Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations

Daniel Brown*, Wonjoon Goo*, Prabhat Nagarajan, and Scott Niekum

TEXAS
The University of Texas at Austin

PeARL
Personal Autonomous Robotics Lab

Inverse Reinforcement Learning

Current approaches …

1. Can’t do better than the demonstrator.

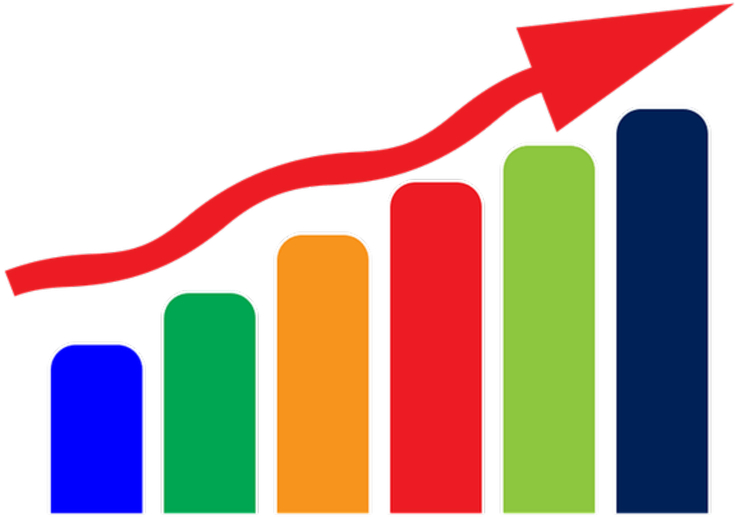
2. Are hard to scale to complex problems.

Inverse Reinforcement Learning

Current approaches …

1. Can’t do better than the demonstrator.
2. Are hard to scale to complex problems.

IRL via Ranked Demonstrations



Inverse Reinforcement Learning

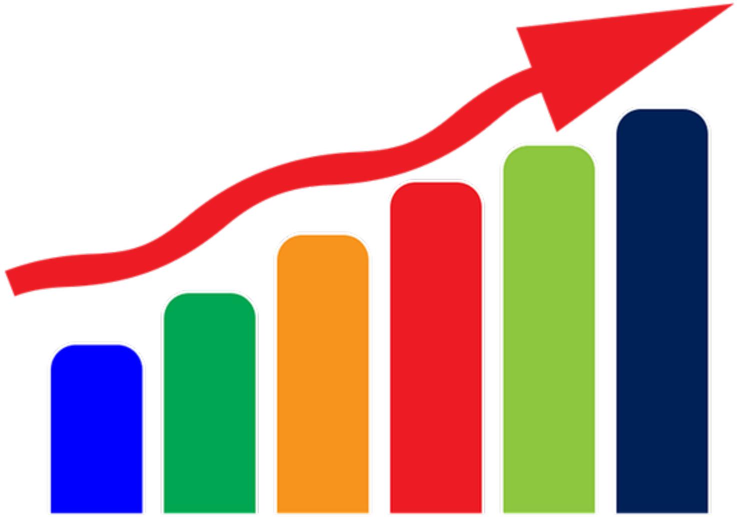
Current approaches …

1. ~~Can't do better than the demonstrator.~~

We find a reward function that explains the ranking, allowing for extrapolation.

2. Are hard to scale to complex problems.

IRL via Ranked Demonstrations



Inverse Reinforcement Learning

Current approaches …

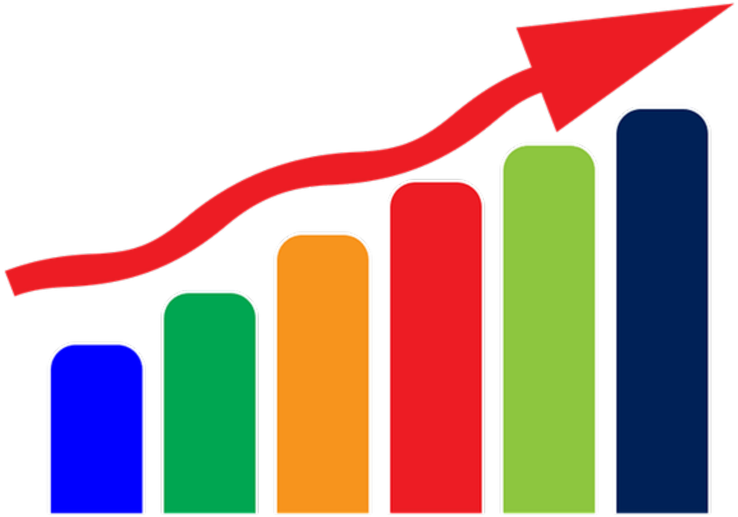
1. ~~Can't do better than the demonstrator.~~

We find a reward function that explains the ranking, allowing for extrapolation.

