UNIVERSITÄT KLAGENFURT

Recommender Systems: Value, Methods, Measurements

Dietmar Jannach, University of Klagenfurt, Austria
dietmar.jannach@aau.at

Presented at the ACM Latin-American Summer School on Recommender Systems
LARS 2019, Fortaleza, Brazil

www.aau.at

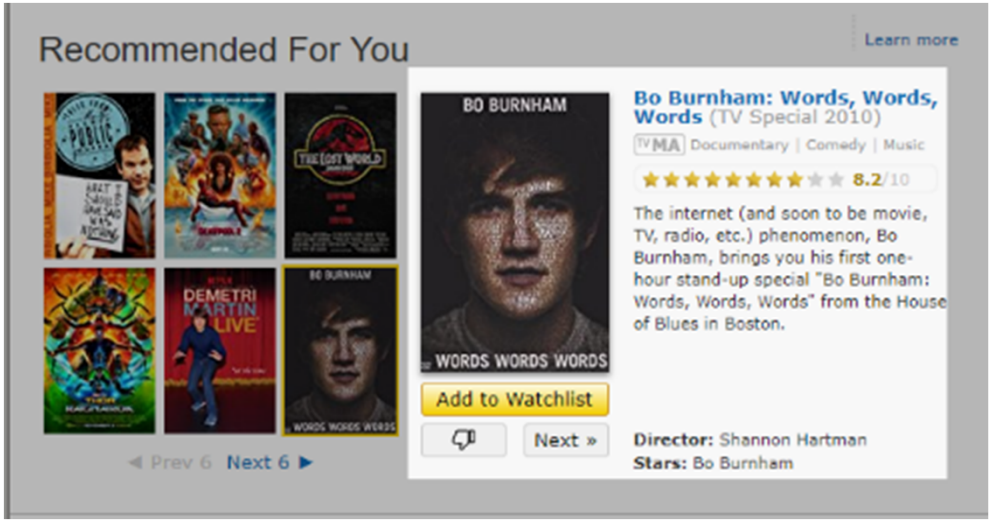
Outline

- What are Recommender Systems and what is their value?
- How do we build Recommender Systems and how do we know they work well?

- (Pointers to other lectures in the summer school)

Recommender Systems

- A pervasive part of our daily online user experience
- One of the most widely used applications of machine learning



Applications

- News
- Books
- Videos
- Music
- Games
- Shopping goods
- Friends
- Groups
- Jobs
- Apps
- Restaurants
- Hotels
- Deals
- Partners
- …
- Cigars
- Software code
- …

UNIVERSITÄT KLAGENFURT

The Value of Recommender Systems

www.aau.at

What's their purpose and value?

- Why should we use recommender systems?
  - Recommenders can have value both for consumers and the providers of the recommendations
  - Academic research (implicitly) mostly focuses on the consumer perspective
  - There can be even more stakeholders
    - Leading to multi-stakeholder recommendation problems
    - See also the lecture on Fairness in Recommender Systems

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

Potential value for the consumer

- Examples:
  - Help users find objects that match their long-term preferences (information filtering)
  - Help users explore the item space and improve decision making
  - Make contextual recommendations, e.g.,
    - Show alternatives
    - Show accessories
  - Remind users of what they liked in the past
  - Actively notify consumers of relevant content
  - Establish group consensus

Potential value for the provider

- Examples:
  - Change user behavior in desired directions
  - Create additional demand
  - Increase (short term) business success
  - Enable item "discoverability"
  - Increase activity on the site and user engagement
  - Provide a valuable add-on service
  - Learn more about the customers

Multi-stakeholder considerations

- When goals are fully aligned
  - Better recommendations can lead to more satisfied, returning customers who find what they need
  - This is one implicit assumption of academic research
- When there can be a goal conflict
  - Not all recommendable items may have the same business value
  - From a business perspective, it might be better to recommend items with a higher sales margin
    - As long as the recommendations are still reasonable

Measuring the business value

- Typical quotes about value

  "35% of Amazon.com's revenue is generated by its recommendation engine."

  "We think the combined effect of personalization and recommendations save us more than $1B per year."

  "Netflix says 80 percent of watched content is based on algorithmic recommendations"

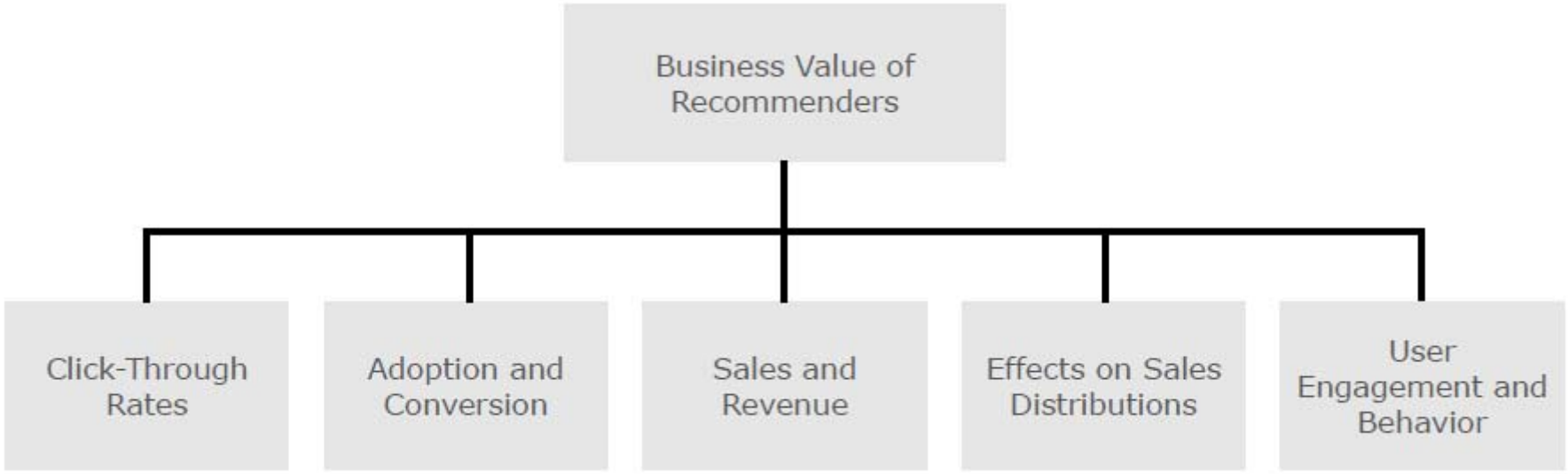
Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix Recommender System: Algorithms, Business Value, and Innovation. ACM Trans. Manage. Inf. Syst. 6, 4, Article 13

Measuring the business value

- Measuring the business value can be difficult
  - What does it tell us that 80% of the watched content comes from the recommendations?
  - Where do the said savings come from?
- The used measures often largely depend on
  - The business model of the provider
  - The intended effects of the recommendations
  - Assumptions about consumer value

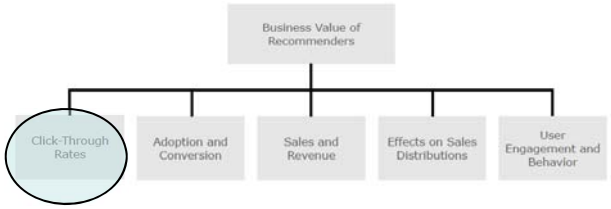
What is measured?

- Considering both the impact and value perspective



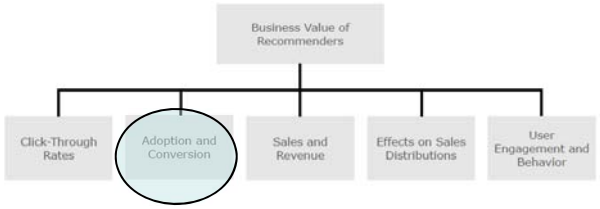
Jannach, D., Jugovac, M.; "Measuring the Business Value of Recommender Systems", arxiv preprint, https://arxiv.org/pdf/1908.08328.pdf

Click-Through Rates



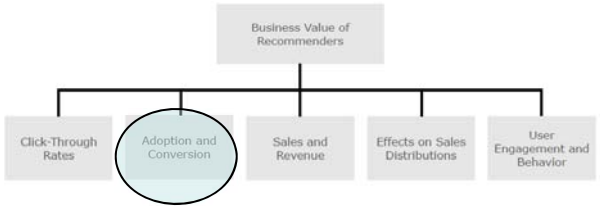
- Measures how many clicks are garnered by recommendations
  - Popular in the news recommendation domain
    - Google News: 38% more clicks compared to popularity-based recommendations
    - Forbes: 37% improvement through better algorithm compared to time-decayed popularity based method
    - swissinfo.ch: Similar improvements when considering only short-term navigation behavior
  - YouTube: Almost 200% improvement through co-visitation method (compared to popular recommendations)

Adoption and Conversion Rates



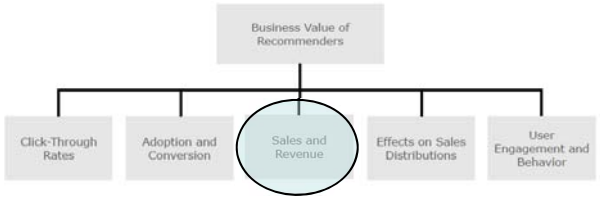
- CTR usually not the ultimate measure
  - Cannot know if users actually liked/purchased what they clicked on (consider also: click bait)
- Therefore
  - Various, domain-specific adoption measures common
- YouTube, Netflix: "Long CTR"/ "Take rate"
  - only count click if certain amount of video was watched

Adoption and Conversion Rates



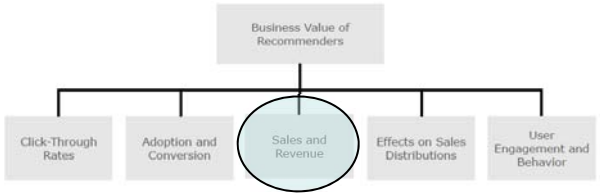
- Alternatives when items cannot be viewed/read:
- eBay:
  - "purchase-through-rate", "bid-through-rate"
- Other:
  - LinkedIn: Contact with employer made
  - Paper recommendation: "link-through", "cite-through"
  - E-Commerce marketplace: "click-outs"
  - Online dating: "open communications", "positive contacts per user"

Sales and Revenue



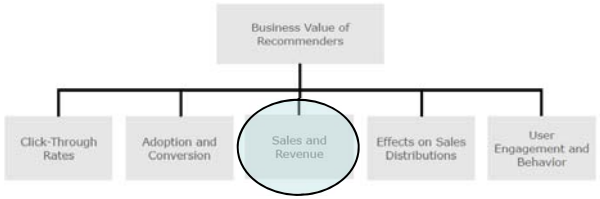
- CTR and adoption measures are good indicators of relevant recommendations
- However:
  - Often unclear how this translates into business value
  - Users might have bought an item anyway
  - Substantial increases might be not relevant for business when starting from a very low basis
- In addition:
  - Problem of measuring effects with flat-rate subscription models (e.g., Netflix).

Sales and Revenue



- Only a few studies, some with limitations
  - Video-on-demand study: 15% sales increase after introduction (no A/B test, could be novelty effect)
  - DVD retailer study:
    - 35% lift in sales when using purchased-based recommendation method compared to “no recommendations”
    - Almost no effects when recommendations were based on view statistics
    - Choice of algorithm matters a lot

Sales and Revenue

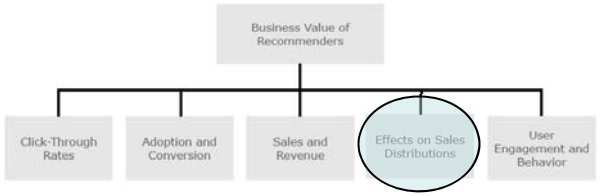


- e-grocery studies:
  - 1.8 % direct increase in sales in one study
  - 0.3 % direct effects in another study
  - However:
    - Up to 26% indirect effects, e.g., where customers were pointed to other categories in the store
    - "Inspirational" effect also observed in music recommendation in our own work
- eBay:
  - 6 % increase for similar item recommendations through largely improved algorithm
  - (500 % increase in other study for specific area)

Sales and Revenue

- Book store study:
  - 28 % increase with recommender compared with “no recommender”; could be seasonal effects
  - Drop of 17 % after removing the recommender
- Mobile games (own study)
  - 3.6 % more purchases through best recommender
  - More possible

