

# 9. Előadás

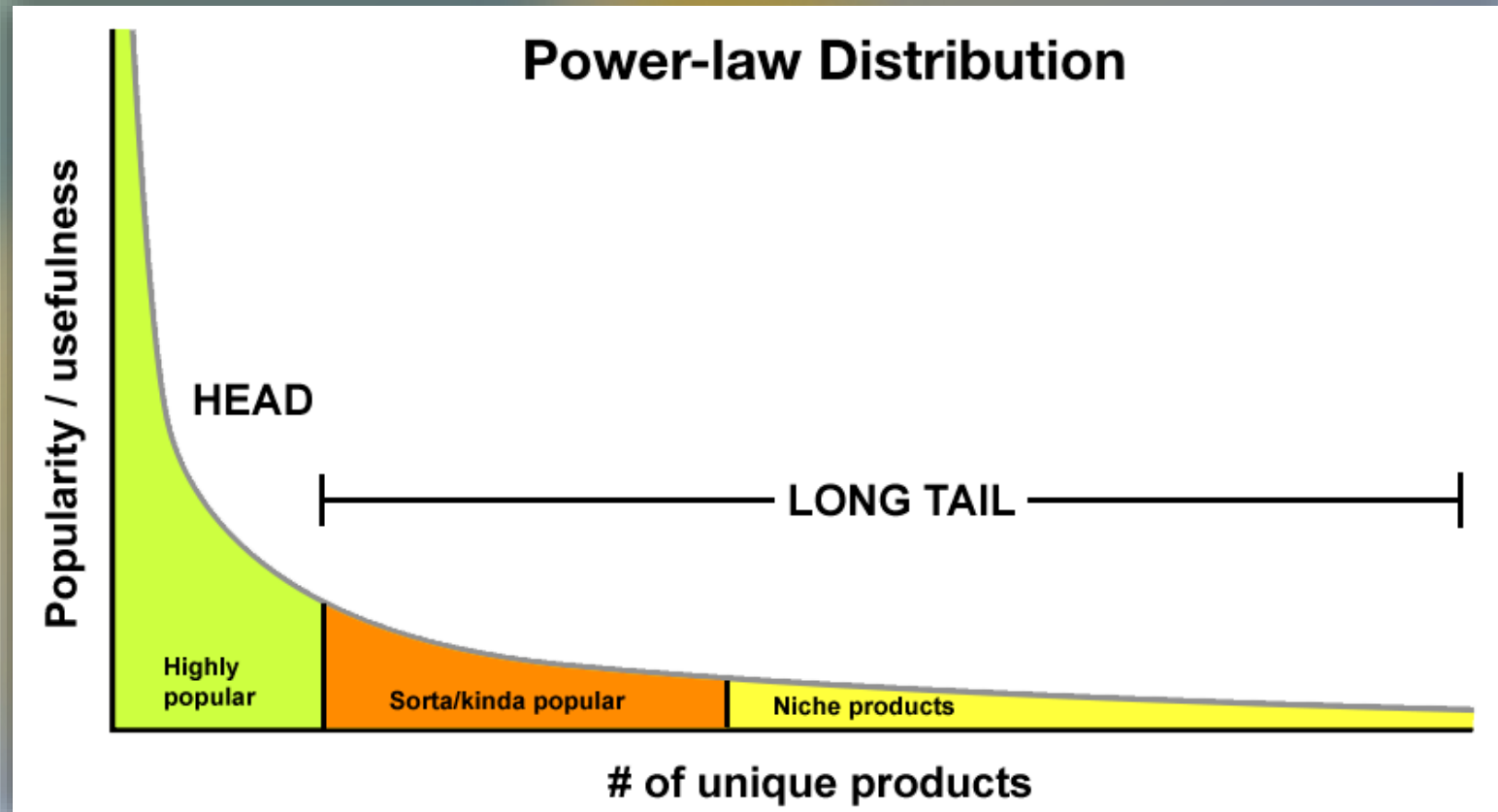
## Ajánló rendszerek

# A hosszú farok eloszlás

🐍 Az internet előtt az volt a jellemző, hogy pár termék generálta a forgalom nagy részét, és mivel a hely is limitált az üzlethelységben, a kevesek által keresett termékek nem kaptak helyet a polcon.

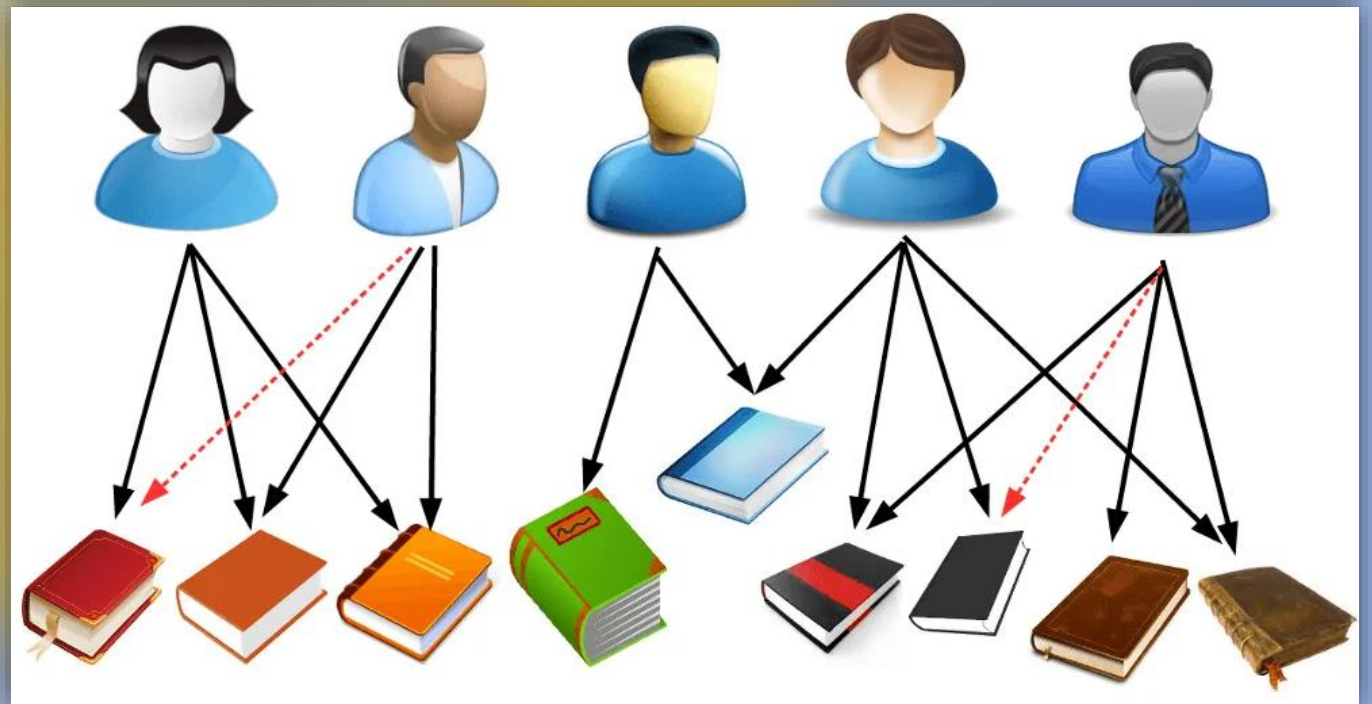
🐍 Az internetes kereskedelem elterjedése helyet adott a vásárlók szűk csoportja számára vonzó, *niche* termékeknek, amik specifikus felhasználásukkal vonzzák be a vásárlókat.

🐍 Ez a vásárlóknak szélesebb termékspektrumot, az eladóknak pedig nagyobb vásárlóközöniséget jelentett.






