I (*don't*) know what you did last summer

A roadmap in session-based inference

November, 17th 2021

Jacopo Tagliabue

Director of A.I.

coveo

360° relevance in Commerce, Service, Website and Workplace

$325M
Capital raised since 2018
for R&D, growth and acquisitions

1,500
Customer Deployments
globally & across multiple use cases

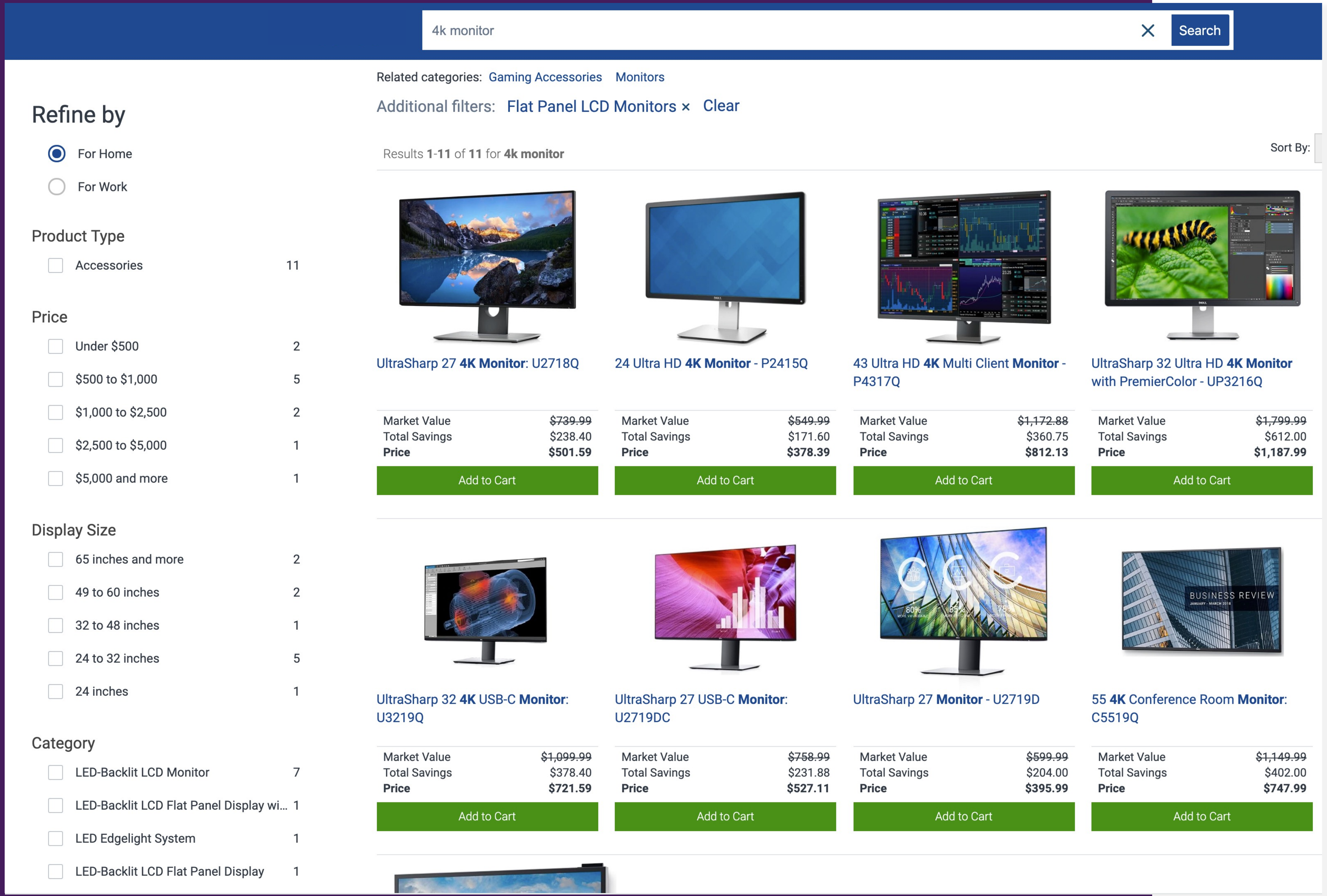
150
Certified SI Partners And Integrations
+strategic alliances & integrations with key ISVs

#1
US/Gartner/Forrester leader
FORRESTER® Gartner.
#1 applied AI platform company in Canada

Global
7 offices, 4 Data Centers around the world
for scale, performance, local compliance

ROI
Proven Customer Success
measured business value

A typical integration for ecommerce



- Search
- Query suggestions
- Recommendations
- Category listing

Significant research roadmap in AI, IR, NLP

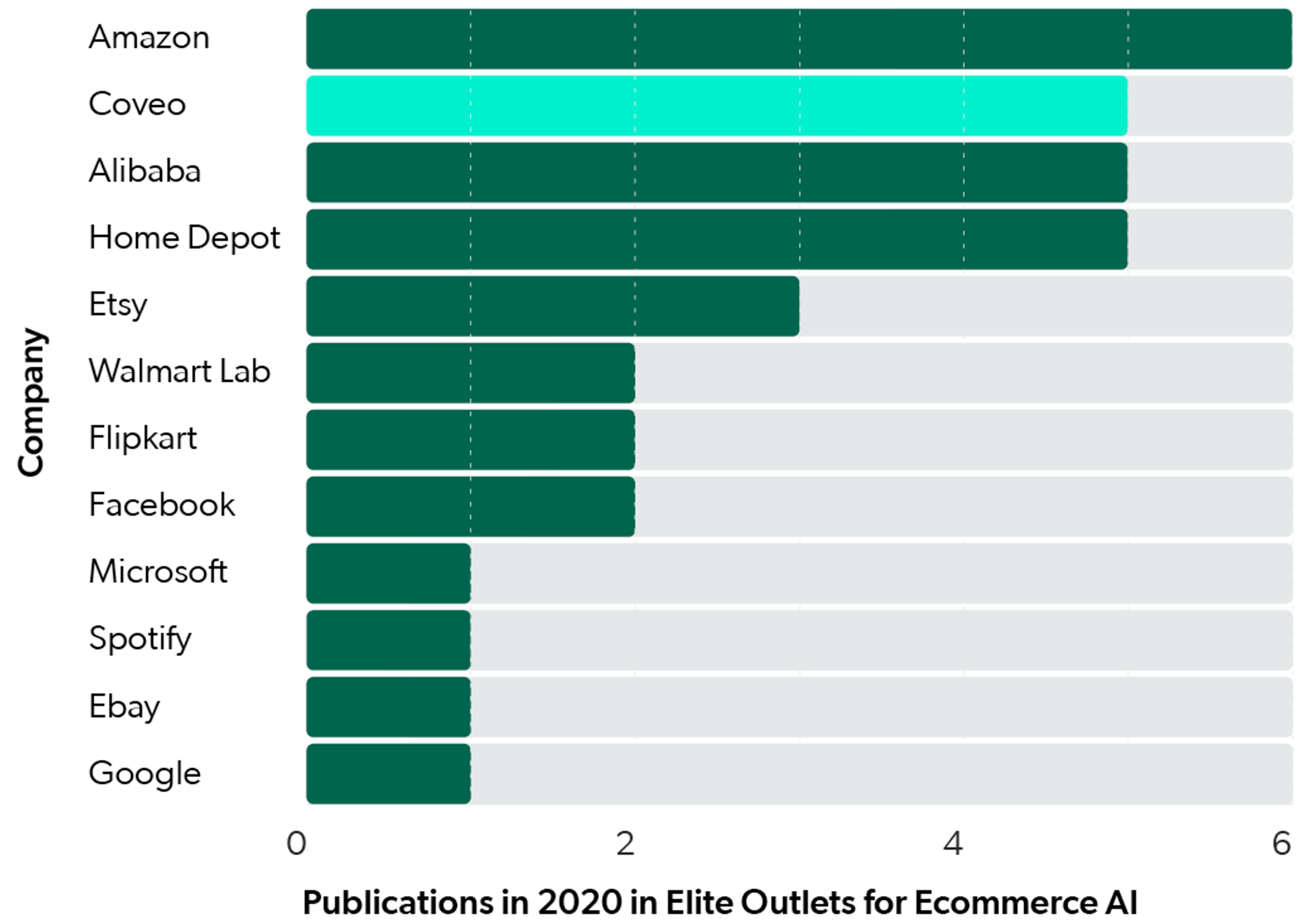
🥳Congratulations to the recipients of the Best Industry Paper Award at #NAACL2021 Industry Track!

Query2Prod2Vec: Grounded Word Embeddings for eCommerce

Federico Bianchi @fb_vinid , Jacopo Tagliabue @jacopotagliabue, Bingqing Yu

Best Industry Paper Award

Congratulations to the recipients of the Best Industry Paper Award at NAACL 2021 Industry Track!



Find the odd one out!

Session-based inference in e-commerce: a case study in applied research

Some hard facts about e-commerce

1

High bounce rates

Around 40%-50% of users leave a website after viewing a single page.

2

Small recurrent user base

Less than 10% of users come back more than 3 times in 12 months.

coveo™

Some hard facts about e-commerce

1

High bounce rates

Around 40%-50% of users leave a website after viewing a single page.

2

Small recurrent user base

Less than 10% of users come back more than 3 times in 12 months.

The model really does NOT know what you did last summer!

coveo™

Constraints for any solution

1

Move fast

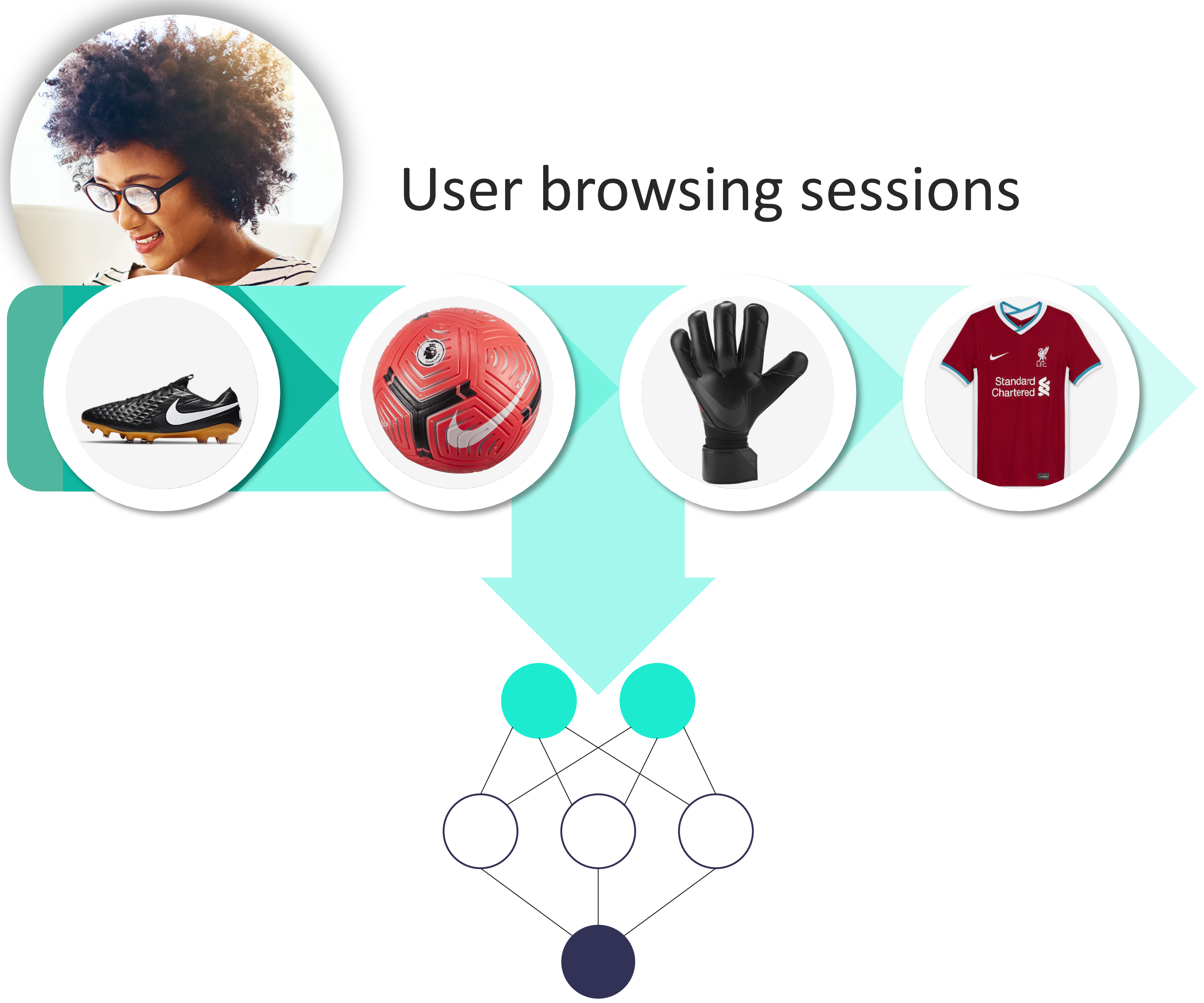
Personalization need to happen *as soon* as possible and with *as little* data as possible.

