Personalized XAI (Explainable AI)

Cristina Conati

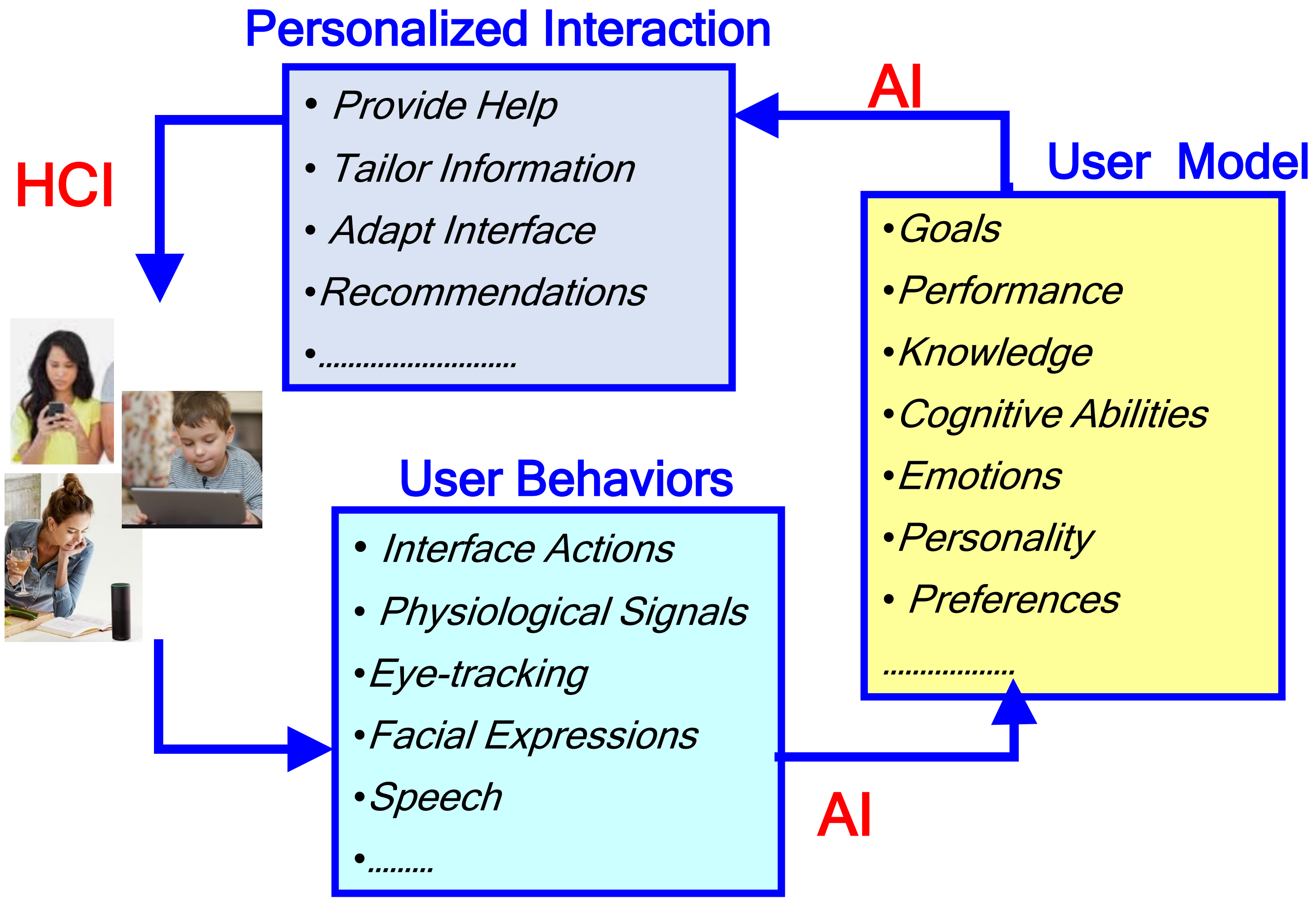
University of British Columbia

AI-Driven Personalization

- Supporting AI-Human collaboration via AI artifacts that can
  - understand relevant properties of their users (e.g.needs/states/abilities)
  - personalize the interaction accordingly

AI-Driven Personalization

- Supporting AI-Human collaboration via AI artifacts that can
  - understand relevant properties of their users (e.g.needs/states/abilities)
  - personalize the interaction accordingly



Personalized Interaction

- Provide Help
- Tailor Information
- Adapt Interface
- Recommendations
- ....

HCI

AI

User Model

- Goals
- Performance
- Knowledge
- Cognitive Abilities
- Emotions
- Personality
- Preferences

....

User Behaviors

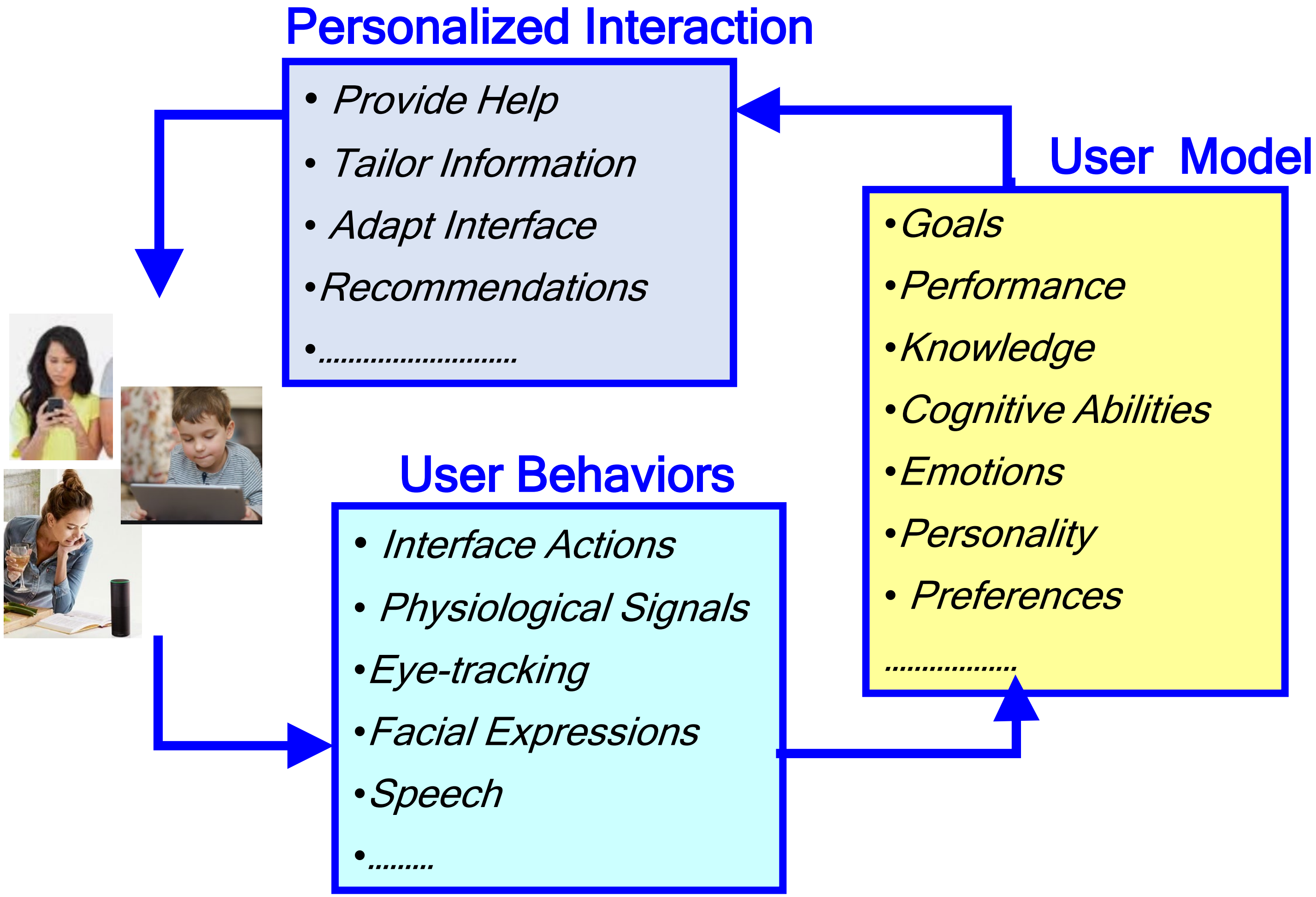
- Interface Actions
- Physiological Signals
- Eye-tracking
- Facial Expressions
- Speech
- ....

AI

- Substantial research for several decades
  - Recommender systems
  - Smart homes
  - Personalized Health
  - Assistive technology
  - Intelligent Tutoring Systems (ITS)
- Explosion of interest in recent years because of the new AI renaissance

AI-Driven Personalization

How to preserve transparency, user control and trust?



Personalized Interaction

- *Provide Help*
- *Tailor Information*
- *Adapt Interface*
- *Recommendations*
- *.........................*

User Model

- *Goals*
- *Performance*
- *Knowledge*
- *Cognitive Abilities*
- *Emotions*
- *Personality*
- *Preferences*

....

User Behaviors

- *Interface Actions*
- *Physiological Signals*
- *Eye-tracking*
- *Facial Expressions*
- *Speech*
- *.........*

Explanations of AI-driven Systems

- Explainable AI (XAI): can we increase AI interpretability, transparency, trust by enabling AI systems to explain their behaviors
  - for model developers
  - for end users
- Some results that explanations can be useful [e.g., Herlocker et al., 2000; Lawlor et al., 2015, Wang et al 2018, work from this group].
- But also evidence that they are not always wanted or effective [e.g. Herlocker et al, 2000, Bunt et al., 2012; 2007. Millecamp et al 2019, Putnam and Conati, 2019]

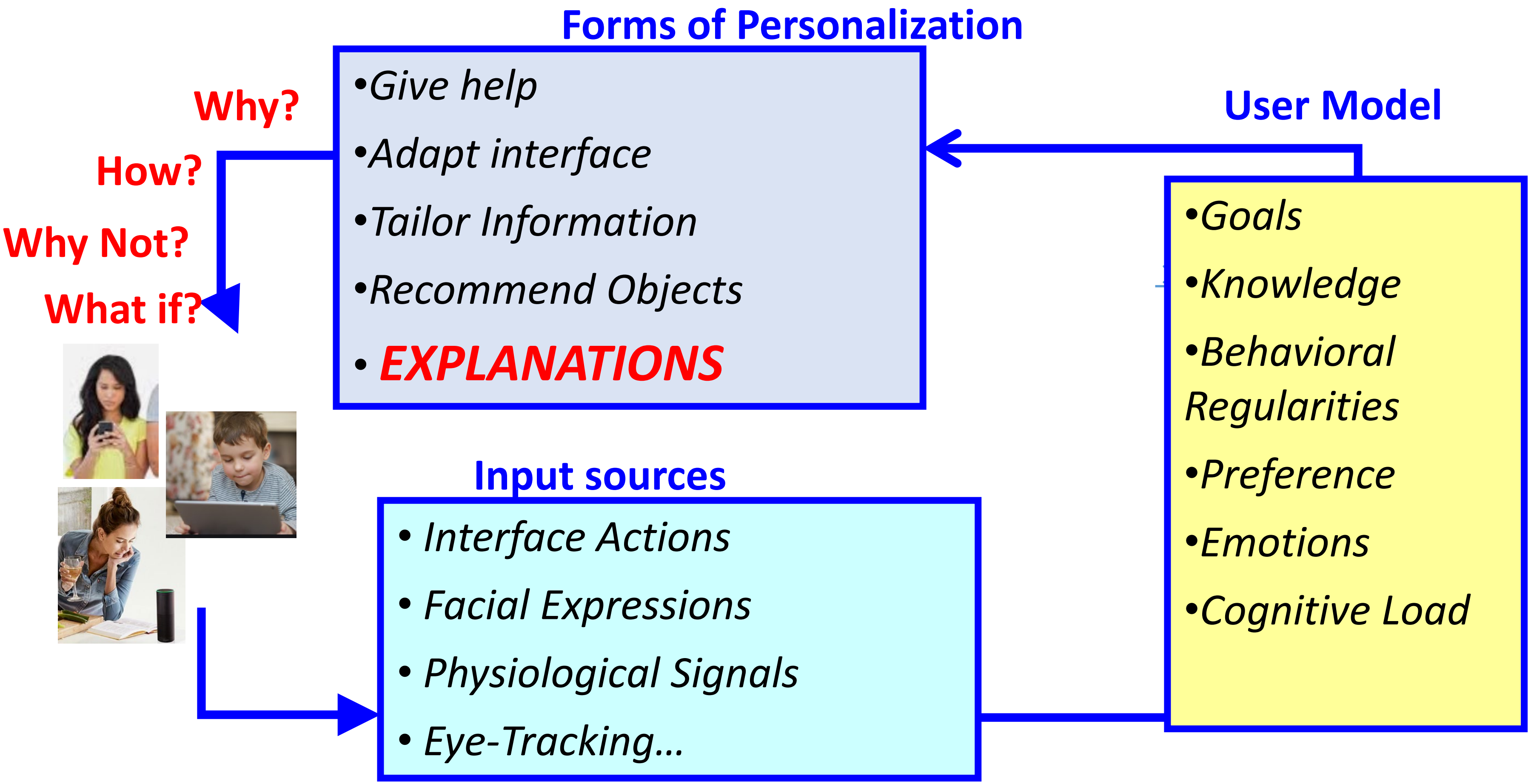
One-size-fits-all AI explanations do not work

Vision: Personalized XAI

- Intelligent systems that understand to whom, when and how to provide explanations of their behaviors.
- Good UI tools to interact with the explanations

Vision: Personalized XAI

- Intelligent systems that understand to whom, when and how to provide explanations of their behaviors.
- Good UI tools to interact with the explanations



Forms of Personalization

- *Give help*
- *Adapt interface*
- *Tailor Information*
- *Recommend Objects*
- ***EXPLANATIONS***

Why?