# Mik azok az ajánló rendszerek?

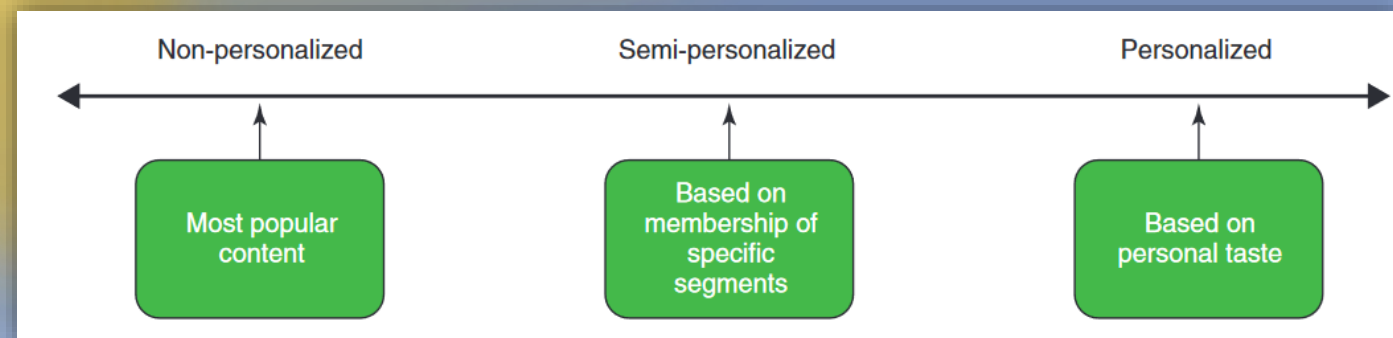
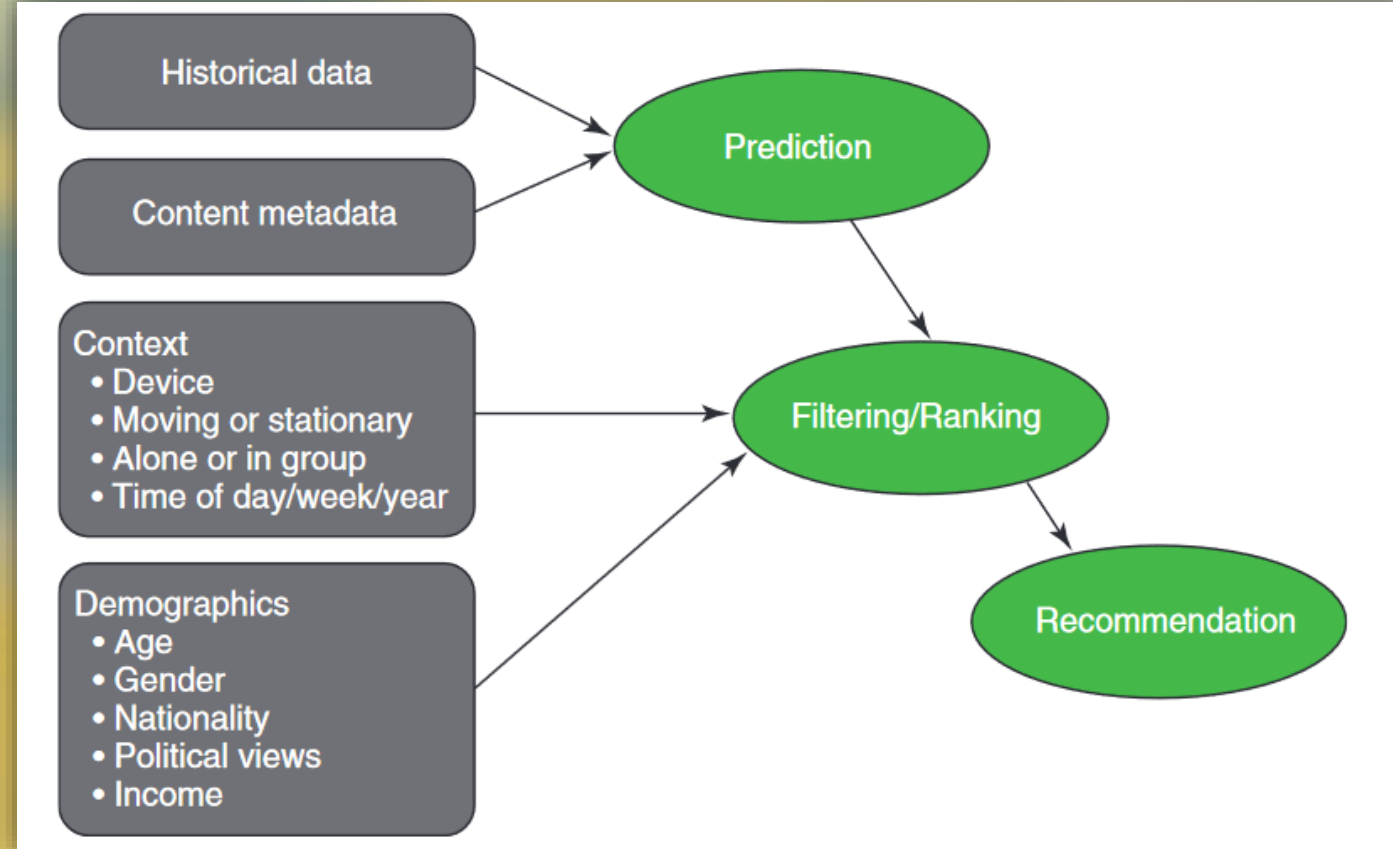
- 🐍 Ez elég magától értetődő: ahogy a név is sugallja, olyan technikák vagy rendszerek, amelyek valamilyen terméket, szolgáltatást vagy entitást javasolnak a számukra rendelkezésre álló információ alapján.
- 🐍 Más szóval: objektumok két csoportja közötti kapcsolatok felderítése.
- 🐍 Pl. Filmek, termékek, könyvek ajánlása vásárlóknak





# A probléma felírása

-  Az alap elgondolás, hogy ajánlásokat szeretnénk tenni rendelkezésre álló adatok alapján.
-  Különböző rendszertípusoknak különböző igénye van az adatok forrásával, milyenségével, folyamatos rendelkezésre állásával szemben.
-  A rendszereket sem mindegy, milyen predikcióra optimalizáljuk. Ez a skála a népszerű, sokaknak megfelelő tartalomtól a személyre szabott, specifikus ajánlásokig terjedhet.



# A predikciós probléma

- A feladatnak ebben a verziójában rendelkezünk egy  $m$  felhasználóból és  $n$  termékből álló mátrixszal. Az  $i$ -edik sor  $j$ -edik oszlopa jelenti azt, hogy adott felhasználó hogyan értékelte a vonatkozó terméket.
- Vegyünk egy komplexebb példát: ha a Netflix adatbázisában 20000 film és 5000 felhasználó található. Ez egy  $20000 * 5000$ -es mátrixot ad eredményül.
- Viszont az egyes felhasználók a filmek töredékét sem látták: ez egy ritka-mátrix (sok üres értékkel).
- A feladat tehát a mátrix hiányzó értékeinek a megbecslése a filmekről és felhasználókról rendelkezésre álló információ alapján.
- Értékelések szempontjából megkülönböztetünk explicit és implicit értékeléseket.

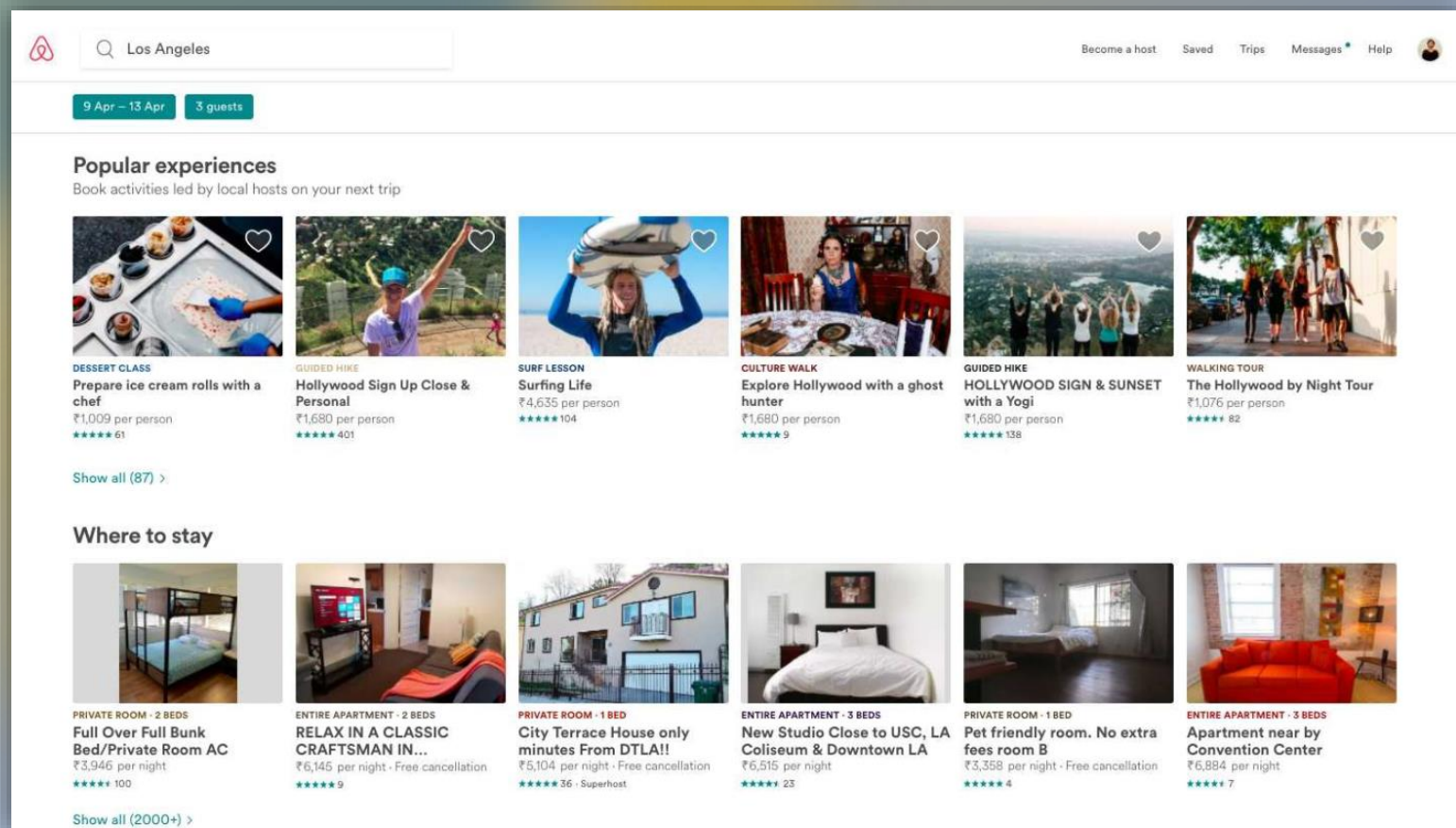
	$i_1$	$i_2$	$i_3$	$i_4$	$i_5$	$i_6$
U1	4	?	3	?	5	?
U2	?	2	?	?	4	1
U3	?	?	1	?	2	5
U4	?	?	3	?	?	1
U5	1	4	?	?	2	5
U6	5	?	2	1	?	4
U7	?	2	3	?	4	5

# A rangsorolási probléma

🐍 A rangsorolás a predikciós problémának az intuitívabb megfogalmazása.

