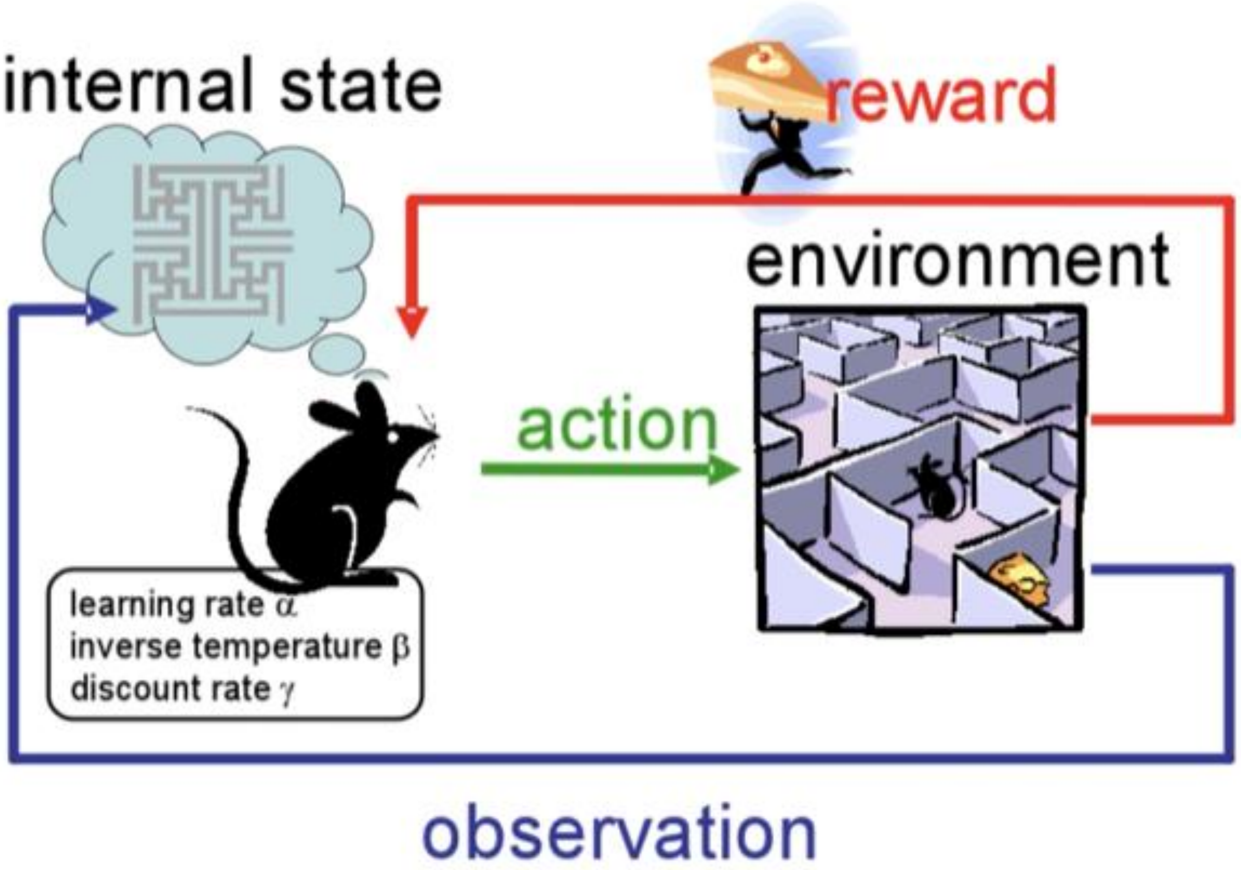
모두를 위한 RL 강좌 : Lecture 1~3

임도연

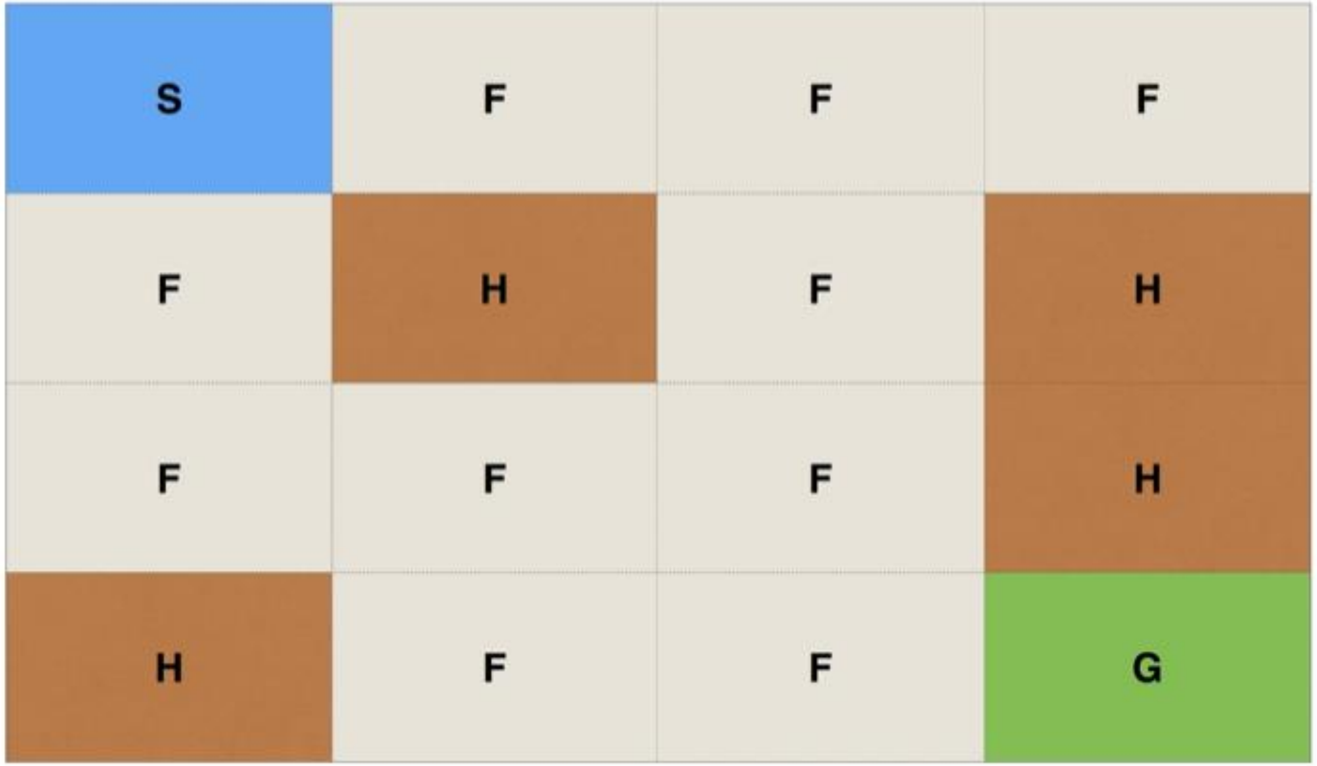
Lecture 1 : RL 수업 소개

1. 강화학습이란?


internal state
reward
environment
action
learning rate α
inverse temperature β
discount rate γ
observation

Lecture2 : Playing OpenAI GYM Games

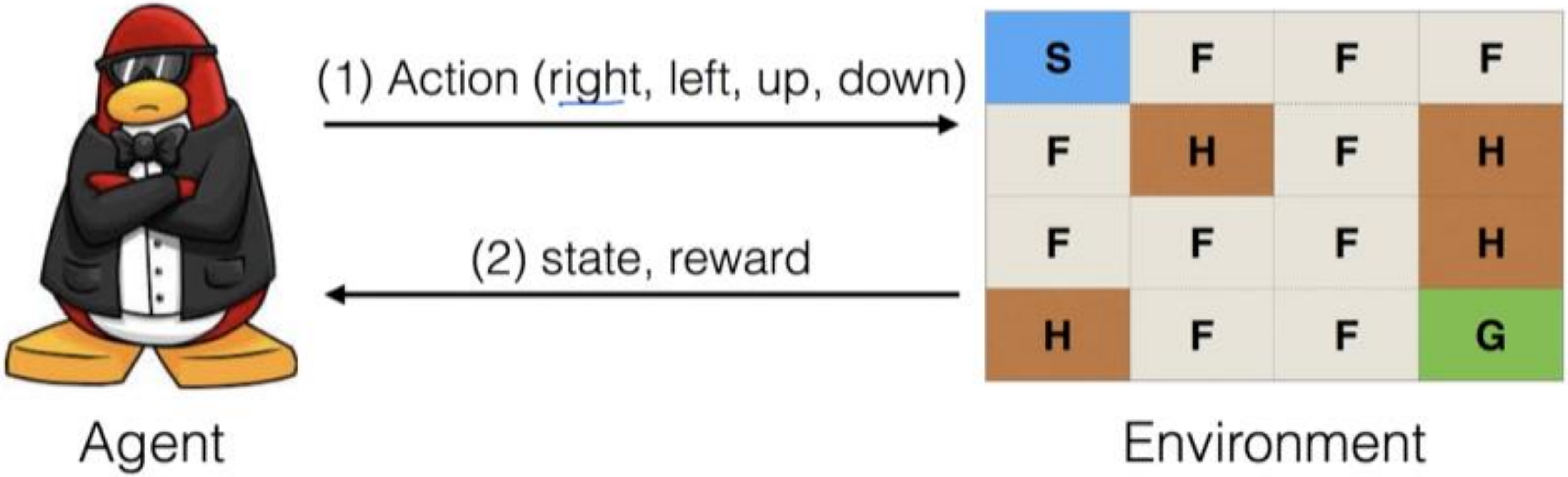
Frozen Lake 게임을 통한 실습



S	F	F	F
F	H	F	H
F	F	F	H
H	F	F	G

S : 시작점 F : 얼어있는 땅 H : 구멍 G : 목표 지점(도착점)

