Reinforcement Learning & Markov Decision Processes

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning

Fall 2023

Today

- Logistics (Welcome!)
- Overview of RL
- Markov Decision Processes
    - Problem statement
    - Policy Evaluation

Course staff introductions

- **Instructors:** Lucas Janson and Sham Kakade
- **TFs:** Benjamin Schiffer
- **CAs:** Luke Bailey, Alex Dazhen Cai, Kevin Yee Du, Kevin Yifan Huang, Saket Joshi, Thomas Kaminsky, Patrick McDonald, Eric Meng Shen, Natnael Mekuria Teshome
- Homework 0 is posted today!
  - This is “review” homework for material you should be familiar with to take the course.

Course Overview

All policies are stated on the course website:
https://shamulent.github.io/CS_Stat184_Fall23.html

- We want u to obtain fundamental and practical knowledge of RL.
- **Grades: Participation; HW0 +HW1-HW4; Midterm; Project**
- Participation (5%): not meant to be onerous (see website)
  - Just attending regularly will suffice (tbd: we'll have some web form per class)
  - If you can't, then increase your participation in Ed/section.
  - Let us know if you some responsibility, let us know via Ed.
- HWs (45%): will have math and programming components.
  - We will have an "embedded ethics lecture" + assignment
- Midterm (20%): this will be in class. Date to be finalized soon.
- Project (30%): 2-3 people per project. Will be empirical.

