Evaluating E-commerce and Marketplace Search User Perception vs. Business Metrics

Haystack EU 2024

Julien Meynet

wallapop

Wallapop is the leading platform in **conscious** and **human consumption**, that aspires to create a unique inventory ecosystem of reused or unique products. We are connecting a community of **+19M** users in Southern Europe **+100M** new listings per year.

Our philosophy:

Opportunity / Sustainability / Accessibility

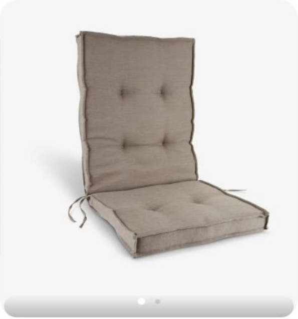
Disclaimer

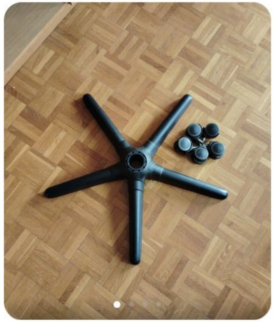
Content of this presentation is not specifically related to Wallapop or how search works at Wallapop

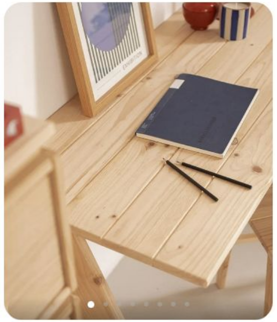
Motivation

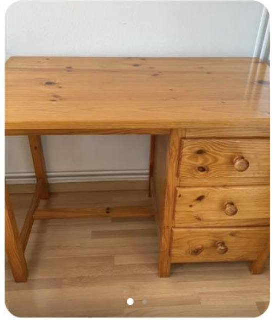
Starter example

🔎 chair



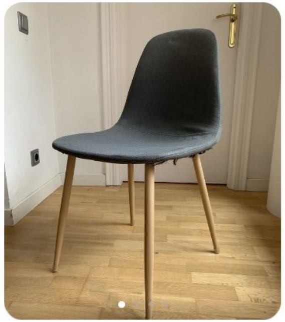




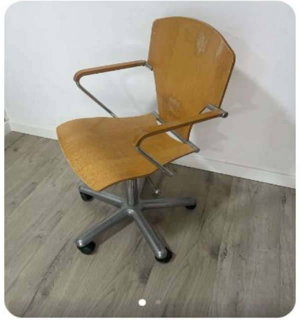


Starter example

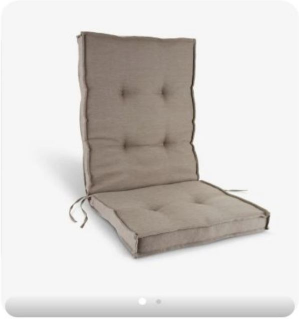
🔎 chair



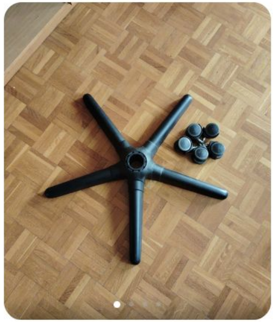
✅



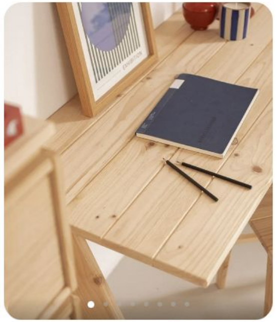
✅



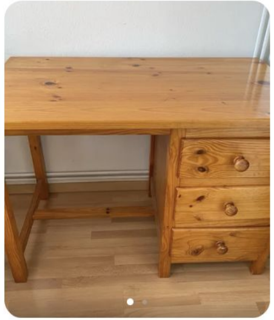
❓



❓



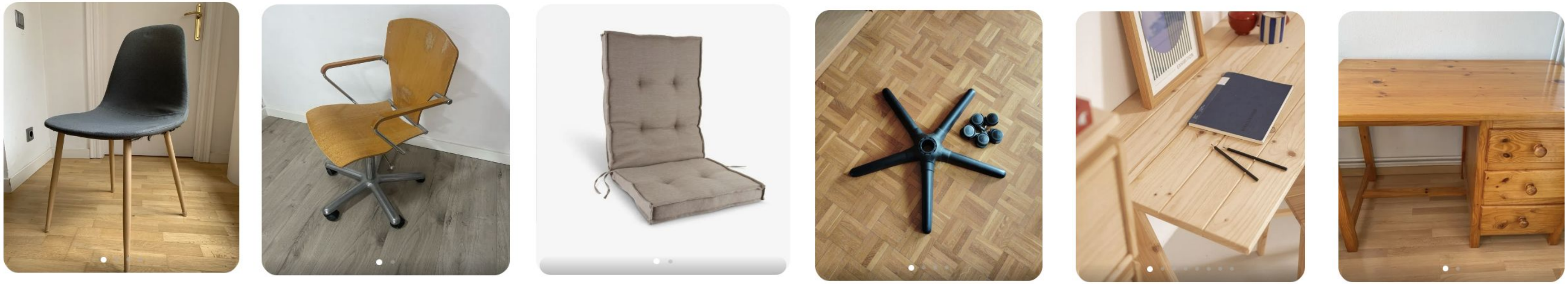
⁉️



⁉️

Quiz

🔎 chair

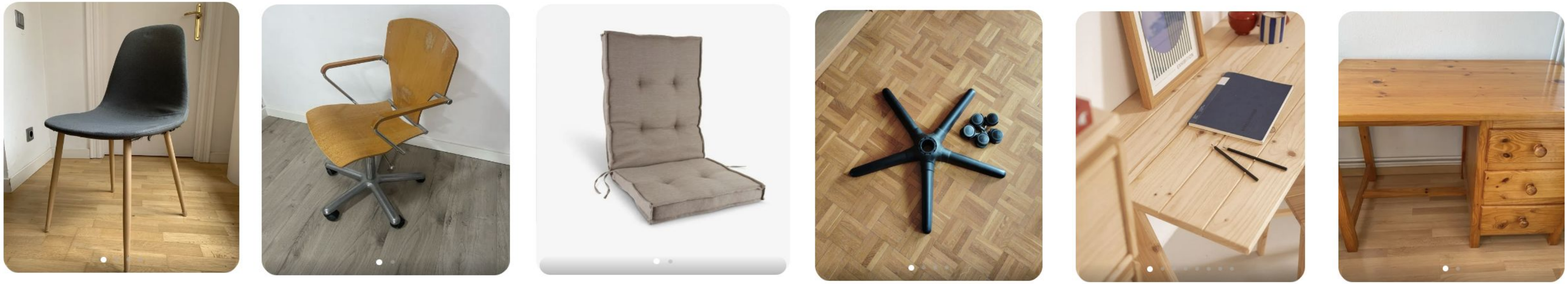


Do you think this result list is “good” ?

