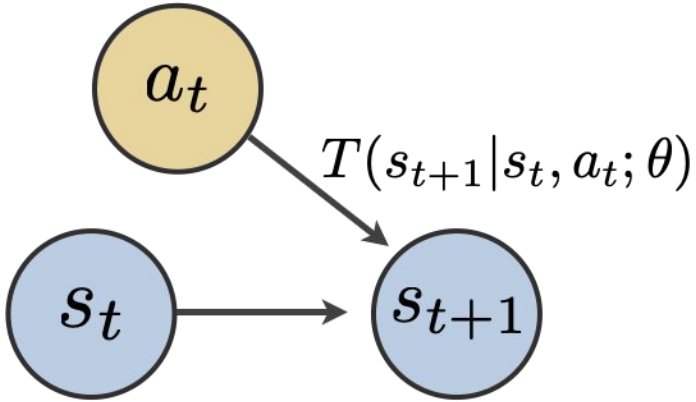
Reinforcement Learning for Healthcare

Taylor W. Killian
University of Toronto / Vector Institute

W'20 CSC 2541 - Lecture 7

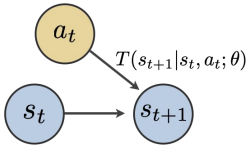
20 February 2020

Foundations*



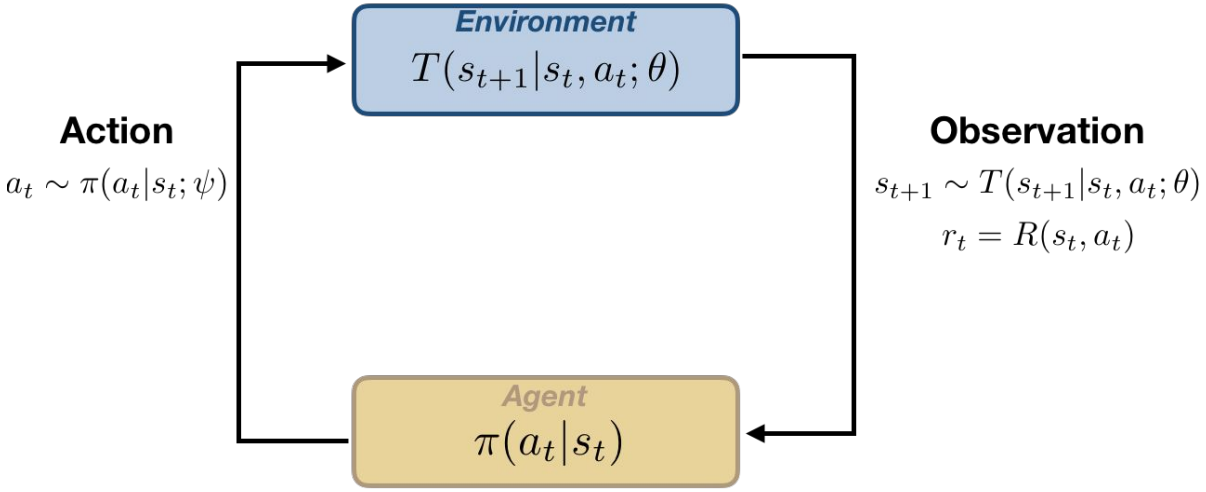
The basic unit of any RL or decision-making problem is the transition between observations after applying some action

* This lecture does not provide an in-depth introduction to RL. For a better and more thorough introduction. Please review Sutton and Barto [2017] or David Silver's RL lectures (on youtube)



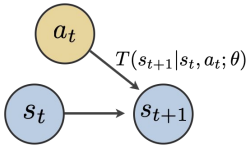
Foundations

$$\mathcal{M} = \{S, A, T, R, \gamma\}$$



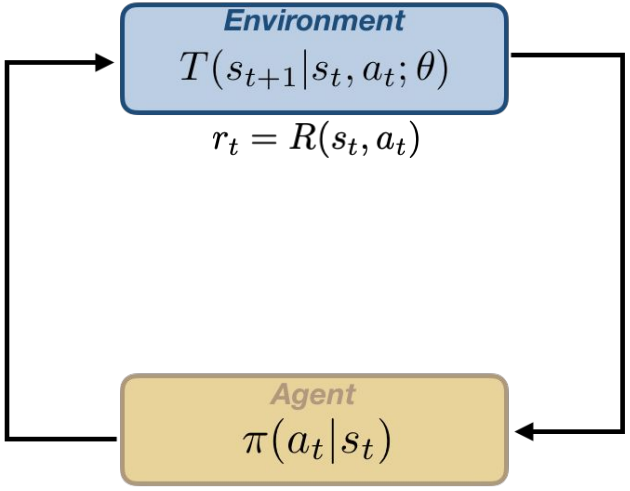
Repeated interactions with an environment introduce a series of observations that can be ordered to accomplish some task.