Frozen Lake 게임을 통한 실습



(1) Action (right, left, up, down)

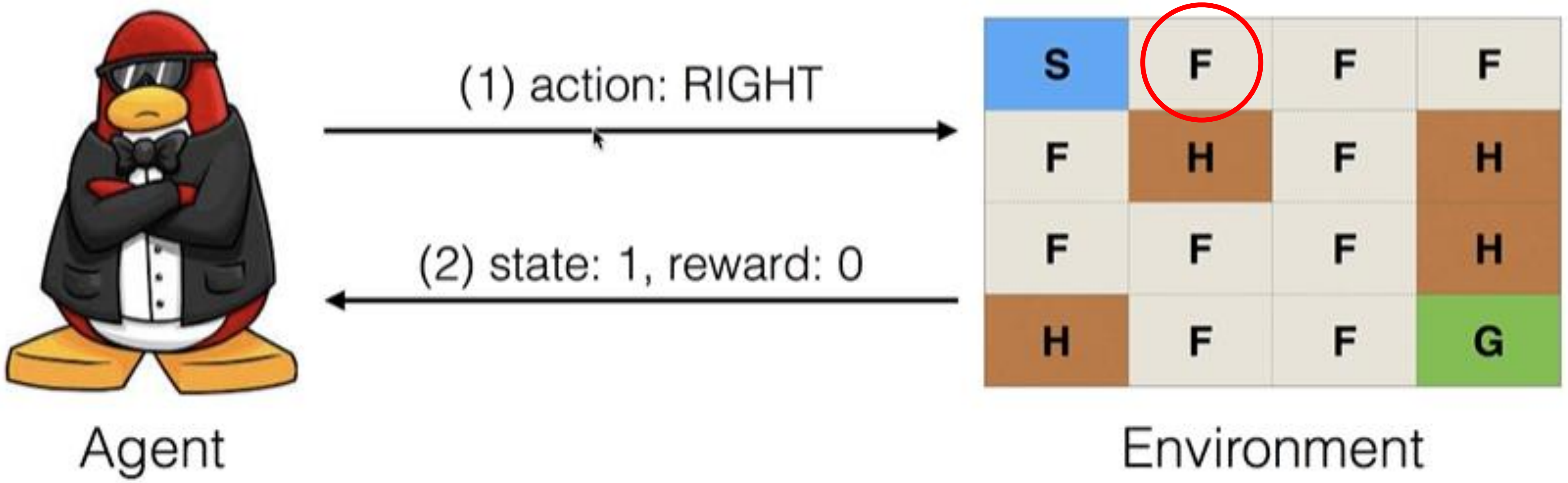
(2) state, reward

Agent

Environment

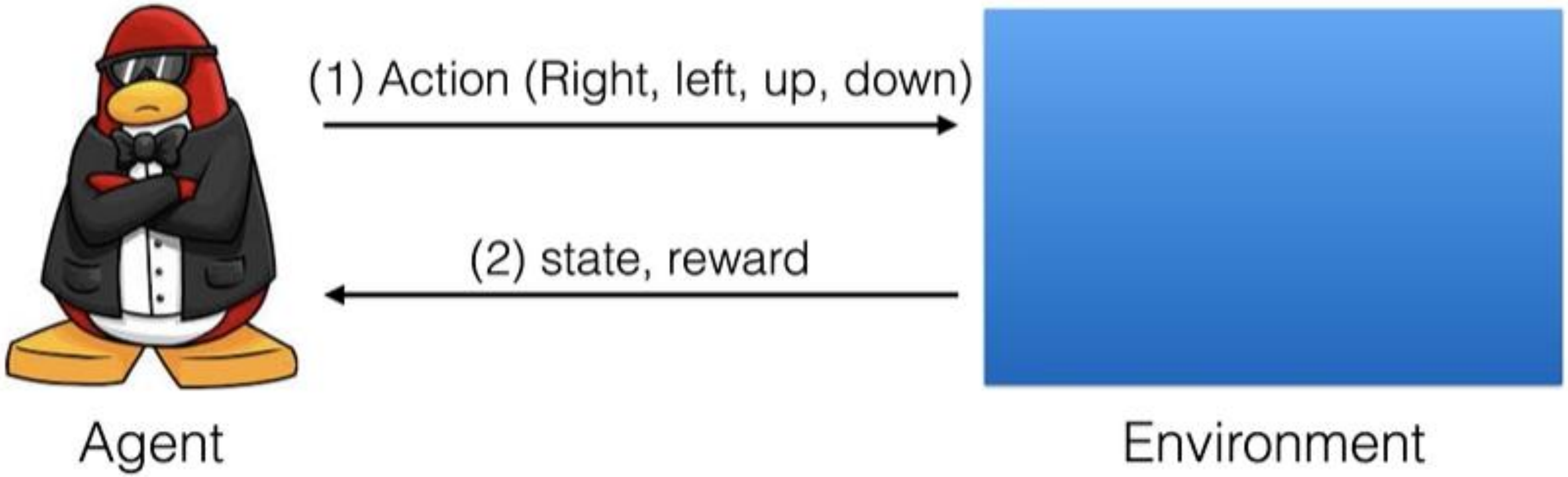
S	F	F	F
F	H	F	H
F	F	F	H
H	F	F	G

Frozen Lake 게임을 통한 실습



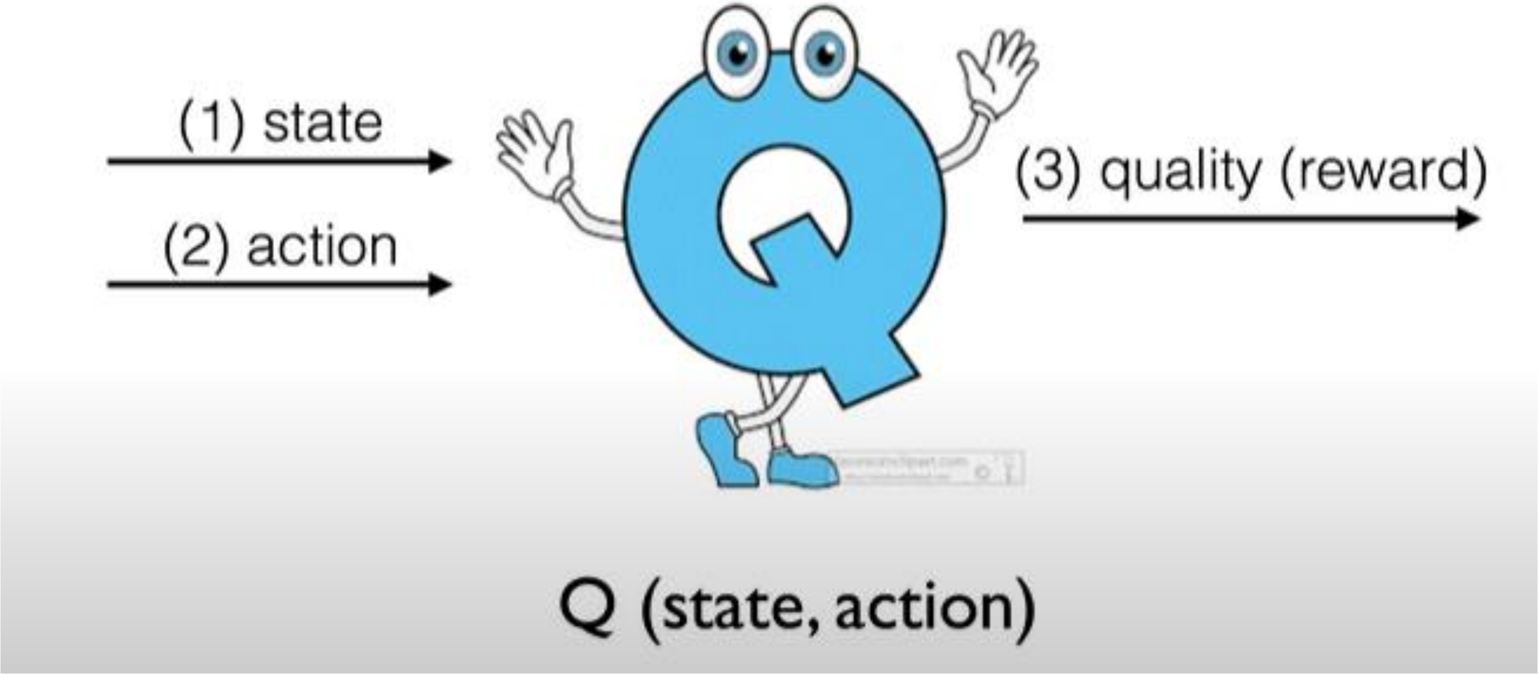
S	F	F	F
F	H	F	H
F	F	F	H
H	F	F	G

Frozen Lake 게임을 통한 실습



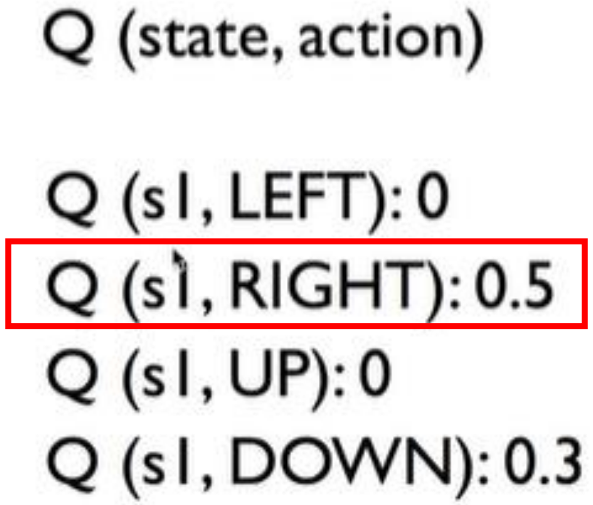
Lecture3 : Dummy Q-learning

Q-learning



Q : 현재 상태에서 취한 행동의 보상에 대한 quality

Q-learning



Q (state, action)

