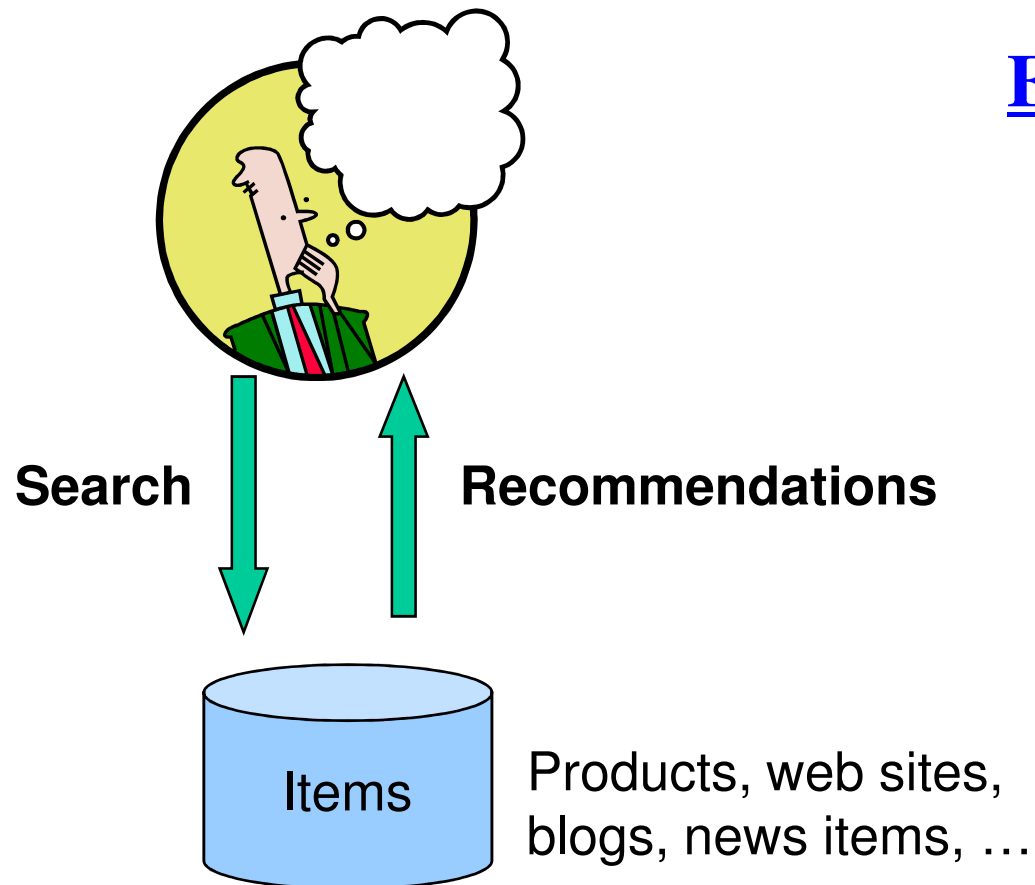


- **Customer X**
  - Buys Metallica CD
  - Buys Megadeth CD



- **Customer Y**
  - Does search on Metallica
  - Recommender system suggests Megadeth from data collected about customer X

# Recommendations



## Examples:

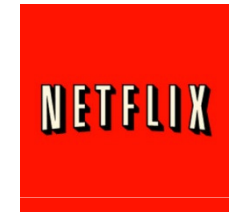
amazon.com.



StumbleUpon



del.icio.us



**m o v i e l e n s**

helping you find the *right* movies

last.fm™  
the social music revolution

Google™  
News

You Tube

XBOX  
LIVE

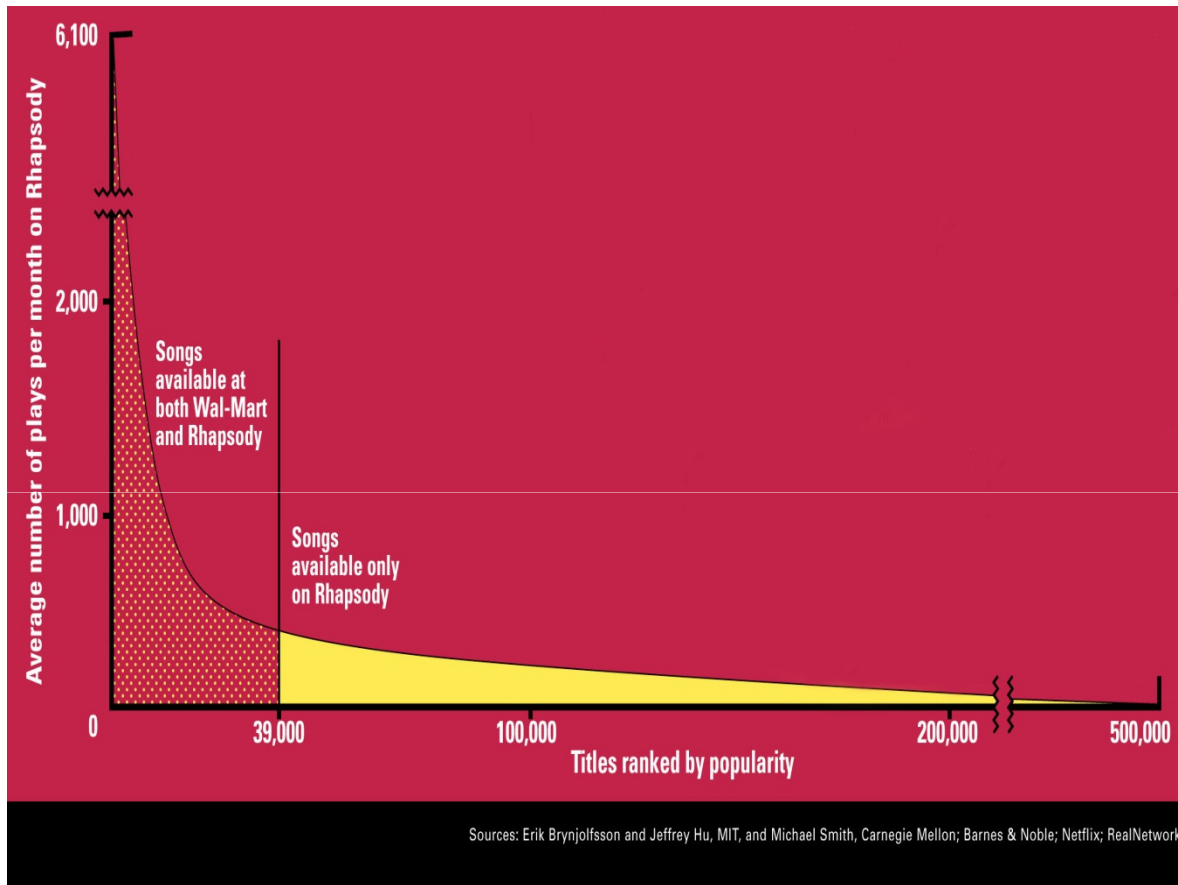
# Γιατί να χρησιμοποιούμε συστήματα συστάσεων;

- Αξία για τον πελάτη
  - Βρίσκει αντικείμενα που είναι ενδιαφέροντα
  - Περιορίζουν το σύνολο των δυνατών επιλογών
  - Βοηθά στην εξερεύνηση του χώρου επιλογών
  - Ανακαλύπτει νέα πράγματα
- Αξία για τον προμηθευτή
  - Επιπρόσθετη και πιθανότατα προσωποποιημένη υπηρεσία για τον πελάτη
  - Αυξάνει την εμπιστοσύνη και την αφοσίωση του πελάτη
  - Αυξάνει τις πωλήσεις, την αναλογία εμφανίσεων/πωλήσεων, κ.α.
  - Ευκαιρίες για προώθηση προϊόντων
  - Αποκτά περισσότερη γνώση για τον πελάτη

# Από την ανεπάρκεια στην αφθονία

- Ο χώρος στα ράφια αποτελεί ένα σπάνιο-ανεπαρκές αγαθό για τους παραδοσιακούς πωλητές
  - Επίσης: τηλεοπτικά δίκτυα, κινηματογράφοι, ...
- Ο παγκόσμιος ιστός επιτρέπει την διασπορά πληροφορίας για προϊόντα με σχεδόν μηδενικό κόστος
  - Από την ανεπάρκεια στην αφθονία
- Περισσότερες επιλογές δημιουργούν την ανάγκη για καλύτερα φίλτρα
  - Μηχανές συστάσεων
  - How Into Thin Air made Touching the Void a bestseller

# Η μεγάλη ουρά (long tail)



- Σύστησε ευρέως άγνωστα αντικείμενα που μπορεί όμως να αρέσουν στους χρήστες!
- 20% των αντικειμένων συγκεντρώνουν το 74% των θετικών βαθμολογιών

# Το πρόβλημα

- Τα συστήματα συστάσεων μας βοηθούν να ταιριάζουμε χρήστες με αντικείμενα
  - Ελαφρύνουμε την υπερφόρτωση πληροφορίας
  - Βοηθός πωλήσεων (καθοδηγεί, συμβουλεύει, πείθει,...)
- Διαφορετικοί σχεδιασμοί συστημάτων
  - Με βάση τη διαθεσιμότητα των δεδομένων που μπορούμε να εκμεταλλευτούμε
  - Έμμεση και άμεση ανάδραση από τον χρήστη
  - Χαρακτηριστικά του πεδίου εφαρμογής

# Στόχος και κριτήρια επιτυχίας

- Διαφορετικές οπτικές γωνίες
  - Εξαρτάται από τον τομέα και τον στόχο
  - Δεν υπάρχει ένα σενάριο που να ταιριάζει σε όλες τις περιπτώσεις
- Σε σχέση με την ανάκτηση
  - Μείωση κόστους αναζήτησης
  - Παροχή «σωστών» συστάσεων
  - Οι χρήστες ξέρουν εκ των προτέρων τι θέλουν
- Σε σχέση με τις συστάσεις
  - Serendipity- αναγνώριση αντικειμένων από το τέλος της ουράς
  - Οι χρήστες δεν ήξεραν για την ύπαρξή τους



# Στόχος και κριτήρια επιτυχίας

- Σε σχέση με την πρόβλεψη
  - Προέβλεψε σε ποιο βαθμό θα αρέσει ένα αντικείμενο στους χρήστες
- Σε σχέση με την αλληλεπίδραση
  - Προσφέρει στους χρήστες «ευχάριστη αίσθηση»
  - Εκπαιδεύει τους χρήστες για τον τομέα του προϊόντος
  - Πείθει τους πελάτες-εξηγεί
- Σε σχέση με την αλλαγή
  - Εμπορικές καταστάσεις
  - Αυξάνει τα «hit», τα “clickthrough” τα «lookers to bookers» ποσοστά
  - Βελτιστοποιεί τα περιθώρια πωλήσεων και κέρδους

# Πώς τα αποτιμούμε?

- Τεστ με πραγματικούς χρήστες
  - A/B tests
  - Μετρήσεις: αύξησης πωλήσεων, ποσοστά εμφάνισης/πώλησης
- Εργαστηριακά πειράματα
  - Ελεγχόμενα πειράματα
  - Μετρήσεις: ικανοποίηση από το σύστημα (π.χ. ερωτηματολόγια)
- Offline πειράματα
  - Βασισμένα σε ιστορικά δεδομένα
  - Μετρήσεις: ακρίβεια πρόβλεψης, κάλυψη

