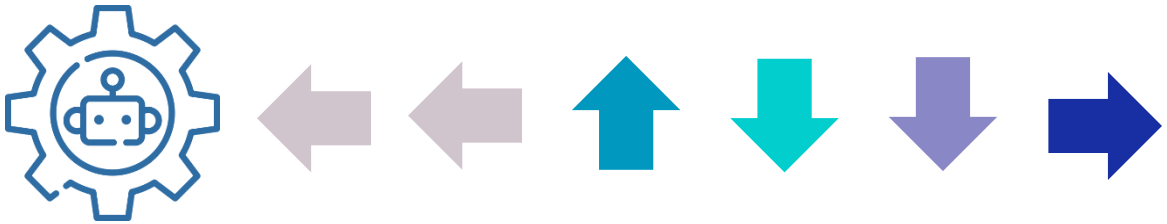
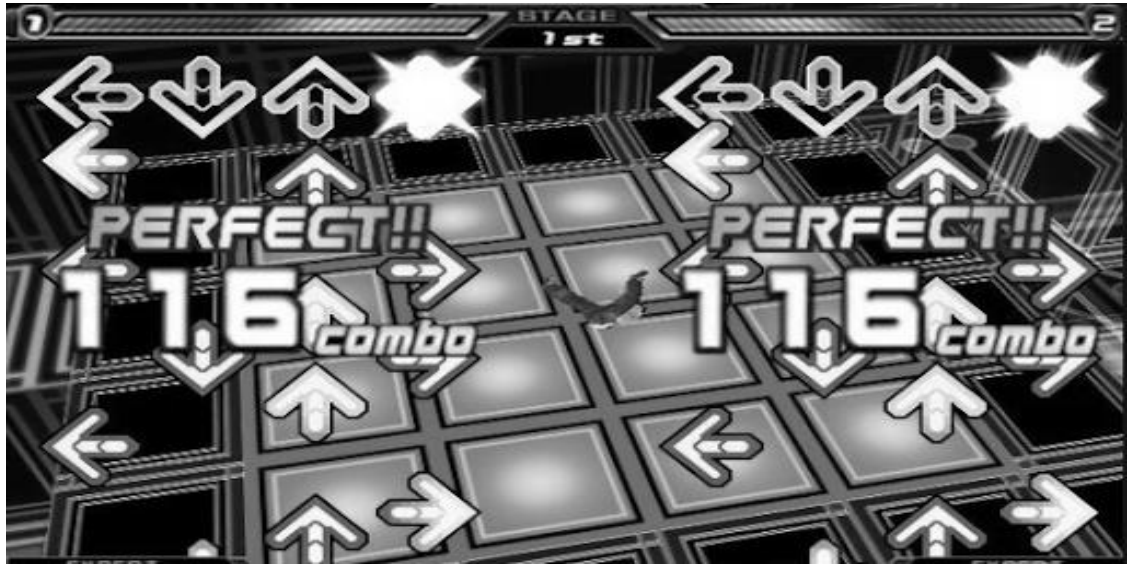
PART 6

# 강화학습

딥러닝 & 강화학습 담당
이재화 강사



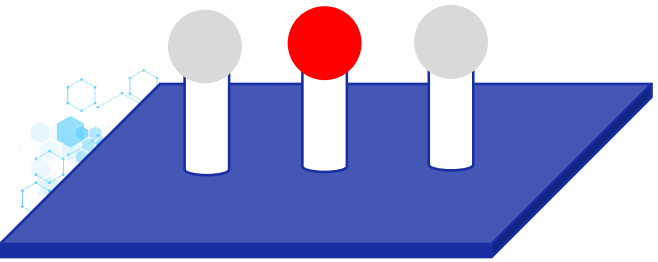
✓ 강화학습은 다른 머신러닝 혹은 딥러닝 분야와는 다르게 순차적으로 행동을 결정해야 하는 문제를 다룬다.

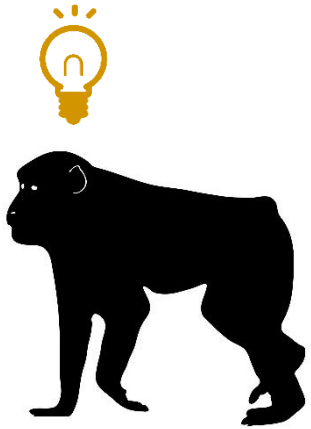



#벨만 기대 방정식

$$V_\pi(s) \\ (Bellman\ Expectation\ Equation) \\ = E_\pi\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s\right]$$

✓ 이러한 문제를 컴퓨터가 풀기 위해서는 문제를 수학적으로 정의.





