

Recommender Systems: Value, Methods, Measurements

Dietmar Jannach, University of Klagenfurt, Austria

dietmar.jannach@aau.at

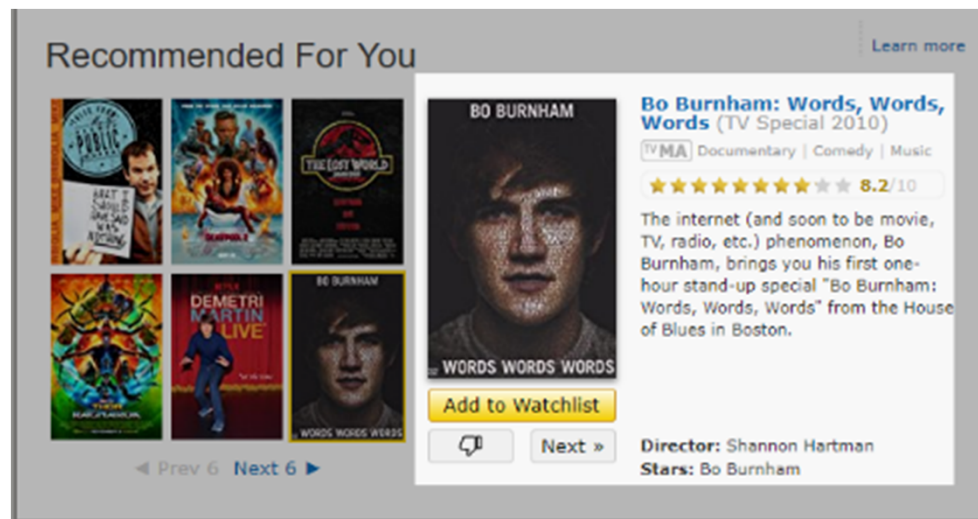
Presented at the ACM RecSys 2019 Summer School on Recommender Systems
Gothenburg, 2019

Outline

- What are Recommender Systems?
 - What is their value?
 - How do we build Recommender Systems?
 - How do we know they work well?
-
- (Pointers to other lectures in the summer school)

Recommender Systems

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



Applications

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

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**
 - See the later talk on multi-stakeholder recommendations

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

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

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.”

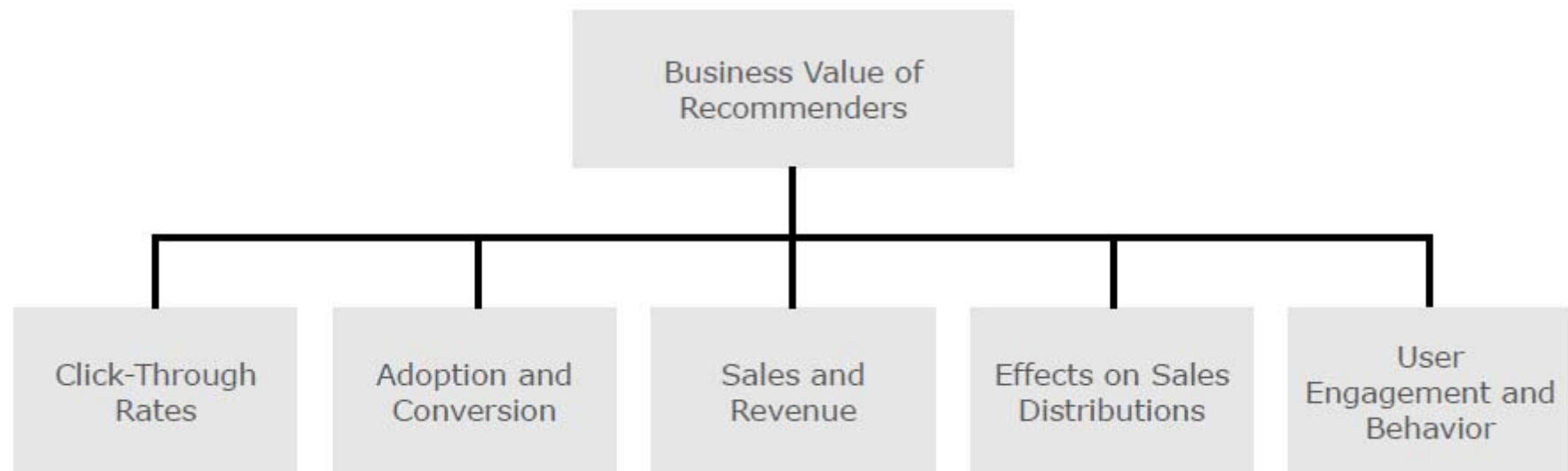
“Netflix says 80 percent of watched content is based on algorithmic recommendations”

Measuring the business value

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

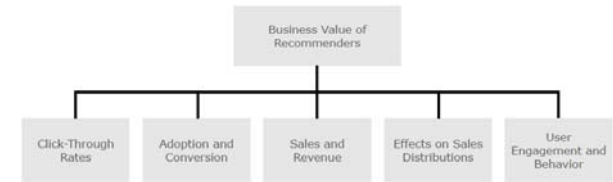
What is measured?

- Considering both the **impact** and **value** perspective



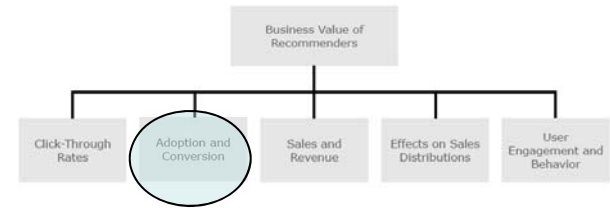
Jannach, D., Jugovac, M.; "Measuring the Business Value of Recommender Systems", arxiv preprint, <https://arxiv.org/pdf/1908.08328.pdf>

Click-Through Rates



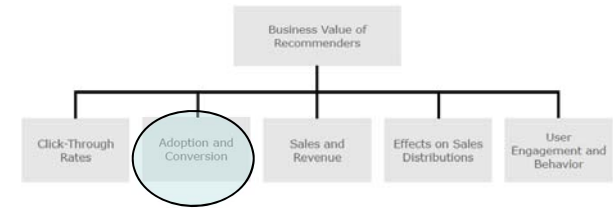
- Measures how many clicks are garnered by recommendations
 - Popular in the news recommendation domain
 - [Google News](#): 38% more clicks compared to popularity-based recommendations
 - [Forbes](#): 37% improvement through better algorithm compared to time-decayed popularity based method
 - [swissinfo.ch](#):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

Adoption and Conversion Rates



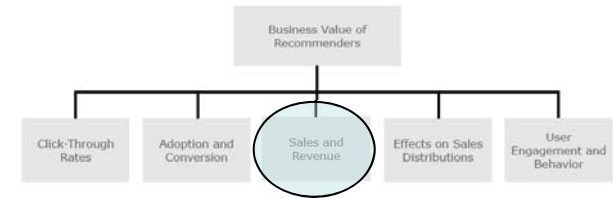
- CTR usually not the ultimate measure
 - Cannot know if users actually liked/purchased what they clicked on (consider also: click baits)
- Therefore
 - Various, domain-specific adoption measures common
- YouTube, Netflix: “Long CTR”/ “Take rate”
 - only count click if certain amount of video was watched

Adoption and Conversion Rates



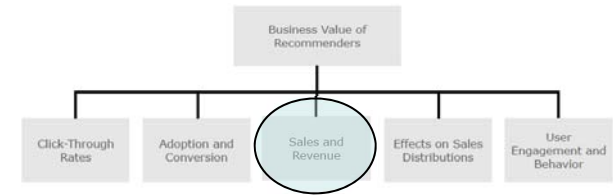
- Alternatives when items cannot be viewed/read:
- eBay:
 - “purchase-through-rate”, “bid-through-rate”
- Other:
 - LinkedIn: Contact with employer made
 - Paper recommendation: “link-through”, “cite-through”
 - E-Commerce marketplace: “click-outs”
 - Online dating: “open communications”, “positive contacts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

Sales and Revenue



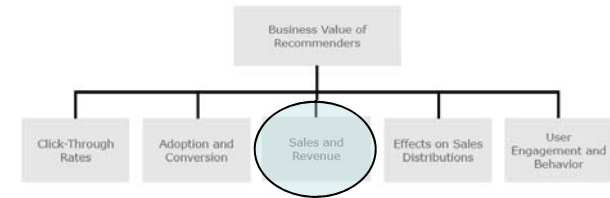
- CTR and adoption measures are good indicators of relevant recommendations
- However:
 - Often unclear how this translates into business value
 - Users might have bought an item anyway
 - Substantial increases might be not relevant for business when starting from a very low basis
- In addition:
 - Problem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



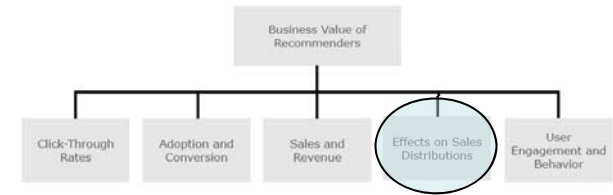
- e-grocery studies:
 - 1.8 % direct increase in sales in one study
 - 0.3 % direct effects in another study
 - However:
 - Up to 26% indirect effects, e.g., where customers were pointed to other categories in the store
 - “Inspirational” effect also observed in music recommendation in our own work
-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



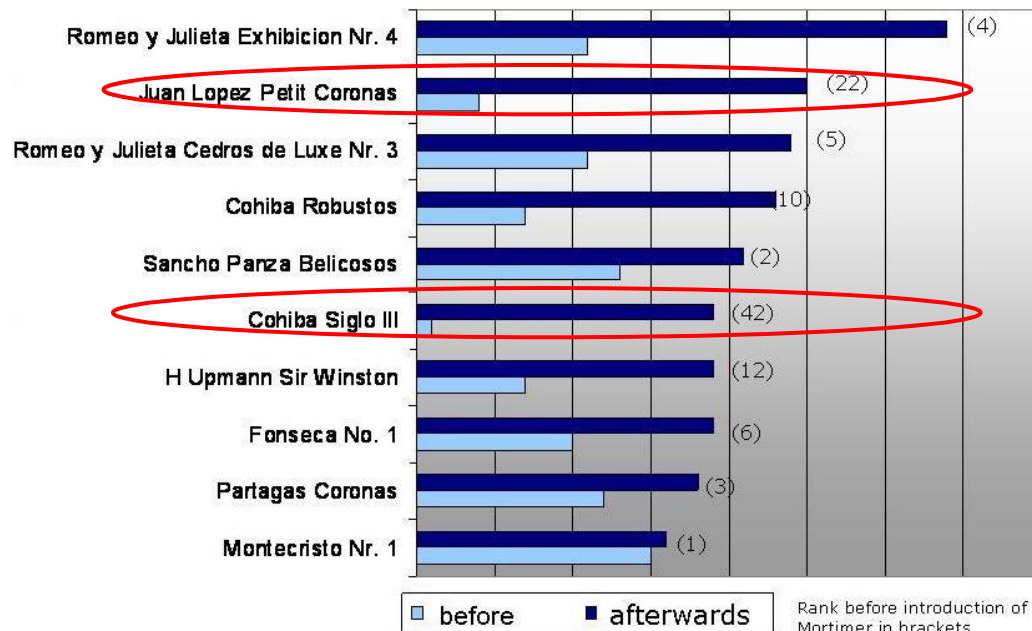
Effects on Sales Distributions



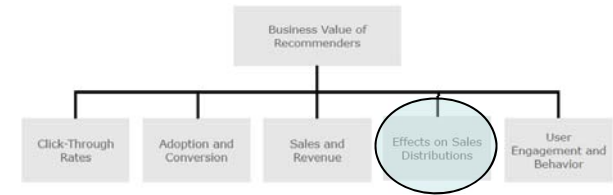
- Goal is maybe not to sell *more* but *different* items
- Influence sales behavior of customers
 - stimulate cross-sales
 - sell off on-stock items
 - promote items with higher margin
 - long-tail recommendations

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

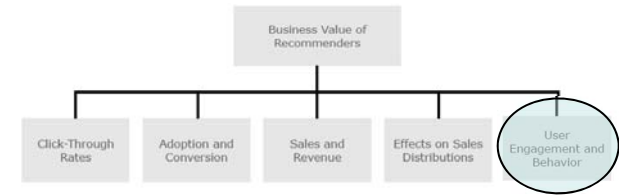


Effects on Sales Distributions



- 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

User Behavior and Engagement



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

Discussion

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

Discussion

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

What to measure?

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

What to measure?

