

Explaining Recommendations to End Users: Simple or Complex?

Keynote at QUARE Workshop at SIGIR '22
Madrid/Online

Dietmar Jannach, University of Klagenfurt, Austria

dietmar.jannach@aau.at

Recommender Systems

- A big success in industry



- Amazon, Spotify, YouTube, Netflix, Facebook, Twitter, Google, ...

Explaining the recommendations

- From Amazon.com
 - Answering potential “why” questions by end users

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

Because you purchased...



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

DVD ~ Charlize Theron



Don't use for recommendations



Why explaining? (1)

Don't use for recommendations

- Transparency
 - Explain how the system works
- Scrutability
 - Allow users to tell the system it is wrong
- Trust
 - Increase users' confidence in the system
- Effectiveness
 - Help users make good decisions

Why explaining? (2)

Don't use for recommendations

- Persuasiveness
 - Convince users to try or buy
- Efficiency
 - Help users make decisions faster
- Satisfaction
 - Increase the ease of use or enjoyment

Tintarev, N., Masthoff, J. : "*Beyond Explaining Single Item Recommendations*". In: Recommender Systems Handbook. Ricci, F., Shapira, B. and Rokach, L. (Eds.), Springer US, 2022

Also: Buchanan, B.G., Shortliffe, E.H.: Explanations as a Topic of AI Research, in Rule-based Systems, pp. 331–337. Addison-Wesley, Massachusetts (1984)

How to explain?

- Various dimensions:
 - What **type** of content/information to show?
 - What **level of detail** or complexity?
 - In which **form**: visual, text-based?
 - **Personalized** or not?
 - **When**: Only when requested? Always?
 - What is the **target**:
 - Explaining the recommendation or **why not** an alternative?

How to explain – Complex or simple

- From Pandora Music

[Tintarev & Masthoff, 2022]

Based on what you've told us so far, we're playing this track because it features solo strings, mystical qualities, minor key tonality, melodic songwriting and intricate melodic phrasing.

Close

*“Based on what you've told us so far, **we're playing this track because** it features solo strings, mystical quality, minor key tonality, melodic songwriting and intricate melodic phrasing.”*

How to explain – Complex or simple?

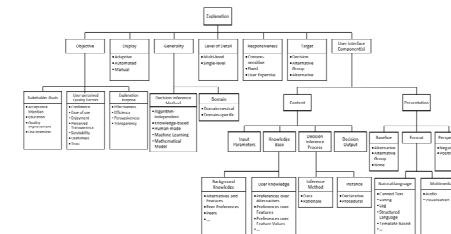
- How done in this case:
 - Information **about** assumed item features and (assumed) user preferences
 - Rather detailed
 - Text-based
 - Personalized
 - Upon request

Based on what you've told us so far, we're playing this track because it features solo strings, mystical qualities, minor key tonality, melodic songwriting and intricate melodic phrasing.

Close

How to explain – Complex or simple?

- Possible alternatives
 - “*Everybody likes Ed Sheeran*”or
 - “*Our recommendations are based on the trendiness of the artist*”
- The key questions
 - Which explanation one is better?
 - According to which purpose?