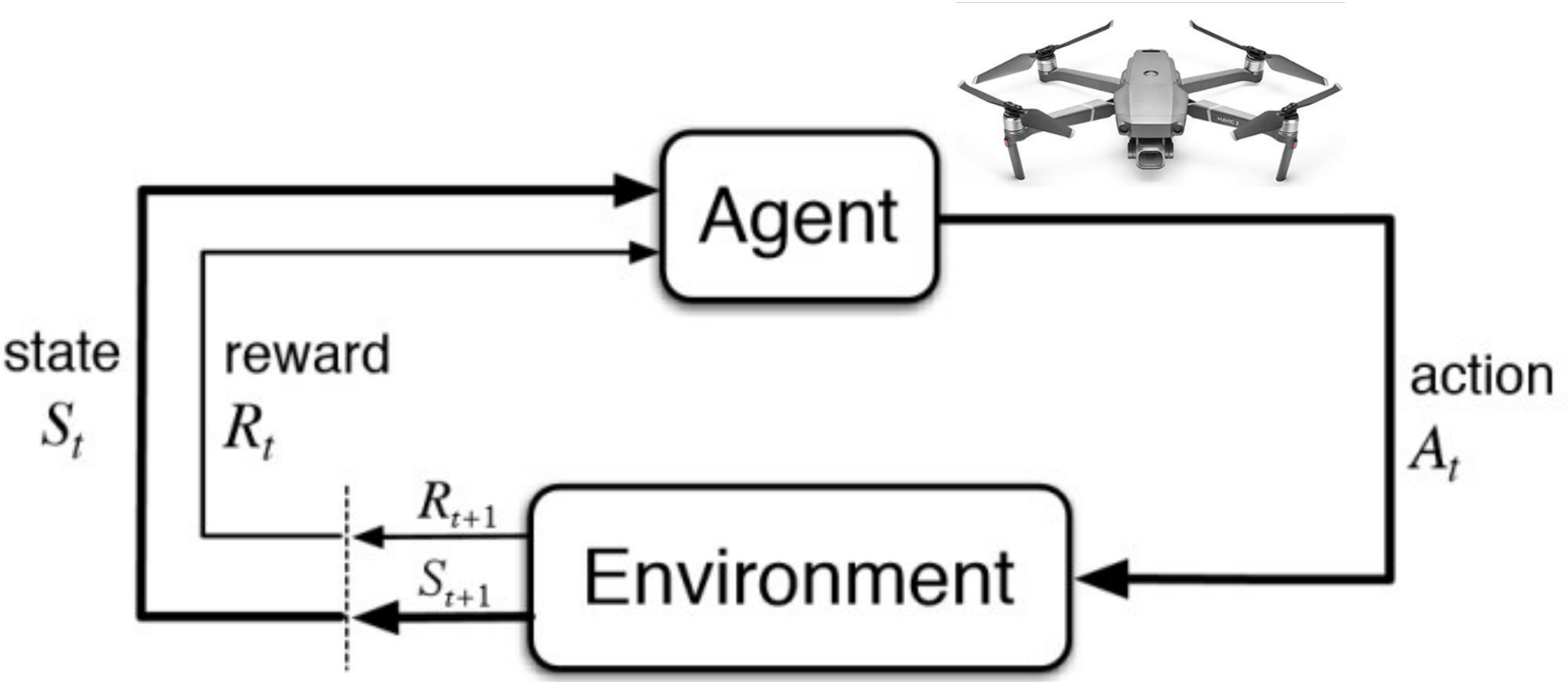
Other Points

- Our policies aim for consistency among all the students.
- Participation: we will have a web-based attendance form (TBD)
- Communication: please only use Ed to contact us
- Late policy (basically): you have 96 cumulative hours of late time.
  - *Please use this to plan for unforeseen circumstances.*
- Regrading: ask us in writing on Ed within a week

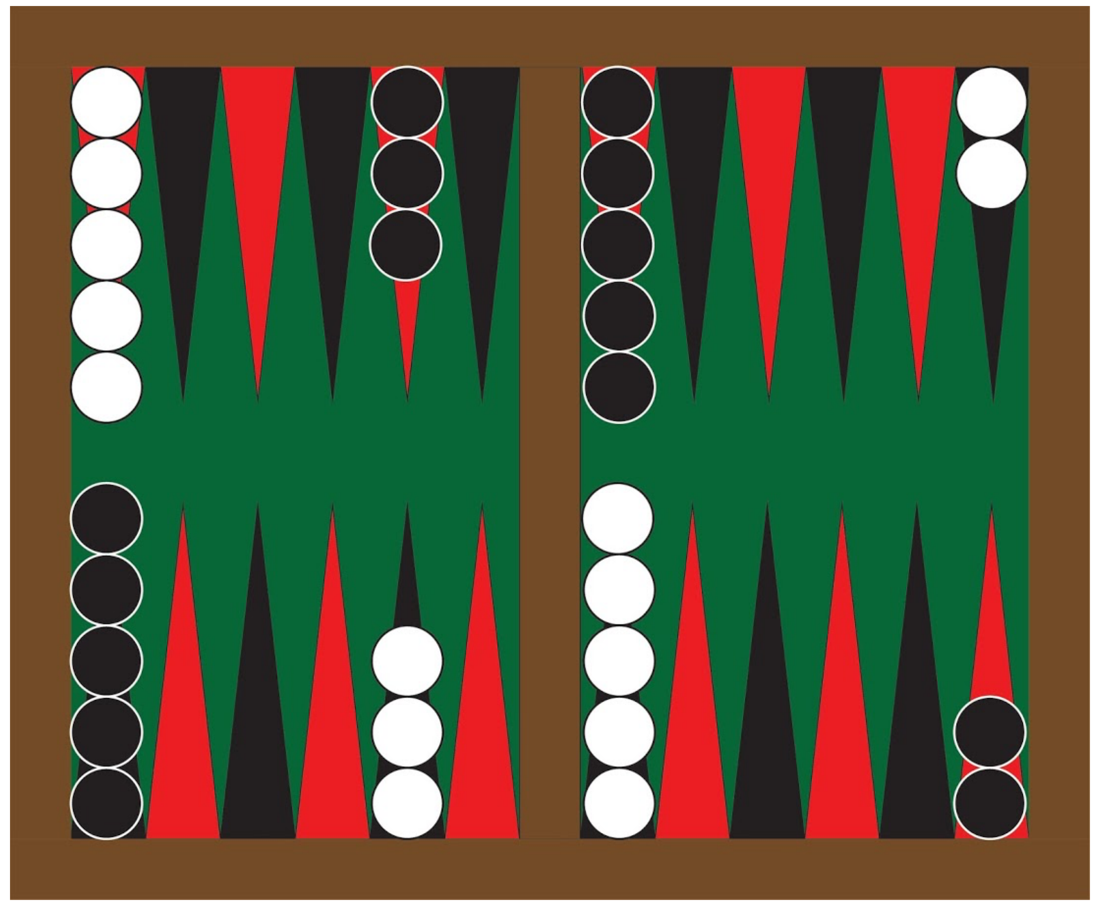
Today

- Logistics (Welcome!)
- Overview of RL
- Markov Decision Processes
    - Problem statement
    - Policy Evaluation

