Welcome to DDA4230!

Baoxiang Wang

Spring 2021

Today's class

- ❑ Course organization
- ❑ Introduction: Sequential decision making and reinforcement learning
- ❑ Preliminaries

Course staff

- **Instructor:**
  - *Baoxiang Wang*
    - **Office:** DY 503
    - **Office Hours:** Mon 3:30pm – 4:30pm
    - **Zoom Office Hours:** Mon 3:30pm – 4:30pm
    - **Zoom ID:** 974 4304 7377
    - **Zoom Passcode:** 181854
    - **Email:** bxiangwang@cuhk.edu.cn

- **Lectures**:
    - Room: CD 104
    - Times: Mon Wed 1:30pm-2:50pm
    - Zoom ID: 974 4304 7377
    - Zoom Passcode: 181854

Course staff (TAs)

• Shaokui Wei:	
Tutorial Schedule:	T1 Tu 7:00-7:50 PM
	T2 Tu 8:00-8:50 PM
(Zoom) Office Hours:	Tu 2:30-3:30 PM
Office:	TD 207
	Zoom ID: 929 5933 5402 Password: 832505
Email:	115010239@link.cuhk.edu.cn

Social distancing policy

Social distancing policy

- We encourage the following social distancing policy:
- 1. Wear masks at all times in lectures, tutorials and office hours
- 2. Keep appropriate distance between you and other classmates, TAs and professors
- 3. Maintain good personal hygiene at all times

Tutorials

- Tutorials begin on the **second** week, i.e. Jan 19th

- All tutorials will be broadcasted concurrently with the following Zoom ID:

- Zoom ID: 929 5933 5402
  Zoom Passcode: 832505

Course contents

- Introduction to reinforcement learning
- Multi-armed bandits and bandit algorithms The simplest RL problem.
- Markov decision processes Math formulation of RL
- Discrete MDPs, policy iteration, value iteration
- Policy evaluation, policy gradient, actor-critic method standard RL algorithms
- Temporal-difference method, SARSA, Q-learning Also standard
- Recent advancements in RL 2010' AlphaGo, Alpha Fold etc.

A large portion

1980' RL first proposed by Rich Sutton

1950' Machine intelligence and optimal control by Alan Turing

1850' Animal intelligence.

Course resources

- Referred courses

ELE524 Princeton: Foundations of Reinforcement Learning by Chi Jin

CSE599 UW: Reinforcement Learning and Bandits by Alekh Agarwal and Sham Kakade

CS285 UCB: Deep Reinforcement Learning by Sergey Levine

CMPUT397/609 Alberta: Reinforcement Learning I/II by Martha White and Rich Sutton

- Referred books

Dimitri P Bertsekas and John N Tsitsiklis. Neuro-dynamic programming. 1996.

Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. 2018

Csaba Szepesvari. Algorithms for reinforcement learning. 2010

Lattimore, Tor, and Csaba Szepesvári. Bandit algorithms. 2020

Reinforcement Learning: Theory and Algorithms. Alekh Agarwal, Nan Jiang, Sham Kakade, Wen Sun. 2020

Assessment Scheme

Assignments: 25% (4 assignments)

Midterm exam: 25%; Mar 24th (Wendesday) 1:30pm –2:50pm, using the lecture time. Total 1 hour and 20 minutes.

Final project: 50%

Assessment Scheme

- Assignments: 25%; 4 times total

First 2 are written assignments on RL foundations

Assignment 3 and 4 are more algorithmic and involve coding and implementations

- If you work on theoretical topics for your final project, you have an option to skip assignment 3 and 4 (all weights go to assignments 1 and 2). Please email Shaokui Wei to request so before Mar 1.
- For written assignments, LaTeX is encouraged (handwritten also fine). For coding, PyTorch is recommended (other programming languages/packages also fine)
- Assignment box TBD; Tolerate 24-hour late with 20% penalty.

Assessment Scheme

- Midterm exam: 25%; Mar 24th (Wendesday) 1:30pm –2:50pm, using the lecture time. Total 1 hour and 20 minutes.
- Written exam on RL basics; 3-4 long questions; open book (paper-based materials, no electronic devices)

Tentative topics: 1) Math formulation of MDPs; 2) Discrete MDPs and policy/value iteration 3) policy gradient and Q-learning 4) RL theory

No coding questions in midterm;

Assessment Scheme

- Final Project: 50%

Including 1) final project proposal; 2) final project report and code archive 3) oral presentation (optional); Complete mainly by yourself.

