UNIVERSITÄT KLAGENFURT

Recommender Systems: Business Value and Measurements

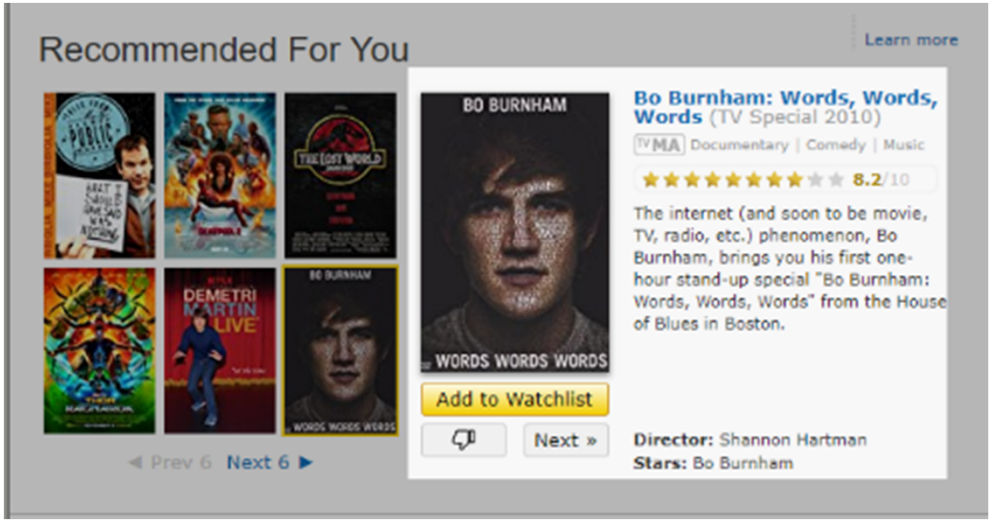
Dietmar Jannach, University of Klagenfurt, Austria
dietmar.jannach@aau.at

Presented at the The 18th Dutch-Belgian Information Retrieval Workshop, Amsterdam, 2019

www.aau.at

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
- …

Roots, Goals, Characteristics

- Roots in various fields
  - e.g., Information Retrieval, Machine Learning, Human Computer Interaction
- Their design can furthermore be influenced by insights from more distant fields
  - e.g., Consumer behavior, Psychology, Marketing
- Typical goals:
  - Avoid information overload (filtering)
  - Active promotion of content
- Personalization often as a central concept

UNIVERSITÄT KLAGENFURT

Purpose and Value

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

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

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

Jannach, D. and Adomavicius, G.: "Price and Profit Awareness in Recommender Systems". In: Proceedings of the ACM RecSys 2017 Workshop on Value-Aware and Multi-Stakeholder Recommendation. Como, Italy, 2017

Multi-stakeholder considerations

- An even more complex example
  - Consider a hotel booking site, where hotels pay commissions when they are booked through the site
- Potential goals for the stakeholders
  - Consumer
    - Find a hotel that matches the needs and represents the best value for money (2 goals already)
  - Booking site
    - Help users find a matching deal, also maximize commission
  - Hotel
    - Maximize revenue and/or maximize occupancy rate

Abdollahpouri, H., et al..: *"Multistakeholder Recommendation: Survey and Research Directions"*. User Modeling and User-Adapted Interaction, forthcoming. 2019

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"

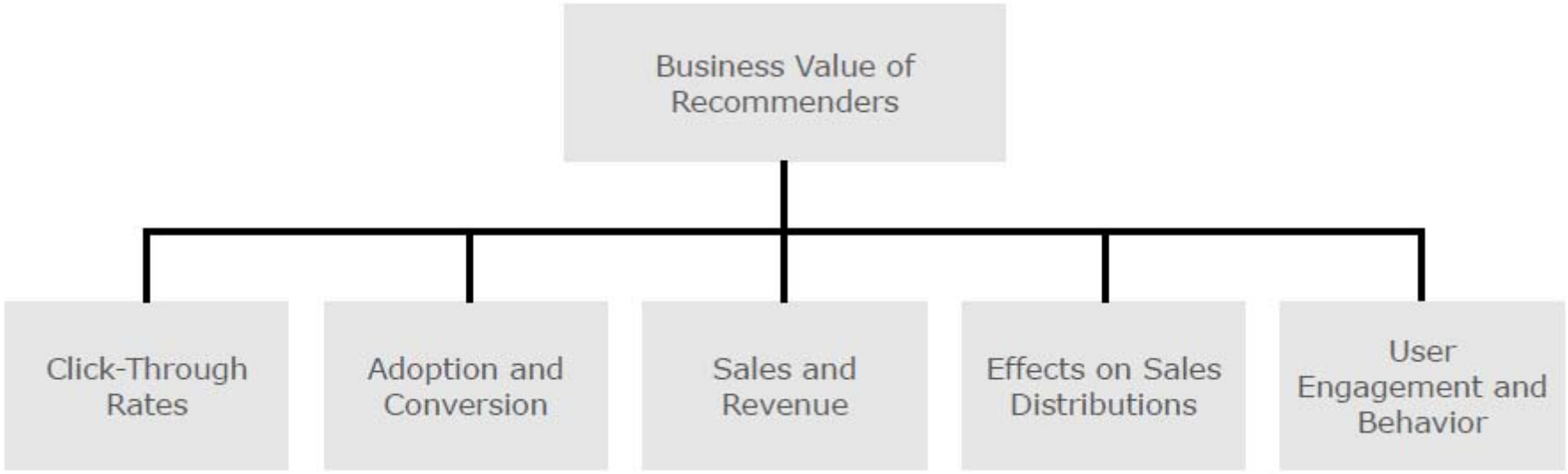
Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix Recommender System: Algorithms, Business Value, and Innovation. ACM Trans. Manage. Inf. Syst. 6, 4, Article 13