# Μέτρα: Ακρίβεια και Ποσοστό Ανάκλησης

- Ακρίβεια: μέτρο ακρίβειας, το ποσοστό των σχετικών αντικειμένων που ανακτώνται σε σχέση με όλα τα ανακτημένα αντικείμενα

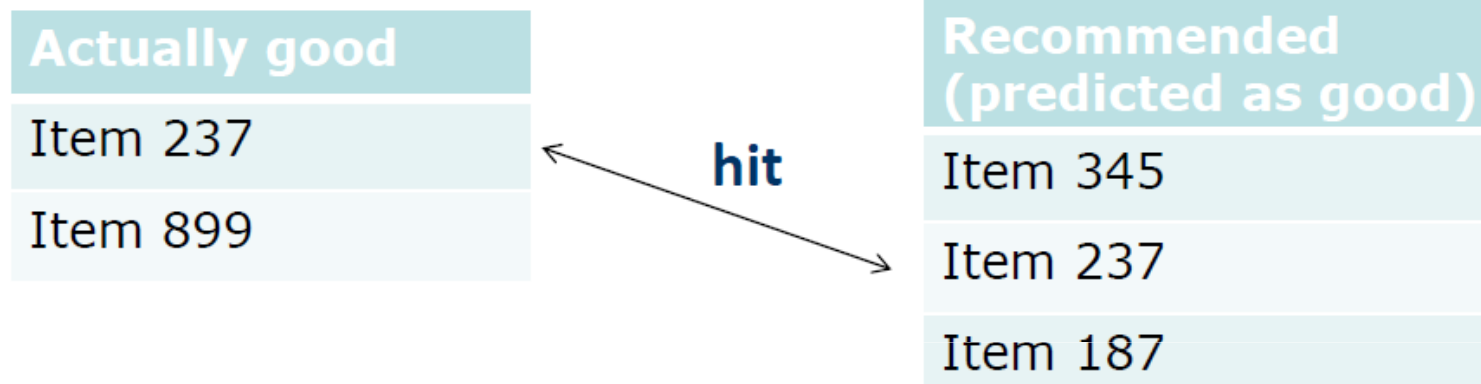
$$\text{Precision} = (\text{true positives}) / (\text{true positives} + \text{false positives})$$

- Ποσοστό ανάκλησης: μέτρο πληρότητας, το ποσοστό των σχετικών αντικειμένων που ανακτώνται σε σχέση με όλα τα σχετικά αντικείμενα

$$\text{Recall} = (\text{true positives}) / (\text{true positives} + \text{false negatives})$$

# Μέτρα: Rank Score – μετράει η θέση

**For a user:**



- Το Rank Score επεκτείνει την ακρίβεια και το ποσοστό ανάκλησης λαμβάνοντας υπόψη τις θέσεις σωστών αντικειμένων σε μια βαθμολογημένη λίστα
  - Πολύ σημαντικό γιατί τα χαμηλά βαθμολογημένα αντικείμενα μπορεί να μη τα δει ο χρήστης

# Μέτρα Ακρίβειας

- Ιστορικές βαθμολογίες χρηστών καθιστούν την «αλήθεια»
- Μέτρηση του λάθους
  - Μέσο Απόλυτο Λάθος (Mean Absolute Error) υπολογίζει την απόκλιση μεταξύ προβλεπόμενων βαθμολογιών και πραγματικών βαθμολογιών

$$MAE = \frac{1}{n} \sum_{i=1}^n |p_i - r_i|$$

# Είδη Συστάσεων

- Του εκδότη
- Απλές συναθροίσεις
  - Τα κορυφαία 10 (Top 10), πιο δημοφιλή, τις πιο πρόσφατες μεταφορτώσεις
- Ραμμένες στα μέτρα του κάθε χρήστη
  - Amazon, Netflix,...

# Συστήματα Συστάσεων

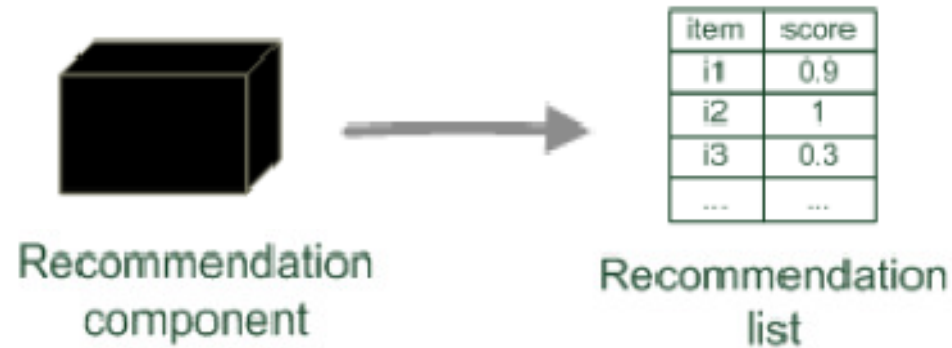
- Ένα σύστημα συστάσεων ως συνάρτηση
- Δεδομένου:
  - Ενός μοντέλου χρήστη (π.χ. βαθμολογίες, προτιμήσεις, δημογραφικά, καταστασιακές πληροφορίες)
  - Αντικειμένων (με ή χωρίς περιγραφή των χαρακτηριστικών τους)
- Βρες:
  - Το σκορ συσχέτισης, το οποίο χρησιμοποιείται για την διάταξη των αποτελεσμάτων
- Τελικά:
  - Σύστησε αντικείμενα που θεωρούνται ως σχετικά
- Αλλά:
  - Προσοχή γιατί η «σχετικότητα» μπορεί να εξαρτάται από το περιβάλλον/κατάσταση/συμφραζόμενα
  - Τα χαρακτηριστικά της ίδιας της λίστας συστάσεων μπορεί να είναι σημαντικά (π.χ. η ποικιλία τους (diversity))

# Φορμαλιστικό μοντέλο

- $C$  = σύνολο πελατών
- $S$  = σύνολο αντικειμένων
- Συνάρτηση οφέλους  $u: C \times S \rightarrow R$ 
  - $R$  = σύνολο βαθμολογιών (ratings)
  - Το  $R$  είναι ένα διατεταγμένο σύνολο
  - Π.χ., 0-5 αστέρια, πραγματικός αριθμός στο  $[0, 1]$