2

Stay in the pocket

A shopping session becomes the natural boundary for our ML models to work effectively.

A research program for session-based inference



Session representation

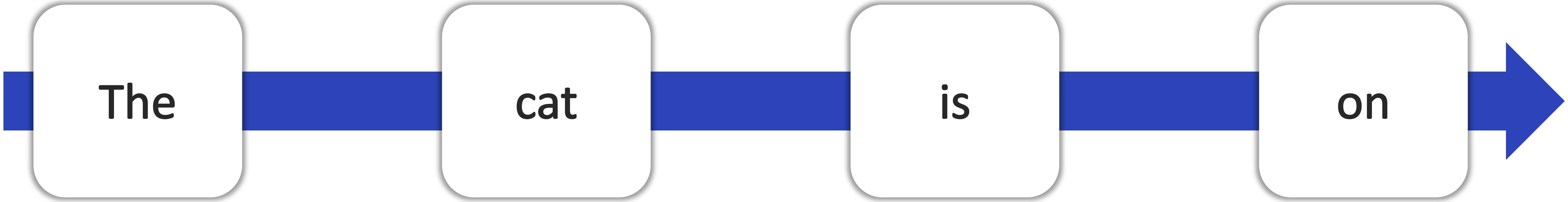
List of functionalities

- Search
- Recommendations
- Query suggestion
- Causal Attribution
- Intent Detection

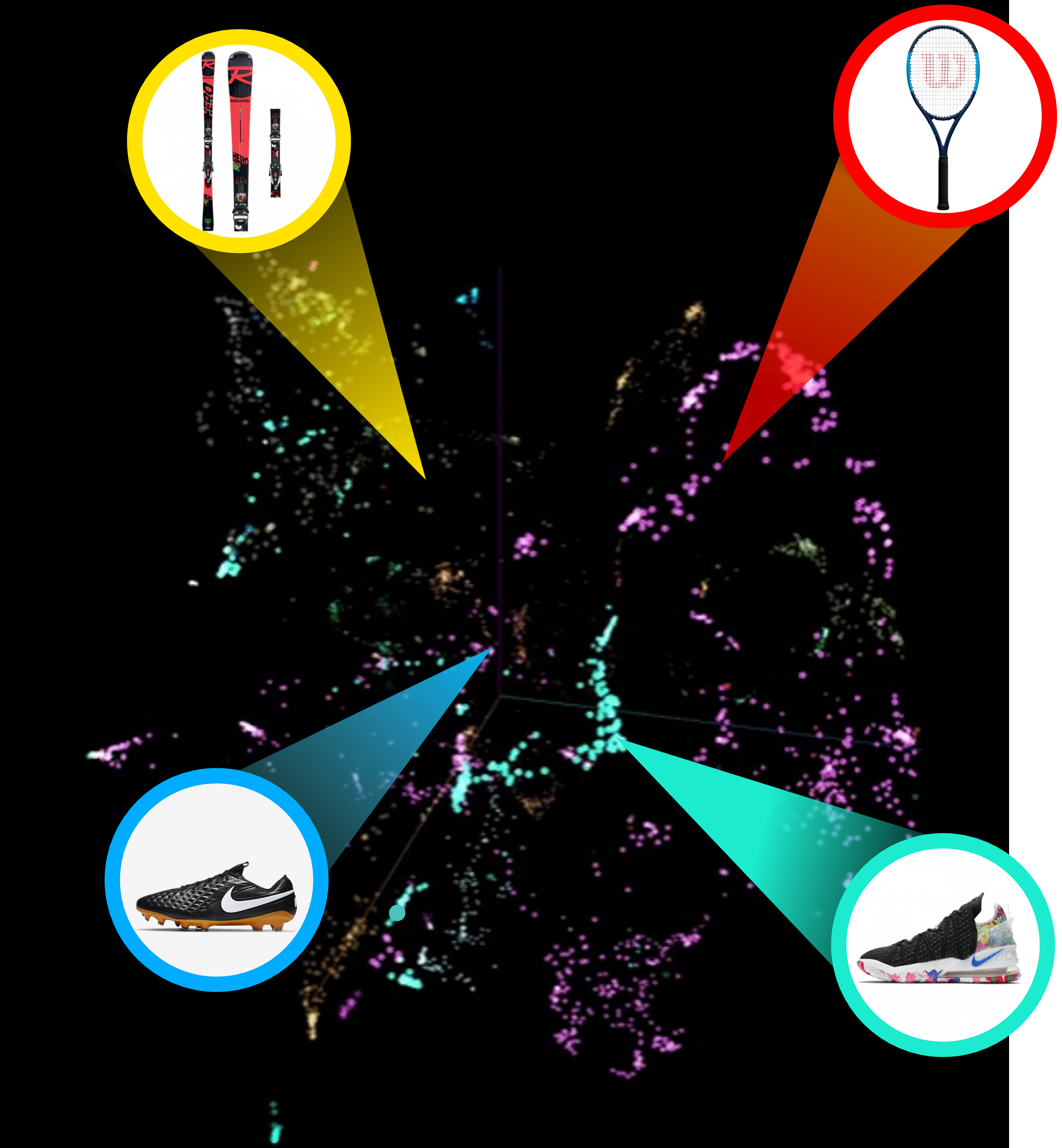
Modeling user intent with product embeddings

Product embeddings

word2vec



prod2vec



The product space

- In-session intent is represented as the products users interact with within a session.
- Products are represented as a multi-dimensional vector space: similar products that are “closer” in the space.
- Building such a space can be done in a purely unsupervised manner.

Session representation

- Session vectors are functions of the product vectors shoppers interact with:

$$SV = f(\, p_1\_v,\, p_2\_v,\, \ldots\, p_n\_v)$$

[f can be the (unweighted/weighted) average, or something more complex]

coveo™

Session-based personalization

- Different users walk in different regions of the space.

Fantastic product embeddings and how to align them

Training product embeddings

- Embedding model is a CBOW with negative sampling.
- Implementation is done with Gensim as a Python library.
- Hyper-parameters need special optimization for this use case.

However:

- Proper quantitative and qualitative validation procedure are needed.
- Representation of low-count items may be sub-optimal (i.e. cold-start).

Check your product embeddings

QUANTITATIVE

- Standard NEP (Next Event Prediction) task: given a session of *n* interactions, take the first *n-1* and predict the n^{th} (kNN / LSTM).

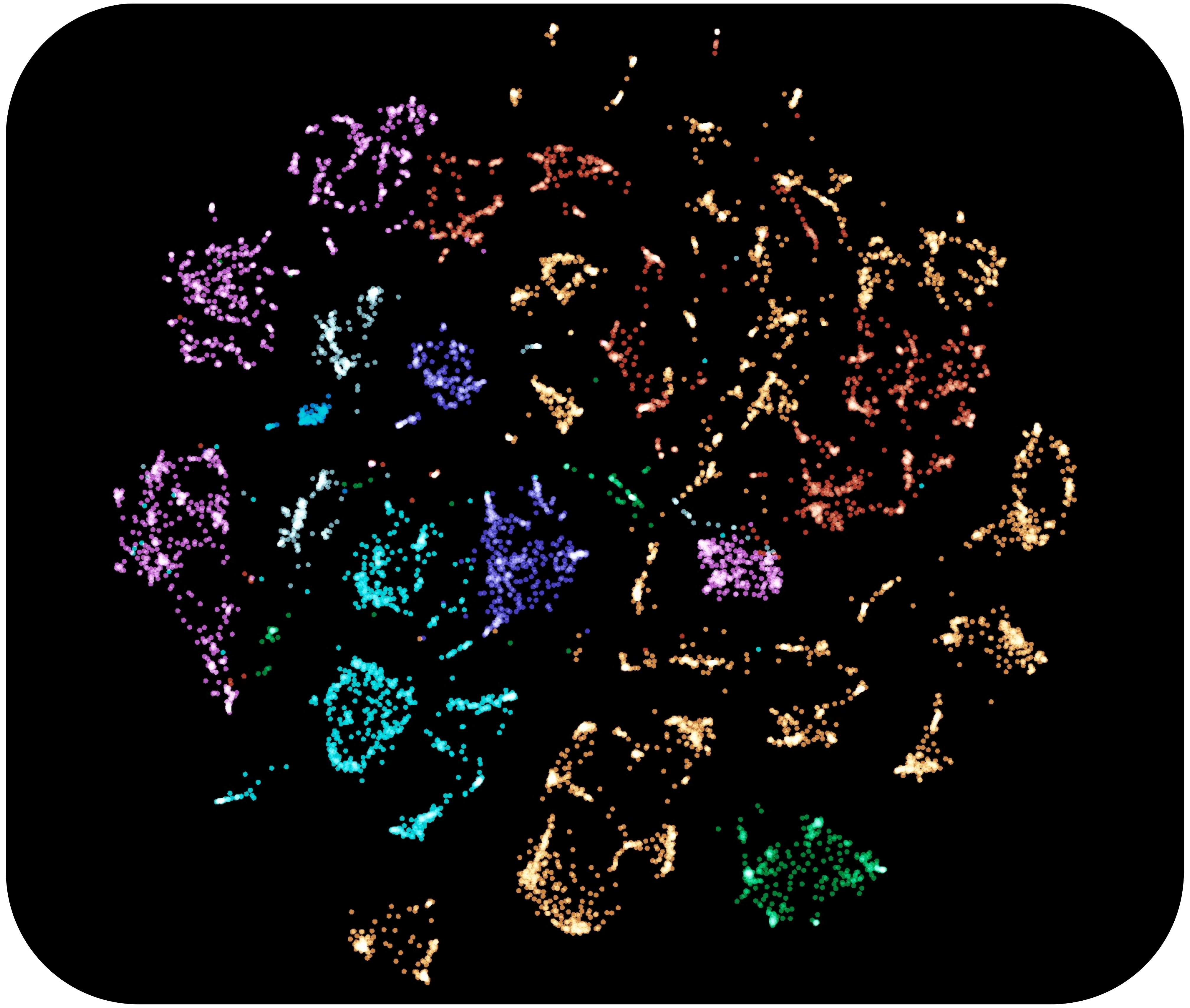
"QUALITATIVE"

- Take merchandising types as specified in the catalog by humans, and learn a classifier from the product space into the taxonomy.

