RL Tutorial 4

Marcelo Mattar

Learning vs. planning

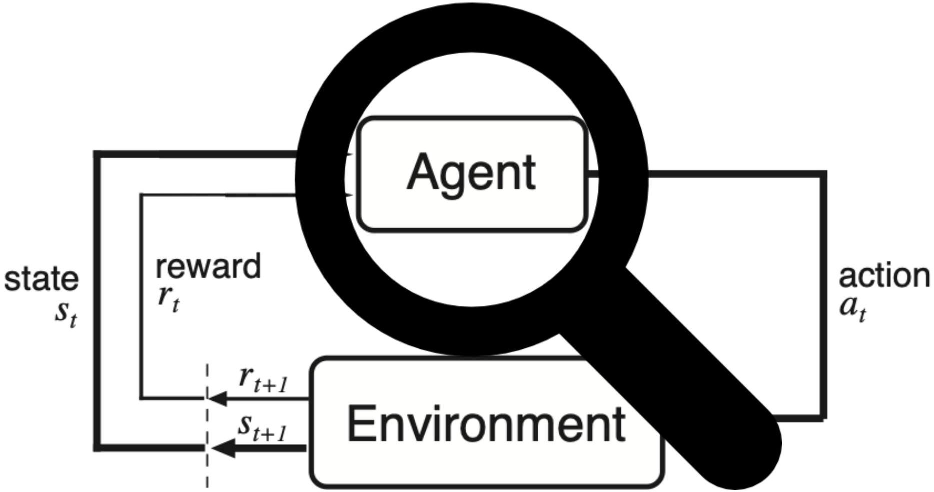
- So far, we've seen how agents learn values by acting in the world and *experiencing* the outcome of their actions
- Methods such as Q-learning are ***model-free*** as they do not require a model

- In this tutorial, we will learn about ***model-based*** methods, which compute action values via planning.
- Instead of learning values from experience, planning is the process of computing action values from a model.

But what is a model?

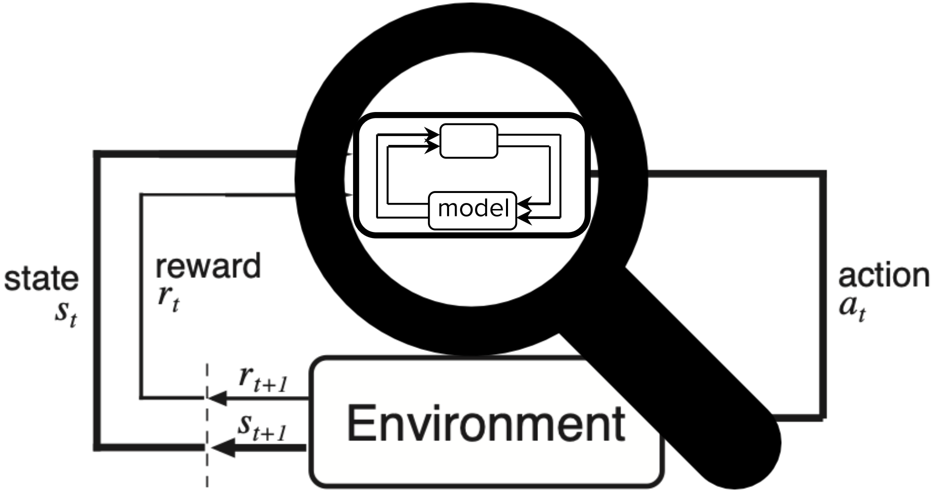
A representation of how the world will respond to the agent's actions

But what is a model?



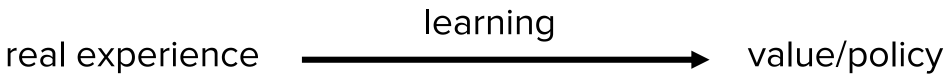
But what is a model?

