In the Name of God

Iran University
of Science and
Technology
دانشگاه علم و صنعت ایران

Reinforcement Learning in Control

Dr. Saeed Shamaghdari

Electrical Engineering Department
Control Group

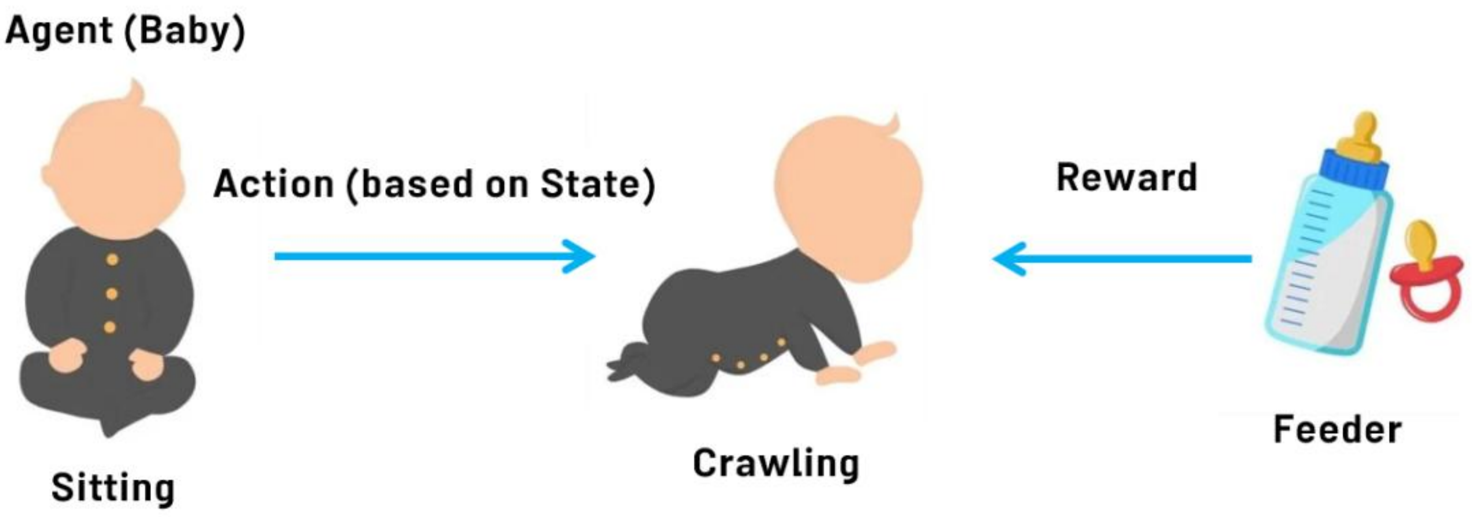
Fall 2025 | 4041

Introduction

Reinforcement Learning: Learning through Interaction

Most natural way of learning: learning from experience

- Quadrupedal Walking
- Human Negotiation
- Driving Car (modern/old)
- Infant Playing



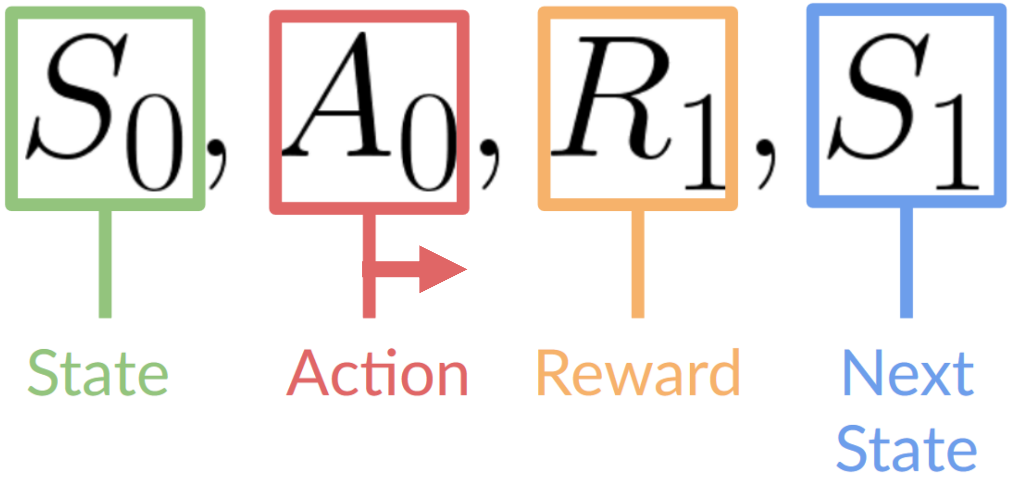
Information Gathering: Understanding the outcome of a series of actions.
Reward Learning: Determining which action to take to achieve the goal.

