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

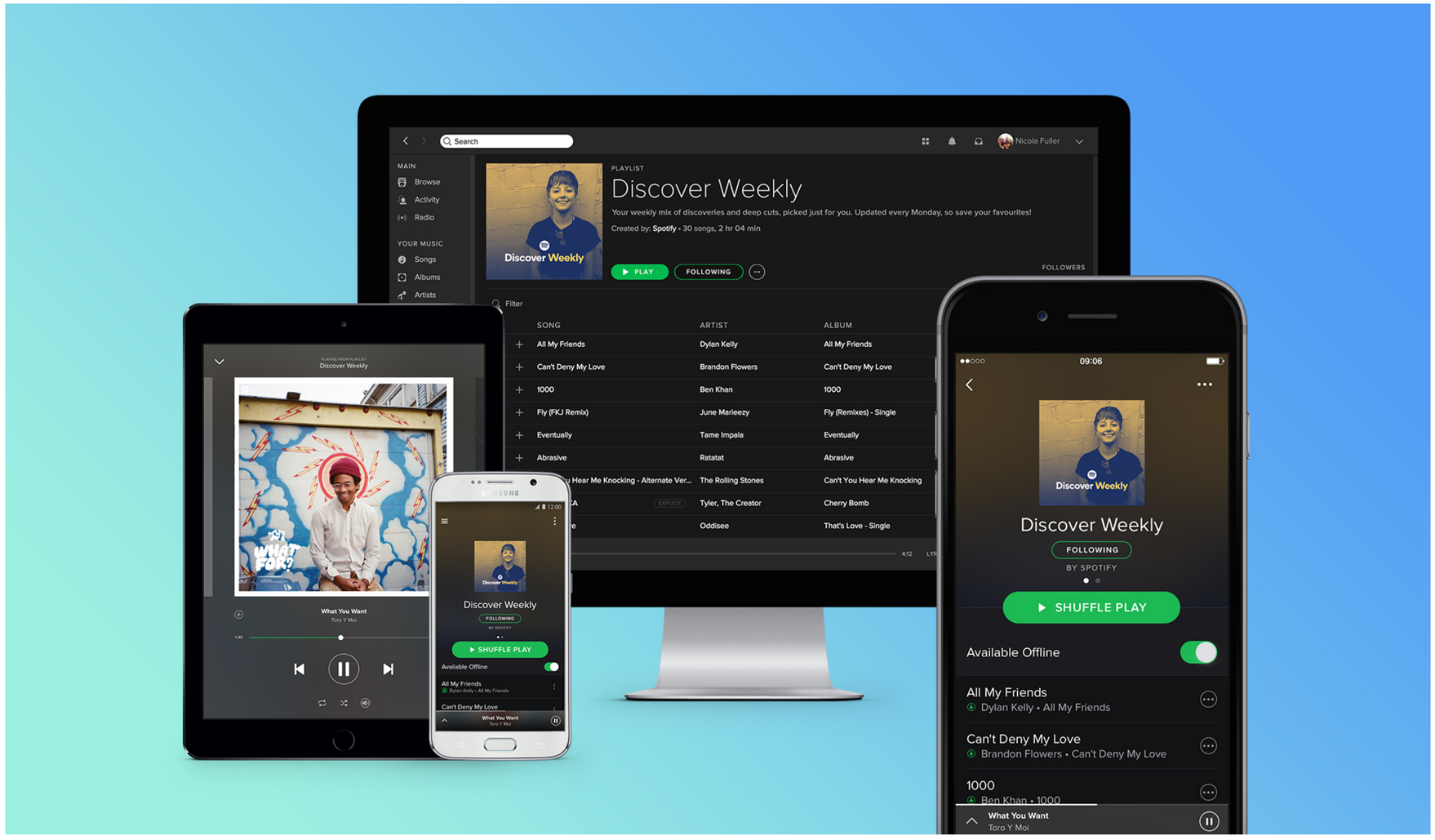
CONTENT-BASED FILTERING

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

SPOTIFY CASE

Spotify Discovery

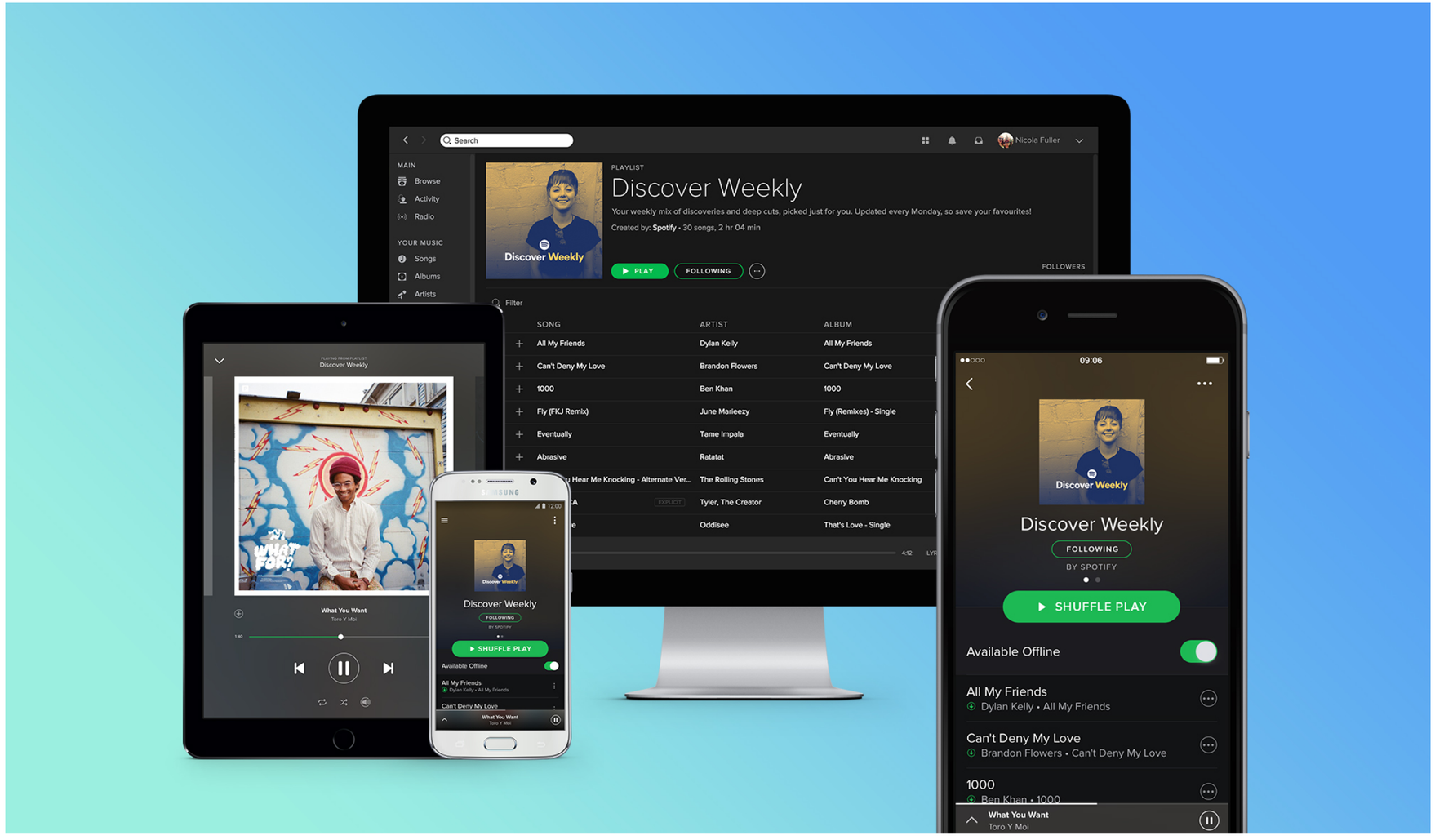
- Why they did it?
- Why it worked so well?
- What were their key findings?
- Downsides nowadays?

