香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
HONG KONG BAPTIST UNIVERSITY
香港浸會大學計算機科學系

Understanding User Feedback on Recommendations in Conversational Systems

Dr. Li Chen

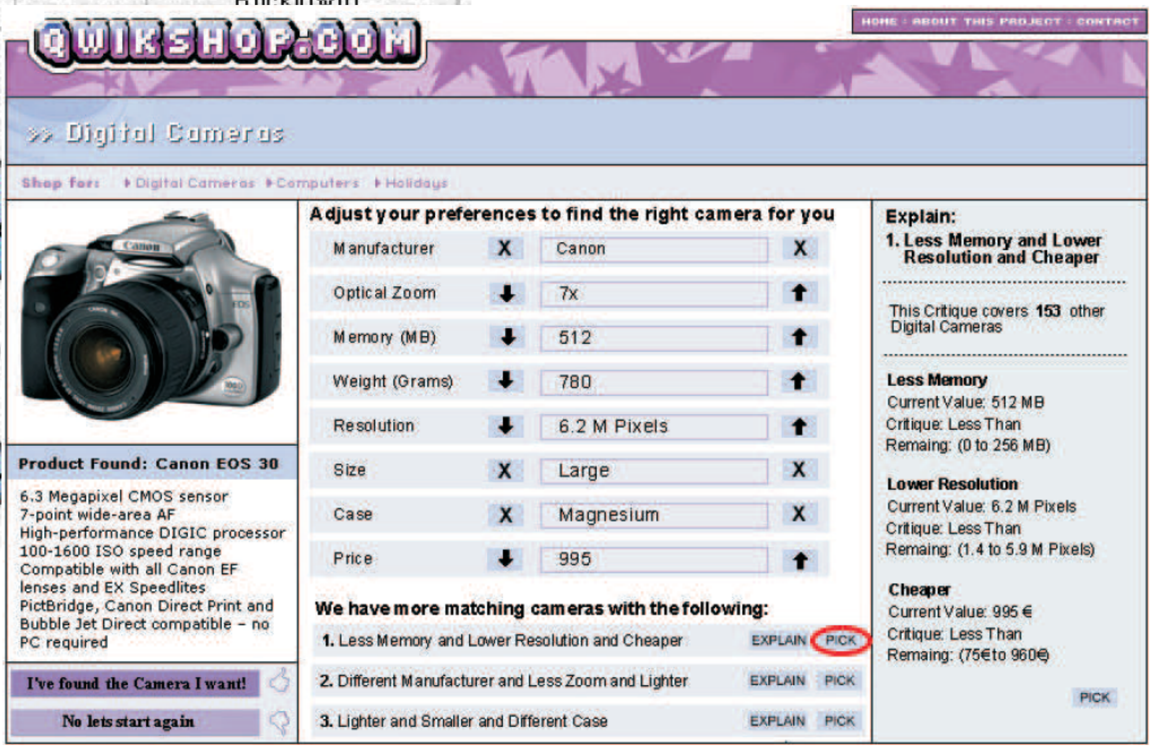
lichen@comp.hkbu.edu.hk

September 25, 2020

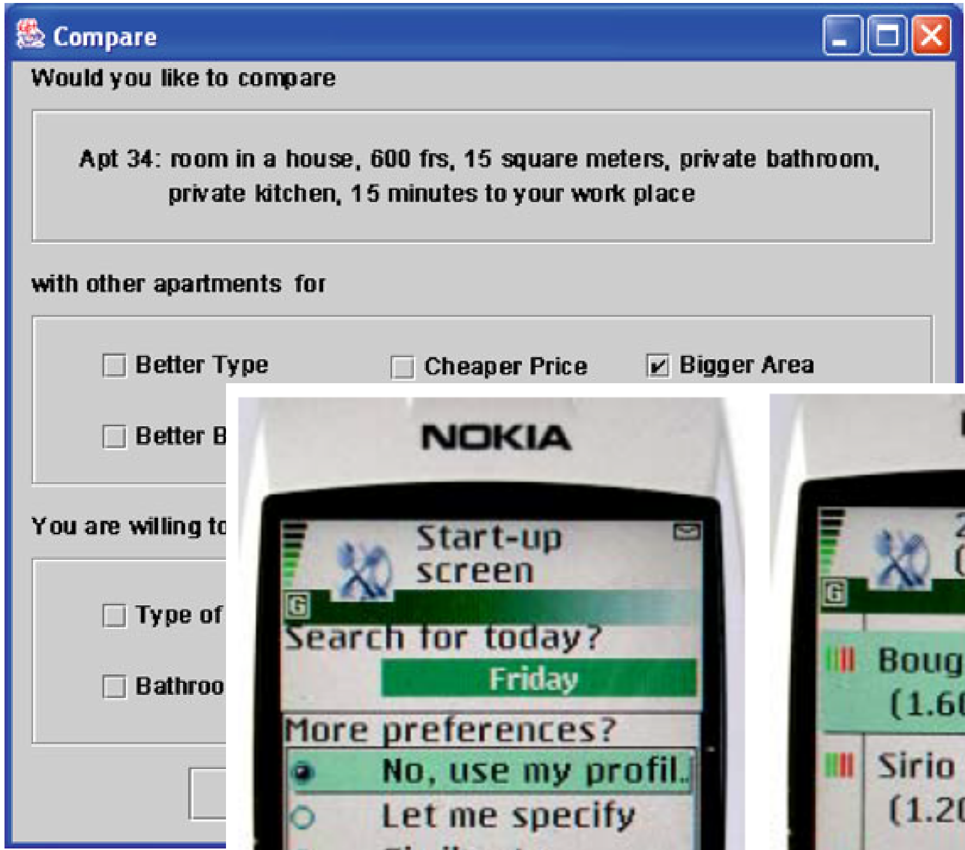
Invited talk for the 2nd International Workshop on Context-Aware Recommender Systems (CARS 2020), in conjunction with RecSys'20

Traditional Conversational Recommender Systems (CRS)

