Decoding Passion: Discovering User Interest Profiles with Hybrid AI

AI Hub @ Allegro

Anna Mikołajczyk — Data Scientist
Anna Pintara — Data Scientist

Discovering user interests profiles

Our hobbies

- belly dance
- healthy lifestyle
- board games

- motherhood
- skiing
- contemporary dance

Hobby as a gateway to inspiration

Moving from individual product data to a hobby insight.

BEHAVIORAL TRACE

Item #329487

Running Shoes

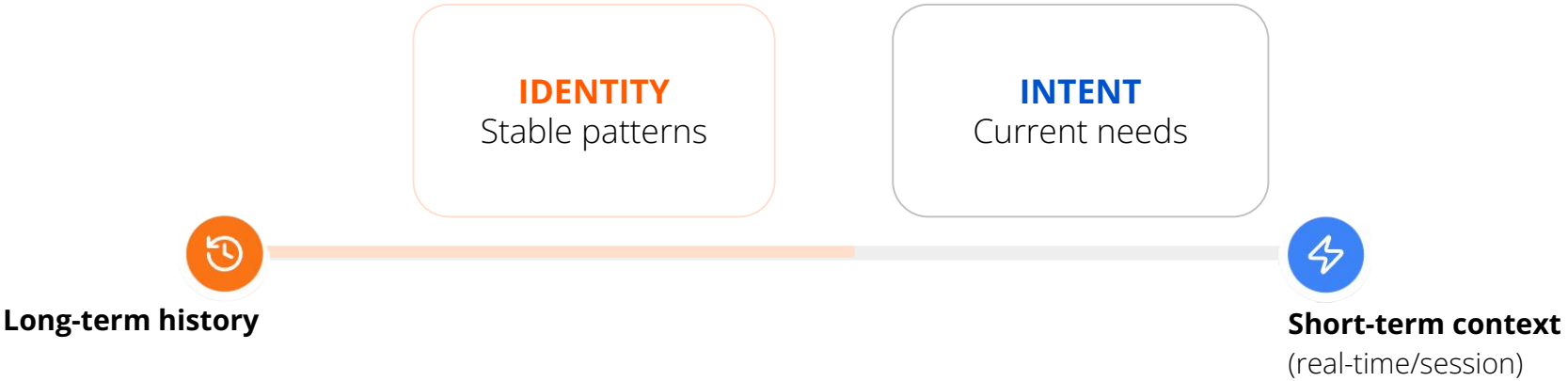
“User bought shoes.”

HOBBY INSIGHT

RUNNING ENTHUSIAST

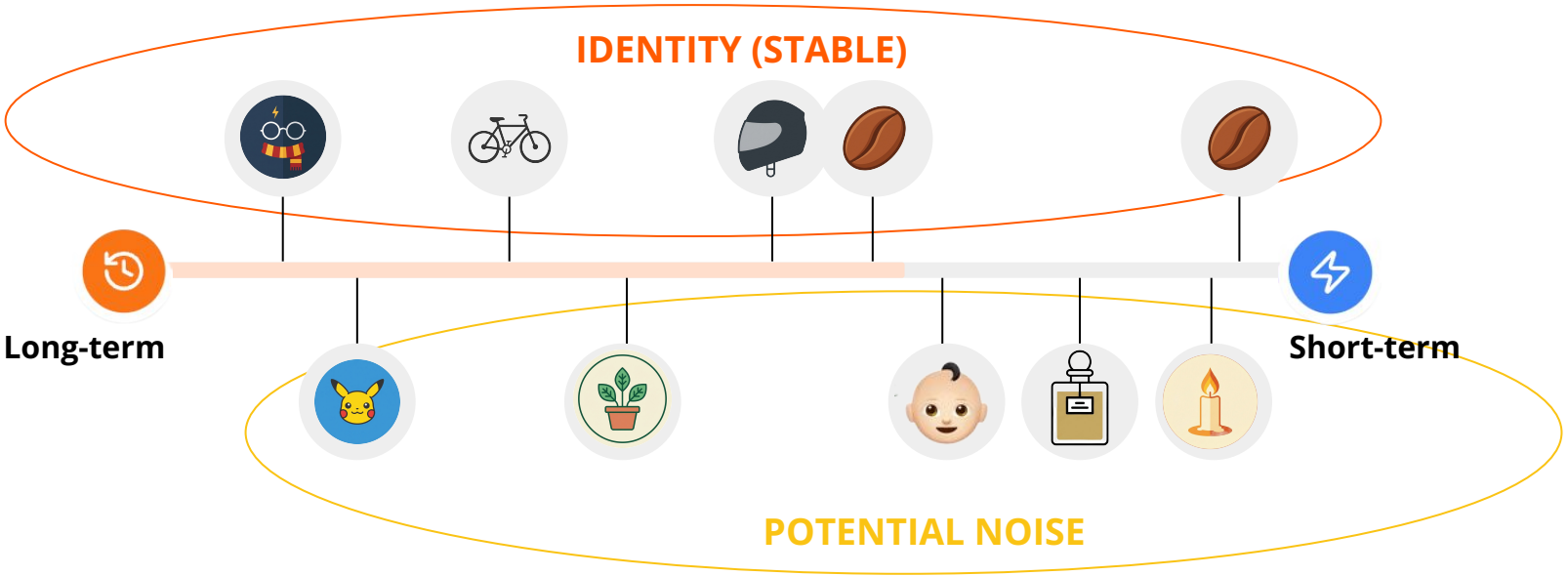
The Horizon of Understanding

The longer we know you the more we understand users **identity**.



The horizon of understanding (illustration)

Distinguishing stable identity from transactional noise across 730 days.



The Definition of Hobby

Lifestyle

PARENTING, GARDENING

Activity

MUSIC, FISHING

Profession

HAIRDRESSER, INFLUENCER

Lifestage

RENOVATING, FESTIVALS

Style

RETRO, MINIMALIST

Cultural affinity

POTTERHEADS, LEGO

The approach: hybrid pipeline

STEP 01

HOBBY GENERATION

Assigning topics to millions of products using textual data.

TECHNOLOGY STACK

Generative AI

STEP 02

HOBBIES STRUCTURATION

Organizing raw tags into coherent, hierarchical tree-based structure.

TECHNOLOGY STACK

Clustering
+ Sentence Transformers
+ Generative AI

STEP 03

USERS PROFILING

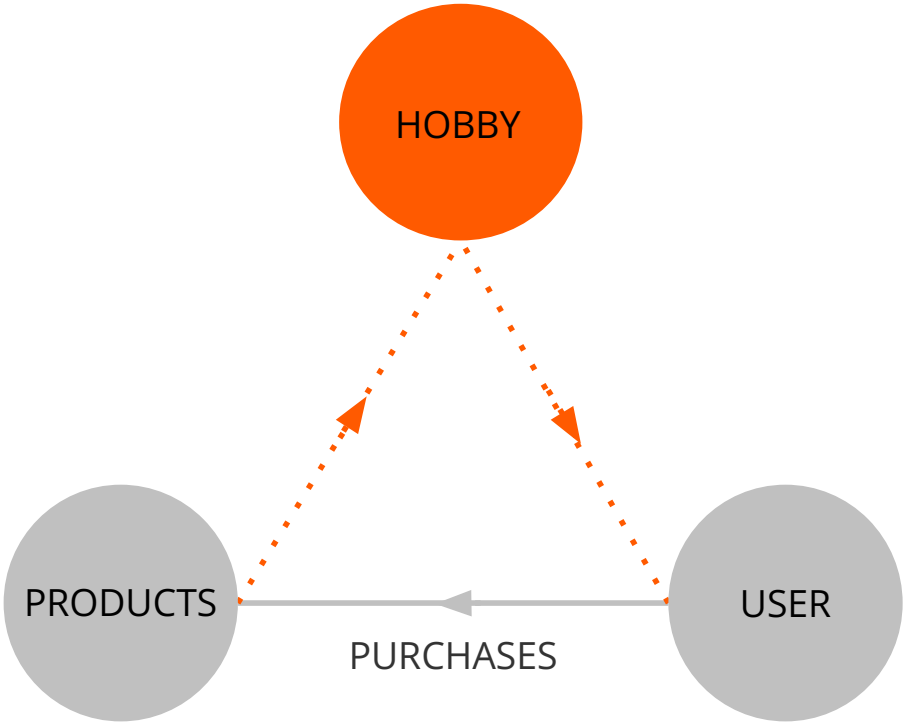
Aggregating history to find the strongest interests using behavioral dynamics.

TECHNOLOGY STACK

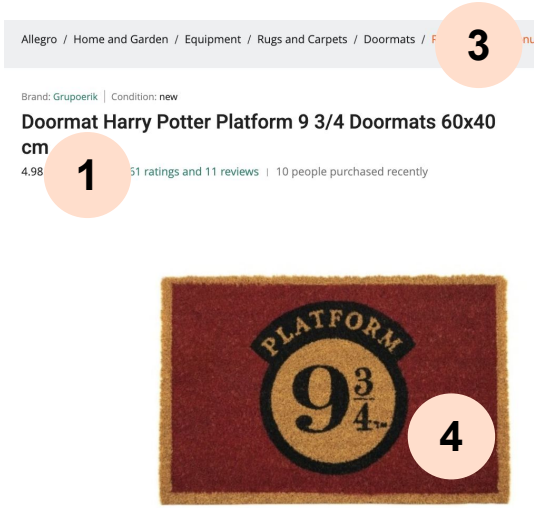
RFM Scoring

WHAT ARE THE HOBBY SIGNALS?

How do we get hobbies?