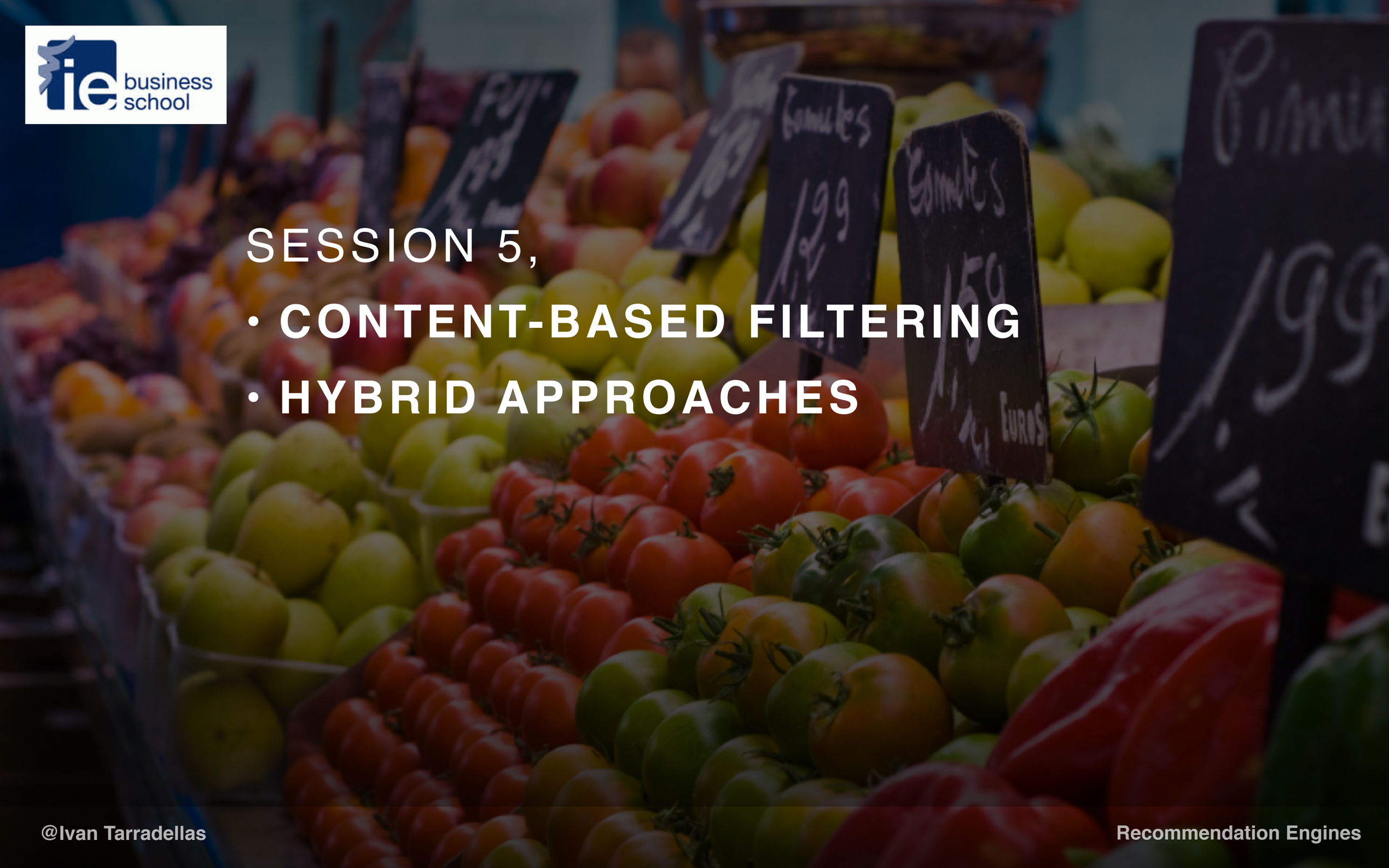


SPOTIFY CASE

Spotify Discovery

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

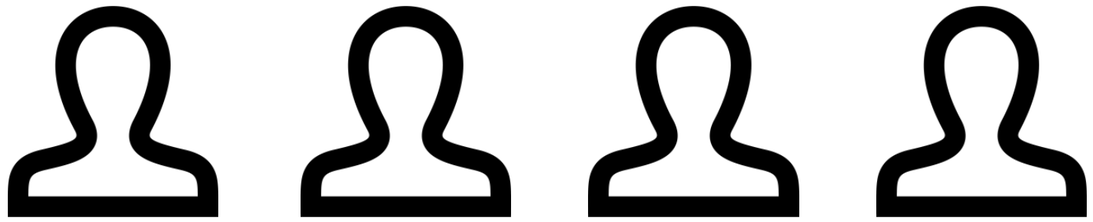




SESSION 5,

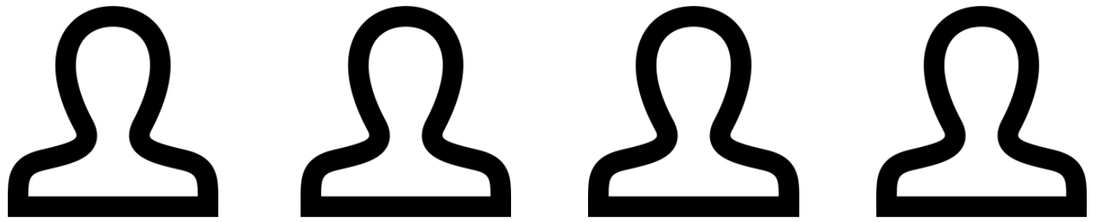
- **CONTENT-BASED FILTERING**
- **HYBRID APPROACHES**

IN COLLABORATIVE FILTERING...



opinions on items

IN CONTENT-BASED FILTERING...



content on items

IN CONTENT-BASED FILTERING...

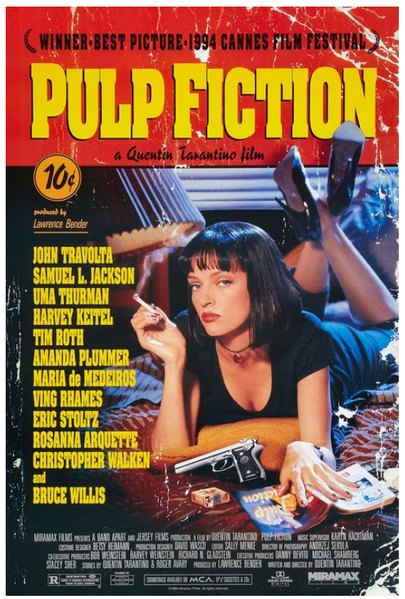
action
woody allen
english

woody allen
comedy
english

comedy
modern
worldwide

terror
woody allen
almodÓvar

tarantino
english



IN CONTENT-BASED FILTERING...

Voilà, we have a RECOMMENDER!

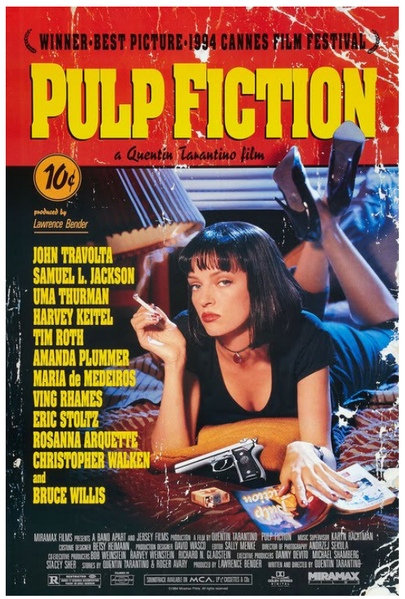
woody allen, english
action

woody allen
comedy
english

action
woody allen
english

english
action, woody allen

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

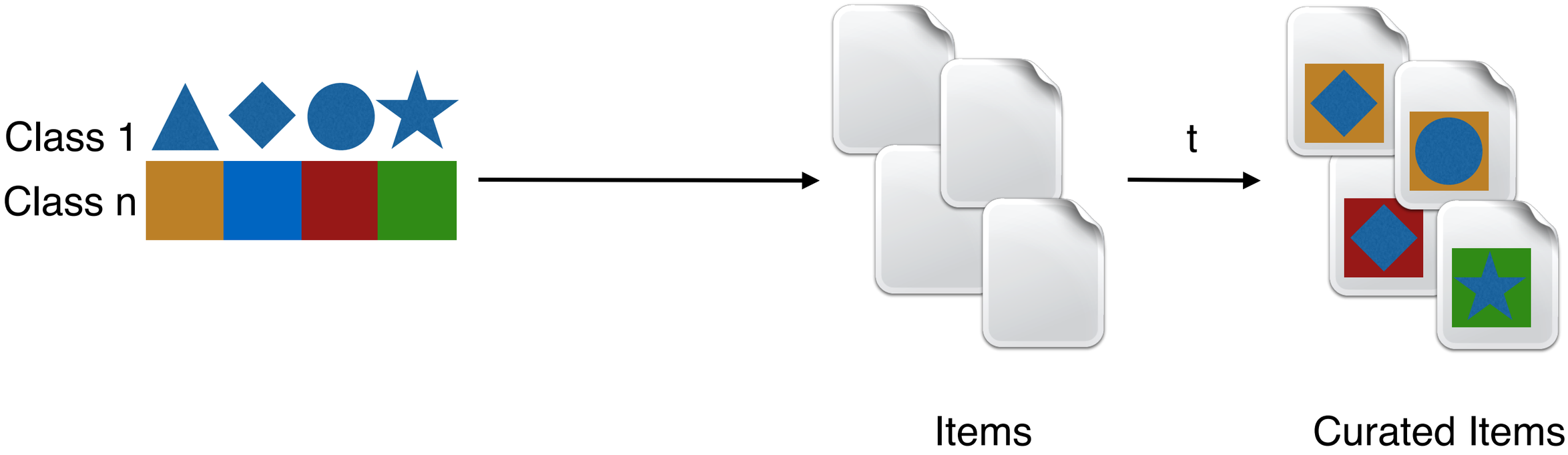
How to build the user preferences?

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

MODELING ITEMS

Define a taxonomy:

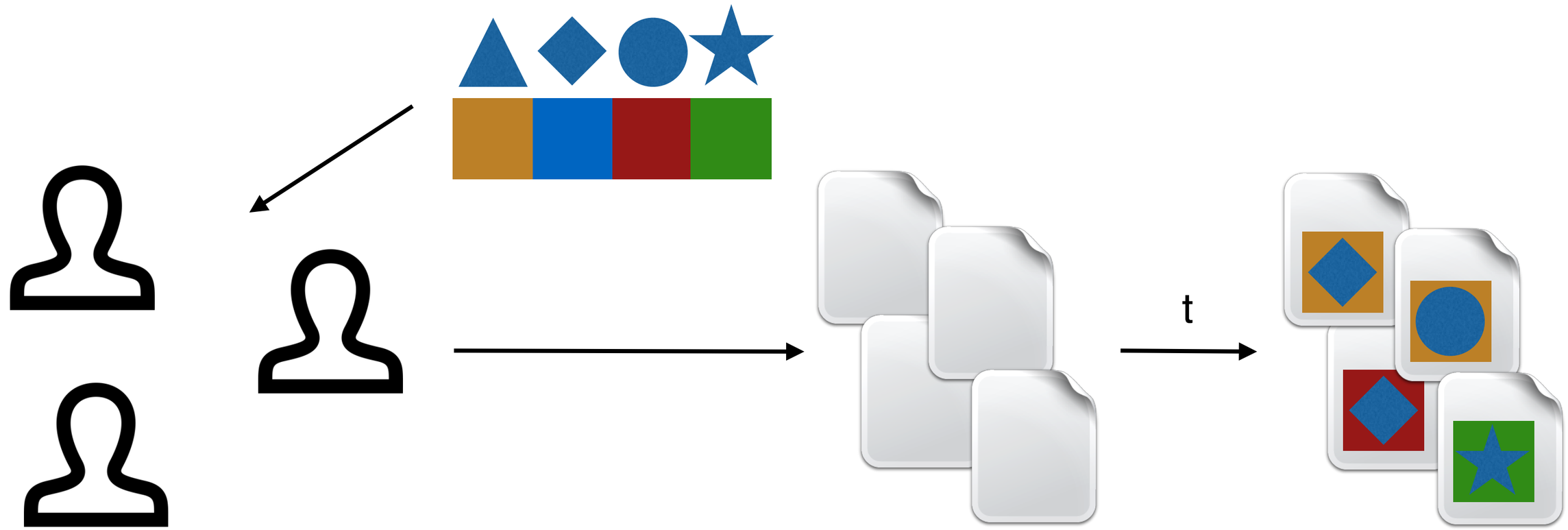
Choose carefully the set of **relevant** attributes to describe the items.
- Easy to understand for the user
- Useful for the recommendations



MODELING ITEMS

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

Taxonomy



Experts / Clients (Few)

New Items (Few)

Curated Items

MODELING ITEMS

Some examples: Booking.com

Property Type

Apartments	1400
Hotels	386
Guesthouses	266
Hostels	102
Bed and Breakfasts	58
Boats	8
Vacation Homes	6
Villas	3
Love Hotels	2

Facility

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

Room Facility

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

MODELING ITEMS

Some examples:

tripadvisor®

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)

Property | Neighborhood | Amenities & Features

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)

