

Machine Learning and Data Analytics

ME 5013- Fall 2019

Lectures 21

- Reinforcement Learning Introduction



The University of Texas at San Antonio™

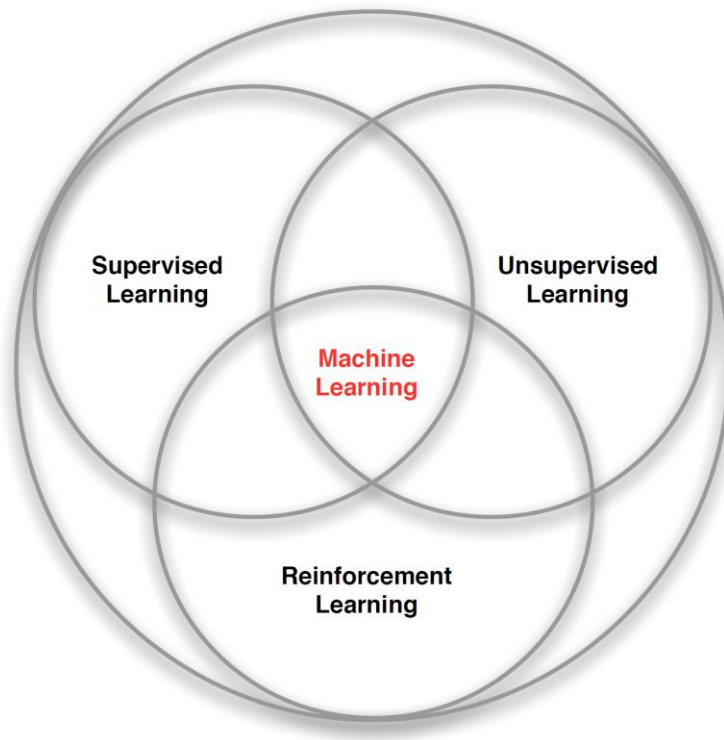
Adel Alaeddini, PhD

Associate Professor of Mechanical Engineering

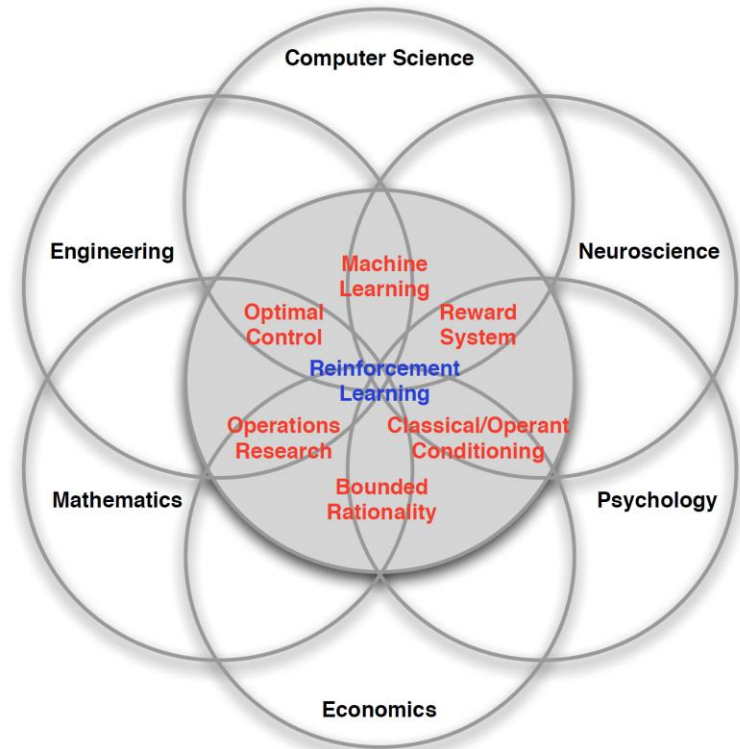
Advanced Data Engineering Lab

adel.alaeddini@utsa.edu

Branches of Machine Learning



