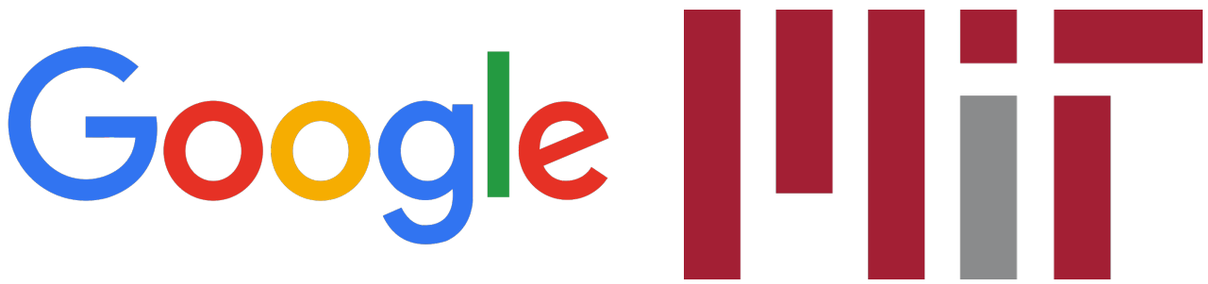
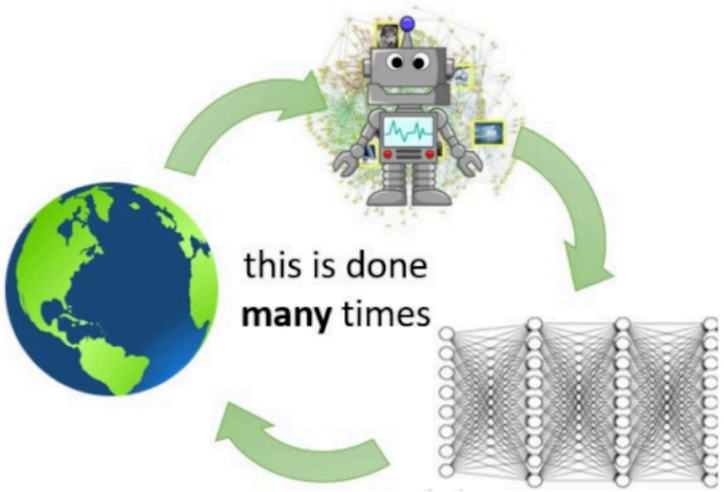
OPAL: **O**ffline **P**rimitive Discovery for **A**ccelerating Offline Reinforcement **L**earning

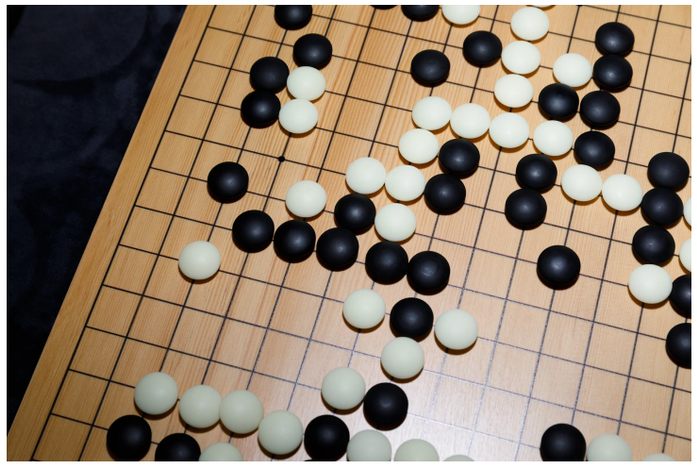
Anurag Ajay
(w/ Aviral Kumar, Pulkit Agrawal, Sergey Levine, Ofir Nachum)