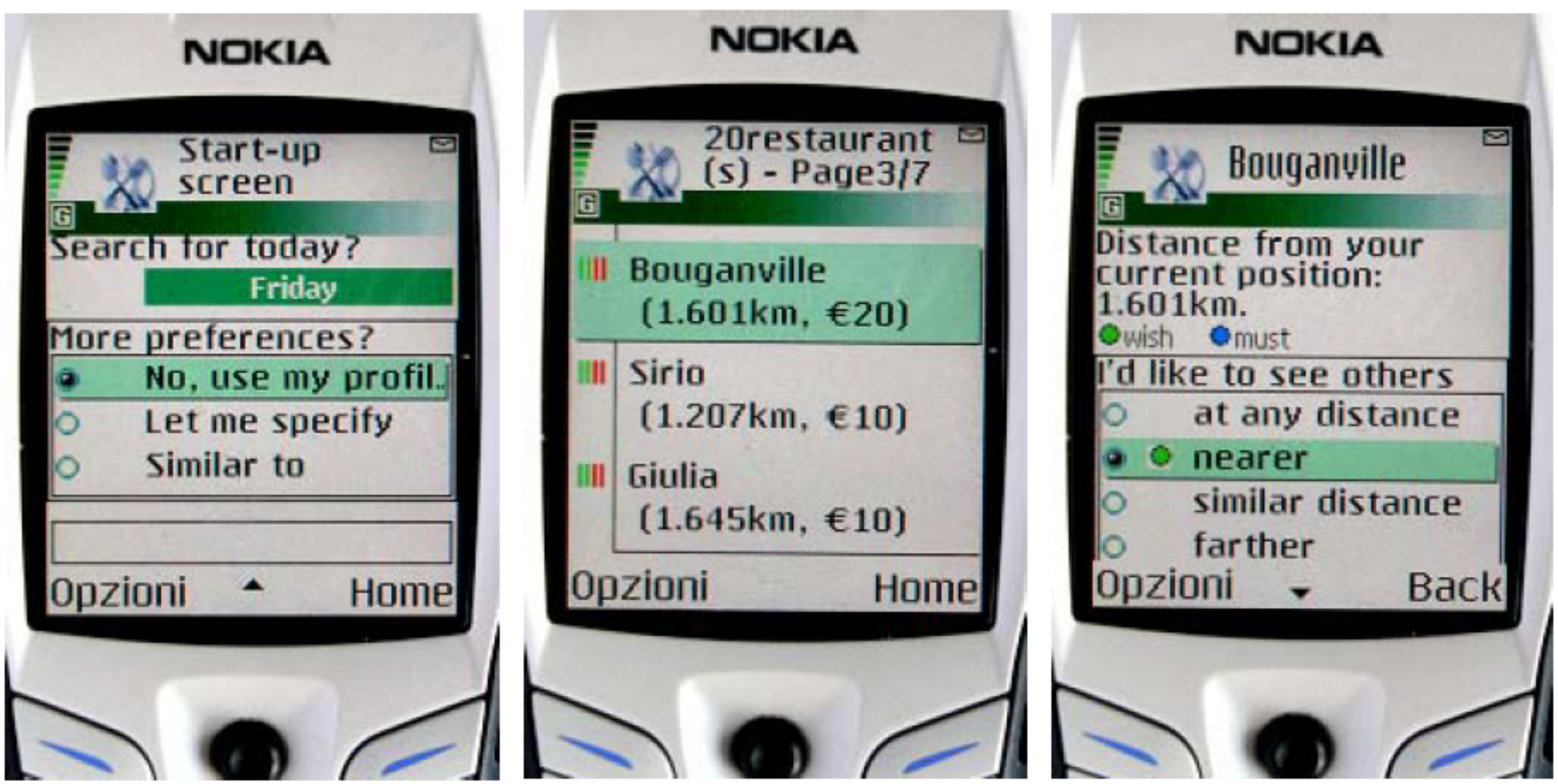
FindMe (Burke et al., 1997)



Dynamic Critiquing (McCarthy et al., 2005)



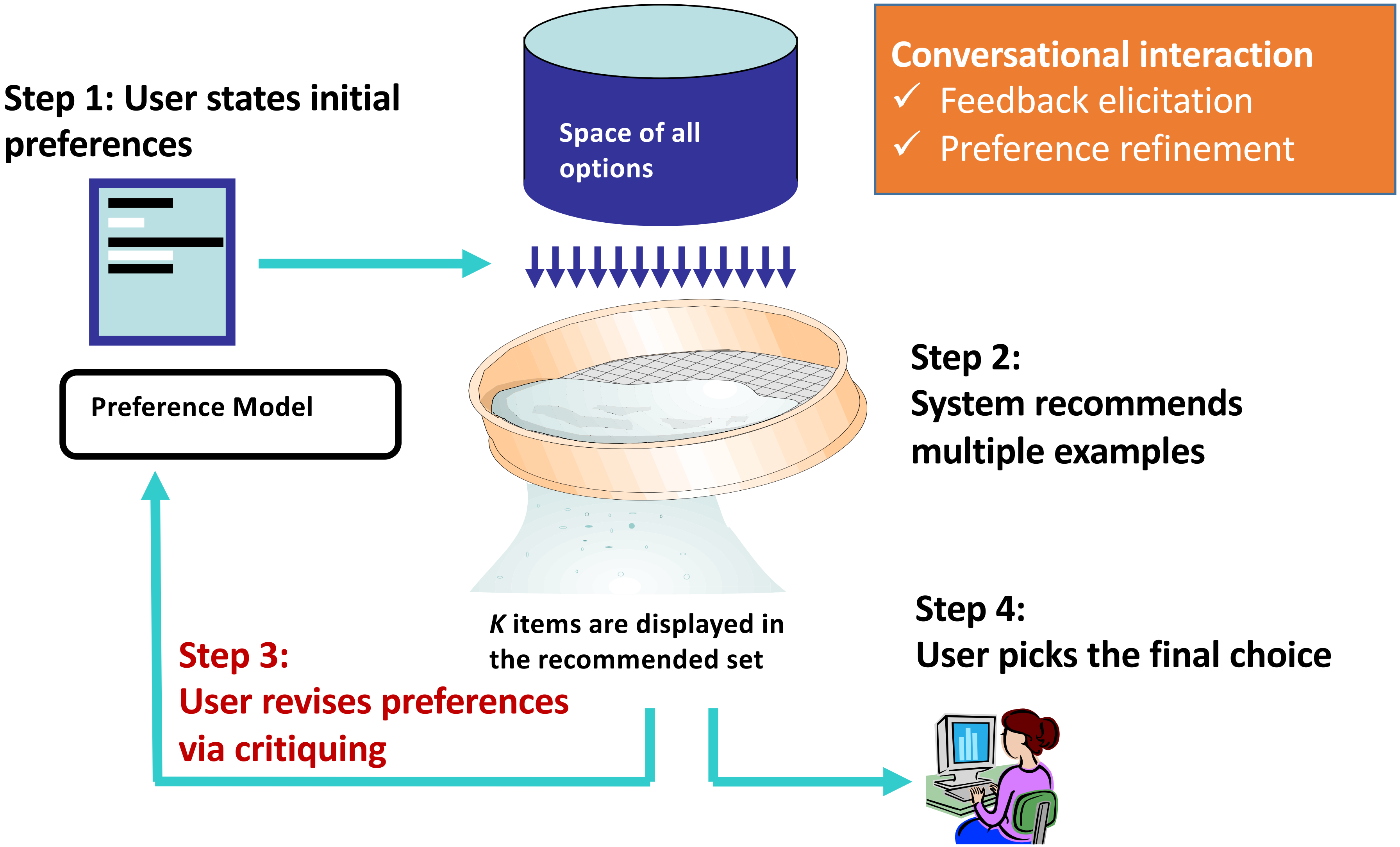
Apartment



(a) (b) (c)

MobyRek with mobile critiquing (Ricci and Nguyen, 2005)

Critiquing-based Recommender Systems



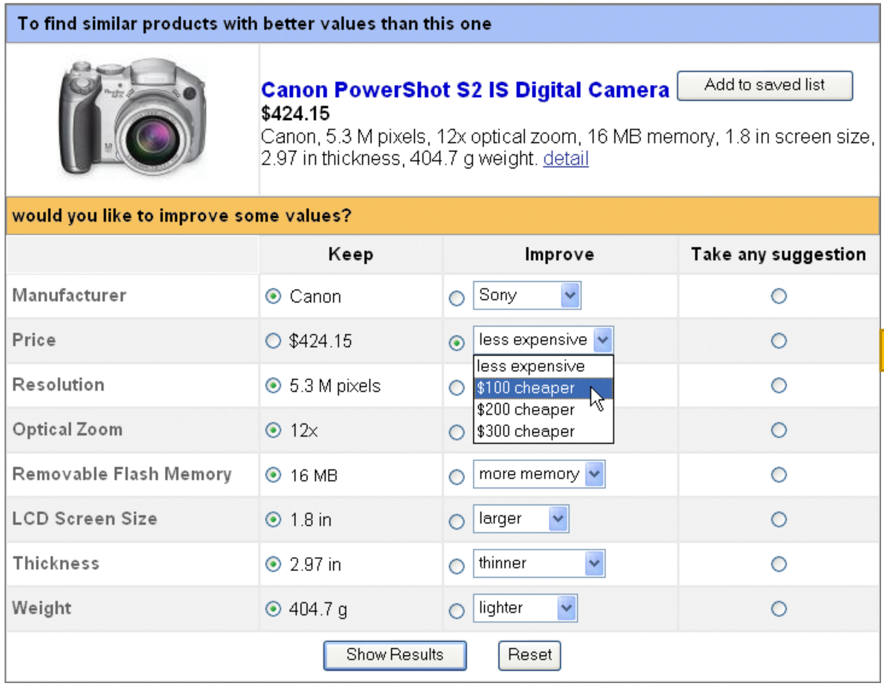
Li Chen and Pearl Pu. Critiquing-based Recommenders: Survey and Emerging Trends. *User Modeling and User-Adapted Interaction Journal (UMUAI)*, vol. 22(1), pages 125-150, 2012.