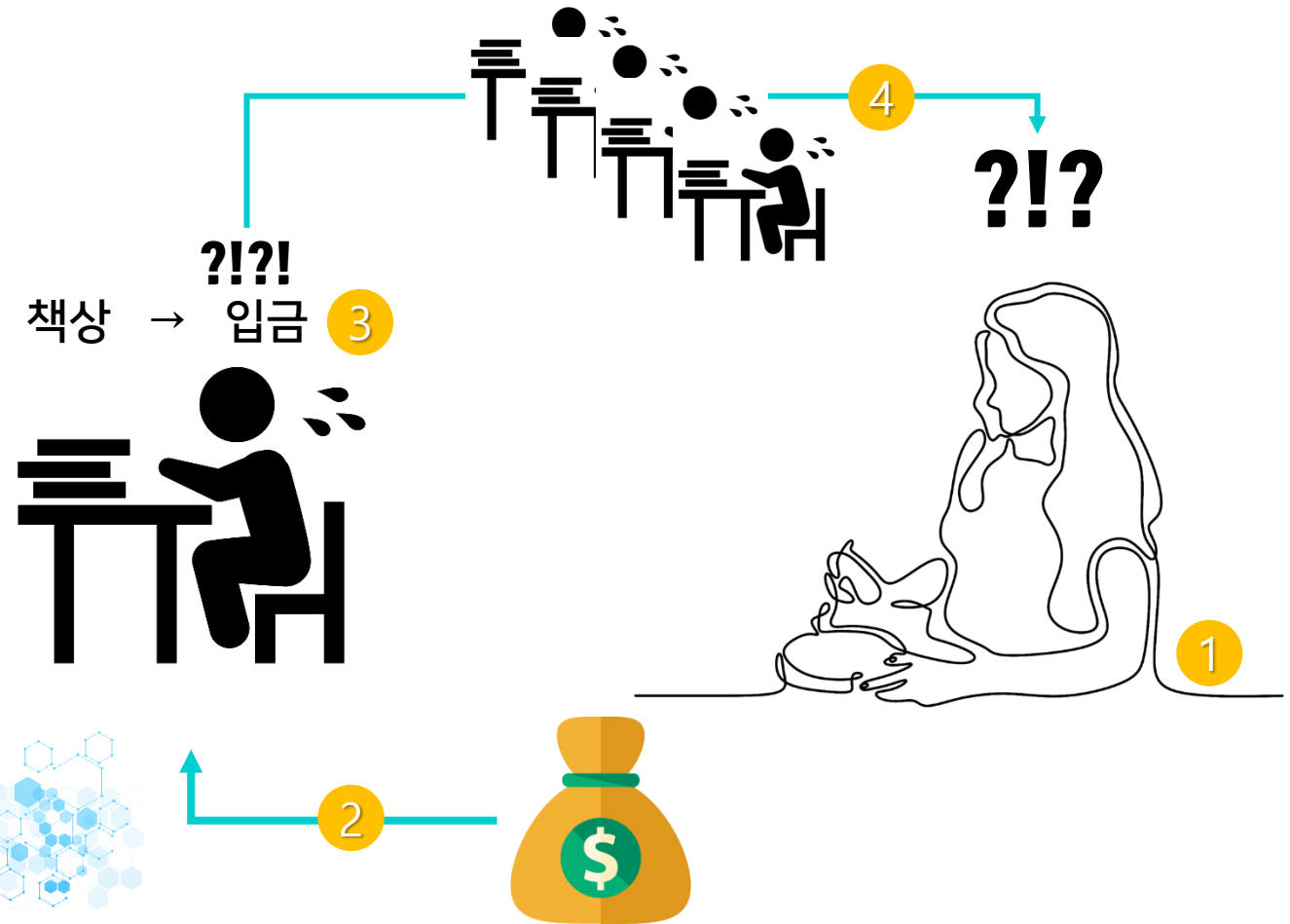
✓ 강화학습의 강화는 시행착오를 통해 학습하는 방법 중 하나

✓ 시행착오 학습은 동물들이 이것저것 시도해보면서 그 결과를 통해 학습을 수행하는 방법을 의미.

# 강화

이전에 배우지는 않았지만, 직접 시도하면서 얻게되는 결과로 행동과 보상사이의 상관관계를 학습하는 것.

그러면서 좋은 보상을 얻게 해주는 행동을 점점 더 많이 하는것을 의미.