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

The academic perspective

- In academia, we aim to
 - abstract from application specifics, and
 - develop generalizable methods
- Abstract computational tasks from the literature
 - Find all or some good items
 - Predict the relevance of unseen items
 - Recommend sequence
 - Just browsing

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

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	<ul style="list-style-type: none"> • Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space • ... 	<ul style="list-style-type: none"> • Change user behavior in desired directions • Create additional demand • Increase activity on the site • ...
Operational Perspective	System Task	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space, ... 	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space, ... 	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space, ... 	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space, ... 	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space, ... 	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space, ... 	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space, ... 	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Help users find objects that match the user's long-term preferences • Show alternatives • Help users explore or understand the item space, ... 	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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

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

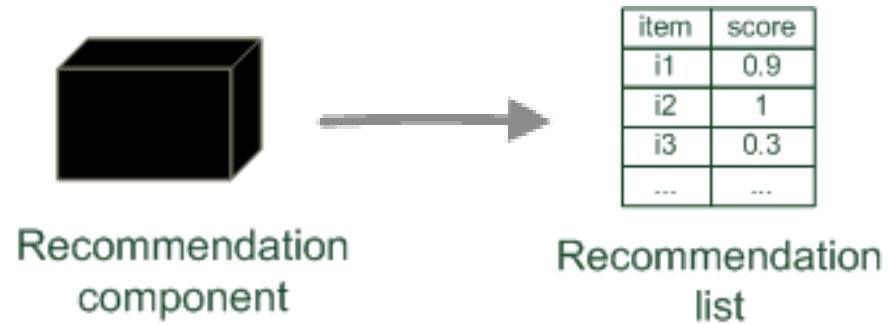
- Content-based Filtering
 - Collaborative Filtering
 - Hybrid Systems
 - Knowledge-based Systems
-
- Interactive Recommendation

Information Filtering roots

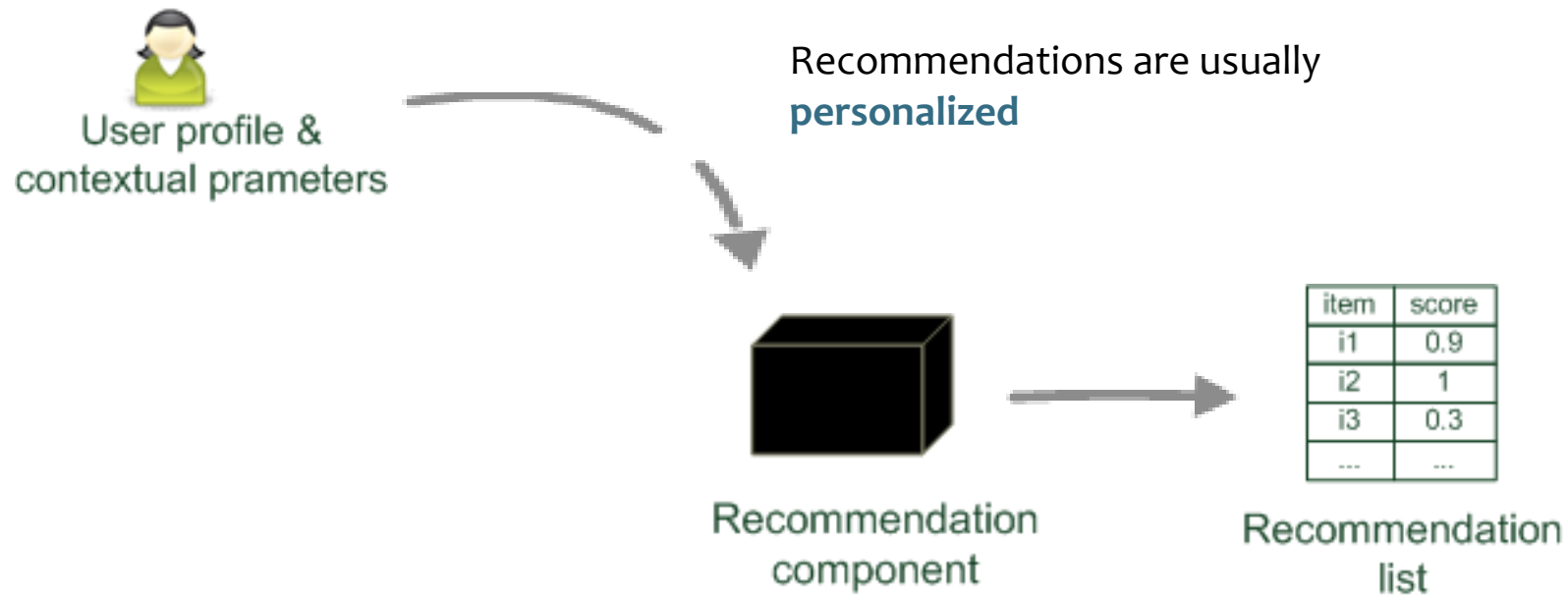
- Information Filtering
 - Systems that filter incoming streams of information in a personalized way
 - Dates back to the late 1960s
 - Early systems use explicitly stated preferences regarding topics or keywords
 - Later on, automated content analysis and user profiling
- Today:
 - “Content-based Filtering”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

Recommendation Principles

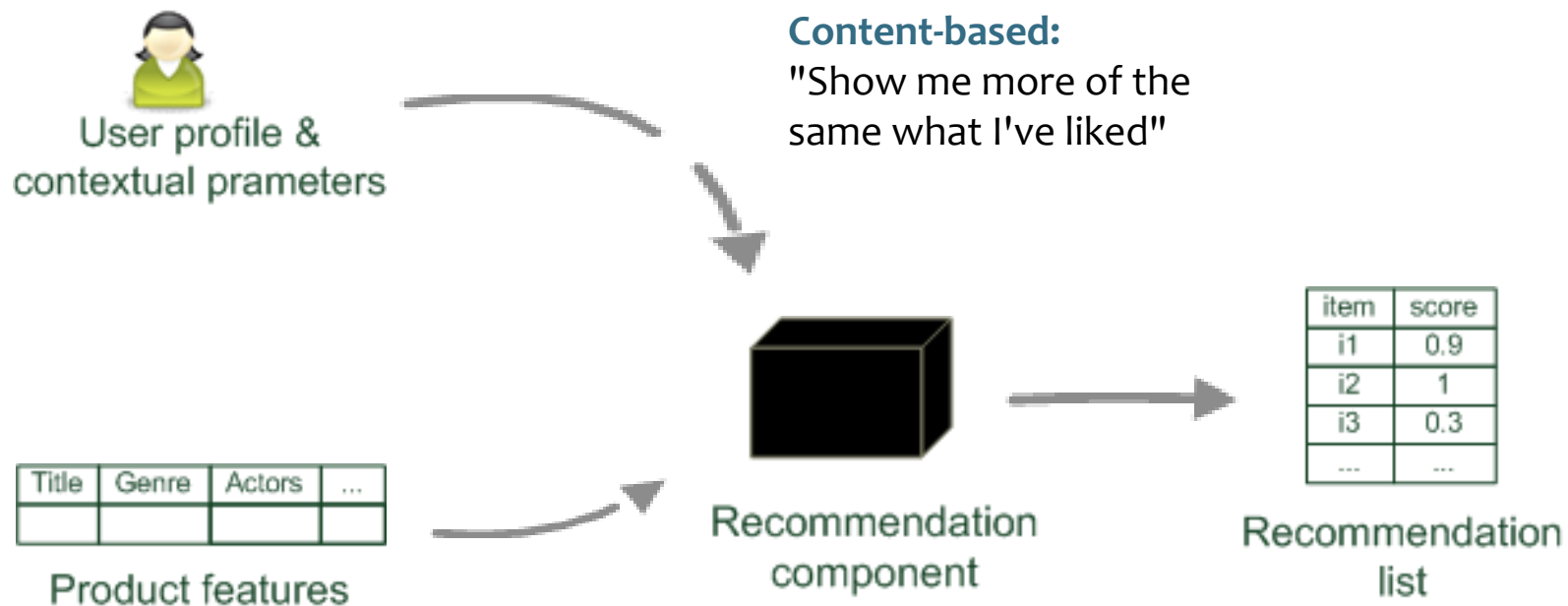
Recommender systems
reduce information
overload by estimating
relevance



Recommendation Principles



Content-based Filtering



- More details: lecture on content-based recommender systems in this summer school

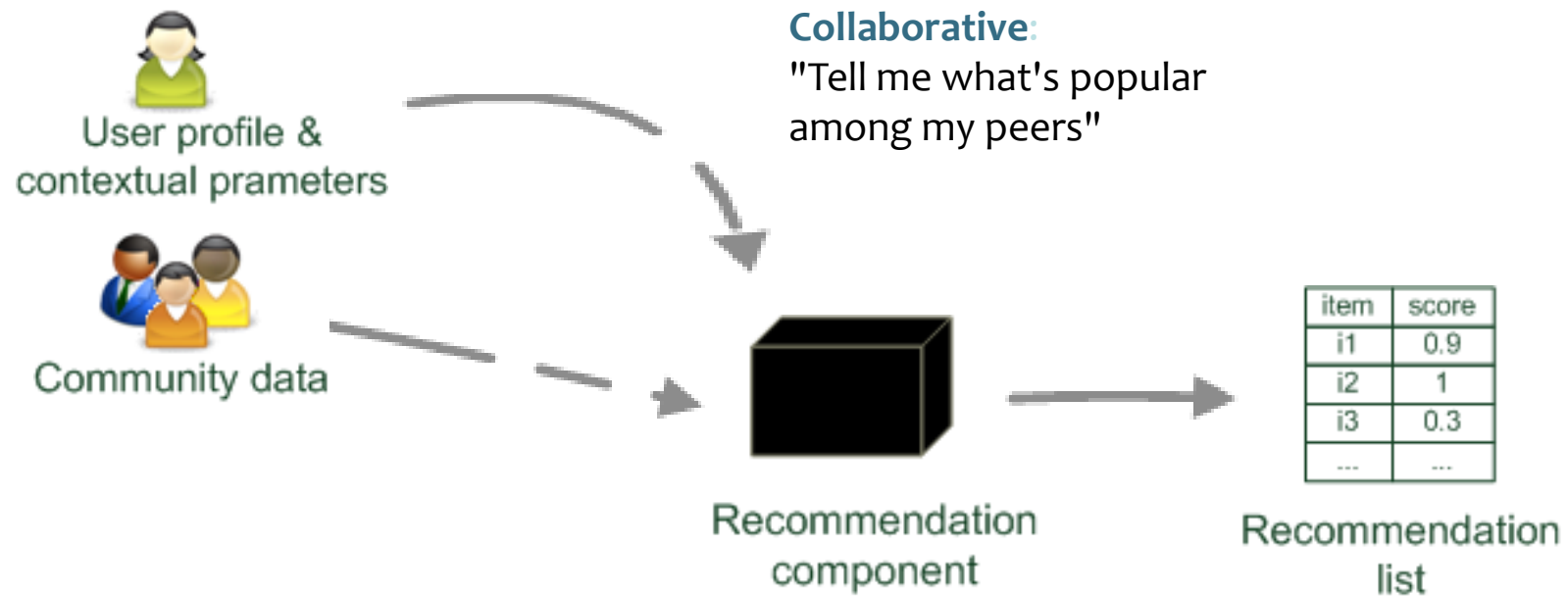
Outline

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

Leveraging the opinions of others

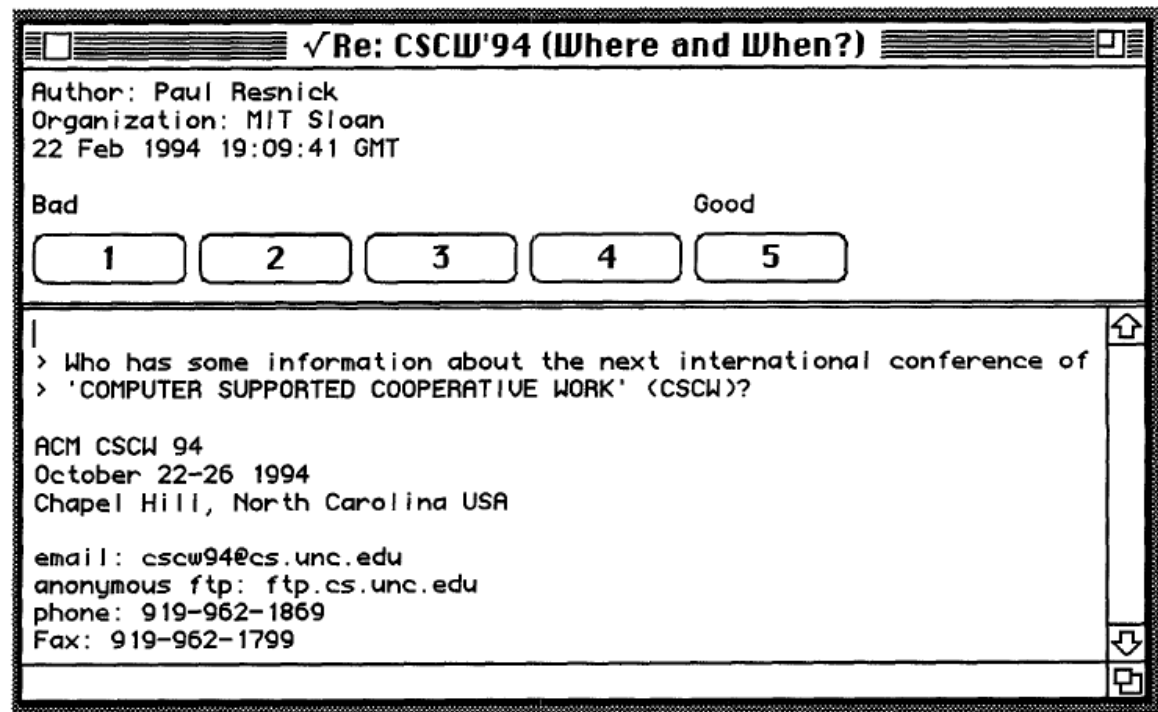
- 1982: ACM president complained about email junk
 - Envisioned a set of “trusted authorities” that assess the quality of the messages
- 1987: Information Lens
 - Based on manual filters, but could also specify people whose opinions they value
- 1992: Tapestry – “Collaborative Filtering”
 - Continued 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