Q (s1, LEFT): 0

Q (s1, RIGHT): 0.5

Q (s1, UP): 0

Q (s1, DOWN): 0.3

max값

Q가 가지는 최대값을 의미

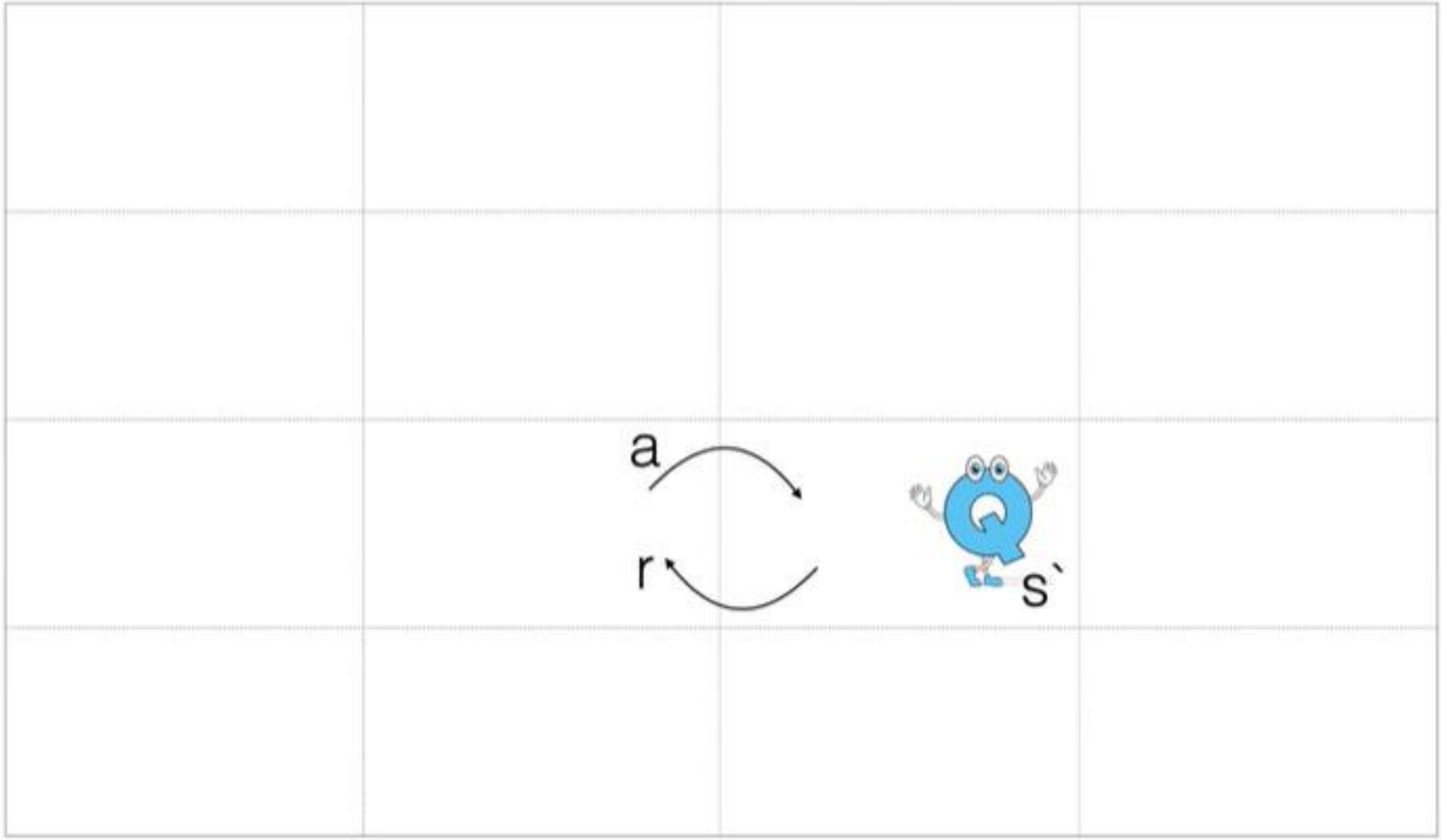
$$\text{Max Q} = \max_{a'} Q(s, a')$$

$$\pi^{\boxed{*}}(s) = \operatorname{argmax}_{a} Q(s, a)$$

최대값일 때 가지게 되는 변수 값

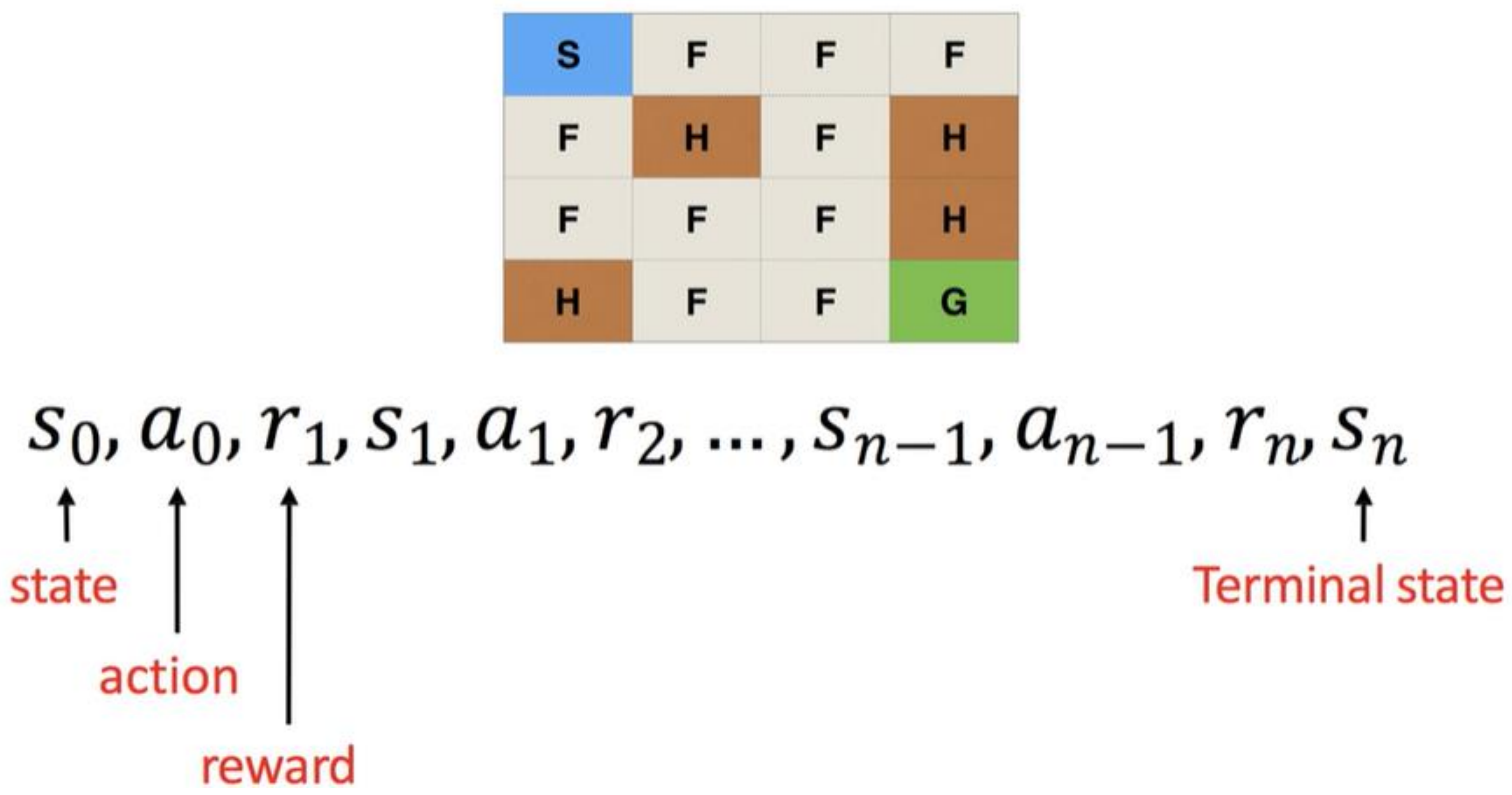
* : 최적의 값을 의미

Q-learning



ⓒSaebyeol Yu. Saebyeol's PowerPoint

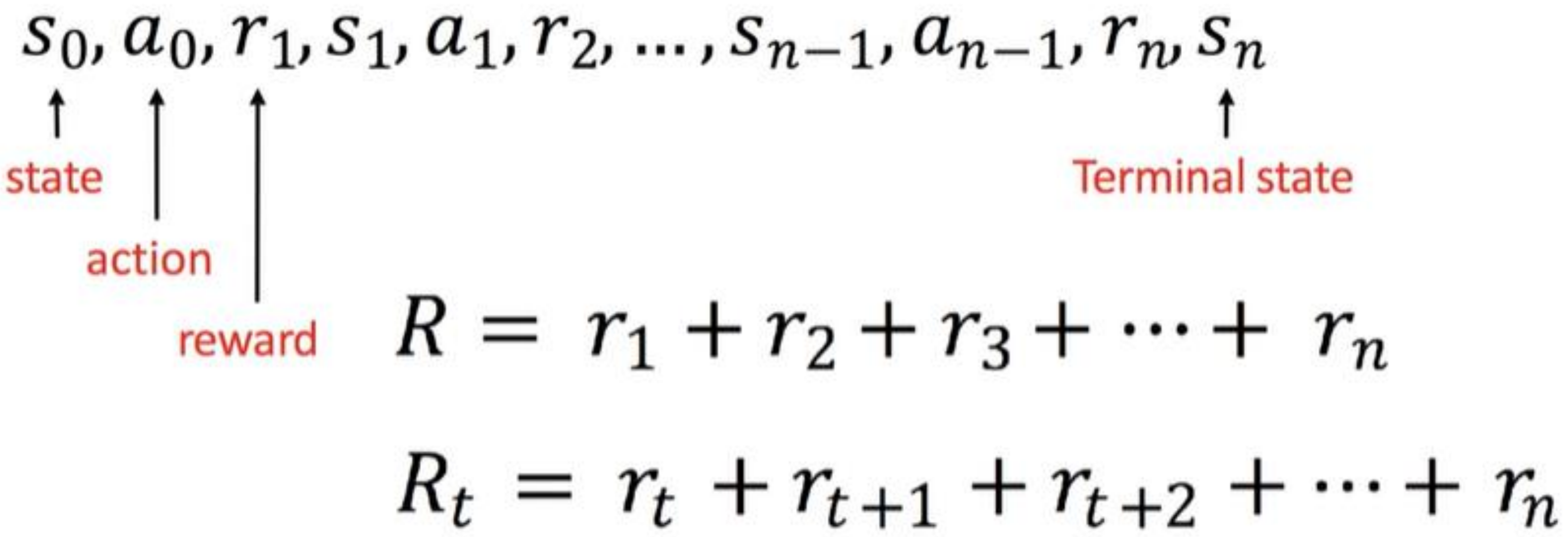