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

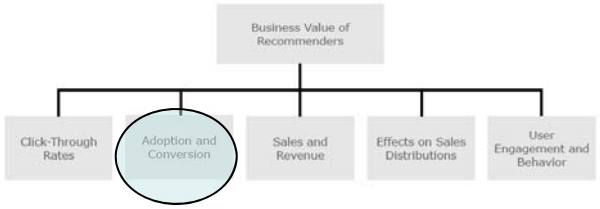


Jannach, D., Jugovac, M.; "Measuring the Business Value of Recommender Systems", arxiv preprint, https://arxiv.org/pdf/1908.08328.pdf, ACM Transactions on Management Information Systems (forthcoming)

Click-Through Rates

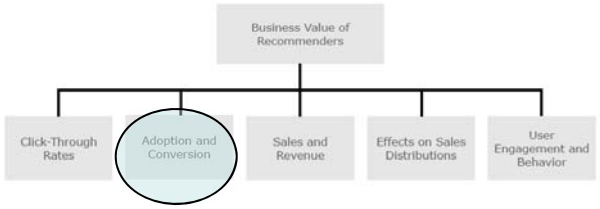
- Measures how many clicks are garnered by recommendations
  - Popular in the news recommendation domain
    - Google News: 38% more clicks compared to popularity-based recommendations
    - Forbes: 37% improvement through better algorithm compared to time-decayed popularity based method
    - swissinfo.ch: 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



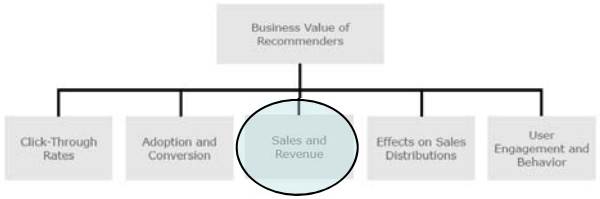
- CTR usually not the ultimate measure
  - Cannot know if users actually liked/purchased what they clicked on (consider also: click bait)
- Therefore
  - Various, domain-specific adoption measures common
- YouTube, Netflix: "Long CTR"/ "Take rate"
  - only count click if certain amount of video was watched

Adoption and Conversion Rates



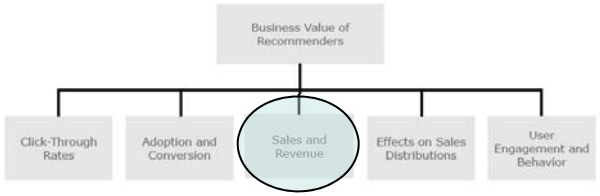
- Alternatives when items cannot be viewed/read:
- eBay:
  - "purchase-through-rate", "bid-through-rate"
- Other:
  - LinkedIn: Contact with employer made
  - Paper recommendation: "link-through", "cite-through"
  - E-Commerce marketplace: "click-outs"
  - Online dating: "open communications", "positive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

Sales and Revenue



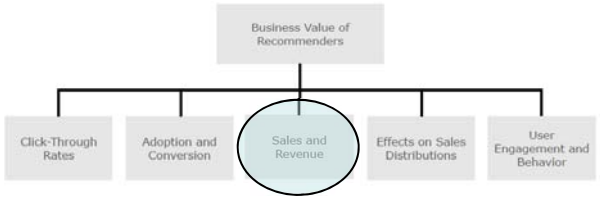
- CTR and adoption measures are good indicators of relevant recommendations
- However:
  - Often unclear how this translates into business value
  - Users might have bought an item anyway
  - Substantial increases might be not relevant for business when starting from a very low basis
- In addition:
  - 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
- Side observation
  - Choice of algorithm can matter a lot
  - But very different algorithms are compared in the discussed papers

