Introduction to Reinforcement Learning

Prof. Junni Zou

Institute of Media, Information and Network
Dept. of Computer Science and Engineering
Shanghai Jiao Tong University
http://min.sjtu.edu.cn

Spring, 2024

Institute of Media, Information, and Network

Outline

Table of Contents

About MIN

Institute of Media, Information, and Network

媒体信息网络研究所
(Media Information Network Institute)

媒体信息网络研究所

- 电子信息与电气工程学院
  电子工程系、
  计算机科学与工程系
- 英文主页:
  http://min.sjtu.edu.cn

主要研究方向:

- 信号处理
- 多媒体通信
- 计算机视觉
- 人工智能
- 生物医学信息

itute of Media,
rmation, and Network

About MIN

Institute of Media, Information, and Network

团队成员

媒体信息网络研究所 （Media Information Network Institute）

- 教授 5 人、副教授 2 人 （ 特聘教授 2 人 ）、助理教授 1 人
  国家杰青 3 人，长江 / 万人 1 人，国家优青 2 人，青年长江 1 人，青年拔尖 1 人
- 在读博士生 30 余人（毕业 19 人）；在读硕士生 20 余人（毕业 100 余人）

李成林 教授
国家优青

邹君妮 教授
国家杰青
国家优青

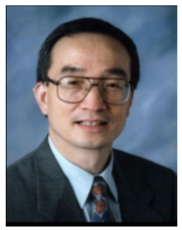
熊红凯 教授
国家杰青 / 长江学者
万人领军

林巍峣 教授
国家杰青
青年长江

戴文睿 副教授
青年拔尖

郑紫阳
助理教授

何广强 教授



Prof. Yuan F. Zheng
美国OSU

Prof. Tsuhan Chen
美国CMU、Cornell

Prof. Pascal Frossard
瑞士 EPFL

Prof. David Taubman
澳大利亚 UNSW

Institute of Media, Information, and Network

About MIN

Institute of Media, Information, and Network

About MIN

Institute of Media, Information, and Network

About MIN

Institute of Media, Information, and Network

About Class

- Instructor: Junni Zou (zoujunni@sjtu.edu.cn), SEIEE Building 3-437
- Course website: http://www.cs.sjtu.edu.cn/~zou-jn/
- Reference book: Richard S. Sutton and Andrew G. Barto, Reinforcement Learning: An Introduction, Second Edition
- TA: Yiheng Jiang (yiheng_j@sjtu.edu.cn)
- Grading Policy:
  Homework: 50%, Final Project: 50%

We use the slides of Prof. David Silver and Prof. Hung-yi Lee as reference.
https://www.davidsilver.uk/teaching/
http://speech.ee.ntu.edu.tw/~tlkagk/courses_ML20.html

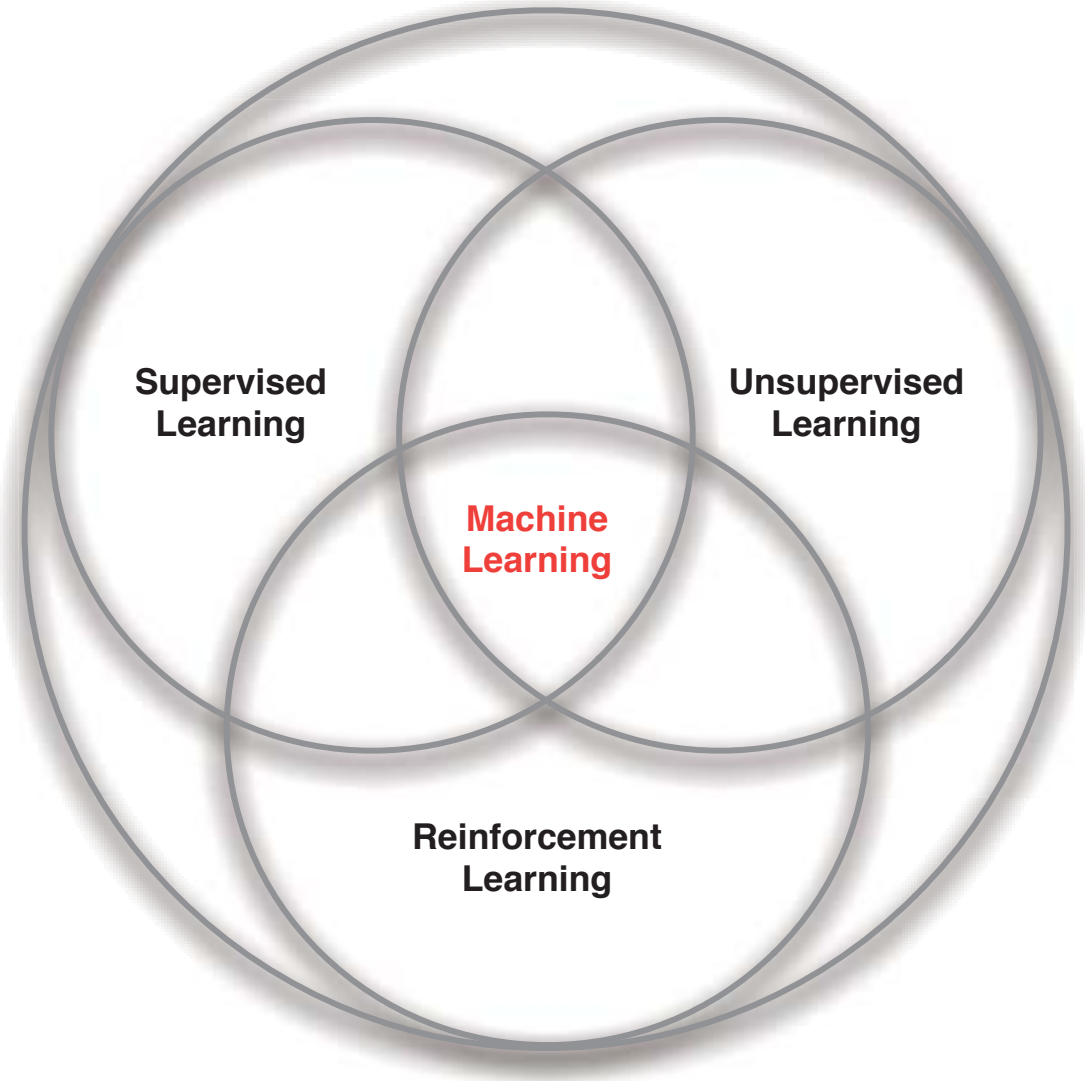
Institute of Media, Information, and Network

Online Courses of Reinforcement Learning

- Reinforcement Learning, David Silver, UCL
- Reinforcement Learning, Emma Brunskill, Stanford
- Deep Reinforcement Learning, Sergey Levine, UC Berkeley
- Deep Reinforcement Learning and Control, Katerina Fragkiadaki, CMU

Table of Contents

Development of Reinforcement Learning



Institute of Media, Information, and Network

Development of Reinforcement Learning

"The brain evolved, not to think or feel, but to control movement."

Daniel Wolpert

https://www.ted.com/talks/daniel_wolpert_the_real_reason_for_brains?language=en

Institute of Media, Information, and Network

Development of Reinforcement Learning

"The brain evolved, not to think or feel, but to control movement."

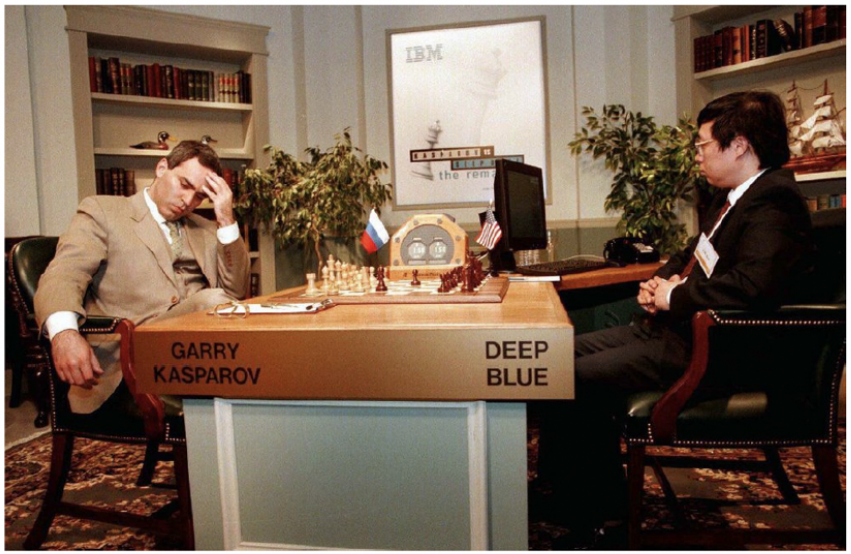
Daniel Wolpert

Sea squirts digest their own brain when they decide not to move anymore

Institute of Media, Information, and Network

Development of Reinforcement Learning

Deep Blue



深蓝重1270公斤，有32个处理器

1997-05-11，深蓝（2胜1负3平）
击败世界冠军卡斯帕罗夫

Q1: Is this a machine learning achievement?

Q2: What is machine learning / artificial intelligence?

A2: The discipline that develops agents that learn and improve with experience (Tom Mitchell)

A1: No, it is not. Brute-force manual development of a board evaluation function

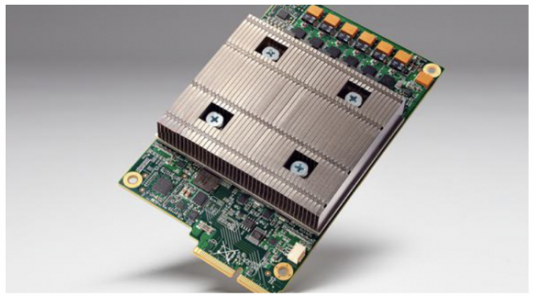
twork

Development of Reinforcement Learning

AlphaGo

2016-03，AlphaGo 4:1击败世界围棋冠军李世石

- Monte Carlo Tree Search with neural nets
- expert demonstrations
- self play



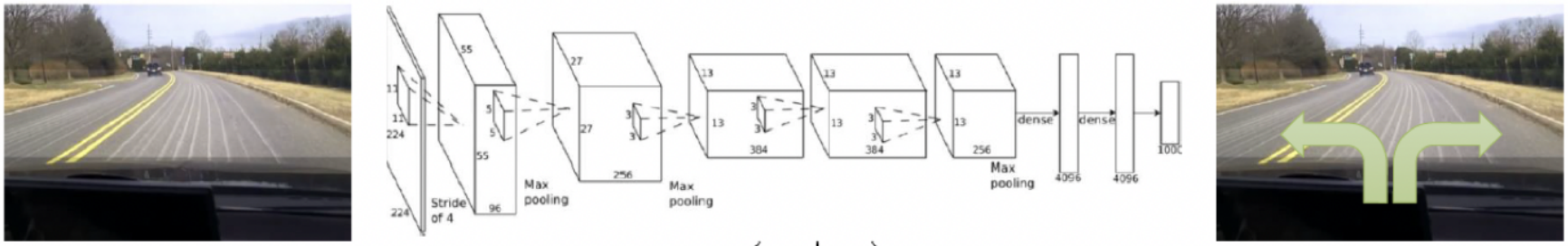
Tensor Processing Unit from Google

- AlphaGo使用4个TPU处理器
- 升级版的AlphaGo仅使用1个TPU
- AlphaGo的计算能力是深蓝的30万倍

Institute of Media, Information, and Network

Development of Reinforcement Learning

Policy: a mapping from observations to actions



$\mathbf{o}_t$ $\pi_\theta(\mathbf{u}_t|\mathbf{o}_t)$ $\mathbf{u}_t$

Why should we study this **now** ?

- Advances in computational capability
- Advances in deep learning
- Advances in reinforcement learning

twork

Development of Reinforcement Learning

- 强化学习算法理论的形成：上个世纪七八十年代
- 2013年12月，DeepMind 展示RL算法在Atari游戏中打败人类专业玩家，其成果在2015年发布于《自然》
- 2014年，谷歌收购DeepMind 团队
- 2016年3月，AlphaGo以4:1击败世界围棋冠军李世石（Alpha Go Lee）
- 2017年5月，AlphaGo以3:0击败世界围棋冠军柯洁（AlphaGo Master）
- 2017年10 月19 日凌晨，DeepMind 在《Nature》发布论文，最新版本AlphaGo Zero，训练3 天以100:0击败了AlphaGo Lee，训练40 天击败AlphaGo Master

Development of Reinforcement Learning

2017 年，DeepMind 宣布开始研究能进行即时战略游戏星际争霸 2 的人工智能——AlphaStar。

AlphaStar与《星际争霸2》比赛直播

2019-01-25，AlphaStar与《星际争霸2》比赛直播

AlphaStar在实况录像的10场均获胜，现场比赛不敌人类，总成绩10：1

k

Development of Reinforcement Learning

2019-10-30，Nature 封面论文

Article

Grandmaster level in StarCraft II using multi-agent reinforcement learning

https://doi.org/10.1038/s41586-019-1724-z

Received: 30 August 2019

Accepted: 10 October 2019

Published online: 30 October 2019

Oriol Vinyals[1,3]*, Igor Babuschkin[1,3], Wojciech M. Czarnecki[1,3], Michaël Mathieu[1,3], Andrew Dudzik[1,3], Junyoung Chung[1,3], David H. Choi[1,3], Richard Powell[1,3], Timo Ewalds[1,3], Petko Georgiev[1,3], Junhyuk Oh[1,3], Dan Horgan[1,3], Manuel Kroiss[1,3], Ivo Danihelka[1,3], Aja Huang[1,3], Laurent Sifre[1,3], Trevor Cai[1,3], John P. Agapiou[1,3], Max Jaderberg[1], Alexander S. Vezhnevets[1], Rémi Leblond[1], Tobias Pohlen[1], Valentin Dalibard[1], David Budden[1], Yury Sulsky[1], James Molloy[1], Tom L. Paine[1], Caglar Gulcehre[1], Ziyu Wang[1], Tobias Pfaff[1], Yuhuai Wu[1], Roman Ring[1], Dani Yogatama[1], Dario Wünsch[2], Katrina McKinney[1], Oliver Smith[1], Tom Schaul[1], Timothy Lillicrap[1], Koray Kavukcuoglu[1], Demis Hassabis[1], Chris Apps[1,3] & David Silver[1,3]*

Institute of Media, Information, and Network

Development of Reinforcement Learning

戴密斯·哈萨比斯（Demis Hassabis），人工智能企业家，DeepMind Technologies公司创始人，人称“阿尔法围棋之父”。4岁开始下国际象棋，8岁自学编程，13岁获得国际象棋大师称号。17岁进入剑桥大学攻读计算机科学专业。在大学里，他开始学习围棋。2005年进入伦敦大学学院攻读神经科学博士，选择大脑中的海马体作为研究对象。两年后，他证明了5位因为海马体受伤而患上健忘症的病人，在畅想未来时也会面临障碍，并凭这项研究入选《科学》杂志的“年度突破奖”。2011年创办DeepMind Technologies公司，以“解决智能”为公司的终极目标。

阿尔法围棋设计团队部分成员

大卫·席尔瓦（David Silver），剑桥大学计算机科学学士、硕士，加拿大阿尔伯塔大学计算机科学博士，伦敦大学学院讲师，Google DeepMind研究员，阿尔法围棋主要设计者之一。

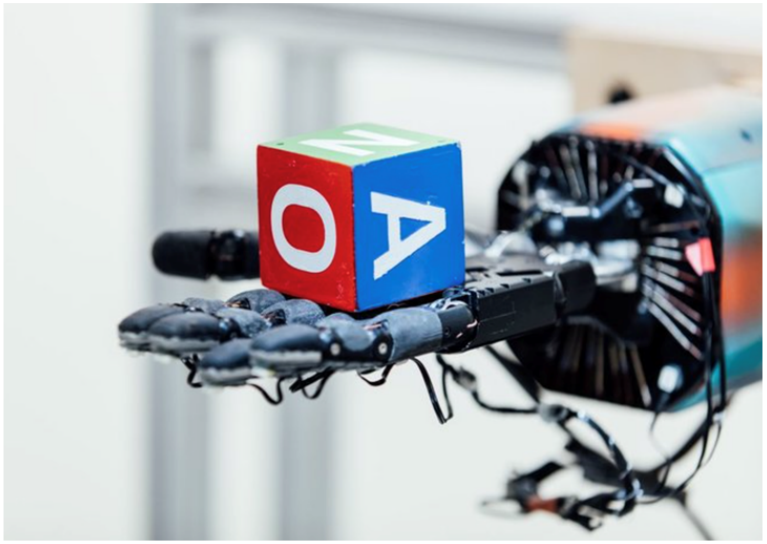
k

Development of Reinforcement Learning

OpenAI

OpenAI，由诸多硅谷大亨联合建立的人工智能非营利组织。

2015年马斯克与其他硅谷科技大亨进行连续对话后，决定共同创建OpenAI，希望能够预防人工智能的灾难性影响，推动人工智能发挥积极作用。特斯拉电动汽车公司与美国太空技术探索公司SpaceX创始人马斯克、Y Combinator总裁阿尔特曼、天使投资人彼得·泰尔（Peter Thiel）以及其他硅谷巨头2014年12月承诺向OpenAI注资10亿美元。



nature

Explore Content | Journal Information | Publish With Us | Subscribe

nature > articles > article

Article | Published: 24 February 2021

First return, then explore

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley & Jeff Clune

Nature **590**, 580–586(2021) | Cite this article

ledia,
and Network

Development of Reinforcement Learning

Reinforcement Learning in UC, Berkeley

Michael I. Jordan
UC, Berkeley

Andrew Ng
Stanford University

Pieter Abbeel
UC, Berkeley

Sergey Levine
UC, Berkeley

Sergey Levine

Associate Professor

Research Areas	Research Centers	Teaching Schedule
Artificial Intelligence (AI)	Berkeley Artificial Intelligence Research Lab (BAIR)	**Fall 2021**
Control, Intelligent Systems, and Robotics (CIR)	CITRIS People and Robots (CPAR)	CS 285. Deep Reinforcement Learning, Decision Making, and Control, MoWe 5:00PM - 6:29PM, Internet/Online
	Berkeley Deep Drive (BDD)	