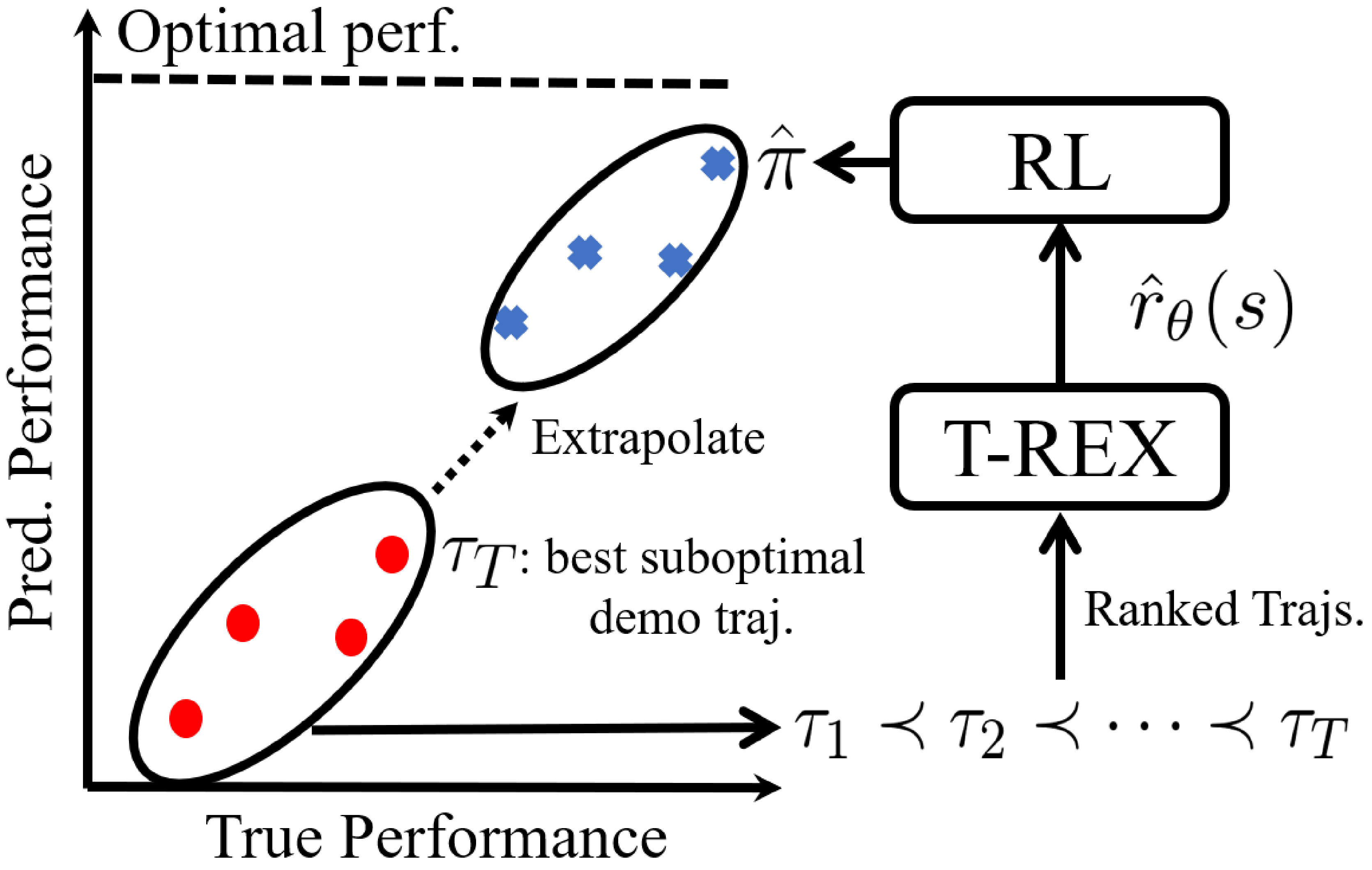
2. ~~Are hard to scale to complex problems.~~

Inverse Reinforcement Learning becomes standard binary classification.

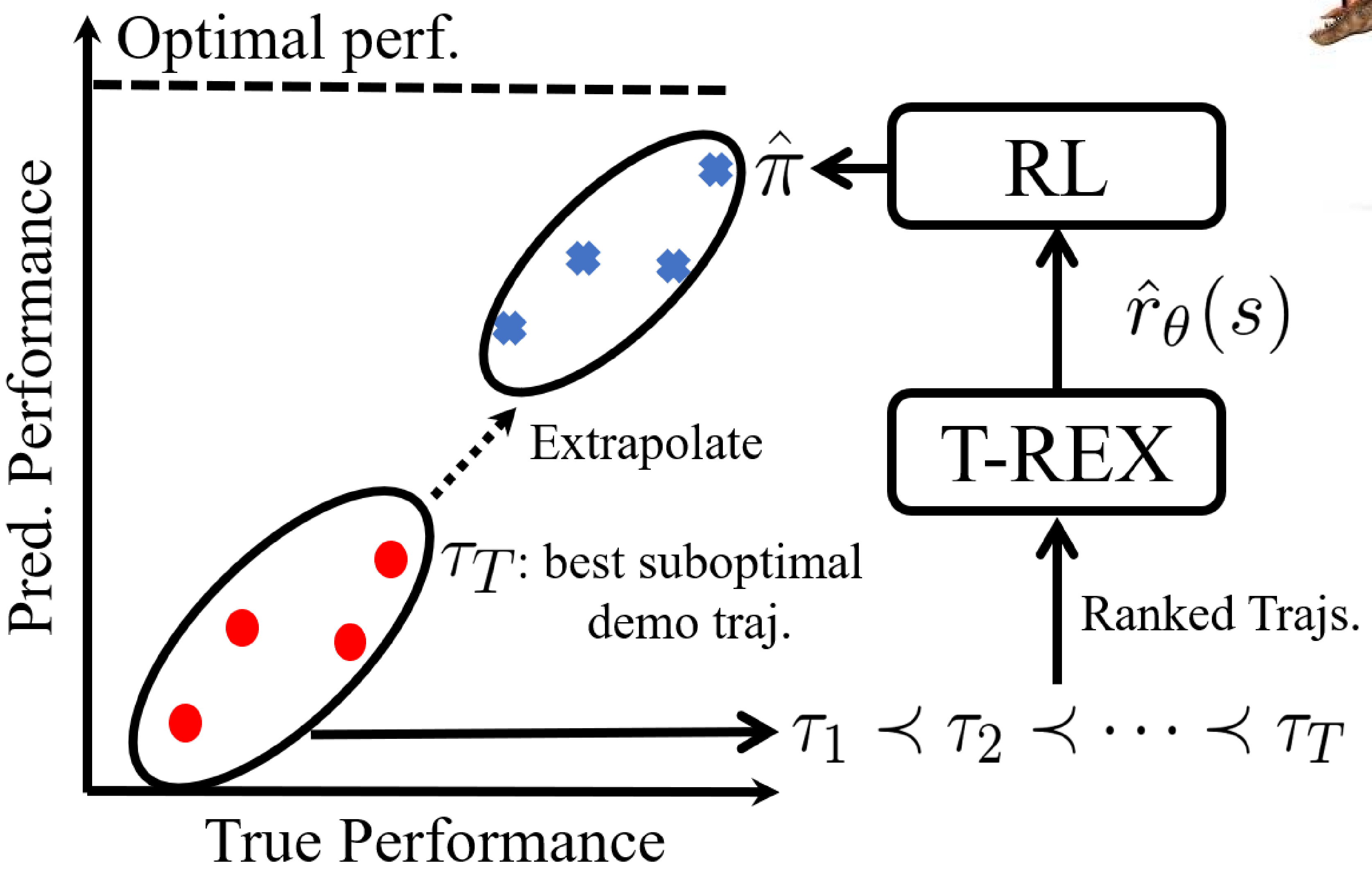
IRL via Ranked Demonstrations

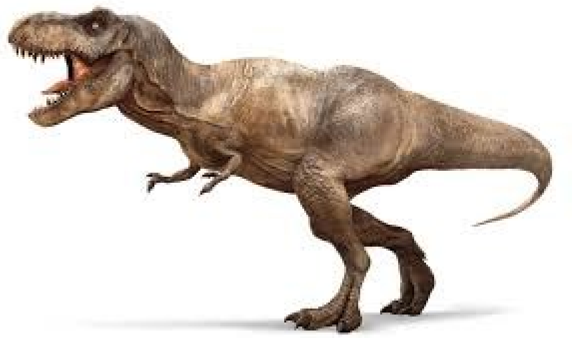


Trajectory-ranked Reward Extrapolation (T-REX)



Trajectory-ranked Reward Extrapolation (T-REX)





Trajectory-ranked Reward Extrapolation (T-REX)

Given ranked demonstrations

$$\tau_1 \prec \tau_2 \prec \cdots \prec \tau_T$$

How do we train the reward function $\hat{r}_\theta(s)$?

