Reinforcement Learning: Introduction

Chen-Yu Wei

RL is a New Regular Course in UVA CS

- From this semester, the CS department makes RL a **regular** course
  - CS 4501 → CS 4771
  - CS 6501 → CS 6771

- There will be Graduate RL **and** Undergraduate RL **every semester**
  - We will use the following pattern in the near future (at least 2 years)

	Fall	Spring
Undergraduate	Prof. Shangtong Zhang	Me
Graduate	Me	Prof. Shangtong Zhang

Platforms

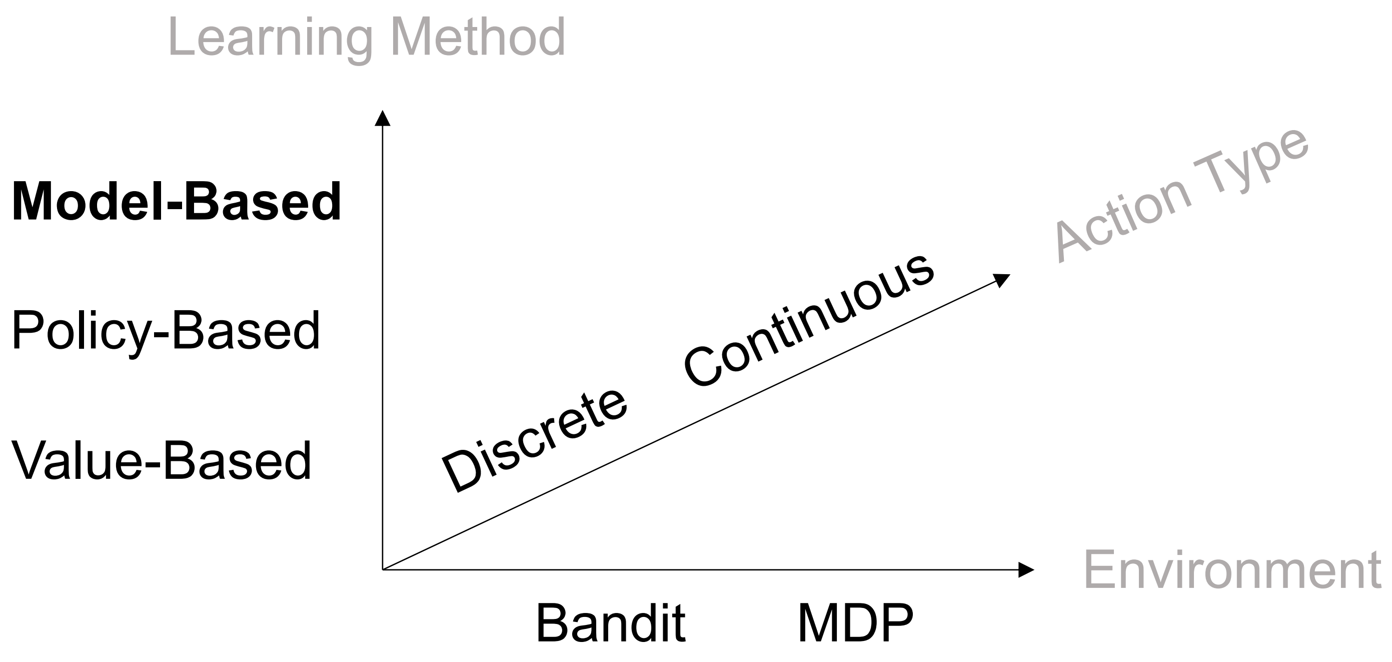
- Course website: https://bahh723.github.io/rl2025fa/
  - Syllabus, announcement, slides, lecture recordings
  - Can be accessed from Lou's List or my personal website
- Gradescope (haven't created)
  - Homework submission
- Piazza (haven't created)
  - Questions and discussions

They will be created once the student list becomes more stable (e.g., next week).

Topics in This Course

- The **math** behind **basic** RL algorithms
- We will follow the previous semester (Spring 2025) closely (link)

Topics in This Course



Core Algorithms

Exploration in MDPs

Inference-Time Algorithms

Imitation Learning

Special Topics

Bold are tentative new topics in this semester

This semester I plan to have less *in-class* math proofs – the proofs go to homework.

Prerequisites

- Linear Algebra, Probability, Calculus, **Machine Learning**
- Convex Optimization
- Python

Assignments are sometimes math heavy (check the HWs in previous semesters)

Recommended Resources (longer list)

- Courses
  - UC Berkeley CS285
- Webpages
  - OpenAI SpinningUp
- Books
  - Sutton and Barto, Reinforcement Learning: An Introduction
  - Agarwal, Jiang, Kakade, and Sun, Reinforcement Learning: Theory and Algorithms
- Implementations
  - OpenAI StableBaseline3
  - ShangtongZhang

Assignments (60%): 4 Problem Sets

A mixture or one of the following:

- Math / algorithm design problems
  - Submission: Latex or hand-writing + taking photo

- Programming tasks (using PyTorch)
  - Might need you to plot results or report numbers
  - Submission: It's usually easier to do them in Latex (I'll release latex template)

Assignments (60%): 4 Problem Sets

- Late policy
  - 10 free late days distributed to all assignments as you like
  - No assignment can be submitted 7 days after its deadline
  - Each additional late day results in 10% deduction in the semester's assignment grade

- Examples
  - HW1: **3** days late, HW2: **6** days late, HW3: **3** days late, HW4: **2** days late
    → HW grade *= 0.6
  - HW1: **8** days late, HW2: **6** days late, HW3: **3** days late, HW4: **2** days late
    → HW1 = 0 points **and** HW grade *= 0.9

Final Project (35%)

- Breakdown
  - Proposal (5%): $\leq$ 3 pages in NeurIPS format
  - Midterm report (5%): $\leq$ 3 pages in NeurIPS format
  - Presentation (10%): Upload a ~10 mins video to Panopto (a shared online space)
  - Online discussions (5%): Discussions on Panopto
  - Final report (10%): $\leq$ 8 pages in NeurIPS format

- Types of projects (basically any)
  - Application, algorithm design, systematic comparison, theoretical understanding, survey…
- Goal: Apply RL techniques to problems you're interested in.

Final Project (35%)

- It may be built on existing projects
  - Describe in the proposal the **current status** of the project and **what’s new** (otherwise the proposal will get no points)
- 2-3 students in a group (no solo project is allowed)
- Proposal deadline: **September 27**

Participation (5%)

- You will get
  - ≥ 1 if attendance rate $\geq 30\%$
  - ≥ 2 if attendance rate $\geq 50\%$
  - ≥ 3 if attendance rate $\geq 80\%$ or if you have occasional interaction in the class or piazza (not including final presentation)
  - $= 5$ if you have very active interaction in the class or piazza

TA & Office Hour

- **TA:** Braham Snyder
  - Email: dqr2ye@virginia.edu
  - Office hour: TBD

- Me
  - Email: chenyu.wei@virginia.edu
  - Office hour: Monday 3:30-4:30pm at Rice 409