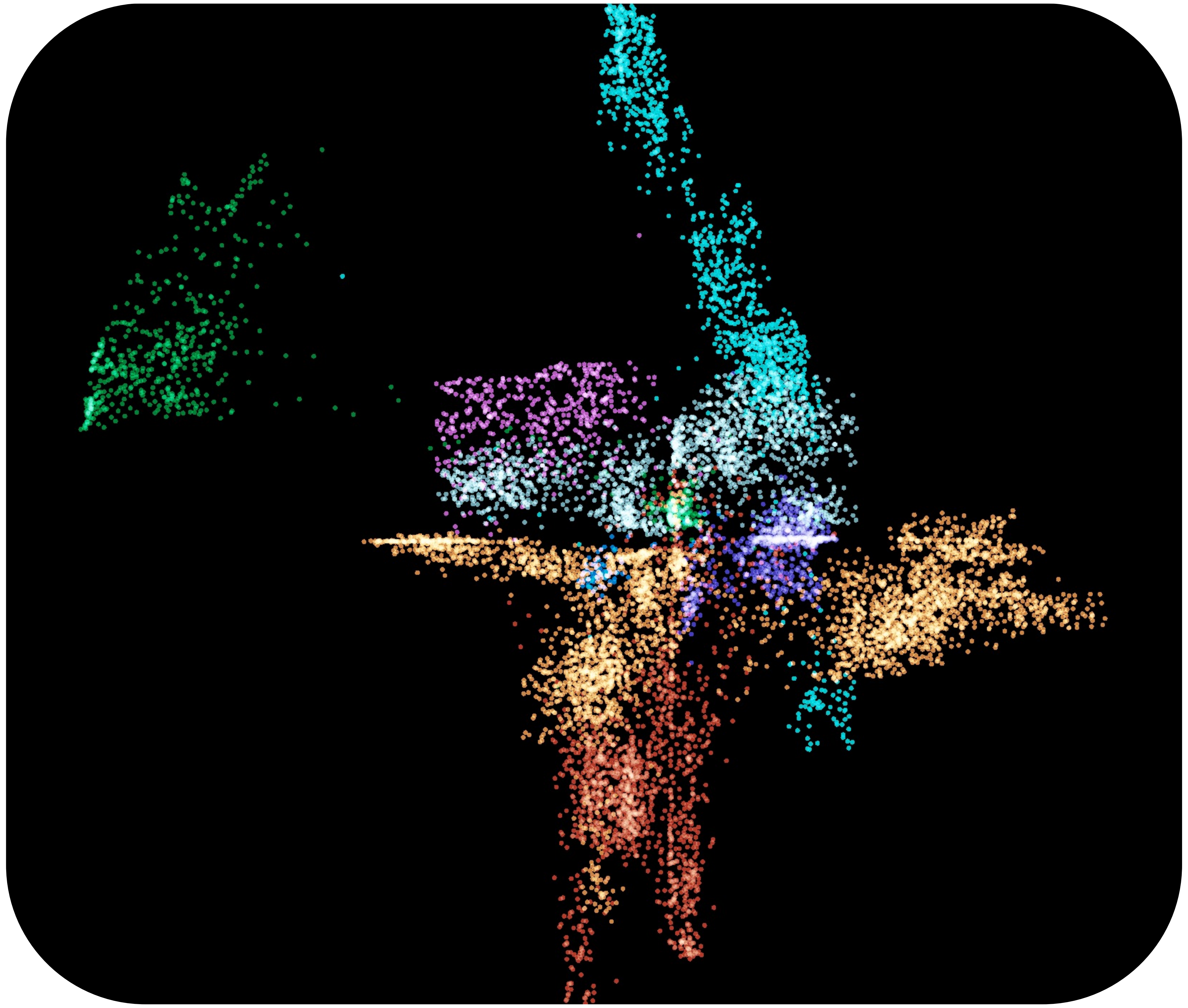
Good vs bad hyperparameters

- Embeddings from a sport apparel e-commerce website (colors represent sport activities).

Good embeddings



Bad embeddings



What about cold items?

kNN of a popular product

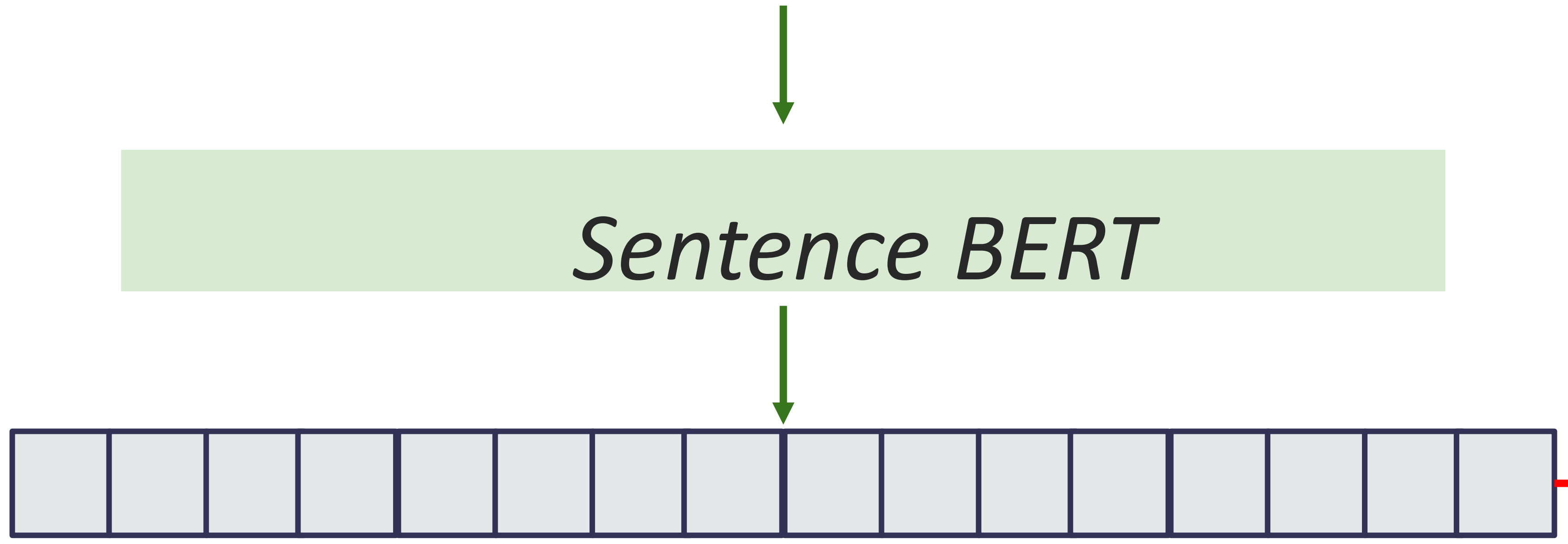
kNN of a rare product

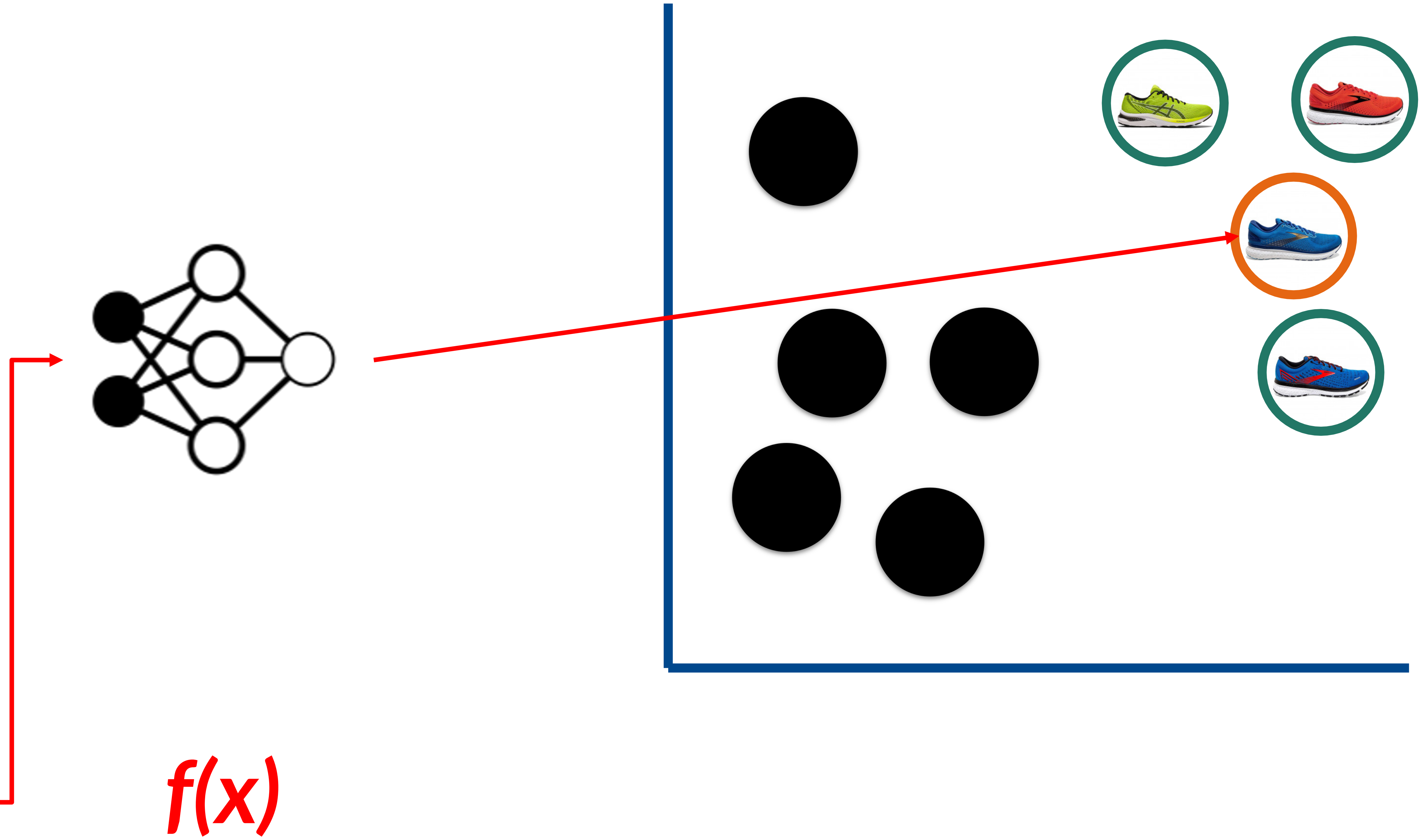
Plus, new products have no interactions!

Content-Embedding Substitution

1. We learn a mapping between product *textual* meta-data and the embedding space by using *popular products only*.

NAME	My Running Shoes
CATEGORY	Running Shoes
BRAND	Mizuno
DESCRIPTION	It’s awesome!



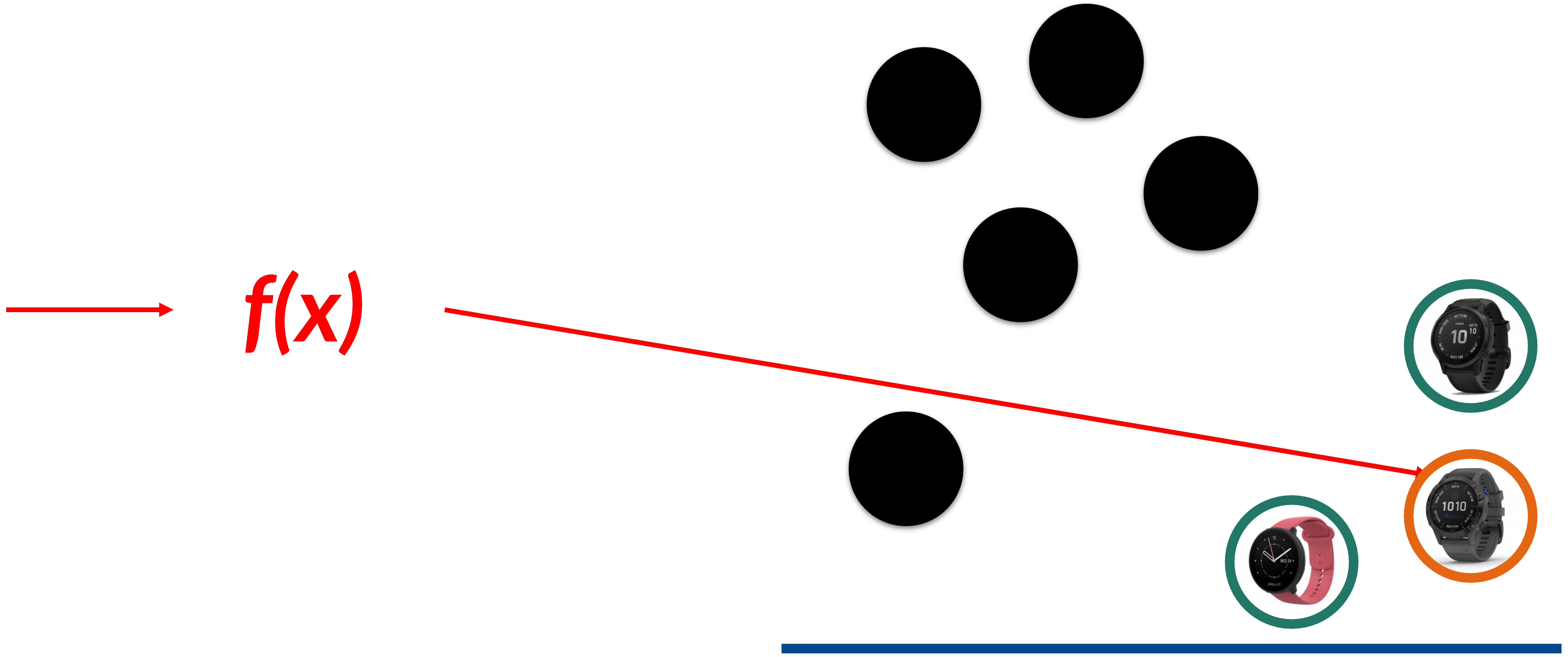


coveo™

Content-Embedding Substitution

1. We learn a mapping between product *textual* meta-data and the embedding space by using *popular products only*.
2. We substitute rare/new product vectors with “simulated” vectors, by applying the learned mapping to their meta-data.