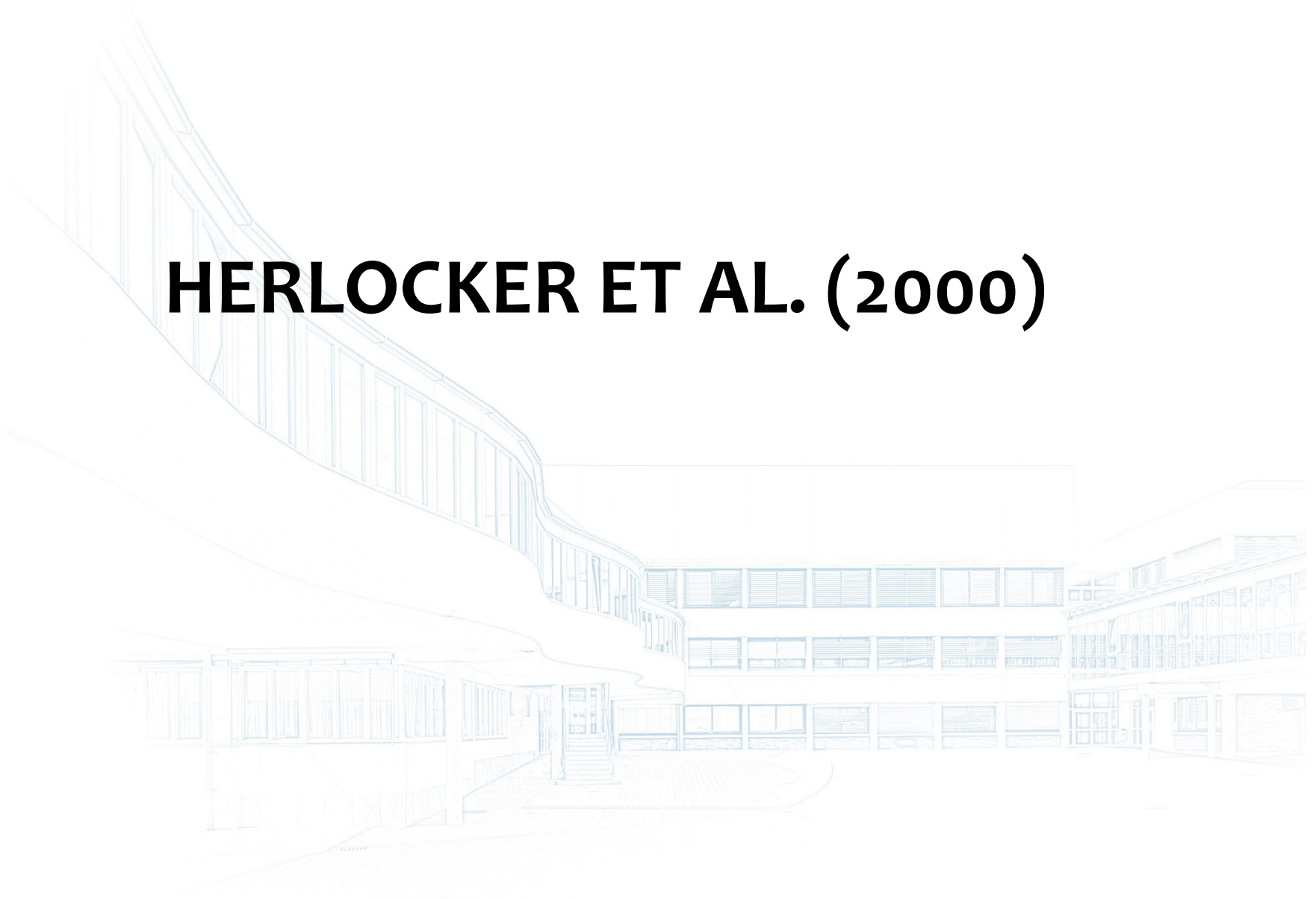


Nunes, I. and Jannach, D.: "A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems". User-Modeling and User-Adapted Interaction, Vol. 27(3-5). Springer, 2017, pp. 393-444

CASE STUDIES FROM “HISTORY” (2000-2014)



HERLOCKER ET AL. (2000)



Herlocker et al. (2000)

- Motivates explanations by transparency
 - Removing the black box
- Possible benefits:
 - Justification (user understanding)
 - User involvement
 - Education (about strength and limitations)
 - Acceptance (of the system as decision aide)

Herlocker et al. (2000)

- RQ 1:

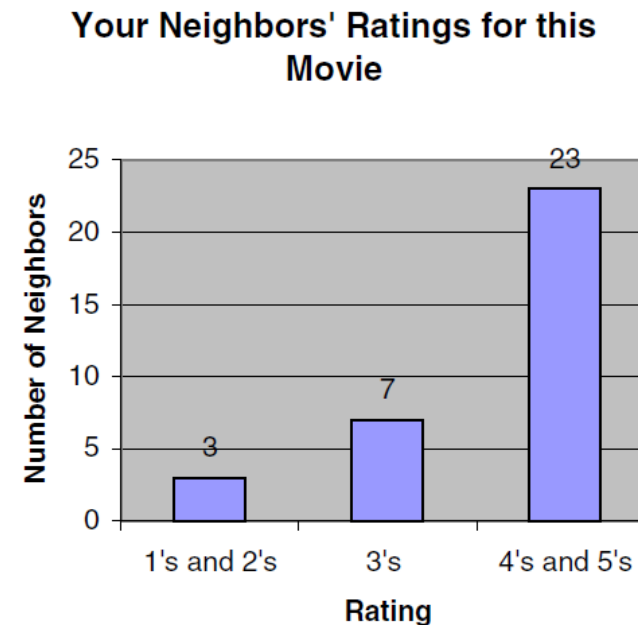
“What models and techniques are effective in supporting explanation in an ACF system?” (Persuasion)

- A user study was performed:

- 78 MovieLens users were shown 21 ways of explaining the (same) recommendation
- They were asked to rate (1-7) “**how likely they would be to go and see the movie**”
- Average responses were compared with base case with no explanations

Herlocker et al. (2000)

- Visual examples (underlying is a kNN Collaborative Filtering model)



Herlocker et al. (2000)

- Text-based examples:
 - Not using past ratings:

“MovieLens has predicted correctly for you 80% of the time in the past”

- Based on similarity:

“This movie is similar to 4 other movies that you rated 4 stars or higher.”

Herlocker et al.

- RQ1 Outcomes

- Histograms can be helpful, but it depends
- Second place is surprising
- Average ratings had a negative effect

#		N	Mean Response	Std Dev
1	Histogram with grouping	76	5.25	1.29
2	Past performance	77	5.19	1.16
3	Neighbor ratings histogram	78	5.09	1.22
4	Table of neighbors ratings	78	4.97	1.29
5	Similarity to other movies rated	77	4.97	1.50
6	Favorite actor or actress	76	4.92	1.73
7	MovieLens percent confidence in prediction	77	4.71	1.02
8	Won awards	76	4.67	1.49
9	Detailed process description	77	4.64	1.40
10	# neighbors	75	4.60	1.29
11	No extra data – focus on system	75	4.53	1.20
12	No extra data – focus on users	78	4.51	1.35
13	MovieLens confidence in prediction	77	4.51	1.20
14	Good profile	77	4.45	1.53
15	Overall percent rated 4+	75	4.37	1.26
16	Complex graph: count, ratings, similarity	74	4.36	1.47
17	Recommended by movie critics	76	4.21	1.47
18	Rating and %agreement of closest neighbor	77	4.21	1.20
19	# neighbors with std. deviation	78	4.19	1.45
20	# neighbors with avg correlation	76	4.08	1.46
21	Overall average rating	77	3.94	1.22

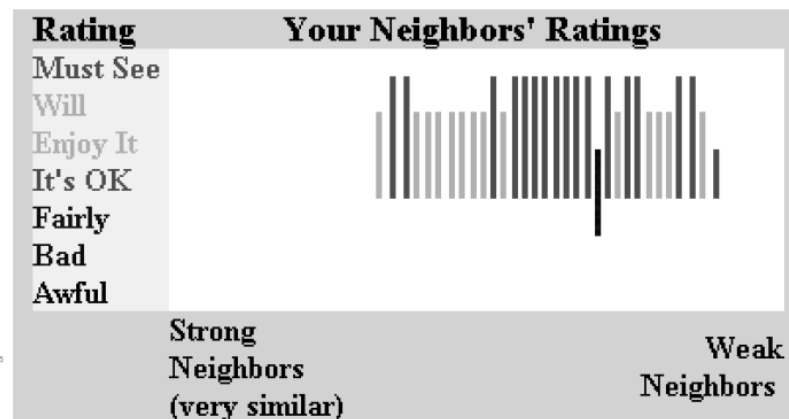
Herlocker et al. (2000)

- RQ2:
 - Can explanations improve acceptance of automated collaborative filtering (ACF) systems and
- RQ3:
 - Can explanations improve the filtering performance of users?

