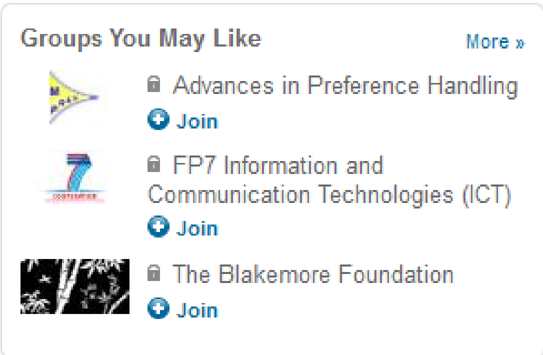
Session-based Recommendation: Challenges and Recent Advances

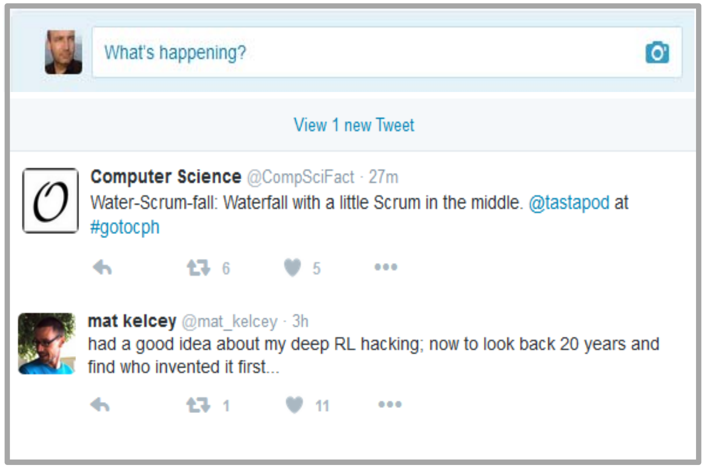
Dietmar Jannach, AAU Klagenfurt, Austria
dietmar.jannach@aau.at

Recommender Systems

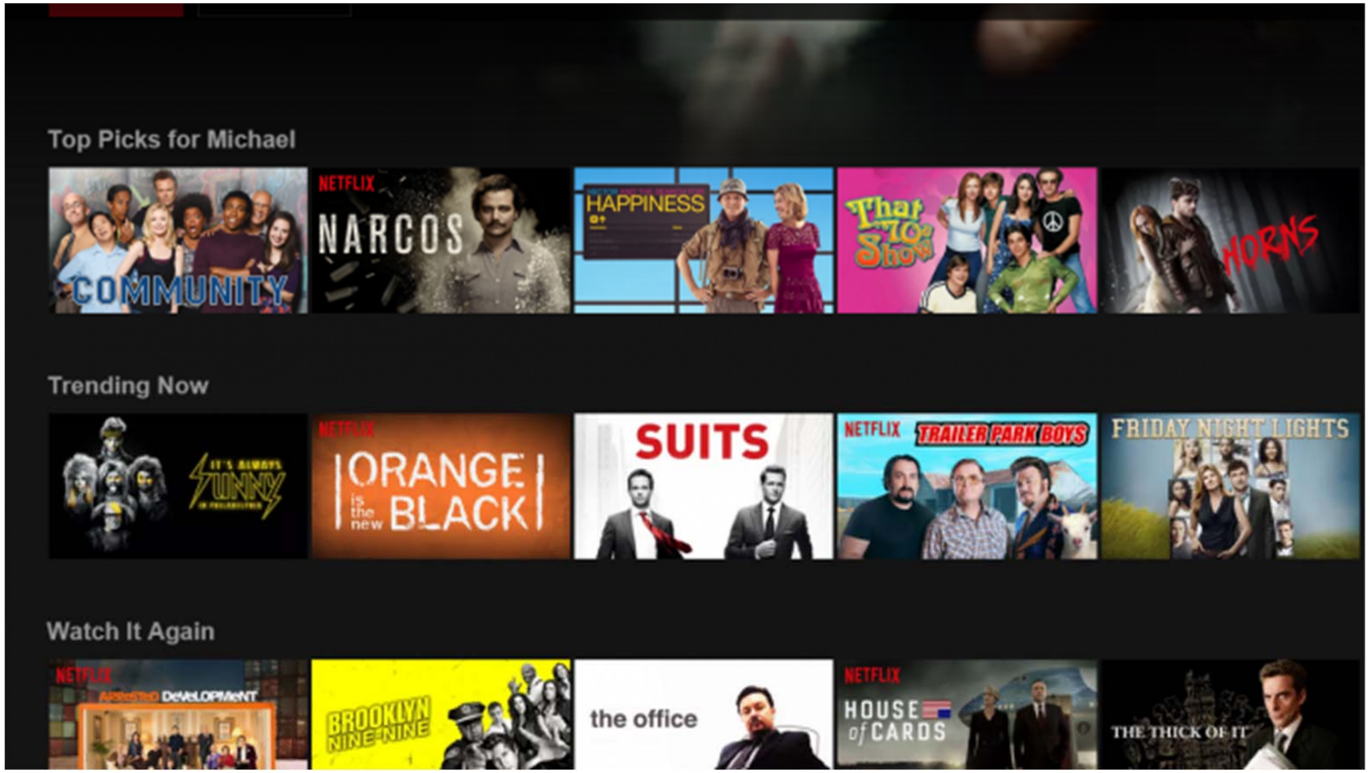
- A central part of our daily user experience
  - They help us locate potentially interesting things
  - They serve as filters in times of information overload
  - They have an impact user behavior and business

