Towards Autonomous RL

Learning to Act with Less Human Supervision

Ben Eysenbach
PhD student in MLD
beysenba@cs.cmu.edu

Abhishek Gupta
PhD student at UC Berkeley
abhigupta@eecs.berkeley.edu

Jan 27, 2021

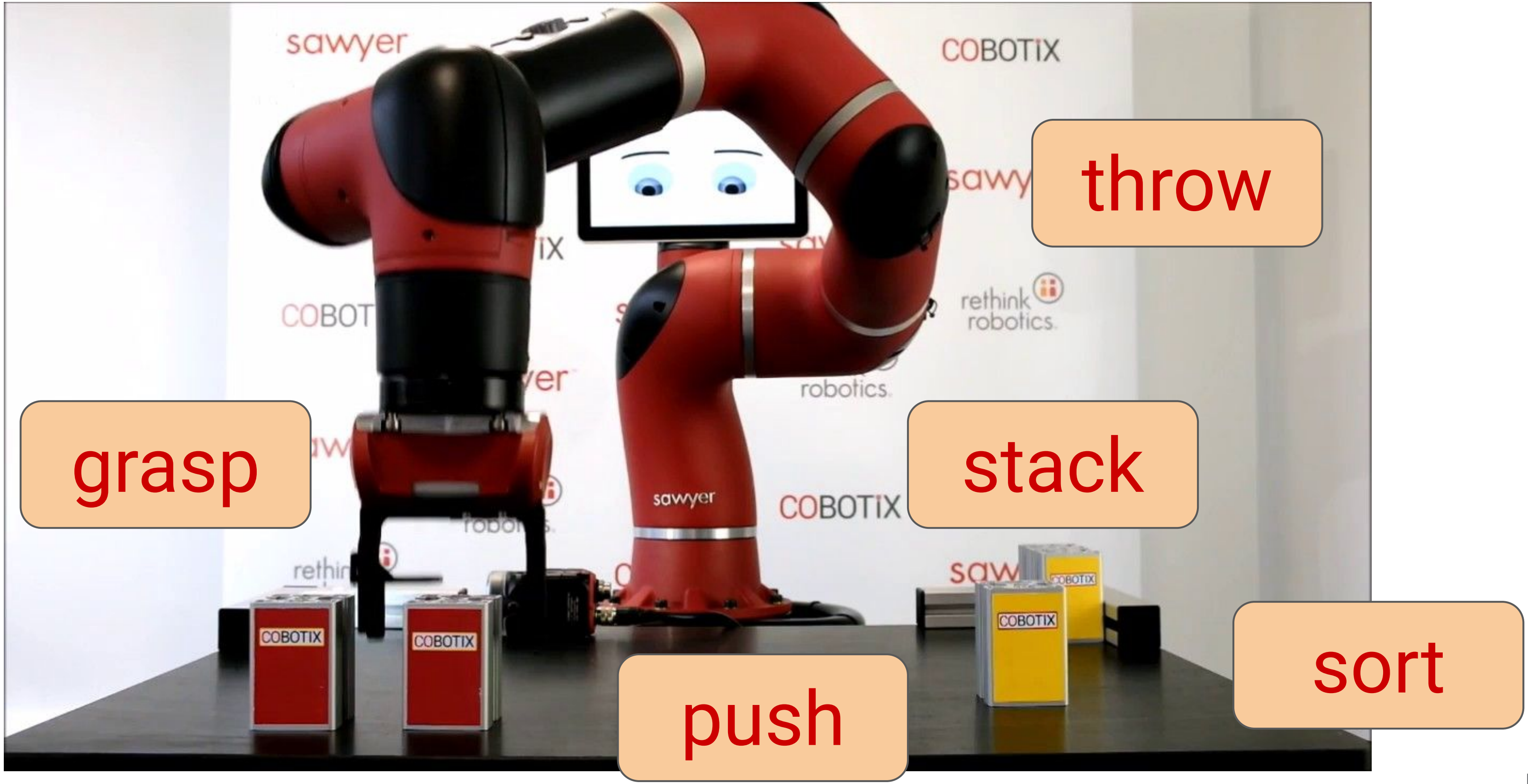
Challenge: Current RL Requires Human Supervision

Need human supervision for:

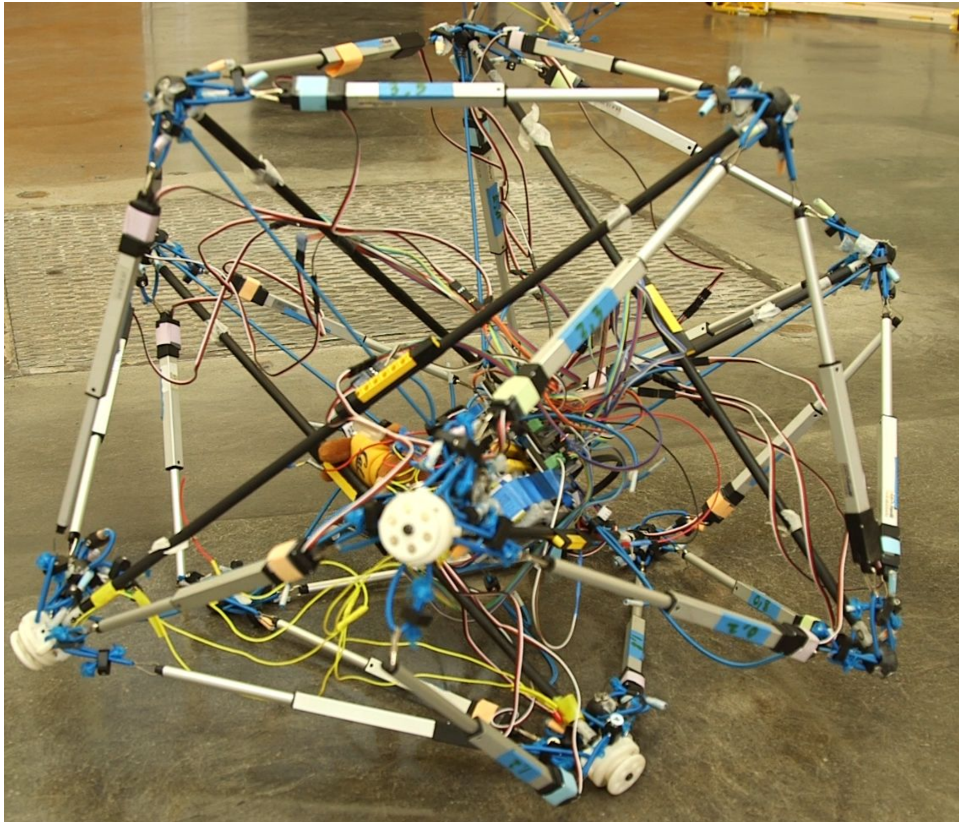
- Designing reward functions
- Specifying useful skills
- Tuning parameters of learning algorithm
- Resetting
- Avoiding dangerous states
- Designing a curriculum

What is a Skill?

What is a Skill?



What is a Skill?



What is a Skill?

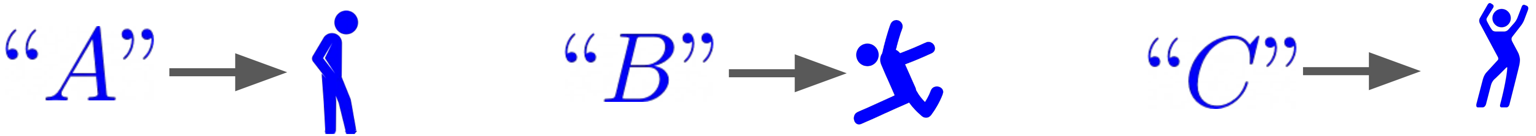
Properties of good skills:

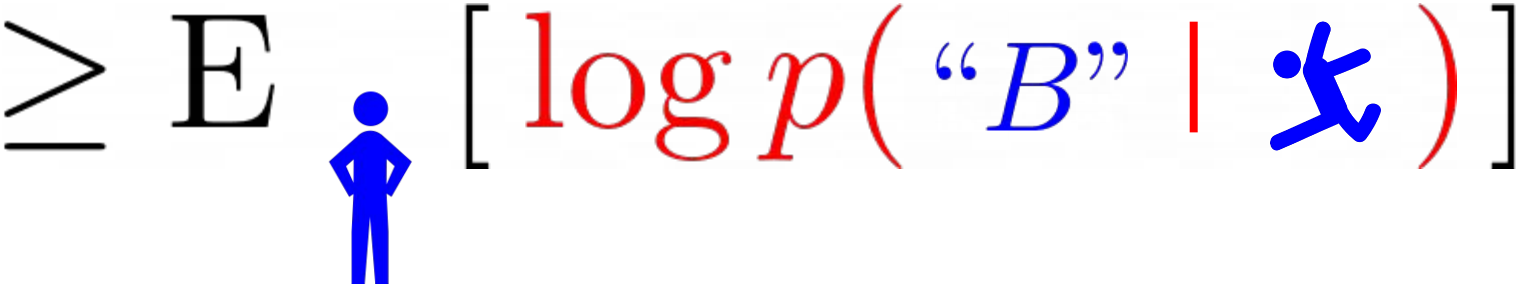
- *Exploration* - at most one skill “dithers”; forces skills to explore large regions of the state space.
- *Predictability* - want to predict what a skill will do (important for hierarchical RL).
- *Interpretability* - easy to infer which skill is being executed at any given point in time.

