Recommender Systems: Value, Methods, Measurements

Dietmar Jannach, University of Klagenfurt, Austria

dietmar.jannach@aau.at

Presented at the ACM RecSys 2023 Summer School on Recommender Systems

Copenhagen, 2023

Outline (slides: https://tinyurl.com/recsys-summer-2023)

- Recommender Systems
  - Why should we have them?
    - Value
  - How do we build them?
    - Methods
  - How do we know they work well?
    - Measurements

- What makes building them difficult?
  - Both in industry and academia

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
(and some measurements)

What's their purpose and value?

- Recommenders can **create 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., item providers (manufacturers, property owners, artists), who my benefit from recommendations
    - They may have competing interests

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

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 may be better to recommend items with a higher margin (as long as the recommendations are still reasonable)
  - Leads to a multi-objective recommendation problem

=> See sessions on multi-stakeholder recommendation and fairness

Zheng, Y. and Wang, D. X. "A survey of recommender systems with multi-objective optimization". Neurocomputing 474, 141–153, 2022

Jannach, D. and Abdollahpouri, H.: "A Survey on Multi-Objective Recommender Systems". Frontiers in Big Data, Vol. 6. 2023

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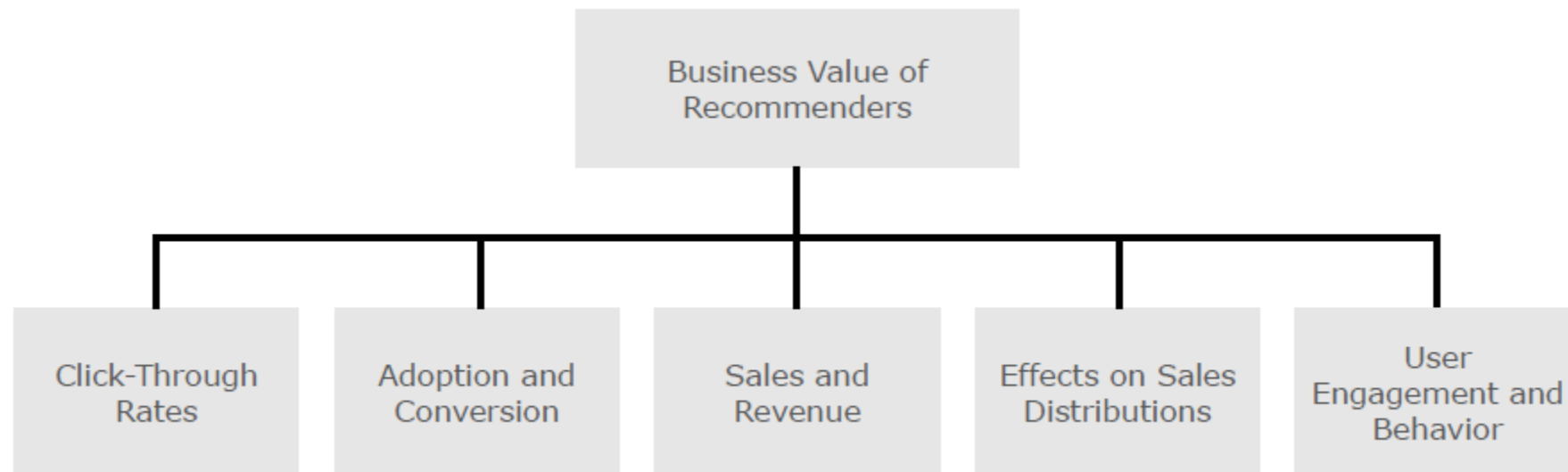
- Measuring the business value can be difficult
  - What does it tell us that 80% of the watched content comes from the recommendations (if everything is a recommendation)?
  - Where do the said savings come from?
- The used measures often largely depend on
  - The business model of the provider
  -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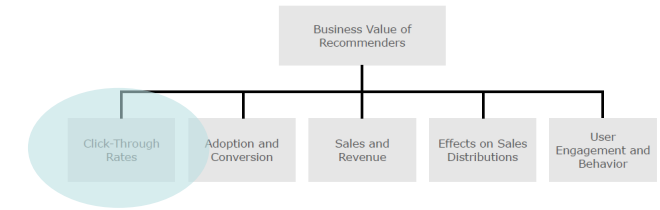 - Assumptions about consumer value

What is measured?

- Considering both the impact and value perspective

Jannach, D. and Jugovac, M.: "Measuring the Business Value of Recommender Systems". ACM Transactions on Management Information Systems, Vol. 10(4). 2019

Click-Through Rates

Business Value of Recommenders

Click-Through Rates | Adoption and Conversion | Sales and Revenue | Effects on Sales Distributions | User Engagement and Behavior

- 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
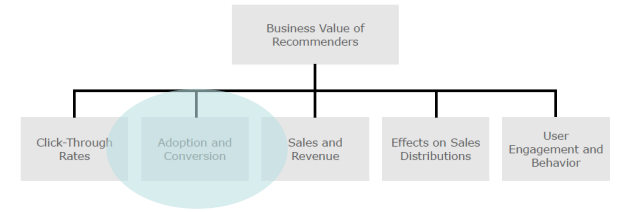
- CTR usually not the ultimate measure

  - 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
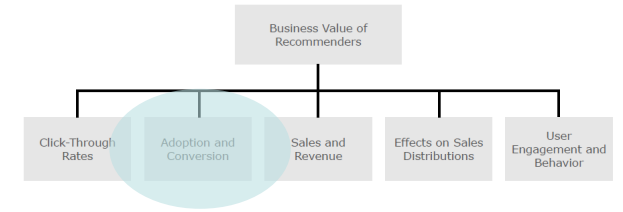
- Alternatives when items cannot be viewed/read:

- eBay:
  - "purchase-through-rate", "bid-through-rate"

- Other:
  - LinkedIn: Contact with employer made
  - Paper recommendation: "link-through", "cite-through"
  - E-Commerce marketplace: "click-outs"
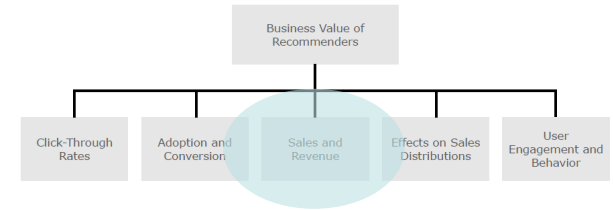 - Online dating: "open communications", "positive contacts per user"

From CTR to Sales and Revenue
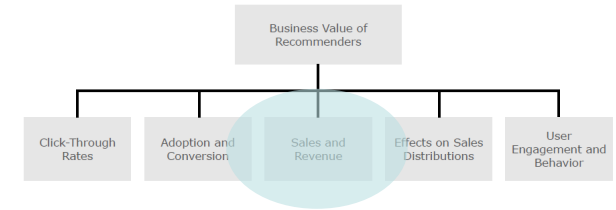
- CTR and adoption measures are good indicators of relevant recommendations

