Reinforcement Learning

Beyond the Markov Property

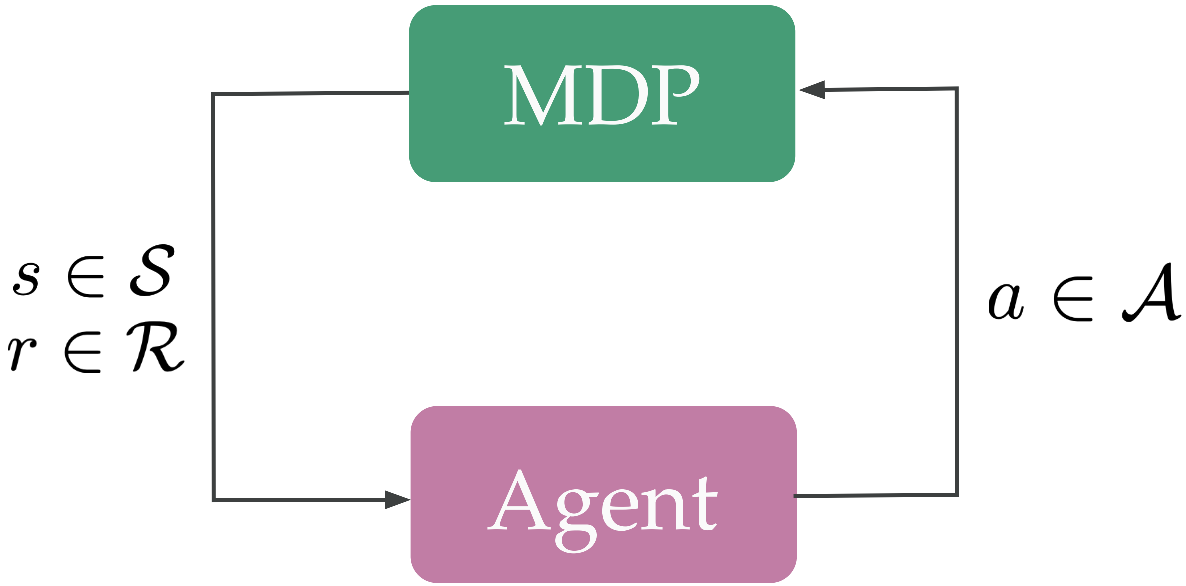
David Abel, Michael Herrmann

11 March, 2025

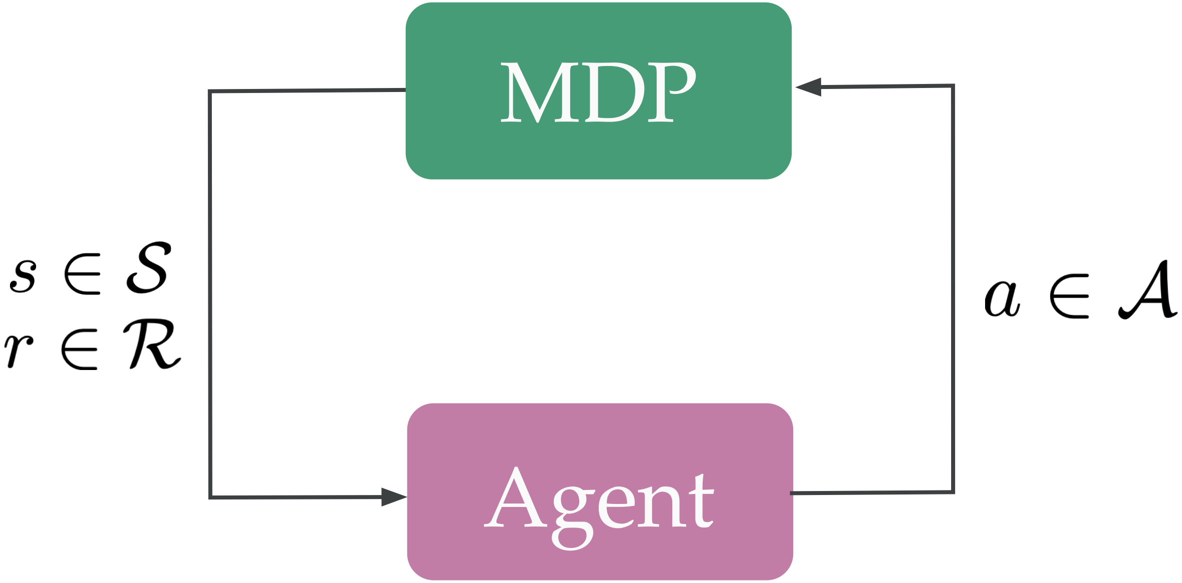
Lecture Overview

1. A Refresh: Markov, State
2. POMDP: Partially Observable MDPs
3. AIXI
4. The Big World Hypothesis

The Markov Property

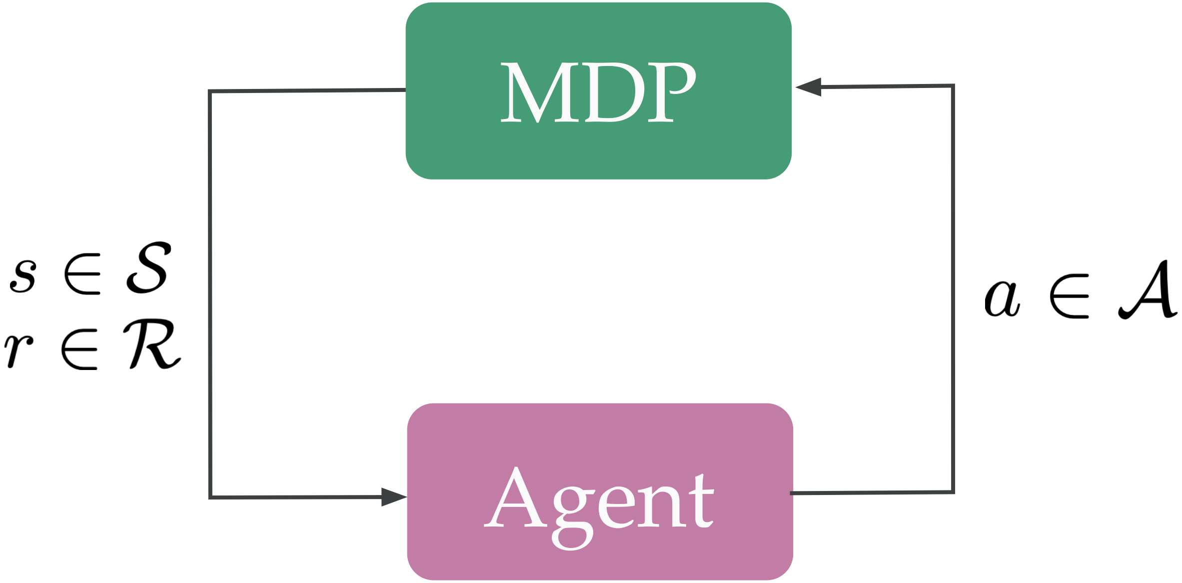


The Markov Property



Q: What does the Markov Property grant us as agent designers?

The Markov Property



$$p(s', r \mid s, a) \implies \begin{matrix} \pi_*(s) \\ v_*(s) \end{matrix}$$

The Markov Property

Definition: RL Problem (MDPs)

???

Definition: Planning Problem (MDPs)

???

$$p(s', r \mid s, a) \quad \boxed{\Longrightarrow} \quad \begin{array}{c} \pi_*(s) \\ v_*(s) \end{array}$$

The Markov Property

Definition: RL Problem (MDPs)

Given: $\mathcal{S}, \mathcal{A}$, *query access to* p

Repeat, for t = 0, 1, …

1. Agent selects action $A_t \in \mathcal{A}$
2. Agent observes

$$S_t, R_t \sim p(s', r \mid s, a)$$

Goal: maximise total reward.

