

# RECOMMENDATION ENGINES

MASTER IN BUSINESS ANALYTICS AND BIG DATA

# ABOUT THIS COURSE... (I/II)

- Session 1, Introduction to Recommendation Engines
- Session 2, Recommendation Methods
- Session 3, Collaborative Filtering
- **Session 4, *Recommendation Engine Labs (Part 1)***
- Session 5, Content-based Filtering and Hybrid Approaches
- **Session 6, *Recommendation Engine Labs (Part 2)***
- Session 7, Building a Recommendation Engine in the Real World
- **Session 8, *Recommendation Engine Labs (Part 3)***
- Session 9 & 10, **Final Project Evaluation**

# TODAY'S AGENDA

- Recapitulation...
- **Session 5, Content-based Filtering & Hybrid Approaches**
  - Content-based Filtering
  - Hybrid Approaches
- **Session 6, Recommendation Engine Labs (Part 2)**
  - Building a Content-based Filtering Engine
  - Building an Hybrid Recommendation Engine

SO FAR...

# RECOMMENDATION METHODS

NON-PERSONALISED

PERSONALISED

COLLABORATIVE FILTERING

CONTENT-BASED FILTERING

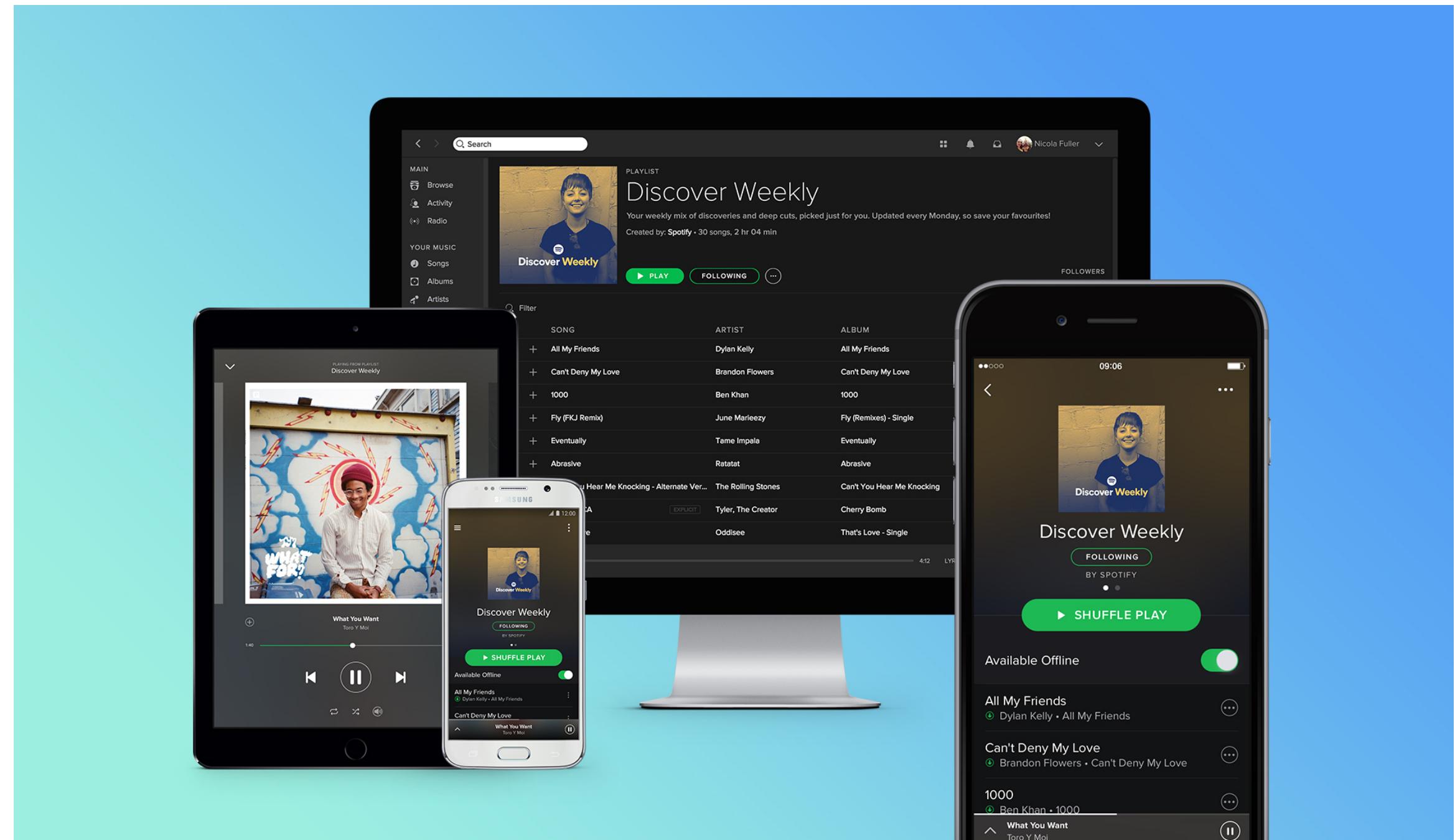
There is NOT a BEST method... it all depends on the domain, goal, data, purpose, ...



# SPOTIFY CASE

## Spotify Discovery

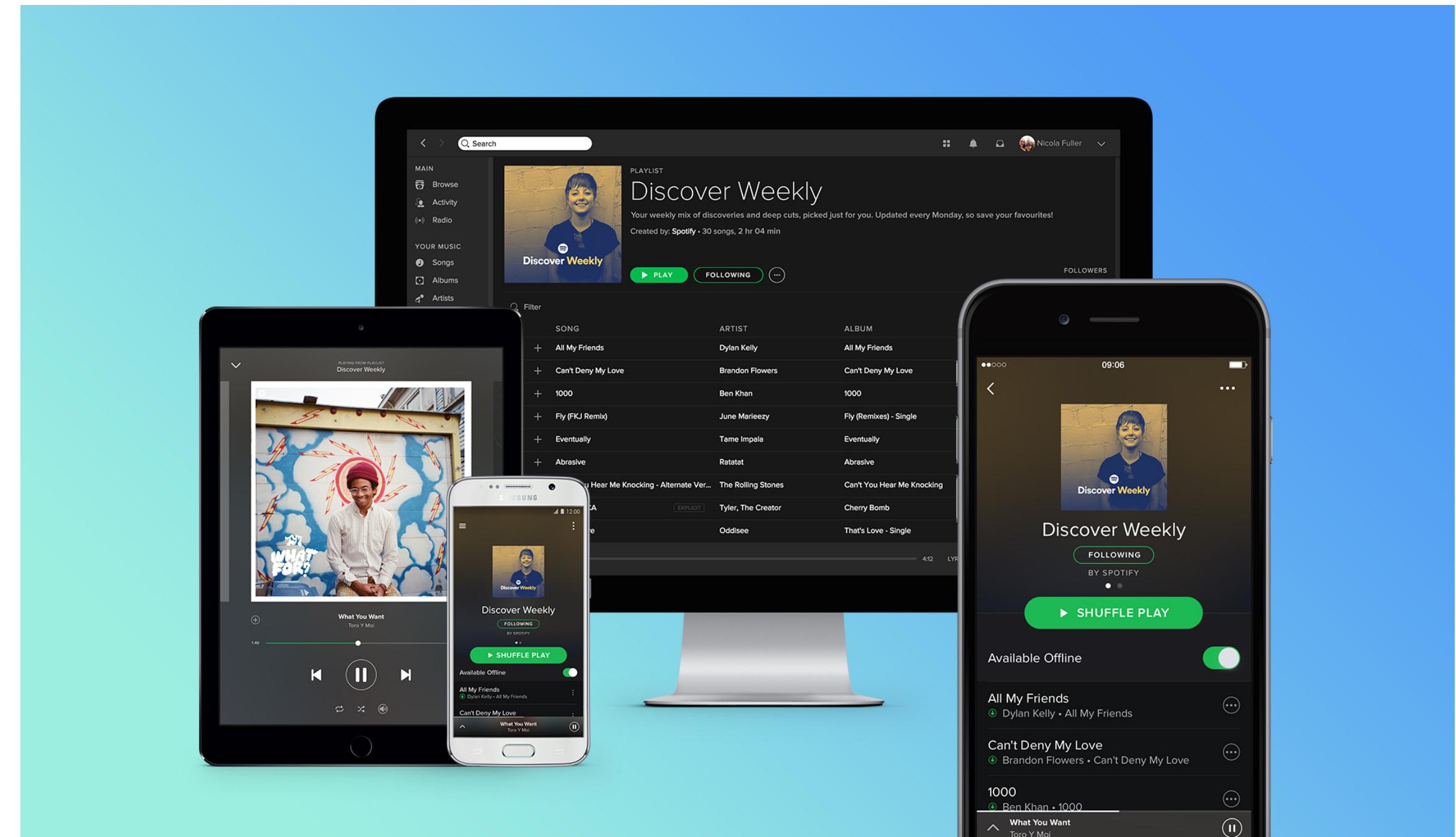
- Why they did it?
- Why it worked so well?
- What were their key findings?
- Downsides nowadays?



# SPOTIFY CASE

## Spotify Discovery

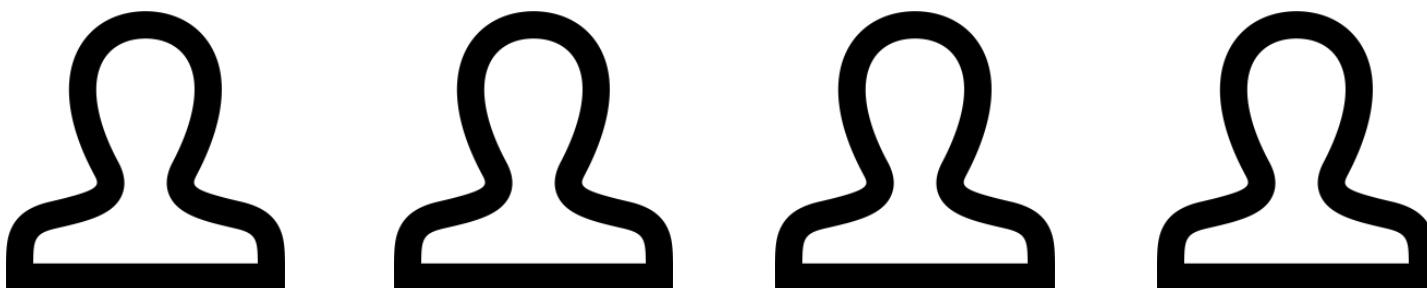
- Reused format
- Quality over quantity
- Habit formation
- Sense of ownership
- UX
- Interdisciplinar
- Feedback loop



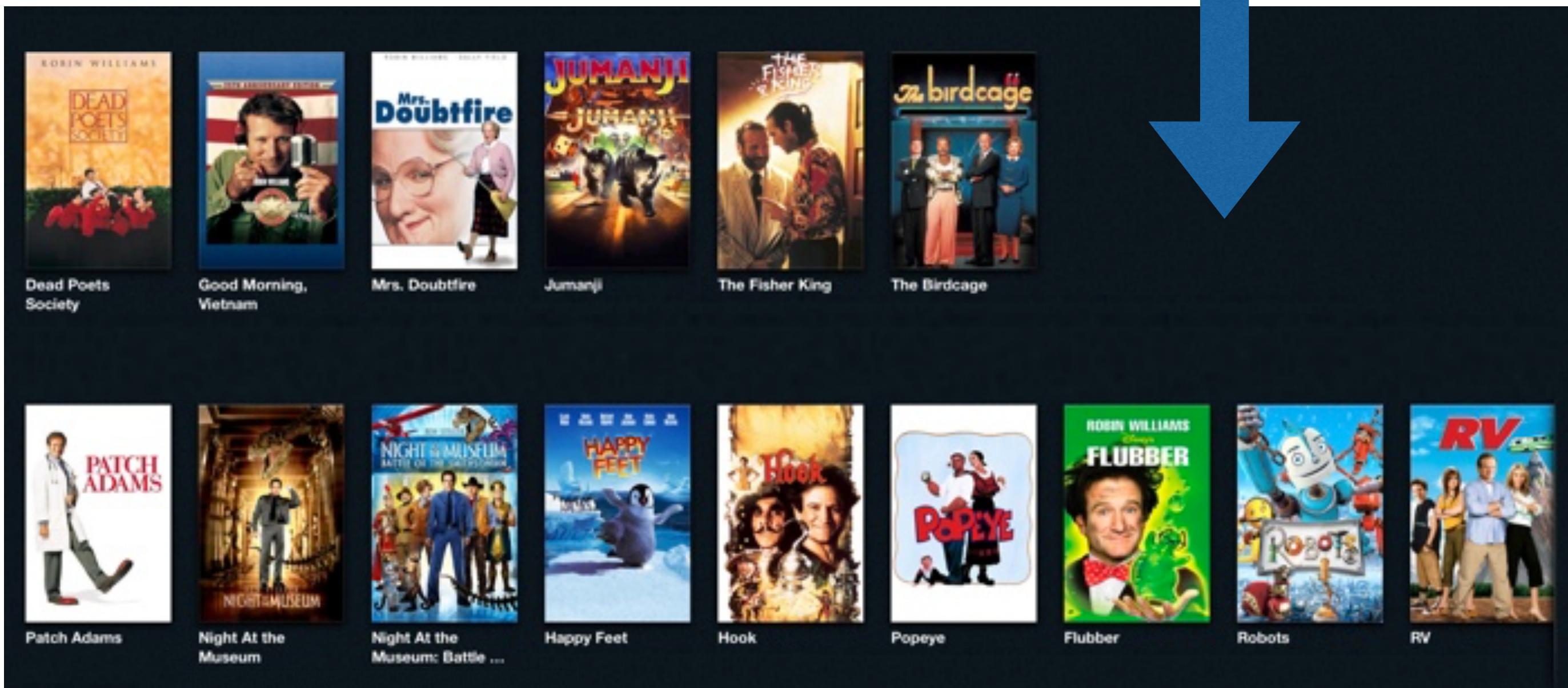
## SESSION 5,

- CONTENT-BASED FILTERING
- HYBRID APPROACHES

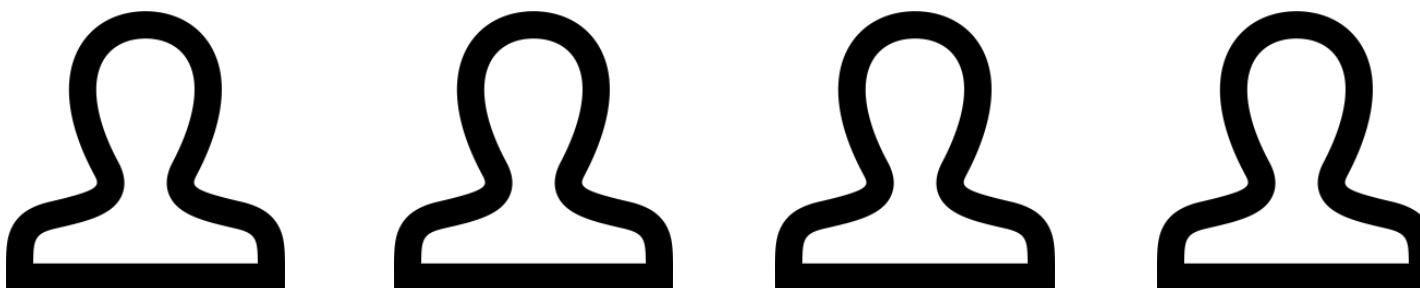
# IN COLLABORATIVE FILTERING...



