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

Niranjani Prasad*, Li-Fang Cheng*, Corey Chivers†, Michael Draugelis†, and Barbara E. Engelhardt*

- 13 Aug 2017 -

Motivation

- Management of routine ICU interventions constitute a major part of intensive care, e.g:

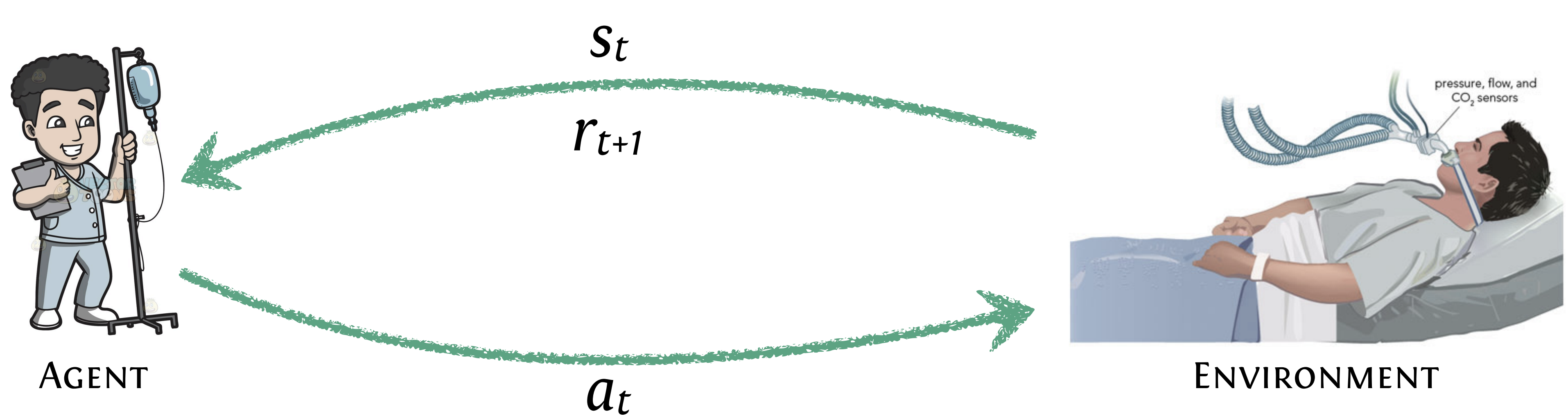
- **Invasive mechanical ventilation**: use of mechanical means to assist or replace spontaneous breathing.
    - 40% of ICU ventilated at any given hour – 12% of US hospital costs.
    - Typically coupled with sedation to maintain comfort and stability.
- Timely intervention can improve outcomes and reduce costs.
- But their effect is often poorly understood – particularly for heterogenous patient populations – so clinical opinion varies.

Ventilation

- **Weaning:** process of liberation from mechanical ventilation.
  - Premature, delayed weaning both associated with worse outcomes.

- We aim to develop a *clinician-in-loop* decision support tool to
  - alert caregivers when a patient is ready for weaning, and
  - recommend sedation and ventilation settings…

…by modeling this as a **Markov Decision Process** (MDP).