Text is all we need



1. Name
2. Description
3. Category path
4. Images
5. Parameters

Hobby generation

product information

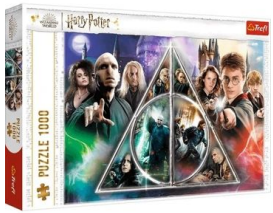
<PROMPT1>
ASSIGN
HOBBY
</PROMPT1>

Assign hobby LLM

> harry potter

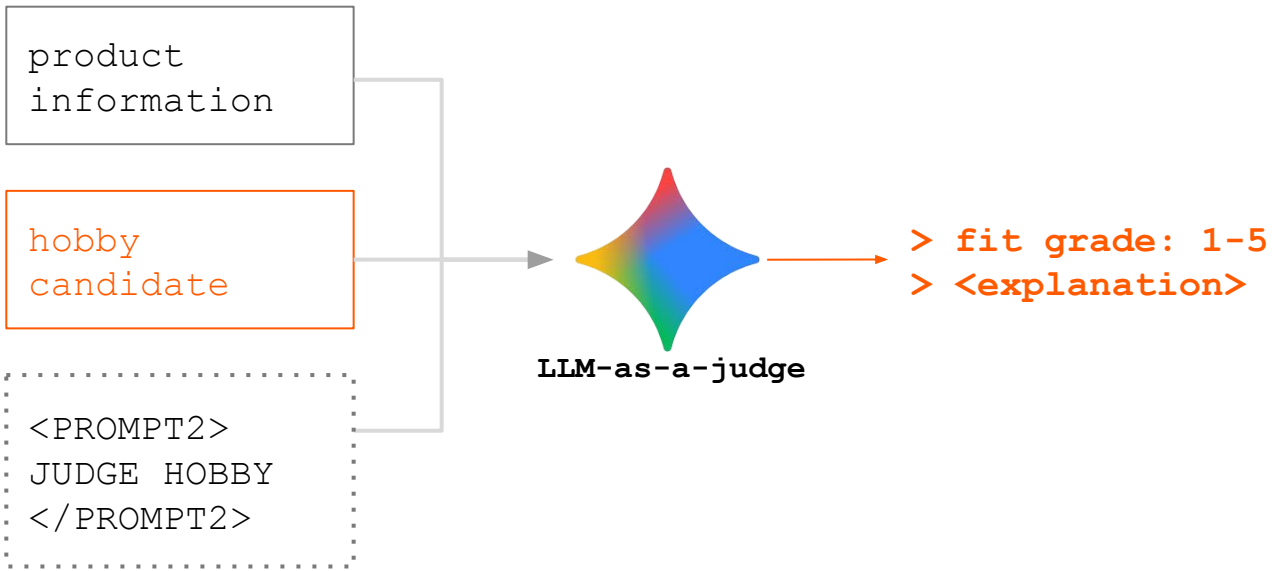
We use Large Language Model to interpret product textual info. Each item receives a **hobby/topic tag**.

Can we trust an LLM?

	Name	Hobby
	*Puzzle 1000 Deathly Hallows Trefl*	**harry potter**
	*Born2Be Black Women's Boots 38 Transitional Rubber*	**festival wear**

	Name	Hobby
	Anti-smog mask size S	**honey extraction**
	SILOO ball medal hanger	**achievements**

Hobby judge to the rescue



Hobby judge verdict

Name	Hobby	Fit grade	Explanation
	honey extraction	1	No clear connection.
	achievements	1	Abstract concept, not a hobby.

How many hobbies per product?

	Name	Hobby
	*Puzzle 1000 Deathly Hallows Trefl*	harry potter problem solving family activity fantasy fandom

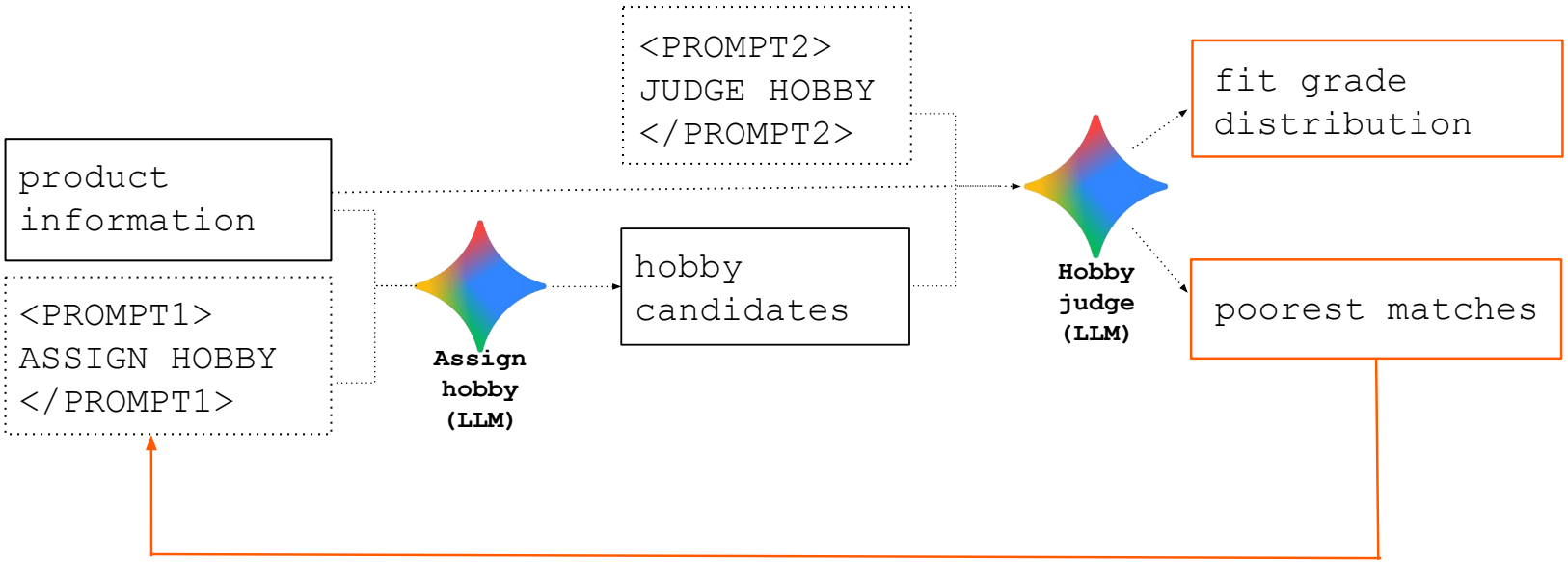
Which input data is relevant?

	Name	Category Path	Description
	*Born2Be Black Women's Boots 38 Transitional Rubber*	*Allegro* *> Fashion* *> Clothes, Footwear, Accessories* *> Footwear* *> Women’s* *> Wellingtons*	*Born2Be Czarne Kalosze damskie 38 Przejściowy Guma* *Kod produktu: 398643* *Kolor: czarny* *Materiał: Guma* *Nosek buta: Okrągły* *Ocieplenie: Nieocieplony* *Rodzaj obcasa: Płaski* *Rozmiarówka: Kup większe* *Sezon: Przejściowy* *Styl: Casual* *Tęgość stopy: Neutralny* *Wkładka/wyściółka: Materiał włókienniczy* *Zapięcie: Wsuwany/otwarty* *TABELA ROZMIARÓW* *Rozmiar 36* *[1] Długość wkładki 22.70 cm* *[2] Szerokość wkładki 7.50 cm* *[3] Wysokość obcasa 2.5 cm* *[6] Wysokość buta 36.5 cm* *[7] Obwód cholewki 38.00 cm* *[8] Obwód łydki 32.00 cm* *Długość stopy 22.20 cm* *Rozmiar 37* *[1] Długość wkładki 22.90 cm* *[2] Szerokość wkładki 7.50 cm* *[3] Wysokość obcasa 2.5 cm* *[6] Wysokość buta 36.5 cm* *[7] Obwód cholewki 38.00 cm* *[8] Obwód łydki 33.00 cm* *Długość stopy 22.40 cm* *Rozmiar 38* *[1] Długość wkładki 24.00 cm* *[2] Szerokość wkładki 8.00 cm* *[3] Wysokość obcasa 2.5 cm* *[6] Wysokość buta 36.5 cm* *[7] Obwód cholewki 38.00 cm* *[8] Obwód łydki 33.00 cm* *Długość stopy 23.50 cm* *Rozmiar 39* *[1] Długość wkładki 24.20 cm* *[2] Szerokość wkładki 8.00 cm* *[3] Wysokość obcasa 2.5 cm* *[6] Wysokość buta 36.5 cm* *[7] Obwód cholewki 39.00 cm* *[8] Obwód łydki 35.00 cm* *Długość stopy 23.70 cm* *Rozmiar 40* *[1] Długość wkładki 24.70 cm* *[2] Szerokość wkładki 8.50 cm* *[3] Wysokość obcasa 2.5 cm* *[6] Wysokość buta 37 cm* *[7] Obwód cholewki 40.00 cm* *[8] Obwód łydki 35.00 cm* *Długość stopy 24.20 cm*

Call for experiments

- **How many** hobbies?
- Which **input data**?
- **Prompt modification** impact?
- **Language** choice?
- … any other

How we managed to compare experiments

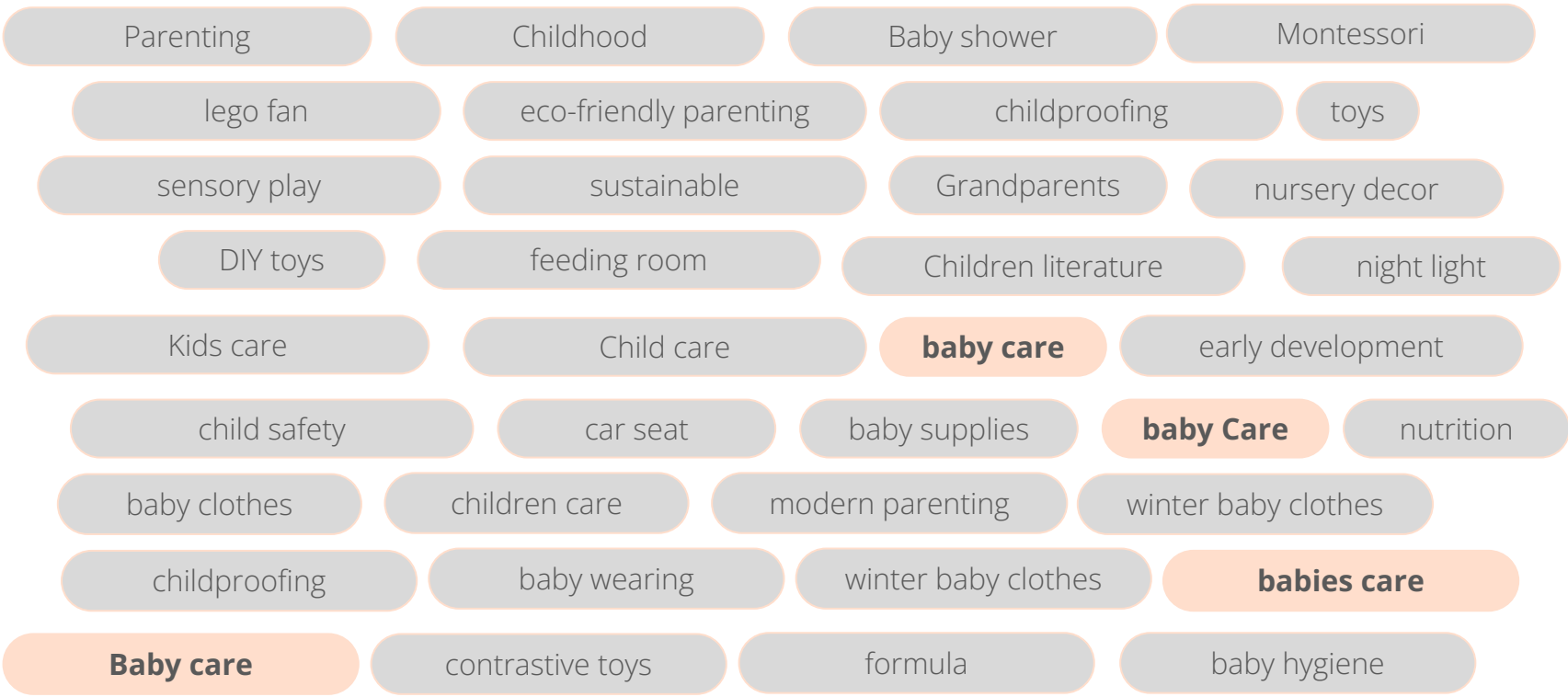


How big is our hobby base?