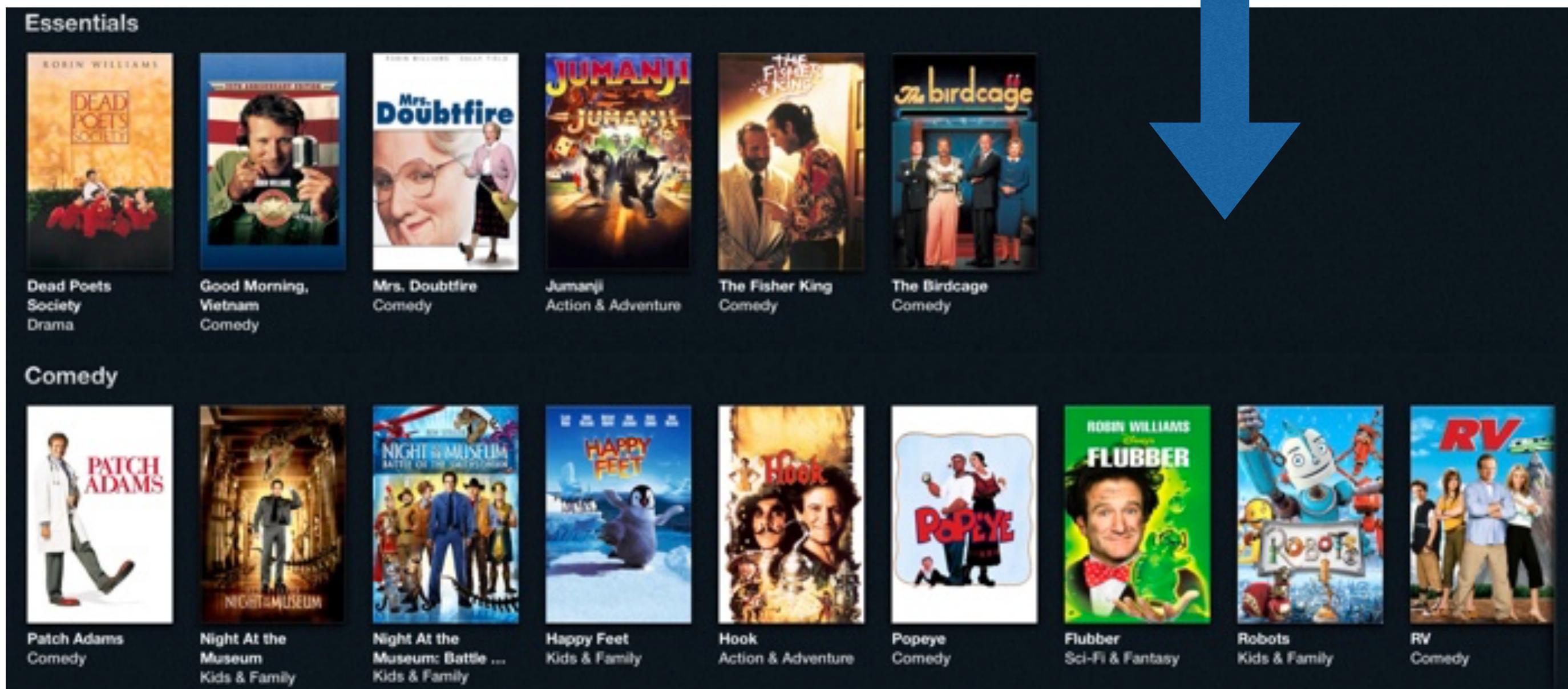
opinions on items



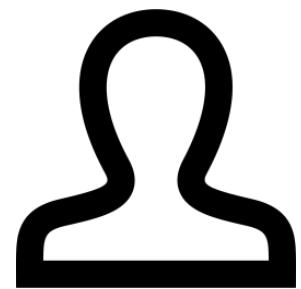
# IN CONTENT-BASED FILTERING...



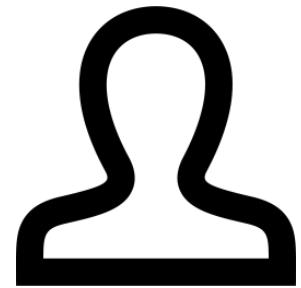
content on items



# IN CONTENT-BASED FILTERING...



action  
woody allen  
english



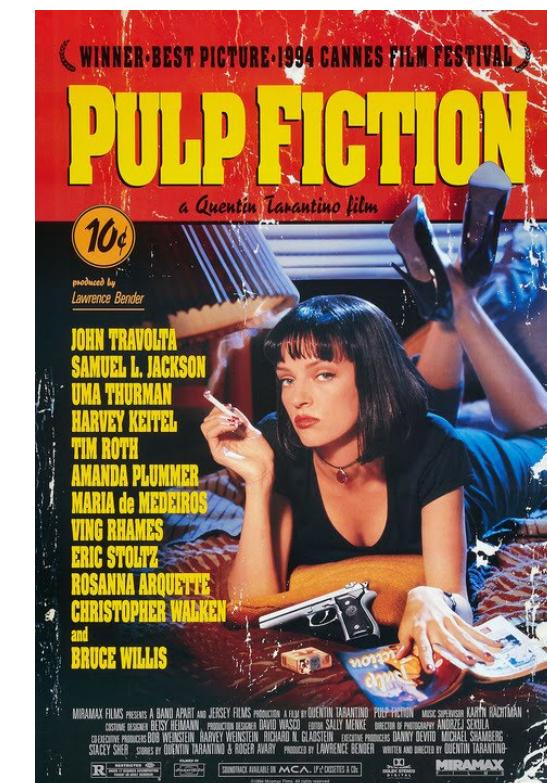
comedy  
modern  
worldwide



terror  
woody allen  
almodÓvar



woody allen  
comedy  
english



tarantino  
english

# IN CONTENT-BASED FILTERING...

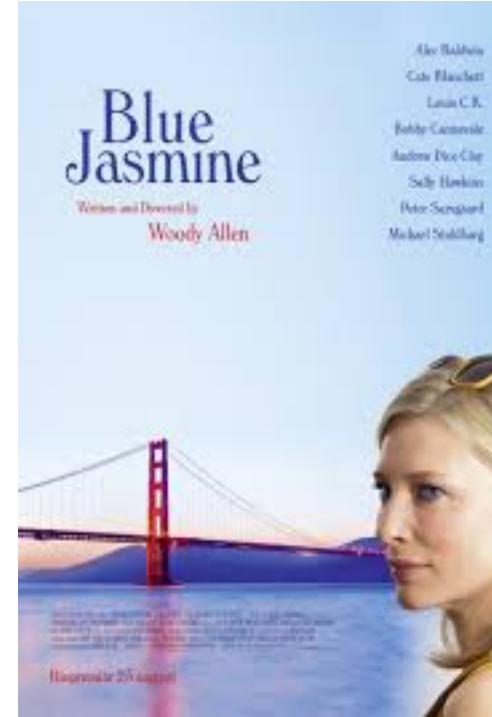
Voilà, we have a RECOMMENDER!



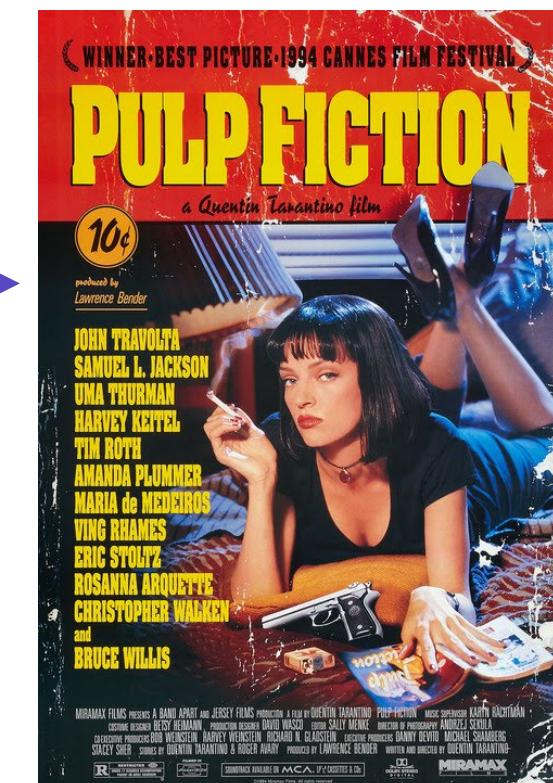
action  
woody allen  
english

woody allen, english  
action

english  
action, woody allen



woody allen  
comedy  
english



tarantino  
english

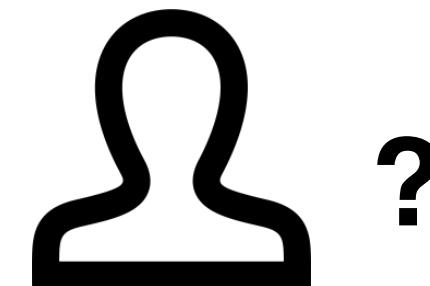
# IN CONTENT-BASED FILTERING...

Questions!



?

- How to model the items
- How to model the users (build their preferences)



?

# IN CONTENT-BASED FILTERING...

The KEY question are:

**How to model the items?**

1. Define a clear taxonomy
2. Assign attributes



**How to build the user preferences?**



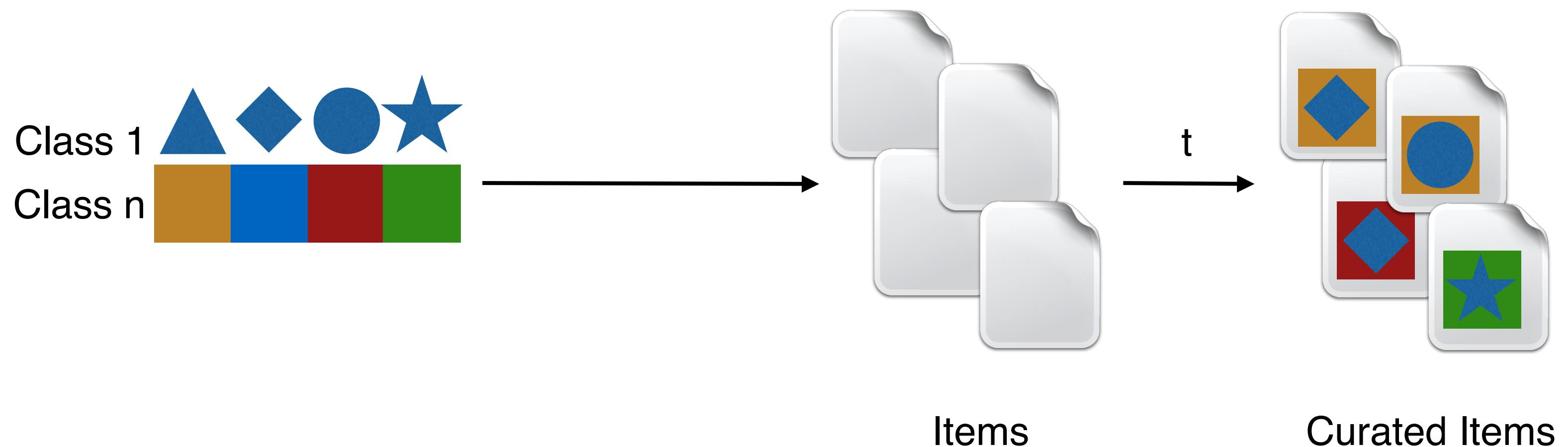
1. Ask explicitly the user: what are your preferences?
2. Deduce user preferences from ratings:
  - explicit ratings
  - implicit ratings

# MODELING ITEMS

Define a taxonomy:

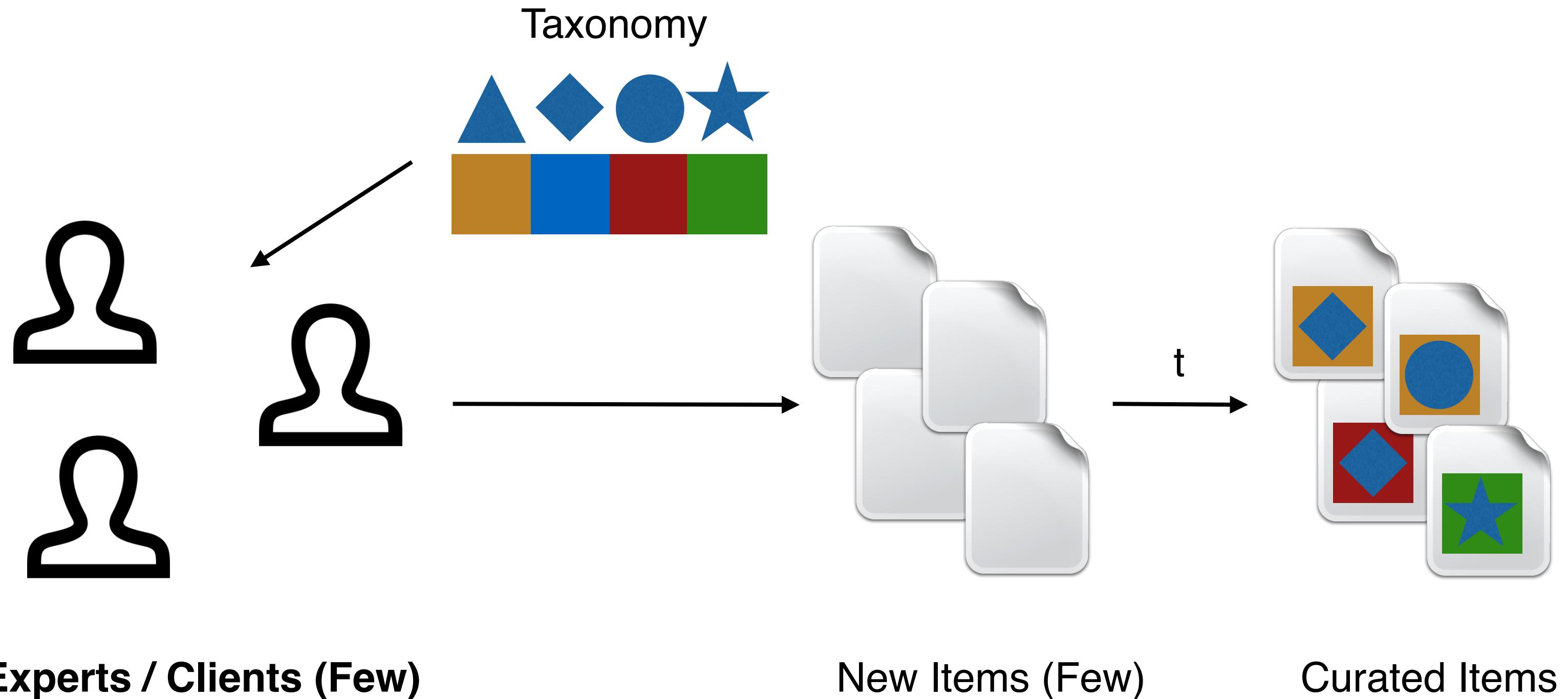
Choose carefully the set of **relevant** attributes to describe the items.

- Easy to understand for the user
- Useful for the recommendations



# MODELING ITEMS

Manual description of items by experts and clients. Good quality / care of the content.



# MODELING ITEMS

Some examples:

**Booking.com**

▼ Property Type

<input type="checkbox"/> Apartments	1400
<input type="checkbox"/> Hotels	386
<input type="checkbox"/> Guesthouses	266
<input type="checkbox"/> Hostels	102
<input type="checkbox"/> Bed and Breakfasts	58
<input type="checkbox"/> Boats	8
<input type="checkbox"/> Vacation Homes	6
<input type="checkbox"/> Villas	3
<input type="checkbox"/> Love Hotels	2

▼ Facility

<input type="checkbox"/> Wi-Fi	2171
<input type="checkbox"/> Parking	1101
<input type="checkbox"/> Airport Shuttle	622
<input type="checkbox"/> Fitness Center	142
<input type="checkbox"/> Non-smoking Rooms	1174
<input type="checkbox"/> Indoor Pool	16
<input type="checkbox"/> Spa	42
<input type="checkbox"/> Family Rooms	1605
<input type="checkbox"/> Outdoor Pool	176
<input type="checkbox"/> Pet Friendly	465
<input type="checkbox"/> Facilities for Disabled Guests	458
<input type="checkbox"/> Restaurant	224

▼ Room Facility

<input type="checkbox"/> Air conditioning	1935
<input type="checkbox"/> Bathtub	715
<input type="checkbox"/> Flat-screen TV	1587
<input type="checkbox"/> Kitchen/kitchenette	1475
<input type="checkbox"/> Patio	188
<input type="checkbox"/> Private pool	24
<input type="checkbox"/> Soundproof	312
<input type="checkbox"/> Spa tub	71
<input type="checkbox"/> Terrace	585
<input type="checkbox"/> View	388
<input type="checkbox"/> Washing machine	1242

# MODELING ITEMS

Some examples:



## Property type

- Vacation Rental House (507)
- Condominium/Apartment (3,785)
- Cabin Vacation Rental (0)
- Private Room (93)
- Resort Vacation Rental (0)
- Specialty Vacation Rental (0)
- Villa (0)

## Payment Method

- Book Online (3941)

## Amenities

- Internet (3,830)
- Air Conditioning (3,336)
- Wi-Fi (2,634)
- Other outdoor space (2,505)
- Washer/Dryer (1,363)
- Elevator/Lift access (756)
- Parking (595)
- Grill (254)
- Garden or Yard (230)
- All pools (194)
- Public pool (144)
- Gym (73)
- Fireplace (59)
- Hot Tub (58)
- Private pool (50)
- Child pool (27)
- Game Room (25)
- Sauna (9)
- Shared Tennis Court (5)

## Neighborhood

## Amenities & Features

## Distinctive Features

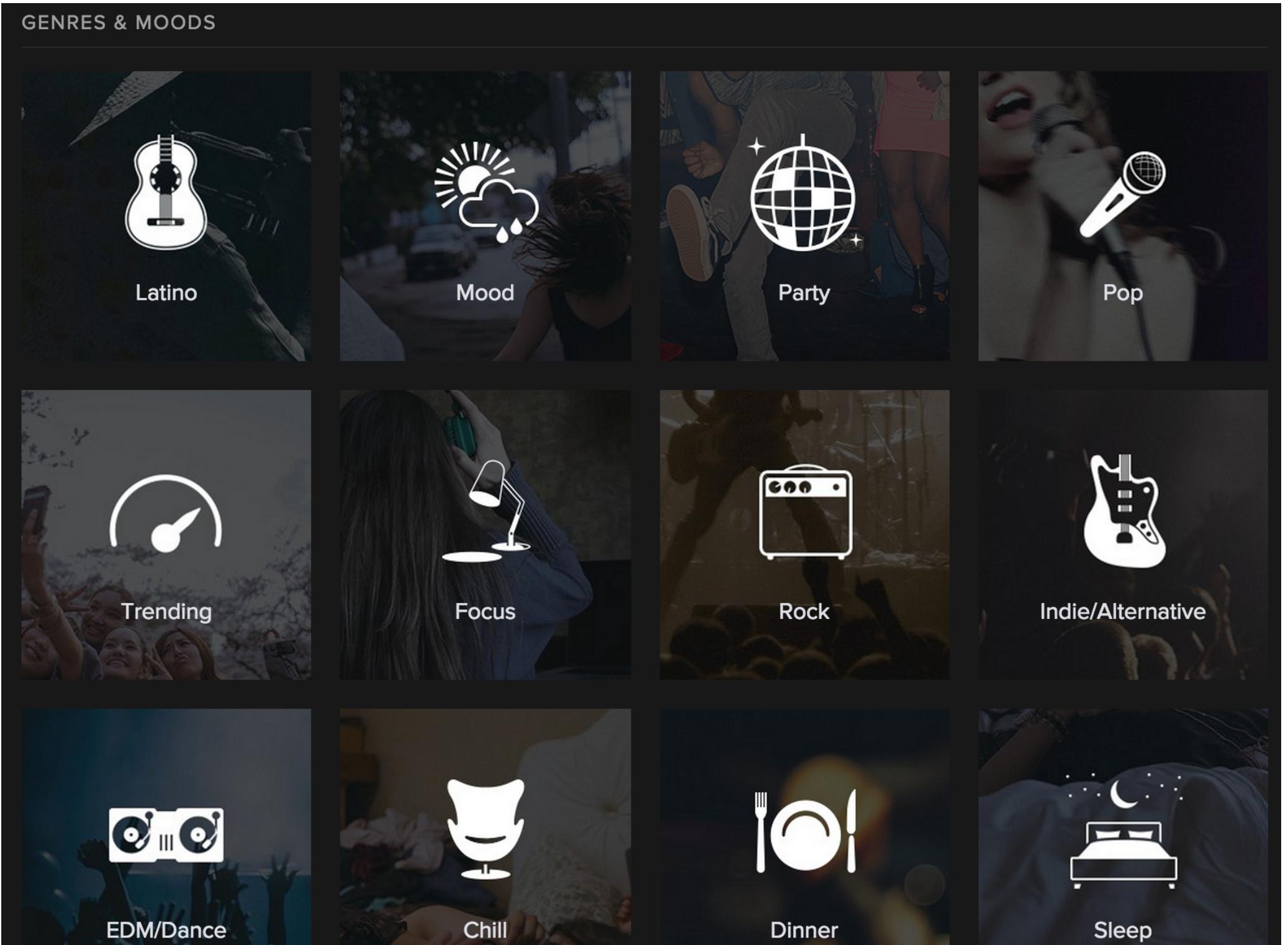
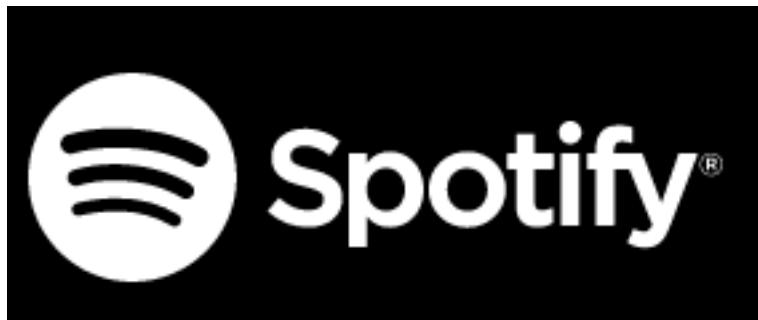
- Waterfront (39)
- Private Fishing Lake/River (1)
- Private Tennis Court (1)
- Boat Available (26)
- Ocean Views (133)
- Mountain Views (147)
- Water Views (56)
- Staffed Property (45)
- Housekeeping Included (142)
- Housekeeping Optional (129)

## Suitability

- Wheelchair access (543)
- Elder access (1,683)
- Suitable for children (2,511)
- Pet friendly (302)
- Smoking allowed (616)

# MODELING ITEMS

Some examples:



# CATEGORIZING CONTENT

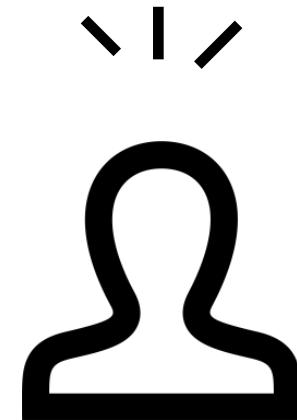
Question!

What are the downsides of manual modeling?

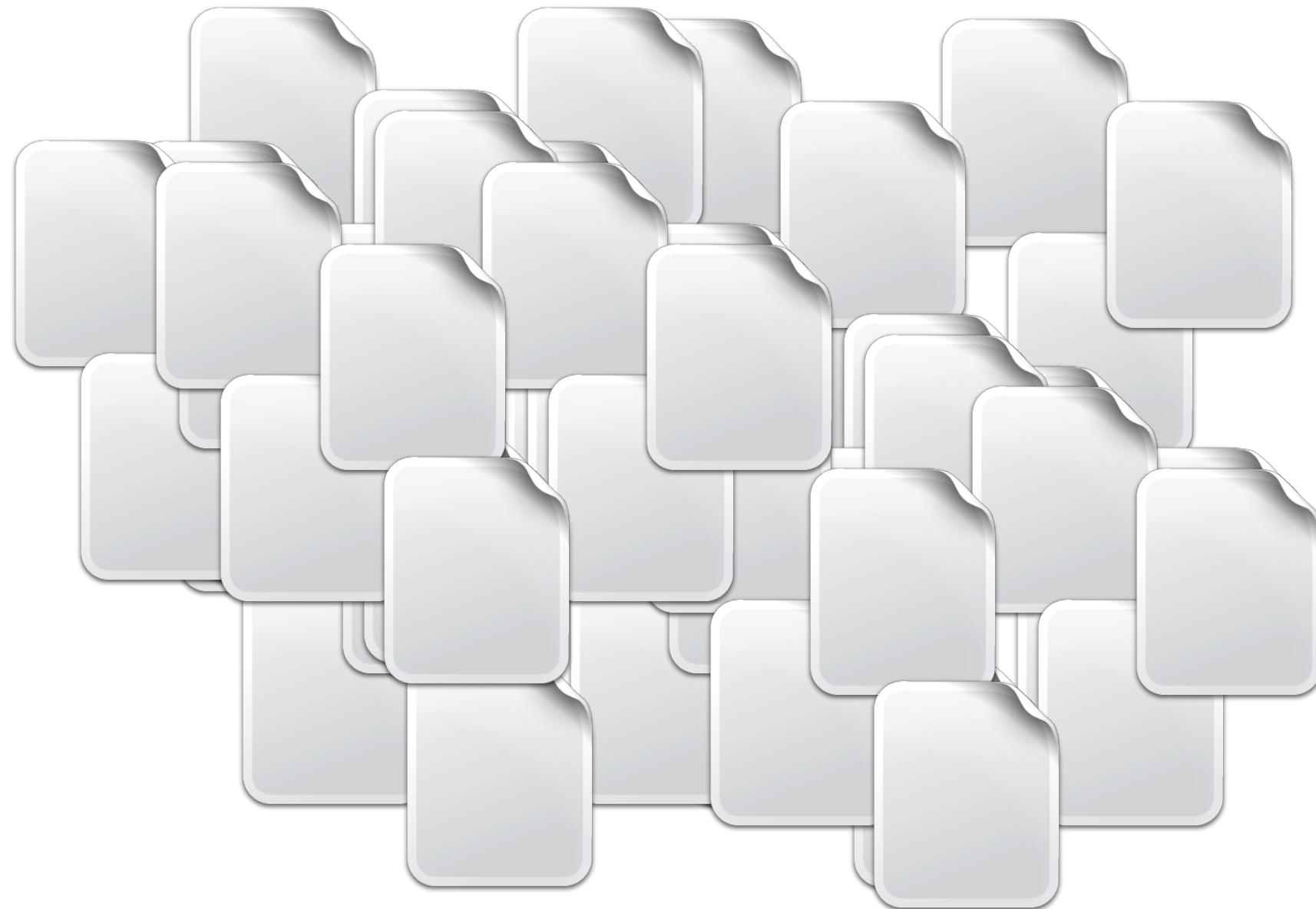


# MODELING ITEMS

If there is too many new items **experts** cannot describe on time each new one. **Scale** problem

