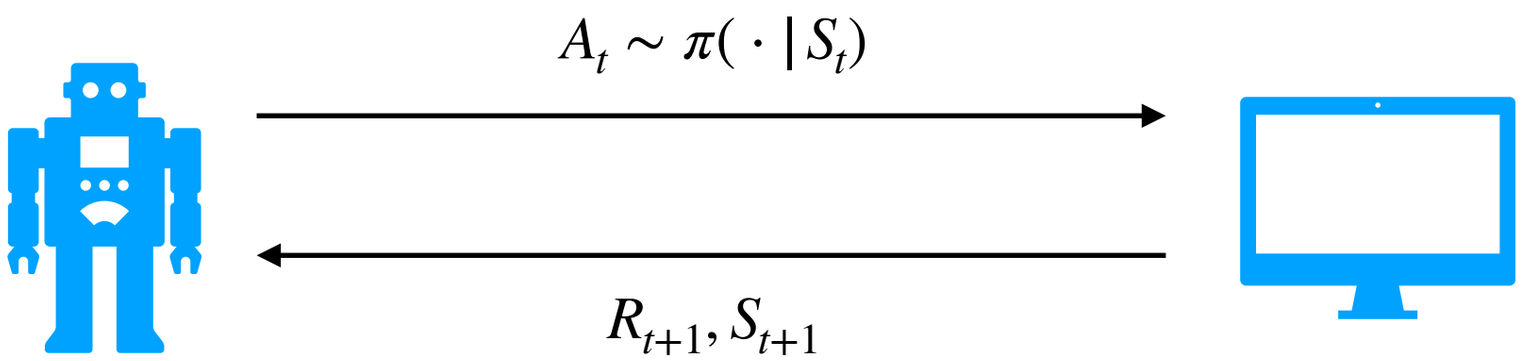
Offline Reinforcement Learning: Current and Future

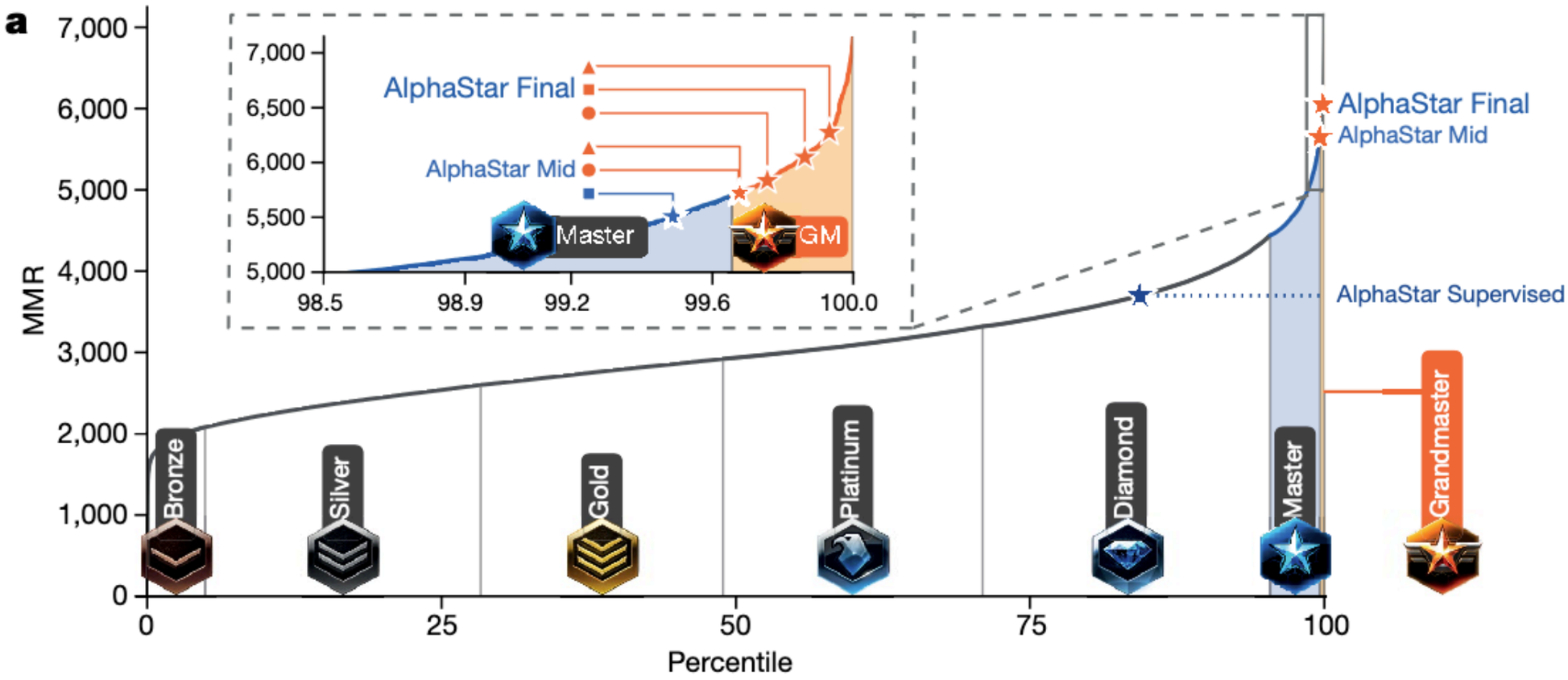
Shangtong Zhang, Assistant Professor

Department of Computer Science
University of Virginia
https://shangtongzhang.github.io/

Canonical RL relies on agent-env interaction



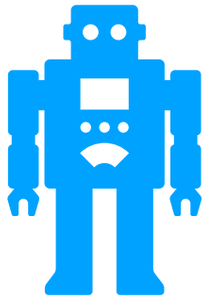
Case study: AlphaStar



(Vinyals et al. 2019)

Trillions of interactions with SCII simulator!