Definition: Planning Problem (MDPs)

Given: an MDP, $(\mathcal{S}, \mathcal{A}, p)$

Output: An optimal policy, π

The Markov Property

Definition: RL Problem (MDPs)

Learn to maximise reward by interaction

Definition: Planning Problem (MDPs)

Compute the optimal policy given complete knowledge of

$$(\mathcal{S}, \mathcal{A}, p)$$

$$p(s', r \mid s, a) \quad \boxed{\Longrightarrow} \quad \begin{matrix} \pi_*(s) \\ v_*(s) \end{matrix}$$

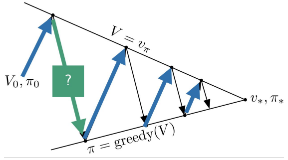
The Markov Property

(One) Algorithm: RL Problem (MDPs)

$$Q_{t+1}(s,a) = (1-\alpha)Q_t(s,a) + \alpha(r + \gamma \max_{a'} Q_t(s',a'))$$

Q-Learning

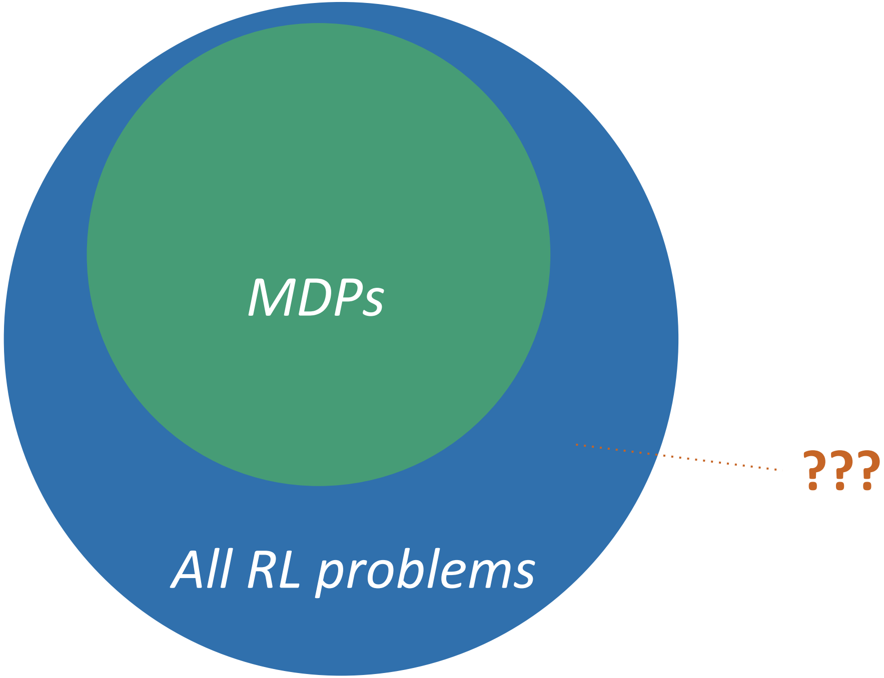
(One) Algorithm: Planning (MDPs)



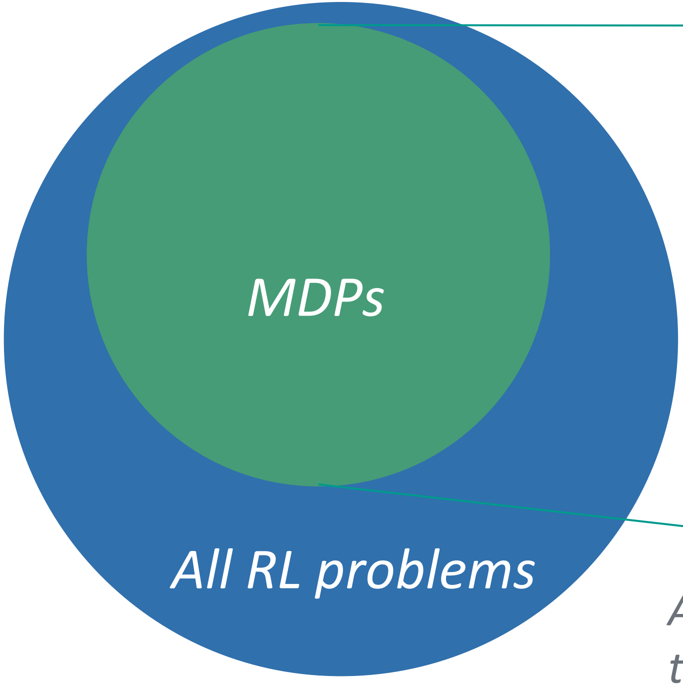
Policy Iteration

$$p(s', r \mid s, a) \quad \Longrightarrow \quad \begin{matrix} \pi_*(s) \\ v_*(s) \end{matrix}$$

Beyond the Markov Property



Beyond the Markov Property



E^3

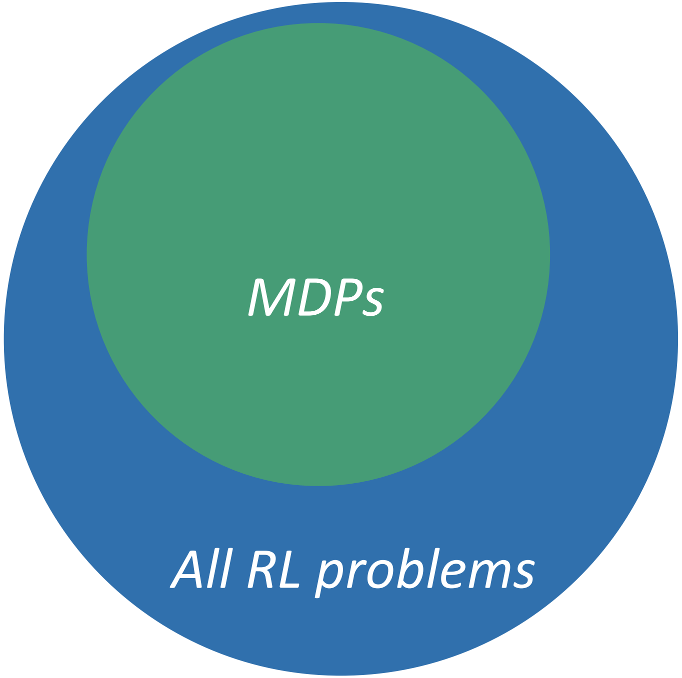
[Kearns and Singh; 2002]

R-MAX

[Brafman and Tenenholtz; 2003]

Algorithms that are guaranteed to do near-optimal RL in MDPs!

Beyond the Markov Property



???

