南京大学 人工智能学院
SCHOOL OF ARTIFICIAL INTELLIGENCE, NANJING UNIVERSITY

QQ：810862334

www.lamda.nju.edu.cn/IntroRL

Reinforcement Learning Introduction

前期知识

概率论
统计
算法
程序设计
机器学习、深度学习

参考材料

Richard S. Sutton and Andrew G. Barto
Reinforcement Learning: An Introduction

David Silver课程视频： Introduction to Reinforcement Learning （10节课）
https://www.bilibili.com/video/BV17x411Z7Zo?zw

参考材料（进阶）

Berkeley CS285 2020课程
https://www.bilibili.com/video/BV1154y1k7ZE?from=search&seid=12966764310094379808

Deepmind 2018课程
https://www.bilibili.com/video/BV16t411y7Gq?p=1

RLChina 2020课程
网址：https://rlchina.org/ 或直接点击下方链接

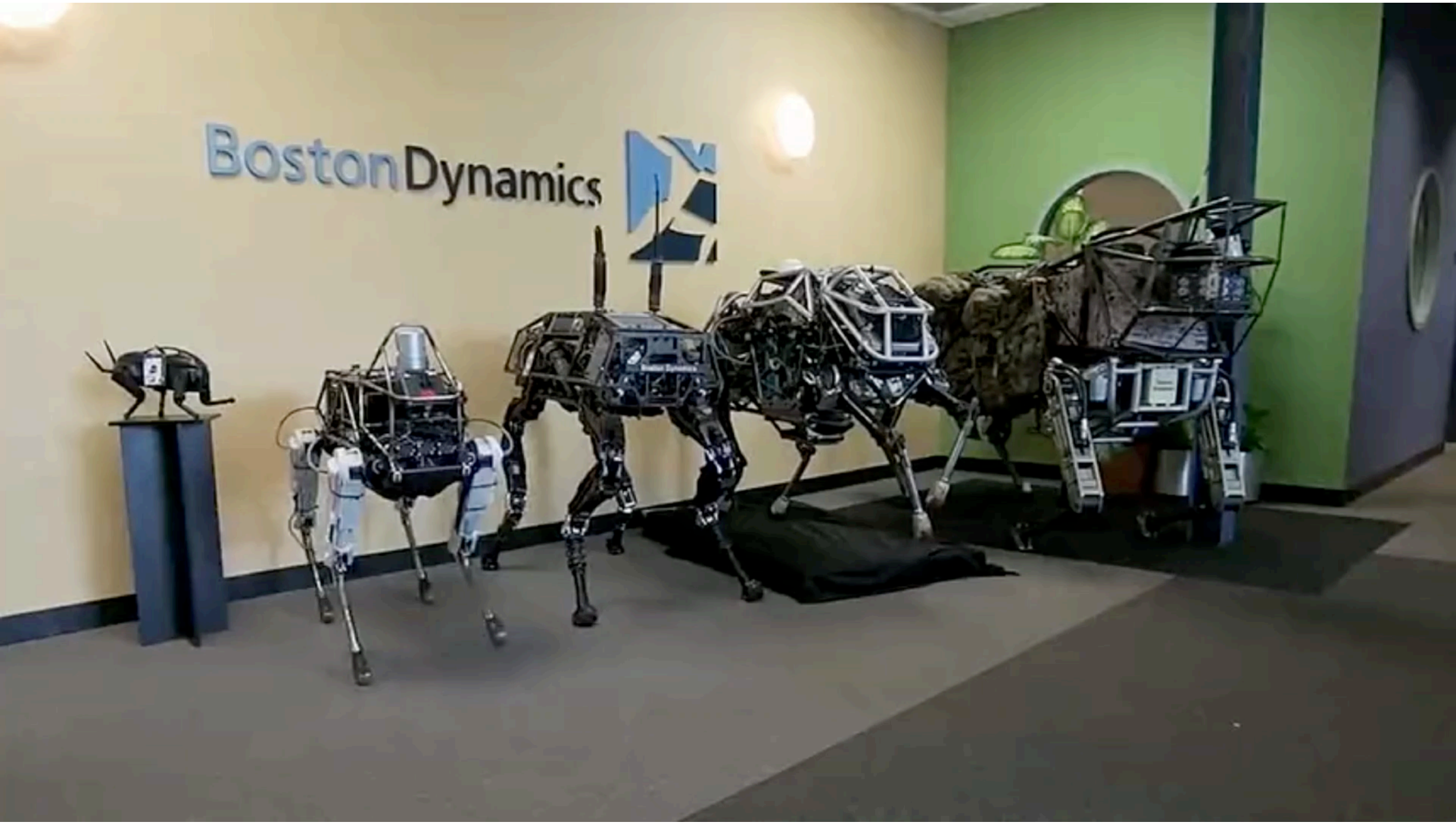
课程成绩

5次作业，每一次20分

作业1：模仿学习：Dagger算法
作业2：强化学习：Q-learning
作业3：深度强化学习：DQN
作业4：model-based强化学习
作业5：offline RL算法