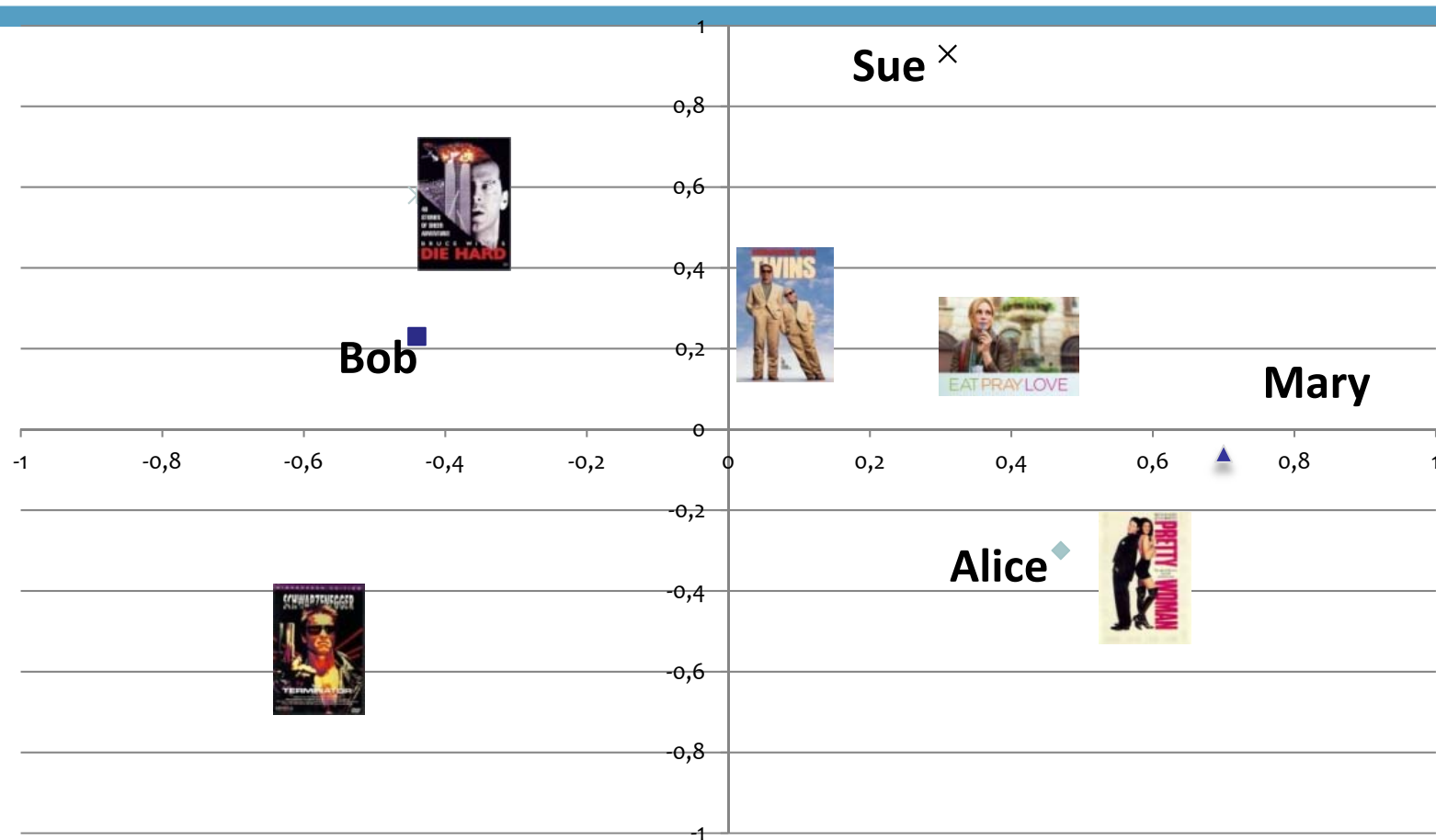
The Netflix Prize (2006-2009)

- 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

- 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



Matrix Factorization

- SVD: $M_k = U_k \times \Sigma_k \times V_k^T$

U_k	Dim1	Dim2
Alice	0.47	-0.30
Bob	-0.44	0.23
Mary	0.70	-0.06
Sue	0.31	0.93

V_k^T					
Dim1	-0.44	-0.57	0.06	0.38	0.57
Dim2	0.58	-0.66	0.26	0.18	-0.36

Σ_k	Dim1	Dim2
Dim1	5.63	0
Dim2	0	3.23

- Prediction: $\hat{r}_{ui} = \bar{r}_u + U_k(\text{Alice}) \times \Sigma_k \times V_k^T(\text{EPL})$
 $= 3 + 0.84 = 3.84$

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
 - More on this later in the summer school
- Recently, also 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**”)

Session-based Recommendation

- Guessing the intention can be difficult



The image shows a product listing for a Minnow Sports Aluminum Baseball Bat. On the left, there are five small thumbnail images showing different views of the bat. The main image shows the bat diagonally, with the Minnow Sports logo and 'Baseball Bat' text. Below the bat, it says '32" ▶ 24 oz'. To the right of the bat, the text 'MINNOW SPORTS' is displayed above a baseball icon. Further right, the product title 'Minnow Sports Aluminum Baseball Bat For Baseball & Teeball' is shown with a 4-star rating and '8 customer reviews'. Below the title, the price is listed as '\$29.99' with a sale price of '\$19.99', indicating a 33% discount. The text 'In Stock.' is followed by a note that the item does not ship to Germany. Below this, it says 'Sold by BBro Store and Fulfilled by Amazon. Gift-wrap available.' A dropdown menu for 'Item Display Length' is set to '32.0 inches'. At the bottom, a list of bullet points describes the bat's features: lightweight aluminum alloy, ultra-thin 32" handle with All Sports grip, stylish design with full rolled-over end, and suitability for all levels of players.

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.

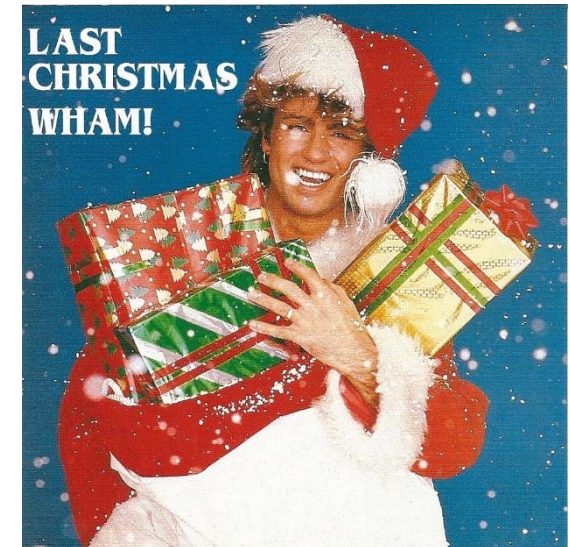
Item Display Length:
32.0 inches

- 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

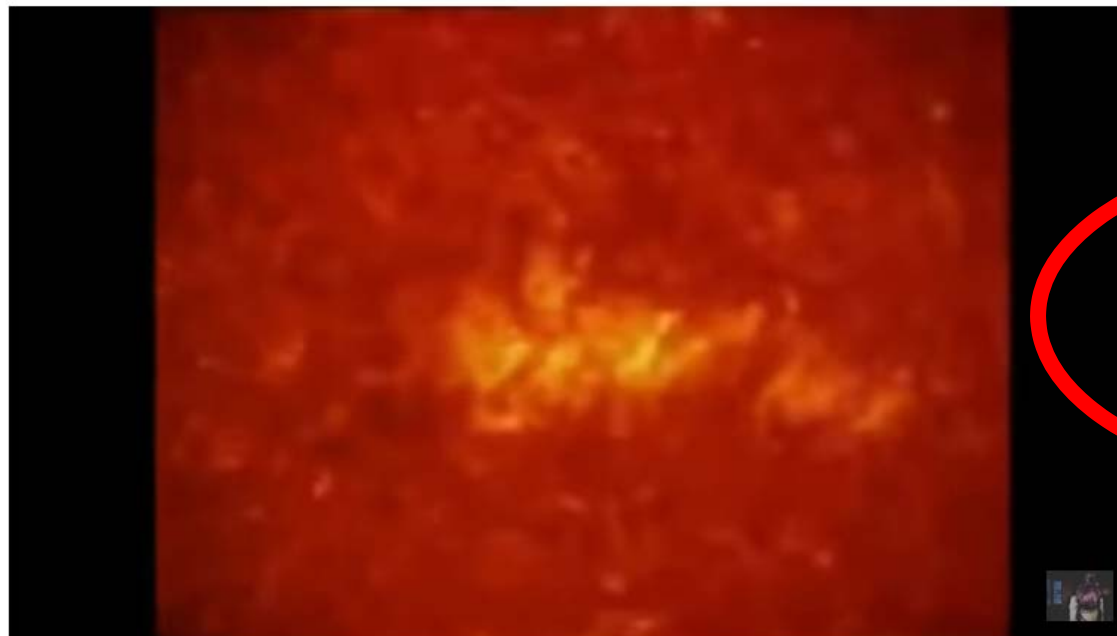


Session-based Recommendation

- 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?



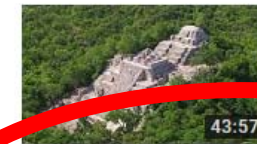
YouTube



DAS WELTALL Beste Doku über das Universum HD Doku

Nächstes Video

AUTOPLAY



Die Kokosinsel - Schatzinsel der Piraten [Doku]

DokuTV
637.100 Aufrufe

43:57



**Bob der Baumeister
Spielzeugautos, Bagger,**

Kinder Spielzeug Kanal
1,4 Mio. Aufrufe

26:36



**Die Kelten 1/3: Europas
vergessene Macht**

Stefan Nährlich
179.000 Aufrufe

52:39



**BLVD 7.0 - Erich von Däniken
im Gespräch mit Ken Jebsen**

KenFM
420.985 Aufrufe

1:36:25

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?

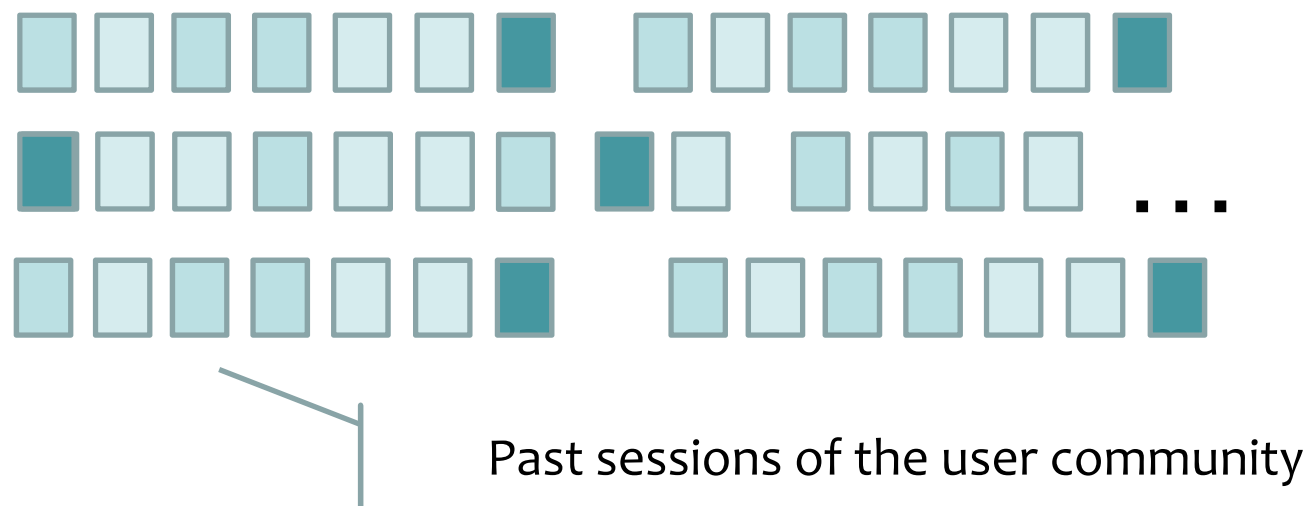
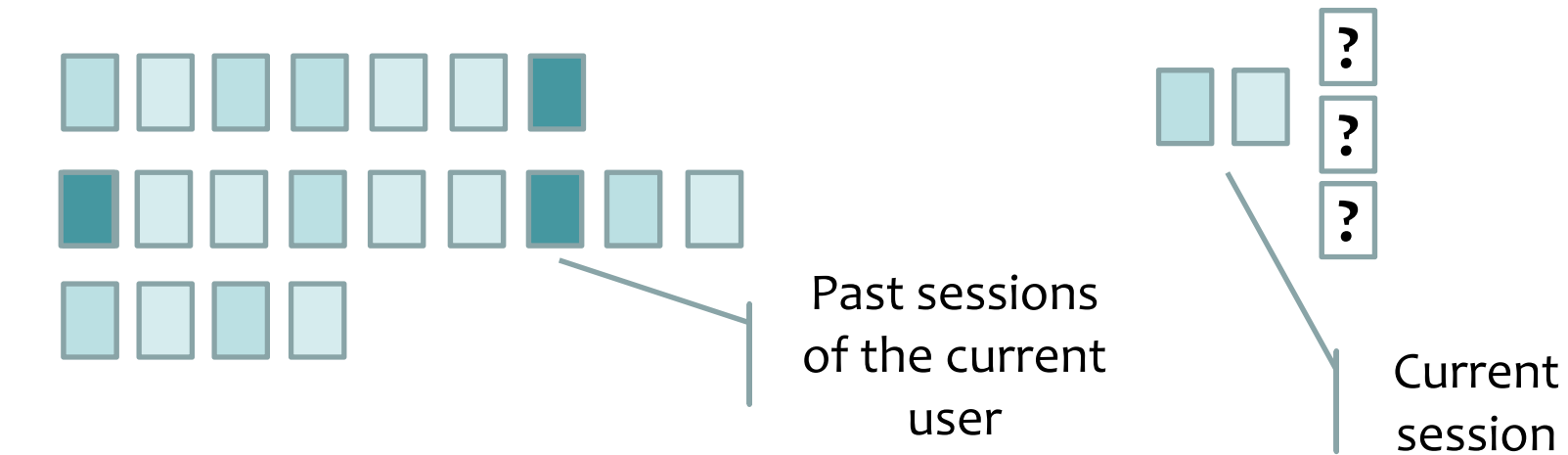
- Some plausible options
 - Only shoes or only watches?
 - Mostly Nike shoes?
 - Maybe also some T-shirts?
- Considerations and observations
 - Using the matrix completion formulation, the system will mostly recommend T-shirts and trousers
 - Research indicates that both models are relevant, but that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