Sales and Revenue

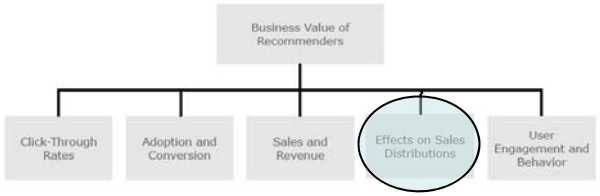


- e-grocery studies:
  - 1.8 % direct increase in sales in one study
  - 0.3 % direct effects in another study
  - However:
    - Up to 26% indirect effects, e.g., where customers were pointed to other categories in the store
    - "Inspirational" effect also observed in music recommendation in our own 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

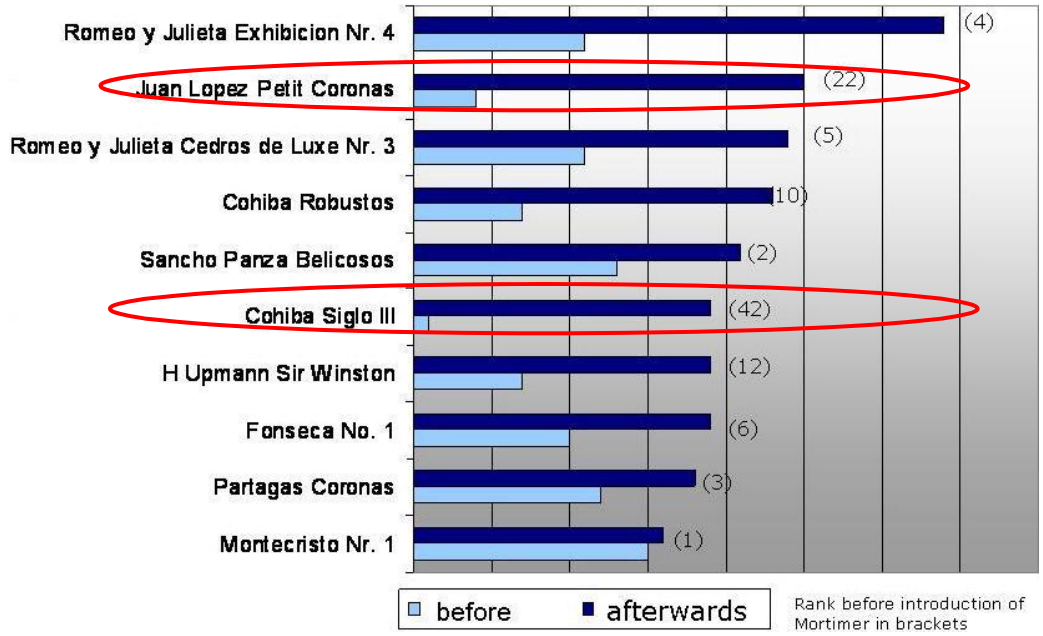
Effects on Sales Distributions



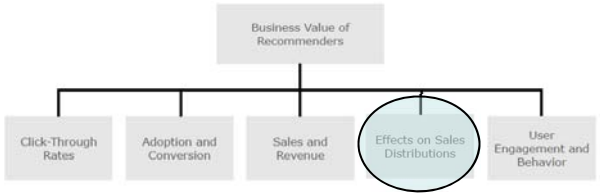
- Goal is maybe not to sell *more* but *different* items
- Influence sales behavior of customers
  - stimulate cross-sales
  - sell off on-stock items
  - promote items with higher margin
  - long-tail recommendations

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



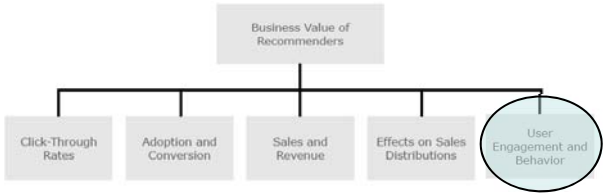
Effects on Sales Distributions



- Netflix:
  - Measure the "effective catalog size", i.e., how many items are actually (frequently) viewed
  - Recommenders lead users away from blockbusters
- 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

Jannach, D., Lerche, L., Kamehkhosh, I. and Jugovac, M.: "What recommenders recommend: an analysis of recommendation biases and possible countermeasures". User Modeling and User-Adapted Interaction, Vol. 25(5). Springer Nature, 2015, pp. 427-491.

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 most preferable
  - i.e., where the business value of can almost be directly measured, e.g., in sales volume increases
- 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 can be misleading
- 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
    - 100% CTR increase reported in Garcin et al. after changing UI

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).

Discussion

- More pitfalls of indirect measurements
- 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

UNIVERSITÄT KLAGENFURT

What **should** we measure?

