Frontiers:
Offline Reinforcement Learning

Sanjiban Choudhury

Cornell Bowers CIS
Computer Science

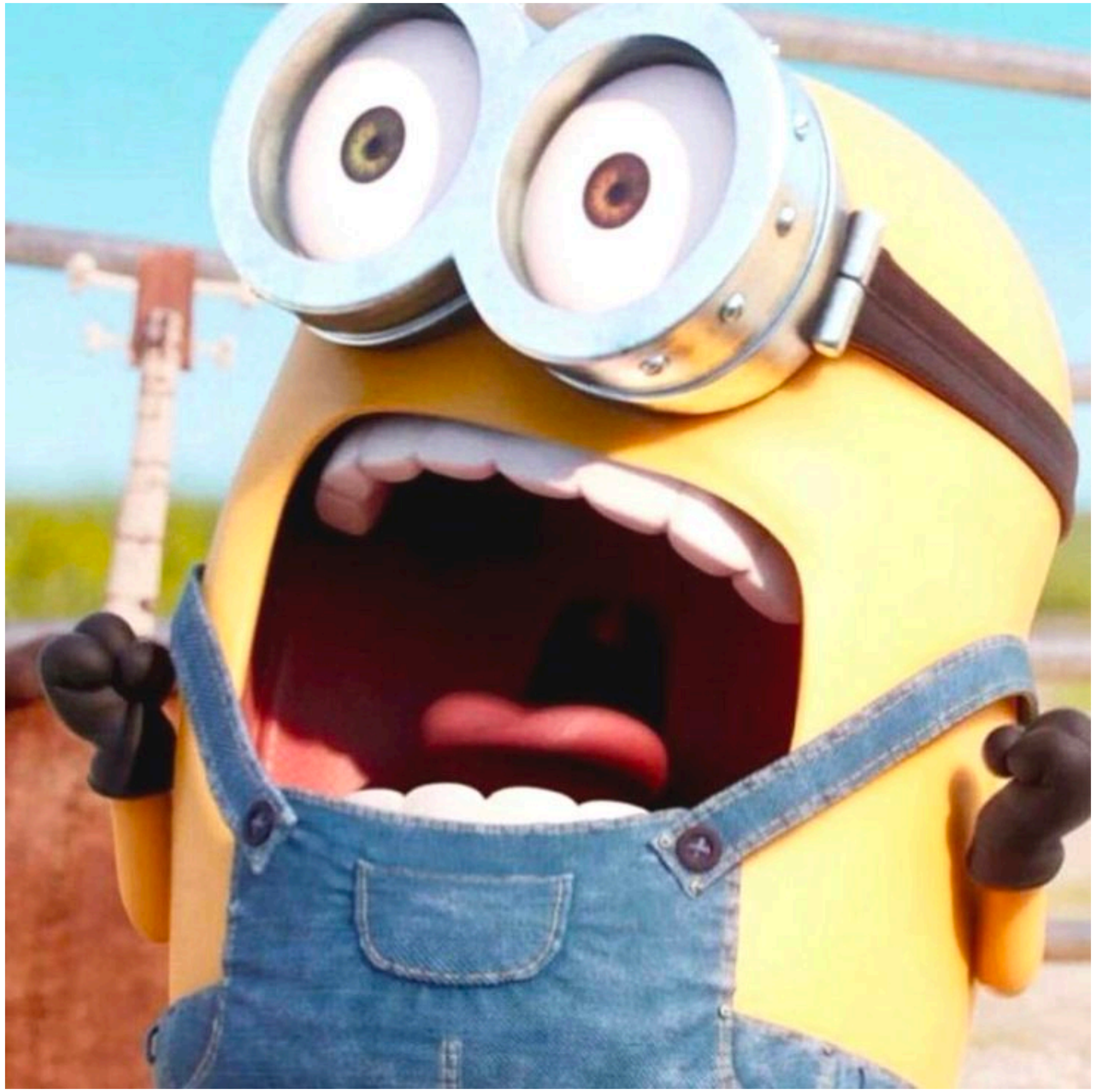
Frontiers

Problem: *Insane* number of papers out there!!

Impossible for outsiders to
find any sort of scaffolding

Many of these papers
recycle old ideas
while butchering the insight

Hope: Sparse set of papers
that give you reach



Our Strategy

Goal: Engage with various frontiers of research on robot decision making

Strategy: Equip you with a sparse "support vector" of papers that gives you maximal reach on the problem

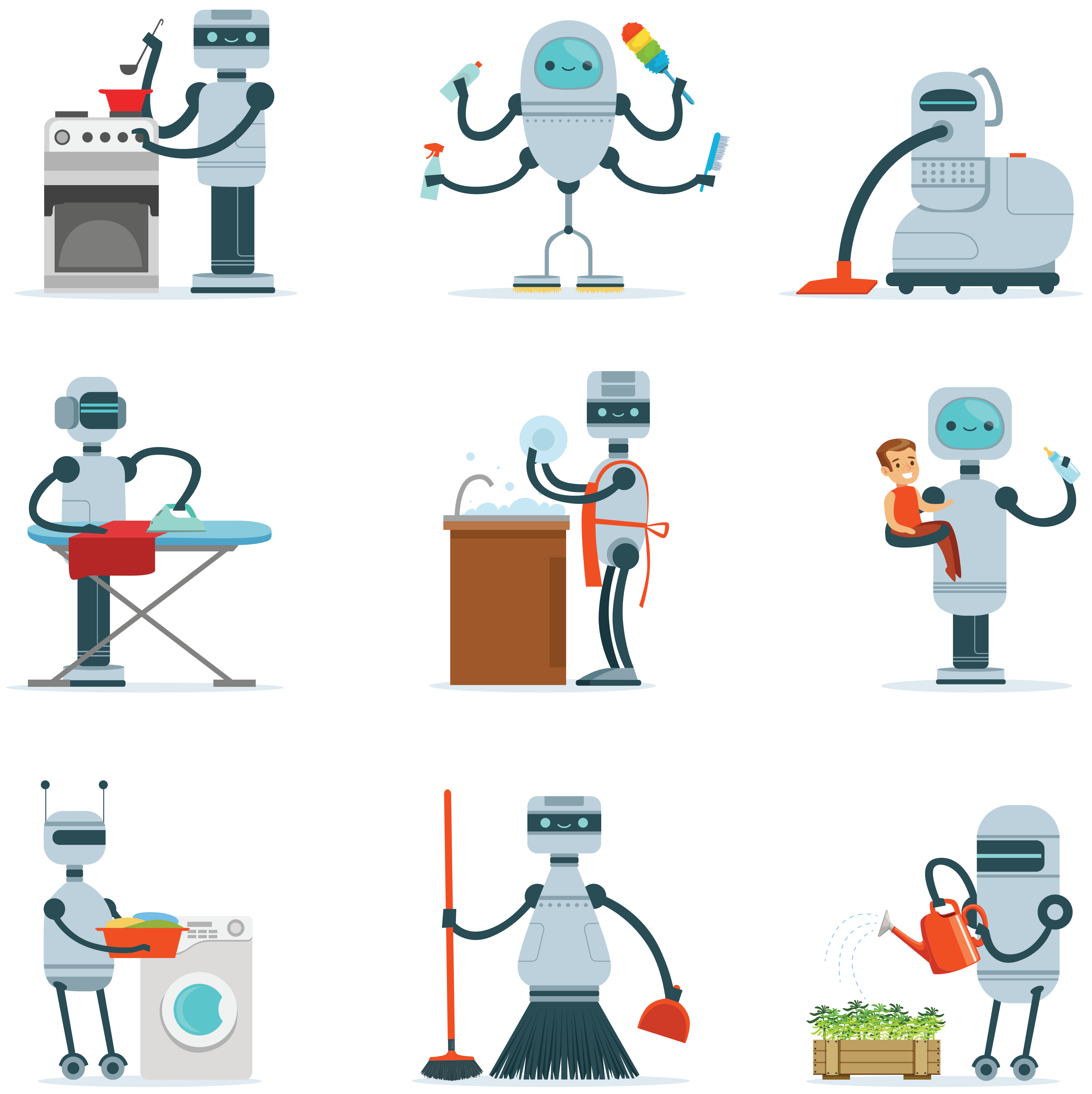
Expectation: For details and concrete implementation, you should be able to look that up

The Problem

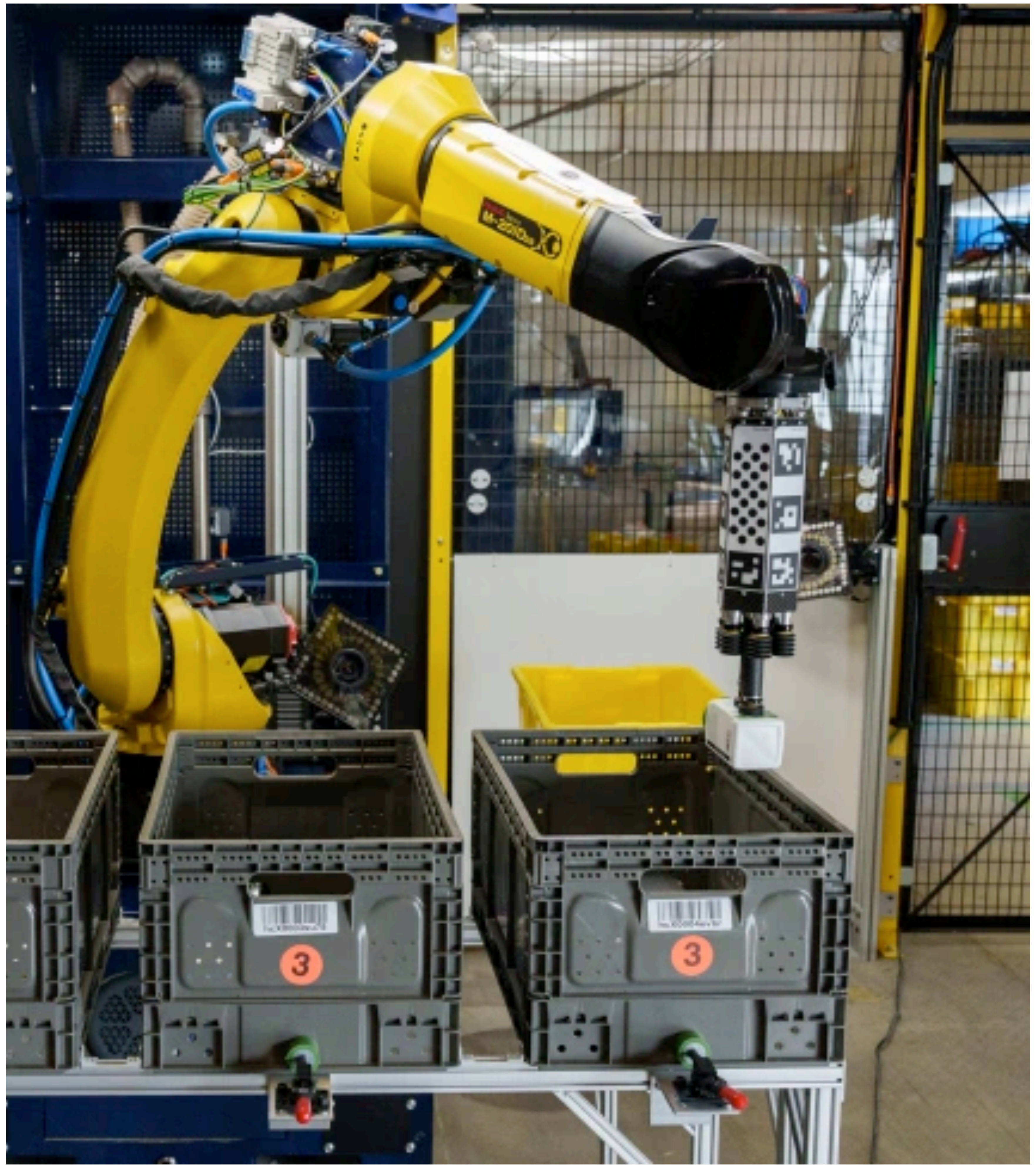
Real World, Real Problems

Robots can augment human capabilities to tackle these problems

Robots only really work in the CLOSED world

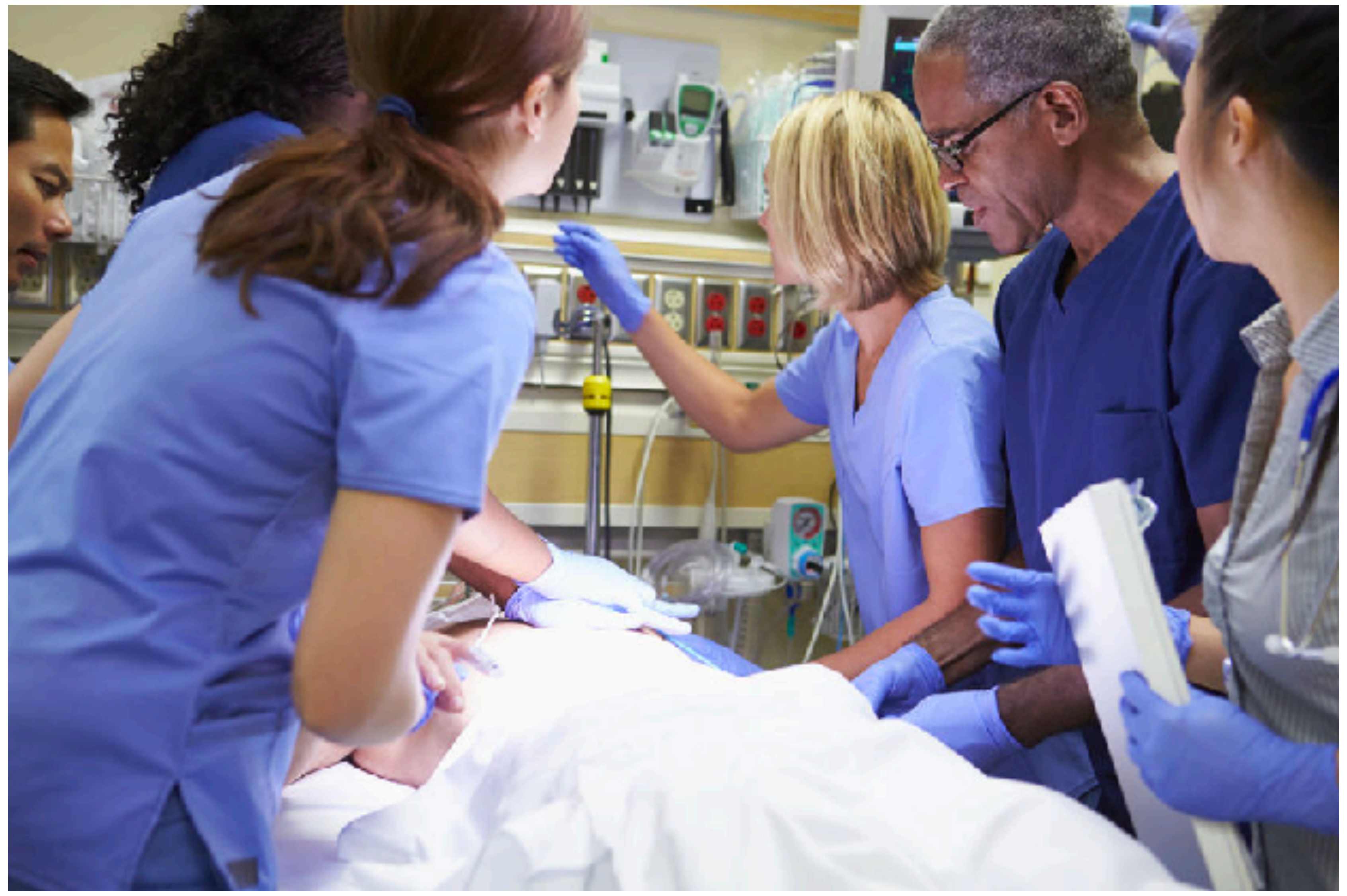


The Dream

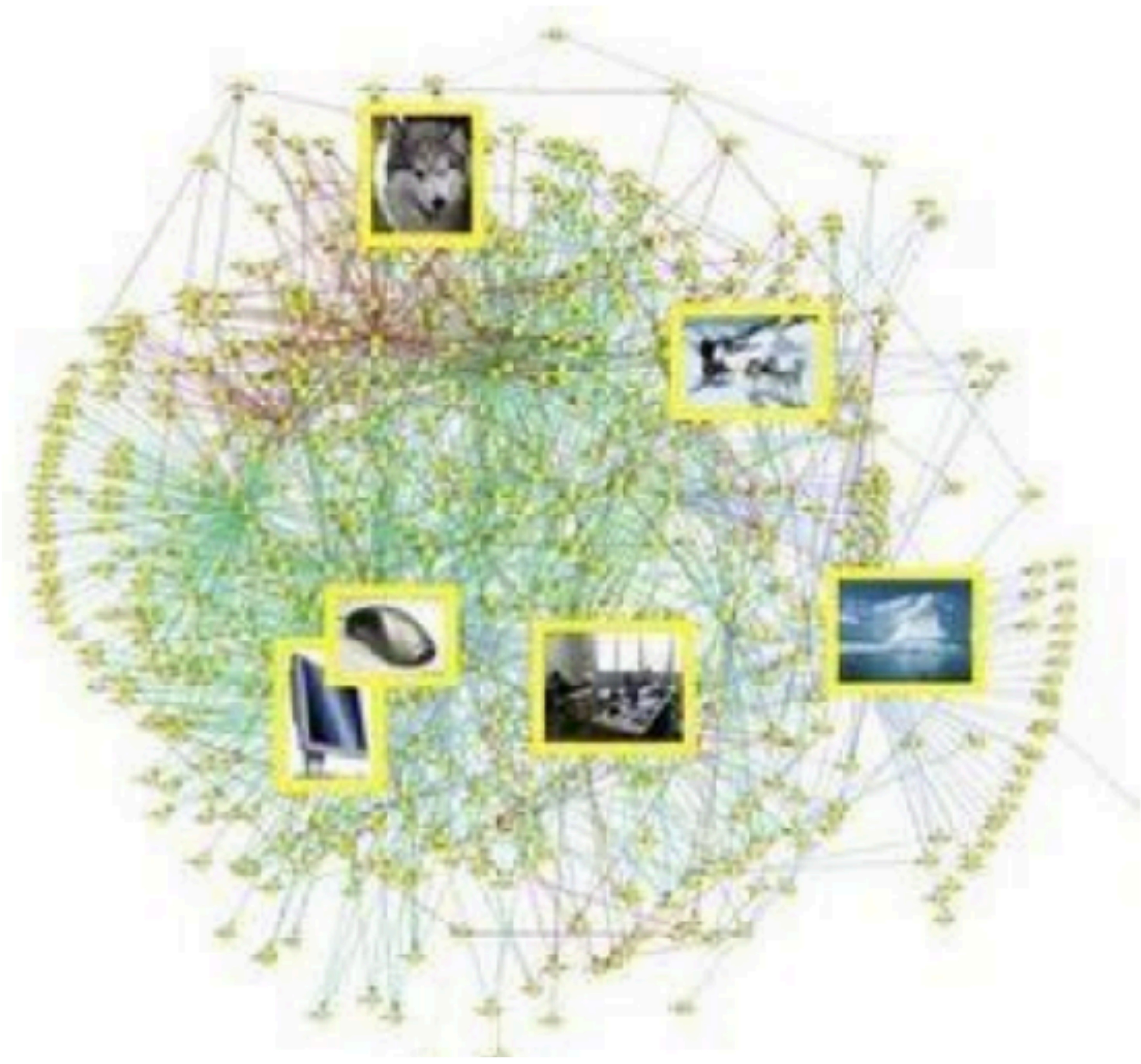


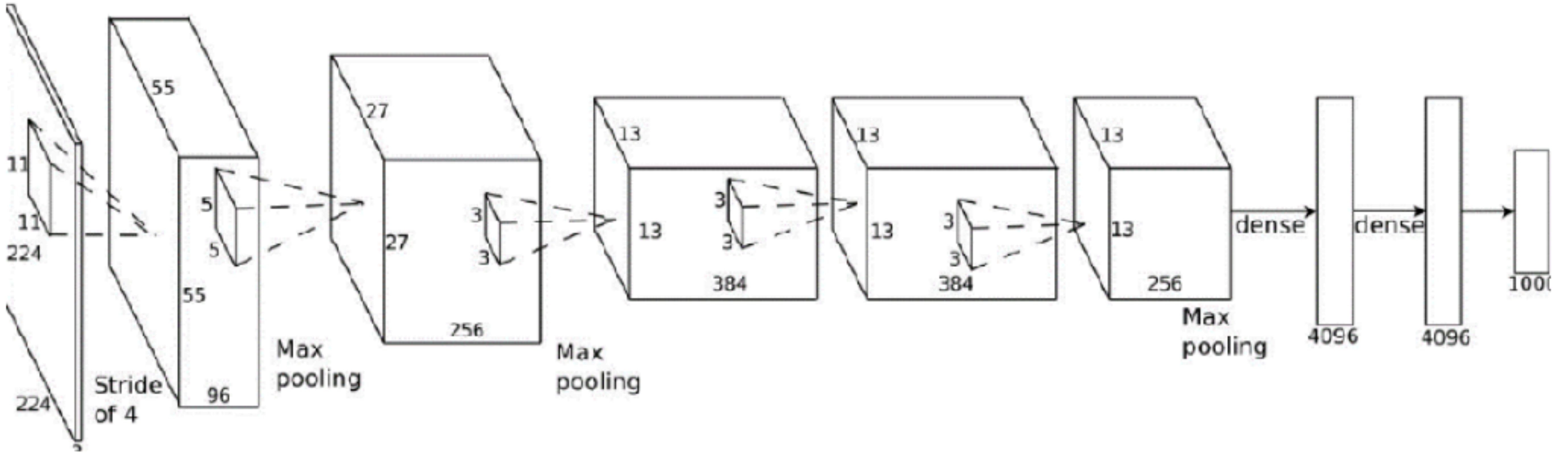
Reality

Generalize to variations of the OPEN world?



Machine learning's answer!



