

# Conversational Recommender Systems

## Tutorial at IJCAI 2022, Vienna

Dietmar Jannach, University of Klagenfurt, Austria

[dietmar.jannach@aau.at](mailto:dietmar.jannach@aau.at)

Markus Zanker, Free University of Bozen, Italy

[markus.zanker@unibz.it](mailto:markus.zanker@unibz.it)

# Welcome

---

- Today's presenters
- Markus Zanker
  - Professor at University of Bozen/Bolzano and University of Klagenfurt, Austria
- Dietmar Jannach
  - Professor at University of Klagenfurt



# Recommender Systems

---

- A big success in industry



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

# Towards Deep Conversational Recommendation (NeurIPS '19)

---

**OURS:** hi, how can i help you ?

**SEEKER:** 2001 : a space odyssey might be a great option. anything else that you would suggest ?

**OURS:** star wars : the force awakens is also a good one return of the jedi all good movies

**SEEKER:** those sound good ! i 'm going to look into those movies.

**OURS:** i hope i was able to help you find a good movie to watch

**SEEKER:** thank you for your help ! have a great night ! good bye

# Agenda

---

1. Introduction
2. Characterization of Conversational Recommender System (CRS)
3. Interaction Modalities
4. Underlying Knowledge and Data
5. Computational Tasks
6. Evaluation of CRS
7. Wrap-up & Discussion

# Basic Reading

---

- Jannach, D., Manzoor, A., Cai, W. and Chen, L.: "A Survey on Conversational Recommender Systems". ACM Computing Surveys, 2021
- Gao, C., Lei, W., He, X., de Rijke, M., Chua, T.: "Advances and Challenges in Conversational Recommender Systems: A Survey". AI Open, 2/2021
- Jannach, D.: "Evaluating Conversational Recommender Systems". Artificial Intelligence Review, 2022

# Agenda

---

1. Introduction
2. Characterization of Conversational Recommender System (CRS)
3. Interaction Modalities
4. Underlying Knowledge and Data
5. Computational Tasks
6. Evaluation of CRS
7. Wrap-up & Discussion

# Already some history (1979)

---

I'D LIKE TO KNOW WHAT SORT OF PERSON YOU THINK YOU ARE. PLEASE TYPE IN A FEW SINGLE WORDS THAT YOU THINK CHARACTERIZE THE MOST IMPORTANT ASPECTS OF YOURSELF

unconventional open direct honest humorous persistent adventurous

HAVE YOU EVER READ EARTH ANGELS BY SUSAN CAHILL?

no

YOU MIGHT LIKE IT.

IT IS ABOUT A GIRL WHO GOES TO CATHOLIC ELEMENTARY SCHOOL, HIGH SCHOOL, AND COLLEGE, THEN GOES ON TO BECOME A NUN, BUT FINALLY BACKS OUT A FEW DAYS BEFORE FINAL VOWS.

DOES THAT SOUND GOOD?

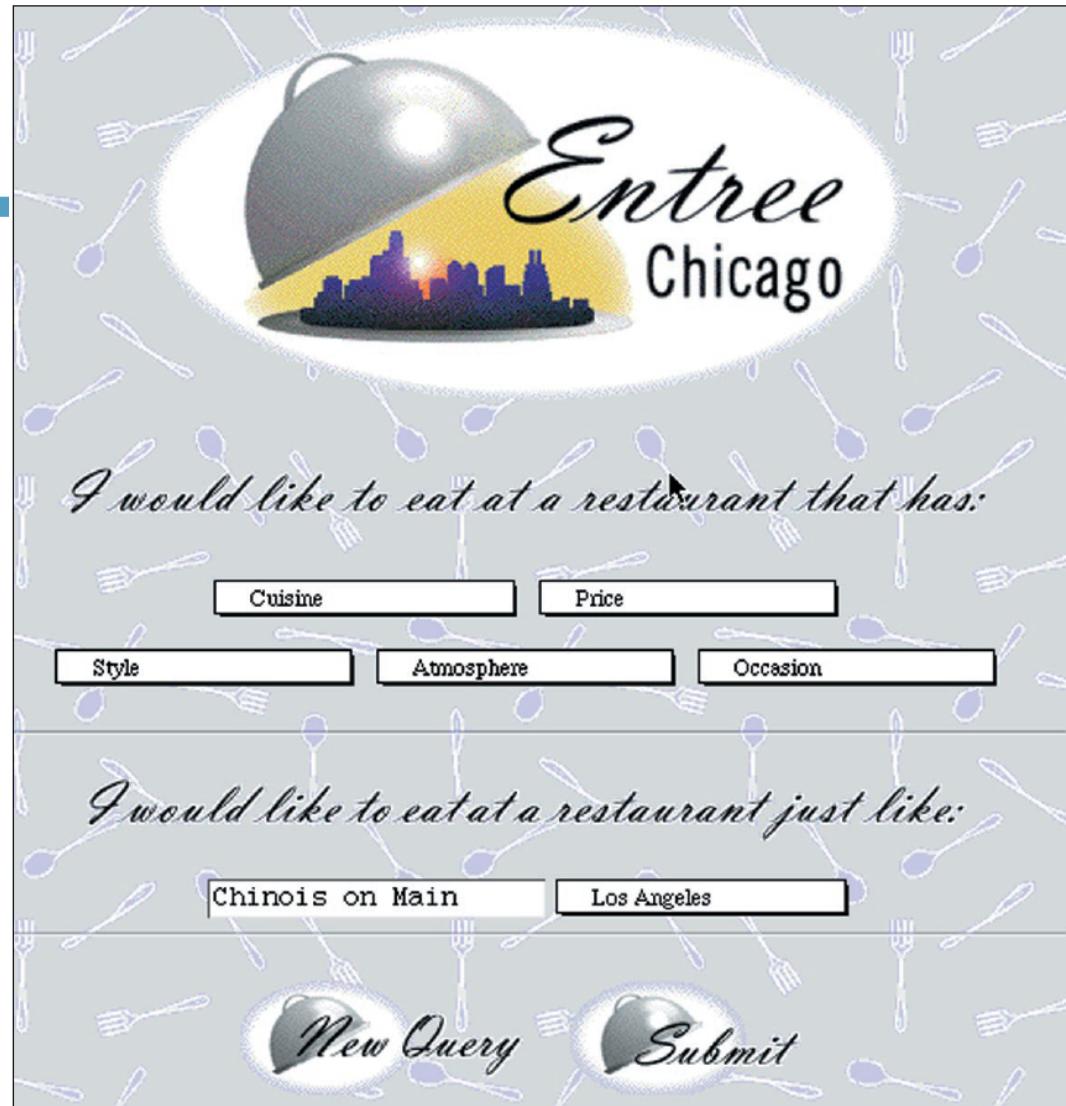
no

WAS IT BECAUSE OF THE DEPTH OF THE CHARACTER DEVELOPMENT? (EITHER TOO MUCH OR NOT ENOUGH)

no

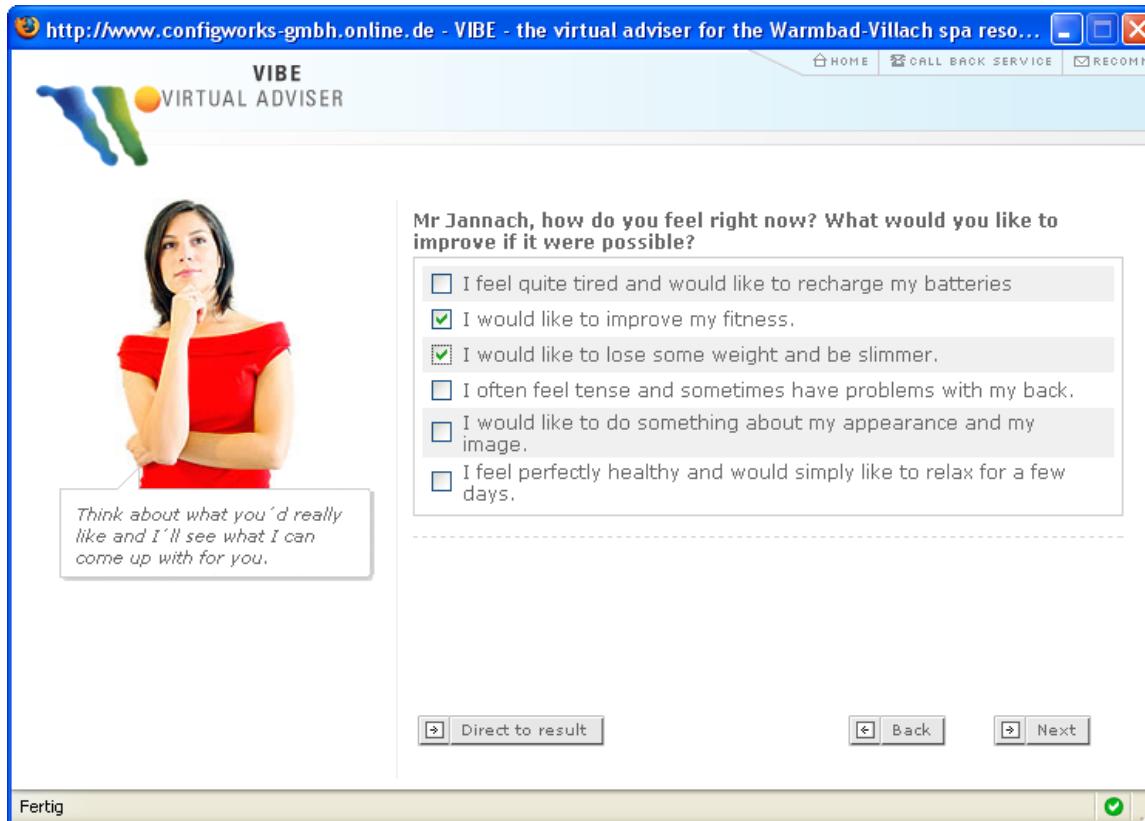
1997

The FindMe  
System:  
“Critiquing”



R. D. Burke, K. J. Hammond, and B. C. Young. The FindMe approach to assisted browsing. IEEE Expert, 12(4):32–40, 1997

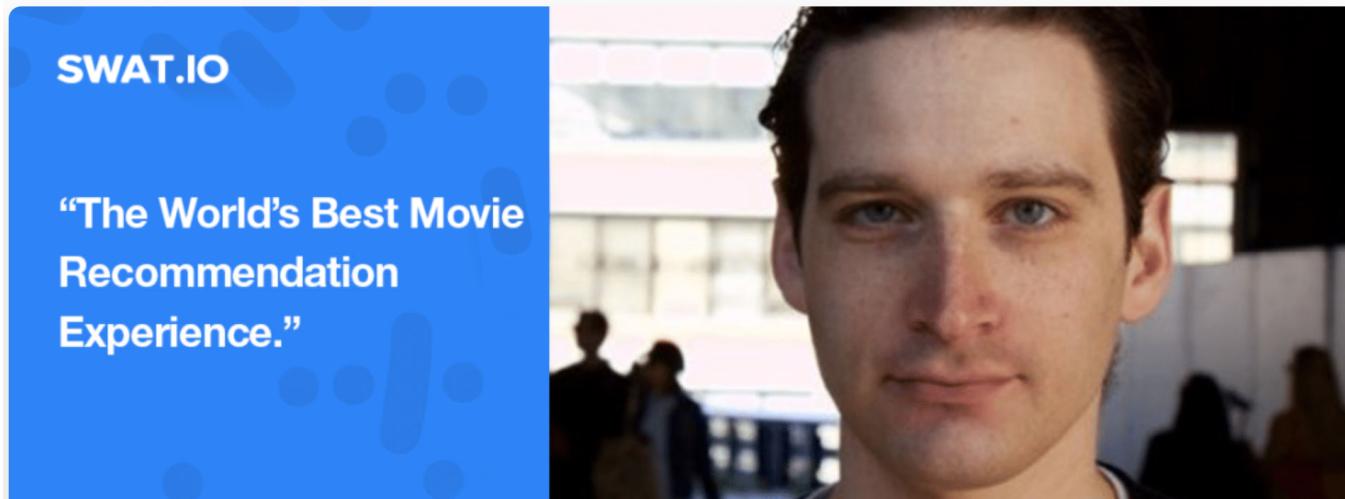
# Dynamic form-based Interactions (early 2000s)



D. Jannach. ADVISOR SUITE – A knowledge-based sales advisory system. In ECAI '04, pages 720–724, 2004.

# In the real world in the early 2020s

---



# And Chill

---

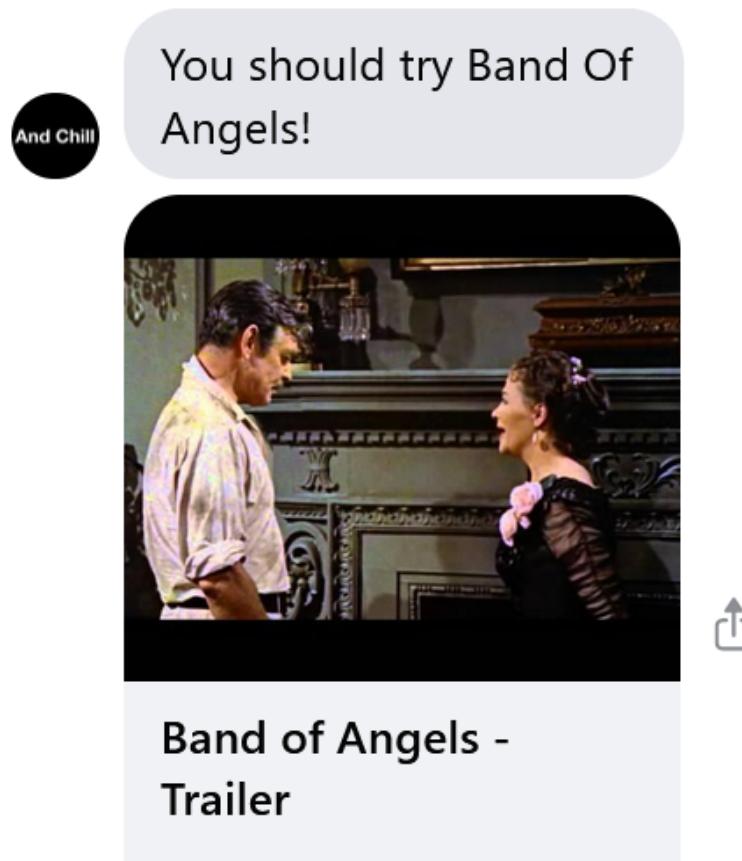


OK, Let's go! The floor  
is yours. Give me ONE  
movie you like, and  
why you like it.

I like rebecca, it is wiht a  
lot of suspense, black and  
white hitchcock-like

# And Chill

---



# And Chill

---

why do you recommend  
band of angels

Gotcha... that title,  
right? There's only  
1.8% of users that have  
mentioned this one. I  
am still generating AI-  
powered movie recs  
based on this film. Can  
you tell me about a  
different one you like?



And Chill

# Towards Deep Conversational Recommendation - Supplemental

---

SEEKER: yes i loved that movie as well . are there any good movies without dinosaurs ?

OURS: i have n't seen that one . i will have to check it out . have you seen jurassic park

SEEKER: hi

OURS: what kind of movies do you like ? what kind of movies do you like ?

