



# Bypassing the Popularity Bias: Repurposing Models for Better Long-Tail Recommendation



**Václav Blahut**

Machine Learning Researcher, Seznam.cz, a.s.



**Karel Koupil**

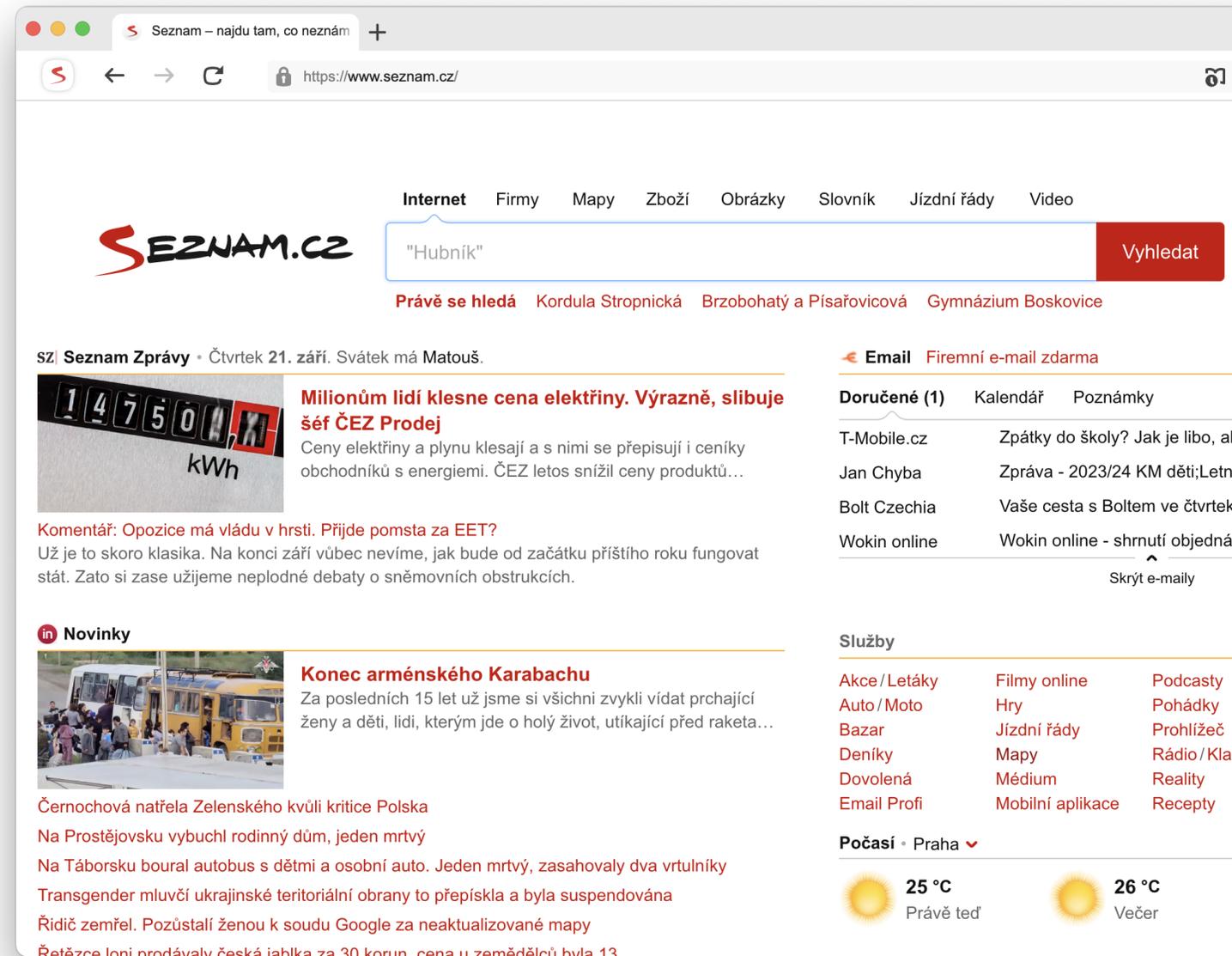
Product Manager, Seznam.cz, a.s.

7th FAccTRec Workshop on Responsible Recommendation at RecSys 2024  
October 14th, Bari, Italy



# Seznam?

- „The homepage of czech internet“
- Search engine, e-mail service, news and entertainment content publisher, video streaming platform, maps, internet browser, TV, radio...
- Private company
- Online advertisement platform
- Czech market



The screenshot shows the homepage of Seznam.cz in a web browser. The browser's address bar displays "https://www.seznam.cz/". The page features a navigation menu with links for "Internet", "Firmy", "Mapy", "Zboží", "Obrázky", "Slovník", "Jízdní řády", and "Video". A search bar contains the text "Hubník" and a red "Vyhledat" button. Below the search bar, a snippet of search results is visible, including "Kordula Stropnická", "Brzobohatý a Písařovicová", and "Gymnázium Boskovice". The main content area is divided into sections: "Seznam Zprávy" with a news article about electricity prices, "Novinky" with a news item about Karabach, and "Služby" with various service links. A weather widget at the bottom right shows "25 °C" and "26 °C".

Seznam.cz logo and navigation menu: Internet, Firmy, Mapy, Zboží, Obrázky, Slovník, Jízdní řády, Video

Search bar: "Hubník" [Vyhledat]

Právě se hledá: Kordula Stropnická, Brzobohatý a Písařovicová, Gymnázium Boskovice

Seznam Zprávy • Čtvrtek 21. září. Svátek má Matouš.

**Milionům lidí klesne cena elektřiny. Výrazně, slibuje šéf ČEZ Prodej**  
Ceny elektřiny a plynu klesají a s nimi se přepisují i ceníky obchodníků s energiemi. ČEZ letos snížil ceny produktů...

Komentář: Opozice má vládu v hrsti. Přijde pomsta za EET?  
Už je to skoro klasika. Na konci září vůbec nevíme, jak bude od začátku příštího roku fungovat stát. Zato si zase užijeme neplodné debaty o sněmovních obstrukcích.

Novinky

**Konec arménského Karabachu**  
Za posledních 15 let už jsme si všichni zvykli vídat prchající ženy a děti, lidem, kterým jde o holý život, utíkající před raketa...