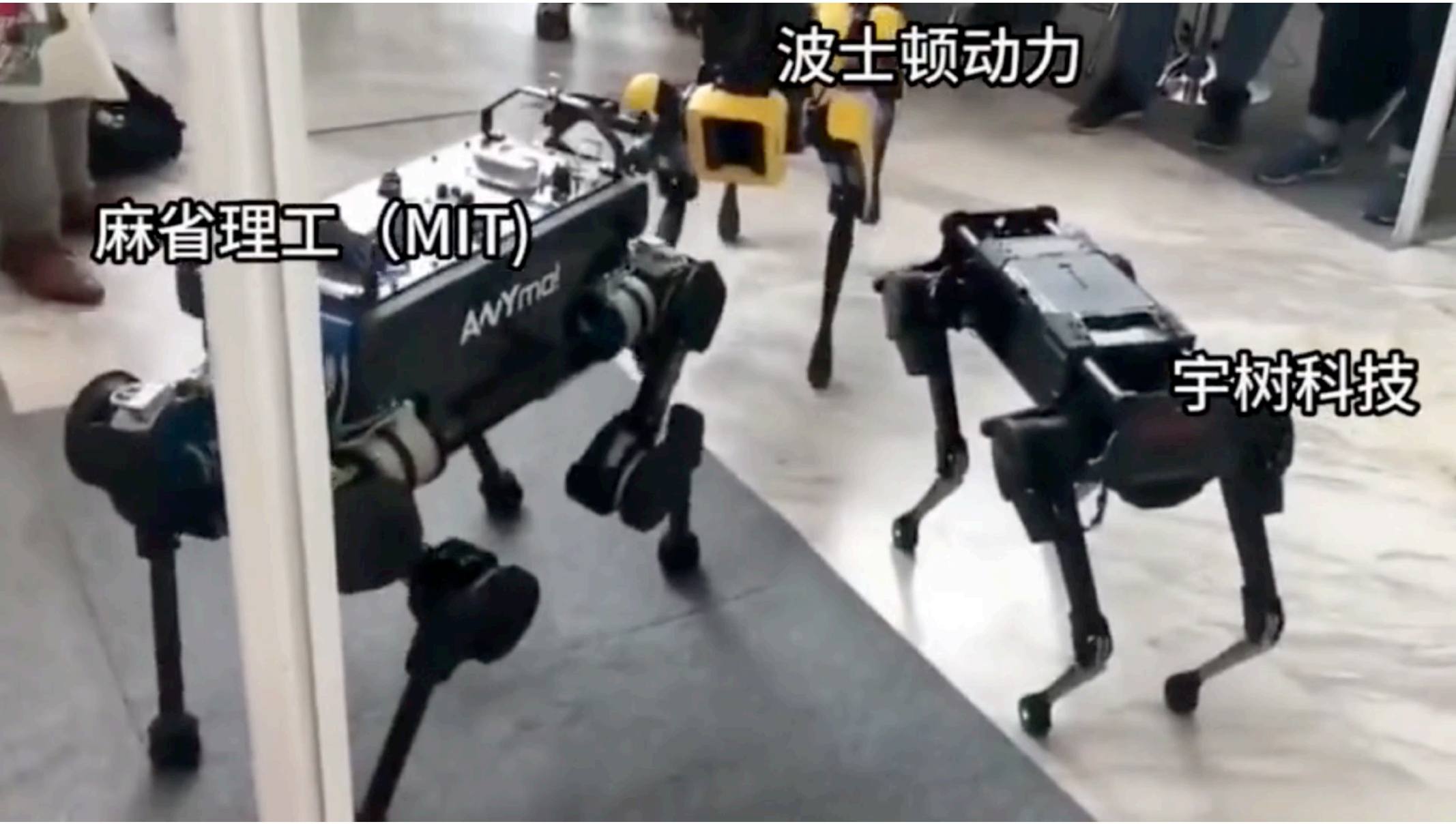
The intelligence of survival

Robotics



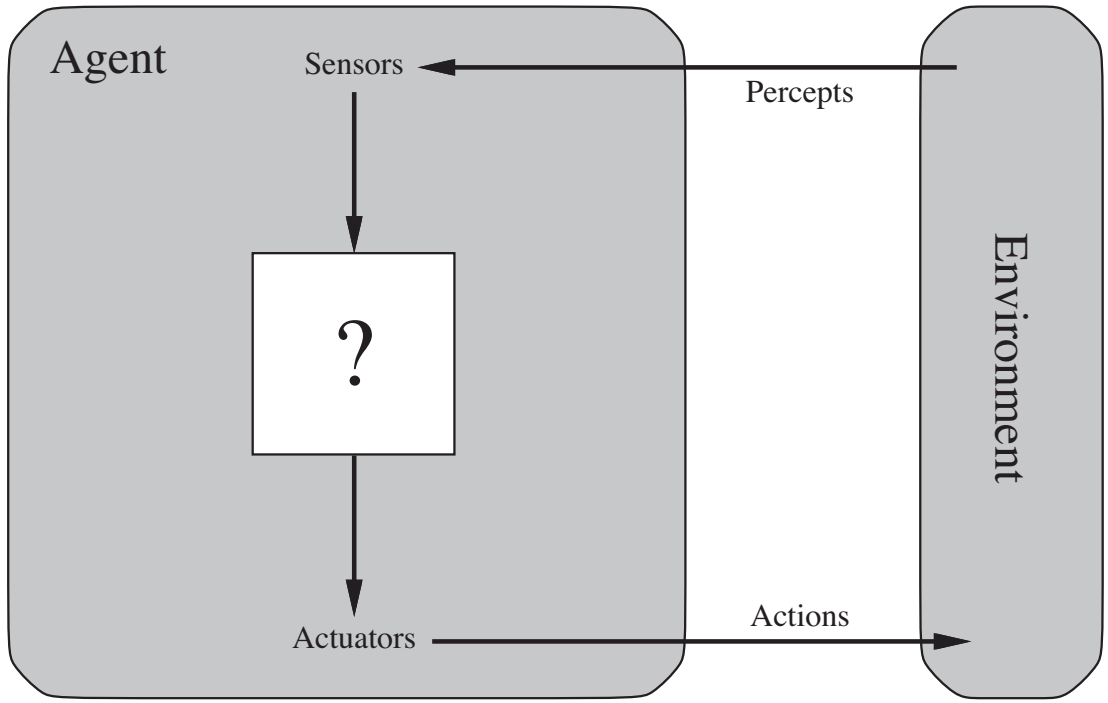
Robotics by RL

Comparison



Comparison with the real dog

Agent



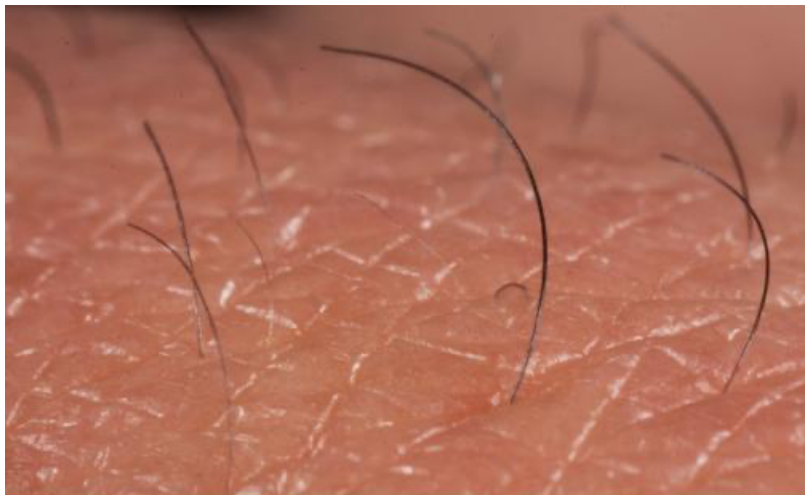
Agents include humans, robots, softbots, thermostats, etc.

