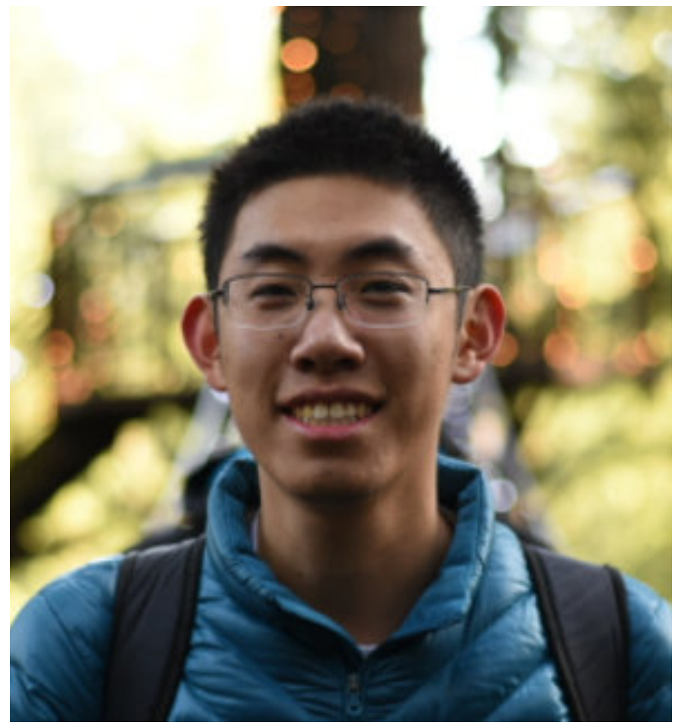
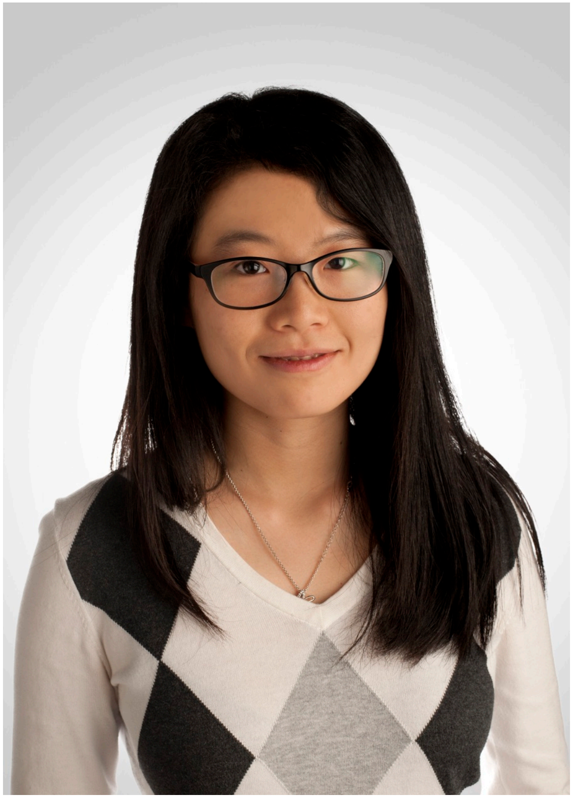
Upper Confidence Reinforcement Learning with Value Targeted Regression