Recommendations everywhere

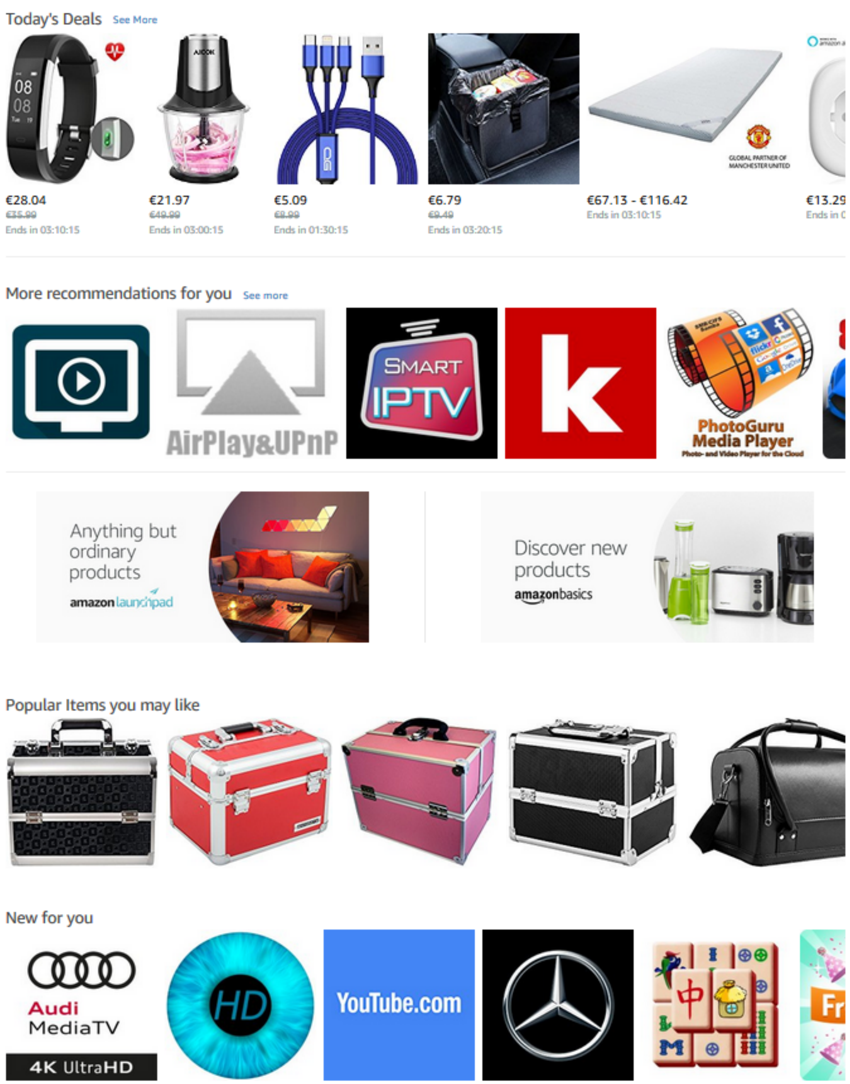




Recommendations everywhere



Recommendations everywhere



A field with a tradition

- 1970s: Early roots in IR and what was called "Selective Dissemination of Information"
- 1990s: A field develops, "content-based" approaches, Collaborative Filtering
- 2000s and beyond: The Netflix Prize and its implications
- Today and the future:
  - Deep learning everywhere
  - But are we focusing on the most important problems?

The recommendation problem

- A very general definition:
  - "Find a good/optimal selection of items to place in the recommendation list(s) of users"
- The corresponding questions:
  - What determines a good/optimal selection?
    - Help users find something new?
    - Show the user alternatives to a certain item?
    - The diversity of the recommendations?
  - Good or optimal for whom?
    - The consumer, the platform or retailer, the manufacturer, all of them?

Jannach, D. and Adomavicius, G.: "***Recommendations with a Purpose***". In: Proceedings of the 10th ACM Conference on Recommender Systems (RecSys 2016). Boston, Massachusetts, USA, 2016, pp. 7-10

A common problem abstraction

- Recommendation as a matrix completion task

	Item1	Item2	Item3	Item4	Item5
Alice	5	?	4	4	?
User1	3	1	?	3	3
User2	?	3	?	?	5
User3	3	?	1	5	4
User4	?	5	5	?	1

- Goal:
  - Learn/Optimize a prediction function from the data
- Quality assessment:
  - Prediction error on the test data

But, think again of about this one

- Past ratings do not play an obvious role
- There's seemingly not even personalization
- Nonetheless, it is a key application example in the literature

Outline

- Characterizing the session-based recommendation problem
- Algorithmic approaches for “next event” predictions
  - Categorization
  - A performance comparison
- Session-aware recommendation in e-commerce
  - On short-term intents, reminders, trends and discounts

Outline

- Characterizing the session-based recommendation problem
- Algorithmic approaches for “next event” predictions
  - Categorization
  - A performance comparison
- Session-aware recommendation in e-commerce
  - On short-term intents, reminders, trends and discounts

Session-based Recommendation

- Instead of a rating matrix, we are given a sequentially ordered log of user interactions
  - e.g., item views, purchases, listening or viewing events, …
- In many cases,
  - user cannot be identified (first-time users, users not logged in)
  - no longer-term preference information is available
  - user interest/intention/preferences must be assessed from a small set of interactions

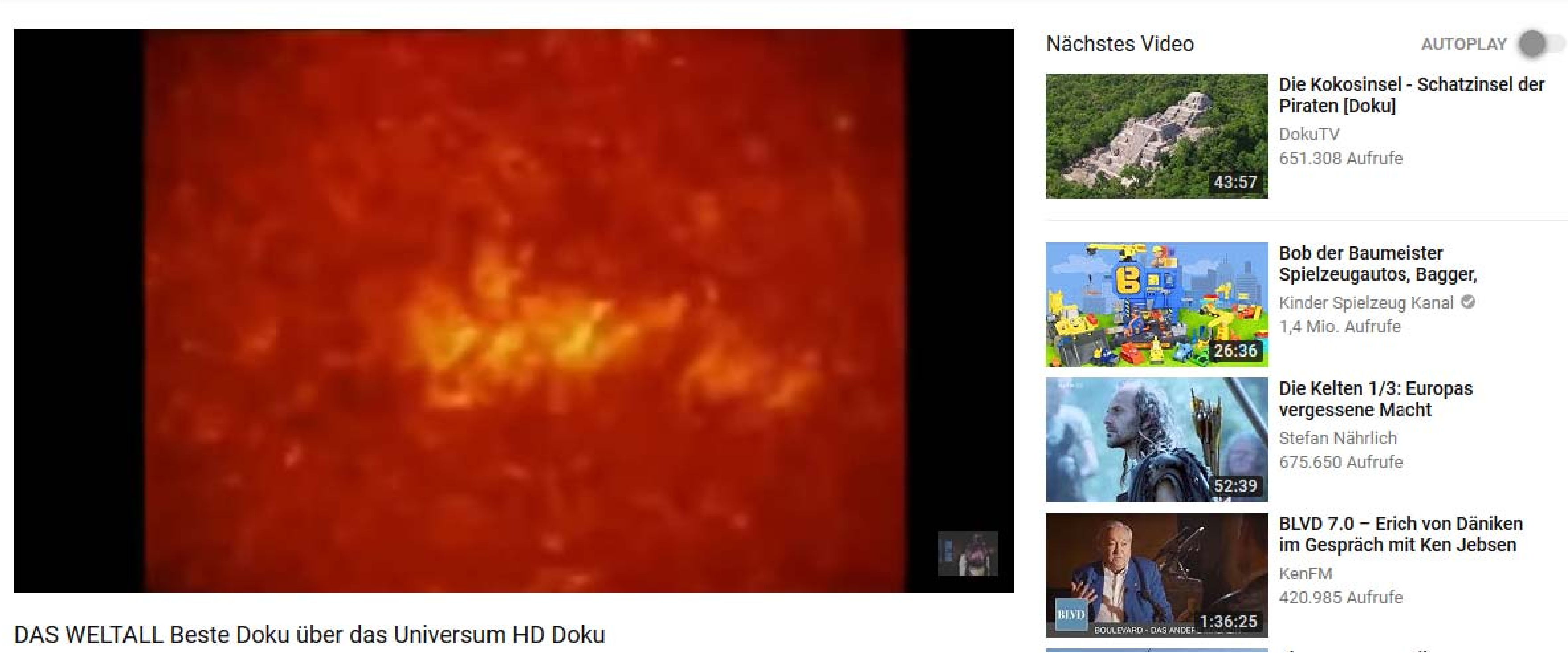
Session-based Recommendation

- Guessing the intention can be difficult

Session-based Recommendation

- Also in online music recommendation
- Our user searched and listened to "Last Christmas" by Wham!
- Should we, …
  - Play more songs by Wham!?
  - More pop Christmas songs?
  - More popular songs from the 1980s?
  - Play more songs with controversial user feedback?

