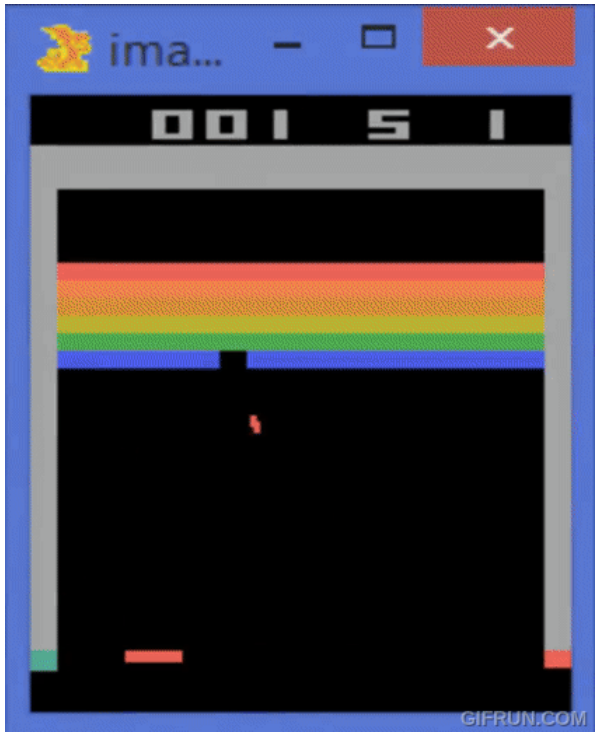
Introduction to the Course

Chen-Yu Wei

Learning To Make Decisions from Interactions

Games

10 mins training

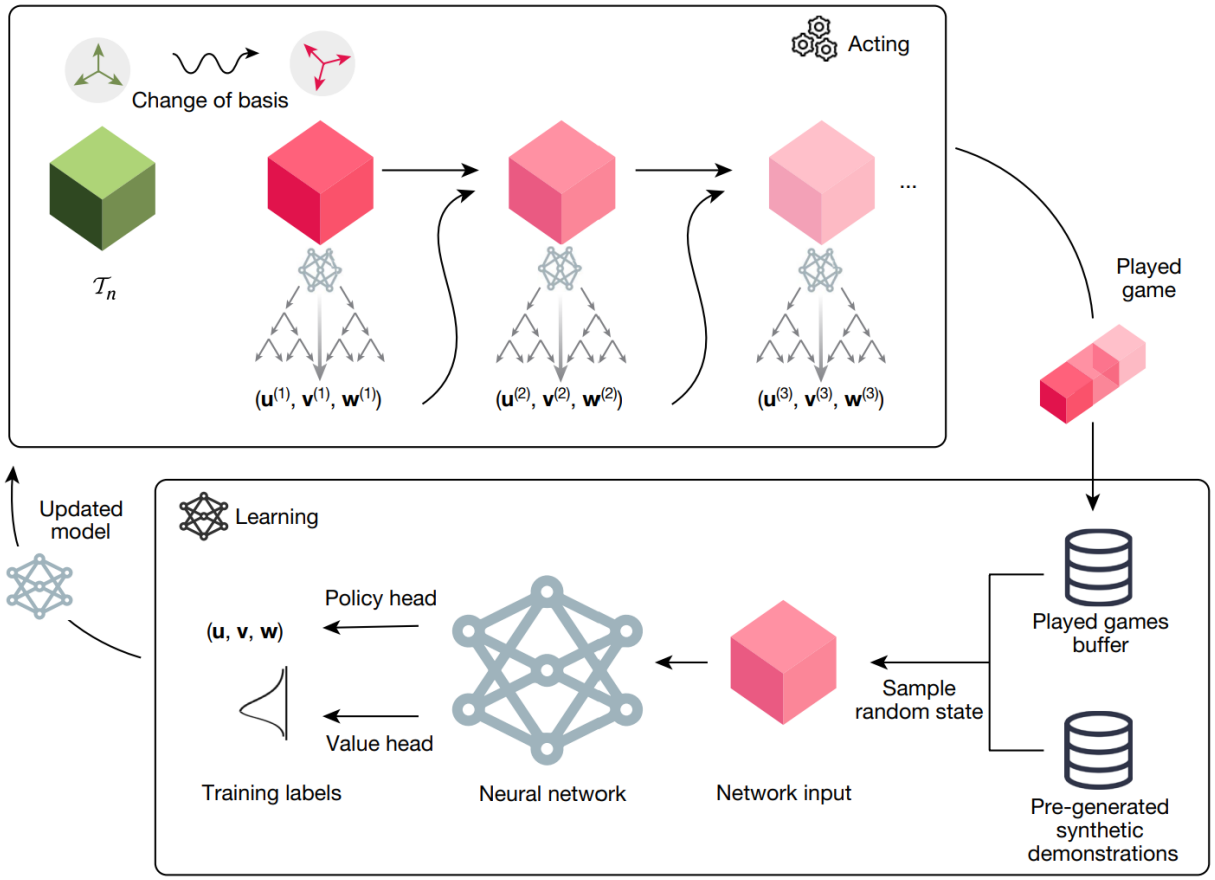


120 mins

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	AlphaTensor rank 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) ⊗ (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) ⊗ (2, 1, 1)	80	76	76

