Integrated Planning and Reinforcement Learning for Compositional Domains

—

Harsha Kokel
Research Scientist
IBM Research

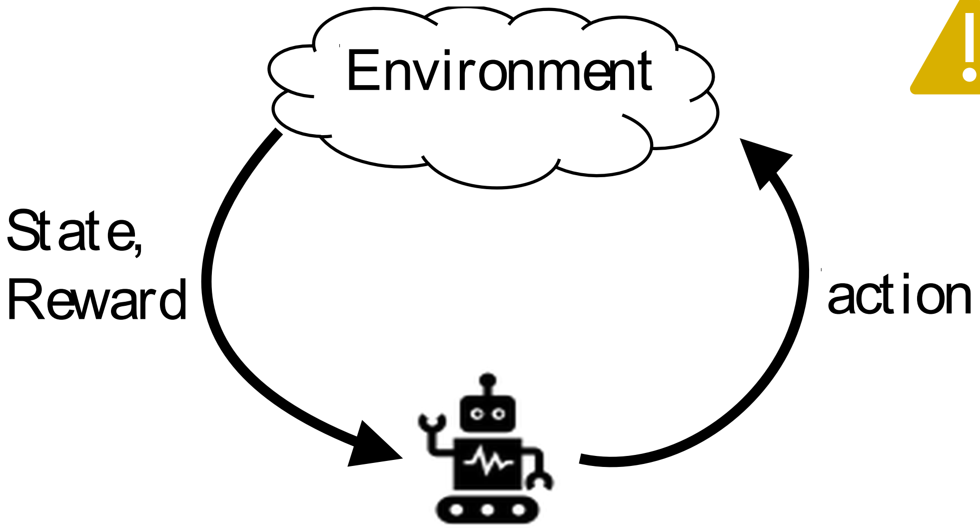
Integrated Planning and Reinforcement Learning for Compositional Domains

—

Harsha Kokel
Research Scientist
IBM Research

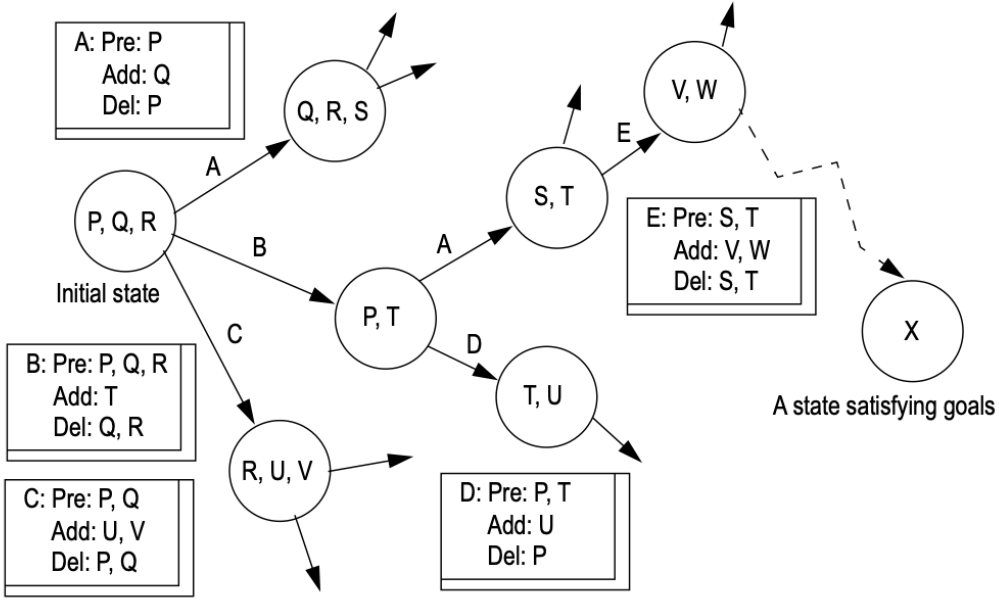
Prof. Sriraam Natarajan (left)

Reinforcement Learning



Relational MDP

Planning



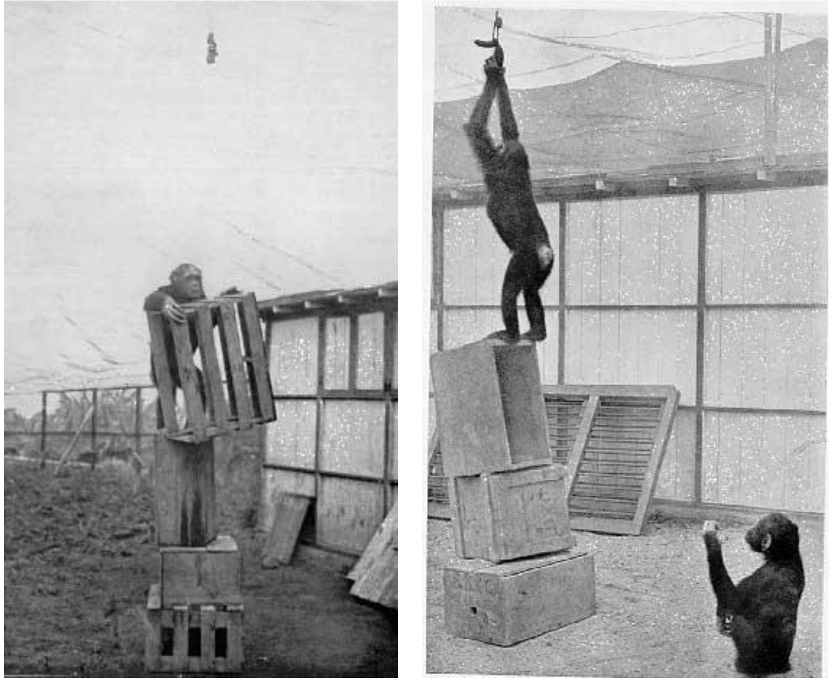
Long and Fox (2002)