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin F. Yang

Co-Authors





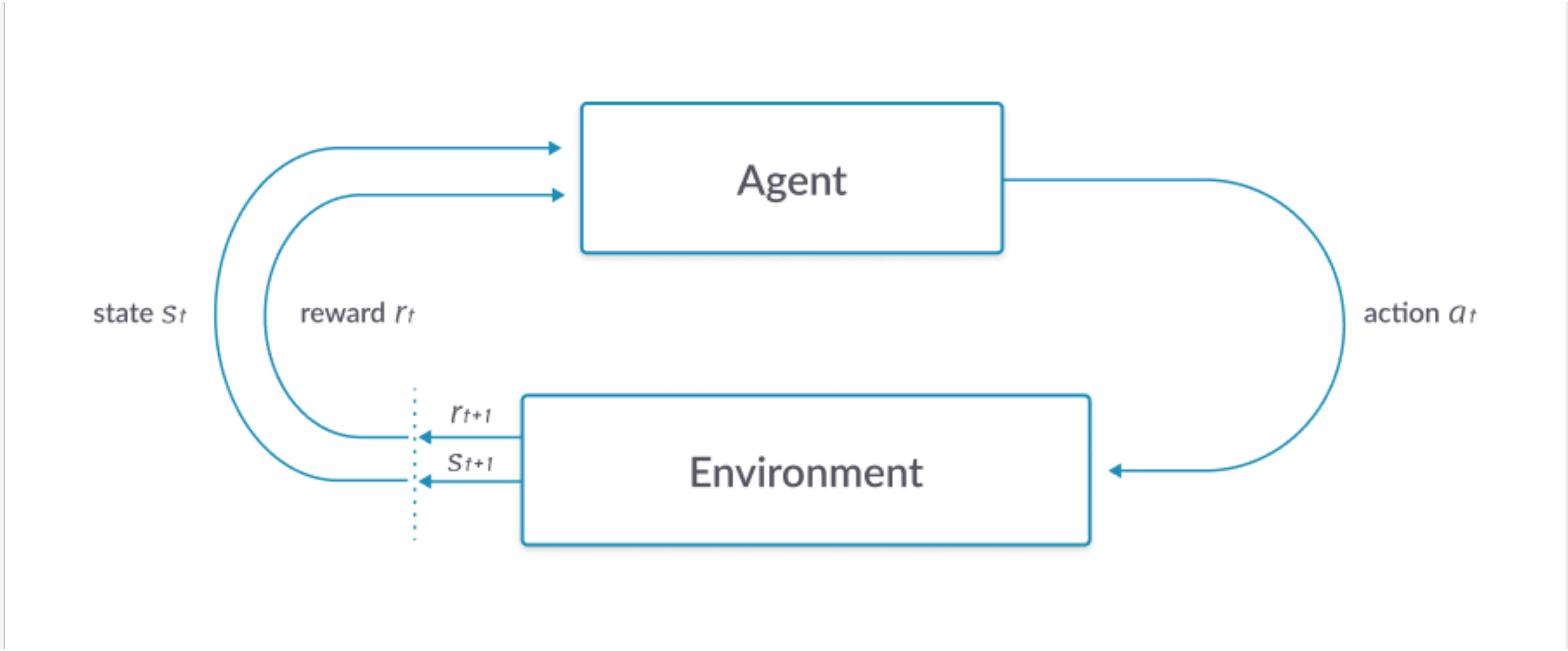
Zeyu Jia
Peking University

Csaba Szepesvari
University of Alberta/Deepmind