NAME	My GPS Watch
CATEGORY	Running Accessories
BRAND	Garmin
DESCRIPTION	Very punctual!



coveo™

Content-Embedding Substitution

ENGINEERING WISE…

- Simple and scalable method, it does not require any change to existing training pipelines, as downstream models won't know which vectors are "real" and which one are synthetic.

PRODUCT WISE…

- Help with "unreasonable mistakes", which are very common in recommender systems and immediately degrade the shopping experience.

References for more details

SIGIR 2020

Fantastic Embeddings and How to Align Them: Zero-Shot Inference in a Multi-Shop Scenario

Federico Bianchi*
Bocconi University
Milano, Italy
f.bianchi@unibocconi.it

Jacopo Tagliabue*†
Coveo Labs
New York, NY
jtagliabue@coveo.com

Bingqing Yu*
Coveo
Montreal, Canada
cyu2@coveo.com

Luca Bigon‡
Coveo
Montreal, Canada
lbigon@coveo.com

Ciro Greco§
Coveo Labs
New York, NY
cgreco@coveo.com

ABSTRACT

This paper addresses the challenge of leveraging multiple embedding spaces for multi-shop personalization, proving that zero-shot inference is possible by transferring shopping intent from one website to another without manual intervention. We detail a machine learning pipeline to train and optimize embeddings *within shops* first, and support the quantitative findings with additional qualitative insights. We then turn to the harder task of using learned embeddings *across shops*: if products from different shops live in the same vector space, user intent - as represented by regions in this space - can then be transferred in a zero-shot fashion across websites. We propose and benchmark unsupervised and supervised methods to "travel" between embedding spaces, each with its own assumptions on data quantity and quality. We show that zero-shot personalization is indeed possible at scale by testing the shared embedding space with two downstream tasks, event prediction and type-ahead suggestions. Finally, we curate a cross-shop

ACM Reference Format:
Federico Bianchi, Jacopo Tagliabue, Bingqing Yu, Luca Bigon, and Ciro Greco. 2020. Fantastic Embeddings and How to Align Them: Zero-Shot Inference in a Multi-Shop Scenario. In *Proceedings of ACM SIGIR Workshop on eCommerce (SIGIR eCom'20).* ACM, New York, NY, USA, 11 pages.

1 INTRODUCTION

Inspired by the similarity between words in sentences and products in browsing sessions, recent work in recommender systems re-adapted the NLP CBOW model [20] to create *product embeddings* [17], i.e. low-dimensional representations which can be used alone or fed to downstream neural architectures for other machine learning tasks. Product embeddings have been mostly investigated as static entities so far, but, exactly as words [10], products are all but static. Since the creation of embeddings is a stochastic process, training embeddings for similar products in different digital shops

RecSys

The Embeddings That Came in From the Cold: Improving Vectors for New and Rare Products with Content-Based Inference

Authors: Jacopo Tagliabue, Bingqing Yu, Federico Bianchi Authors Info & Affiliations

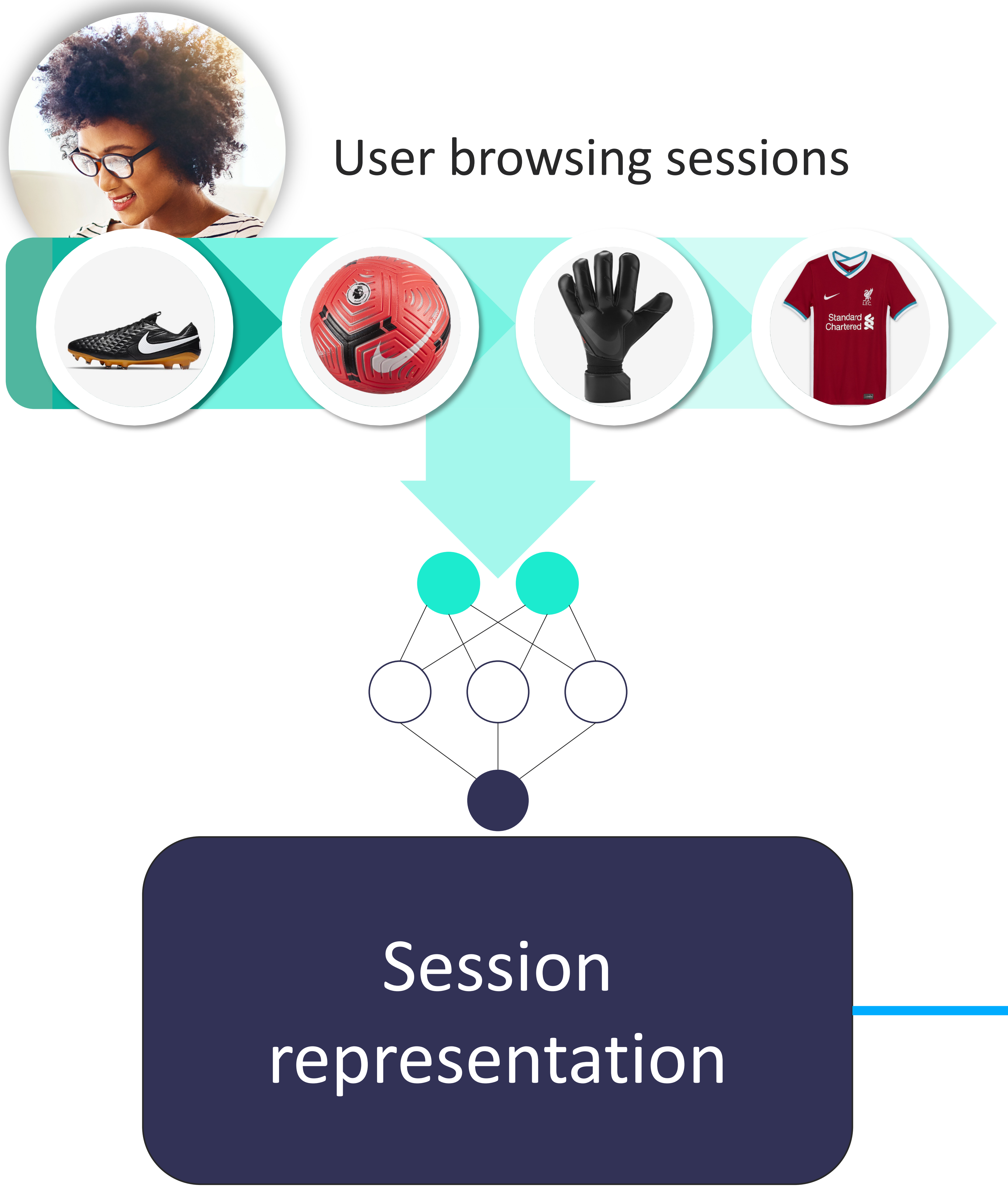
Publication: RecSys '20: Fourteenth ACM Conference on Recommender Systems • September 2020 • Pages 577–578 • https://doi.org/10.1145/3383313.3411477

ABSTRACT

Training product embeddings in a multi-tenant scenario involves solving the challenges of ever changing catalogs across dozens of deployments, without supervision. In this work, we detail how we deal with new and rare products when building neural representations at scale: we show how to inject product knowledge into behavior-based embeddings to provide the best accuracy with minimal engineering changes in existing infrastructure and without additional

Injecting personalization in downstream NLP systems

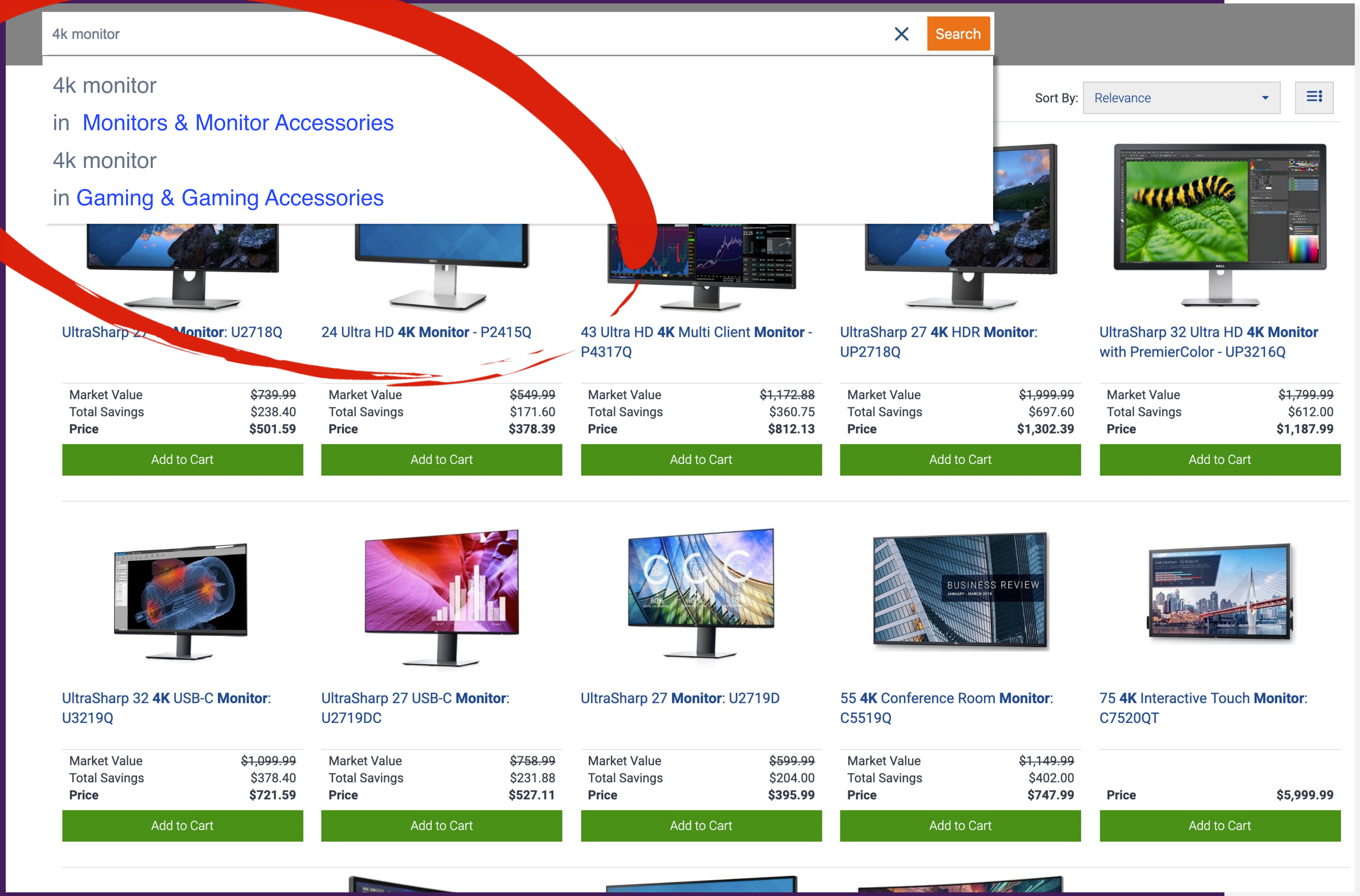
A research program for session-based inference