Planning

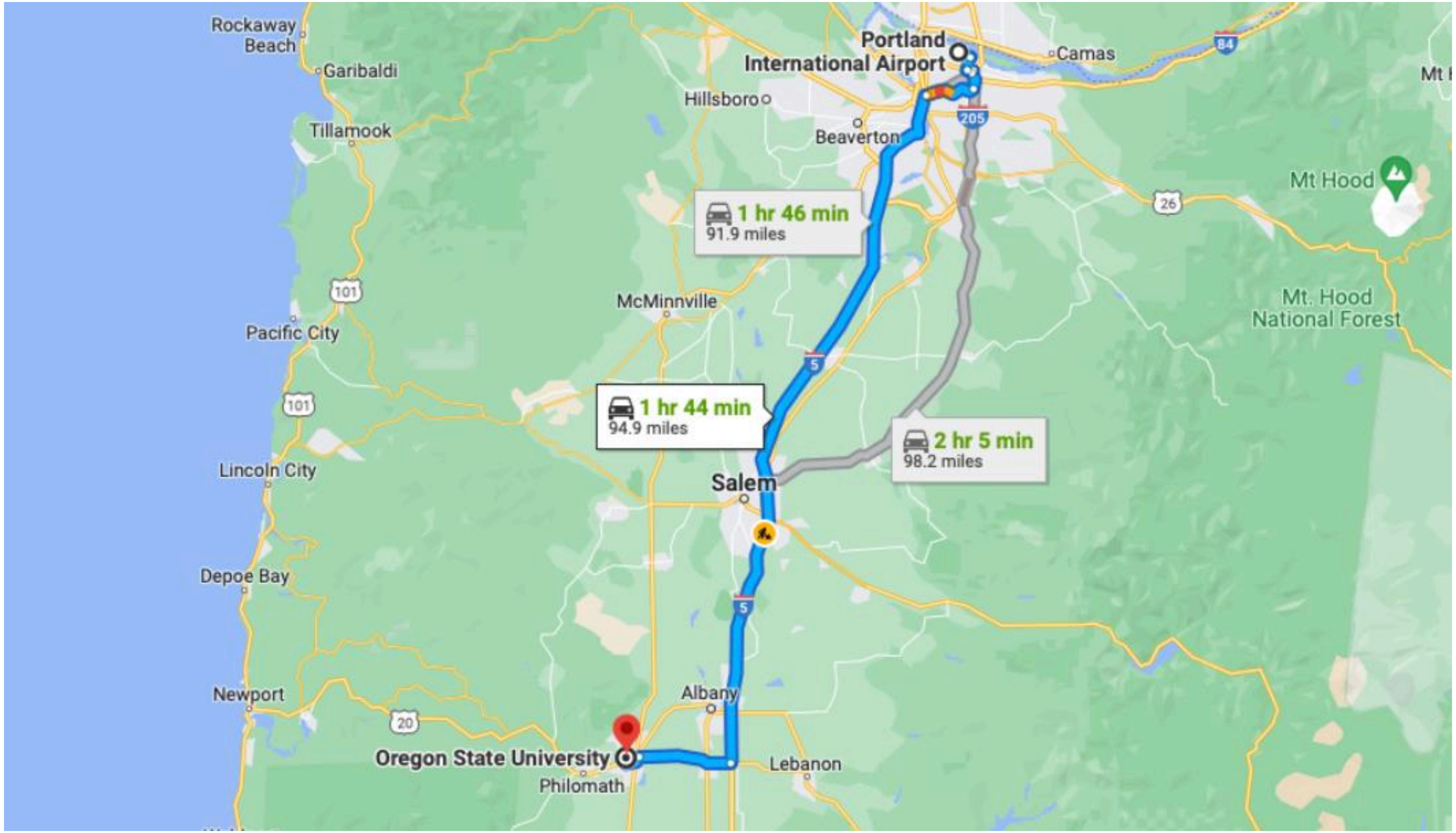
- Search through space of states
- Relies on an explicit model of the environment

RL

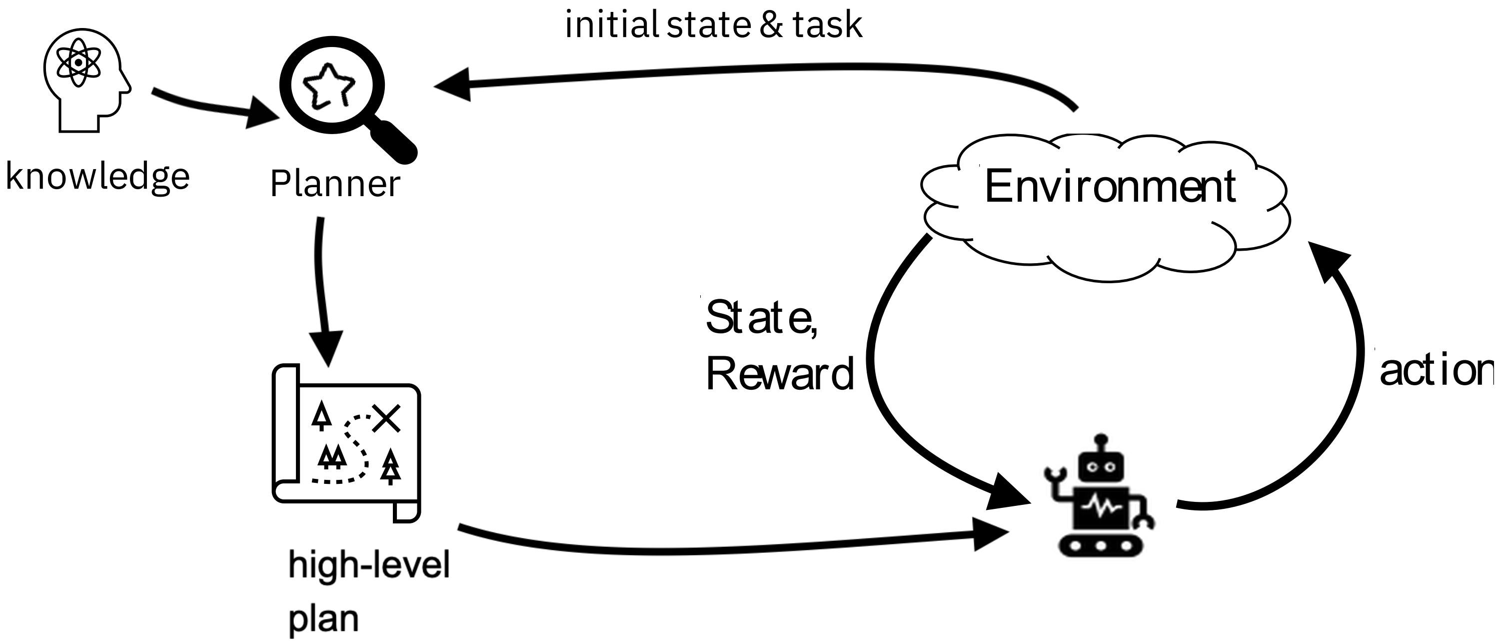
- Search through space of policies
- Relies on trial & error by interaction



Köhler (1948)


Rockaway Beach
Garibaldi
Tillamook
Pacific City
Lincoln City
Depoe Bay
Newport
Portland International Airport
Hillsboro
Beaverton
Camas
Mt Hood
Mt. Hood National Forest
1 hr 46 min
91.9 miles
McMinnville
1 hr 44 min
94.9 miles
2 hr 5 min
98.2 miles
Salem
Albany
Lebanon
Oregon State University
Philomath