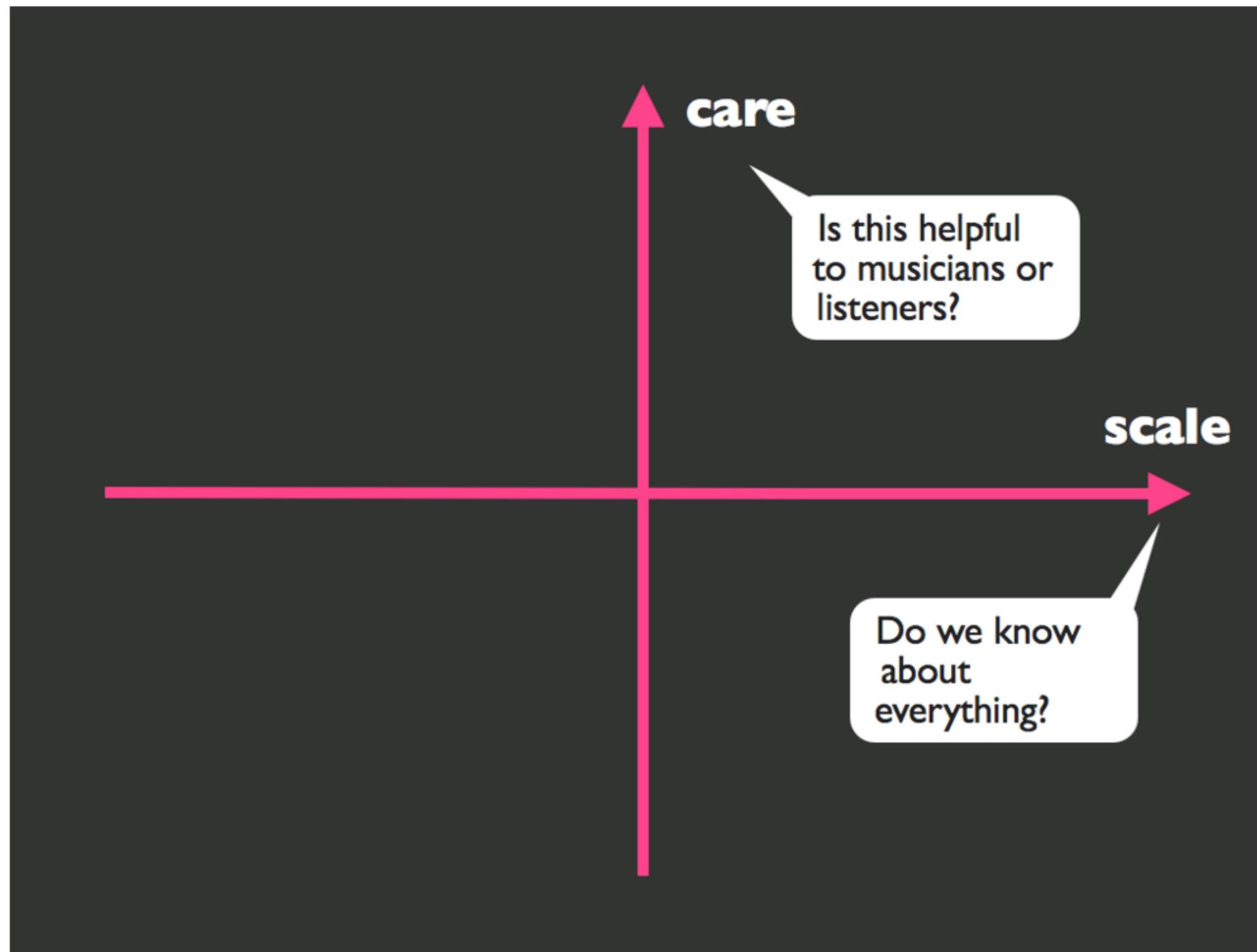


Experts (Few)



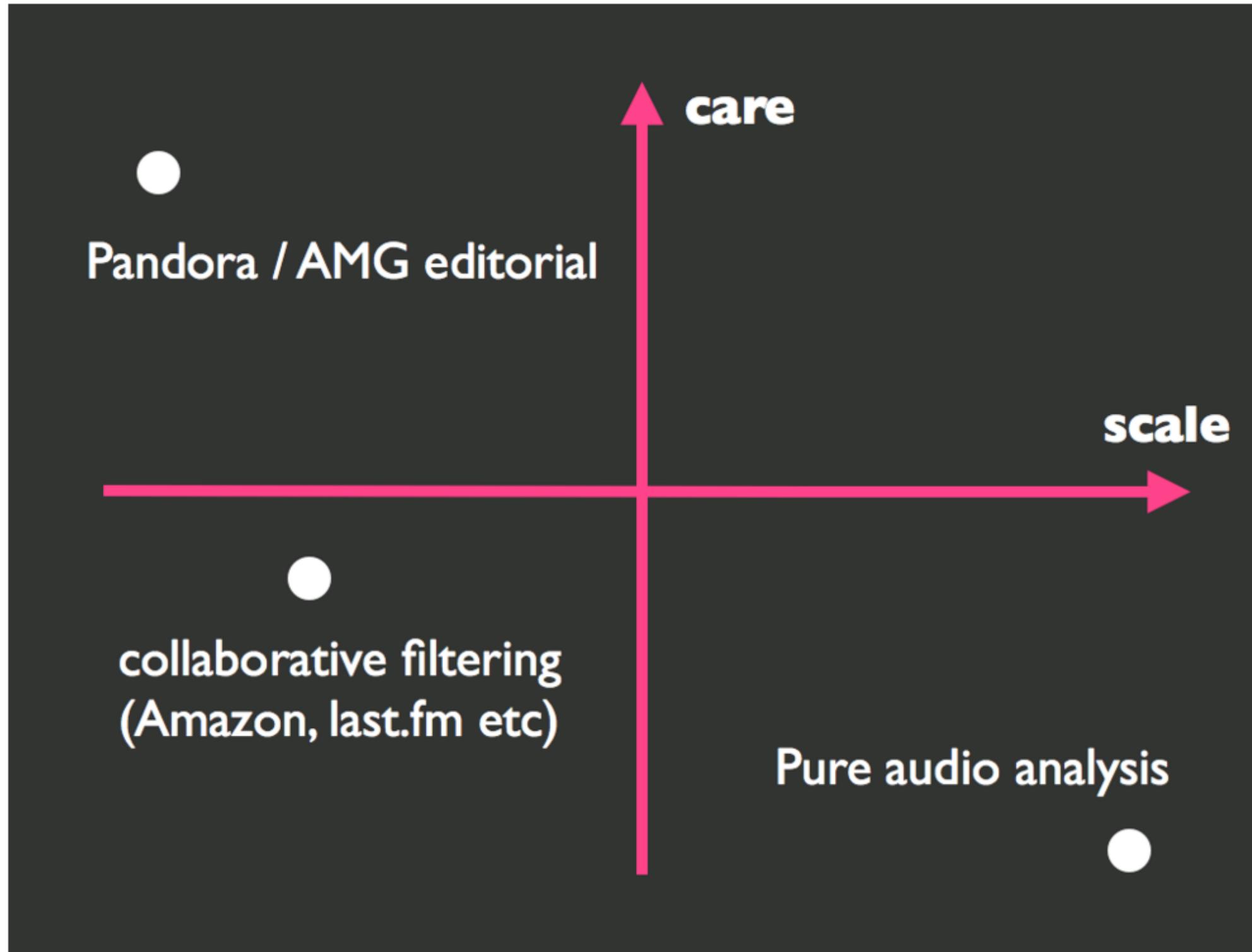
Too many new Items

# ECHO NEST CASE



<http://notes.variogr.am/post/37675885491/how-music-recommendation-works-and-doesnt-work>

# ECHO NEST CASE



# ECHO NEST CASE

CARE:

ARTIST

The Beatles

PLAY FOLLOW ...

8,009,976  
MONTHLY LISTENERS

OVERVIEW RELATED ARTISTS ABOUT

POPULAR

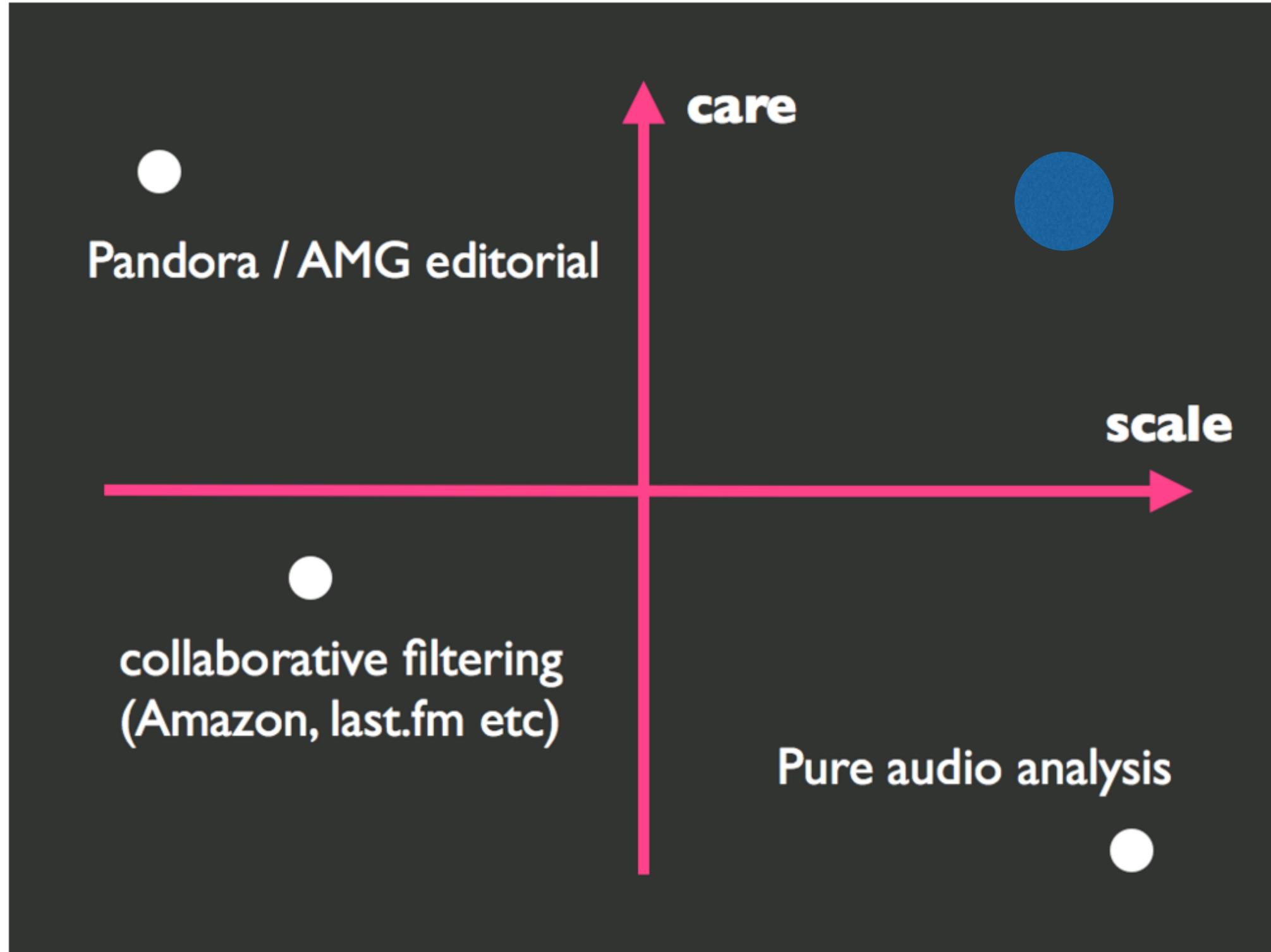
Rank	+	Song	Listeners
1	+	Here Comes The Sun - Remastered 2009	10,367,879
2	+	Let It Be - Remastered 2009	10,541,318
3	+	Hey Jude - Remastered 2015	8,546,130
4	+	I Want To Hold Your Hand - Remastered 2015	7,467,827
5	+	Come Together - Remastered 2009	11,614,937

SHOW 5 MORE

RELATED ARTISTS

- John Lennon
- Paul McCartney
- George Harrison
- Ringo Starr

# ECHO NEST CASE



how artist and song is represented by listeners?

**NLP**

*“computer ‘read’ all that was going on across the internet”*

*“We crawl the web constantly, scanning over 10 million music related pages a day. We throw away spam and non-music related content through filtering, we try to quickly find artist names in large amounts of text and parse the language around the name.”*

+

how a song sounds?

