MusicBot: Evaluating Critiquing-Based Music Recommenders with Conversational Interaction

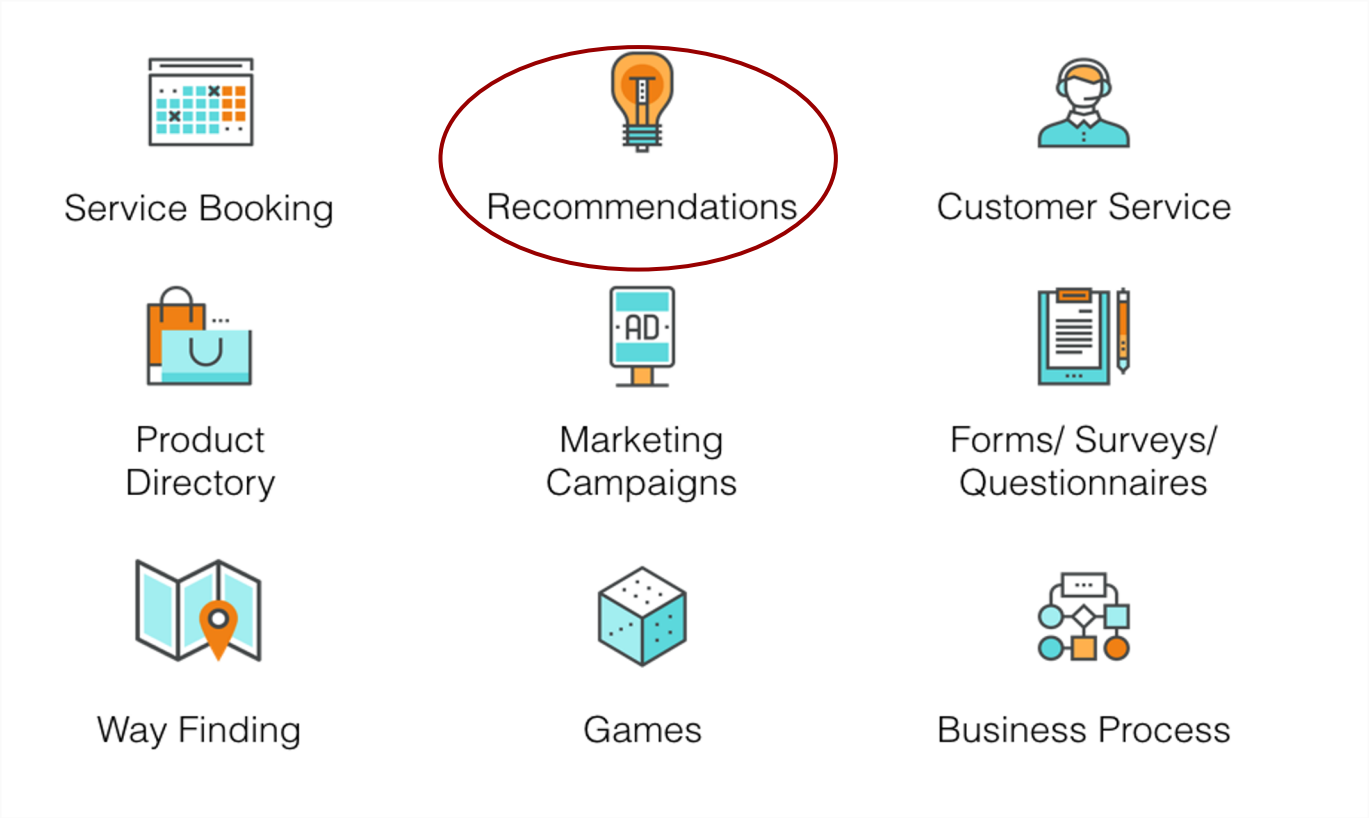
Yucheng Jin
Lenovo Research