👤 From **user** perspective

Quiz

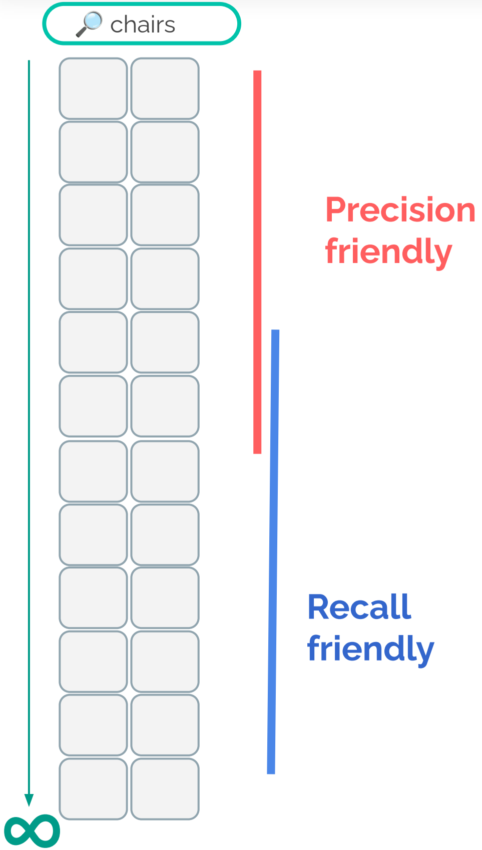
🔎 chair



Do you think this result list is “good” ?

💰 From **business** perspective
(especially if some of those are promoted)?

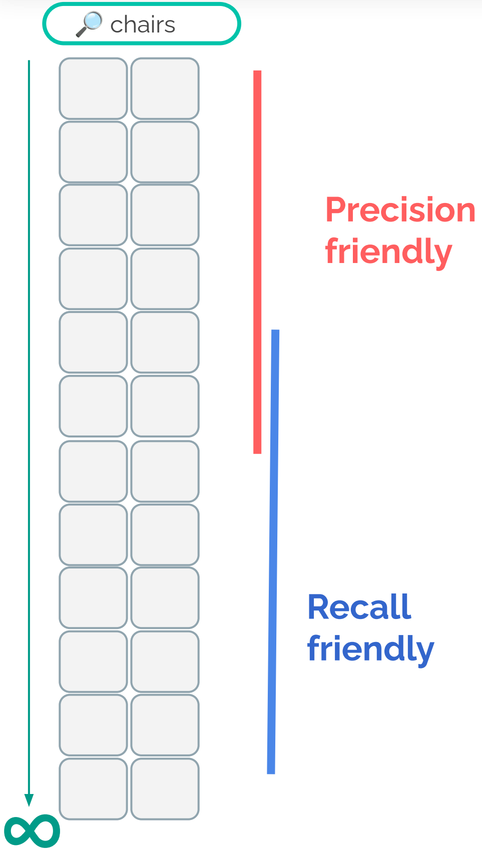
Precision & recall revisited



An old school problem, but still important today, especially for e-commerce and marketplaces

Many search systems show **infinite feeds** of scrollable results

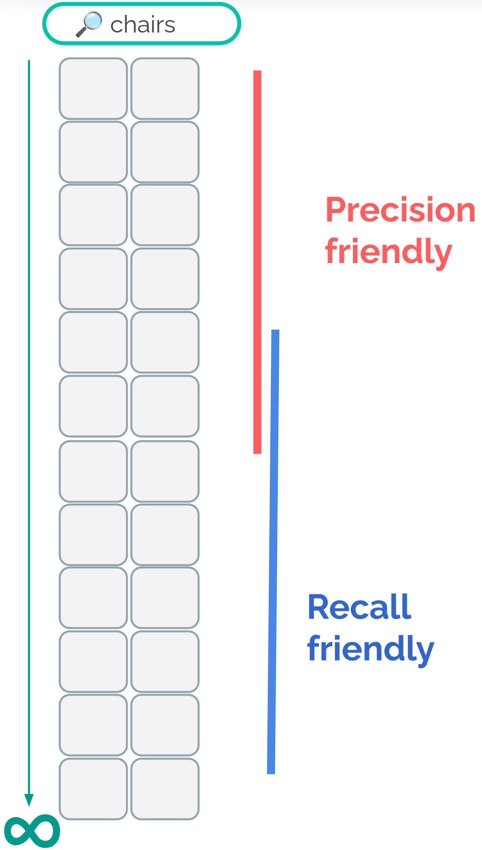
Precision & recall revisited



Why ?

- **Discovery**: we want to keep users engaged
- We want to **avoid 0 / few results** which provides poor experience
- Search often framed as 2 phases:
  - candidate retrieval
  - (multi-stage) re-ranking

Precision & recall revisited



chairs

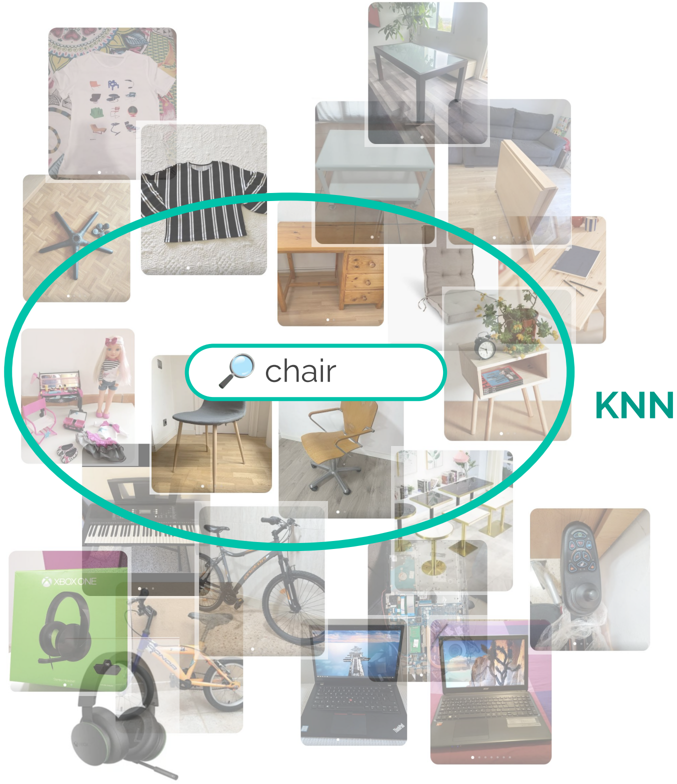
Precision friendly

Recall friendly

What is the problem?

- Many users scroll **a lot** and might end up seeing “not so relevant” results (especially for specific queries)
- “Non-relevance” sorts (price, newest, etc..) might end up showing random results

Precision & recall revisited



Dense retrieval (Vector Search) also changed the game, being extremely good at recall … almost too good 🫣

In this talk we will revisit precision & recall from 2 perspectives:

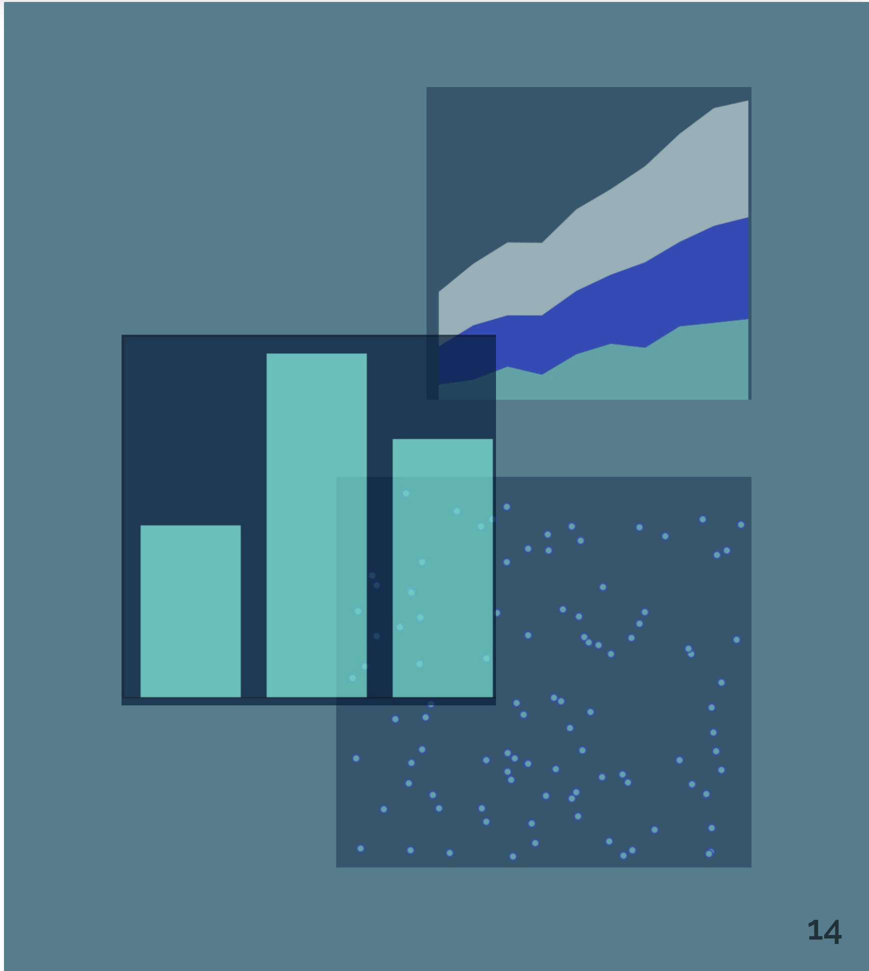
- **User perception**
- **Business metrics**