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 – a challenge

- Note:
  - The same set of recommendations can be good or not, depending on the purpose, context, and application
- Examples
  - 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
- Abstract computational tasks from the literature
  - Find all or some good items
  - Predict the relevance of unseen items
  - Recommend sequence
  - Just browsing

Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. 22, 1

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

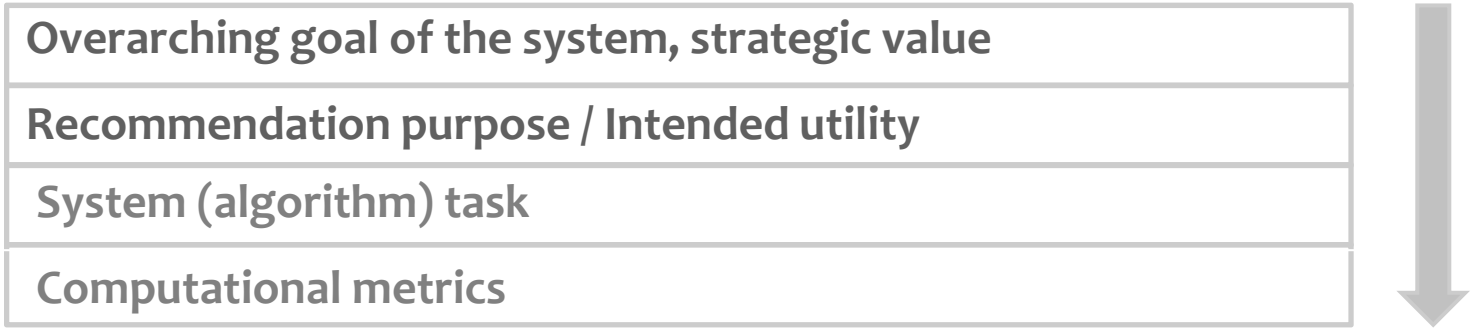
Benefits & Limitations

- 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

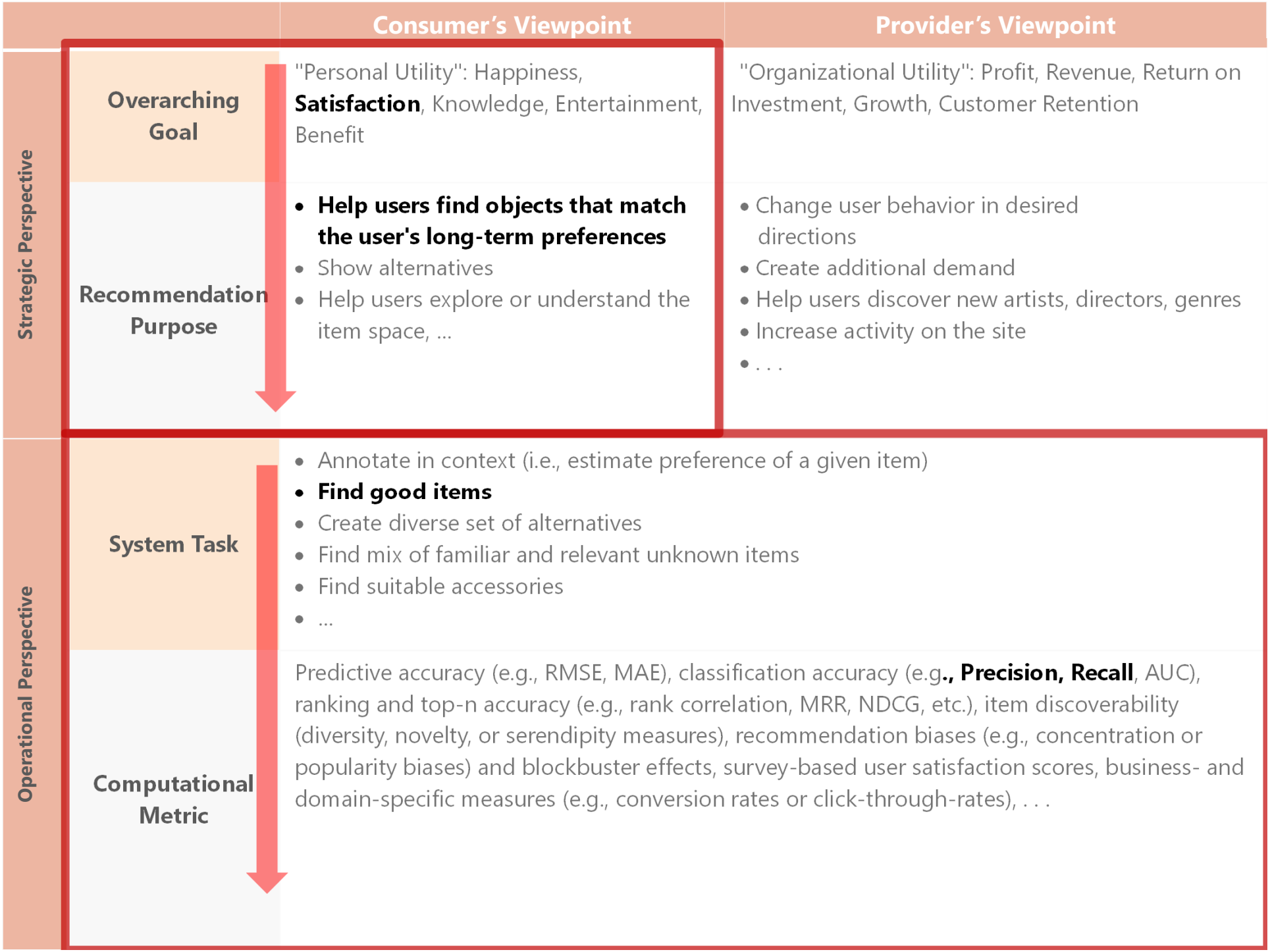


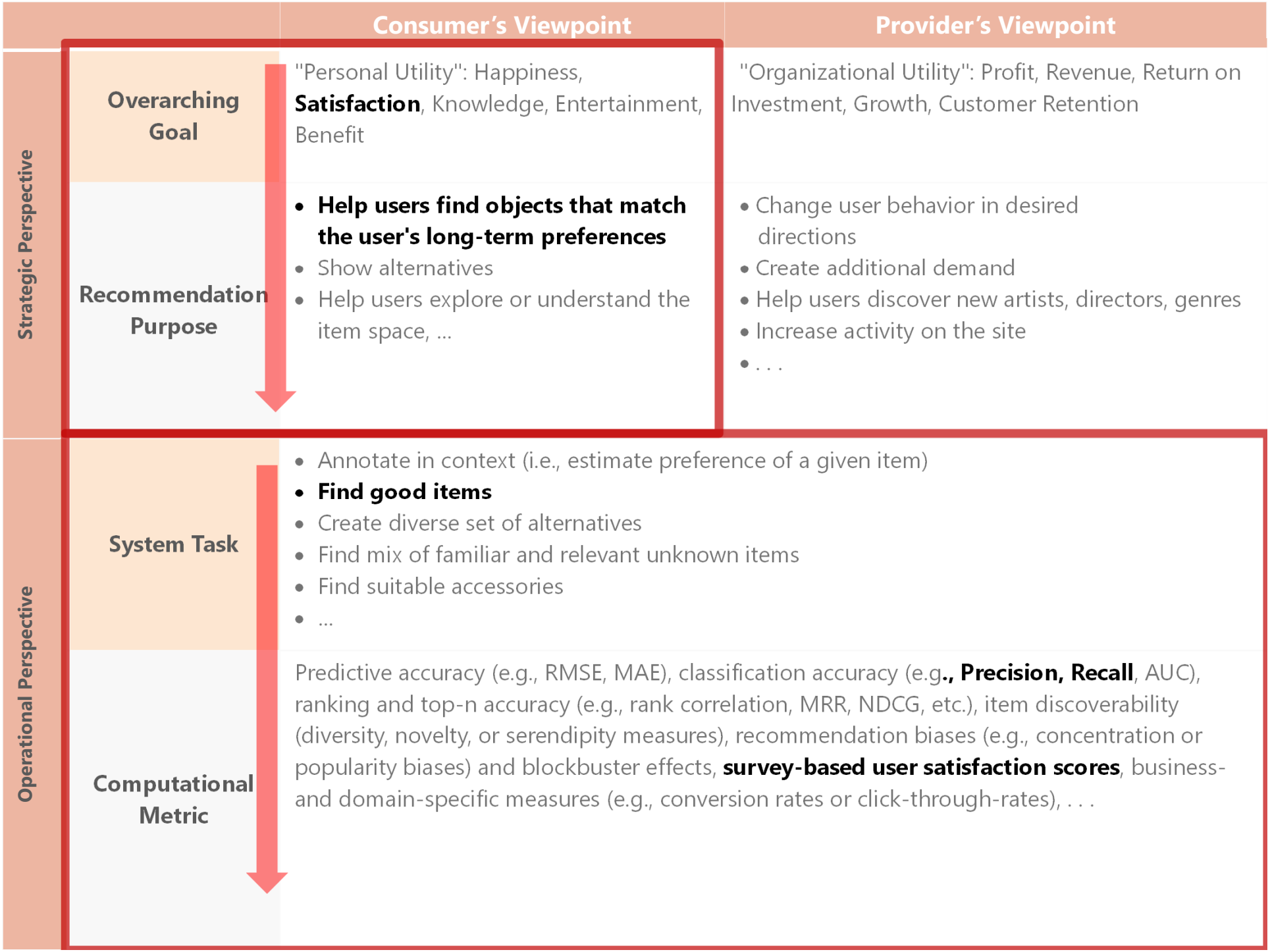
Jannach, D. and Adomavicius, G.: *"Recommendations with a Purpose"*. In: Proceedings of the 10th ACM Conference on Recommender Systems (RecSys 2016). Boston, Massachusetts, USA, 2016, pp. 7-10

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), . . .