- ✓ 왜 용돈이 들어왔는지 가늠은 할 수는 있지만, 정확하게 알 수는 없다.
- ✓ 이러한 행동을 통해 보상이 들어왔고, 이러한 행동이 보상과 연결된다는 것을 확인.

NEW!

Reinforcement learning

Unsupervised Learning?

- 주어진 데이터에 대해 학습하는것은 아님

Supervised Learning?

- 정답이 주어지는것은 아님

✓ 강화학습은 보상을 통해 학습을 수행

✓ 보상은 컴퓨터가 선택한 행동에 대한 환경의 반응

✓ 보상은 직접적인 정답은 아니지만,

컴퓨터에게는 간접적인 정답의 역할을 하게 됨.

강화학습에서는 자신의 행동의 결과로 나타나는 보상을 통해 학습을 수행

지도학습에서는 직접적인 정답을 통해 오차를 계산해서 학습



✓ 자신이 놓인 환경에서 자신의 상태를 인식한 후 행동


✓ 환경은 에이전트에게 행동에 맞는 보상을 주고
그 다음 자신이 처한 상태를 알려준다.

보상을 통해 에이전트는 어떤 행동이 좋은 행동인지 알게된다.

이 행동이 반복되며 지속적인 보상을 얻게 된다면 좋은 행동을 학습.

# 강화학습의 목적

에이전트가 환경을 탐색하면서 행동을 통해 얻는 보상들의 합을 최대화 하는

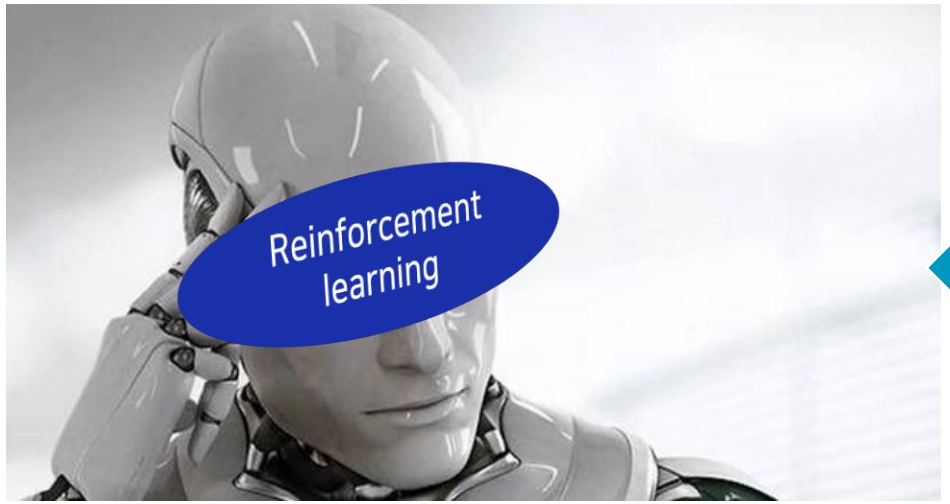
"최적의 행동양식 또는 정책"을 학습하는 것.

남은 시간(음수의 보상 방식)


남은 HP (양수의 보상 방식)


- ✓ 보상을 적절히 융합하여 효과적인 학습을 수행.
- ✓ 이 문제는 생각보다 강화학습에서 중요한 문제.
- ✓ 실질적으로 문제에 적용할 때 신중하게 고려되야할 부분.

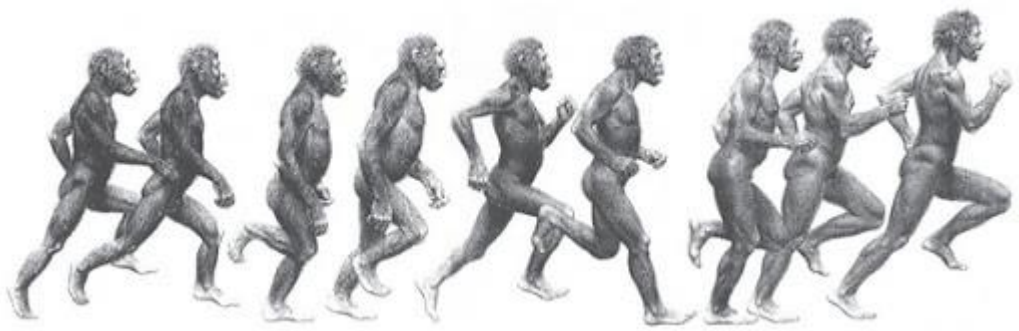
# #강화학습의 장점




생각의 전환


✓ 데이터가 없으니 당연히 학습시간이 길어짐

✓ 에이전트가 움직이는 그 모든것이 학습데이터화



#발전과정을 통한 새로운 아이디어 확보

# Part 6. 강화


- ✓ 사람처럼 환경과 상호작용하면서 스스로 학습을 하는 방식
- ✓ 문제 자체에 대해 잘 이해하지 않으면 내가 풀고자 하는 문제자체에 대해 정확한 지식을 갖고 있지 않는다면 엉뚱한 결과를 불러옴
- ✓ 강화학습은 선택 즉, 에이전트의 행동 결정 순서를 순차적으로 선택해야하는 문제에 적합