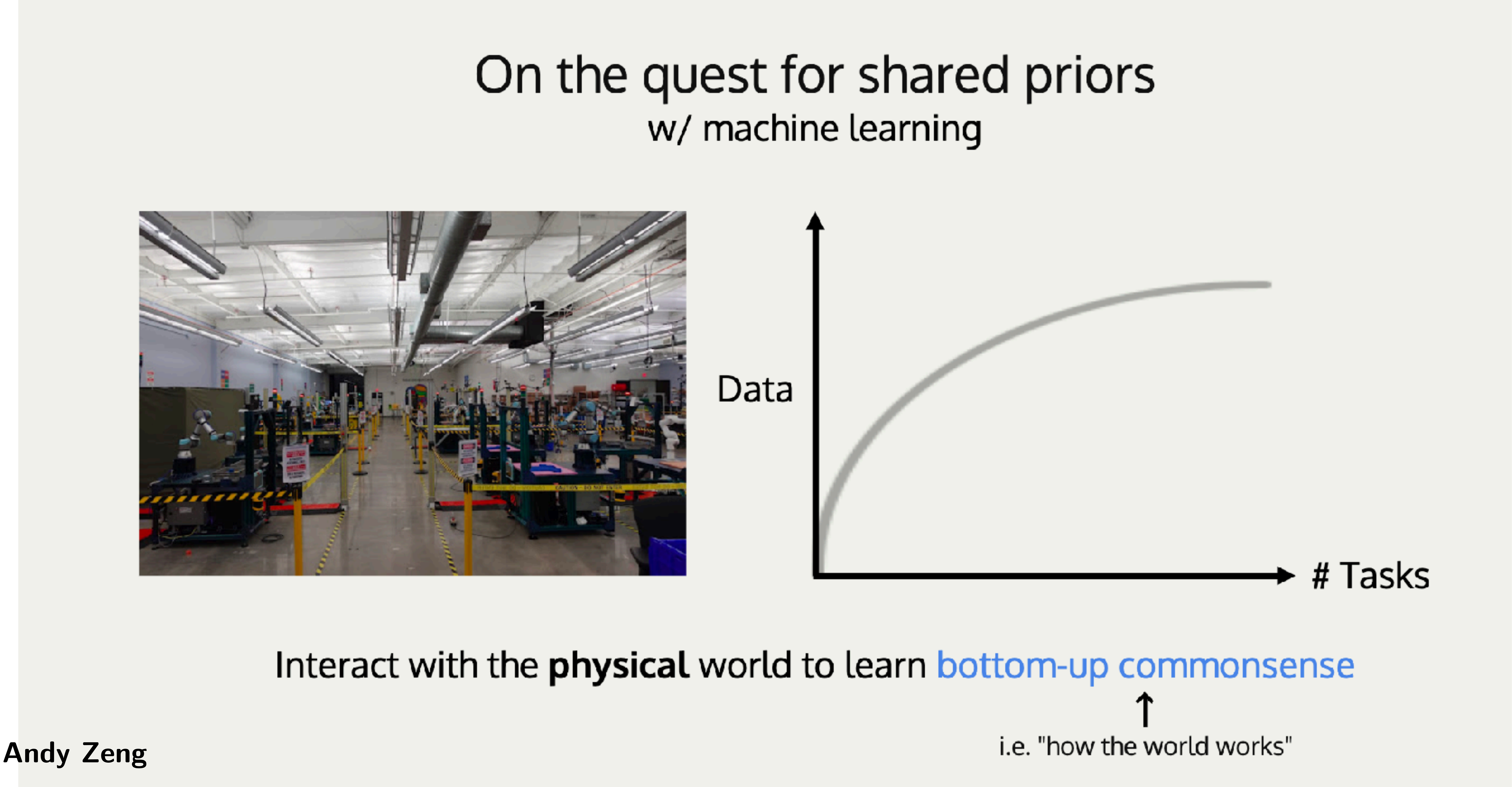


Big Data

Big Models

Credit: **Sergey Levine** "Offline RL lecture"

Hasn't quite been true so far robotics ...

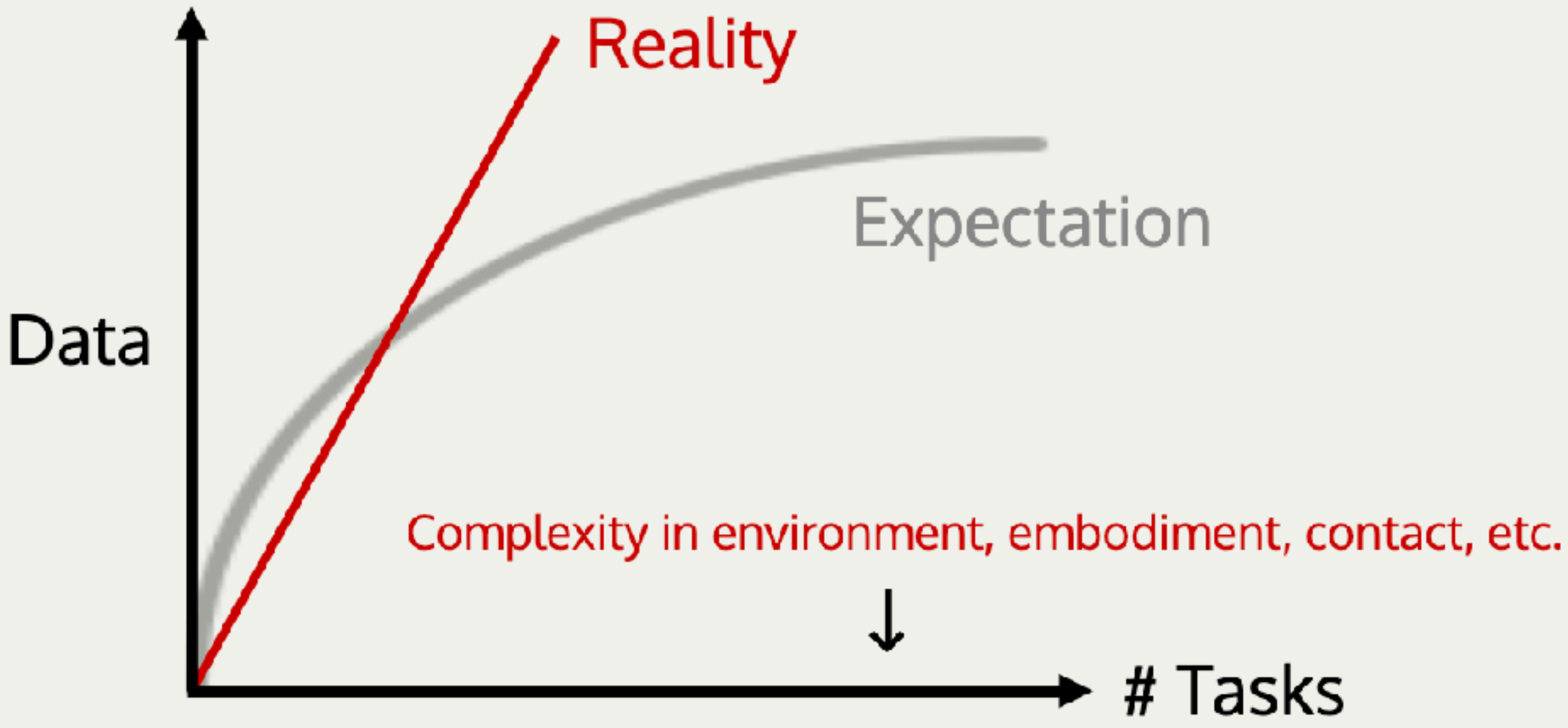


Credit: **Andy Zeng**

Hasn't quite been true so far robotics ...

On the quest for shared priors

w/ machine learning



Interact with the **physical** world to learn bottom-up commonsense

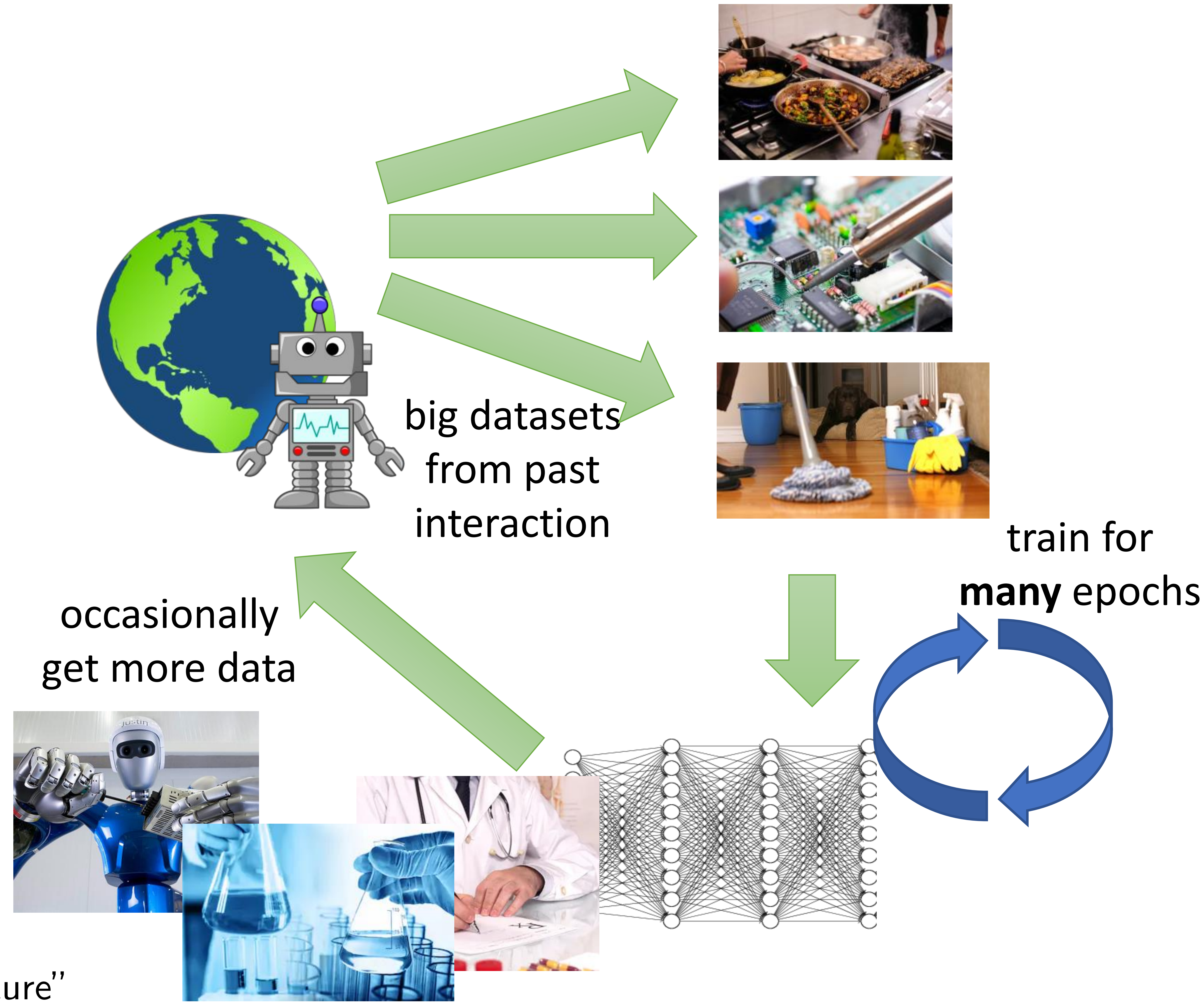
i.e. "how the world works"

Credit: **Andy Zeng**

But for today, let's pretend we can collect a ton of data

How can we learn "optimal" from large data collected by *any* policy?

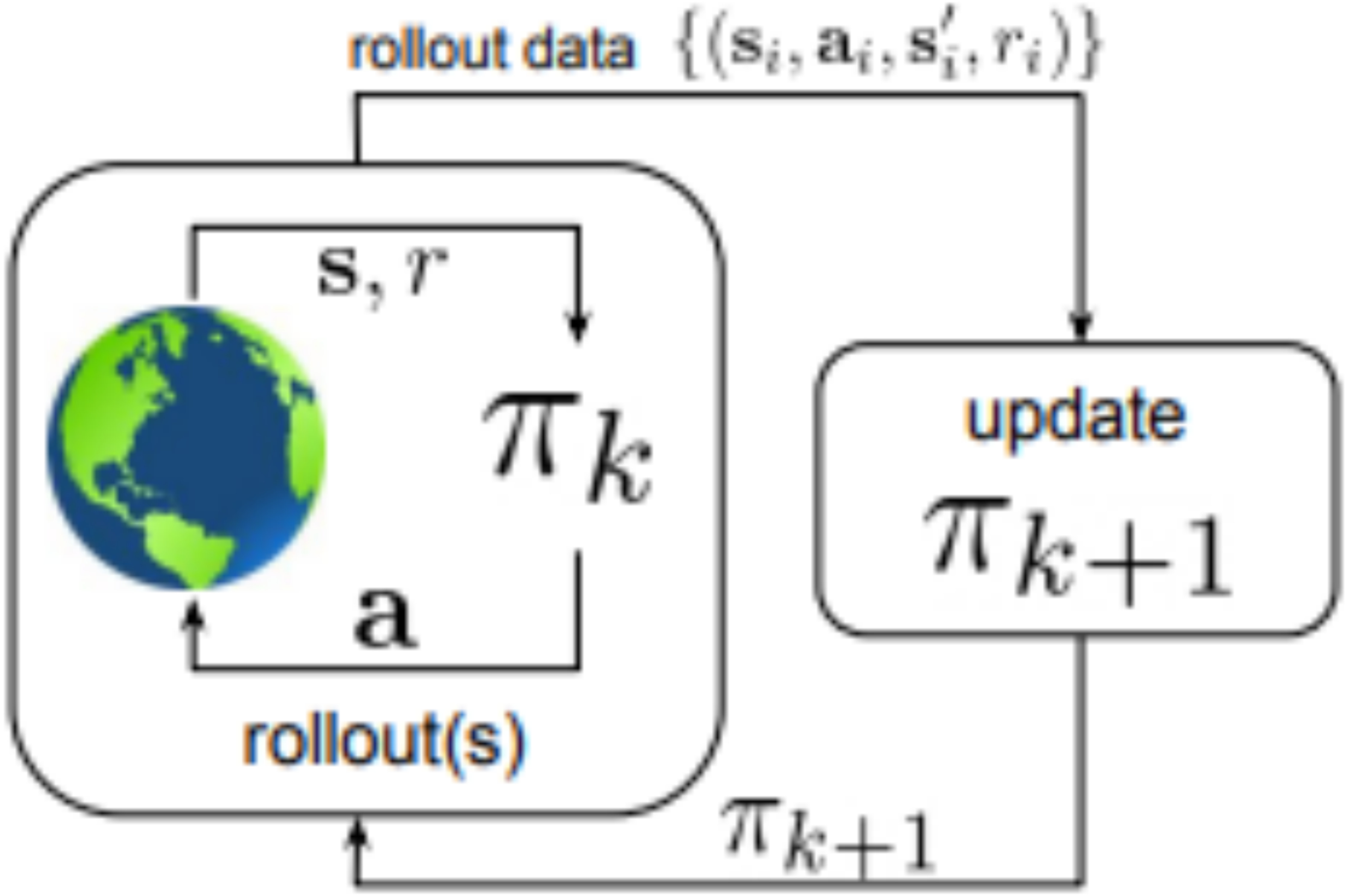
Goal: Offline Reinforcement Learning



Credit: **Sergey Levine** “Offline RL lecture”

Different paradigms of RL

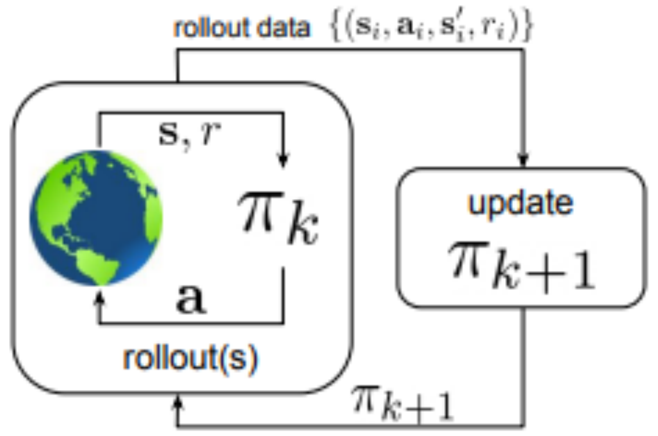
on-policy RL



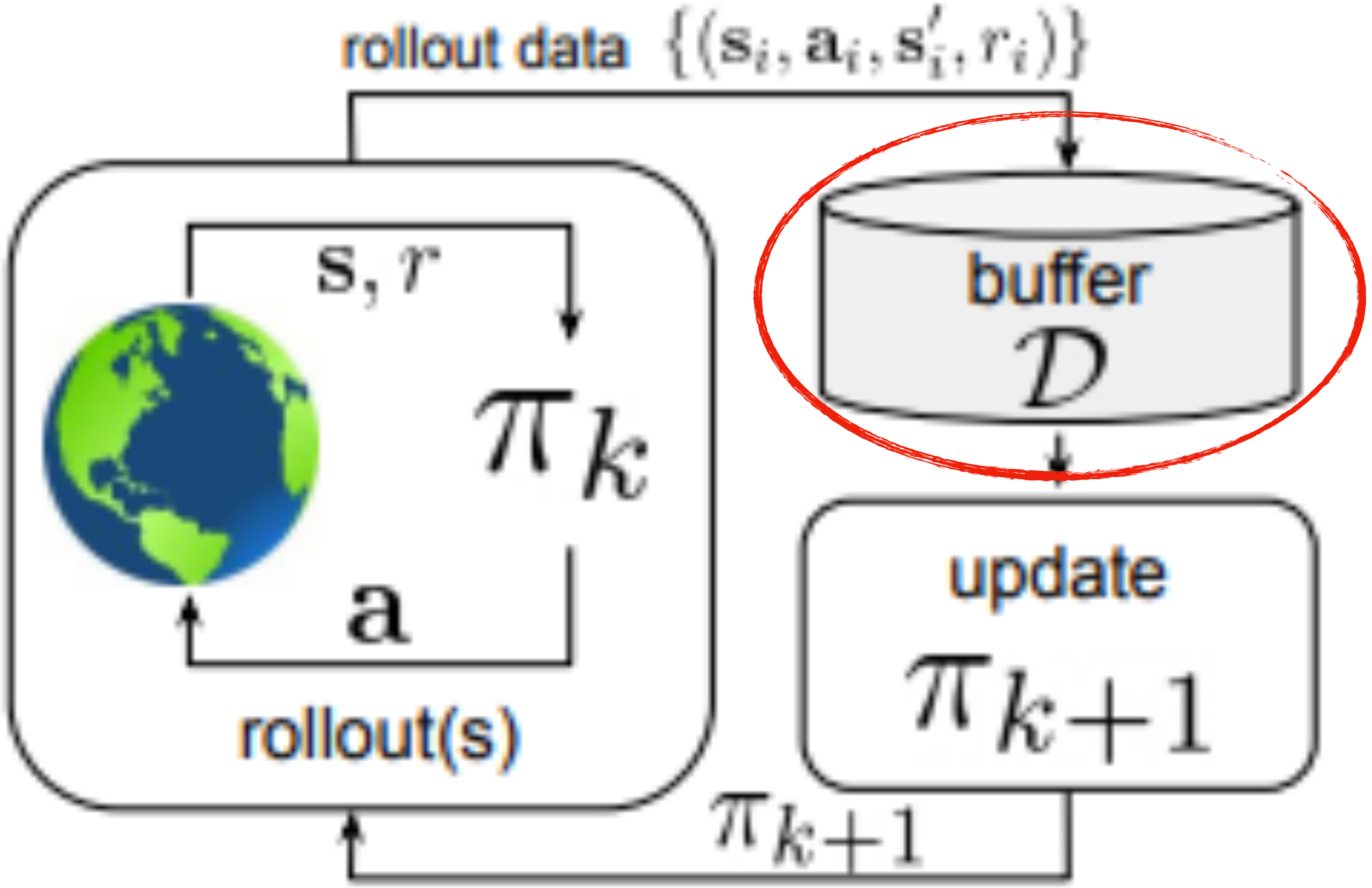
Credit: **Sergey Levine** "Offline RL lecture"

Different paradigms of RL

on-policy RL



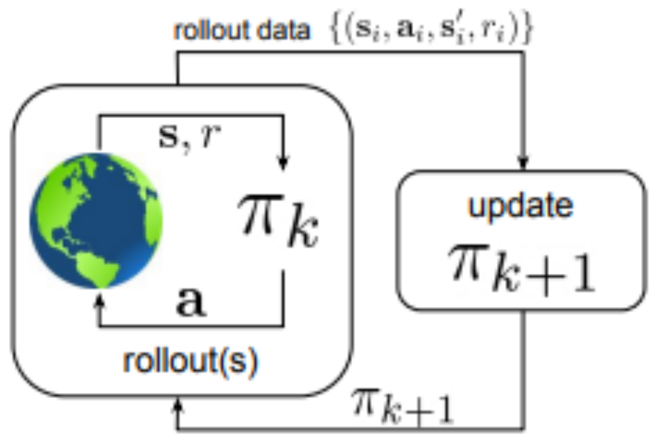
off-policy RL



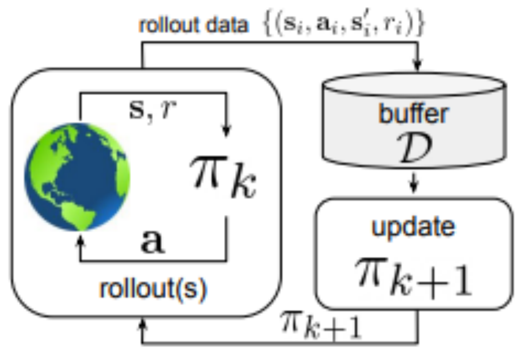
Credit: **Sergey Levine** "Offline RL lecture"

Different paradigms of RL

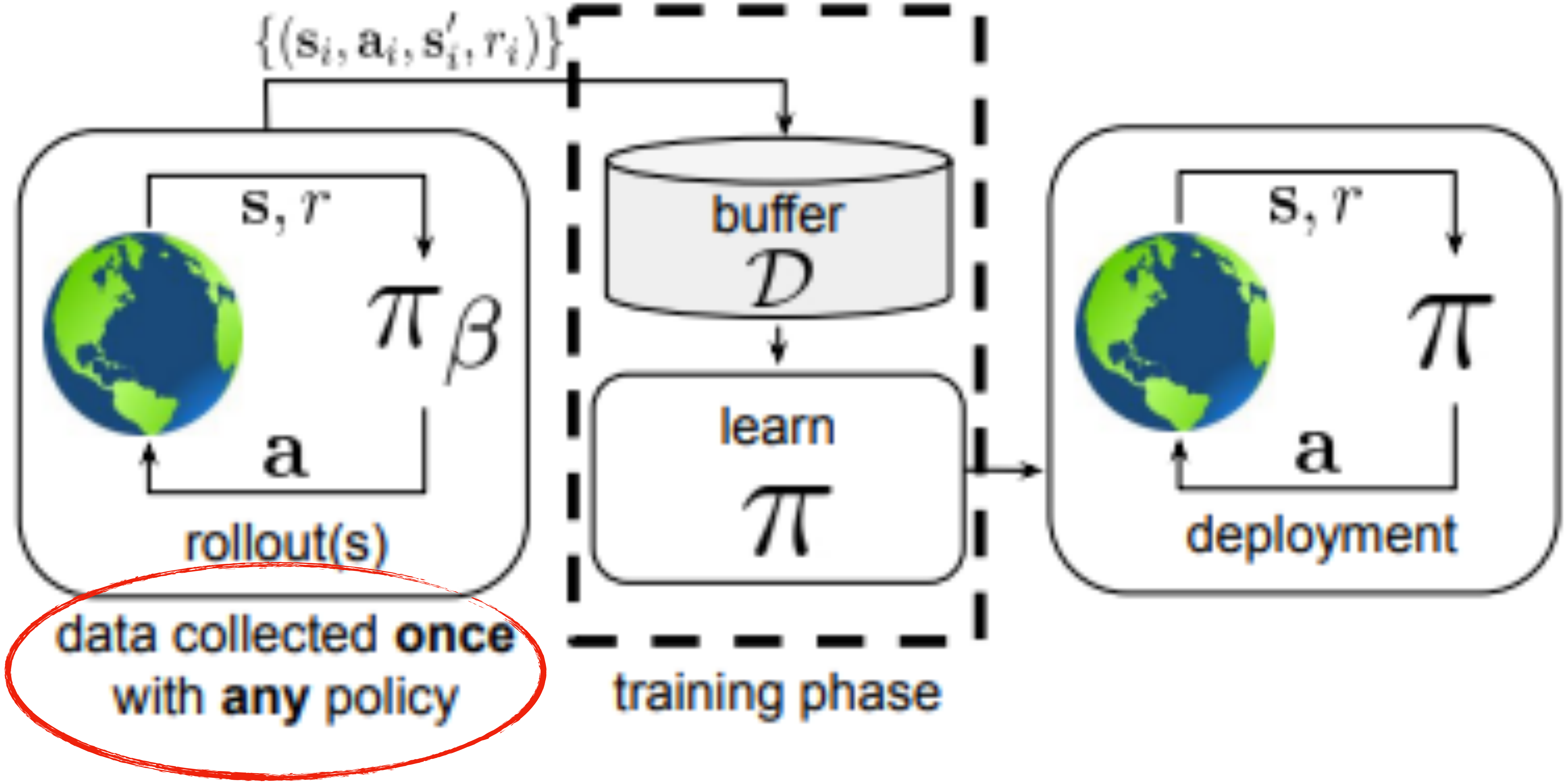
on-policy RL



off-policy RL



offline reinforcement learning



Credit: **Sergey Levine** "Offline RL lecture"

Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems

Sergey Levine[1,2], Aviral Kumar[1], George Tucker[2], Justin Fu[1]
[1]UC Berkeley, [2]Google Research, Brain Team

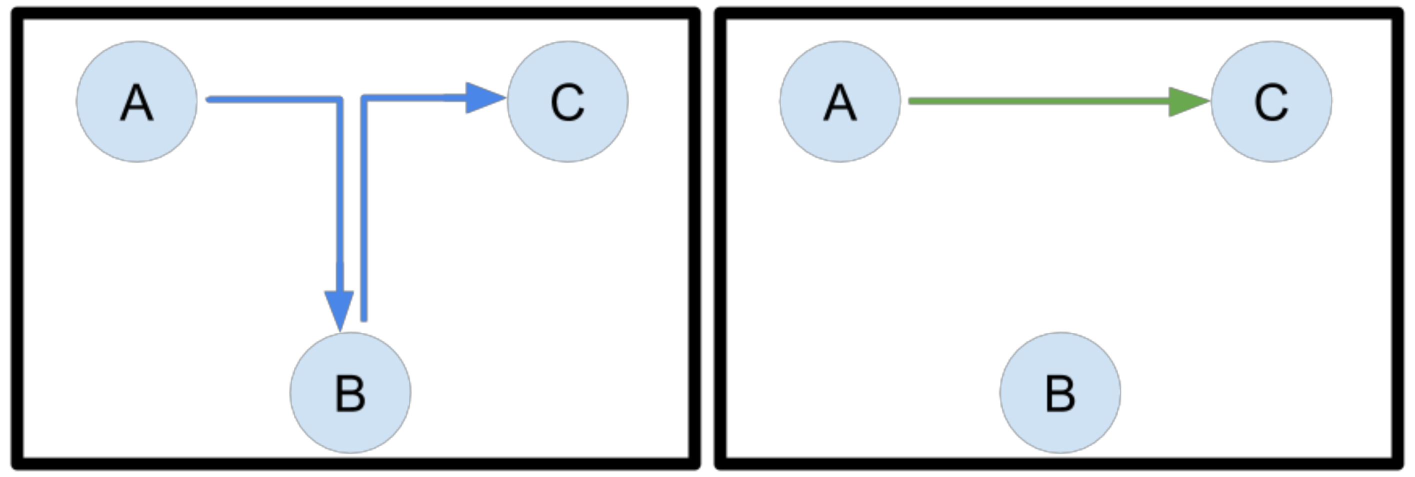
Fun collab tutorial: https://colab.research.google.com/drive/1oJOYlAIOl9d1JjlutPY66KmfPkwPCgEE?usp=sharing

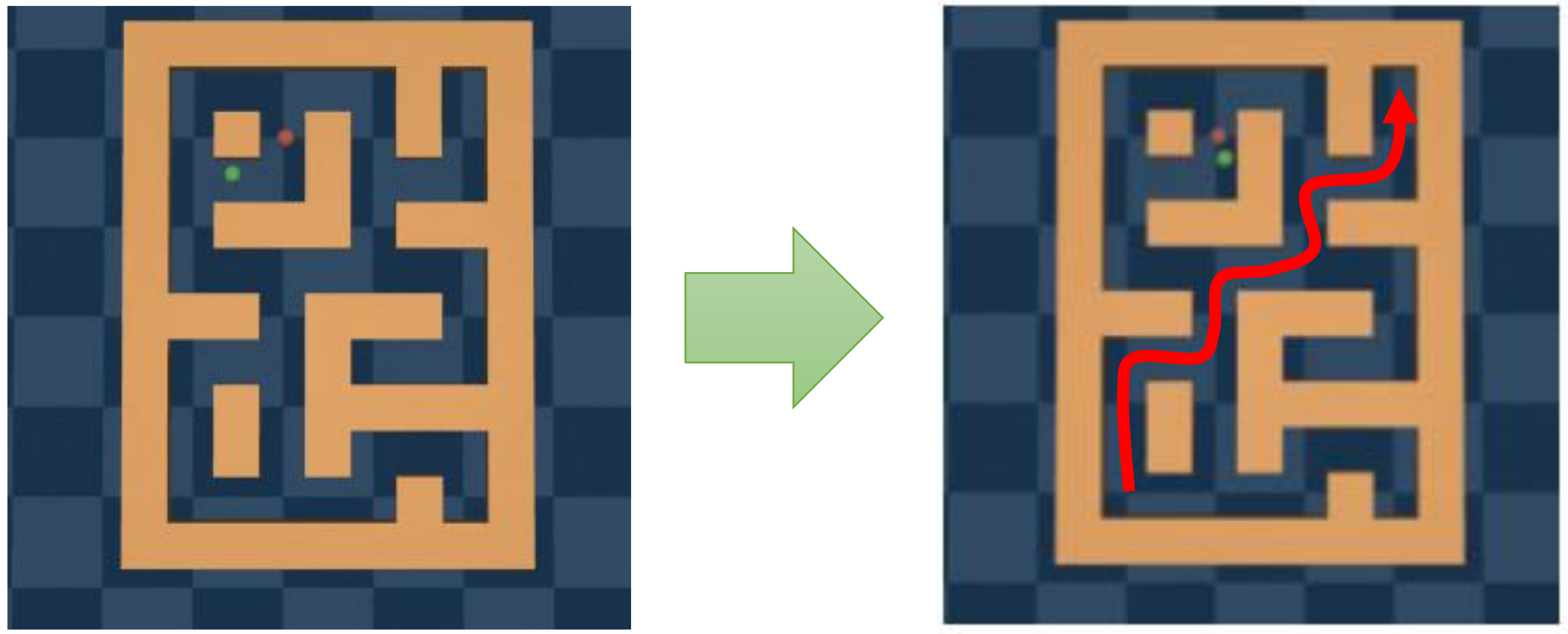
How is this even possible?

1. Find the "good stuff" in a dataset full of good and bad behaviors

2. Generalization: good behavior in one place may suggest good behavior in another place

3. "Stitching": parts of good behaviors can be recombined



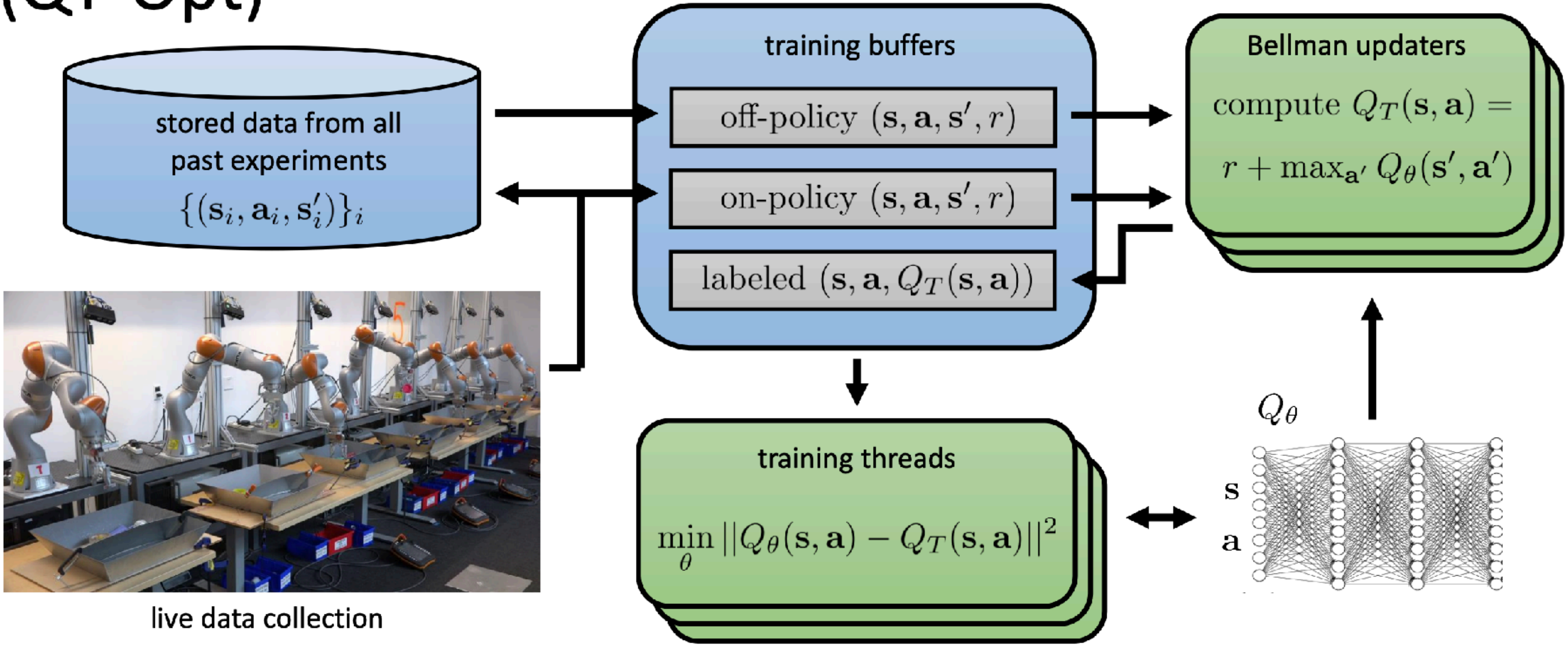


Credit: **Sergey Levine** "Offline RL lecture"

Does it work?

Sometimes*

Large-scale Q-learning with continuous actions (QT-Opt)



Kalashnikov, Irpan, Pastor, Ibarz, Herzong, Jang, Quillen, Holly, Kalakrishnan, Vanhoucke, Levine. **QT-Opt: Scalable Deep Reinforcement Learning of Vision-Based Robotic Manipulation Skills**

Optimal Insulin Dose

Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes

Harry Emerson [a,*], Matthew Guy [a,b], Ryan McConville [a]

[a] *University of Bristol, 1 Cathedral Square, Bristol, BS1 5TS, United Kingdom*
[b] *University Hospital Southampton, Tremona Road, Southampton, SO16 6YD, Hampshire, United Kingdom*

Table 2
The mean performance of the offline RL algorithms: BCQ, CQL and TD3-BC against the online RL approach SAC-RNN and the control baseline PID. TD3-BC can be seen to significantly improve the proportion of TIR when compared to the PID and the SAC-RNN algorithms. This is done so without any associated increase in risk (reward) or TBR. Statistical significance was confirmed via a Friedman rank test for all glucose metrics ($p < 0.05$). †, ‡ and § indicate an offline RL, online RL and classical control algorithm respectively, with the best performing algorithm highlighted in bold.

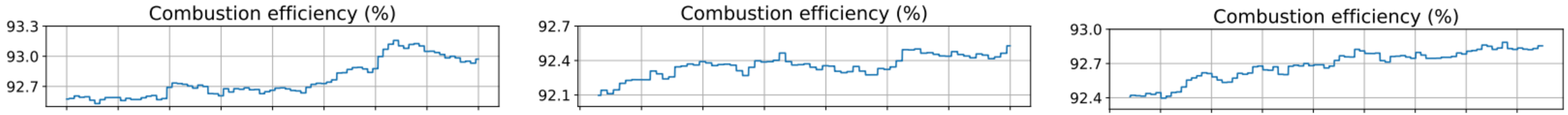
Algorithm	Reward	TIR (%)	TBR (%)	CV (%)	Failure (%)
BCQ†	−41,034 ± 1,060	**65.8 ± 0.6**	1.0 ± 0.1	35.1 ± 0.4	**0.00**
CQL†	−45,259 ± 1,071	56.2 ± 0.5	**0.1 ± 0.1**	30.3 ± 0.3	**0.00**
TD3-BC†	**−37,955 ± 547**	65.3 ± 0.5	0.2 ± 0.1	33.3 ± 0.2	**0.00**
SAC-RNN‡	−93,480 ± 71,826	34.9 ± 3.1	4.1 ± 0.7	**29.6 ± 1.3**	13.3
PID§	−49,077 ± 556	61.6 ± 0.3	0.4 ± 0.1	33.5 ± 0.2	**0.00**

Combustion control in power stations

DeepThermal: Combustion Optimization for Thermal Power Generating Units Using Offline Reinforcement Learning

Xianyuan Zhan[1]*, Haoran Xu[2,3,4]*, Yue Zhang[2,3], Xiangyu Zhu[2,3], Honglei Yin[2,3], Yu Zheng[2,3,4]

[1] Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China
[2] JD iCity, JD Technology, Beijing, China
[3] JD Intelligent Cities Research, Beijing, China
[4] Xidian University, Xi'an, China
{zhanxianyuan, ryanxhr, zhangyuezjx, zackxiangyu, yinhonglei93}@gmail.com, msyuzheng@outlook.com



(a) 270 MW Experiment (b) 290MW Experiment (c) 310MW Experiment

Figure 4: Real-world experiments at CHN Energy Nanning Power Station

Goal-directed conversation

CHAI: A CHatbot AI for Task-Oriented Dialogue with Offline Reinforcement Learning

Siddharth Verma
UC Berkeley

Justin Fu
UC Berkeley

Mengjiao Yang
UC Berkeley

Sergey Levine
UC Berkeley

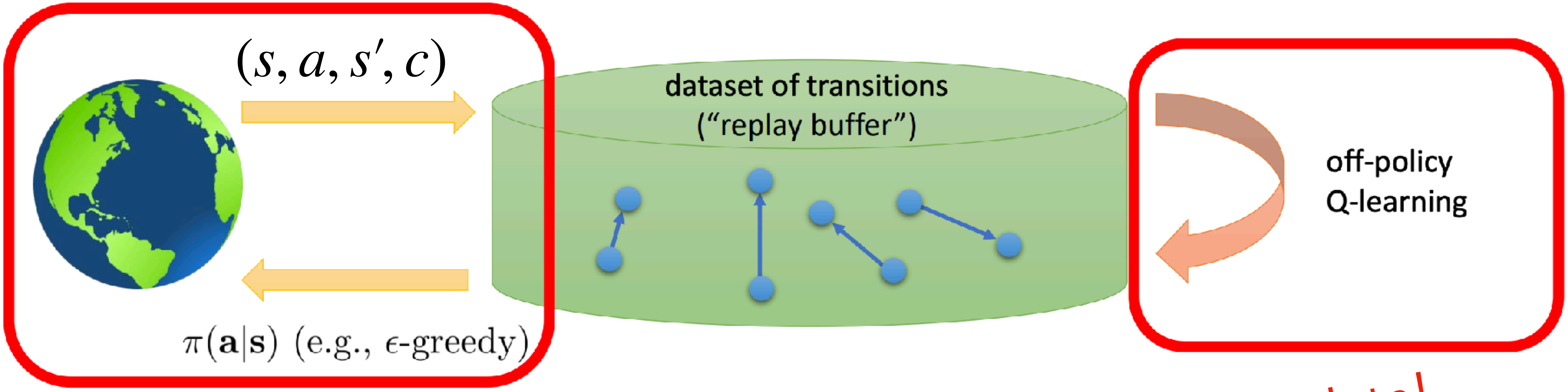
Metric	Fluency	Coherency	On-Topic	Human-Likeness	Total
CHAI-prop	$\mathbf{4.31 \pm 0.97}$	$\mathbf{3.91 \pm 1.17}$	$\mathbf{4.16 \pm 0.99}$	$\mathbf{3.47 \pm 1.27}$	$\mathbf{15.84 \pm 3.86}$
He et al. (2018) (Utility)	3.56 ± 1.34	2.47 ± 1.39	3.09 ± 1.40	2.13 ± 1.13	11.25 ± 4.50
Lang. Model	4.06 ± 1.11	2.66 ± 1.36	3.63 ± 1.18	2.50 ± 1.10	12.84 ± 3.66

Table 2: Human evaluation scores comparing CHAI, He et al. (2018), and language model (higher is better). Numbers are reported as means and standard deviations over 32 trials. CHAI scores the highest across all metrics.

We have already covered a fundamental algorithm in class that can learn from offline data.

What is it?

Q-learning



For every (s_t, a_t, c_t, s_{t+1})

Can learn from any data!

$$Q^*(s_t, a_t) = Q^*(s_t, a_t) + \alpha(c(s_t, a_t) + \gamma \min_{a'} Q^*(s_{t+1}, a') - Q^*(s_t, a_t))$$

Q-learning

For every $\left(s_t, a_t, c_t, s_{t+1}\right)$

$$Q^*(s_t, a_t) = Q^*(s_t, a_t) + \alpha(c(s_t, a_t) + \gamma \min_{a'} Q^*(s_{t+1}, a') - Q^*(s_t, a_t)$$

Notice we are *not* approximating $Q^\pi(s_t, a_t)$

We don't even care about π

We can learn from any data!

Q-learning

For every $\left(s_t, a_t, c_t, s_{t+1}\right)$

$$Q^*(s_t, a_t) = Q^*(s_t, a_t) + \alpha(c(s_t, a_t) + \gamma \min_{a'} Q^*(s_{t+1}, a') - Q^*(s_t, a_t)$$

Conditions for convergence

1. Each state-action pair is visited infinite times
2. $\lim_{k \to \infty} \sum_{k=0}^{\infty} \alpha_k = \infty$
3. $\lim_{k \to \infty} \sum_{k=0}^{\infty} \alpha_k^2 < \infty$,

Q-learning

For every $\left(s_t, a_t, c_t, s_{t+1}\right)$

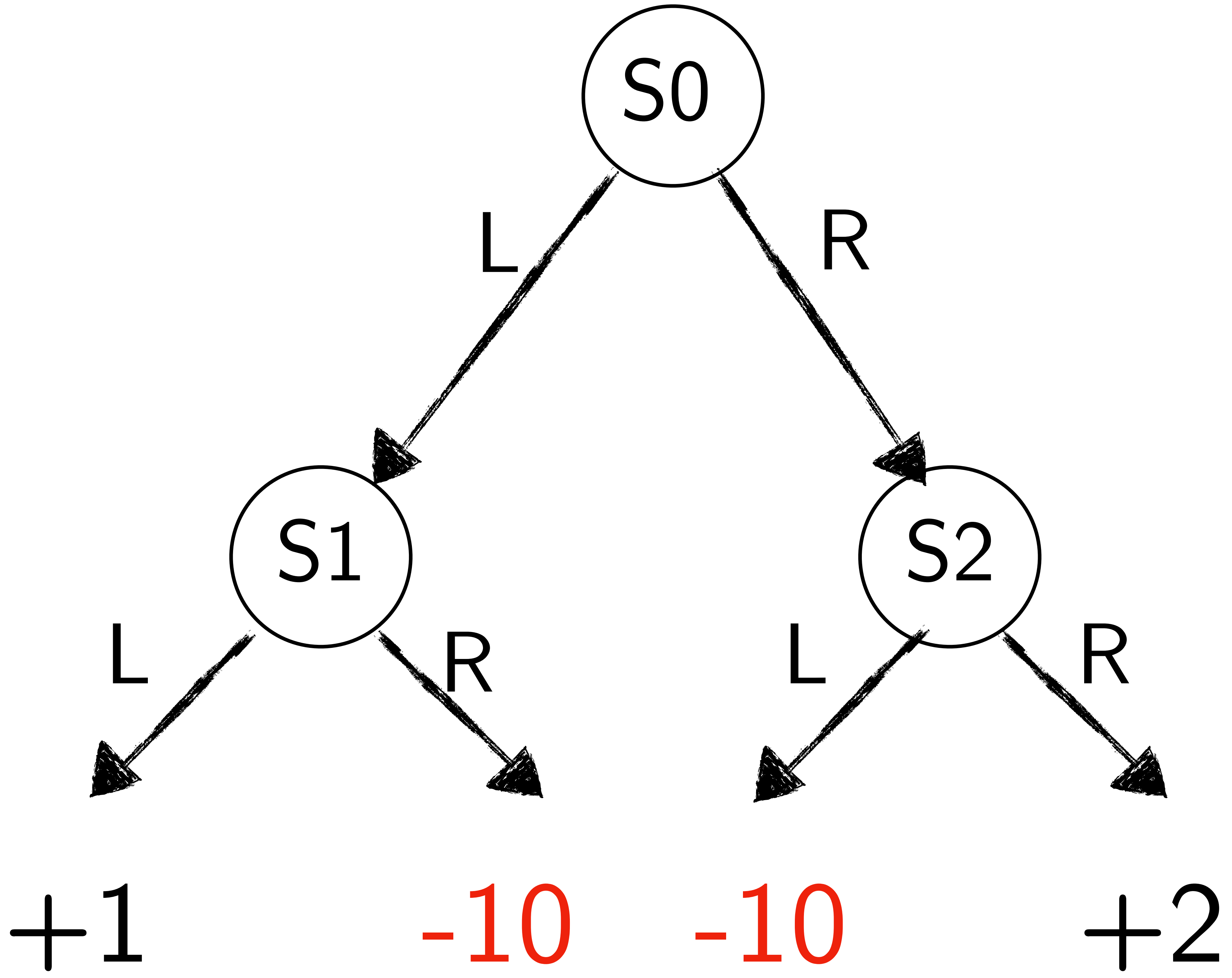
$$Q^*(s_t, a_t) = Q^*(s_t, a_t) + \alpha(c(s_t, a_t) + \gamma \min_{a'} Q^*(s_{t+1}, a') - Q^*(s_t, a_t)$$

What happens when this is not met?

