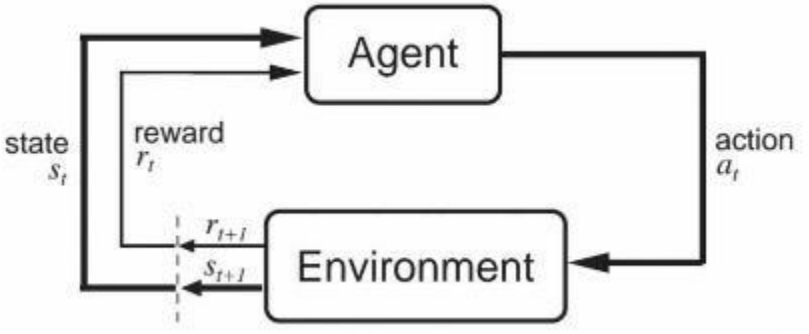
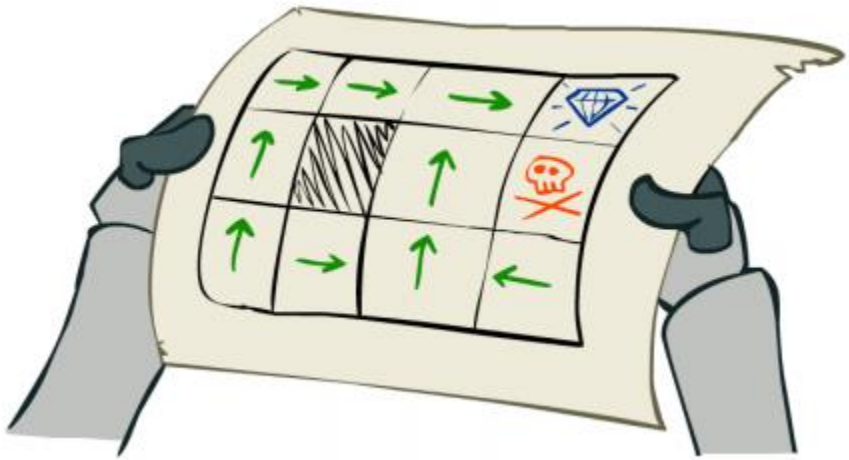
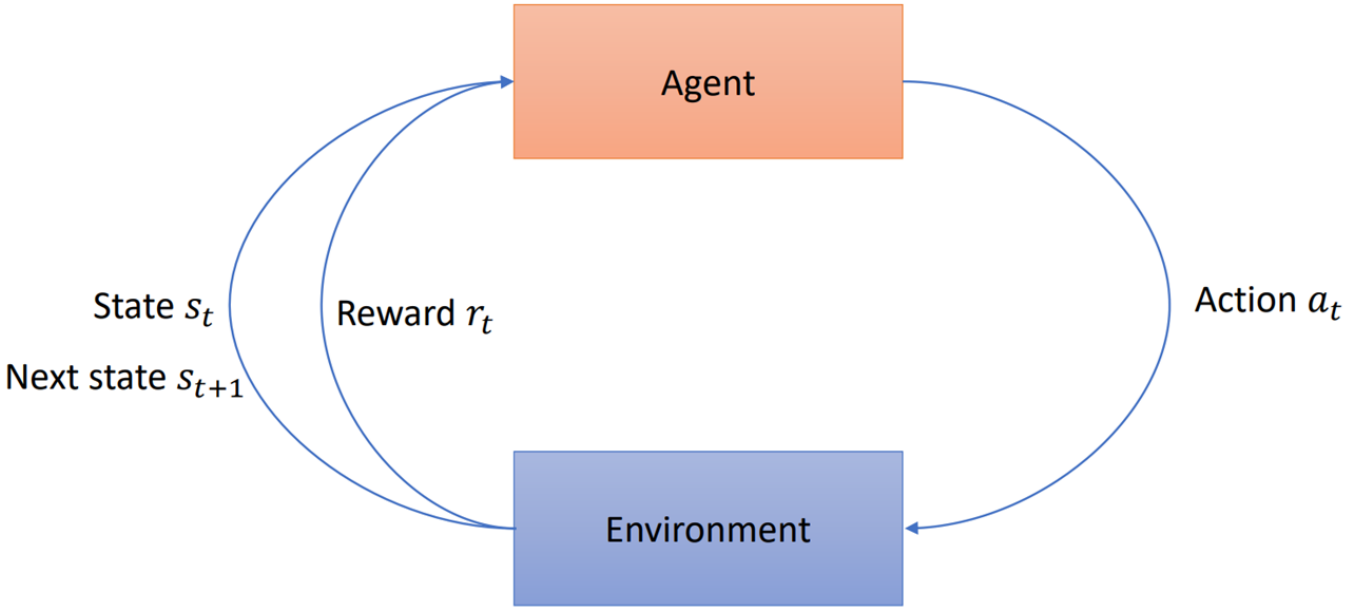
강화학습
Model-free Reinforcement Learning