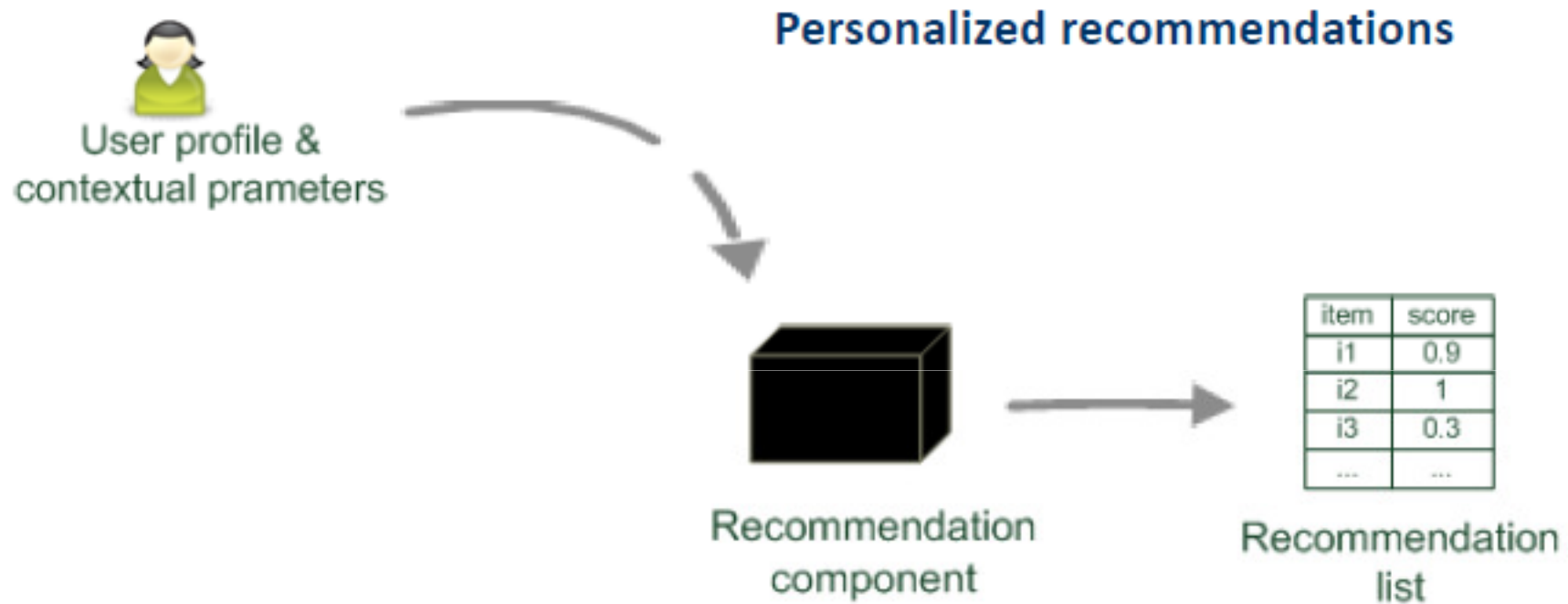


# Λειτουργία Συστημάτων Συστάσεων

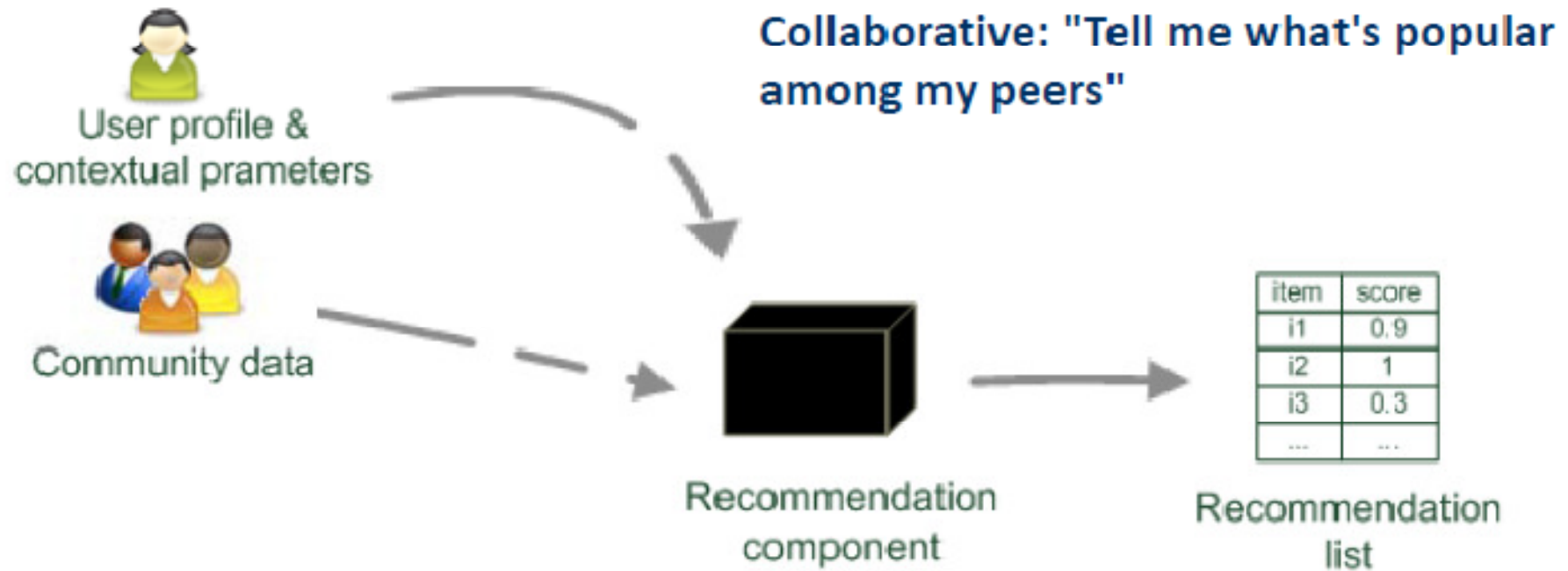


- Μειώνουν την υπερφόρτωση πληροφορίας εκτιμώντας τη σχετικότητα

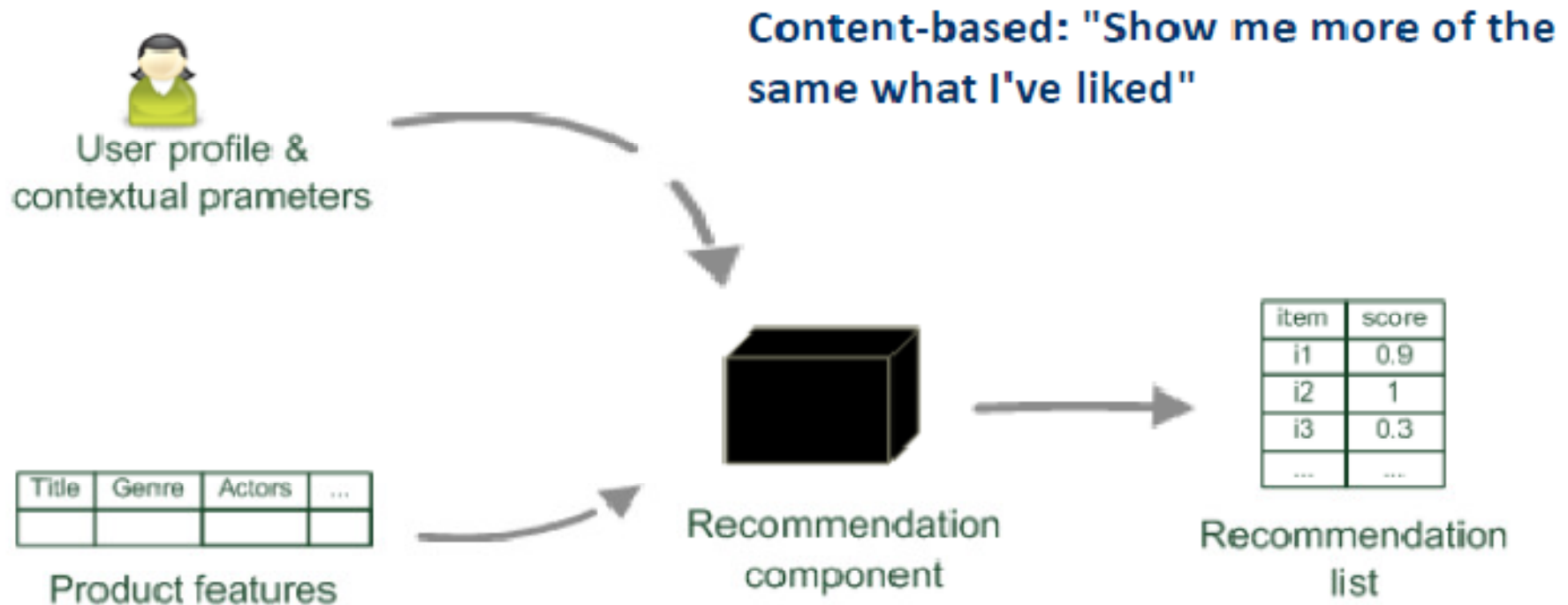
# Είδη Συστημάτων Συστάσεων



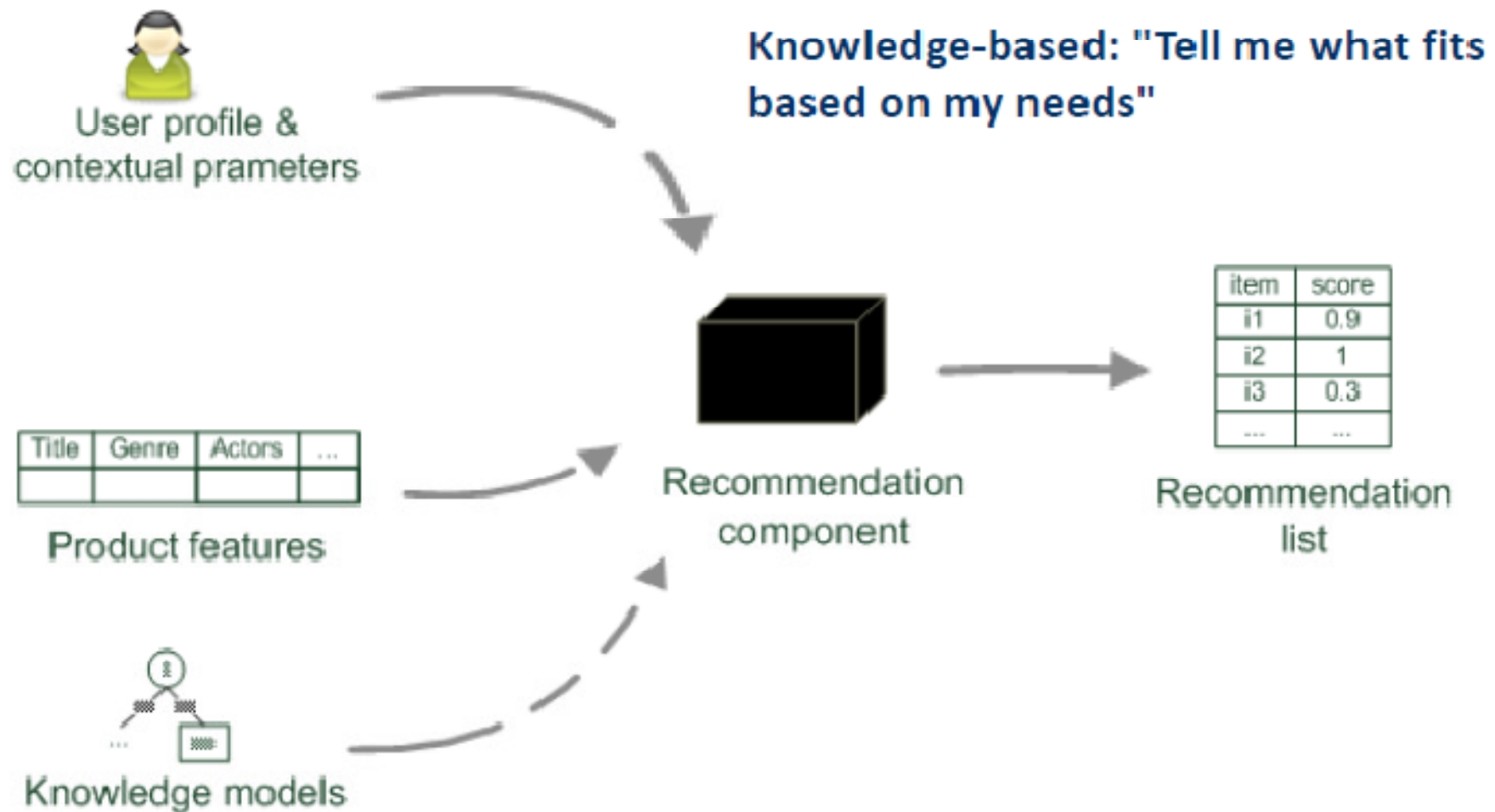
# Είδη Συστημάτων Συστάσεων



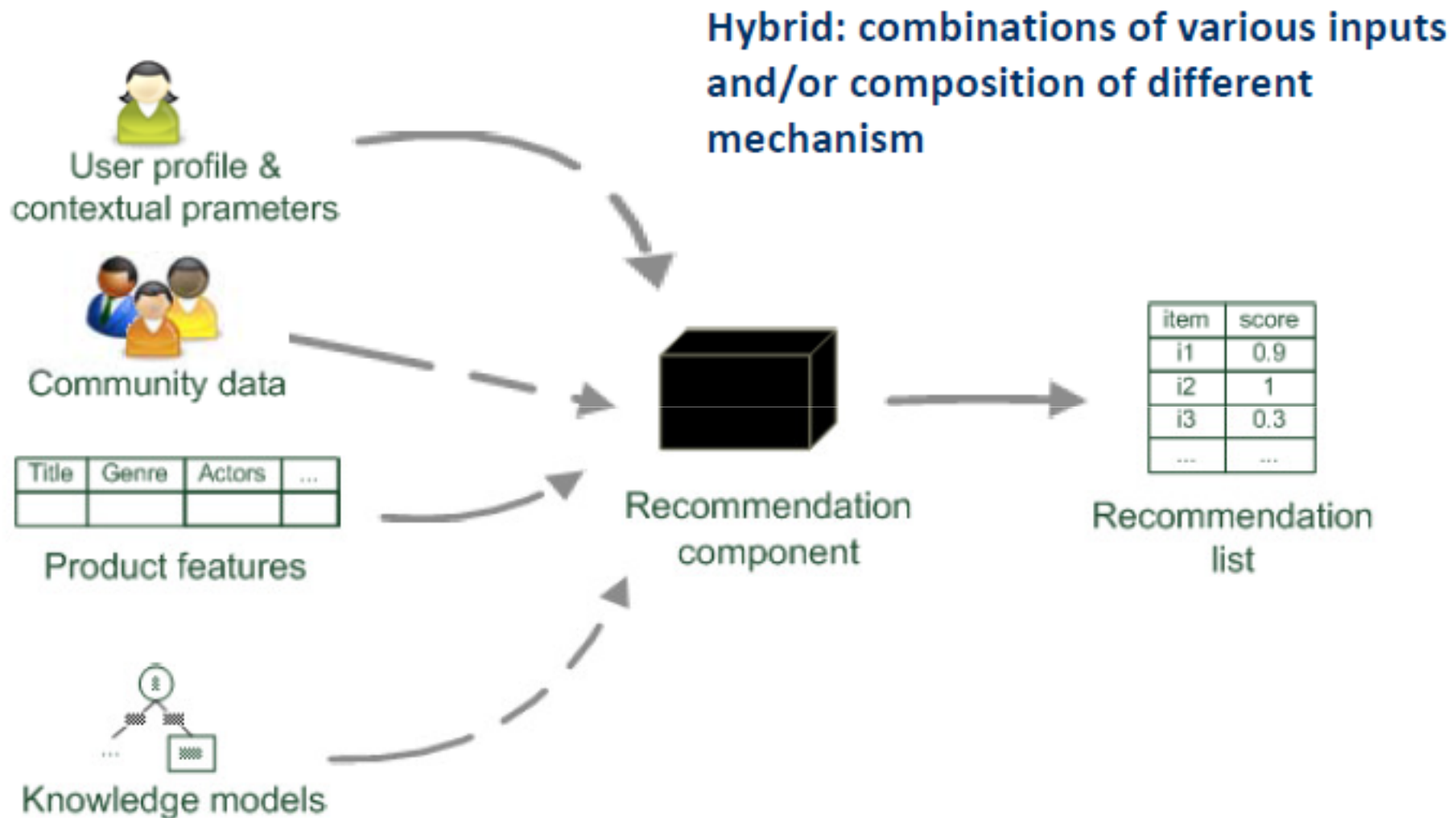
# Είδη Συστημάτων Συστάσεων



# Είδη Συστημάτων Συστάσεων



# Είδη Συστημάτων Συστάσεων



# Πίνακας Οφέλους

	King Kong	LOTR	Matrix	Nacho Libre
Alice	1		0.2	
Bob		0.5		0.3
Carol	0.2		1	
David				0.4