Trajectory-ranked Reward Extrapolation (T-REX)

$$\tau_1 \prec \tau_2 \prec \cdots \prec \tau_T$$

Trajectory-ranked Reward Extrapolation (T-REX)

$$\boxed{\tau_1} \prec \boxed{\tau_2} \prec \cdots \prec \tau_T$$

Trajectory-ranked Reward Extrapolation (T-REX)

$$\boxed{\tau_1} \prec \boxed{\tau_2} \prec \cdots \prec \tau_T$$

$$\sum_{s \in \tau_i} \hat{r}_\theta(s) < \sum_{s \in \tau_j} \hat{r}_\theta(s)$$

Trajectory-ranked Reward Extrapolation (T-REX)

$$\boxed{\tau_1} \prec \boxed{\tau_2} \prec \cdots \prec \tau_T$$

$$\sum_{s \in \tau_i} \hat{r}_\theta(s) < \sum_{s \in \tau_j} \hat{r}_\theta(s)$$

$$\mathcal{L}(\theta) \approx - \sum_{\tau_i \prec \tau_j} \log \frac{\exp \sum_{s \in \tau_j} \hat{r}_\theta(s)}{\exp \sum_{s \in \tau_i} \hat{r}_\theta(s) + \exp \sum_{s \in \tau_j} \hat{r}_\theta(s)}$$

Trajectory-ranked Reward Extrapolation (T-REX)

$$\boxed{\tau_1} \prec \tau_2 \prec \boxed{\cdots} \prec \tau_T$$

$$\sum_{s \in \tau_i} \hat{r}_\theta(s) < \sum_{s \in \tau_j} \hat{r}_\theta(s)$$

$$\mathcal{L}(\theta) \approx - \sum_{\tau_i \prec \tau_j} \log \frac{\exp \sum_{s \in \tau_j} \hat{r}_\theta(s)}{\exp \sum_{s \in \tau_i} \hat{r}_\theta(s) + \exp \sum_{s \in \tau_j} \hat{r}_\theta(s)}$$

Trajectory-ranked Reward Extrapolation (T-REX)

$$\boxed{\tau_1} \prec \tau_2 \prec \cdots \prec \boxed{\tau_T}$$

$$\sum_{s \in \tau_i} \hat{r}_\theta(s) < \sum_{s \in \tau_j} \hat{r}_\theta(s)$$

$$\mathcal{L}(\theta) \approx -\sum_{\tau_i \prec \tau_j} \log \frac{\exp \sum_{s \in \tau_j} \hat{r}_\theta(s)}{\exp \sum_{s \in \tau_i} \hat{r}_\theta(s) + \exp \sum_{s \in \tau_j} \hat{r}_\theta(s)}$$

Trajectory-ranked Reward Extrapolation (T-REX)

$$\tau_1 \prec \boxed{\tau_2} \prec \boxed{\cdots} \prec \tau_T$$

$$\sum_{s \in \tau_i} \hat{r}_\theta(s) < \sum_{s \in \tau_j} \hat{r}_\theta(s)$$

$$\mathcal{L}(\theta) \approx -\sum_{\tau_i \prec \tau_j} \log \frac{\exp \sum_{s \in \tau_j} \hat{r}_\theta(s)}{\exp \sum_{s \in \tau_i} \hat{r}_\theta(s) + \exp \sum_{s \in \tau_j} \hat{r}_\theta(s)}$$

Trajectory-ranked Reward Extrapolation (T-REX)

$$\tau_1 \prec \boxed{\tau_2} \prec \cdots \prec \boxed{\tau_T}$$

$$\sum_{s \in \tau_i} \hat{r}_\theta(s) < \sum_{s \in \tau_j} \hat{r}_\theta(s)$$

$$\mathcal{L}(\theta) \approx -\sum_{\tau_i \prec \tau_j} \log \frac{\exp \sum_{s \in \tau_j} \hat{r}_\theta(s)}{\exp \sum_{s \in \tau_i} \hat{r}_\theta(s) + \exp \sum_{s \in \tau_j} \hat{r}_\theta(s)}$$

Trajectory-ranked Reward Extrapolation (T-REX)

