Summer School in Foundations of RL
Lecture 1: What Reinforcement Learning?

Enrique Mallada

Goals:

- Give an overview of school; describe course objectives
- Introduce reinforcement learning and showcase its successes
- Formulate the Reinforcement Learning Problem

Outline

- Course Administration
- Motivation and Success Stories
- The Reinforcement Learning Problem
- Course Outline and Goals

Online Resources and Books

- **Online Course Materials:**
  - Course Website: http://mallada.ece.jhu.edu/2025-summer-school/

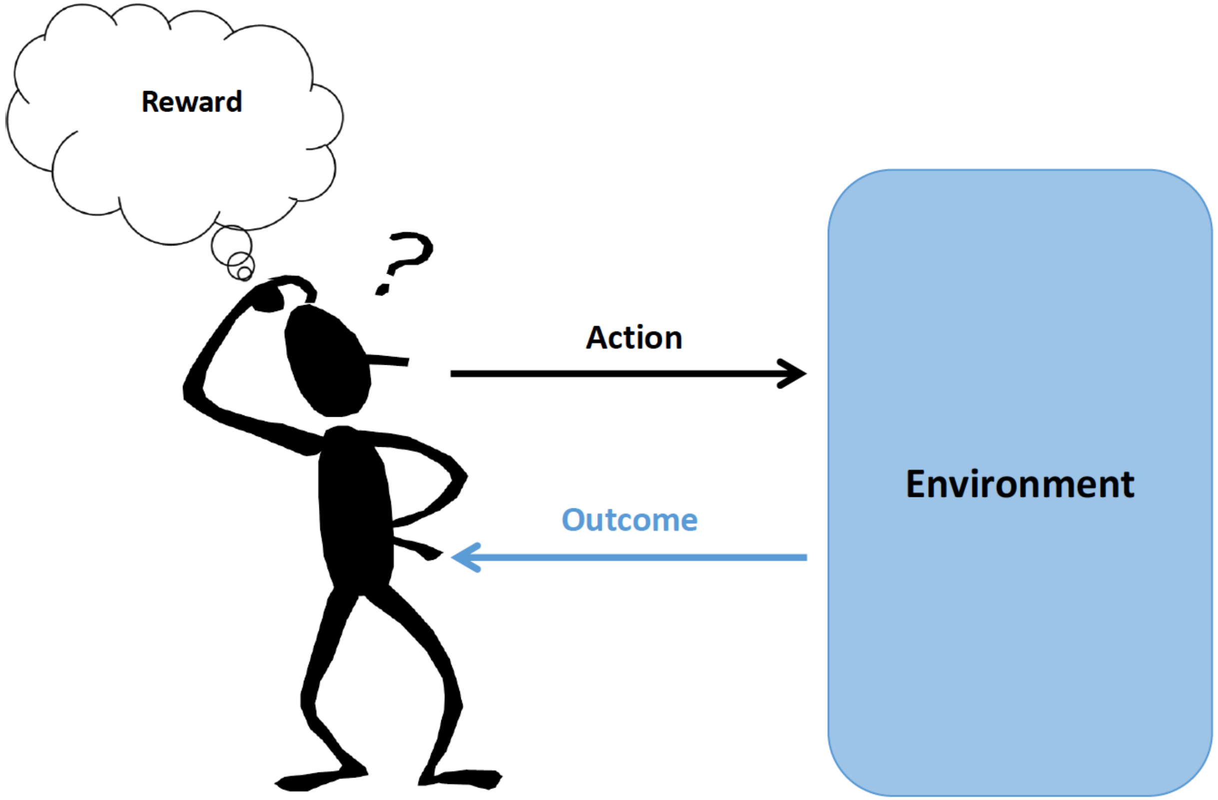
- **Closely Followed Books:**
  - Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction* (2nd Edition) ISBN-13: 978-0262039246.
    Online: http://www.incompleteideas.net/book/the-book-2nd.html
  - Csaba Szepesvári, *Algorithms for Reinforcement Learning,*
    Online: http://www.ualberta.ca/~szepesva/papers/RLAlgsInMDPs.pdf
  - Alekh Agarwal, Nan Jjiang, Sham M. Kakade, Wen Sun, *Reinforcement Learning: Theory and Algorithms (working draft)*
    Online: https://rltheorybook.github.io/rltheorybook_AJKS.pdf

- **Other Recommended Books:**
  - Sean Meyn, *Control Systems and Reinforcement Learning*
    Online: https://meyn.ece.ufl.edu/control-systems-and-reinforcement-learning/
  - Dimitri P. Bertsekas, *Dynamic Programming and Optimal Control, Vol. I* (4th Ed.) *and Vol. II* (4th Ed. Approx. Dynamic Programming).
  - Torr Lattimore and Csaba Szepesvári, *Bandits Algorithms*
    Online: https://tor-lattimore.com/downloads/book/book.pdf
  - David F. Anderson, Timo O. Seppäläinen and Benedek Valkó, Introduction to Probability

Outline

- Course Administration
- Motivation and Success Stories
- The Reinforcement Learning Problem
- Course Outline and Goals

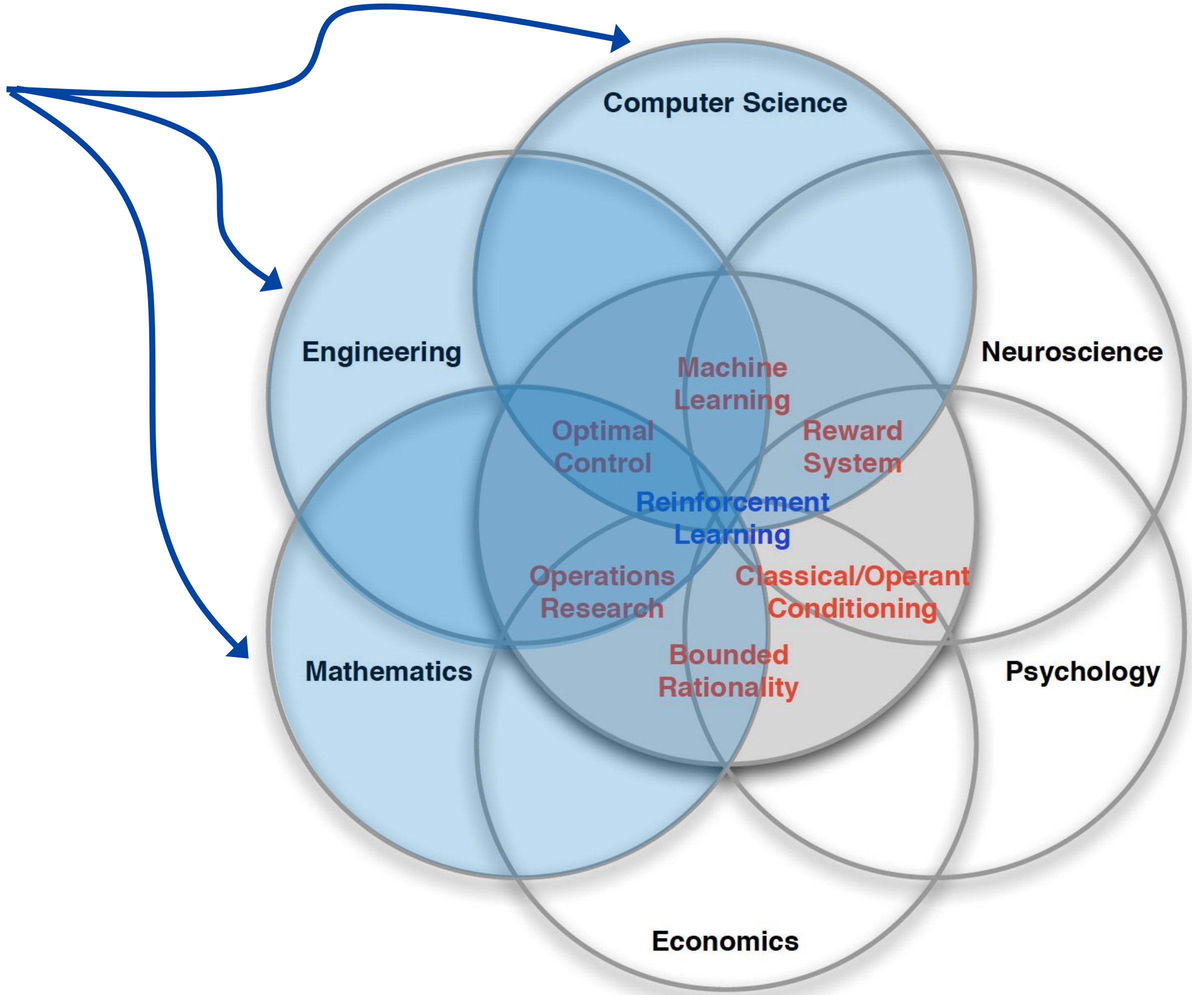
Reinforcement Learning



Learning to **attain a goal** through **sequential interactions** with a **poorly understood environment**