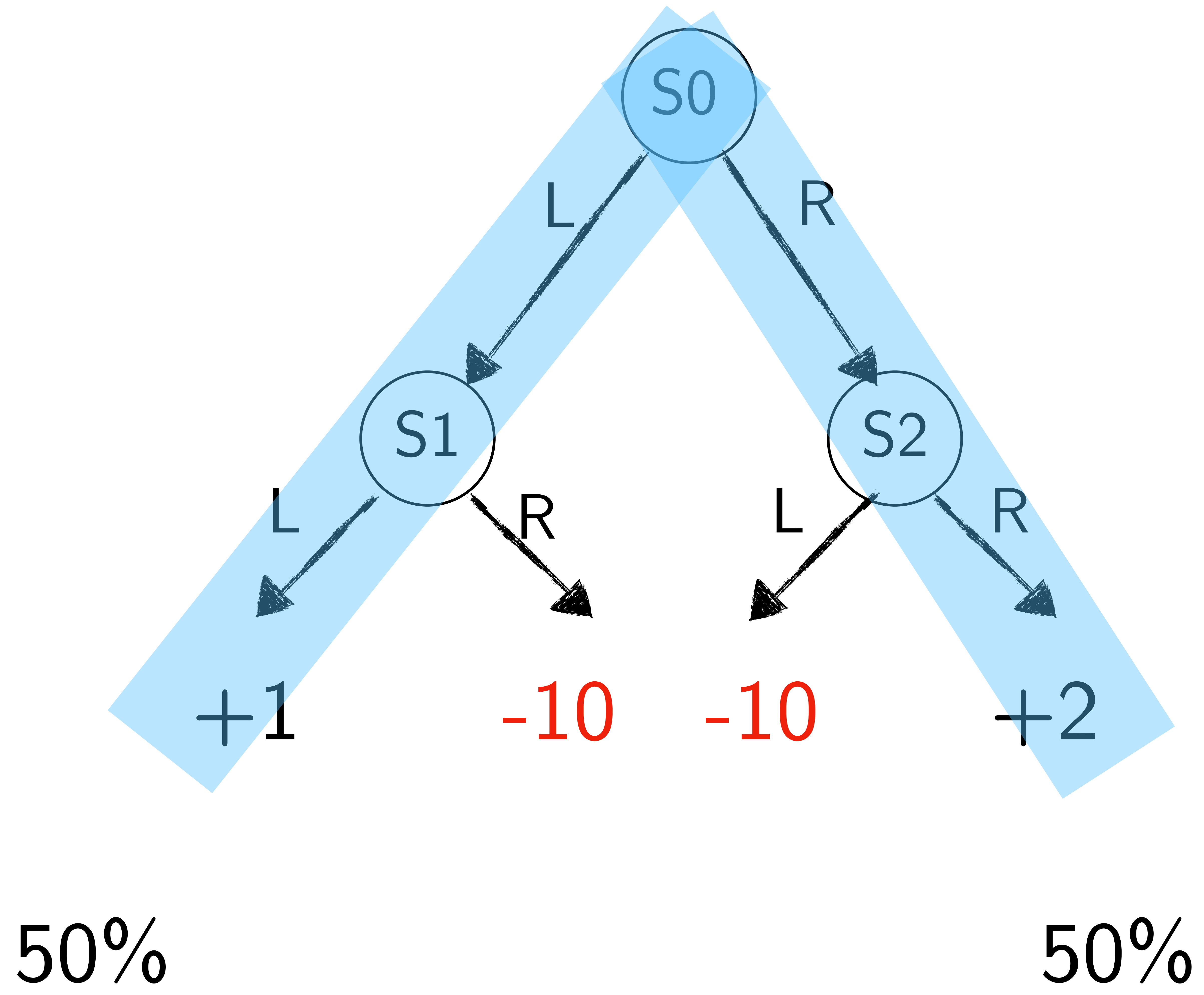
1. $\boxed{\text{Each state-action pair is visited infinite times}}$
2. $\lim_{k\to\infty} \sum_{k=0}^{\infty} \alpha_k = \infty$
3. $\lim_{k\to\infty} \sum_{k=0}^{\infty} \alpha_k^2 < \infty,$

Activity!

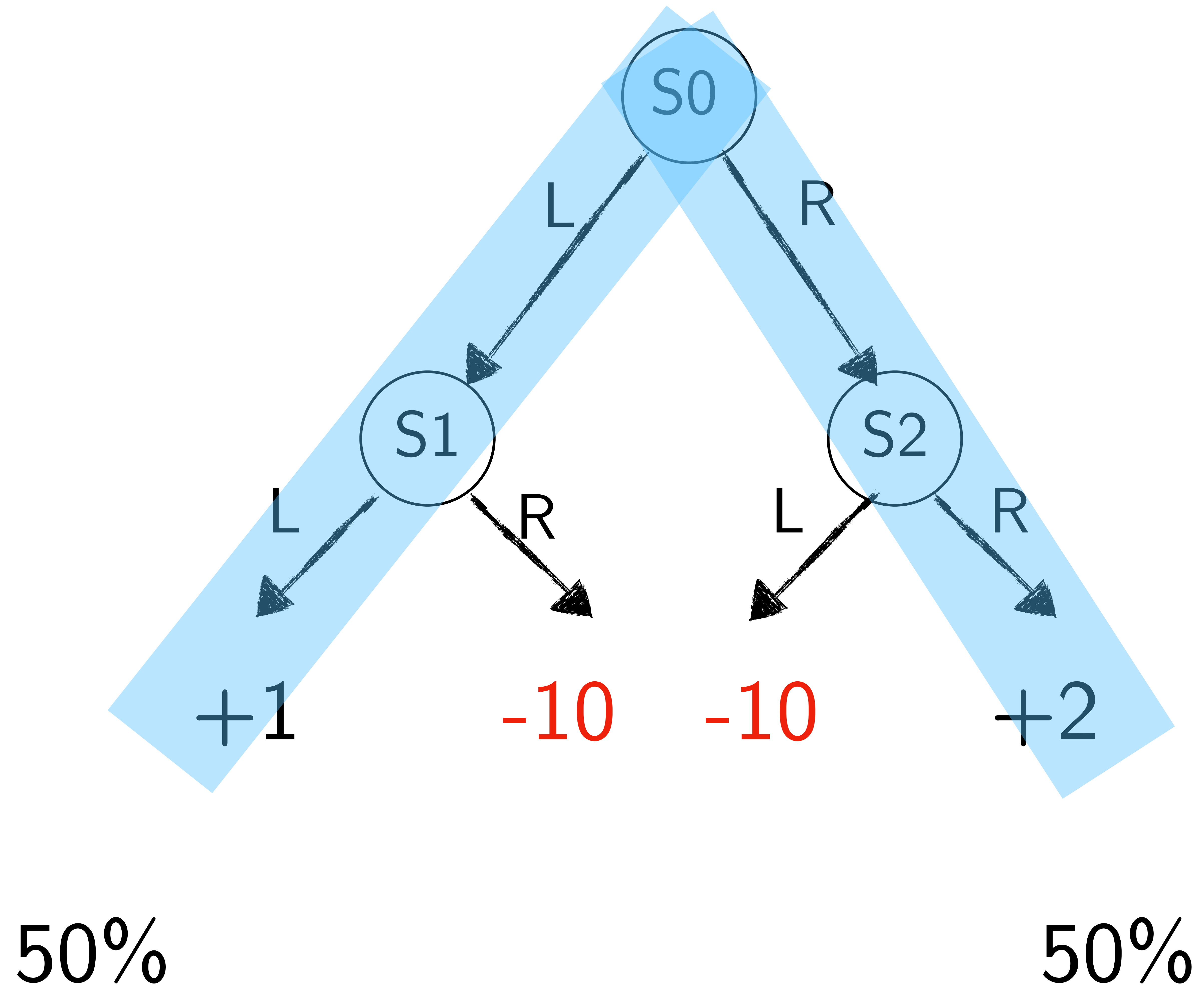
Consider the following MDP



Let's say I collected some data from the MDP



What would happen if I did Q-learning with this data?

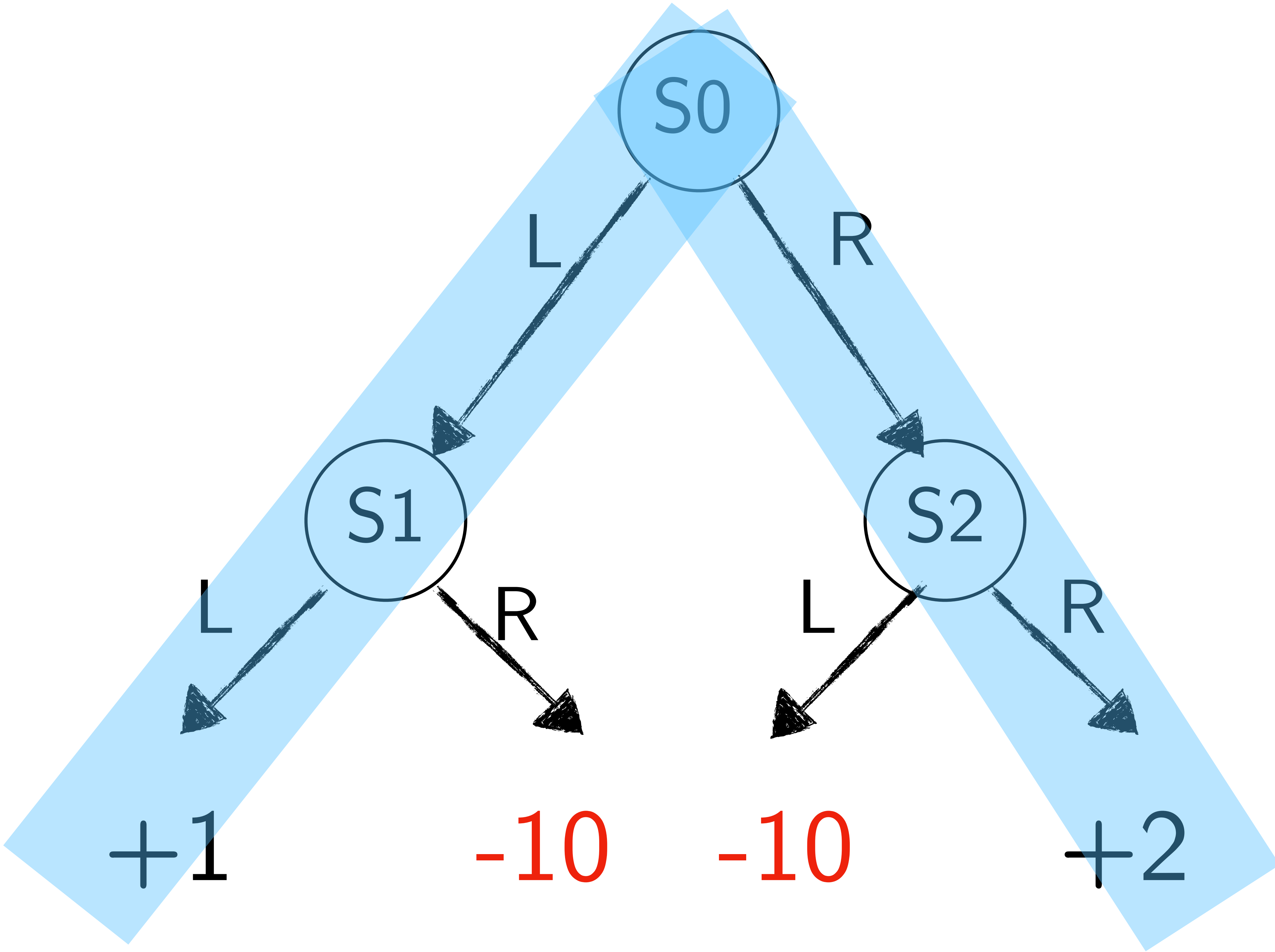


Think-Pair-Share!

Think (30 sec): What would happen if we did Q-learning with this data? Ideas on how to fix it.

Pair: Find a partner

Share (45 sec): Partners exchange ideas

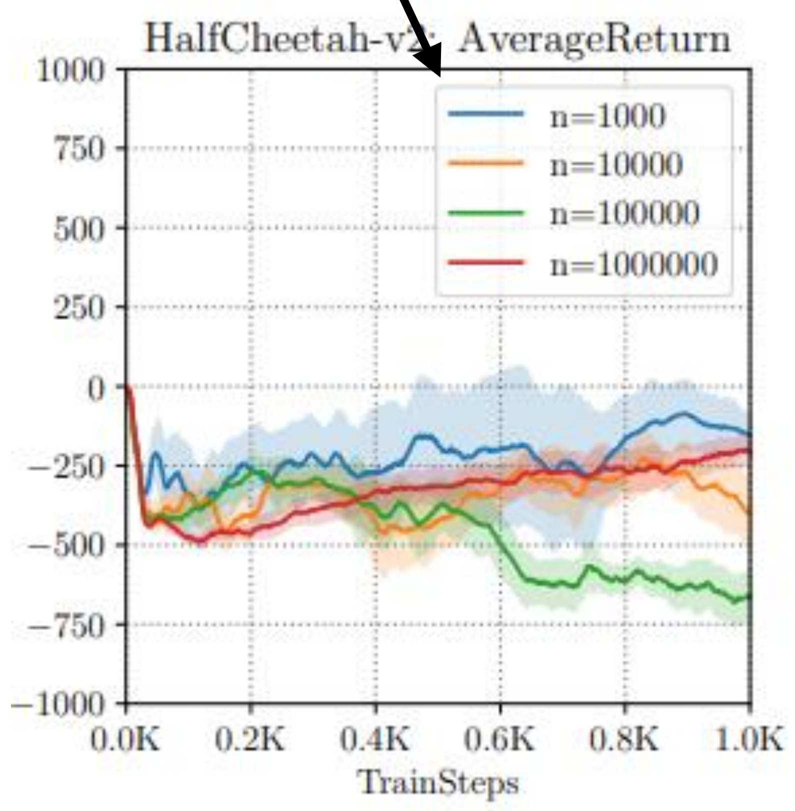


Credit: **Sergey Levine** "Offline RL lecture"

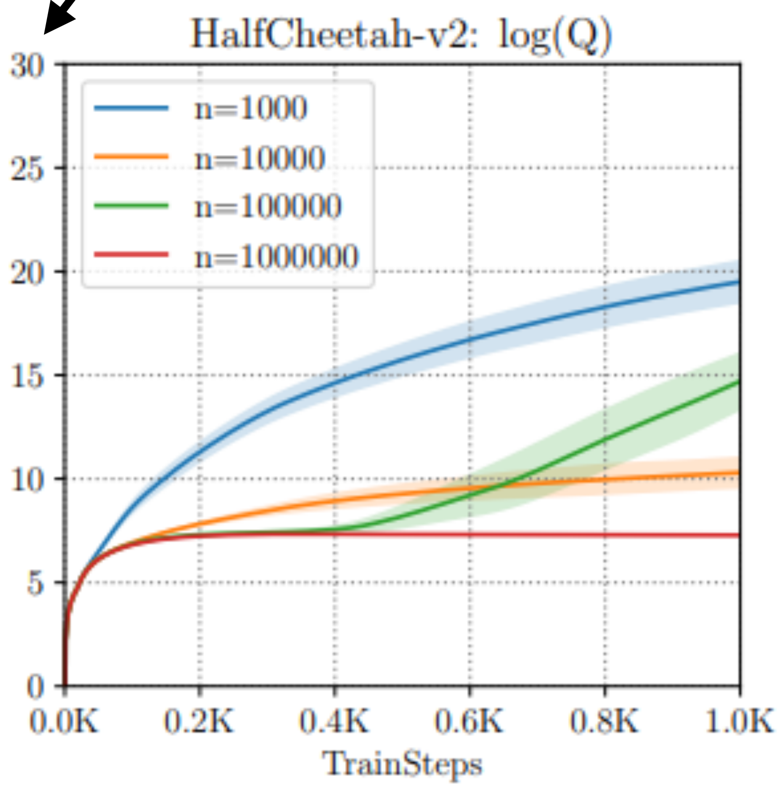
Why is offline RL hard?

amount of data

log scale (massive overestimation)



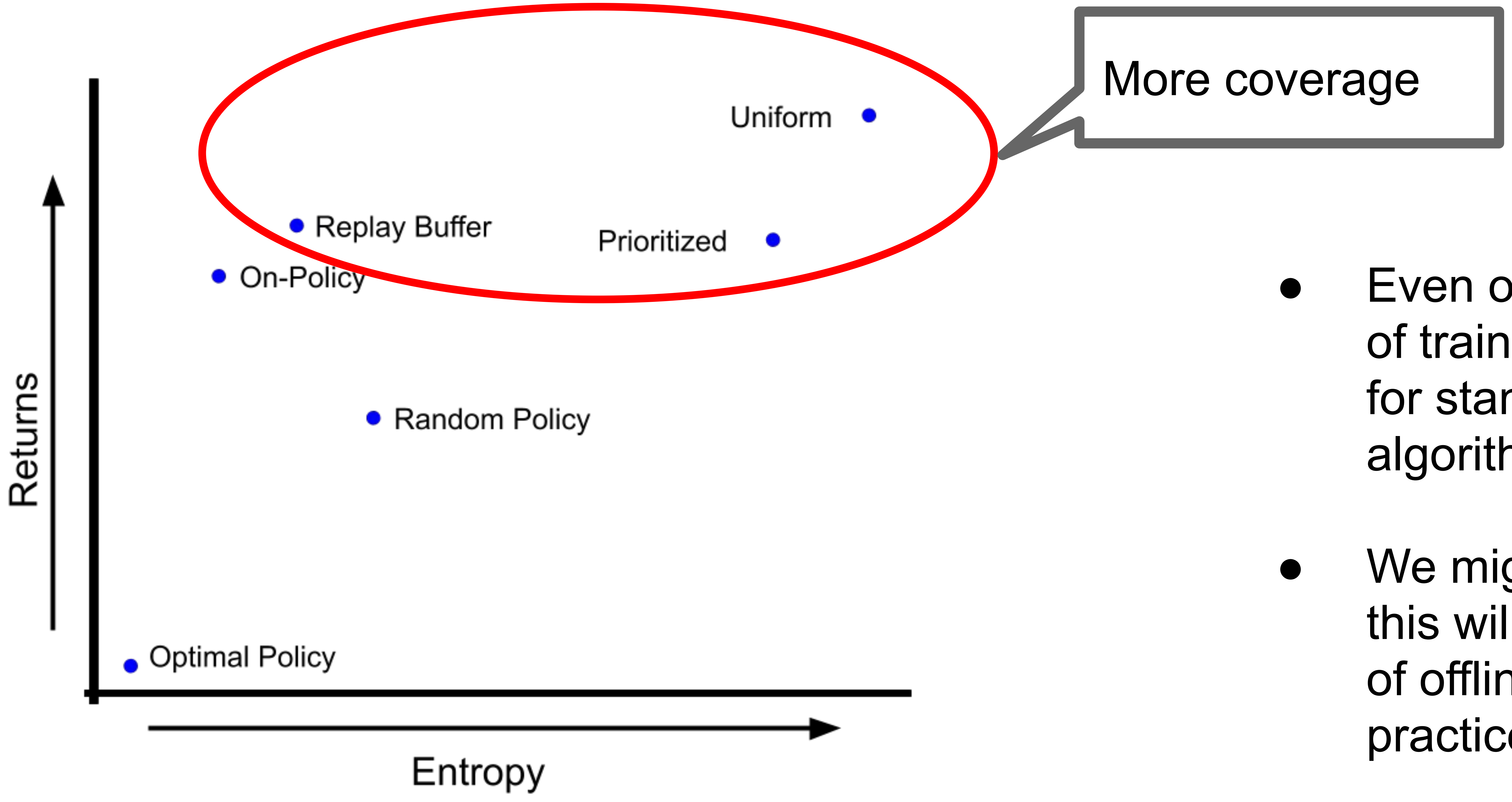
how well it does



how well it *thinks* it does (Q-values)

Kumar, Fu, Tucker, Levine. **Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction.** NeurIPS '19

Illustration of Distribution Shift



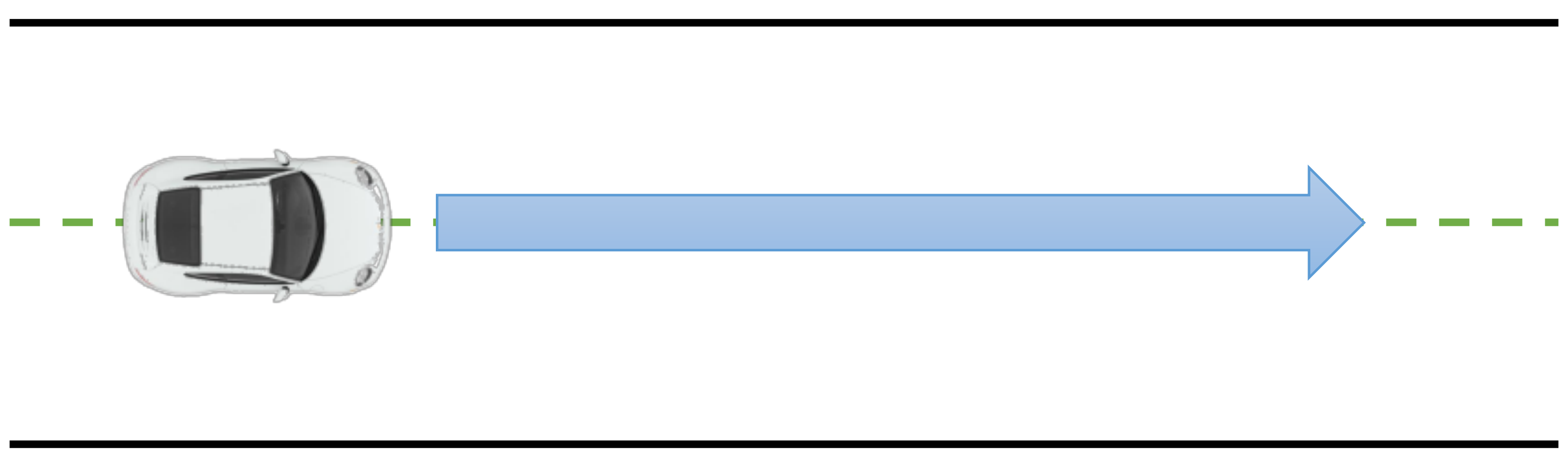
- Even on gridworlds, the choice of training distribution matters for standard Q-learning algorithms.
- We might only wonder how this will affect the performance of offline RL algorithms in practice.

Figure from *Diagnosing Bottlenecks in Deep Q-learning Algorithms*. **K.*,** Fu*, Levine. ICML 2019

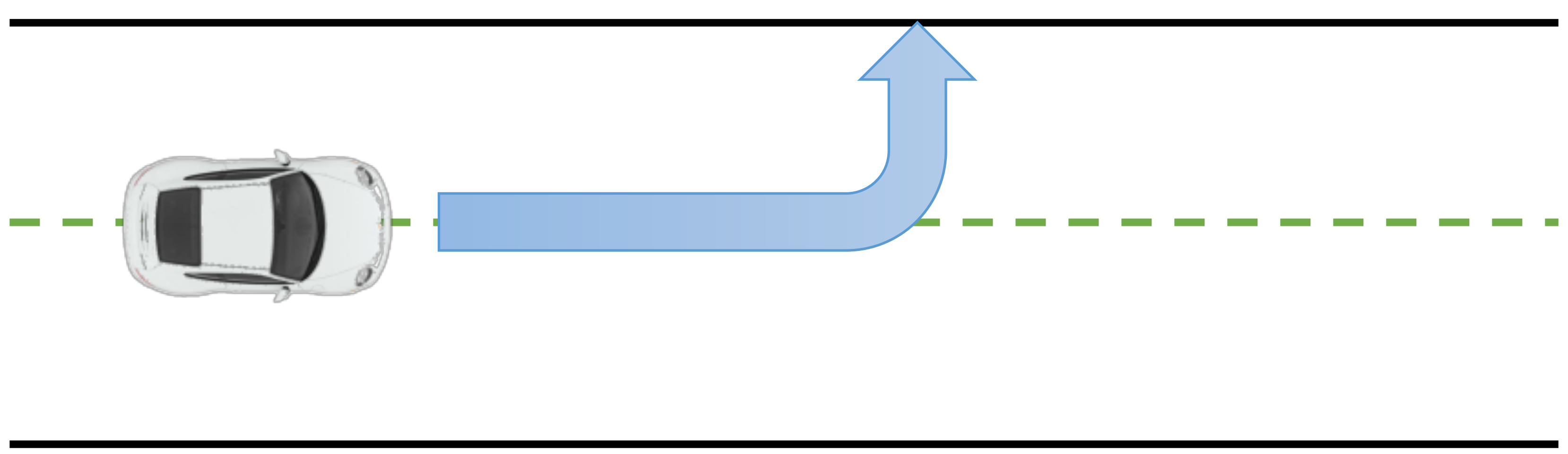
Why is offline RL hard?