🐍 Ha adott  $n$  elem halmaza, a rangsorolási probléma megpróbálja megkülönböztetni a legjobban javasolható  $k$  elemet, amit javasolhat egy felhasználónak a rendelkezésre álló információ alapján.

🐍 Nem nehéz belátni, hogy a predikciós probléma gyakran a rangsorolási problémához vezet vissza.





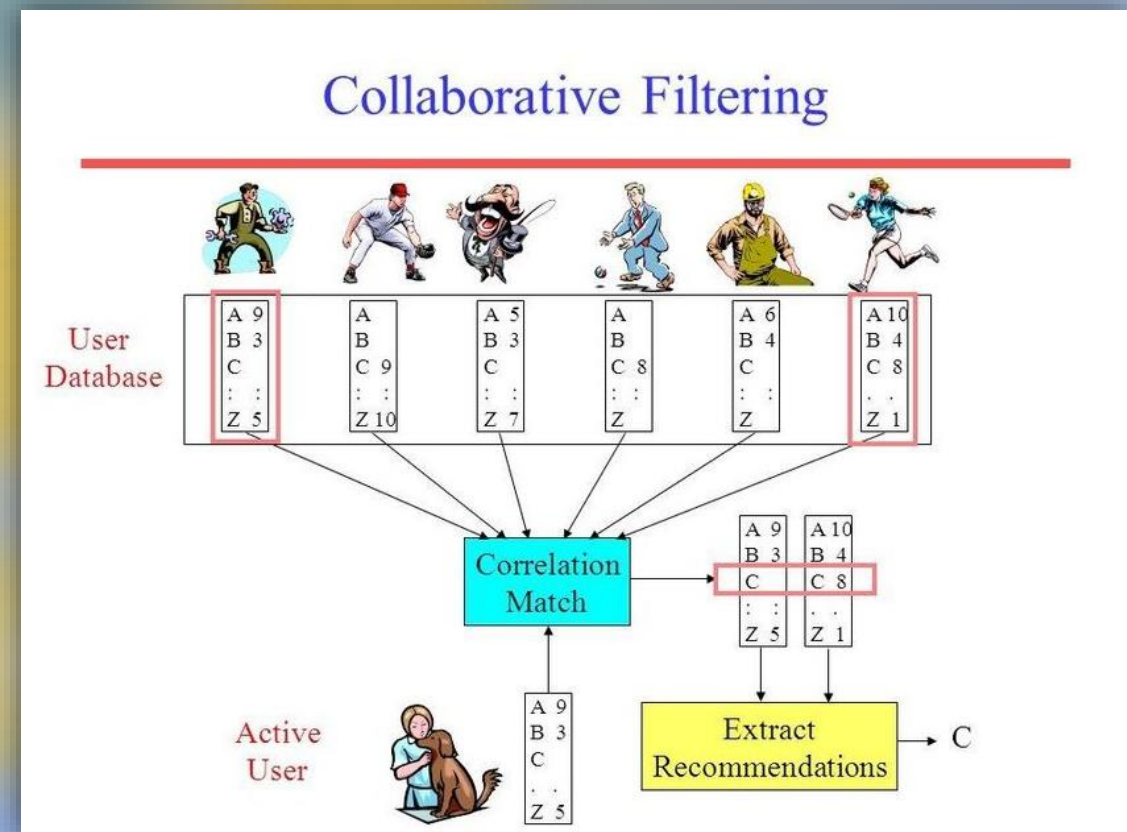
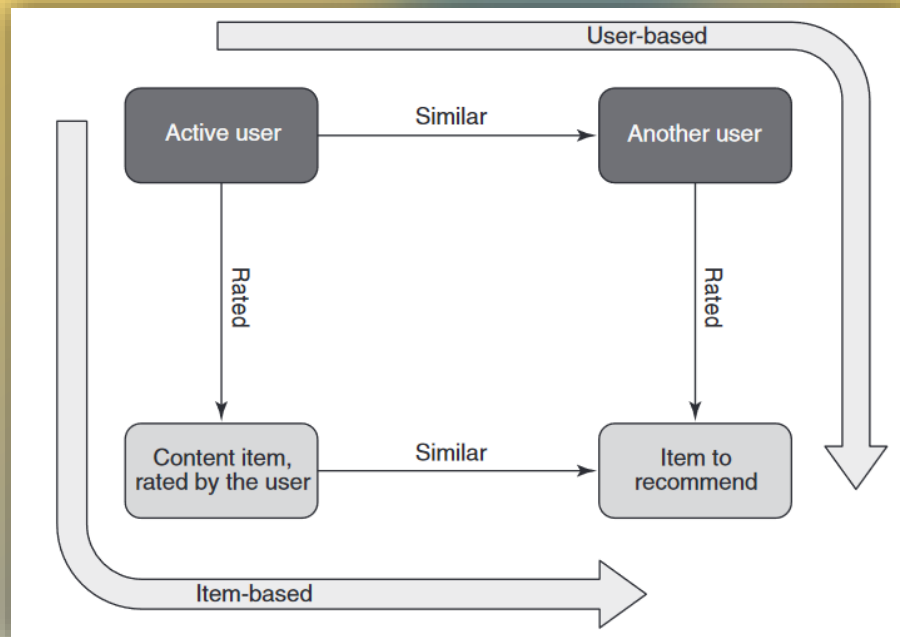
# AR típusok 1: Kollaboratív szűrők

🐍 Azokat a rendszereket, amelyek a közösség által adott értékelésekből, használati metrikából állítanak össze javaslatokat, kollaboratív szűrőknek nevezzük.

🐍 Két típusát lehet megkülönböztetni:

🐍 Felhasználó-alapú: azok a felhasználók, akik hasonló termékeket vásároltak, hasonló termékeket fognak vásárolni a jövőben is.

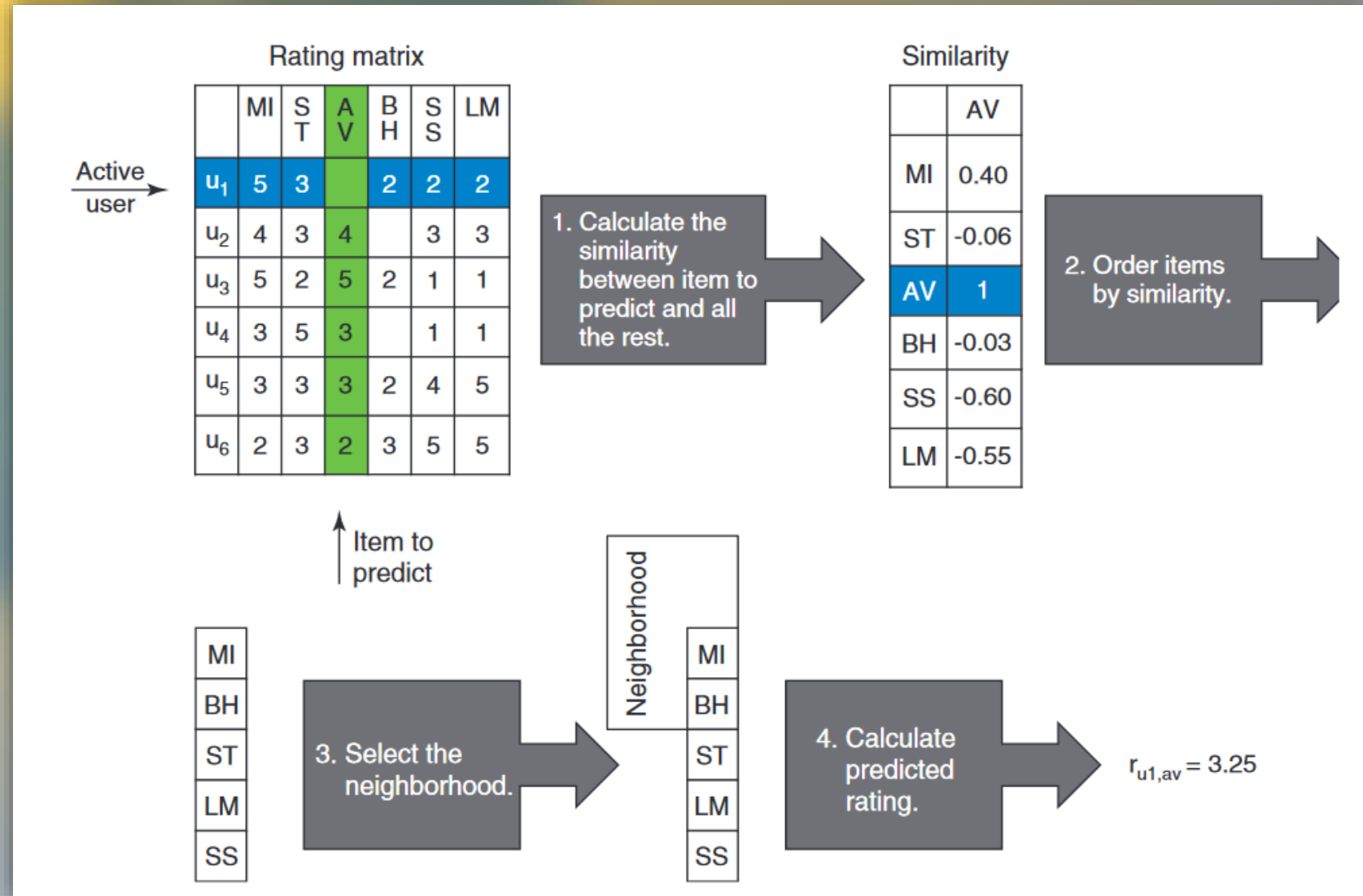
🐍 Termék-alapú: ha a felhasználók a termékeket hasonlóan értékelték, hasonlóan kell lenniük.