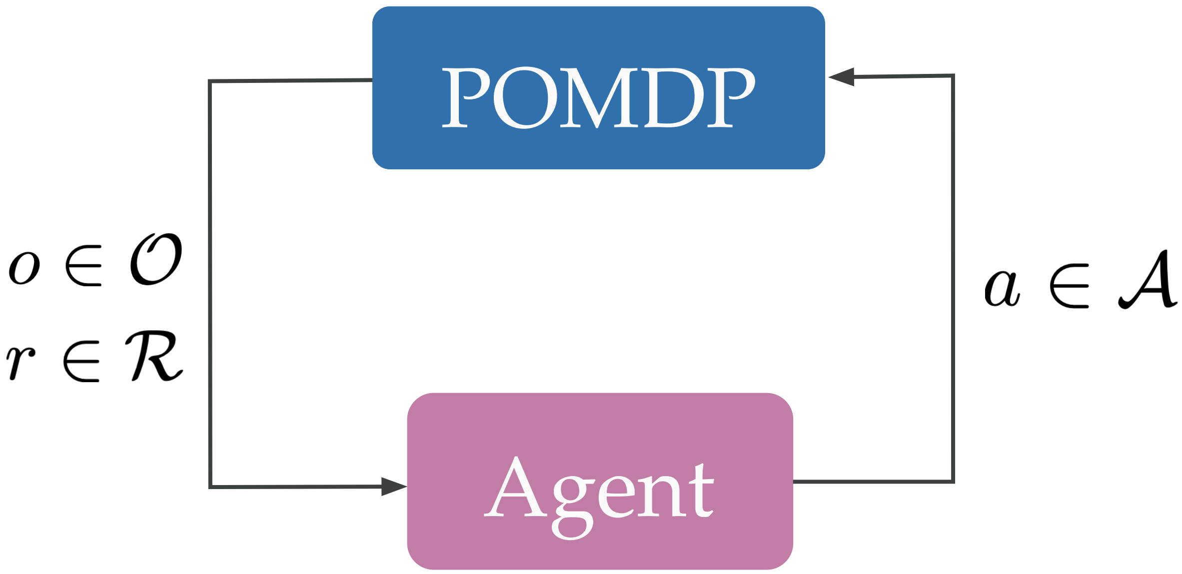
Partially Observable MDPs (POMDPs)

$\mathcal{S}$	A set of states
$\mathcal{A}$	A set of actions
$p(s' \mid s, a)$	State transition function
$r(s, a)$	Reward function

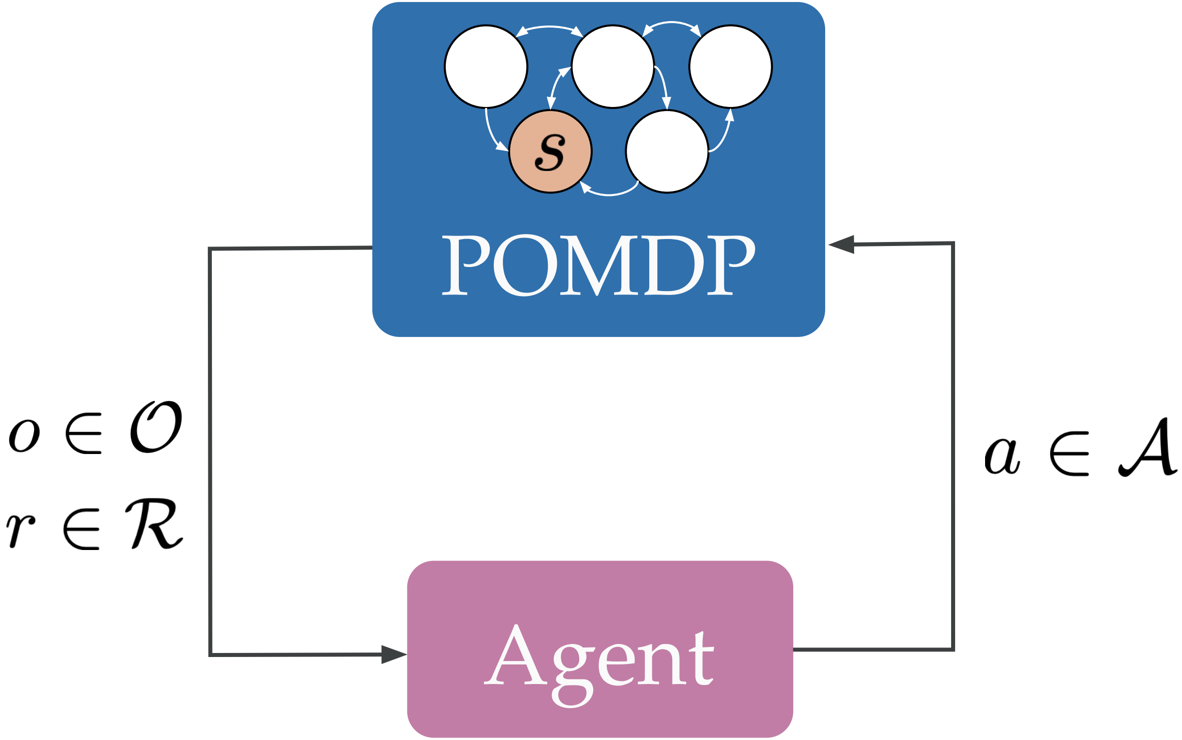
Partially Observable MDPs (POMDPs)

$\mathcal{S}$	A set of states
$\mathcal{A}$	A set of actions
$\mathcal{O}$	A set of observations
$p(s' \mid s, a)$	State transition function
$r(s, a)$	Reward function
$\omega(o \mid s, a)$	Observation function

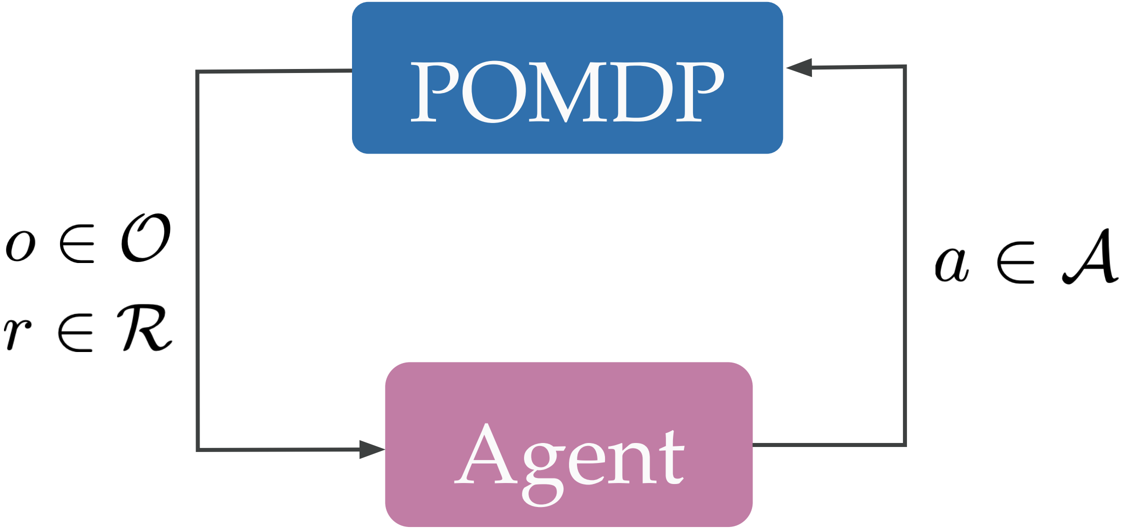
Beyond the Markov Property



Beyond the Markov Property



Solving POMDPs: RL and Planning



Definition: RL Problem (POMDPs)

Definition: Planning Problem (POMDPs)

Partially Observable MDPs (POMDPs)

Definition: RL Problem (POMDPs)

???

Definition: Planning Problem (POMDPs)

???

Discussion (2 minutes):

What is the RL problem in POMDPs?

What is the planning problem in POMDPs?

(if time): Do our MDP algorithms work in either?

