HARVARD
School of Engineering
and Applied Sciences

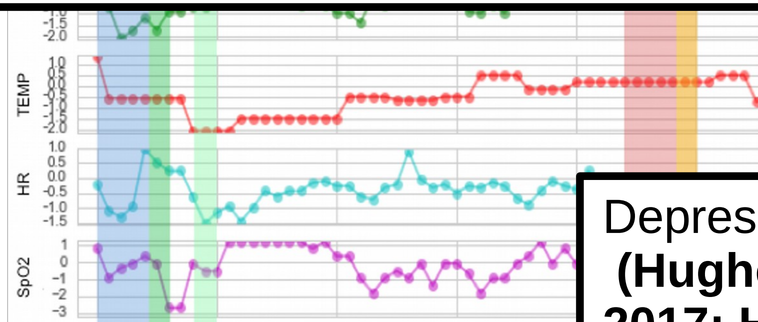
Towards Improving Health Decisions with Reinforcement Learning

Finale Doshi-Velez

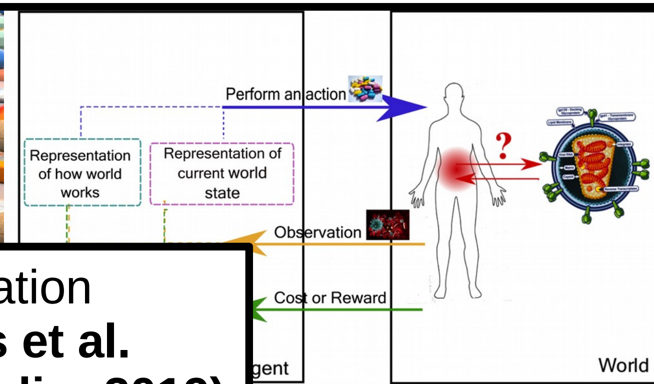
Collaborators: Sonali Parbhoo, Maurizio Zazzi, Volker Roth, Xuefeng Peng, David Wihl, Yi Ding, Omer Gottesman, Liwei Lehman, Matthieu Komorowski, Aldo Faisal, David Sontag, Fredrik Johansson, Leo Celi, Aniruddh Raghu, Yao Liu, Emma Brunskill, and the CS282 2017 Course

Our Lab: ML Towards Effective, Interpretable Health Interventions

Predicting and Optimizing Interventions in ICU (**Wu et al. 2015; Ghassemi et al. 2017; Peng 2018; Raghu 2018; Gottesman 2018; Gottesman 2019**)



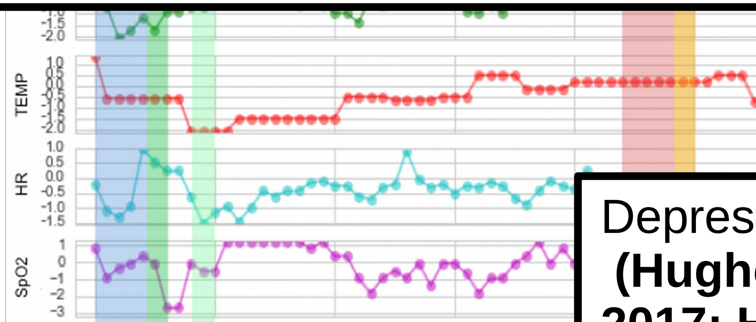
HIV Management Optimization **(Parbhoo et al., 2017, Parbhoo et al. 2018)**



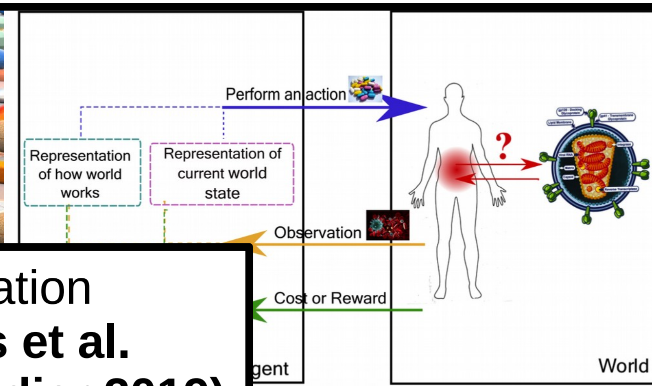
Depression Treatment Optimization **(Hughes et al., 2016; Hughes et al. 2017; Hughes et al. 2018; Pradier 2019)**

Our Lab: ML Towards Effective, Interpretable Health Interventions

Predicting and Optimizing Interventions in ICU (**Wu et al. 2015; Ghassemi et al. 2017; Peng 2018; Raghu 2018; Gottesman 2018; Gottesman 2019**)



HIV Management Optimization **(Parbhoo et al., 2017, Parbhoo et al. 2018)**

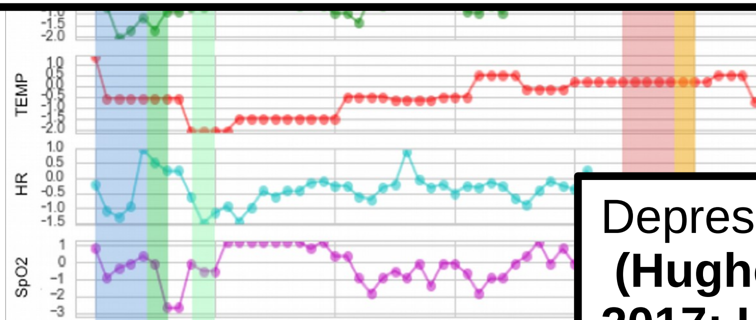


Depression Treatment Optimization **(Hughes et al., 2016; Hughes et al. 2017; Hughes et al. 2018; Pradier 2019)**

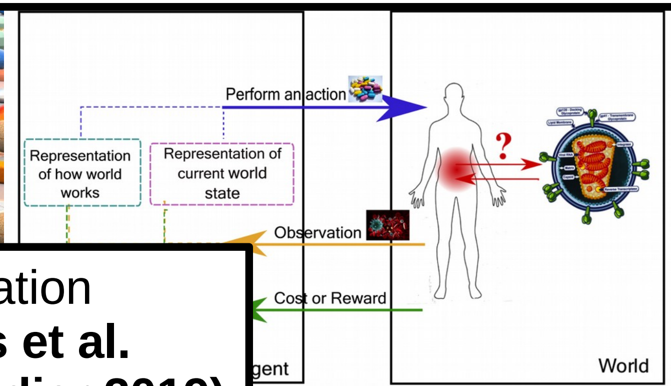
Today: How can reinforcement learning help solve problems in healthcare?

Our Lab: ML Towards Effective, Interpretable Health Interventions

Predicting and Optimizing Interventions in ICU (**Wu et al. 2015; Ghassemi et al. 2017; Peng 2018; Raghu 2018; Gottesman 2018; Gottesman 2019**)



HIV Management Optimization **(Parbhoo et al., 2017, Parbhoo et al. 2018)**



Depression Treatment Optimization **(Hughes et al., 2016; Hughes et al. 2017; Hughes et al. 2018; Pradier 2019)**

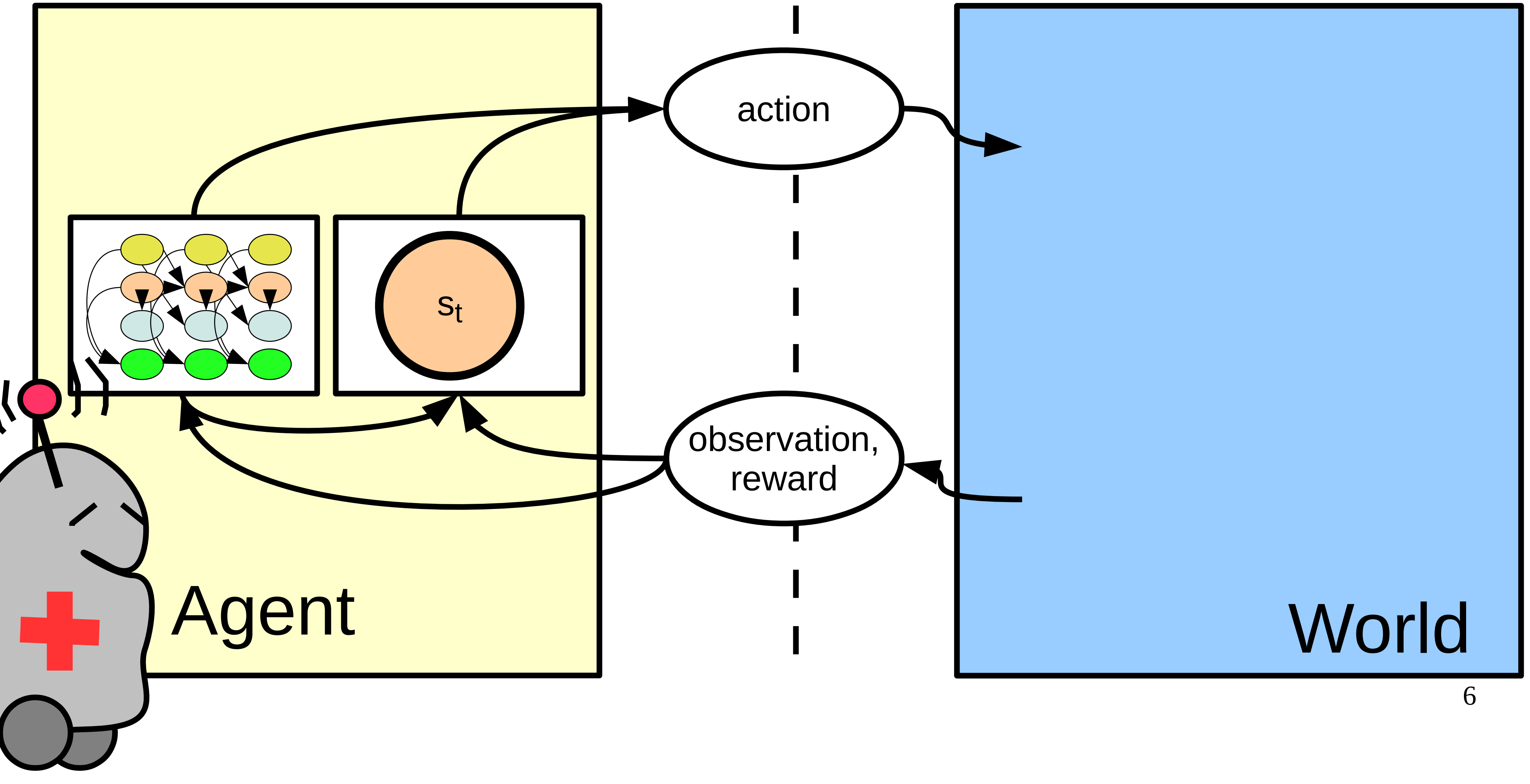
Focus: Situations that require a **sequence of decisions**

Challenges in the Health Space

- The data are typically available only in batch
  - No control over the clinician policy!
- The data give very partial views of the process
  - Measurements, confounds missing
  - Intents missing
- Success is not always easy to quantify

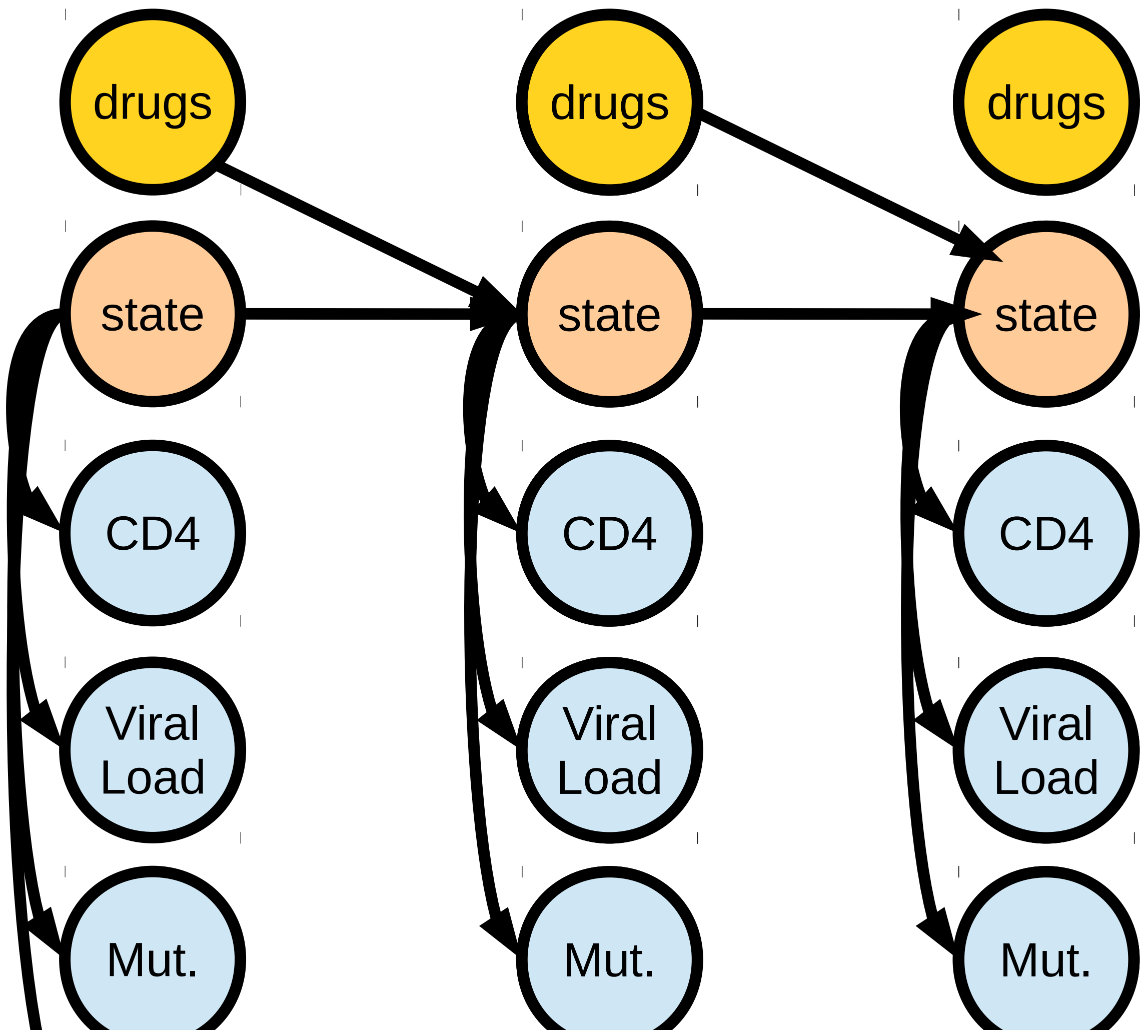
BUT: We still want to extract as much from these data as we can!

Problem Set-Up



Solutions: Train Model/Value Function

Solves the long-term problem (e.g. Ernst 2005; Parbhoo 2014; Marivate 2015), often in simulation/simplified settings.



Rewards:

If $V_t > 40$:

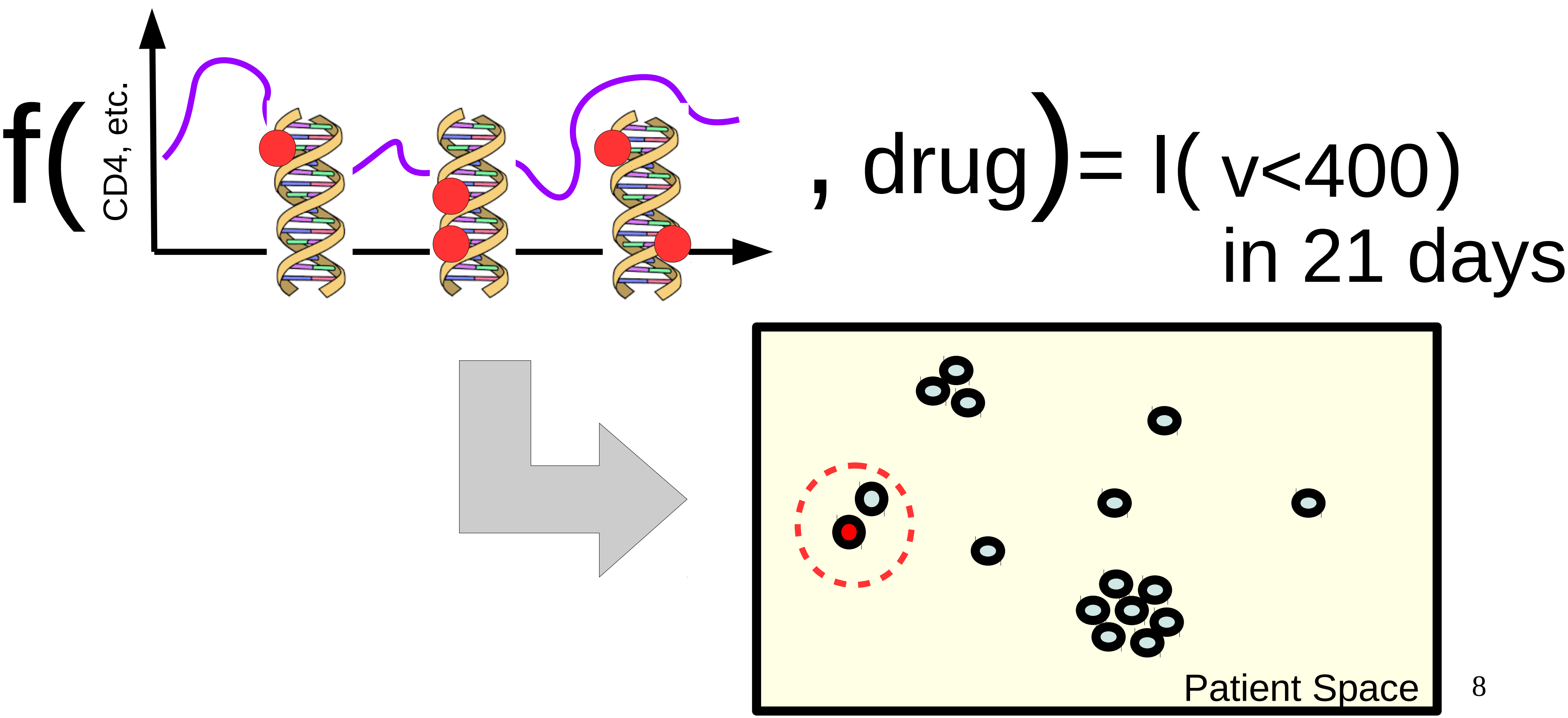
$r_t = -.7 \log V_t + .6 \log T_t - .2|M_t|$

Else:

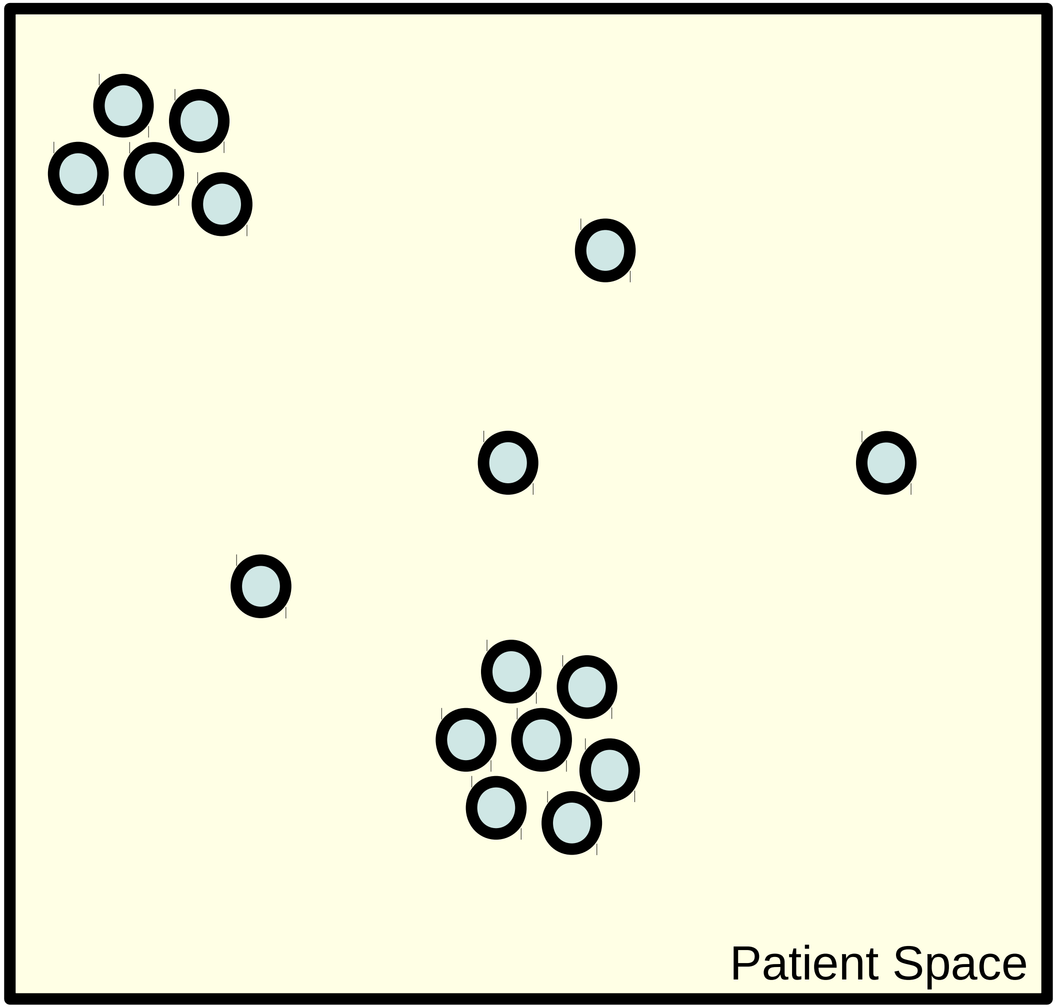
$r_t = 5 + .6 \log T_t - .2|M_t|$

Solutions: Nonparametric

Use the full patient history to predict immediate outcomes (e.g. Bogojeska 2012), but often ignore long term effects.

