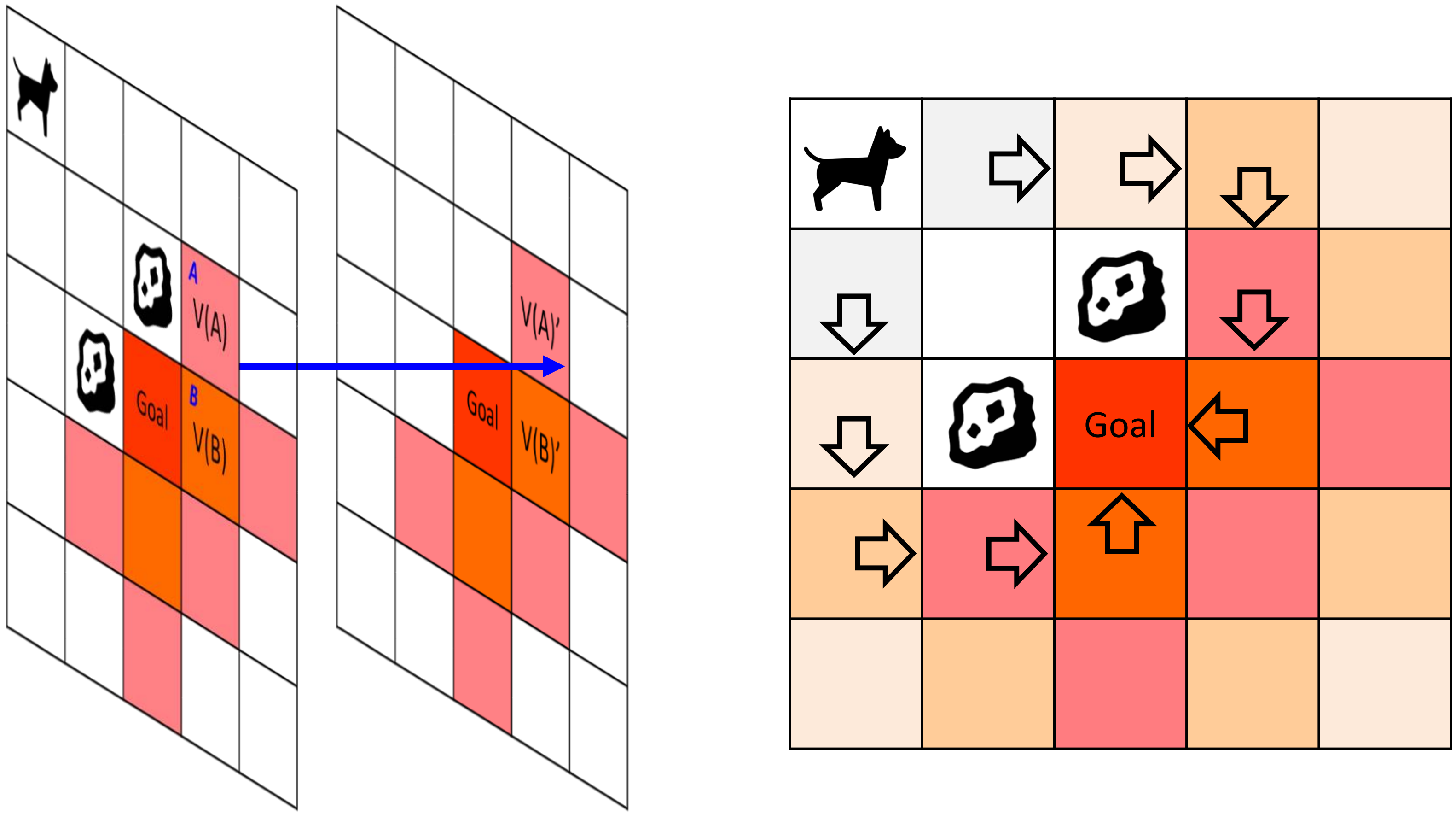
Markov Decision Process

Sangkeun Jung

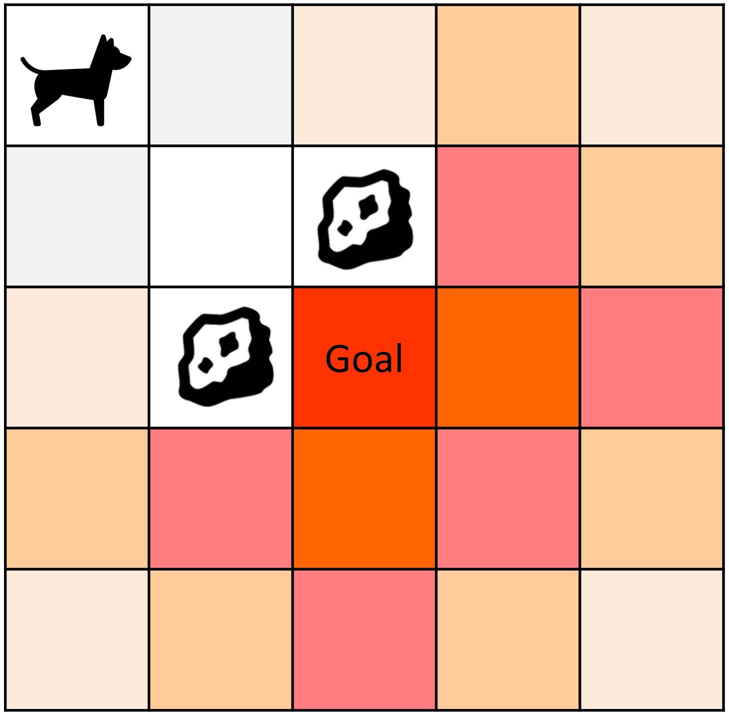
Last class,



We implemented simple value-update method to find path.

What is the limitation of the method?

Assumptions what we made



✓ **Agent는 Goal 의 Value를 알고 있다.**

➔ 가치를 알 수 없는 경우도 많고, 그 값의 scale에 따라 결과가 크게 달라질 것이다.

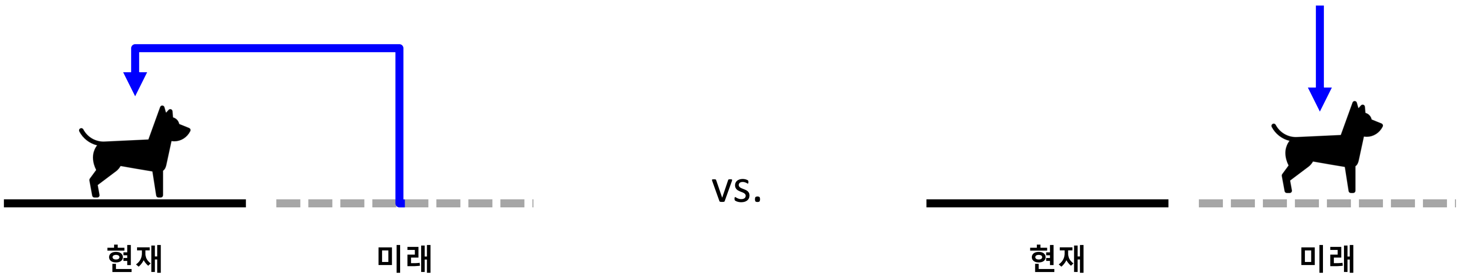
✓ **Agent는 땅을 옮기기 전에, 옮기고 나서의 땅이 함정인지 아닌지 알고 있다.**

➔ 실제로 수행해보지 않고는 모르는 경우가 대부분이다.