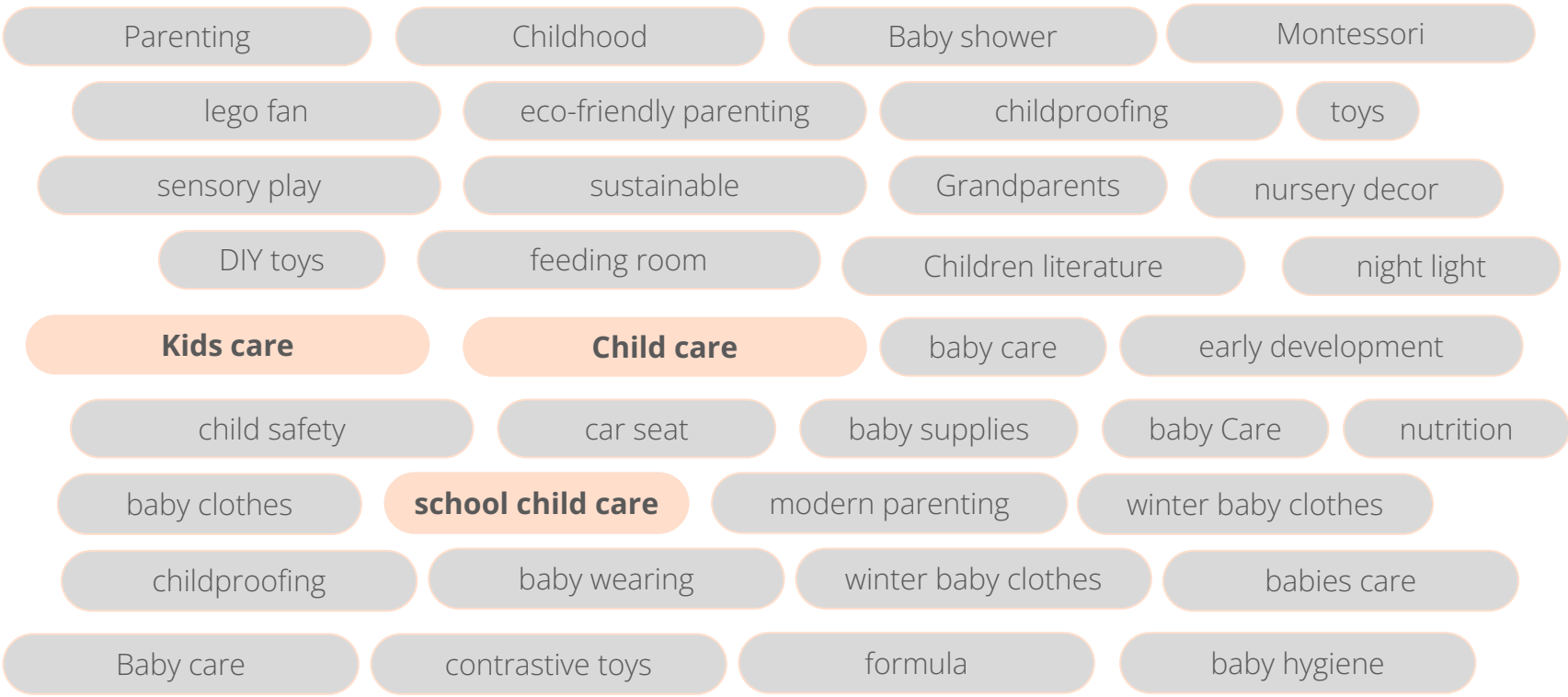
~115k	~10%	~100k
generated hobbies	removed by hobby judge	remaining hobby candidates

Highly unstructured data: 350 children topics

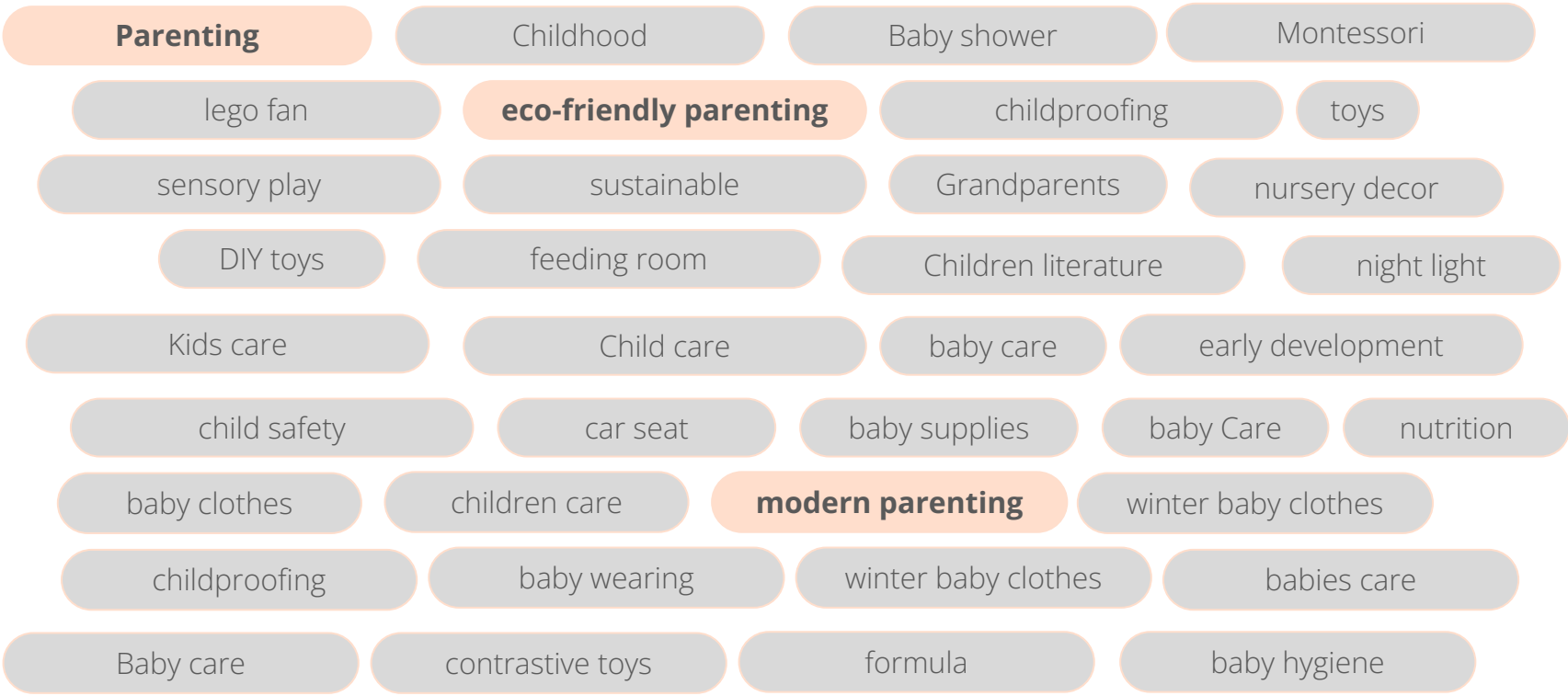
Highly unstructured data: Inconsistent forms



Highly unstructured data: synonyms



Highly unstructured data: inconsistent granularities



HOW TO ORGANIZE TOPIC CHAOS?

From natural language to mathematical language

NATURAL LANGUAGE

SEMANTIC VECTOR SPACE
(768 dimensional)

Embeddings reflect semantic meaning, so we expect:

> jogging to be closer to triathlon than tennis and painting
> painting far away from all three other as not related to sport

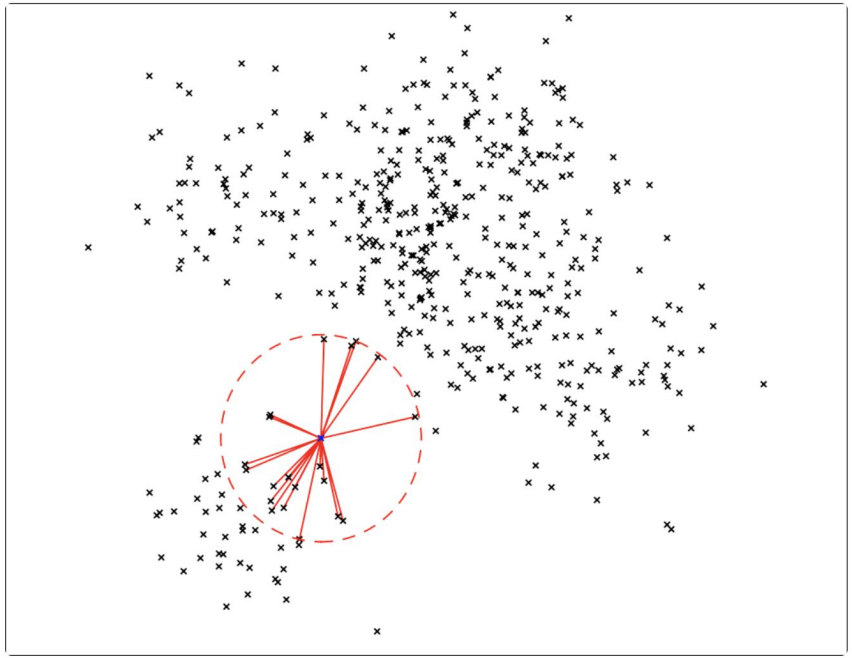
Nearest Neighbor Search

What is it?