RL and Planning in POMDPs

Definition: RL Problem (POMDPs)

Given: $(\mathcal{S}, \mathcal{A}, \mathcal{O})$, *query access to* e

Repeat, for t = 0, 1, …

1. Agent selects action $A_t \in \mathcal{A}$
2. Agent observes

$$O_t, R_t \sim e(o, r \mid s, a)$$

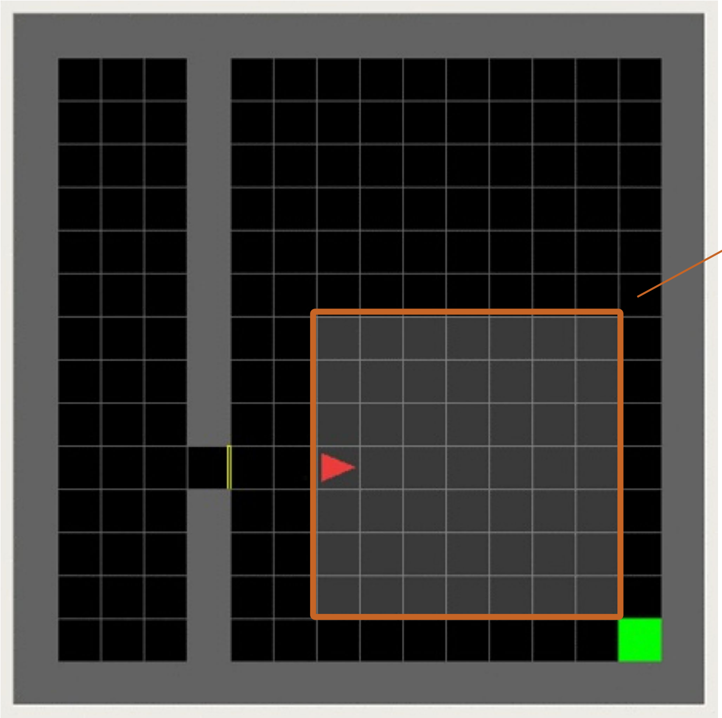
Goal: maximise total reward.

Definition: Planning Problem (POMDPs)

Given: an POMDP, $(\mathcal{S}, \mathcal{A}, \mathcal{O}, p, r, \omega)$

Output: An optimal policy, π

POMDP Example



14 x 14 grid

Uncertainty
about state!

state = (x,y, agent-direction)
observation = (7x7 grid, agent-direction)

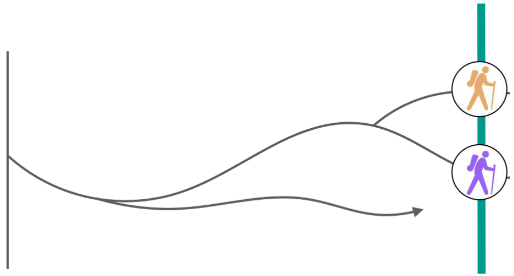
Q: How many environment states *could* the agent be in, while seeing this o, roughly?

Finite vs. Infinite Horizon

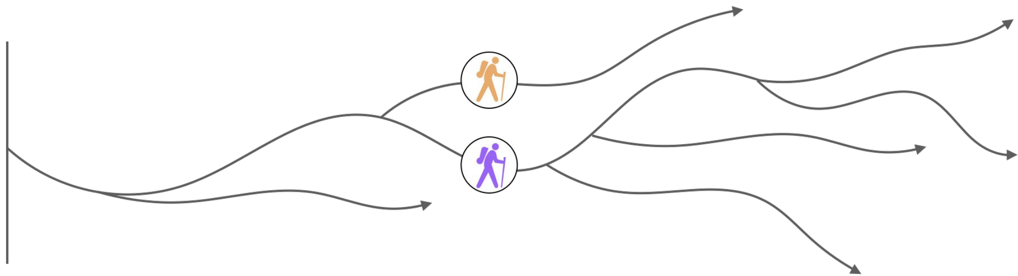
Definition: Planning Problem (POMDPs)

Given: an POMDP, $(\mathcal{S}, \mathcal{A}, \mathcal{O}, p, r, \omega)$

Output: An optimal policy, π



Finite Horizon



Infinite Horizon

Planning: A Few Classic POMDP Results

Theorem. Planning in **infinite horizon** POMDPs is *undecidable*.

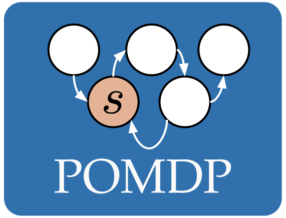
[O. Madani, S. Hanks, A. Condon; 2003]

Theorem. Planning in **finite horizon** POMDPs, is $\texttt{PSPACE}$-complete

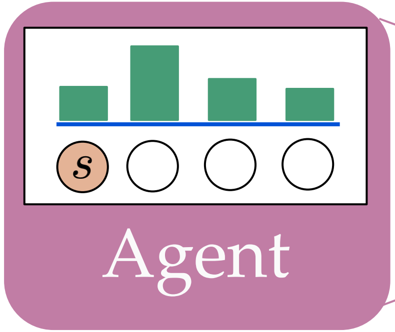
[Papadimitriou, Tsitsiklis; 1987]

Takeaway: Planning in POMDPs is hard!

Learning in POMDPs: Approximate



- Belief state complexity can be independent of env.
- Maintained via Bayes
- Value function can be function of belief state

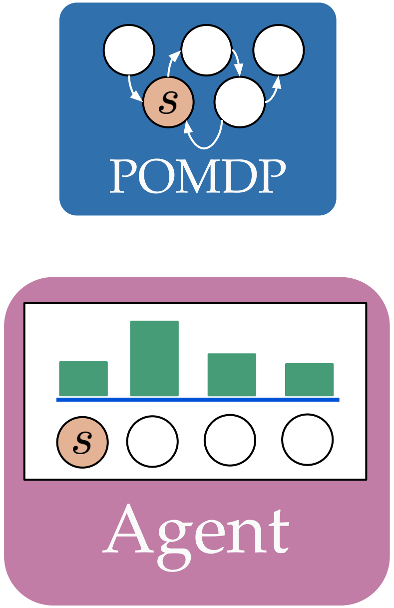


$$b(s) = p(s \mid o, a, b')$$

Belief State, Belief MDPs

[Kaelbling, Littman, Cassandra; 1996]

