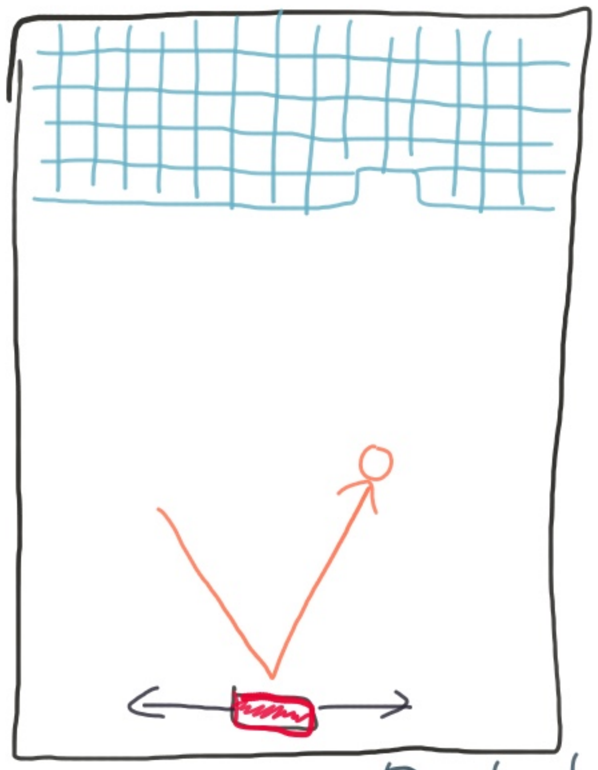
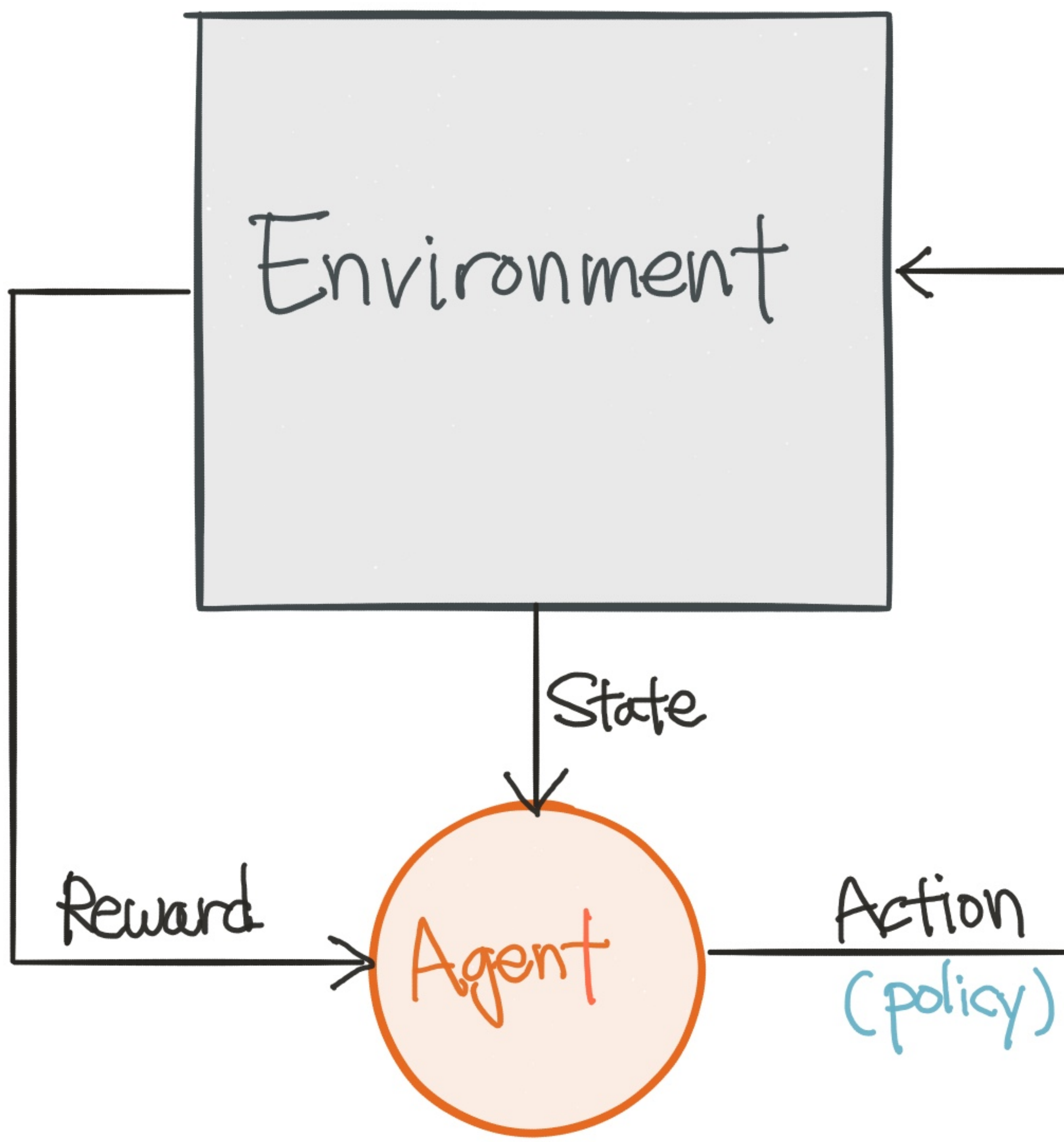
Reinforcement Learning