✓ **Agent는 모든 땅의 존재를 알고 있다.**

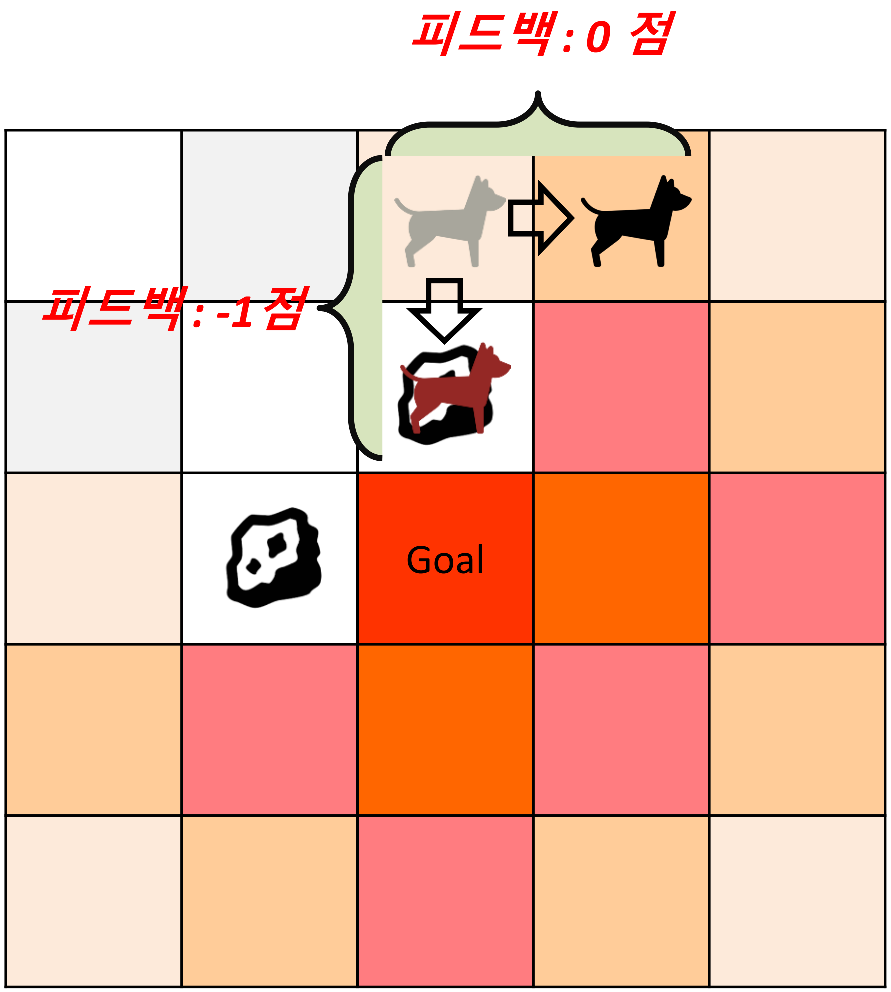
➔ 갈 수 있는 땅의 존재를 알기도 힘들고, 그 수가 무한대에 가까울 것