# Termék-alapú rendszerek eljárása

Ahhoz, hogy megtaláljuk valamely termékhez a leginkább hasonlót, először kiszámoljuk, mennyiben hasonlít a többi termékhez, berendezzük növekvő sorrendbe, majd kiválasztjuk a hozzá legközelebb taláhatót.

Előnye, hogy a döntéseit könnyen meg lehet indokolni: „ha tetszett ez, tetszeni fog az”.





# De hogyan számíttódik ki a hasonlóság?

- 🐍 Feltételezzük, hogy a hasonlóság statikus, tehát előzetesen, offline ki lehet őket számolni.
- 🐍 Először az Amazon alkalmazta és publikálta ennek a módszertanát Greg Linden publikálta. A termék-termék kollaboratív szűrésre ezt az algoritmust hozta:

```
For each item in product catalog,  $I_1$ 
  For each customer C who purchased  $I_1$ 
    For each item  $I_2$  purchased by customer C
      Record that a customer purchased  $I_1$  and  $I_2$ 
    For each item  $I_2$ 
      Compute the similarity between  $I_1$  and  $I_2$ 
```




- 🐍 Ennek az eredménye egy olyan adatszerkezet, amelyben termékenként lehet keresni a hasonló egyedek között.

# Példa hasonlósági tábla összeállítására

🐍 Vegyünk egy hasonlósági mátrixot, és csináljuk meg hasonlósági táblát.

🐍 Mindenki látta a MiB-et, ezért végig kell menni az összes felhasználón. Az első Sara, aki látta a Star-Trek-et, Braveheart-ot, Sense and Sensibility-t, és a Les Miserables-t.

🐍 Ezeket a filmeket hozzáadjuk a MiB-bel együtt értékeltekhez, és végig megyünk az összes felhasználón. Az eredmény:

						
	Comedy	Action	Comedy	Action	Drama	Drama
Sara	5	3		2	2	2
Jesper	4	3	4		3	3
Therese	5	2	5	2	1	1
Helle	3	5	3		1	1
Pietro	3	3	3	2	4	5
Ekaterina	2	3	2	3	5	5

- MIB: [ST, B, SS, LM, AV]
- ST: [MIB, B, SS, LM, AV]
- B: [MIB, ST, SS, LM, AV]
- SS: [MIB, ST, B, LM, AV]
- LM: [MIB, ST, SS, B, AV]
- AV: [MIB, ST, B, SS, LM]

# Két egyed hasonlóságának kiszámítása

🐍 Ehhez a módosított koszinusz hasonlósági függvényt kell használnunk. Azért a módosítottat, mert normalizálja az eredményeket, hogy  $-1$  és  $1$  közé essenek. Egyébként nagyon hasonlóan működik a Pearson korrelációs együtthatóhoz.

🐍 A számításokhoz szükséges:

🐍  $r_{i,u}$ : az  $i$ -edik film értékelése  $u$  felhasználótól

🐍  $\bar{r}_u$ :  $u$  felhasználó átlagos értékelése

🐍  $nr_{i,u} = r_{i,u} - \bar{r}_u$ : normalizáláshoz szükséges tényező

$$\text{Sim}(\text{"MIB"}, \text{"ST"}) = \frac{\sum_u nr_{MIB,u} nr_{ST,u}}{\sqrt{\sum_u nr_{MIB,u}^2} \sqrt{\sum_u nr_{ST,u}^2}}$$

Jesper értékelései

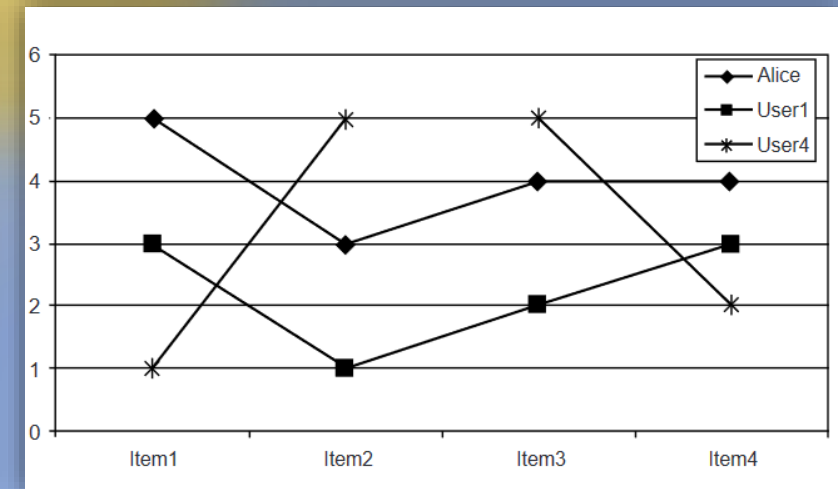
$$\frac{(2.2 * 0.2) + (0.6 * -0.4) + (2.33 * -0.67) + (0.4 * 2.4) + (-0.33 * -0.3) + (-1.33 * -0.3)}{\sqrt{2.2^2 + 0.6^2 + 2.33^2 + 0.4^2 + -0.33^2 + -1.33^2} * \sqrt{0.2^2 + -0.4^2 + -0.67^2 + 2.4^2 + -0.33^2 + -0.33^2}}$$

Jesper értékelése a MIB-ről

Hasonlóság

Jesper értékelése az ST-ről


$$= \frac{0.1467}{\sqrt{3.559} * \sqrt{2.574}} = 0.016$$





# Korrelációs mátrix létrehozása

🐍 Az előbb felvázolt módszerrel felírható minden filmnek a normalizált értékelése a felhasználó átlagos értékelései szerint. A pozitív értékek a felhasználónál jobbnak számítanak, mint az átlagos, a negatívak pedig rosszabbat.

						
	Comedy	Action	Comedy	Action	Drama	Drama
Sara	2.20	0.20		-0.80	-0.80	-0.80
Jesper	0.60	-0.40	0.60		-0.40	-0.40
Therese	2.33	-0.67	2.33	-0.67	-1.67	-1.67
Helle	0.40	2.40	0.40		-1.60	-1.60
Pietro	-0.33	-0.33	-0.33	-1.33	0.67	1.67
Ekaterina	-1.33	-0.33	-1.33	-0.33	1.67	1.67

🐍 Ha a felhasználókra számított normalizált értékelésekre korrelációs együttthatót számítunk, megkapjuk a film-film hasonlósági mátrixot. Az 1-es jelenti a tökéletes hasonlóságot, a  $-1$  a tökéletes különbözőséget.

						
MIB	1	0.63	1	-0.21	-0.88	-0.83
ST	0.63	1	0.35	-0.47	-0.64	-0.62
AV	1	0.35	1	0.01	-0.89	-0.83
B	-0.21	-0.47	0.01	1	-0.23	-0.32
SS	-0.88	-0.64	-0.89	-0.23	1	0.96
LM	-0.83	-0.62	-0.83	-0.32	0.96	1