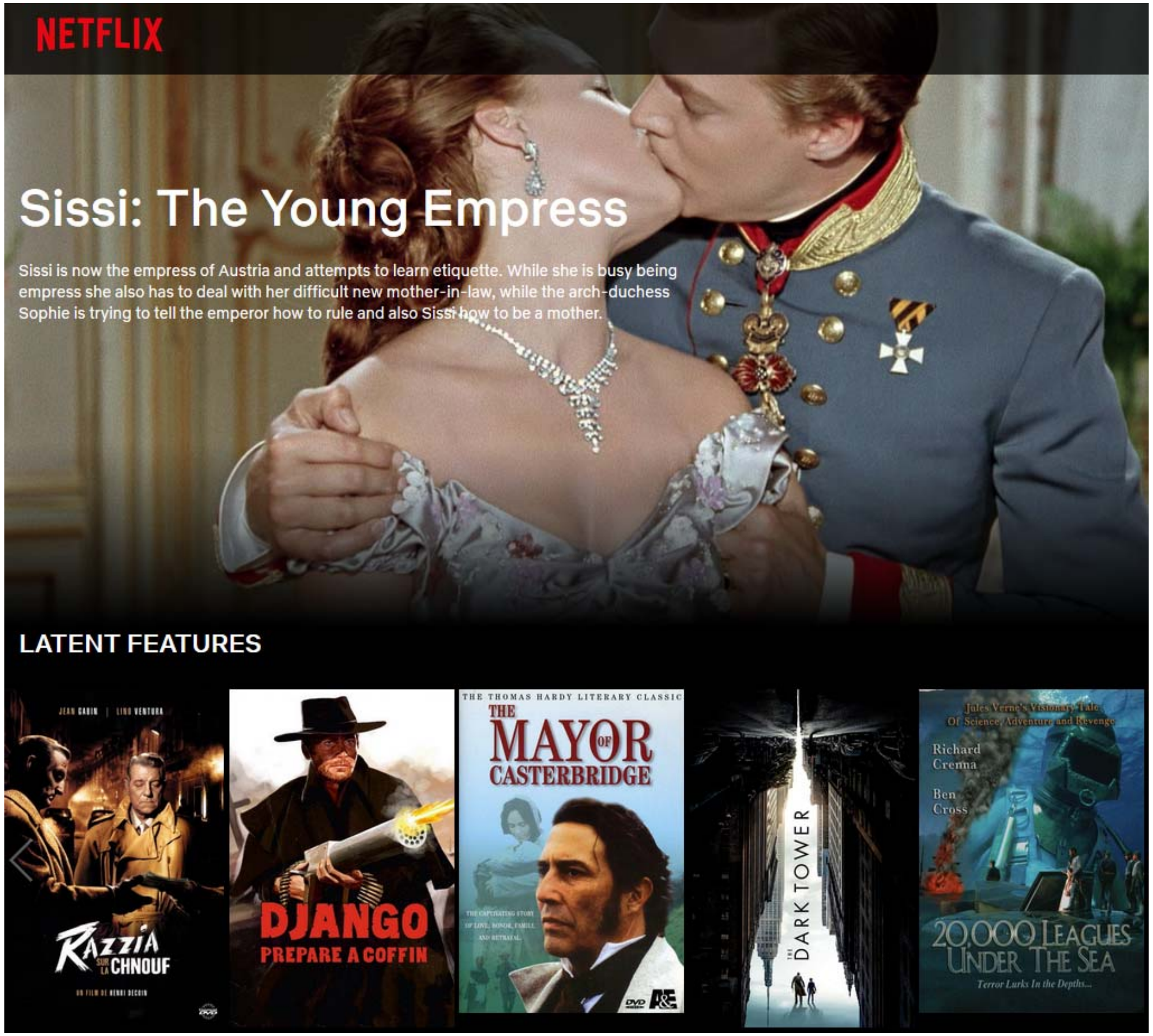
YouTube


Nächstes Video
AUTOPLAY
Die Kokosinsel - Schatzinsel der Piraten [Doku]
DokuTV
651.308 Aufrufe
43:57
Bob der Baumeister Spielzeugautos, Bagger,
Kinder Spielzeug Kanal
1,4 Mio. Aufrufe
26:36
Die Kelten 1/3: Europas vergessene Macht
Stefan Nährlich
675.650 Aufrufe
52:39
BLVD 7.0 – Erich von Däniken im Gespräch mit Ken Jebsen
KenFM
420.985 Aufrufe
1:36:25
DAS WELTALL Beste Doku über das Universum HD Doku

YouTube


YouTube
Suchen
Nächstes Video
AUTOPLAY
Der Schatz im Keltengrab
G.L.
118.235 Aufrufe
1:26:22
Mission Mars - Europas Raumfahrt zwischen Vision un...
ARTEde
64.579 Aufrufe
52:33
ERDE
AUCH DIE ERDE
13 der schaurigsten Theorien der Menschheit
SONNENSEITE
42.149 Aufrufe
Neu
11:14
Der wilde Pazifik - Deutsch 4K DOKU
BlackAngel1001
105.100 Aufrufe
52:05
Vom Rand der Erkenntnis • Stringtheorie • GUT • Weltform...
Urknall, Weltall und das Leben
702.028 Aufrufe
1:04:47
Flussmonster Teil 47 Der Humboldt Kalmar
Anime Mojo Movies 2018
381.531 Aufrufe
44:25
Feuerwehrmann Sam Unboxing: Jupiter Feuerwehrautos & neu...
Kinder Spielzeug Kanal
8,3 Mio. Aufrufe
10:31
1:19:04 / 3:42:38
DAS WELTALL Beste Doku über das Universum HD Doku
1.106.628 Aufrufe
1929
568
TEILEN
Lucy's Doku Channel
Am 10.02.2018 veröffentlicht
ABONNIEREN 6641