**Acoustic Analysis**

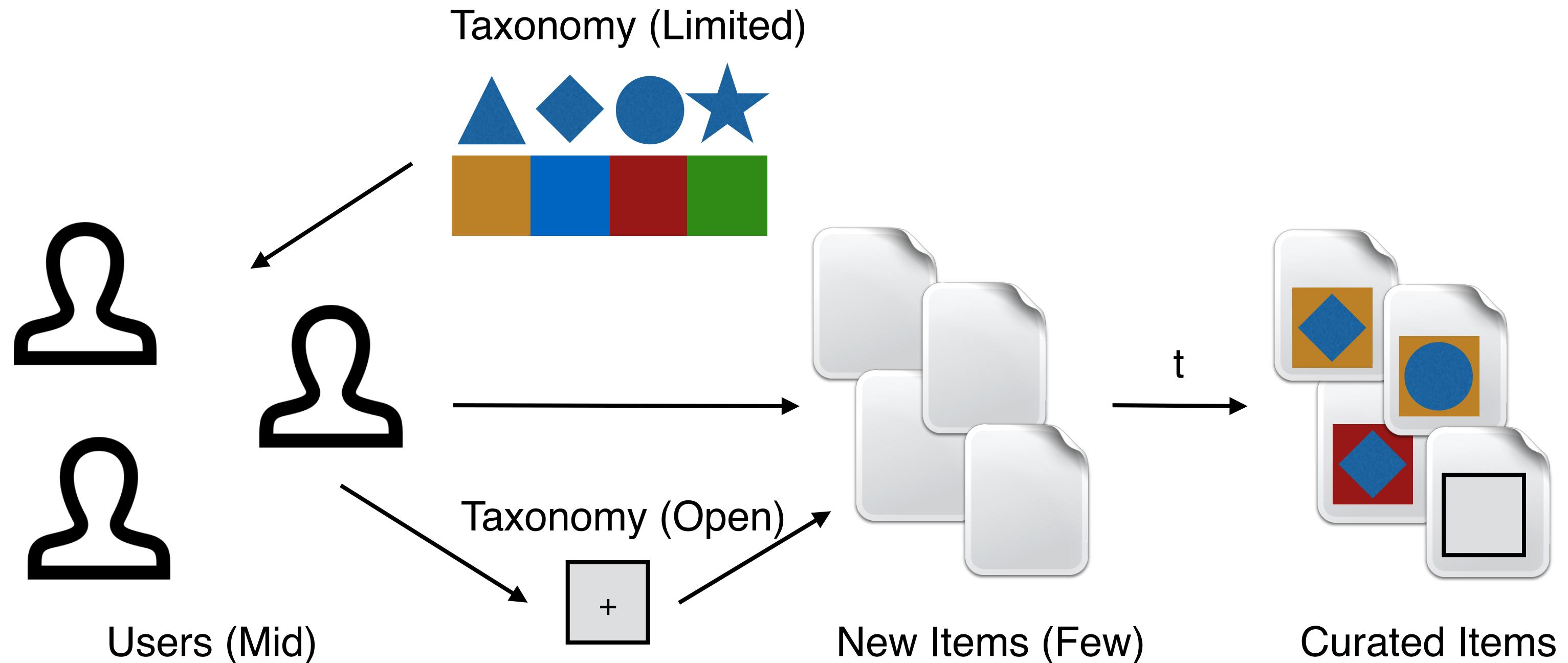
# ECHO NEST CASE

How popular services know about music

Service	Source of data
Pandora	Musicologists take surveys
Songza	Editors or music fans make playlists
Last.fm	Activity data, tags on artists and songs, acoustic analysis <a href="#">[1]</a>
All music guide	Music editors & writers
Amazon	Purchase & browsing history
iTunes Genius	Purchase data, activity data from iTunes <a href="#">[2]</a>
Echo Nest	Acoustic analysis, text analysis

# MODELING ITEMS

Collective classification of content created by the **end users**. Need **quality control**.



# MODELING ITEMS

Some examples:

## Edit Topics

Add and remove topics for this question, or set a context topic that appears before the question to clarify any ambiguous question text.

- x Laughing
- x Fun
- x Jokes
- x Comedy
- x Humor
- x Survey Question

Select Topic

Cancel

Done

Edit Topics



Quora

Search

Home Write Notifications

24 WANT ANSWERS

Latest activity: 42m ago

QUESTION TOPICS

Ballet  
Dance (activity)  
The Human Race and Condition  
Philosophy  
Psychology  
Philosophy of Everyday Life  
Life  
Psychology of Everyday Life

What are the greatest pleasures of human existence?

Write Question Details

Want Answers | 24 Comment 1 Share Downvote ...

58 ANSWERS ASK TO ANSWER

Ivan Tarradellas  
Edit Biography • Make Anonymous

Write your answer, or answer later

Mona Huang  
134 upvotes by Linda Ianovna Blokhina Houston, Effie Mihaloew, Lauren Glenn, (more)

To pee after holding it for an hour.  
Written 30 Apr. 20,519 views.

Upvote | 134 Downvote Comments 7 Share ...

Arqam Ahmad, Searching the lost beauty in humans  
191 upvotes by Marcus Souza, Kyle Murao, Shivam Nitin Babubhai, (more)

For me the greatest pleasure is in

# MODELING ITEMS

Collective classification of content created by a external **task force**. Need **quality control**.

The screenshot shows the Amazon Mechanical Turk homepage. At the top, there's a navigation bar with tabs for 'Your Account', 'HITs' (which is selected), and 'Qualifications'. Below the navigation bar are links to 'Introduction', 'Dashboard', 'Status', and 'Account Settings'. A large yellow banner in the center states 'Mechanical Turk is a marketplace for work.' It explains that businesses and developers get access to an on-demand, scalable workforce, and workers select from thousands of tasks whenever it's convenient. It also mentions '211,687 HITs available.' Below the banner, there are two main sections: 'Make Money by working on HITs' for workers and 'Get Results from Mechanical Turk Workers' for requesters. Both sections include descriptions, icons, and a 'Get Started' button.

**Make Money by working on HITs**

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. [Find HITs now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

**Find an interesting task** → **Work** → **Earn money**

[Find HITs Now](#)

or [learn more about being a Worker](#)

**Get Results from Mechanical Turk Workers**

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. [Get Started.](#)

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

**Fund your account** → **Load your tasks** → **Get results**

[Get Started](#)

# CATEGORIZING CONTENT

Automatic classification of content. Need quality control.

Unsupervised Learning Problems

i.e. Text analysis to group news or to classify emails on spam / not spam, market segmentation ...

Top Stories

**US diplomatic offices attacked in Egypt, Libya**

CBS News - 17 minutes ago Updated at 6:04 pm ET (CBS/AP) CAIRO - A film attacking Islam's prophet, Muhammad, has sparked aggressive protests against US interests in both Egypt and Libya.

Egyptian protesters scale US Embassy wall in Cairo [The Associated Press](#)  
Angry protesters scale the walls of US embassy in Cairo [Daily News Egypt](#)

Your preferred source: [Mysterious Anti-Muslim Movie Prompts Protest in Egypt](#) [New York Times](#)

Highly Cited: [Protesters storm US Embassy in Cairo](#) [CNN](#)  
From Egypt: [Egyptian foreign ministry commits to protecting embassies](#) [Ahram Online](#)

Related  
[Cairo »](#)  
[Muhammad »](#)  
[Egypt »](#)

**Israeli Leader Sharpens Call on US to Set Limits on Iran**

New York Times - 40 minutes ago WASHINGTON - Prime Minister Benjamin Netanyahu of Israel inserted himself into the most contentious foreign policy issue of the American presidential campaign on Tuesday, criticizing the Obama administration for refusing to set clear "red lines" on ...

**In Chicago Strike, 2 Sides Differ on How Much They Differ**

New York Times - 1 hour ago CHICAGO - Contract negotiations continued Tuesday between Chicago Public Schools officials and the city's teachers' union as 350000 students stayed out of classes for a second day.

Poll: 47% of Chicago registered voters support teachers in strike [Chicago Sun-Times](#)  
Negotiators return to bargaining table as Chicago teachers strike rolls into ... [Fox News](#)

Your preferred source: [Striking Chicago teachers not close to a deal, union says](#) [CNN](#)

# ECHO NEST CASE

## Text Analysis

n2 Term	Score	np Term	Score	adj Term	Score
dancing queen	0.0707	dancing queen	0.0875	perky	0.8157
mamma mia	0.0622	mamma mia	0.0553	nonviolent	0.7178
disco era	0.0346	benny	0.0399	swedish	0.2991
winner takes	0.0307	chess	0.0390	international	0.2010
chance on	0.0297	its chorus	0.0389	inner	0.1776
swedish pop	0.0296	vous	0.0382	consistent	0.1508
my my	0.0290	the invitations	0.0377	bitter	0.0871
s enduring	0.0287	voulez	0.0377	classified	0.0735
and gimme	0.0280	something's	0.0374	junior	0.0664
enduring appeal	0.0280	priscilla	0.0369	produced	0.0616

*Echo Nest Cultural vectors*

# BUILDING THE USER PREFERENCES

Building a **shirt recommender** with explicit preferences:



1. Ask explicitly the user: what are your preferences?

- Blue, red, white, **black**, pink
- No sleeves, **short sleeves**, long sleeves
- **Cotton**, synthetic, others

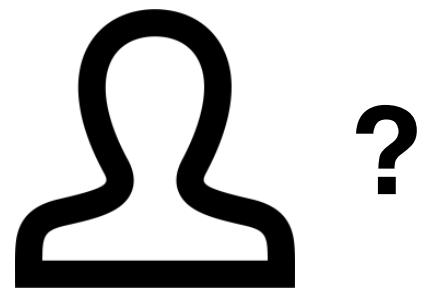
2. Recommend based on the user preferences:

- Red, no sleeves, synthetic: 0
- Black, long sleeves, cotton: 2
- Blue, short sleeves, synthetic: 1

# BUILDING THE USER PROFILE

Question!

Downsides of explicit preferences?



# BUILDING THE USER PREFERENCES

Building a **shirt recommender** with deducing preferences:



1. Collect rated items:

- like: (blue, short sleeves, cotton)
- like: (red, short sleeves, synthetic)
- like: (white, short sleeves, cotton)

2. Deduce preferences:

- (blue: 1, red: 1, white: 1)
- (short sleeves: 3, long sleeves: 0, no sleeves: 0)
- (cotton: 2, synthetic: 1, others: 0)

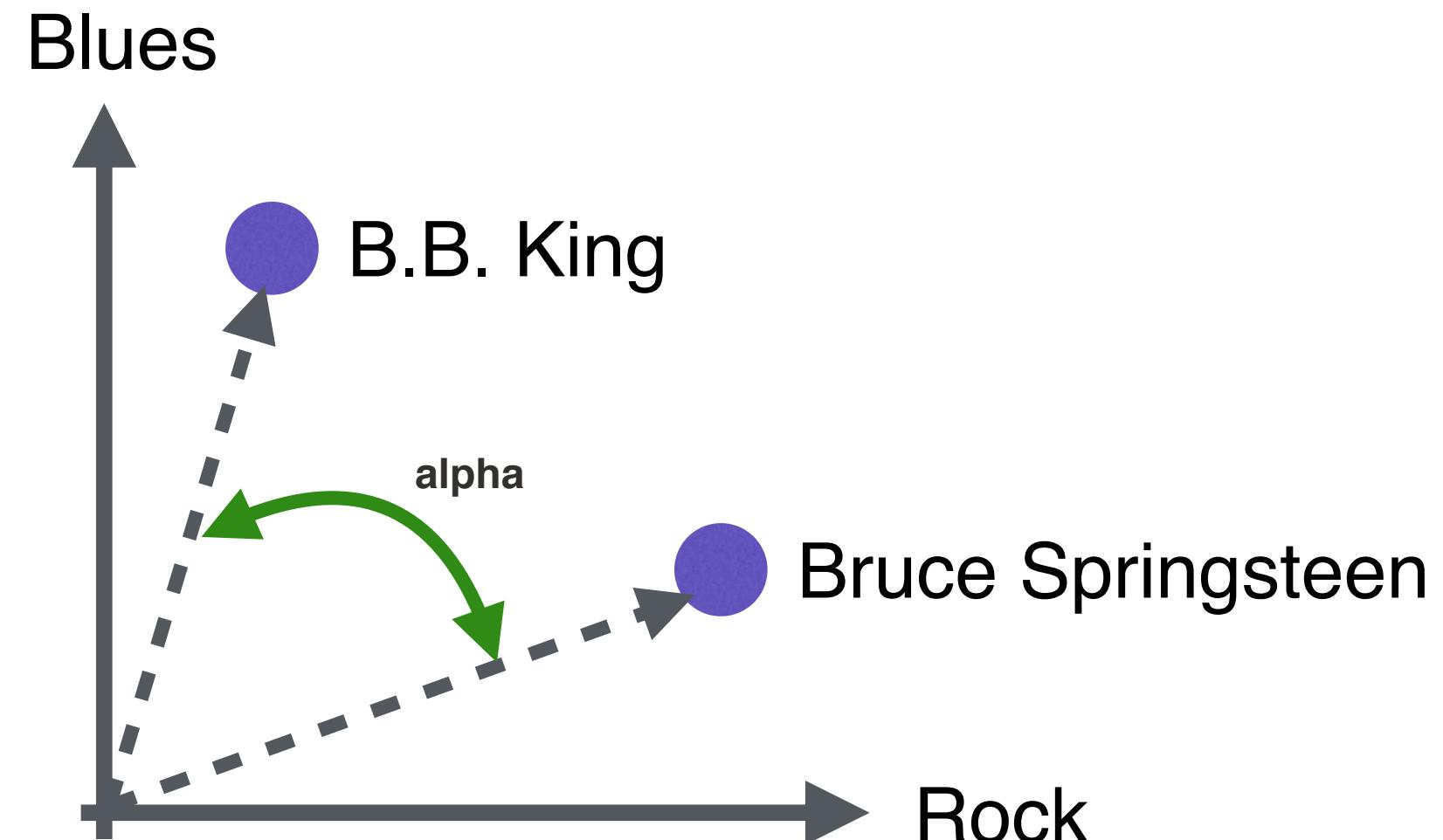
# BUILDING THE USER PREFERENCES

A better approach

# KEYWORD VECTORS FOR ITEMS

- All *Keywords* define the *Vector Space*
- *Keywords are Dimensions*
- *Items* are *Vectors*