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

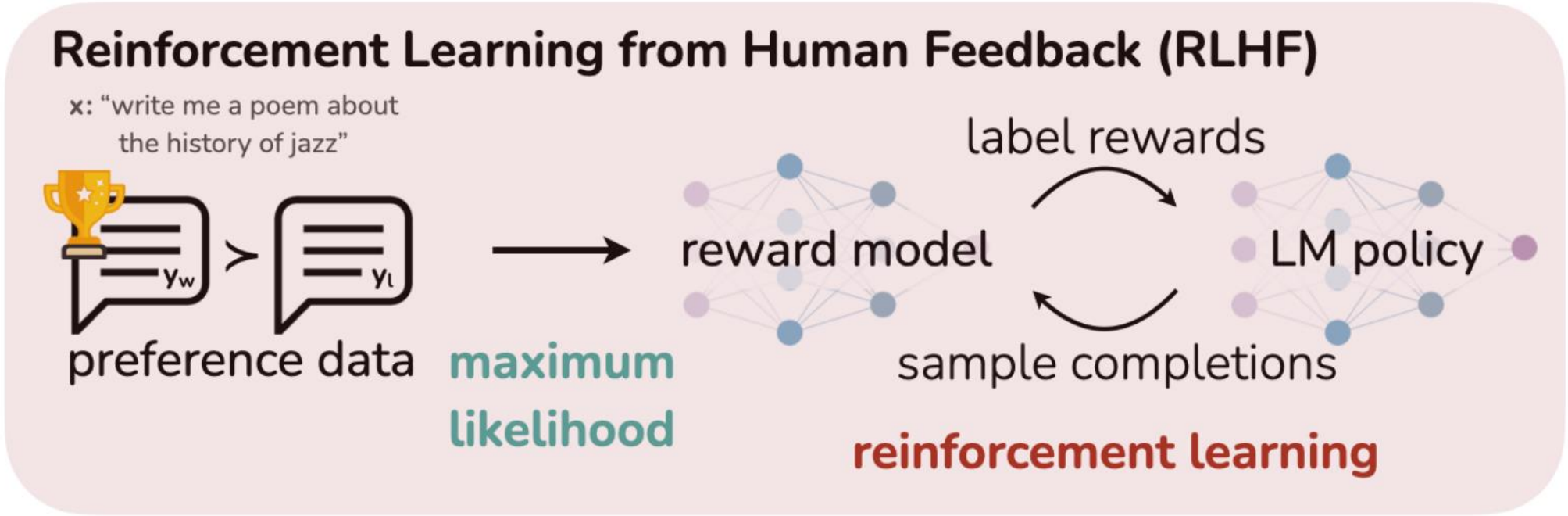
Autonomous Driving

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

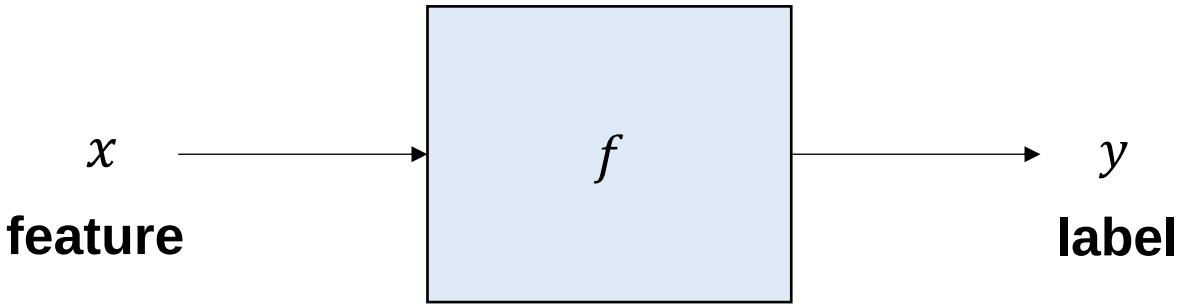
Languages



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

Closer Look at Reinforcement Learning

Supervised Learning



f () = Cat

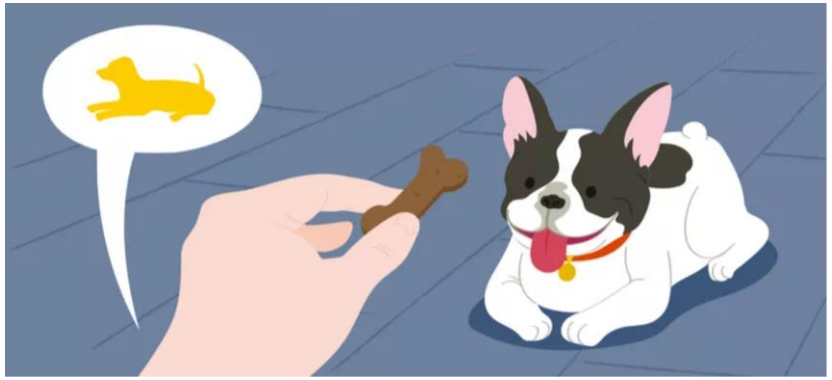
f (temperature, humidity,...) = 1000mm precipitation