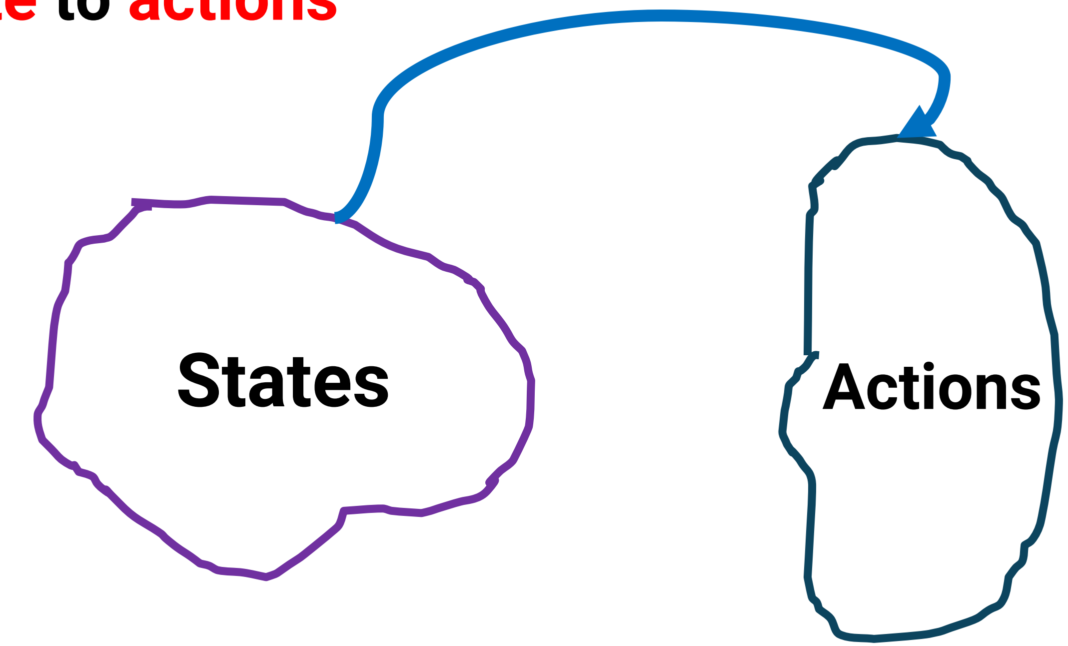
RL: Goal Directed Learning

Reinforcement Learning: Mapping from state to actions

Goal: Maximizing long-term reward
Using optimal control ideas

Impact of Action Choice:
Next **Reward**, Next **State**





Supervised Learning (Regression, Classification)

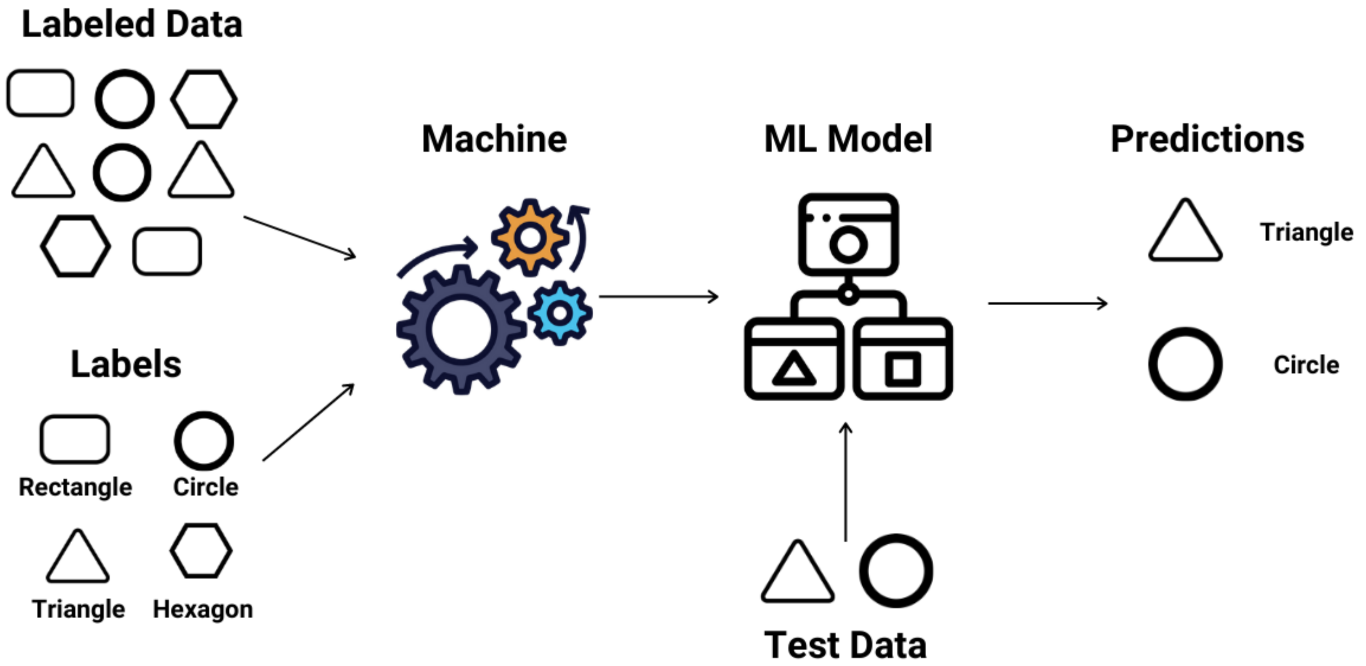
Training Datasets:

Text Doc.
Images
Sound

→ Feature Vectors

Labels → Machine

Prediction Model



RL vs. Supervised Learning:

No labeled dataset required

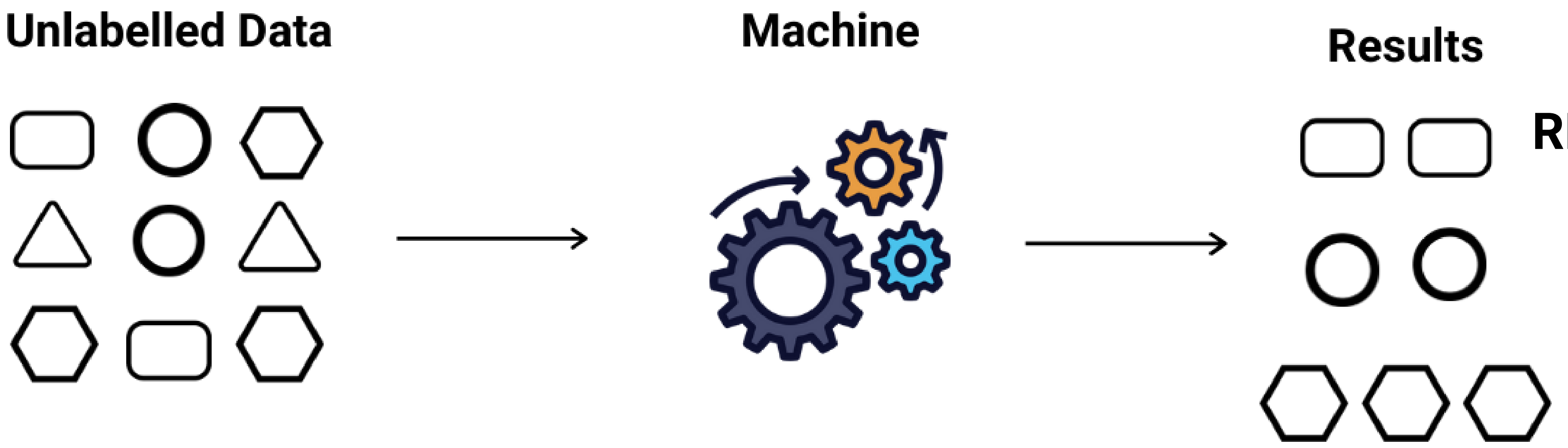
No access to correct answers

Unsupervised Learning (Clustering, Principal Component)

Training Dataset → Feature Vectors → ML Algorithm

ML Algorithm → Model

New Data → Feature Vectors → Model → Probability, Cluster ID



RL vs. Unsupervised Learning:

No need to discover structure

Focus on reward maximization (vs. analyzing data patterns)

In Summary …

	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Data	Labeled data	Unlabeled data	Environment and feedback
Goal	Learn mapping between input data and output labels	Discover patterns, relationships, or groupings	Learn policy to maximize cumulative reward

Reinforcement Learning: Life-Long Learning

Requirements of an Agent in Reinforcement Learning:

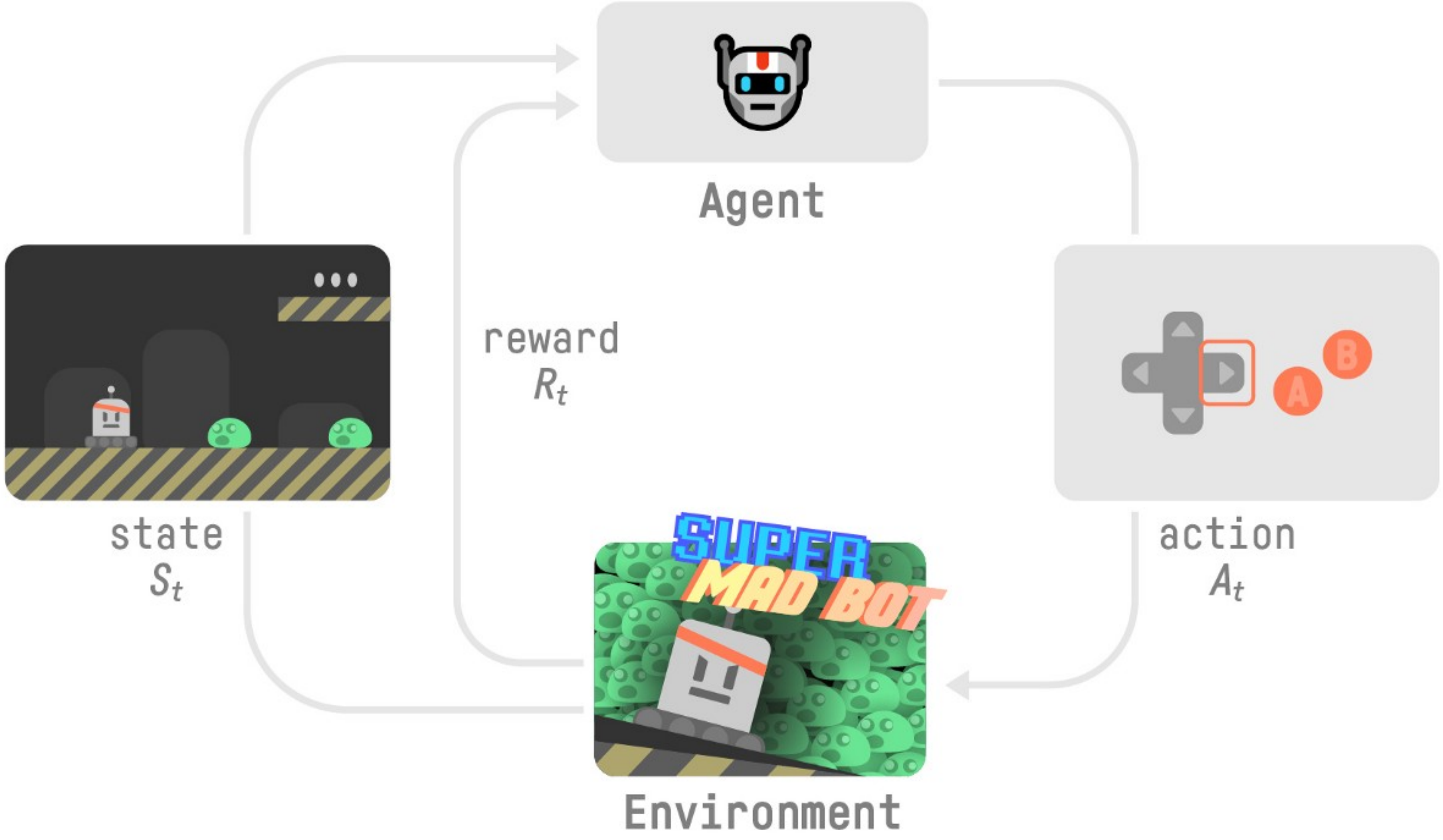
Sense

↓

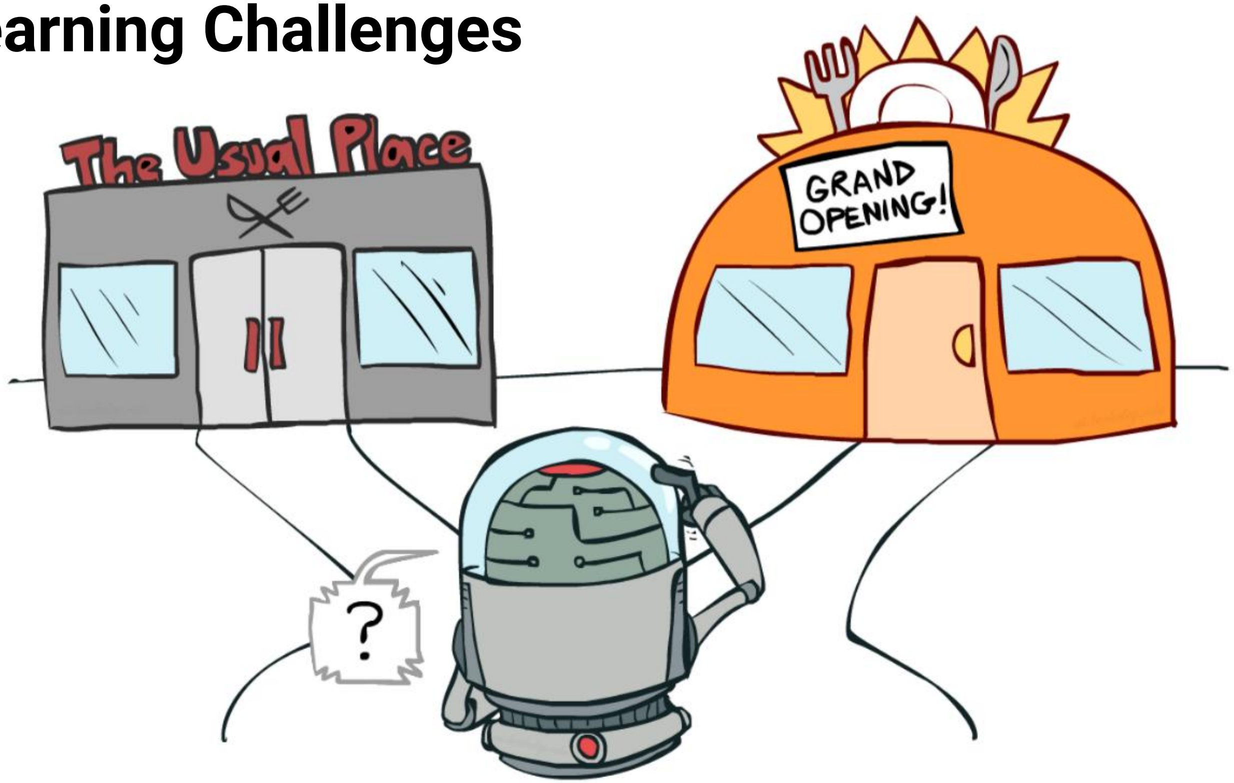
Act

↓

Goal



Reinforcement Learning Challenges



Exploitation is exploiting known information to maximize the reward.

Exploration is exploring the environment (deterministic/stochastic) by trying random actions in order to find more information about the environment.

Reinforcement Learning Challenges

Exploitation is exploiting known information to maximize the reward.