Mengdi Wang
Princeton/Deepmind

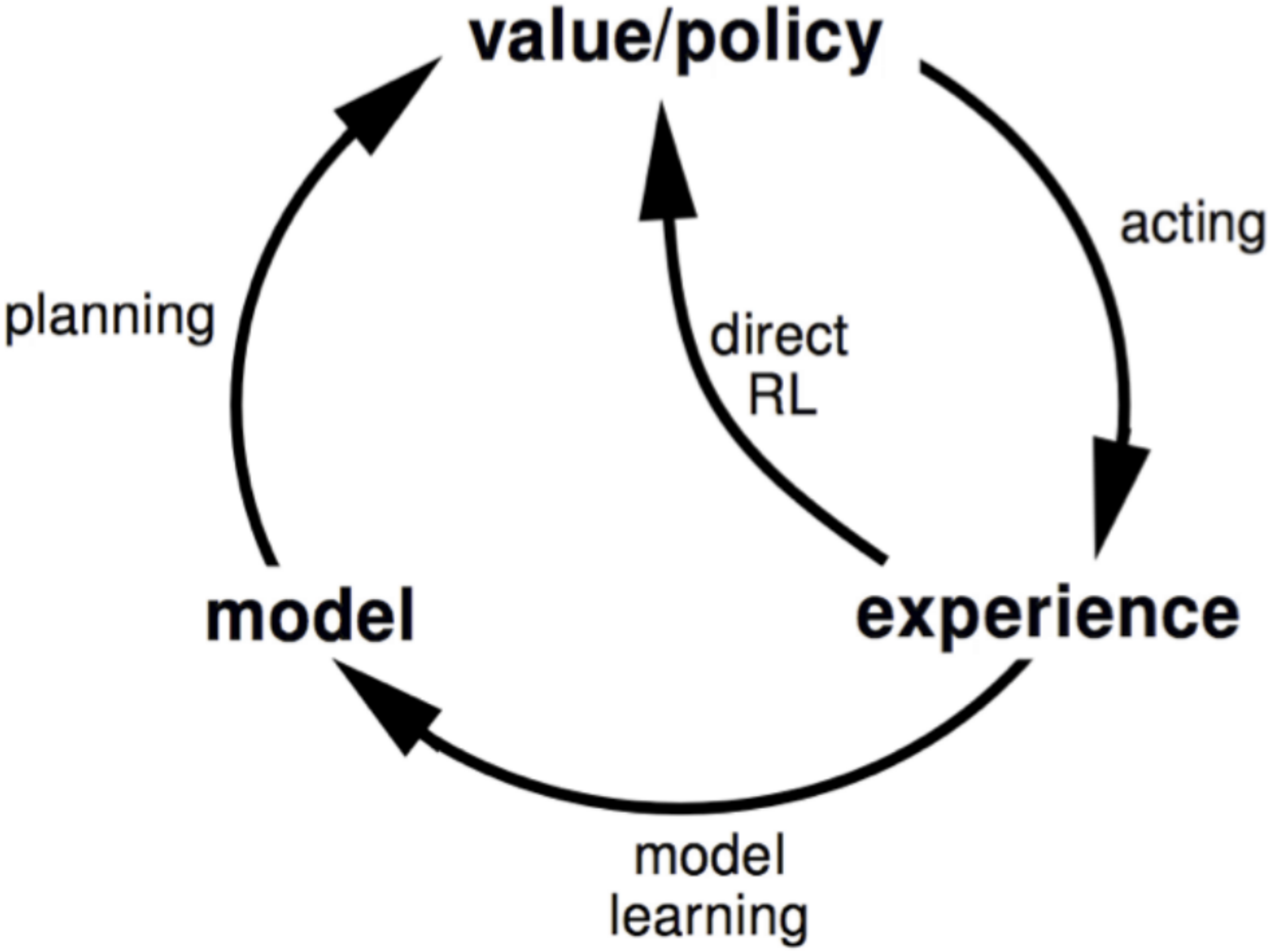
Lin F. Yang
UCLA

The Reinforcement Learning (RL) Problem



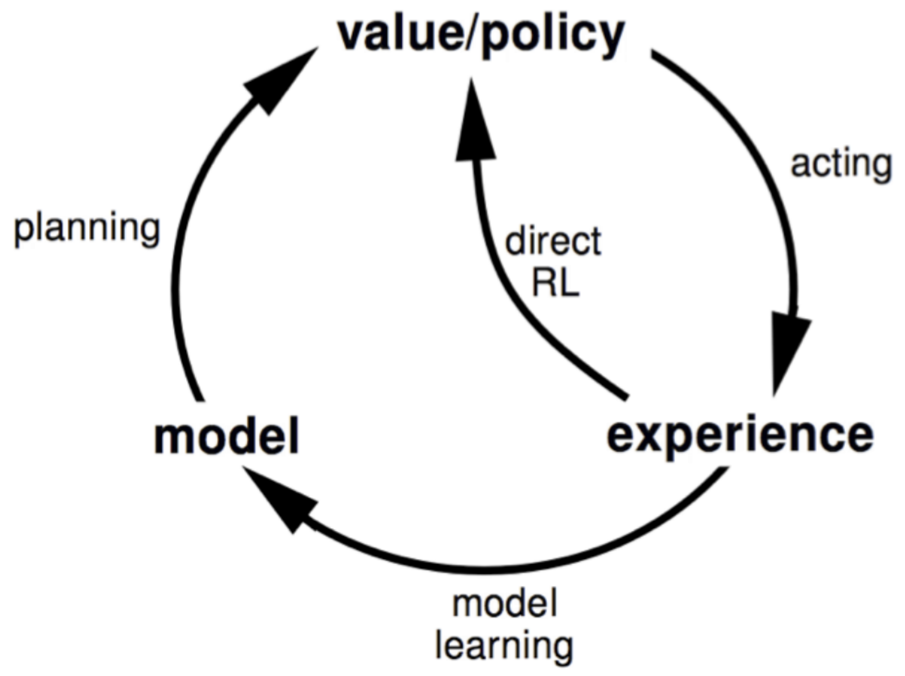
Model-Based RL (MBRL)