Reinforcement Learning

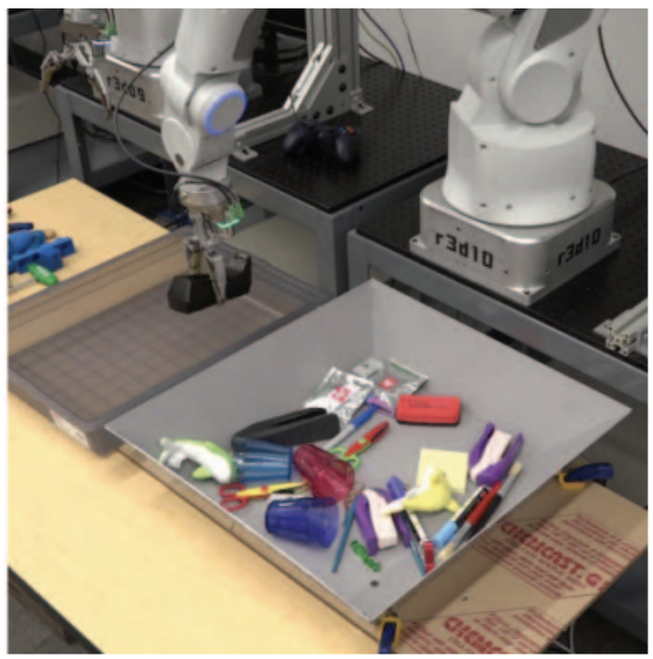


Not always practical

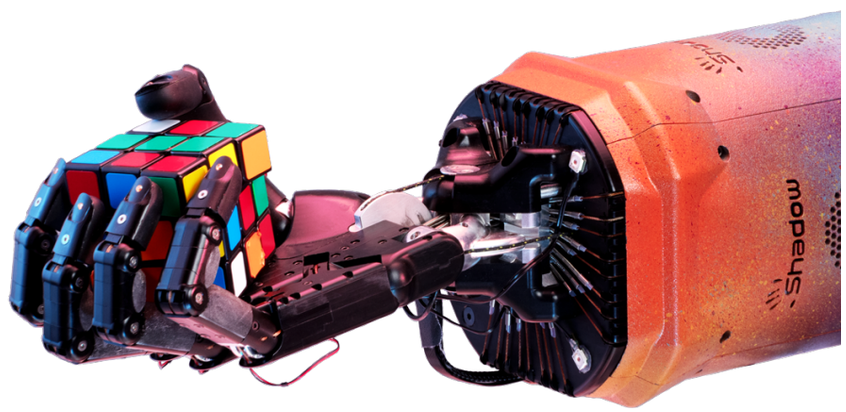


(Silver et al)

(Vinyals et al)

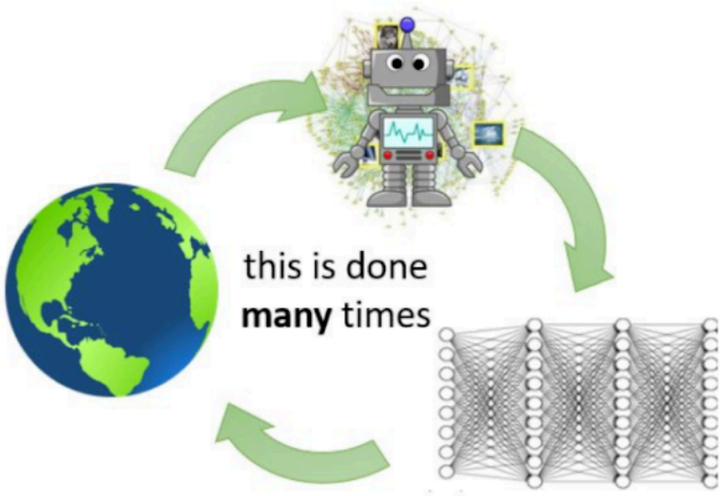


(Levine et al)