이러한 순차적 행동 결정 문제를 MDP
(Markov decision process)

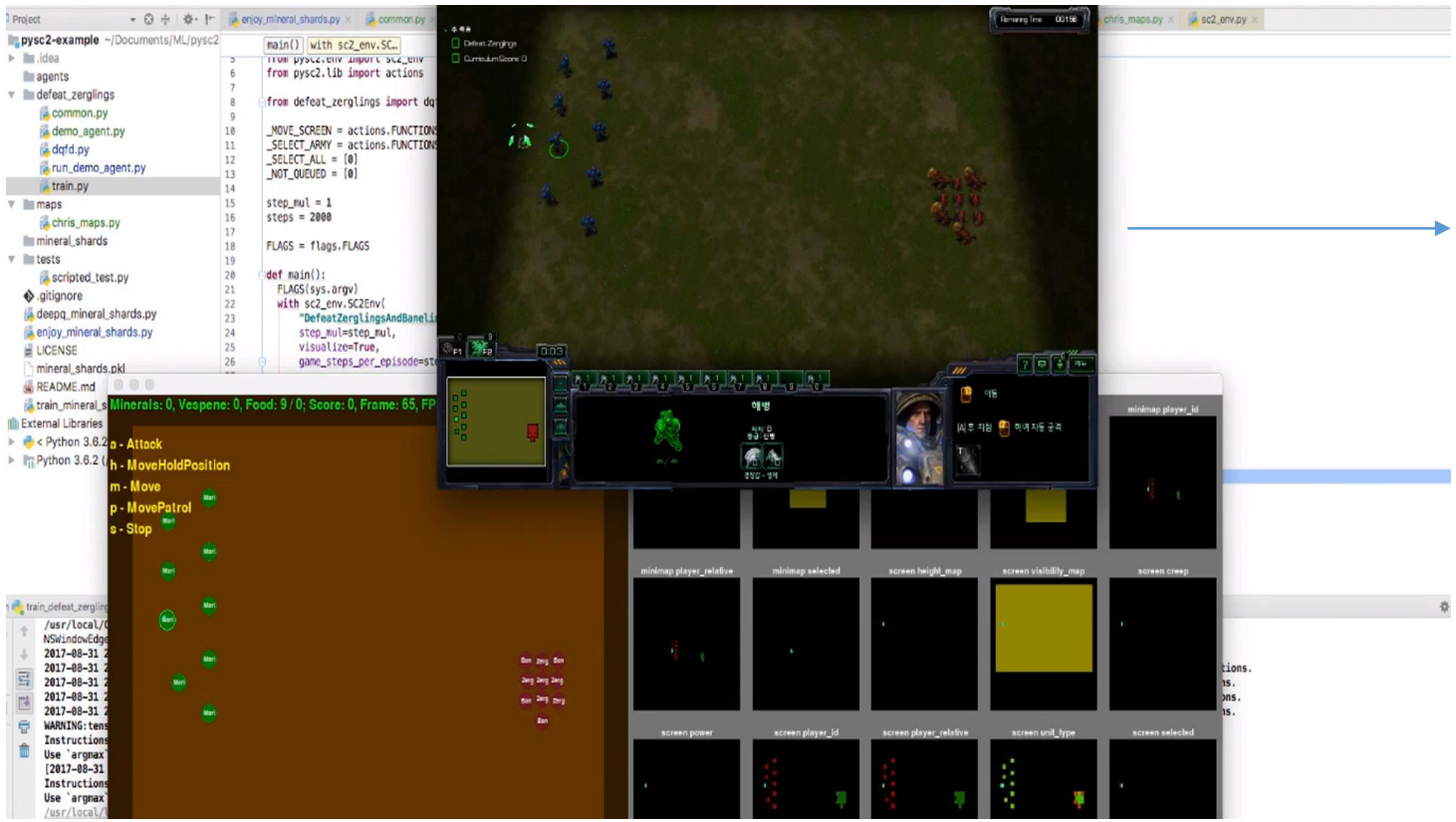
| 상태 | 행동 |
|---|---|
| 보상 | 정책 |

#상태
- 현재 나의 HP 혹은 캐릭터의 특징
- 내가 움직이는 속도 등

## 상태는 '정의'가 중요!

에이전트의 상태를 통해 상황을 판단.