# A Grand Challenge

|  | DeepCRS   | KBRD      |
|--|-----------|-----------|
| Number of dialogues                        | 70        | 70        |
| Generated sentences (overall)              | 399       | 359       |
| Sentences labeled as meaningful            | 277 (69%) | 209 (58%) |
| Sentences labeled as <i>not</i> meaningful | 122 (31%) | 150 (42%) |
| Dialogues without problems                 | 5         | 5         |
| Chit-chat sentences                        | 132       | 88        |
| Chit-chat labeled as meaningful            | 112 (85%) | 77 (87%)  |
| Number of recommendations                  | 106       | 119       |
| Recs. labeled as meaningful                | 63 (60%)  | 66 (55%)  |
| Nb. dialogues with no meaningful recs.     | 25 (36%)  | 20 (28%)  |
| Nb. dialogues with no rec. made.           | 7 (10%)   | 6 (8.5%)  |

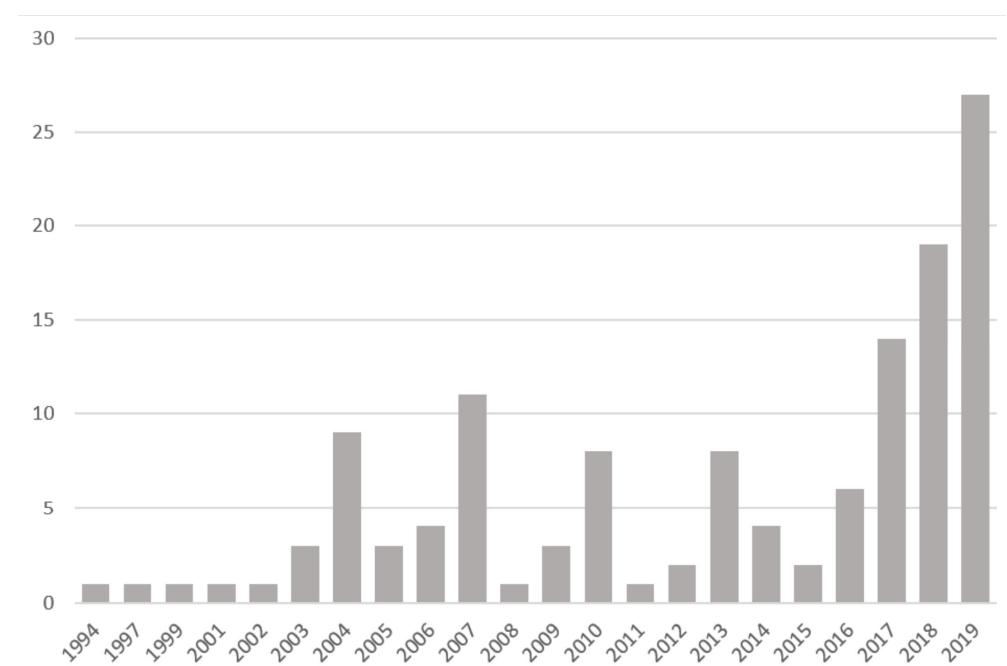
Table 2: Analysis of Dialogue and Recommendation Quality

Manzoor, A., Jannach, D.: Conversational Recommendation based on End-to-end Learning: How Far Are We? Computers and Human Behavior Reports, 2021

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

# Growing Research Interest

- Advances in speech-enabled devices
- Advances in natural language processing
- Advances in machine learning in general



Number of papers on CRS per year found in "A Survey on Conversational Recommender Systems". ACM Computing Surveys, 2021

# Agenda

---

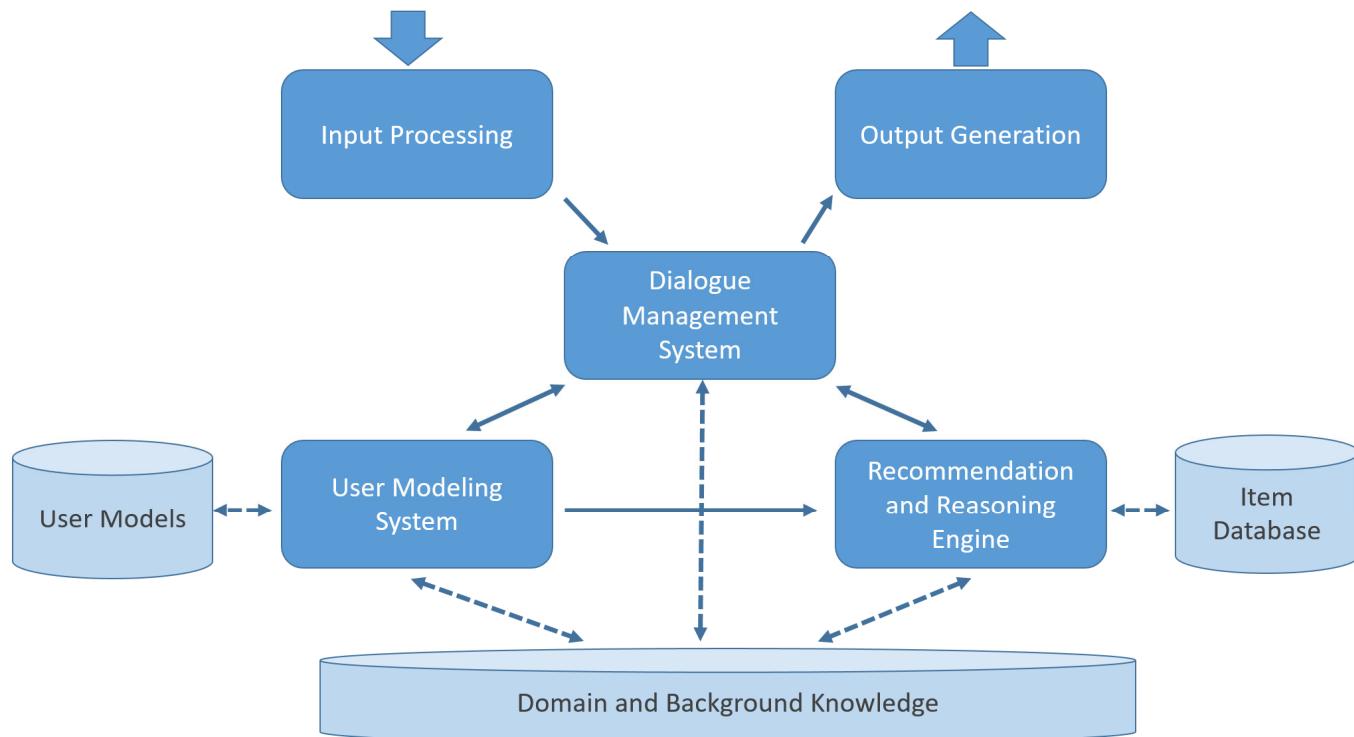
1. Introduction
2. Characterization of Conversational Recommender System (CRS)
3. Interaction Modalities
4. Underlying Knowledge and Data
- Short break --
5. Computational Tasks
6. Evaluation of CRS
7. Wrap-up & Discussion

# Characterization of a CRS

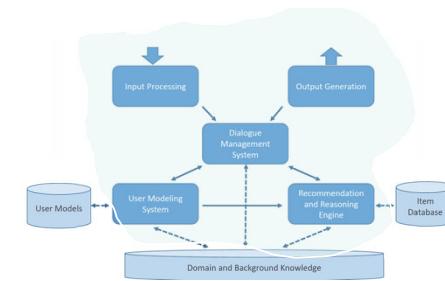
---

- “A CRS is a software system that supports its users in achieving recommendation-related *goals* through a *multi-turn dialogue*.”  
(Jannach et al., CSUR 2021)
- Characteristics
  - Task orientation
    - Recommendation, decision support, preference elicitation, explanation
  - Multi-turn interaction
    - Compare: Q&A systems, Apple’s Siri
    - Requires form of *state management*
  - Not limited to certain modality, e.g., natural language

# Conceptual Architecture

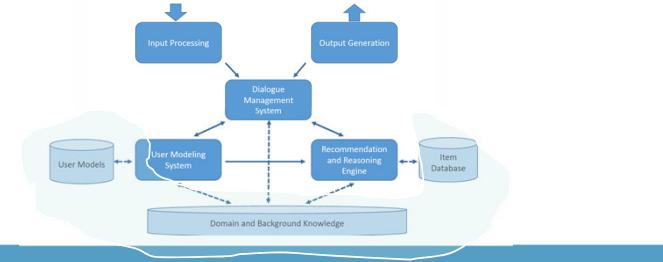


# Conceptual Architecture



- Computational Elements
  - Dialogue Management (state tracker)
  - User Modeling System
  - Recommendation and Reasoning Engine
    - May include the generation of explanations or the decision about the next conversational move (including intent recognition)
  - Input and Output Processing
    - Speech-to-text conversion, voice output, named entity recognition, ...

# Conceptual Architecture



- Knowledge Elements
  - Item Database
    - Catalog of recommendable items, including meta-data
  - Domain & Background Knowledge
    - Dialogue Knowledge
      - Possible states and transitions, supported intents
    - World Knowledge
      - e.g., from DBpedia
    - Recorded dialogues for learning

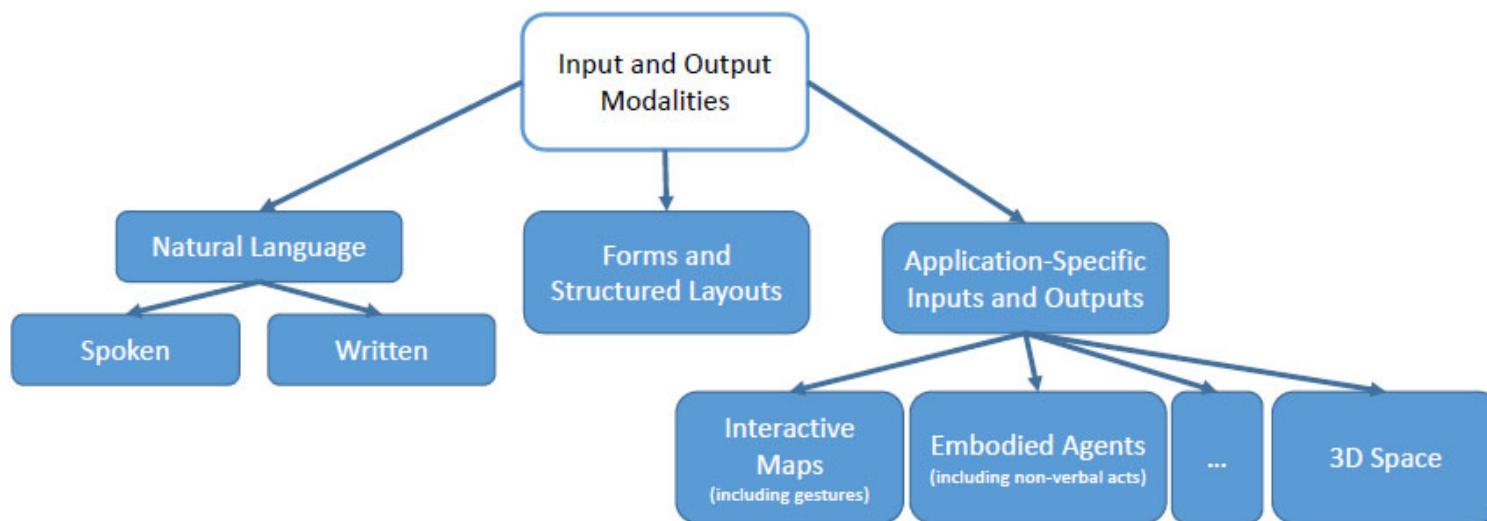
# Agenda

---

1. Introduction
2. Characterization of Conversational Recommender System (CRS)
- 3. Interaction Modalities**
4. Underlying Knowledge and Data
5. Computational Tasks
6. Evaluation of CRS
7. Wrap-up & Discussion

# Interaction Modalities

---



# Single Modality or Combined

---

- Forms (e.g., buttons, radio-buttons) and structured layouts
  - Common for critiquing-based CRS and web-based interactive advisory systems
- Natural language based interaction
  - Task-oriented dialogue systems and more recent (deep) learning-based approaches
  - Spoken-text-only approaches are often implemented on smart speakers (e.g., Amazon Alexa or Google Home)
- Hybrid approach
  - Combining natural language with other modalities

# Example: FindMe

UKRAINIAN VILLAGE. TWO bedroom rehab garden apartment. Lr, Eurokitchen, hwfl, excellent security, forced air, lots of closets, laundry in building. Garage space included. Dogs OK.  
Available immediately. \$600/ mo. 312-489-1554. / ;

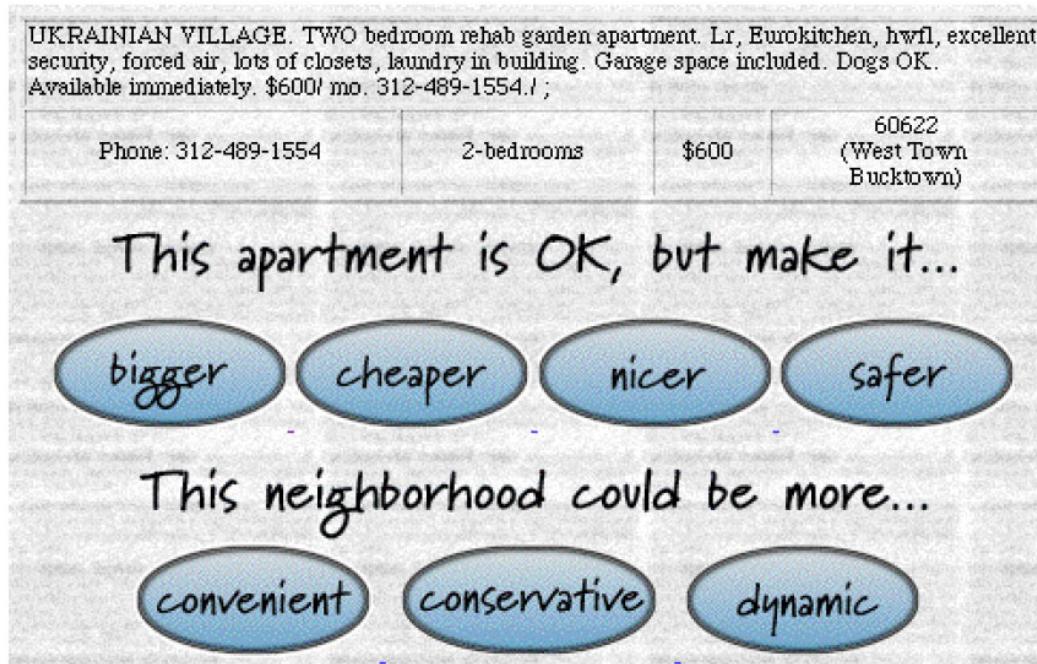
|                     |            |       |                                  |
|---------------------|------------|-------|----------------------------------|
| Phone: 312-489-1554 | 2-bedrooms | \$600 | 60622<br>(West Town<br>Bucktown) |
|---------------------|------------|-------|----------------------------------|

This apartment is OK, but make it...

bigger      cheaper      nicer      safer

This neighborhood could be more...

convenient      conservative      dynamic



Burke, R., Hammond, K., and Young, B.: The FindMe Approach to Assisted Browsing. IEEE Expert: Intelligent Systems and Their Applications 12(4):32-40, 1997.