With a specified task, each interaction with the environment can be defined to provide some auxiliary signal reflecting the utility (cost) of subsequent actions



Foundations

$$\mathcal{M} = \{S, A, T, R, \gamma\}$$

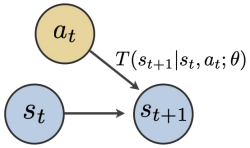


$$V^{\pi}(s_t) = \mathbb{E}\left[\sum_{i=t}^{T} \gamma^{i-1} r_i\right]$$

$$V^{\pi}(s_t) = \max_a \; Q^{\pi}(s_t, a_t)$$

$$Q^{\pi}(s_t, a_t) = R(s_t, a_t) + \gamma \max_a Q^{\pi}(s_{t+1}, a)$$

Policies π are optimized by maximizing the "expected future reward" based on some discounted horizon of the reward gained in subsequent interactions with the environment



Foundations

$$\mathcal{M} = \{S, A, T, R, \gamma\}$$

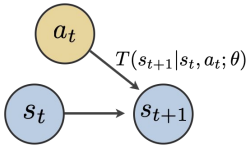
Policy Iteration
Policy Gradients

$$V^{\pi}(s_t) = \mathbb{E}\left[\sum_{i=t}^{T} \gamma^{i-1} r_i\right]$$

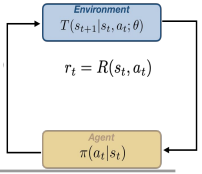
Value Iteration
Q-Learning

$$Q^{\pi}(s_t, a_t) = R(s_t, a_t) + \gamma \max_a Q^{\pi}(s_{t+1}, a)$$

Based on the choice of policy optimization algorithm, you will utilize one or the other value function representation



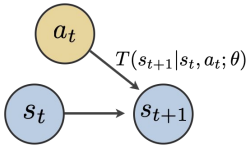
What's Learnable in RL?



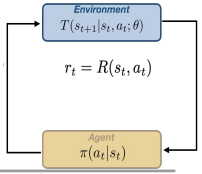
$$\mathcal{M} = \{S, A, T, R, \gamma\}$$

$$\pi(a_t|s_t; \psi)$$

Beyond parameters for the policy, depending on your particular research question and the format of your data, you may be able to learn or infer other elements of the MDP



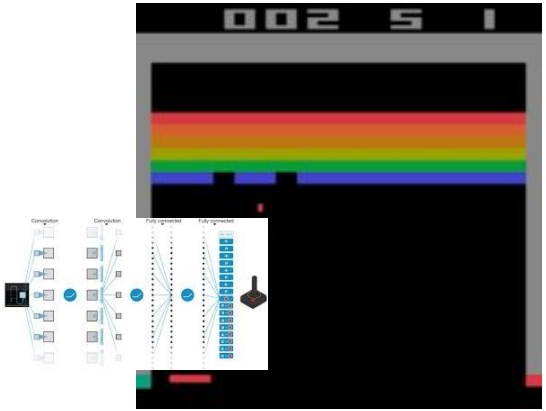
Open Questions in RL



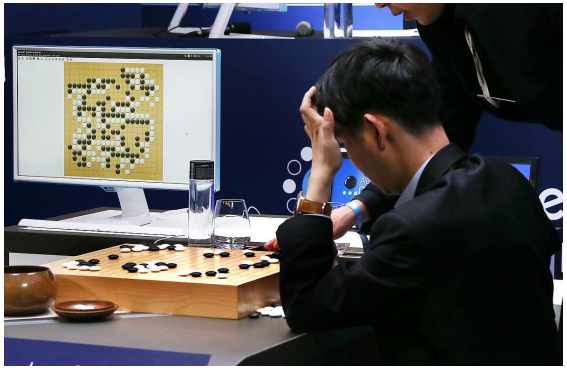
- Sample complexity
- Exploration v. Exploitation
- Representation Learning
- IRL vs. interactive feed-forward design
- Safety / Causality

  (i.e. is a developing policy guaranteed to “do no harm”)
- Transfer Learning
- Reward design
- Off-policy vs on-policy

Why RL?