Fundamental problem: counterfactual queries

Training data



What the policy wants to do



Is this good? Bad?

How do we know if we didn’t see it in the data?

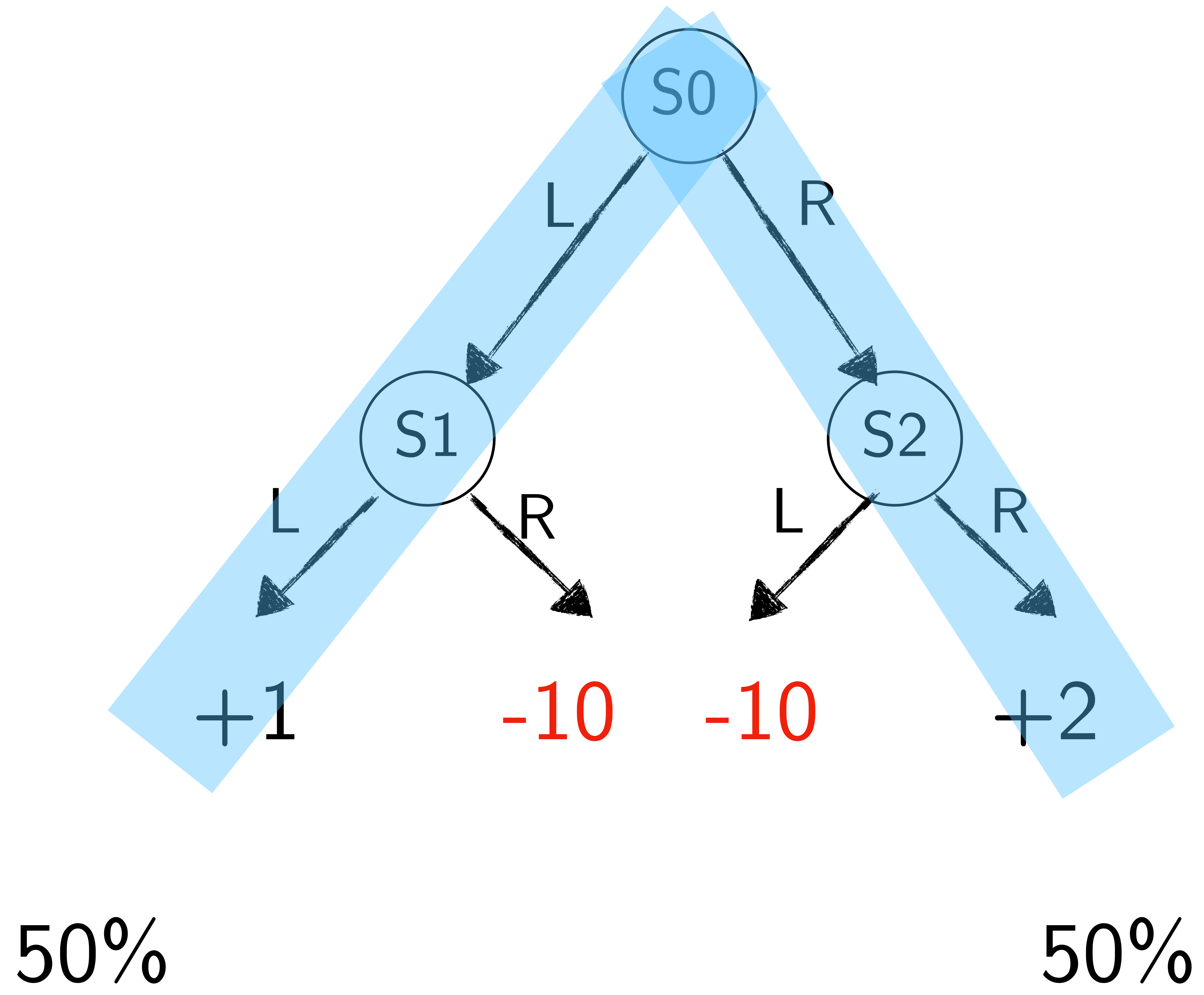
Online RL algorithms don’t have to handle this, because they can simply **try** this action and see what happens

Offline RL methods must somehow account for these unseen (“out-of-distribution”) actions, ideally in a safe way

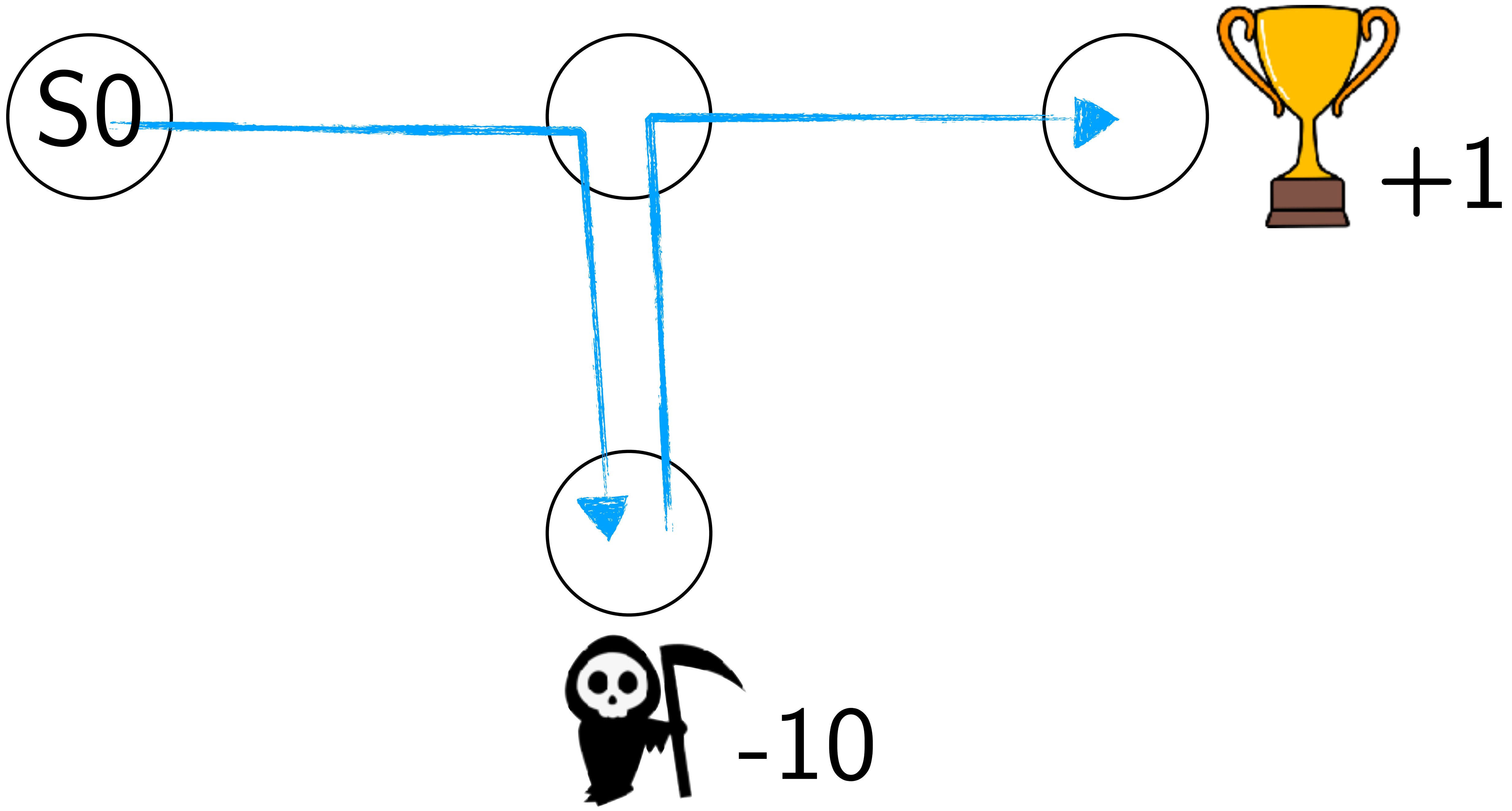
…while still making use of generalization to come up with behaviors that are better than the best thing seen in the data!

Levine, Kumar, Tucker, Fu. **Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems.** ‘20

Why not just do imitation learning?



Now consider this MDP



What is the optimal policy? What would imitation learning do?

Pessimism

Pessimism as a policy constraint

Don't deviate too much from the data collecting policy

$$Q(\mathbf{s}, \mathbf{a}) \leftarrow r(\mathbf{s}, \mathbf{a}) + E_{\mathbf{a}' \sim \pi_{\text{new}}}[Q(\mathbf{s}', \mathbf{a}')]$$

$$\pi_{\text{new}}(\mathbf{a}|\mathbf{s}) = \arg\max_{\pi} E_{\mathbf{a} \sim \pi(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s}, \mathbf{a})] \text{ s.t. } \boxed{D_{\text{KL}}(\pi \| \pi_{\beta}) \leq \epsilon}$$

Choose any divergence, e.g. KL!

Credit: **Sergey Levine** "Offline RL lecture"

TD3+BC: Most simple and effective offline RL!

A Minimalist Approach to Offline Reinforcement Learning

Scott Fujimoto[1,2] **Shixiang Shane Gu**[2]
[1]Mila, McGill University
[2]Google Research, Brain Team
scott.fujimoto@mail.mcgill.ca

$$\cancel{\pi = \operatorname*{argmax}_{\pi} \mathbb{E}_{(s,a)\sim\mathcal{D}}[Q(s,\pi(s))].}$$

$$\pi = \operatorname*{argmax}_{\pi} \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\lambda Q(s,\pi(s)) - (\pi(s)-a)^2\right],$$

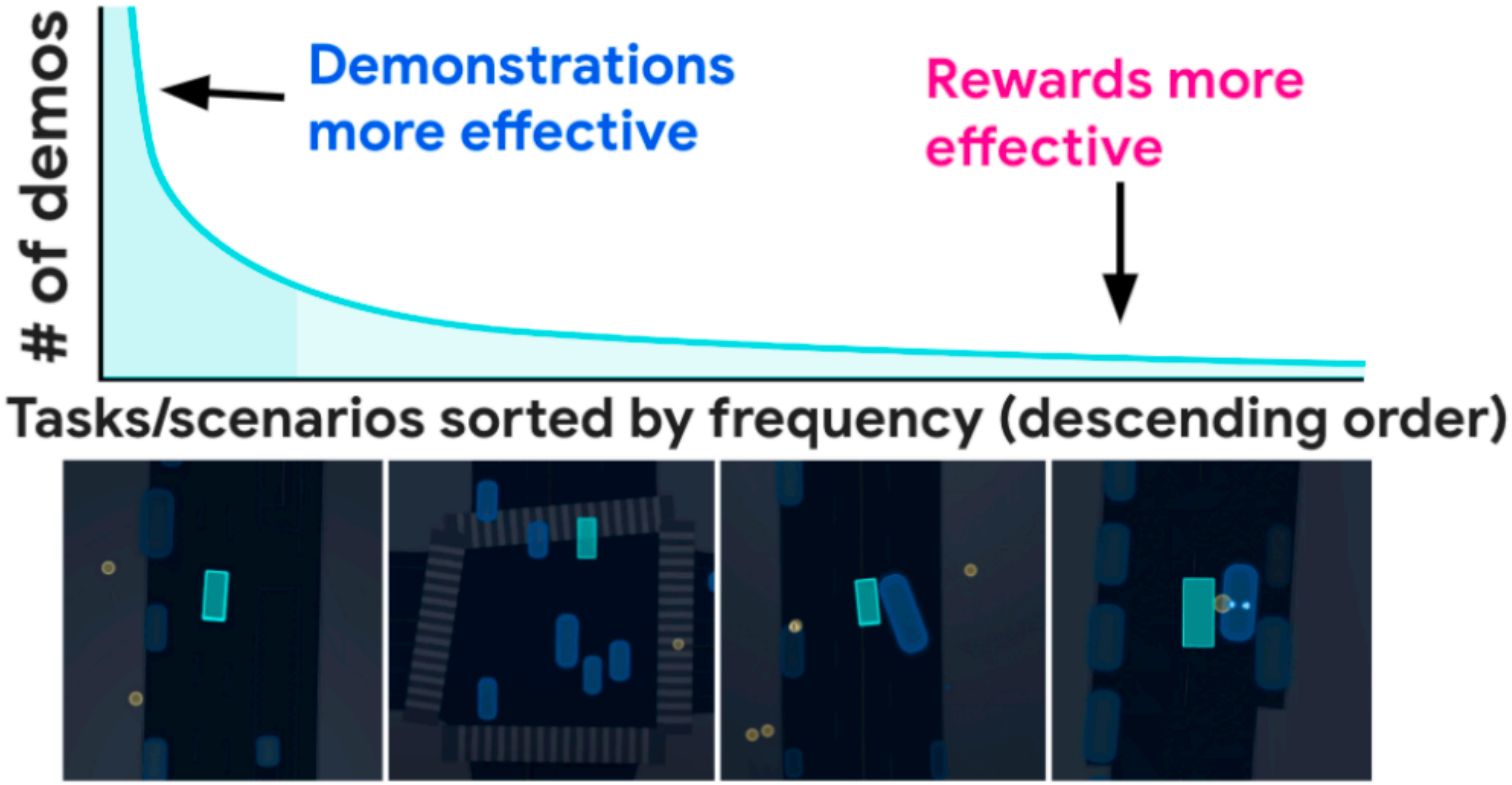
		BC	BRAC-p	AWAC	CQL	Fisher-BRC	TD3+BC
Random	HalfCheetah	2.0 ±0.1	23.5	2.2	21.7 ±0.9	32.2 ±2.2	10.2 ±1.3
	Hopper	9.5 ±0.1	11.1	9.6	10.7 ±0.1	11.4 ±0.2	11.0 ±0.1
	Walker2d	1.2 ±0.2	0.8	5.1	2.7 ±1.2	0.6 ±0.6	1.4 ±1.6
Medium	HalfCheetah	36.6 ±0.6	44.0	37.4	37.2 ±0.3	41.3 ±0.5	42.8 ±0.3
	Hopper	30.0 ±0.5	31.2	72.0	44.2 ±10.8	99.4 ±0.4	99.5 ±1.0
	Walker2d	11.4 ±6.3	72.7	30.1	57.5 ±8.3	79.5 ±1.0	79.7 ±1.8
Medium Replay	HalfCheetah	34.7 ±1.8	45.6	-	41.9 ±1.1	43.3 ±0.9	43.3 ±0.5
	Hopper	19.7 ±5.9	0.7	-	28.6 ±0.9	35.6 ±2.5	31.4 ±3.0
	Walker2d	8.3 ±1.5	-0.3	-	15.8 ±2.6	42.6 ±7.0	25.2 ±5.1
Medium Expert	HalfCheetah	67.6 ±13.2	43.8	36.8	27.1 ±3.9	96.1 ±9.5	97.9 ±4.4
	Hopper	89.6 ±27.6	1.1	80.9	111.4 ±1.2	90.6 ±43.3	112.2 ±0.2
	Walker2d	12.0 ±5.8	-0.3	42.7	68.1 ±13.1	103.6 ±4.6	101.1 ±9.3
Expert	HalfCheetah	105.2 ±1.7	3.8	78.5	82.4 ±7.4	106.8 ±3.0	105.7 ±1.9
	Hopper	111.5 ±1.3	6.6	85.2	111.2 ±2.1	112.3 ±0.2	112.2 ±0.2
	Walker2d	56.0 ±24.9	-0.2	57.0	103.8 ±7.6	79.9 ±32.4	105.7 ±2.7
	Total	595.3 ±91.5	284.1	-	764.3 ±61.5	974.6 ±108.3	979.3 ±33.4

Table 2: Average normalized score over the final 10 evaluations and 5 seeds. The highest performing scores are highlighted. CQL and Fisher-BRC are re-run using author-provided implementations to ensure an identical evaluation process, while BRAC and AWAC use previously reported results. ± captures the standard deviation over seeds. TD3+BC achieves effectively the same performances as the state-of-the-art Fisher-BRC, despite being much simpler to implement and tune and more than halving the computation cost.

Works on real self-driving problems!

Imitation Is Not Enough: Robustifying Imitation with Reinforcement Learning for Challenging Driving Scenarios

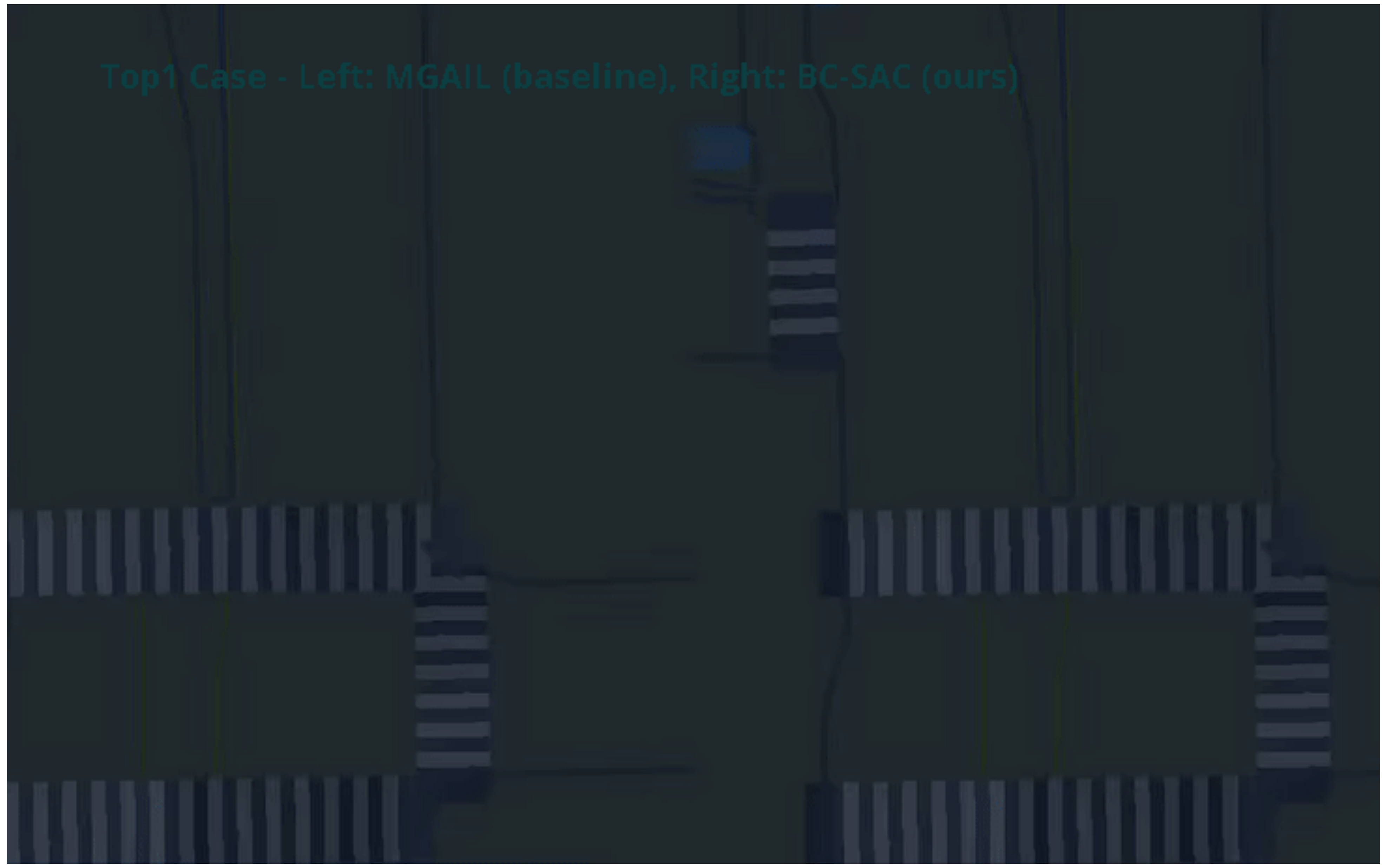
Yiren Lu[1], Justin Fu[1], George Tucker[2], Xinlei Pan[1], Eli Bronstein[1], Rebecca Roelofs[2], Benjamin Sapp[1], Brandyn White[1], Aleksandra Faust[2], Shimon Whiteson[1], Dragomir Anguelov[1], Sergey Levine[2,3]



Works on real self-driving problems!

Imitation Is Not Enough: Robustifying Imitation with Reinforcement Learning for Challenging Driving Scenarios

Yiren Lu[1], Justin Fu[1], George Tucker[2], Xinlei Pan[1], Eli Bronstein[1], Rebecca Roelofs[2], Benjamin Sapp[1], Brandyn White[1], Aleksandra Faust[2], Shimon Whiteson[1], Dragomir Anguelov[1], Sergey Levine[2,3]



https://waymo.com/research/imitation-is-not-enough-robustifying-imitation-with-reinforcement-learning/

But choosing a divergence seems arbitrary?

$$Q(\mathbf{s},\mathbf{a}) \leftarrow r(\mathbf{s},\mathbf{a}) + E_{\mathbf{a}'\sim\pi_{\text{new}}}[Q(\mathbf{s}',\mathbf{a}')]$$

$$\pi_{\text{new}}(\mathbf{a}|\mathbf{s}) = \arg\max_{\pi} E_{\mathbf{a}\sim\pi(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s},\mathbf{a})] \text{ s.t. } \boxed{D_{\text{KL}}(\pi\|\pi_\beta) \leq \epsilon}$$

Another notion of pessimism

Can we make the Q-value itself pessimistic on actions it has not seen?

Conservative Q-Learning (CQL)

Conservative Q-Learning for Offline Reinforcement Learning

Aviral Kumar[1], **Aurick Zhou**[1], **George Tucker**[2], **Sergey Levine**[1,2]
[1]UC Berkeley, [2]Google Research, Brain Team
aviralk@berkeley.edu

Conservative Q-Learning (CQL)

1. Learn $\hat{Q}^{\pi}$ using offline data $\mathcal{D}$.
2. Optimize policy w.r.t. $\hat{Q}^{\pi}$: $\pi \leftarrow \arg\max_{\pi} \mathbb{E}_{\pi}[\hat{Q}^{\pi}] - \alpha D(\pi_{\phi}, \pi_{\beta})$

Approach 2: Directly modify the Q-function to be pessimistic

- Key idea behind CQL: Learn lower-bounds on Q-values

CQL Algorithm:

1. Learn $\hat{Q}^{\pi}_{\text{CQL}}$ using offline data $\mathcal{D}$.
2. Optimize policy w.r.t. $\hat{Q}^{\pi}_{\text{CQL}}$: $\pi \leftarrow \arg\max_{\pi} \mathbb{E}_{\pi}[\hat{Q}^{\pi}_{\text{CQL}}]$.

