Temporally Coherent Clustering of Student Data

Severin Klingler, Tanja Käser, Barbara Solenthaler and Markus Gross
ETH Zurich, Switzerland

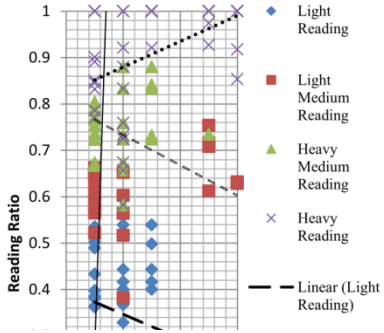
Clustering in EDM

Clustering sequential data to detect behavior patterns

Clustering in EDM

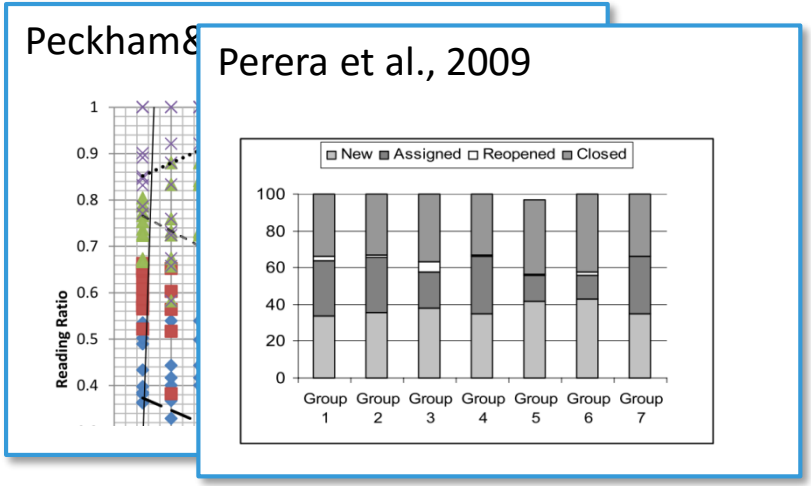
Clustering sequential data to detect behavior patterns

Peckham& McCall, 2012



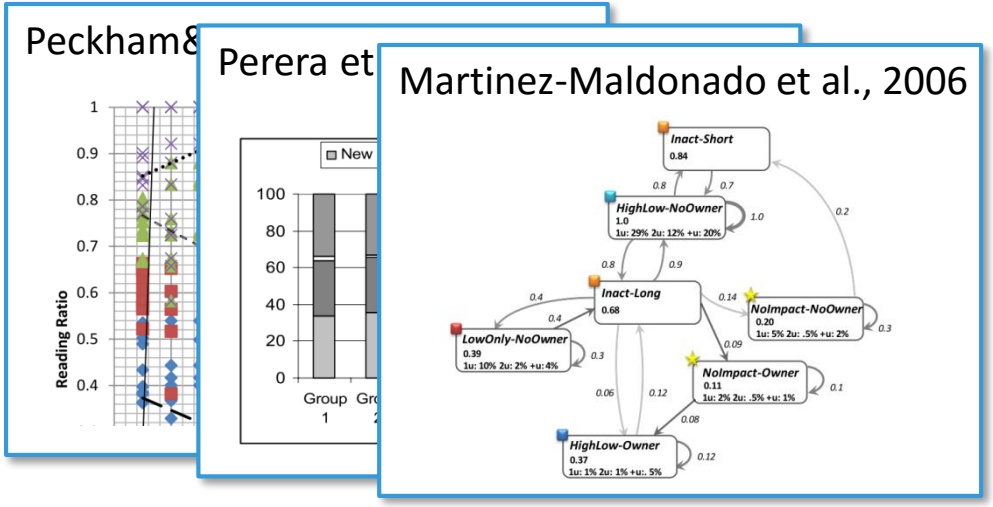
Clustering in EDM

Clustering sequential data to detect behavior patterns



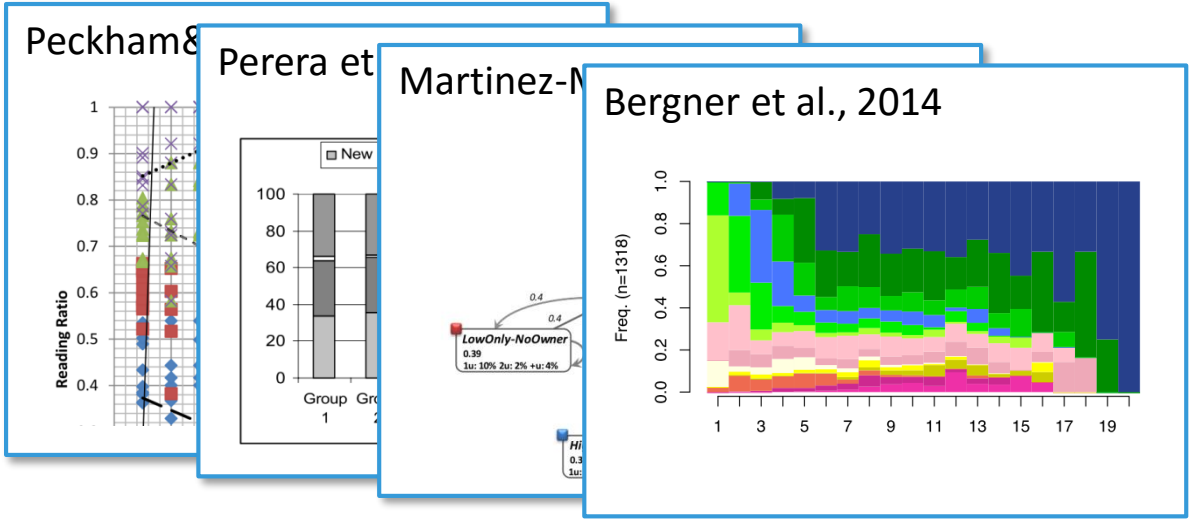
Clustering in EDM

Clustering sequential data to detect behavior patterns



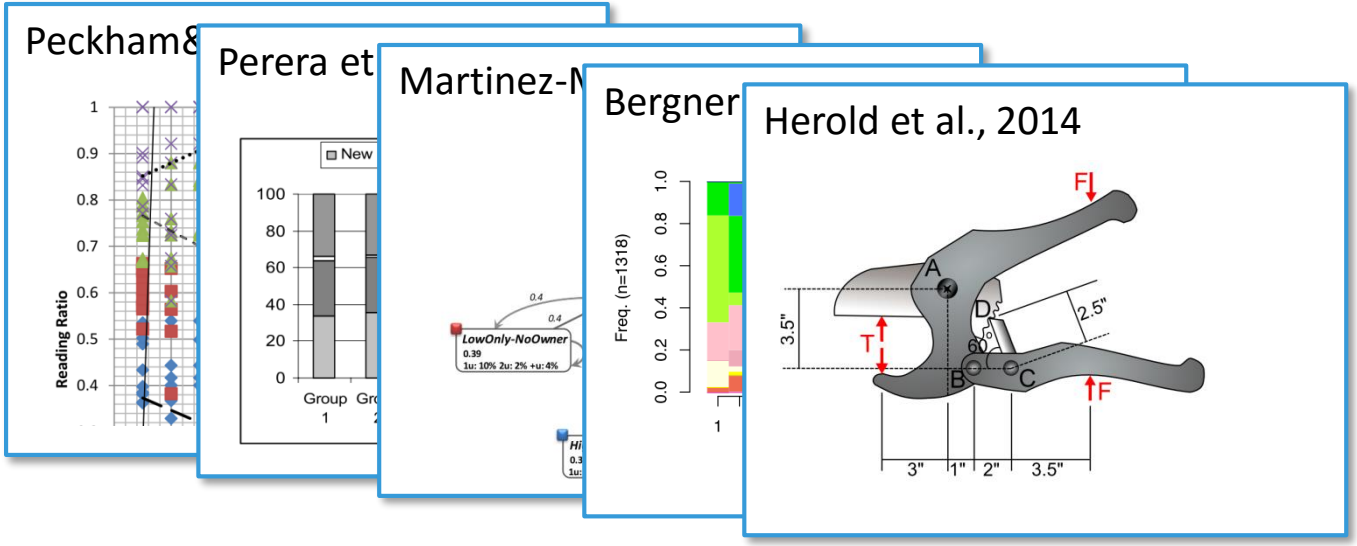
Clustering in EDM

Clustering sequential data to detect behavior patterns

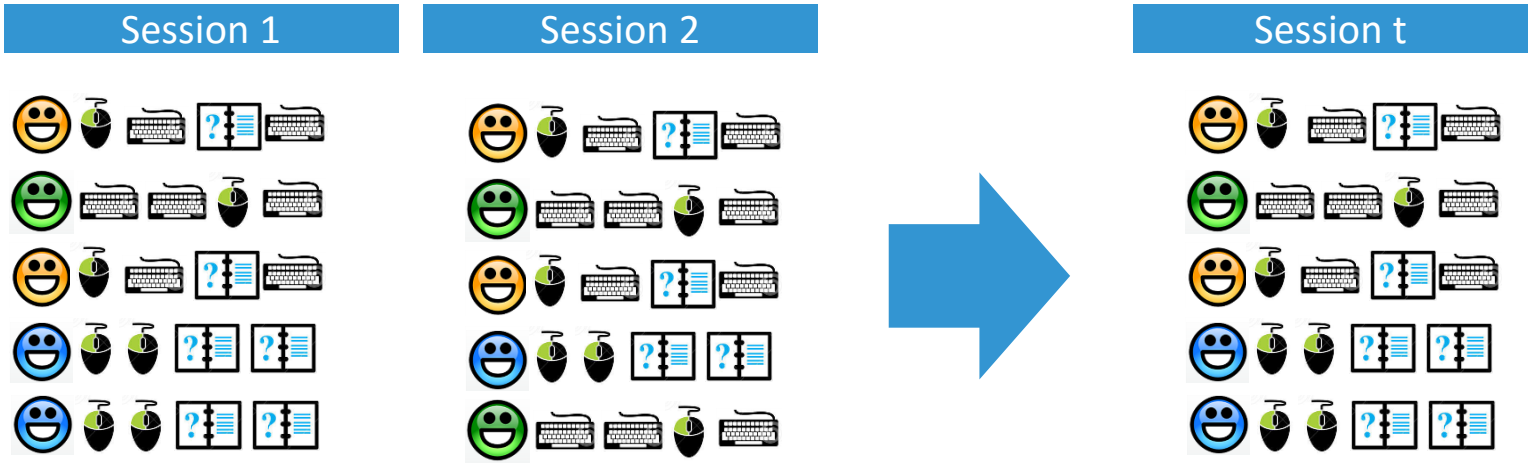


Clustering in EDM

Clustering sequential data to detect behavior patterns



Evolution of behavior patterns

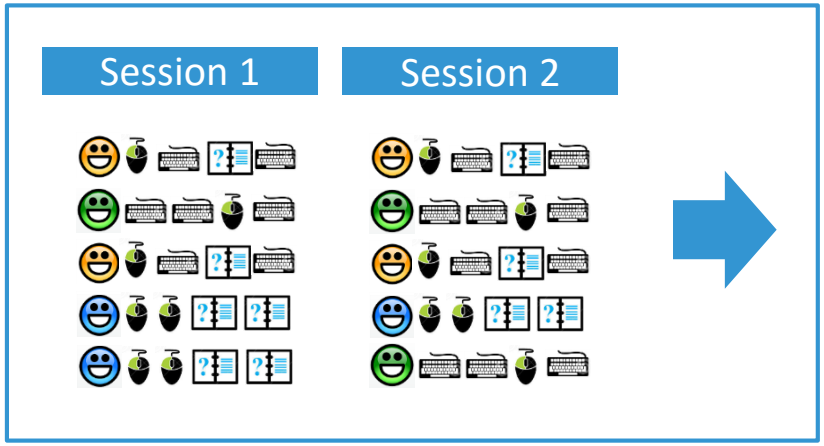


Related work

Entire sequence

e.g. [Bergner et al., 2014], [Martinez-Maldonado et al., 2013], [Herold et al., 2013]

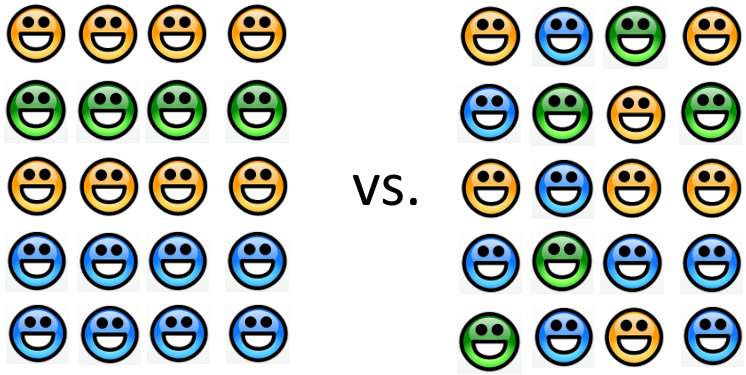
Evolutionary analysis



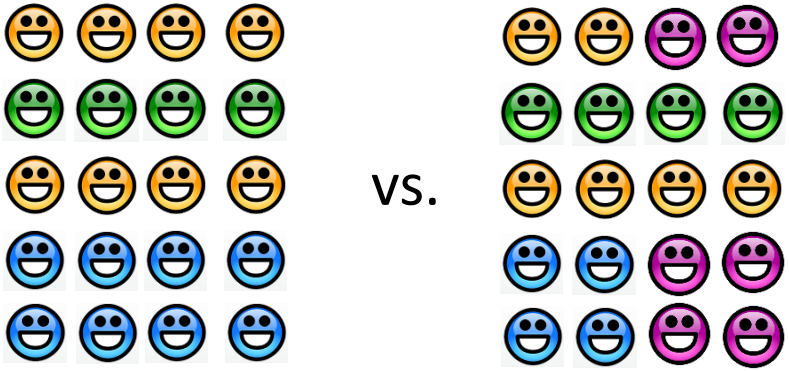
[Kinnebrew et al., 2013]

Challenges of evolutionary clustering

Temporal consistency



Cluster changes
(e.g. size and numbers)



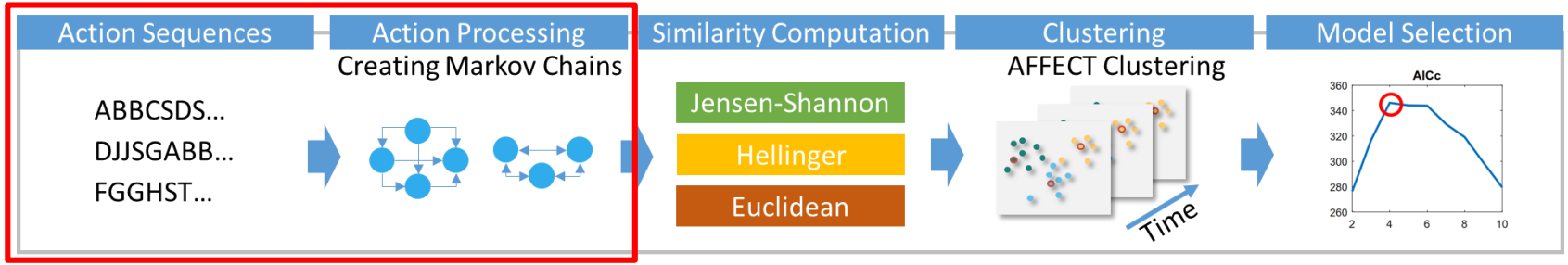
Contribution

Complete processing pipeline for evolutionary clustering based on AFFECT algorithm[Xu et al., 2014]