 4

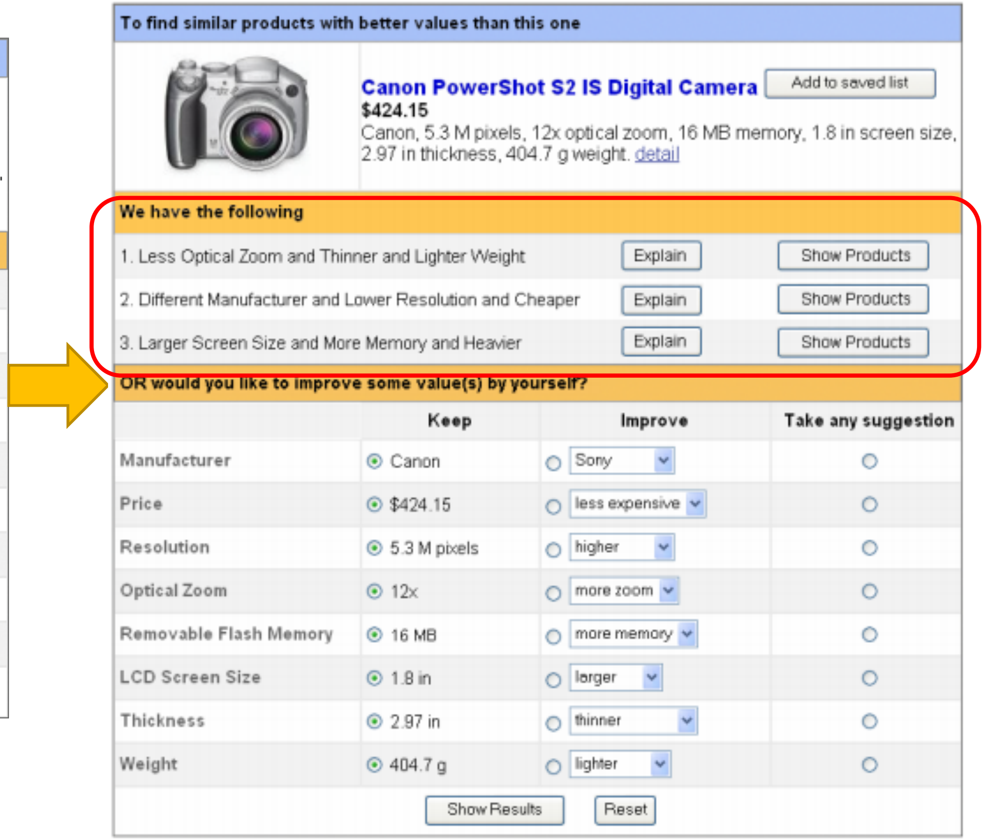
Motivated by **Adaptive Decision Theory**

- Users are likely to construct their preferences in a context-dependent and adaptive fashion during the decision process (Payne et al., 1993; Carenini and Poole, 2000).
- Users become aware of their latent preferences only when proposed solutions violate them (Pu and Faltings, 2000 & 2002).
- Compensatory decision strategy (i.e., tradeoff making) normally leads to rational and high-quality decision (Frisch and Clemen, 1994)

Unfamiliar product domain



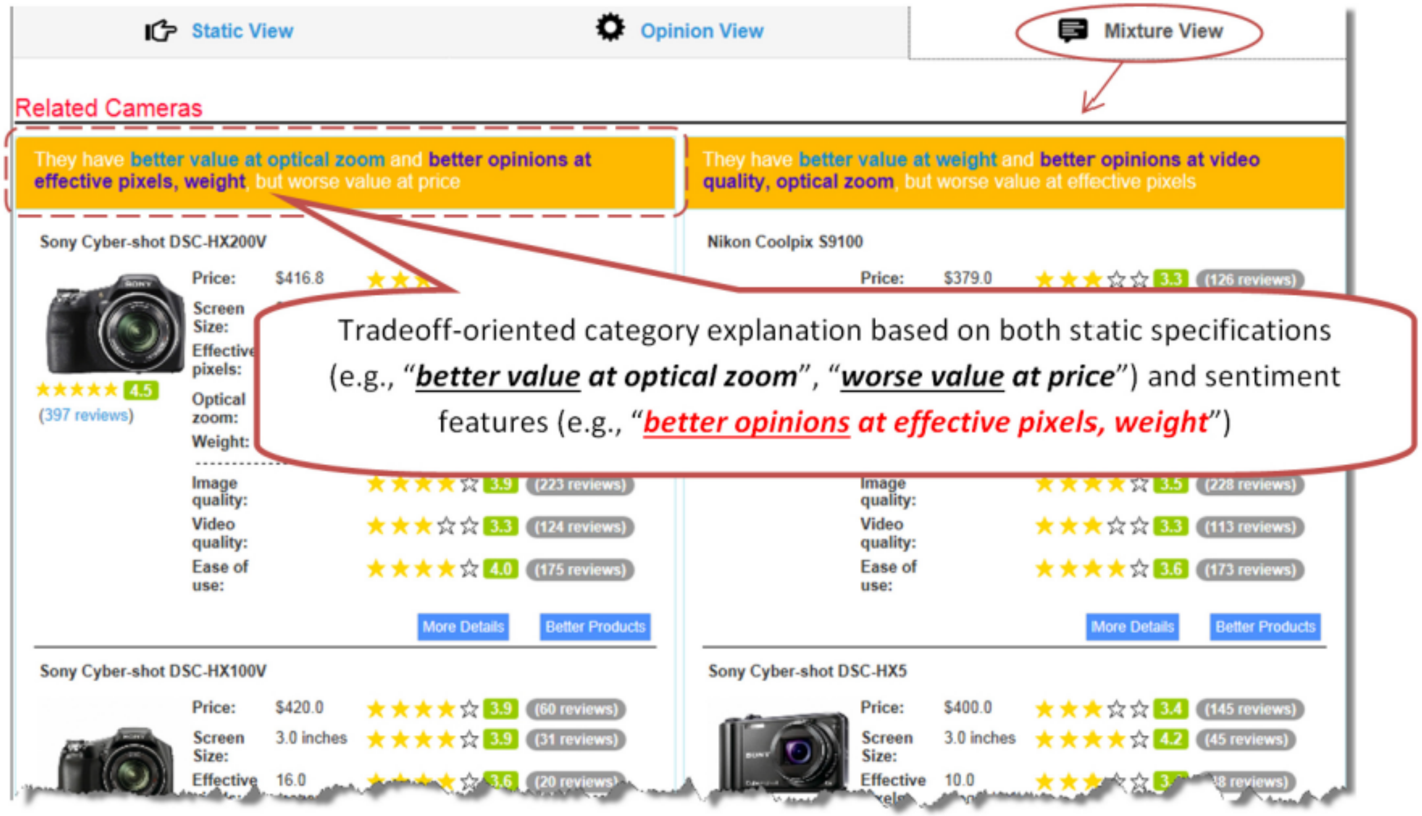
User-initiated critiquing: Unit or compound (Chen and Pu, AAAI'06)