Similar Item Recommendations

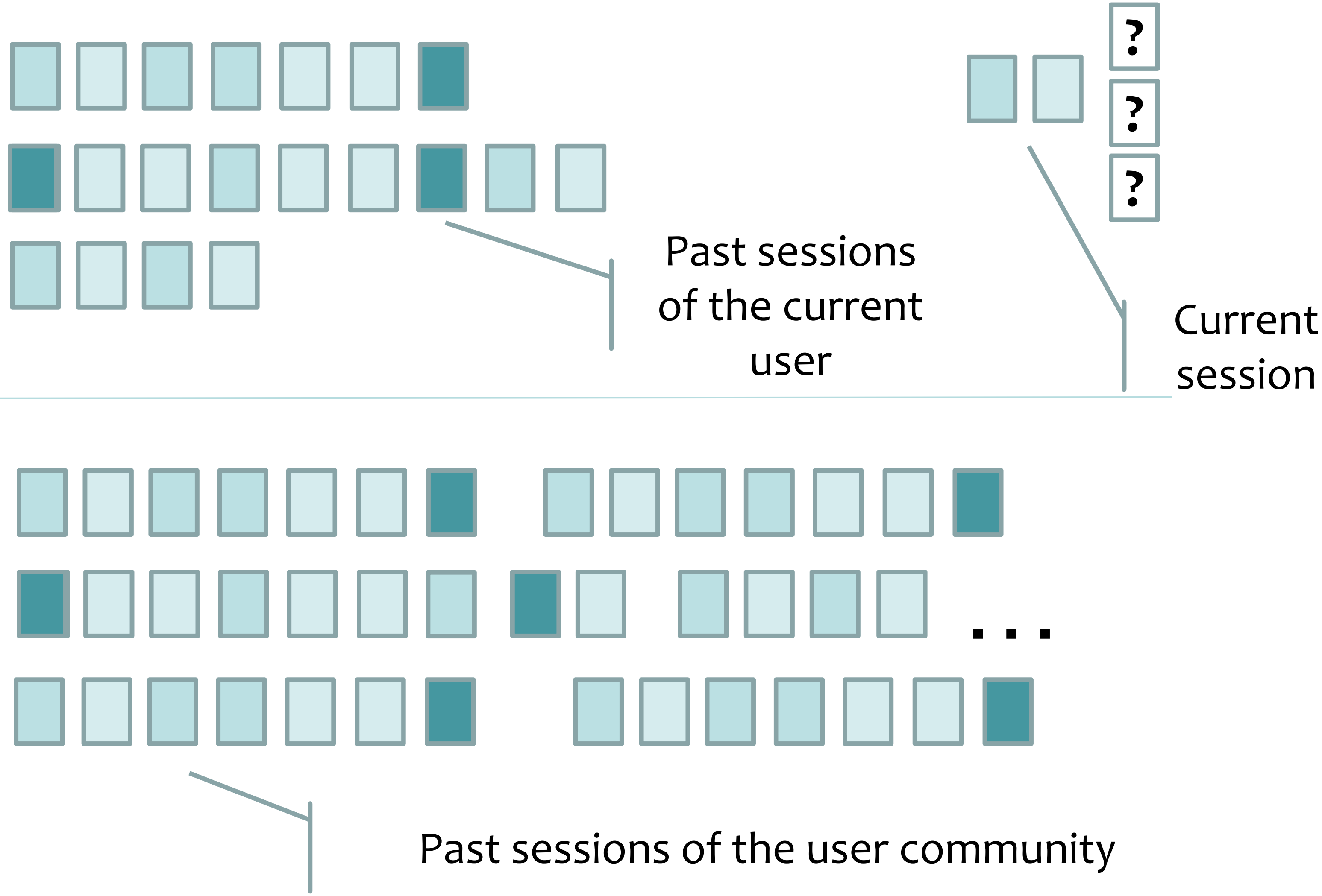


Session-aware Recommendation

- In some domains, past sessions of the current user are also known,
  - potential for personalization
  - possibility to remind users of objects
- We call "session-aware" recommendation
- Generally,
  - both session-based and session-aware recommendation as specific subtasks for sequence-aware recommenders

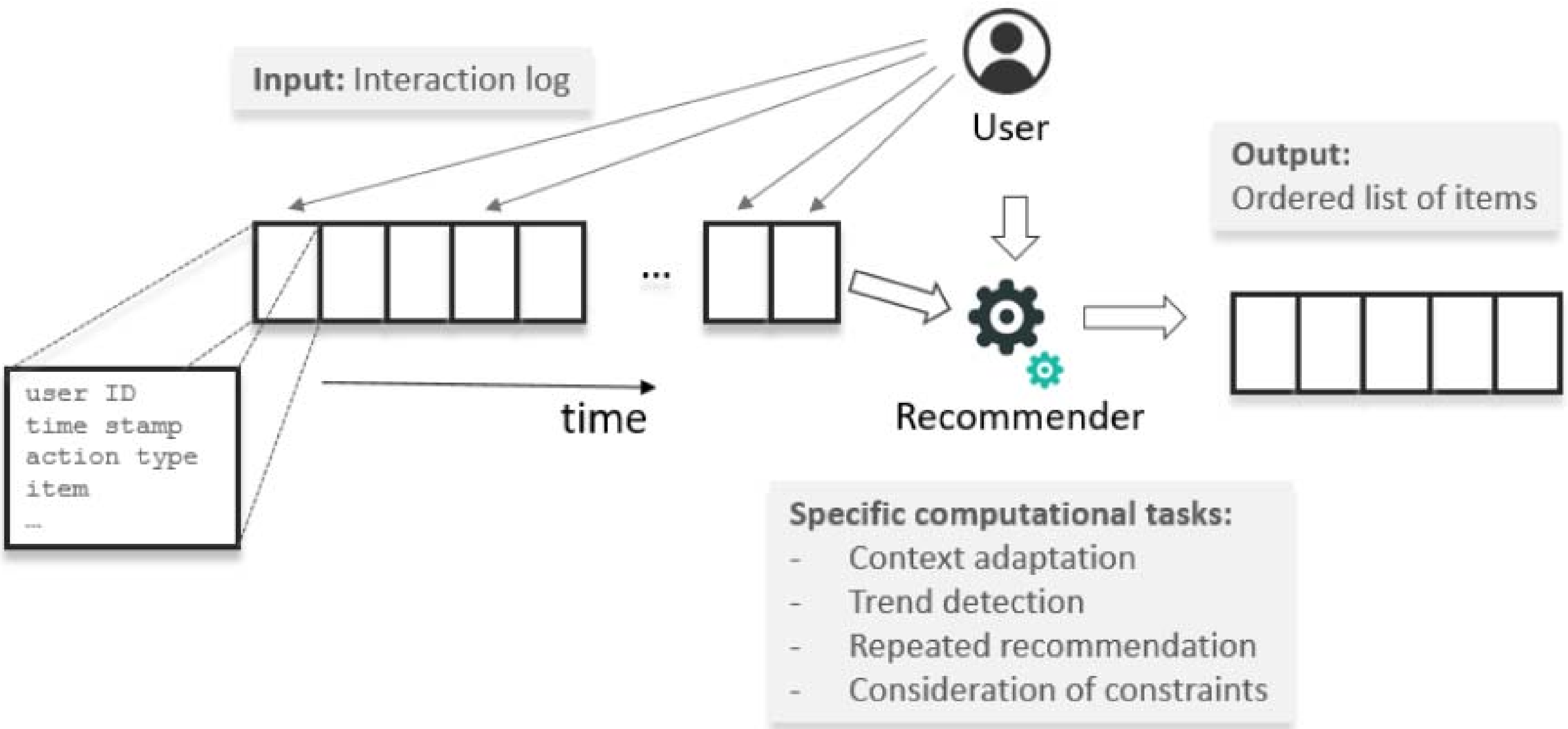
Quadrana, M., Cremonesi, P. and Jannach, D.: "**Sequence-Aware Recommender Systems**". ACM Computing Surveys. ACM, 2018