We propose several extensions to tailor method for educational data sets

Pipeline can be used as a black box for any ITS

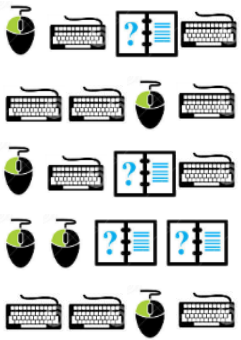
Pipeline overview



Action sequences

Input to our system

The only part that is system dependent



event mapping

Input, Backspace, Change View, Invalid Input

Input, Input, Input, Change View

Invalid Input, Backspace, Invalid Input, Backspace

Action processing

Action sequences

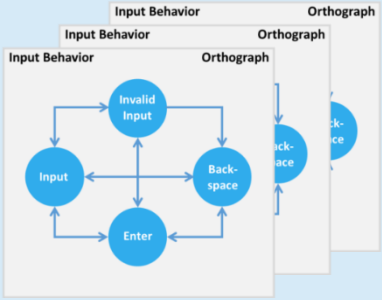
✔ provide rich temporal information

✘ exhibit considerable amount of noise

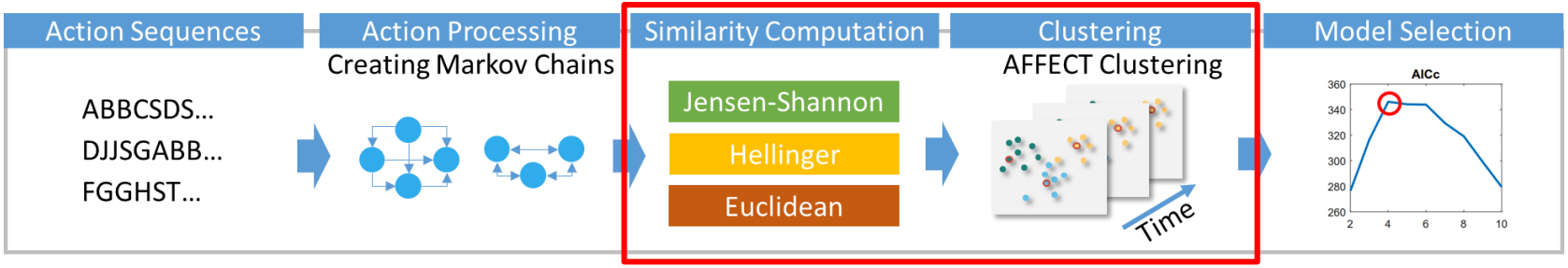
Markov chains

actions = states

transition probabilities = frequencies

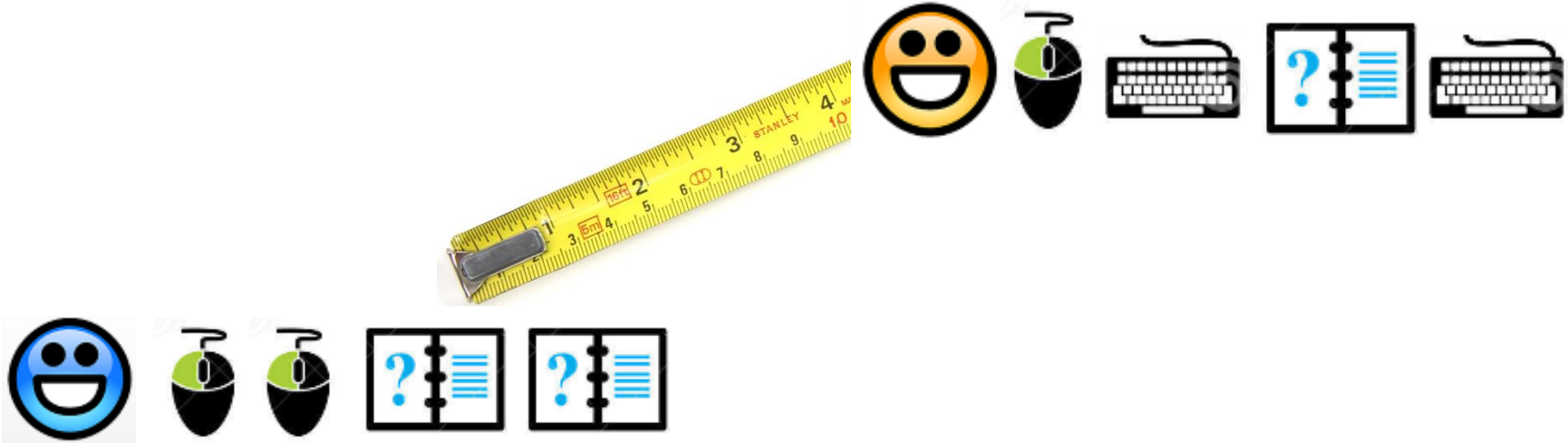


Pipeline overview



Similarity computation

Suitable similarity measure between students?



Similarity computation

Action sequence based

Longest common subsequences

[Bergner et al., 2014]

$$S_1 = AAACCGTGAGTTATTCGTTCTAGAA$$
$$S_2 = CACCCCTAAGGTACCTTTGGTTC$$

Levenshtein distance

[Desmarais & Lemieux, 2013]

I	N	T	E	*	N	T	I	O	N

*	E	X	E	C	U	T	I	O	N

Similarity computation

Action sequence based

Longest common subsequences

[Bergner et al., 2014]

$$S_1 = AAACCGTGAGTTATTCGTTCTAGAA$$
$$S_2 = CACCCCTAAGGTACCTTTGGTTC$$

Levenshtein distance

[Desmarais & Lemieux, 2013]

I	N	T	E	*	N	T	I	O	N

*	E	X	E	C	U	T	I	O	N

Markov chain based

Euclidean distance [Köck & Paramythis, 2011]

$$\mathrm{EUC}(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

Similarity computation

Action sequence based

Longest common subsequences

[Bergner et al., 2014]

$$S_1 = AAACCGTGAGTTATTCGTTCTAGAA$$
$$S_2 = CACCCCTAAGGTACCTTTGGTTC$$

Levenshtein distance

[Desmarais & Lemieux, 2013]

I	N	T	E	*	N	T	I	O	N

*	E	X	E	C	U	T	I	O	N

Markov chain based

Euclidean distance [Köck & Paramythis, 2011]

$$\mathrm{EUC}(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

Jenson-Shannon

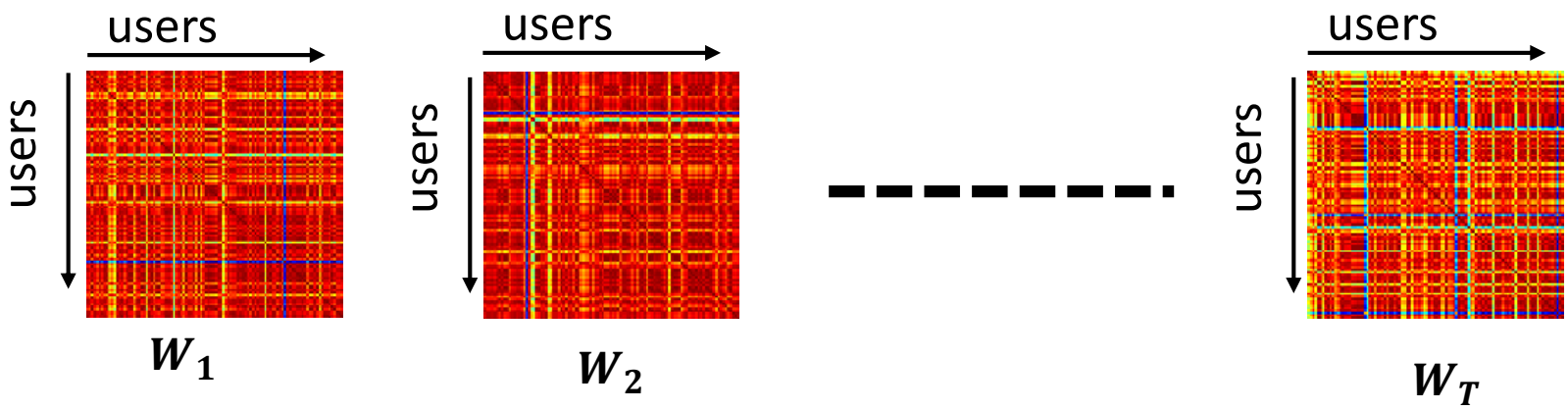
$$\mathrm{SD}(p,q) = \frac{1}{2}\sum_i p_i \frac{p_i}{q_i} + q_i \frac{q_i}{p_i}$$

Hellinger

$$\mathrm{HD}(p,q) = \frac{1}{\sqrt{2}}\sqrt{\sum_i (\sqrt{p_i} - \sqrt{q_i})^2}$$

Clustering

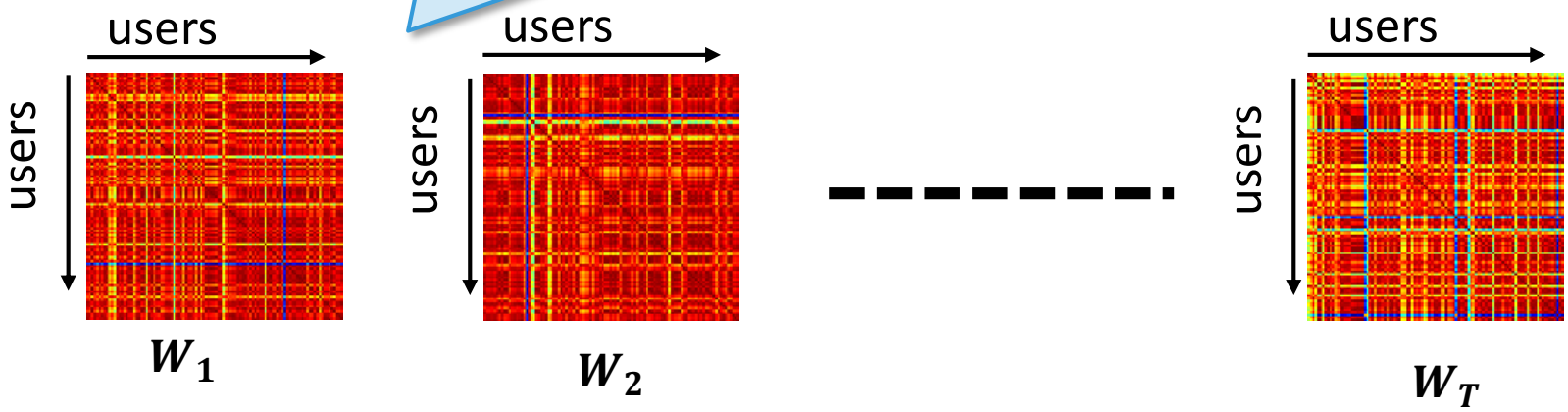
How to cluster the pairwise similarity matrices?



Clustering

How to cluster the pairwise similarity matrices?

Standard clustering at each time step?

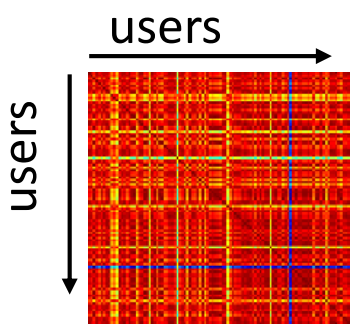


Clustering

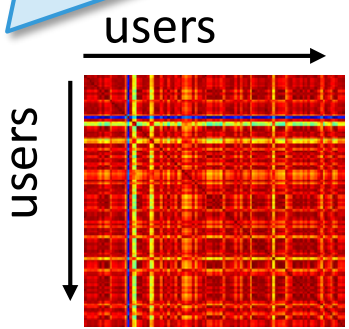
How to cluster the pairwise similarity matrices?

Standard clustering at each time step?

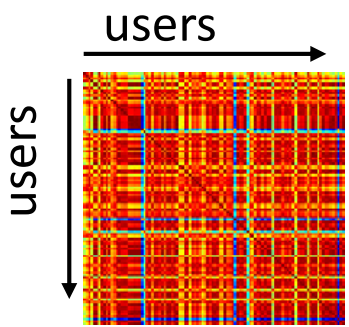
Does not use temporal information



W_1



W_2



W_T

AFFECT clustering [Xu et al.,2014]

Assumption

$$W^t = \Psi^t + N^t$$

observed similarities

AFFECT clustering [Xu et al.,2014]

Assumption

$$W^t = \Psi^t + N^t$$

observed similarities

true similarities

AFFECT clustering [Xu et al.,2014]

Assumption

$$W^t = \Psi^t + N^t$$

observed similarities

true similarities

random noise

AFFECT clustering [Xu et al.,2014]

Assumption

$$W^t = \Psi^t + N^t$$

Smoothed similarity matrix proposed

$$\hat{\Psi}^t = \alpha^t \hat{\Psi}^{t-1} + (1 - \alpha^t) W^t$$

AFFECT clustering [Xu et al.,2014]

Assumption

$$W^t = \Psi^t + N^t$$

Smooth proposed

previous best estimate of similarities

$$\hat{\Psi}^t = \alpha^t \hat{\Psi}^{t-1} + (1 - \alpha^t) W^t$$

AFFECT clustering [Xu et al.,2014]

Assumption

$$W^t = \Psi^t + N^t$$

Smooth proposed

previous best estimate of similarities

noisy observation

$$\hat{\Psi}^t = \alpha^t \hat{\Psi}^{t-1} + (1 - \alpha^t) W^t$$

AFFECT clustering [Xu et al.,2014]

Assumption

$$W^t = \Psi^t + N^t$$

Smooth proposed

previous best estimate of similarities

noisy observation

$$\hat{\Psi}^t = \alpha^t \hat{\Psi}^{t-1} + (1 - \alpha^t) W^t$$

controls amount of smoothing

AFFECT clustering [Xu et al.,2014]

Optimal α alpha as a trade-off:

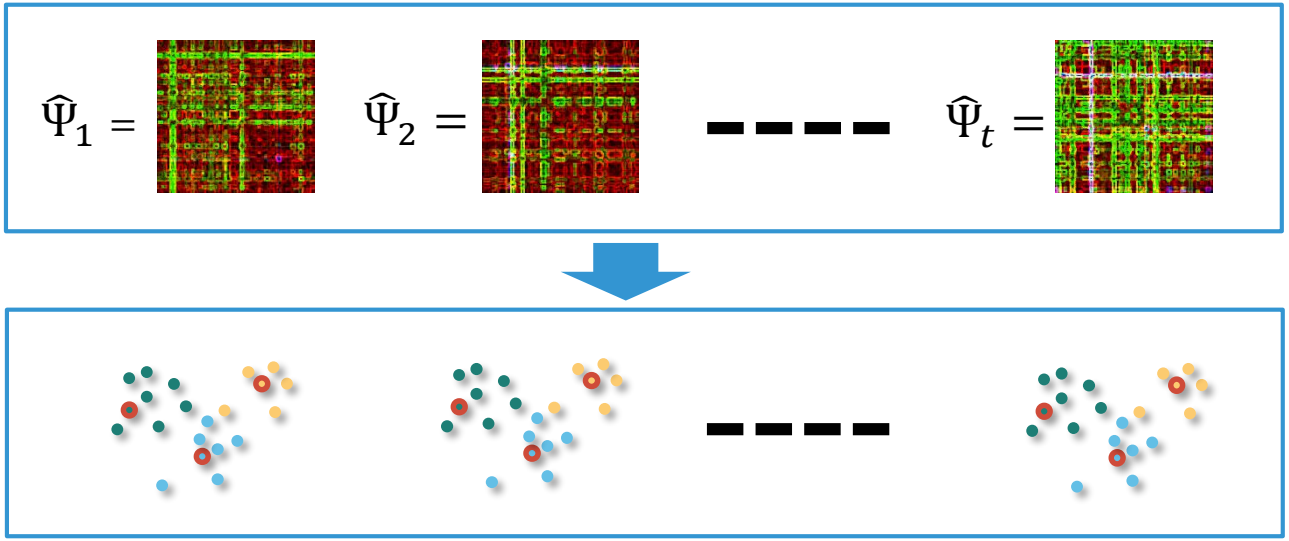
estimated noise

$$\alpha^t = \frac{\sum_i \sum_j \mathrm{var}(n_{ij}^t)}{\sum_i \sum_j (\hat{\psi}_{ij}^{t-1} - \psi_{ij}^t)^2 + \mathrm{var}(n_{ij}^t)}$$

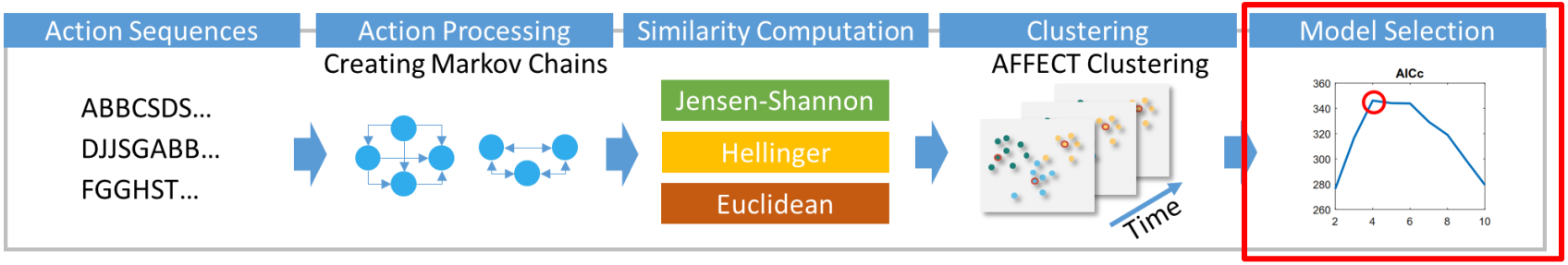
Amount of new information

AFFECT clustering [Xu et al.,2014]

Based on the estimates $\widehat{\Psi}_t$ we apply static **k means clustering**



Pipeline overview



Model selection

We expect clusters to change over time

- growth and shrinkage
- dissolving and forming

Determine the number of clusters at each time step

Model selection

$$AICc = -2\ln(LL) + 2P + \frac{2P(P+1)}{n-P-1}$$

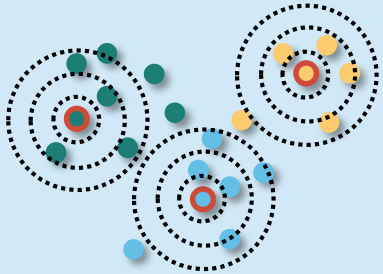
Model selection

$$AICc = -2\ln(LL) + 2P + \frac{2P(P+1)}{n-P-1}$$

Likelihood [Pelleg & Moore, 2000]

spherical Gaussians

based on empirical variance



Model selection

$$AICc = -2\ln(LL) + 2P + \frac{2P(P+1)}{n-P-1}$$

Number of parameters

based on effective dimensionality

[Krikpatrick, 2000]

Model selection

$$AICc = -2\ln(LL) + 2P + \frac{2P(P+1)}{n-P-1}$$

Correction for finite sample size

[Burnham & Anderson, 2002]

Evaluation

Synthetic experiments

- Performance evaluation of our method based on ground truth
- Robustness to noise

Exploratory data analysis

- Cluster extraction on real world data
- Comparison across ITS

Evaluation

Synthetic experiments

- Performance evaluation of our method based on ground truth
- Robustness to noise

