The Effective Horizon Explains Deep RL Performance in Stochastic Environments

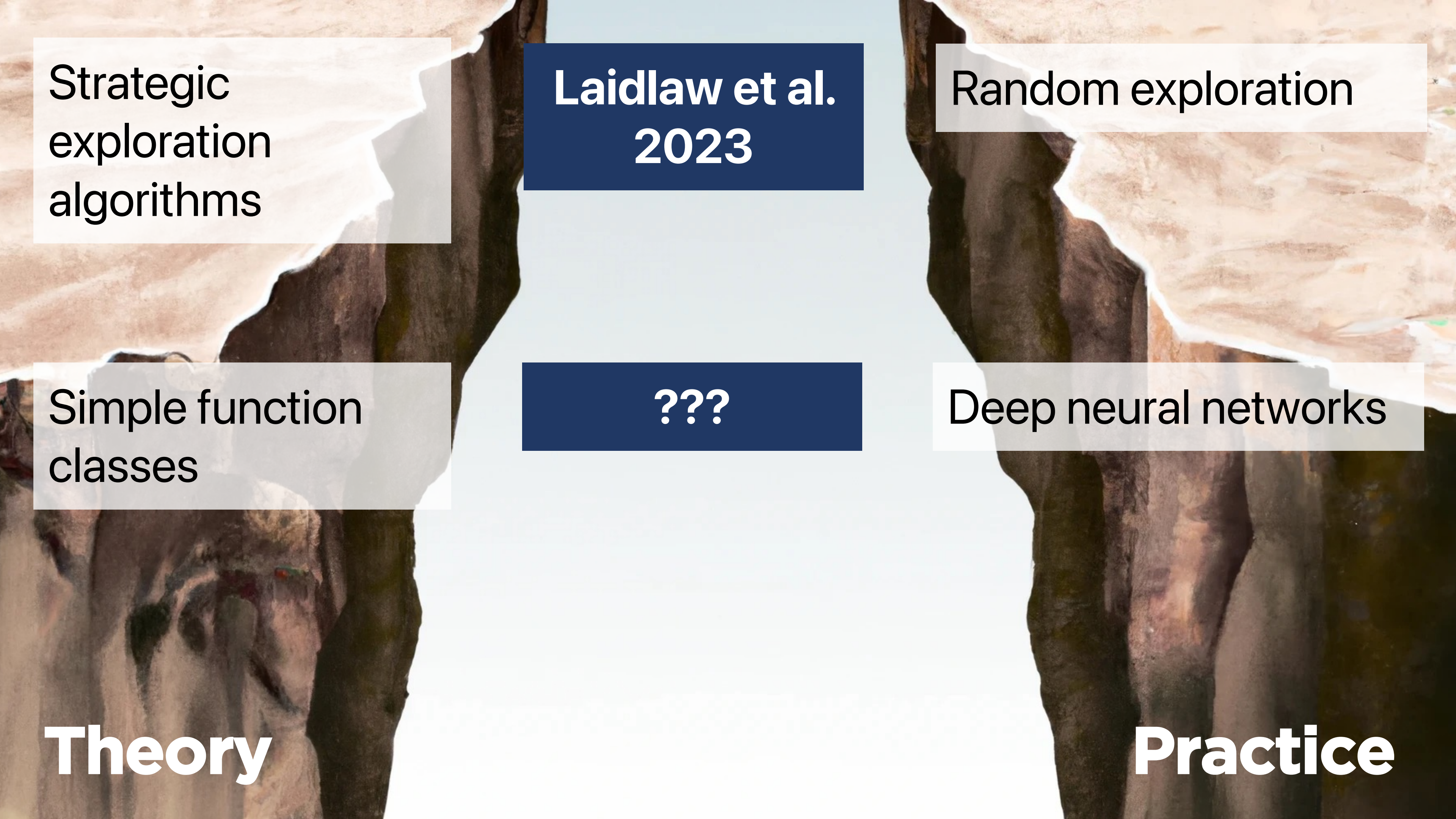
Cassidy Laidlaw

with Banghua Zhu, Anca Dragan, and Stuart Russell

Berkeley
UNIVERSITY OF CALIFORNIA

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

CHAI
Center for Human-Compatible Artificial Intelligence

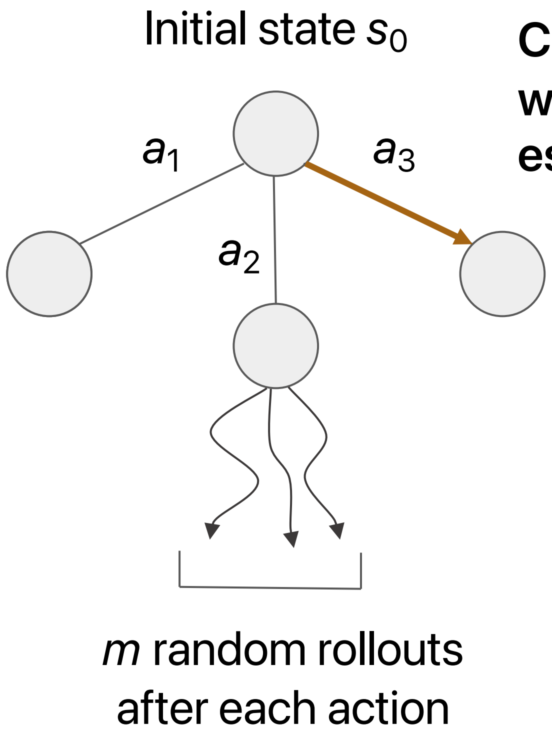

Strategic exploration algorithms
Laidlaw et al. 2023
Random exploration
Simple function classes
???
Deep neural networks
Theory
Practice

Previous work

In many common MDPs, acting greedily with respect to the **random policy's Q-function** gives an **optimal policy**.

Previous work: the Greedy Over Random Policy (GORP) algorithm

If acting greedily on the random policy's Q-function is optimal, GORP will find an optimal policy.