Wanling Cai, Li Chen
Hong Kong Baptist University

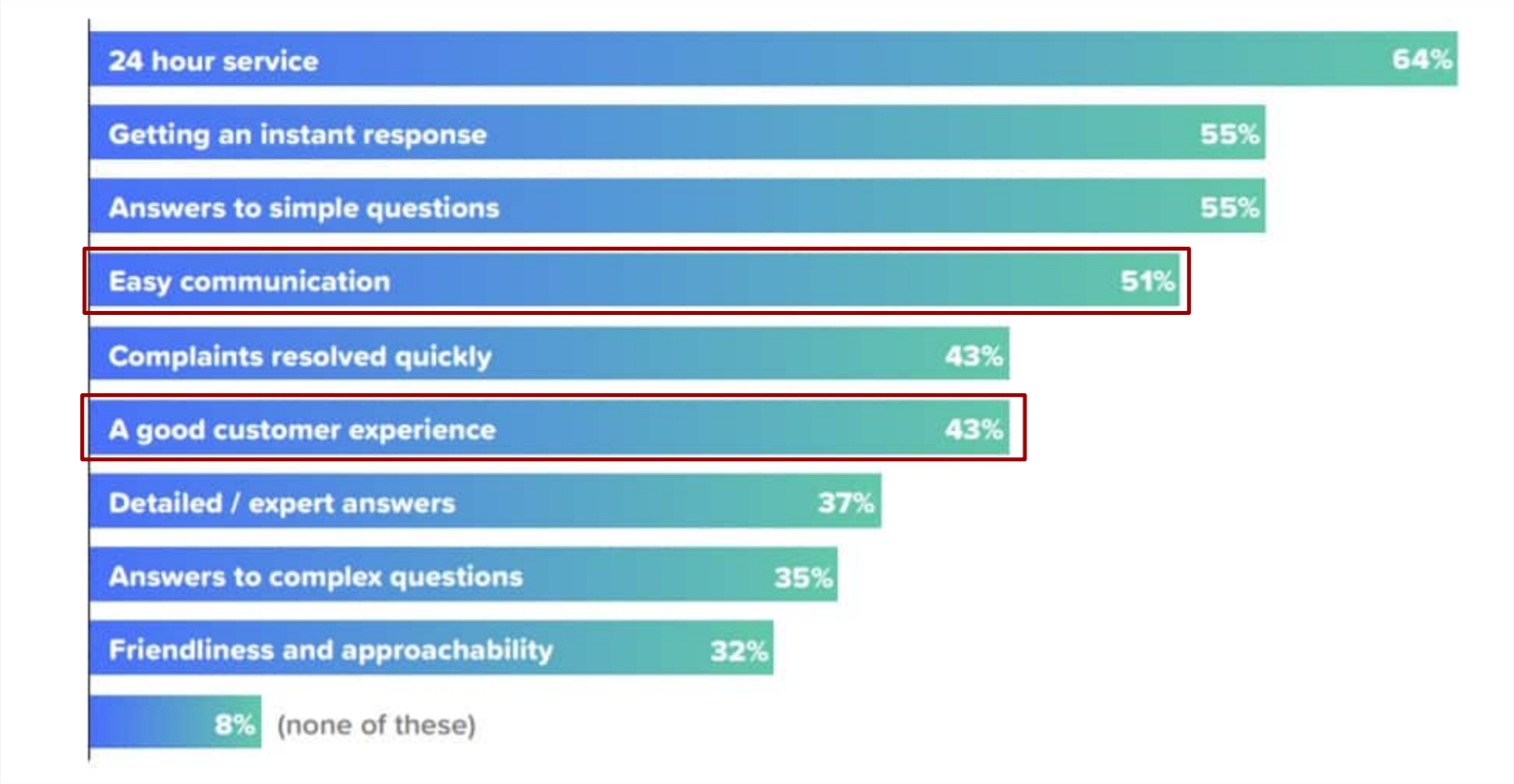
Nyi Nyi Thun, Katrien Verbert
University of Leuven