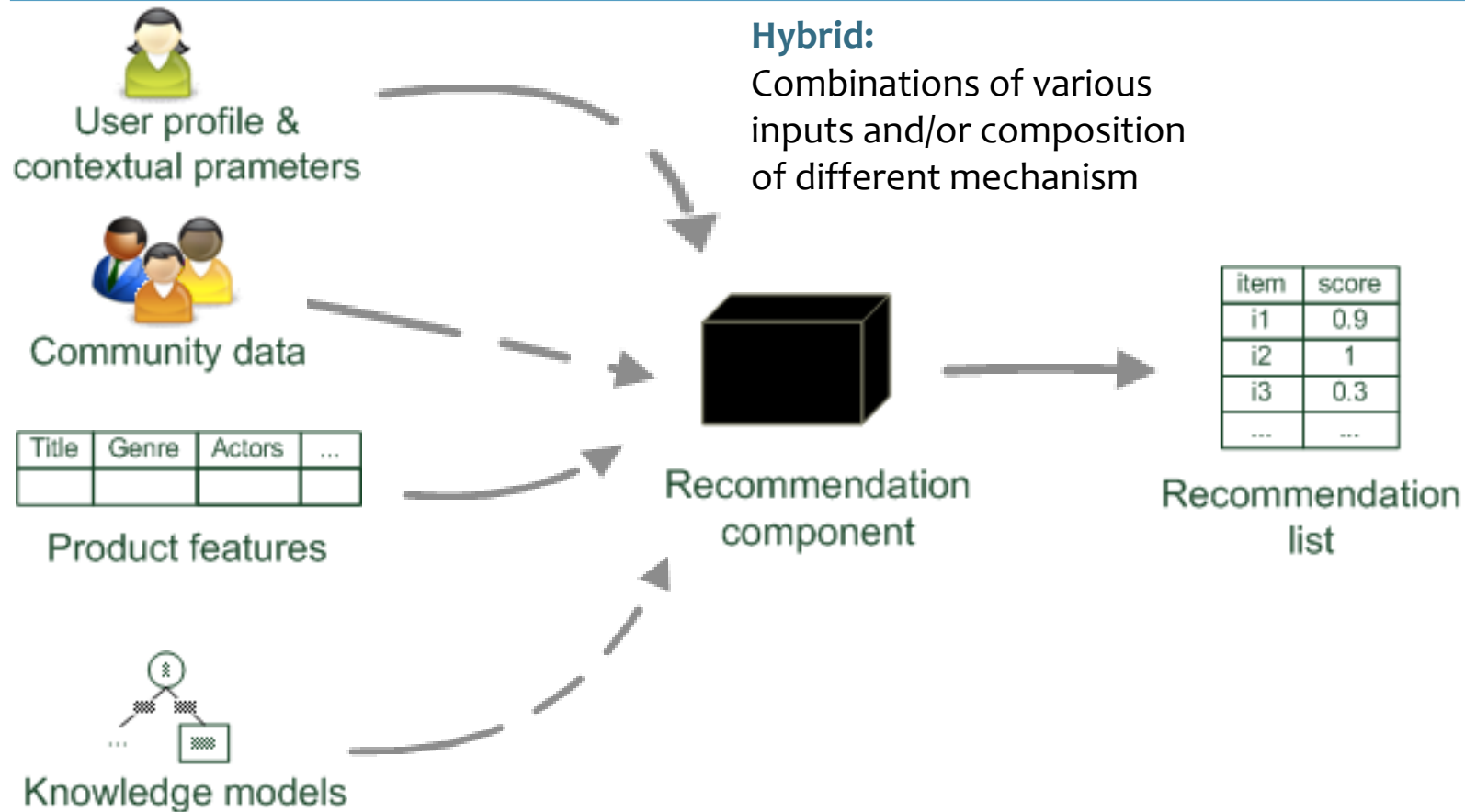
Outline

- 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



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

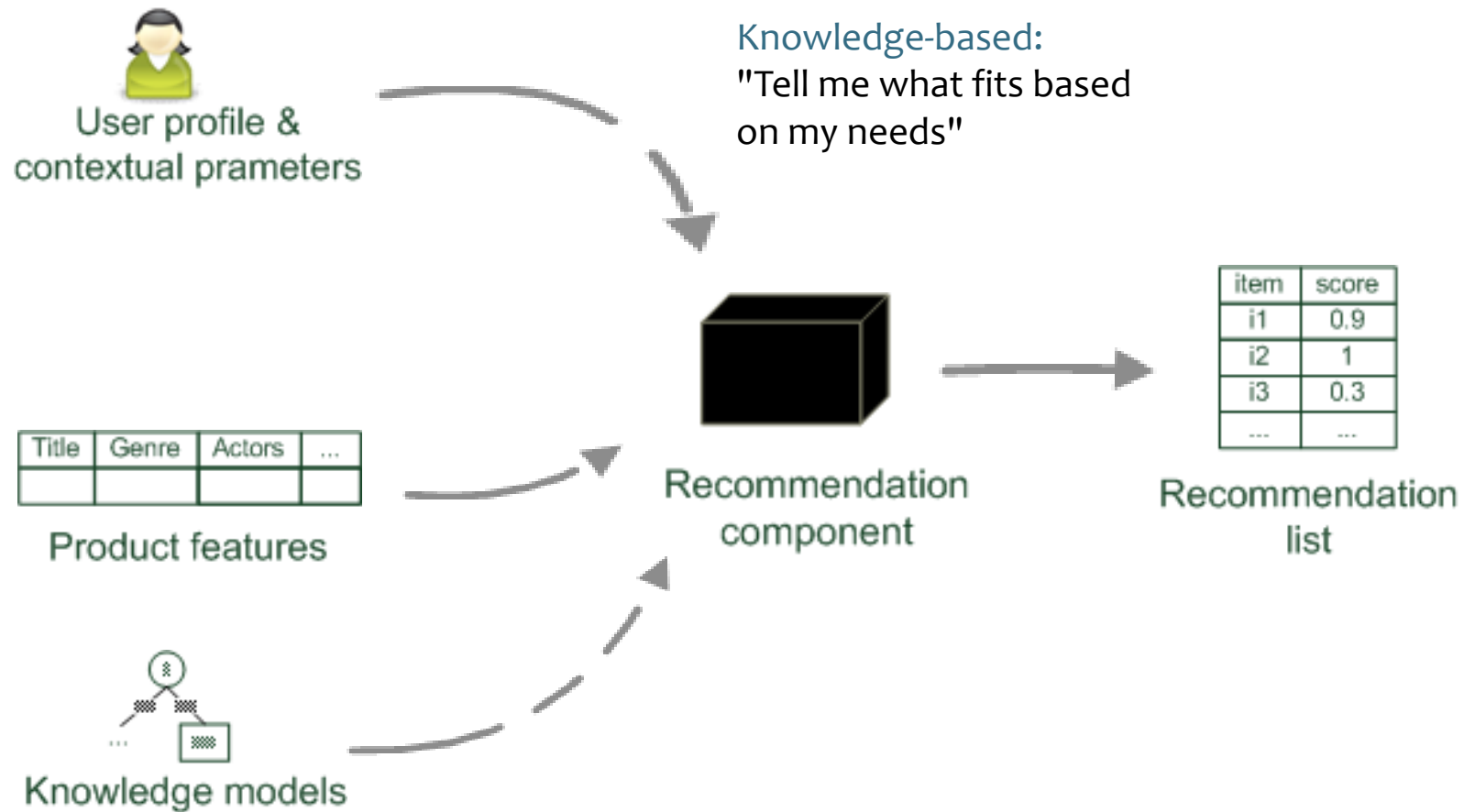
Similar Item Recommendations

- A common recommendation feature on many online platforms
 - Find similar movies, artists, recipes, ...
 - movies with same actors, from the same director etc.
- A relevant area for content-based techniques, but popularity of items is also important
- Another under-explored problem area
 - What determines the similarity of two items?
 - What makes a similar item recommendation useful?

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?

http://www.configworks-gmbh.online.de - VIBE - the virtual adviser for the Warmbad-Villach spa reso...

VIBE
VIRTUAL ADVISER

HOME CALL BACK SERVICE RECOMMENDATION



Mr Jannach, how do you feel right now? What would you like to improve if it were possible?

- ☐ I feel quite tired and would like to recharge my batteries
- ☒ I would like to improve my fitness.
- ☒ I would like to lose some weight and be slimmer.
- ☐ I often feel tense and sometimes have problems with my back.
- ☐ I would like to do something about my appearance and my image.
- ☐ I feel perfectly healthy and would simply like to relax for a few days.

Direct to result Back Next


Fertig

Is this even a recommender?

http://www.configworks-gmbh.online.de - VIBE - the virtual adviser for the Warmbad-Villach spa reso...

VIBE
VIRTUAL ADVISER

Did you know that...



Wonderful, we've now got to your final selection. Here's my recommendation for you ...

❖ **Feel well week**

Length of stay:	per week (7 nights) per person
Meals:	Half board
Accommodation:	The Warmbaderhof
Dates:	At any season
Rate in single room:	from € 1595
Rate in double room:	from € 1595

[Details](#)
[Why?](#)

I can also recommend the following packages:

- You can book a personal massage or a whole massage programme for your stay at any time.

❖ **Golf & Spa**

Length of stay:	per week (7 nights) per person
Meals:	Half board
Accommodation:	The Warmbaderhof
Dates:	01.04.2008-31.10.2008

[Details](#)
[Why?](#)

[Back](#) [Restart](#) [Print](#) [Online-request](#)

Fertig

Is this even a recommender?

The screenshot shows a web browser window with the address bar displaying <http://www.configworks-gmbh.online.de> and the page title "VIBE - the virtual adviser for the Warmbad-Villach spa reso...". The page header includes the "VIBE VIRTUAL ADVISER" logo and navigation links for "HOME", "CALL BACK SERVICE", and "RECOMMENDATIONS".

On the left side, there is a photograph of a woman in a red dress pointing upwards. A speech bubble next to her contains the text: "You're bound to ask yourself why I recommended the following. I'll be happy to explain..."

The main content area is titled "My arguments specially for you." and contains a list of arguments:

- I am happy to have found autumn packages for you, as you wished. If you want more suggestions for a specific date, you'll have to use the detailed advice option (more questions).
- We have a whole range at the Warmbad-Villach spa resort to suit your request Leisure and activities programme & Long walks. Ask about them.
- Our comprehensive supporting programme of cultural events (Carinthian Summer Music Festival, Villach Carnival, exhibitions at the Warmbad culture club, Jazz Over Villach, etc.) all year round and attractions in the vicinity will round off your stay at the
- Do you want to feel fit and healthy? Our sports and activities programmes respond to your wishes

At the bottom right of the content area is a "Back" button. The bottom status bar of the browser window shows the word "Fertig" on the left and a green checkmark icon on the right.

Outline

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

From Algorithms to User Experience

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

Konstan, J.A. & Riedl, J.. “Recommender systems: from algorithms to user experience”
User Model User-Adap Inter (2012) 22: 101.

Garcin, F., Faltings, B., Donatsch, O., Alazzawi, A., Bruttin, C., and Huber, A. 2014. Offline and online evaluation of news recommender systems at swissinfo.ch. In *Proceedings of the 8th ACM Conference on Recommender systems (RecSys '14)*.

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

Customers Who Bought This Item Also Bought



[Star Wars Trilogy Episodes I-III \(Blu-ray + DVD\)](#)

Hayden Christiansen

★★★★☆ 2,042

Blu-ray

\$34.96 ✓Prime



Star Wars: The Force Awakens (Blu-ray/DVD/Digital HD)

Harrison Ford

★★★★☆ 10,002

Blu-ray

\$24.41 ✓Prime



Star Wars: Episode I - The Phantom Menace (Widescreen Edition)

Ewan McGregor

★★★★☆ 3,533

DVD

\$53.24 ✓Prime



[Harry Potter: Complete 8-Film Collection \[Blu-ray\]](#)

Daniel Radcliffe

★★★★☆ 6,945

Blu-ray

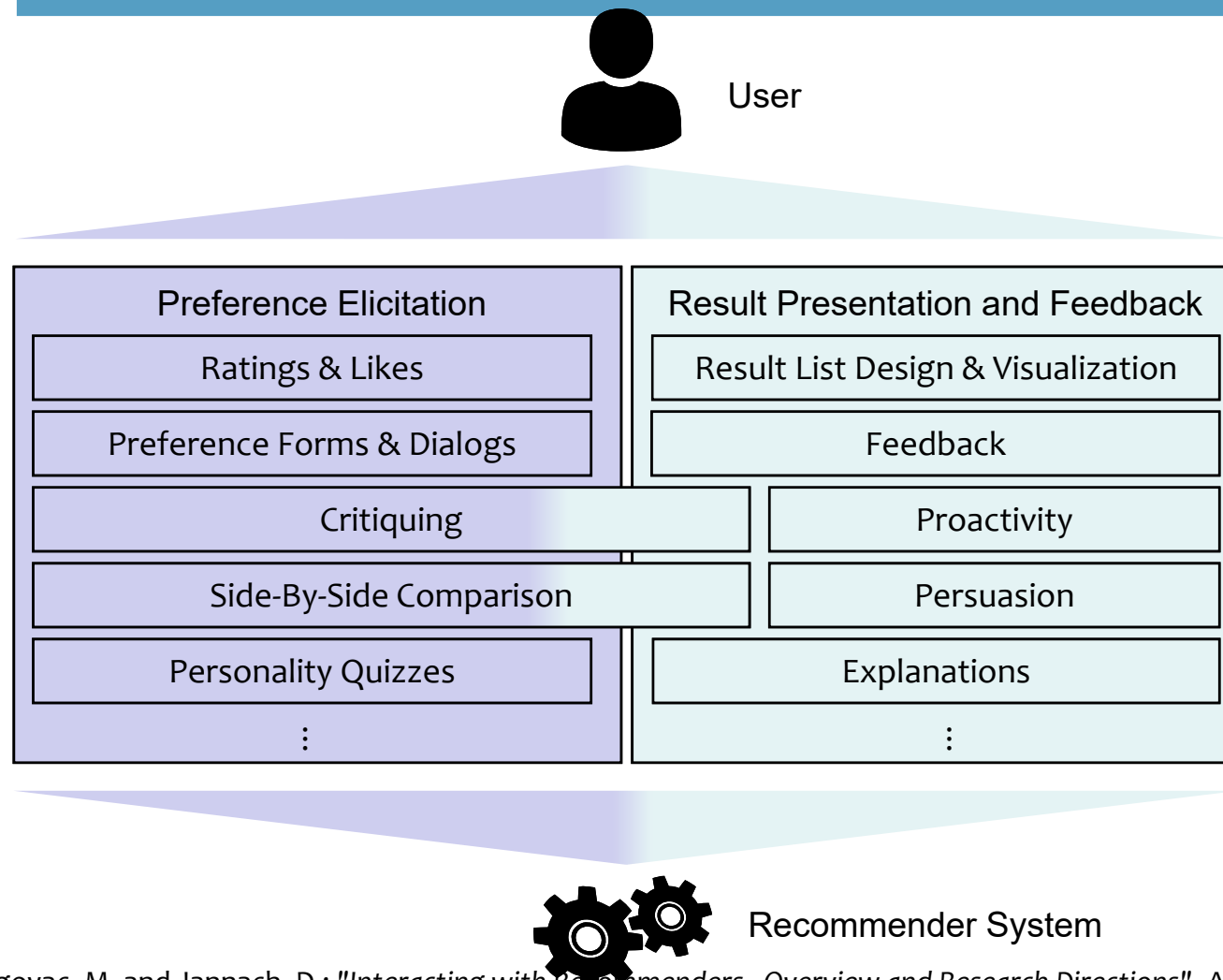
\$65.00 ✓Prime

Source: Amazon.com

UI research for Recommenders

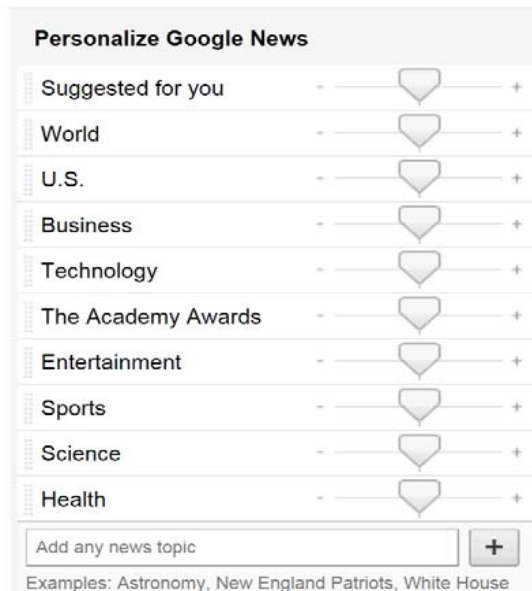
- HCI research is one of the main roots of recommender systems research
- Nonetheless, UI-related aspects seem less explored than algorithmic questions
 - One reason lies in the difficulty of evaluating new proposals
 - Existing research is also largely scattered