Free to choose any topic in RL theory/algorithms/applications

- Proposal due on Apr 2. Optional to make appointment to discuss your proposal; In rare case, proposals can be rejected and the student will be asked to write a new proposal.
- Final report due on May 14. Mark will be based on report and code.
- Oral presentation on May 10 – May 14. Optional. You can choose to make a 4-minute oral presentation to help us understand your work and highlight your contribution.

Assessment Scheme

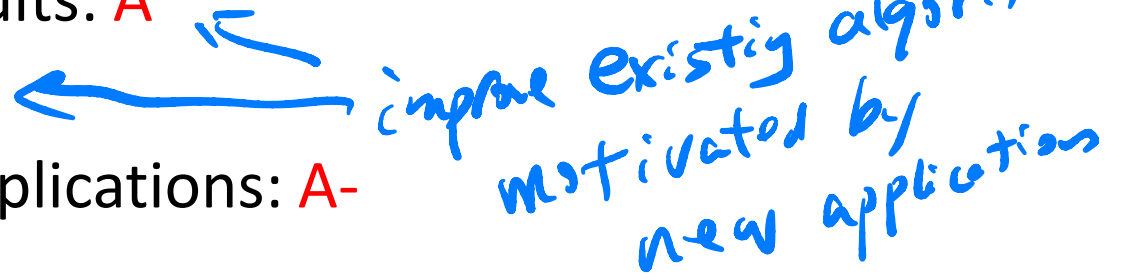
- Final Project: 50%
- Research-oriented project with freedom to explore novel topics in RL
- Marking scheme

Novel, significant, and publishable results: A

Novel but incremental results: A

Existing algorithm/theory with new applications: A-

Reproduction: B+, B, or B-



- Complete mainly by yourself. You might have collaborators but 1) your contribution need to be more than 50% 2) you need to acknowledge them appropriately

Final Project Topics

- Theory

Bandit algorithms: combinatorial bandits, online learning to rank, online influence maximization, matching bandits, strategic bandits, bandit applications in recommendation systems

Discrete MDP theory: sample complexity bounds, regret bounds, RL with UC-exploration, exploration and covering of RL

Other theoretical topics: Privacy in RL, fairness in RL, constrained RL, optimization methods (e.g. trust region methods, distributionally robust methods), theoretical connection of existing RL algorithms,

Final Project Topics

- Algorithmic

Sample Efficiency and Variance Reduction: Statistical RL (with advanced statistics) for variance reduction

SC2 minigame for computing power

Multi-agent RL: game theoretic approaches, centralized training and decentralized execution, multi-agent games like StarCraft II and DOTA 2

Strategic RL: RL with strategic environment, RL with network games

Off-policy policy optimization and policy evaluation

Offline RL and medical treatment and recommendation systems

Model-based RL and search algorithms

Final Project Topics

Gamification

- Applications (cast an application to an RL game)

RL methods in robotic control

RL methods to play games

RL for medical treatment

RL for recommendation systems

RL for mathematical problems: e.g. TSP, optimization, linear algebra

RL for equity trading and execution

About Research-oriented Projects

- Encourage new topics
- Research – explore the uncharted
- Differences with teaching-based projects

Important dates

- Jan 18 (Jan 29 due): Assignment 1
- Feb 22 (Mar 5 due): Assignment 2
- Mar 8 (Mar 19 due): Assignment 3
- Mar 22 (Apr 9 due): Assignment 4
- Mar 24: Midterm
- Apr 2: Final proposal due
- May 10 - 14: week of presentation
- May 14: Final report due

Today’s class

☑Course organization

❑Introduction: Sequential decision making and reinforcement learning

❑Preliminaries

RL versus Supervised Learning

Supervised learning

- Given an *assumption* of the data generation process, (e.g. i.i.d. with Gaussian noise), figure *prediction*

Reinforcement learning

- Given *observations* of what happened, figure out *the action to take*

no train dataset

no teacher

learn from scratch

RL versus Optimal Control

Go game: when you place a stone, there will be a stone

Optimal control

- Given a *model* of the real world (called the world model), analytically figure out *the best action to take*