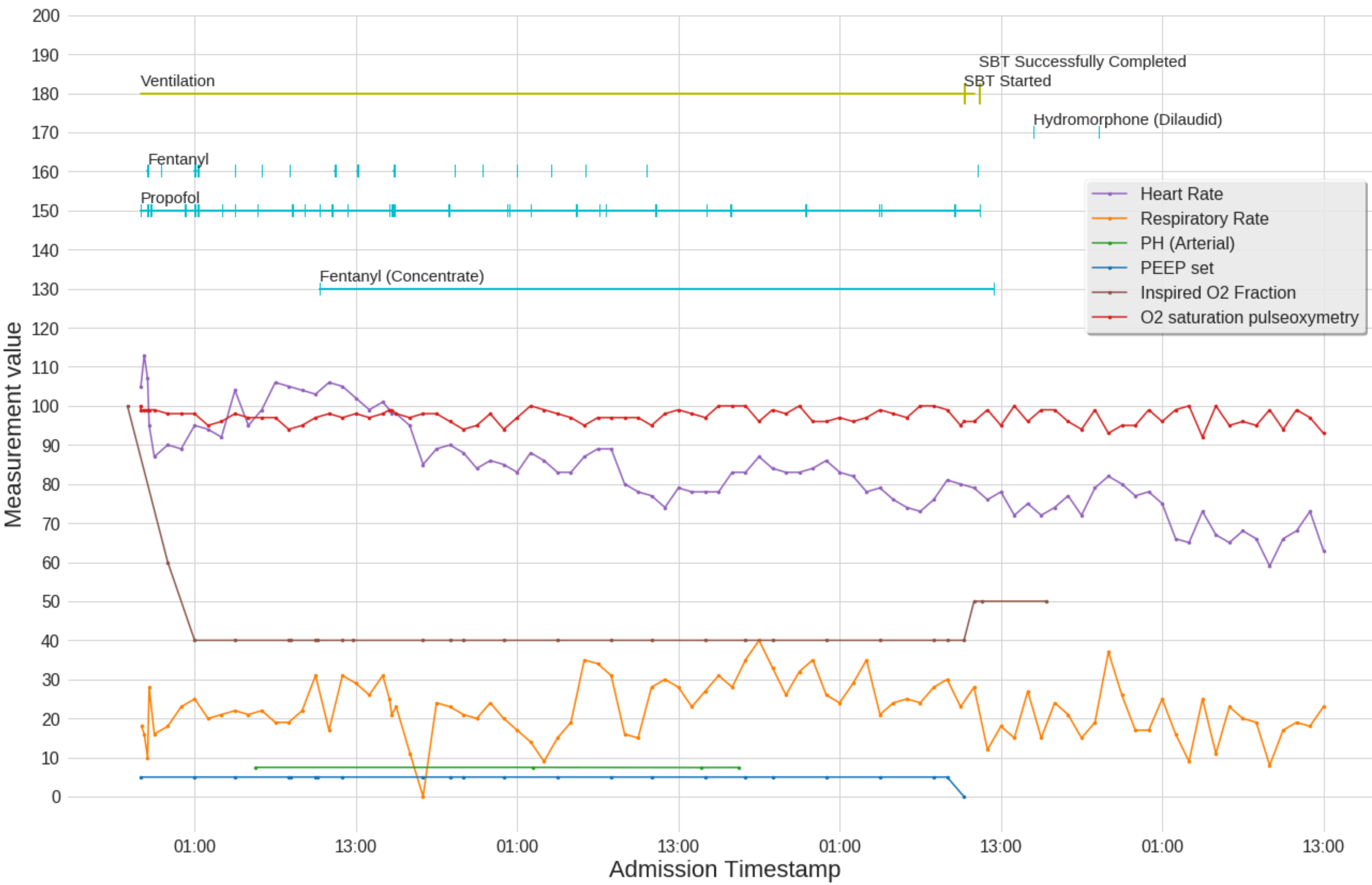


- Offline, off-policy RL to learn optimal policy given sub-optimal histories.

Why Reinforcement Learning?

- Fundamentally a **sequential decision making problem**:
    - Choose the best action at each point in a stochastic process,
    - Capture delayed effects of actions, and uncertainty in transitions and outcomes.