Distinctive Features

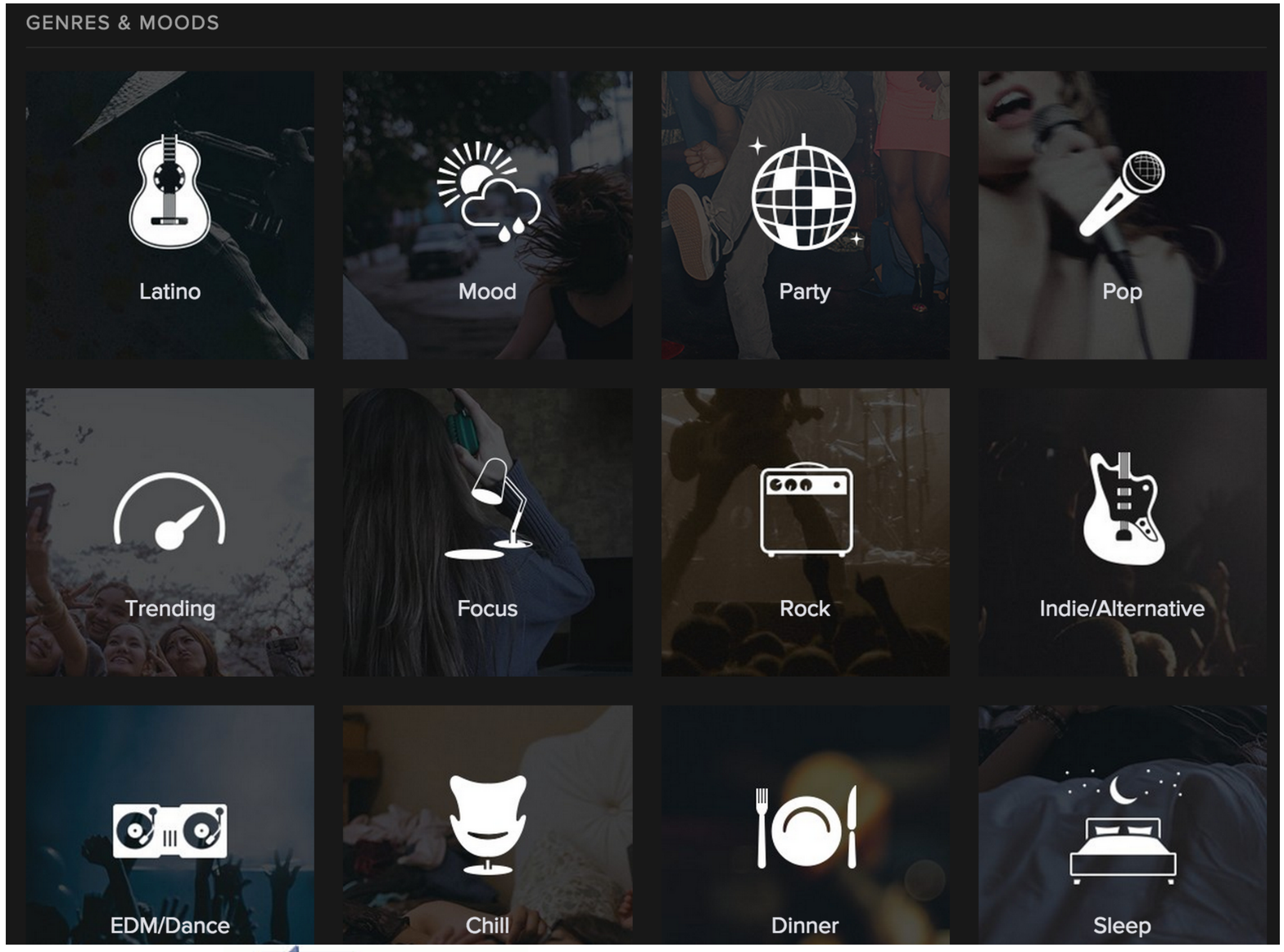
- Waterfront (39)
- Private Fishing Lake/River (1)
- Private Tennis Court (1)
- Boat Available (26)
- Ocean Views (133)
- Mountain Views (147)
- Water Views (56)
- Staffed Property (45)
- Housekeeping Included (142)
- Housekeeping Optional (129)

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

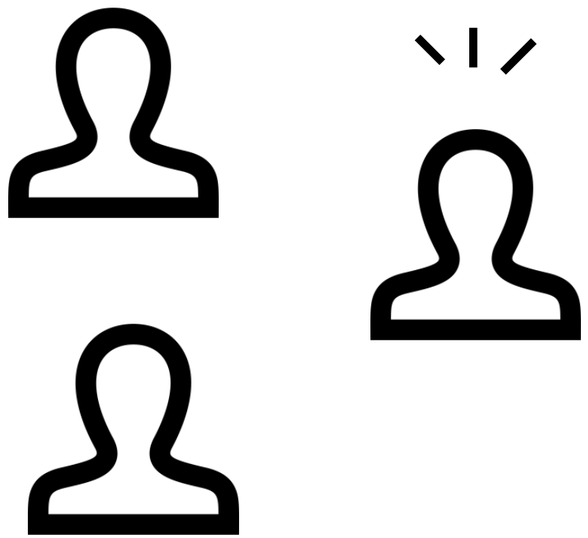
Question!

What are the downsides of manual modeling?

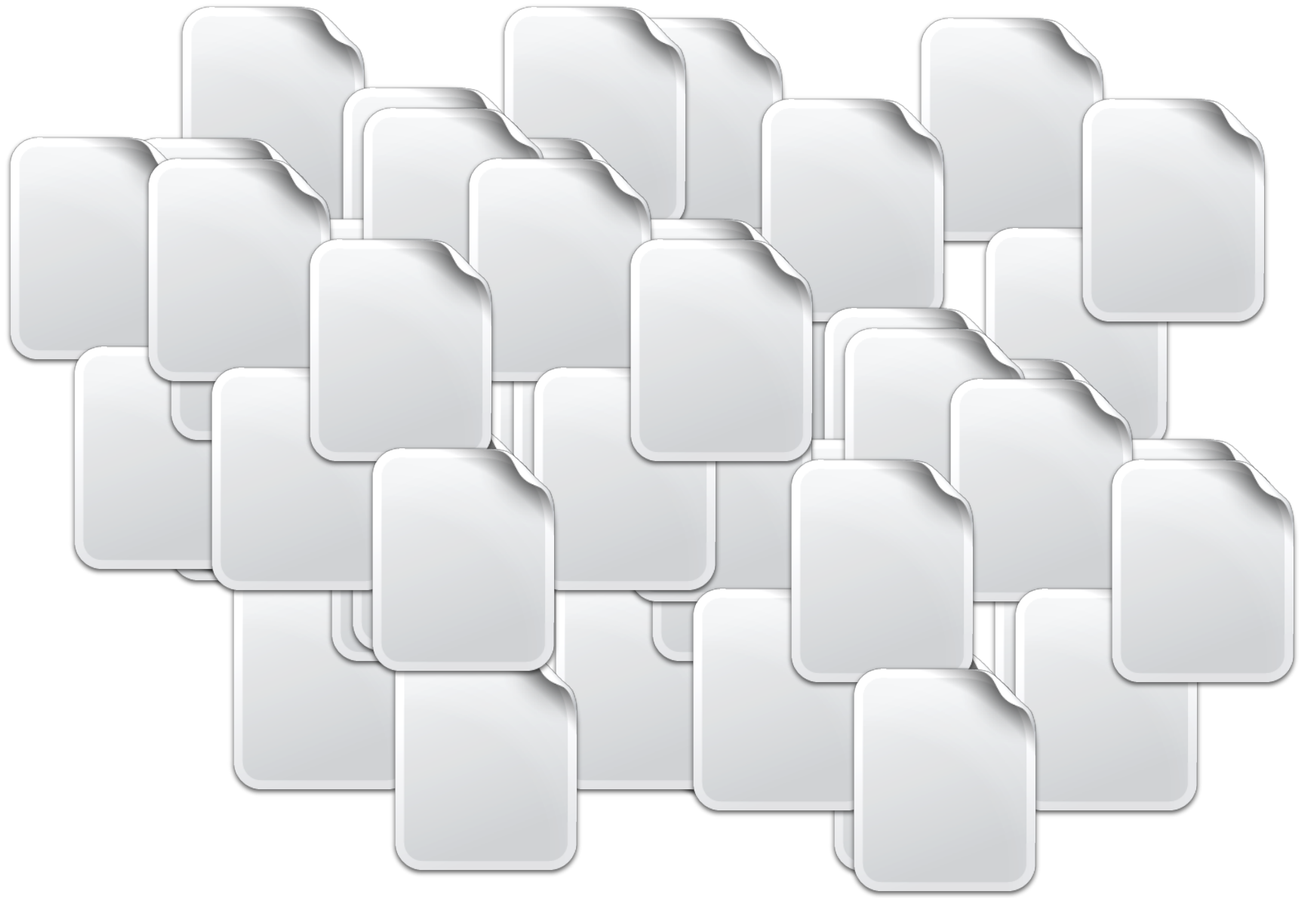
?