Faces of reinforcement learning

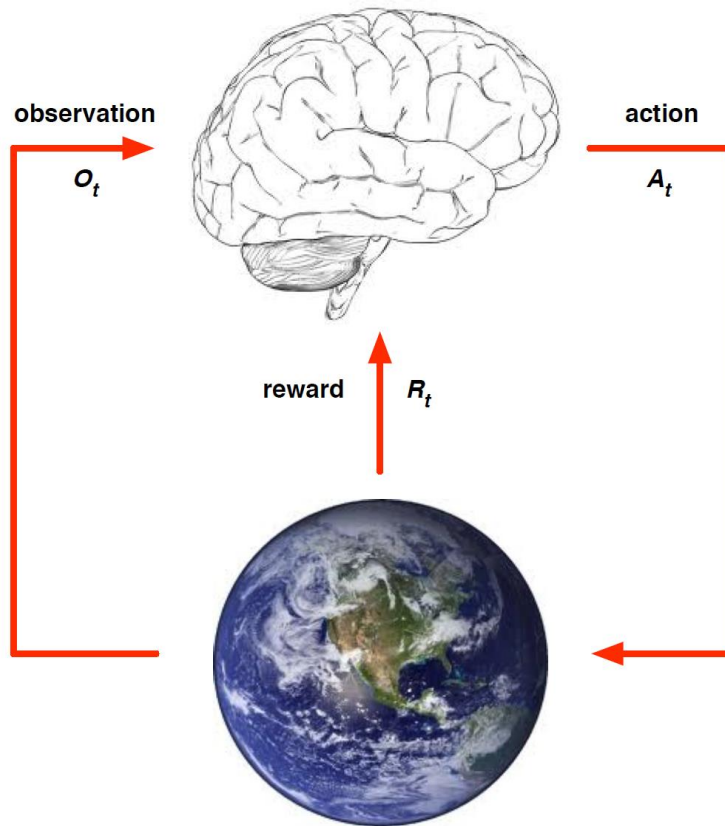


What makes reinforcement learning different from other machine learning paradigms?

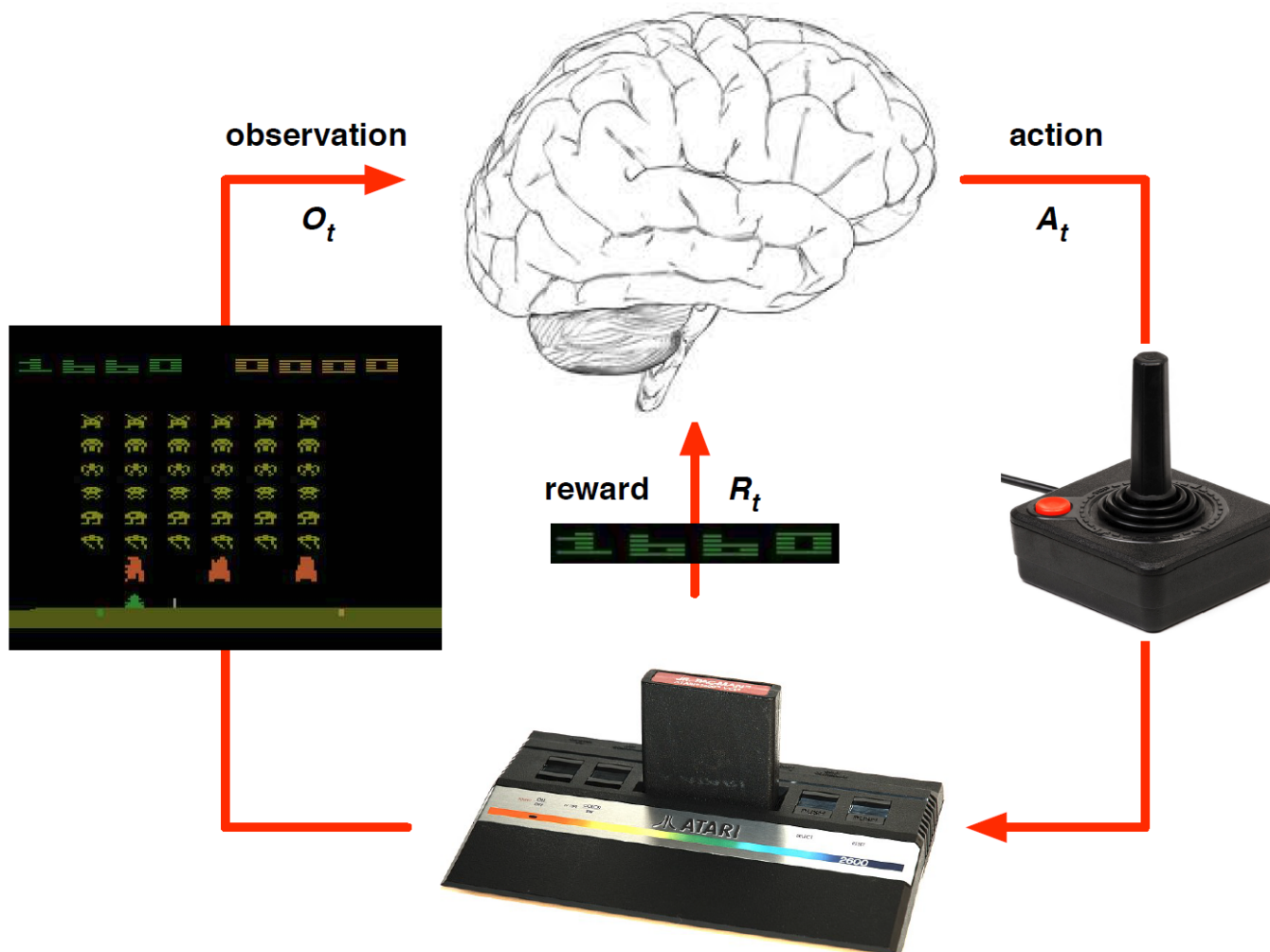
- There is no supervisor, only a *reward* signal
 - Feedback is delayed, not instantaneous
 - Time really matters (sequential, non i.i.d data)
 - Agent's actions affect the subsequent data it receives
-