Evaluating Search

Evaluating search

Let's review some different **metric types**:

- Precision & recall
- User experience metrics
- Business metrics



Precision & recall - reminder

Precision

- The percentage of returned items that are relevant
- **Extreme case**: return only the best possible item, if any

VS ⚡

Recall

- The percentage of relevant items that are returned
- **Extreme case**: return all catalogue

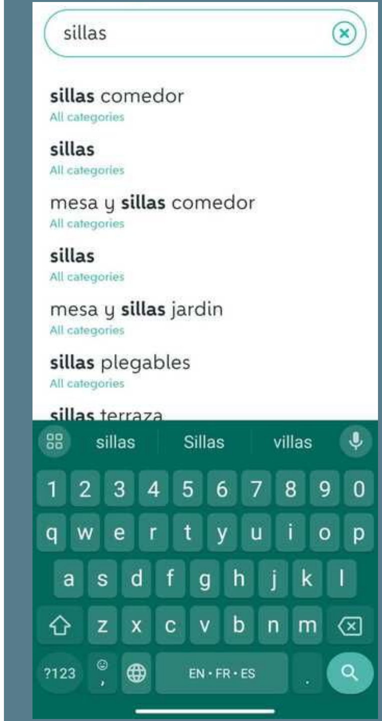
Assumes relevance is binary, which is not true in practice in 90% of cases

User experience metrics (UX metrics)

Measure **user satisfaction** when interacting with a search system.

Wait, why bother? we already have conversions from search, CTRs , NDCGs, etc..

Search is more than just retrieval and ranking. **User satisfaction depends on the whole search experience**



sillas

sillas comedor
All categories

sillas
All categories

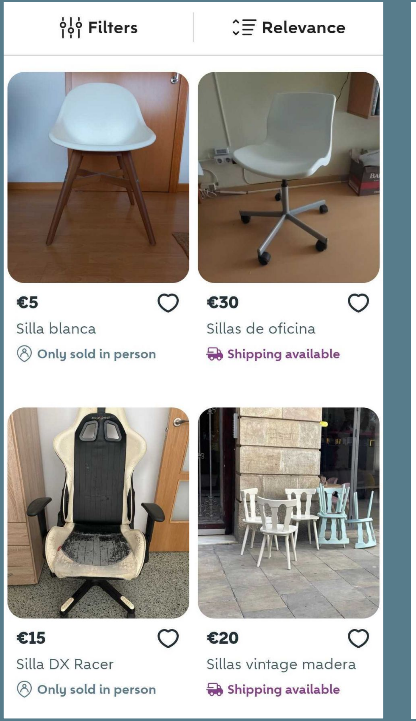
mesa y **sillas** comedor
All categories

sillas
All categories

mesa y **sillas** jardin
All categories

sillas plegables
All categories

sillas terraza

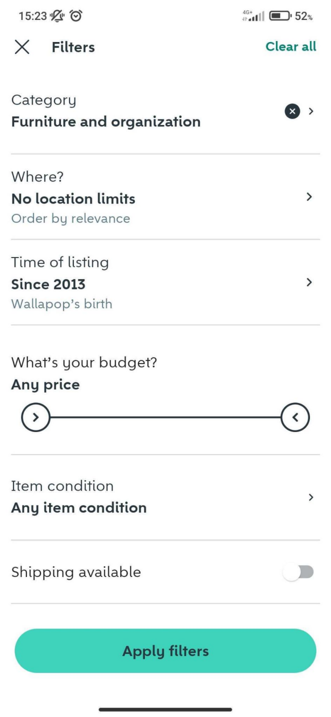


€5
Silla blanca
Only sold in person

€30
Sillas de oficina
Shipping available

€15
Silla DX Racer
Only sold in person

€20
Sillas vintage madera
Shipping available



15:23 52%

Filters Clear all

Category
Furniture and organization

Where?
No location limits
Order by relevance

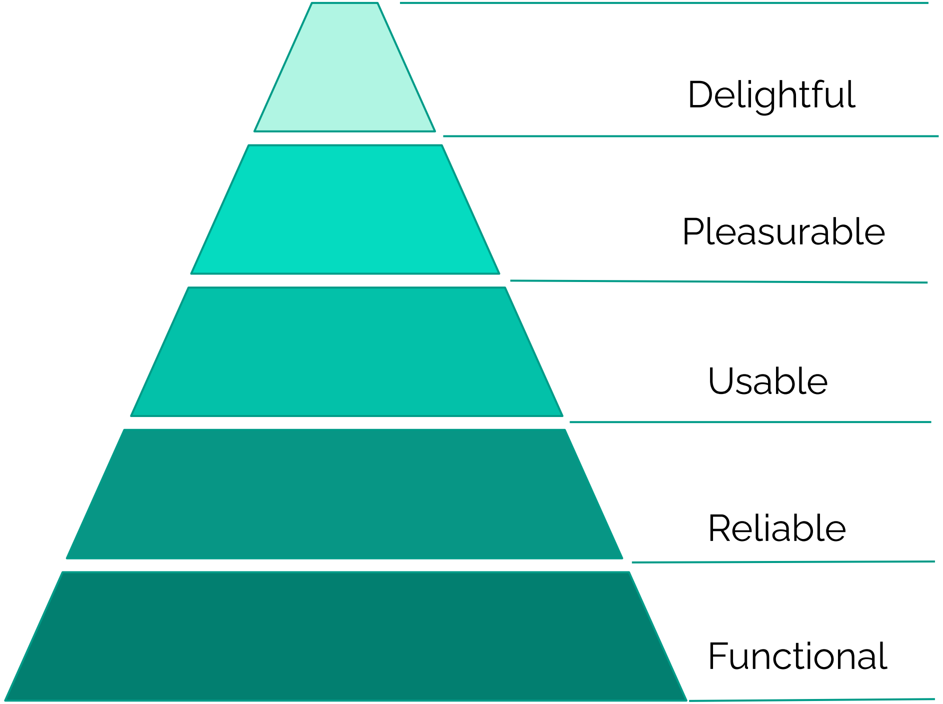
Time of listing
Since 2013
Wallapop's birth

What's your budget?
Any price

Item condition
Any item condition

Shipping available

UX - Pyramid of user needs



Search is a core functionality

What do users care about?

