CSCE 636 Deep Learning

Lecture 18: Deep Reinforcement Learning

Anxiao (Andrew) Jiang

Based on the interesting lecture of Prof. Hung-yi Lee "Deep Reinforcement Learning"
https://www.youtube.com/watch?v=W8XF3ME8G2I&list=PLJV_el3uVTsPy9oCRY30oBPNLCo89yu49&index=33

Deep Reinforcement Learning

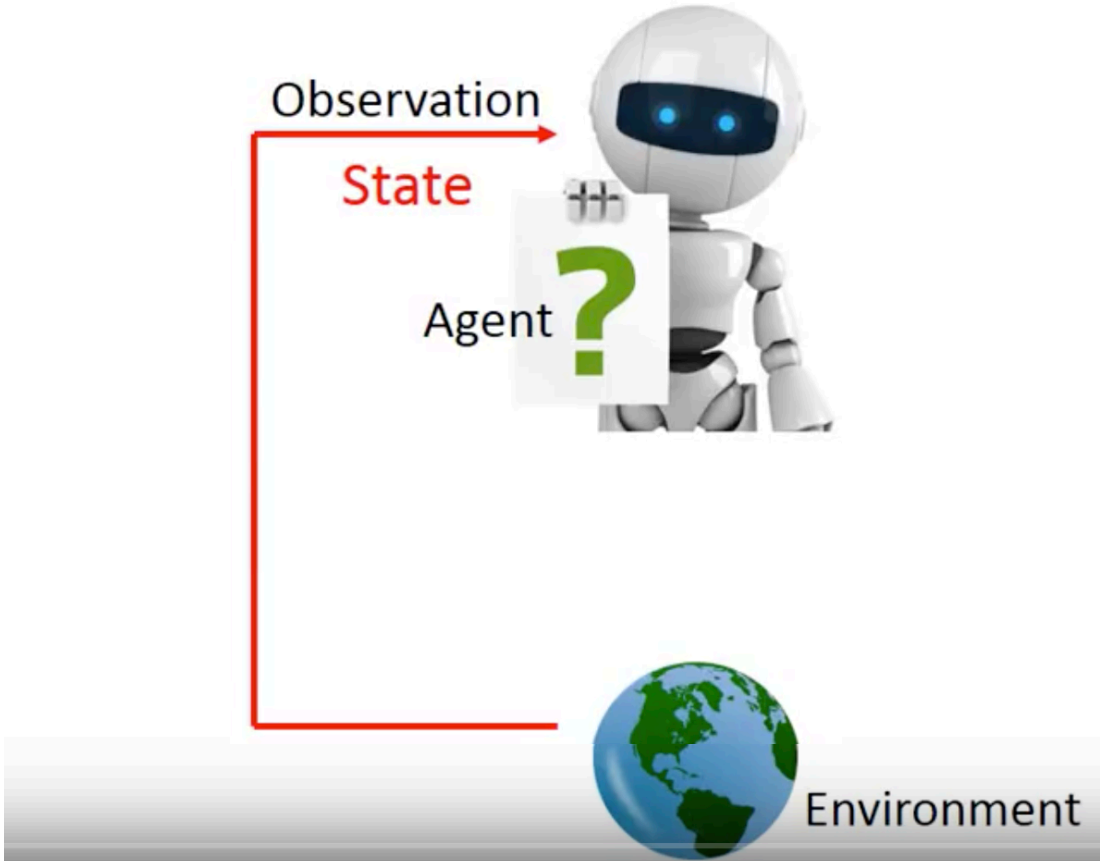
Scratching the surface

Deep Reinforcement Learning

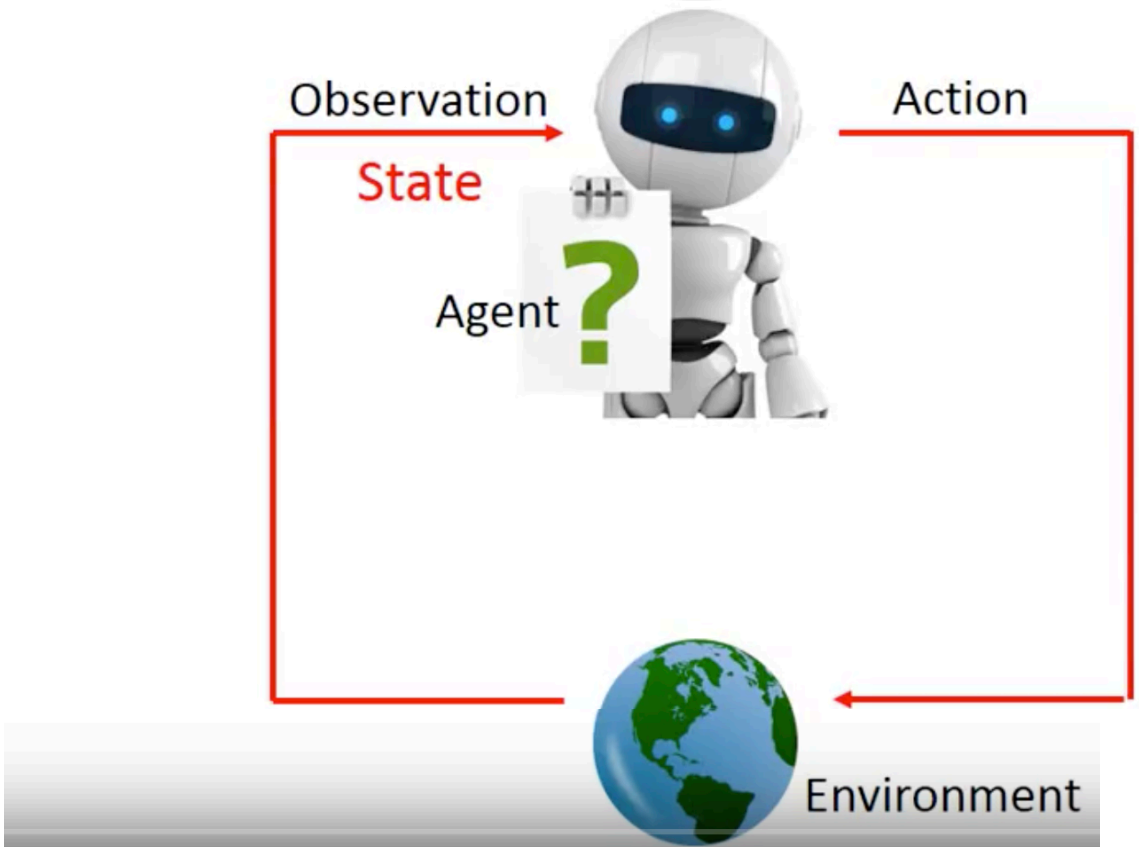
Scenario of Reinforcement Learning

Environment

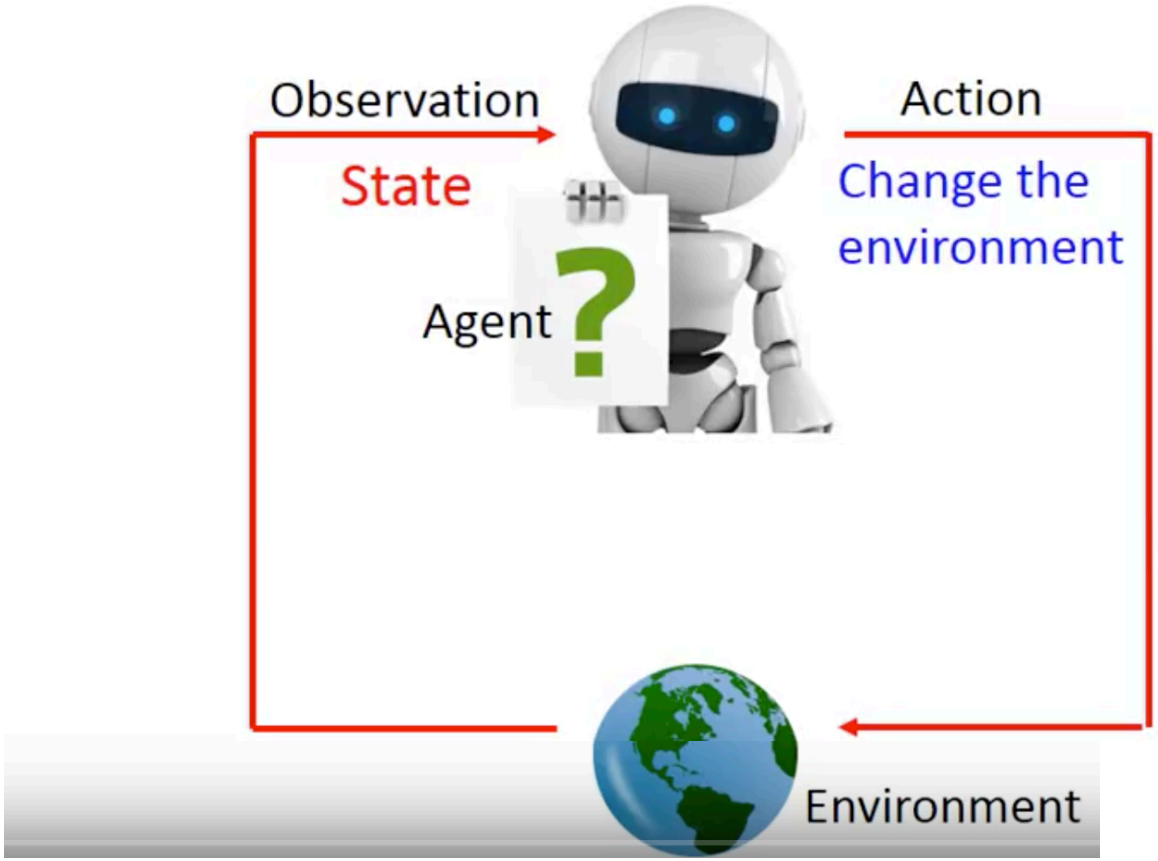
Scenario of Reinforcement Learning



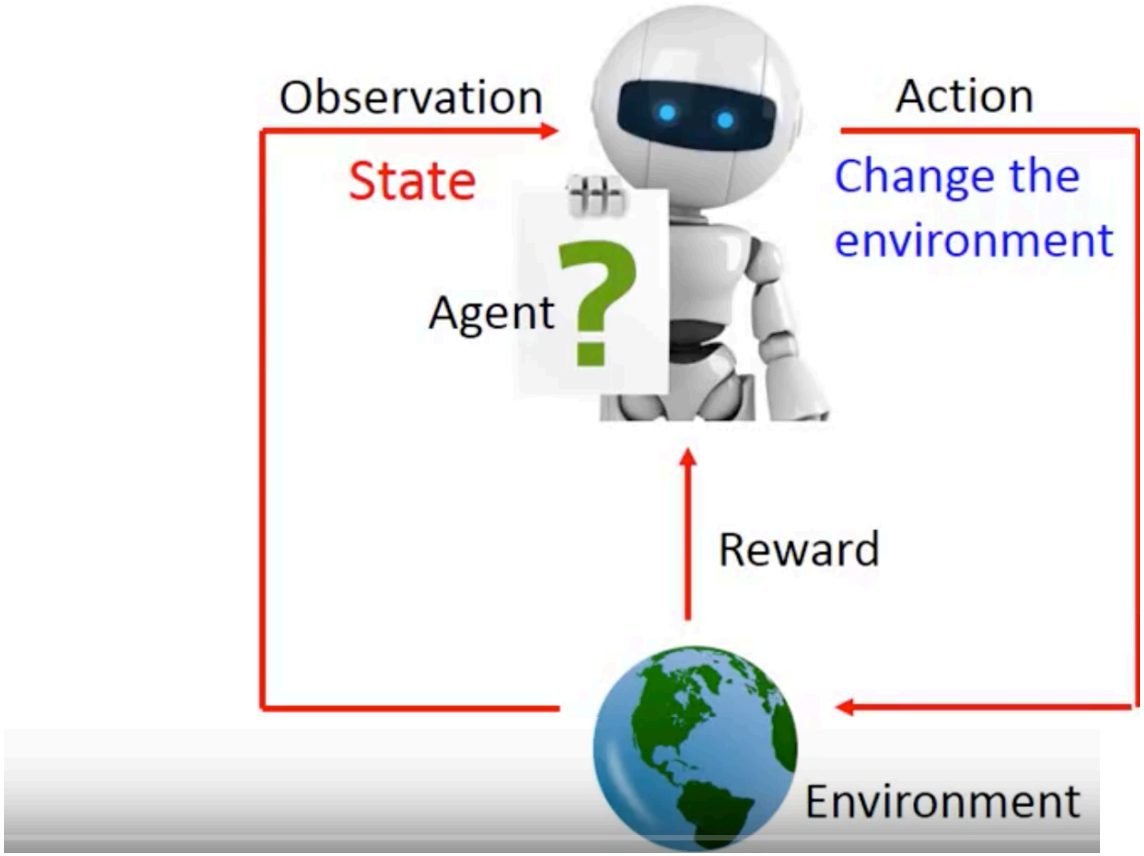
Scenario of Reinforcement Learning



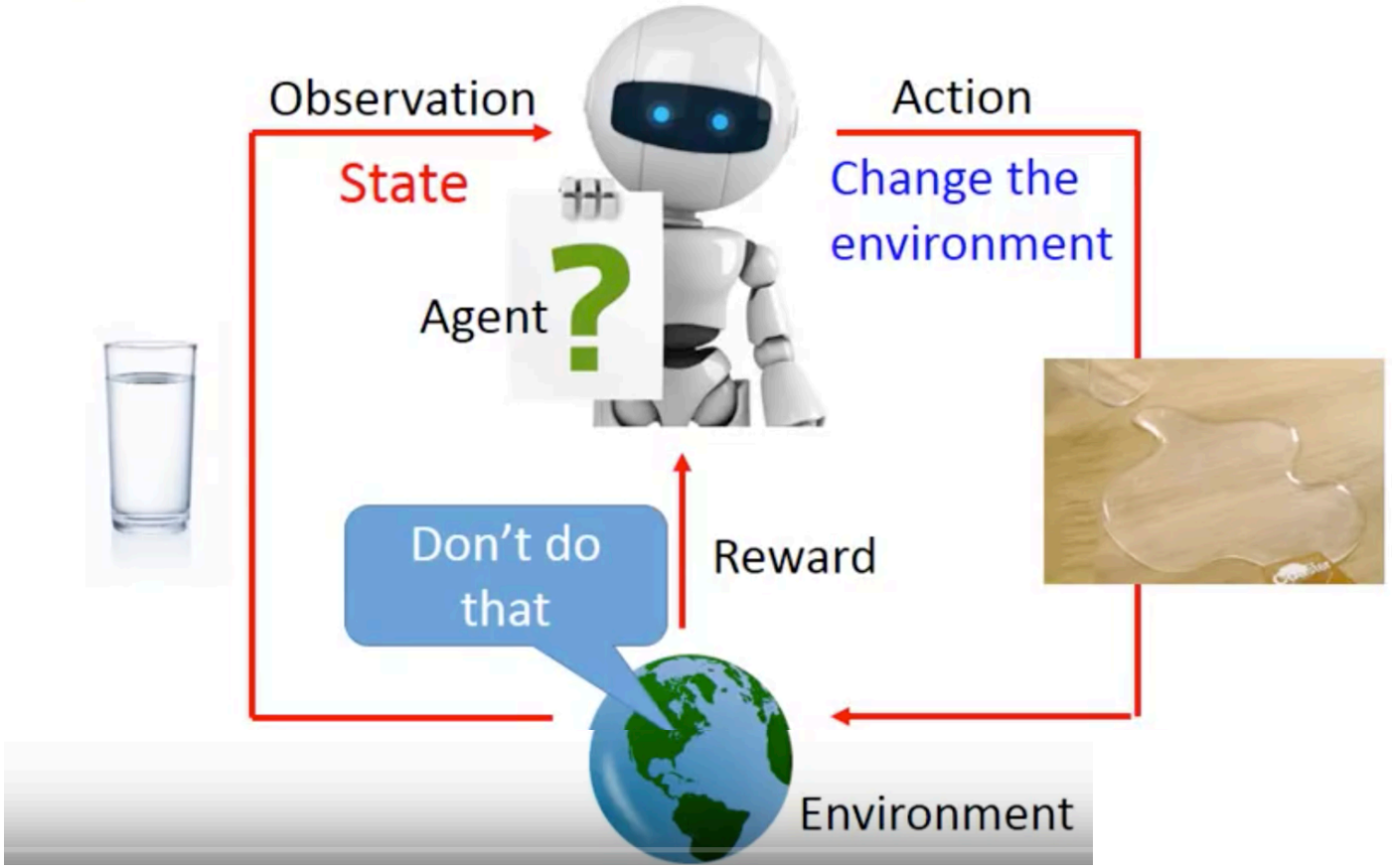
Scenario of Reinforcement Learning