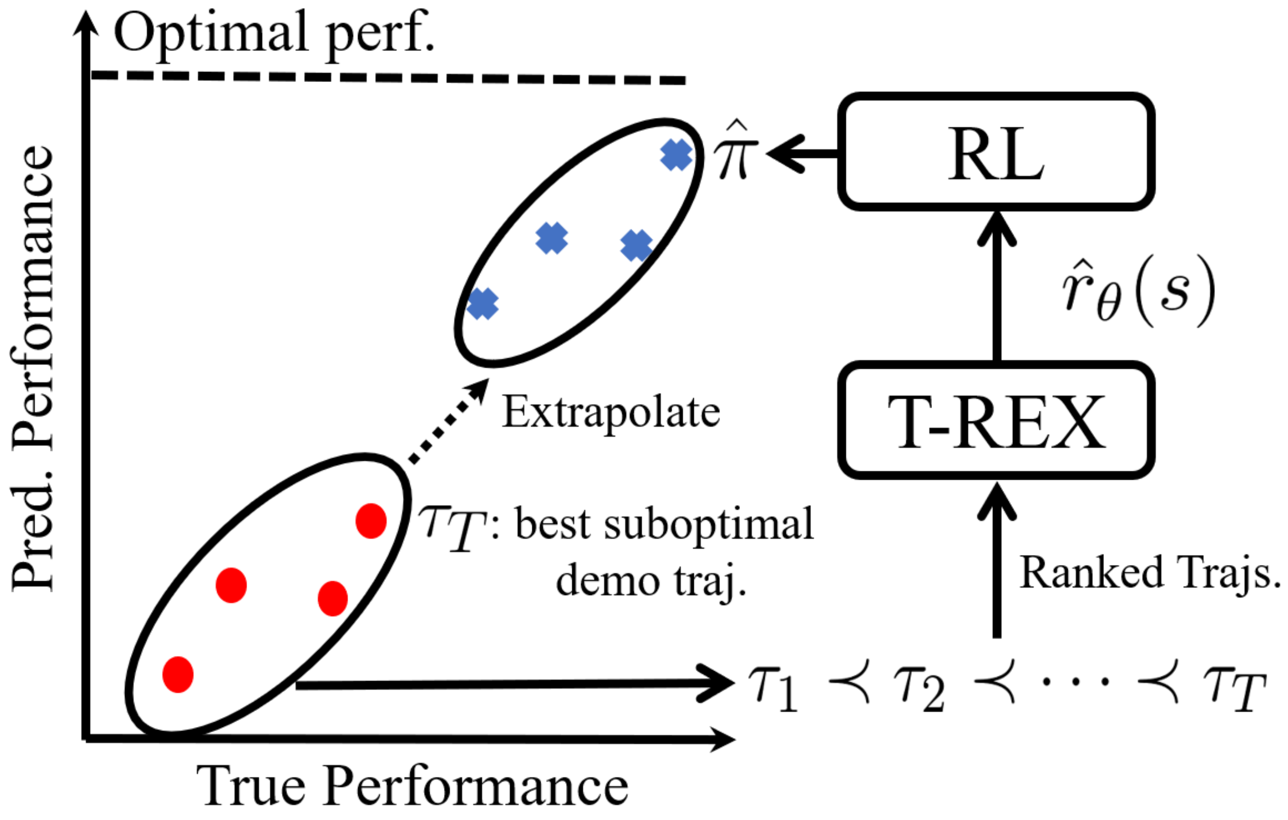
$$\tau_1 \prec \tau_2 \prec \boxed{\cdots} \prec \boxed{\tau_T}$$

$$\sum_{s \in \tau_i} \hat{r}_\theta(s) < \sum_{s \in \tau_j} \hat{r}_\theta(s)$$

$$\mathcal{L}(\theta) \approx -\sum_{\tau_i \prec \tau_j} \log \frac{\exp \sum_{s \in \tau_j} \hat{r}_\theta(s)}{\exp \sum_{s \in \tau_i} \hat{r}_\theta(s) + \exp \sum_{s \in \tau_j} \hat{r}_\theta(s)}$$

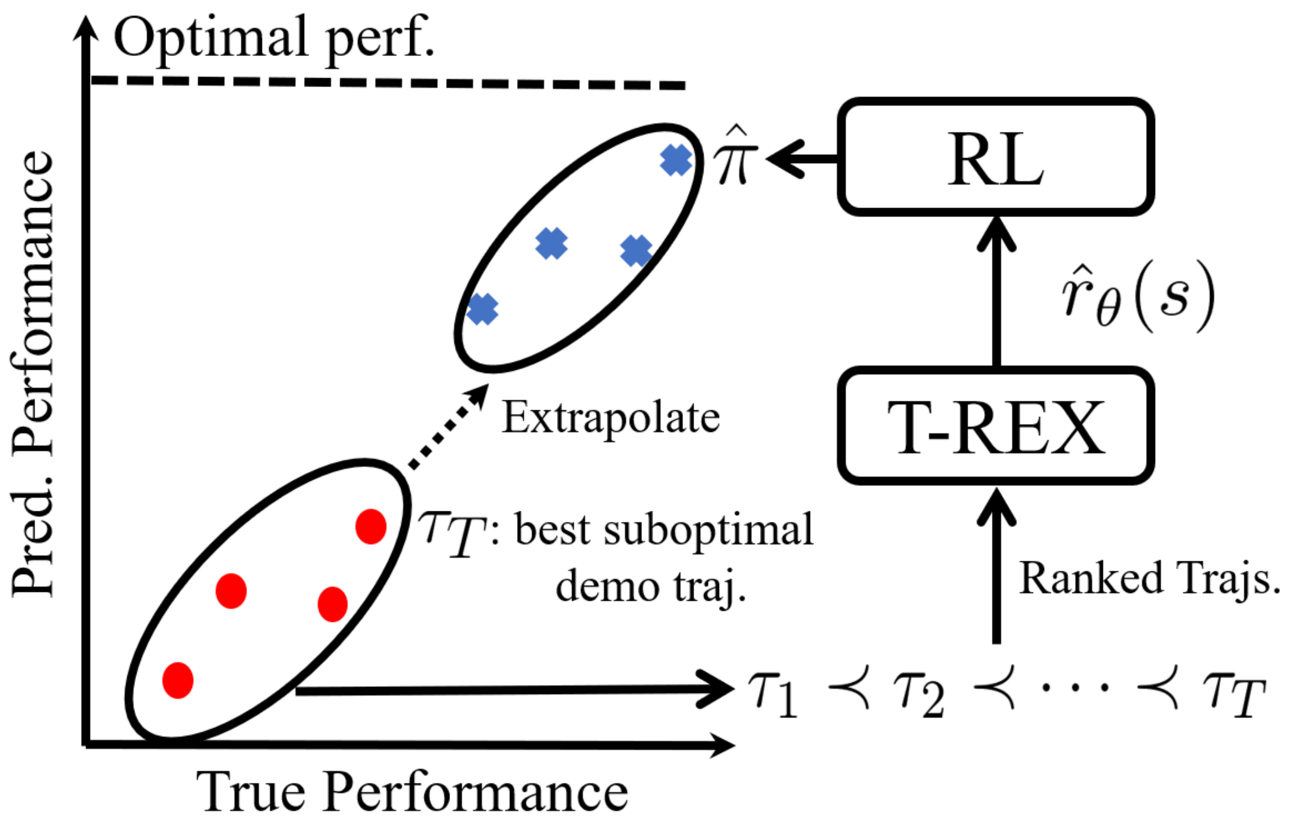
We subsample trajectories to create a large dataset of weakly labeled pairs!