Background

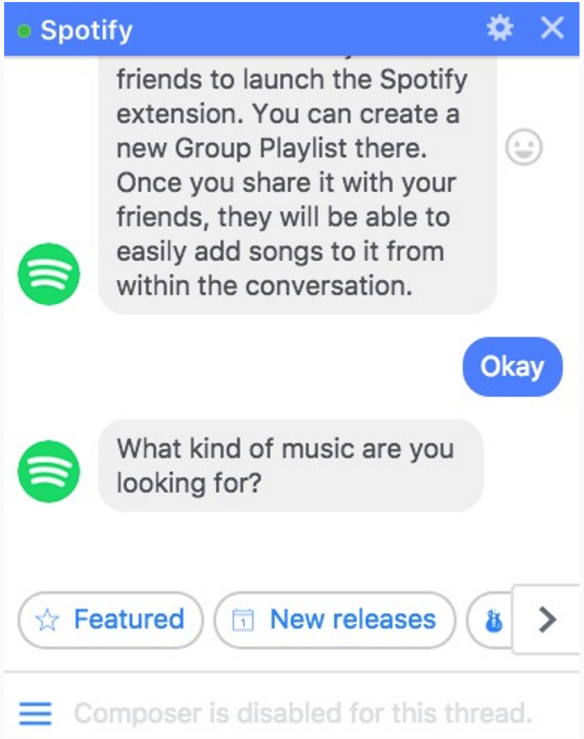
Chatbots



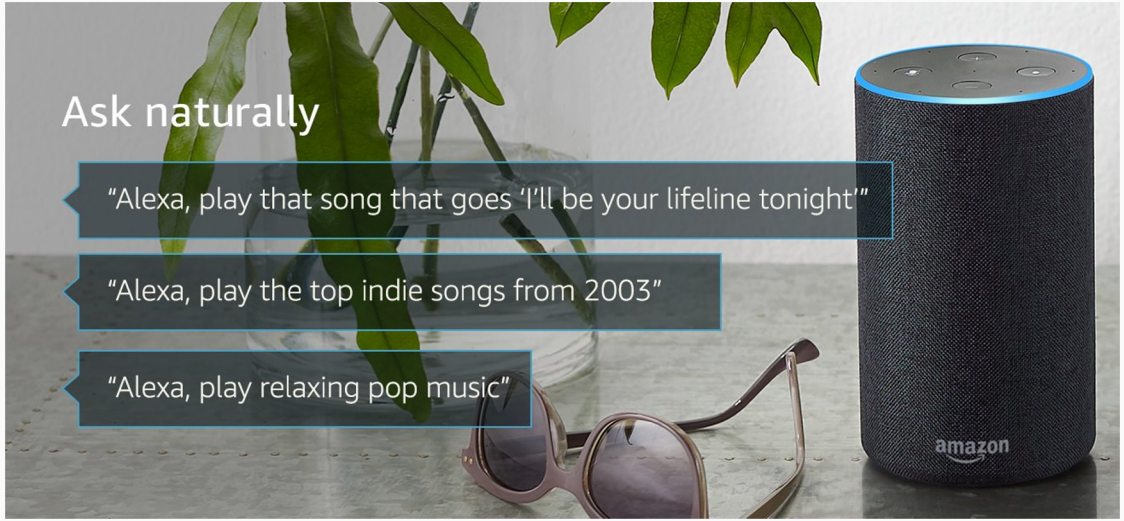
https://chatbot.com.hk/img/use-cases.png



https://blog.aimultiple.com/chatbot-benefits/



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



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

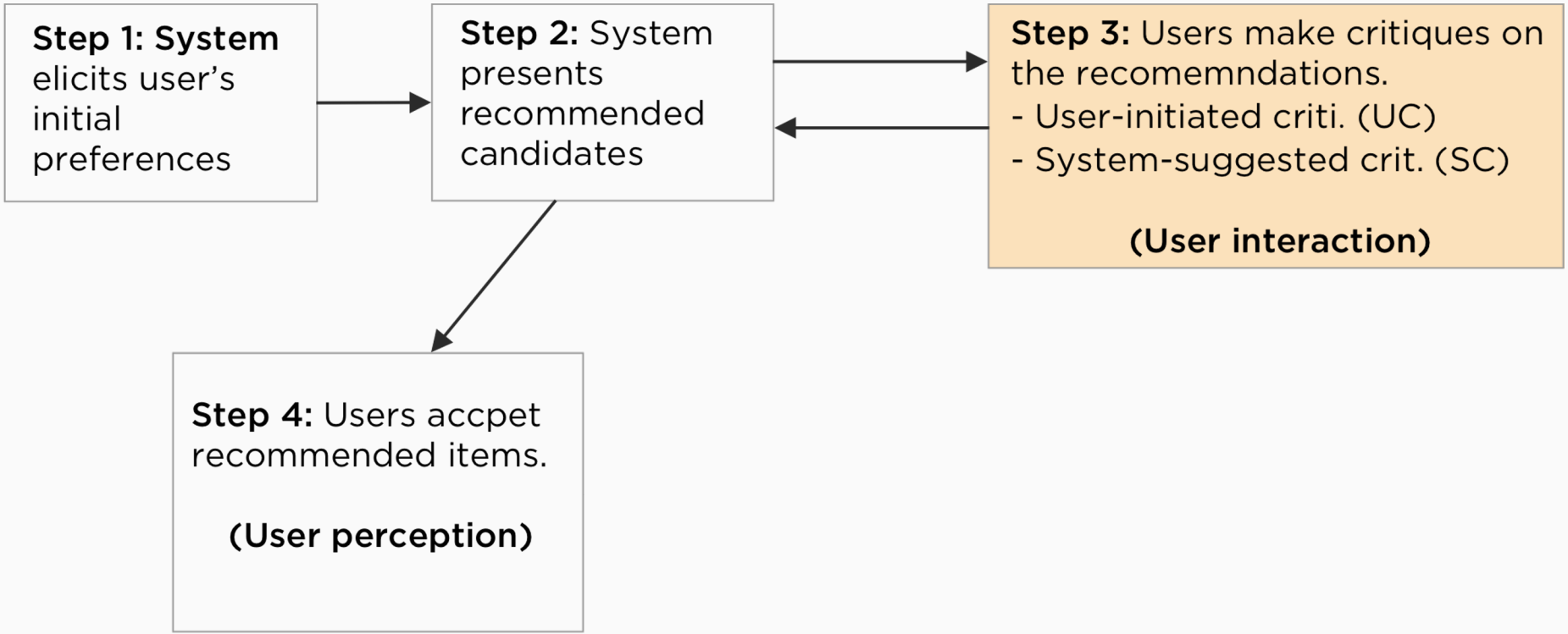
Limitations:

- User-initiated
- Simple commands
- Single round interaction

Critiquing-Based Recommenders

*"Critiquing-based recommender systems **elicit users' feedback**, called **critiques**, which they made on the recommended items. Through the use of the critiquing feedback, the recommender systems are able to **more accurately** learn the users' profiles, and therefore suggest **better recommendations**."* (Chen and Pu, 2011)

A typical interaction flow of critiquing-based recommender systems



(a)

To find similar products with better values than this one

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

The product being critiqued

(b)

We have the following

1. Less Optical Zoom and Thinner and Lighter Weight	Explain	Show Products
2. Different Manufacturer and Lower Resolution and Cheaper	Explain	Show Products
3. Larger Screen Size and More Memory and Heavier	Explain	Show Products

System-suggested compound critiques

(c)

OR would you like to improve some value(s) by yourself?

	Keep	Improve	Take any suggestion
Manufacturer	Canon	Sony	
Price	$424.15	less expensive	
Resolution	5.3 M pixels	higher	
Optical Zoom	12x	more zoom	
Removable Flash Memory	16 MB	more memory	
LCD Screen Size	1.8 in	larger	
Thickness	2.97 in	thinner	
Weight	404.7 g	lighter	

Show Results Reset

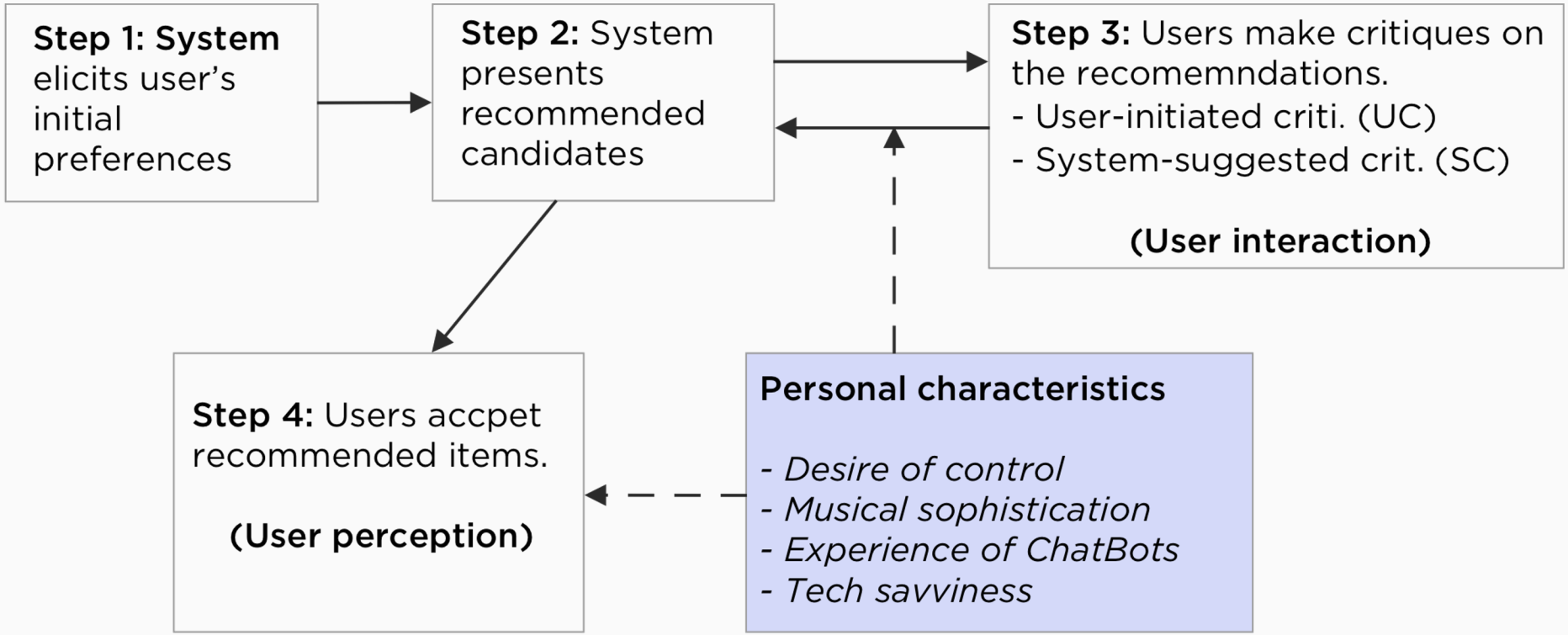
User-initiated critiquing facility

+ confidence in decision making, decision accuracy

-- objective effort of making decision

(Chen and Pu, 2007)

