UNIVERSITÄT KLAGENFURT

Recommender Systems: Value, Methods, Measurements

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

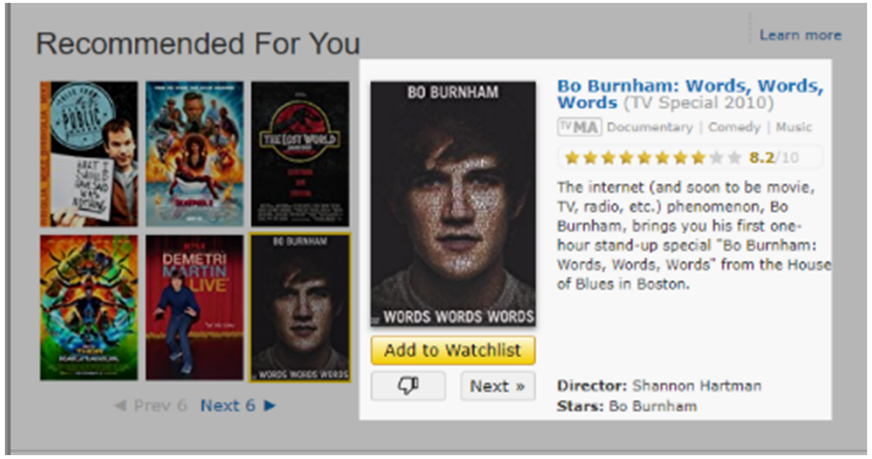
Presented at the ESSIR 2022 Summer School
Lisbon, 2022

Outline

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

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
- …

Part I: 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
    - e.g., on a hotel booking platform, see later

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

A Complex Multi-stakeholder 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, Vol. 30. 2020, pp. 127–158

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"

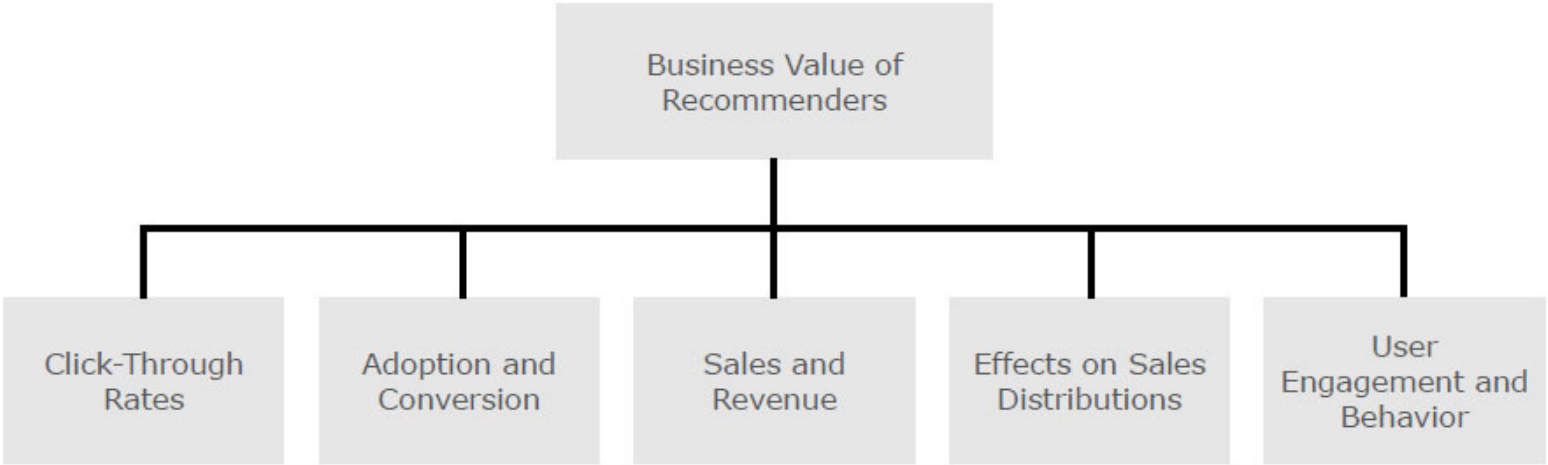
Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix Recommender System: Algorithms, Business Value, and Innovation.
ACM Trans. Manage. Inf. Syst. 6, 4, Article 13

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

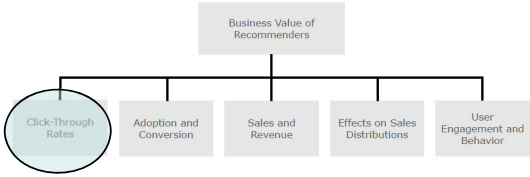
What is measured?

- Considering both the impact and value perspective



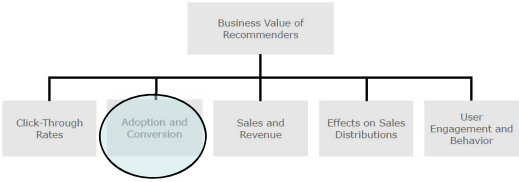
Jannach, D. and Zanker, M.: "*Impact and Value of Recommender Systems*". In: Recommender Systems Handbook. Ricci, F., Shapira, B. and Rokach, L. (Eds.), Springer US, 2021

Click-Through Rates



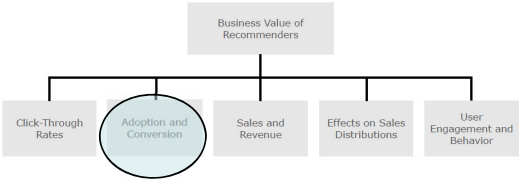
- Measures how many clicks are garnered by recommendations
  - Popular in the news recommendation domain
    - Google News: 38% more clicks compared to popularity-based recommendations
    - Forbes: 37% improvement through better algorithm compared to time-decayed popularity based method
    - swissinfo.ch: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



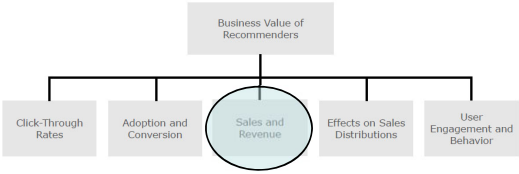
- CTR usually not the ultimate measure
  - Cannot know if users actually liked/purchased what they clicked on (consider also: click baits)
- Therefore
  - Various, domain-specific adoption measures common
- YouTube, Netflix: "Long CTR"/ "Take rate"
  - only count click if certain amount of video was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

Adoption and Conversion Rates



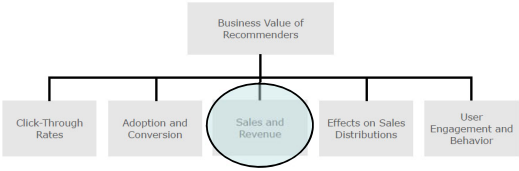
- Alternatives when items cannot be viewed/read:
- eBay:
  - "purchase-through-rate", "bid-through-rate"
- Other:
  - LinkedIn: Contact with employer made
  - Paper recommendation: "link-through", "cite-through"
  - E-Commerce marketplace: "click-outs"
  - Online dating: "open communications", "positive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

Sales and Revenue



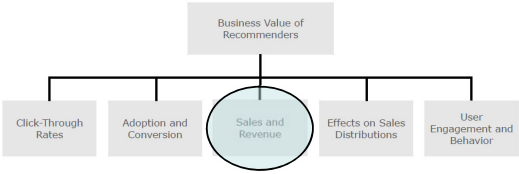
- CTR and adoption measures are good indicators of relevant recommendations
- However:
  - Often unclear how this translates into business value
  - Users might have bought an item anyway
  - Substantial increases might be not relevant for business when starting from a very low basis
- In addition:
  - 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



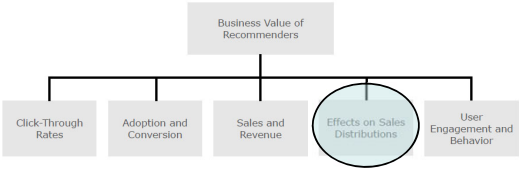
- e-grocery studies:
  - 1.8 % direct increase in sales in one study
  - 0.3 % direct effects in another study
  - However:
    - Up to 26% indirect effects, e.g., where customers were pointed to other categories in the store
    - "Inspirational" effect also observed in music recommendation in our own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

Jannach, D. and Hegelich, K.: *"A Case Study on the Effectiveness of Recommendations in the Mobile Internet"*.
In: Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys 2009). New York City, New York, 2009, pp. 205-208

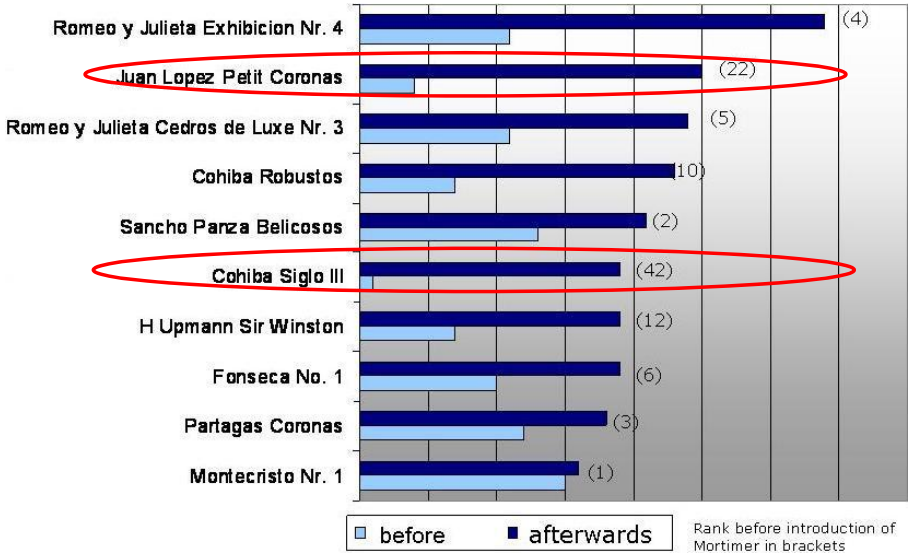
Effects on Sales Distributions



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

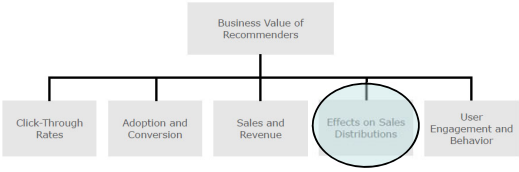
Effects on Sales Distributions

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



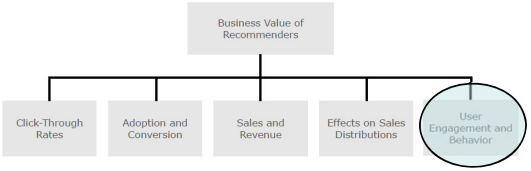
Zanker, M., Bricman, M., Gordea, S., Jannach, D. and Jessenitschnig, M.: "*Persuasive Online-Selling in Quality and Taste Domains*". In: 7th International Conference on Electronic Commerce and Web Technologies (EC-Web 2006). Krakow, Poland, 2006, pp. 51-60

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

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 a given 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
- 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

Overarching goal of the system, strategic value
Recommendation purpose / Intended utility
System (algorithm) task
Computational metrics

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 first part

- 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

Part II: Methods

A bit of history

- 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

Recommender Systems and IR

- Partially shared goals:
  - Help users find relevant content
- Similar task:
  - Determine a ranked list of items
- Related algorithms and techniques
  - Document encoding, ranking
- Main differences:
  - Explicit queries (IR) vs. learned user profile (RS)
  - Personalization is central in RS