Trajectory-ranked Reward Extrapolation (T-REX)



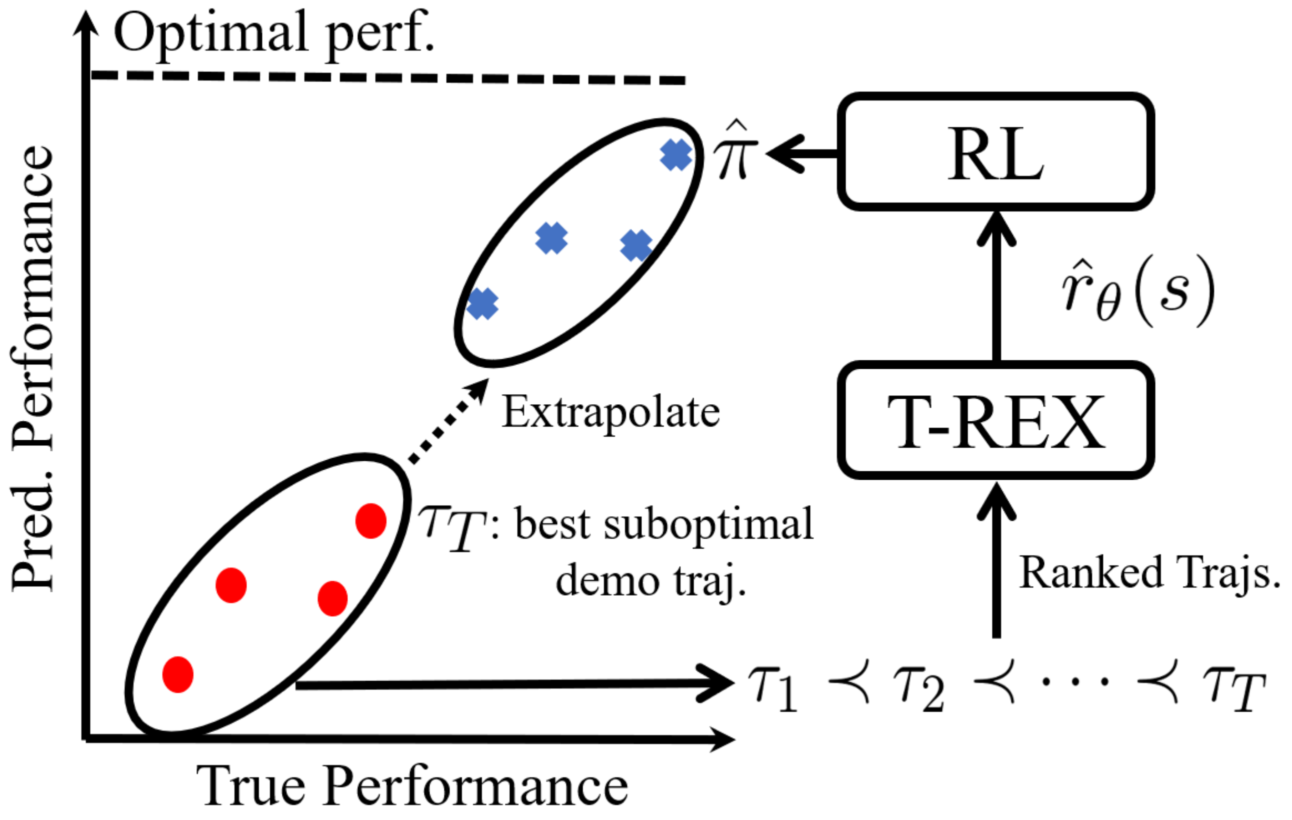
- Simple:
  - IRL as binary classification.
  - No human supervision during policy learning.
  - No inner-loop MDP solver.
  - No inference time data collection (e.g. GAIL).
  - No action labels required.

Trajectory-ranked Reward Extrapolation (T-REX)



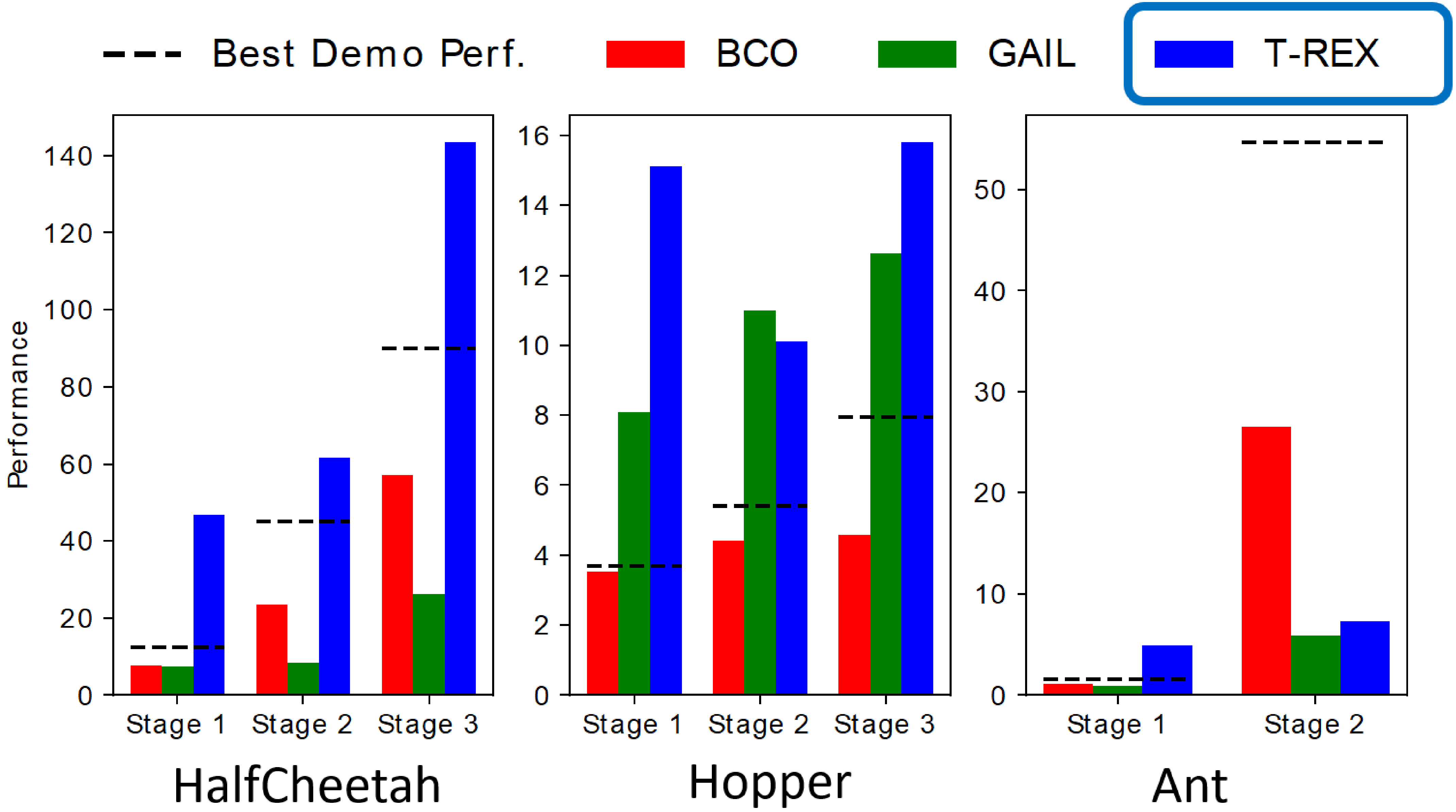
- Simple:
  - IRL as binary classification.
  - No human supervision during policy learning.
  - No inner-loop MDP solver.
  - No inference time data collection (e.g. GAIL).
  - No action labels required.
- Scales to high-dimensional tasks (e.g. Atari games)