How?

Why Not?

What if?

Input sources

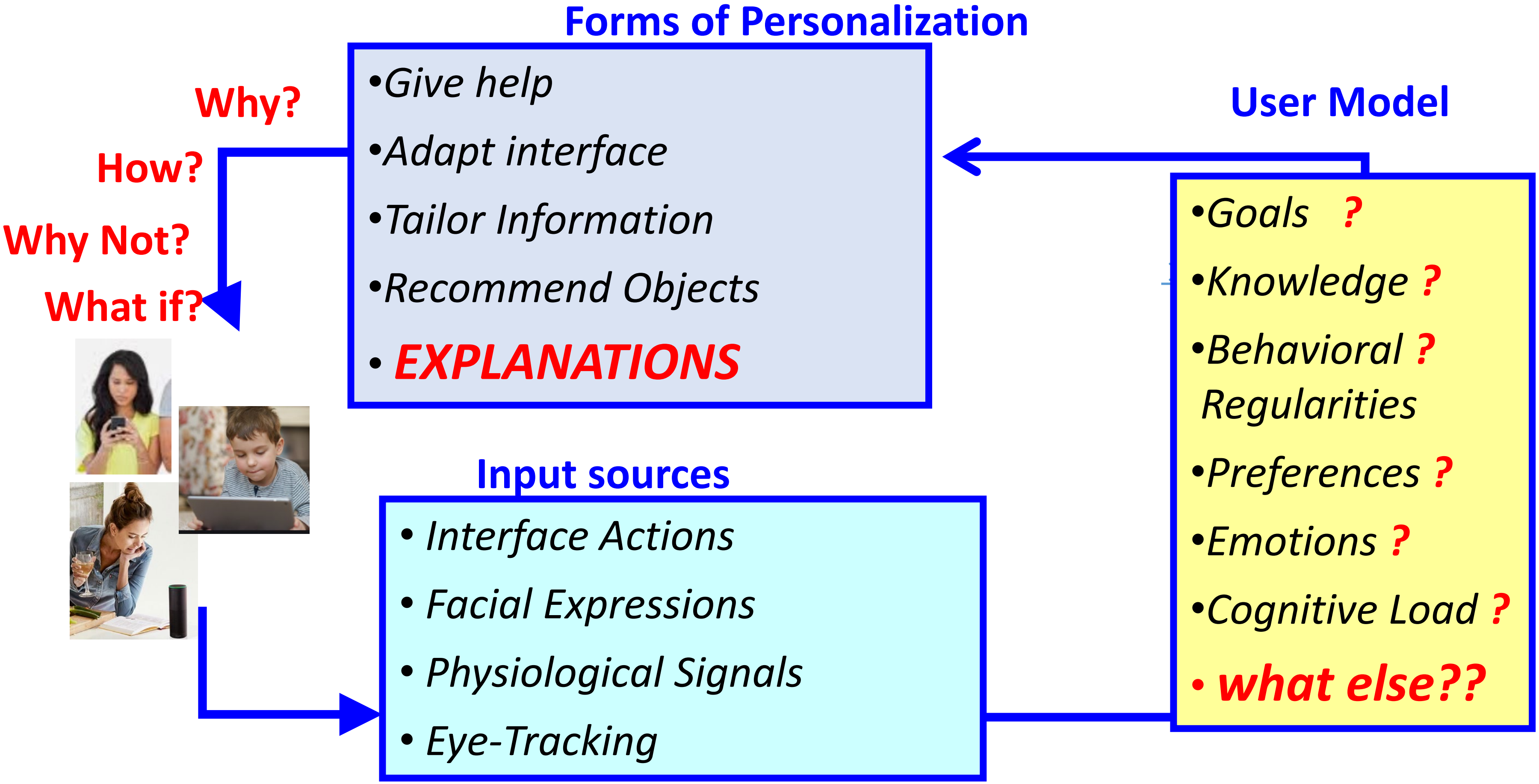
- *Interface Actions*
- *Facial Expressions*
- *Physiological Signals*
- *Eye-Tracking…*

User Model

- *Goals*
- *Knowledge*
- *Behavioral Regularities*
- *Preference*
- *Emotions*
- *Cognitive Load*

Vision: Personalized XAI

- Intelligent systems that understand to whom, when and how to provide explanations of their behaviors.
- Good UI tools to interact with the explanations



Forms of Personalization

- *Give help*
- *Adapt interface*
- *Tailor Information*
- *Recommend Objects*
- ***EXPLANATIONS***

Why?

How?

Why Not?

What if?

User Model

- *Goals ?*
- *Knowledge ?*
- *Behavioral ? Regularities*
- *Preferences ?*
- *Emotions ?*
- *Cognitive Load ?*
- ***what else??***

Input sources

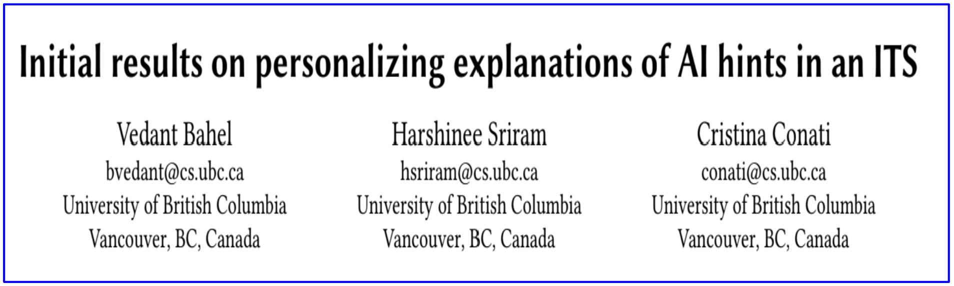
- *Interface Actions*
- *Facial Expressions*
- *Physiological Signals*
- *Eye-Tracking*

Current Work

Investigate role of long-term user traits in Personalized XAI: personality traits and cognitive skills

Some exciting results on explanations for

Hints in an Intelligent Tutoring System (ITS)

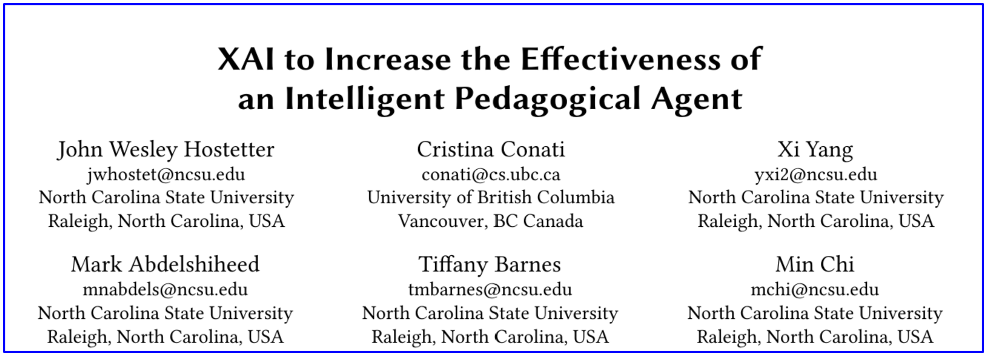


Initial results on personalizing explanations of AI hints in an ITS

Vedant Bahel
bvedant@cs.ubc.ca
University of British Columbia
Vancouver, BC, Canada

Harshinee Sriram
hsriram@cs.ubc.ca
University of British Columbia
Vancouver, BC, Canada

Cristina Conati
conati@cs.ubc.ca
University of British Columbia
Vancouver, BC, Canada



XAI to Increase the Effectiveness of an Intelligent Pedagogical Agent

John Wesley Hostetter
jwhostet@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Cristina Conati
conati@cs.ubc.ca
University of British Columbia
Vancouver, BC Canada

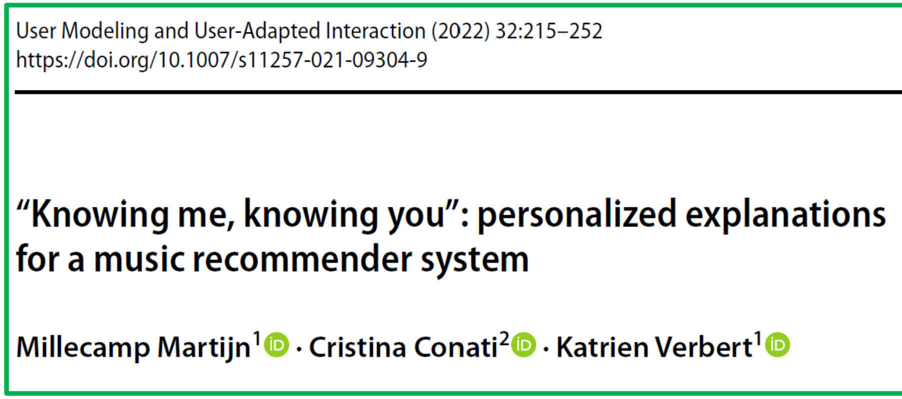
Xi Yang
yxi2@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Mark Abdelshiheed
mnabdels@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Tiffany Barnes
tmbarnes@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Min Chi
mchi@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Music Recommendations



User Modeling and User-Adapted Interaction (2022) 32:215–252
https://doi.org/10.1007/s11257-021-09304-9

"Knowing me, knowing you": personalized explanations for a music recommender system

Millecamp Martijn[1] · Cristina Conati[2] · Katrien Verbert[1]

Current Work

Investigate role of long-term user traits in Personalized XAI: personality traits and cognitive skills

Some exciting results on explanations for

Hints in an Intelligent Tutoring System (ITS)

Initial results on personalizing explanations of AI hints in an ITS

Vedant Bahel
bvedant@cs.ubc.ca
University of British Columbia
Vancouver, BC, Canada

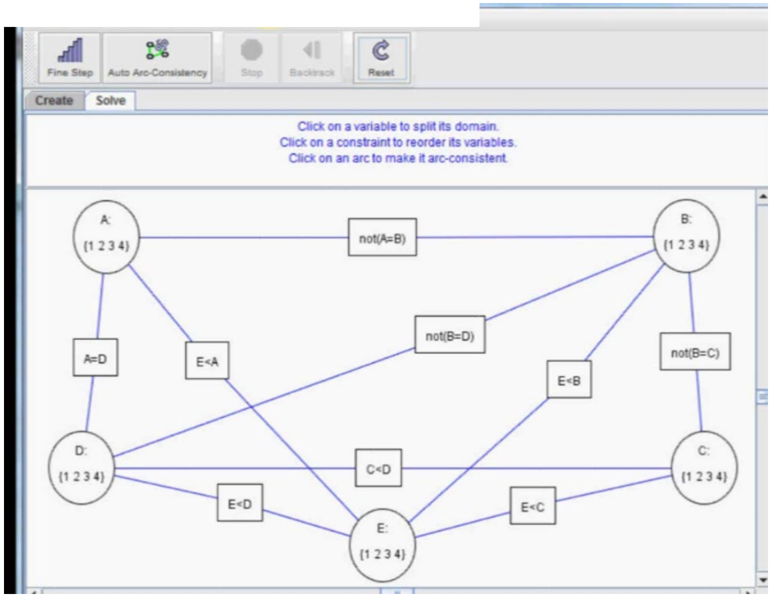
Harshinee Sriram
hsriram@cs.ubc.ca
University of British Columbia
Vancouver, BC, Canada