Commit to the action with the highest estimated Q-value

From deterministic to stochastic

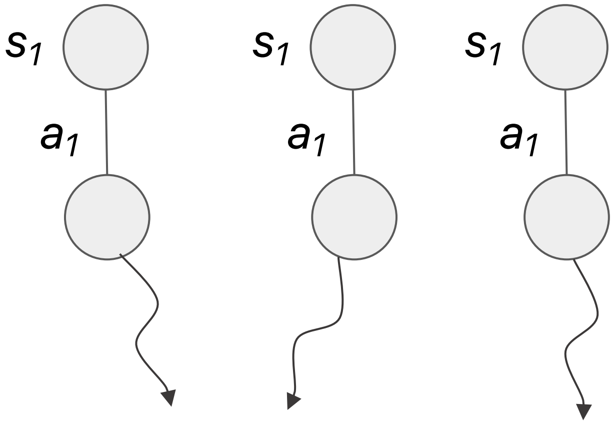
GORP solves **73%** of the environments in BRIDGE.

But when we add sticky actions to make these environments **stochastic**, GORP only solves **19%.**

From deterministic to stochastic

- In deterministic environments, an **open-loop plan or sequence of actions** is enough.
- But in stochastic environments, we need **a closed loop plan—a *policy***.

From deterministic to stochastic



Fit $\hat{Q}_{rand}$ to $(s_1, a_1, \Sigma\ R)$ triples via **regression**

If our regressed Q-function generalizes well in-distribution, then at most initial states we can choose an optimal action!