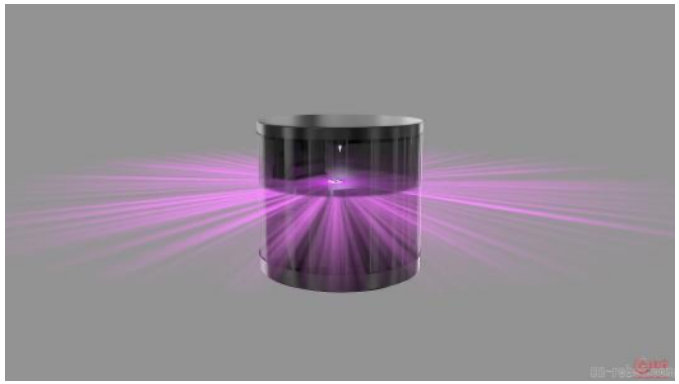
The agent function maps from percept histories to actions:

$$f : \mathcal{P}^* \rightarrow \mathcal{A}$$

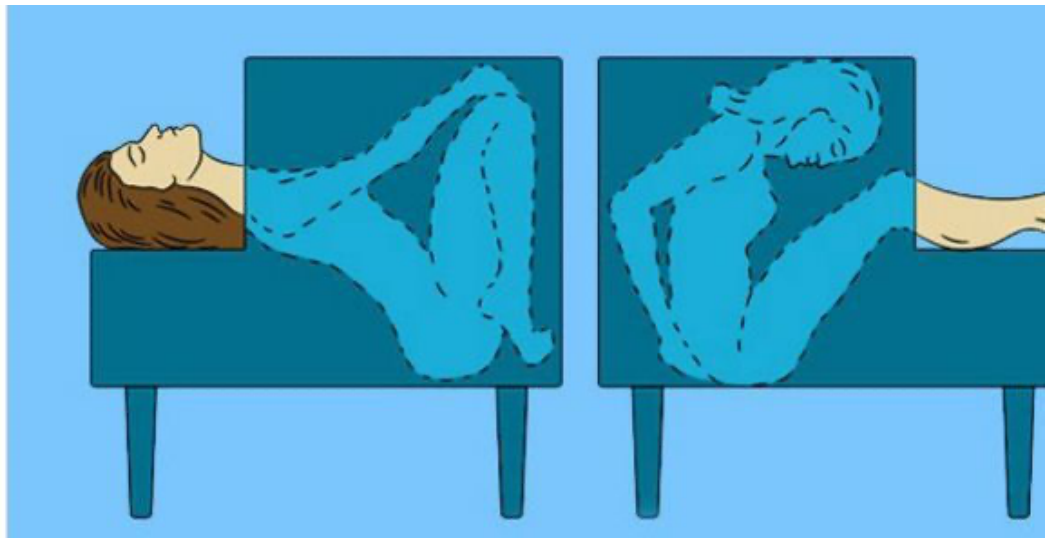
The agent program runs on the physical architecture to produce f