# Example: Dynamic Critiquing

QUIKSHOP.COM

Home | About This Project | Contact

>> Digital Cameras

Shop For: > Digital Cameras > Computers > Holidays



**Product Found: Canon EOS 30**

6.3 Megapixel CMOS sensor  
7-point wide-area AF  
High-performance DIGIC processor  
100-1600 ISO speed range  
Compatible with all Canon EF lenses and EX Speedlites  
PictBridge, Canon Direct Print and Bubble Jet Direct compatible - no PC required

**Adjust your preferences to find the right camera for you**

|                |   |              |   |
|----------------|---|--------------|---|
| Manufacturer   | X | Canon        | X |
| Optical Zoom   | ↓ | 7X           | ↑ |
| Memory (MB)    | ↓ | 512          | ↑ |
| Weight (Grams) | ↓ | 780          | ↑ |
| Resolution     | ↓ | 6.2 M Pixels | ↑ |
| Size           | X | Large        | X |
| Case           | X | Magnesium    | X |
| Price          | ↓ | 995          | ↑ |

**Explain:**

**1. Less Memory and Lower Resolution and Cheaper**

This Critique covers 153 other Digital Cameras

**Less Memory**  
Current Value: 512 MB  
Critique: Less Than  
Remaining: (0 to 256 MB)

**Lower Resolution**  
Current Value: 6.2 M Pixels  
Critique: Less Than  
Remaining: (1.4 to 5.9 M Pixels)

**Cheaper**  
Current Value: 995 €  
Critique: Less Than  
Remaining: (75€ to 980€)

We have more matching cameras with the following:

|   |         |      |
|---|---------|------|
| 1. Less Memory and Lower Resolution and Cheaper     | EXPLAIN | PICK |
| 2. Different Manufacturer and Less Zoom and Lighter | EXPLAIN | PICK |
| 3. Lighter and Smaller and Different Case           | EXPLAIN | PICK |

I've found the Camera I want! 

No lets start again 

PICK

McCarthy, K., Reilly, J., McGinty, L., and Smyth, B.: Experiments in Dynamic Critiquing. In Proceedings of the Tenth International Conference on Intelligent User Interfaces, 175-182, New York: ACM Press, 2005.

# Example: Example Critiquing

The figure consists of two side-by-side screenshots of user interfaces for critiquing products.

**Left Screenshot (Apartment Comparison):**

- Title:** Compare
- Text:** Would you like to compare  
Apt 34: room in a house, 600 frs, 15 square meters, private bathroom, private kitchen, 15 minutes to your work place
- Section:** with other apartments for
  - Better Type
  - Cheaper Price
  - Bigger Area
  - Better Bathroom
  - Better Kitchen
  - Closer Distance
- Section:** You are willing to compromise on the following attributes:
  - Type of Apartment
  - Price
  - Area
  - Bathroom
  - Kitchen
  - Distance
- Buttons:** Cancel, Show Results

**Right Screenshot (Product Comparison for Canon PowerShot S2 IS Digital Camera):**

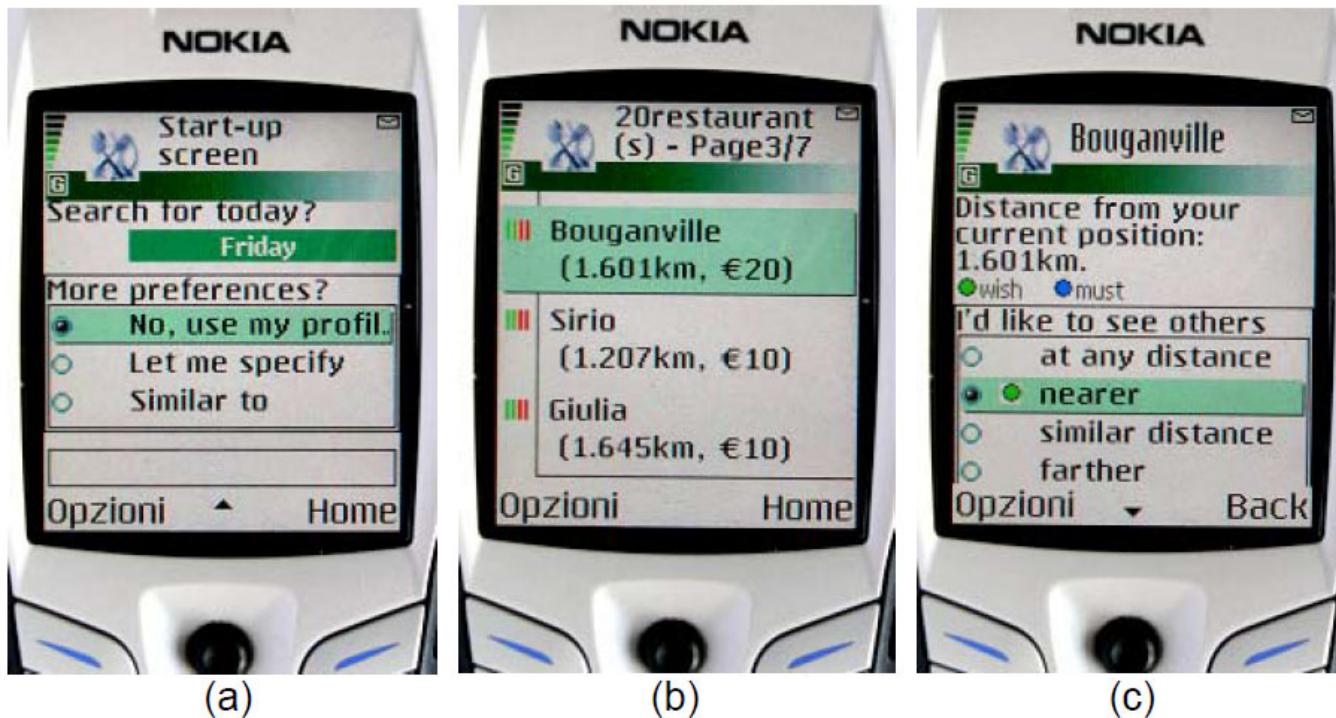
- Title:** To find similar products with better values than this one
- Image:** Canon PowerShot S2 IS Digital Camera
- Text:** Canon PowerShot S2 IS Digital Camera Add to saved list  
\$424.15  
Canon, 5.3 M pixels, 12x optical zoom, 16 MB memory, 1.8 in screen size, 2.97 in thickness, 404.7 g weight. [detail](#)
- Section:** would you like to improve some values?

|                        | Keep  | Improve   | Take any suggestion   |
|------------------------|---|---|-----------------------|
| Manufacturer           | <input checked="" type="radio"/> Canon        | <input type="radio"/> Sony                      | <input type="radio"/> |
| Price                  | <input type="radio"/> \$424.15                | <input checked="" type="radio"/> less expensive | <input type="radio"/> |
| Resolution             | <input checked="" type="radio"/> 5.3 M pixels | <input type="radio"/> \$100 cheaper             | <input type="radio"/> |
| Optical Zoom           | <input checked="" type="radio"/> 12x          | <input type="radio"/> \$200 cheaper             | <input type="radio"/> |
| Removable Flash Memory | <input checked="" type="radio"/> 16 MB        | <input type="radio"/> \$300 cheaper             | <input type="radio"/> |
| LCD Screen Size        | <input checked="" type="radio"/> 1.8 in       | <input type="radio"/> more memory               | <input type="radio"/> |
| Thickness              | <input checked="" type="radio"/> 2.97 in      | <input type="radio"/> larger                    | <input type="radio"/> |
| Weight                 | <input checked="" type="radio"/> 404.7 g      | <input type="radio"/> thinner                   | <input type="radio"/> |
|                        |   | <input type="radio"/> lighter                   | <input type="radio"/> |
- Buttons:** Show Results, Reset

Pu, P. and Chen, L.: Integrating Tradeoff Support in Product Search Tools for E-Commerce Sites. In Proceeding of ACM Conference on Electronic Commerce (EC'05), pages 269-278, Vancouver, Canada, June 5-8, 2005.

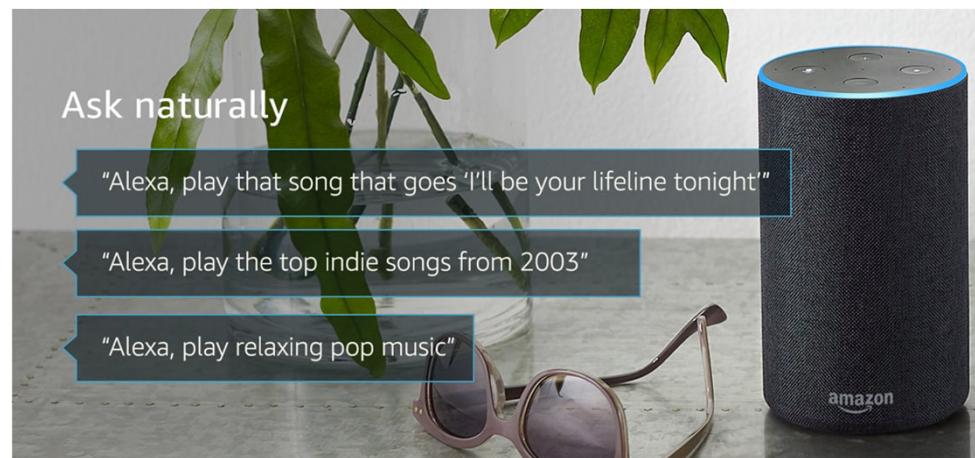
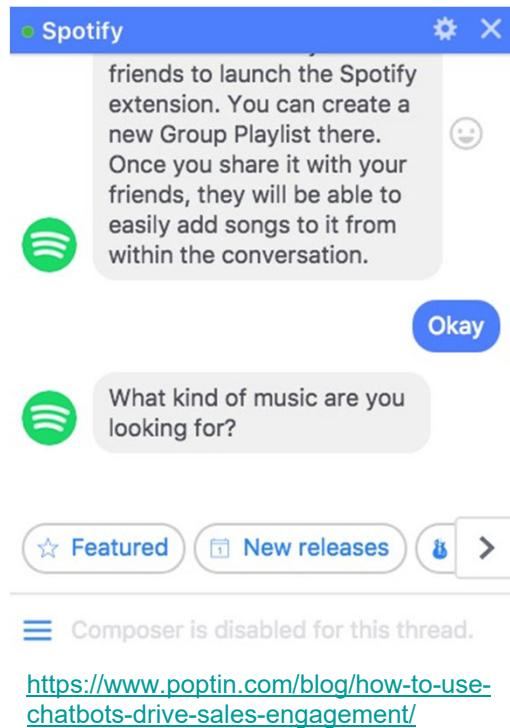
Pu, P. and Chen, L.: Evaluating Critiquing-based Recommender Agents. In Proceedings of Twenty-first National Conference on Artificial Intelligence (AAAI'06), pages 157-162, Boston, USA, July 16-20, 2006.

# Example: Mobile Critiquing

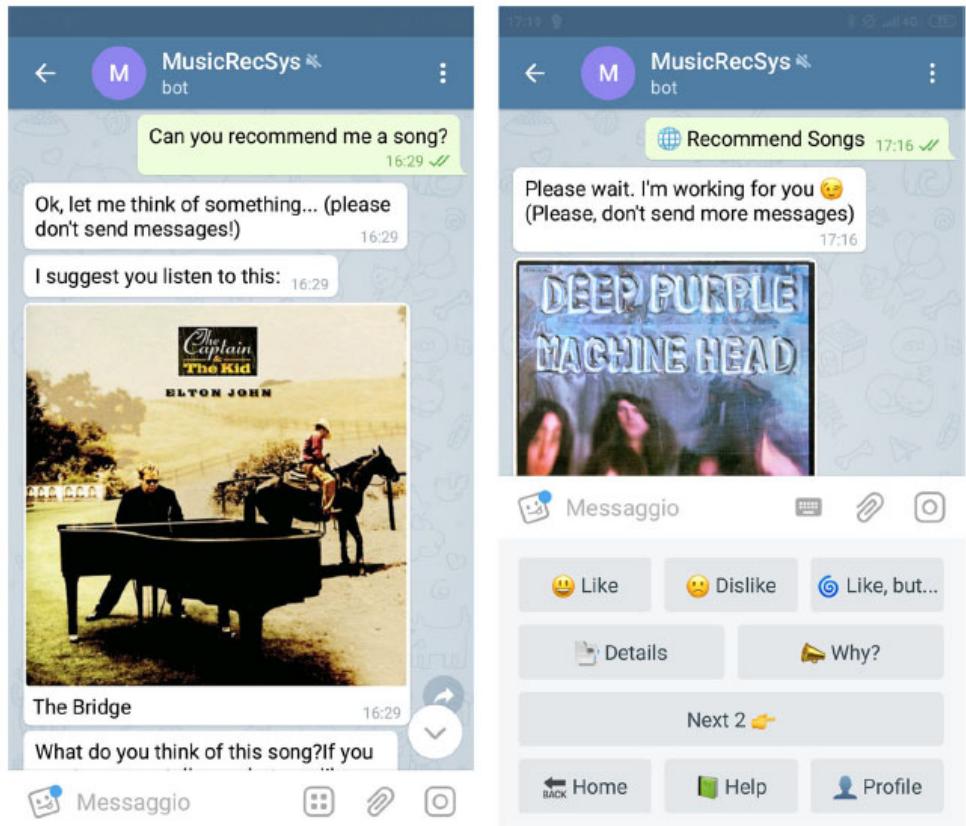


Ricci, F. and Nguyen, Q. N.: Acquiring and Revising Preferences in a Critique-Based Mobile Recommender System,  
in IEEE Intelligent Systems, vol. 22, no. 3, pp. 22-29, May-June 2007.

# Natural Language Interaction



# Combined Modalities



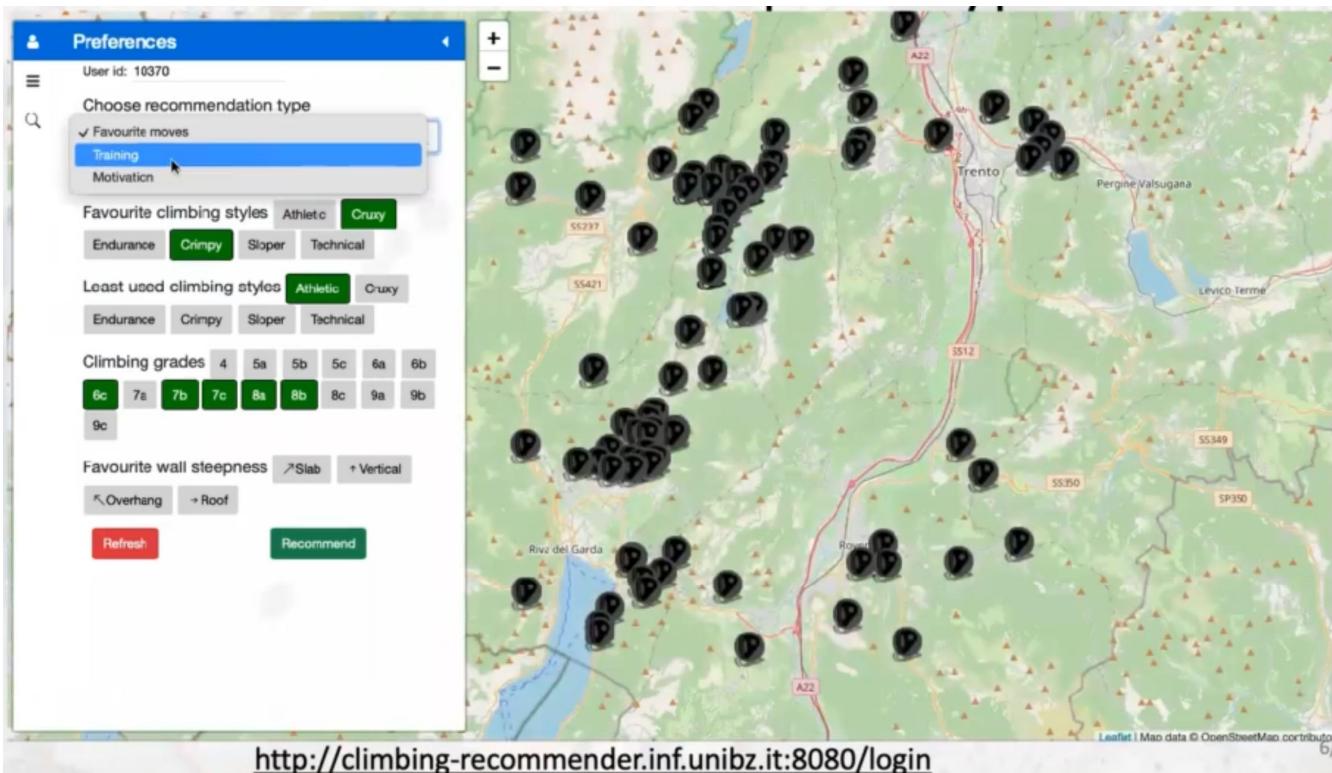
Narducci, F., Basile, P., de Gemmis, M., Lops, P., and Semeraro, G.: An investigation on the user interaction modes of conversational recommender systems for the music domain. UMUAI '19, pages 1–34, 2019.