A typical interaction flow of critiquing-based recommender systems



Personal Characteristics

- Desire for Control (DFC)
  Degree of control individuals perceive towards outcomes (Burger 1986)
  DFC → (+) task performance

- Musical Sophistication (MS)
  Assessing musical sophistication index for general population (Müllensiefen et al., 2014)
  MS → (+) acceptance and perceived diversity (Jin et al., 2019)

- Tech Savviness (TS)
  Confidence in trying new tech
  TS → mobile information search behaviour (Dewan and Benckendorff 2013)

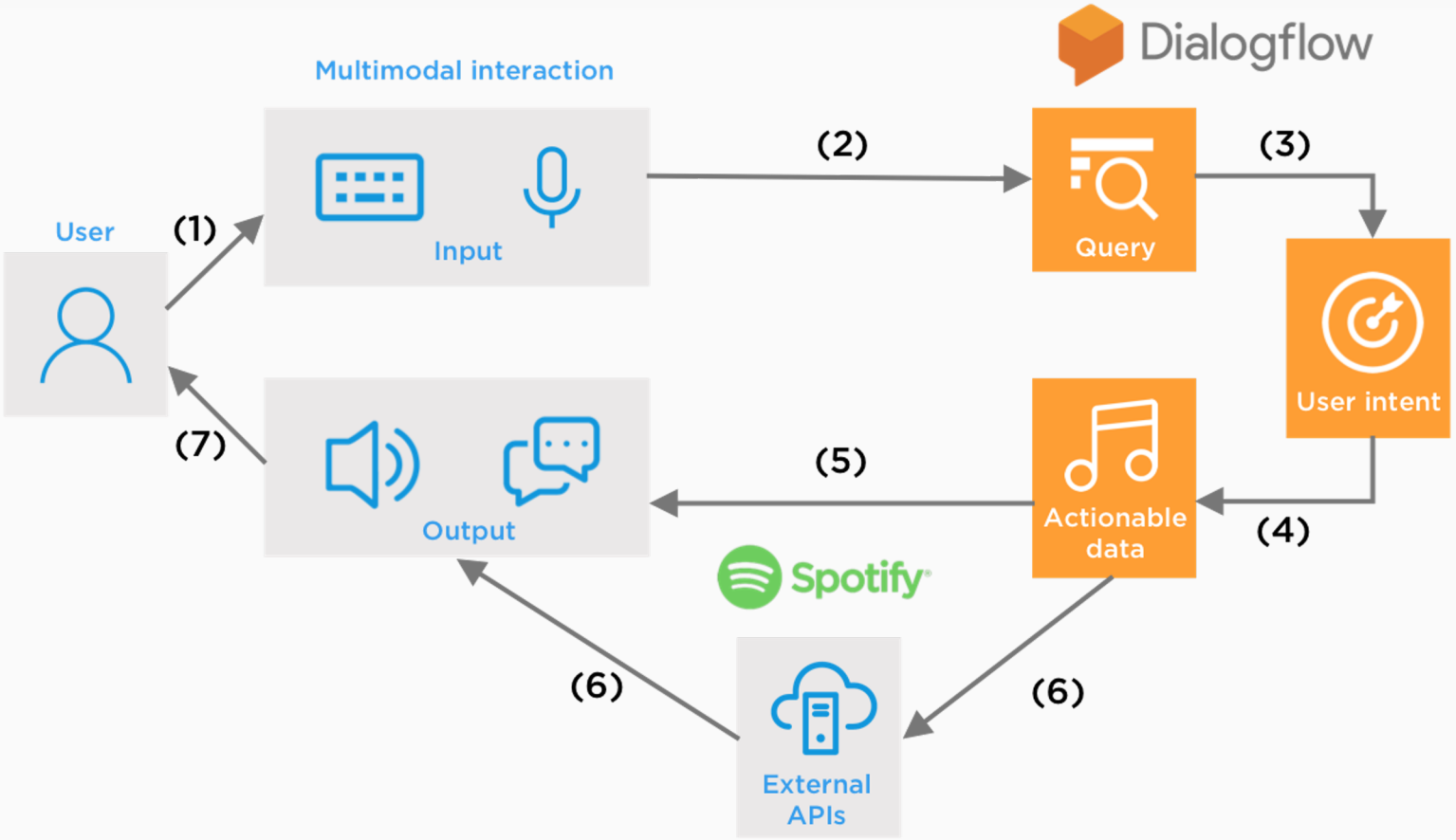
- Chatbot Experience (CE)

RQ1: Which critiquing setting, user-initiated critiquing (UC) versus hybrid critiquing (HC), is better suited for controlling music recommendations?

