Lecture 1: Introduction to RL

王志　南京大学

2025-02-17

What we will cover

- Preliminaries of machine learning
  - Supervised learning: Regression, classification
  - Unsupervised learning: clustering, dimensionality reduction
- Finite Markov decision process (MDP)
- Tabular RL algorithms
  - Dynamic programming
  - Monte-Carlo methods
  - Temporal-difference learning
- Deep RL algorithms
  - Policy gradients
  - Actor-critics
  - Value function-based methods
- State-of-the-arts
  - Offline RL
  - Meta-RL, multi-task RL
  - LLMs and RL

Pre-requisite

- Probability theory (概率论)
  - Probability distribution (概率分布), random variable (随机变量), expectation (期望), variance (方差)
  - probability density (概率密度) , conditional probability (条件概率), Bayes rule (贝叶斯公式)
- Linear algebra
  - Matrix multiplication (矩阵乘法), eigen-vector (特征向量)
  - Inverse matrix (逆矩阵), Hessian matrix (海赛矩阵)
- Basic programming
  - Python

Grading

- Homework: 40%
  - Policy gradient (20%)
  - DDPG (20%)
- Final project: 60%
  - Make a presentation (with slides in English) of one research paper selected from a given list
- Teaching assistant: Zican Hu (胡紫灿)
  - All homeworks and final projects should be submitted to huzican0419@gmail.com

Useful links and references

- Course website
  - Heyuanmingong.github.io/teaching.html
- Books
  - Richard S. Sutto and Andrew G. Barto, *Reinforcement Learning: An Introduction*, 2nd Edition.
  - http://www.incompleteideas.net/book/the-book-2nd.html
- Famous open classes
  - CS 295 at UC Berkeley, *Deep Reinforcement Learning* (http://rail.eecs.berkeley.edu/deeprlcourse/)
  - CS 234 at Stanford, *Reinforcement Learning* (http://web.stanford.edu/class/cs234/)

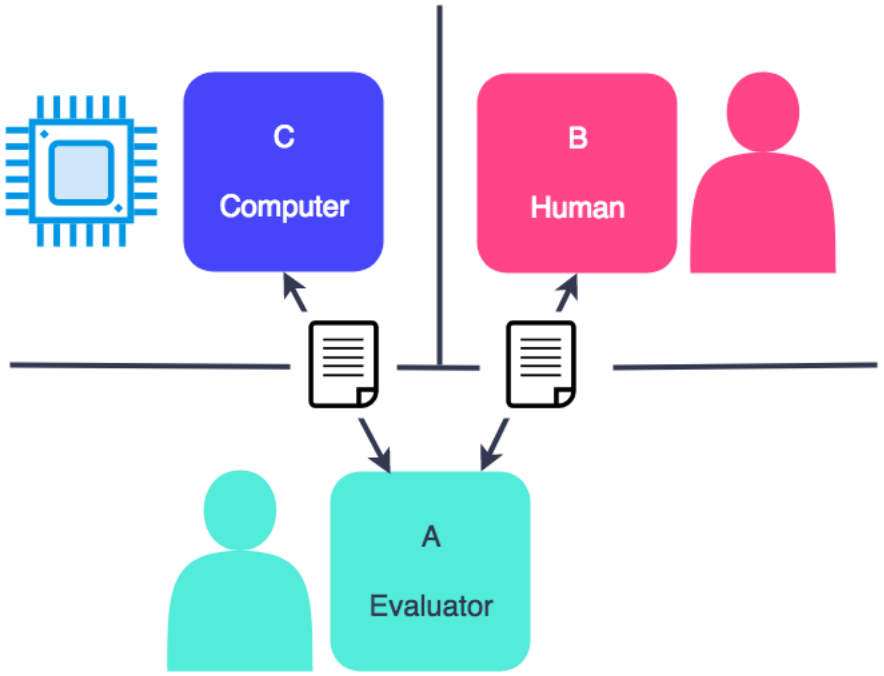
Content

Artificial Intelligence (AI, 人工智能)

- The simulation of human intelligence processes by machines, especially computer systems
- The ability to learn, understand, and think about things in a logical way
- Artificial intelligence is a type of technology that mimics the human thought, empowering machines to act on their own and to perform functions similar to human intelligence, such as the ability to perceive, learn, reason, and act

Birth of AI

Turing test (图灵测试)



Alan Turing
(1912~1954)

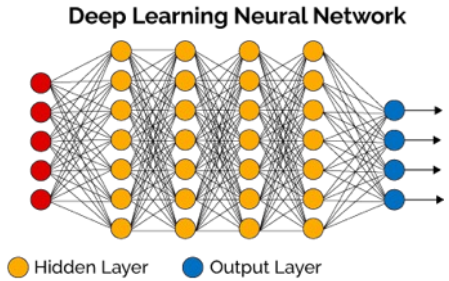
Beyond Turing: AI may exceed human intelligence

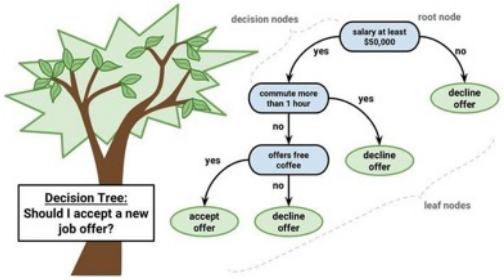
Reasons for Rise of AI

- **Exponential growth of computing power**




- **Rapid development of AI algorithms**



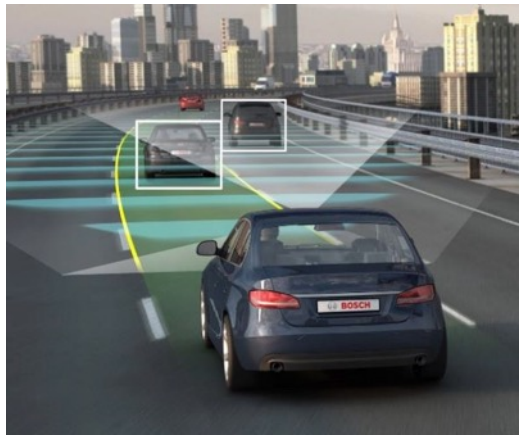


- **Massive data collection from connection of things**

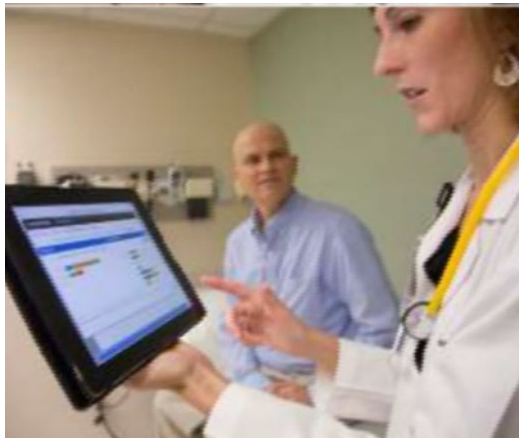
Applications of AI

Natural language processing

Autonomous driving



Medical diagnosis



Robotics

E-commerce

Public security

Content

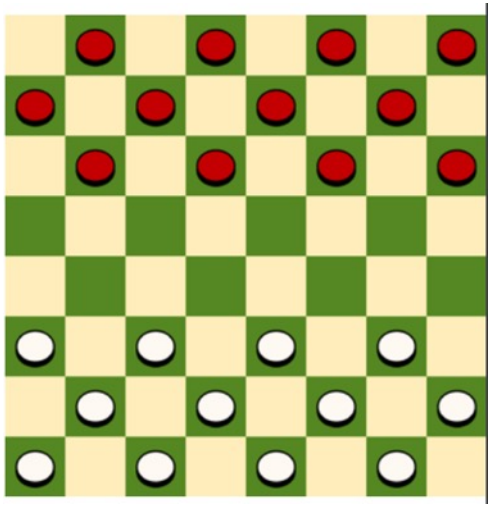
Machine Learning (机器学习)

Arthur Samuel (1959):

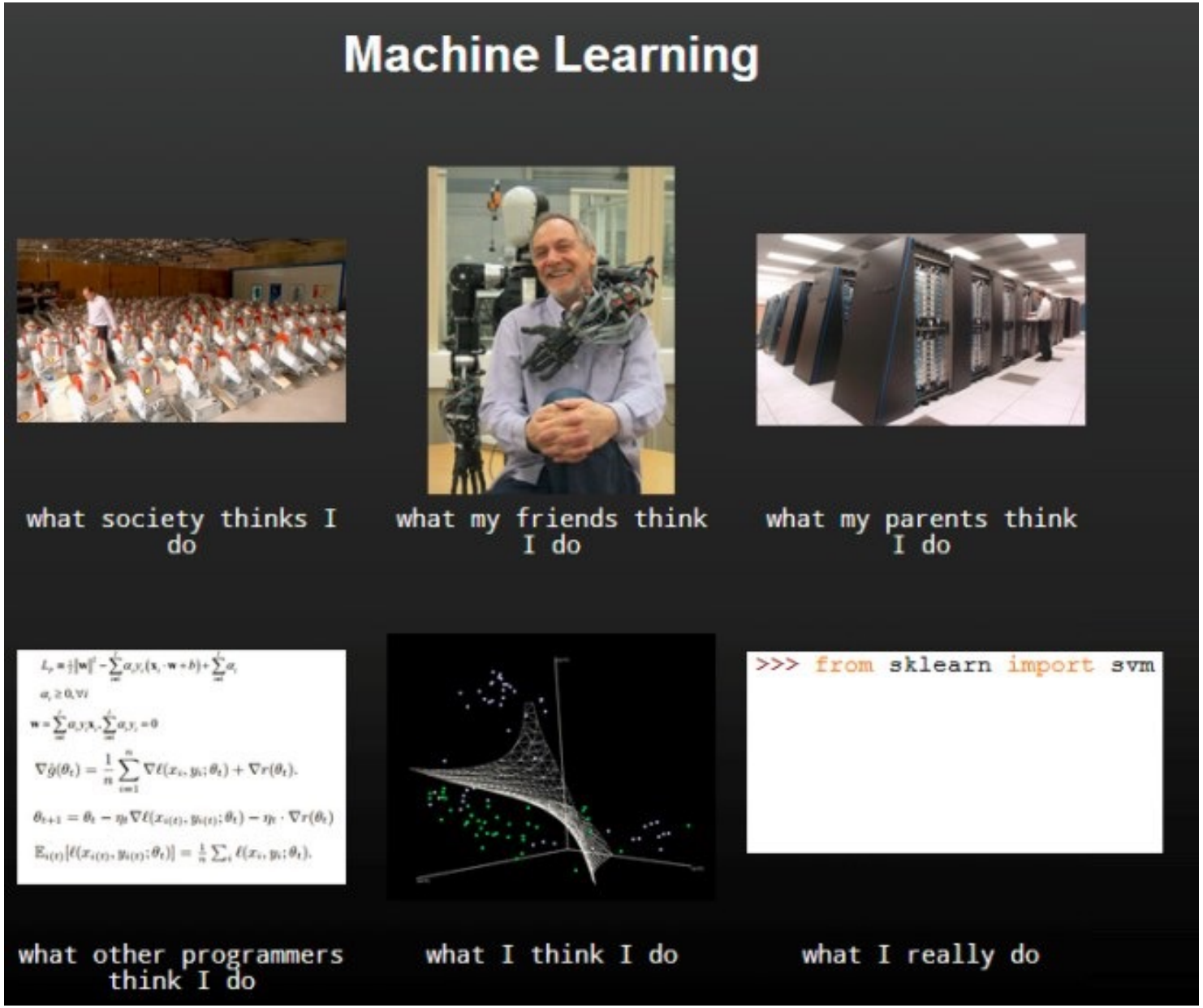
Machine Learning is the field of study that gives the computer the ability to learn without being explicitly programmed.