# Application-specific Modalities

---

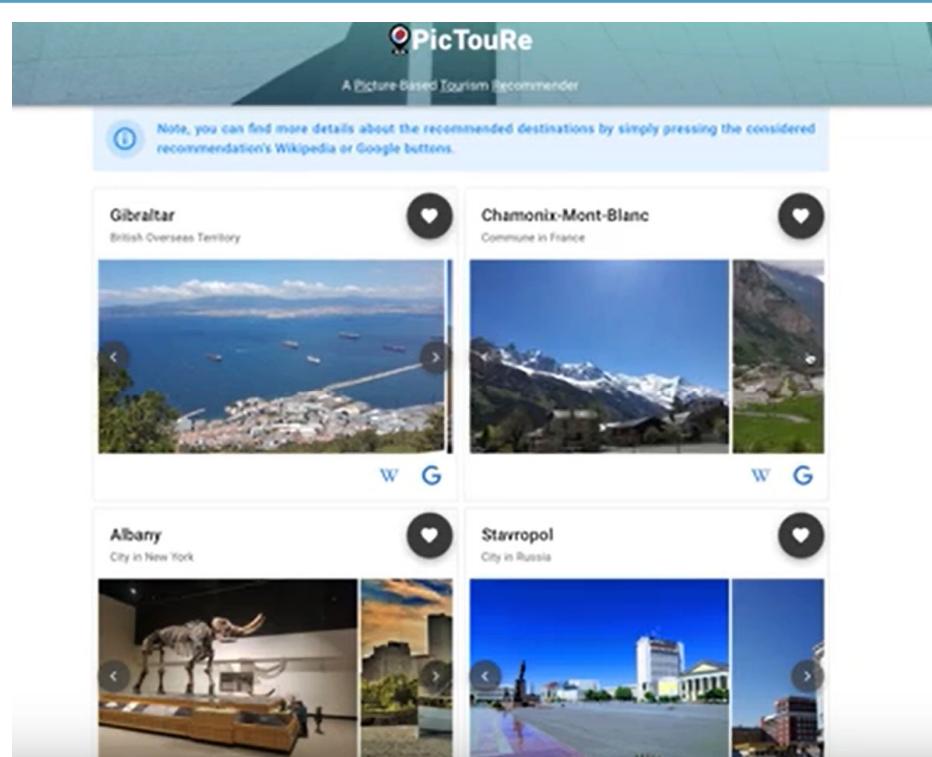
- Visual inputs on geographic map
- Pen gestures
- Handwritten input
- Body postures
- Gestures
- Facial expressions
- Speech prosody
- Etc.

# Example: Climbing recommendations



Iustina Alekseevna Ivanova, Attaullah Burio, and Francesco Ricci. Map and Content-Based Climbing Recommender System. In Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22 Adjunct), pp. 41–45, 2022.

# Example: Picture-based



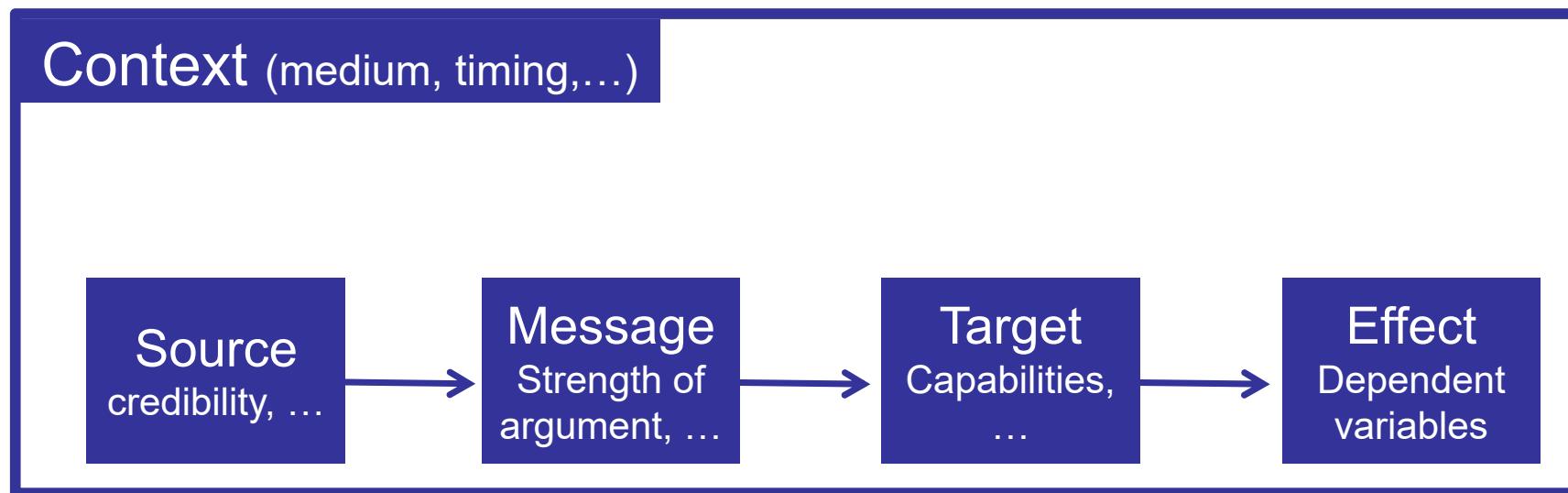
Mete Sertkan, Julia Neidhardt, and Hannes Werthner. 2020. PicTouRe - A Picture-Based Tourism Recommender. Fourteenth ACM Conference on Recommender Systems, pp. 597–599.

# Example: Mixed-mode disease identification

The screenshot shows the DSSApple web application interface. At the top, there is a navigation bar with the logo 'DSSApple' (an apple with a magnifying glass), 'Glossary', 'About', and a language selector ('UK'). Below the navigation bar, there are two horizontal tabs: 'Outer symptoms' (highlighted in green) and 'Inner symptoms'. A series of photographs of apples with various lesions are displayed under these tabs. On the right side, there is a question: 'Do you see any halo surrounding the lesion on the apple?' followed by three radio button options: 'Yes', 'No', and 'I don't know'. Below the question, there are two close-up photographs of apple lesions with blue arrows pointing to specific features. At the bottom, there are navigation buttons for 'Back' and 'Next'.

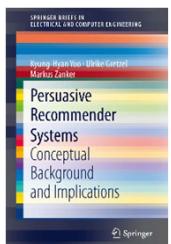
Gabriele Sottocornola, Sanja Baric, Maximilian Nocker, Fabio Stella, Markus Zanker:  
Picture-based and conversational decision support to diagnose post-harvest apple diseases. Expert Syst. Appl. 189:  
116052, 2022.

# Complex view: Persuasive RS



- Persuasion is generally seen as the *intended inducing* of another person to *believe something, to do something or to change attitudes, mood and behaviour*

Kyung-Hyan Yoo, Ulrike Gretzel & Markus Zanker, Persuasive Recommender Systems Conceptual Background and Implications.  
Springer, 2013.



# Embodied Conversational Agents

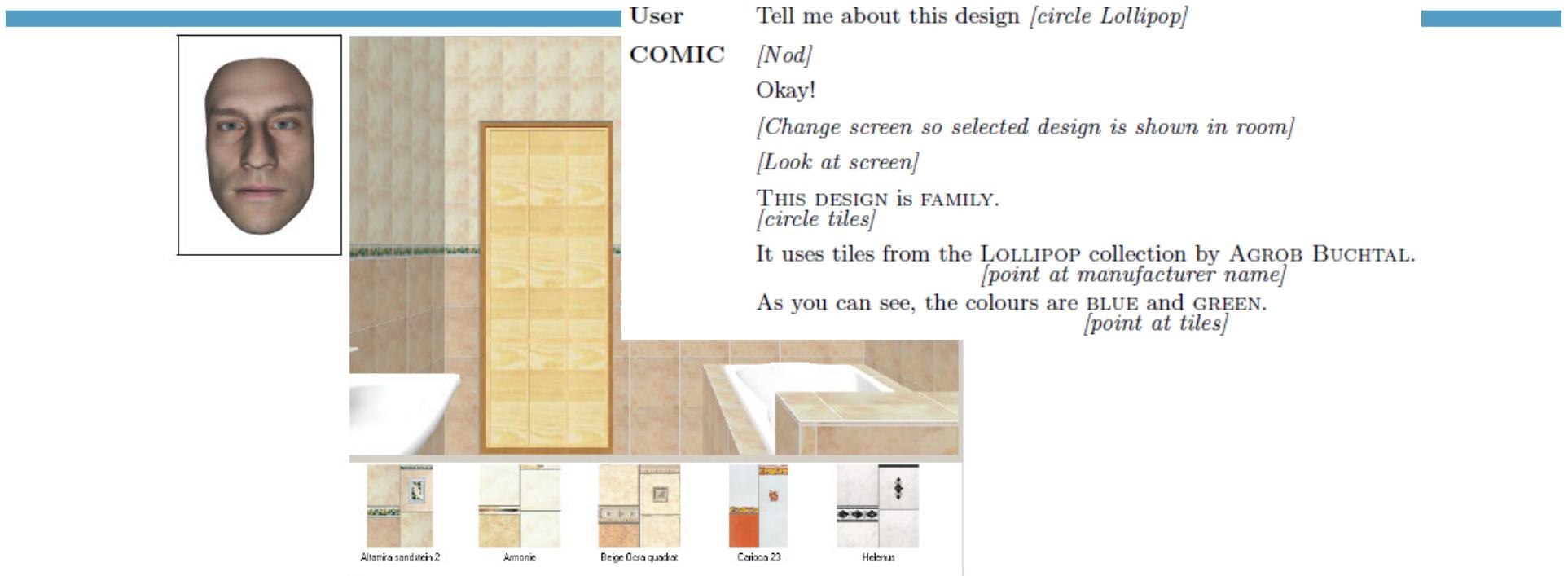


Figure 1. Components of the COMIC system (tile-browsing phase)

Foster, M. E., and Oberlander, J.: User preferences can drive facial expressions: Evaluating an embodied conversational agent in a recommender dialogue system. *User Modeling and User-Adapted Interaction*, 20(4):341–381, 2010.

# Virtual 3-D Space



(a) Three users are currently working individually, each one with its own panel ( $t_1$  in Figure 4).



(b) Tester 2 likes the product on Tester 3's panel and presses "I like" button. Tester 3 and Tester 2 are collaborating and they use chat messages for deciding a critique ( $t_2$  in Figure 4).



Fig. 3: Screenshot of the recommendation panel.

Contreras, D., Salamo, M., Rodriguez, I., and Puig, A.: Shopping decisions made in a virtual world: Defining a state-based model of collaborative and conversational user-recommender interactions. *IEEE Consumer Electronics Magazine*, 7(4): 260–35, 2018.

# Application Environment

---

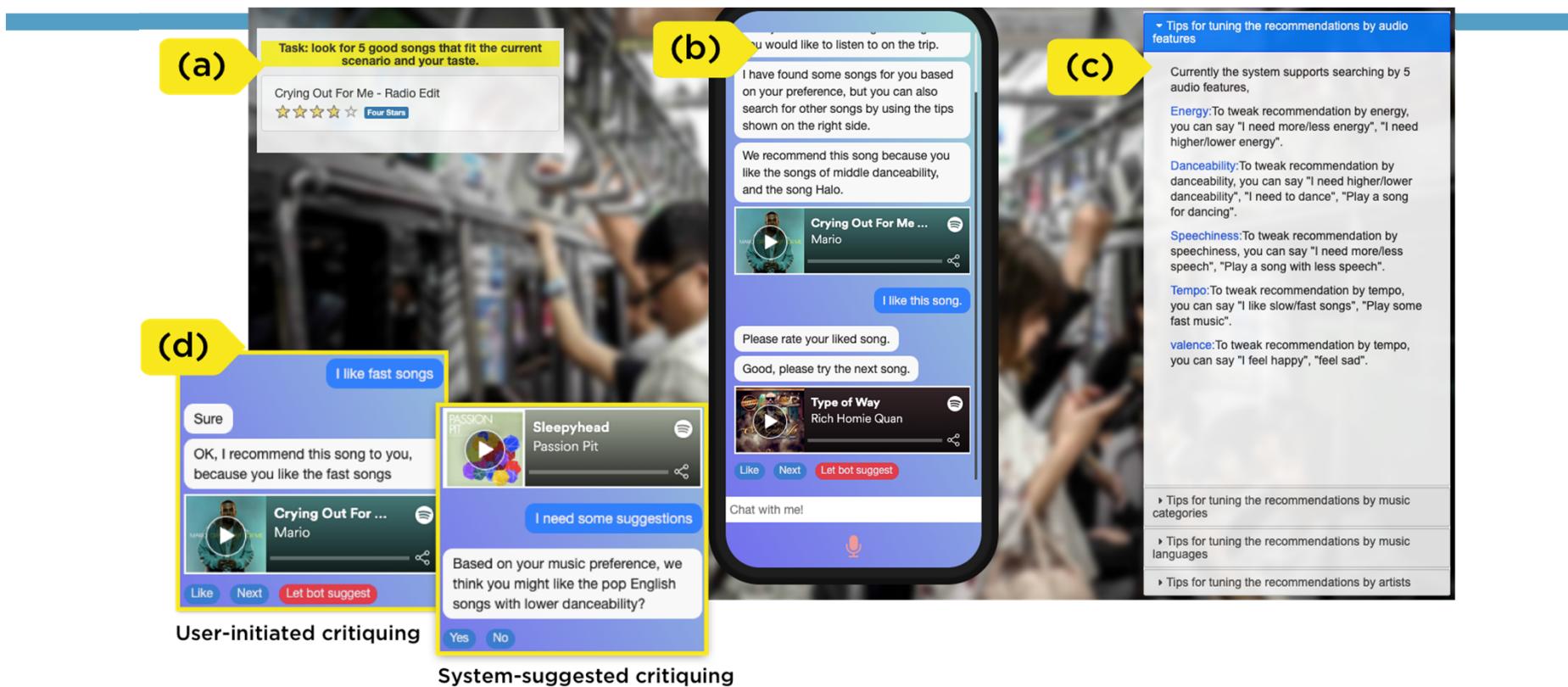
- Stand-alone vs. Embedded
  - Stand-alone
    - Recommendation is the central functionality of the system
    - E.g., mobile tourist guides, interactive e-commerce advisory systems
  - Embedded
    - Providing recommendations is only one of many functionalities the device is capable of
    - E.g., the use of a CRS on voice-based home assistants