Hybrid critiquing: User-initiated critiquing + system-suggested critiques (Chen and Pu, IUI'07)

- Critiquing-based system can significantly improve users' **decision accuracy** by **up to 57%**, against non-critiquing based
- Hybrid critiquing can achieve the **desired user control** and effectively save users' interaction effort

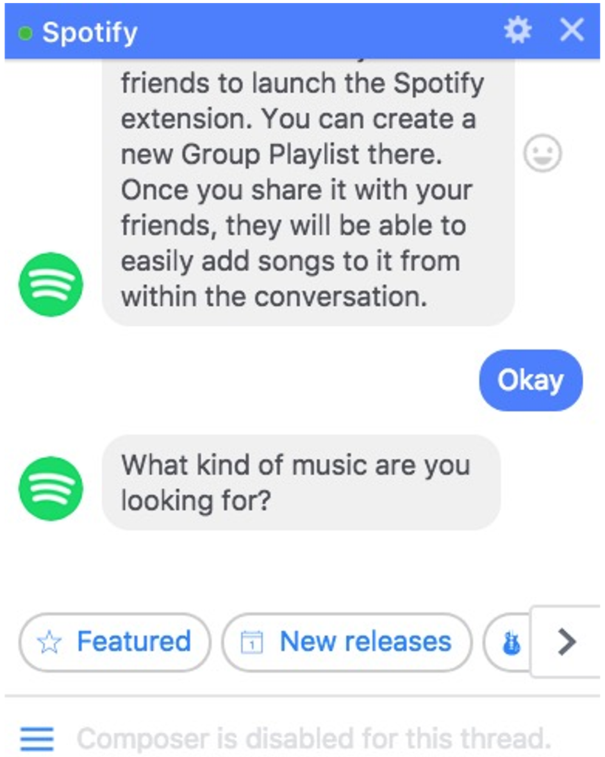
Sentiment-based critiquing



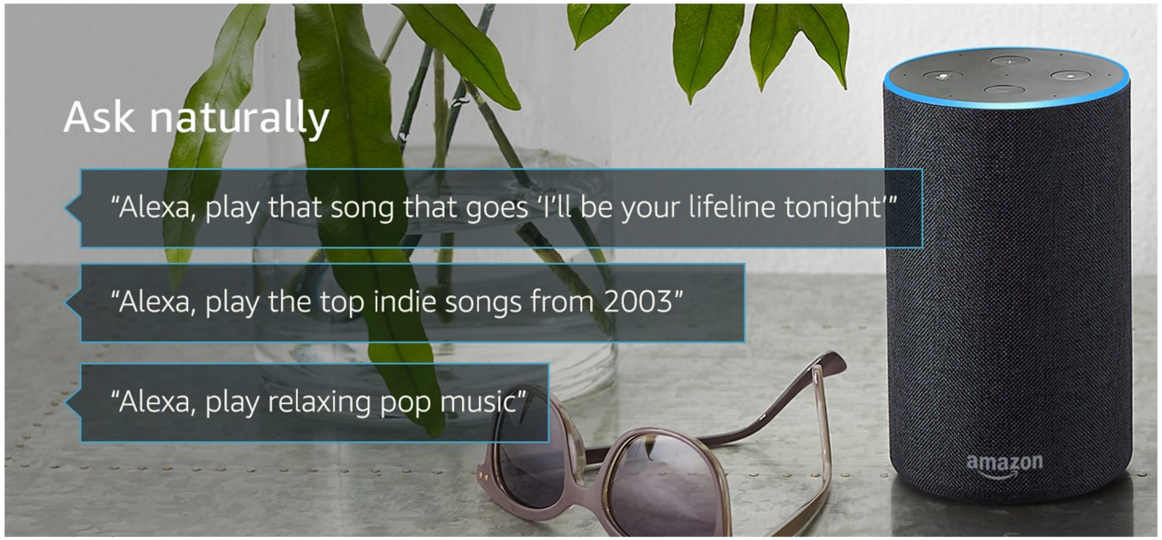
Incorporation of sentiment features into the critiquing interface can improve users' **product knowledge** and **preference certainty**

Li Chen, Dongning Yan, and Feng Wang. User Evaluations on Sentiment-based Recommendation Explanations. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 9(4), Article 20, 2019.

Dialogue-based CRS (DCRS)



https://www.poptin.com/blog/how-to-use-chatbots-drive-sales-engagement/



https://www.amazon.co.uk/b?ie=UTF8&node=11368385031

Challenges

- **Dialogue-based CRS**: Users can freely express their preference in a way that they feel at ease
- *But,*
  - in such less controlled setting, how to elicit their feedback on recommendation?
  - can we accurately understand their intents behind utterances?
  - can we predict their satisfaction with recommendation?
- Little work has investigated these issues in a multi-turn, mixed initiative dialogue-based CRS

Our Focuses

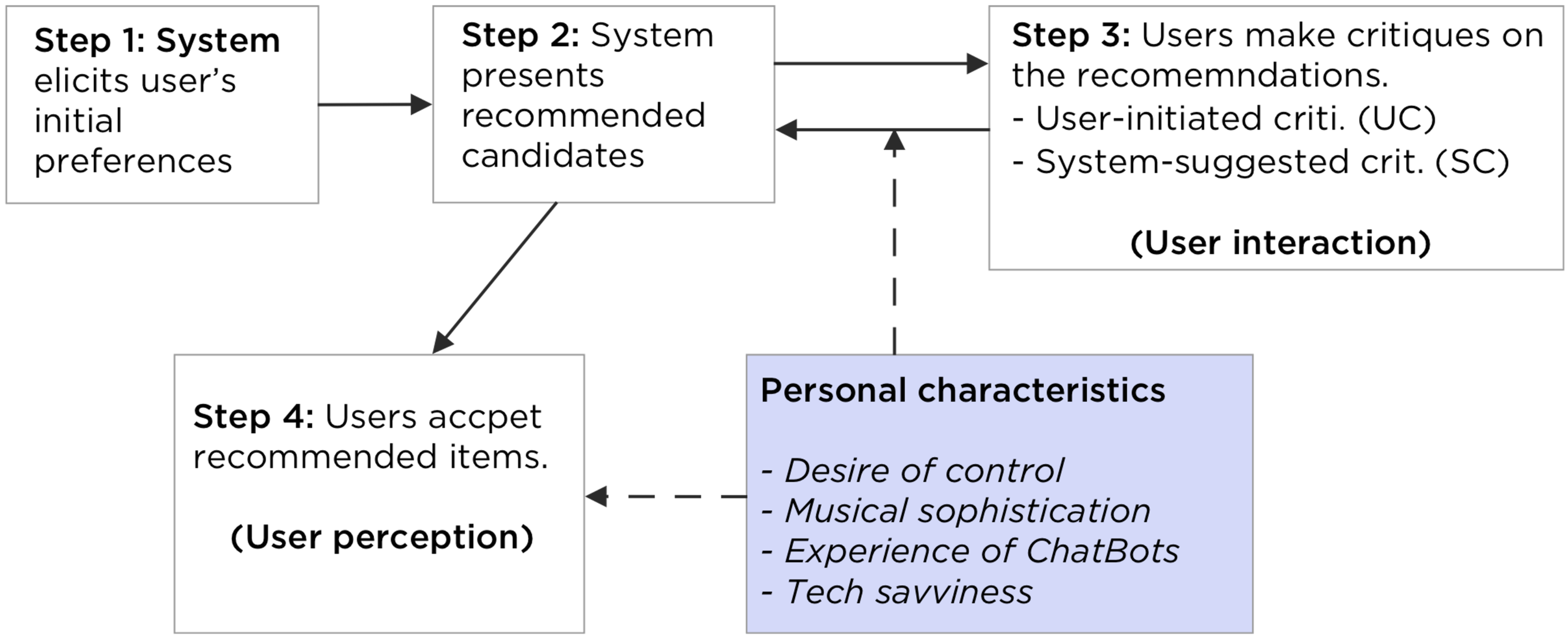
Empirical study: User perception of and interaction with critiquing-integrated DCRS