Scope of Reinforcement Learning

This class touches upon these



*source David Silver Deepmind

Early Success of Reinforcement Learning

‘92 Checkers: Chinook vs. Marion Tinsley

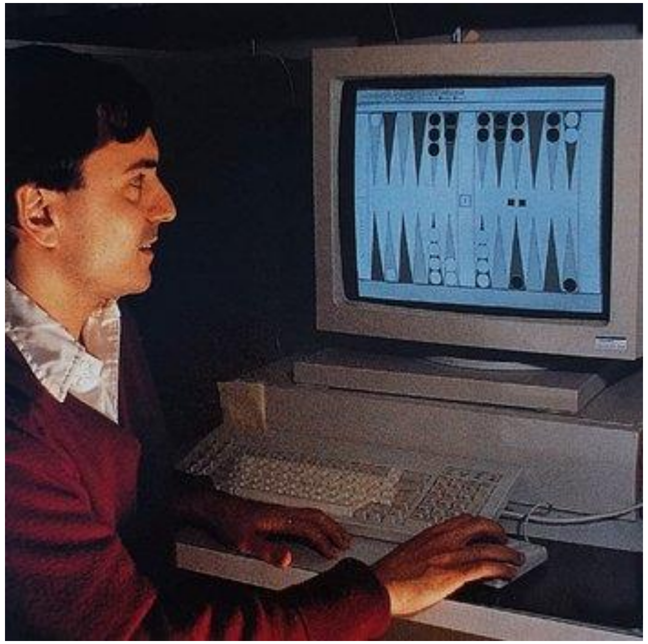
Chinook team (August 1992). From left to right: Duane Szafron, Joe Culberson, Paul Lu, Brent Knight, Jonathan Schaeffer, Rob Lake, and Steve Sutphen. Our checkers expert, Norman Treloar, is missing.

‘96 Chess: Deep Blue vs Gary Kasparov

Chess grandmaster beaten by AI predicts it will 'destroy' most jobs - Business Insider

Early Success of Reinforcement Learning

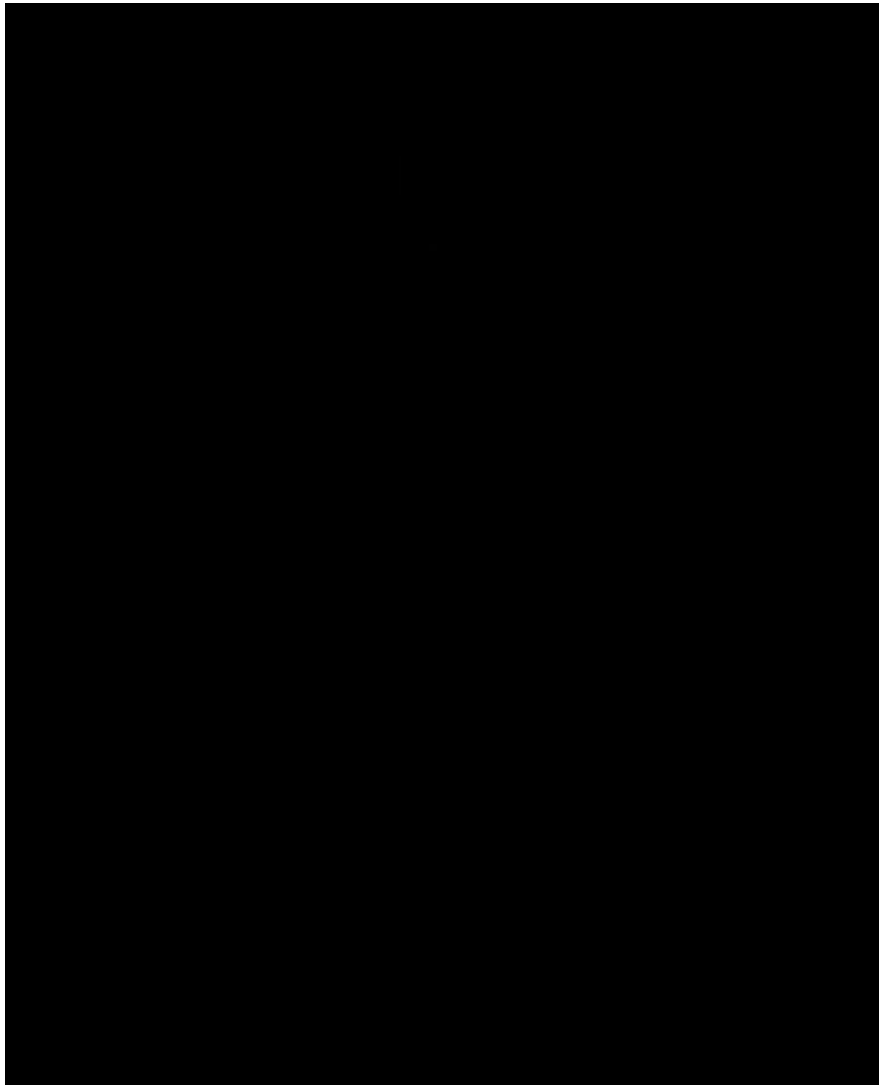
‘95 Backgammon: TD-Gammon “Knowledge Free Training”



Backgammon expert Kit Woolsey found that TD-Gammon's positional judgement, especially its weighing of risk against safety, was superior to his own or any human's.

TD-Gammon's excellent positional play was undercut by occasional poor endgame play. The endgame requires a more analytical approach, sometimes with extensive lookahead

The Advent of Deep-Learning – Google Deepmind's DQN



LETTER

doi:10.1038/nature14236

Human-level control through deep reinforcement learning

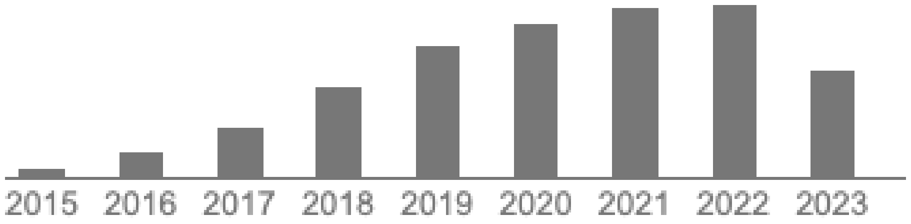
Volodymyr Mnih[1]*, Koray Kavukcuoglu[1]*, David Silver[1]*, Andrei A. Rusu[1], Joel Veness[1], Marc G. Bellemare[1], Alex Graves[1], Martin Riedmiller[1], Andreas K. Fidjeland[1], Georg Ostrovski[1], Stig Petersen[1], Charles Beattie[1], Amir Sadik[1], Ioannis Antonoglou[1], Helen King[1], Dharshan Kumaran[1], Daan Wierstra[1], Shane Legg[1] & Demis Hassabis[1]

The theory of reinforcement learning provides a normative account[1], deeply rooted in psychological[2] and neuroscientific[3] perspectives on animal behaviour, of how agents may optimize their control of an environment. To use reinforcement learning successfully in situations approaching real-world complexity, however, agents are confronted

agent is to select actions in a fashion that maximizes cumulative future reward. More formally, we use a deep convolutional neural network to approximate the optimal action-value function

$$Q^*(s,a) = \max_{\pi} \mathbb{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots | s_t = s, a_t = a, \pi\right],$$