RQ2: Which personal characteristics might influence user’s perception and interaction of recommendations?

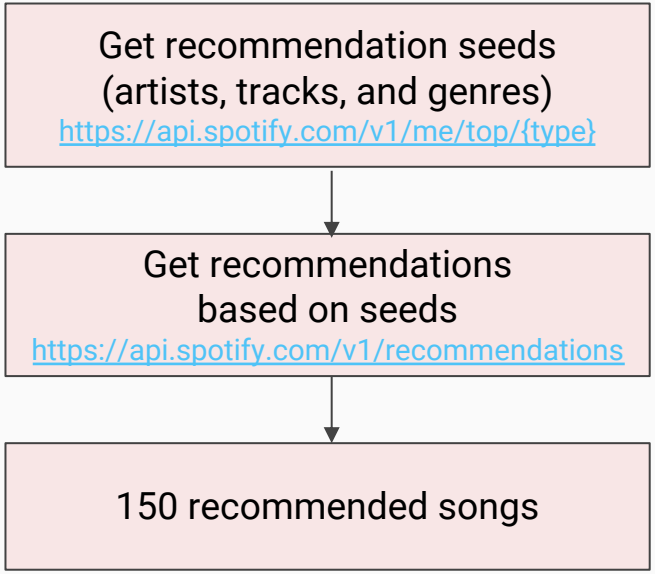
System Design

System Architecture



Algorithms

Recommendation Algorithm



Spotify WEB API

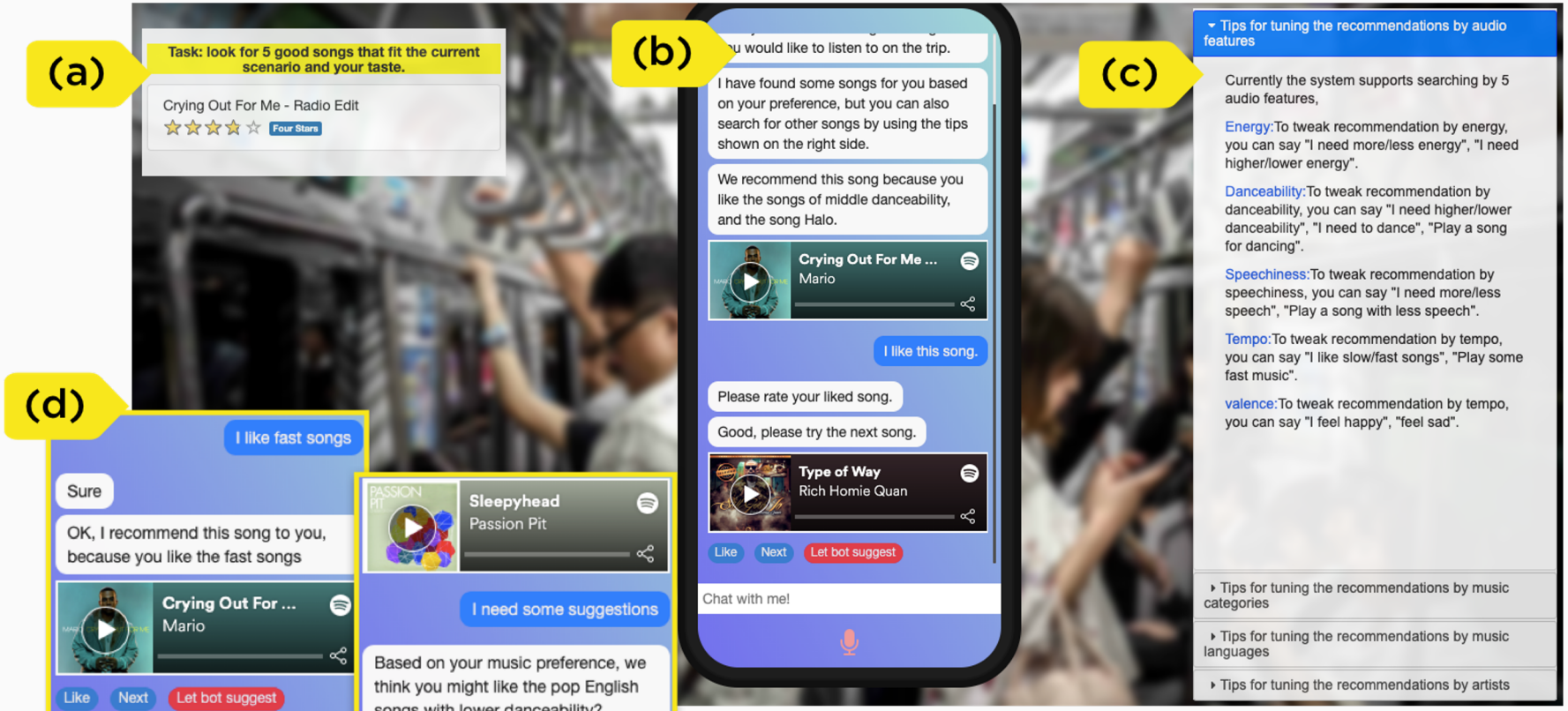
Critiquing-based Algorithm

- User-initiated critiquing
  E.g., “I need lower energy"
  (genre, language, artist, danceability, speechiness, energy, valence, and tempo)