- However:
  - Often unclear how this translates into business value
  - Users might have bought an item anyway (i.e., without recommendation)
  - Substantial increases might be not relevant for business when starting from a very low basis

- Sales and Revenue figures are more direct value measurements

Sales and Revenue

Business Value of Recommenders

Click-Through Rates | Adoption and Conversion | Sales and Revenue | Effects on Sales Distributions | User Engagement and Behavior

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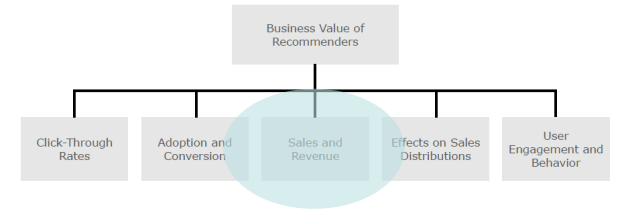
- e-grocery studies:
  - 1.8 % direct increase in sales in one study
  - 0.3 % direct effects in another study
  - However:
    - Up to 26% indirect effects, e.g., where customers were pointed to other categories in the store
    - "Inspirational" effect also observed in music recommendation in our own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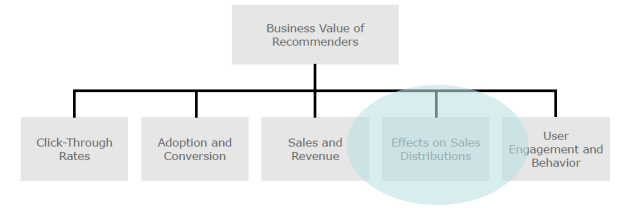- eBay:
  - 6 % increase for similar item recommendations through largely improved algorithm
  - (500 % increase in other study for specific area)

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)
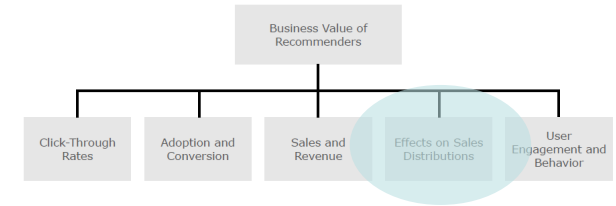 - 3.6 % more purchases through best recommender
  - More was possible

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

- ## 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

Effects on Sales Distributions

Business Value of Recommenders

Click-Through Rates | Adoption and Conversion | Sales and Revenue | Effects on Sales Distributions | User Engagement and Behavior
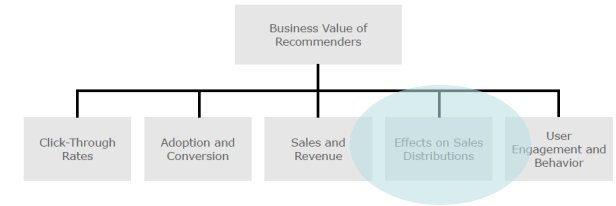
- ## Premium cigars study:
  - – Interactive advisory system installed
  - – Measurable shift in terms of what is sold
    - e.g., due to better-informed customers

Rank before introduction of Mortimer in brackets

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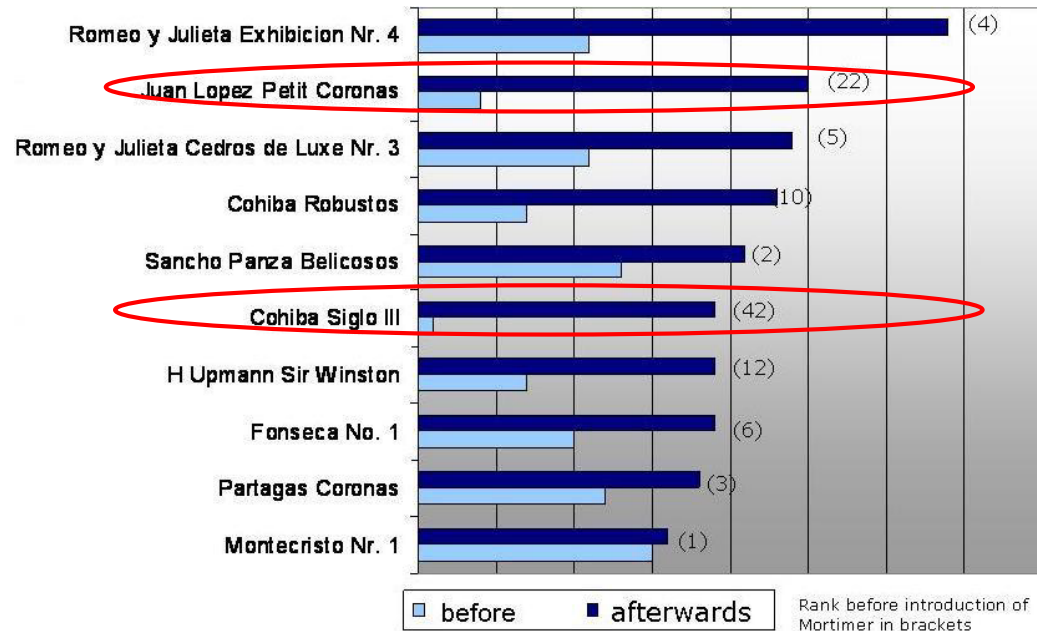
User Behavior and Engagement
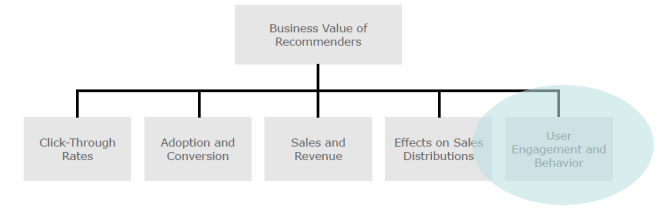
- Assumption:
  - Higher engagement leads to higher re-subscription rates (e.g., at Spotify)
- News domain studies:
  - 2.5 times longer sessions, more sessions when there is a recommender
- Music domain study:
  - Up to 50% more user activity
- LinkedIn:
  - More clicks on job profiles after recommender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

Why measuring value is difficult

- Operational aspects
  - Certain measures may actually be misleading
    - in particular CTR
  - Longitudinal effects often unclear
    - even when A/B tests last for weeks
  - A/B testing can be expensive and risky
    - Finding good "offline" proxy metrics is challenging

Why measuring value is difficult

- Strategic/organizational aspects
  - It may be actually unclear what the relevant business KPIs are
    - There may also be competing department interests
  - There may be multiple competing objectives
    - How to balance them, e.g., short-term vs long-term, consumer vs provider benefit

- What to measure then?
  - It all depends on the intended purpose and value of the system

The academic perspective

- In academia, we aim to
  - abstract from application specifics, e.g., business models, 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).