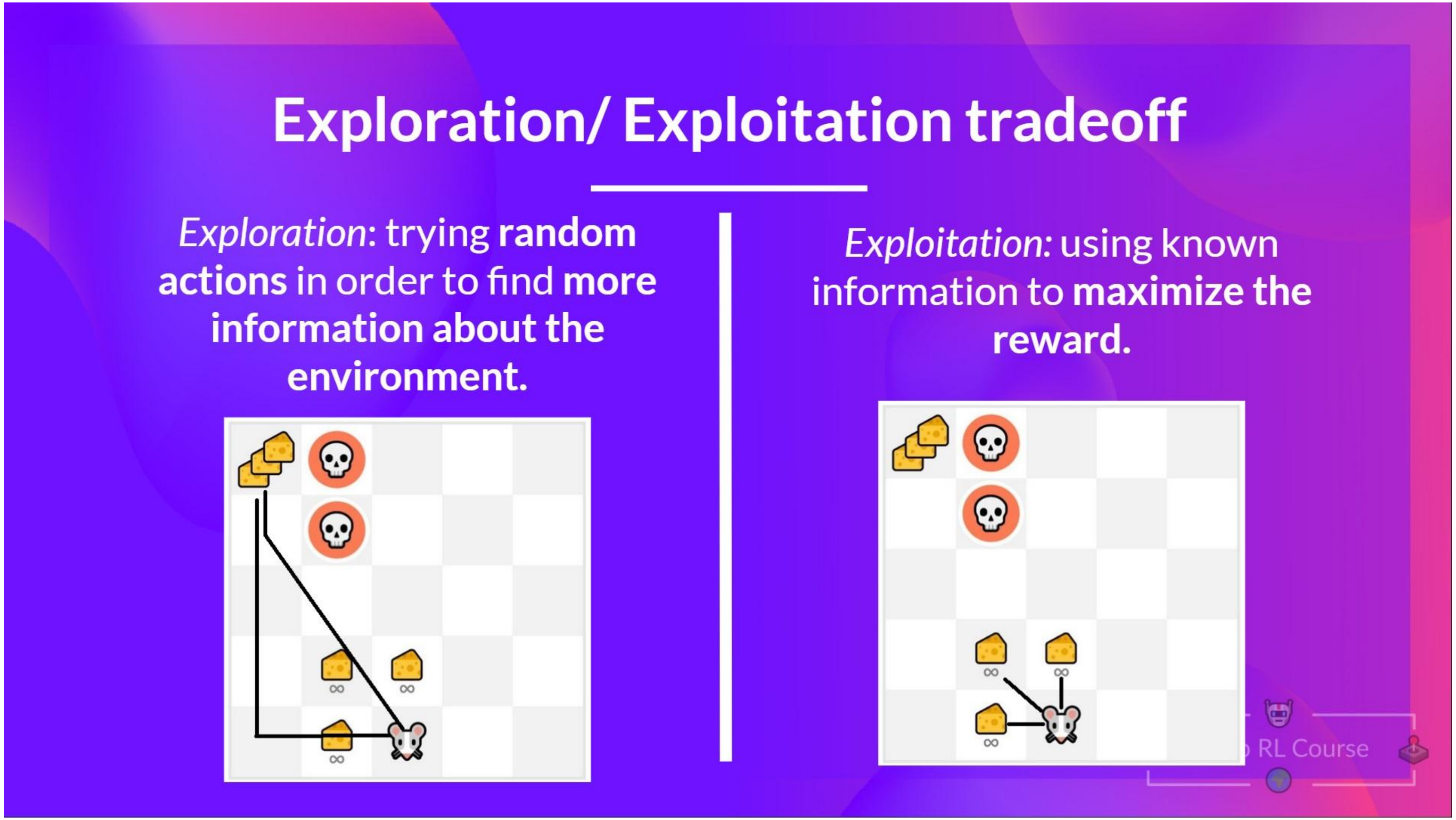
Exploration is exploring the environment (deterministic/stochastic) by trying random actions in order to find more information about the environment.

Selecting better actions

In Stochastic Environments:

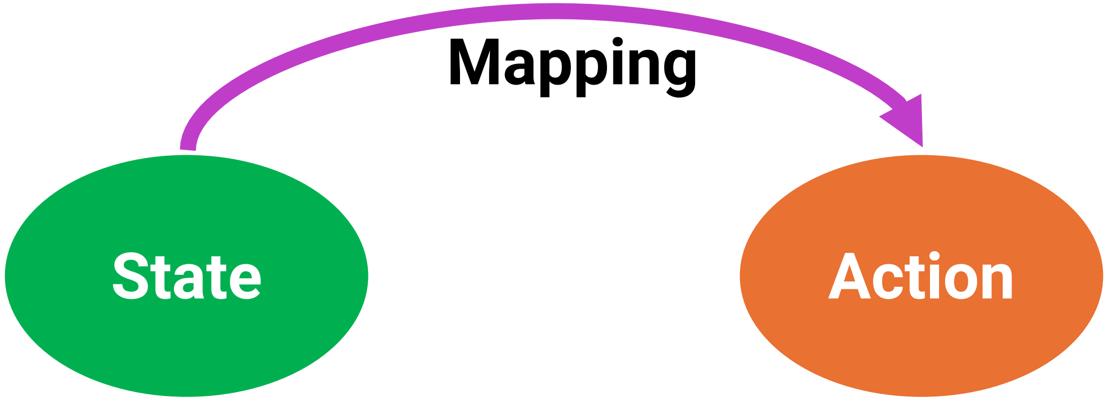
Repeated execution of an action to estimate its *Expected Reward*

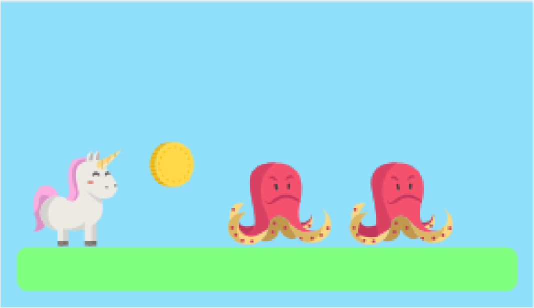
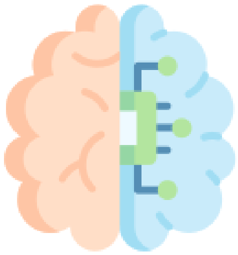
To recap…


Exploration/ Exploitation tradeoff
Exploration: trying **random actions** in order to find **more information about the environment.**
Exploitation: using known information to **maximize the reward.**
RL Course