- Ability to interact with the system in an intuitive way
- Finding what they are looking for
- Stability (no technical glitches, acceptable latency, etc..)

Search users usually expect reliable, accurate results, not necessarily emotional or delightful experiences.

wallapop

How to measure UX metrics?

Several commonly used methods for collecting UX metrics

Surveys

Send surveys to group of users and ask them questions like, "Are the search results relevant?" or "How would you rate your overall search experience?"

User Interviews

Conduct user interviews and interact with them to understand their main pain points when using search.

Feedback component

Collect feedback within search results (👍 / 👎 or free text input)

App review

Read / analyze comments users leave in app reviews. Spoiler: they often complain about search 😉

User perception vs user utility

user perception != user utility

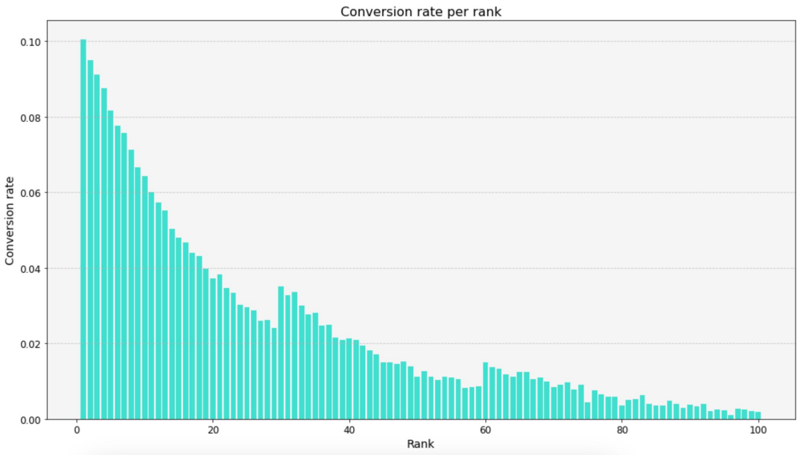
Business metrics

Measure how good search is **for our business**, i.e closer to “conversions”

Also, incorporate different business objectives (e.g, sponsored sellers, or results), not only relevance for the searcher.

Examples

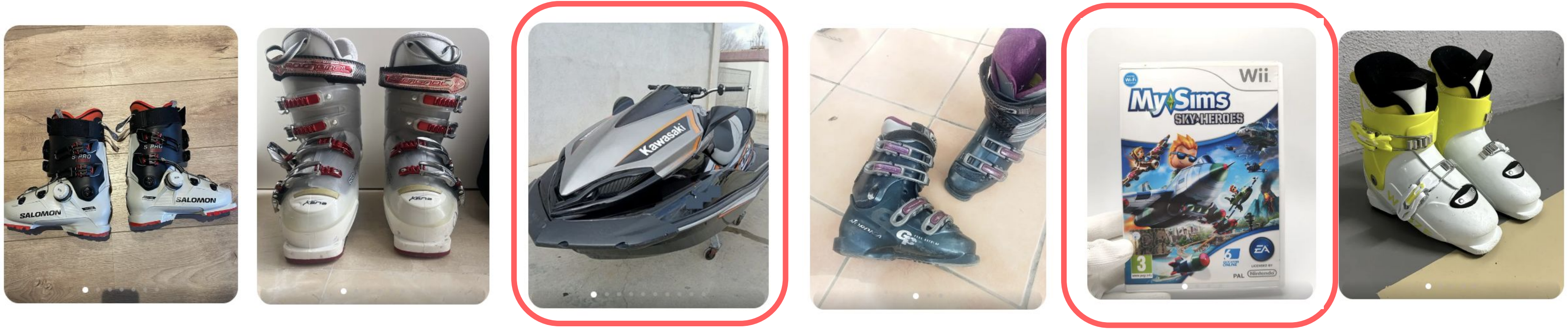
- search conversions (clicks, orders, chats, transactions)
- user engagement / retention
- zero results



The hidden cost of showing **irrelevant** results

🔎 ski boots

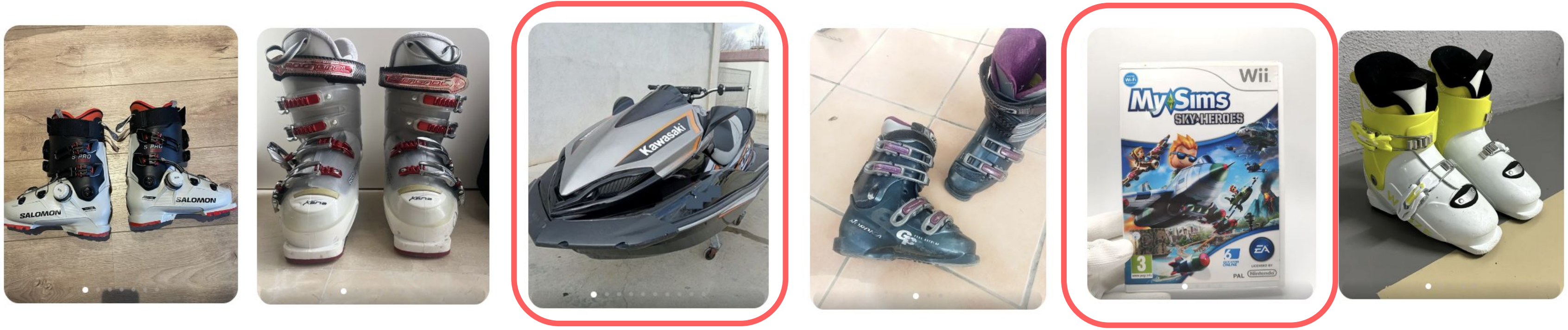
Let's consider a search result list with (very) irrelevant results



The hidden cost of showing **irrelevant** results

🔎 ski boots

Let's consider a search result list with (very) irrelevant results



Cost for User Perception

"Search is broken" / "Please fix the search" / "Urgent need to improve search accuracy"

Cost for Business

There will likely be no conversion on those items, **not a big deal**. Equivalent to having 2 relevant results not clicked.

The hidden cost of **not** showing **relevant** results

🔎 ski boots

Let's consider a search result list with missing relevant results

The hidden cost of **not** showing **relevant** results

🔎 ski boots

Let's consider a search result list with missing relevant results

Cost for User Perception

No impact at all on user perception. Users can't guess there were relevant results if they don't see them!! They might feel catalog is not rich, but won't think search is doing a poor job.

Cost for Business

Potentially **big opportunity missed** depending on how many relevant results are missing. If 1 result eventually ends up converting → big impact

Correlation

User perception of search

~

Precision

VS

⚡

Business metrics

~

Recall

Similar to precision & recall, user perception of search and business metrics might be contradictory, and finding the right trade off is key

Open (controversial) question: could it be that business metrics are more correlated to user **utility**??