Černochová natřela Zelenského kvůli kritice Polska  
Na Prostějovsku vybuchl rodinný dům, jeden mrtvý  
Na Táborsku boural autobus s dětmi a osobní auto. Jeden mrtvý, zasahovaly dva vrtulníky  
Transgender mluvčí ukrajinské teritoriální obrany to přepískla a byla suspendována  
Řidič zemřel. Pozůstalí ženou k soudu Google za neaktualizované mapy  
Řetězce loni prodávaly česká jablka za 30 korun, cena u zemědělců byla 13

Email: Firemní e-mail zdarma

Doručené (1) | Kalendář | Poznámky

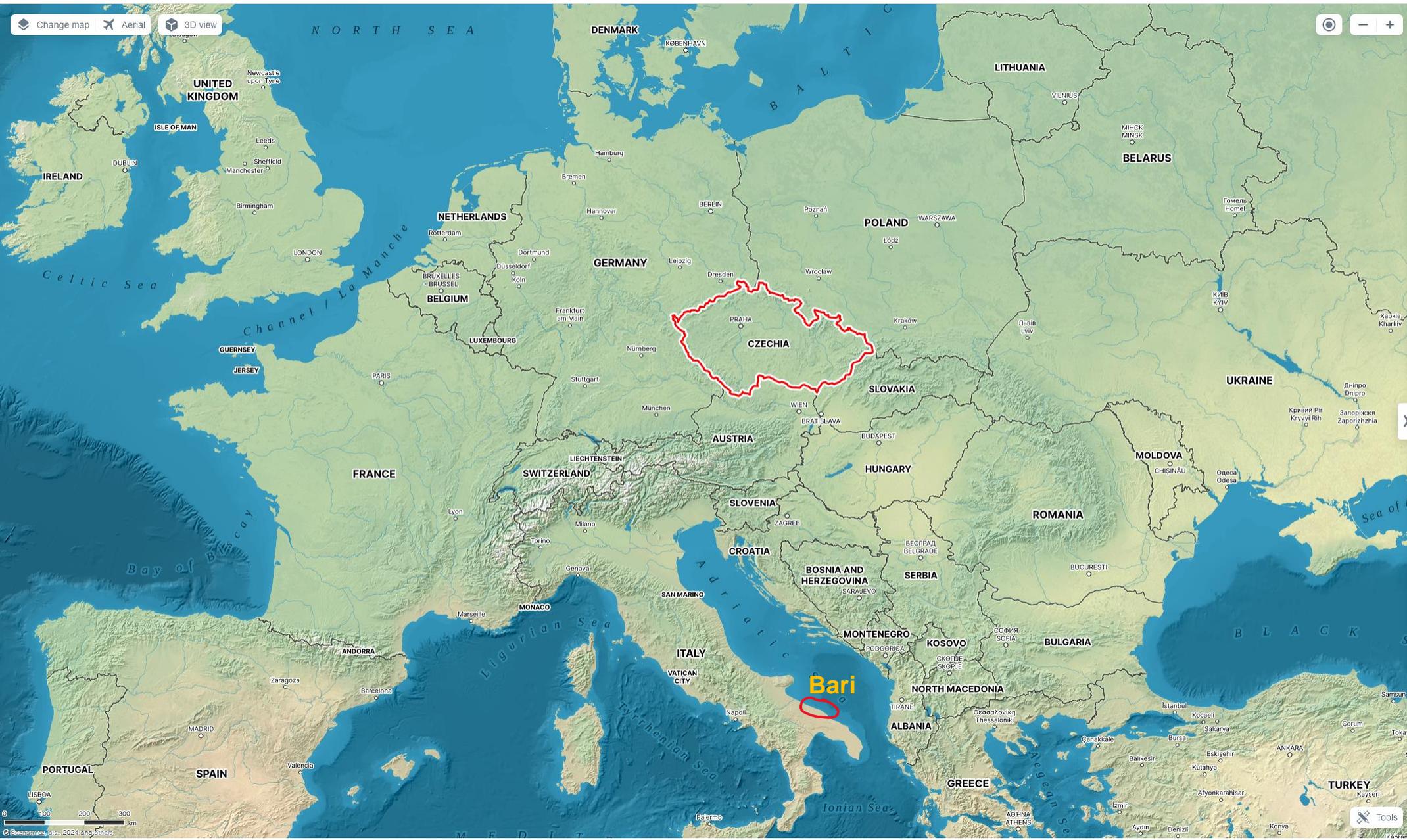
T-Mobile.cz | Zpátky do školy? Jak je líbo, a  
Jan Chyba | Zpráva - 2023/24 KM děti;Letn  
Bolt Czechia | Vaše cesta s Boltem ve čtvrtek  
Wokin online | Wokin online - shrnutí objedná  
Skrýt e-maily

Služby

Akce / Letáky | Filmy online | Podcasty  
Auto / Moto | Hry | Pohádky  
Bazar | Jízdní řády | Prohlížeč  
Deníky | Mapy | Rádio / Kla  
Dovolená | Médium | Reality  
Email Profi | Mobilní aplikace | Recepty

Počasí • Praha ▼

25 °C | Právě teď | 26 °C | Večer



Search Directions Log in My maps



**Czechia**  
 Česko  
 Country  
 Europe

Directions Save Share Travel tips

Find hotel nearby Booking.com

www.czechia.eu Suggest data modification

The Czech Republic, also known as Czechia, is a landlocked country in Central Europe. Historically known as Bohemia, it is bordered by Austria to the south, Germany to the west, Poland to the northeast, and Slovakia to the southeast. The Czech Republic has a hilly landscape that covers an area of 78871 km2 with a mostly temperate continental and oceanic climate.

[Wikipedia](#)

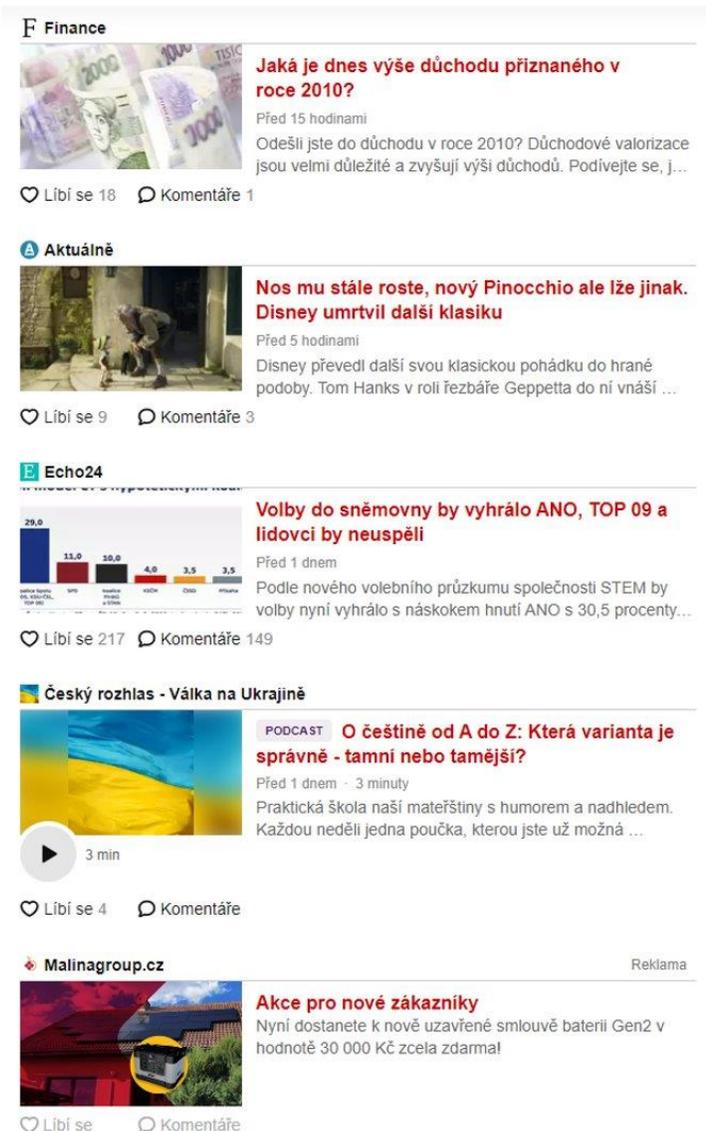
Capital:	Prague
Currency:	Czech koruna
Country calling code:	+420
Population:	10,900,555
Area:	78,866 km²
Head of state:	Petr Pavel
Head of government:	Petr Fiala
Basic form of government:	parliamentary republic
Official language:	Czech