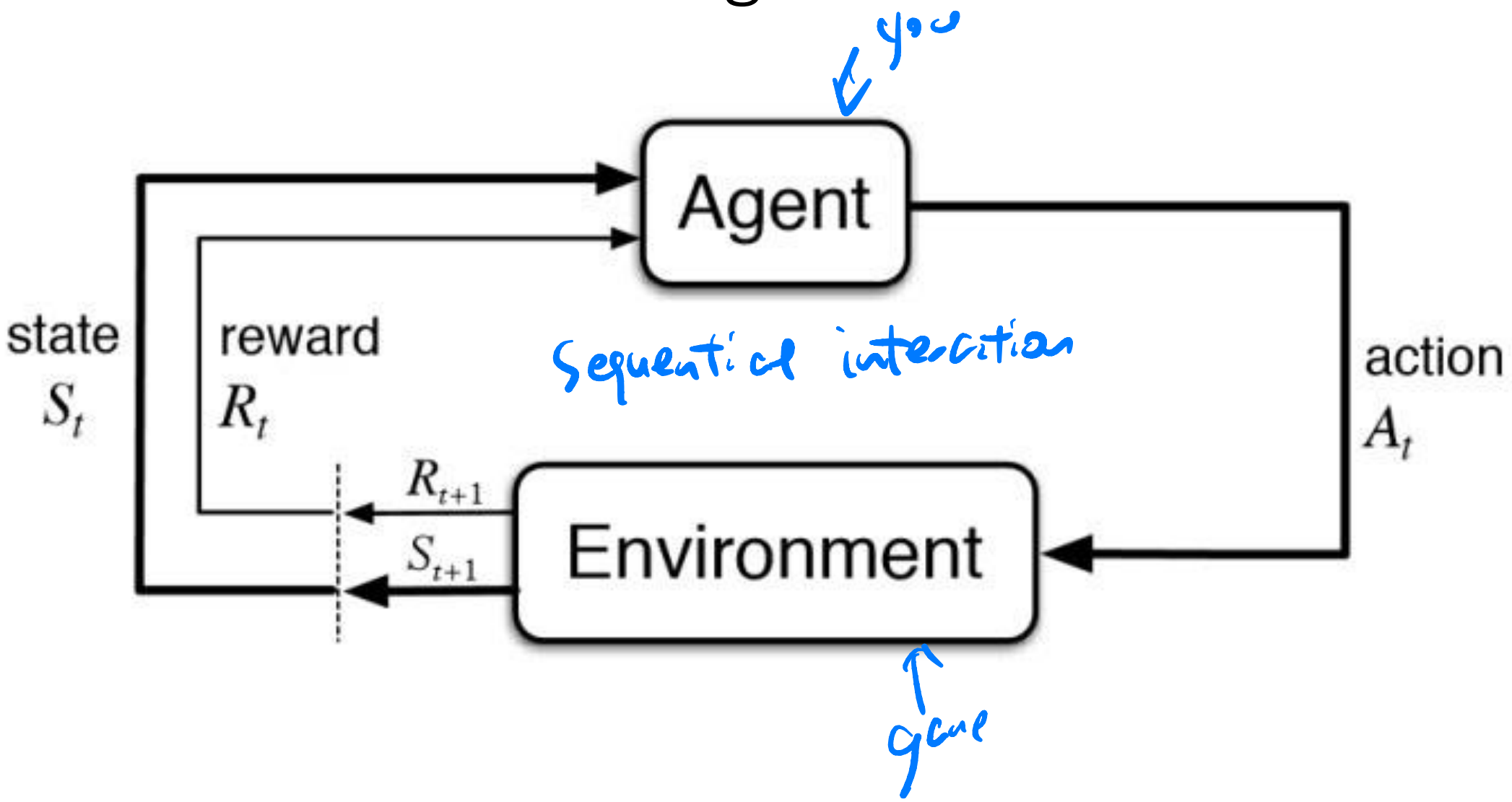
Reinforcement learning

RL = Optimal control + trial and error

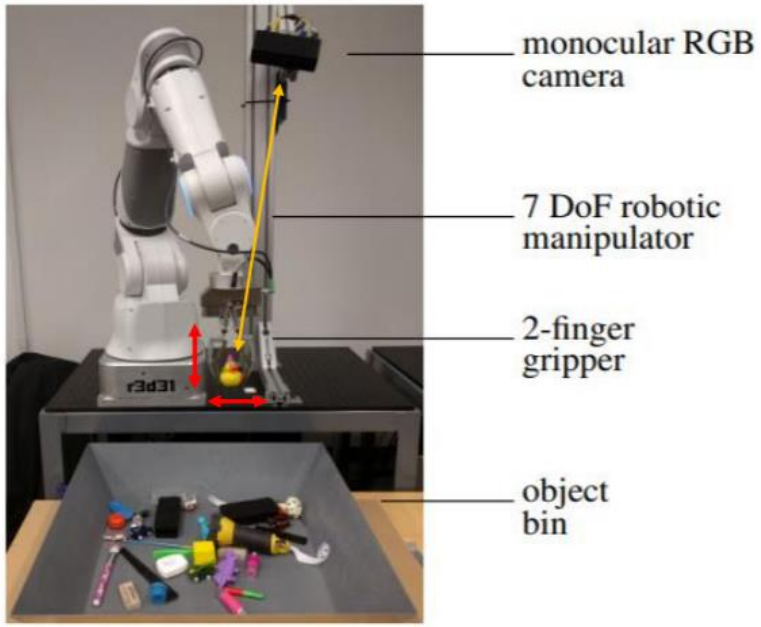
- Given *observations* of what happened, through trial and error, estimate *the best action to take*