Classification of user intents for dialogue-based conversational recommendations

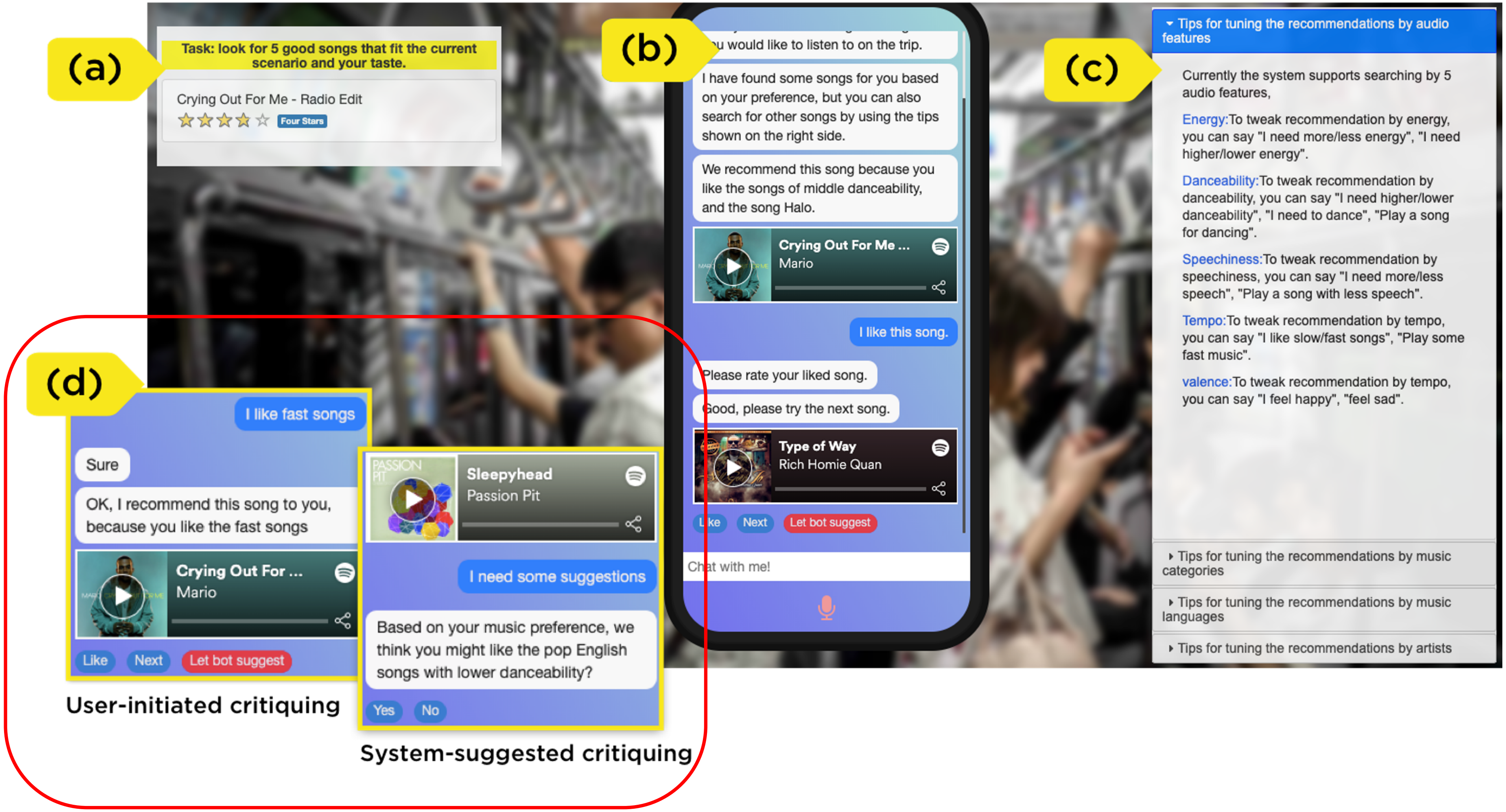
Prediction of user intents and satisfaction

Yucheng Jin, Wanling Cai, Li Chen, Nyi Nyi Htun, and Katrien Verbert. MusicBot: Evaluating Critiquing-based Music Recommenders with Conversational Interaction. In *Proceedings of 28th ACM International Conference on Information and Knowledge Management (CIKM'19)*, pages 951–960, Beijing, China, November 3-7, 2019.

Critiquing-based interaction in dialogue system



Interface design of our MusicBot



User Experiment

- Participants: 45 valid (19 female)
- User initiated (UC) critiquing vs. Hybrid critiquing (UC + SC)
- Experimental task

Find 5 ❤ songs in two scenarios and give ratings

subway

UC

party

HC

Watch Video Tutorial → Build User Profile → Pre-Study Questionnaire → Warm Up → Interact with MusicBot → Post-Study Questionnaire

Measurements

Question items
Q1: The items recommended to me matched my interests.
Q2: I easily found the songs I was looking for.
Q3: Looking for a song using this interface required too much effort (reverse scale).
Q4: The songs recommended to me are diverse.
Q5: I found it easy to inform the system if I dislike/like the recommended song.
Q6: I felt in control of modifying my taste using *MusicBot*.
Q7: I am confident I will like the songs recommended to me.
Q8: I like to give feedback on the music I am listening.
Q9: This music chatbot can be trusted.
Q10: I found the system easy to understand in this conversation.
Q11: In this conversation, I knew what I could say or do at each point of the dialog.
Q12: The system worked in the way I expected in this conversation.
Q13: I will use this music chatbot again.
Q14: Overall, I am satisfied with the chatbot.

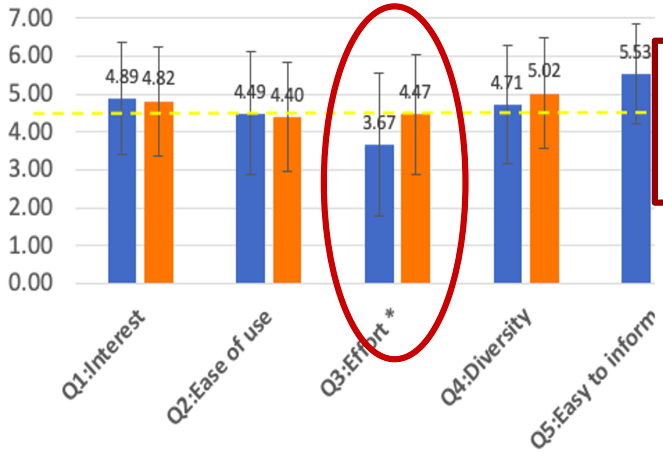
- Rating (stars) for the selected songs
- Completion time
- Dialog turns
- Listened songs
- Button clicks
- Messages by typing
- Messages by voice
- Words per utterances
- Unknown utterances