Reinforcement Learning Main Elements

1. Policy



 →  → 

$$\text{State} \rightarrow \pi(\text{State}) \rightarrow \text{Action}$$

Function
Look-up Table
Search

Policy

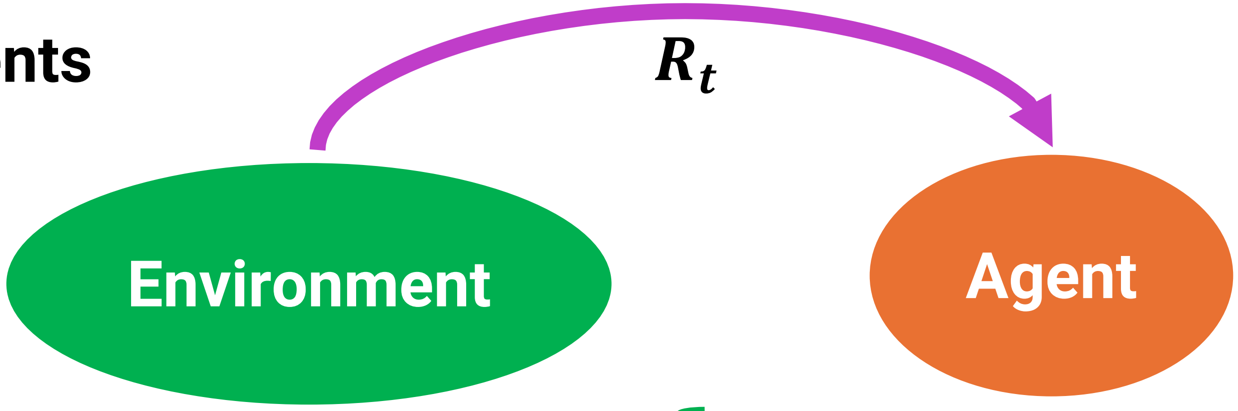
Deterministic
Stochastic

Reinforcement Learning Main Elements

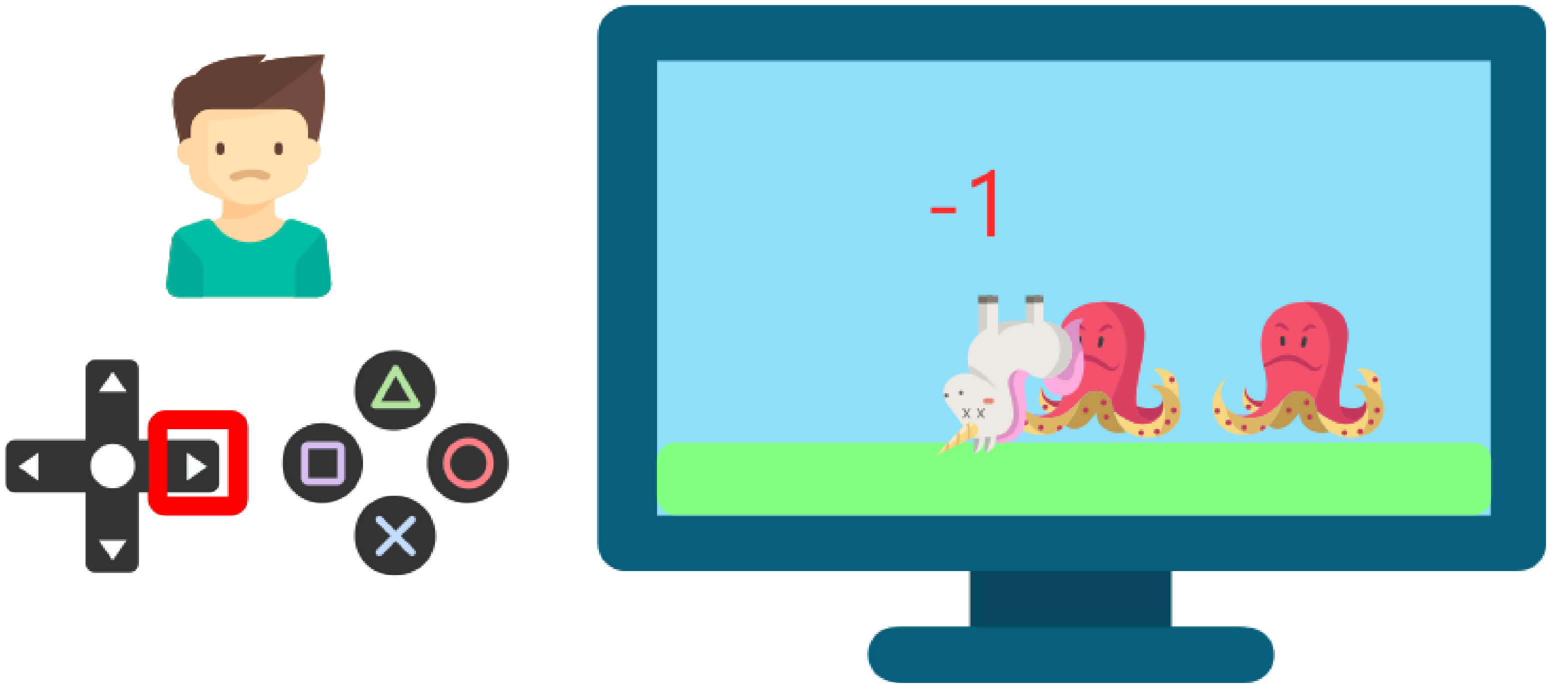
2. Reward Signal

Agent Goal:

Maximizing total reward



Reward $\begin{cases} \text{Deterministic} \\ \text{Stochastic} \end{cases}$



Reinforcement Learning Main Elements

3. Value Function

Reward: Instantaneous reward (momentary goodness)

Value: Long-term reward (long-term goodness)