- Starting from the next week

Learning To Make Decisions from Interactions

Games

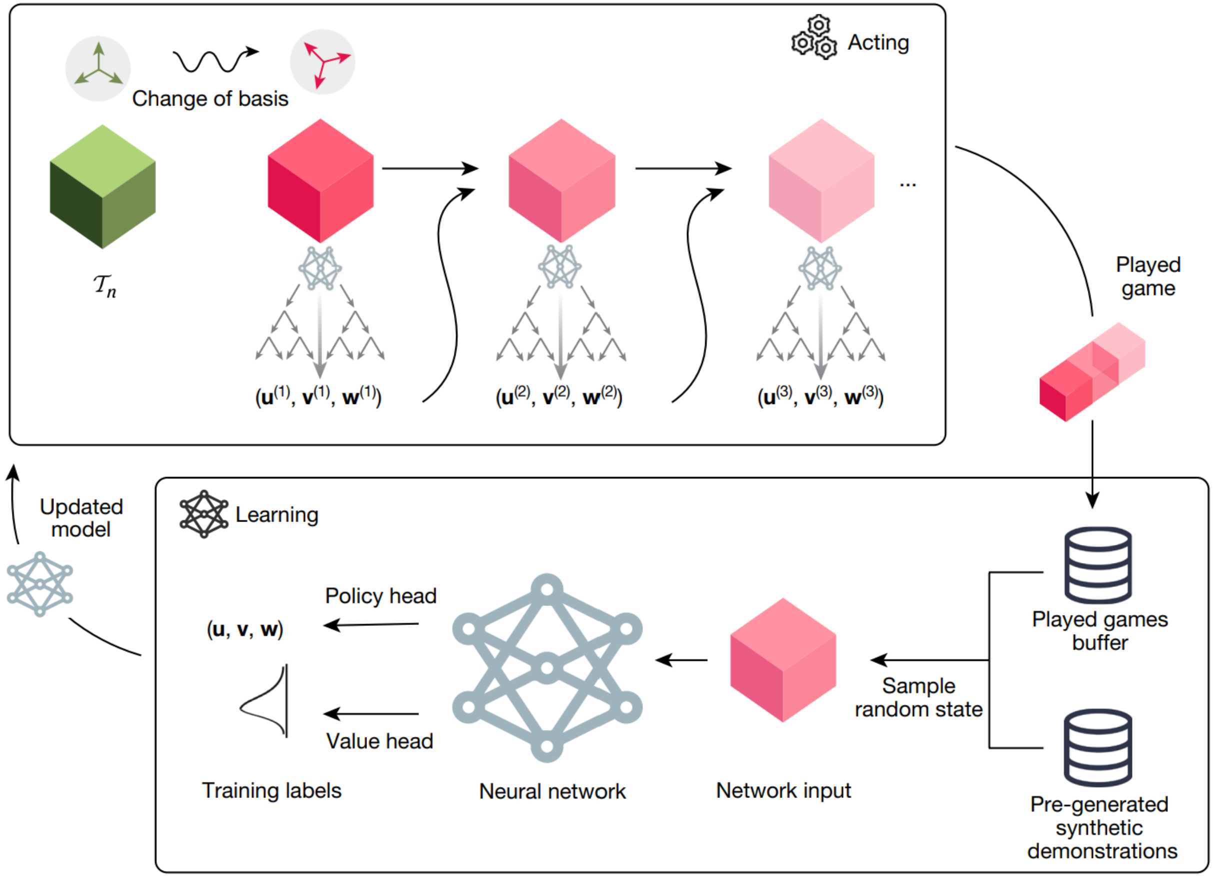
10 mins training

120 mins

240 mins

Mnih et al., Playing Atari with Deep Reinforcement Learning, 2015

Algorithm Discovery (faster matrix multiplication)



Size (n, m, p)	Best method known	Best rank known	AlphaTensor rank Modular	Standard
(2, 2, 2)	(Strassen, 1969)[2]	7	7	7
(3, 3, 3)	(Laderman, 1976)[15]	23	23	23
(4, 4, 4)	(Strassen, 1969)[2] $(2, 2, 2) \otimes (2, 2, 2)$	49	47	49
(5, 5, 5)	(3, 5, 5) + (2, 5, 5)	98	96	98
(2, 2, 3)	(2, 2, 2) + (2, 2, 1)	11	11	11
(2, 2, 4)	(2, 2, 2) + (2, 2, 2)	14	14	14
(2, 2, 5)	(2, 2, 2) + (2, 2, 3)	18	18	18
(2, 3, 3)	(Hopcroft and Kerr, 1971)[16]	15	15	15
(2, 3, 4)	(Hopcroft and Kerr, 1971)[16]	20	20	20
(2, 3, 5)	(Hopcroft and Kerr, 1971)[16]	25	25	25
(2, 4, 4)	(Hopcroft and Kerr, 1971)[16]	26	26	26
(2, 4, 5)	(Hopcroft and Kerr, 1971)[16]	33	33	33
(2, 5, 5)	(Hopcroft and Kerr, 1971)[16]	40	40	40
(3, 3, 4)	(Smirnov, 2013)[18]	29	29	29
(3, 3, 5)	(Smirnov, 2013)[18]	36	36	36
(3, 4, 4)	(Smirnov, 2013)[18]	38	38	38
(3, 4, 5)	(Smirnov, 2013)[18]	48	47	47
(3, 5, 5)	(Sedoglavic and Smirnov, 2021)[19]	58	58	58
(4, 4, 5)	(4, 4, 2) + (4, 4, 3)	64	63	63
(4, 5, 5)	$(2, 5, 5) \otimes (2, 1, 1)$	80	76	76

Deepmind, “Discovering faster matrix multiplication algorithms with reinforcement learning”, 2022