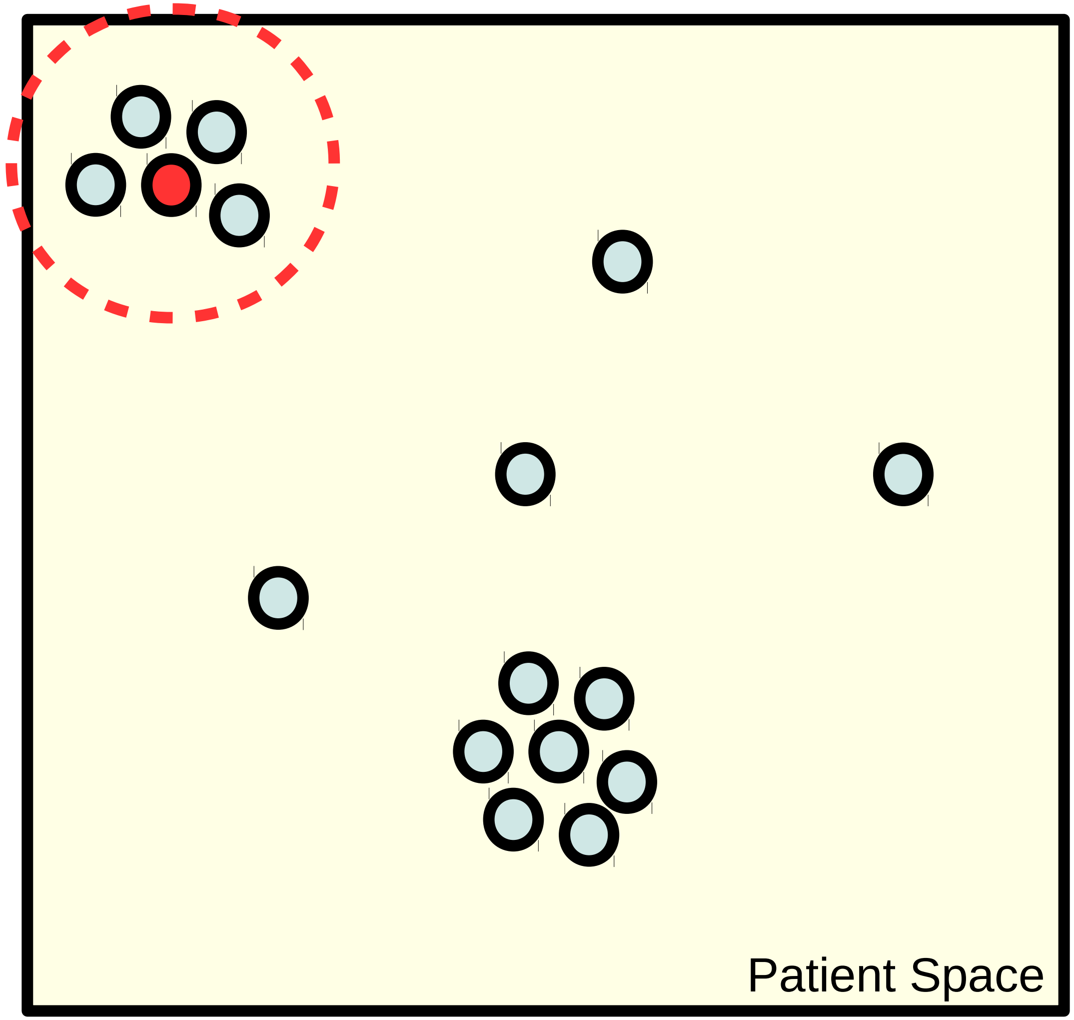


Our insight: These approaches have complementary strengths!

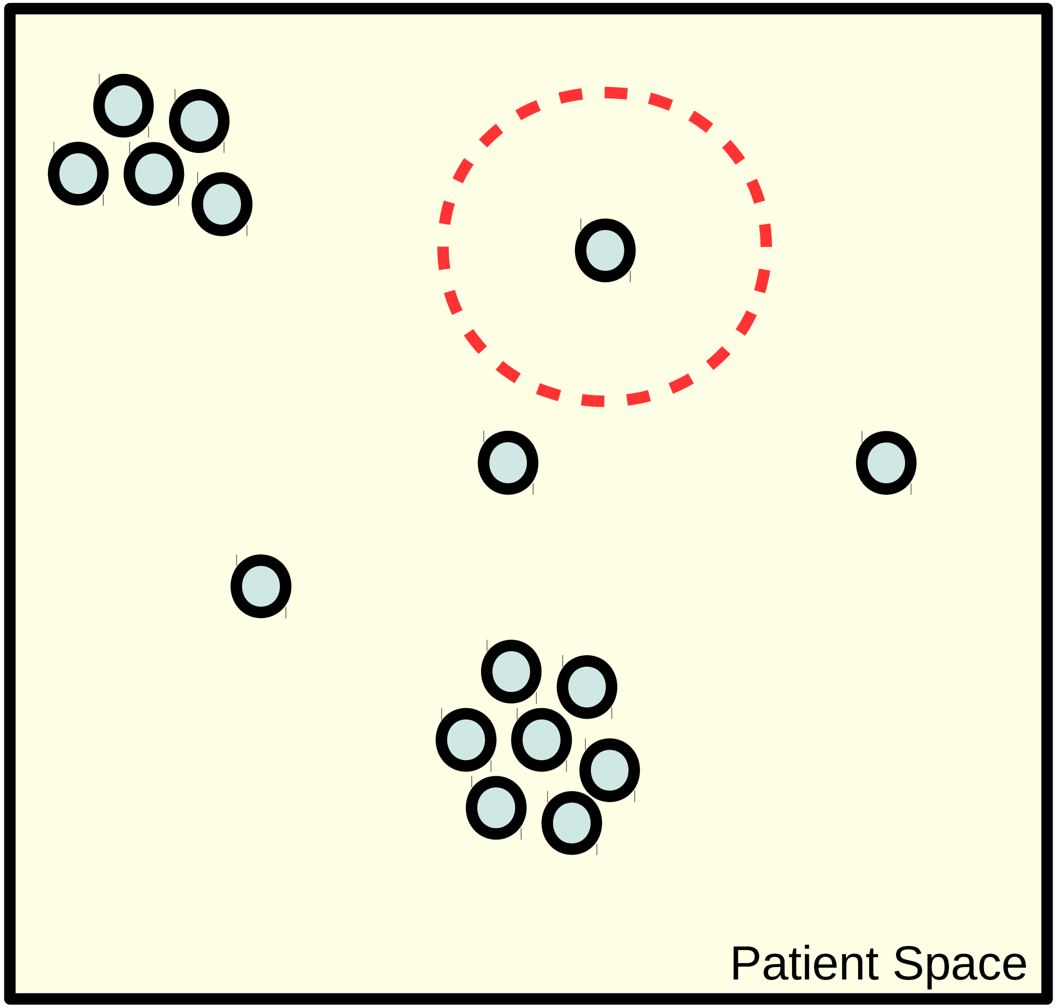


Our insight: These approaches have complementary strengths!

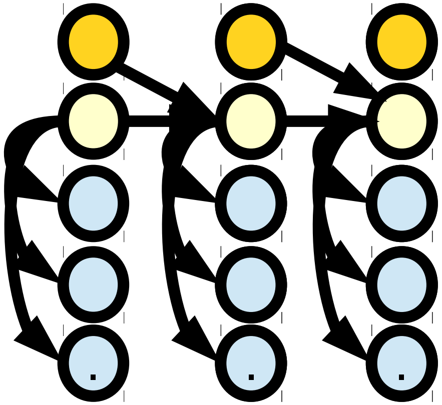
Patients in clusters may be best modeled by their neighbors



Our insight: These approaches have complementary strengths!



Patients without neighbors may be better modeled with a parametric model



New Solution: Ensemble the Predictors

Kernel Action

h(POMDP Action) = Actual Action

Patient Statistics

Application to HIV Management

- 32,960 patients from EU Resist Database; hold out 3,000 for testing.
- Observations: CD4s, viral loads, mutations
- Actions: 312 common drug combinations (from 20 drugs)

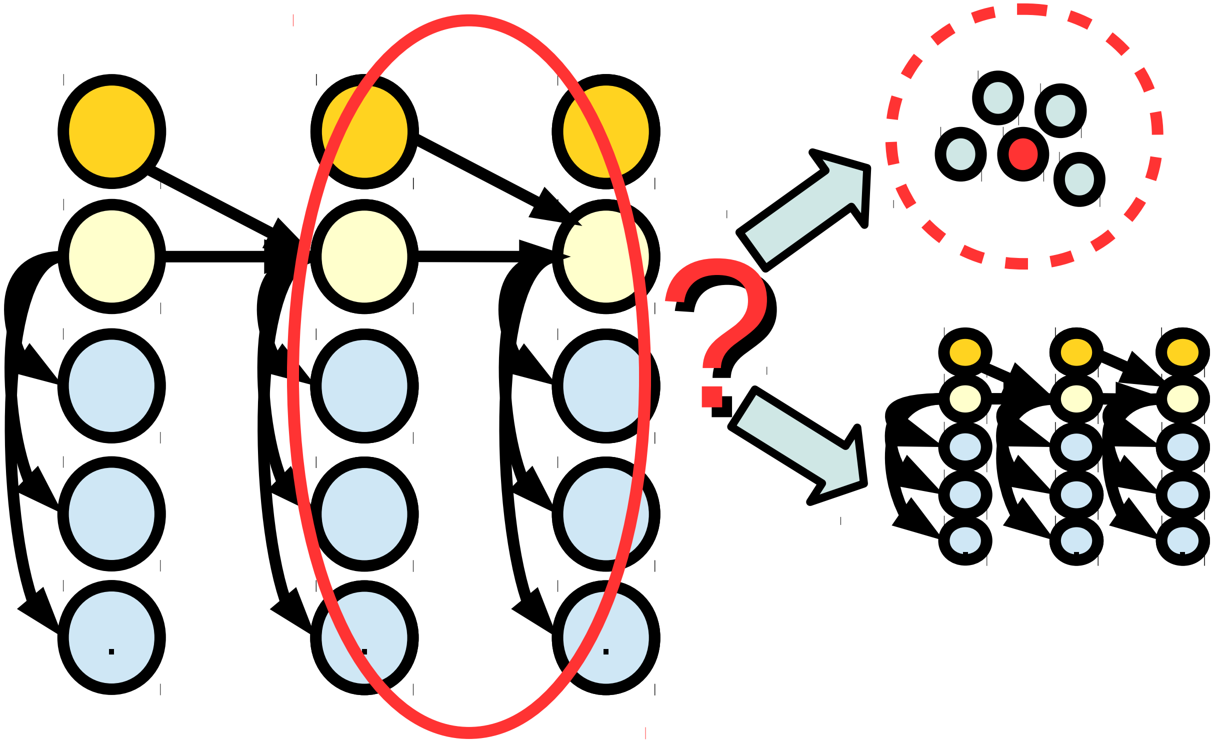
Approach	DR Reward
Random Policy	-7.31 ± 3.72
Neighbor Policy	9.35 ± 2.61
Model-Based Policy	3.37 ± 2.15
Policy-Mixture Policy	11.52 ± 1.31
Model-Mixture Policy	12.47 ± 1.38

Parbhoo et al., AMIA 2017; Parbhoo et al., PLoS Medicine 2018

Application to HIV Management

- 32,960 pa
  Resist Da
  out 3,000
- Observat
  viral load
- Actions: 3
  drug com
  20 drugs)

Extension: Putting the mixing in the model.



ward
± 3.72
2.61
2.15
± 1.31
± 1.38

Parbhoo et al., AMIA 2017; Parbhoo et al., PLoS Medicine 2018

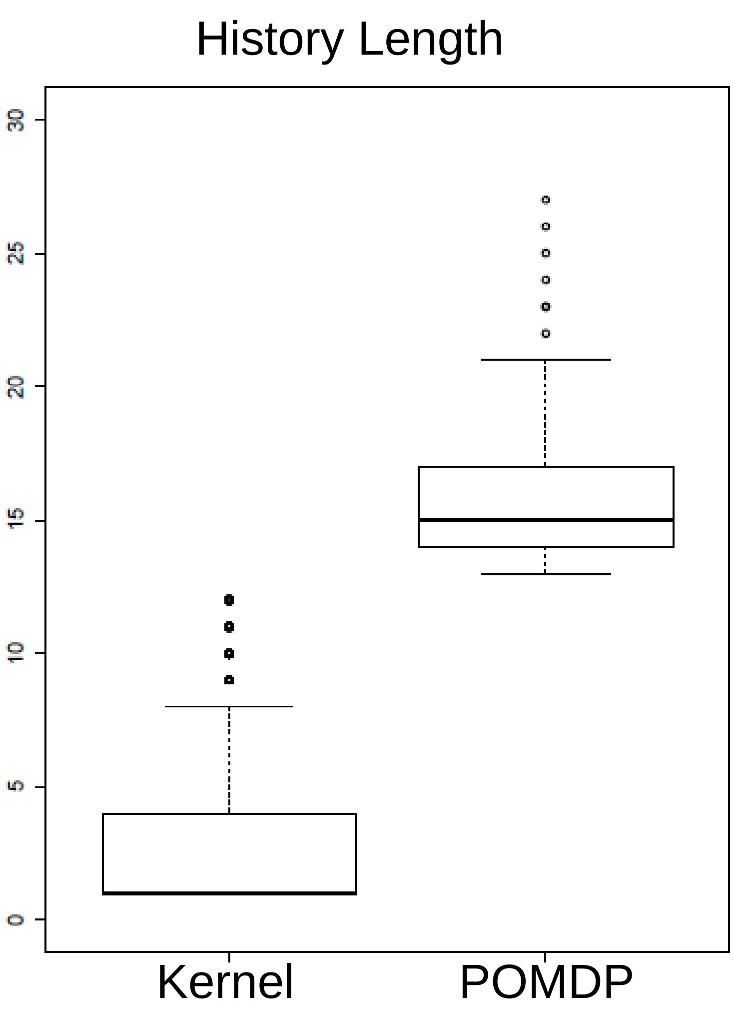
Application to HIV Management

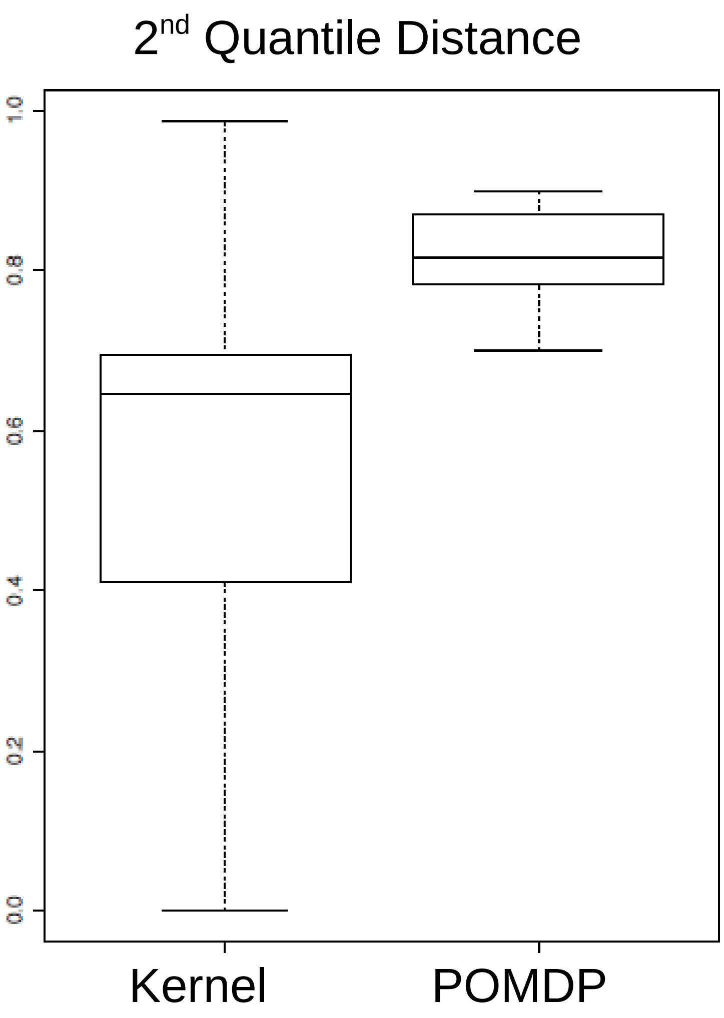
- 32,960 patients from EU Resist Database; hold out 3,000 for testing.
- Observations: CD4s, viral loads, mutations
- Actions: 312 common drug combinations (from 20 drugs)

Approach	DR Reward
Random Policy	-7.31 ± 3.72
Neighbor Policy	9.35 ± 2.61
Model-Based Policy	3.37 ± 2.15
Policy-Mixture Policy	11.52 ± 1.31
Model-Mixture Policy	12.47 ± 1.38

Parbhoo et al., AMIA 2017; Parbhoo et al., PLoS Medicine 2018

And: Our hypothesis was correct! Model used when neighbors are far





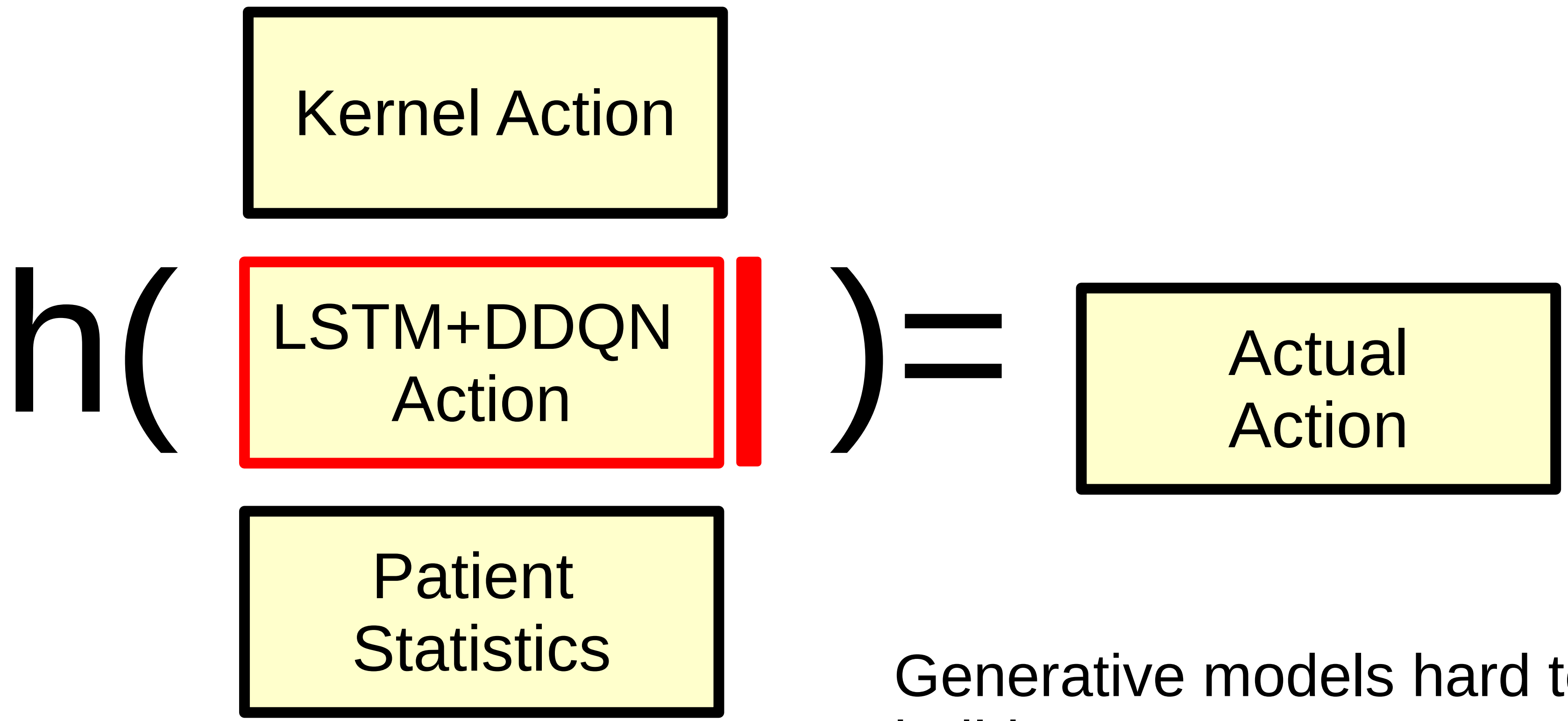
Application to Sepsis Management

- Cohort of 15,415 patients with sepsis from the MIMIC dataset (same as Raghu et al. 2017); contains vitals and some lab tests.
- Actions: focus on vasopressors and fluids, used to manage circulation.
- Goal: reduce 30-day mortality; rewards based on probability of 30-day mortality:

$$r(o,a,o') = -\log \frac{f(o')}{1-f(o')} f(o') + \log \frac{f(o)}{1-f(o)}$$

Peng et al., AMIA 2018

Minor Adjustment: Values, not Models

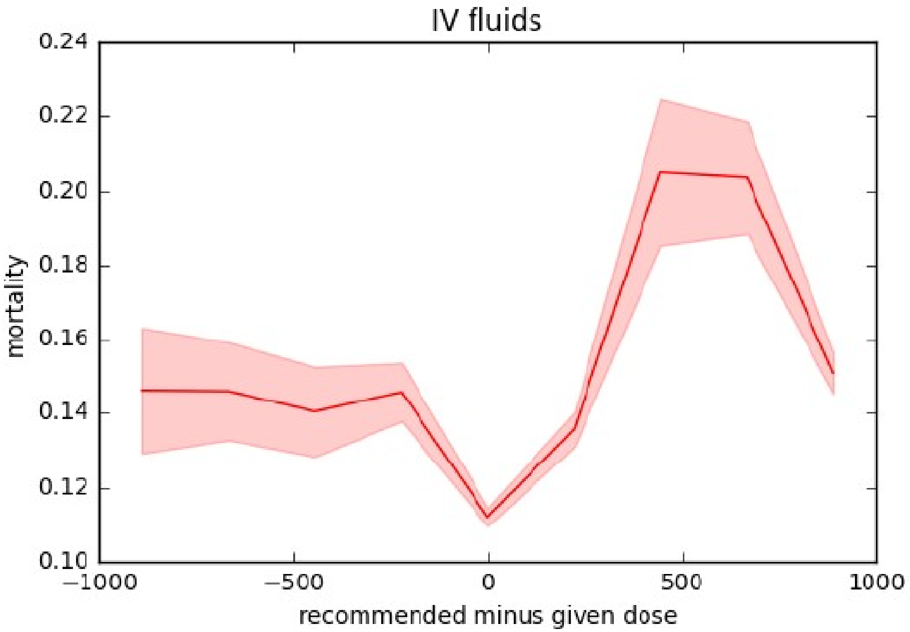


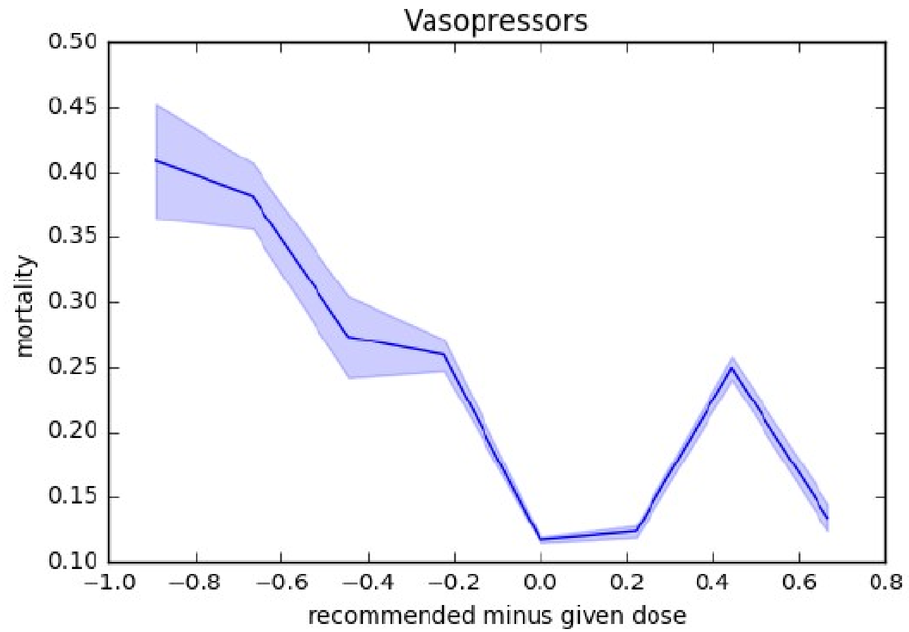
Generative models hard to build → LSTM+DDQN

LSTM+DDQN suggests never-taken actions → hard cap.

Application to Sepsis Management

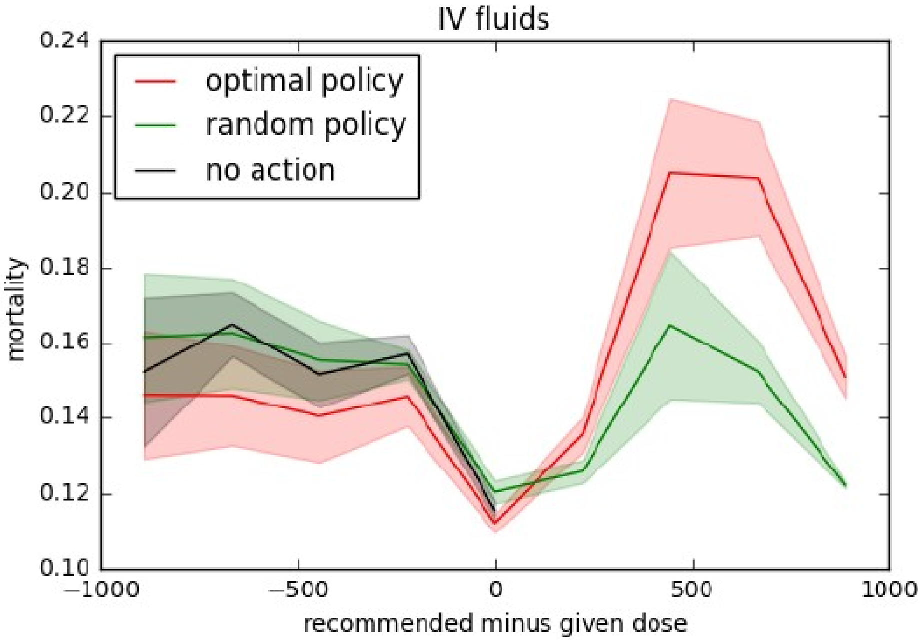
	Physician	Kernel	DQN	MoE_{V_d,Q_d}	MoE_{V_b,Q_b}
non-recurrent encoded	3.76	3.73	4.06	3.93	4.31
recurrent encoded	3.76	4.46	4.23	5.03	5.72

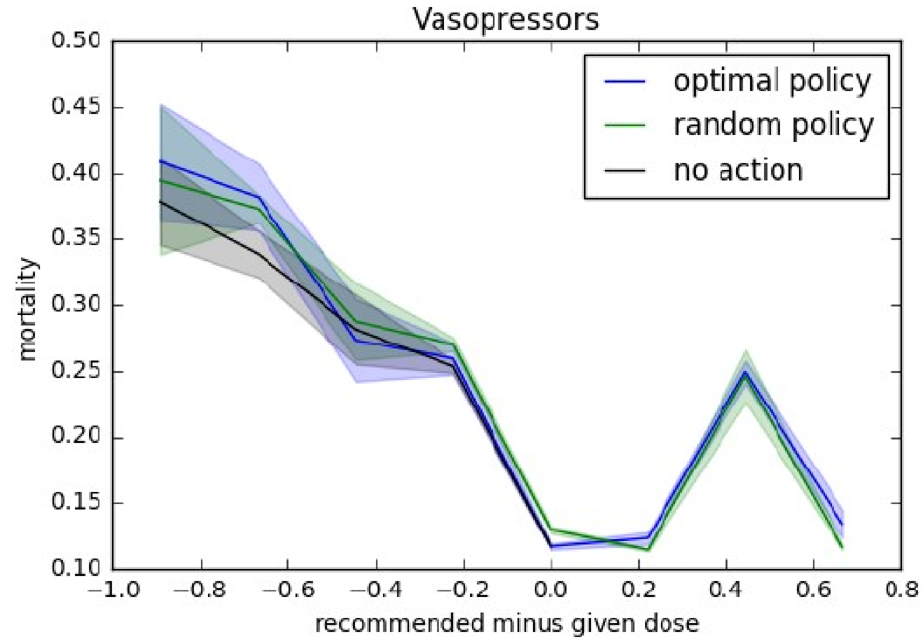




Application to Sepsis Management

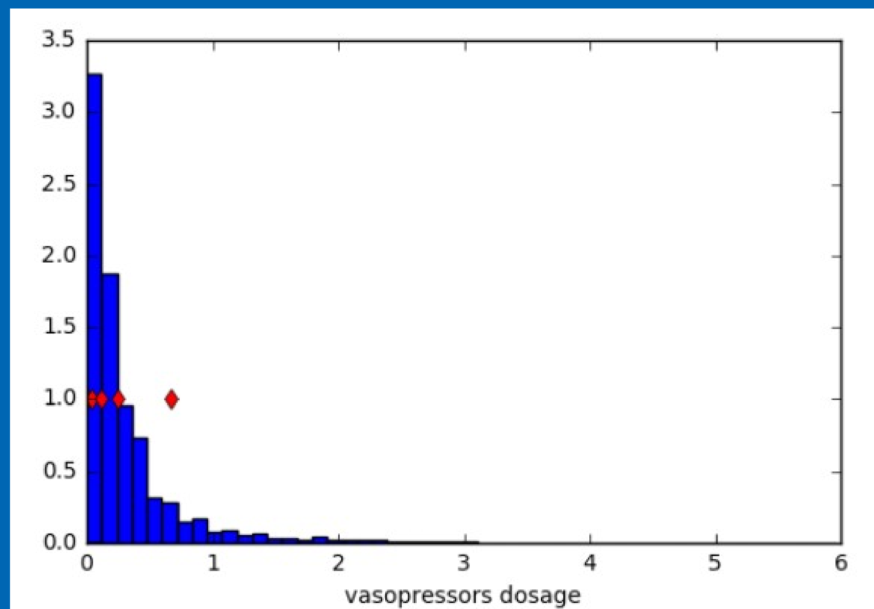
	Physician	Kernel	DQN	MoE_{V_d,Q_d}	MoE_{V_b,Q_b}
non-recurrent encoded	3.76	3.73	4.06	3.93	4.31
recurrent encoded	3.76	4.46	4.23	5.03	5.72

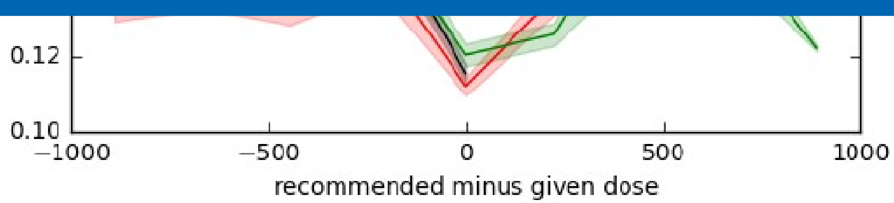


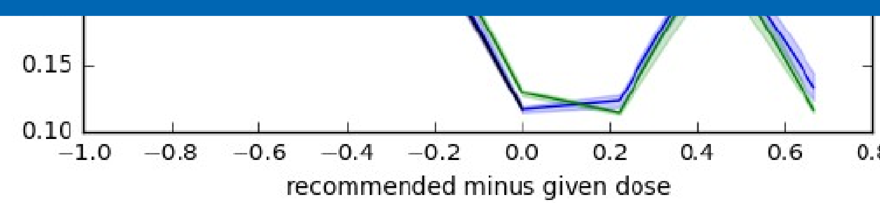


Application to Sepsis Management

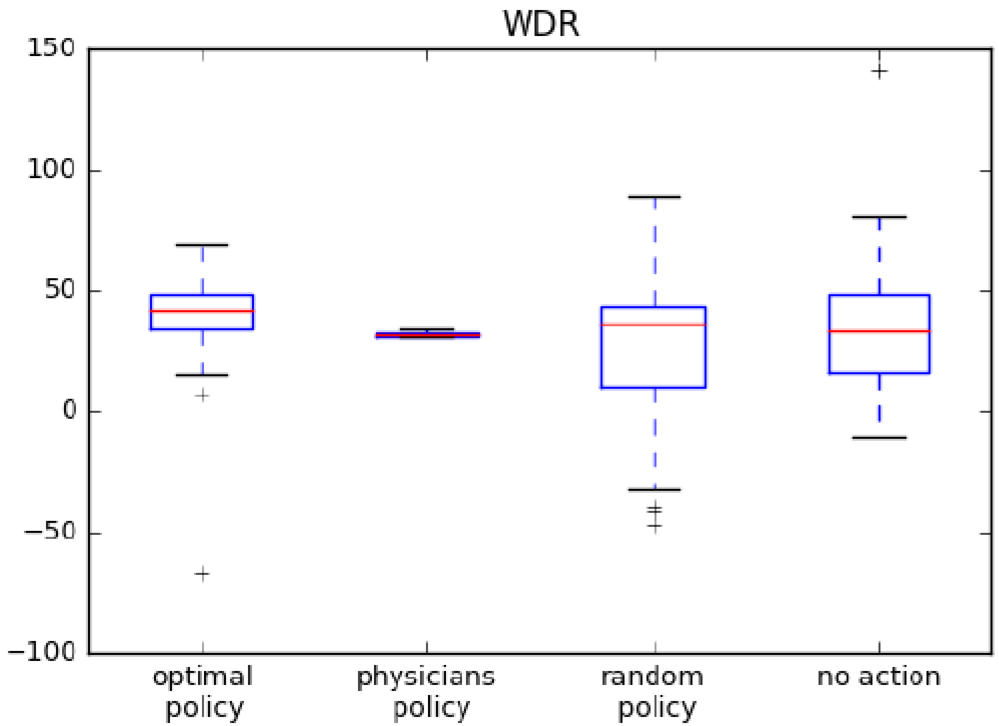
	Physician	Kernel	DQN	[illegible]	[illegible]





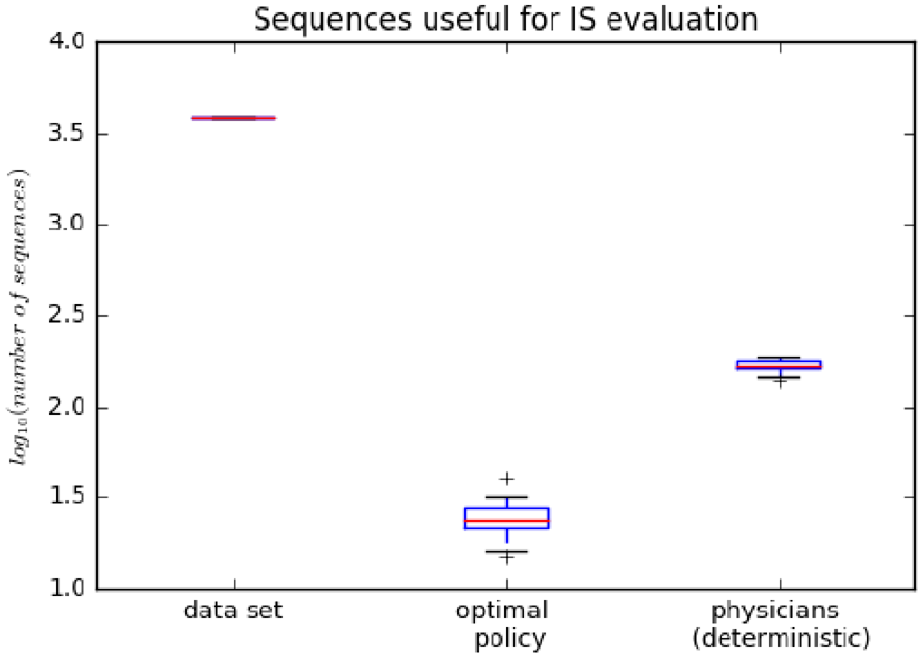


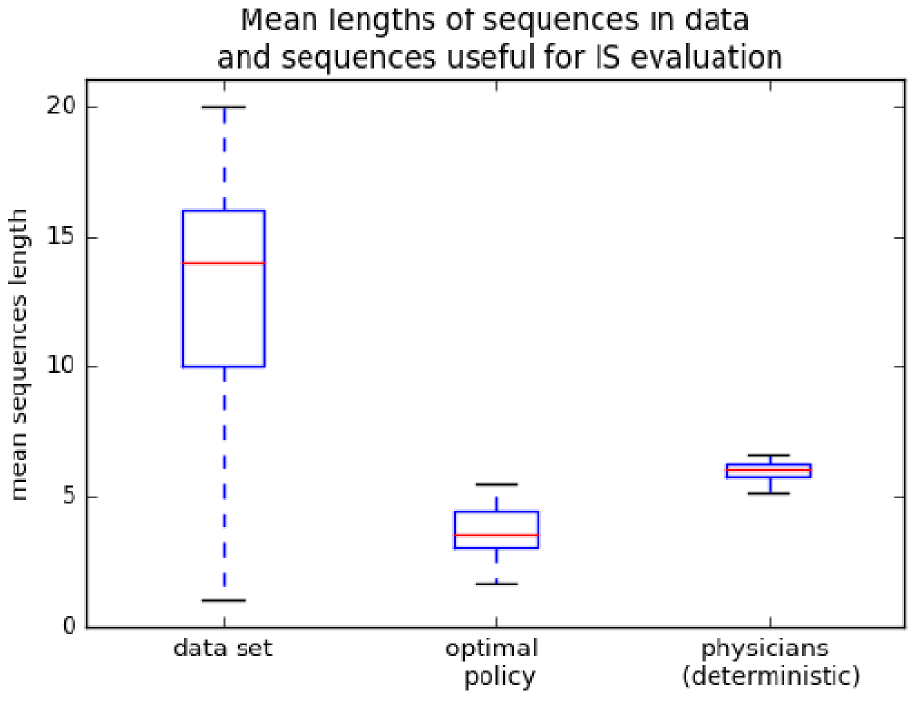
Just the start: Statistical Methods have high variance



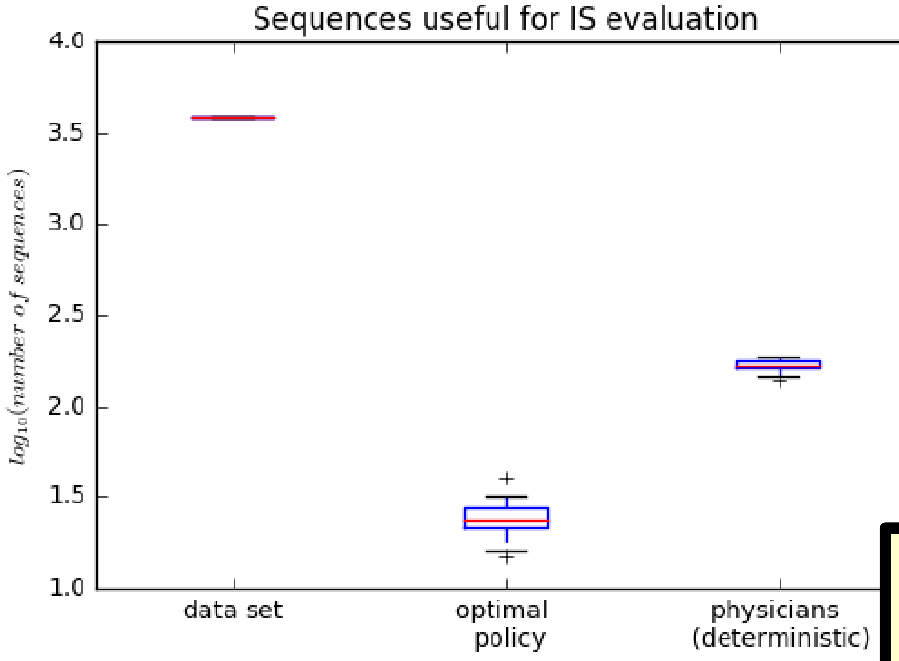
Gottesman et al. 2018, Raghu et al. 2018