MODELING ITEMS

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

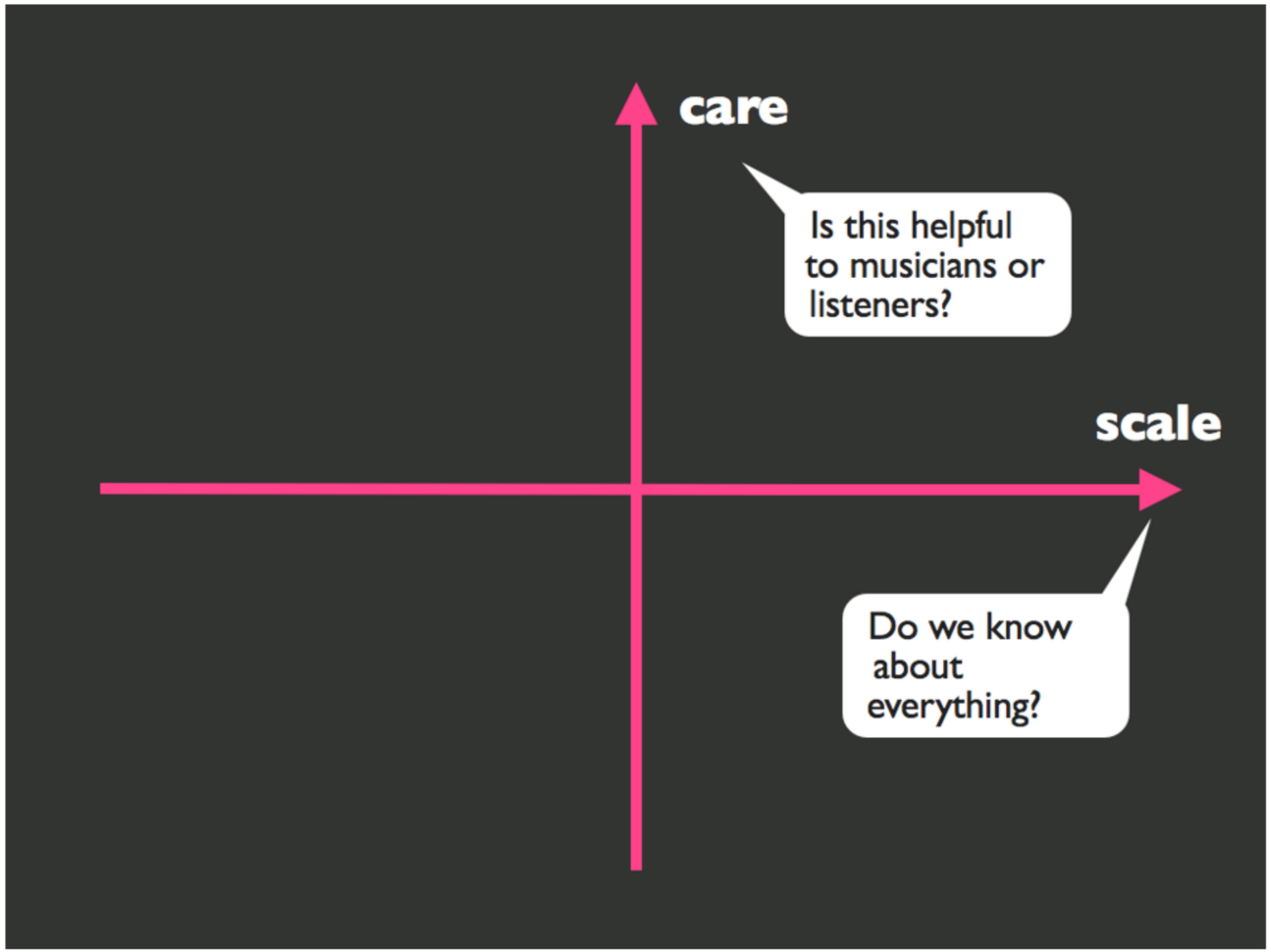


Experts (Few)



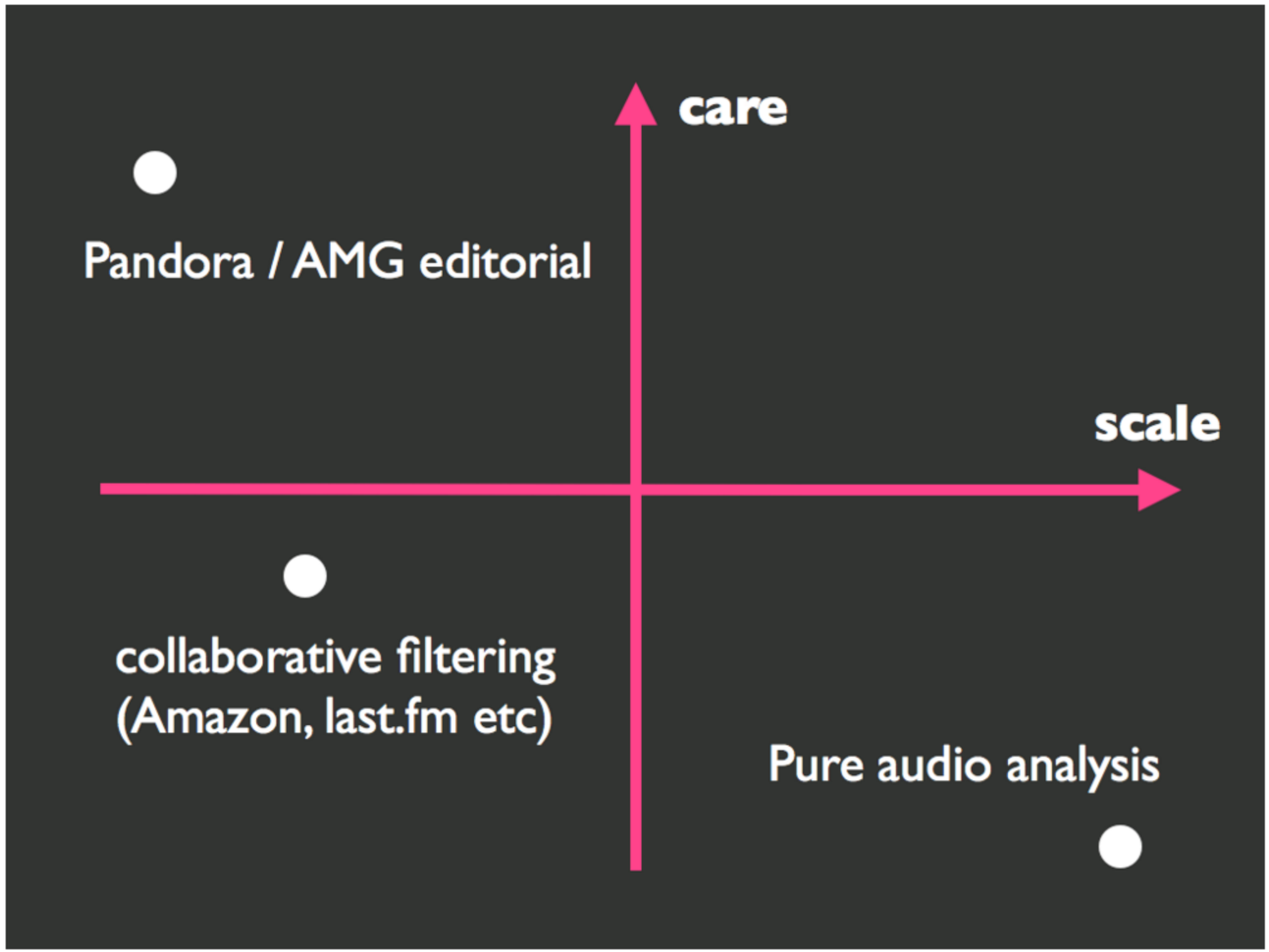
Too many new Items

ECHO NEST CASE



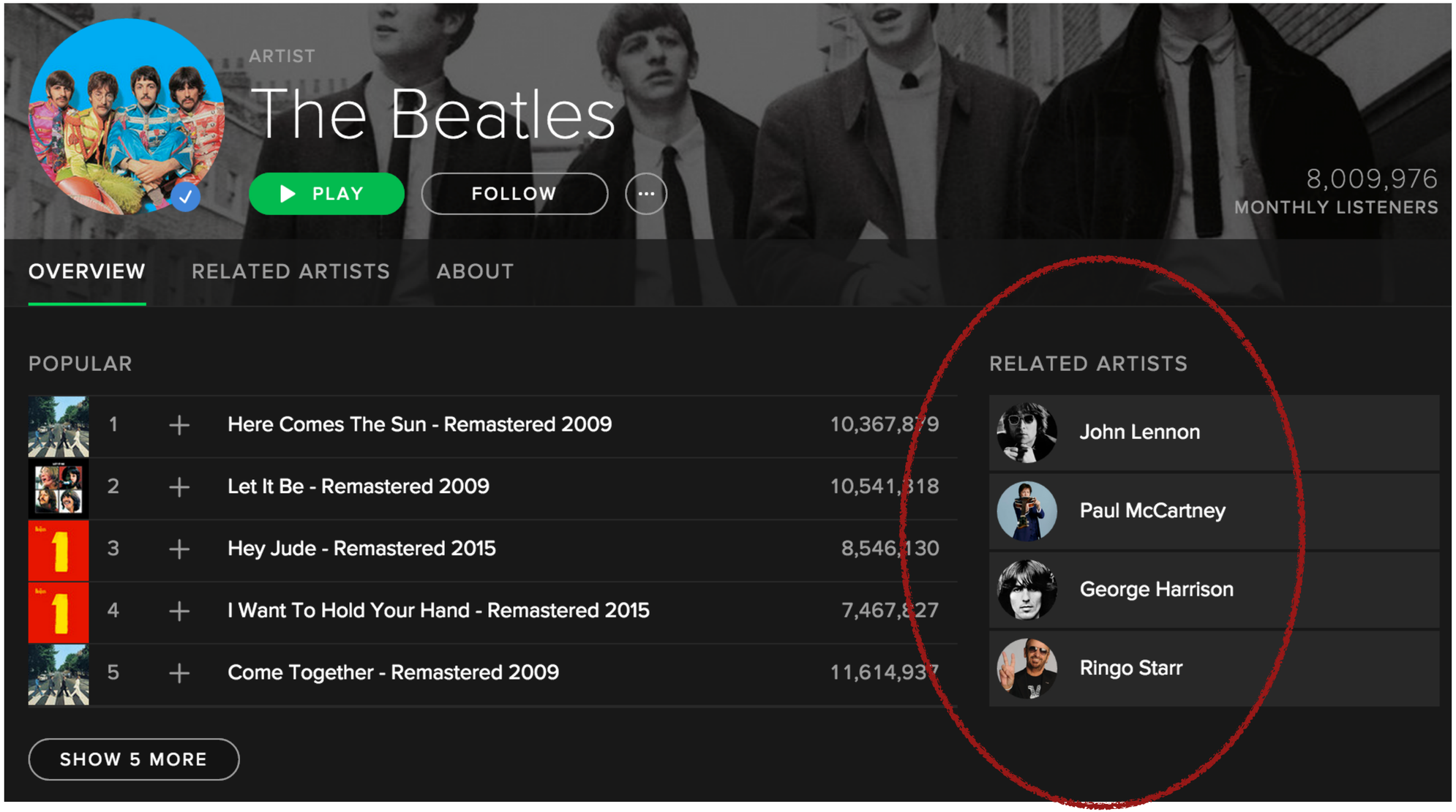
http://notes.variogr.am/post/37675885491/how-music-recommendation-works-and-doesnt-work