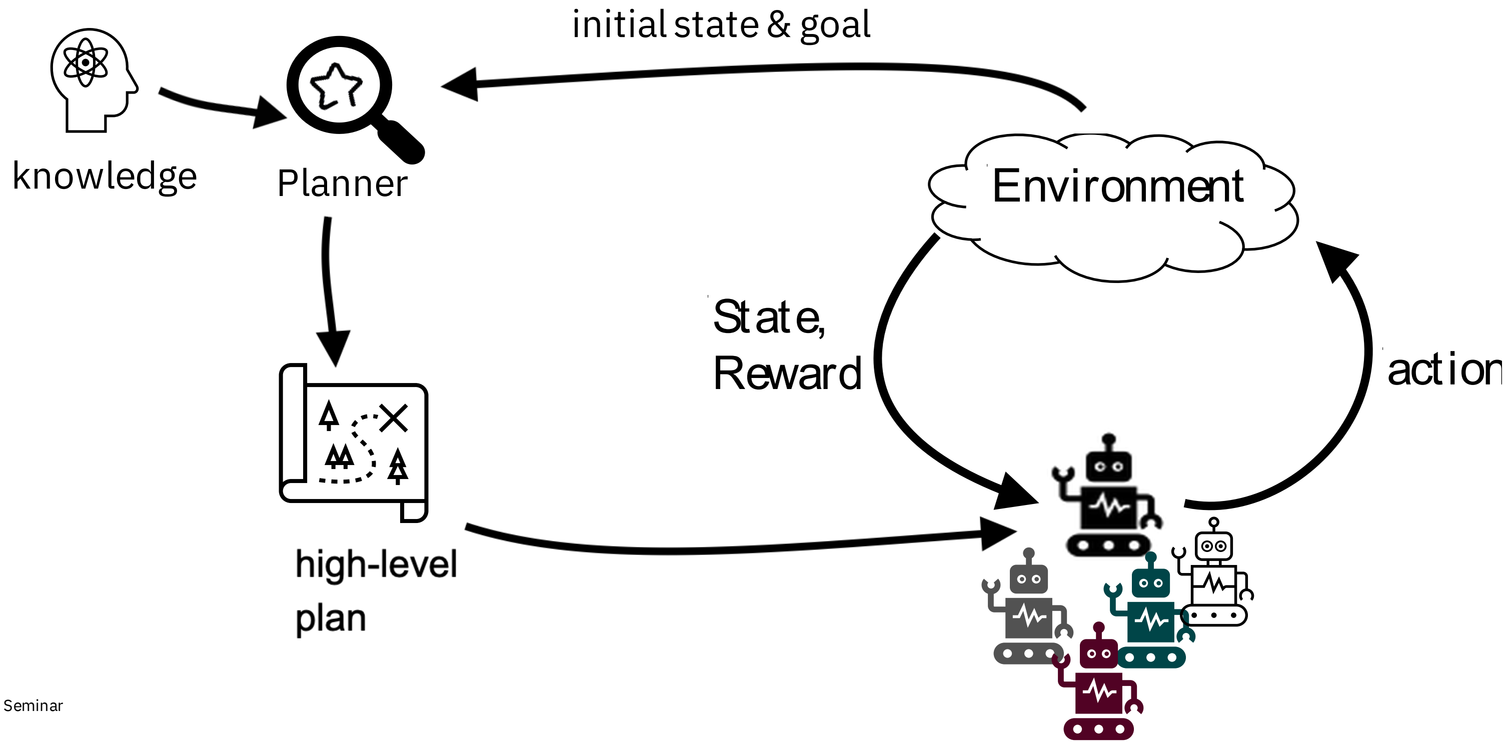
Integrating Planning and RL




EXIT 288
5
NORTH
Portland
205
NORTH
Oregon City
EXIT
ONLY
ROAD WORK AHEAD


LEFT TURN YIELD ON FLASHING ARROW
SPEED LIMIT 35

Integrating Planning and RL



Decomposing GRMDP

Goal-directed RMDP:

$\langle S, A, P, R, \gamma, G \rangle$

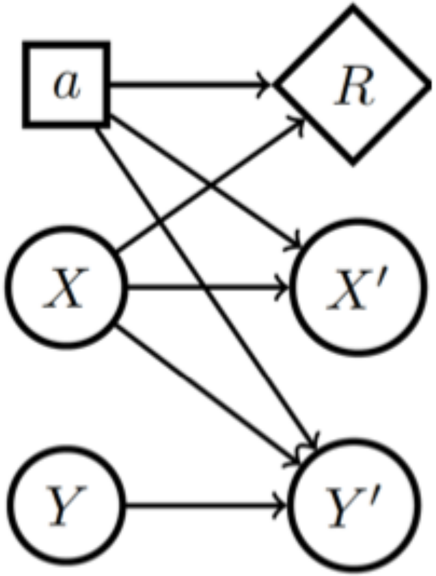
$\langle S, A, P_1, R_1, \gamma \rangle$ $\langle S, A, P_2, R_2, \gamma \rangle$ ... $\langle S, A, P_n, R_n, \gamma \rangle$

Multiple subgoal RMDP:

$$R_o(s,a,s') = \begin{cases} t_R + R(s,a,s') & \textit{if } s' \in \beta(o) \textit{ and } s \notin \beta(o) \\ 0 & \textit{if } s' \in \beta(o) \textit{ and } s \in \beta(o) \\ R(s,a,s') & \textit{otherwise} \end{cases}$$

$$P_o(s,a,s') = \begin{cases} 0 & \textit{if } s \in \beta(o) \textit{ and } s' \notin \beta(o) \\ 1 & \textit{if } s \in \beta(o) \textit{ and } s' \in \beta(o) \\ P\,(s,a,s') & \textit{otherwise} \end{cases}$$

Irrelevant variables



Factored MDP represented as Dynamic Bayesian Network (DBN)

Model-agnostic Abstraction

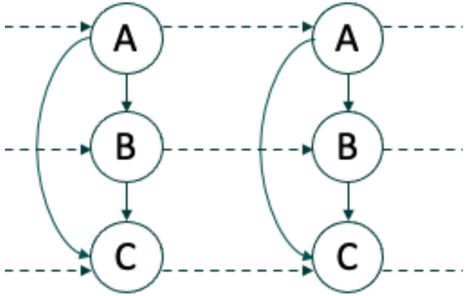
A **model-agnostic abstraction** $\phi(s)$ is such that for any action a and an abstract state $\bar{s}$, $\phi(s_1)$=$\phi(s_2)$ if and only if

$$\sum_{\{s_1' \mid \phi(s_1')=\bar{s}\}} R_o(s_1, a, s_1') = \sum_{\{s_2' \mid \phi(s_2')=\bar{s}\}} R_o(s_2, a, s_2')$$