Data sources

GPS Tools

# Seznam Newsfeed

- Personalized, infinite feed of content
- News, entertainment articles, videos, podcasts, photos, user comments...
- In-house produced content
- **External partner content**

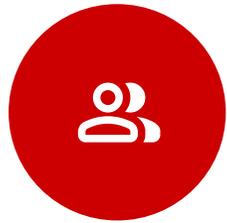


The screenshot displays a newsfeed with five items:

- Finance:** Article titled "Jaká je dnes výše důchodu přiznaného v roce 2010?". Includes a photo of Euro banknotes and a play button icon.
- Aktuálně:** Article titled "Nos mu stále roste, nový Pinocchio ale lze jinak. Disney umrtvil další klasiku". Includes a photo of a man and a dog.
- Echo24:** Article titled "Volby do sněmovny by vyhrálo ANO, TOP 09 a lidovci by neuspěli". Includes a bar chart showing election results.
- Český rozhlas - Válka na Ukrajině:** Podcast article titled "O češtině od A do Z: Která varianta je správně - tamní nebo tamější?". Includes a photo of the Ukrainian flag and a play button icon.
- Malinagroup.cz:** Advertisement titled "Akce pro nové zákazníky". Includes a photo of a red building and a yellow car.

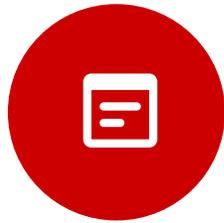


# Seznam Newsfeed in numbers



**1.8 M**

daily active users



**2 000**

daily published items



**450**

publishers

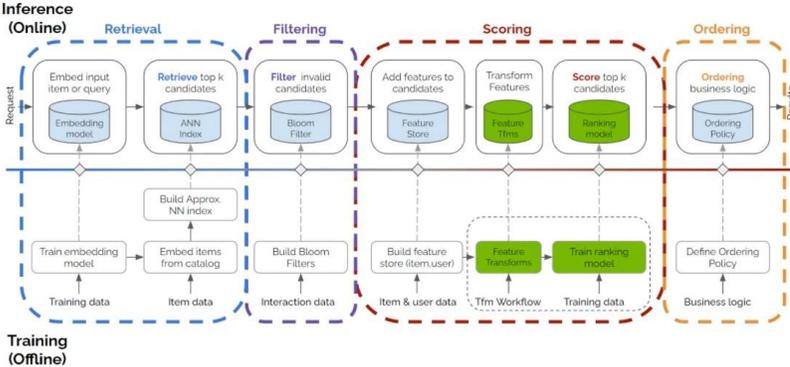
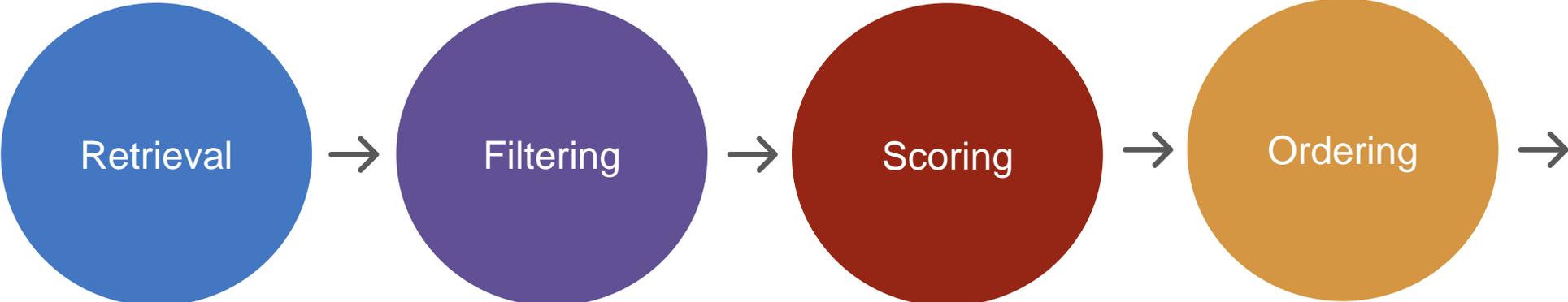


**3 000**

peak requests per second



# 4-stage recommender system



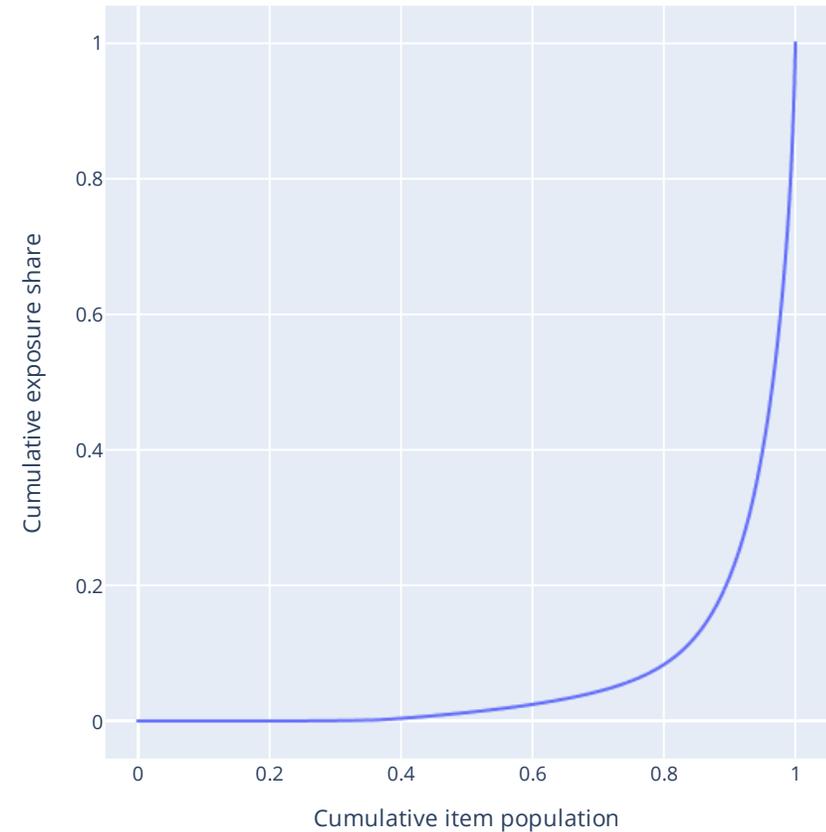
Source: Karl Higley, Even Oldridge, Ronay Ak, Sara Rabhi, and Gabriel de Souza Pereira Moreira. 2022. Building and Deploying a Multi-Stage Recommender System with Merlin. In Proceedings of the 16th ACM Conference on Recommender Systems (RecSys '22). Association for Computing Machinery, New York, NY, USA, 632–635. <https://doi.org/10.1145/3523227.3551468>