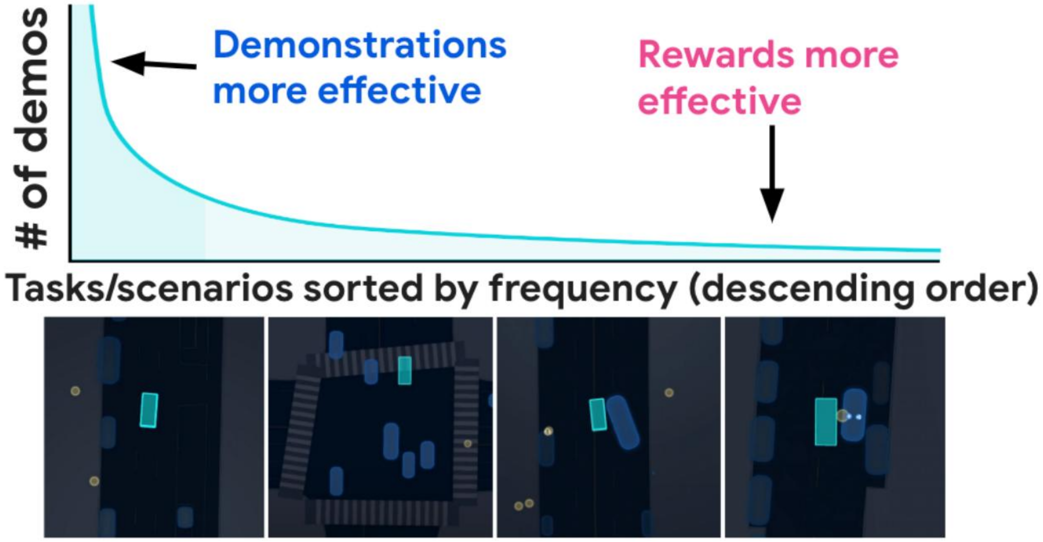
Autonomous Driving

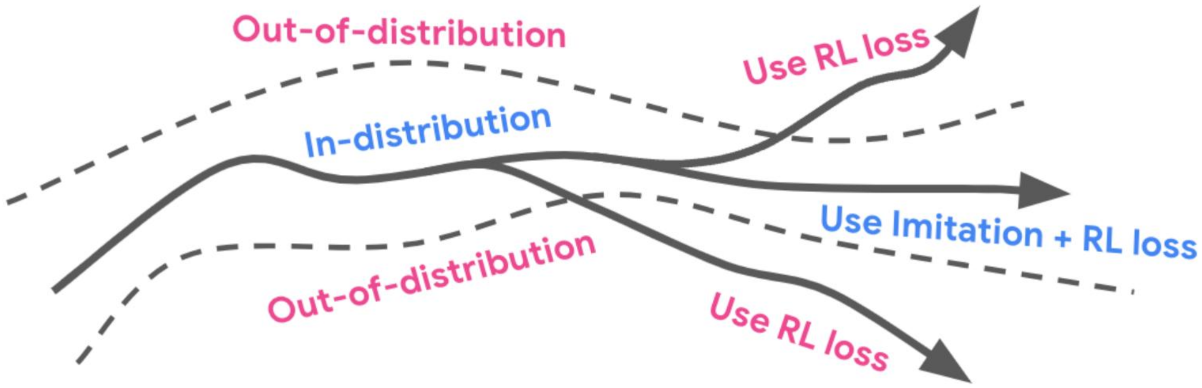
RL in simulators

Self-driving on the road

Amini et al., “VISTA 2.0: An Open, Data-driven Simulator for Multimodal Sensing and Policy Learning for Autonomous Vehicles”, 2021

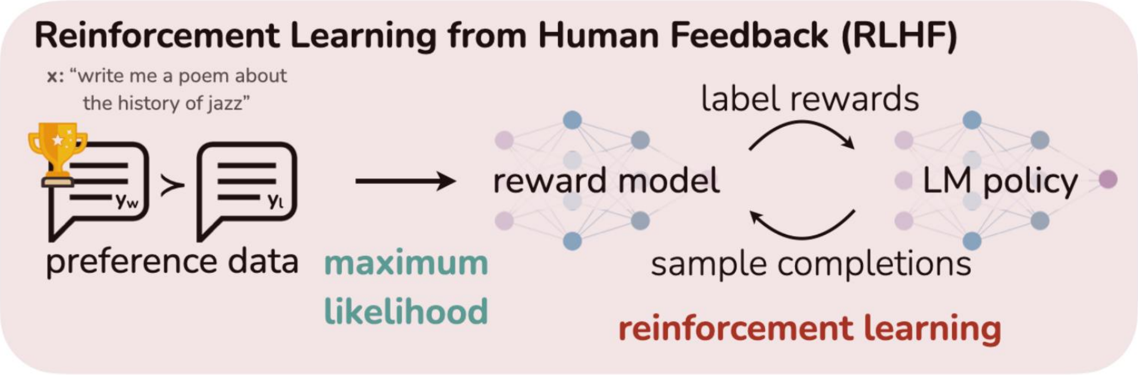
Autonomous Driving



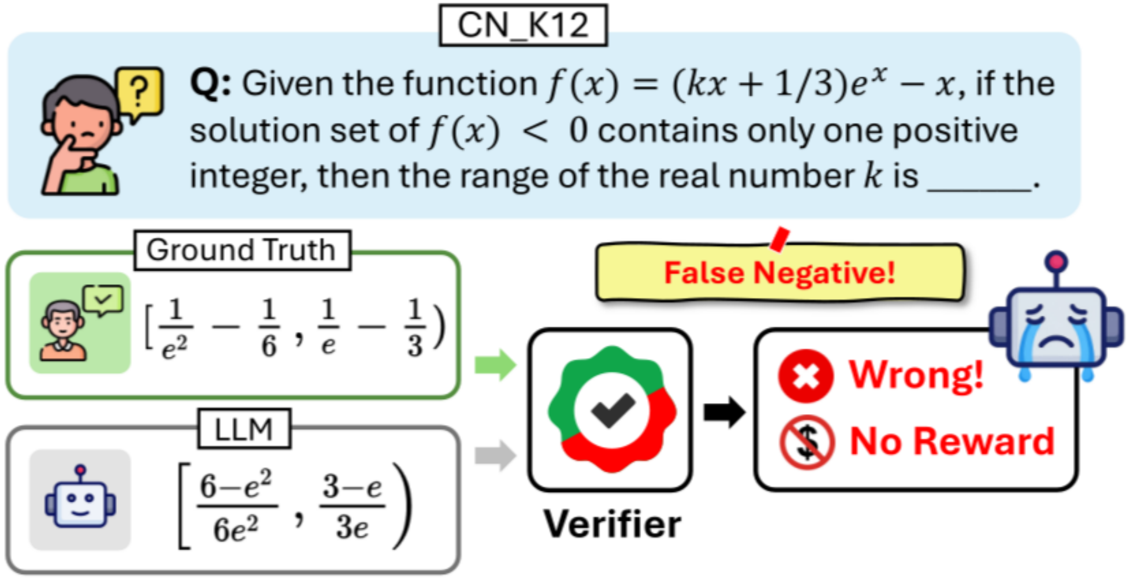


Lu et al., “Imitation Is Not Enough: Robustifying Imitation with Reinforcement Learning for Challenging Driving Scenarios”, 2022

Post-Training Large Language Models



Rafailov et al., "Direct Preference Optimization: Your Language Model is Secretly a Reward Model", 2023



Xu et al. "TinyV: Reducing False Negatives in Verification Improves RL for LLM Reasoning", 2025

Closer Look at Reinforcement Learning

Supervised Learning

$$x \longrightarrow \boxed{f} \longrightarrow y$$

feature **label**

f () = Cat

f (temperature, humidity,...) = 1000mm precipitation

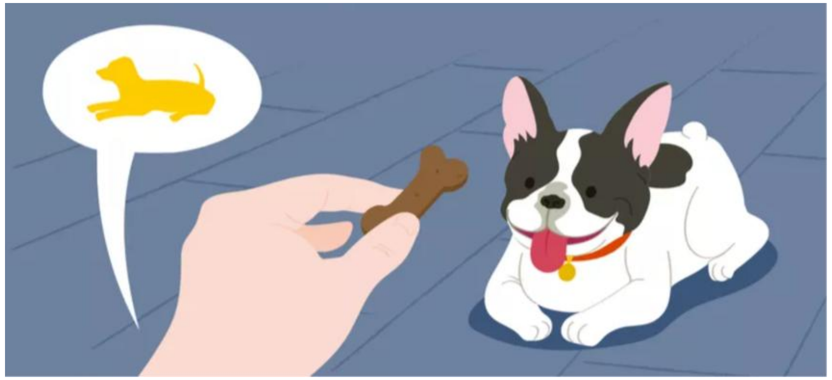
Given a lot of (x, y) pairs, find an f that such that $f(x) \approx y$

Reinforcement Learning

- Reinforce?