List of functionalities

- Search
- Recommendations
- **Query suggestion**
- Causal Attribution
- Intent Detection

Recall and sorting

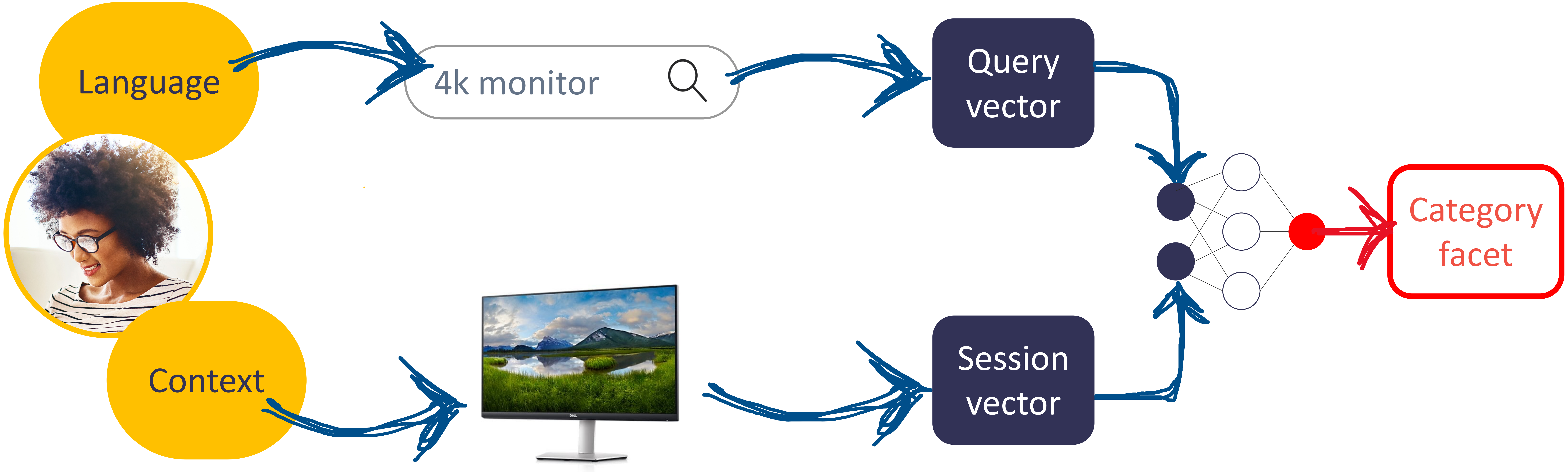


- Excessive recall can affect negatively the user experience.
- Sort by price vs. sort by relevance.
- Query scoping through search suggestions as a countermeasure.

coveo™

Machine learning to optimize query scoping

Given query ambiguity, category selection is a function of language *but also* shopping context.

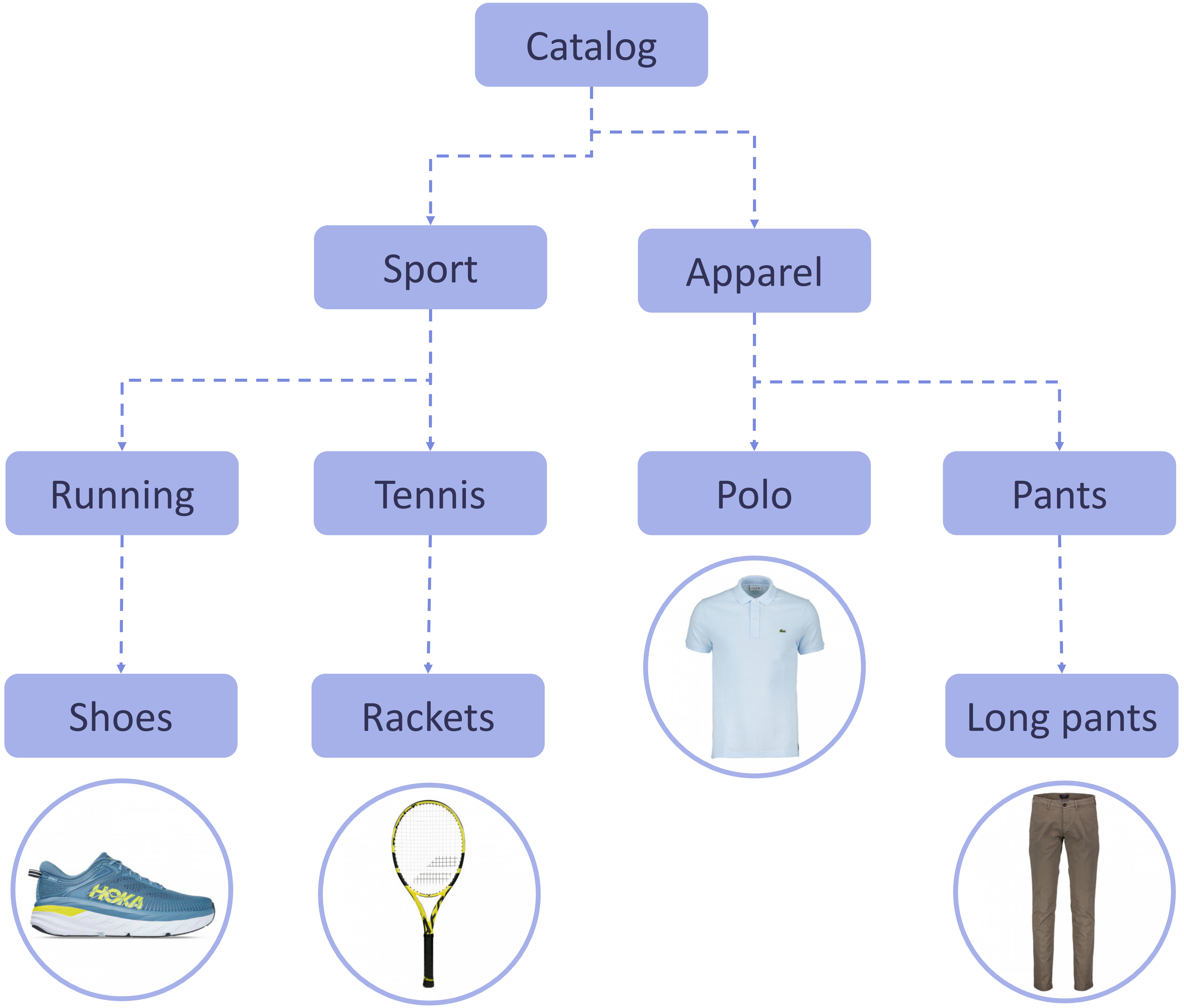


Dell 27 4K UHD Monitor S2721QS

Modelling inference through the underlying domain

Catalogs are hierarchical

- E-commerce catalogs are organized in hierarchical taxonomies.
- Their nodes tell us the structural relations between products and categories.
- Can we use this information with the session information to personalize query scoping for the query suggestion?



From multi-class to multi-path classification

- Category prediction is generally framed as a **multi-class** problem, but we can make it a **multi-path** one.

MULTI-CLASS

```
shoes
      in nike
```

```
shoes
      in adidas
```

```
shoes
      in mizuno
```

```
s
```

```
shoes
      in ??????
```

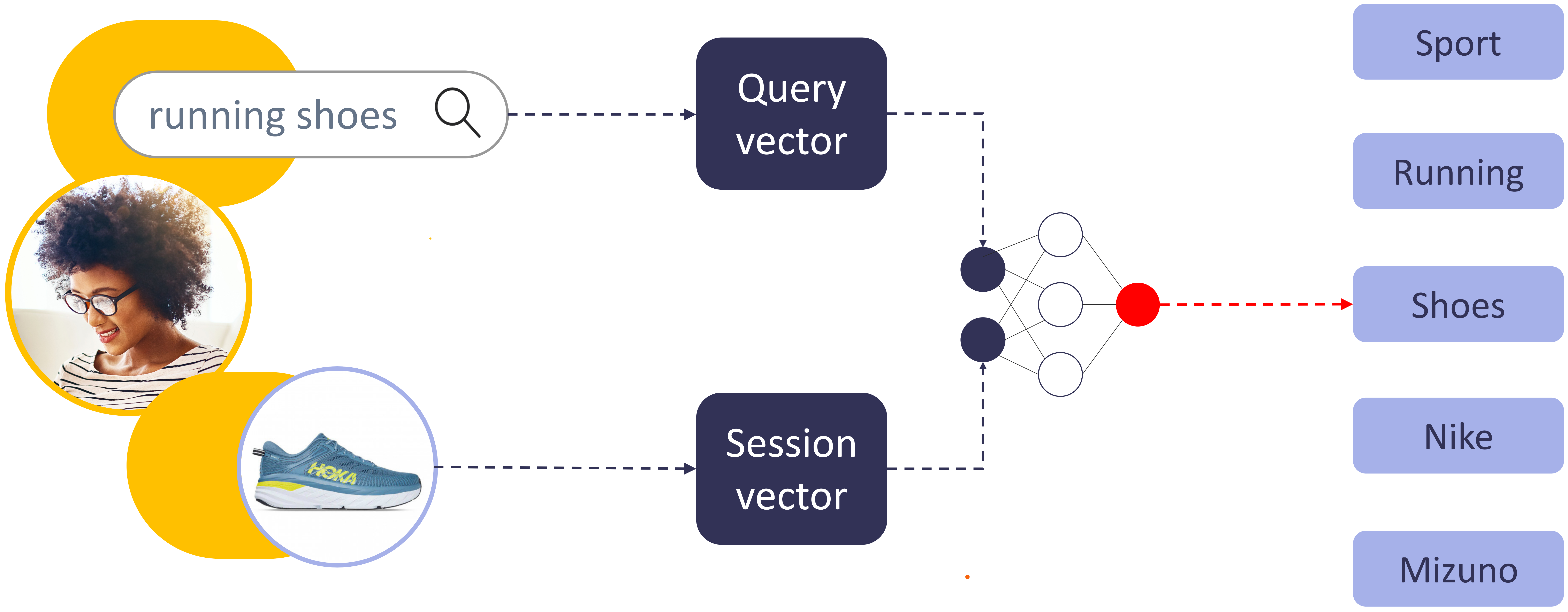
MULTI-PATH

```
shoes
      in nike/tennis
```

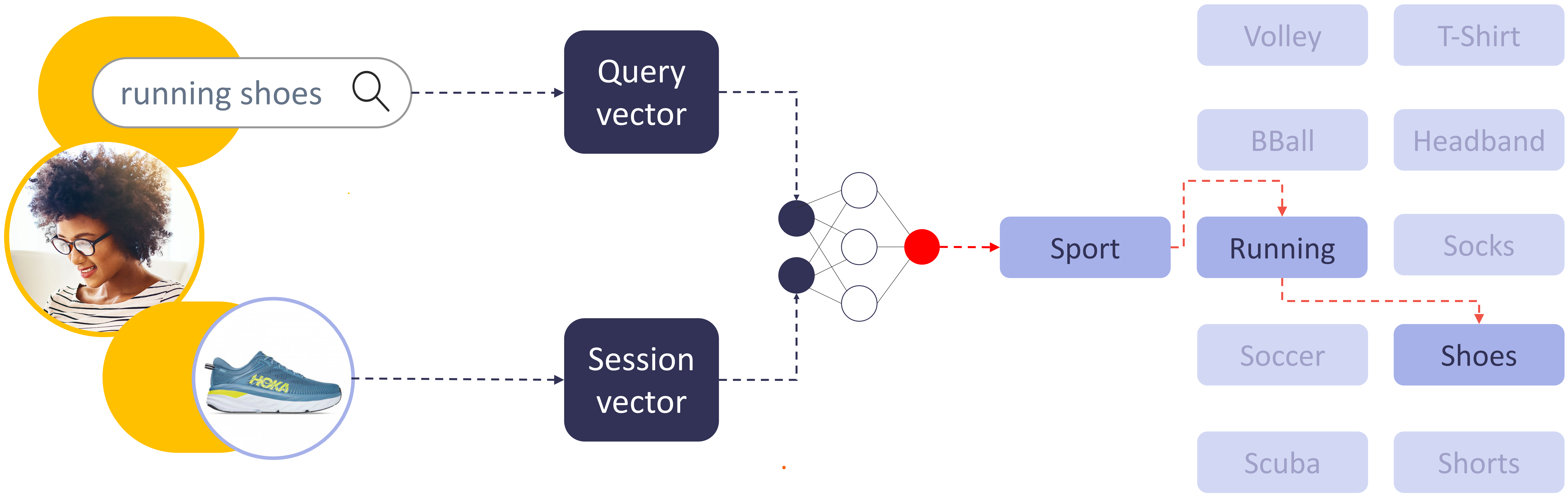
```
shoes
      in nike/soccer
```

```
shoes
      in mizuno
```

From multi-class...