Online interaction can be slow

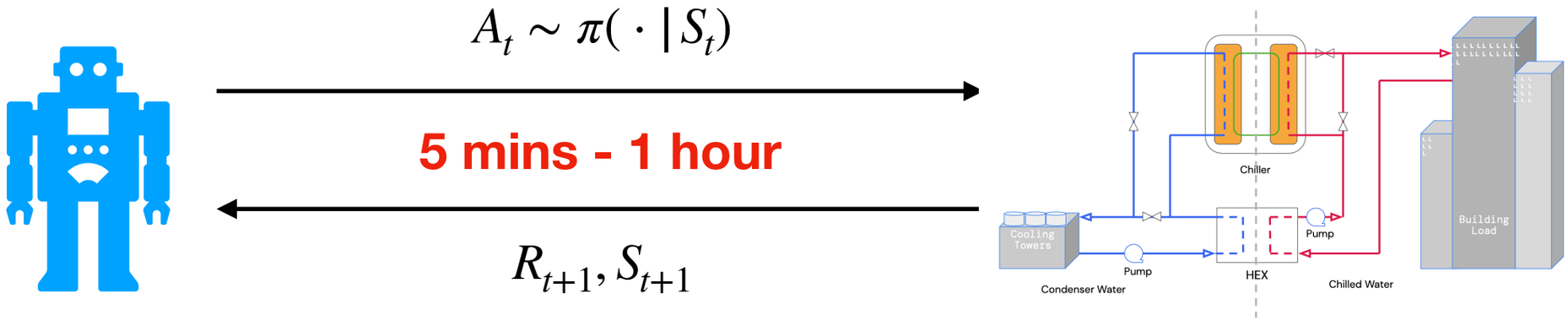


$A_t \sim \pi(\cdot \mid S_t)$

R_{t+1}, S_{t+1}

DMV

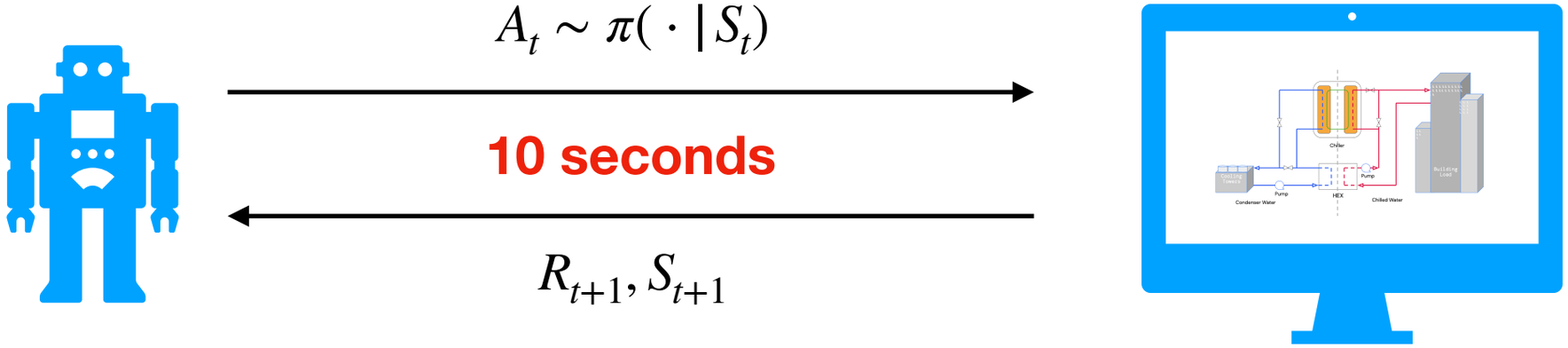
Case study: industrial cooling system



(Chervonyi et al. 2022)

1M training steps is nothing in RL

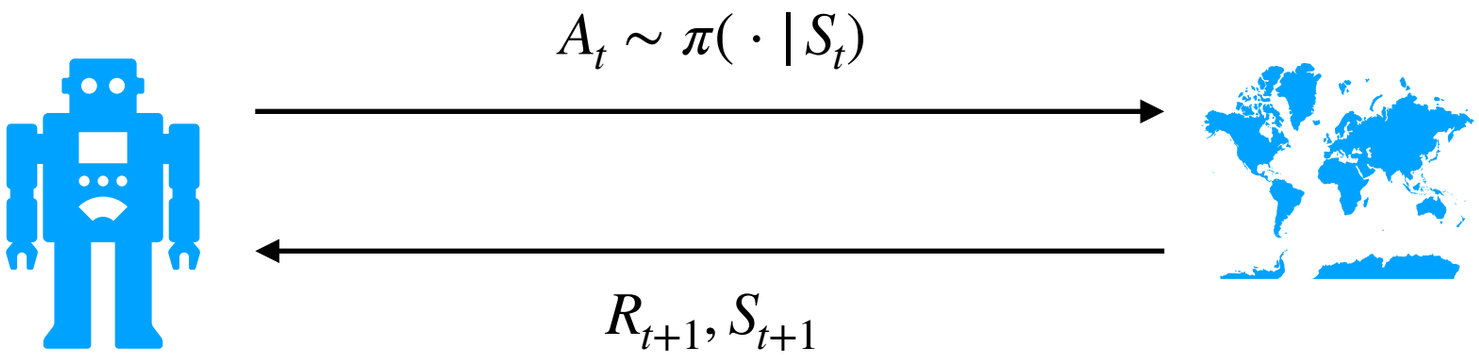
Case study: industrial cooling system



(Chervonyi et al. 2022)