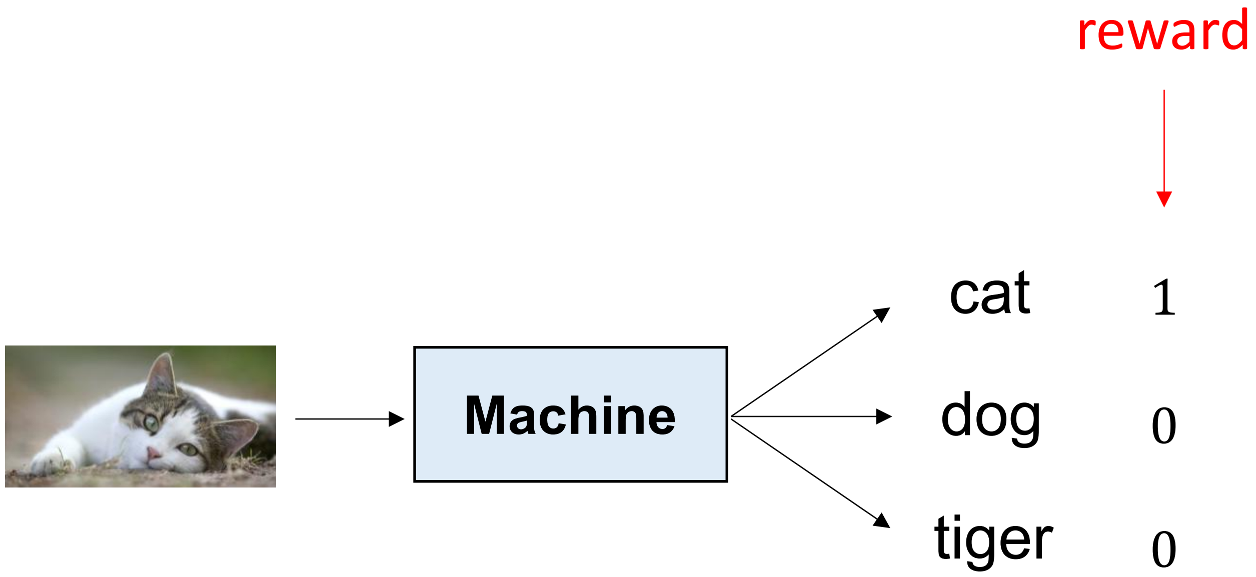
 

- Reinforce?



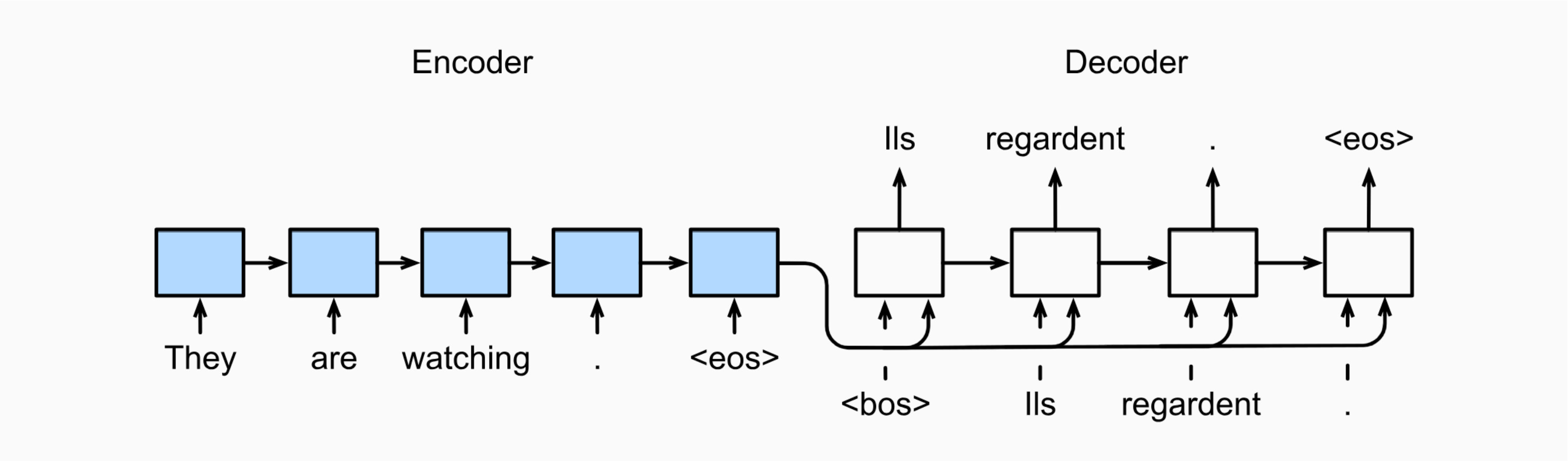
Reinforcement Learning

- Learning from reward feedback?



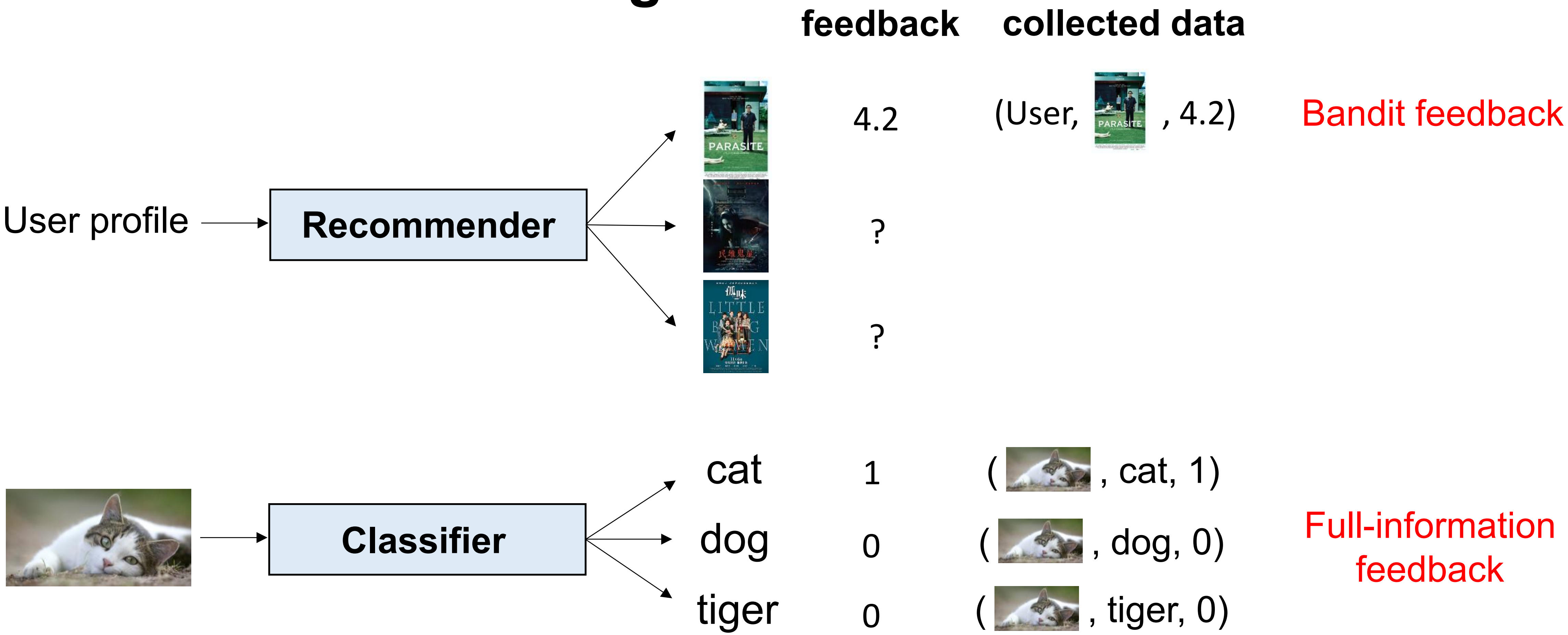
Reinforcement Learning

- Learning sequential decision making?