***ResQue:* User-centric** evaluation framework for recommender systems (Pu et al., 2011)

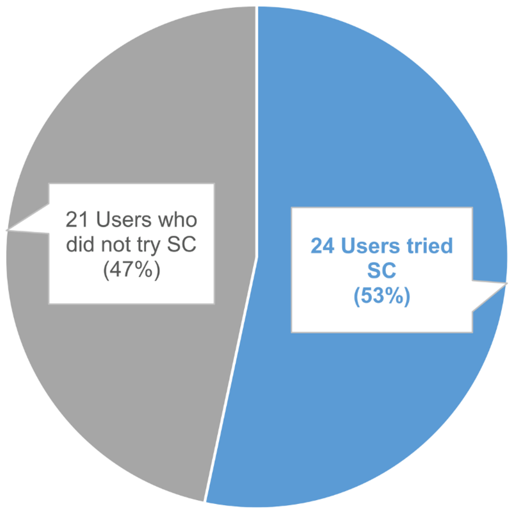
PARADISE: Evaluation framework for **spoken dialogue agents** (Walker et al., 1997)

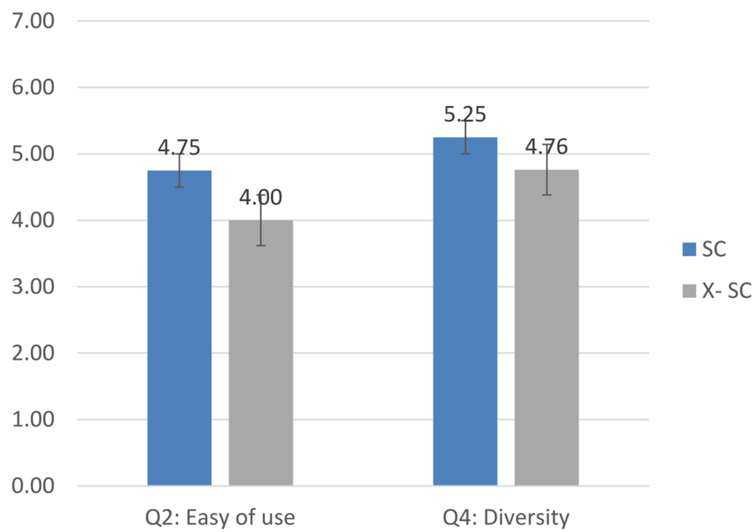
Objective behavioral variables

UC: User-initiated Critiquing
HC: Hybrid Critiquing (UC + SC)



Interaction metrics	UC (mean,sd)	HC (mean,sd)
Rating (stars)	(4.05, 0.47)	(4.08, 0.44)
Completion time* (minutes)	(5.40, 4.19)	(6.98, 4.16)
#Listened songs**	(10.67, 4.99)	(13.13, 6.09)
#Turns(times)**	(12.29, 8.21)	(16.11, 9.35)
#Btn(times)***	(9.18, 3.38)	(12.64, 7.07)
#Typing(times)	(3.09, 4.78)	(3.07, 4.21)
#Voice(times)	(1.24, 7.90)	(0.71, 2.97)
#Words	(2.13, 1.92)	(2.28, 1.84)
#Unknown utterances	(1.78, 6.46)	(0.78, 1.80)





Users who tried SC tend to perceive higher ease of use and diversity.

Effect of personal characteristics on user perceptions

PC	Q1:Interest	Q2:Ease of use	Q3:Effort	Q4:Diversity	Q5:Easy to inform	Q6:Control	Q7:Confidence
CE	0.15 (0.33)	0.14 (0.37)	0.07 (0.66)	0.03 (0.84)	-0.03 (0.86)	0.11 (0.46)	0.05 (0.73)
TS	-0.01 (0.98)	-0.13 (0.40)	**0.36 (0.02)***	0.10 (0.51)	-0.08 (0.59)	-0.19 (0.21)	-0.12 (0.43)
MS	**0.40 (0.01)***	0.25 (0.10)	-0.22 (0.14)	0.17 (0.26)	0.10 (0.53)	**0.31 (0.04)***	0.29 (0.05)
DFC	0.23 (0.14)	0.03 (0.84)	0.13 (0.41)	0.24 (0.11)	0.22 (0.15)	**0.35 (0.02)***	0.25 (0.10)

PC	Q8:Feedback	Q9:Trust	Q10:Understand	Q11:Difficulty	Q12:Expected	Q13:Intent to reuse	Q14:Satisfaction
CE	0.06 (0.70)	-0.01 (1.00)	-0.07 (0.65)	0.02 (0.88)	0.06 (0.69)	0.21 (0.17)	0.10 (0.52)
TS	0.16 (0.29)	0.07 (0.66)	-0.12 (0.42)	-0.04 (0.77)	0.04 (0.78)	-0.12 (0.42)	-0.19 (0.10)
MS	**0.55 (<0.001)*****	**0.37 (0.01)***	0.09 (0.57)	0.13 (0.38)	0.23 (0.14)	**0.31 (0.04)***	0.22 (0.15)
DFC	0.06 (0.68)	0.16 (0.29)	**0.30 (0.04)***	**0.38 (0.01)***	0.22 (0.14)	0.28 (0.06)	0.20 (0.19)

Music Sophistication (+): Interest matching, Control, Trust, Intention to Give Feedback and Reuse
Desire For Control (+): Control, Easy to Understand and Use

Summary

- Combining UC and SC in a conversational user interface may increase user engagement and likelihood of finding more (diverse) songs.
- Designers should consider MS and DFC as key personal characteristics in interaction design for critiquing-based music recommenders.

- ***Limitations***
  - Small-scale user data
  - Not "smart" enough to understand user intentions

Our Focuses

Empirical study: User perception of and interaction with critiquing support in DCRS

Classification of user intents for dialogue-based conversational recommendations

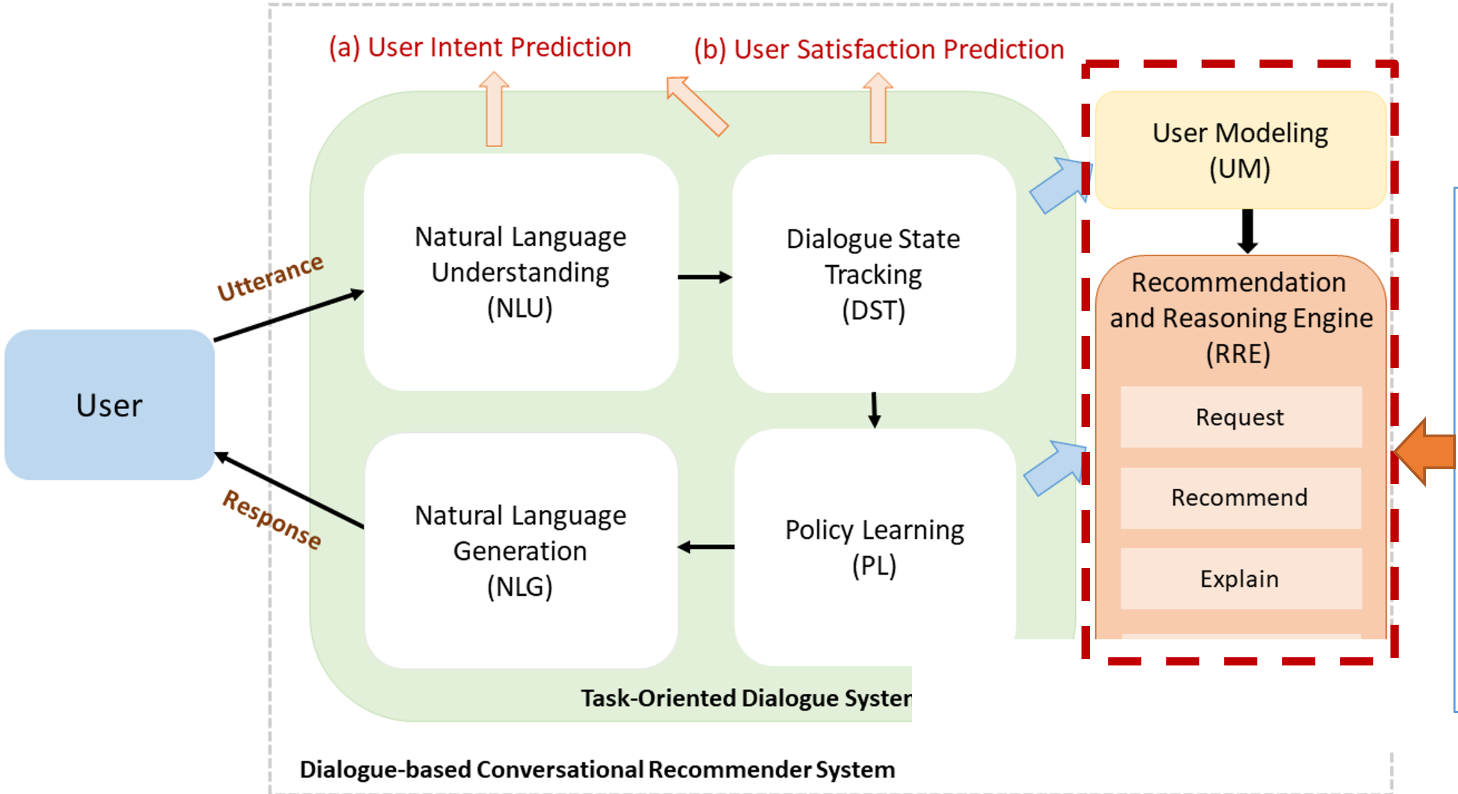
Prediction of user intents and satisfaction