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

Reinforcement Learning

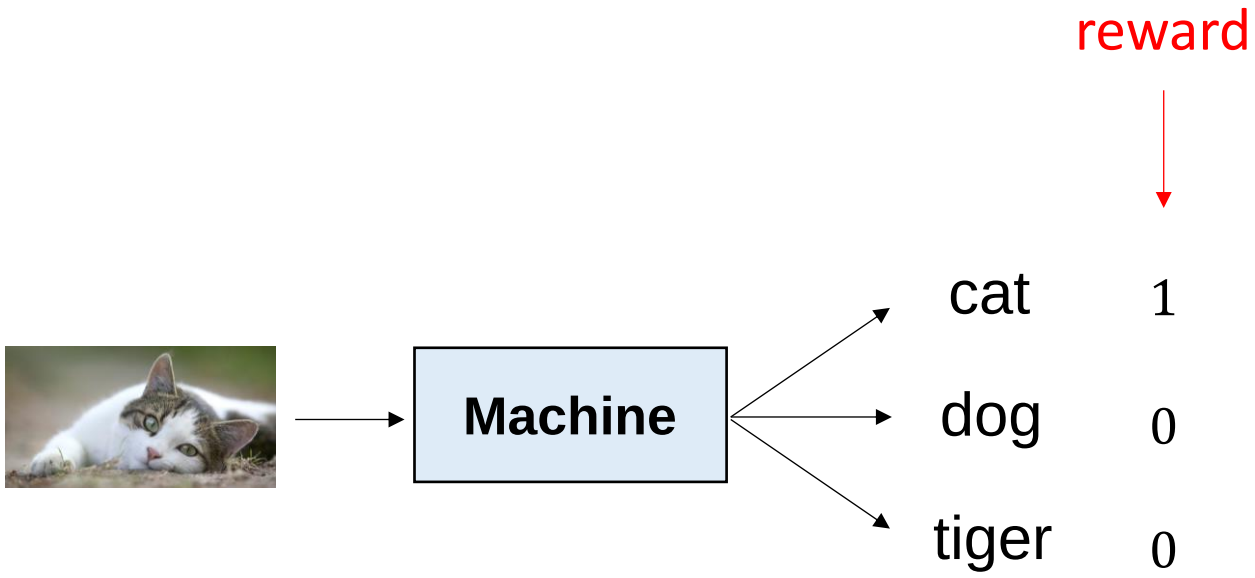
- Reinforce?

- Reinforce?



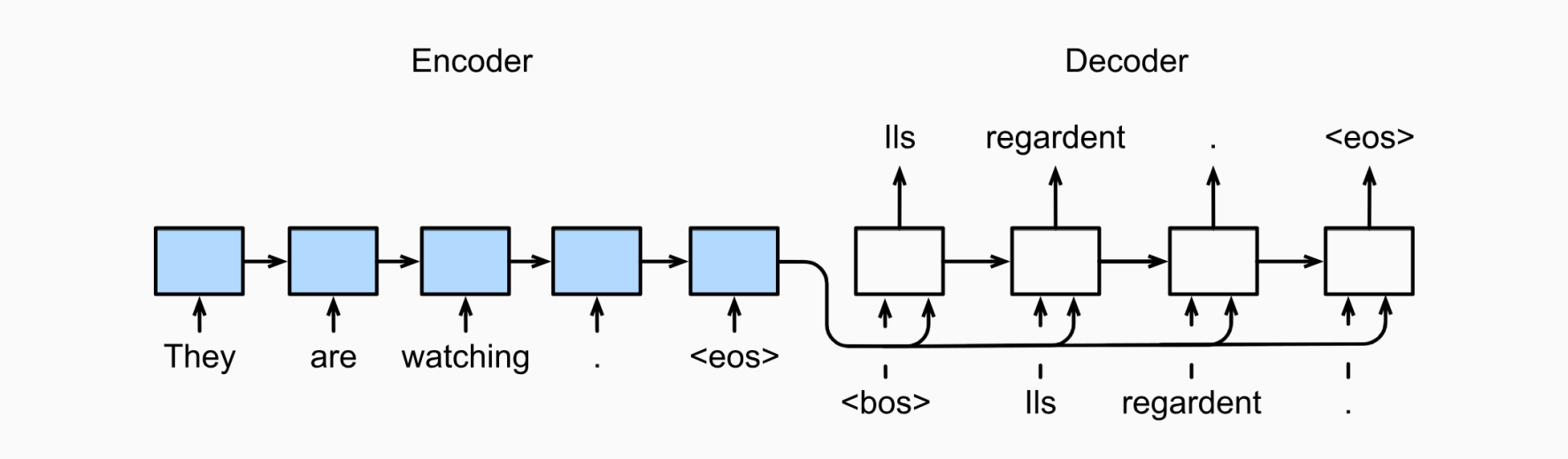
Reinforcement Learning

- Learning from reward feedback?