What to measure: 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
Strategic Perspective — **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 • ...	
Operational Perspective — **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
Strategic Perspective — **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 • ...	
Operational Perspective — **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**
Strategic Perspective — **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 ...	
Operational Perspective — **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**
Strategic Perspective — 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 • ...	
Operational Perspective — 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**
Strategic Perspective — 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 • ...	
Operational Perspective — 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 and value is generally not trivial
  - It all depends on the purpose
    - Choice of the evaluation measure matters a lot
    - CTR can be misleading
- "Metric-Task-Purpose-Fit" to be considered

Part II: Methods

History and roots

- 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, based on the used information

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

- Interactive and Conversational Recommendation

The Basic Principle (simplified)*

User profile &
Contextual parameters

Recommendations are usually
personalized

item	score
i1	0.9
i2	1
i3	0.3
...	...

Recommendation
component

Recommendation
list
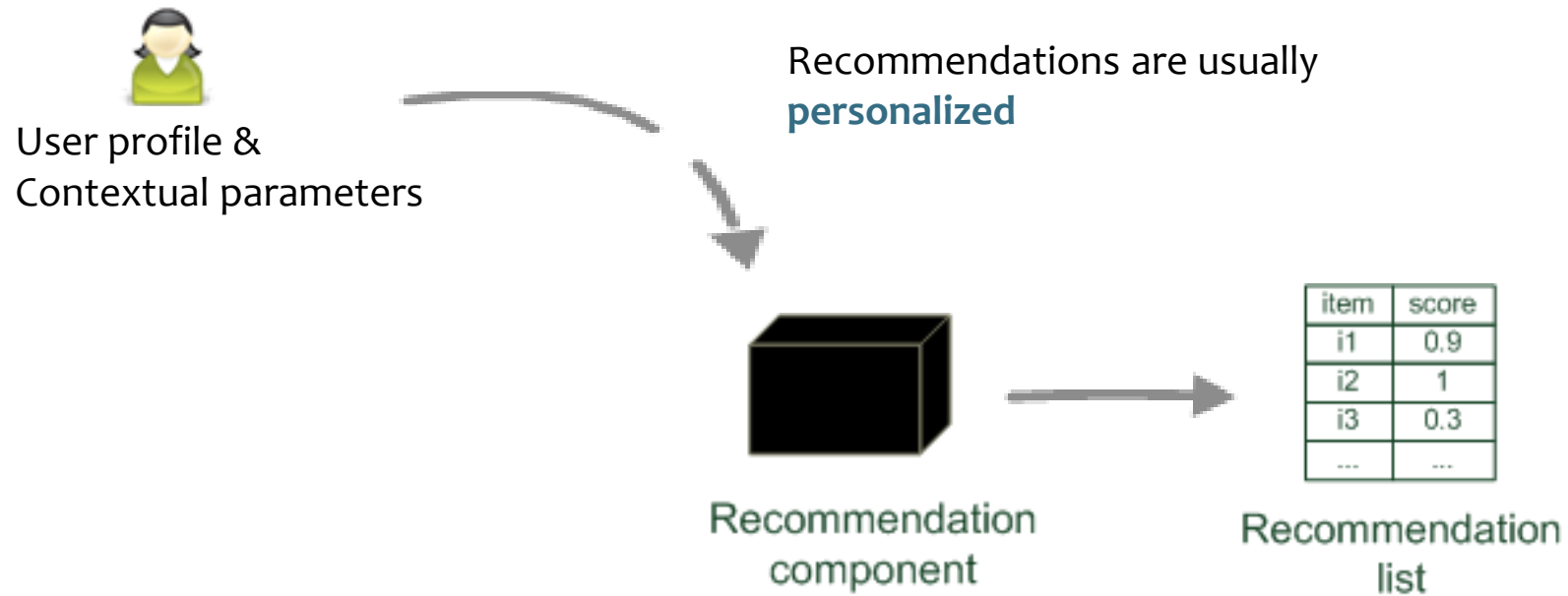
*One-shot scenario, no further interaction, no explanations etc.

Content-based Filtering

Content-based:
"Show me more of the same what I've liked"

User profile & contextual prameters

Title	Genre	Actors	...

Product features

Recommendation component

item	score
i1	0.9
i2	1
i3	0.3
...	...

Recommendation list
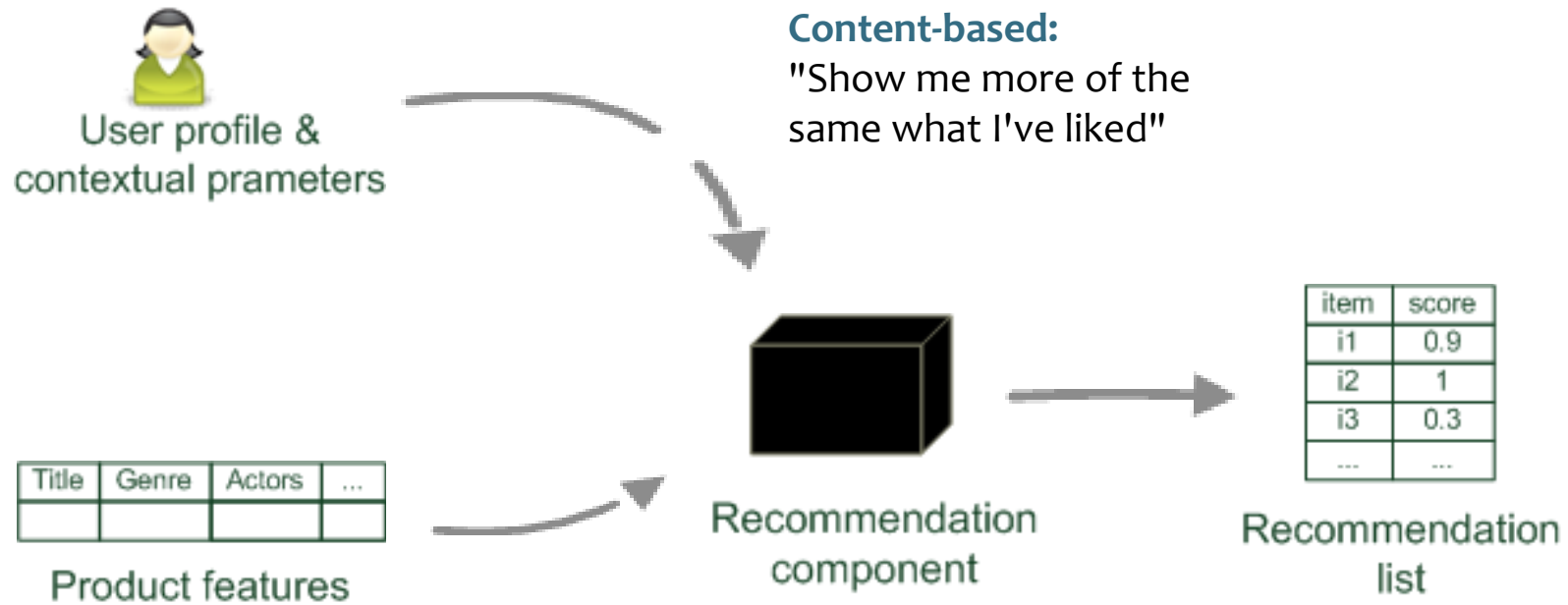
Content-based Filtering

- Basic approach
  - Automated construction of a user profile based on characteristics of items the user previously liked/consumed/etc.
  - Various types of item information considered in the literature
  - Recommending: Matching of user profile with item profiles
    - By design leads to "more of the same" recommendations and limited discovery
- Long history since the 1960s
  - Similar to personalized information retrieval, the term "content-based" stems from early use cases such as news recommendation.