Wanling Cai and Li Chen. 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. [Best Student Paper Award]

User Intent and Satisfaction Prediction

User intent indicates the **goal** or **intention** that users have during their interaction with the system (Rose and Levinson, WWW 2004)

User satisfaction indicates **if the user's goal is fulfilled** to some extent (Hashemi *et al.*, CIKM 2018)



Implications:

1. More accurately model users' preference
2. Allow the system to select more appropriate action

Recommendation Dialogue Data

Recommender:	Hi how are you today? I heard you might be interested in a movie. Any particular genre?
Seeker:	Hi, I'm good, just looking for a nice horror movie. Nothing to gory, I liked Beetlejuice.
Recommender:	hmm. I don't know too many horror movies. I did watch The Birds.
Seeker:	Yeah I've seen the birds it was okay but I felt like it was too old for my tastes.
Recommender:	border line with suspense might be something like Hannibal or The Silence of the Lambs.
Seeker:	I didn't like any of those movies, too much talking.
Recommender:	okay. Well, how about Saw ?
Seeker:	Something more like Final Destination.
Recommender:	Do you like any other genres?
Seeker:	The Saw was okay, I felt like it was too violent. I really love like fantasy horror, maybe Ghost.
Recommender:	I've heard that is a good one. Have you seen Signs ?
Seeker:	I heard about that but didn't watch it.
Recommender:	Mel Gibson in it. I've heard it is excellent.
Seeker:	okay, great I will check it out. thank you.

ReDial Dataset

human-human dialogues

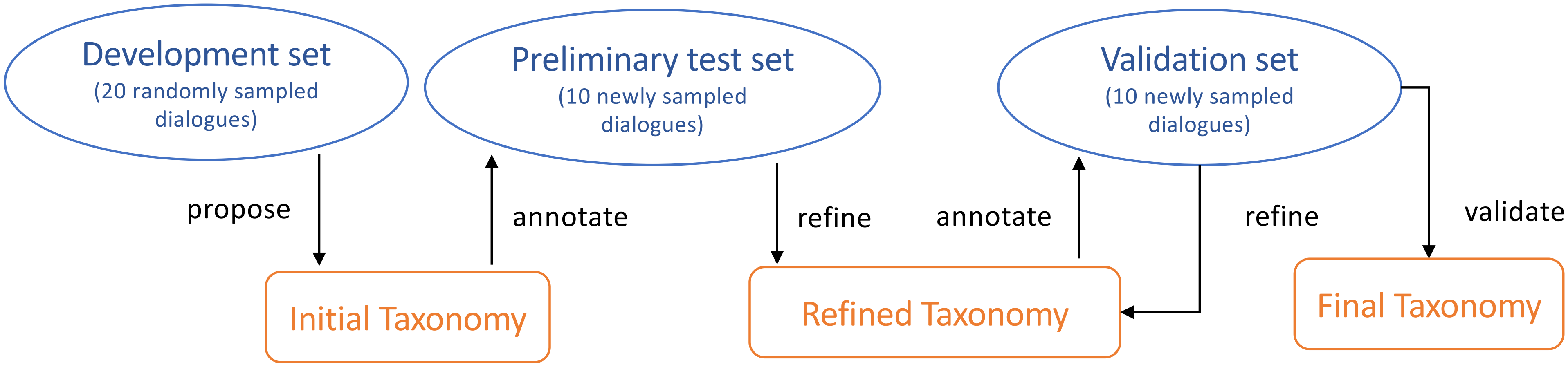
centered around movie recommendations

(Li *et al.*, NIPS 2018)

ReDial dataset: https://redialdata.github.io/website/

Statistics of our selected dialogue data

Items	**SAT-Dial** (with user-satisfied recommendation)	**unSAT-Dial** (without user-satisfied recommendation)
# Conversations	253	83
# Human seekers	125 (# utterances: 1,711)	59 (# utterances: 550)
# Human recommenders	151 (# utterances: 1,747)	68 (# utterances: 575)
# Suggested movies per dialogue	4.57	4.51
# Turns per dialogue	mean=6.58, min=3, max=19	mean=6.49, min=3, max=12
# Words per utterance	mean=11.29, min=1, max=72	mean=10.72, min=1, max=69



Intent (Code)	Description	Percentage
Ask for Recommendation		**18.26%**
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%
Add Details		**18.58%**
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%
Give Feedback		**61.92%**
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%
Others	Greetings, gratitude expression, or chit-chat utterances.	14.55%

User Intent Prediction

- Multi-label Classification Problem

  E.g., "I did see that one, but I didn't really like it. I do love 80s movies though." -> two intents: Reject and Critique-Add

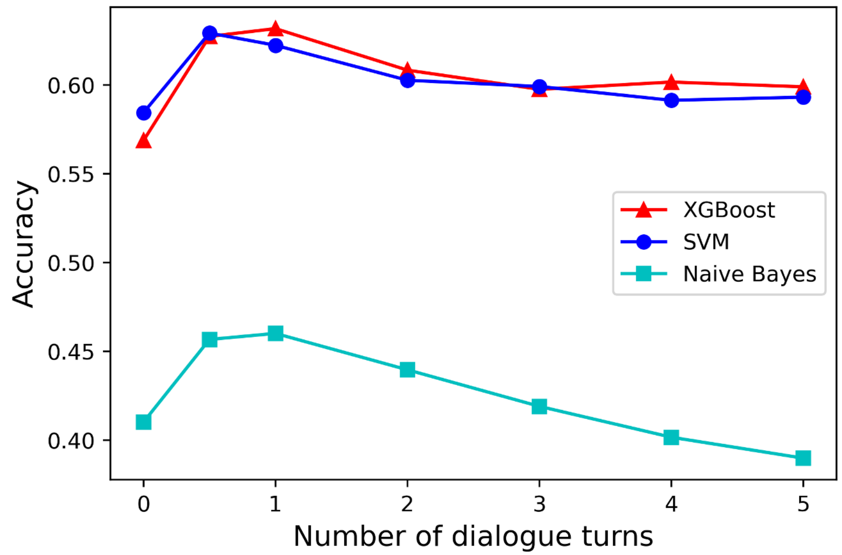
- *Classification Models*
  - 8 Machine Learning Models (e.g., LR, SVM, Naive Bayes, XGBoost, MLP, etc.) and 2 Deep Learning Models (CNN and Bi-LSTM)