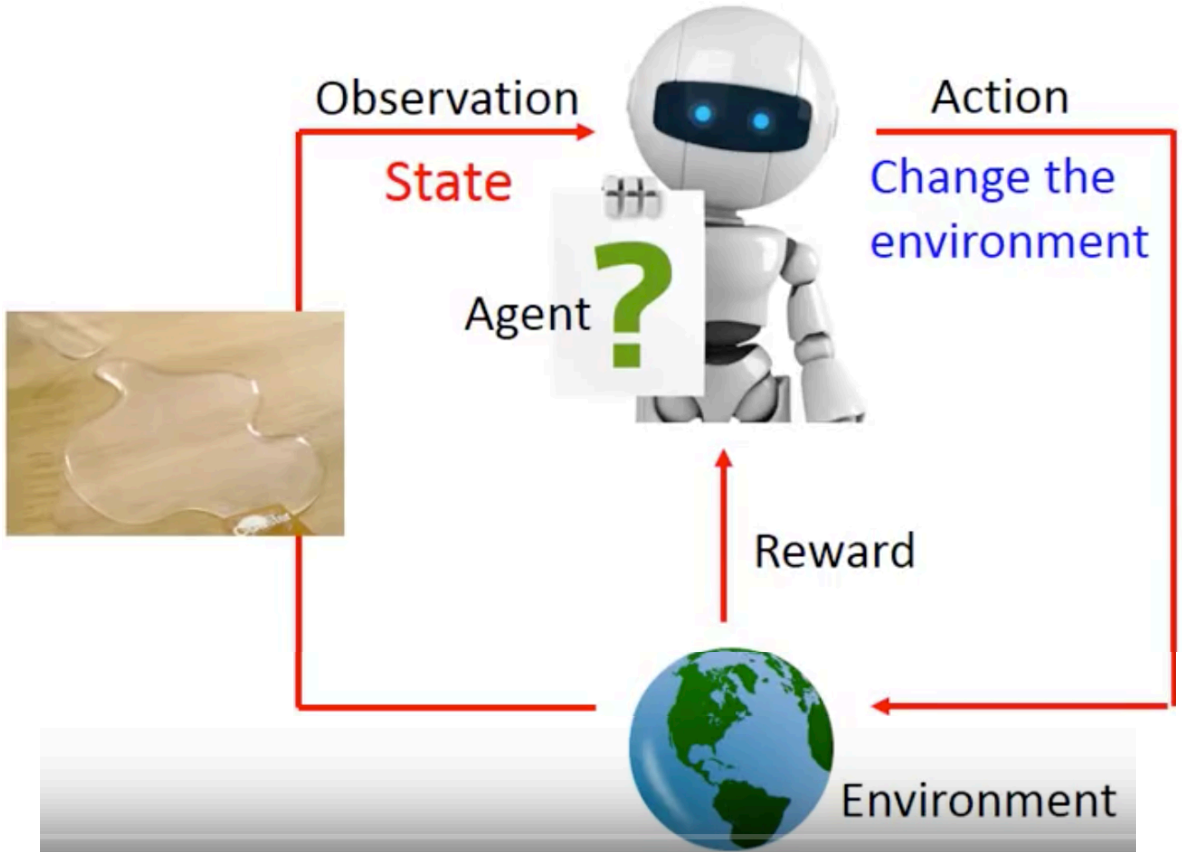
Scenario of Reinforcement Learning



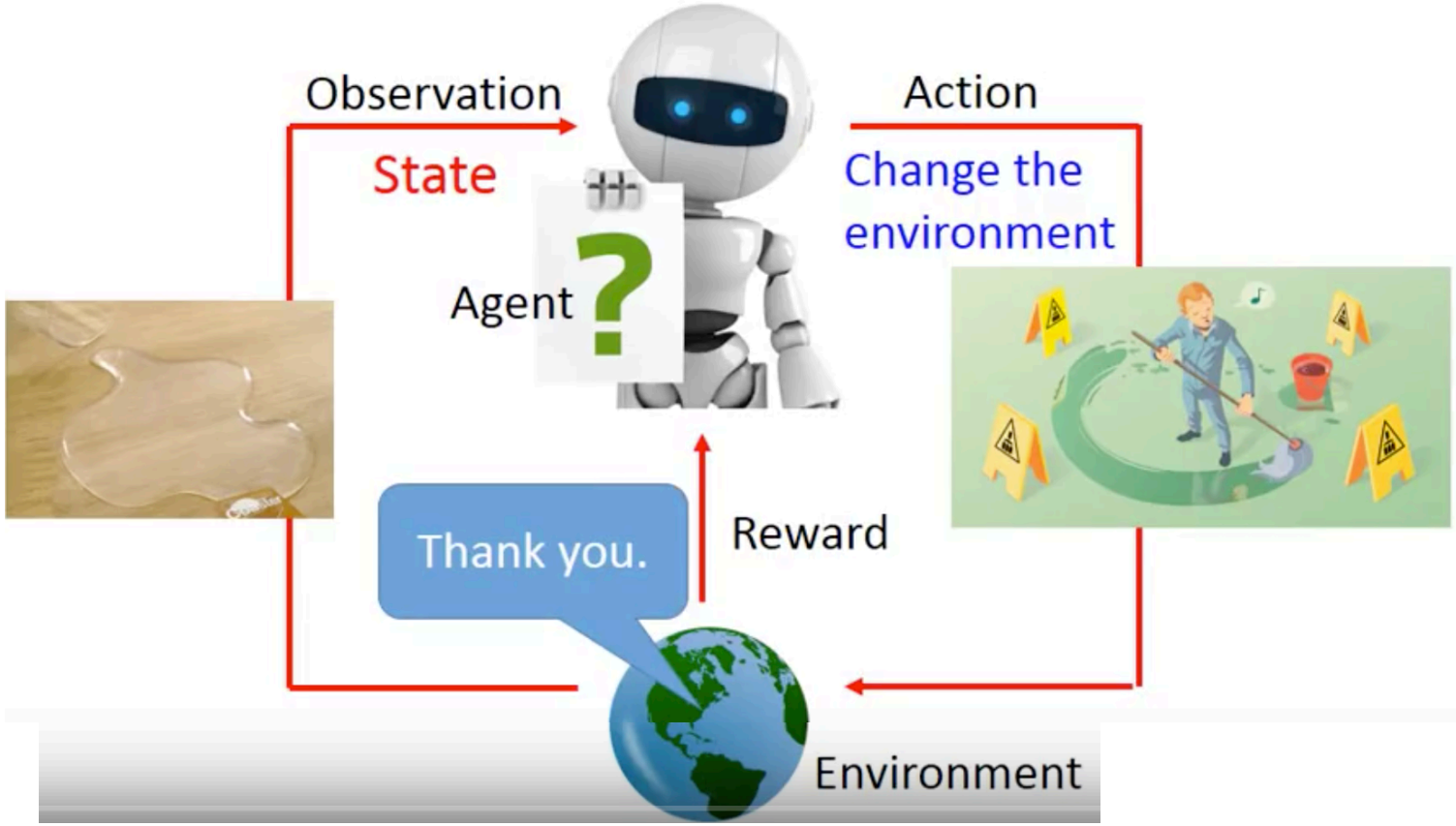
Scenario of Reinforcement Learning



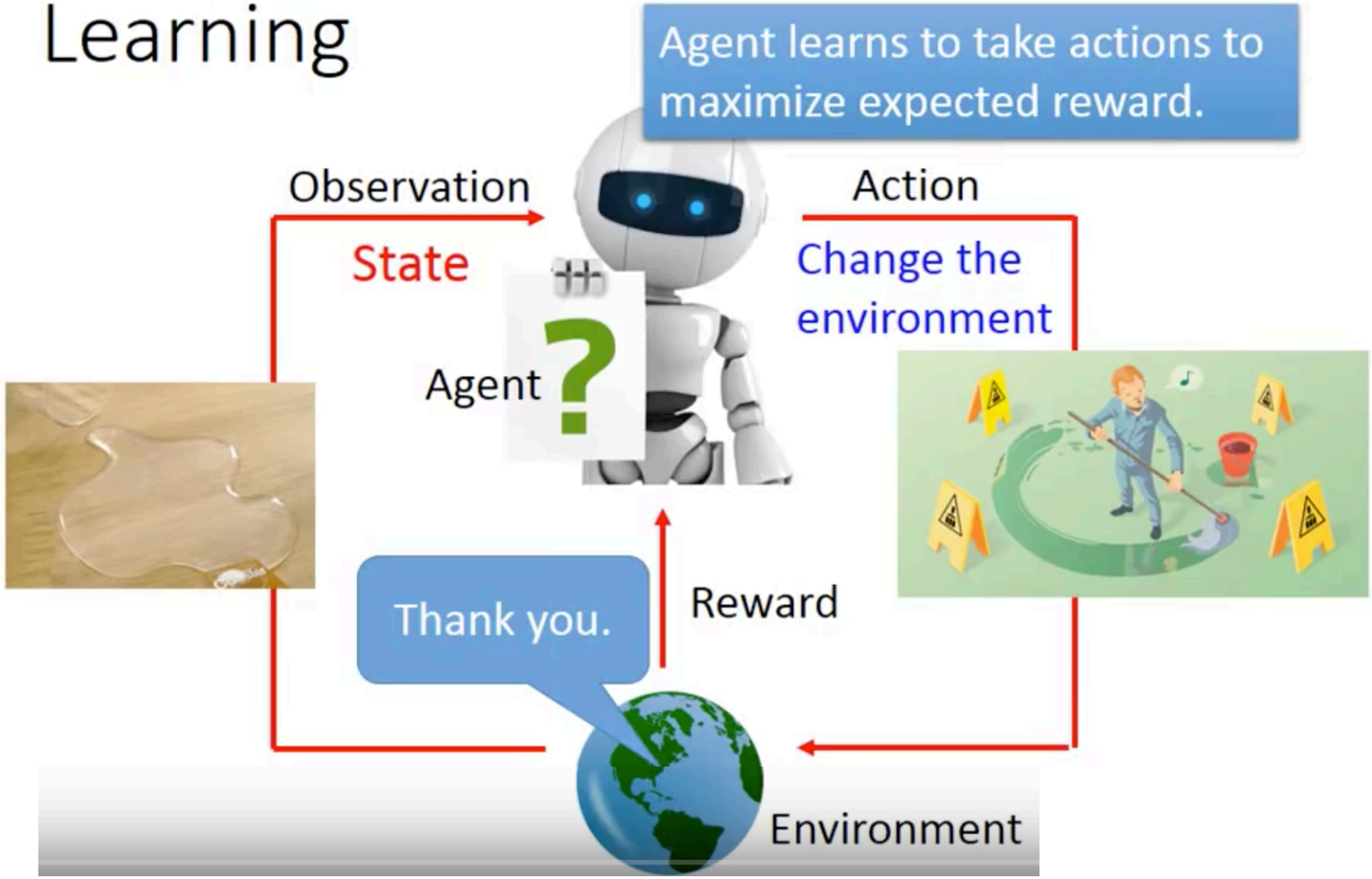
Scenario of Reinforcement Learning



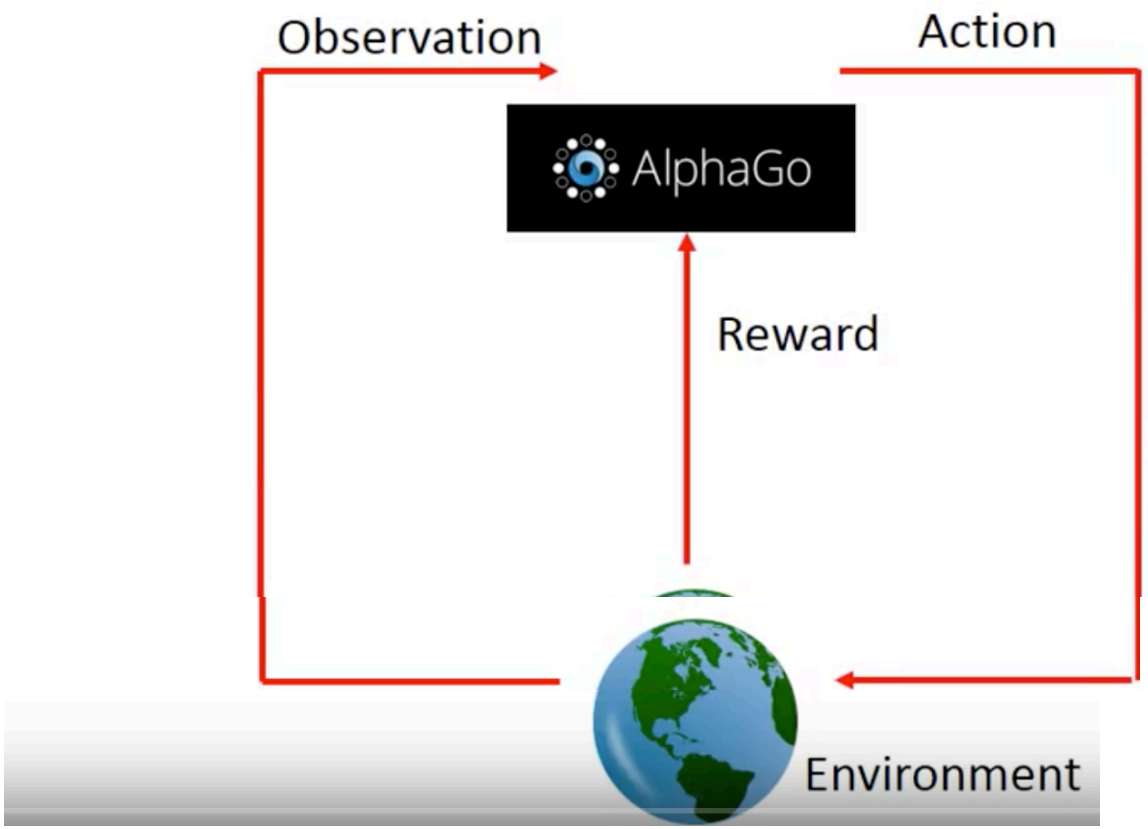
Scenario of Reinforcement Learning



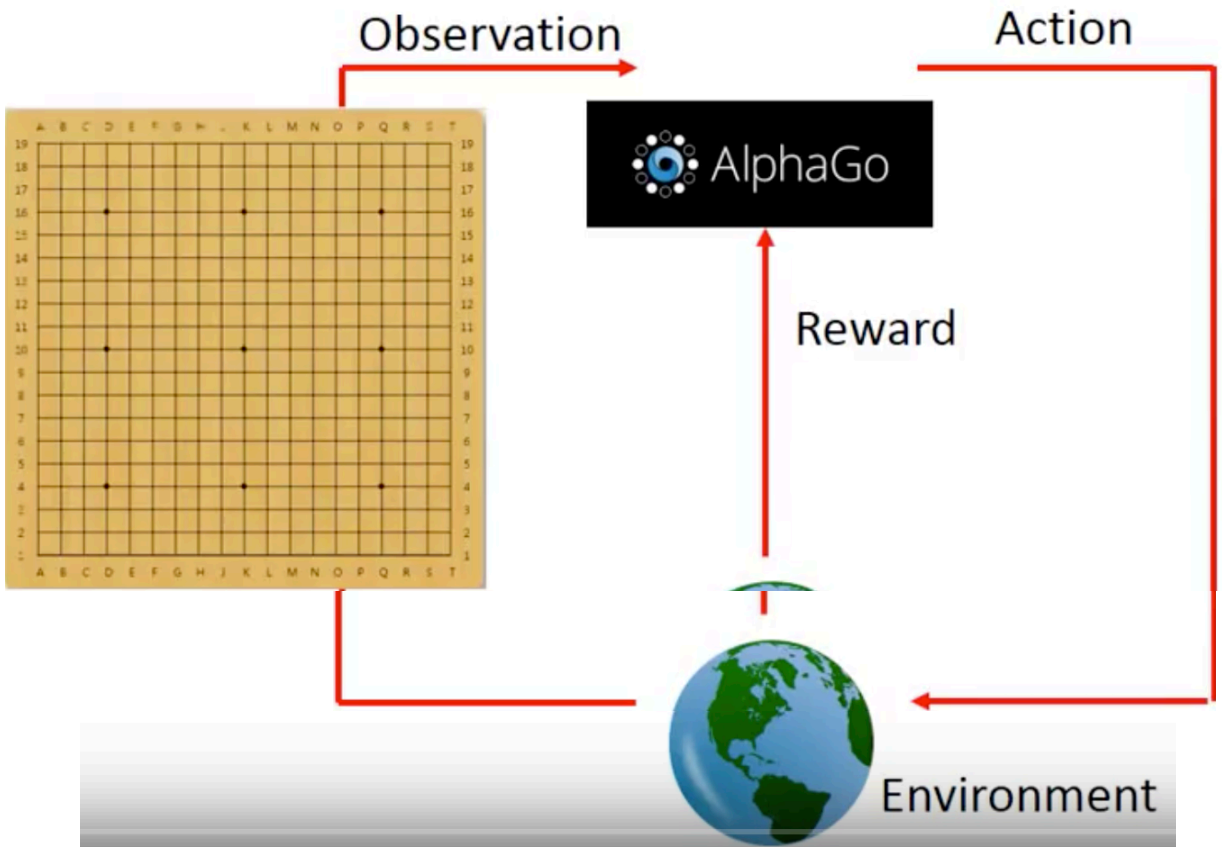
Scenario of Reinforcement Learning



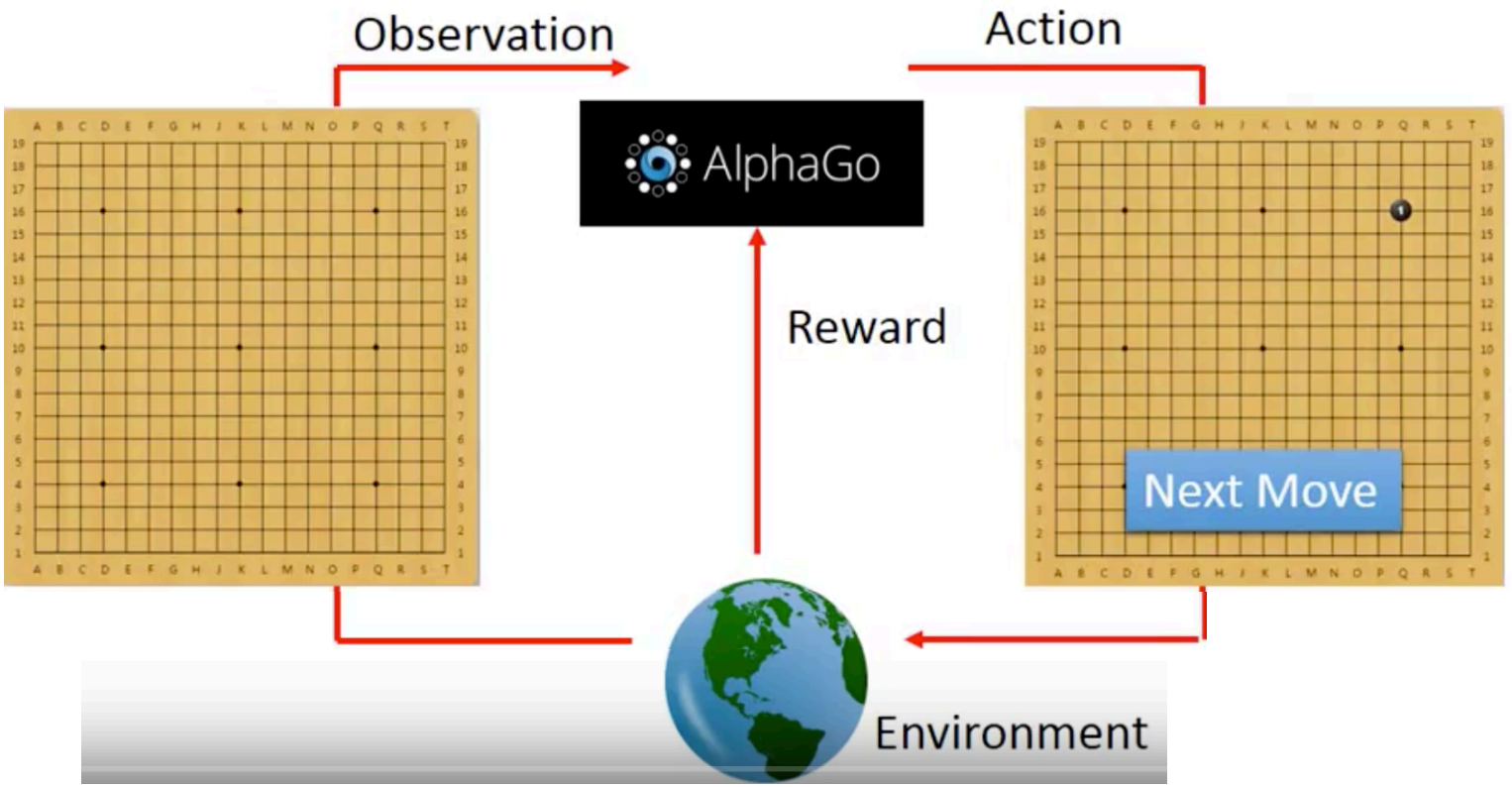
Learning to paly Go



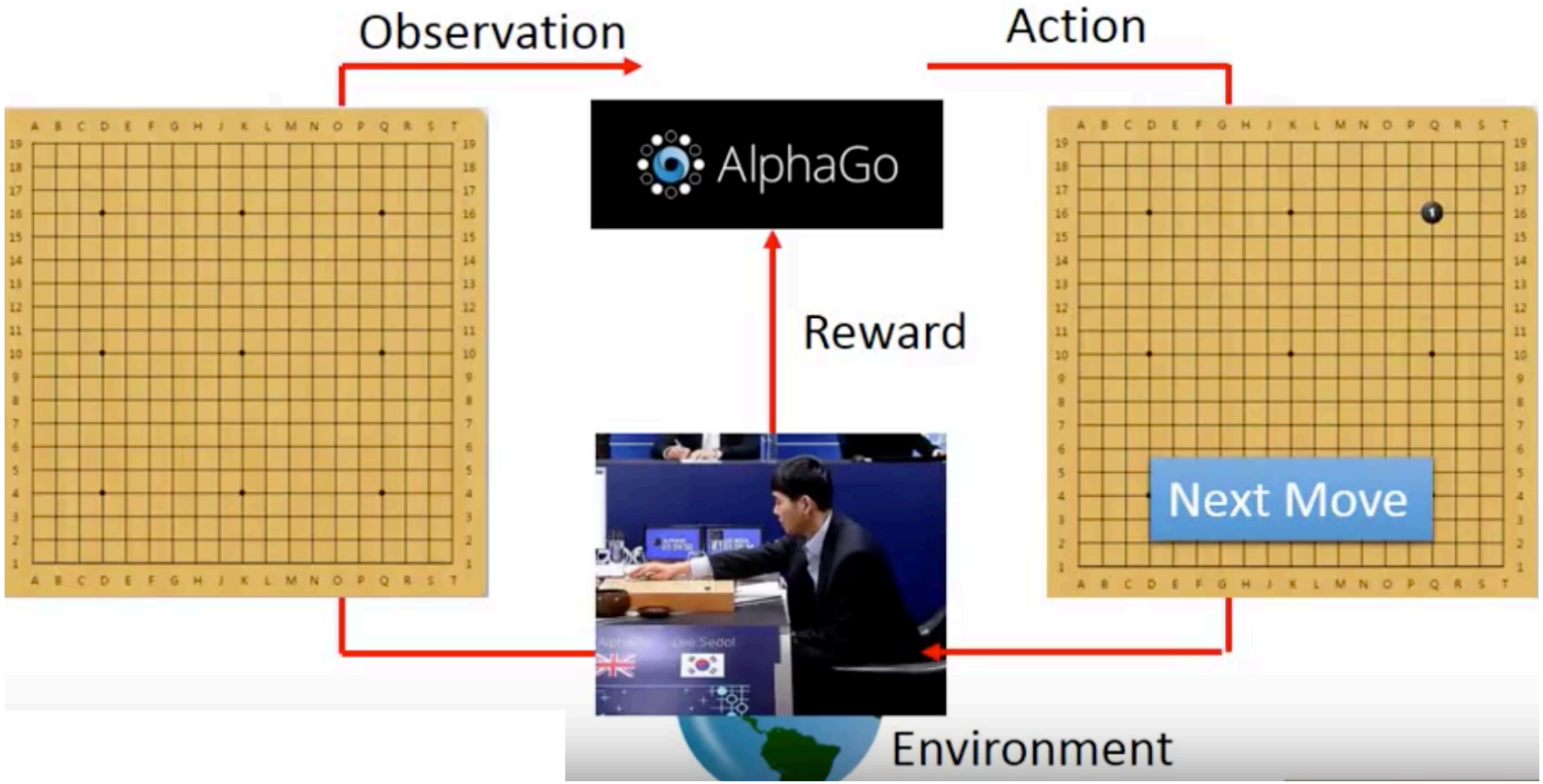
Learning to paly Go



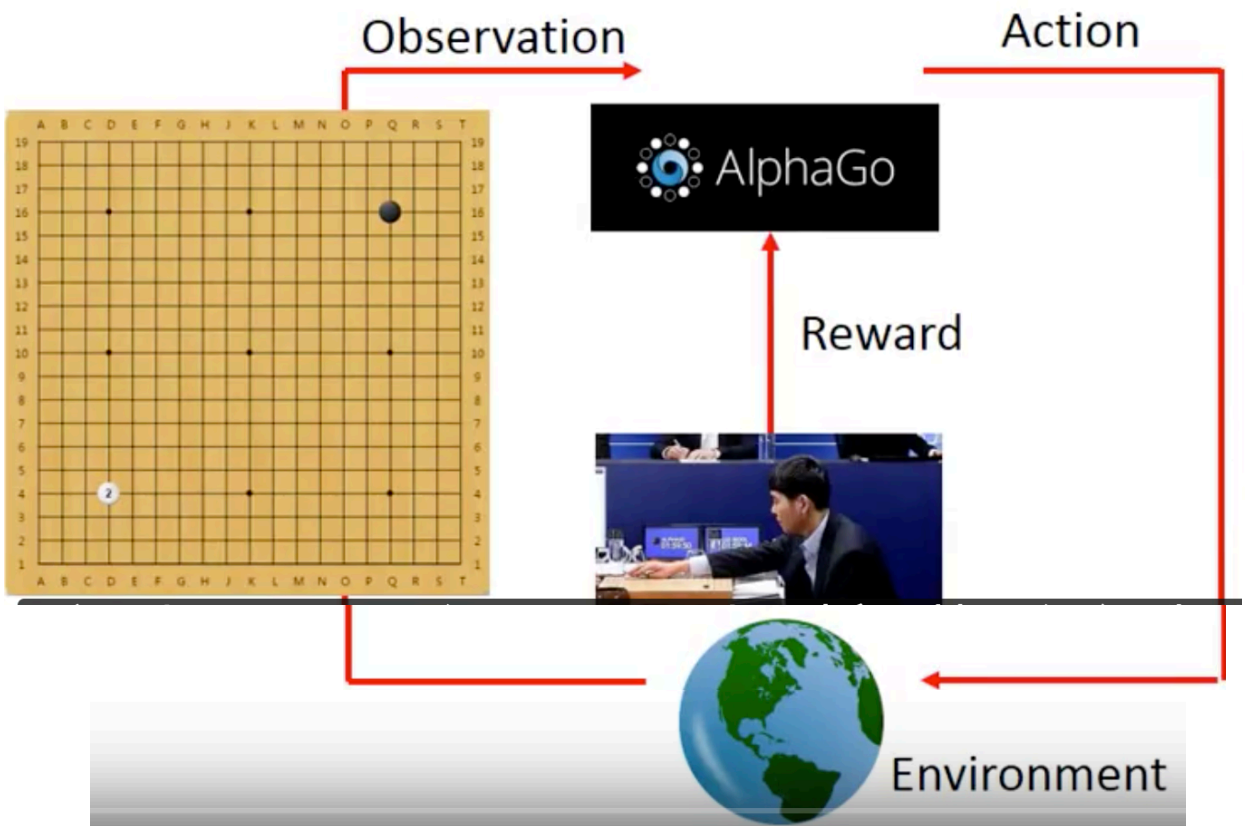
Learning to paly Go



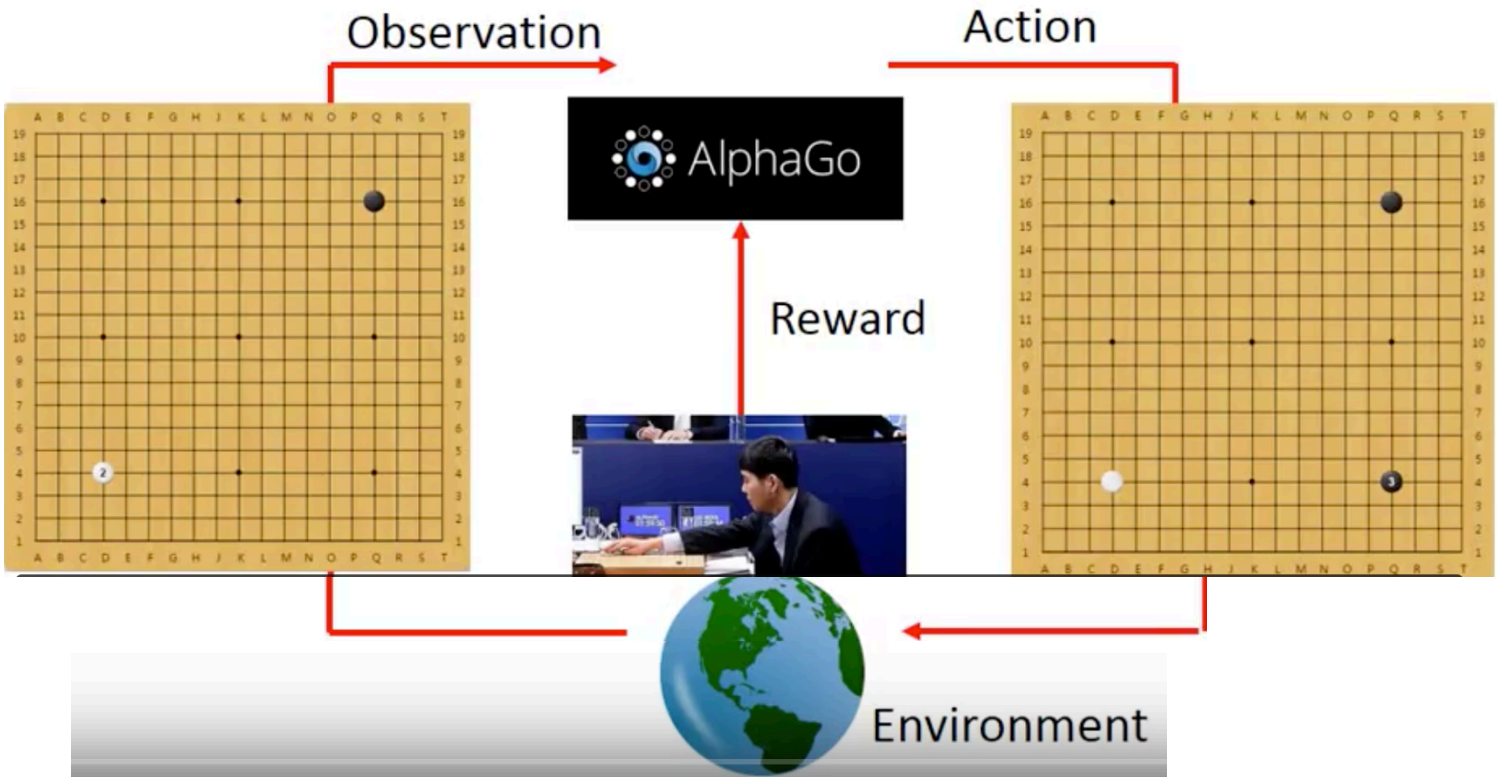
Learning to paly Go



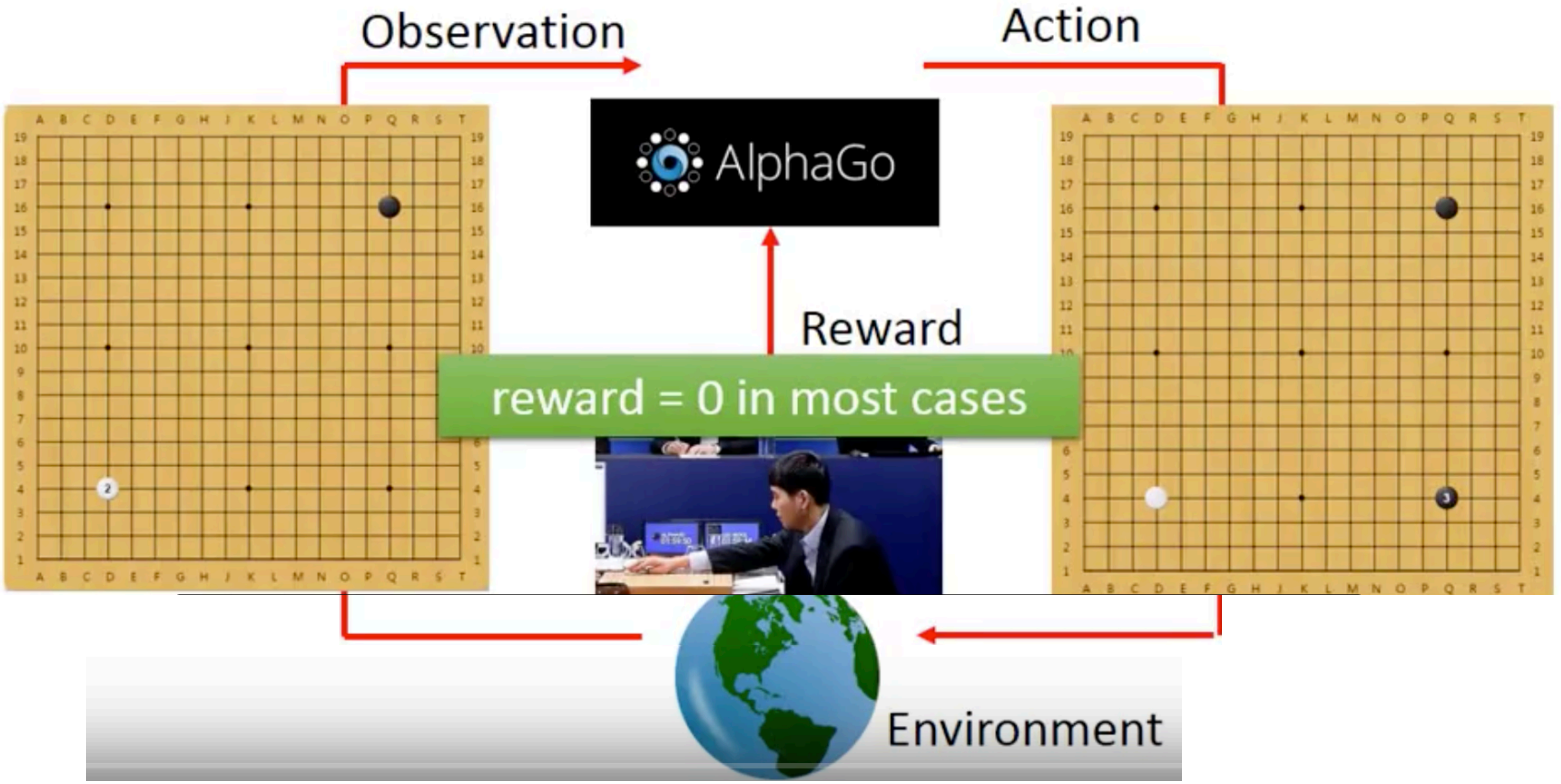
Learning to paly Go