Structuring Existing Works



Design Space Examples

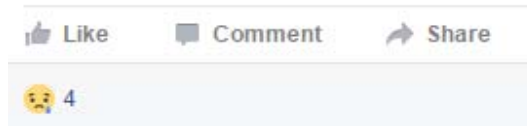
- Telling the system **explicitly** what you like
 - Global settings
 - Ratings
 - But how many options? How many categories?



Rate this item

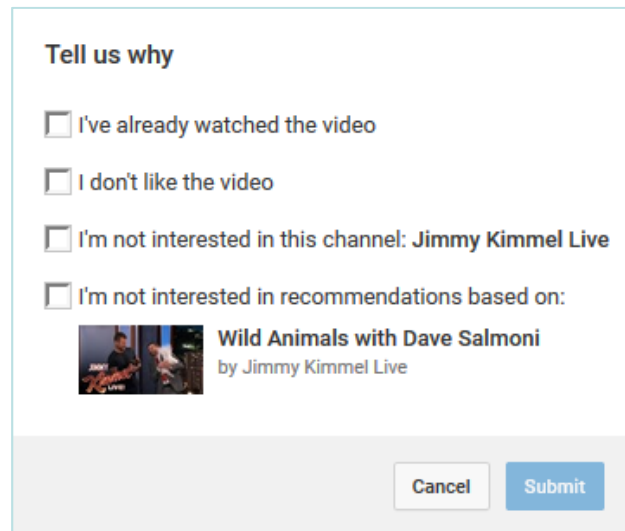


Sources: Facebook.com,
Google.com




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?



Tell us why

- ☐ I've already watched the video
- ☐ I don't like the video
- ☐ I'm not interested in this channel: Jimmy Kimmel Live
- ☐ I'm not interested in recommendations based on:
 **Wild Animals with Dave Salmoni**
by Jimmy Kimmel Live

Cancel Submit

Source: Youtube.com

List Design Considerations

Customers Who Bought This Item Also Bought

The screenshot displays three product recommendations from Amazon. Each item is shown with its image, title, star rating, number of reviews, price, and Prime eligibility. Annotations with leader lines point to specific elements: 'List label' points to the section header, 'Item description' points to the product title, 'Community rating' points to the star rating and review count, 'Highlighting' points to the 'Best Seller' badge, and 'Number of options' points to the entire list of items.

Item	Image	Description	Rating	Reviews	Price	Prime
Nikon AF-S FX NIKKOR 50mm f/1.8G Lens with Auto Focus for Nikon DSLR Cameras		Nikon AF-S FX NIKKOR 50mm f/1.8G Lens with Auto Focus for Nikon DSLR Cameras	★★★★★	1,505	\$216.95	✓Prime
LowePro Adventura 140 Camera Shoulder Bag for DSLR or Camcorder		LowePro Adventura 140 Camera Shoulder Bag for DSLR or Camcorder	★★★★☆	178	\$26.99	✓Prime
Lexar Professional 633x 64GB SDXC UHS-I Card w/Image Rescue 5 Software...		Lexar Professional 633x 64GB SDXC UHS-I Card w/Image Rescue 5 Software...	★★★★★	592	\$22.50	✓Prime

Source: amazon.com

Annotations:

- List label
- Item description
- Community rating
- Highlighting
- Number of options

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?

Recommended for you



Guardians of the Galaxy [Blu-ray]

Blu-ray ~ Chris Pratt (8 Jan 2015)

In stock

Price: EUR 9,99


73 used & new from EUR 8,75

Rate this item



☐ I own it

☐ Not interested

 Add to Cart

 Add to Wish List

Because you purchased...



Mad Max: Fury Road [Blu-ray] (Blu-ray)

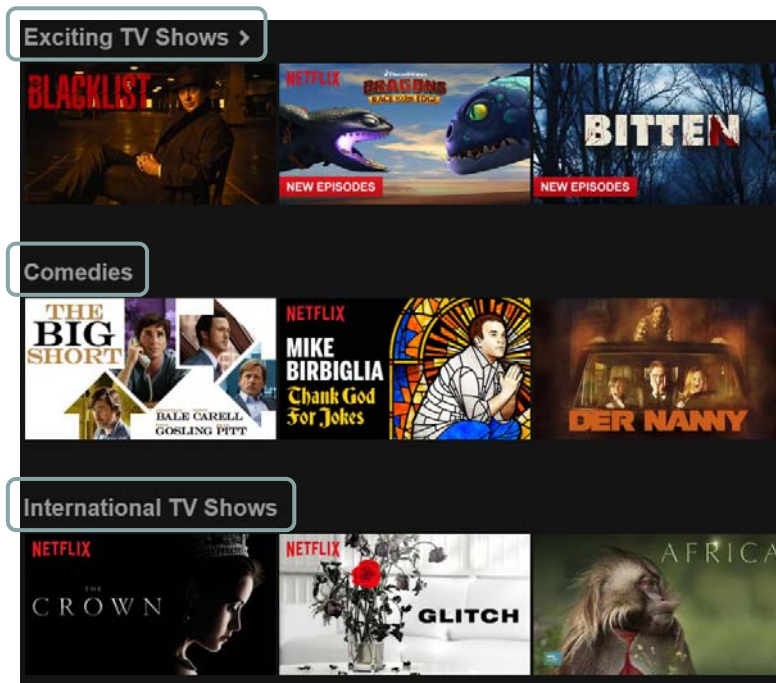
DVD ~ Charlize Theron



☐ Don't use for recommendations

List Content

- Grouping items in a list
Semantically



Source: netflix.com

Statistically

Frequently Bought Together



Customers Who Bought This Item Also Bought



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?

Twitter Messages

Hashtags
Loaded top 100 of 2560
Sorted by frequency

Sources
Loaded top 100 of 11103
Sorted by frequency

Messages
Loaded top 100 of 22589
Sorted by Time

Anomaly Detector

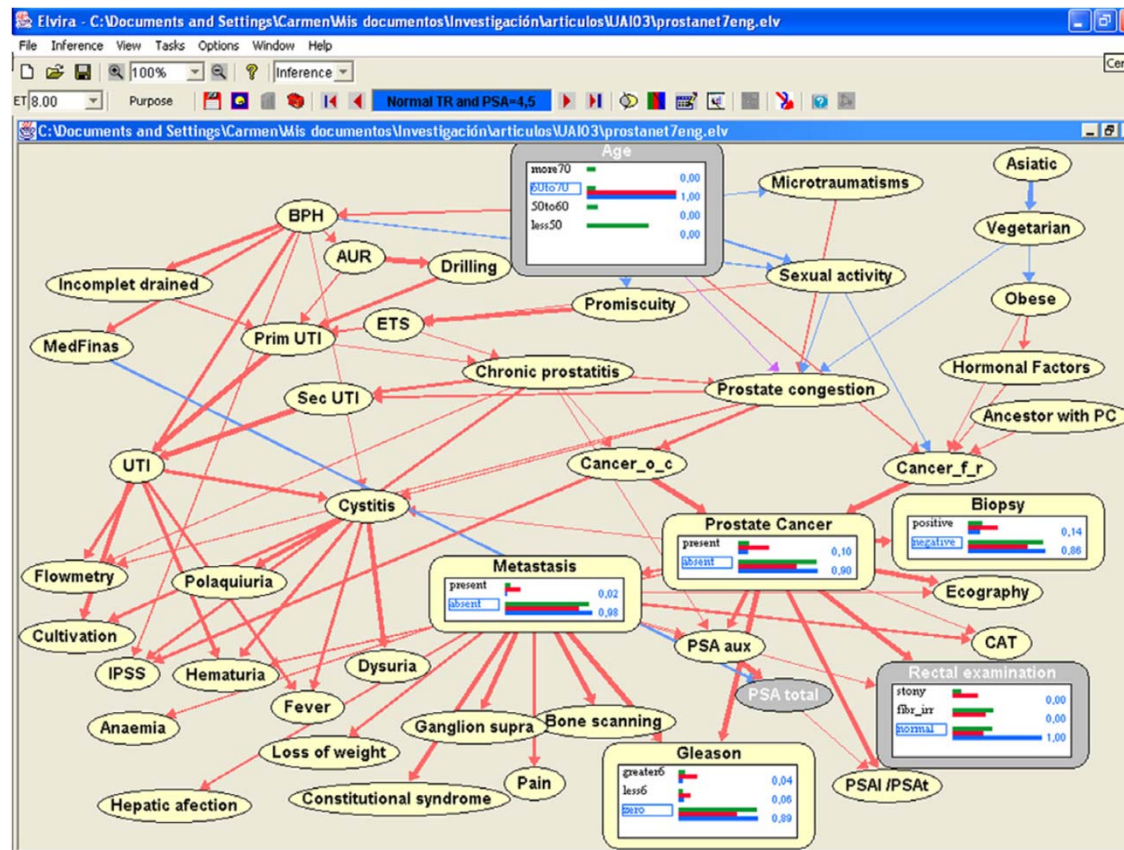
Topics
Loaded top 100 of 3974
Sorted by frequency

Time Chunks
Loaded top 43 of 43
Sorted by name

Anomaly Reports
Loaded top 100 of 4145
Sorted by score

The diagram illustrates the workflow of an Anomaly Detector. It starts with a 'Twitter Messages' panel on the left, which contains 'Hashtags' and 'Sources' sections. The 'Messages' section shows a list of tweets with a time axis. Arrows point from the 'Messages' section to an 'Anomaly Detector' panel on the right. The 'Anomaly Detector' has three sections: 'Topics', 'Time Chunks', and 'Anomaly Reports'. The 'Topics' section shows a list of topics. The 'Time Chunks' section shows a list of time chunks. The 'Anomaly Reports' section shows a list of reports. The diagram illustrates how tweets are processed into topics and time chunks, which are then used to generate anomaly reports.

Another academic example



Industry example

The screenshot shows the TripAdvisor page for the Clontarf Castle Hotel. The page features a green header with the TripAdvisor logo and navigation links. The main content area includes the hotel's name, a 4.5-star rating based on 1,985 reviews, and a ranking of #20 out of 174 hotels in Dublin. A 'Certificate of Excellence' badge is also displayed. The address is listed as Castle Avenue, Clontarf, Dublin D3, Ireland. A sidebar on the left provides a comparison of hotel amenities, showing that the hotel is better than 60% of alternatives for Bar/Lounge, 90% for Free Parking, and 70% for Restaurant. It also lists reasons to avoid the hotel, such as Airport Transportation and Leisure Centre. The main image shows the hotel's exterior at night, with a quote from a traveler: "... bar with a great atmosphere ...", "Enjoy a drink in the lovely relaxing lounge", "...don't miss the music in the bar area ...". On the right, there are links to traveler photos (1224), professional photos, and a map to browse nearby locations. At the bottom, there are tags for Family-friendly, Luxury, Best Value, and Free Wifi.

tripadvisor IRELAND Clontarf Castle Hotel Reviews, Dublin

Hi, Barry EUR

Dublin Hotels Flights Holiday Rentals Restaurants Things to Do Best of 2015 Your Friends More Write a Review

Europe Ireland Province of Leinster County Dublin Dublin Dublin Hotels Search for a city, hotel, etc.