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

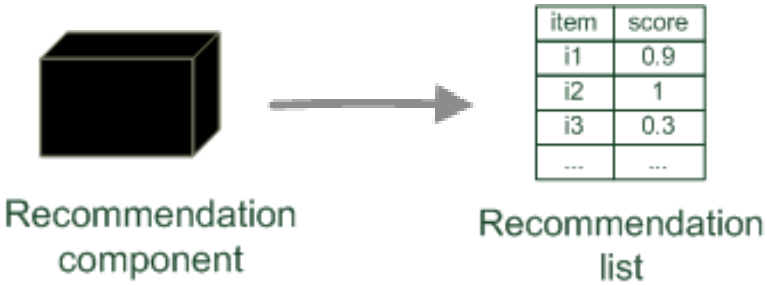
- Interactive Recommendation

Information Filtering roots

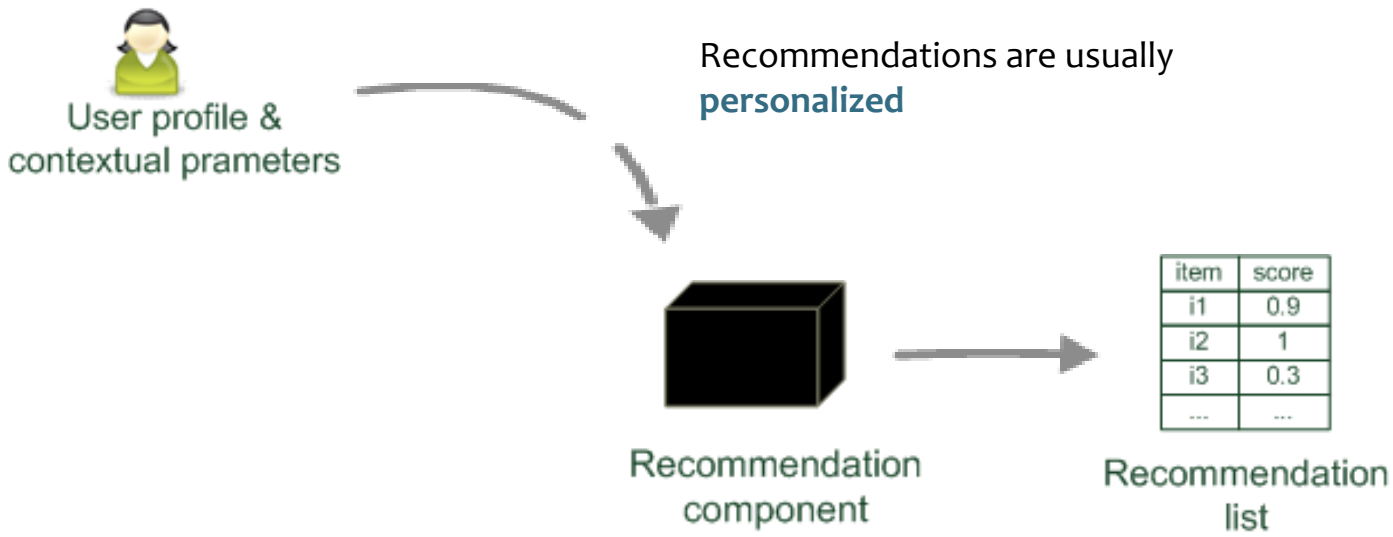
- Information Filtering
  - Systems that filter incoming streams of information in a personalized way
  - Dates back to the late 1960s
  - Early systems use explicitly stated preferences regarding topics or keywords
  - Later on, automated content analysis and user profiling
- Today:
  - “Content-based Filtering” recommender techniques
  - Personalized Information Retrieval

Recommendation Principles

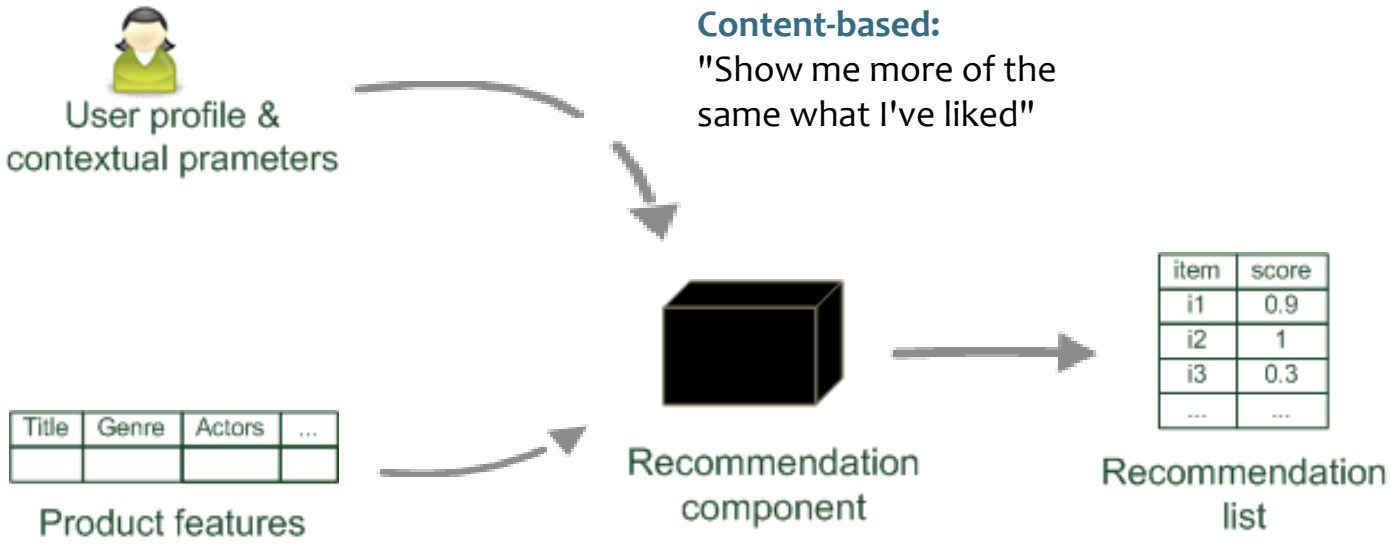
Recommender systems reduce information overload by estimating relevance



Recommendation Principles



Content-based Filtering



Outline

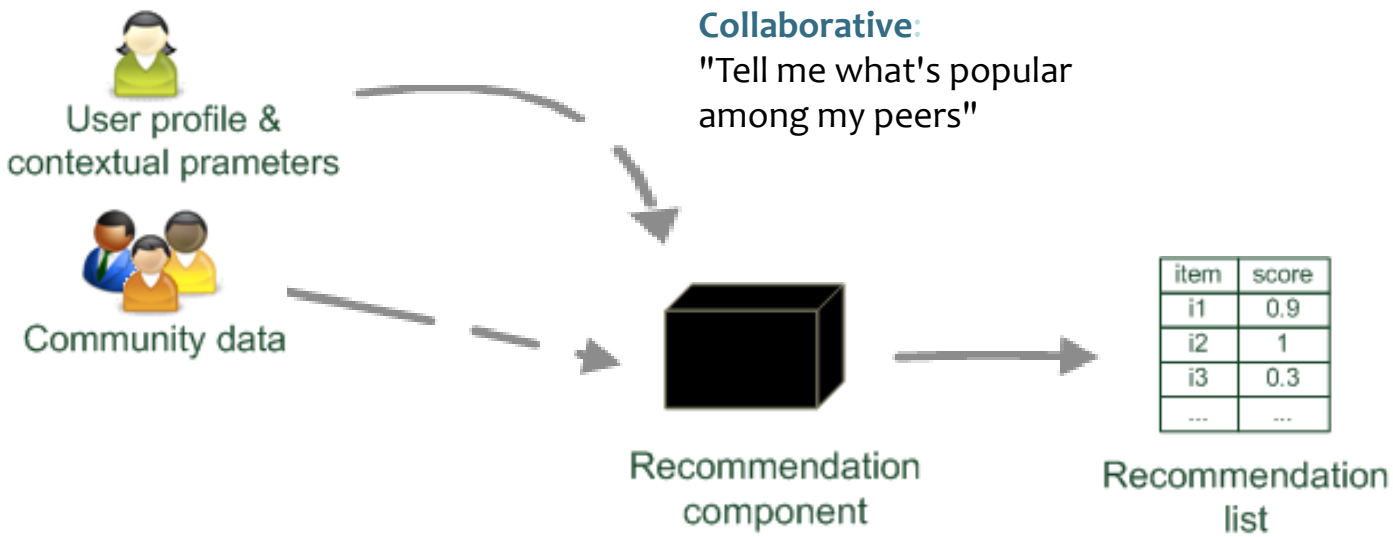
- Content-based Filtering
- Collaborative Filtering
- Hybrid Systems
- Knowledge-based Systems

- Interactive Recommendation

Leveraging the opinions of others

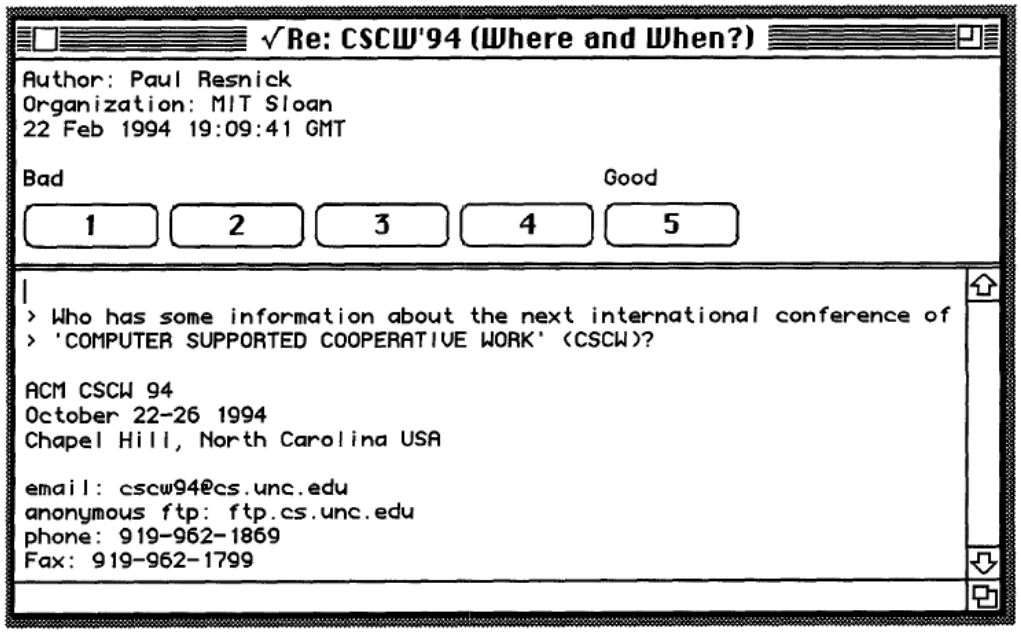
- 1982: ACM president complained about email junk
  - Envisioned a set of "trusted authorities" that assess the quality of the messages
- 1987: Information Lens
  - Based on manual filters, but could also specify people whose opinions they value
- 1992: Tapestry – "Collaborative Filtering"
  - Continued Information Lens ideas, introduced idea of considering ratings, but still a manual process
- 1994: GroupLens and others
  - System automatically predicted ratings of users

Collaborative Filtering



Collaborative Filtering

- The predominant approach since 1994
- 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

- Recommendation considered as matrix completion ("matrix filling") problem

	Item1	Item2	Item3	Item4	Item5
Alice	5	?	4	4	?
User1	3	?	2	3	?
User2	?	3	4	?	?
User3	?	3	1	?	4
User4	1	?	5	2	1

- GroupLens
  - Relies on a user-based nearest-neighbor method (User-KNN)

User-KNN

- Given an "active user" (Alice) and an item I not yet seen by Alice
  - The *goal is to estimate Alice's rating for this item*, e.g., by
    - find a set of users (peers) who liked the same items as Alice in the past **and** who have rated item I
    - use, e.g., the average of their ratings to predict, if Alice will like item I
    - do this for all items Alice has not seen and recommend the best-rated

KNN Methods

- Some questions
  - How do we measure similarity?
  - How many neighbors should we consider?
  - How do we generate a prediction from the neighbors' ratings?
  - How to make this scalable?