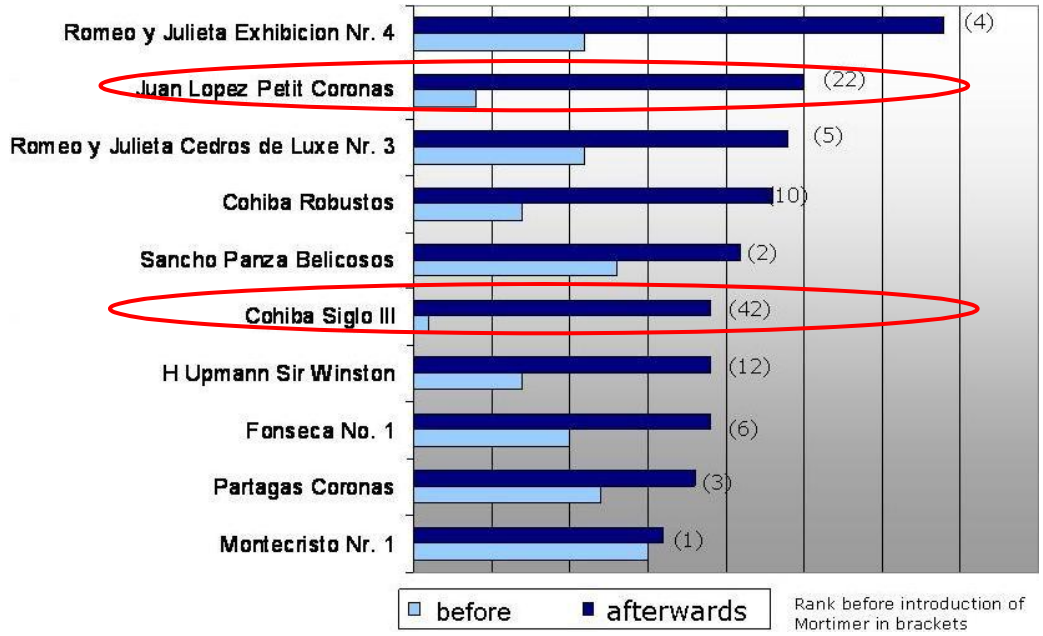
Effects on Sales Distributions



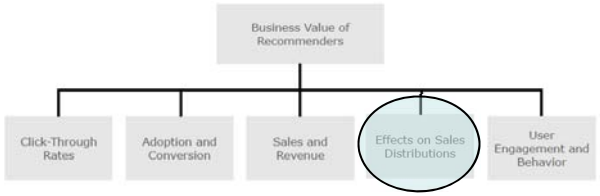
- Goal is maybe not to sell *more* but *different* items
- Influence sales behavior of customers
  - stimulate cross-sales
  - sell off on-stock items
  - promote items with higher margin
  - long-tail recommendations

Effects on Sales Distributions

- Premium cigars study:
  - Interactive advisory system installed
  - Measurable shift in terms of what is sold
    - e.g., due to better-informed customers

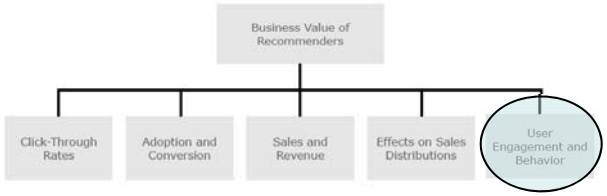


Effects on Sales Distributions



- Netflix:
  - Measure the "effective catalog size", i.e., how many items are actually (frequently) viewed
  - Recommenders lead users away from blockbusters
    - Could also be beneficial in terms of license costs
- Online retailer study:
  - Comparison of different algorithms on sales diversity
  - Outcomes
    - Recommenders tend to **decrease** the overall diversity
    - Might increase diversity at individual level though

User Behavior and Engagement



- Assumption:
  - Higher engagement leads to higher re-subscription rates (e.g., at Spotify)
- News domain studies:
  - 2.5 times longer sessions, more sessions when there is a recommender
- Music domain study:
  - Up to 50% more user activity
- LinkedIn:
  - More clicks on job profiles after recommender introduced

Discussion

- Direct measurements:
  - Business value can almost be directly measured
  - Limitations
    - High revenue might be easy to achieve (promote discounted products), but not the business goal
    - Field tests often last only for a few weeks; field tests sometimes only with new customers (e.g., at Netflix)
    - Long-term indirect effects might be missed

Discussion

- Indirect measurements:
  - CTR considered harmful
    - Recommendations as click-bait, but long term dissatisfaction possible
    - CTR optimization not in line with optimization for customer relevance
    - CTRs and improvements often easy to achieve, e.g., by changing the user interface or by focusing on already popular items
  - Adoption and conversion
    - Mobile game study: Clicks and certain types of conversions were not indicative for business value
  - Engagement
    - Difficult to assess when churn rates are already low

What to measure?

- The underlying questions:
  - What is the intended purpose of the system?
  - What kind of value should it create?
- Leading to:
  - What is a good recommendation in this context, i.e. one that serves any or all of these goals?

What to measure?

- Beware:
  - The same set of recommendations can be good or not, depending on the purpose, context, and application, e.g.,
    - Recommending already popular items can be good for the business or not
    - Recommending things, for example musical songs, that the user already knows can be desirable or not, depending on the user's mood
    - Recommending a set of items that are very similar to each other might be helpful for the user or not, depending on their stage in the decision making process

The academic perspective

- In academia, we aim to
  - abstract from application specifics, and
  - develop generalizable methods

The predominant approach

- Most common task: “Find good items”
- Most common method: “offline experimentation” and accuracy optimization
- Approach
  - Find or create a dataset that contains historical information about which recommendable items were considered “good” for individual users
  - Hide some of the information
  - Predict the hidden information
  - Measure the accuracy of the predictions

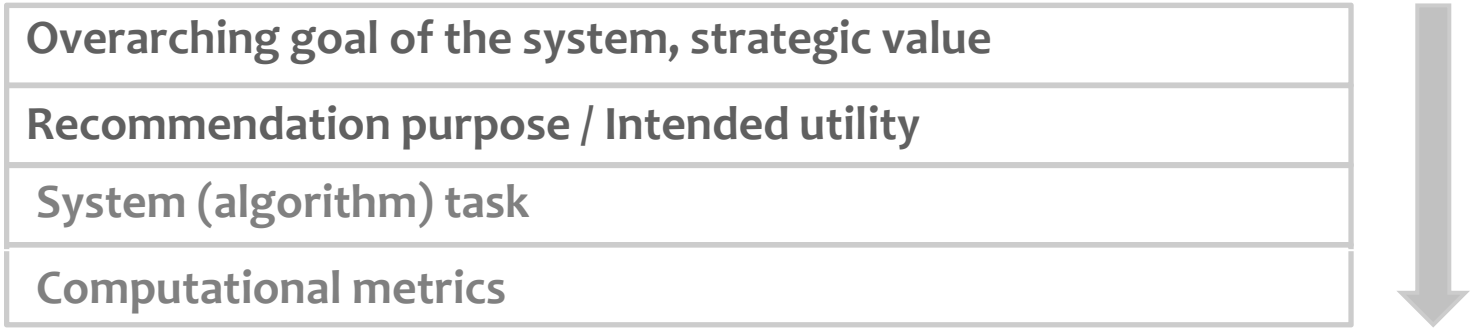
Benefits & Limitations

- Benefits of this approach
  - Well-defined problem
  - Continuous improvement
  - Comparability & reproducibility
- Potential limitations
  - Being accurate is not enough, and higher accuracy not necessarily means better value for the user
  - The value for other stakeholders is not considered
  - Over-simplification of the problem

Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06).

A conceptual framework

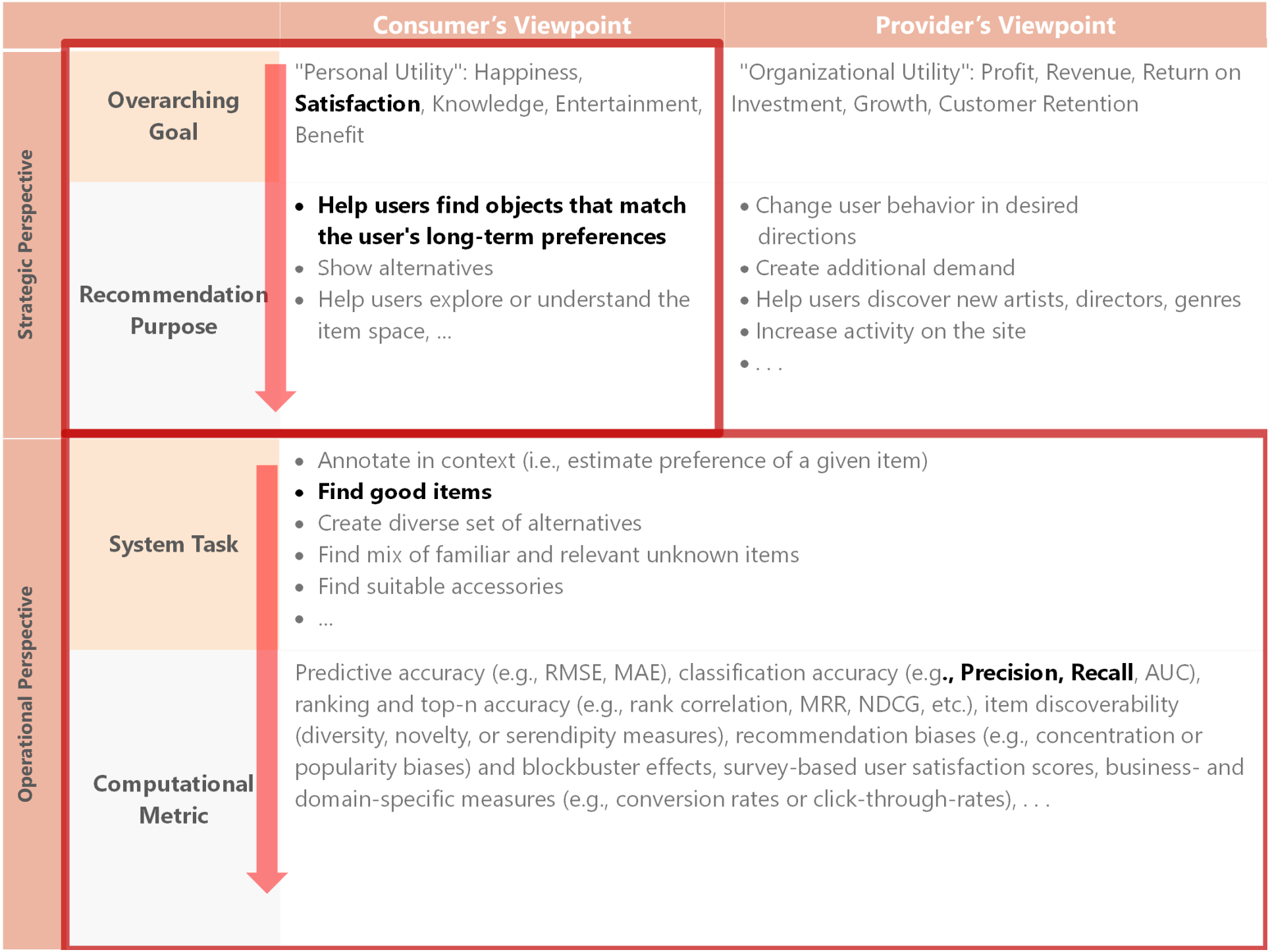
- Should help to decide what and how to measure (both in academia and industry)
- Layered structure – strategic to operational
- Considers two viewpoints

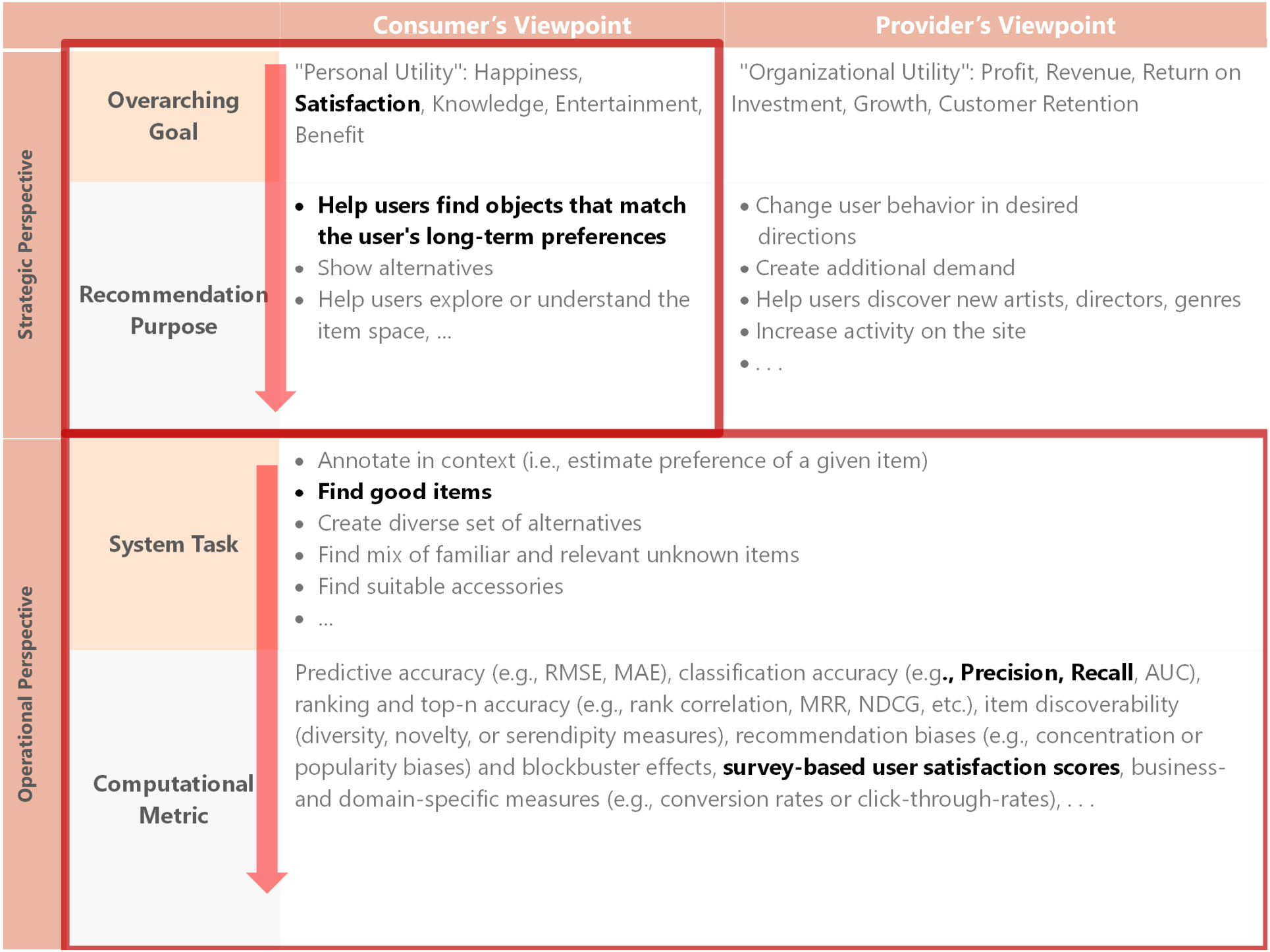


Framework overview

		Consumer's Viewpoint	Provider's Viewpoint
Strategic Perspective	Overarching Goal	"Personal Utility": Happiness, Satisfaction, Knowledge, ...	"Organizational Utility": Profit, Revenue, Growth, ...
	Recommendation Purpose	• Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space • ...	• Change user behavior in desired directions • Create additional demand • Increase activity on the site • ...
Operational Perspective	System Task	• Annotate in context (i.e., estimate preference of a given item) • Find good items • Create diverse set of alternatives • Find suitable accessories • Retrieve novel but relevant items • ...	
	Computational Metric	Predictive accuracy (e.g., RMSE, MAE), classification accuracy (e.g., precision, recall, AUC), ranking and top-n accuracy (e.g., rank correlation, MRR, NDCG, etc.), item "discoverability" (diversity, novelty, or serendipity measures), recommendation biases (e.g., concentration or popularity biases) and blockbuster effects, survey-based user satisfaction scores, business- and domain-specific measures (e.g., conversion rates or click-through-rates), ...	