    - Handle data collected from biased policies.

MIMIC III Dataset



Preprocessing Data

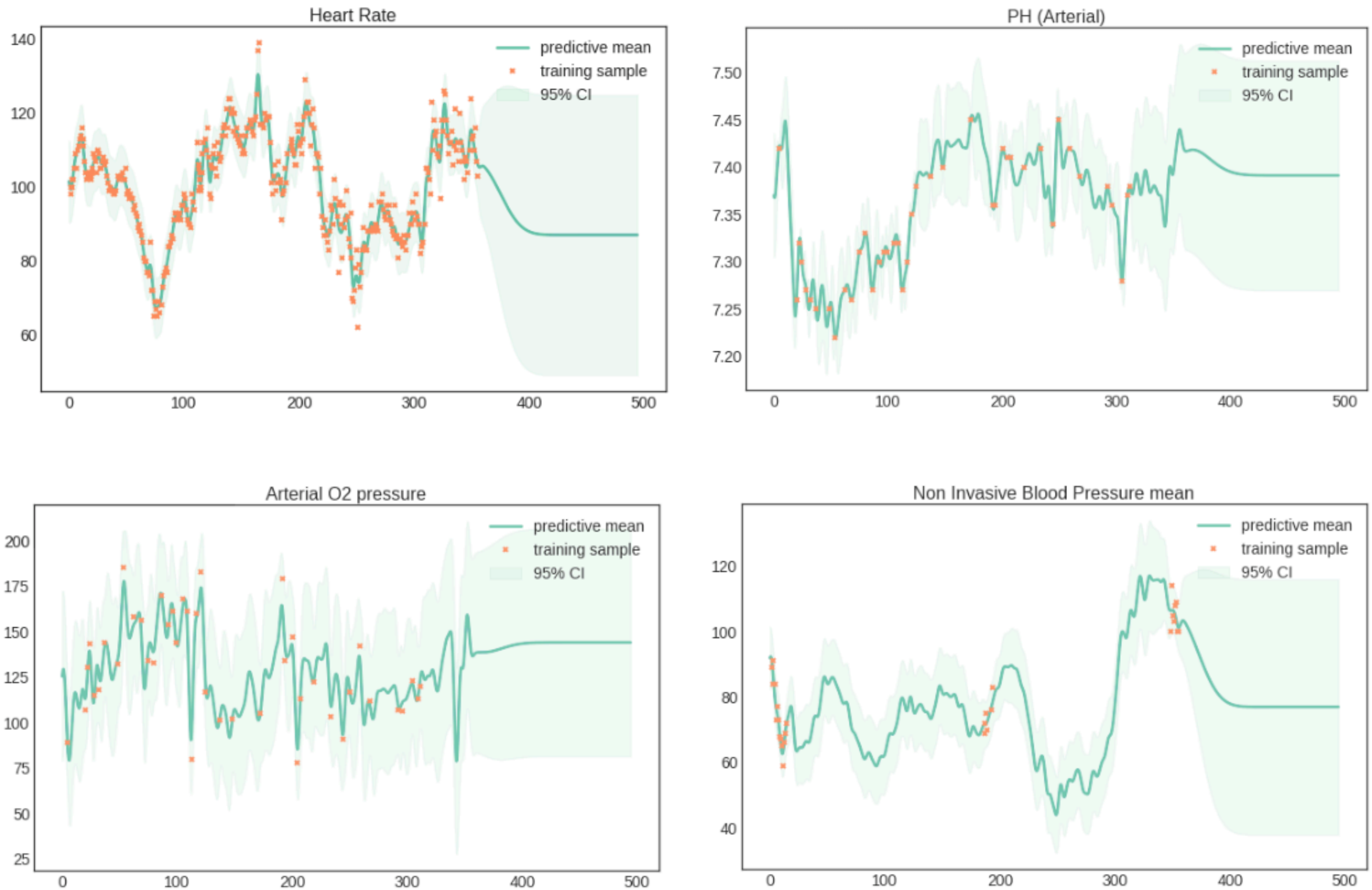
- Measurements tend to be sparse, irregularly sampled and error-prone
- We tackle this by using **multi-output Gaussian Processes** (GPs) to jointly model vitals by estimating covariance structures between them [Cheng'17].