수행전에 가치를 알고 결정 ➔ **수행 후 피드백을 받도록**



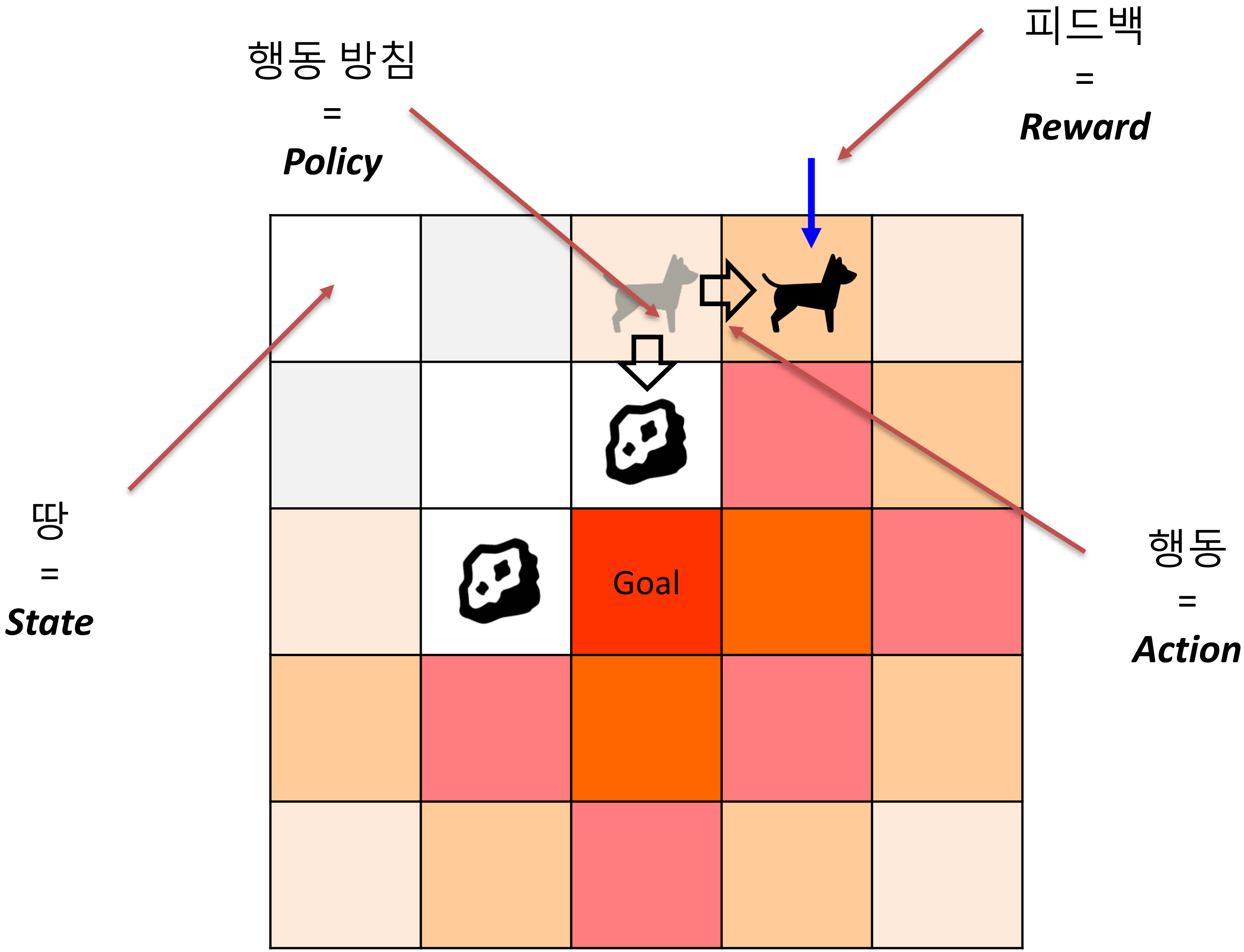
현재
미래
vs.
현재
미래

관점 전환

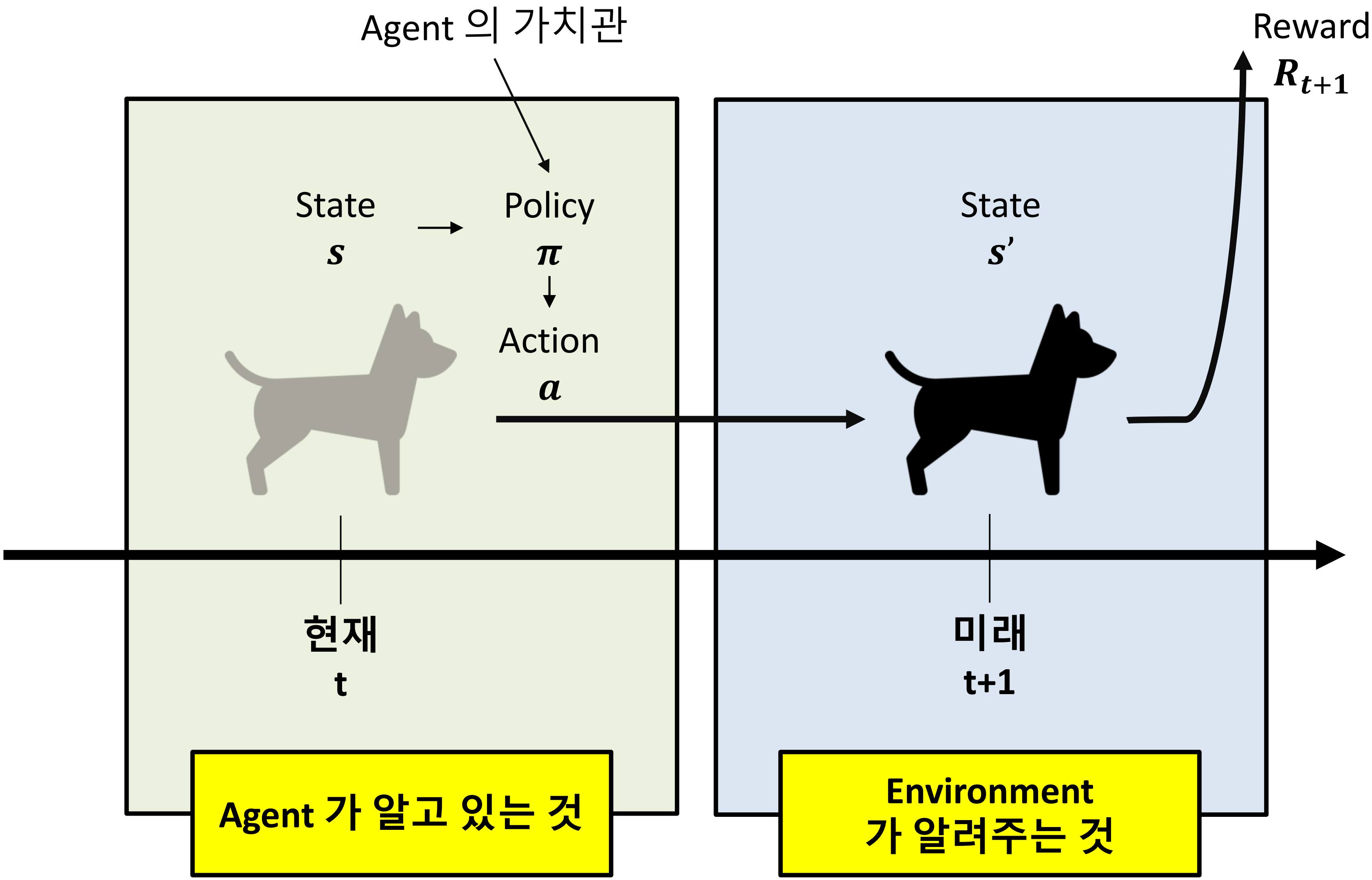


Agent 가 먼저 수행해 보고(Exploration)
Feedback 을 주는 방식

Problem Settings

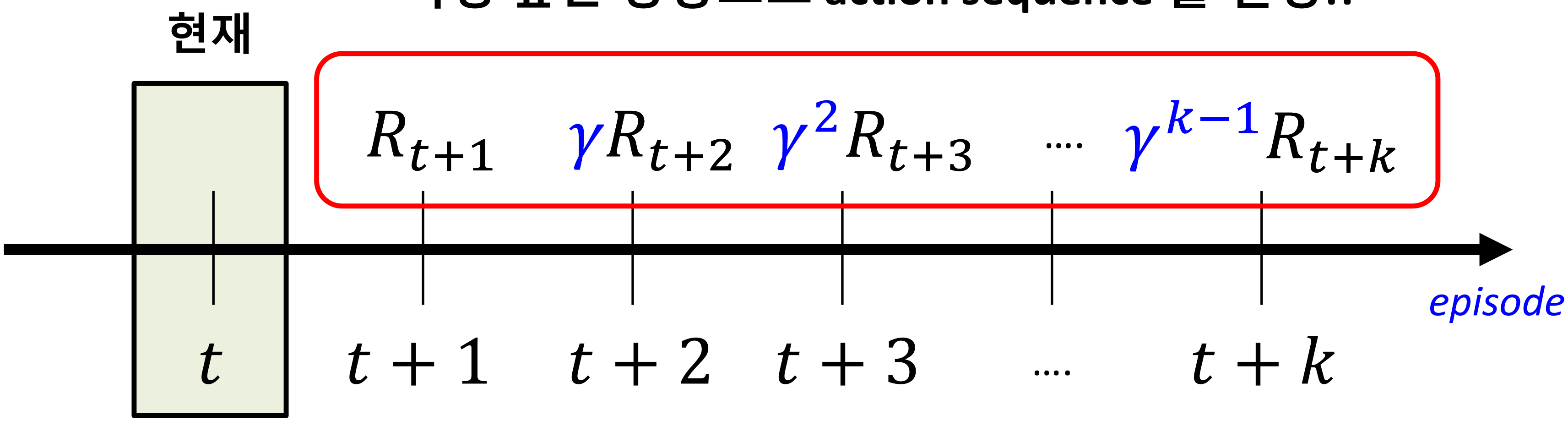


Problem Settings



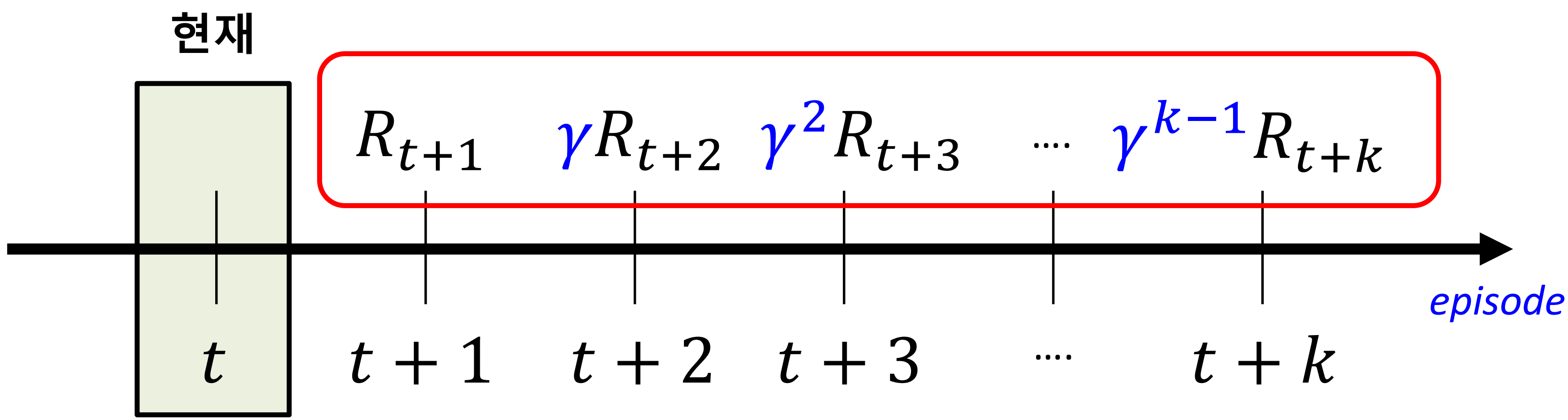
Problem Settings

모든 보상의 합이
가장 높은 방향으로 action sequence 를 결정!!



초기의 결정이 더 큰 효과를 일으키게 되므로,
초기의 결정에 더 큰 보상이 올 수 있게

Concept | *Return*



G_t : Agent 가 t 시점이후를 탐험하면서 얻은 총 보상의 합

$$G_1 = R_2 + \gamma R_3 + \gamma^2 R_4 + \gamma^3 R_5 + \gamma^4 R_6$$

$$G_2 = R_3 + \gamma R_4 + \gamma^2 R_5 + \gamma^3 R_6$$

$$G_3 = R_4 + \gamma R_5 + \gamma^2 R_6$$

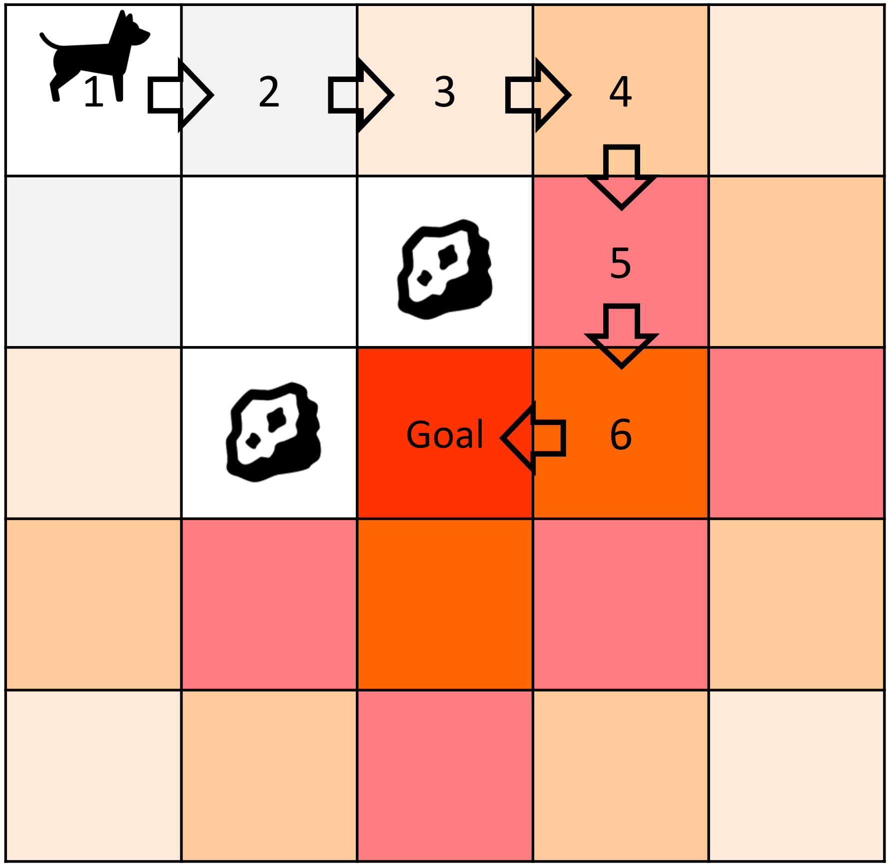
$$G_4 = R_5 + \gamma R_6$$

$$G_5 = R_6$$

1 Episode

$$G_1 = R_2 + \gamma R_3 + \gamma^2 R_4 + \gamma^3 R_5 + \gamma^4 R_6 + \gamma^5 R_7$$