DeepMind “solves” Atari



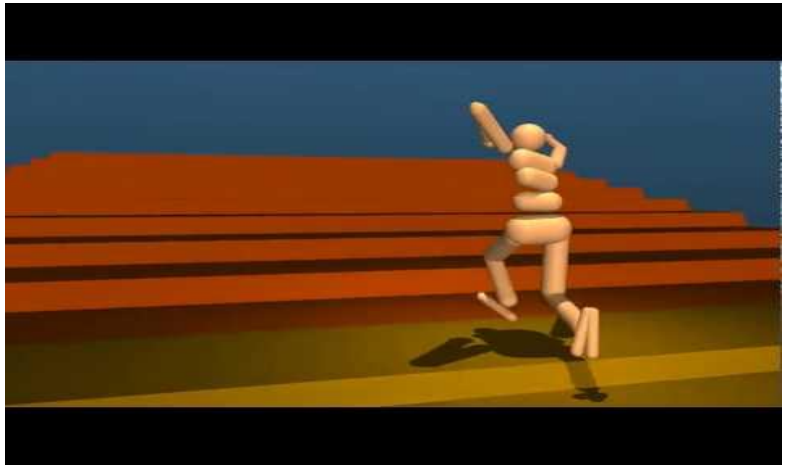
AlphaGo defeats Lee Sedol

OpenAI beats Humans at Dota II

Recent exciting demonstrations of RL successes have showed the flexibility and representational power when learning complex behaviors from sequential observations and guiding utility (cost) signal

Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." *Nature* 518.7540 (2015): 529-533.
Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." *Nature* 529.7587 (2016): 484.
Berner, Christopher, et al. "Dota 2 with Large Scale Deep Reinforcement Learning." *arXiv preprint arXiv:1912.06680* (2019).

Why RL?



Emergent Locomotion Behavior

Complex Motor Skills in the Real World

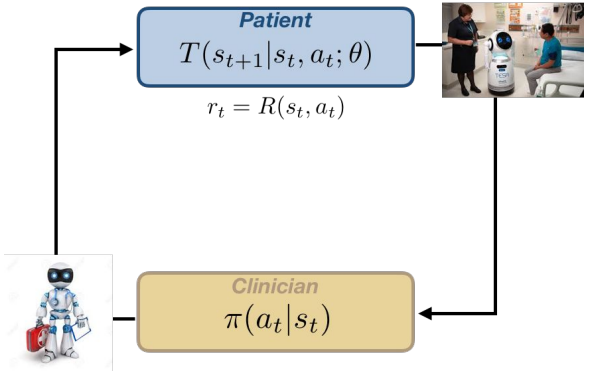
Recent exciting demonstrations of RL successes have showed the flexibility and representational power when learning complex behaviors from sequential observations and guiding utility (cost) signal

Heess, Nicolas, et al. "Emergence of locomotion behaviours in rich environments." *arXiv preprint arXiv:1707.02286* (2017).
Zhang, Marvin, et al. "SOLAR: deep structured representations for model-based reinforcement learning." *arXiv preprint arXiv:1808.09105* (2018).

Healthcare as Sequential Decision Making

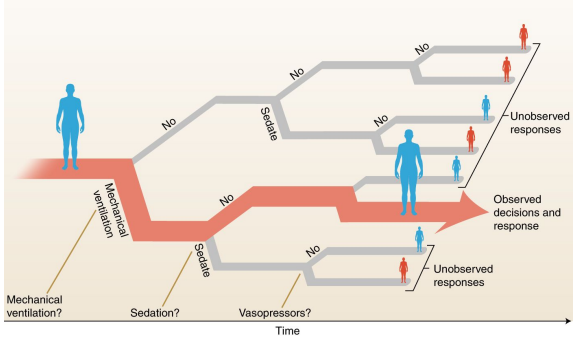
The practice of medicine is inherently a sequential decision making problem:

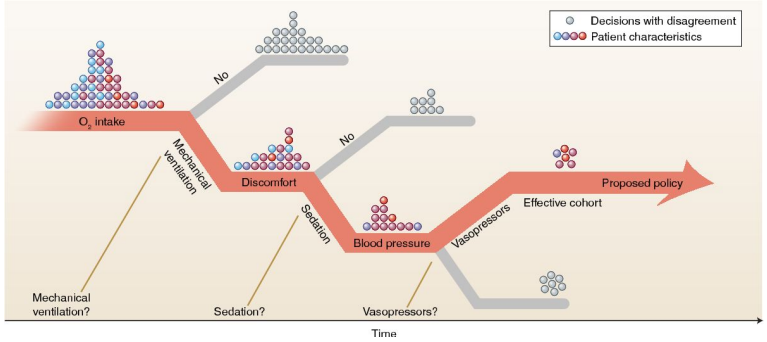
- **Clinicians, with their best understanding of a patient's status, propose a treatment.**
- **The patient's status may or may not change as a result of the prescribed treatment.**
- **Eventual outcomes are a noisy measure of the affect the prior treatment decisions had on the patient.**



Can we utilize the Reinforcement Learning framework to describe/explain and augment clinician decision making?

Limitations of RL in Healthcare





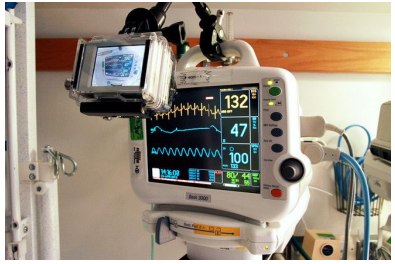
Learning optimal treatment policies from observational data--*an offline and off-policy RL task*--is complicated by:

1. the inability to explore, and
2. a shrinking volume of training observations as top strategies are discovered

These two limitations severely complicate the ability to develop proactive RL algorithms/policies that suggest ***what to do***

Image Credits: *Gottesman, Omer, et al. "Guidelines for reinforcement learning in healthcare." Nat Med 25.1 (2019): 16-18.*

Limitations of RL in Healthcare



Beyond the inability to explore and diminishing training support there are other significant challenges to using RL algorithms for healthcare:

1. Unclear objectives
   a. What is a stable and clinically relevant reward?
   b. What motivates the clinician when there are competing priorities?

2. Biased measurements and noisy partial observations
   a. Oftentimes routine tests and measurements are missing which may indicate the clinician's belief about the patient's condition

3. Clinical practice varies widely between doctors and institutions.
   a. There is no clear understanding of what the best "expert" policy is to learn from.
   b. Sets of observations may differ between institutions

Limitations of RL in Healthcare (RL4H)

Beyond the inability to explore and diminishing training support there are other significant challenges to using RL algorithms for healthcare:

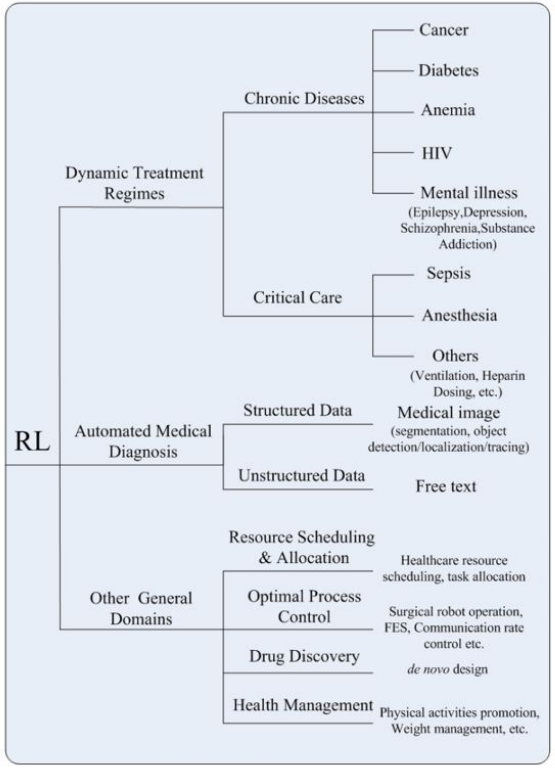
1. Unclear objectives
   a. What is a stable and clinically relevant reward?