$$\mathbf{v} = f(\mathbf{t}) + \boldsymbol{\varepsilon},$$

$$f(\mathbf{t}) \sim \mathcal{GP}(m(\mathbf{t}), \kappa(\mathbf{t}, \mathbf{t}'))$$

- We set $m(\mathbf{t}) = 0$ and $k(\mathbf{t}, \mathbf{t}')$ as kernel in linear coregionalization model with the spectral kernel as the basis function.

Preprocessing Data

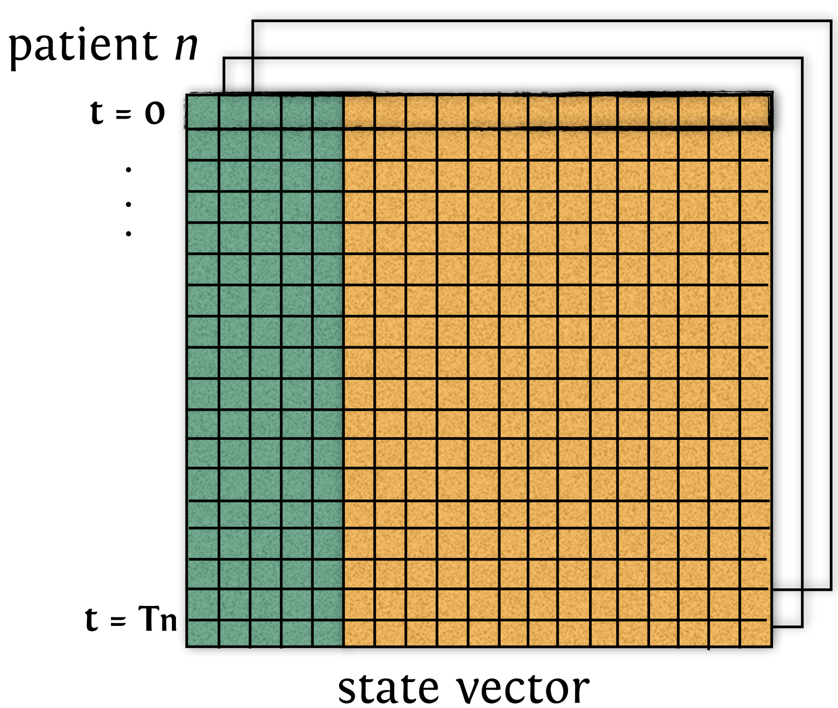


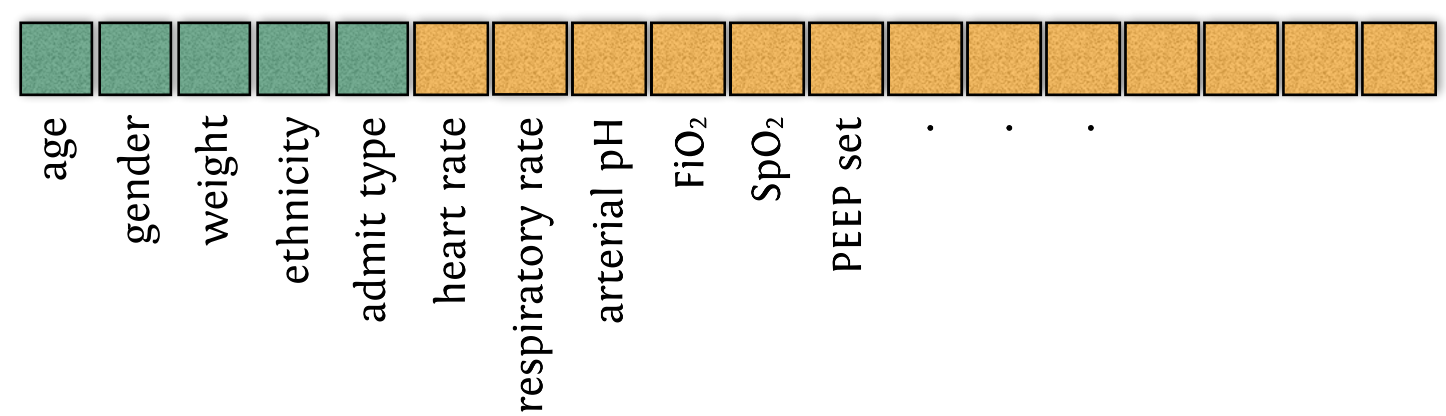
MDP Formulation

- Given histories $\langle s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3, s_3 \ldots \rangle^n$, we wish to solve for the following **objective function**:

$$\max_{\pi(s):\mathcal{S}\rightarrow\mathcal{A}} \sum_{n=1}^{N} R(s_t^n, \pi), \text{ where } R(s,\pi) = \lim_{T\rightarrow\infty} \sum_{t=0}^{T} \gamma^t r(s_t, \pi(s_t))$$

- **State variable** $s_t \in S$: 32-dim feature vector comprising demographic data, time-varying vitals, sedatives, vent duration, reintubation number.





MDP Formulation

- Given histories $\langle s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3, s_3 \ldots \rangle^n$, we wish to solve for the following **objective function**:

$$\max_{\pi(s):\mathcal{S}\to\mathcal{A}} \sum_{n=1}^{N} R(s_t^n, \pi), \text{ where } R(s,\pi) = \lim_{T\to\infty} \sum_{t=0}^{T} \gamma^t r(s_t, \pi(s_t))$$

- **State variable** $s_t \in S$: *32*-dim feature vector comprising demographic data, time-varying vitals, sedatives, vent duration, reintubation number.

- Action or **decision variable** at each time step is chosen from a discrete action space of vent settings and dosage levels.

$$\mathcal{A} = \left\{ \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}0\\2\end{bmatrix}, \begin{bmatrix}0\\3\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}1\\3\end{bmatrix} \right\}$$

All administered sedatives are mapped to single discretized scale.

MDP Formulation

- Given histories $\langle s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3, s_3 \ldots \rangle^n$, we wish to solve for the following **objective function**:

$$\max_{\pi(s):\mathcal{S}\to\mathcal{A}} \sum_{n=1}^{N} R(s_t^n, \pi), \text{ where } R(s, \pi) = \lim_{T\to\infty} \sum_{t=0}^{T} \gamma^t r(s_t, \pi(s_t))$$

- **State variable** $s_t \in S$: *32*-dim feature vector comprising demographic data, time-varying vitals, sedatives, vent duration, reintubation number.

- Action or **decision variable** at each time step is chosen from a discrete action space of vent settings and dosage levels.

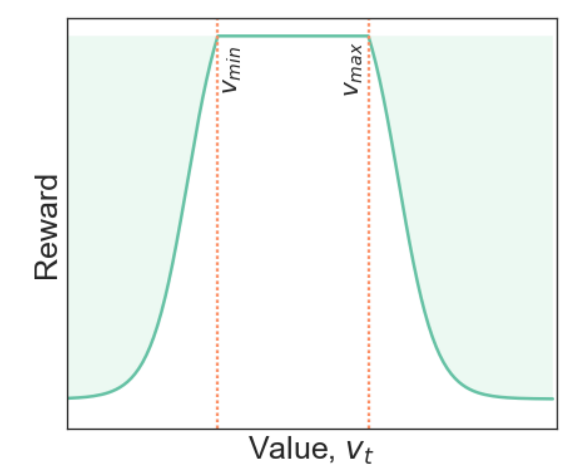
- Exogenous information comes in the form of the **reward function**.