강화학습



$$s_0, a_0, r_0, s_1, a_1, r_1, \ldots$$

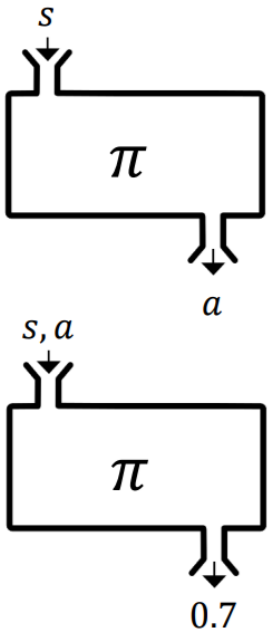
(Infinite-horizon) MDP

$$\{S, A_s, p(\cdot\,|s,a), r(s,a), \gamma\colon\ s \in S, a \in A_s\}$$

- S: 상태공간 (state space)
- A_s: 행동공간 (action space)
- $p(s'|s,a)$: 상태전이확률 (state transition probabilities)
- $r(s,a)$: 보상 (rewards)
- γ: 감가율 (discount factor)

(Infinite-horizon) MDP

- **정책 (Policy) π**
  - 상태(state)를 행동(action)에 맵핑 (mapping)
  - 확정적 정책 (Deterministic policy)
    - $\pi(s) \rightarrow a$
  - 확률적 정책 (Stochastic policy)
    - $\pi(a|s) = P(A_t = a|S_t = s)$



(Infinite-horizon) MDP

- **최적 정책 (optimal policy) π^*란?**
  - 현 보상 최대화?
  - 모든 미래 보상들의 합 최대화?
  - **감가율이 반영된 모든 미래 보상합의 최대화!**

$s_0, a_0, r_0, s_1, a_1, r_1, \ldots$  $\boldsymbol{\pi^* = \arg\max_{\pi} E\left[\sum_{t\geq 0} \gamma^t r_t \,|\pi\right]}$

$a_t \sim \pi(a|s_t)$

$s_{t+1} \sim p(\cdot\,|s_t, a_t)$

(Infinite-horizon) MDP

- **정책 평가 (Policy evaluation)**
  - 정책 π가 주어졌을 때 상태 s의 가치함수 $v^{\pi}(s)$
  - 벨만 기대 방정식
    - (확정적 정책) $v^{\pi}(s) = r(s,\pi(s)) + \gamma \sum_{s'} P(s'|s,\pi(s))\, v^{\pi}(s')$
    - (확률적 정책) $v^{\pi}(s) = \sum_{a \in A} \pi(a|s) \left[r(s,a) + \gamma \sum_{s'} P(s'|s,a)\, v^{\pi}(s') \right]$

$$v^{\pi}(s) = E\left[\sum_{t \geq 0} \gamma^t r_t \,|s_0 = s, \pi\right]$$

- **벨만 최적 방정식**

$$v^*(s) = \max_{\pi} v^{\pi}(s) \quad \Rightarrow \quad v^*(s) = \max_{a} \left\{ r(s,a) + \gamma \sum_{s'} P(s'|s,a)\, v^*(s') \right\}$$

(Infinite-horizon) MDP

- **최적 가치함수와 최적 정책**
  - 가치 반복 (value iteration)

$$v^*(s) = \max_a \left\{ r(s,a) + \gamma \sum_{s'} P(s'|s,a)\, v^*(s') \right\}$$

$$v_0 \equiv 0 \rightarrow v_1 \rightarrow v_2 \rightarrow \cdots$$

  - 정책 반복 (policy iteration)

$$\pi_0 \rightarrow v_0 \rightarrow \pi_1 \rightarrow v_1 \rightarrow \pi_2 \rightarrow v_2 \rightarrow \cdots$$

예측 (prediction)

제어 (control)