		Consumer's Viewpoint	Provider's Viewpoint
Strategic Perspective	**Overarching Goal**	"Personal Utility": Happiness, Satisfaction, Knowledge, Entertainment, Benefit	"Organizational Utility": Profit, Revenue, Return on Investment, Growth, Customer Retention
	Recommendation Purpose	• Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space, ...	• Change user behavior in desired directions • Create additional demand • Help users discover new artists, directors, genres • Increase activity on the site • . . .
Operational Perspective	**System Task**	• Annotate in context (i.e., estimate preference of a given item) • **Find good items** • Create diverse set of alternatives • Find mix of familiar and relevant unknown items • Find suitable accessories • ...	
	Computational Metric	Predictive accuracy (e.g., RMSE, MAE), classification accuracy (e.g**., Precision, Recall**, AUC), ranking and top-n accuracy (e.g., rank correlation, MRR, NDCG, etc.), item discoverability (diversity, novelty, or serendipity measures), recommendation biases (e.g., concentration or popularity biases) and blockbuster effects, survey-based user satisfaction scores, business- and domain-specific measures (e.g., conversion rates or click-through-rates), . . .	


Consumer's Viewpoint
Provider's Viewpoint
Strategic Perspective
Overarching Goal
"Personal Utility": Happiness, Satisfaction, Knowledge, Entertainment, Benefit
"Organizational Utility": Profit, Revenue, Return on Investment, Growth, Customer Retention
Recommendation Purpose
• Help users find objects that match the user's long-term preferences
• Show alternatives
• Help users explore or understand the item space, ...
• Change user behavior in desired directions
• Create additional demand
• Help users discover new artists, directors, genres
• Increase activity on the site
• . . .
Operational Perspective
System Task
• Annotate in context (i.e., estimate preference of a given item)
• Find good items
• Create diverse set of alternatives
• Find mix of familiar and relevant unknown items
• Find suitable accessories
• ...
Computational Metric
Predictive accuracy (e.g., RMSE, MAE), classification accuracy (e.g., Precision, Recall, AUC), ranking and top-n accuracy (e.g., rank correlation, MRR, NDCG, etc.), item discoverability (diversity, novelty, or serendipity measures), recommendation biases (e.g., concentration or popularity biases) and blockbuster effects, survey-based user satisfaction scores, business- and domain-specific measures (e.g., conversion rates or click-through-rates), . . .


Consumer's Viewpoint
Provider's Viewpoint
Strategic Perspective
Overarching Goal
"Personal Utility": Happiness, Satisfaction, Knowledge, Entertainment, Benefit
"Organizational Utility": Profit, Revenue, Return on Investment, Growth, Customer Retention
Recommendation Purpose
• Help users find objects that match the user's long-term preferences
• Show alternatives
• Help users explore or understand the item space, ...
• Change user behavior in desired directions
• Create additional demand
• Help users discover new artists, directors, genres
• Increase activity on the site
• . . .
Operational Perspective
System Task
• Annotate in context (i.e., estimate preference of a given item)
• Find good items
• Create diverse set of alternatives
• Find mix of familiar and relevant unknown items
• Find suitable accessories
• ...
Computational Metric
Predictive accuracy (e.g., RMSE, MAE), classification accuracy (e.g., Precision, Recall, AUC), ranking and top-n accuracy (e.g., rank correlation, MRR, NDCG, etc.), item discoverability (diversity, novelty, or serendipity measures), recommendation biases (e.g., concentration or popularity biases) and blockbuster effects, survey-based user satisfaction scores, business- and domain-specific measures (e.g., conversion rates or click-through-rates), . . .

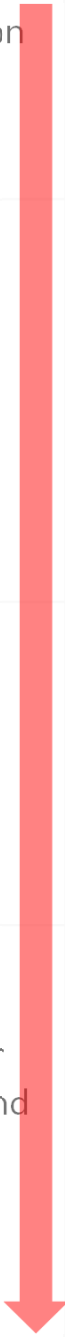
		Consumer's Viewpoint	Provider's Viewpoint
Strategic Perspective	**Overarching Goal**	"Personal Utility": Happiness, Satisfaction, Knowledge, Entertainment, Benefit	"Organizational Utility": Profit, Revenue, Return on Investment, Growth, Customer Retention
	Recommendation Purpose	• Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space, ...	• Change user behavior in desired directions • Create additional demand • Help users discover new artists, directors, genres • Increase activity on the site • . . .
Operational Perspective	**System Task**	• Annotate in context (i.e., estimate preference of a given item) • Find good items • Create diverse set of alternatives • Find mix of familiar and relevant unknown items • Find suitable accessories • ...	
	Computational Metric	Predictive accuracy (e.g., RMSE, MAE), classification accuracy (e.g., Precision, Recall, AUC), ranking and top-n accuracy (e.g., rank correlation, MRR, NDCG, etc.), item discoverability (diversity, novelty, or serendipity measures), recommendation biases (e.g., concentration or popularity biases) and blockbuster effects, survey-based user satisfaction scores, business- and domain-specific measures (e.g., conversion rates or click-through-rates), . . .	

		Consumer's Viewpoint	Provider's Viewpoint
Strategic Perspective	**Overarching Goal**	"Personal Utility": Happiness, Satisfaction, Knowledge, Entertainment, Benefit	"Organizational Utility": Profit, Revenue, Return on Investment, Growth, **Customer Retention**
	Recommendation Purpose	• Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space, ...	• Change user behavior in desired directions • Create additional demand • Help users discover new artists, directors, genres • Increase activity on the site • . . .
Operational Perspective	**System Task**	• Annotate in context (i.e., estimate preference of a given item) • Find good items • Create diverse set of alternatives • Find mix of familiar and relevant unknown items • Find suitable accessories • ...	
	Computational Metric	Predictive accuracy (e.g., RMSE, MAE), classification accuracy (e.g., Precision, Recall, AUC), ranking and top-n accuracy (e.g., rank correlation, MRR, NDCG, etc.), item discoverability (diversity, novelty, or serendipity measures), recommendation biases (e.g., concentration or popularity biases) and blockbuster effects, survey-based user satisfaction scores, business- and domain-specific measures (e.g., conversion rates or click-through-rates), . . .	

		Consumer's Viewpoint	Provider's Viewpoint
Strategic Perspective	**Overarching Goal**	"Personal Utility": Happiness, Satisfaction, Knowledge, Entertainment, Benefit	"Organizational Utility": Profit, Revenue, Return on Investment, Growth, **Customer Retention**
	Recommendation Purpose	• Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space, ...	• Change user behavior in desired directions • Create additional demand • **Help users discover new artists, directors, genres** • Increase activity on the site • . . .
Operational Perspective	**System Task**	• Annotate in context (i.e., estimate preference of a given item) • Find good items • Create diverse set of alternatives • Find mix of familiar and relevant unknown items • Find suitable accessories • ...	
	Computational Metric	Predictive accuracy (e.g., RMSE, MAE), classification accuracy (e.g., Precision, Recall, AUC), ranking and top-n accuracy (e.g., rank correlation, MRR, NDCG, etc.), item discoverability (diversity, novelty, or serendipity measures), recommendation biases (e.g., concentration or popularity biases) and blockbuster effects, survey-based user satisfaction scores, business- and domain-specific measures (e.g., conversion rates or click-through-rates), . . .	

		Consumer's Viewpoint	Provider's Viewpoint
Strategic Perspective	**Overarching Goal**	"Personal Utility": Happiness, Satisfaction, Knowledge, Entertainment, Benefit	"Organizational Utility": Profit, Revenue, Return on Investment, Growth, **Customer Retention**
	Recommendation Purpose	• Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space, ...	• Change user behavior in desired directions • Create additional demand • **Help users discover new artists, directors, genres** • Increase activity on the site • . . .
Operational Perspective	**System Task**	• Annotate in context (i.e., estimate preference of a given item) • Find good items • Create diverse set of alternatives • **Find mix of familiar and relevant unknown items** • Find suitable accessories • ...	
	Computational Metric	Predictive accuracy (e.g., RMSE, MAE), classification accuracy (e.g., Precision, Recall, AUC), ranking and top-n accuracy (e.g., rank correlation, MRR, NDCG, etc.), item discoverability (diversity, novelty, or serendipity measures), recommendation biases (e.g., concentration or popularity biases) and blockbuster effects, survey-based user satisfaction scores, business- and domain-specific measures (e.g., conversion rates or click-through-rates), . . .	

		Consumer's Viewpoint	Provider's Viewpoint
Strategic Perspective	**Overarching Goal**	"Personal Utility": Happiness, Satisfaction, Knowledge, Entertainment, Benefit	"Organizational Utility": Profit, Revenue, Return on Investment, Growth, **Customer Retention**
	Recommendation Purpose	• Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space, ...	• Change user behavior in desired directions • Create additional demand • **Help users discover new artists, directors, genres** • Increase activity on the site • . . .
Operational Perspective	**System Task**	• Annotate in context (i.e., estimate preference of a given item) • Find good items • Create diverse set of alternatives • **Find mix of familiar and relevant unknown items** • Find suitable accessories • ...	
	Computational Metric	**Predictive accuracy** (e.g., RMSE, MAE), classification accuracy (e.g., Precision, Recall, AUC), ranking and top-n accuracy (e.g., rank correlation, MRR, NDCG, etc.), item discoverability (diversity, **novelty, or serendipity measures**), recommendation biases (e.g., concentration or popularity biases) and blockbuster effects, survey-based user satisfaction scores, business- and domain-specific measures (e.g., conversion rates or click-through-rates), . . .**?**	



Summary of value considerations

- Demonstrated business value of recommenders in many domains
- Size of impact however depends on many factors like baselines, domain specifics etc.
- Measuring impact is generally not trivial
  - Choice of the evaluation measure matters a lot
  - CTR can be misleading
- “Metric-Task-Purpose-Fit” to be considered

UNIVERSITÄT KLAGENFURT

Methods

A common categorization

- Content-based Filtering
- Collaborative Filtering
- Hybrid Systems
- Knowledge-based Systems

Outline

- Content-based Filtering
- Collaborative Filtering
- Hybrid Systems
- Knowledge-based Systems

- Interactive Recommendation

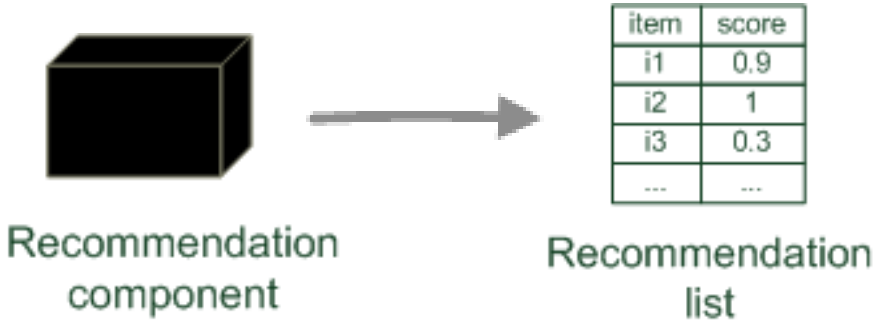
Outline

- Content-based Filtering
- Collaborative Filtering
- Hybrid Systems
- Knowledge-based Systems