In order to make RL in Healthcare feasible, we need to fundamentally rethink how we utilize standard RL forms and frameworks to develop sequentially relevant insights into clinical decision making

institutions.

a. There is no clear understanding of what the best "expert" policy is to learn from.
b. Sets of observations may differ between institutions

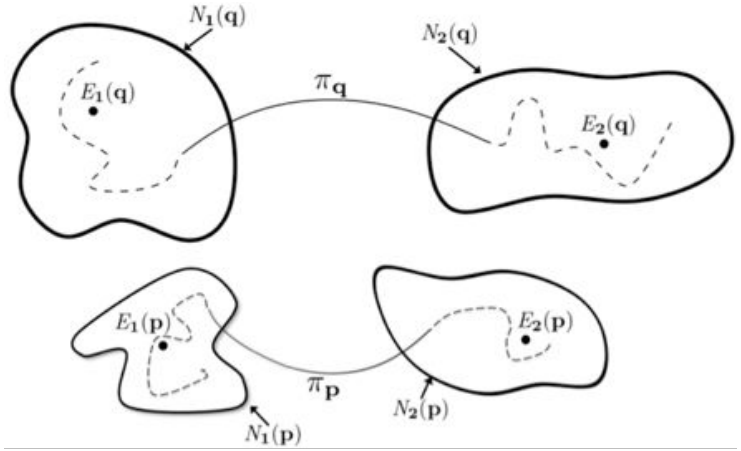
RL within Healthcare is not Totally New



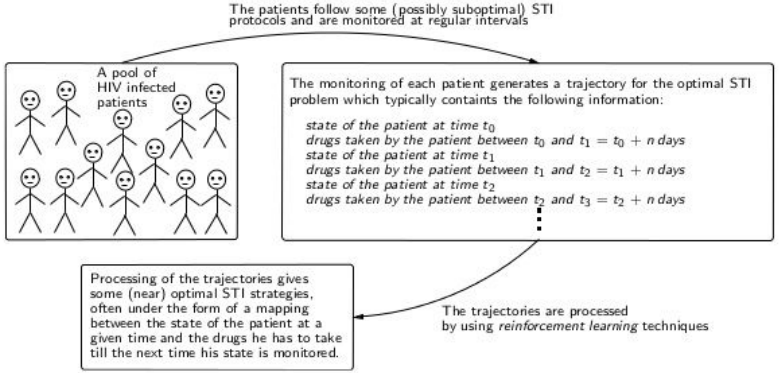
From "Reinforcement Learning in Healthcare: A Survey"; Yu, Liu and Nemati [2019]

https://arxiv.org/abs/1908.08796

Clinical Data Based Optimal STI Strategies for HIV



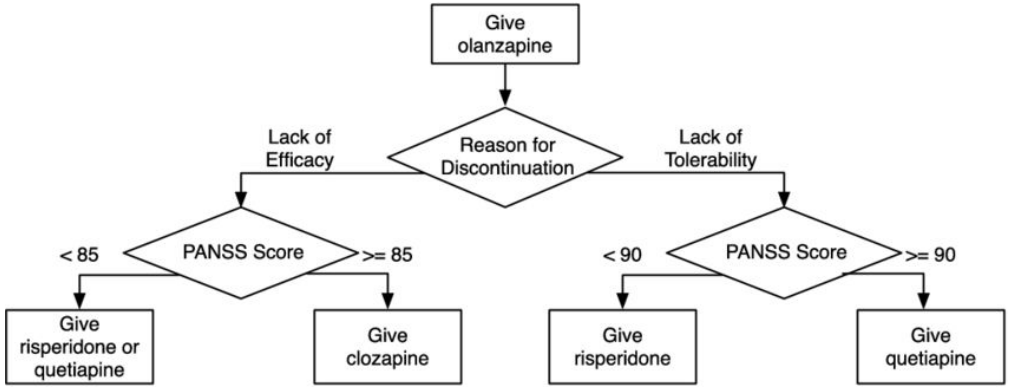
Adapted from Adams, et al. (2004)



Ernst, et al [2006] use FQI with tree-based function approximators (Ernst, et al [2005]) in a batch setting to develop treatment strategies for simulated HIV patients

Ernst, Damien, et al. "Clinical data based optimal STI strategies for HIV: a reinforcement learning approach." *Proceedings of the 45th IEEE Conference on Decision and Control*. IEEE, 2006.