**Problem:  
Popularity bias  
and long-tail content exposure**

# Publisher and item exposure

- Item exposure – # of visible impressions
- Popularity bias ~ long-tail items
- Publisher diversity is important

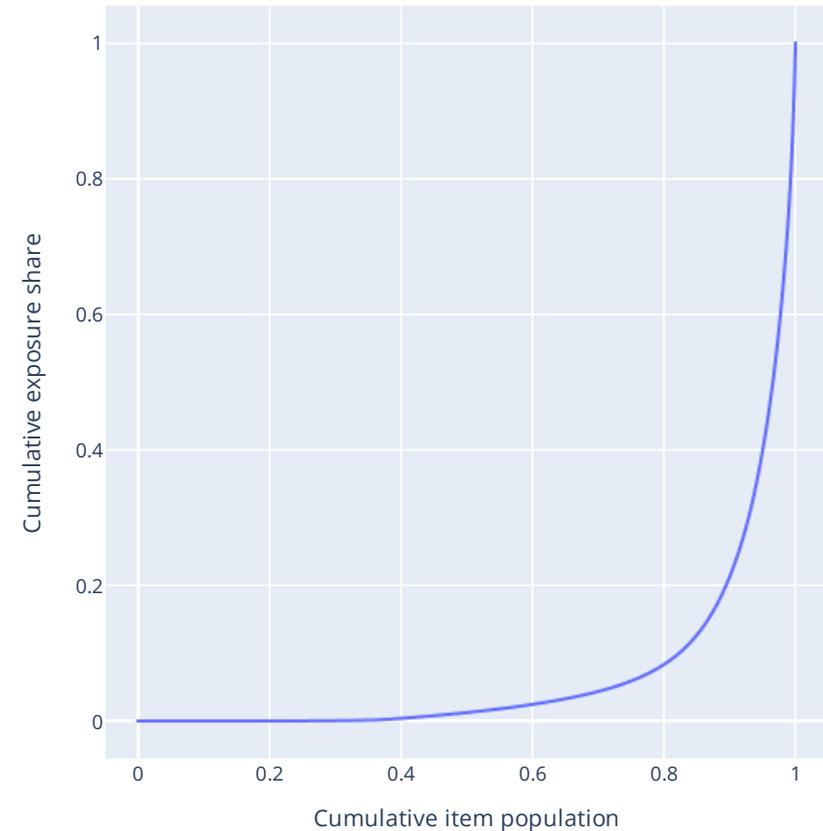


# Publisher and item exposure

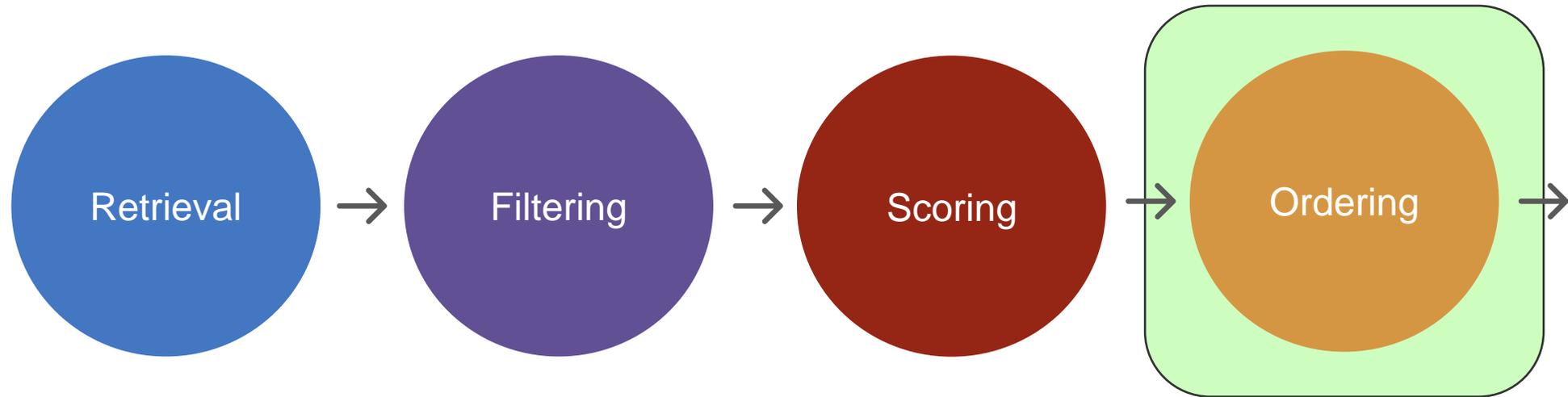
- Item exposure – # of visible impressions
- Popularity bias ~ long-tail items
- Publisher diversity is important

**Goal:**

**Deliver relevant exposure to unfairly underexposed items**



# Randomization after ranking



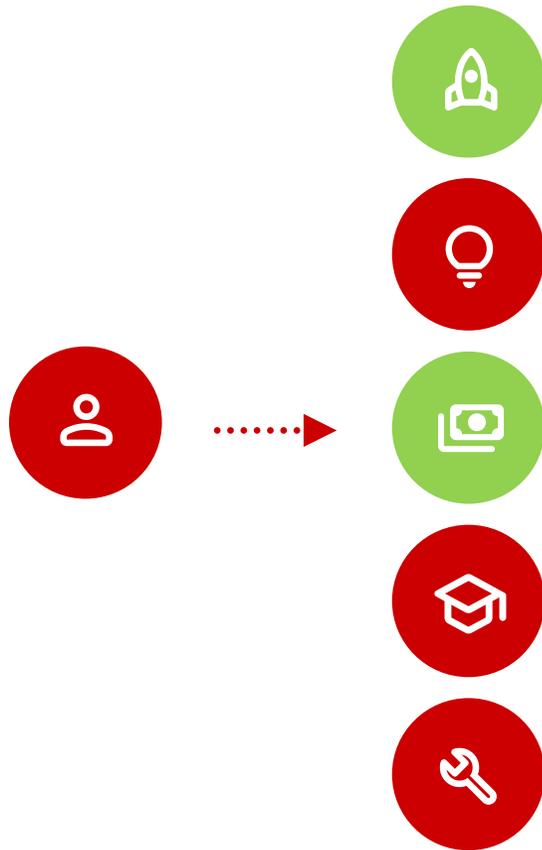
Source: Karl Higley, Even Oldridge, Ronay Ak, Sara Rabhi, and Gabriel de Souza Pereira Moreira. 2022. Building and Deploying a Multi-Stage Recommender System with Merlin. In Proceedings of the 16th ACM Conference on Recommender Systems (RecSys '22). Association for Computing Machinery, New York, NY, USA, 632–635. <https://doi.org/10.1145/3523227.3551468>