- We fit a model to $(s_t, a_t, s_{t+1}, r_{t+1})$

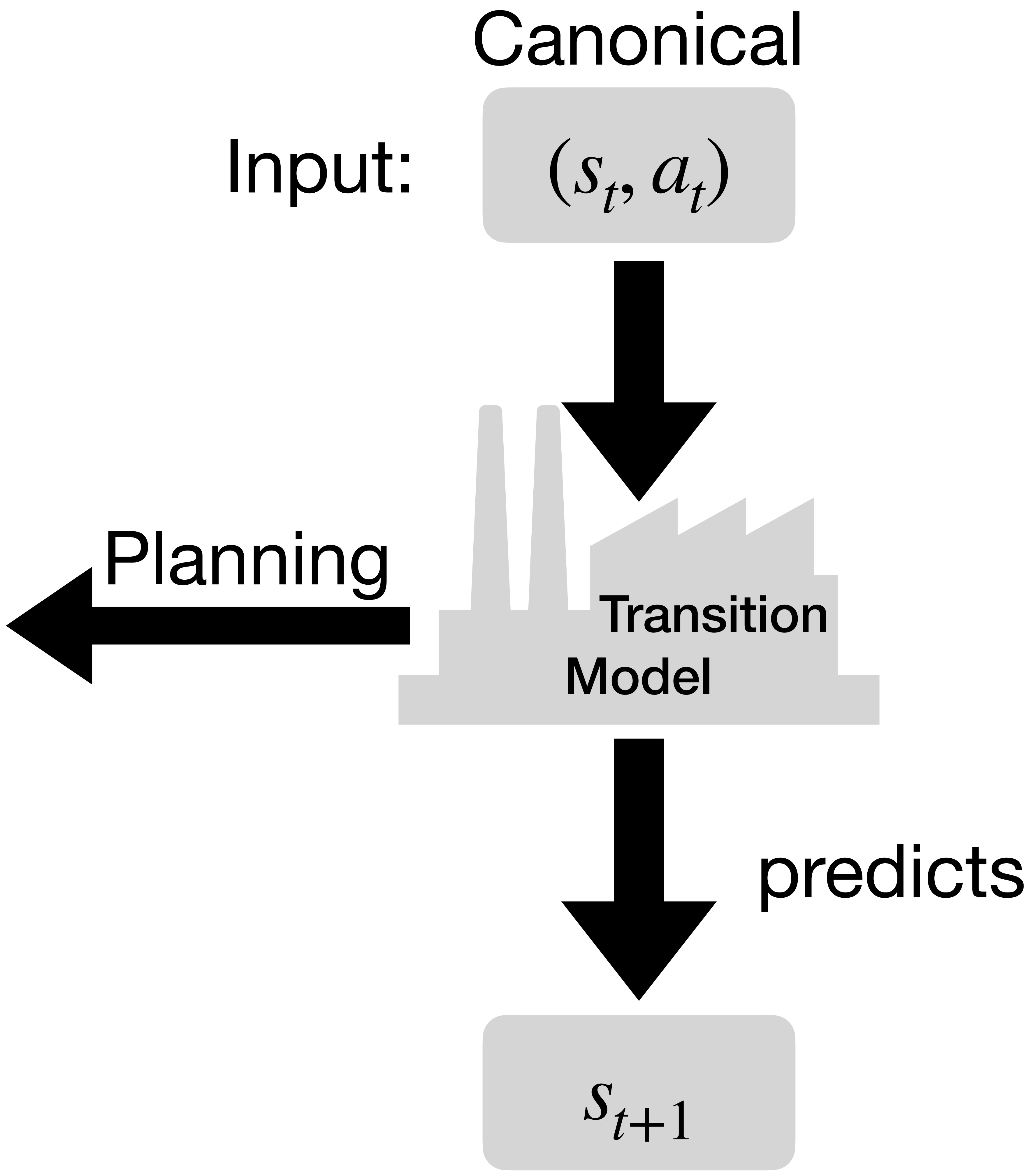


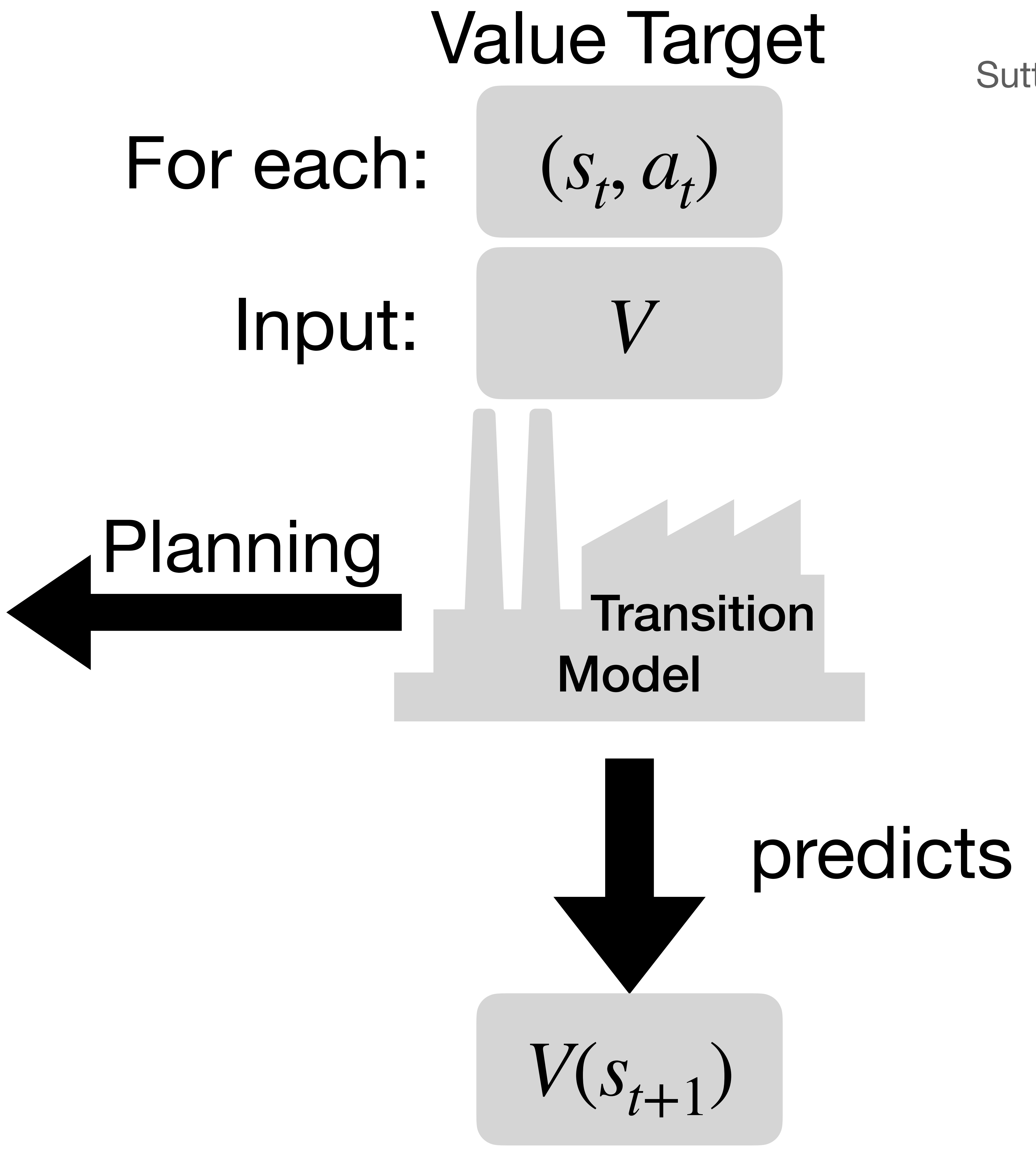
Sutton and Barto (2018)

How to Learn a Transition Model ?



