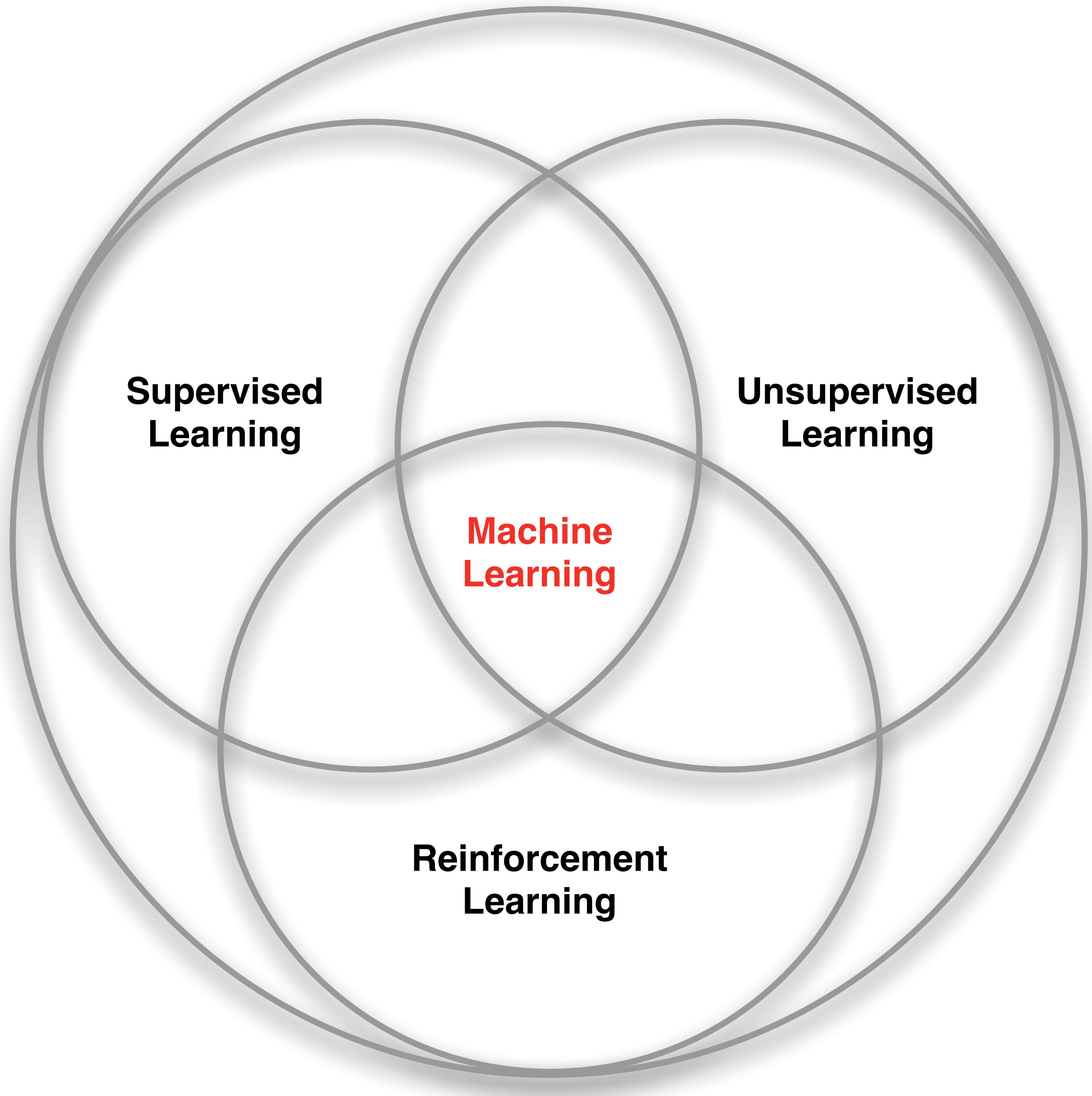
Introduction to Machine Learning (CS419M)

Lecture 24: Reinforcement learning & Course wrap-up

Apr 18, 2018

Slides on RL borrowed from David Silver's lectures.

Branches of ML



What is Reinforcement Learning?