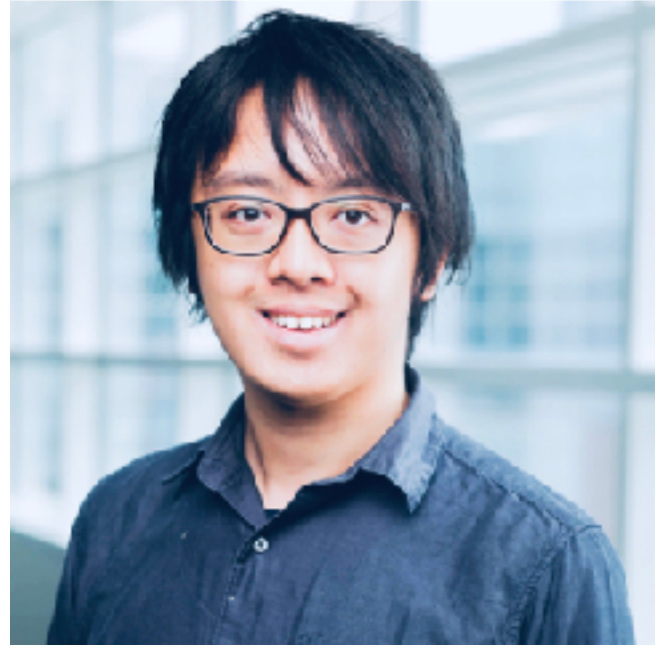
Credit: **Aviral Kumar** "Conservative Q learning"

Many ways to construct a conservative Q value

Original CQL paper proposed one such way

Recent work has come up with more unified frameworks

Adversarially Trained Actor Critic (ATACL)



Adversarially Trained Actor Critic for Offline Reinforcement Learning

Ching-An Cheng [*1] **Tengyang Xie** [*2] **Nan Jiang** [2] **Alekh Agarwal** [3]

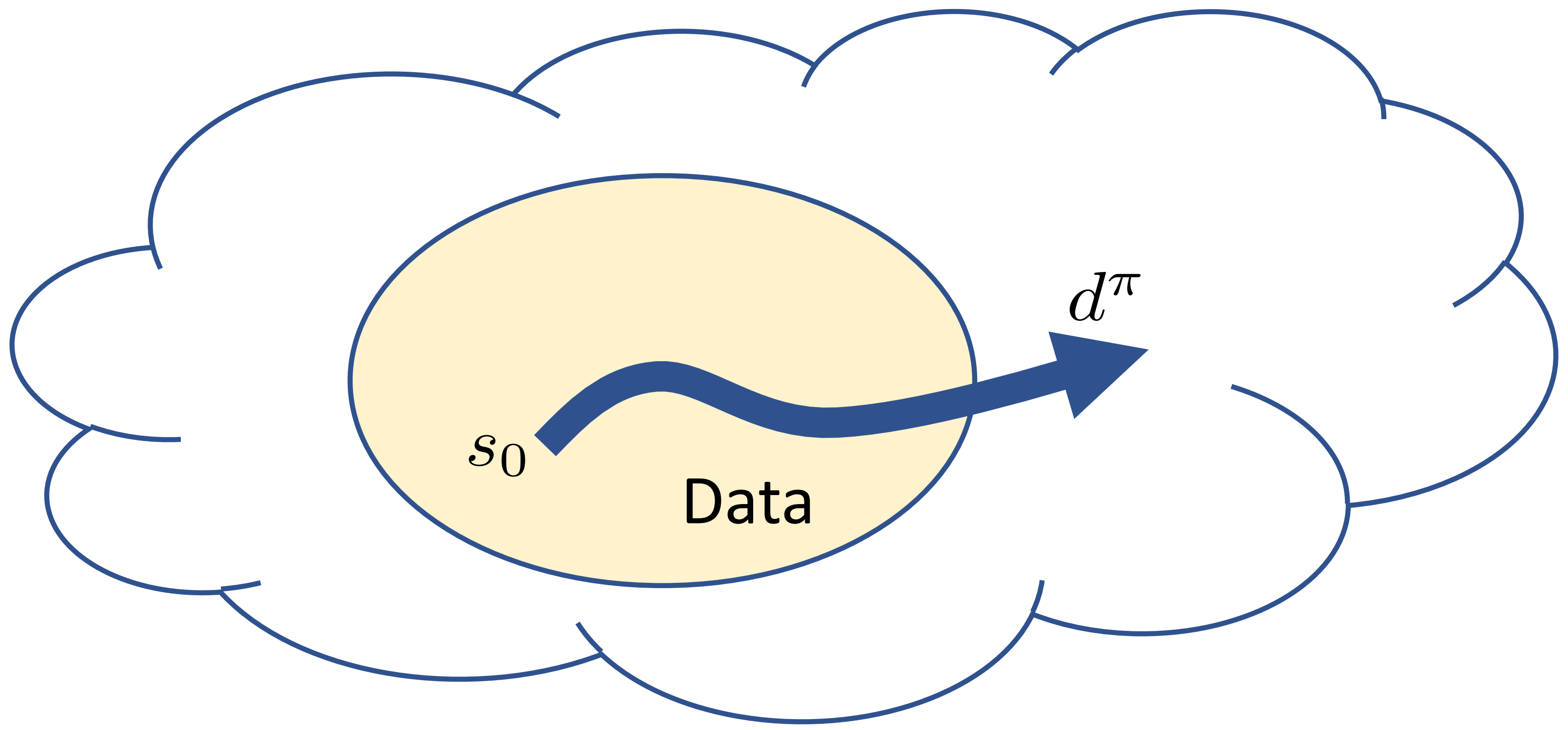
Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

Key Idea: Relative Pessimism

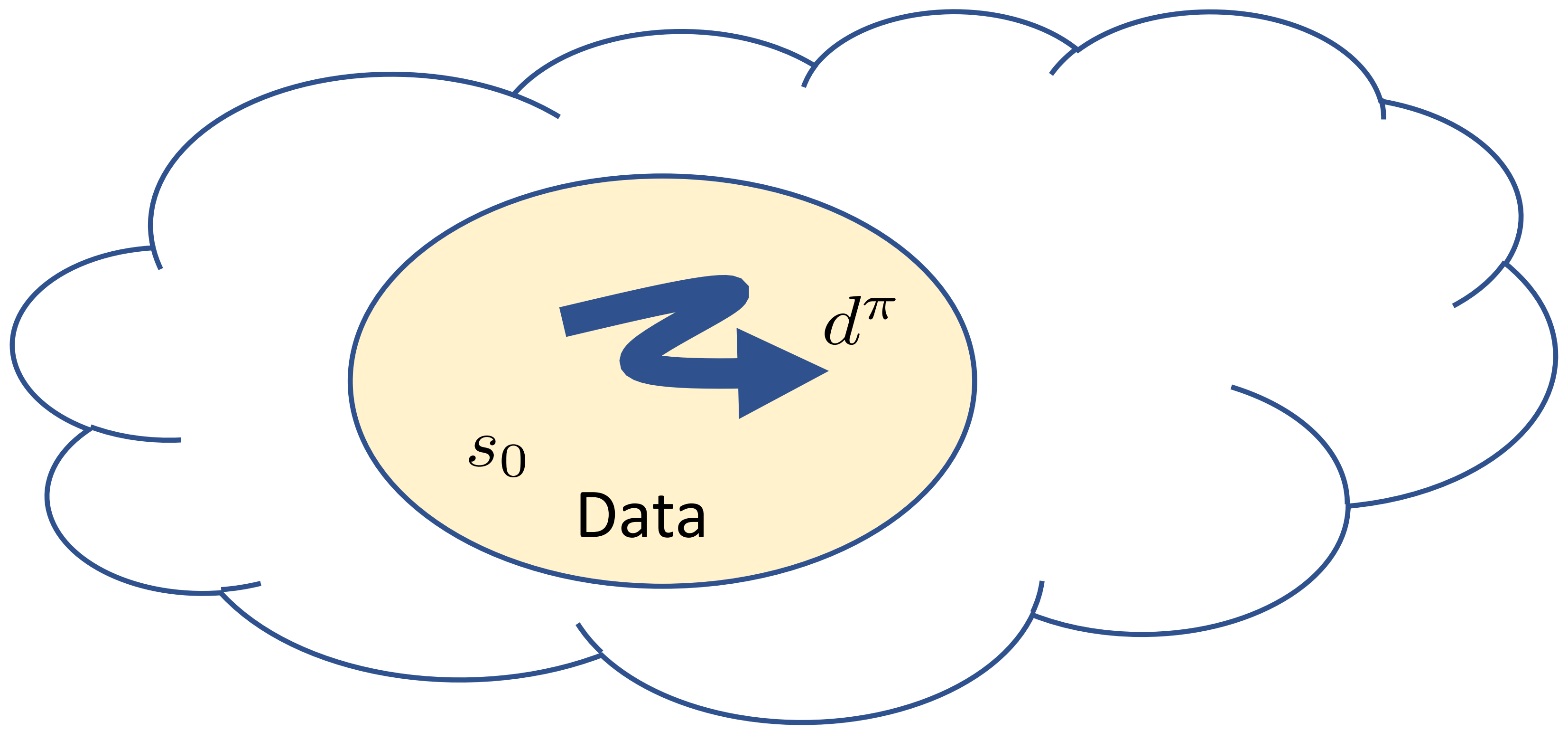
- Optimize for the **worst-case** performance compared with the **behavior policy $\boldsymbol{\mu}$.**

$$\widehat{\pi}^* \in \arg\max_{\pi \in \Pi} \boxed{\text{Lower bound of } J(\pi) - J(\mu)}$$

Lower bound $< J(\pi) - J(\mu)$



Lower bound $\approx J(\pi) - J(\mu)$



Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

Key Idea: Relative Pessimism

- Optimize for the **best worst-case** performance compared with the **behavior policy μ.**

$$\widehat{\pi}^* \in \arg\max_{\pi \in \Pi} \boxed{\text{Lower bound of } J(\pi) - J(\mu)}$$

ATAC frames this problem as a Stackelberg game (i.e., bilevel optimization)

Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

A Stackelberg Game for Offline RL

- ATAC optimizes for relative pessimism via solving a Stackelberg game

Leader (policy)

$$\widehat{\pi}^* \in \arg\max_{\pi \in \Pi} \boxed{\mathcal{L}_\mu(\pi, f^\pi)}$$

Follower (critic) s.t.

$$f^\pi \in \arg\min_{f \in \mathcal{F}} \mathcal{L}_\mu(\pi, f) + \beta \mathcal{E}_\mu(\pi, f)$$

Leader plays first then the Follower

Bilinear Payoffs of relative pessimism

Bellman error

Trade-off conservatism vs. generalization

$$\mathcal{L}_\mu(\pi, f) := \mathbb{E}_\mu[f(s, \pi) - f(s, a)]$$
$$\mathcal{E}_\mu(\pi, f) := \mathbb{E}_\mu[((f - \mathcal{T}^\pi f)(s, a))^2].$$

Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

A Stackelberg Game for Offline RL

- ATAC optimizes for relative pessimism via solving a Stackelberg game

Leader (policy)
$$\widehat{\pi}^* \in \arg\max_{\pi \in \Pi} \boxed{\mathcal{L}_\mu(\pi, f^\pi)} \geq \mathcal{L}(\pi, Q^\pi) \equiv J(\pi) - J(\mu), \forall \beta \geq 0$$

Follower (critic)
$$\text{s.t.} \quad f^\pi \in \arg\min_{f \in \mathcal{F}} \mathcal{L}_\mu(\pi, f) + \beta \mathcal{E}_\mu(\pi, f)$$

Leader plays first then the Follower

Bilinear Payoffs of relative pessimism

Bellman error

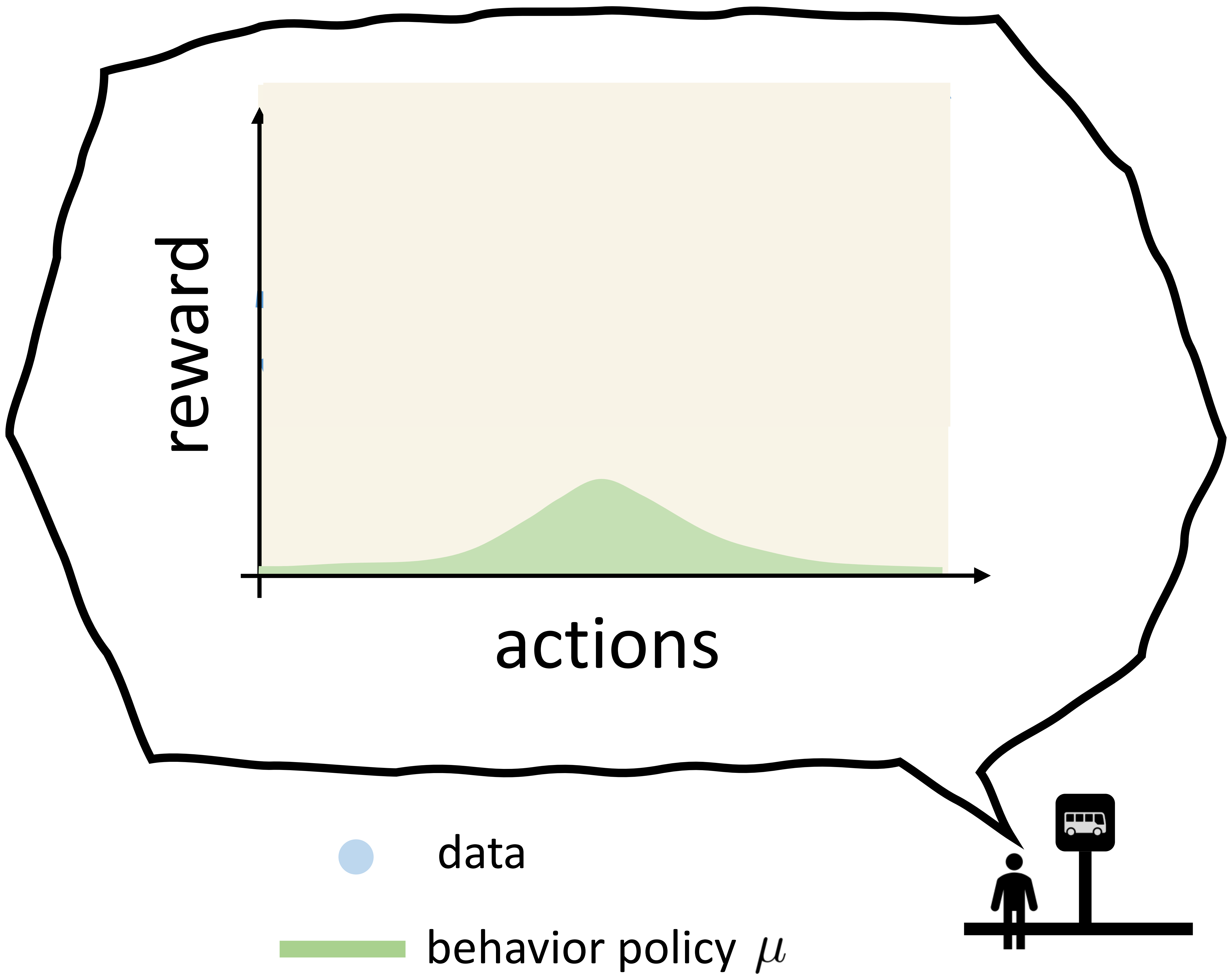
Trade-off conservatism vs. generalization

Robust Policy Improvement Property

For all $\beta \geq 0$, the ATAC policy is always no worse than the behavior policy that collected the data.

$$\mathcal{L}_\mu(\pi, f) := \mathbb{E}_\mu[f(s, \pi) - f(s, a)]$$
$$\mathcal{E}_\mu(\pi, f) := \mathbb{E}_\mu[((f - \mathcal{T}^\pi f)(s, a))^2].$$

Let's look at a simple example

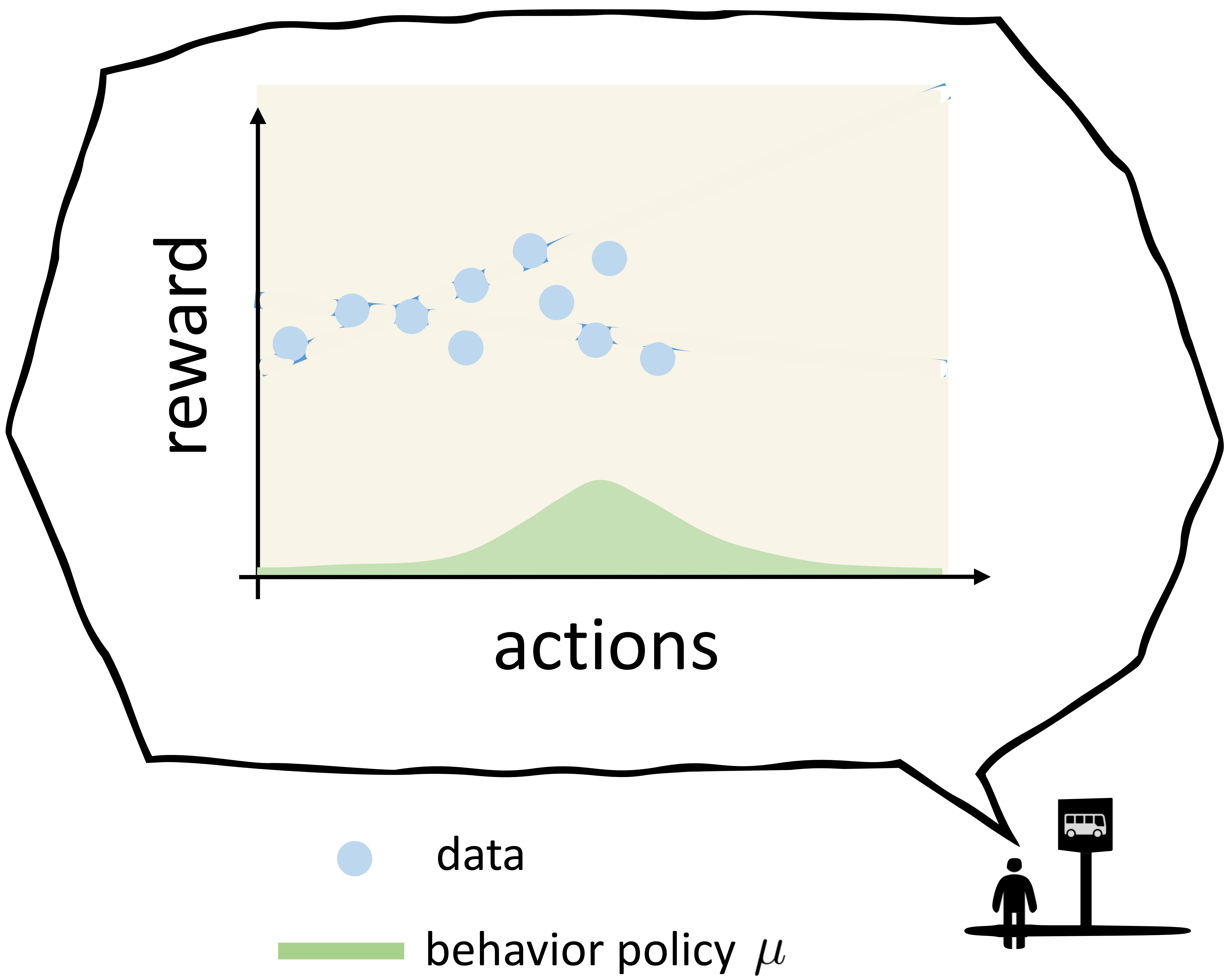


Let's say the time horizon T=1

(Multi-armed bandit!)

Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

Let's look at a simple example

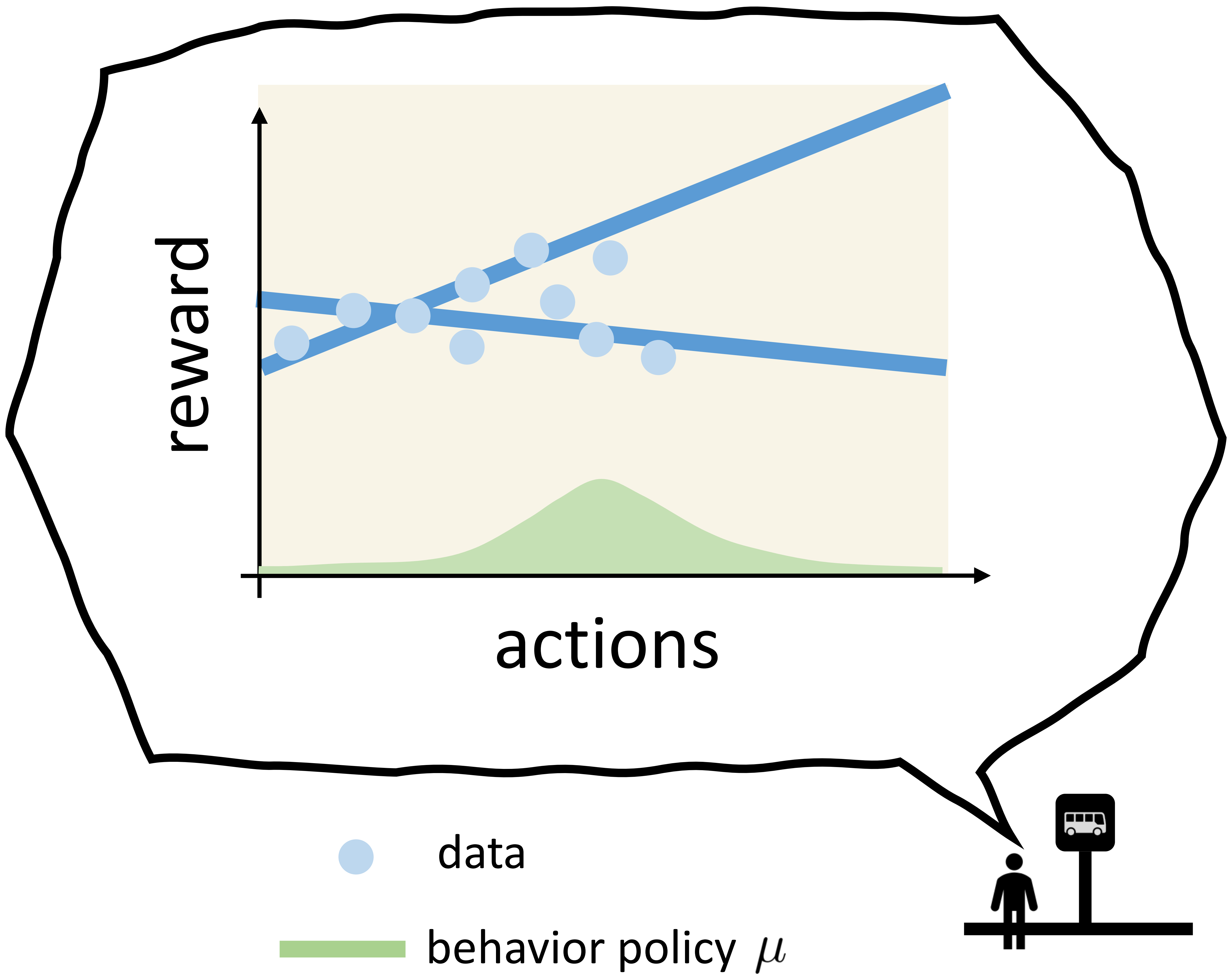


Let's say the time horizon T=1

(Multi-armed bandit!)