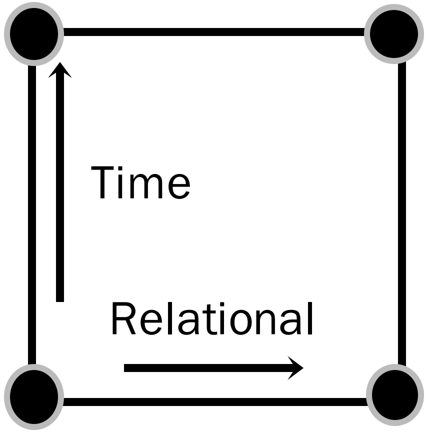
$$\sum_{\{s_1' \mid \phi(s_1')=\bar{s}\}} P_o(s_1, a, s_1') = \sum_{\{s_2' \mid \phi(s_2')=\bar{s}\}} P_o(s_2, a, s_2')$$

Dietterich NeurIPS 2000; Ravindran and Barto IJCAI 2003;
Givan, Dean, and Greig AI 2003; Li, Walsh, and Littman ISAIM 2006

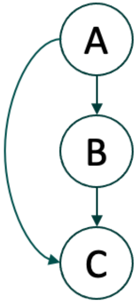
Graphical models



Dynamic BN

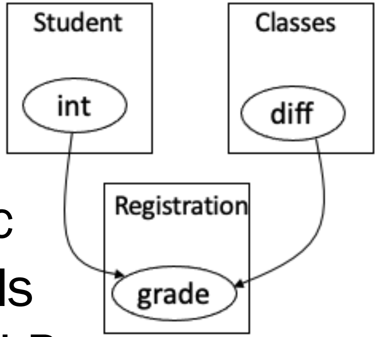


Dynamic PLM



Bayesian Network (BN)

Probabilistic Logic Models (PLM, PRM, BLP, LBN, DAPER)



Koller and Friedman 2009;
Getoor and tasker 2007; Raedt et al. 2016

D-FOCI

First Order Conditional Influence (FOCI) statements

$$\texttt{if}\ \langle condition \rangle \quad \texttt{then}\ \langle influents \rangle \quad \texttt{QINF}\ \langle resultant \rangle$$

Dynamic FOCI statements

$$[\langle subgoal \rangle]\text{:}\ \langle influents \rangle \ \xrightarrow{[+1]}\ \langle resultant \rangle$$

Natarajan, Tadepalli, Dietterich, and Fern 2008

D-FOCI

$$\{\texttt{action},\texttt{taxi\_at}(X)\} \xrightarrow{+1} \texttt{taxi\_at}(X) \quad (3a)$$

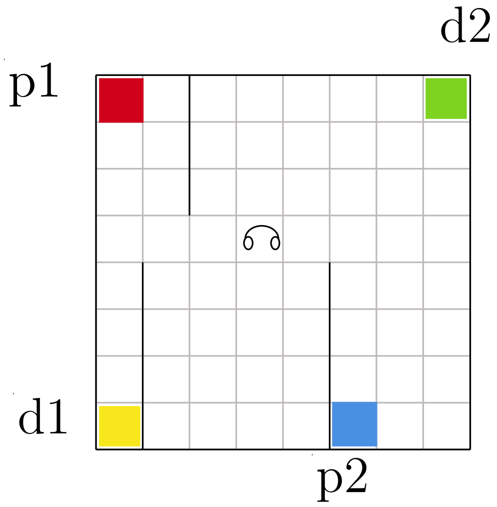
$$\texttt{pick}(P) : \{\texttt{action}, \texttt{taxi\_at}(X), \texttt{at}(P, Y), \texttt{in\_taxi}(P)\} \xrightarrow{+1} \texttt{in\_taxi}(P) \quad (3b)$$

$$\texttt{pick}(P) : \{\texttt{in\_taxi}(P)\} \longrightarrow Reward \quad (3c)$$

$$\texttt{drop}(P) : \{\texttt{at\_dest}(P)\} \longrightarrow Reward \quad (3d)$$

$$\texttt{drop}(P) : \{\texttt{at}(P, X), \texttt{dest}(P, D), \texttt{at\_dest}(P)\} \longrightarrow \texttt{at\_dest}(P) \quad (3e)$$

$$\texttt{drop}(P) : \{\texttt{action}, \texttt{taxi\_at}(X), \texttt{at}(P, Y), \texttt{in\_taxi}(P)\} \xrightarrow{+1} \texttt{at}(P, K) \quad (3f)$$



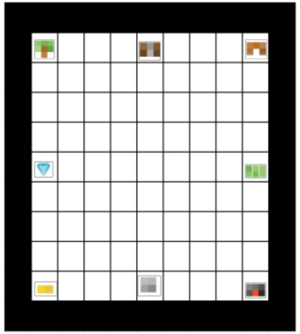
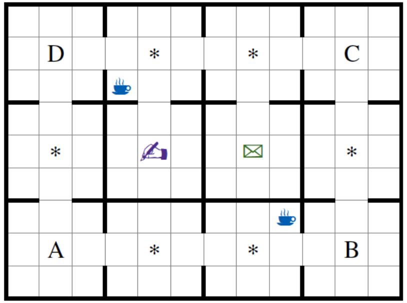
Experiments

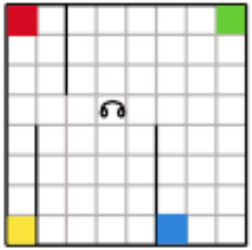
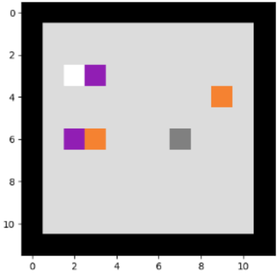
Domains

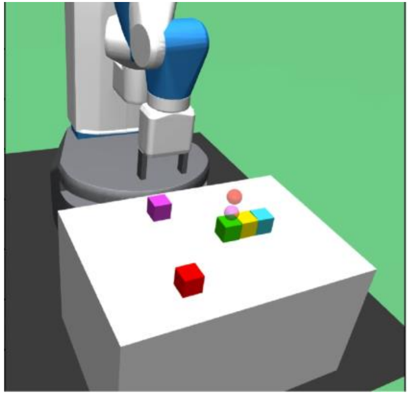
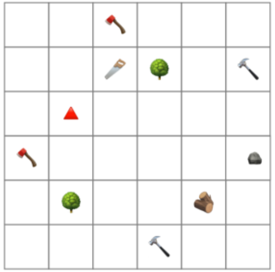
- Office World
- Minecraft World
- Relational Taxi
- Relational Box World
- Craft World
- Robotic Fetch domain

Baselines

- HRL (options framework)
- Tabular Q-learning
- Deep RL (DDQN, HDQN, SAC)
- Deep Relational RL
- Planning+RL (Taskable RL)