ECHO NEST CASE

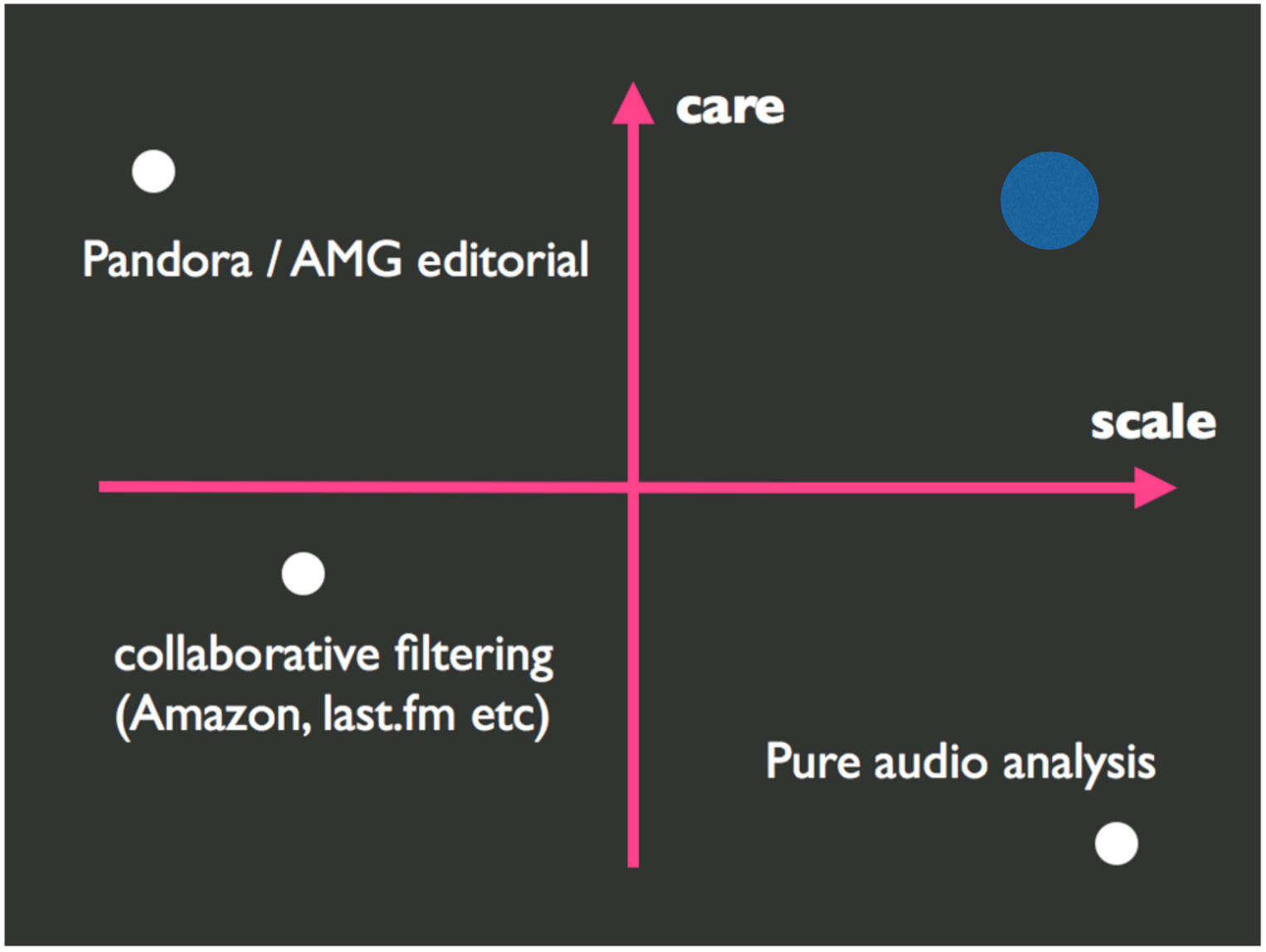


ECHO NEST CASE

CARE:



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?

Accoustic Analysis

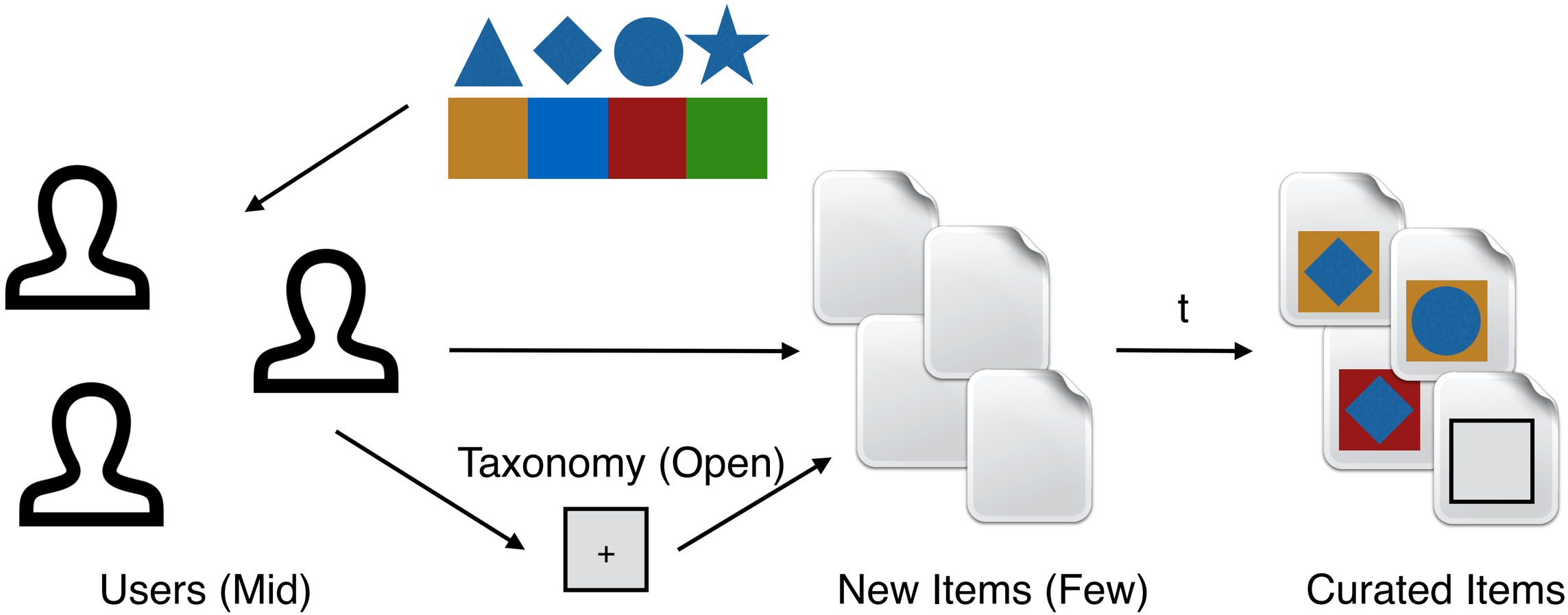
ECHO NEST CASE

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[1]
All music guide	Music editors & writers
Amazon	Purchase & browsing history
iTunes Genius	Purchase data, activity data from iTunes[2]
Echo Nest	Acoustic analysis, text analysis

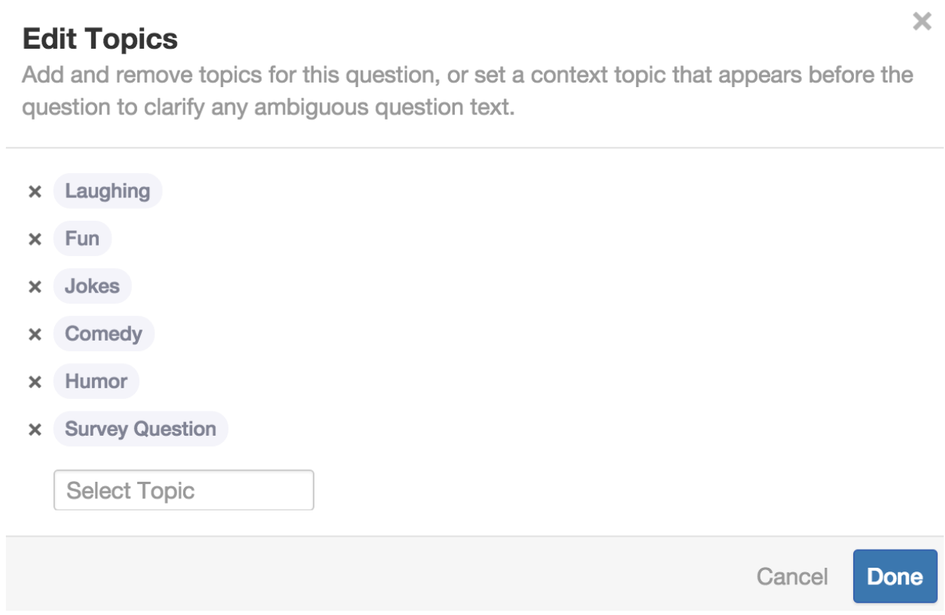
MODELING ITEMS

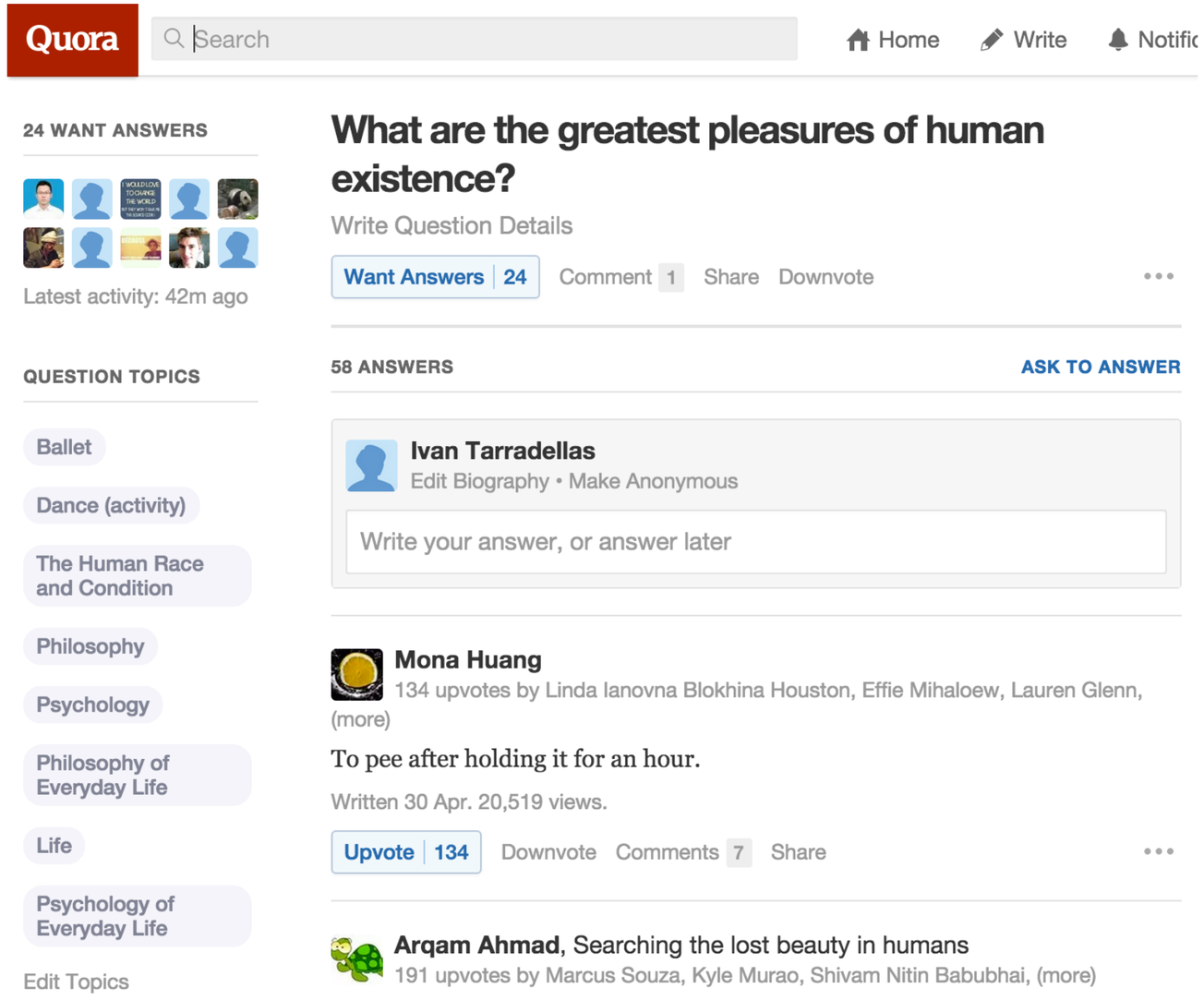
Collective classification of content created by the **end users.** Need **quality control**.