The RL Setting, basically



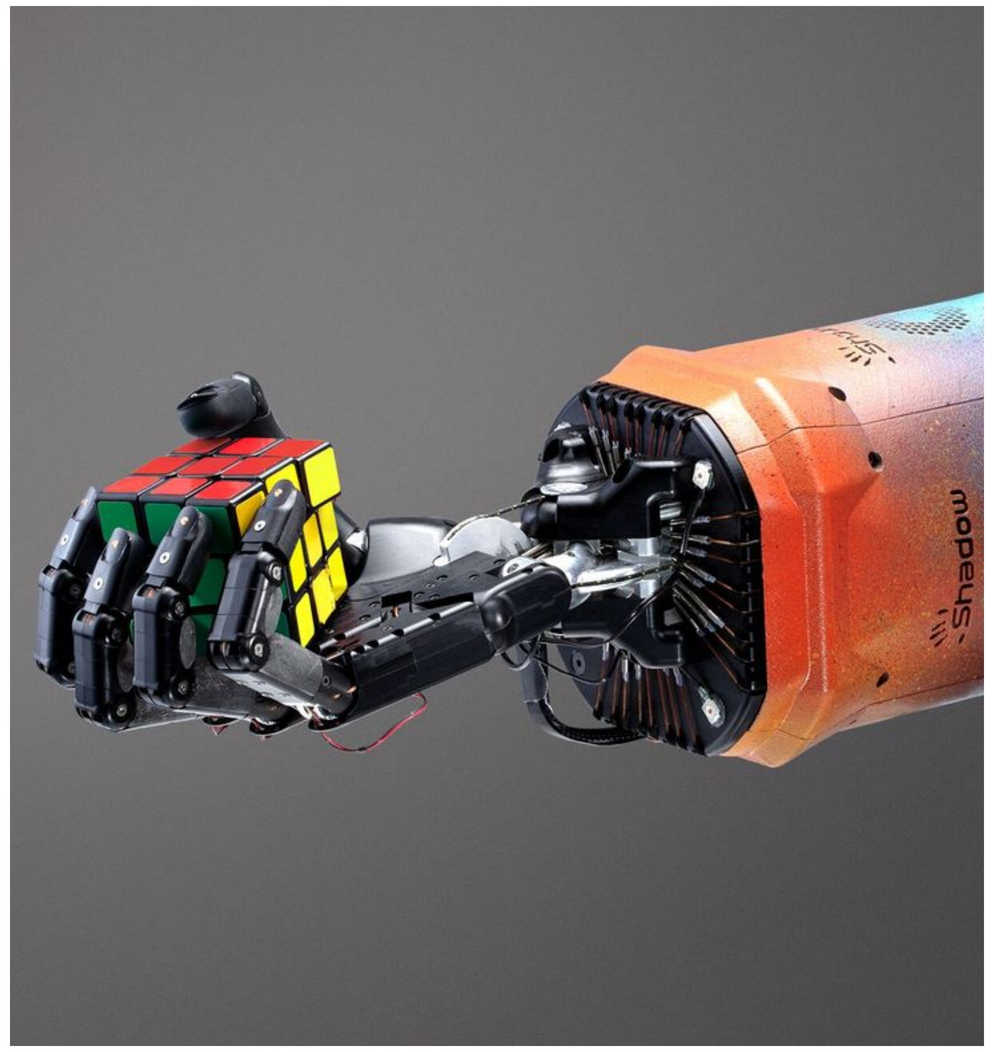
Many RL Successes



TD GAMMON [Tesauro 95]

[AlphaZero, Silver et.al, 17]

[OpenAI Five, 18]