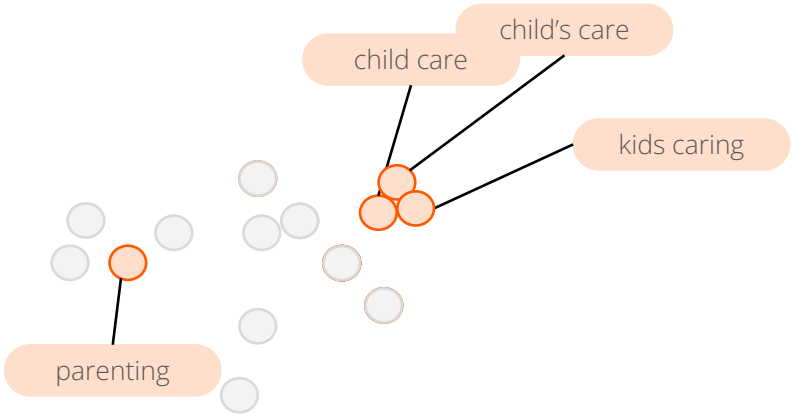
- Finds closest data points

Core Idea:

- calculate distances with all data points
- smallest distances = **neighbors**



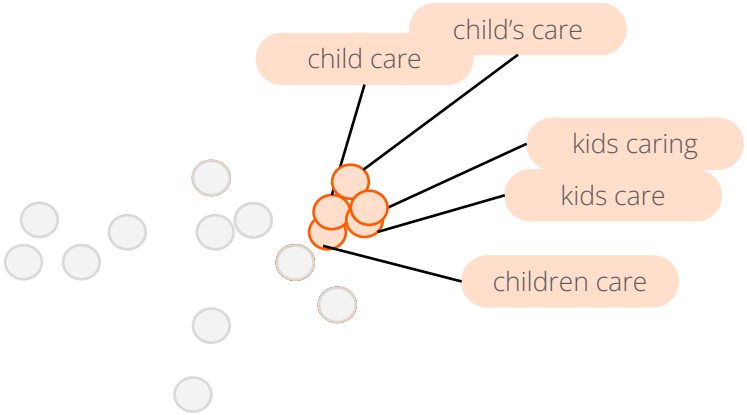
Hobbies pre-processing: synonyms & multiple forms

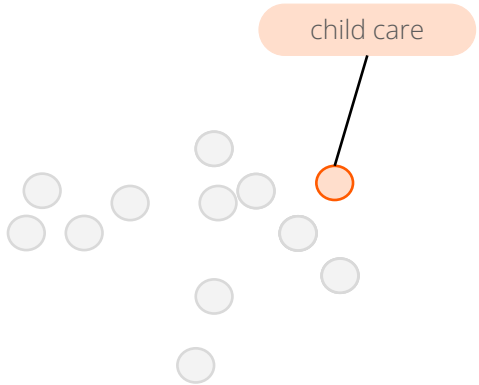


		cosine similarity
child care	kid care	0.99
child care	child's care	0.99
kids caring	child's care	0.98
		similarity threshold
....		....
child care	parenting	0.74

We want to get rid of the neighbors that are too close to each other (cosine similarity above threshold) by **merging them into only one**!

Hobbies pre-processing: synonyms & multiple forms





Pre-processed hobby base

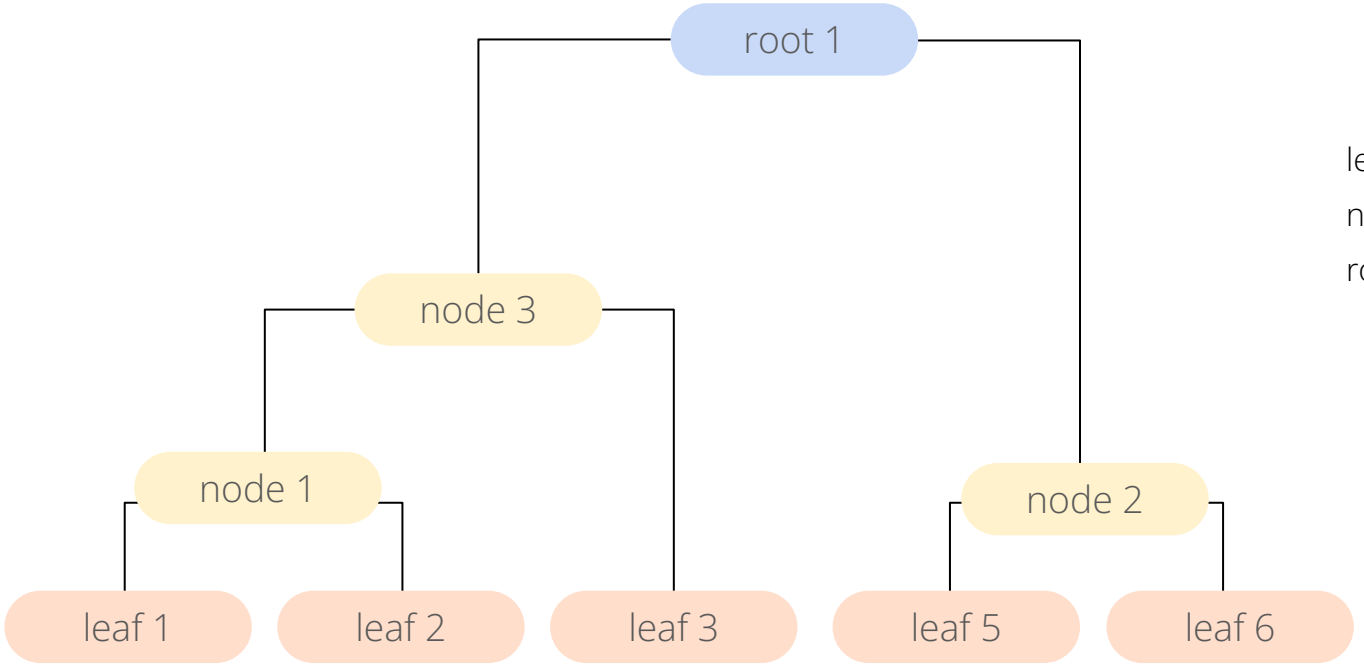
-70%

hobby set reduction
(pre-processing)