Q-learning



S	F	F	F
F	H	F	H
F	F	F	H
H	F	F	G

$$s_0, a_0, r_1, s_1, a_1, r_2, \ldots, s_{n-1}, a_{n-1}, r_n, s_n$$

Q-learning



$$s_0, a_0, r_1, s_1, a_1, r_2, \dots, s_{n-1}, a_{n-1}, r_n, s_n$$

$$R = r_1 + r_2 + r_3 + \cdots + r_n$$

$$R_t = r_t + r_{t+1} + r_{t+2} + \cdots + r_n$$

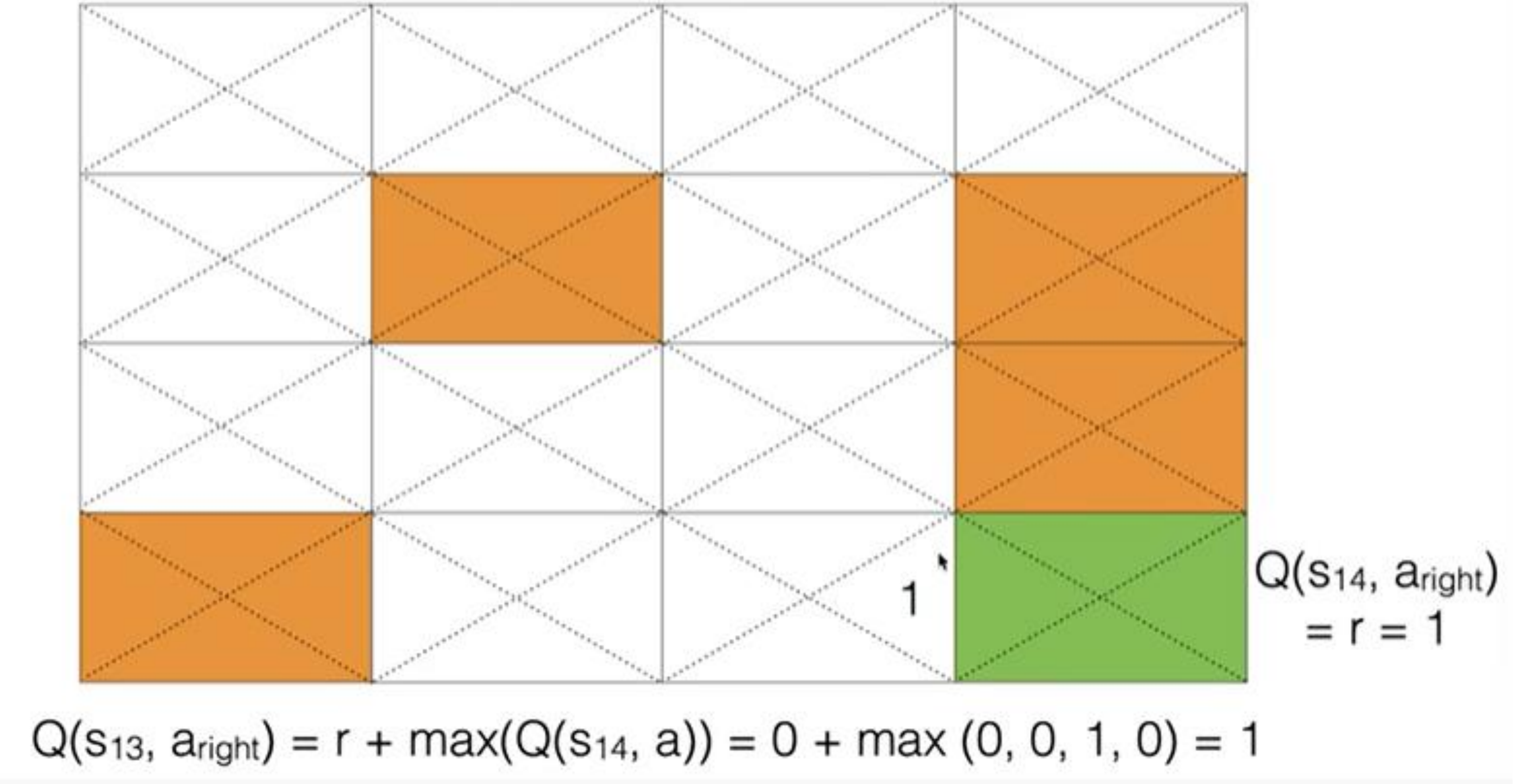
R(t+1) = $r_{t+1} + r_{t+2} + \cdots + r_n$