Idea: Learn a **set** of skills that is **as diverse as possible**.


10
20
30
40
50

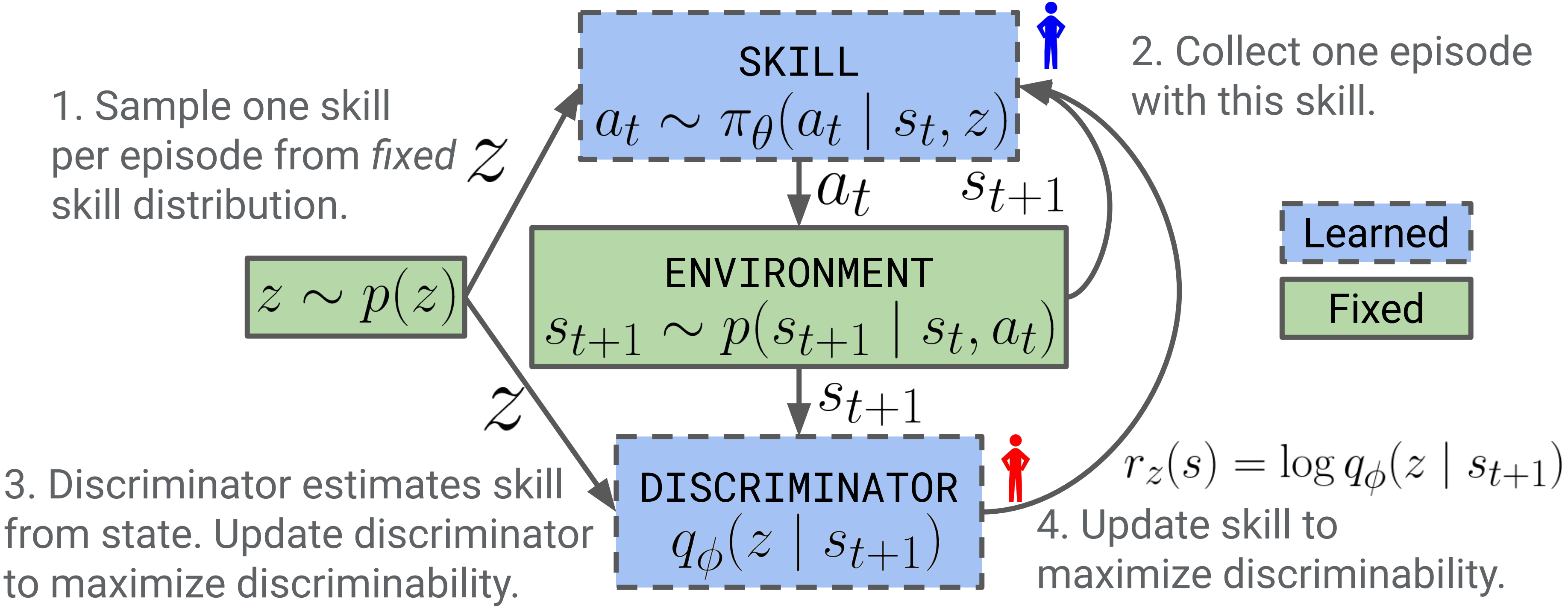
How many bits of information can communicate to ?





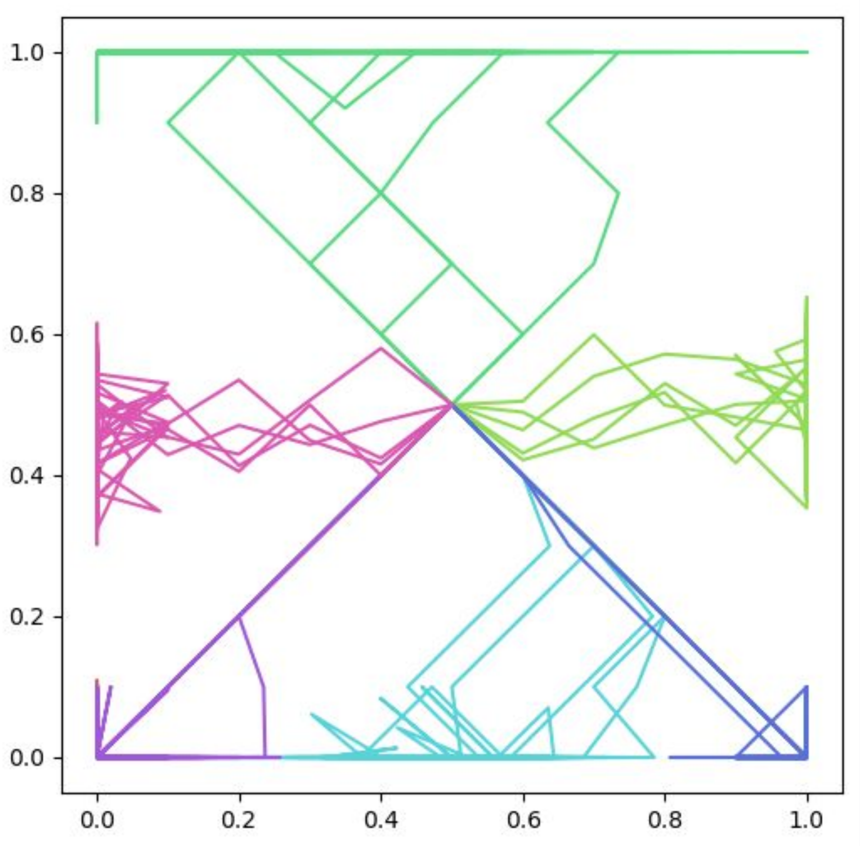
Diversity Is All You Need [DIAYN]

DIAYN: How does the algorithm work?