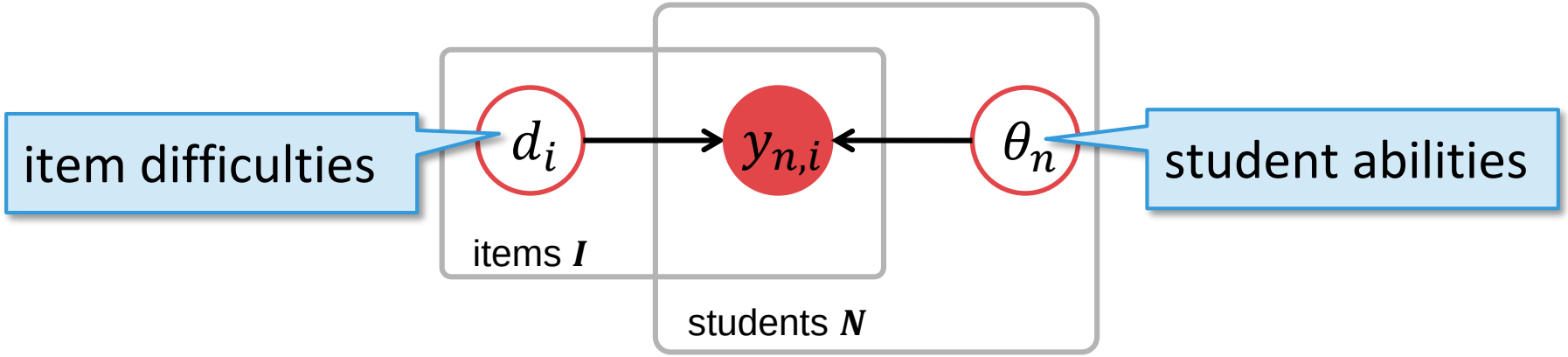
Exploratory data analysis

- Cluster extraction on real world data
- Comparison across ITS

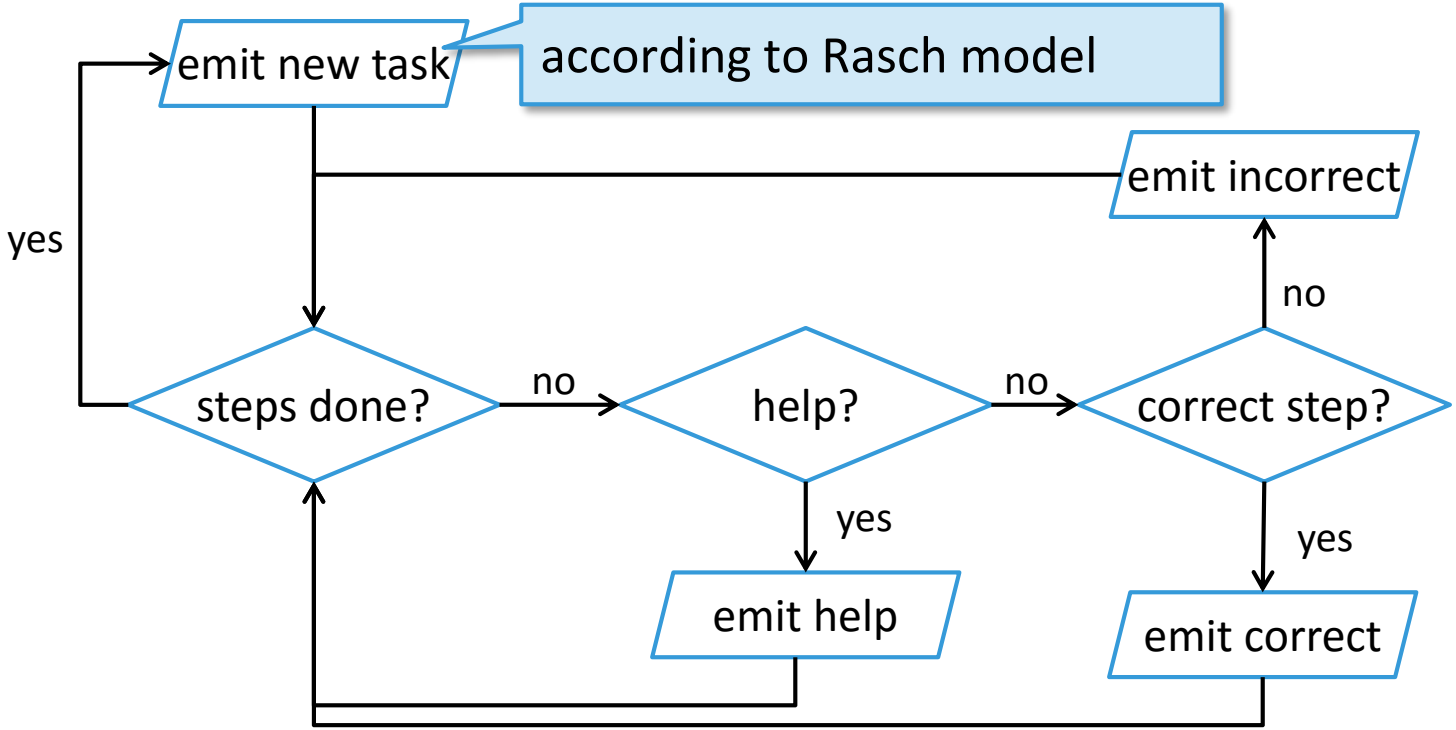
Synthetic data generation

80 students over 50 sessions solving 20 tasks (8 steps)

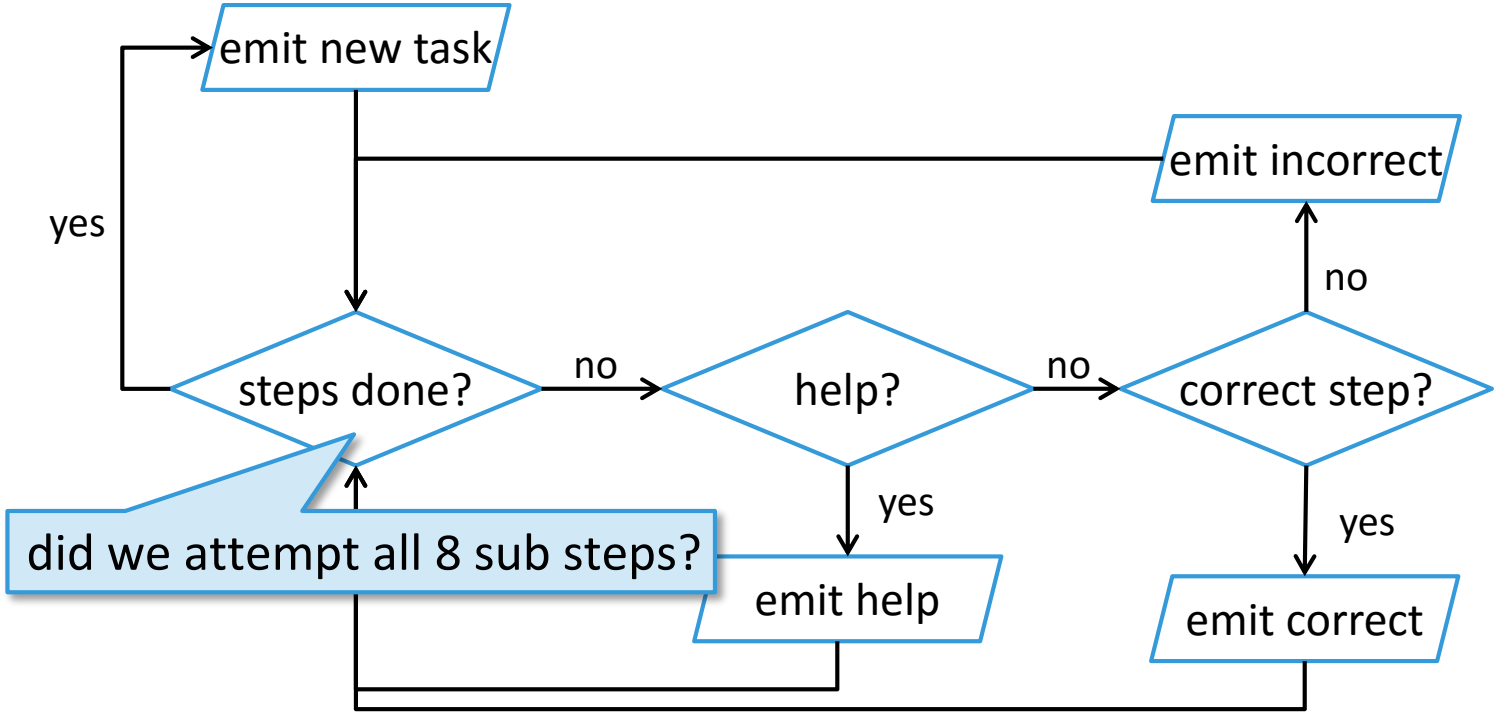
Rasch model for probability of correctly solving a task



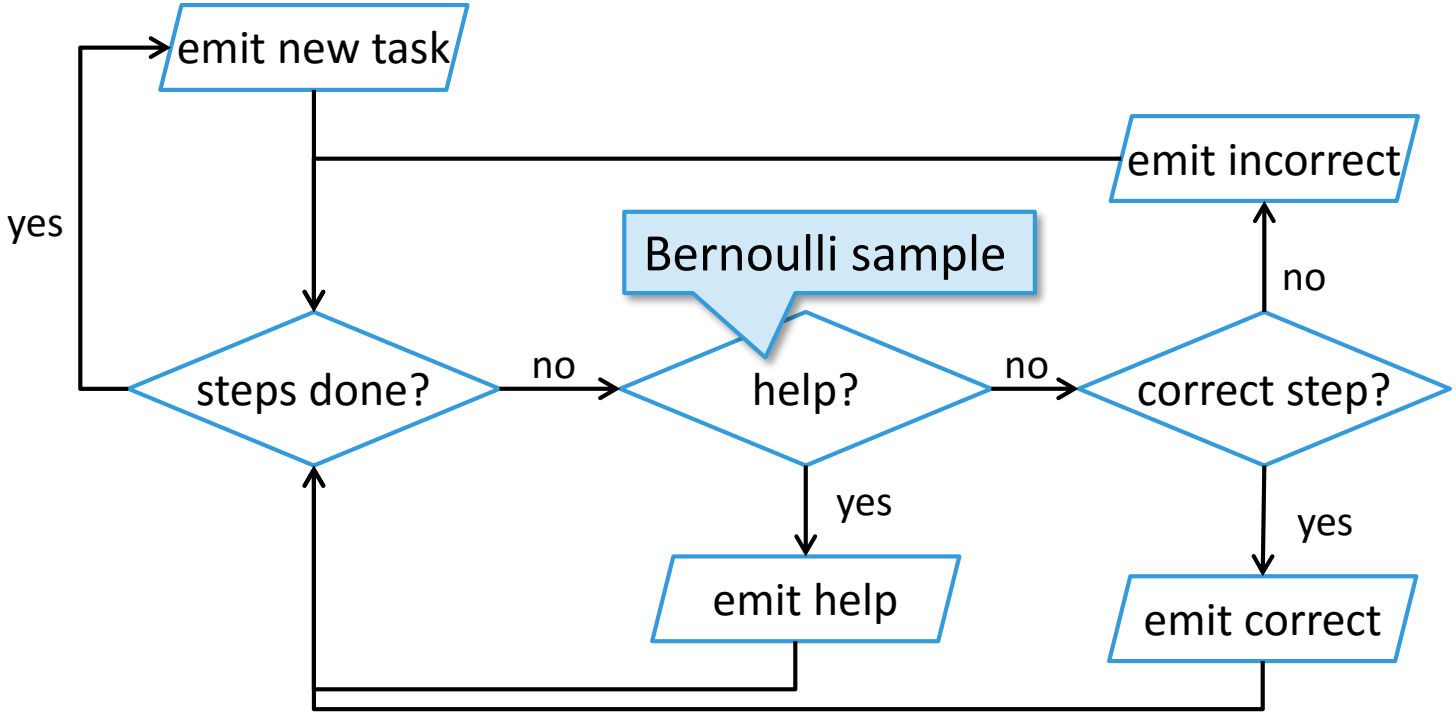
Synthetic data generation



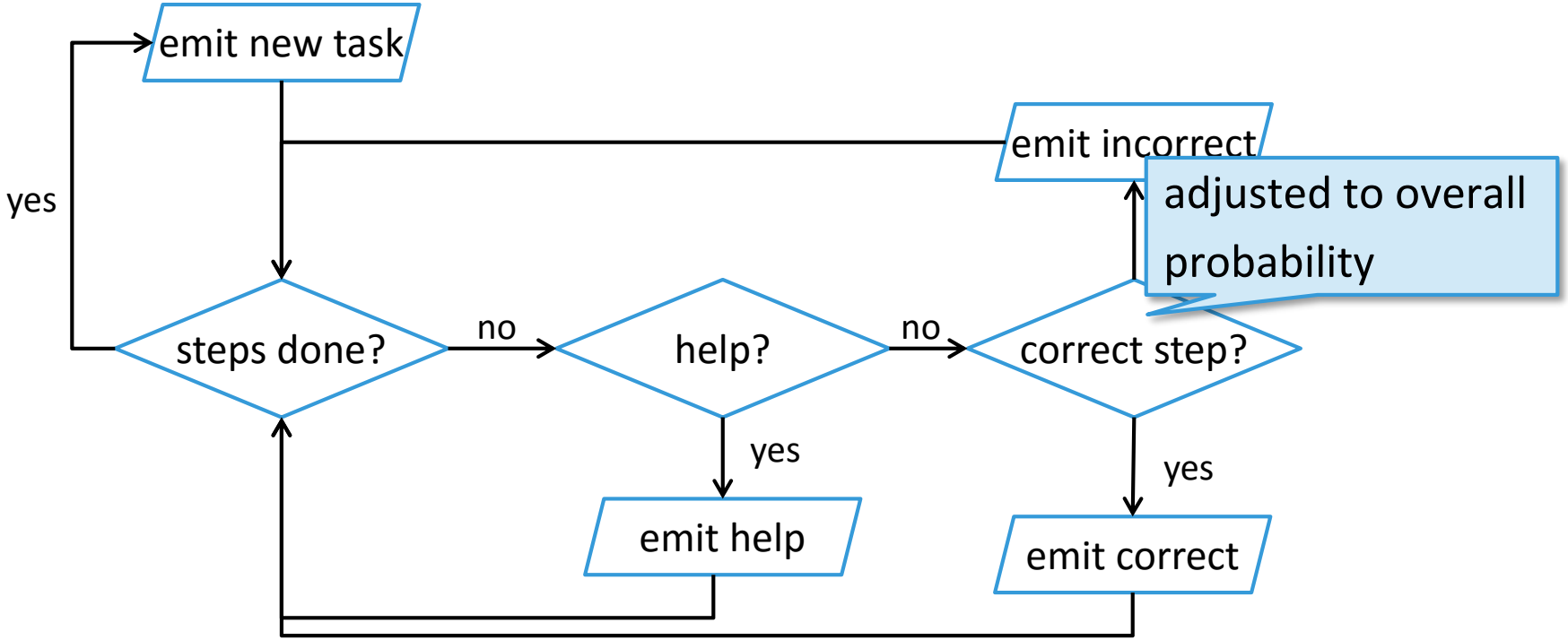
Synthetic data generation



Synthetic data generation



Synthetic data generation



Synthetic data generation

We simulated four student groups with different behavior

Good performance	Frequent help request	θ	p_H
✗	✗	-1	0.05
✓	✗	1	0.05
✗	✓	-1	0.2
✓	✓	1	0.2

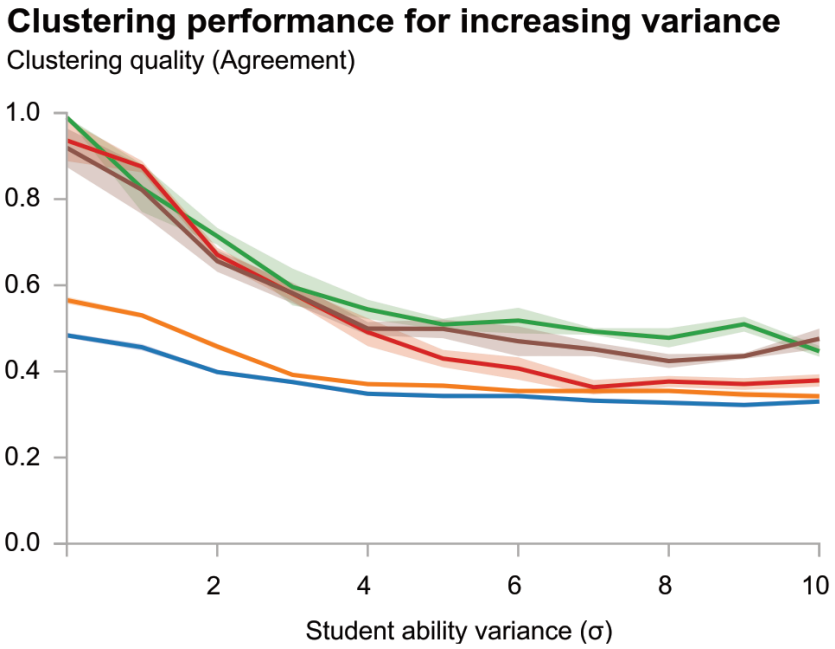
Clustering Quality & Robustness

We compare to the following methods (cross-validated)