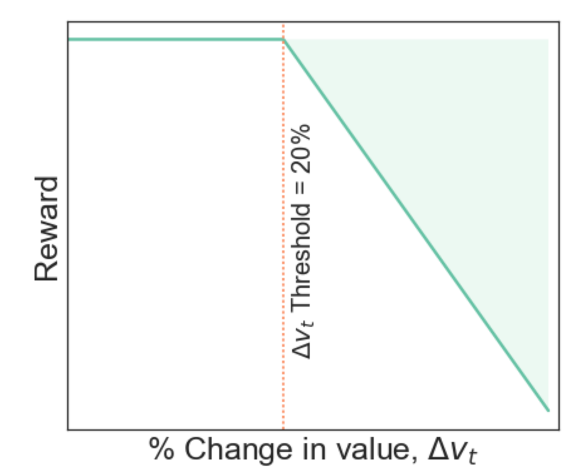
Defining the Reward function

- Wean guidelines from Hospital of the University of Pennsylvania:

Physiological Stability	**Oxygenation Criteria**
Respiratory Rate ≤ 30	PEEP (cm H_2O) ≤ 8
Heart Rate ≤ 130	SpO_2 (%) ≥ 88
Arterial pH ≥ 7.3	Inspired O_2 (%) ≤ 50

- We want to penalize:
  - prolonged ventilation
  - vitals exceeding desired ranges
  - sharp changes in vitals
  - failed spontaneous breathing trials
  - reintubation within ICU admission





Fitted Q-iteration

- Approximation of the Q-function all over the state-action space must be determined from finite, sparse sets of transitions.

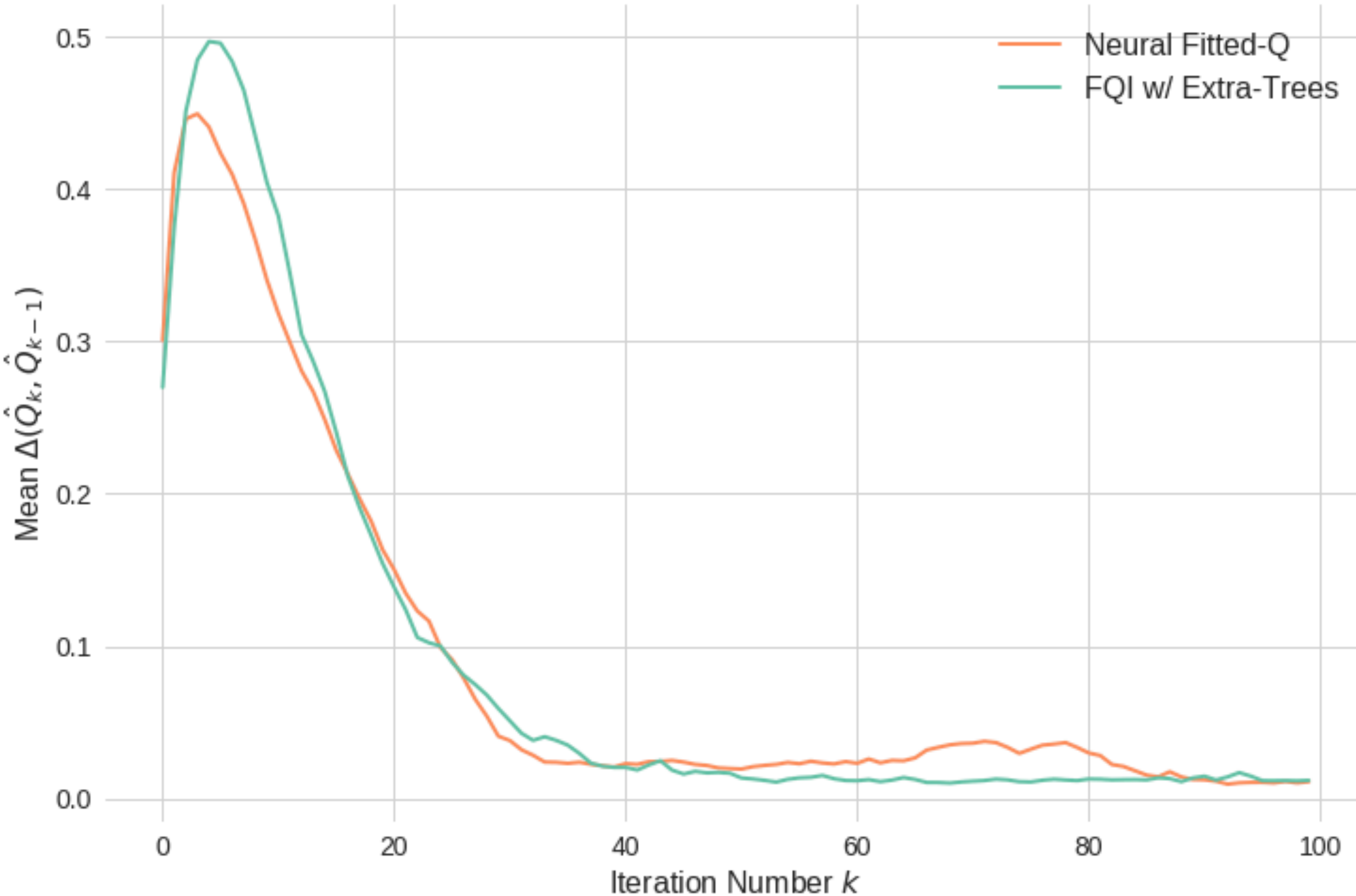
- **Fitted Q Iteration** (FQI) is a form of off-policy batch-mode RL that uses one-step transitions $\mathcal{F} = \{(s_t^n, a_t^n, s_{t+1}^n)\}_{n=1:|\mathcal{F}|}$ to learn a sequence $\hat{Q}_1, \hat{Q}_2 ... \hat{Q}_K$, by solving a series of *K* supervised learning problems.