Sensors

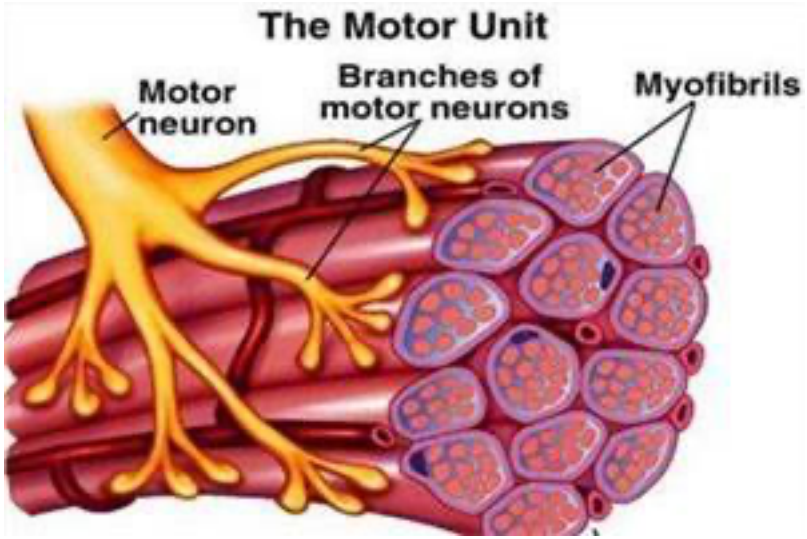


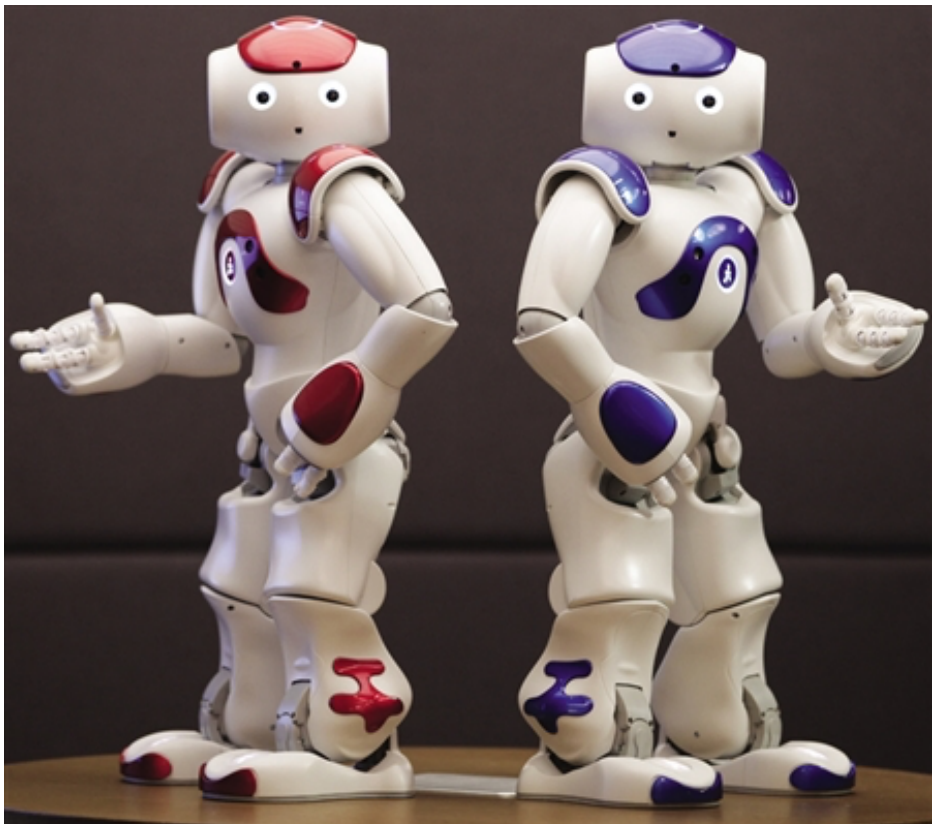


Observation vs state

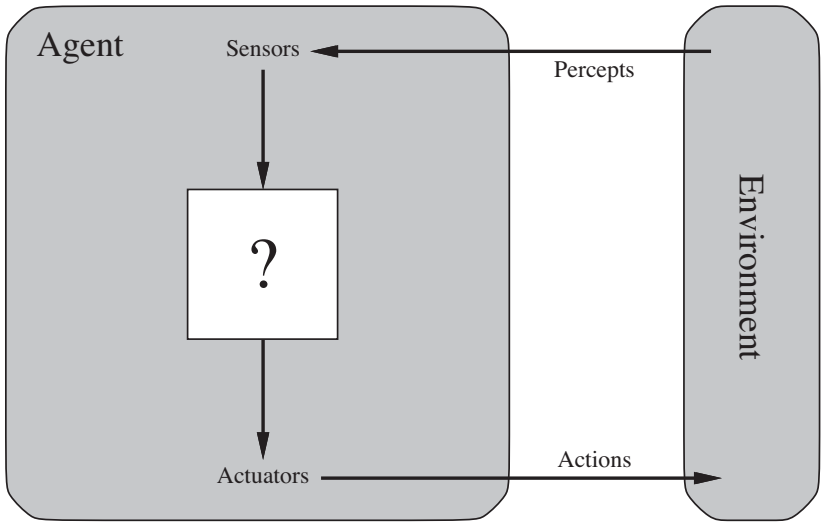


Actuators


The Motor Unit
Motor neuron
Branches of motor neurons
Myofibrils



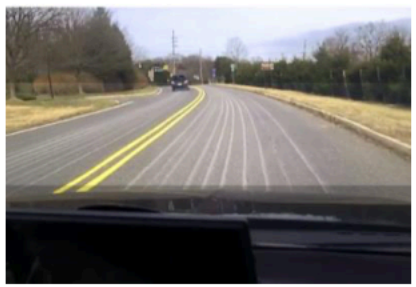
Function / Policy

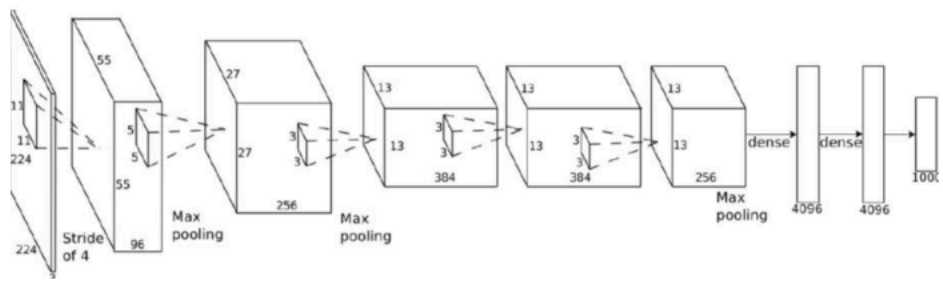


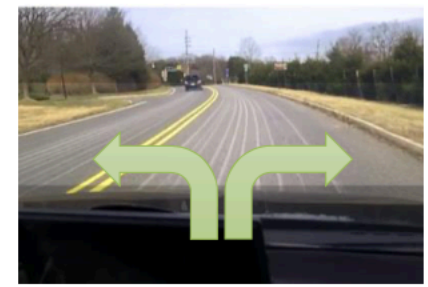
$$\pi(s) \rightarrow a$$

Policy: $\pi : S \times A \rightarrow \mathbb{R}, \quad \sum_{a \in A} \pi(a|s) = 1$