- System-suggested critiquing (Chen and Pu, 2007)
  E.g., “Based on your music preference, we think you might like English songs with higher danceability and higher energy."

1. Critique pattern vector (e.g., {(energy, higher), (danceability, similar)})
2. Association rule mining algorithm (i.e., Apriori algorithm)
3. Multi-attribute utility theory (MAUT)
4. A set of personalized and diversified critiques

Conversational User Interface



User-initiated critiquing (UC)

System-suggested critiquing (SC)

Experimental Design

User-initiated critiquing System (UC)

VS

Hybrid-critiquing System (HC)

Support both UC and
System-suggested critiquing (SC)

User Experiment

Participants: 51(45)

Recruitment

- Personal contacts
- Research groups
- University contacts

Reward

A prize draw
(each voucher: 10 USD)

Age

- 20-30 (36)
- 30-40 (6)
- 41-50 (1)
- > 50 (2)

Gender

- Female = 19
- Male = 26

User Experiment

Experimental task

Find 5 ❤ songs
in two scenarios
and give ratings

subway

UC

party

HC

Desire for Control (DFC)
(Bugger et al., 1979)
Musical Sophistication (MS)
(Gold-MSI, Müllensiefen et al., 2014)
Tech Savviness (TS)
("I am confident when it comes to try new technology.")
Chatbot Experience (CE)
("I often use a chatbot (such as Siri, Cortana) on my personal devices.")

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

Subjective Experience (Post-study Questionnaire *7-point Likert Scale*)

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.

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

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

User Interaction Behavior (Interaction Logs)

- Rating (stars) for the selected songs
- Completion time
- Dialog turns

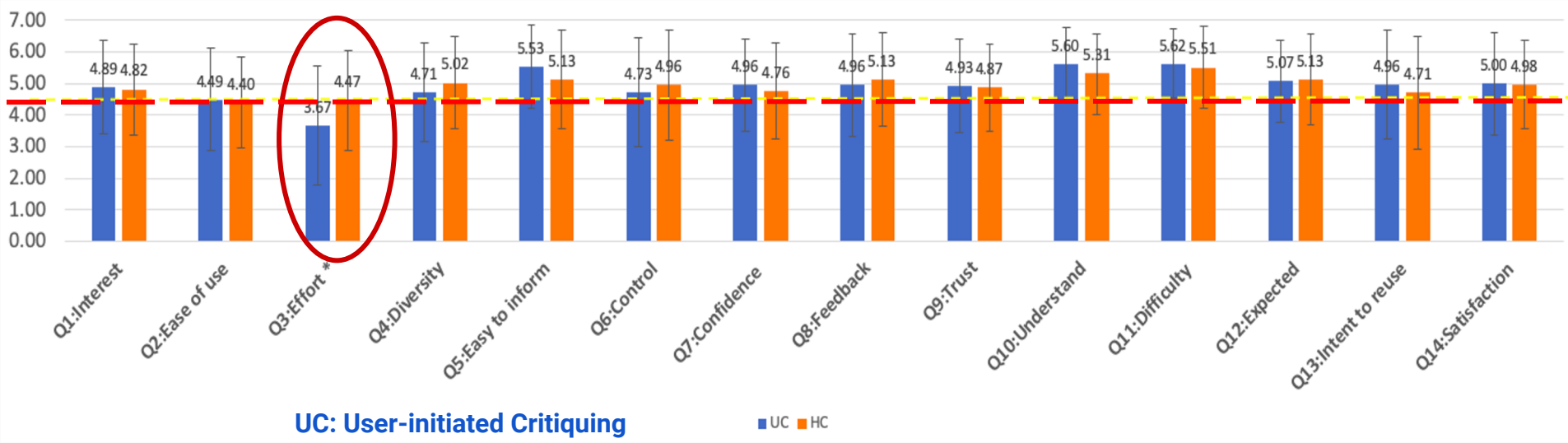
The number of

- Listened songs
- Button clicks
- Messages by typing
- Messages by voice
- Words per utterances
- Unknown utterances