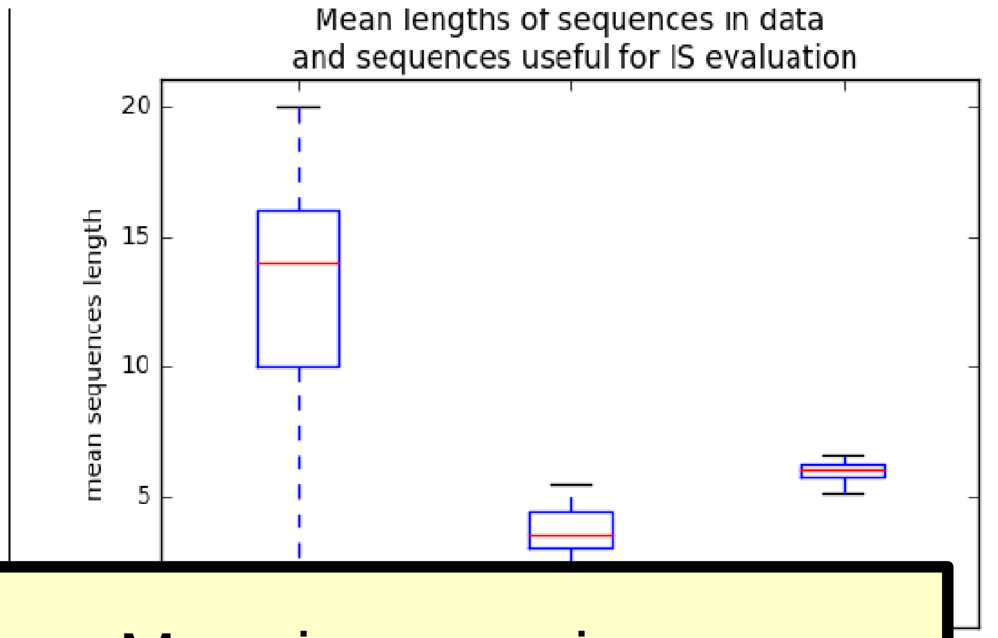
And select non-representative cohorts





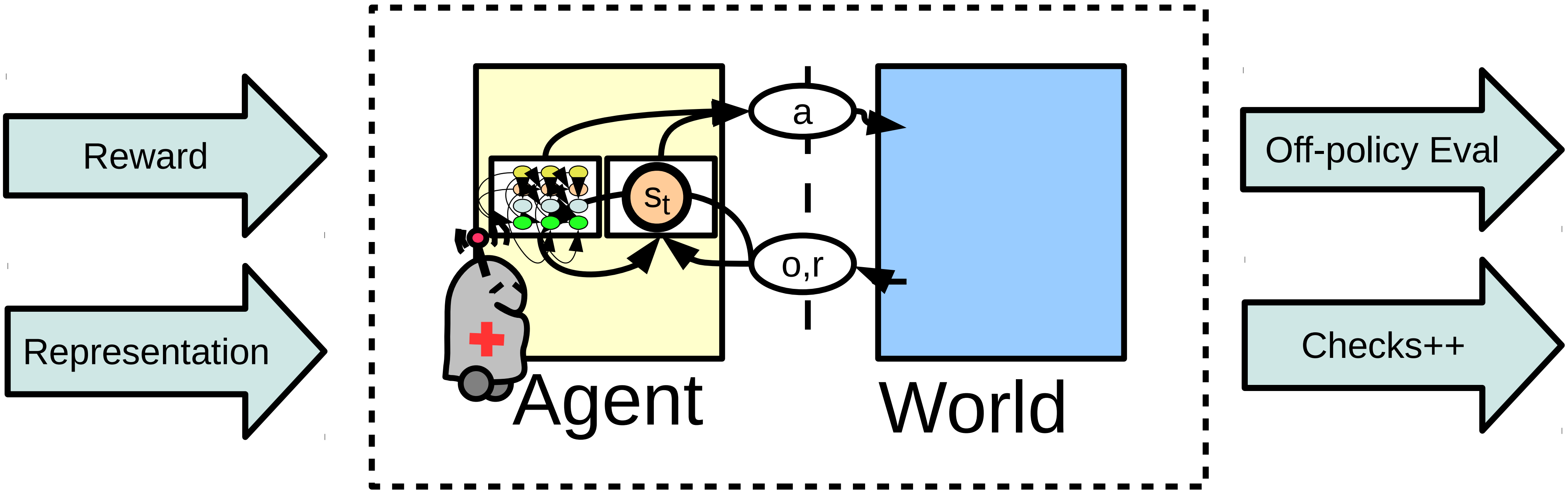
And select non-representative cohorts



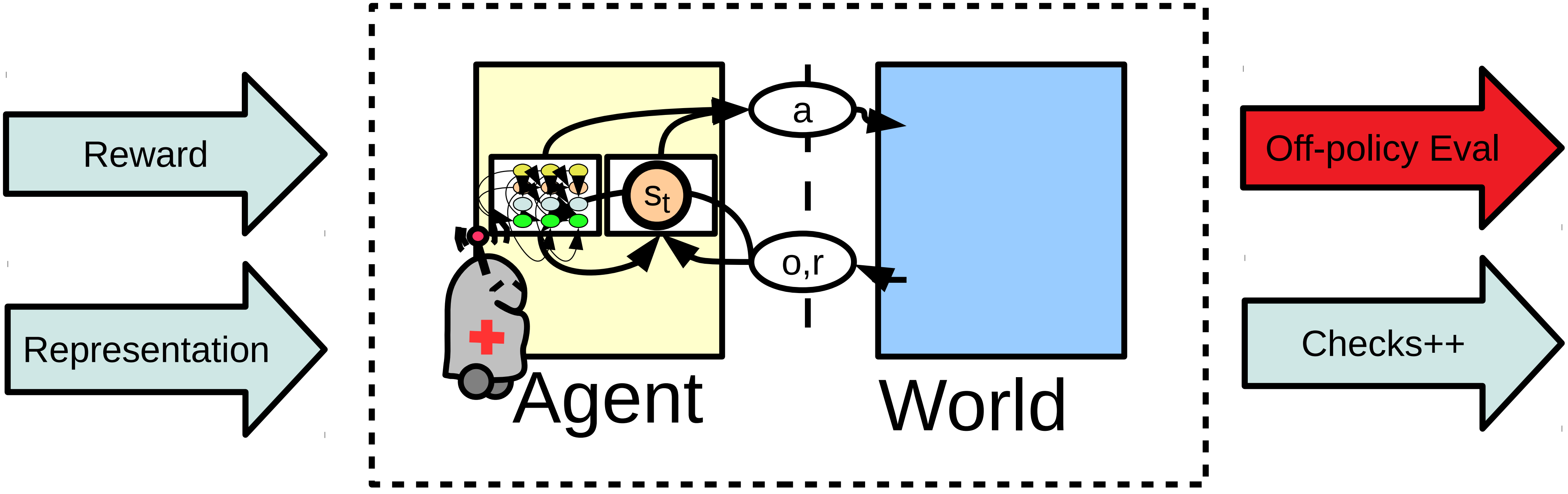


More issues arise with poor representation choices and poor reward functions

How can increase confidence in our results?



Off-Policy Evaluation



Off-Policy Evaluation

Core question: Given data collected under some behavior policy π_b, can we estimate the value of some other evaluation policy π_e?

Three main kinds of approaches:

- Importance-sampling: reweight current data (high variance)

$$\rho_n = \prod_t \frac{\pi_e(a_{tn}|s_{tn})}{\pi_b(a_{tn}|s_{tn})}$$

- Model-based: build model with current data, simulate (high bias)
- Value-based: apply value evaluation to current data (high bias)

Off-Policy Evaluation

Core question: Given data collected under some behavior policy π_b, can we estimate the value of some other evaluation policy π_e?

Three main kinds of approaches:

- Importance-sampling: reweight current data (high variance)

$$\rho_n = \prod_t \frac{\pi_e(a_{tn} | s_{tn})}{\pi_b(a_{tn} | s_{tn})}$$

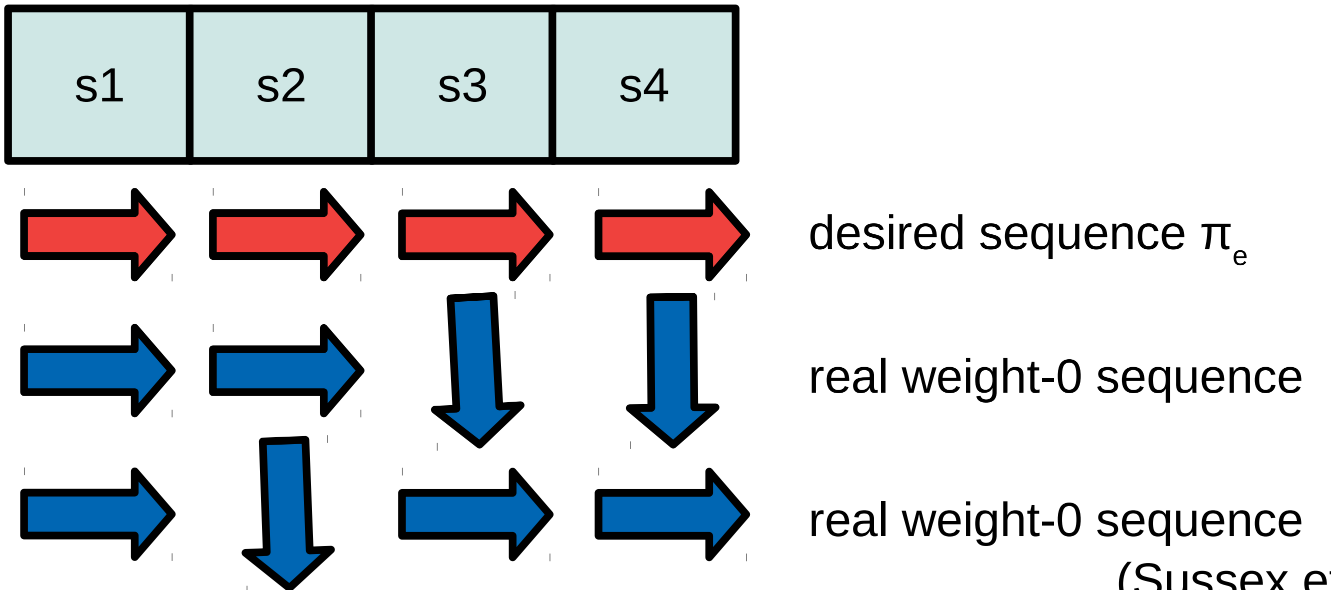
- Model-based: build model with current data, simulate (high bias)
- Value-based: apply value evaluation to current data (high bias)

Stitching to Increase Sample Sizes

Importance sampling-based estimators suffer because importance weights most importance weights get small very fast:

$$\rho_n = \prod_t \frac{\pi_e(a_{tn}|s_{tn})}{\pi_b(a_{tn}|s_{tn})}$$

One way to ameliorate the issue: “stitch” trajectories with zero weight to get more non-zero weight trajectories.



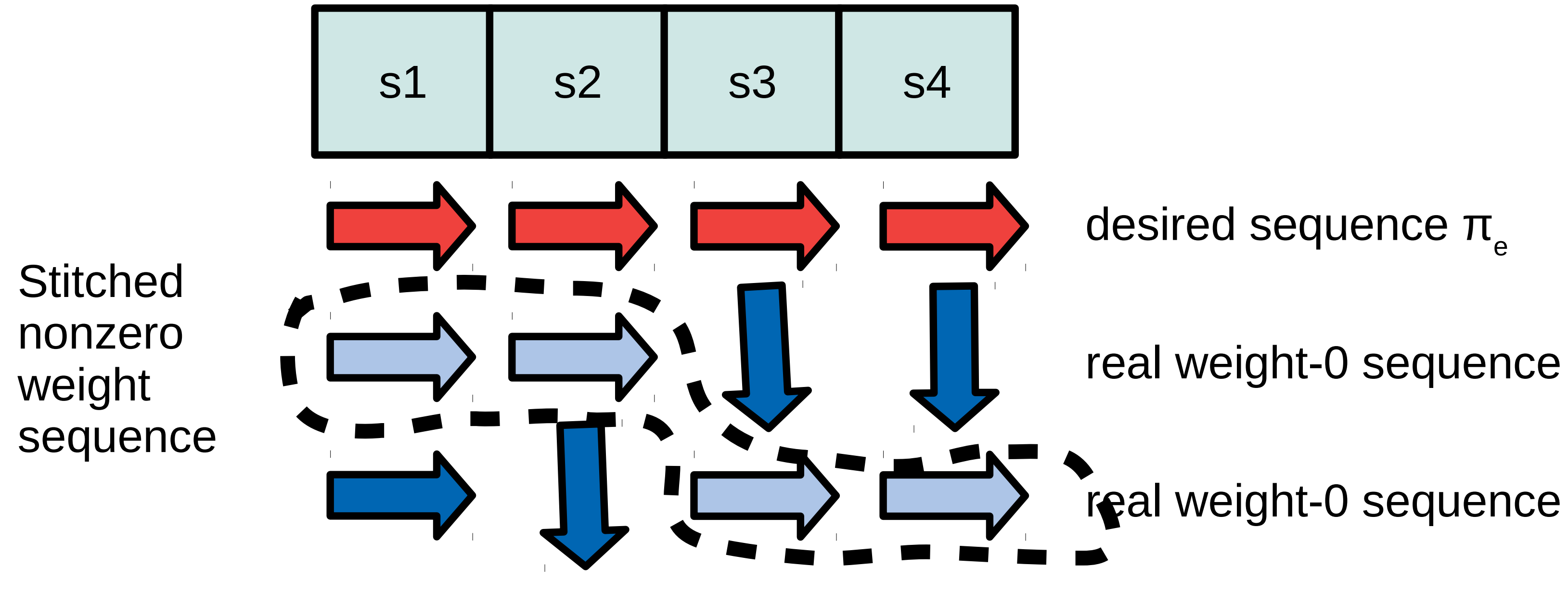
(Sussex et al, ICMLWS 2018)

Stitching to Increase Sample Sizes

Importance sampling-based estimators suffer because importance weights most importance weights get small very fast:

$$\rho_n = \prod_t \frac{\pi_e(a_{tn}|s_{tn})}{\pi_b(a_{tn}|s_{tn})}$$

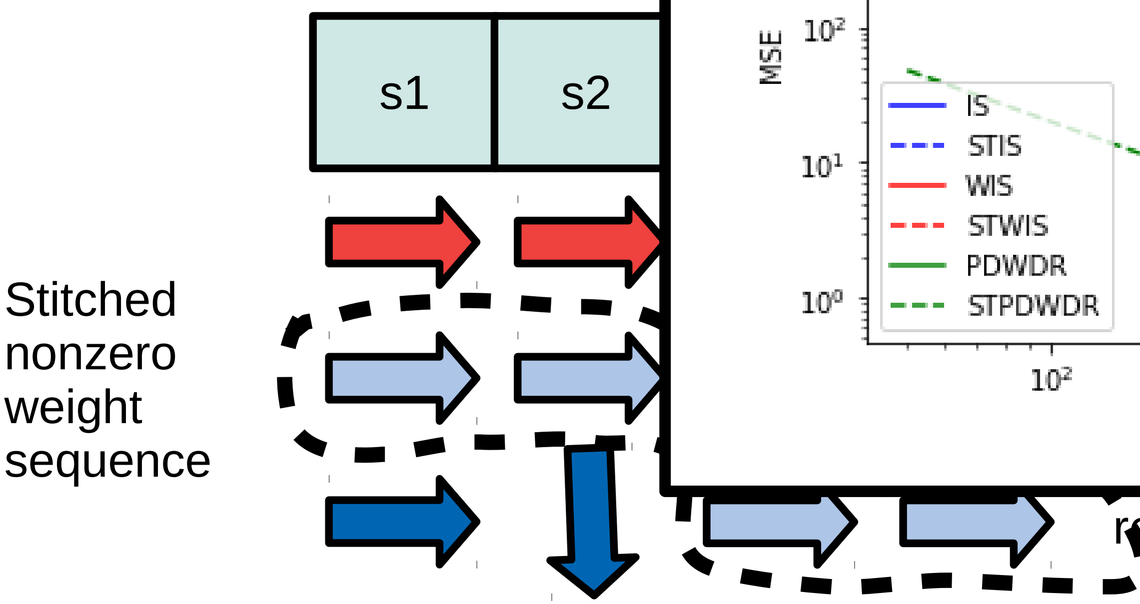
One way to ameliorate the issue: "stitch" trajectories with zero weight to get more non-zero weight trajectories.

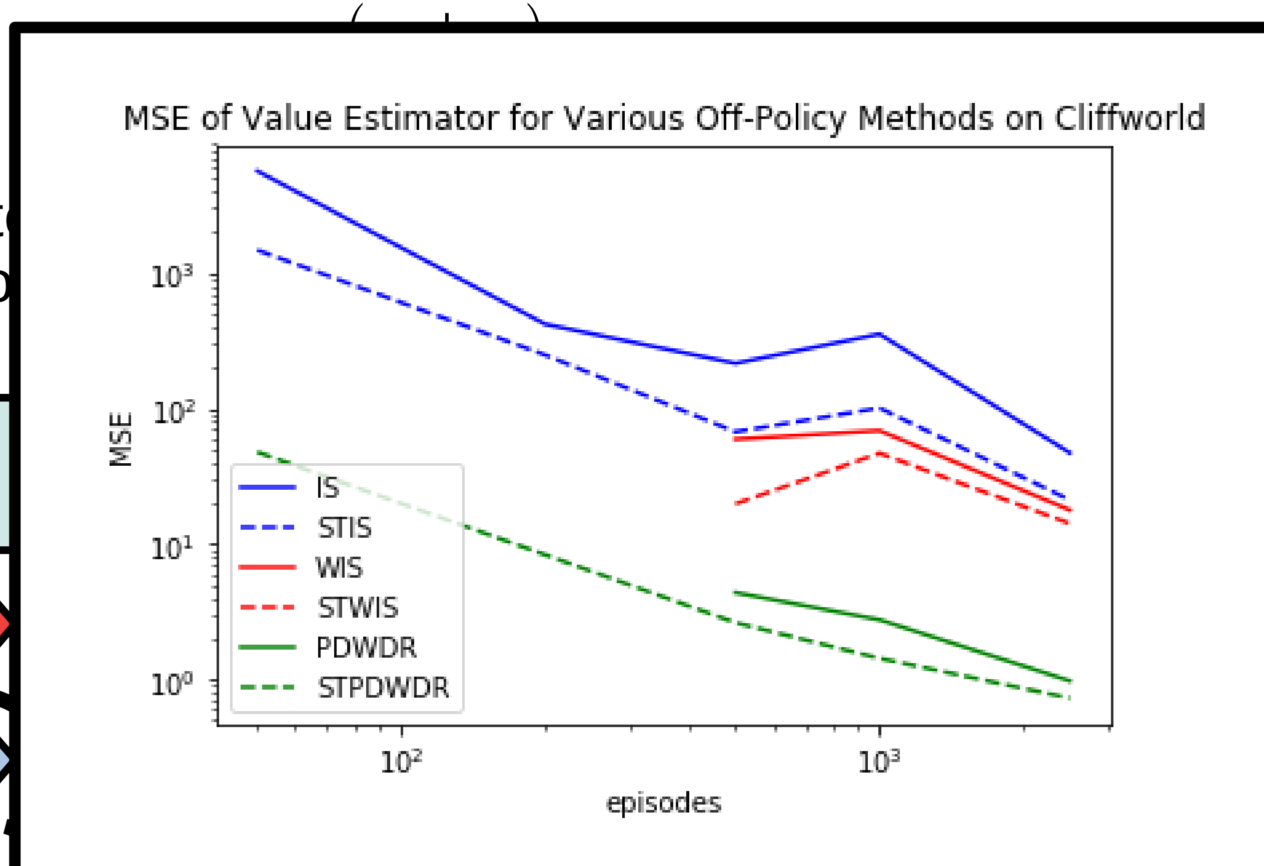


Stitching to Increase Sample Sizes

Importance sampling-based estimators suffer because importance weights most importance weights get small very fast:

One way to ameliorat
weight to get more no





Off-Policy Evaluation

Core question: Given data collected under some behavior policy π_b, can we estimate the value of some other evaluation policy π_e?