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

Matrix Completion

- Recommendation as matrix completion
  - Problem often reduced to learn parameters of a function to predict the missing entries
  - Algorithms can be compared by their prediction (post-diction) of some known, but held-out ratings
  - Measures, e.g., Root Mean Square Error

$$RMSE = \sqrt{\frac{\sum_{(u,i)\in K} (\hat{r}_{ui} - r_{ui})^2}{|K|}}$$

Collaborative Filtering success (CF)

- 1998:
  - Dimensionality reduction for CF, clustering
  - Collaborative/Content-based Hybrids
- 1999: It works in e-commerce!
  - First reports on successful applications in practice (e-commerce, music, video)
- 2000: Item-to-item collaborative filtering
- 2003: Amazon.com
  - Report on the successful use of recommendations at Amazon.com using item-to-item filtering

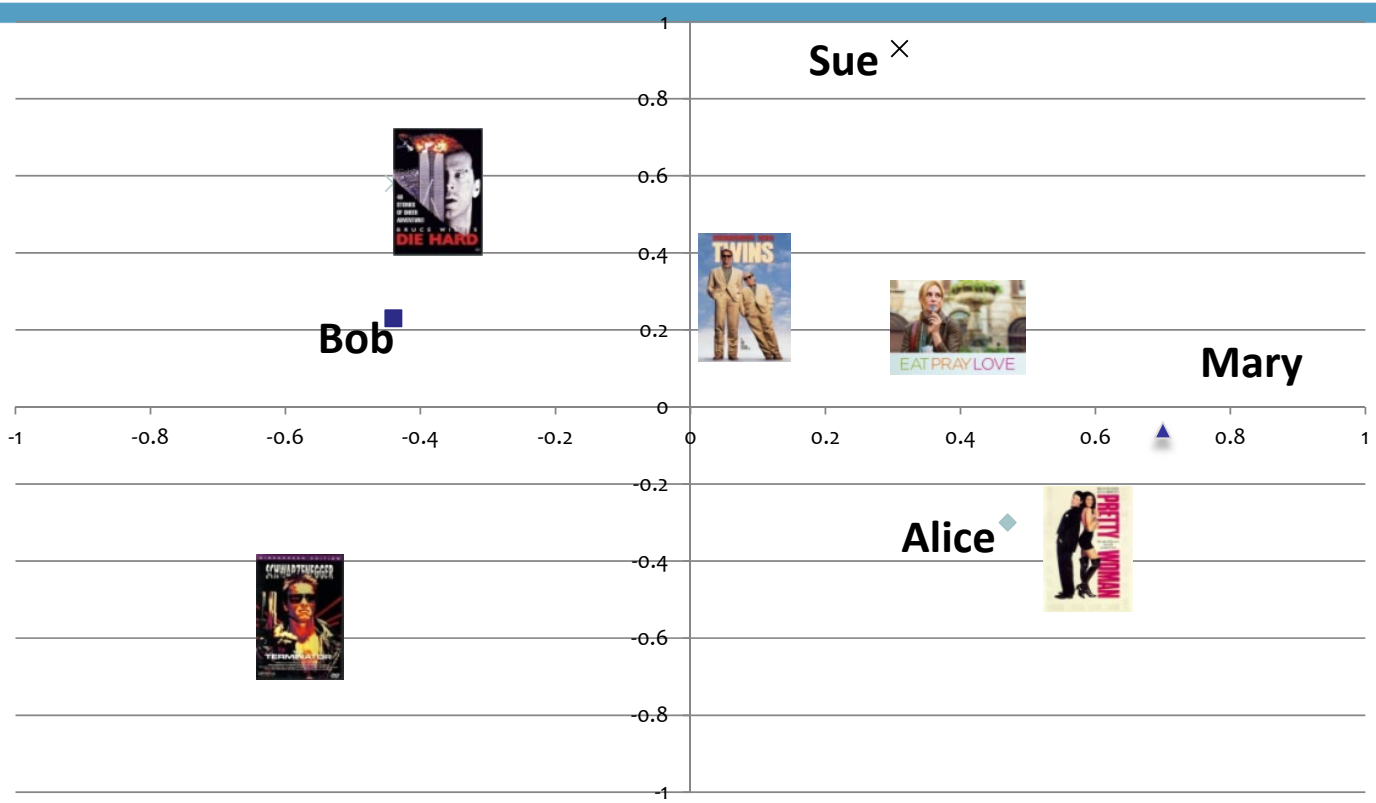
The Netflix Prize (2006-2009)

- Netflix announced a 1 million dollar prize in 2006
  - For beating their system by 10% in terms of the prediction error
  - Provided at that time huge dataset
- Effects
  - Further boosted research on matrix completion
- Contest ended in 2009, some winning ingredients:
  - Matrix factorization, ensemble methods

Matrix Factorization (MF)

- 2000: Early experiments with Singular Value Decomposition
  - Use SVD for dimensionality reduction
  - Capture the most important factors/aspects in the data
  - Should also help to reduce noise
- 2006 and later: MF variants using, e.g., gradient descent optimization

Projection into lower-dim. space



Post-Netflix-Prize Developments

- Rating prediction increasingly considered irrelevant in practice
  - Item relevance prediction still important
- Various ranking-based methods ("learning-to-rank") proposed around 2009
- More focus on situations where only implicit feedback is available
- Probably hundreds of CF algorithms per year
- In the last seven years, mostly using deep learning techniques

Matrix Completion - Benefits

- Benefits of the problem abstraction
  - Problem abstraction is domain-independent
  - Fosters design of algorithms that are not tied to a certain application
  - Established evaluation procedures exist
  - Reproducibility of results, in theory, is easy
  - A number of public datasets exist

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

- An alternative problem abstraction
  - Aims to address different various real-world application problems
  - Input is not a rating matrix, but a sequential log of recorded user interactions
    - Item views, purchases, listening events
  - Most common problem is to predict items that are relevant in the user's ongoing session
    - Often, users are anonymous and the user's intent must be guessed from a small set of interactions ("session-based recommendation")

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

Session-based Recommendation

- Guessing the intention can be difficult

Roll over image to zoom in

Minnow Sports

Minnow Sports Aluminum Baseball Bat For Baseball & Teeball

8 customer reviews

Price: ~~$29.99~~
Sale: $19.99
You Save: $10.00 (33%)

In Stock.

This item does not ship to **Germany**. Please check other sellers who may ship internationally. Learn more

Sold by BBro Store and Fulfilled by Amazon. Gift-wrap available.

Item Display Length:

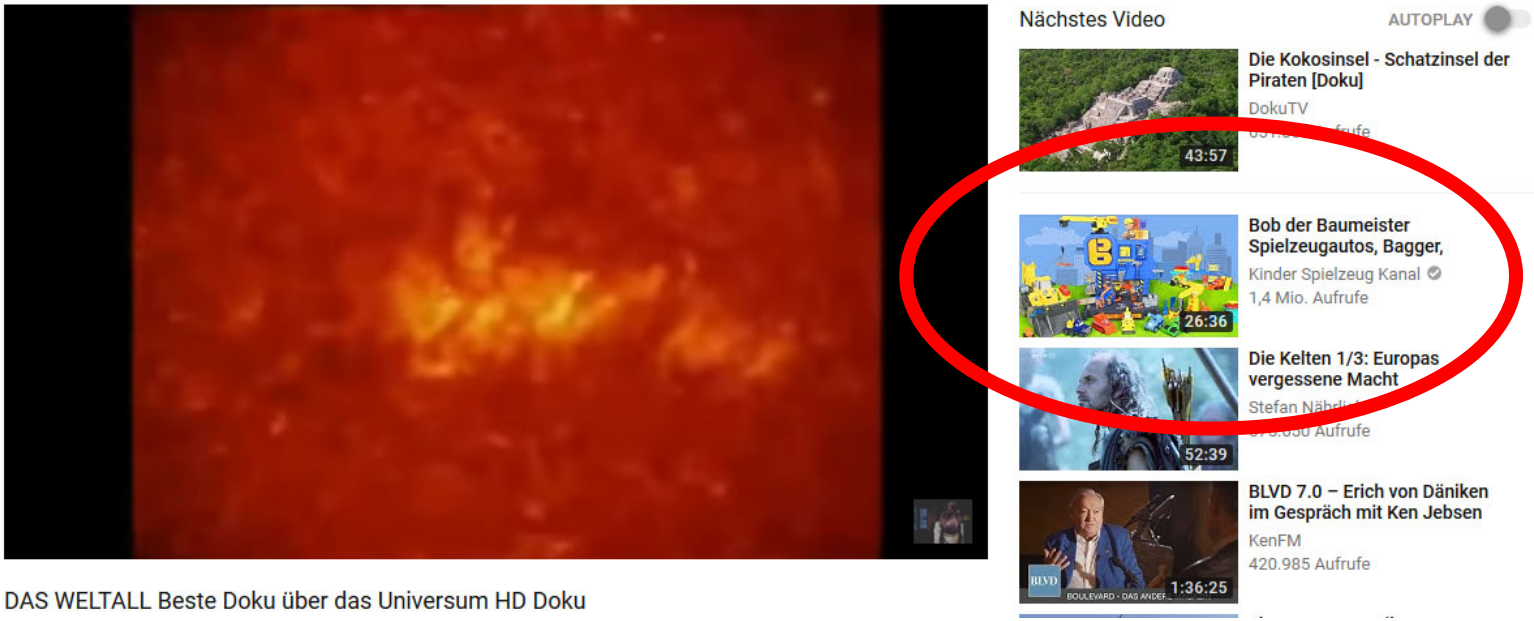
32.0 inches

- Made from lightweight high grade Aluminum alloy for faster swing speed
- Ultra-thin 32" handle with All Sports grip for increased stability and accuracy
- Stylish design featuring full rolled-over end for ultimate performance
- Ideal for all levels of baseball players from practice to matches
- 32 inches in length & 24 ounces

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
43:57
Bob der Baumeister Spielzeugautos, Bagger,
Kinder Spielzeug Kanal
1,4 Mio. Aufrufe
26:36
Die Kelten 1/3: Europas vergessene Macht
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

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 this problem "session-aware" recommendation
- One main problem is to effectively combine long-term and short-term preference models