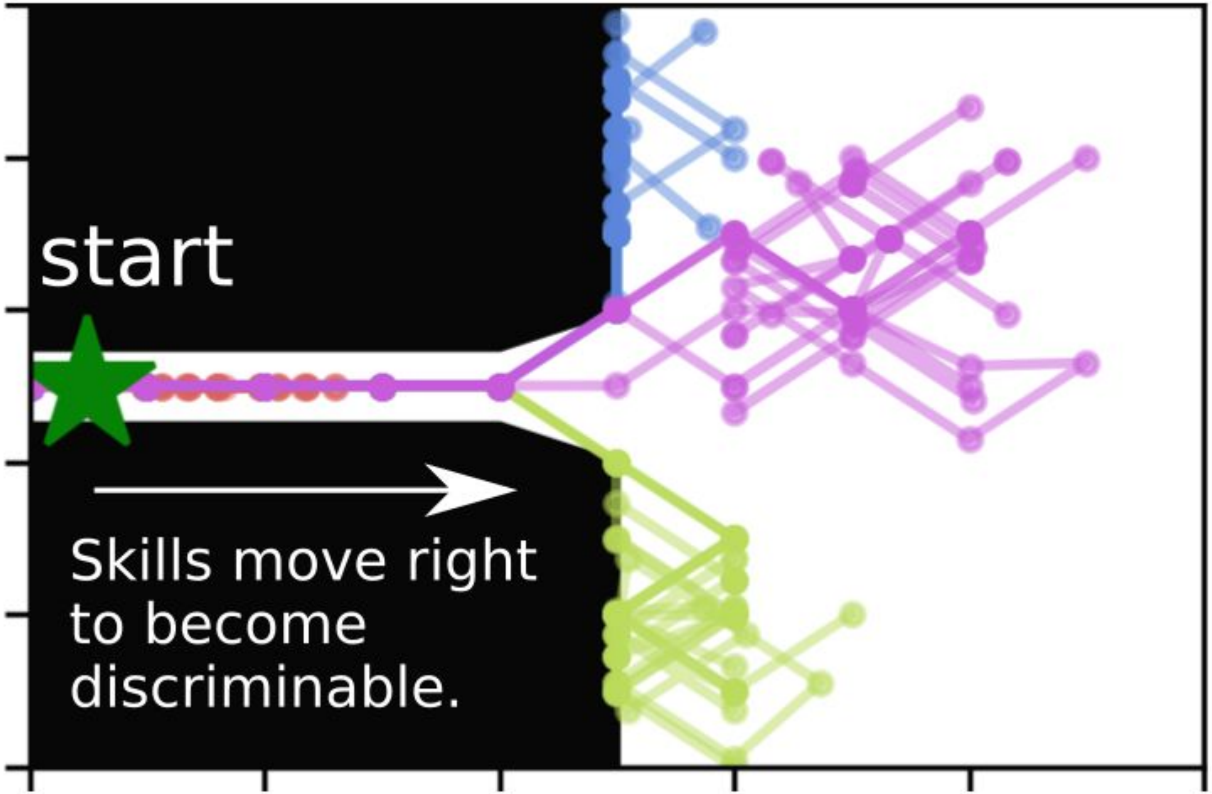


Visualizing DIAYN

- *Exploration*
- *Predictability*
- *Interpretability*

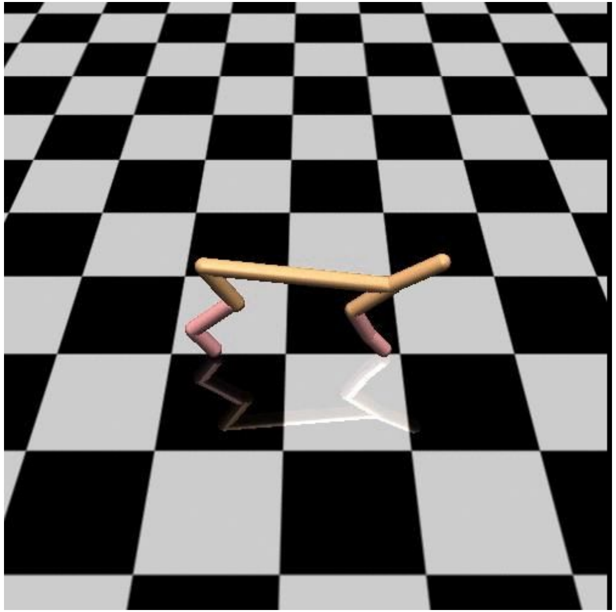


DIAYN maximizes *future* diversity.

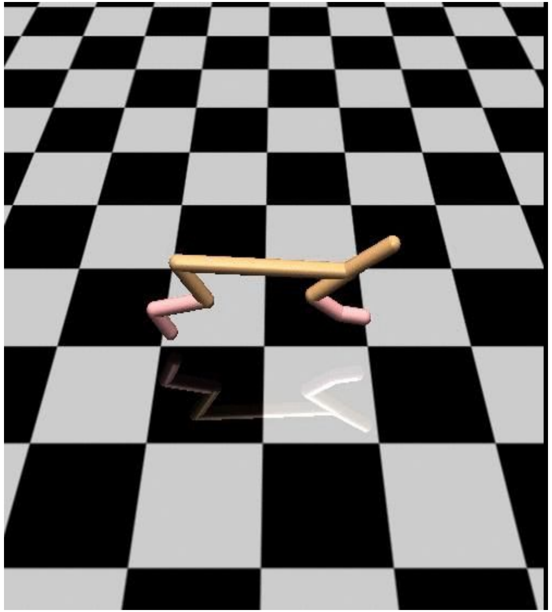


What Skills are Learned?

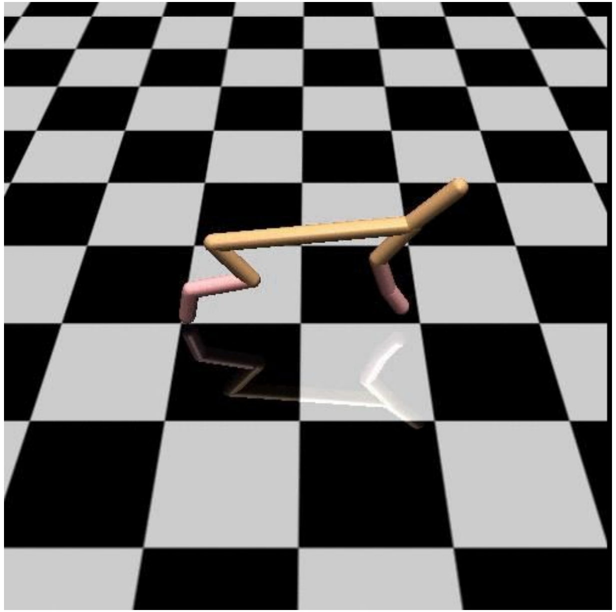
Walking forwards



Running Backwards

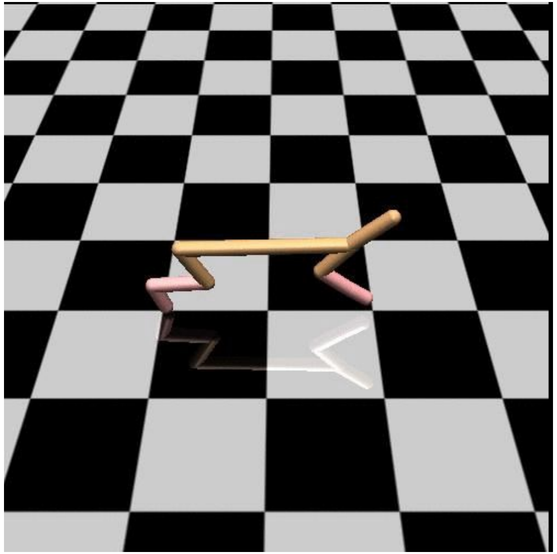


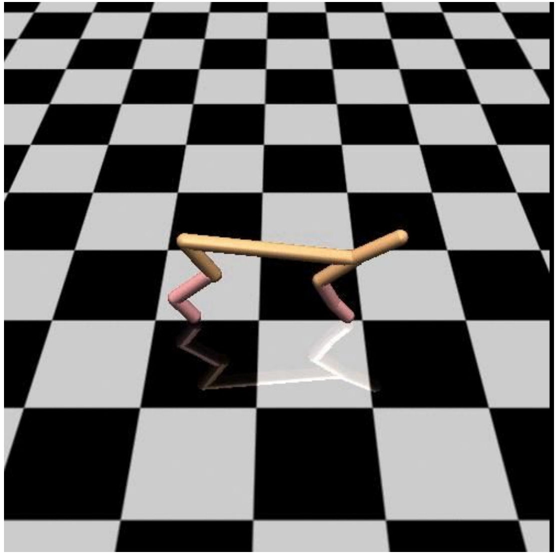
Front flips

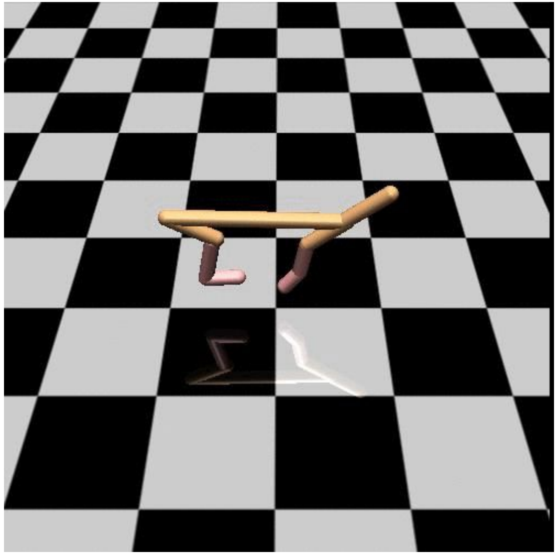


What Skills are Learned?

Skills for different forward gaits

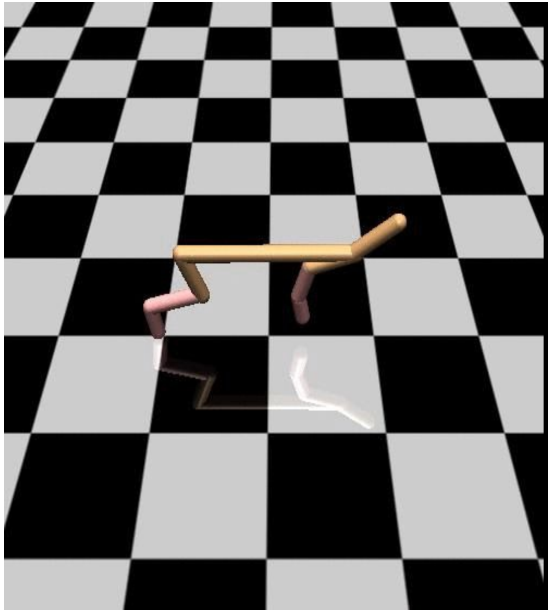
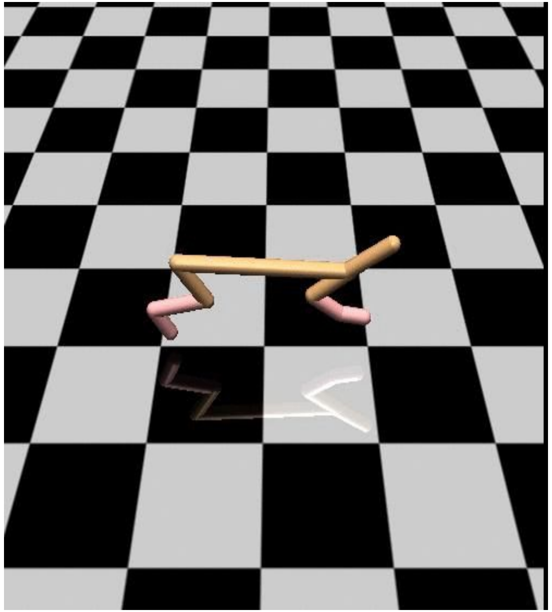
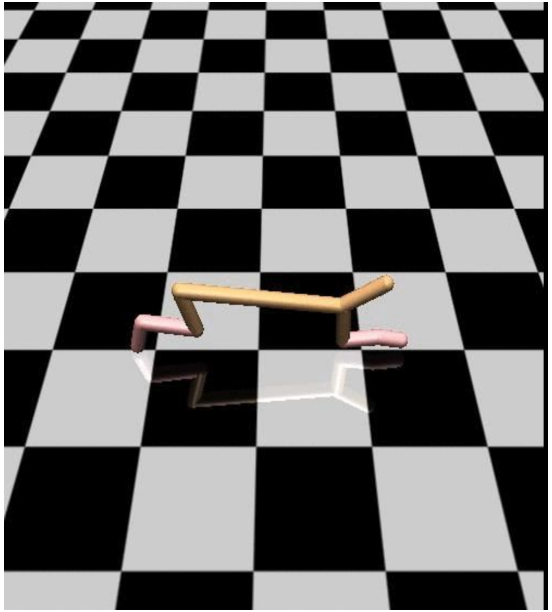