다음 행동을 결정하기에 충분한 정보를 제공해야 할 수 있어야 함.




## 스타2의 세계

- 현재 움직이는 저글링들의 특성정보만 알고,
  업그레이드 현황, 체력 등을 모른다면
  에이전트는 사실상 이 게임을 지속할 수 없다.

- 이러한 에이전트가 교전 승리를 학습하려면은
  저글링의 위치, 속도, 공격력 등 이러한 정보가 필요.

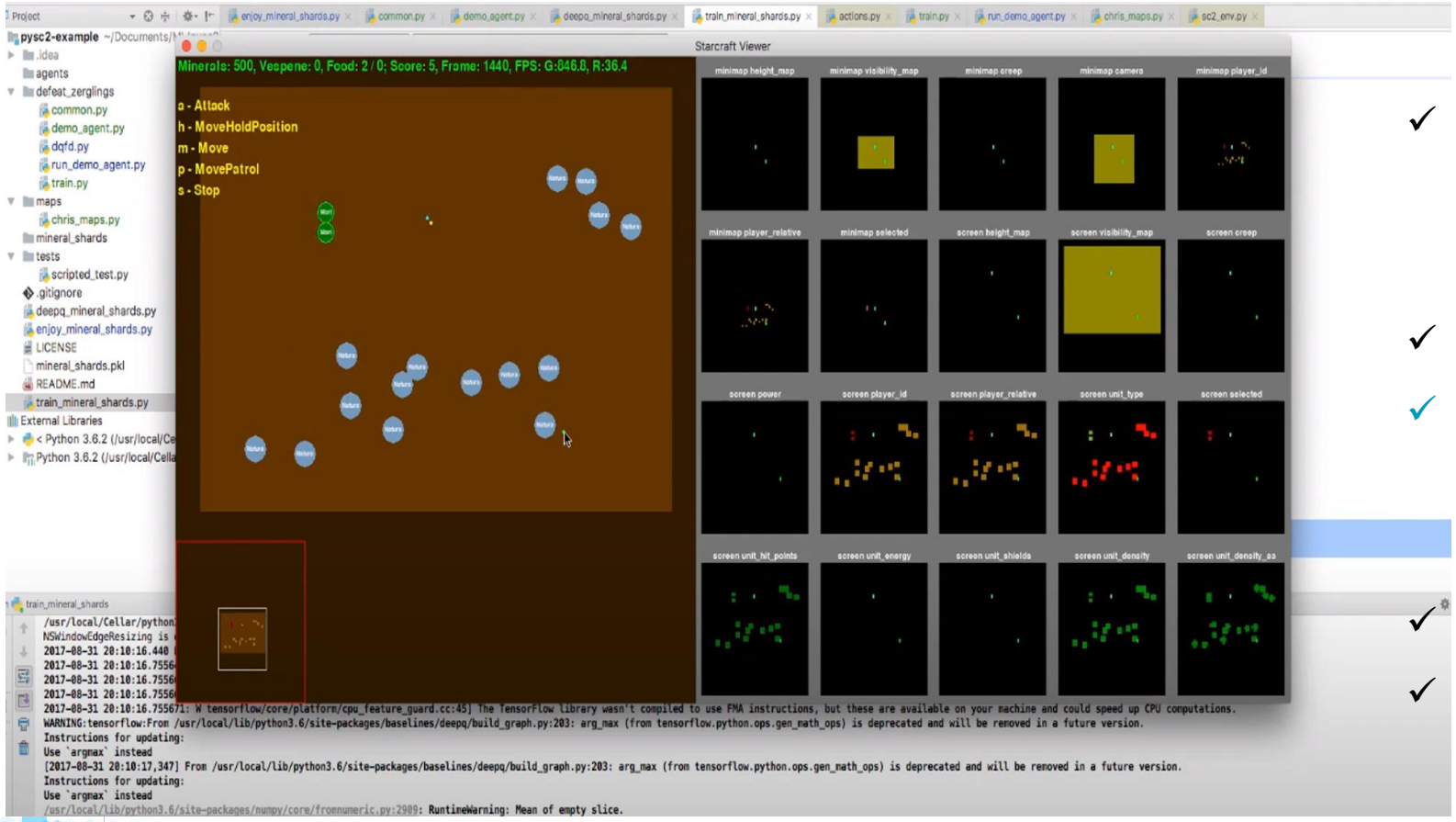
# Part 6. 강화

## #행동

✓ 에이전트가 어떠한 상태에서 취할 수 있는 행동.