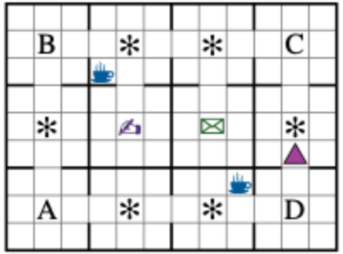



Experiments

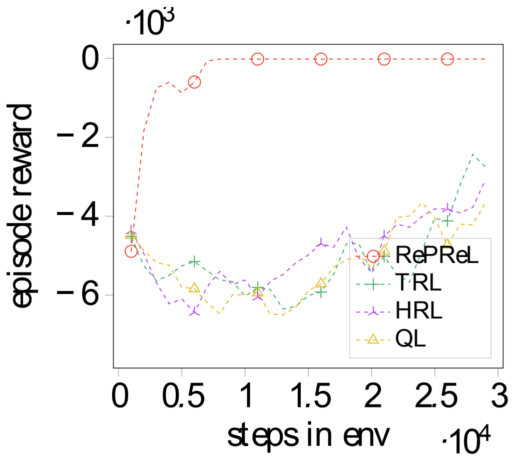
Office World

Sample efficiency

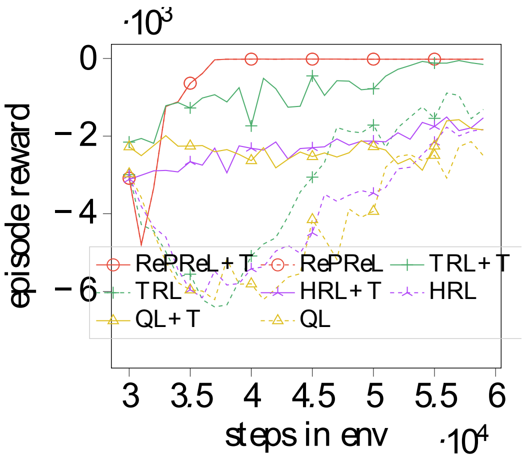
Transfer across task



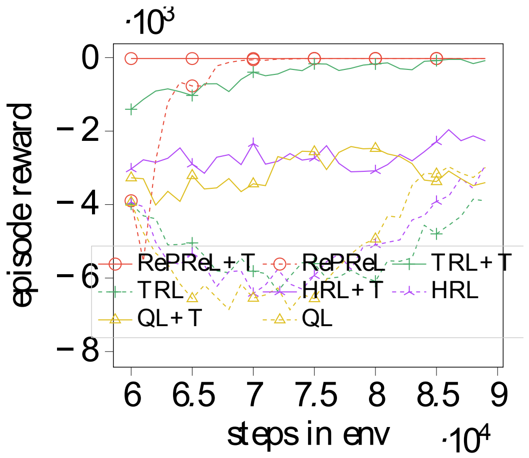
Symbol	Meaning
▲	Agent
✻	Furniture
☕	Coffee machine
✉	Mail room
✍	Office
A, B, C, D	Marked locations