Three main kinds of approaches:

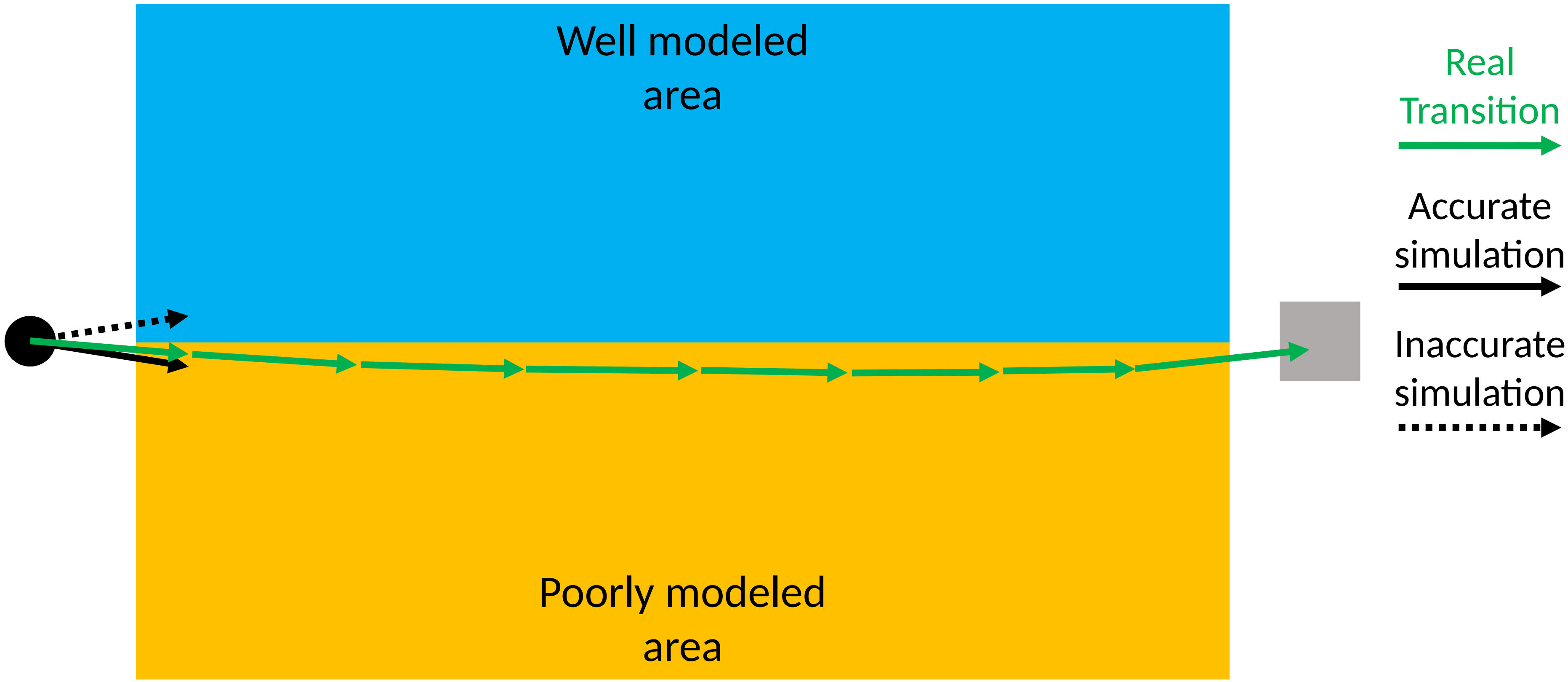
- Importance-sampling: reweight current data (high variance)

$$\rho_n = \prod_t \frac{\pi_e(a_{tn}|s_{tn})}{\pi_b(a_{tn}|s_{tn})}$$

- Model-based: build model with current data, simulate (high bias)
- Value-based: apply value evaluation to current data (high bias)

Better Models: Mixtures help again!

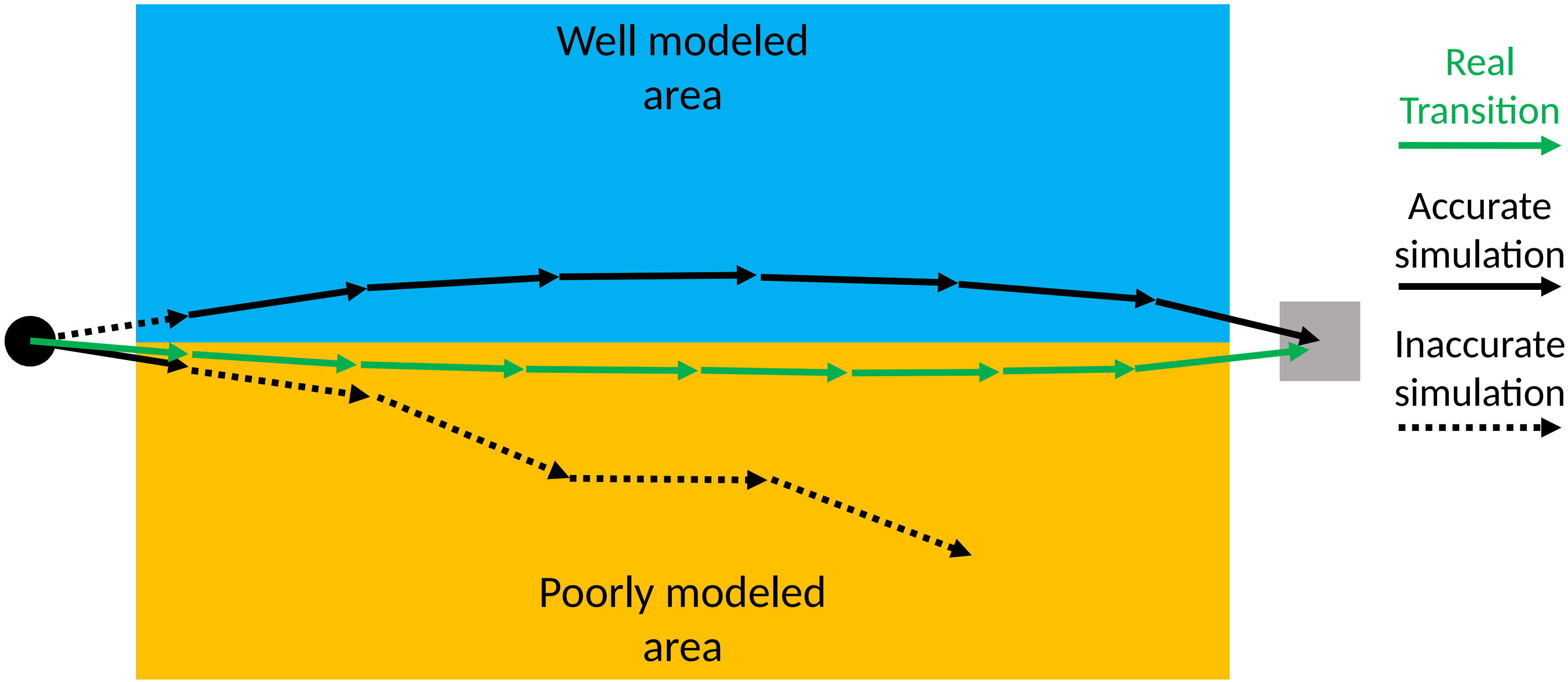
We use RL to bound the long-term accuracy of the value estimate.



(Gottesman et al, ICML 2019)

Better Models: Mixtures help again!

We use RL to bound the long-term accuracy of the value estimate.



Bound on the Quality

$$\left|g_T - \hat{g}_T\right| \leq L_r \sum_{t=0}^{T} \gamma^t \sum_{t'=0}^{t-1} \left(L_t\right)^{t'} \varepsilon_t\left(t - t' - 1\right) + \sum_{t=0}^{T} \gamma^t \varepsilon_r(t)$$

Total return error

Error due to state estimation

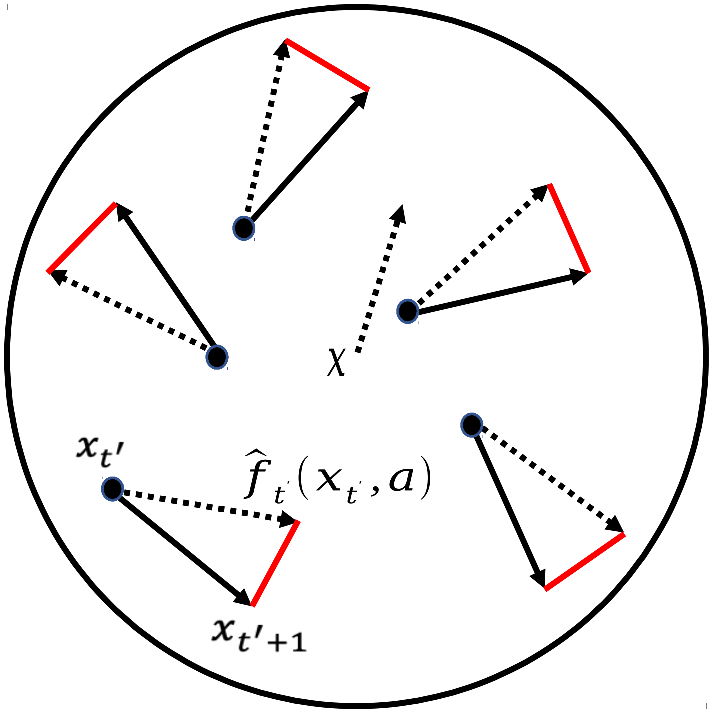
Error due to reward estimation

$L_{t/r}$ - Lipschitz constants of transition/reward functions
$\varepsilon_{t/r}(t)$ - Bound on model errors for transition/reward at time t
T - Time horizon
γ - Reward discount factor
$g_T \equiv \sum_{t=0}^{T} \gamma^t r(t)$ - Return over entire trajectory

Closely related to bound in - *Asadi, Misra, Littman. "Lipschitz Continuity in Model-based Reinforcement Learning." (ICML 2018).*

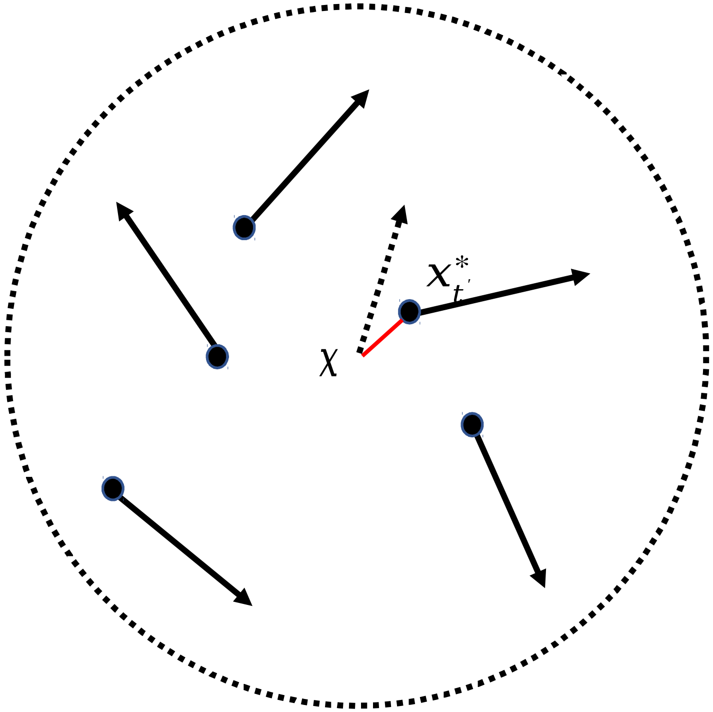
Estimating Errors

Parametric



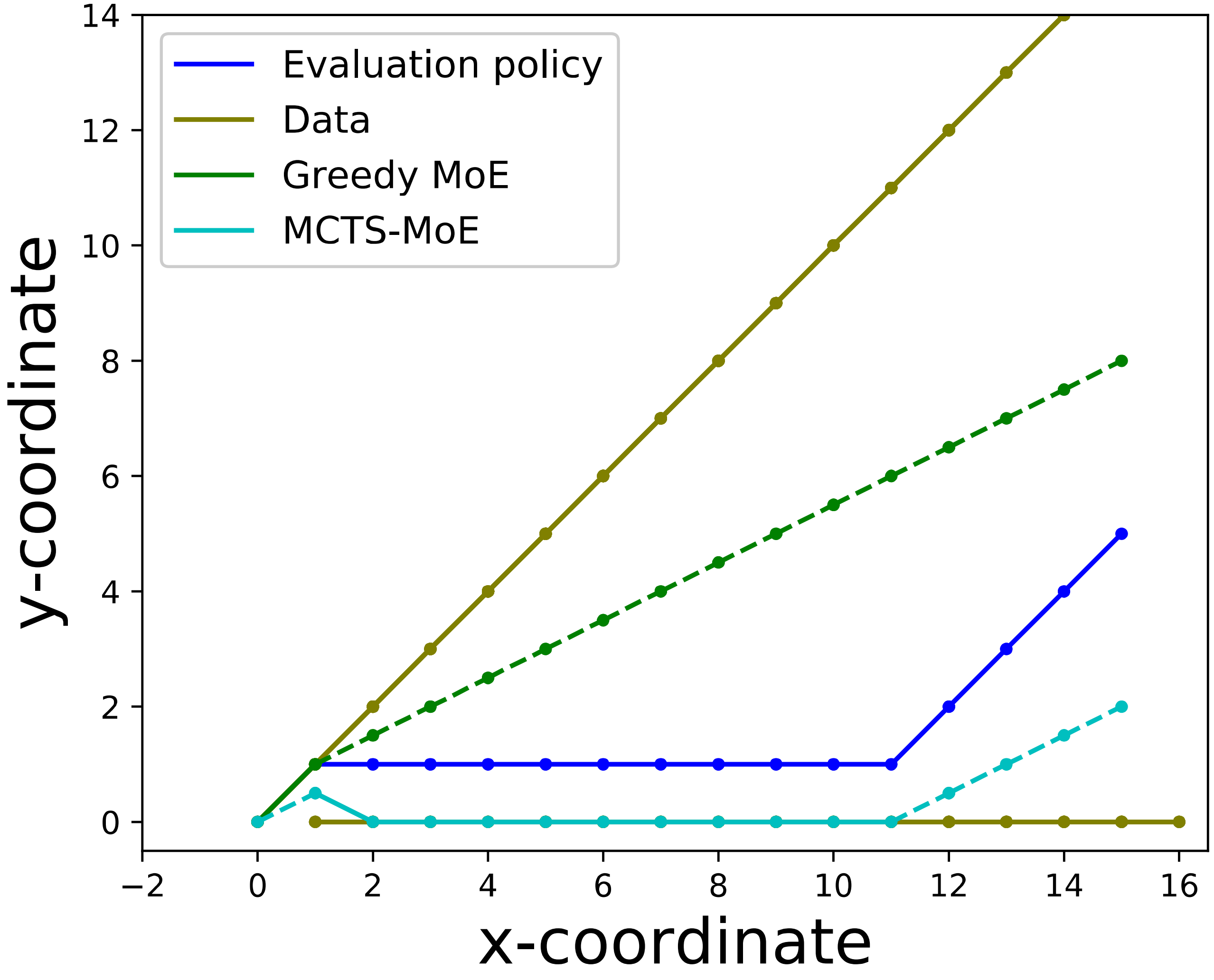
$$\hat{\varepsilon}_{t,p} \approx \max \Delta(x_{t'+1}, \hat{f}_t(x_{t'}, a))$$

Nonparametric

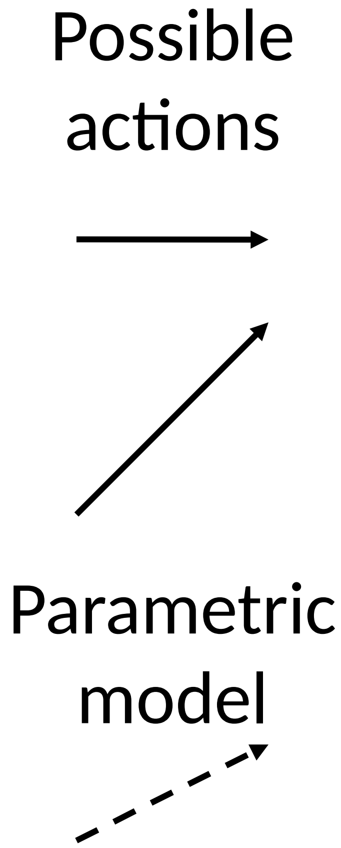


$$\hat{\varepsilon}_{t,np} \approx L_t \cdot \Delta(x, x^*_{t'})$$

Toy Example



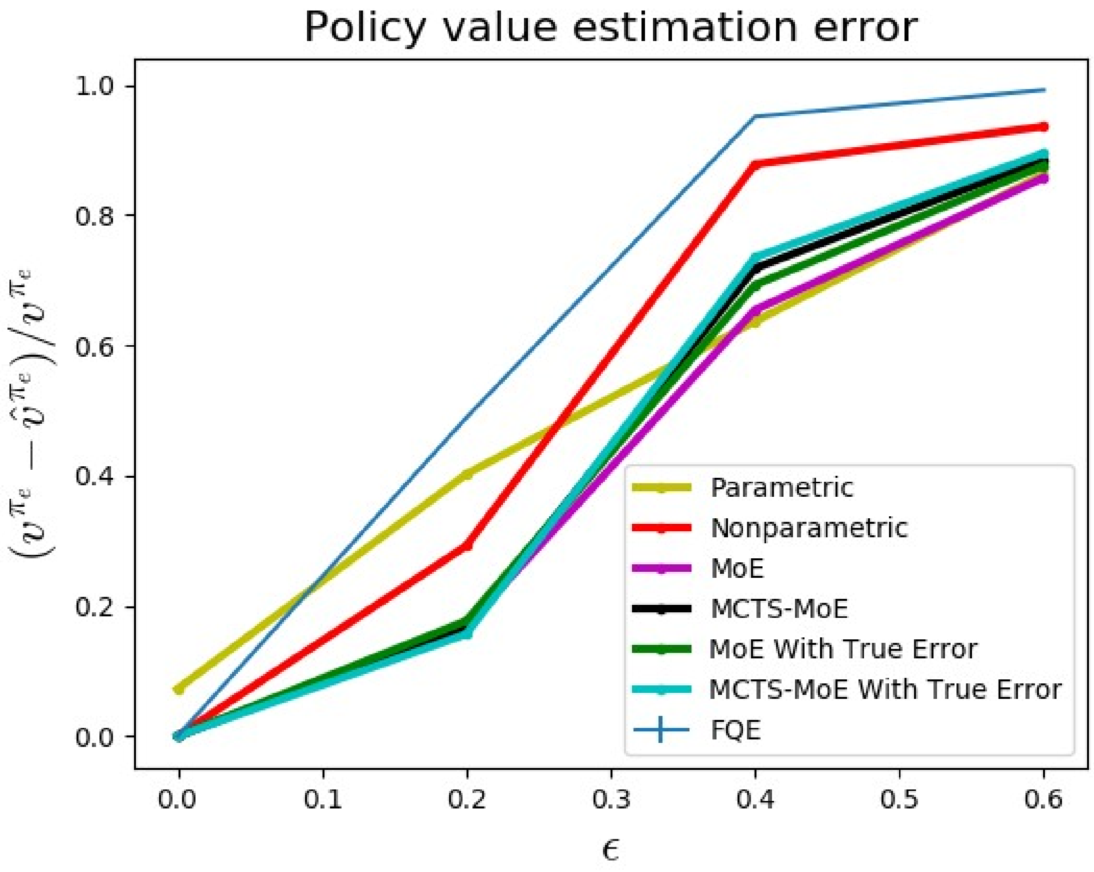
Evaluation policy
Data
Greedy MoE
MCTS-MoE
y-coordinate
x-coordinate
0
2
4
6
8
10
12
14
−2
16



Possible actions
Parametric model

Example with HIV Simulator

We use RL to bound the long-term accuracy of the value estimate.



Better Models: Designed for Evaluation

Main objective: find a model that will minimize error in individual treatment effects:

$$\cancel{\left(E_{s_0}\left[V^{\pi}(s_0)\right]-E_{s_0}\left[\hat{V}^{\pi}(s_0)\right]\right)^2}$$

$$E_{s_0}\left[\left(V^{\pi}(s_0)-\hat{V}^{\pi}(s_0)\right)^2\right]$$

where the value function is estimated via trajectories from an approximated model M. Question: Can we do better than just optimizing M for p(M|data)?