A Problem Abstraction



Sequence-aware Recommendation

- High-level overview
  - specific types of inputs, specific computational tasks



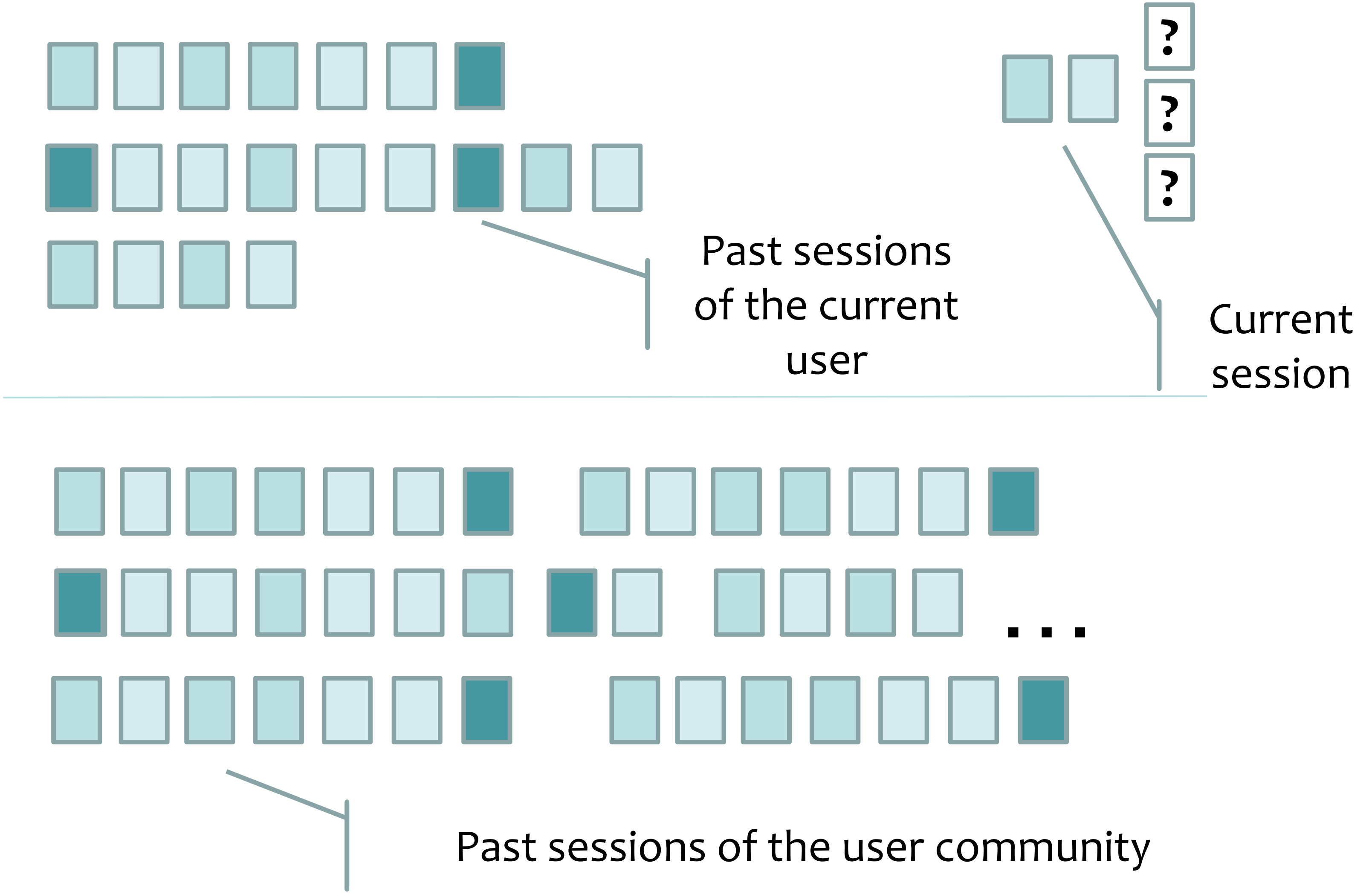
Outline

- Characterizing the session-based recommendation problem
- Algorithmic approaches for “next event” predictions
  - Categorization
  - A performance comparison
- Session-aware recommendation in e-commerce
  - On short-term intents, reminders, trends and discounts

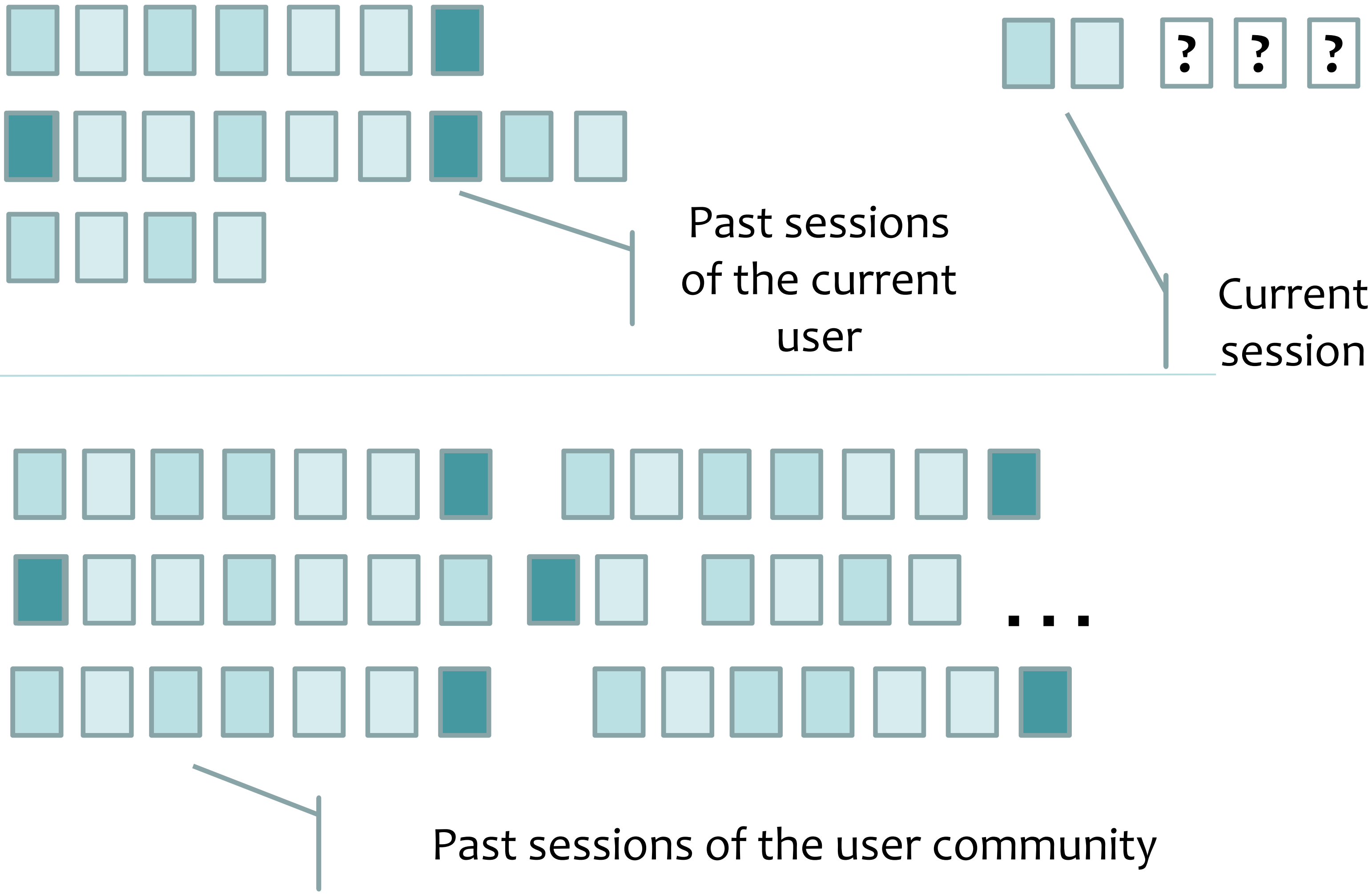
Operationalizing the Research Problem

- Background
  - Intention of user is not known, not clear if we should recommend more similar items or, e.g., accessories
- Computational task, simplified to:
  - Predict subsequent user action(s), given
    - the last N actions by the user (e.g., in the current session)
    - other types of information (community behavior, metadata, …)
- Evaluation
  - Use standard IR measures (precision, recall, MRR , …)
  - Some interaction log datasets are publicly available
    - But can be biased