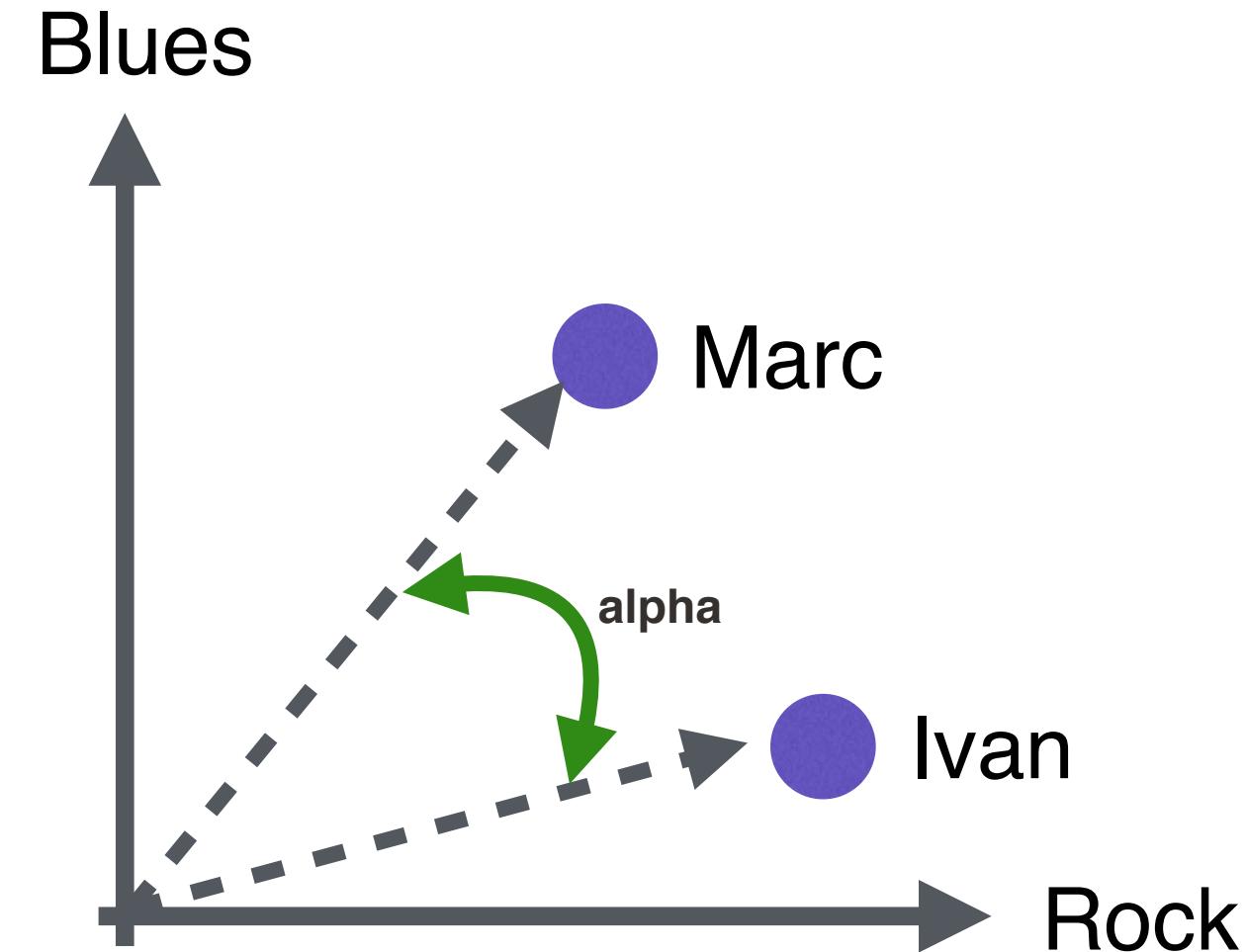
`simil(BBKing, Bruce)=  
cosine(alpha)`



# KEYWORD VECTORS FOR USER PREFERENCES

- All **Keywords** define the **Vector Space**
- **Keywords are Dimensions**
- **User Preferences** are **Vectors**

`simil(Ivan, Marc)=  
cosine(alpha)`



# WHAT IS THE CONCEPT BEHIND CB FILTERING?

Question!

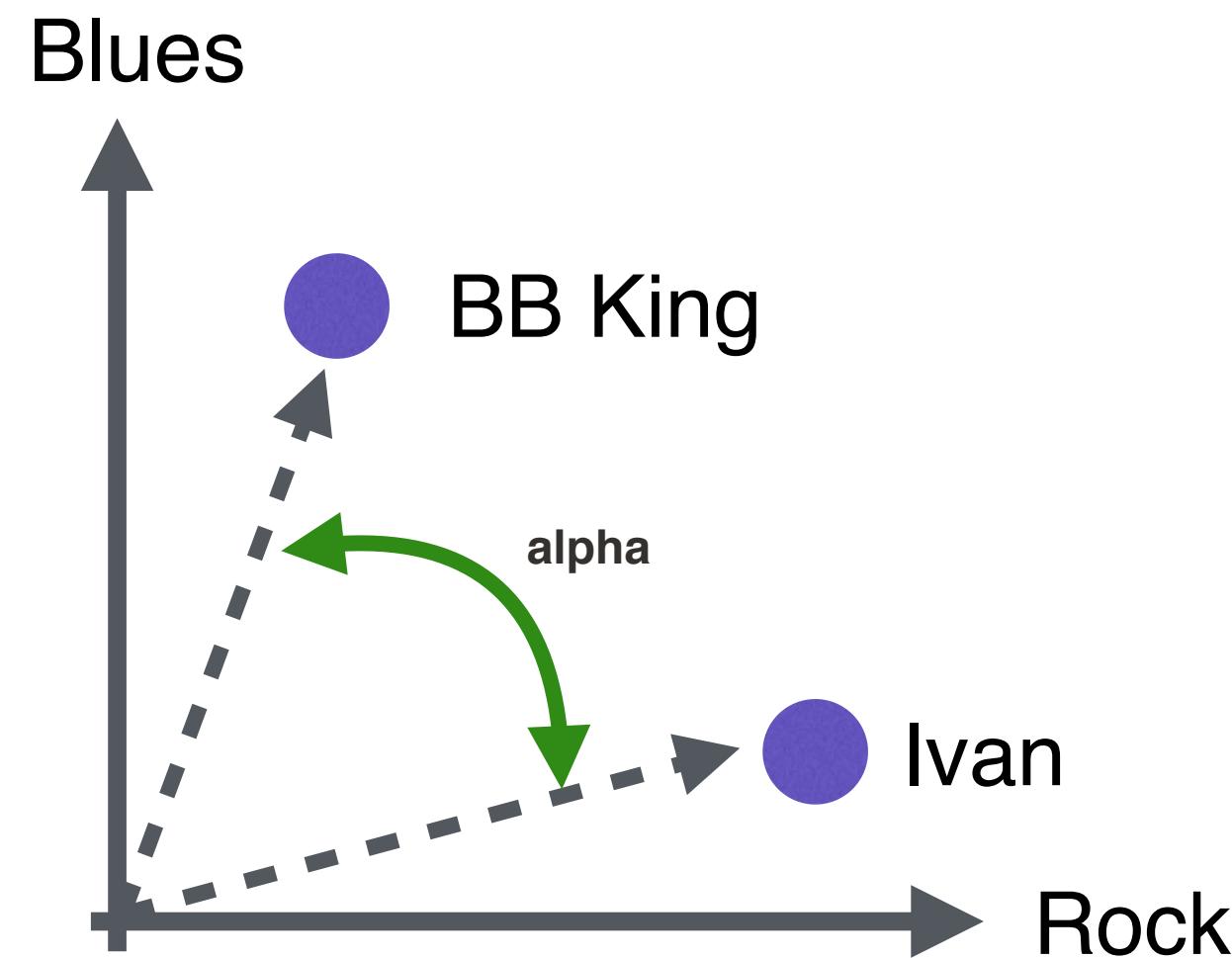
We have seen similarity between items based on attributes.  
We have seen similarity between users based on preferences.

How then you would calculate the similarity between an item and a user?

# WHAT IS THE CONCEPT BEHIND CB FILTERING?

HOW SIMILAR ARE THE TWO VECTORS!

**rating(Ivan, BBKing) =**  
 $\text{cosine}(\alpha)$



# DIFFERENT WAYS TO BUILD KEYWORD VECTORS...

- Boolean vectors (the color of a car)
- Integer number for countable things (rated tags!)
- Real number for relevance/intensity/degrees (rock, blues, jazz, ...)
- *TFIDF concept = term frequency \* log (#doc/#doc with term)*
  - BB King has been tagged 33 times into **Blues** and 10 into **Rock**
  - **Blues** is in 340 artists out of 2,500 artists
  - **Rock** is in 2,000 artists out of 2,500 artists
  - **BBKing: (Blues, 33 \* log (2,500 / 340)**
  - **BBKing: (Rock, 10 \* log (2,500 / 2,000)**
- Usually those vectors are normalised

# HOW DO WE BUILD USER PREFERENCES?

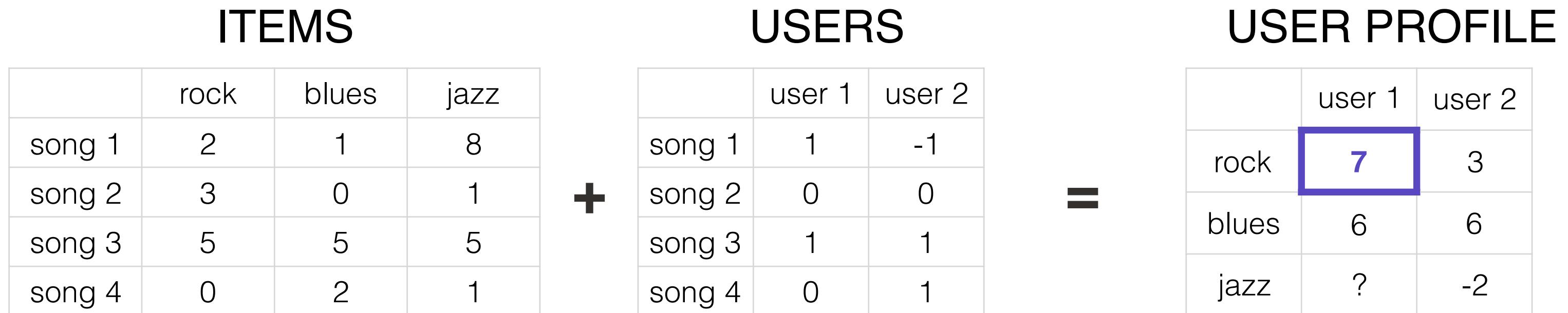
Questions?

# HOW DO WE BUILD USER PREFERENCES?

- **Aggregating items vectors** that are rated (purchased, listened to, watched, liked, ...)
- Different ways to do that:
  - Simply unary
  - Unary with threshold
  - Weight (only positives)
  - Weight (including negatives)
  - ...

# HOW DO WE BUILD USER PREFERENCES?

Aggregating items vectors through Simply Unary

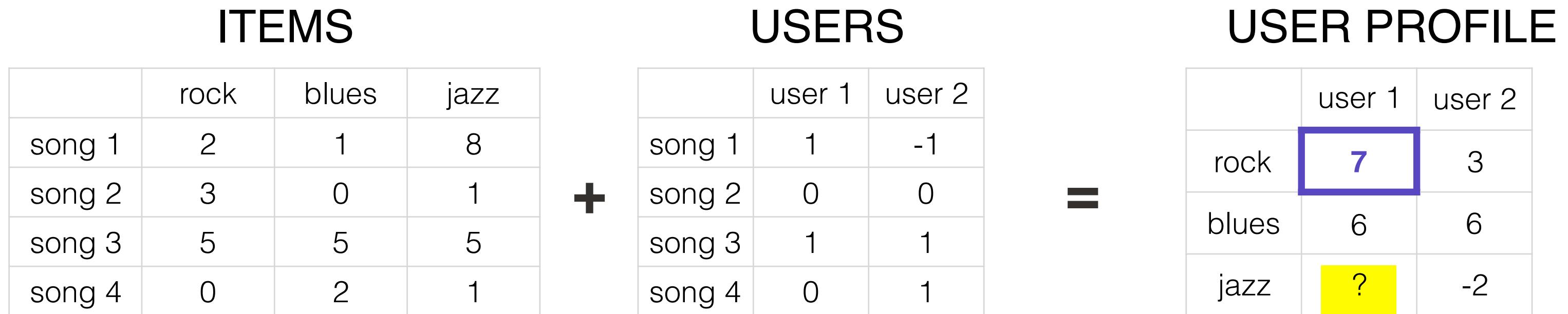


ex: (user 1, rock) =  $1*2 + 0*3 + 1*5 + 0*0 = 7$

- It is **not normalised** by number of votes/rates present in an item
- All keywords are considered **equally significant** (The Beatles vs Serrat)

# HOW DO WE BUILD USER PREFERENCES?

Aggregating items vectors through Simply Unary



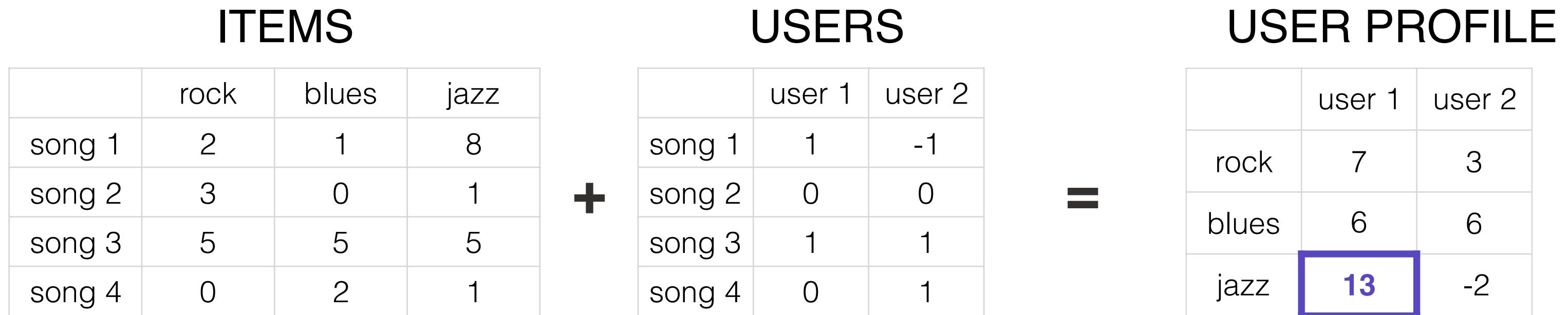
$$\text{ex: (user 1, rock)} = 1*2 + 0*3 + 1*5 + 0*0 = 7$$

Question!