- Learning by interacting with an environment to achieve a goal
- Learning by trial-and-error with only some form of evaluative feedback (aka "reward”)
- Agent’s actions affect the subsequent data it receives

Rewards

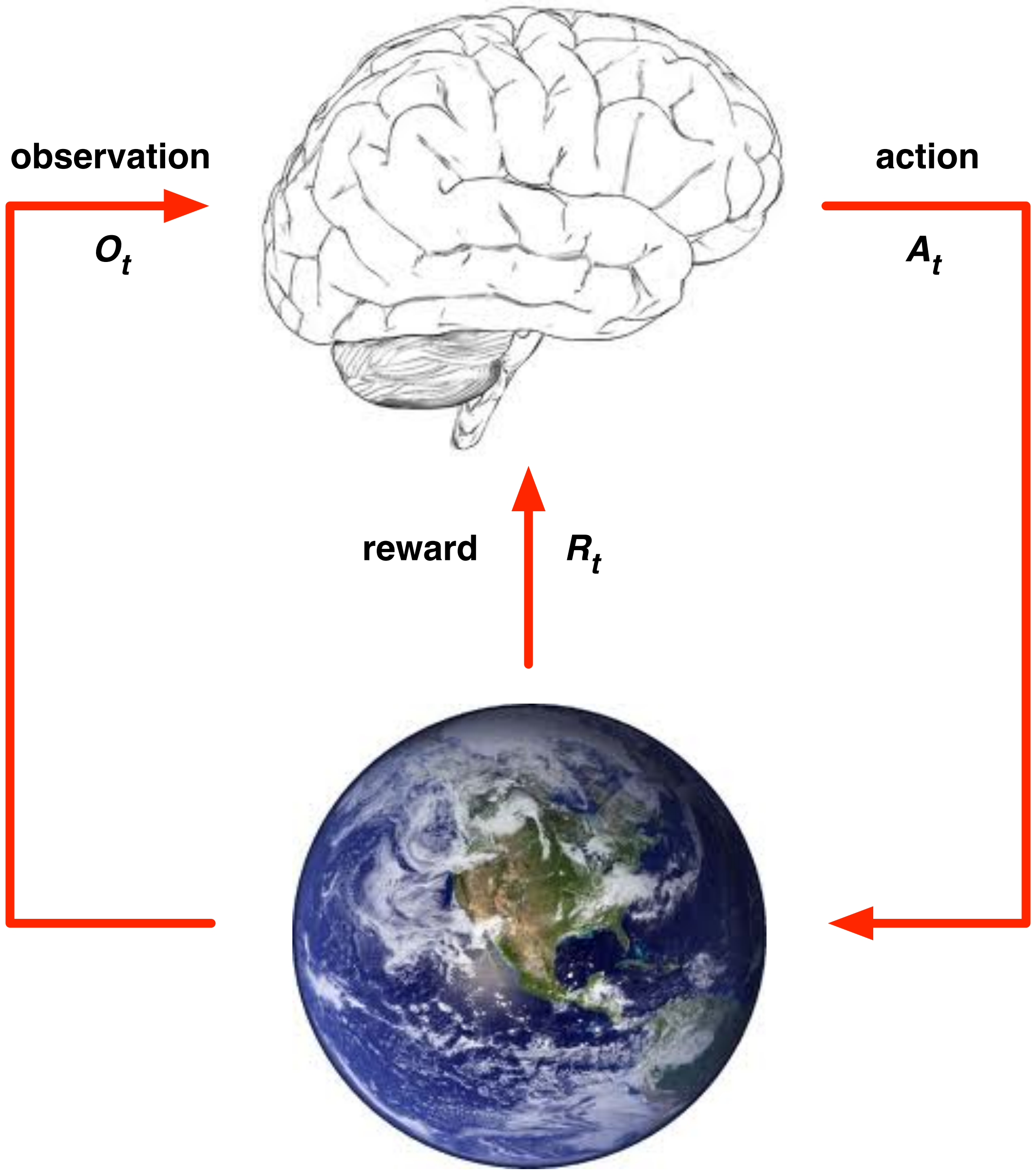
- A reward R_t is a scalar feedback signal
- Indicates how well agent is doing at step t
- The agent's job is to maximise cumulative reward