😱

Takeaways

Case 1

You have been (over) optimizing your CTRs, NDCGs, CRs for years using the most advanced ML models

=> Ask your users what they think about your search! You might discover some clear dysfunctional aspects the experience

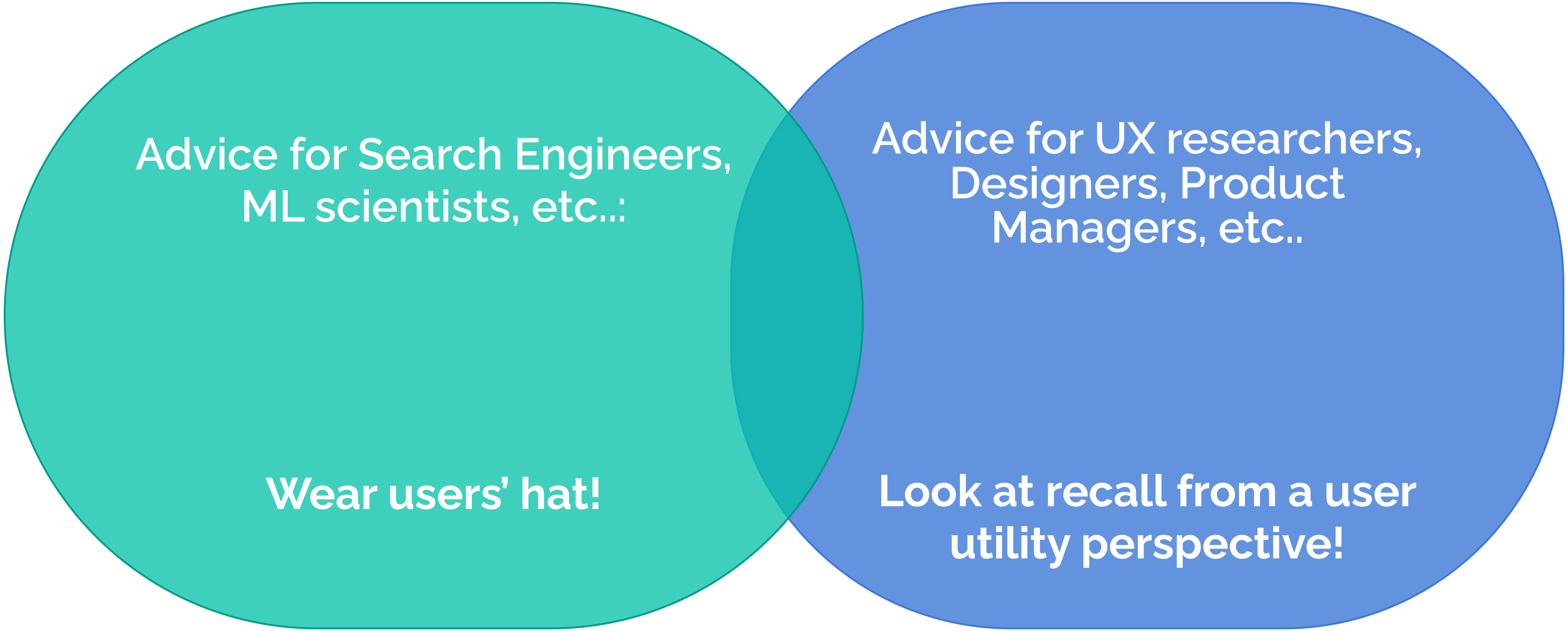
Case 2

You are 100% user centric 🥰, app reviews, NPS, user surveys is your obsession and no single user complains about "**search accuracy**" (or any fancy synonym).

=> Sounds like your search is highly focused on precision… Think about "utility" for your users. You might have huge opportunities if you improve recall

Neither case is ideal, both very likely corresponding to broken search systems 💔

Takeaways



Advice for Search Engineers, ML scientists, etc..:

Wear users' hat!

Advice for UX researchers, Designers, Product Managers, etc..

Look at recall from a user utility perspective!

An illustration with query specificity

Query specificity

🔍 fender 12432 results

🔍 fender stratocaster 4337 results

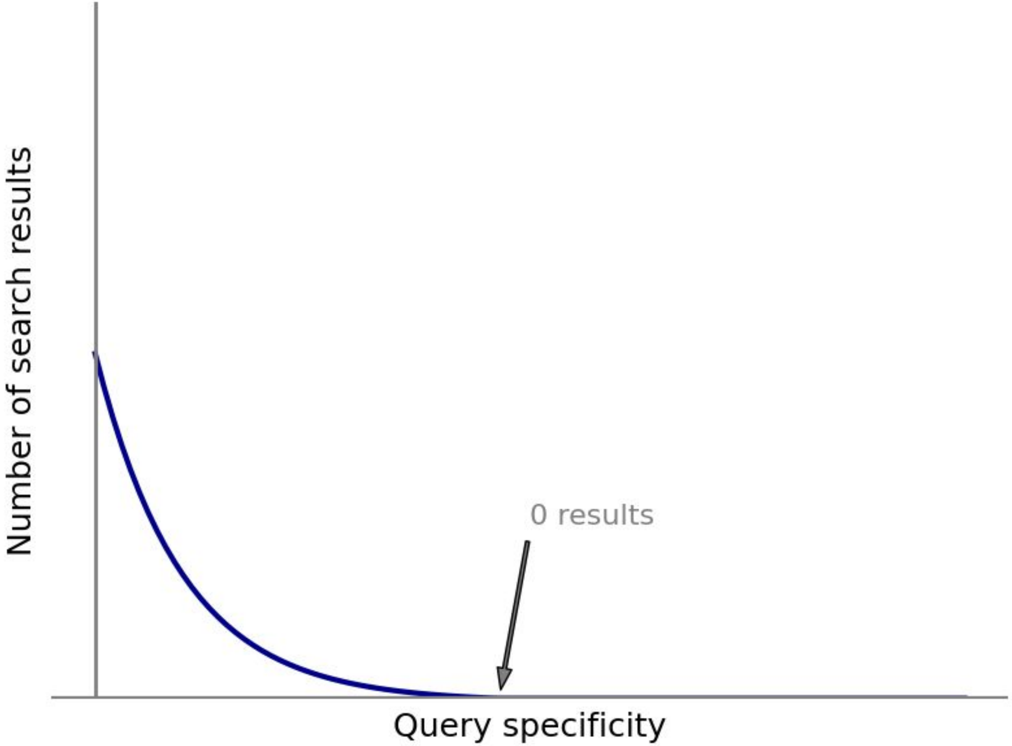
🔍 fender stratocaster custom shop 61 4 results

Intuition (& user expectations)*: each time you add a term, your query is becoming more specific and you should see less results

* still prevalent in ecommerce / marketplaces search. Natural Language queries are increasing but still not so frequent