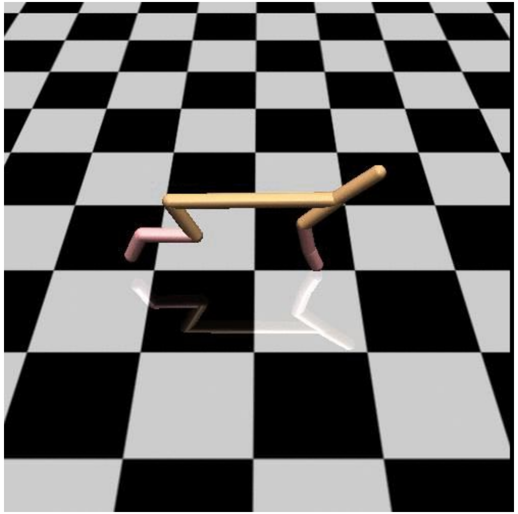
What Skills are Learned?

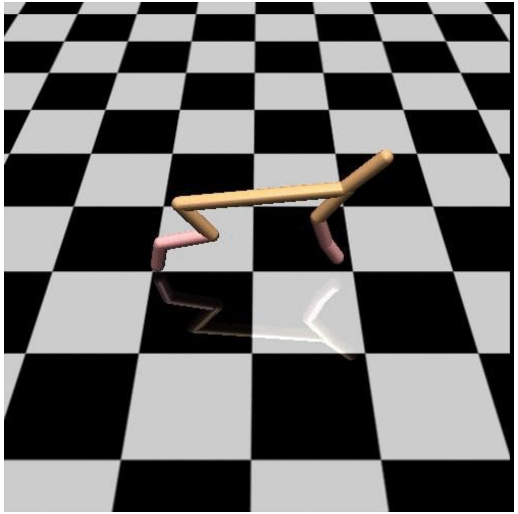
Skills for different backward gaits

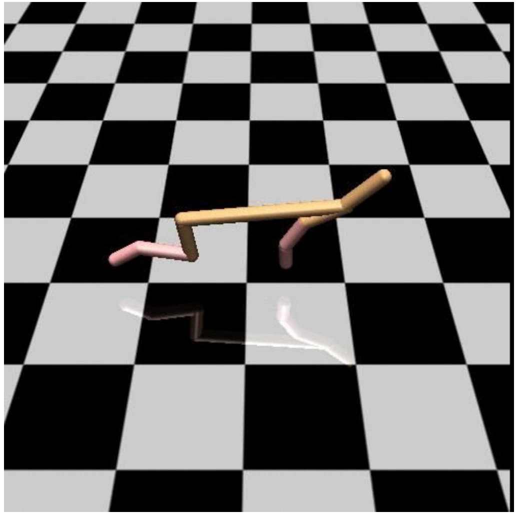




What Skills are Learned?

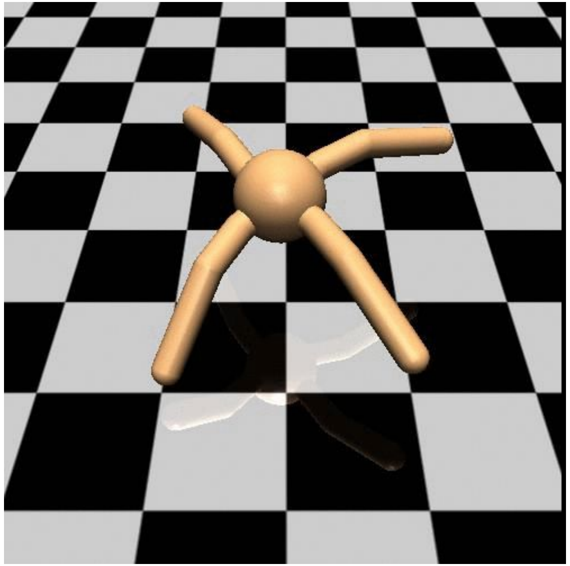
Skills for different front flips

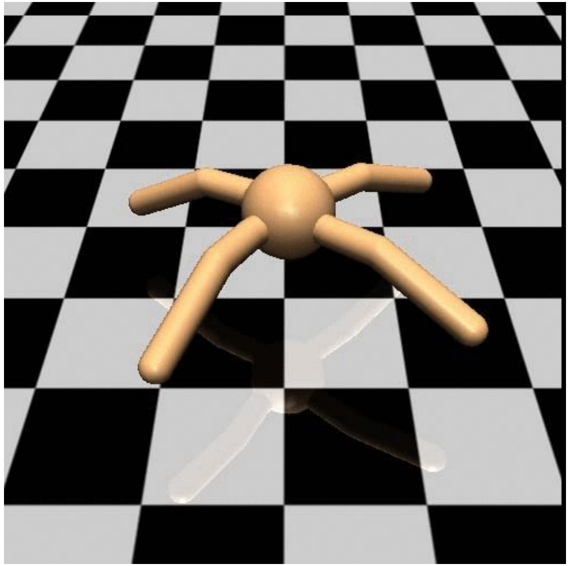


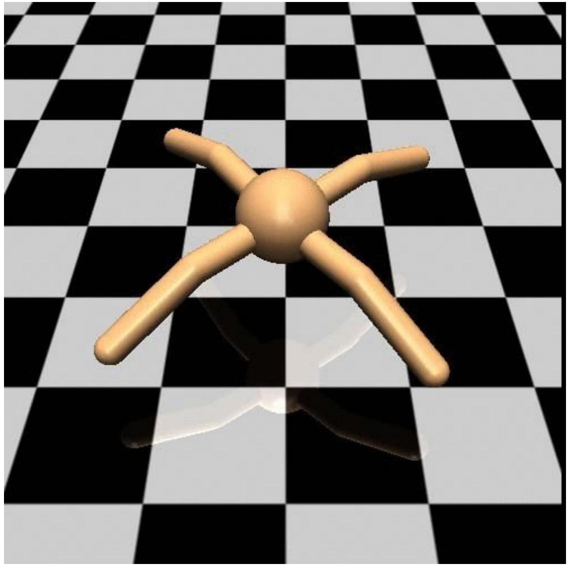




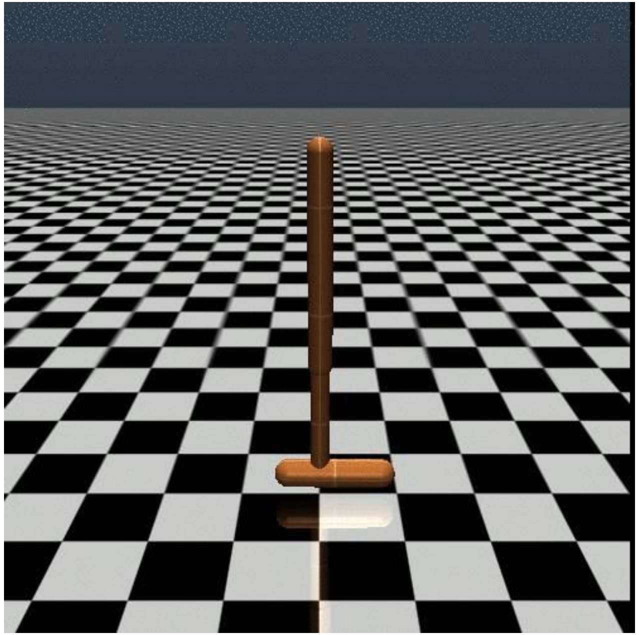
What Skills are Learned?