=> See sessions on content-based filtering and media recommendation

Collaborative Filtering

Collaborative:
"Tell me what's popular among my peers"

User profile &
Contextual parameters

Community data

Recommendation component

item	score
i1	0.9
i2	1
i3	0.3
...	...

Recommendation list
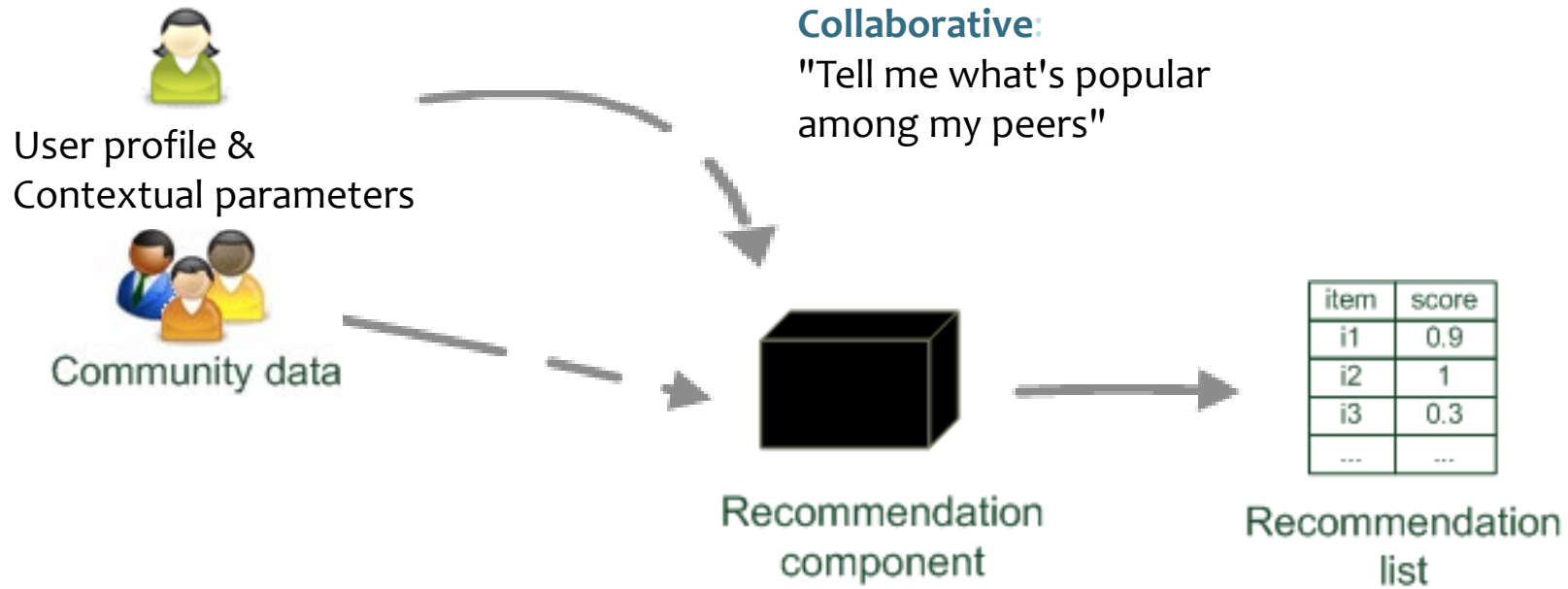
Collaborative Filtering

- Basic Approach
  - Leverage the opinion or observed behavior of other users when recommending
  - Assumption (simplified): When users agreed in their opinion in the past, they will probably agree in the future
- History since early 1990s
  - Term introduced in 1992 with experimental "Tapestry" system
  - "GroupLens" system and related systems proposed in 1994
    - Fully automated prediction of user preferences

Collaborative Filtering in GroupLens

- The GroupLens system
  - User-item ratings as the only input
  - Recommendations based on nearest-neighbor approach
  - Original paper proposed a system, and not only an algorithm

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.
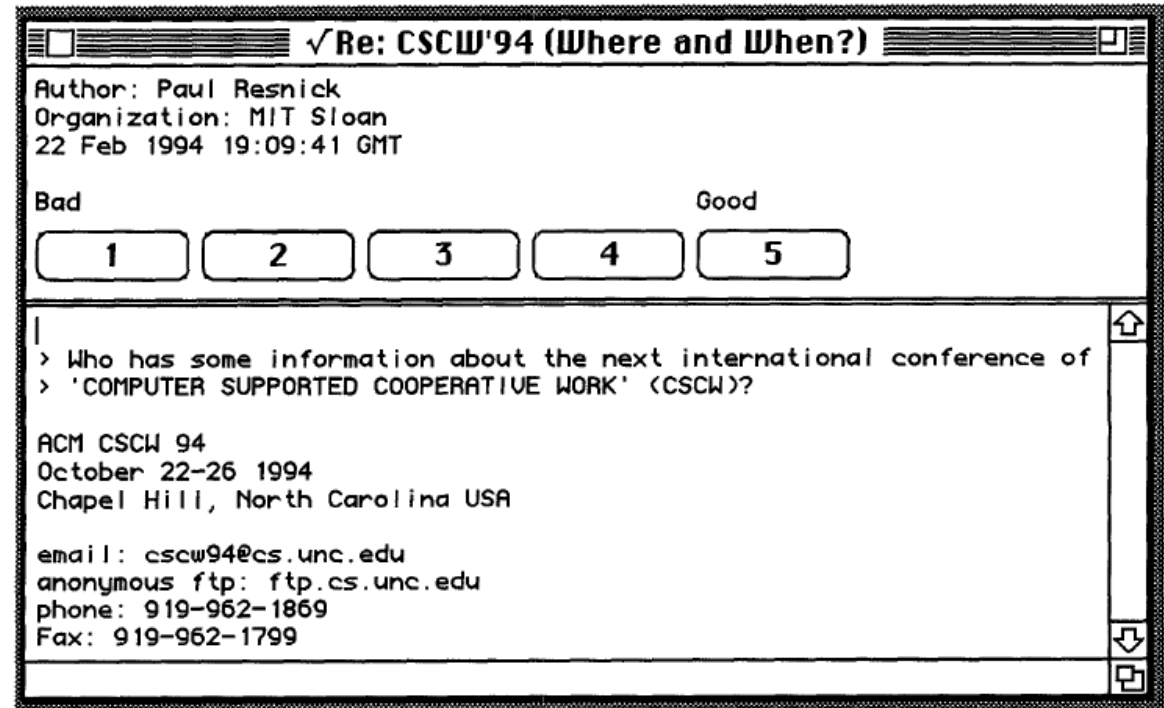
Matrix Completion