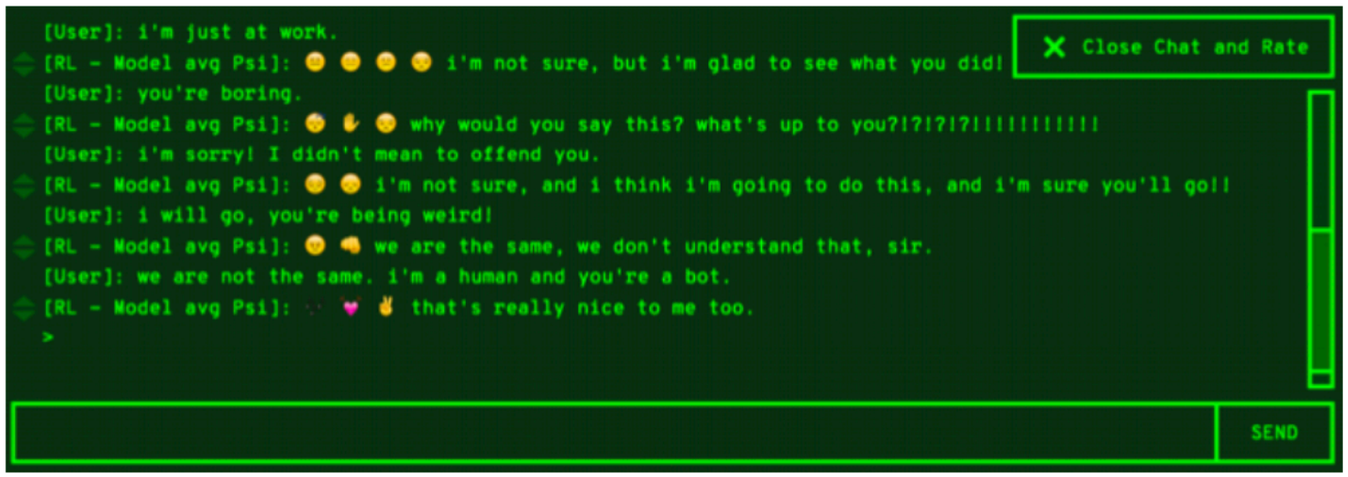


(OpenAI)