LCS_KM*	Longest common subsequence	->	k-means
MC_EUC_KM**	Markov chains -> Euclidean dist.	->	k-means
Ours_HD	Markov chains -> Hellinger dist.	->	AFFECT clustering
Ours_SD	Markov chains -> Shannon div .	->	AFFECT clustering
Ours_EUC	Markov chains -> Euclidean dist.	->	AFFECT clustering

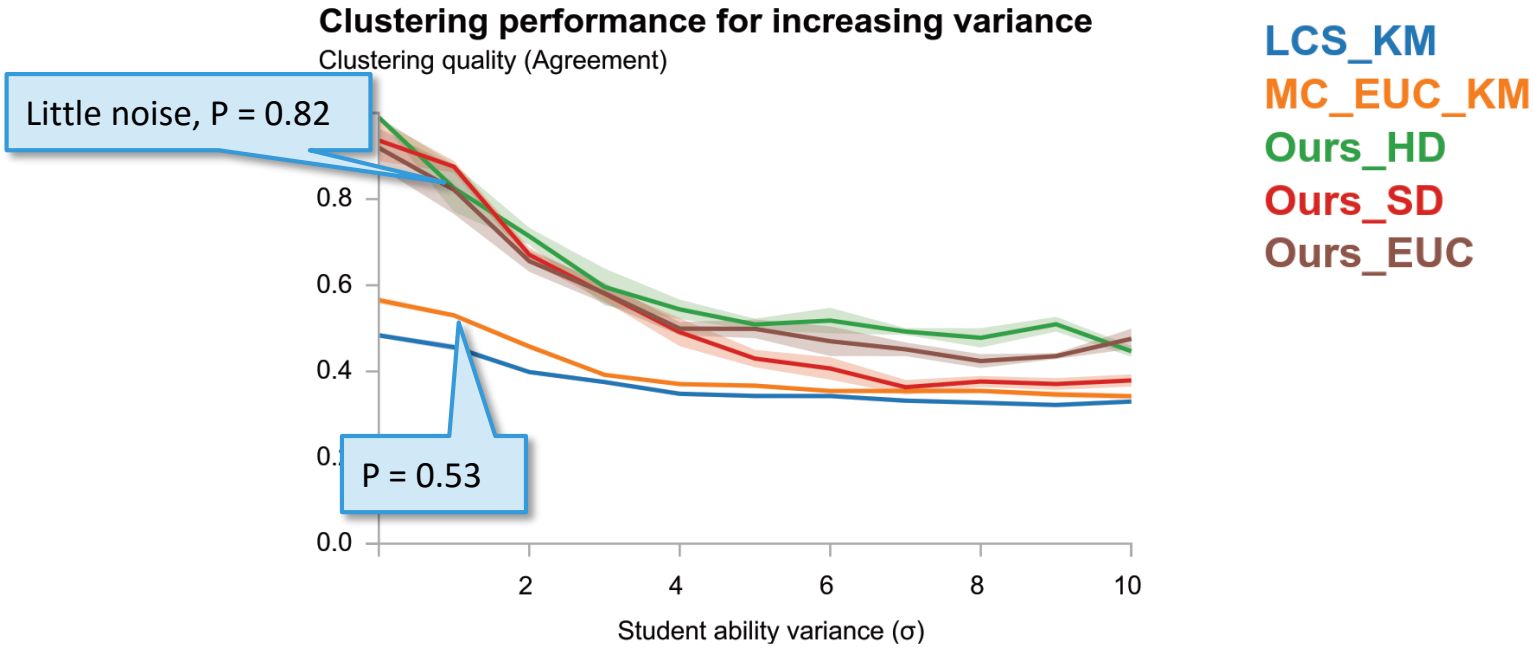
* [Bergner et al., 2014] **[Köck & Paramythis, 2011]

Clustering Quality & Robustness

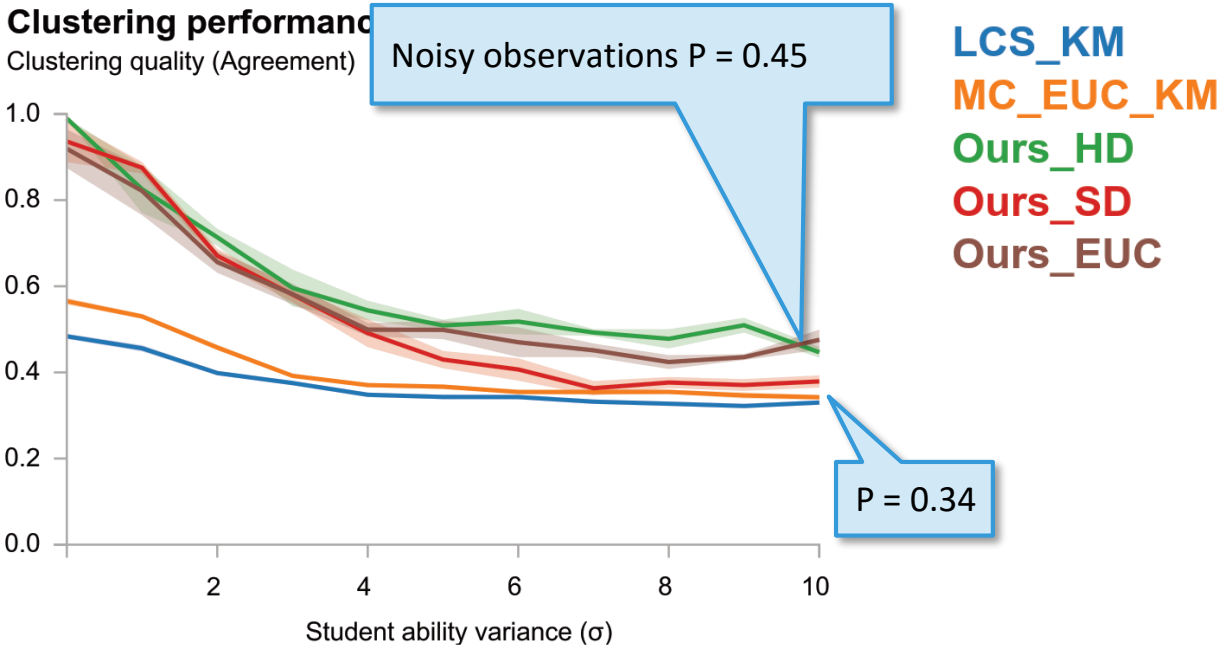


LCS_KM
MC_EUC_KM
Ours_HD
Ours_SD
Ours_EUC

Clustering Quality & Robustness

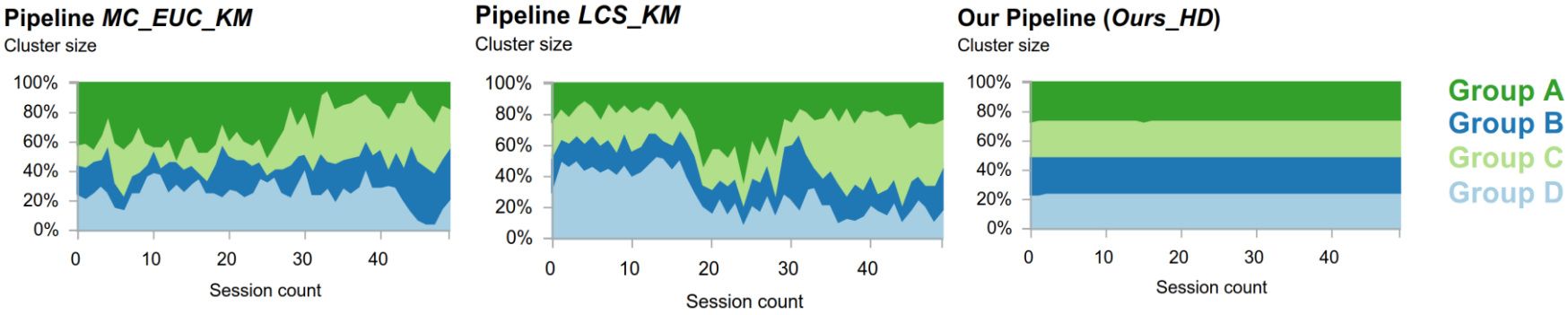


Clustering Quality & Robustness



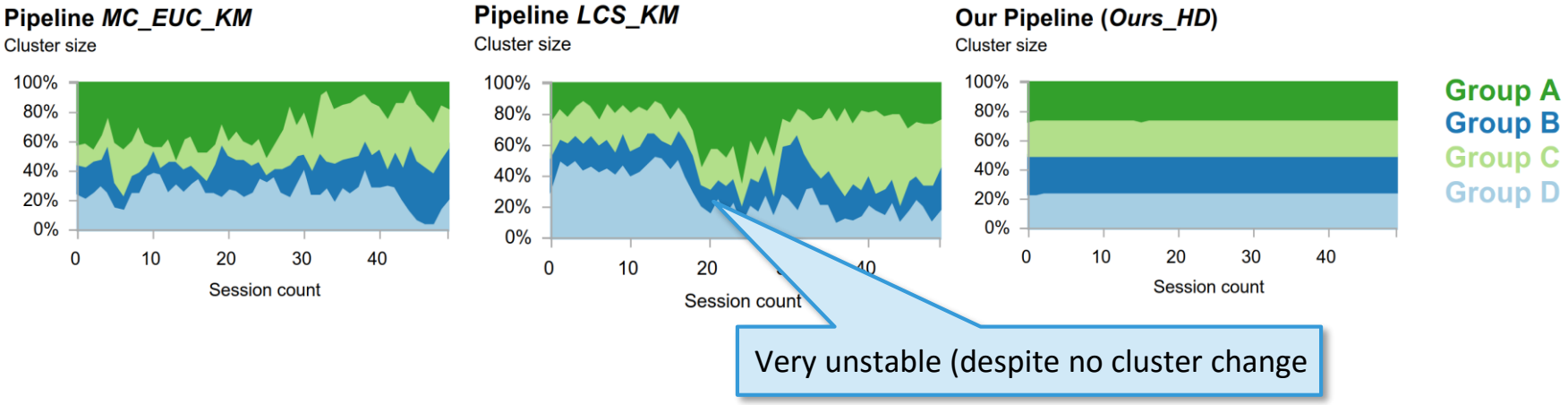
Stability

Temporal stability over 50 simulated sessions



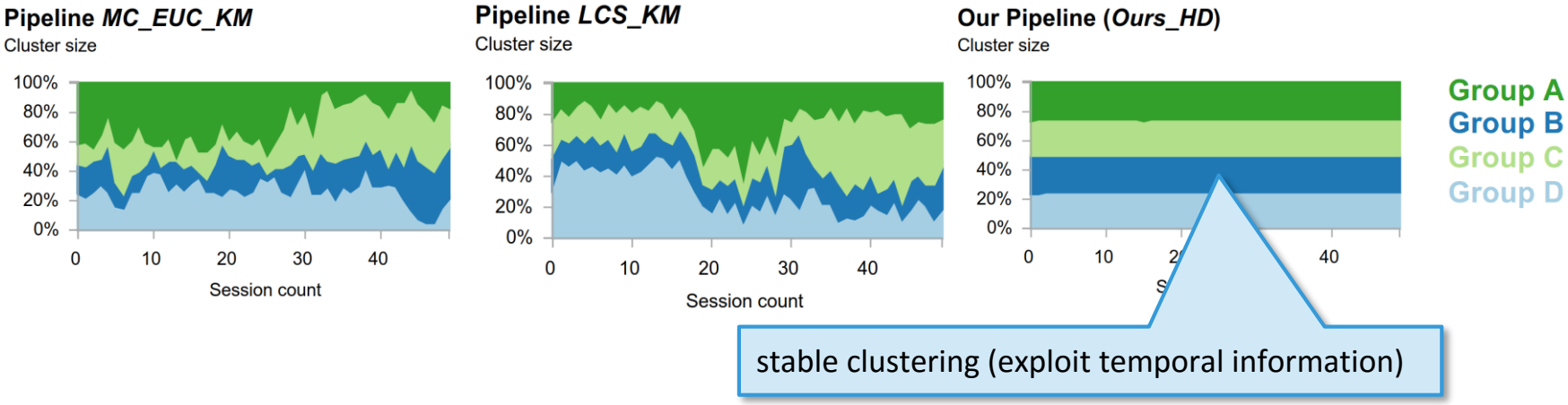
Stability

Temporal stability over 50 simulated sessions



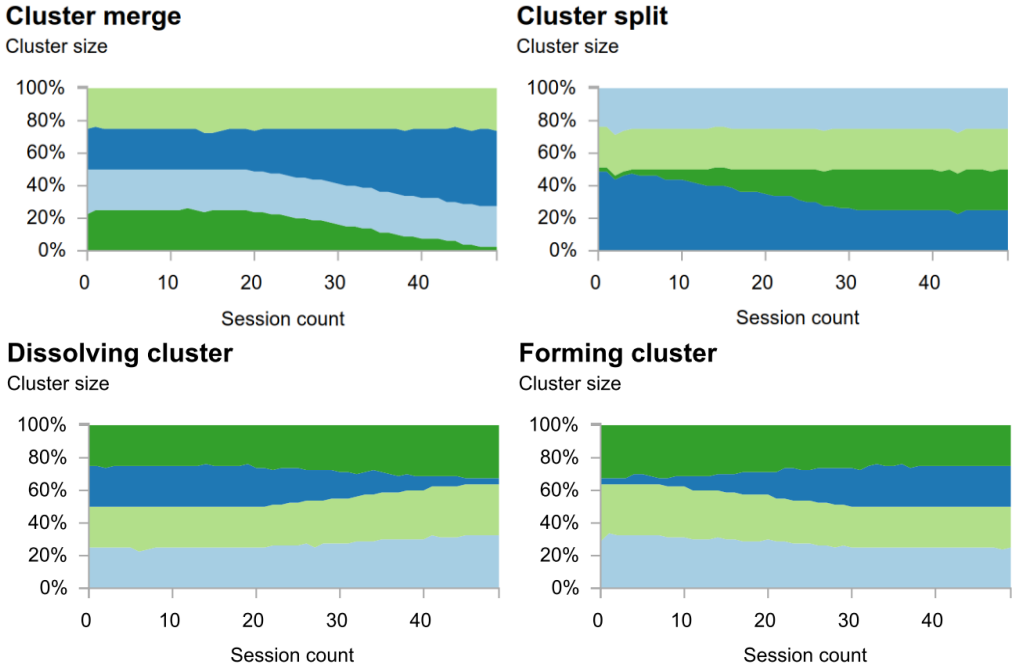
Stability

Temporal stability over 50 simulated sessions



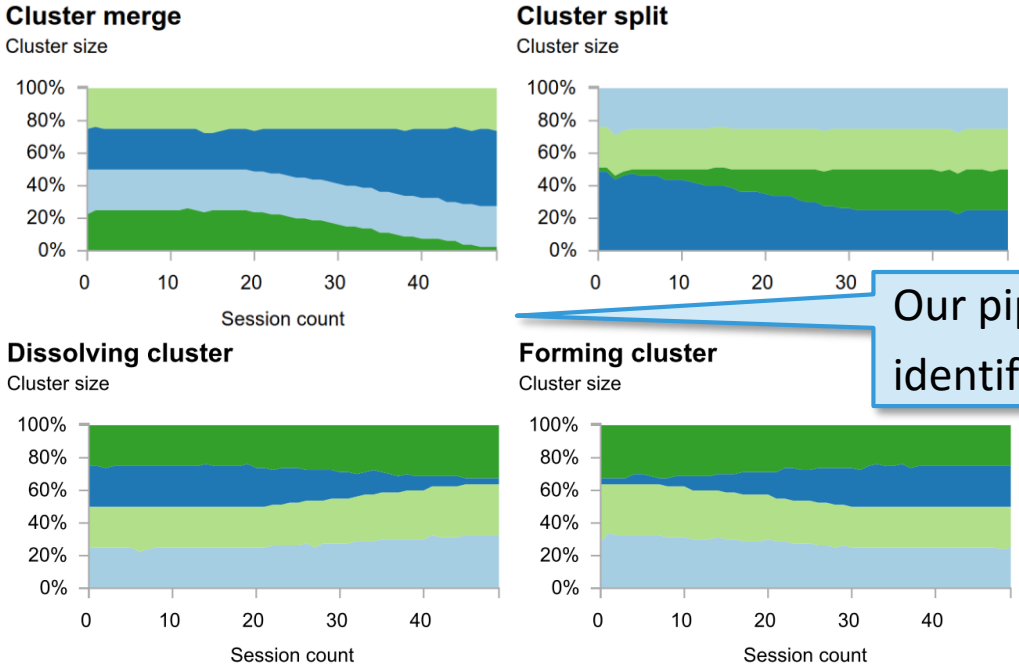
Interpretability

Identification of cluster numbers and sizes



Interpretability

Identification of cluster numbers and sizes



Our pipeline correctly identifies cluster events.