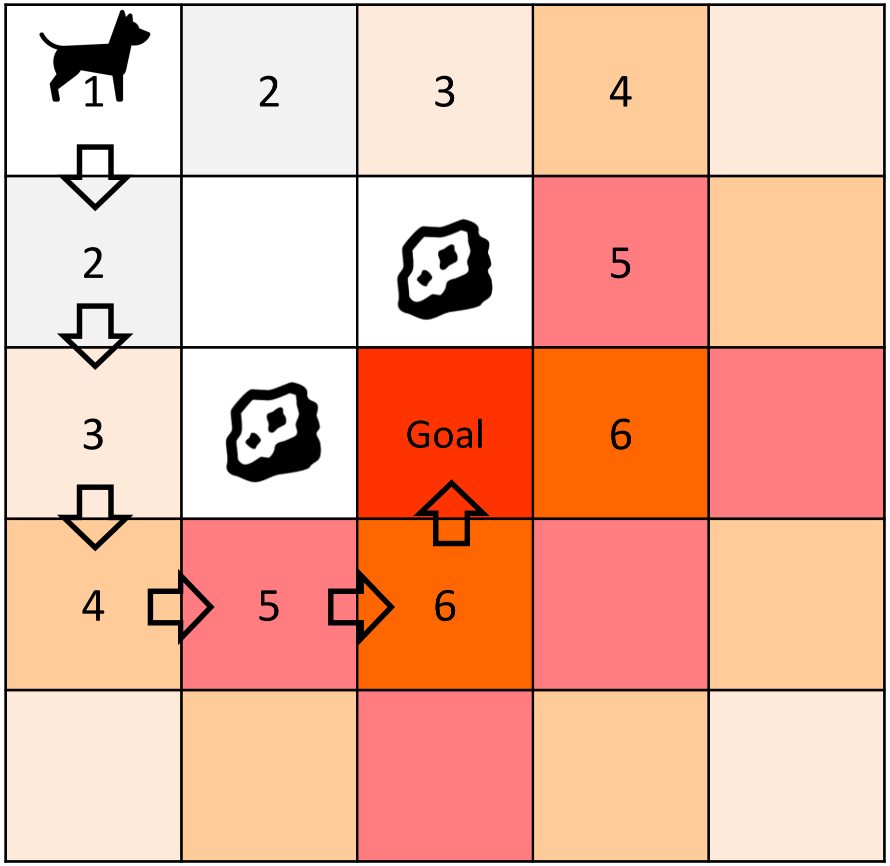
Agent 가 아래의 Path 를 따라갔을 때의 총 보상 값



1 Episode

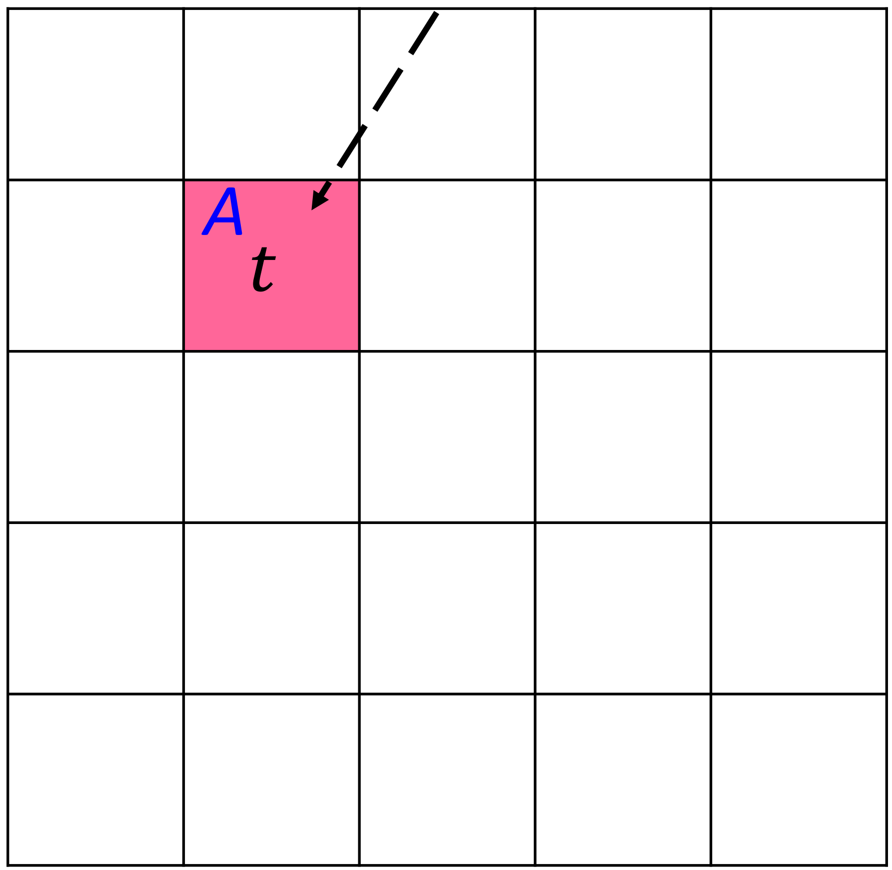
$$G_1 = R_2 + \gamma R_3 + \gamma^2 R_4 + \gamma^3 R_5 + \gamma^4 R_6 + \gamma^5 R_7$$

Agent 가 아래의 Path 를 따라갔을 때의 총 보상 값

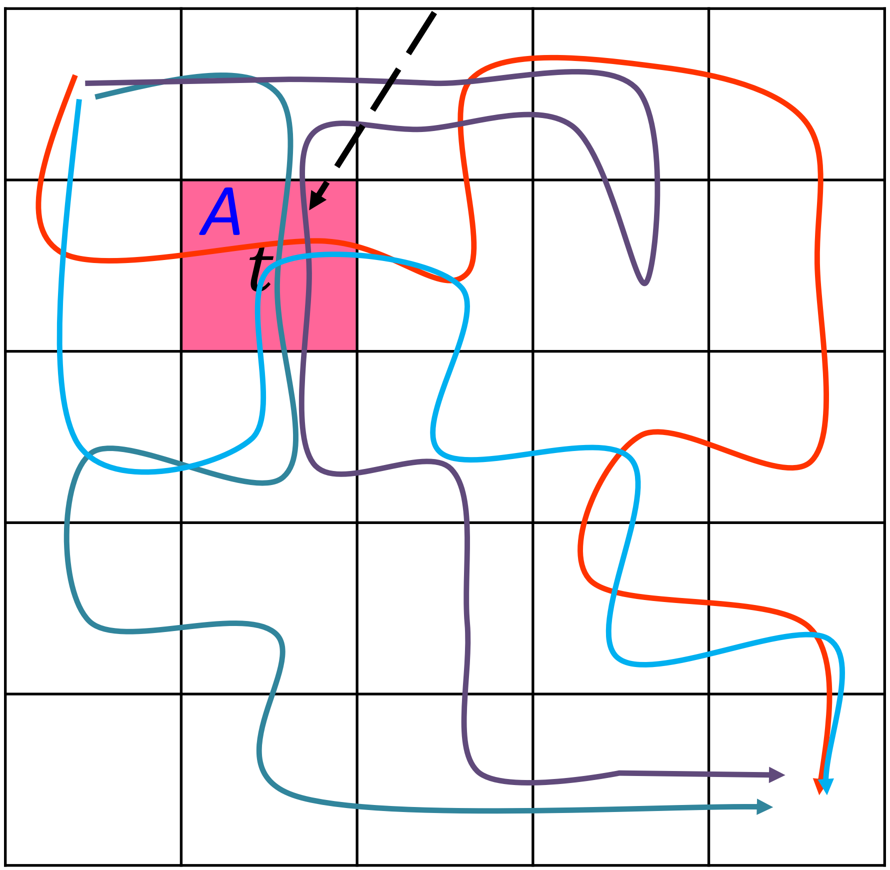


[Question : 5 min]