te of Media,
nation, and Network

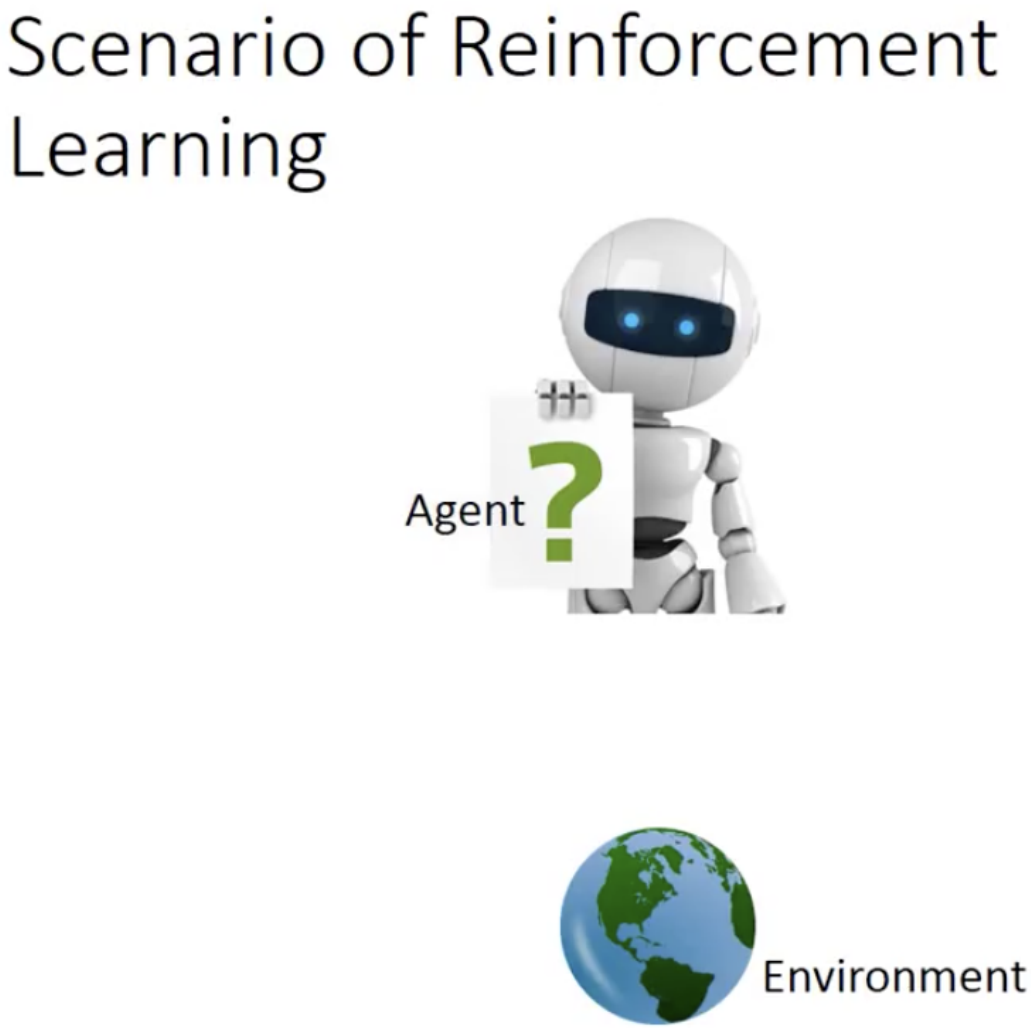
Development of Reinforcement Learning

Deep Reinforcement Learning

Deep Reinforcement Learning: AI = RL + DL

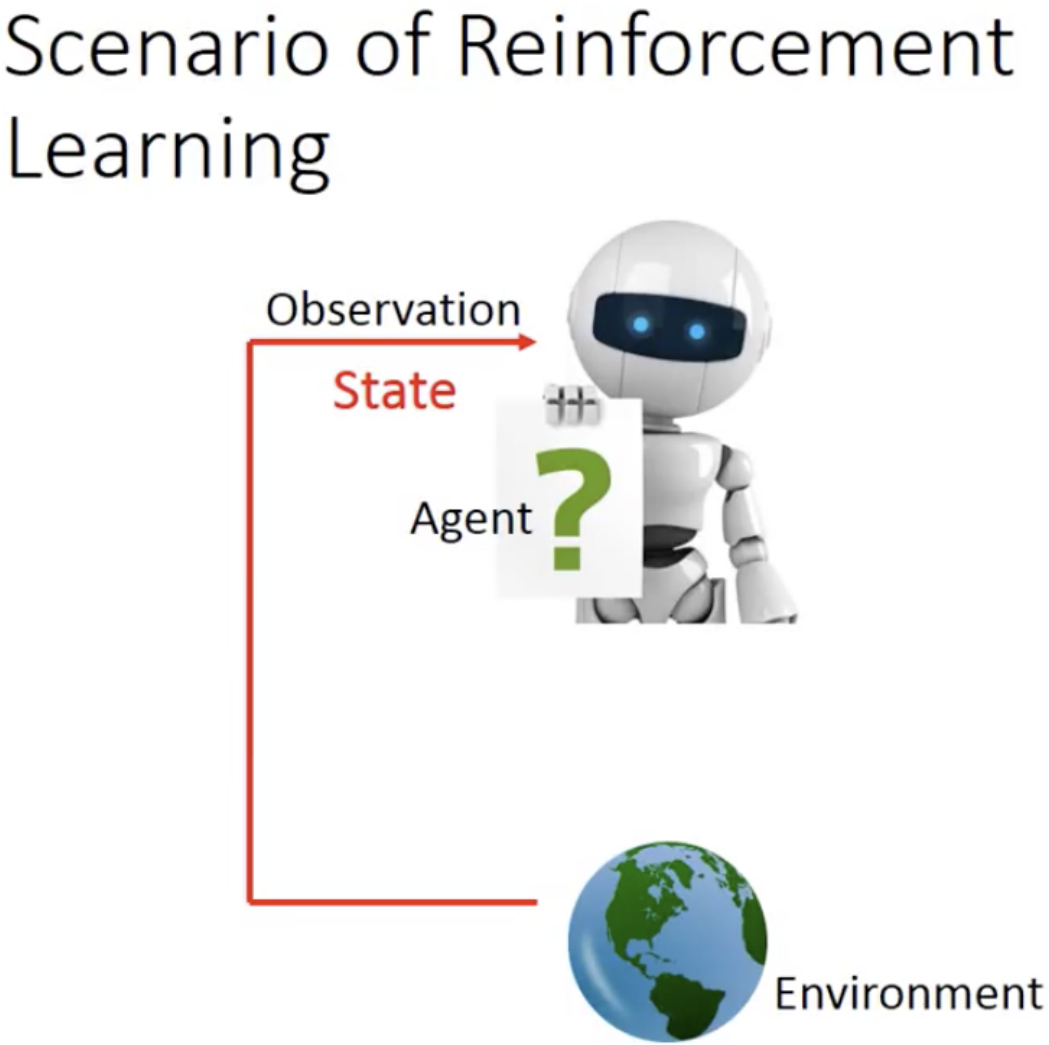
of Media,
on, and Network

Scenario of Reinforcement Learning



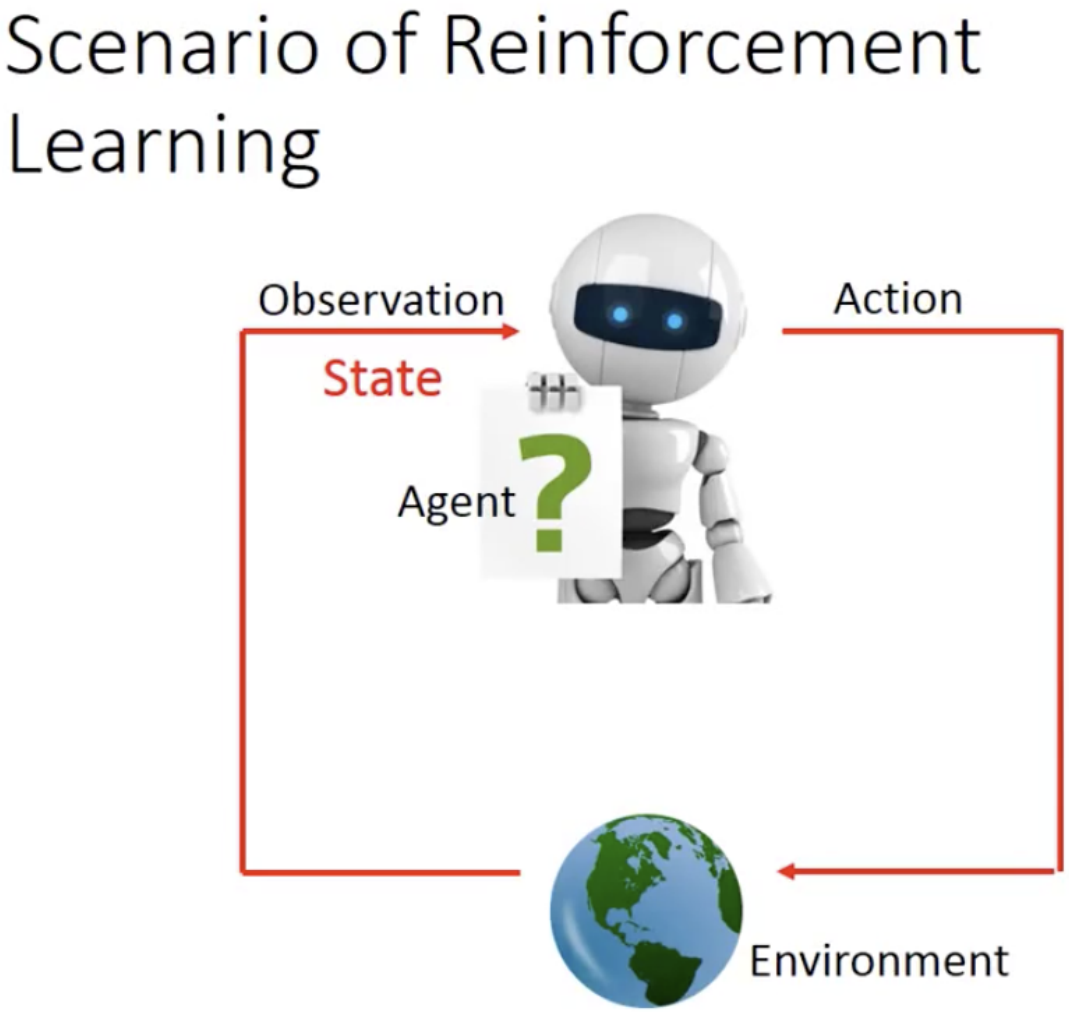
Institute of Media, Information, and Network

Scenario of Reinforcement Learning



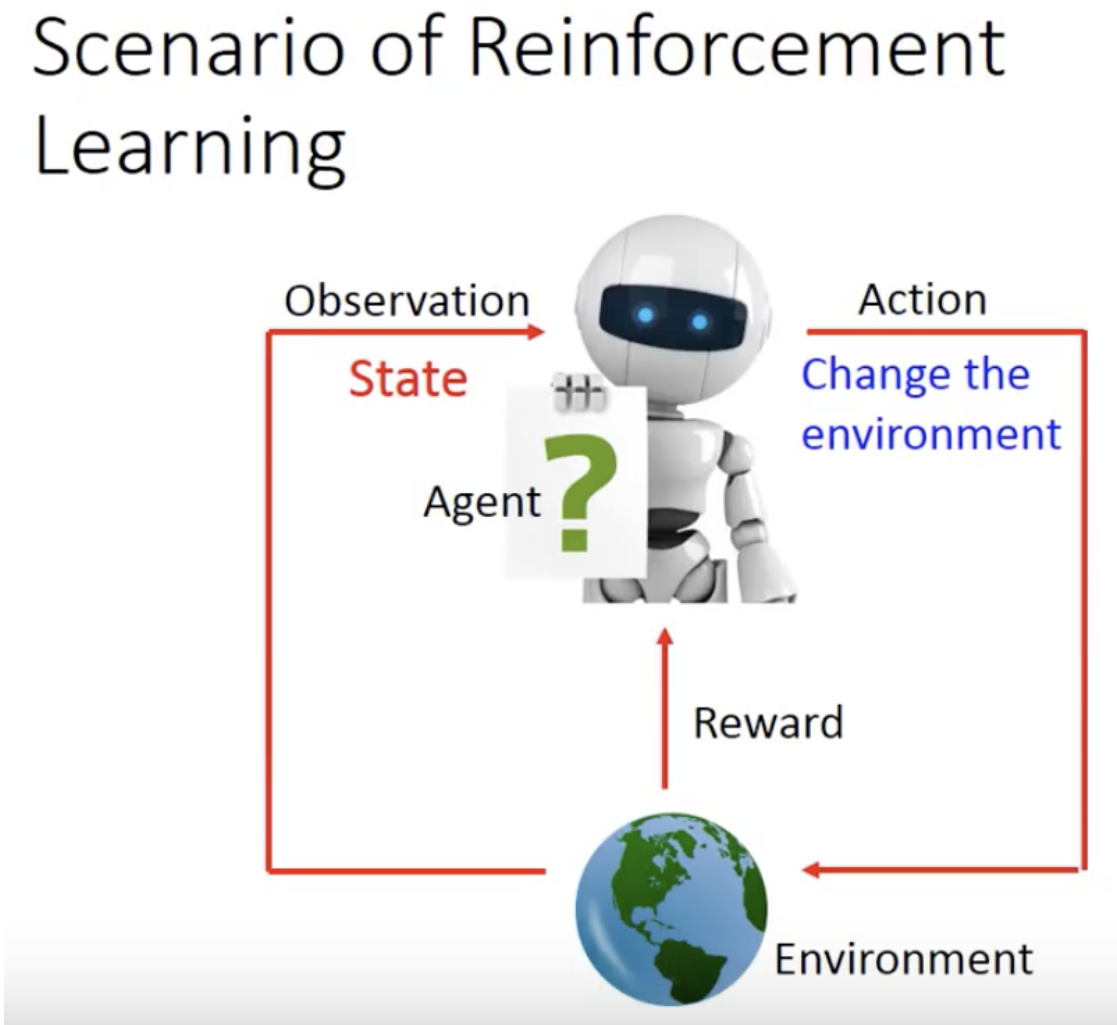
Institute of Media, Information, and Network

Scenario of Reinforcement Learning



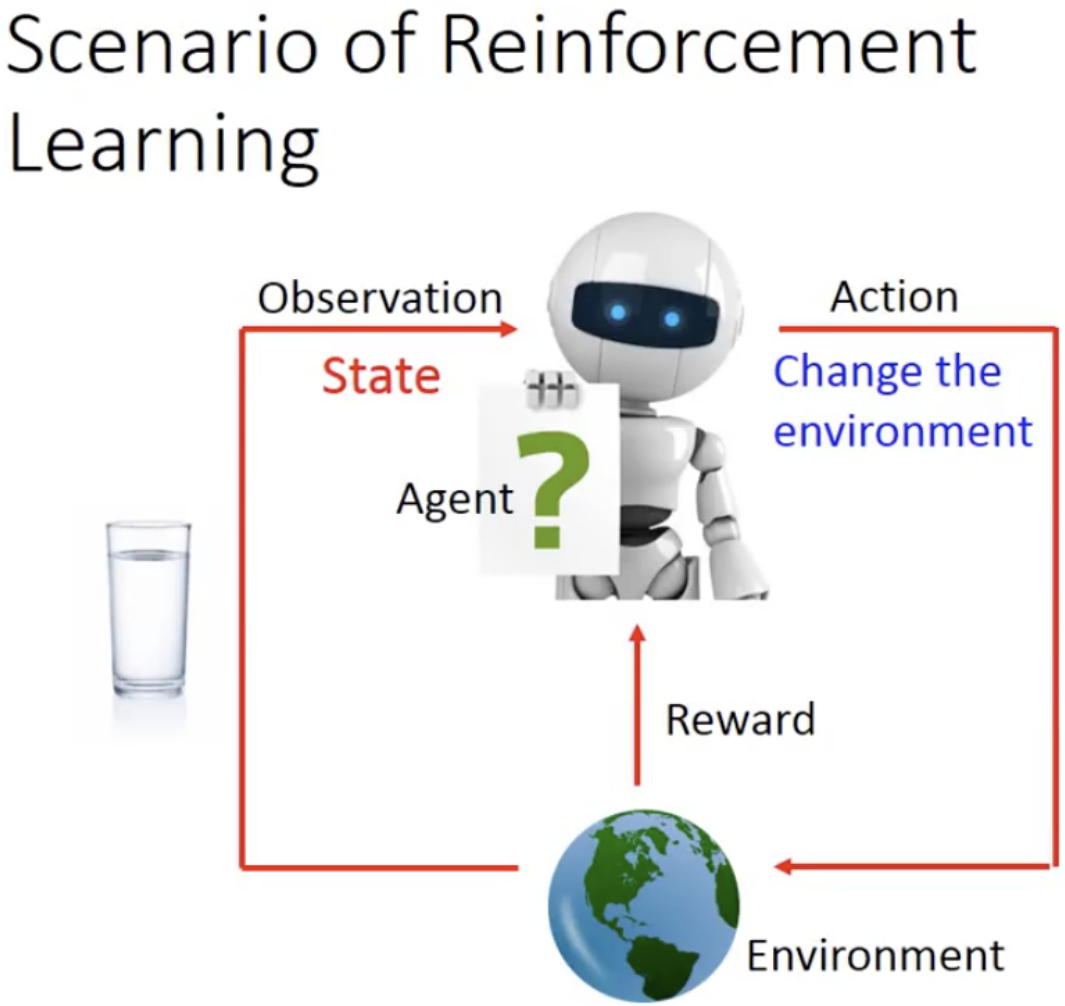
Institute of Media, Information, and Network

Scenario of Reinforcement Learning



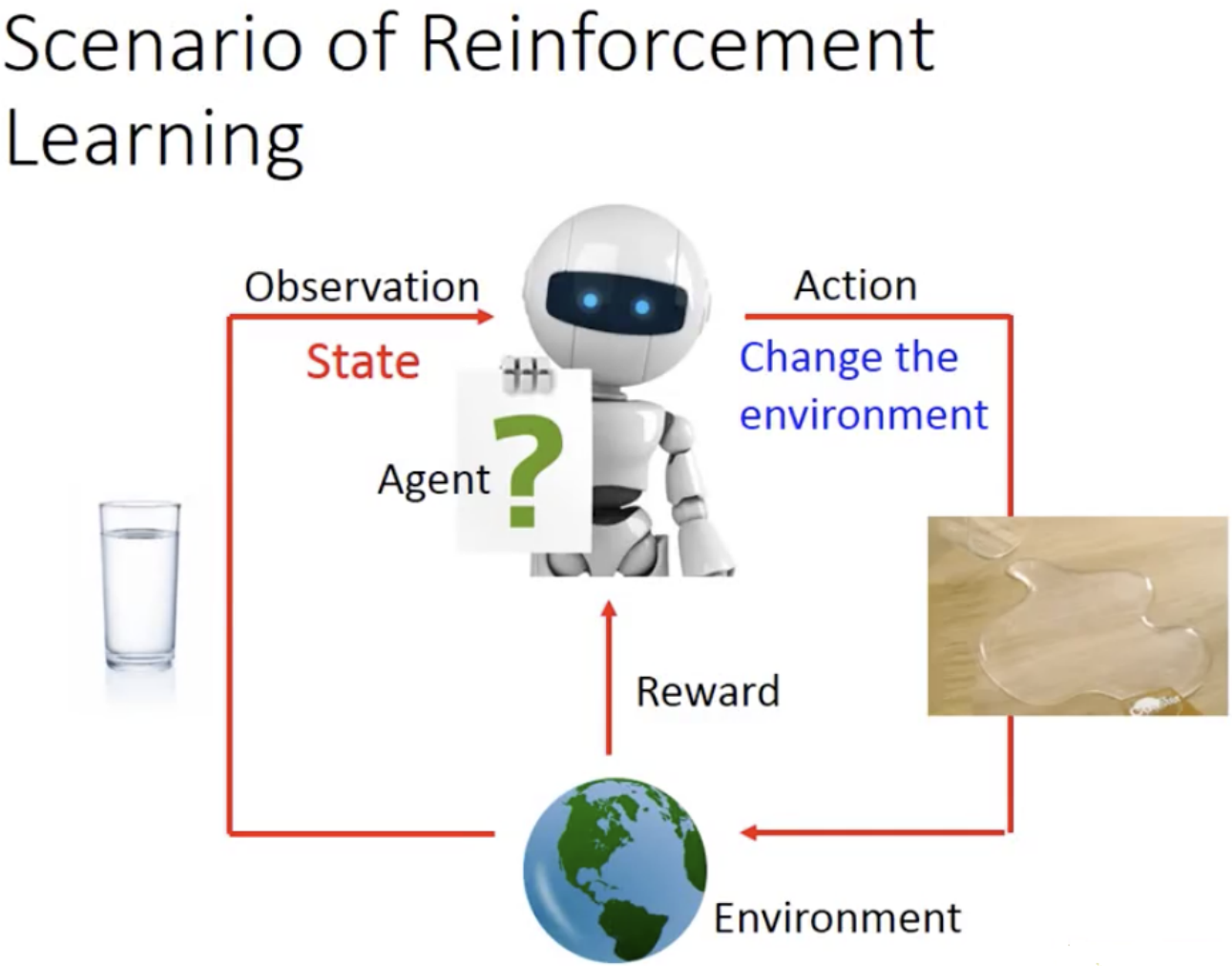
Institute of Media, Information, and Network