강화학습

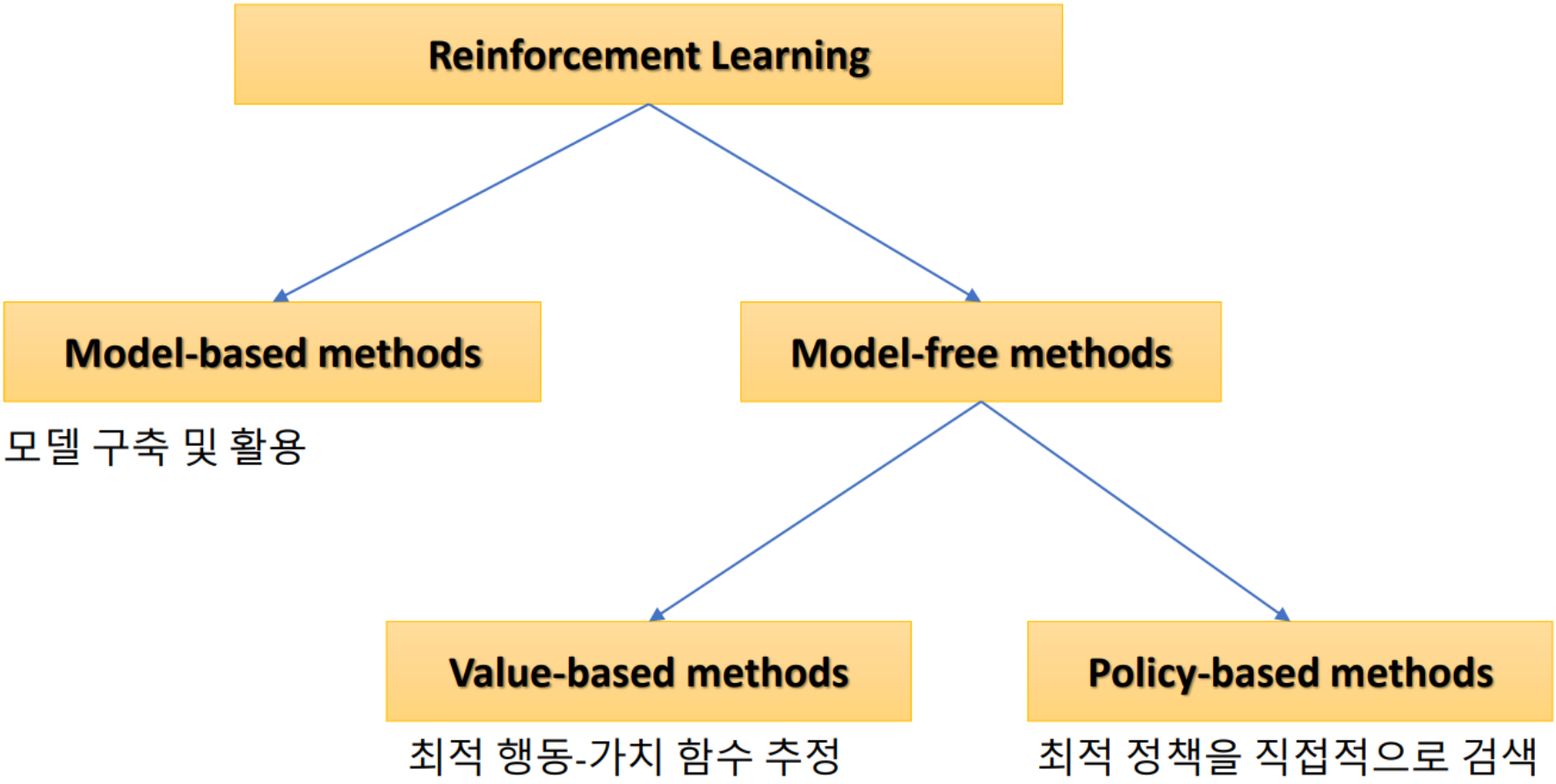
- **모델(환경에 대한 가정)을 모를 때 학습하는 방법**

$$\{S, A_s, p(\cdot\,|s,a), r(s,a), \gamma\colon\ s \in S, a \in A_s\}$$

  - 즉, $p(\cdot\,|s,a)$와 $r(s,a)$를 모를 경우, 경험을 통한 학습

- **매우 단순한 모델기반 강화학습**
  - 이러한 모델을 추정할 수 있게 되면 이전에 살펴보았던 MDP 모델을 해결함으로써 가치함수를 추정할 수 있으며 정책을 도출할 수 있게 됨

강화학습 방법론 분류


Reinforcement Learning
Model-based methods
Model-free methods
모델 구축 및 활용
Value-based methods
Policy-based methods
최적 행동-가치 함수 추정
최적 정책을 직접적으로 검색