~30k

remaining hobbies for
grouping

Hierarchical clustering

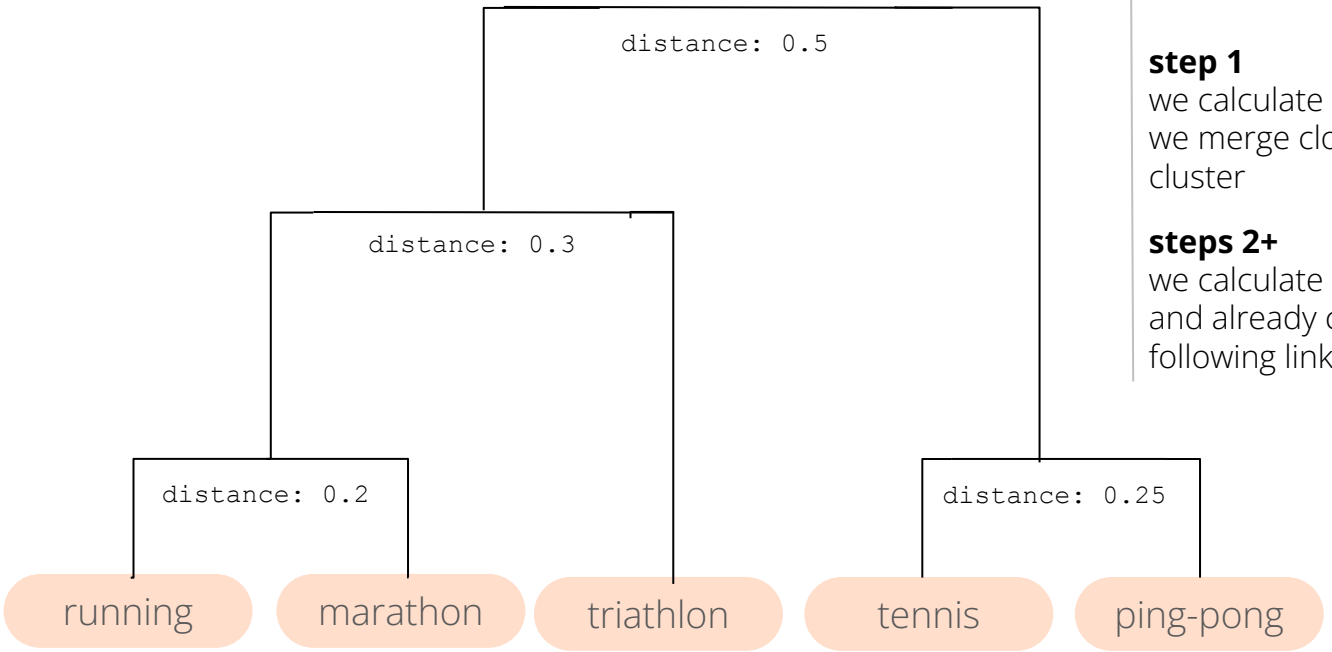


leaf = hobby

node = hobby

root = hobby

Topic taxonomy - agglomerative bottom-up approach

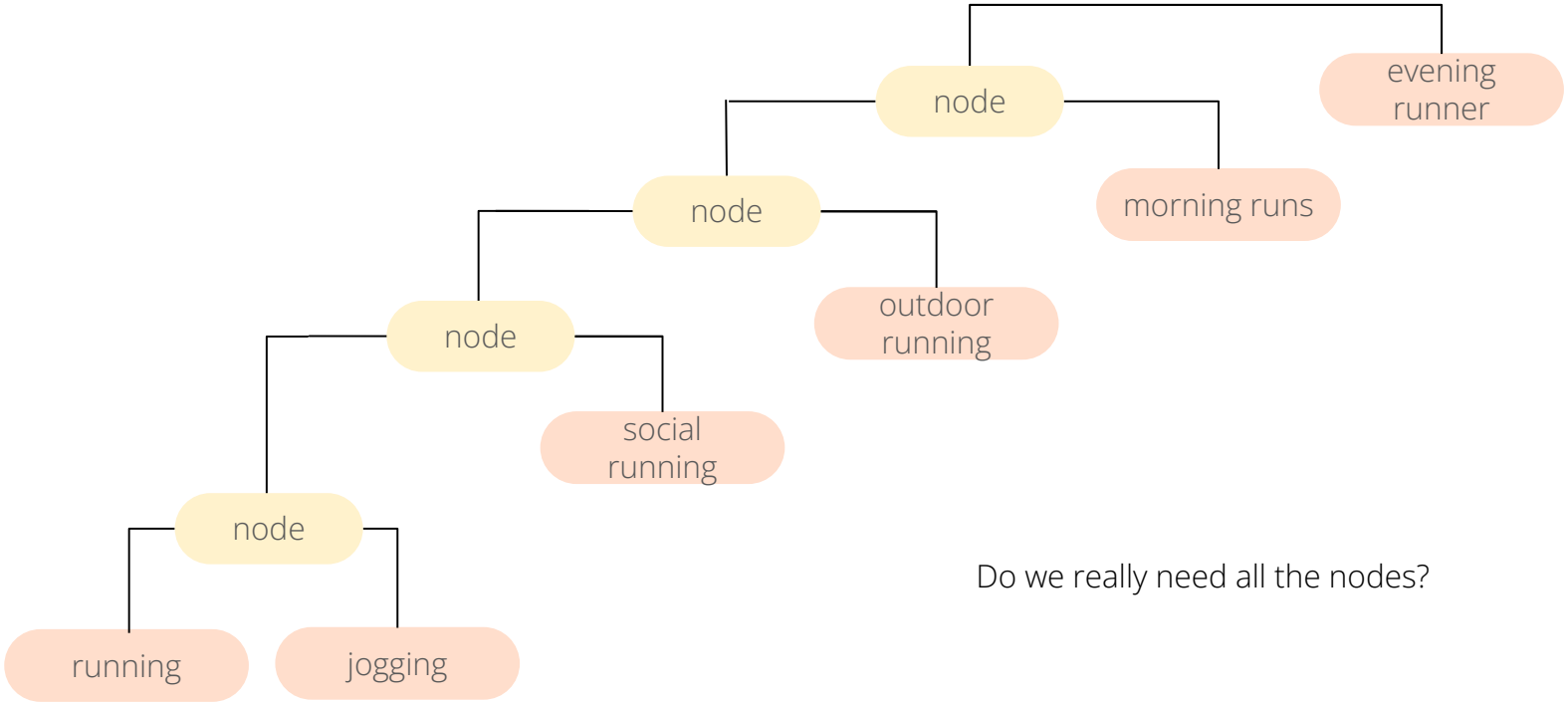


step 0
each topic makes separate cluster

step 1
we calculate distance of each pair and we merge closest ones into one cluster

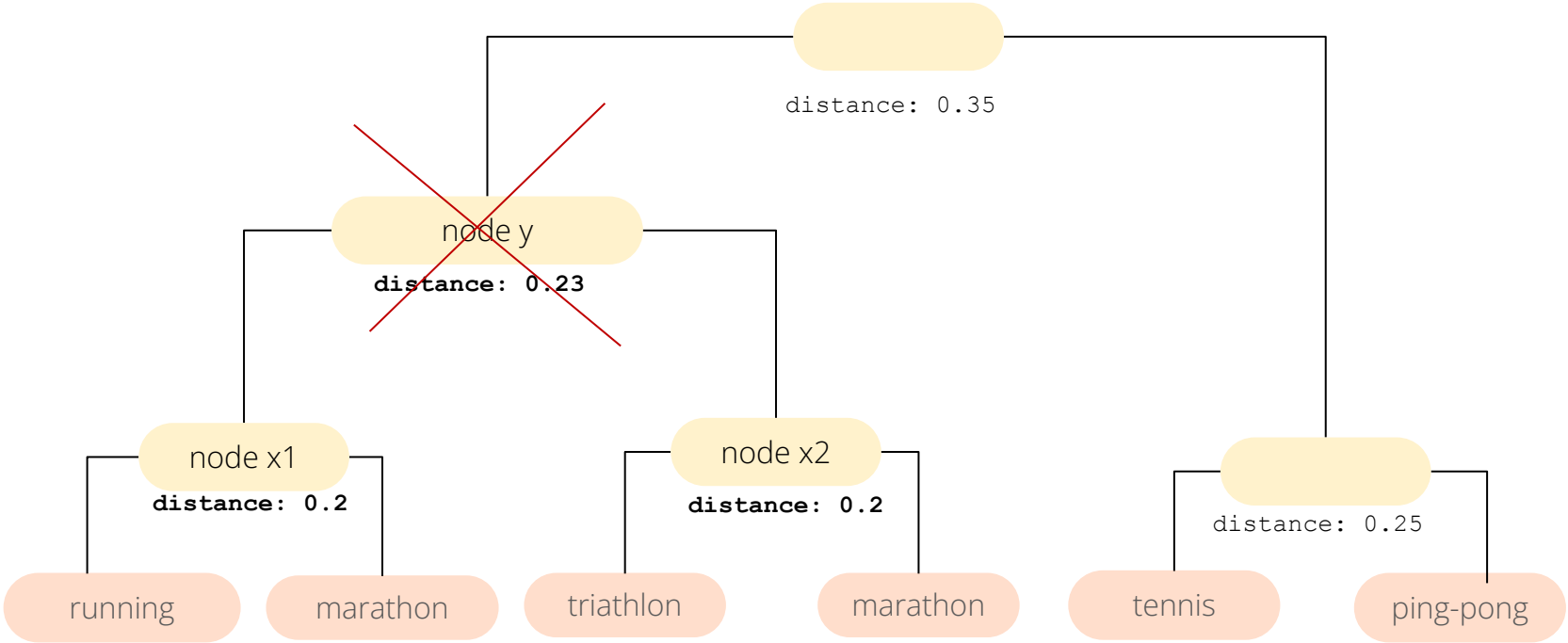
steps 2+
we calculate distance of remaining topics and already created clusters and make following links

Custom post-processing: nodes relevance



Do we really need all the nodes?

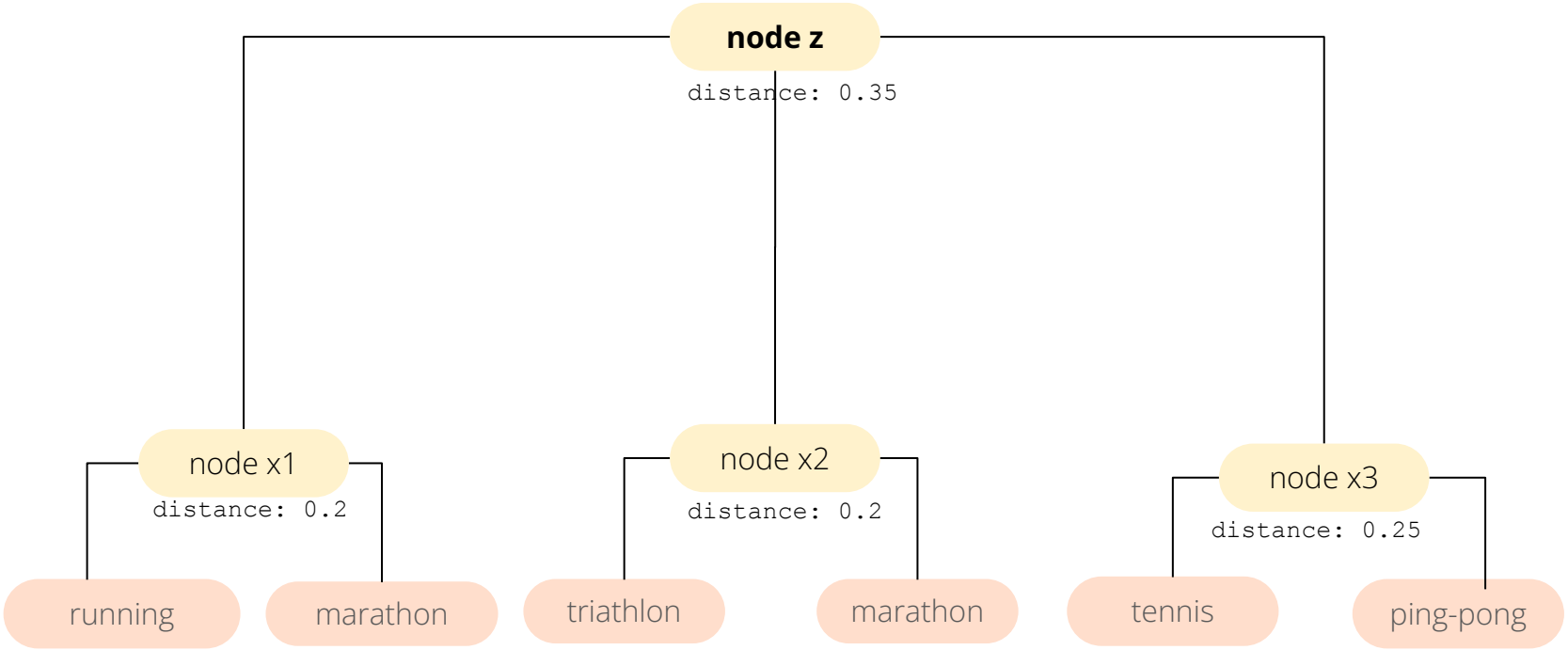
Custom post-processing: removing irrelevant nodes



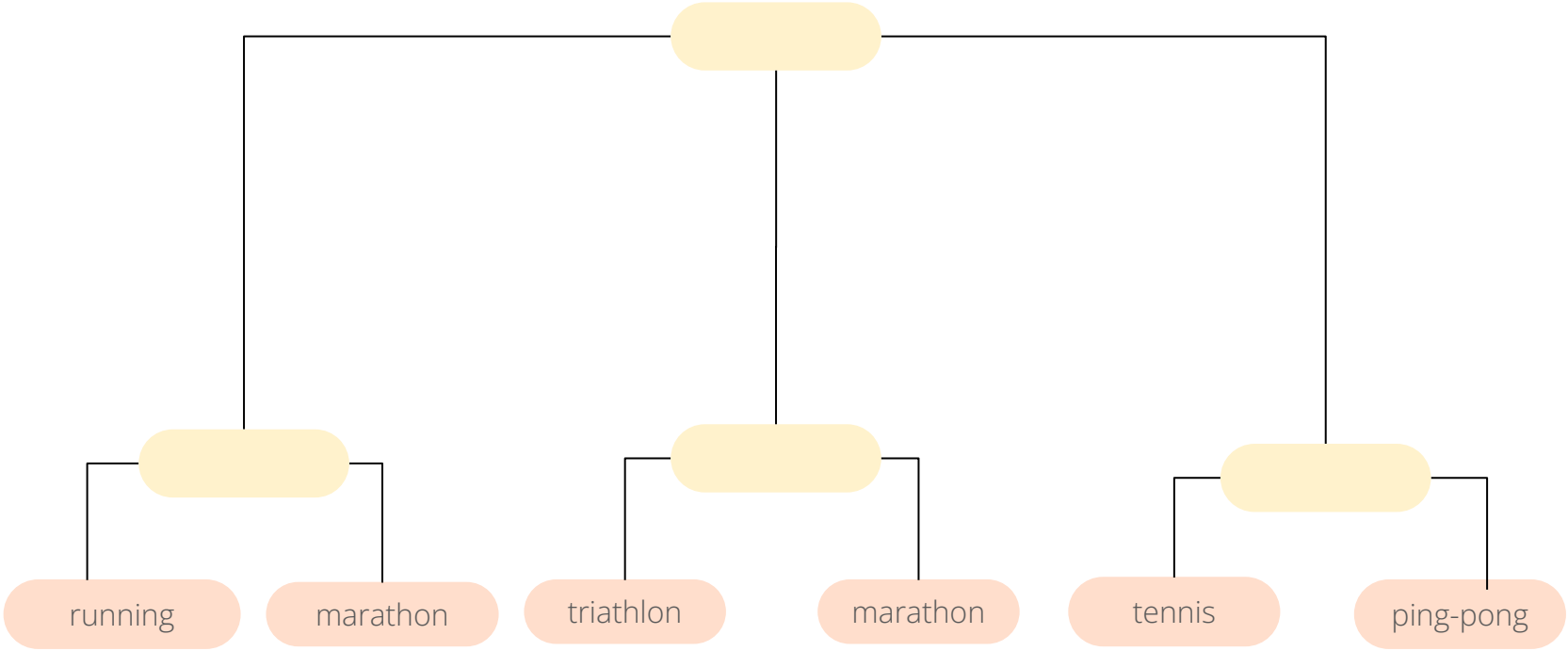
Is 0.23 greater than avg(0.2, 0.2) by more than 20%?