R(t) = r_t + R(t+1)

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

$R(t)^* = r_t$ +max R(t+1)

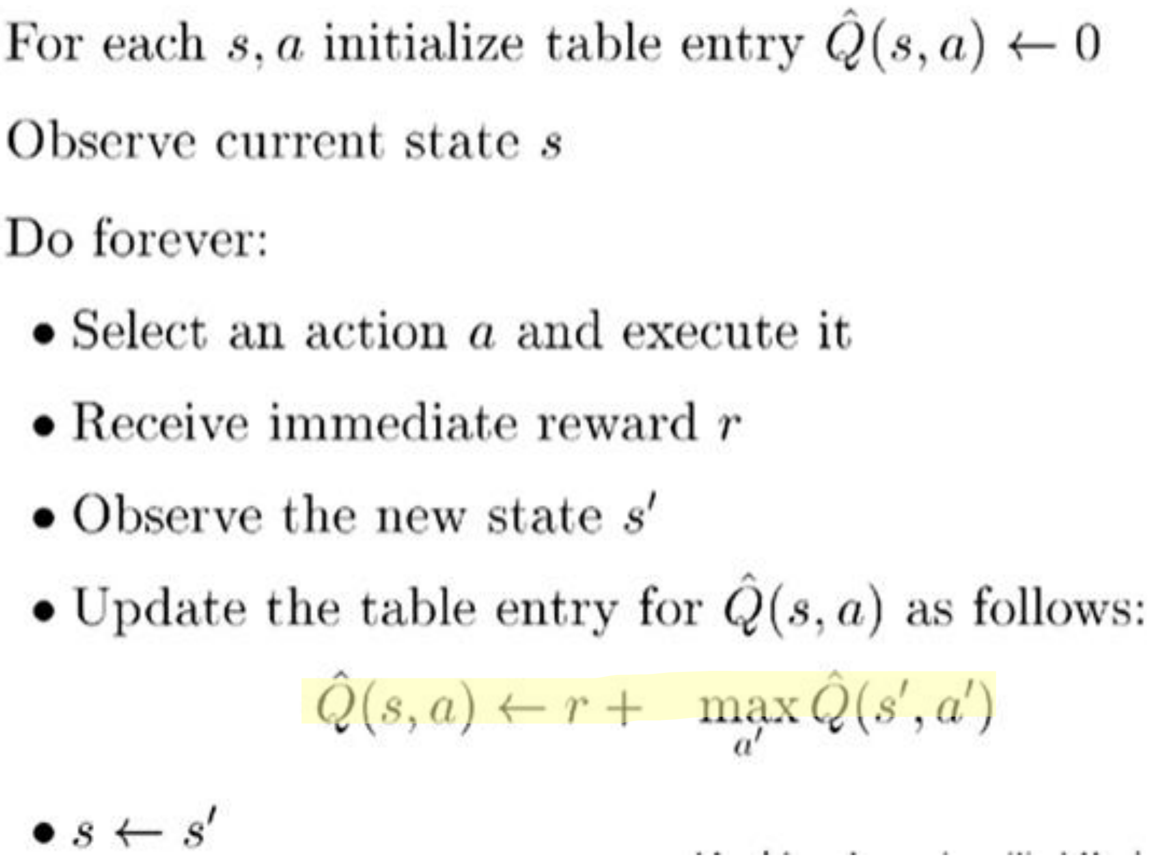
Q-learning



$Q(s_{14}, a_{right})$
$= r = 1$

$Q(s_{13}, a_{right}) = r + \max(Q(s_{14}, a)) = 0 + \max(0, 0, 1, 0) = 1$

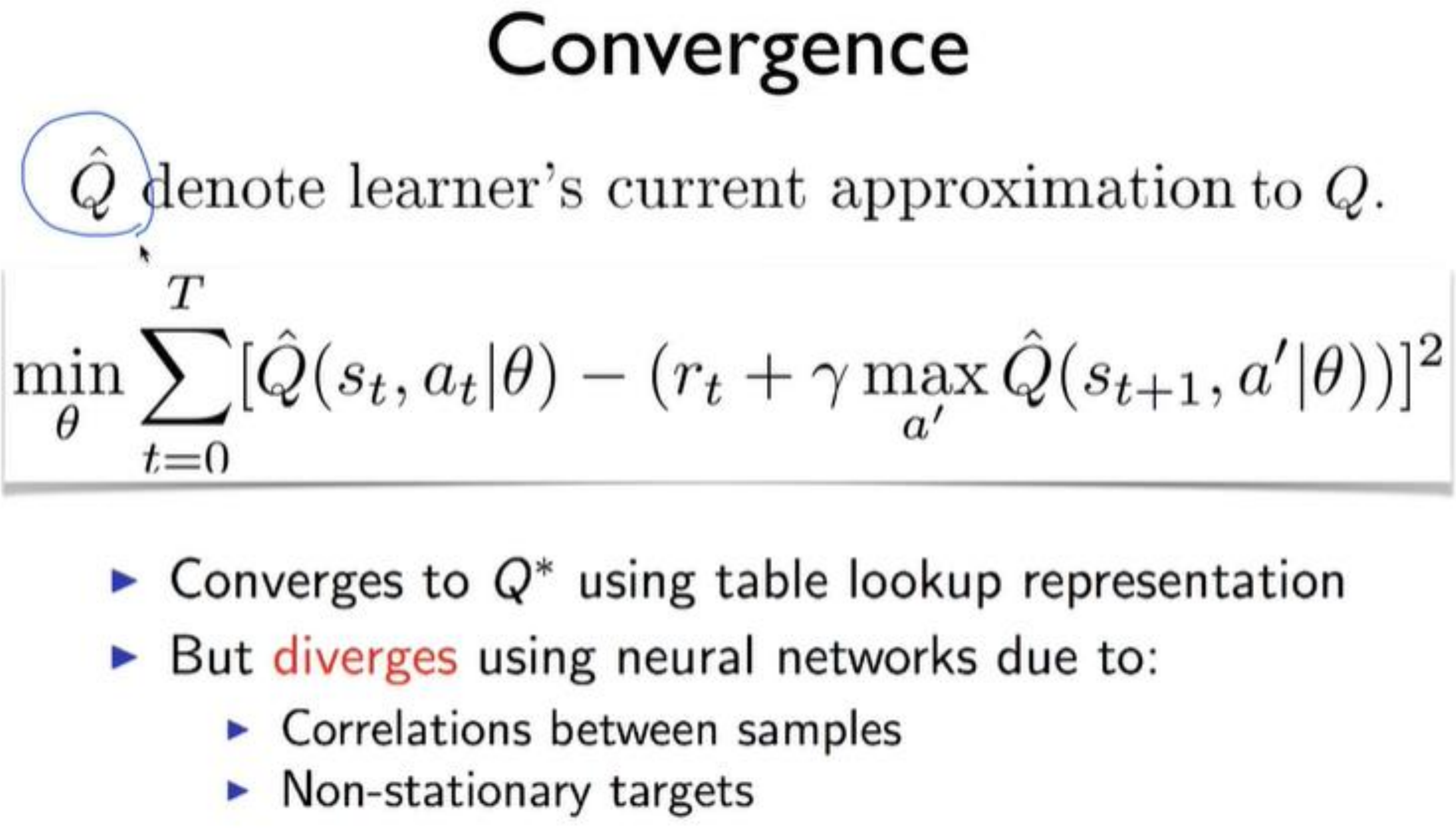
Q-learning



```
For each $s, a$ initialize table entry $\hat{Q}(s, a) \leftarrow 0$
Observe current state $s$
Do forever:
  • Select an action $a$ and execute it
  • Receive immediate reward $r$
  • Observe the new state $s'$
  • Update the table entry for $\hat{Q}(s, a)$ as follows:
        $\hat{Q}(s, a) \leftarrow r + \max_{a'} \hat{Q}(s', a')$
  • $s \leftarrow s'$
```

Lecture7 : DQN

DQN



Convergence

$\hat{Q}$ denote learner's current approximation to Q.

$$\min_{\theta} \sum_{t=0}^{T} [\hat{Q}(s_t, a_t|\theta) - (r_t + \gamma \max_{a'} \hat{Q}(s_{t+1}, a'|\theta))]^2$$

- Converges to Q^* using table lookup representation
- But diverges using neural networks due to:
  - Correlations between samples
  - Non-stationary targets