- It is **not normalised** by number of votes/rates present in an item
- All keywords are considered **equally significant** (The Beatles vs Serrat)

# HOW DO WE BUILD USER PREFERENCES?

Aggregating items vectors through Simply Unary

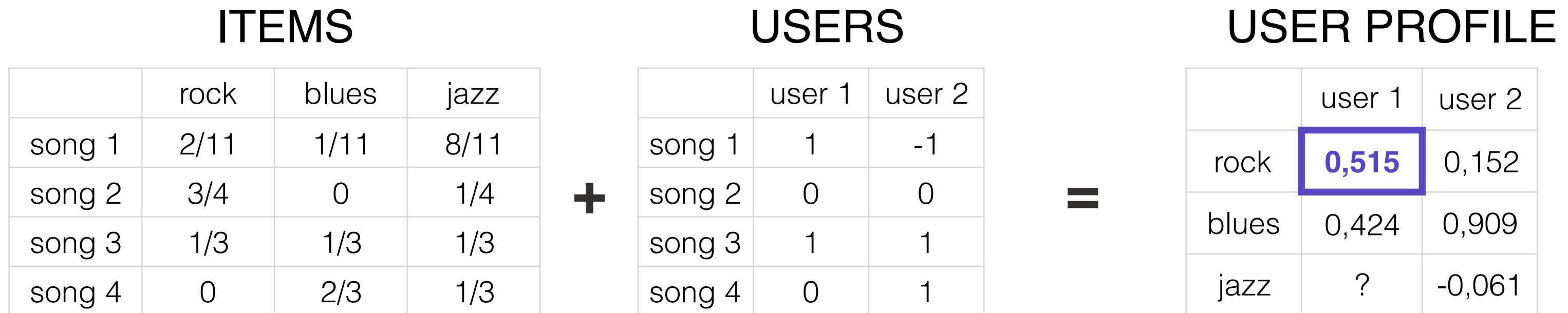


$$\text{ex: } (\text{user 1, jazz}) = 1*8 + 0*1 + 1*5 + 0*1 = \boxed{13}$$

- It is **not normalised** by number of votes/rates present in an item
- All keywords are considered **equally significant** (The Beatles vs Serrat)

# HOW DO WE BUILD USER PREFERENCES?

Aggregating items vectors through Unit Weight



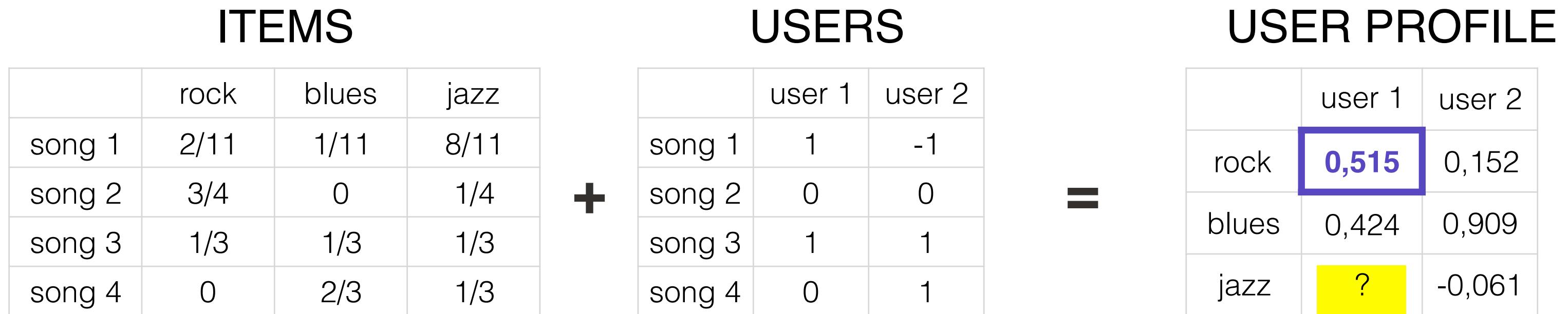
ex: (user 1, rock) =  $1 \cdot 2/11 + 0 \cdot 3/4 + 1 \cdot 1/3 + 0 \cdot 0 = 0,515$

✓ It is **not normalised** by number of votes/rates present in an item

- All keywords are considered **equally significant** (The Beatles vs Serrat)

# HOW DO WE BUILD USER PREFERENCES?

Aggregating items vectors through Unit Weight



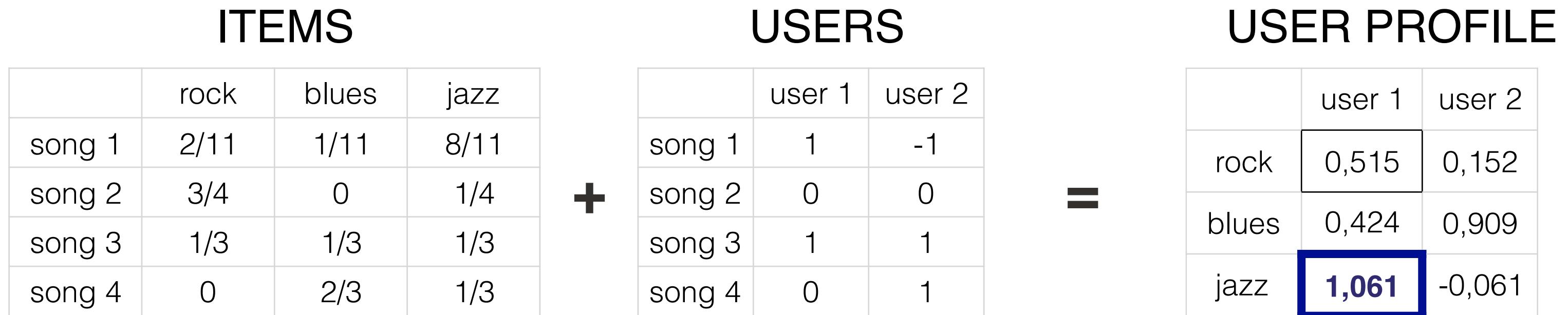
ex: (user 1, rock) =  $1 \cdot 2/11 + 0 \cdot 3/4 + 1 \cdot 1/3 + 0 \cdot 0 = 0,515$

Question!

- ✓ It is **not normalised** by number of votes/rates present in an item
- All keywords are considered **equally significant** (The Beatles vs Serrat)

# HOW DO WE BUILD USER PREFERENCES?

Aggregating items vectors through Unit Weight



$$\text{ex: } (\text{user 1, jazz}) = 1 * 8/11 + 0 * 1/4 + 1 * 1/3 + 0 * 1/3 = \boxed{1,061}$$

✓ It is **not normalised** by number of votes/rates present in an item

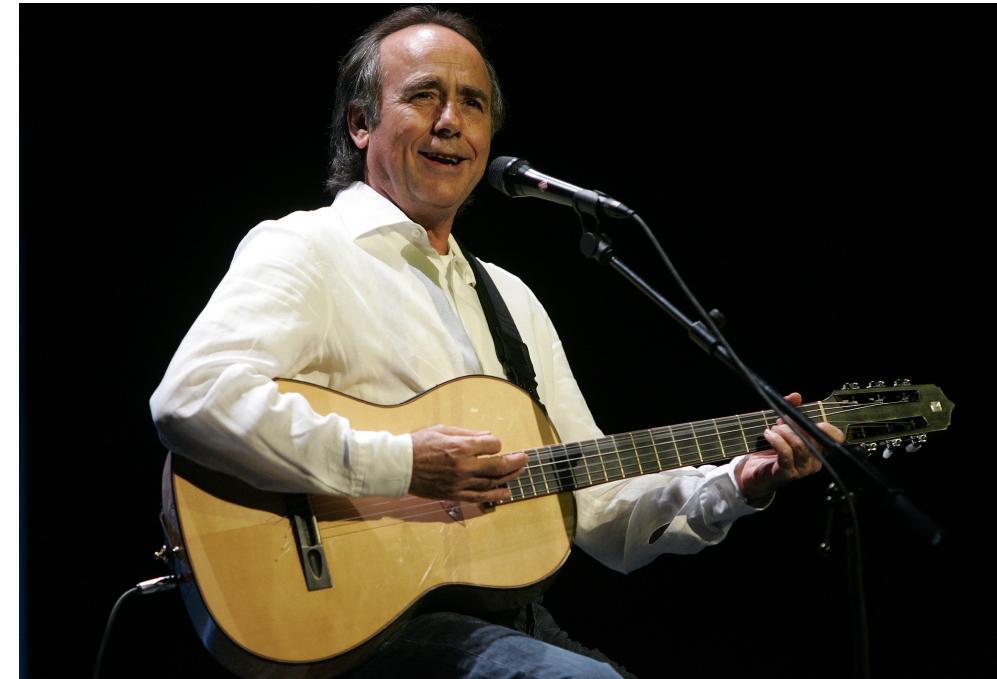
- All keywords are considered **equally significant** (The Beatles vs Serrat)

# ARE ALL PREFERENCES/ATTRIBUTES EQUALLY RELEVANT?

If a user likes **The Beatles**



If a user likes **Serrat**



vs

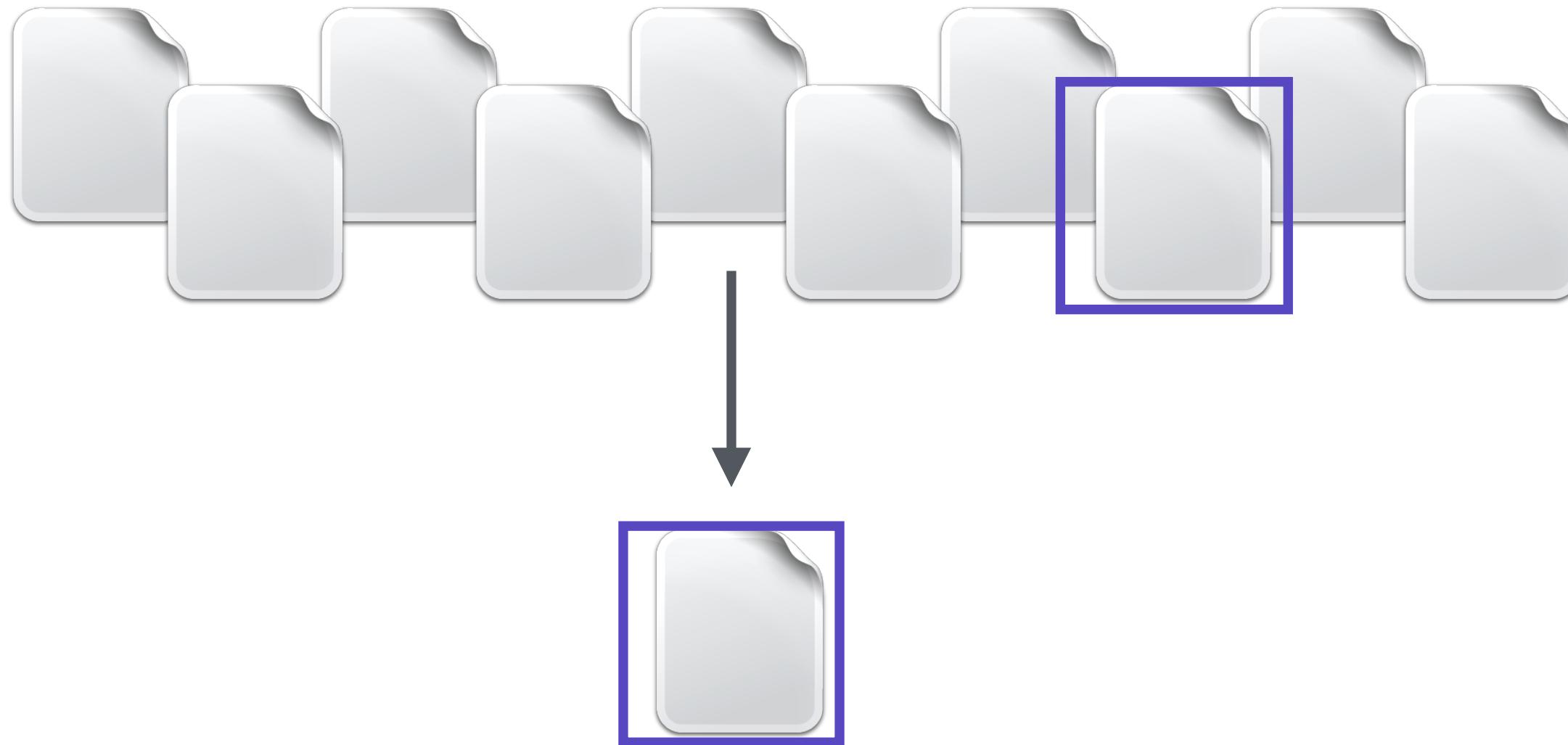
thinking...