A Problem Abstraction



A Problem Abstraction



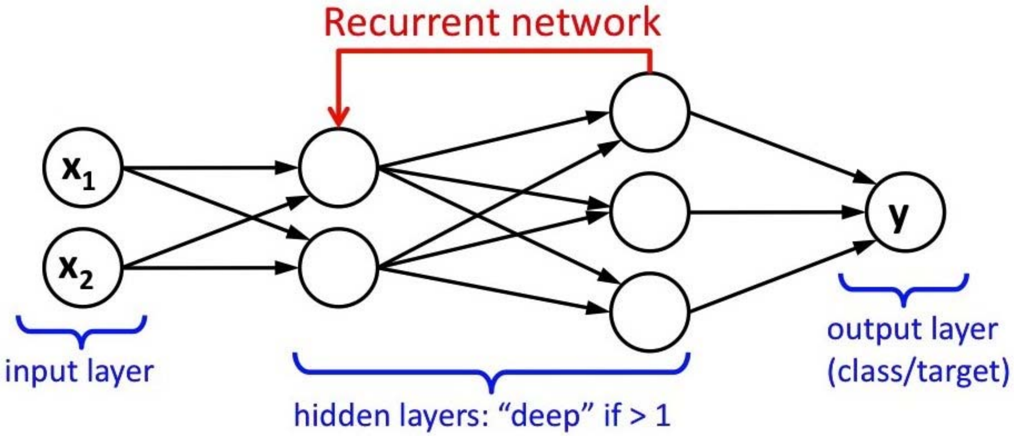
Basic Algorithmic Approaches

- Item co-occurrences in individual sessions
  - "Customers who bought", association rules of size 2
- Simple Markov Chains and "Sequential Rules"
  - Count how often appeared (immediately) after another in the training data
- Session-based nearest neighbors
  - Look for similar past sessions
  - Use a weighted prediction scheme
  - Different similarity functions possible
    - Sequence-agnostic and sequence-aware ones

Advanced Algorithmic Approaches

- Sequence-learning techniques
  - Frequent pattern mining
    - Frequent item sets, frequent sequential patterns
  - Sequence modeling
    - Markov Models, Recurrent Neural Networks
  - Distributed item representations
    - Distributional and Latent Markov embeddings
- Sequence-aware matrix factorization
- Hybrids
  - Factorized Markov Chains, others

Deep Learning



- Recurrent Neural Networks (RNN)
  - can learn from sequential data
  - a "natural choice" for the problem
- Recent algorithm: GRU4REC
  - Proposed by Hidasi et al. (2016)
    - Multiple improvements since then
  - Uses Gated Recurrent Units
  - Several technical innovations to ensure scalability

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. In ICLR '16.
Balázs Hidasi and Alexandros Karatzoglou. 2017. Recurrent Neural Networks with Top-k Gains for Session-based Recommendations, RecSys 2017

Performance Comparison

- Background
  - Today, no “standard” for benchmarking session-based algorithms exist
  - Researchers often use
    - different evaluation protocols and measures
    - different datasets
    - different baselines
  - This makes the assessment of the true value of new approaches difficult

On research methodology in ML

- Kaggle machine learning competitions
  - Defined dataset for training
  - Test dataset not revealed
  - Defined measures
  - Many competitors
  - (Code has to be made public)

- Academic machine learning research
  - Researcher picks dataset (often non-public)
  - Researcher knows test data
  - Researcher picks evaluation measure
  - Researcher picks competitors (baselines)
  - Researcher does not share code

On research methodology in ML

- Ranking of algorithms depends on
  - Dataset characteristics
  - Choice of measure and evaluation protocol
  - Choice of baselines
  - Parameterization of baselines
- “Improvement” not always clear

- Academic machine learning research
  - Researcher picks dataset (often non-public)
  - Researcher knows test data
  - Researcher picks evaluation measure
  - Researcher picks competitors (baselines)
  - Researcher does not share code

Literature

- **“Troubling Trends in Machine Learning Scholarship” by Lipton & Steinhardt:**
  - https://arxiv.org/abs/1807.03341
- **“Machine Learning that Matters” by Wagstaff**
  - ICML 2012
  - (Same for Deep Reinforcement Learning, AAAI 2018)
- **“Improvements that don't add up: ad-hoc retrieval results since 1998” by Armstrong et al.**
  - CIKM 2009

Performance Comparison

- Recently conducted an extensive set of experiments, including
  - both simple and sophisticated algorithms,
  - different datasets (e-commerce, music, news), and
  - different performance measures
    - including computational complexity
  - single-split and multiple-split evaluation protocol

Jannach, D. and Ludewig, M.: "**When Recurrent Neural Networks meet the Neighborhood for Session-Based Recommendation**". In: Proceedings of the 11th ACM Conference on Recommender Systems (RecSys 2017). Como, Italy, 2017

Ludewig, M., Jannach, D.: "**Evaluation of Session-based Recommendation Algorithms**", User Modeling and User-Adapted Interaction (forthcoming) https://arxiv.org/abs/1803.09587, 2018

Scalability Issues

- Naïve nearest-neighbor methods do not scale
- Our approach
  - Use compact in-memory data structures
    - Supporting fast look-up of possible neighbors
  - Use data sampling
    - E.g., 1,000 out of several million past sessions
    - Focus on most recent events
    - Only small accuracy compromises in many cases
- Results
  - Prediction time, e.g., at about 30 ms per request
  - Immediate "model updates" possible

Scalability Issues

- Complex methods
  - can require substantial amount of resources for training
  - but are usually fast at prediction time
- Example
  - E-commerce dataset, about 7 million sessions
  - A few hours for training (GPU-based)
    - Challenge lies in parameter optimization
    - New data usually requires full re-training
  - Main memory requirements largely depend on catalog size

Performance Comparison

- Main outcomes
  - In almost all configurations, one of the simple or almost trivial methods outperformed the most recent sequence learning method GRU4REC (2.0)
  - Much room for improvement regarding the development of more complex methods
    - e.g., hybrids

Ludewig, M., Jannach, D.: "**Evaluation of Session-based Recommendation Algorithms**". https://arxiv.org/abs/1803.09587, 2018

Side observations / remarks

- Careful choice of baselines needed
- Finding a good baselines is a not trivial
  - Ranking of algorithms varies across datasets
- Neighborhood-based baselines should be included in future experiments as well

Outline

- Characterizing the session-based recommendation problem
- Algorithmic approaches for “next event” predictions
  - Categorization
  - A performance comparison
- Session-aware recommendation in e-commerce
  - On short-term intents, reminders, trends and discounts

Session-Aware Recommendation

- Investigated various aspects regarding the success of recommenders in e-commerce
  - based on data shared to us by a large retailer
  - partly based on field test (A/B study)
- Specific questions
  - How important are long-term models?
  - Should we remind users of already seen items?
  - Can we leverage community trends in the recommendation process?
  - Should we recommend discounted items?