Each time you hit a brick, your score increases. (Reward)

The goal of RL is to find an optimal policy that maximizes the expected sum of reward.

Reinforcement Learning



Markov Decision Process

State World modelling $s \in S$

Action Possible actions $a \in A$

Reward $R(s,a) : S \times A \to \mathbb{R}$

Policy $\pi(s) : S \to A$

Transition model $T(s,a,s') : S \times A \to S$

Q function $Q(s,a) : S \times A \to \mathbb{R}$

Value function $V(s) : S \to \mathbb{R}$

Episode $s_0, a_0, r_0, s_1, a_1, r_1, \dots s_t, a_t, r_t$

Markov Decision Process

Terminology

To perform well in the long-term, we need to take into account the future rewards we are going to get

$$R = r_1 + r_2 + \cdots + r_n$$

$$R_t = r_t + r_{t+1} + \cdots + r_n$$

But, our world is stochastic!

$$R_t = r_t + \underbrace{\gamma r_{t+1} + \cdots + \gamma^{n-t} r_n}_{\text{discounted future reward}}$$

So we give more emphasis on current reward.

$$R_t = r_t + \gamma R_{t+1}$$

Discounted Future Reward

Q Learing

$$Q(s_t, a_t) = \max R_{t+1} = \max\left(r_{t+1} + \gamma r_{t+2} + \cdots\right)$$

Q function is 1) Maximum discounted future reward when we perform action a in state s.

2) The best possible score at the end of the game.

