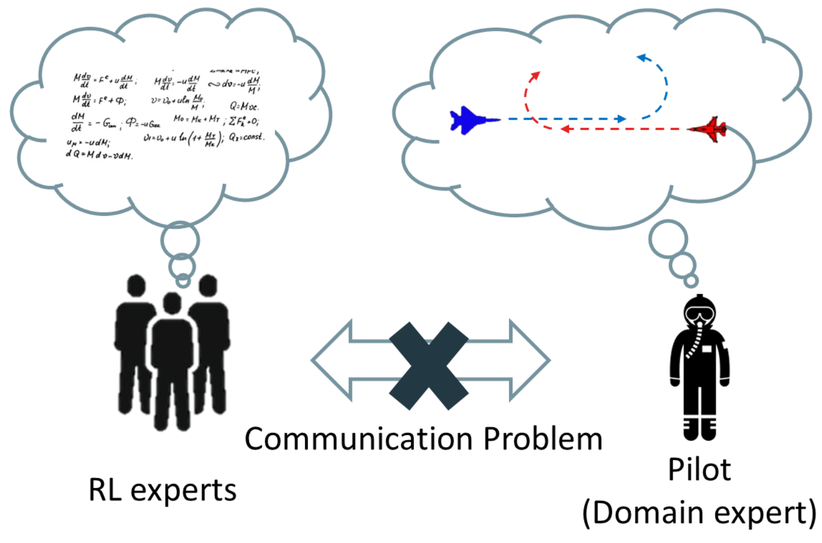
State Trajectory Abstraction and Visualization Method for Explainability in Reinforcement Learning

Yoshiki Takagi, Roderick Tabalba, Jason Leigh
Laboratory for Advanced Visualization & Applications, University of Hawaii at Manoa

Motivation

- Create a visual representation of RL agents' behavior understandable to domain experts without ML expertise