...to multi-path classification



coveo™

Modelling meaning with custom embeddings

The limits of BERT(s)

- While large pre-trained contextual models (e.g. BERT) have dominated NLP in recent years, Information Retrieval applied to products is different:
  - Queries are very short: consequently, the contextual advantage is smaller.
  - Industry specific jargon and its semantics are not always captured by training datasets.
  - The bigger the model, the slower and more expensive it is to serve.

Query2Prod2Vec: a grounded language model

- Since queries are **about** products, why not use products to ground the meaning of queries?
- Shoppers searching for ”cap” generate a distribution of clicks over products p_1, p_2, … p_n.
- Clicked products are mapped to their embeddings e_1, e_2, … e_n in the *prod2vec* space.
- Finally, the linguistic vector for “cap” is the average of e_1, e_2, … e_n weighted by the clicks.

hat

1	1	9	2
2	2	5	2
5	5	5	7

Weighted average of product vectors

Product vectors

coveo™

References for more details

NAACL 2021

Language in a (Search) Box: Grounding Language Learning in Real-World Human-Machine Interaction

Federico Bianchi* Bocconi University Milano, Italy f.bianchi@unibocconi.it

Ciro Greco Coveo Labs New York, USA cgreco@coveo.com

Jacopo Tagliabue Coveo Labs New York, USA jtagliabue@coveo.com

Abstract

We investigate grounded language learning through real-world data, by modelling a teacher-learner dynamics through the natural interactions occurring between users and search engines; in particular, we explore the emergence of semantic generalization from unsupervised dense representations outside of synthetic environments. A grounding domain, a denotation function and a composition function are learned from user data only. We show how the resulting semantics for noun phrases exhibits compositional properties while being fully learnable without any explicit labelling. We benchmark our grounded semantics on compositionality and zero-shot inference tasks, and we show that it provides better

that language may be learned based on its usage and that learners draw part of their generalizations from the observation of teachers' behaviour (Tomasello, 2003). These ideas have been recently explored by work in grounded language learning, showing that allowing artificial agents to access human actions providing information on language meaning has several practical and scientific advantages (Yu et al., 2018; Chevalier-Boisvert et al., 2019).

While most of the work in this area uses toy worlds and synthetic linguistic data, we explore grounded language learning offering an example in which unsupervised learning is combined with a language-independent grounding domain in a real-world scenario. In particular, we propose to use the interaction of users with a search engine as a setting

NAACL 2021

Query2Prod2Vec Grounded Word Embeddings for eCommerce

Federico Bianchi Bocconi University Milano, Italy f.bianchi@unibocconi.it

Jacopo Tagliabue* Coveo Labs New York, USA jtagliabue@coveo.com

Bingqing Yu Coveo Montreal, Canada cyu2@coveo.com

Abstract

We present **Query2Prod2Vec**, a model that grounds lexical representations for product search in product embeddings: in our model, *meaning* is a mapping between words and a latent space of products in a digital shop. We leverage shopping sessions to learn the underlying space and use merchandising annotations to build lexical analogies for evaluation: our experiments show that our model is more accurate than known techniques from the NLP and IR literature. Finally, we stress the importance of data efficiency for product search outside of retail giants, and highlight how **Query2Prod2Vec** fits with practical constraints faced by most practitioners.

1 Introduction

industry-specific jargon (Bai et al., 2018), low-resource languages; moreover, specific embedding strategies have often been developed in the context of high-traffic websites (Grbovic et al., 2016), which limit their applicability in many practical scenarios. In *this* work, we propose a sample efficient word embedding method for IR in eCommerce, and benchmark it against SOTA models over industry data provided by partnering shops. We summarize our contributions as follows:

1. we propose a method to learn dense representations of words for eCommerce: we name our method **Query2Prod2Vec**, as the mapping between words and the latent space is mediated by the product domain;

Experiments

Dataset and benchmarks

- We test several query embeddings strategies and three inference methods (simple count-based baseline **CM**, **MLP**, full enc-dec), reporting accuracy at different depth in the catalog tree.
- Given our multi-tenant nature, we check for robustness by running all tests on two shops, differing in products, categories, traffic and vertical.

Shop	Queries (with context)	Products
Shop 1	270K (185K)	29.699
Shop 2	270K (227K)	93.967

Model	D=1	D=2	D=last
CM	0.63	0.53	0.22
MLP+BERT	0.72	0.59	0.33
SP+BERT	0.77	0.64	0.40
SP+LSTM	0.79	0.68	0.43
SP+W2V	0.82	0.71	0.46
SP+SV	**0.87**	**0.79(0.01)**	**0.55**
CM	0.41	0.34	0.24
MLP+BERT	0.61	0.50	0.39
SP+BERT	0.66	0.55	0.45
SP+LSTM	0.67	0.57	0.46
SP+W2V	0.69	0.59	0.47
SP+SV	**0.80**	**0.71**	**0.59**

Table 2: Accuracy scores for $depth = 1$, $depth = 2$, $depth = last$, divided by **Shop 1** (*top*) and **Shop 2** (*bottom*). We report the mean over 5 runs, with SD if $SD \geq 0.01$.

The role of inductive bias and context

- **SP+SV** even with only 1/10th of samples outperforms all other models.
- By leveraging the bias encoded in the hierarchical structure of the products, **SP+SV** allows paths that share nodes (*sport, sport / basketball*) to also share statistical evidence.
- Session information helps the most with *unseen* queries at test time (unsurprisingly).

Model (D=last)	1/10	1/4
CM	0.18	0.20
MLP+BERT	0.28	0.30
SP+BERT	0.31	0.34
SP+SV	**0.51**	**0.53**

Table 3: Accuracy scores (**D=last**) when training on portions of the original dataset for **Shop 1**.

coveo™

Tuning the black box with an interpretable decision module

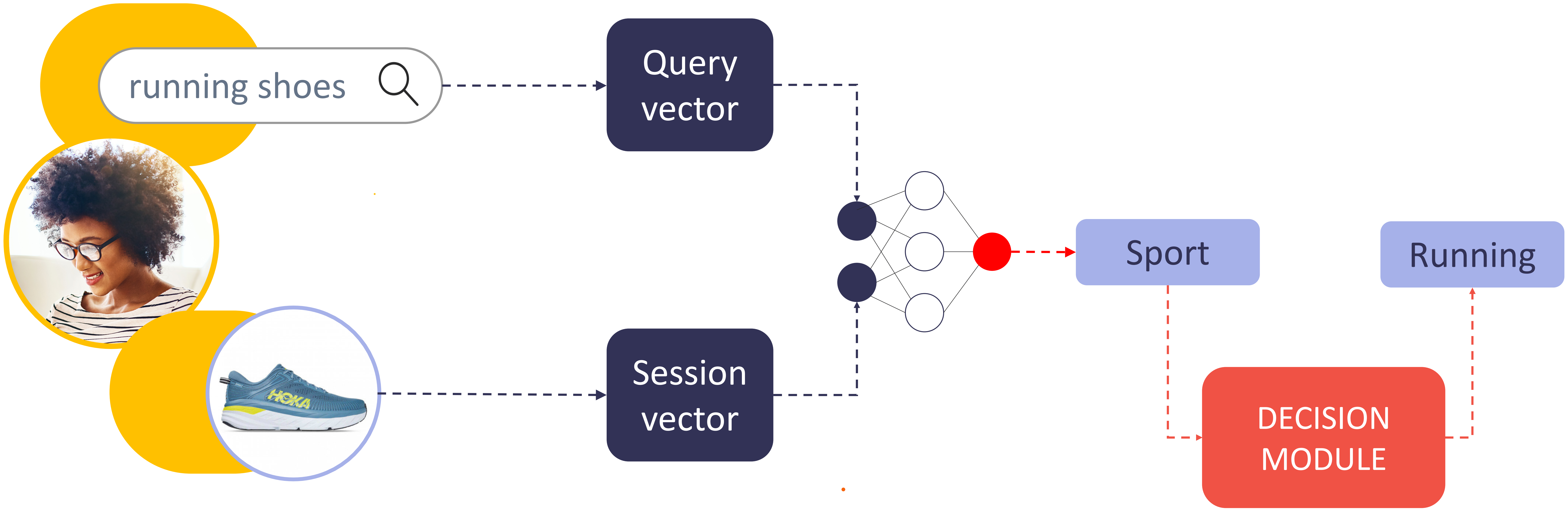
coveo™

Precision and recall in the eye of the beholder

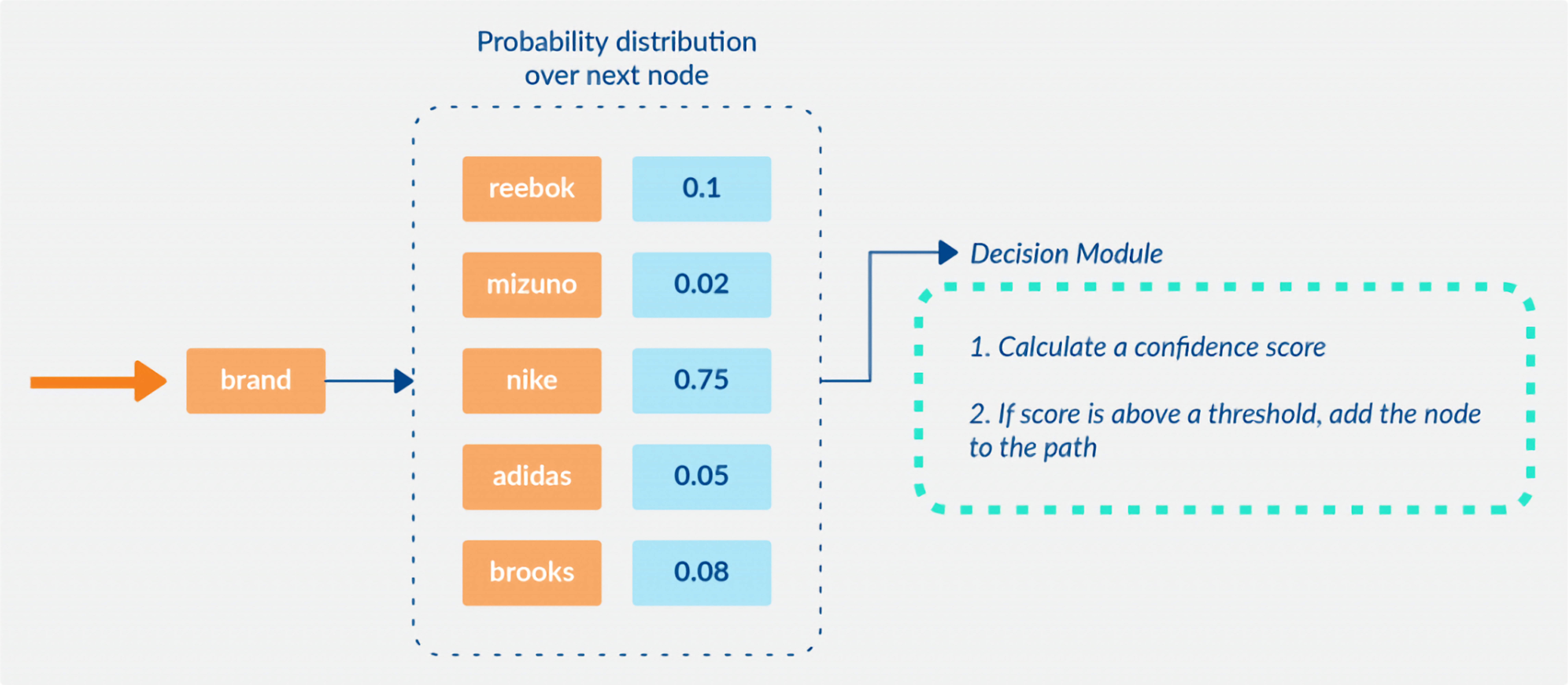
- Given a **query** and a **session**, SessionPath may generate a path 3 levels deep.
- In the case #1, the result set is cut at "nike", leaving more choice to the shopper; in the case #2, the result set is not cut, narrowing down on basketball-specific items.
- Different industries have different sensibilities on *precision vs recall*: there is no “right answer”.