Evaluation

Synthetic experiments

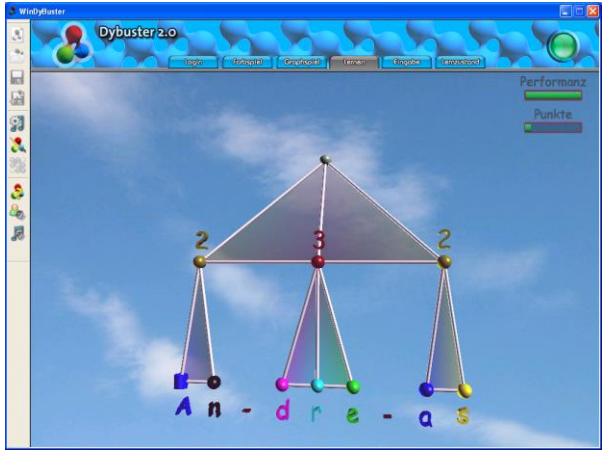
- Performance evaluation of our method based on ground truth
- Robustness to noise

Exploratory data analysis

- Cluster extraction on real world data
- Comparison across ITS

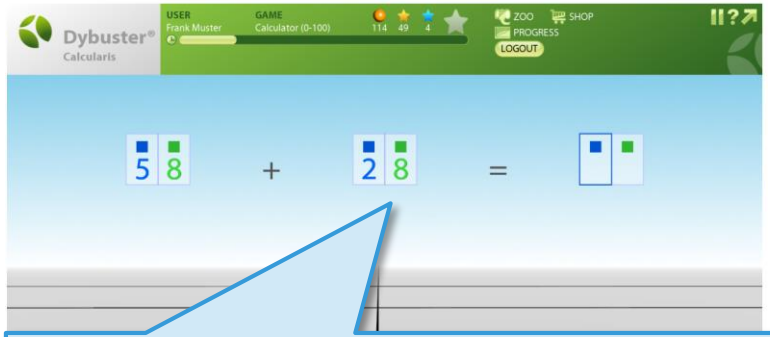
Exploratory data analysis

Clustering student interactions in two different ITS



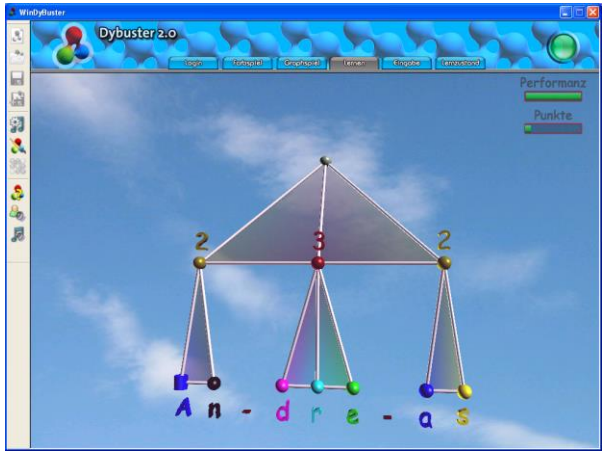
Exploratory data analysis

Clustering student interactions in two different ITS



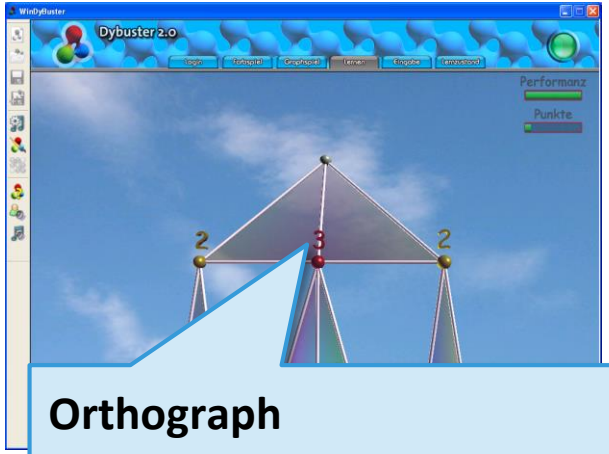
Calcularis

Data from 134 students
Intelligent tutoring system
Children with difficulties in mathematics



Exploratory data analysis

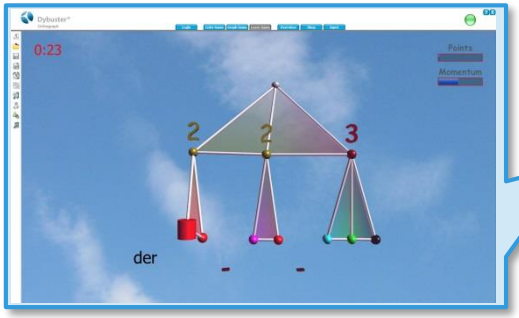
Clustering student interactions in two different ITS