# Βασικά Προβλήματα

1. Συλλογή «γνωστών» βαθμολογιών για την κατασκευή του πίνακα
2. Εξαγωγή/ συμπερασμός άγνωστων βαθμολογιών από τις γνωστές βαθμολογίες
  - Κυρίως ενδιαφερόμαστε για υψηλές άγνωστες βαθμολογίες
3. Αποτίμηση μεθόδων συμπερασμού



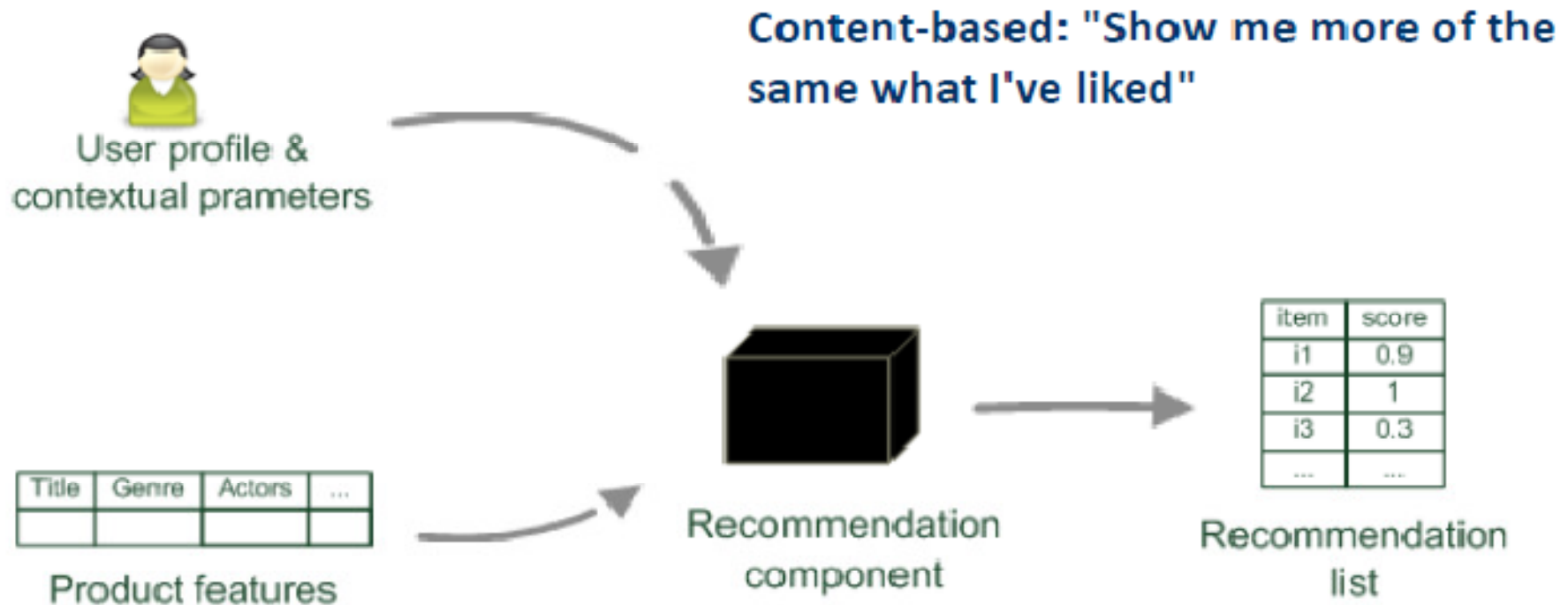
# Συλλογή Βαθμολογιών

- Ρητά
  - Ζητάμε από τον κόσμο να βαθμολογεί αντικείμενα
  - Δεν δουλεύει καλά στην πράξη – ο κόσμος βαριέται
- Έμμεσα
  - Μαθαίνουμε τις βαθμολογίες από τις ενέργειες των χρηστών
  - Π.χ., η αγορά σημαίνει μεγάλο βαθμό
  - Πώς καταλαβαίνουμε χαμηλές βαθμολογίες;

# Εργαλεία Συμπερασμού

- **Βασικό πρόβλημα:** ο πίνακας U είναι αραιός
  - Οι περισσότεροι άνθρωποι δεν έχουν βαθμολογήσει τα περισσότερα αντικείμενα
  - **Cold start:**
    - Τα νέα αντικείμενα δεν έχουν βαθμολογίες
    - Οι νέοι χρήστες δεν έχουν ιστορικό
- Διαφορετικές προσεγγίσεις:
  - Προσωποποιημένες
  - Με βάση το περιεχόμενο
  - Με βάση τη γνώση
  - Συνεργατικές
  - Υβριδικές

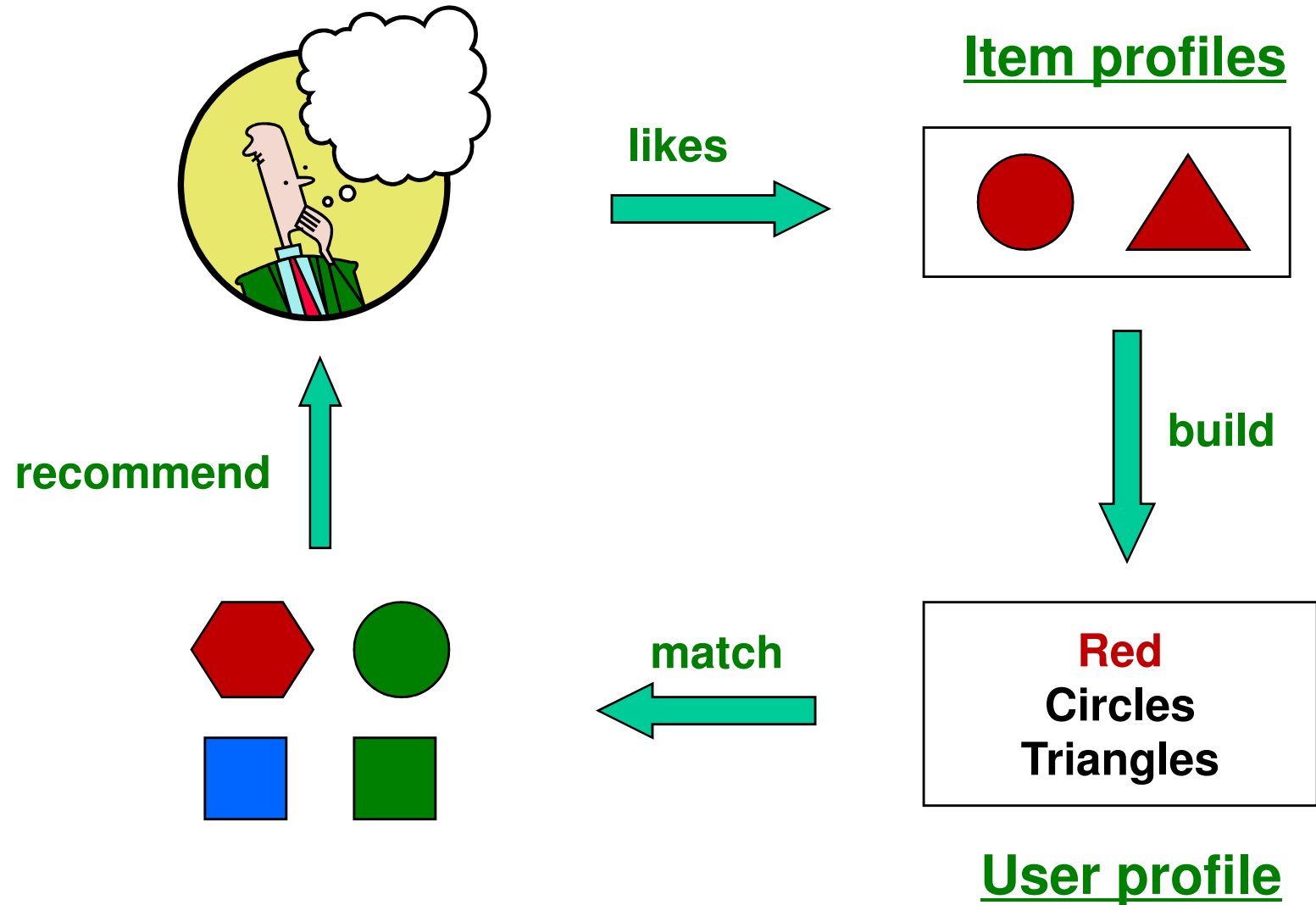
# Είδη Συστημάτων Συστάσεων



# Συστάσεις με βάση το περιεχόμενο


- Βασική ιδέα: συστήνουμε αντικείμενα στον πελάτη  $c$  παρόμοια με άλλα αντικείμενα που βαθμολόγησε προηγουμένως με υψηλό βαθμό ο  $c$
- Συστάσεις ταινιών
  - Σύστησε ταινίες με τον ίδιο(ους) ηθοποιό(ους), σκηνοθέτη, είδος....
- Ιστότοπους, *blogs*, νέα
  - Σύστησε άλλους ιστότοπους με «παρόμοιο» περιεχόμενο
- Τι χρειαζόμαστε:
  - Κάποιες πληροφορίες για τα διαθέσιμα αντικείμενα όπως είδος, σκηνοθέτης...
  - Κάποιο είδος προφίλ χρήστη που περιγραφεί τι αρέσει στους χρήστες (τις προτιμήσεις τους)
- Η προσέγγιση:
  - Μάθε τις προτιμήσεις των χρηστών
  - Εντόπισε /σύστησε αντικείμενα που είναι «όμοια» με τις προτιμήσεις του χρήστη

# Σχέδιο Δράσης



# Προφίλ Αντικειμένων

- Για κάθε αντικείμενο, δημιουργήσε ένα προφίλ αντικειμένου
- Το προφίλ είναι ένα σύνολο από χαρακτηριστικά
  - Για ταινίες: σεναριογράφος, τίτλος, ηθοποιός, σκηνοθέτης,...
  - Για κείμενο: σύνολο «σημαντικών» λέξεων σε ένα έγγραφο
- Πώς επιλέγουμε τις σημαντικές λέξεις;
  - Η συνήθης ευριστική είναι το TF- IDF (term frequency  $\times$  Inverse Document Frequency, συχνότητα όρου  $\times$  ανάστροφη συχνότητα εγγράφου)