Cristina Conati
conati@cs.ubc.ca
University of British Columbia
Vancouver, BC, Canada

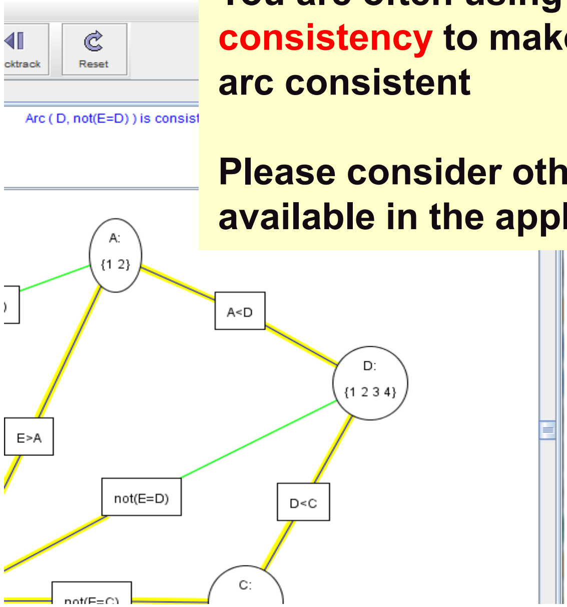
AI-Driven Hints for the ACSP Tutor

[Amershi and Conati 2009; Kardan et al., 2012; 2015; Fratamico et al., 2017; Lallé et al., 2020, 2021]

- The Adaptive CSP Applet (ACSP) helps students learn with an interactive simulation for a constraint satisfaction algorithm
- Detects when students are not using the simulation well for learning
- Generates AI-driven hints to guide a student towards behaviors effective for learning



Click on a variable to split its domain.
Click on a constraint to reorder its variables.
Click on an arc to make it arc-consistent.

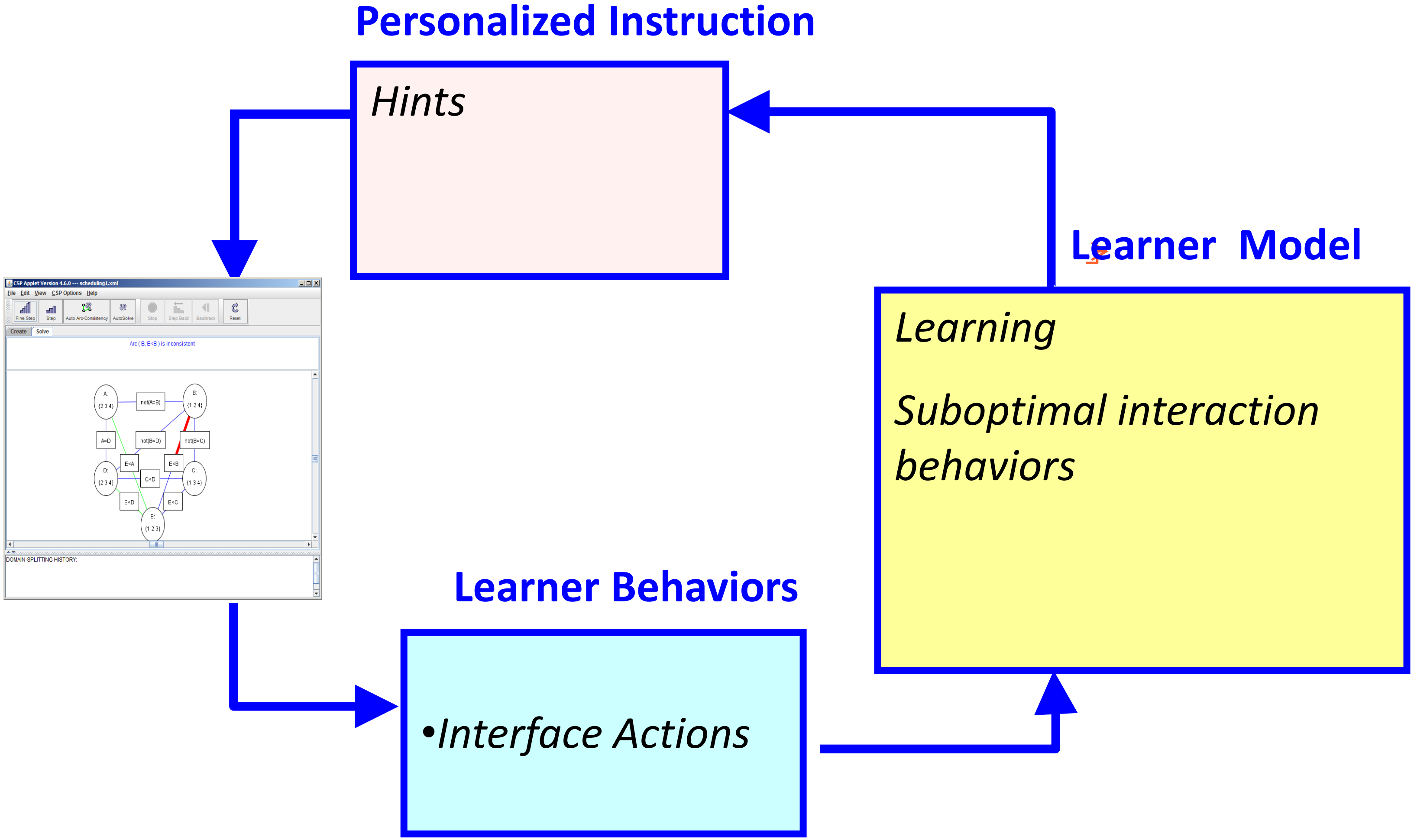


Arc (D, not(E=D)) is consist

You are often using Auto Arc-consistency to make the graph arc consistent

Please consider other option available in the applet

Personalization Loop in ACSP



Personalized Instruction

Hints

Learner Model

Learning

Suboptimal interaction behaviors

Learner Behaviors

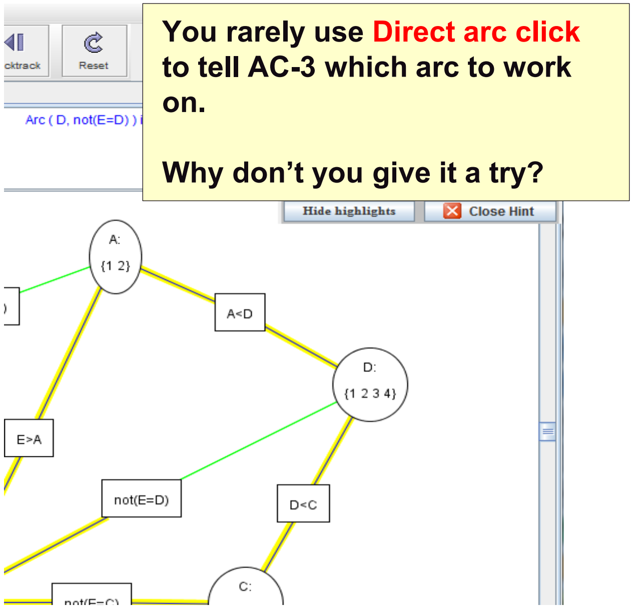
•*Interface Actions*

CSP Applet Version 4.6.0 --- scheduling1.xml

Arc (B, E<B) is inconsistent

DOMAIN-SPLITTING HISTORY:

Explaining ACSP Hints



You rarely use Direct arc click to tell AC-3 which arc to work on.

Why don't you give it a try?

Adaptive CSP (ACSP)

The ACSP hints improve student learning

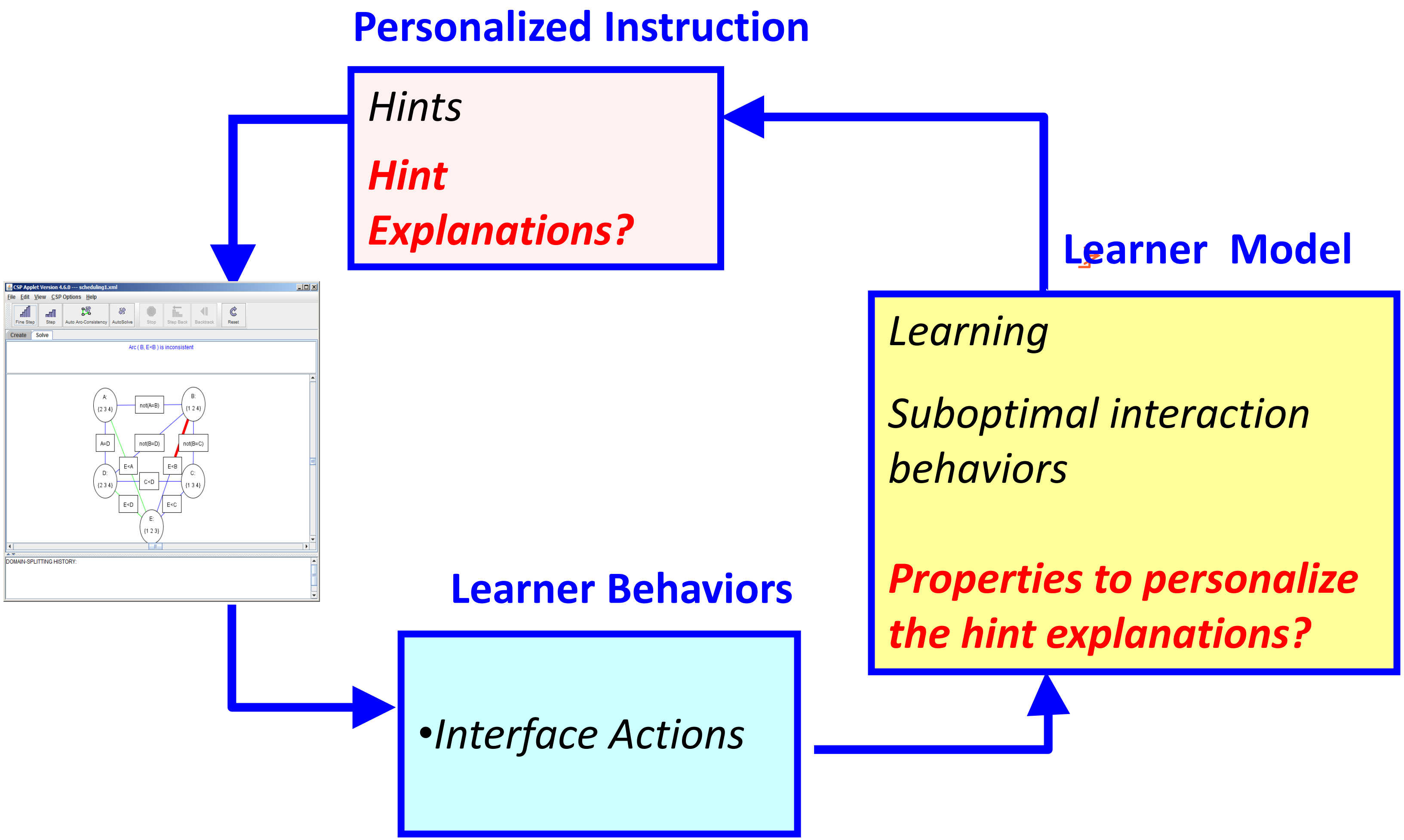
[Kardan and Conati, CHI 2015]

Could these hints be even more effective if the system can explain why and how they were generated?

Does this depend on specific student characteristics?

Conati et al., AI Journal 2021

Personalization Loop in ACSP



Personalized Instruction

Hints

Hint Explanations?

Learner Model

Learning

Suboptimal interaction behaviors

Properties to personalize the hint explanations?