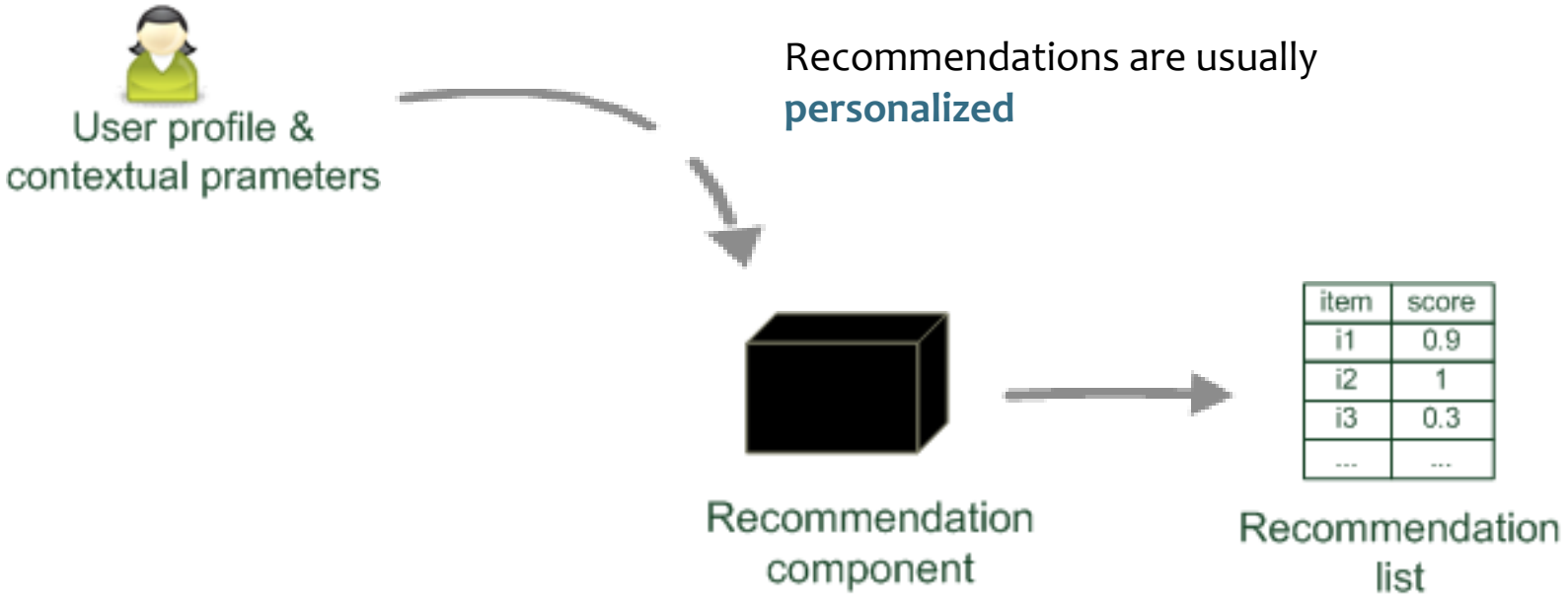
- Interactive Recommendation

Recommendation Principles

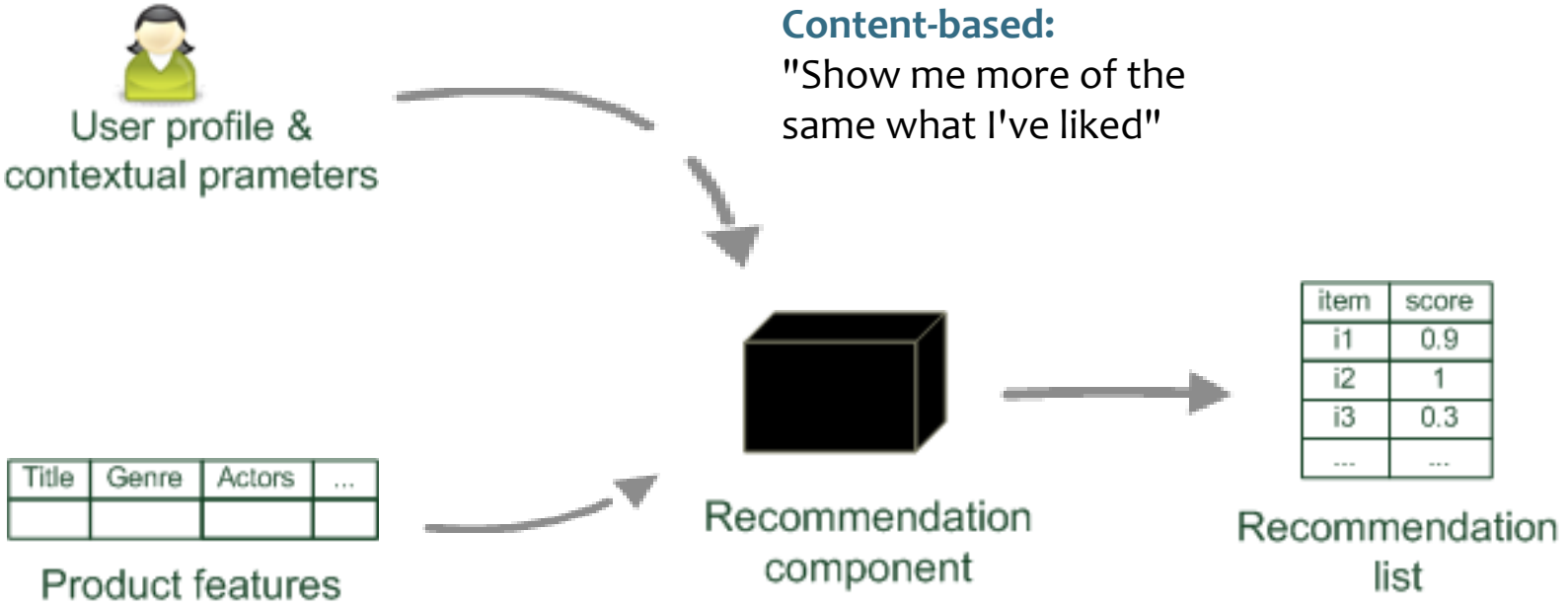
Recommender systems reduce information overload by estimating relevance



Recommendation Principles



Content-based Filtering

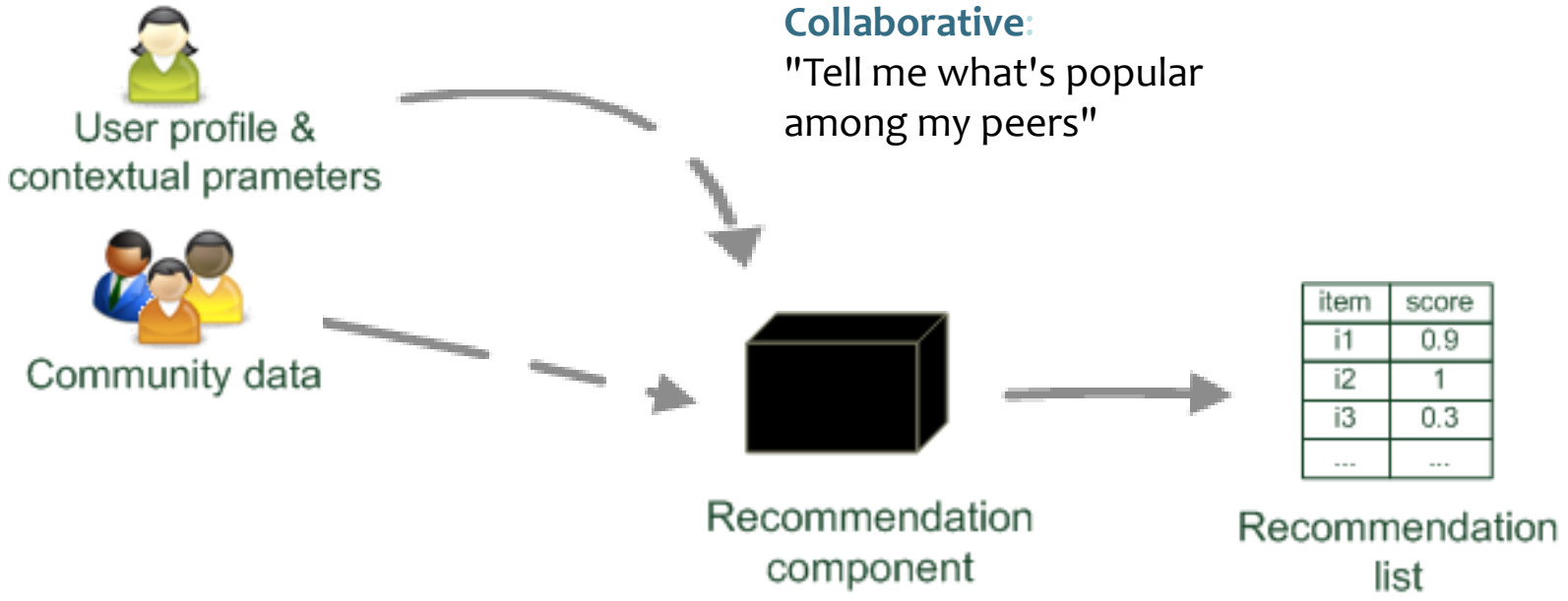


Outline

- Content-based Filtering
- Collaborative Filtering
- Hybrid Systems
- Knowledge-based Systems

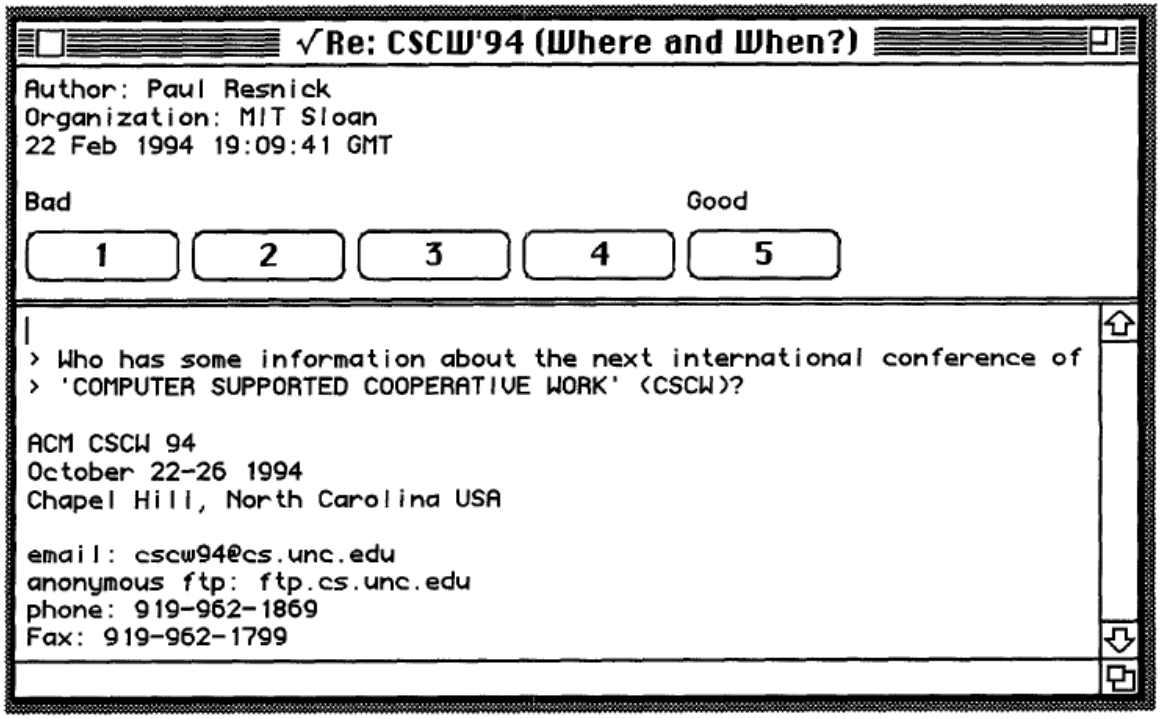
- Interactive Recommendation

Collaborative Filtering



Collaborative Filtering

- The predominant approach since 1994
- Recent advances in later lecture in summer school
- The GroupLens system
  - User-item ratings as the only input



Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. *GroupLens: an open architecture for collaborative filtering of netnews.* In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work* (CSCW '94). 175-186.

Matrix Completion - Limitations

- Amazon's contextual recommendations are a guiding scenario in the literature
  - But there are no ratings
  - There apparently is not even personalization

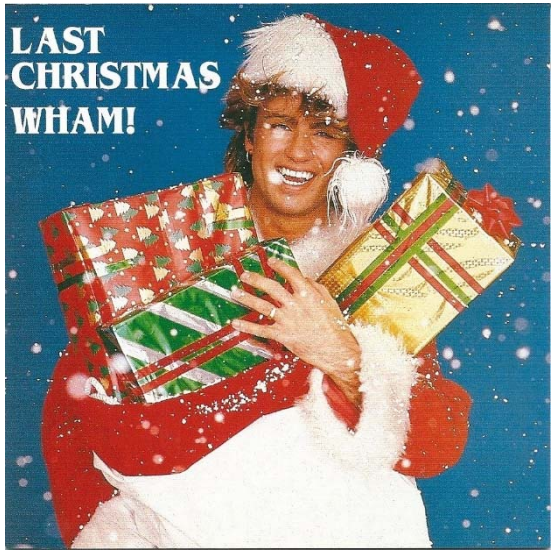
Sequence-aware Recommenders

- Timely research topic
  - Consider interaction logs as input (in contrast to rating matrix)
- Session-based recommendation
  - Recommend to anonymous users, given only a few interactions
- Session-aware recommendation
  - Recommend to known users in the context of an ongoing session

Quadrana, M., Cremonesi, P. and Jannach, D.: "*Sequence-Aware Recommender Systems*". ACM Computing Surveys, Vol. 51(4). ACM, 2018

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?