Trajectory-ranked Reward Extrapolation (T-REX)

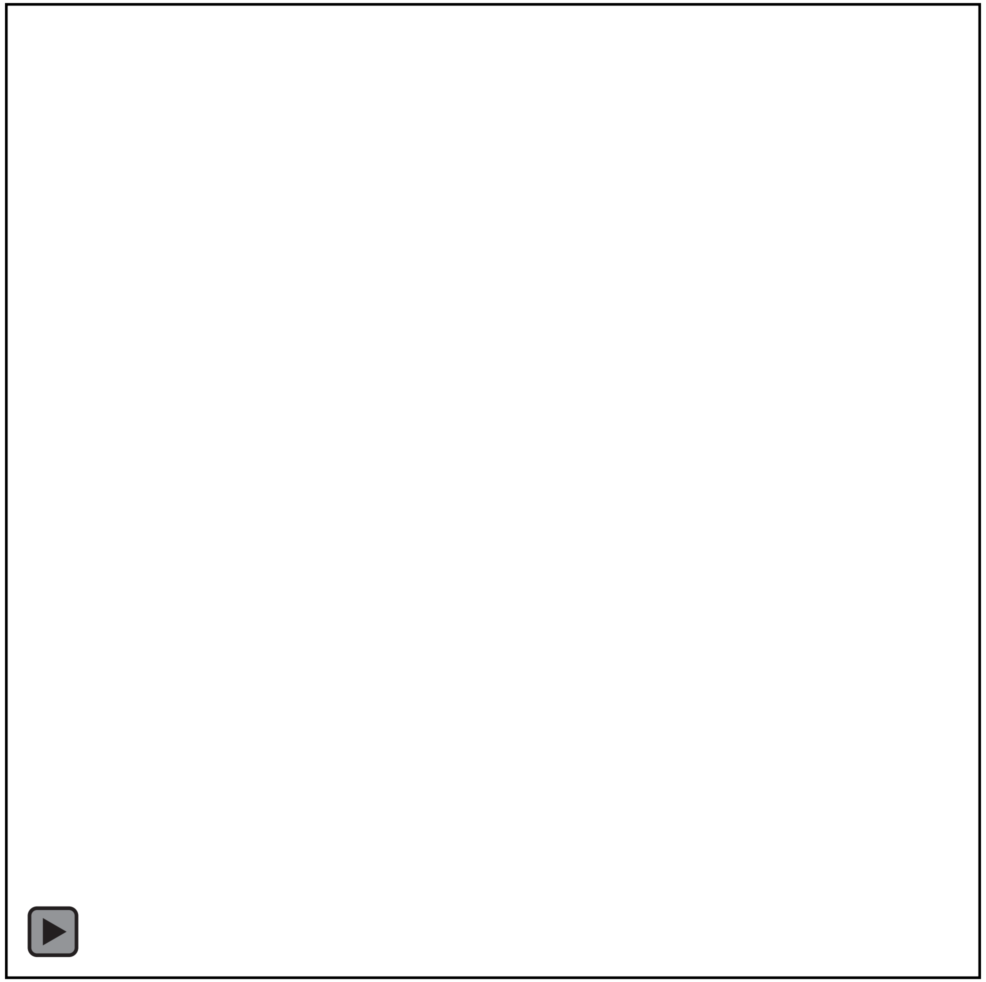


- Simple:
  - IRL as binary classification.
  - No human supervision during policy learning.
  - No inner-loop MDP solver.
  - No inference time data collection (e.g. GAIL).
  - No action labels required.
- Scales to high-dimensional tasks (e.g. Atari games)
- Can produce policies much better than demonstrator