- 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

- Items are recommended based on predicted ratings

GroupLens: User-based K-nearest-neighbors (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
  - Assessing the predictions without users (offline):
    - Hide & predict some ratings, compute the average prediction error, e.g., based on the Root Mean Square Error

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

kNN Methods: Variations and hyperparameters

- How to measure similarity?
  - e.g., Pearson correlation or Cosine Similarity
- How many neighbors should we consider?
  - e.g., decide on this hyperparameter based on RMSE
- How do we generate a prediction from the neighbors' ratings?
  - e.g., use a weighted approach of neighbor ratings as proposed in GroupLens

	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

Collaborative Filtering (CF): A Success Story

- **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-dimensional spaces
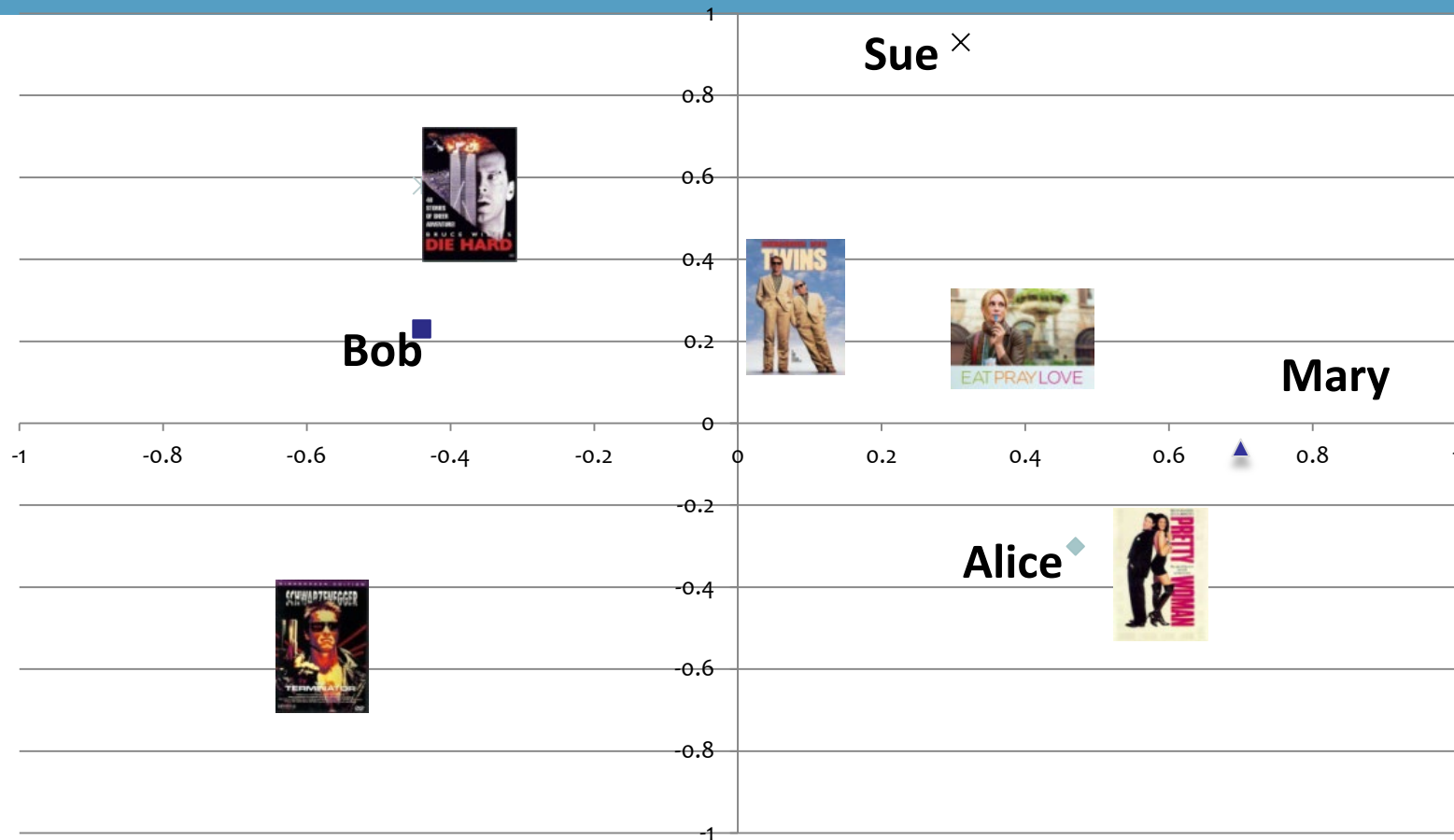
- Nowadays called "embeddings"

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
- During the last few years, deep learning techniques dominate the landscape

Pablo Castells and Dietmar Jannach, „Recommender Systems: A Primer", to appear in „Advanced Topics in Information Retrieval", ACM Press, 2023
https://arxiv.org/abs/2302.02579

Recommendation as a Supervised ML Problem

- While the techniques change over time, the underlying machine learning (ML) problem formulation remained mostly constant, where the goal is to (Adomavicius & Tuzhilin, 2005)

  - learn a function from noisy data, which

  - given a user *u* and an item *i*,

  - predicts the relevance of *item* i *for user u.*

- Learning-to-rank models may work a bit differently, but ultimately also predict relevance scores for a given set of items

Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. on Knowl. and Data Eng. 17, 6 (June 2005), 734–749

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

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?

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

Past sessions
of the current
user

?
?
?