- The training set for the *k*th problem is defined by:

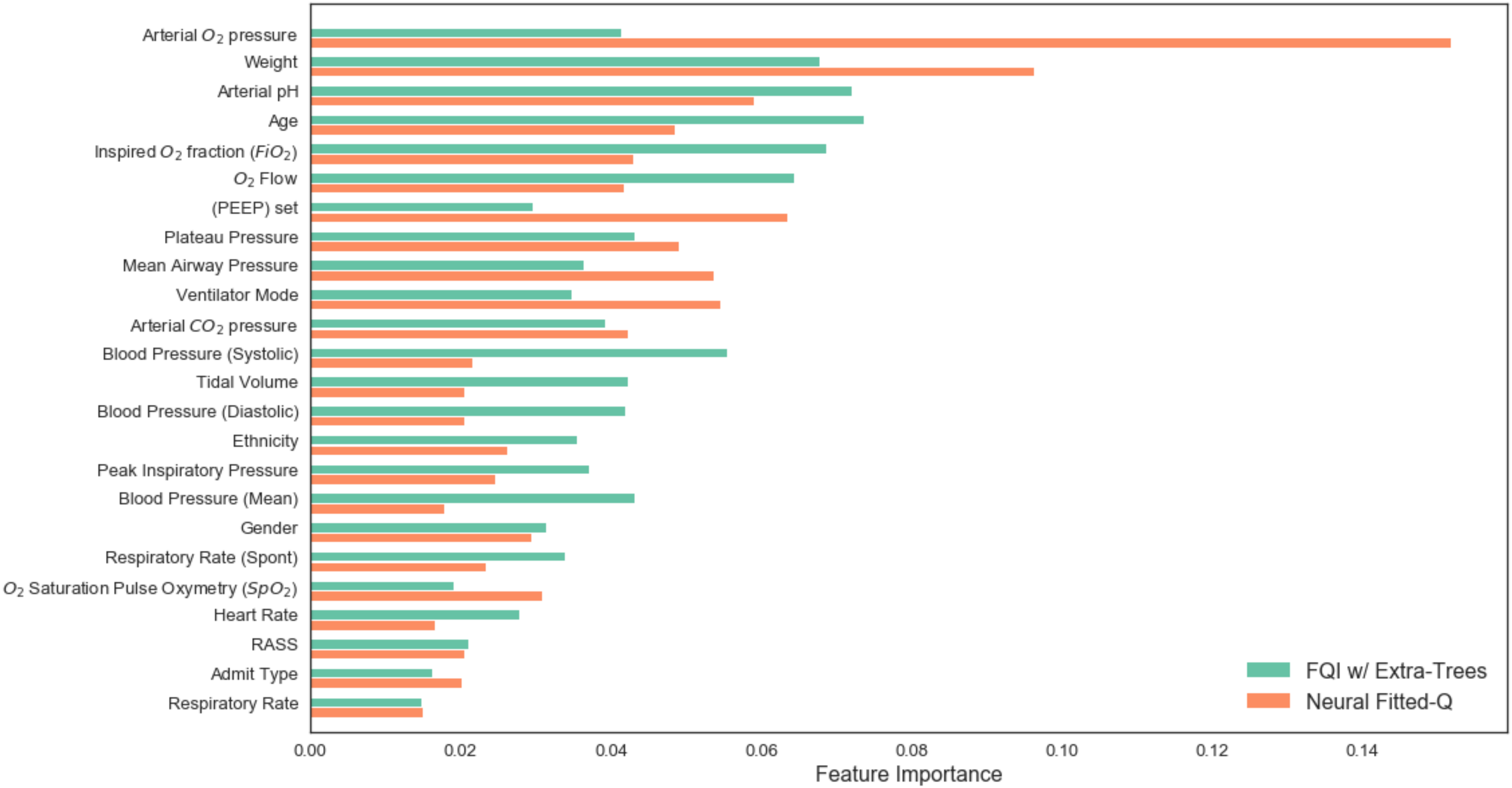
$$\{(s_t^n, a_t^n), r(s_t^n, a_t^n) + \gamma \max_{a \in A} \hat{Q}_{k-1}(s_{t+1}^n, a)\}_{n=1:|\mathcal{F}|}$$