Title	Genre	Author	Type	Price	Keywords
The Night of the Gun	Memoir	David Carr	Paperback	29.90	Press and journalism, drug addiction, personal memoirs, New York
The Lace Reader	Fiction, Mystery	Brunonia Barry	Hardcover	49.90	American contemporary fiction, detective, historical
Into the Fire	Romance, Suspense	Suzanne Brockmann	Hardcover	45.90	American fiction, murder, neo-Nazism

# Αναπαράσταση περιεχομένου και ομοιότητα αντικειμένων

## Item representation

Title	Genre	Author	Type	Price	Keywords
The Night of the Gun	Memoir	David Carr	Paperback	29.90	Press and journalism, drug addiction, personal memoirs, New York
The Lace Reader	Fiction, Mystery	Brunonia Barry	Hardcover	49.90	American contemporary fiction, detective, historical
Into the Fire	Romance, Suspense	Suzanne Brockmann	Hardcover	45.90	American fiction, murder, neo-Nazism

## User profile

Title	Genre	Author	Type	Price	Keywords
...	Fiction	Brunonia, Barry, Ken Follett	Paperback	25.65	Detective, murder, New York

$keywords(b_f)$   
describes Book  $b_f$   
with a set of  
keywords

## Simple approach

- Compute the similarity of an unseen item with the user profile based on the keyword overlap (e.g. using the Dice coefficient)



$$\frac{2 \times |keywords(b_i) \cap keywords(b_f)|}{|keywords(b_i)| + |keywords(b_f)|}$$

- Or use and combine multiple metrics

# TF-IDF

- Η απλή αναπαράσταση με λέξεις κλειδιά έχει τα προβλήματά της—ειδικά όταν η εξαγωγή των λέξεων κλειδιών γίνεται αυτόματα
  - Δεν έχουν όλες οι λέξεις την ίδια σημασία
  - Μεγαλύτερα έγγραφα έχουν μεγαλύτερη πιθανότητα να έχουν επικάλυψη με το προφίλ χρήστη
- Σύνηθες μέτρο:
  - Κωδικοποιεί τα έγγραφα κειμένου ως διανύσματα όρων με βάρη
  - TF: μετράει πόσο συχνά εμφανίζεται ένας όρος (πυκνότητα σε ένα έγγραφο)
    - Υποθέτει ότι πιο σημαντικοί όροι εμφανίζονται περισσότερες φορές
    - Κανονικοποίηση γίνεται για να λάβουμε υπόψη και το μέγεθος του κειμένου
  - IDF: στοχεύει στο να μειώσει το βάρος όρων που εμφανίζονται σε όλα τα έγγραφα



# TF - IDF

- $f_{ij}$  = συχνότητα του όρου  $t_i$  στο έγγραφο  $d_j$

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$$

- $n_i$  = πλήθος εγγράφων που αναφέρουν τον όρο  $i$
- $N$  = συνολικός αριθμός εγγράφων

$$IDF_i = \log \frac{N}{n_i}$$

- Σκορ TF.IDF  $w_{ij} = TF_{ij} \times IDF_i$
- Προφίλ εγγράφων = σύνολο λέξεων με τα μεγαλύτερα σκορ TF.IDF, μαζί με αυτά τα σκορ

# Παράδειγμα

	<b>Antony and Cleopatra</b>	<b>Julius Caesar</b>	<b>The Tempest</b>	<b>Hamlet</b>	<b>Othello</b>	<b>Macbeth</b>
<b>Antony</b>	157	73	0	0	0	0
<b>Brutus</b>	4	157	0	1	0	0
<b>Caesar</b>	232	227	0	2	1	1
<b>Calpurnia</b>	0	10	0	0	0	0
<b>Cleopatra</b>	57	0	0	0	0	0
<b>mercy</b>	1.51	0	3	5	5	1
<b>worser</b>	1.37	0	1	1	1	0

# Παράδειγμα

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4					
Caesar	23					
Calpurnia	0					
Cleopatra	57					
mercy	1.5					
worser	1.3					
	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	5.25	3.18	0	0	0	0.35
Brutus	1.21	6.1	0	1	0	0
Caesar	8.59	2.54	0	1.51	0.25	0
Calpurnia	0	1.54	0	0	0	0
Cleopatra	2.85	0	0	0	0	0
mercy	1.51	0	1.9	0.12	5.25	0.88
worser	1.37	0	0.11	4.15	0.25	1.95

# Βελτιώσεις του μοντέλου

- Τέτοια διανύσματα είναι συνήθως μεγάλα και αραιά
- Αφαίρεση stop words
  - Εμφανίζονται σε όλα τα έγγραφα
  - Π.χ. άρθρα, προθέσεις, κ.α...
- Χρήση stemming
  - Στοχεύει στην αφαίρεση εναλλακτικών λέξεων με την ίδια ρίζα
  - Π.χ. Stemming-stem
- Αποκοπή μεγέθους
  - Χρησιμοποίησε μόνο τις κορυφαίες η αντιπροσωπευτικές λέξεις για να αφαιρέσουμε τον «θόρυβο» από τα δεδομένα
  - Π.χ. τις 100 top λέξεις

# Βελτιώσεις του μοντέλου

- Χρήση λεξικών, χρήση πιο έξυπνων μεθόδων για την επιλογή χαρακτηριστικών
  - Αφαίρεση λέξεων που δεν σχετίζονται με τον τομέα
- Αναγνώριση φράσεων ως όρους
  - Πιο περιγραφικές για ένα κείμενο από μία μεμονωμένη λέξη
  - Π.χ., Ηνωμένα Έθνη
- Περιορισμοί
  - Παραμένει άγνωστο το σημασιολογικό περιεχόμενο
  - Παράδειγμα: χρήση λέξης με αρνητικό νόημα
    - «δεν υπάρχει τίποτα στο μενού που θα άρεσε σε έναν χορτοφάγο»
    - Η λέξη «χορτοφάγος» θα λάβει εφאלμένα μεγαλύτερο βάρος από το επιθυμητό
    - Αποτέλεσμα: ένα ανεπιθύμητο ταίριασμα με έναν χρήστη που ενδιαφέρεται για χορτοφαγικά εστιατόρια

# Προφίλ Χρηστών και προβλέψεις

- Πιθανά προφίλ χρηστών:
  - Ζυγισμένος μέσος όρος των βαθμολογημένων προφίλ αντικειμένων
  - Παραλλαγή: βάρος με βάση τη διαφορά από το μέσο βαθμό του αντικειμένου
  - ....
- Ευριστική πρόβλεψης:
  - Δεδομένου ενός προφίλ χρήστη  $c$  και ενός προφίλ αντικειμένου  $s$ , εκτίμησε το  $u(c,s) = \cos(c,s) = c \cdot s / (|c||s|)$
  - Χρειαζόμαστε μια αποδοτική μέθοδο για να εντοπίζουμε αντικείμενα με μεγάλο όφελος

# Απλή μέθοδος: κοντινότερος γείτονας

- Δεδομένου ενός συνόλου εγγράφων  $D$  που έχει βαθμολογήσει ένας χρήστης
  - Είτε άμεσα μέσω κάποιας διεπαφής χρήστη
  - Είτε έμμεσα παρακολουθώντας τη συμπεριφορά του χρήστη
- Βρες τους  $n$  κοντινότερους γείτονες στο  $D$  για ένα αντικείμενο  $i$  που δεν έχει δει ακόμα ο χρήστης
  - Χρήση μέτρων ομοιότητας
- Καλή μέθοδος για τις βραχυπρόθεσμες συστάσεις /follow-up stories

# Αναζήτηση βάση ερώτησης

- Η ποιότητα εξαρτάται από την ατομική ικανότητα να θέτουμε ερωτήσεις με τις σωστές λέξεις κλειδιά
- Στους χρήστες επιτρέπεται να βαθμολογούν ως σχετικά/άσχετα τα έγγραφα που επιστρέφονται (ανάδραση)
- Το σύστημα μαθαίνει ένα πρότυπο για σχετικά και άσχετα έγγραφα
- Οι ερωτήσεις επεκτείνονται στη συνέχεια με επιπρόσθετους όρους/βάρη από σχετικά έγγραφα
- Απαιτεί έναν ικανοποιητικό αριθμό βαθμολογημένων αντικειμένων για να μάθει
  - Δυνατότητα έμμεσης αυτοματοποιημένης βαθμολογίας
- Η αλληλεπίδραση με το χρήστη ανοίγει νέους δρόμους
  - Αλληλεπιδραστική βελτίωση των ερωτήσεων
  - Βοηθά τους χρήστες να μάθουν το λεξιλόγιο που θα πρέπει να χρησιμοποιούν για να ικανοποιήσουν τις ανάγκες τους



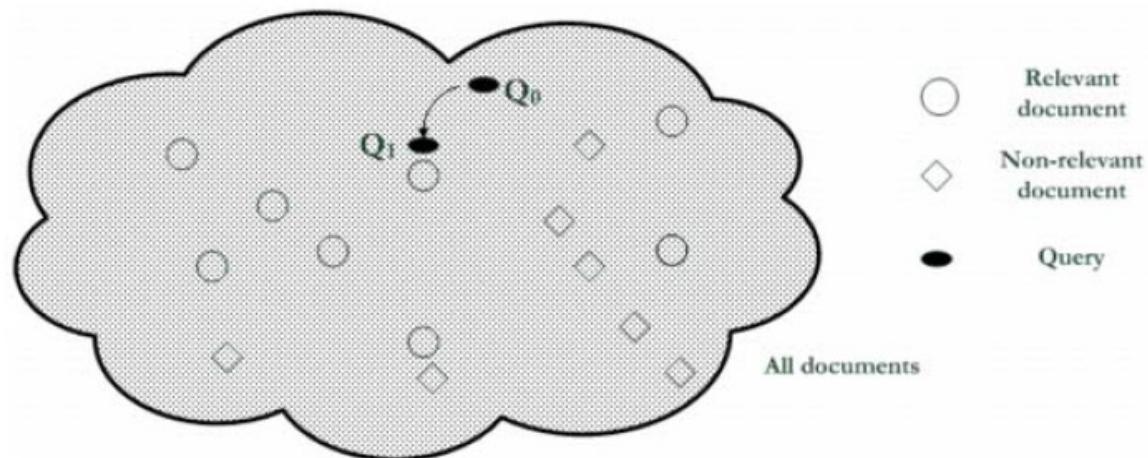
## Rocchio details

---

- Document collections  $D^+$  and  $D^-$
- $\alpha, \beta, \gamma$  used to fine-tune the feedback
- often only positive feedback is used



$$Q_{i+1} = \alpha * Q_i + \beta \left( \frac{1}{|D^+|} \sum_{d^+ \in D^+} d^+ \right) - \gamma \left( \frac{1}{|D^-|} \sum_{d^- \in D^-} d^- \right)$$



# Προσεγγίσεις με βάση πιθανοτικά μοντέλα

- Για κάθε χρήστη, μάθε έναν ταξινομητή που ταξινομεί τα αντικείμενα σε βαθμολογικές κλάσεις
  - Που αρέσουν στο χρήστη, που δεν αρέσουν στο χρήστη
  - Τεχνικές/μοντέλα τεχνητής νοημοσύνης/μηχανικής μάθησης
- Εφάρμοσε τον ταξινομητή σε κάθε αντικείμενο για να βρεις υποψήφιους προς σύσταση
- Πρόβλημα: κλιμάκωση