Scenario of Reinforcement Learning



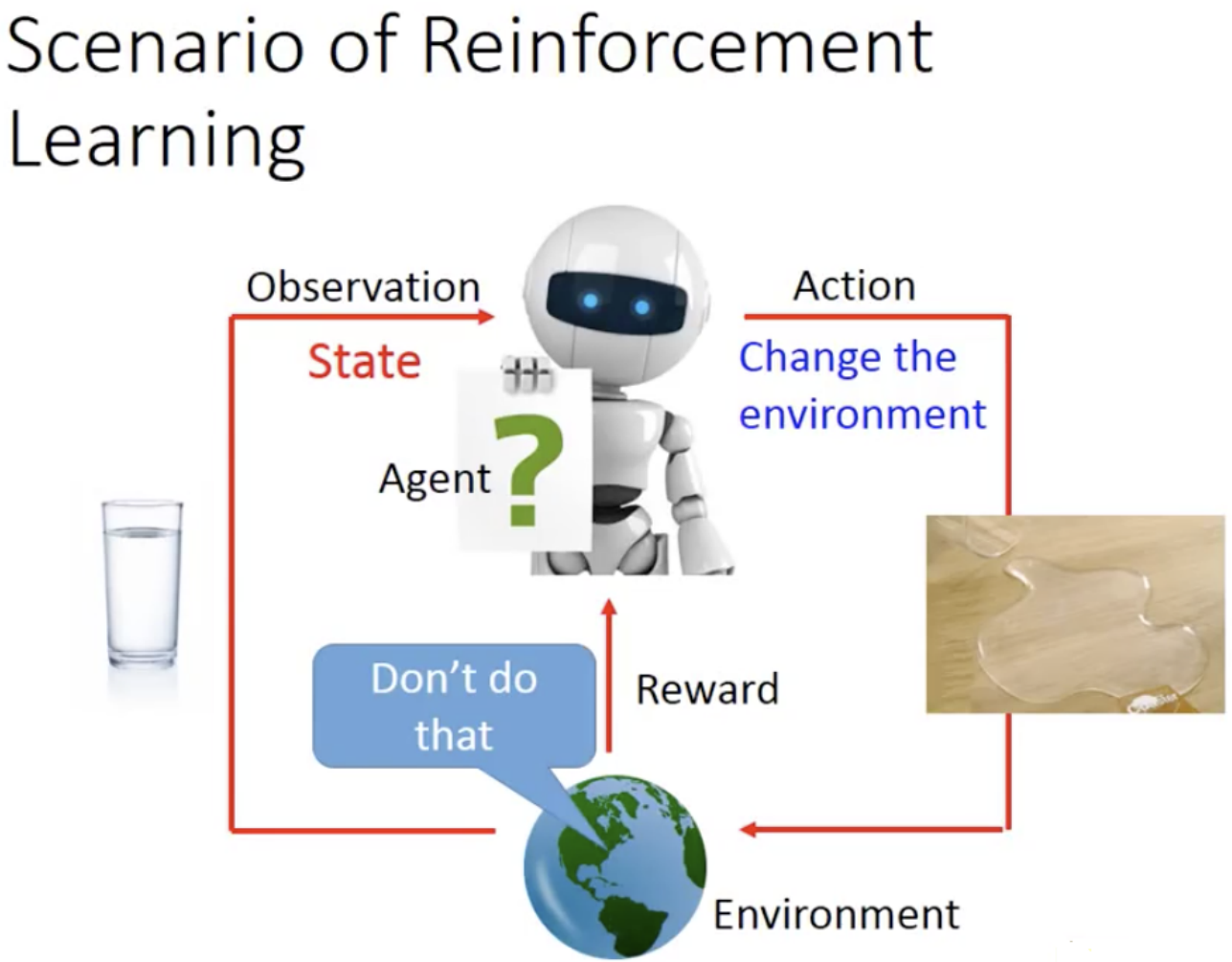
Institute of Media, Information, and Network

Scenario of Reinforcement Learning



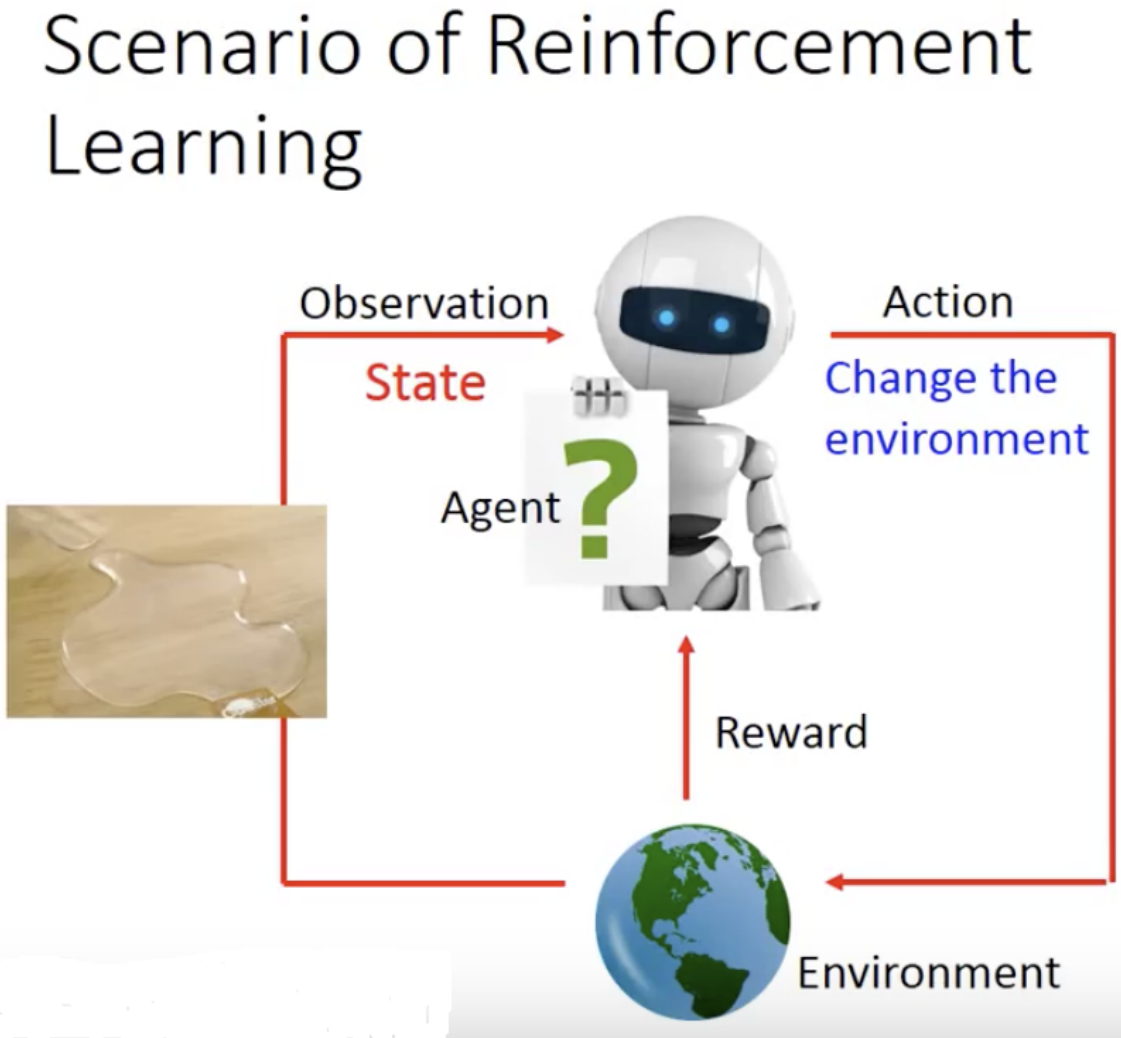
Institute of Media, Information, and Network

Scenario of Reinforcement Learning



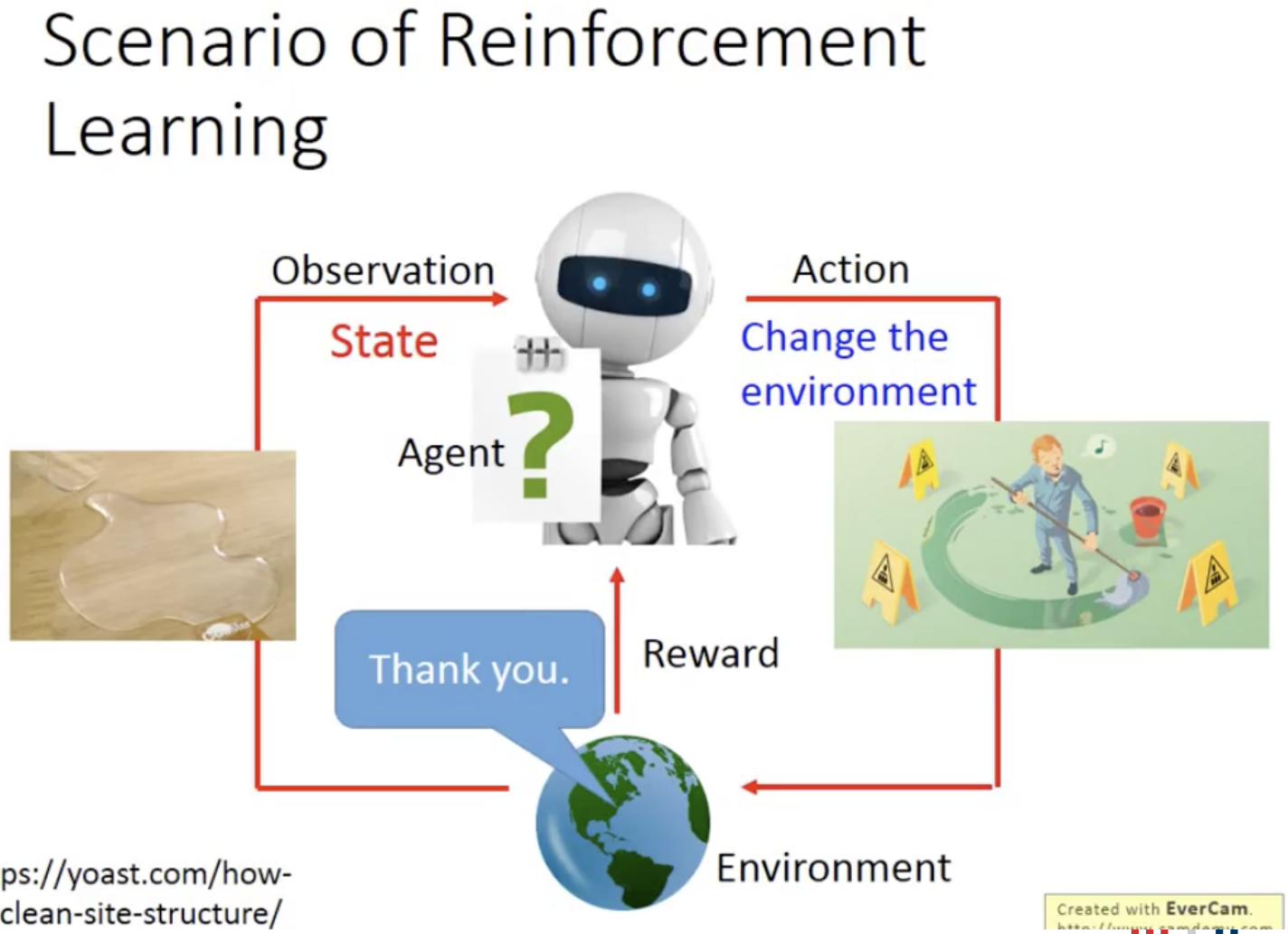
Institute of Media, Information, and Network

Scenario of Reinforcement Learning



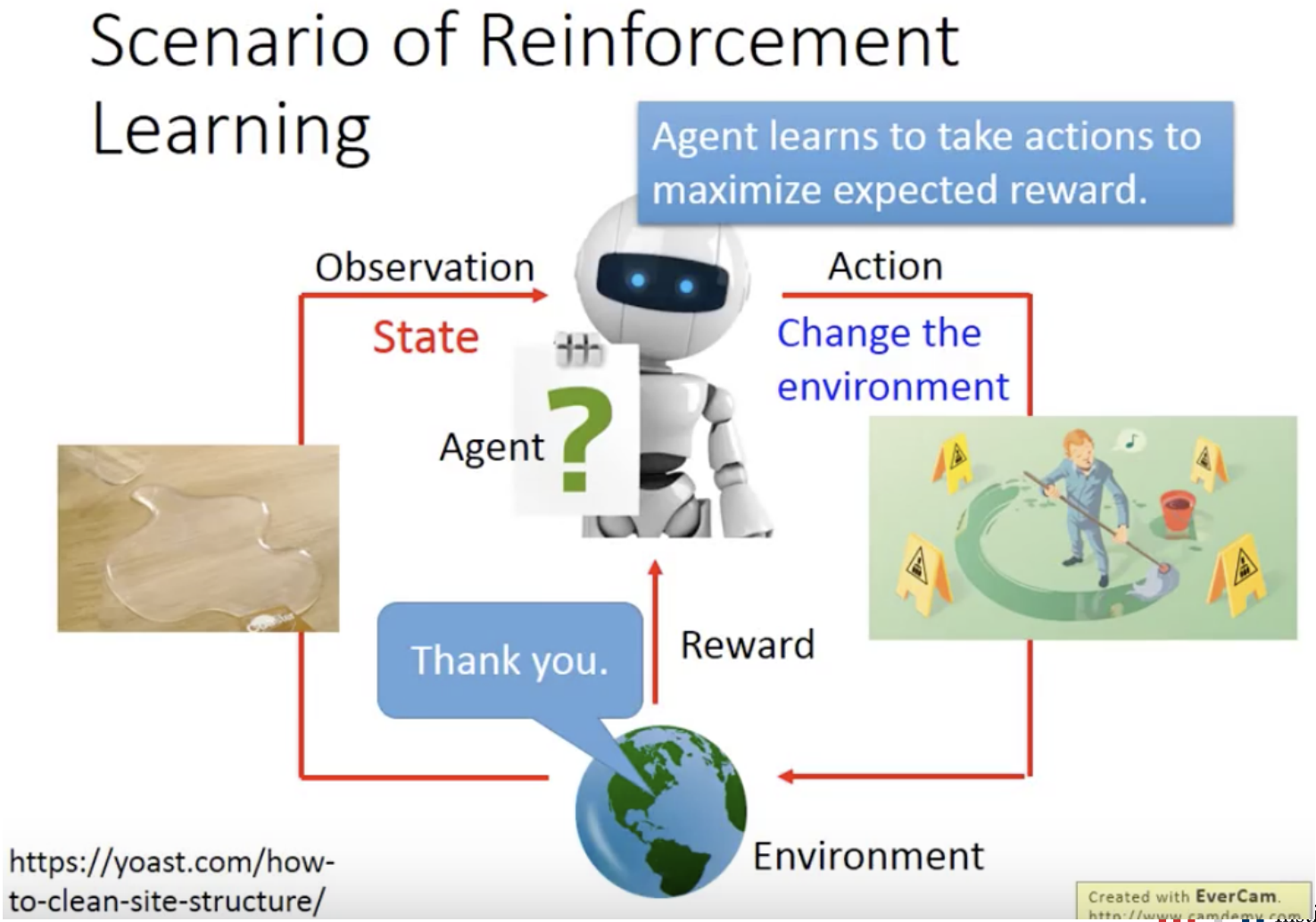
Institute of Media, Information, and Network

