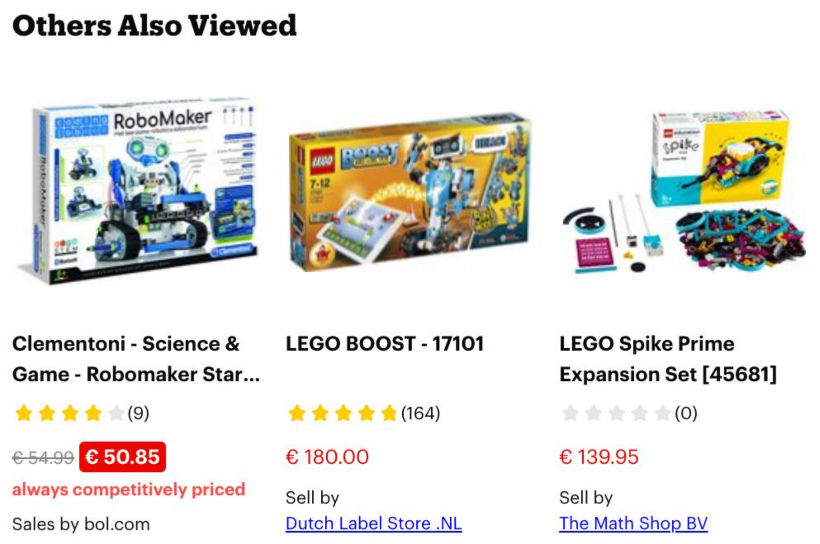
Serving Low-Latency Session-Based Recommendations at bol.com

Barrie Kersbergen[1,2], Olivier Sprangers[1], Sebastian Schelter[1]
[1]*University of Amsterdam* [2]*bol.com*

What is Session-Based Recommendation?

- Session-Based Recommendation (SBR): Predict the next item with which a user will interact in a session
- E-commerce / bol.com:

- Given a sequence of items $s = [s_1, s_2, \ldots, s_n]$, predict the next item s_{n+1}



Scalability Challenges in SBR

- Recommendations too costly to precompute due to the large number of potential sessions
  - SBR system needs to **compute recommendations online** and **maintain state**
- **Low latency response time** (p90 < 50ms) required in real-world scenarios
- **High throughput** (>1000 predictions/second)
- High-dimensional, extremely sparse click data from e-commerce platforms (**33 million distinct items**, very short sessions)

Efficient and High-Quality Recommendations

Experimental study on bol.com data

- **Replicated experimental study** from Ludewig et al: "Performance comparison of neural and non-neural approaches to session-based recommendation", RecSys'19 **on bol.com click data**
- **Vector-Session kNN (VS-kNN) approach outperformed neural approaches both in terms of prediction quality and inference time**
- Published as "Learnings from a Retail Recommendation System on Billions of Interactions at bol.com" at ICDE'21

Design of the **VMIS-kNN** algorithm

- Adaptation of VS-kNN
- Leverages a **precomputed index over historical click data** for fast inference
- Minimises intermediate results during nearest neighbor search
- **Highly tuned implementation in Rust**

Datasets for Offline Evaluation

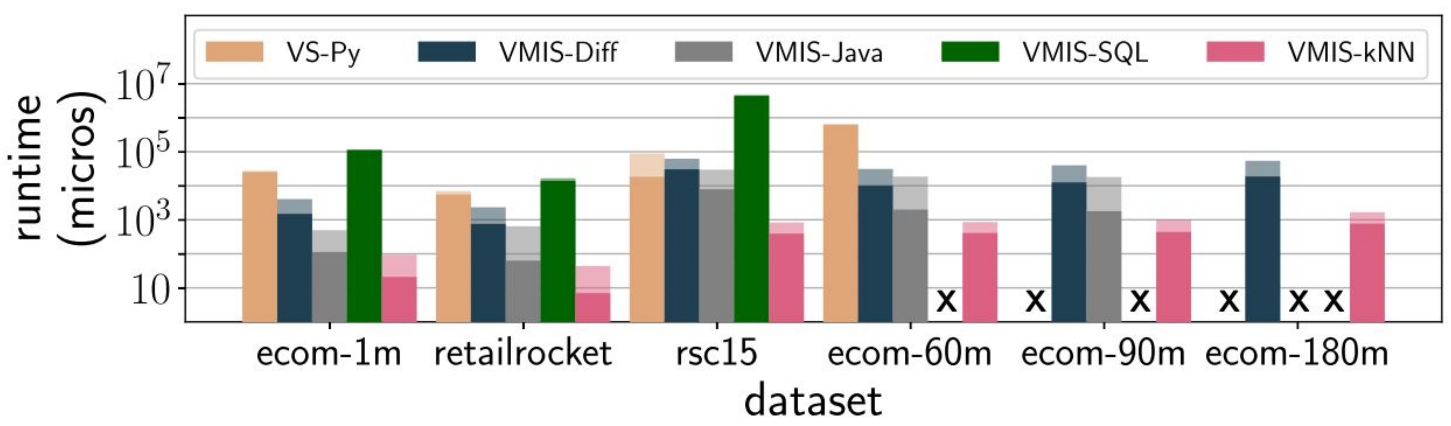
	retailr	*rsc15*	*ecom-1m*	*ecom-60m*	*ecom-90m*	*ecom-180m*
clicks	86,635	31,708,461	1,152,438	67,017,367	89,883,761	189,317,506
sessions	23,318	7,981,581	214,490	10,679,757	13,799,762	28,824,487
items	21,276	37,483	110,988	1,760,602	2,263,670	3,305,412
days	10	181	30	29	91	91
public?	yes	yes	no	no	no	no
clicks per session						
p25	2	2	2	2	2	2
p50	2	3	4	4	4	4
p75	4	4	6	7	7	7
p99	19	19	28	36	38	39