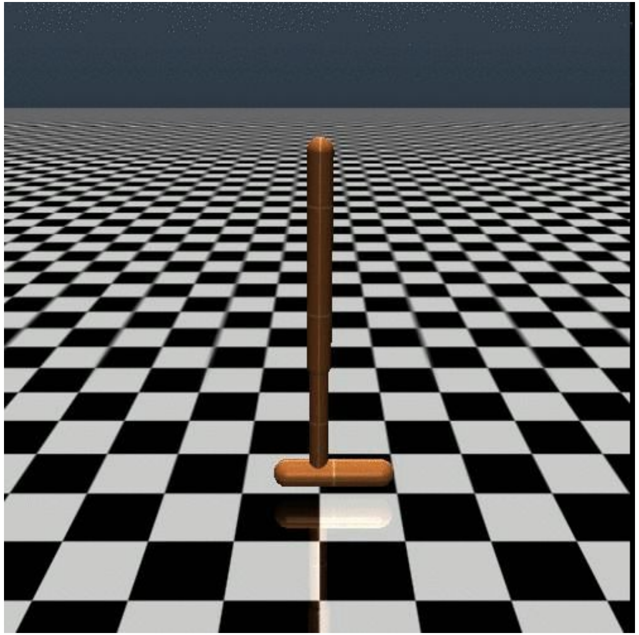




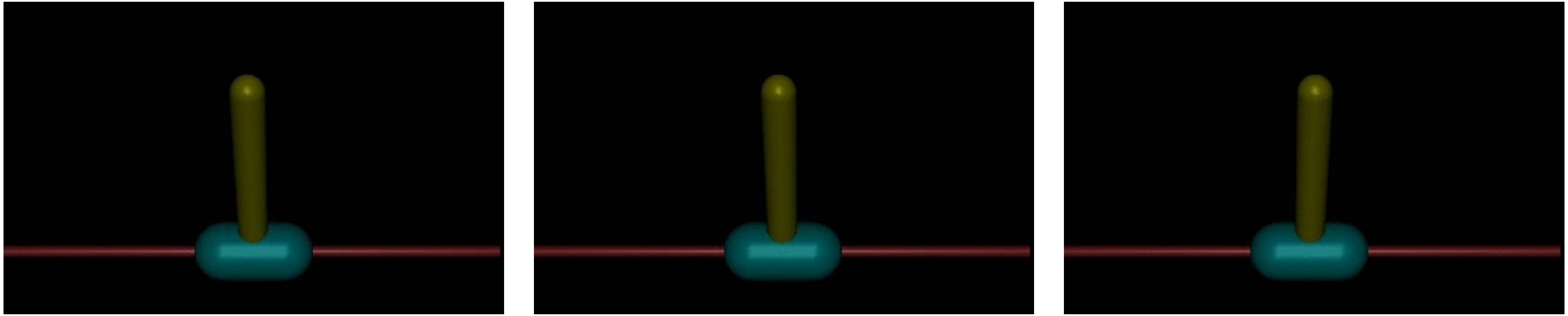


What Skills are Learned?





What Skills are Learned?



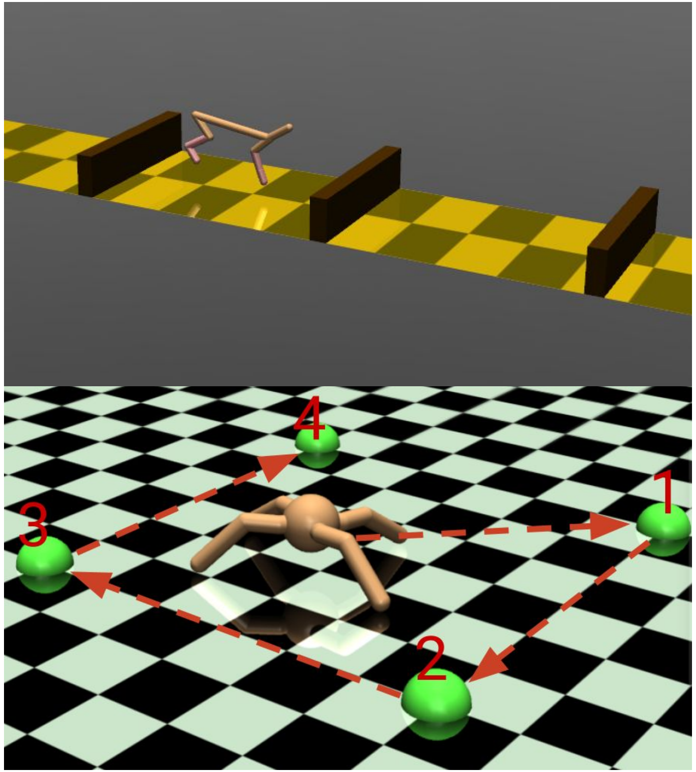
Why is DIAYN useful?

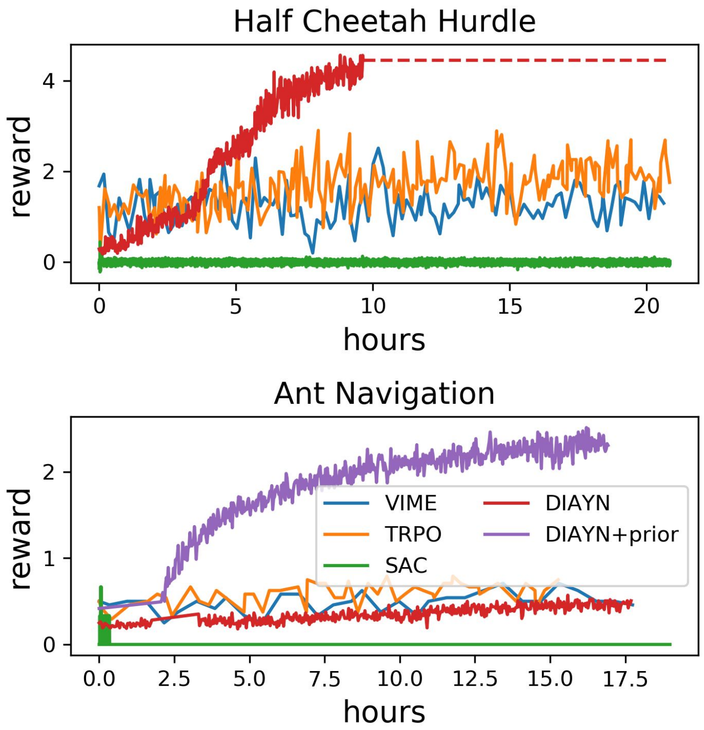
Returns a policy $\pi_\theta(a \mid s, z)$ *with a low dimensional knob that spans a large set of behaviors.*

Applications of DIAYN:

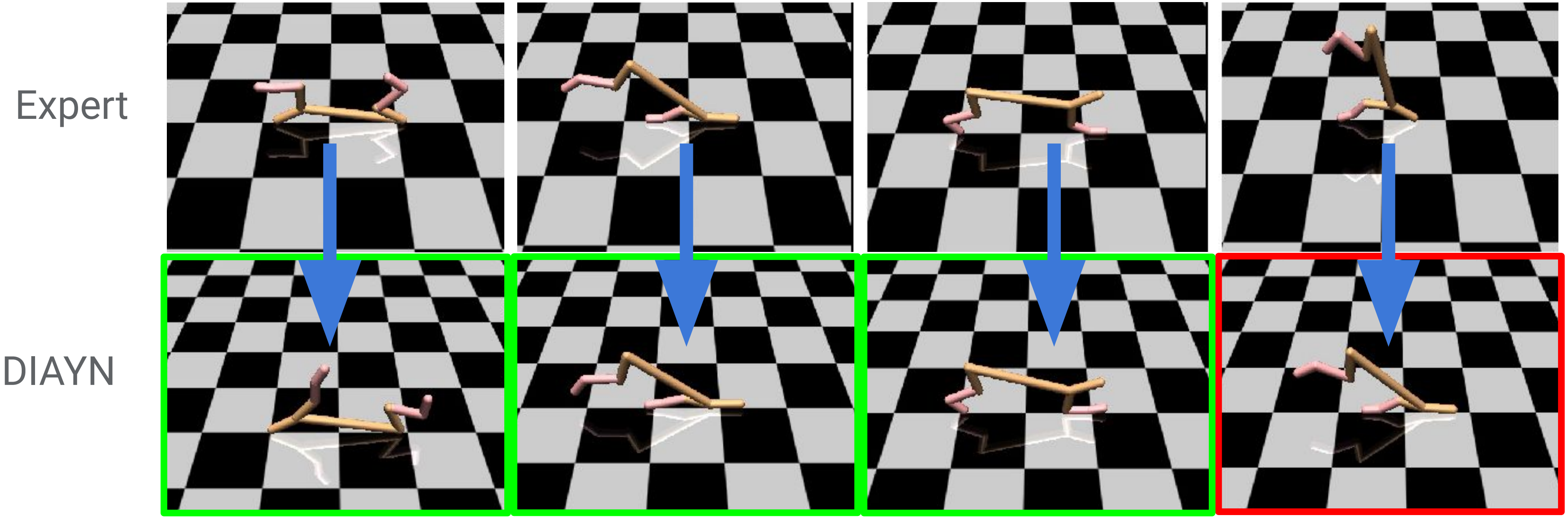
- **Hierarchical RL**
- **Imitation Learning**
- Learn an environment-specific policy initialization
- **Unsupervised Meta-Learning**

DIAYN for Hierarchical RL





DIAYN for 0-Shot Imitation Learning



See [Pathak '18]

Is Diversity *Really* All You Need?