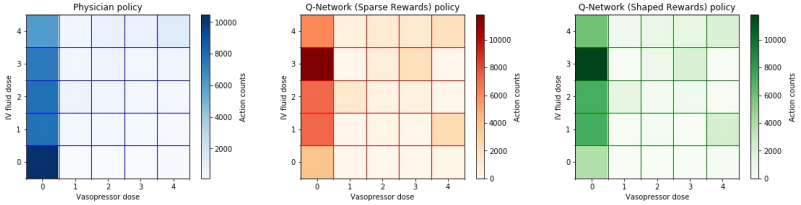
Informing Sequential Clinical Decision Making through Reinforcement Learning

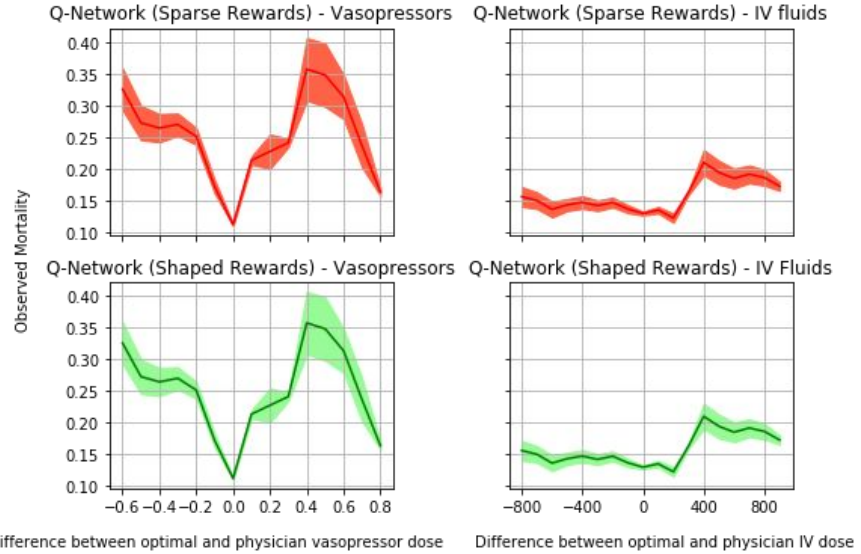


- **Shortreed, et al [2011] provide one of the first rigorous studies of how RL may be leveraged in a healthcare setting.**
- **They perform empirical investigations of how Q-learning may be used in observational settings, finally applying their insights to the treatment of Schizophrenia.**

Shortreed, Susan M., et al. "Informing sequential clinical decision-making through reinforcement learning: an empirical study." *Machine learning* 84.1-2 (2011): 109-136.

Continuous State-Space Models for Optimal Sepsis Treatment - a Deep Reinforcement Learning Approach