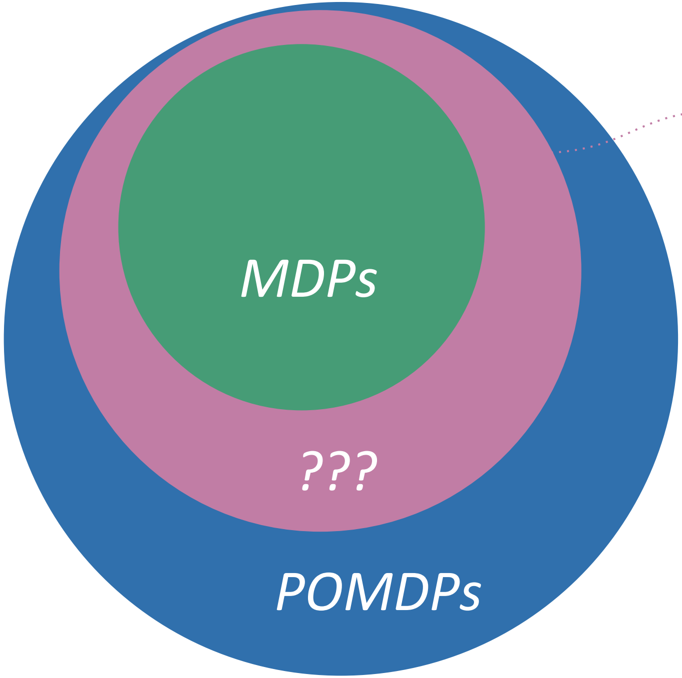
Learning in POMDPs: Approximate



Belief State, Belief MDPs

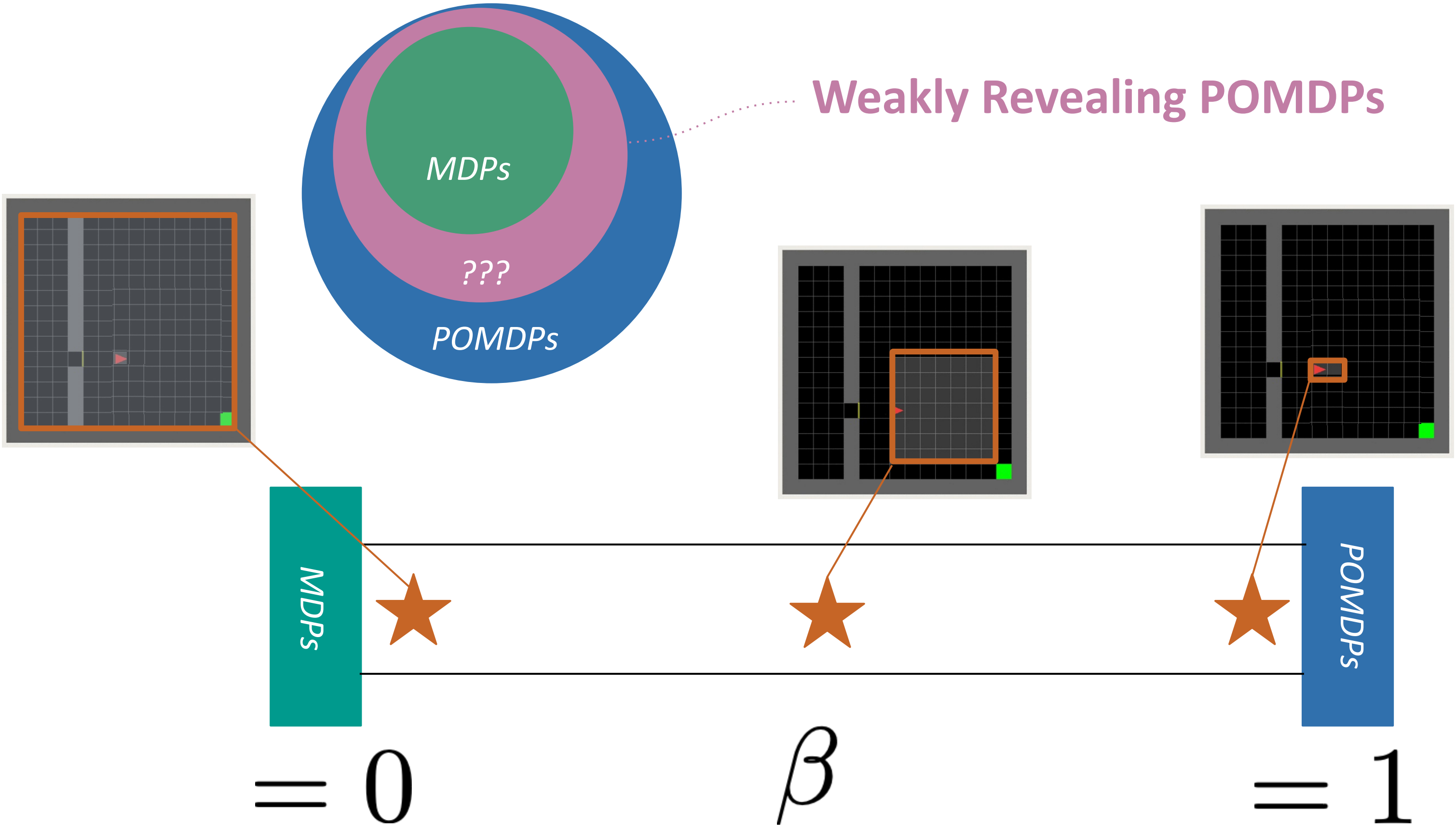
Predictive State

What Makes a POMDP Hard?



Weakly Revealing POMDPs
[Jin et al.; 2023]

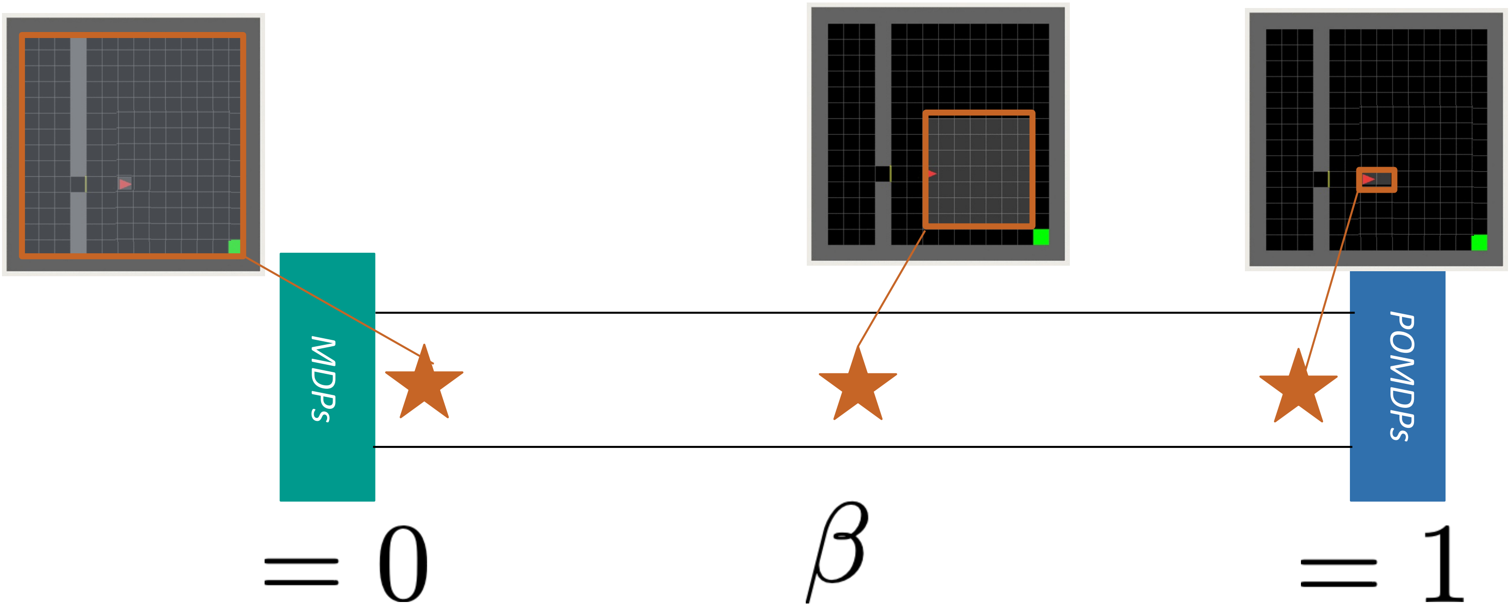
Weakly Revealing POMDPs



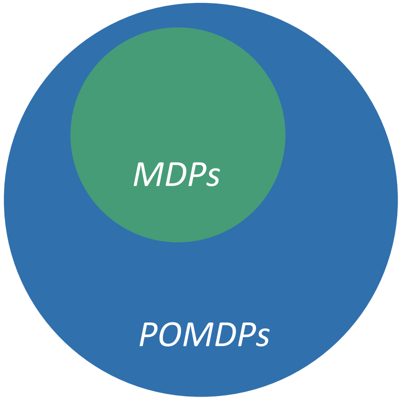
Weakly Revealing POMDPs

Theorem. Learning difficulty in POMDPs scales as a function of β

[Jin et al.; 2023]