Deliver mail



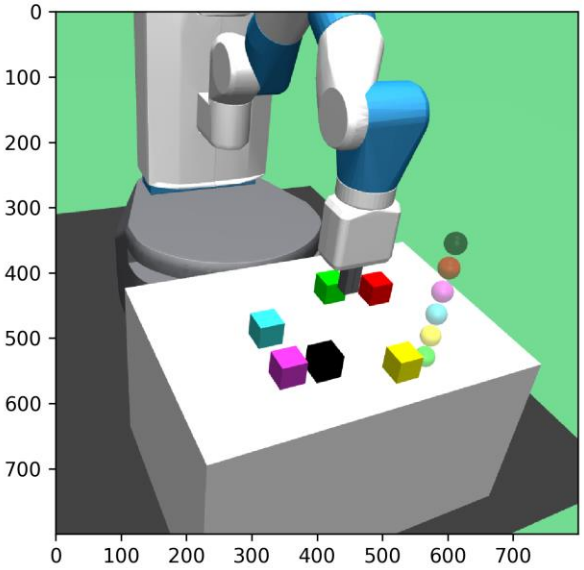
Deliver coffee

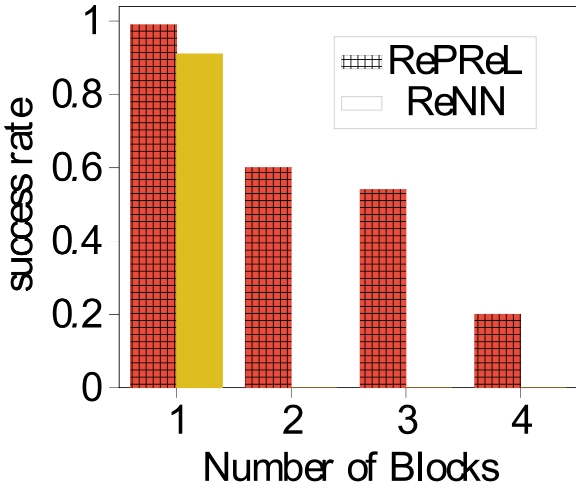


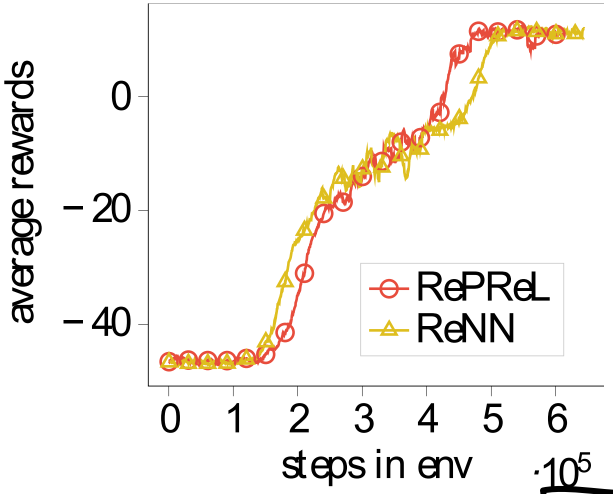
Deliver mail and coffee

Kokel et al. ICAPS 2021

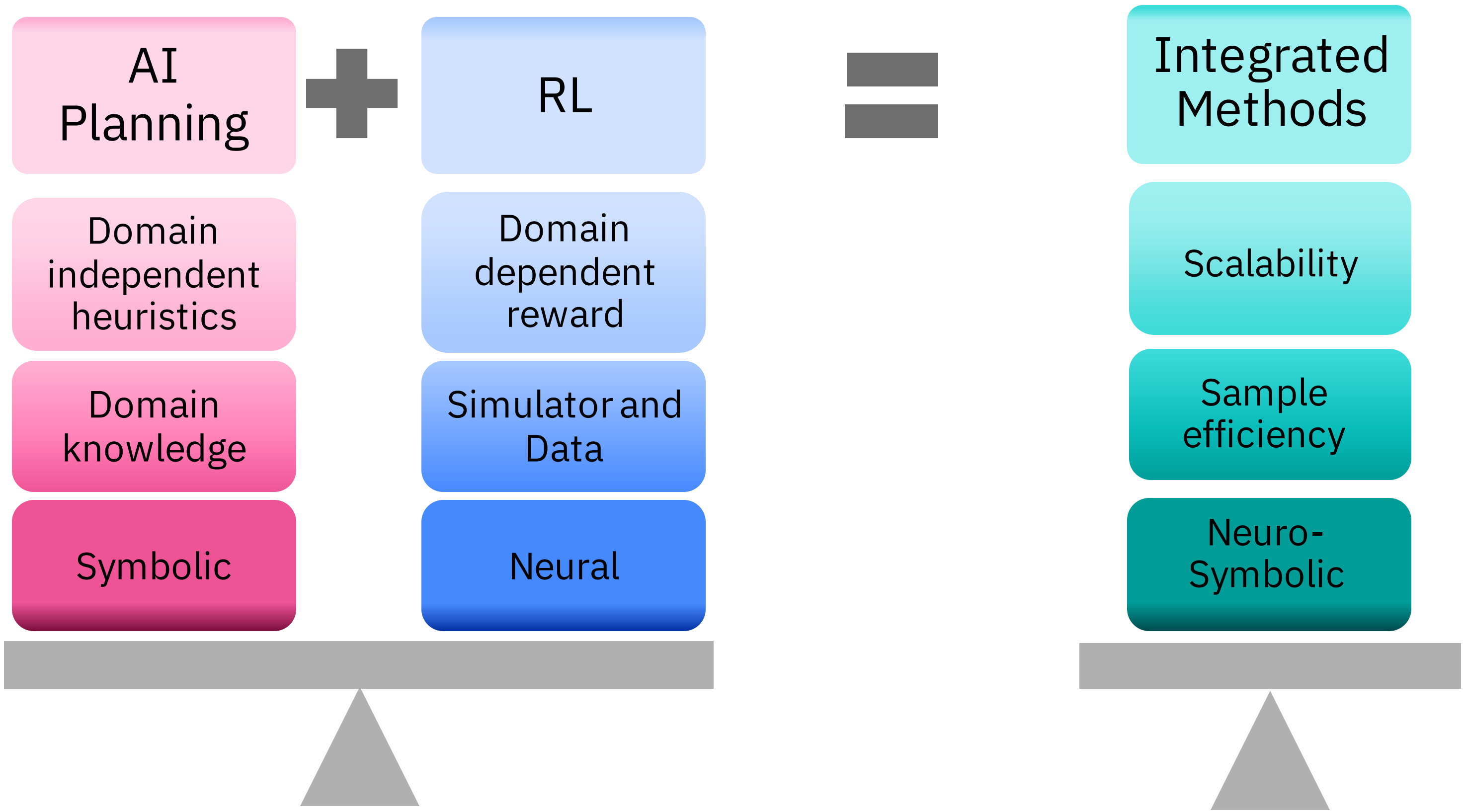
Deep Relational RL







[1]Kokel et al. NCAA 2022a
[2]Li et al. ICRA 2022


AI Planning
Domain independent heuristics
Domain knowledge
Symbolic
RL
Domain dependent reward
Simulator and Data
Neural
Integrated Methods
Scalability
Sample efficiency
Neuro-Symbolic

Bridging the Gap
Planning & RL

Planning

- Search through space of states
- Relies on an explicit model of the environment
- PDDL Task

RL

- Search through space of policies
- Relies on trial & error by interaction
- MDP

PDDL Task

$\langle \mathcal{L}, \mathcal{O}, I, G \rangle$

Lifted Action Models $\mathcal{O}$

```
(:action move
:parameters (?curpos ?nextpos ?dir)
:precondition (and (place ?curpos)
   (place ?nextpos) (at-robot ?curpos)
   (conn ?curpos ?nextpos ?dir) (open ?nextpos))
:effect (and (at-robot ?nextpos)
          (not (at-robot ?curpos))))
```

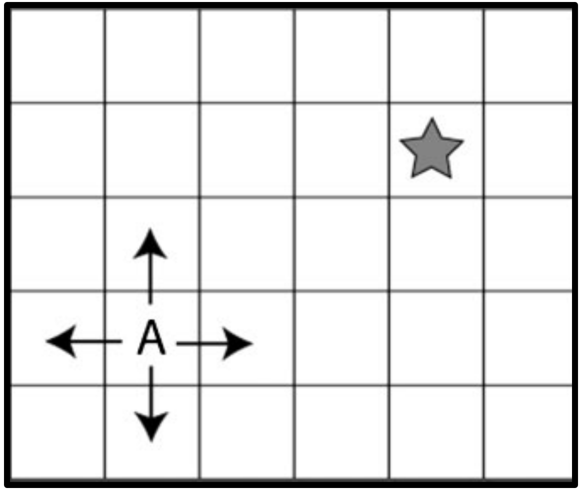
move(c_1_1, c_1_2, east),
move(c_2_2, c_2_1, west), ...

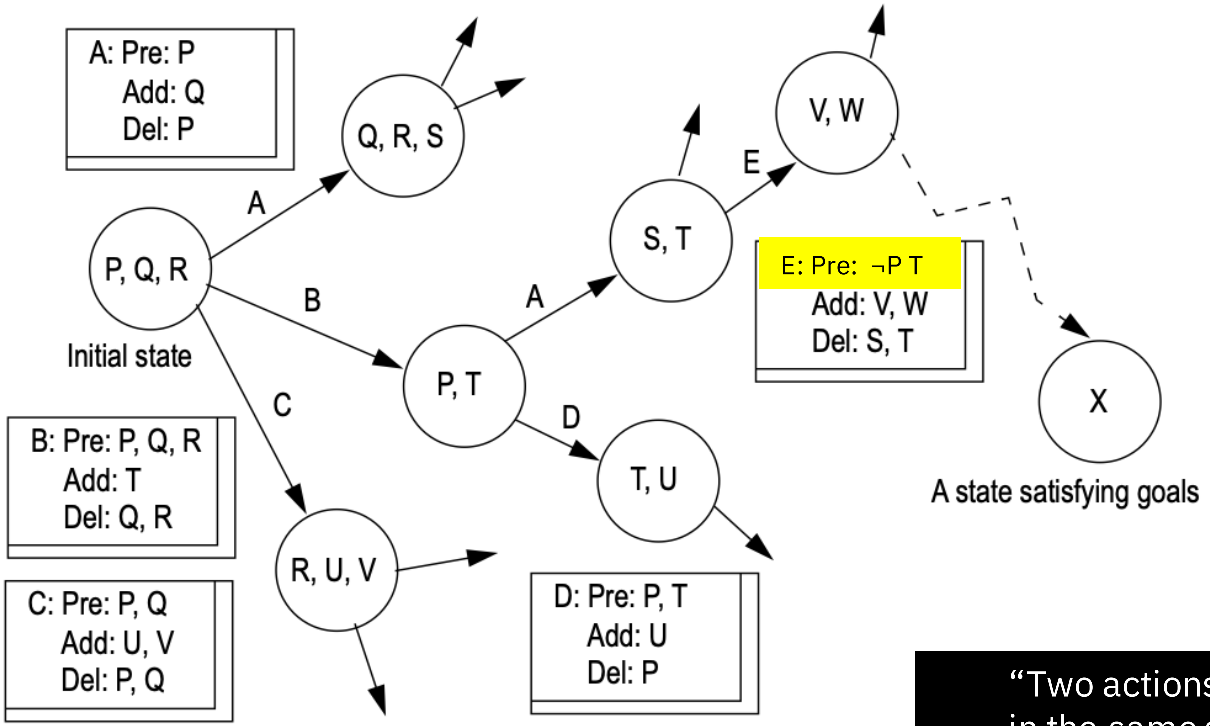
MDP

$\langle S, A, T, R, \gamma \rangle$

Actions

[east, west, north, south]