MODELING ITEMS

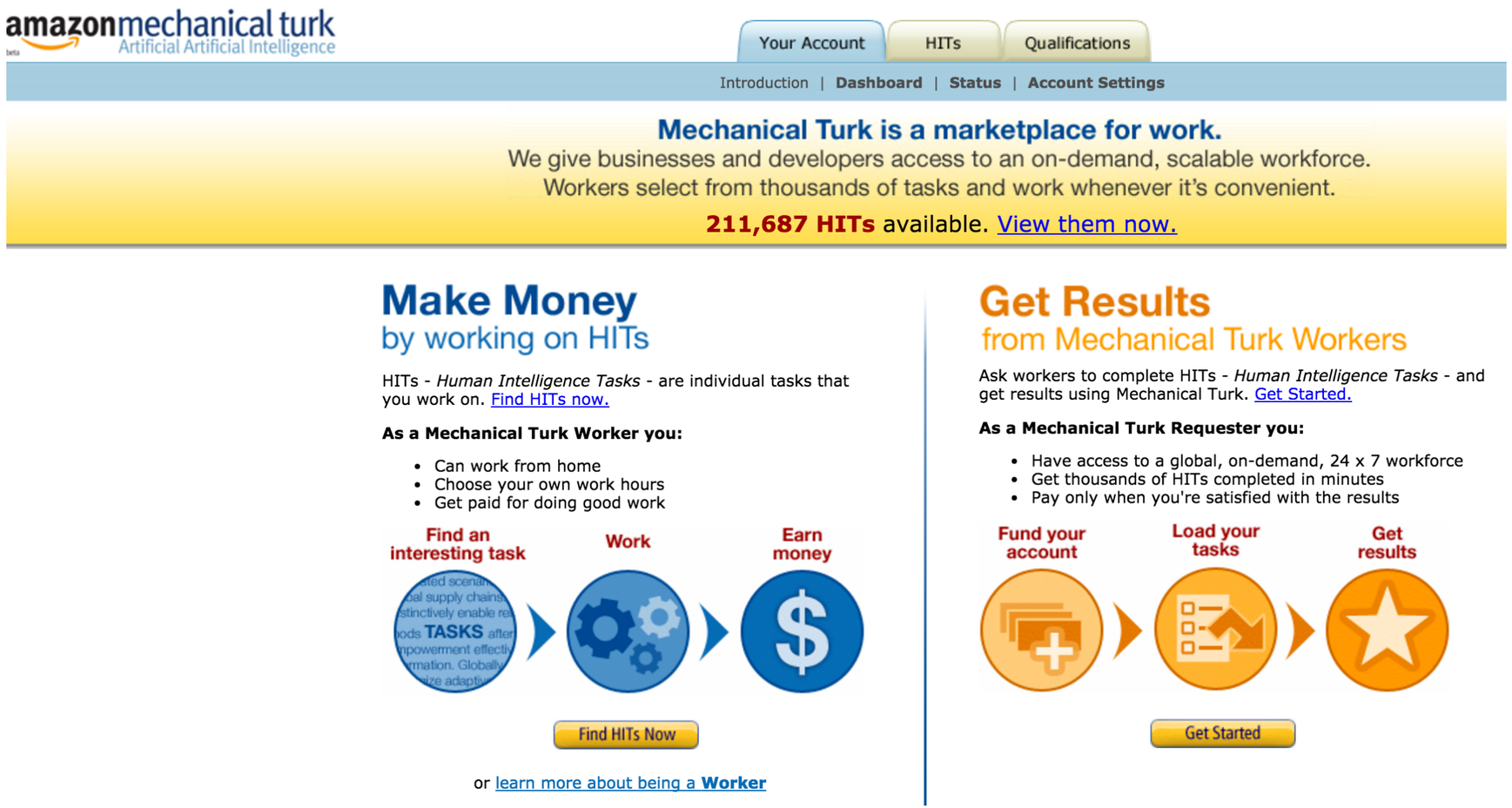
Some examples:





MODELING ITEMS

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

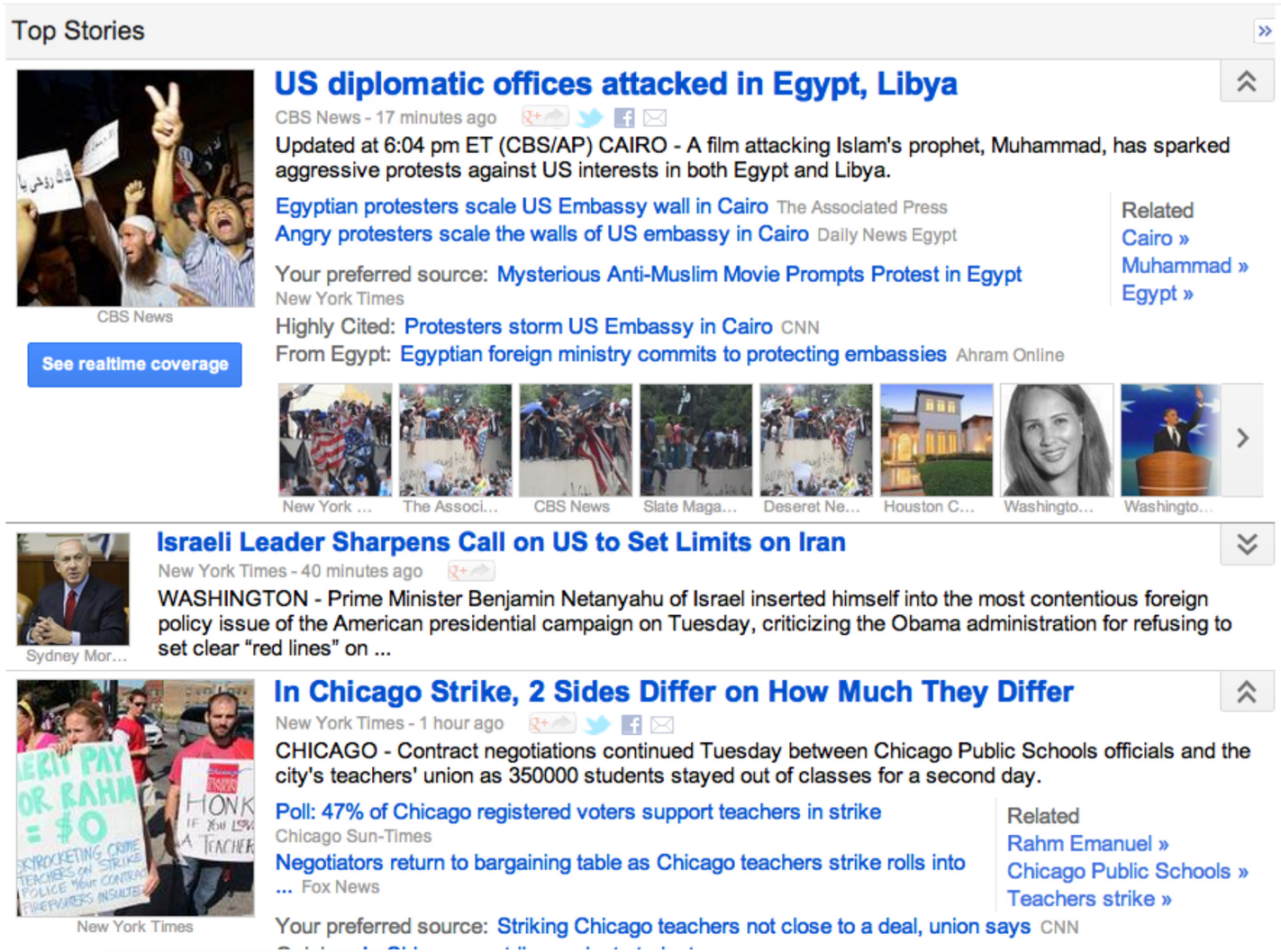


CATEGORIZING CONTENT

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 …



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

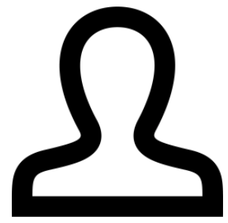
Building a **shirt recommender** with **explicit preferences**:

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)