Learner Behaviors

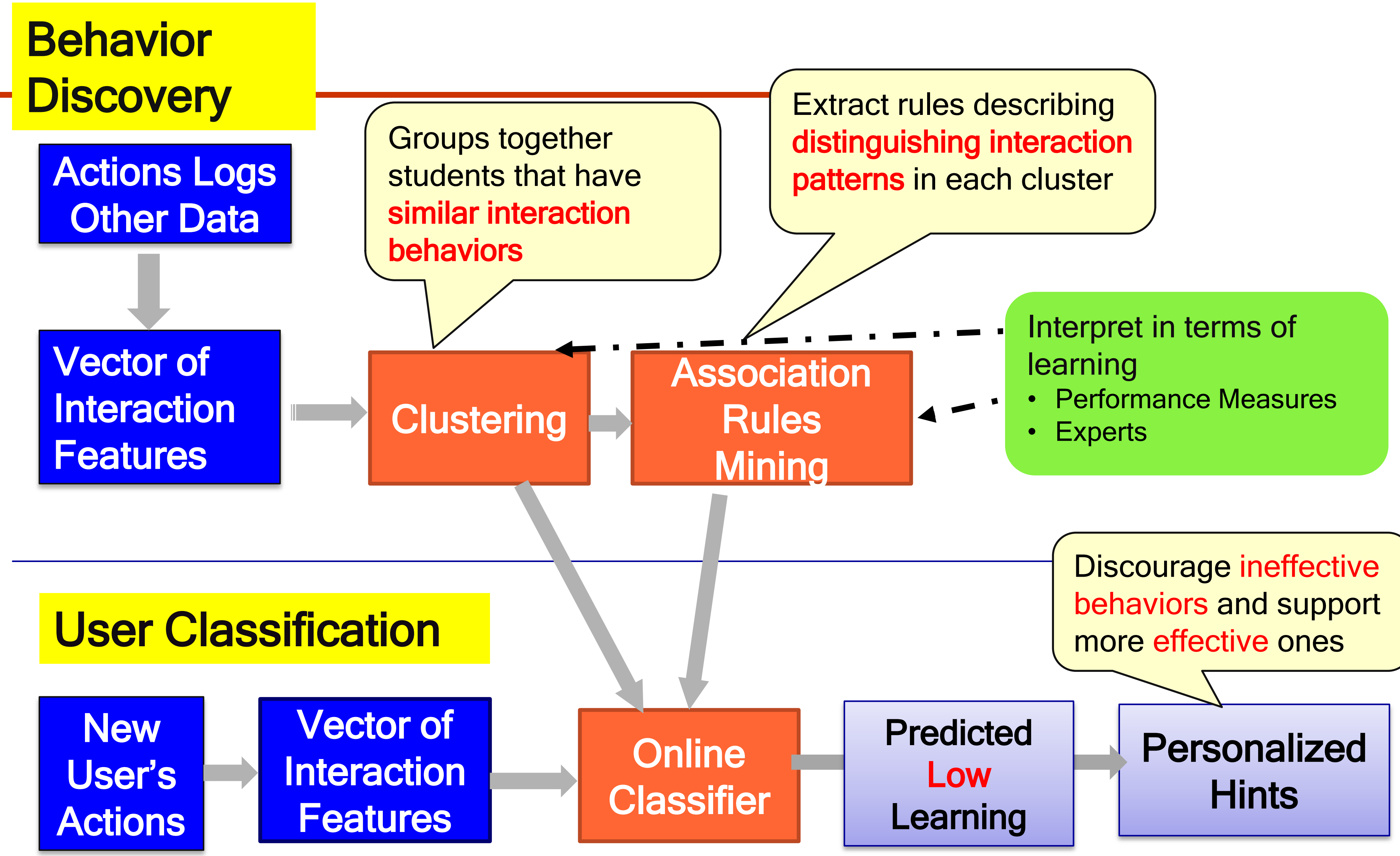
•Interface Actions

CSP Applet Version 4.6.0 --- scheduling1.xml

Arc (B, E<B) is inconsistent

DOMAIN-SPLITTING HISTORY:

Explaining ACSP Hints: Underlying AI



Behavior Discovery

Actions Logs
Other Data

Vector of Interaction Features

Groups together students that have similar interaction behaviors

Clustering

Association Rules Mining

Extract rules describing distinguishing interaction patterns in each cluster

Interpret in terms of learning
- Performance Measures
- Experts

User Classification

New User's Actions

Vector of Interaction Features

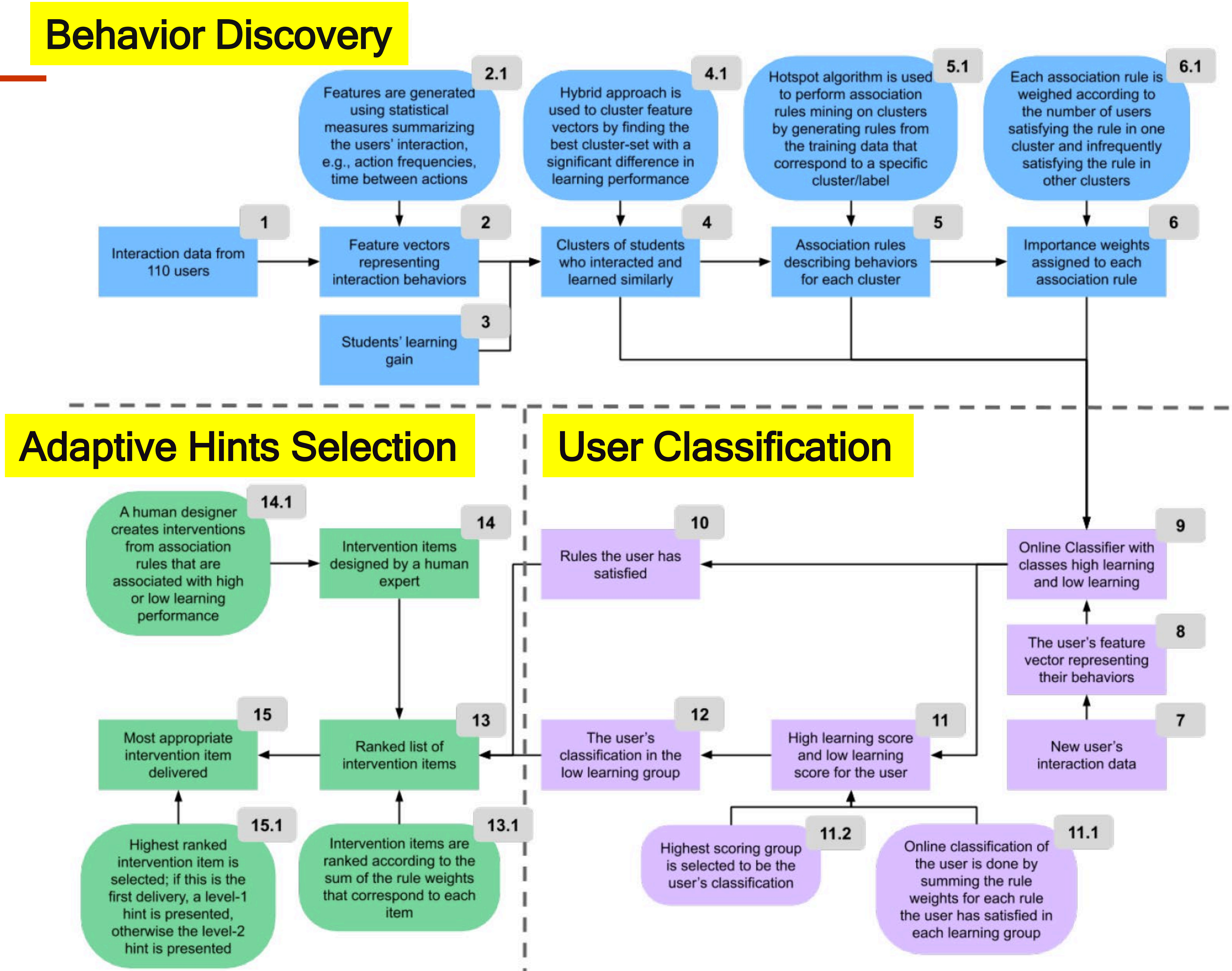
Online Classifier

Predicted Low Learning

Personalized Hints

Discourage ineffective behaviors and support more effective ones

Explaining ACSP Hints: Underlying AI



Behavior Discovery

1 Interaction data from 110 users

2 Feature vectors representing interaction behaviors

2.1 Features are generated using statistical measures summarizing the users' interaction, e.g., action frequencies, time between actions

3 Students' learning gain

4 Clusters of students who interacted and learned similarly

4.1 Hybrid approach is used to cluster feature vectors by finding the best cluster-set with a significant difference in learning performance

5 Association rules describing behaviors for each cluster

5.1 Hotspot algorithm is used to perform association rules mining on clusters by generating rules from the training data that correspond to a specific cluster/label

6 Importance weights assigned to each association rule

6.1 Each association rule is weighed according to the number of users satisfying the rule in one cluster and infrequently satisfying the rule in other clusters

Adaptive Hints Selection

14 Intervention items designed by a human expert

14.1 A human designer creates interventions from association rules that are associated with high or low learning performance

13 Ranked list of intervention items

13.1 Intervention items are ranked according to the sum of the rule weights that correspond to each item

15 Most appropriate intervention item delivered

15.1 Highest ranked intervention item is selected; if this is the first delivery, a level-1 hint is presented, otherwise the level-2 hint is presented

User Classification

7 New user's interaction data

8 The user's feature vector representing their behaviors

9 Online Classifier with classes high learning and low learning

10 Rules the user has satisfied

11 High learning score and low learning score for the user

11.1 Online classification of the user is done by summing the rule weights for each rule the user has satisfied in each learning group

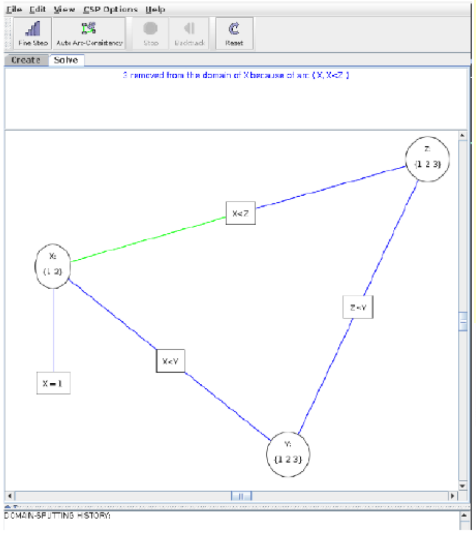
11.2 Highest scoring group is selected to be the user's classification

12 The user's classification in the low learning group

Design Criteria for Explanations

[Kulesza et al., 2015].

- **Iterative**: accessible at different levels, at user's will.
- **Truthful**: conveying an accurate description of relevant mechanisms.
- **Not overwhelming**



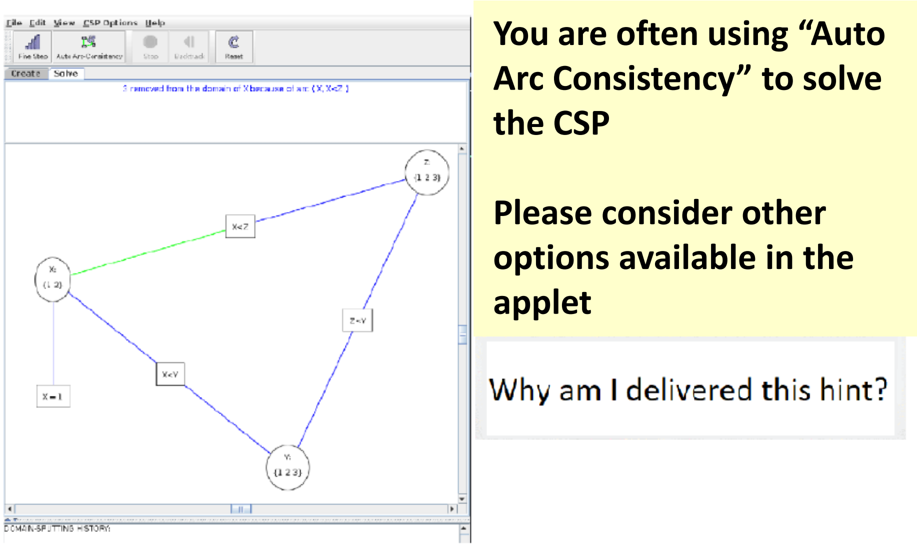
3 removed from the domain of X because of arc {X, X<Z}

You are often using "Auto Arc Consistency" to solve the CSP

Please consider other options available in the applet

Why am I delivered this hint?