$10 \times 10^6 \div 3600 \div 24 \approx 116$ **days**

Online interaction can be dangerous

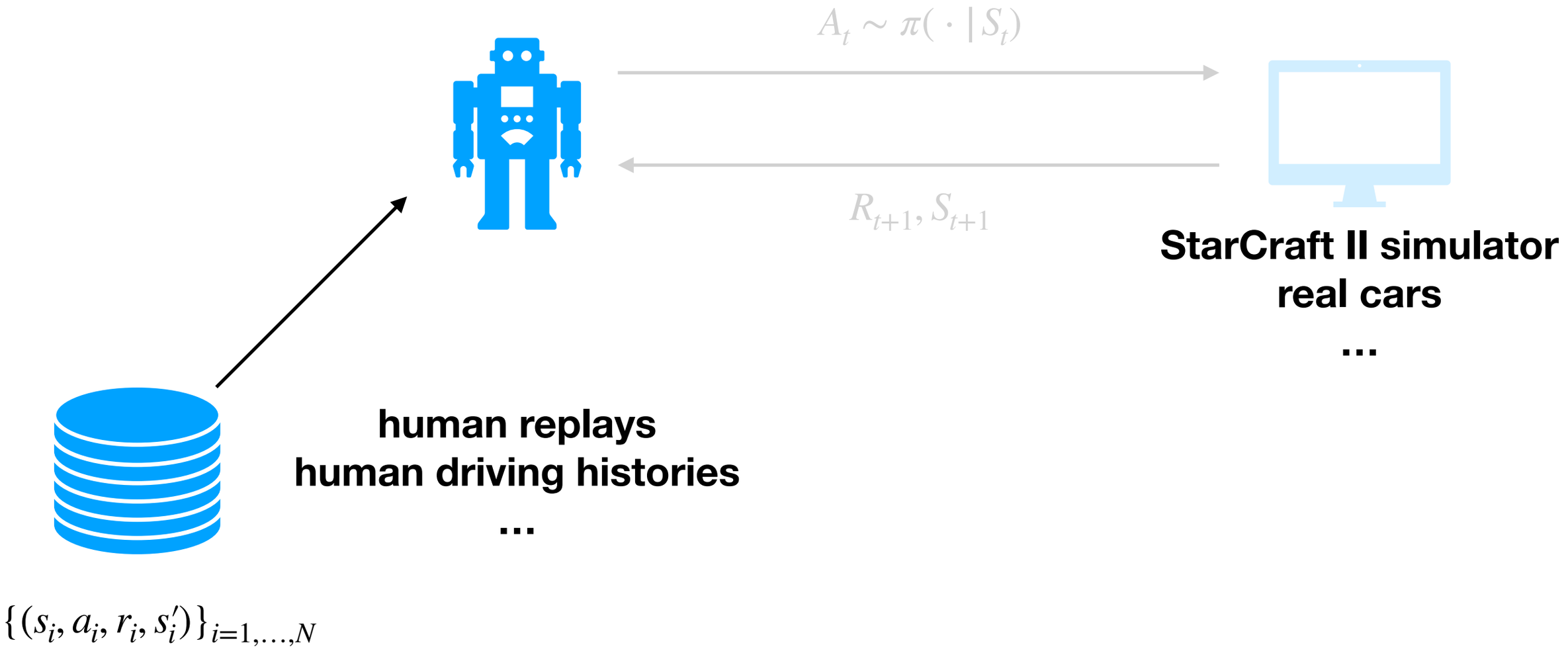


TECH NEWS

Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk

The automated car lacked "the capability to classify an object as a pedestrian unless that object was near a crosswalk," an NTSB report said.

Offline RL uses previously logged data



Case study: Offline AlphaStar

StarCraft II Unplugged: Large Scale Offline Reinforcement Learning

Michaël Mathieu* **Sherjil Ozair*** **Srivatsan Srinivasan** **Caglar Gulcehre**

Shangtong Zhang **Ray Jiang** **Tom Le Paine** **Konrad Żołna** **Richard Powell**

Julian Schrittwieser **David Choi** **Petko Georgiev** **Daniel Toyama** **Aja Huang**

Roman Ring **Igor Babuschkin** **Timo Ewalds** **Mahyar Bordbar** **Sarah Henderson**

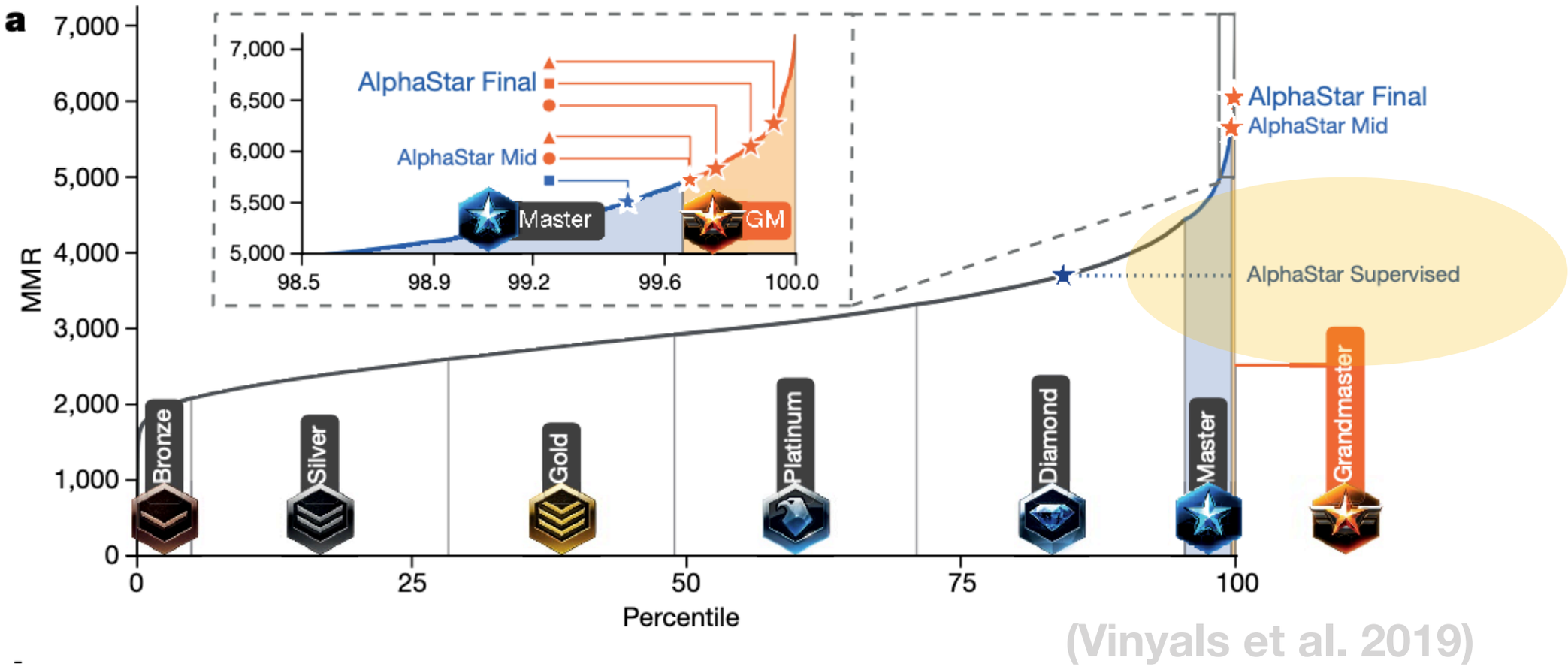
Sergio Gómez Colmenarejo **Aäron van den Oord** **Wojciech Marian Czarnecki**