Reinforcement learning is based on the reward hypothesis

Definition (Reward Hypothesis)

All goals can be described by the maximisation of expected cumulative reward

Agent and Environment



- 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

Major components of an RL Agent

- 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

Policy

- 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 Functions

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

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

Value Functions

- A value function is a prediction of future reward
  - "How much reward will I get from action a in state s?"
- Q-value function gives expected total reward
  - from state s and action a
  - under policy π
  - with discount factor γ

$$Q^{\pi}(s, a) = \mathbb{E}\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots \mid s, a\right]$$

- Value functions decompose into a Bellman equation

$$Q^{\pi}(s, a) = \mathbb{E}_{s', a'}\left[r + \gamma Q^{\pi}(s', a') \mid s, a\right]$$

Optimal Value Function

- An optimal value function is the maximum achievable value

$$Q^*(s, a) = \max_{\pi} Q^{\pi}(s, a) = Q^{\pi^*}(s, a)$$

- Once we have Q^* we can act optimally,

$$\pi^*(s) = \operatorname*{argmax}_{a} Q^*(s, a)$$

Exploration and Exploitation

- Reinforcement learning is like trial-and-error learning
- The agent should discover a good policy
- From its experiences of the environment
- Without losing too much reward along the way

- *Exploration* finds more information about the environment
- *Exploitation* exploits known information to maximise reward
- It is usually important to explore as well as exploit

Examples

- Restaurant Selection

  Exploitation Go to your favourite restaurant
  Exploration Try a new restaurant

- Online Banner Advertisements

  Exploitation Show the most successful advert
  Exploration Show a different advert

- Oil Drilling

  Exploitation Drill at the best known location
  Exploration Drill at a new location

- Game Playing

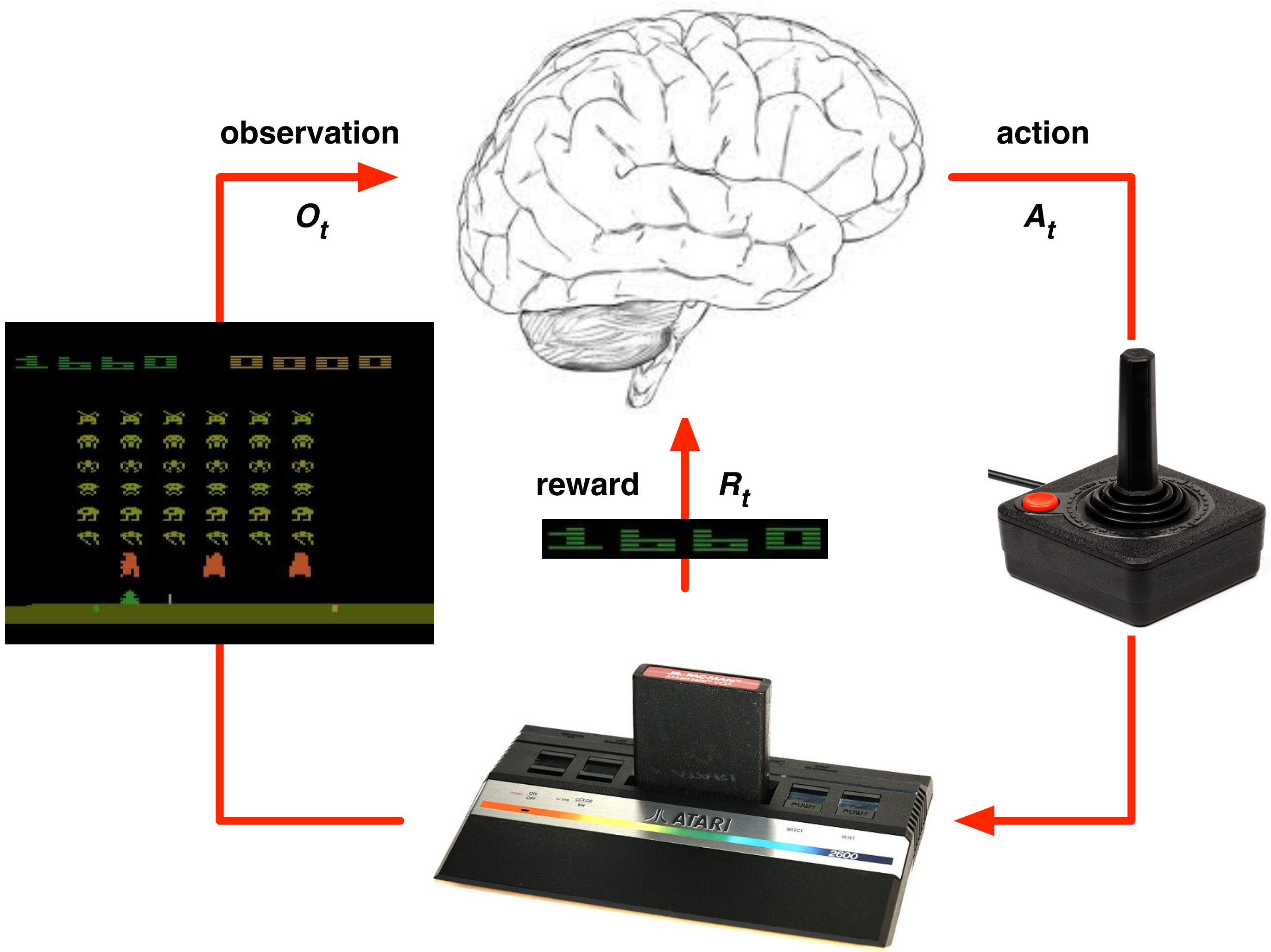
  Exploitation Play the move you believe is best
  Exploration Play an experimental move