- **Iterative**: accessible at different levels, at user’s will.
- **Truthful**: conveying an accurate description of relevant mechanisms.
- **Not overwhelming**

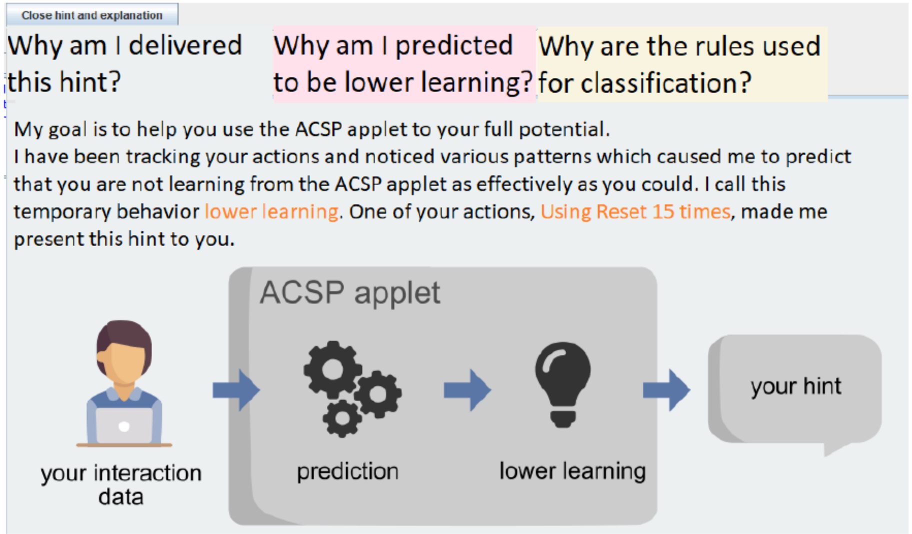


3 removed from the domain of X because of arc (X, X<Z)

You are often using “Auto Arc Consistency” to solve the CSP

Please consider other options available in the applet

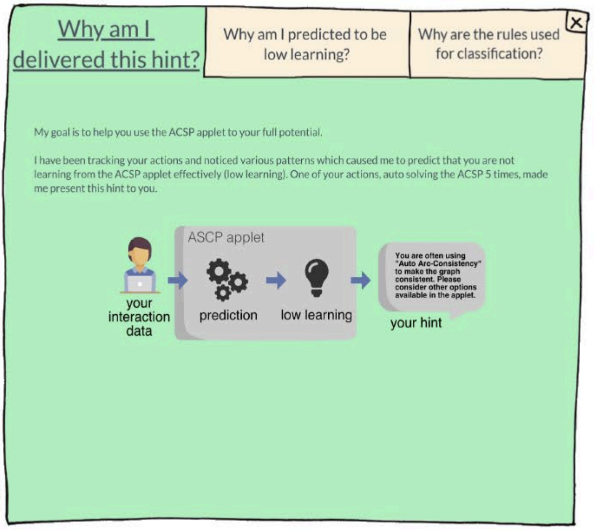
Why am I delivered this hint?


Close hint and explanation
Why am I delivered this hint?
Why am I predicted to be lower learning?
Why are the rules used for classification?

My goal is to help you use the ACSP applet to your full potential.
I have been tracking your actions and noticed various patterns which caused me to predict that you are not learning from the ACSP applet as effectively as you could. I call this temporary behavior lower learning. One of your actions, Using Reset 15 times, made me present this hint to you.

ACSP applet
your interaction data
prediction
lower learning
your hint

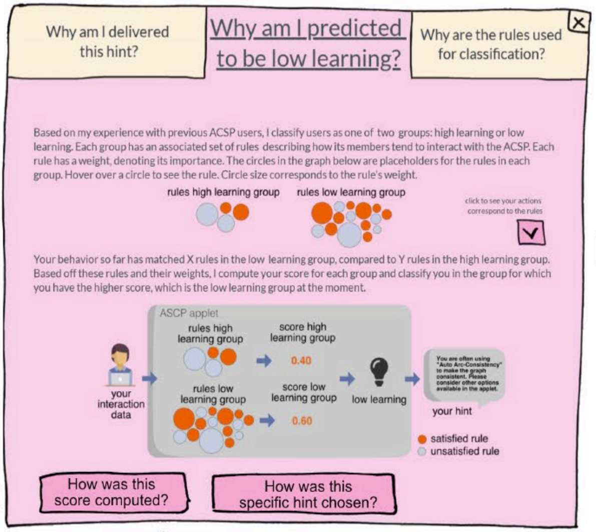
Explanations in ACSP Tutor


Why am I delivered this hint?
Why am I predicted to be low learning?
Why are the rules used for classification?
ASCP applet
your interaction data
prediction
low learning
You are often using "Auto Arc-Consistency" to make the graph consistent. Please consider other options available in the applet.
your hint

(A)

My goal is to help you use the ACSP applet to your full potential.

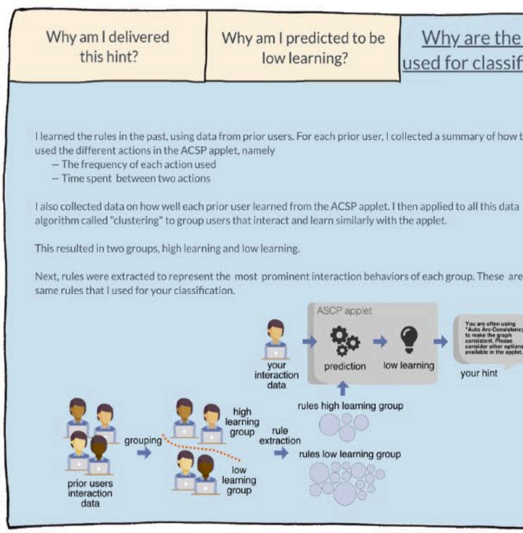
I have been tracking your actions and noticed various patterns which caused me to predict that you are not learning from the ACSP applet effectively (low learning). One of your actions, auto solving the ACSP 5 times, made me present this hint to you.


Why am I delivered this hint?
Why am I predicted to be low learning?
Why are the rules used for classification?
rules high learning group
rules low learning group
click to see your actions correspond to the rules
ASCP applet
rules high learning group
score high learning group
0.40
rules low learning group
score low learning group
0.60
your interaction data
low learning
You are often using "Auto Arc-Consistency" to make the graph consistent. Please consider other options available in the applet.
your hint
● satisfied rule
○ unsatisfied rule
How was this score computed?
How was this specific hint chosen?

(B)

Based on my experience with previous ACSP users, I classify users as one of two groups: high learning or low learning. Each group has an associated set of rules describing how its members tend to interact with the ACSP. Each rule has a weight, denoting its importance. The circles in the graph below are placeholders for the rules in each group. Hover over a circle to see the rule. Circle size corresponds to the rule's weight.

Your behavior so far has matched X rules in the low learning group, compared to Y rules in the high learning group. Based off these rules and their weights, I compute your score for each group and classify you in the group for which you have the higher score, which is the low learning group at the moment.


Why am I delivered this hint?
Why am I predicted to be low learning?
Why are the used for classif
ASCP applet
your interaction data
prediction
low learning
You are often using "Auto Arc-Consistency" to make the graph consistent. Please consider other options available in the applet.
your hint
rules high learning group
prior users interaction data
grouping
high learning group
low learning group
rule extraction
rules low learning group

(C)

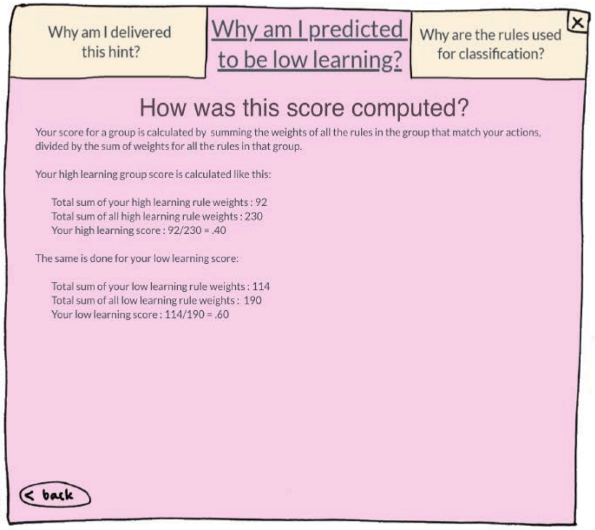
I learned the rules in the past, using data from prior users. For each prior user, I collected a summary of how t used the different actions in the ACSP applet, namely

- The frequency of each action used
- Time spent between two actions

I also collected data on how well each prior user learned from the ACSP applet. I then applied to all this data algorithm called "clustering" to group users that interact and learn similarly with the applet.

This resulted in two groups, high learning and low learning.

Next, rules were extracted to represent the most prominent interaction behaviors of each group. These are same rules that I used for your classification.


Why am I delivered this hint?
Why am I predicted to be low learning?
Why are the rules used for classification?
< back

(D)

How was this score computed?

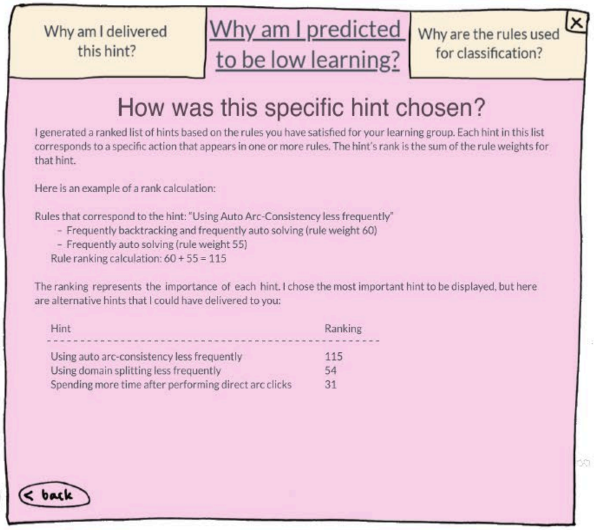
Your score for a group is calculated by summing the weights of all the rules in the group that match your actions, divided by the sum of weights for all the rules in that group.

Your high learning group score is calculated like this:

Total sum of your high learning rule weights : 92
Total sum of all high learning rule weights : 230
Your high learning score : 92/230 = .40

The same is done for your low learning score:

Total sum of your low learning rule weights : 114
Total sum of all low learning rule weights : 190
Your low learning score : 114/190 = .60


Why am I delivered this hint?
Why am I predicted to be low learning?
Why are the rules used for classification?
< back

(E)

How was this specific hint chosen?

I generated a ranked list of hints based on the rules you have satisfied for your learning group. Each hint in this list corresponds to a specific action that appears in one or more rules. The hint's rank is the sum of the rule weights for that hint.

Here is an example of a rank calculation:

Rules that correspond to the hint: "Using Auto Arc-Consistency less frequently"

- Frequently backtracking and frequently auto solving (rule weight 60)
- Frequently auto solving (rule weight 55)

Rule ranking calculation: 60 + 55 = 115

The ranking represents the importance of each hint. I chose the most important hint to be displayed, but here are alternative hints that I could have delivered to you:

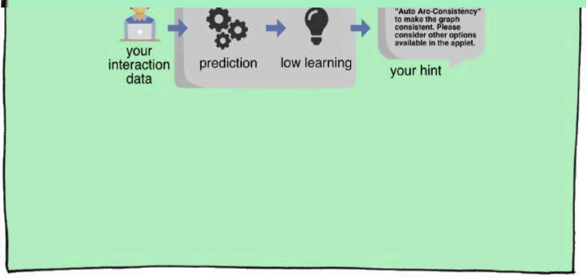
Hint	Ranking
Using auto arc-consistency less frequently	115
Using domain splitting less frequently	54
Spending more time after performing direct arc clicks	31

Explanations in ACSP Tutor

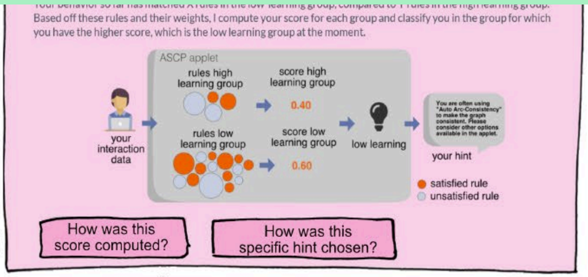
Why am I delivered this hint?

Why am I predicted to be low learning?

Why are the rules used for classification?


your interaction data
prediction
low learning
"Auto Arc-Consistency" to make the graph consistent. Please consider other options available in the applet.
your hint

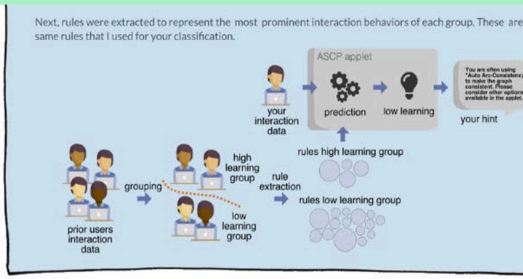
(A)



Your behavior so far has matched X rules in the low learning group, compared to Y rules in the high learning group. Based off these rules and their weights, I compute your score for each group and classify you in the group for which you have the higher score, which is the low learning group at the moment.