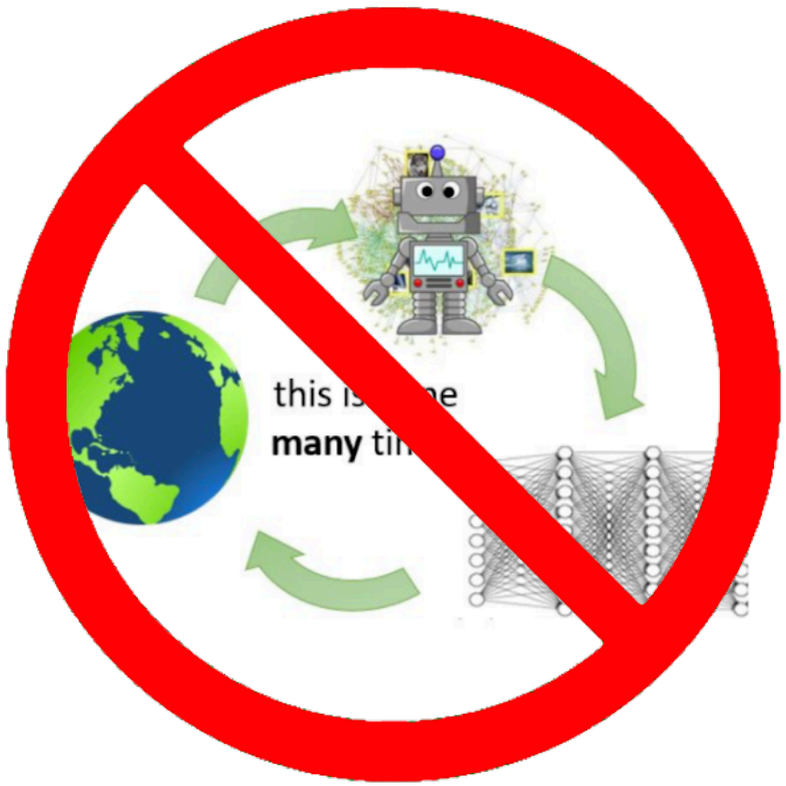
Reinforcement Learning

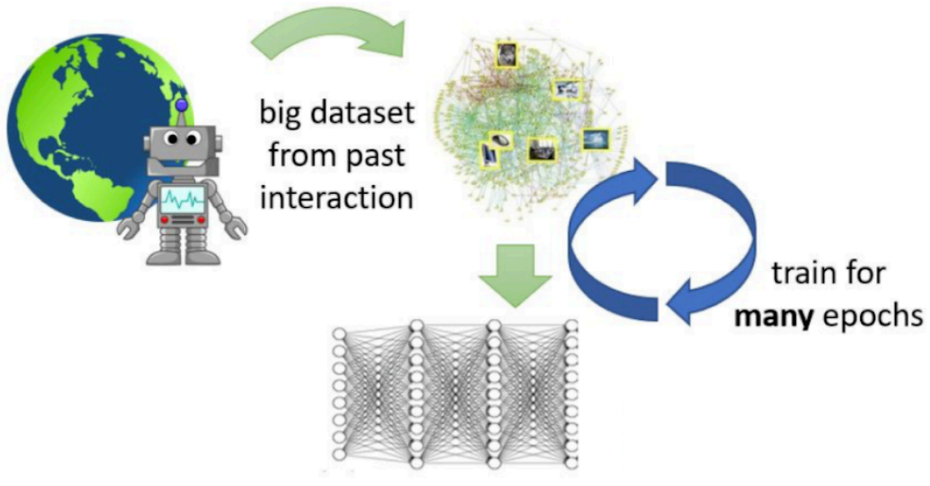


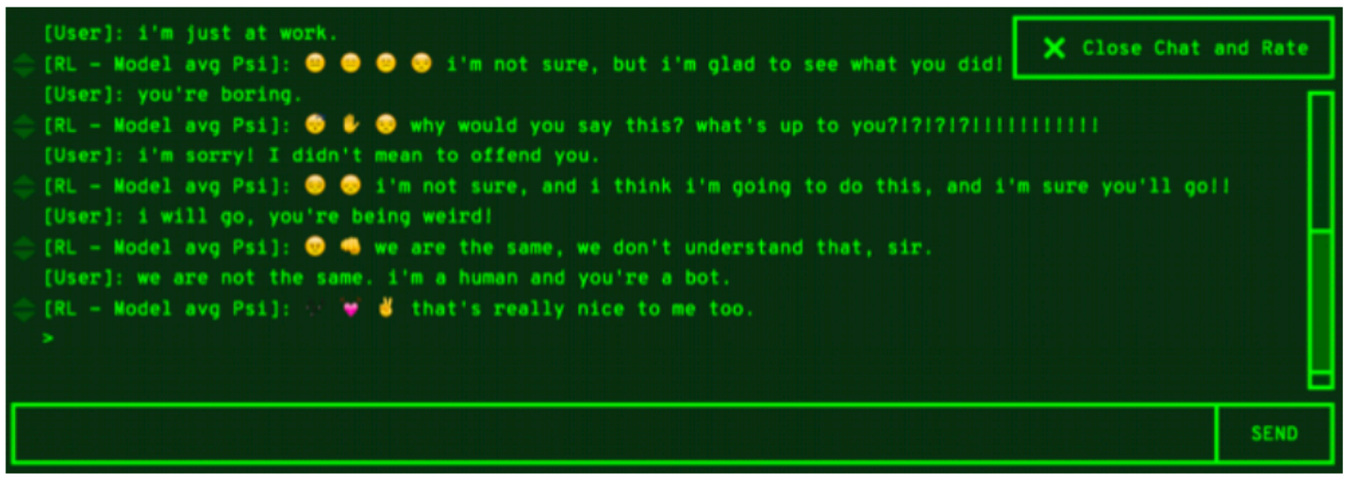
Not always practical



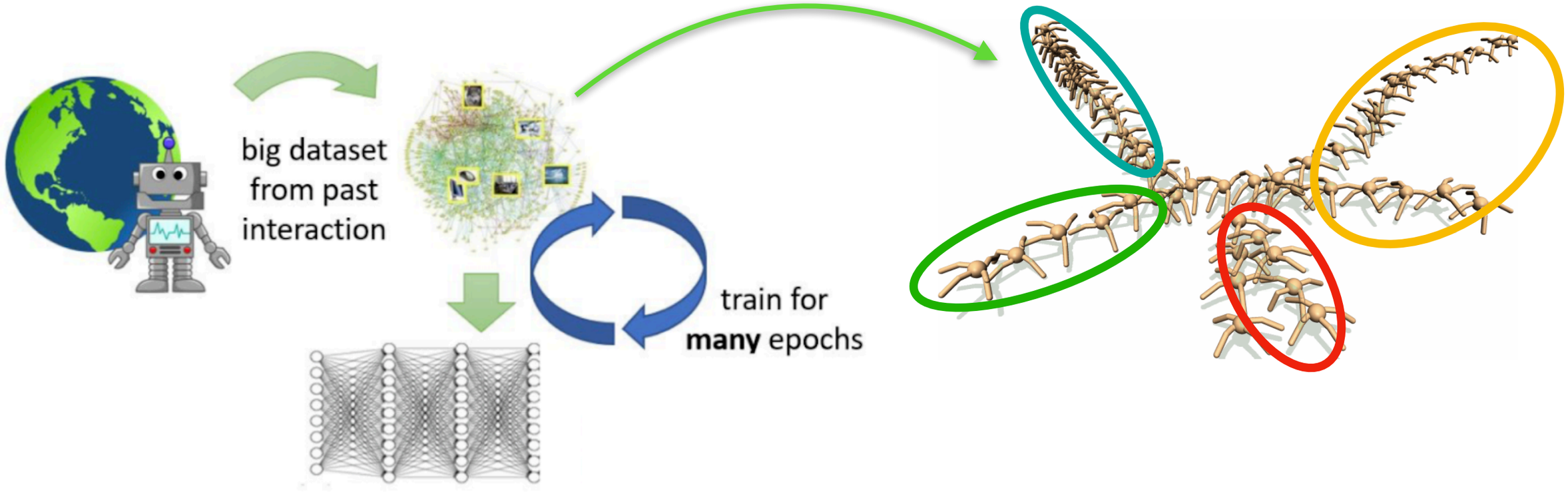
(Offline) Reinforcement Learning