Results & Discussion

Subjective Experience

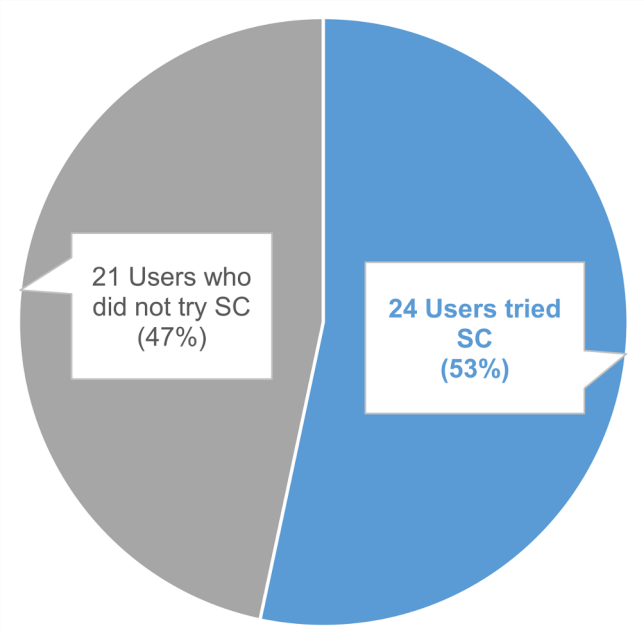
Usability and User Assessment Results

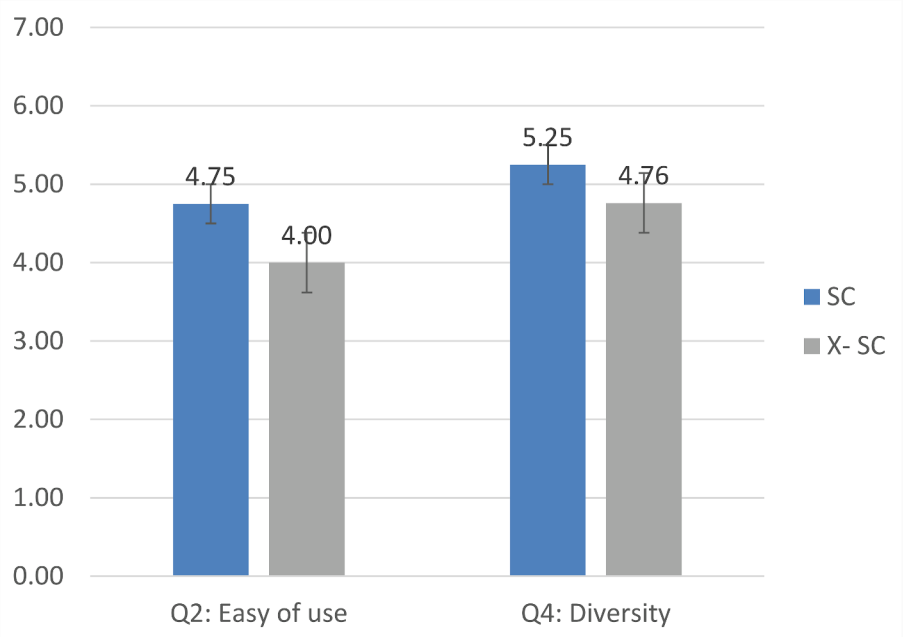


UC: User-initiated Critiquing

HC: Hybrid Critiquing (UC + SC)

Further Analysis in HC





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

User Interaction Behavior

Descriptive Statistics for User Interaction Data

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)

HC leads to more dialogue turns, more completion time, more listened songs.

RQ1: Which critiquing setting, UC versus HC, is better suited for controlling music recommendations?

Our Suggestion

Combining UC and SC in a conversational user interface may increase ***user engagement*** *and likelihood of* ***finding more (diverse) songs****.*

Personal Characteristics

Table 5: The Effect of PC on Users' Perceptions of Recommendations measured by Pearson correlation coefficient.

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)

MS(+): Interest matching, Control, Trust, Intention to Give Feedback and Reuse.

DFC(+): Control, Easy to Understand and Use.

RQ2: Which personal characteristics might influence the user's perception and interaction of recommendations?

Our Suggestion

*Designers should consider **MS** and **DFC** as key personal characteristics in conversational interaction design for critiquing-based music recommenders.*

Limitations and Conclusion

Limitations

- Not “Smart” enough to understand user intentions
- Small-scale user data

Conclusion

- *Online evaluation* of conversational agent for music recommender combing *two critiquing techniques*
- HC tends to increase users' *perceived diversity and user engagement (time spend on the system)*
- Two influential characteristics, *musical sophistication (MS) and desire for control (DFC)*

Thanks!
Any questions?

Yucheng Jin
jinyc2@lenovo.com

Wanling Cai
cswlcai@comp.hkbu.edu.hk

Acknowledgement
Special thanks to SIGIR Student Grants