3) The quality of certain action in given state

$$\pi(s) = \arg\max_a Q(s,a)$$

Q Learning

Given $\langle s, a, r, s' \rangle$

Bellman Equation

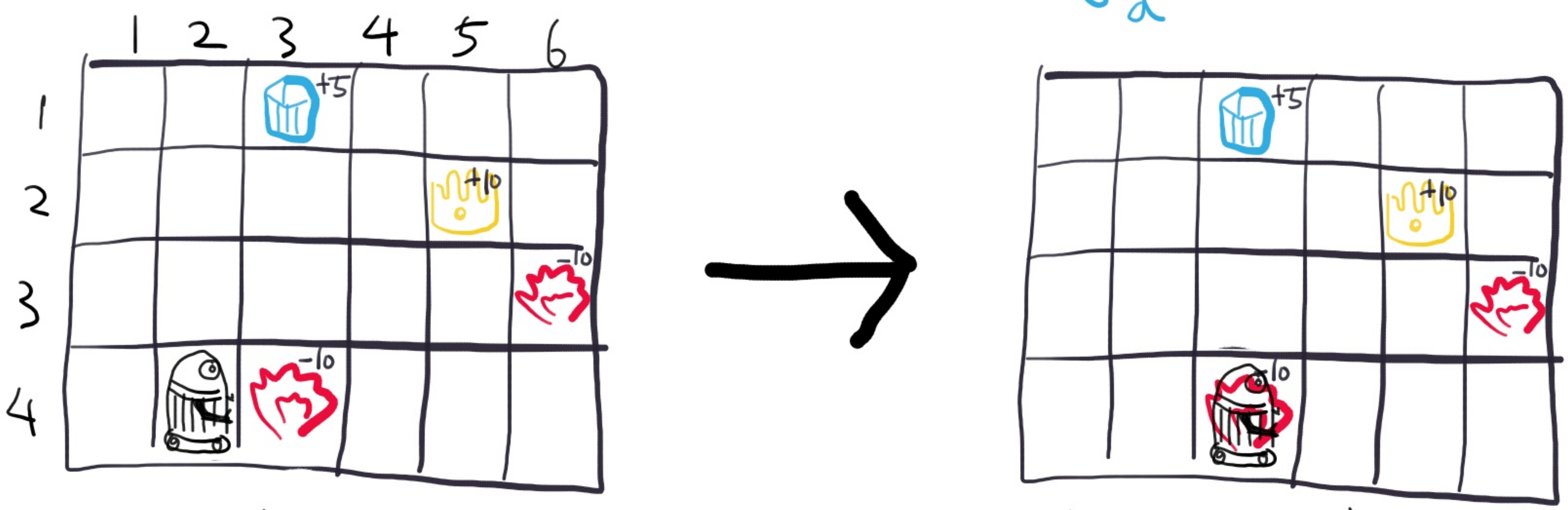
immediate reward

next state

$$Q(s,a) = r + \gamma \max_{a'} Q(s',a')$$

discount factor

$a' = \arg\max_{a} Q(s',a)$



$\langle s=[4,2],\ a=\rightarrow,\ r=-10,\ s'=[4,3] \rangle$

Bellman Equation

Initialize $Q(s,a)$

Observe initial state s

repeat

select and carry out an action a

observe reward r and new state s'

$$Q(s,a) = Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)$$

$s = s'$

until terminated

Q Learning Algorithm

Initialize $Q(s,a)$

Observe initial state s

repeat

$a = \arg\max_{a'} Q(s,a')$
or
Randomly

select and carry out an action a

Comes from some oracle

observe reward r and new state s'