Jannach, D., Ludewig, M. and Lerche, L.: "**Session-based Item Recommendation in E-Commerce: On Short-Term Intents, Reminders, Trends, and Discounts**". User-Modeling and User-Adapted Interaction, Vol. 27(3-5). Springer, 2017, pp. 351-392

Long-term and short-term models

- Being able to predict which kinds of things a certain user generally likes, is important
- Here’s what the customer looked at or purchased during the last weeks



- Now, he or she return to the shop and browse these items



What to recommend?

- Some plausible options
  - Only shoes or only watches?
  - Mostly Nike shoes?
  - Maybe also some T-shirts?
- Using the matrix completion formulation
  - One trains a model based only on past actions
  - The context of the user’s current shopping intent is considered only in “context-aware” recommenders
  - Without the context, the algorithm will probably most recommend mostly T-shirts and trousers. Is this what you expect?

On short-term intents

- Research question:
  - What is the relative importance of adapting recommendations to users' short-term intents (shopping goals) when they visit the site?
- Measurement approach
  - "Hide-and-predict" simulation experiments on log data from a large online shop (Zalando)
  - Compare capability of session-aware and session-agnostic algorithms of predicting the purchased items in a given session

Jannach, D., Lerche, L. and Jugovac, M.: "**Adaptation and Evaluation of Recommendations for Short-term Shopping Goals**". In: RecSys 2015, pp. 211-218

Contextualization Strategies

- Various comparably simple “real-time” strategies tested, e.g.,
  - CoOccur, i.e., “Customers who bought … also bought”
- Feature Matching (FM)
  - Rank items up when they have features in common with those from the current session (e.g., same brand)
- Recently Viewed (RV)
  - Recommend recently viewed items in reverse chronological order

Technical approach

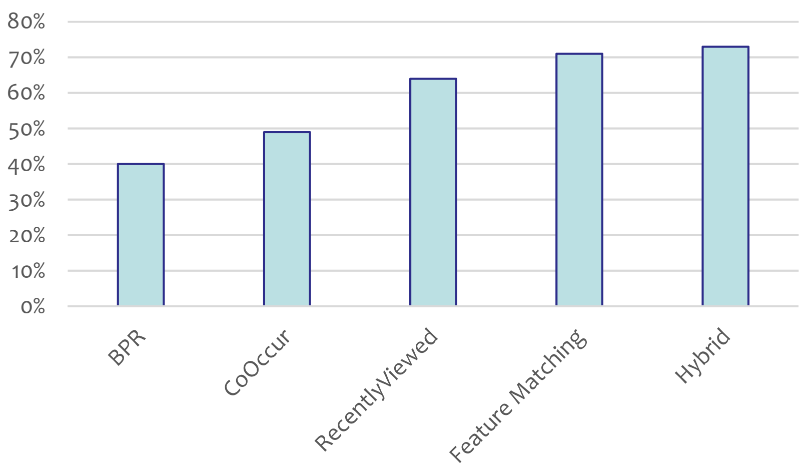
- Simple two-stage approach
  - Stage 1: Learn a long-term user profile, rank items
  - Stage 2: Filter or re-rank items based on the assumed short-term situation or intents
- Stage 1 can be done offline
  - Various algorithms were tested
    - Bayesian Personalized Ranking, Factorization Machines, Item-item-Nearest-Neighbors, Popularity-based and Random Baselines
- Stage 2 must be "real-time"
  - Can still be based on complex models, e.g., RNNs

Dataset

- Evaluations mostly based on an e-commerce log dataset
  - By Zalando, a major European online fashion retailer
  - Dataset contains sample of user activity logs
    - 1 million purchases
    - 20 million view events
    - 170,000 sessions
    - 800,000 users (many non-purchasers)
    - 150,000 different items
- Dataset is very sparse
  - Many users without any purchase

Empirical Results

- Observations for dense dataset (example)
  - Recall of best baseline method (BPR): 40%
  - Other:
    - Customers who bought … : 49%
    - Just show me what I have seen : 64%
    - Show me similar things : 71%
    - Combining long- and short-term: 73%



Insights

- Combination of various short-term and long-term signals as the most effective strategy
- Choice of baseline ranking method is relevant
  - Better baseline ranking in most cases leads to stronger overall results
- Importance of short-term adaptation
  - Contextualization-only methods often already better than the best long-term profile
  - Becomes more and more relevant, the more is known for the current session
  - Reminding is a very effective strategy

More on reminders

- Follow-up study
  - Deeper analysis of reminders
    - Using again the Zalando dataset
  - Development of more intelligent reminding strategies
  - Evaluation of reminding strategy in field test

Lerche, L., Jannach, D. and Ludewig, M.: ***"On the Value of Reminders within E-Commerce Recommendations"***. In: UMAP 2016, 2016.

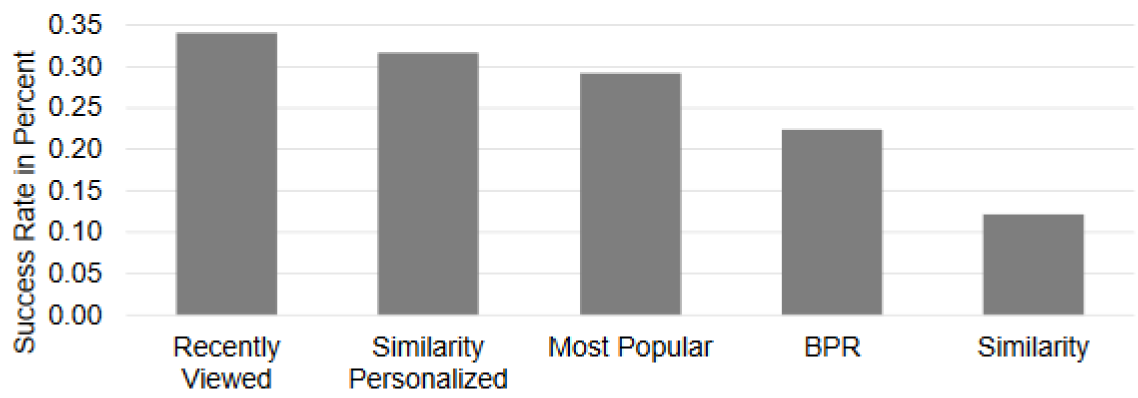
A field study on the business value

- A/B-tested different strategies on an e-commerce site for electronic gadgets
- Competing strategies
  - BPR as a learning-to-rank model
  - Similarity-based recommendation (using a reference item)
  - A personalized similarity-based approach
  - Popularity-based baseline
  - Present recently viewed items
    - In reverse chronological order