Policy (deterministic): $\pi : S \rightarrow A$







Environment

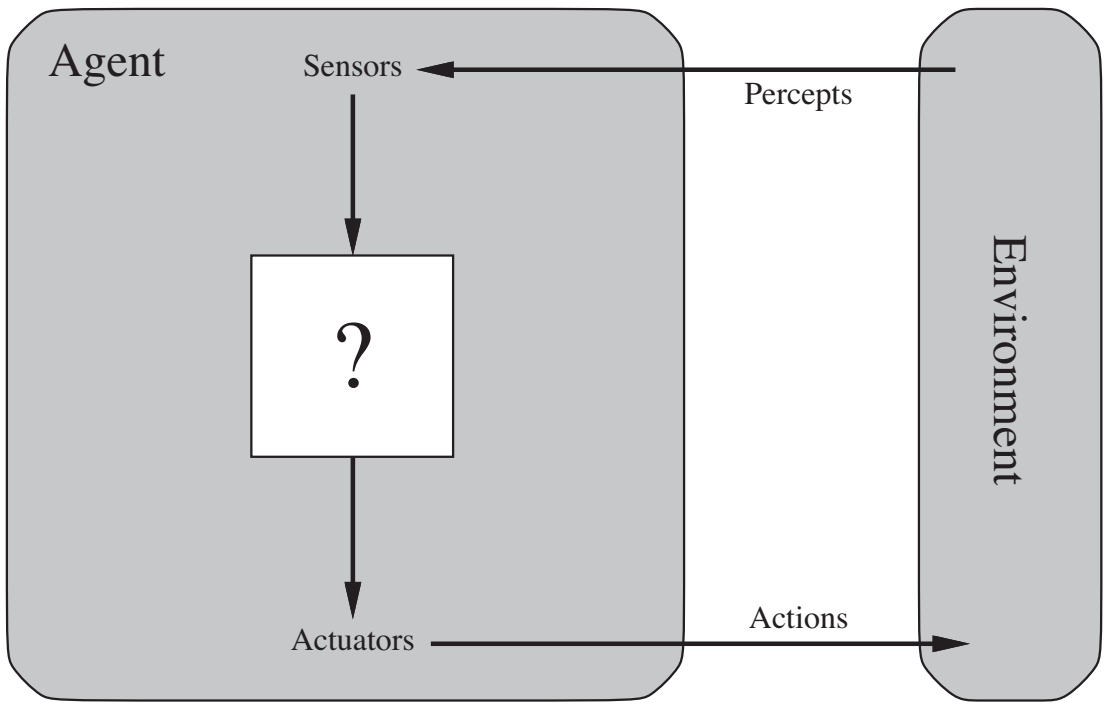
Environment

environment is influenced by the actions

$$s_t, a_t \rightarrow P \rightarrow s_{t+1}$$

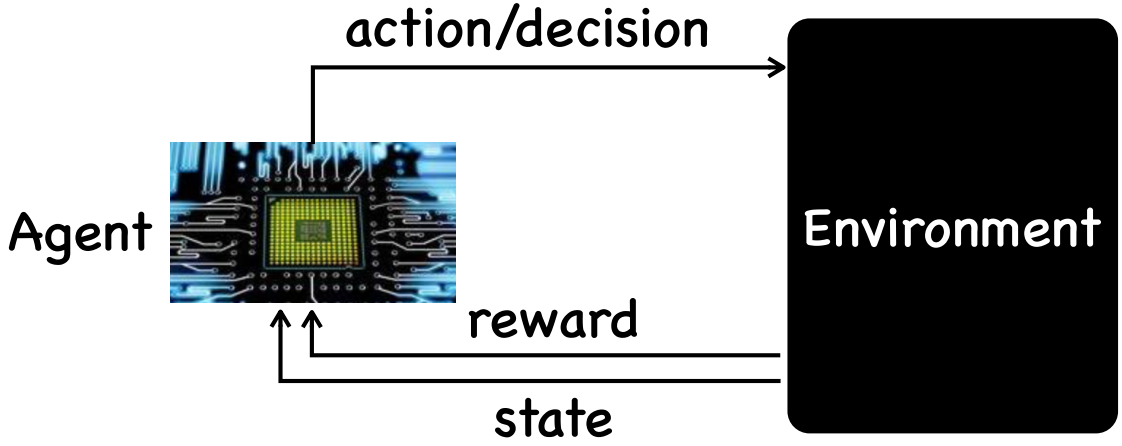
$$P(s_{t+1}|s_t, a_t)$$

Agent



The intelligence of survival

Reinforcement learning



Agent's goal: learn a policy to maximize long-term total reward

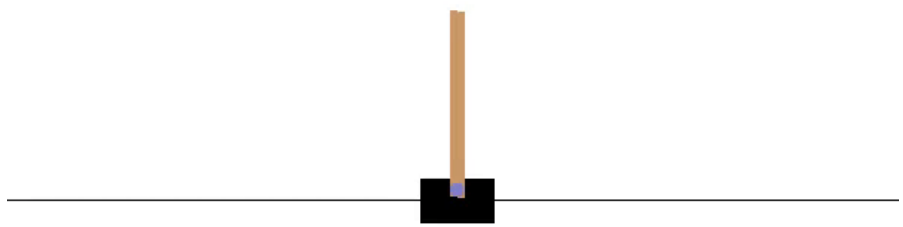