Improving Offline RL with *primitives*



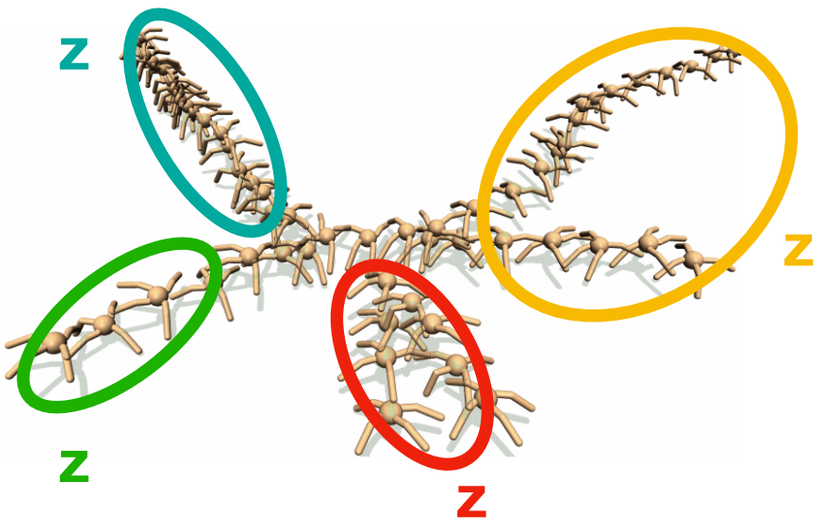
<u>Idea:</u> Improve Offline RL by leveraging *primitives* present in a *multi-modal* dataset

Intuition for OPAL

+1 for reaching goal

Given a task (i.e. a reward function)