animal learning

Reinforcement learning

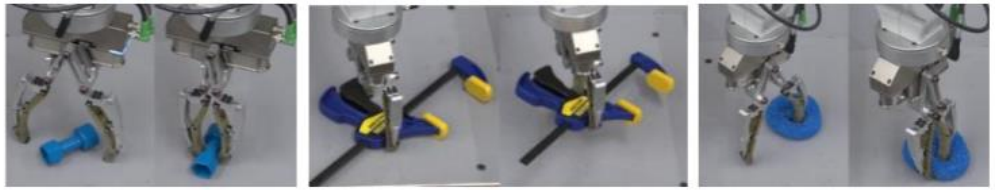


RL versus SL: Examples (source: CS285 UCB)



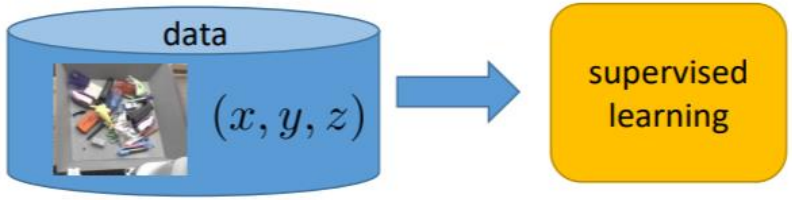
Option 1:

Understand the problem, design a solution



Option 2:

Set it up as a machine learning problem



RL versus SL: Examples (source: CS285 UCB)

Standard (supervised) machine learning:

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

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

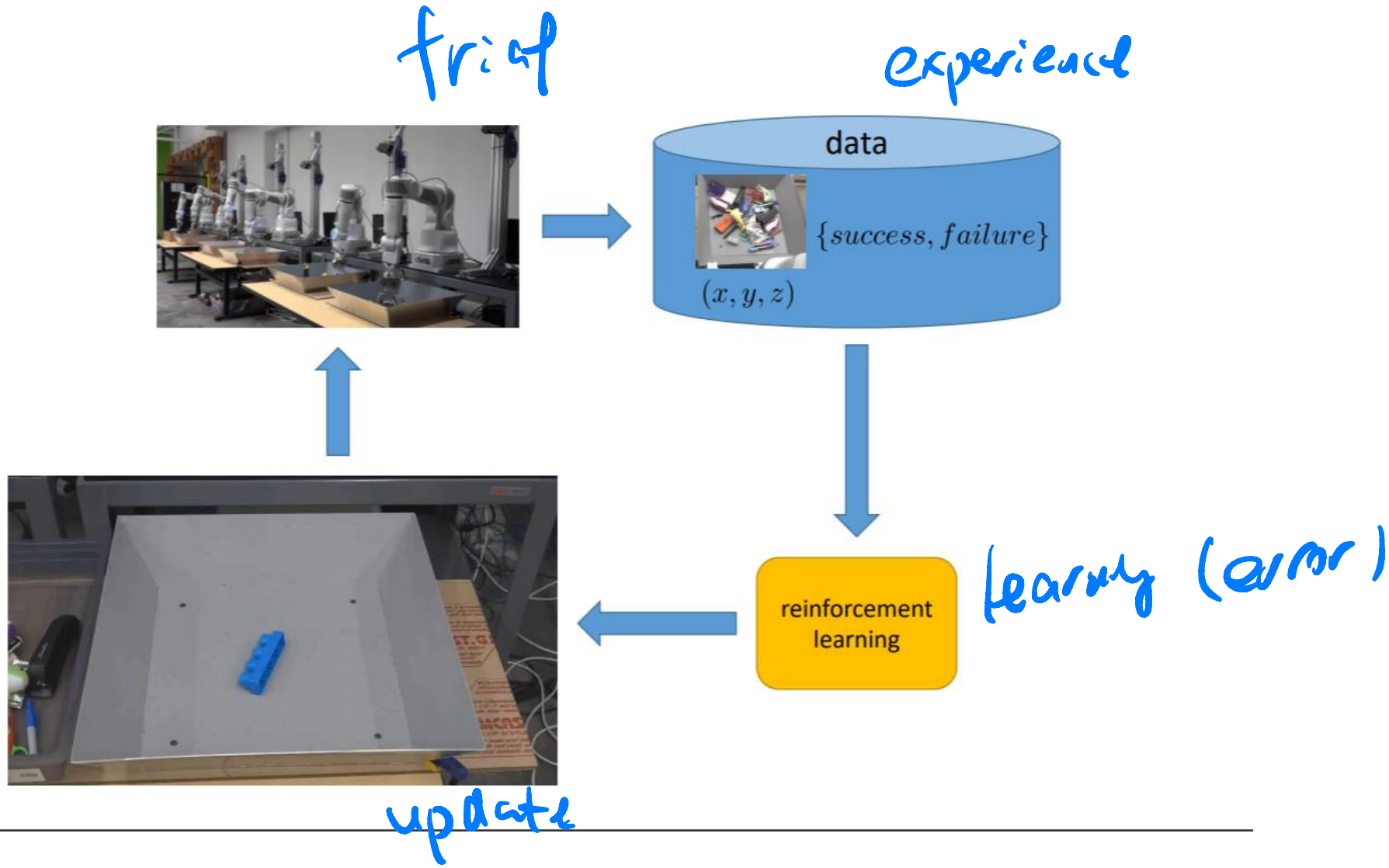
Usually assumes:

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

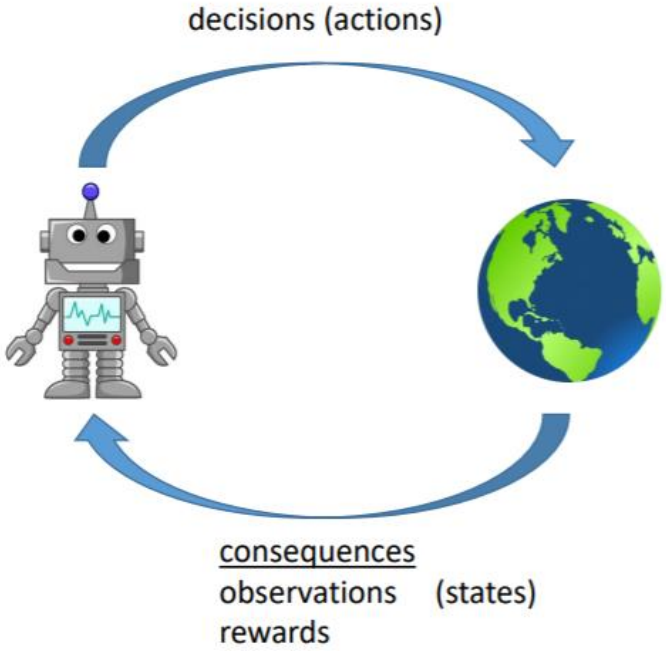
Reinforcement learning:

- 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

RL versus SL: Examples (source: CS285 UCB)



RL versus SL: Examples (source: CS285 UCB)