Scenario of Reinforcement Learning

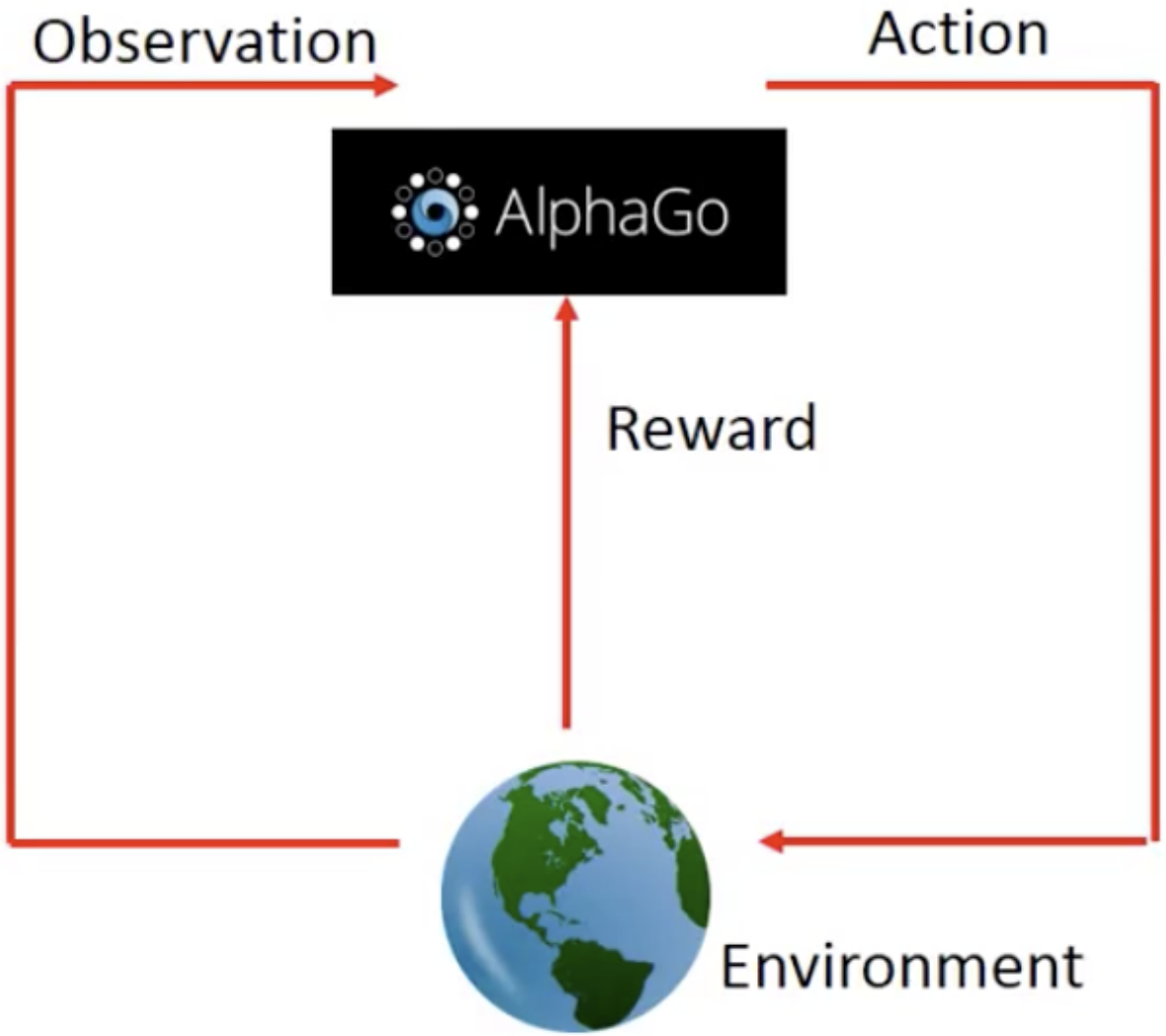


Information, and Network

Scenario of Reinforcement Learning

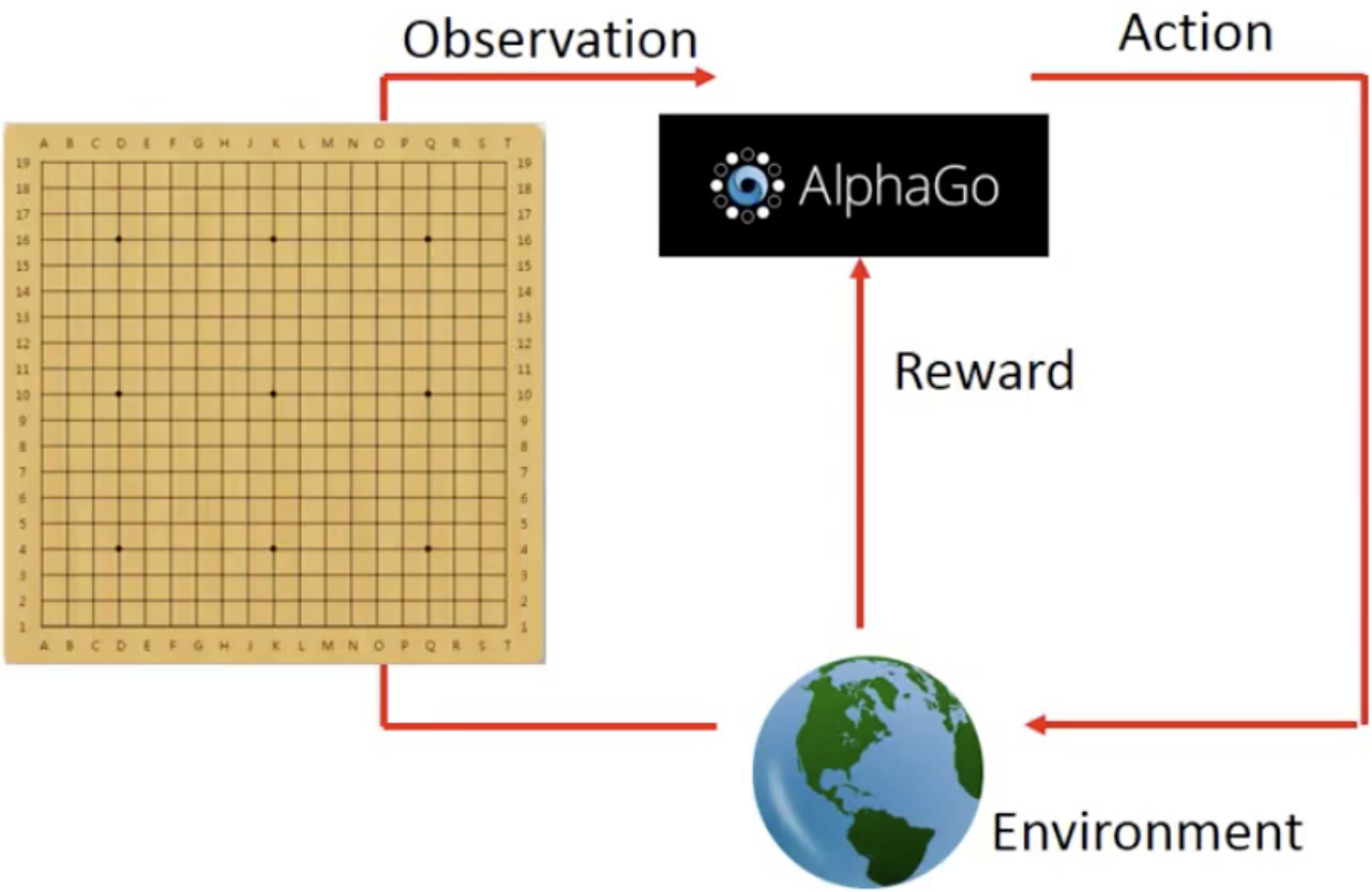


Scenario of Reinforcement Learning

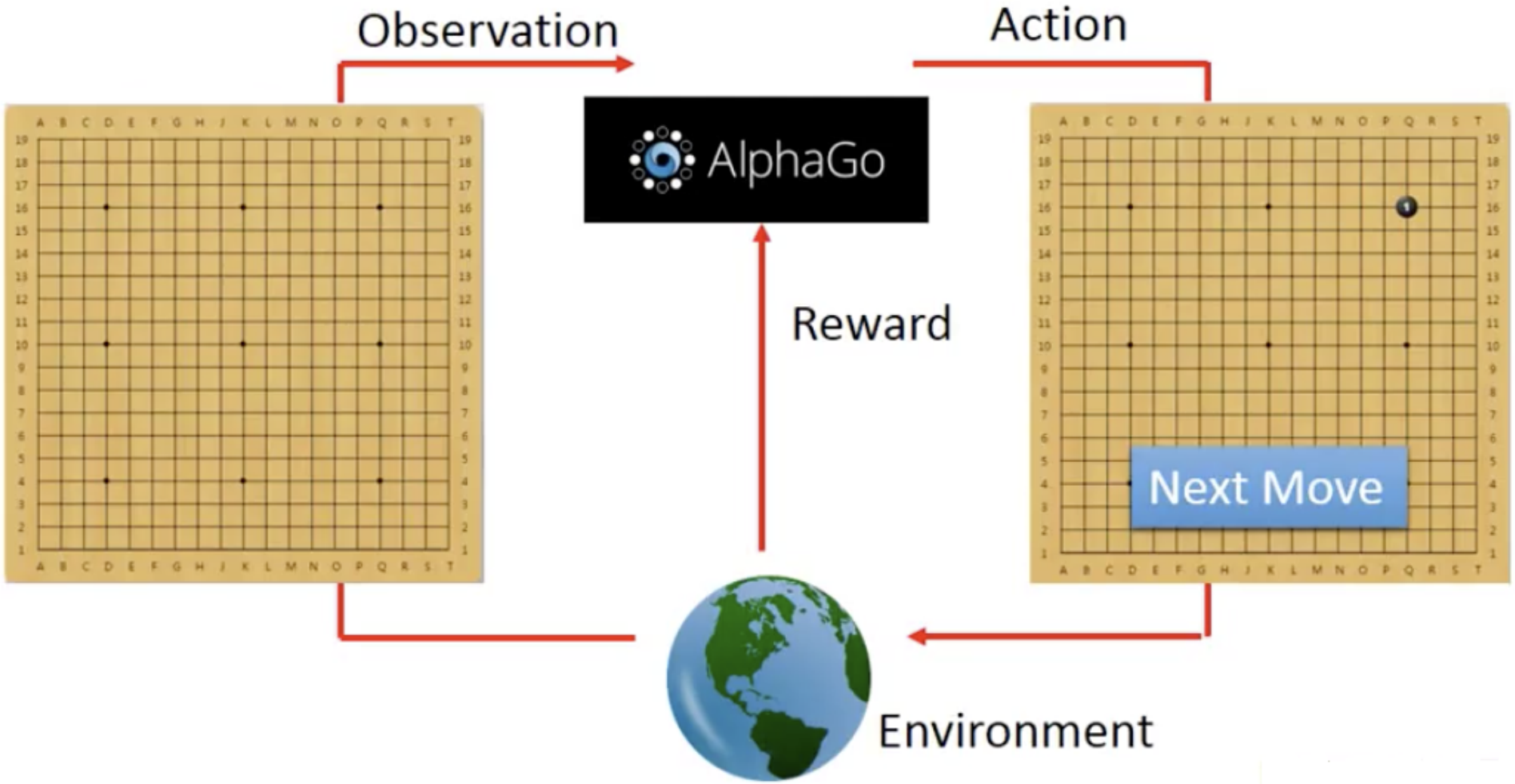


Institute of Media, Information, and Network

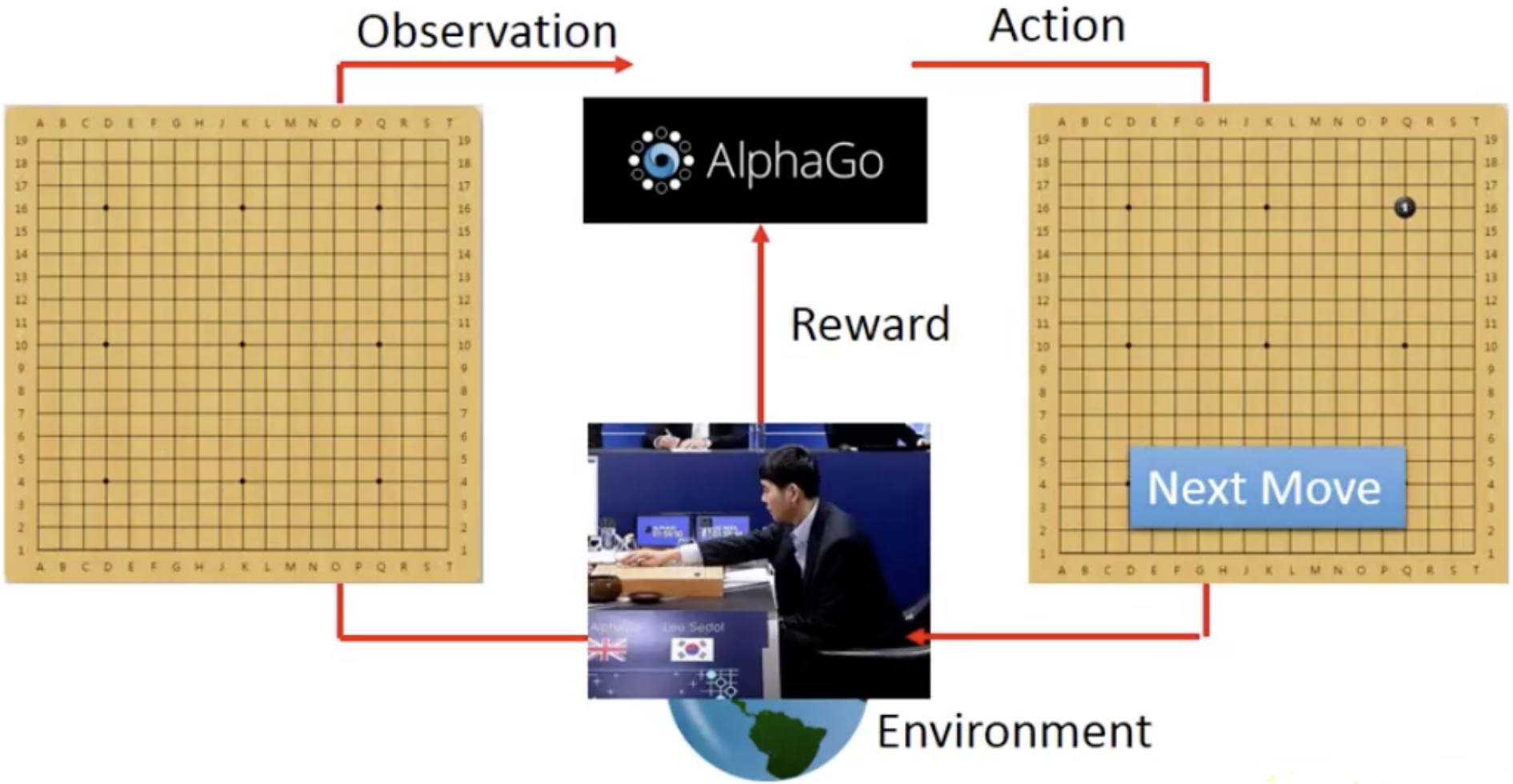
Scenario of Reinforcement Learning



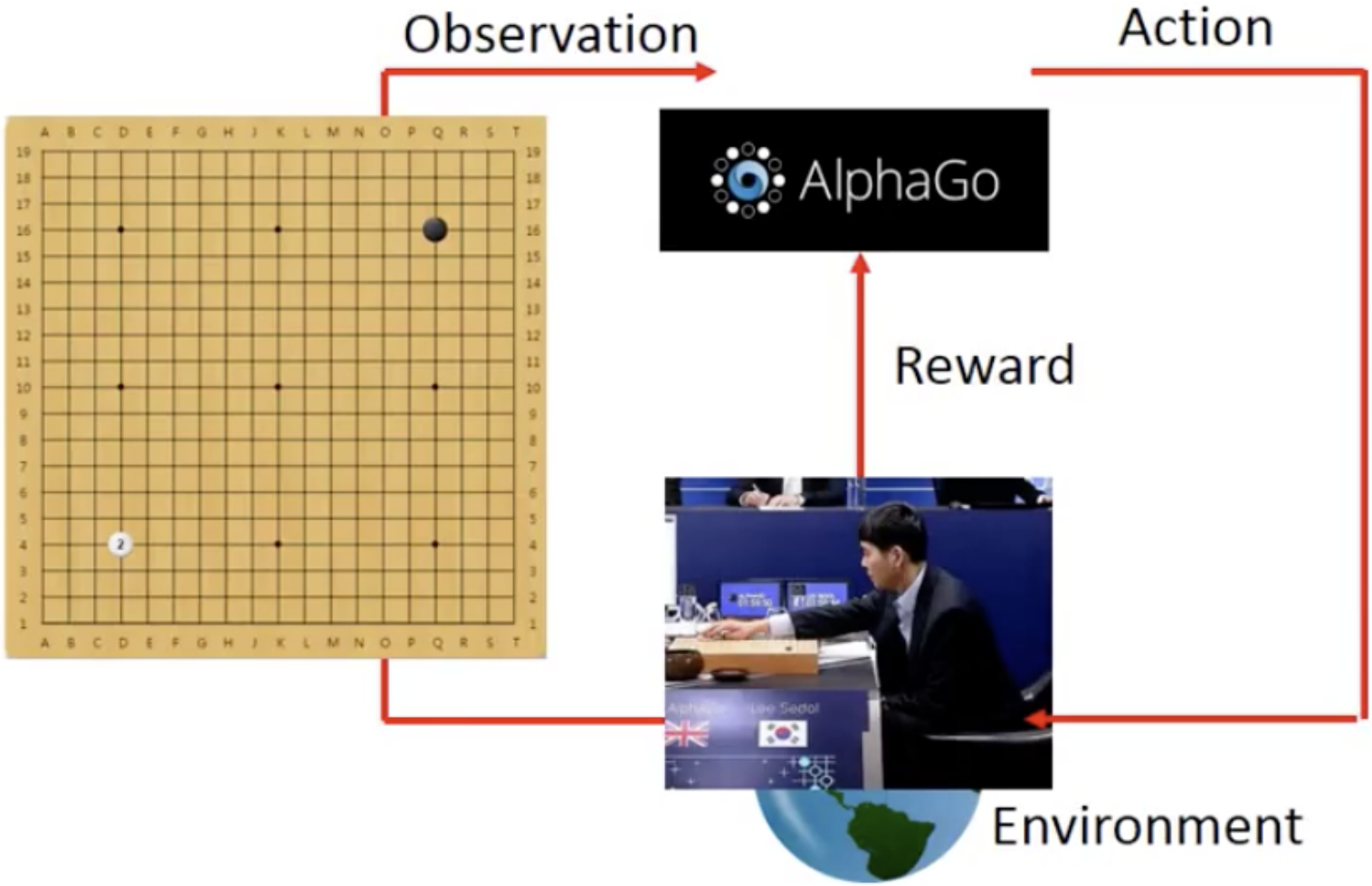
Scenario of Reinforcement Learning



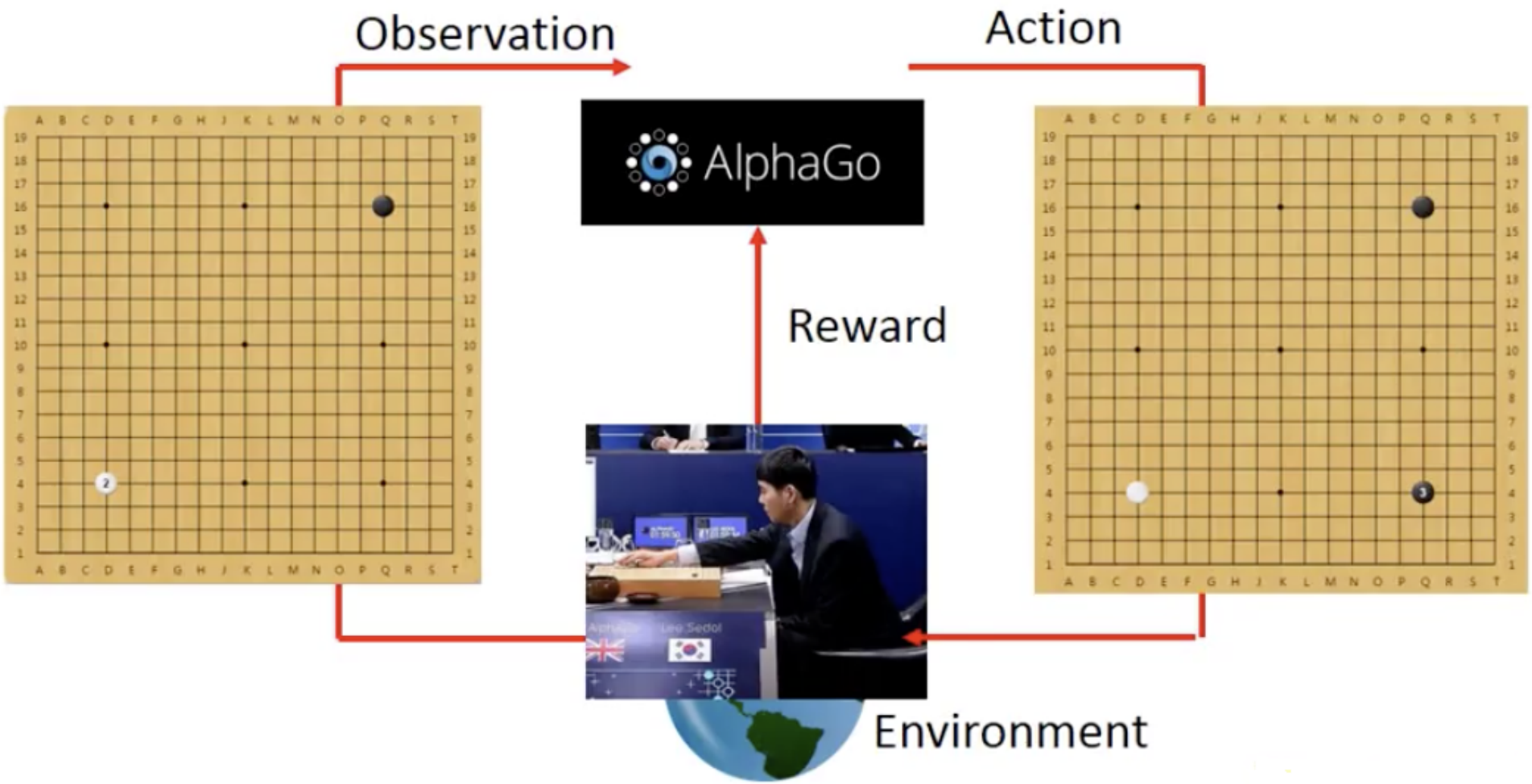
Scenario of Reinforcement Learning



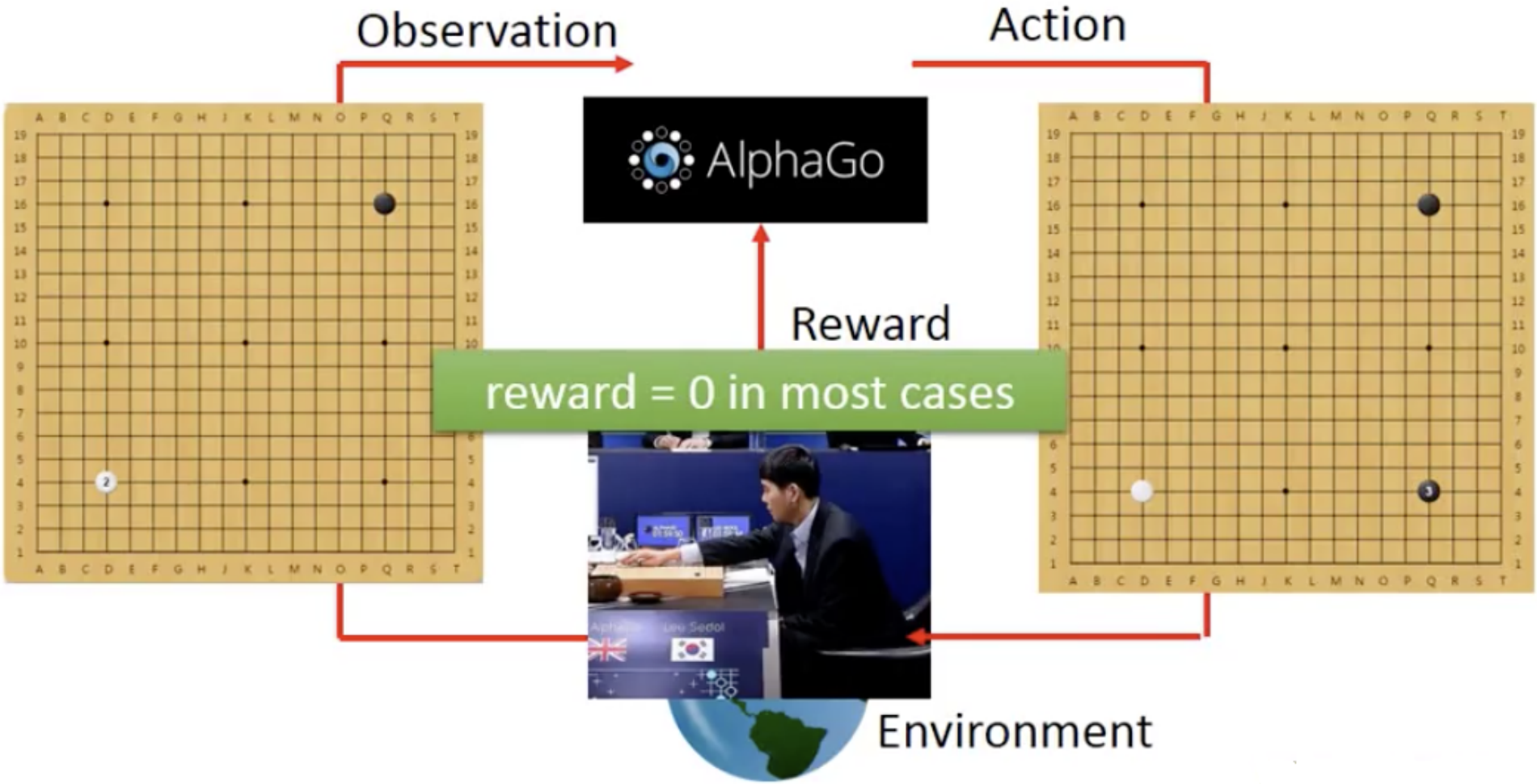
Scenario of Reinforcement Learning



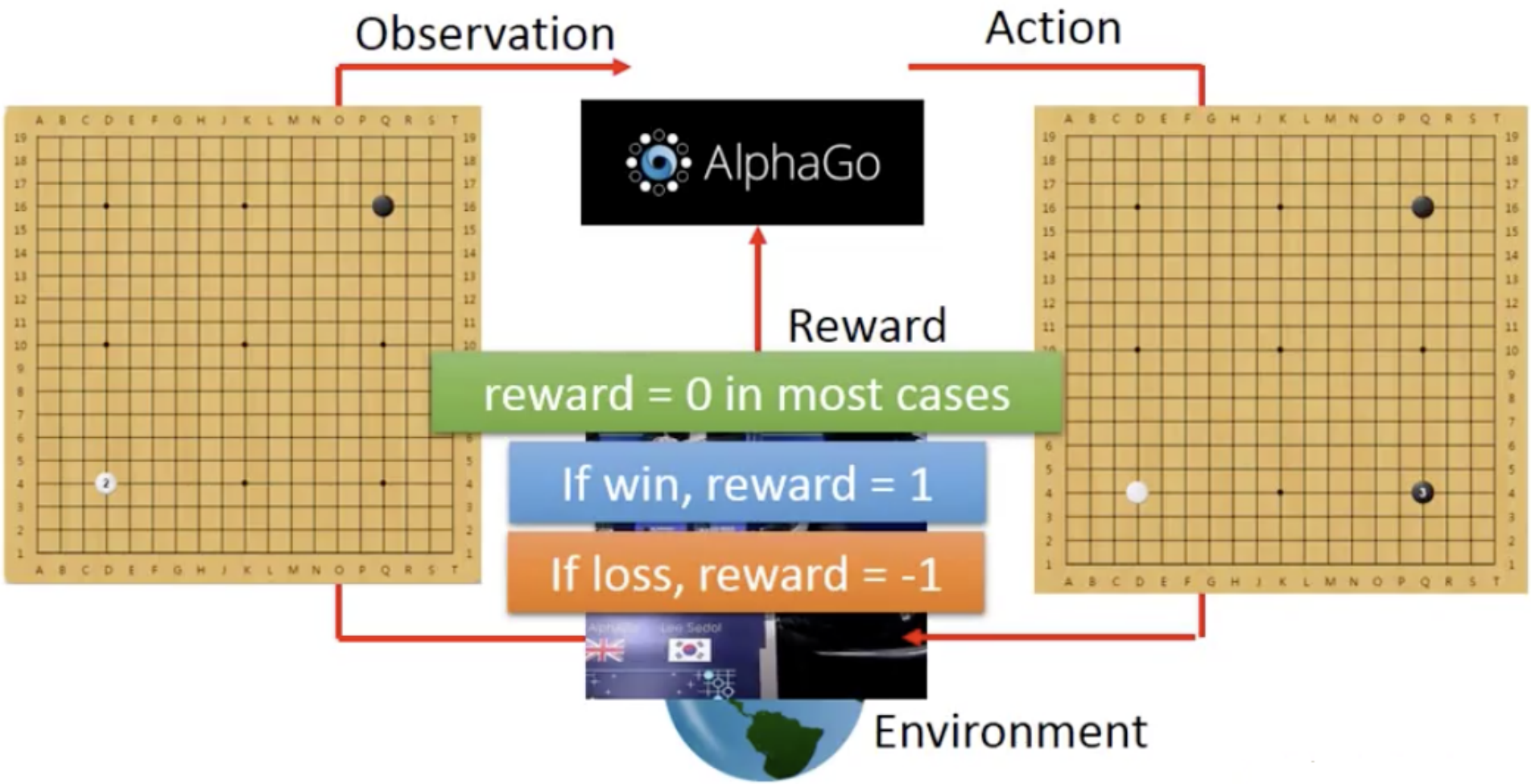
Scenario of Reinforcement Learning



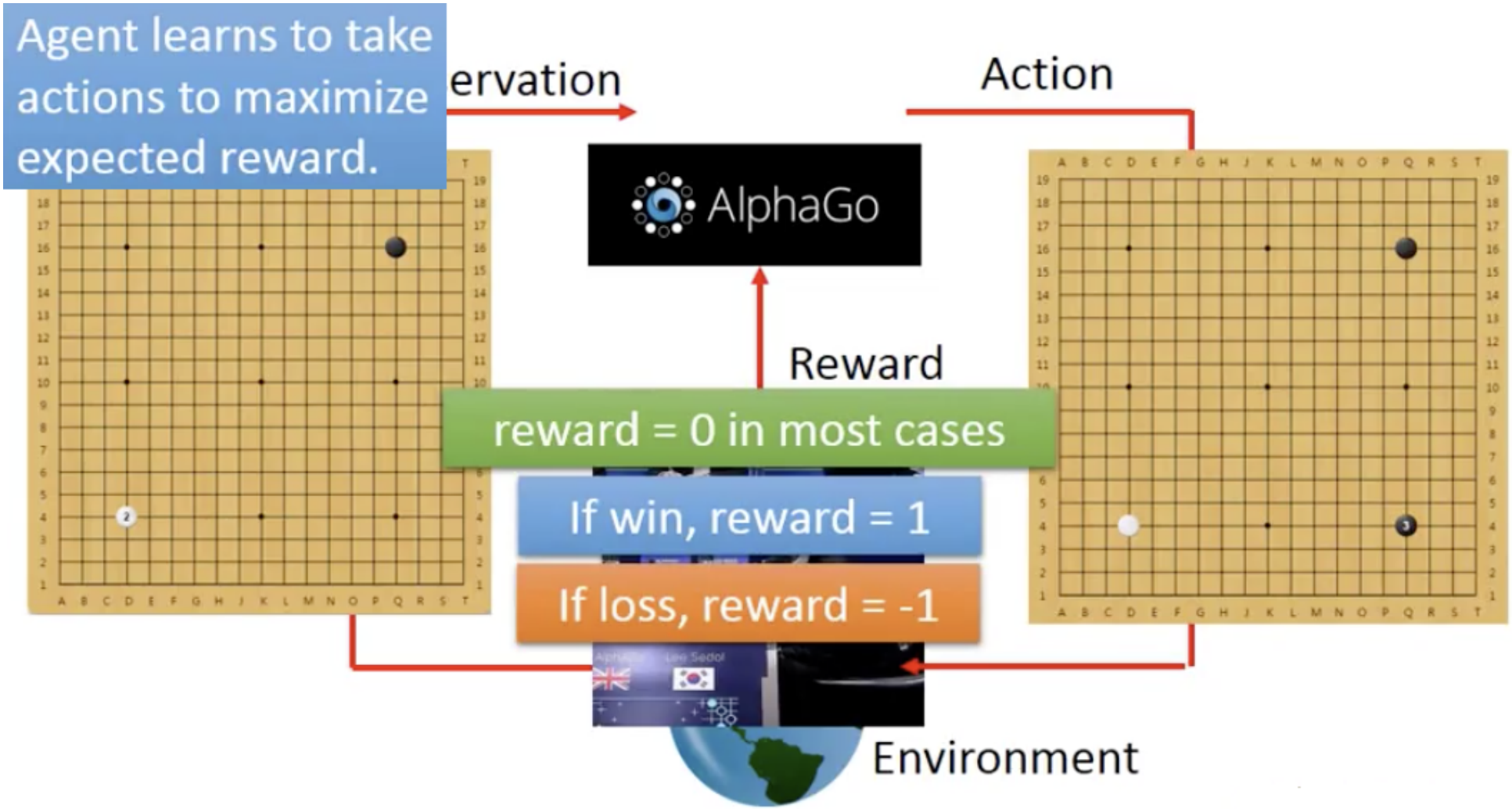
Scenario of Reinforcement Learning



Scenario of Reinforcement Learning



Scenario of Reinforcement Learning



Scenario of Reinforcement Learning

- Supervised v.s. Reinforcement

- Supervised:

- Reinforcement Learning

Scenario of Reinforcement Learning

- Supervised v.s. Reinforcement

- Supervised:

Next move:
"5-5"

Next move:
"3-3"

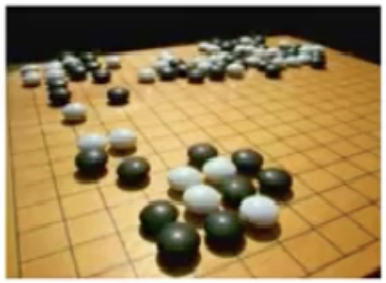
- Reinforcement Learning

Scenario of Reinforcement Learning

- Supervised v.s. Reinforcement

- Supervised: Learning from teacher

Next move: "5-5"



Next move: "3-3"

- Reinforcement Learning