General POMDPs



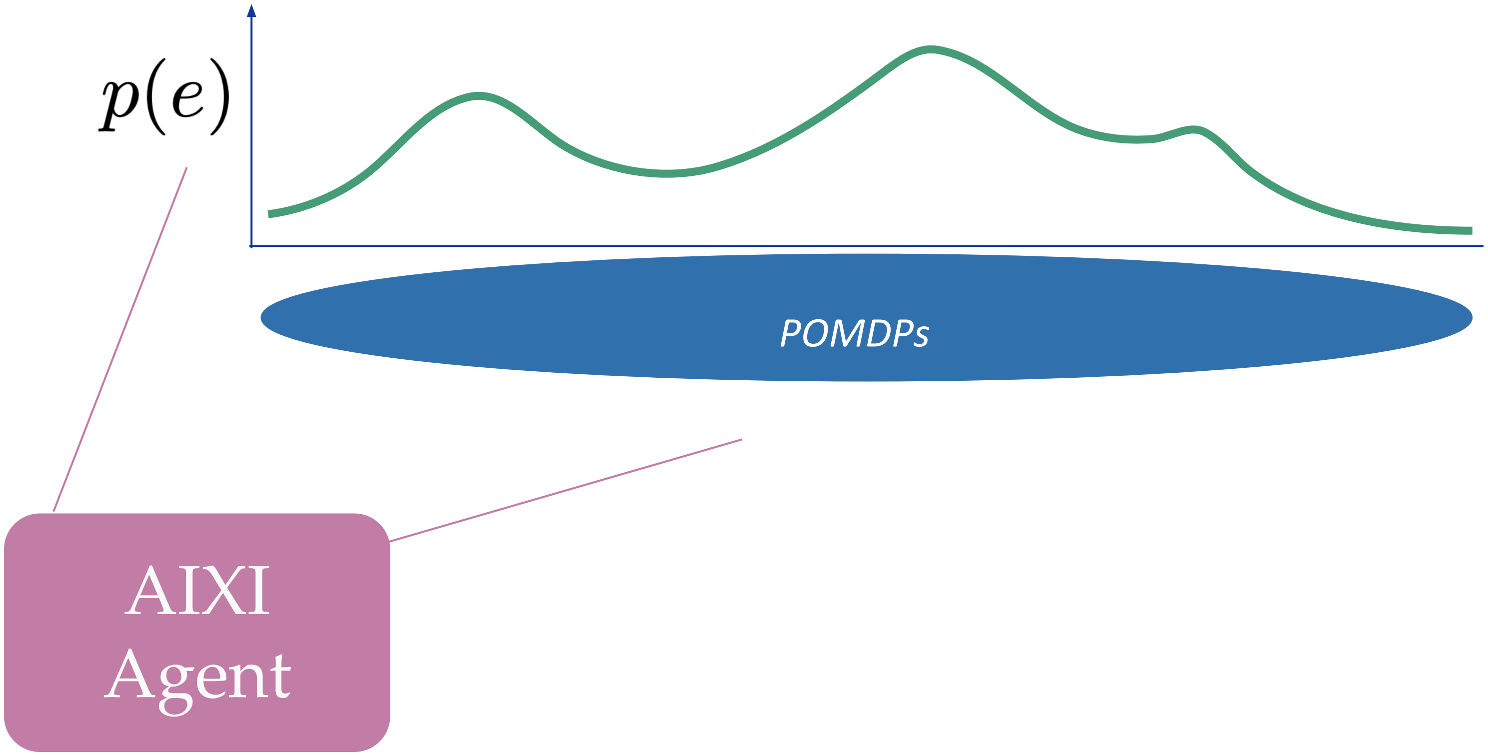
Q: Learning in *all* POMDPs?
Any RL problem?

AIXI

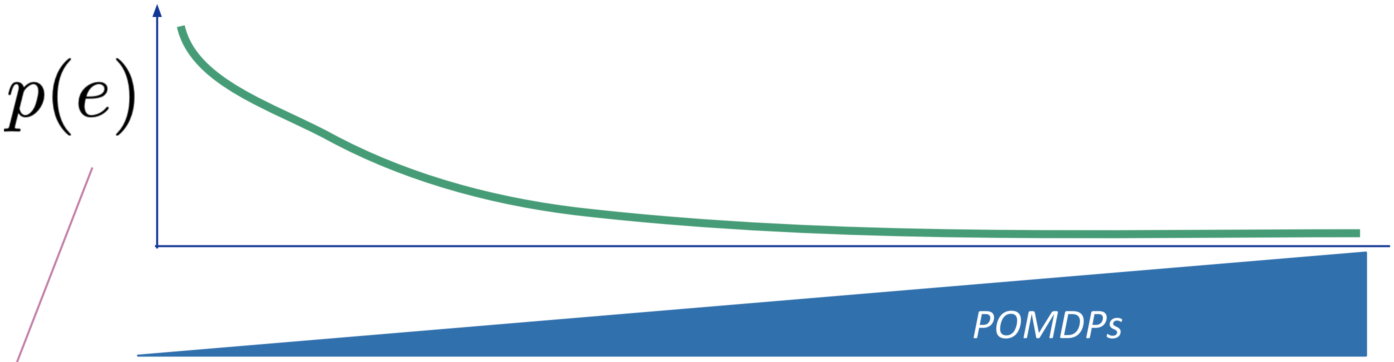
[Hutter; 2000]

(1) No constraints on agent

AIXI



AIXI

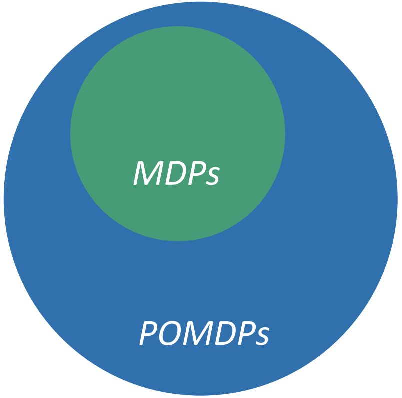


Ordered from simplest to most complex

AIXI Agent

$$a = \arg\max_{a \in \mathcal{A}} Q_{p(e)}(o_1 o_2 \ldots, a)$$

General POMDPs



Q: Learning in *all* POMDPs?
Any RL problem?

AIXI

[Hutter; 2000]