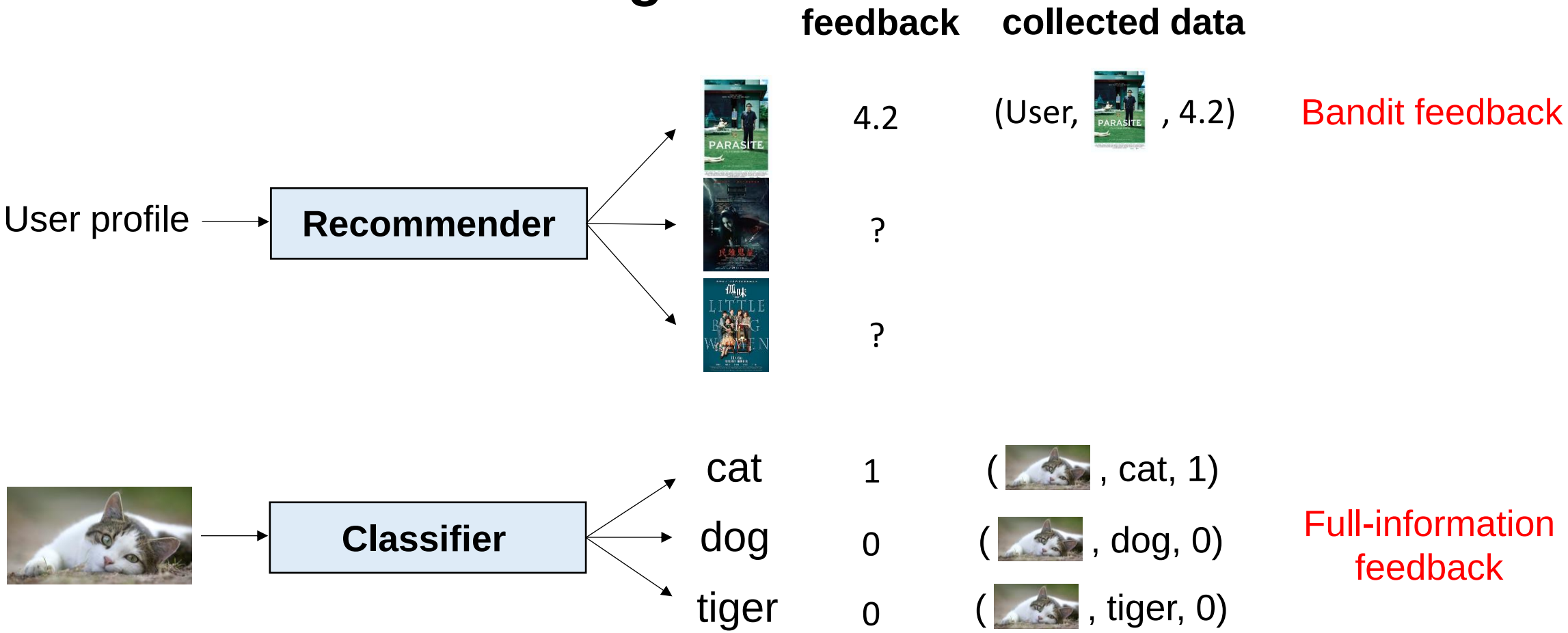
Reinforcement Learning

- Learning sequential decision making?



"Dive into Deep Learning"