coveo™

Hybrid architecture: adding a decision module

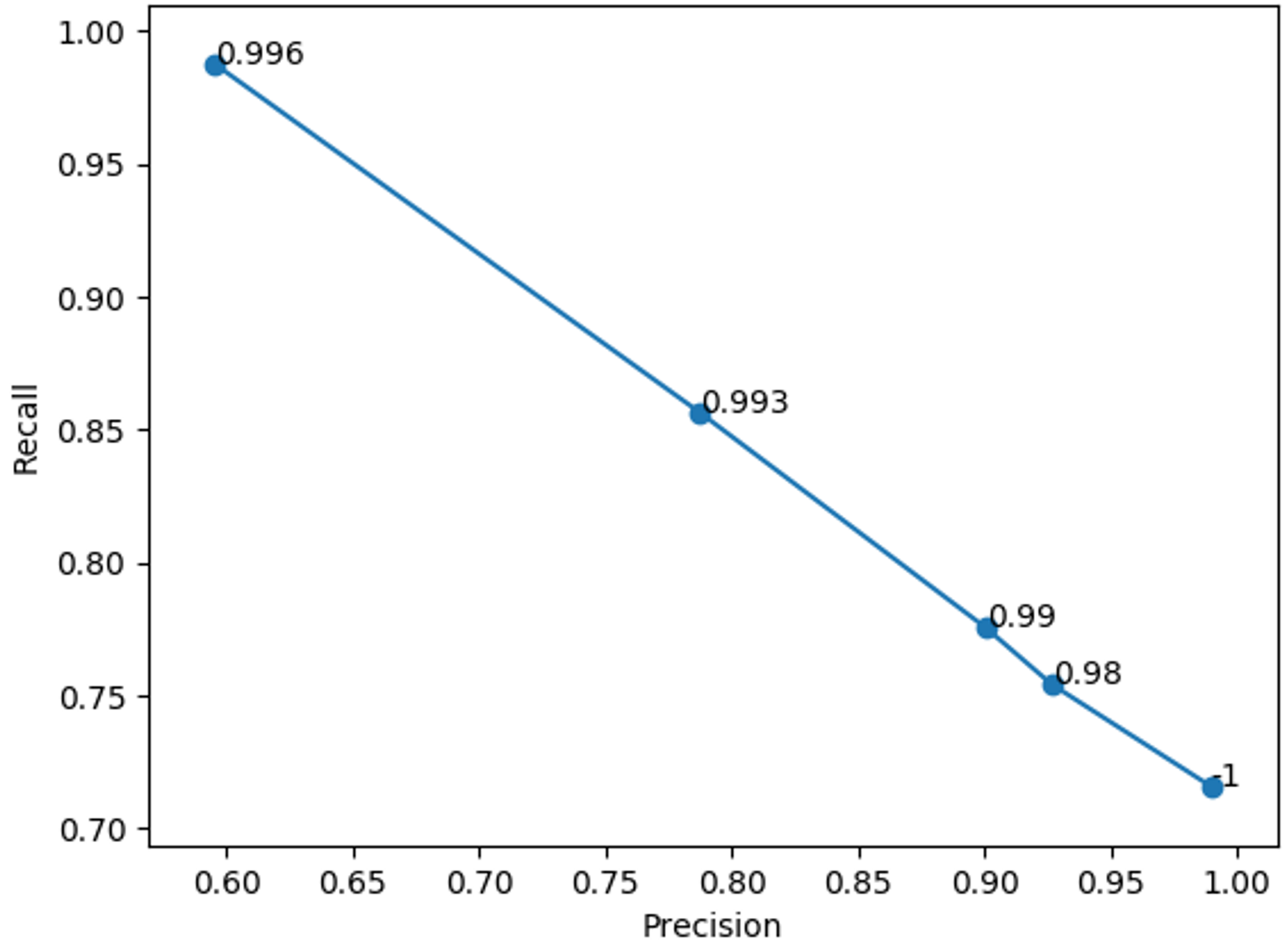


Hybrid architecture: adding a decision module



Hybrid architecture: adding a decision module

- By setting our decision threshold, we can pick the precision vs recall trade-off we want.



GINI THRESHOLD	PRECISION	RECALL
0.996	0.65	0.99
0.993	0.82	0.91
0.990	0.93	0.77
0.980	0.99	0.74

SessionPath at work

- Model at work with different thresholds: **A**, empty session, **B**, containing an interaction with a running shoes; **A** defaults to the most common path, **B** showcases both session conditioning *and* a flexible path depth.

n

nike shoes in ??????

A

EMPTY SESSION

TR=0.98 shoes/sneakers/men

TR=0.97 shoes/sneakers/men

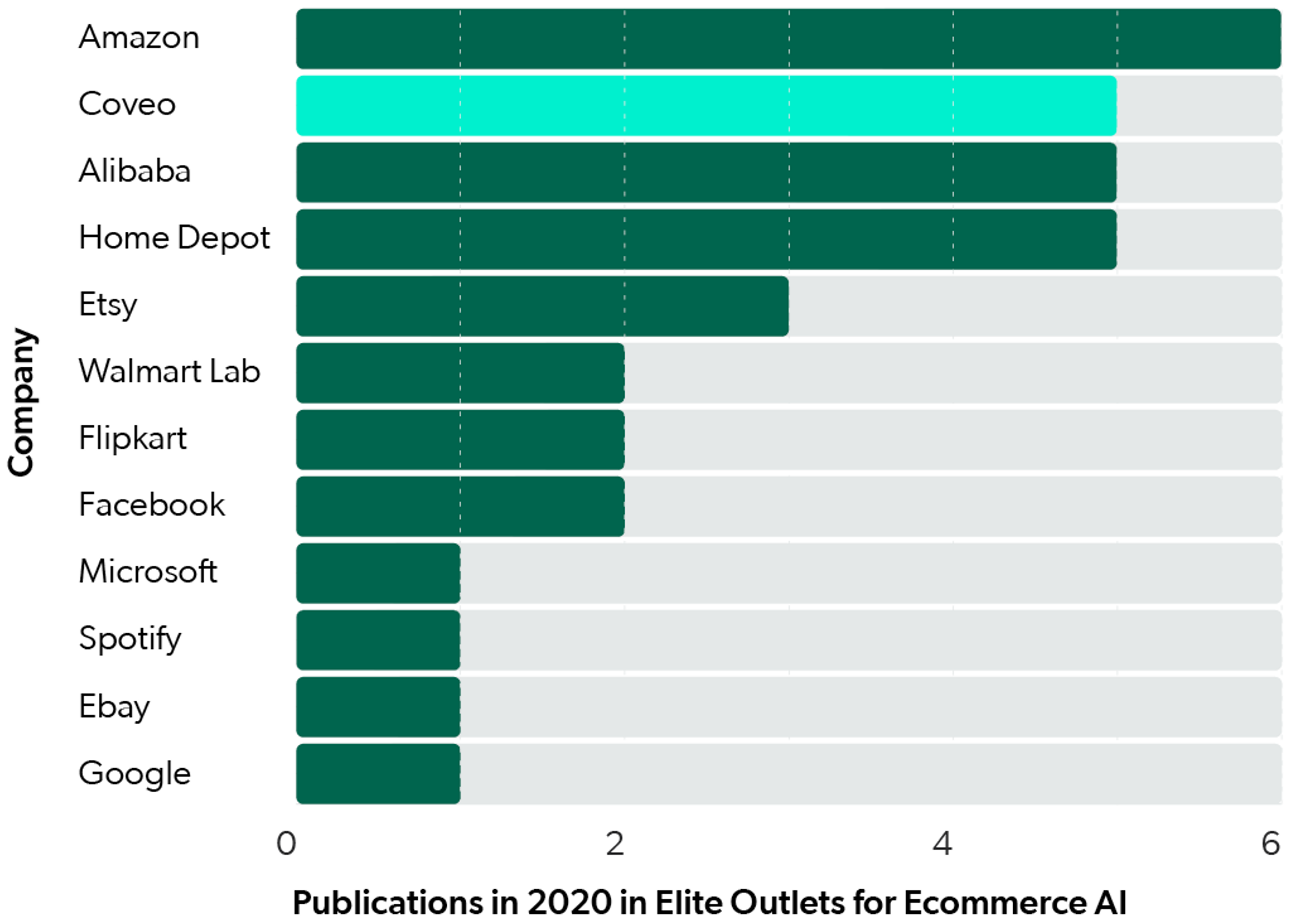
B

TR=0.98 shoes/running/shoes

TR=0.97 shoes/running/shoes/a3

coveo™

Doing cutting-edge ML at reasonable scale



KDD2020 SIGIR 2020 XI'AN·CHINA RecSys acm.org Association for Computational Linguistics THE WEB CONFERENCE April 20-24 TAIPEI 2020