“Dive into Deep Learning”

Reinforcement Learning

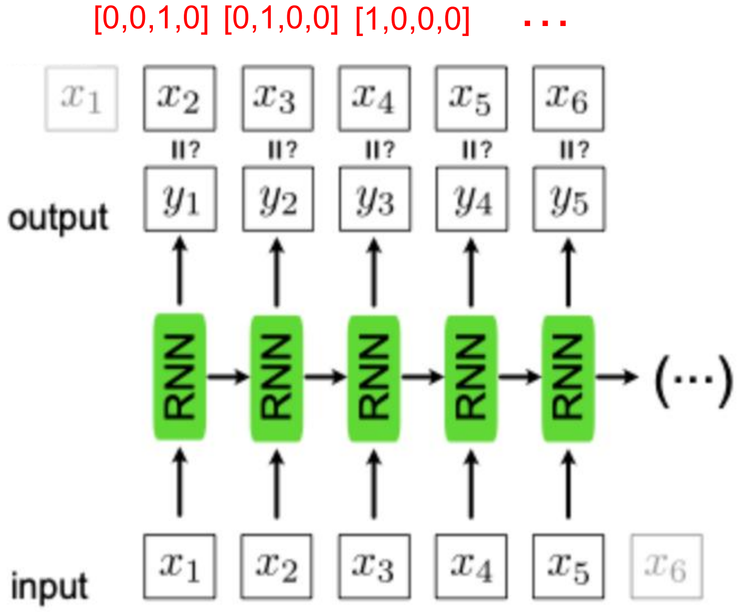


RL usually deals with bandit feedback

Bandit Feedback

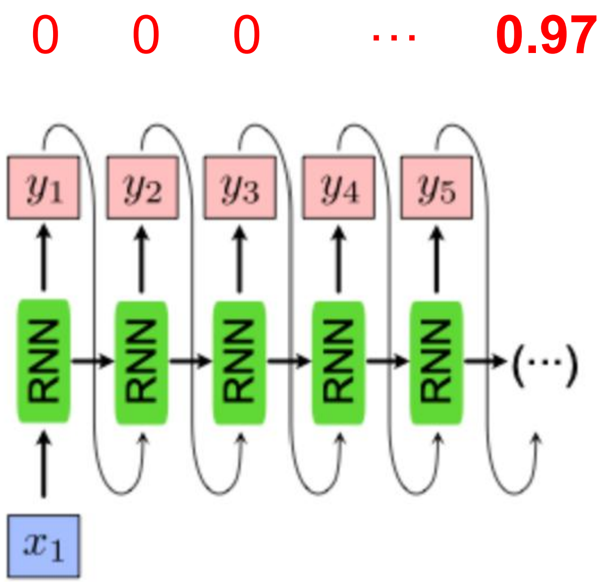
- Needs **exploration**

RL in Sequential Decision Making



(Machine Learning for Scientists)

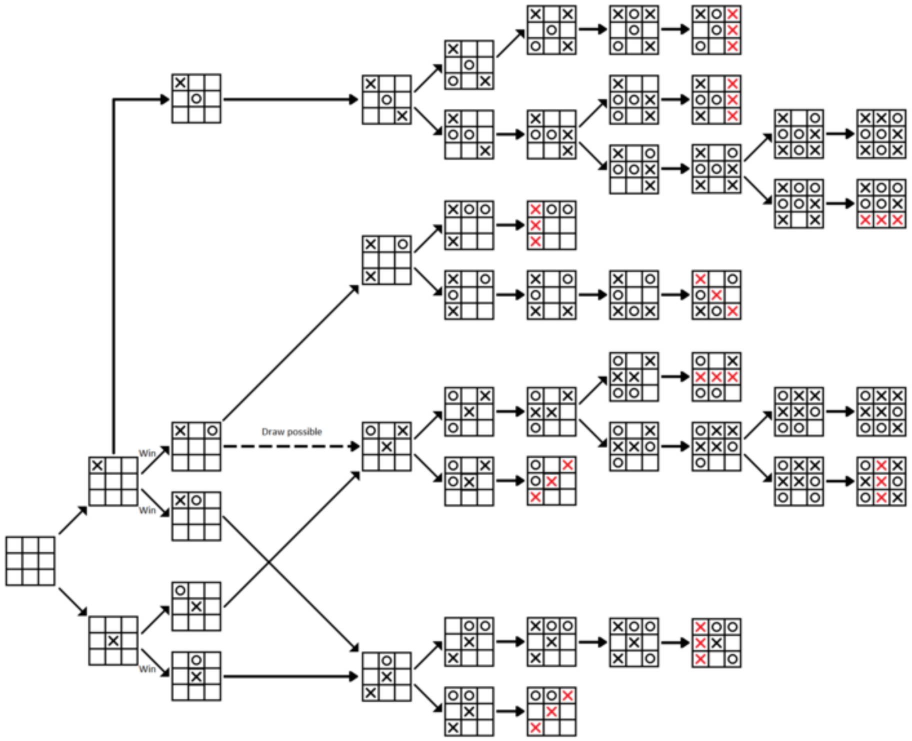
overall score for $(y_1, y_2, \ldots)$



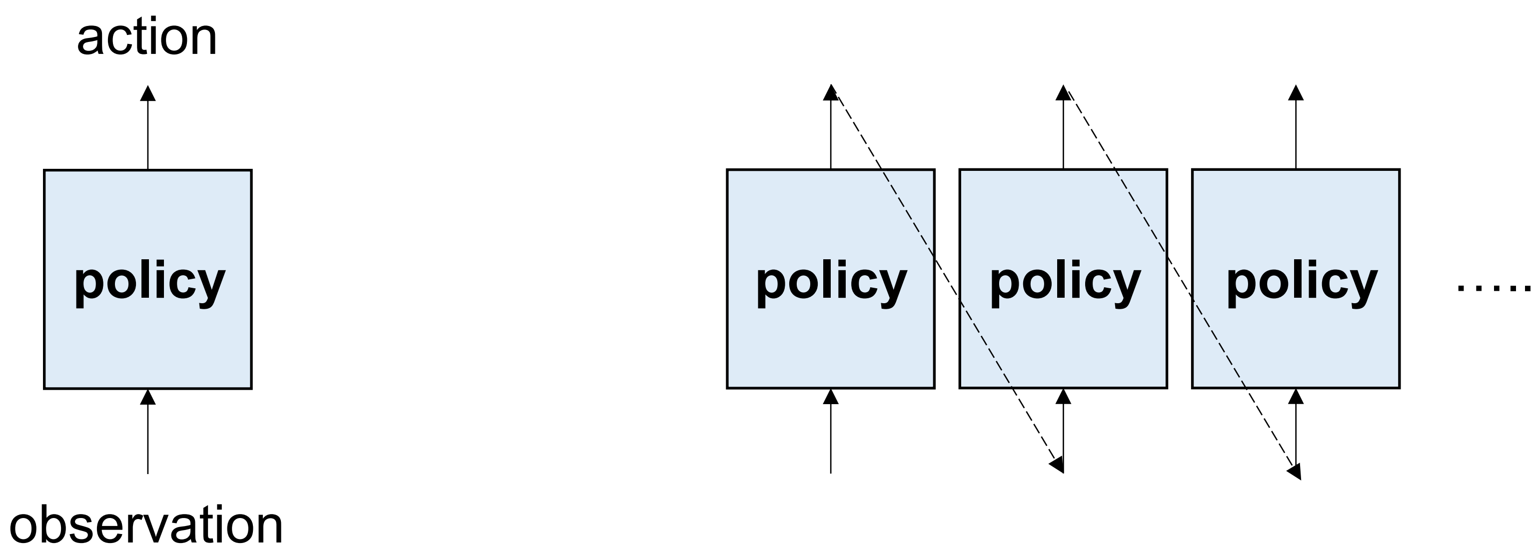
Bandit + **Delayed and Aggregated** Feedback