**Proposed solution:  
Inverse recommendation**



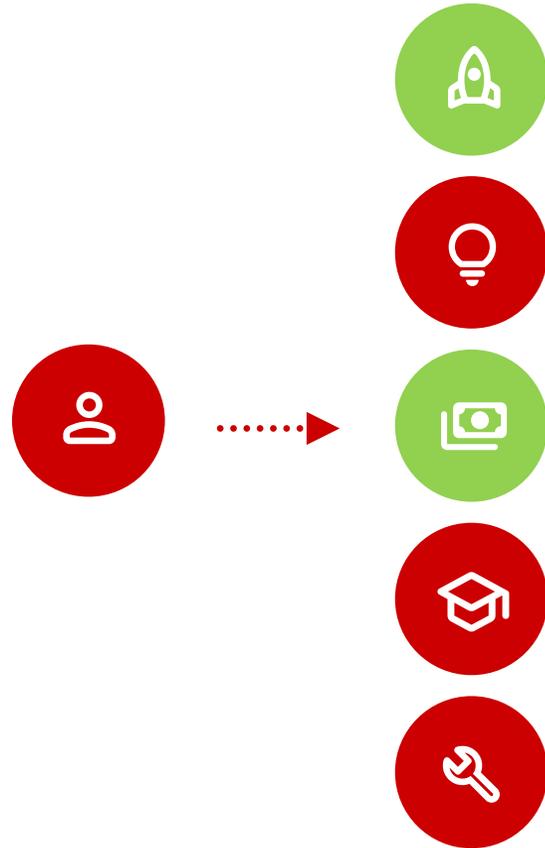
# Traditional recommendation



Top items for a user

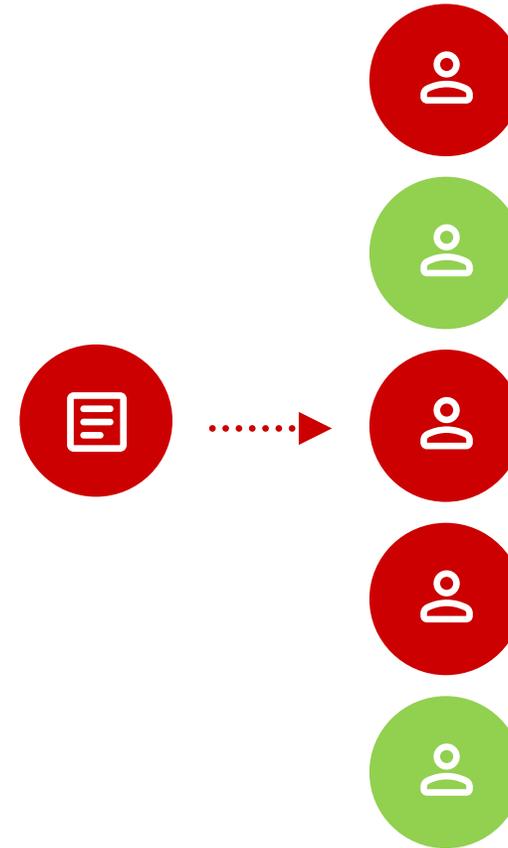


# Traditional recommendation



Top items for a user

# Inverse recommendation a.k.a. reverse k-MIPS [0]

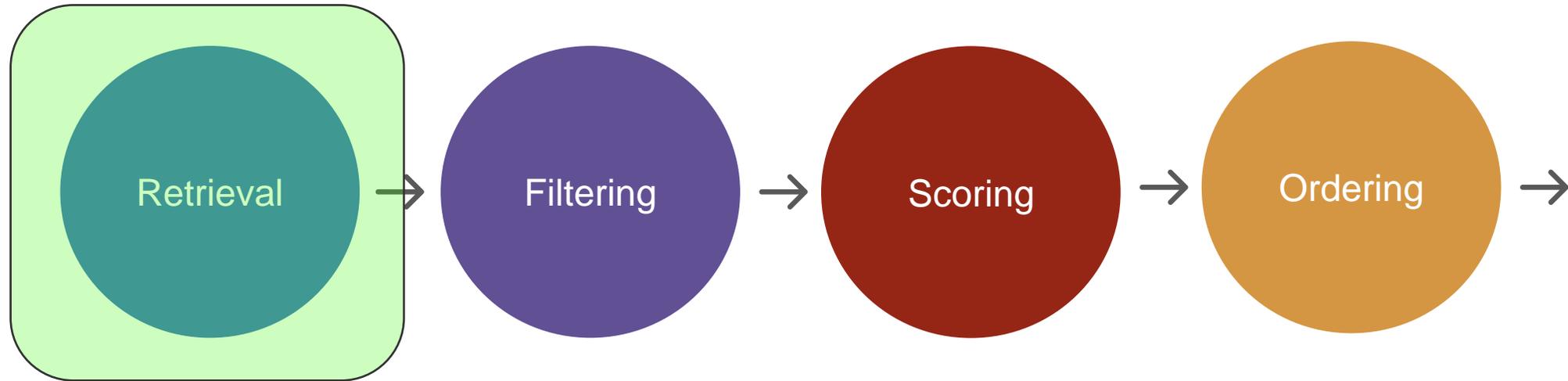


Top users for an item

[0] Daichi Amagata and Takahiro Hara. 2021. Reverse Maximum Inner Product Search: How to efficiently find users who would like to buy my item? In Proceedings of the 15th ACM Conference on Recommender Systems (RecSys '21). Association for Computing Machinery, New York, NY, USA, 273–281. <https://doi.org/10.1145/3460231.3474229>