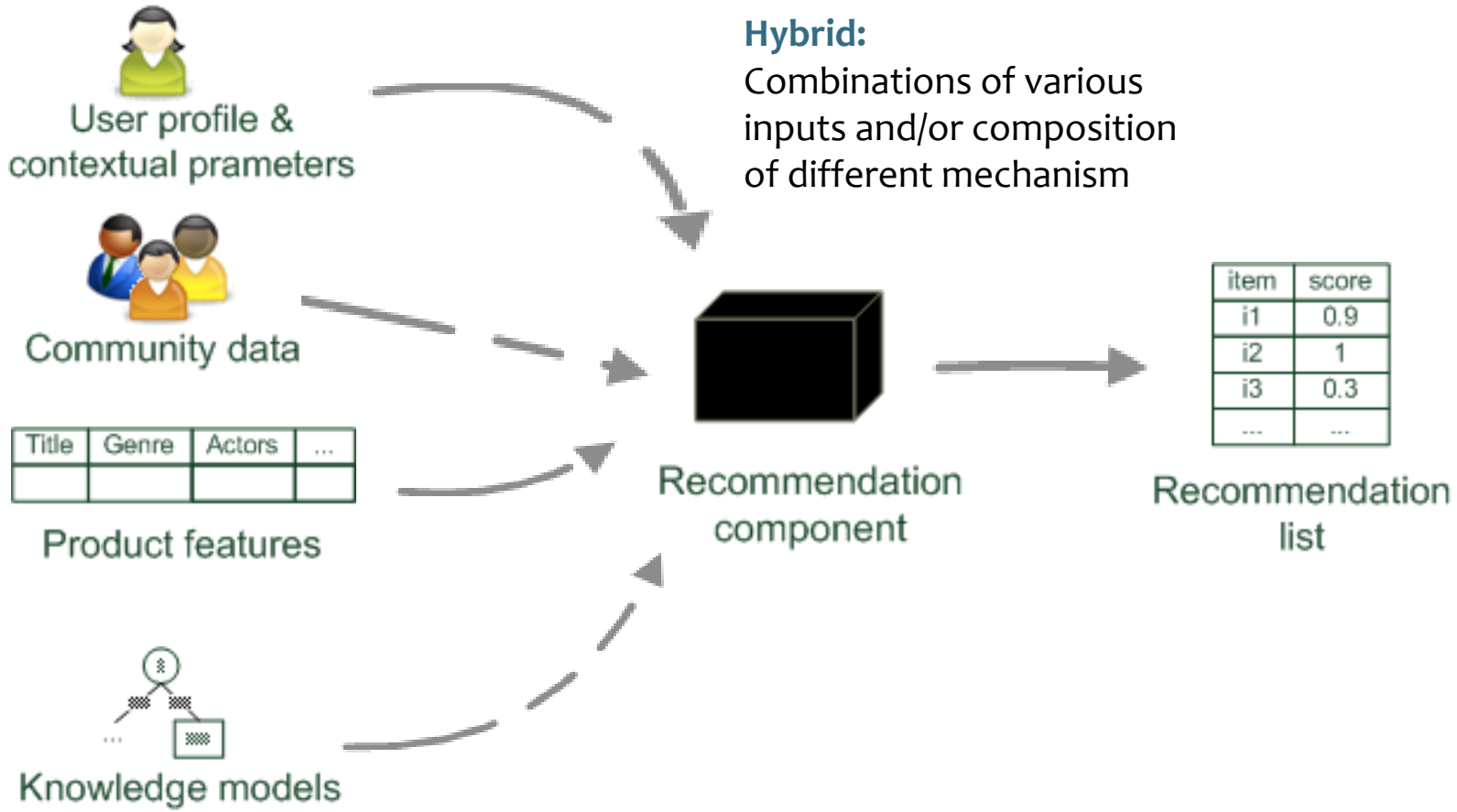


Outline

- Content-based Filtering
- Collaborative Filtering
- Hybrid Systems
- Knowledge-based Systems

- Interactive Recommendation

Hybrid Recommendation Approach

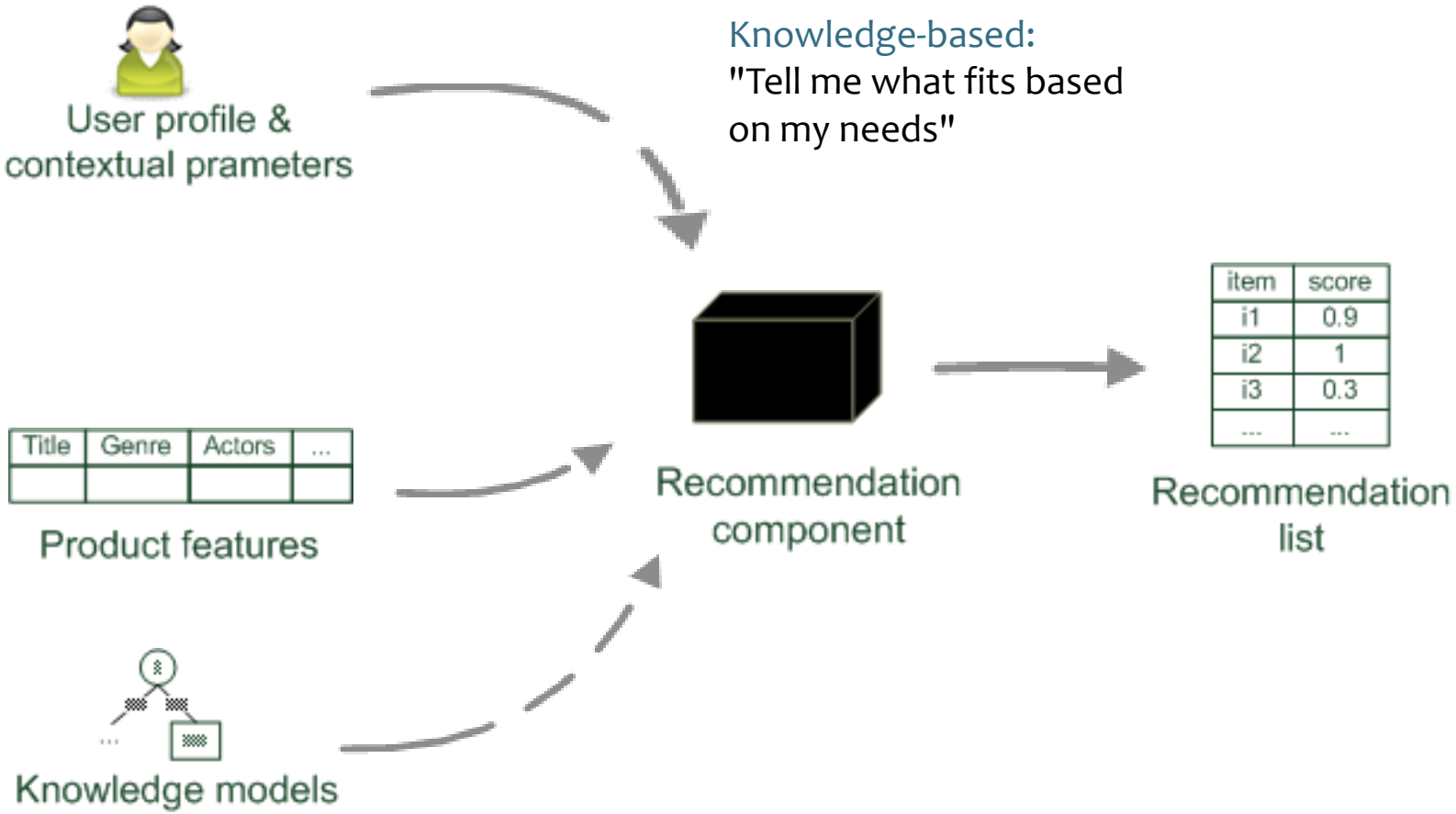


Outline

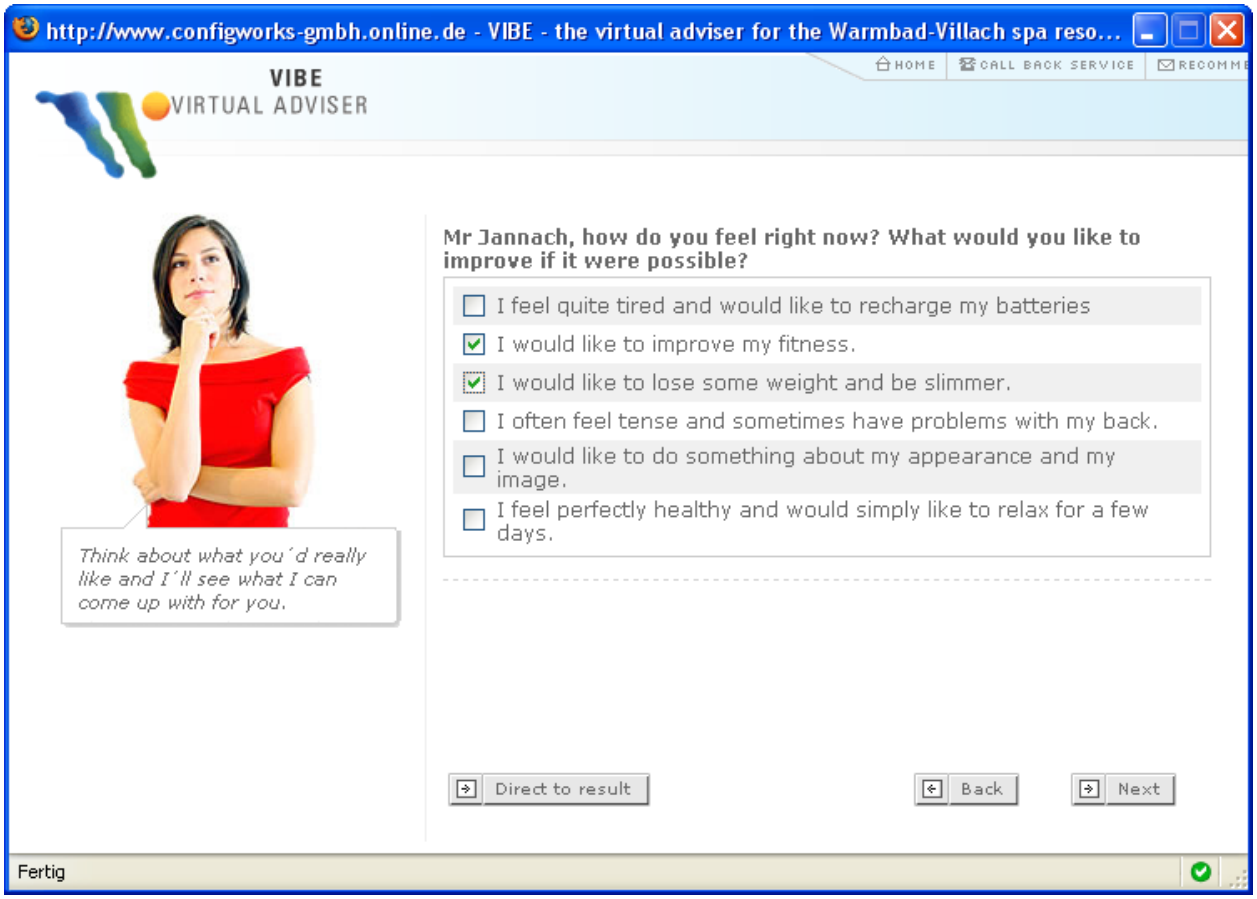
- Content-based Filtering
- Collaborative Filtering
- Hybrid Systems
- Knowledge-based Systems