A. L. Samuel*

Some Studies in Machine Learning
Using the Game of Checkers. II—Recent Progress



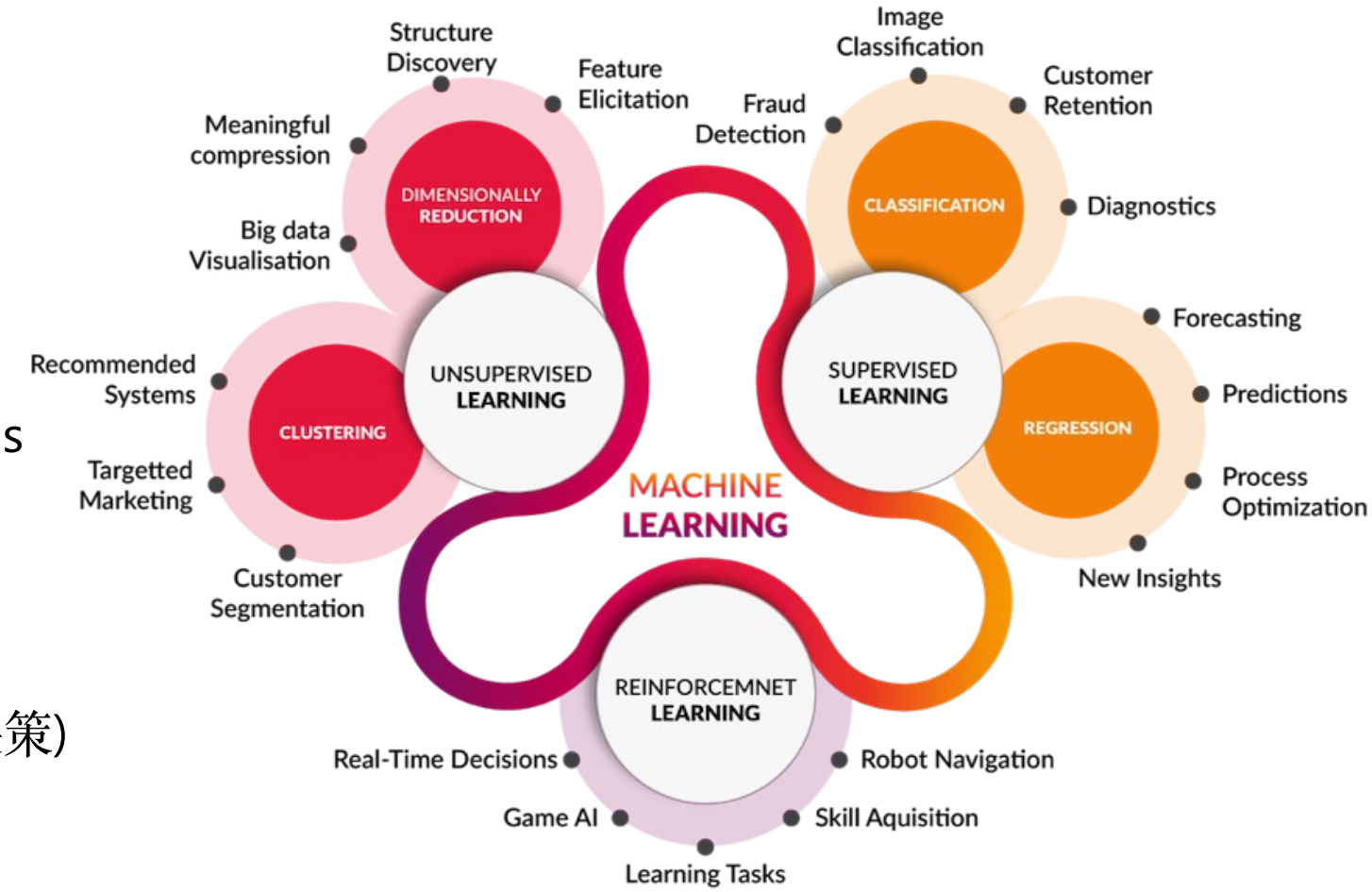
Machine Learning



Machine Learning

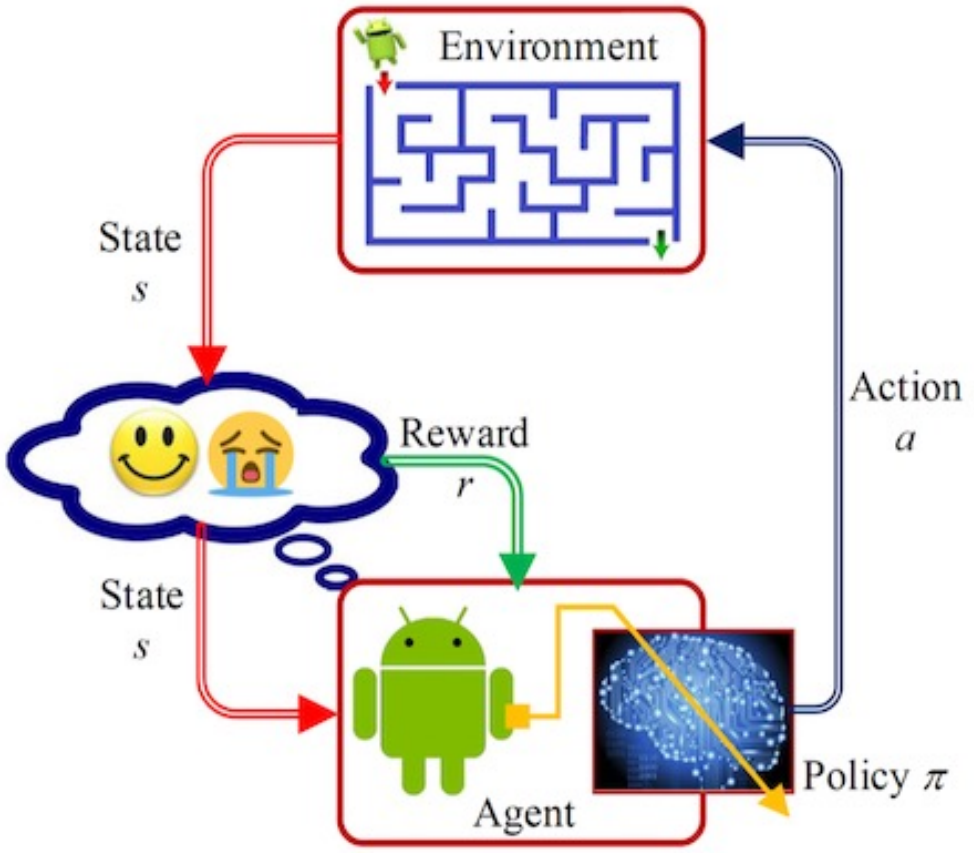
Taxonomy of Machine Learning

- Supervised learning (监督)
  - A mapping from input to label
  - Regression, classification
- Unsupervised learning (非监督)
  - Learn hidden patterns without labels
  - Clustering, dimension reduction
- Reinforcement learning (强化)
  - sequential decision-making (序贯决策)



Reinforcement Learning (RL, 强化学习)

- Mathematical formalism for learning-based decision-making
- Reinforce behaviors through rewards received from interacting with the world in a trial-and-error way
- Approach for sequential decision-making and control from experience

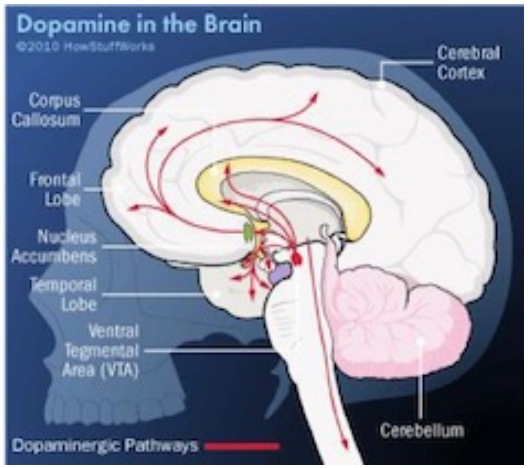


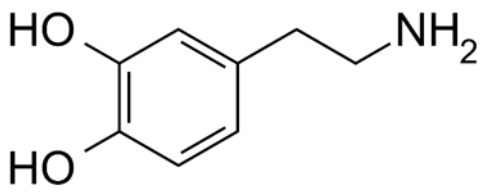
Inspiration of RL

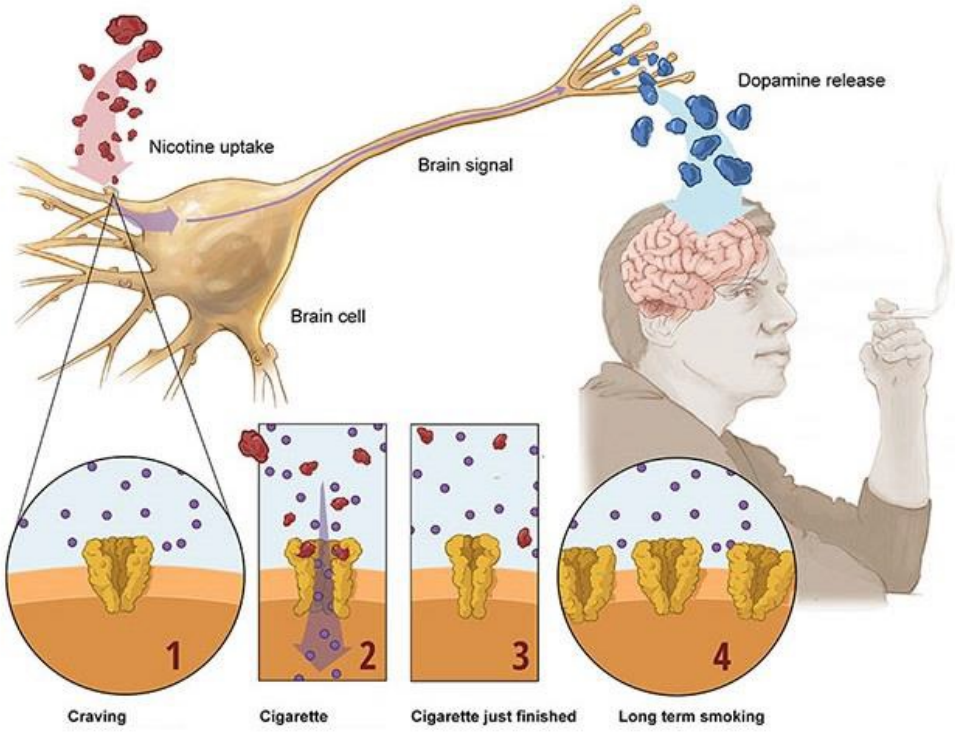
- A biologically inspired training method
  - Originate from mimicking learning behaviors of animals
  - Repeat behaviors that are rewarded, avoid behaviors that are punished

Inspiration of RL

- Animal behavior is rewarded by brain dopamine (多巴胺)
  - Dopamine is neurotransmitter (神经递质)
  - Responsible for reward-motivation, motor control, and arousal