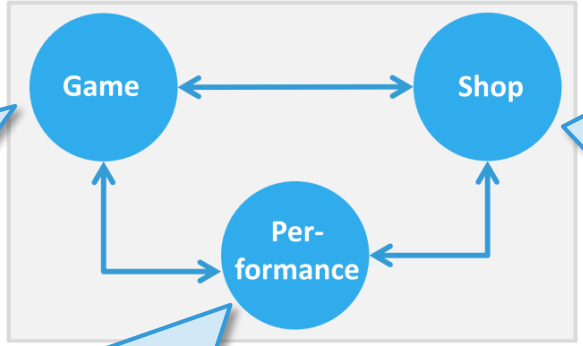


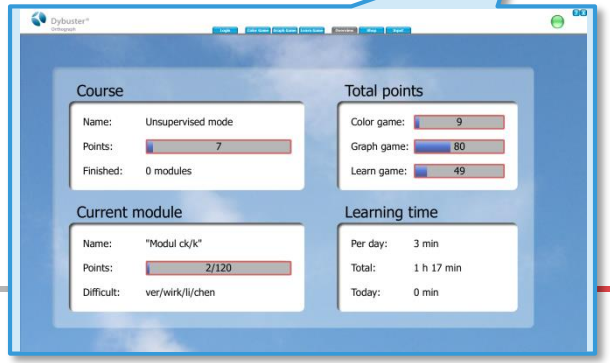
Orthograph

Data from 106 students
Computer-based training
Children with dyslexia

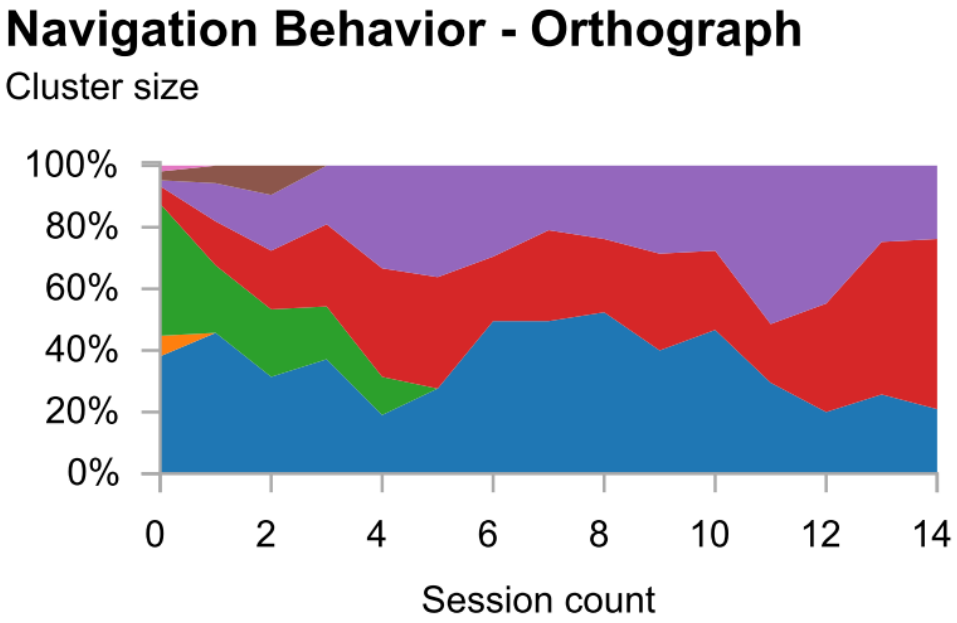
Example: Navigation behavior



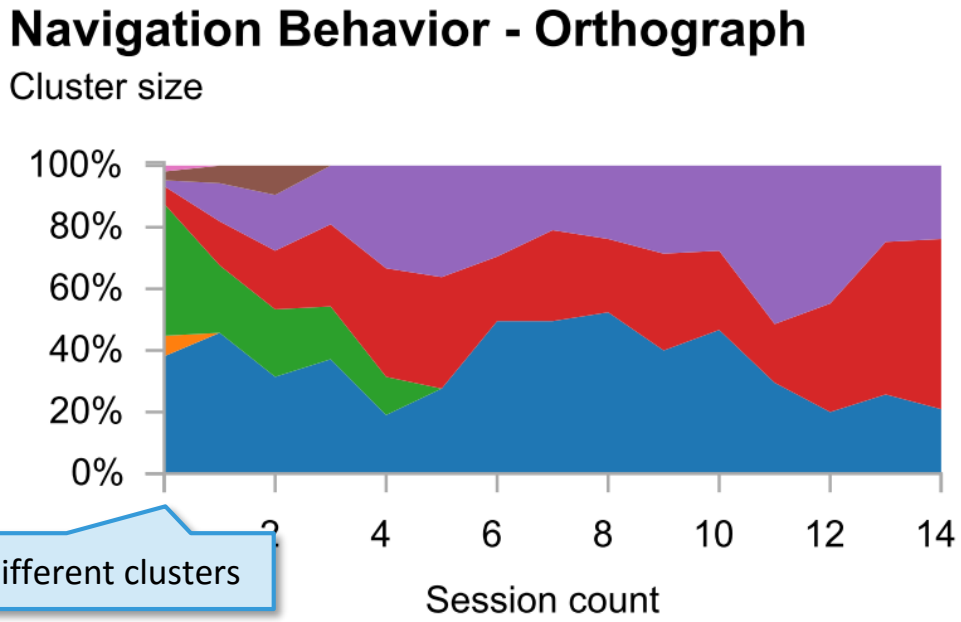




Navigation behavior

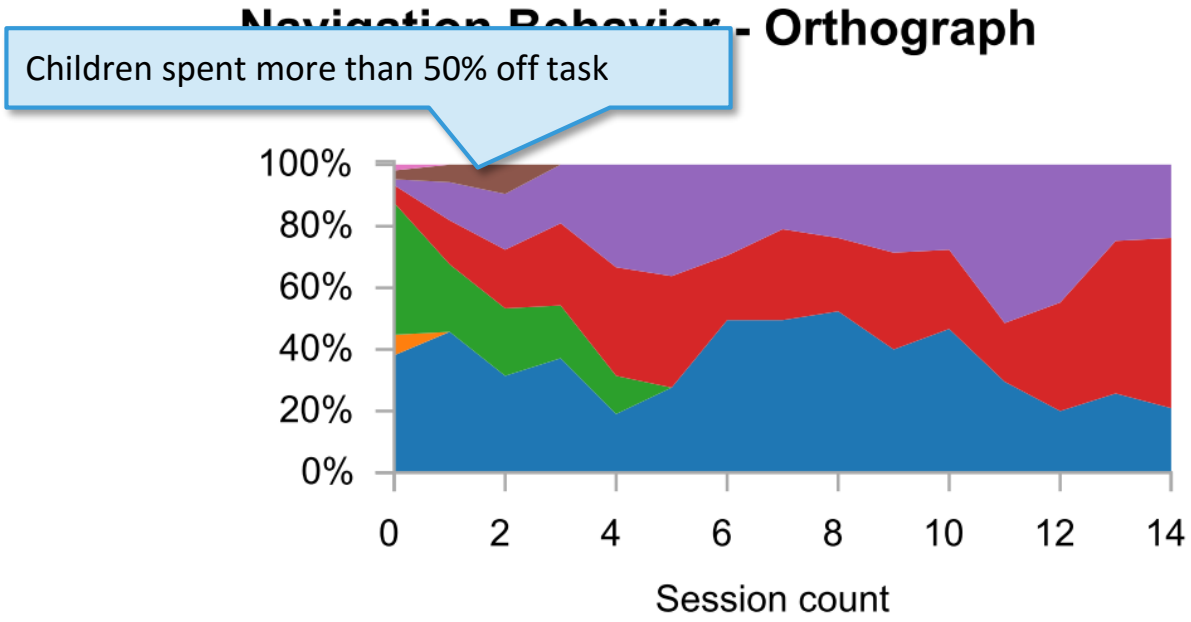


Navigation behavior

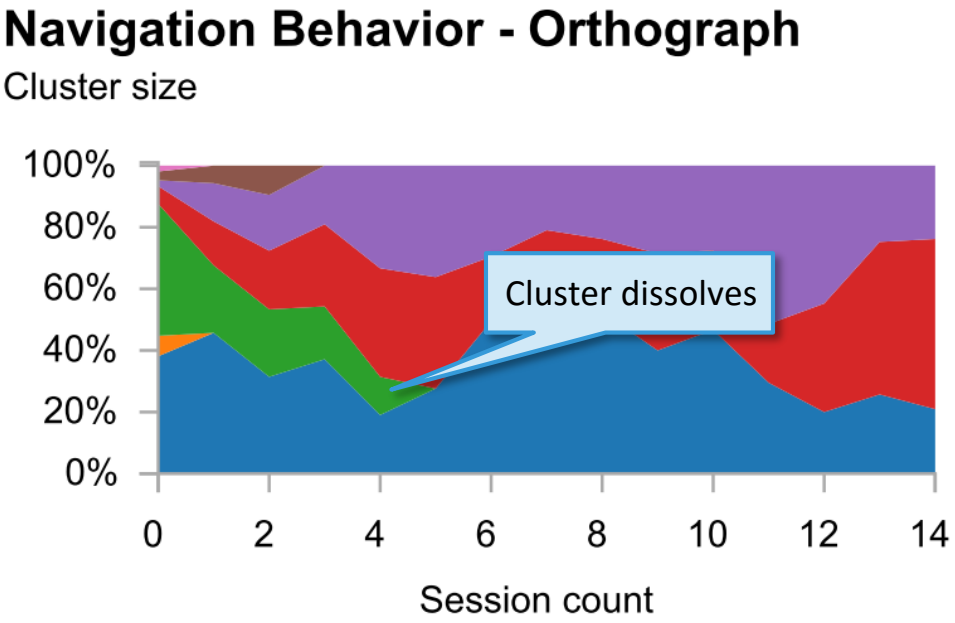


Beginning 7 different clusters

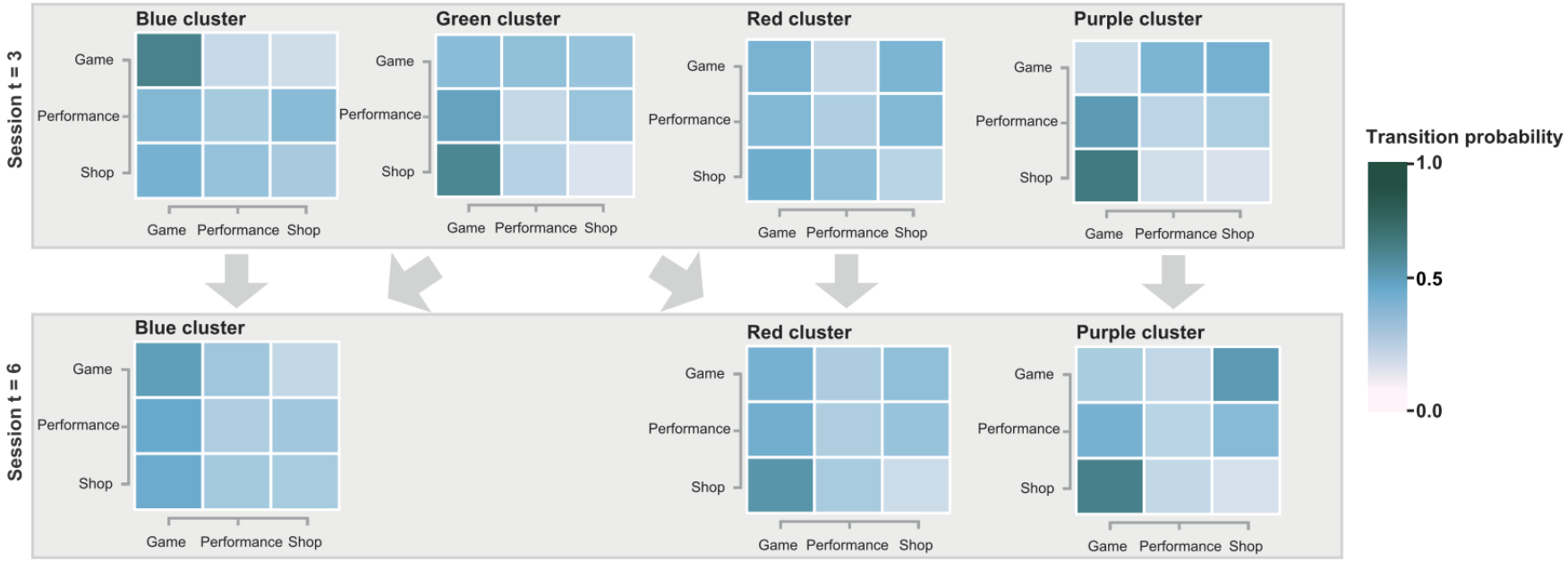
Navigation behavior



Navigation behavior

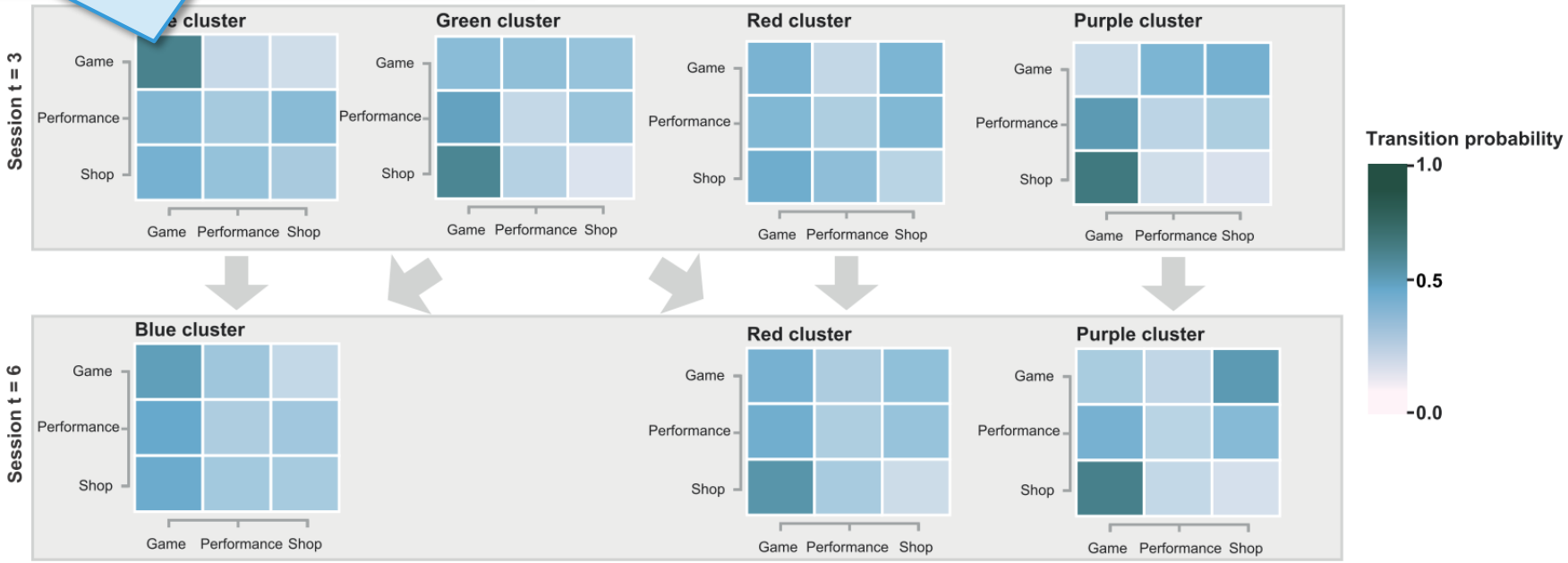


Navigation behavior

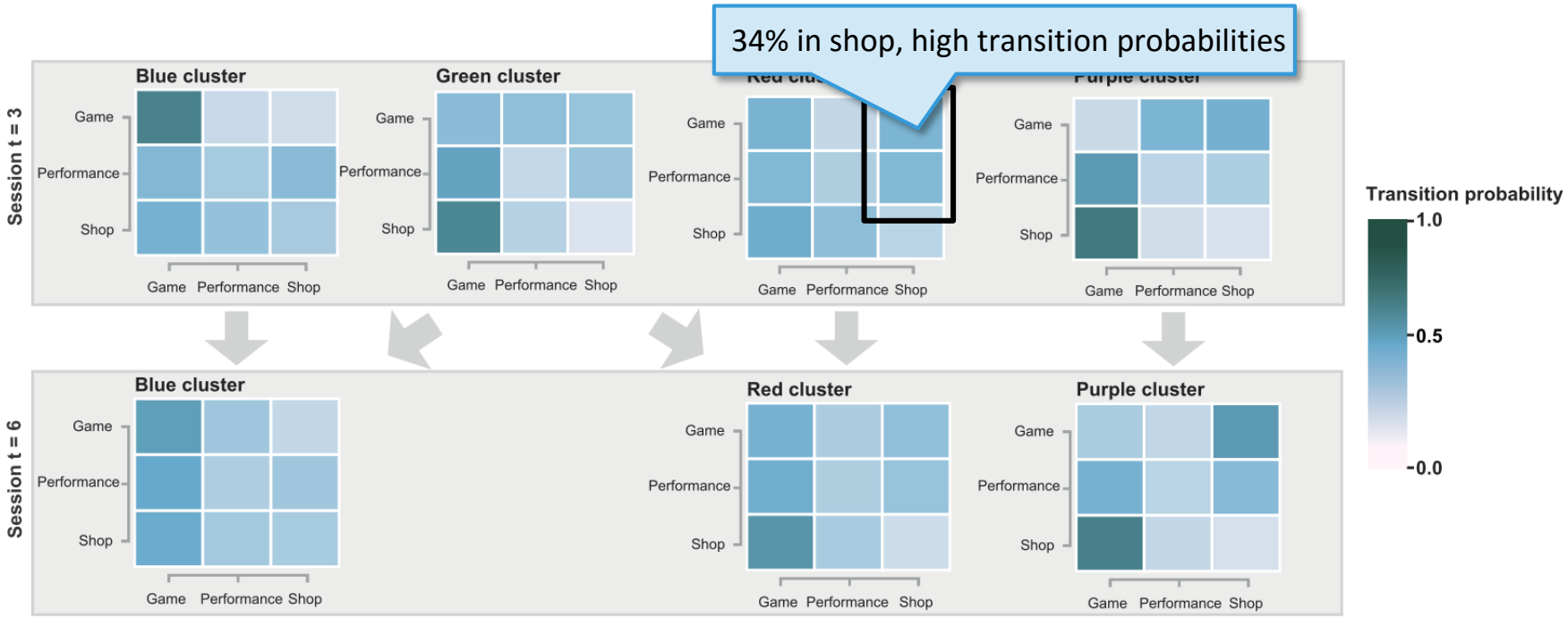


Navigation behavior

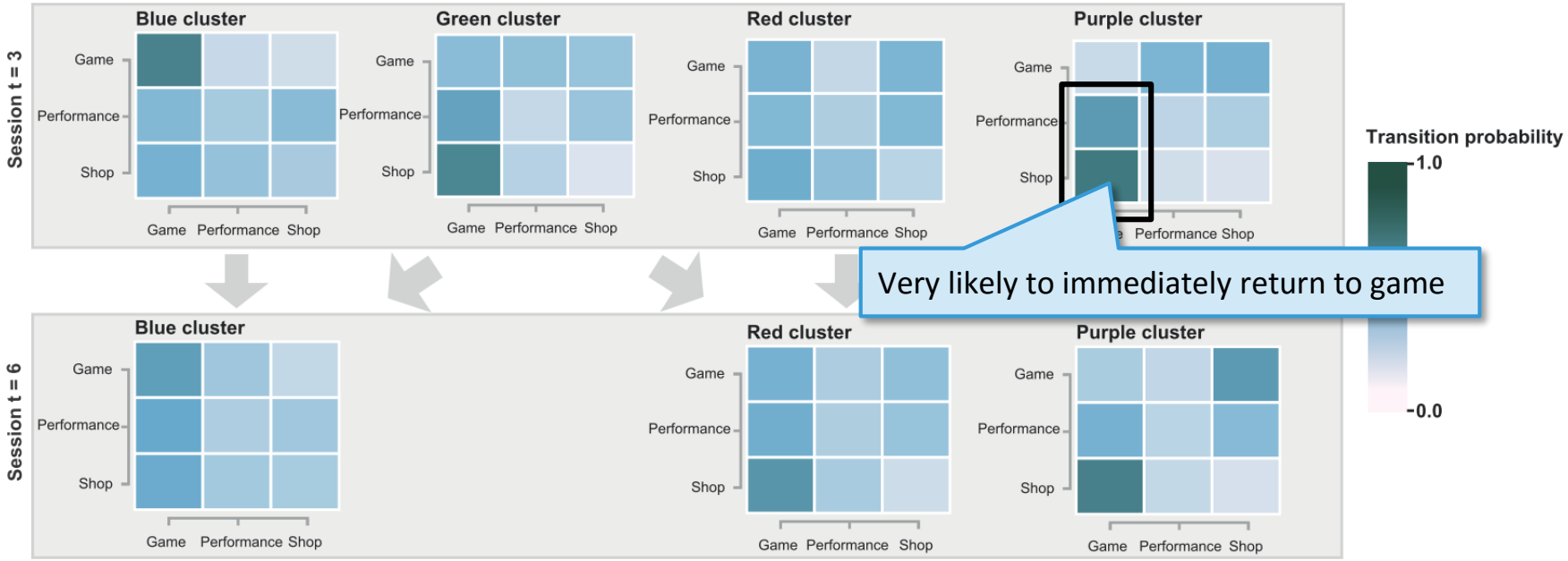
Very focused on training (80% in training)