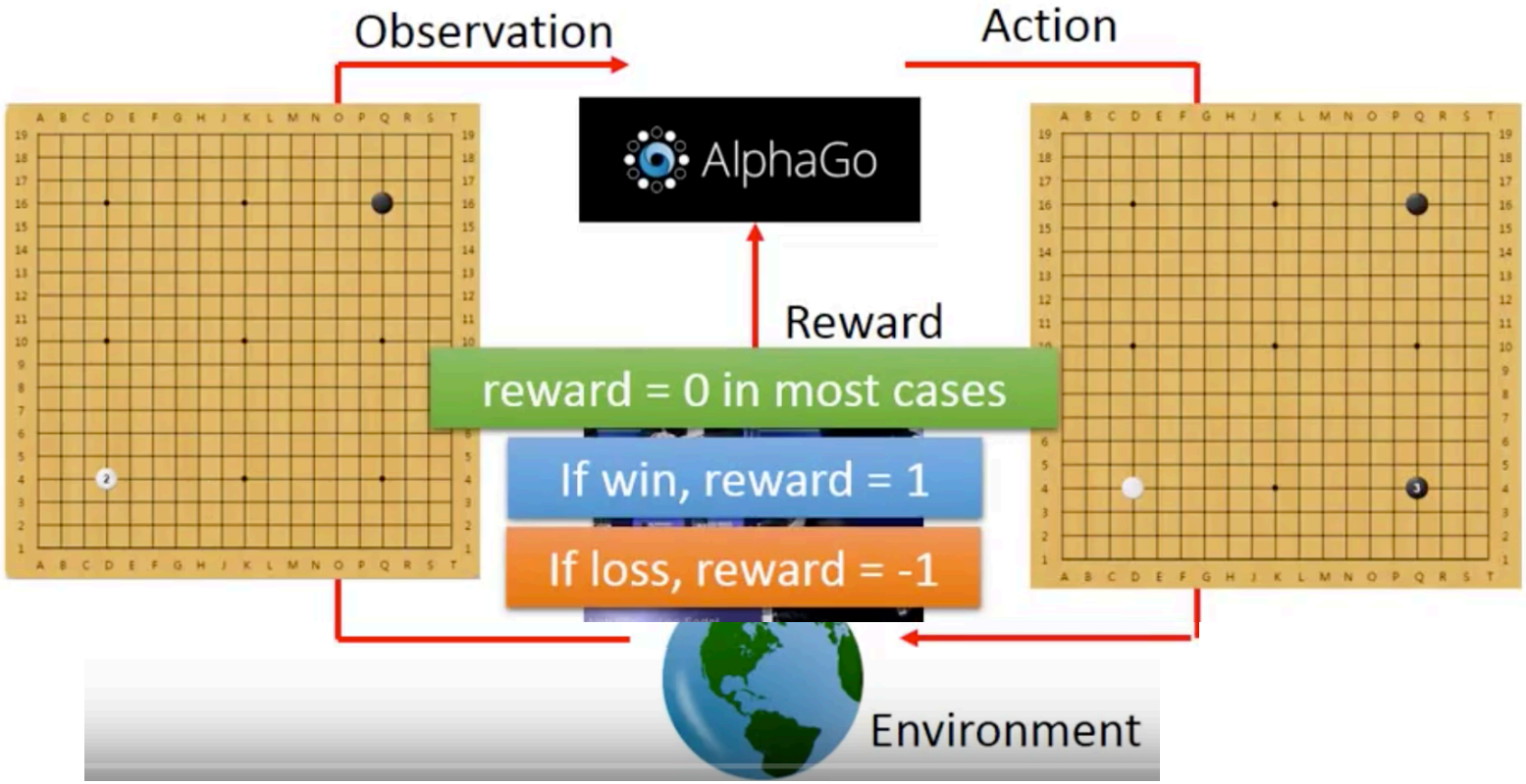
Learning to paly Go



Learning to paly Go

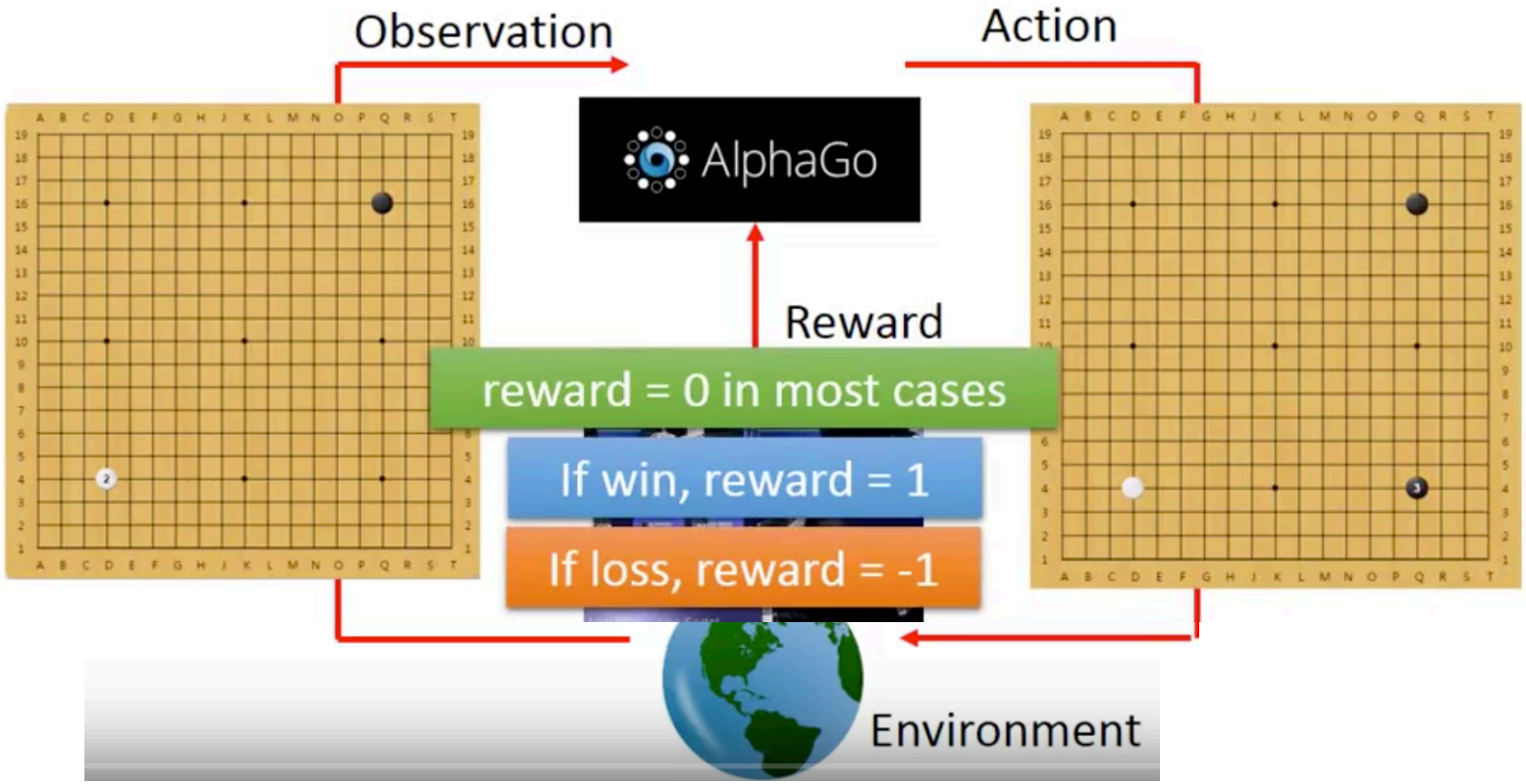


Learning to paly Go



Learning to paly Go

Agent learns to take actions to maximize expected reward.



Learning to paly Go
- Supervised v.s. Reinforcement

- Supervised:

- Reinforcement Learning

Learning to paly Go
- Supervised v.s. Reinforcement

- Supervised:

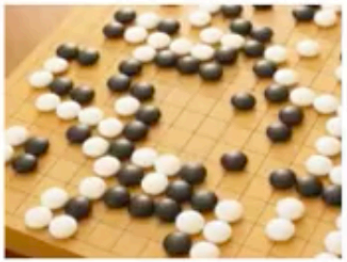
Next move:
"5-5"

Next move:
"3-3"

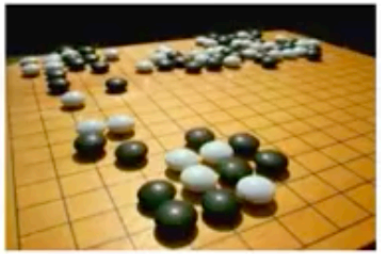
- Reinforcement Learning

Learning to paly Go
- Supervised v.s. Reinforcement

- Supervised: Learning from teacher



Next move:
"5-5"



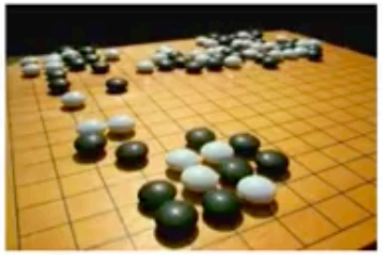
Next move:
"3-3"

- Reinforcement Learning