Nando de Freitas **Oriol Vinyals**

DeepMind

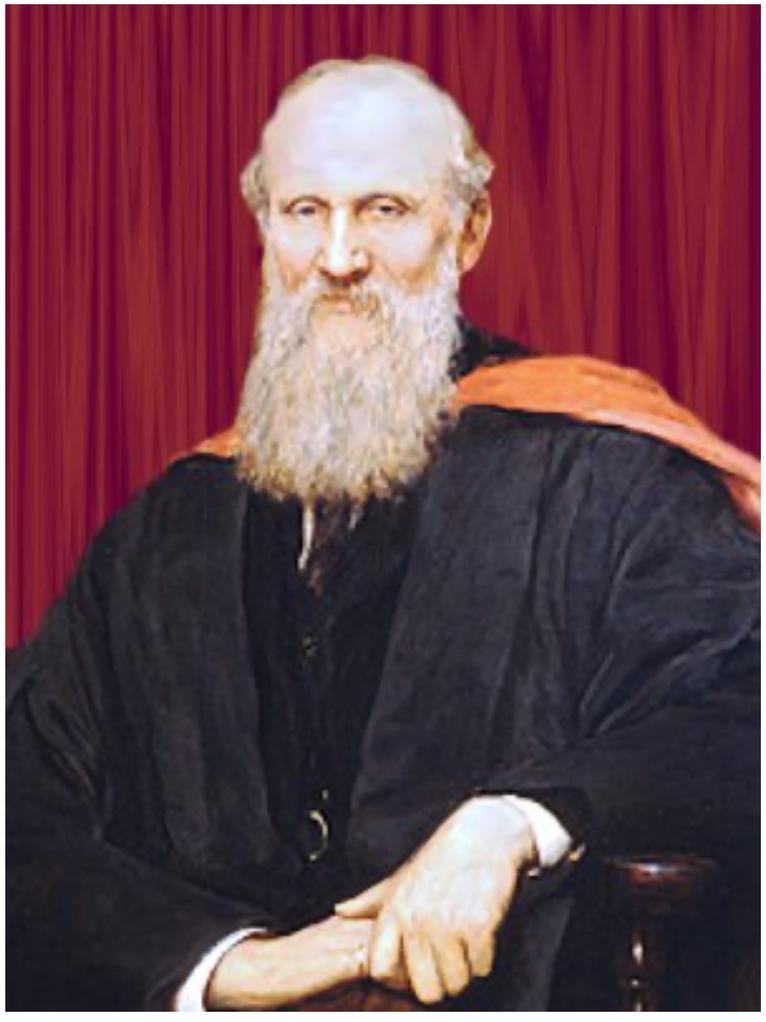
https://github.com/deepmind/alphastar
https://openreview.net/forum?id=Np8Pumfoty

Case study: Offline AlphaStar



(Vinyals et al. 2019)

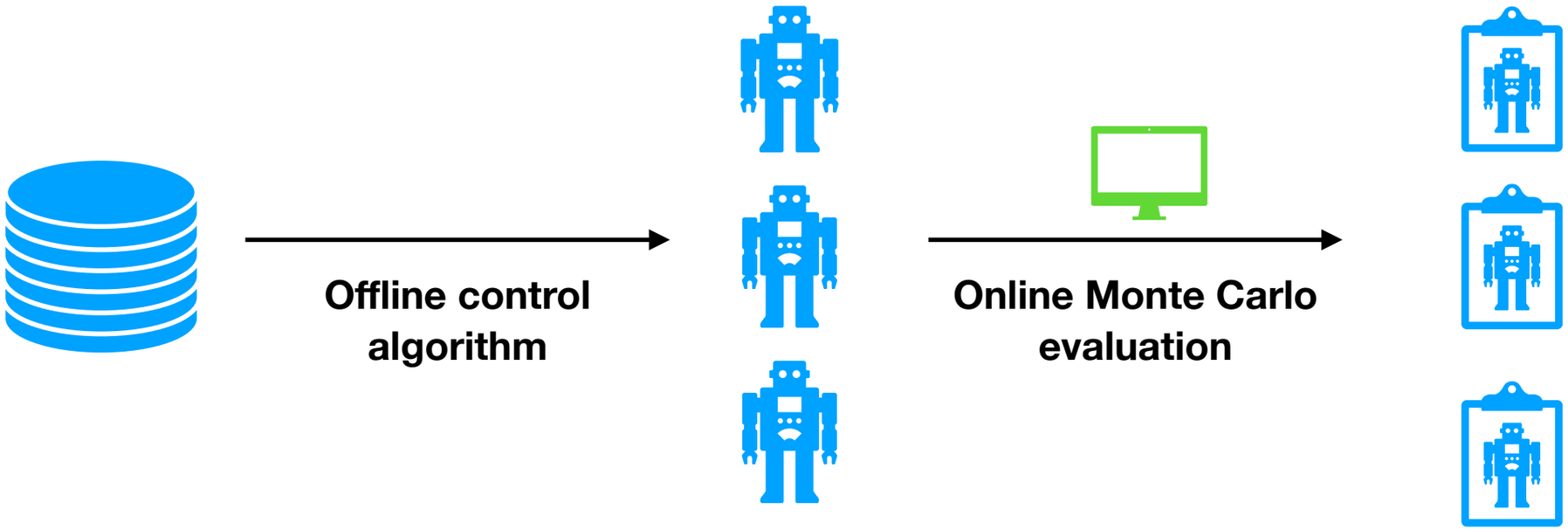
Offline AlphaStar has more than 90% win-rate against AlphaStar Supervised.



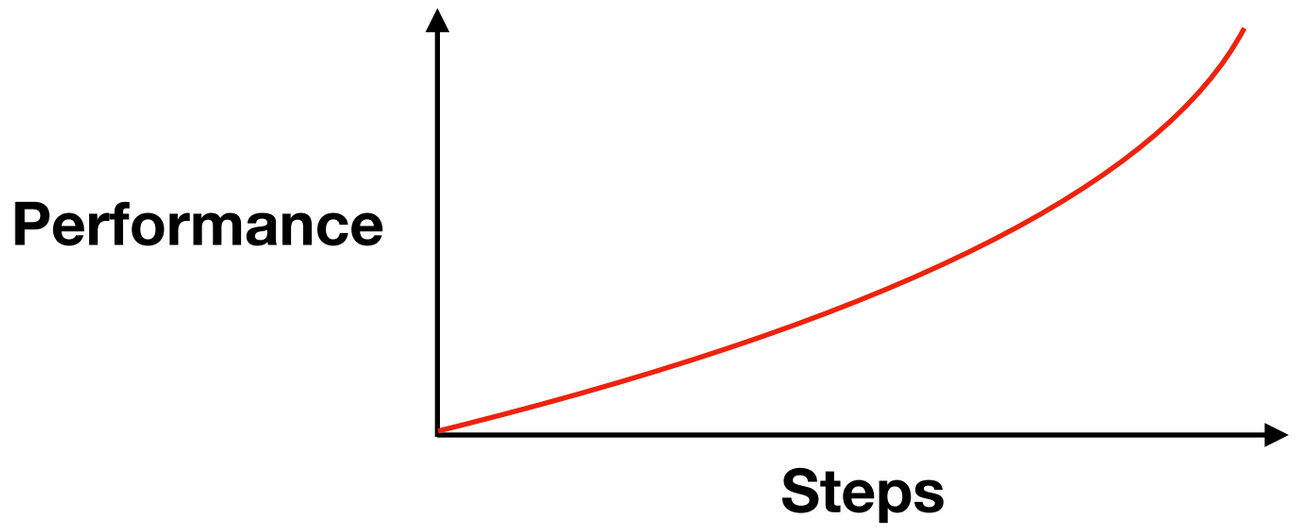
William Thomson, Lord Kelvin
1824 - 1907

Only two small clouds remained on the horizon of knowledge in physics.