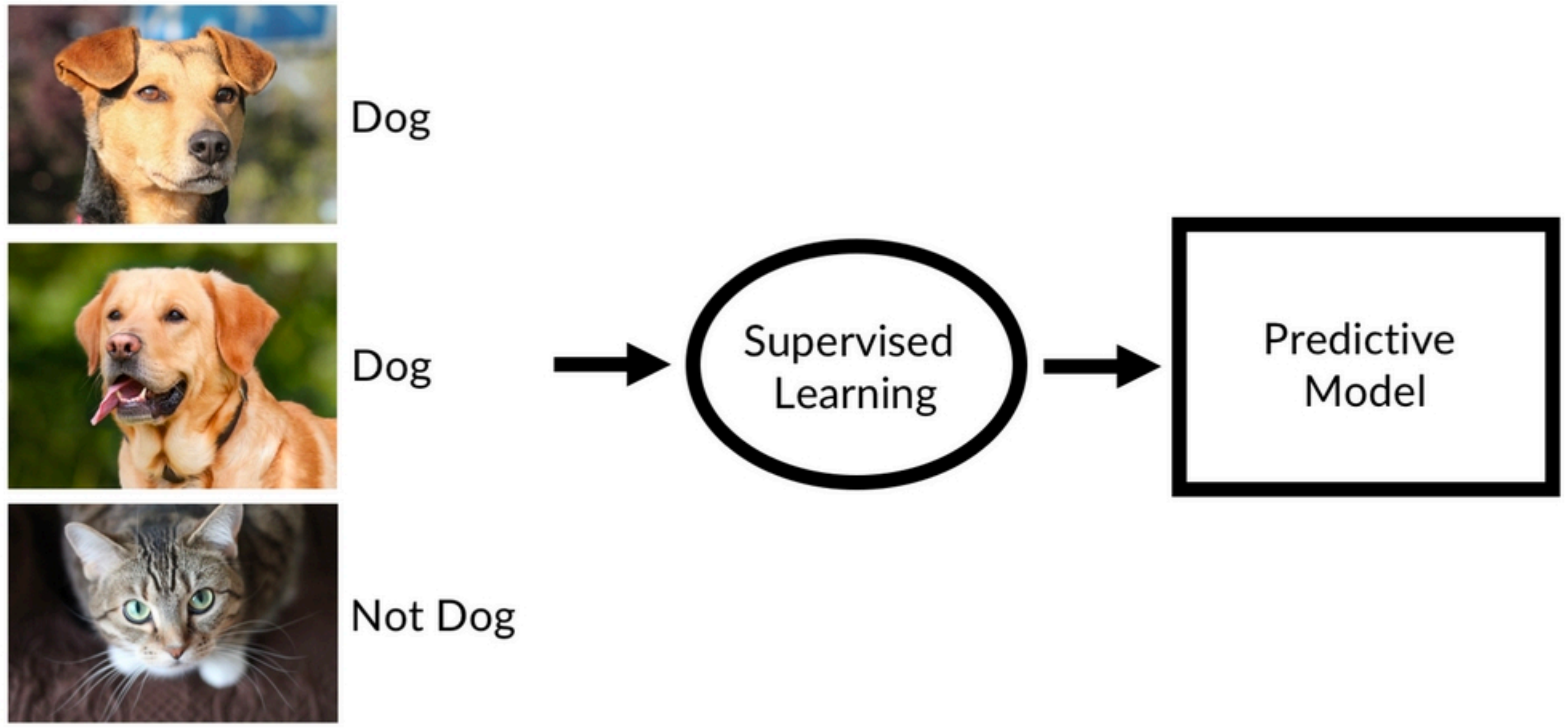
[OpenAI,19]

Supply Chains [Madeka et al '23]

Many Future RL Challenges

Vs Other Settings

	Learn from Experience	Generalize	Interactive	Exploration	Credit assignment
Supervised Learning	✓	✓			
Bandits ("horizon 1"-RL)	✓	✓	✓	✓	
Reinforcement Learning	✓	✓	✓	✓	✓



Point/Counterpoint: Why Should/Shouldn’t You Study RL?

Point: An elegant formulation!
Counterpoint: seen the notation?

Point: Tackles (Nearly) the Most General Problem
Counterpoint: Maybe *too* general?

Point: pivotal for AGI?
Counterpoint: AGI could just be Big Data + Scale?

Point: Exploration is fun!
Counterpoint: Exploitation is fun too!