Learning to paly Go
- Supervised v.s. Reinforcement

- Supervised: Learning from teacher

Next move:
"5-5"



Next move:
"3-3"

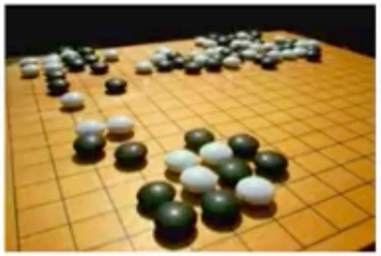
- Reinforcement Learning

First move ⇨ …… many moves ……

Learning to paly Go
- Supervised v.s. Reinforcement

- Supervised: Learning from teacher

Next move:
"5-5"



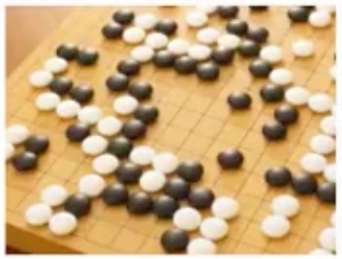
Next move:
"3-3"

- Reinforcement Learning

First move → ..... many moves ..... → Win!

Learning to paly Go
- Supervised v.s. Reinforcement

- Supervised: Learning from teacher



Next move:
"5-5"

Next move:
"3-3"

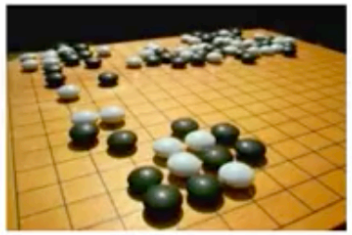
- Reinforcement Learning Learning from experience

First move → ..... many moves ..... → Win!