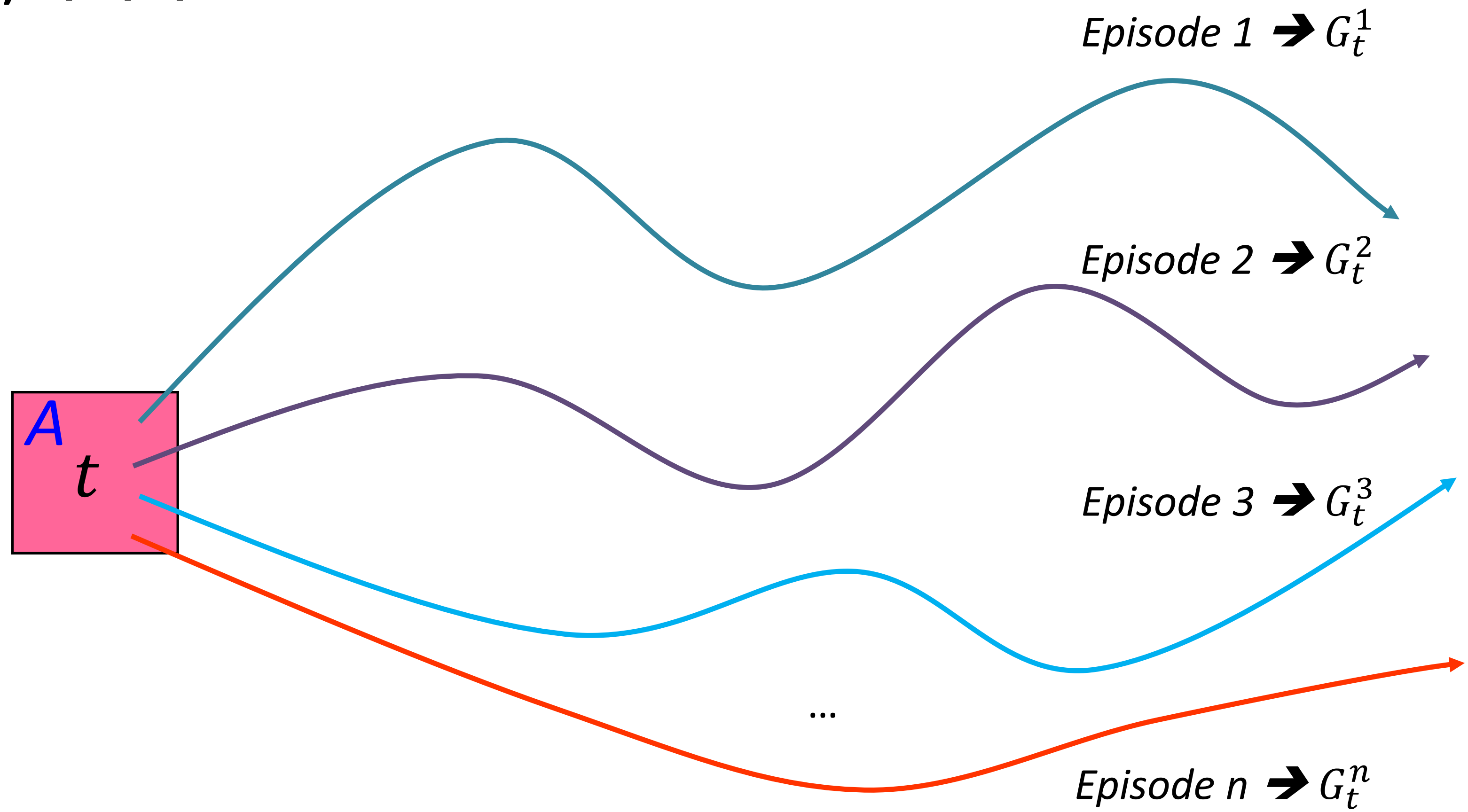
시간 t 에서의, 땅 A의 가치는 얼마일까?



시간 t 에서의, 땅 A의 가치는 얼마일까?



State (땅) 의 가치



$$v(A) \approx \frac{1}{n}\left(\sum_{k}^{n} G_t^k\right)$$

시간 t 에서, 해당 state를 거쳐가는 모든 Return 값들의 평균이
*땅의 가치*라고 생각해보면 어떨까?

가치 기대값

모든 에피소드를 다 해봐야만 Return 값이 나오기 때문에
위의 방법은 **현실성**이 없다.

즉 실제로 모두 해보지 않고, '확률'적인 기대값을 활용하면
State의 가치를 아래처럼 Estimation 해 볼 수 있다.

$$v(s) = E[G_t | S_t = s]$$

땅(State) s 의 가치 기대값

기대값(Expectation)

$$1 * \frac{1}{6} + 2 * \frac{1}{6} + 3 * \frac{1}{6} + 4 * \frac{1}{6} + 5 * \frac{1}{6} + 6 * \frac{1}{6} = 3.5$$

$$E[X] = \sum_i p_i x_i$$

확률　가치

Value function

$$v(s) = E[G_t | S_t = s]$$

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

$$v(s) = E[R_{t+1} + \underline{\gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots} | S_t = s]$$

$$v(s) = E[R_{t+1} + \underline{\gamma (R_{t+2} + \gamma R_{t+3} + \cdots})| S_t = s]$$

$$v(s) = E[R_{t+1} + \gamma \underline{G_{t+1}} | S_t = s]$$

G_{t+1}는 S_{t+1} 라는 땅을 거쳐갔을 때의 전체 보상값이 되므로. 이 값을 S_{t+1}의 땅의 가치라고 생각할 수 있다.

$$v(s_t) = E[R_{t+1} + \gamma v(s_{t+1}) | S_t = s]$$

Bellman Expectation Equation

$$v(s) = E[R_{t+1} + \gamma v(s_{t+1}) | S_t = s]$$

땅의 가치와 그 가치에 기반한 행동방침(Policy) 에 따라 전체 Episode 를 진행할 것이기 때문에, 위 수식은 엄밀하게 아래처럼 표현된다.

$$\boxed{v_\pi(s) = E_\pi[R_{t+1} + \gamma v_\pi(s_{t+1}) | S_t = s]}$$

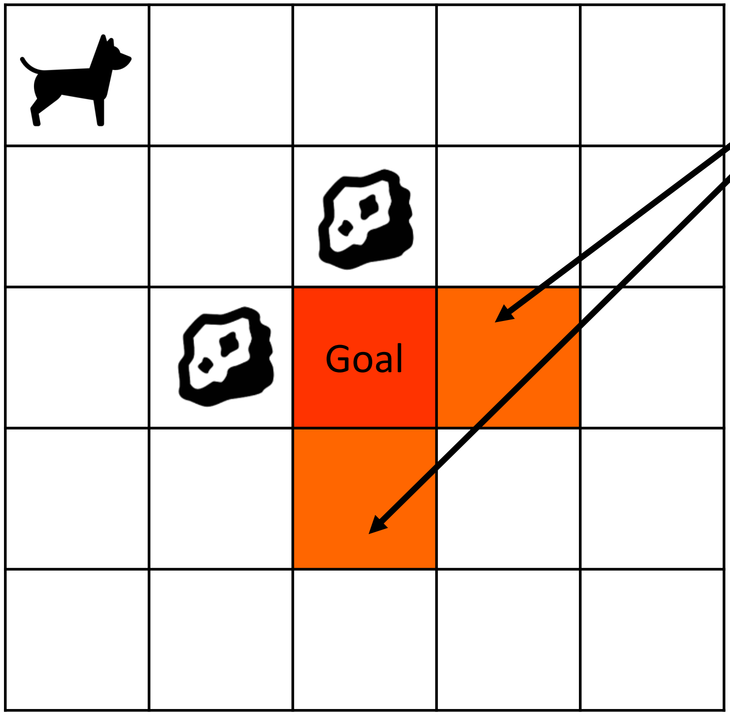
!!!!! The Most Important Formula in RL !!!!!

[Question : 5min]
Why is it so important?

$$v_{\pi}(s) = E_{\pi}[R_{t+1} + \gamma v_{\pi}(s_{t+1}) | S_t = s]$$

(remind)

(remind) How update value?

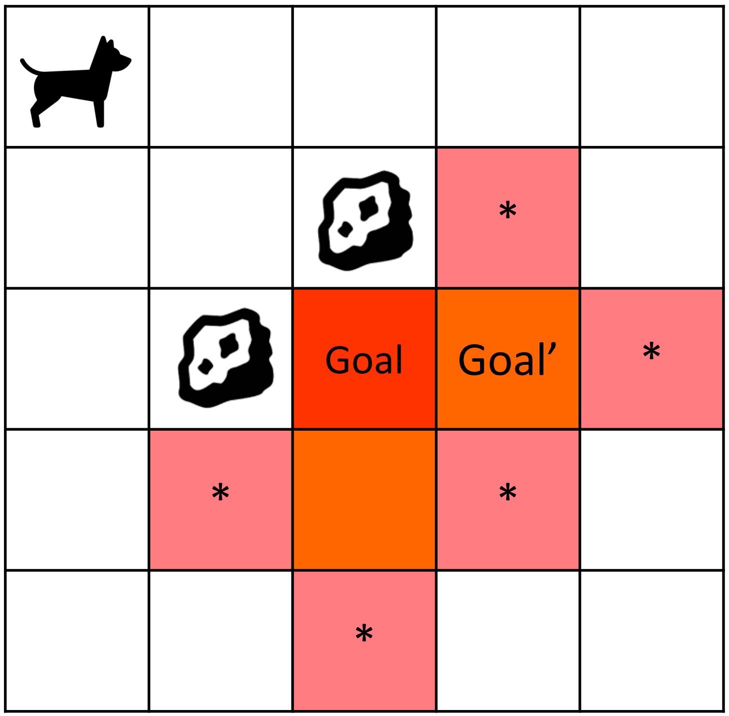


이 두 땅의 가치는 매우 높아야 함
(그러나 Goal 보다는 낮아야 함)

이 두 땅의 가치는 Goal 에 도착할 수 있기 때문

그렇다면, 저 두 땅에만 도착할 수 있다면
Goal 에 도착할 가능성이 올라감

(remind) How update value?