Cited by 25626



The Advent of Deep-Learning – Google Deepmind’s DQN

2017 AlphaZero – Chess, Shogi, Go

2019 AlphaStar – Starcraft II

Article

Grandmaster level in StarCraft II using multi-agent reinforcement learning

https://doi.org/10.1038/s41586-019-1724-z

Received: 30 August 2019

Accepted: 10 October 2019

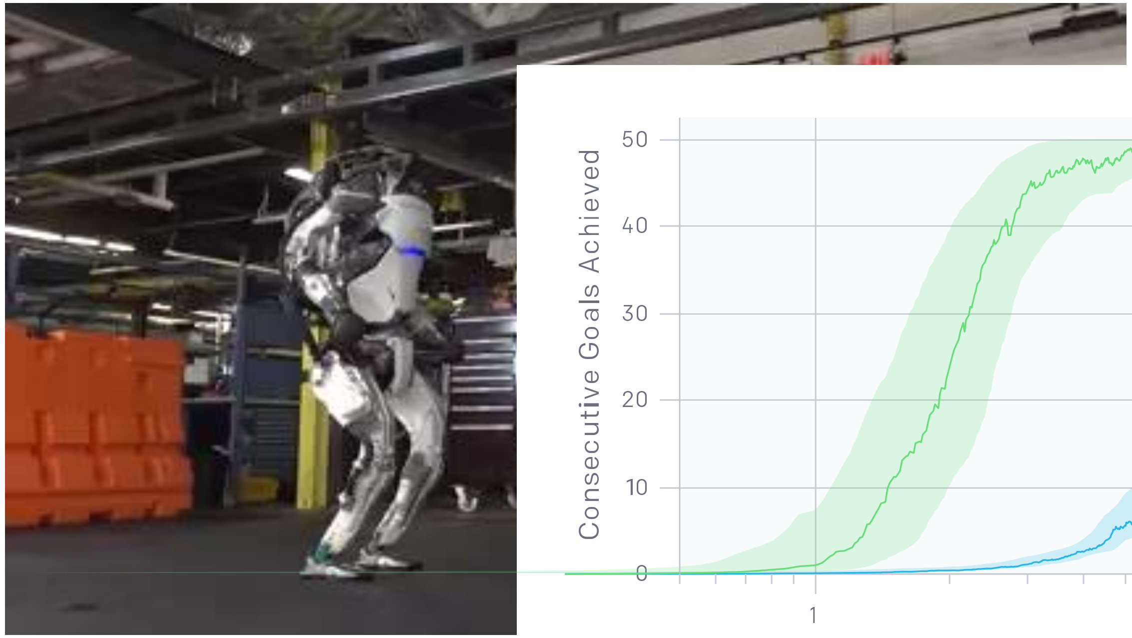
Published online: 30 October 2019

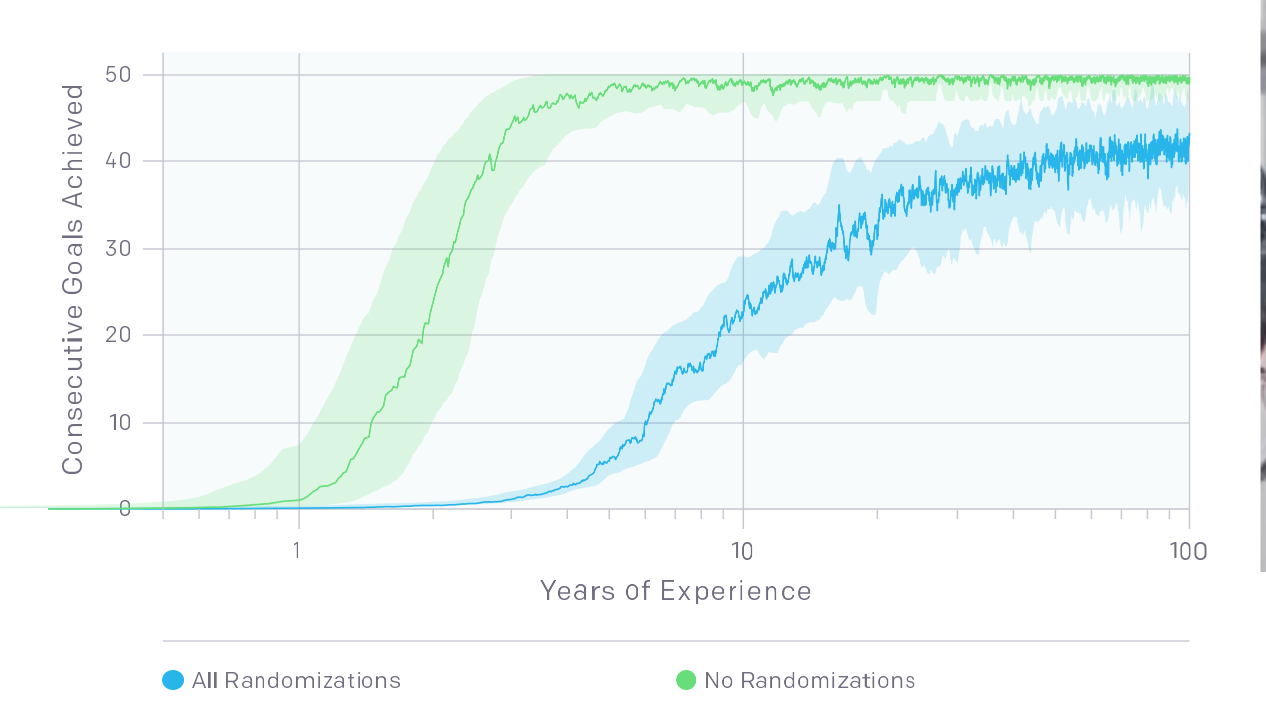
Oriol Vinyals[1,3*], Igor Babuschkin[1,3], Wojciech M. Czarnecki[1,3], Michaël Mathieu[1,3], Andrew Dudzik[1,3], Junyoung Chung[1,3], David H. Choi[1,3], Richard Powell[1,3], Timo Ewalds[1,3], Petko Georgiev[1,3], Junhyuk Oh[1,3], Dan Horgan[1,3], Manuel Kroiss[1,3], Ivo Danihelka[1,3], Aja Huang[1,3], Laurent Sifre[1,3], Trevor Cai[1,3], John P. Agapiou[1,3], Max Jaderberg[1], Alexander S. Vezhnevets[1], Rémi Leblond[1], Tobias Pohlen[1], Valentin Dalibard[1], David Budden[1], Yury Sulsky[1], James Molloy[1], Tom L. Paine[1], Caglar Gulcehre[1], Ziyu Wang[1], Tobias Pfaff[1], Yuhuai Wu[1], Roman Ring[1], Dani Yogatama[1], Dario Wünsch[2], Katrina McKinney[1], Oliver Smith[1], Tom Schaul[1], Timothy Lillicrap[1], Koray Kavukcuoglu[1], Demis Hassabis[1], Chris Apps[1,3] & David Silver[1,3*]

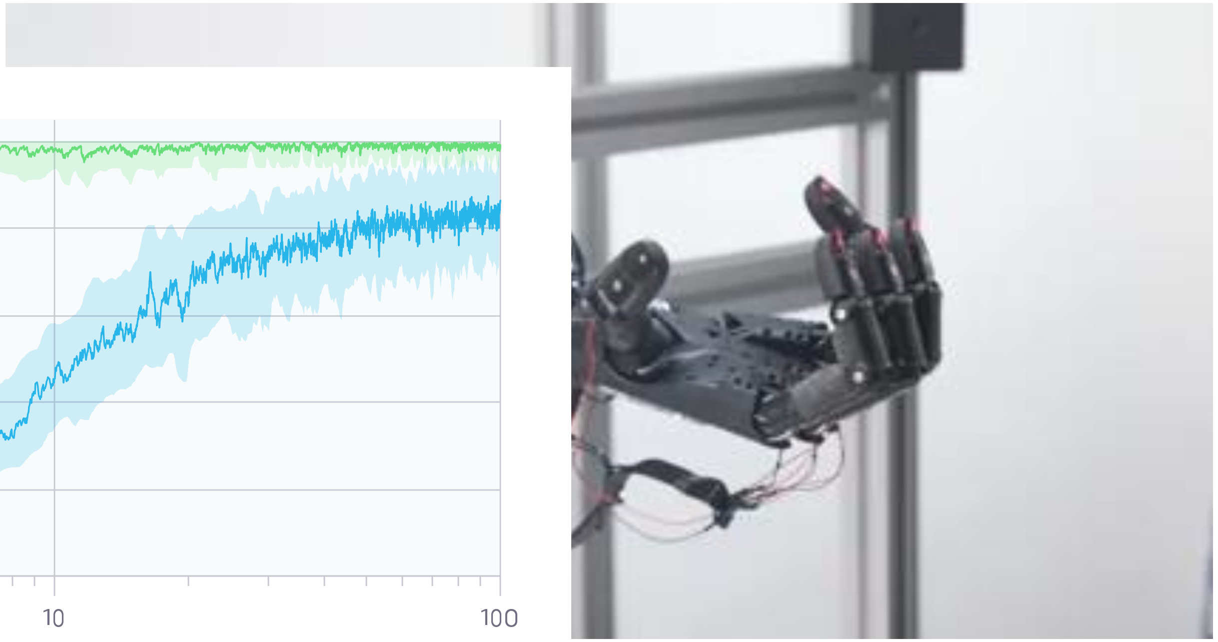
Expanding Horizons – Robotics?