Learning to paly Go
- Supervised v.s. Reinforcement

- Supervised: Learning from teacher

Next move:
"5-5"



Next move:
"3-3"

- Reinforcement Learning Learning from experience

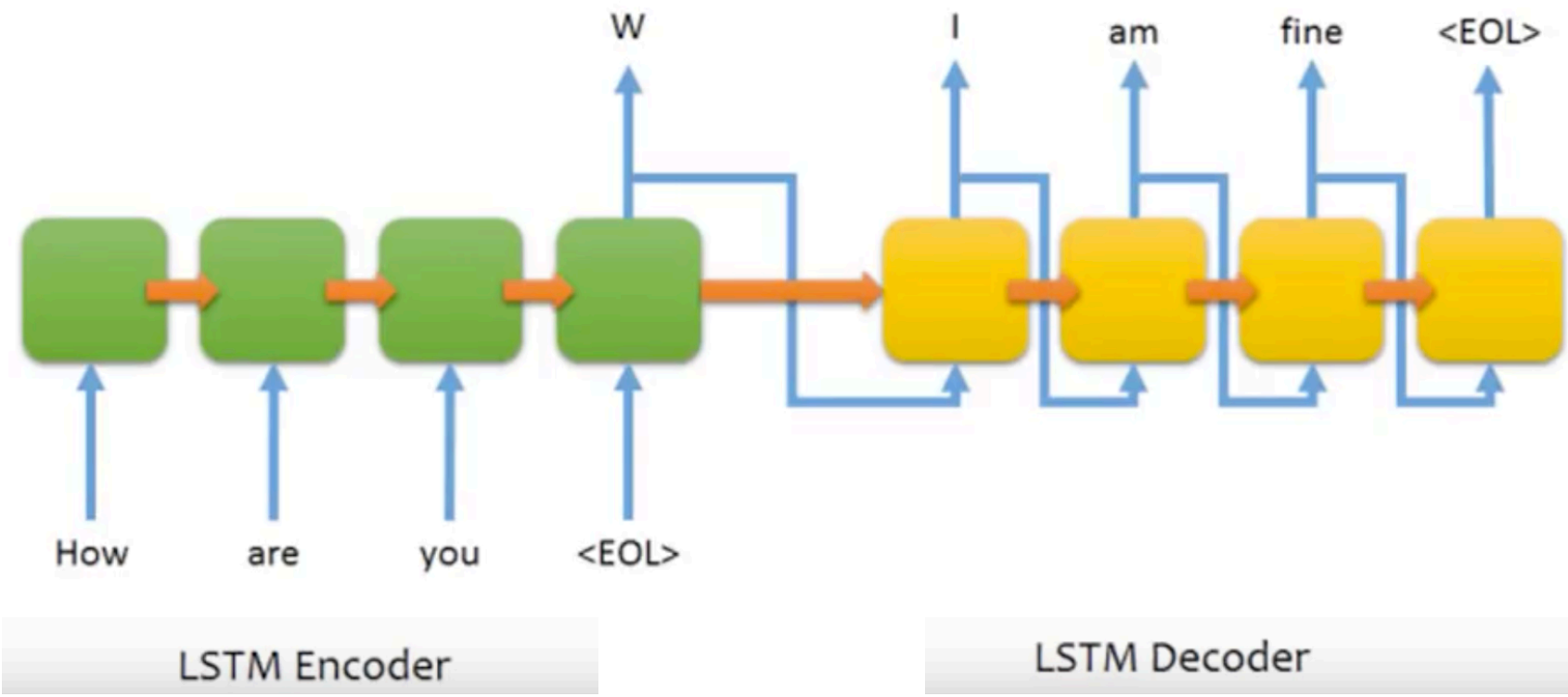
First move → ..... many moves ..... → Win!

(Two agents play with each other.)

Alpha Go is supervised learning + reinforcement learning.

Learning a chat-bot

- Sequence-to-sequence learning



Learning a chat-bot
- Supervised v.s. Reinforcement

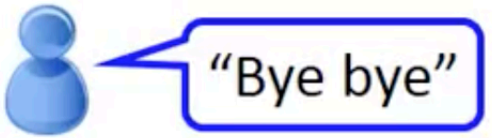
- Supervised

- Reinforcement

Learning a chat-bot
- Supervised v.s. Reinforcement

- Supervised

Say "Hi"



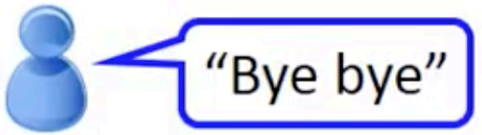
Say "Good bye"

- Reinforcement

Learning a chat-bot
- Supervised v.s. Reinforcement

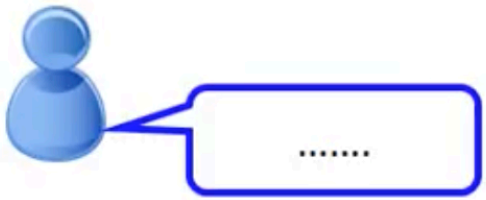
- Supervised

Say "Hi"



Say "Good bye"

- Reinforcement



.....

Hello ☺

.....

Bad

More applications

- Flying Helicopter
  - https://www.youtube.com/watch?v=0JL04JJjocc
- Driving
  - https://www.youtube.com/watch?v=0xo1Ldx3L5Q
- Google Cuts Its Giant Electricity Bill With DeepMind-Powered AI
  - http://www.bloomberg.com/news/articles/2016-07-19/google-cuts-its-giant-electricity-bill-with-deepmind-powered-ai
- Text generation
  - Hongyu Guo, “Generating Text with Deep Reinforcement Learning”, NIPS, 2015
  - Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, Wojciech Zaremba, “Sequence Level Training with Recurrent Neural Networks”, ICLR, 2016

Example: Playing Video Game

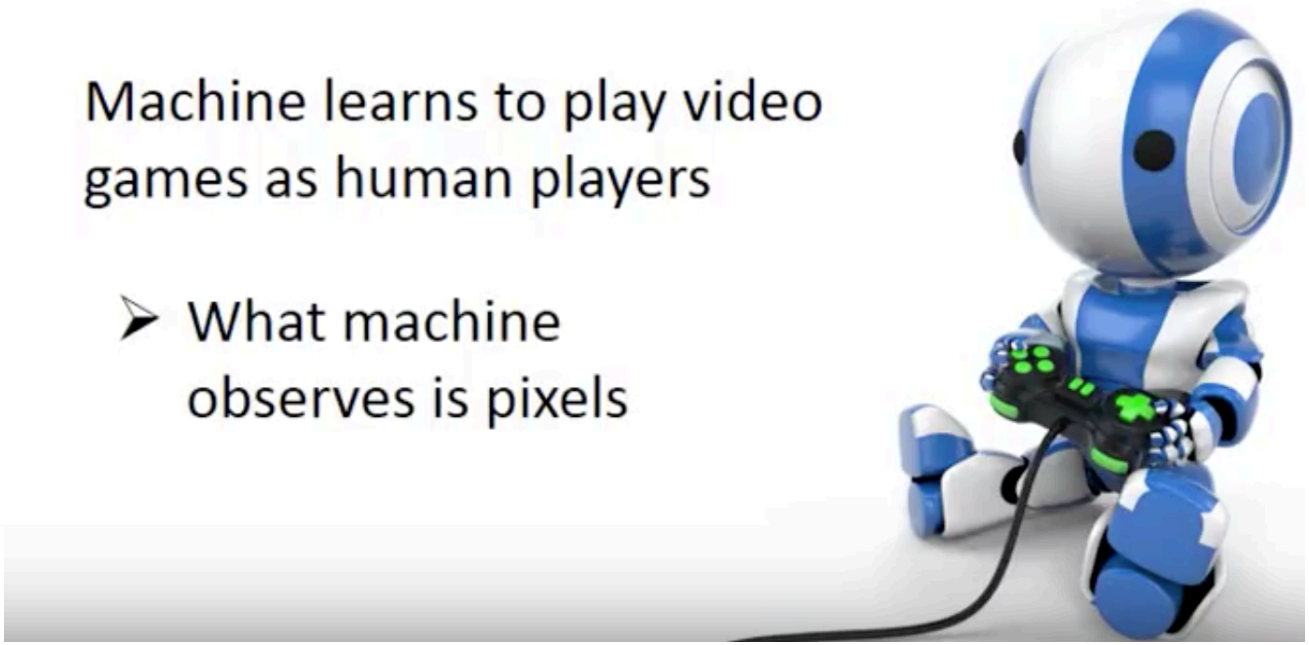
- Widely studies:
  - Gym: https://gym.openai.com/
  - Universe: https://openai.com/blog/universe/

Example: Playing Video Game

- Widely studies:
  - Gym: https://gym.openai.com/
  - Universe: https://openai.com/blog/universe/

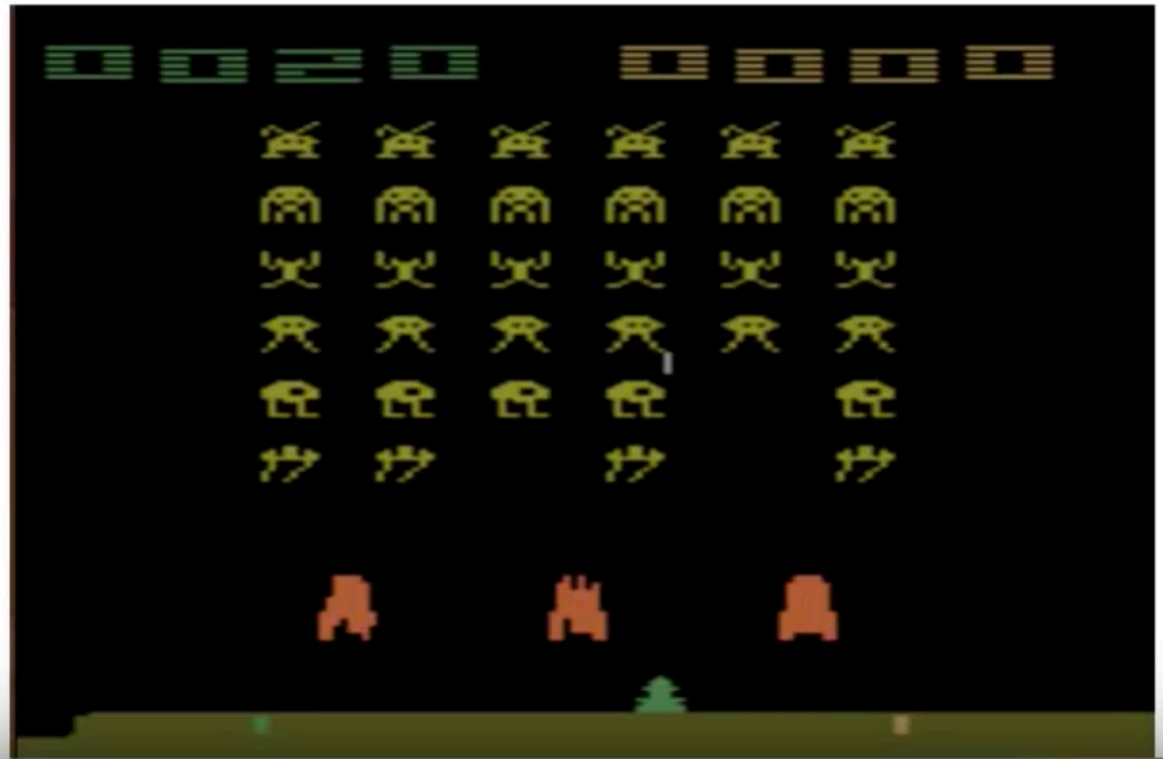
Machine learns to play video games as human players