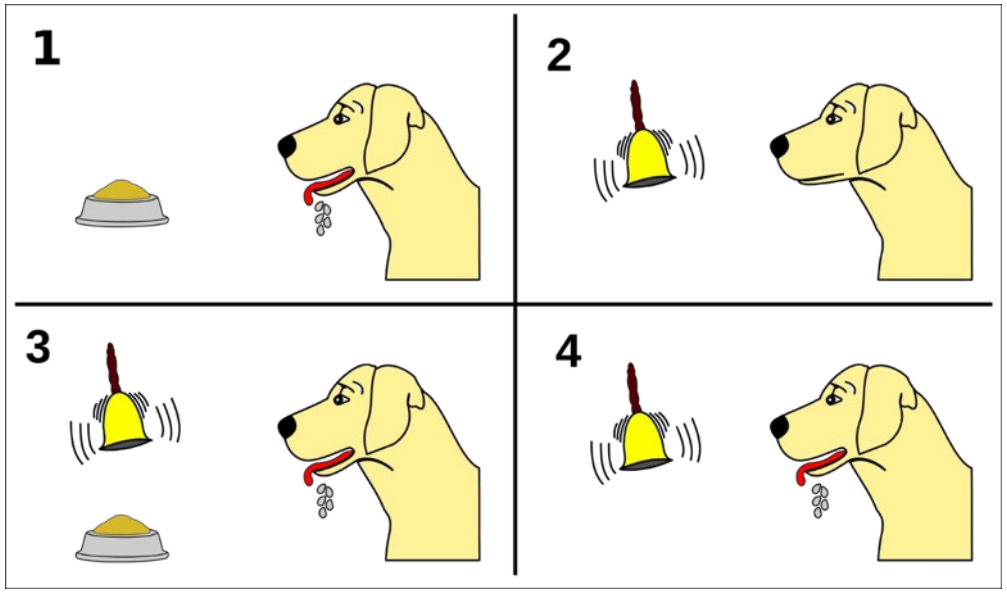


Inspiration of RL

- Pavlov's dog experiment
  - Ivan Pavlov, Russian physiologist, 1901
  - Humans and animals learn through rewards

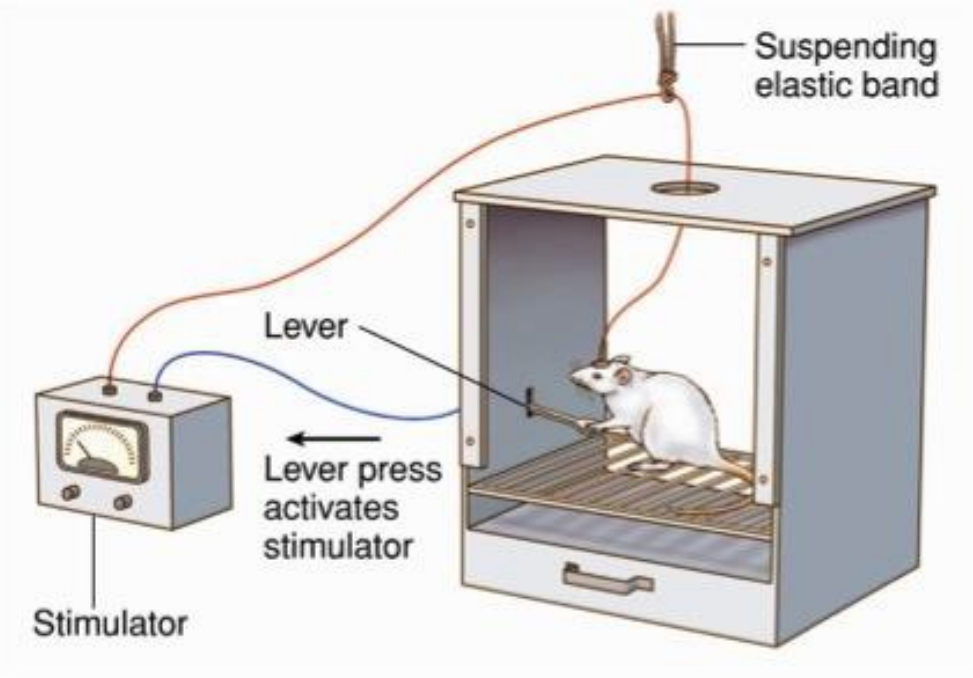
Training: Bell ➔ Food

Testing: Bell ➔ Salivate (分泌唾液)



Inspiration of RL

- Brain stimulation reward experiment
  - J Olds & P Milner, McGill University, 1954
  - Rats would continually press a lever in return for receiving a brief pulse of electrical stimulation in a particular region of rat's brain

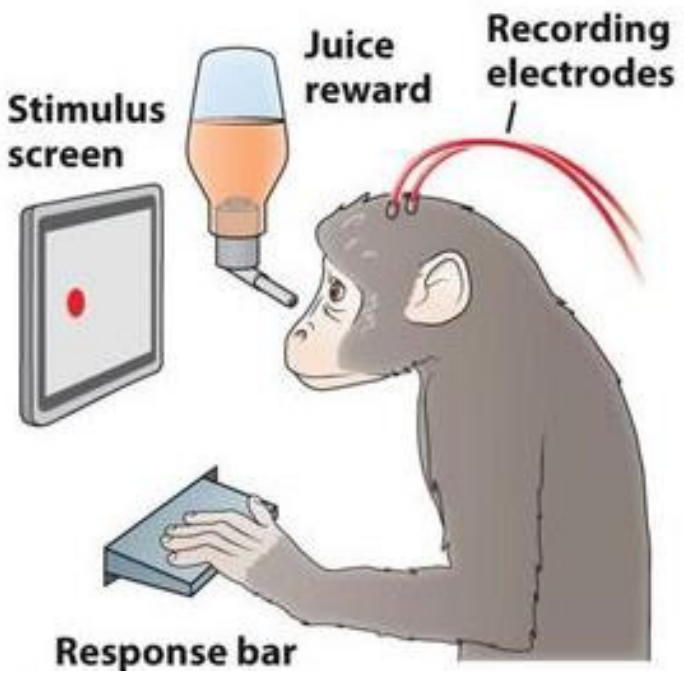


Result: Existence of a reward center in the brain

Inspiration of RL

- Dopamine releasing mechanism experiment
  - Schultz et al., Cambridge University, 1997

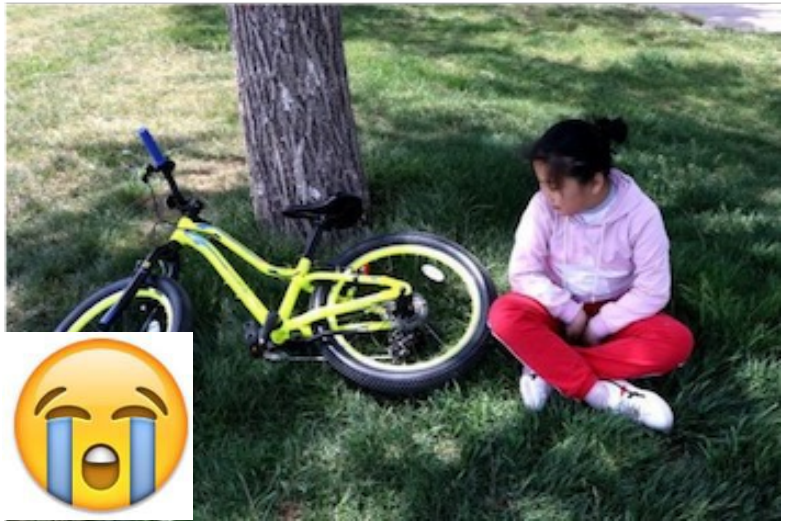
Num	Predicted reward	Actual reward	Error	Dopamine release
1	No	Yes	+	↑
2	Yes	Yes	0	→
3	Yes	No	-	↓



- Result: Dopamine release comes from the difference or error between actual reward and predicted reward

Inspiration of RL

- How a child learns to ride a bicycle
  - Initially, an adult will show how to ride (imitation learning, 模仿学习)
  - Then, the child should try it by herself from the reward or penalty
  - By learning from trial and errors, she finally masters it



Markov Decision Process (MDP, 马尔科夫决策过程)

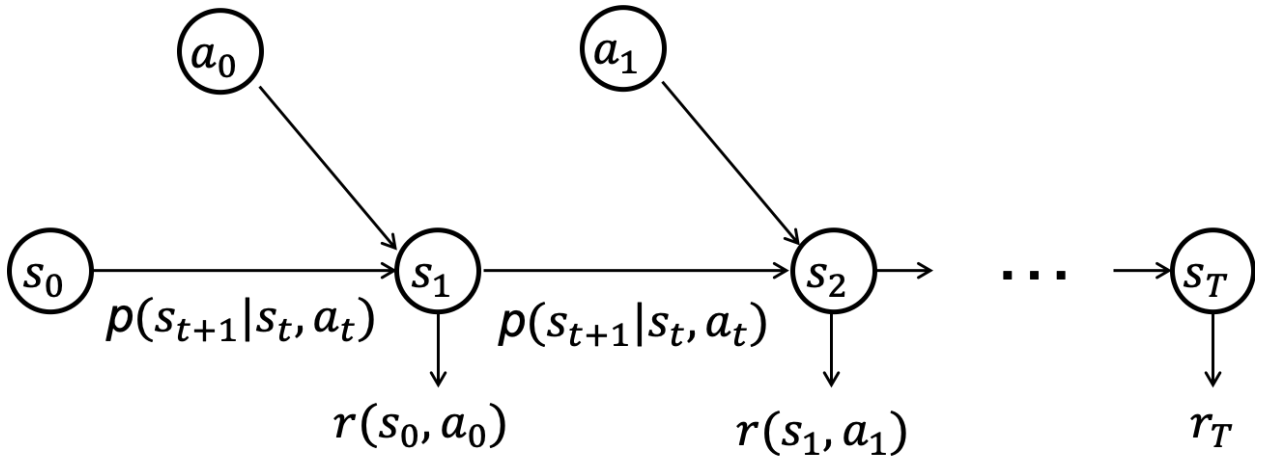
$$\boldsymbol{M} = \langle \boldsymbol{S}, \boldsymbol{A}, \boldsymbol{T}, \boldsymbol{R} \rangle$$

S: State space — State $s \in S$ (discrete/continuous)

A: Action space — Action $a \in A$ (discrete/continuous)

T: Transition operator — $T_{i,j,k} = p(s_{t+1} = j | s_t = i, a_t = k)$

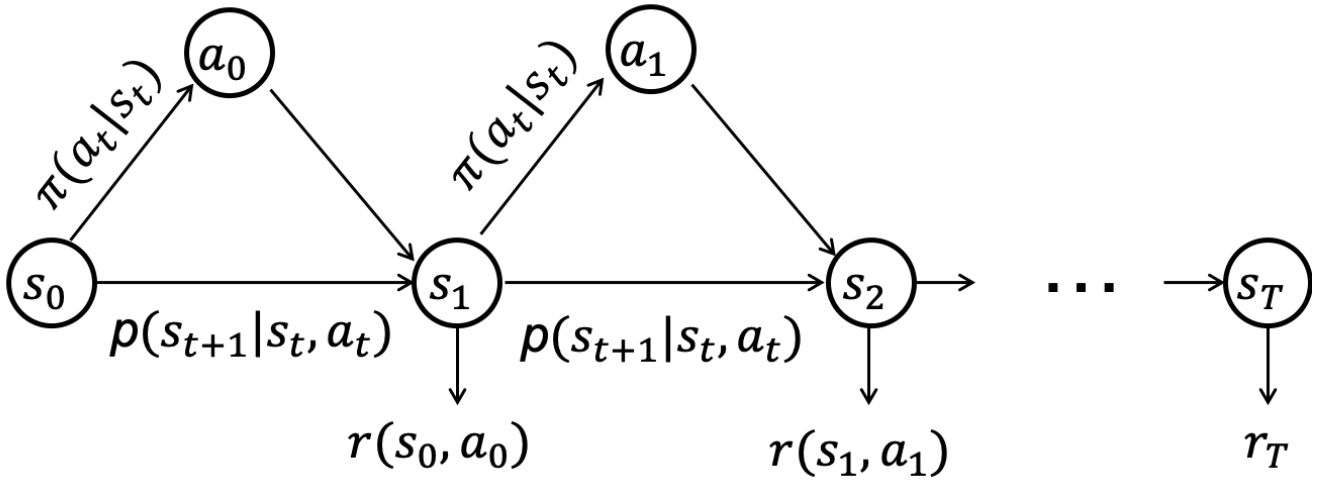
R: Reward function — $R_{i,j,k} = r(s_{t+1} = j | s_t = i, a_t = k)$