# Interaction Initiative

---

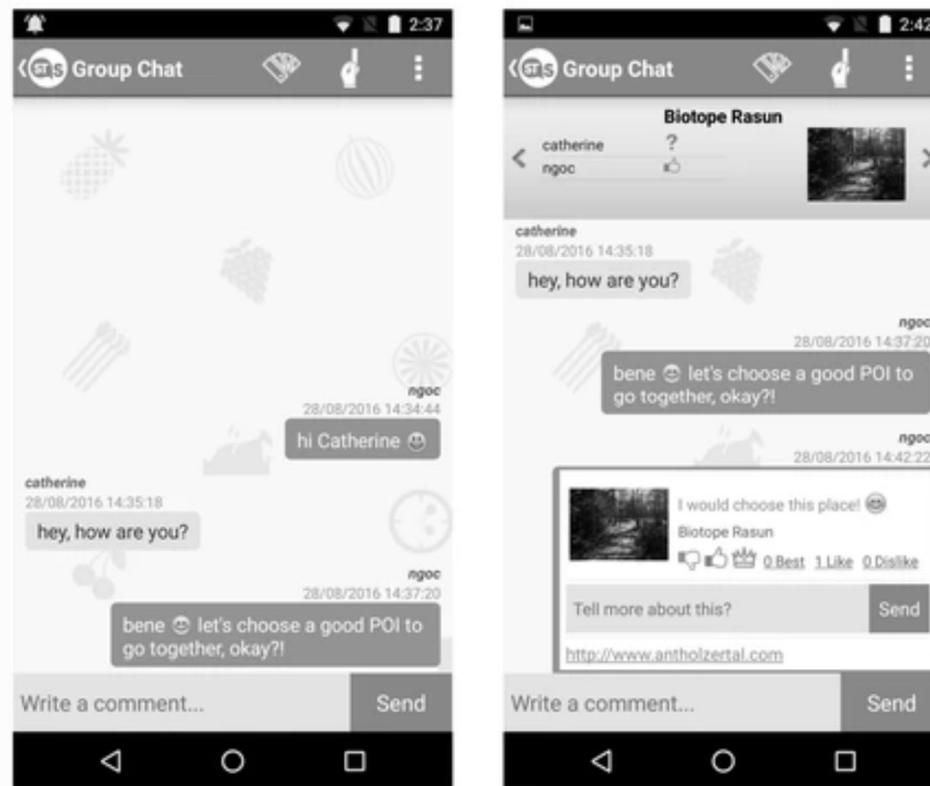
- System-driven
  - E.g., critiquing-based systems, form-based interactive advisory systems
- User-driven
  - “User-asks, system-responds”
  - Rarely entirely user-driven
- Mixed-initiative
  - Most of existing CRS

# Mixed-initiative: Music Chatbot



Cai, W., Jin, Y., and Chen, L.: Critiquing for Music Exploration in Conversational Recommender Systems. In: Proceedings of 26th International Conference on Intelligent User Interfaces (IUI'21), College Station, TX, USA, April 14–17, 2021.

# Mixed-initiative: Group chat with recs



Nguyen, T.N., Ricci, F. (2017). A Chat-Based Group Recommender System for Tourism. In: Schegg, R., Stangl, B. (eds) Information and Communication Technologies in Tourism 2017.

# Discussion

---

- Natural language interaction vs. form-based inputs
  - Pure natural language interfaces in principle provide the opportunity to elicit preferences in a more natural way.
  - Users might be better acquainted and feel more comfortable with more traditional interaction mechanisms (forms and buttons).
  - A mix of a natural language interface and buttons led to the best user experience (Iovine et al., 2020).

A. Iovine, F. Narducci, and G. Semeraro. Conversational recommender systems and natural language: A study through the ConveRSE framework. *Decision Support Systems*, 131:113250–113260, 2020.

# Discussion, cont.

---

- New application scenarios
  - Interactive wall
  - Service robot
  - In-car
  - Etc.
- General challenges
  - Privacy considerations
  - Aspects of technology acceptance
  - Application-specific ones (e.g., safety considerations in an in-car setting)

# Agenda

---

1. Introduction
2. Characterization of Conversational Recommender System (CRS)
3. Interaction Modalities
- 4. Underlying Knowledge and Data**
5. Computational Tasks
6. Evaluation of CRS
7. Wrap-up & Discussion

# Types of Knowledge and Data

---

- Supported user intents
  - Users' particular information needs and intents that can occur in conversations
  - Pre-defined or automatic detection (e.g., in NLP-based systems)
- User modeling
- Dialogue states
- Background knowledge

# High-level Overview

A high-level overview of domain-independent user intents

| Intent Name                       | Intent Description                               |
|-----------------------------------|--|
| <i>Initiate Conversation</i>      | Start a dialogue with the system.                |
| <i>Chit-chat</i>                  | Utterances unrelated to the recommendation goal. |
| <i>Provide Preferences</i>        | Share preferences with the system.               |
| <i>Revise Preferences</i>         | Revise previously stated preferences.            |
| <i>Ask for Recommendation</i>     | Obtain system suggestions.                       |
| <i>Obtain Explanation</i>         | Learn more about why something was recommended.  |
| <i>Obtain Details</i>             | Ask about more details of a recommended object.  |
| <i>Feedback on Recommendation</i> | Give feedback on the provided recommendation(s). |
| <i>Restart</i>                    | Restart the dialogue.                            |
| <i>Accept Recommendation</i>      | Accept one of the recommendations.               |
| <i>Quit</i>                       | Terminate the conversation.                      |

# Hierarchical Taxonomy of User Intents

| Intent (Code)                 | Description  | Percentage    |
|-------------------------------|--|---------------|
| <b>Ask for Recommendation</b> |  | <b>18.26%</b> |
| Initial Query (IQU)           | Seeker asks for a recommendation in the first query.                             | 12.91%        |
| Continue (CON)                | Seeker asks for more recommendations in the subsequent query.                    | 3.10 %        |
| Reformulate (REF)             | Seeker restates her/his query with or without clarification/further constraints. | 1.50%         |
| Start Over (STO)              | Seeker starts a new query to ask for recommendations.                            | 0.84%         |
| <b>Add Details</b>            |  | <b>18.58%</b> |
| Provide Preference (PRO)      | Seeker provides specific preference for the item s/he is looking for.            | 12.30%        |
| Answer (ANS)                  | Seeker answers the question issued by the recommender.                           | 4.91%         |
| Ask Opinion (ASK)             | Seeker asks the recommender's personal opinions.                                 | 2.39%         |
| <b>Give Feedback</b>          |  | <b>61.92%</b> |
| Seen (SEE)                    | Seeker has seen the recommended item before.                                     | 21.14%        |
| Accept (ACC)                  | Seeker likes the recommended item.   | 18.89%        |
| Reject (REJ)                  | Seeker dislikes the recommended item.  | 11.50%        |
| Inquire (INQ)                 | Seeker wants to know more about the recommended item.                            | 6.55%         |
| Critique-Feature (CRI-F)      | Seeker makes critiques on specific features of the current recommendation.       | 6.50%         |
| Critique-Add (CRI-A)          | Seeker adds further constraints on top of the current recommendation.            | 5.35%         |
| Neutral Response (NRE)        | Seeker does not indicate her/his preference for the current recommendation.      | 4.29%         |
| Critique-Compare (CRI-C)      | Seeker requests sth similar to the current recommendation in order to compare.   | 1.55%         |
| <b>Others</b>                 | Greetings, gratitude expression, or chit-chat utterances.                        | 14.55%        |

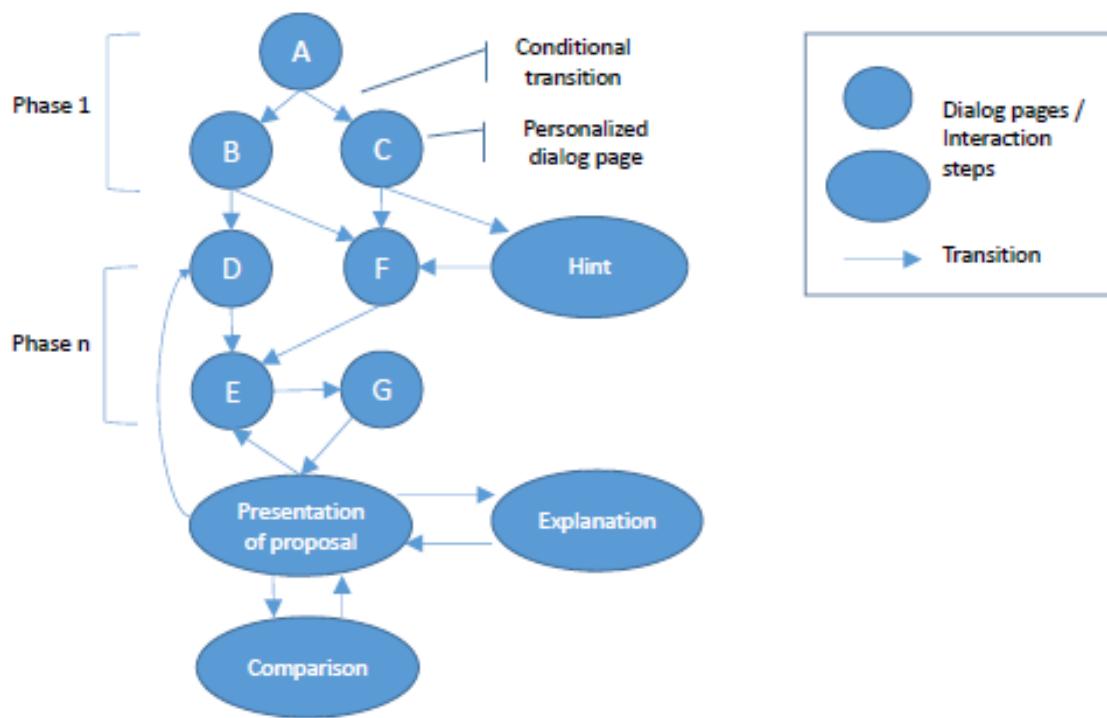
Cai, W., and Chen, L.: Predicting User Intents and Satisfaction with Dialogue-based Conversational Recommendations.  
 In Proceedings of 28th Conference on User Modeling, Adaptation and Personalization (UMAP'20), pages 33–42, July 14-17, 2020.

# User Modeling

---

- User profile
  - Preference expressions or estimates regarding **individual items**, e.g., ratings, like and dislike statements
  - Preferences regarding **individual item facets**, e.g., the genre of a movie or the desired functionalities
- Long-term preference
  - Long-term and supposedly more stable preferences (e.g., for non-smoking rooms in restaurants) from multiple sessions

# Dialogue States: Pre-defined

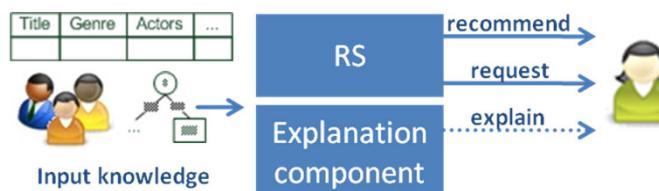


Jannach, D., and Kreutler, G.: Rapid development of knowledge-based conversational recommender applications with Advisor Suite. Journal of Web Engineering, 6(2):165–192, June 2007.

# Pre-defined States: Generation of explanations

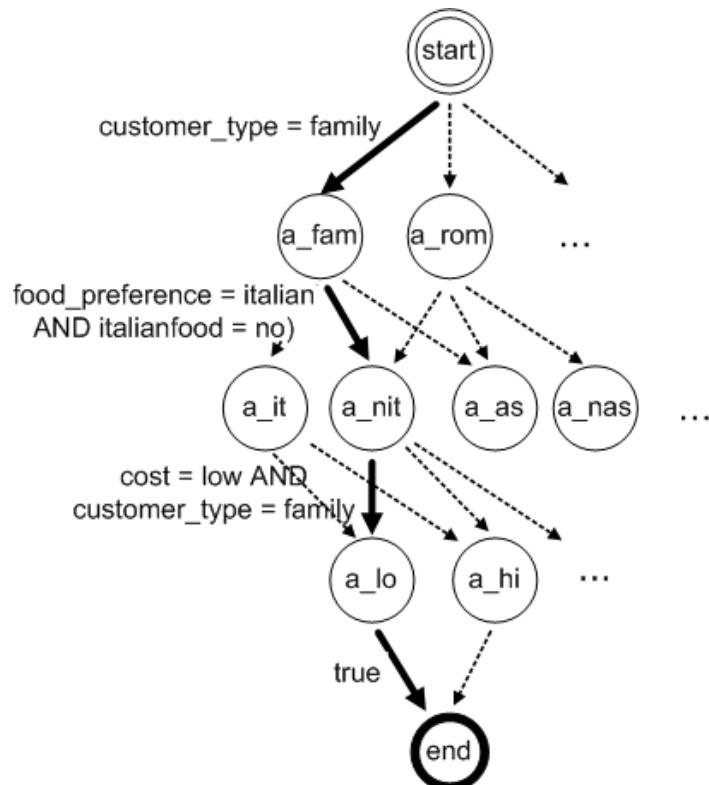
The screenshot shows a recommendation interface. At the top, there's a blue header bar with a search icon and the text "Suche". Below it, a card displays a spa resort named "Terme VIVAT" located in "Slowenien" (Slovenia) at "Moravske Toplice". A green icon of a building with water is labeled "Recommended spa". Below the card, a text box states: "Diese Therme entspricht zu 88% den von Ihnen gesuchten Kriterien" (This spa matches 88% of your search criteria). A question "Warum wurde Ihnen diese Therme empfohlen?" (Why was this spa recommended to you?) is followed by a detailed explanation: "The spa resort VIVAT is very apt for families with children. They are offering for instance day nursery and animators for children. The spa hotel offers 364 beds with a connecting passage to the spa area. The spa itself is rather small and of agreeable size. Its spa water promises positive effects when suffering from insufficient perfusion and rheumatism." To the right of the card, a photo of a couple in a swimming pool is labeled "Matching degree » zur Therme" (Matching degree to the spa). Below the photo, a blue button labeled "NSPIELE" is visible. A small blue robot-like character says "Hallo! Ich bin Aquarius. Kann ich Ihnen helfen?" (Hello! I am Aquarius. Can I help you?).

- Recommended items are accompanied with knowledgeable explanations



Zanker, M., and Ninaus, D.: Knowledge explanations for recommender systems. IEEE/WIC/ACM Conference on Web Intelligence and Intelligent Agent Technology (WI/IAT), pp:657–660, 2010.

# Pre-defined States: Generation of explanations



- Knowledge-based explanation model is represented by a layered directed acyclic graph (DAG)
- Contains a distinguished start and an end node
- Each layer represents a property of the item
- A valid explanation consists of all arguments along a path from the start to the end node, where no transition constraint is violated

Zanker, M., and Ninaus, D.: Knowledge explanations for recommender systems. IEEE/WIC/ACM Conference on Web Intelligence and Intelligent Agent Technology (WI/IAT), pp:657–660, 2010.