Current
session

Past sessions of the user community
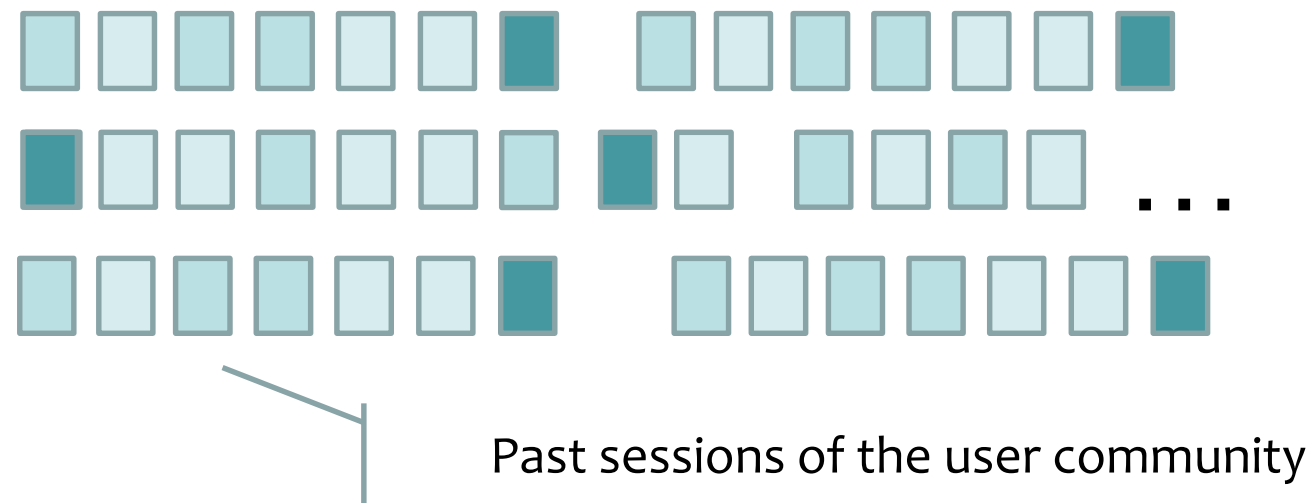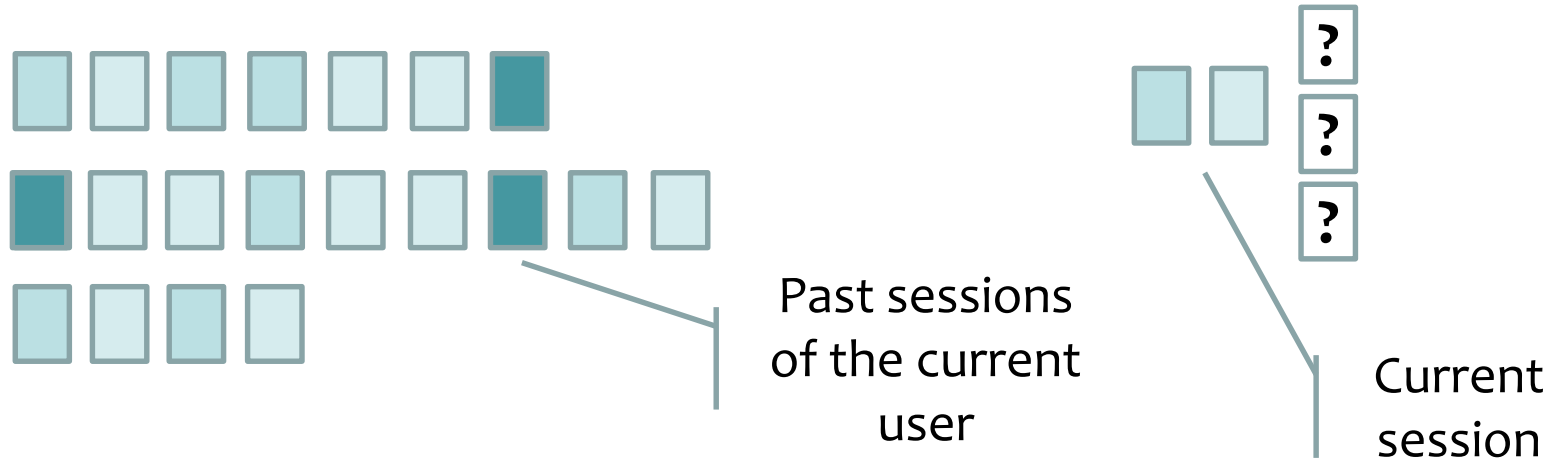
Technical Approaches

- 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, Attention/Transformer

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

Hybrid Recommendation Approaches

User profile & Contextual parameters

Community data

Title	Genre	Actors	...

Product features

Knowledge models

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

Recommendation component

item	score
i1	0.9
i2	1
i3	0.3
...	...

Recommendation list
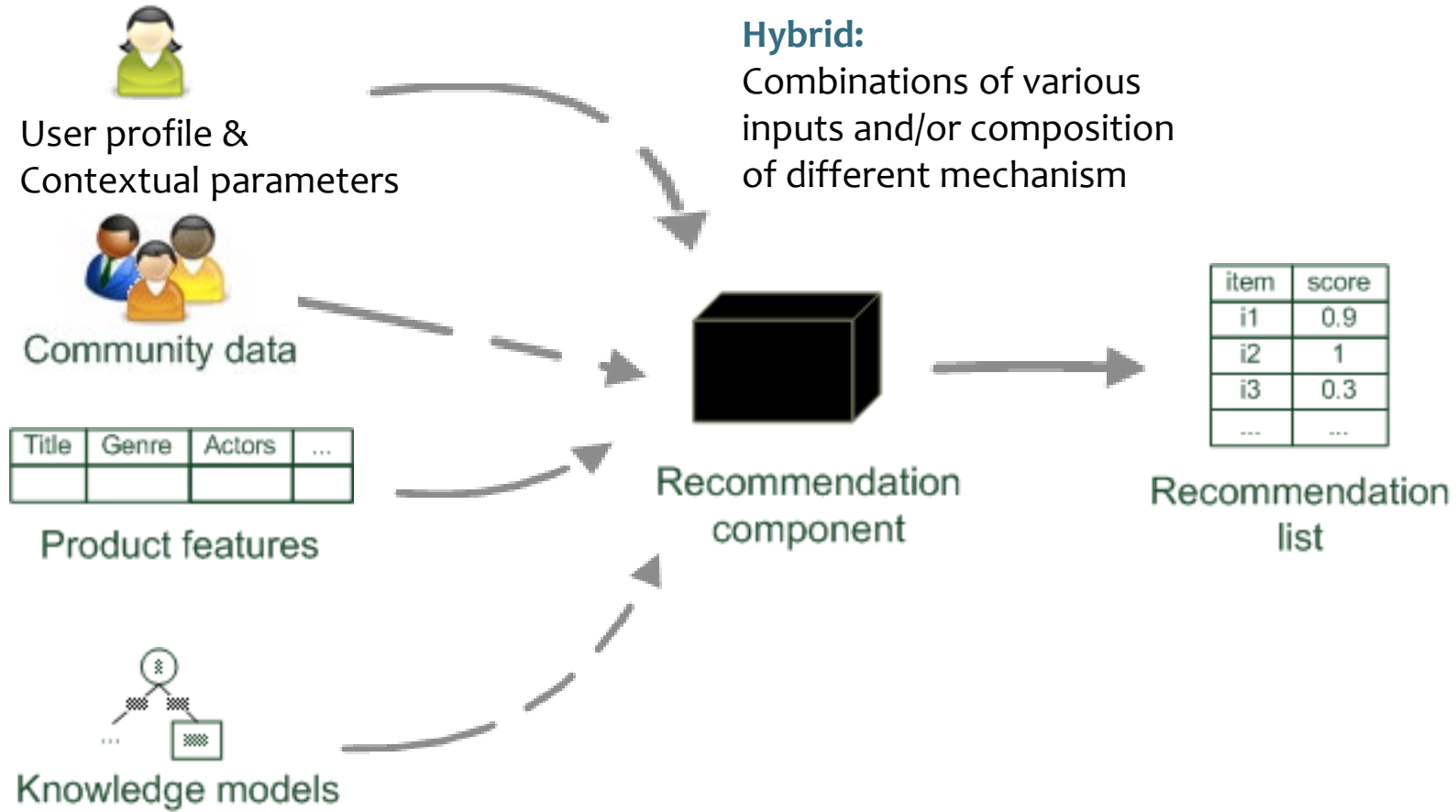
Hybridization Designs

- Various forms proposed in the literature
  - Monolithic exploiting different features
    - E.g., Combining different signals in one system
  - Parallel use of several systems
    - E.g., switching, based on recorded user interactions
  - Pipelined invocation of different systems
    - E.g., use one recommender for filtering items and another one for ranking what remains

Robin Burke. 2002. Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction 12, 4 (November 2002), 331–370.