Applications

- Fly stunt manoeuvres in a helicopter
- Defeat the world champion at Backgammon
- Manage an investment portfolio
- Control a power station
- Make a humanoid robot walk
- Play many different Atari games better than humans



- At each step t the agent:
 - Executes action A_t
 - Receives observation O_t
 - Receives scalar reward R_t
- The environment:
 - Receives action A_t
 - Emits observation O_{t+1}
 - Emits scalar reward R_{t+1}
- t increments at env. step



- A **reward** R_t is a scalar feedback signal
- Indicates how well agent is doing at step t
- Goal: *select actions to maximise total future reward*
- Actions may have long term consequences
- Reward may be delayed
- It may be better to sacrifice immediate reward to gain more long-term reward
- Examples:
 - A financial investment (may take months to mature)
 - Refuelling a helicopter (might prevent a crash in several hours)
 - Blocking opponent moves (might help winning chances many moves from now)

- **State** is the information used to determine what happens next
- Formally, state is a function of the history:

$$S_t = f(H_t)$$

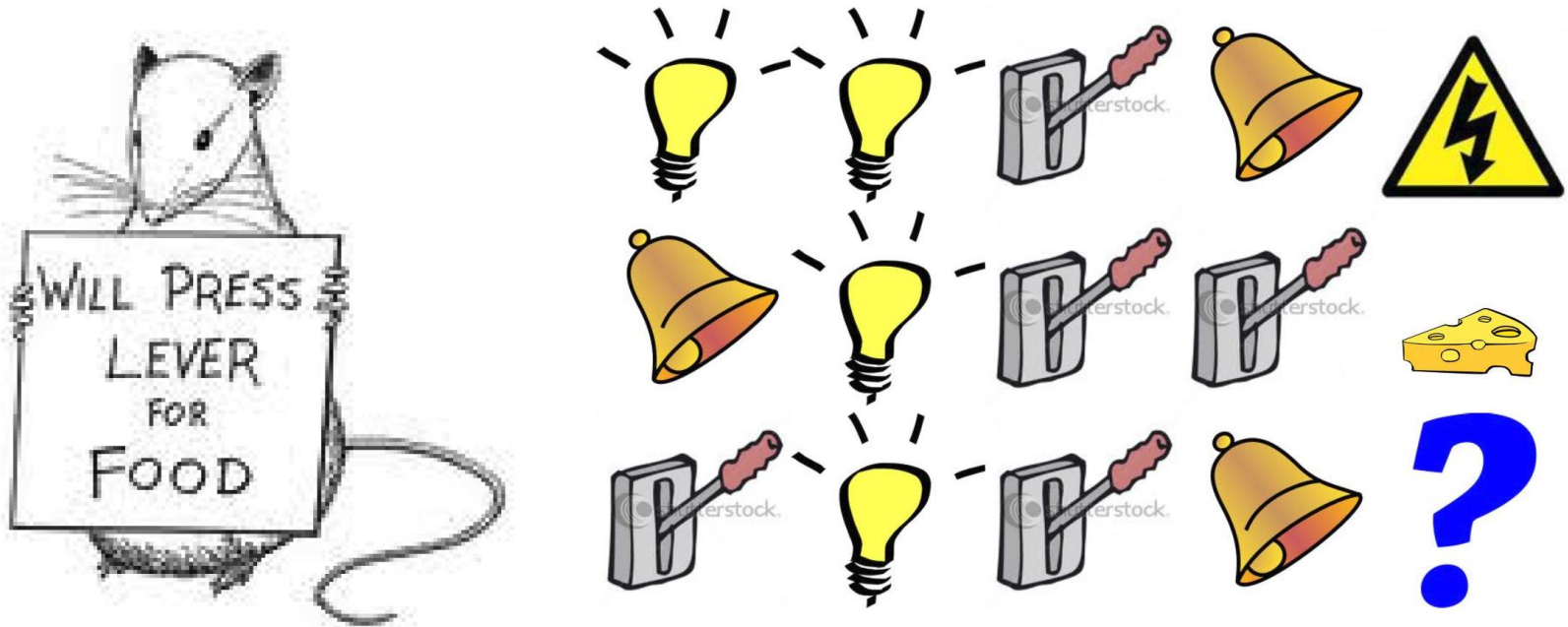
An **information state** (a.k.a. **Markov state**) contains all useful information from the history.

Definition

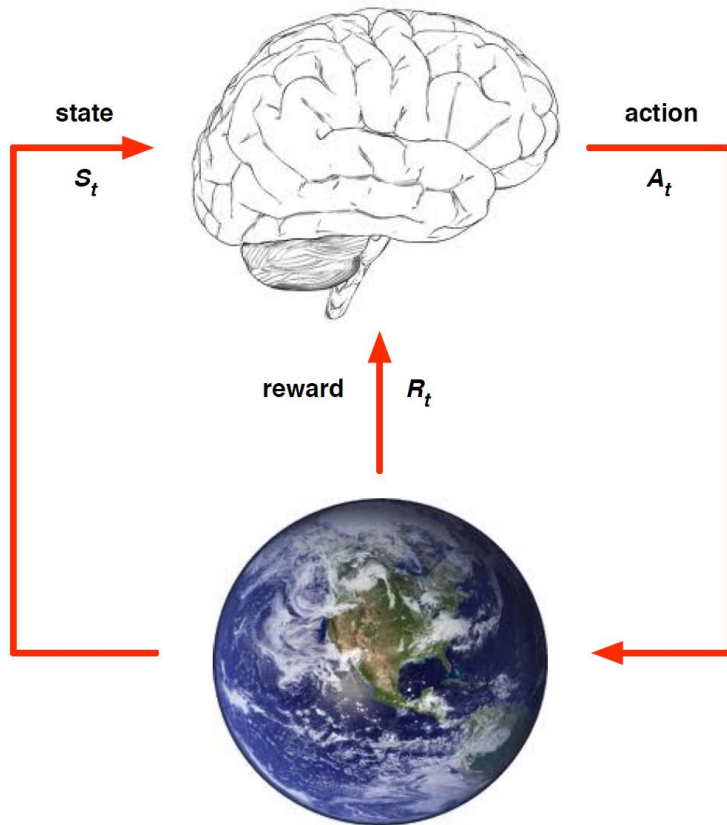
A state S_t is **Markov** if and only if

$$\mathbb{P}[S_{t+1} \mid S_t] = \mathbb{P}[S_{t+1} \mid S_1, \dots, S_t]$$

- Once the state is known, the history may be thrown away
- i.e. The state is a sufficient statistic of the future
- The environment state S_t^e is Markov
- The history H_t is Markov



- What if agent state = last 3 items in sequence?
- What if agent state = counts for lights, bells and levers?
- What if agent state = complete sequence?