Given a state and an action, a model returns the next state and next reward

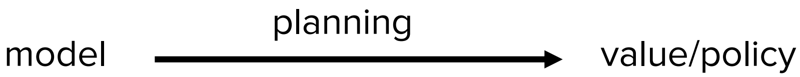


Model-free and Model-based RL

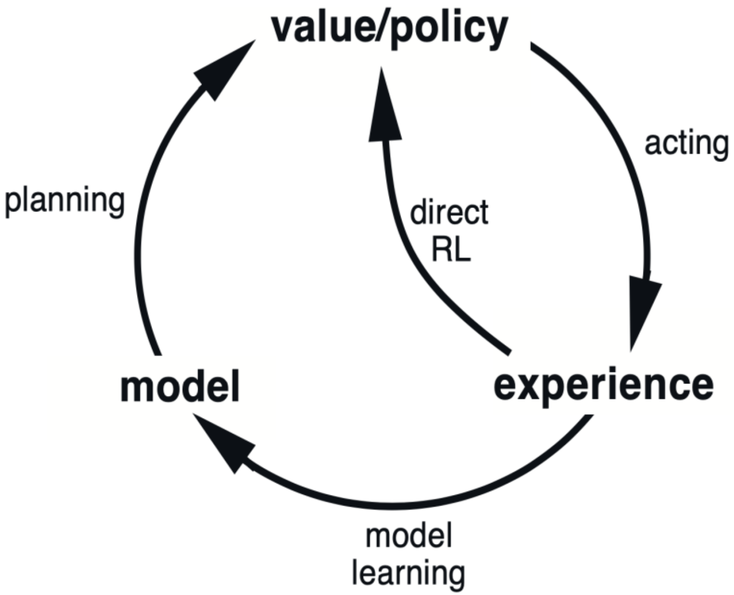
- Model-free RL:



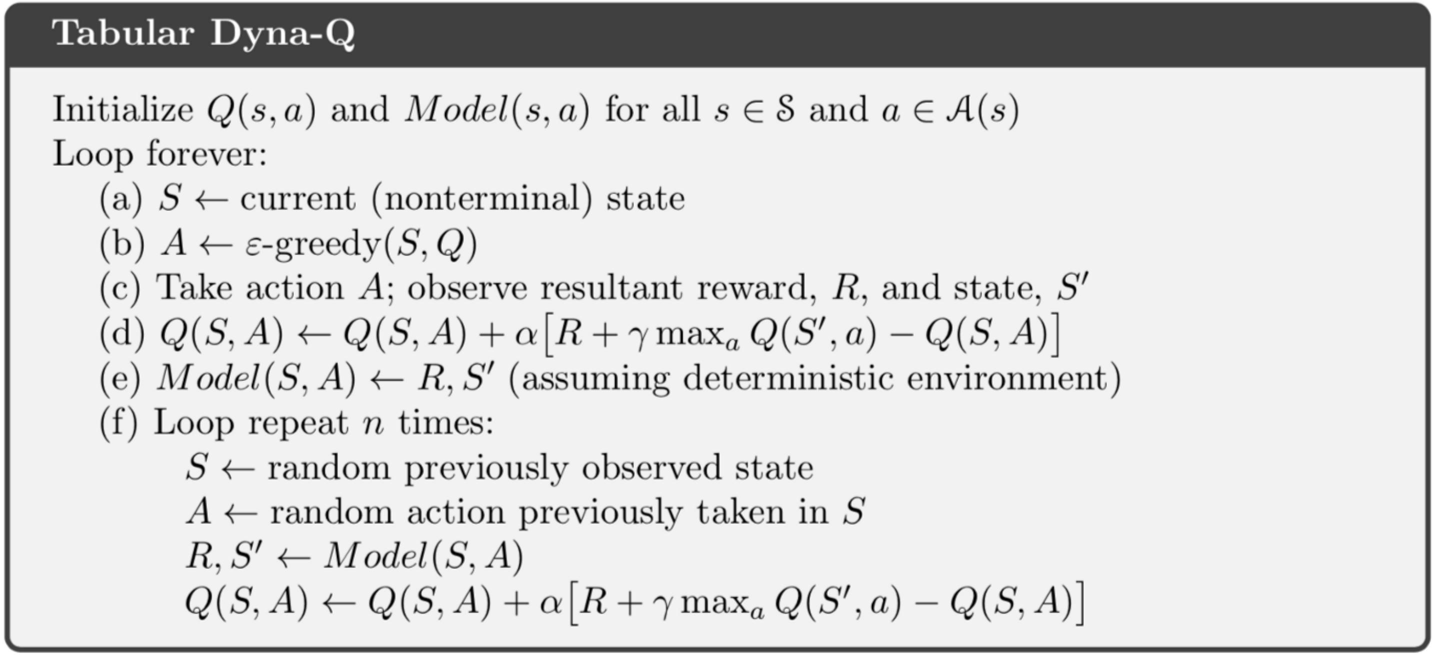
- Model-based RL:



Integrating planning and learning



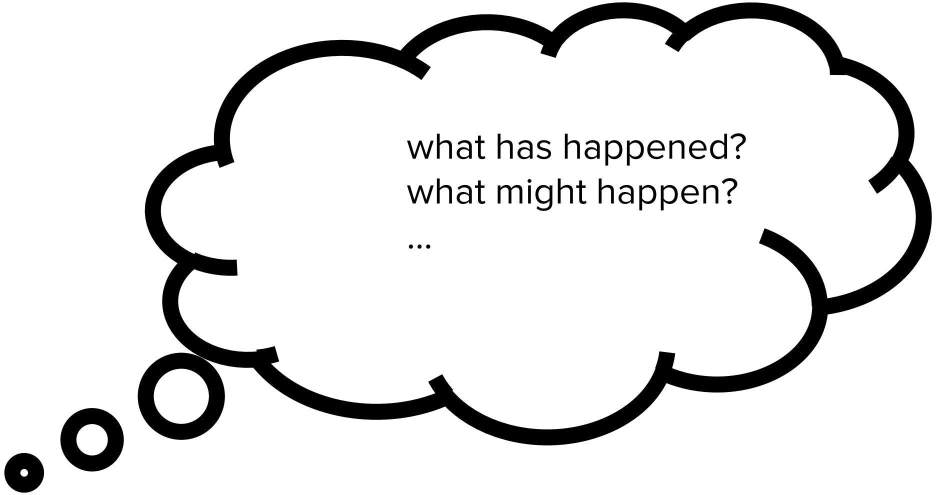
Dyna-Q architecture



Tabular Dyna-Q

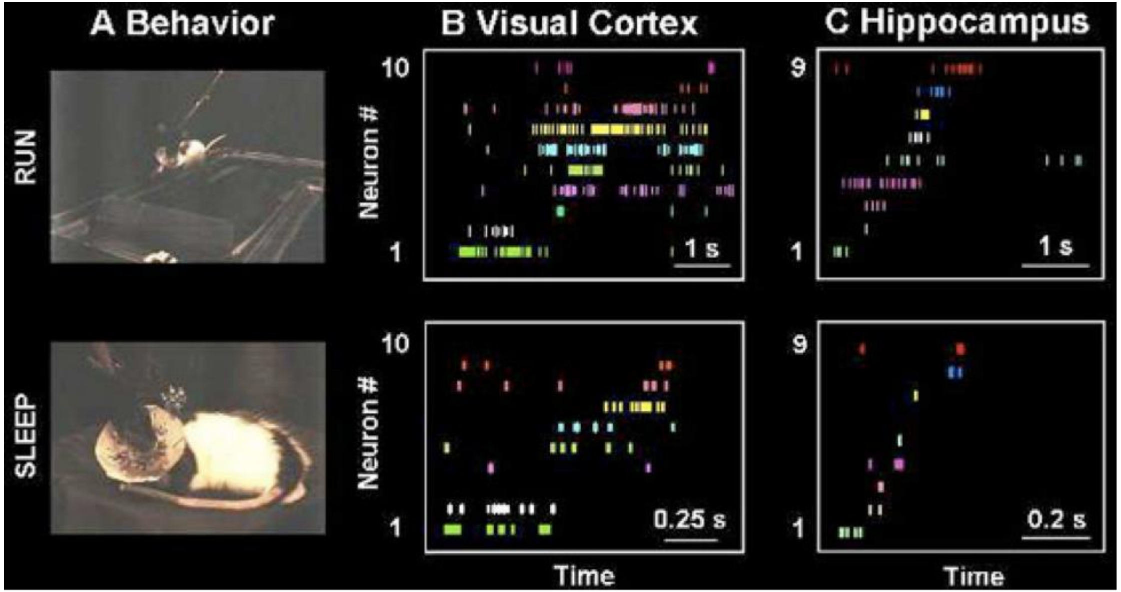
```
Initialize $Q(s,a)$ and $Model(s,a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$
Loop forever:
    (a) $S \leftarrow$ current (nonterminal) state
    (b) $A \leftarrow \varepsilon$-greedy$(S, Q)$
    (c) Take action $A$; observe resultant reward, $R$, and state, $S'$
    (d) $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',a) - Q(S,A)\big]$
    (e) $Model(S,A) \leftarrow R, S'$ (assuming deterministic environment)
    (f) Loop repeat $n$ times:
            $S \leftarrow$ random previously observed state
            $A \leftarrow$ random action previously taken in $S$
            $R, S' \leftarrow Model(S,A)$
            $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',a) - Q(S,A)\big]$
```