Point: Enabled Real-world successes!
Counterpoint: Those Deployments Come with "Hacks"

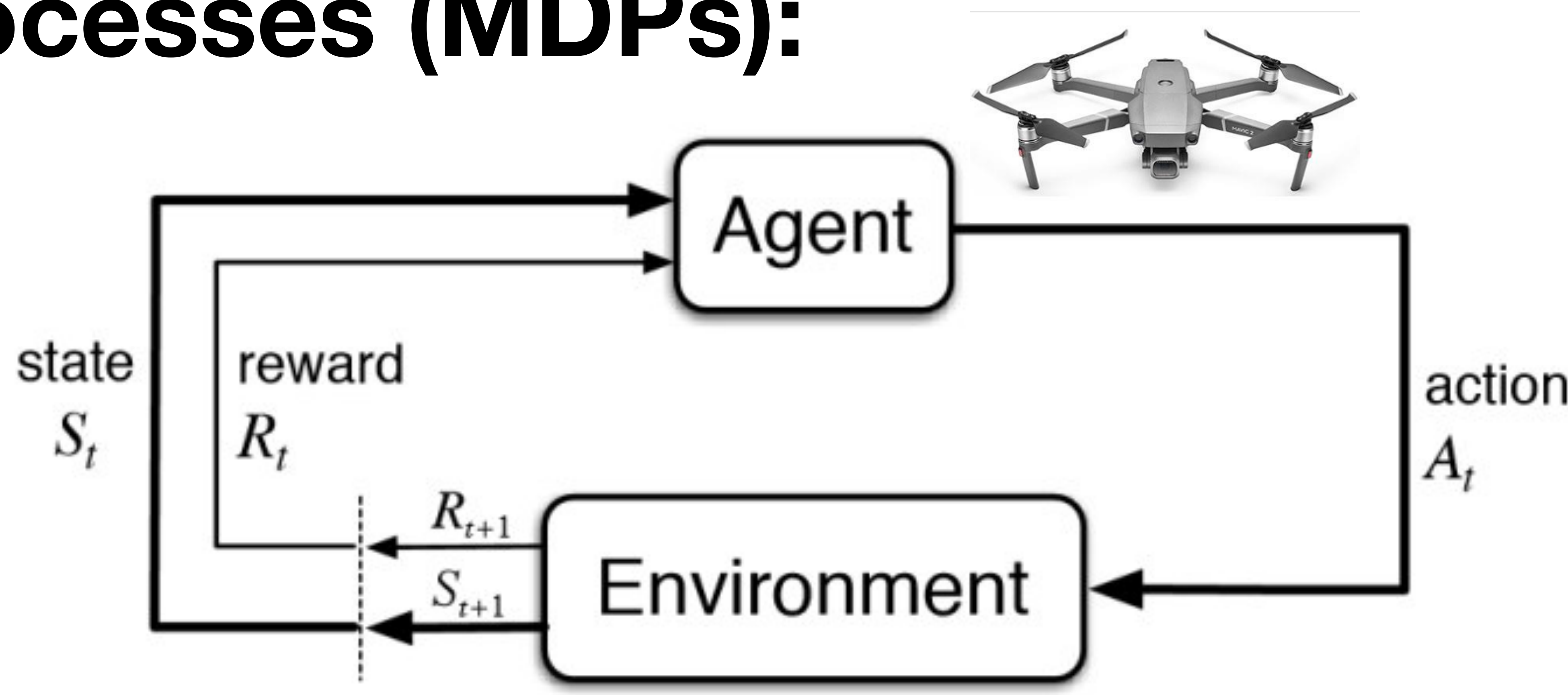
Point: Yann said “abandon RL”!
Counterpoint: Yann also said “abandon generative models”, “abandon probabilistic models”, and “abandon contrastive learning”!

The class will be challenging, and we hope you will enjoy it!