Recommendations in A/B test

Field study outcomes

- "Success rate" as business measure
  - Click on recommendation and click on outgoing link to external retailer
  - Pure reminders led to best business value in this specific situation

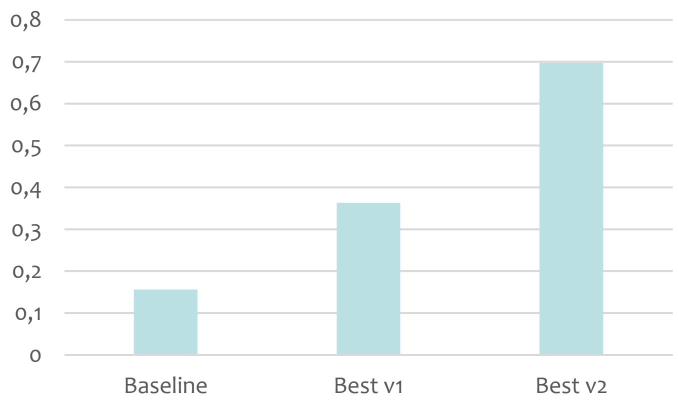


Better reminders?

- General filtering strategy
  - Do not remind users of items in categories where recently a purchase was made
- Designed different "adaptive" strategies
  - Recency-based baseline: Use reverse chronological order
  - Intensity-based ranking: Rank reminder items based on the number of past clicks
  - Item-similarity ranking: Select reminder items based on their fit for the current session
  - Session-similarity ranking: Select reminders based in their occurrence in similar past sessions

Empirical Evaluation

- Baseline ranking method:
  - Session-based nearest neighbors
    - Configured to include reminders as well
- Results (hit rate, example, 2 evaluation variants)
    - v1 hides view event for target item, v2 reveals them



- Adaptive reminders better than simple reminders

On Trends and Discounts

- More general question
  - What are factors that determine the success of a recommendation?
- Our dataset includes additional information:
  - For each view event, the three recommendations displayed on the item detail page
  - Click events on the recommended items
  - Information about discounts (visible to customers) at the time of recommendation
  - (Purchase information)

Research Goals

- Goals
  - Analyze which recommendations are successful, i.e., lead to a purchase event later on
  - Operationalize these insights in new recommendation algorithms

Analyzing the effectiveness of the recommendations on the site

- Reminders:
  - Only 10% of the recommendations seen before
  - But 44% of the successful ones were already known
- Short-term intents
  - Recommendations are more likely to be successful when from the same brand, category etc.
- Trends
  - Success rate of four times higher when the recommended item is trending on that day
- Discounts
  - Recommending on-sale items boosts the success rate

A systematic feature analysis

- Engineered about 90 features to predict the success of a recommendation
  - Framed as a classification problem
  - Systematically determined feature importance values

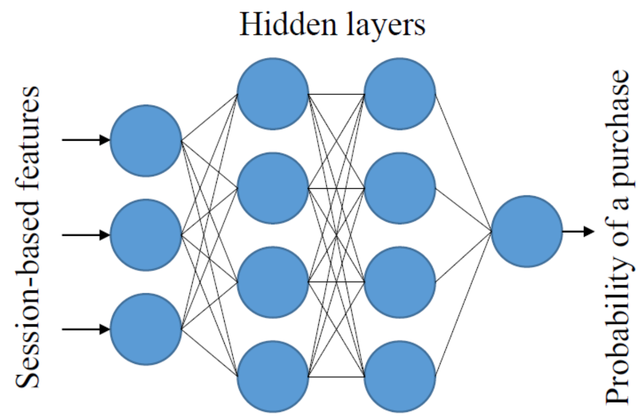
Feature	Gain Ratio	Chi Squared
Discount level	0.439	0.556
Current popularity (day)	0.371	1.000
Discount flag	0.325	0.556
Viewed before	0.286	0.435
Current popularity (week)	0.242	0.785
Distance to first item view (in days)	0.232	0.428
Distance to last item view (in days)	0.217	0.441
Distance to first item view (in sessions)	0.214	0.428
Distance to last item view (in sessions)	0.210	0.443
Current popularity (month)	0.201	0.563

Operationalization in algorithms

- Approach 1: A weighted two-phase approach
  - Create a candidate list of, e.g., 200 items using a baseline technique, in our case nearest-neighbors
  - Re-Rank items based on a weighted scoring scheme
  - Scoring functions
    - Content-wise similarity
    - Appearance in recent history (reminders)
    - Recent popularity on site
    - Level of discount
  - Importance weights have to be fine-tuned

Operationalization in algorithms

- Approach 2: A classification based approach
  - Framed scoring problem as a classification problem
  - Engineered 32 features in a similar way as was done for the feature importance analysis
    - Slightly different problem formulation
    - We are not interested analyzing success factors in general, but to predict the purchase probability for the given session
  - Used a deep neural-network approach for classification (H2O.ai library)
    - Outperformed Random Forests and manually tuned weights



Results

- Deep Learning based method led to best results
  - Independent of the chosen baseline ranking technique
- Random Forests were not better than the manually tuned weighted hybrid

Baseline	C-KNN		C-CoOcc		BPR	
Metric@10	HR	MRR	HR	MRR	HR	MRR
No post-processing	0.268	0.091	0.123	0.046	0.062	0.021
FM	0.281	0.093	0.145	0.052	0.119	0.046
IRec-FM	0.306	0.097	0.266	0.096	0.262	0.111
DR-FM	0.316	0.177	0.242	0.120	0.168	0.094
RPOP-FM	0.361	0.187	0.233	0.103	0.216	0.096
RFPREDICT	0.381	0.248	0.274	0.150	0.241	0.119
WR(RPOP,DR,0.5)-FM	0.382	0.220	0.262	0.121	0.225	0.100
DEEPPREDICT	**0.405**	**0.284**	0.322	0.205	0.301	0.188

General insights

- First approach in academia to “reconstruct” success factors of recommendations from log data
- Could successfully operationalize the insights in a new prediction method
- Feature engineering is important
- Domain-dependent aspects should be considered
  - Reminding or not
  - Recommending discounted items or not
  - Recommending trending items or not

Summary

- Session-based recommendation as a highly relevant problem in practice
- Recently increased interest
  - public datasets
  - “success” of RNNs
- Agreed-upon benchmark setup still needed
  - protocols, measures, baselines
- Domain-specific characteristics can be important (e.g., short-term community trends)

Outlook

- Better situation-dependent recommendations
  - Try to guess the customer’s decision and consumption situation
    - e.g., trying to learn what the options are, forming a choice set
    - e.g. listening to favorite artists, willing to explore new
  - Appropriate evaluation procedures needed, e.g., based on user studies
- More sophisticated algorithms for next-event prediction

Kamehkhosh, I. and Jannach, D.: "**User Perception of Next-Track Music Recommendations**". In: The 2017 Conference on User Modeling Adaptation and Personalization (UMAP 2017). Bratislava, Slovakia, 2017

- Thank you for your attention
- dietmar.jannach@aau.at

- Thanks to:
  - Paolo Cremonesi, Michael Jugovac, Iman Kamehkhosh, Lukas Lerche, Malte Ludewig, Massimo Quadrana