Reinforcement Learning

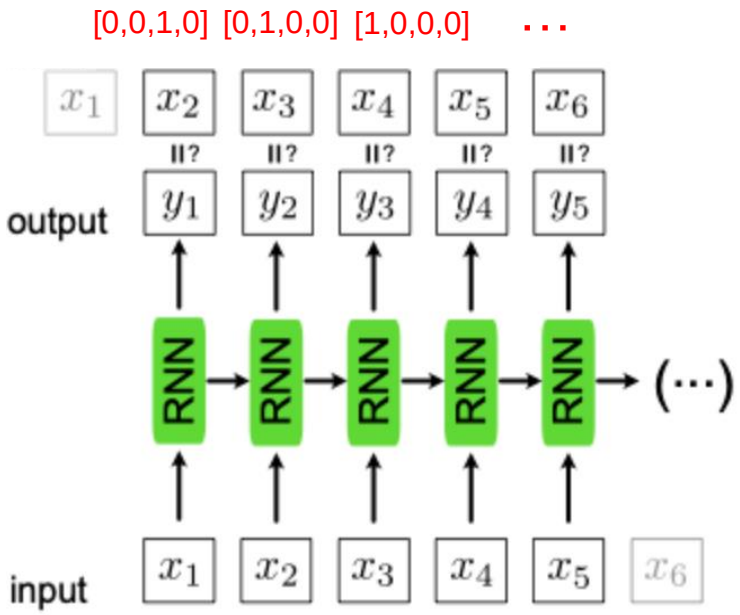


RL usually deals with bandit feedback

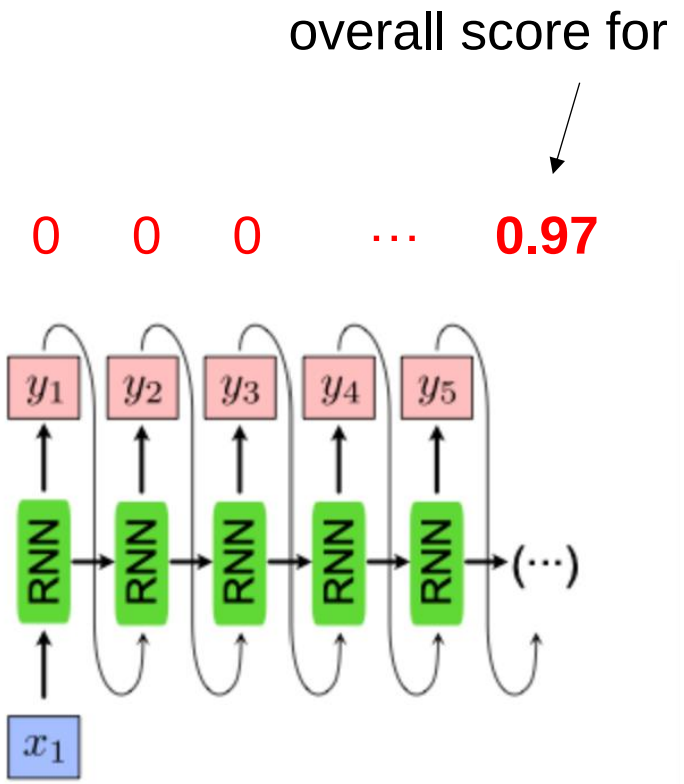
Bandit Feedback

- Needs **exploration**

RL in Sequential Decision Making



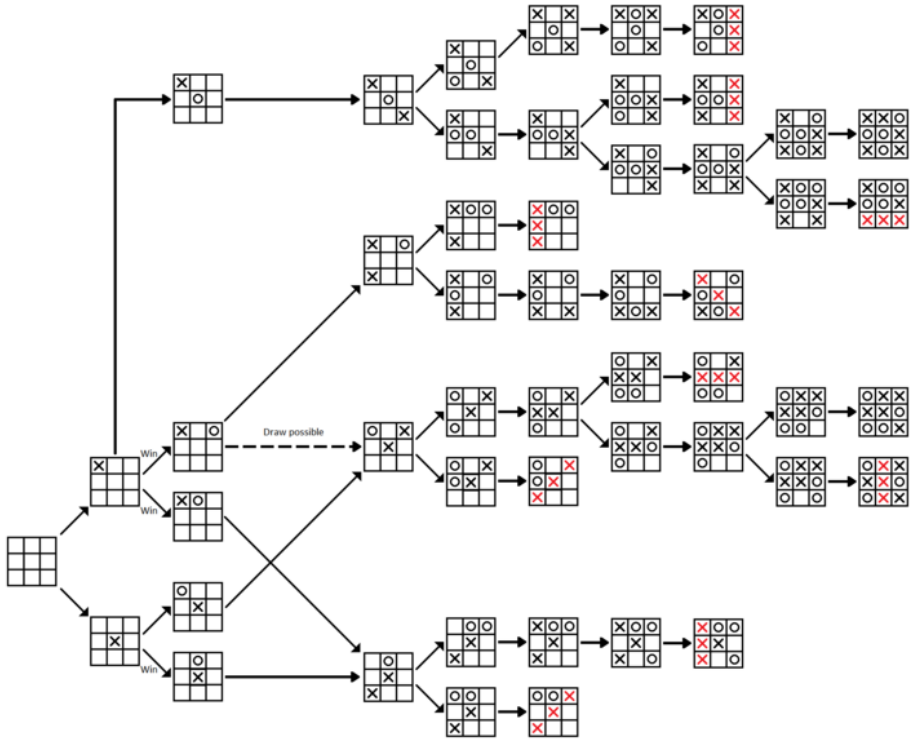
(Machine Learning for Scientists)



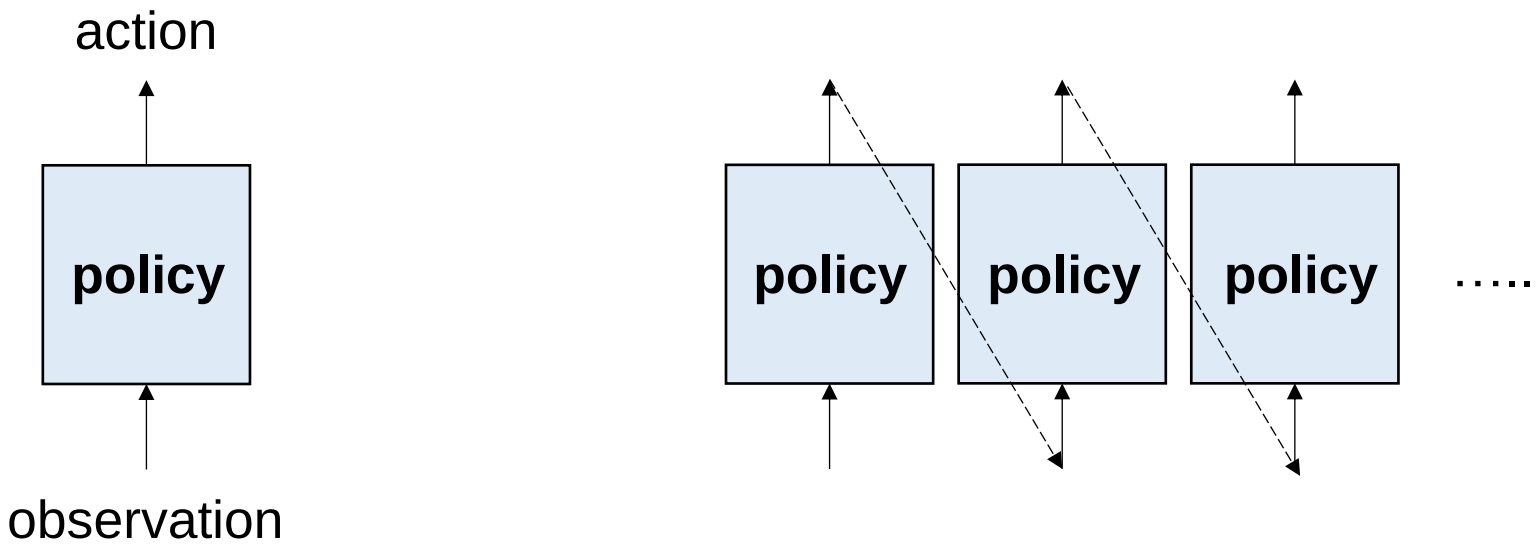
Bandit + **Delayed and Aggregated** Feedback