Task Policy
$\pi_\psi(z|s)$

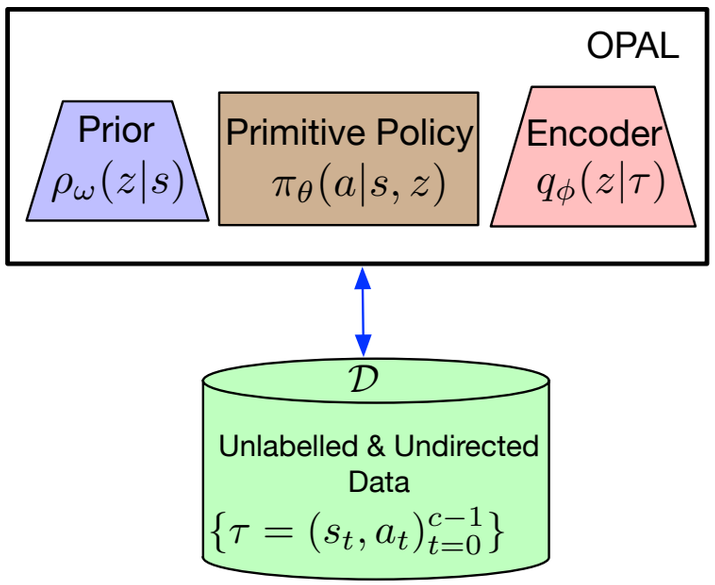


Primitive Policy
$\pi_\theta(a|s,z)$

Reduced task horizon & Compact action space → Better Credit Assignment

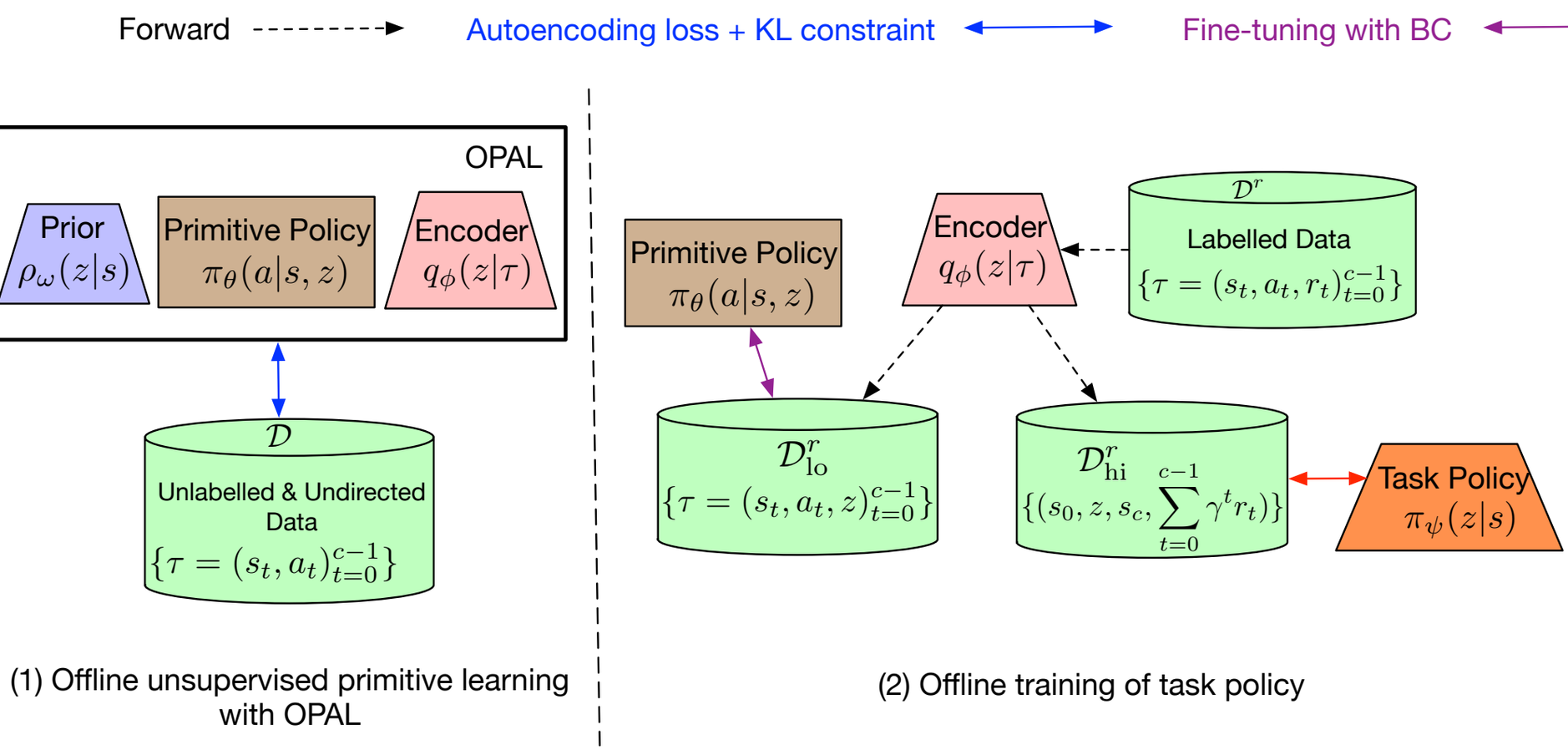
OPAL:**O**ffline **P**rimitive Discovery for **A**ccelerating RL

Forward ---------▸ Autoencoding loss + KL constraint ◂——▸ Fine-tuning with BC ◂——▸ Offline RL ◂——▸