The goal of RL

- Find optimal policies to maximize cumulative reward
  - In a trial-and-error manner
  - A general optimization framework for sequential decision-making

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_t r(s_t, a_t)\right]$$

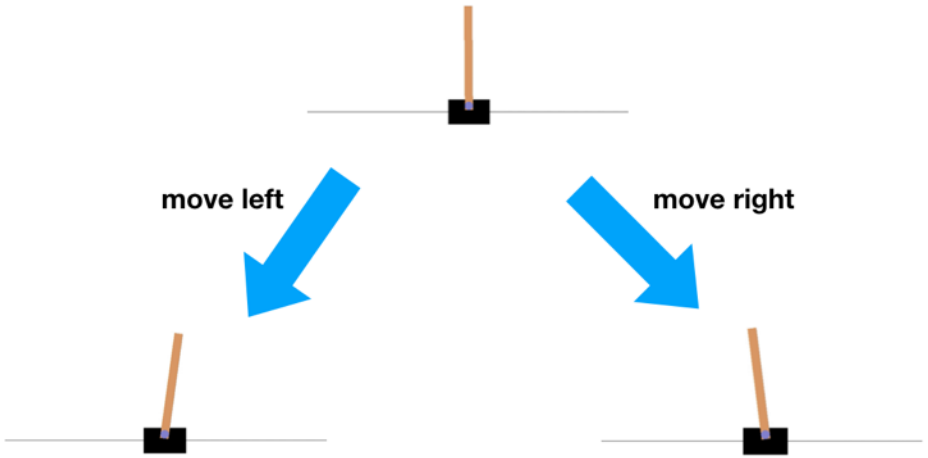


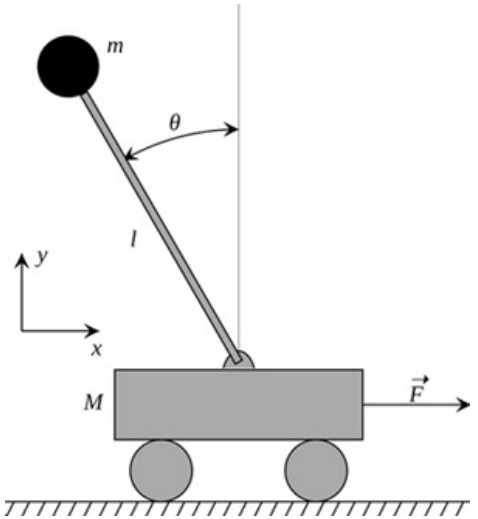
How is RL different from other machine learning topics?

- Standard supervised learning:
  - given $D = \sum_i (x_i, y_i)$
  - learn to predict y from x, $f(x) \approx y$
- Usually assumes:
  - Independent and identically distributed (i.i.d., 独立同分布)
  - Known ground truth outputs in training

- RL
  - Data is not i.i.d., previous outputs influence future inputs
  - Ground truth answer is not known, only known if we succeed or failed (more generally, we know the reward)

Example: Inverted pendulum (倒立摆)

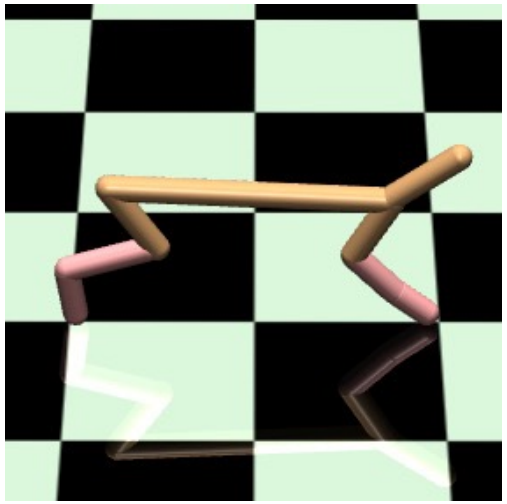




- x: position along the x-axis
- θ: angle of the pendulum

- State: $(x, \dot{x}, \theta, \dot{\theta})$
- Action: $\{-1, 0, +1\}$
- Reward: +1 if stand up, -0.01 otherwise

Example: Robot locomotion (机器人运动学)



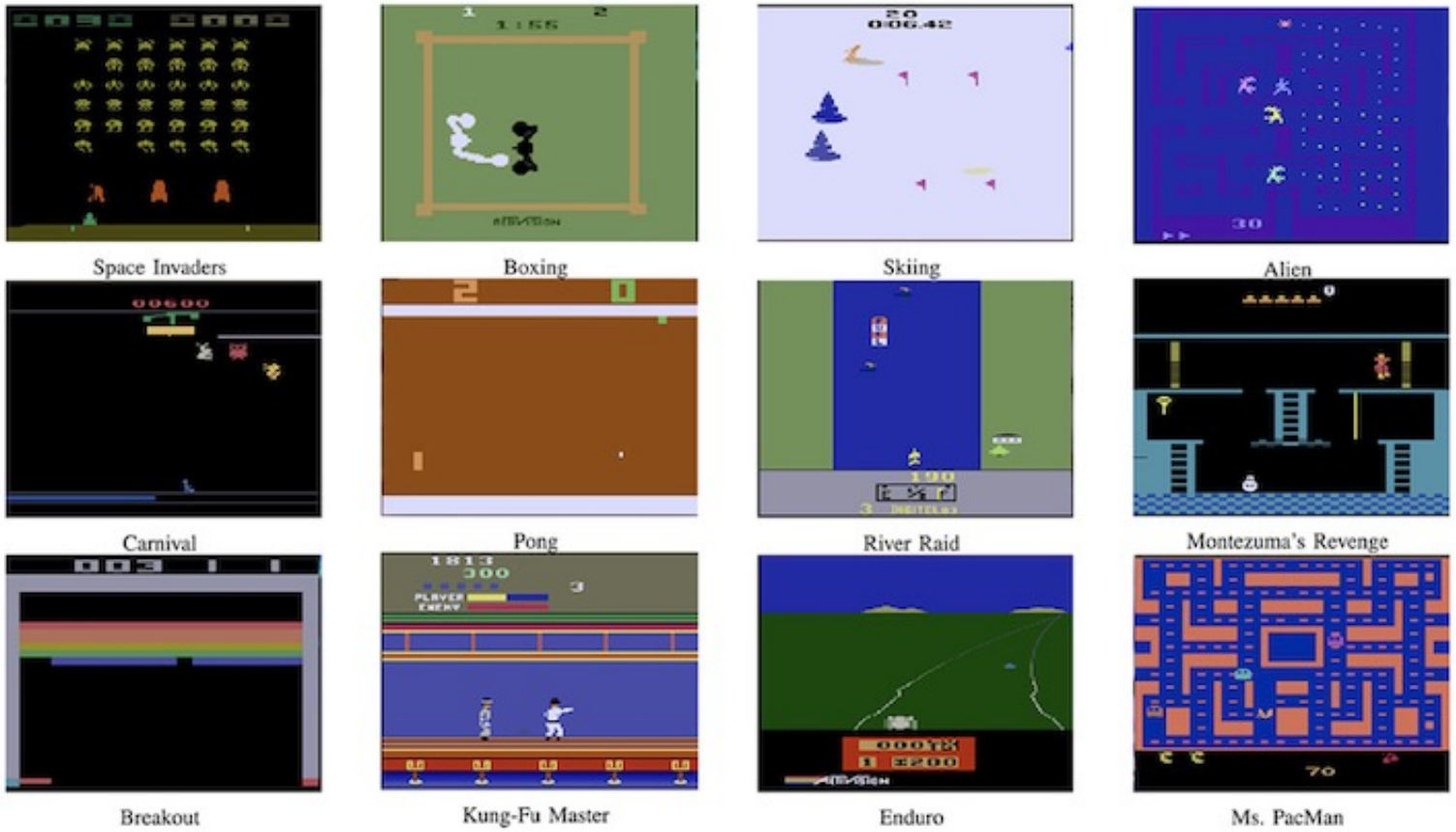
Half Cheetah

Ant

Humanoid

- Make the robots run forward, or navigation
  - Observations/states: positions, velocities, angular velocities
  - Actions: torques at the joint
  - Rewards: velocities, goal

Example: Atari games



- Observations/states: raw images
- Actions: control signals
- Rewards: win/lose

In general...

States: sight, smell
Actions: muscle contractions
Rewards: food

States: camera images
Actions: motor torque
Rewards: task success measure

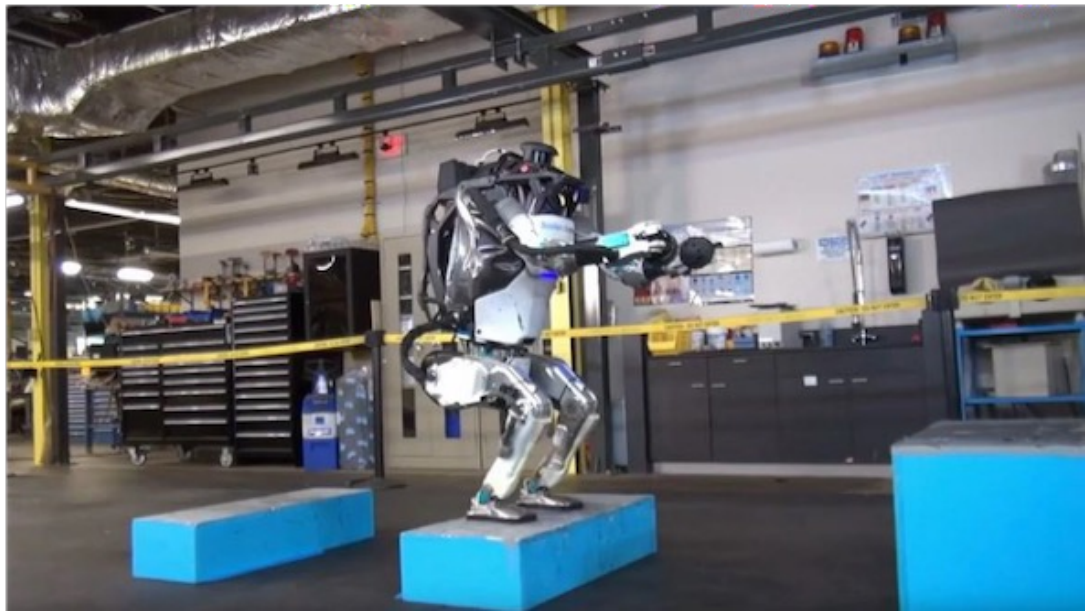
States: inventory (库存) level
Actions: what to purchase
Rewards: profit

Content

1. Artificial Intelligence (AI)
2. Reinforcement Learning (RL)
3. Why should we care about (deep) RL?
4. How to build intelligent machines?
5. Beyond learning from reward