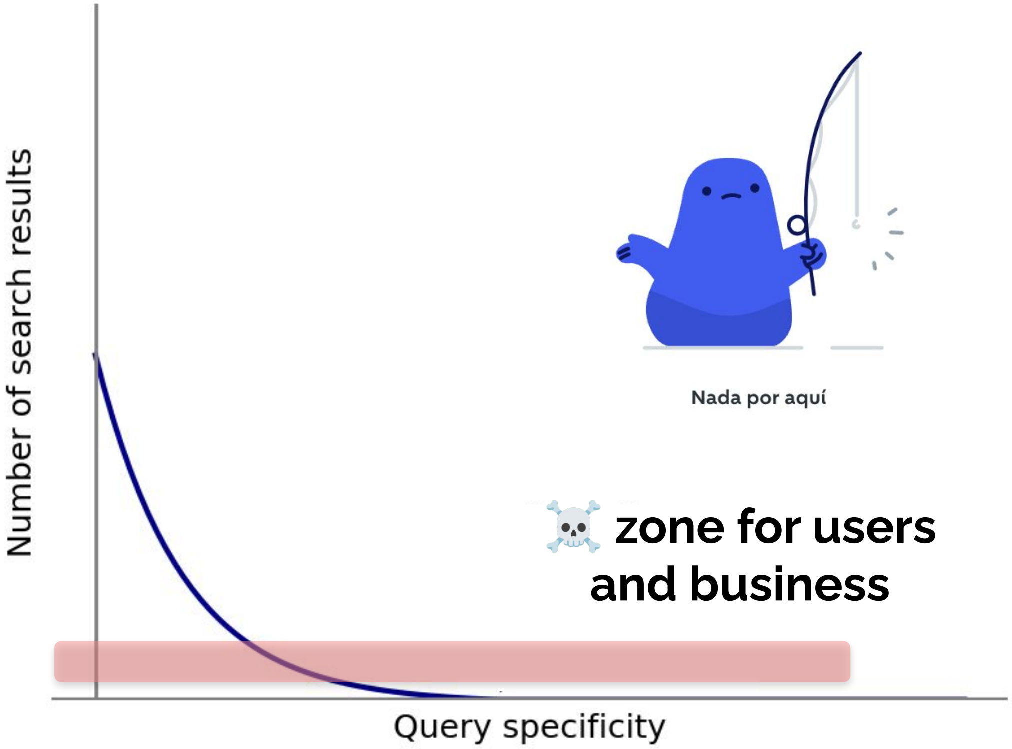
Query specificity

🔍 fender stratocaster custom shop 61



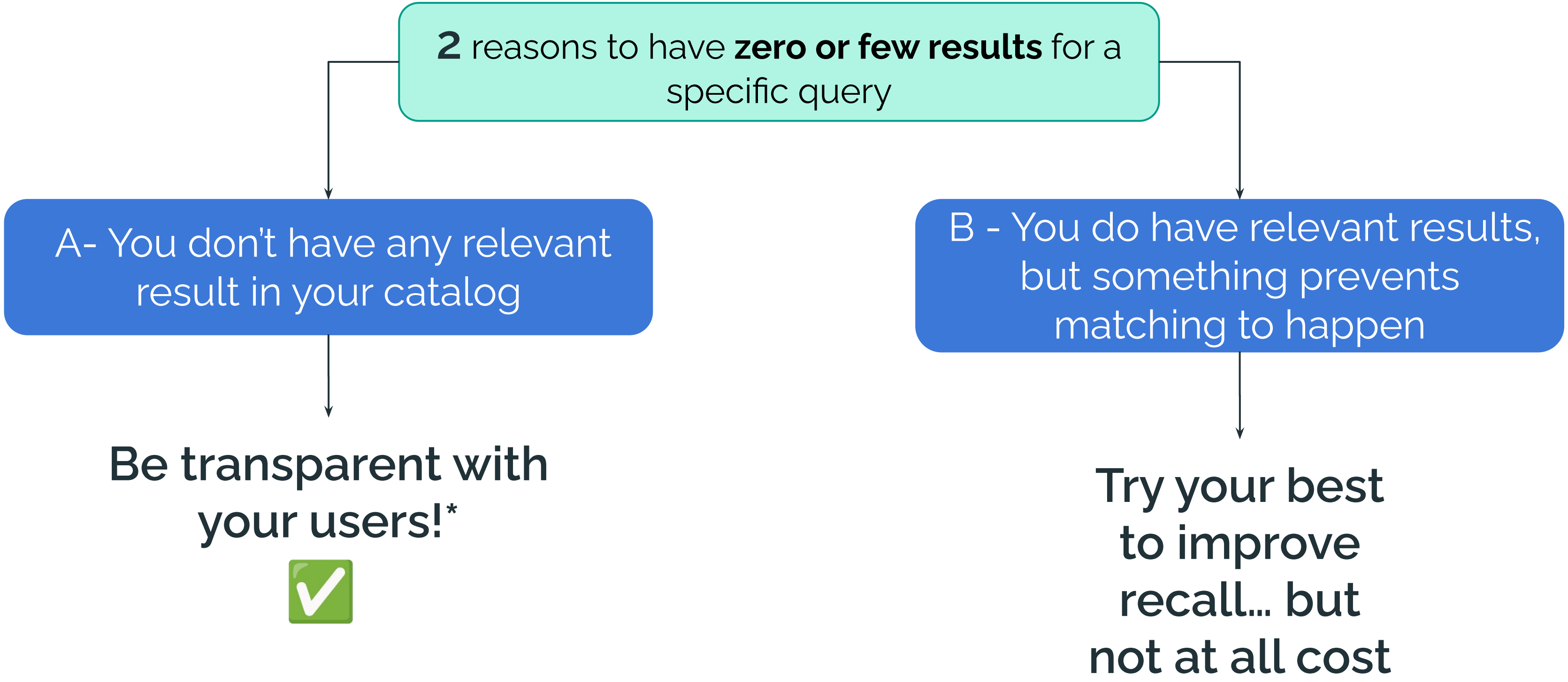
Query specificity

🔎 fender stratocaster custom shop 61



☠️ **zone for users and business**

Zero or few results



2 reasons to have **zero or few results** for a specific query

A- You don't have any relevant result in your catalog

Be transparent with your users!*

✅

B - You do have relevant results, but something prevents matching to happen

Try your best to improve recall… but not at all cost

*E. Pugh "When Zero Search Results is the right answer"

Matching failed

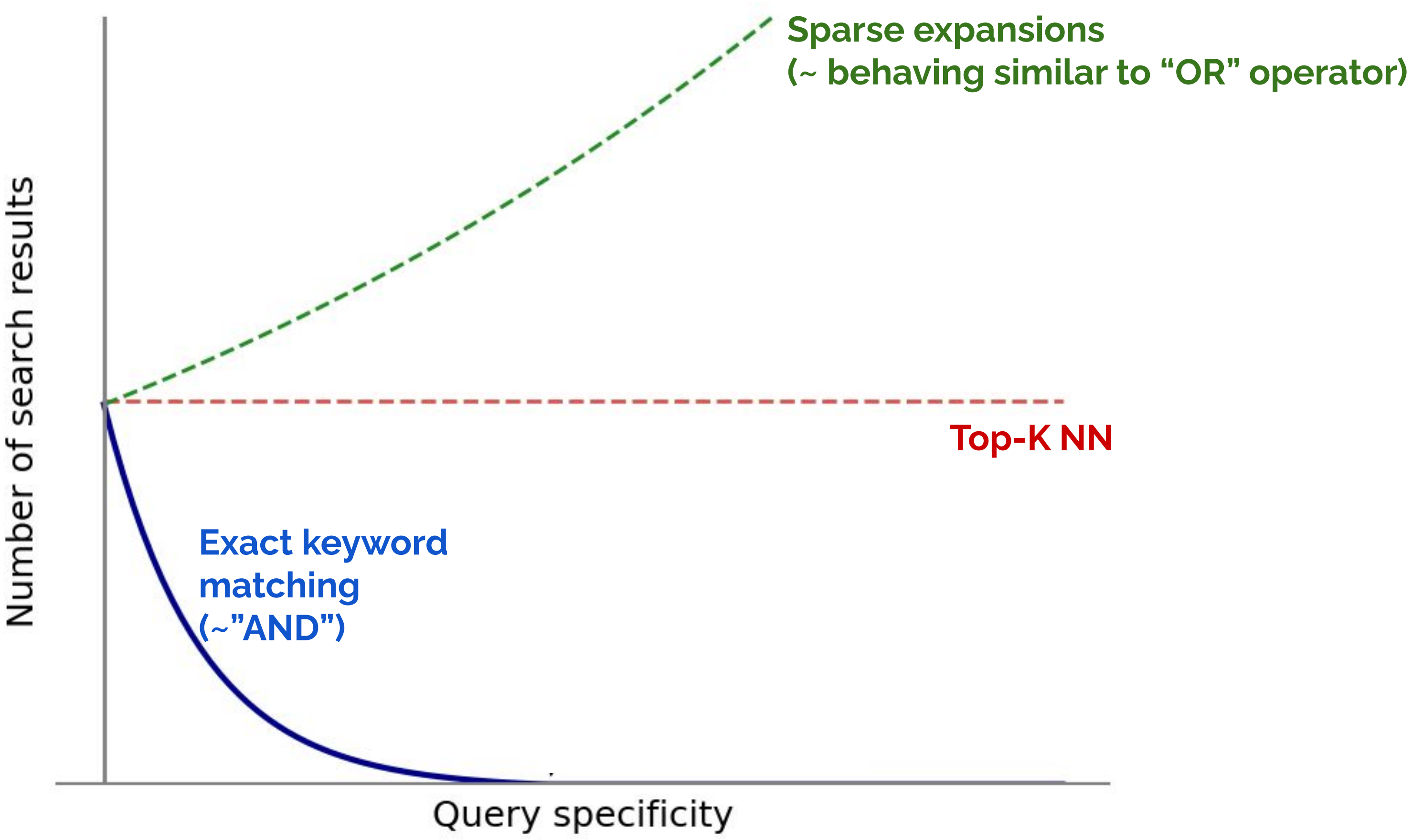
B - You do have relevant results, but something prevents matching to happen