Offline AlphaStar uses online Monte Carlo for model selection

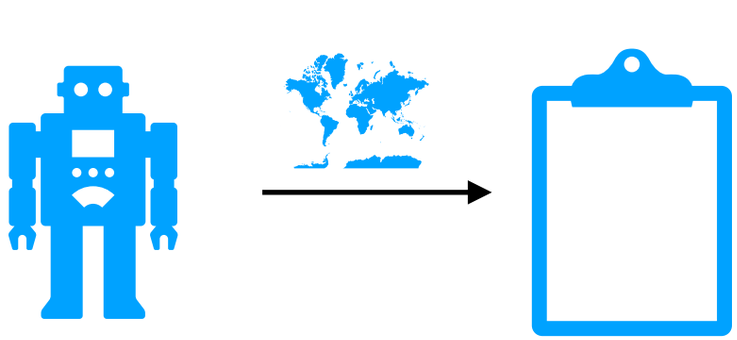


Monte Carlo dominates RL evaluation

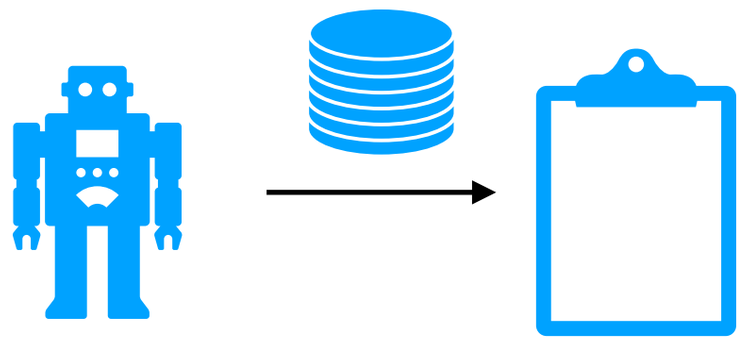


99% of such curves in RL papers are generated by online Monte Carlo

It is desired to do evaluation with offline data

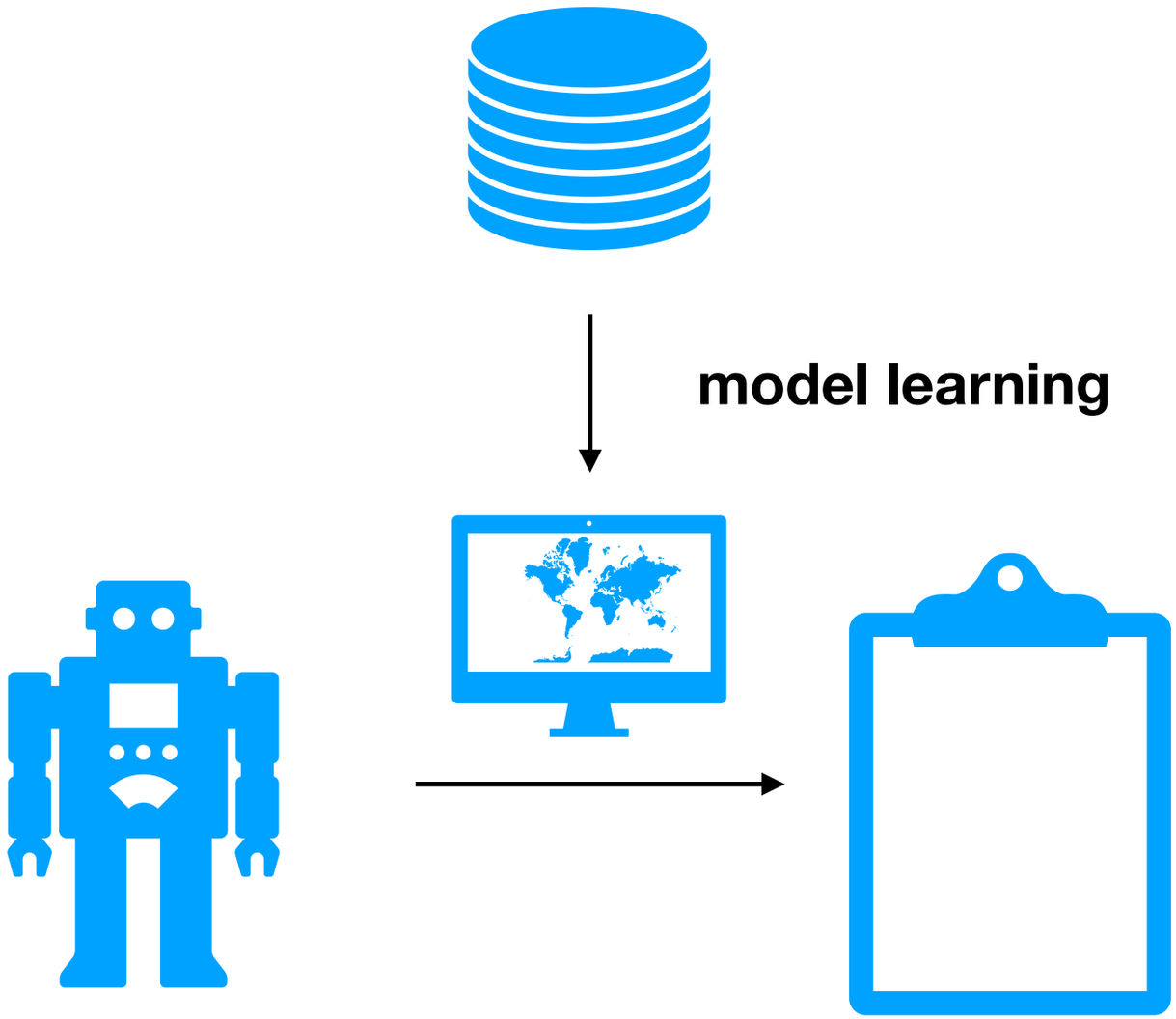


Online Monte Carlo



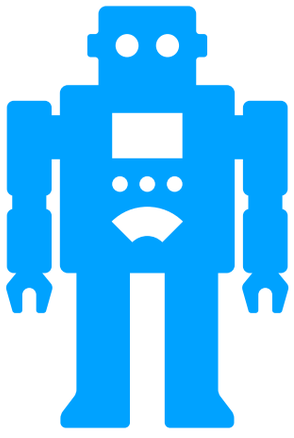
Offline data

Model-based offline evaluation reduces to simulator



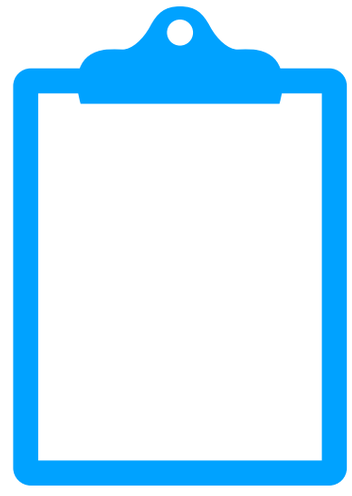
Model-free offline evaluation reduces to model-selection