Boston Dynamics

OpenAI – Rubik's Cube





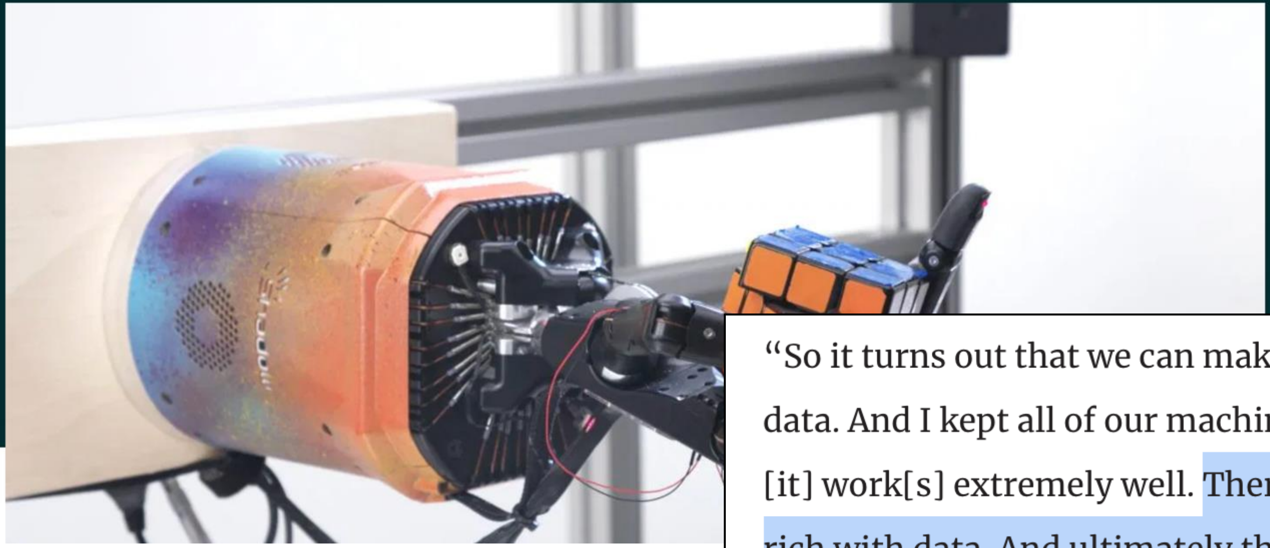


Based on simulated environments and/or require guidance from experts

Expanding Horizons – Robotics is HARD

OpenAI disbands its robotics research team

Kyle Wiggers @Kyle_L_Wiggers July 16, 2021 11:24 AM



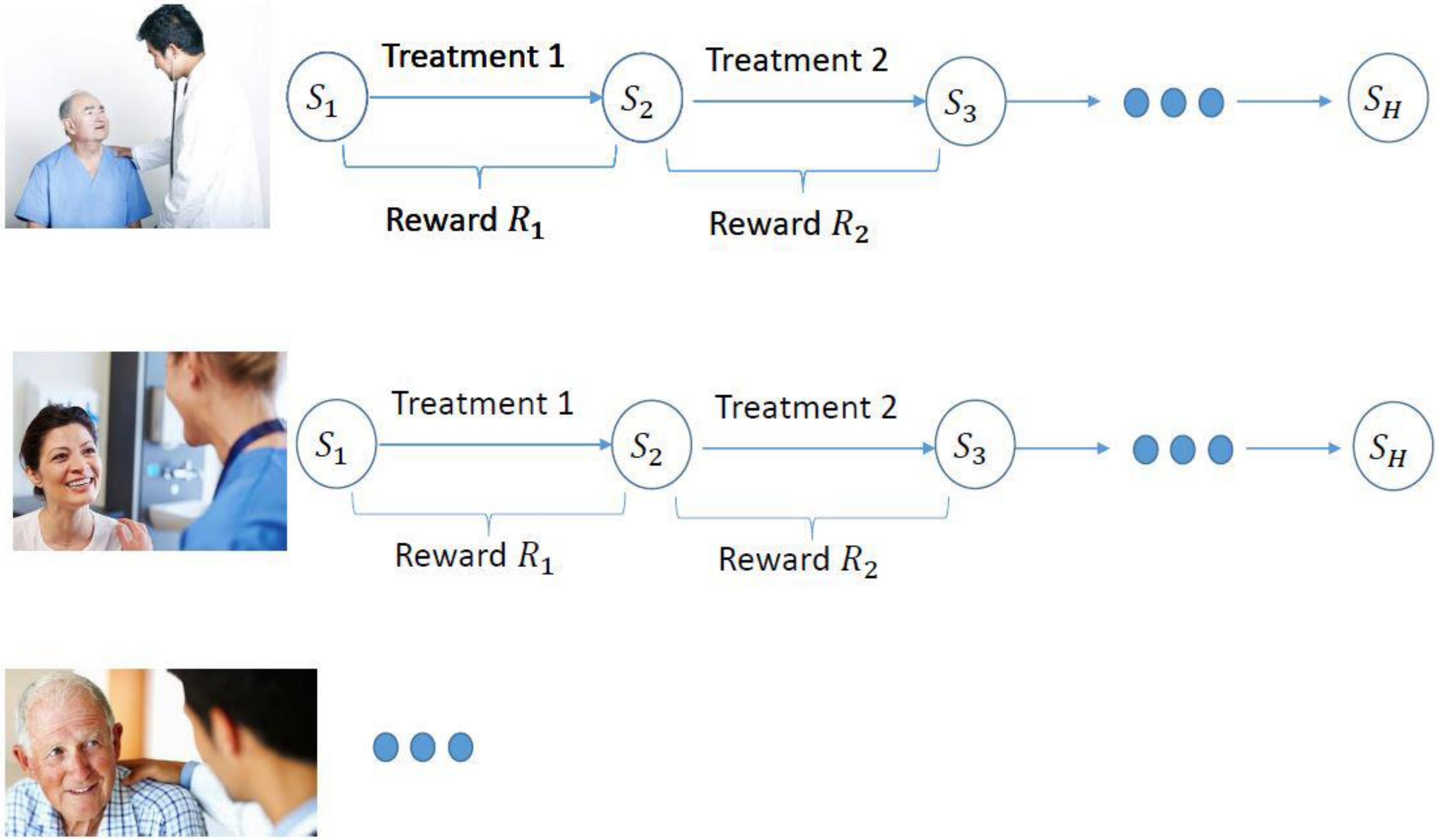
"So it turns out that we can make a gigantic progress whenever we have access to data. And I kept all of our machinery unsupervised, [using] reinforcement learning — [it] work[s] extremely well. There [are] actually plenty of domains that are very, very rich with data. And ultimately that was holding us back in terms of robotics," Zaremba said. "The decision [to disband the robotics team] was quite hard for me. But I got the realization some time ago that actually, that's for the best from the perspective of the company."

Based on simulated environments and/or require guidance from experts

Expanding Horizons – Robotics is HARD

Expanding Horizons – Robotics is HARD

Expanding Horizons – Management of Chronic Diseases?



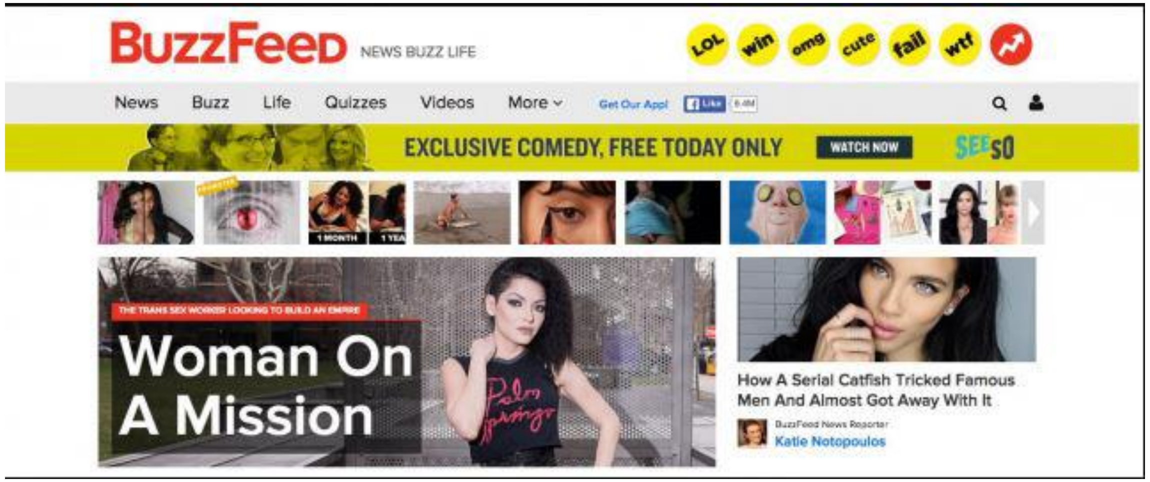
Various researchers are working on mobile health interventions

Expanding Horizons – Intelligent Tutoring Systems?