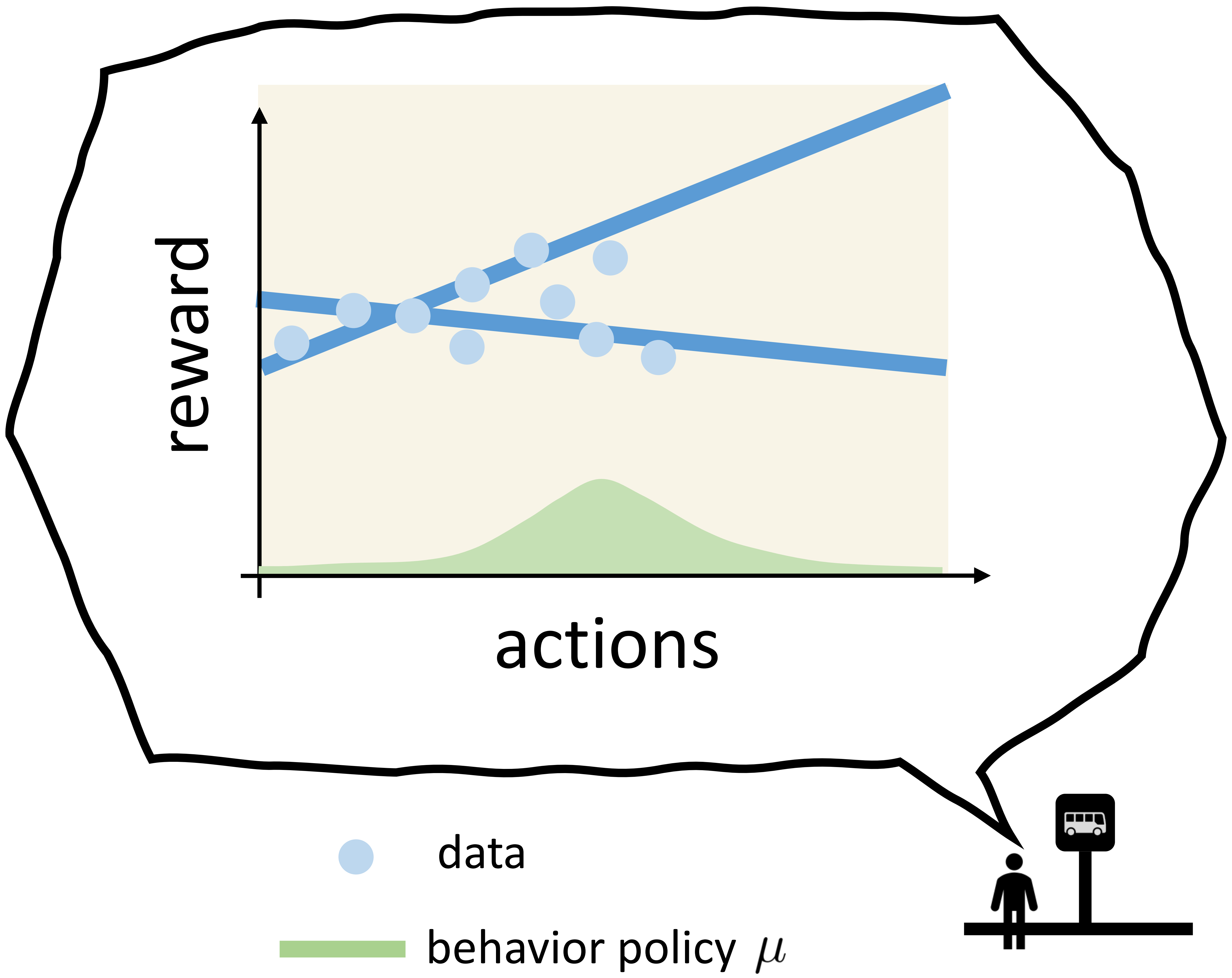
Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

Let's look at a simple example



Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

A Stackelberg Game for Offline RL



ATAC

$$\widehat{\pi}^* \in \arg\max_{\pi \in \Pi} \mathcal{L}_\mu(\pi, f^\pi)$$

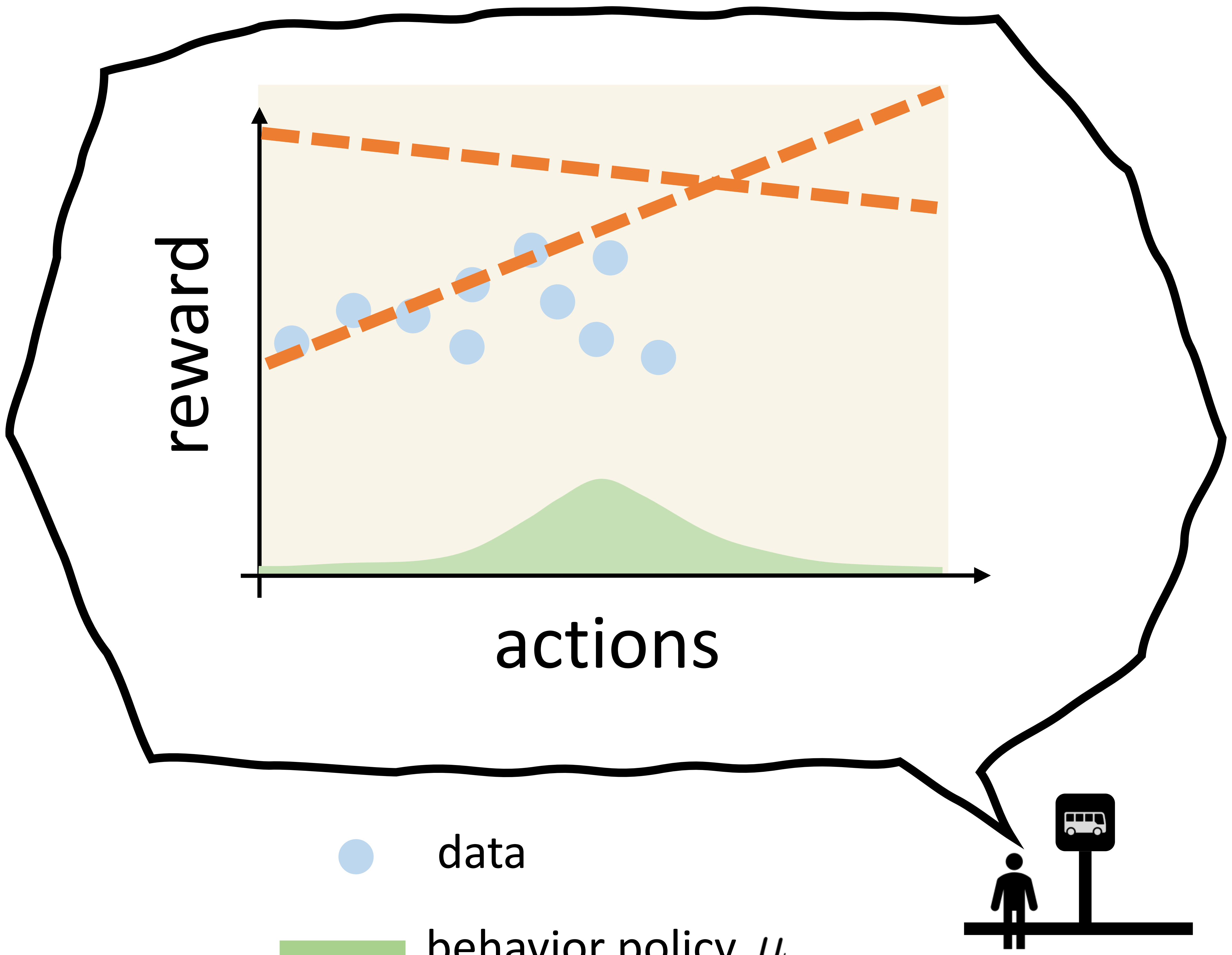
$$\text{s.t.} \quad \boxed{f^\pi} \in \arg\min_{f \in \mathcal{F}} \mathcal{L}_\mu(\pi, f) + \beta \mathcal{E}_\mu(\pi, f)$$

hypothesis $f(s, \cdot)$ with small $\beta\mathcal{E}_\mu$

Functions that are consistent with the reward and the dynamics on the behavior data

Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

A Stackelberg Game for Offline RL



ATAC

$$\widehat{\pi}^* \in \arg\max_{\pi\in\Pi} \boxed{\mathcal{L}_\mu(\pi, f^\pi)}$$
$$\text{s.t.} \quad f^\pi \in \arg\min_{f\in\mathcal{F}} \mathcal{L}_\mu(\pi, f) + \beta\mathcal{E}_\mu(\pi, f)$$

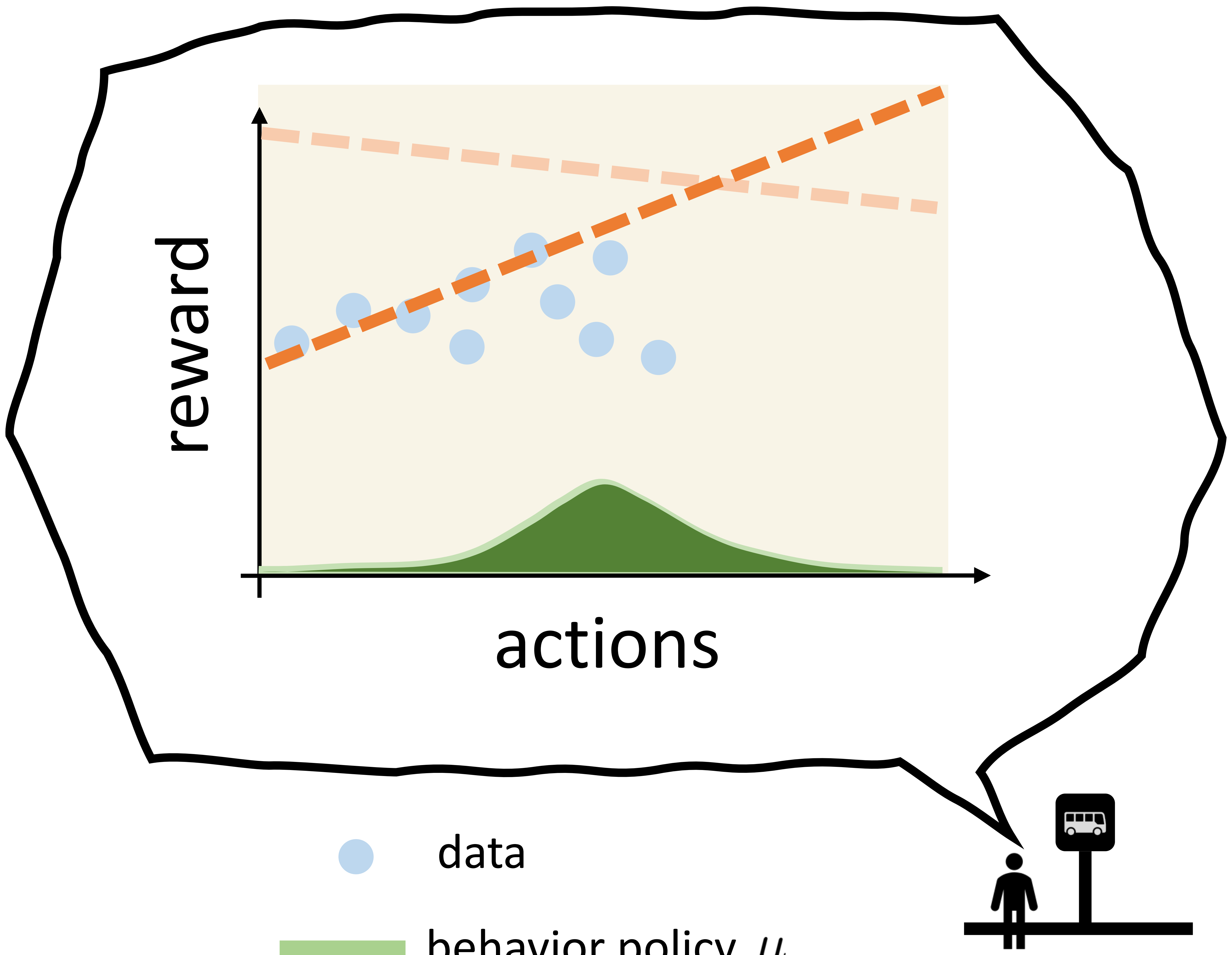
value difference hypothesis $f(s,\cdot) - f(s,\mu)$

Functions shifted from the original hypotheses

What is the solution to the Stackelberg game?

Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

A Stackelberg Game for Offline RL



data

behavior policy μ

ATAC

$$\widehat{\pi}^* \in \arg\max_{\pi\in\Pi} \boxed{\mathcal{L}_\mu(\pi, f^\pi)}$$
$$\text{s.t.} \quad f^\pi \in \arg\min_{f\in\mathcal{F}} \mathcal{L}_\mu(\pi, f) + \beta\mathcal{E}_\mu(\pi, f)$$

value difference hypothesis $f(s,\cdot) - f(s,\mu)$

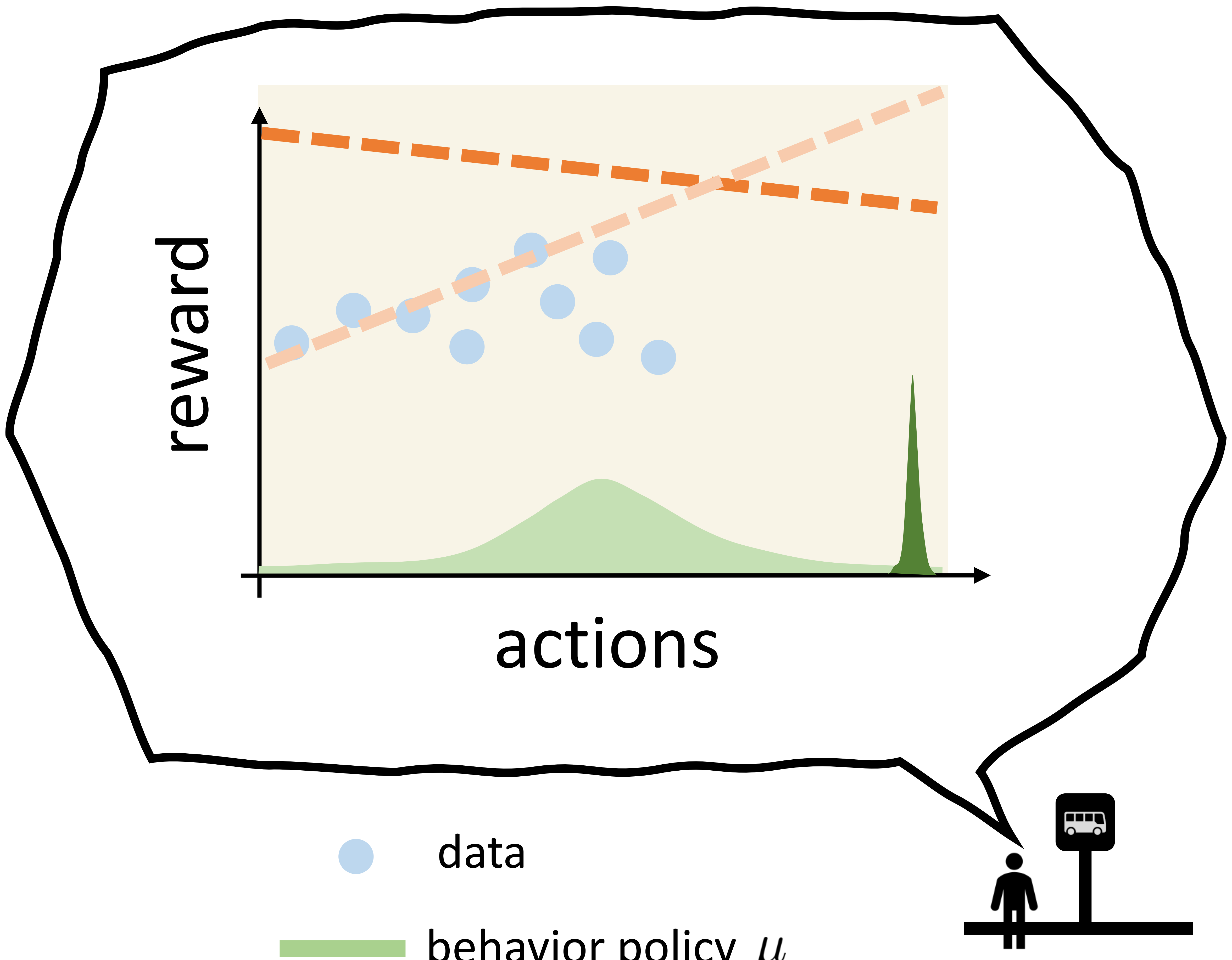
inactive value difference hypothesis

decision policy π

Not the behavior policy in this case…

Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

A Stackelberg Game for Offline RL



ATAC

$$\widehat{\pi}^* \in \arg\max_{\pi\in\Pi} \boxed{\mathcal{L}_\mu(\pi, f^\pi)}$$

$$\text{s.t.} \quad f^\pi \in \arg\min_{f\in\mathcal{F}} \mathcal{L}_\mu(\pi, f) + \beta\mathcal{E}_\mu(\pi, f)$$

value difference hypothesis $f(s,\cdot) - f(s,\mu)$

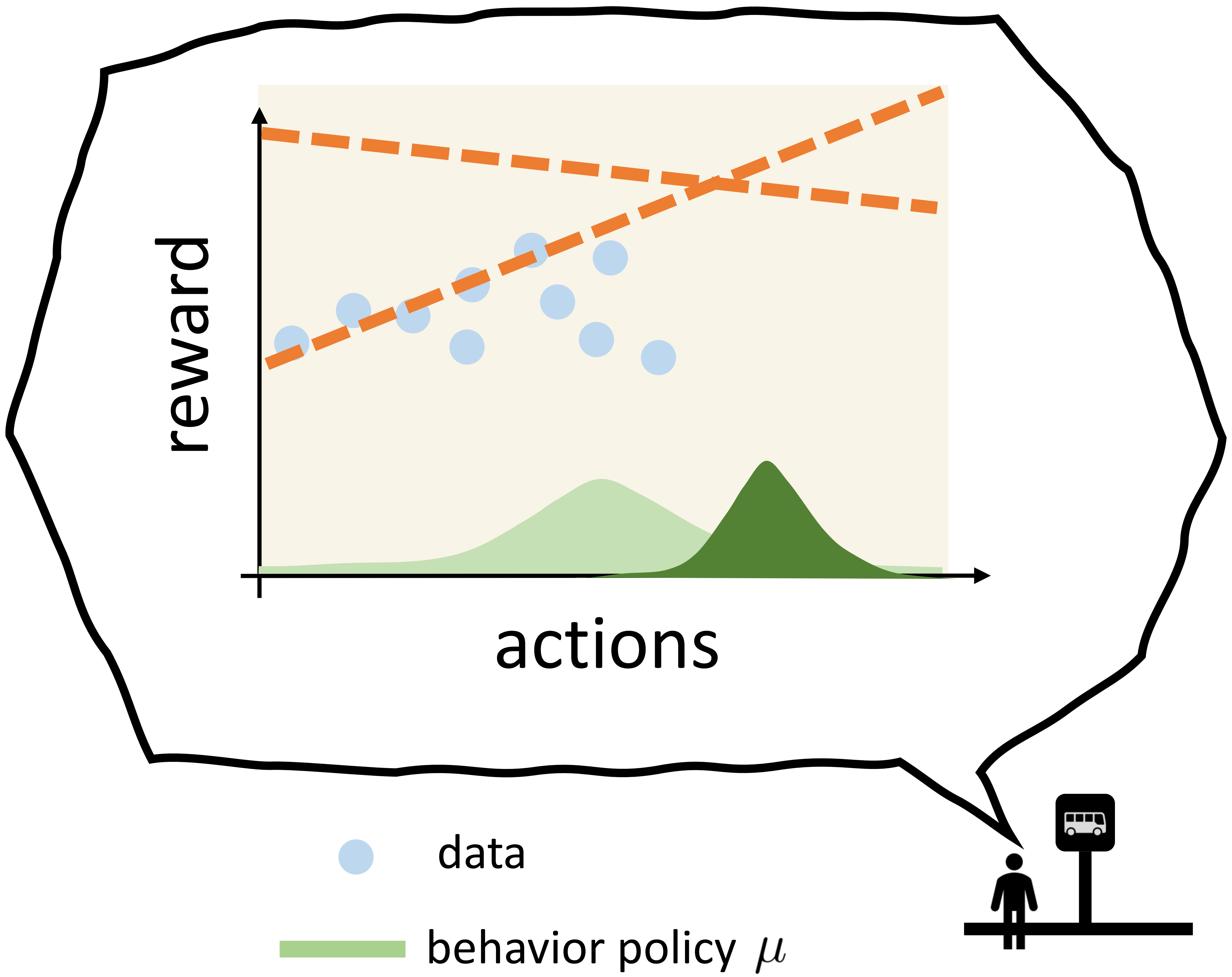
inactive value difference hypothesis

decision policy π

Not the policy that maximizes a single hypothesis...

Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

A Stackelberg Game for Offline RL



ATAC

$$\boxed{\widehat{\pi}^*} \in \arg\max_{\pi\in\Pi} \mathcal{L}_\mu(\pi, f^\pi)$$

$$\text{s.t.} \quad f^\pi \in \arg\min_{f\in\mathcal{F}} \mathcal{L}_\mu(\pi, f) + \beta\mathcal{E}_\mu(\pi, f)$$

value difference hypothesis $f(s,\cdot) - f(s,\mu)$

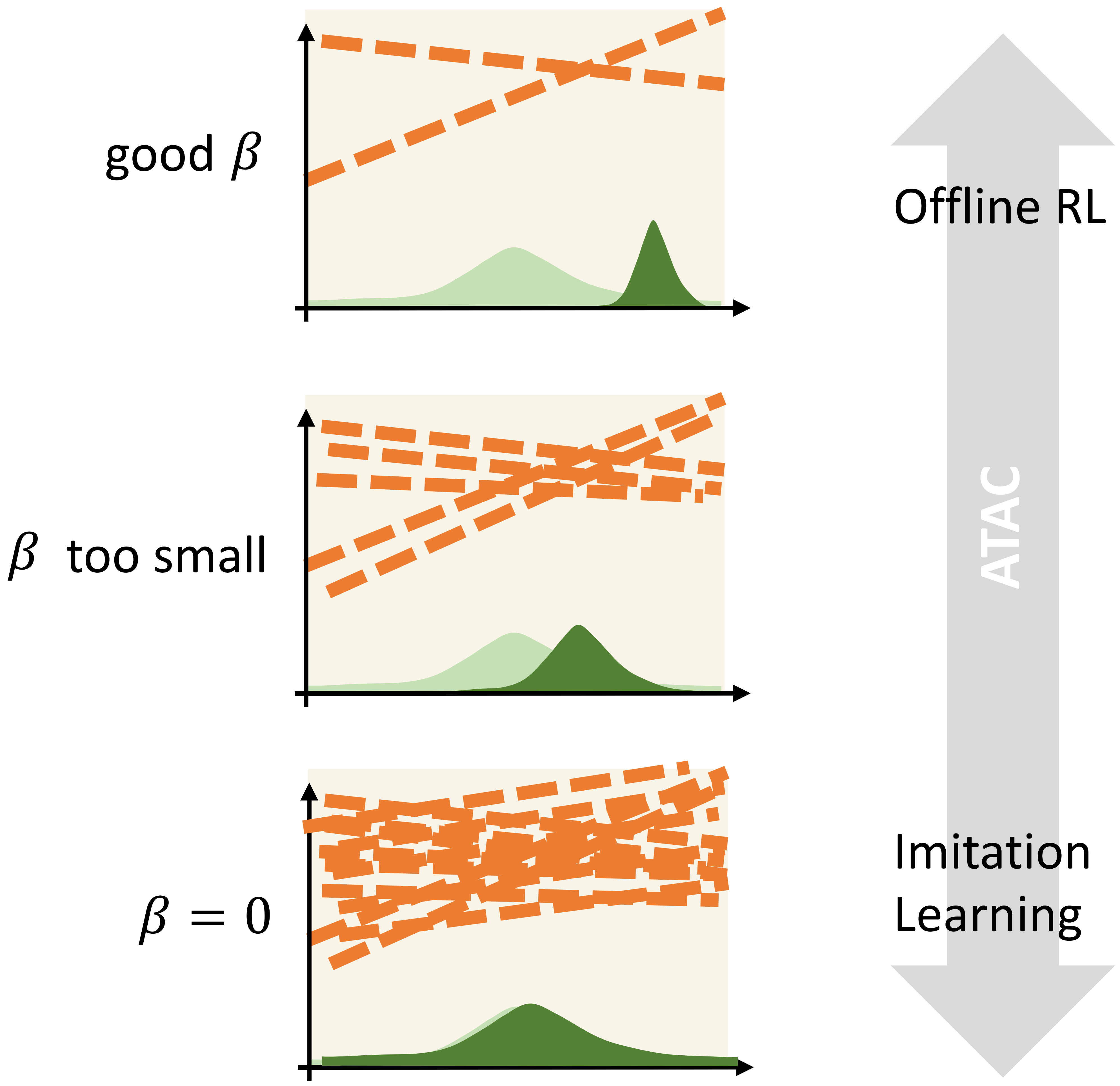
inactive value difference hypothesis

decision policy π

The optimal decision balances multiple hypotheses

Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

A Stackelberg Game for Offline RL



Leader = Actor = Conditional generator

Follower = Critic = Discriminator

ATAC

$$\widehat{\pi}^* \in \arg\max_{\pi \in \Pi} \mathcal{L}_\mu(\pi, f^\pi)$$

$$\text{s.t.} \quad f^\pi \in \arg\min_{f \in \mathcal{F}} \mathcal{L}_\mu(\pi, f) + \beta \mathcal{E}_\mu(\pi, f)$$

ATAC provides a bridge between offline RL and imitation learning with IPM via the lens of generative adversarial networks (GAN)

Offline RL + Relative Pessimism = IL + Bellman Regularization

Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

Solving the Stackelberg Game

losses are approximated by samples

$$f_k \in \arg\min_{f \in \mathcal{F}} \mathcal{L}_\mu(\pi_k, f) + \beta \mathcal{E}_\mu(\pi_k, f)$$

No-Regret + Best Response Scheme

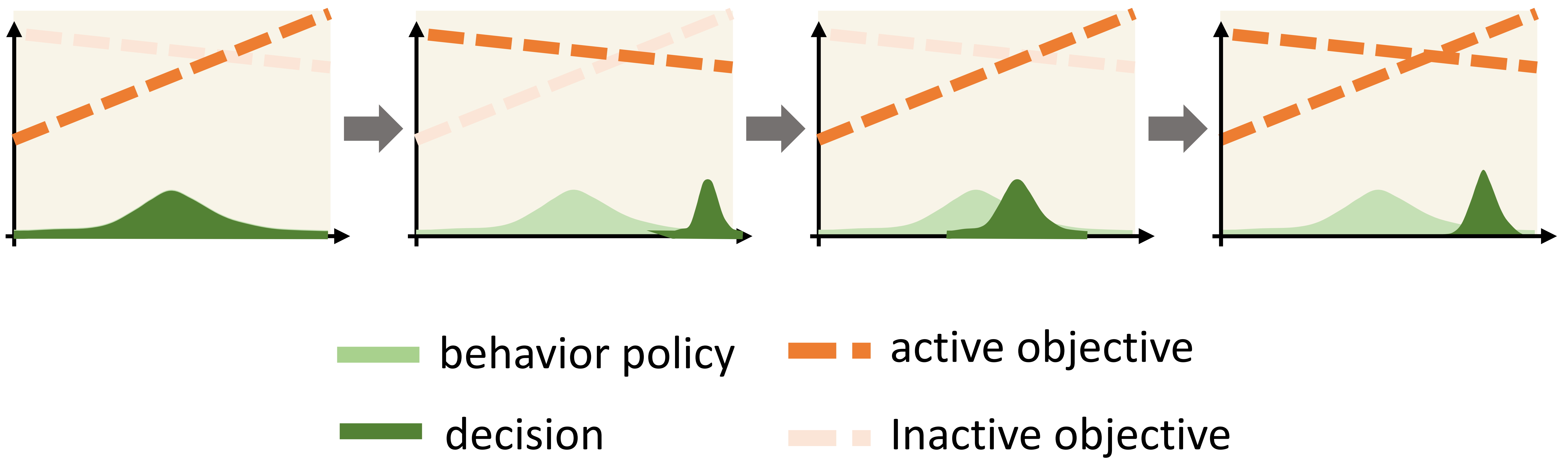
Repeat for K iterations

small policy update

$$\pi_{k+1} = \mathrm{NoRegret}(\pi_k, f_k)$$

Output uniform mixture of policies (theory) or the last policy (practice)

In practice, the above is implemented by two-timescale SGD updates



Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

ATAC Theory (Informal)

Learning Optimality

Assume $\mathcal{F}$ satisfies realizability and completeness.

Given dataset $\mathcal{D}$ s.t. $|\mathcal{D}| = N$. With $\beta = \Theta(N^{2/3})$. Then $\forall \pi \in \Pi$,

$$J(\pi) - J(\hat{\pi}) \leq \mathcal{O}\left(\frac{1}{(1-\gamma)N^{1/3}}\right) + \epsilon_{\text{generalization}}(\mathcal{F}, \pi, \mathcal{D})$$

average Bellman error of f_t on the distribution of π

*With a **well tuned** β, ATAC can compete with any policy within the data coverage.*

ATAC Theory (Informal)

Robust Policy Improvement

Assume $\mathcal{F}$ satisfies **realizability** without the need of completeness.

$$\text{If } \mu \in \Pi, \text{ then } \; J(\mu) - J(\bar{\pi}) \leq \mathcal{O}\left(\frac{1}{(1-\gamma)N^{1/2}} + \frac{\beta}{(1-\gamma)N} \right)$$

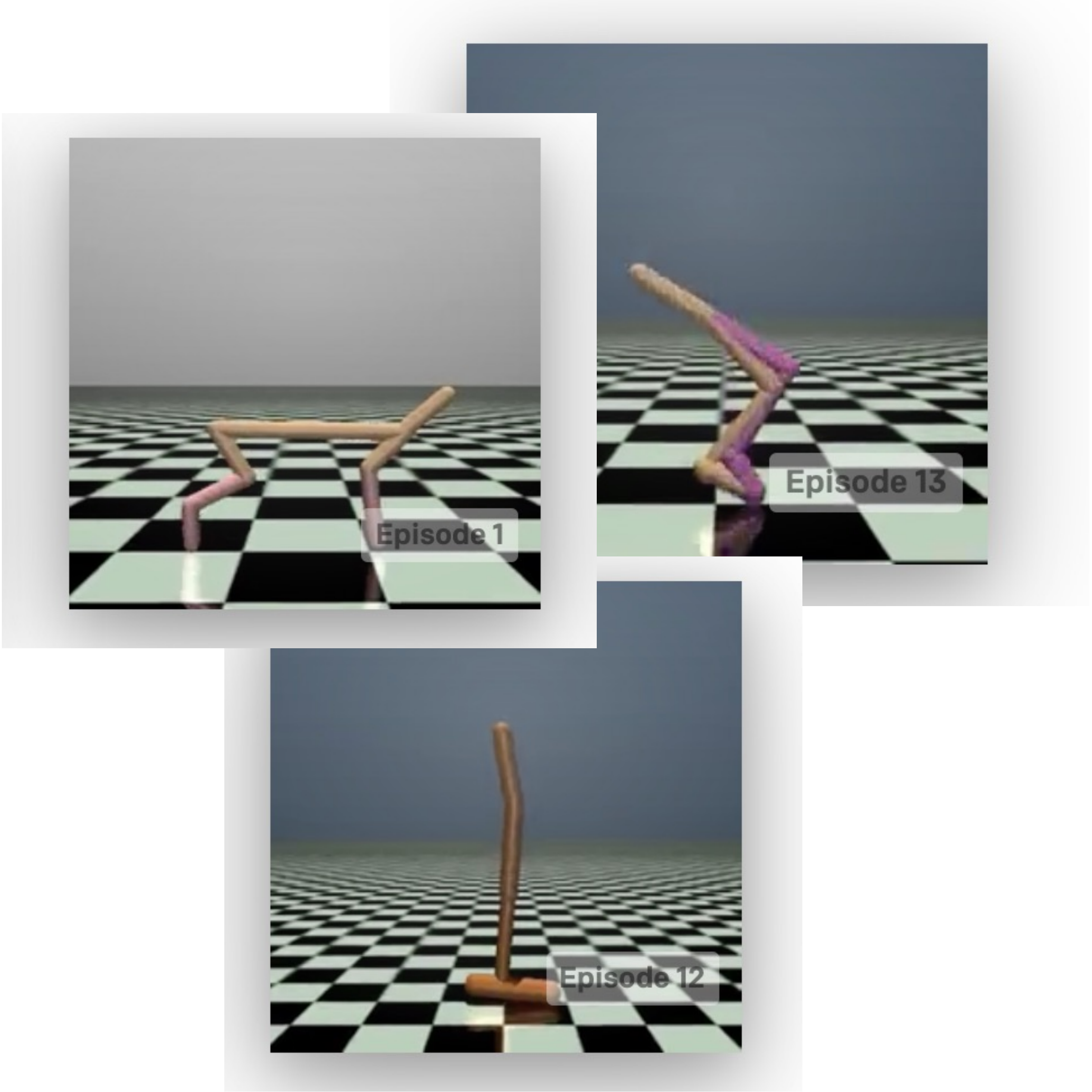
faster rate

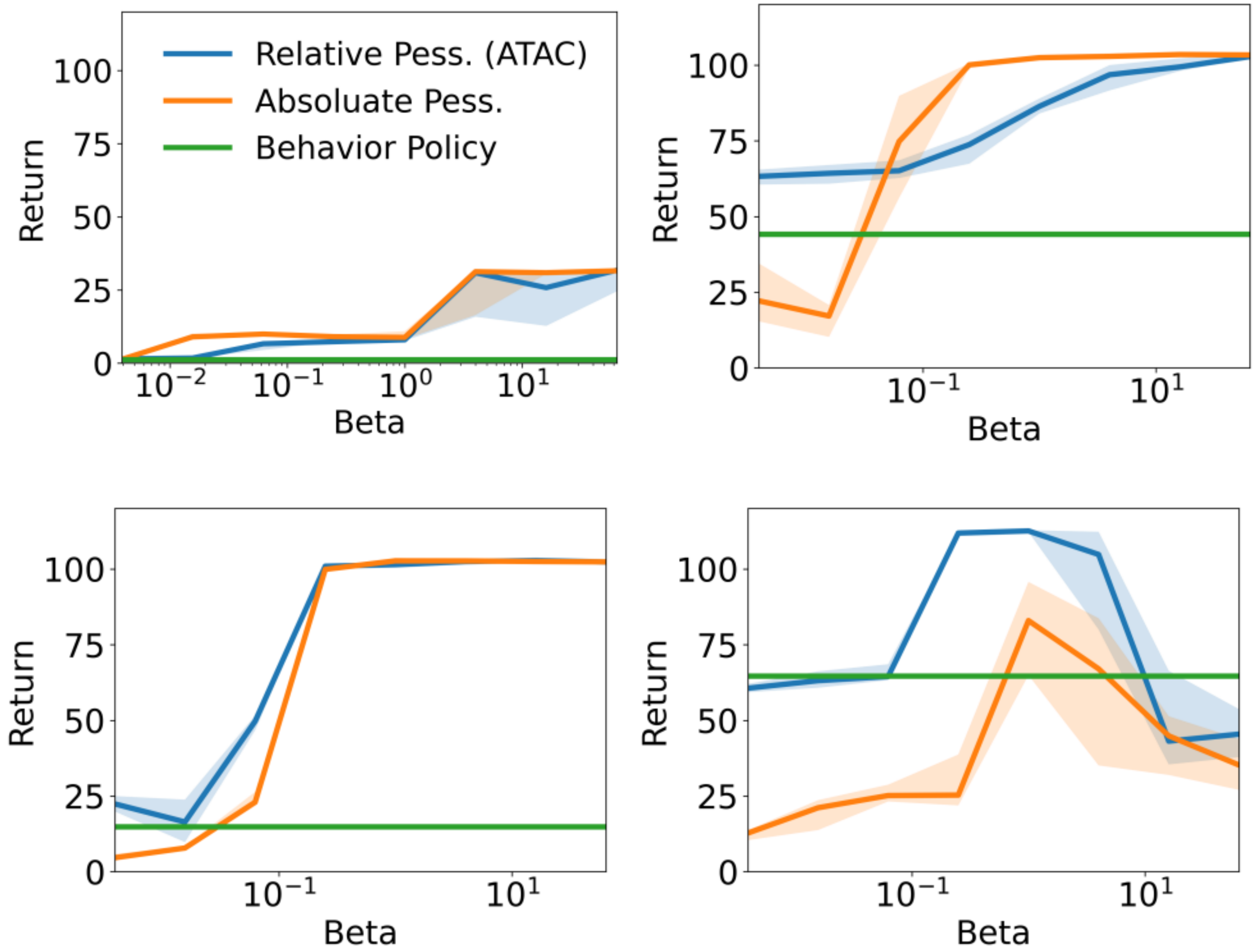
ATAC always improves over the behavior policy so long as $\beta = o(N)$.

Experimental Results

Robust Policy Improvement

ATAC's robustness property enables online HP selection. We can gradually increase β to tune its performance without breaking the baseline performance.

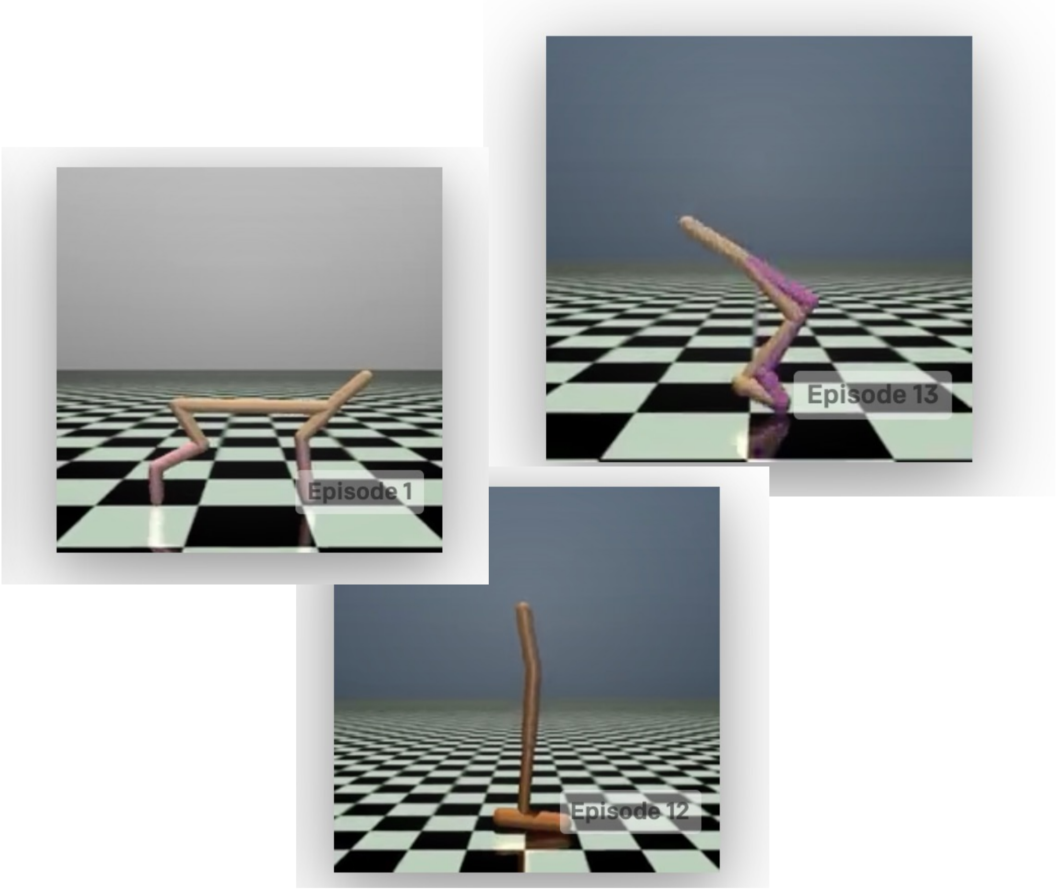


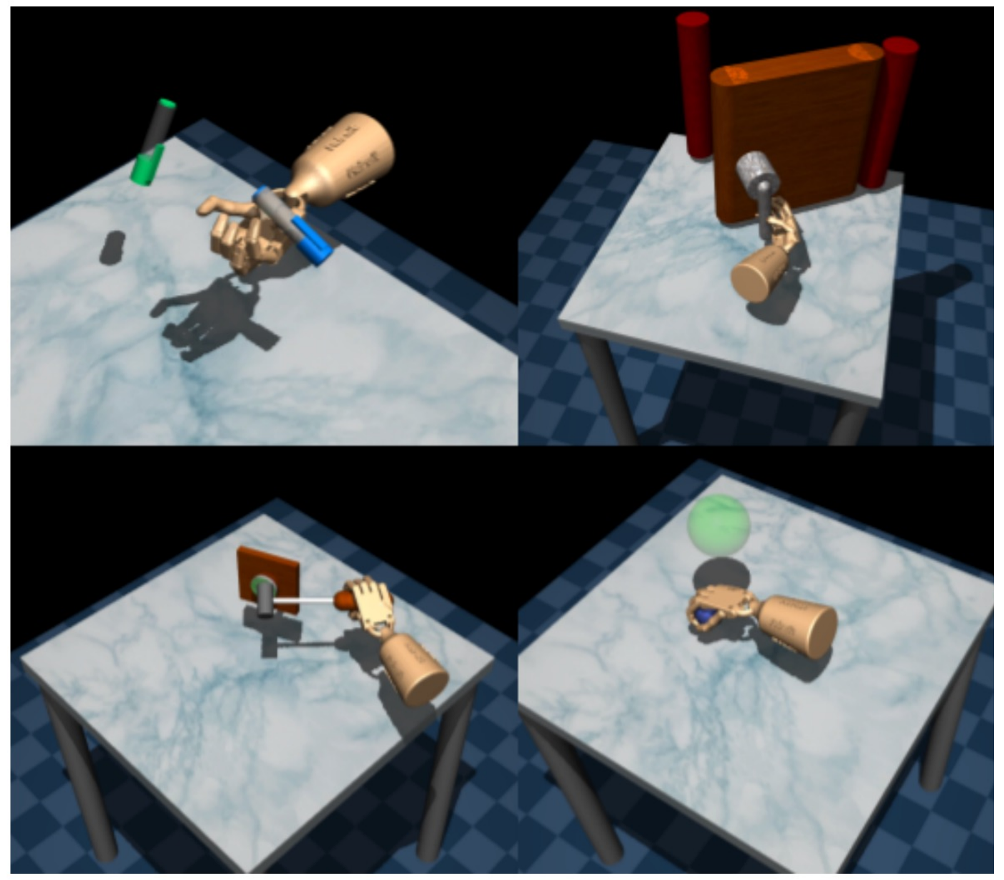


Credit: **Ching-An Cheng** "Adversarially Trained Actor Critic for Offline Reinforcement Learning" ICML 2022

Experimental Results

ATAC achieves SOTA performance, outperforming baseline algorithms in most datasets





	Behavior	ATAC*	CQL	COMBO	TD3BC	IQL	BC
halfcheetah-rand	-0.1	4.8	35.4	**38.8**	10.2	-	2.1
walker2d-rand	0.0	**8.0**	7.0	7.0	1.4	-	1.6
hopper-rand	1.2	**31.8**	10.8	17.9	11.0	-	9.8
halfcheetah-med	40.6	**54.3**	44.4	**54.2**	42.8	47.4	36.1
walker2d-med	62.0	**91.0**	74.5	75.5	79.7	78.3	6.6
hopper-med	44.2	**102.8**	86.6	94.9	**99.5**	66.3	29.0
halfcheetah-med-replay	27.1	49.5	46.2	**55.1**	43.3	44.2	38.4
walker2d-med-replay	14.8	**94.1**	32.6	56.0	25.2	73.9	11.3
hopper-med-replay	14.9	**102.8**	48.6	73.1	31.4	94.7	11.8
halfcheetah-med-exp	64.3	**95.5**	62.4	90.0	**97.9**	86.7	35.8
walker2d-med-exp	82.6	**116.3**	98.7	96.1	101.1	**109.6**	6.4
hopper-med-exp	64.7	**112.6**	**111.0**	**111.1**	**112.2**	91.5	**111.9**
pen-human	207.8	79.3	37.5	-	-	71.5	34.4
hammer-human	25.4	**6.7**	**4.4**	-	-	1.4	1.5
door-human	28.6	8.7	**9.9**	-	-	4.3	0.5
relocate-human	86.1	**0.3**	**0.2**	-	-	**0.1**	**0.0**
pen-cloned	107.7	73.9	39.2	-	-	37.3	56.9
hammer-cloned	8.1	2.3	2.1	-	-	2.1	0.8
door-cloned	12.1	**8.2**	0.4	-	-	1.6	-0.1
relocate-cloned	28.7	**0.8**	**-0.1**	-	-	**-0.2**	**-0.1**
pen-exp	105.7	**159.5**	107.0	-	-	-	85.1
hammer-exp	96.3	**128.4**	86.7	-	-	-	**125.6**
door-exp	100.5	**105.5**	**101.5**	-	-	-	34.9
relocate-exp	101.6	**106.5**	95.0	-	-	-	**101.3**

Datasets where ATAC is the best performing algorithm, with 9% improvement (median) compared with the best baseline algorithm.