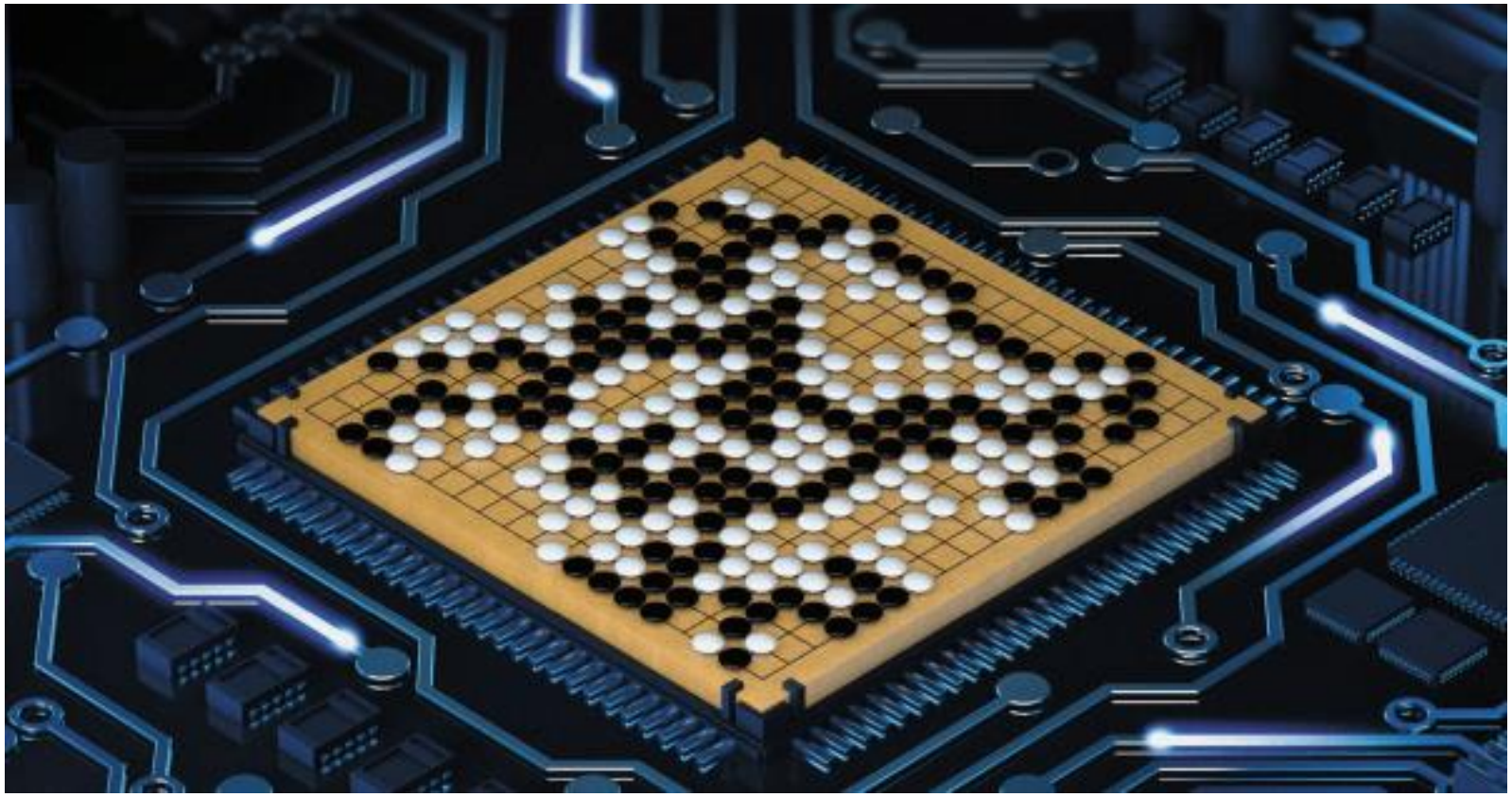


Actions: muscle contractions
Observations: sight, smell
Rewards: food

Actions: motor current or torque
Observations: camera images
Rewards: task success measure (e.g., running speed)