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

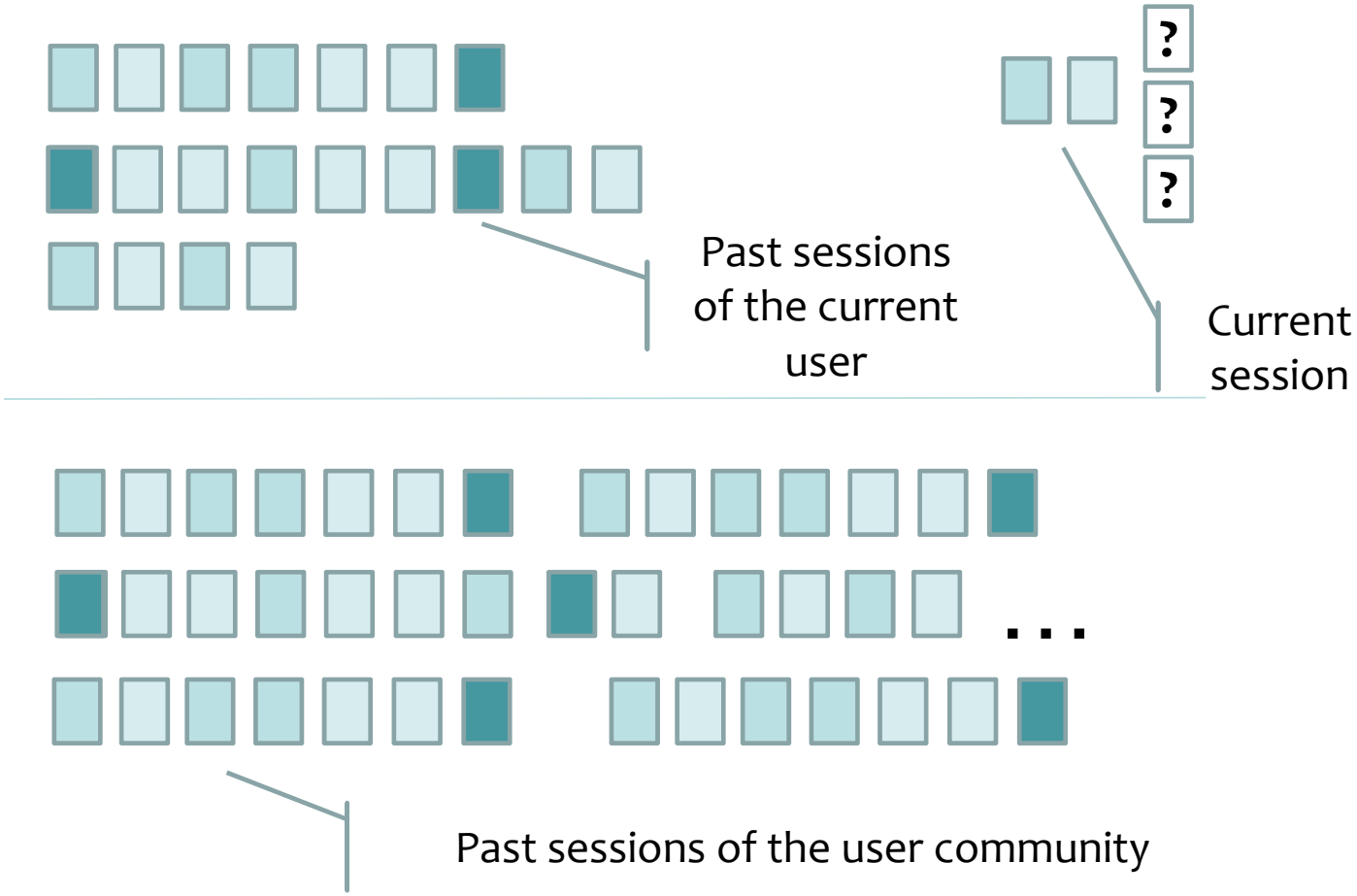
What to recommend?

- Some plausible options
  - Only shoes or only watches?
  - Mostly Nike shoes?
  - Maybe also some T-shirts?
- Considerations and observations
  - Using the matrix completion formulation, the system will mostly recommend T-shorts and trousers
  - Research indicates that both models are relevant, but that the short-term model is much more important

Quadrana, M., Karatzoglou, A., Hidasi, B., Cremonesi, P.: Personalizing Session-based Recommendations with Hierarchical Recurrent Neural Networks. RecSys 2017: 130-137

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

A Problem Abstraction



Technical Approaches

- Basic techniques
  - Item co-occurrences: “Customers who bought … also bought”
  - Markov Chains and Sequential Rules
- Nearest neighbors
  - Find past sessions that are similar to the current (ongoing) one, predict items from neighbor sessions
- Sequence learning / modeling
  - Embeddings, Recurrent Neural Networks

Applications and History

- Early applications for next-page prediction in web browsing
- Next-track music recommendations and automated radio stations, video playlists
- Next-POI recommendation in travel and tourism applications
- E-commerce applications, increasingly since 2015
  - In particular many neural methods proposed recently
  - Publicly available datasets

Outline

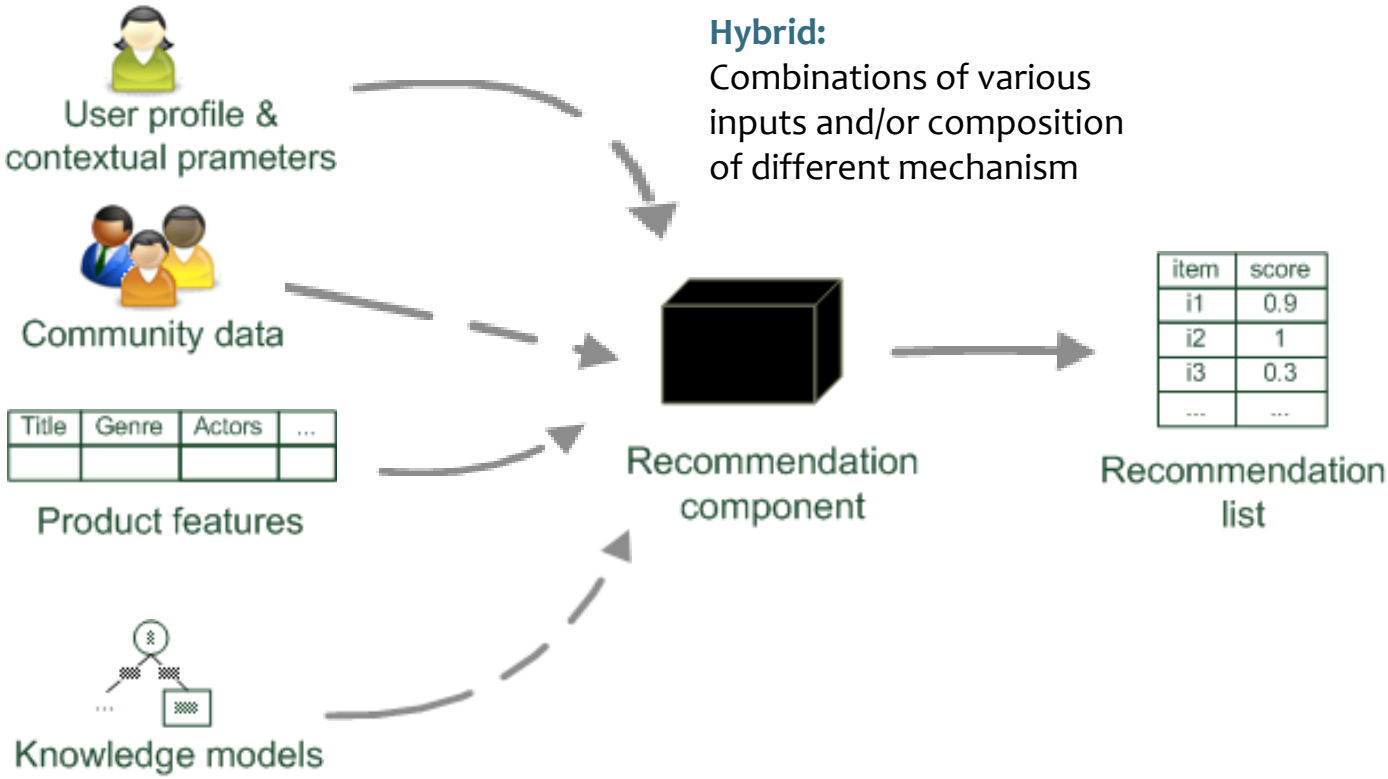
- Content-based Filtering
- Collaborative Filtering
- Hybrid Systems
- Knowledge-based Systems

- Interactive Recommendation

Hybrid Systems

- Leveraging other types of knowledge
- Independent of problem abstraction
- Typical combinations, e.g.,
  - Collaborative and content-based approaches
  - Considering demographics
- Idea is to combine advantages of individual approaches, e.g.,
  - Use side information when there is no collaborative information (yet) for some users

Hybrid Recommendation Approach



Hybrid:
Combinations of various inputs and/or composition of different mechanism

Content-based Methods and Hybrids

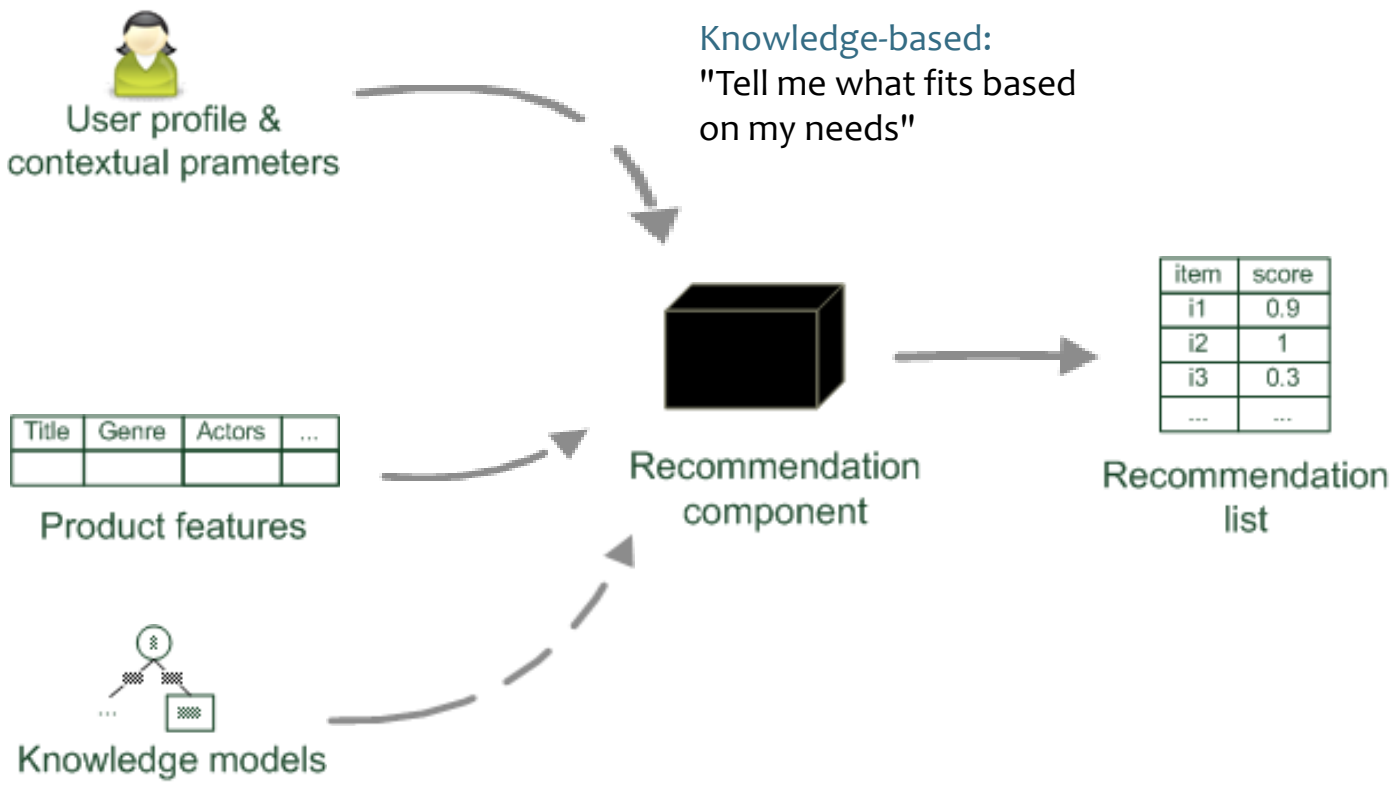
- Pure content-based techniques are rarely used for recommendation
  - They are limited to finding similar items
  - Content encodings (e.g., TFIDF, embeddings) tell us little about the general quality of the items
    - Recommendations can be obscure or too niche
- Very common, however:
  - Leverage information about items or users in combination with collaborative filtering approaches
  - In particular helpful for cold-start scenarios