Model

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

$$\mathcal{P}_{ss'}^{a} = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$$
$$\mathcal{R}_{s}^{a} = \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right]$$

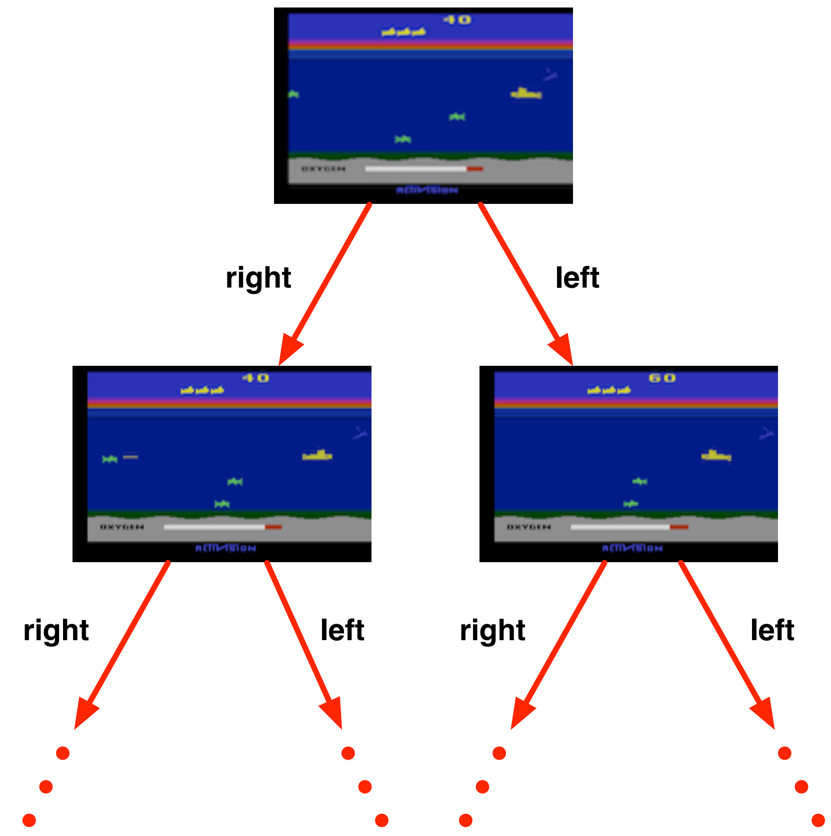
Atari: Reinforcement Learning



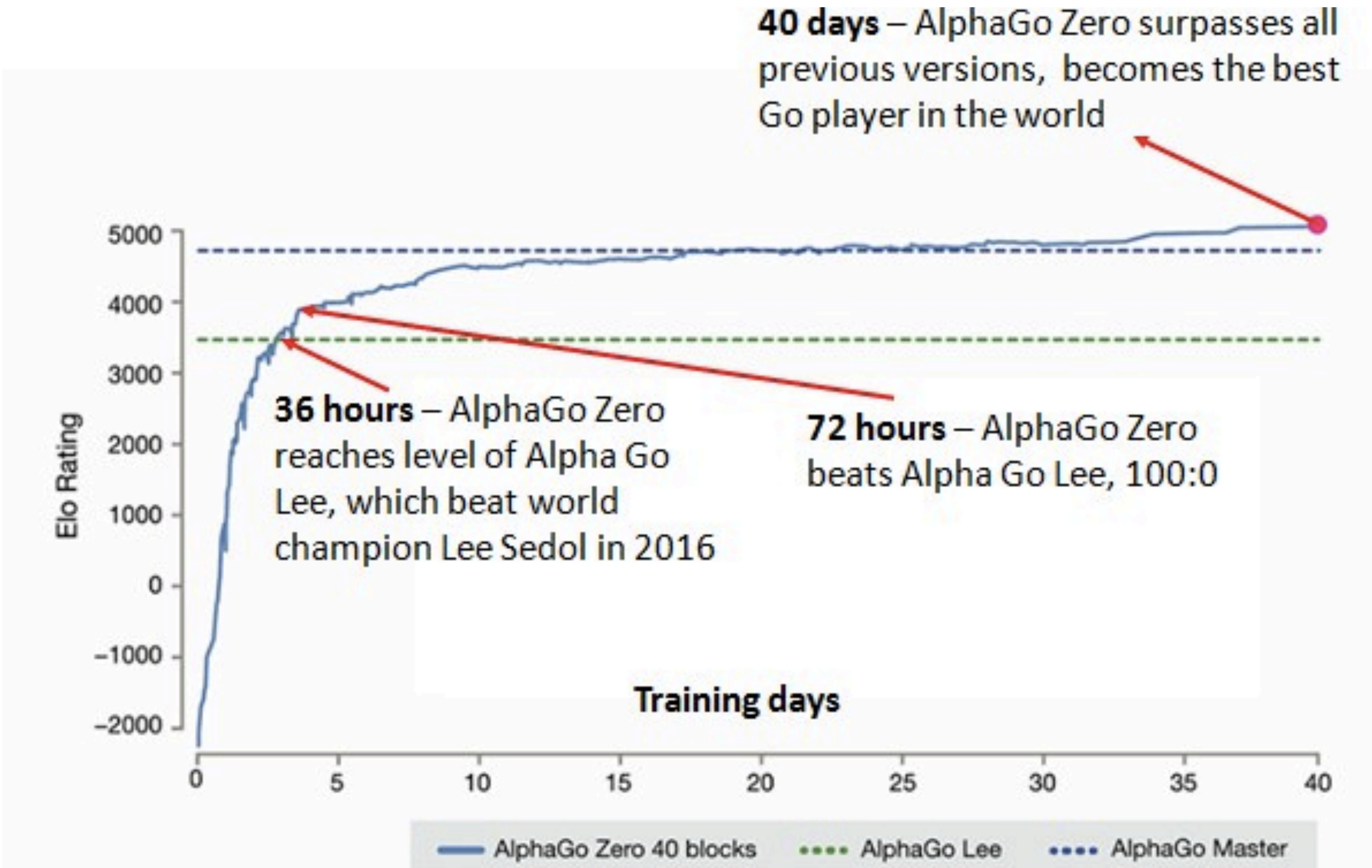
- Rules of the game are unknown
- Learn directly from interactive game-play
- Pick actions on joystick, see pixels and scores

Atari: Planning

- Rules of the game are known
- Can query emulator
    - perfect model inside agent's brain
- If I take action *a* from state *s*:
    - what would the next state be?
    - what would the score be?
- Plan ahead to find optimal policy
    - e.g. tree search



AlphaGo Zero



From: "Mastering the Game of Go without Human Knowledge"

Excellent freely available textbook on RL!

Reinforcement Learning:
An Introduction

Second edition, in progress

****Complete Draft****

November 5, 2017

Richard S. Sutton and Andrew G. Barto

http://incompleteideas.net/book/bookdraft2017nov5.pdf