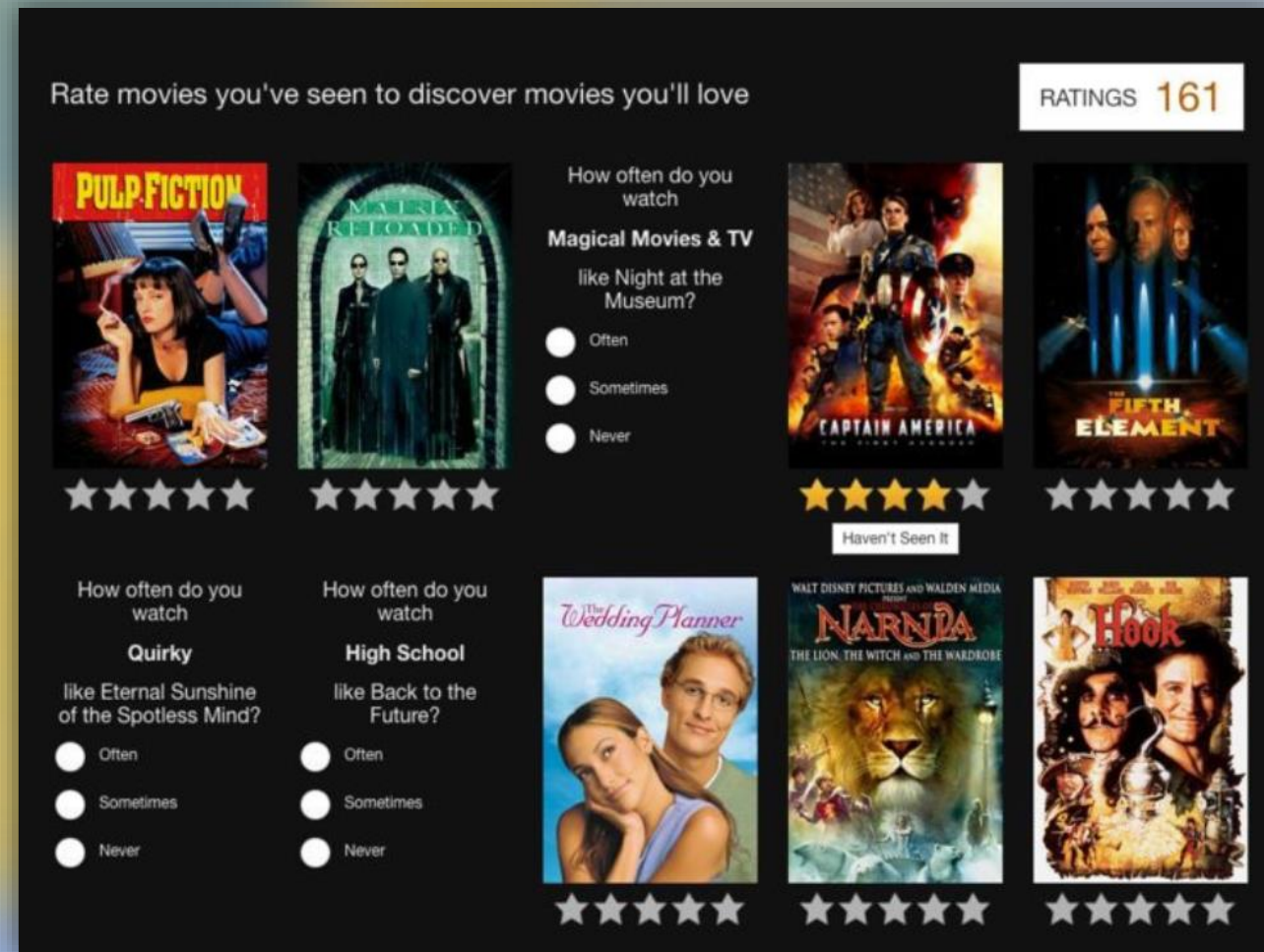
# AR típusok 2: Tartalom alapú rendszerek

🐍 A kollaboratív szűrőkkel ellentétben a tartalom alapú szűrőknek nincs szüksége információra múltbeli vásárlásokról és információról. Ehelyett a javaslatokat a felhasználók profiljai és a termékek metaadatai alapján állítják össze.

🐍 Példa erre a Netflix ajánló rendszere: mikor először bejelentkezünk, megkér rá, hogy értékeljünk olyan filmeket, amiket már korábban láttunk.

🐍 Ezeknek a filmeknek a metaadatai alapján fogja tudni összeállítani a javaslatokat.

🐍 **Probléma:** a tartalom alapú rendszerek nem használják ki a közösség adta lehetőségeket, ezért a predikciók gyakran nyilvánvalóak.



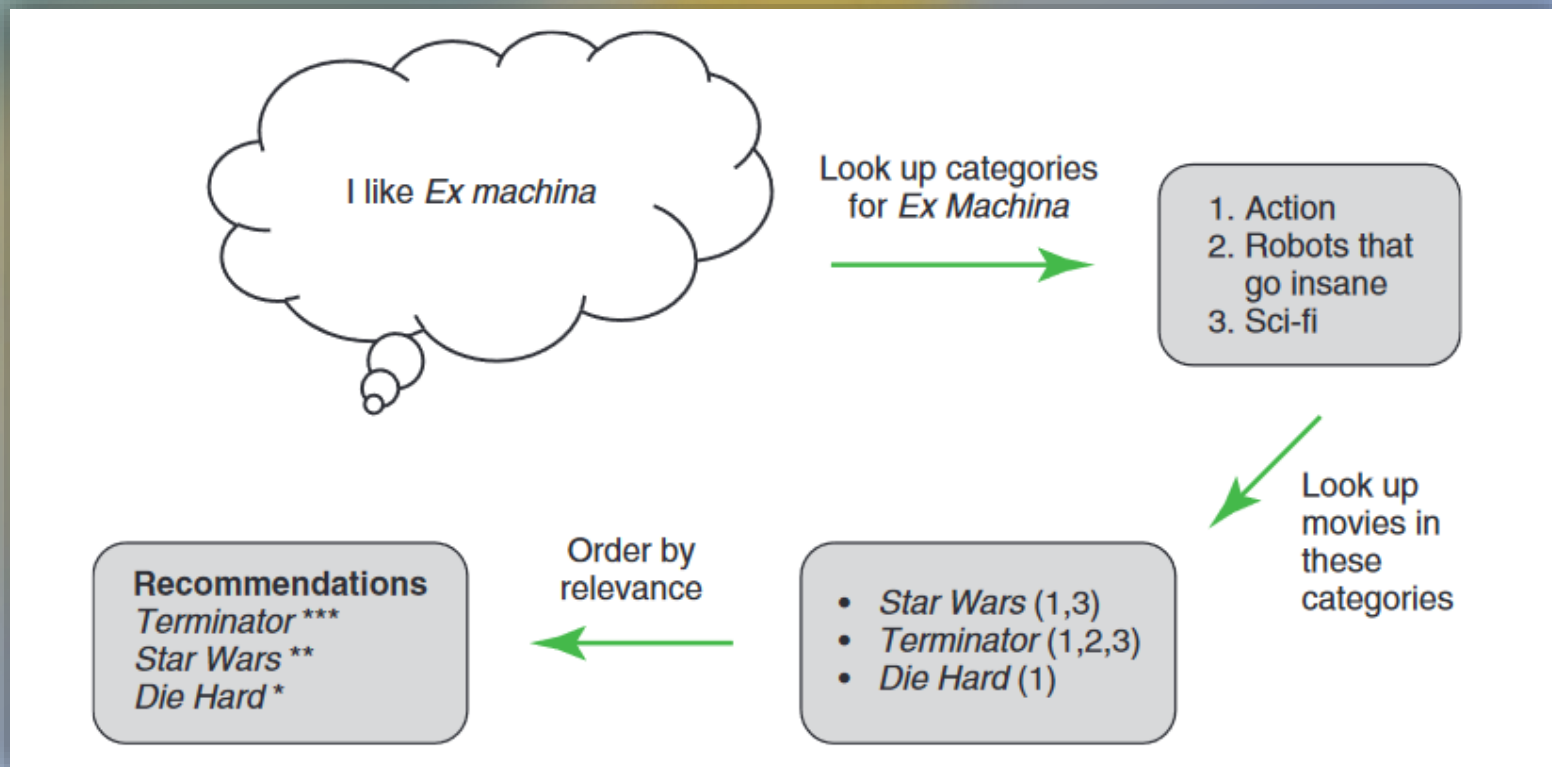


# Tartalom alapú rendszerek csővezetéke

🐍 Az alap elgondolás az, hogy ahogy 33a felhasználó érintkezik a rendszerrel, az megpróbál bizonyos elemekhez hasonlóakat mutatni neki anélkül, hogy a felhasználó bármikor véleményt adott volna a tartalommal kapcsolatban.




🐍 A releváns tartalom megtalálására egy gyakori megoldás a **tag**: a Web2.0 idején megjelent weboldal tartalmára utaló kulcsszavak. Illetve a **fact**, ami tényszerűen írja le az elemet pl.: év.