$$Q(s,a) = Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)$$

learning rate

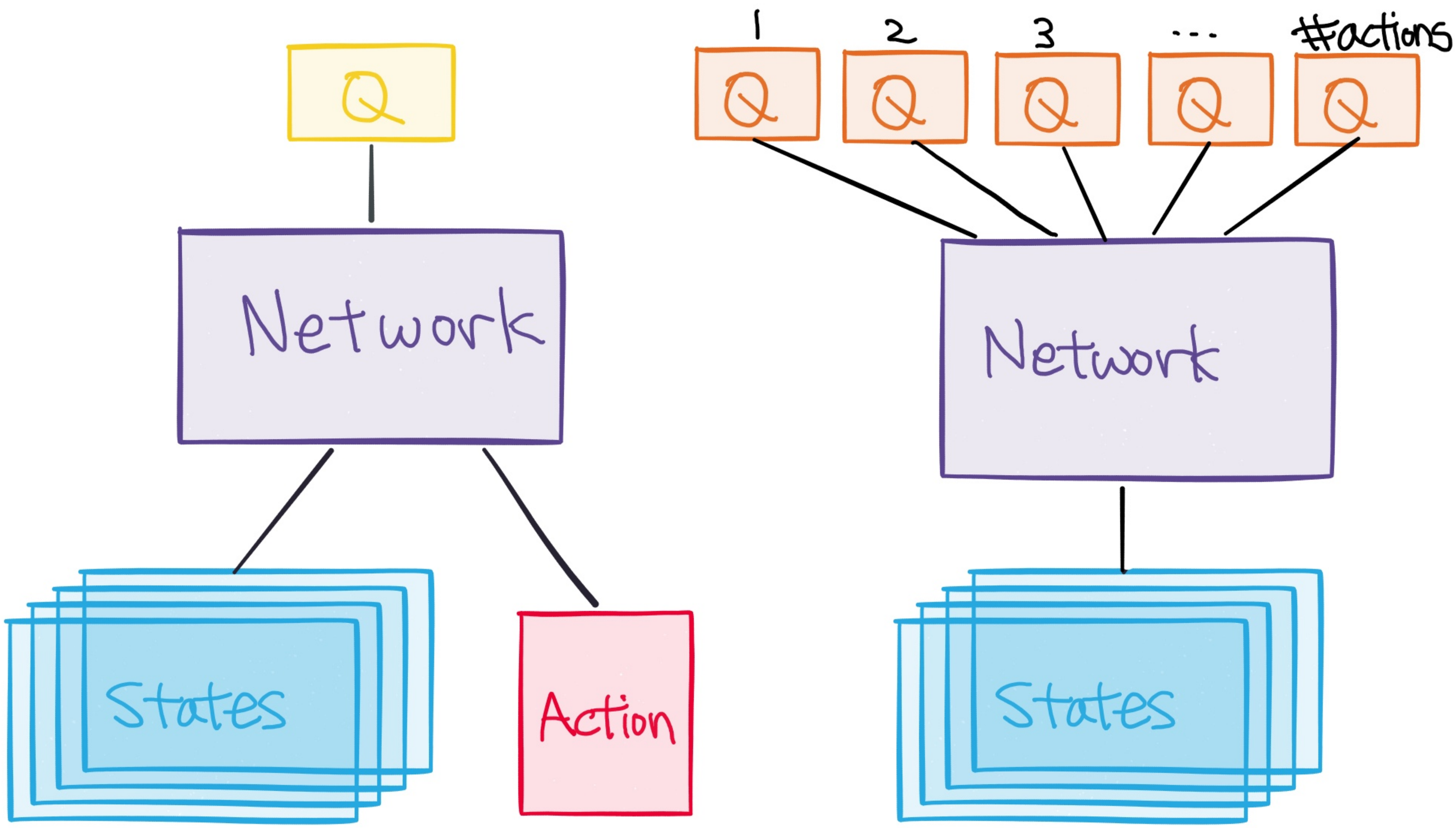
discount factor

$s = s'$

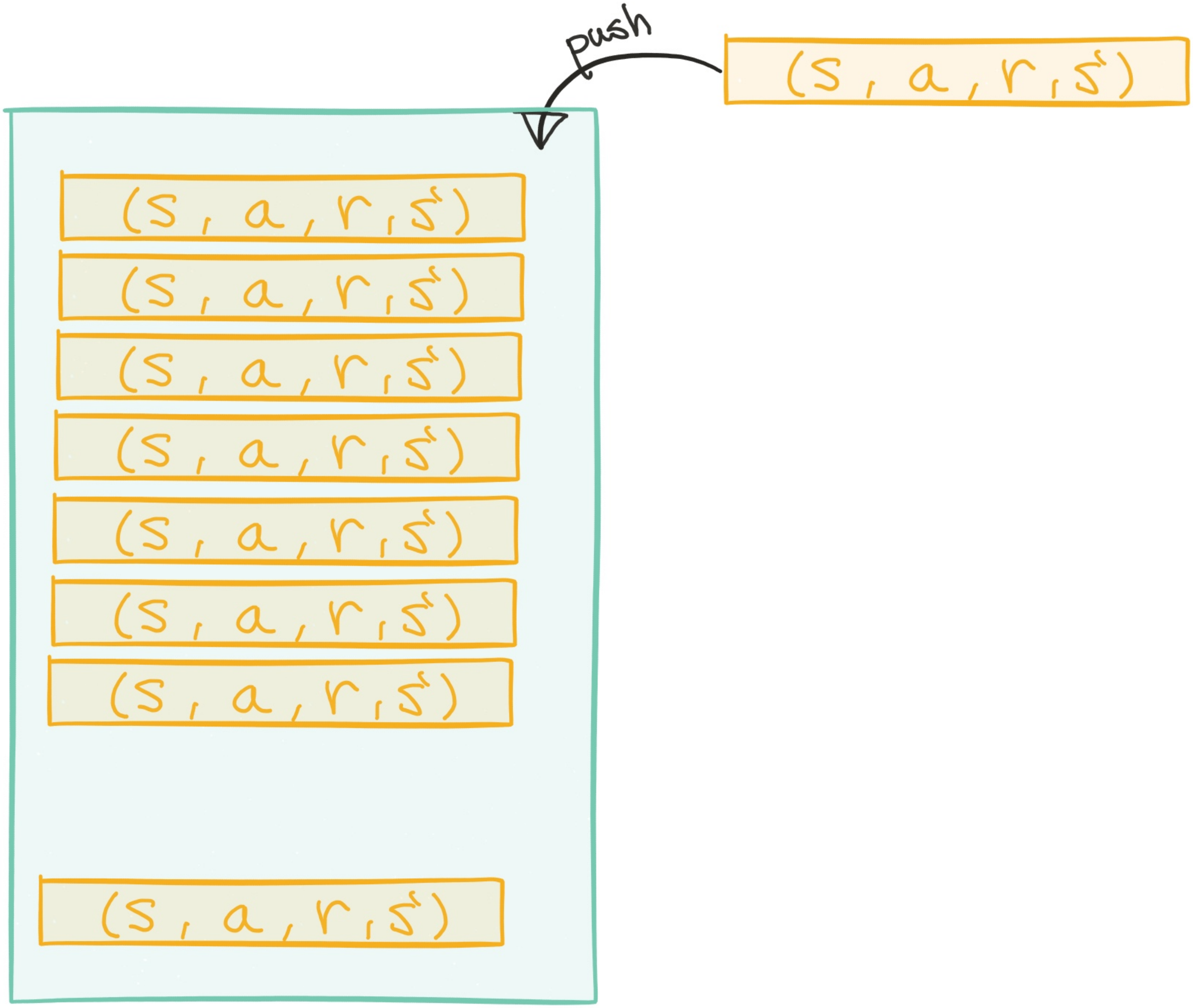