# Απλή Πιθανοτική Προσέγγιση

- 2 κλάσεις: αρέσει/δεν αρέσει
- Απλή λογική αναπαράσταση εγγράφων
- Υπολογισμός της πιθανότητας ένα έγγραφο να αρέσει ή να μην αρέσει με βάση το θεώρημα του Bayes

Doc-ID	recommender	intelligent	learning	school	Label
1	1	1	1	0	1
2	0	0	1	1	0
3	1	1	0	0	1
4	1	0	1	1	1
5	0	0	0	1	0
6	1	1	0	0	?

$$\begin{aligned}P(X|Label = 1) &= P(recommender = 1|Label = 1) \\&\times P(intelligent = 1|Label = 1) \\&\times P(learning = 0|Label = 1) \\&\times P(school = 0|Label = 1) \\&= \frac{3}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} \approx 0.149\end{aligned}$$

# Μοντέλα αποφάσεων

- Χρήση δέντρου απόφασης για προβλήματα συστάσεων
  - Οι εσωτερικοί κόμβοι έχουν ως ετικέτα τα χαρακτηριστικά των αντικειμένων (λέξεις κλειδιά)
  - Χρησιμοποιούνται για το διαμερισμό των παραδειγμάτων ελέγχου
    - Ύπαρξη ή μη μιας λέξης κλειδιού
  - Στα φύλλα εμφανίζονται δύο κλάσεις
    - Ενδιαφέρον –μη ενδιαφέρον
  - Το δέντρο μπορεί να κατασκευαστεί αυτόματα από τα δεδομένα εκπαίδευσης
  - Δουλεύει καλύτερα για λίγα χαρακτηριστικά
  - Χρησιμοποιεί μετά-δεδομένα αντί για το TF-IDF

# Επιλογή Χαρακτηριστικών

- Διαδικασία επιλογής υποσυνόλου διαθέσιμων όρων
- Διαφορετικές στρατηγικές υπάρχουν για την χρήση των χαρακτηριστικών
  - Με βάση λεξικών και γνώσης του τομέα
  - Επιλογή με βάση τη συχνότητα για την απαλοιφή λέξεων που εμφανίζονται πολύ συχνά και πολύ σπάνια

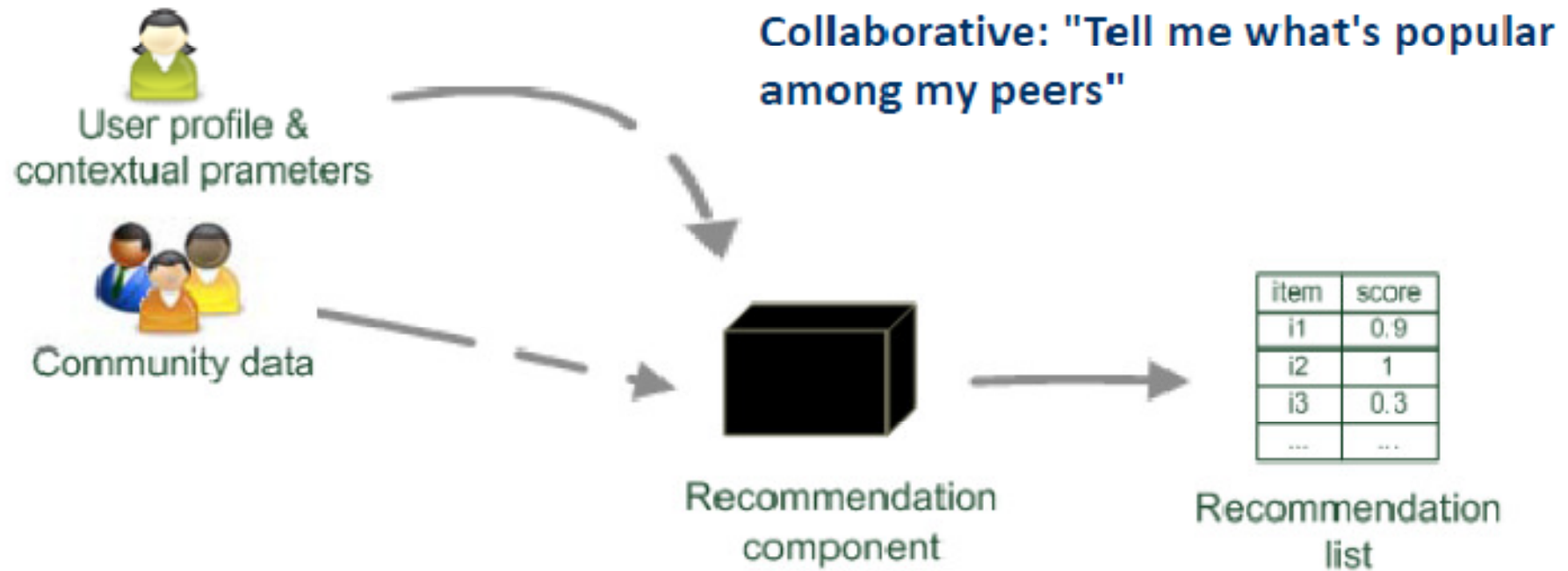
## Πλεονεκτήματα συστάσεων με βάση το περιεχόμενο

- **Δεν υπάρχει ανάγκη για δεδομένα από άλλους χρήστες**
  - Δεν εμφανίζεται το πρόβλημα του cold-start ή του αραιού πίνακα
- **Μπορεί να δώσει συστάσεις σε χρήστες με μοναδικά γούστα**
- **Μπορεί να συστήσει νέα και μη δημοφιλή αντικείμενα**
  - Δεν εμφανίζεται το πρόβλημα της πρώτης βαθμολογίας (first-rater problem)
- **Μπορεί να παρέχει εξηγήσεις**
  - Μπορεί να δώσει εξηγήσεις για τα αντικείμενα που συστήνει δίνοντας τα χαρακτηριστικά του περιεχομένου που προκάλεσαν τη σύσταση

# Περιορισμοί προσεγγίσεων με βάση το περιεχόμενο

- **Εύρεση των κατάλληλων χαρακτηριστικών**
  - Π.χ., εικόνες, ταινίες, μουσική
- **Υπερεξειδίκευση**
  - Ποτέ δεν προτείνει αντικείμενα έξω από το προφίλ περιεχομένου του χρήστη
  - Οι άνθρωποι έχουν πολλαπλά ενδιαφέροντα
- **Συστάσεις για νέους χρήστες**
  - Πώς κατασκευάζουμε το προφίλ του χρήστη;
- **Οι λέξεις κλειδιά μπορεί να μην είναι επαρκείς για να κρίνουμε τη σχετικότητα ενός εγγράφου ή ιστότοπου**
  - Ενημερότητα, χρηστικότητα, αισθητική
  - Το περιεχόμενο μπορεί να είναι πολύ περιορισμένο
  - Περιεχόμενο μπορεί να μην εξάγεται αυτόματα

# Είδη Συστημάτων Συστάσεων





# Συνεργατικό Φιλτράρισμα – Collaborative Filtering (CF)

- Η πιο βασική προσέγγιση για τη δημιουργία συστάσεων
  - Χρησιμοποιείται από μεγάλες, εμπορικές τοποθεσίες ηλεκτρονικού εμπορίου
  - Καλά κατανοητή προσέγγιση, με πολλούς διαθέσιμους αλγορίθμους και παραλλαγές
  - Εφαρμόσιμη σε πολλούς τομείς (βιβλία, ταινίες, DVDs,...)
- Προσέγγιση:
  - Χρησιμοποίησε τη «σοφία του πλήθους» για να συστήσεις αντικείμενα
- Βασική ιδέα και υπόθεση:
  - Οι χρήστες δίνουν βαθμολογίες (έμμεσα ή και άμεσα) για να καταλογογραφήσουν τα αντικείμενα
  - Οι πελάτες που είχαν παρόμοια γούστα στο παρελθόν, θα έχουν παρόμοια γούστα και στο μέλλον

# Συνεργατικό Φιλτράρισμα

- Θεωρήστε έναν χρήστη  $c$  και ένα αντικείμενο  $i$  που δεν έχει βαθμολογήσει ο  $c$
- Βρες ένα σύνολο  $D$  από άλλους χρήστες των οποίων οι βαθμολογίες σε άλλα αντικείμενα είναι «παρόμοιες» με τις βαθμολογίες του  $c$
- Εκτίμησε τη βαθμολογία για το  $i$  του χρήστη  $c$  με βάση τις βαθμολογίες για το  $i$  των άλλων χρηστών στο  $D$

# Βασική Τεχνική: κοντινότερου γείτονα

- Δεδομένου ενός ενεργού χρήστη (της Αλίκης) και ενός αντικειμένου  $i$  που δεν έχει δει ακόμα η Αλίκη
- Στόχος να εκτιμηθεί ο βαθμός της Αλίκης για το  $i$ 
  - Βρες ένα σύνολο χρήστες που τους άρεσαν τα ίδια αντικείμενα με την Αλίκη στο παρελθόν και έχουν βαθμολογήσει το αντικείμενο  $i$
  - Χρησιμοποίησε το μέσο όρο των βαθμολογιών τους για να προβλέψεις αν το αντικείμενο θα αρέσει στην Αλίκη
  - Κάνε αυτά τα βήματα για όλα τα αντικείμενα που δεν έχει δει η Αλίκη και πρότεινε τα κορυφαία
- Βασική υπόθεση:

Οι προτιμήσεις των χρηστών παραμένουν σταθερές και συνεπείς στο χρόνο

# Παράδειγμα

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

# Βασικές ερωτήσεις

- Πώς μετράμε την ομοιότητα;
- Πόσους γείτονες πρέπει να λάβουμε υπόψη;
- Πώς παράγουμε μια πρόβλεψη από τις βαθμολογίες των γειτόνων;

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

# Μέτρα Ομοιότητας

- Έστω  $r_x$  το διάνυσμα των βαθμολογιών του χρήστη  $x$

$$\underline{r}_x = [*, \_, \_, *, ***]$$

$$\underline{r}_y = [*, \_, **, **, \_]$$

- Jaccard ομοιότητα

$$J(r_x, r_y) = \frac{r_x \cap r_y}{r_x \cup r_y}$$

$\underline{r}_x, \underline{r}_y$  as sets:

$$\underline{r}_x = \{1, 4, 5\}$$

$$\underline{r}_y = \{1, 3, 4\}$$

- **Πρόβλημα:** Αγνοεί την τιμή της βαθμολογίας
- Το μέτρο της ομοιότητας συνημίτονου (cosine similarity)

$$\text{sim}(\mathbf{x}, \mathbf{y}) = \cos(\mathbf{r}_x, \mathbf{r}_y) = \frac{\mathbf{r}_x \cdot \mathbf{r}_y}{\|\mathbf{r}_x\| \cdot \|\mathbf{r}_y\|}$$