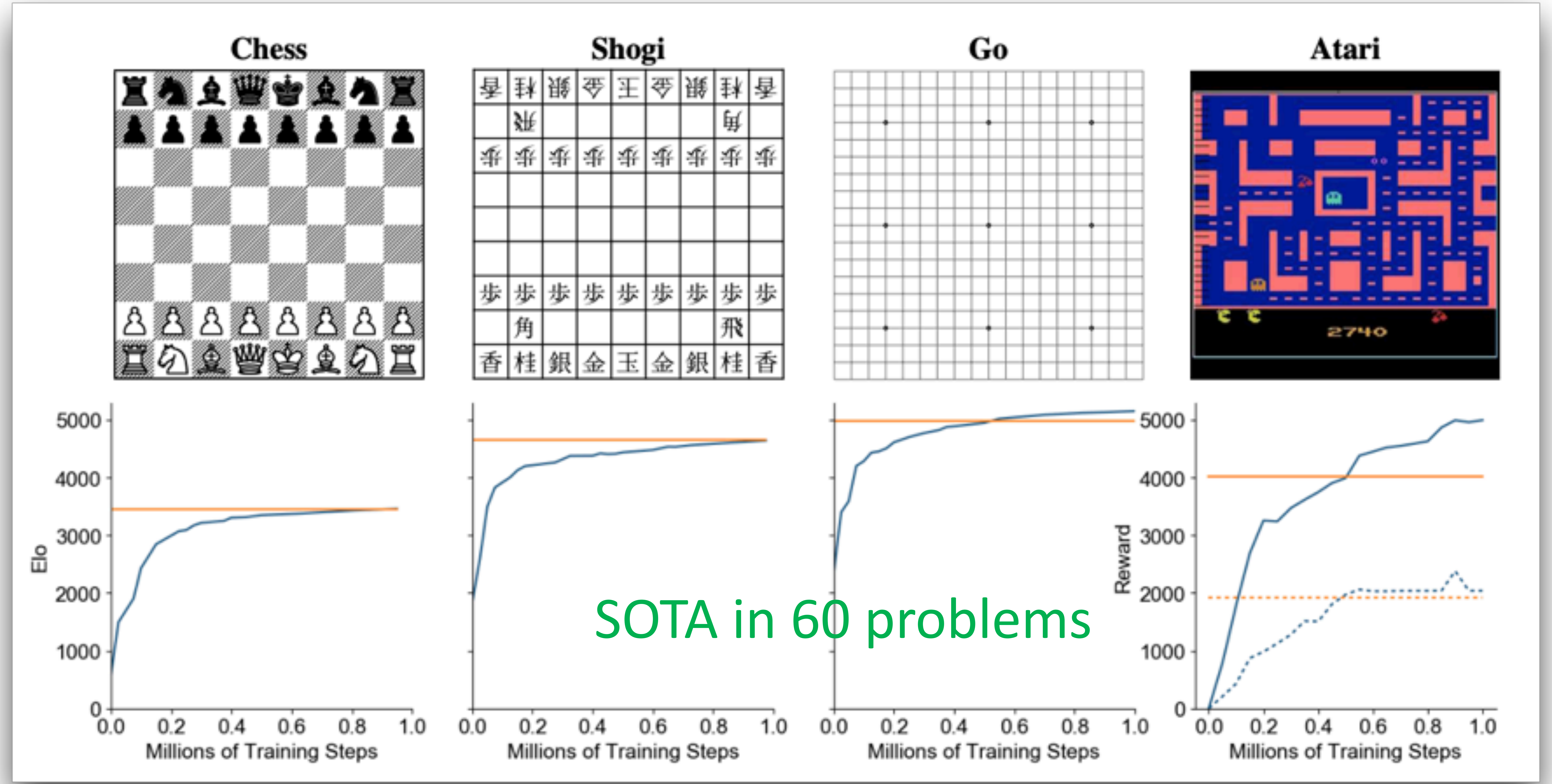
Sutton and Barto (2018)





Using neural nets: MuZero!

Objective:
Keep $||(P - \tilde{P})\tilde{V}||_{n,2}^2$ small



SOTA in 60 problems

Is value-targeted learning sufficient and efficient for model-based online RL?

Episodic Markov Decision Process (MDP)

- MDP: $M = (\mathcal{S}, \mathcal{A}, P, r, H, s_o)$
- The Value Function of policy π is defined: $V_h^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{i=h}^{H} r(s_i, \pi(s_i)) \,|\, s_h = s\right]$
- The goal is to minimize the total regret:
  - $R(T) = \sum_{k=1}^{K} V_1^{*}(s_1^k) - \sum_{k=1}^{K}\sum_{h=1}^{H} r\left(s_h^k, a_h^k\right)$, where $T = KH$.

Assumptions about our Problem Setting

- Assumption 1 (Known Transition Model Family)
  - $P \in \mathscr{P}$
  - $\mathscr{P}$ is known
- Definition 1 (Linear Mixture Models)
  - $P(s' \mid s, a) = \sum_{j=1}^{d} \theta_j P_j(s' \mid s, a)$
  - where $\theta_1, \ldots, \theta_d$ are unknown.

Model-Based Optimistic Planning

- We want $P = \arg\max_{P' \in B} V^*_{P',1}\left(s_1\right)$
- We compute the optimal policy, π^*_P, according to P.
- Then we follow π^*_P in the current episode.
- How to construct B?

Value Targeted Regression for Confidence Set Construction

- Confidence Set: $B = \{P' \in \mathscr{P} \,|\, \tilde{L}(P') \leq \tilde{\beta}\}$
    - where: $\tilde{L}(P') = \sum_{k'=1}^{k} \sum_{h=1}^{H} \left(P'(\cdot \,|\, s_h^{k'}, a_h^{k'})^\top V_{h+1,k'} - y_{h,k'}\right)^2$
    - and $y_{h,k'} = V_{h+1,k'}(s_{h+1}^{k'}),\ h \in [H]$ and $k' \in [k]$

Theoretical Analysis

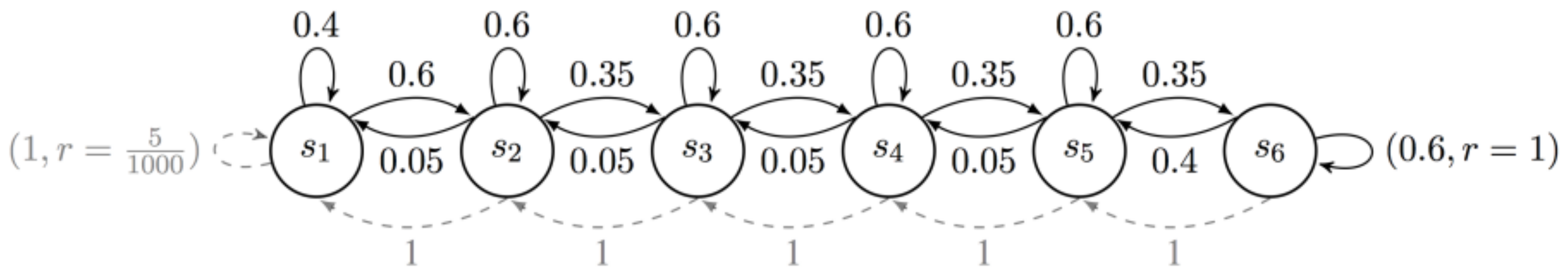
- Eluder Dimension - Length of the longest independent sequence
  - In the game “Battleship”, how long before you hit your opponents ship?
- $\mathcal{F} = \left\{ f \,\middle|\, \exists P \in \mathcal{P} \text{ s.t. for any } (s, a, v) \in \mathcal{S} \times \mathcal{A} \times \mathcal{V},\, f(s, a, v) = \int P(ds' \,|\, s, a) v(s') \right\}$.
- Regret for Known Transition Model Family (Assumption 1),
  - $R(K) \leq O\left(\text{poly}(d_E, d, H)/\sqrt{K}\right)$