Full observability: agent directly observes environment state

$$O_t = S_t^a = S_t^e$$

- Agent state = environment state = information state
- Formally, this is a **Markov decision process** (MDP)

- An RL agent may include one or more of these components:
 - Policy: agent's behaviour function
 - Value function: how good is each state and/or action
 - Model: agent's representation of the environment

- A **policy** is the agent's behaviour
- It is a map from state to action, e.g.
- Deterministic policy: $a = \pi(s)$
- Stochastic policy: $\pi(a|s) = \mathbb{P}[A_t = a | S_t = s]$

- Value function is a prediction of future reward
- Used to evaluate the goodness/badness of states
- And therefore to select between actions, e.g.

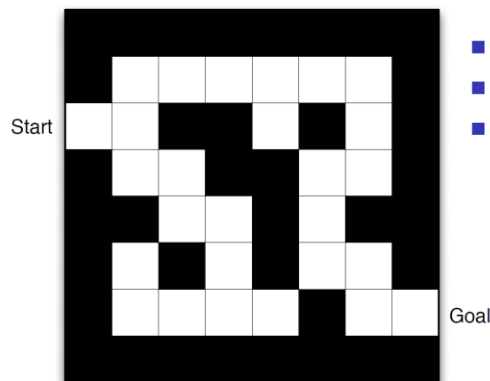
$$v_{\pi}(s) = \mathbb{E}_{\pi} [R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s]$$

- A **model** predicts what the environment will do next
- \mathcal{P} predicts the next state
- \mathcal{R} predicts the next (immediate) reward, e.g.

$$\mathcal{P}_{ss'}^a = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$$

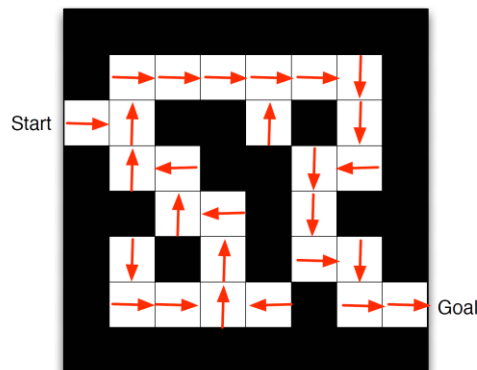
$$\mathcal{R}_s^a = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$$

Value function



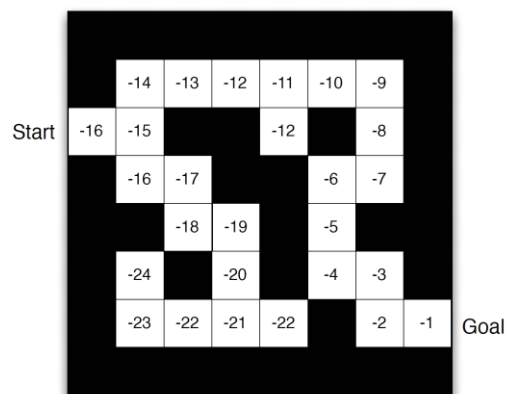
- Rewards: -1 per time-step
- Actions: N, E, S, W
- States: Agent's location

Policy



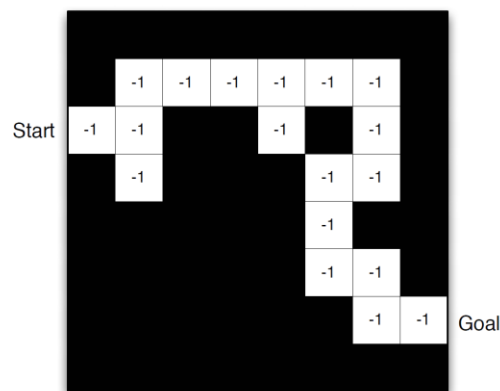
- Arrows represent policy $\pi(s)$ for each state s

Value function



- Numbers represent value $v_{\pi}(s)$ of each state s

Model



- Agent may have an internal model of the environment
- Dynamics: how actions change the state
- Rewards: how much reward from each state
- The model may be imperfect
- Grid layout represents transition model $\mathcal{P}_{ss'}^a$
- Numbers represent immediate reward \mathcal{R}_s^a from each state s (same for all a)