Delayed and Aggregated Feedback

- Need for **credit assignment**



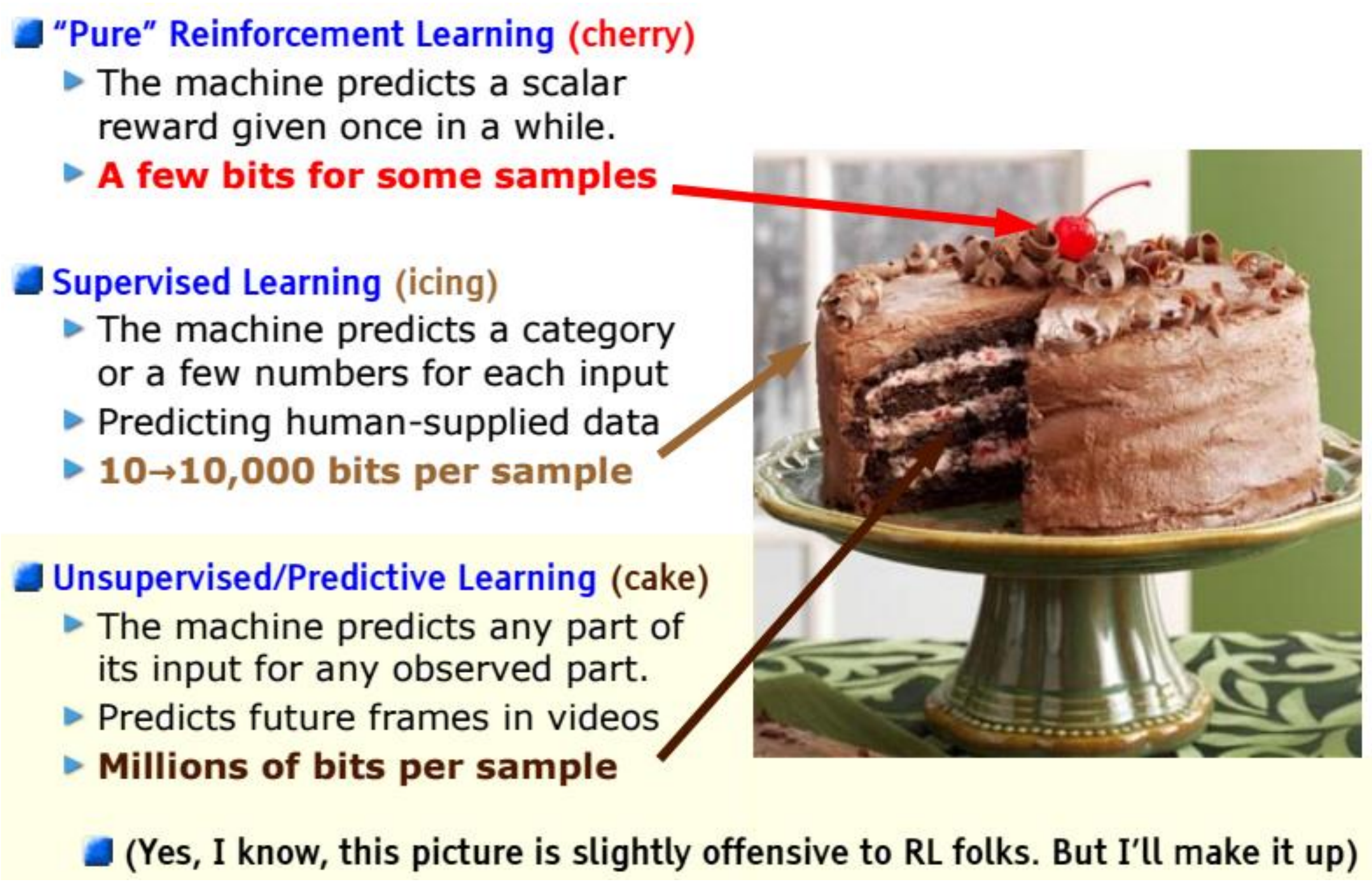
RL vs SL



SL feedback: "what to do in each step" (full-information, immediate)

RL feedback: "how you're doing overall" (bandit, delayed)

RL Signal Can Be Very Sparse



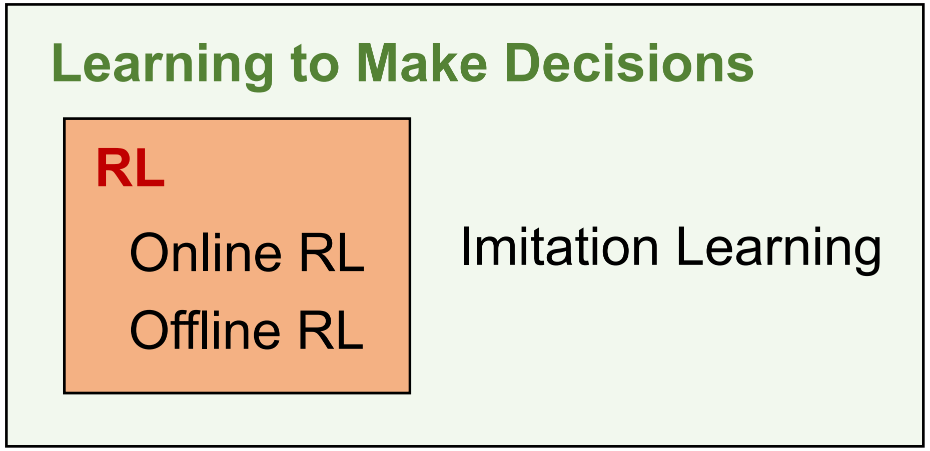
(Yann LeCun, 2016 NIPS)