- Can take advantage of generalization capabilities of any regression method:
    - Extremely Randomized Trees [Ernst'05]
    - Feedforward Neural Networks [Riedmiller'05]

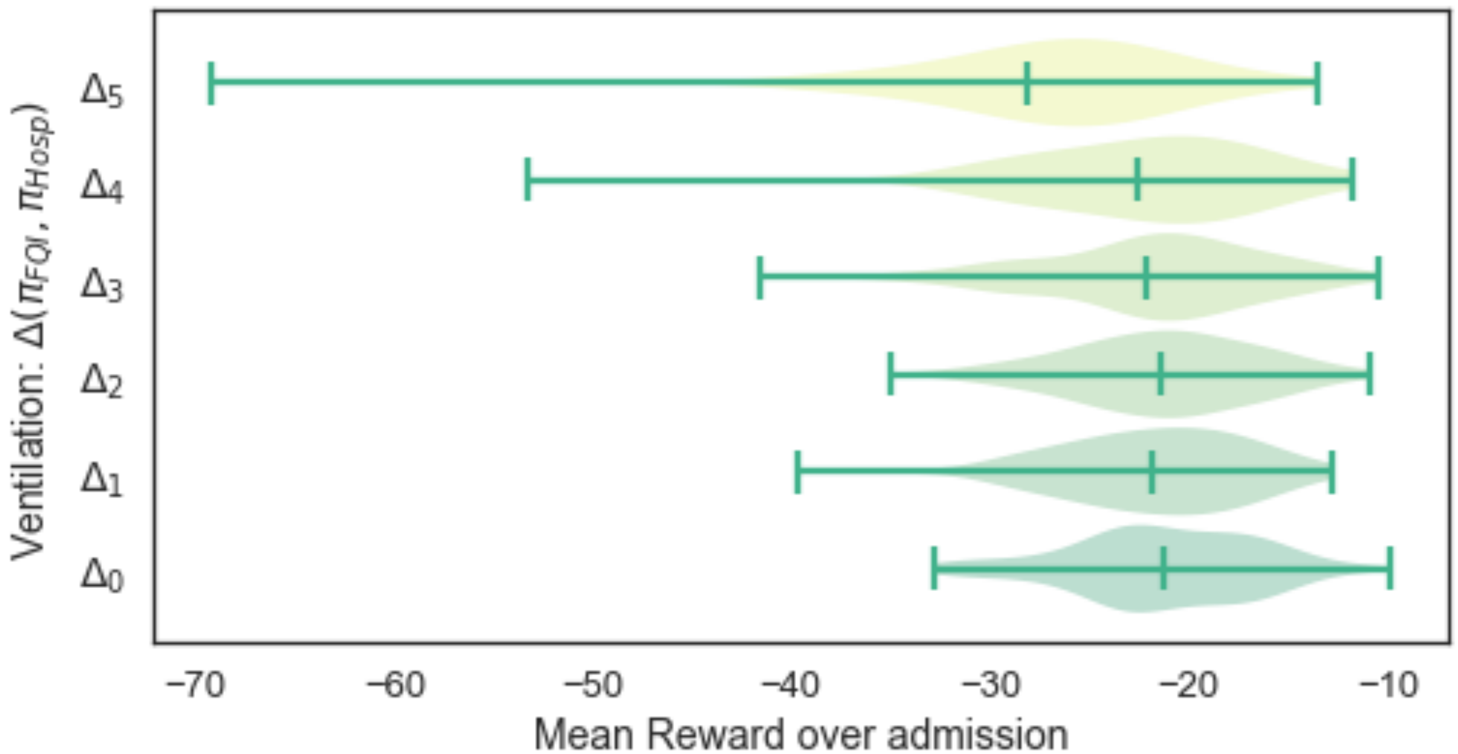
Convergence of Fitted-Q iteration

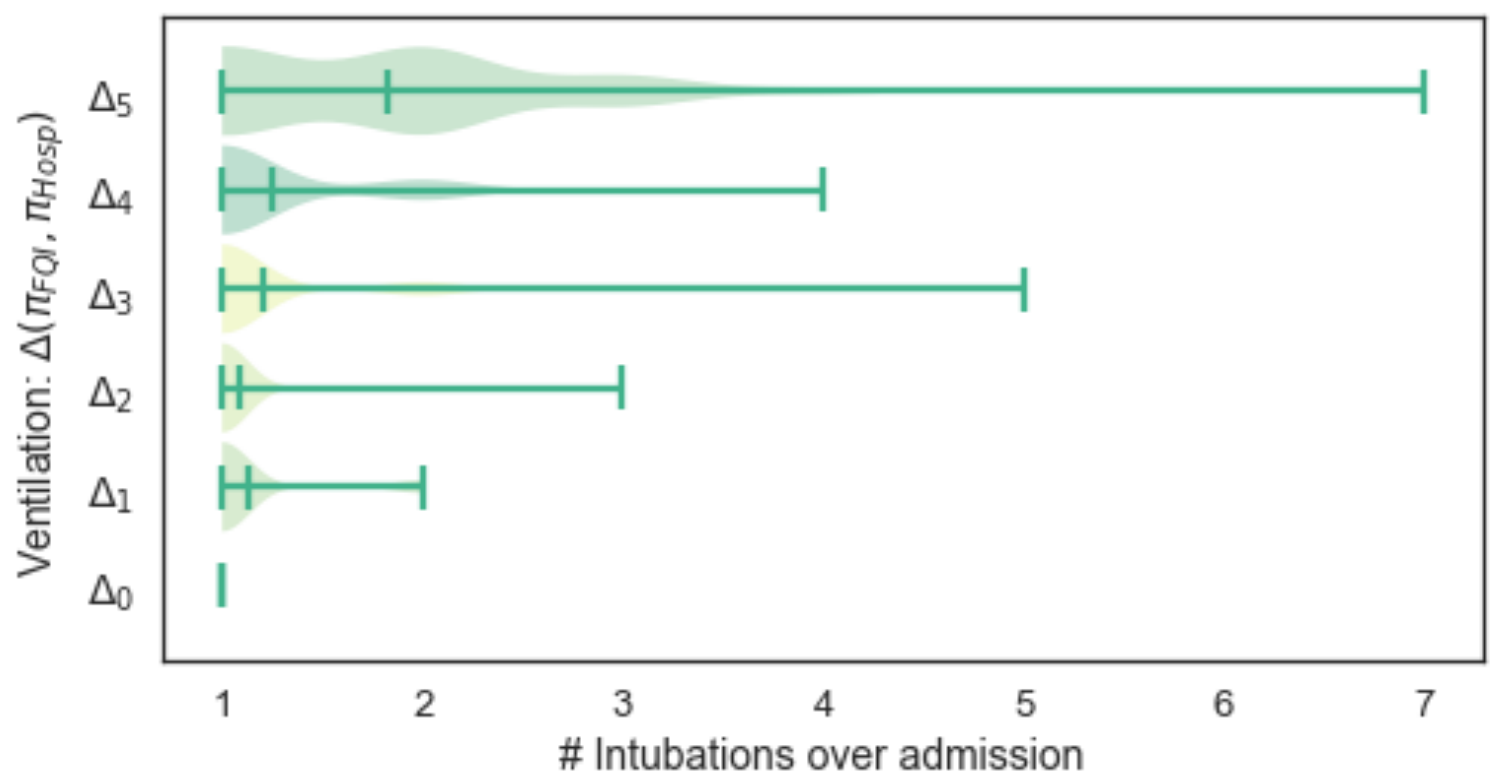


Policy Estimation



Evaluating performance of Policies





Conclusions

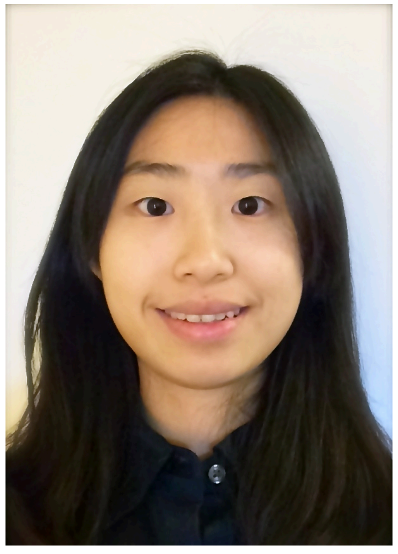
- Proposes a data-driven approach to the weaning from ventilation in the ICU.
- Patient admissions are modeled as MDPs, with clinically driven definitions of state, action, and reward.
- Reinforcement learning with FQI is then used to learn a simple weaning policy from examples in historical data.
- Capable of extracting meaningful indicators for patient readiness.
- Recommendations appear to outperform clinical practice on average, in terms of regulation of vitals and reintubations.

Future Directions

- Controlling Policy sensitivity to reward shaping
  - Inferring clinician's priorities using inverse RL [Abbeel'04]
  - Optimization over multiple objectives [Lizotte'12]
- Accounting for Partial Observability
- Quantifying uncertainty
  - Probabilistic estimates of Q using GP regression [Chowdhary'14]
- Controlling for Intervention bias from sub-optimal histories

THANK YOU!

Acknowledgements:



Li-Fang Cheng
Princeton University

Corey Chivers
Penn Medicine

Michael Draugelis
Penn Medicine

Barbara E. Engelhardt
Princeton University

Come by our poster!

QUESTIONS?