(1) Offline unsupervised primitive learning with OPAL

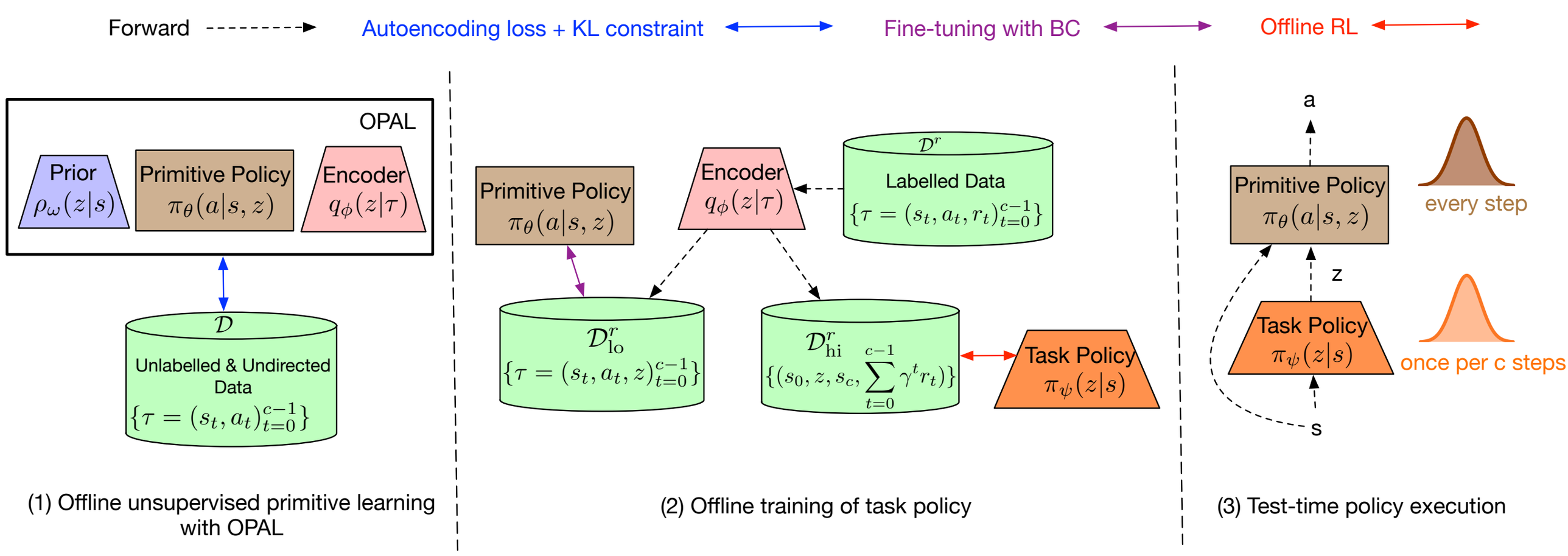
OPAL:**O**ffline **P**rimitive Discovery for **A**ccelerating RL



(1) Offline unsupervised primitive learning with OPAL

(2) Offline training of task policy

OPAL:**O**ffline **P**rimitive Discovery for **A**ccelerating RL



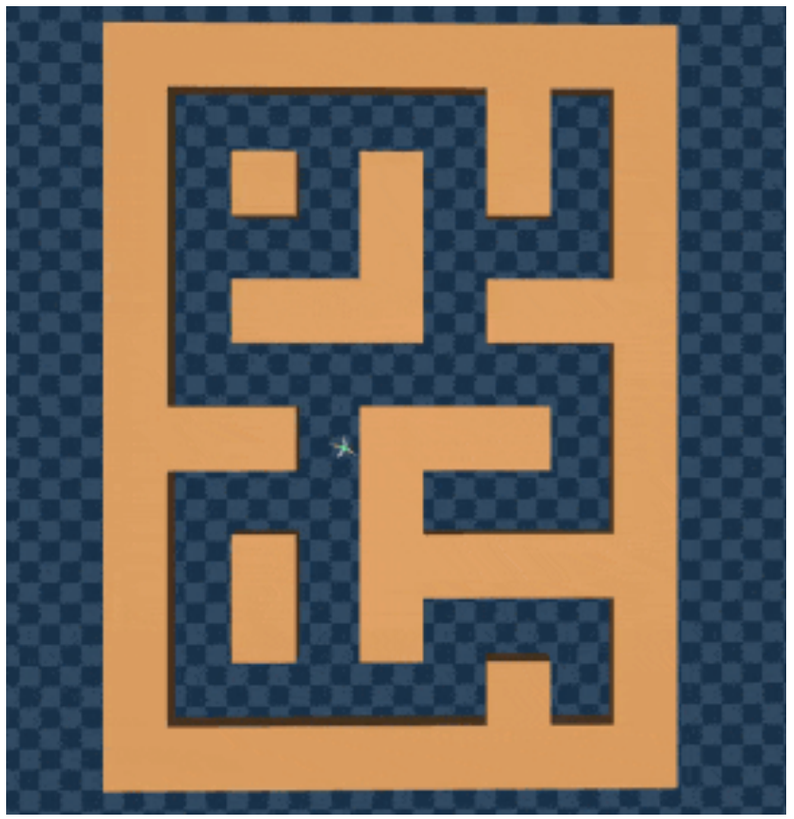
(1) Offline unsupervised primitive learning with OPAL