- Features

Category	*Features*
Content	TF-IDF, Name Entity, # Relevant Items
Discourse	POS, 5W1H Question, Question Mark, Exclamation Mark, Utterance Length
Sentiment	Thanks, Sentiment Score, Opinion Lexicon
Context	Absolute Position, Utterance Similarity, Previous user intents & recommendation actions

Methods	Binary Relevance				Classification Chain				Label Powerset			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Logistic Regression	0.5796	0.7160	0.6148	0.6612	0.6111	0.6898	0.6322	0.6596	0.6198	0.6791	0.6053	0.6400
SVM	0.5597	0.6701	0.6047	0.6332	**0.6293**	0.7179	0.6340	0.6730	0.6048	0.6004	0.6123	0.6056
Naive Bayes	0.4438	0.5137	0.5705	0.5400	0.4567	0.5137	0.5793	0.5439	0.5365	0.5989	0.5542	0.5755
Decision Tree	0.5264	0.5187	0.6778	0.5871	0.5356	0.5513	0.6325	0.5887	0.4515	0.4706	0.4755	0.4729
Random Forest	0.5742	0.5962	**0.7029**	0.6449	0.5968	0.6372	**0.6817**	0.6583	0.4794	0.4748	0.5096	0.4913
XGBoost	**0.5970**	**0.8169**	0.6007	**0.6919**	0.6274	**0.7957**	0.6268	**0.7010**	**0.6199**	**0.6868**	**0.6109**	**0.6463**
MLP	0.4773	0.7922	0.4743	0.5928	0.5079	0.7780	0.5045	0.6115	0.6157	0.6837	0.6029	0.6407

	Cont	Disc	Sent	Context	Acc	Prec	Rec	F1
1 Category	✓				**0.4726**	**0.7165**	**0.4868**	**0.5793**
		✓			0.3918	0.5224	0.3841	0.4426
				✓	0.3407	0.5020	0.3343	0.4011
			✓		0.1993	0.3241	0.2044	0.2498
2 Categories	✓			✓	**0.5603**	**0.7669**	**0.5627**	**0.6488**
		✓		✓	0.5438	0.6946	0.5346	0.6039
	✓	✓			0.5291	0.7381	0.5350	0.6201
	✓		✓		0.4921	0.7289	0.5067	0.5972
			✓	✓	0.4587	0.6209	0.4518	0.5229
		✓	✓		0.4268	0.5553	0.4208	0.4787
3 Categories	✓	✓		✓	**0.6119**	**0.7913**	**0.6112**	**0.6896**
	✓		✓	✓	0.5870	0.7760	0.5887	0.6692
		✓	✓	✓	0.5698	0.7188	0.5569	0.6275
	✓	✓	✓		0.5415	0.7418	0.5500	0.6313
All	✓	✓	✓	✓	**0.6274**	**0.7957**	**0.6268**	**0.7010**



Context features can help boost the prediction performance

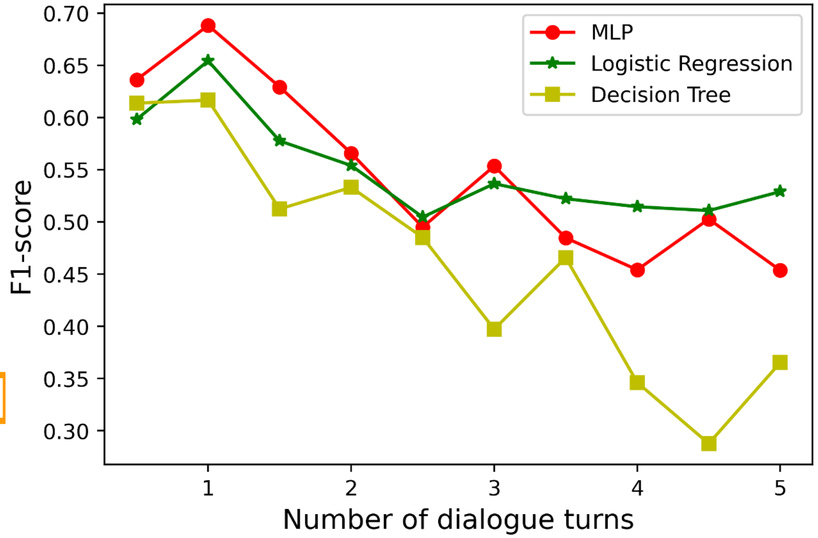
User Satisfaction Prediction

- Binary Classification Problem
- Classification Models
  - 8 Machine Learning Models: LR, SVM, Naive Bayes, XGBoost, MLP, etc.
- Features
  - Dialogue behavior features (i.e., user intents and recommender actions)
  - Utterance-level features (i.e., content, discourse, and sentiment features)

Category	*Features*
Content	TF-IDF, Name Entity, # Relevant Items
Discourse	POS, 5W1H Question, Question Mark, Exclamation Mark, Utterance Length
Sentiment	Thanks, Sentiment Score, Opinion Lexicon

Methods	Cont	Disc	Sent	Dial	Prec	Rec	F1
Logistic Regression	✓	✓		✓	0.8488	**0.5806**	0.6795
SVM		✓		✓	0.8778	0.5556	0.6629
Naive Bayes				✓	0.8833	0.5556	0.6651
Decision Tree				✓	0.7109	0.5528	0.6167
Random Forest				✓	0.8862	0.5306	0.6503
XGBoost				✓	0.7897	0.5653	0.6426
MLP				✓	**0.8990**	0.5681	**0.6884**
KNN			✓	✓	0.8850	0.5181	0.6427

Comparison of Classification Models



Method	Cont	Disc	Sent	Dial	Prec	Rec	F1
MLP				✓	**0.8990**	**0.5681**	**0.6884**
	✓				0.6551	0.4944	0.5501
		✓			0.5570	0.3486	0.4122
			✓		0.6067	0.2681	0.3606
	✓	✓	✓	✓	0.7995	0.5444	0.6292

Comparison of Feature Categories

- Classification Models: MLP (best precision & F1)
- Effective Features: Dialogue behavior features (i.e., user intents and recommender actions)

Summary

- Taxonomy established for user intents in dialogue-based CRS
- User intent prediction: XGBoost and SVM can achieve outperforming accuracy by unifying four feature categories (i.e., content, sentiment, discourse, and context)
- User satisfaction prediction: Leveraging both user intents and recommender actions enables some model like MLP to achieve competitive accuracy

Future Work

- Prediction
  - To verify the taxonomy's generalizability to other domains
  - To measure the performance of deep learning (DL) methods with more labelled dialogue data
  - To investigate the temporal sequence of utterances/responses within a dialogue, for further improving the prediction accuracy
- System design
  - To integrate more feedback/critiquing aids to match to users' intents
  - To study how system-suggested critiques could guide users to explore (diverse) items

Thanks!

Ms. Wanling Cai

Dr. Yucheng Jin

Contact info:

Dr. Li CHEN

lichen@comp.hkbu.edu.hk

Homepage:

http://www.comp.hkbu.edu.hk/~lichen/

Intent Annotation of Recommendation Dialogue (**IARD**) dataset is publicly available at:
https://www.comp.hkbu.edu.hk/~lichen/download/IARD_dataset.html