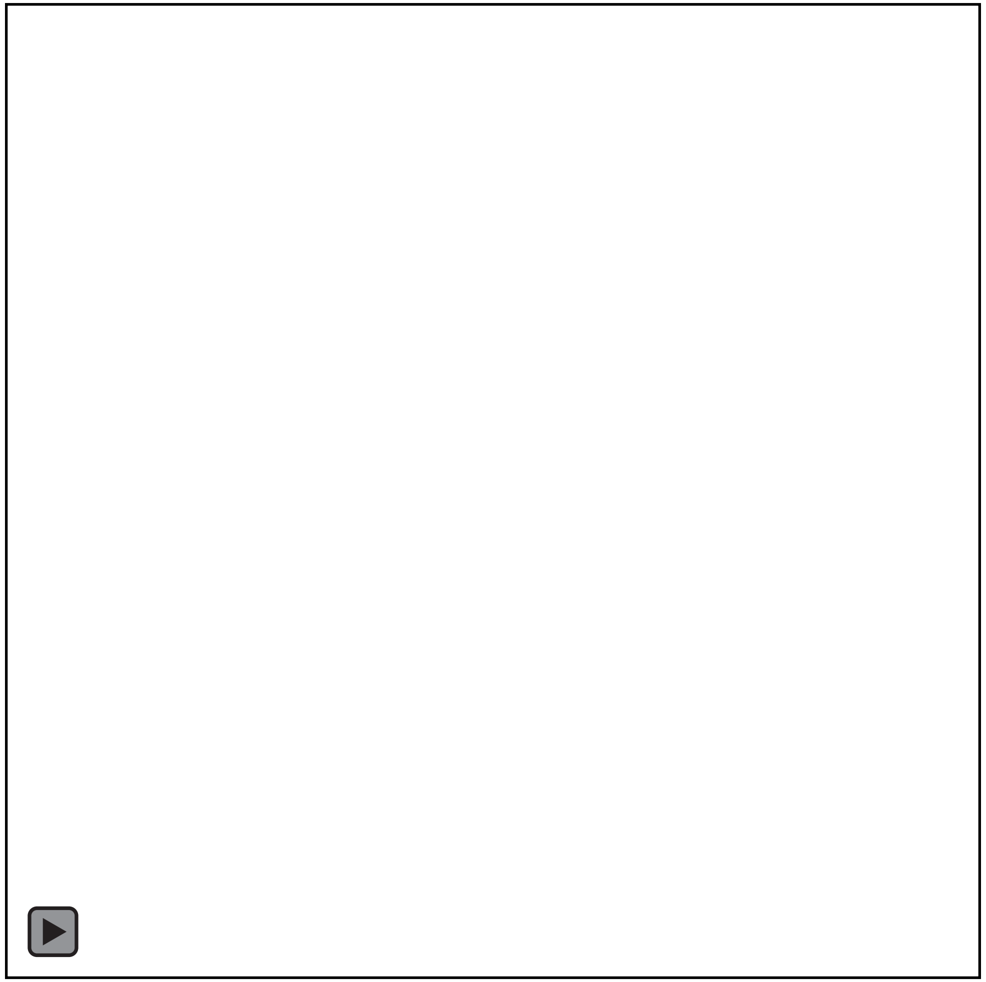
T-REX Policy Performance



T-REX on HalfCheetah



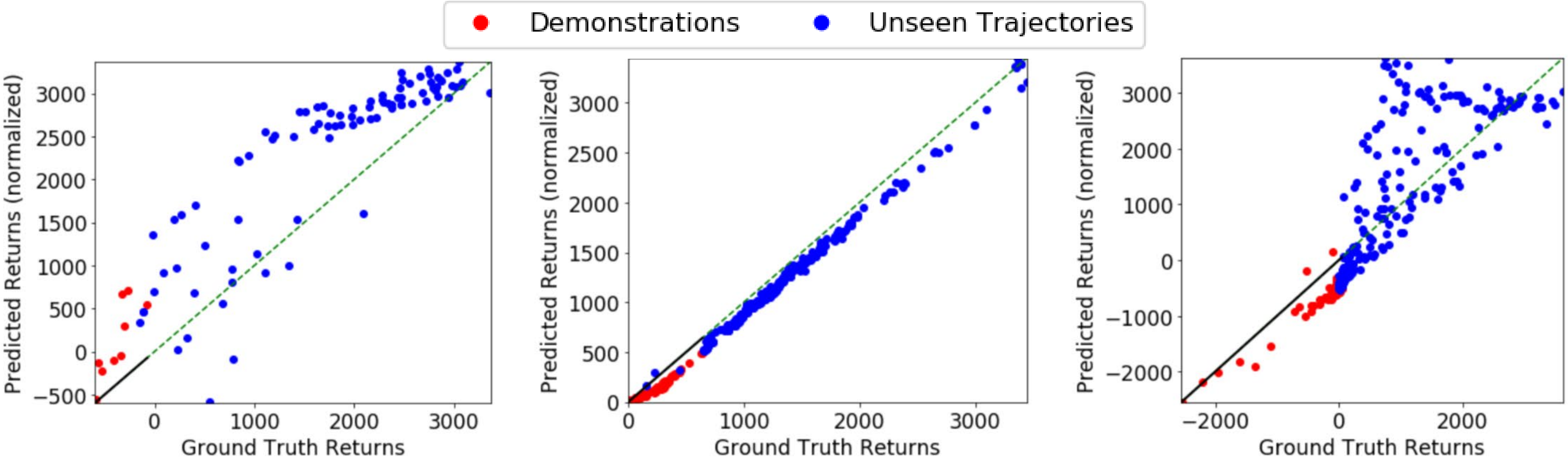
Best demo (88.97)



T-REX (143.40)