# Dialogue States: NLP-based Systems

---

- Mainly two phases in NLP-based conversational preference elicitation systems
  - Phase 1: Asking questions
  - Phase 2: Presenting a recommendation list
- Implicit states
  - Based on the implemented intents
  - **Or** encoded in a complex neural model as trained on a corpus of recorded human conversations, e.g., in the end-to-end learning CRS

# Background Knowledge: Item-related Knowledge

| Domain       | Description  |
|--------------|--|
| Movies       | Traditional movie rating databases from MovieLens, EachMovie, Netflix, used for example in [73, 166, 166].   |
| Electronics  | A product database with more than 600 distinct products was collected from various retailers [46].<br>A smartphone database consisting of 1721 products with multiple features [34].<br>Amazon electronics review dataset <sup>5</sup> containing millions of products, user reviews and product meta-data [164].  |
| Travel       | A dataset consisting of 120 personal computers, each with 8 features [128].<br>More than 100 sightseeing spots in Japan with 25 different features [52].<br>A database of restaurants in the San Francisco area covering 1,900 items with multiple features like cuisine, ratings, price, location, or parking [135].<br>Search logs and reviews of 3,549 users of a restaurant review provider, focusing on locations in Cambridge [29].<br>A travel destinations dataset, crawled from online platforms containing 5,723,169 venues in 180 cities around the globe [53]. |
| Food Recipes | A restaurants dataset crawled for Dublin city, which consists of 632 restaurants with 28 different features [89].  |
| E-commerce   | A food recipe dataset containing dishes and their ingredients [162].   |
| Music        | A product database of 11M products and logged data from the search engine of an e-commerce website was collected. The logged data consists of 3,146,063 unique questions [156].<br>A music dataset crawled from multiple online sources, containing 2,778 songs with 206k explanatory statements and 22 user tags [165].   |

# Background Knowledge: Dialogue Corpora

| Domain                 | Name       | Description  |
|------------------------|------------|--|
| Movies                 | ReDial     | Crowdworkers from Amazon Mechanical Turk (AMT) were used to collect over 10,000 dialogues centered around the theme of providing movie recommendations [73]. A paired mechanism was used where one person acts as a <i>recommendation seeker</i> and the other as a <i>recommender</i> . |
|                        | CCPE-M     | A Wizard-of-Oz (WoZ) approach <sup>b</sup> is taken to elicit movies preferences from crowdworkers within natural conversations. The dataset consists of over 500 dialogues that contain over 10,000 preference statements [111].  |
|                        | GoRecDial  | This dataset consists of 9,125 dialogue interactions and 81,260 conversation turns collected through pairs of human workers; here also one plays the role of a movie seeker and the other as a recommender [63].   |
|                        | bAbI       | In [95], the authors used a general movie dialogue dataset provided by Facebook Research [39] to build a CRS. The dataset contains task-based conversations in a question-answering style. It consists of 6,733 and 6,667 dialogue conversations for training and testing respectively.  |
| Restaurants and Travel | CRM        | An initial dataset containing 385 dialogues is collected using a pre-defined dialogue template through AMT [133]. Using this dataset, a larger synthetic dataset of 875,721 simulated dialogues is created.  |
|                        | ParlAI     | A goal-oriented, extended version of the bAbI dataset that was collected using a bot and users. It consists of three datasets (training, development and testing), each comprising 6,000 dialogues. [62].  |
|                        | MultiWOZ   | A large human-human dialogue corpus, which covers 7 domains and consists of 8,438 multi-turn dialogues around the themes of travel & planning recommendation [154].  |
| Fashion                | MMD        | A dataset consisting of 150,000 conversations between shoppers and a large number of expert sales agents is collected. 9 dialogue states were identified in the resulting dataset [123].   |
| Multi-domain           | OpenDialKG | A dataset of chat conversations between humans is collected, consisting of 15,000 dialogues and 91,000 conversation turns on movies, books, sports, and music [93].  |

# Background Knowledge: Logged Interaction Histories

---

| Domain     | Description   |
|------------|---|
| Movies     | A dialogue dataset involving 347 users was collected in [64] during the experimental evaluation of a recommender system.  |
|            | A subset of the ReDial dataset was analyzed and annotated in [16] to classify the user feedback types in 200 dialogues at the utterance level.  |
|            | A dialogue corpus was collected in [146] for the purpose of dialogue quality analysis from experimental session logs consisting of 226 complete dialogue turns with 20 users.   |
|            | A user study was conducted in [147], where a <i>movie seeker</i> and a <i>human recommender</i> converse with each other. The dialogue corpus consists of 2,684 utterances and 24 complete dialogues.   |
| Travel     | A dataset containing preferences for hotel, flight, car rental searches was collected in [4] involving 200 users of a content-based recommender system that supports multiple tasks (i.e., hotel, car, flight booking) in the same dialogue.                        |
| Fashion    | A user study was conducted using a virtual shopping system, where the goal was to find a wedding dress. A non-verbal feedback (e.g., gestures, facial expressions, voices) dataset involving 345 subjects was collected and then annotated for model training [18]. |
| E-commerce | A dataset containing conversation logs of users with a chatbot of an online customer service center (Alibaba.com) was collected in [109]. It consists of over 91,000 Q&A pairs as a knowledge base used for the information retrieval task.                         |

# Background Knowledge: Lexicons and World Knowledge

---

| Source Name | Description   |
|-------------|---|
| Wikipedia   | A dataset crawled from online sources (Wikipedia and Wikitravel) for the purpose of entity recognition in the travel domain [71].   |
| WordNet     | WordNet <sup>7</sup> is used in order to compute the semantic distance between entities and keywords mentioned in the conversation [71, 78].  |
| Wikiquote   | A quote dataset crawled from two online sources, Wikiquote <sup>8</sup> and the Oxford Concise Dictionary of Proverbs [68].   |
| Citysearch  | In [78], a dataset of 137,000 users reviews on 24,000 restaurants was harvested from two online sources (Citysearch <sup>9</sup> and MenuPages <sup>10</sup> ) to generate a dictionary of mappings between semantic representations of cuisines and dialogue concepts. |

# Discussion

---

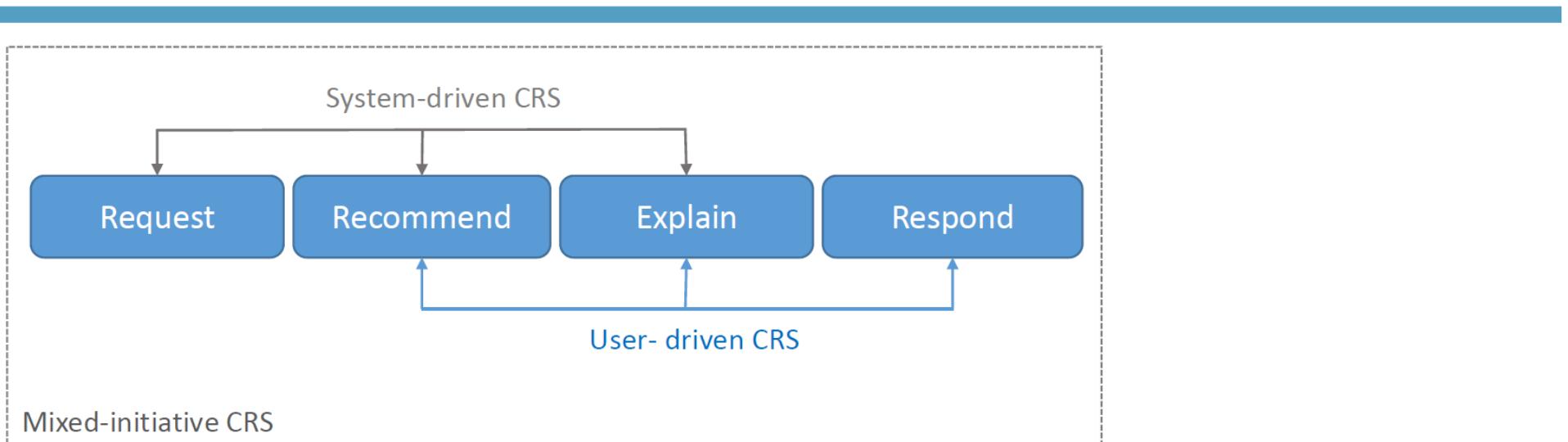
- Pre-defined knowledge vs. Learning approaches
  - Form-based interaction
    - Typically pre-defined in terms of possible dialogue states, supported user intents, or user profile attributes to acquire
  - NLP-based interaction
    - More dynamic in terms of the possible dialogue flow, and relying on additional knowledge sources
- Intent engineering
  - The set of supported intents determines how rich and varied the resulting conversations can be
  - Challenge: how to anticipate or learn over time which intents the users might have

# Agenda

---

1. Introduction
2. Characterization of Conversational Recommender System (CRS)
3. Interaction Modalities
4. Underlying Knowledge and Data
- 5. Computational Tasks**
6. Evaluation of CRS
7. Wrap-up & Discussion

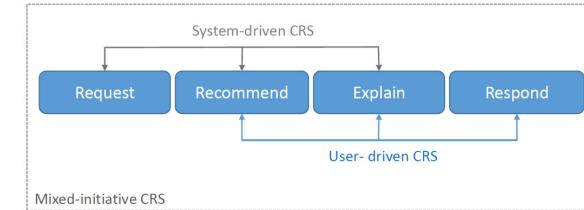
# Main Computational Tasks



- A conceptual, generic view
  - Not all tasks (explicitly) supported in all systems

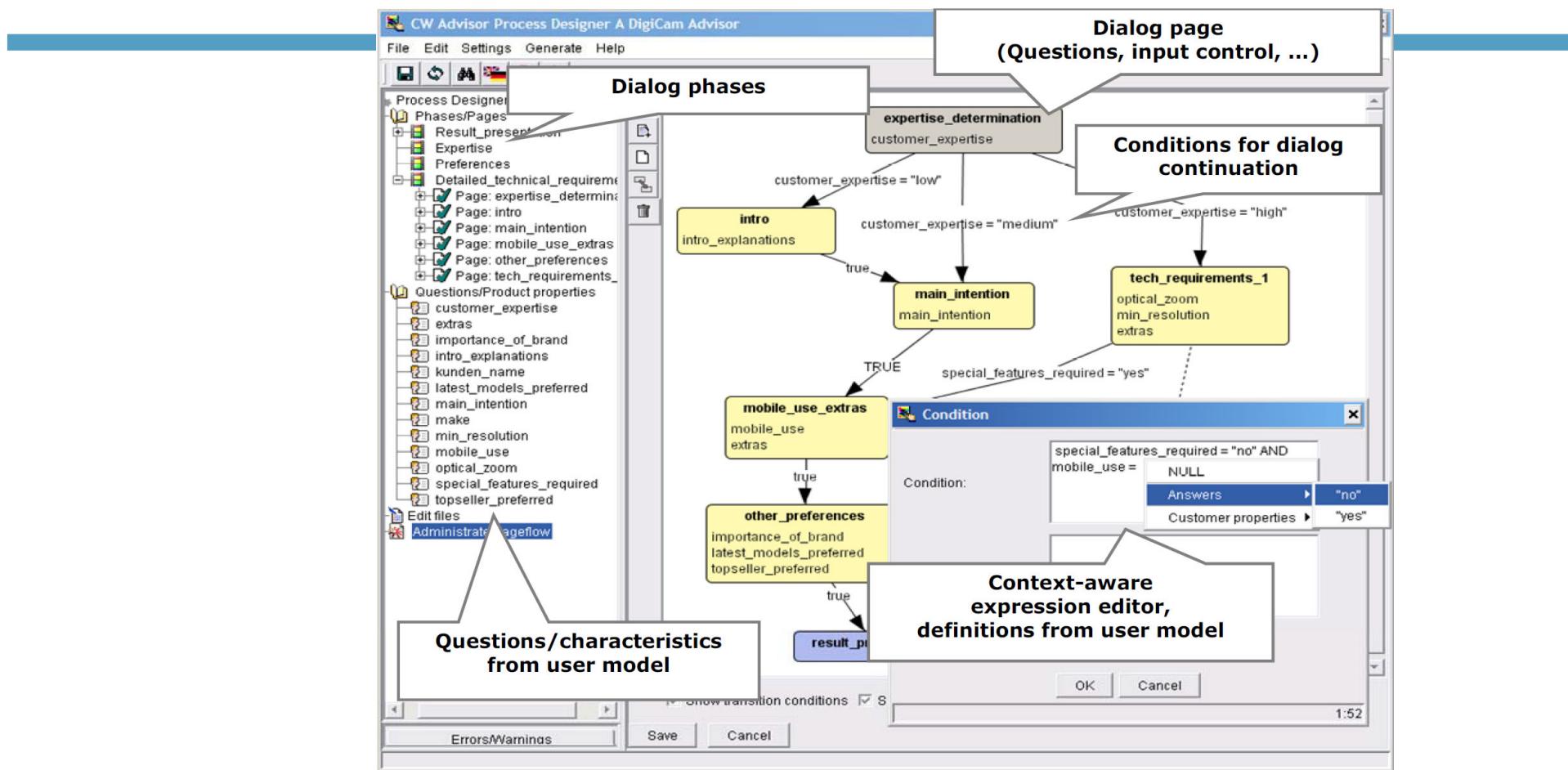
# Main Tasks:

## Request (examples)

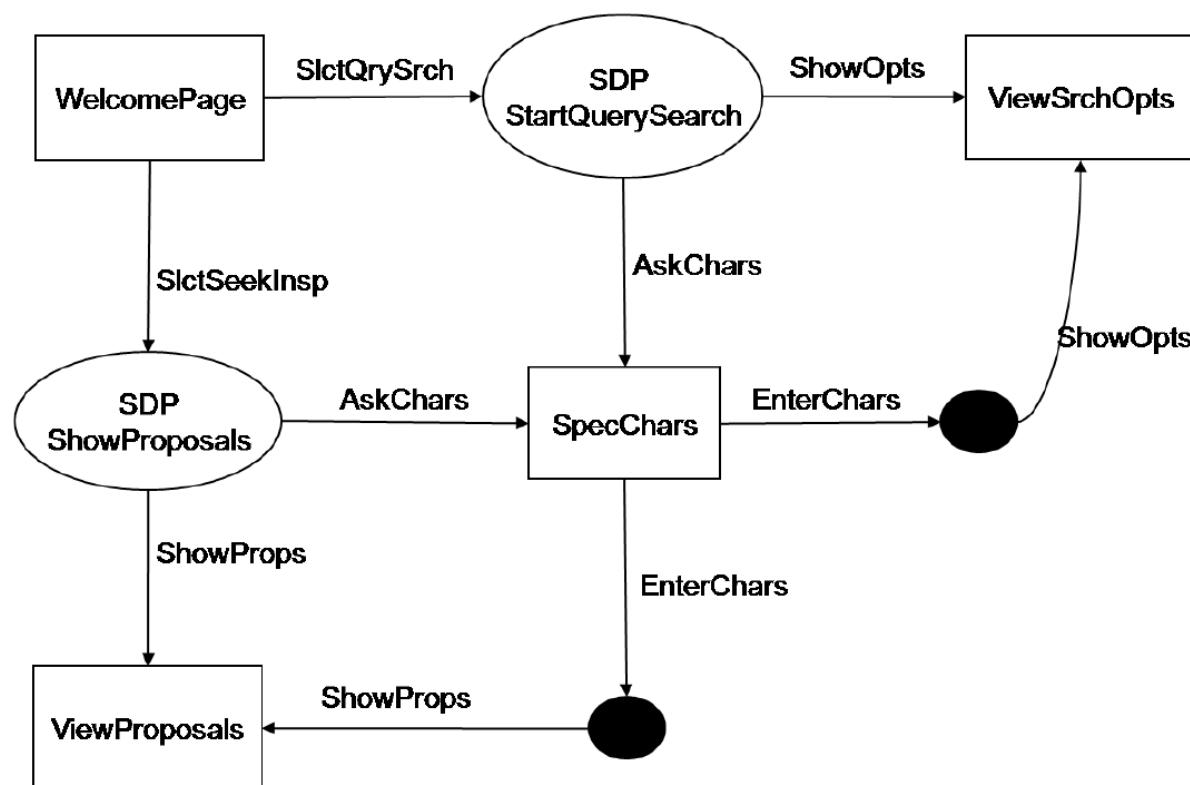


- Slot-filling
  - A common approach to acquire preferences
  - Determine next question to ask
    - Questions often relate to desired item features/attributes
    - Computation may be based, e.g., on set of remaining items in critiquing approaches, to reduce interaction time
- Reasoning about dialogue state
  - Determine next “conversational move” (e.g., ask more, explain, recommend)
    - e.g., based on reinforcement learning / bandit approaches
    - Or based on explicit rules

# Request: A rule-based Approach



# Request: Learning the next move

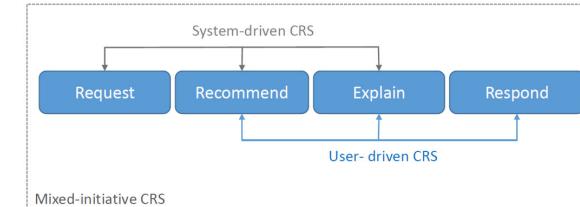


Tariq Mahmood and Francesco Ricci. 2009. Improving recommender systems with adaptive conversational strategies. ACM Conference on Hypertext and Hypermedia (HT '09).