Eluder Dimension

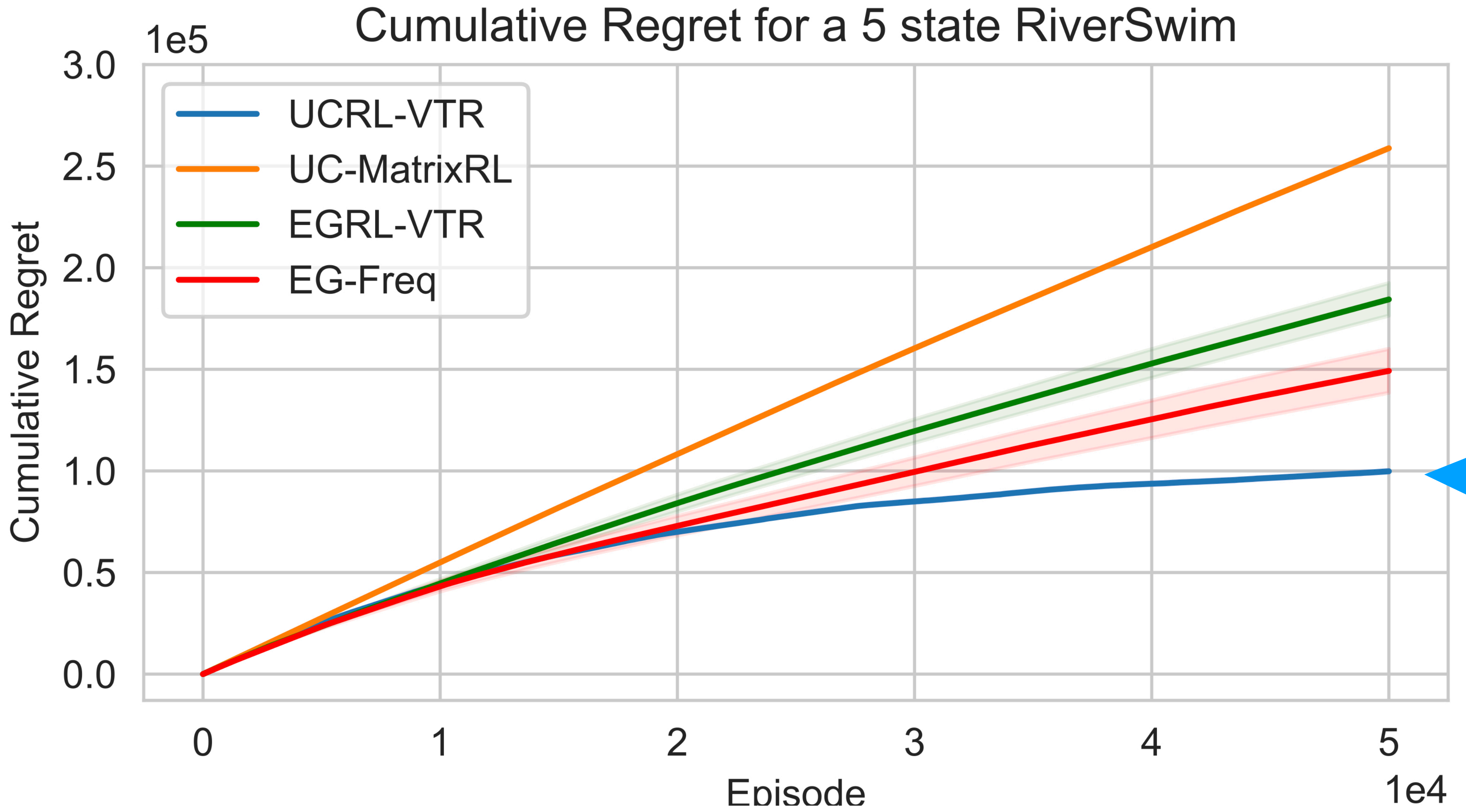
- Regret for Linearly-Parameterized Transition Model (Defn 1)
    - $R(K) = \tilde{O}\left(d\sqrt{H^3 K \log(1/\delta)}\right)$
- Lower Bound: $R(K) \geq \Omega\left(H\sqrt{dK}\right)$