Clontarf Castle Hotel

1,985 Reviews #20 of 174 Hotels in Dublin Certificate of Excellence

Castle Avenue | Clontarf, Dublin D3, Ireland Hotel amenities

Reasons for you to choose this hotel:

- Bar/Lounge (better than 60% of alternatives)
- Free Parking (better than 90% of alternatives)
- Restaurant (better than 70% of alternatives)

Reasons for you to avoid this hotel:

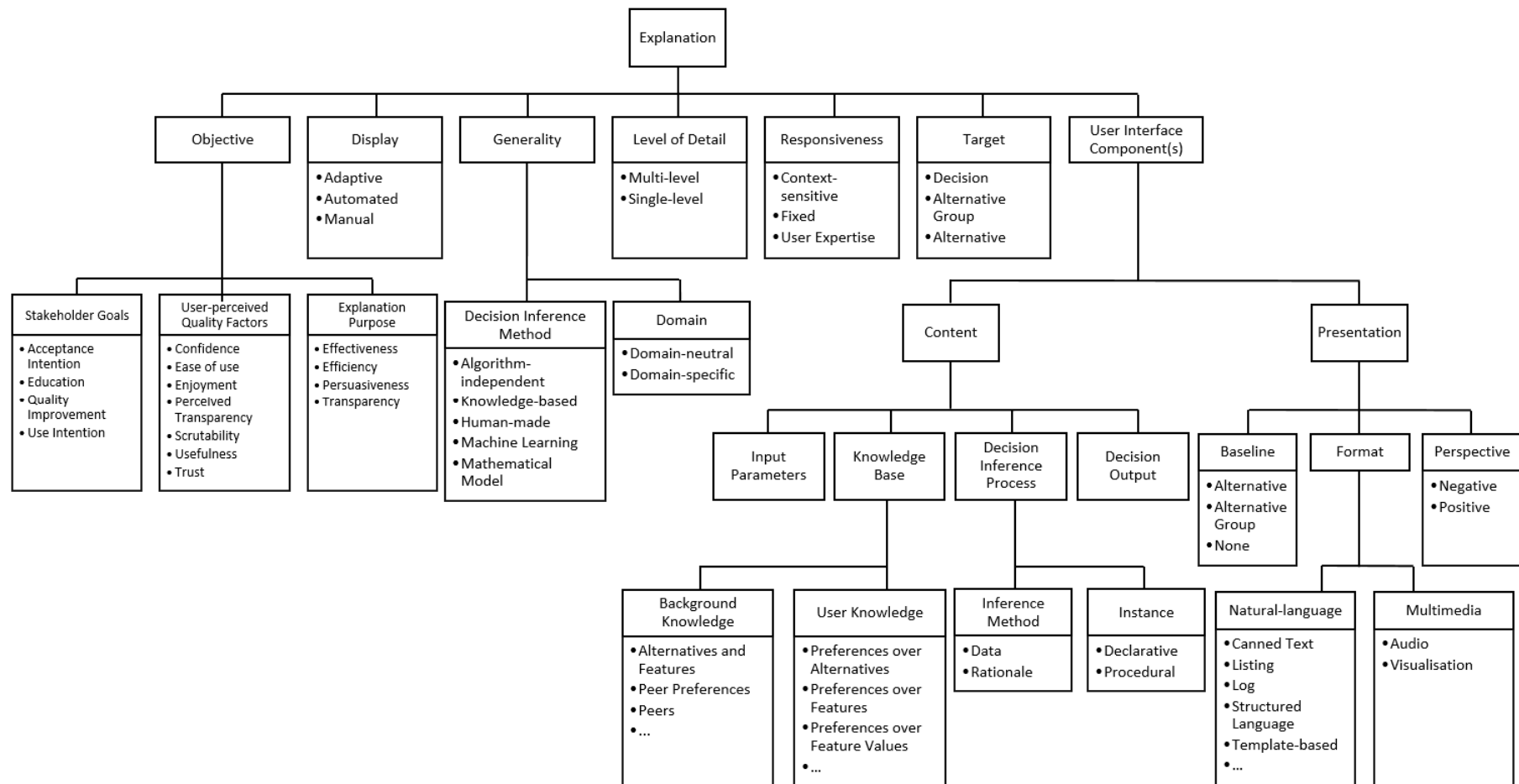
- Airport Transportation (worse than 90% of alternatives)
- Leisure Centre (worse than 75% of alternatives)

This explanation has been generated based on things that matter to you. Click here to see additional features.

★★★★☆ Family-friendly Luxury Best Value Free Wifi

Traveller photos 1224 Professional photos Browse nearby

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), see later lecture
- Offline experiments
 - Simulations using on historical data (measuring, e.g., prediction accuracy, coverage)
- Theoretical analyses
 - For example, regarding scalability

Offline experiments

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

Sounds safe?

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

Progress can still be limited

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

Potential issues w/ research practice

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

A slightly exaggerated comparison

- | | |
|---|---|
| <ul style="list-style-type: none">• Kaggle machine learning competitions<ul style="list-style-type: none">– Defined dataset for training– Test dataset not revealed– Defined measures– Many competitors– (Sometimes code has to be made public) | <ul style="list-style-type: none">• Academic machine learning research<ul style="list-style-type: none">– 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 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

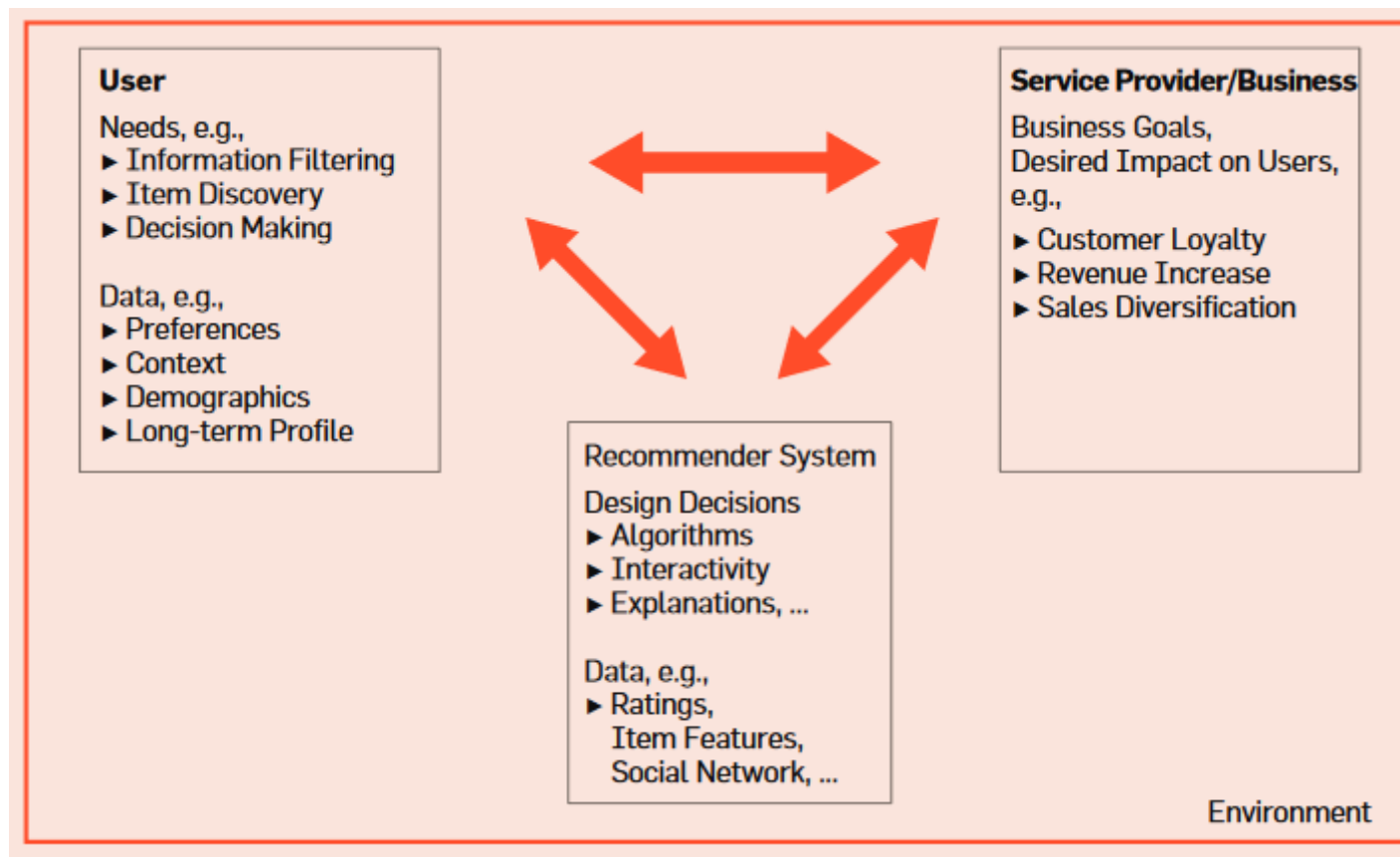


Possible steps forward

- 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

Possible steps forward

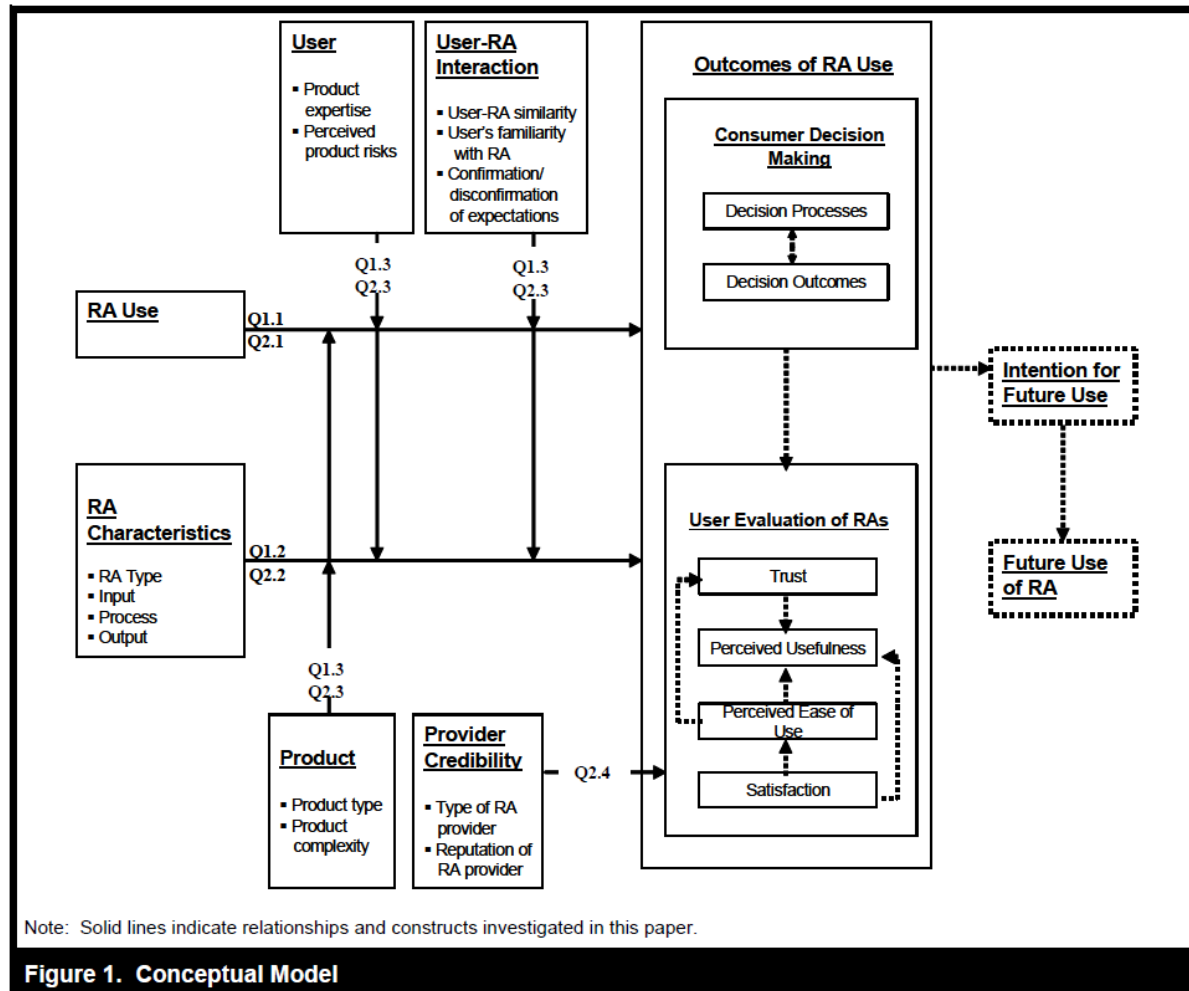
- “From algorithms to systems”