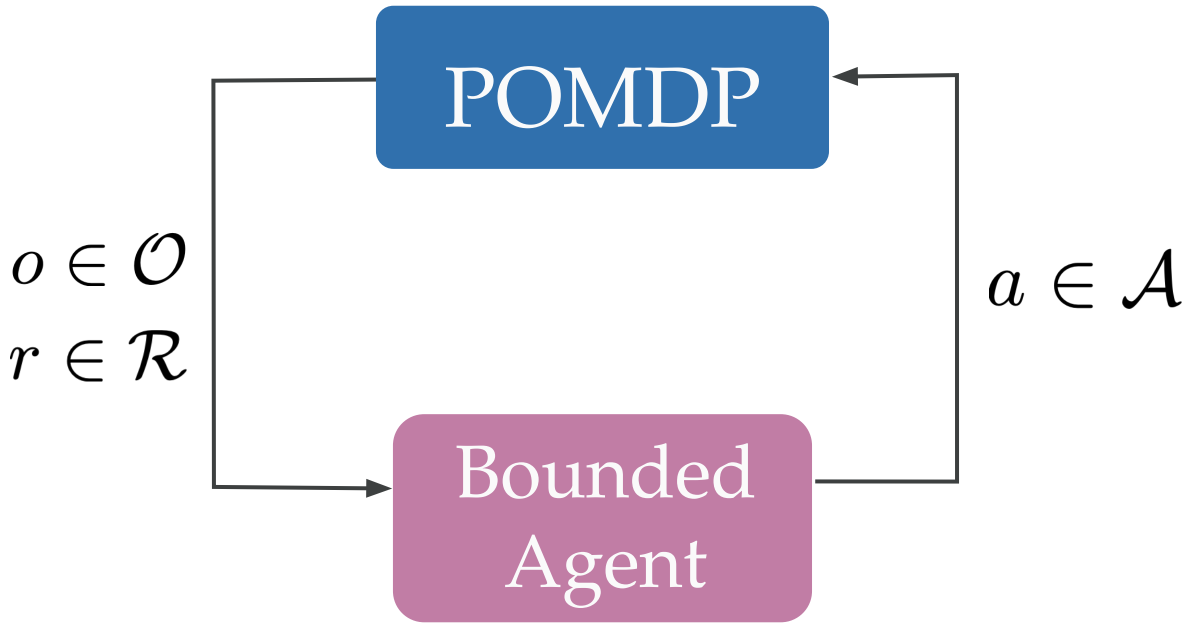
(1) No constraints on agent

Simple Agent,
Complex Environment

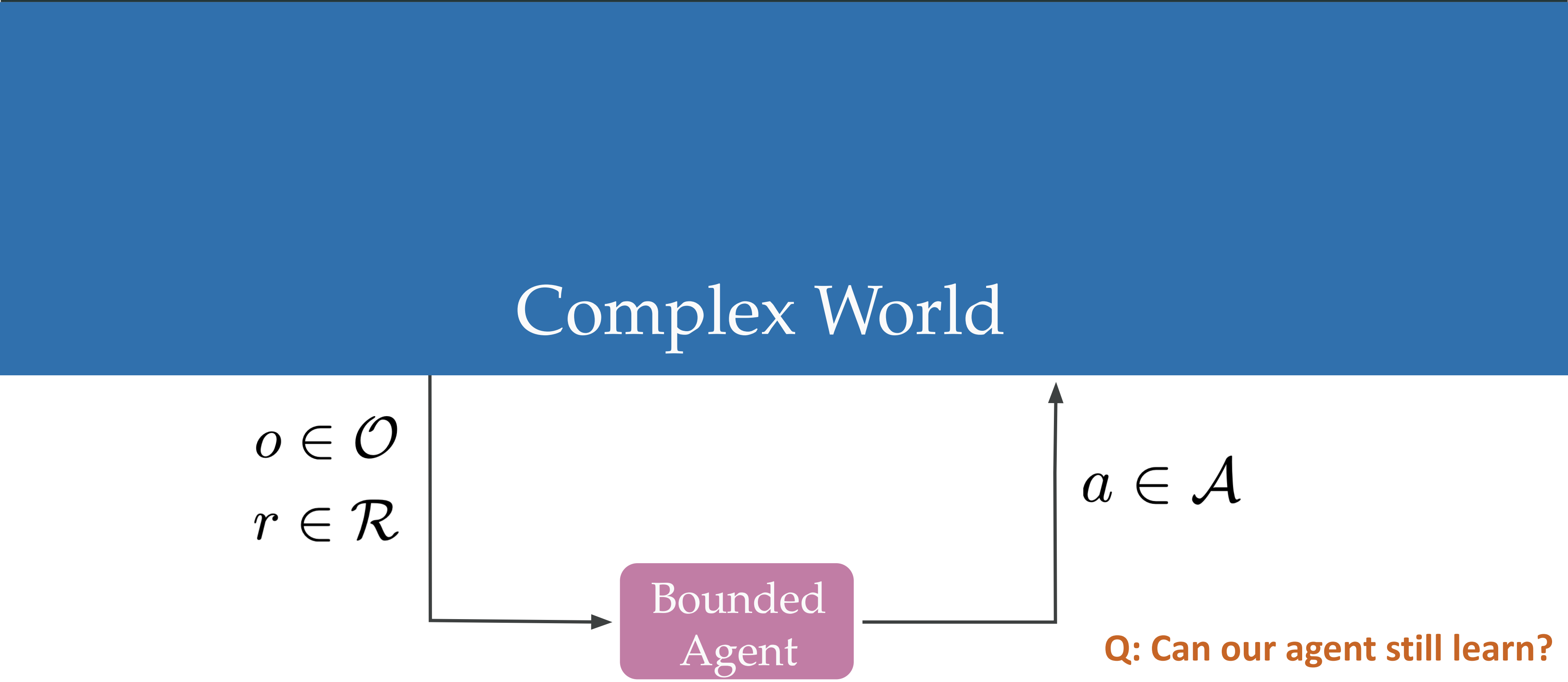
[Dong et al.; 2022]

(2) Finite agent

Simple Agent, Complex Environment by Dong et al. (2022)

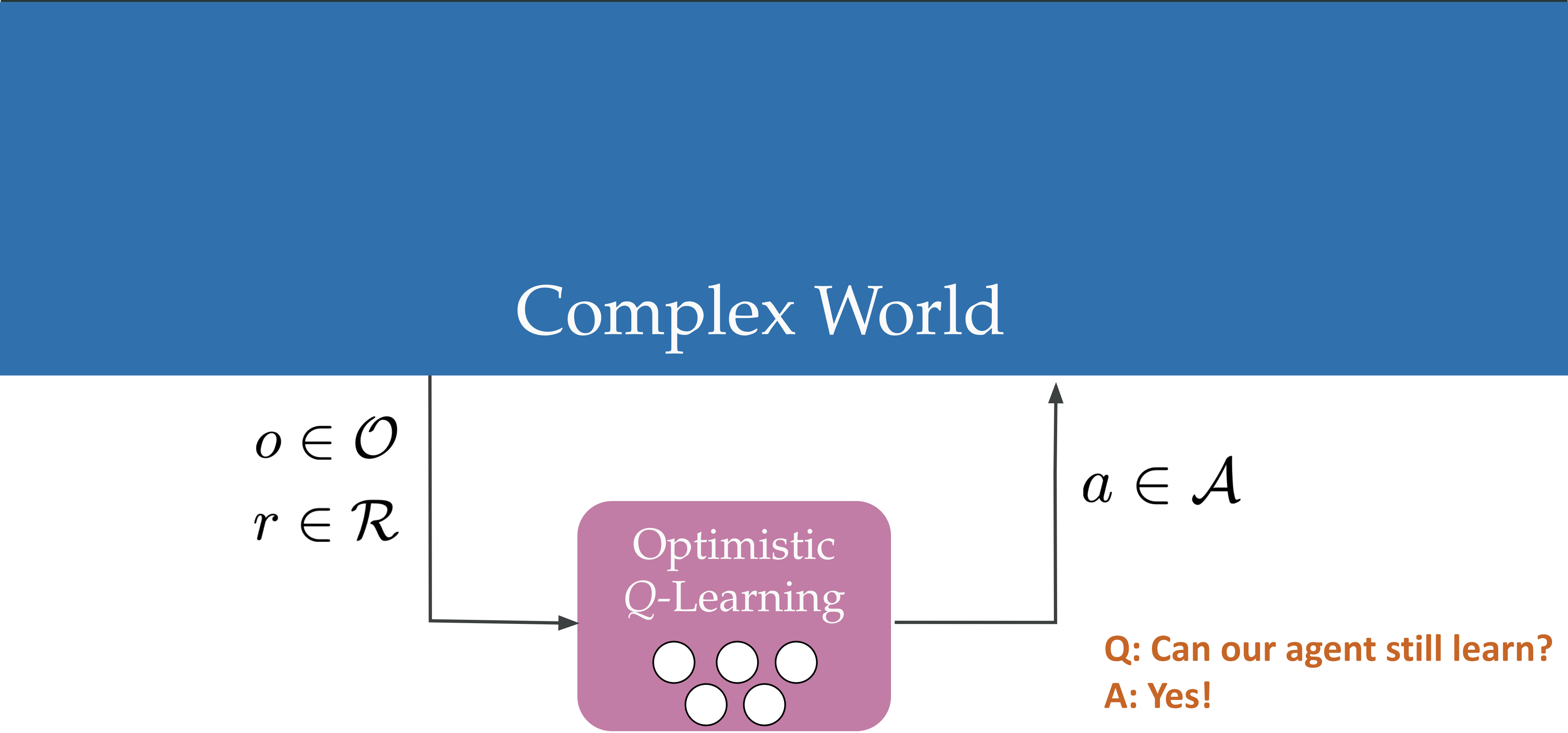


Simple Agent, Complex Environment by Dong et al. (2022)



Q: Can our agent still learn?