*Emma Brunskill's, Stanford

Expanding Horizons – Beyond Myopia in E-Commerce?

- Online marketplaces and web services have repeated interactions with users but are deigned to optimize the next interaction.
- RL provides a framework for optimizing the cumulative value generated by such interactions.
- How useful will this turn out to be?



*Daniel Russo, Columbia

Outline

- Course Administration
- Motivation and Success Stories
- The Reinforcement Learning Problem
- Course Outline and Goals

Reinforcement Learning Involves

- **Exploration**
- Delayed consequences
- Optimization
- Generalization

Exploration

- Learning about the world by making decisions
  - Agent as scientist
  - Learn to ride a bike by trying (and failing)
- Censored data
  - Only get a reward (label) for decision made
  - Lack of counter factual
- Decisions impact what we learn about
  - Should I do a PhD?

Reinforcement Learning Involves

- Exploration
- **Delayed consequences**
- Optimization
- Generalization

Delayed consequences

- Decisions now can impact things much later...
  - Saving for retirement
  - Finding a key in video game Montezuma's revenge
- Introduces two challenges
  - **When planning:** decisions involve reasoning about not just **immediate** benefit of a decision but also its **longer-term** ramifications
  - **When learning:** temporal credit assignment is hard (what caused later high or low rewards?)

Reinforcement Learning Involves

- Exploration
- Delayed consequences
- **Optimization**
- Generalization

Optimization

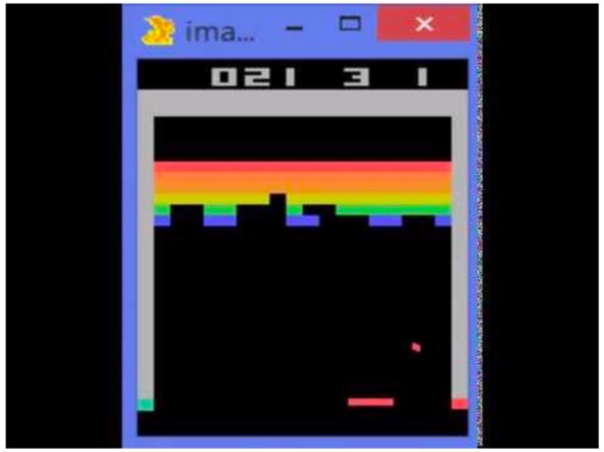
- Goal is to find an optimal way to make decisions
  - Yielding best outcomes or at least very good outcomes
- Explicit notion of utility of decisions
- Example: Finding minimum distance route between two cities given network of roads

Reinforcement Learning Involves

- Exploration
- Delayed consequences
- Optimization
- **Generalization**

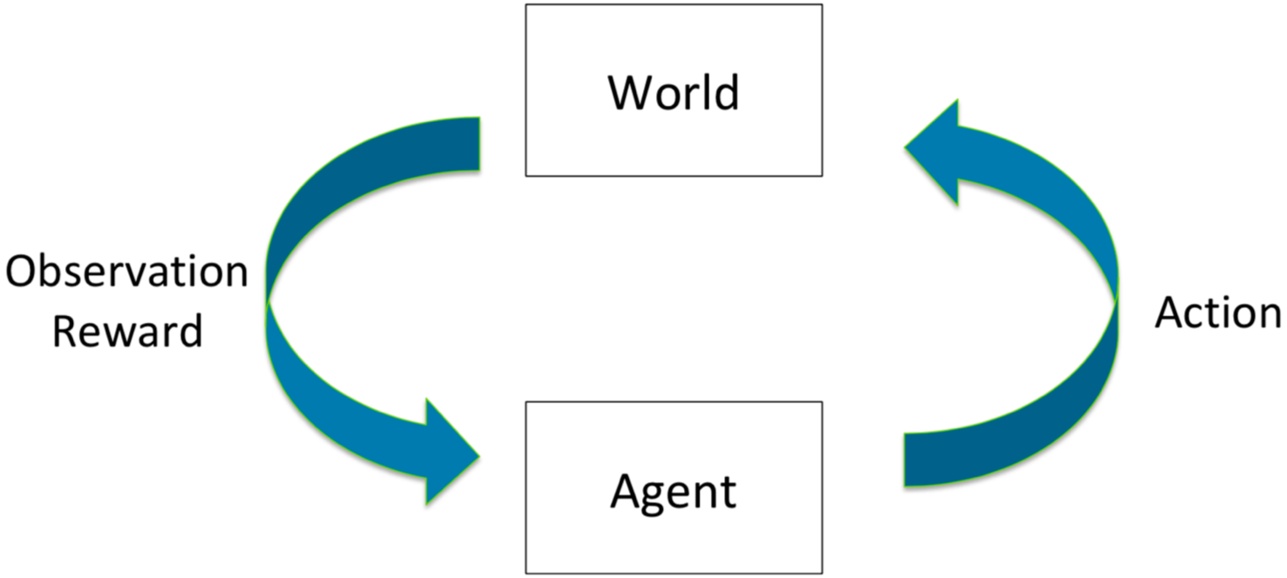
Generalization

- Policy is mapping from past experience to action
- Why not just pre-program a policy?



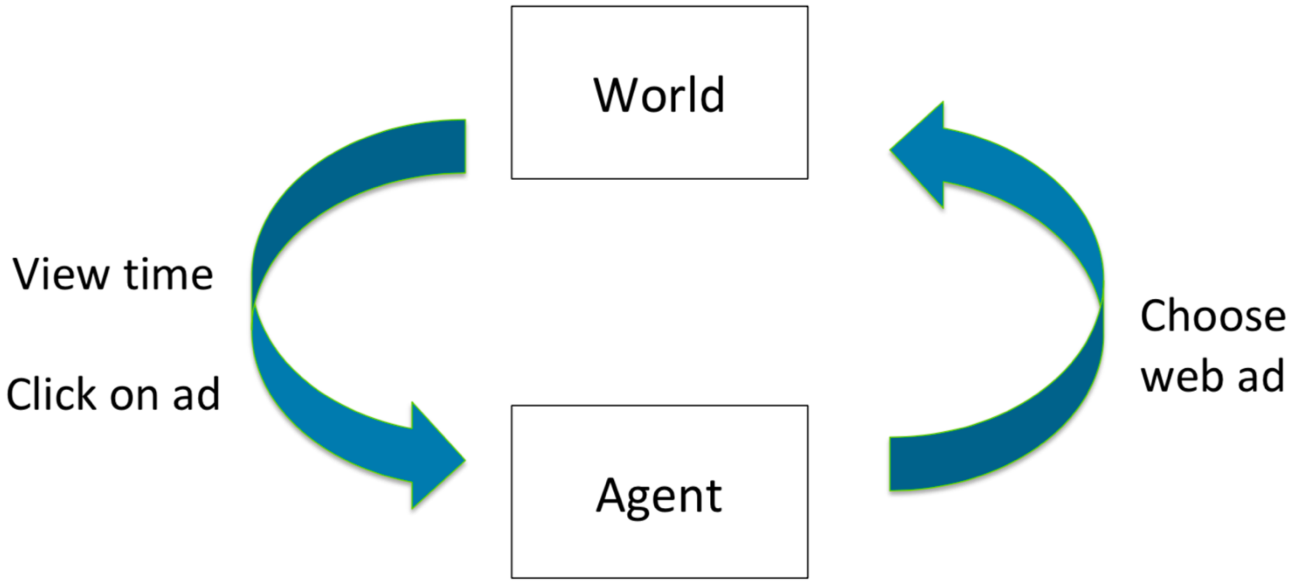
- How many possible images are there?
  - $(256^{100\times200})^3$