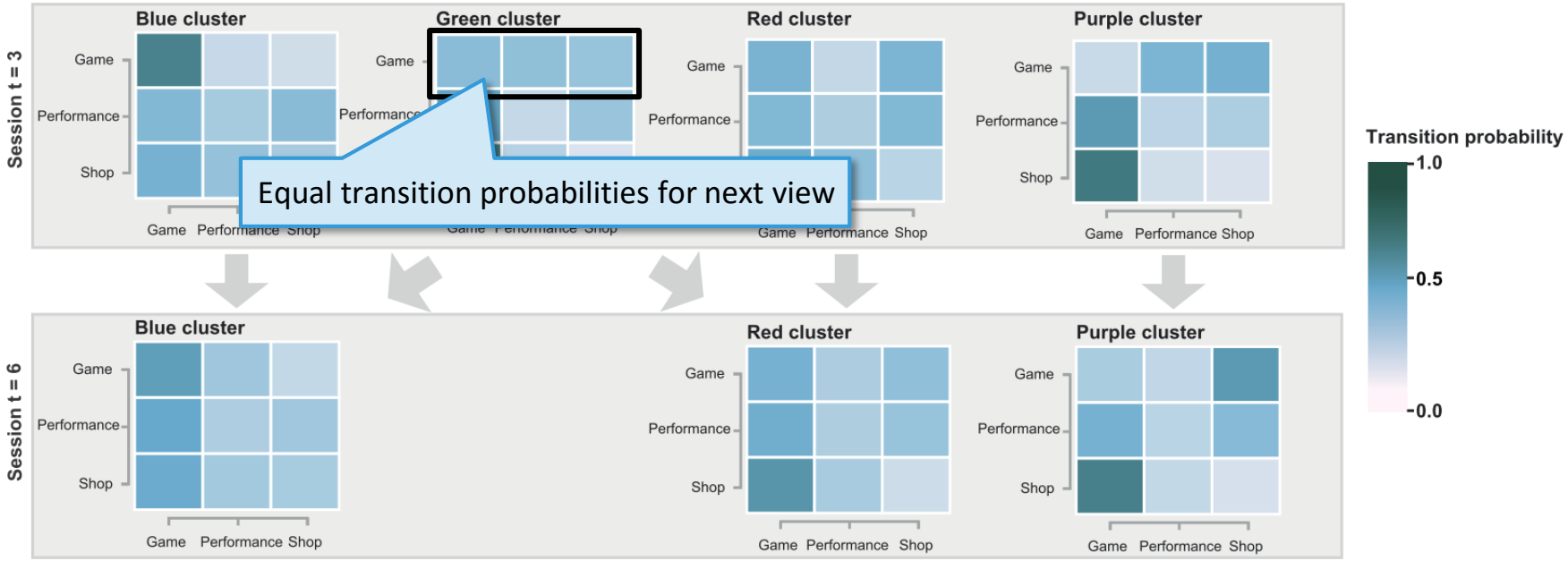
Navigation behavior



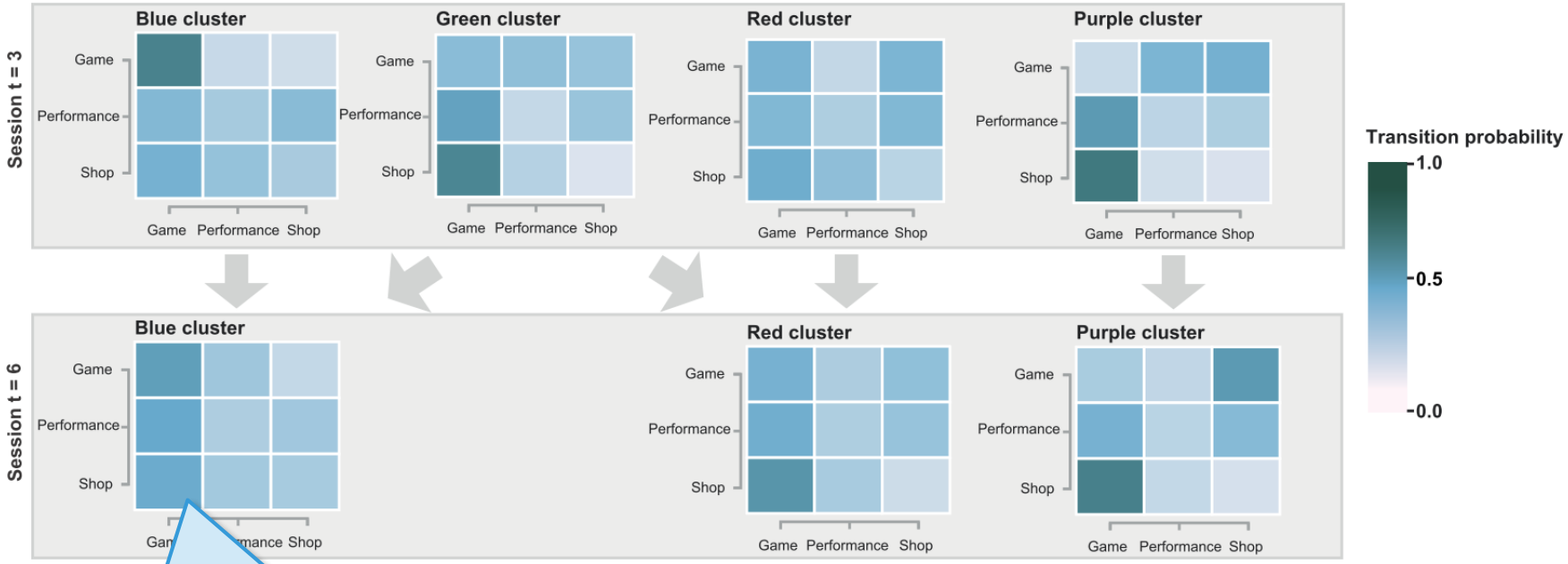
Navigation behavior



Navigation behavior

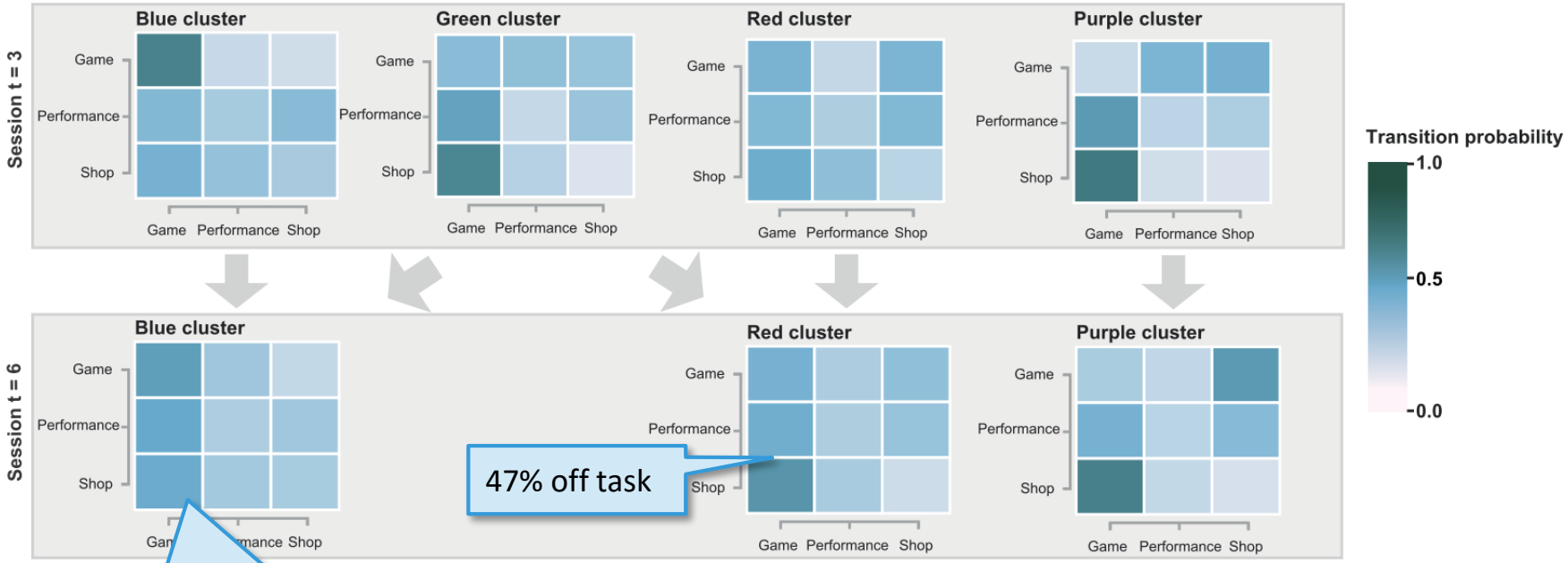


Navigation behavior

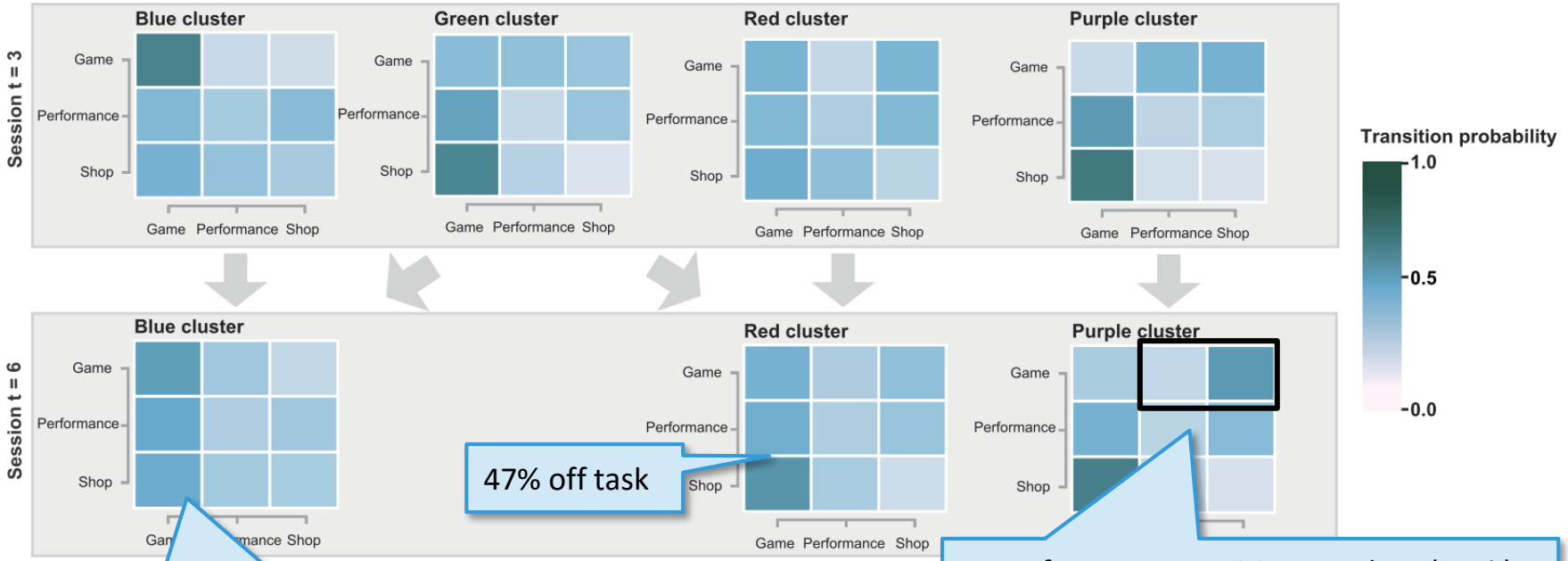


Still focused, 76% in training

Navigation behavior



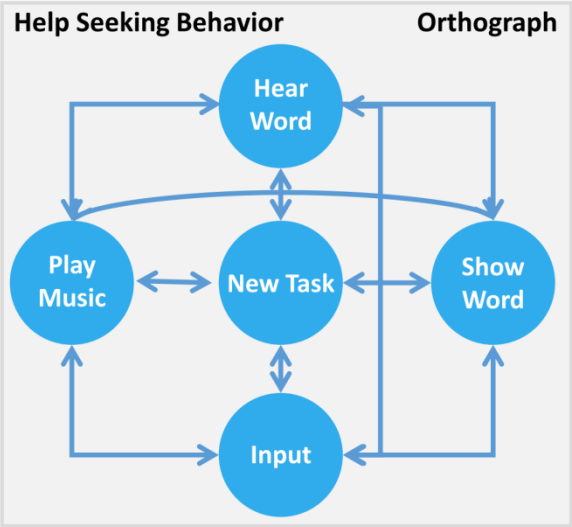
Navigation behavior

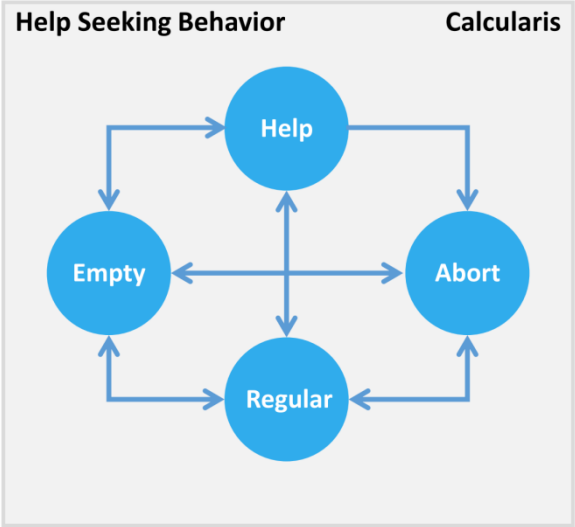


Still focused, 76% in training

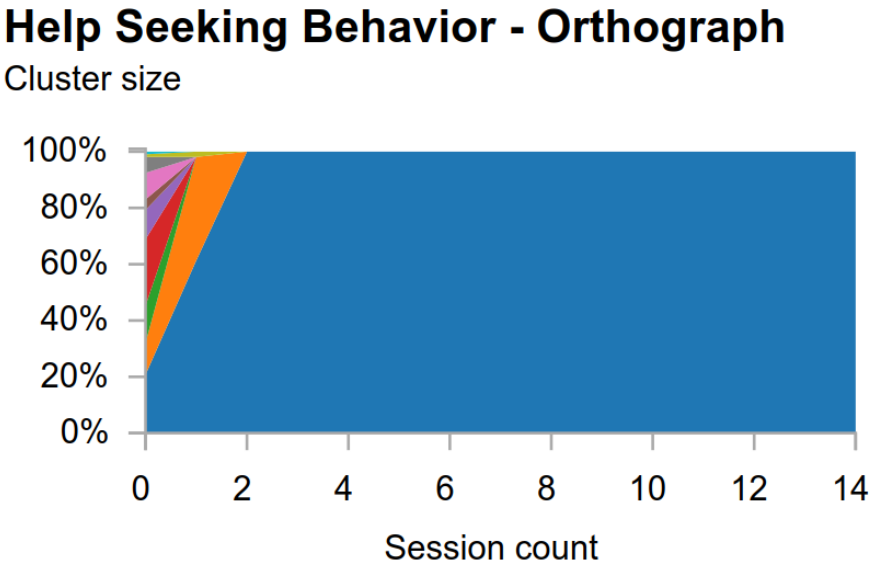
More frequent transitions to shop (58%) than performance (17%)