NOPE! We remove node y!

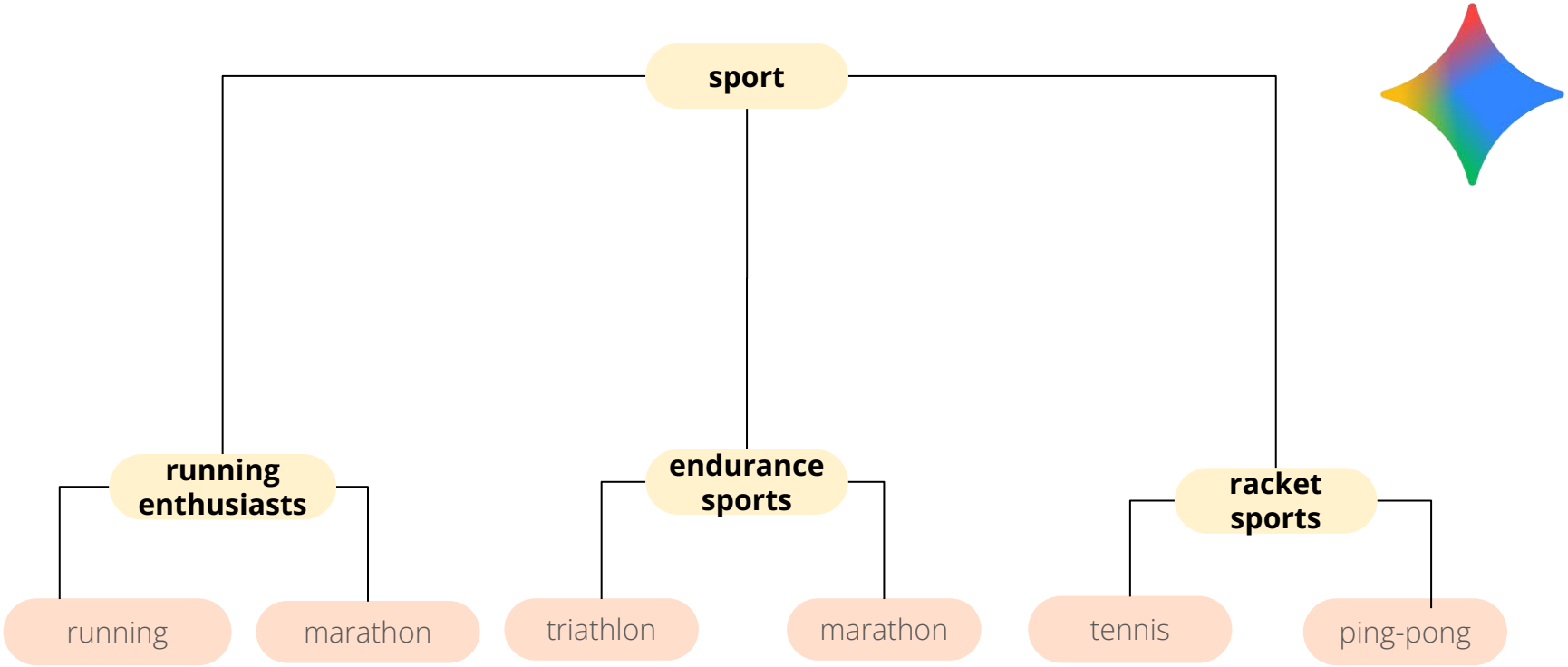
Custom post-processing direction



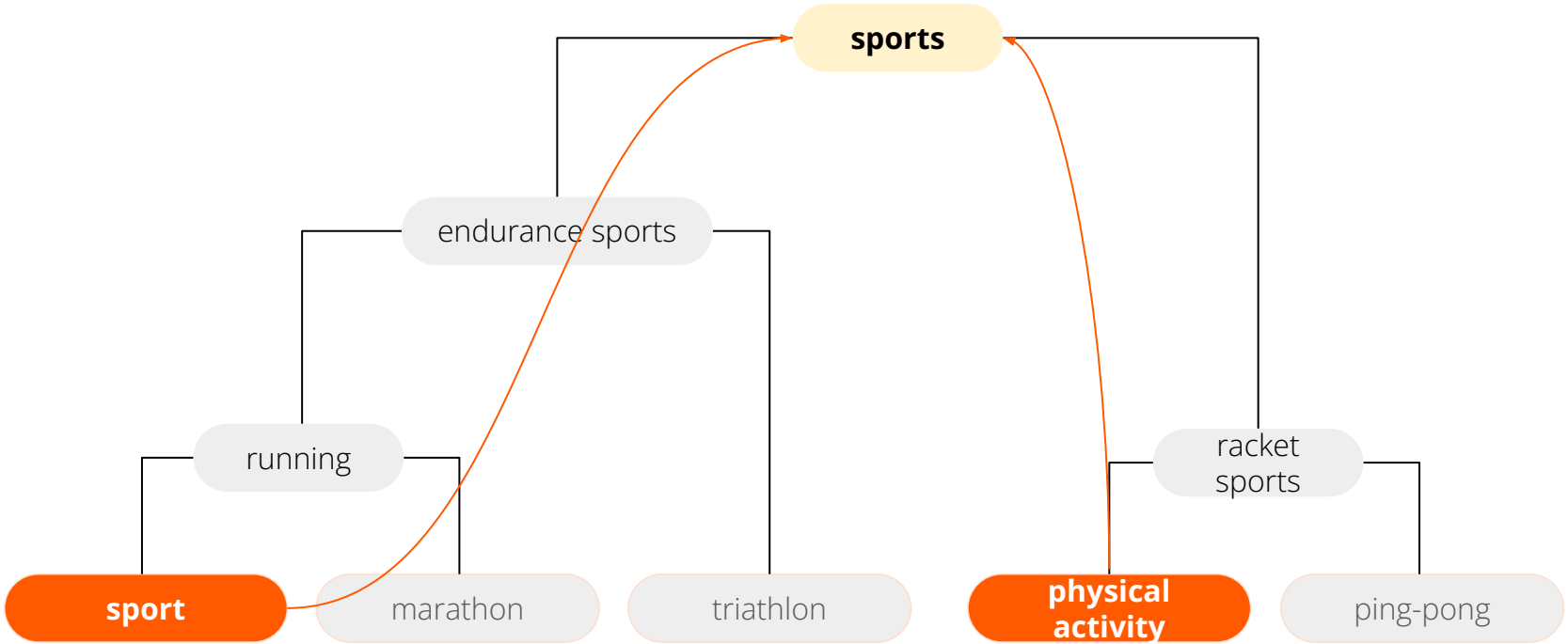
Tree post-processing: What's missing?



Custom post-processing: Naming nodes

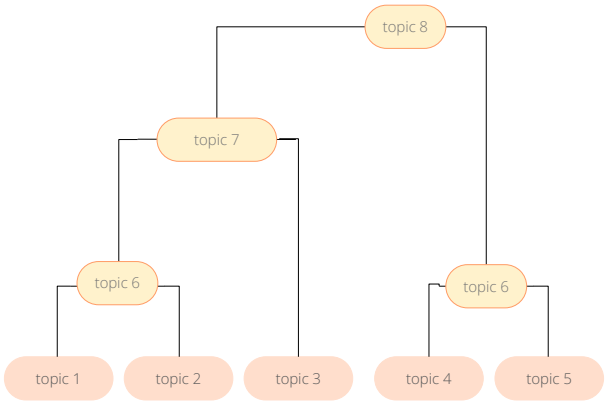


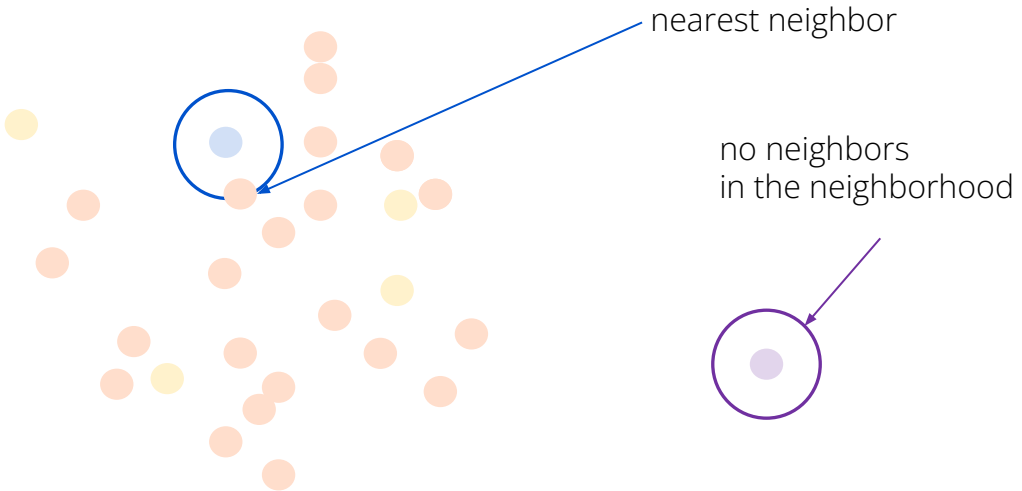
Custom post-processing: making discovered structure coherent



What about new topics?