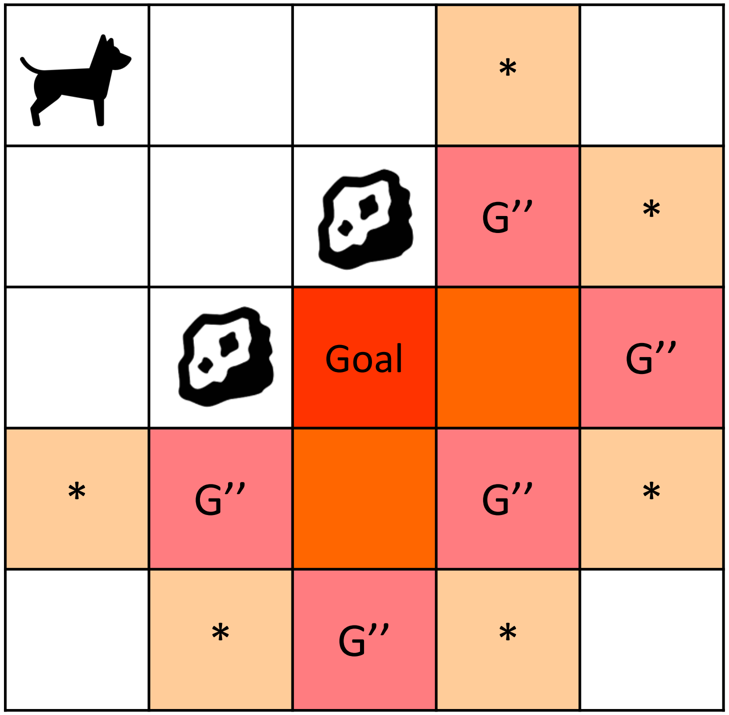


* 표시를 한 땅은 가치가 다른 땅보다 높아야 함
(그러나 Goal’보다는 낮아야 함)

* 땅을 거쳐야만, Goal’ 에 도착할 수 있기 때문

(반복) * 땅에 도착할 수 있다면
Goal 에 도착할 가능성이 올라감

(remind) How update value?



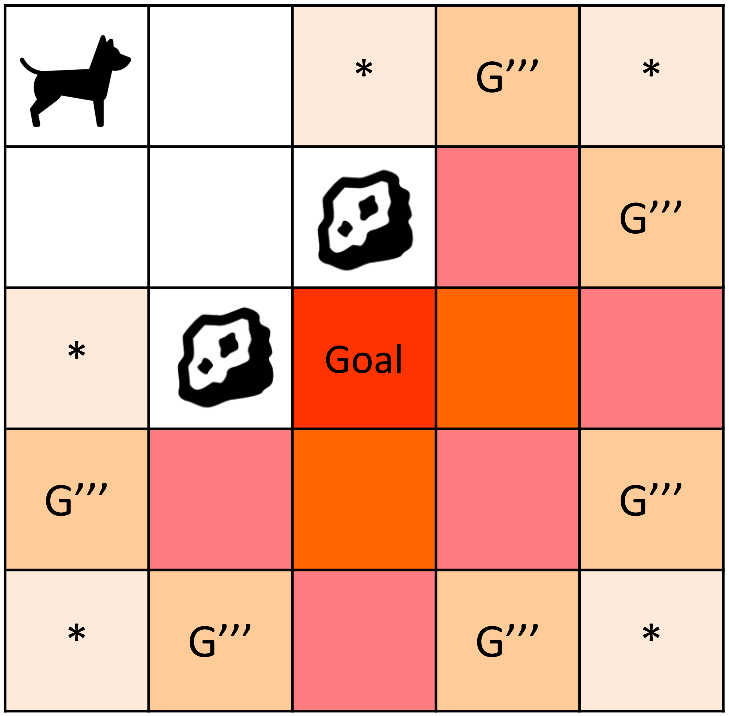
G’’ = Goal’’

* 표시를 한 땅은 가치가 다른 땅보다 높아야 함
(그러나 Goal’’ 보다는 낮아야 함)

* 땅을 거쳐야만, Goal’’ 에 도착할 수 있기 때문

(반복) * 땅에 도착할 수 있다면
Goal 에 도착할 가능성이 올라감

(remind) How update value?



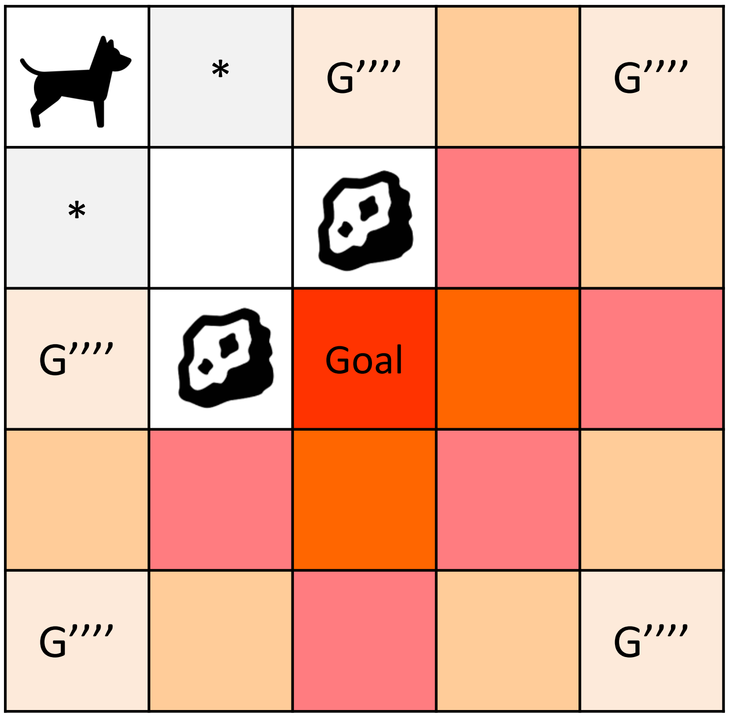
G’’’ = Goal’’’

* 표시를 한 땅은 가치가 다른 땅보다 높아야 함
(그러나 Goal’’’ 보다는 낮아야 함)

* 땅을 거쳐야만, Goal’’’ 에 도착할 수 있기 때문

(반복) * 땅에 도착할 수 있다면
Goal 에 도착할 가능성이 올라감

(remind) How update value? (final)



G’’’’ = Goal’’’’

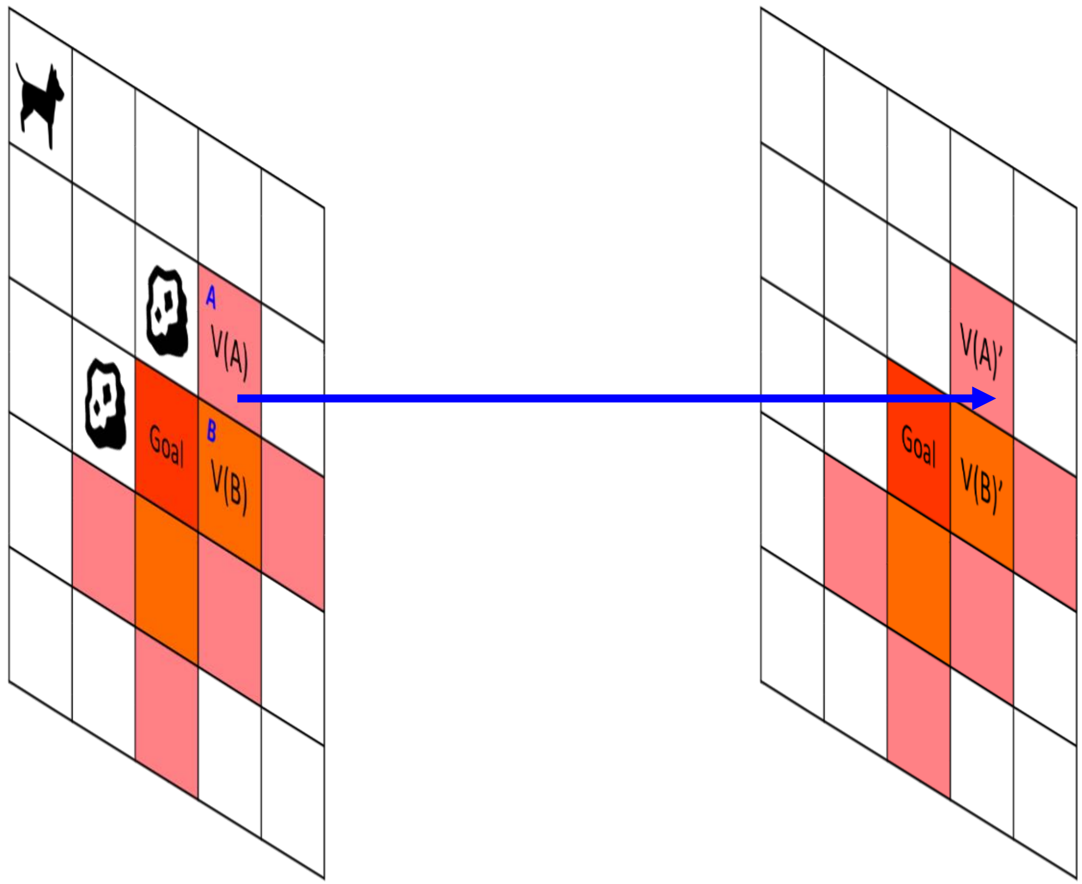
* 표시를 한 땅은 가치가 다른 땅보다 높아야 함
(그러나 Goal’’’’ 보다는 낮아야 함)