Reward Extrapolation

T-REX can extrapolate beyond the performance of the best demo



HalfCheetah

Hopper

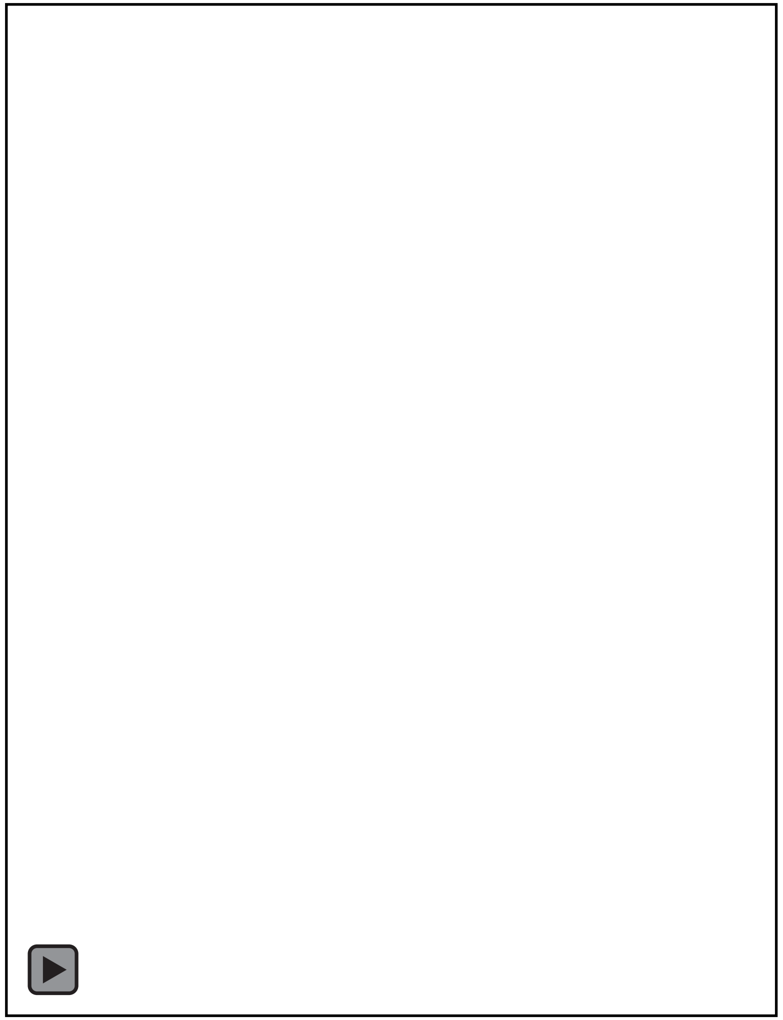
Ant

Results: Atari Games

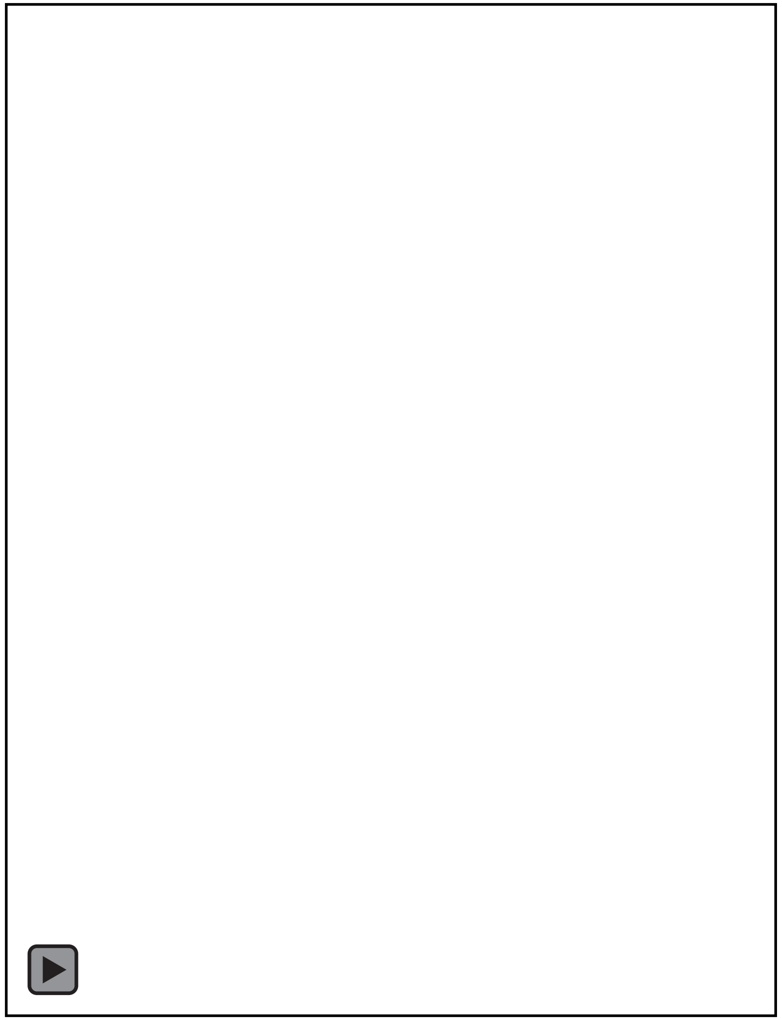
T-REX outperforms best demonstration on 7 out of 8 games!

	Ranked Demonstrations		LfD Algorithm Performance		
Game	Best	Average	T-REX	BCO	GAIL
Beam Rider	1,332	686.0	**3,335.7**	568	355.5
Breakout	32	14.5	**221.3**	13	0.28
Enduro	84	39.8	**586.8**	8	0.28
Hero	**13,235**	6,742.0	0	2,167	0
Pong	-6	-15.6	**-2.0**	-21	-21
Q*bert	800	627	**32,345.8**	150	0
Seaquest	600	373.3	**747.3**	0	0
Space Invaders	600	332.9	**1,032.5**	88	370.2

T-REX on Enduro



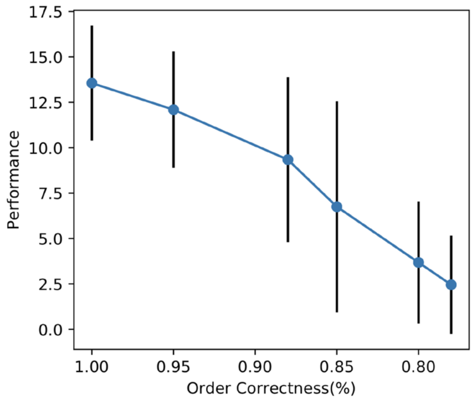
Best demo (84)



T-REX (520)

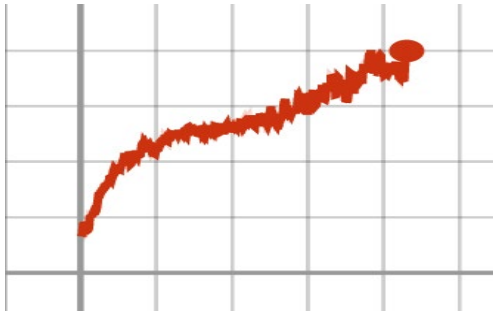
Come see our poster @ Pacific Ballroom #47

Robust to noisy ranking labels

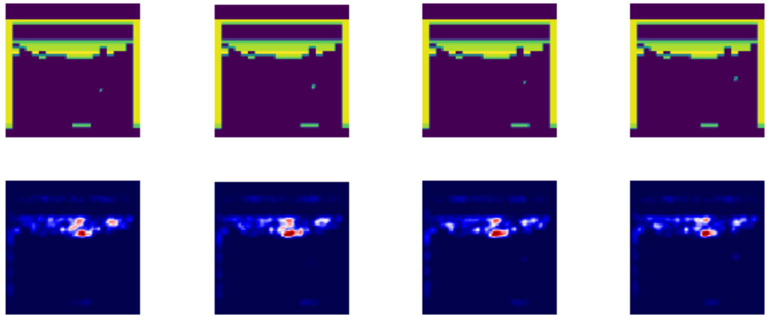


Human demos / ranking labels

Automatic ranking by watching a learner improve at a task

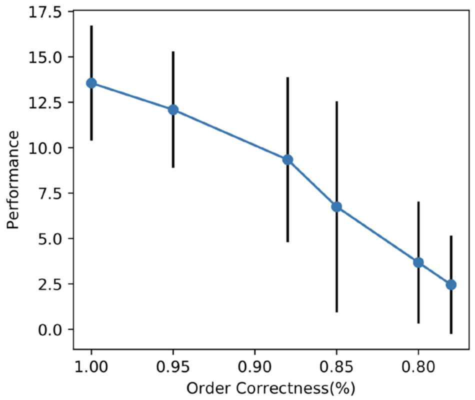


Reward function visualization

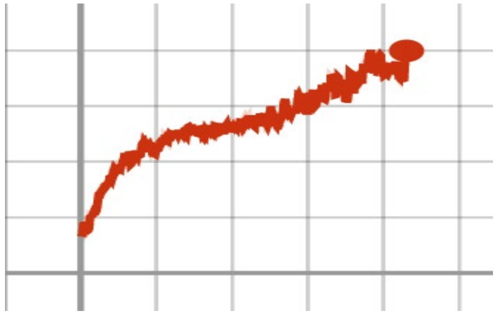


Come see our poster @ Pacific Ballroom #47

Robust to noisy ranking labels



Automatic ranking by watching a learner improve at a task



T-REX

Human demos / ranking labels

Reward function visualization