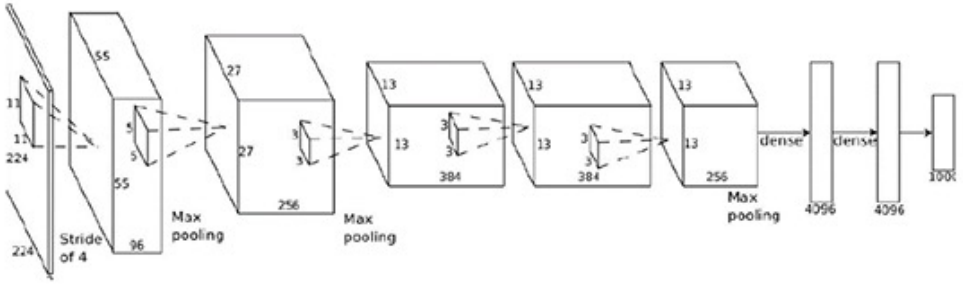
Intelligent machines

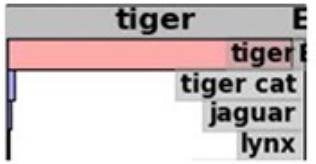
Intelligent machines must able to adapt

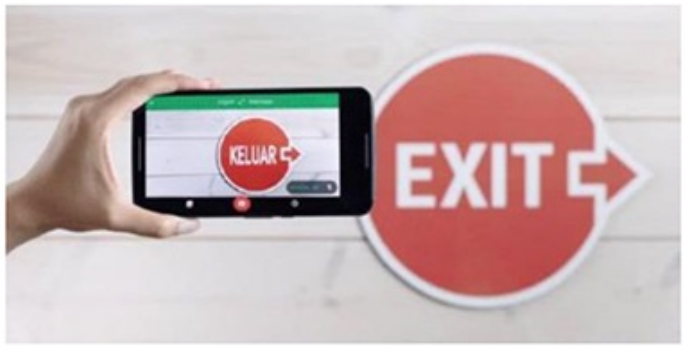


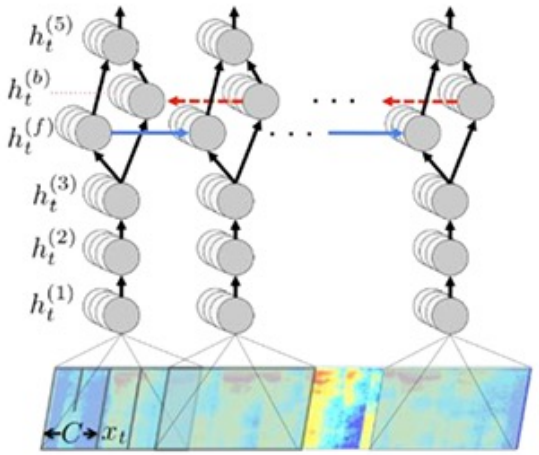
Deep learning

- Deep learning helps us handle unstructured environments

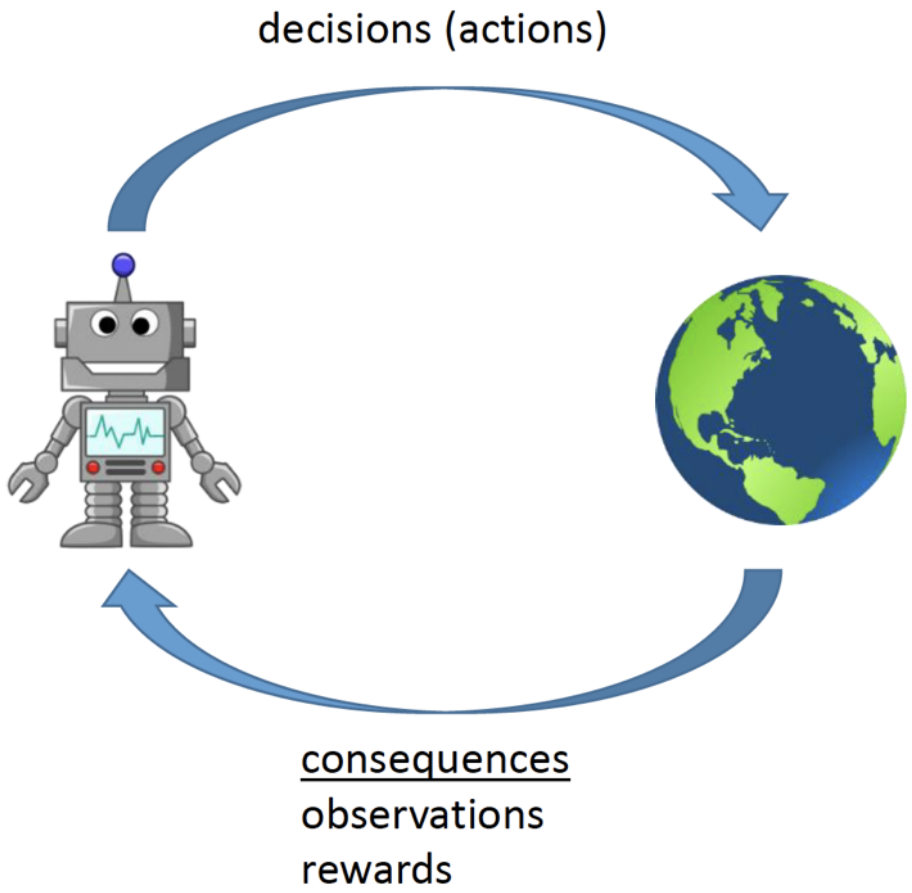


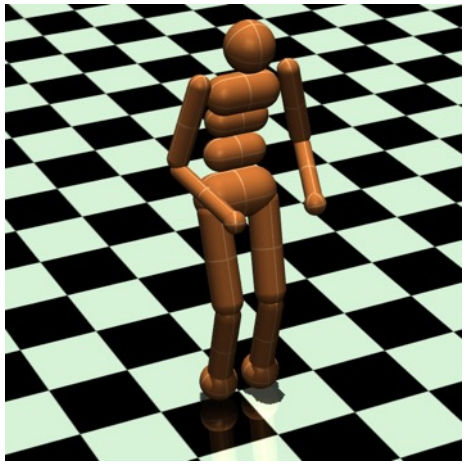


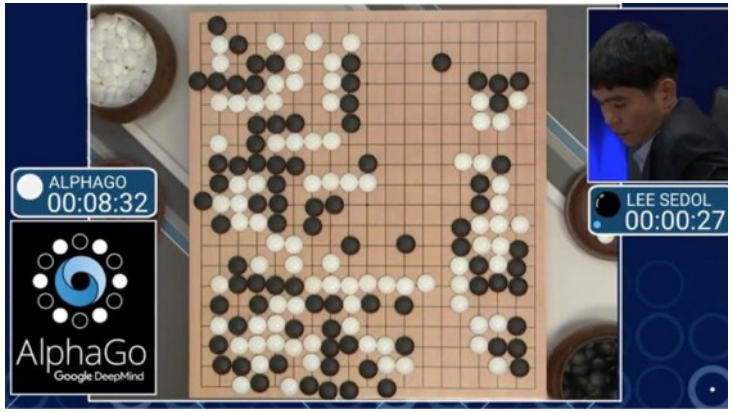




- RL provides a formalism for behavior/decision-making

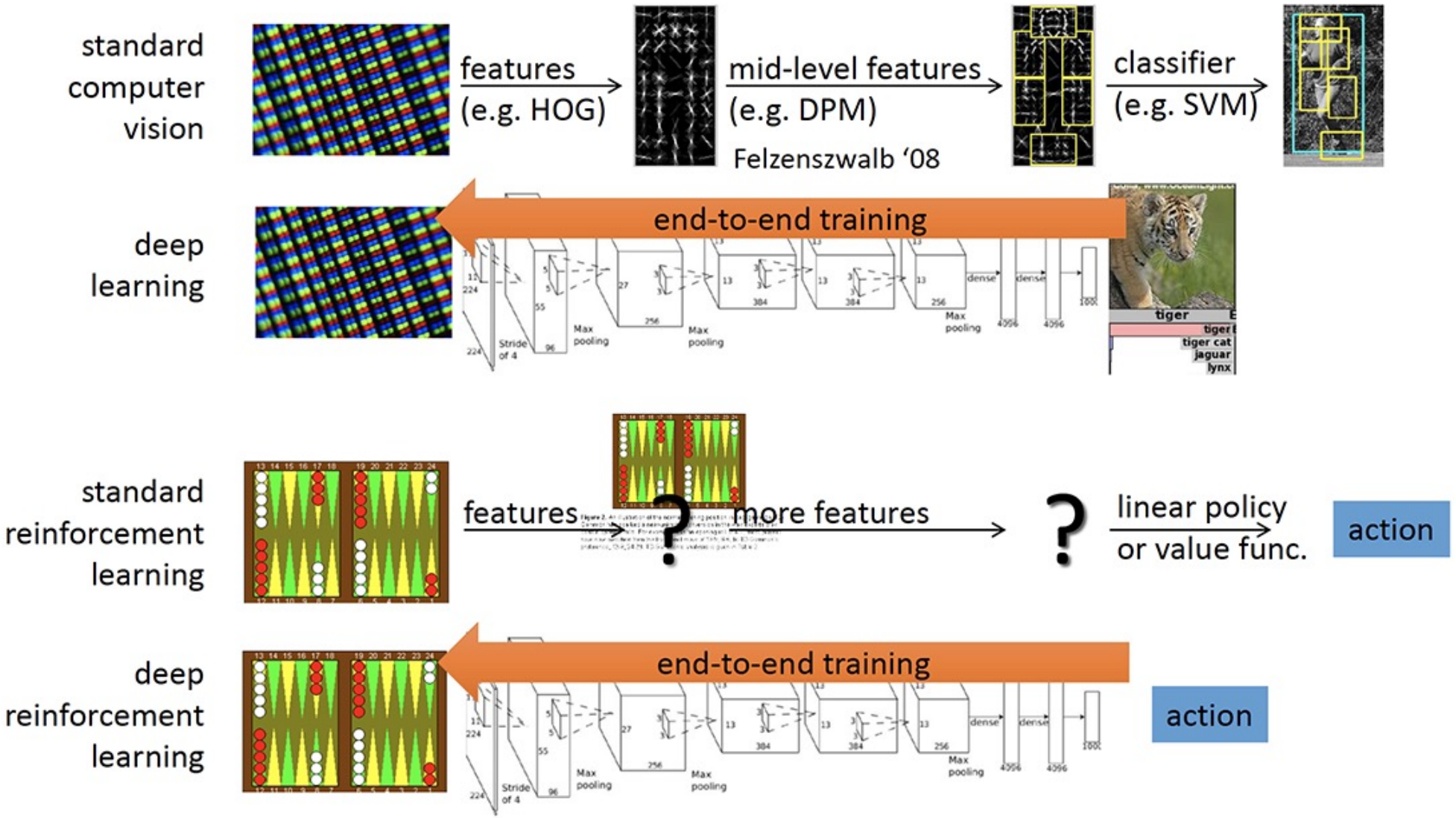




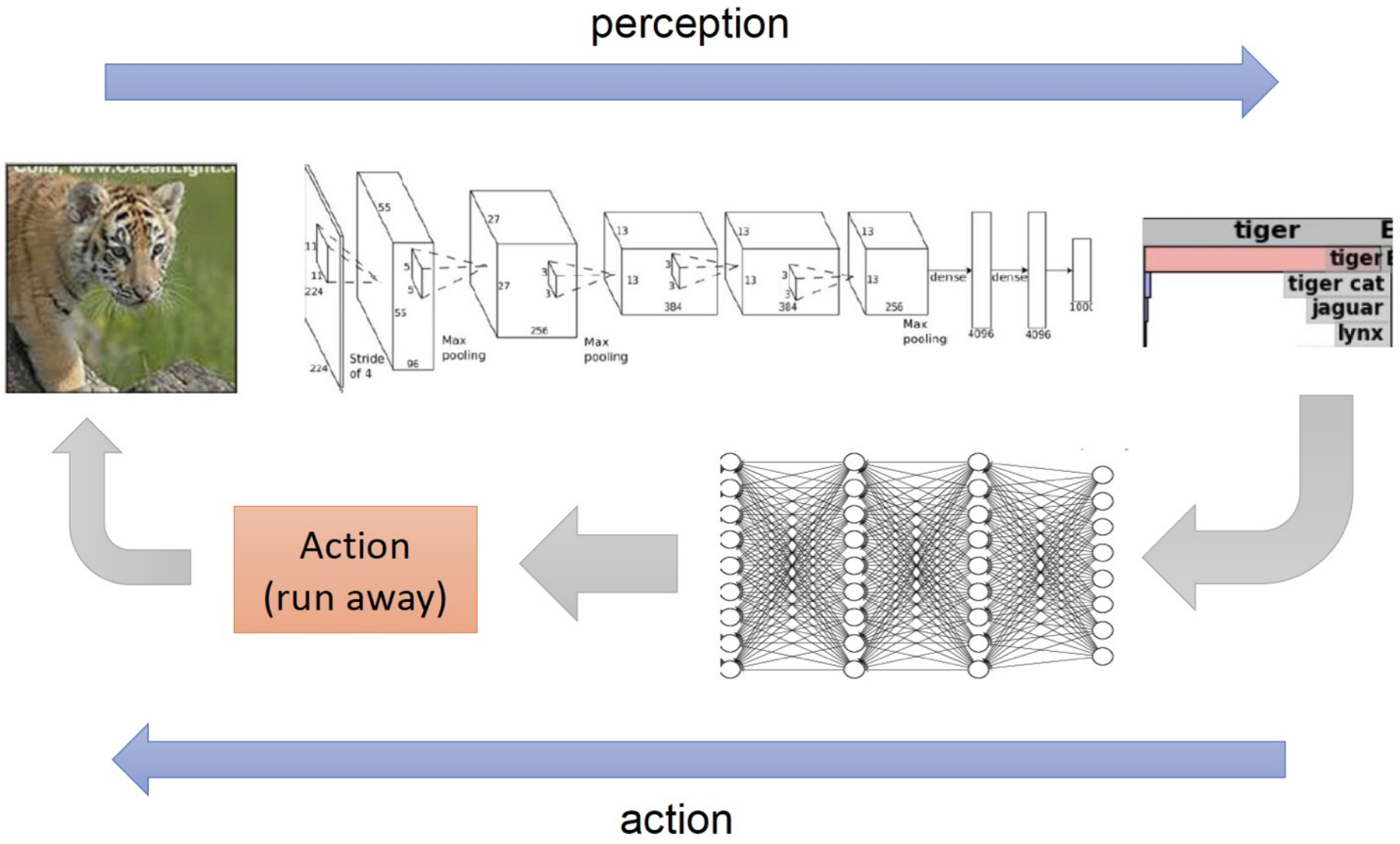