Actions: what to purchase
Observations: inventory levels
Rewards: profit

RL Milestone Applications



AlphaGo

AlphaGo Zero

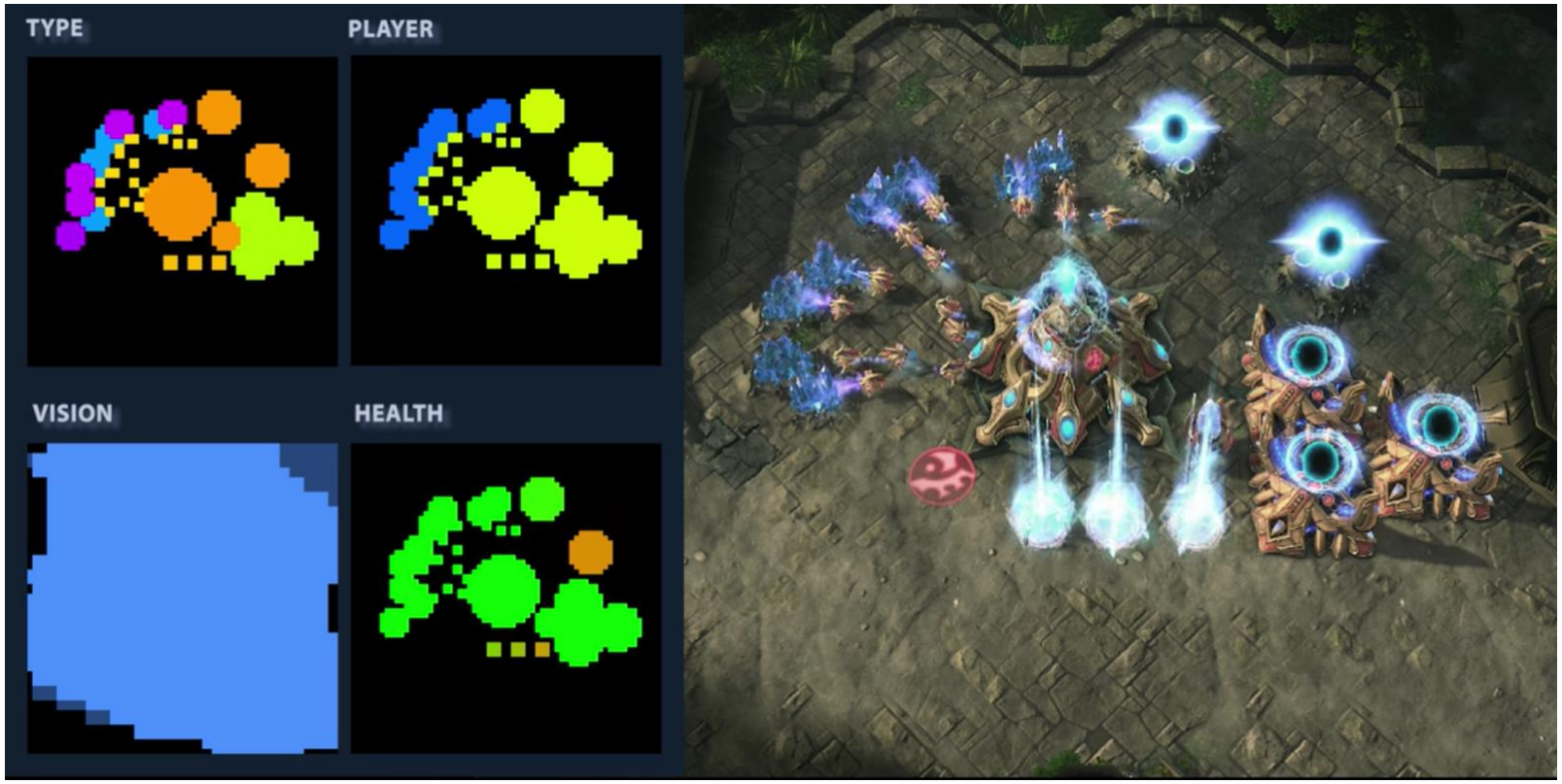
AlphaZero

MuZero

No person can beat it

Used widely in pre training

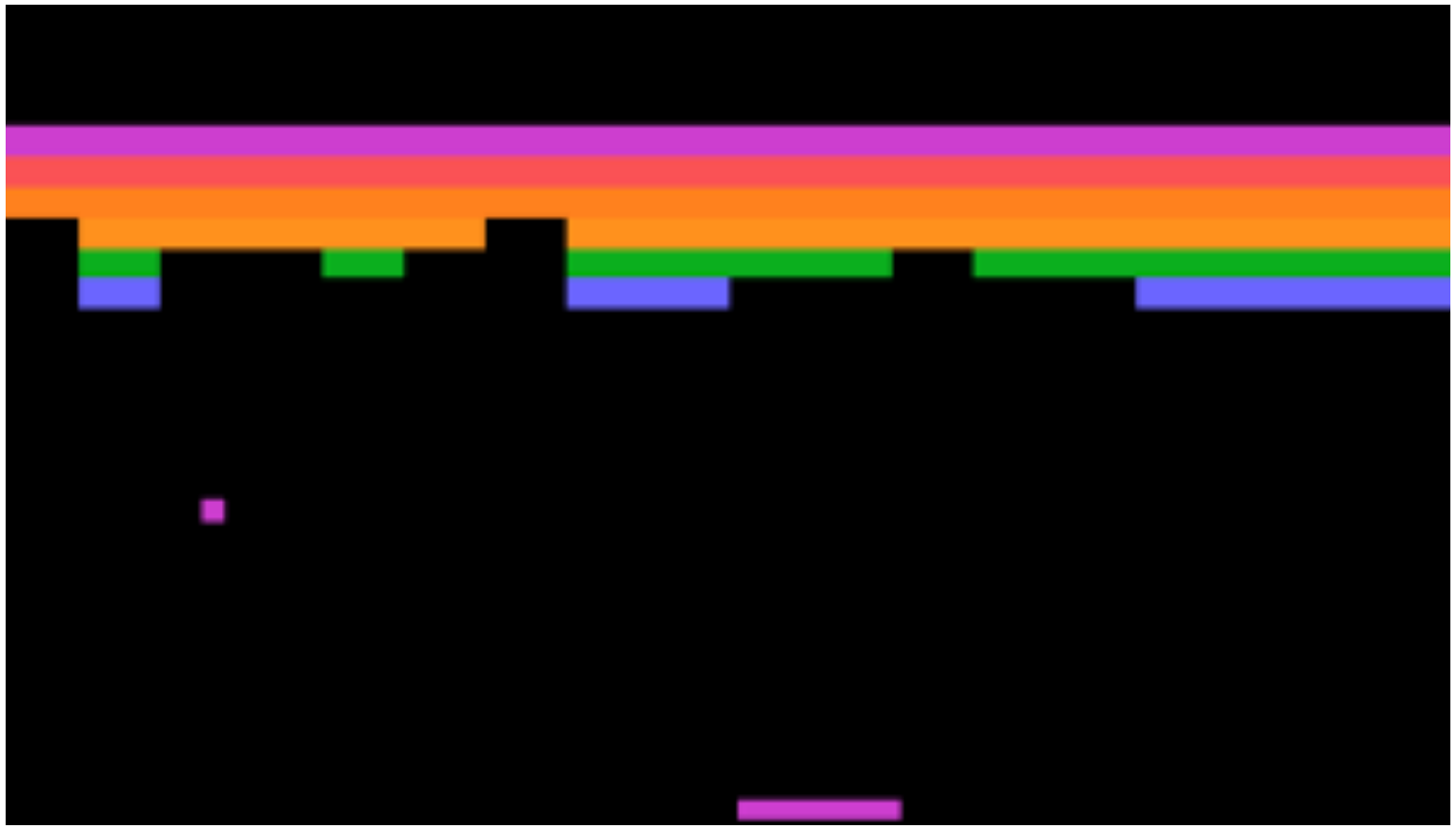
RL Milestone Applications



Alphastar

Still open

RL Milestone Applications



NIPS '13
Nature '15

Solved.
Standard.

RL Milestone Applications

More game theoretical et al.
by Mike Bowling '15 17 Nature
by Noam et al. '17 Science

RL Milestone Applications

AlphaFold '20

RL Industrial Applications (personal opinion)

RL has been deployed to the following domains in industry

- Game AI
- Stock trading/execution and market making

} deployed

RL has some industrial pilots

- Autonomous driving — sensitivity
- Recommendation systems — offline
- Medical treatment and healthcare — offline
- Robotics — Safety

Today's class

☑Course organization

☑Introduction: Sequential decision making and reinforcement learning

❑Preliminaries

- ❑Analysis, linear algebra (multivariate)
- ❑Probability, statistics