# Main Tasks:

## Recommend

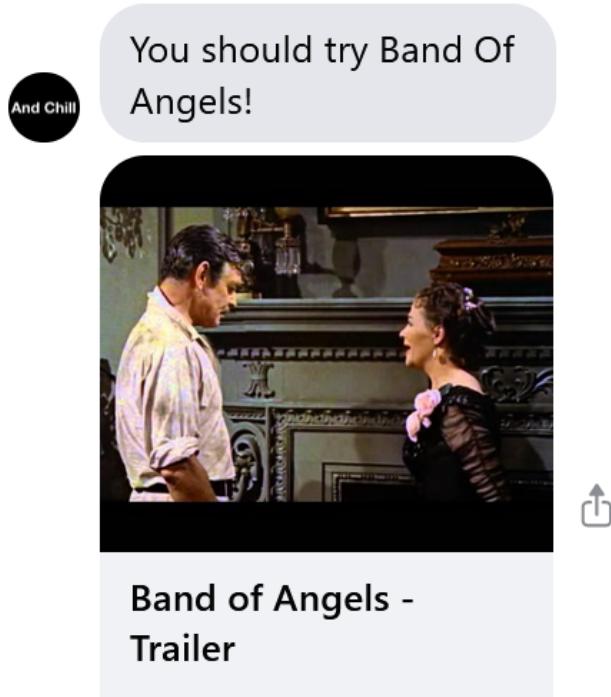
- Various technical approaches possible:
  - Collaborative, content-based, hybrid, and in particular: knowledge-based
  - Knowledge-based:
    - Match elicited feature preferences with available items
      - Constraint-based, case-based, utility-based
    - Often only based on short-term preferences acquired during the usage session
  - Some models combine long-term and short-term user models



# Main Tasks: Recommend

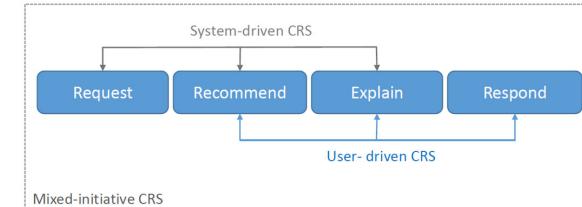
---

- Often only one recommendation at a time  
(in contrast to traditional recommenders)



# Main Tasks: Explain

- Explanations are widely recognized as a trust-building feature of recommenders
  - Little research in CRS, though
  - Practical systems also have their limitations



why do you recommend  
band of angels

Gotcha... that title,  
right? There's only  
1.8% of users that have  
mentioned this one. I  
am still generating AI-  
powered movie recs  
based on this film. Can  
you tell me about a  
different one you like?



# Explain: Trust-inspiring interfaces

- Critiquing-based Recommendation

| The most popular product                                     |       |                 |              |                  |                     |              |                 |
|--|-------|-----------------|--------------|------------------|---------------------|--------------|-----------------|
| Manufacturer   | Price | Processor speed | Battery life | Installed memory | Hard drive capacity | Display size | Weight          |
| C  | —     | \$2'095.00      | 1.67 GHz     | 4.5 hour(s)      | 512 MB              | 80 GB        | 38.6 cm 2.54 kg |
| We also recommend the following products because             |       |                 |              |                  |                     |              |                 |
| they are cheaper and lighter, but have lower processor speed |       |                 |              |                  |                     |              |                 |
| Manufacturer   | Price | Processor speed | Battery life | Installed memory | Hard drive capacity | Display size | Weight          |
| C  | —     | \$1'499.00      | 1.5 GHz      | 5 hour(s)        | 512 MB              | 80 GB        | 33.8 cm 1.91 kg |
| C  | —     | \$1'739.99      | 1.5 GHz      | 4.5 hour(s)      | 512 MB              | 80 GB        | 38.6 cm 2.49 kg |
| C  | —     | \$1'625.99      | 1.5 GHz      | 5 hour(s)        | 512 MB              | 80 GB        | 30.7 cm 2.09 kg |

P. Pu and L. Chen. Trust building with explanation interfaces. In IUI '06, pages 93–100, 2006.

# Explain: User-tailored Descriptions

---

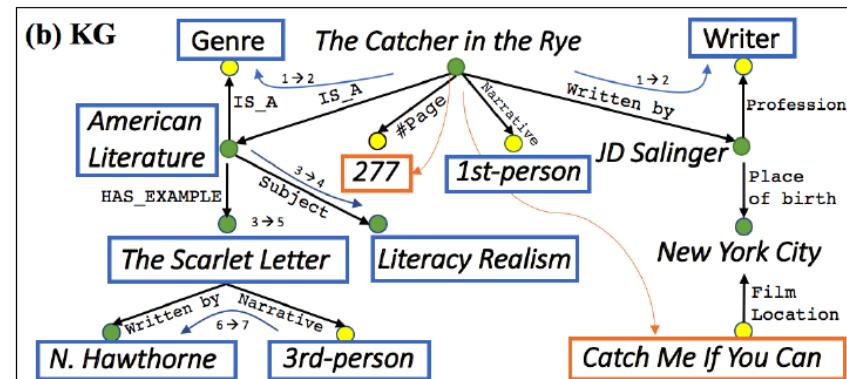
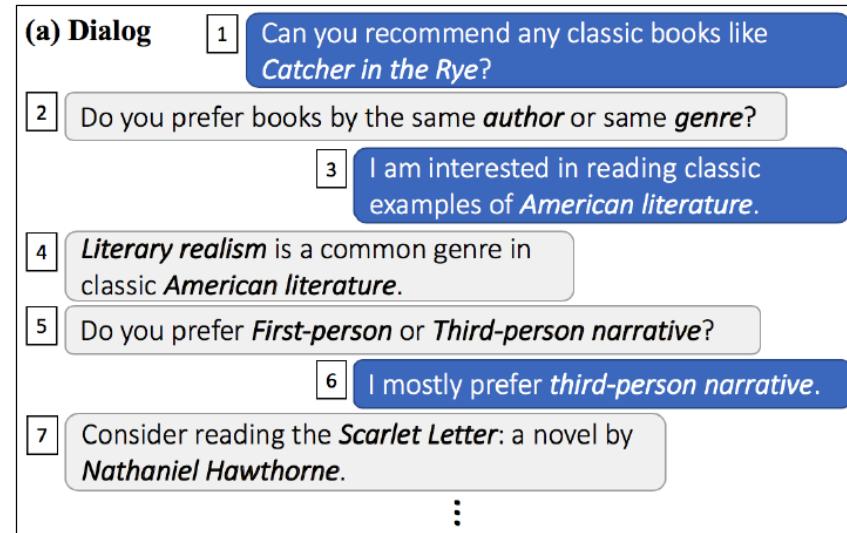
- Varying levels of conciseness

| User | Z value | Output  |
|------|---------|---|
| CK   | 0.3     | Bond Street has the best overall value among the selected restaurants. Bond Street has excellent food quality.                        |
| BA   | 0.3     | Komodo has the best overall value among the selected restaurants. Komodo's price is \$29. It's a Japanese, Latin American restaurant. |
| VM   | 0.3     | Komodo has the best overall value among the selected restaurants. Komodo's price is \$29 and it has very good food quality.           |

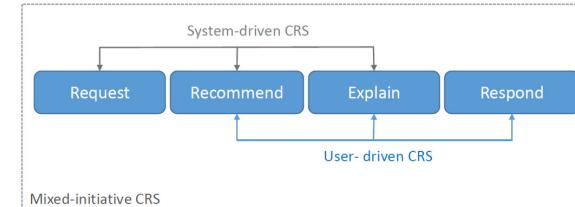
# Explain: Knowledge-enhanced model

- Using a Knowledge Graph to recommend and explain

S. Moon, P. Shah, A. Kumar, and R. Subba. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In ACL '19, pages 845–854, 2019.



# Main Tasks: Respond



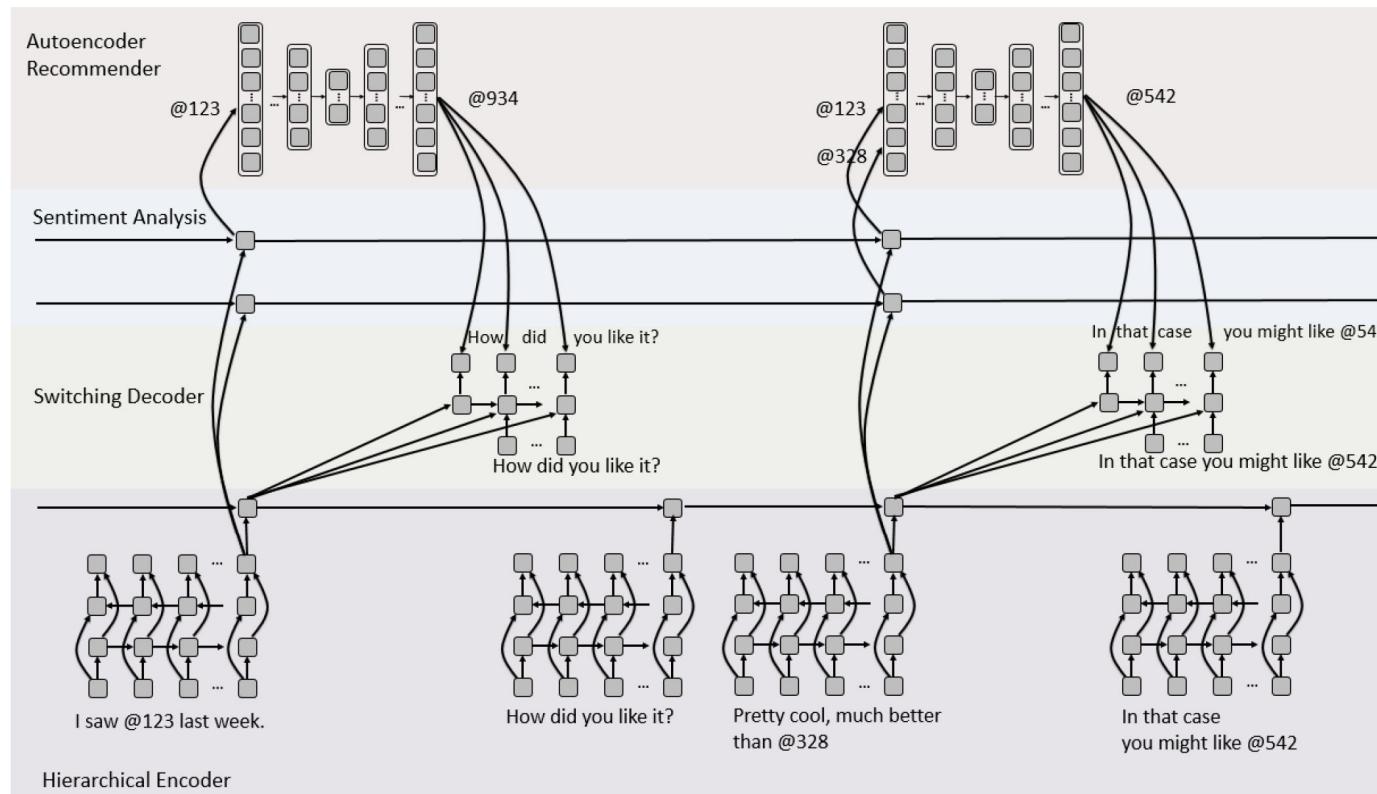
- A generic category
  - Relevant in user-driven and mixed-initiative CRS
  - React when user proactively leads the dialogue
- Properly respond to certain actions, e.g.,
  - when the user states a preference without being asked, or refines the requirements
    - e.g., “I like Pulp Fiction, but not Quentin Tarantino”
  - when user wants to restart the dialogue
    - e.g., because of a conversational breakdown
  - when the user asks for more information
    - e.g. “How about Huawei P9?”

# Main Tasks: Respond

---

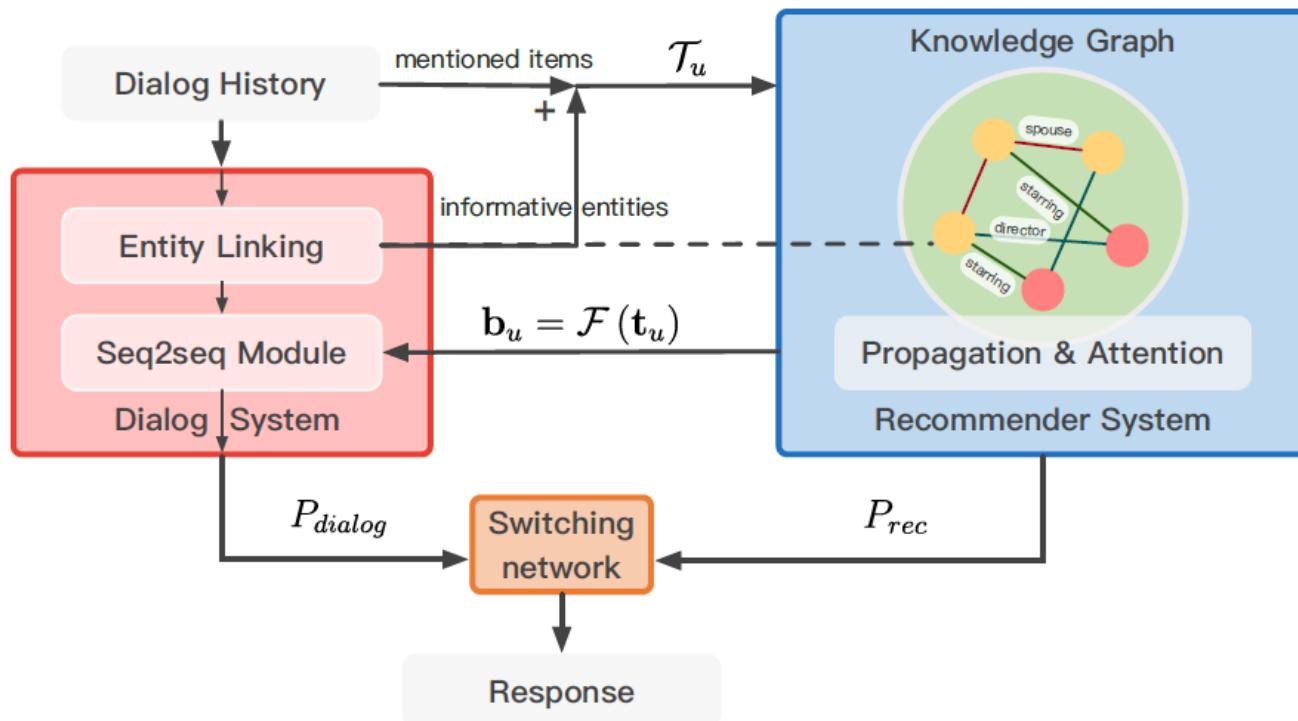
- Alternative technical approaches
- 1) Intent-detection / mapping
  - Based on a pre-defined list of supported user intents and an engineered or learned mapping logic
- 2) End-to-end learning
  - Automatically learn how to respond based on training data consisting of recorded dialogues

# Respond: Deep Learning



R. Li, S. E. Kahou, H. Schulz, V. Michalski, L. Charlin, and C. Pal. Towards deep conversational recommendations. In NIPS '18, pages 9725–9735, 2018.