✓ “상.하.좌.우”, “대쉬”, “점프”, “돌격”, “프로펠러 동작” 같은 것을 의미.

✓ 스타크래프트, 대전 격투, 카트라이더 이와 같은 게임에서의 행동은 마우스나 키보드를 통해 줄 수 있는 입력.




✓ 학습이 아직 수행되지 않은 에이전트는
어떠한 움직임이 좋은 행동인지 알 수 없음.
(무작위 움직임)

✓ 무작위 행동을 통해 어쩌다 좋은 보상을 획득.

✓ 에이전트는 좋은 보상에 대한 학습을 수행하면서,
좋은 보상을 획득할 행동들에 대해 확률을 높이게 됨.

✓ 그 행동을 취하면 환경은 에이전트에게 보상.

✓ 에이전트의 다음 상태 즉, 행동을 수행한 다음의
에이전트에 대한 상태를 다시 관찰해서 알려줍니다

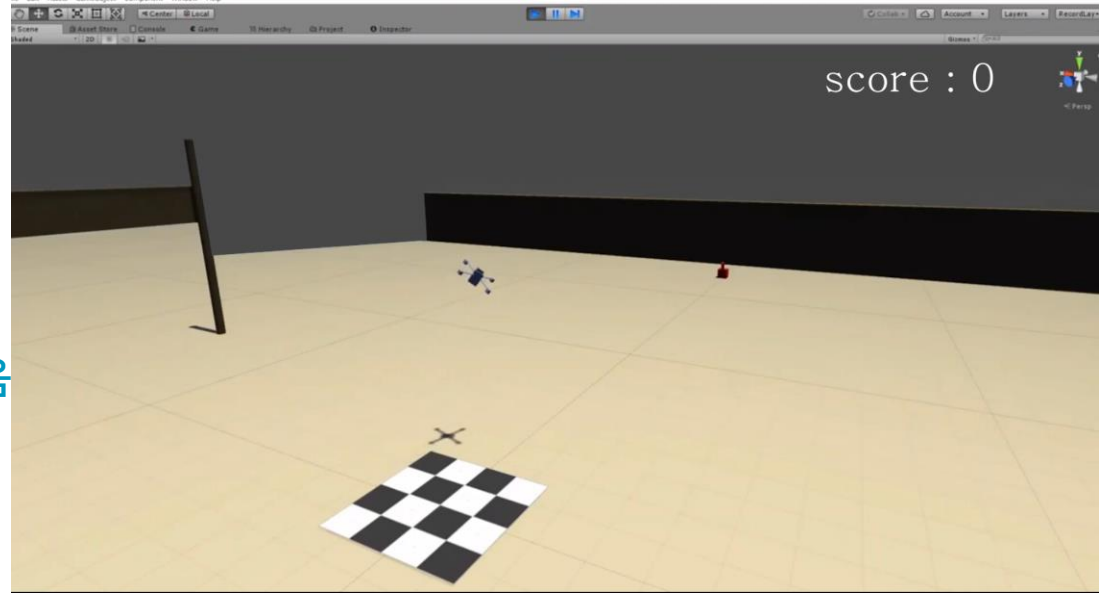
Part 6. 강화

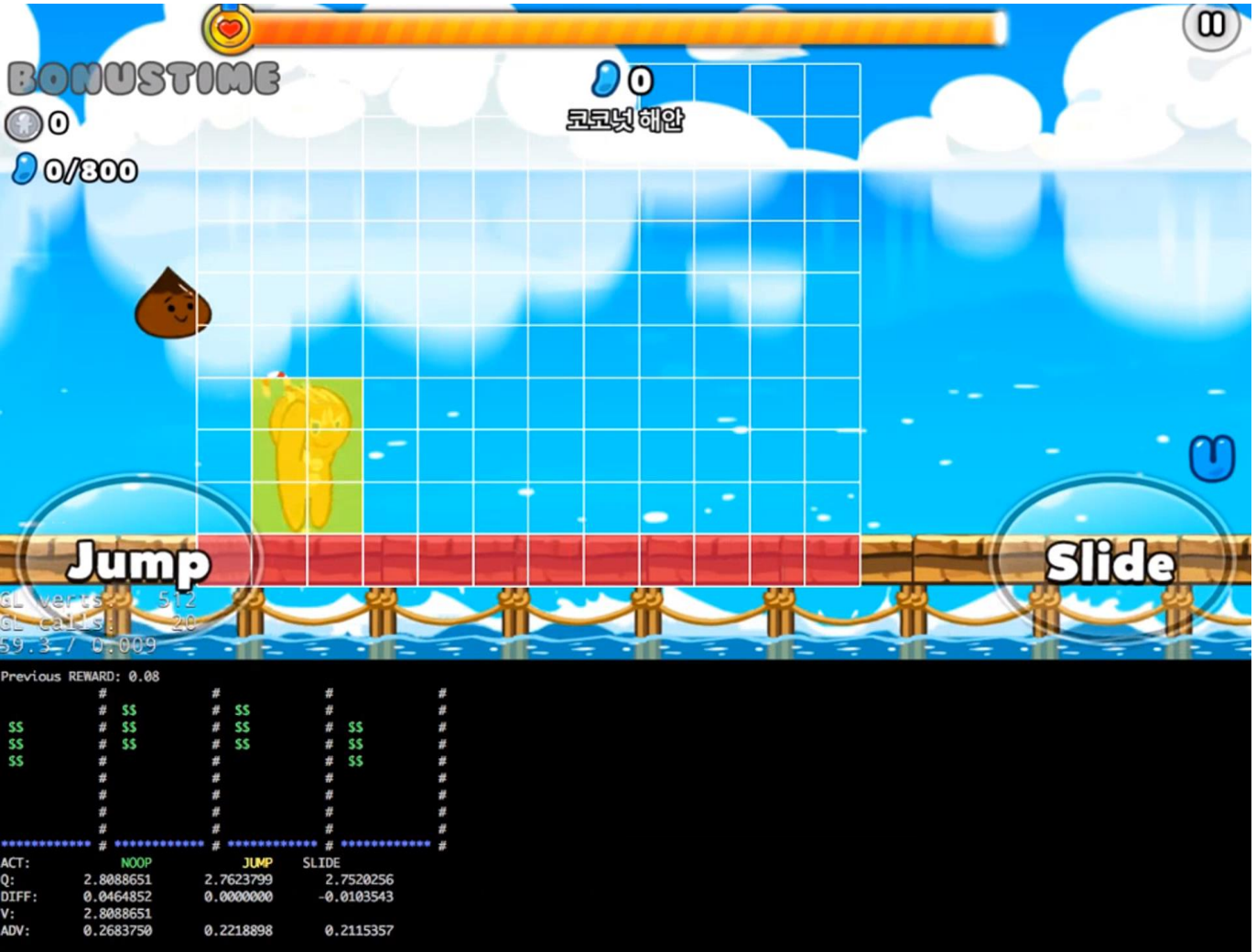
#보상

✓ 다른 머신러닝 기법과 다르게 만들어주는 가장 핵심적인 요소

✓ 보상은 에이전트가 학습할 수 있는 유일한 정보

✓ 보상이라는 정보를 통해 에이전트는 자신이 했던 행동이 좋은 행동인지 알 수 있음




## 강화학습의 목표

시간에 따라 얻는 보상들의 합을 최대로 하는 정책을 찾는 것

이 보상은 에이전트에 속하지 않는 환경의 일부

어떠한 상황에서 얼마의 보상이 나오는지 미리 알수 없음

# Part 6. 강화

## #정책

✓ 순차적 행동 결정문제에서 구해야 하는 답은 바로 정책

✓ 특정 상태가 아닌 모든 상태에 대해 어떠한 행동을 해야할지 에이전트는 알아야 함

✓ 모든 상태에 대해 에이전트가 어떤 행동을 해야하는지 정해놓은 것


✓ 순차적 행동 결정 문제를 풀었다고 한다면 에이전트는 제일 좋은 정책을 얻게 됨.

✓ 제일 좋은 정책은 optimal policy

✓ optimal policy에 따라 행동했을 때, 보상의 합을 최대로 받을 수 있음.