Long and Fox (2002)

"Two actions that are applicable in the same state cannot have the same label"

(**:action** pickup

:parameters (?k - key ?r - room)

:precondition (**and** (at ?k ?r)

(at-agent ?r)

(empty-hand))

:effect (**and** (**not** (at ?k ?r))

(**not** (empty-hand))

(carry ?k))

)

# of grounding = #of keys * # of rooms

Are all the parameters of LAM relevant?*

*Do they define different grounded actions that can be applied in a single state?

Relevant parameters

Know,

(at key1 room1) $\oplus$ (at key1 room2)

So,

(pickup key1 room1) $\oplus$

(pickup key1 room2)

$\oplus$: Mutually exclusive

(**:action** pickup

:parameters (?k - key ?r - room)

:precondition (**and** (at ?k ?r)

(at-agent ?r)

(empty-hand))

:effect (**and** (**not** (at ?k ?r))

(**not** (empty-hand))

(carry ?k))

)

Applicable Action Mutex Group (AAMG)

(pickup key1 room1),
(pickup key1 room2),
(pickup key1 room3),
.
.
.

(pickup key2 room1),
(pickup key2 room2),
(pickup key2 room3),
.
.
.

(pickup ?k - key ?r - room)

(pickup ?k – key ␣)

(pickup key1 ␣)

(pickup key2 ␣)

Parameter Seed Set of pickup
{?k – key }

Action Space Reduction

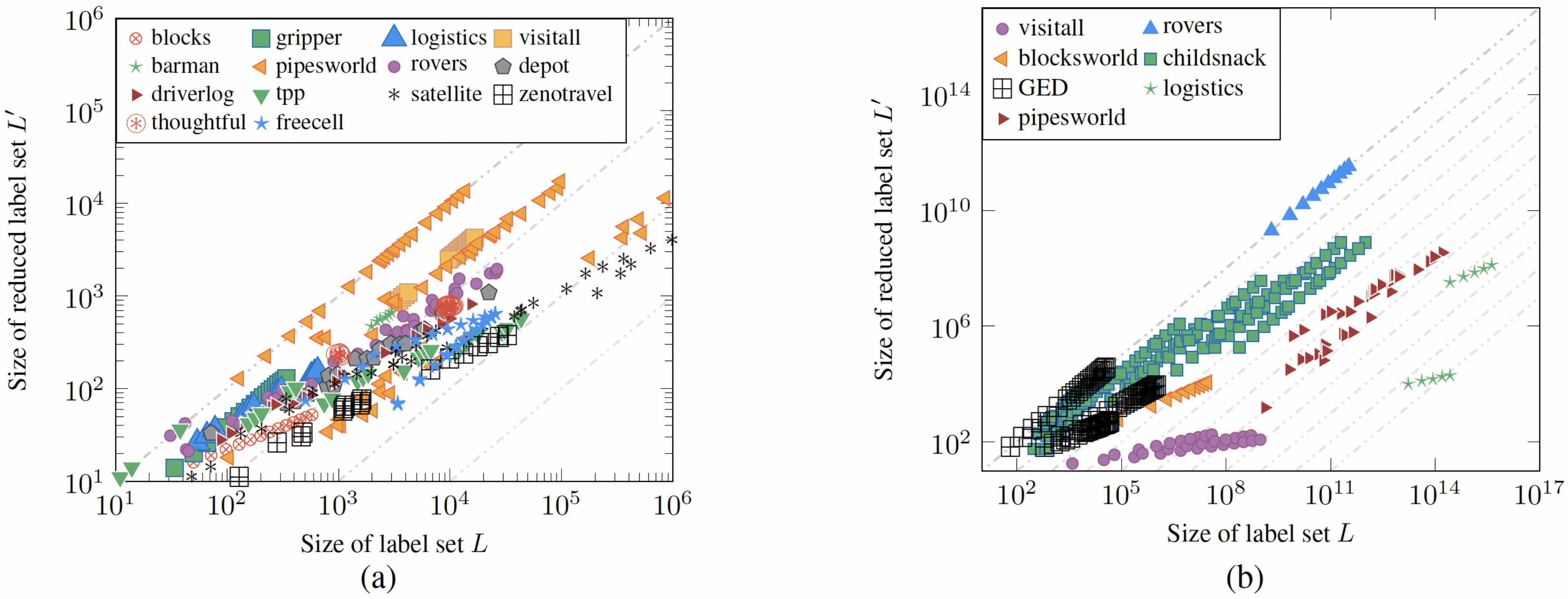


Figure 2: Comparison of label set sizes on (a) 14 IPC STRIPS domains and (b) 7 HTG domains.