# Respond: Deep Learning



Q. Chen, J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, and J. Tang. Towards knowledge-based recommender dialog system. In EMNLP-IJCNLP '19, pages 1803–1813, 2019.

# Respond: Retrieval-based Methods

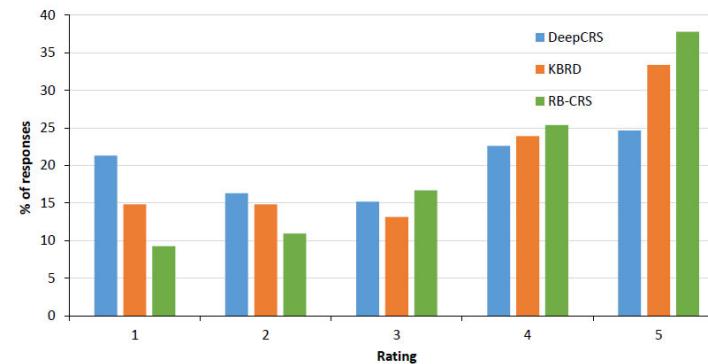
---

- Language-generation-based techniques
  - Dominate the research landscape
  - Have the potential to creatively respond to user utterances
- Retrieval-based approaches
  - Idea is to retrieve (and adapt) existing utterances from a given corpus, given a current query
  - Potential plus: Retrieved sentences are semantically and logically correct
  - Not much research done

# Respond: Retrieval-based Methods

- Recent user study
  - Comparing DeepCRS, KBRD, and simple retrieval-based system
- Let users (crowdworkers) assess different system responses

|                | DeepCRS | KBRD | RB-CRS      |
|----------------|---------|------|-------------|
| Avg. score     | 3.13    | 3.46 | <b>3.71</b> |
| Std. deviation | 1.49    | 1.45 | <b>1.32</b> |



A. Manzoor and D. Jannach. Generation-based vs. Retrieval-based Conversational Recommendation: A User-Centric Comparison. In RecSys '21, 2021  
Manzoor, A. and Jannach, D.: "Towards retrieval-based conversational recommendation". Information Systems, Vol. forthcoming. 2022

# Agenda

---

1. Introduction
2. Characterization of Conversational Recommender System (CRS)
3. Interaction Modalities
4. Underlying Knowledge and Data
5. Computational Tasks
- 6. Evaluation of CRS**
7. Wrap-up & Discussion

# Evaluation

---

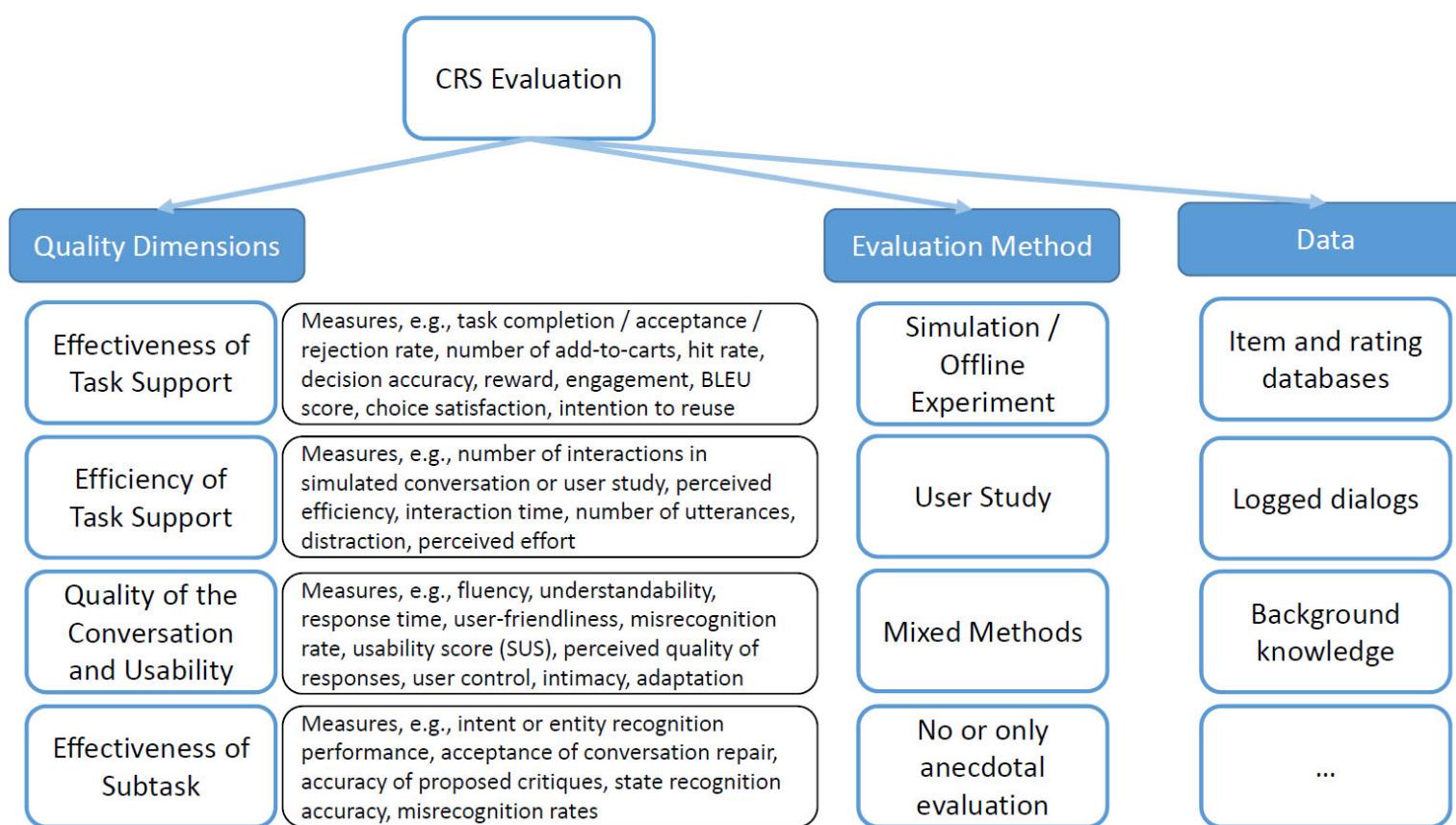
- General evaluation paradigms recommenders
  - Experimental research with users
    - Live (A/B) test of a deployed system
    - Controlled user studies with treatment and control groups
  - Research [without users](#) / simulations
    - “Offline” experiments regarding prediction or classification accuracy (e.g., for recommendations, entity recognition , etc.)
  - Non-experimental research
    - Quasi-experiments
    - Observational research
    - Qualitative research

# Evaluation Dimensions

---

- Effectiveness of Task Support
  - To what extent does the system help users find something new or make good decisions?
- Efficiency of Task Support
  - To what extent does the system help users make decisions faster (without compromising their choice satisfaction etc.)?
- Quality of the Conversation and Dialogue, Usability
  - For example, is the conversation natural? Does the system respond quickly and fluently?
- Effectiveness of Subtask
  - For example, how accurate is the entity recognition module?

# Overview of Evaluation Dimensions and Methods



# Effectiveness of Task Support

- Objective Measures

- **Offline:** Accuracy measures (RMSE, MAE, Precision, Recall, NDCG, MRR), proportion of simulated users accepting a recommendation
- **With users:** Task success, adoption of recommendations, if users changed their mind afterwards

- Subjective Measures

- Often based on the **ResQue** model, e.g., perceived accuracy, attractiveness, novelty, diversity, context compatibility, ...

## Quality Dimensions

Effectiveness of Task Support

Efficiency of Task Support

Quality of the Conversation and Usability

Effectiveness of Subtask

# Efficiency of Task Support

- Objective Measures

- Number of interaction cycles (dialogue turns)
  - Often used in simulations and user studies, but not so frequently in language-based interaction approaches
- Task completion time (time to make a decision)
  - Assumption is that shorter interactions are desirable; but may not always be the case

- Subjective Measures

- Perceived effort, cognitive effort
  - Often as part of usability assessments

#### Quality Dimensions

Effectiveness of Task Support

Efficiency of Task Support

Quality of the Conversation and Usability

Effectiveness of Subtask

# Quality of the Conversation

- .. and more usability
- Objective Measures
  - Linguistics: BLEU and NIST scores, perplexity, fluency
- Subjective Measures
  - Ease-of-use, task-ease, perceived level of user control, perceived transparency, adaptation of the system, expected behavior (intuitiveness), entertainment, mutual attentiveness, positivity, fluency, engagingness, overall dialogue quality, generation performance, ...

# Effectiveness of Subtask

- Objective Measures
  - Regret/reward for RL-based selection of next action
  - Number of applied critiques (of those generated and proposed)
  - Entity and intent recognition accuracy
- Subjective Measures
  - Interpretation performance

Quality Dimensions

Effectiveness of Task Support

Efficiency of Task Support

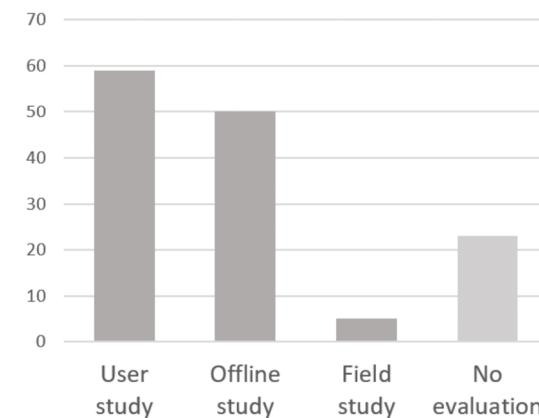
Quality of the Conversation and Usability

Effectiveness of Subtask

# Evaluation: Discussion

---

- Wide range of methods applied
- No standards exist, often only offline
  - What do research apply?
  - Correspondence of offline metrics with user perceptions sometimes unclear (see BLEU)
- Fresh off the press:
  - "Evaluating Conversational Recommender Systems". Artificial Intelligence Review, 2022  
<https://doi.org/10.1007/s10462-022-10229-x>



# Agenda

---

1. Introduction
2. Characterization of Conversational Recommender System (CRS)
3. Interaction Modalities
4. Underlying Knowledge and Data
5. Computational Tasks
6. Evaluation of CRS
7. Wrap-up & Discussion

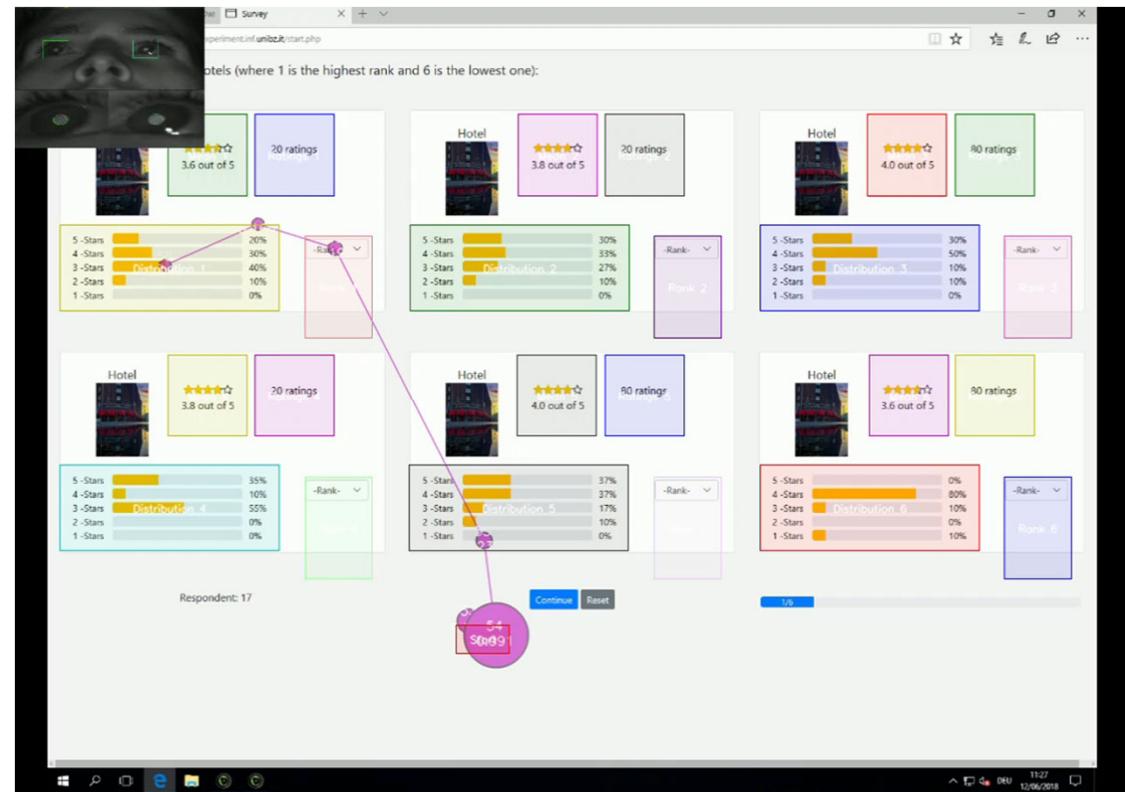
# Future Directions, examples

---

- Modalities
  - Not entirely clear which modality is the best in a context
  - Understanding non-verbal communication acts
- Mixed-method Evaluation, Exploratory Research
  - Need better understanding of what users expect and how recommendation dialogues should be structured
- Explanations
  - Not much explored yet, but might be expected
- Business Value of CRS
  - Almost no studies

# Example

- Mixed-method evaluation
- Eye-tracking to observe users' attention
- Compensatory and non-compensatory decision making strategies



Ludovik Coba, Laurens Rook, Markus Zanker, and Panagiotis Symeonidis. Decision making strategies differ in the presence of collaborative explanations: two conjoint studies. 24th International Conference on Intelligent User Interfaces (IUI '19), pp. 291–302.

# Thanks for the attention

---

- Time for questions
- Contact
  - [dietmar.jannach@aau.at](mailto:dietmar.jannach@aau.at)
  - [markus.zanker@unibz.it](mailto:markus.zanker@unibz.it)