ASCP applet
your interaction data
rules high learning group
score high learning group
0.40
rules low learning group
score low learning group
0.60
low learning
You are often using "Auto Arc-Consistency" to make the graph consistent. Please consider other options available in the applet.
your hint
satisfied rule
unsatisfied rule
How was this score computed?
How was this specific hint chosen?

(B)



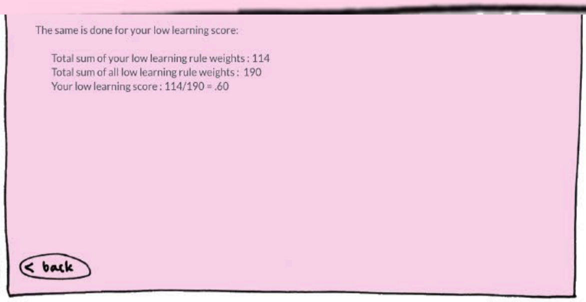
Next, rules were extracted to represent the most prominent interaction behaviors of each group. These are same rules that I used for your classification.

ASCP applet
your interaction data
prediction
low learning
You are often using "Auto Arc-Consistency" to make the graph consistent. Please consider other options available in the applet.
your hint
prior users interaction data
grouping
high learning group
low learning group
rule extraction
rules high learning group
rules low learning group

(C)

How was this score computed?

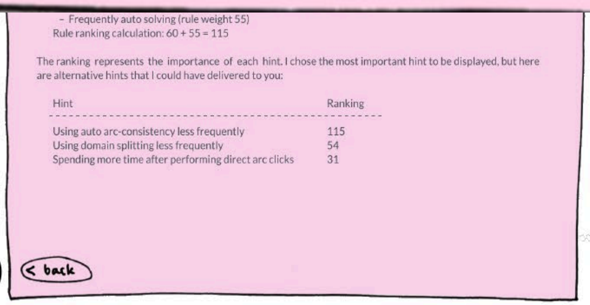
How was this specific hint chosen?



The same is done for your low learning score:

Total sum of your low learning rule weights : 114
Total sum of all low learning rule weights : 190
Your low learning score : 114/190 = .60

(D)



- Frequently auto solving (rule weight 55)

Rule ranking calculation: 60 + 55 = 115

The ranking represents the importance of each hint. I chose the most important hint to be displayed, but here are alternative hints that I could have delivered to you:

Hint	Ranking
Using auto arc-consistency less frequently	115
Using domain splitting less frequently	54
Spending more time after performing direct arc clicks	31

(E)

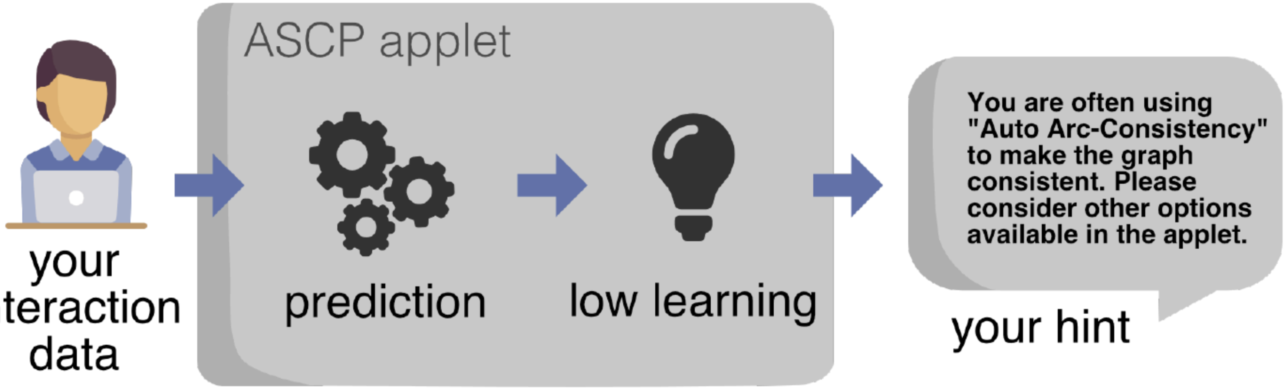
Why am I delivered this hint?

Why am I predicted to be low learning?

Why are the rules used for classification?

My goal is to help you use the ACSP applet to your full potential.

I have been tracking your actions [why, 7] and noticed various patterns [why, 10] which caused me to predict that you are not learning from the ACSP applet effectively (low learning) [why, 12]. One of your actions, auto solving the ACSP 5 times, made me present this hint to you.

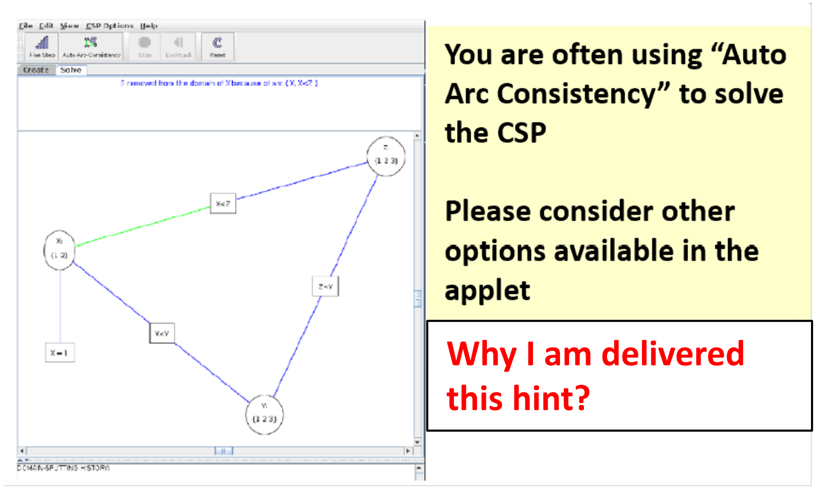


You are often using "Auto Arc-Consistency" to make the graph consistent. Please consider other options available in the applet.

User Study

Compared the effectiveness of the ACSP hints

With Explanations



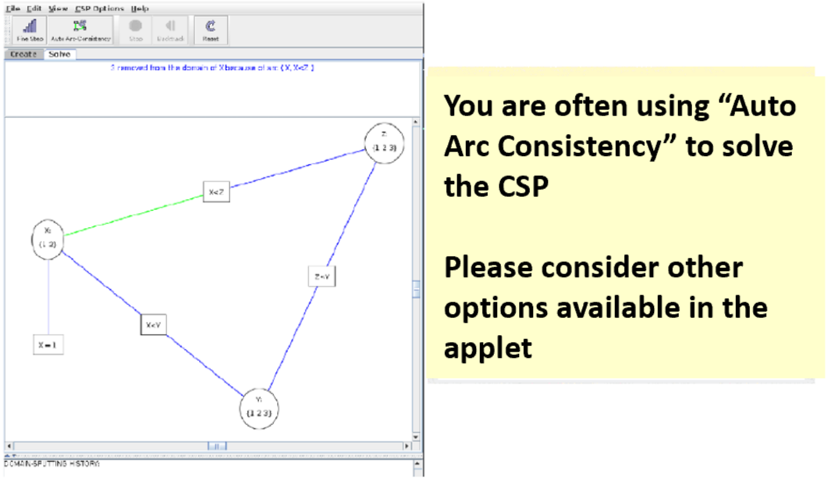
3 removed from the domain of X because of arc {X, X<Z }

You are often using "Auto Arc Consistency" to solve the CSP

Please consider other options available in the applet

Why I am delivered this hint?

Without Explanations



3 removed from the domain of X because of arc {X, X<Z }

You are often using "Auto Arc Consistency" to solve the CSP

Please consider other options available in the applet

Checked for the possible impact of various user traits

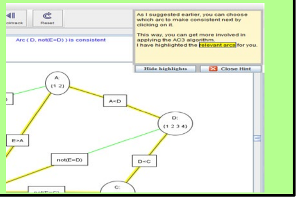
User Study

Test users for possibly relevant **user characteristics**

- 5 Factor Personality Traits
- Need for Cognition (N4C)
- Reading Proficiency
- Perceptual speed, Visual working memory

→

Two groups worked with the ACSP with and without explanation



Arc (D, not(E=D)) is consistent

As I suggested earlier, you can choose which arc to make consistent next by clicking on it.

This way, you can get more involved in applying the AC3 algorithm.
I have highlighted the relevant arcs for you.

- Explanation: N = 24;
- Control: N=16

→

Questionnaires on perception of hints and explanations

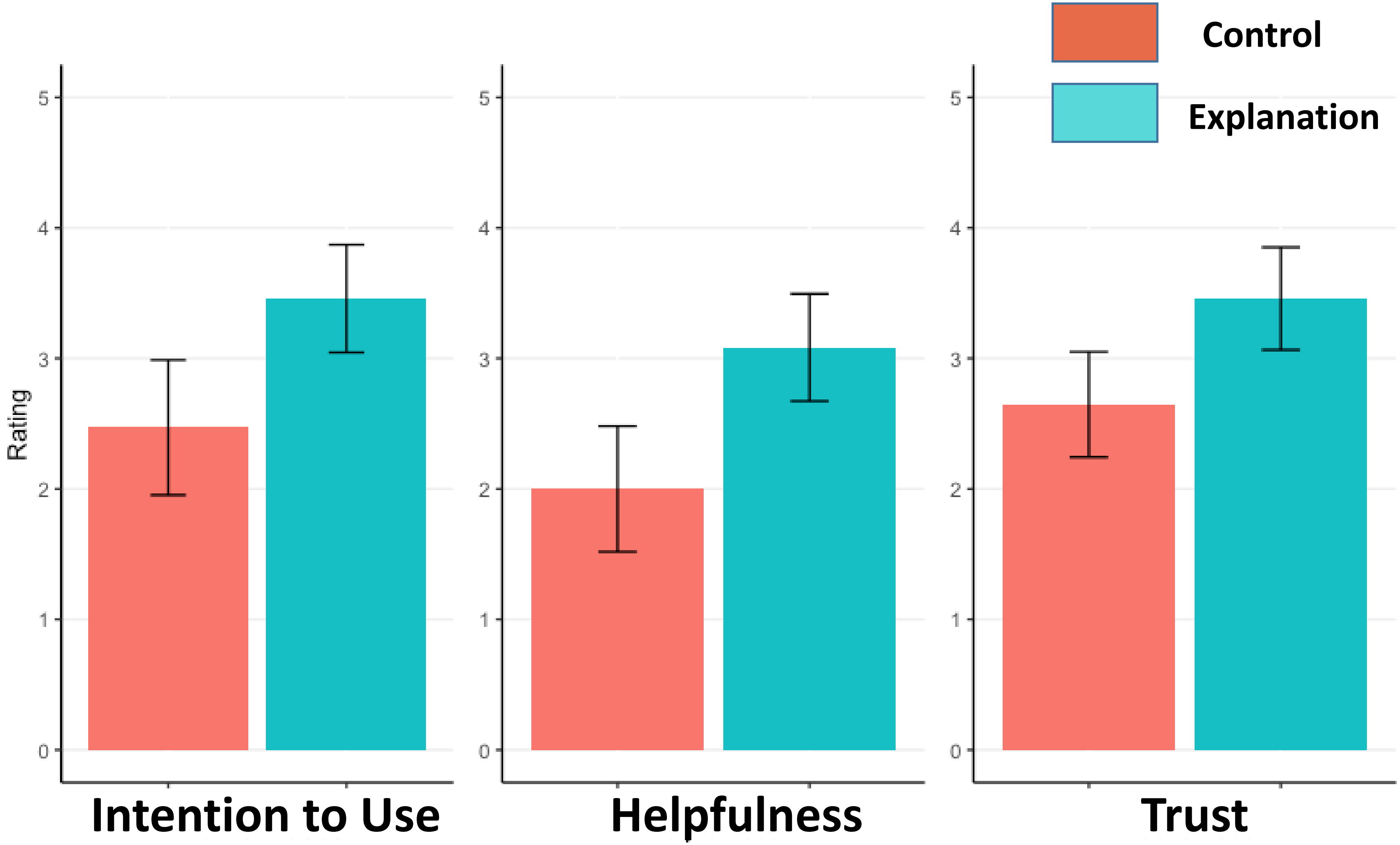
Tracked participants gaze with a Tobii T120