Simple Agent, Complex Environment by Dong et al. (2022)

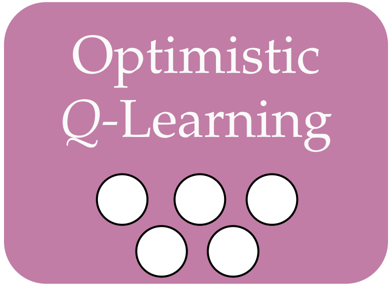


Q: Can our agent still learn?
A: Yes!

Simple Agent, Complex Environment by Dong et al. (2022)

Theorem. Bounded optimistic *Q*-learning will perform only f(**agent size**, **env. complexity**) worse than the *best unbounded agent*.

[Dong et al.; 2022]



Q: Can our agent still learn?

A: Yes!

The Big World Hypothesis

"In many decision-making problems the agent is orders of magnitude smaller than the environment"

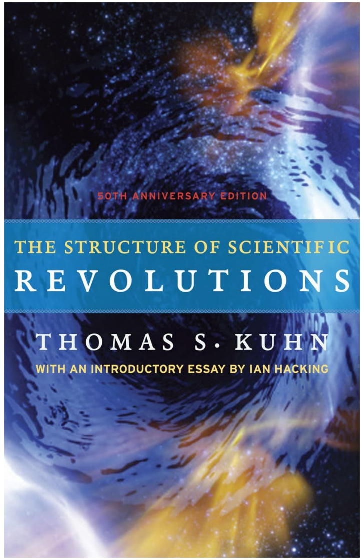
- Javeed, Sutton (2024)

Bounded rationality
-Simon

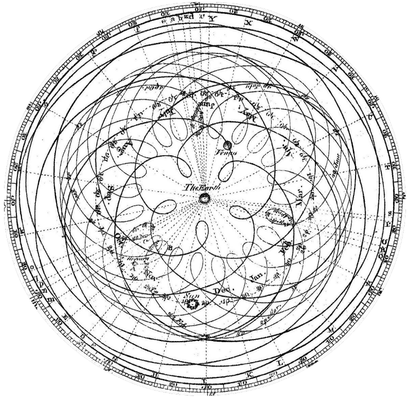
All models are wrong, some are useful
-Box

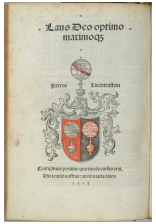
Open-endedness
-Lehman and Stanley

RL: The Road Ahead

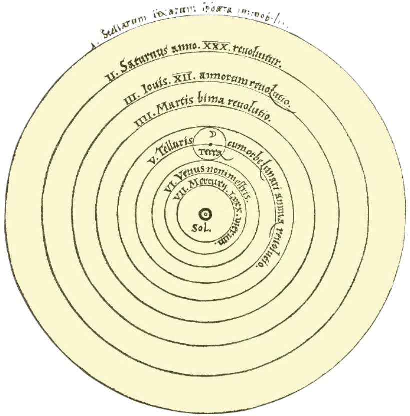


Geocentric
(≈150 AD)



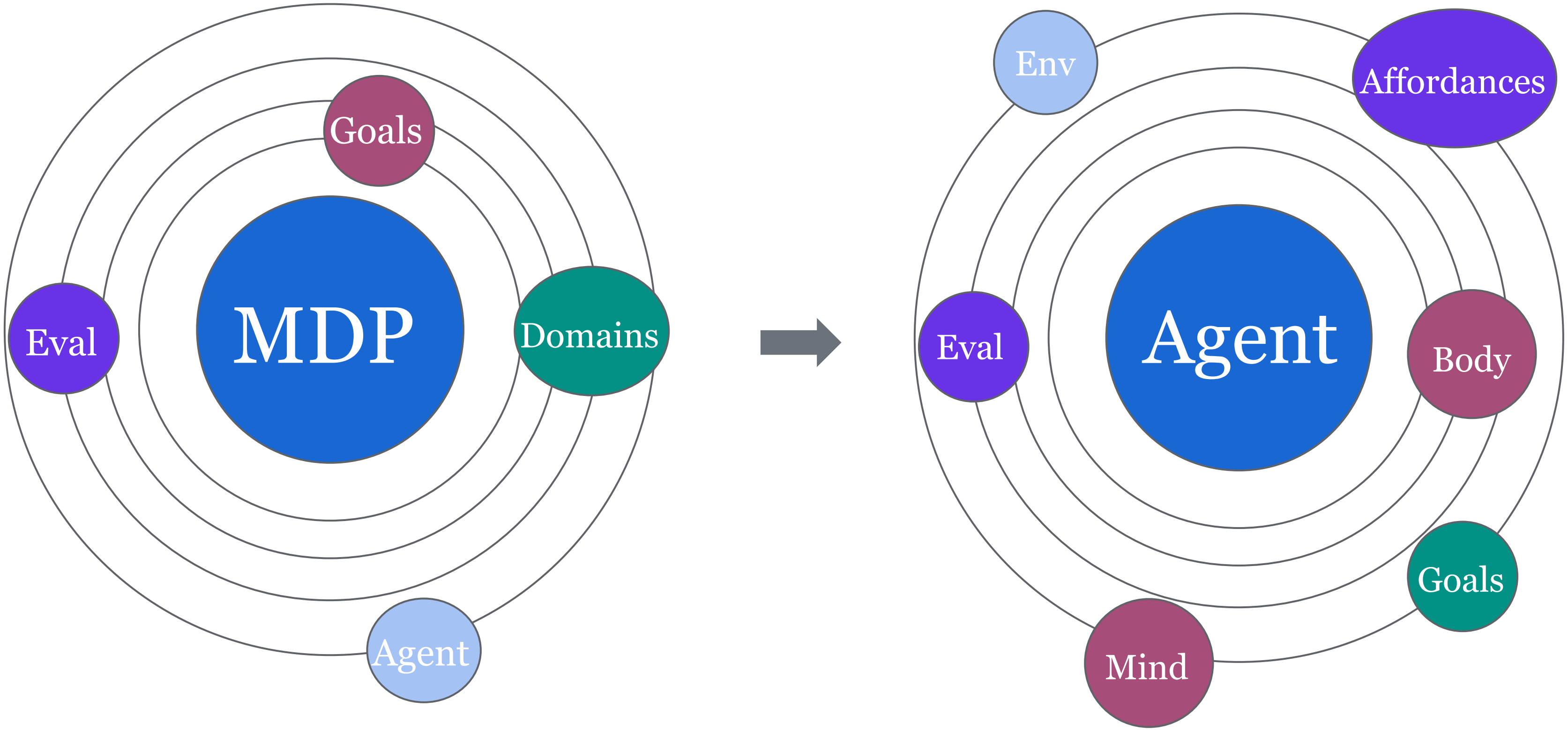


Heliocentric
(≈1500 AD)



\+ *Mechanics, germ theory, …*

RL: The Road Ahead



RL: The Road Ahead

RL as the science of *learning to act*

My Last Lecture: Feedback

Feedback: tinyurl.com/dave-feedback

Thank you!

Optional Reading

POMDPs

- *Learning Without State-Estimation in Partially Observable Markovian Decision Processes* by Singh, Jaakola, Jordan (1994)
- *Planning and Acting in Partially Observable Stochastic Domains* by Kaelbling, Littman, Cassandra (1996)
- *Predictive Representations of State* by Littman, Sutton, Singh (2001)

Big Worlds, AIXI

- *A Monte Carlo AIXI Approximation* by Veness et al. (2010)
- *Simple Agent, Complex Environment* by Dong, Zhou, Van Roy (2022)
- *The Big World Hypothesis and its Ramifications for AI* by Javeed and Sutton (2024)
- *The Alberta Plan*, by Sutton, Pilarski, Bowling (2022)