* 땅을 거쳐야만, Goal’’’’ 에 도착할 수 있기 때문

(반복) * 땅에 도착할 수 있다면
Goal 에 도착할 가능성이 올라감

(remind) Value Update | V(A) → V(A)’

A,B는 각각
땅의 이름



땅 A에서 <u>갈 수 있는 모든 땅 중</u>,
가장 가치가 높은 땅의 가치를,
땅 A의 가치로 Update

$$V(A)' = \gamma V(B)$$

그러나, 땅 B 보다는 조금 가치가 낮게

Discount factor: 0.0 ~ 1.0

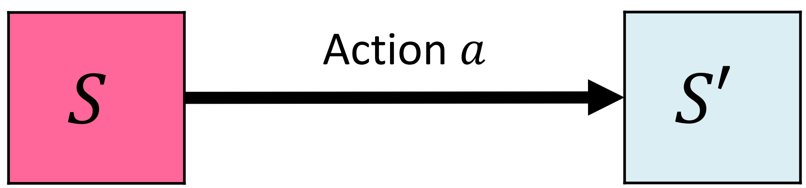
$$v_\pi(s) = E_\pi[R_{t+1} + \gamma v_\pi(s_{t+1}) | S_t = s]$$

$$v_\pi(s_t) = \cdots\ v_\pi(s_{t+1}) \cdots$$

$$v_\pi(s) = \cdots\ v_\pi(s') \cdots$$

땅 s 의 가치

땅 s' 의 가치

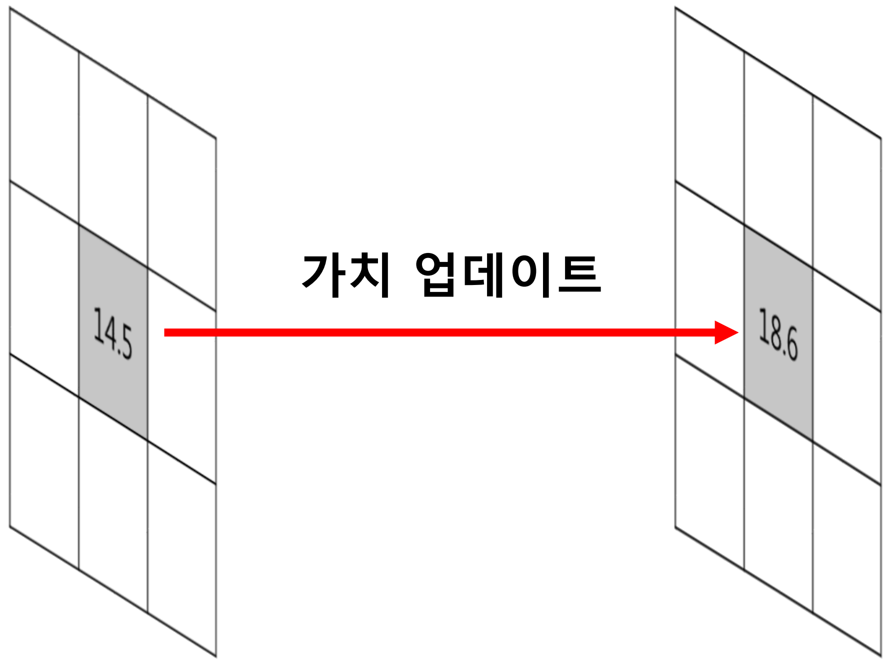


현재(t)

미래(t+1)

땅 s 의 가치를 Agent 가 행동한 후 가게 되는 땅 s' 의 가치를 이용해서 계산(Update) 할 수 있다.

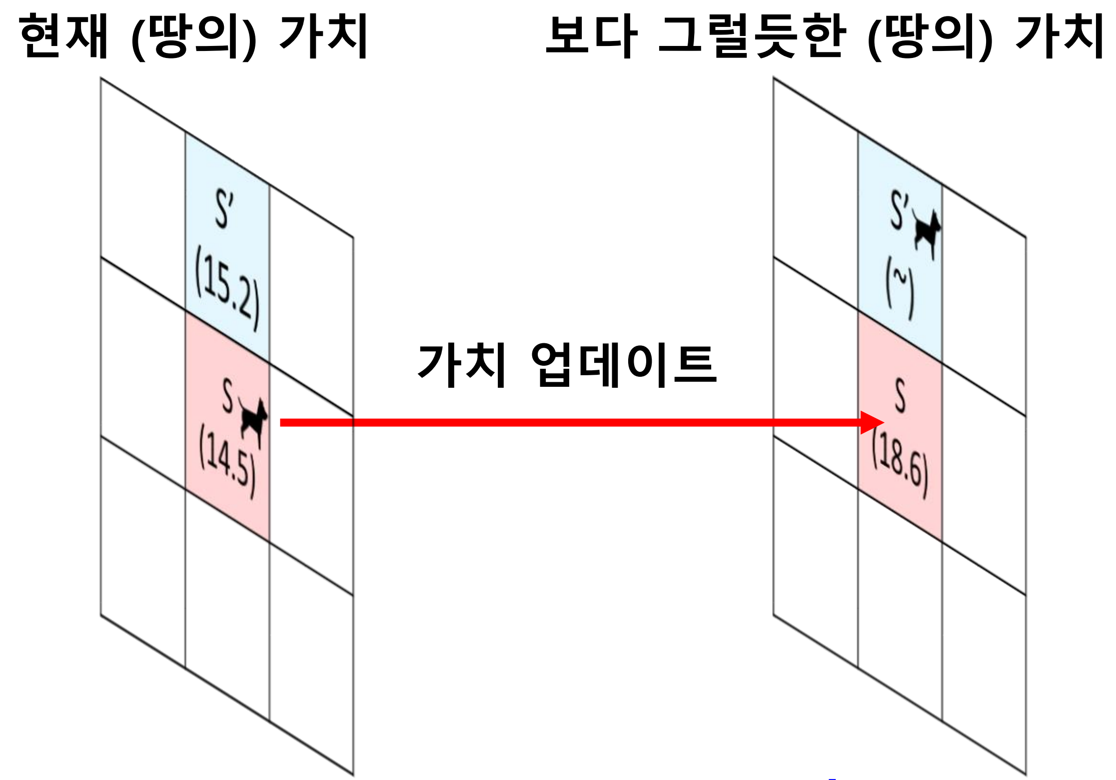
(Remind)



현재 (땅의) 가치　　　**보다 그럴듯한 (땅의) 가치**

- 최초에는 땅의 가치를 모르기 때문에(동등, uniform)
- 목표점까지 갈 수 있는 땅들이 높은 값을 가질 수 있게
- 가치를 업데이트 할 수 있는 방법이 필요하다.