Effect of explanations

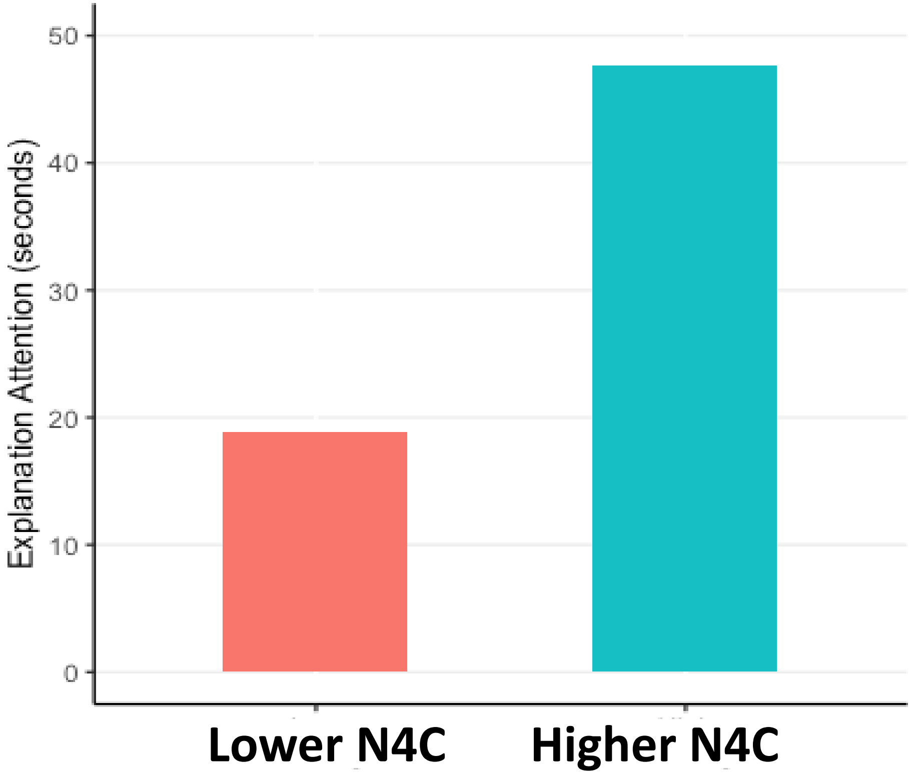
- Perception of AI-driven hints
- Learning (learning gains from pre to posttest)

Explanations and Hints Perception

- Significant effects of explanations on Intention to use, Helpfulness, Trustworthiness of Hints



Impact of Individual Differences



❑ Users with higher Need for Cognition (N4C) show higher attention to explanations than lower N4C users

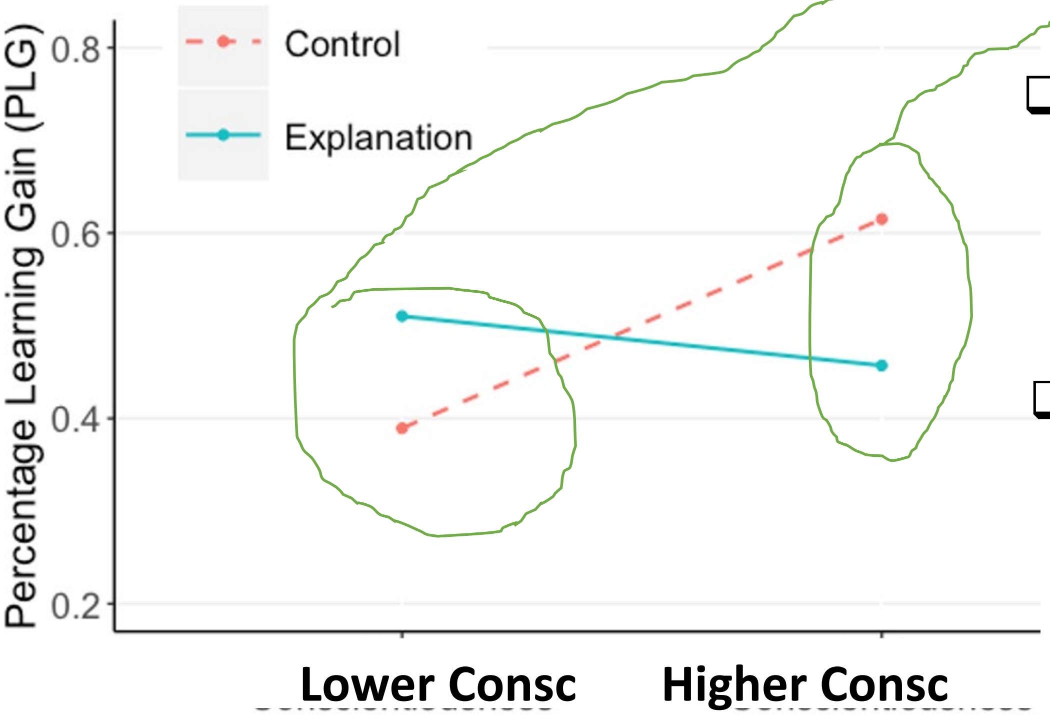
❑ Consistent with higher N4C relating to enjoying effortful cognitive activities

❑ Possible personalization:

- encourage interaction with explanations for low N4C users

Impact of Individual Differences

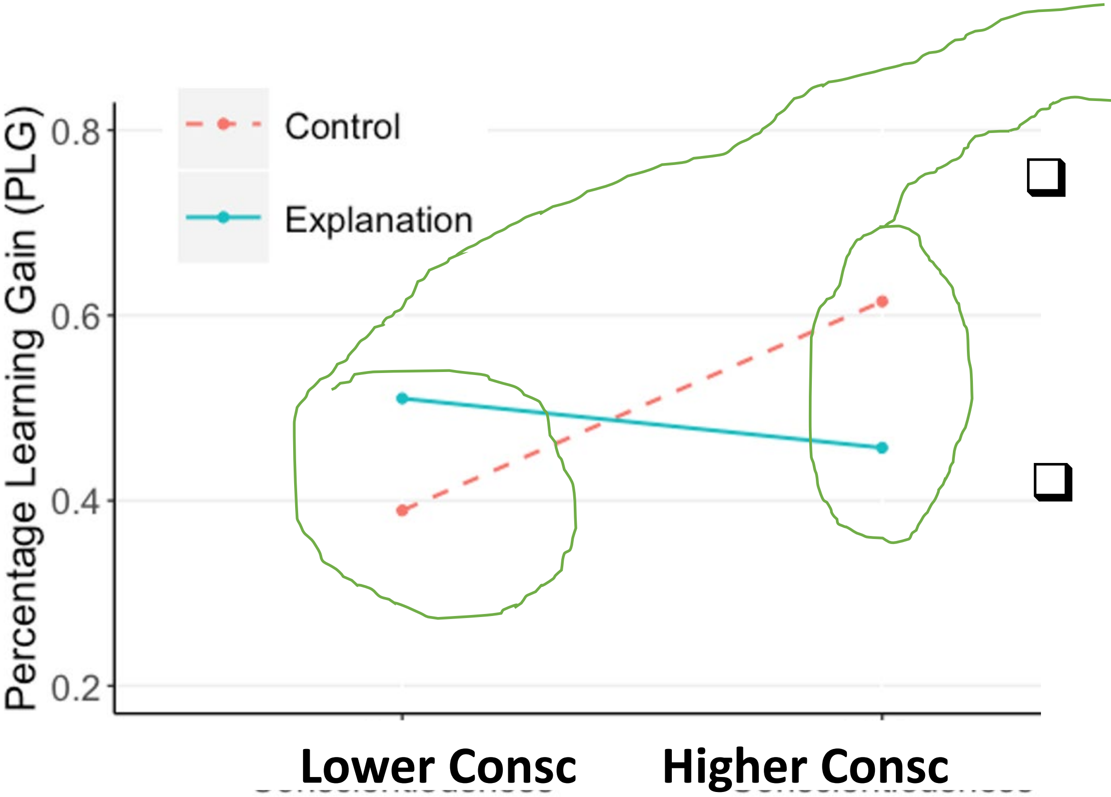
- ❑ Interaction effect of explanation and Conscientiousness (Consc) on learning gains



- ❑ Users with lower Consc learn more with explanations
  - Less if they have higher Consc
- ❑ Explanations provide motivation for lower Consc users to follow the hints?
- ❑ Possible personalization:
  - Encourage interaction with explanations for low Consc users

Impact of Individual Differences

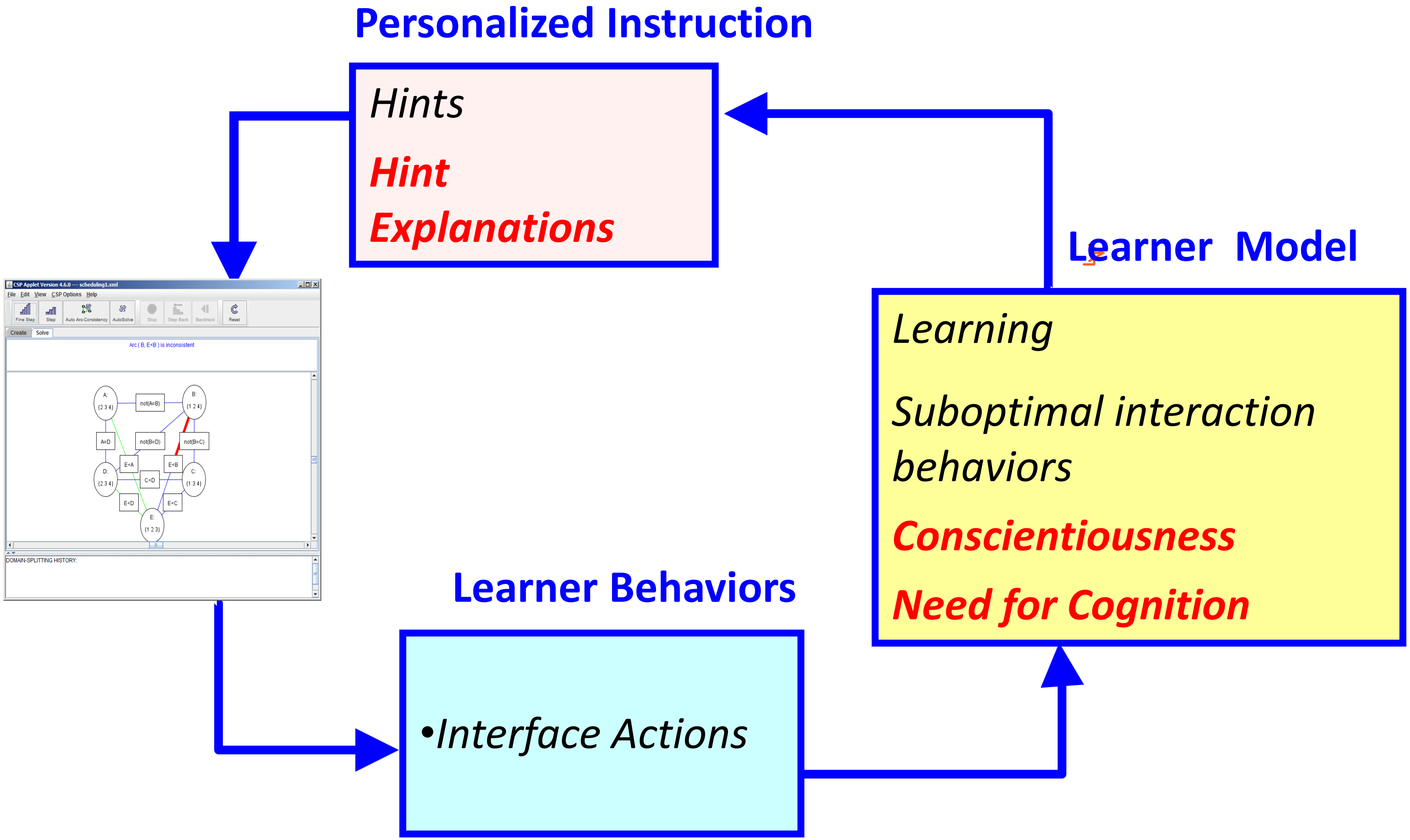
- ❑ Interaction effect of explanation and Conscientiousness (Consc) on learning gains



- ❑ Users with lower Consc learn more with explanations
  - Less if they have higher Consc
- ❑ Explanations provide motivation for lower Consc users to follow the hints?