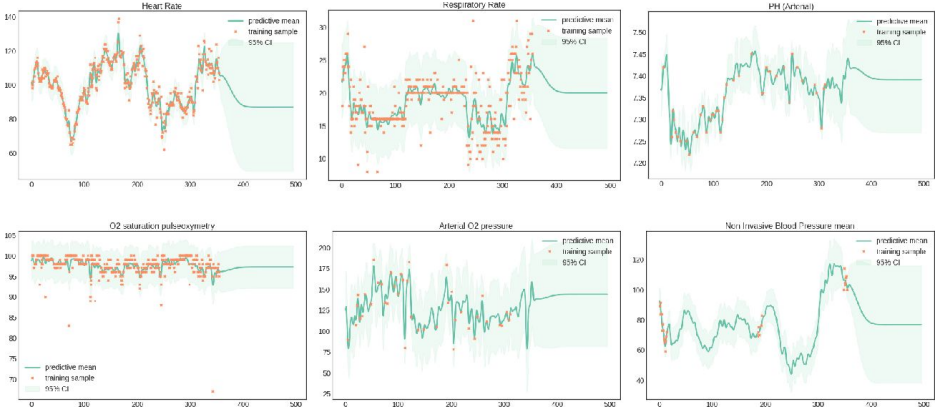


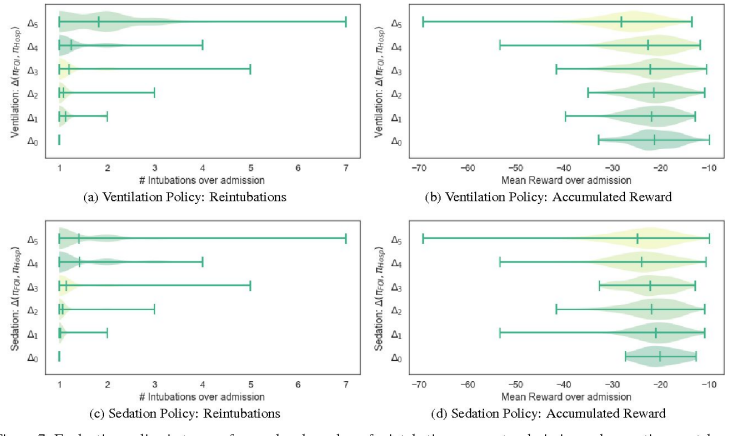


Raghu, et al [2017] implement Deep Q-learning methodologies to develop treatment strategies for septic patients in Intensive Care Units, comparing with how learned strategies differ from actual doctor decisions.

Raghu, Aniruddh, et al. "Continuous state-space models for optimal sepsis treatment-a deep reinforcement learning approach." *arXiv preprint arXiv:1705.08422* (2017).

A Reinforcement Learning Approach to Weaning of Mechanical Ventilation in Intensive Care Units





(a) Ventilation Policy: Reintubations
(b) Ventilation Policy: Accumulated Reward
(c) Sedation Policy: Reintubations
(d) Sedation Policy: Accumulated Reward

Prasad, et al [2017] utilize GP regression for missing data imputation and then leverage FQI and NNs to develop policies for extubation of patients who are on mechanical ventilation.

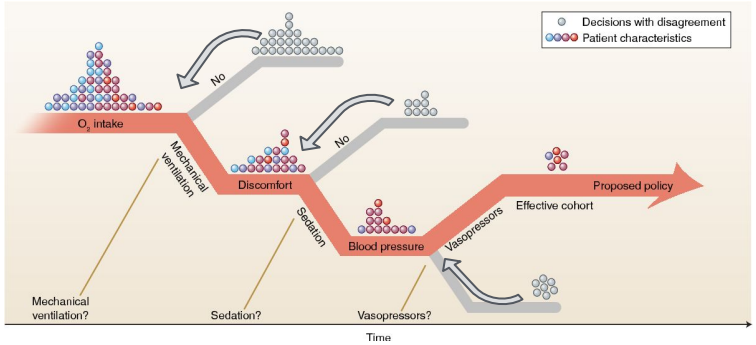
Prasad, Niranjani, et al. "A reinforcement learning approach to weaning of mechanical ventilation in intensive care units." *arXiv preprint arXiv:1704.06300* (2017).

Focusing on Value Functions

$$V^{\pi}(s_t) = \mathbb{E}\left[\sum_{i=t}^{T} \gamma^{i-1} r_i\right]$$

$$V^{\pi}(s_t) = \max_a \; Q^{\pi}(s_t, a_t)$$

$$Q^{\pi}(s_t, a_t) = R(s_t, a_t) + \gamma \max_a Q^{\pi}(s_{t+1}, a)$$



Traditionally, RL seeks to maximize the value function by approximating the effect of subsequent actions from the current observation (the Q-function) in developing a policy

The max operator excludes potentially valuable signal from the outcomes of actions that aren't locally optimal. Without exploration, the outcomes following these actions aren't incorporated into the value estimate for corresponding states.

Image Credits: *Gottesman, Omer, et al. "Guidelines for reinforcement learning in healthcare." Nat Med 25.1 (2019): 16-18.*

Summary

- The practice of medicine is inherently a sequential decision making problem.
- While there are some complications with utilizing RL in observational settings, there is great promise with the framework being able to better describe the sequential nature of the patient-clinician decision problem
    - We cannot blindly implement SOTA RL approaches and expect the same kind of "superhuman" results → We need to be vigilant and thoughtful about how we develop RL research in healthcare settings

Largely, open problems in RL map neatly into open problems in ML4H. There's a lot of exciting developments to come in the near future!

Taylor's Opinion