$\underline{r}_x, \underline{r}_y$  as points:

$$\underline{r}_x = \{1, 0, 0, 1, 3\}$$

$$\underline{r}_y = \{1, 0, 2, 2, 0\}$$

- **Πρόβλημα:** Θεωρεί την απουσία βαθμολογίας ως αρνητική βαθμολογία

# Παράδειγμα

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

- **Intuitively we want:**  $\text{sim}(A, B) > \text{sim}(A, C)$

- **Jaccard similarity:**  $1/5 < 2/4$

- **Cosine similarity:**  $0.386 > 0.322$

– Considers missing ratings as “negative”

– **Solution: subtract the (row) mean**

**sim A,B vs. A,C:**  
**0.092 > -0.559**

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	2/3			5/3	-7/3		
B	1/3	1/3	-2/3				
C				-5/3	1/3	4/3	
D		0					0

# Μέτρα Ομοιότητας

- Συντελεστής Pearson (**Pearson correlation coefficient**)

- $S_{xy}$  = αντικείμενα που βαθμολογούνται και από τους δύο χρήστες  $x$  και  $y$
- Ομοιότητα μεταξύ  $-1$  και  $1$

$$sim(x,y) = \frac{\sum_{s \in S_{xy}} (r_{xs} - \bar{r}_x)(r_{ys} - \bar{r}_y)}{\sqrt{\sum_{s \in S_{xy}} (r_{xs} - \bar{r}_x)^2} \sqrt{\sum_{s \in S_{xy}} (r_{ys} - \bar{r}_y)^2}}$$

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1



sim = 0,85

sim = 0,00

sim = 0,70

sim = -0,79



# Προβλέψεις βαθμολογίας

- Έστω  $D$  το σύνολο των  $k$  χρηστών που είναι πιο όμοιοι στον  $c$  και έχουν βαθμολογήσει το αντικείμενο  $s$
- Πιθανές συναρτήσεις πρόβλεψης (για το αντικείμενο  $s$ ):
  - $r_{cs} = \frac{1}{k} \sum_{d \in D} r_{ds}$
  - $r_{cs} = \frac{(\sum_{d \in D} sim(c, d) \times r_{ds})}{\sum_{d \in D} sim(c, d)}$
- Πολλές διαθέσιμες επιλογές

# Βελτιώσεις της συνάρτησης πρόβλεψης

- Πιθανόν οι βαθμολογίες όλων των γειτόνων να μην έχουν την ίδια αξία
  - Η συμφωνία για γενικά (κοινά) αρεστά αντικείμενα δεν έχει τόση πληροφοριακή αξία όσο η συμφωνία σε αμφιλεγόμενα αντικείμενα
  - Πιθανή λύση: Δίνουμε περισσότερο βάρος σε αντικείμενο που εμφανίζουν μεγαλύτερη διακύμανση
- Αξία του πλήθους των κοινών βαθμολογημένων αντικειμένων
  - Χρήση «βάρους σημαντικότητας», για παράδειγμα γραμμική μείωση του βάρους όταν το πλήθος των κοινών βαθμολογημένων αντικειμένων είναι χαμηλό
- Επιλογή γειτονιάς
  - Χρήση κατωφλιού ομοιότητας ή προκαθορισμένου αριθμού γειτόνων

# Πολυπλοκότητα

- Το ακριβό βήμα είναι η εύρεση των  $k$  πιο όμοιων πελατών
  - $O(|U|)$
- Πολύ ακριβό για να γίνει την ώρα της εκτέλεσης
  - Επιβάλλει τον προ-υπολογισμό
- Η αφελής προσέγγιση απαιτεί χρόνο:  $O(N|U|)$
- Μπορεί να γίνει χρήση συσταδοποίησης, αλλά πέφτει η ποιότητα

# Συνεργατικό φιλτράρισμα με βάση τα αντικείμενα

- Μέχρι τώρα: συνεργατικό φιλτράρισμα **χρήστη-χρήστη**  
Χρήση της ομοιότητας μεταξύ αντικειμένων (και όχι χρηστών)  
για τις προβλέψεις
- Μια άλλη όψη
  - Για ένα αντικείμενο  $s$ , βρες άλλα παρόμοια αντικείμενα
  - Εκτίμησε τον βαθμό του αντικειμένου με βάση βαθμολογίες παρόμοιων αντικειμένων
  - Μπορούμε να χρησιμοποιήσουμε τις ίδιες μετρικές ομοιότητας και πρόβλεψης όπως και στο μοντέλο χρήστη-χρήστη
- Στην πράξη, έχει παρατηρηθεί ότι το μοντέλο αντικείμενο-αντικείμενο δουλεύει καλύτερα από το μοντέλο χρήστη-χρήστη
- Θεωρούνται πιο σταθερές οι ομοιότητες των αντικειμένων από τις ομοιότητες χρηστών

# Παράδειγμα

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

# Πλεονεκτήματα και μειονεκτήματα συνεργατικού φιλτραρίσματος

- Δουλεύει για οποιοδήποτε είδος αντικειμένου
  - Δεν απαιτείται καμία επιλογή χαρακτηριστικού
- Πρόβλημα νέου χρήστη
- Πρόβλημα νέου αντικειμένου
- Αραιός πίνακας βαθμολογιών
  - Συσταδοποίηση για λείανση

# Ρητές Βαθμολογίες

- Πιθανότατα οι πιο ακριβείς
- Πιο συχνά χρησιμοποιούμενες (1 έως 5, 1 έως 7)
- Ερευνητικά ζητήματα:
  - Η βέλτιστη κοκκικοποίηση, π.χ. για ταινίες καλύτερα από 1 έως 10
  - Πολυδιάστατες βαθμολογίες
- Βασικά προβλήματα:
  - Οι χρήστες δεν θέλουν να βαθμολογούν αντικείμενα
  - Πώς θα κινητοποιήσουμε τους χρήστες για να βαθμολογούν περισσότερο;

# Έμμεσες Βαθμολογίες

- Συνήθως συλλέγονται από ένα web shop ή μια εφαρμογή στην οποία ενσωματώνεται το σύστημα συστάσεων
- Όταν ο πελάτης αγοράσει ένα αντικείμενο, πολλά συστήματα ερμηνεύουν την πράξη ως μια θετική βαθμολόγηση
- Τα κλικ, οι επισκέψεις, ο χρόνος επίσκεψης, τα downloads...
- Έμμεσες βαθμολογίες μπορούν να συλλέγονται συνεχώς και δεν απαιτούν πρόσθετη προσπάθεια από την πλευρά του χρήστη
- Βασικό πρόβλημα
  - Δεν μπορούμε να είμαστε σίγουροι ότι ερμηνεύουμε σωστά τη συμπεριφορά των χρηστών
  - Για παράδειγμα, σε έναν χρήστη μπορεί να μην αρέσουν όλα τα βιβλία που αγόρασε, ή μπορεί να αγόρασε και βιβλία για άλλους
- Καλύτερα να χρησιμοποιούνται βοηθητικά στις ρητές βαθμολογίες

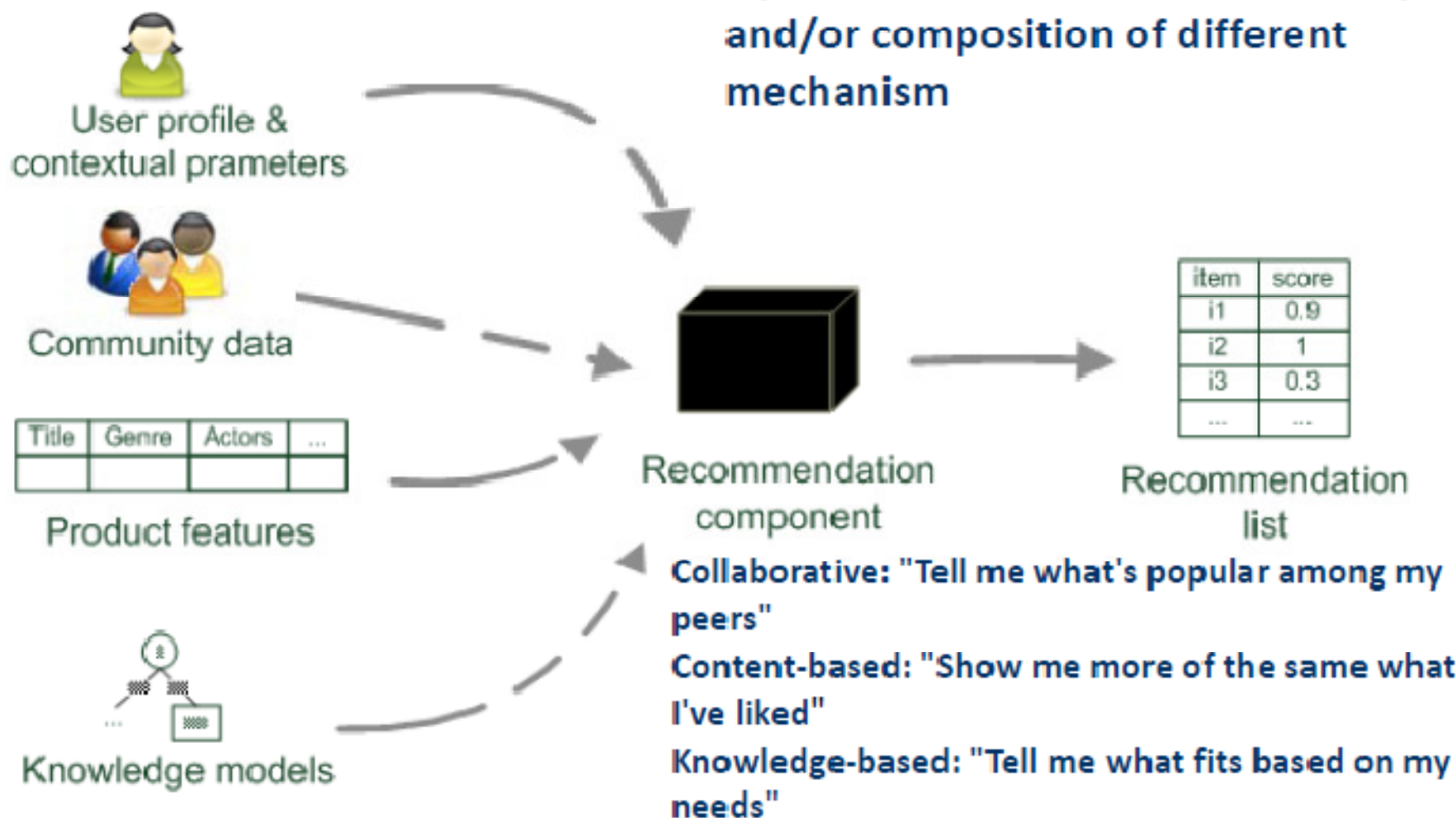


# Υβριδικές Μέθοδοι

- Υλοποίηση δύο ξεχωριστών συστημάτων/μοντέλων συστάσεων και συνδυασμός προβλέψεων
- Πρόσθεση μεθόδων με βάση το περιεχόμενο στο συνεργατικό φιλτράρισμα
  - Προφίλ αντικειμένων για το πρόβλημα νέων αντικειμένων
  - Δημογραφικά για την αντιμετώπιση του προβλήματος νέου χρήστη
  - Κοινωνικά χαρακτηριστικά