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

- "Pure" Reinforcement Learning (cherry)
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- Supervised Learning (icing)
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- Unsupervised/Predictive Learning (cake)
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

- (Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)

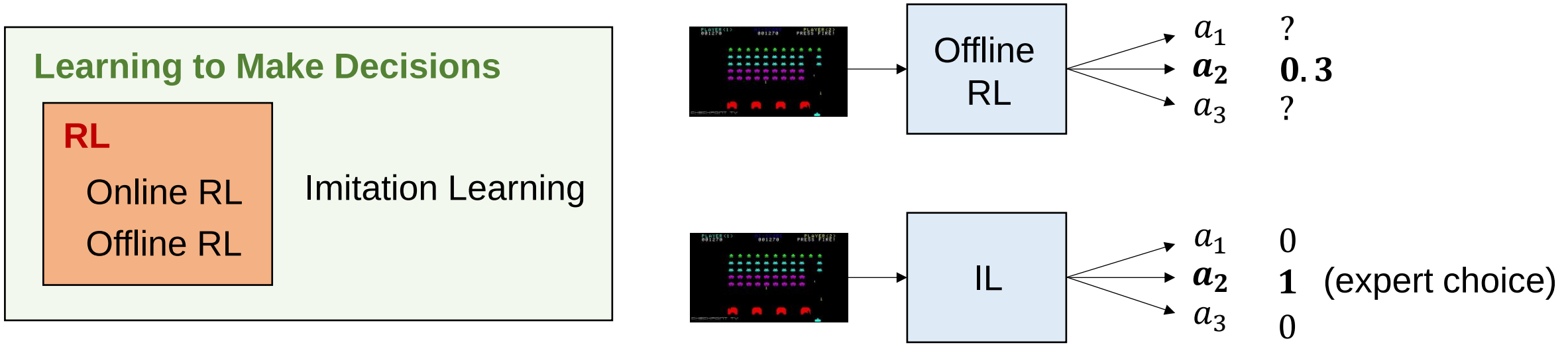
(Yann LeCun, 2016 NIPS)

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 (not in our scope)



When Is IL (SL) Insufficient?

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

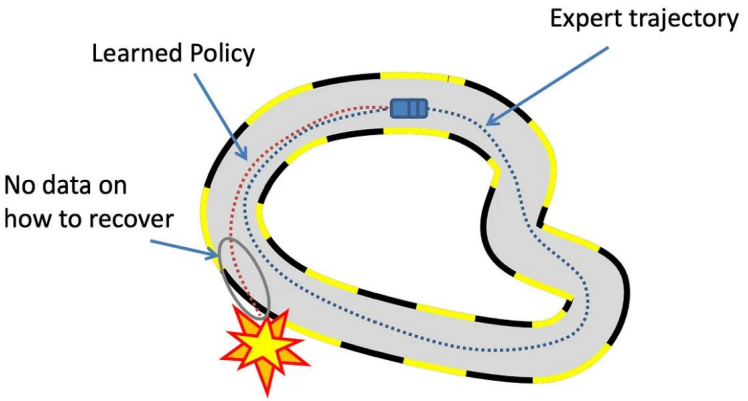
  ⇒ RL can **search** for better solutions

- 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

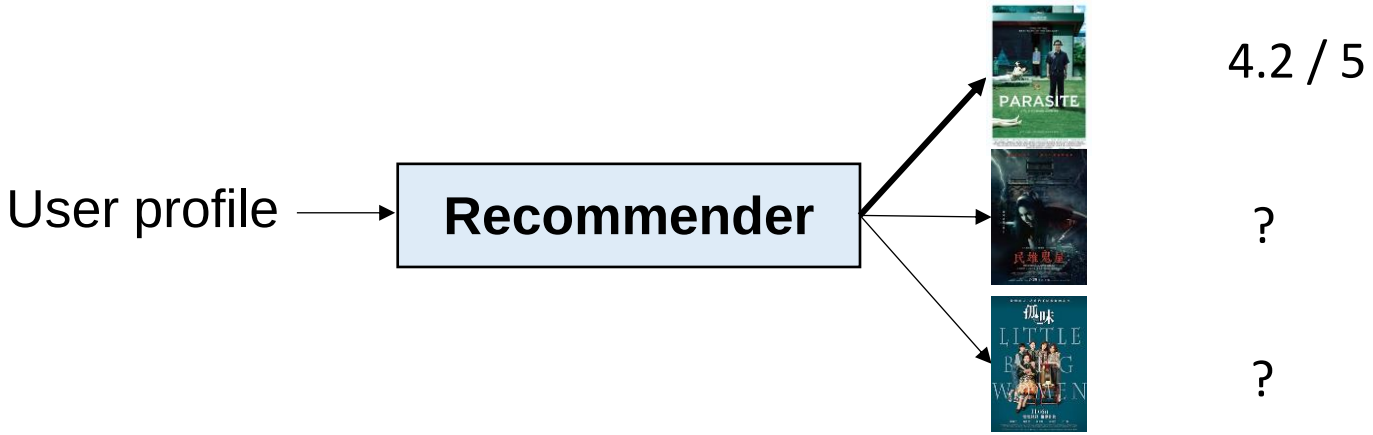
Challenges in RL (1)