“learning”

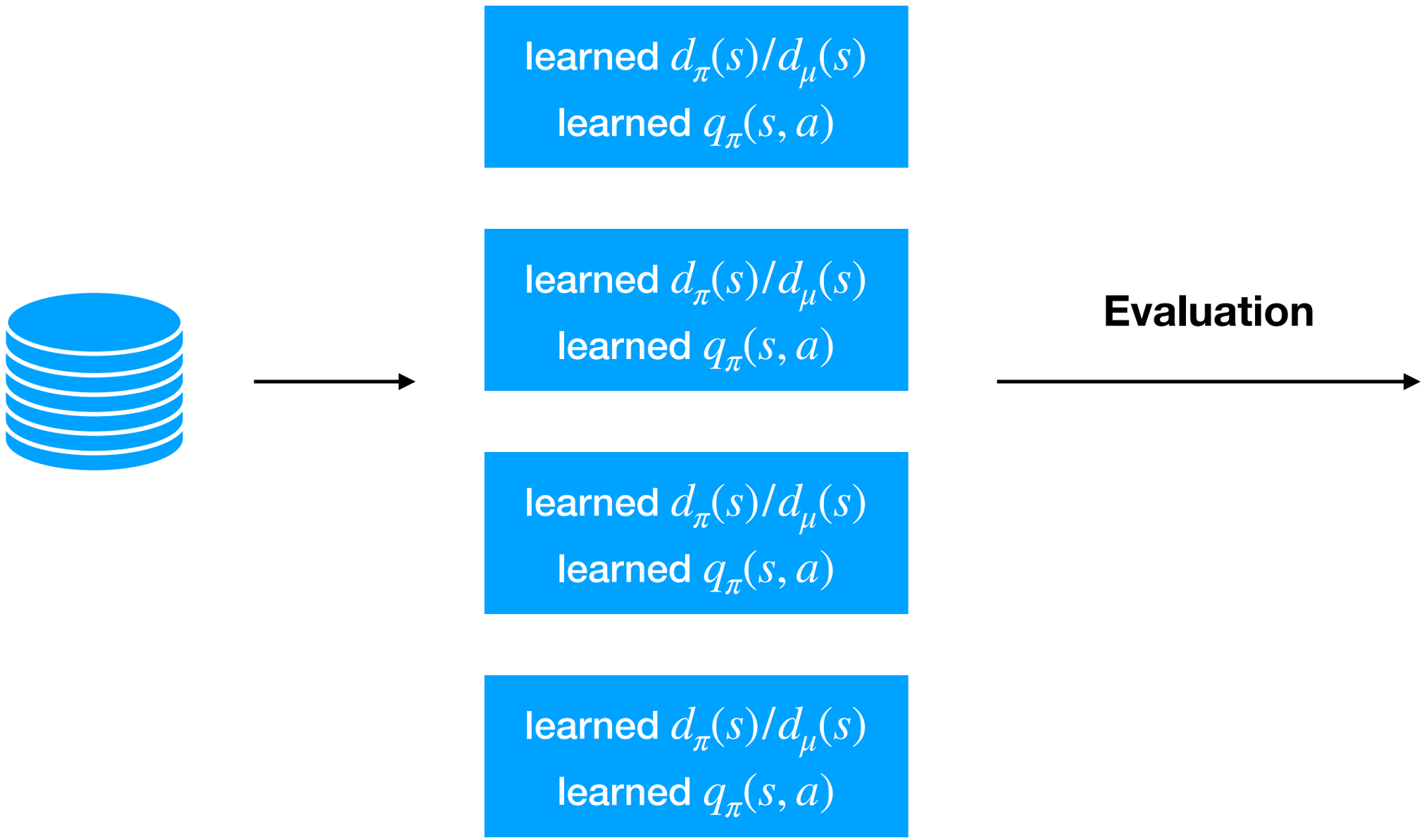


density ratio $d_\pi(s)/d_\mu(s)$

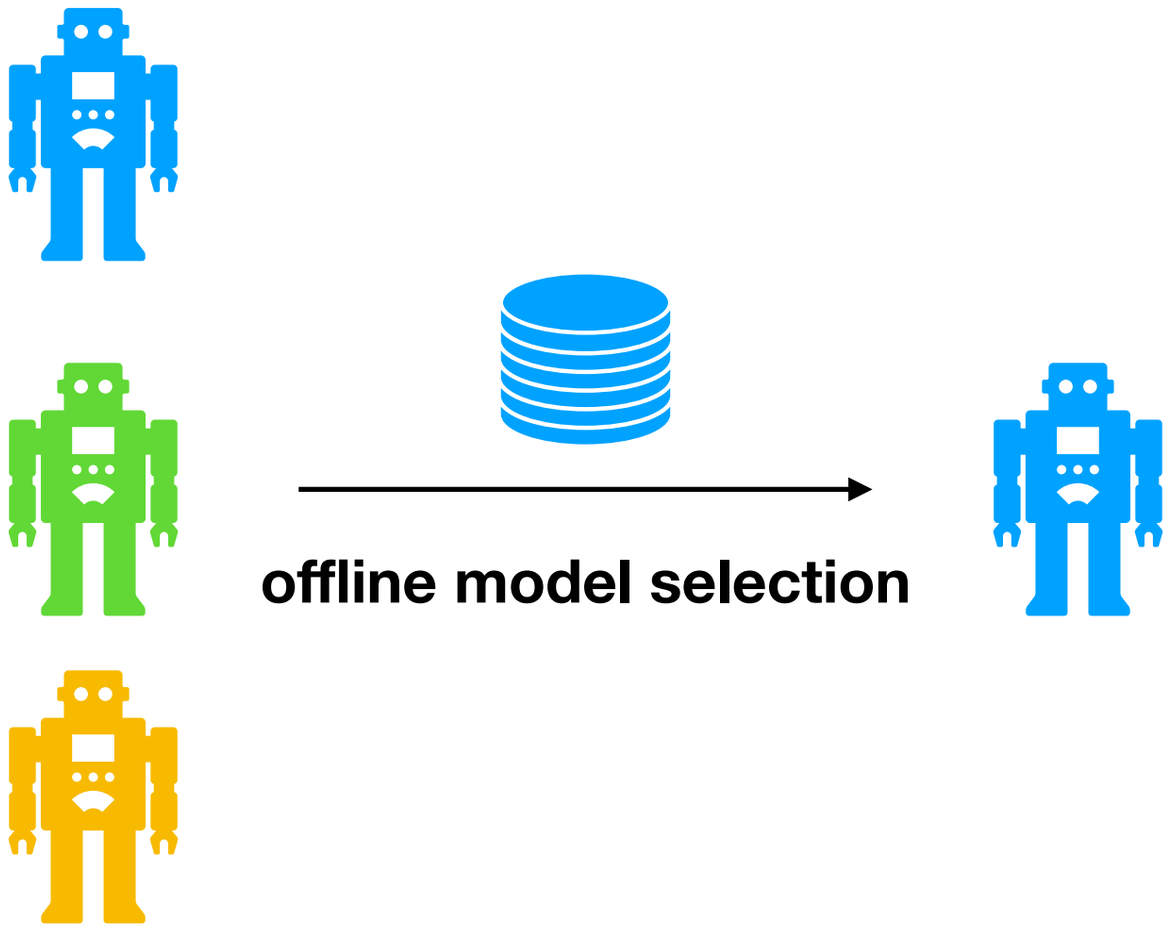