Herlocker et al. (2000)

- Another experimental study involving MovieLens users
 - Treatment saw explanation interfaces
 - e.g., this more complex one

Ratings for *Sixth Sense, The (1999)* by your Neighbors



Click on a bar to see that neighbor's profile!

Herlocker et al. (2000)

- Study design for RQ2/RQ3 (filtering perf.):
 - MovieLens users were asked to return to the system whenever they saw a new movie in real life
 - A mini-survey was done, including, e.g.,
 - Did you consult MovieLens before going?
 - ...
 - How much did MovieLens influence your decision?
 - Was the movie worth seeing?
- Main outcome
 - No statistical significant “filtering performance”
 - But 86% said that they would like to see the explanation interface added to MovieLens

BILGIC & MOONEY (2005)



Bilgic & Money (2005)

- Build in Herlocker et al.'s observations:
 - Explanation helps “promotion”, but not necessarily “satisfaction”
- Propose better explanation interfaces
 - Keyword style
 - Influence style
- Evaluate in book recommendation scenario
 - Based on hybrid CF/CB recommender

Bilgic & Money (2005)

- Keyword style

Slot	Word	Count	Strength	Explain
DESCRIPTION	HEART	2	94.14	Explain
DESCRIPTION	BEAUTIFUL	1	17.07	Explain
DESCRIPTION	MOTHER	3	11.55	Explain
DESCRIPTION	READ	14	10.63	Explain
DESCRIPTION	STORY	16	9.12	Explain

- Compares content features of recommended items with user profile features
- “Explain” goes one step further and lets users inspect where the values came from

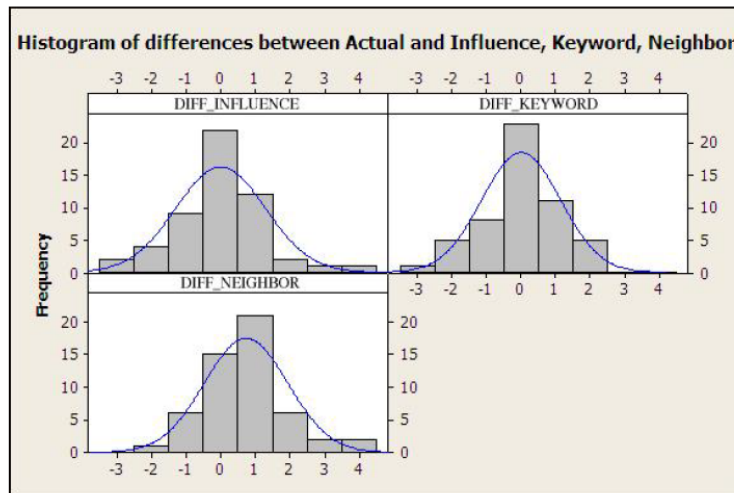
Bilgic & Money (2005)

- Experimental approach
- Assumption
 - *A good explanation will be effective in helping users understand if the recommendation is a good match for them*
- Procedure in user study

1. Get sample ratings from the user.
2. Compute a recommendation r .
3. For each explanation system e
 - 3.1 Present r to the user with e 's explanation.
 - 3.2 Ask the user to rate r
4. Ask the user to try r and then rate it again.

Bilgic & Money (2005)

- Results (34 subjects, 53 trials)



- Differences between actual and influenced ratings smallest for new explanation types
- Nearest-neighbor method from Herlocker et al. leads to *overestimate* of the quality of the recommended item

GEDIKLI ET AL. (2014)



Gedikli et al. (2014)

- Investigate the effects of explanations in **five dimensions** in parallel:
 - efficiency, effectiveness, persuasiveness, perceived transparency, and satisfaction
- Aim to obtain deeper understanding of **tag clouds** as an explanation mechanism
- Compare ten different explanation styles
 - Including personalized and non-personalized ones
 - Including several from Herlocker et al.
 - (Not including Keyword-Style Expl.; were tested earlier)

Gedikli et al. (2014)

- Personalized and non-personalized tag clouds



Tag-based explanations: Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: explaining recommendations using tags. In Proceedings of the 14th international conference on Intelligent user interfaces (IUI '09).