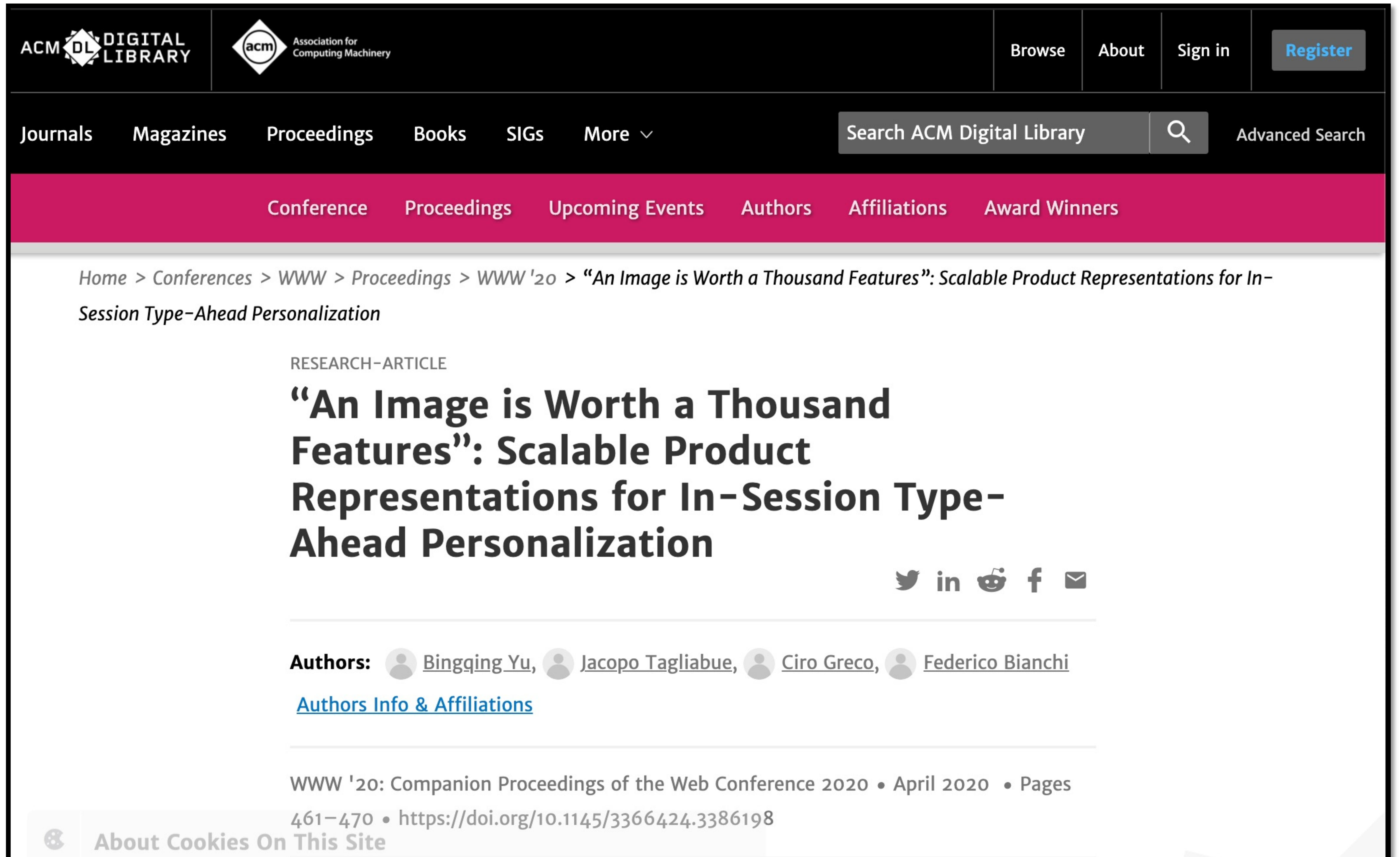
ML is still hard outside of few players!

ML at “reasonable scale”

- Lack of massive computing and massive user base
- Lack of representative models / datasets
- Lack of talent
- Lack of engineering / tooling best practices

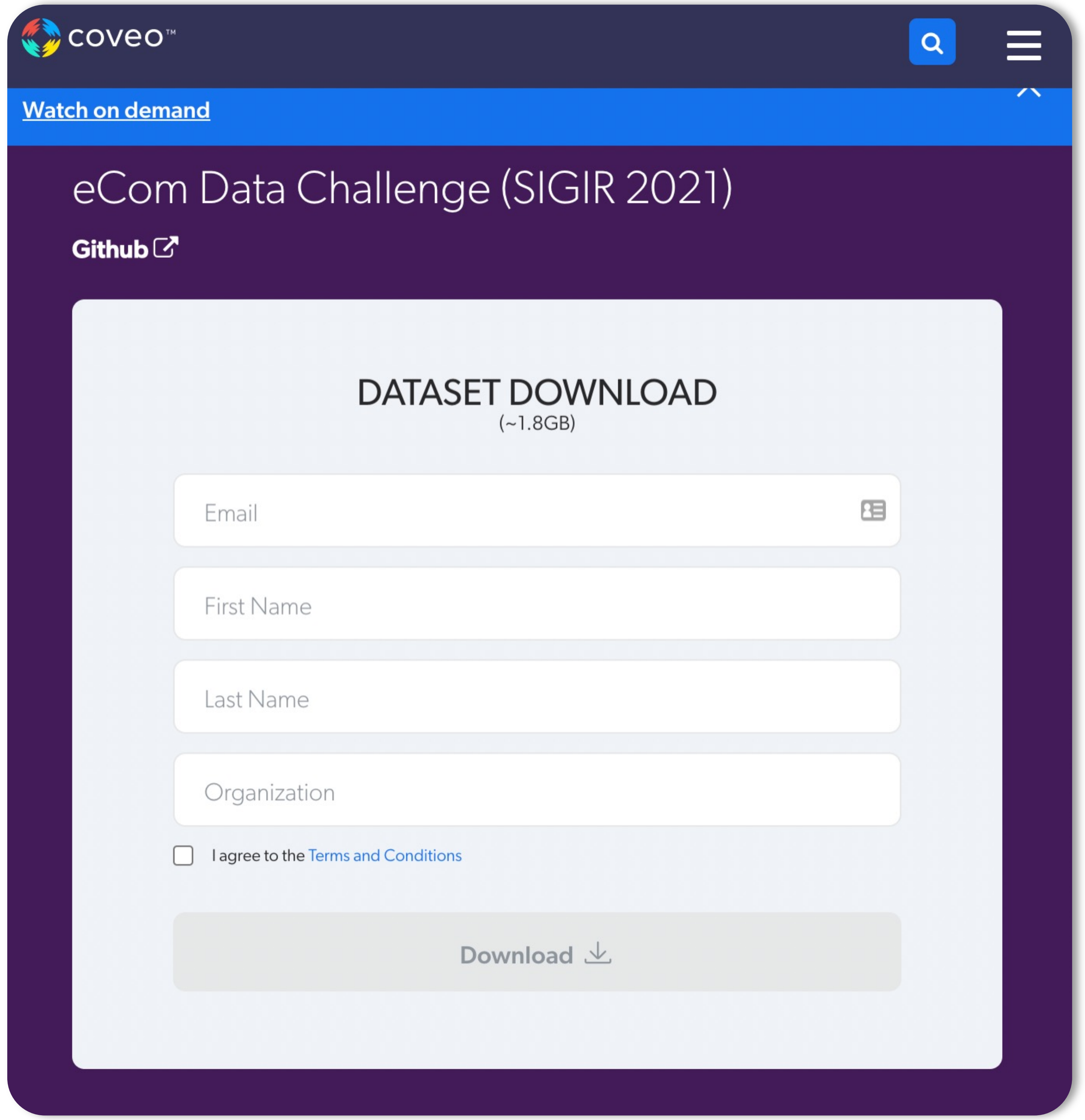
ML at “reasonable scale”

- Lack of massive computing and massive user base
  - Turn business constraints into a research question (e.g. how do make inference within a session?). There are tons of interesting problems at “reasonable scale”!



ML at "reasonable scale"

- Lack of representative models / datasets
  - Release datasets and open source code that work across organizations of all / most sizes.



- More than 30M browsing events, fully anonymized and hashed, generated over ~5M millions of shopping sessions produced by real users on real ecommerce sites.
- https://github.com/coveooss/SIGIR-ecom-data-challenge

ML at “reasonable scale”

- Lack of talent
  - Build a shared roadmap with academia, especially young researchers: share the “cost” and the “awards” of exploring ideas together.

ML at “reasonable scale”

- Lack of engineering / tooling best practices
  - Evangelize the field with code and best practices to build end-to-end systems and make ML teams productive.

README.md

You Don't Need a Bigger Boat

An end-to-end (Metaflow-based) implementation of an intent prediction flow for kids who can't MLOps good and wanna learn to do other stuff good too.

This is a WIP - check back often for updates.

Philosophical Motivations

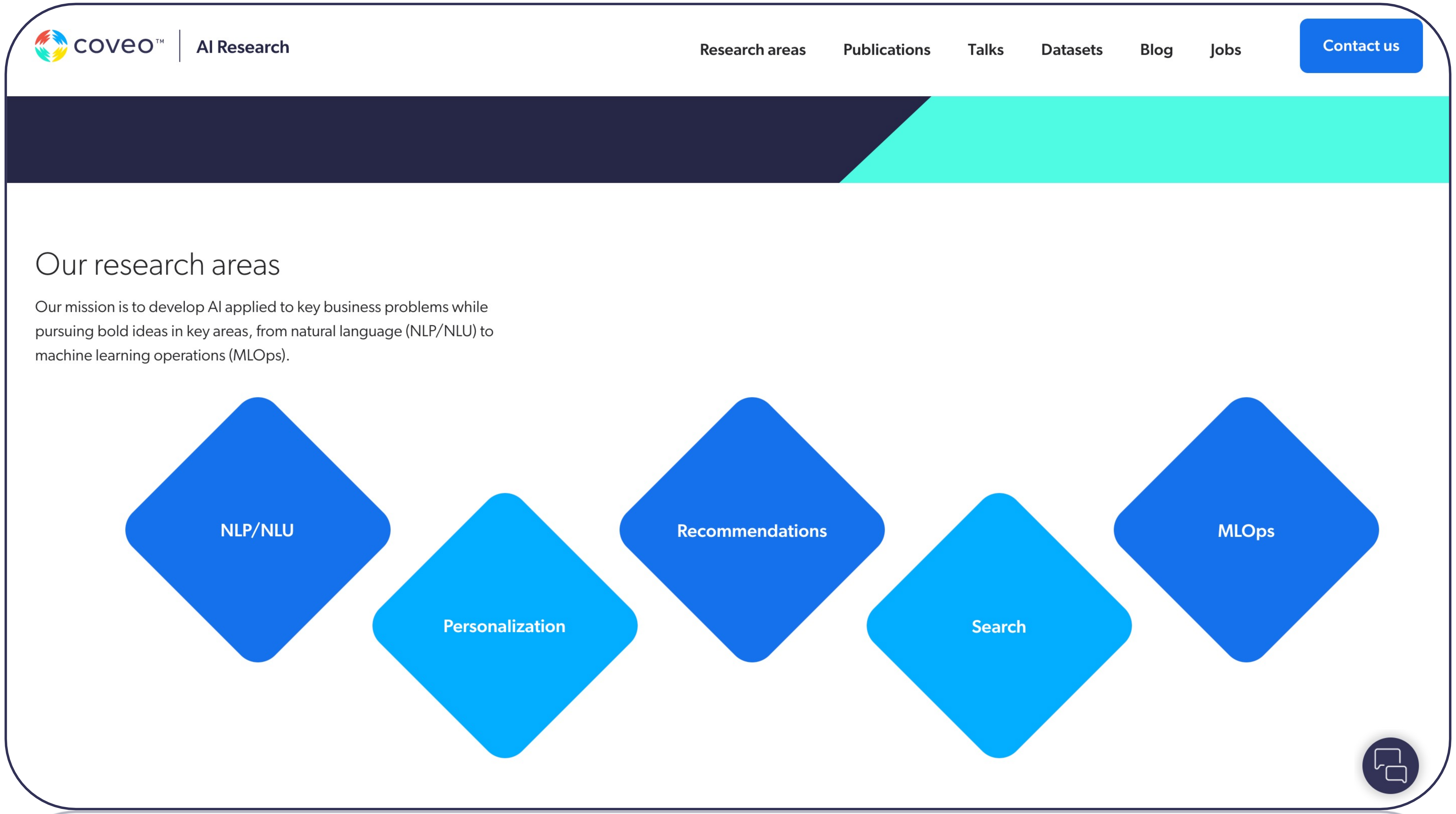
There is plenty of tutorials and blog posts around the Internet on data pipelines and tooling. However:

- they (for good pedagogical reasons) tend to focus on *one* tool / step at a time, leaving us to wonder how the rest of the pipeline works;
- they (for good pedagogical reasons) tend to work in a toy-world fashion, leaving us to wonder what would happen when a real dataset and a real-world problem enter the scene.

This repository (and soon-to-be-drafted written tutorial) aims to fill these gaps. In particular:

- we provide open-source working code that glues together what we believe are some of the best tools in the ecosystem, going *all the way* from raw data to a deployed endpoint serving predictions;
- we run the pipeline under a realistic load for companies at "reasonable scale", leveraging a huge open

Find us at: research.coveo.com



- Peer-reviewed papers
- Open datasets
- Talks / lectures
- More soon…

coveo™

See you, space cowboys.

coveo™