action value function $q_\pi(s,a)$



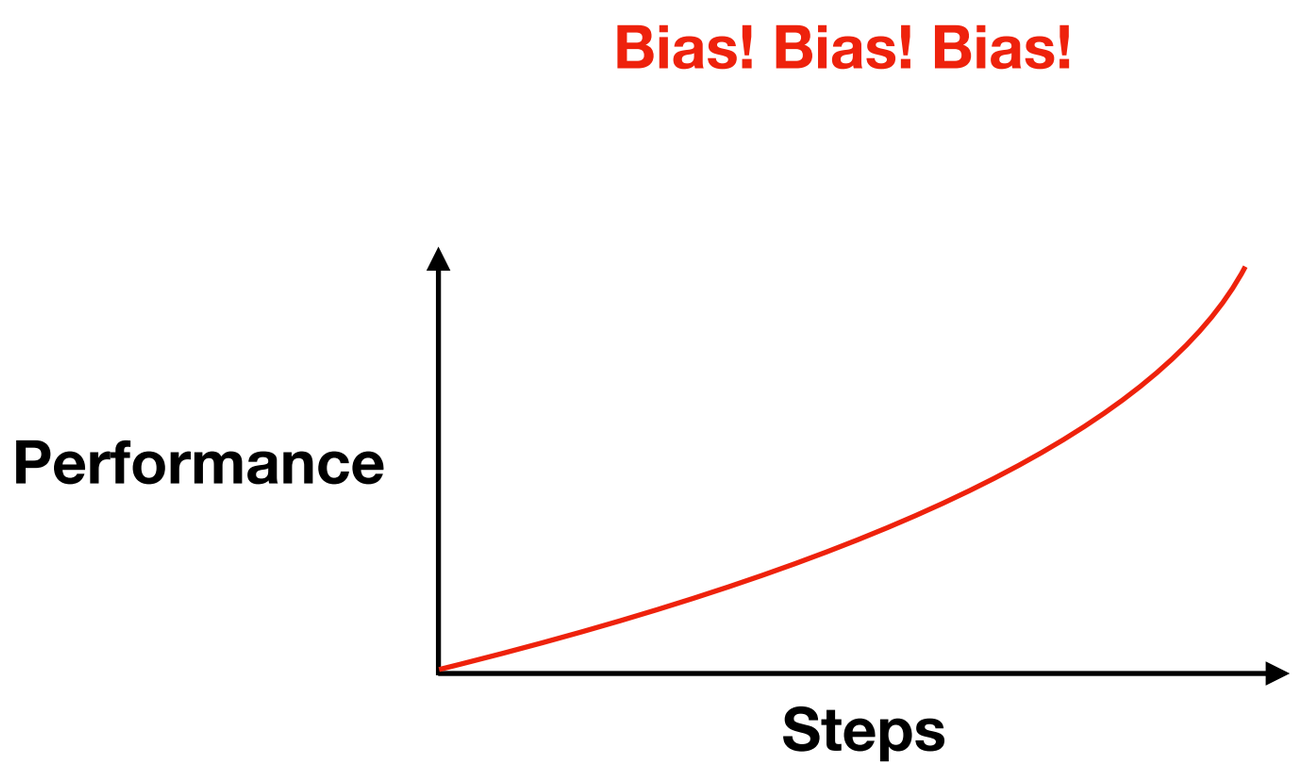
Model-free offline evaluation reduces to model-selection