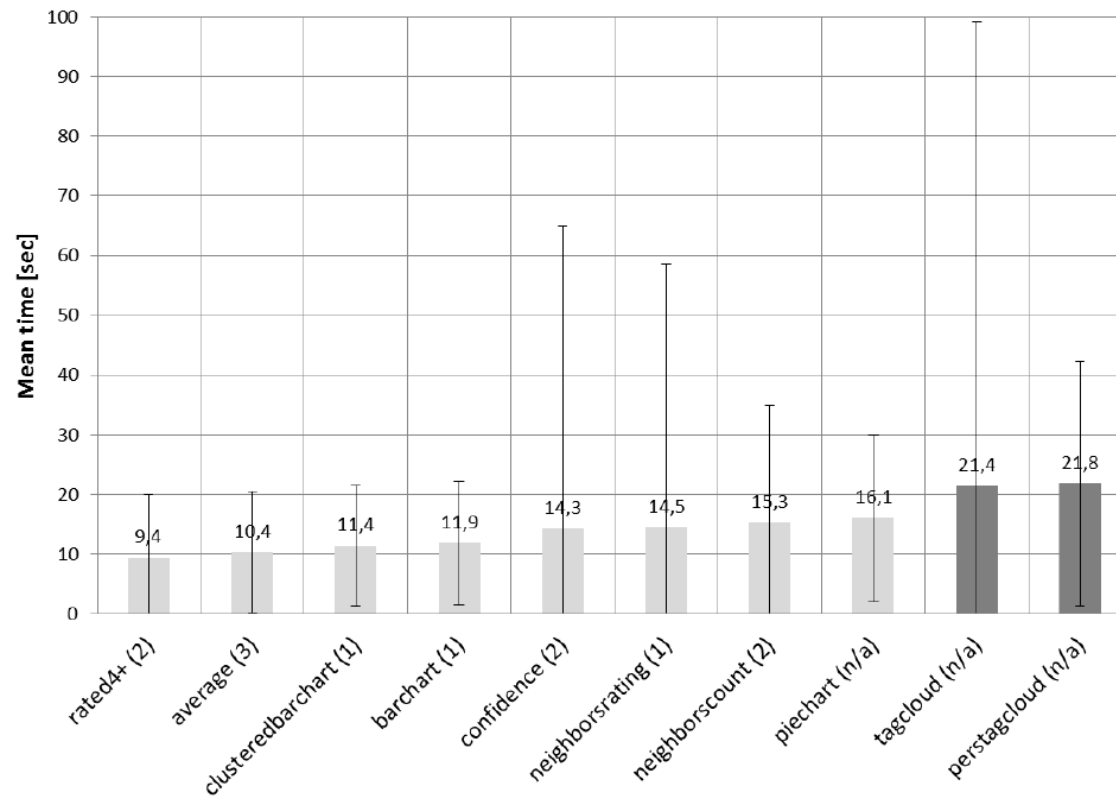
Personalized explanations were also studied earlier in: Tintarev, N., Masthoff, J. Evaluating the effectiveness of explanations for recommender systems. *User Model User-Adap Inter* 22, 399–439 (2012).

Gedikli et al. (2014)

- Experimental procedure, from Bilgic & Mooney:
 1. Let users assess item based on explanation
 2. Show details about item
 3. Let users rate the item again
 4. Ask users to rate the explanation interface
- Measurements
 - Satisfaction & Transparency: Self report
 - Efficiency: Time needed
 - Effectiveness and direction of persuasiveness: Difference between ratings
- Study size = 105

Gedikli et al. (2014)

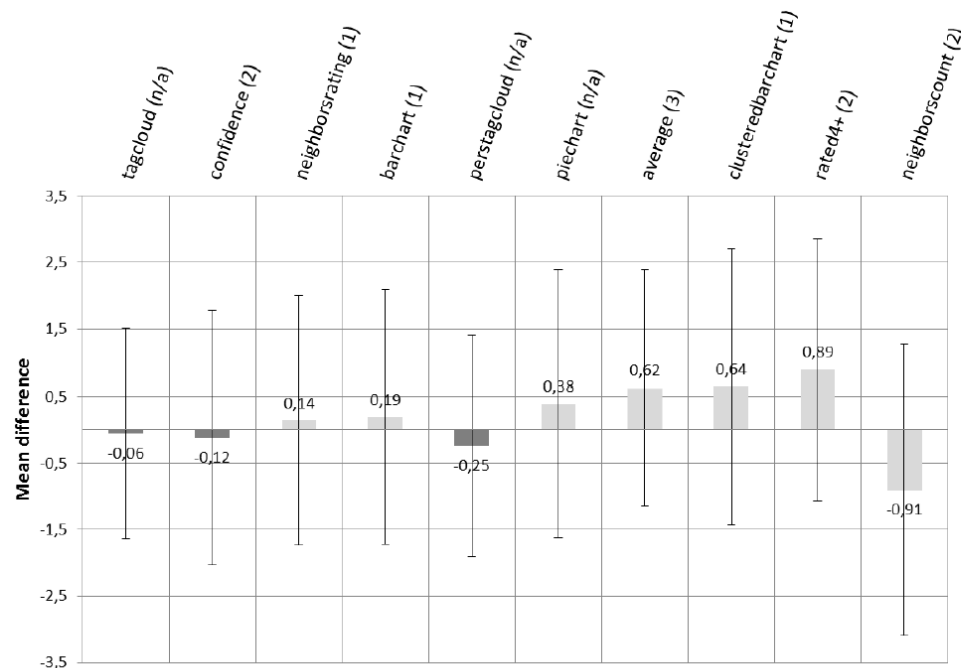
- Efficiency findings:
 - Processing tag clouds needs more time



Gedikli et al. (2014)