# THE INFORMATION FILTERING PROBLEM

In the **Search Problem...**

*“The Stethoscope”*

Corpus:



# THE INFORMATION FILTERING PROBLEM

In the **Search Problem**, usually we consider two factors:

- Term Frequency
- Not all Terms are equally relevant

# THE INFORMATION FILTERING PROBLEM

TFIDF (Term Frequency Inverse Document Frequency)

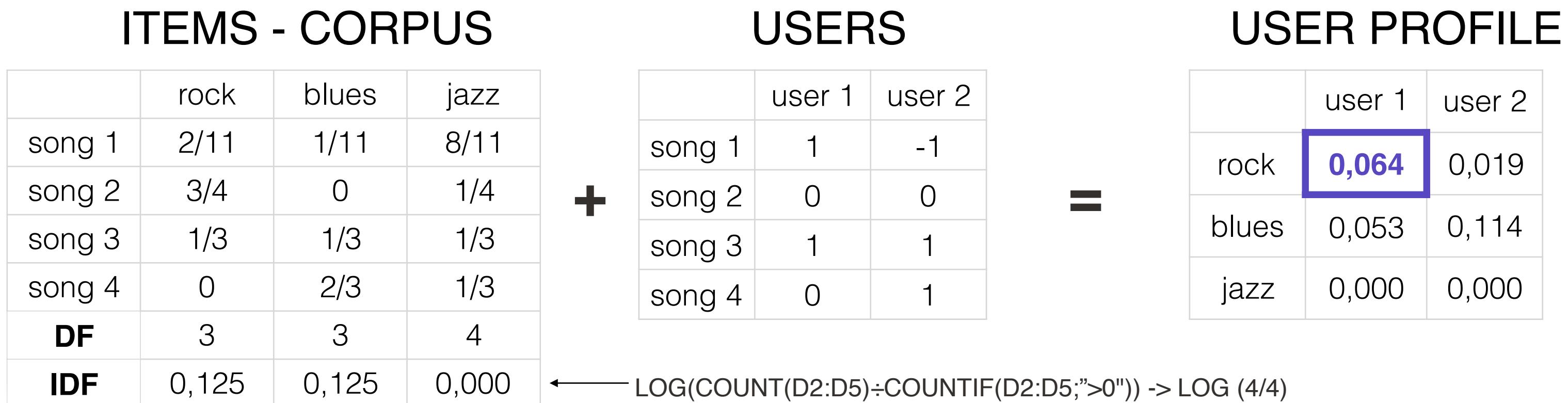
- **TFIDF = TF \* IDF**
  - **TF** = #occurrences of term
    - *How frequent is the term occurring in the document?*
  - **IDF** = log (#documents / #documents with term)
    - *How rare is the term to be in a document?*
- 

“*The Stethoscope*”

*Stethoscope*: IDF is very high  
*The*: IDF is very low

# HOW DO WE BUILD USER PREFERENCES?

Aggregating items vectors through TFIDF (Inverse Document Frequency)



$$\text{ex: (user 1, rock)} = (1 * 2/11 + 0 * 3/4 + 1 * 1/3 + 0 * 0) * 0,125 = \textcolor{blue}{0,064}$$

- ✓ It is **not normalised** by number of votes/rates present in an item
- ✓ All keywords are considered equally significant (The Beatles vs Serrat)

# HOW DO WE RECOMMEND?

- **Prediction** is the cosine of (profile , item)!
- Dot-product of normalised vectors = dot-product / length of two vectors
- Cosine goes from -1 to 1, 1 being perfect match!

# HOW DO WE RECOMMEND?

USER PROFILE

	user 1	user 2
rock	0,064	0,019
blues	0,053	0,114
jazz	0,000	0,000

ITEM

	song
rock	0,40
blues	0,20
jazz	0,80

ITEM PREDICTION

(user 1, song)	0,476
(user 2, song)	0,287

With **ITEM PREDICTIONS**,

we can order unknown items to build **RECOMMENDATIONS**

# HOW DO WE RECOMMEND?

USER PROFILE

	user 1	user 2
rock	0,064	0,019
blues	0,053	0,114
jazz	0,000	0,000

ITEM

	song
rock	0,40
blues	0,20
jazz	0,80

ITEM PREDICTION

O	A	B
1	(user 1, song)	0,476
2	(user 2, song)	0,287
=		

• fx ▾ SUMPRODUCT ▾ Table 1-1-1::B2:B4 ▾; Table 1-1-1-1::B2:B4 ▾ ÷ SUMPRODUCT ▾  
SQRT ▾ SUMPRODUCT ▾ Table 1-1-1::B2:B4 ▾; Table 1-1-1::B2:B4 ▾; SQRT ▾  
SUMPRODUCT ▾ Table 1-1-1-1::B2:B4 ▾; Table 1-1-1-1::B2:B4 ▾

# HOW DO WE RECOMMEND?

USER PROFILE

	Marc	Ivan
rock	0,05	0,90
blues	0,90	0,10

ITEM

	BBKing	Oasis
rock	0,10	0,90
blues	0,80	0,00

ITEM PREDICTION

(Marc, BBKing)	0,998
(Ivan, BBKing)	0,233

(Marc, Oasis)	0,055
(Ivan, Oasis)	0,994

# LAB 2

## PART 1!



# TODAY'S AGENDA

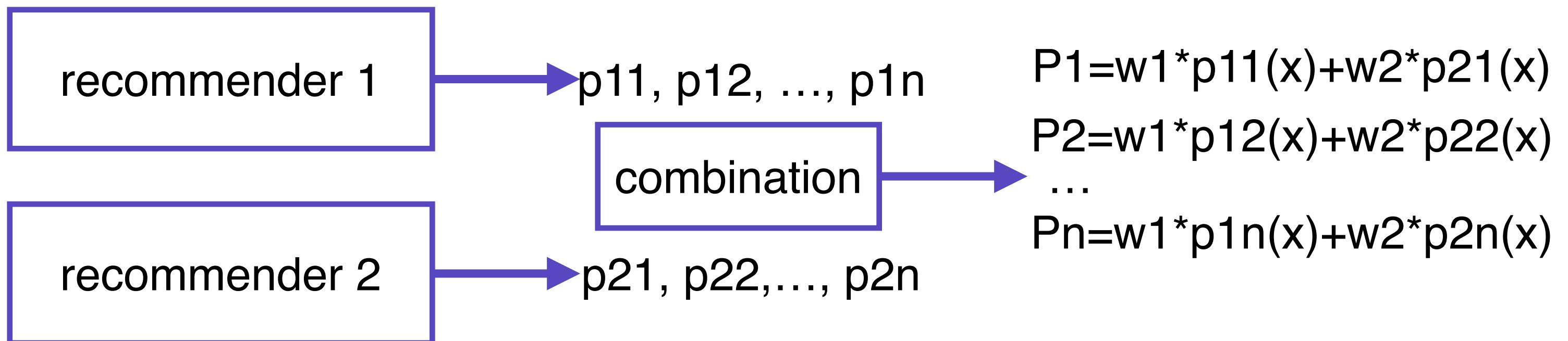
- Recapitulation...
- Session 5, Content-based Filtering & Hybrid Approaches
  - Content-based Filtering
  - Hybrid Approaches

# HYBRID APPROACHES

- Very common to combine different techniques in real industry problems
- **COLD START PROBLEM** is often solved by combining different methods
- Different options for specific domains, goals, problems, etc.

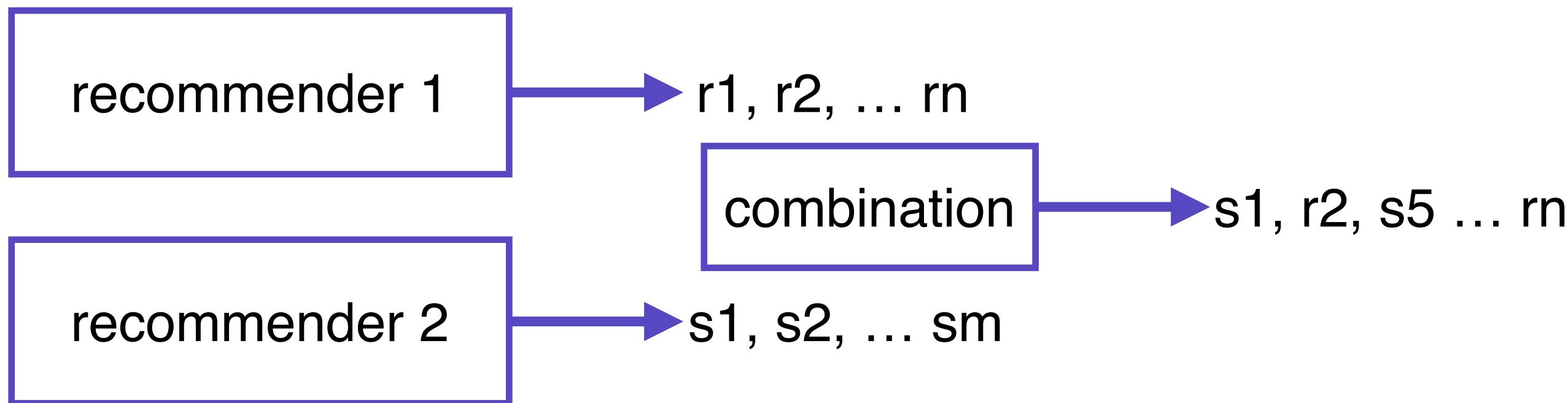
# HYBRID APPROACHES

- Weighting



# HYBRID APPROACHES

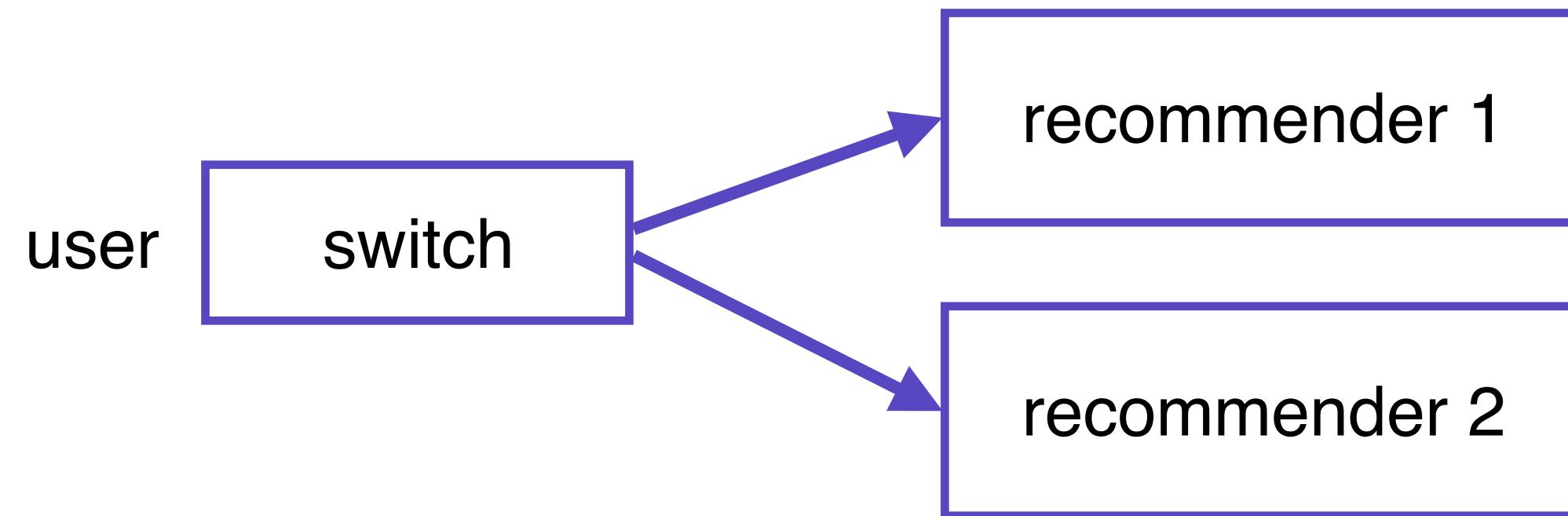
- Mixed



- Example: one gives you recommendations to build confidence. The other gives you recommendations to discover new items. Combine them!

# HYBRID APPROACHES

- Switching



# HYBRID APPROACHES

- Concatenating



# HYBRID APPROACHES

Be creative... whatever fits your domain and specific problem!

# IN-CLASS WORKSHOP

# IN-CLASS WORKSHOP

Share and discuss the **1 actionable idea** you came up with to **improve the Booking.com platform**.

The product change itself must make sense within our current subject of study.

For example, making a copy change for all users would not be an adequate example for this exercise, but developing a model to drive which copy to show to a user would be.

# IN-CLASS WORKSHOP

Please prepare for this by carefully looking through the **Booking.com** product and think about specific suggestions. Imagine it's a working situation and think about how you would go about implementing it and measuring the success.

## Some questions you could ask yourself when preparing:

- \* What is your hypothesis for each idea?
- \* What data would you look to validate your hypothesis and before running an A/B test?
- \* What metrics would you look at to evaluate the performance of the A/B test?
- \* How would you prioritize these ideas against each other and why?

LAB 2

PART 2 - HOMEWORK!

HYBRID GROUP CHALLENGE