From deterministic to stochastic

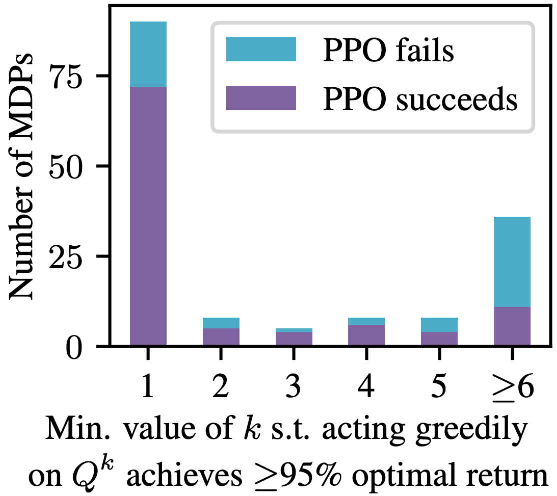
By committing to **acting greedily on the estimated Q-function** for the first timestep, we can reach a **fixed distribution** over states at the second timestep.

- Repeat the process to regression the random policy's Q-function at the second timestep, and so on!

Generalizing the surprising property

- Let Q_1 be the Q-function of the random policy
- Let Q_{k+1} be the result of applying one step of Q value iteration (QVI) to Q_k

We say an MDP is ***k*-QVI-solvable** if acting greedily with respect to Q_k is optimal.



Shallow Q-iteration via RL (SQIRL)

GORP

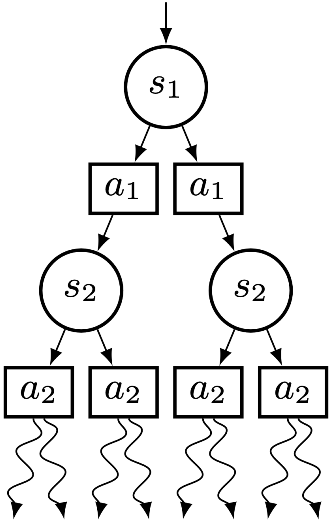
$$\pi_1 \leftarrow \arg\max_a \hat{Q}_1$$

$$\hat{Q}_1 \leftarrow R_1 + \hat{V}_2$$

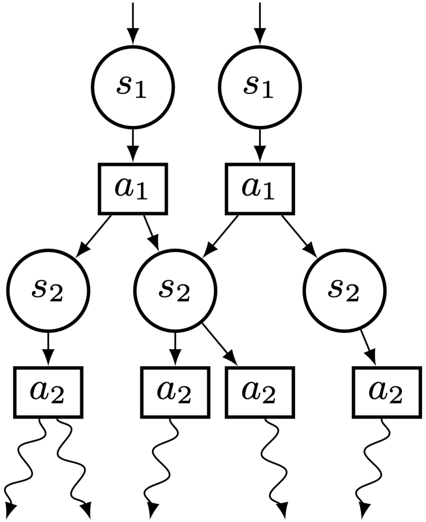
$$\hat{V}_2 \leftarrow \max_a \hat{Q}_2$$

$$\hat{Q}_2 \leftarrow \textsc{Avg}(y^i)$$

$$y^i \leftarrow \sum_{t=2}^{T} R_t$$



SQIRL



$$\pi_1 \leftarrow \arg\max_a \hat{Q}_1$$

$$\hat{Q}_1 \leftarrow \textsc{Regress}(R_1 + \hat{V}_2)$$

$$\hat{V}_2 \leftarrow \max_a \hat{Q}_2$$

$$\hat{Q}_2 \leftarrow \textsc{Regress}(y^i)$$

$$y^i \leftarrow \sum_{t=2}^{T} R_t$$

Generalizing the surprising property

- Let Q_1 be the Q-function of the random policy
- Let Q_{k+1} be the result of applying one step of Q value iteration (QVI) to Q_k

We say an MDP is ***k*-QVI-solvable** if acting greedily with respect to Q_k is optimal.