Bellman Equation for Value Update



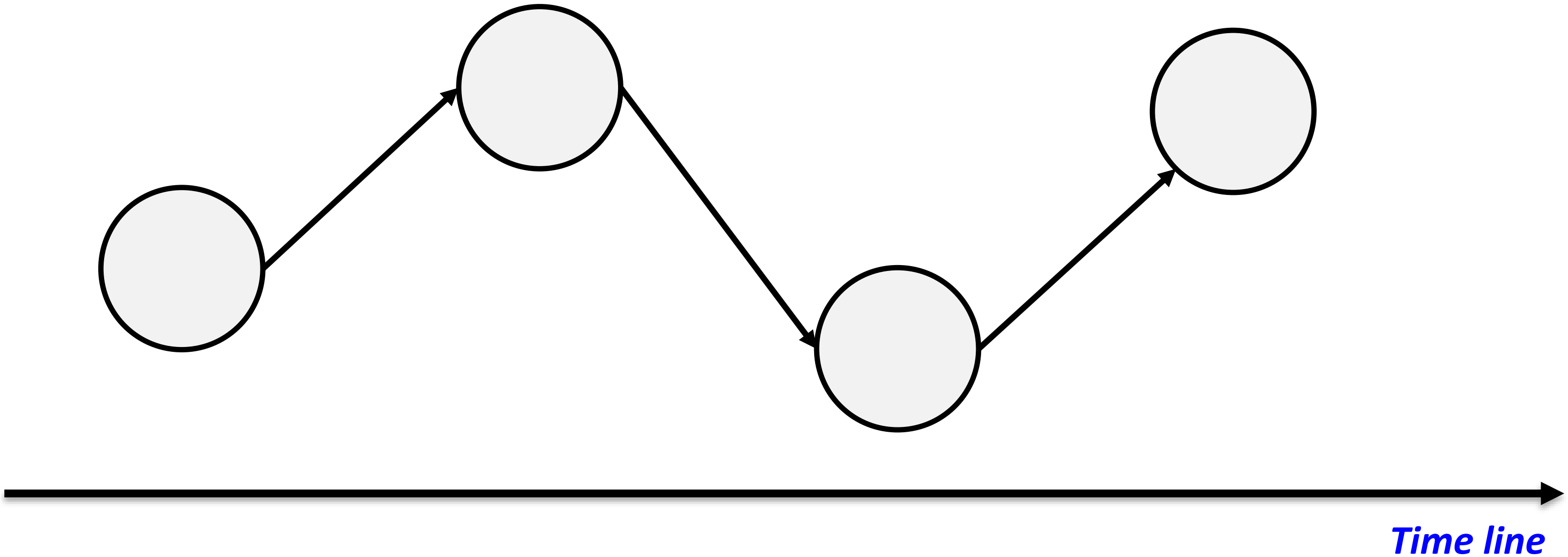
$$v_\pi(s) = \cdots\ v_\pi(s')\cdots$$

S 에 있던 Agent 가 현재의 Policy 를 활용해 S' 로 이동

S' 이후로 부터 얻게 되는 모든 보상들을 이용해 땅 S' 의 가치를 알 수 있고

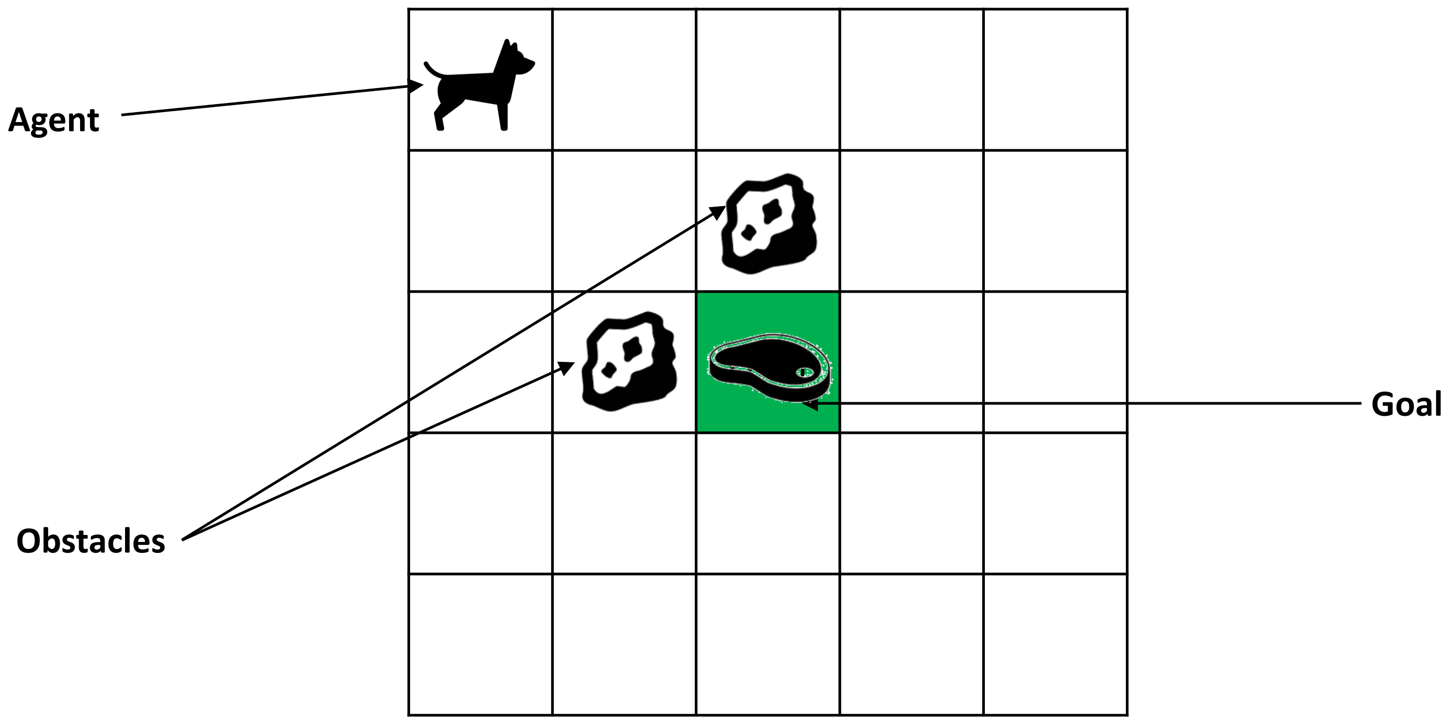
수식을 활용해, 땅 S의 가치를 땅 S' 의 가치를 통해 가치갱신 할 수 있다.

(Remind) Big Picture

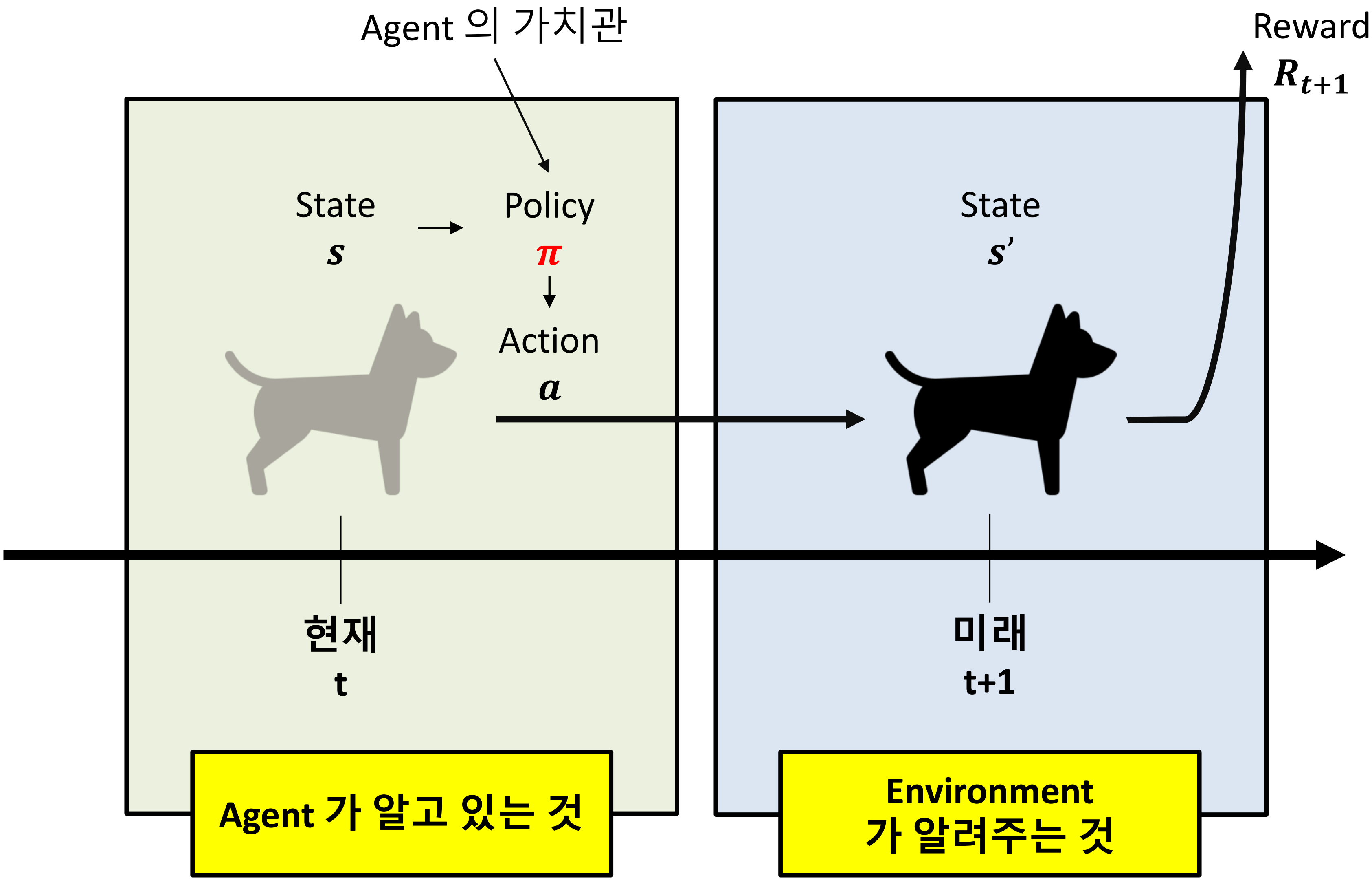


Sequential Decision Making

In artificial intelligence, sequential decision making refers to algorithms that take the dynamics of the world into consideration.


Agent
Obstacles
Goal

Problem Settings



TARGET : “이런 상황에서는 어떻게 해야하지?” 를 배우는 것

Bellman Equations

Bellman Expectation Equation

$$v_\pi(s) = E_\pi[R_{t+1} + \gamma v_\pi(s_{t+1}) | S_t = s]$$

최적의 정책을 찾은 후
이를 통해 계산된
최적 Value function

Bellman Optimality Equation

$$v_*(s) = \max_a E_\pi[R_{t+1} + \gamma v_*(s_{t+1}) | S_t = s, A_t = a]$$

Note that!

True Value ≠ Optimal Value

현재의 정책에 기반했을 때
반복적으로 계산해내어
수렴한 땅의 값

최적의 정책을 통해 구해낸
최적의 땅의 값