Outline

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

- Interactive Recommendation

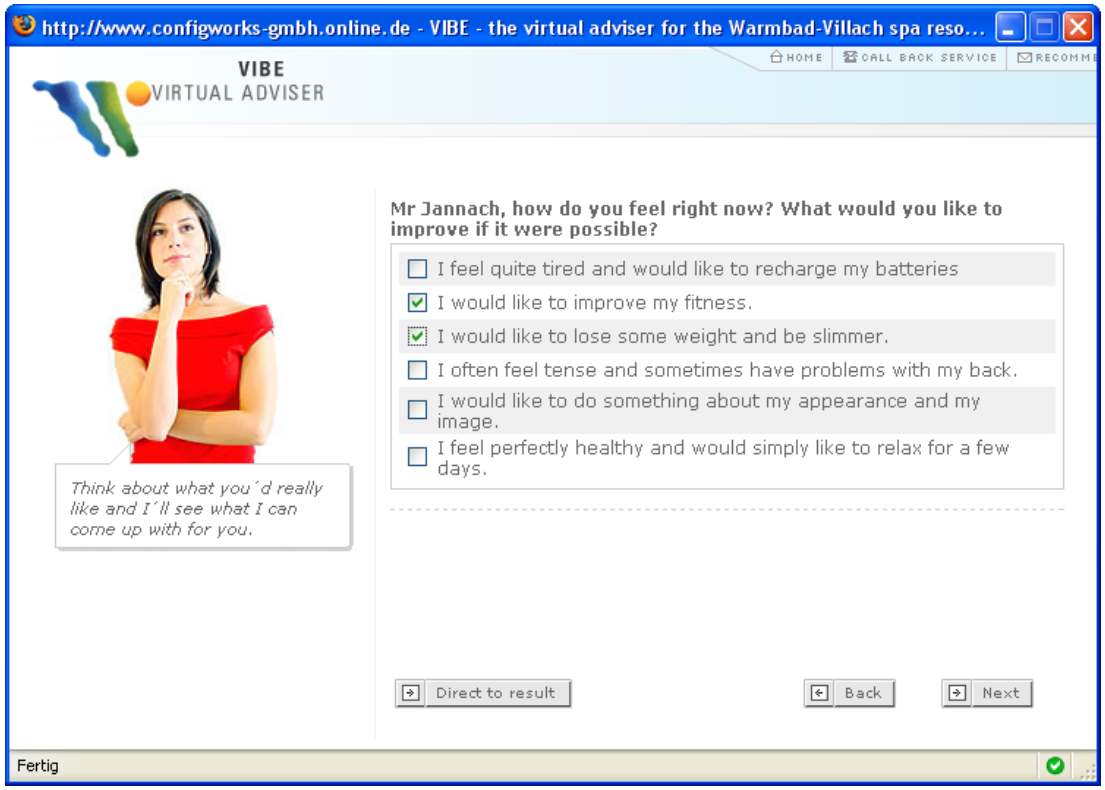
Knowledge-based Systems



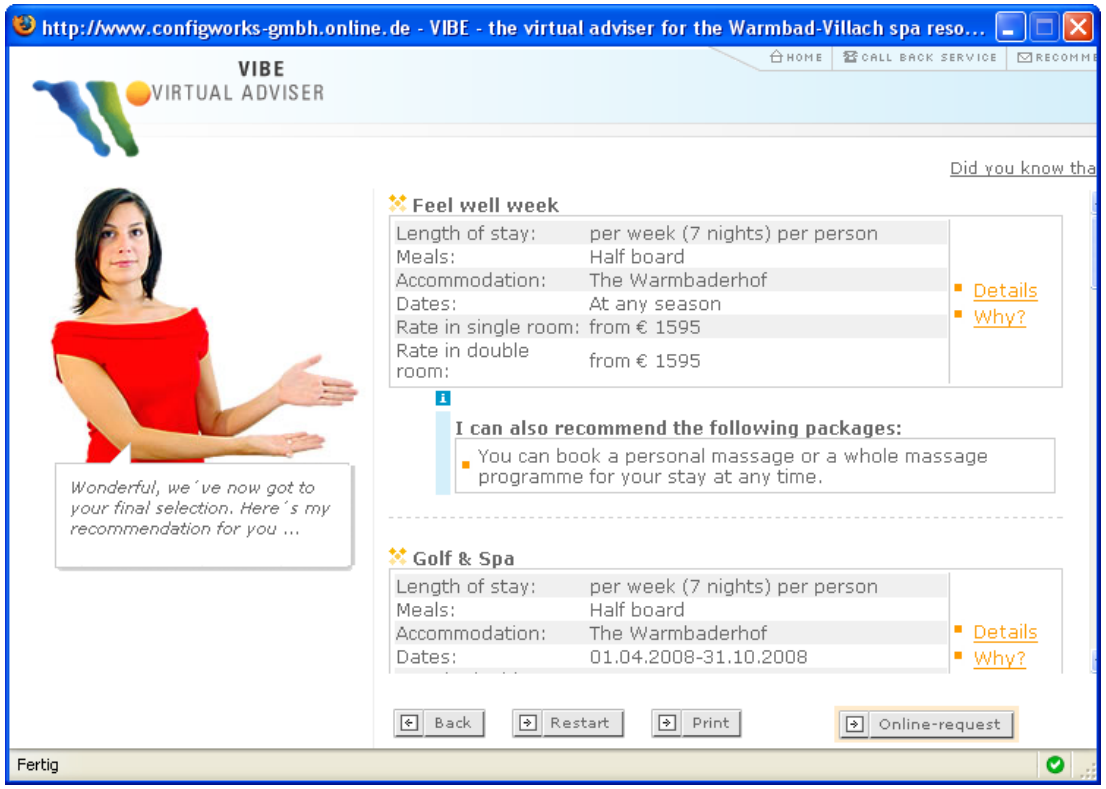
Knowledge-based Systems

- Explicitly encode recommendation knowledge
- Usually no learning, but knowledge engineering
- Used for certain application domains, e.g.,
  - One-time investments and decisions
  - Domains where technical constraints have to be considered
  - **Interactive/conversational** recommendations, chat bots

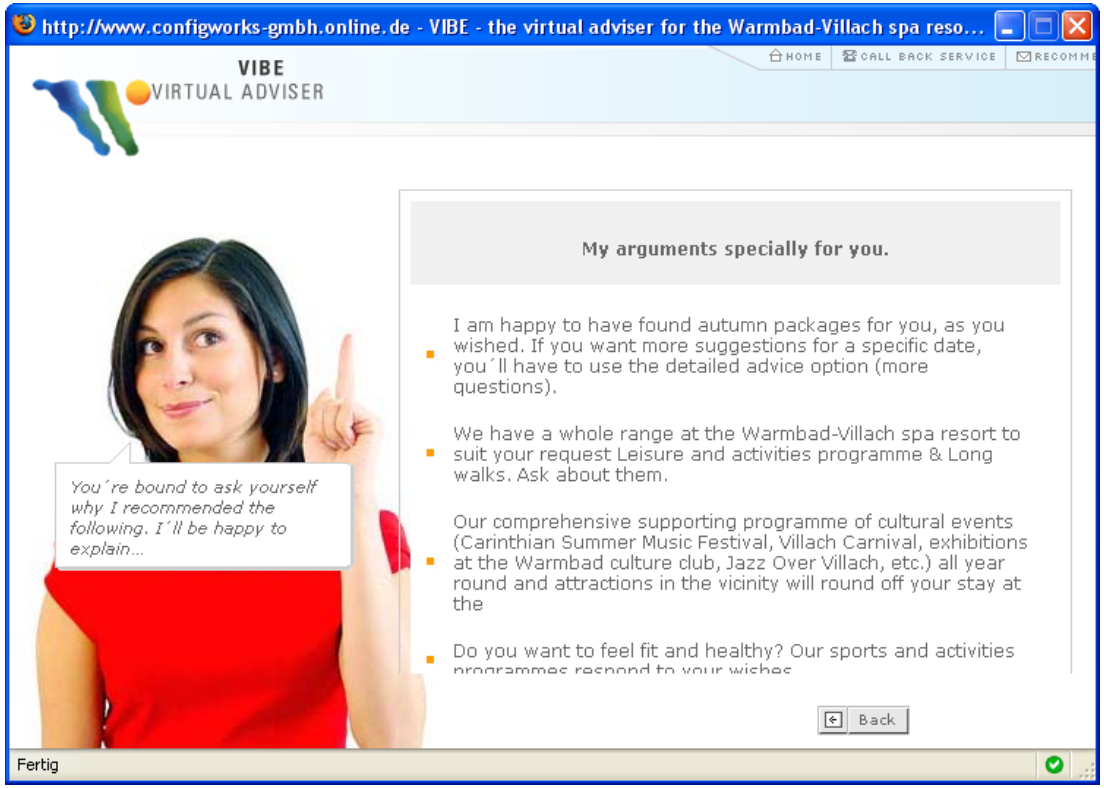
Is this even a recommender?



Is this even a recommender?



Is this even a recommender?



Outline

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

- Interactive Recommendation

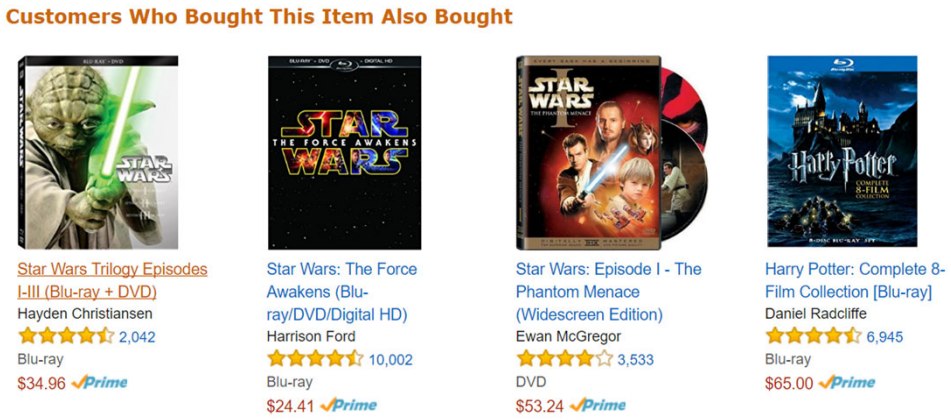
From Algorithms to User Experience

- 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
- 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).

Interactive Recommender Systems

- But: A common assumption in many research works: Which user interaction?
  - The system monitors what I do
  - And then shows me stuff
  - Which I can click on



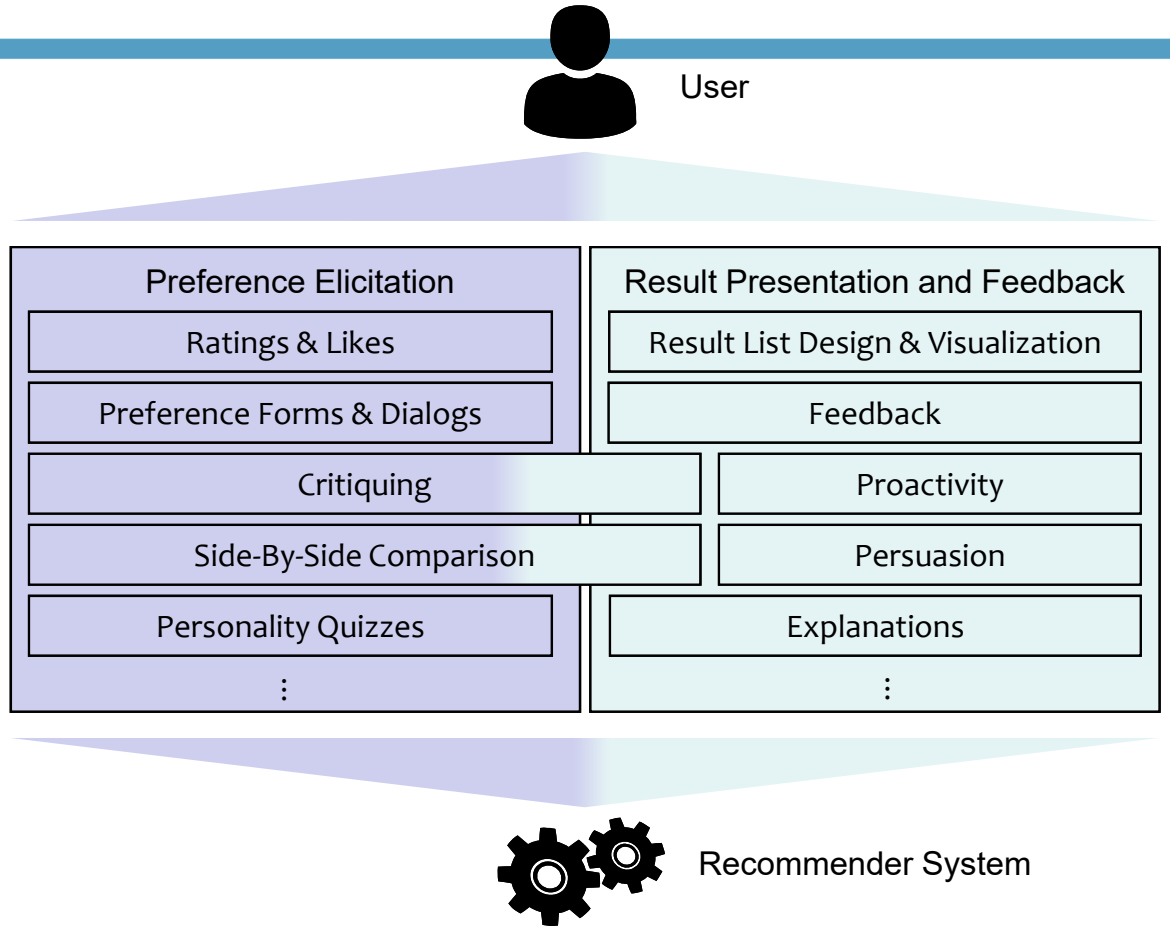
Source: Amazon.com

UI research for Recommenders

- HCI research is one of the main roots of recommender systems research
- Nonetheless, UI-related aspects seem less explored than algorithmic questions
  - One reason lies in the difficulty of evaluating new proposals
  - Existing research is also largely scattered
- Recent revival: Conversational Recommender Systems
  - "A Grand AI Challenge"