T-step: $\sum_{t=0}^{T} r_t$ average: $\frac{1}{T}\sum_{t=0}^{T} r_t$ discounted: $\sum_{t=0}^{\infty} \gamma^t r_t$

a way of programming agents

Cart Pole

States: (pole angle, angular velocity)
Actions: move left, move right
Rewards:

0 stands, -1 failed



Video games

States: video
Actions: gamepad buttons
Rewards:

game score

Game of Go

States: board

Actions: (x,y)

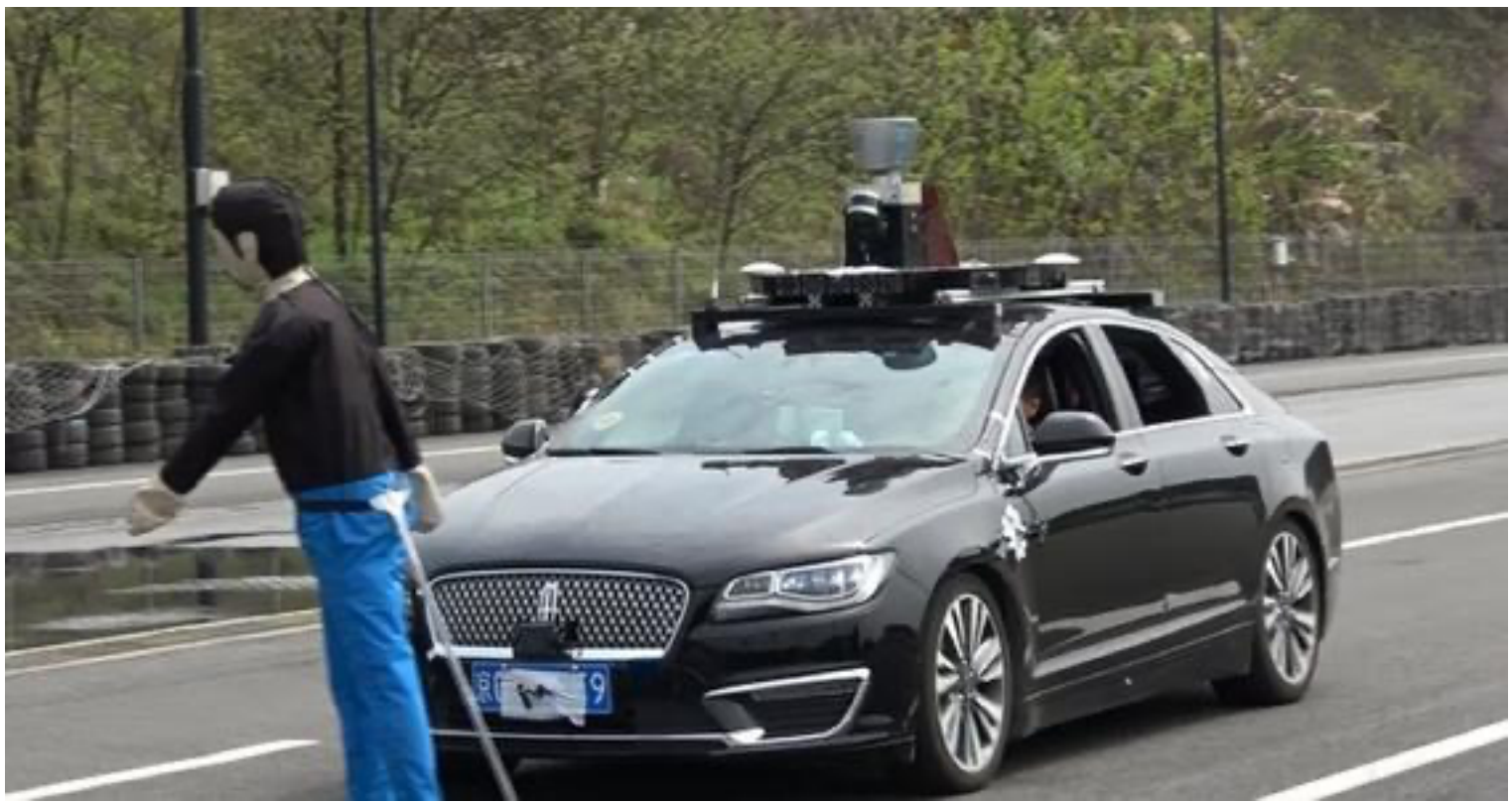
Rewards:

0 win, -1 lose

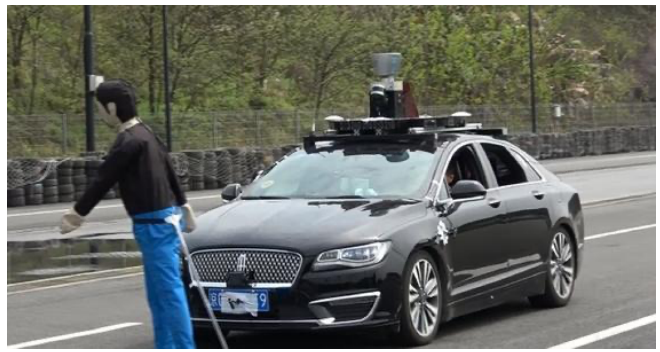
Self-driving car

States: surrounding
Actions: steering wheel, accelerator, brake
Rewards:

+100 fast and safe, -100 collisions, …



Go vs self-driving car



- Known environment vs Unknown environment
- Discrete vs Continuous states/actions
- One goal vs many goals
- Simple reward vs rewards to be detected
- Safe vs killing

Notations

Richard Bellman

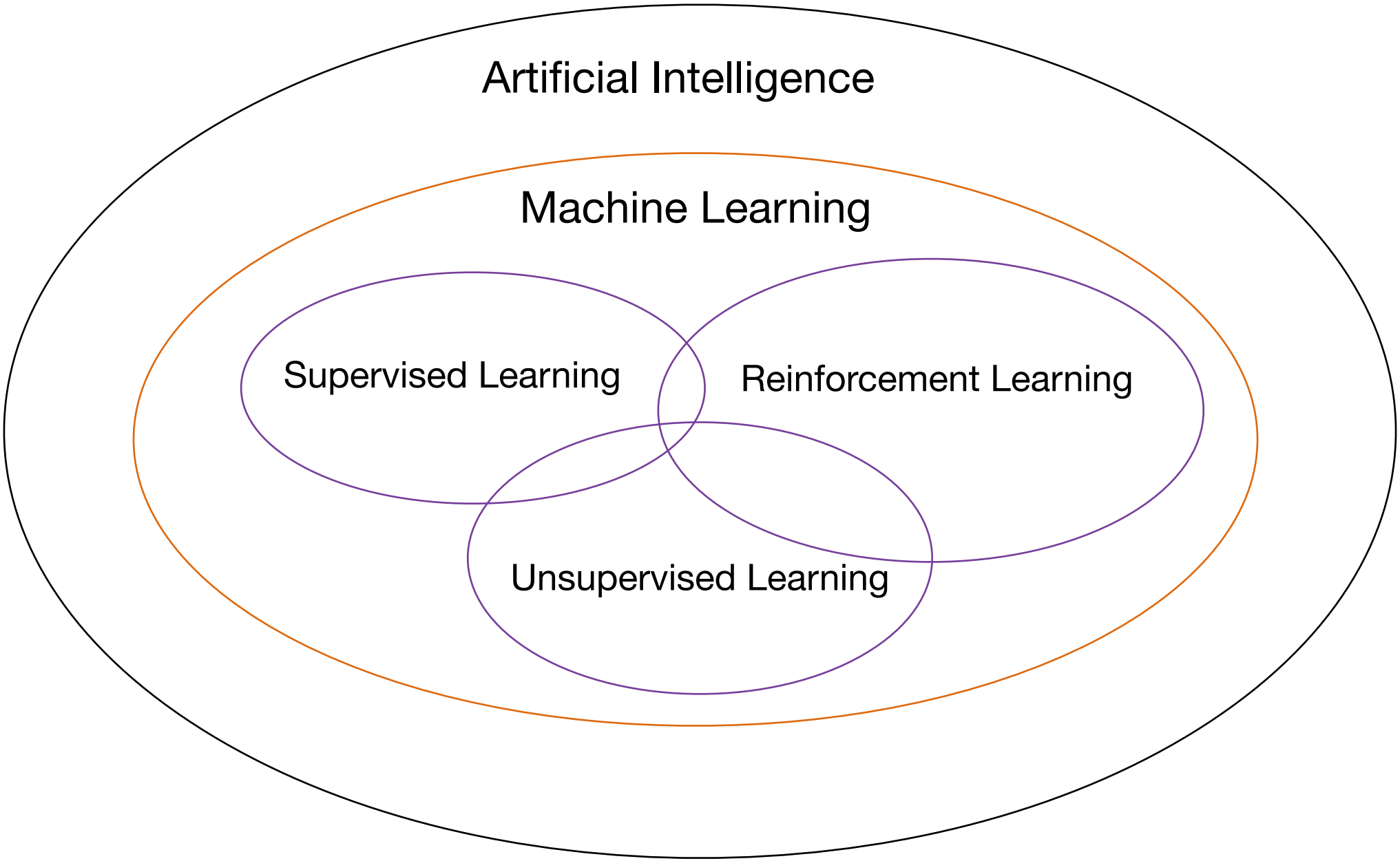
Lev Pontryagin

actions a_t
states s_t
rewards r_t
dynamics $p(s_{t+1} \mid s_t, a_t)$
observations o_t

actions u_t
states x_t
costs $c(x_t, u_t)$
dynamics $p(x_{t+1} \mid x_t, u_t)$

Diagram from Sergey Levine

Machine Learning



Difference between RL and SL?

both learn a model …

Prediction:

what is this?