Jannach, D., Zanker, M., Felfernig, A. and Friedrich, G.: "Recommender Systems - An Introduction". Cambridge University Press, 2010.

Collaborative Filtering with Side Information

- 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

Zhu Sun, Qing Guo, Jie Yang, Hui Fang, Guibing Guo, Jie Zhang, Robin Burke, "Research commentary on recommendations with side information: A survey and research directions", Electronic Commerce Research and Applications, Volume 37, 2019

Knowledge-based Systems

Knowledge-based:
"Tell me what fits based on my needs"

User profile & contextual prameters

Title	Genre	Actors	...

Product features

Knowledge models

Recommendation component

item	score
i1	0.9
i2	1
i3	0.3
...	...

Recommendation list

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?

A common categorization, based on the used information

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

- **Interactive and Conversational 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
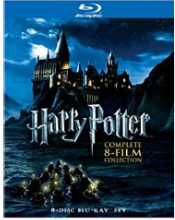
Konstan, J.A. & Riedl, J.. "Recommender systems: from algorithms to user experience"
User Model User-Adap Inter (2012) 22: 101.

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

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

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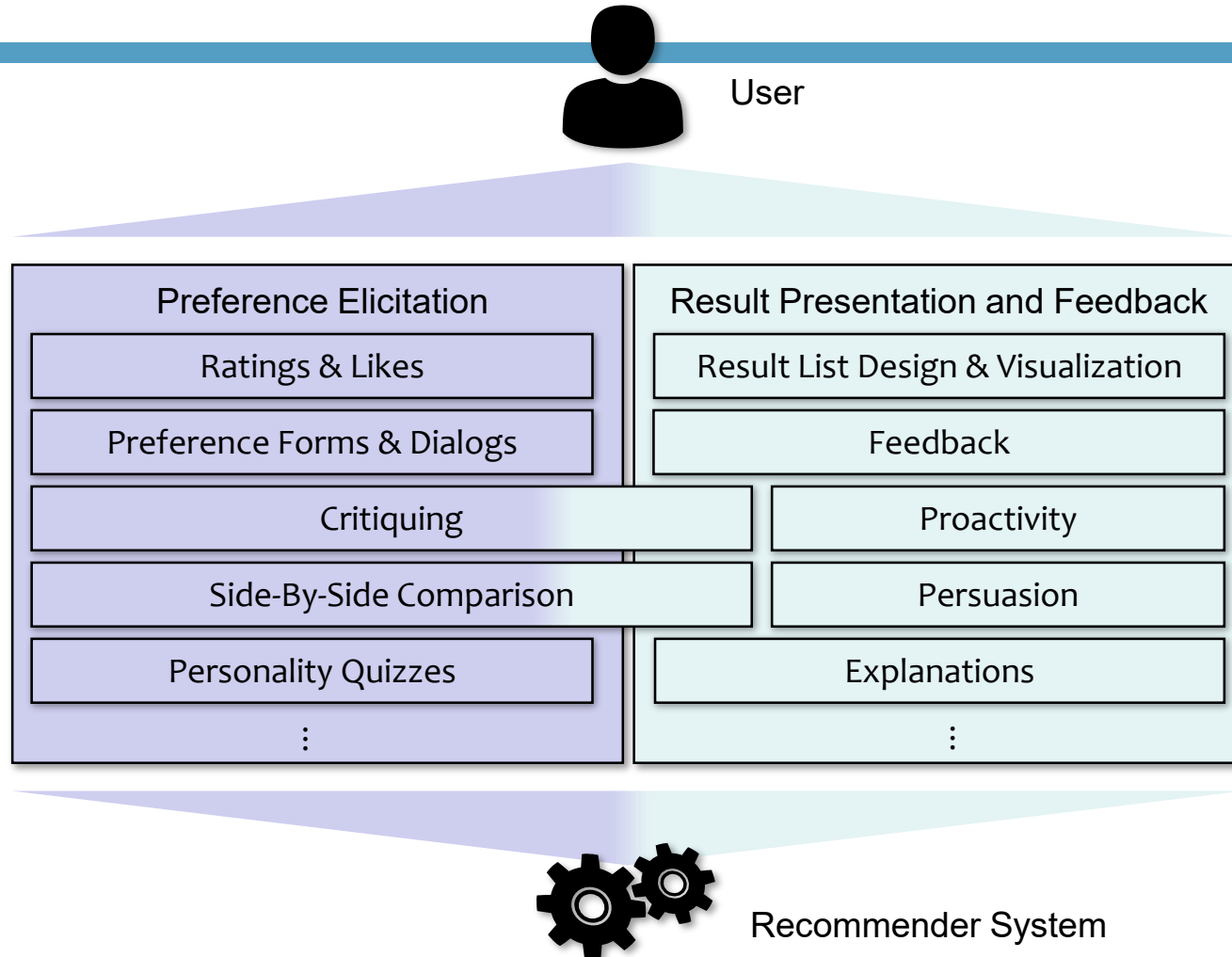
Design Space Examples

- Telling the system explicitly what you like
  - Global settings
  - Ratings
    - But how many options? How many categories?
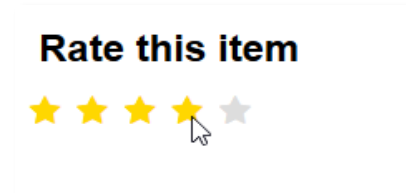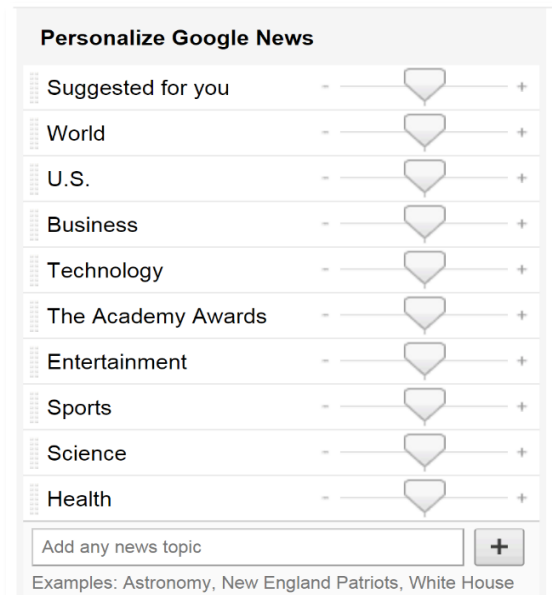
Sources: Facebook.com, Google.com

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?