Jannach, D. and Chen, L.: "*Conversational Recommendation: A Grand AI Challenge*". AI Magazine, Vol. 43(2). 2022

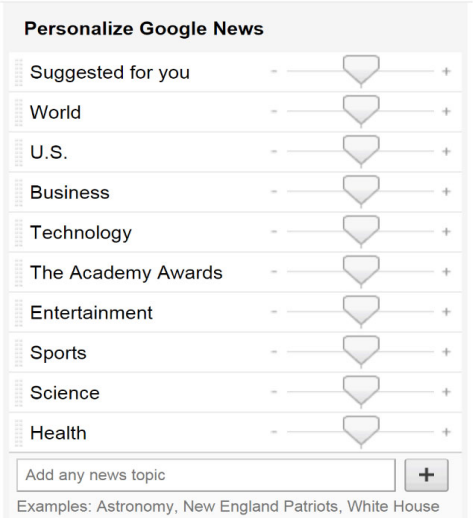
Structuring Existing Works

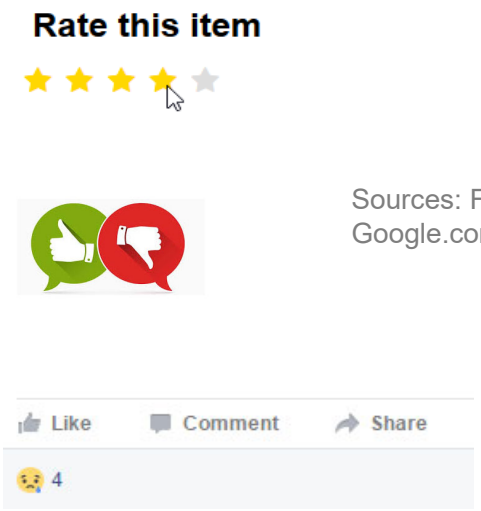


Jugovac, M. and Jannach, D.: *"Interacting with Recommenders - Overview and Research Directions"*. ACM Transactions on Intelligent Interactive Systems (ACM TiiS), Vol. 7(3). 2017

Design Space Examples

- Telling the system explicitly what you like
  - Global settings
  - Ratings
    - But how many options? How many categories?

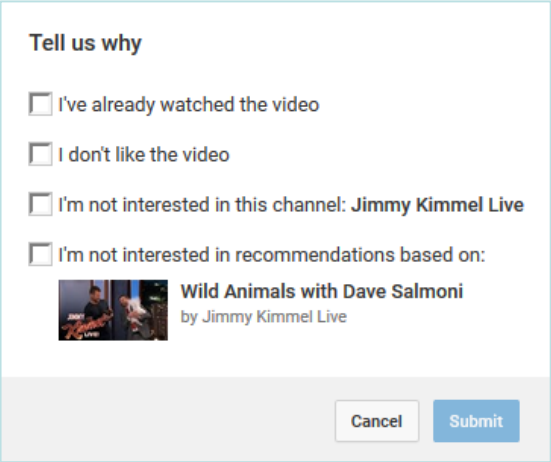




Sources: Facebook.com, Google.com

Design Space Examples

- What to display as recommendation?
  - The items of course
    - How many? Where on the screen? Multiple lists?
- Should users be able to give feedback?
    - Like/Dislike?
    - Or more?

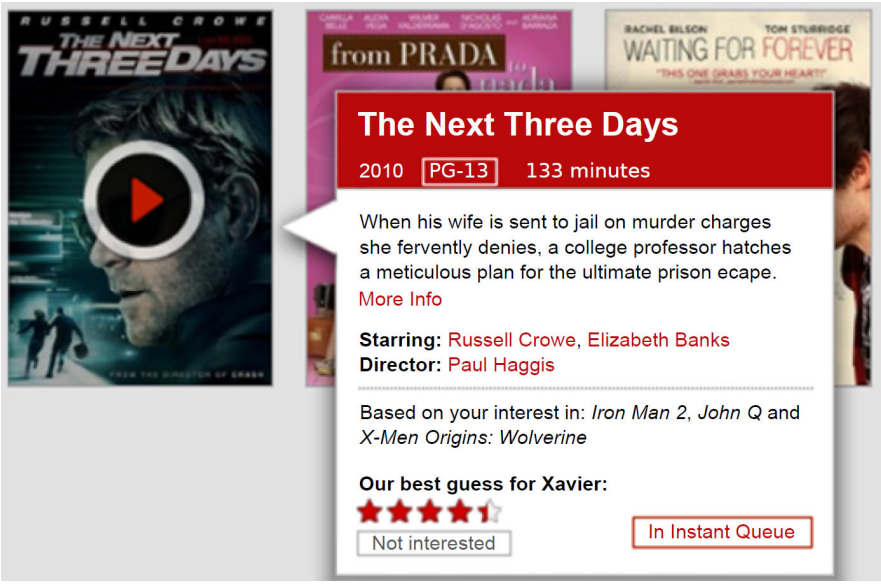


Source: Youtube.com

List Design Considerations

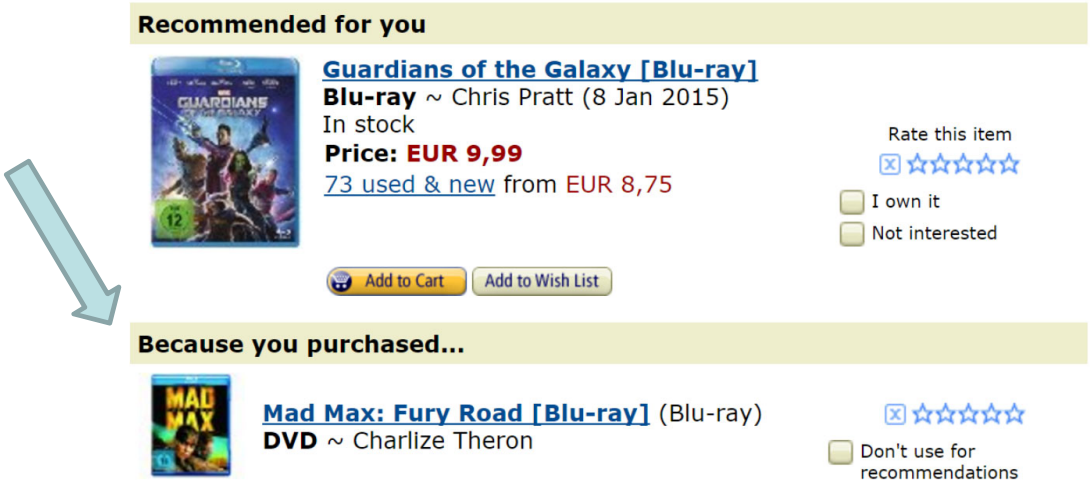
What else to show?

- What to display in addition to a nice picture?
  - Maybe some explanation, but which one?
  - A predicted rating?



Explanations and Control

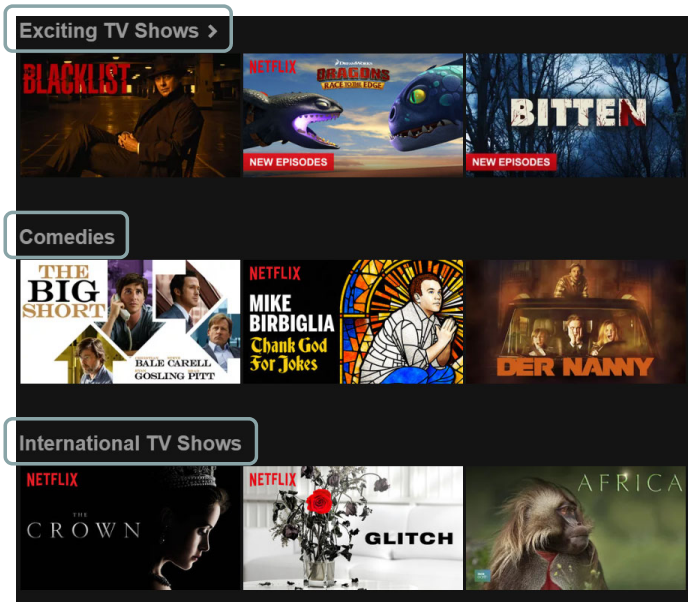
- What to display in addition to a nice picture?
  - Maybe some explanation, but which one?
  - Or our logic to recommend this?



List Content

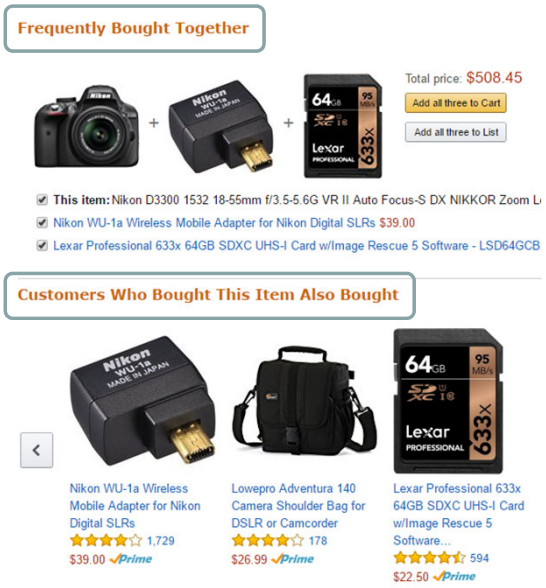
- Grouping items in a list

Semantically



Source: netflix.com

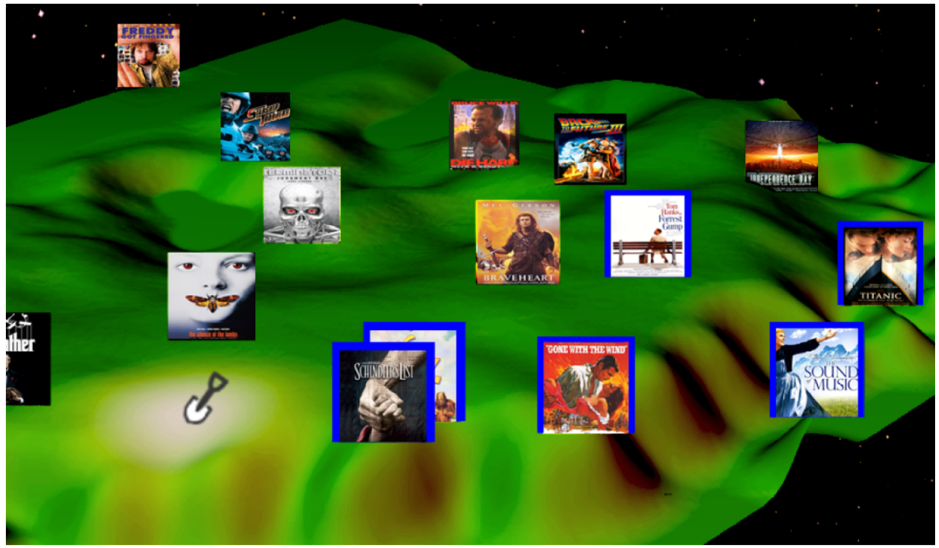
Statistically



Source: amazon.com

Recommendations in 3D space

- Additional 3rd dimension for extra information (e.g. user profile)