Offline model selection hardly has correctness guarantee

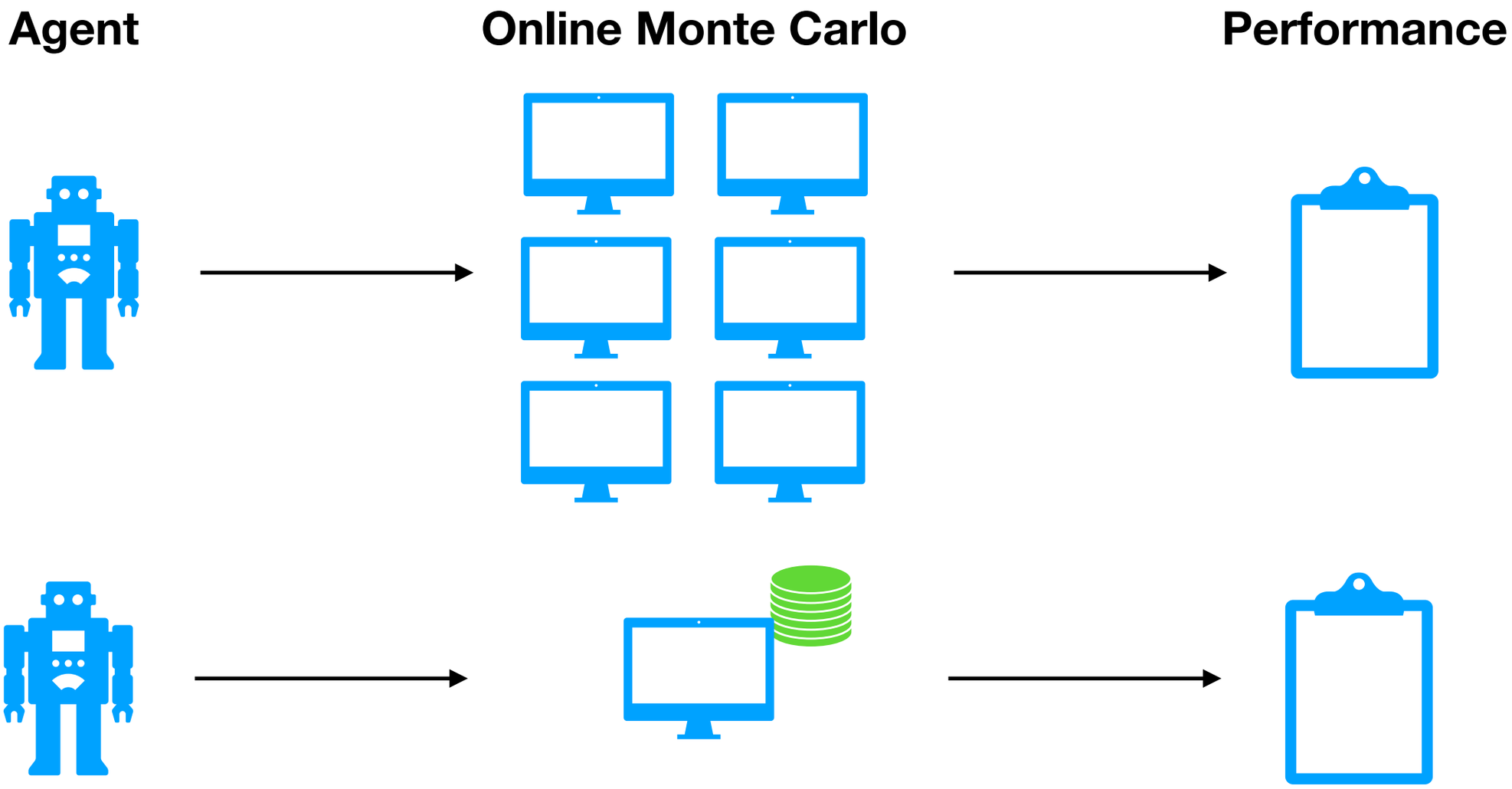


Pitfalls of offline evaluation methods



99% of such curves in RL papers are generated by online Monte Carlo

Improve Monte Carlo with offline data while maintaining its unbiasedness



Build intuition with STAT 101

- Estimating an expectation $\mathbb{E}_{X\sim p}[f(X)]$
- Monte Carlo $\frac{1}{N}\sum_{i=1}^{N} f(X_i), \quad X_i \sim p$
- Importance sampling

$$\mathbb{E}_{X\sim p}\left[f(X)\right] = \mathbb{E}_{X\sim q}\left[\frac{p(X)}{q(X)} f(X)\right] \qquad \frac{1}{N}\sum_{i=1}^{N} \frac{p(X_i)}{q(X_i)} f(X_i), \quad X_i \sim q$$

- Optimal sampling distribution minimizing the variance

$$q(x) = \frac{p(x)\,|f(x)|}{\sum_y p(y)\,|f(y)|}$$

A detailed look at the “optimal” sampling distribution

Importance sampling $\mathbb{E}_{X\sim p}\left[f(X)\right] = \mathbb{E}_{X\sim q}\left[\frac{p(X)}{q(X)}f(X)\right]$

“Optimal” sampling distribution $q(x) = \frac{p(x)\,|f(x)|}{\sum_y p(y)\,|f(y)|}$

1. $q(x)$ dose not necessarily cover $p(x)$

2. $q(x)$ is not necessarily optimal $\begin{cases} f(x_1) = -10 \\ f(x_2) = 2 \\ f(x_3) = 2 \end{cases}, \quad \begin{cases} p(x_1) = 0.1 \\ p(x_2) = 0.5, \\ p(x_3) = 0.4 \end{cases} \quad \begin{cases} q_*(x_1) = 0 \\ q_*(x_2) = 0 \\ q_*(x_3) = 1 \end{cases}$

$\mathbb{E}_{X\sim p}[f(X)] = 0.8 = \mathbb{E}_{X\sim q_*}[\frac{p(x)}{q_*(x)}f(X)]$

From STAT 101 to RL 999

- Data in RL $\{S_0, A_0, R_1, S_1, A_1, R_2, \ldots, R_T\} \sim \mu$

- Per-decision importance sampling ratio Monte Carlo estimator

$$\sum_{t=1}^{T} \left(\prod_{i=0}^{t-1} \frac{\pi(A_i \mid S_i)}{\mu(A_i \mid S_i)} \right) R_t$$

"Improving Monte Carlo Evaluation with Offline Data."
Shuze Liu, Shangtong Zhang
arXiv:2301.13734, 2023.

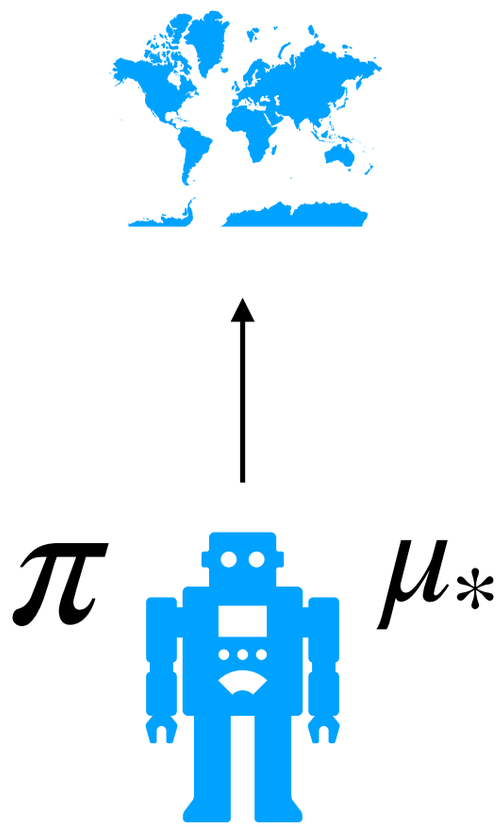
- A **provably** variance reducing behavior policy for the per-decision MC estimator

$$\sum_{t=1}^{T}\left(\prod_{i=0}^{t-1}\frac{\pi(A_i\,|\,S_i)}{\mu(A_i\,|\,S_i)}\right)R_t$$

- A computationally efficient and model-free method to learn this behavior policy from offline data

"Improving Monte Carlo Evaluation with Offline Data."
Shuze Liu, Shangtong Zhang
arXiv:2301.13734, 2023.

- A bandit algorithm to adaptively switch between target and behavior policies for data collection



Searching good behavior policies for off-policy MC is not new

- Some cannot exploit offline data and require new online data
- Some assume special structure of the MDP and need to learn a model

Take home message

- Improve Monte Carlo evaluation with offline data might be the next agenda in offline RL

- Vinyals, Oriol, et al. "Grandmaster level in StarCraft II using multi-agent reinforcement learning." Nature 575.7782 (2019): 350-354.

- Chervonyi, Yuri, et al. "Semi-analytical Industrial Cooling System Model for Reinforcement Learning." arXiv preprint arXiv:2207.13131 (2022).

Thanks