Show this can be optimized via a transfer-learning type objective:

$$L(M)=\underbrace{\sum_{nt} l(M,n,t)}_{\text{“on-policy” loss}}+\underbrace{\sum_{nt}\rho_{nt}\, l(M,n,t)}_{\text{“reweighted for }\pi_e\text{” loss}}+\ldots$$

(Liu et al, NIPS 2018)

Better Models: Designed for Evaluation

Main objective: find a model that will minimize error in individual treatment effects:

$$\cancel{\left(E_{s_0}[V^{\pi}(s_0)]-E_{s_0}[\hat{V}^{\pi}(s_0)]\right)^2}$$

$$E_{s_0}\left[\left(V^{\pi}(s_0)-\hat{V}^{\pi}(s_0)\right)^2\right]$$

where the value

approximated m

optimizing M for

Show this can be

$$L(M)$$

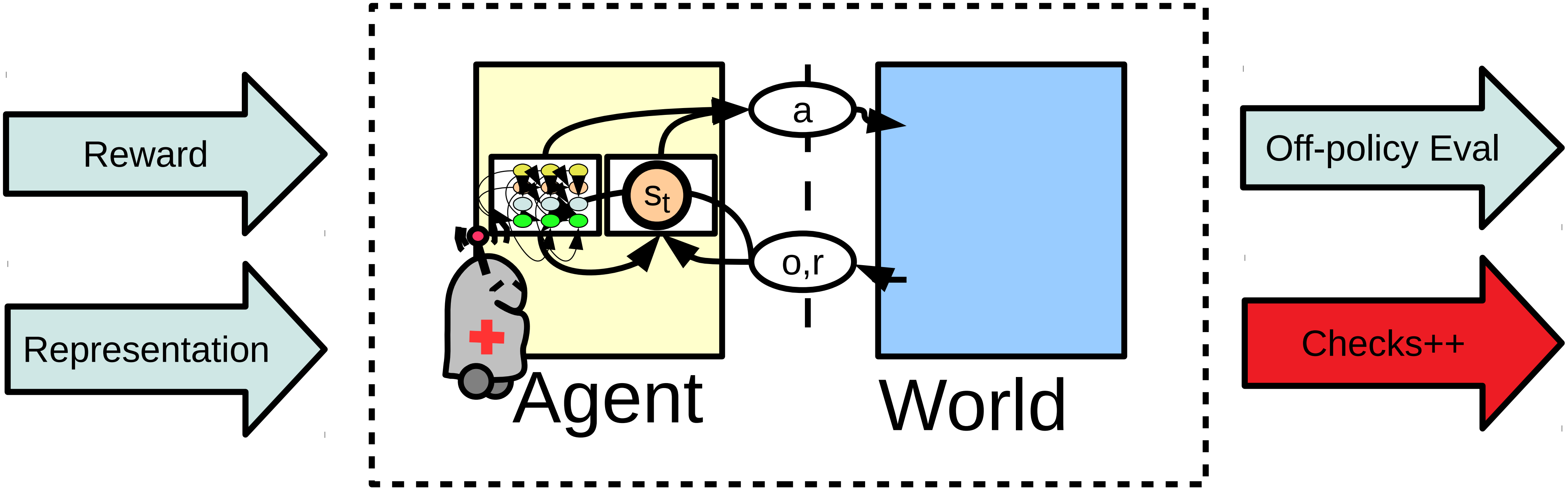
Table 1: Root MSE for Cart Pole

Long Horizon	RepBM	DR	AM	DR(AM)	AM(π)	MRDR Q	MRDR	IS
Mean	**0.4121**	1.359	0.7535	1.786	41.80	151.1	202	194.5
Individual	**1.033**	-	1.313	-	47.63	151.9	-	-
Short Horizon	**RepBM**	**DR**	**AM**	**DR(AM)**	**AM(π)**	**MRDR Q**	**MRDR**	**IS**
Mean	0.07836	**0.02081**	0.1254	0.0235	0.1233	3.013	0.258	2.86
Individual	**0.4811**	-	0.5506	-	0.5974	3.823	-	-

Table 2: Root MSE for Mountain Car

	RepBM	DR	AM	DR(AM)	AM(π)	MRDR Q	MRDR	IS
Mean	**12.31**	135.8	17.15	141.6	72.61	135.4	172.7	149.7
Individual	**31.38**	-	36.36	-	79.46	138.1	-	-

Checking the reasonableness of our policies



Some Basic Digging

Positive Evidence: Reproducing across sites (robust to covariate shift)

Our HIV results hold across two distinct cohorts.

EU Resist

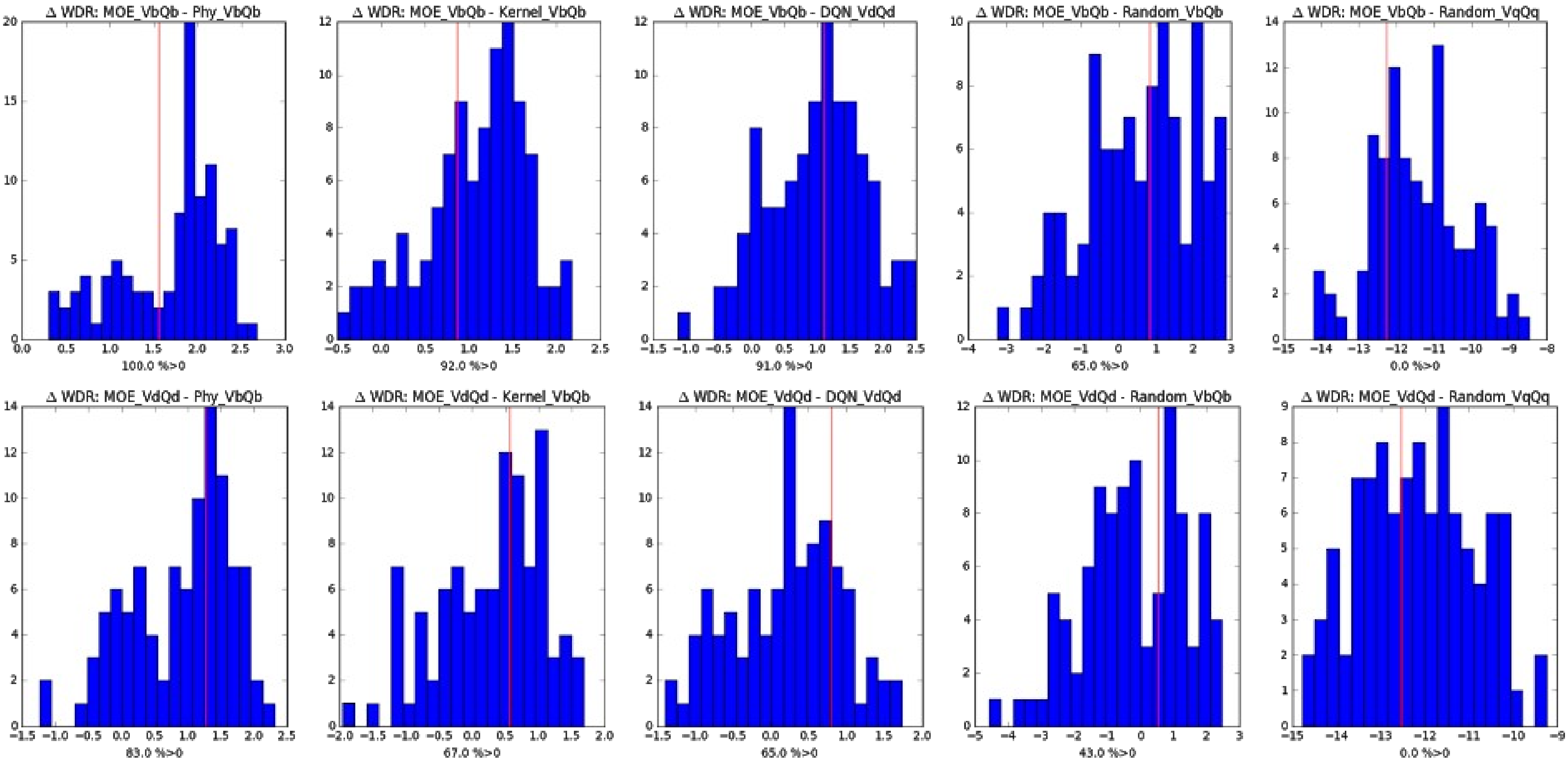
	Doubly Robust	**Importance Sampling**	**Weighted Importance**
Random Policy	–2.31 ± 1.42	–3.48 ± 1.36	–2.80 ± 1.27
Short-term Kernel	2.17 ±1.4	2.18 ±1.20	2.16 ±1.71
Long-term Kernel	9.47 ± 1.70	5.72 ± 1.81	6.97 ± 1.29
POMDP	6.04 ± 2.18	4.15 ± 2.28	6.67 ± 1.74
Mixture-of-experts	**11.83 ± 1.26**	**12.50 ± 1.19**	**11.07 ± 1.21**

Swiss HIV Cohort

	Doubly Robust	**Importance Sampling**	**Weighted Importance**
Random Policy	–6.33 ± 3.47	–5.57 ± 2.17	–6.18 ± 3.24
Short-term Kernel	1.64 ±1.86	2.03 ±1.81	2.17 ±1.74
Long-term Kernel	9.67 ± 1.49	7.38 ± 1.72	7.64 ± 1.92
POMDP	5.46 ± 2.05	6.72 ± 2.88	7.76 ± 2.10
Mixture-of-experts	**10.73 ± 1.02**	**13.59 ± 1.57**	**11.83 ± 1.31**

Positive Evidence:
Check importance weights, variances

Sepsis: results hold with different control variates



Ask the Experts

Asking the Doctors

- HIV: Checking against standard of care:

	NNRTIs	NRTIs	PIs	Fusion/Entry Inhibitors
First-line therapy	12 157	3 054	774	128
Second-line therapy	4 068	8 764	6 082	1 042

- As well as three expert clinicians:

	Clinician 1	Clinician 2	Clinician 3
Agree	18	15	13
Partially Agree	10	11	13
Disagree	2	4	4

Asking the Doctors

- HIV: Checking against standard of care:

		nhibitors
First-		
Secon		

What's the best way
to "ask the doctors"?

- As well as three expert clinicians:

	Clinician 1	Clinician 2	Clinician 3
Agree	18	15	13
Partially Agree	10	11	13
Disagree	2	4	4

Detour: Summarizing a Treatment Policy

How can we best communicate a treatment policy to a clinical expert? Formalize as the following game:

Us: Present expert with some state-action pairs

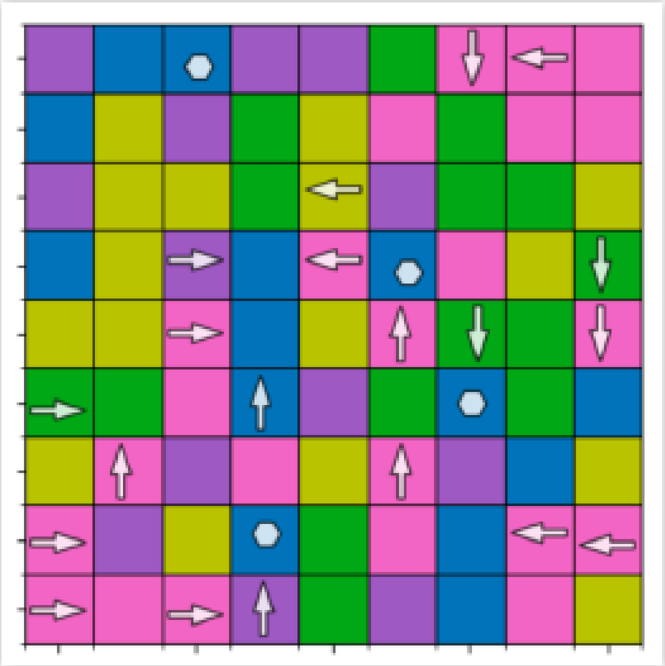
Expert: Predict the agent's action in a new state, s'

Our Goal: choose the state-action pairs so the expert predicts the best.

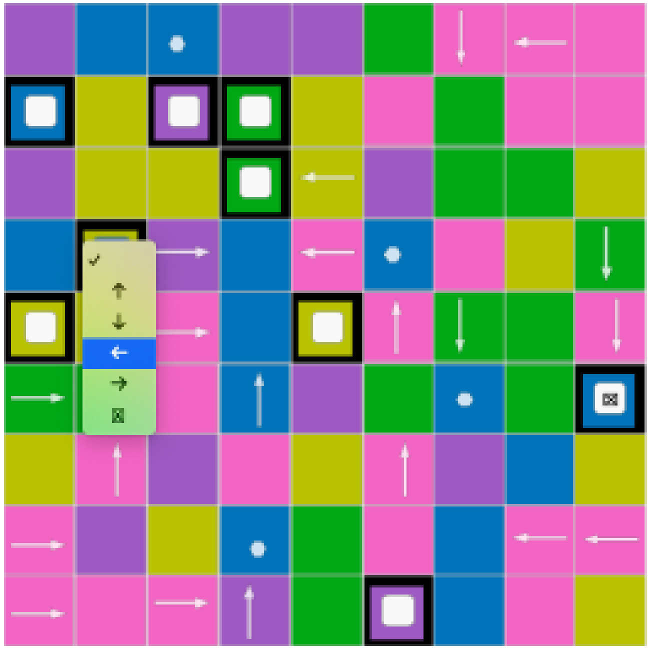
(Lage et al, IJCAI 2019)

Example 1: Gridworld

Given:

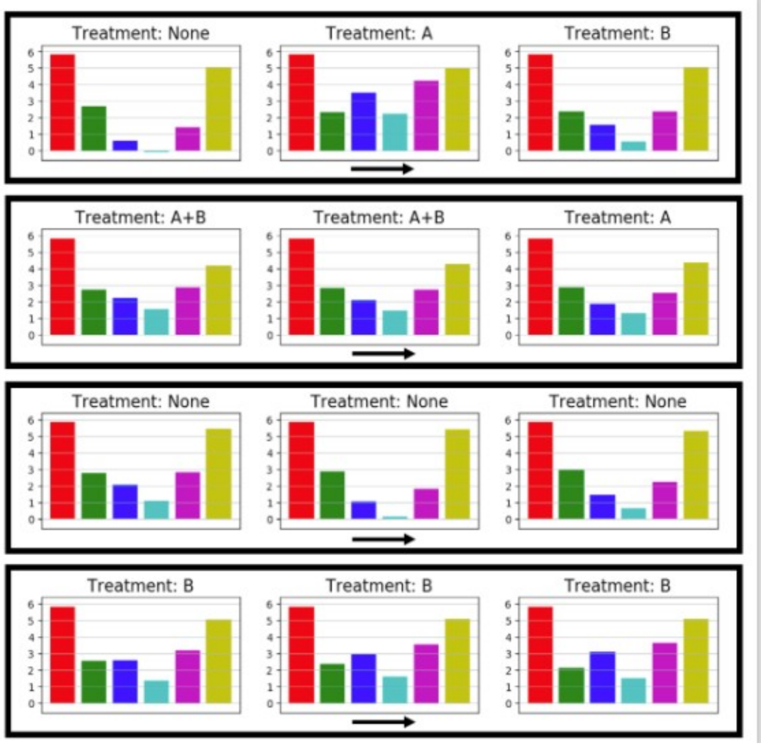


What happens in states like:

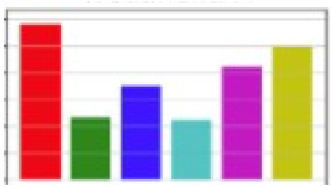


Example 2: HIV Simulator

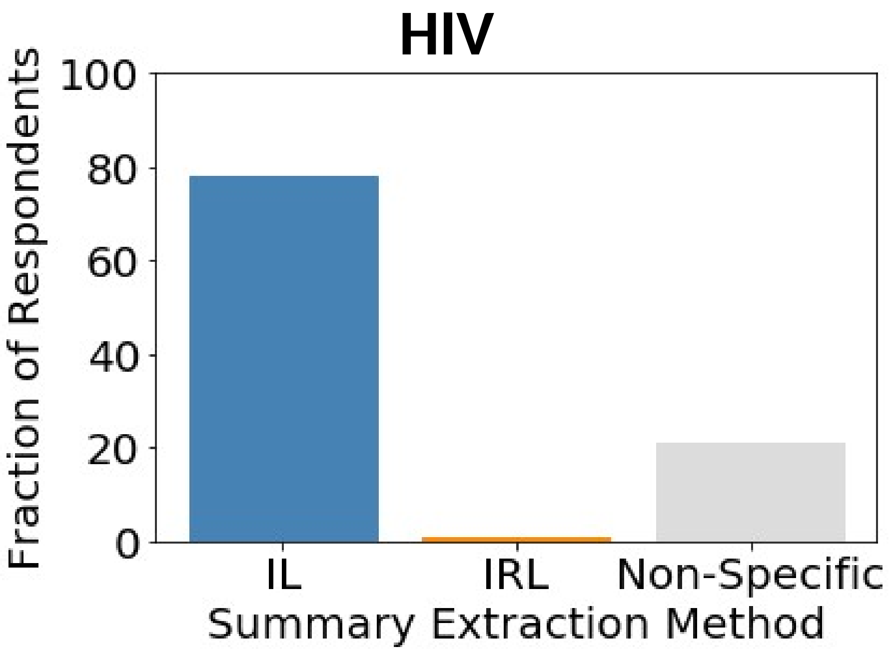
Given:

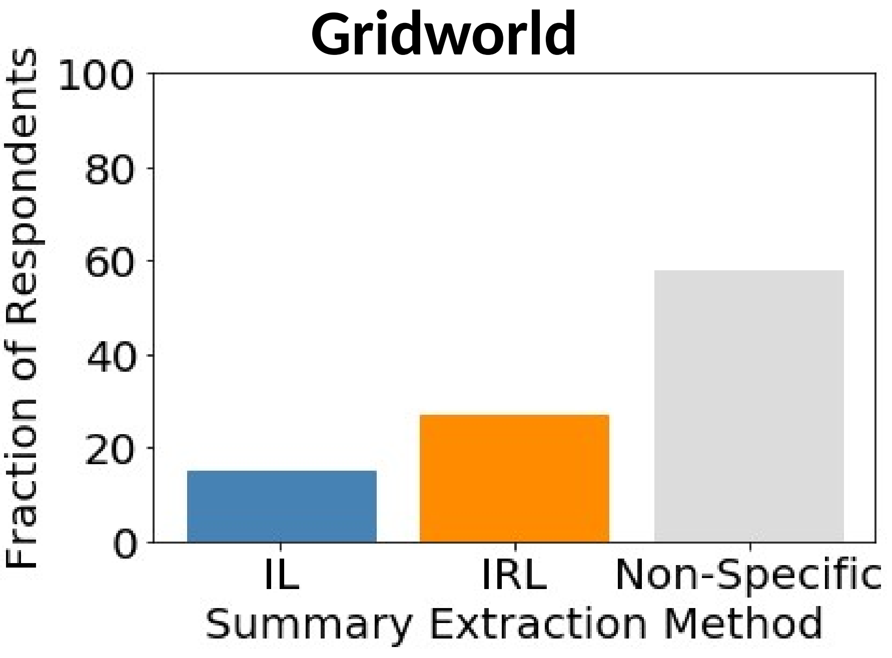


What happens in states like:

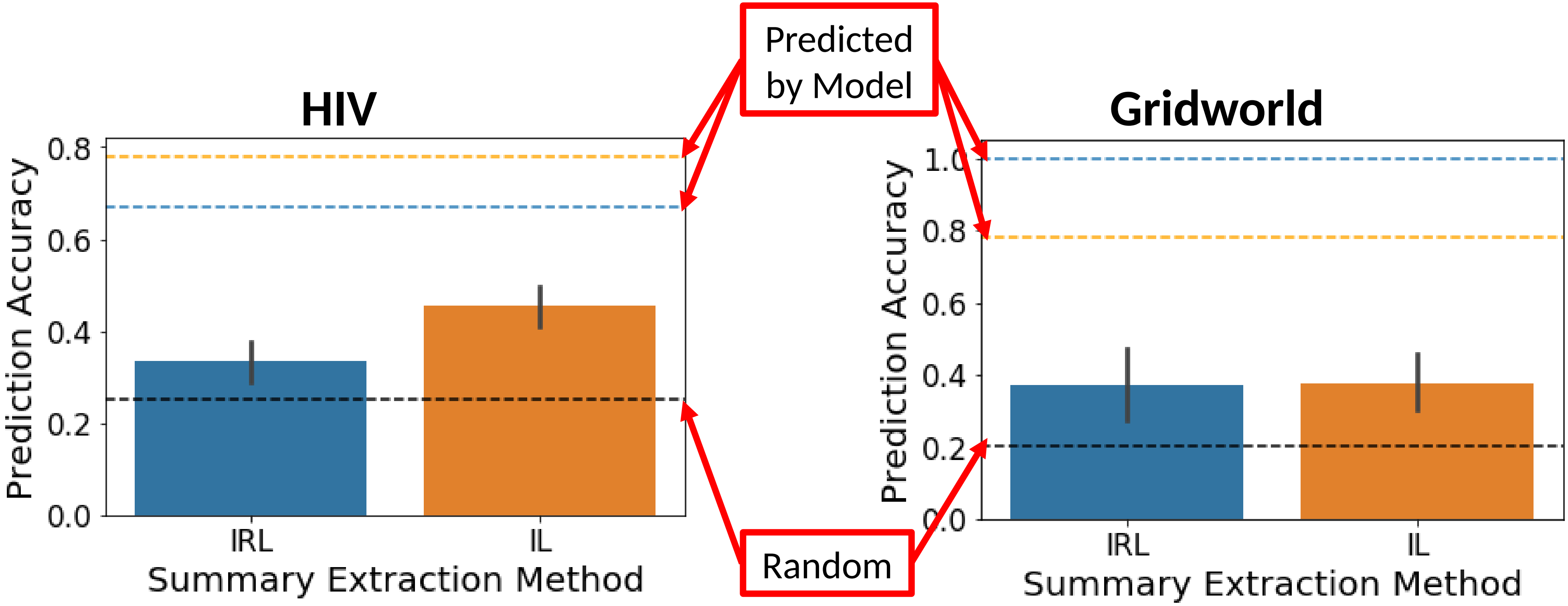


Finding: Humans use different methods in different scenarios





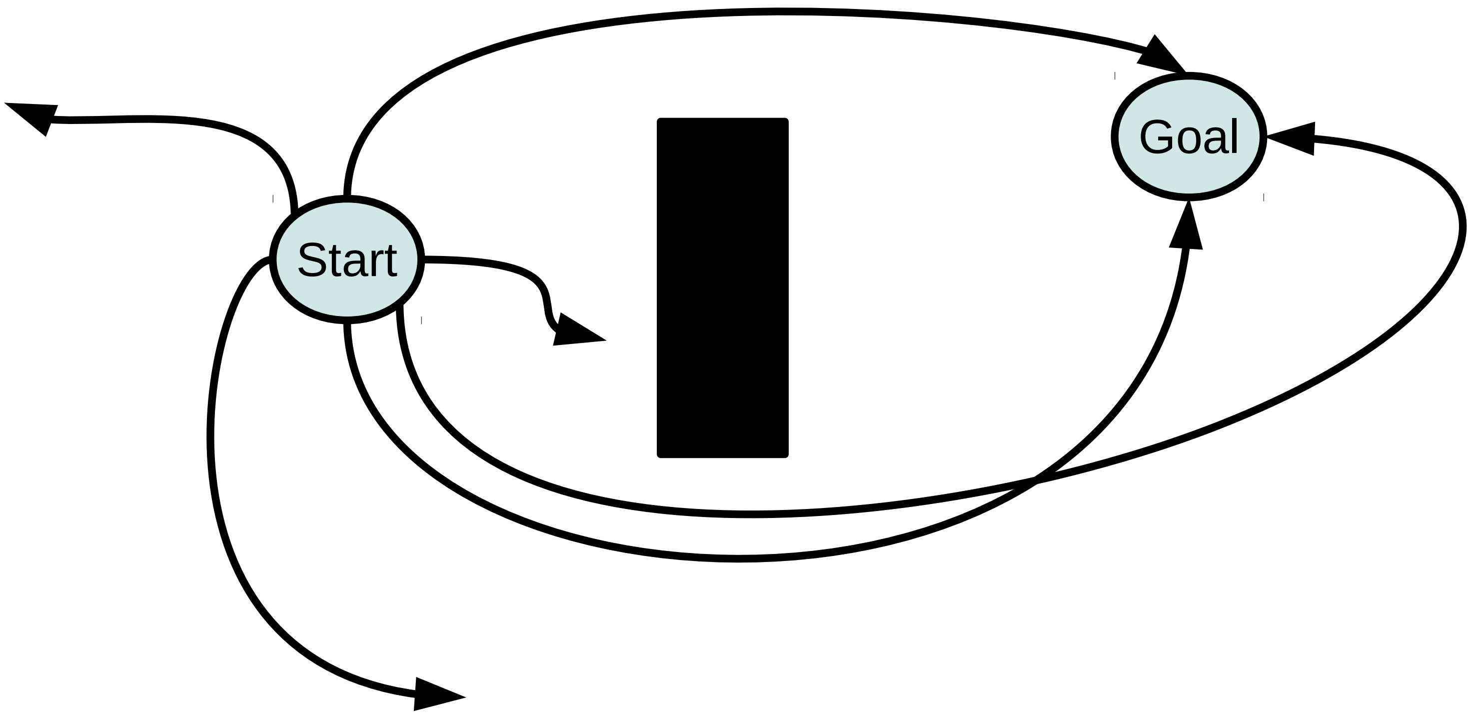
…and it's important to account for it!