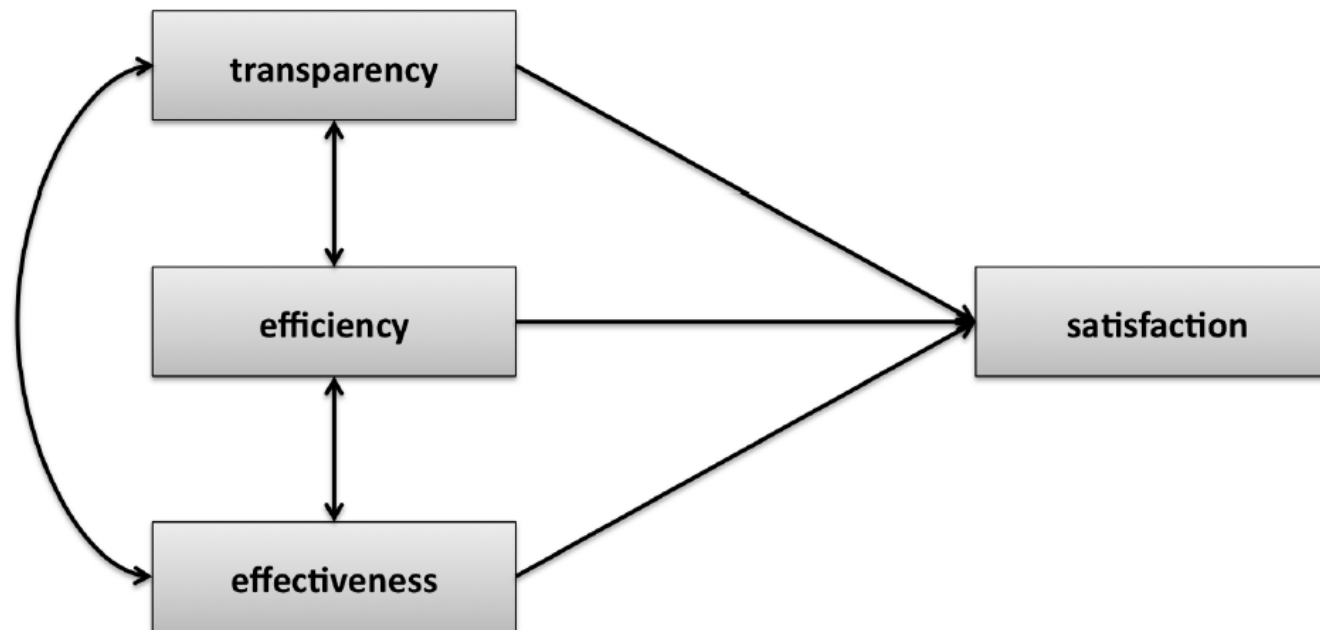
- Effectiveness

- Tag clouds lead to a good estimate; others sometimes lead to over- and underestimates
- Personalization not helpful, comp. Tintarev & Masthoff



Gedikli et al. (2014)

- Path analysis:
 - Transparency impacts satisfaction



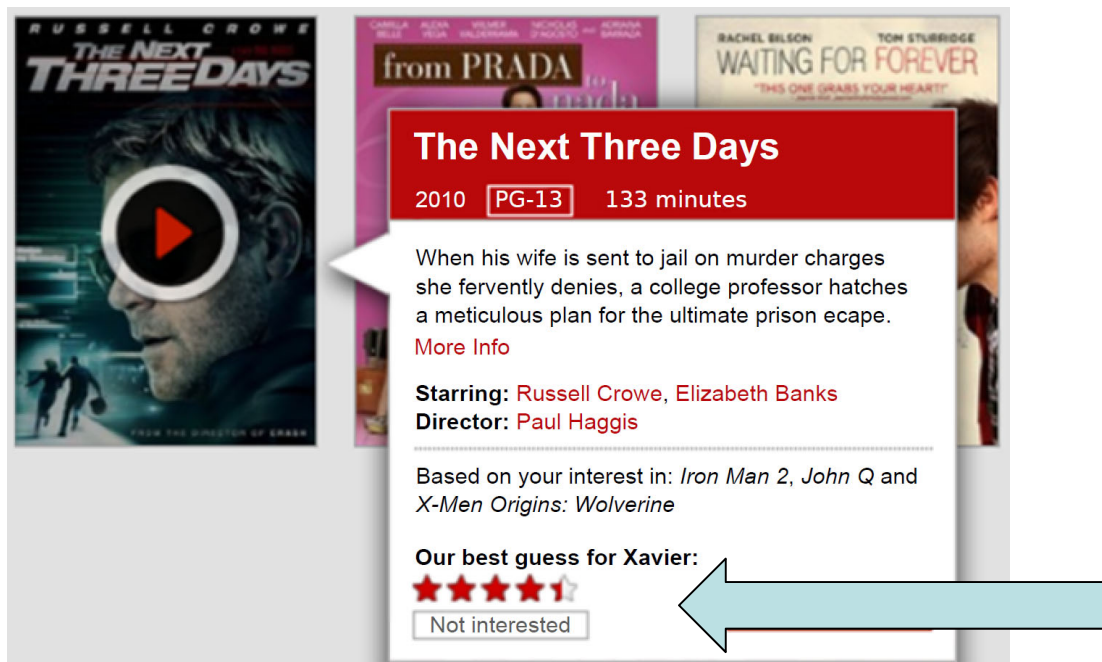
See also: Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (RecSys '11).

Commonalities of selected studies

- All were investigating with respect to one or more **explanation purposes**
- Evaluations were based on **user studies**
- All published in pre-neural times
 - Often using nearest-neighbor technique as baseline recommendation method

A multitude of proposals

- From very simple (industry) ...



The screenshot shows a movie recommendation card for "The Next Three Days" by Paul Haggis. The card features a play button icon on the movie poster, a red header with the title, and a white text box with a description and a "More Info" link. Below the text box, it lists the starring cast and director. At the bottom, there is a "Based on your interest in:" section with movie titles and a "Our best guess for Xavier:" section with a star rating and a "Not interested" button. A large light blue arrow points from the right towards the star rating.

The Next Three Days
2010 **PG-13** 133 minutes

When his wife is sent to jail on murder charges she fervently denies, a college professor hatches a meticulous plan for the ultimate prison escape.
[More Info](#)

Starring: Russell Crowe, Elizabeth Banks
Director: Paul Haggis

Based on your interest in: *Iron Man 2*, *John Q* and *X-Men Origins: Wolverine*

Our best guess for Xavier:
★★★★☆

A multitude of proposals

- to a bit more complex simple (industry) ...