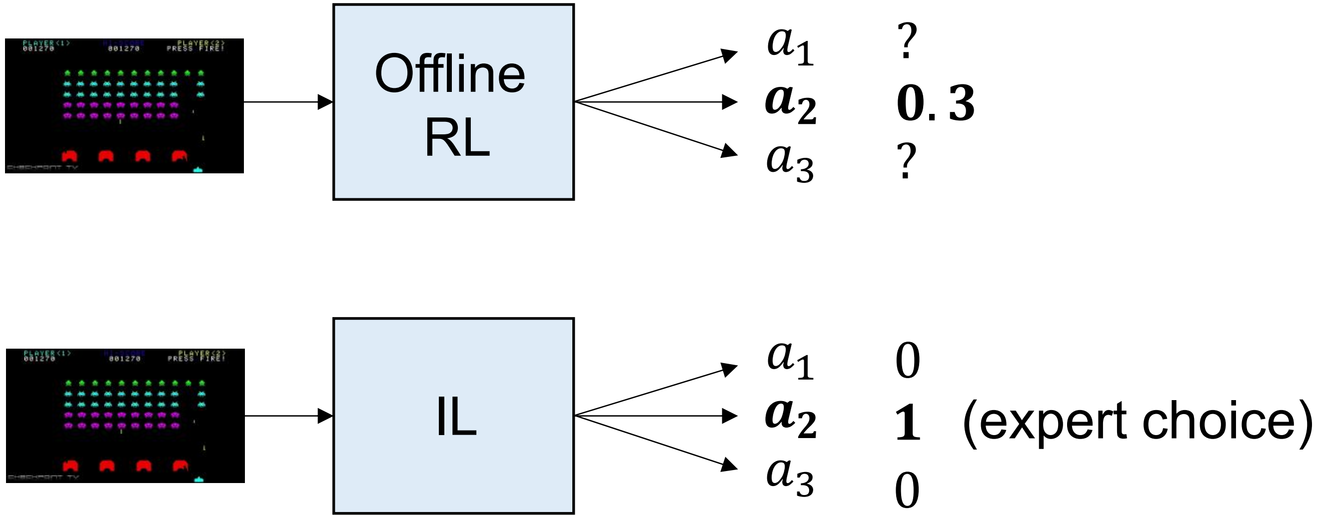
The Scope of This Course

Online RL: through interactions, under bandit / delayed feedback

Offline RL: through existing data, under bandit / delayed feedback

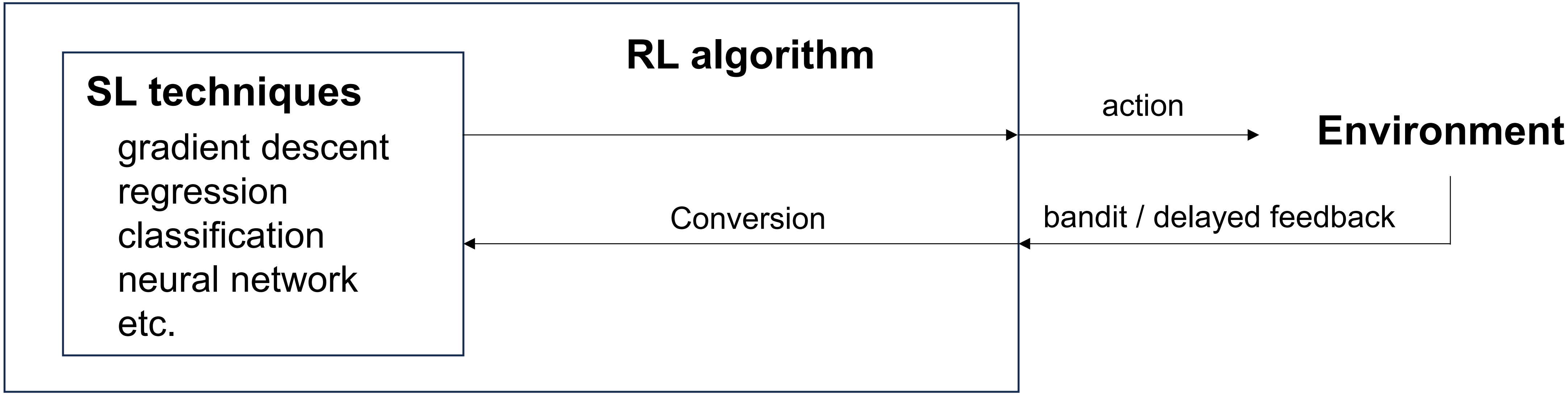
Imitation Learning: through expert data, under label feedback





"RL" in This Course

- Set of techniques to handle bandit / delayed feedback in machine learning.
- A common approach (not always)
  - Reuse / reduce to supervised learning techniques



When Is IL (SL) Insufficient?

- The truly best policy is unknown / expert is imperfect
  - Atari game, Go
  - Faster matrix multiplication

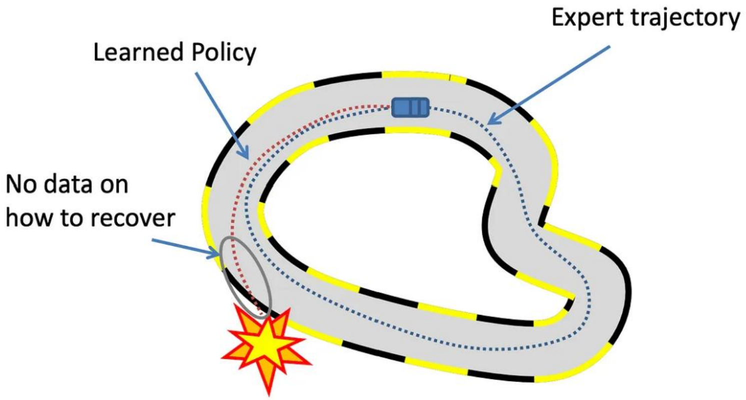
  ⇒ RL can **search** for better solutions

- RL signal may more faithfully reflect our real objective
  - RL from Human Feedback

  ⇒ RL can provide **alignment** to the real objective

- The expert data has limited coverage
  - Autonomous driving

  ⇒ RL can explore edge cases and **robustify** solutions



Challenges in RL

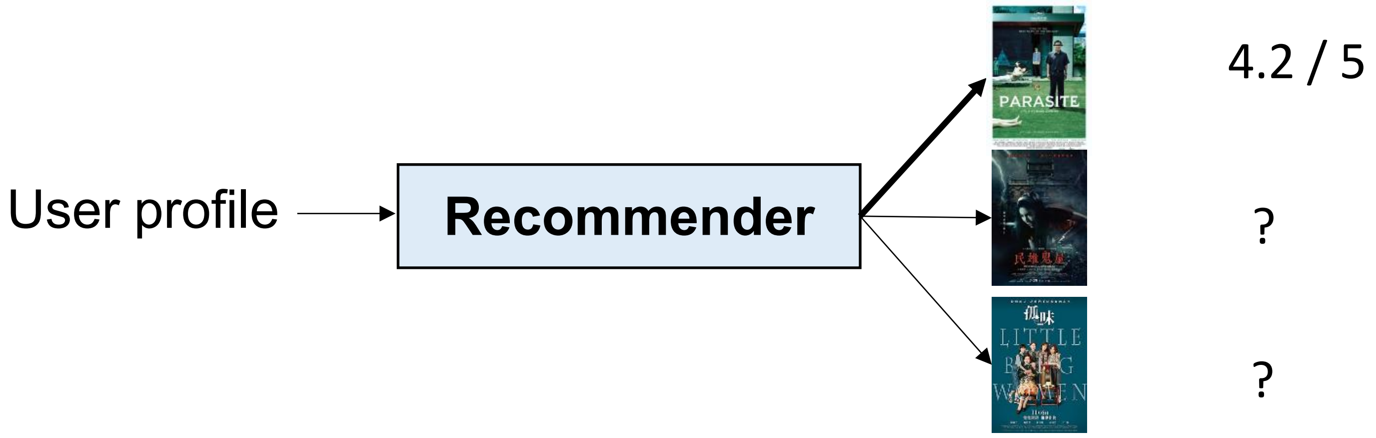
Challenges in RL (1)