- Interactive Recommendation

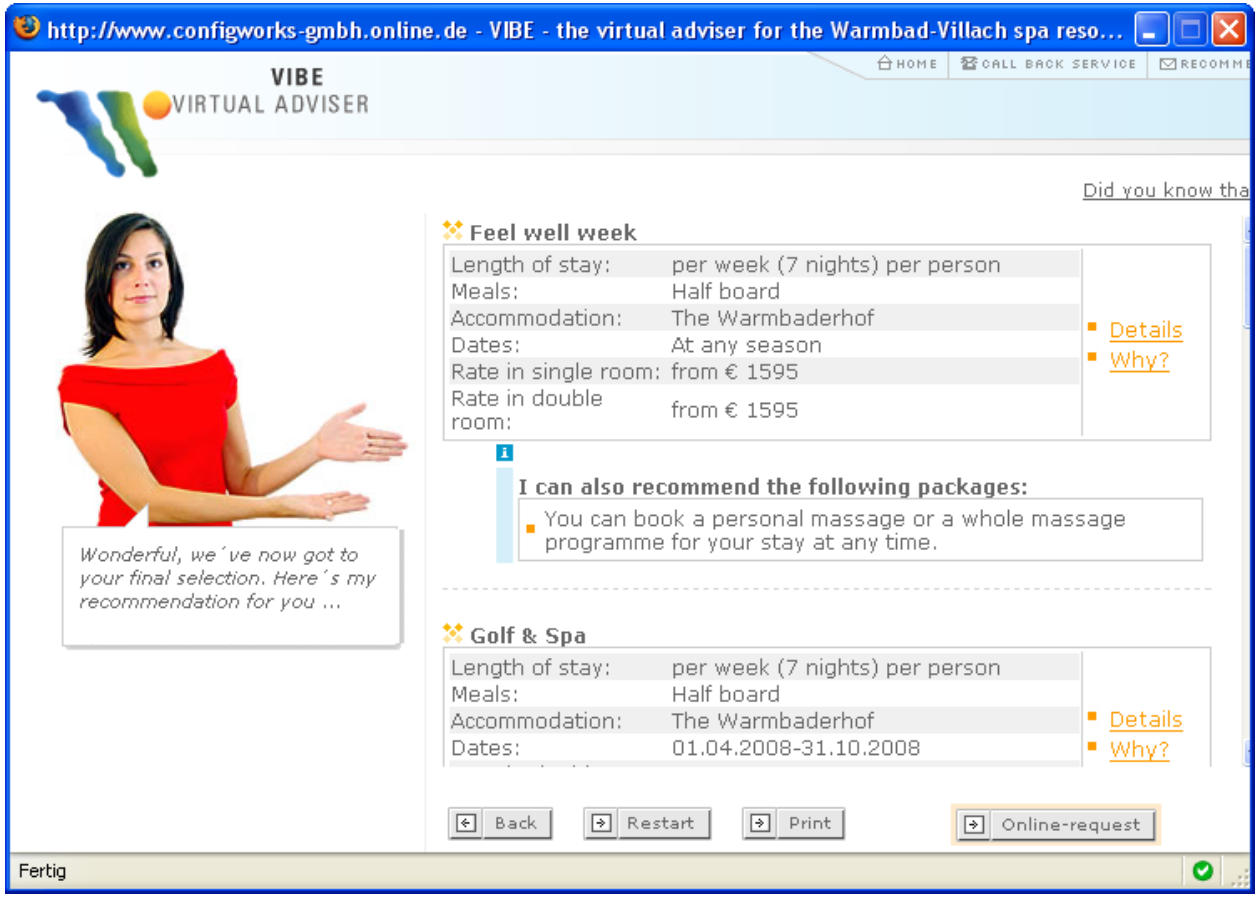
Knowledge-based Systems



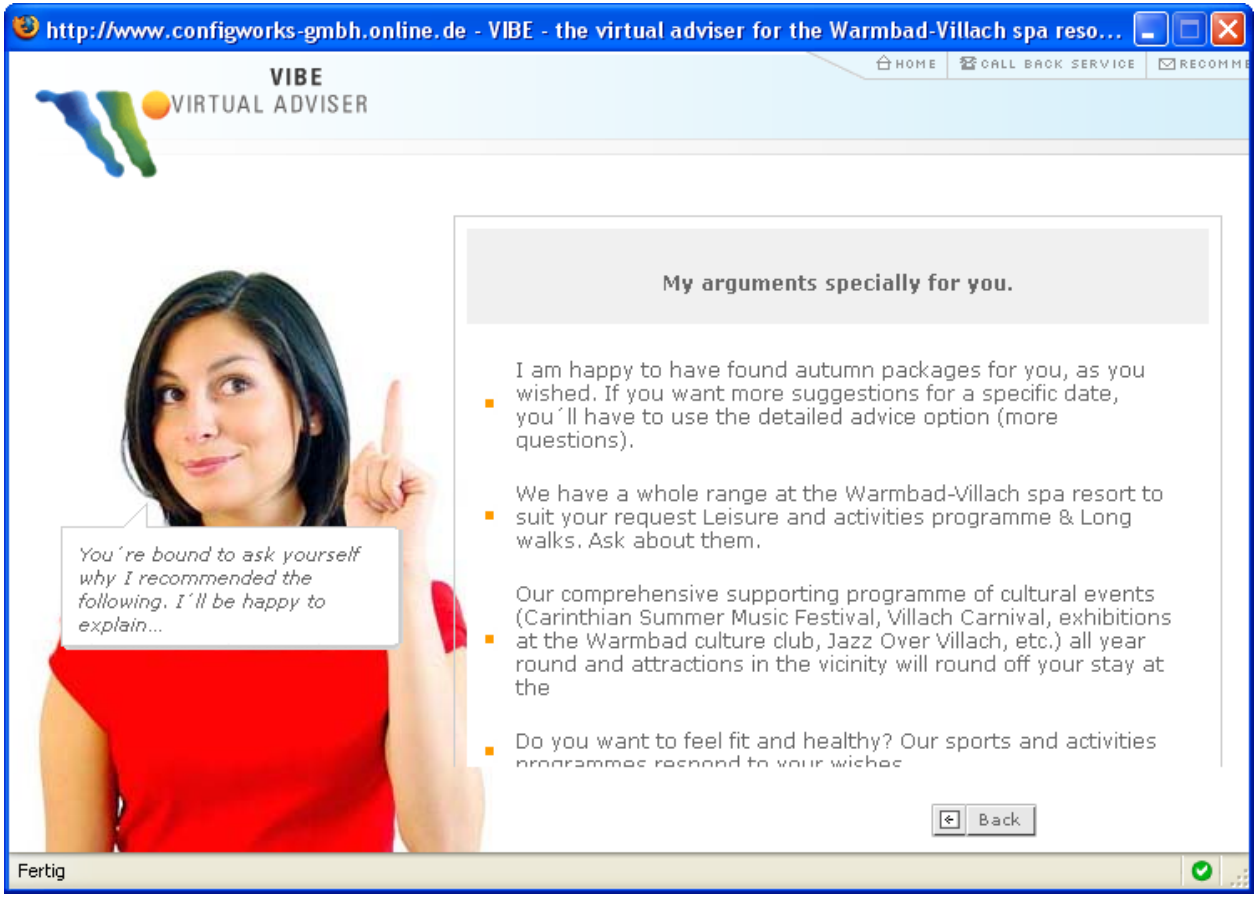
Is this even a recommender?



Is this even a recommender?



Is this even a recommender?



Outline

- Content-based Filtering
- Collaborative Filtering
- Hybrid Systems
- Knowledge-based Systems

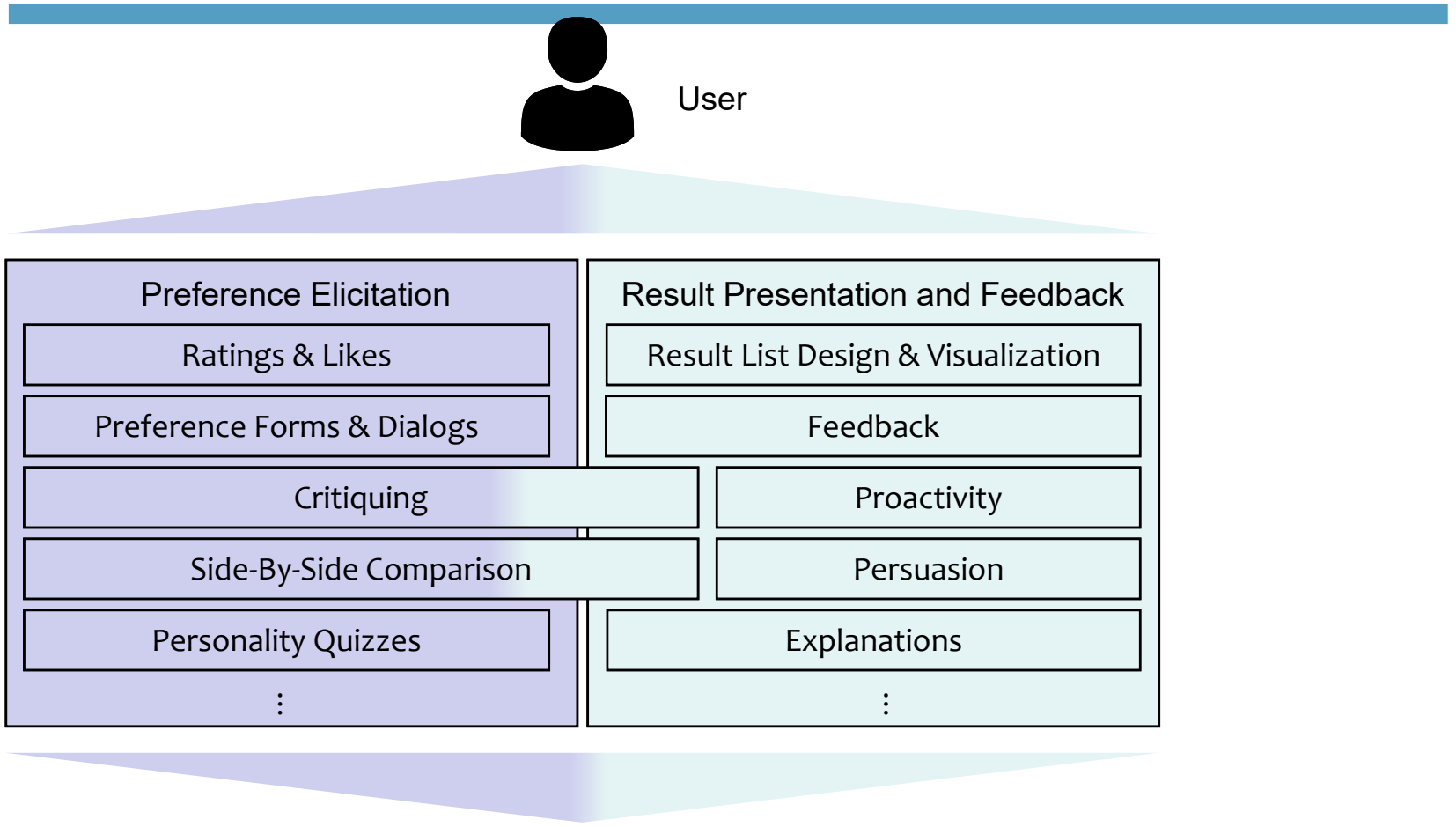
- Interactive Recommendation

From Algorithms to User Experience

- Most academic research focuses on algorithmic aspects
  - e.g., learning to predict / "post-dict" hidden ratings
- But a recommender *system* is more than the algorithm, see later lectures
- The UI can have a huge impact on adoption
  - Garcin et al., for example, report a more than 100% increase in the CTR when changing the position of the recommendations

Konstan, J.A. & Riedl, J.. "Recommender systems: from algorithms to user experience"
User Model User-Adap Inter (2012) 22: 101.
Garcin, F., Faltings, B., Donatsch, O., Alazzawi, A., Bruttin, C., and Huber, A. 2014. Offline and online evaluation of news recommender systems at swissinfo.ch. In *Proceedings of the 8th ACM Conference on Recommender systems* (RecSys '14).

Structuring Existing Works



Jugovac, M. and Jannach, D.: *"Interacting with Recommenders - Overview and Research Directions"*. ACM Transactions on Intelligent Interactive Systems (ACM TiiS), Vol. 7(3). 2017

Summary of methods

- We found algorithmic works based on collaborative filtering to be dominant
  - Recently, sequence-aware recommenders were more in the focus
- In contrast, many questions regarding the design of a recommender system remain open
- The design space for the user interface, for example, is huge, but the literature is comparably scarce

UNIVERSITÄT KLAGENFURT

Measurements

Evaluation approaches

- Testing a real application with real users
  - A/B tests (measuring, e.g., sales increase, CTR)
- Laboratory studies
  - Controlled experiments (measuring, e.g., satisfaction with the system), see later lecture
- Offline experiments
  - Simulations using on historical data (measuring, e.g., prediction accuracy, coverage)
- Theoretical analyses
  - For example, regarding scalability

Offline experiments

- Such experiments are, by far, the most common form of empirical research in the CS literature
- Main ingredients:
  - One or two historical dataset containing ratings or implicit feedback
  - A number of existing algorithms to compare the new proposal with
  - A number of established accuracy metrics (RMSE, Precision, Recall) and evaluation procedures to determine the metrics (e.g., cross-validation)

Sounds safe?

- All seems okay, “proving” progress in a reproducible way seems straightforward
  - At least one dataset should be public nowadays, so that others can replicate the results
  - The evaluation protocol and the metrics are well accepted and broadly known
  - The algorithmic proposals are usually laid out in great depth in the papers. Sometimes, even the source code is shared

Progress can still be limited

- Reason 1: “Proving” progress by finding a better model for a very specific experimental setup can be relatively easy
- Reason 2: The used metrics are not necessarily helpful to measure improvements as perceived by users in the first place

Potential issues w/ research practice

- Applied ML research often obsessed with accuracy and the hunt for the "best model"
  - "leaderboard chasing"
- But, there probably is no best model. The ranking of algorithms can depend on:
  - Given dataset
  - Used pre-processing steps
  - Evaluation measure
  - Choice of baselines
  - Optimization of baselines

Worrying observations

- Sometimes, it remains unclear if we truly make progress
  - Armstrong et al. (2009) find that there was not much progress within the previous ten years for a given Information Retrieval Task
  - Lin (2019) and Yang et al. (2019) found that ten years later problems with the choice of baselines still exist for deep learning methods
  - Rendle et al. (2019) run new experiments for classical recommendation tasks and find that recent methods are not necessarily better than previous ones

Worrying observations

- Makridakis (2018) compared various ML methods for time-series prediction, concluding that existing statistics-based methods are often better
- Ludewig et al. (2018-2019) evaluated various session-based recommendation techniques, finding that simple methods are often very competitive
- Ferrari Dacrema et al. (2019) examined recent neural top-n recommendation techniques and found potential issues in terms of the choice and optimization of baselines

Potential ways forward

- Further increasing reproducibility is advocated
  - Reproducibility should be easy to establish
    - Many researchers use free software tools
    - Sharing images of the experimental environment is easy
    - Code should include everything from algorithm, over data-pre-processing and evaluation
- Choice and optimization of baselines as main problem
  - Often not clear what represents the state-of-the-art
  - Validation against optimized existing methods

Potential ways forward

- Toward more "theory-guided" research
  - Choice of dataset/pre-processing often seems arbitrary
  - Choice of evaluation procedures often seems arbitrary and not guided by an application problem
    - Various forms of measures used, cut-off lengths between one and several hundred, cross-validation/leave-one-out …

Offline experiments and computational metrics in general

- Reason 2 from above: The used metrics are not necessarily helpful to measure improvements as perceived by users in the first place
- Generally:
  - Being able to accurately predict the relevance of items for users is and will be a central problem of recommender systems research
  - Increasing the prediction accuracy therefore can be a relevant goal of research

