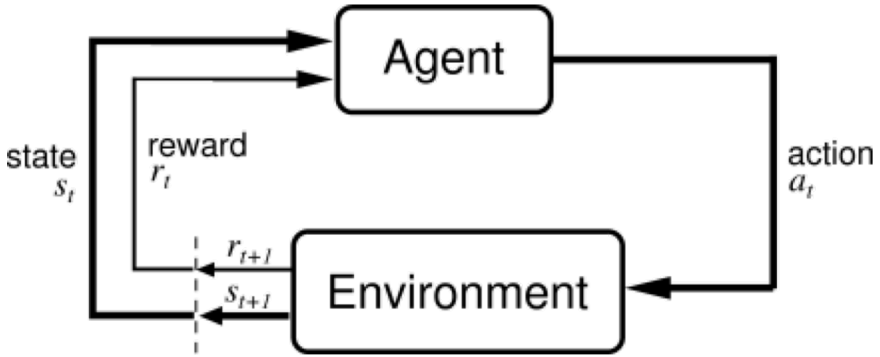
Reinforcement Learning (RL)



$$\max_{\theta} \quad \mathrm{E}[\sum_{t=0}^{H} R(s_t)|\pi_\theta]$$

- RL: agent learns from repeated interaction with environment
- Model-free RL:
  - interaction with real world

  → Improve learned policy or Q function
- Model-based RL:
  - interaction with real world

  → Improve learned environment simulator

  → interaction with learned simulator

  → Improve policy or Q function

[Image credit: Sutton and Barto 1998]

Canonical Model-Based RL

- for iter = 1, 2, …
  - collect data under current policy
  - improve learned simulator from all past data
  - improve policy by RL in learned simulator

Anticipated benefit?
– much better sample efficiency

So why not used all the time?
-- not achieving same asymptotic performance as model-free methods
-- “overfitting” (“model-bias”)

Model-based RL Asymptotic Performance

- Because learned (ensemble of) models imperfect
  - Resulting policy good in simulation(s), but not optimal in real world

- Attempted fix: learn better dynamics model
  - Such efforts have so far proven insufficient

Overfitting in Model-based RL

- Standard overfitting (in supervised learning)
  - Neural network performs well on training data, but poorly on test data
    - E.g. on prediction of s_next from (s, a)

- New overfitting challenge in Model-based RL
  - policy optimization tends to exploit regions where insufficient data is available to train the model, leading to catastrophic failures
  - = "model-bias" (Deisenroth & Rasmussen, 2011; Schneider, 1997; Atkeson & Santamaria, 1997)

Recall: Domain Randomization + Few-Shot RL!

- Key idea:
  - No need to learn an accurate model
  - Suffices to learn a set of models representative of the real world
  - And then run few-shot RL in that set of models

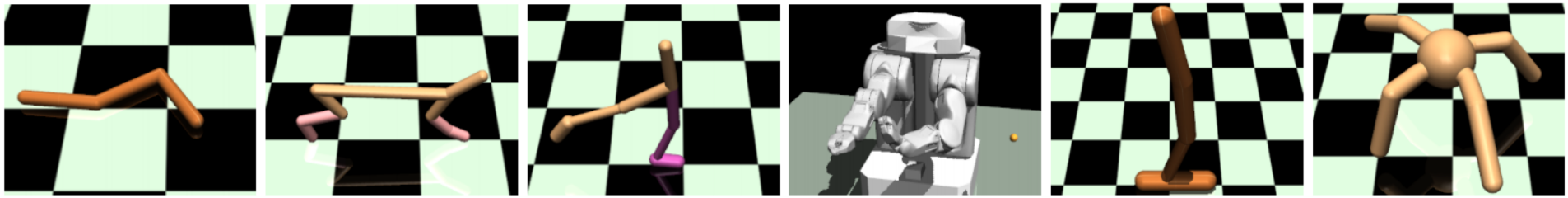
Model-Based RL via Meta Policy Optimization (MB-MPO)

for iter = 1, 2, …

- collect data under current adaptive policies $\pi_{\theta'_1}, \ldots, \pi_{\theta'_K}$
- learn **ENSEMBLE** of K simulators from all past data
- **meta-policy optimization over ENSEMBLE**
  - → **new meta-policy** π_θ
  - → **new adaptive policies** $\pi_{\theta'_1}, \ldots, \pi_{\theta'_K}$

[Clavera*, Rothfuss*, Schulman, Fujita, Asfour, Abbeel, CoRL 2018]

MB-MPO Evaluation



[Clavera*, Rothfuss*, Schulman, Fujita, Asfour, Abbeel, CoRL 2018]