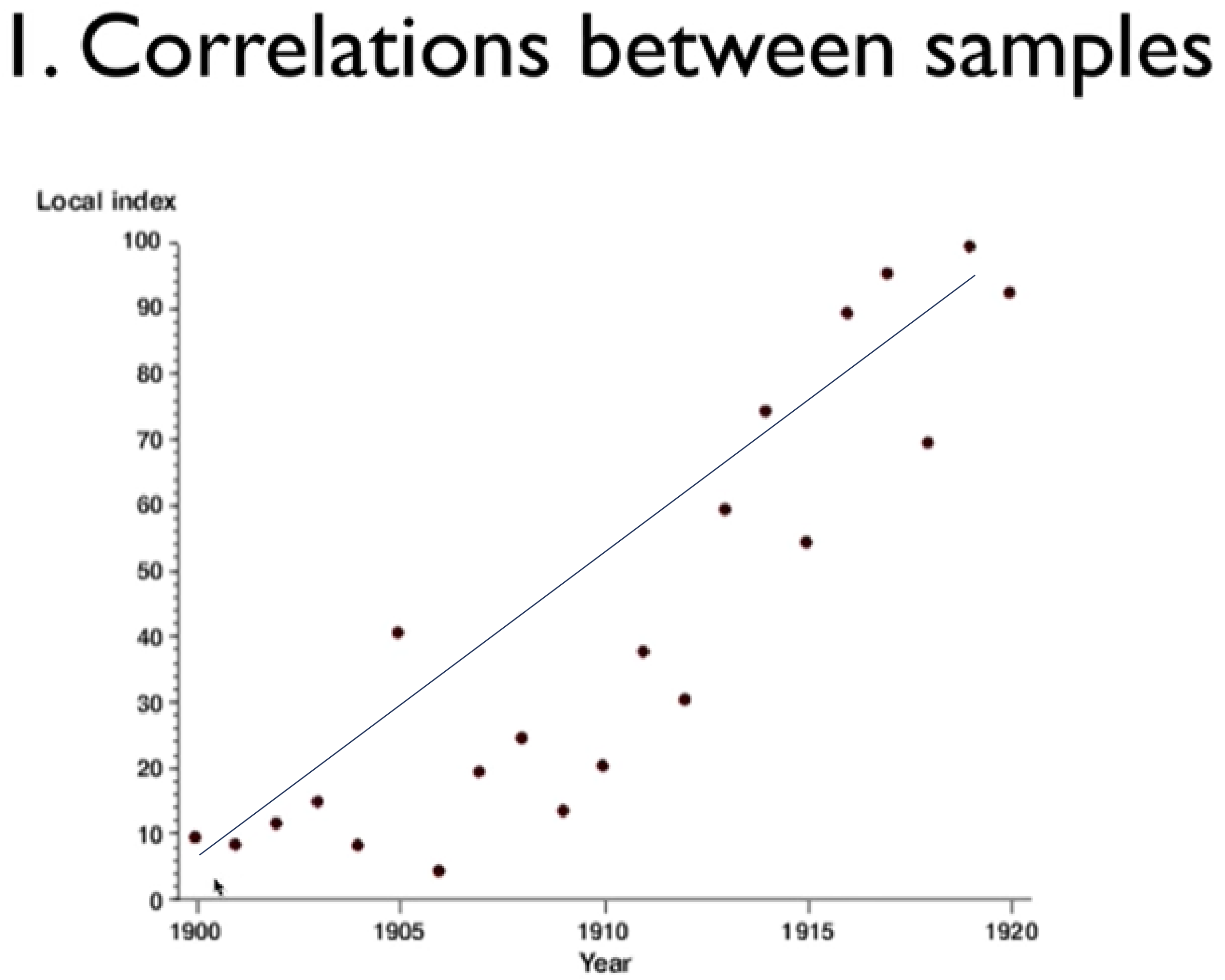
DQN

DQN paper
www.nature.com/articles/nature14236

DQN source code:
sites.google.com/a/deepmind.com/dqn/

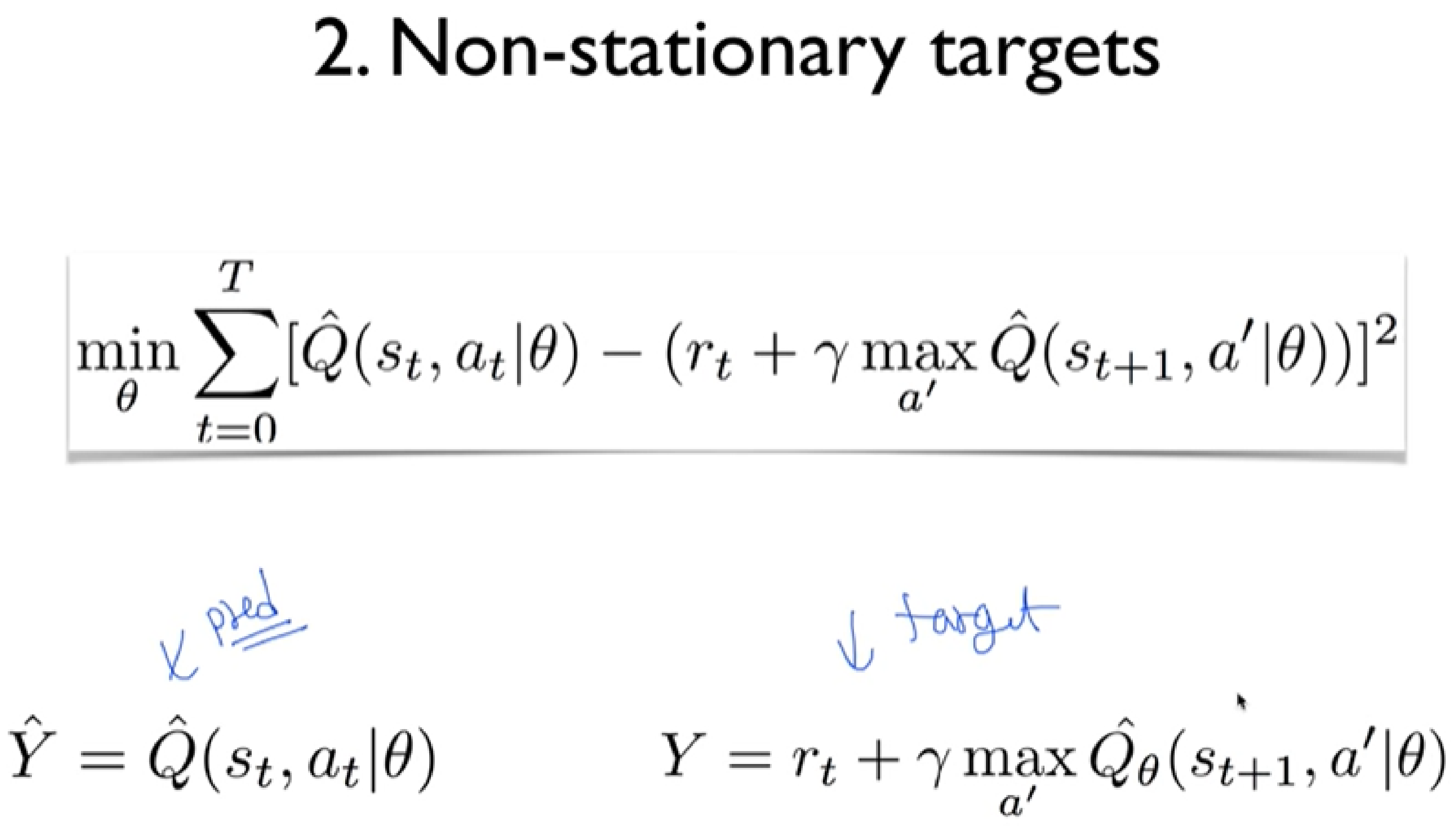
nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
LEARNING CURVE
Self-taught AI software attains human-level performance in video games
SHARE DATA IN OUTBREAKS
A GIANT IN THE EARLY UNIVERSE
TELEPORTATION FOR TWO

DQN



1. Correlations between samples

DQN



2. Non-stationary targets

$$\min_{\theta} \sum_{t=0}^{T} [\hat{Q}(s_t, a_t|\theta) - (r_t + \gamma \max_{a'} \hat{Q}(s_{t+1}, a'|\theta))]^2$$