The problems with accuracy

- Accuracy alone is not enough
  - Recommending items that the user might have bought anyway might be of little business value
  - Focusing on accuracy alone can lead to monotone recommendations (e.g., only movies from the Star Wars series) and limited discovery
  - Optimizing for accuracy might lead to recommendations that are considered too "obscure" for users
    - Familiarity with some recommendations might be important to increase the user's trust in a system

Multi-metric evaluations

- One possible way forward
- Offline experimentation can assess multiple, possibly competing, goals in parallel (see later lecture in the summer school)
  - Accuracy
  - Diversity
  - Novelty
  - Serendipity
  - Long-term effects, e.g., on reinforcement effects
  - Business value for multiple stakeholders
  - Scalability …

The problems of offline experiments

- Are offline experiments actually predictive of the perceived value?
  - Gomez-Uribe and Hunt (2015), Netflix, found that offline experiments were **not** found "*to be as highly predictive of A/B test outcomes as we would like.*"
  - In fact, a number of user studies did **not** find that algorithms with higher prediction accuracy led to better quality perceptions by study participants

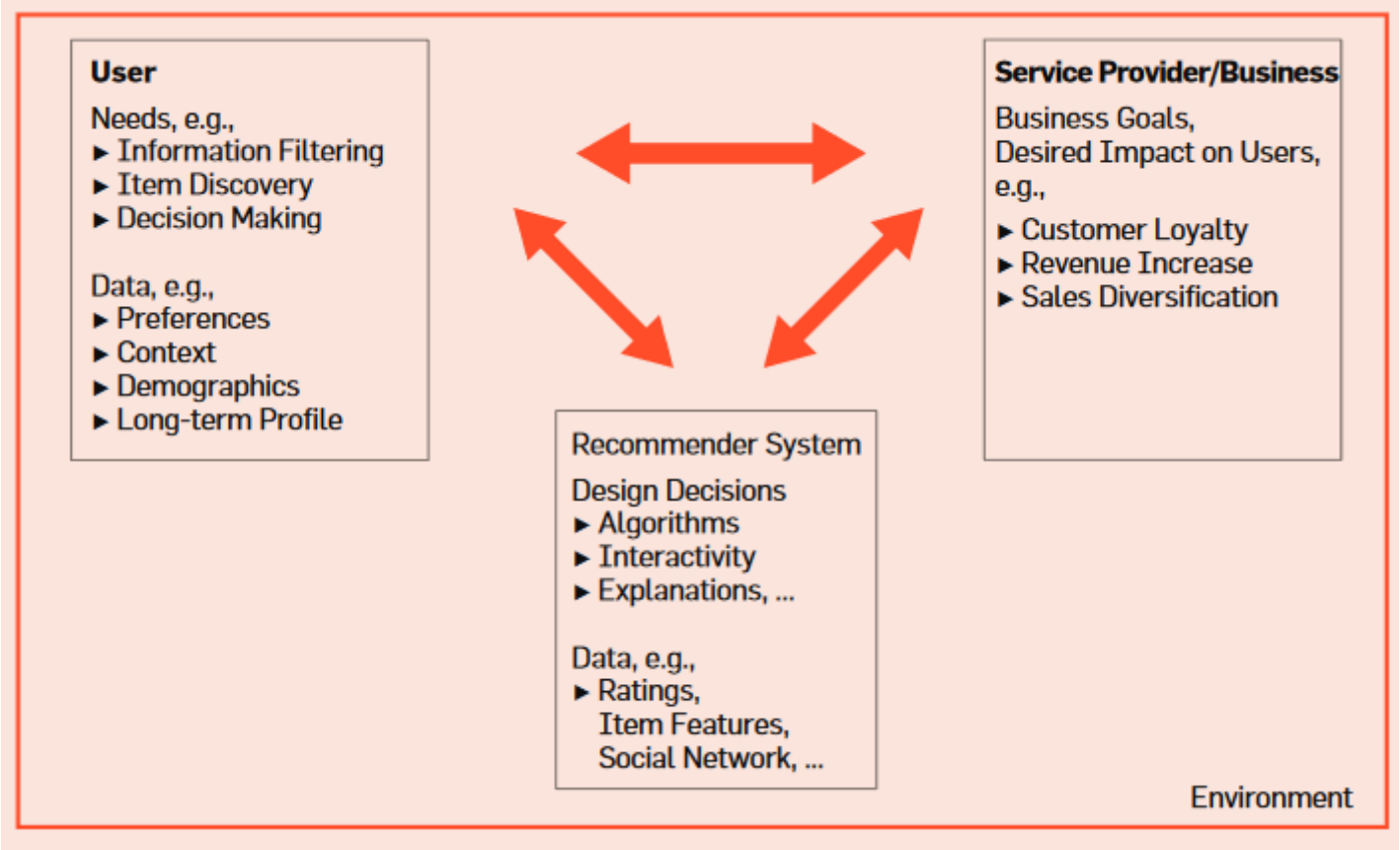
Possible steps forward

- Toward a more comprehensive approach to recommender systems research
  - Considering the user in the loop
  - Considering the business value for one or more stakeholders
  - Use a richer methodological repertoire
- See later lecture in this summer school

Jannach, D., Resnick, P., Tuzhilin, A. and Zanker, M.: "*Recommender Systems - Beyond Matrix Completion*". Communications of the ACM, Vol. 59(11). Association for Computing Machinery (ACM), 2016, pp. 94-102

Possible steps forward

- "From algorithms to systems"

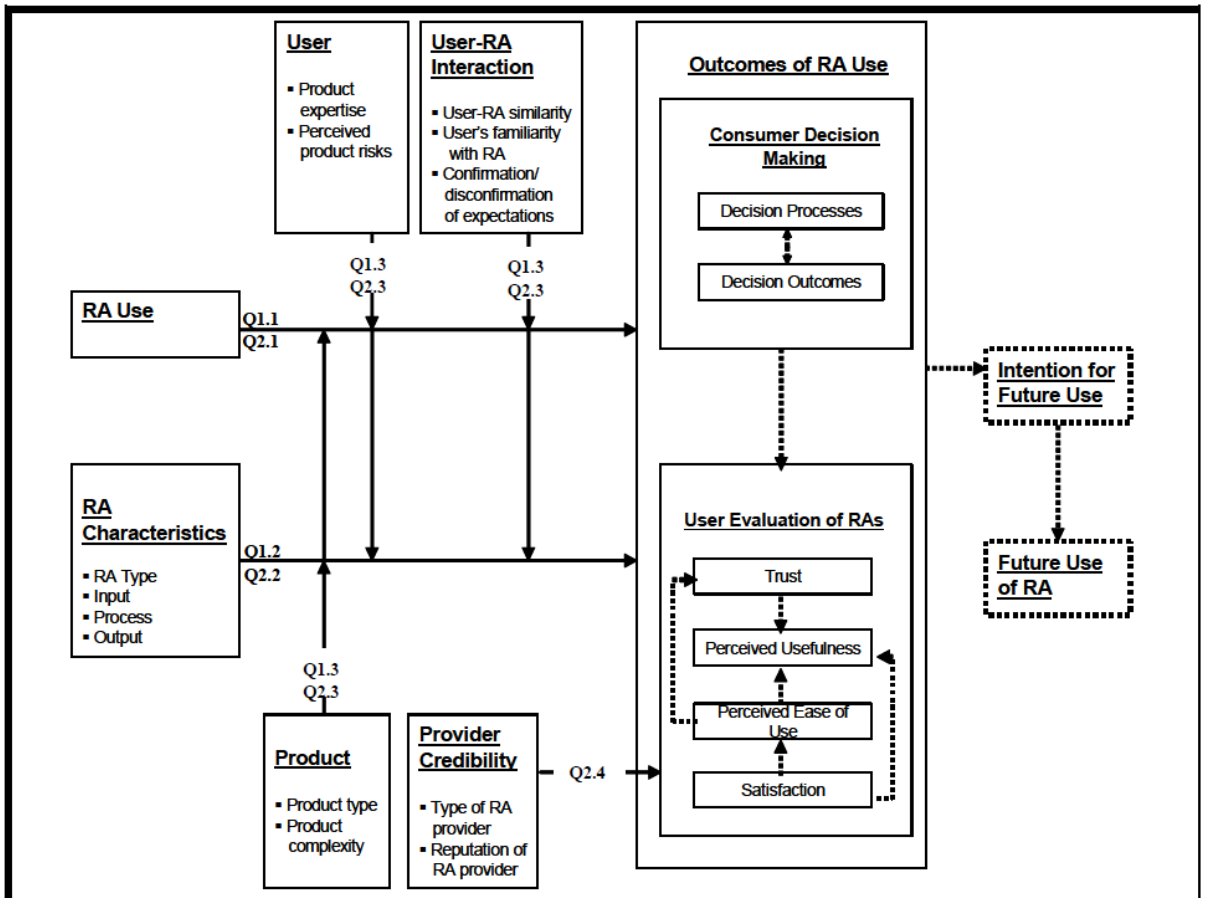


User-centric research

- Much richer conceptual models of recommender systems and their impact exist in the field of Information Systems
  - Algorithms are only one of many components
  - Apparently limited knowledge of these works in the computer science community

Xiao, B. and Benbasat, I.: „E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact“, MISQ, 2007

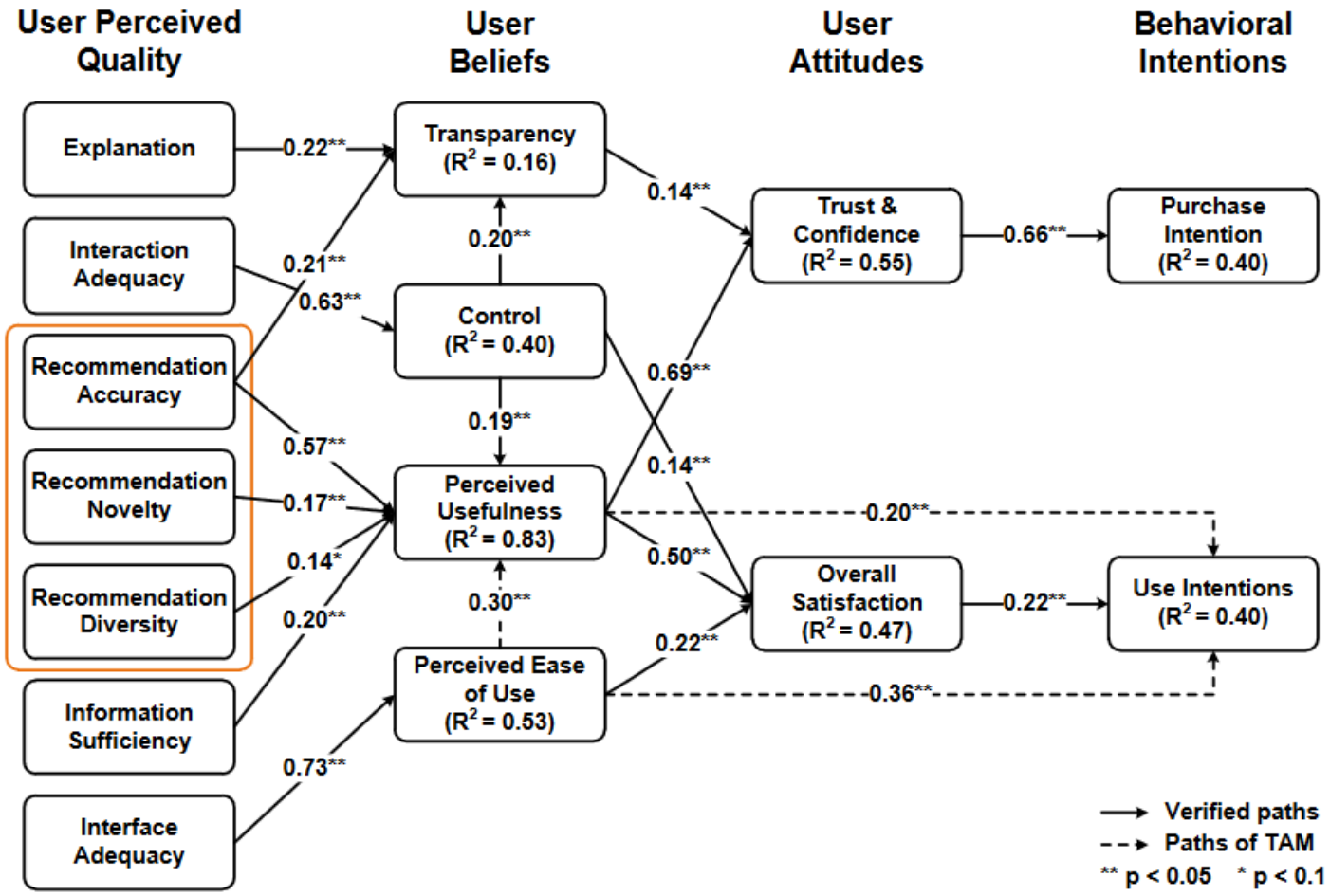
A conceptual model



Note: Solid lines indicate relationships and constructs investigated in this paper.