A mathematical framework for cognition?



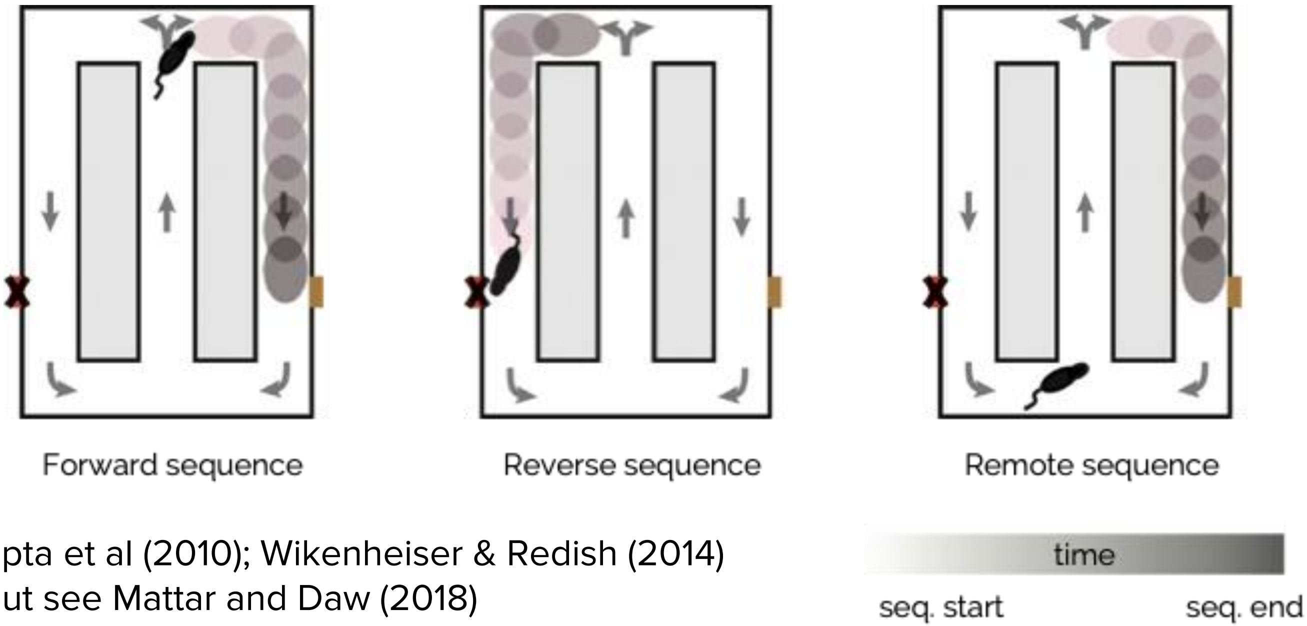
what has happened?
what might happen?
...

Replay for consolidation



Ji, D. & Wilson, M.A. (2007)

Replay for planning



Gupta et al (2010); Wikenheiser & Redish (2014)
...but see Mattar and Daw (2018)