Probability of this to happen is also increasing significantly with query length

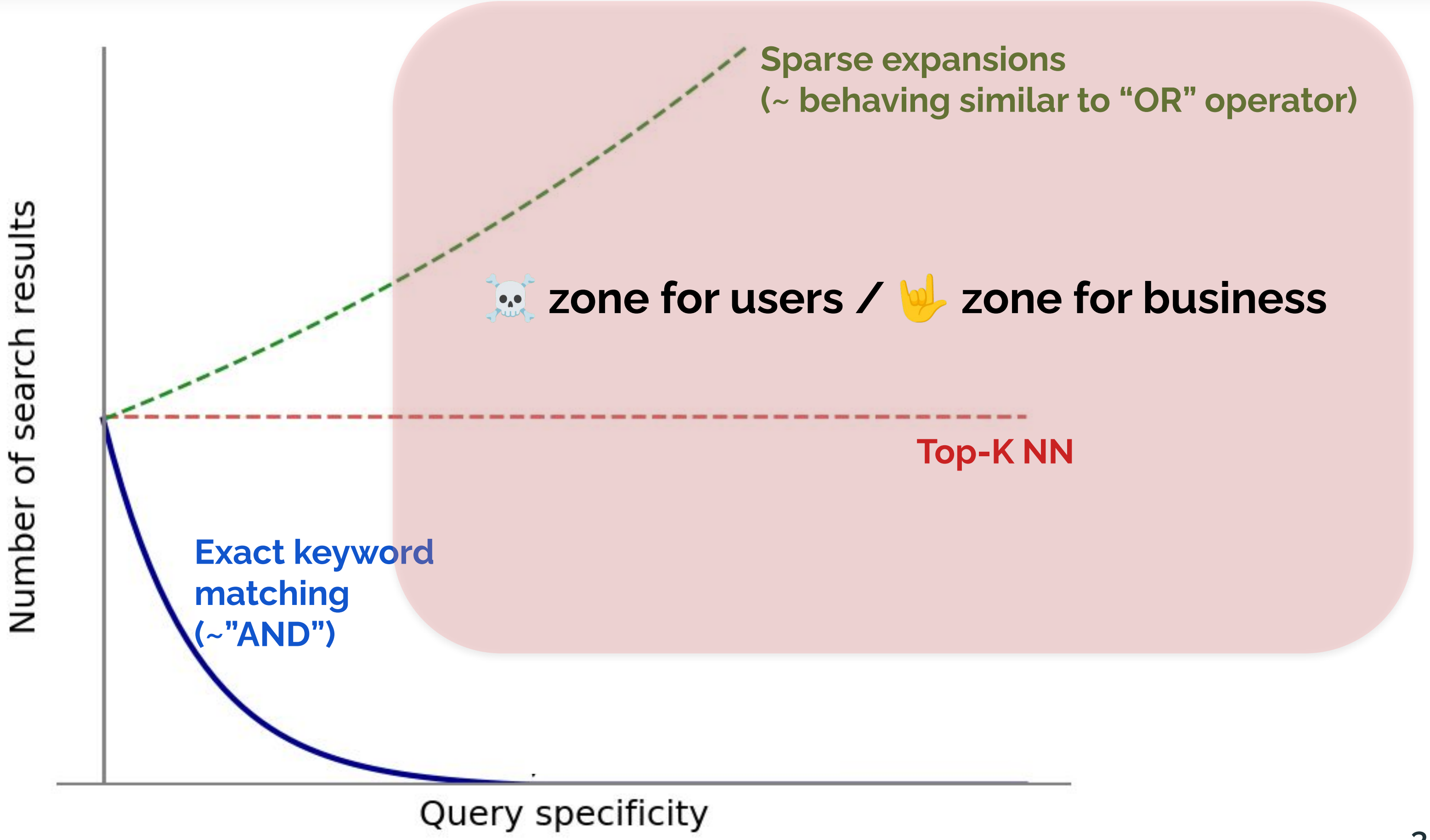
- Typos
- Synonyms
- Vocabulary gap
- Missing attributes in catalogue
- Over specification (e.g Car Volkswagen)

...