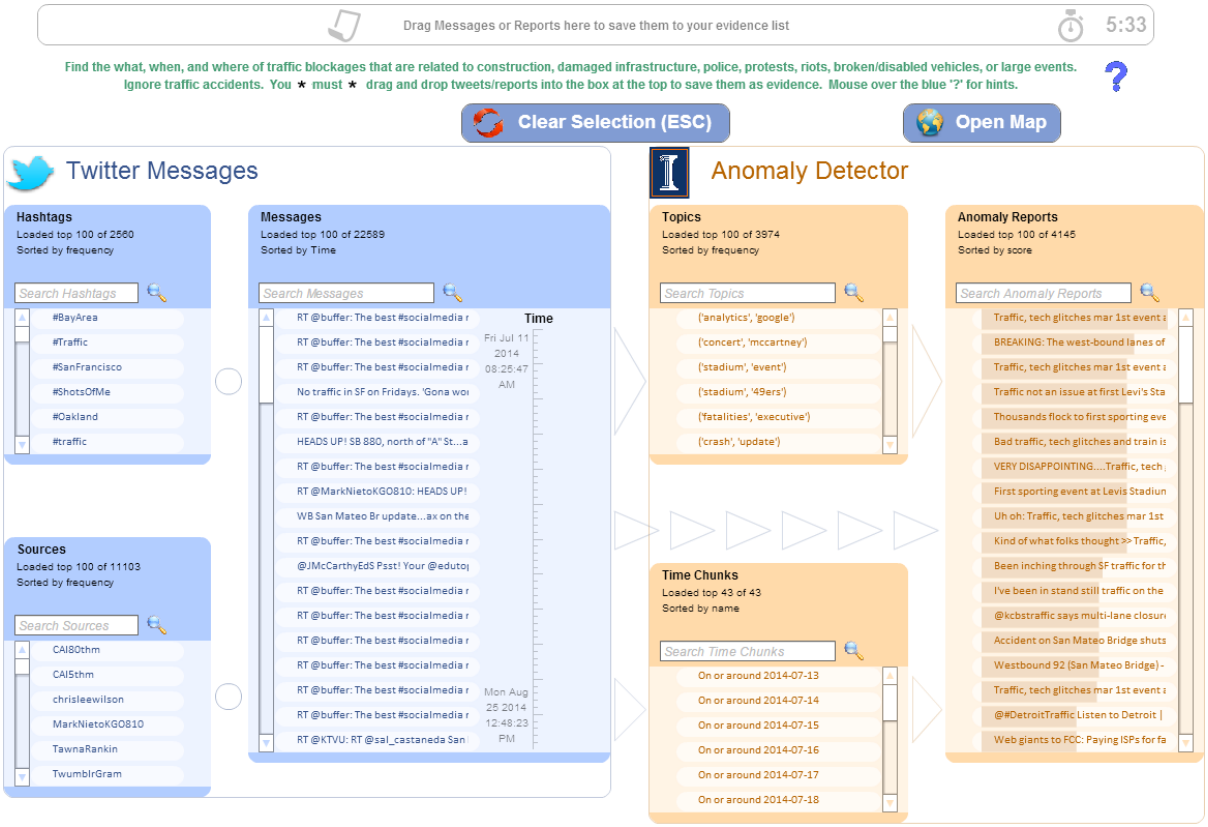


Source: [1]

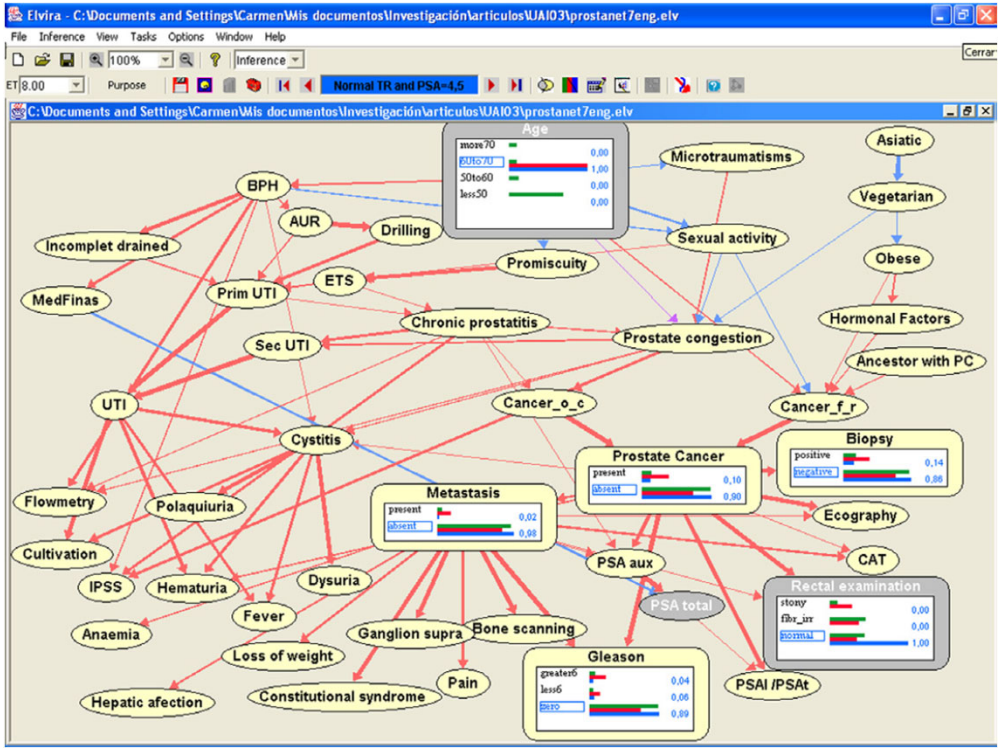
More on explanations

- Should we explain the recommendations?
- What would be the purpose of the explanations?
  - Persuade, increase trust, increase decision efficiency, help users make better decisions?
- How should we visualize the explanations?
- Should we personalize the explanations?

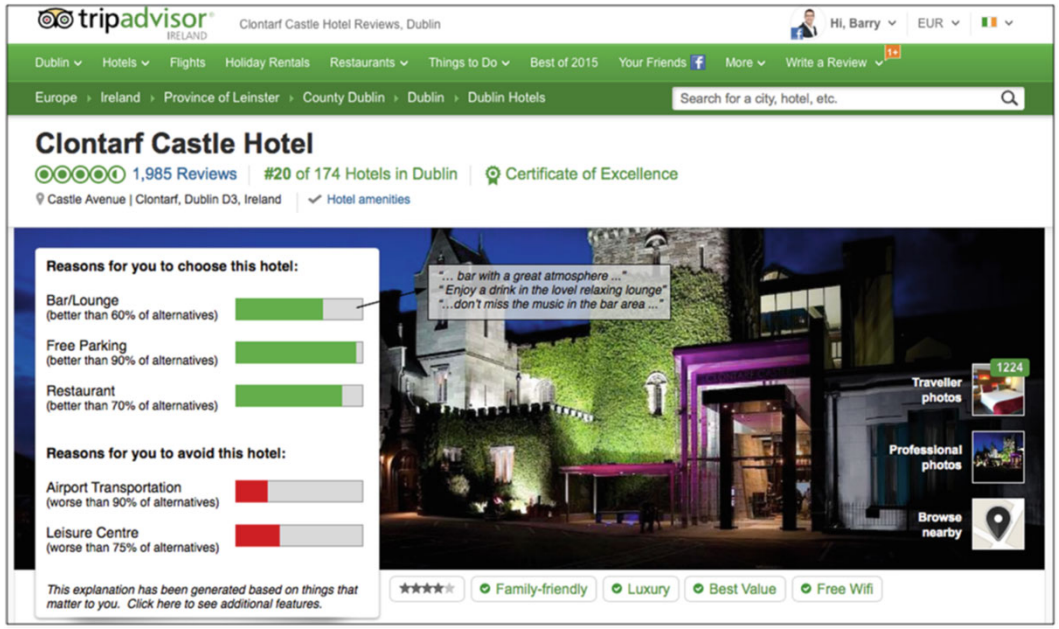
An academic example



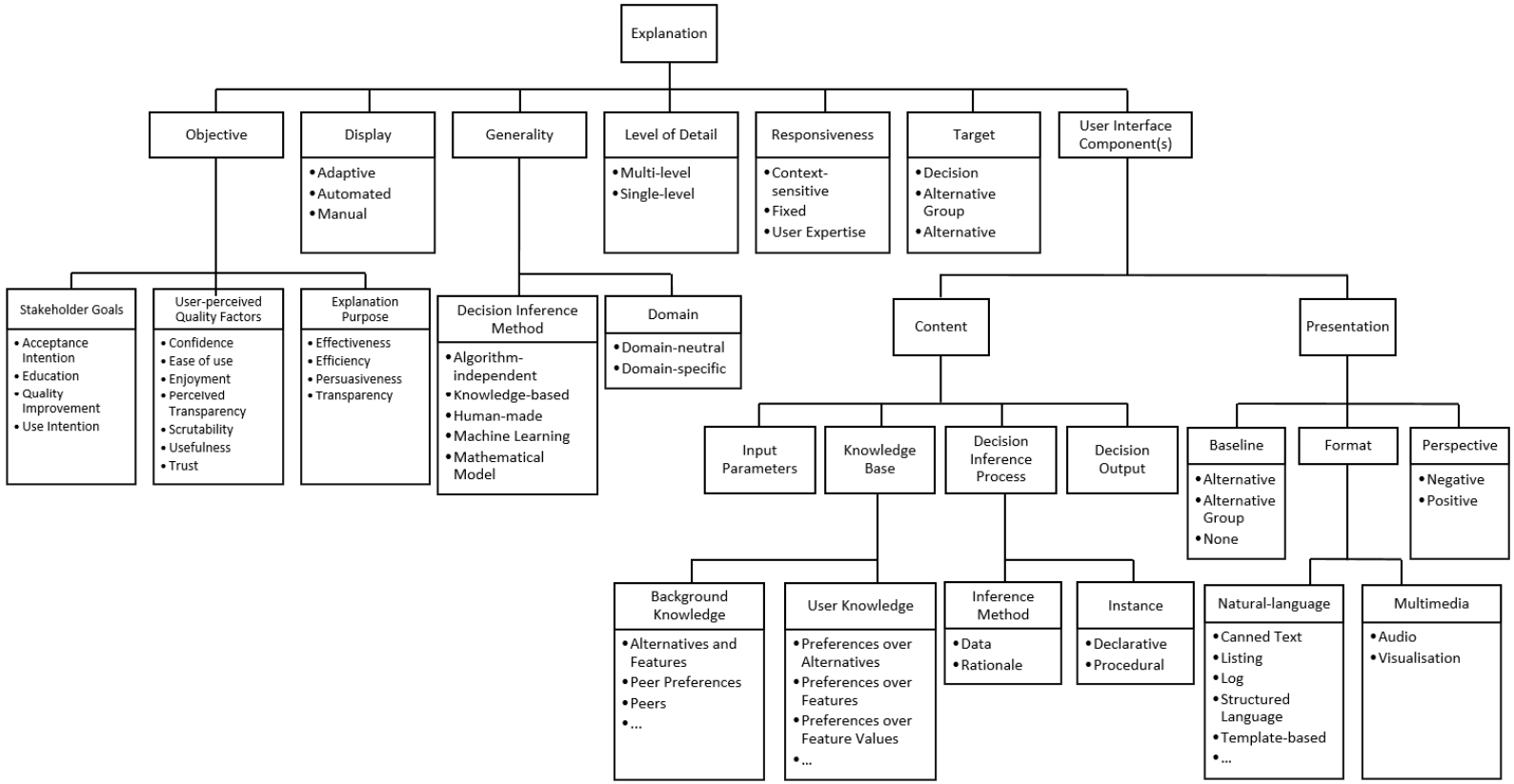
Another academic example



Industry example



Possible aspects to consider



Nunes, I. and Jannach, D.: *"A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems"*. User-Modeling and User-Adapted Interaction, Vol. 27(3-5). Springer, 2017, pp. 393-444