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

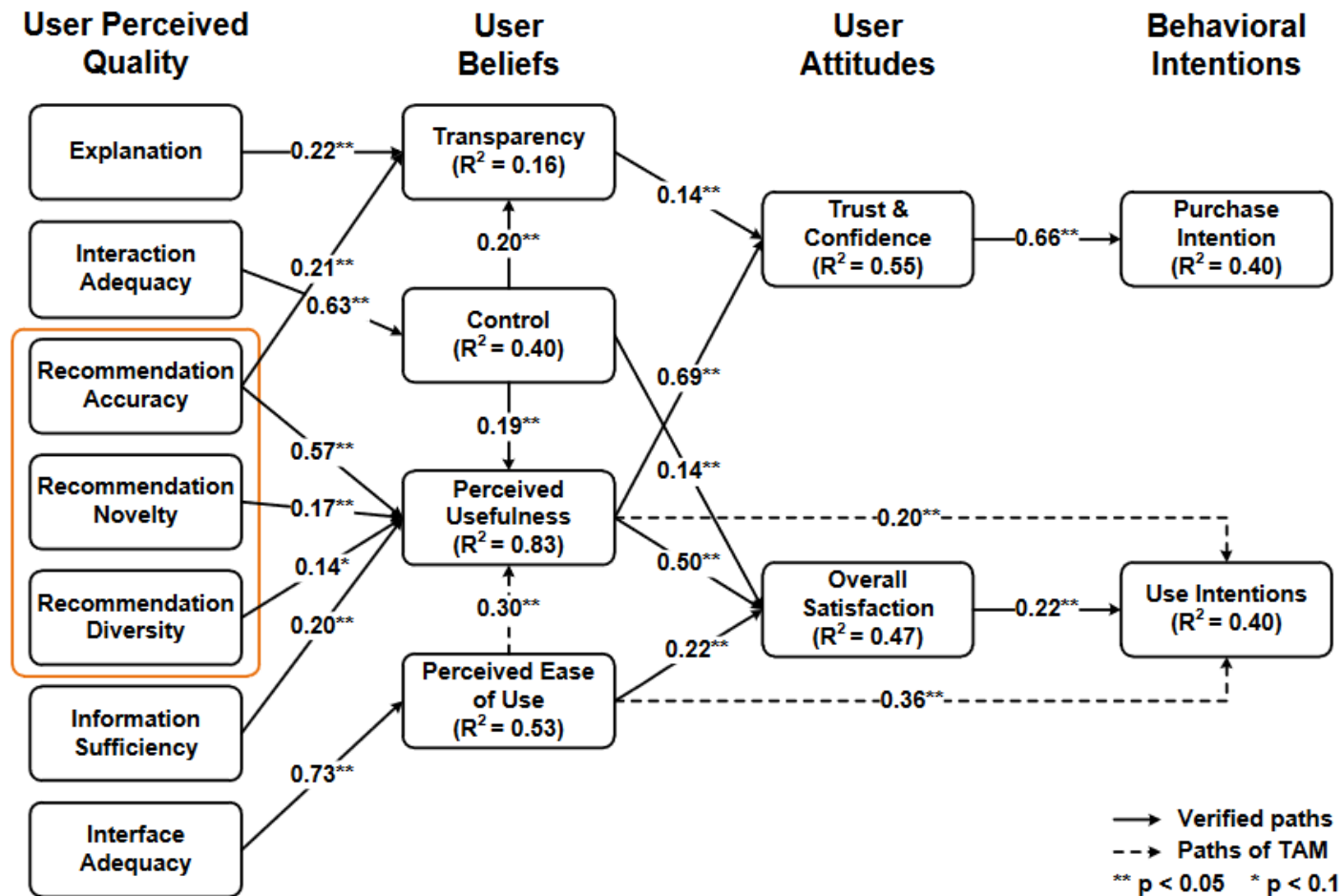
A conceptual model



User-centric research

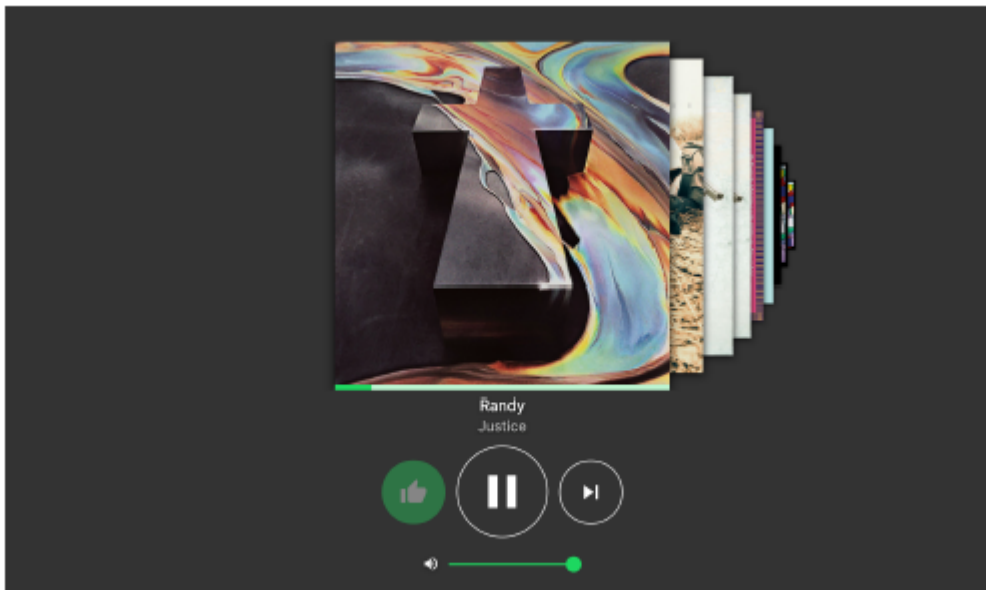
- Different evaluation frameworks exist, e.g.,
 - Pu et al. (RecSys 2011, UMUAI 2012)
 - Knijnenburg et al. (UMUAI 2012)
- Frameworks describe relevant quality criteria
 - e.g., perceived accuracy, novelty, diversity, context compatibility, interface adequacy, information sufficiency and explainability,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

Example validation



User studies: Examples

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



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

Background

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

Motivation and setup

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

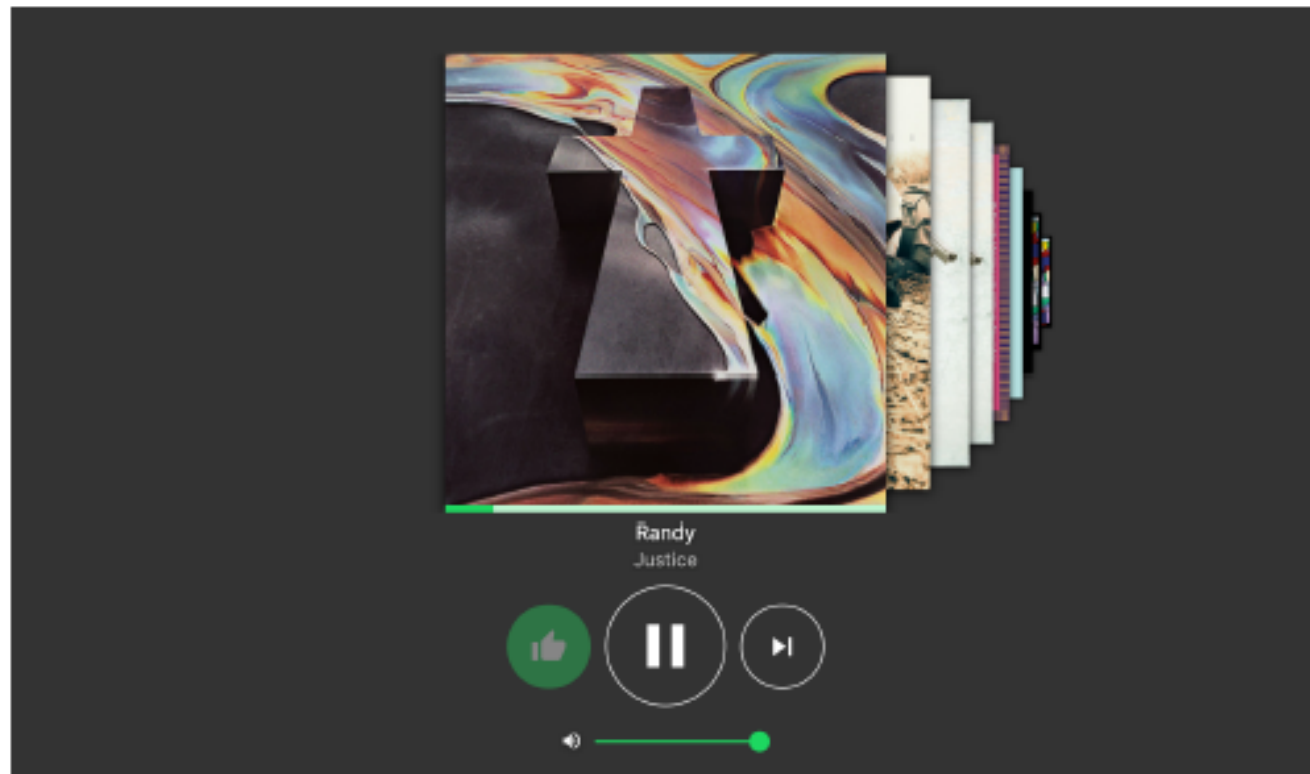
Experimental details

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

Experiment details

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

User interface



User interface

Do You know the track?* Yes ☒ No

Completely Disagree Completely Agree

Does the track match the previously liked tracks?*

○ — 2 — ○ — ○ — ○ — ○ — ○

Do you like the track in general?*

○ — ○ — ○ — ○ — ○ — 6 — ○

Finish Study

Questionnaire

Question

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

Running the experiment

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

Result analysis

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

Result analysis

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

Result analysis

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

Result analysis

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

Offline Results

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

Limitations

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

Summary of main findings

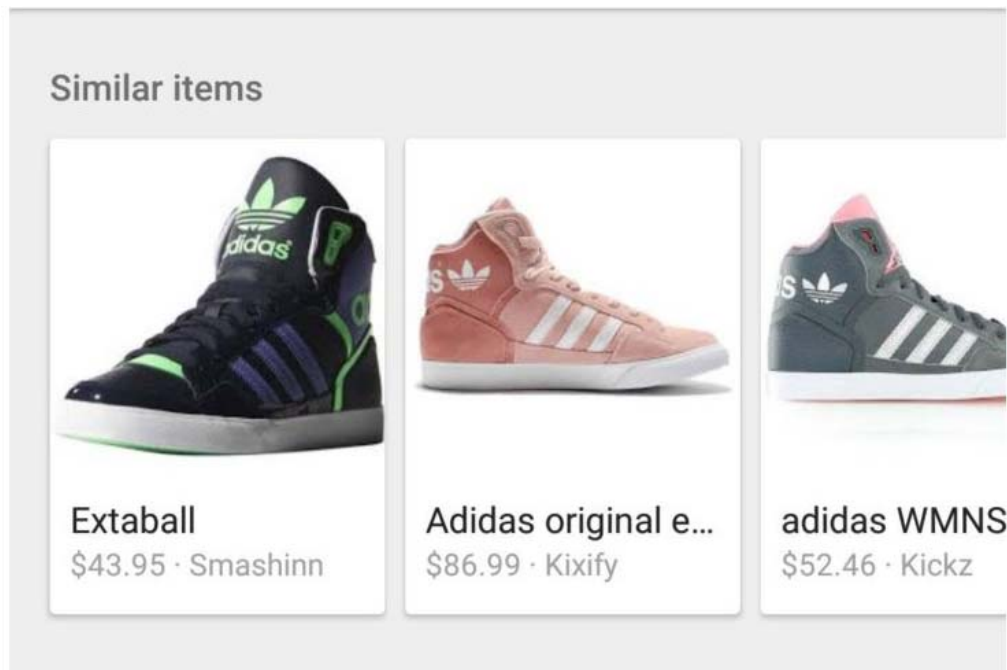
- Spotify
 - These recommendations would have led to terrible performance values in offline experiments
 - Still, they were well-received by the users
 - Spotify's recommendations help the purpose of discovery, which seems central for such an application
- S-KNN
 - was not only good in the offline setting,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

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

User studies: Examples

- **Example 2:** User perception of similar item recommendations



Trattner, C. and Jannach, D.: "Learning to Recommend Similar Items from Human Judgements, forthcoming

Background

- The recommendation of similar items is a common feature on many websites, e.g.,
 - Shopping goods
 - Artists
 - Movies with same actors
- Similarity functions play a central role here
 - like in content-based recommenders
 - In the literature, the design of this function is often based on [experience/intuition](#)
 - e.g., cosine distance of TF-IDF-encoded plot summaries

Background

- Similar items vs. related items
 - “Customers who bought” might lead to similar item recommendations, but they can also be accessories
- Recent research by Yao and Harper
 - Based on collecting (many) human similarity judgements
 - They analyze existing recommendation strategies with respect to their performance when recommending similar objects

Our goals

1. Automatically learn a useful similarity function
 - Based on pairwise human similarity judgments
 - Based on automatically extracted “content features”
2. Validate that the learned similarity function is indeed helpful in different ways
 - Being able to recommend objects that are perceived to be similar
 - Being able to produce recommendation lists that are “useful”

Experimental procedure

- Step 1: Collect similarity statements
 - Use a specifically designed application
 - Use crowdworkers to provide similarity judgements
 - Two domains
 - Cooking recipes
 - Movies
 - Show certain features that are often considered relevant
 - Name, image, cooking directions, ingredients
 - Title, cover image, plot summary, ...
 - Also ask participants which criteria they generally consider for their judgements

Step 1: User interface

[Task 1 / 10]

To what extent are the two recipes shown below similar?



(Scroll to the end of page to get to the next question)

Linguine Pasta with Shrimp and Tomatoes



Ingredients

2 tablespoons olive oil
3 cloves garlic, minced
4 cups diced tomatoes
1 cup dry white wine
2 tablespoons butter
salt and black pepper to taste
1 (16 ounce) package linguine pasta
1 pound peeled and deveined medium shrimp
1 teaspoon Cajun seasoning
2 tablespoons olive oil

Directions

Heat 2 tablespoons of olive oil in a large saucepan over medium heat. Stir in the garlic, cook 2 minutes. Add the tomatoes, and wine. Bring to a simmer and cook 30 minutes, stirring frequently. Once the tomatoes have simmered into a sauce, stir in the butter and season with salt and pepper. Fill a large pot with lightly-salted water, bring to a rolling boil, stir in the linguine and return to a boil. Cook the pasta uncovered, stirring occasionally, until the pasta has cooked through but is still firm to the bite,

Hudson's Baked Tilapia with Dill Sauce



Ingredients

4 (4 ounce) fillets tilapia
salt and pepper to taste
1 tablespoon Cajun seasoning, or to taste
1 lemon, thinly sliced
1/4 cup mayonnaise
1/2 cup sour cream
1/8 teaspoon garlic powder
1 teaspoon fresh lemon juice
2 tablespoons chopped fresh dill

Directions

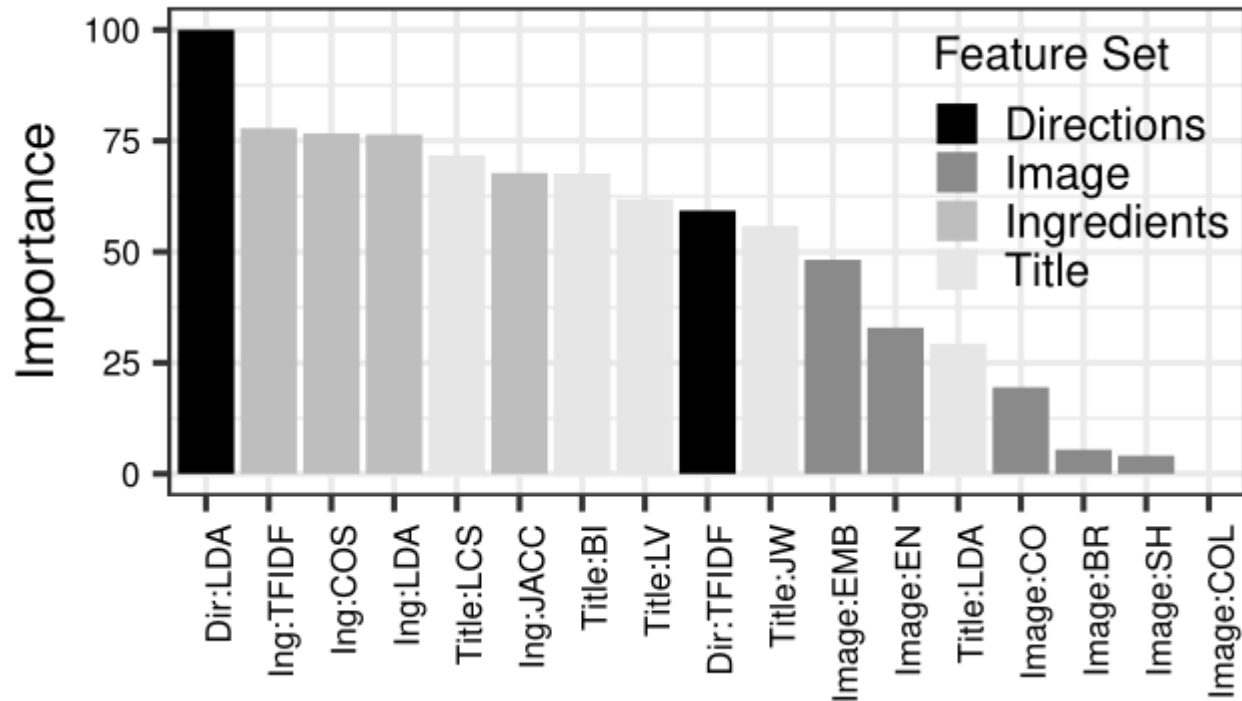
Preheat the oven to 350 degrees F (175 degrees C). Lightly grease a 9x13 inch baking dish.
Season the tilapia fillets with salt, pepper and Cajun seasoning on both sides. Arrange the seasoned fillets in a single layer in the baking dish. Place a layer of lemon slices over the fish fillets. I usually use about 2 slices on each piece so that it covers most of the surface of the fish. Bake uncovered for 15 to 20 minutes in the preheated oven, or until fish flakes easily with a fork.