Table 1: Public and proprietary datasets for evaluation.

Micro Benchmark for Inference

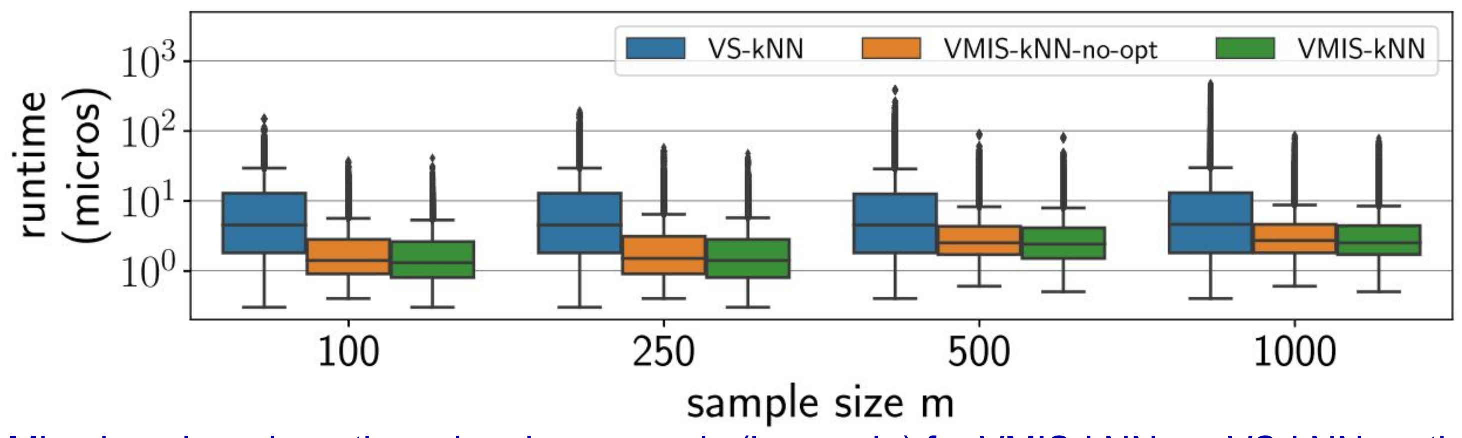
Benchmark for inference performance of different implementations

- Original VS-kNN (Python)
- VMIS-kNN in Dataflow systems (Differential Dataflow / RDBMS)
- VMIS-kNN in Rust and Java



Experimental results

- VMIS-kNN (Rust) shows low prediction latency **p90 =< 1.7ms for all datasets**
- VMIS-kNN (Rust) at least an **order of magnitude faster than baselines**
- Several baselines encounter memory issues for larger datasets



Microbenchmark runtimes in microseconds (log-scale) for VMIS-kNN vs. VS-kNN on the ecom-1m dataset with k=100

Serenade

- **Offline computation of session index**
  - 180 days of click data
  - (2.3 billion user-item interactions)
- Data-parallel computation with pySpark

- **Online serving of next-item recommendations**
  - **Replication** of index (13 GB of memory required per machine)
  - Needs 2 vCPU's in total to handle 1000 req/s
- **Colocation** of evolving **sessions** with recommendation **requests** and session **updates** (using session affinity)
  - Maintenance of session state via a local key-value store (RocksDB)