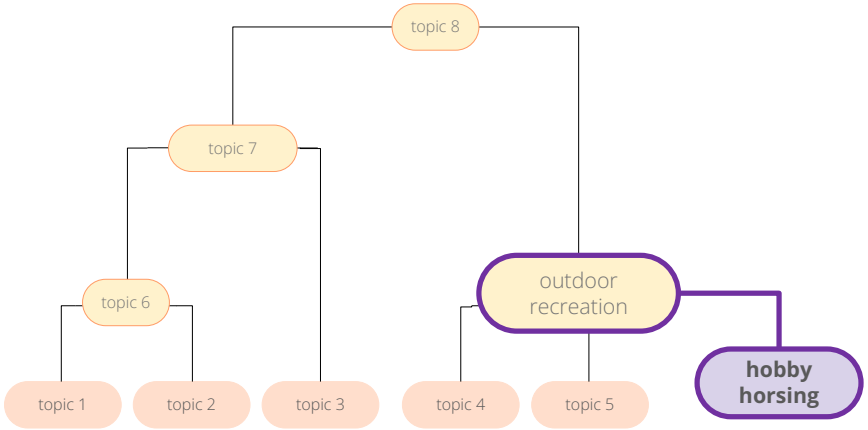
NEW TOPIC	TOPIC FROM THE TREE
# outdoor play	# outdoor play
# active family	# family activities
# hobby horsing	

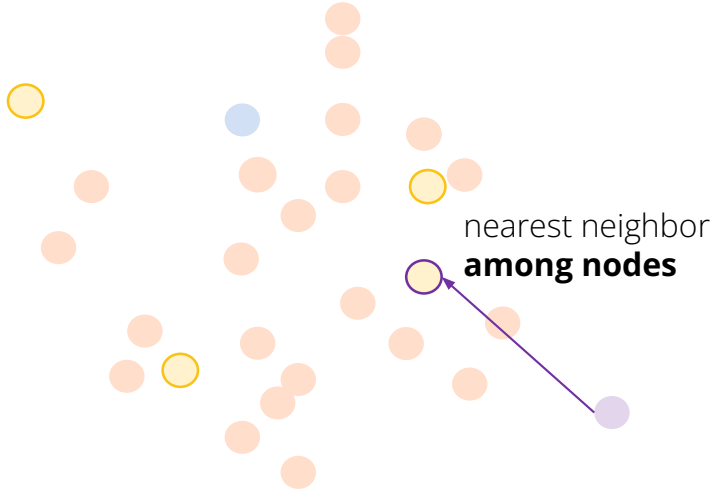




What about new topics?

NEW TOPIC	TOPIC IN THE TREE
# outdoor play	# outdoor play
# active family	# family activities
# hobby horsing	# hobby horsing





The approach: hybrid pipeline

STEP 01

HOBBY GENERATION

Assigning topics to millions of products using textual data.

TECHNOLOGY STACK

Generative AI

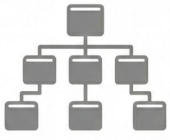
STEP 02

HOBBIES STRUCTURATION

Organizing raw tags into coherent, hierarchical tree-based structure.

TECHNOLOGY STACK

Clustering + Embedding Model + Generative AI



STEP 03

USERS PROFILING

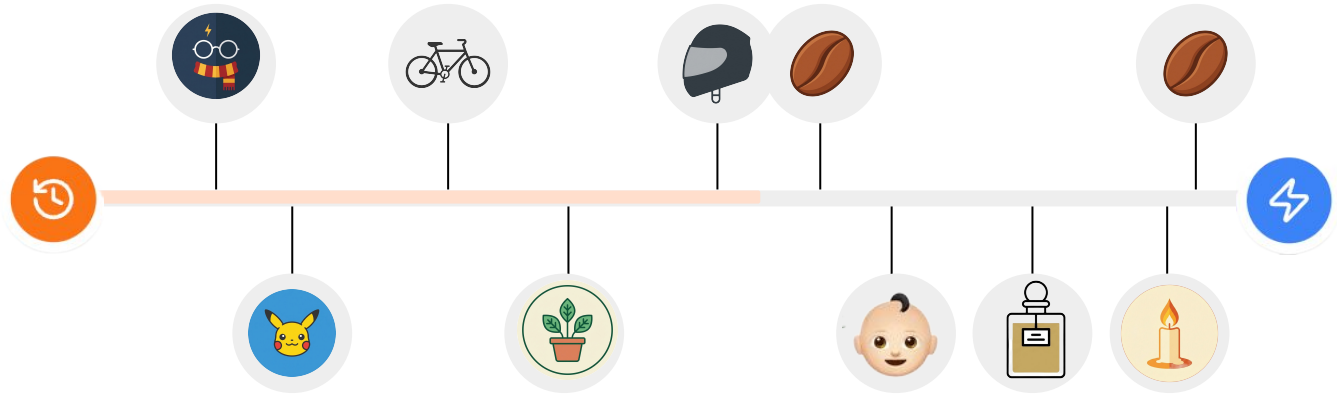
Aggregating history to find the strongest interests using behavioral dynamics.

TECHNOLOGY STACK

Recency Frequency Monetary Scoring

HOW TO GET USER PROFILES?

User profiling input



+ hobbies per each product

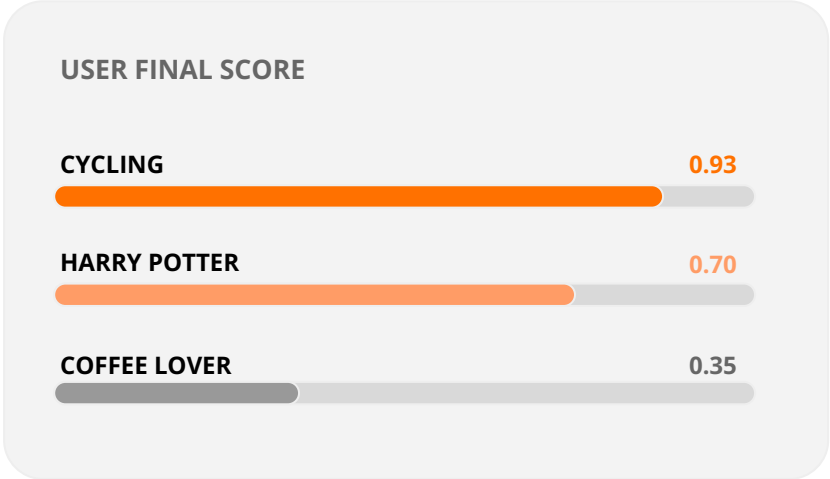
User Profiling Approach

Calculating the strength of interests of each user using behavioral dynamics.

R **Recency**
How long ago was the last interaction?

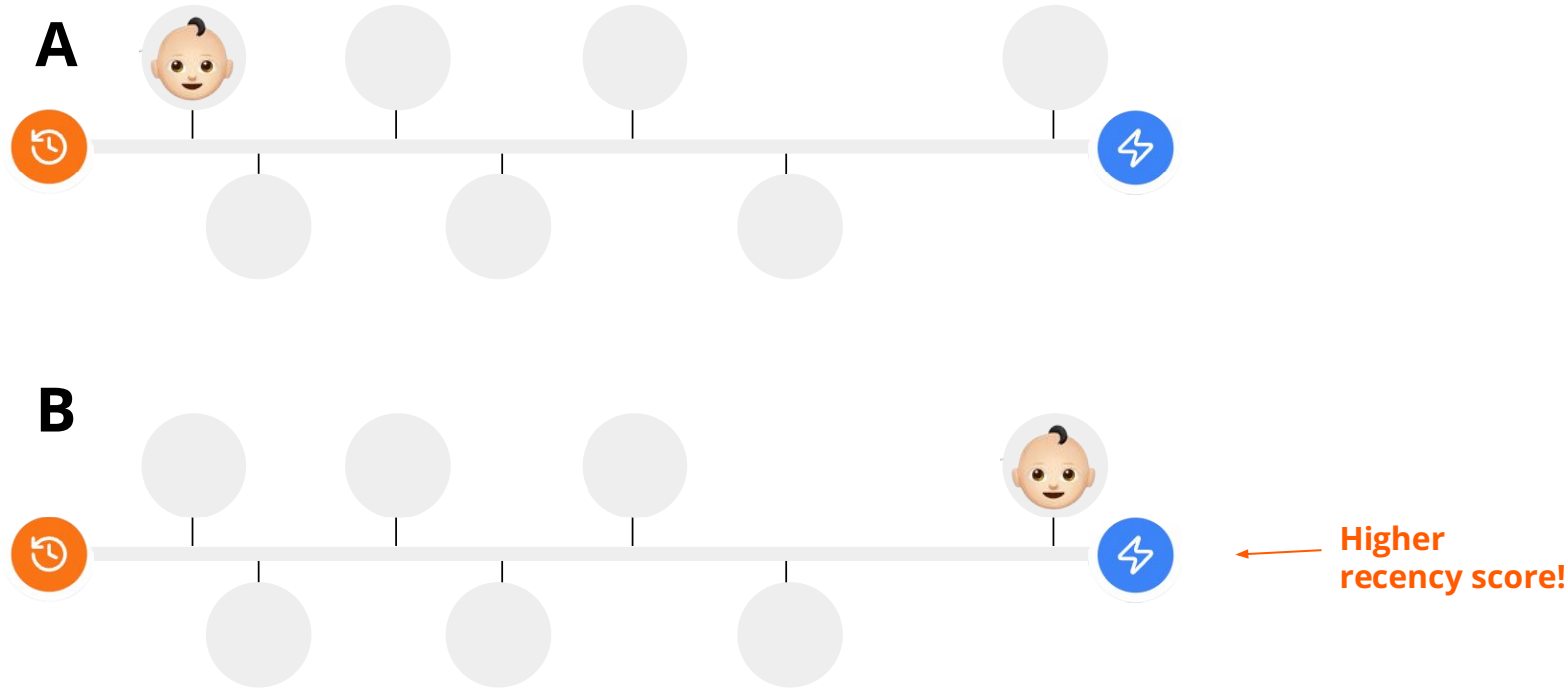
F **Frequency**
How often does the user engage with this topic?

M **Monetary**
What is the value of items in this hobby?