Scenario of Reinforcement Learning

- Supervised v.s. Reinforcement

- Supervised: Learning from teacher

Next move: "5-5"

Next move: "3-3"

- Reinforcement Learning

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

Network

Scenario of Reinforcement Learning

- 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!

Scenario of Reinforcement Learning

- 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.)

Scenario of Reinforcement Learning

- 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.

Network

Scenario of Reinforcement Learning

- Supervised v.s. Reinforcement

Standard (supervised) machine learning:

given $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$

learn to predict y from $\mathbf{x}$ $\quad f(\mathbf{x}) \approx y$

Usually assumes:

- i.i.d. data
- known ground truth outputs in training

Reinforcement learning:

- Data is **not** i.i.d.: previous outputs influence future inputs!
- Ground truth answer is not known, only know if we succeeded or failed
  - more generally, we know the reward

Institute of Media, Information, and Network

Scenario of Reinforcement Learning

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

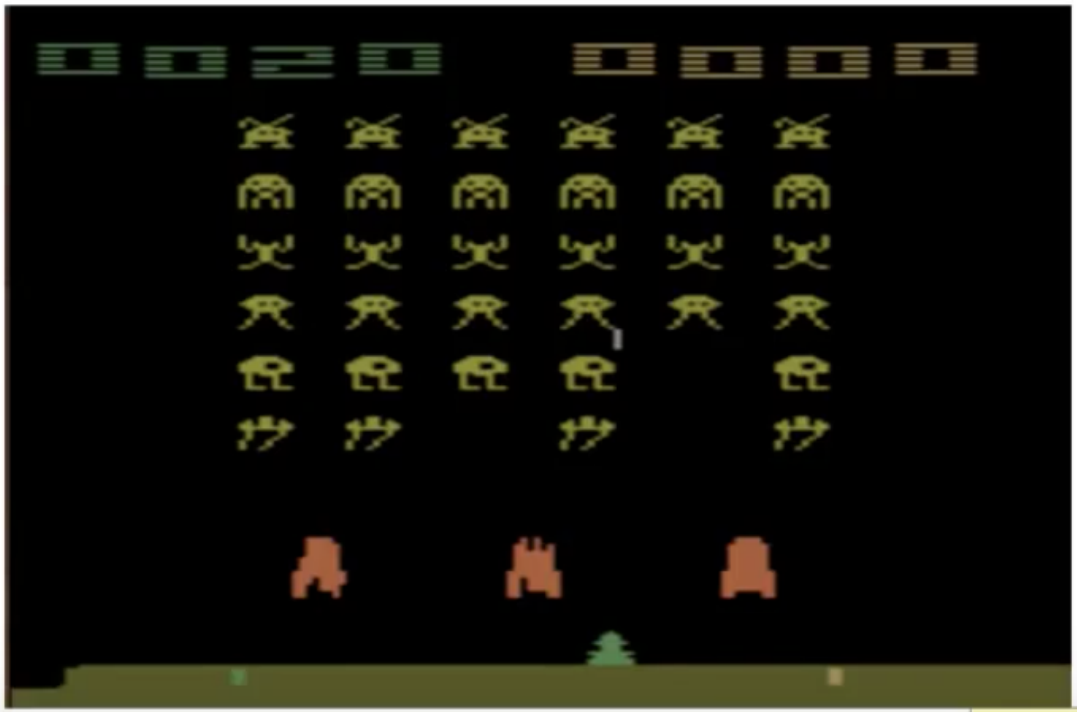
- What machine observes is pixels
- Machine learns to take proper action itself