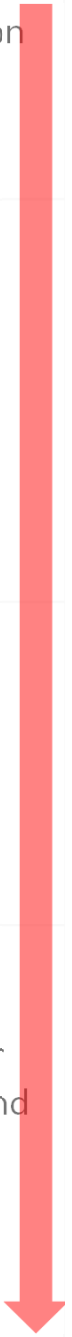
		Consumer's Viewpoint	Provider's Viewpoint
Strategic Perspective	**Overarching Goal**	"Personal Utility": Happiness, Satisfaction, Knowledge, Entertainment, Benefit	"Organizational Utility": Profit, Revenue, Return on Investment, Growth, Customer Retention
	Recommendation Purpose	• Help users find objects 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 so far

- 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

Limitations of today's research practice and ways forward

Evaluation aspects

- Computer Science research in this context is mostly about building "better" recommenders
  - i.e., systems or algorithms that serve a particular purpose better than alternative approaches
    - Often not about understanding what makes things better
- Typical purposes could be (see earlier slides)
  - Rank relevant items higher in the list
  - Make sure that the list is not monotonous
  - …
  - Increase the user's trust in the system
  - Provide a more convenient user interface

How can we know we are better?

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

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 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

A slightly exaggerated comparison

- Kaggle machine learning competitions
  - Defined dataset for training
  - Test dataset not revealed
  - Defined measures
  - Many competitors
  - (Sometimes 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 not necessarily share code

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 the classical recommendation task of rating prediction 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
    - Sometimes, researchers claim that their method is suited to make better recommendations
    - Then they use a rating dataset and transform all ratings to ones for evaluating an implicit feedback method
    - What is measured then, however, is how good we are at predicting who will rate what. Which does not necessarily mean better recommendations
  - 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: 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

Accuracy of recommenders

- In some domains, higher prediction accuracy almost directly leads to better systems
  - Language translation tasks
  - Image recognition tasks
- This analogy not necessarily holds for recommender systems
  - A small accuracy increase in a certain offline experiment might not tell us a lot about the quality of the resulting recommendations

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 and limited discovery
  - Optimizing for accuracy might lead to recommendations that are considered too "obscure" for users
    - Familiarity can be important for trust

Multi-metric evaluations

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

Multi-metric evaluations

- A number of works nowadays consider trade-offs (e.g., accuracy vs. diversity)
- However, limited work exists that actually validates the used computational metrics
  - e.g., whether increasing Intra-List-Diversity based on some content features actually increases the diversity *perception* of users