Sequential Decision Making



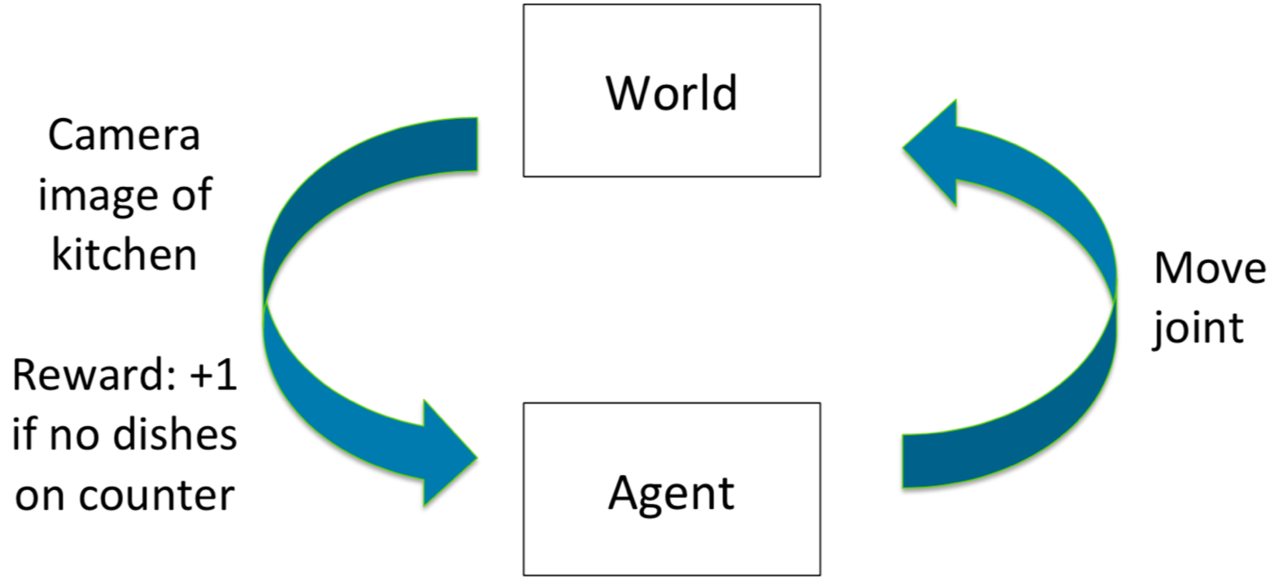
- Goal: Select actions to maximize total expected future reward
- May require balancing immediate & long-term rewards

Example: Web Advertising



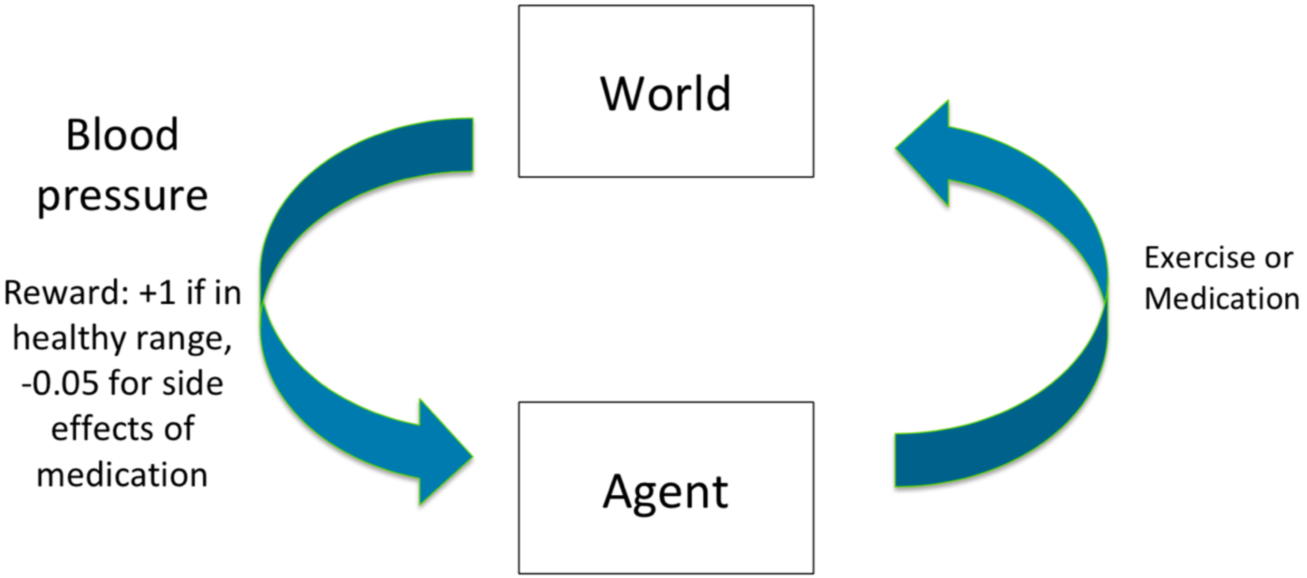
- Goal: Select actions to maximize total expected future reward
- May require balancing immediate & long-term rewards

Example: Robot Unloading Dishwasher



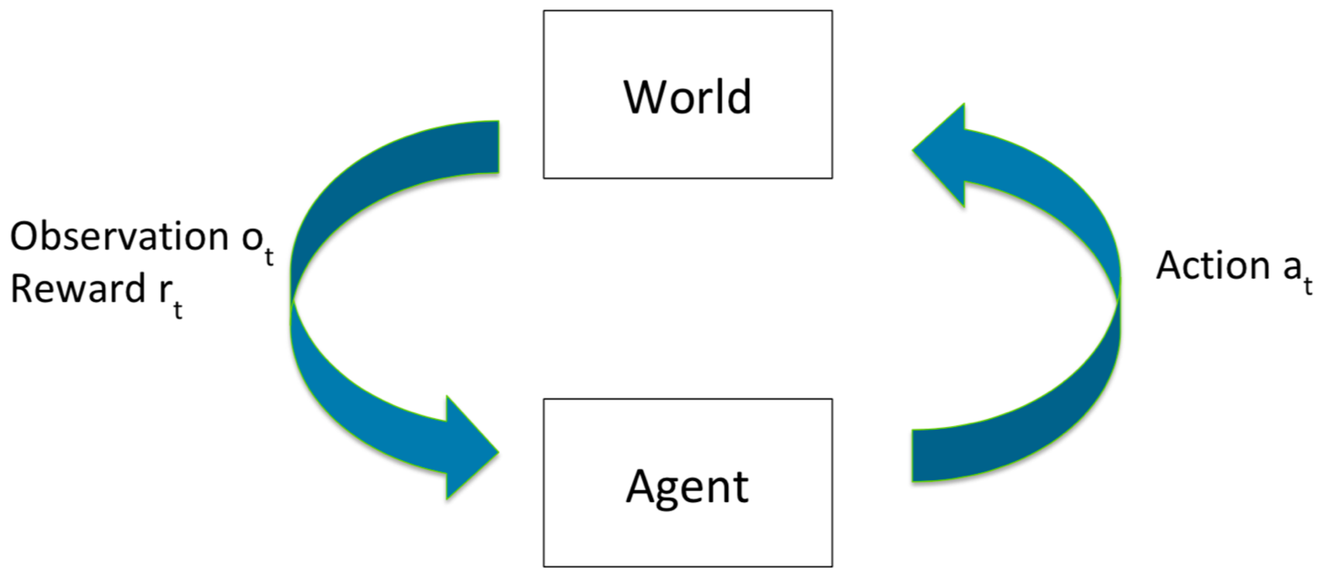
- Goal: Select actions to maximize total expected future reward
- May require balancing immediate & long-term rewards