# Part 6. 강화

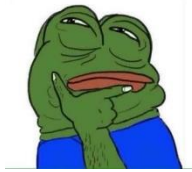
*value function*

: 앞으로 받을 것이라 예상하는 보상


에이전트는 실제로 받은 보상을 토대로 가치함수와 정책을 수정

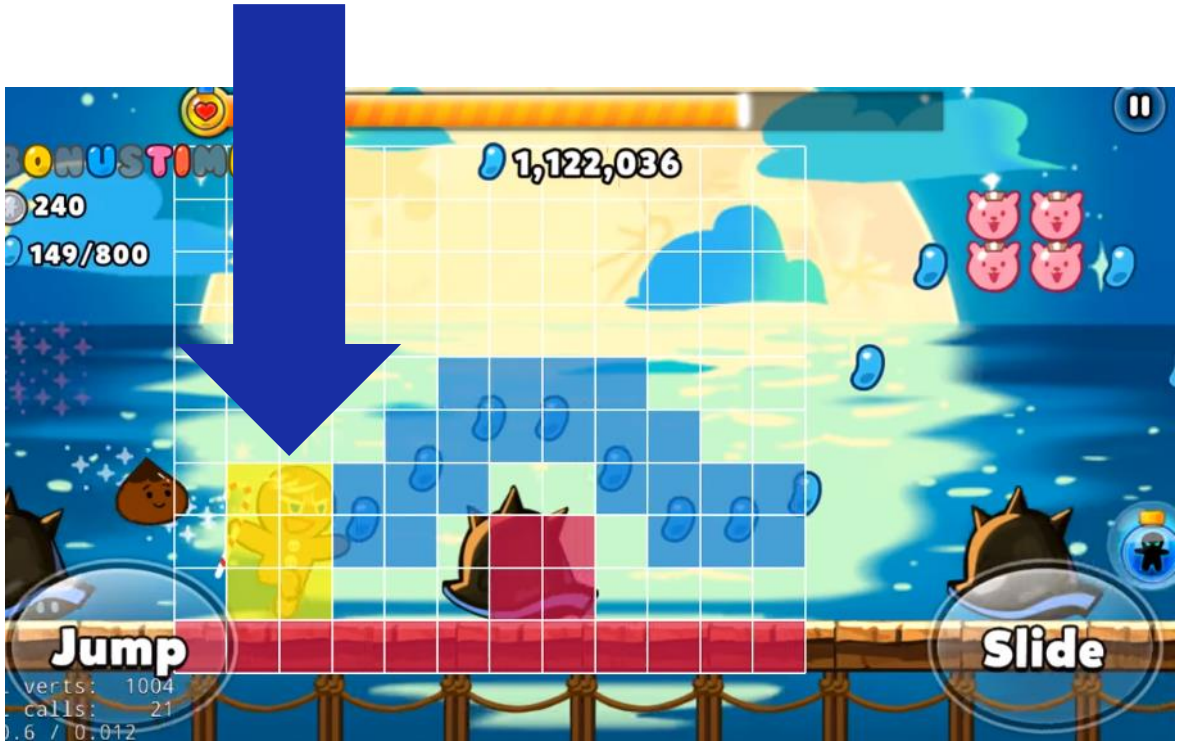
학습과정을 충분히 반복한다면 가장 많은 보상을 받게하는 정책을 학습

그렇다면 어떠한 방식을 통해 최적의 정책을 찾게 되는걸까요?



Part 6. 강화

# 어떠한 특정 상태의 에이전트




Q. 다음에는 어떤 행동을 하는것이 좋지?

A_1 슬라이딩 -> 점수 획득 및 게임 종료

A_2 점프 -> 점수 획득

A_3 달리기 -> 점수 획득 및 게임 종료

Q. 이걸 어떻게 알지?

A. 앞으로 받을 보상에 대한 개념이 바로 가치함수

#반환값의 등장

# # 에이전트의 탐험


가치함수의 개념

: 반환값에 대한 기댓값으로 특정 상태의 가치를 판단할 수 있게된다.

상태의 가치를 고려하는 이유는

현재 에이전트가 갈 수 있는 상태들의 가치를 안다면

그 중에서 가장 가치가 제일 높은 상태를 선택할 수 있기 때문.

- ✓ 반환값은 에이전트가 실제로 환경을 탐험하면서 받은 보상의 합.
- ✓ 에이전트는 환경과 정해진 시간 동안 상호작용, 마지막 상태가 되면 그때 반환값을 계산할 수 있음.
- ✓ 즉, 에이전트가 에피소드가 끝난 후에 보상을 정산하는것이 반환값.

상태$_{t-1}$

상태$_t$

- ✓ MDP로 정의되는 문제에서 가치함수는 항상 정책에 의존.
- ✓ 정책을 고려한 가치함수를 벨만 기대 방정식

$$V_\pi(s)$$
$$(Bellman\ Expectation\ Equation)$$
$$= E_\pi\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s\right]$$

상태$_{t-1}$에서 상태$_t$로 넘어갔으니 상태$_t$에 맞게 행동하자!

**#에이전트의 정책**

- ✓ 벨만 기대 방정식은 현재 상태의 가치함수와 다음 상태의 가치함수 사이의 관계를 말해주는 방정식.
- ✓ 강화학습은 벨만 방정식을 어떻게 풀어가느냐가 관건.