Media,
, and Network

Scenario of Reinforcement Learning

Example: Playing Video Game

- Space invader

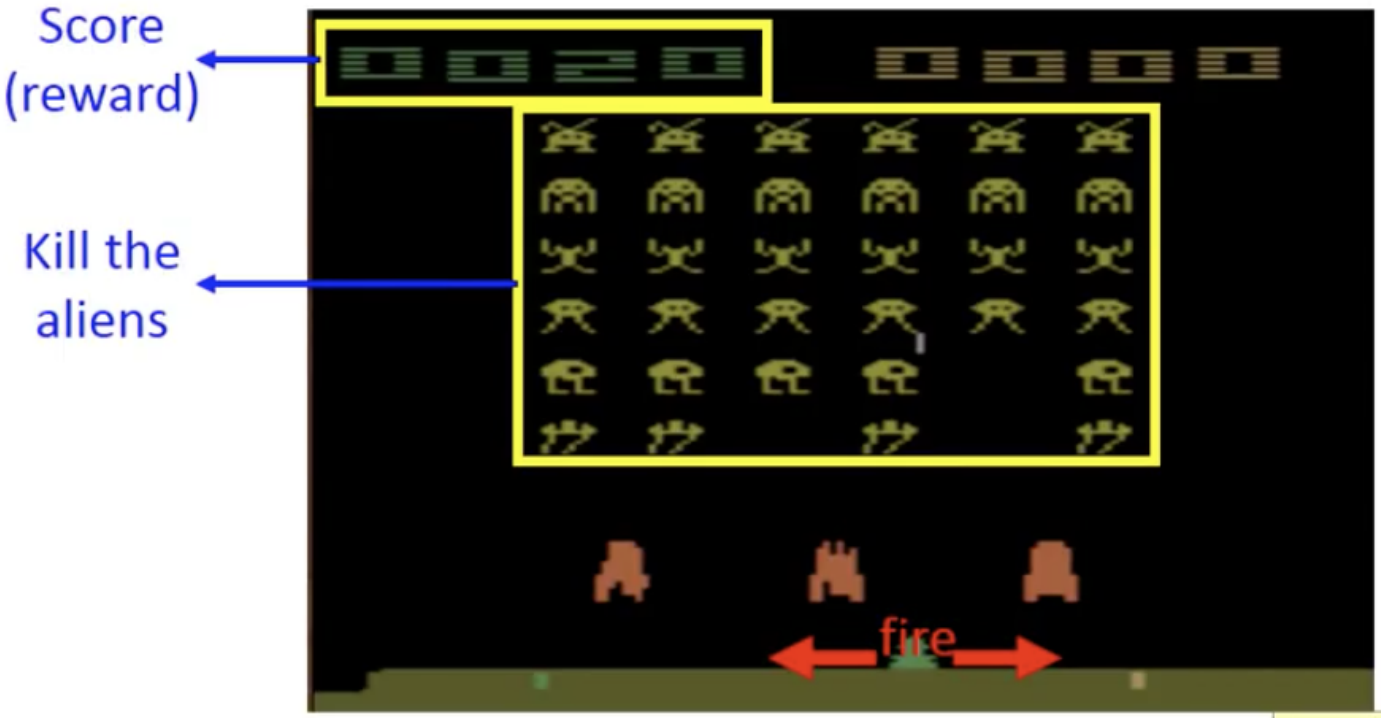


of Media,
on, and Network

Scenario of Reinforcement Learning

Example: Playing Video Game

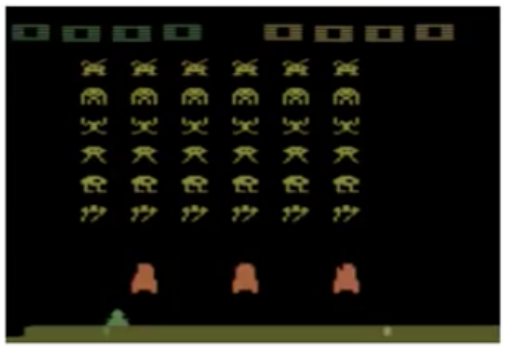
- Space invader



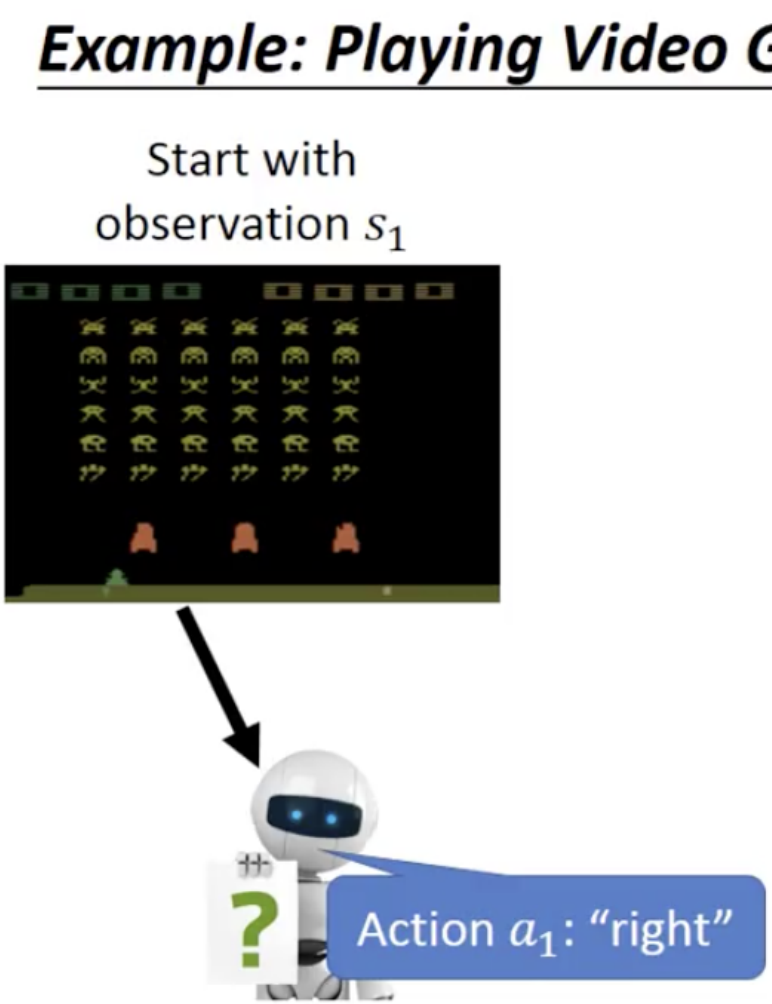
Scenario of Reinforcement Learning

Example: Playing Video Game

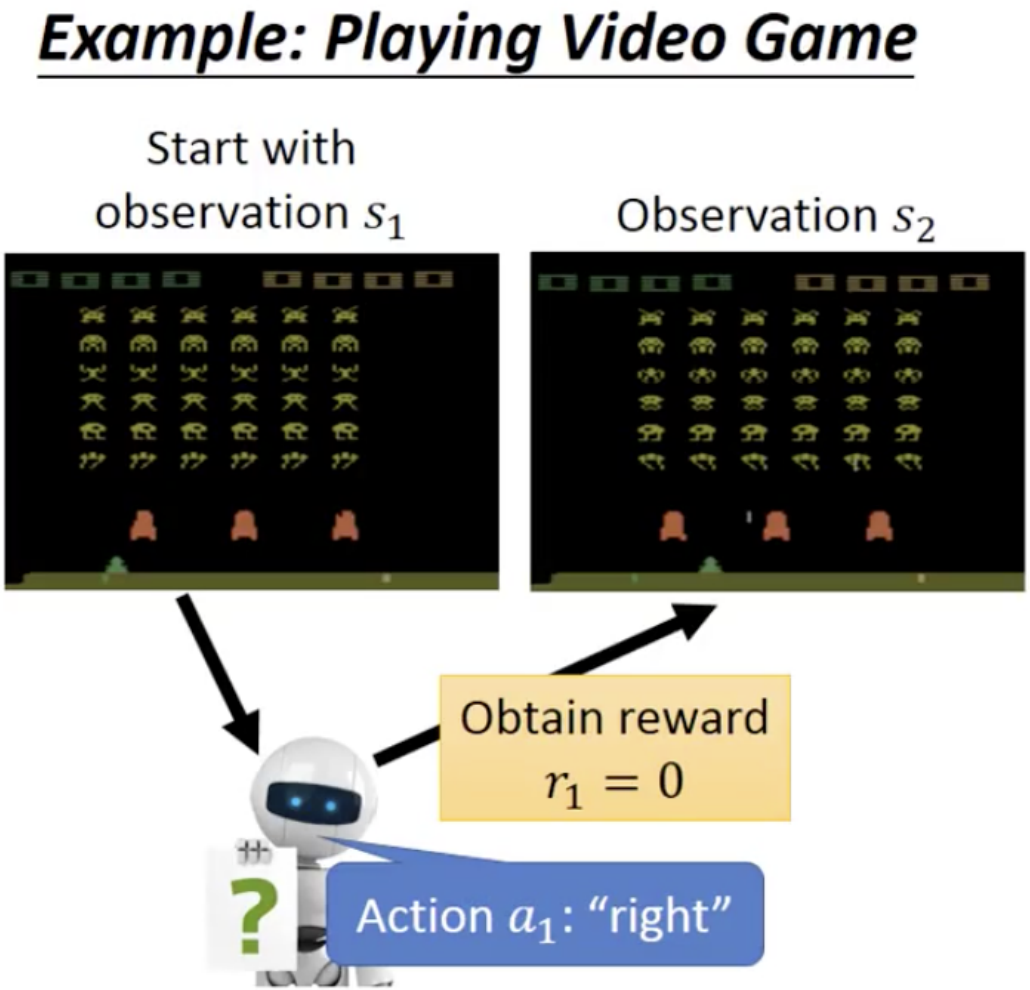
Start with observation s_1



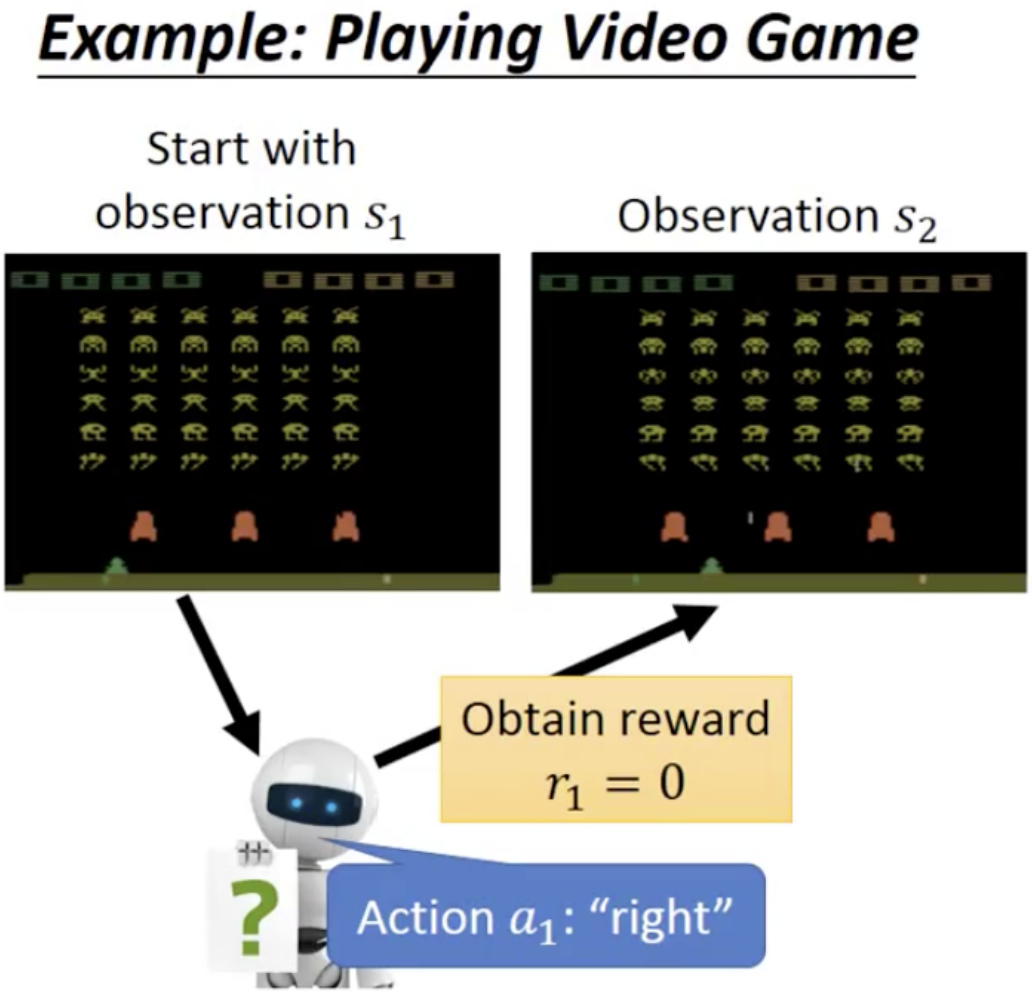
Scenario of Reinforcement Learning



Scenario of Reinforcement Learning

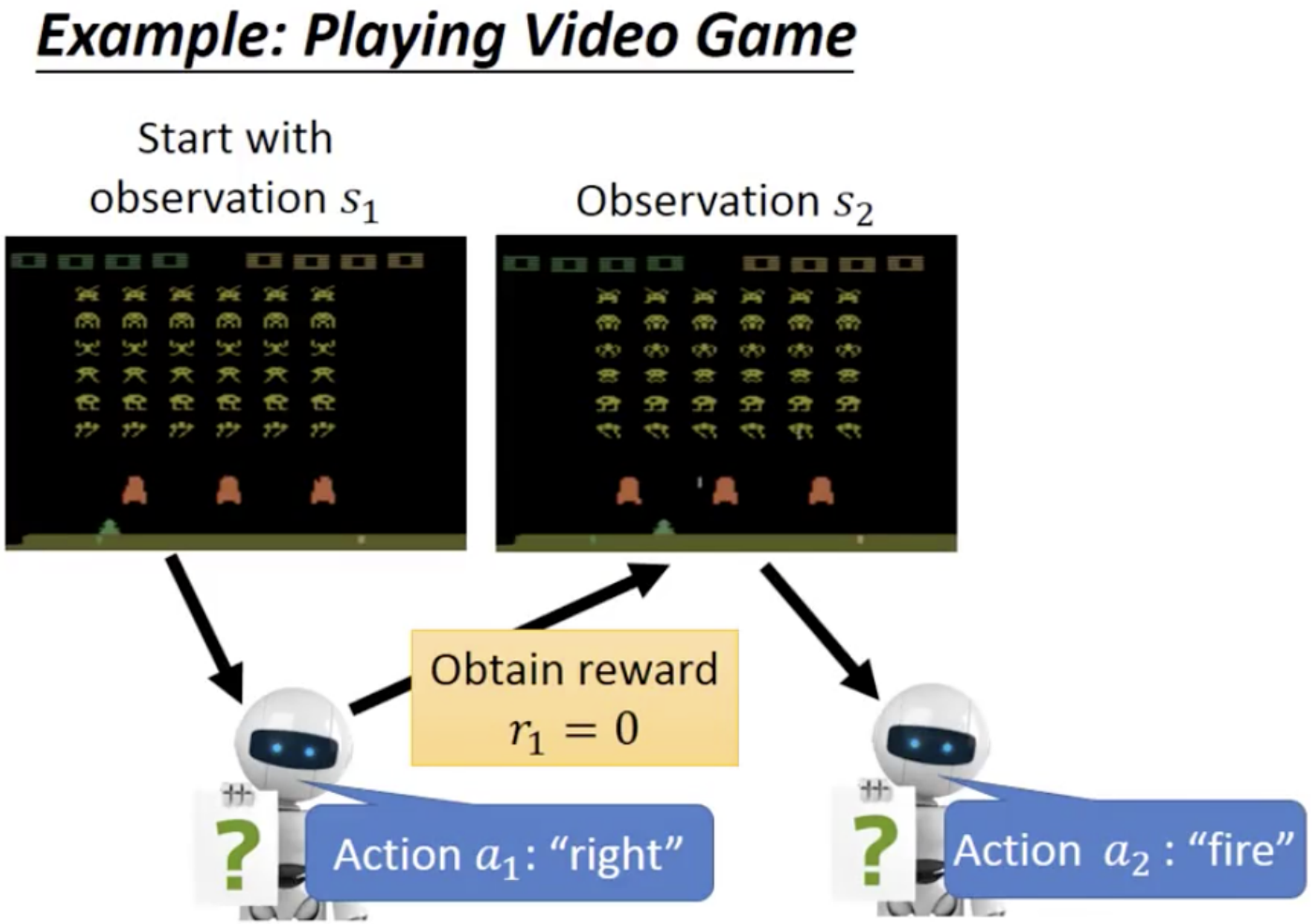


Scenario of Reinforcement Learning



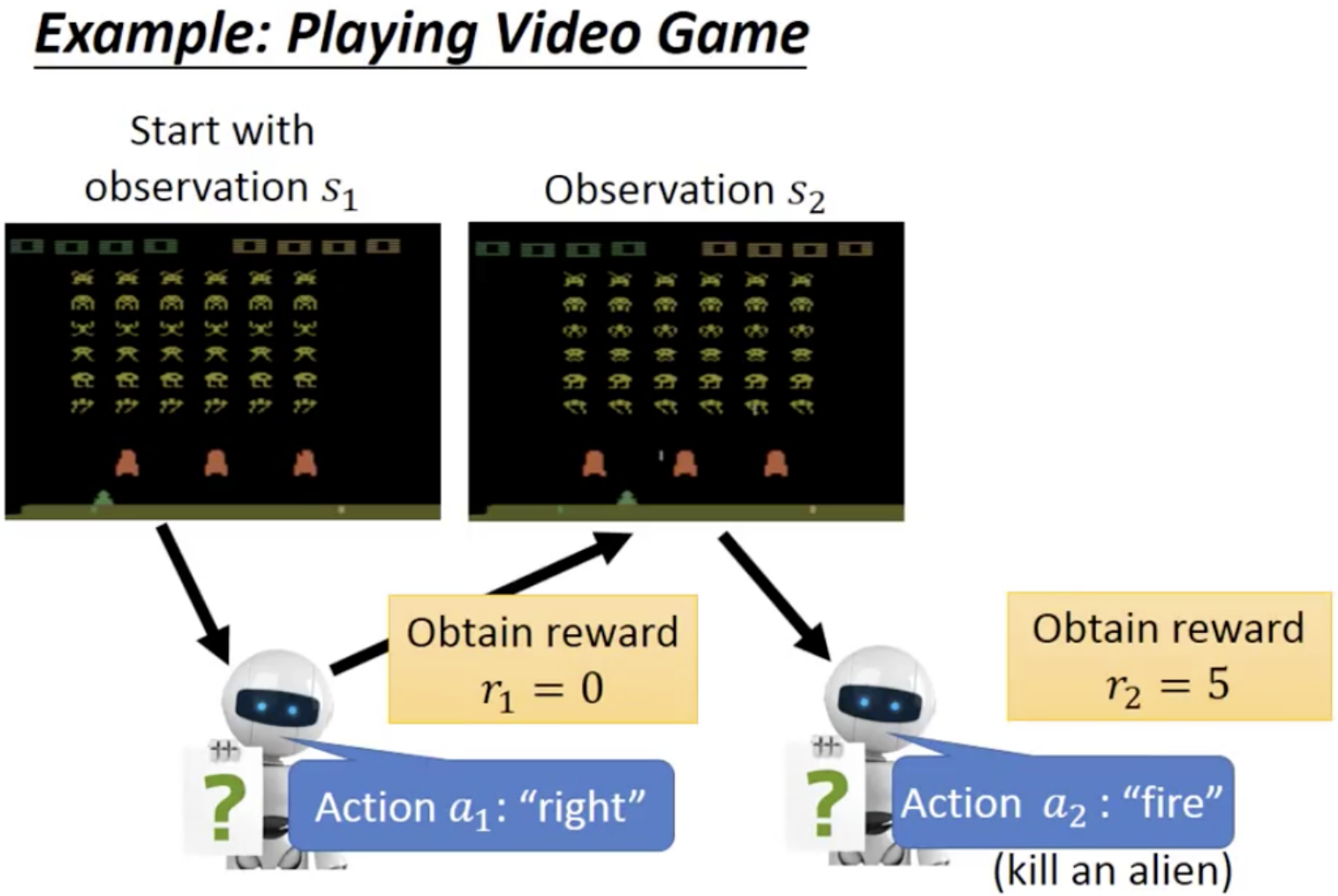
Usually there is some randomness in the environment