Generalization: a key challenge in all machine learning paradigms

(Khosravian and Amirkhani, 2022)

Challenges in RL (2)

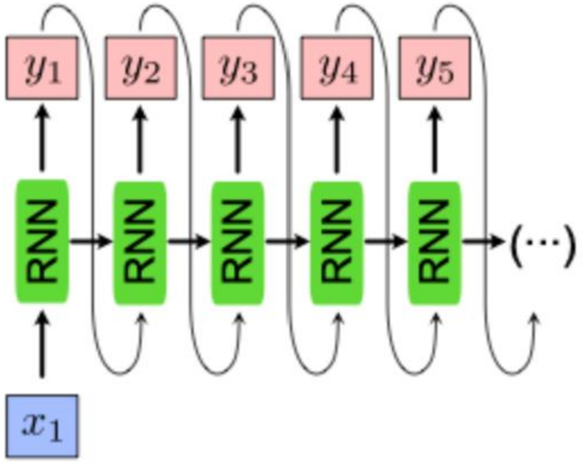
Exploration and exploitation tradeoff (due to bandit feedback)

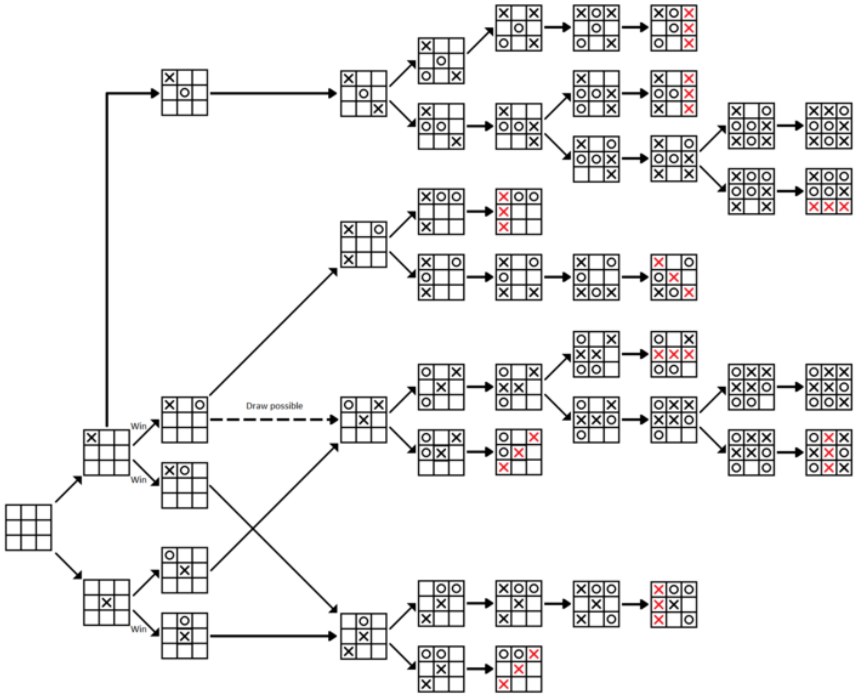


Challenges in RL (3)

Credit assignment (due to delayed and aggregated feedback)

0 0 0 $\cdots$ **0.97**

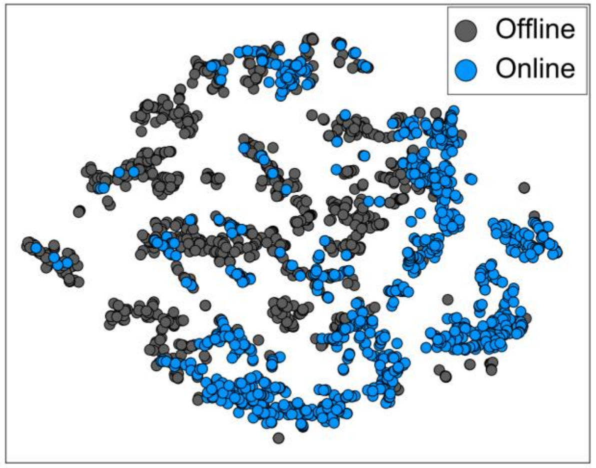


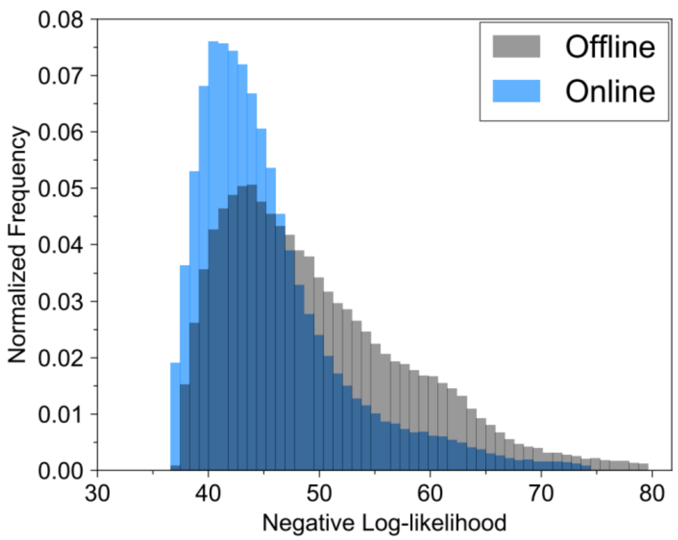


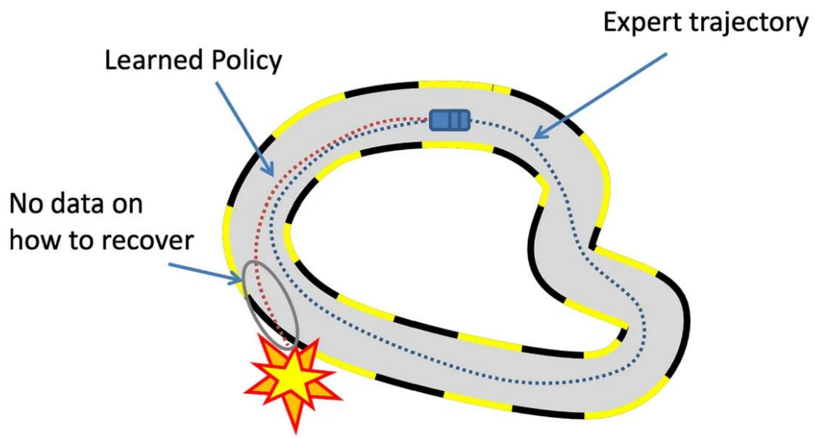
Identify the contribution of each action to the outcome

Challenges in RL (4)

Distribution mismatch / shift (especially in offline RL)







Lee et al., Addressing Distribution Shift in Online Reinforcement Learning with Offline Datasets

Other Challenges

- Reward design
- Safety
- Robustness under attacks

…