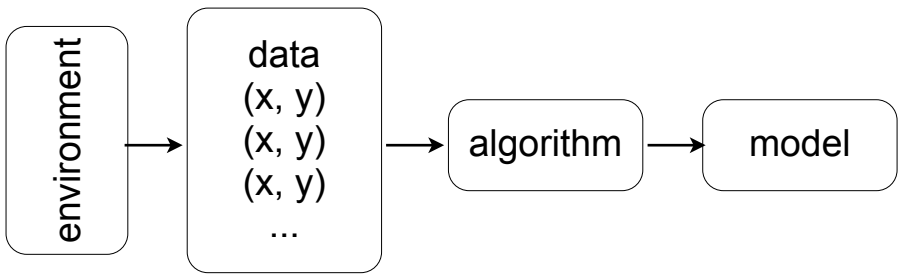
this is a tumor!

Decision-making:

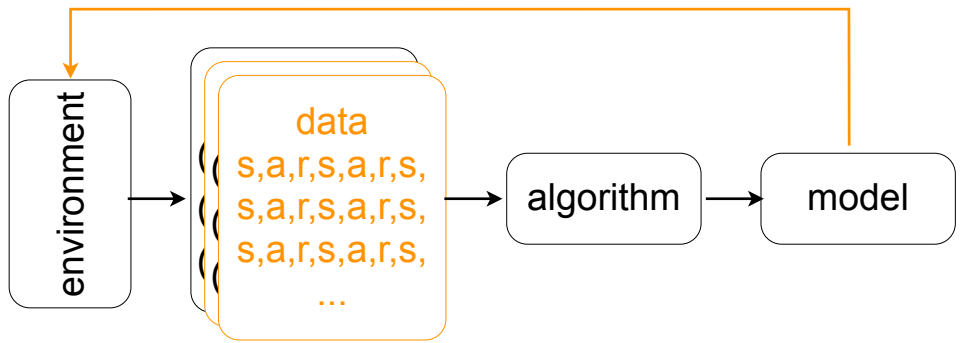
how to cue?

turn-left, cut, …, tumor-removed

Supervised learning process



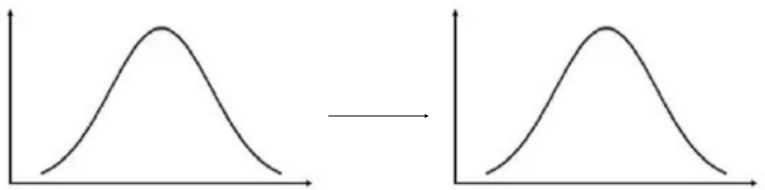
Reinforcement learning process



Difference between RL and SL?

Supervised learning objective

$$\arg\min_{\theta} E_{x\sim\mathcal{D}}\ \text{loss}(f_\theta(x), y(x))$$



training distribution test distribution

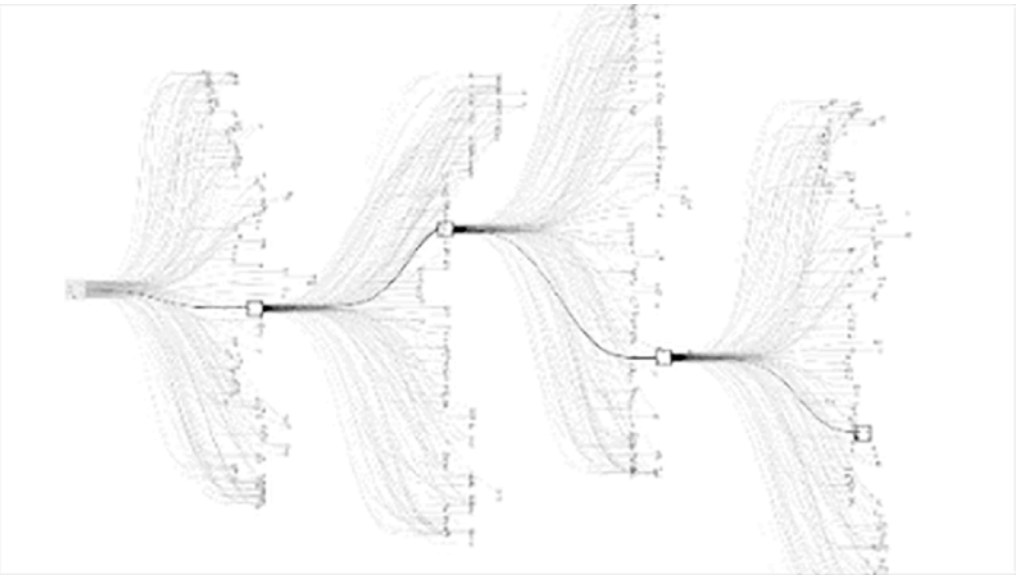
train/test data are sampled i.i.d.

future is the same as the past

train on past data

Reinforcement learning objective

$$\arg\max_{\theta} E_{s\sim\mathcal{D}^{\pi_\theta}}\ \text{reward}(s, \pi_\theta(s))$$



future is to be chosen

train by trial-and-error