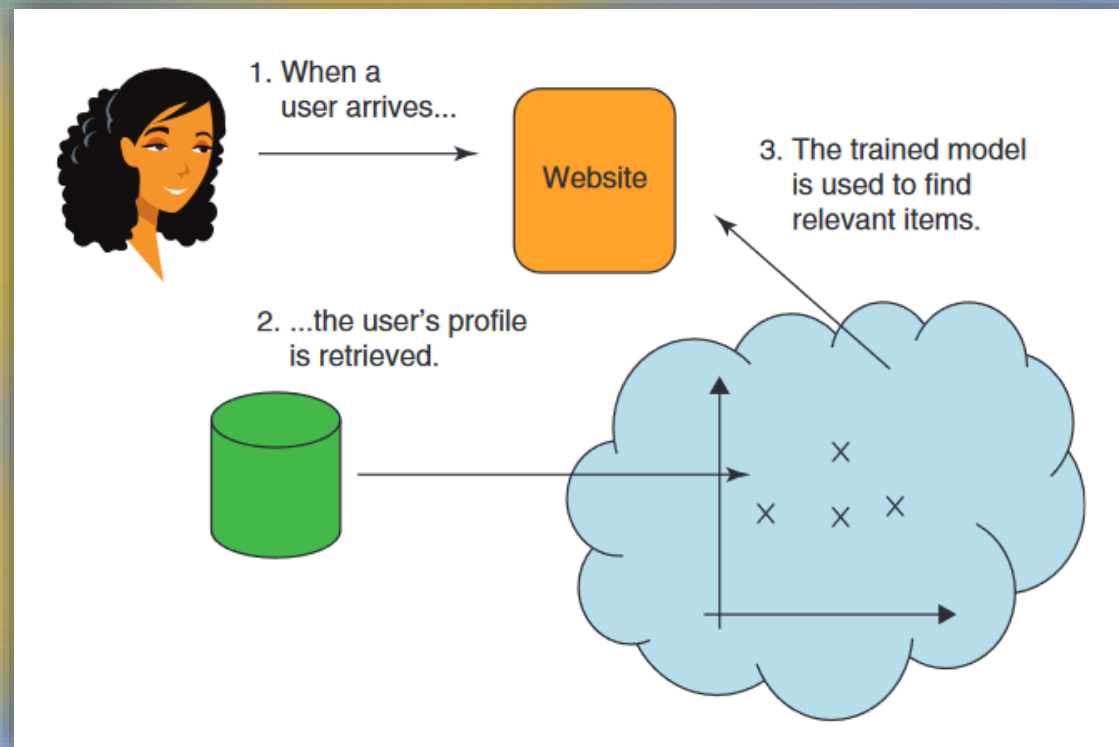
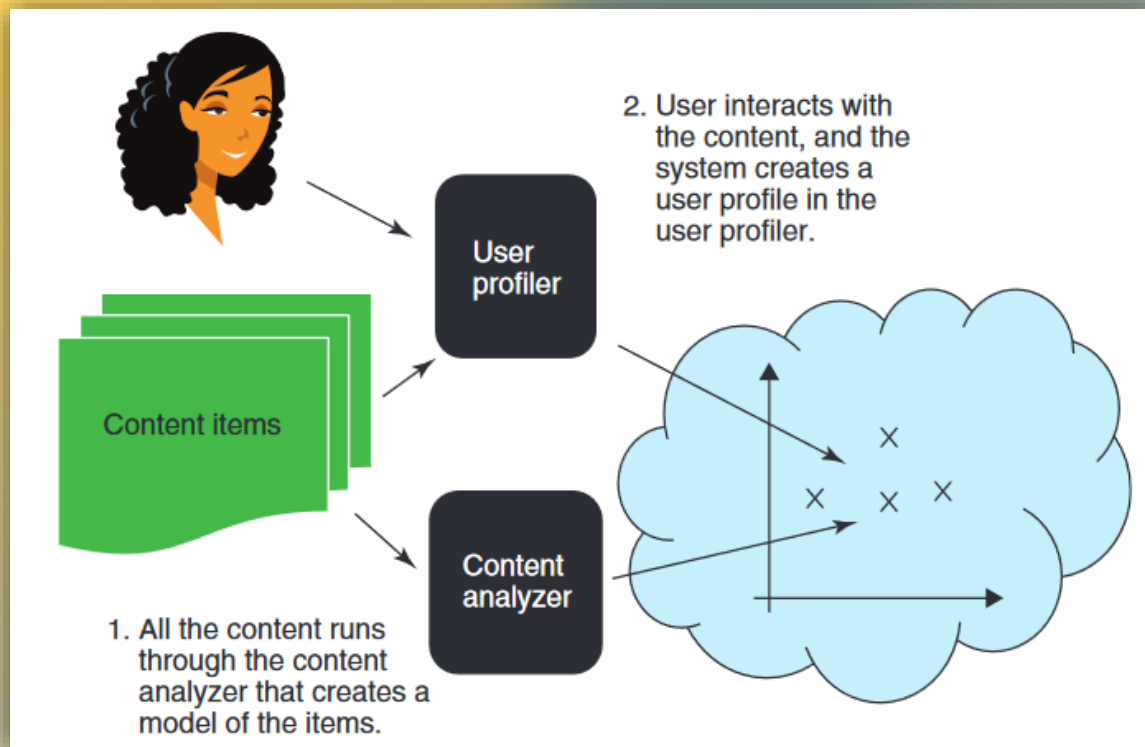
🐍 Egy másik lehet a **TF-IDF vektorizáció**: egy numerikus mutató, ami arra utal, hogy egy szó mennyire fontos adott dokumentumban egy szövegtörzsön vagy gyűjteményen belül.





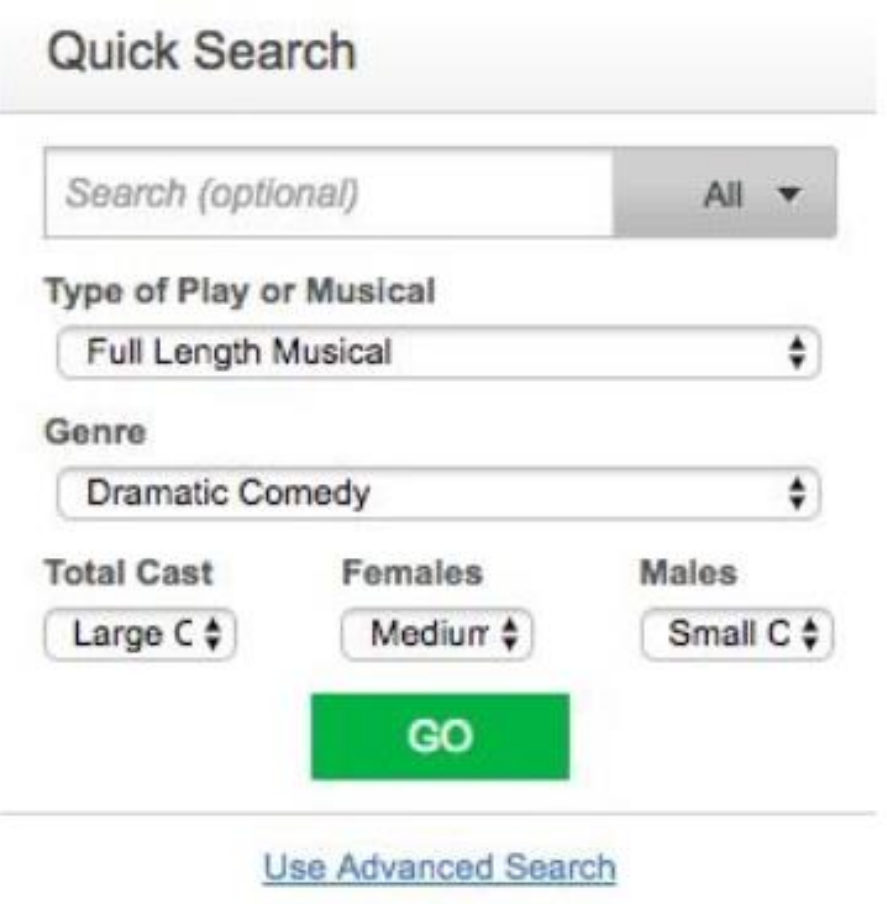
# Tartalom alapú rendszerek részei

-  Tartalomelemző: a tartalom alapján készít modelleket az elérhető elemekről.
-  Felhasználó profilozó: felhasználói profilokat készít. Ez gyakran egy egyszerű lista azokról az elemekről, amiket a felhasználó fogyasztott.
-  Elem visszakereső: ennek a komponensnek fel kell keresnie a releváns tartalmakat azáltal, hogy a felhasználói profilokat összehasonlítja az elemek profiljaival.



# AR típusok 3: Tudásalapú rendszerek

- 🐍 Az ajánló rendszereknek ez a fajtája olyan tételek esetén használatos, amiket az emberek nagyon ritkán vásárolnak.
- 🐍 Ebből az okból kifolyólag lehetetlen, hogy az előbb felsorolt típusok közül valamelyikbe bekerüljenek ezek az egyedek.  
Például: ingatlan vásárlások.
- 🐍 Ebben a rendszerben való kereséshez a felhasználó megadja a termék elvárt paramétereit, mint pl. szobák száma, alapterület stb..., és a rendszer megadja azokat az egyedeket, amelyekre jellemzőek az elvárt paraméterek.
- 🐍 Hátránya, hogy a keresés nem fog váratlan, újszerű eredményeket adni, kiszámítható a működése.



The image shows a 'Quick Search' form with the following fields and options:

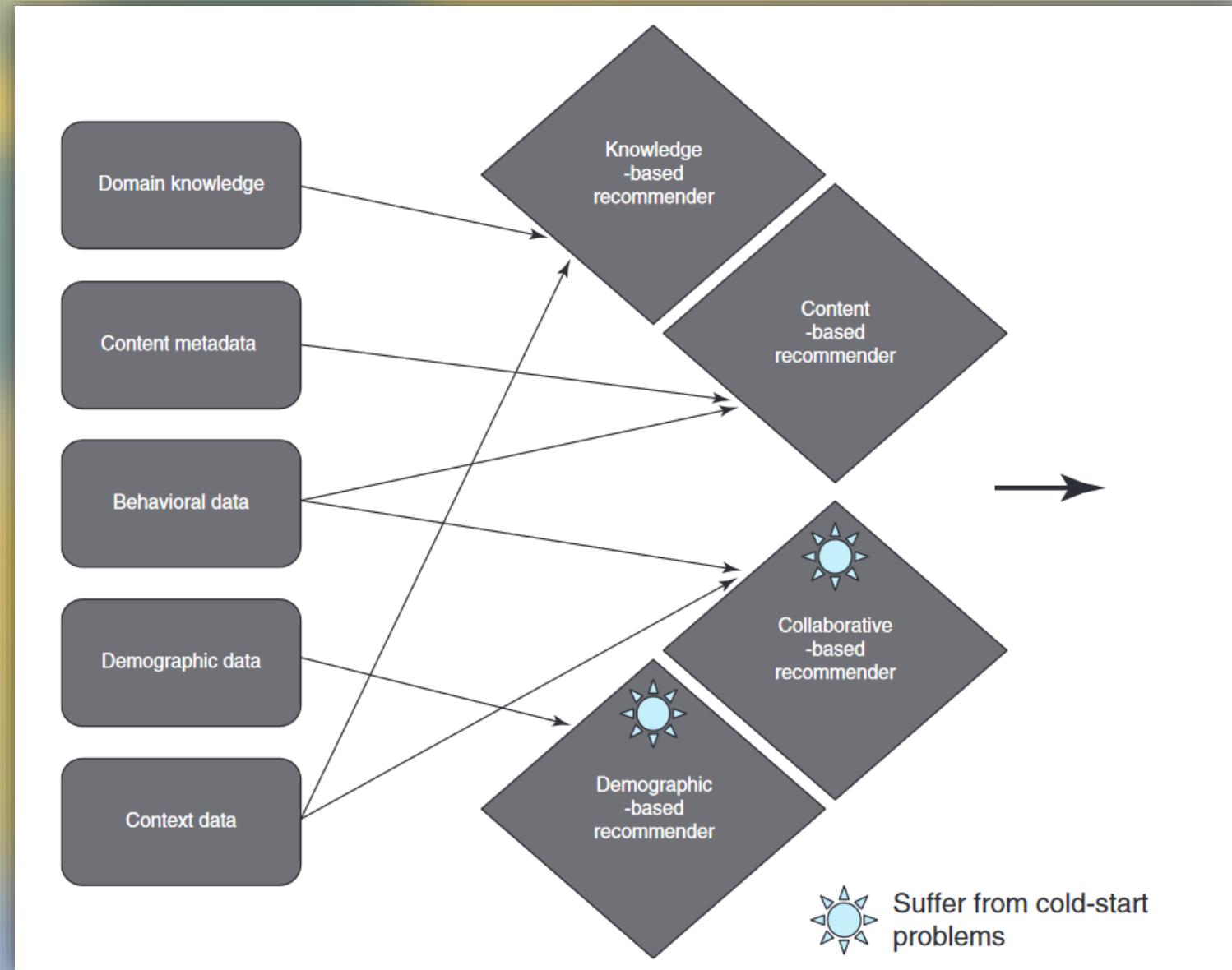
- Search (optional)**: A text input field.
- All**: A dropdown menu.
- Type of Play or Musical**: A dropdown menu with 'Full Length Musical' selected.
- Genre**: A dropdown menu with 'Dramatic Comedy' selected.
- Total Cast**: A dropdown menu with 'Large C' selected.
- Females**: A dropdown menu with 'Medium' selected.
- Males**: A dropdown menu with 'Small C' selected.
- GO**: A green button to submit the search.
- Use Advanced Search**: A link at the bottom.

# AR típusok 4: hibrid megközelítések

🐍 Ahogy azt a név is sugallja, a hibrid rendszerek ötvözik az eddig felsorolt típusokat. Ahogy már láttuk az előző példákból, minden modell fajtának megvannak az előnyei és a hátrányai. A hibrid rendszerek megpróbálják ezek előnyeit megtartani, és a hátrányait kihagyni a rendszerből.

🐍 Három típusa:

- 🐍 Monolitikus
- 🐍 Együttes
- 🐍 Kevert

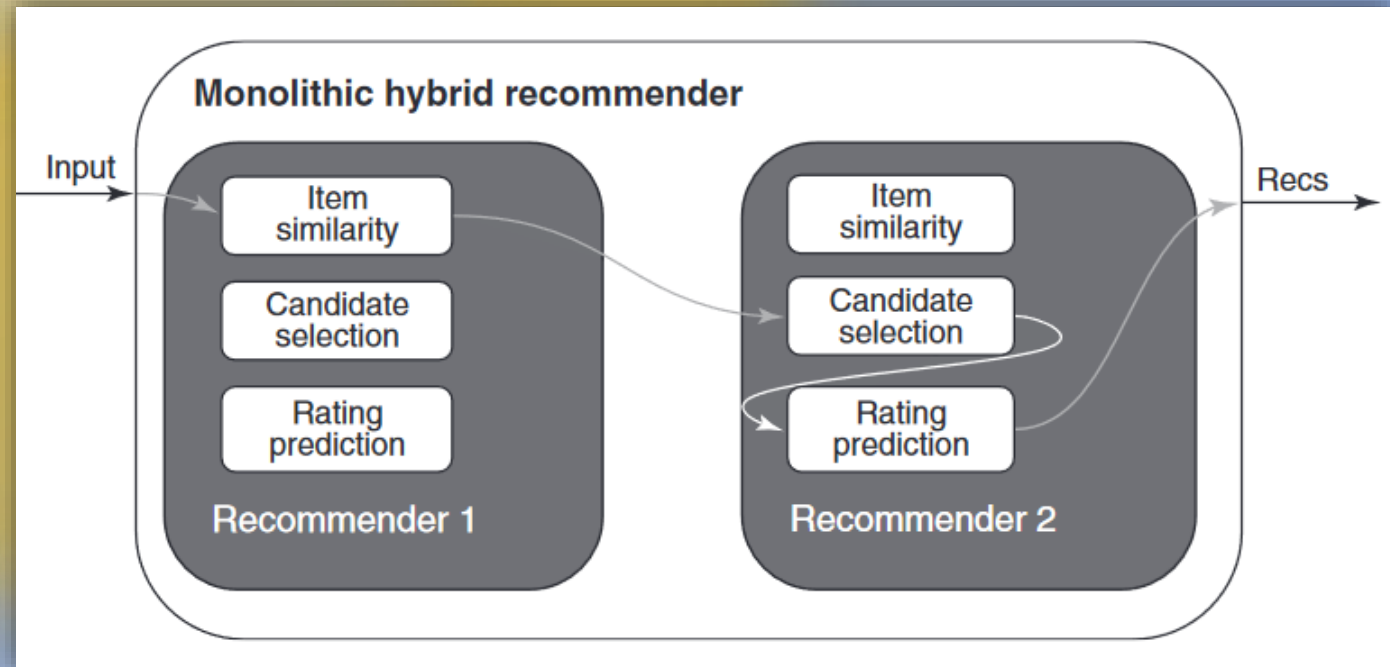




# Monolitikus hibrid ajánló rendszerek

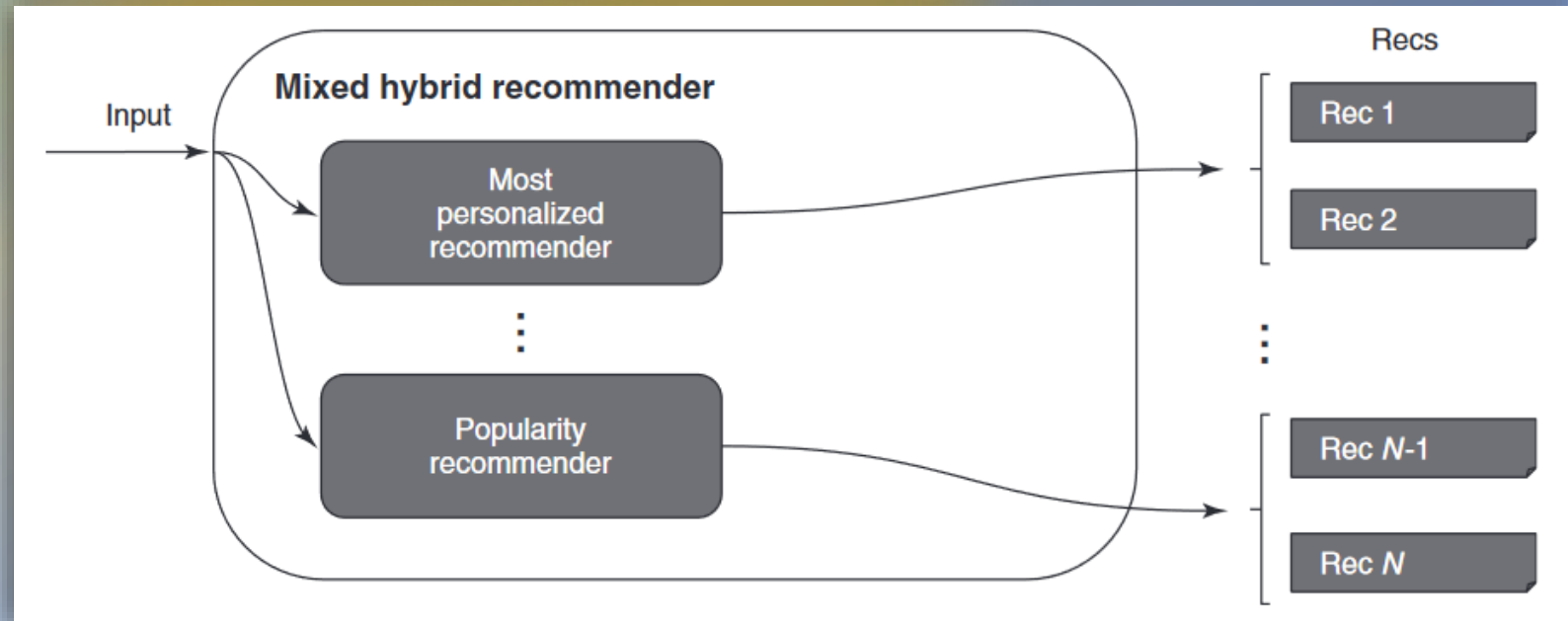
- Ezek az ajánló rendszerek Frankensteinjei. Egy ajánló rendszer általánosságban több komponenst tartalmaz különböző feladatok elvégzéséhez, mint a hasonlóság-számító, elemkiválasztó.
- Egy monolitikus ajánló különböző rendszerek komponenseit használja fel egy csővezetékbe építve. Gyakori, hogy extra lépéseket is hozzáad a folyamathoz annak érdekében, hogy javítsa a predikciókat.
- Például, tartalomalapú adatok keverése viselkedéshez köthető javaslatokkal:

	Sci-fi 1	Sci-fi 2	Sci-fi 3	Sci-fi 4
User 1	4	4		
User 2	5	4		
User 3			2	4
Sci-fi lover	5	5	5	5



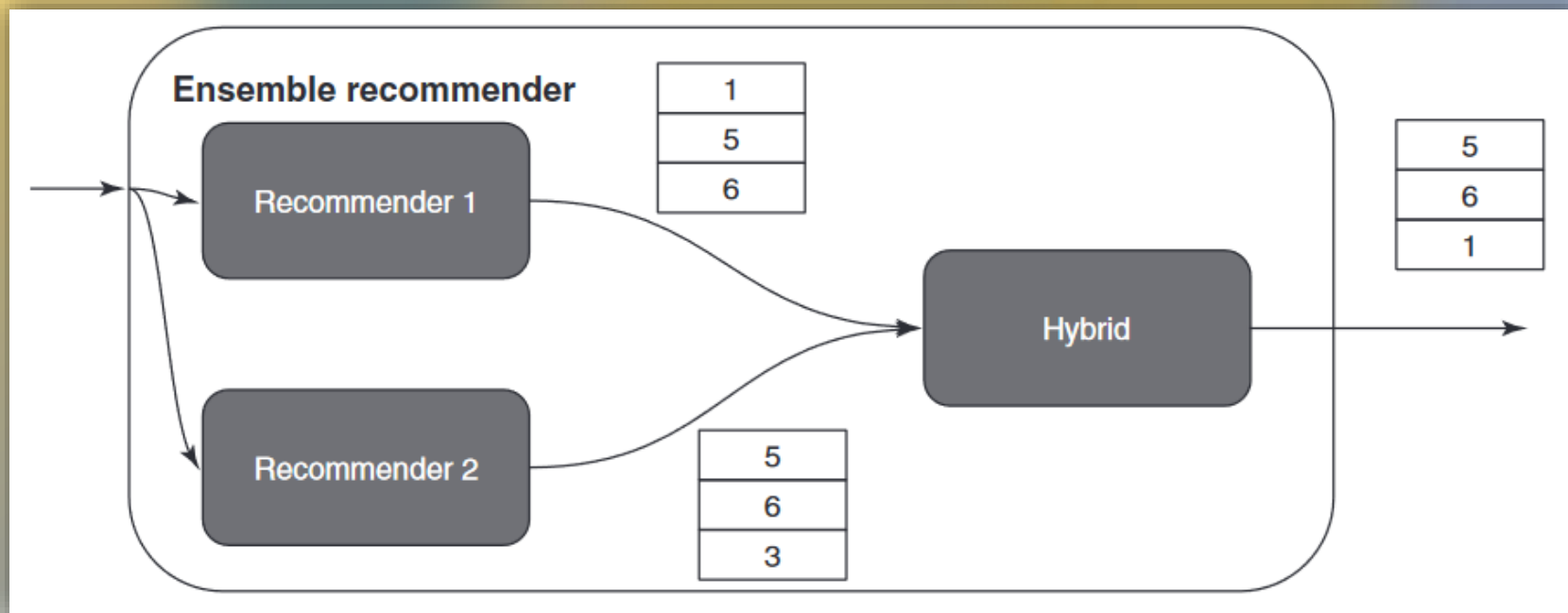
# Kevert hibrid ajánlók

- 🐍 A kevert rendszerek valójában nem csinálnak túl sok keverést. Ebben az az eljárás, hogy a több, szigetszerűen működő ajánló rendszer predikcióinak unióját téríti vissza.
- 🐍 Az ajánló rendszert fel lehet fogni úgy, mint a személyreszabottság mértékét. Az elsőt a lehetőleg személy specifikusabb predikciók jellemzik, míg az utolsót a legnépszerűbbek. Gyakran a személyreszabott rendszerek 1-2 predikciót adnak, míg a népszerűek sokkal többet.
- 🐍 Ha minden rendszer egy pontszámot térít vissza, ezt normalizálva rendezett listát lehet belőlük készíteni, amely megadja a fontosság sorrendjét.



# Együttes hibrid ajánlók

- Ahogy az együttes tanulás esetében, úgy az együttes ajánlók is több különálló modell predikcióit kombinálják össze egygé. A válaszok aggregálása történhet különböző módszerekkel, mint a szavazás, súlyozás, kapcsolás.
- Az ábrán erre látunk egy példát: a Recommender 1 predikciói [1,5,6], a Recommender 2 predikciói [5,6,3]. Ekkor a hibrid végső eredménye [5,6,1] lesz.
- Ez az eredmény változhat attól függően, hogy hogyan számítjuk a döntetlent.

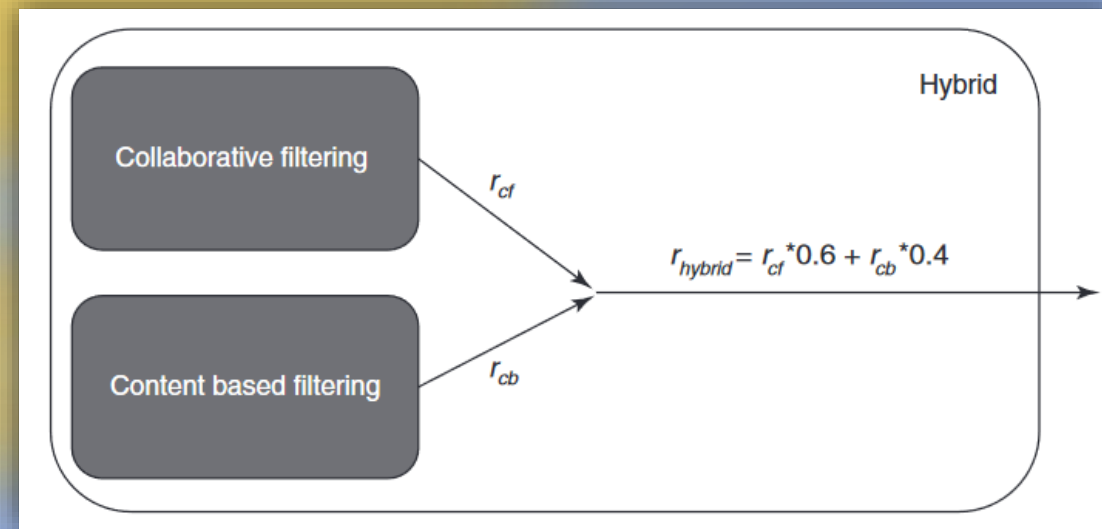
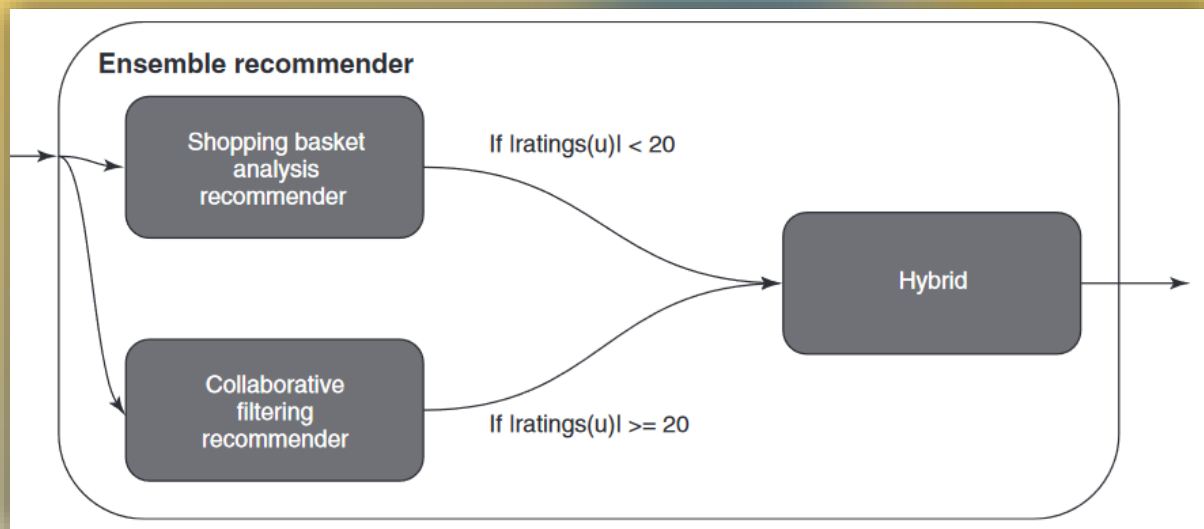




# Kapcsolt és súlyozott együttesek

A bal oldali ábrán egy kapcsolt modellt láthatunk, ami attól függően választ a modellek predikciói közül, hogy a kérdéses felhasználó kevesebb, vagy több terméket értékelt. A mögöttes elgondolás, hogy aki keveset értékelt, annak a vásárlói kosara több releváns információt tartalmazhat.

A jobb oldali pedig egy súlyozott együttes: a benne jelen levő ajánlók megkülönböztetett súllyal számítanak bele a predikcióba. A probléma felvetése, hogy a tartalomalapú rendszer nem tesz különbséget jó és rossz minőség között, a kollaboratív szűrés pedig nem tesz különbséget fontosságban.



# Egy komplexebb ajánló rendszer ökoszisztémája

A felhasználói interakció a bal felső sarokban kezdődik, majd kerül bele a rendszerbe.

Ez egy nem teljes ábra, hiányoznak pl. az adatgyűjtő, modelleket tanító komponensek.