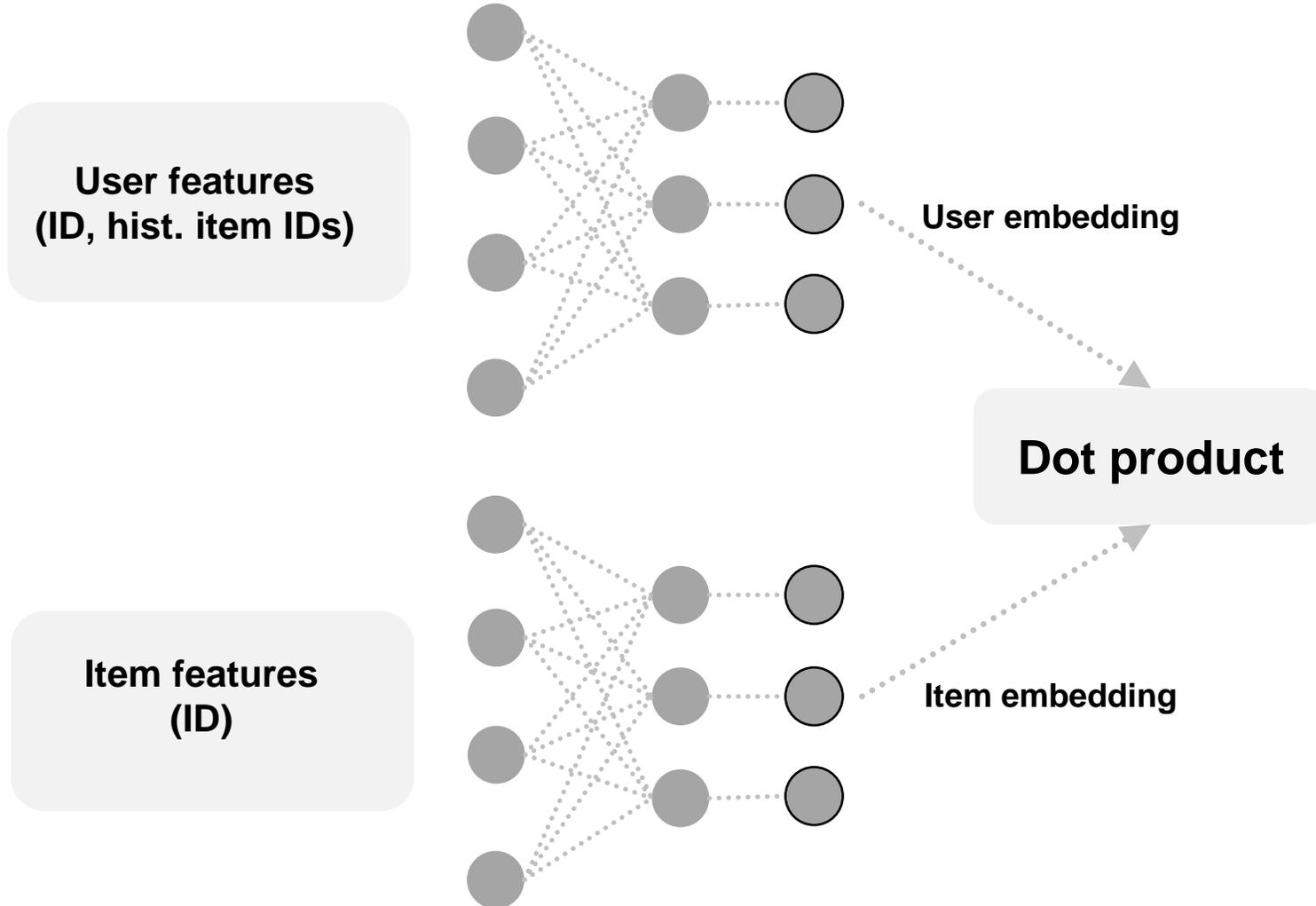
# Retrieval model



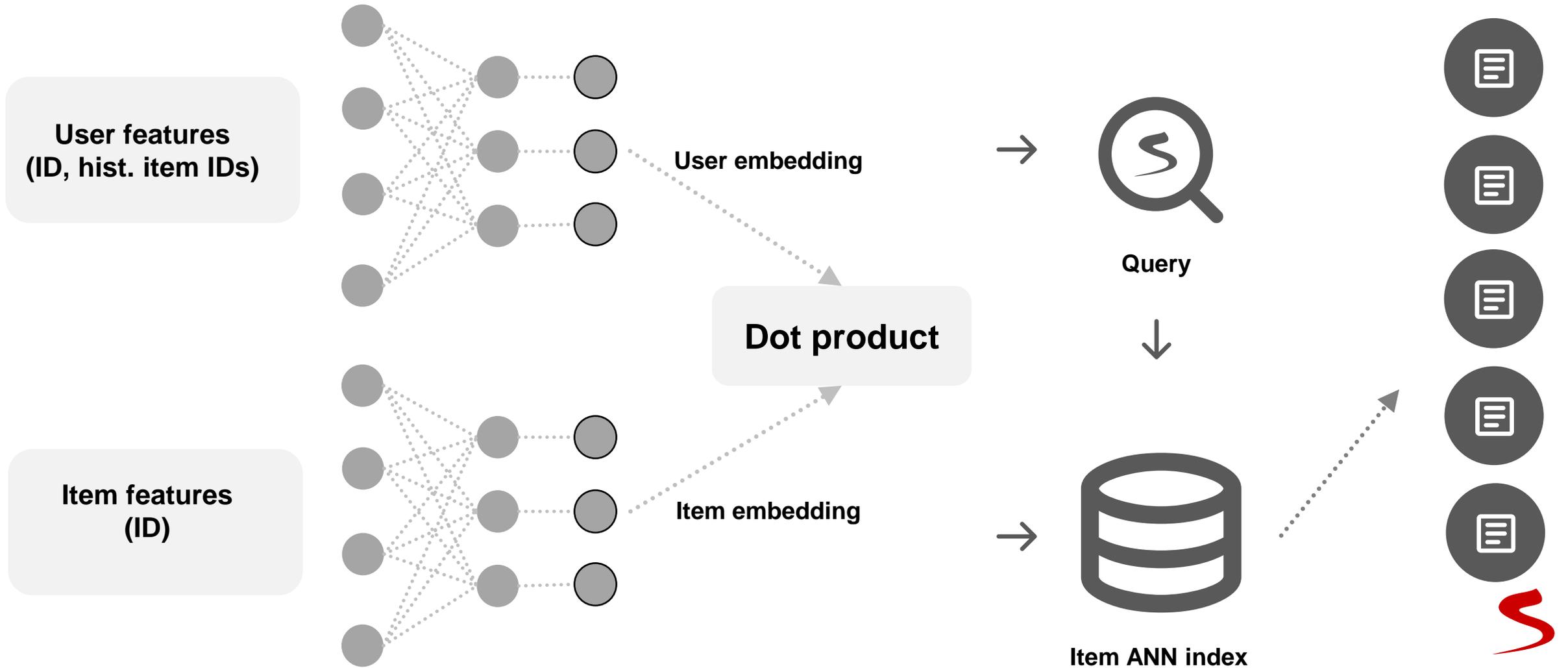
Source: Karl Higley, Even Oldridge, Ronay Ak, Sara Rabhi, and Gabriel de Souza Pereira Moreira. 2022. Building and Deploying a Multi-Stage Recommender System with Merlin. In Proceedings of the 16th ACM Conference on Recommender Systems (RecSys '22). Association for Computing Machinery, New York, NY, USA, 632–635. <https://doi.org/10.1145/3523227.3551468>