MB-MPO Evaluation

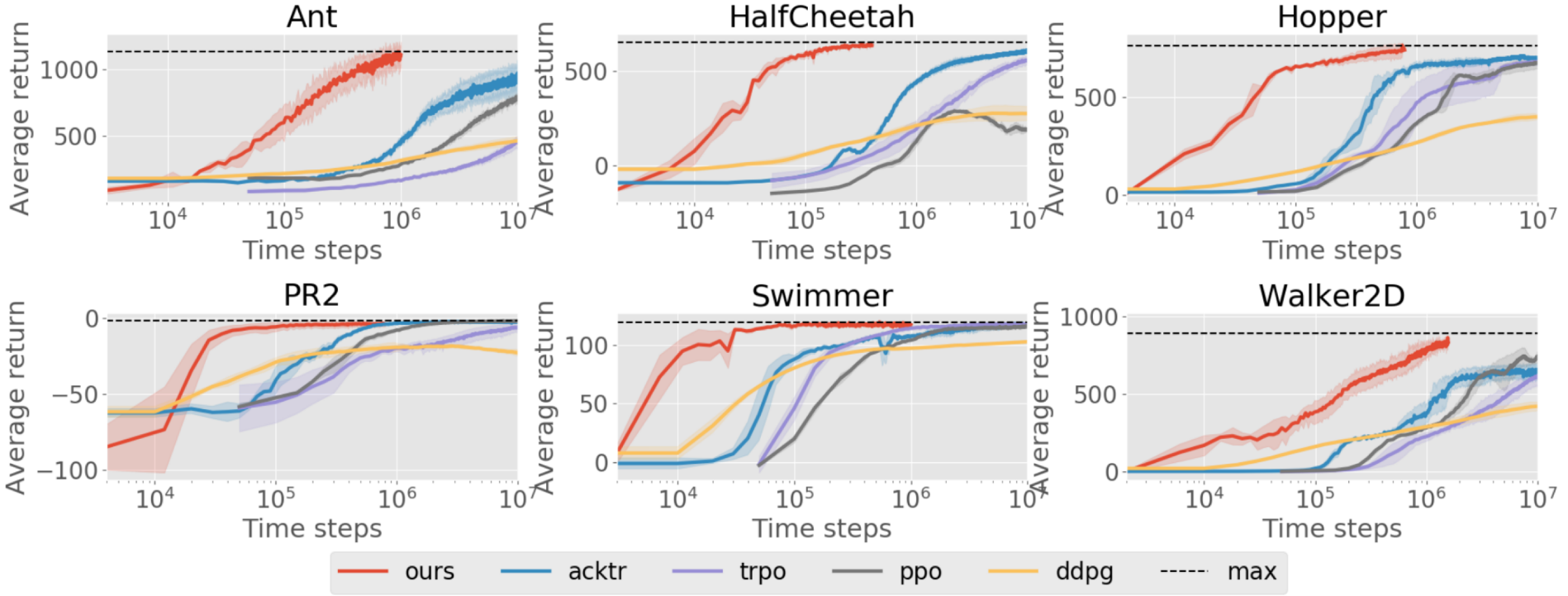
[Clavera*, Rothfuss*, Schulman, Fujita, Asfour, Abbeel, CoRL 2018]

MB-MPO Evaluation

[Clavera*, Rothfuss*, Schulman, Fujita, Asfour, Abbeel, CoRL 2018]

MB-MPO Evaluation

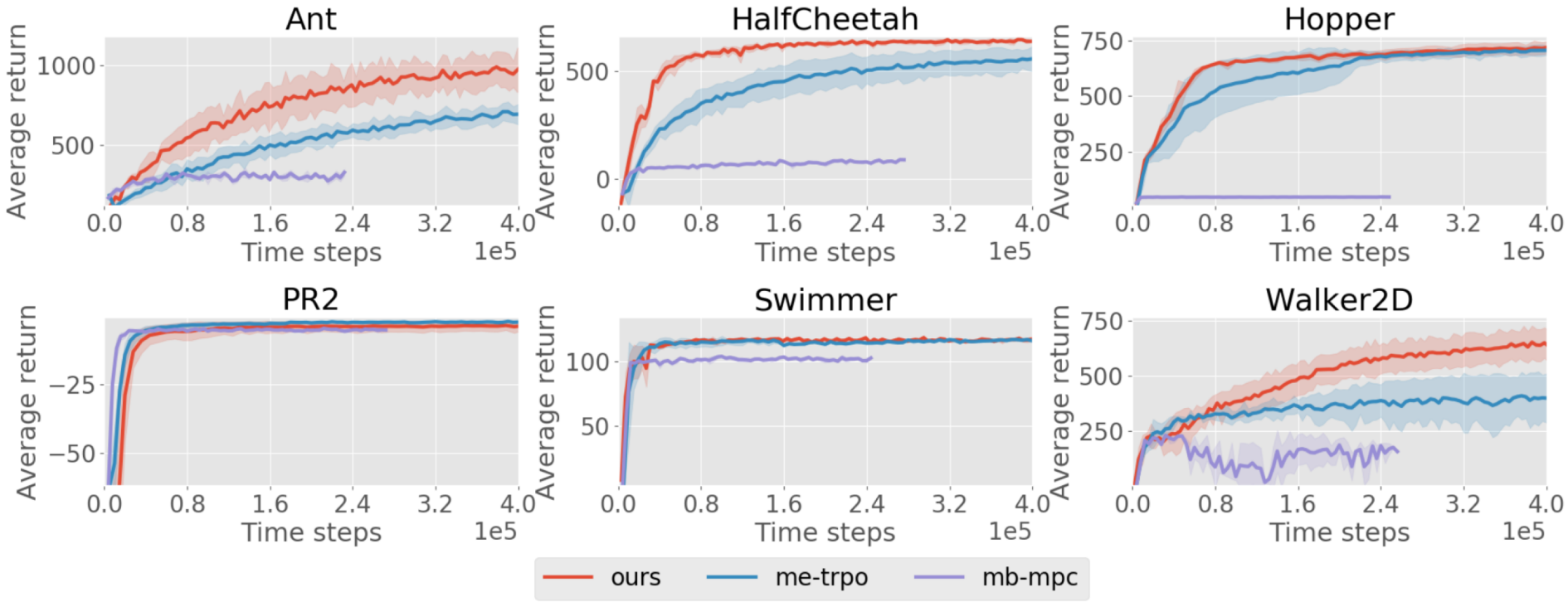
- Comparison with state of the art model-free



[Clavera*, Rothfuss*, Schulman, Fujita, Asfour, Abbeel, CoRL 2018]

MB-MPO Evaluation

- Comparison with state of the art model-based



[Clavera*, Rothfuss*, Schulman, Fujita, Asfour, Abbeel, CoRL 2018]