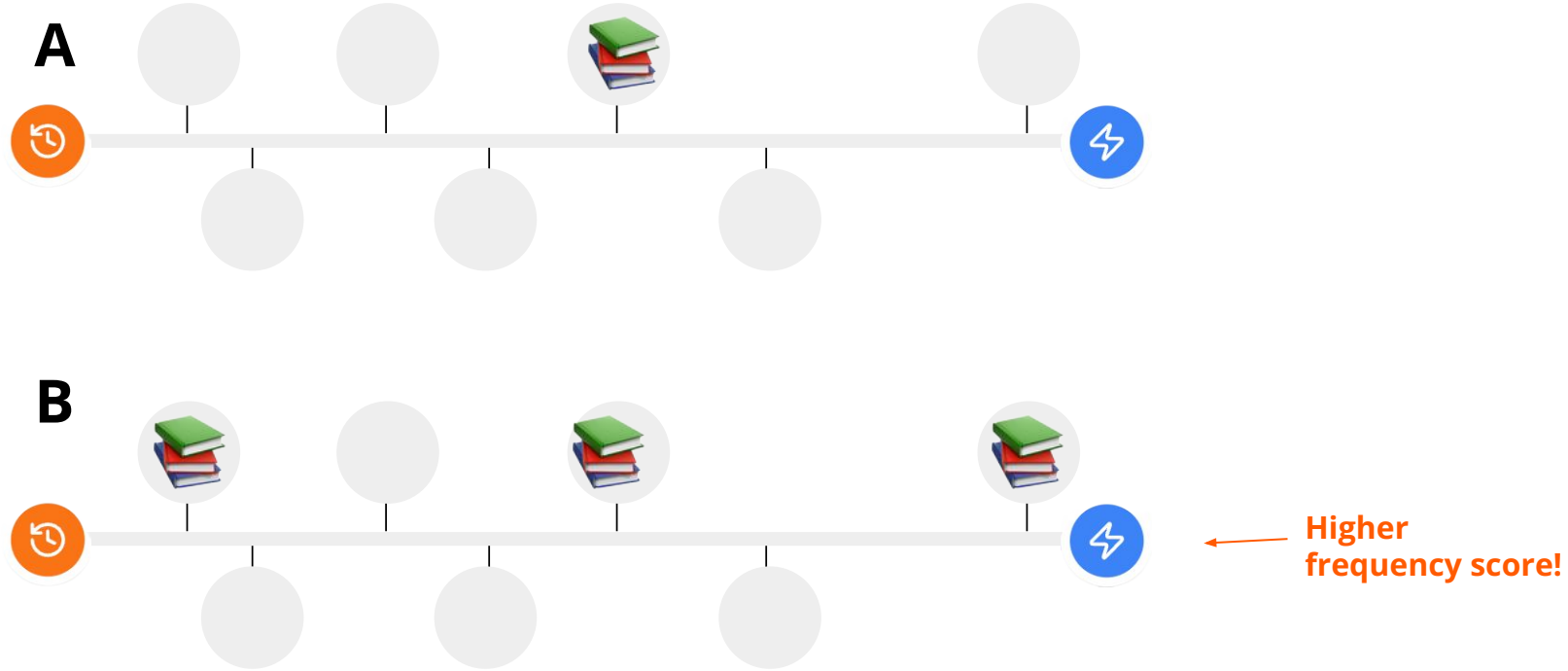
Why recency?

How long ago was the last interaction?



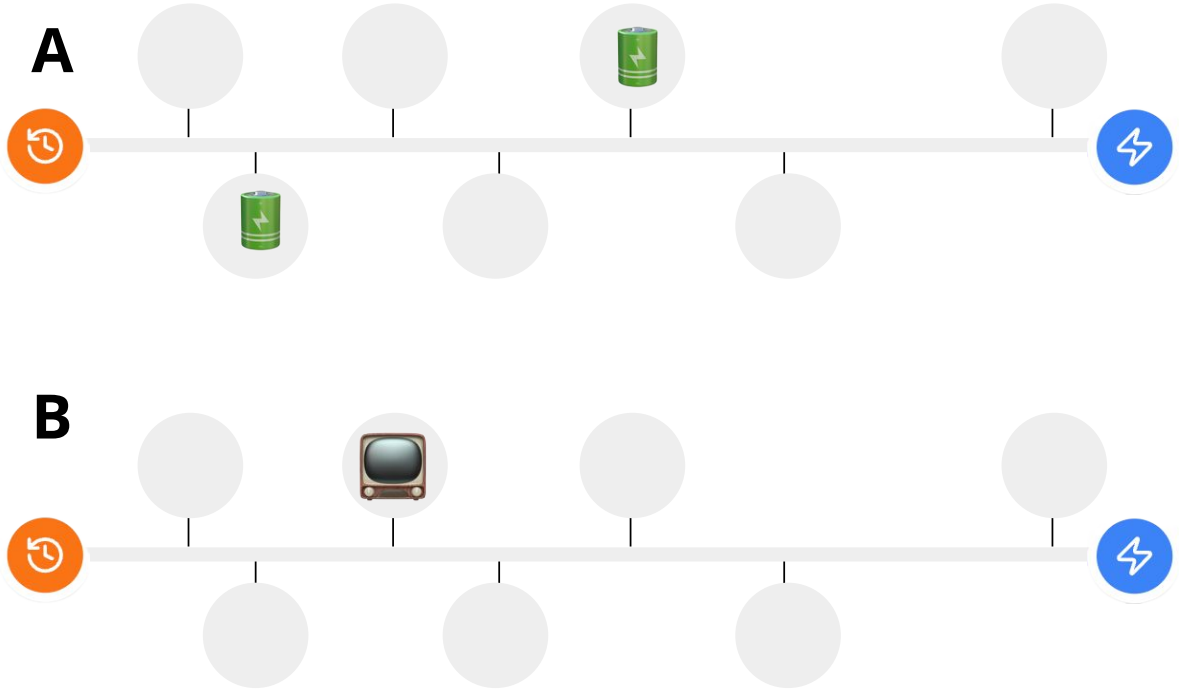
Why frequency?

How often does the user engage with this topic?



Why monetary?

What is the value of items in this hobby?



Importance of scores calibration

BEST USER FOR A TOPIC

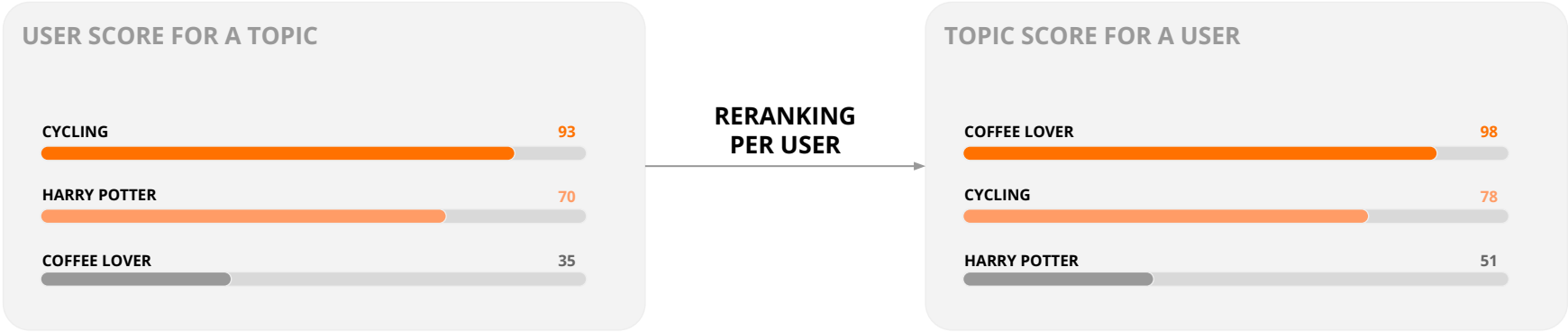
≠

BEST TOPIC FOR A USER

RFM score is based on **relativity to all the other users** engaged in a topic.

Must be relative to the other topics of a user.

How to rerank the hobbies?



- How **much spent** on hobby vs other hobbies ?
- How **many unique products** related to a hobby?
- How **long** am I interested in a hobby?
- Am I relatively **engaged** in a hobby? (RFM)

OUR INTERESTS PROFILES

Our passions vs hobby profiles

motherhood

skiing

contemporary dance

HEART HEALTH	0.82
BABY CARE	0.82
WATER SPORTS	0.8
CAREGIVING	0.8
PUZZLES	0.79
PERSONAL HYGIENE	0.79
KIDS INTERESTS	0.77
CHILD HYGIENE	0.76
HEALTHY LIFESTYLE	0.76
SKINCARE	0.75

Our passions vs hobby profiles

motherhood

skiing

contemporary dance

HEART HEALTH	0.82
BABY CARE	0.82
WATER SPORTS	0.8
CAREGIVING	0.8
PUZZLES	0.79
PERSONAL HYGIENE	0.79
KIDS INTERESTS	0.77
CHILD HYGIENE	0.76
HEALTHY LIFESTYLE	0.76
SKINCARE	0.75

Our passions vs hobby profiles

belly dance

healthy lifestyle

board games

WELLNESS	0.97
BEAUTY	0.84
HAIR CARE	0.81
HEALTH	0.77
TEXTILE ARTS	0.76
GADGET	0.75
ORAL HYGIENE	0.73
COSMETICS	0.73
HOME ORGANIZATION	0.69
FANTASY	0.66

Our passions vs hobby profiles

belly dance

healthy lifestyle

board games

WELLNESS	0.97
BEAUTY	0.84
HAIR CARE	0.81
HEALTH	0.77
TEXTILE ARTS	0.76
GADGET	0.75
ORAL HYGIENE	0.73
COSMETICS	0.73
HOME ORGANIZATION	0.69
FANTASY	0.66

Our passions vs hobby profiles

belly dance

healthy lifestyle

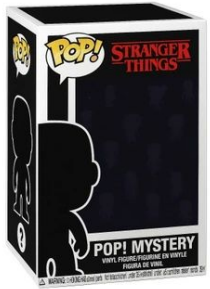
board games

WELLNESS	0.97
BEAUTY	0.84
HAIR CARE	0.81
HEALTH	0.77
TEXTILE ARTS	0.76
GADGET	0.75
ORAL HYGIENE	0.73
COSMETICS	0.73
HOME ORGANIZATION	0.69
FANTASY	0.66

HOW DO WE UNLOCK BUSINESS VALUE?

Targeted campaigns: Stranger Things Edition

Koszulka Hawkins Fashion



Figurka Pop Collectibles

Koszulka Hawkins Home

Kinder Joy Food

PRODUCT SELECTION

How do we identify the right inventory?

Prioritize products with Stranger Things hobby tags.

AUDIENCE LOGIC

What are the ideal recipients?

Target users having a high affinity score for stranger things fandom or 80s nostalgia/sci-fi genre.

Thematic collections

Harry Potter Fandom

Pokemon Collection

Labubu Series

SELECTION LOGIC

Which collection to show to user X?

We can sort collections using topic-user ranking score.

Key takeaways

- hobby judge enables iterative prompt optimization
- it’s worth to assign > 1 hobby to a product
- hobby generation does not have to be expensive
- inconsistent granularity needs attention
- proper aggregation level is crucial to achieve diverse hobby profile
- new topics need attention
- scoring users per topic != scoring topic per user

Thank you for your attention!

allegro