Generalization: a key challenge in all machine learning paradigms

(Khosravian and Amirkhani, 2022)

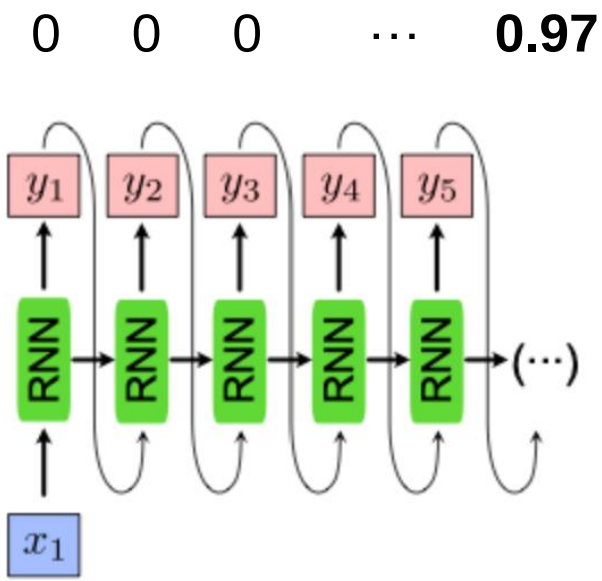
Challenges in RL (2)

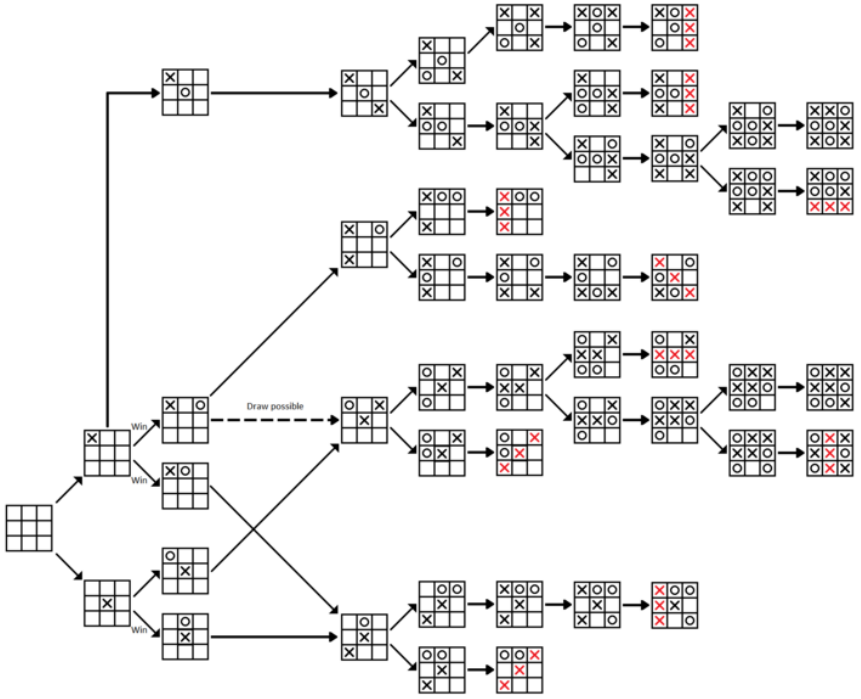
Exploration and exploitation tradeoff (due to bandit feedback)



Challenges in RL (3)

Credit assignment (due to delayed and aggregated feedback)