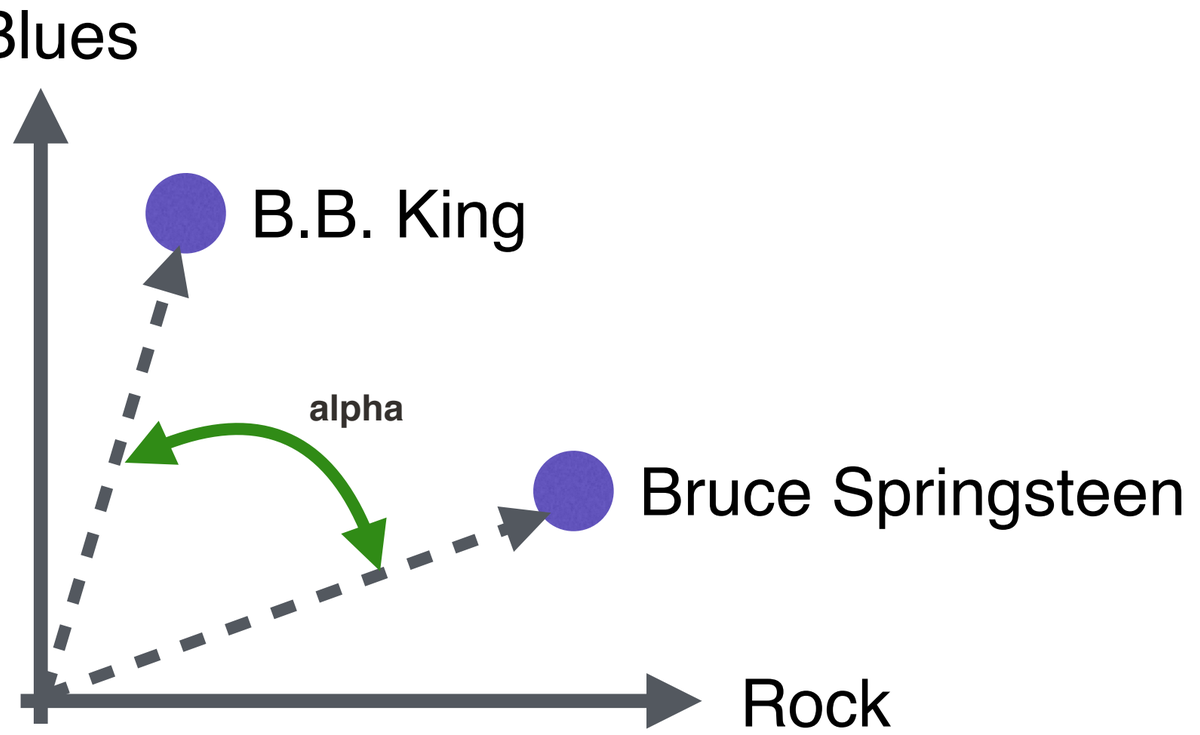
BUILDING THE USER PREFERENCES

A better approach

KEYWORD VECTORS FOR ITEMS

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

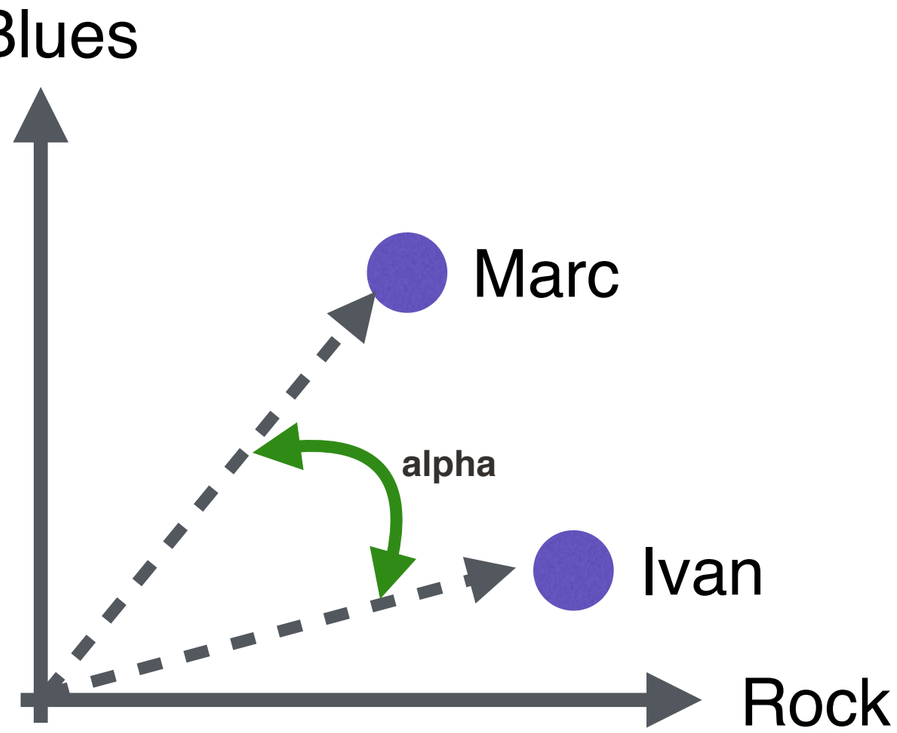
```
simil(BBKing, Bruce)=
cosine(alpha)
```



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!

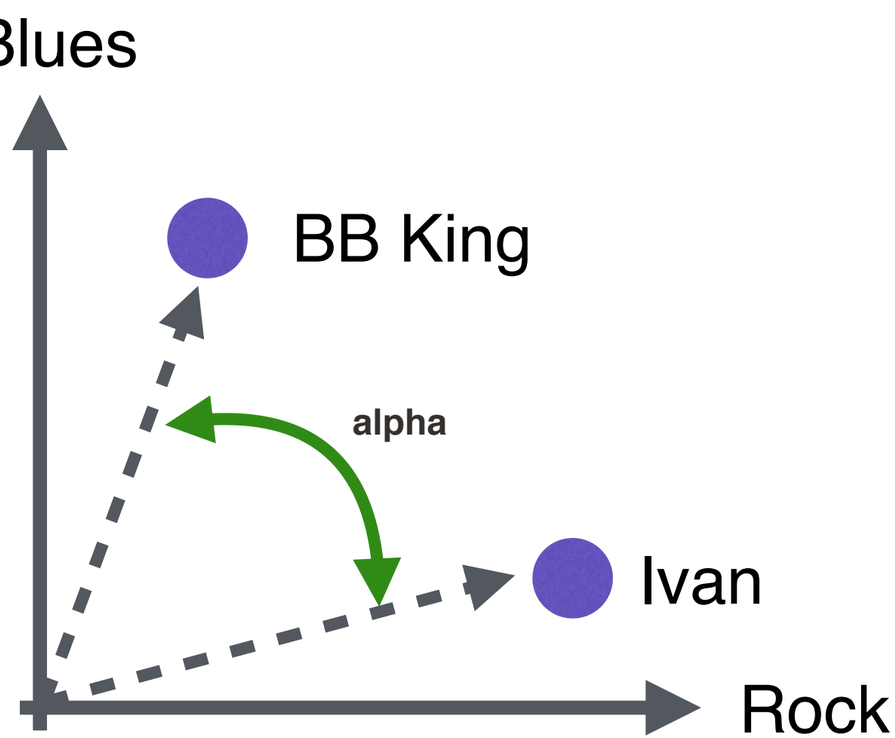
We have seen similarity between items based on attributes.
We have seen similarity between users based on preferences.

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)=
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**

ITEMS

	rock	blues	jazz
song 1	2	1	8
song 2	3	0	1
song 3	5	5	5
song 4	0	2	1

+

USERS

	user 1	user 2
song 1	1	-1
song 2	0	0
song 3	1	1
song 4	0	1

=

USER PROFILE

	user 1	user 2
rock	7	3
blues	6	6
jazz	?	-2

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**

ITEMS

	rock	blues	jazz
song 1	2	1	8
song 2	3	0	1
song 3	5	5	5
song 4	0	2	1

+

USERS

	user 1	user 2
song 1	1	-1
song 2	0	0
song 3	1	1
song 4	0	1

=

USER PROFILE

	user 1	user 2
rock	7	3
blues	6	6
jazz	?	-2

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**

ITEMS

	rock	blues	jazz
song 1	2	1	8
song 2	3	0	1
song 3	5	5	5
song 4	0	2	1

+

USERS

	user 1	user 2
song 1	1	-1
song 2	0	0
song 3	1	1
song 4	0	1

=

USER PROFILE

	user 1	user 2
rock	7	3
blues	6	6
jazz	**13**	-2

ex: (user 1, jazz) = 1*8 + 0*1 + 1*5 + 0*1 = **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**

ITEMS

	rock	blues	jazz
song 1	2/11	1/11	8/11
song 2	3/4	0	1/4
song 3	1/3	1/3	1/3
song 4	0	2/3	1/3

+

USERS

	user 1	user 2
song 1	1	-1
song 2	0	0
song 3	1	1
song 4	0	1

=

USER PROFILE

	user 1	user 2
rock	**0,515**	0,152
blues	0,424	0,909
jazz	?	-0,061

ex: (user 1, rock) = 1*2/11 + 0*3/4 + 1*1/3 + 0*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**

ITEMS

	rock	blues	jazz
song 1	2/11	1/11	8/11
song 2	3/4	0	1/4
song 3	1/3	1/3	1/3
song 4	0	2/3	1/3

+

USERS

	user 1	user 2
song 1	1	-1
song 2	0	0
song 3	1	1
song 4	0	1

=

USER PROFILE

	user 1	user 2
rock	**0,515**	0,152
blues	0,424	0,909
jazz	?	-0,061

ex: (user 1, rock) = 1*2/11 + 0*3/4 + 1*1/3 + 0*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**

ITEMS

	rock	blues	jazz
song 1	2/11	1/11	8/11
song 2	3/4	0	1/4
song 3	1/3	1/3	1/3
song 4	0	2/3	1/3

+

USERS

	user 1	user 2
song 1	1	-1
song 2	0	0
song 3	1	1
song 4	0	1

=

USER PROFILE

	user 1	user 2
rock	0,515	0,152
blues	0,424	0,909
jazz	**1,061**	-0,061

ex: (user 1, jazz) = 1*8/11 + 0*1/4 + 1*1/3 + 0*1/3 = **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**