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

Nikon AF-S FX NIKKOR 50mm f/1.8G Lens with Auto Focus for Nikon DSLR Cameras
1,505
#1 Best Seller in Camera & Camcorder Lens...
$216.95 Prime

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...
592
$22.50 Prime

Source: amazon.com

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

=> See session on explanations

Interactive and Conversational Systems

- Increased research on Conversational Recommender Systems in recent years
  - Mostly chatbot-like systems, supporting natural language conversations
  - New horizons with ChatGPT

Jannach, D., Manzoor, A., Cai, W. and Chen, L.: "A Survey on Conversational Recommender Systems". ACM Computing Surveys, Vol. 54(5). 2021, pp. 1-26

Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, Tat-Seng Chua,"Advances and challenges in conversational recommender systems: A survey", AI Open, Volume 2, 2021

=> See session on conversational recommendations

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"
- But, there probably is no best model. The ranking of algorithms can depend on:
  - Given dataset
  - Used pre-processing steps
  - Evaluation measure
  - Choice of baselines
  - **Optimization of baselines**

Worrying observations

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

Worrying observations

- Makridakis (2018) compared various ML methods for time-series prediction, concluding that existing statistics-based methods are often better

- Ludewig et al. (2018-2020) evaluated various session-based recommendation techniques, finding that simple methods are often very competitive

- Ferrari Dacrema et al. (2019/2021) examined recent neural top-n recommendation techniques and found potential issues in terms of the choice and optimization of baselines

Problems piling up, also in other ML areas

Armstrong, T.G., A. Moffat, W. Webber, and J. Zobel. 2009. "Improvements That Don't Add Up: Ad-hoc Retrieval Results Since 1998." In Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09), 601–10.

Lin, J. 2019. "The Neural Hype and Comparisons Against Weak Baselines." ACM SIGIR Forum 52(2): 40–51.

Makridakis, S., E. Spiliotis, and V. Assimakopoulos. 2018. "Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward." PloS one 13(3): 1–26.

Rendle, S., W. Krichene, L. Zhang, and J. Anderson. 2020. "Neural Collaborative Filtering vs. Matrix Factorization Revisited." In Proceedings of the 14th ACM Conference on Recommender Systems (RecSys '20).

Ferrari Dacrema, M., P. Cremonesi, and D. Jannach. 2019. "Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches." In Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19), 101–9.

Ferrari Dacrema, M., S. Boglio, P. Cremonesi, and D. Jannach. 2021. "A troubling analysis of reproducibility and progress in recommender systems research." ACM Transactions on Information Systems (TOIS), 39(2): 1–49.

Ludewig, M., S. Latifi, N. Mauro, and D. Jannach. "Empirical Analysis of Session-Based Recommendation Algorithms." User Modeling and User-Adapted Interaction 31(3): 49–181.

Lipton, Z. C., and J. Steinhardt. 2019. "Research for Practice: Troubling Trends in Machine-Learning Scholarship." Communications of the ACM 62(6): 45–53.

Beyond recommender systems

Thread

Tim Dettmers
@Tim_Dettmers

I got excited about a paper, implement stuff and then see they cheated: (1) Copy baseline results from other paper, (2) do much more hyperparam tuning on their own method, (3) accepted to EMNLP. Results look good, but their method is crap! Why waste people's time like this?

Please Commit More Blatant Academic Fraud

Explicit academic fraud is, of course, the natural extension of the sort of mundane, day-to-day fraud that most academics in our community commit on a regular basis. **Trying that shiny new algorithm out on a couple dozen seeds, and then only reporting the best few. Running a big hyperparameter sweep on your proposed approach but using the defaults for the baseline. Cherry-picking examples where your model looks good, or cherry-picking whole datasets to test on, where you've confirmed your model's advantage**. Making up new problem settings, new datasets, new objectives in order to claim victory on an empty playing field. Proclaiming that your work is a "promising first step" in your introduction, despite being fully aware that nobody will ever build on it. Submitting a paper to a conference because it's got a decent shot at acceptance and you don't want the time you spent on it go to waste, even though you've since realized that the core ideas aren't quite correct.

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 …

Potential ways forward

- In the long run: Need better education and awareness
  - Among students, teachers, researchers, policy makers, …
  - Choice of dataset/pre-processing often seems arbitrary

Bauer, C., Fröbe, M., Jannach, D., Kruschwitz, U., Rosso, P., Spina, D. and Tintarev, N.: "Overcoming Methodological Challenges in Information Retrieval and Recommender Systems through Awareness and Education". In: Dagstuhl Seminar 23031: Frontiers of Information Access Experimentation for Research and Education. 2023

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
  - …

Multi-metric over-simplification

- Using some diversity metric along with accuracy may not be sufficient either

  - We need to validate that the metric matches user perceptions

- Moreover, 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

Jesse, M., Bauer, C. and Jannach, D.: "Intra-List Similarity and Human Diversity Perceptions of Recommendations: The Details Matter". User Modeling and User Adapted Interaction, Vol. forthcoming. 2022

General problems of offline experiments
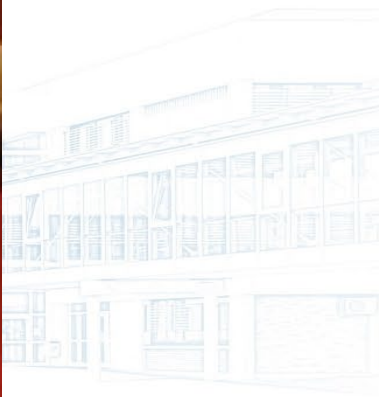
- Are offline experiments actually predictive of the perceived value?
  - Gomez-Uribe and Hunt (2015), Netflix, found that offline experiments were **not** found "*to be as highly predictive of A/B test outcomes as we would like.*"
  - In fact, a number of user studies did **not** find that algorithms 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

- Problem: We measure (only) what we can easily measure

  - The McNamara Fallacy

Jannach, D. and Bauer, C.: "Escaping the McNamara Fallacy: Towards more Impactful Recommender Systems Research". AI Magazine, Vol. 41(4). 2020, pp. 79-95

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

- "From Algorithms to Systems"

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

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

=> See session on user-centric experimentation

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

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

Wrap-up

Summary

- Discussed the business value of recommender systems
  - These are systems to have a real-world impact
  - It is great to work on this topic!
- Reviewed briefly how we can build such systems
  - From algorithms to the user experience
- Discussed limitations of current research practices
  - With an outlook on potential ways forward

- Thank you for your attention
- Enjoy the summer school!

- Contact:

dietmar.jannach@aau.at

- Slides:

https://tinyurl.com/recsys-summer-2023