The screenshot displays the TripAdvisor page for Clontarf Castle Hotel. The header includes the TripAdvisor logo, the location 'Clontarf Castle Hotel Reviews, Dublin', and user information 'Hi, Barry'. The navigation bar lists various travel services like Dublin, Hotels, Flights, etc. The main content area shows the hotel's name, a 4.5-star rating based on 1,985 reviews, and its ranking as the #20 of 174 hotels in Dublin. A 'Certificate of Excellence' badge is also present. Below this, there are location details and a link to 'Hotel amenities'. The central part of the page features a comparison chart titled 'Reasons for you to choose this hotel:' and 'Reasons for you to avoid this hotel:'. The chart compares the hotel's performance in various categories against alternatives. A photo gallery is visible on the right, showing a night view of the hotel's entrance. At the bottom, there are several tags: 'Family-friendly', 'Luxury', 'Best Value', and 'Free Wifi'. A small text box at the bottom left explains that the comparison is based on user reviews.

Reasons for you to choose this hotel:

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

Reasons for you to avoid this hotel:

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

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

... bar with a great atmosphere ...
... Enjoy a drink in the lovely relaxing lounge ...
... don't miss the music in the bar area ...


Traveller photos 1224
Professional photos
Browse nearby

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

A multitude of proposals

- to even more complex (academia) ...

Content-based image retrieval at the end of the early years



Journal Paper

transactions | feature types | salient points | system architectures | retrieval research | content-based retrieval | computer vision | image | machine | content-based image | december | ieee | early years | image retrieval | display space | image databases | salient features | retrieval | ieee transactions | pattern analysis | analysis | vol | information retrieval | broad domain | machine intelligence | weak segmentation | query space | sensory gap | semantic gap | interactive session | retrieval systems | pattern | image processing | narrow domains |

OVERALL RANKING fair

coverage
relevance
similarity

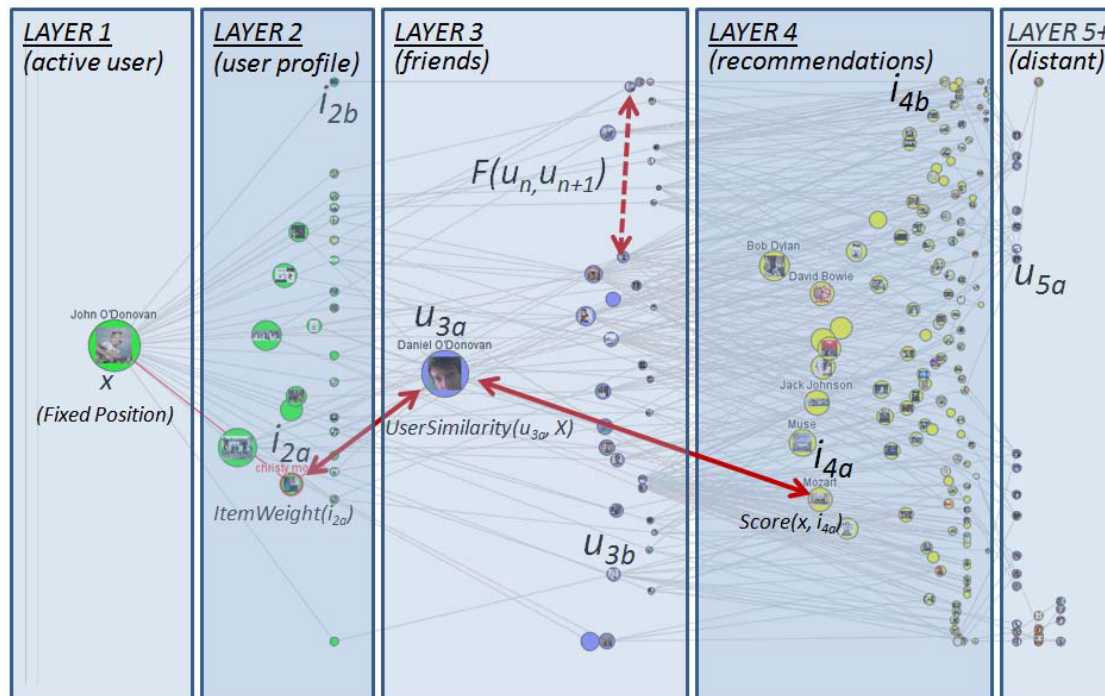
TITLE Content-based image retrieval at the end of the early years

AUTHOR Arnold W. M. Smeulders and Marcel Worring and Simone Santini and Amarnath Gupta and