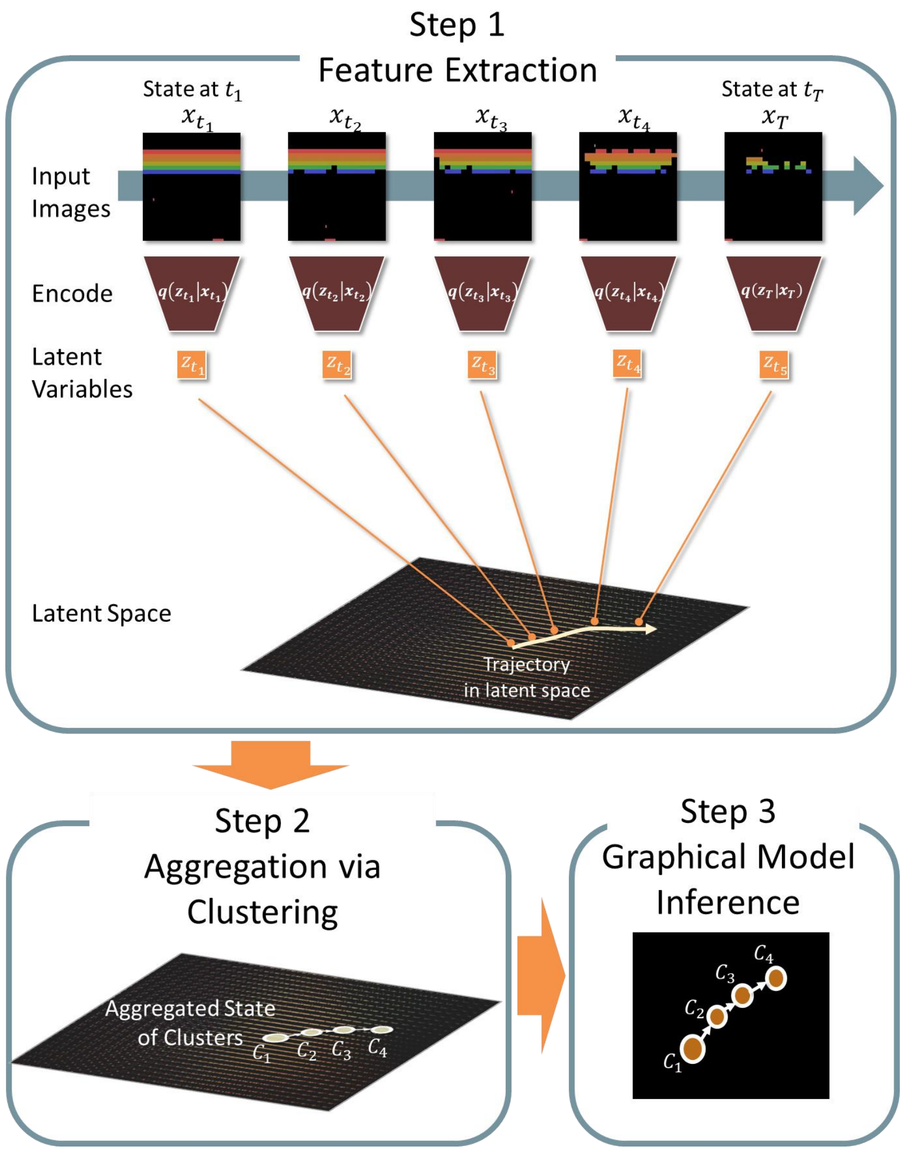


Approach

- Use replay to provide non descriptive knowledge about an agent
- Abstract & visualize it as a trajectory

Methodology

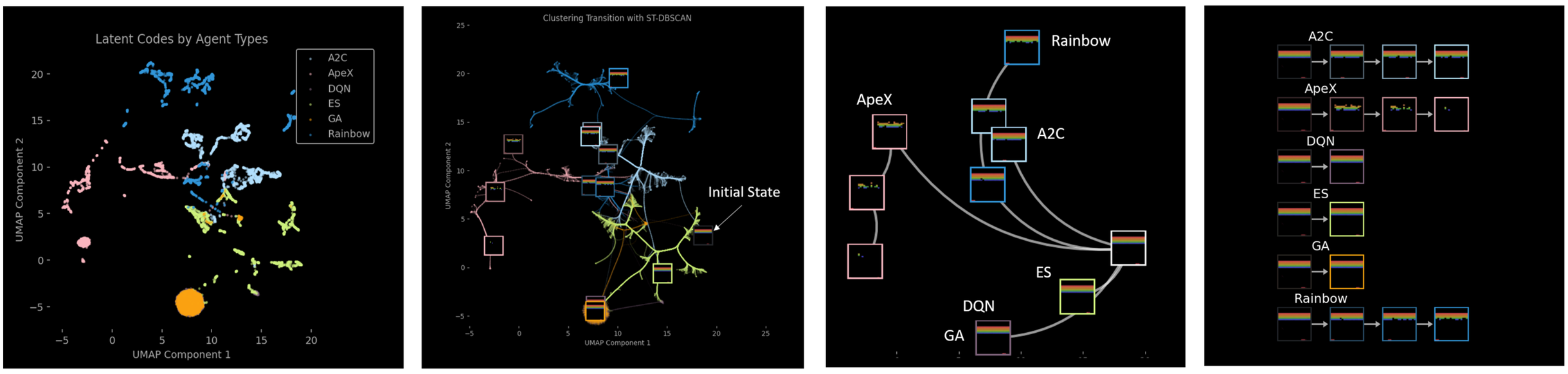
- Used β-VAE & ST-DBSCAN
- Tested to various Atari applications



Result

- Extracted trajectories of several pre-trained agents with β-VAE
- Explored visualization ideas of abstracted trajectories

Raw trajectory → **Clustering & Edge bundling** → **Abstracted trajectory #1** → **Abstracted trajectory #2**



User Study Plan

- Evaluate how well a user's mental model obtained from the proposed abstract trajectory agrees with agents' complete trajectory

Evaluation

- ✓ Accuracy
- ✓ Response time
- ✓ Confidence
- ✓ Preference