Experimental procedure

- Step 2: Learn a regression function
 - To predict the similarity of items based on metadata
 - e.g., 17 different features for the recipe domain, relating to directions, ingredients, title, and image
 - Use collected human judgements as gold standard
 - Do feature importance analysis

Analysis of feature importance

- Recipe domain



Analysis of feature importance

- Movie domain
 - Almost all used features correlate strongly with the human similarity judgements
 - Also measured correlations using “social” information in the movie domain
 - Tag Genome similarity
 - Latent vector similarity computed on ratings (SVD)
 - Social information cues also correlate strongly with user perception

Validating the function

- Additional user study, where participants were presented
 - with a reference recipe
 - and five similar item recommendations
- Between-subjects design
 - Selection of similar item varied
 - All extracted features, only directions features, only ingredient features etc.
 - All extracted feat, plot descriptions, title, etc. plus social recommendations (Tags, SVD)

[Task 1 / 5]

Have a look at the reference recipe and the recommended similar recipe list!

(Scroll down to answer the survey questions)

Reference Recipe

Juiciest Hamburgers Ever



Ingredients

2 pounds ground beef
1 egg, beaten
3/4 cup dry bread crumbs
3 tablespoons evaporated milk
2 tablespoons Worcestershire sauce
1/8 teaspoon cayenne pepper
2 cloves garlic, minced

Directions

Preheat grill for high heat.
In a large bowl, mix the ground beef, egg, bread crumbs, evaporated milk, Worcestershire sauce, cayenne pepper, and garlic using your hands.
Form the mixture into 8 hamburger patties.
Lightly oil the grill grate. Grill patties 5 minutes per side, or until well done.

Recommended Similar Recipes

Hamburgers by Eddie

To what extent is this recipe similar to the reference recipe?



How likely is it that you will try this recipe?



Ingredients

1 pound ground beef
1 egg
2 teaspoons minced garlic
1 tablespoon steak sauce (e.g. A-1), or to taste

Directions

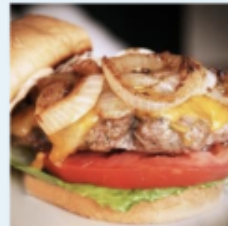
Preheat an outdoor grill for high heat.
In a medium bowl, mix together the ground beef, egg, and garlic. Mix in steak sauce until mixture is sticky

Best Hamburger Ever

To what extent is this recipe similar to the reference recipe?



How likely is it that you will try this recipe?



Ingredients

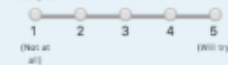
1 1/2 pounds lean ground beef
1/2 onion, finely chopped
1/2 cup shredded Colby Jack or Cheddar cheese
1 teaspoon soy sauce
1 teaspoon Worcestershire sauce
1 egg
1 (1 ounce) envelope dry onion soup mix
1 clove garlic, minced
1 tablespoon garlic powder
1 teaspoon dried parsley

Garlic and Onion Burgers

To what extent is this recipe similar to the reference recipe?



How likely is it that you will try this recipe?



Ingredients

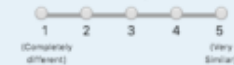
2 pounds ground beef
1 tablespoon Worcestershire sauce
3 cloves garlic, minced
1/2 cup minced onion
1 teaspoon salt
1/2 teaspoon ground black pepper
1 teaspoon Italian-style seasoning

Directions

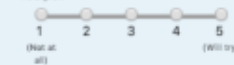
In a large bowl, mix together the beef, Worcestershire sauce, garlic, onion, salt, pepper and Italian

Juicy Lucy Burgers

To what extent is this recipe similar to the reference recipe?



How likely is it that you will try this recipe?



Ingredients

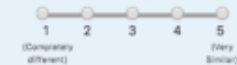
1 1/2 pounds ground beef
1 tablespoon Worcestershire sauce
3/4 teaspoon garlic salt
1 teaspoon black pepper
4 slices American cheese (such as Kraft®)
4 hamburger buns, split

Directions

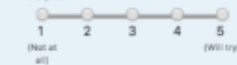
Combine ground beef, Worcestershire sauce, garlic salt, and pepper in a large bowl, mix well.

Biggest Bestest Burger

To what extent is this recipe similar to the reference recipe?



How likely is it that you will try this recipe?



Ingredients

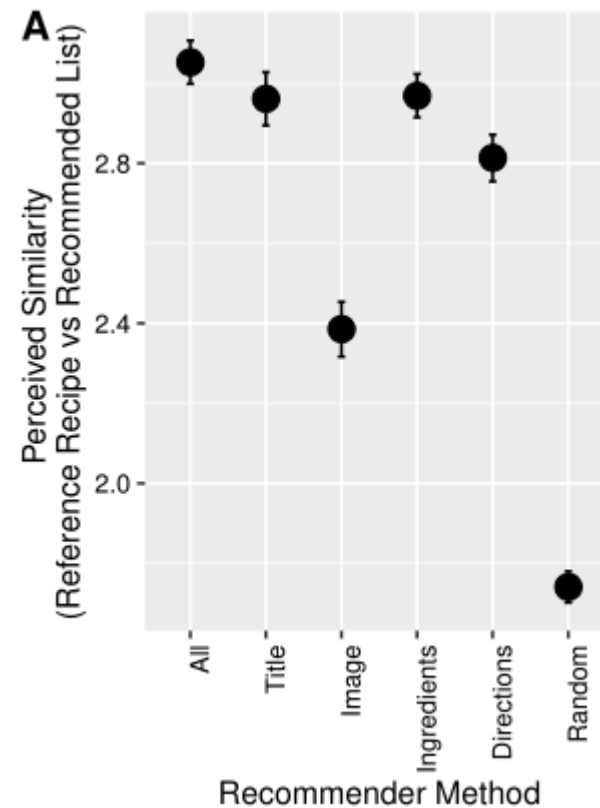
2 pounds ground beef
1 onion, chopped
1 teaspoon salt
1 teaspoon ground black pepper
1 teaspoon dried basil
1/4 cup Italian seasoned bread crumbs
1 tablespoon grated Parmesan cheese
1/3 cup teriyaki sauce
6 slices American cheese
6 onion rolls

Study Design

- For each recommendation, participants stated
 - a) if they consider the item similar to the reference item
 - b) if they plan to try out the recipe
- Post-task questionnaire
 - Asking additional questions about helpfulness, diversity, surprisingness, and excitingness of the list as a whole

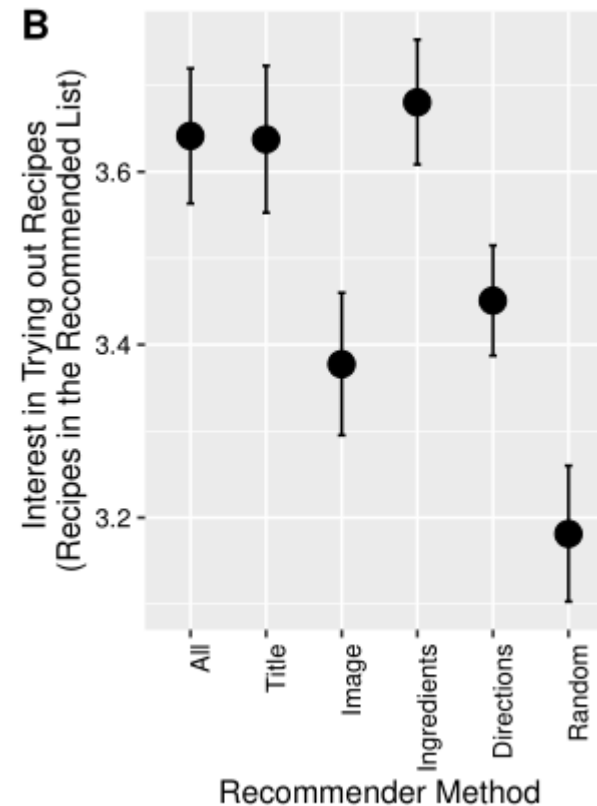
Result analysis (recipes)

- Combining features led to the best perceived similarity



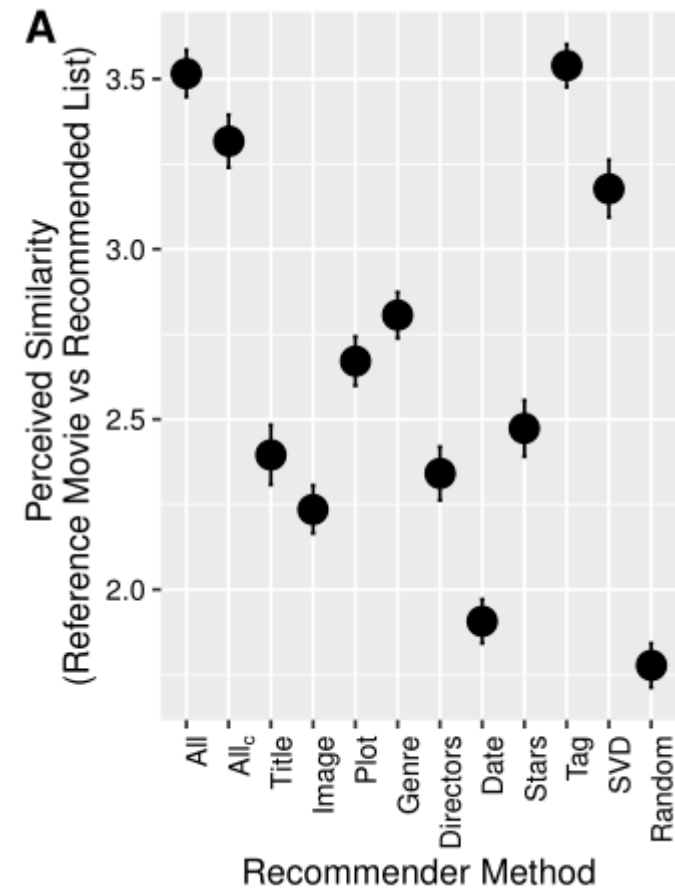
Result analysis (recipes)

- The most similar items are not necessarily the most inspirational ones



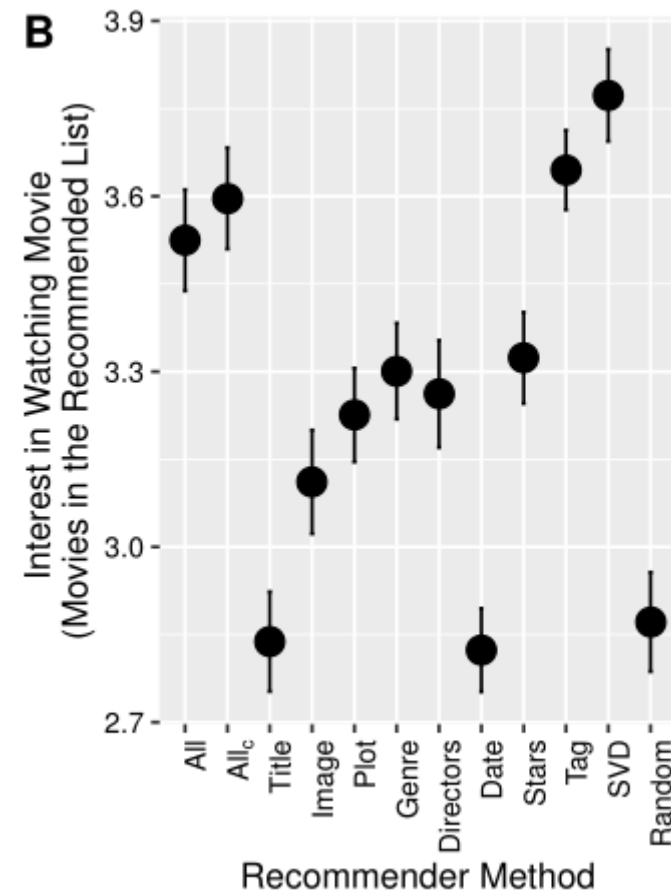
Result analysis (movies)

- A combination of automatically extracted features is almost as good as the social tags



Result analysis (movies)

- Too similar item recommendations are less inspirational than SVD-based recommendations



Some insights

- Automatically extracted features can be helpful to learn similarity functions without the need for social information (tags or ratings)
- If available, social information should be considered
- Choosing the right level of similarity depends on the application domain
 - Too similar options might be considered of less value than more inspirational recommendations

Summing up 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.

-
- 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