Experimental Results

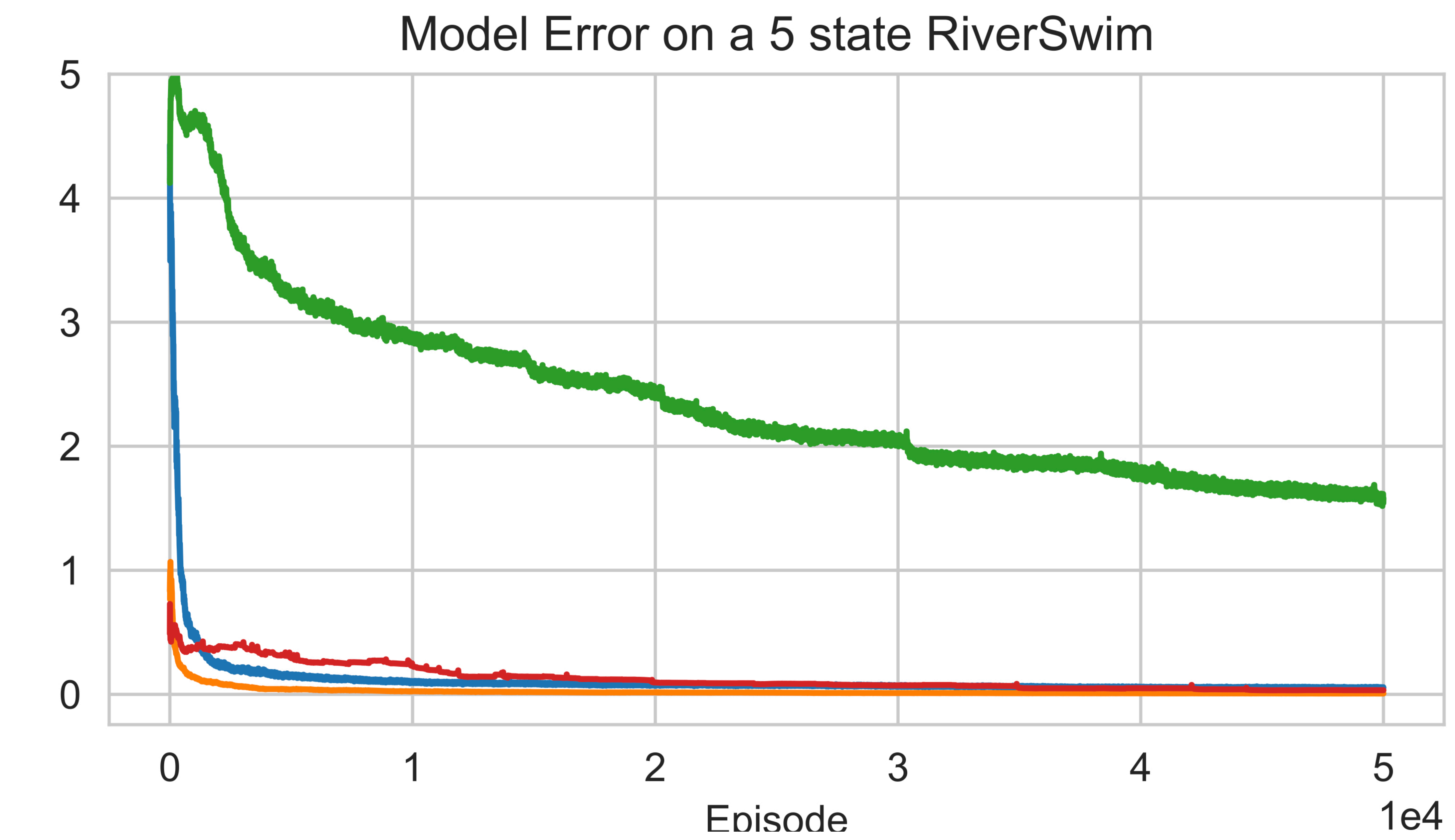
Exploration/ Targets	Optimism	Dithering
Next States	UC-MatrixRL	EG-Freq
Values	UCRL-VTR	EGRL-VTR