- What machine observes is pixels



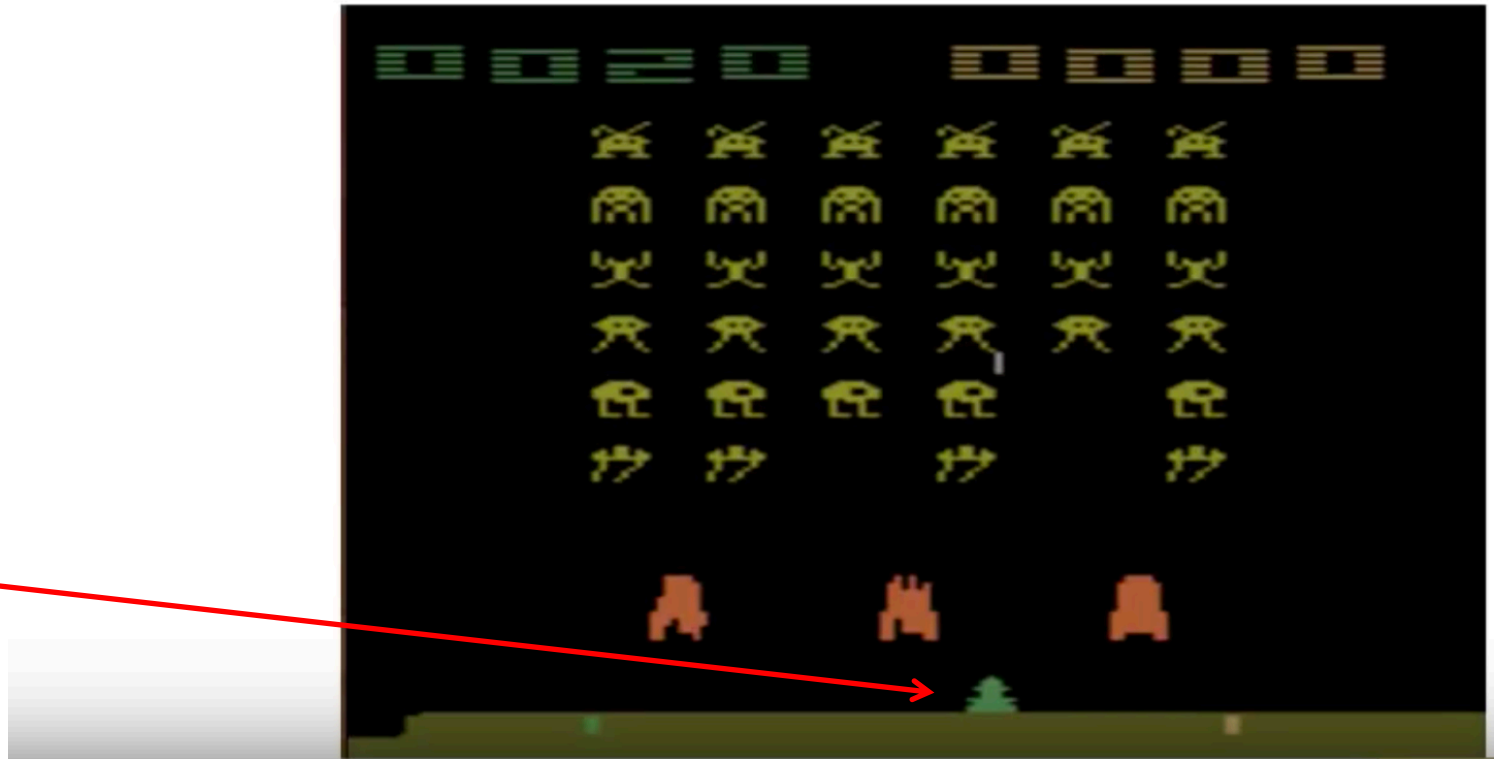
Example: Playing Video Game

- Space invader



Example: Playing Video Game

- Space invader

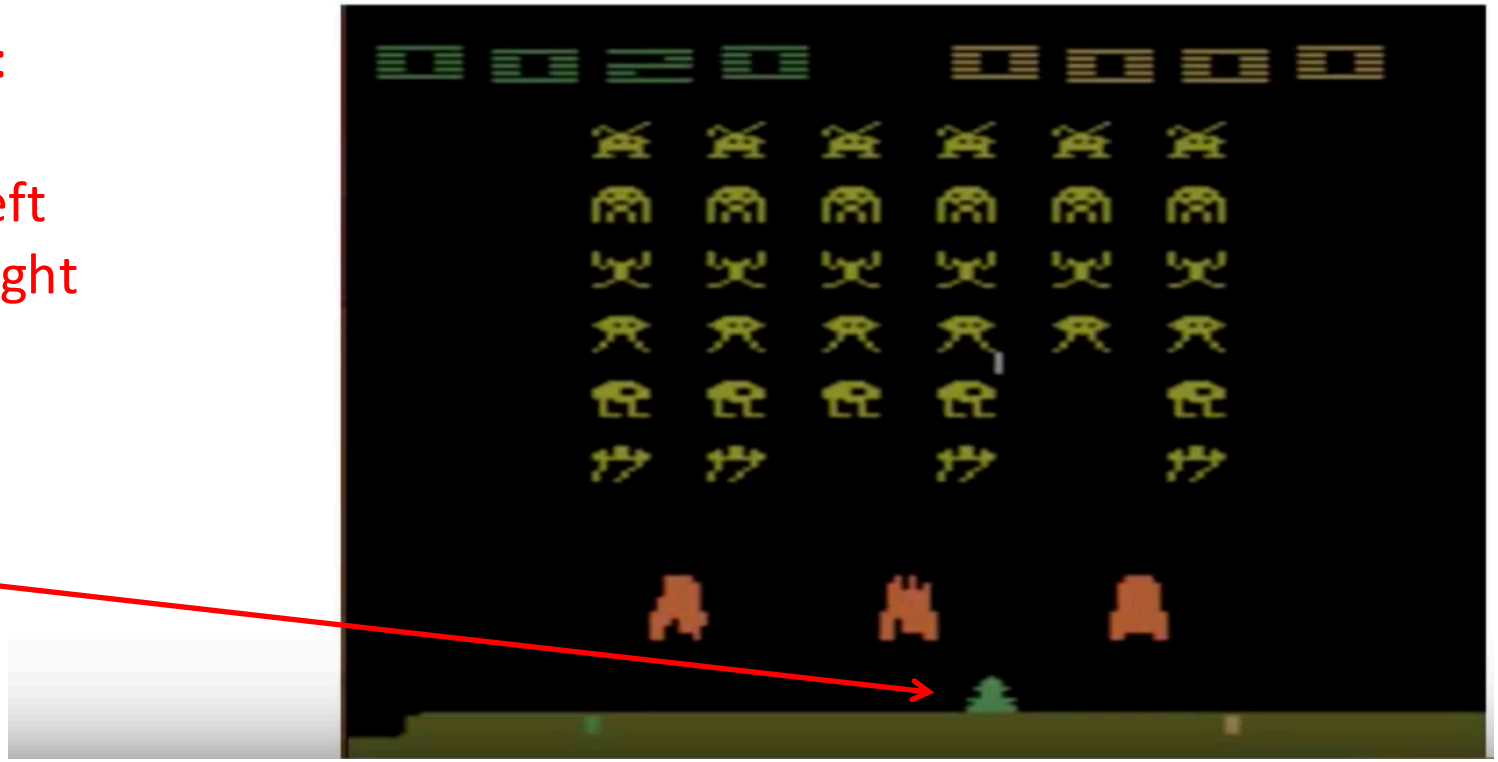


Example: Playing Video Game

- Space invader

Actions:

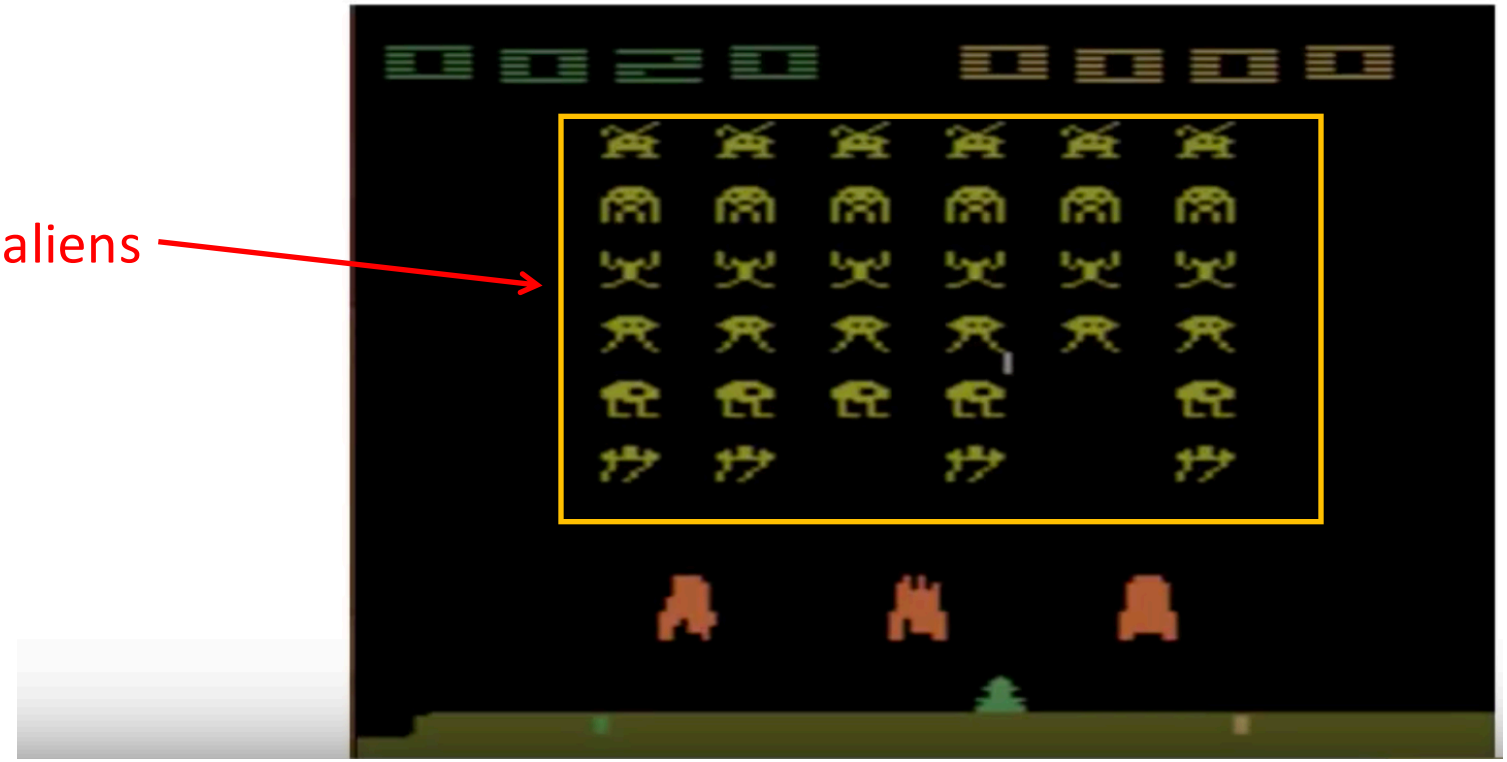
Move left
Move right
Fire



Agent

Example: Playing Video Game

- Space invader

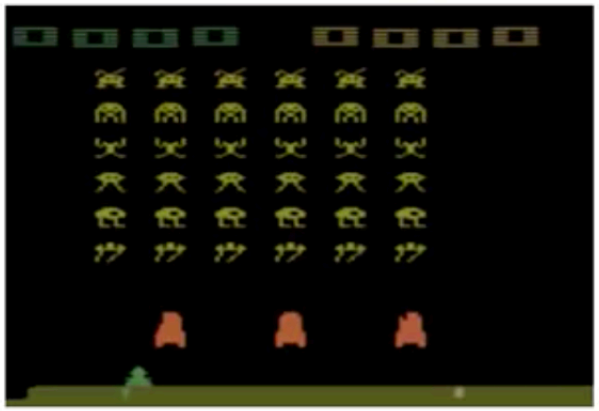


Example: Playing Video Game

https://www.youtube.com/watch?v=_ftVrgJTl4w

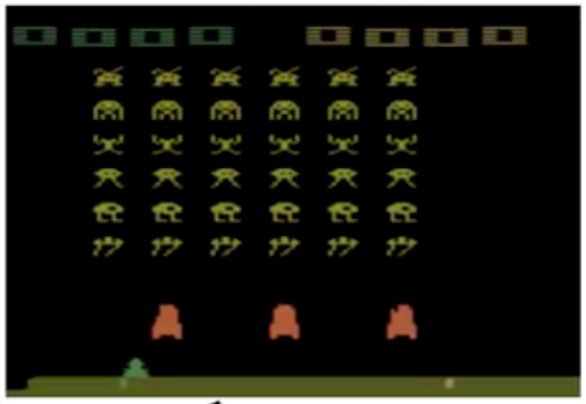
Example: Playing Video Game

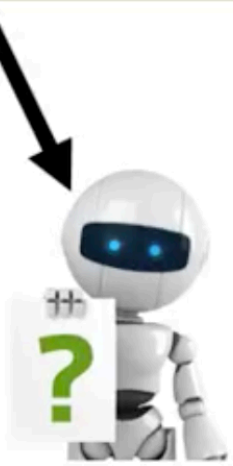
Start with observation s_1



Example: Playing Video Game

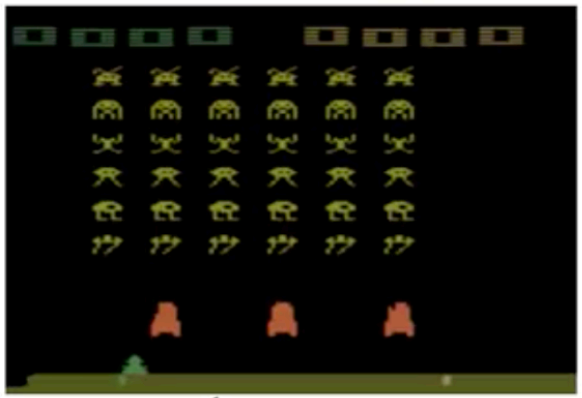
Start with observation s_1





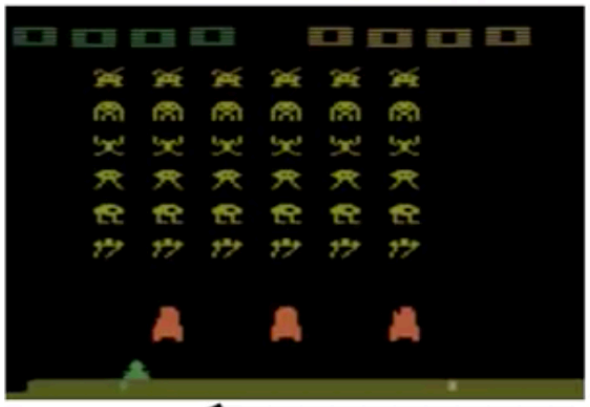
Example: Playing Video Game

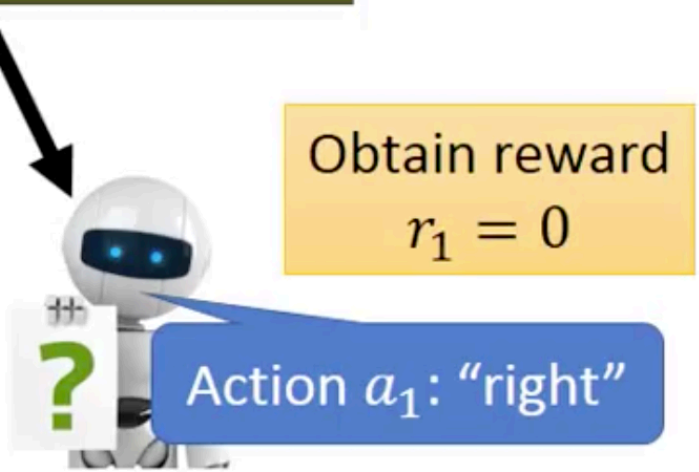
Start with observation s_1



Example: Playing Video Game

Start with observation s_1

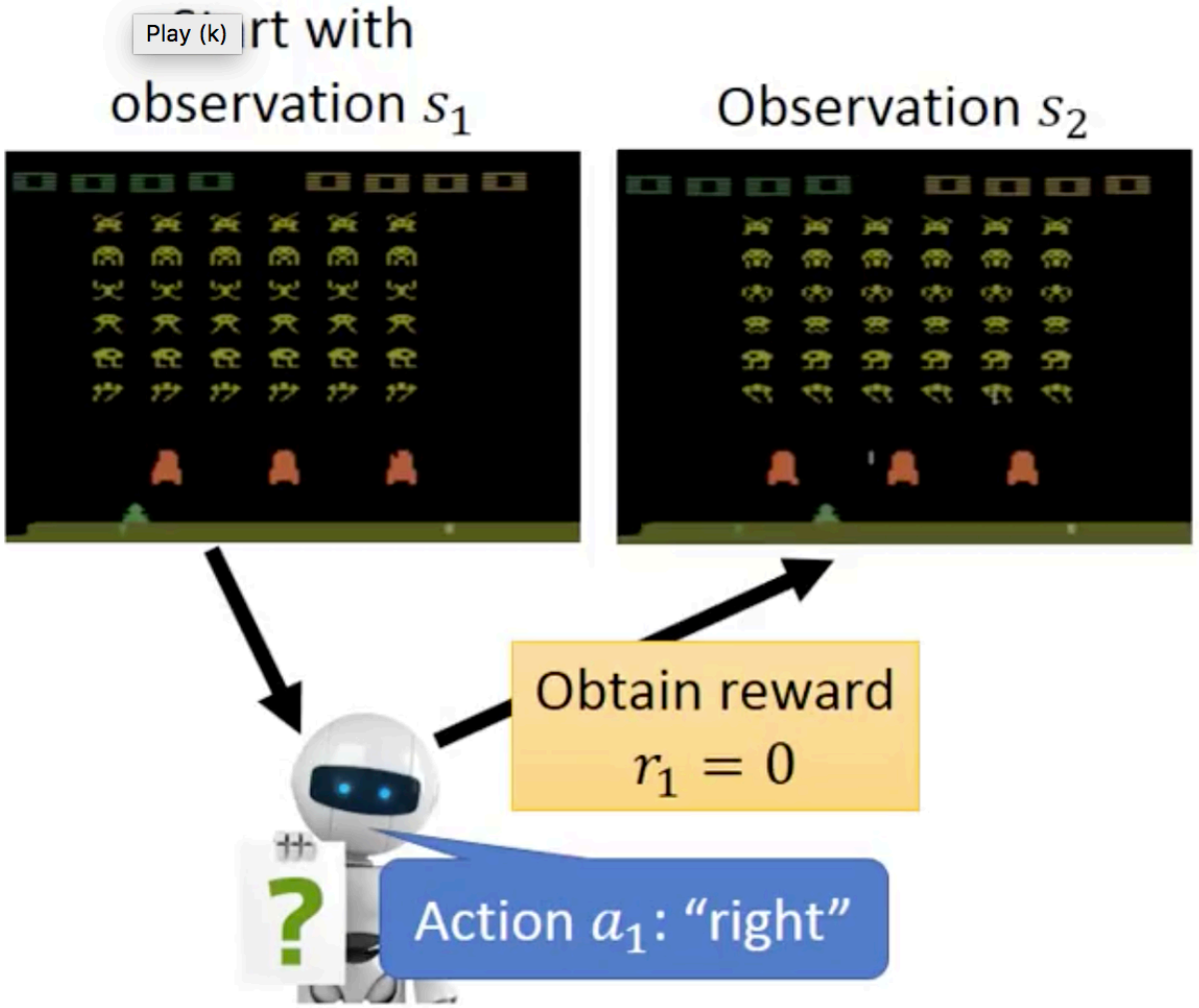




Obtain reward $r_1 = 0$

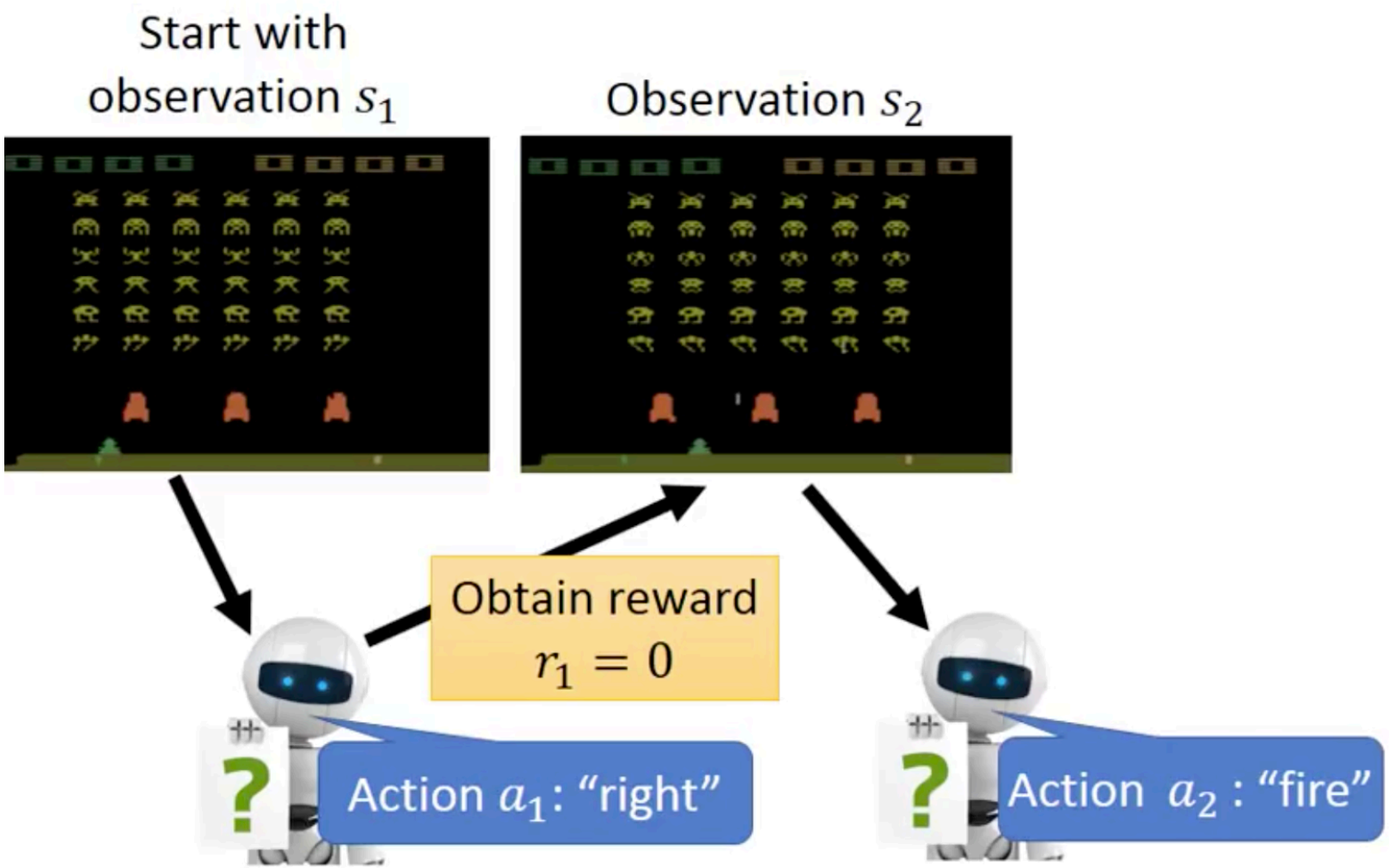
Action a_1: "right"

Example: Playing Video Game

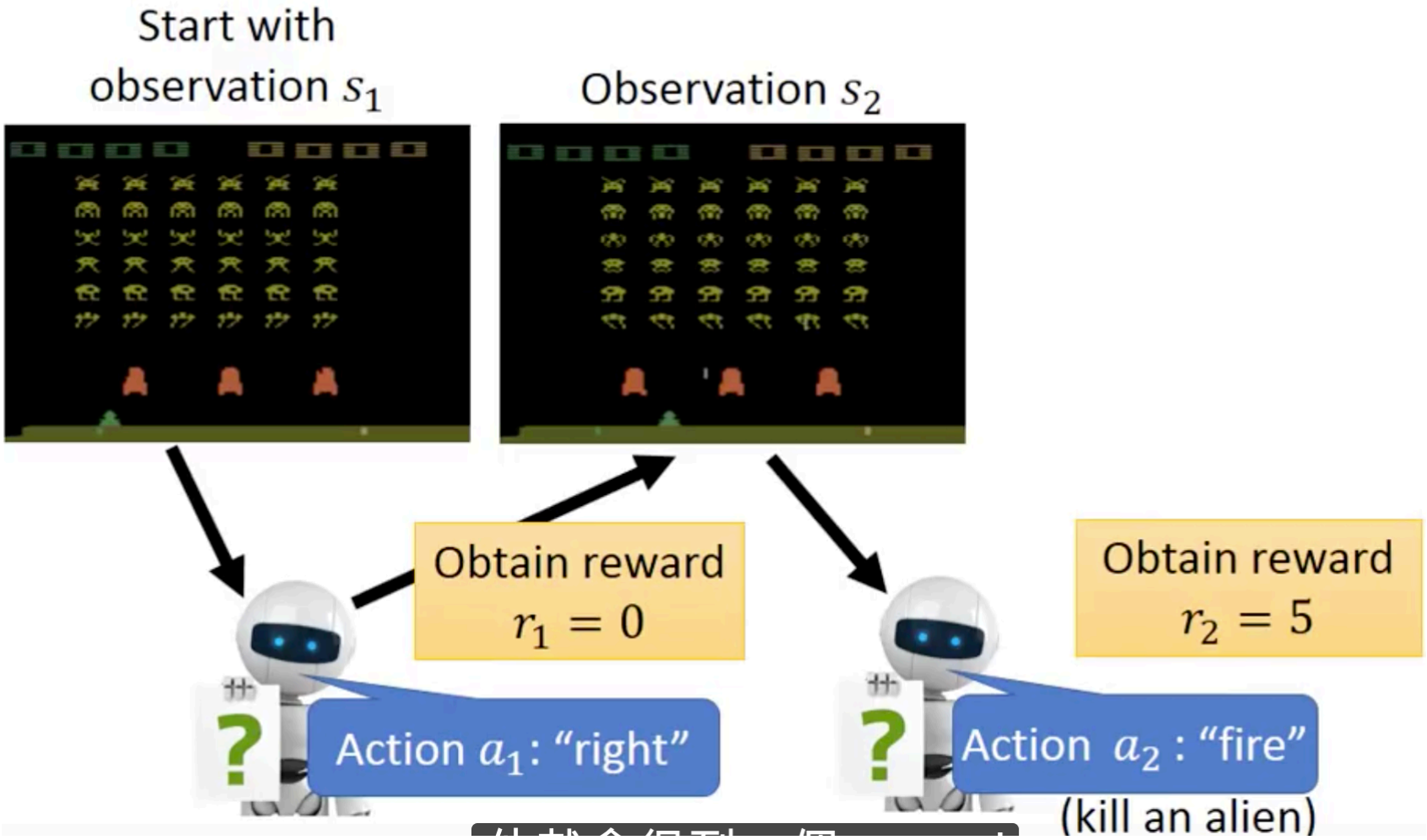


Example: Playing Video Game

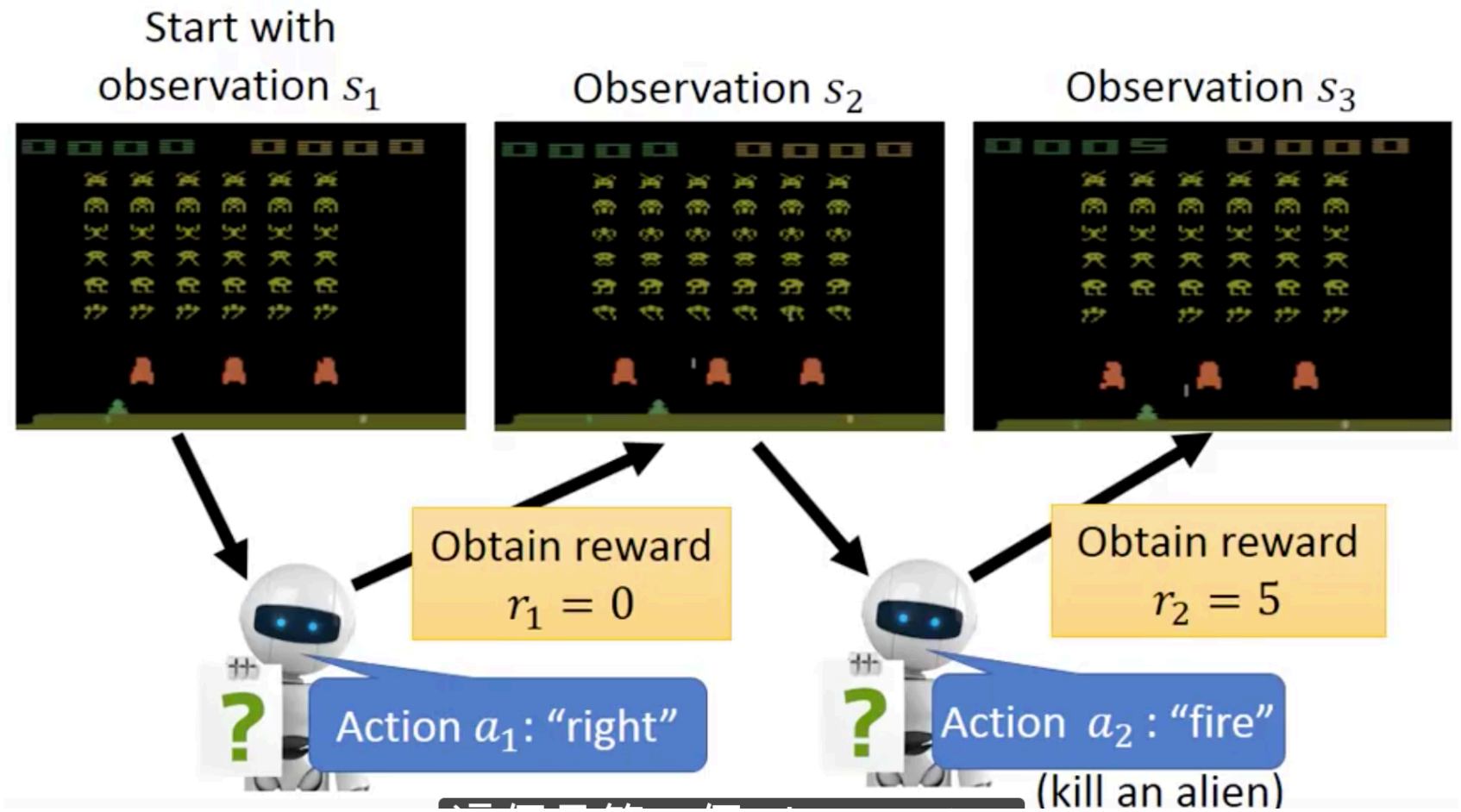
Play (k)



Example: Playing Video Game

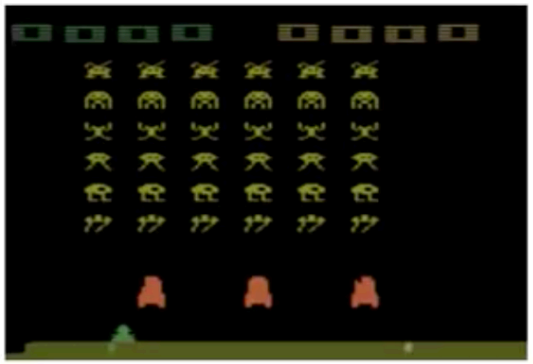


Example: Playing Video Game

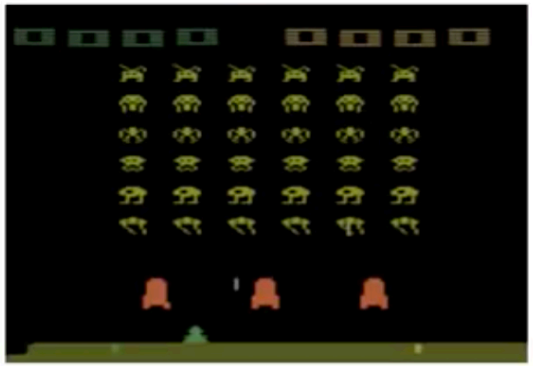


Example: Playing Video Game

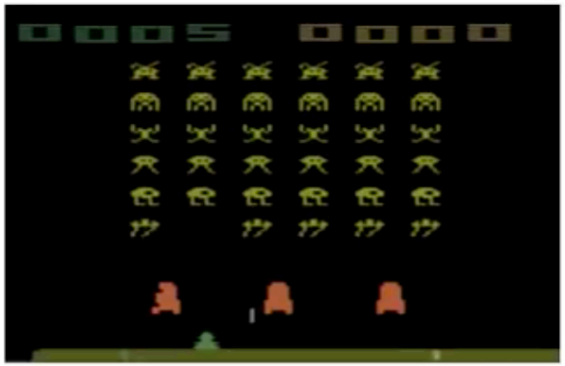
Start with observation s_1



Observation s_2



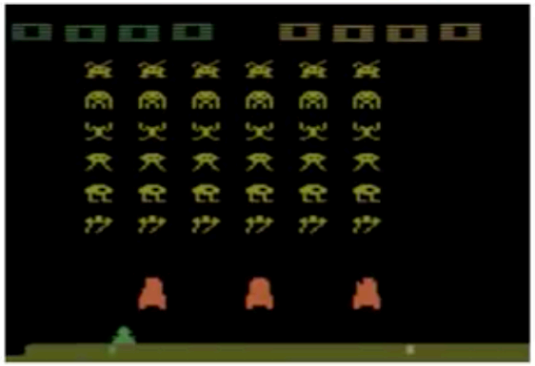
Observation s_3



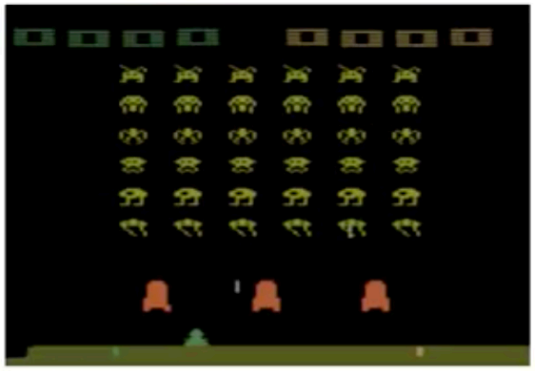
After many turns

Example: Playing Video Game

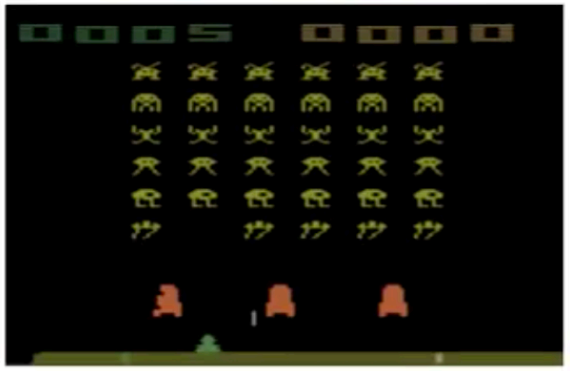
Start with observation s_1



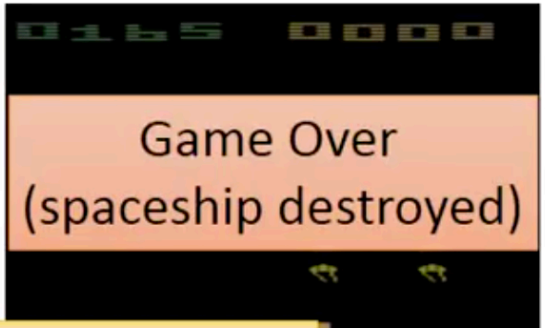
Observation s_2



Observation s_3



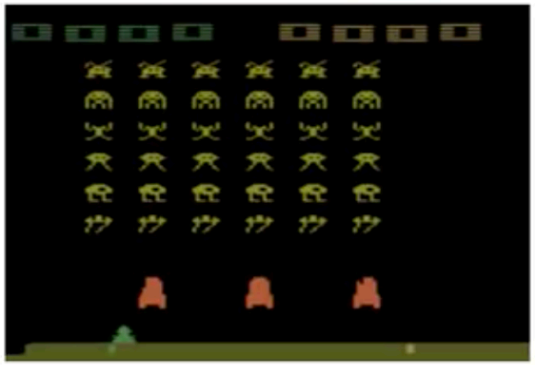
After many turns



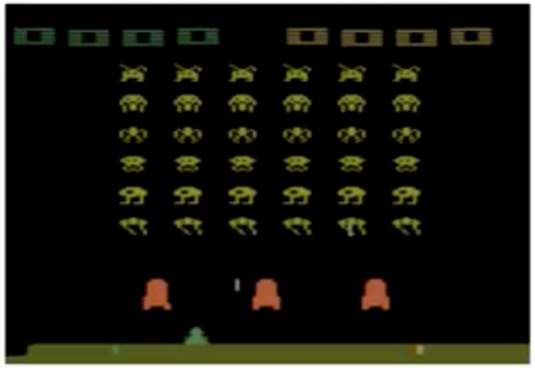
One “Episode”