Today

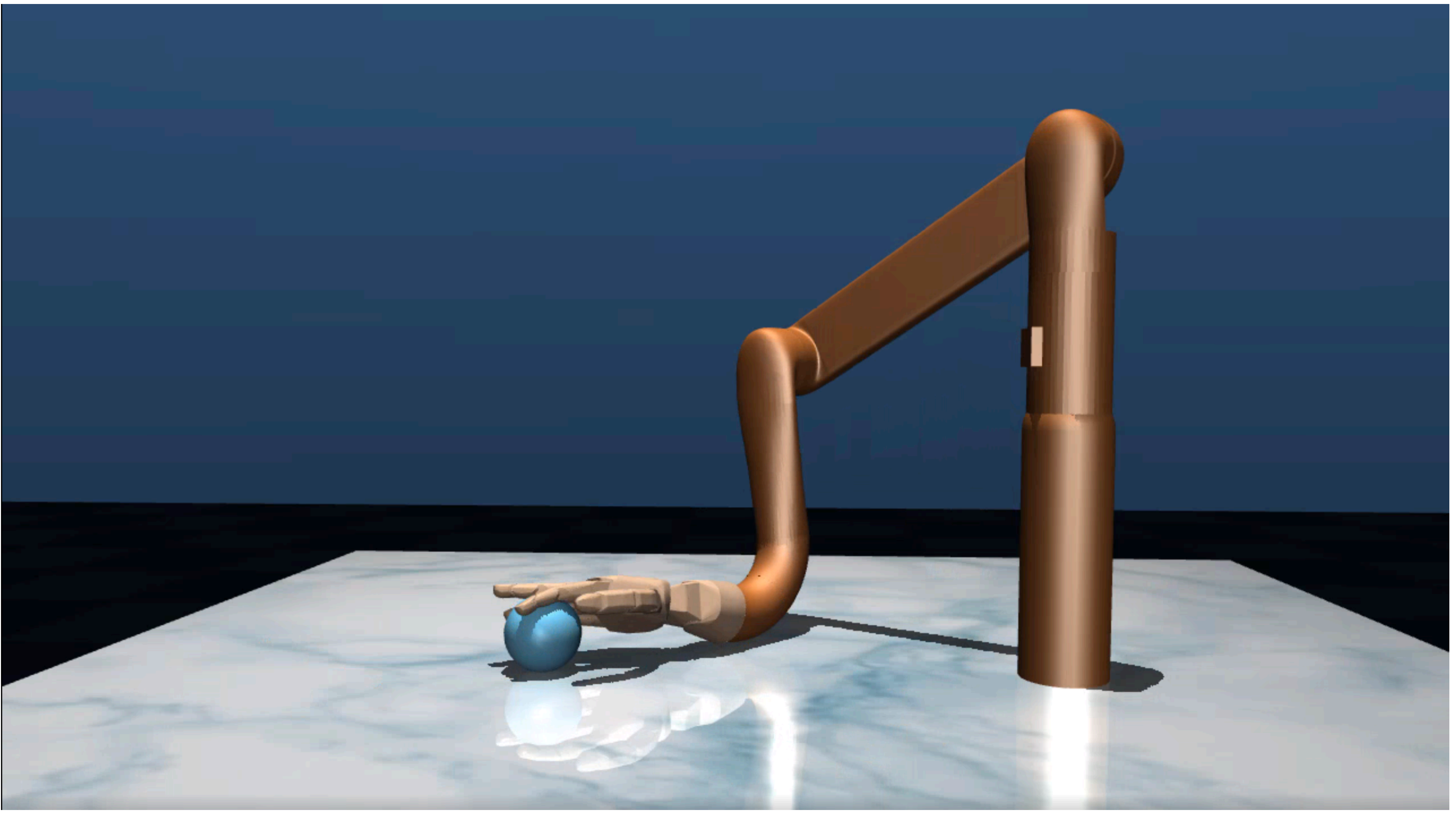
- Logistics (Welcome!)
- Overview of RL
- Markov Decision Processes
    - Problem statement
    - Policy Evaluation

Finite Horizon Markov Decision Processes (MDPs):



- An MDP: $\mathscr{M} = \{\mu, S, A, P, r, H\}$
  - μ is a distribution over initial states (sometimes we assume we start a given state s_0)
  - S a set of states
  - A a set of actions
  - $P : S \times A \mapsto \Delta(S)$ specifies the dynamics model, i.e. $P(s' \mid s, a)$ is the probability of transitioning to s' form states s under action a
  - $r : S \times A \to [0,1]$
    - For now, let's assume this is a deterministic function
    - (sometimes we use a cost $c : S \times A \to [0,1]$)
  - A time horizon $H \in \mathbb{N}$

Example:
robot hand needs to pick the ball and hold it in a goal (x,y,z) position



State s: robot configuration (e.g., joint angles) and the ball's position

Action a: Torque on joints in arm & fingers

Transition $s' \sim P(\cdot \mid s, a)$: physics + some noise

policy $\pi(s)$: a function mapping from robot state to action (i.e., torque)

reward/cost:

$r(s, a)$: immediate reward at state (s, a)