Offering Options

In Progress: Displaying Diverse Alternatives

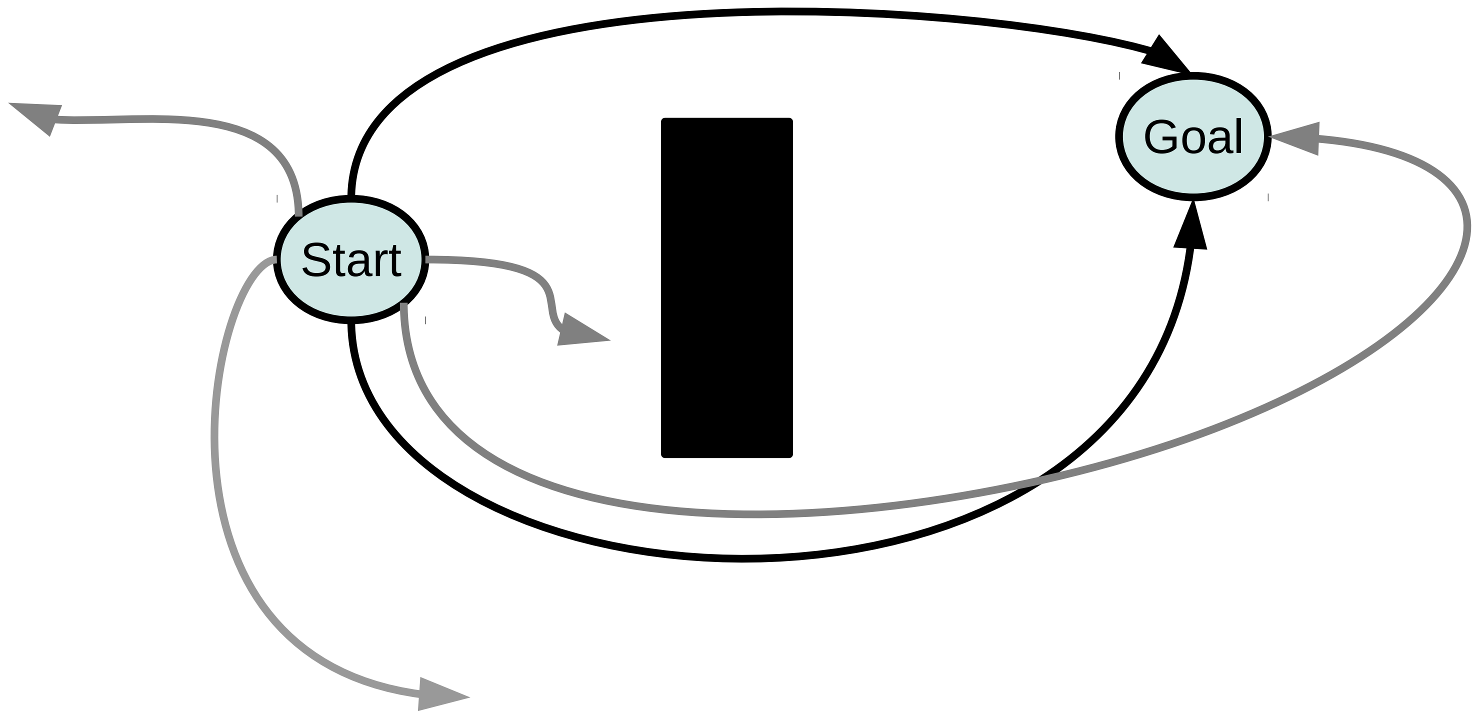
If policies can't be statistically differentiated, share all the options.



(Masood et al, IJCAI 2019)

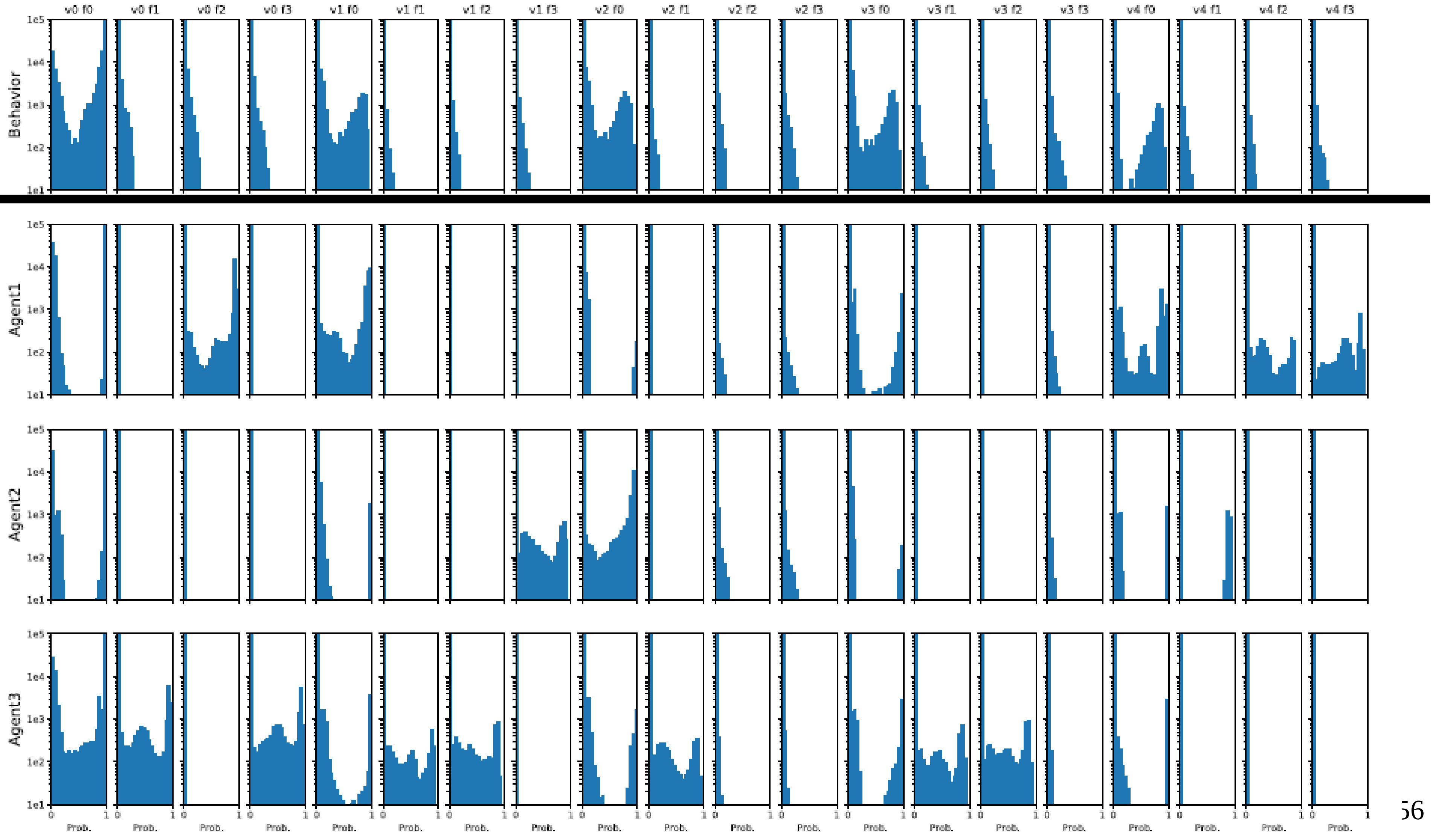
In Progress: Displaying Diverse Alternatives

If policies can't be statistically differentiated, share all the options.



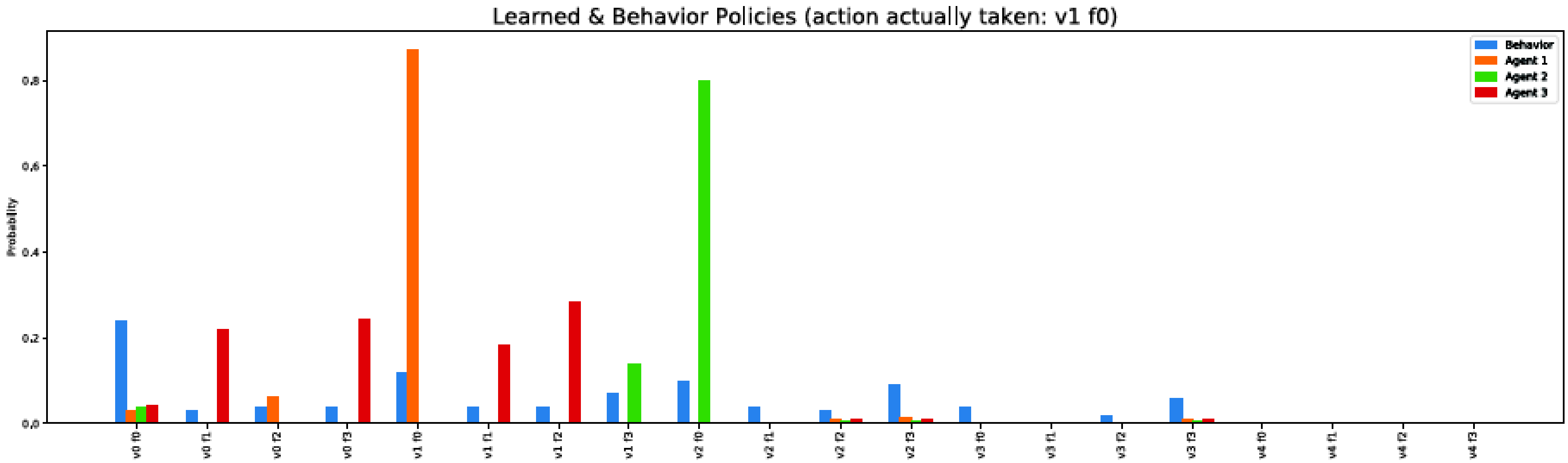
(Masood et al, IJCAI 2019)

Applied to Hypotension Management

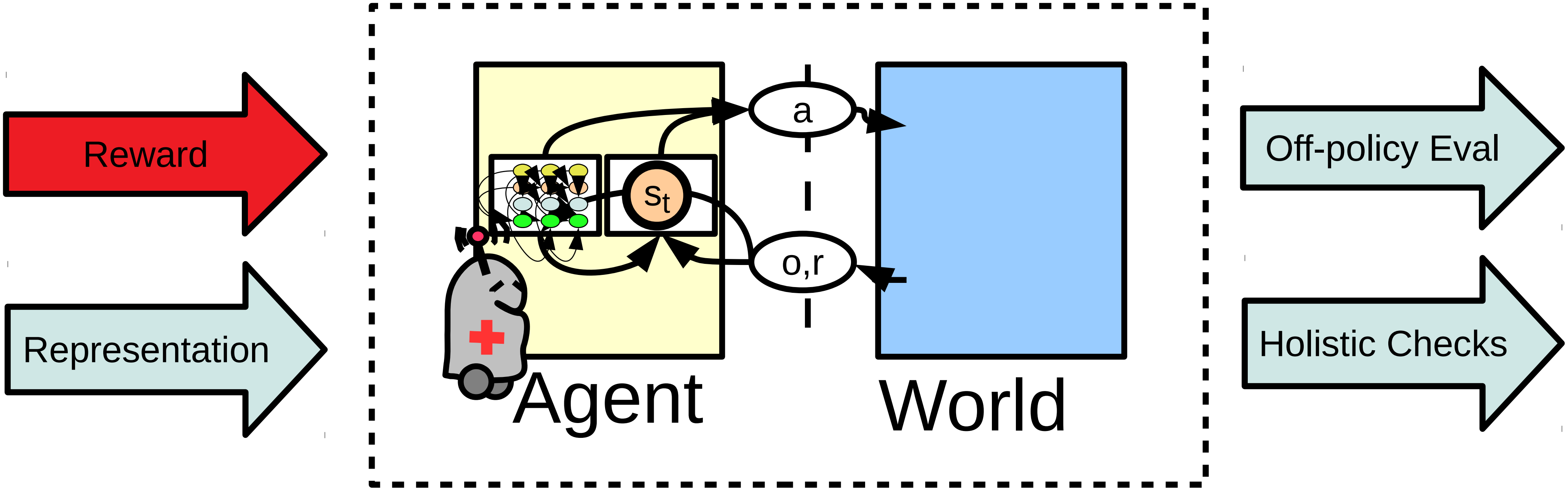


Applied to Hypotension Management

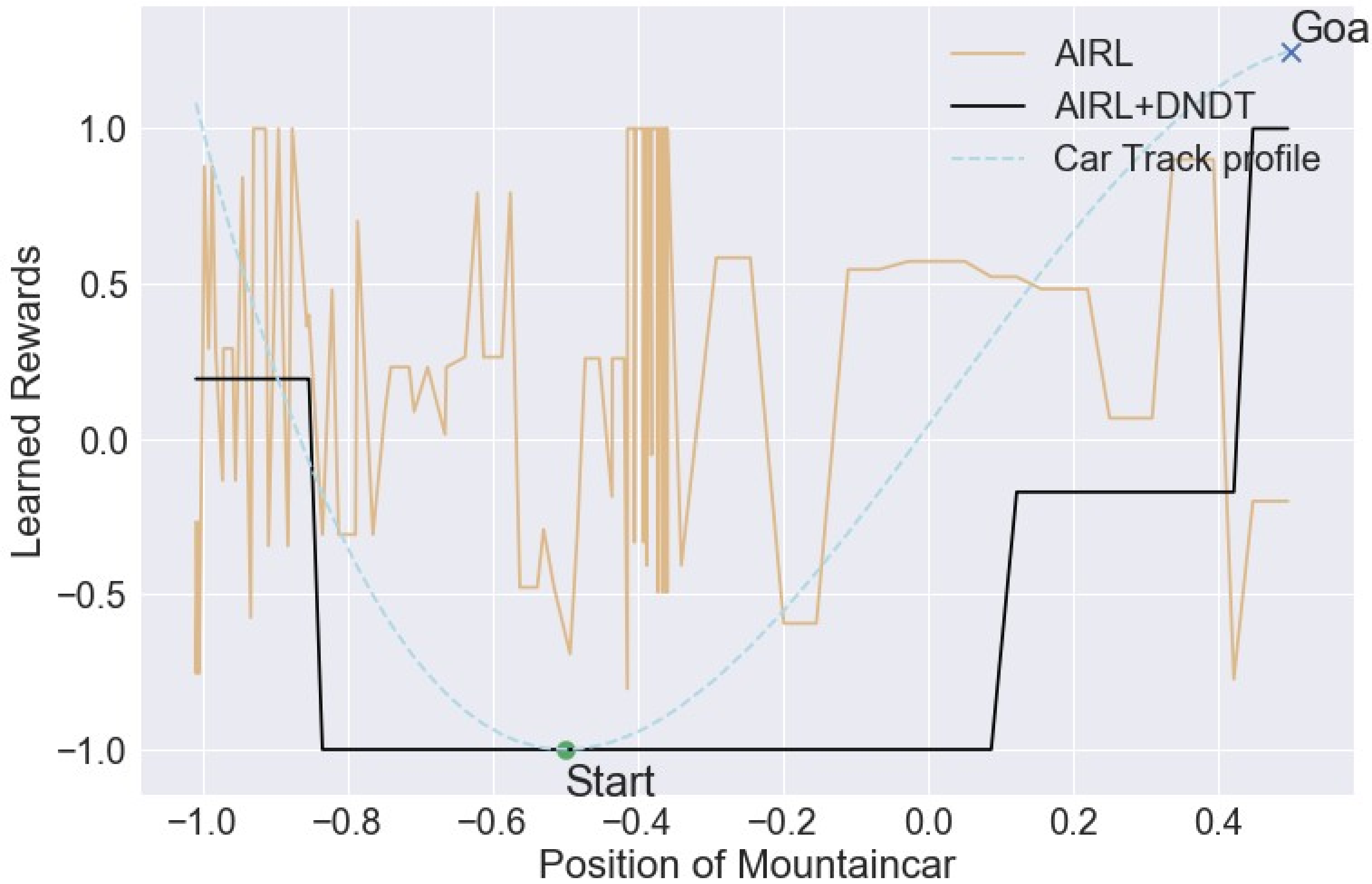
Example for a single decision point



Reward Design



In Progress: IRL to Identify Rewards



(Lee and Srinivasan et al, IJCAI 2019; Srinivasan, in submission)

Going Forward

RL in the health space is tricky, but has potential in several settings. Let's

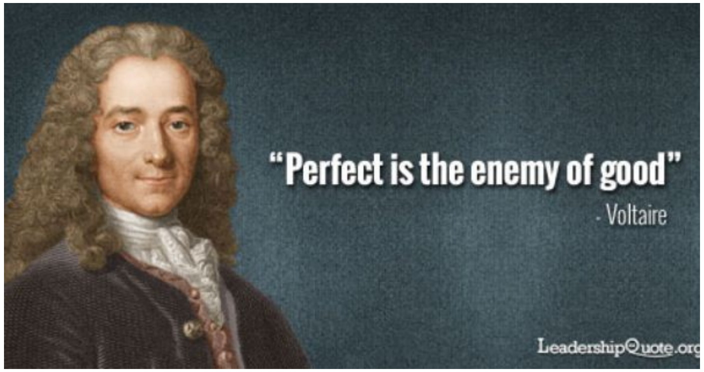
- Think holistically about how RL can provide value in a human-agent system.
- Be careful with analyses but not turn away from messy problems!