If an MDP is k-QVI-solvable, define its ***k*-gap** as

$$\Delta_k = \inf_{(t,\, s) \in [T] \times \mathcal{S}} \left[\max_{a^* \in \mathcal{A}} Q_k^t(s, a^*) - \max_{a \notin \arg\max Q_k^t(s,\, a)} Q_k^t(s, a) \right].$$

The stochastic effective horizon

$$\bar{H}_k = k + \log_A 1/\Delta_k^2$$

The stochastic effective horizon is

$$\bar{H} = \min_k \bar{H}_k$$

Proposition: the (deterministic) effective horizon is upper bounded by the stochastic E.H. up to log factors.

The sample complexity of SQIRL

Theorem: if an MDP is k-QVI-solvable, then the sample complexity of SQIRL for finding an ε-optimal policy is at most

$$\tilde{O}(k\, T^3\, A^{\bar{H}_k}\, \underbrace{\boxed{\alpha^{2(k-1)}\, D}}_{\text{Constants depending on "regression oracle"}} / \varepsilon)$$

For comparison, GORP's sample complexity is $\tilde{O}(k\, T^2\, A^{\bar{H}_k})$.

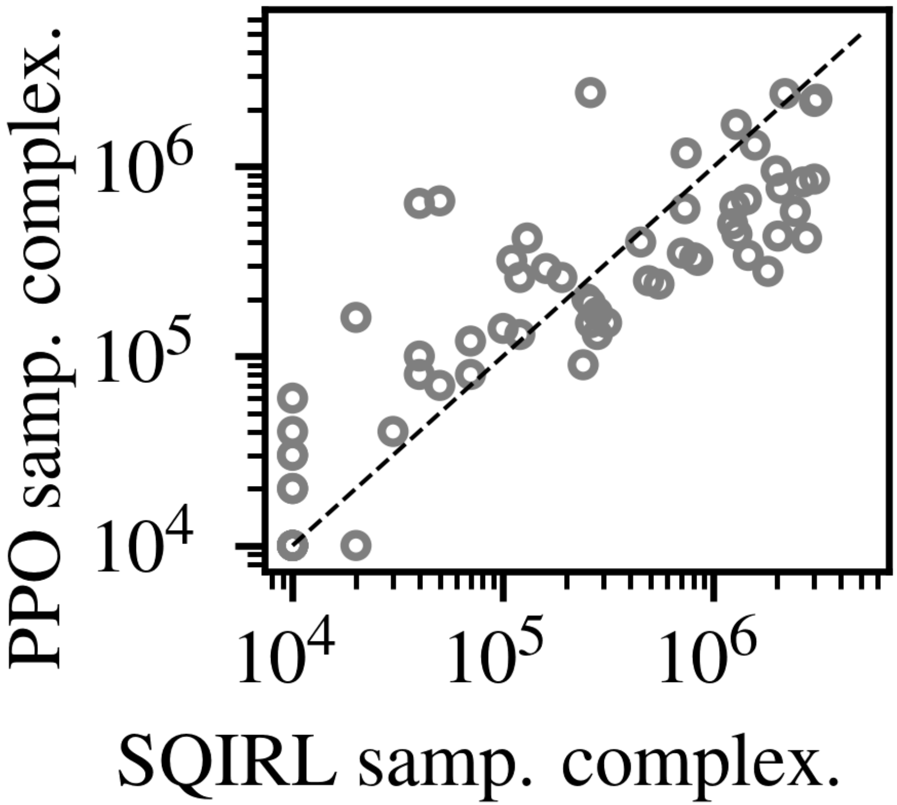
The sample complexity of SQIRL

Setting	Sample complexity bounds	
	Strategic exploration	SQIRL (ours)
Tabular MDP	$\tilde{O}(TSA/\varepsilon^2)$	$\tilde{O}(kT^3SA^{\bar{H}_k+1}/\varepsilon)$
Linear MDP	$\tilde{O}(T^2d^2/\varepsilon^2)$	$\tilde{O}(kT^3dA^{\bar{H}_k}/\varepsilon)$
Q-functions with finite pseudo-dimension	—	$\tilde{O}(k5^kT^3dA^{\bar{H}_k}/\varepsilon)$