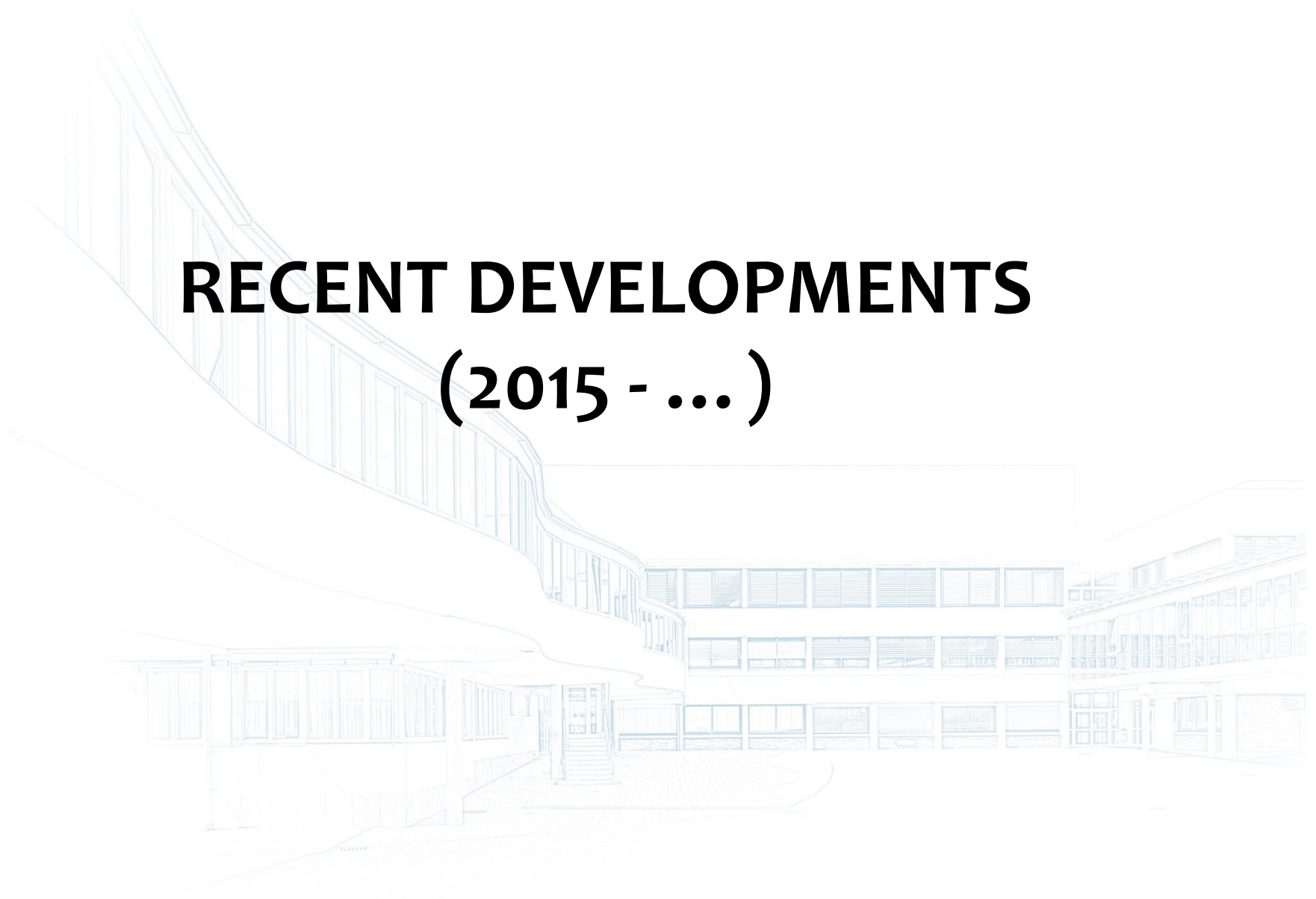
Topics of your interest included in this paper			Other topics included in this paper	
TOPIC	IMPORTANCE	DELETE	TOPIC	INCLUDE
feedback	<input type="range"/>	<input type="checkbox"/>	transactions	<input type="checkbox"/>
content-base	<input type="range"/>	<input type="checkbox"/>	feature types	<input type="checkbox"/>
intelligence	<input type="range"/>	<input type="checkbox"/>	salient points	<input type="checkbox"/>
information	<input type="range"/>	<input type="checkbox"/>	system architectures	<input type="checkbox"/>

A multitude of proposals

- to ultimate transparency (academia)



RECENT DEVELOPMENTS (2015 - ...)

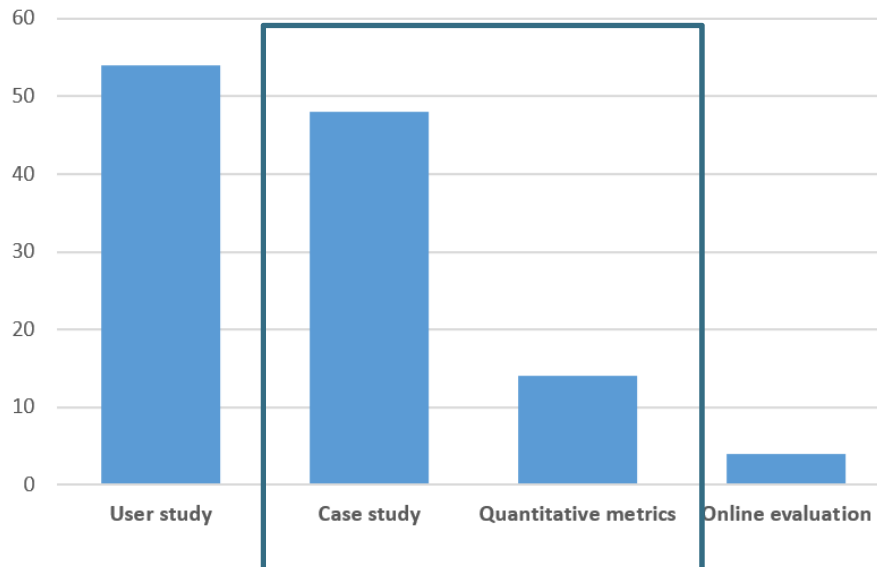


Recent Developments

- Deep Learning Booms
 - Super-charged algorithm research in recommender systems
 - Significant advances in natural language processing
 - Generation of text-based explanations, conversational systems
- Explainable AI
 - AI recognized as also being potentially harmful
 - Desire for transparency
- Policy & Legal Aspects
 - *General Data Protection Regulation, Digital Services Act*

Chen et al. 2022

- Recent survey on the evaluation of explanations
 - 100 papers, mostly very recent ones
 - Statistics: Applied methodology