- ❑Limit theorems, concentration bounds, machine learning, optimization

Analysis

- Multivariate calculus

Useful as a general tool; Specifically useful to derive occupancy distribution-related algorithms

- Multivariate linear algebra

Useful as a general tool; Useful for gradient-based algorithms (this is a large class of algorithms)

- Functional Analysis

Helpful to understand functional variables conceptually. Useful occasionally for kernel methods, Gaussian processes etc.

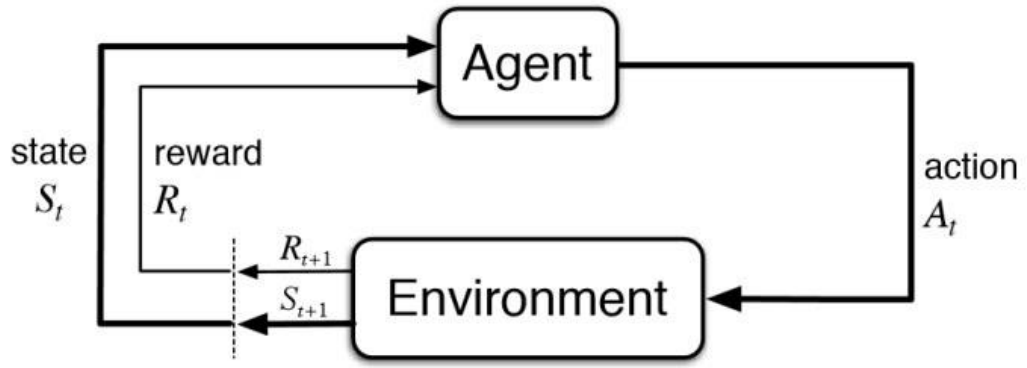
Probability and Statistics

- Many variables are stochastic

State transition

Reward

Policy



- Formulations of RL will be based on stochastic variables
- Unbiased estimation, variance reduction etc. for algorithms

Advanced Tools

- Limit theorems and concentration bounds

Mostly used tools for bandits and discrete MDPs

- Machine learning

Policy evaluation, Q-learning, off-policy learning, offline RL

- Optimization

Policy gradient and actor-critic (e.g. TRPO, PPO)

Thanks!

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