$c(s, a)$: torque magnitude + dist to goal

horizon: timescale H or discount factor γ

$$\pi^\star = \arg\min_{\pi} \mathbb{E}\left[c(s_0, a_0) + c(s_1, a_1) + c(s_2, a_2) + \ldots c(s_{H-1}, a_{H-1}) \,\middle|\, s_0, \pi \right]$$

The Episodic Setting and Trajectories

- Policy $\pi := \left\{\pi_0, \pi_1, \ldots, \pi_{H-1}\right\}$
  - deterministic policies: $\pi_t : S \mapsto A$; stochastic policies: $\pi_t : S \mapsto \Delta(A)$
  - we also consider time-dependent policies (but not a function of the history)
- Sampling a trajectory τ on an episode: for a given policy π
  - Sample an initial state $s_0 \sim \mu$:
  - For $t = 0,1,2,\ldots H-1$
    - Take action $a_t \sim \pi_t(\cdot \mid s_t)$
    - Observe reward $r_t = r(s_t, a_t)$
    - Transition to (and observe) s_{t+1} where $s_{t+1} \sim P(\cdot \mid s_t, a_t)$
  - The sampled trajectory is $\tau = \left\{s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{H-1}, a_{H-1}, r_{H-1}\right\}$

The Probability of a Trajectory & The Objective

- Probability of trajectory: let $\rho_{\pi,\mu}(\tau)$ denote the probability of observing trajectory $\tau = \{s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{H-1}, a_{H-1}, r_{H-1}\}$ when acting under π with $s_0 \sim \mu$.
  - Shorthand: we sometimes write ρ or ρ_π when π and/or μ are clear from context.
  - The rewards in this trajectory must be $r_t = r(s_t, a_t)$ (else $\rho_\pi(\tau) = 0$).
  - For π stochastic:
    $\rho_\pi(\tau) = \mu(s_0)\pi(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\ldots\pi(a_{H-2} \mid s_{H-2})P(s_{H-1} \mid s_{H-2}, a_{H-2})\pi(a_{H-1} \mid s_{H-1})$
  - For π deterministic:
    $\rho_\pi(\tau) = \mu(s_0)\mathbf{1}\left(a_0 = \pi(s_0)\right)P(s_1 \mid s_0, a_0)\ldots P(s_{H-1} \mid s_{H-2}, a_{H-2})\mathbf{1}\left(a_{H-1} = \pi(s_{H-1})\right)$

- Objective: find policy π that maximizes our expected cumulative episodic reward:

$$\max_{\pi} \mathbb{E}_{\tau \sim \rho_\pi}\left[r(s_0, a_0) + r(s_1, a_1) + \ldots + r(s_{H-1}, a_{H-1})\right]$$

Today

- Logistics (Welcome!)
- Overview of RL
- Markov Decision Processes
  - Problem statement
  - Policy Evaluation

Value function and Q functions:

Quantities that allow us to reason policy's long-term effect:

- Value function $V_h^\pi(s) = \mathbb{E}\left[\sum_{t=h}^{H-1} r(s_t, a_t) \middle| s_h = s\right]$

- Q function $Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{t=h}^{H-1} r(s_t, a_t) \middle| (s_h, a_h) = (s, a)\right]$

- At the last stage, what are:

$Q_{H-1}^\pi(s, a) =$

$V_{H-1}^\pi(s) =$

Value function and Q functions:

Quantities that allow us to reason policy's long-term effect:

- Value function $V_h^\pi(s) = \mathbb{E}\left[\sum_{t=h}^{H-1} r(s_t, a_t) \middle| s_h = s\right]$