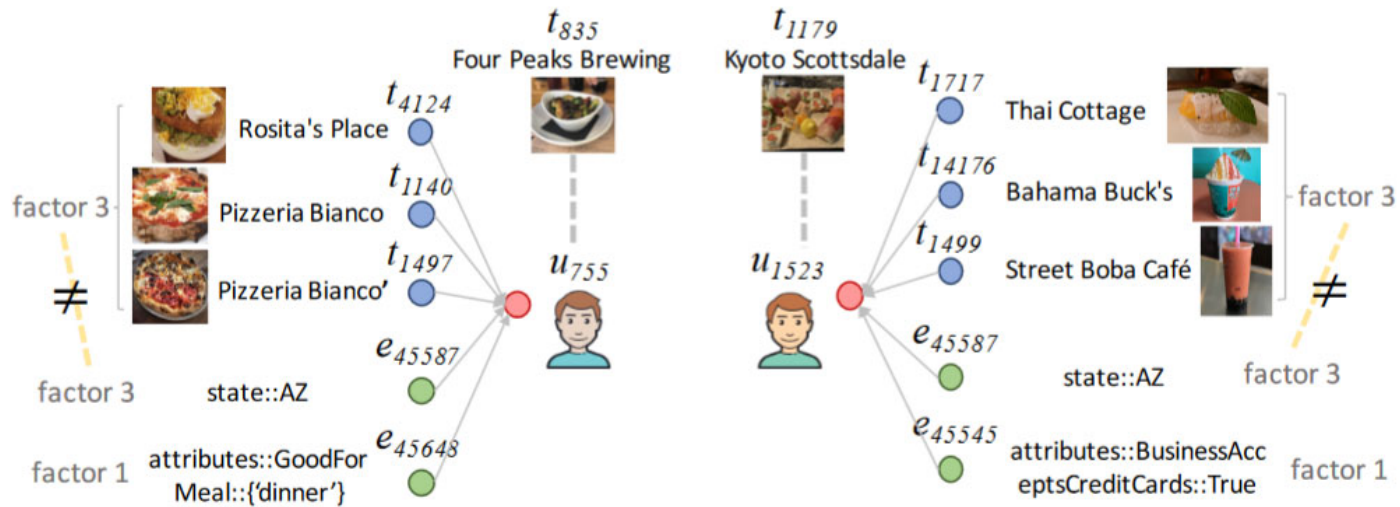
Chen, X., Zhang, Y. and Wen J-R., “Measuring “Why” in Recommender Systems: a Comprehensive Survey on the Evaluation of Explainable Recommendation, <https://arxiv.org/abs/2202.06466v1>

Case Studies

- Common for in-depth algorithmic papers
- Often more on “explainability/interpretability”
 - Not designed for end users
 - But may be the basis for explanation interfaces for end users
 - Usually not evaluated for or with end users
 - Left to demonstrate that this information is truly helpful
 - Usually also **no experimental setting** when human judges are involved
 - No treatment and control, no random assignment, no information if judges are representative, how recruited, ...

Case Studies

- Relations



Computational Metrics

- Various forms used, see Chen et al., 2022
 - Focus on generated natural language explanations
- Often consider user reviews as ground truth
 - e.g. reviews from websites such as TripAdvisor
- Metrics, e.g.,
 - BLEU and ROUGE, keyword/feature overlap
 - Partially taken from machine translation
 - Compare generated explanations with ground truth
 - Diversity metrics, specialized metrics

Computational Metrics

- Assumption of one single ground truth
 - But there be many good explanations
- Also
 - To what extent are explanations personalized?
 - What about grammatical aspects?
- Generally
 - How can be sure that these metrics correspond to **human perceptions**?
 - How can we be sure that the explanations serve any of the potential **purposes**?

Computational Metrics

- Insights from the evaluation of Dialog Systems, EMNLP 2016
 - Sometimes limited correspondence of metrics w. perceptions

How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation

**Chia-Wei Liu^{1*}, Ryan Lowe^{1*}, Iulian V. Serban^{2*}, Michael Noseworthy^{1*},
Laurent Charlin¹, Joelle Pineau¹**

¹ School of Computer Science, McGill University

{chia-wei.liu, ryan.lowe, michael.noseworthy}@mail.mcgill.ca

{lcharlin, jpineau}@cs.mcgill.ca

² DIRO, Université de Montréal

iulian.vlad.serban@umontreal.ca

Conversational Recommenders

- Recently interest and progress
 - Advances in NLP
 - Advances in deep learning
 - Often based on recorded dialogs between humans
- A “natural” test-bed for explanations
 - However, most of today’s system cannot answer why-questions
 - The underlying dialog datasets often do not contain explanations

What next?

- More studies on end user perceptions
 - Also considering the smaller details, e.g., the optimal length of natural language explanations
- Considering also questions of user control
 - Explanations are often an entry point to let users give feedback on explanations or “correct” the system
- Avoid “abstraction traps”
 - By overly relying on unvalidated/unsuitable computational metrics

What next?

- Explanation is to a large extent a problem of **human computer interaction**
- We must build on existing insights to provide effective explanations



ELSEVIER

Artificial Intelligence
Volume 267, February 2019, Pages 1-38



Explanation in artificial intelligence: Insights from the social sciences

Tim Miller 

Show more 

Social Sciences

- Inmates running the asylum:
 - “... *leaving decisions about what constitutes a good explanation of complex decision-making models to the experts who understand these models the best is likely to result in failure in many cases.*”
- Small study by Miller et al.: 17 XAI papers
 - Only four references social sciences research
 - Only one tried to build a model on this

Finally, a reality check

- What are the real use cases and goals?

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

- Transparency, Trust, Fairness, Persuasion, ...?

Finally, a reality check

- Who of you has ever considered Amazon's explanations?
 - To understand,
 - to exert control
- In which applications do we **really** want or need them?
- Why isn't industry using it **more**?

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

Add to Cart Add to Wish List

Rate this item

★★★★★

I own it

Not interested

Because you purchased...



Mad Max: Fury Road [Blu-ray] (Blu-ray)
DVD ~ Charlize Theron

★★★★★

Don't use for recommendations

Future Direction: Explanations and User Control

- Showing explanations may not be the end of the interaction
- There are many more exciting questions!
- Requires data science / HCI collaboration

Thank you!

- Time for questions
- Contact
 - dietmar.jannach@aau.at