Example: Blood Pressure Control



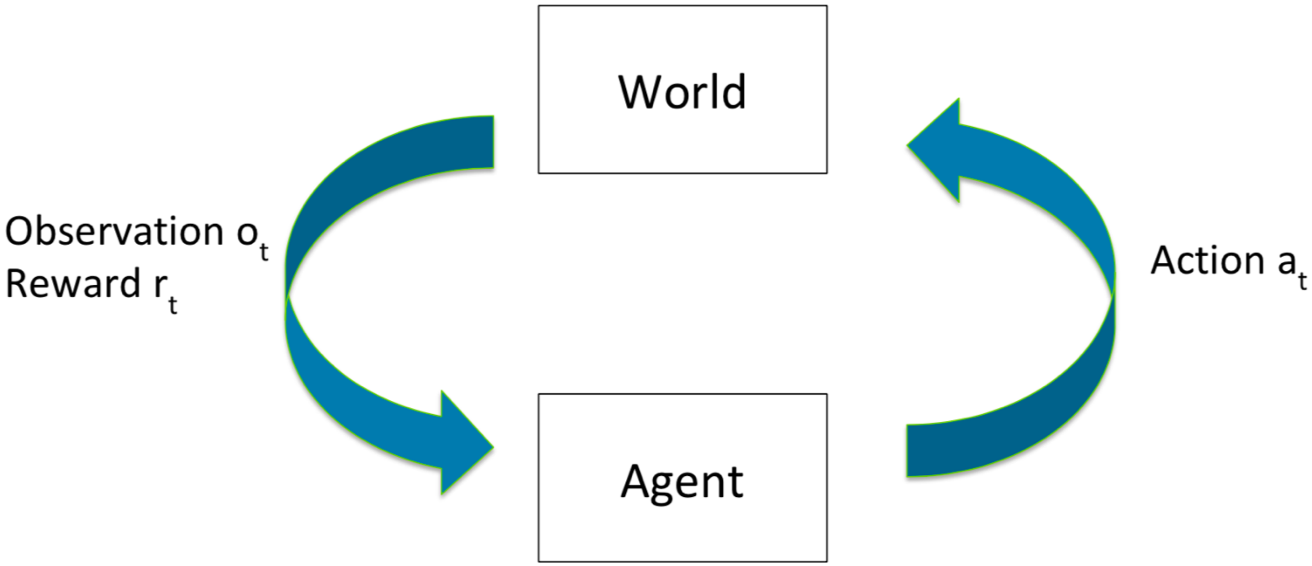
- Goal: Select actions to maximize total expected future reward
- May require balancing immediate & long-term rewards

Agent and the World, in Discrete Time



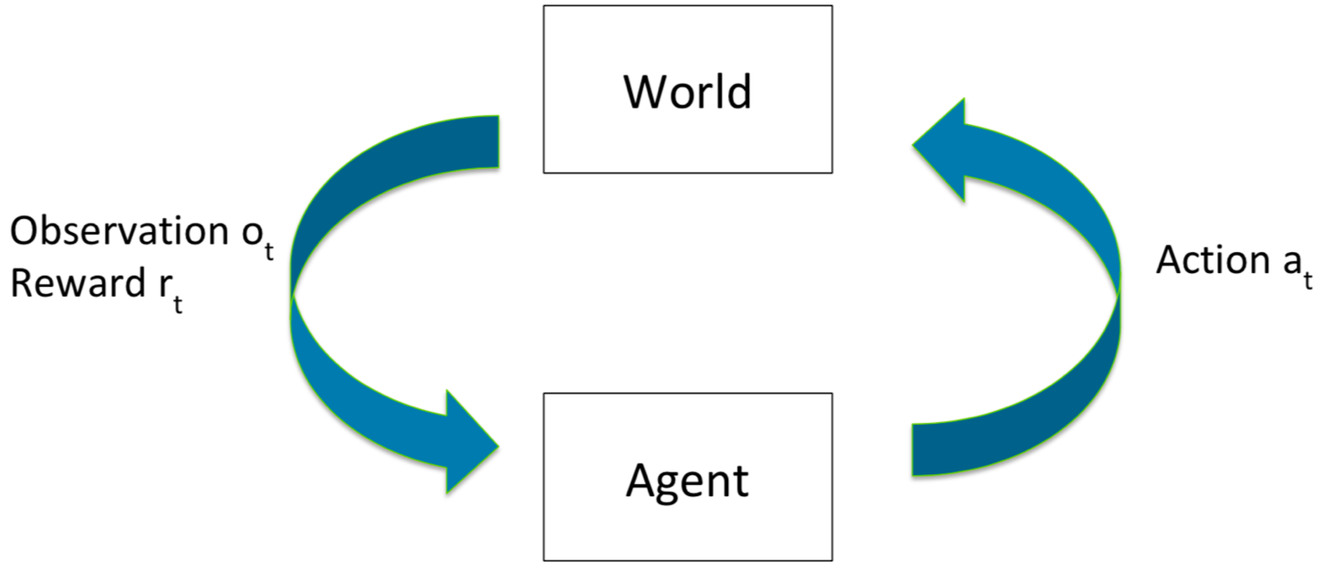
- Each time step t:
  - Agent takes an action a_t
  - World updates given action a_t , emits observation o_t and reward r_t
  - Agent receives observation o_t and reward r_t

History: Sequence of Past Observations, Actions & Rewards



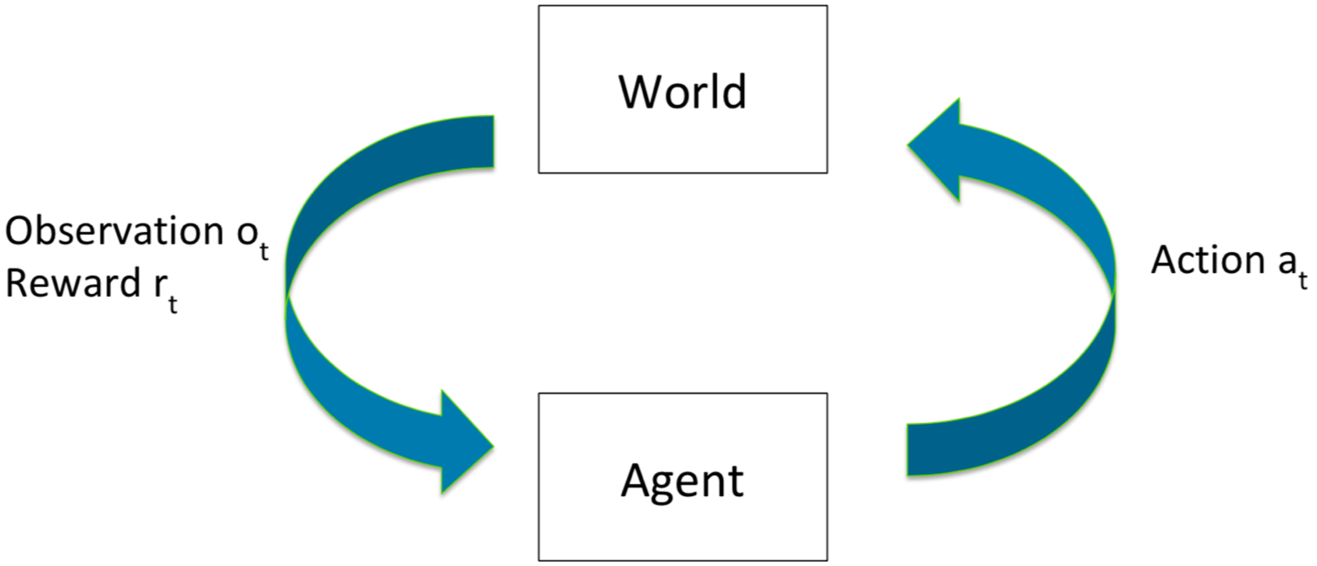
- History $h_t = (a_1, o_1, r_1, \dots, a_t, o_t, r_t)$
  - Agent chooses action based on history
  - State is information assumed to determine what happens next
  - Function of history: $s_t = f(h_t)$

World State



- This is true state of the world used to determine how world generates next observation and reward
- Often hidden or unknown to agent
- Even if known may contain information not needed by agent

Agent State: Agent's Internal Representation



- What the agent / algorithm uses to make decisions about how to act
- Generally, a function of the history: $s_t = f_a(h_t)$
- Could include meta information like state of algorithm (how many computations executed, etc.) or decision process (how many decisions left until an episode ends)

Markov Assumption

- Information state: sufficient statistic of history
- State s_t is Markov if and only if:

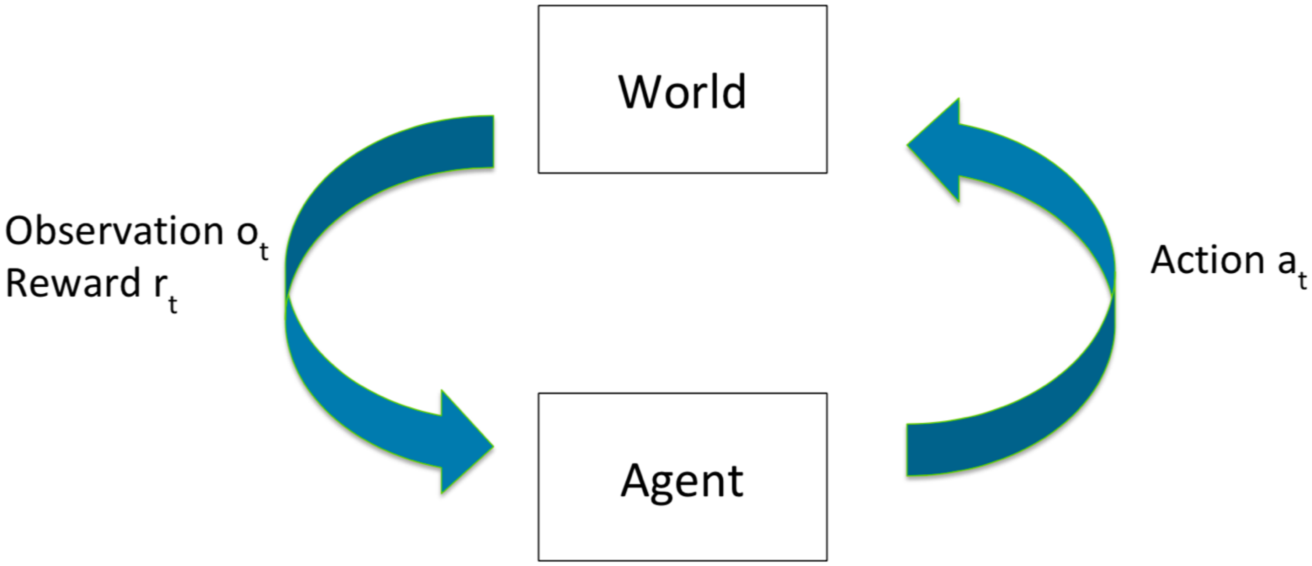
$$p(s_{t+1}|s_t\,,a_t) \;=\; p(s_{t+1}|h_t\,,a_t)$$

- Future is independent of past given present

Why is Markov Assumption Popular?

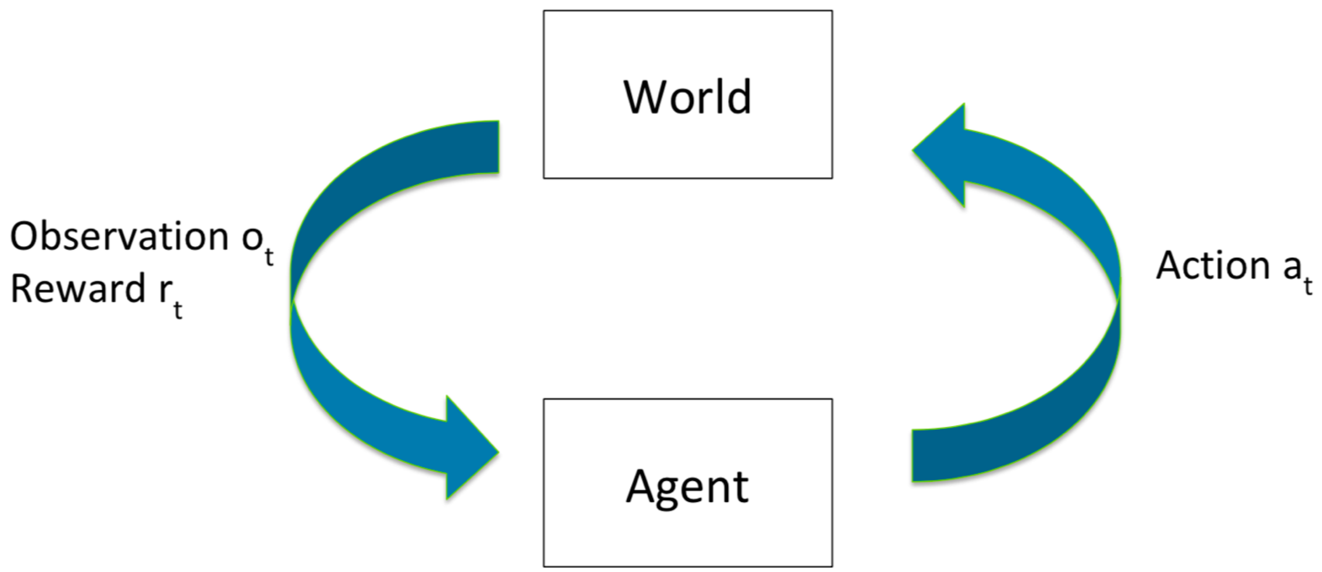
- Can always be satisfied
  - Setting state as history always Markov: $s_t = h_t$
- In practice often assume most recent observation is sufficient statistic of history: $s_t = o_t$
- State representation has big implications for:
  - Computational complexity
  - Data required
  - Resulting performance

Full Observability / Markov Decision Process (MDP)



- Environment and world state $s_t = o_t$

Types of Sequential Decision Processes



- Is state Markov? Is world partially observable? (POMDP)
- Are dynamics deterministic (Lin or Non-Lin Systems) or stochastic (MDP/POMDP)?
- Do actions influence only immediate reward (Bandits) or reward and next state (MDP)?

Outline

- Course Administration
- Motivation and Success Stories
- The Reinforcement Learning Problem
- Course Outline and Goals

About This Course

- **Course Goals:** Specific Outcomes for this course are that
  - Students will learn to model and solve optimal control problems using RL.
  - Students will learn the mathematical theory and conditions that guarantee optimality and convergence RL algorithms.
  - Students will learn different algorithms including on- and off-policy methods, Monte Carlo methods, temporal difference, policy gradient, etc.

- **Warning**
  - This is a graduate level course with a focus on Theory. Though there will be a balance with practical experience.
  - We will prove theorems when we can. The emphasis will be on precise understand of why methods work and why they may fail completely in simple cases.
  - There are tons of engineering tricks to Deep RL. I won't cover these.

Tentative Course Outline

- Exploration
- Delayed consequences
- Optimization
- Generalization

- Model-based Decision Making
  - 1 Markov Decision Processes
  - 2 Dynamic Programming
  - 3 Model-based Policy Evaluation
  - 4 Model-based Policy Improvement
- Elementary Reinforcement Learning
  - 1 Decision Making under Uncertainty
  - 2 Multi-armed Bandits
  - 3 Model-Free Prediction
  - 4 Model-Free Control
- Reinforcement Learning in Practice
  - 1 Value Function Approximation
  - 2 Policy Gradient Methods
  - 3 Integrating Learning and Planning

Tentative Course Outline

- **Exploration**
- Delayed consequences
- Optimization
- Generalization

- Model-based Decision Making
  - 1 Markov Decision Processes
  - 2 Dynamic Programming
  - 3 Model-based Policy Evaluation
  - 4 Model-based Policy Improvement
- Elementary Reinforcement Learning
  - **1 Decision Making under Uncertainty**
  - **2 Multi-armed Bandits**
  - **3 Model-Free Prediction**
  - 4 Model-Free Control
- Reinforcement Learning in Practice
  - 1 Value Function Approximation
  - 2 Policy Gradient Methods
  - 3 Integrating Learning and Planning

Tentative Course Outline

- Exploration
- **Delayed consequences**
- Optimization
- Generalization

- Model-based Decision Making
  - 1 Markov Decision Processes
  - **2 Dynamic Programming**
  - **3 Model-based Policy Evaluation**
  - **4 Model-based Policy Improvement**
- Elementary Reinforcement Learning
  - 1 Decision Making under Uncertainty
  - 2 Multi-armed Bandits
  - 3 Model-Free Prediction
  - 4 Model-Free Control
- Reinforcement Learning in Practice
  - 1 Value Function Approximation
  - 2 Policy Gradient Methods
  - 3 Integrating Learning and Planning

Tentative Course Outline

- Exploration
- Delayed consequences
- **Optimization**
- Generalization

- Model-based Decision Making
  - 1 Markov Decision Processes
  - **2 Dynamic Programming**
  - 3 Model-based Policy Evaluation
  - **4 Model-based Policy Improvement**
- Elementary Reinforcement Learning
  - 1 Decision Making under Uncertainty
  - 2 Multi-armed Bandits
  - 3 Model-Free Prediction
  - **4 Model-Free Control**
- Reinforcement Learning in Practice
  - 1 Value Function Approximation
  - **2 Policy Gradient Methods**
  - 3 Integrating Learning and Planning

Tentative Course Outline

- Exploration
- Delayed consequences
- Optimization
- **Generalization**

- Model-based Decision Making
  - 1 Markov Decision Processes
  - 2 Dynamic Programming
  - 3 Model-based Policy Evaluation
  - 4 Model-based Policy Improvement
- Elementary Reinforcement Learning
  - 1 Decision Making under Uncertainty
  - 2 Multi-armed Bandits
  - 3 Model-Free Prediction
  - 4 Model-Free Control
- **Reinforcement Learning in Practice**
  - **1 Value Function Approximation**
  - **2 Policy Gradient Methods**
  - **3 Integrating Learning and Planning**

Thanks!

Enrique Mallada
mallada@jhu.edu
http://mallada.ece.jhu.edu