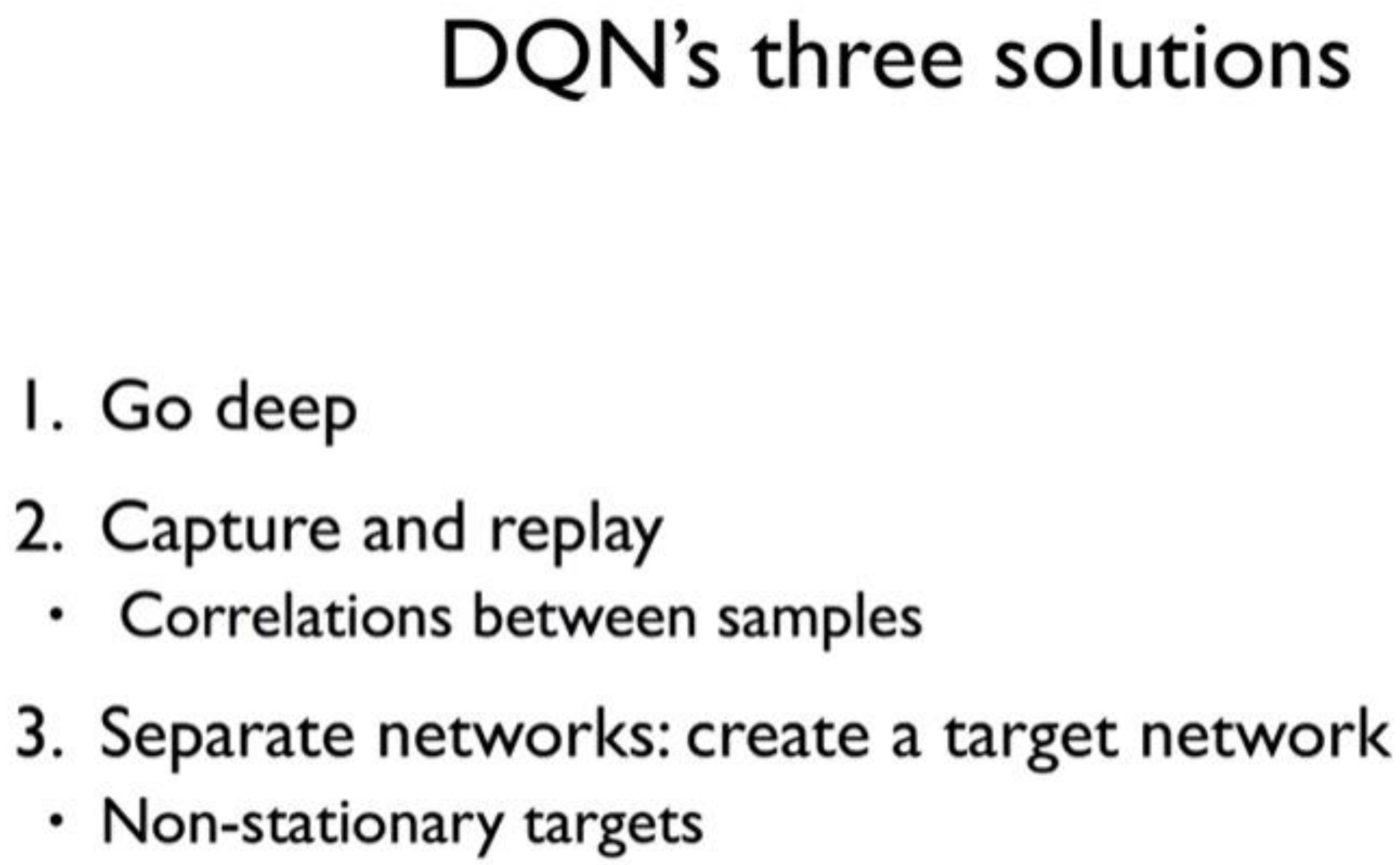
↙ pred

↓ target

$$\hat{Y} = \hat{Q}(s_t, a_t|\theta) \qquad Y = r_t + \gamma \max_{a'} \hat{Q}_\theta(s_{t+1}, a'|\theta)$$

업데이트 하는 과정에서 target이 움직임

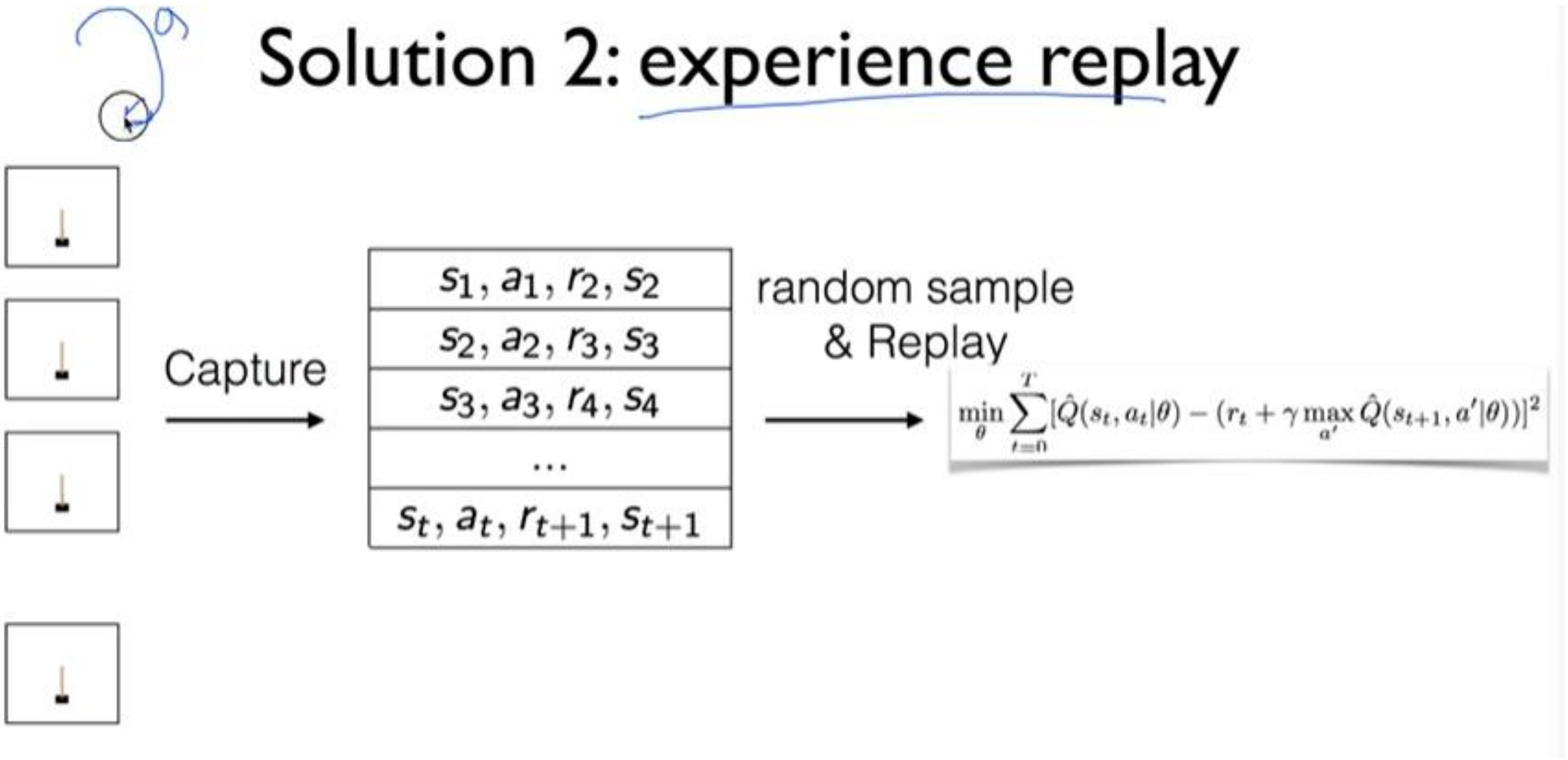
DQN



DQN’s three solutions

1. Go deep
2. Capture and replay
   - Correlations between samples
3. Separate networks: create a target network
   - Non-stationary targets