DIAYN

- learns a latent-conditioned policy that can span many skills
- Does not explicitly optimize for future fine-tuning

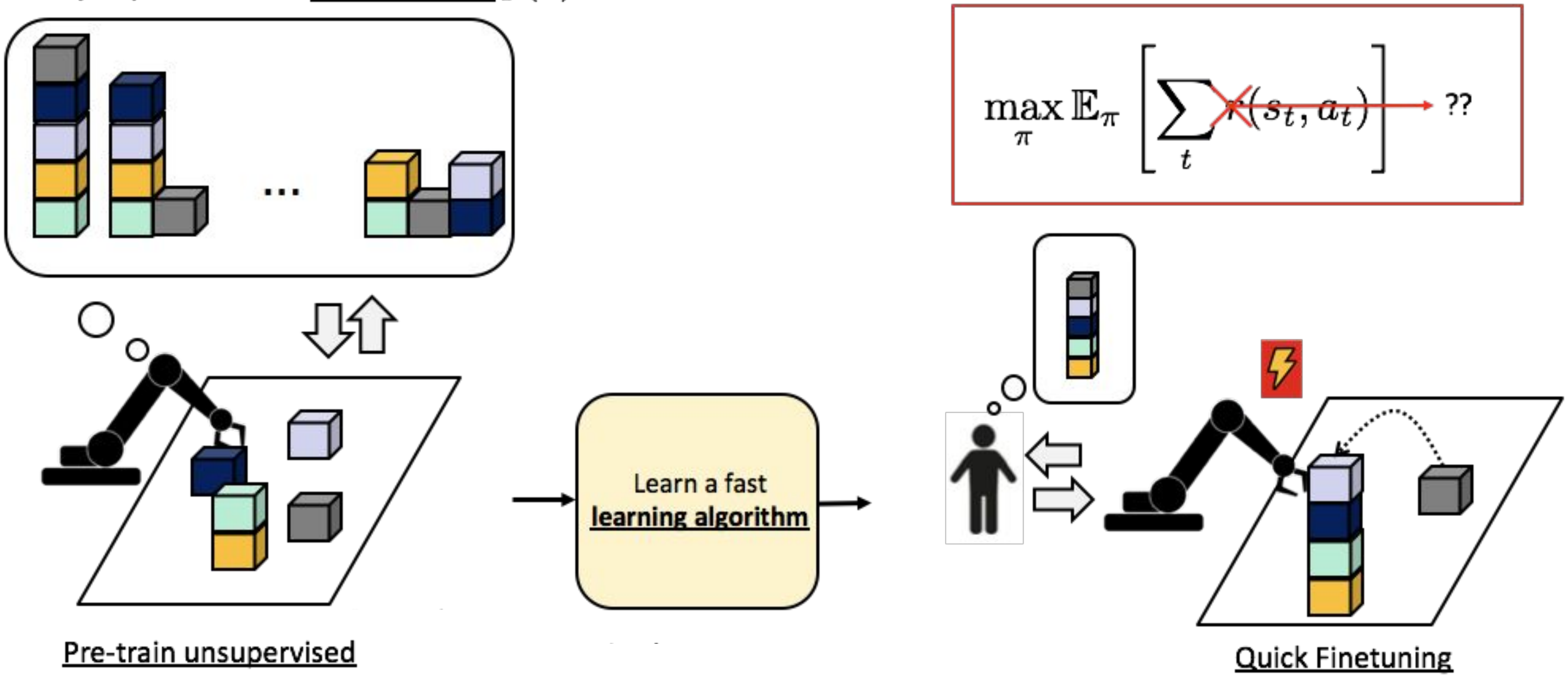
How can we best prepare for the future?

+

Meta-Learning?

A Principled Framework for Unsupervised Meta-RL

Self-propose a task **distribution** $p(\tau)$ to learn on



Formalizing Unsupervised Task Proposals

Learning procedure

Task distribution

Returns

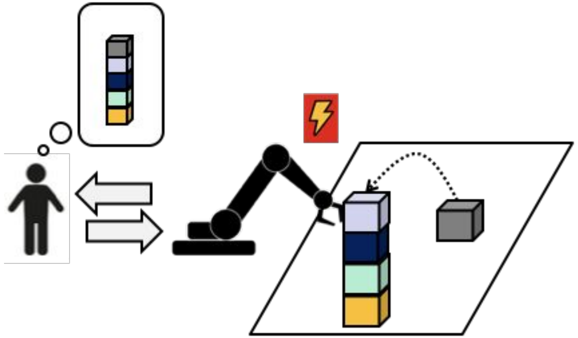
$$\text{Regret}(f, p) = E_{\text{task}\sim p(T)} \sum_i R(\pi_i, \text{task}) - R(\pi_i^*, \text{task})$$

$$\pi_i = f(\pi_{i-1}, \text{task})$$ Update of a learning procedure.

$$\pi_i^* = f^*(\pi_{i-1}^*, \text{task})$$ Update of the optimal learning procedure.

Regret

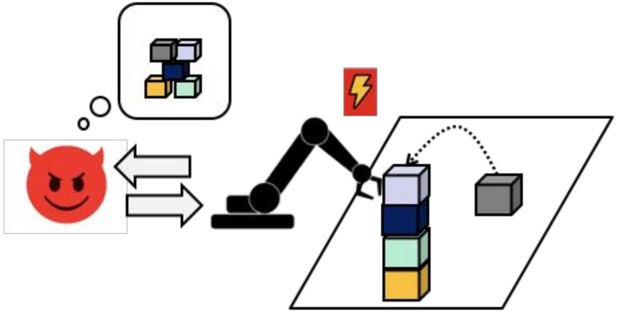
$$\min_f E_{\mathcal{T}\sim p(\mathcal{T})}[\text{Regret}(f, \mathcal{T})]$$



Known test task distribution

Worst Case Regret

$$\min_f \max_p E_{\mathcal{T}\sim p(\mathcal{T})}[\text{Regret}(f, \mathcal{T})]$$



Adversarial worst-case test task distribution

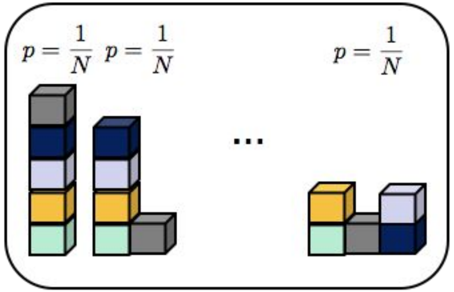
Optimizing Worst Case Regret

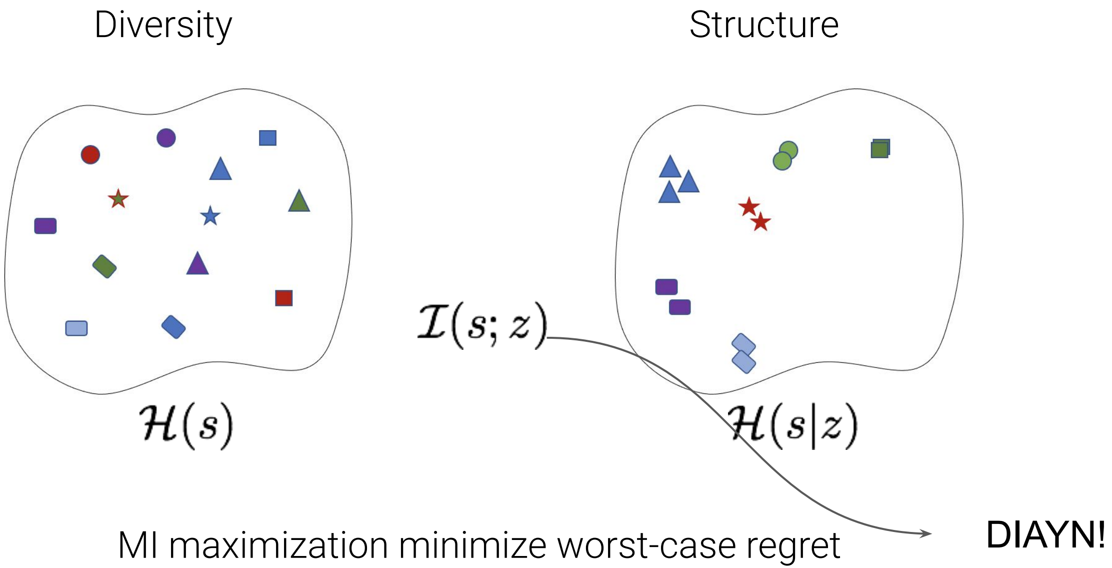
Worst Case Regret

$$\min_{f} \max_{p} E_{\mathcal{T} \sim p(\mathcal{T})}[\text{Regret}(f, \mathcal{T})]$$