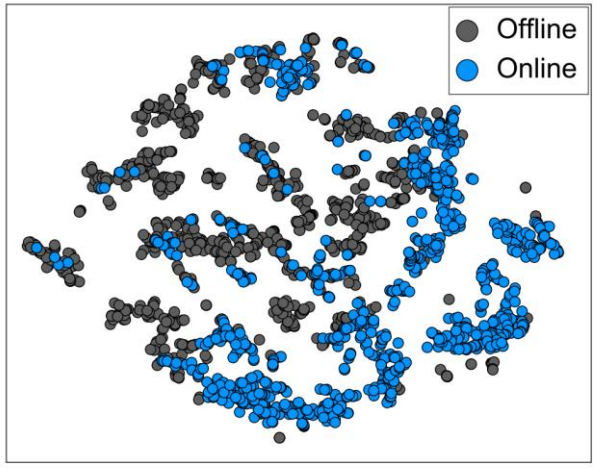


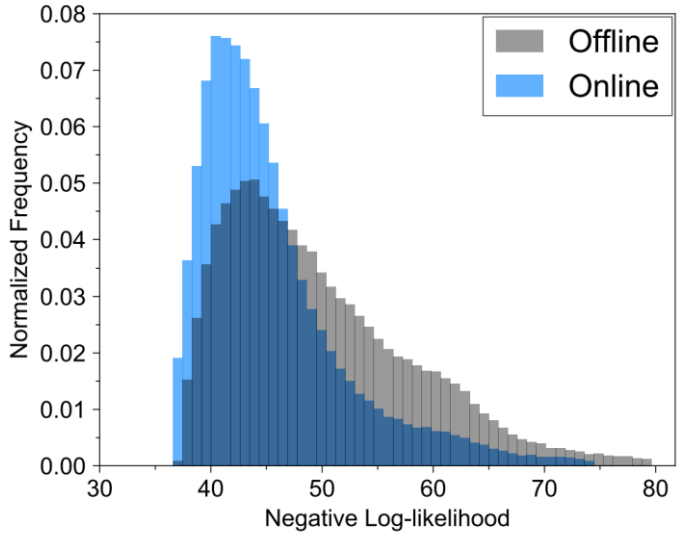


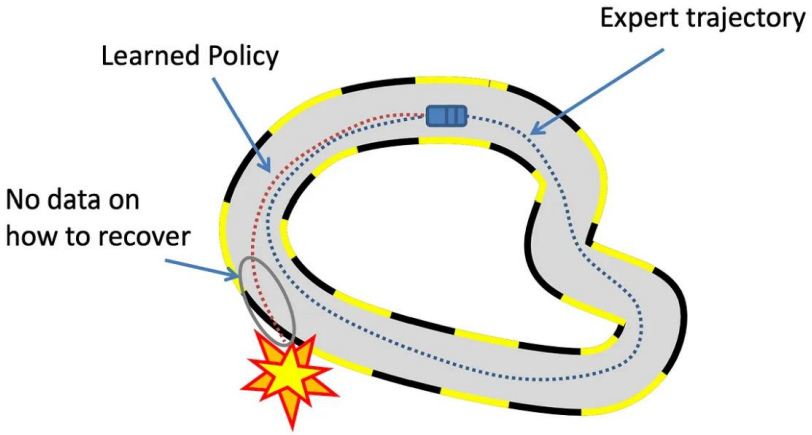
Identify the contribution of each action to the outcome

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

…

Course Content

Platforms

- Course website: https://bahh723.github.io/rl2025sp/
  - Syllabus, announcement, slides, (lecture recordings)
  - Can be accessed from Lou's List or my personal website
- Gradescope
  - Homework submission
- Piazza
  - Questions and discussions

Course Content

(Focusing on exploration-exploitation tradeoff)

Part I. Learning in Bandits

- Multi-armed bandits
- Linear bandits
- Contextual bandits
- Adversarial multi-armed bandits
- Adversarial linear bandits

Part II. Basics of MDPs

- Bellman (optimality) equations
- Value iteration
- Policy iteration

(Focusing on credit assignment and distribution mismatch)

Part III. Learning in MDPs

- Approximate value iteration and variants
  - Least-square value iteration
  - Q-Learning
  - DQN
- Policy evaluation
  - Temporal difference
  - Monte Carlo
- Approximate policy iteration and variants
  - Least-square policy iteration
  - (Natural) policy gradient and actor-critic
  - REINFORCE, A2C, PPO
  - DDPG, SAC

(Focusing on distribution mismatch)

Part IV. Offline RL

Student Project Presentation

Theme of This Course

- The **math** behind **basic** RL algorithms

- The course might be more helpful if the goal is to learn
  - Underlying theory and principles of basic RL
  - Mathematical tools for analyzing ML algorithms

- The course might less helpful if the goal is to learn
  - Advanced topics, e.g., multi-agent RL, distributional RL, hierarchical RL
  - Many practical tricks in RL implementation

Prerequisites

- Linear Algebra, Probability, Calculus
- (Optional but helpful) Machine Learning, Convex Optimization
- Python

Try to work on HW0 **(No submission needed)**.

- Test your understanding for the prerequisites
- May lightly use google
- Consult me or reconsider taking the course if you're stuck in ≥ 2 problems

Online Resources

- Courses
  - UC Berkeley CS285
  - DeepMind x UCL RL Lectures
- Courses (theoretical ones)
  - Csaba Szepesvari
  - Nan Jiang, Wen Sun, Chi Jin
  - Dylan Foster & Sasha Rakhlin
- Books
  - Sutton & Barto, Reinforcement Learning: An Introduction
  - Agarwal et al., Reinforcement Learning: Theory and Algorithms
  - Lattimore & Szepesvari, Bandit Algorithms (bandit)
- Implementations
  - OpenAI SpinningUp
  - OpenAI StableBaseline3
  - ShangtongZhang

Assignments (70%)

- **4 written assignments (40%)**
  - Math / algorithm design problems
  - Latex OR hand-writing + taking photo

- **3 programming assignments (30%)**
  - Programming tasks (using PyTorch)
  - PyTorch tutorial: https://www.youtube.com/watch?v=c36lUUr864M

- 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

Final Project (30%)

- Breakdown
  - Proposal (5%)
  - Midterm report (5%)
  - Presentation (10%)
  - Final report (10%)
- Types of projects (basically any)
  - Application, algorithm design, systematic comparison, theoretical understanding, survey…
- **Goal:** Apply RL techniques to problems you're interested in.
- You're welcome to build it on existing project
  - Describe in the proposal the current status of the project
- 1-3 students in a group
- Proposal deadline: **Feb.16**

TA & Office Hour

- **TA:** Braham Snyder
  - Email: dqr2ye@virginia.edu
  - Office hour: Friday 4-5pm at Rice 442
- Me
  - Email: chenyu.wei@virginia.edu
  - Office hour: Tuesday 3:30-4:30pm at Rice 409

Questions?