Expected Rewards

Human

Reward: Instantaneous pleasure or discomfort

Value: Long-term judgment of satisfaction/dissatisfaction

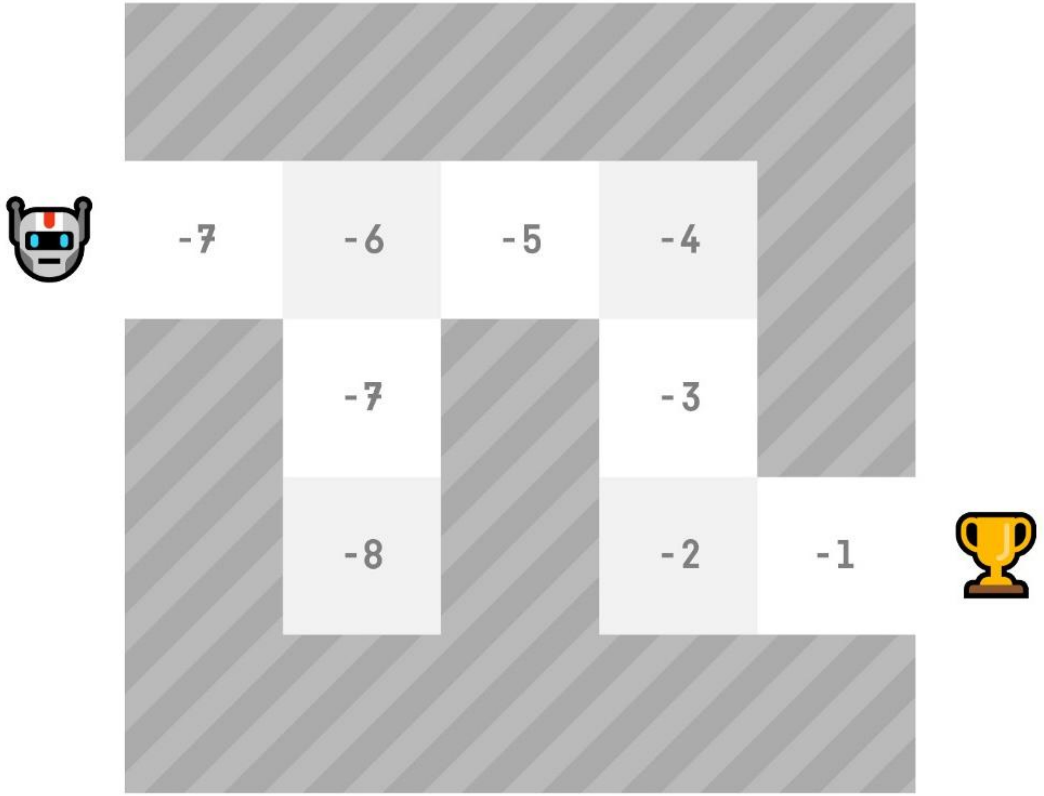
وَعَسَى أَنْ تَكْرَهُوا شَيْئًا وَهُوَ خَيْرٌ لَكُمْ وَعَسَى أَنْ تُحِبُّوا شَيْئًا وَهُوَ شَرٌّ لَكُمْ

Yet it may be that you dislike something, which is good for you, and it may be that you love something, which is bad for you.

Reinforcement Learning Main Elements

3. Value Function

Action Selection Criteria: Value or Reward?



Value Challenge:

Calculation/Estimation method

Reinforcement Learning Main Elements

4. Model

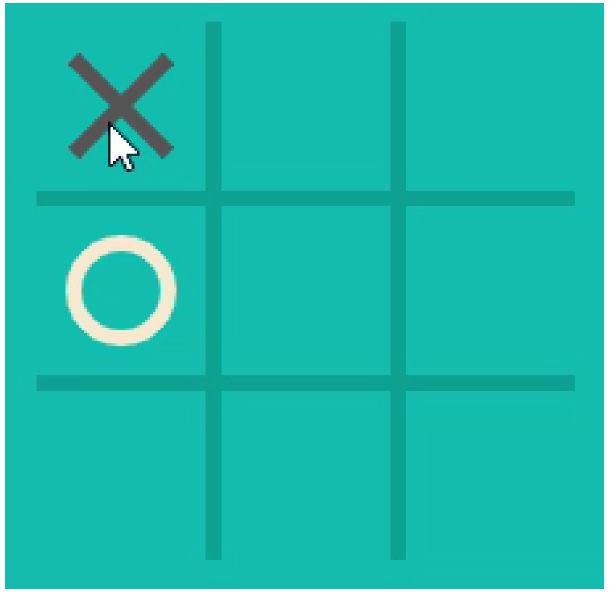
$$\left.\begin{array}{l}\text{Deterministic}\\ \text{Stochastic}\end{array}\right\}\ \text{Behavior of the Environment}\ \left\{\begin{array}{l}\text{Known}\\ \text{Unknown}\end{array}\right.$$

$$\boldsymbol{S_t} \rightarrow \boldsymbol{A_t} \xrightarrow{\textbf{Model}} \left\{\begin{array}{l}\boldsymbol{S_{t+1}}\\ \boldsymbol{A_{t+1}}\end{array}\right.$$

Reinforcement Learning Main Elements

Reinforcement Learning vs. Evolutionary Methods

Lack of attention to policy details in evolutionary methods

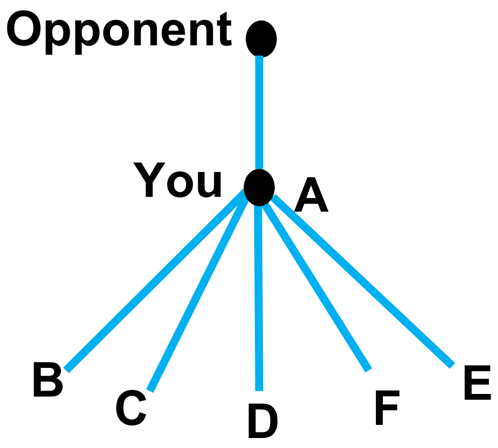


Performance of a skilled agent?
Based on game theory …

Assumption: The opponent is not professional

→ *What is the definition of a state in the game of X-O (Tic-Tac-Toe)?*
Positions of the pieces + whose turn it is?