$$\Big\downarrow$$

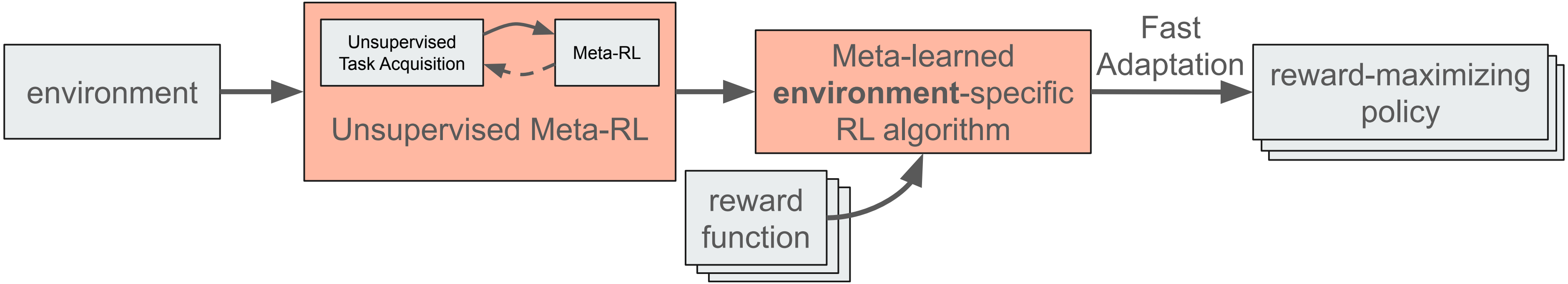
Uniform Distribution $\mathcal{U}(\tau)$





Preparing for the Future: Meta-RL

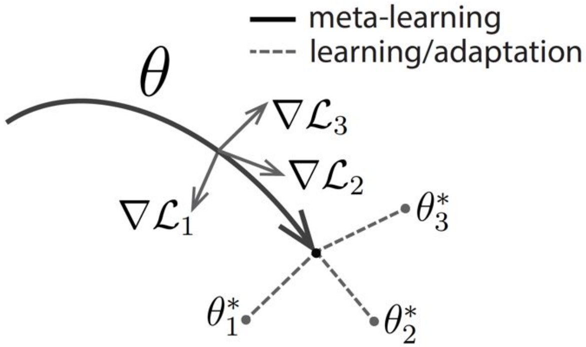
Idea: Do meta-learning on DIAYN skills to learn a good, environment-specific learning algorithm.



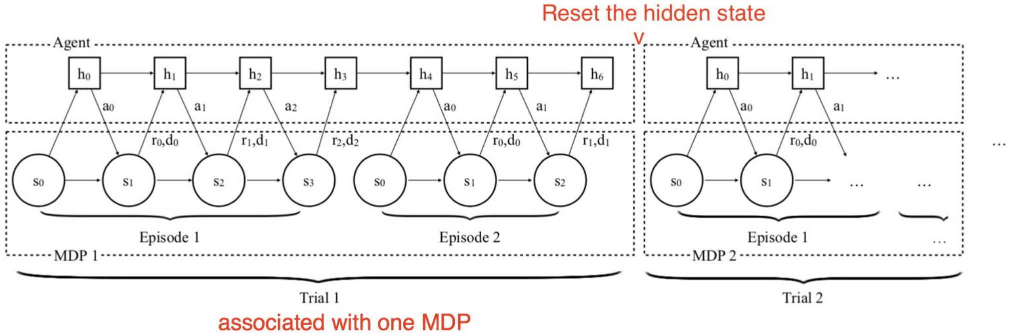
Meta-Learning with Unsupervised Task Distributions

What meta-learning algorithm is suitable?

Gradient Based Meta-Learners

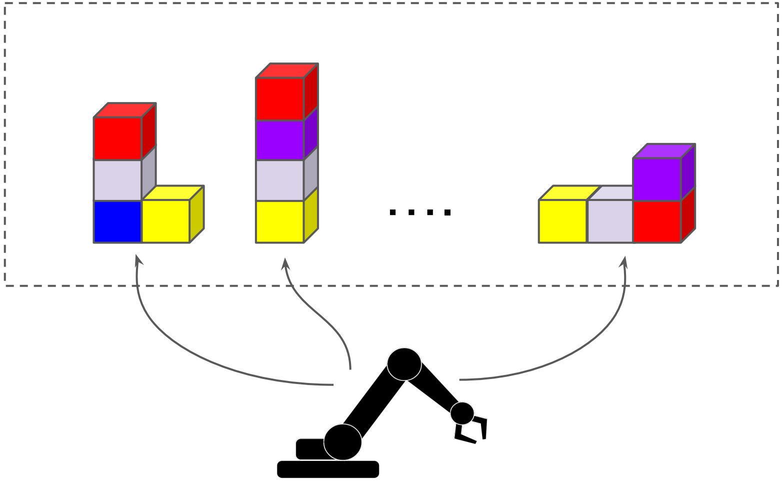


Recurrent Meta-Learners



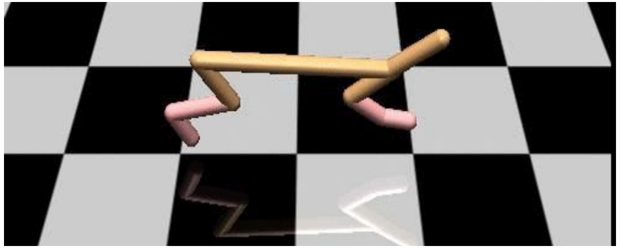
What about No Free Lunch?

Why would this help at all?

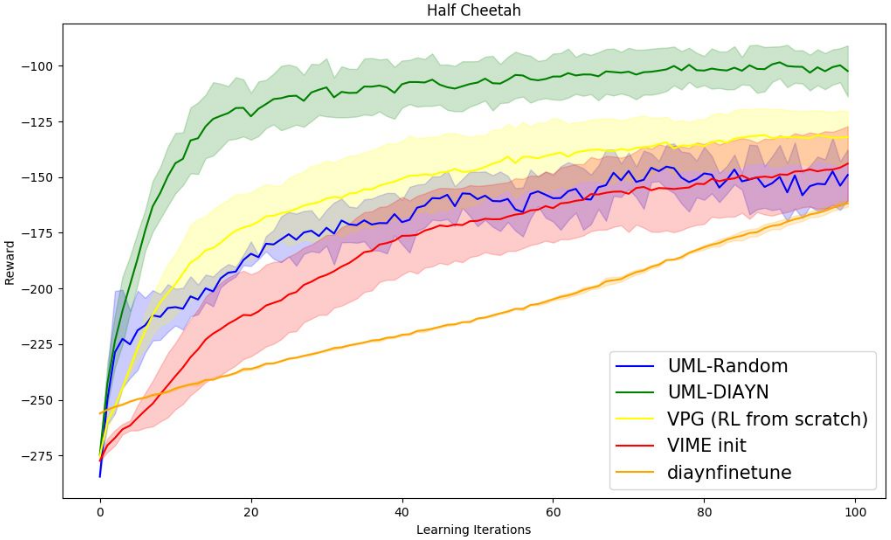


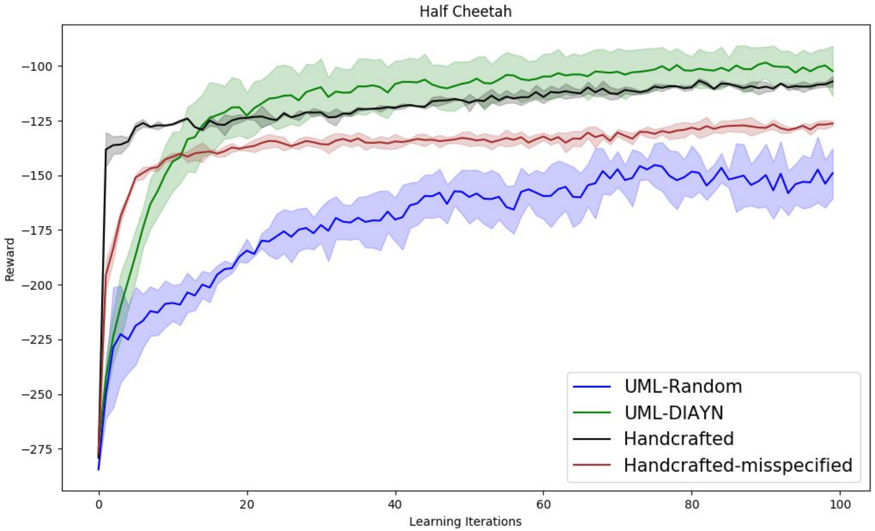
Exploring in the same environment provides the free lunch!