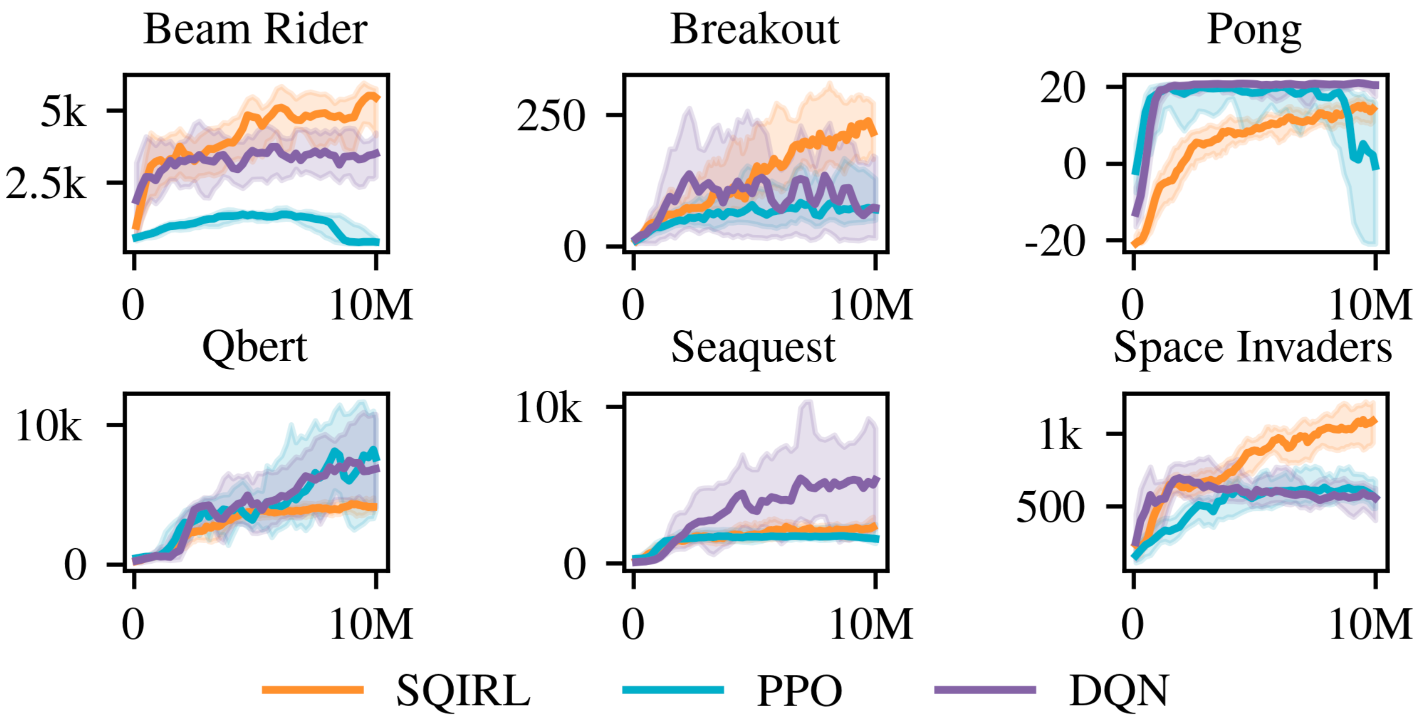
Experimental results

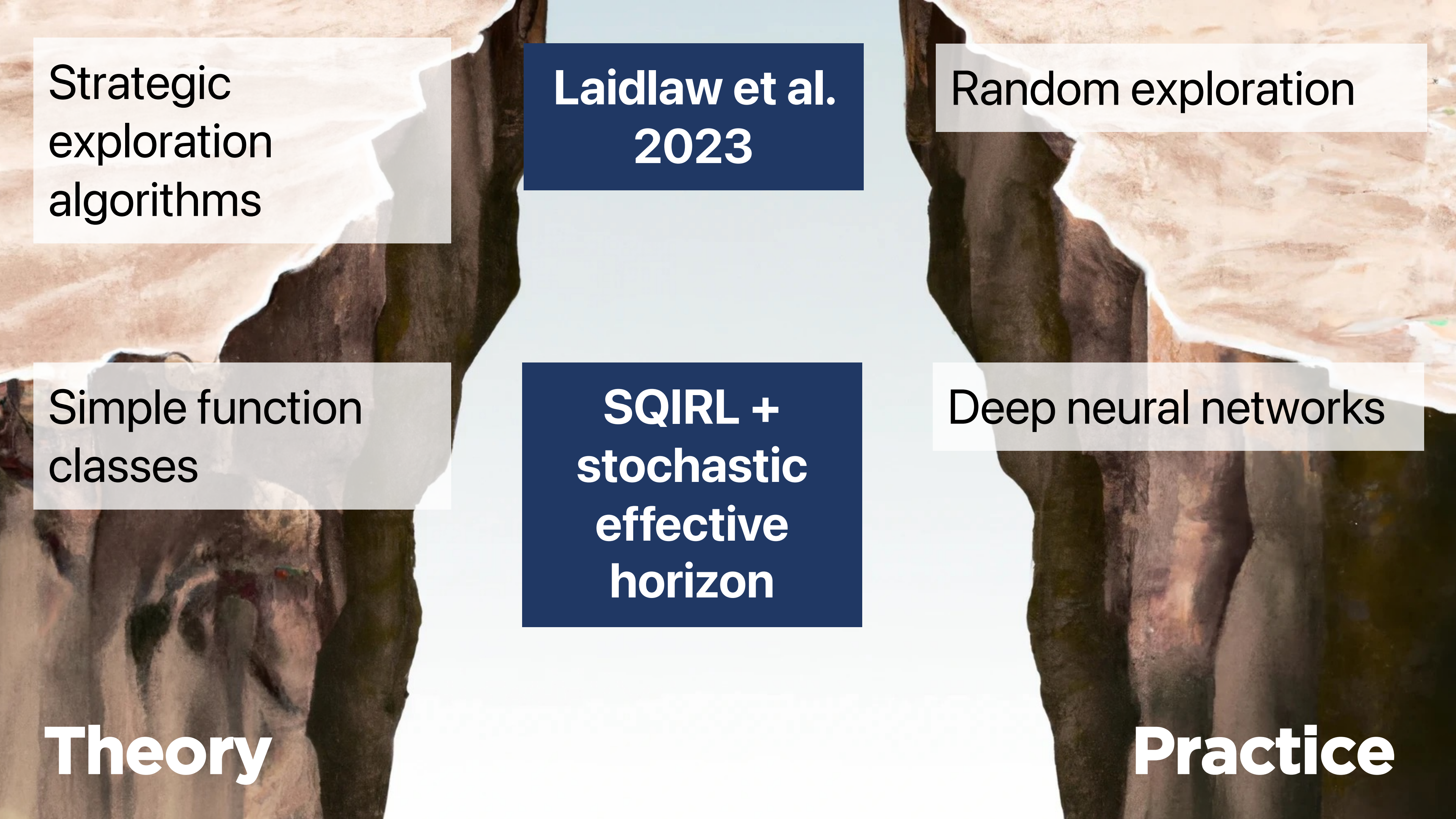
Algorithm	Num. envs. solved
PPO	98
DQN	78
SQIRL	77
GORP	29

Experimental results



Experimental results




Strategic exploration algorithms
Laidlaw et al. 2023
Random exploration
Simple function classes
SQIRL + stochastic effective horizon
Deep neural networks
Theory
Practice

The Effective Horizon Explains Deep RL Performance in Stochastic Environments

Cassidy Laidlaw, Banghua Zhu, Stuart Russell, and Anca Dragan

arxiv.org/abs/2312.08369