Example: Playing Video Game

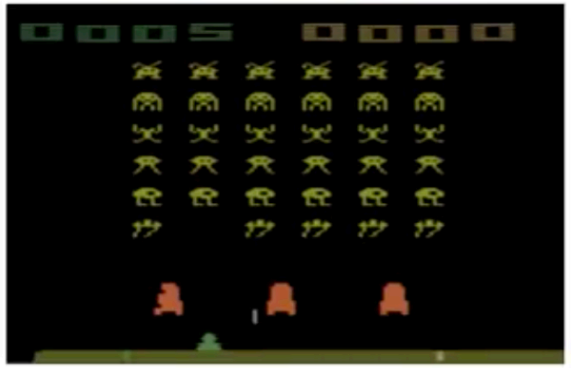
Start with observation s_1



Observation s_2



Observation s_3



After many turns

This is an ***episode***.

Learn to maximize the expected cumulative reward per episode

Difficulties of Reinforcement Learning

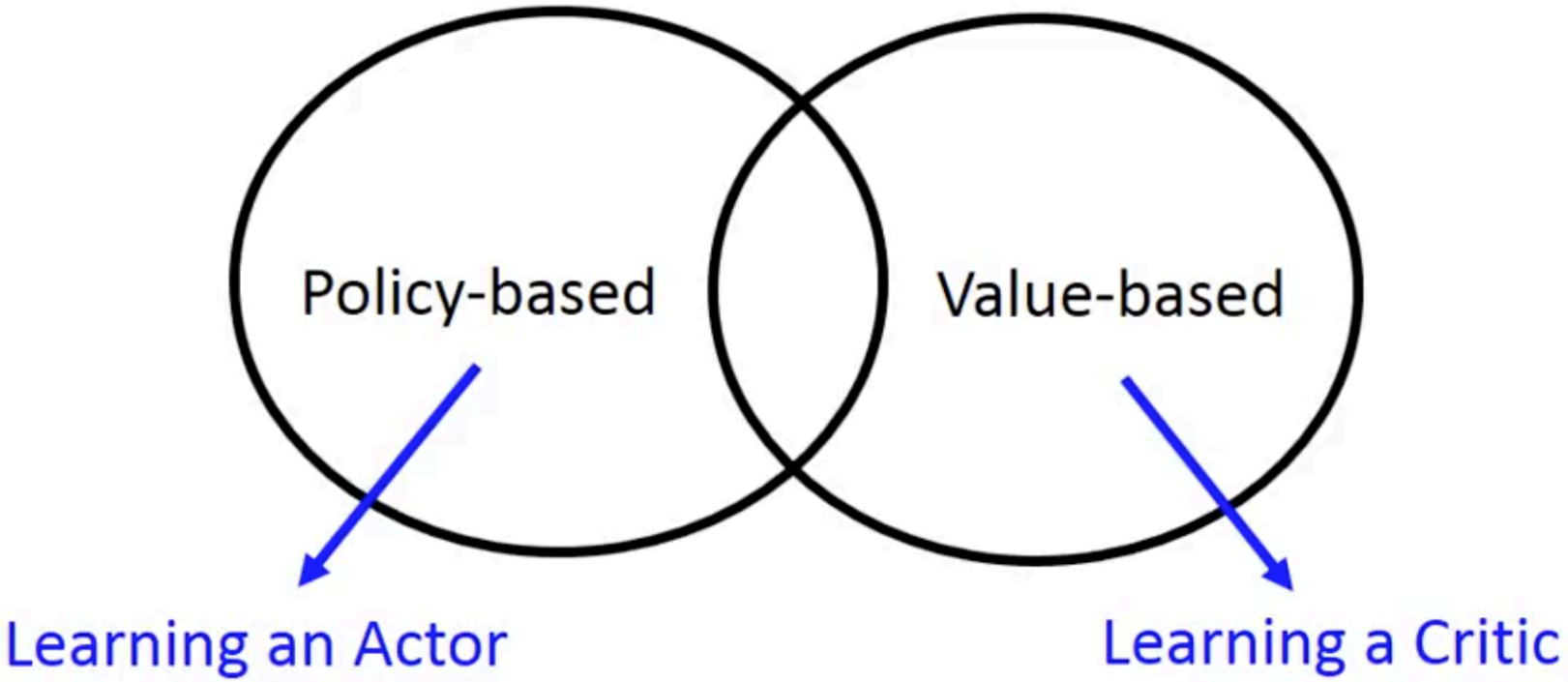
Difficulties of Reinforcement Learning

- Reward delay
  - In space invader, only “fire” obtains reward
    - Although the moving before “fire” is important
  - In Go playing, it may be better to sacrifice immediate reward to gain more long-term reward

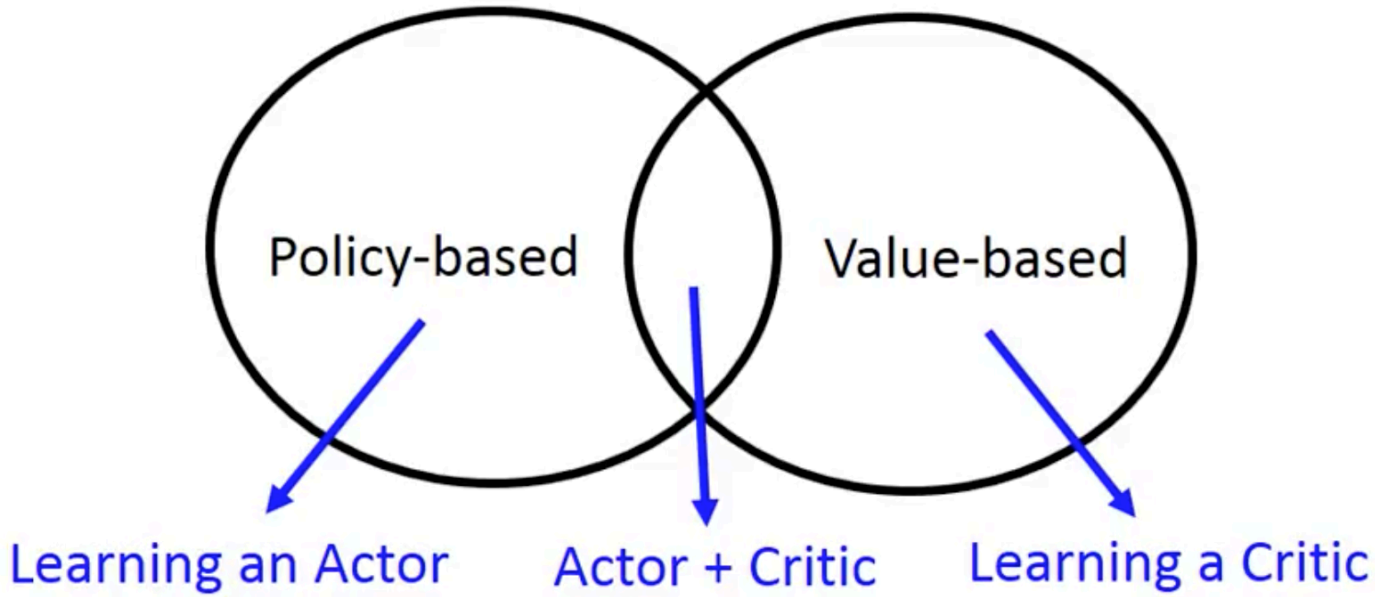
Difficulties of Reinforcement Learning

- Reward delay
  - In space invader, only “fire” obtains reward
    - Although the moving before “fire” is important
  - In Go playing, it may be better to sacrifice immediate reward to gain more long-term reward
- Agent’s actions affect the subsequent data it receives

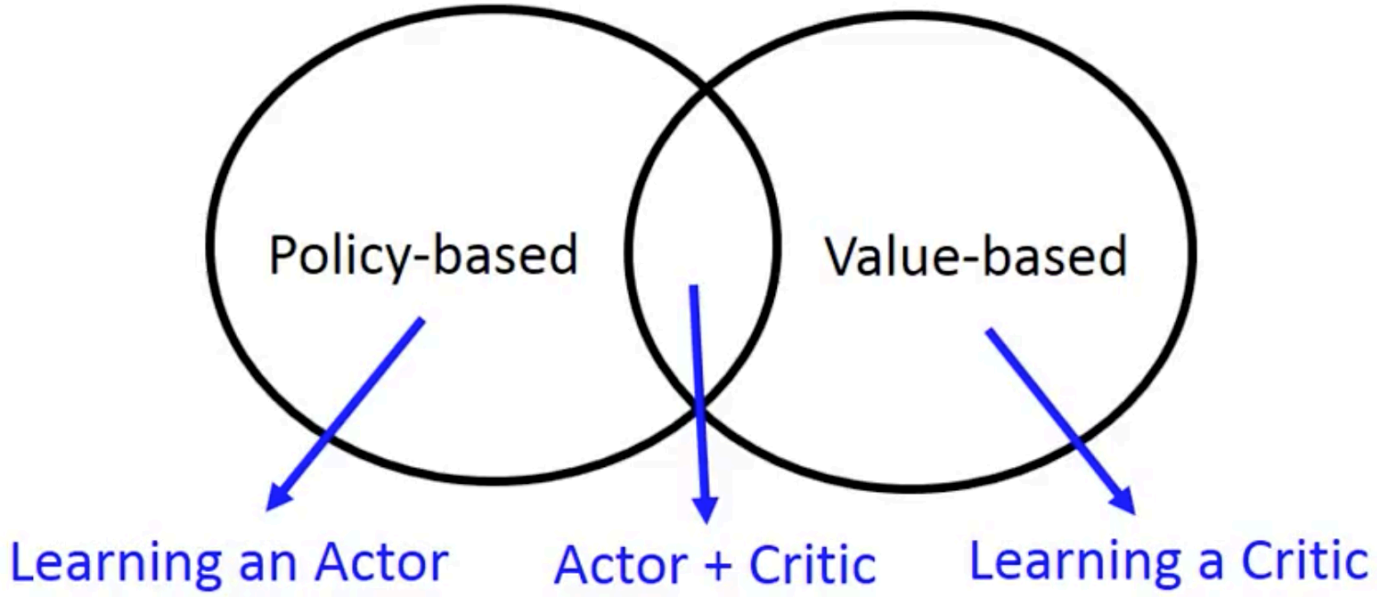
Outline


Policy-based
Value-based
Learning an Actor
Learning a Critic

Outline


Policy-based
Value-based
Learning an Actor
Actor + Critic
Learning a Critic

Outline

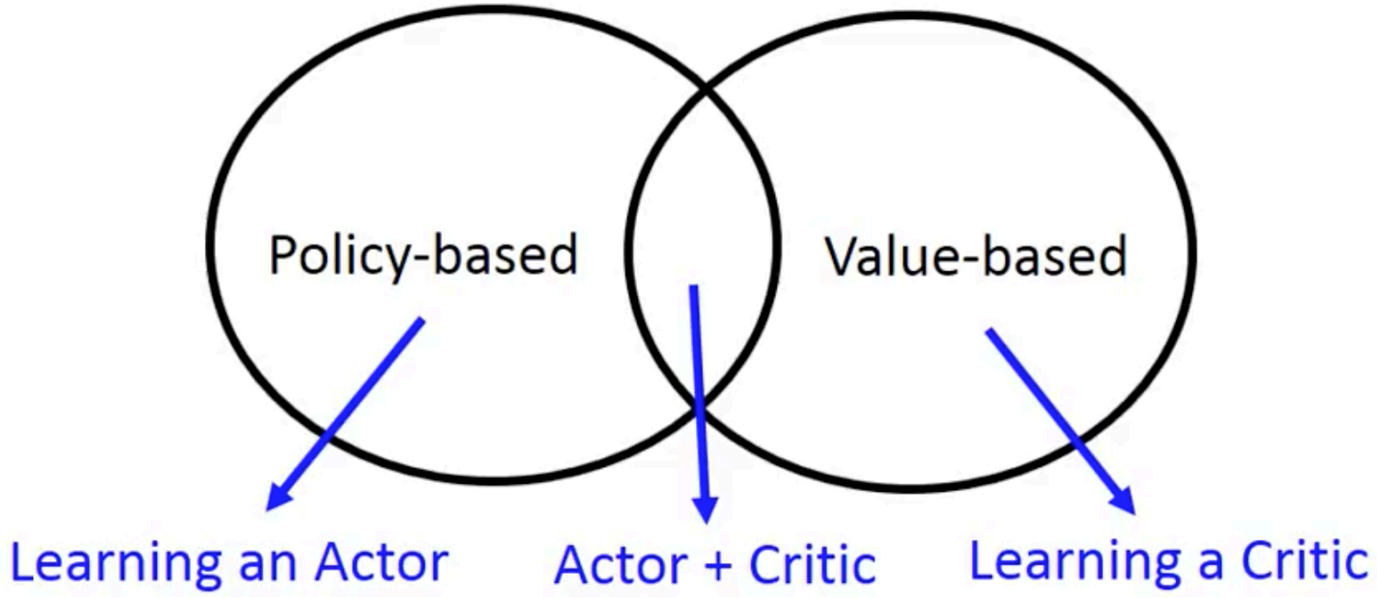


Asynchronous Advantage Actor Critic (A3C)

Volodymyr Mnih et al., "Asynchronous Methods for Deep Reinforcement Learning," ICML, 2016.

Outline

Alpha Go: policy-based + value-based + model-based



Asynchronous Advantage Actor Critic (A3C)

Volodymyr Mnih et al., "Asynchronous Methods for Deep Reinforcement Learning," ICML, 2016.

To learn deep reinforcement learning ……

- Textbook: Reinforcement Learning: An Introduction
  - https://webdocs.cs.ualberta.ca/~sutton/book/the-book.html
- Lectures of David Silver
  - http://www0.cs.ucl.ac.uk/staff/D.Silver/web/Teaching.html (10 lectures, 1:30 each)
  - http://videolectures.net/rldm2015_silver_reinforcement_learning/ (Deep Reinforcement Learning)
- Lectures of John Schulman
  - https://youtu.be/aUrX-rP_ss4

Policy-based Approach

Learning an Actor

Note: Actor means “Agent”