Learning Quickly with Unsupervised Meta-Learning

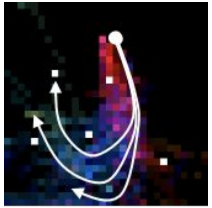


Quicker fine-tuning with provided rewards!

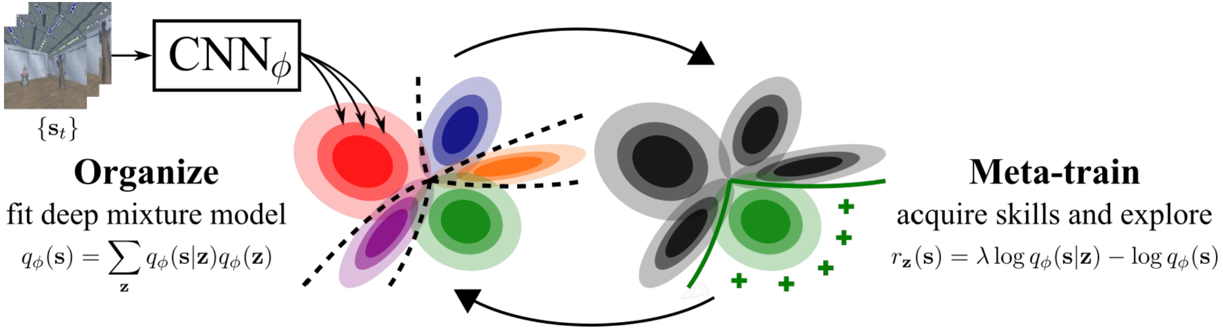




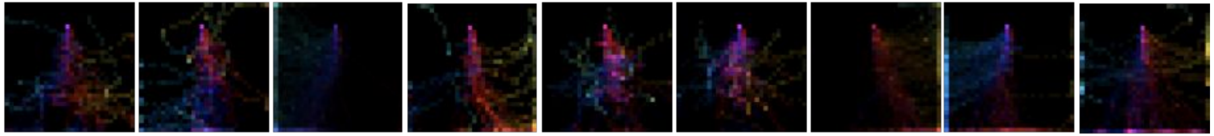
Learning Unsupervised Curricula



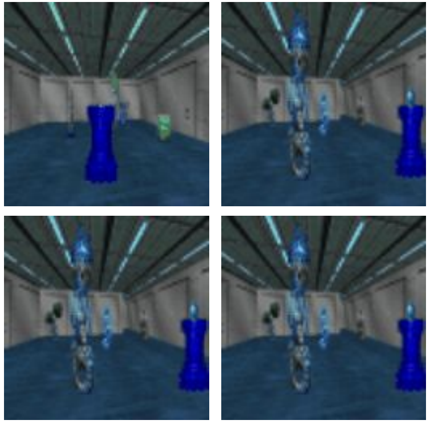
Direction encoded as color



Step 1

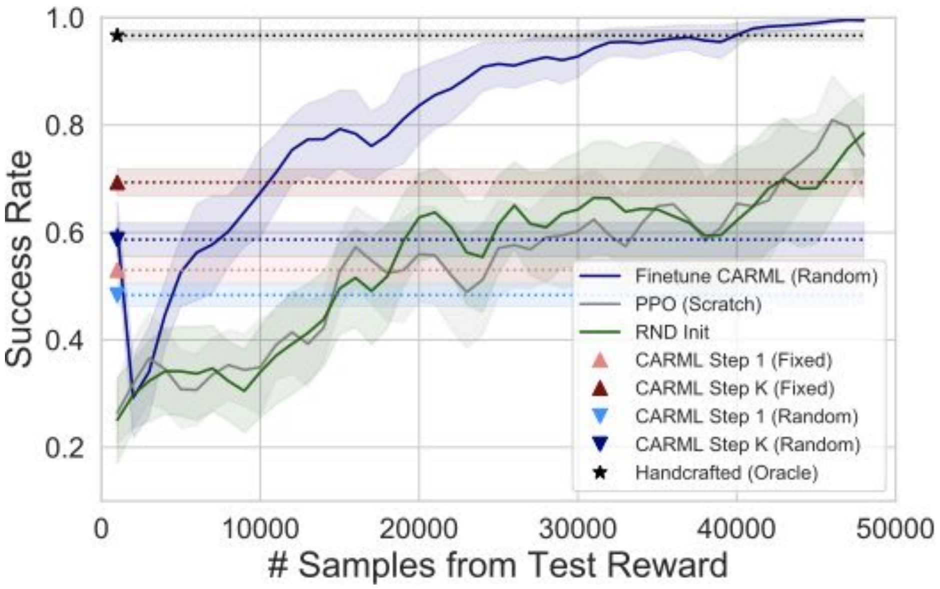


Step 5

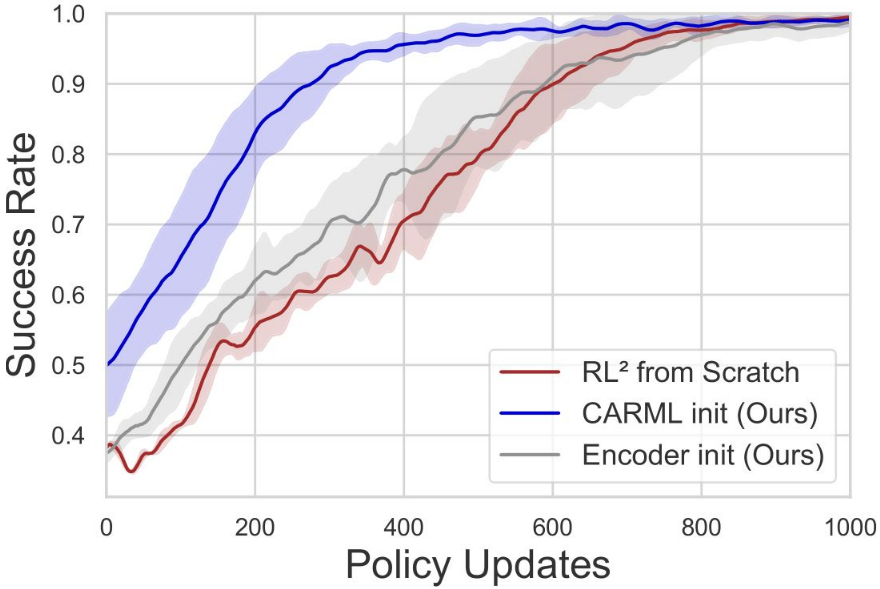


Learning Unsupervised Curricula

Faster fine-tuning



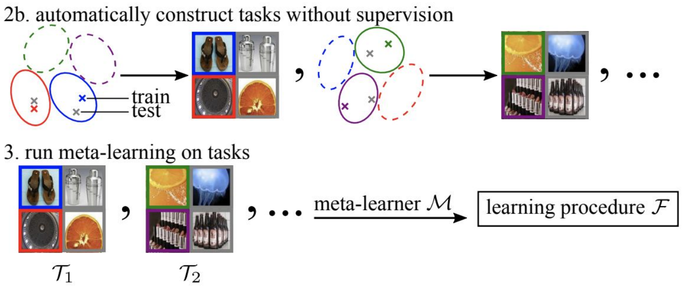
Faster meta-learning



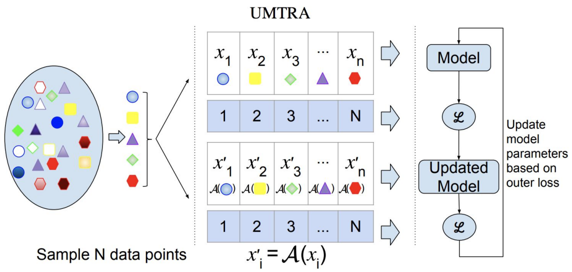
Perspectives on Unsupervised RL

1. Unsupervised RL can obtain semantically meaningful skills without rewards
2. Skills can help solve harder tasks, learn from demonstrations and improve fine-tuning
3. Combining with meta-learning can help prepare for the future!

CACTUS (Hsu et al)



UMTRA (Khodadadeh et al)



Open Problems and Conclusion

Learning without Rewards:

1. *Chicken-and-Egg Problem*: Skills learn to be diverse by using the discriminator's decision function, but the discriminator cannot learn to discriminate skills if they are not diverse.
2. *Application to the Real World*: How can we learn skills unsupervised and reset free?
3. *Semi-supervised RL:* How can we leverage small amounts of supervision with large amounts of unsupervised interaction?

Thanks!

Diversity Is All You Need: https://arxiv.org/abs/1802.06070
Unsupervised Meta-Learning for Reinforcement Learning https://arxiv.org/abs/1806.04640
Unsupervised Curricula for Visual Reinforcement Learning https://arxiv.org/abs/1912.04226