Space Invaders

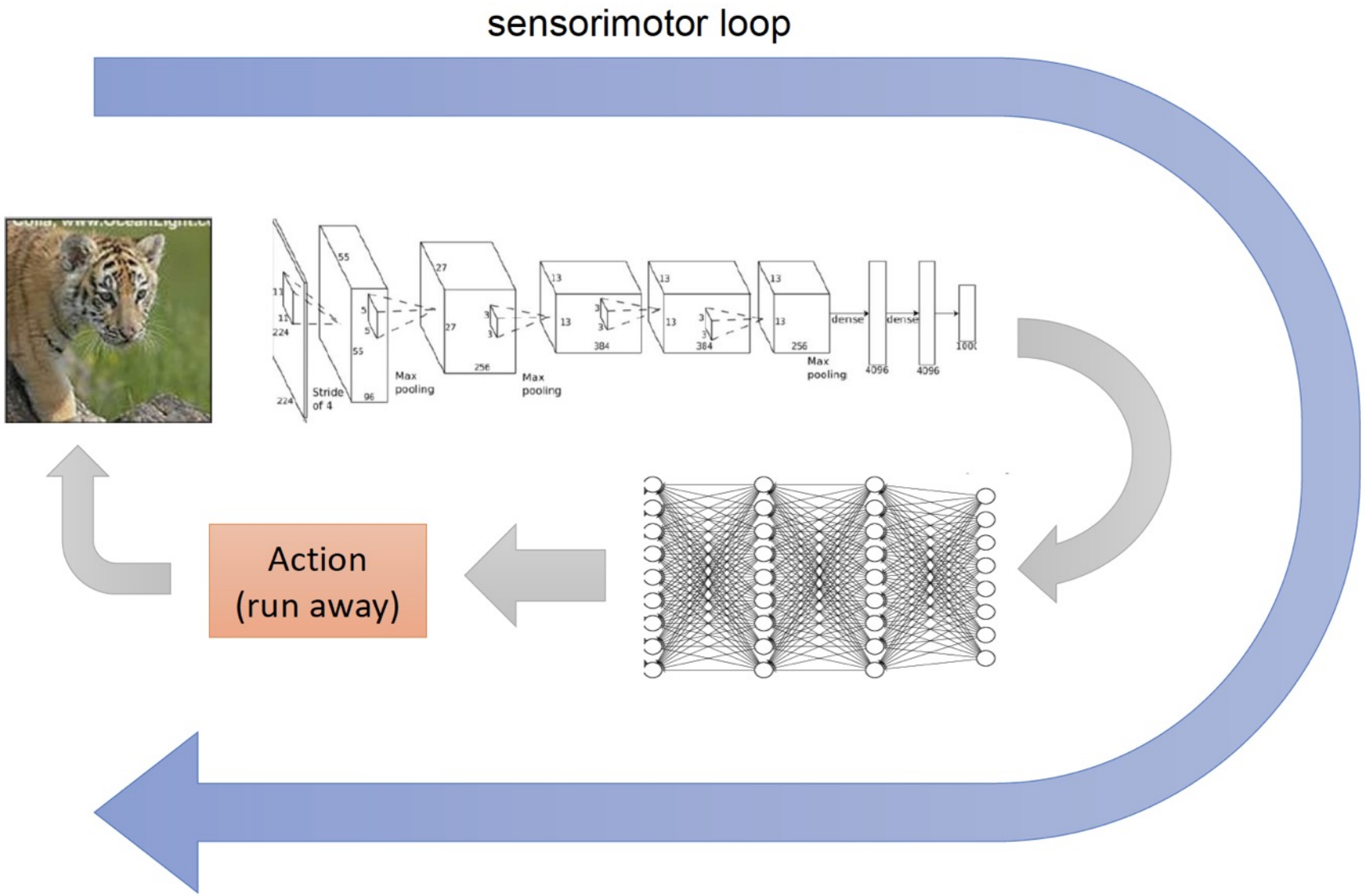
Deep RL = Deep learning + RL



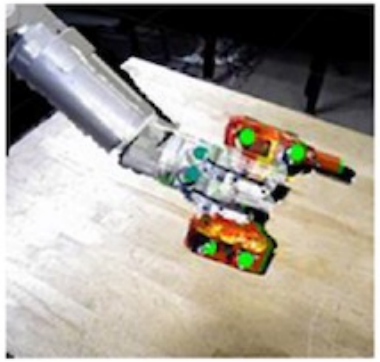
End-to-end learning for sequential decision-making

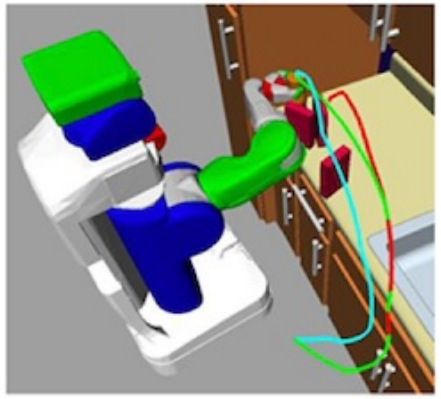


End-to-end learning for sequential decision-making



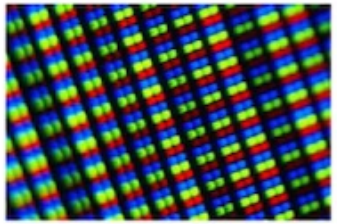
Example: Robotics

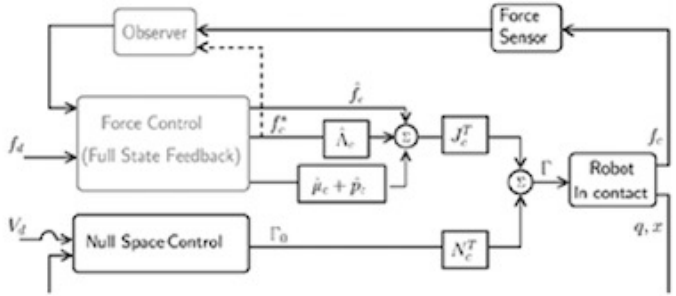




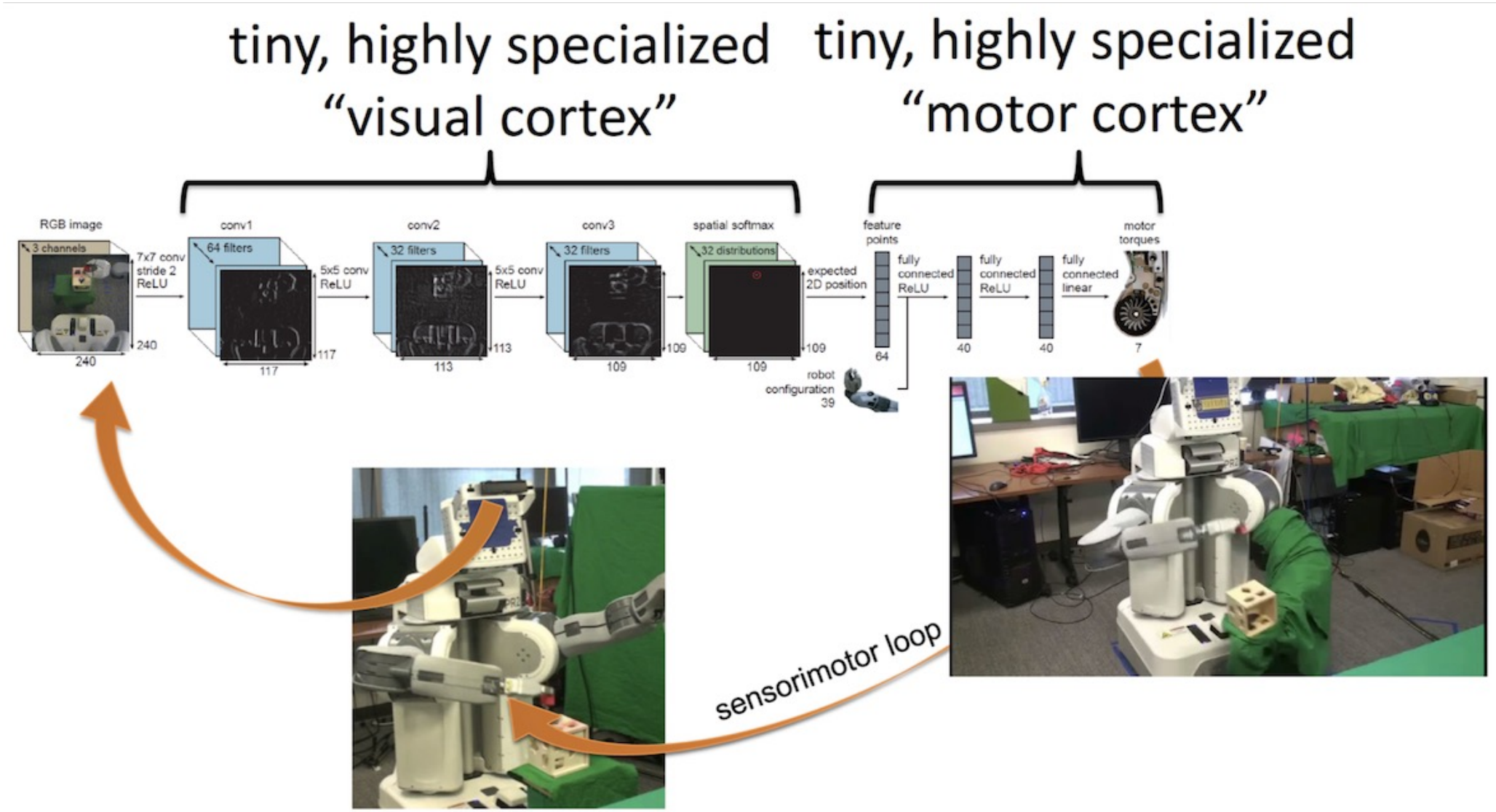
robotic control pipeline

observations → state estimation (e.g. vision) → modeling & prediction → planning → low-level control → controls