Query specificity in modern times



Query specificity in modern times

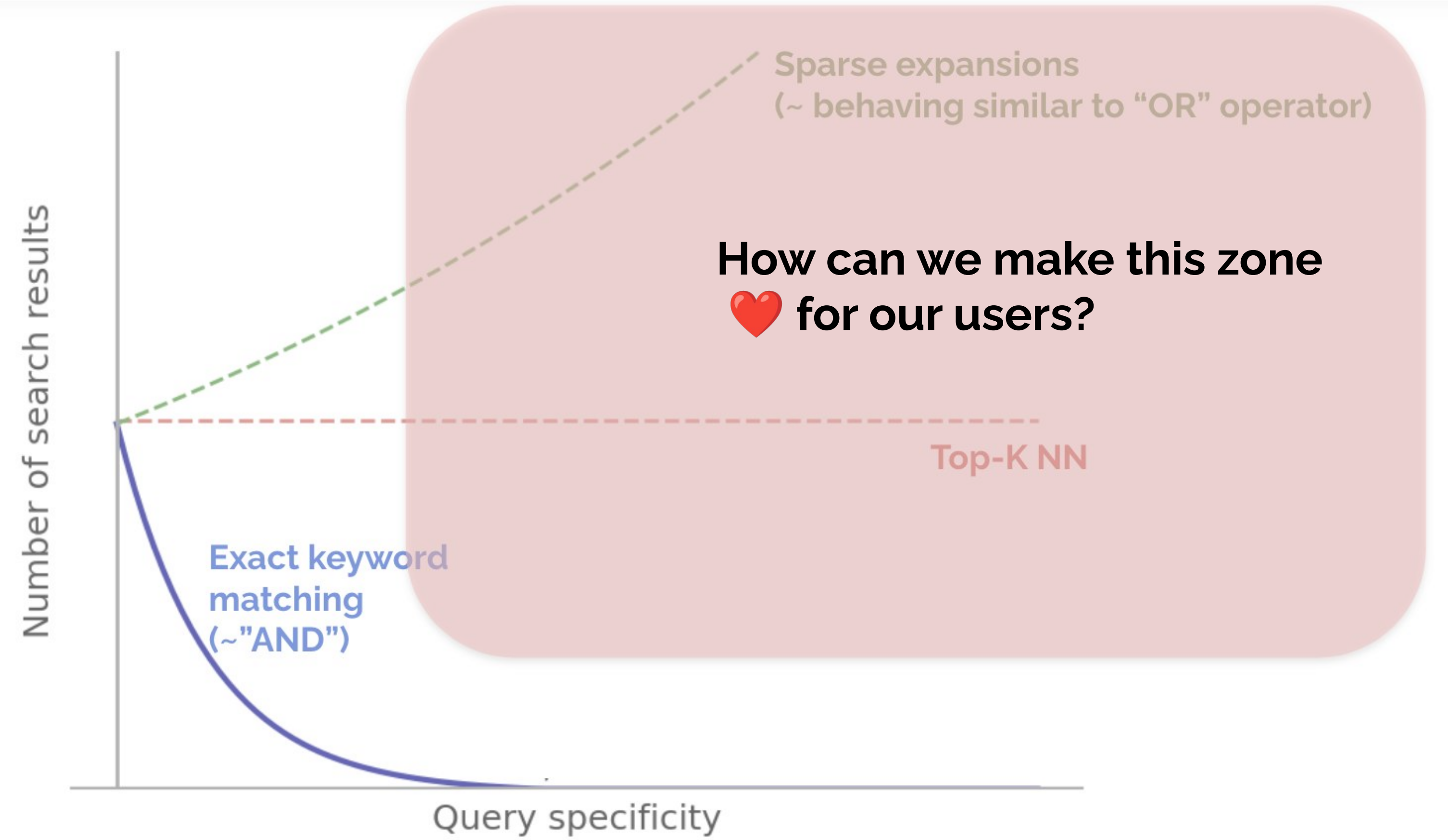


Takeaways

Aggressively improving recall for specific queries

- ✅ will quickly help you tackle zero / few results issues
- ✅ will likely bring impact to your business
- ❓ might not be so great for user perception

Next steps



How can we make this zone ❤️ for our users?

Solutions: How to improve recall while preserving user perception?

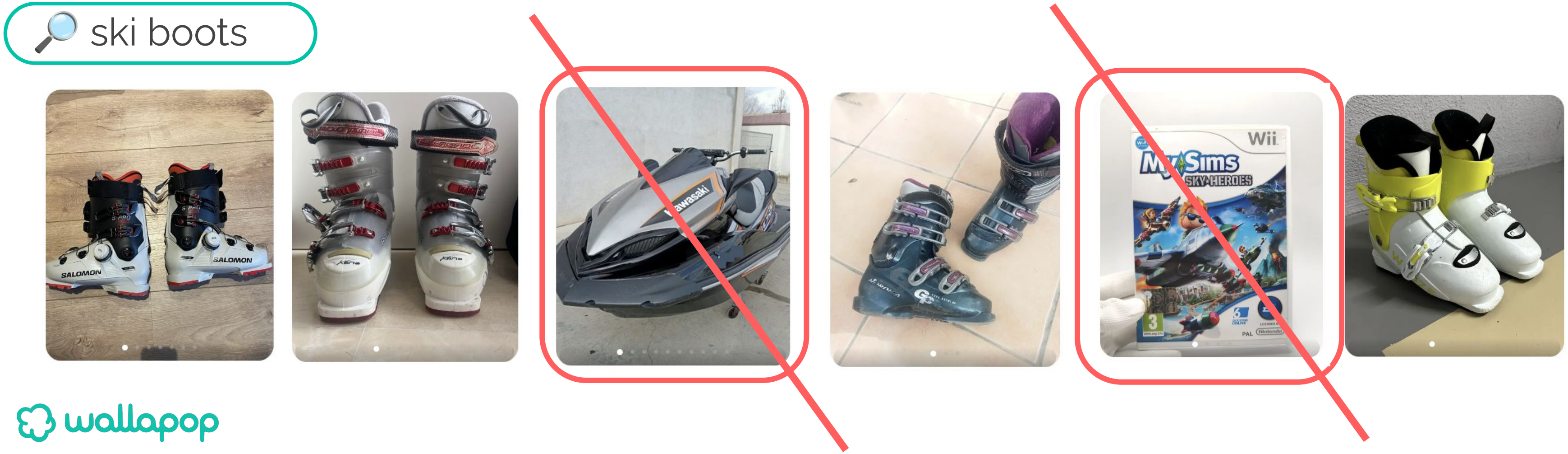
wallapop

Reduce risk of completely irrelevant results

Post-filtering

- "Relevance" depends on many factors, not just "***semantic*** relevance"
- Make sure to filter-out defects, i.e results that are clearly irrelevant semantically

✅ Reduces risk of bad user perception
✅ Less "random" result in other sorts
❌ How to set the trade-off?
❌ Query specificity behavior not fixed
❌ Promoted content?



Search system != retrieval & ranking algorithm

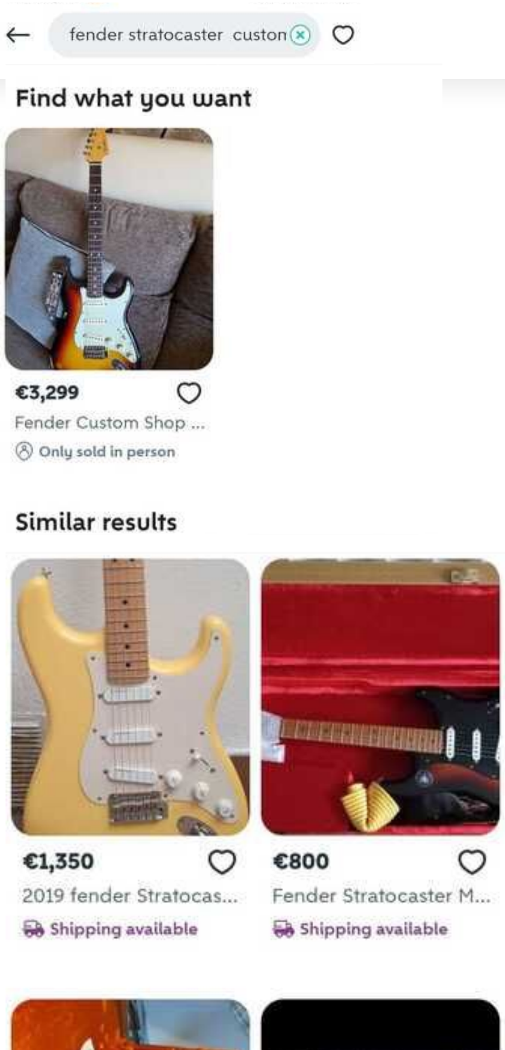
Think about adapting your UI

For example, have clear separation between high precision results and more recall oriented results.

Why ?

- ✅ 1st section has "query specificity" property again
- ✅ Also a good strategic change: recall section is less risky, you can try Vector Search / Personalization, etc..
- ✅ Other sorts will still give precise results! 🪄

- ❌ Adds complexity compared to 1 single result list (e.g duplications, position biases, etc..)
- ❌ Where to cut, what is precise enough ?



fender stratocaster custon

Find what you want

€3,299
Fender Custom Shop ...
Only sold in person

Similar results

€1,350
2019 fender Stratocas...
Shipping available

€800
Fender Stratocaster M...
Shipping available

Users are -not- all identical

Average user does not exist 👤 (but we often tune our search for this average user)

- Some users will prefer simplicity of search box (more natural language)
- Some users will always be hard core filter / sorts / facets users

Don't make assumptions about how they should use your search! Let them choose.

Give them possibility to your user to filter, e.g title only, exact match only, those are basic features, but could be game changer for some users. etc..

Summary

Takeaways

User metrics and Business metrics are **both** key when building search systems

- Business metrics ~ recall
- User perception ~ precision
- User perception **!=** user utility

Don’t blindly trust your business metrics!
Don’t blindly trust your user perception metrics!

Takeaways

There are (simple) ways we can ensure recall is optimized without compromising intuitiveness of the experience.

More recommendations:

- Nothing is set on stone, depending on maturity, priorities, etc..
- Experimenting, testing. measuring is key, always

Gracias!

Questions?

wallapop