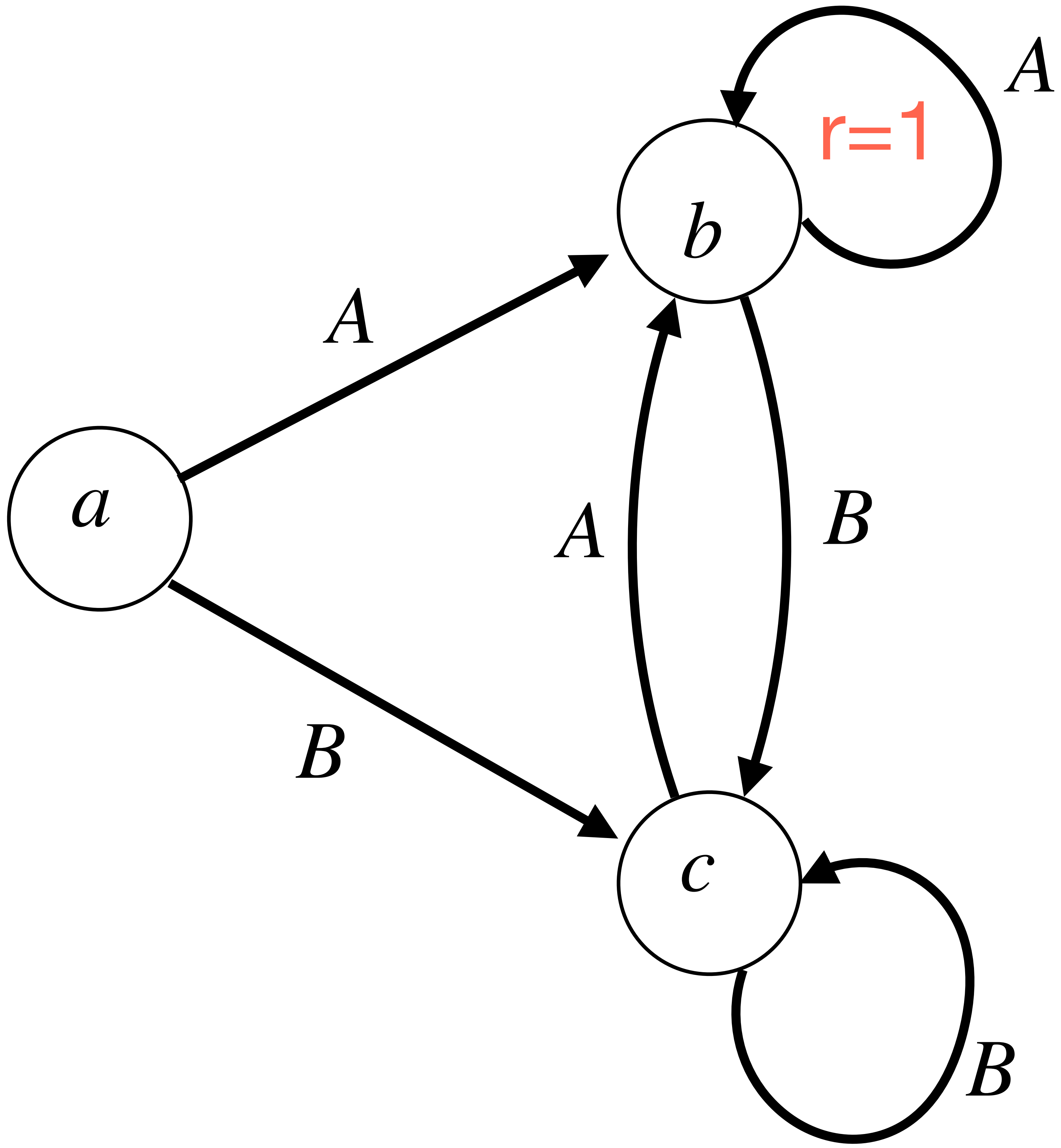
- Q function $Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{t=h}^{H-1} r(s_t, a_t) \middle| (s_h, a_h) = (s, a)\right]$

- At the last stage, for a stochastic policy,:

$$Q_{H-1}^\pi(s, a) = r(s, a) \qquad V_{H-1}^\pi(s) = \sum_a \pi_{H-1}(a \mid s) r(s, a)$$

Example of Policy Evaluation (e.g. computing V^π and Q^π)

Consider the following **deterministic** MDP w/ 3 states & 2 actions, with $H = 3$



Reward: $r(b, A) = 1$, & 0 everywhere else

- Consider the deterministic policy
  $\pi_0(s) = A, \pi_1(s) = A, \pi_2(s) = B, \forall s$

- What is V^π?

$$V_2^\pi(a) = 0, V_2^\pi(b) = 0, V_2^\pi(c) = 0$$

$$V_1^\pi(a) = 0, V_1^\pi(b) = 1, V_1^\pi(c) = 0$$

$$V_0^\pi(a) = 1, V_0^\pi(b) = 2, V_0^\pi(c) = 1$$

Summary:

- **Finite horizon MDPs (a framework for RL):**
- Key concepts:

  V and Q functions; sampling a trajectory $\rho_\pi(\tau)$; Bellman consistency equations;