Impact on learning RL policies

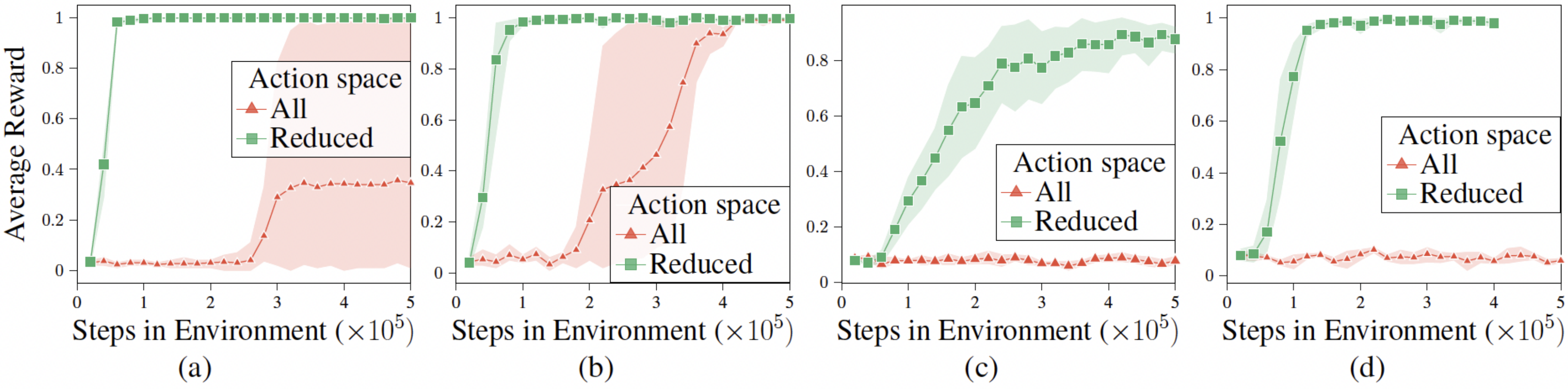
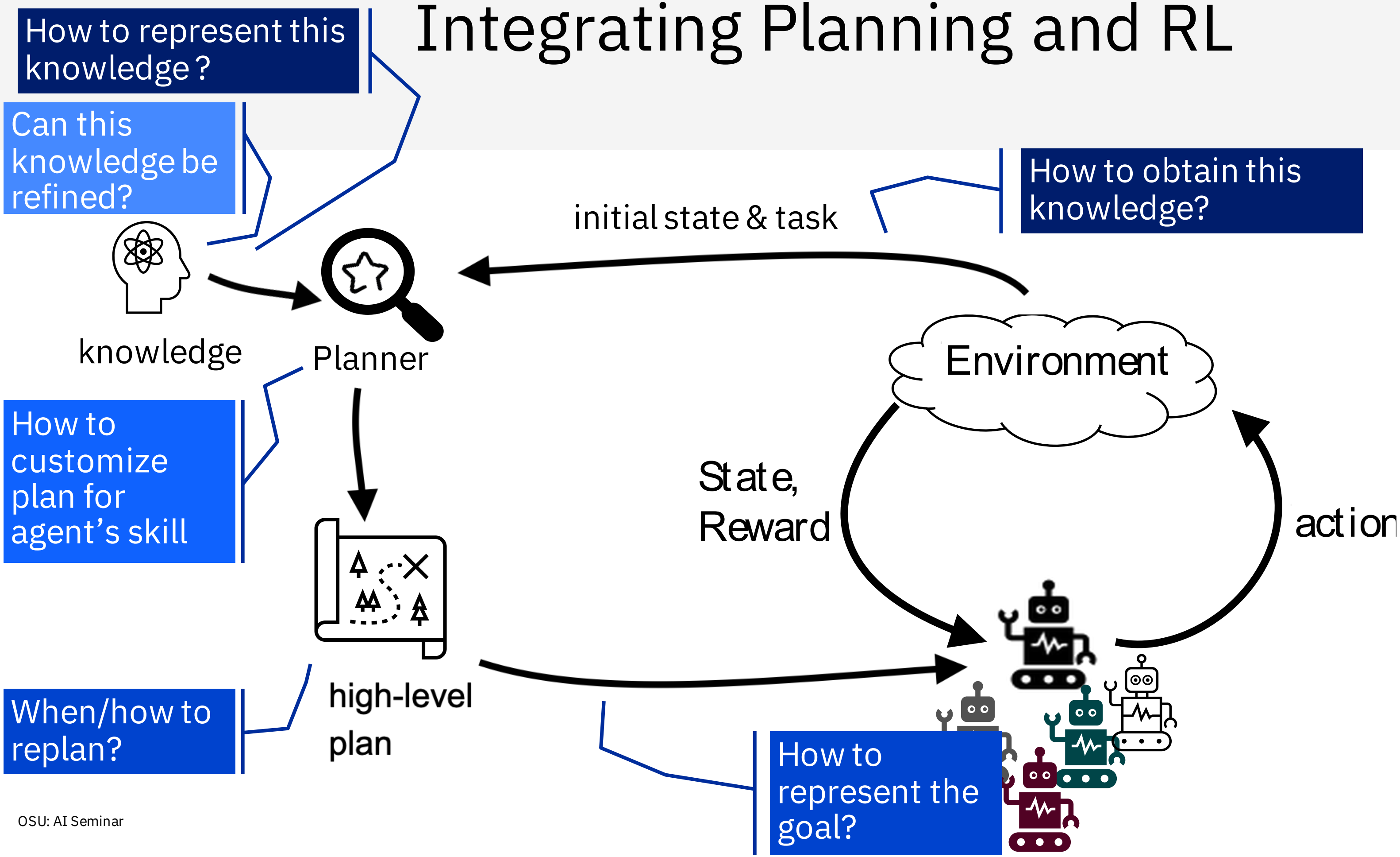


Figure 3: Learning curve in the (a) ferry, (b) gripper, (c) blocks, and (d) logistics; with and without action label reduction.

Integrating Planning and RL



Reference

Harsha Kokel, Arjun Manoharan, Sriraam Natarajan, Balaraman Ravindran, Prasad Tadepalli, *RePReL: Integrating Relational Planning and Reinforcement Learning for Effective Abstraction*, In **ICAPS 2021a.**

Harsha Kokel, Mayukh Das, Rakibul Islam, Julia Bonn, Jon Cai, Soham Dan, Anjali Narayan-Chen, Prashant Jayannavar, Janardhan Rao Doppa, Julia Hockenmaier, Sriraam Natarajan, Martha Palmer, Dan Roth, *Human-guided Collaborative Problem Solving: A Natural Language based Framework*, In **ICAPS (demo track) 2021b**.

Harsha Kokel, Sriraam Natarajan, Balaraman Ravindran, Prasad Tadepalli, RePReL: *A Unified Framework for Integrating Relational Planning and Reinforcement Learning for Effective Abstraction in Discrete and Continuous Domains*, In **NCAA 2022a**.

Harsha Kokel, Nikhilesh Prabhakar, Sriraam Natarajan, Balaraman Ravindran, Prasad Tadepalli, *Hybrid Deep RePReL: Integrating Relational Planning and Reinforcement Learning for Information Fusion*, In **FUSION 2022b**.

Harsha Kokel, Mayukh Das, Rakibul Islam, Julia Bonn, Jon Cai, Soham Dan, Anjali Narayan-Chen, Prashant Jayannavar, Janardhan Rao Doppa, Julia Hockenmaier, Sriraam Natarajan, Martha Palmer, Dan Roth, *Lara -- Human-guided collaborative problem solver: Effective integration of learning, reasoning and communication*, In **ACS 2022c**.

Harsha Kokel, Junkyu Lee, Michael Katz, Kavitha Srinivas, Shirin Sohrabi, Action Space Reduction for Planning Domains, IJCAI 2023

Sriraam Natarajan

Nikhilesh Prabhakar

UT DALLAS
The University of Texas at Dallas

Ravindran Balaraman

Arjun Manoharan

IIT MADRAS
Indian Institute of Technology Madras

Prasad Tadepalli

Oregon State University

Eric Blasch

AFRL
THE AIR FORCE RESEARCH LABORATORY

Junkyu Lee

Kavitha Srinivas

Michael Katz

Shirin Sohrabi

IBM **Research**

Questions?

T