Scenario of Reinforcement Learning

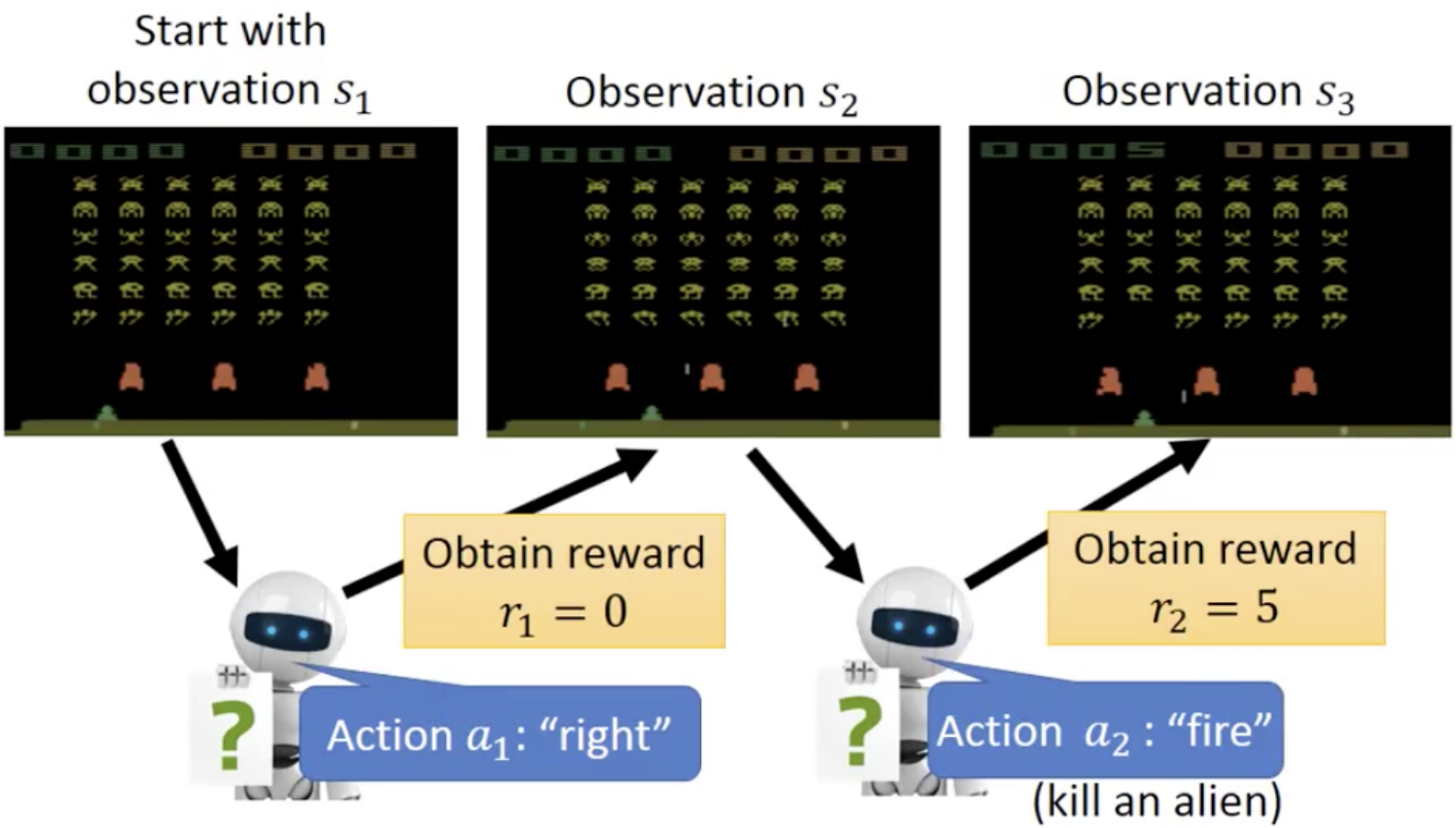


Usually there is some randomness in the environment

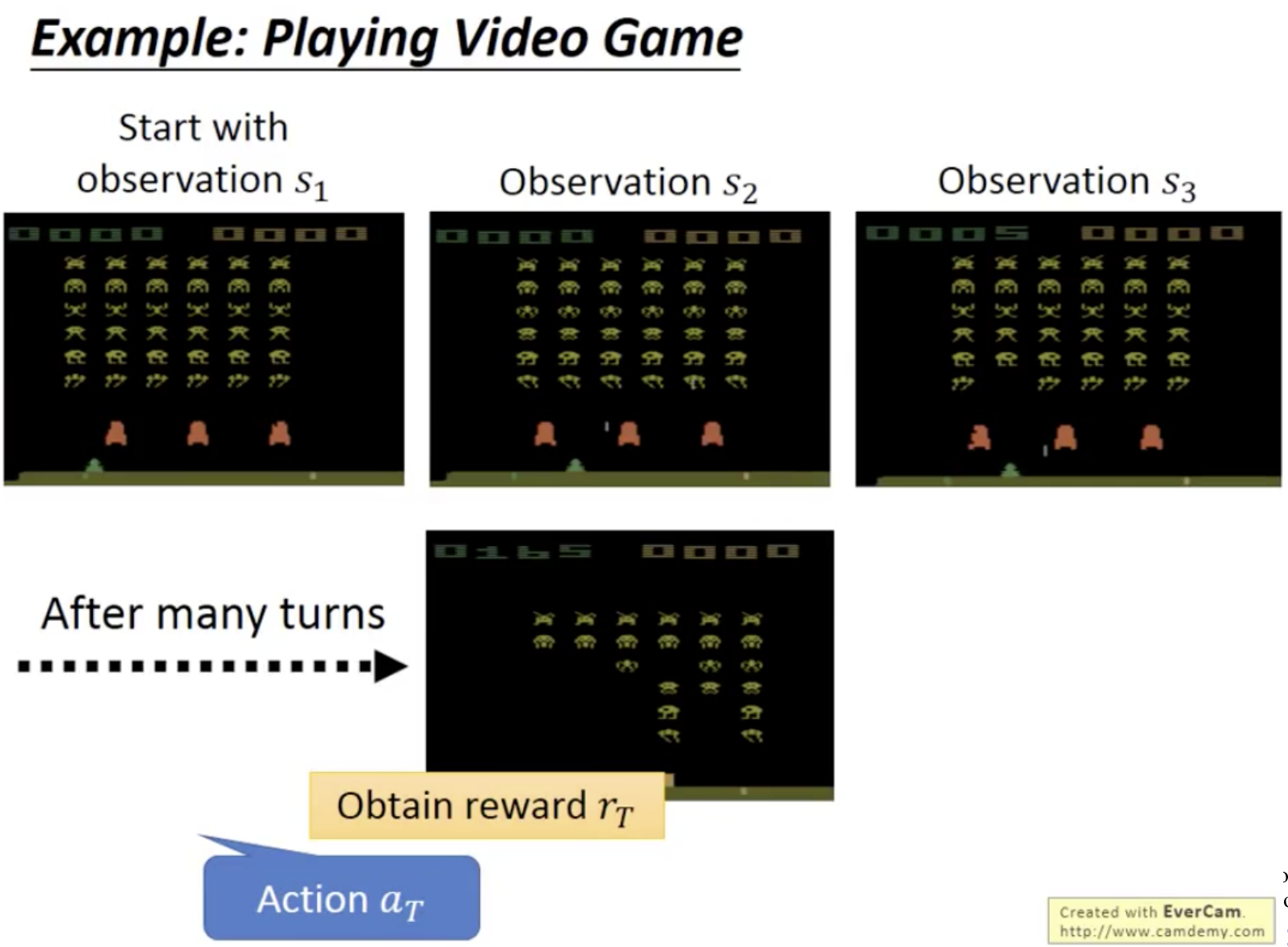
Scenario of Reinforcement Learning



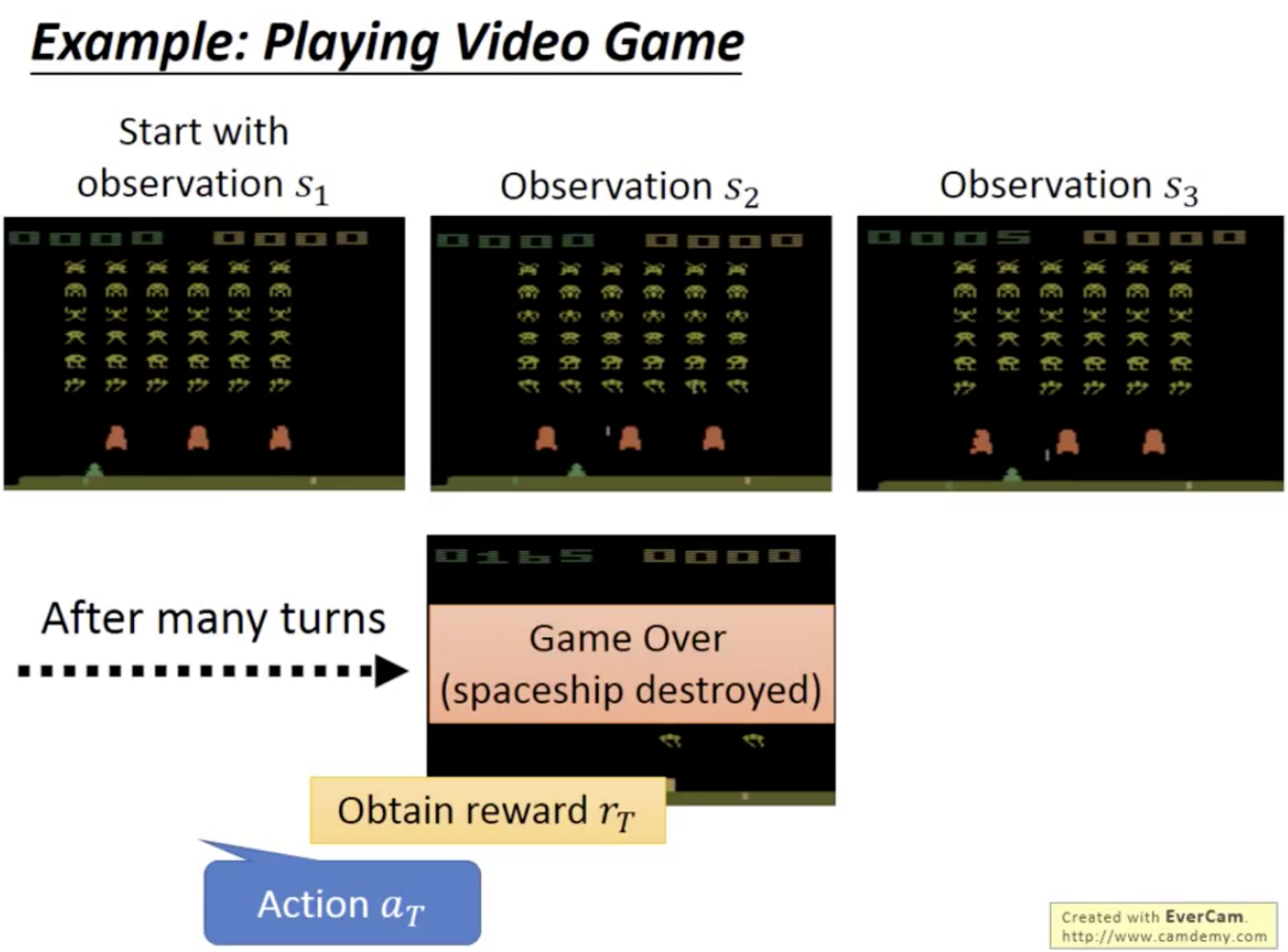
Scenario of Reinforcement Learning



Scenario of Reinforcement Learning

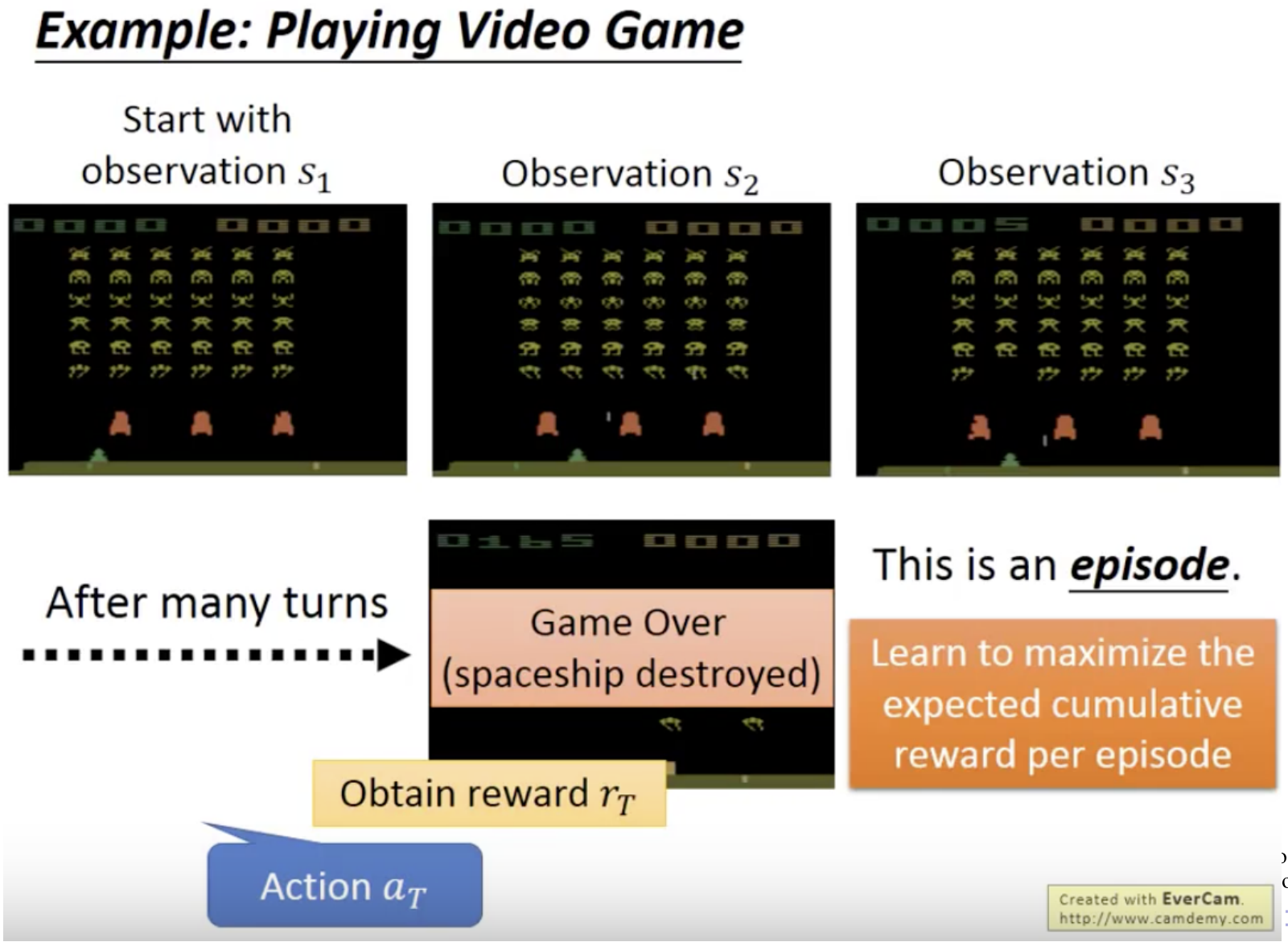


Scenario of Reinforcement Learning



of Media, on, and Network

Scenario of Reinforcement Learning



of Media, on, and Network

Scenario 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
- Agent's actions affect the subsequent data it receives
  - E.g. Exploration

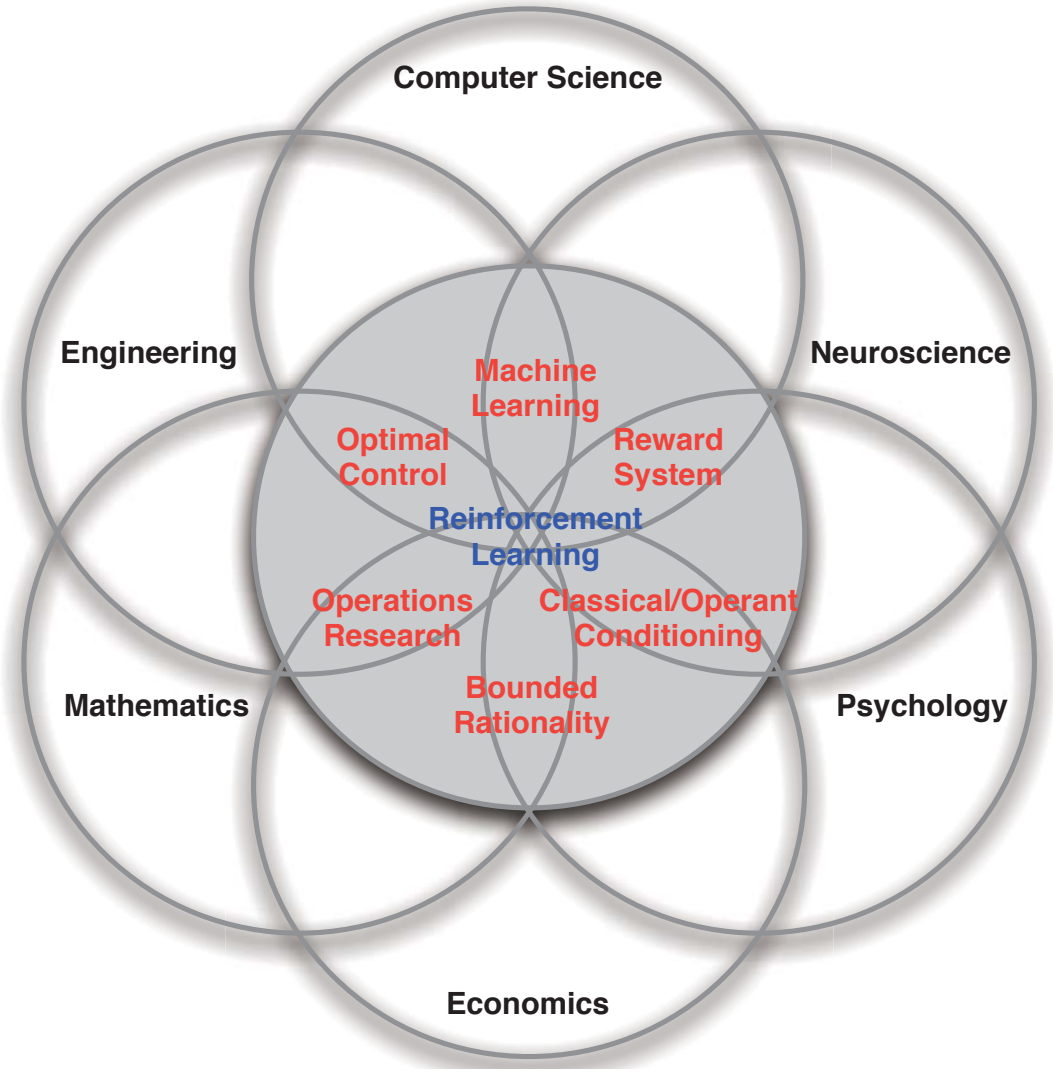
Media,
, and Network

Characteristics of Reinfrocement Learning

What makes reinforcement learning different from other machine learning paradigms?

- There is no supervisor, only a *reward* signal
- Feedback is delayed, not instantaneous
- Time really matters (sequential, non i.i.d data)
- Agent's actions affect the subsequent data it receives

Many Faces of Reinfrocement Learning



Institute of Media, Information, and Network

Examples of RL

- Defeat the world champion at Backgammon
- Manage an investment portfolio
- Control a power station
- Make a humanoid robot walk
- Play many different Atari games better than humans

Table of Contents

Rewards

A *reward signal* defines the goal in a reinforcement learning problem

- A reward R_t is a scalar feedback signal
- The agent's sole objective is to maximize the total reward it receives over the long run (cumulative reward)
- Indicates how well the agent is doing at step t

Reinforcement learning is based on the reward hypothesis

Definition (Reward Hypothesis)

All goals can be described by the maximization of expected cumulative reward

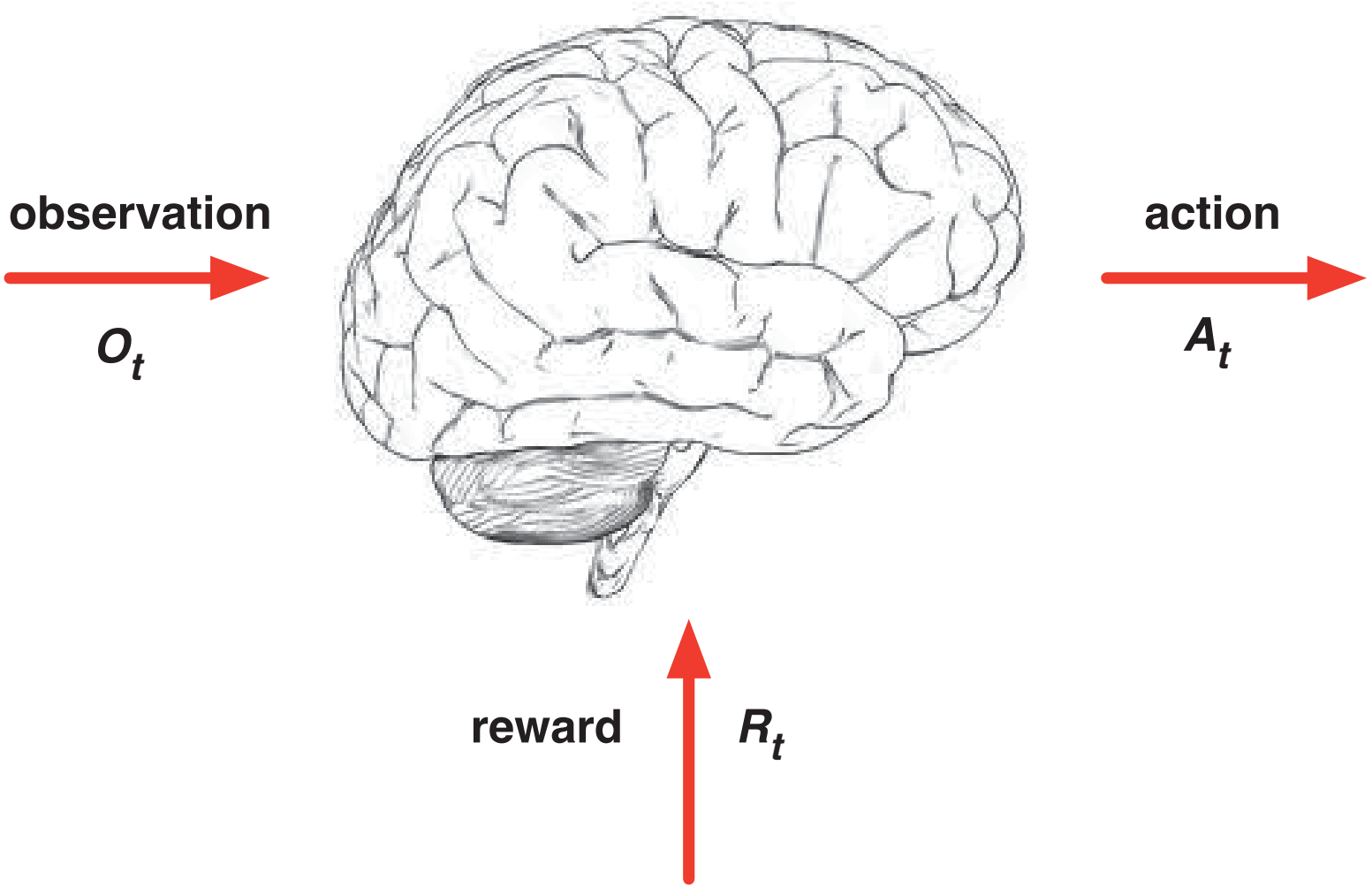
Examples of Rewards

- Fly stunt manoeuvres in a helicopter
  - + value for following desired trajectory
  - − value for crashing
- Defeat the world champion at Backgammon
  - +/− value for winning/losing a game
- Manage an investment portfolio
  - + value for each $ in bank
- Control a power station
  - + value for producing power
  - − value for exceeding safety thresholds
- Make a humanoid robot walk
  - + value for forward motion
  - − value for falling over
- Play many different Atari games better than humans
  - +/− value for increasing/decreasing score

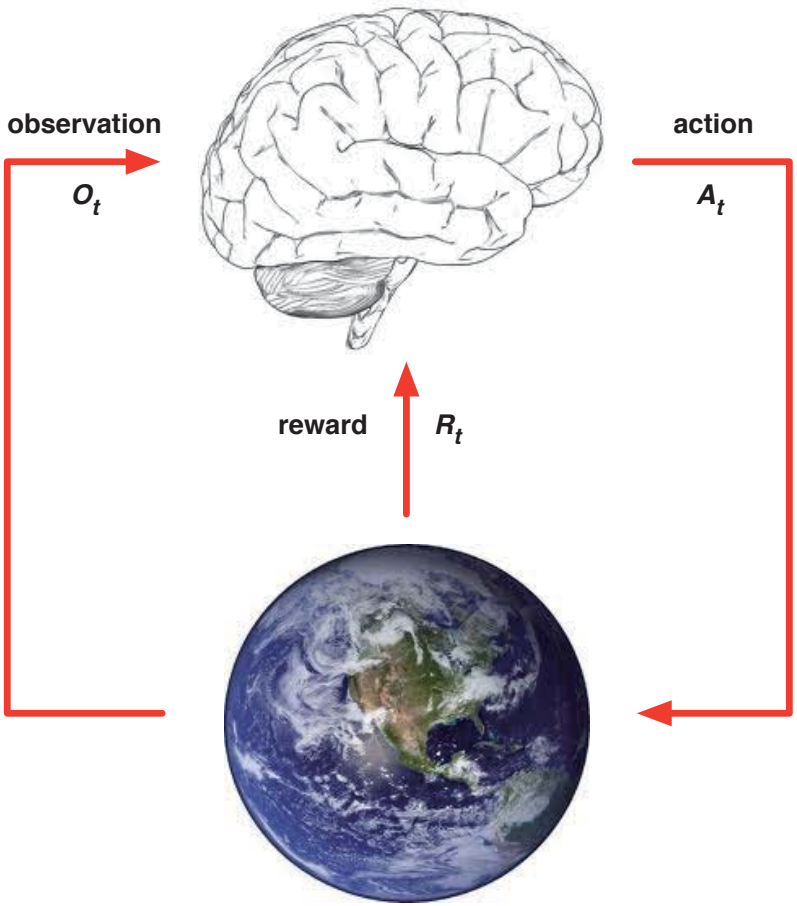
Sequential Decision Making

- Goal: select actions to maximize total future reward
- Actions may have long term consequences
- Reward may be delayed
- It may be better to sacrifice immediate reward to gain more long-term reward
- Examples
  - A financial investment (may take months to mature)
  - Refuelling a helicopter (might prevent a crash in several hours)
  - Blocking opponent moves (might help winning chances many moves from now)

Agent and Environment



Agent and Environment



- At each step t the agent:
  - Executes action A_t
  - Receives observation O_t
  - Receives scalar reward R_t
- The environment:
  - Receives action A_t
  - Emits observation O_{t+1}
  - Emits scalar reward R_{t+1}
- t increments at every step

Institute of Media, Information, and Network

History and State

- The history is the sequence of observations, actions, rewards

$$H_t = O_1, R_1, A_1, \cdots, A_{t-1}, O_t, R_t$$

- i.e. all observable variables up to time t
- i.e. the sensorimotor stream of a robot or embodied agent
- What happens next depends on the history:
  - The agent selects actions
  - The environment selects observations/rewards
- State is the information used to determine what happens next
- Formally, state is a function of the history:

$$S_t = f(H_t)$$

Information State

An information state (a.k.a. Markov) contains all useful information from the history.