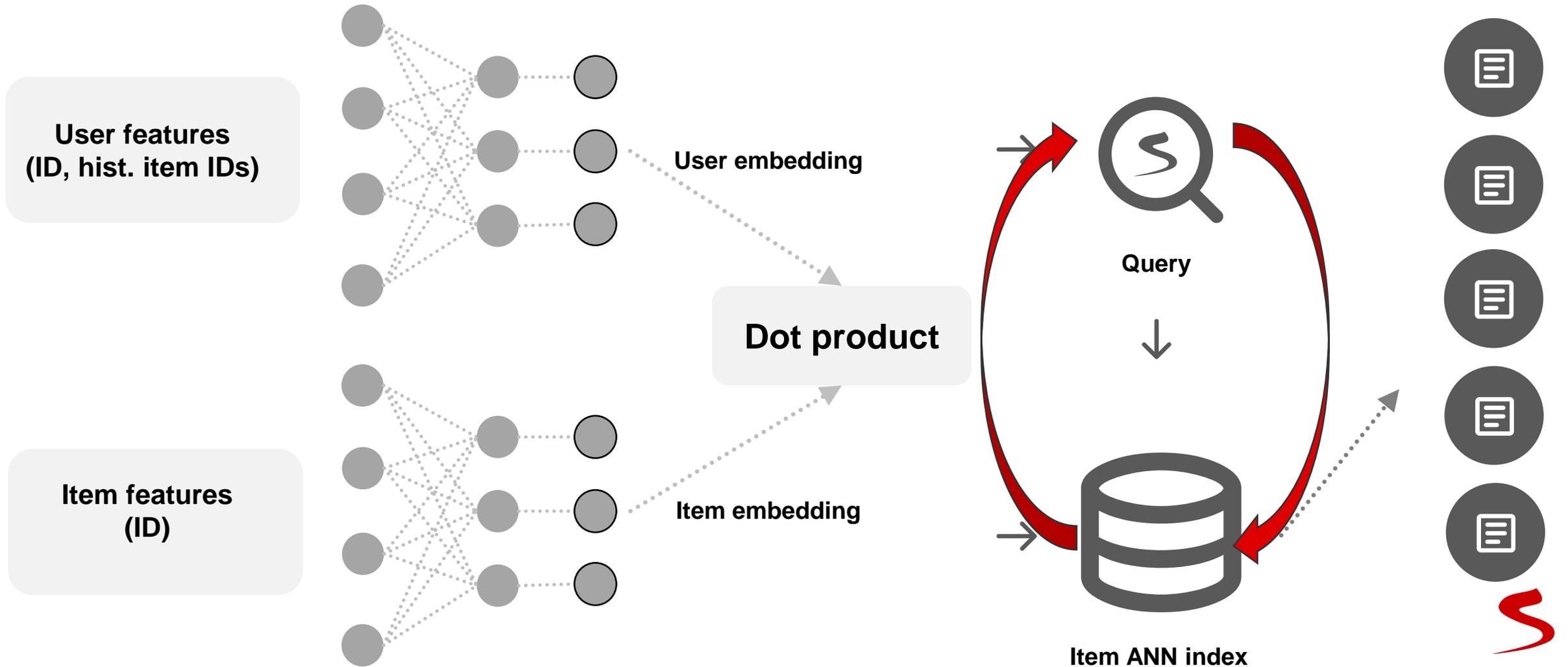
# Two-tower retrieval model



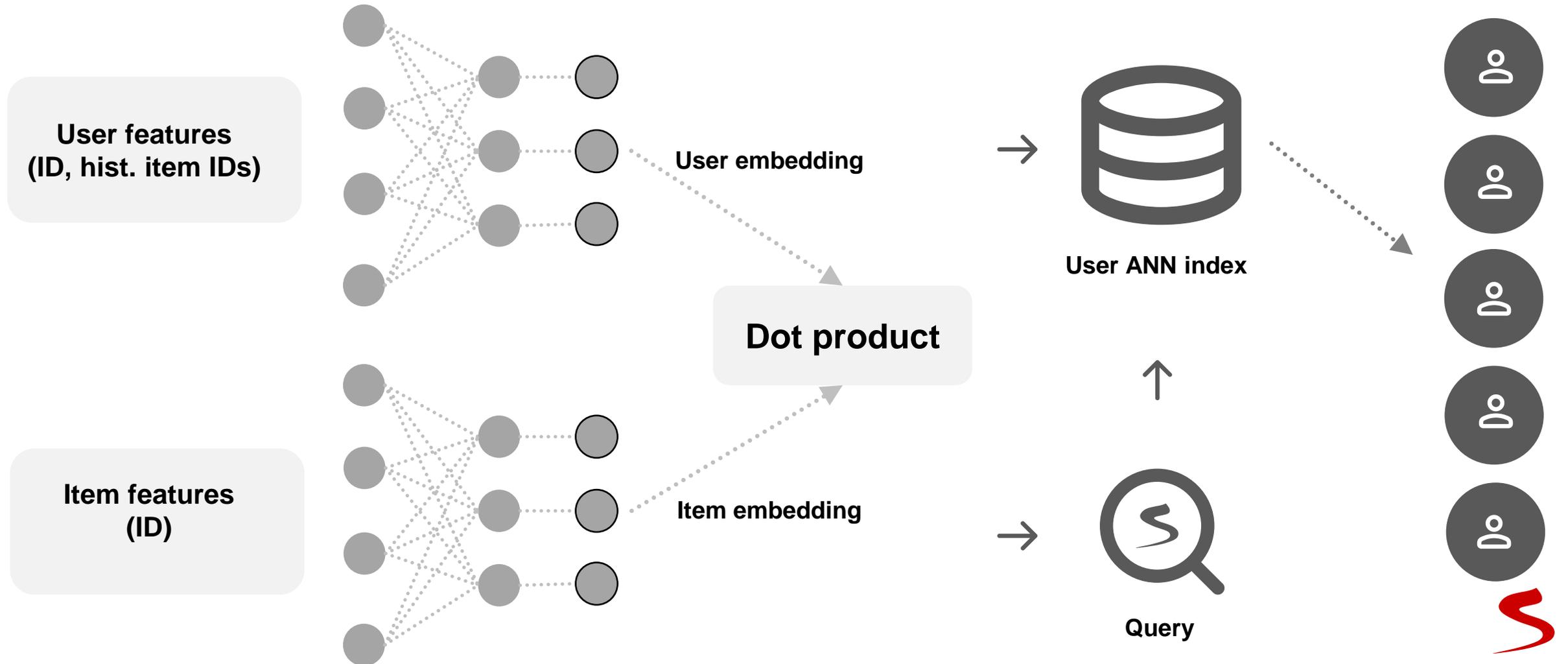
# Two-tower retrieval model



# Two-tower retrieval model



# Two-tower *inverse* retrieval model (InvR)



# Results ordering

- Multiple items per user
- **Score sort**
  - Use the similarity (dot product) value
  - Biased



# Results ordering

- Multiple items per user
- **Score sort**
  - Use the similarity (dot product) value
  - Biased
- **Random sort**
  - Most fair?