- ❑ Possible personalization:
  - Encourage interaction with explanations for low Consc users

Personalization Loop in ACSP



Personalized Instruction

Hints

Hint

Explanations

Learner Model

Learning

Suboptimal interaction behaviors

Conscientiousness

Need for Cognition

Learner Behaviors

- *Interface Actions*

CSP Applet Version 4.6.0 --- scheduling1.xml

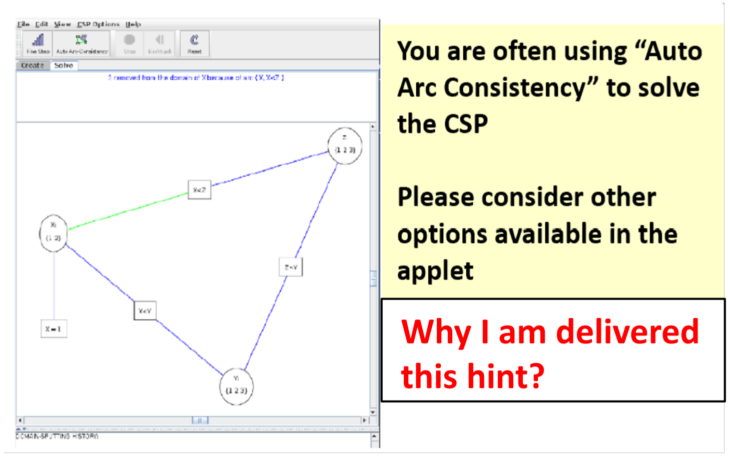
Arc (B, E<B) is inconsistent

DOMAIN-SPLITTING HISTORY:

Testing Personalized Explanations

Encourage interaction with explanations for Low Consc, Low N4C students (Bahel et al. UMAP 2024, arXiv:2403.04035)

Instead of explanations being on-demand



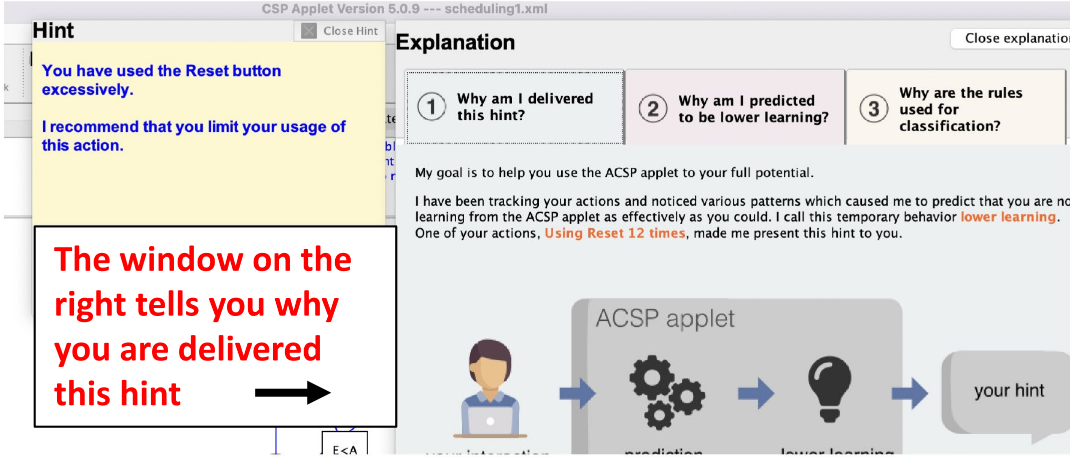
3 removed from the domain of X because of arc ⟨X, X<Z⟩

You are often using “Auto Arc Consistency” to solve the CSP

Please consider other options available in the applet

Why I am delivered this hint?

The explanation interface is opened upfront when a hint is given



You have used the Reset button excessively.

I recommend that you limit your usage of this action.

The window on the right tells you why you are delivered this hint →

① Why am I delivered this hint?

② Why am I predicted to be lower learning?

③ Why are the rules used for classification?

My goal is to help you use the ACSP applet to your full potential.

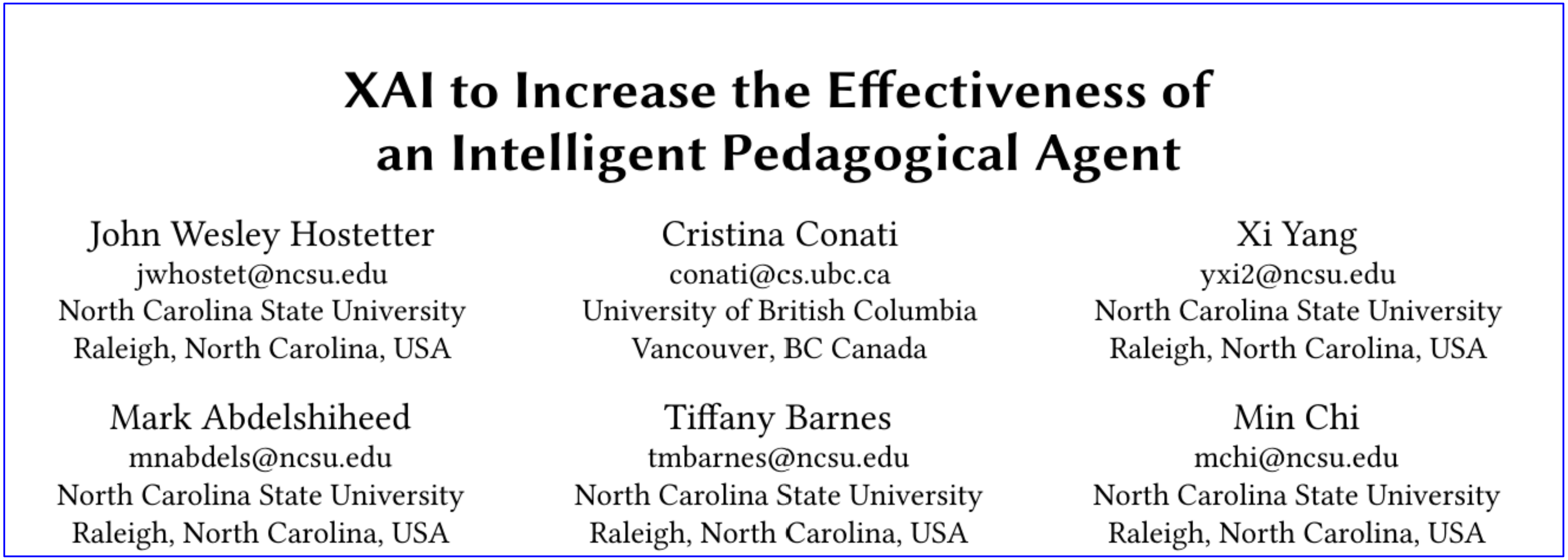
I have been tracking your actions and noticed various patterns which caused me to predict that you are no learning from the ACSP applet as effectively as you could. I call this temporary behavior lower learning. One of your actions, Using Reset 12 times, made me present this hint to you.

Students are prompted to stay if they try to leave the explanation interface too early

- Significant increased attention to explanations
- Significant positive effect on learning

Similar Results with a Different ITS

- Evidence that personalizing explanations of the ITS suggestions to student learning attitude (performance vs learning oriented) improves learning



XAI to Increase the Effectiveness of an Intelligent Pedagogical Agent

John Wesley Hostetter
jwhostet@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Cristina Conati
conati@cs.ubc.ca
University of British Columbia
Vancouver, BC Canada

Xi Yang
yxi2@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

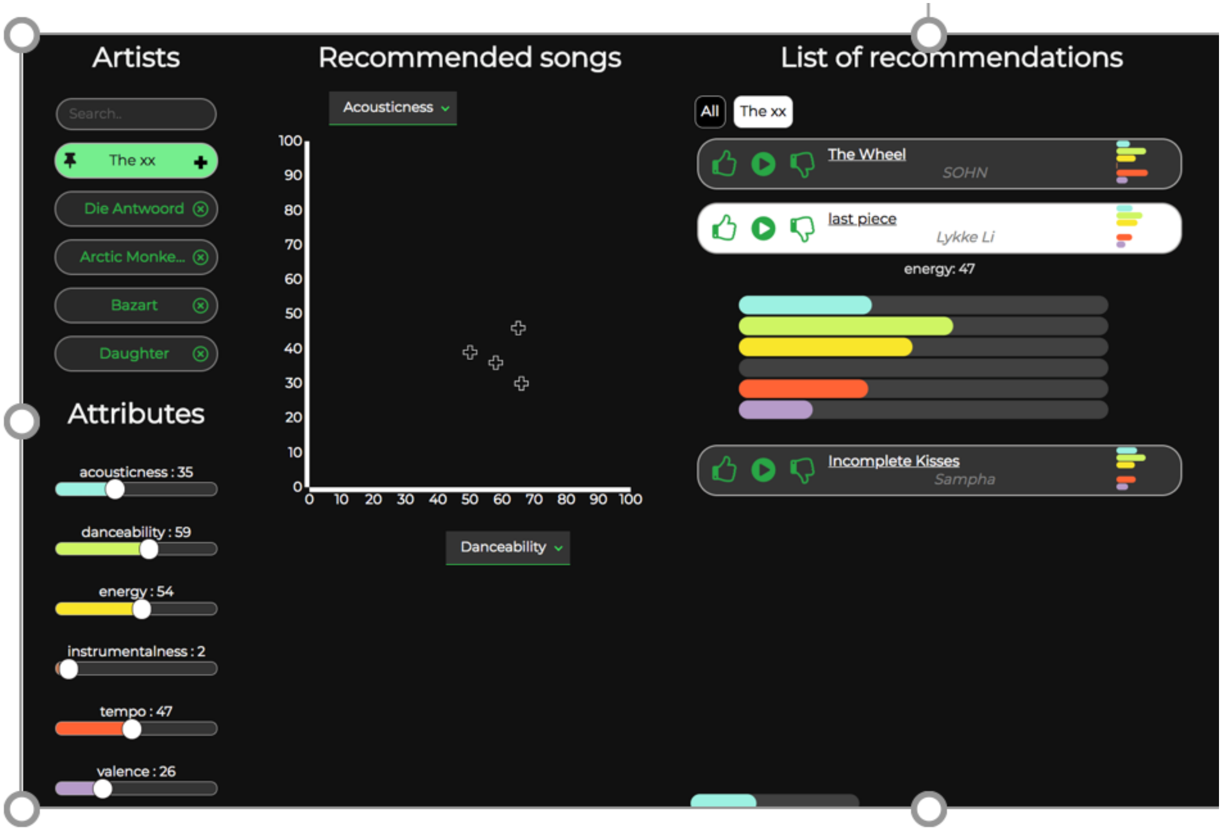
Mark Abdelshiheed
mnabdels@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Tiffany Barnes
tmbarnes@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Min Chi
mchi@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Intelligent Virtual Agents 2023

Similar Results for a Music Recommender System



acousticness : 35

danceability : 59

energy : 54

instrumentalness : 2

tempo : 47

valence : 26

energy: 47

- Evidence that some personality traits and cognitive skills impact explanation effectiveness [Millecamp et al., IUI 2019, UMAP 2020
  - Insights on how to personalize explanations to these user traits]
- Evidence that the personalization works [Millecamp et al., UMUAI J. 2022]

Conclusions and Future work

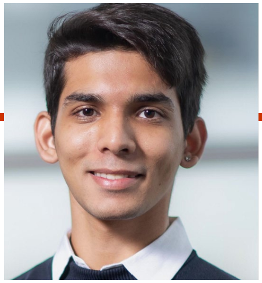
Initial evidence for the need to personalize AI explanations to long-term user traits

What’s next?

- ❑ Real-time classification of the relevant user characteristics
- ❑ Run similar studies with different applications and stakeholders
- ❑ Look at short-term user states (e.g., emotions, cognitive load)
- ❑ Investigate interplay between user characteristics and explanation properties (e.g. type, level of detail, delivery method)

Thanks To

Oswald Barral

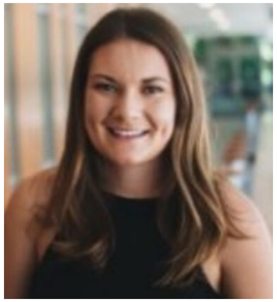


Vedant Bahel

Lauren Fratamico

Sebastien Lalle

Samad Kardan



Vanessa Putnam

Universität
Augsburg
University

Lea Riegel

Harshinee Sriram

Dereck Toker

Nilay Yalcin

And to all of you for your kind attention !