Summary of second part

- We reviewed the history of technical approaches to build recommenders
- 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

Part III: Measurements

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 Part I)
  - 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 with the system)
- 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

Cremonesi, P. and Jannach, D.: *"Progress in recommender systems research: Crisis? What crisis?"*. AI Magazine, Vol. 42(3). 2021, pp. 43-54

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

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

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)
- **"Data Set Selection"**
  - https://www.semanticscholar.org/paper/Data-Set-Selection-LaLoudouana/bb4d9c628314b650b1dab8afe06d02c0551ecc89
  - https://tinyurl.com/y5bov3pm

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 (e.g., using Docker) 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 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
  - 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

Accuracy, again

- 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

Multi-metric evaluation, again

- 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
  - An interesting area for future work

Summary

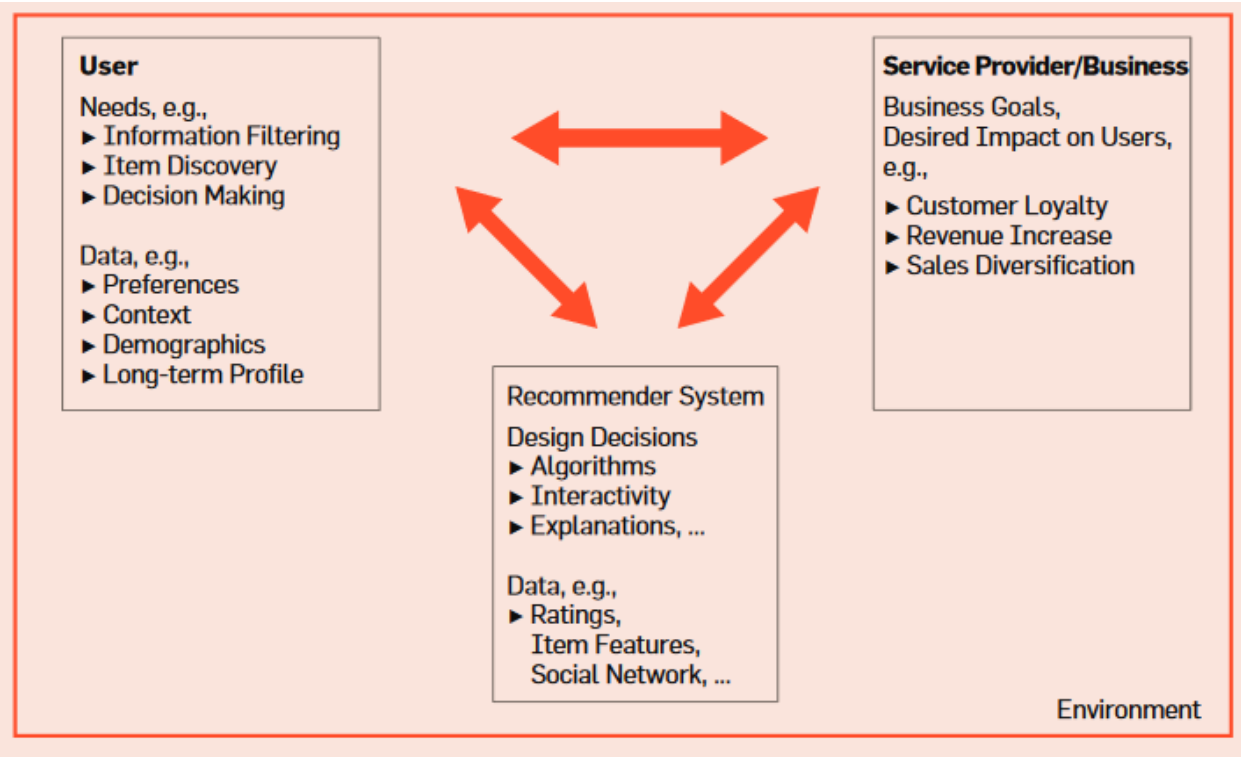
Possible steps forward

- 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
  - Considering a recommender system a sociotechnical system

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"

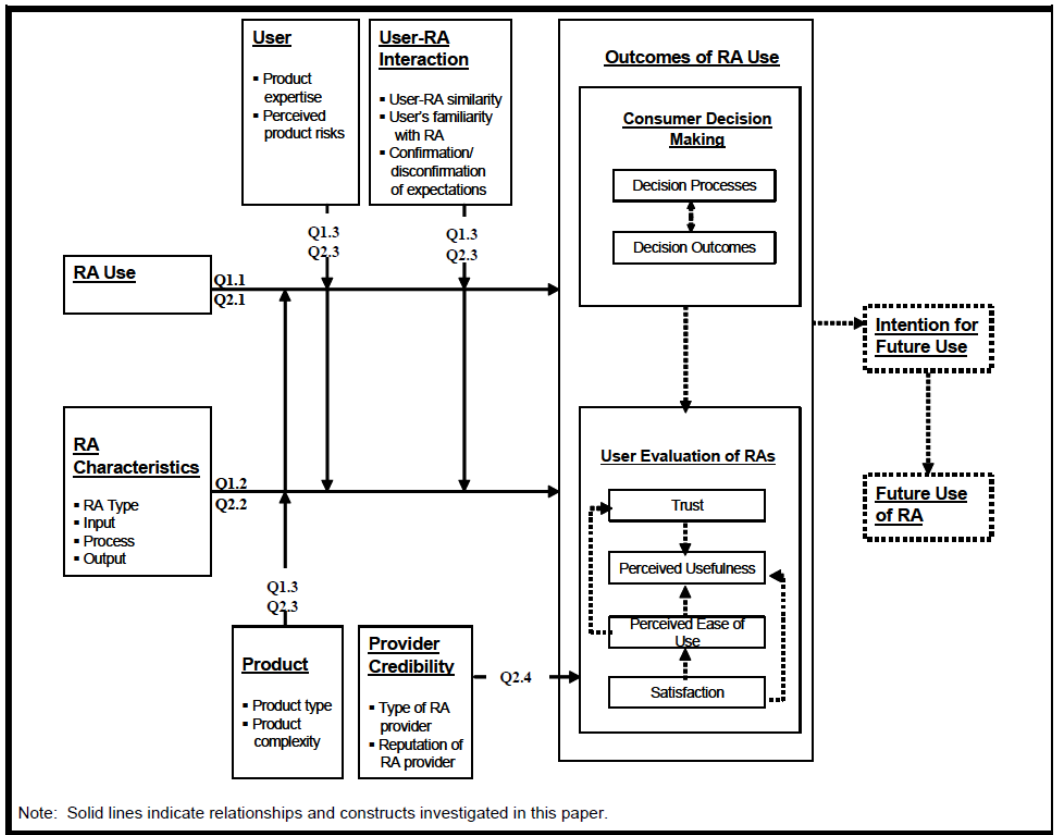


The Information Systems Perspective

- 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



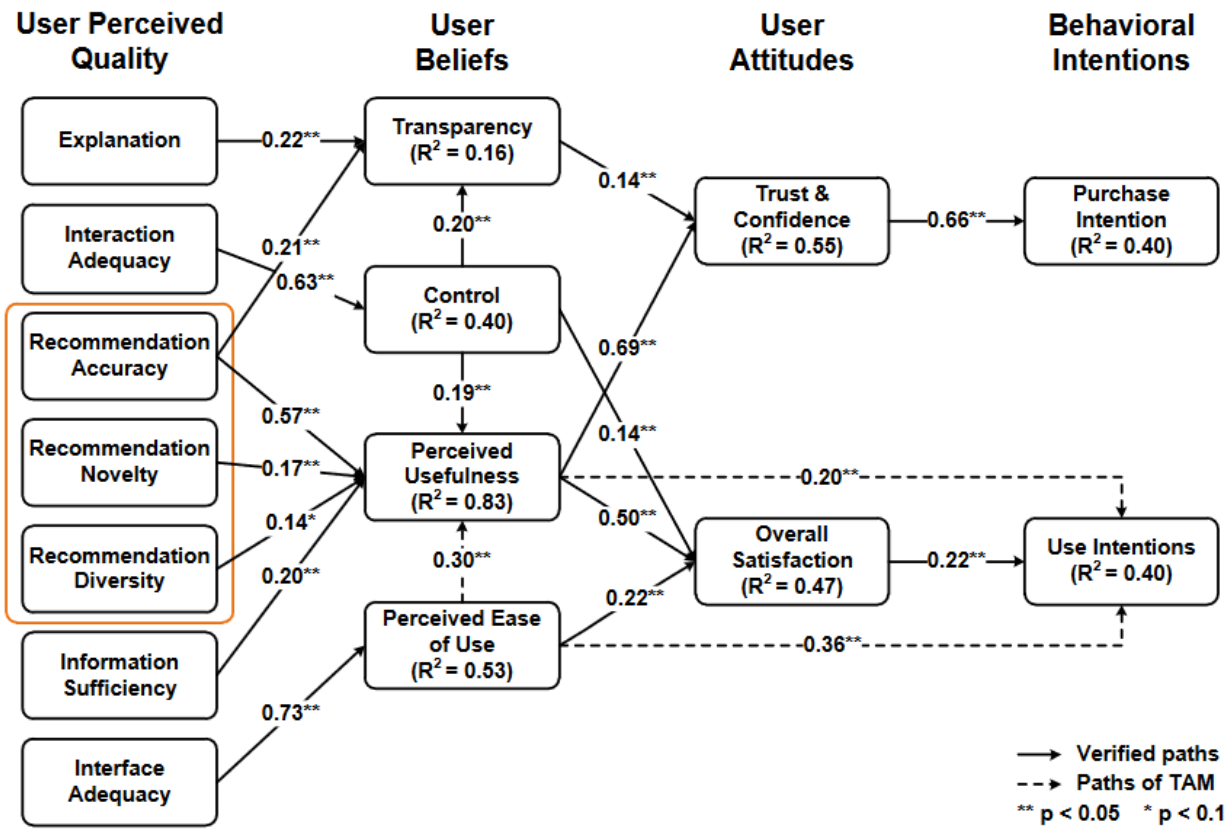
Note: Solid lines indicate relationships and constructs investigated in this paper.

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



Reflecting on user studies

- User studies are often considered difficult
- But they are necessary to understand the foundations
- Abstract computational measures might not correspond to user perceptions or business value

Summed up on Twitter

Twittern

Darren L Dahly
@statsepi

Two things that more than a few "experts" don't seem to get:

1. You can improve a useless prediction, even a lot, and still have a useless prediction.

2. To understand the utility of any prediction, you must understand the specific context where it will be deployed.

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

Trending topics

What's hot? (A subjective selection)

- Explanations (XAI) and Trustworthy Recommendation
- Fairness, Biases, and Responsible Recommendation
- Causality in Recommendations, Counterfactual Approaches, Off-policy Evaluation
- Conversational and Interactive Recommendation
- To some extent, still:
  - Session-based Recommendation, sequential recommendation
- What is dominating?
  - Deep learning for item ranking

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 Dacrema et al.**
  - RecSys 2019