Figure 1. Conceptual Model

Example validation



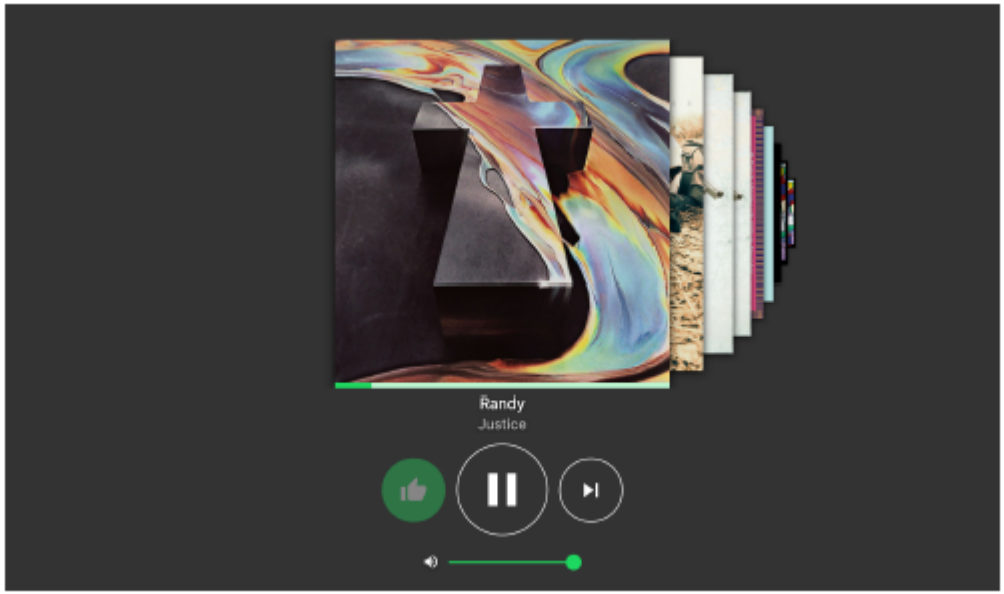
Takeaways

- Computer Science research is mostly focused on algorithms
- But the value of improvements in terms of abstract computational measures is limited or non-existent
  - E.g., due to the used research methodology
- There are many more interesting and relevant questions than algorithms

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

User studies: Examples

- Example 1: User perception of session-based music recommendations



Ludewig, M. and Jannach, D.: "*User-Centric Evaluation of Session-Based Recommendations for an Automated Radio Station*". In: Proceedings of the 2019 ACM Conference on Recommender Systems (RecSys 2019). Copenhagen, 2019

Background

- Various methods for session-based recommendation proposed in recent years
- Competing offline accuracy evaluation results:
  a. Method based on RNNs better than certain baselines using item-based nearest neighbors (Hidasi et al., 2015 and later)
  b. Simple heuristic and session-based nearest neighbors often better than RNNs (Ludewig et al. 2017 and later)

Motivation and setup

- Assess how users perceive the recommendation quality in different dimensions
- Experimental setup:
  - Develop an online application for study participants to interact with
  - Participants select a start track and the application creates and plays a playlist
  - Participants can skip or like tracks, leading to updates of the playlist
  - Participants fill out a questionnaire at the end

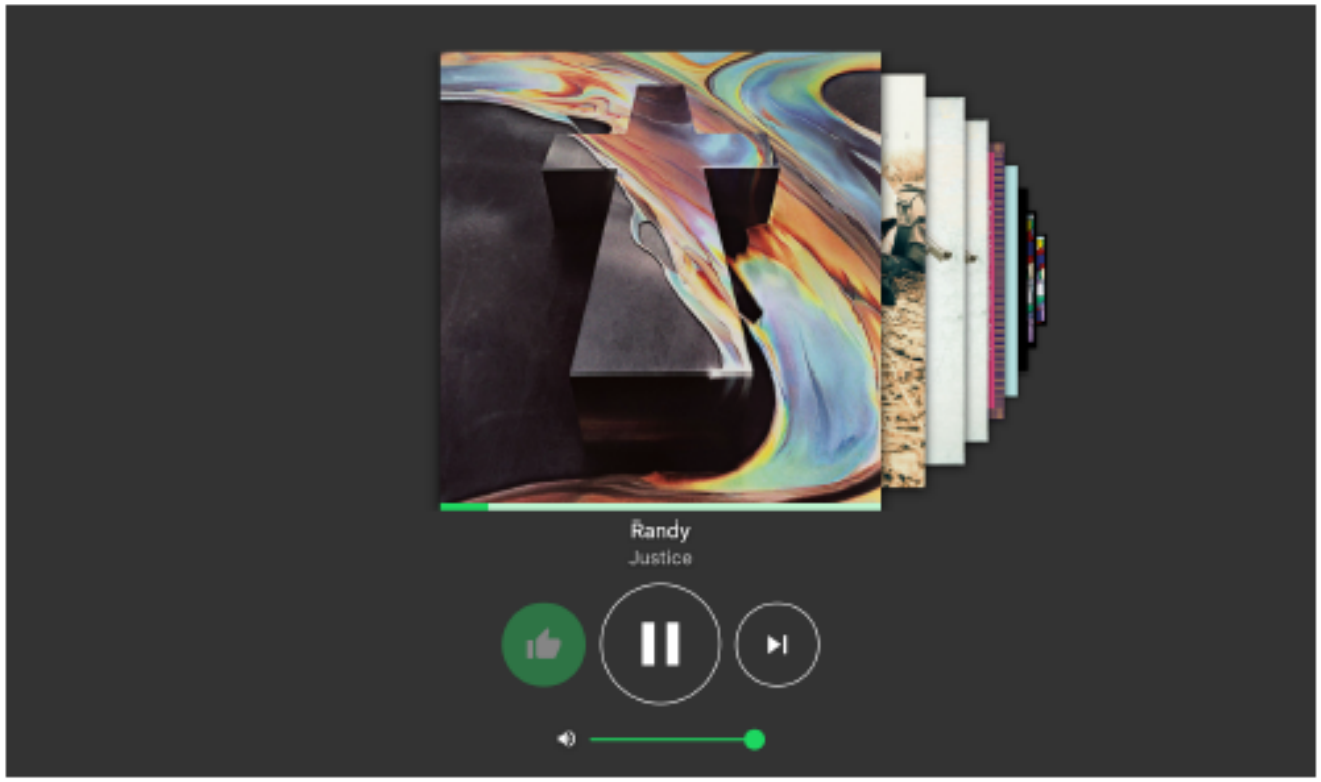
Experimental details

- Different recommendation algorithms tested
  - Simple association rules AR (“customers who bought”)
  - Collocated Artists Greatest Hits (CAGH)
  - GRU4REC: An RNN-based method
  - S-KNN: A session-based nearest neighbor method
  - SPOTIFY: Recommendations were retrieved only through Spotify’s API

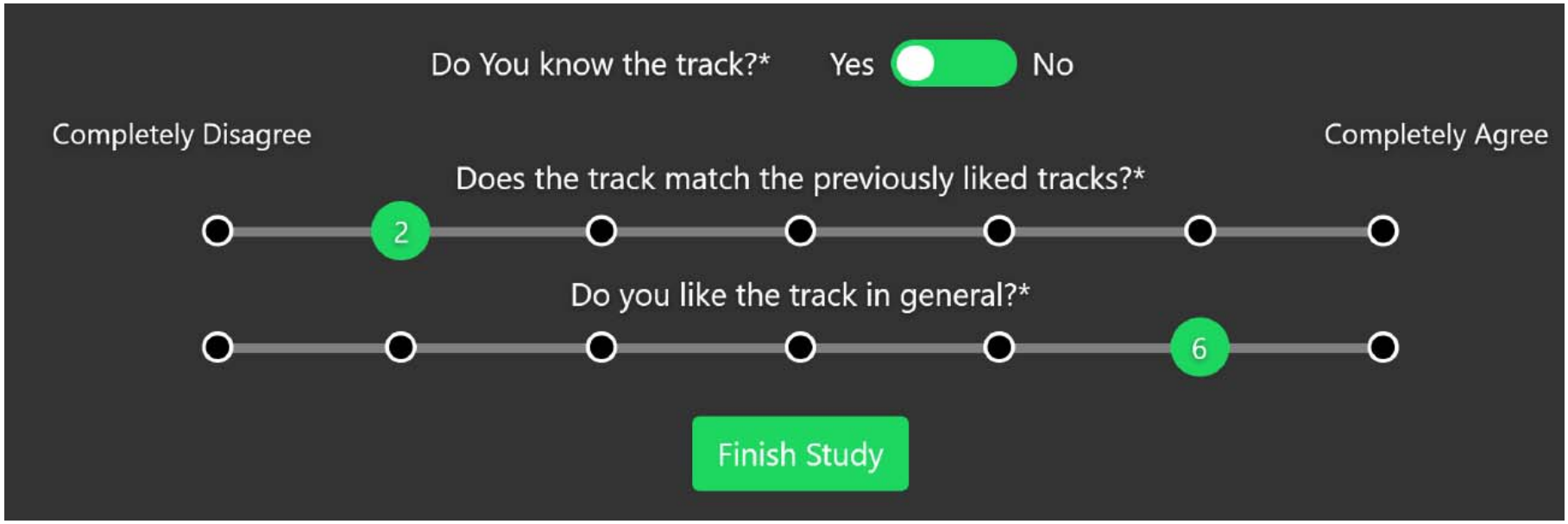
Experiment details

- All user actions are recorded
- Feedback for each track collected
- Post-task questionnaire covers, e.g., aspects of
  - suitability of the tracks with respect to the start track
  - the adaptation of the playlist to the preferences
  - the diversity of the recommendations
  - the novelty of the recommendations
  - the intention to reuse the system
- Feedback was collected using 7-point Likert scale items

User interface



User interface



Questionnaire

Question	
Q1	I liked the automatically generated radio station.
Q2	The radio suited my general taste in music.
Q3	The tracks on the radio musically matched the track I selected in the beginning.
Q4	The radio was tailored to my preferences the more positive feedback I gave.
Q5	The radio was diversified in a good way.
Q6	The tracks on the radio surprised me.
Q7	I discovered some unknown tracks that I liked in the process.
Q8	I am participating in this study with care so I change this slider to two.
Q9	I would listen to the same radio station based on that track again.
Q10	I would use this system again, e.g., with a different first song.
Q11	I would recommend this radio station to a friend.
Q12	I would recommend this system to a friend.

Running the experiment

- Used Amazon Mechanical Turk crowdworkers
  - 50 for reach treatment group in the end
  - Removed quite a number of non-attentive participants to ensure high quality
  - Applied additional quality criteria in advance
- Task details
  - Participants had to listen to at least 15 tracks (30 secs excerpts)
  - Average pure listening time of 5.5 minutes

Result analysis

- Number of Likes:
  - From 4.48 (Spotify) to 6.48 (AR)
- Popularity of recommendations:
  - Spotify and GRU4REC with the least popular / novel recommendations
  - Popularity highly correlates with number of Likes
- Match of next track with previous ones
  - S-KNN and CAGH work best, AR has the weakest scores

Result analysis

- Ratings for tracks
  - Even though AR received the most likes, they received, on average, the lowest rating scores
  - Reason: Many 1-star ratings for apparently bad recommendations
  - Some insights:
    - Optimizing for likes can be misleading
    - One should consider the role of (too) bad recommendations

Result analysis

- Selected questionnaire results:
  - S-KNN recommendations were generally more liked than those of AR, GRU4REC, and Spotify
  - S-KNN recommendations were often considered a good match for the selected seed tracks
  - AR works poor in many dimensions
  - No differences in terms of diversification and surprise were found
  - Spotify excelled in terms of discovery
  - In terms of intention to reuse, S-KNN, CAGH, and Spotify scored highest

Result analysis

- Additional indications:
  - High ratings and/or many likes are not the only factors contributing to system reuse
  - Discovery appears to be a central factor
  - Participants stated that they will re-use the Spotify-based system despite the higher novelty and the lower prediction accuracy
    - Running offline experiments revealed that Spotify scored very, very low on typical measures like Precision and Recall

Offline Results

Algorithm	P@5	R@5	HR@5	MRR@5
S-KNN	**0.271**	**0.044**	0.137	0.077
GRU4REC	0.161	0.028	**0.151**	**0.096**
AR	0.234	0.037	0.135	0.081
CAGH	0.172	0.024	0.052	0.026
SPOTIFY	0.009	0.001	0.002	0.001

Limitations

- Key challenges of user studies lie, e.g., in
  - controlling the experimental conditions
  - making sure that the findings are generalizable to at least a certain subset of the user population
- In our case, e.g.,:
  - Participants did not use a real-world system and they were not listening in a “natural” environment
  - The motivation of participants might be varying
  - The representativeness of the participant sample from Mechanical Turk might not be entirely clear

Summary of main findings

- Spotify
  - These recommendations would have led to terrible performance values in offline experiments
  - Still, they were well-received by the users
  - Spotify's recommendations help the purpose of discovery, which seems central for such an application
- S-KNN
  - was not only good in the offline setting, but led to good results also in terms of the quality perception
- AR
  - Good in terms of likes, but many poor recommendations

Literature

- **“The Neural Hype and Comparisons Against Weak Baselines” by Lin**
  - SIGIR Forum52, 2 (Jan. 2019), 40–51u
- **“Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models” by Yang et al.**
  - SIGIR 2019
- **“On the Difficulty of Evaluating Baselines: A Study on Recommender Systems” by Rendle et al.**
  - arxiv.org (https://arxiv.org/abs/1905.01395), 2019
- **“Statistical and Machine Learning forecasting methods: Concerns and ways forward” by Makridakis et al.**
  - PLOS ONE, 2018

Literature

- **“Evaluation of Session-based Recommendation Algorithms”, “Performance Comparison of Neural and Non-Neural Approaches to Session-based Recommendation” by Ludewig et al.**
  - UMUAI 2018, RecSys 2019
- **“Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches” by Dacrema et al.**
  - RecSys 2019