DQN



Solution 2: experience replay

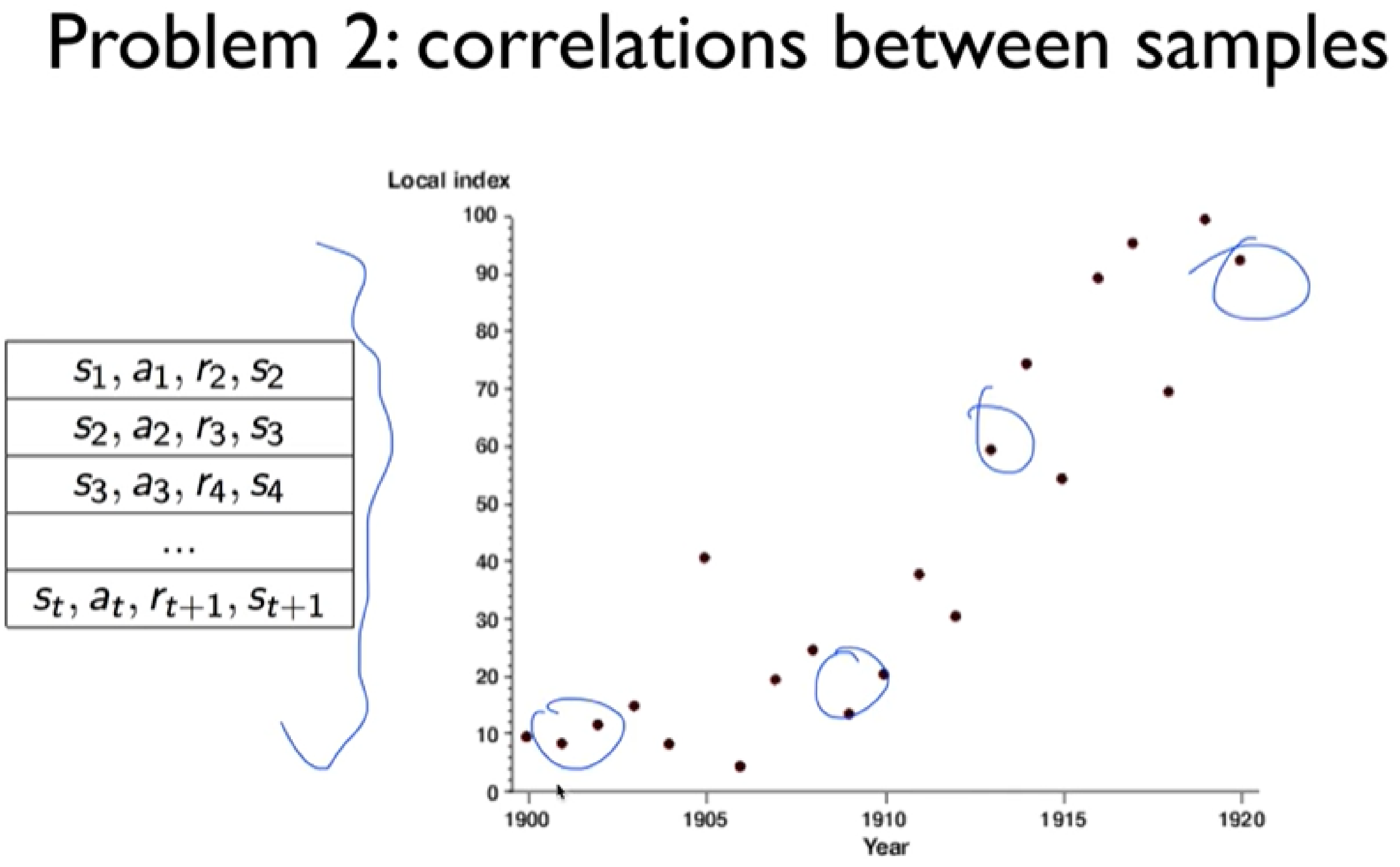
Capture

s_1, a_1, r_2, s_2
s_2, a_2, r_3, s_3
s_3, a_3, r_4, s_4
$\ldots$
$s_t, a_t, r_{t+1}, s_{t+1}$

random sample
& Replay

$$\min_{\theta} \sum_{t=0}^{T} [\hat{Q}(s_t, a_t|\theta) - (r_t + \gamma \max_{a'} \hat{Q}(s_{t+1}, a'|\theta))]^2$$

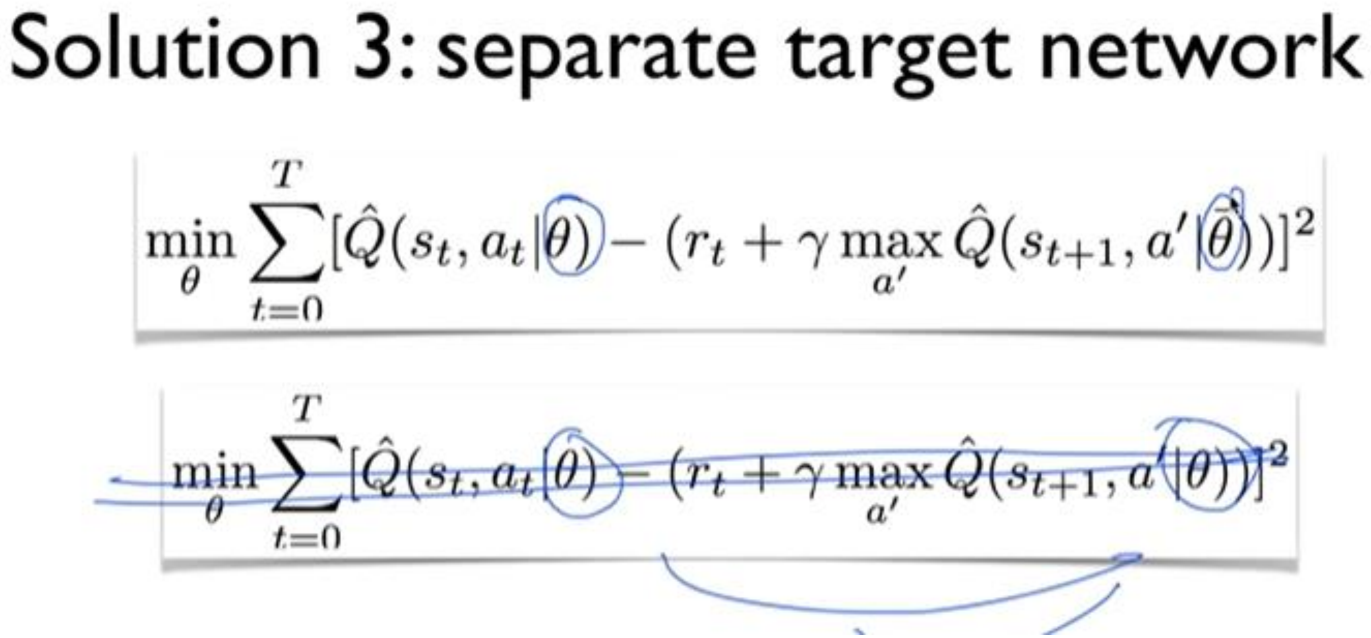
DQN



Problem 2: correlations between samples

s_1, a_1, r_2, s_2
s_2, a_2, r_3, s_3
s_3, a_3, r_4, s_4
$\ldots$
$s_t, a_t, r_{t+1}, s_{t+1}$

DQN

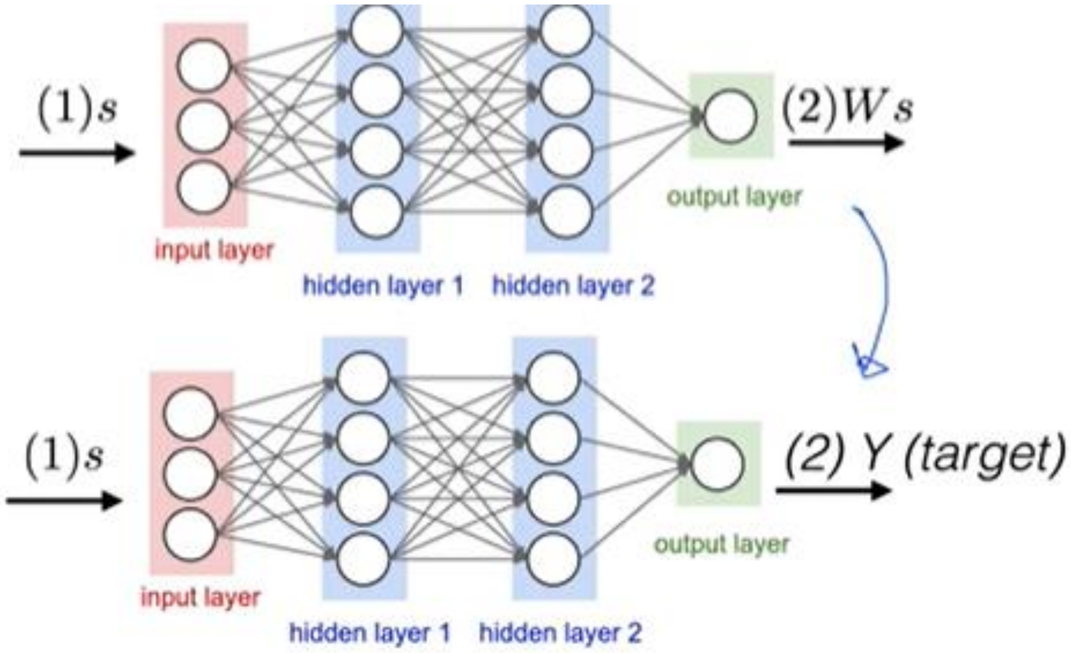


Solution 3: separate target network

$$\min_{\theta}\sum_{t=0}^{T}[\hat{Q}(s_t,a_t|\theta)-(r_t+\gamma\max_{a'}\hat{Q}(s_{t+1},a'|\bar{\theta}))]^2$$

$$\cancel{\min_{\theta}\sum_{t=0}^{T}[\hat{Q}(s_t,a_t|\theta)-(r_t+\gamma\max_{a'}\hat{Q}(s_{t+1},a'|\theta))]^2}$$

DQN


(1)s
input layer
hidden layer 1
hidden layer 2
output layer
(2)W s
(1)s
input layer
hidden layer 1
hidden layer 2
output layer
(2) Y (target)

Q&A