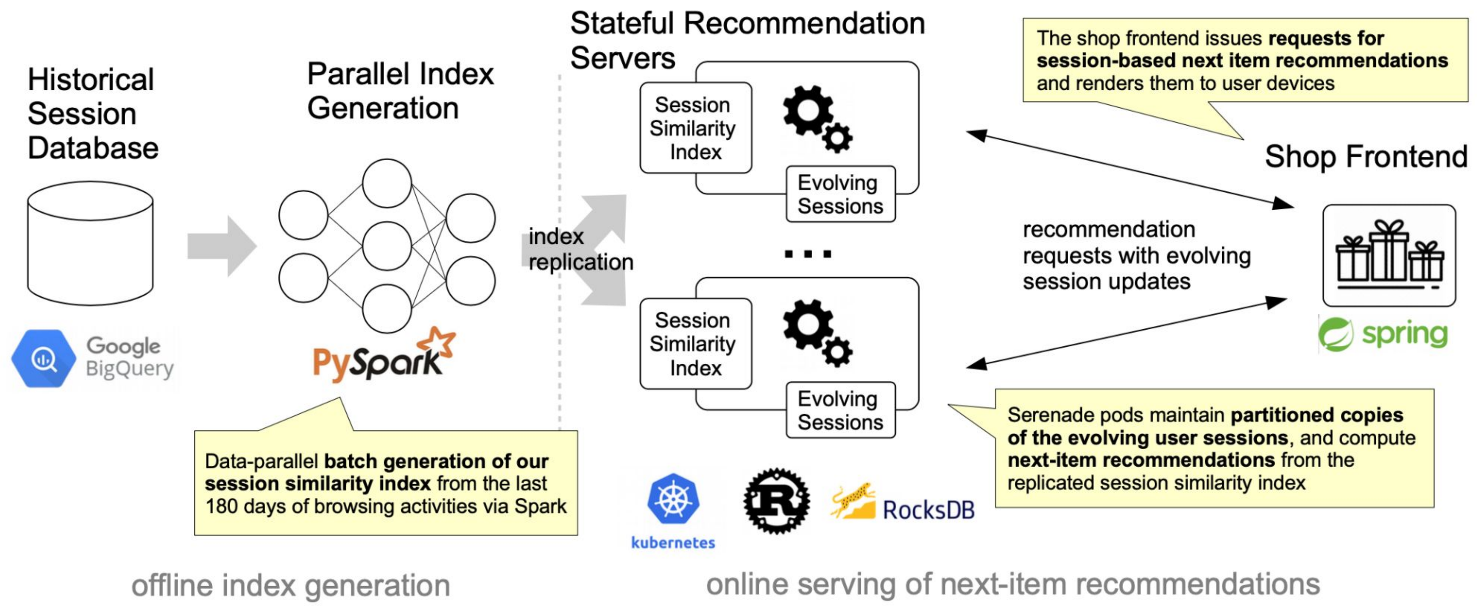


Figure 1: High level architecture of the Serenade recommendation system. The offline component (left) generates a session similarity index ❶ from several billion historical click events via a parallel Spark job in regular intervals. The online serving machines (right) maintain state about the evolving user sessions ❷, and leverage the session similarity index to compute next item recommendations with VMIS-kNN in response to recommendation requests from the shopping frontend ❸.

Online A/B Test on the Live Platform

Experimental setup

- **Serenade vs legacy system** (item-to-item CF)
  - **3 week long A/B test** for 'other also viewed' recommendations on the product detail page
  - Training data: 582 millions user-item interactions after pruning, extracted from 180 days of data, **6.5M distinct items**

Results

- Test included **45 million user sessions**
- Load varied **between 200 and 600 requests per second**
- Response latency: **p90 around 5ms, p99.5 < 10ms**
- 2.85% increase in relevant business metric

Summary

- Design and implementation of a **kNN-based real-world SBR system**
- **Deployed in production** at bol.com
- Scalability achieved via:
  - Scalable variant of a well working kNN approach, based on a **precomputed index** for fast inference
  - **Colocation** of sessions with recommendation requests on serving machines
- A/B test showed low-latency response times and increase in business metrics

- **Open sourced** at https://github.com/bolcom/serenade

- **Systems publication:** Barrie Kersbergen, Olivier Sprangers, Sebastian. Schelter: "Serenade - Low-Latency Session-Based Recommendation in e-commerce at Scale," ACM SIGMOD, 2022 (to appear)

- More about our research on **https://bkersbergen.github.io**

References

[1] Q. Liu *et al.*, "Stamp: short-term attention/memory priority model for session-based recommendation," *KDD*, 2018.

[2] M. Ludewig, N. Mauro, S. Latifi, and D. Jannach, "Performance comparison of neural and non-neural approaches to session-based recommendation," in *RECSYS*, 2019.

[3] B. Kersbergen and S. Schelter, "Learnings from a retail recommendation system on billions of interactions at bol.com," *ICDE*, 2021.

[4] I. Arapakis, X. Bai, and B. B. Cambazoglu, "Impact of response latency on user behavior in web search," in *SIGIR*, 2014.

[5] B. Kersbergen, O. Sprangers, and S. Schelter, "Serenade - low-latency session-based recommendation in e-commerce at scale," *SIGMOD*, 2022 (to appear).

UNIVERSITY OF AMSTERDAM
bol.com
THE AI FOR RETAIL LAB AMSTERDAM
INDE lab