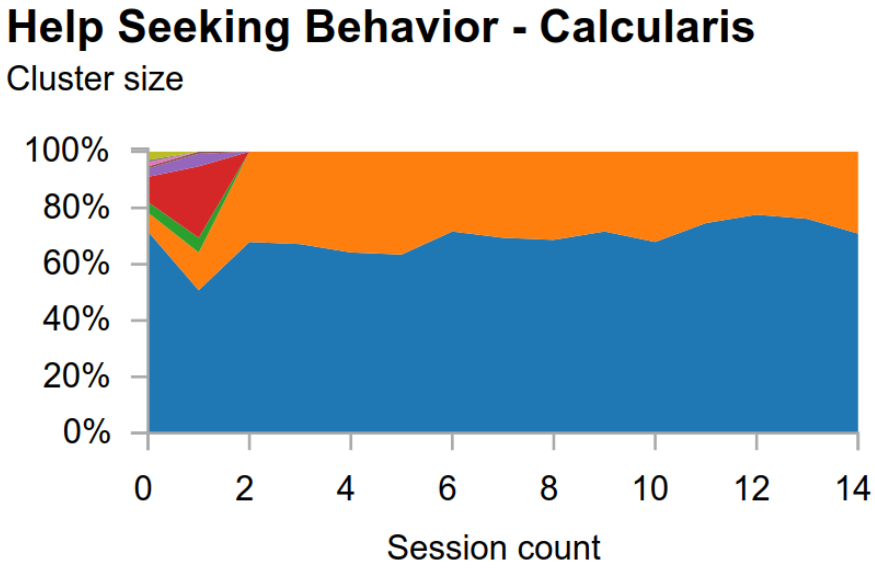
Help Seeking Behavior



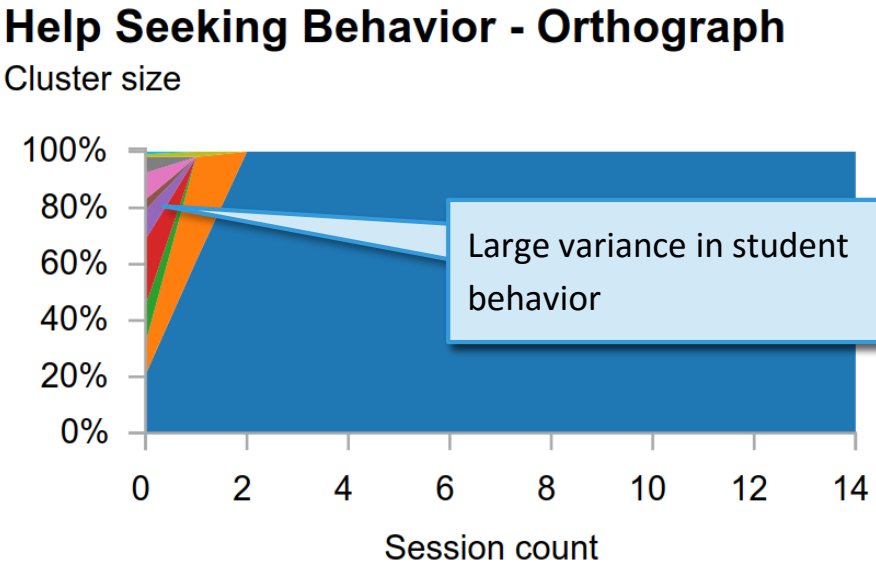


Help Seeking Behavior

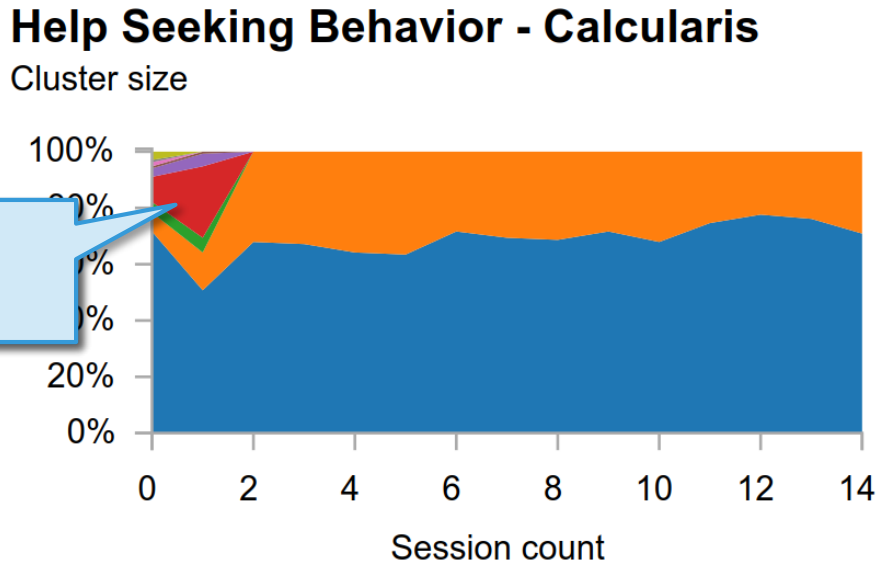




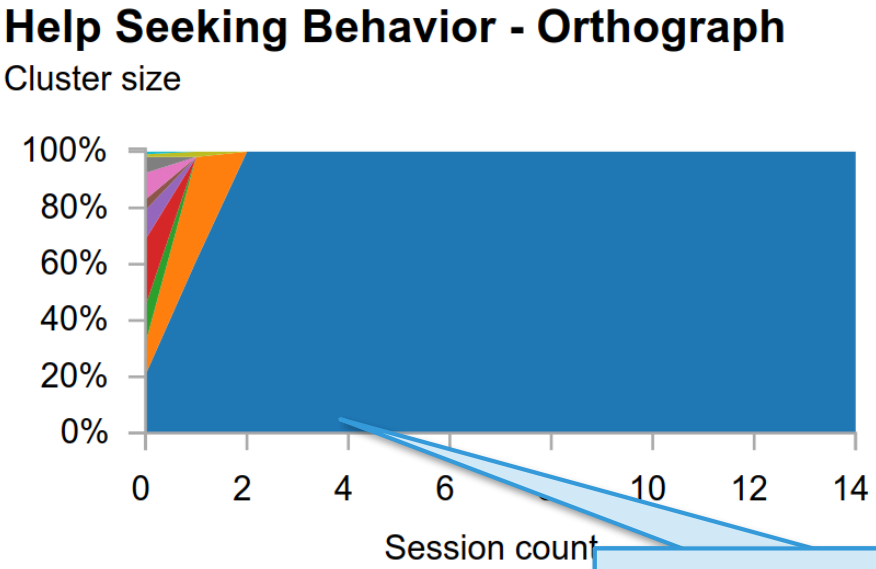
Help Seeking Behavior

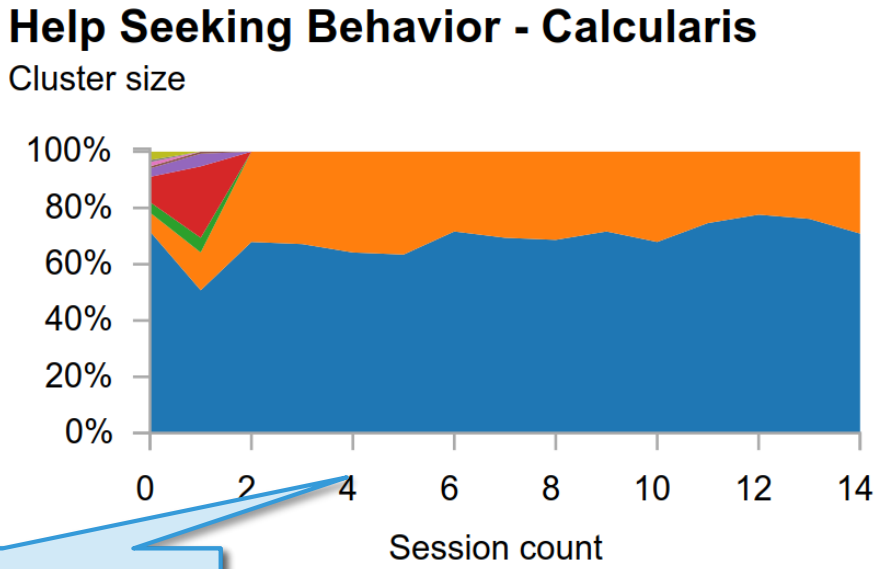


Large variance in student behavior



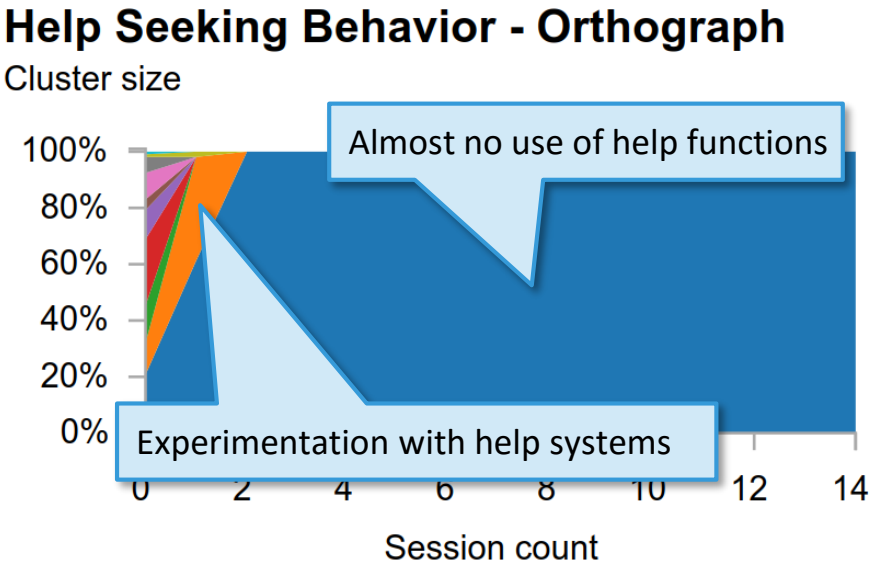
Help Seeking Behavior

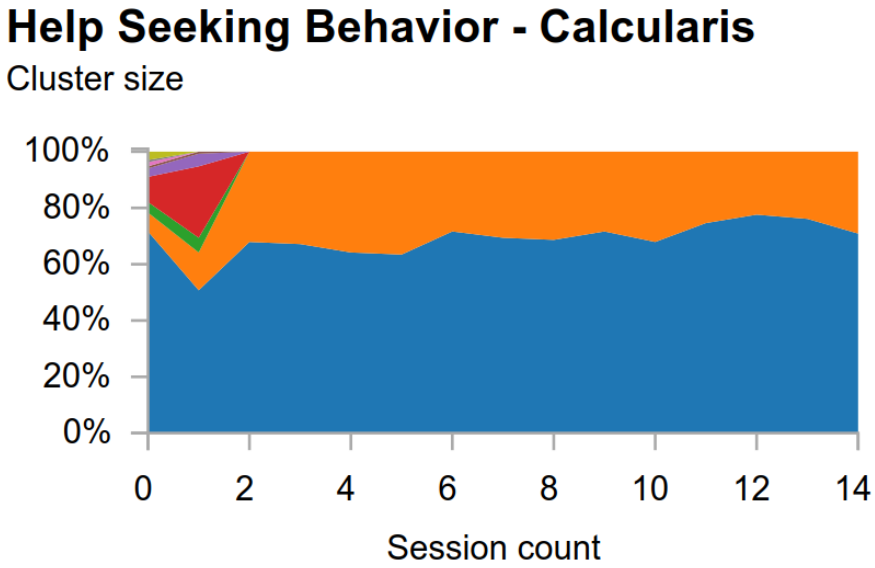




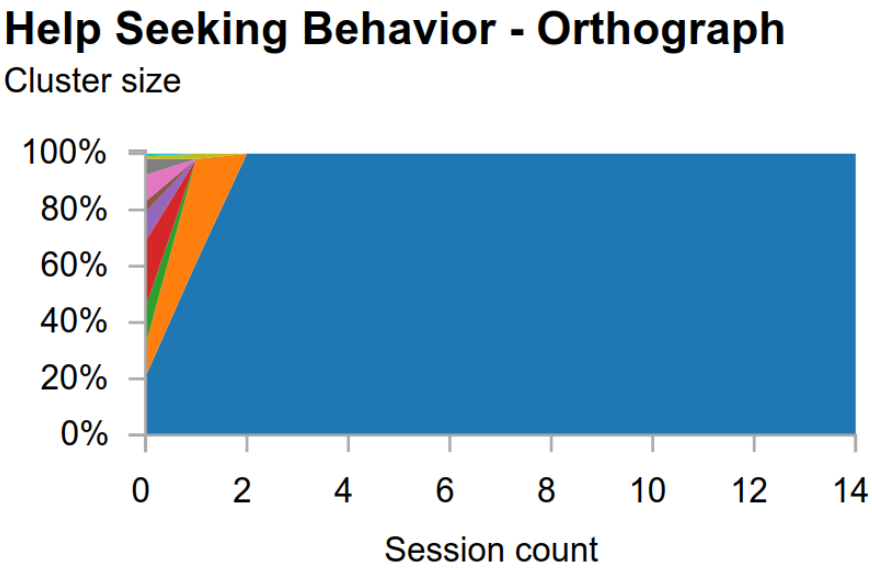
Diversity disappears

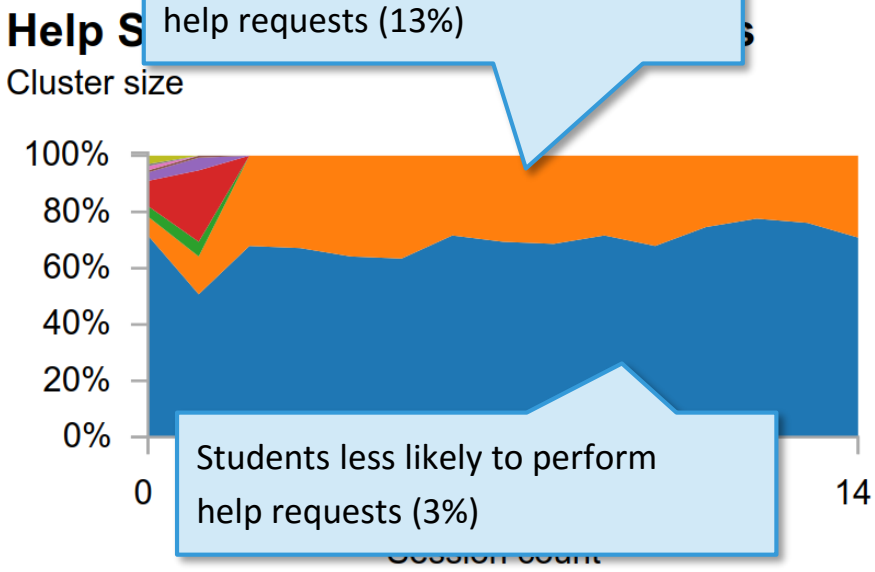
Help Seeking Behavior





Help Seeking Behavior





Students more likely to perform help requests (13%)

Students less likely to perform help requests (3%)

Conclusion

Enforcing temporal coherence beneficial for

- detection of student behavior
- stable detection of cluster events

Exploratory analysis demonstrated

- Reveal interesting properties about student behavior
- Pipeline can be used as a black box for any ITS

Thank you.

References

[Bergner et al., 2014] Y. Bergner, Z. Shu, and A. A. Von Davier. Visualization and Confirmatory Clustering of Sequence Data from a Simulation-Based Assessment Task. In Proc. EDM, 2014.

[Köck & Paramythis, 2011] M. Köck and A. Paramythis. Activity sequence modelling and dynamic clustering for personalized e-learning. UMUAI, 2011.

[Desmarais & Lemieux, 2013] M. Desmarais and F. Lemieux. Clustering and visualizing study state sequences. In Proc. EDM, 2013.

[Xu et al., 2014] K. S. Xu, M. Kliger, and A. O. Hero Iii. Adaptive evolutionary clustering. Data Mining and Knowledge Discovery, 2014.

[Peckham& McCall, 2012] T. Peckham and G. McCalla. Mining Student Behavior Patterns in Reading Comprehension Tasks. In Proc. EDM, 2012.

[Perera et al., 2009] D. Perera, J. Kay, I. Koprinska, K. Yacef, and O. R. Zaıane. Clustering and sequential pattern mining of online collaborative learning data. TKDE, 2009.

[Martinez-Maldonado et al., 2013] R. Martinez-Maldonado, K. Yacef, and J. Kay. Data mining in the classroom: Discovering groups' strategies at a multi-tabletop environment. In Proc. EDM, 2013.

[Herold et al., 2014] J. Herold, A. Zundel, and T. F. Stahovich. Mining meaningful patterns from students' handwritten coursework. In Proc. EDM, 2013.

[Krikpatrick, 2000] M. Kirkpatrick. Patterns of quantitative genetic variation in multiple dimensions. Genetica, 2006.

[Burnham & Anderson, 2002] Burnham, K. P.; Anderson, D. R. (2002), Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed.), Springer-Verlag

[Pelleg & Moore, 2000] D. Pelleg and A. Moore. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In Proc. ICML, 2000.

[Kinnebrew et al., 2013] J. S. Kinnebrew, D. L. Mack, and G. Biswas. Mining temporally-interesting learning behavior patterns. In Proc. EDM, 2013.

Discussion

- ✓ Selected features
  - cover broad range of characteristics
  - in accordance with the literature on DD
- ✓ High sensitivity (0.91) and specificity (0.91)
- ✓ Good construct validity
- ✓ Reliability

approximation only

Parameter influence

Analysis of parameter effects

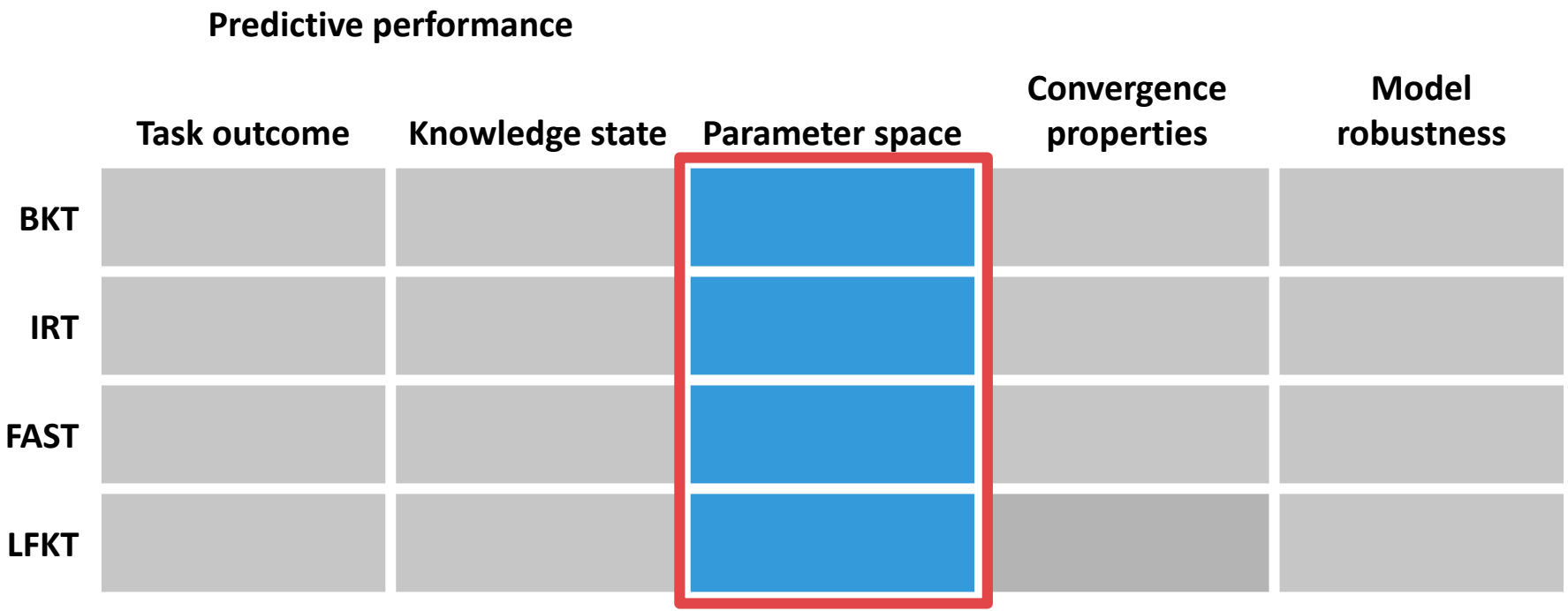
Study effect on performance using linear regression

Variables

All 6 model parameters

additionally: ***correct ratio*** and ***average number of tasks***

Parameter space



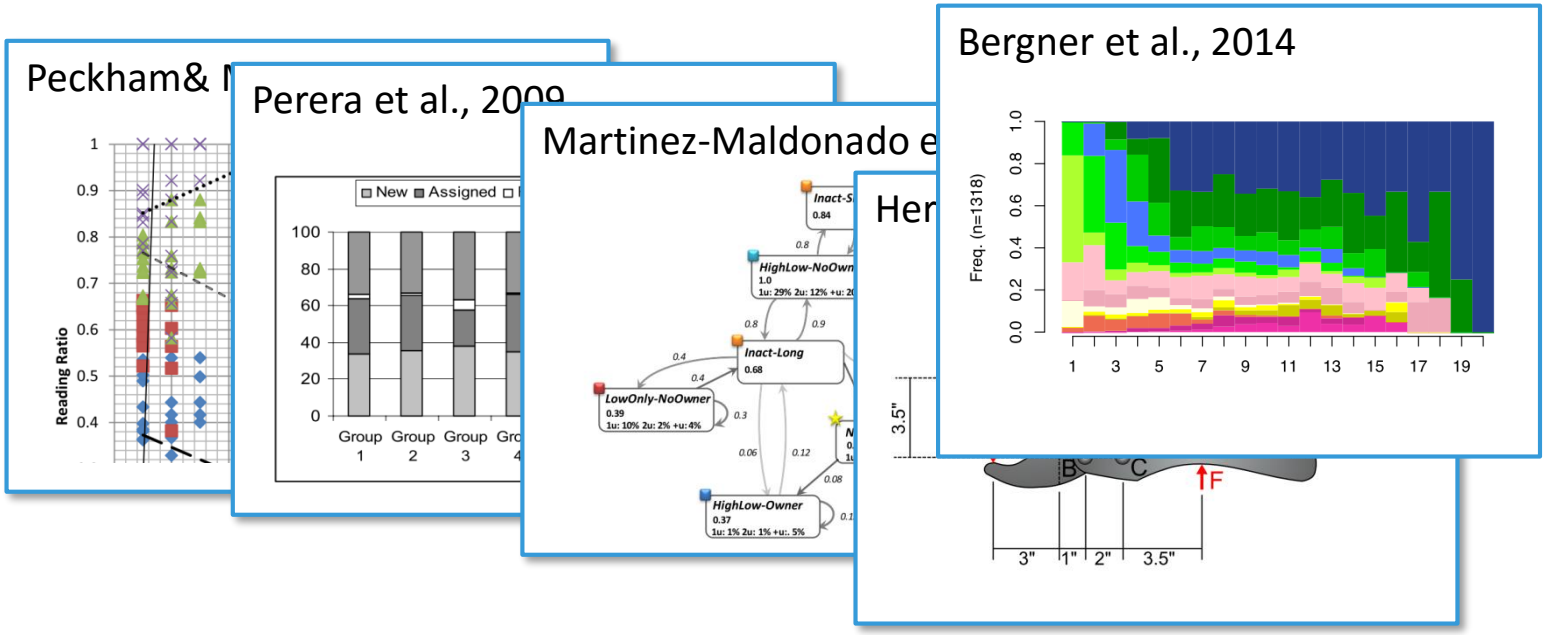
Experimental setup

Parameters configuration designed to match real world conditions

Parameter	Configurations	Reference
BKT parameters	Sampled around clusters	[Ritter et al. 2009]
Student abilities	N(0,1)	[Harris 1989]
Range of item difficulties	[0,3]	[Harris 1989]
Feature weights	[0,1.5]	According to item difficulties

Clustering in EDM

Clustering sequential data to detect behavior patterns



AFFECT clustering [Xu et al.,2014]

Optimal α alpha as a trade-off:

estimated noise

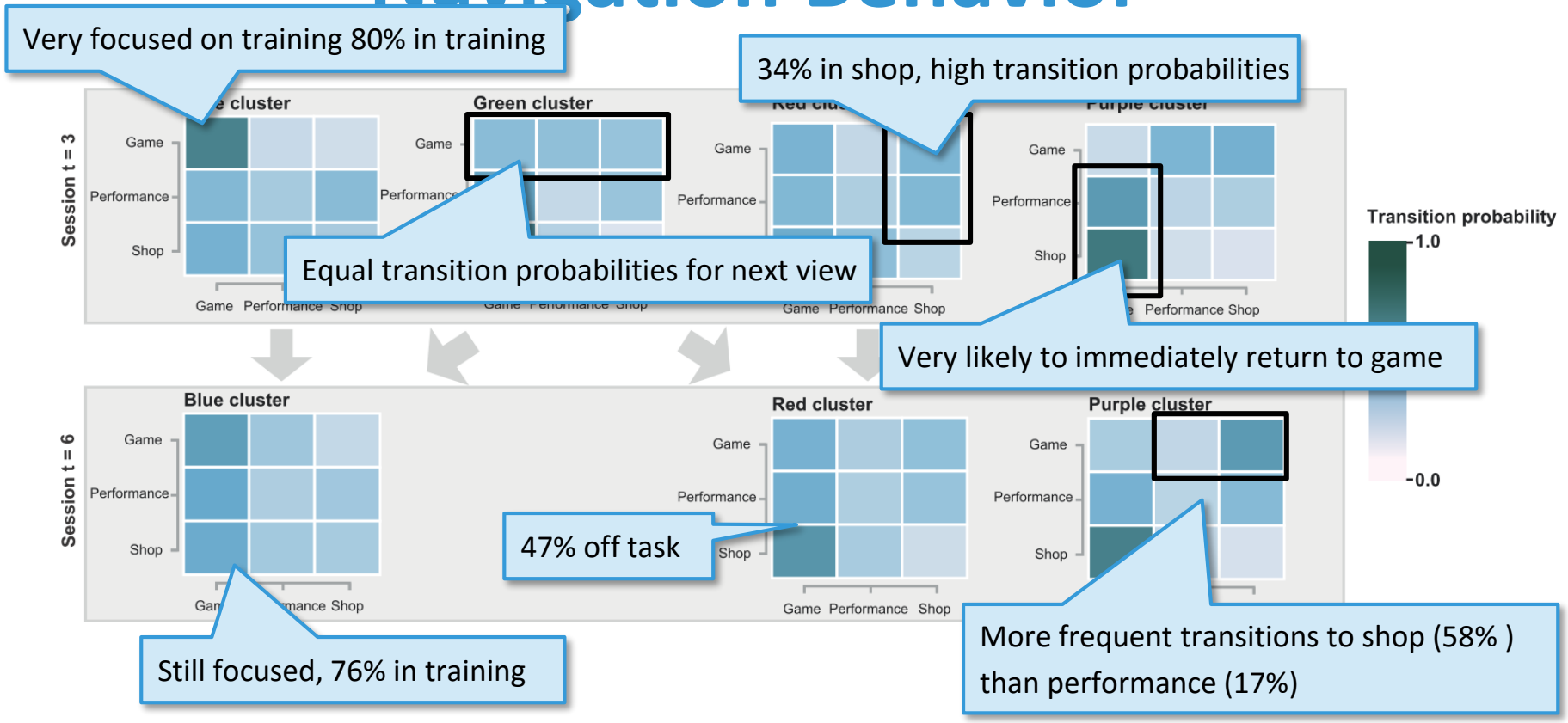
$$\alpha^t = \frac{\sum_i \sum_j \mathrm{var}(n_{ij}^t)}{\sum_i \sum_j (\hat{\psi}_{ij}^{t-1} - \psi_{ij}^t)^2 + \mathrm{var}(n_{ij}^t)}$$

Amount of new information

[Xu,2014] propose estimate

Problem: $\mathrm{var}(\mathrm{n}_{ij})$ and $\psi^t{}_{ij}$ are unknown

Navigation Behavior



Limitations

Similarity computations and clustering is O(n^2)

Investigation of different clustering algorithms

More analysis tools (such as integrated PCA)