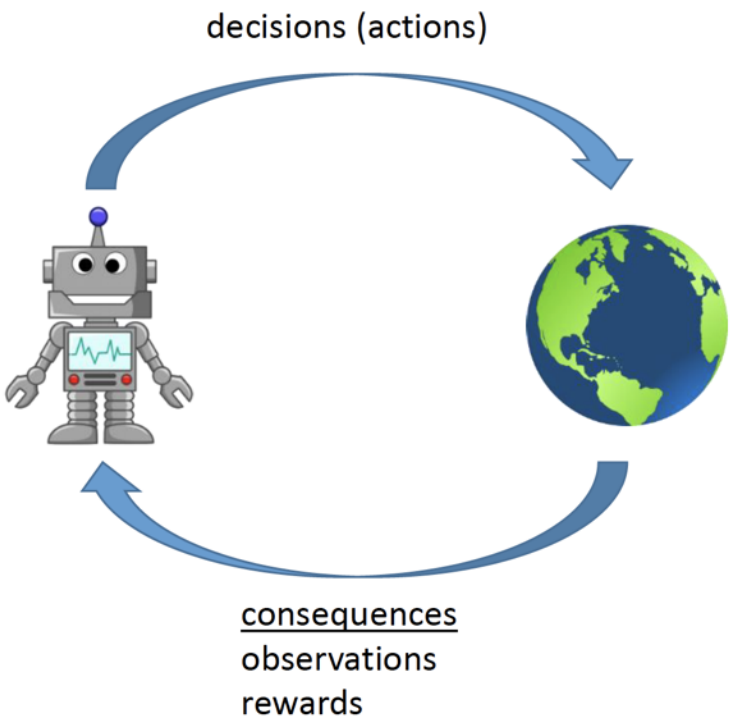


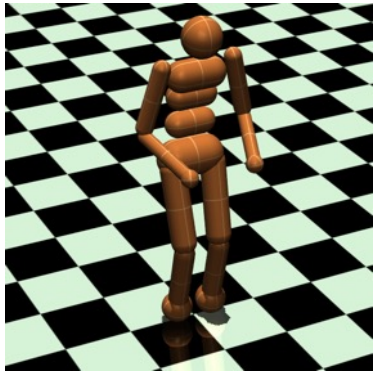
End-to-end learning

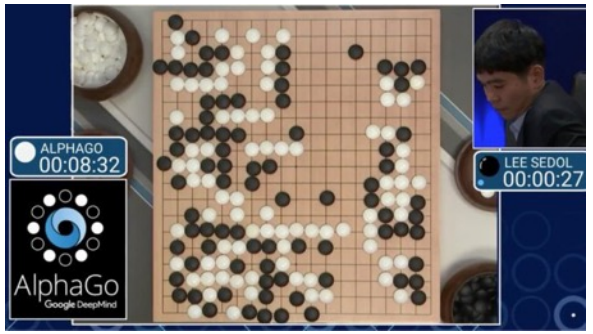


The RL problem is the AI problem

- Deep models allow RL algorithms to solve complex problems end-to-end
  - Deep learning: perception, handle unstructured environments
  - RL: decision, optimize its behavior



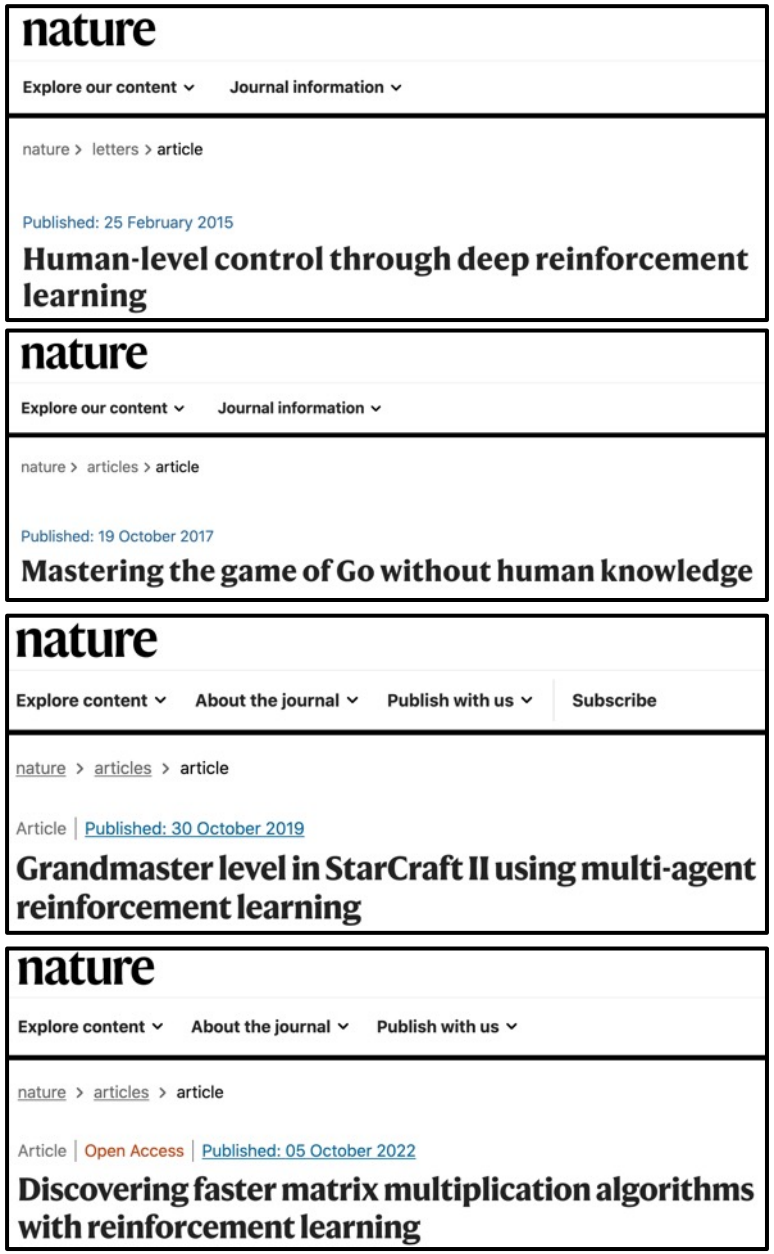




Space Invaders

The RL problem is the AI problem

- Advances in deep learning
- Advances in RL
- Advances in computational capability



The Era of RL

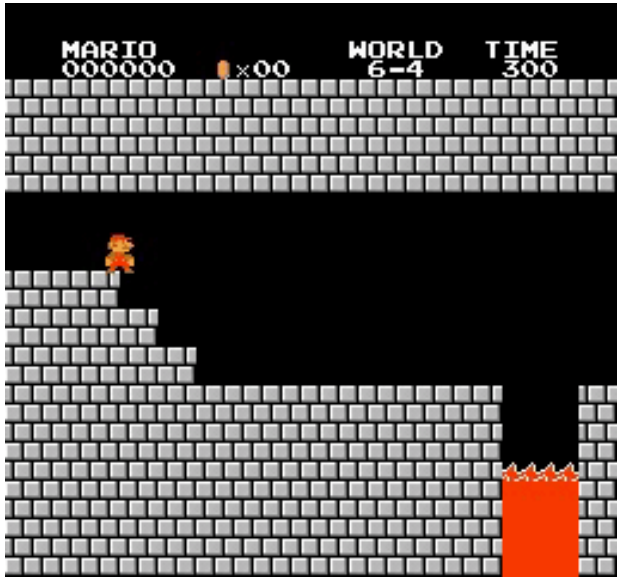
- Video games: Human-level control through DRL, Nature 2015 (视频游戏)
- AlphaGo, Nature 2016; AlphaGo Zero, Nature 2017 (围棋)
- AlphaStar in StarCraft II, Nature 2019 (星际争霸II)
- DRL for legged robots, Science Robotics 2019 (机器人学习)
- Superhuman AI for multiplayer poker, Science 2019 (德州扑克，多人非完全信息博弈)
- Discovering faster matrix multiplication algorithms, Nature 2022 (矩阵相乘算法发现，基础数学)
- Magnetic control of tokamak plasmas, Nature 2022 (可控核聚变控制)
- Outracing champion Gran Turismo drivers, Nature 2022 (赛车模拟控制)
- Safety validation of autonomous vehicles, Nature 2023 (无人驾驶安全验证)
- Faster sorting algorithms discovering, Nature 2023 (排序算法发现，基础信息科学)
- Champion-level drone racing, Nature 2023 (无人机竞速)
- Avoiding fusion plasma tearing instability, Nature 2024 (可控核聚变)
- OpenAI o1, DeepSeek-R1

The Era of RL


MARIO
000000
×00
WORLD
6-1
TIME
400


MARIO
000000
×00
WORLD
6-2
TIME
400


MARIO
000000
×00
WORLD
6-3
TIME
300


MARIO
000000
×00
WORLD
6-4
TIME
300

The Era of RL

The Era of RL

Generative adversarial imitation learning (GAIL)

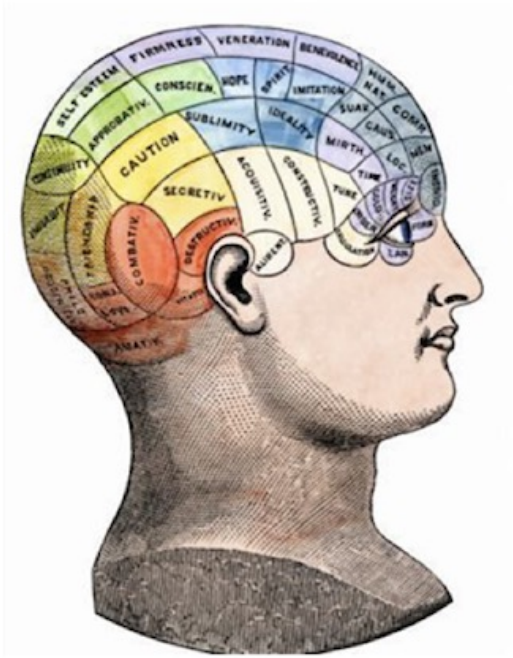
Learning biomimetic behaviors

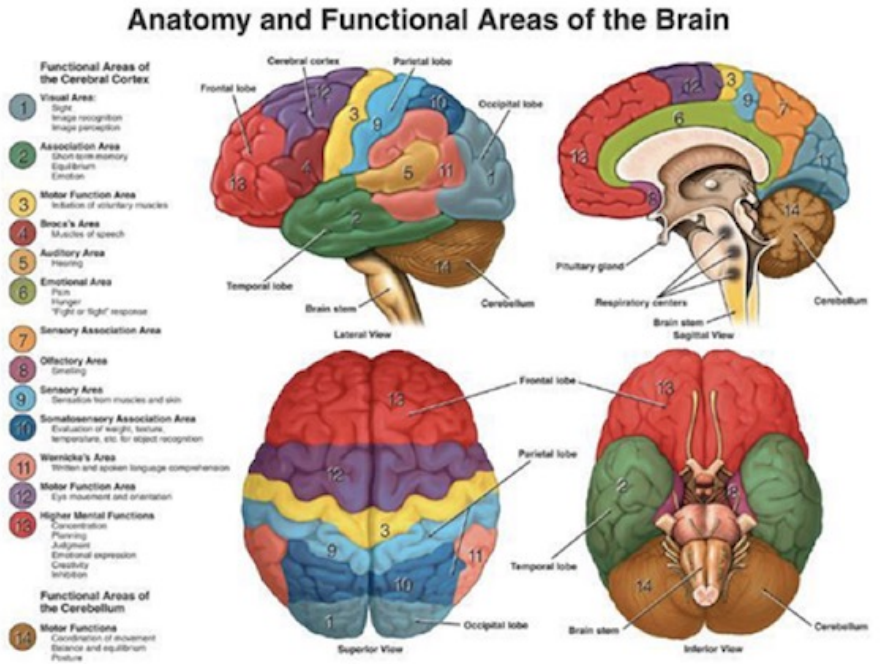
Content

1. Artificial Intelligence (AI)
2. Reinforcement Learning (RL)
3. Why should we care about (deep) RL?
4. How to build intelligent machines?
5. Beyond learning from reward

How do we build intelligent machines?