vs

If a user likes **Serrat**

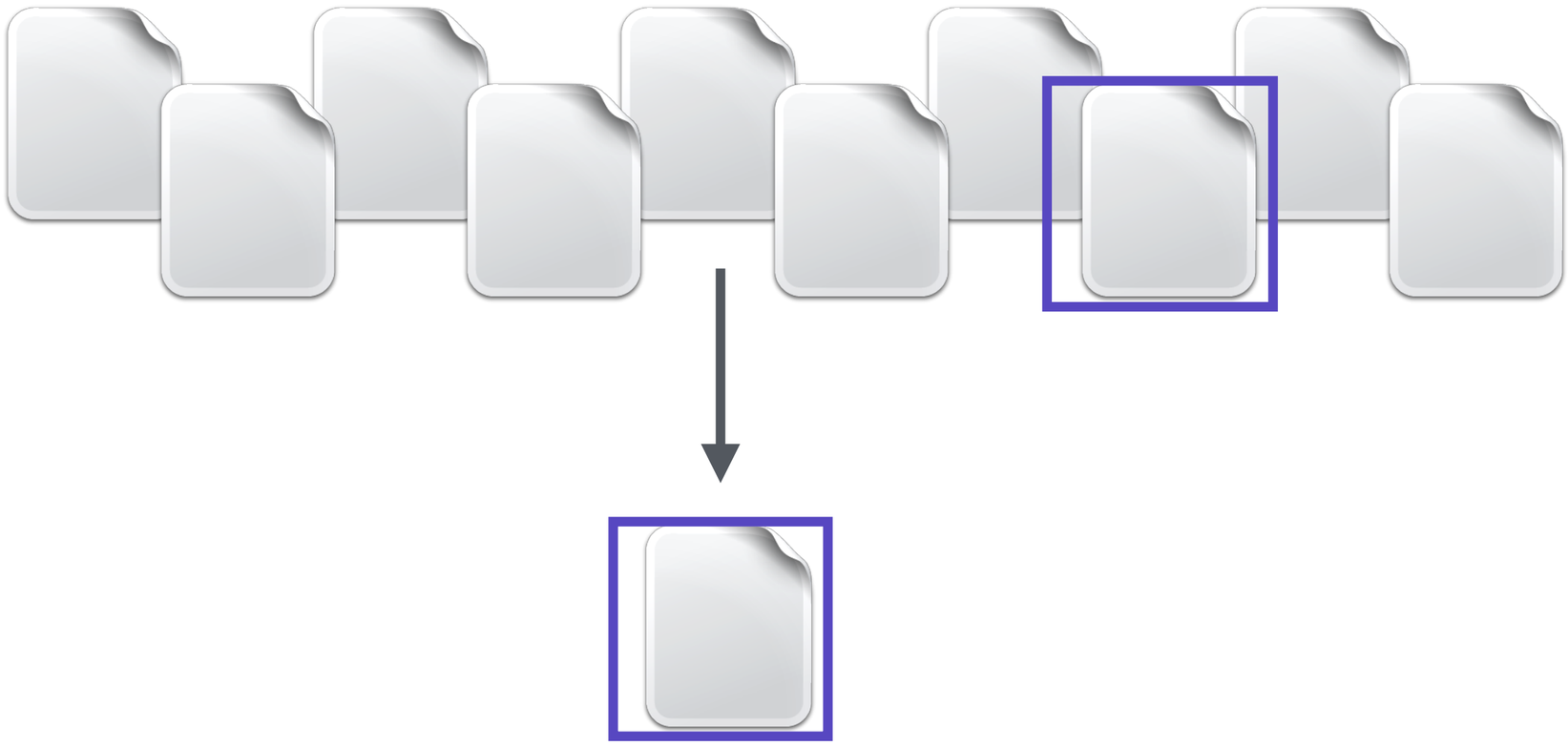
thinking...

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)

ITEMS - CORPUS

	rock	blues	jazz
song 1	2/11	1/11	8/11
song 2	3/4	0	1/4
song 3	1/3	1/3	1/3
song 4	0	2/3	1/3
DF	3	3	4
IDF	0,125	0,125	0,000

+

USERS

	user 1	user 2
song 1	1	-1
song 2	0	0
song 3	1	1
song 4	0	1

=

USER PROFILE

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

← LOG(COUNT(D2:D5)÷COUNTIF(D2:D5;">0")) -> LOG (4/4)

ex: (user 1, rock) =(1*2/11 + 0*3/4 + 1*1/3 + 0*0) * 0,125 = **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

	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…**

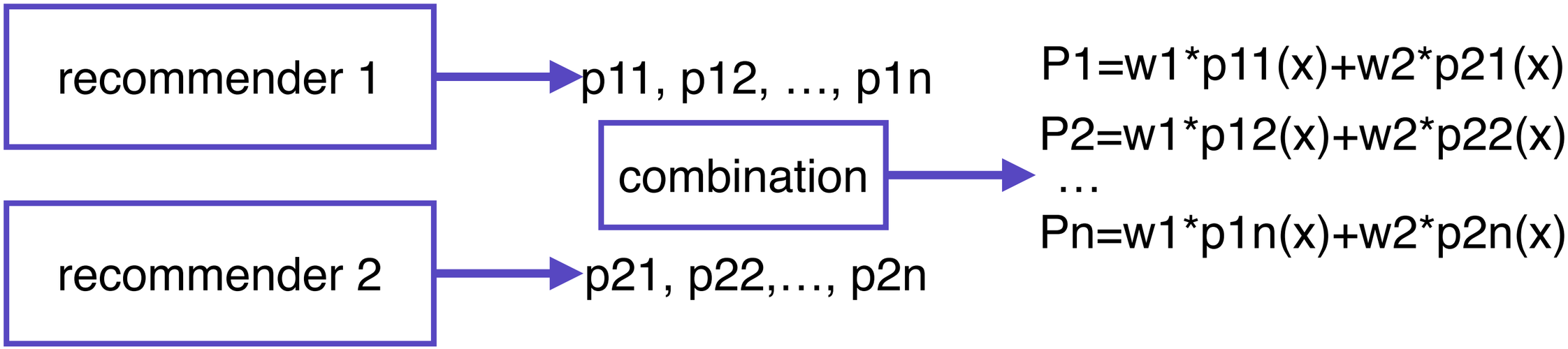
- **Session 5, Content-based Filtering & Hybrid Approaches**
  - Content-based Filtering
  - **Hybrid Approaches**

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.

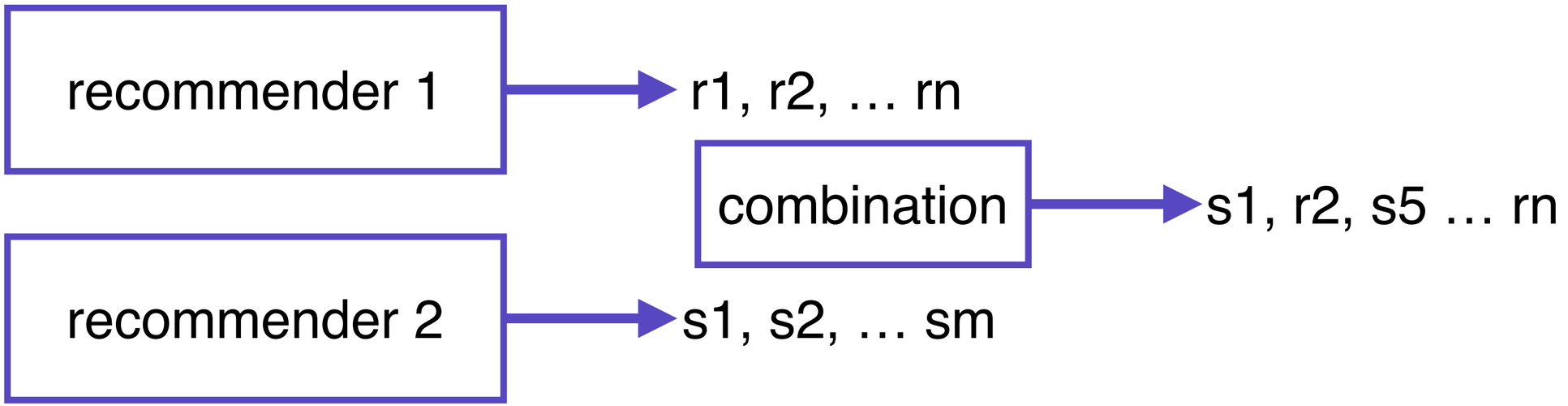
HYBRID APPROACHES

- Weighting



HYBRID APPROACHES

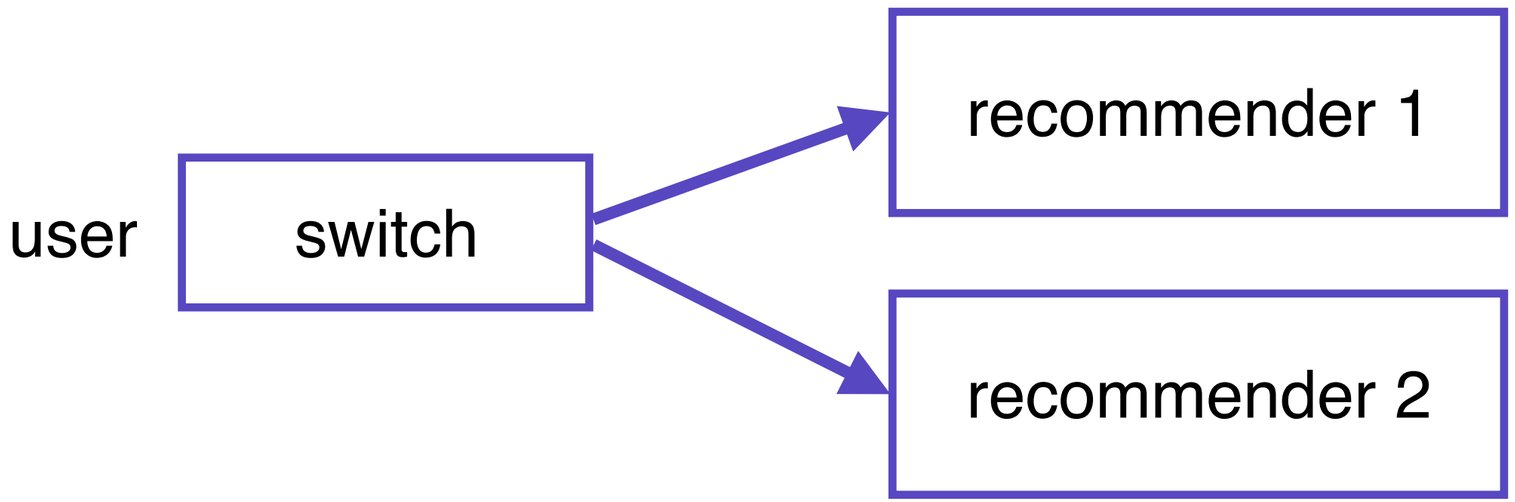
- Mixed



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

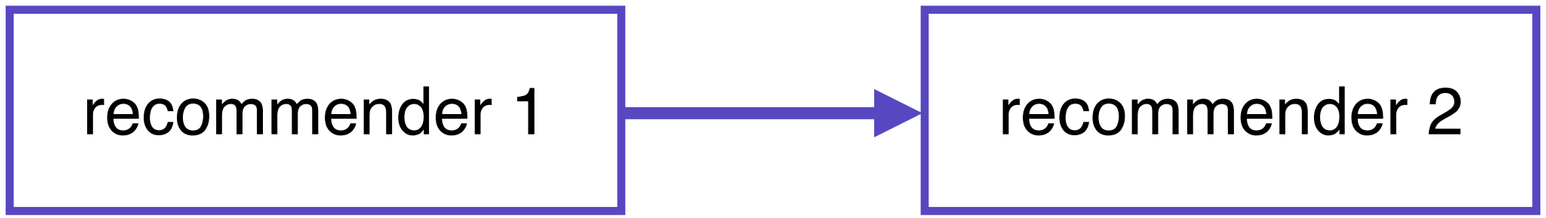
HYBRID APPROACHES

- 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