Collaborators: Sonali Parbhoo, Maurizio Zazzi, Volker Roth, Xuefeng Peng, David Wihl, Yi Ding, Omer Gottesman, Liwei Lehman, Matthieu Komorowski, Aldo Faisal, David Sontag, Fredrik Johansson, Leo Celi, Aniruddh Raghu, Yao Liu, Emma Brunskill, and the CS282 2017 Course

Going Forward

RL in the health space is tricky, but has potential in several settings. Let's

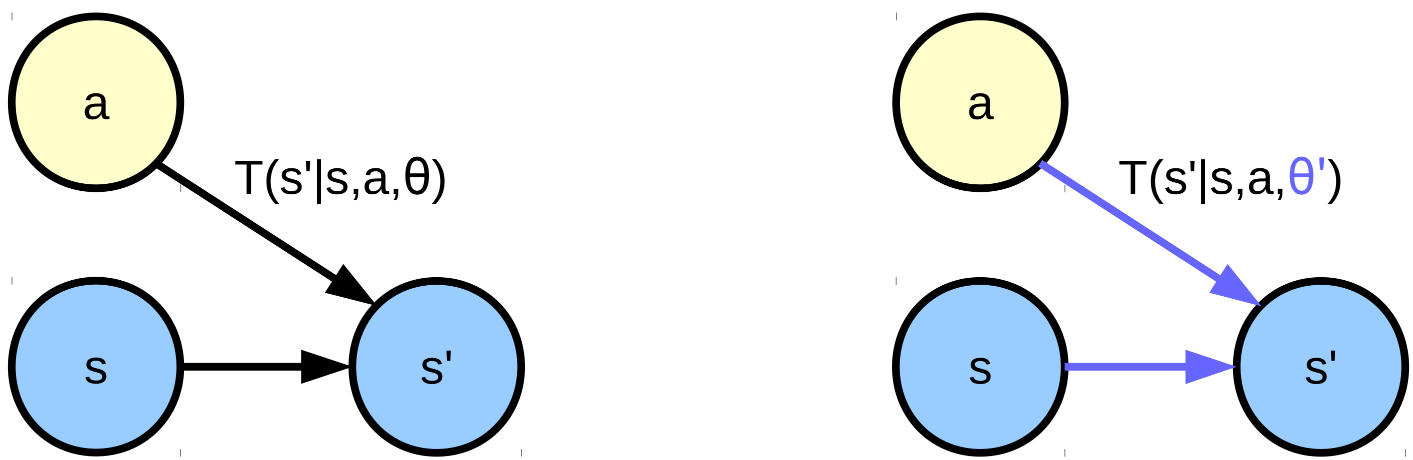
- Think holistically about how RL can provide value in a human-agent system.
- Be careful with analyses but not turn away from messy problems!



Collaborators: Sonali Parbhoo, Maurizio Zazzi, Volker Roth, Xuefeng Peng, David Wihl, Yi Ding, Omer Gottesman, Liwei Lehman, Matthieu Komorowski, Aldo Faisal, David Sontag, Fredrik Johansson, Leo Celi, Aniruddh Raghu, Yao Liu, Emma Brunskill, and the CS282 2017 Course

Modeling Improvement #2: **Personalizing** to patient dynamics

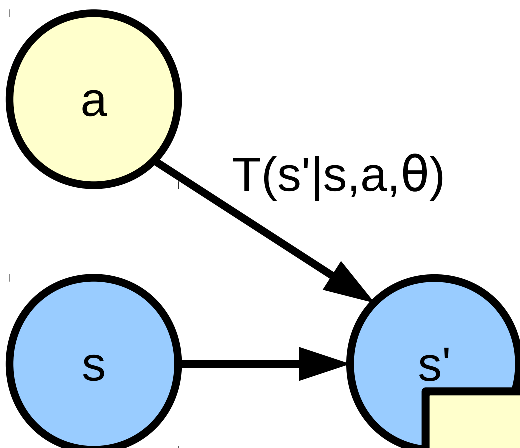
Assume that there exists some small latent vector that would allow us to personalize to the patient's dynamics (HiP-MDP).

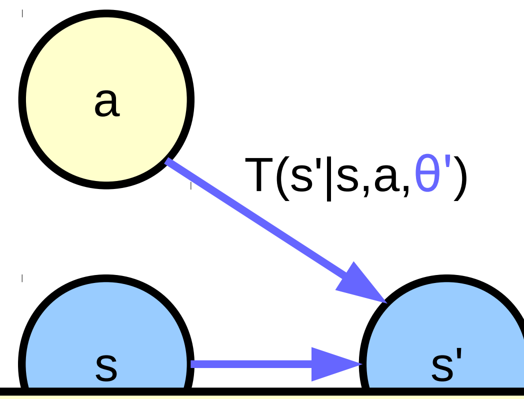


Killian et al. NIPS 2017; Yao et al. ICML LLARLA workshop 2018

Modeling Improvement #2: **Personalizing** to patient dynamics

Assume that there exists some small latent vector that would allow us to personalize to the patient's dynamics (HiP-MDP).





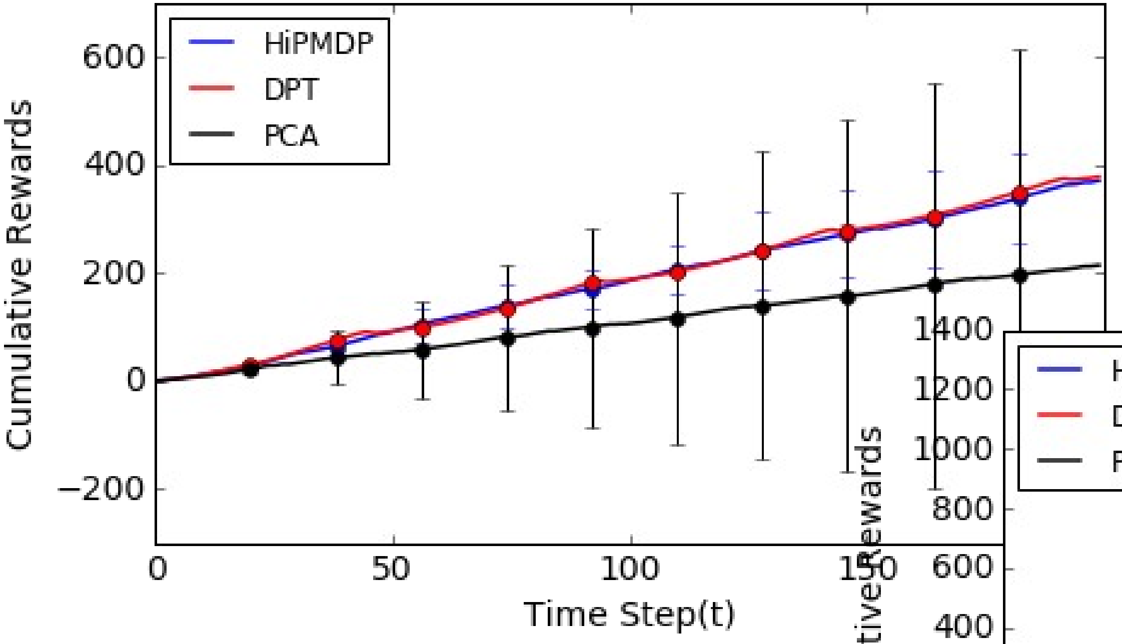
Consider two planning approaches:

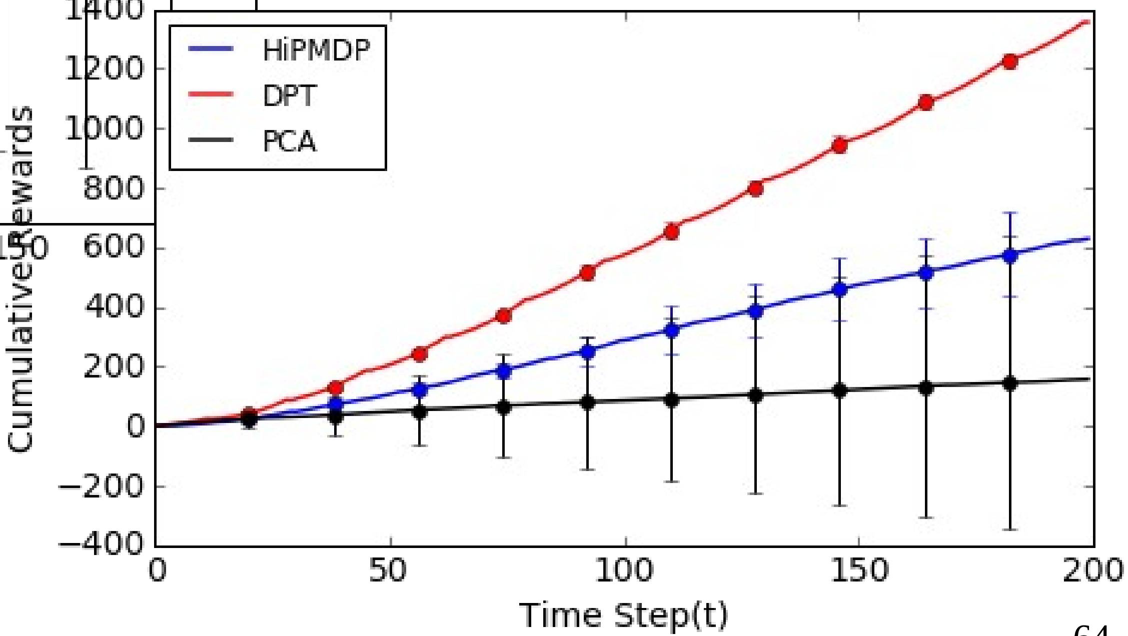
1. Plan given $T(s'|s,a,\theta)$
2. Directly learn a policy $a = \pi(s,\theta)$

Killian et al. NIPS 2017; Yao et al.

Modeling Improvement #2: **Personalizing** to patient dynamics

Results with a (simple) HIV simulator

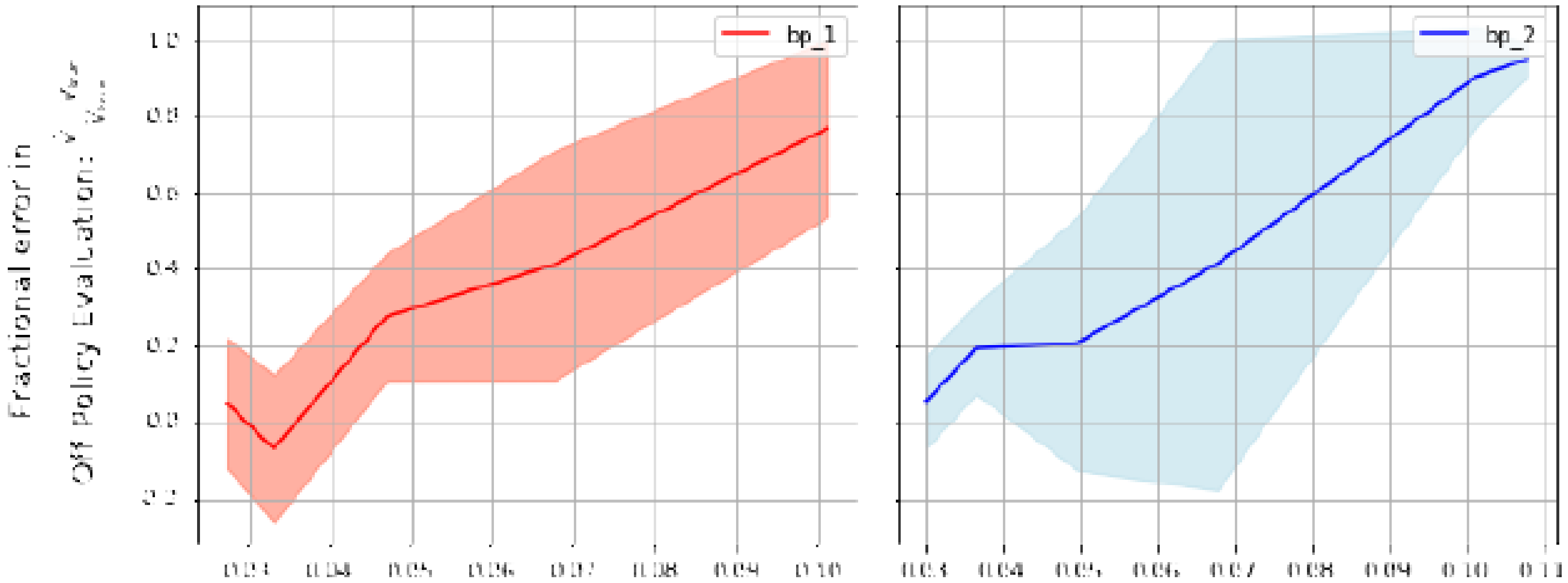




Killian et al. NIPS 2017; Yao et al. ICML LLARLA workshop 2018

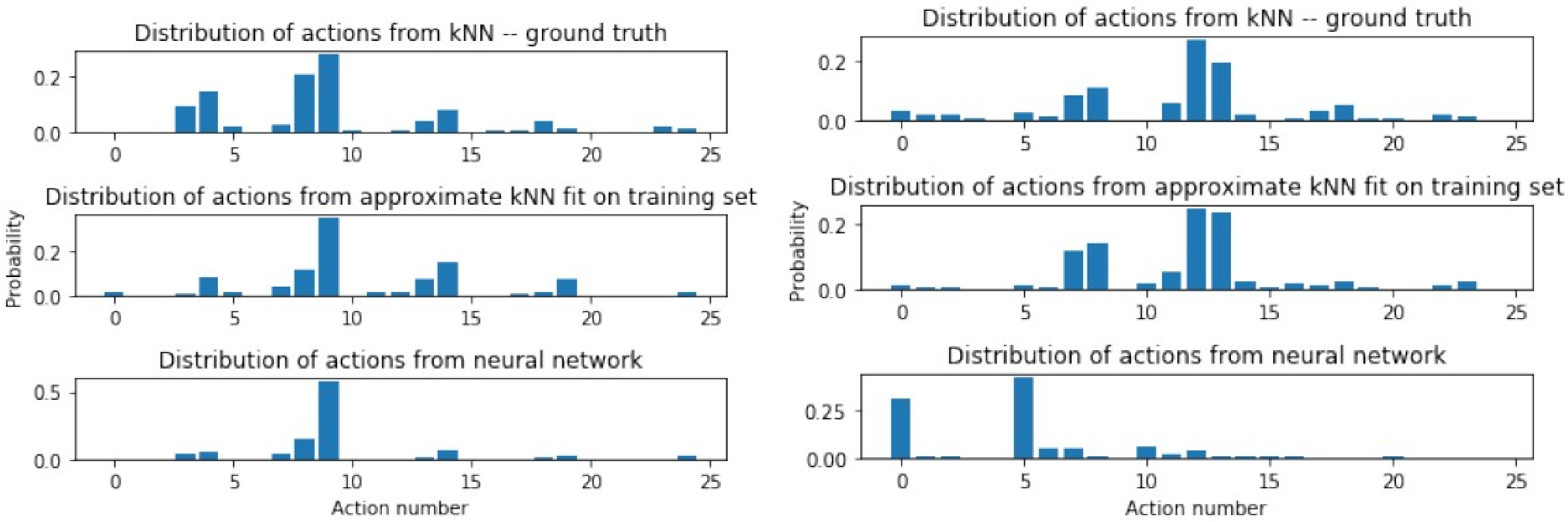
Off-policy Evaluation Challenges: Sensitive to Algorithm Choices

$$\mathrm{WDR}(D) := \sum_{i=1}^{I}\sum_{t=0}^{T}\gamma^t w_i^t r_t^{H_i} - \sum_{i=1}^{I}\sum_{t=0}^{T}\gamma^t\left(w_t^i \hat{Q}^{\pi_e}(S_t^{H_i}, A_t^{H_i}) - w_{t-1}^i \hat{V}^{\pi_e}(S_t^{H_i})\right)$$



Off-policy Evaluation Challenges: Sensitive to Algorithm Choices

$$\mathrm{WDR}(D) := \sum_{i=1}^{I}\sum_{t=0}^{T}\gamma^t w_i^t r_t^{H_i} - \sum_{i=1}^{I}\sum_{t=0}^{T}\gamma^t \left(w_t^i \hat{Q}^{\pi_e}(S_t^{H_i}, A_t^{H_i}) - w_{t-1}^i \hat{V}^{\pi_e}(S_t^{H_i})\right)$$



(a) Overconfident predictions

(b) Incorrect predictions

Sepsis: Neural networks definitely not calibrated...

Off-policy Evaluation Challenges: Sensitive to Algorithm Choices

$$\mathrm{WDR}(D) := \sum_{i=1}^{I}\sum_{t=0}^{T}\gamma^t w_i^t r_t^{H_i} - \sum_{i=1}^{I}\sum_{t=0}^{T}\gamma^t (w_t^i \hat{Q}^{\pi_e}(S_t^{H_i}, A_t^{H_i}) - w_{t-1}^i \hat{V}^{\pi_e}(S_t^{H_i}))$$

kNN is more calibrated

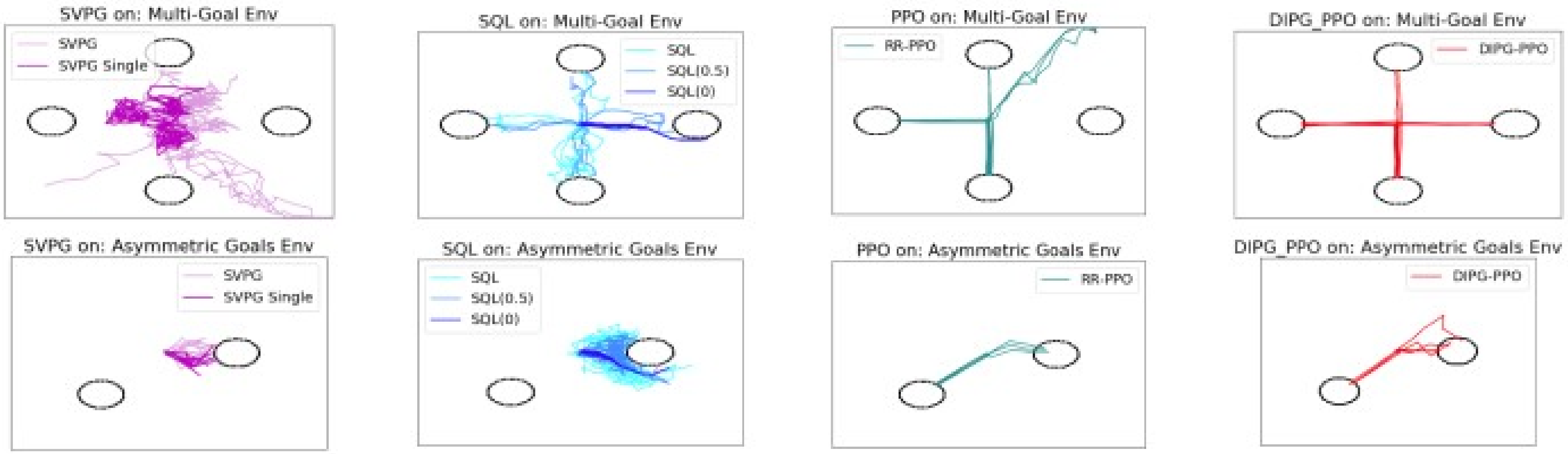
Severity	LR	RF	NN	Approx kNN
0 - 4	0.249	0.214	0.213	**0.129**
5 - 9	0.269	0.254	0.246	**0.152**
10 - 13	0.309	0.309	0.399	**0.210**
14 - 23	0.356	0.337	0.426	**0.199**

Calibration helps

Behaviour Policy Model	**MDP Approximate Model**	**MSE**
Approximate kNN	Fitted Q Iteration	**3.05**
Approximate kNN	Kernel-based RL	6.54
Approximate kNN	Discrete SARSA	6.53
Neural network	Fitted Q Iteration	3.53
Neural network	Kernel-based RL	10.2

In Progress: Displaying Diverse Alternatives

If policies can't be statistically differentiated, give plausible alternatives.



(Masood et al, IJCAI 2019)

SODA-RL Applied to Hypotension Management

Quantitative Results: Safety, quality are important to consider

	Setting			Quantitative Metrics						
	Diversity Weight	Safety Mask?	Quality Term	# Kept Agents	CWPDIS Value	CE w/ Beh. Actions	SymKL w/ Beh. Action Probabilities	ESS	SymKL btw pairs of agents	# Times Agents Allowed Unseen Actions
Diverse and Safe	High	Yes	CE	3	34.25 ± 0.07	1.03 ± 0.04	0.58 ± 0.06	352.2 ± 94.5	1.95 ± 0.21	0 ± 0
	High	Yes	SymKL	3	35.43 ± 1.45	1.13 ± 0.11	0.62 ± 0.13	221.5 ± 102.4	2.05 ± 0.23	0 ± 0
	Low	Yes	CE	0	-	-	-	-	-	-
	Low	Yes	SymKL	4	36.70 ± 0.10	0.52 ± 0.00	0.06 ± 0.00	282.9 ± 30.8	0.00 ± 0.00	0 ± 0
	High	Yes	None	4	35.86 ± 1.51	2.44 ± 0.65	1.39 ± 0.47	310.7 ± 180.9	3.27 ± 0.00	0 ± 0
Diverse, not Safe	High	No	CE	0	-	-	-	-	-	-
	High	No	SymKL	2	41.74 ± 0.36	1.14 ± 0.15	0.92 ± 0.32	234.7 ± 146.1	2.90 ± 0.00	29230 ± 12387
	Low	No	CE	0	-	-	-	-	-	-
	Low	No	SymKL	0	-	-	-	-	-	-
	High	No	None	0	-	-	-	-	-	-
Safe, not Diverse	None	Yes	CE	4	38.29 ± 0.32	0.52 ± 0.00	0.08 ± 0.00	96.1 ± 18.8	0.01 ± 0.00	0 ± 0
	None	Yes	SymKL	4	36.74 ± 0.08	0.52 ± 0.00	0.06 ± 0.00	284.1 ± 27.2	0.00 ± 0.00	0 ± 0
Not Safe or Diverse	None	No	CE	0	-	-	-	-	-	-
	None	No	SymKL	0	-	-	-	-	-	-