Definition

A state S_t is Markov if and only if

$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, \cdots, S_t]$$

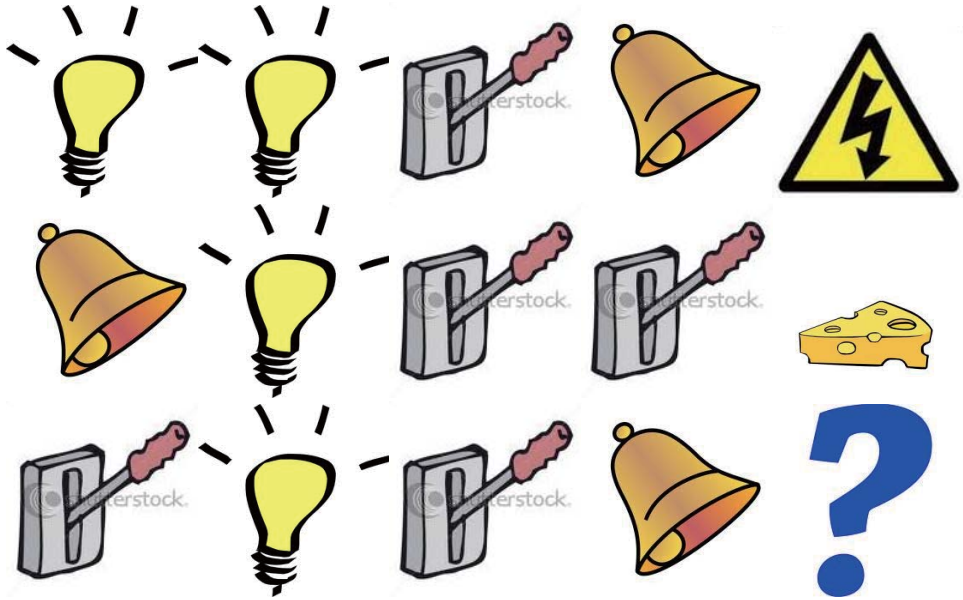
- *The future is independent of the past given the present*

$$H_{1:t} \to S_t \to H_{t+1:\infty}$$

- Once the state is known, the history may be thrown away
- i.e. The state is a sufficient statistic of the future
- The environment state S_t^e is Markov
- The history H_t is Markov

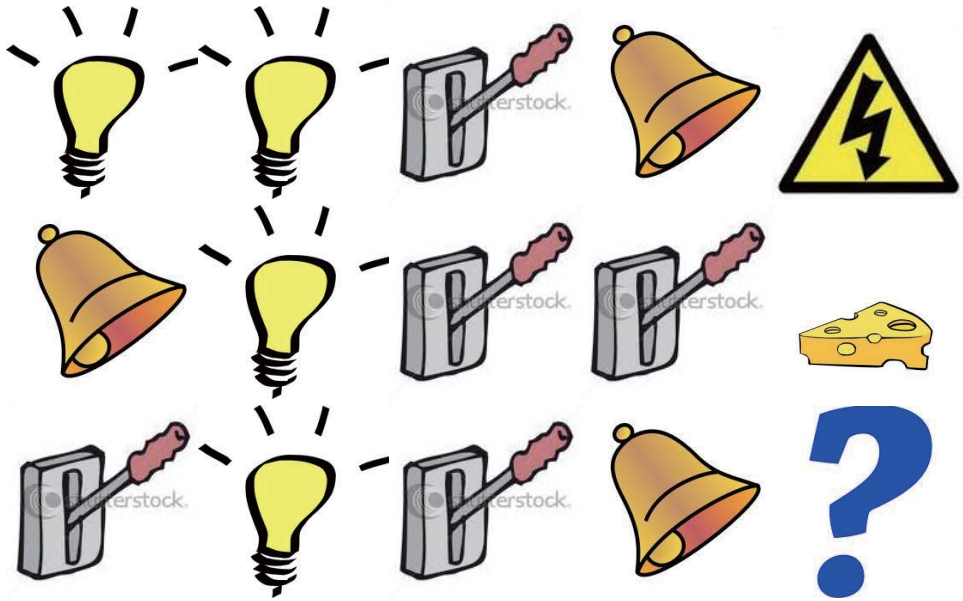
Institute of Media, Information, and Network

Rat Example



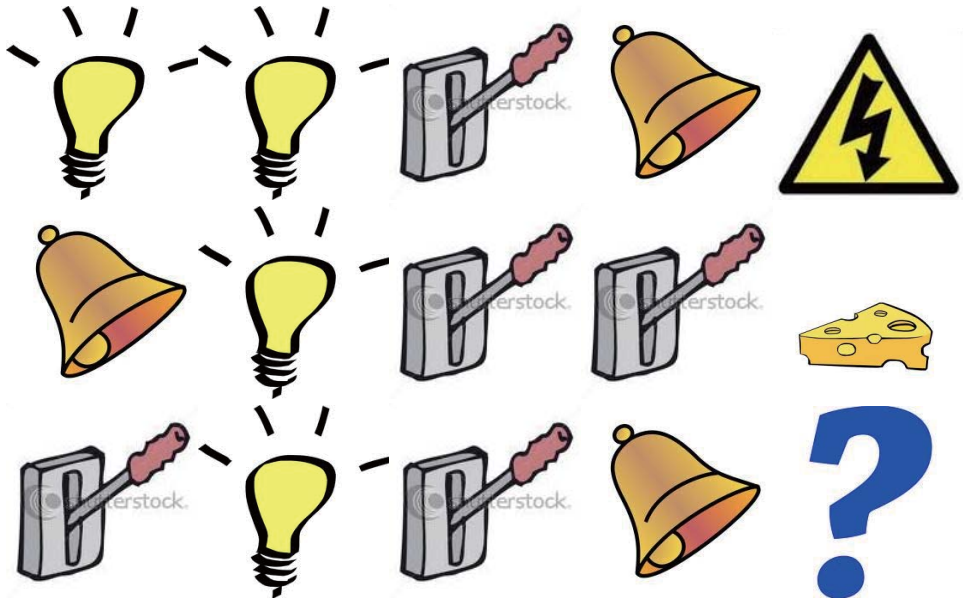
Institute of Media, Information, and Network

Rat Example



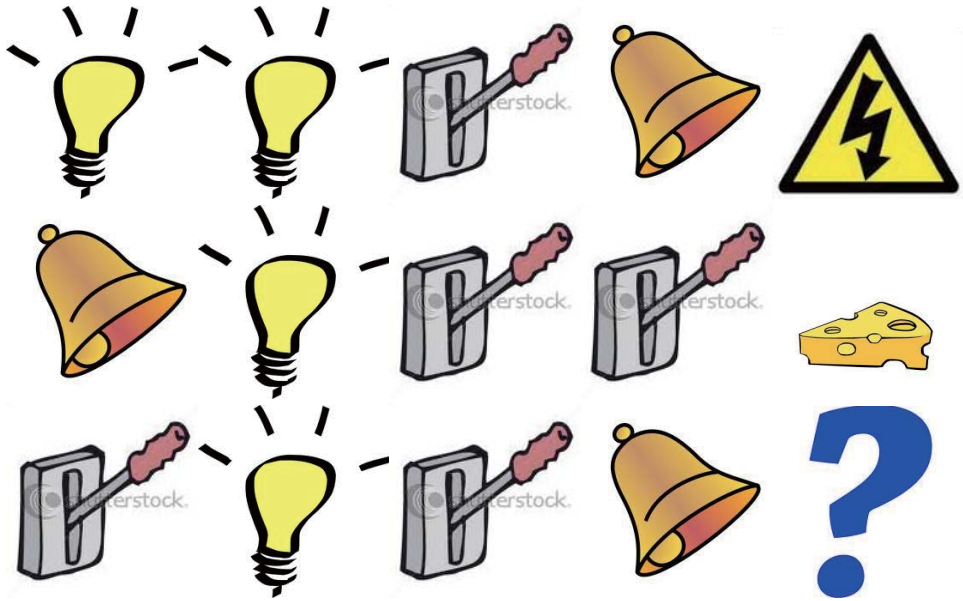
- What if agent state = last 3 items in sequence?

Rat Example



- What if agent state = last 3 items in sequence?
- What if agent state = counts for lights, bell and levers?

Rat Example



- What if agent state = last 3 items in sequence?
- What if agent state = counts for lights, bell and levers?
- What if agent state = complete sequence?

Institute of Media, Information, and Network

Table of Contents

Major Components of an RL Agent

An RL agent may include one or more of these components:

- Policy: agent's behaviour function
- Value function: how good is each state and/or action
- Model: agent's representation of the environment

Policy

- A policy is the agent's behaviour
- It is a map from state to action, e.g.
- Deterministic policy: $a = \pi(s)$
- Stochastic policy: $\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]$

Value Function

- Value function is a prediction of future reward
- Used to evaluate the goodness/ badness of states
- And therefore to select between actions, e.g.

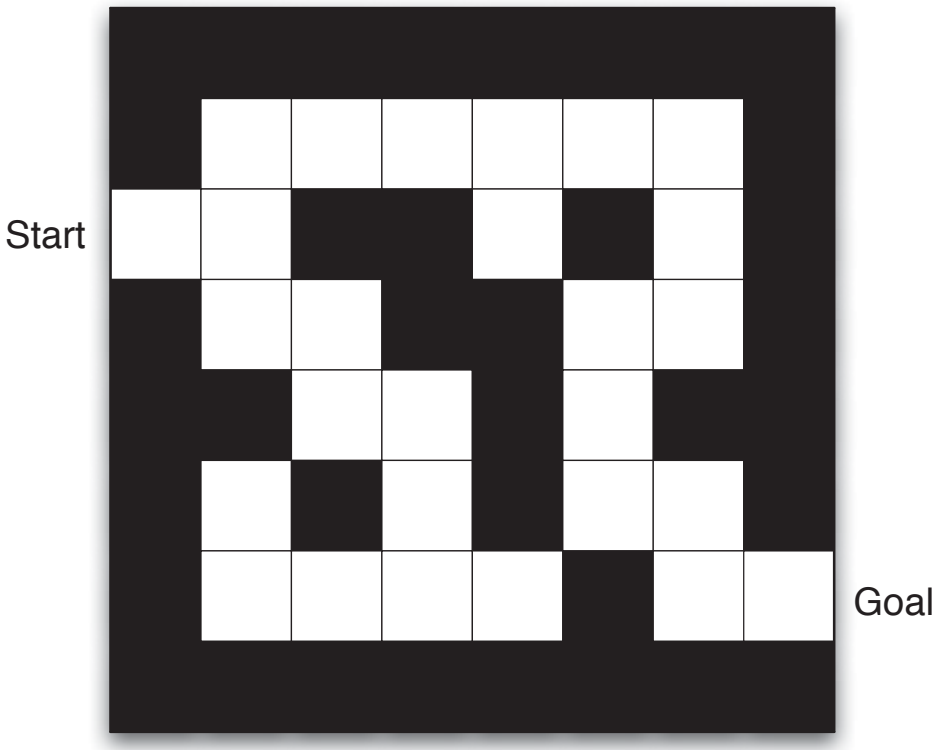
$$v_\pi(s) = \mathbb{E}_\pi\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = s\right]$$

Model

- A model predicts what the environment will do next
- Transitions: $\mathcal{P}$ predicts the next state
- Rewards: $\mathcal{R}$ predicts the next (immediate) reward, e.g.

$$\mathcal{P}^a_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$$
$$\mathcal{R}^a_s = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$$

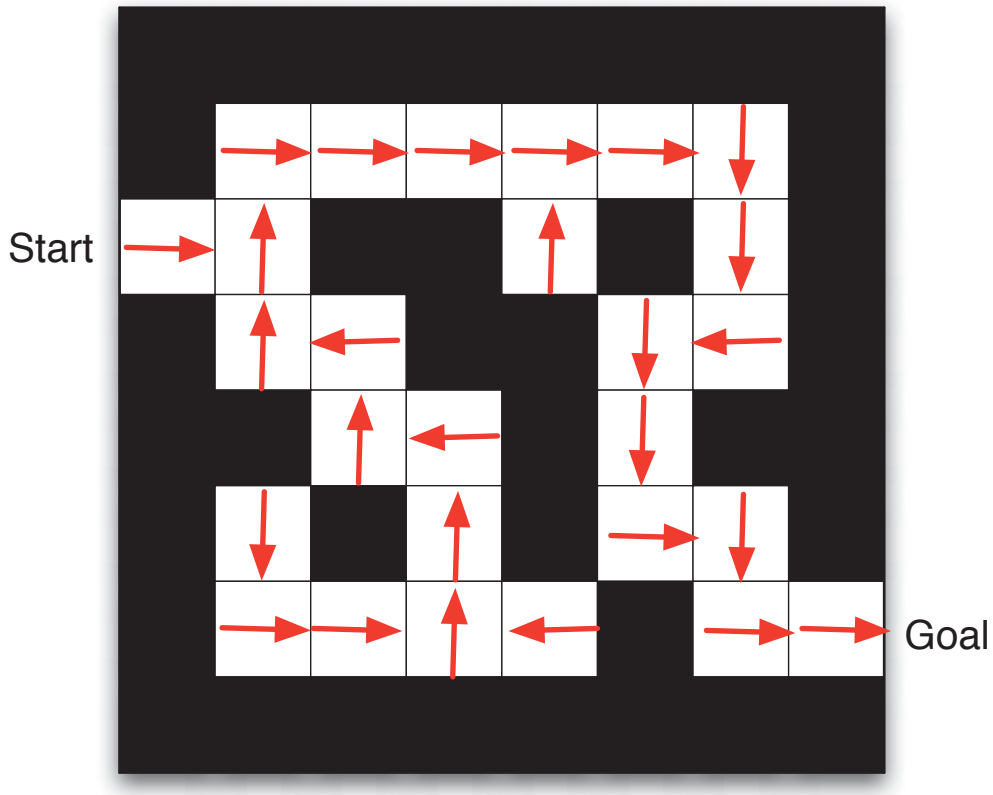
Maze Example



- Rewards: -1 per time-step
- Actions: N, E, S, W
- States: Agent's location

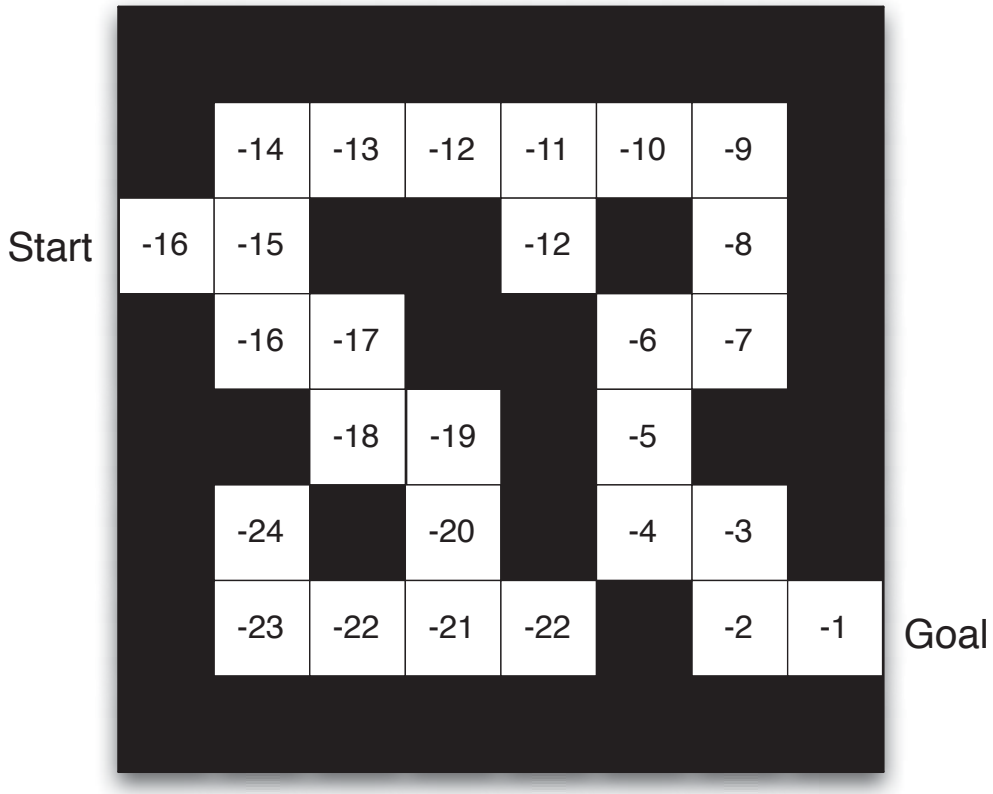
Institute of Media, Information, and Network

Maze Example: Policy



Arrows represent policy $\pi(s)$ for each state s

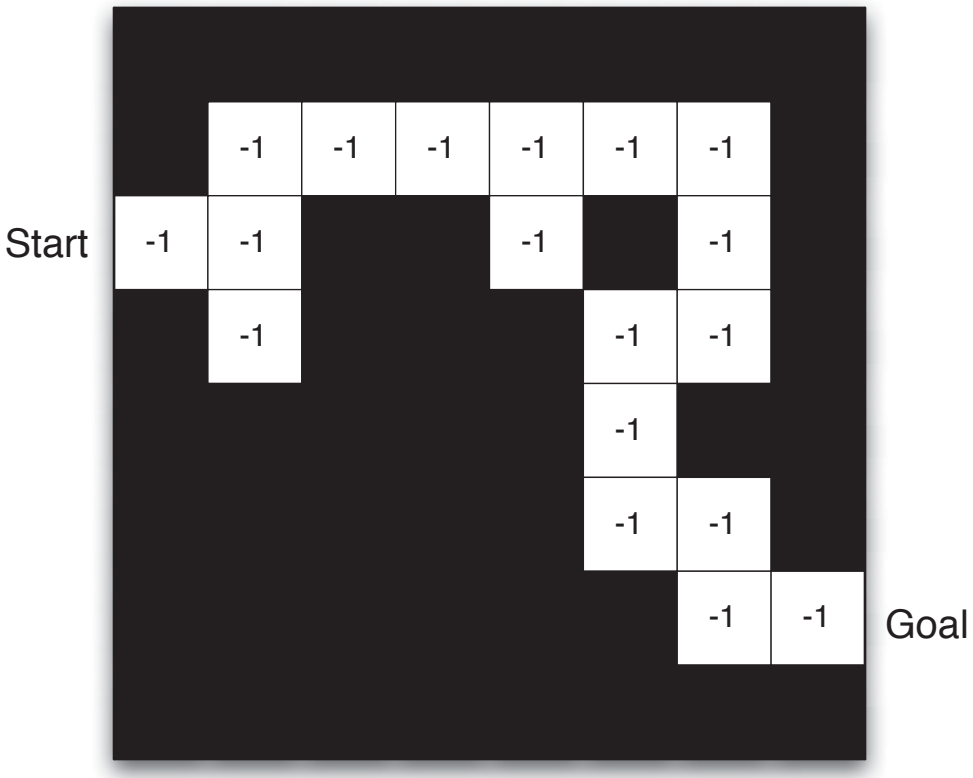
Maze Example: Value Function



Numbers represent value function $v_\pi(s)$ for each state s

Institute of Media, Information, and Network

Maze Example: Model



- Agent may have an internal model of the environment
- Dynamics: how actions change the state
- Rewards: how much reward from each state
- The model may be imperfect
- Grid layout represents transition model $\mathcal{P}^a_{ss'}$
- Numbers represent immediate reward $\mathcal{R}^{\dashv}_{\int}$ from each state s (same for all a)

Institute of Media, Information, and Network

Categorizing RL Approaches (1)

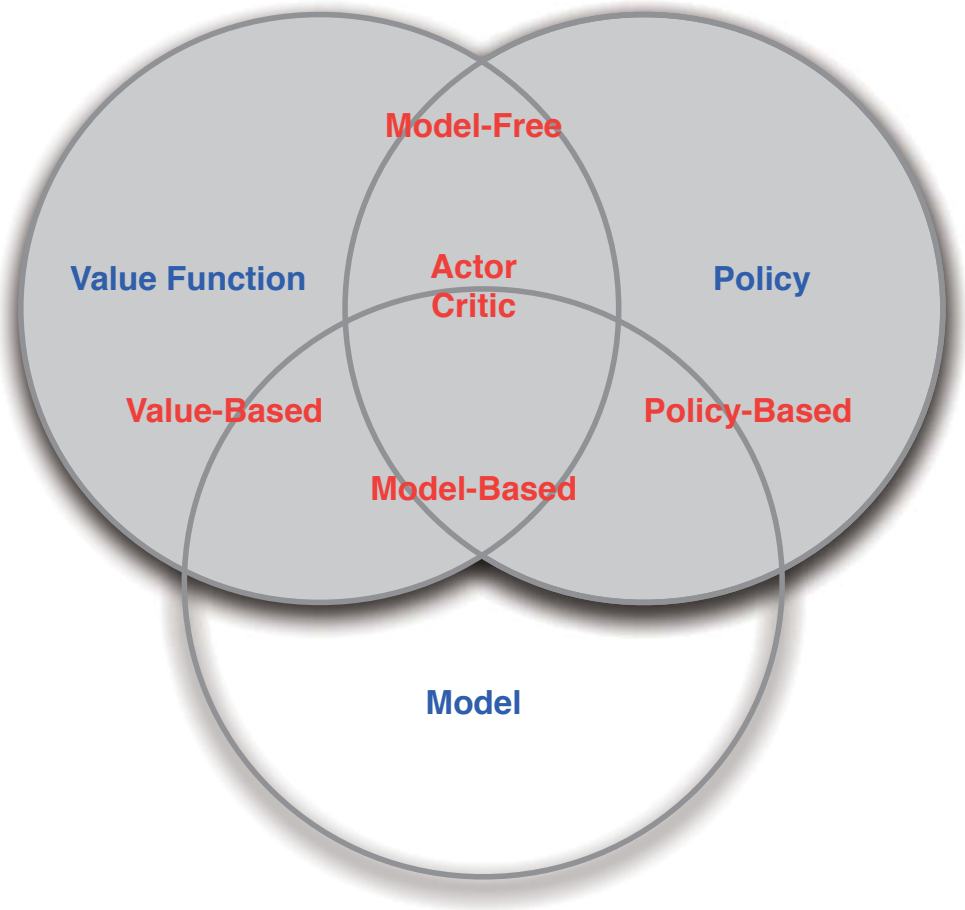
How to choose actions?

- Value Based
  - No Policy (Implicit)
  - Value Function
- Policy Based
  - Policy
  - No Value Function
- Actor-Critic
  - Policy
  - Value Function

Whether can we learn a model?

- Model free
  - Policy and/or Value Function
  - No Model
- Model Based
  - Policy and/or Value Function
  - Model

Categorizing RL Approaches (2)



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

Institute of Media, Information, and Network