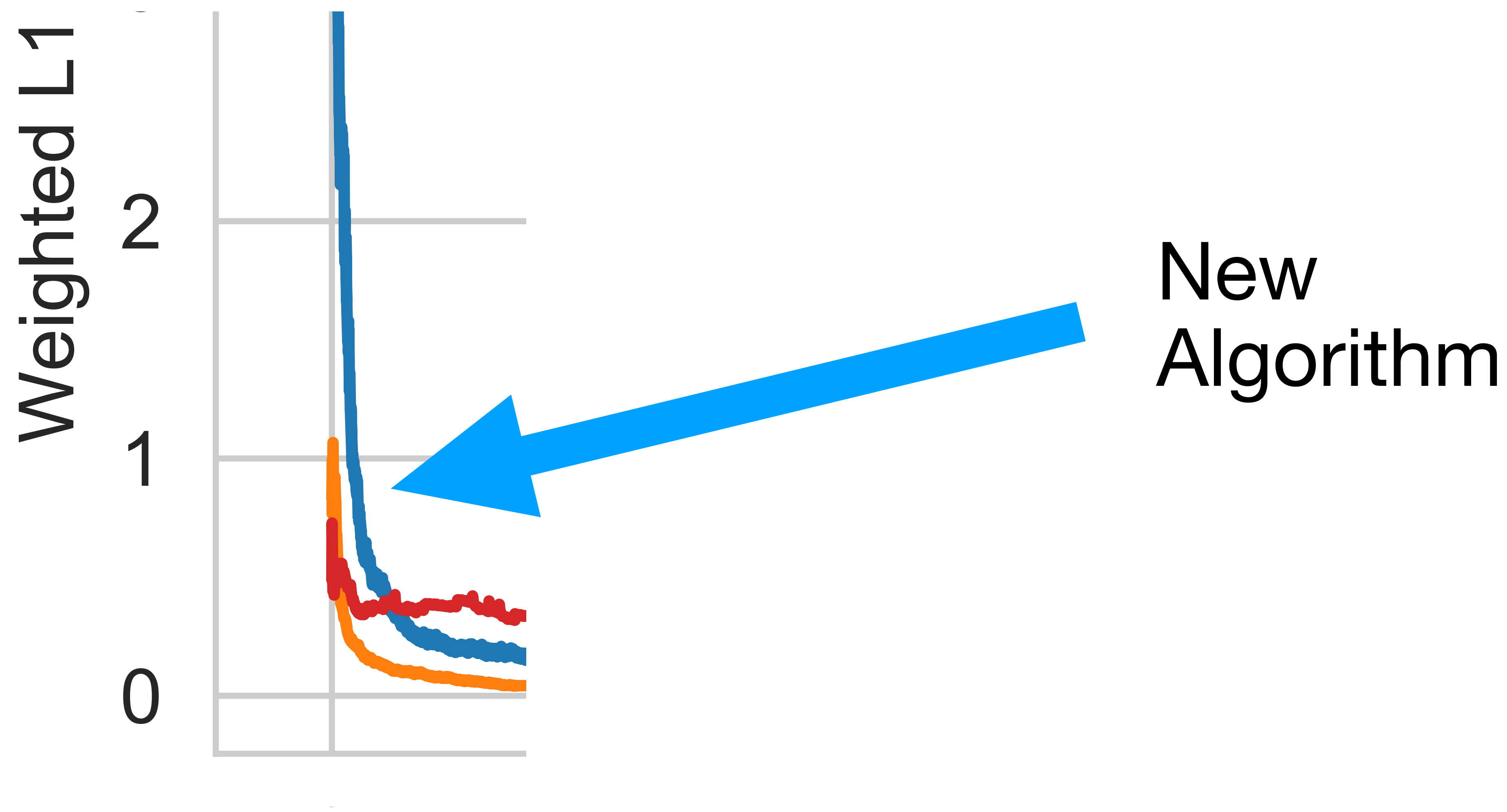
RiverSwim by Strehl & Littman

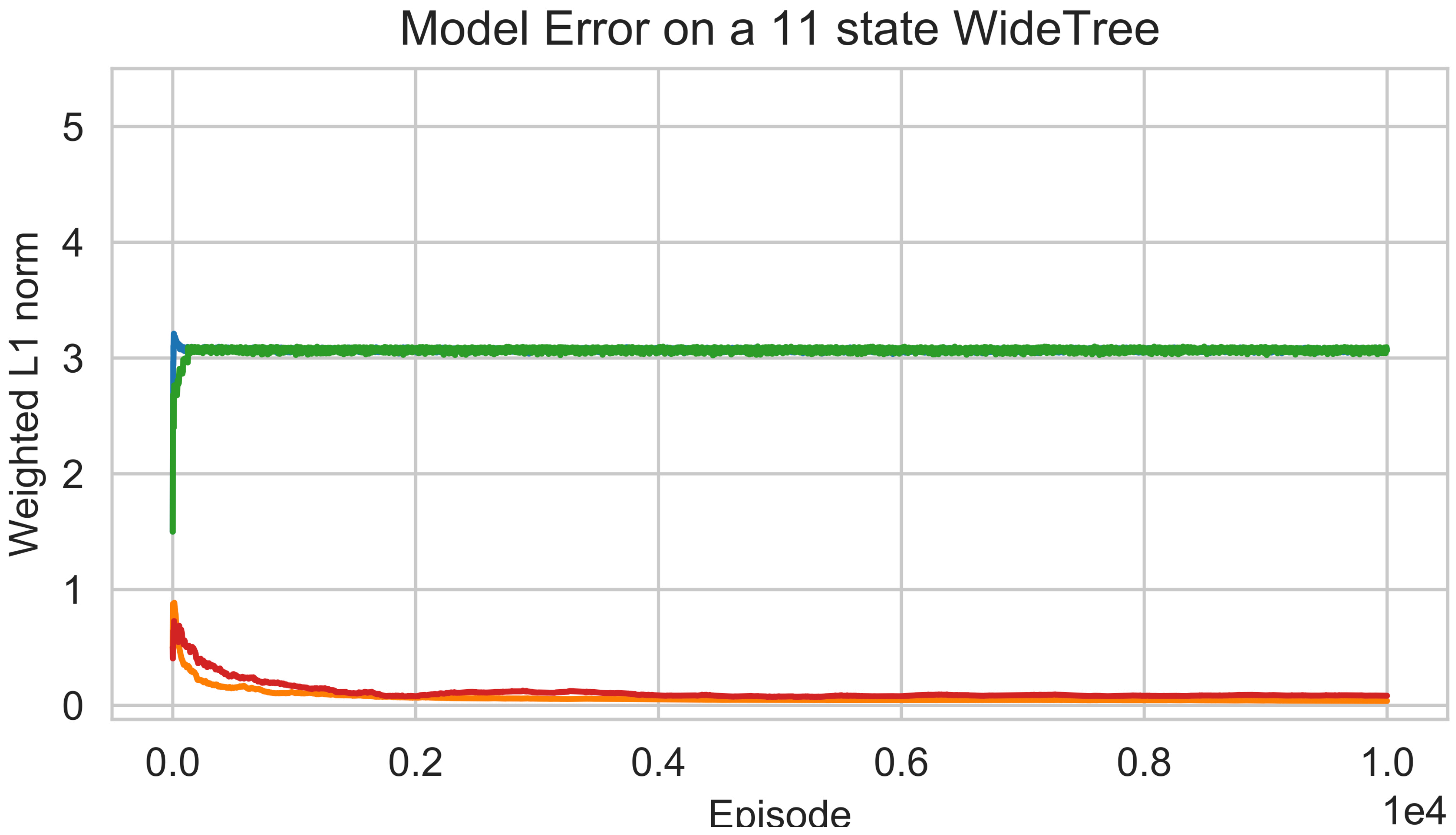


New Algorithm



Exploration/ Targets	Optimism	Dithering
Next States	UC-MatrixRL	EG-Freq
Values	UCRL-VTR	EGRL-VTR





Exploration/ Targets	Optimism	Dithering
Next States	UC-MatrixRL	EG-Freq
Values	UCRL-VTR	EGRL-VTR

Conclusions

- Value-Targeted Regression is efficient/sufficient for MBRL.
- VTR outperforms canonical transition models both theoretically and experimentally
- Computation is expensive and future work is needed to come up with computationally feasible methods to compute the VTR model.