until terminated

immediate reward

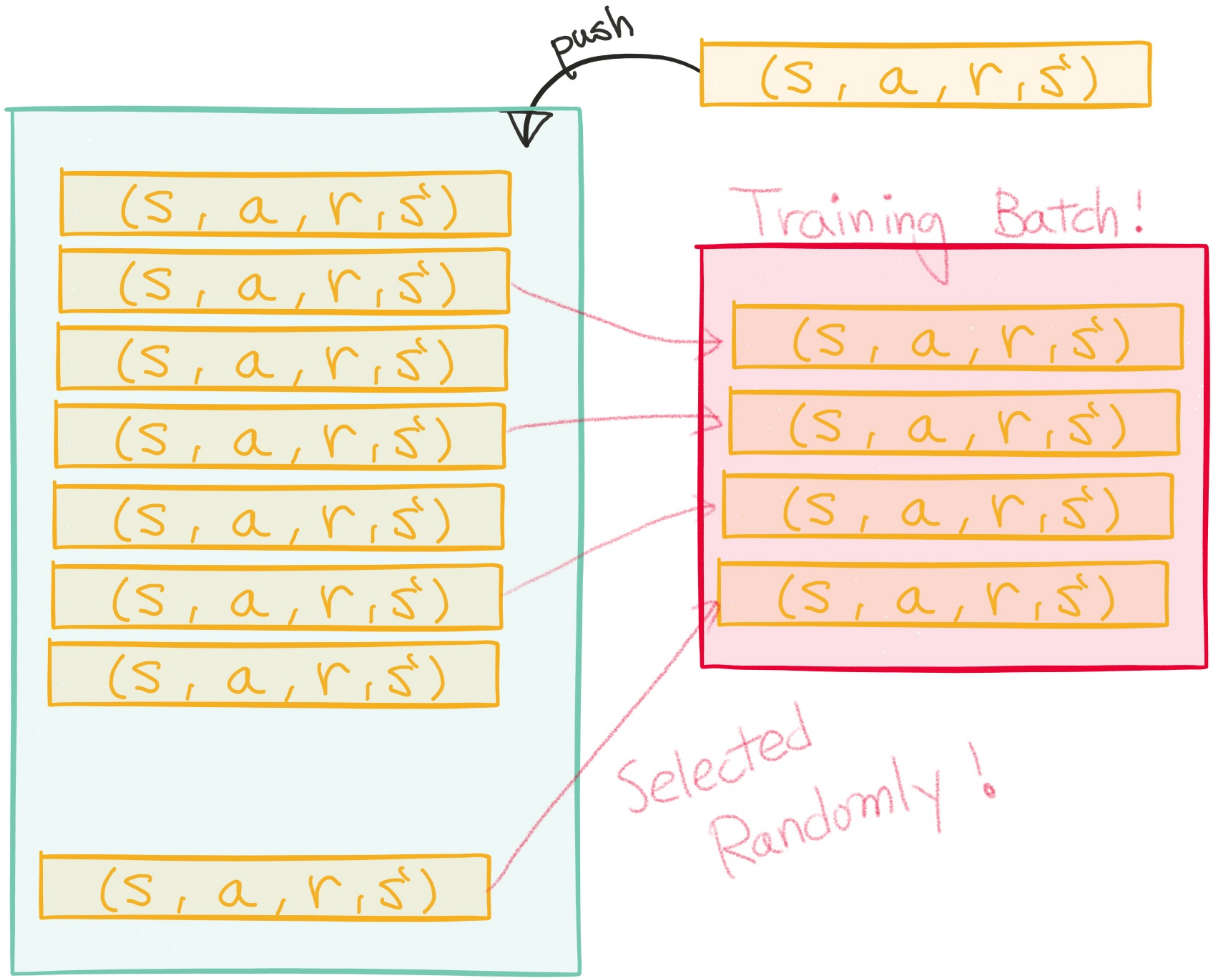
Q Learning Algorithm



Deep Q Learning



Experience Replay



Experience Replay