Constructing the Game Tree



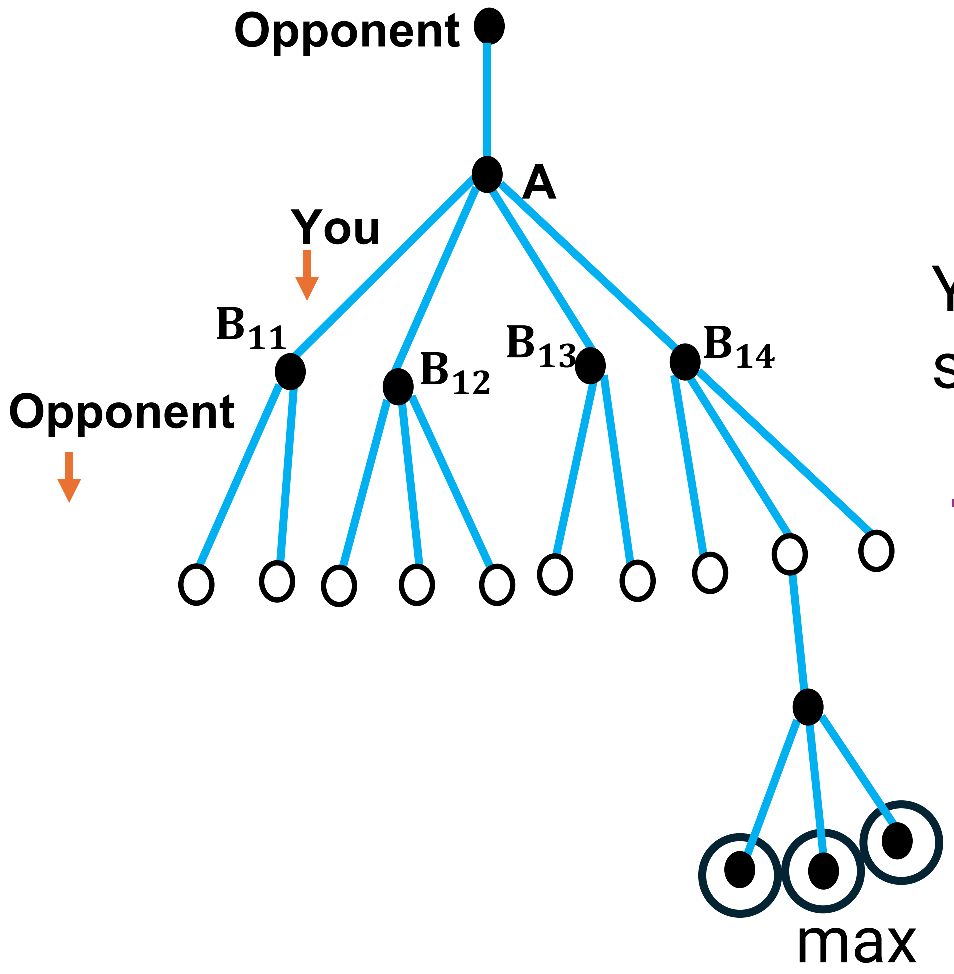
→ If any of B to F wins, then A is the winner.
→ If all of B to F lose, then A is the loser.

Dynamic Programming approach:

Start from the terminal (final) states in the game tree and move bottom-up.

Determine the optimal choice to win.

Constructing the Game Tree



Your probability of winning starting from this position.

P_{14}

Note: The last row represents the definite winning probability.
Dynamic Programming: Calculating probabilities from bottom to top.

Solution with Reinforcement Learning

Temporal Difference

Create a Value table where each row corresponds to a State

Initialize the table (Value)

(example:
a row with O: Pr=0
a row with X: Pr=1
others: Pr=0.5)

Select a policy

Update the Value table based on observations

Game algorithm based on TD (Temporal Difference):

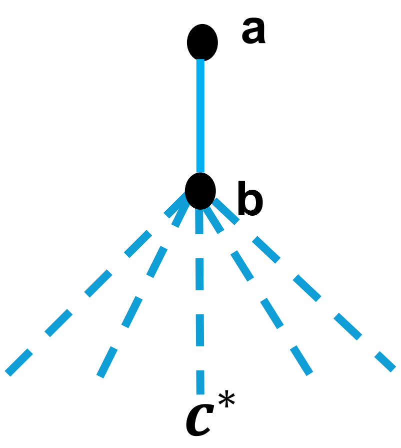
Opponent's move from a to b

Estimate the Value function for the move from b

Greedy selection: move from b to c*

Receive new Reward and calculate V(S(t+1))

Update the table based on game observations:



$$V(S_t) \leftarrow V(S_t) + \alpha\Big[V(S_{t+1}) - V(S_t)\Big]$$

Performing Exploration: randomly selecting suboptimal moves

Example: (going to a restaurant)

Note: no table update

Proof of convergence?

Recap: Solution with RL