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

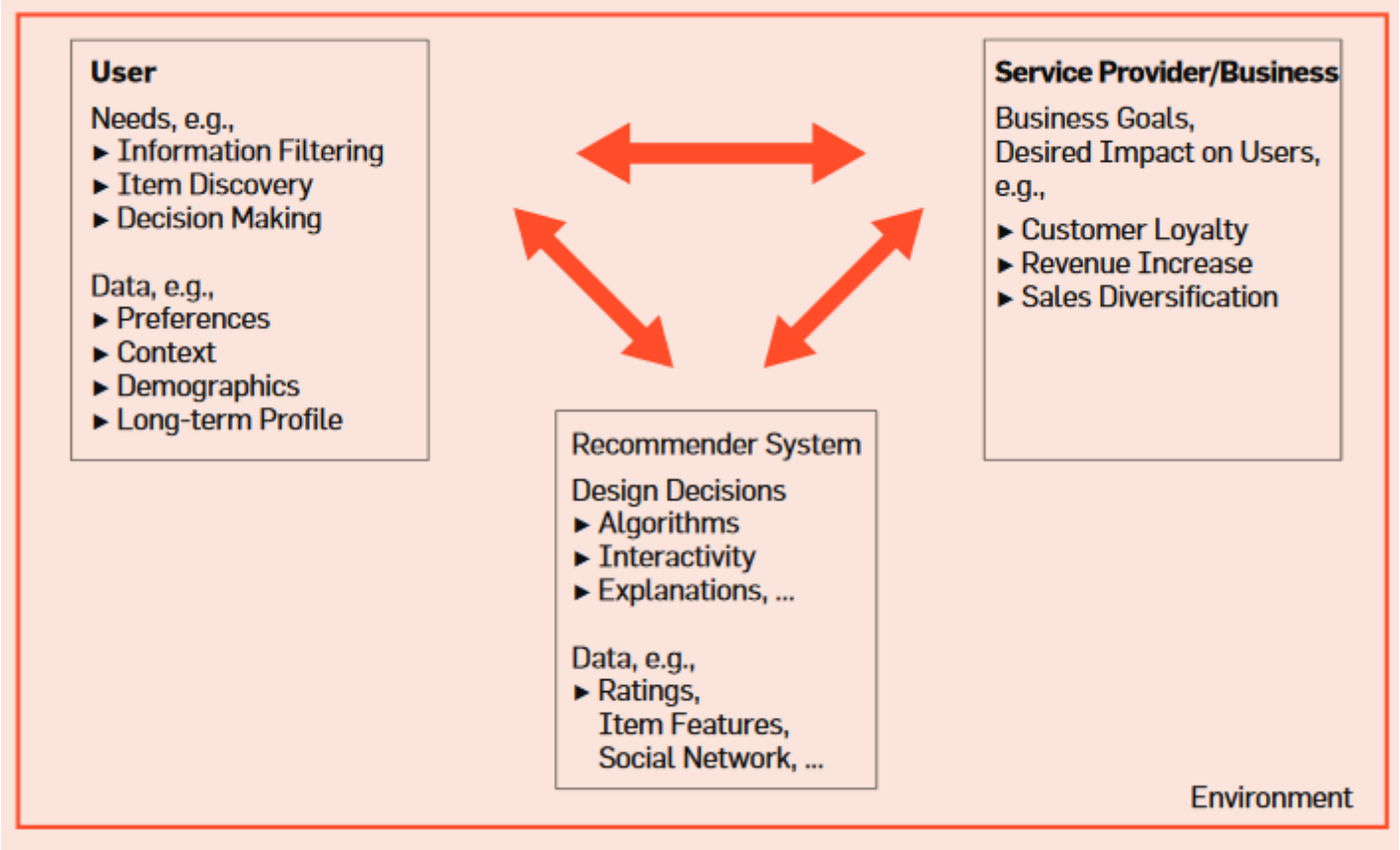
Possible steps forward

- 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
  - Use a multi-modal evaluation approach
    - including novel offline analyses, user studies, and qualitative research instruments

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"