- Imagine you have to build an intelligent machine, where do you start?





“Learning” as the basis of intelligence

- Some things we can all do (e.g., walking)
- Some things we can only learn (e.g. driving a car)
- We can learn a huge variety of things, including very difficult things
- Therefore, our learning mechanisms are likely powerful enough to do everything we associate with intelligence
  - But it may still be very convenient to “hard code” a few really important bits

What must a single algorithm do?

- Interpret (rich) sensory inputs

- Choose (complex) actions

Why (deep) RL?

- Deep = can process complex sensory inputs
  - Compute really complex functions

- RL = can choose complex) actions
  - Reinforce its behaviors by rewards

Some evidence for RL

Reinforcement learning in the brain

Yael Niv
Psychology Department & Princeton Neuroscience Institute, Princeton University

- Basal ganglia (基底神经节) appears to be related to reward system
- Model-free RL-like adaptation is often a good fit for experimental data of animal adaptation (but not always)

Some evidence for RL

Quantum reinforcement learning during human decision-making

Ji-An Li, Daoyi Dong, Zhengde Wei, Ying Liu, Yu Pan, Franco Nori & Xiaochu Zhang

Nature Human Behaviour **4**, 294–307(2020) | Cite this article

- RL has been widely applied in neuroscience (神经科学) and psychology (心理学)
- Value-based decision-making can be illustrated by quantum RL (量子强化学习) at both the behavioral and neural levels

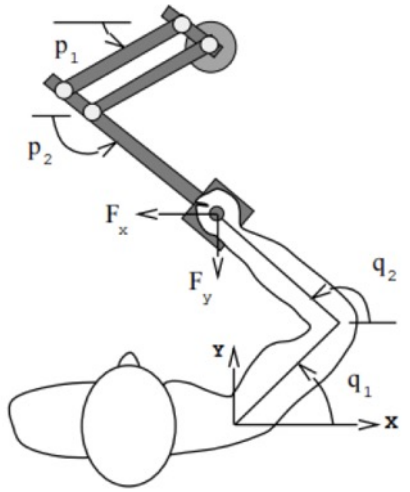
What can (deep) RL do?

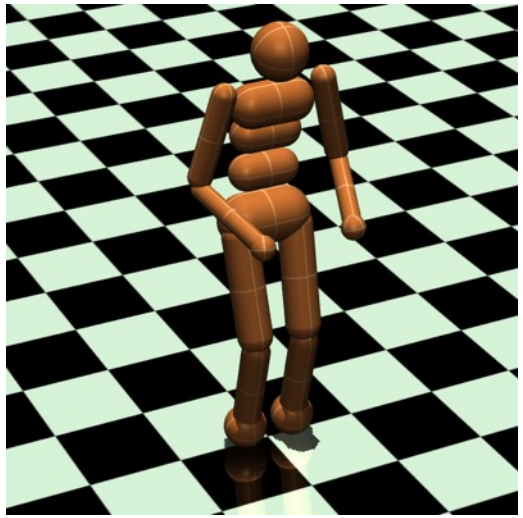
- Acquire high degree of proficiency in domains governed by simple, known rule
- Learn simple skills with raw sensory input given enough experience
- Learn from imitating enough human provided expert behavior (模仿学习)

Still in the infancy, and a promising future!

What has proven challenging so far?

- Humans can learn incredibly quickly
  - (Deep) RL methods are usually slow
- Humans can reuse past knowledge
  - Transfer learning in (deep) RL is an open problem
- Not clear what the reward function should be
- Not clear what the role of prediction should be

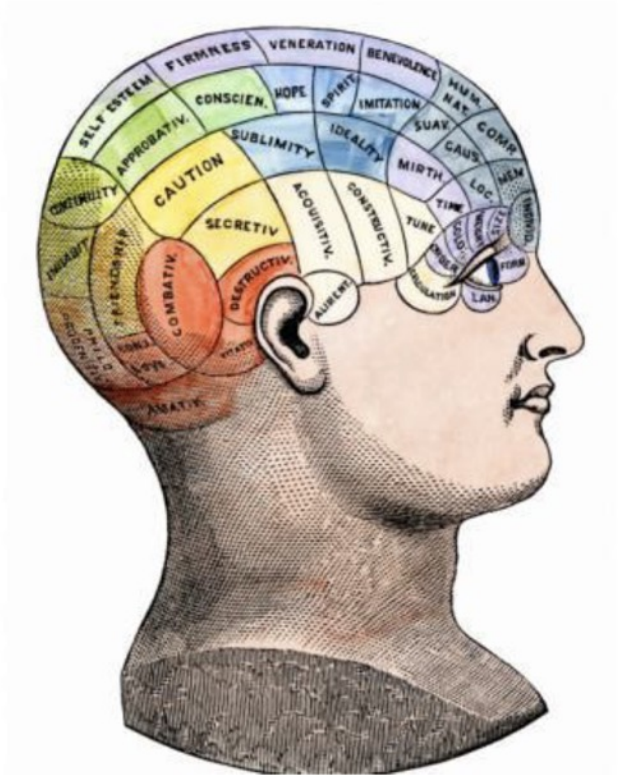


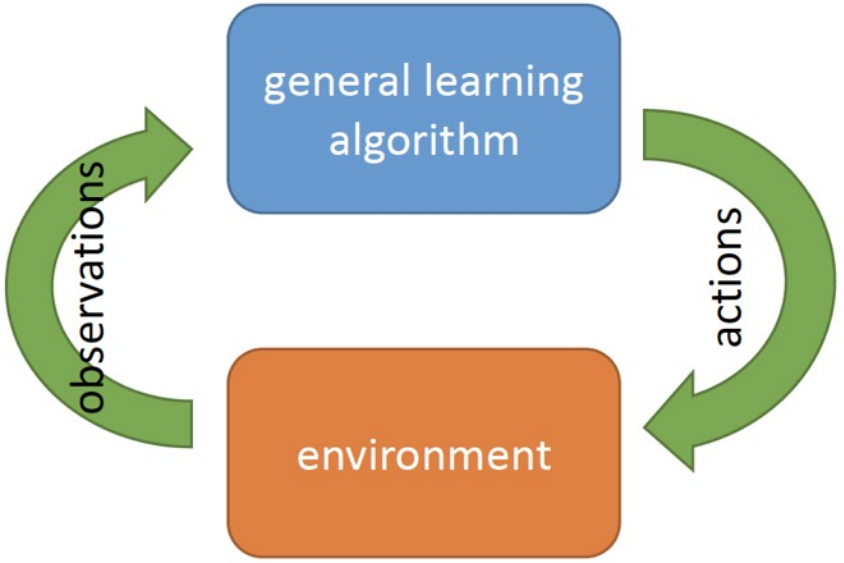


RL, decision-making, human behavior

Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain.

- Alan Turing





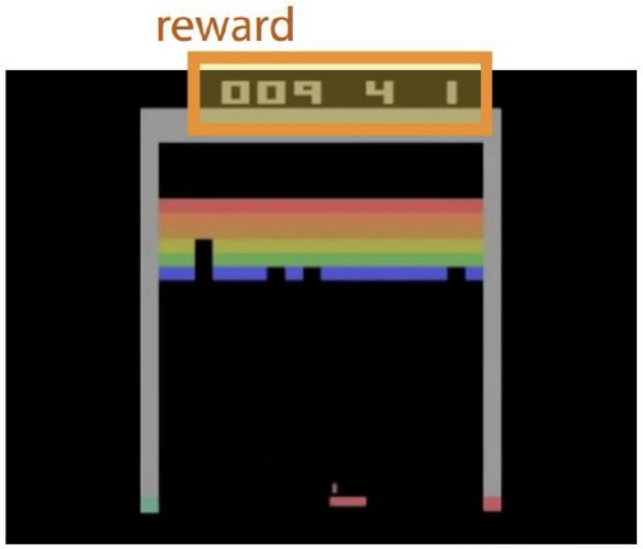
Content

1. Artificial Intelligence (AI)
2. Reinforcement Learning (RL)
3. Why should we care about (deep) RL?
4. How to build intelligent machines?
5. Beyond learning from reward

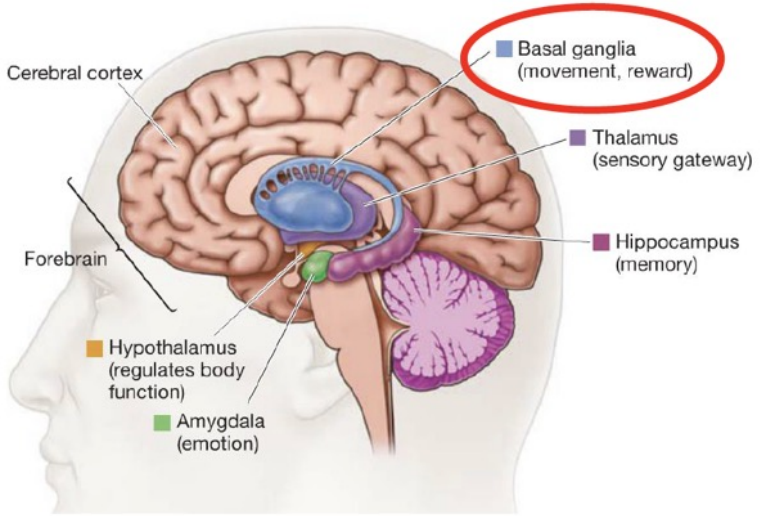
Enabling real-world sequential decision-making

- Basic RL deals with maximizing rewards
- This is not the only problem that matters for sequential decision-making!
- More advanced topics
  - Transferring knowledge between domains (transfer learning, meta-learning)
  - Learning to predict and using prediction to act (model-based RL)
  - Learning reward functions from example (inverse RL)

What do rewards come from?



reinforcement learning agent



- As human beings, we are accustomed to operating with rewards that are so sparse that we only experience them once or twice in a lifetime, if at all.

Are there other forms of supervision?

- Learning from demonstrations
  - Directly copying observed behavior (imitation learning)
  - Inferring rewards from observed behavior (inverse RL)
- Learning from observing the reward
  - Learning to predict
  - Unsupervised learning, self-supervised learning
- Learning from other tasks
  - Transfer learning
  - Meta-learning: learning-to-learn

Learning objectives of this lecture

- You should be able to…
  - Understand the basic concepts about sequential decision-making, the differences between supervised learning and RL
  - How RL provides a formalism for behavior and decision-making

THE END