# Υβριδικές Μέθοδοι

Hybrid: combinations of various inputs and/or composition of different mechanism



# Παράδειγμα

Recommender 1		
Item1	0.5	1
Item2	0	
Item3	0.3	2
Item4	0.1	3
Item5	0	

Recommender 2		
Item1	0.8	2
Item2	0.9	1
Item3	0.4	3
Item4	0	
Item5	0	

Recommender weighted(0.5:0.5)		
Item1	0.65	1
Item2	0.45	2
Item3	0.35	3
Item4	0.05	4
Item5	0.00	

# Παράδειγμα

Recommender 1		
Item1	0.5	1
Item2	0	
Item3	0.3	2
Item4	0.1	3
Item5	0	

Removing no-go items

Recommender 2		
Item1	0.8	2
Item2	0.9	1
Item3	0.4	3
Item4	0	
Item5	0	

Ordering and refinement

Recommender 3		
Item1	0.80	1
Item2	0.00	
Item3	0.40	2
Item4	0.00	
Item5	0.00	

# Συστήματα Συστάσεων: Βασικές Τεχνικές

	Πλεονεκτήματα	Μειονεκτήματα
Collaborative	Δεν απαιτείται άντληση γνώσης Serendipity αποτελεσμάτων Μαθαίνει τμήματα της αγοράς	Απαιτεί κάποιου είδους βαθμολογημένης ανάδρασης Πρόβλημα cold-start για νέους χρήστες και αντικείμενα
Content-based	Δεν απαιτείται κοινότητα Δυνατή η σύγκριση μεταξύ αντικειμένων	Απαιτεί περιγραφές περιεχομένου Πρόβλημα cold-start για νέους χρήστες Καμία έκπληξη
Knowledge-based	Ντετερμινιστικές συστάσεις Εξασφαλισμένη ποιότητα Δεν υπάρχει πρόβλημα cold-start Μπορεί να μοιάζει με διάλογο πωλήσεων	Απαιτείται διαδικασία μάθησης/μοντελοποίησης γνώσης για την αρχικοποίηση Βασικά στατικό Δεν αντιδρά σε βραχυπρόθεσμες τάσεις

# Εξηγήσεις στα συστήματα συστάσεων

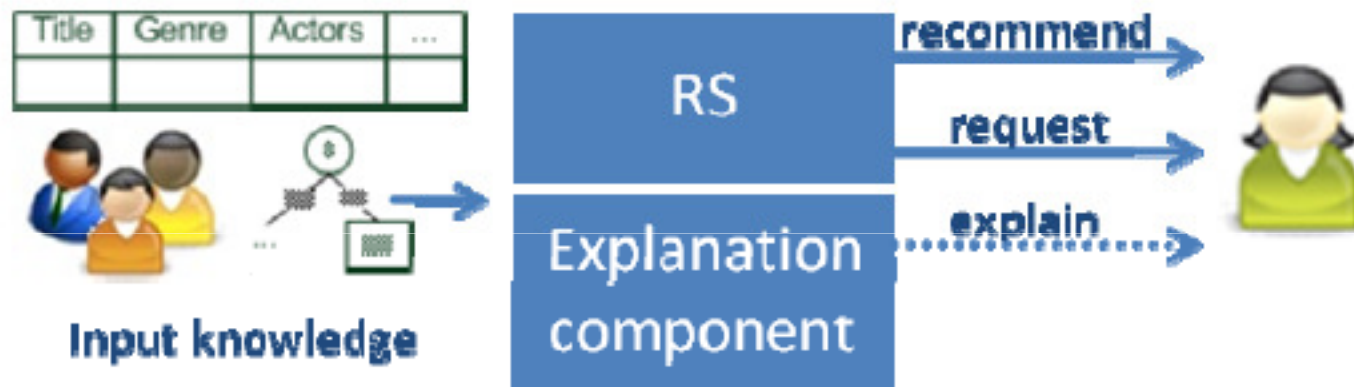
- «Πρέπει να αγοράσεις την κάμερα ΑΒΓΔ γιατί...»
- Γιατί να πρέπει να παρέχουν εξηγήσεις τα συστήματα συστάσεων?
  - Ο πωλητής μπορεί να θέλει να προωθήσει συγκεκριμένα προϊόντα
  - Ο αγοραστής μπορεί να ενδιαφέρεται για την σωστή απόφαση αγοράς

# Τύποι Εξηγήσεων

- Λειτουργικές
  - Ο τύπος αυτοκινήτου Jumbo-Family-Van της μάρκας Rising Sun θα ήταν χρήσιμος για την οικογένεια σου γιατί έχεις τέσσερα παιδιά και το αμάξι έχει 7 θέσεις
- Αιτιατικές
  - Η λάμπα φέγγει γιατί την άναψες
- Πρόθεσης
  - Έπλυνα τα πιάτα επειδή την τελευταία φορά το έκανε ο αδερφός μου
  - Πρέπει να κάνεις τις εργασίες σου γιατί έτσι είπε ο πατέρας σου
- Επιστημονικές εξηγήσεις
  - Εκφράζουν σχέσεις μεταξύ εννοιών που διατυπώνονται σε διάφορα επιστημονικά πεδία και συνήθως αποδεικνύονται θεωρητικά

# Σε ένα σύστημα συστάσεων

- Πρόσθετη πληροφορία για την εξήγηση της εξόδου του συστήματος ακολουθώντας κάποιους σκοπού





# Στόχοι εξηγήσεων (επεξηγήσεων)

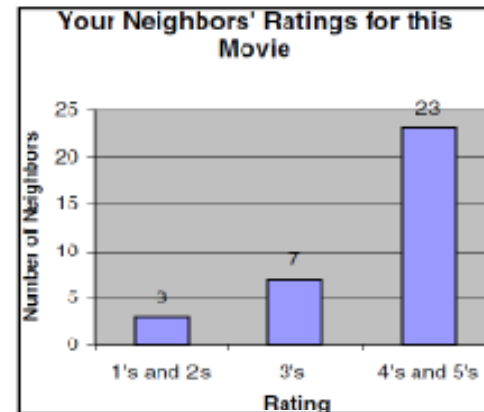
- Διαφάνεια
- Εγκυρότητα
- Αξιοπιστία
- Πειστικότητα
- Αποτελεσματικότητα
- Απόδοση
- Ικανοποίηση
- Σχετικότητα
- Κατανόηση
- Εκπαίδευση

# Παράδειγμα

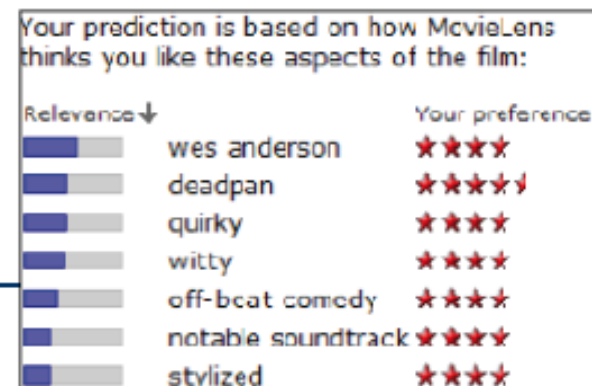
- Ομοιότητα μεταξύ αντικειμένων



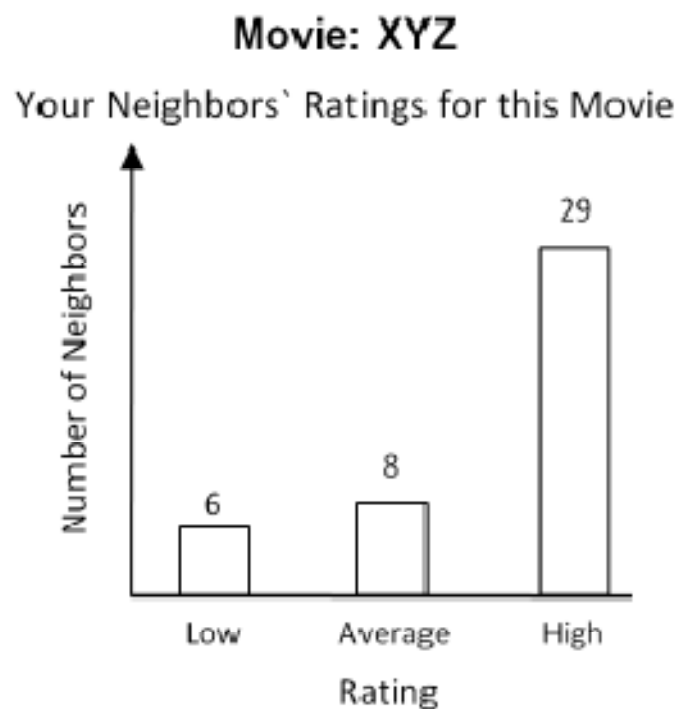
- Ομοιότητα μεταξύ χρηστών



- Ετικέτες



# Παράδειγμα



**Movie: XYZ**  
Personalized Prediction: \*\*\*\*  
Your Neighbors' Ratings for this Movie

Rating	Number of Neighbors
★	2
★★	4
★★★	8
★★★★	20
★★★★★	9

# Αποτίμηση Προβλέψεων

- Σύγκριση προβλέψεων με γνωστές βαθμολογίες
  - Μέσο τετραγωνικό σφάλμα
- Άλλη μέθοδος: μοντέλο 0/1
  - Κάλυψη
    - Αριθμός αντικειμένων/χρηστών για τους οποίους το σύστημα μπορεί να δώσει προβλέψεις
  - Ακρίβεια
    - Ακρίβεια στις προβλέψεις
  - Χαρακτηριστικό λειτουργίας αποδέκτη
    - Συμβιβασμός μεταξύ false positives και false negatives

# Προβλήματα Μέτρων Αποτίμησης

- Περιοριστική σκοπιά με έμφαση μόνο στην ακρίβεια πολλές φορές χάνει άλλες σημαντικές παραμέτρους
  - Διαφοροποίηση πρόβλεψης
  - Συμφραζόμενα πρόβλεψης
  - Διάταξη προβλέψεων
- Στην πράξη, μας νοιάζει να προβλέπουμε μόνο υψηλές βαθμολογίες

# Συμβουλή: Πρόσθεσε δεδομένα

- Εκμεταλλεύσου όλα τα δεδομένα
  - Μην προσπαθείς να μειώσεις το μέγεθος των δεδομένων για να μπορέσεις να τρέξεις πολύπλοκους αλγορίθμους
  - Απλές μέθοδοι σε πολλά δεδομένα έχουν τα καλύτερα αποτελέσματα
- Πρόσθεσε κι άλλα δεδομένα
  - Εκμεταλλεύσου εξωτερικές πηγές για να εμπλουτίσεις τα δεδομένα
- Τα περισσότερα δεδομένα κερδίζουν τους καλύτερους αλγόριθμους

# Εύρεση παρόμοιων διανυσμάτων

- Κοινό πρόβλημα που εμφανίζεται σε διάφορα θέματα
- Δεδομένου ενός μεγάλου αριθμού  $N$  από διανύσματα σε κάποιον χώρο υψηλών διαστάσεων, βρες ζεύγη διανυσμάτων που έχουν υψηλή ομοιότητα (cosine similarity)

# Αναφορές

- <http://www.recommenderbook.net/teaching-material/slides>
- <http://infolab.stanford.edu/~ullman/mmds.html>
- D. Jannach, G. Friedrich. Tutorial: Recommender Systems, International Joint Conference on Artificial Intelligence