# Results ordering

- Multiple items per user
- **Score sort**
  - Use the similarity (dot product) value
  - Biased
- **Random sort**
  - Most fair?
- **User rank sort**
  - Use the rank of user w.r.t. item



# Results ordering

- Multiple items per user
- **Score sort**
  - Use the similarity (dot product) value
  - Biased
- **Random sort**
  - Most fair?
- **User rank sort**
  - Use the rank of user w.r.t. item



score	user rank
0.51	1 924
<b>0.85</b>	5 124
0.46	<b>265</b>
0.64	945
0.28	1 324



# Incorporating in recommendations

- Offline InvR pipeline
- Inserted into slate after business logic



Source: Karl Higley, Even Oldridge, Ronay Ak, Sara Rabhi, and Gabriel de Souza Pereira Moreira. 2022. Building and Deploying a Multi-Stage Recommender System with Merlin. In Proceedings of the 16th ACM Conference on Recommender Systems (RecSys '22). Association for Computing Machinery, New York, NY, USA, 632–635. <https://doi.org/10.1145/3523227.3551468>



# Business-level hyperparameters

- Set of treated publishers
- Minimum exposure
- Slate positions
- Capping



# Evaluation



# A/B testing

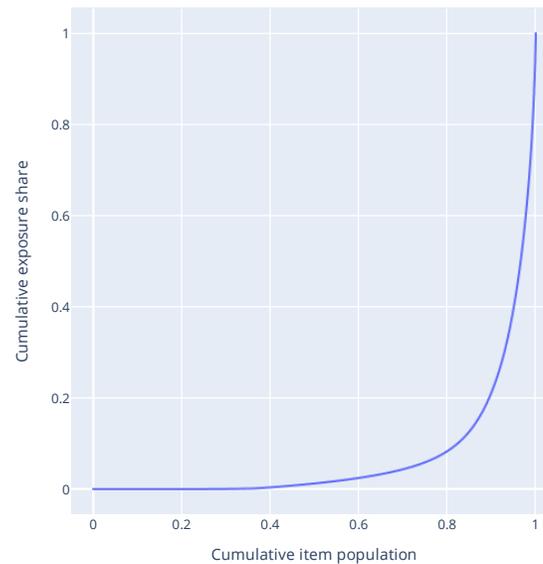
- Randomized A/B tests on live traffic
- A/A variants included
- Empirically sufficient variant sizes and duration
- Randomization unit = user



# Metrics

## Fairness-oriented

- Bottom 50 % share (B50PS)
  - Whole item set
- Percentage of sufficiently exposed items (PSEI)
  - Treated items only
- Top 1 % share (T1PS)
  - Whole item set



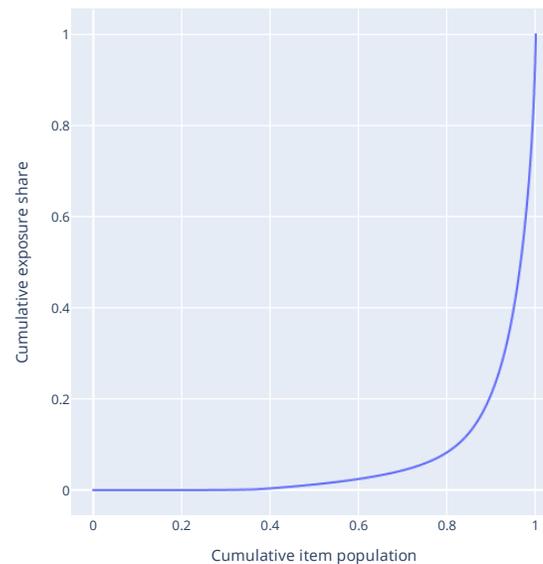
# Metrics

## Fairness-oriented

- Bottom 50 % share (B50PS)
  - Whole item set
- Percentage of sufficiently exposed items (PSEI)
  - Treated items only
- Top 1 % share (T1PS)
  - Whole item set

## KPIs (InvR-only)

- CTR
  - per user
- Clicks
  - per user



# Variants

- Baseline
  - No InvR
- Random
  - **Ablation study**
  - Pick users for the item randomly
- Score sort
- Random sort
- User rank sort



# Results



# Main results

Variant	B50PS	PSEI	T1PS	CTR (InvR)	Clicks (InvR)
Baseline	0 %	0 %	0 %	-	-
Random	+9.2 %	<b>+181 %</b>	-1.0 %	0 %	0 %
InvR Random	+9 %	+43 %	-0.5 %	+271 %	+106 %
InvR Score	+9.9 %	+41 %	-0.9 %	+261 %	+103 %
InvR User rank	<b>+33.3 %</b>	+45 %	<b>-1.6 %</b>	<b>+300 %</b>	<b>+120 %</b>

All reported changes are relative.

B50PS – Bottom 50 % share

PSEI – Percentage of sufficiently exposed items

T1PS – Top 1 % share

CTR (InvR) – click-through rate per user, InvR recommendations only

Clicks (InvR) – clicks per user, InvR recommendations only



# Cost of deployment

InvR User rank variant

**-1.04 %**

CTR per user

**-1.67 %**

clicks per user

All reported changes are relative.



# Side-effects

**+18 %**

unique items recommended  
by main recommender

All reported changes are relative.



# Long-term experience



- Only 40 % of min. exposure generated by InvR



# Considerations, limitations, future work

- Cold-start item embeddings
  - Employ content-based solution instead of IDs
- Popularity bias in user set
  - Truncated user representation
- Model hyper-opt
- Treatment of all items
- User-centered analysis



# Thank you!

# Questions?

Paper at:



<https://arxiv.org/abs/2410.02776>



**Václav Blahut**

Machine Learning Researcher, Seznam.cz, a.s.  
vaclav.blahut@firma.seznam.cz



**Karel Koupil**

Product Manager, Seznam.cz, a.s.

Copyright © 1996–2024 Seznam.cz, a. s.

The logo for Seznam.cz, featuring a stylized red 'S' followed by the text 'EZNAM.CZ' in black. In the bottom right corner of the slide, there is also a decorative image of a light-colored, bushy animal tail.