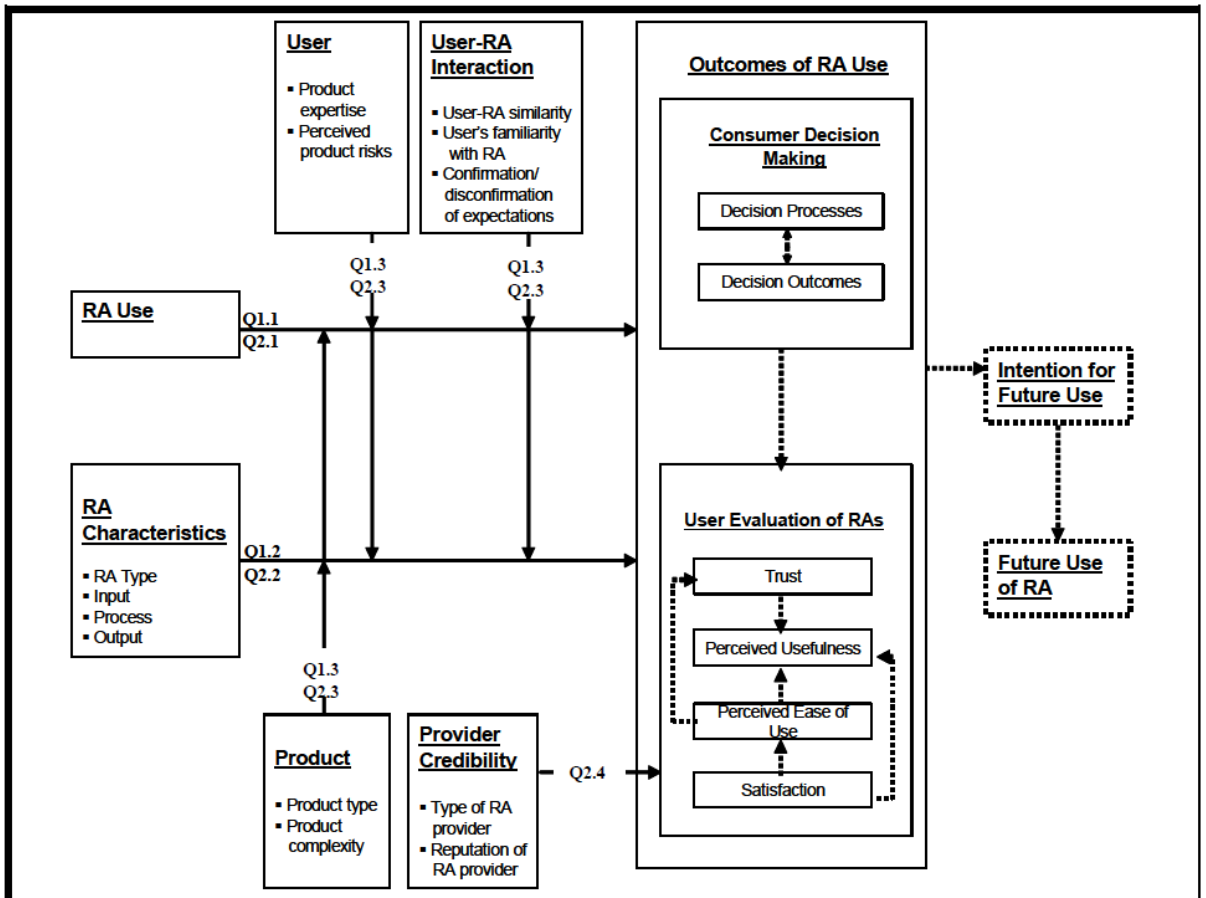


User-centric research

- 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



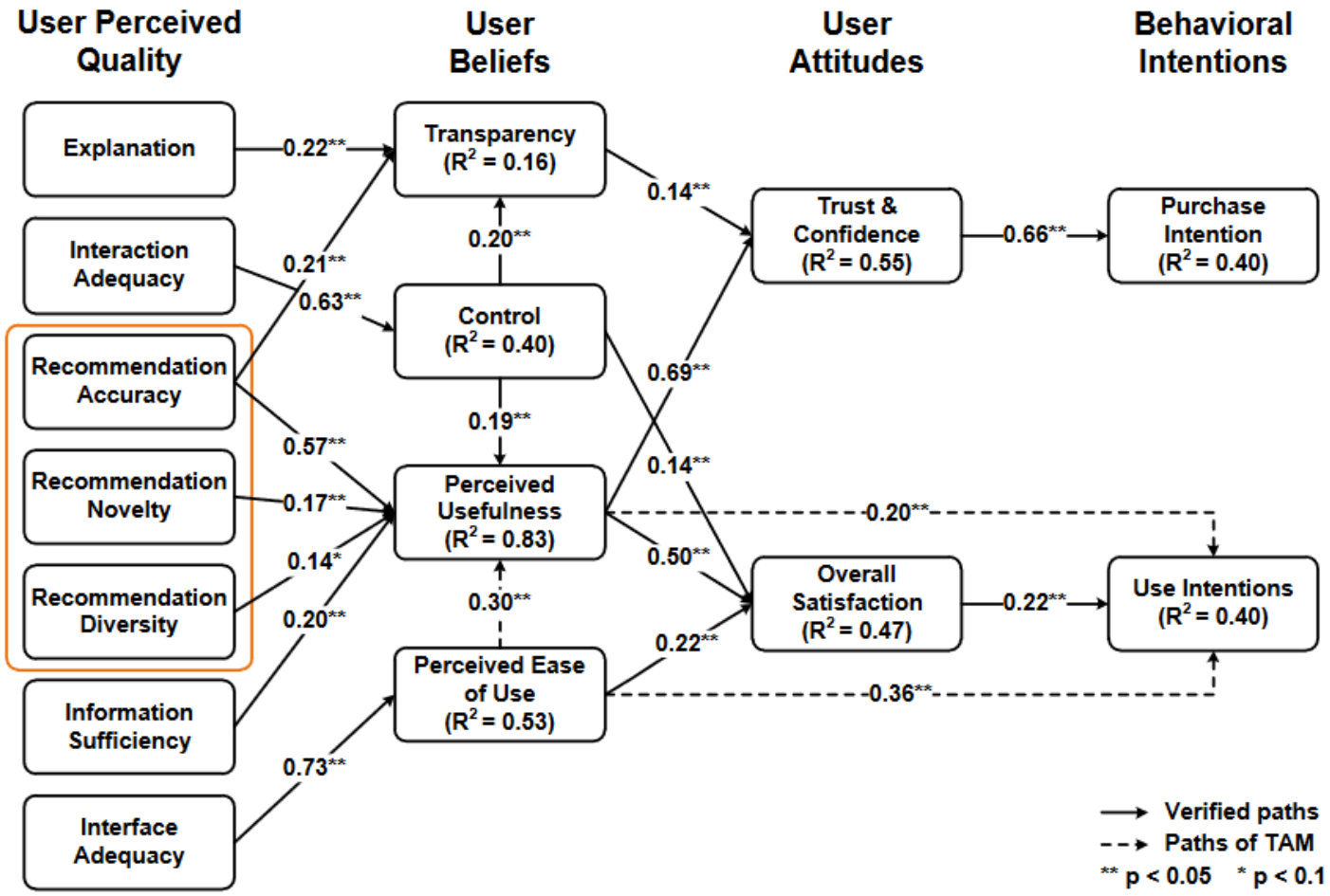
Note: Solid lines indicate relationships and constructs investigated in this paper.

Figure 1. Conceptual Model

User-centric research

- Different evaluation frameworks exist, e.g.,
  - Pu et al. (RecSys 2011, UMUAI 2012)
  - Knijnenburg at al. (UMUAI 2012)
- Frameworks describe relevant quality criteria
  - e.g., perceived accuracy, novelty, diversity, context compatibility, interface adequacy, information sufficiency and explainability, usefulness, ease of use
- and evaluation approaches
  - e.g., in terms of questionnaires

Example validation



Summary

- Business value of recommenders can be immense
- Definition of measurement is sometimes difficult
- Academic research often focused too much on algorithms
- A more comprehensive, impact-oriented approach is advisable

Jannach, D., Sar Shalom, O., Konstan, J.A., „Towards More Impactful Recommender Systems Research “, First Workshop on the Impact of Recommender Systems (impactRS@RecSys 2019), Copenhagen, 2019

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

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 Ferarri Dacrema et al.**
  - RecSys 2019