(2) Offline training of task policy

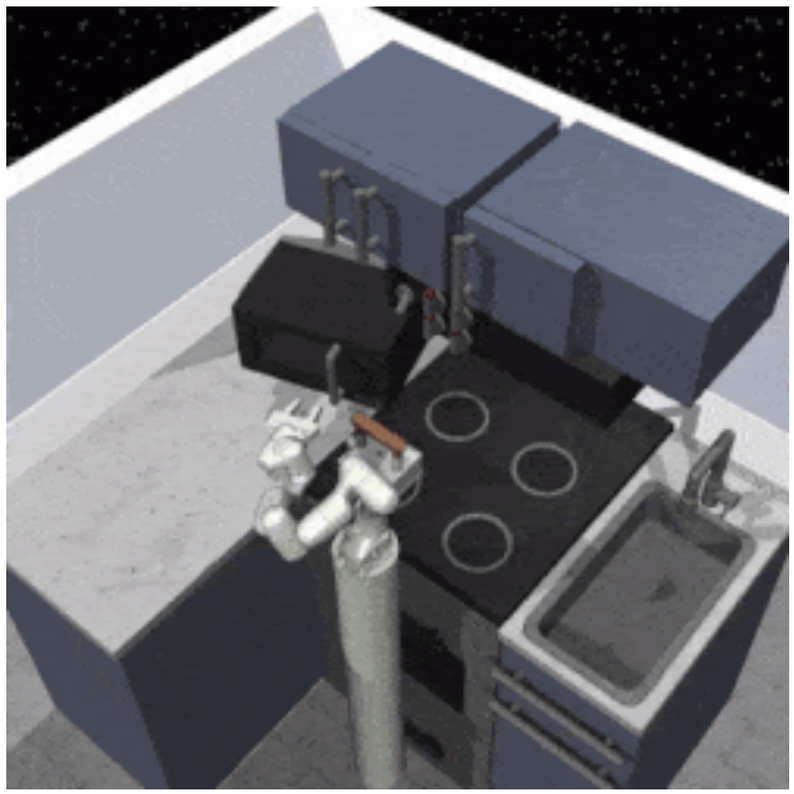
(3) Test-time policy execution

Environments

Antmaze medium



Antmaze large



kitchen

State (joint angles + xy pose): 29 dim

Action (joint torques): 8 dim

State (joint angles + xy pose): 60 dim

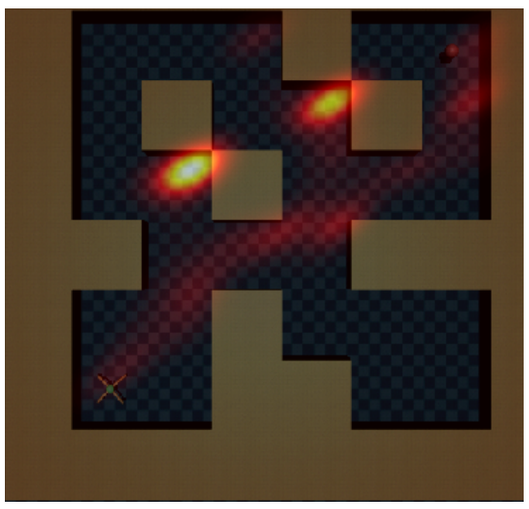
Action (joint torques): 9 dim

(from D4RL (fu et. al.))

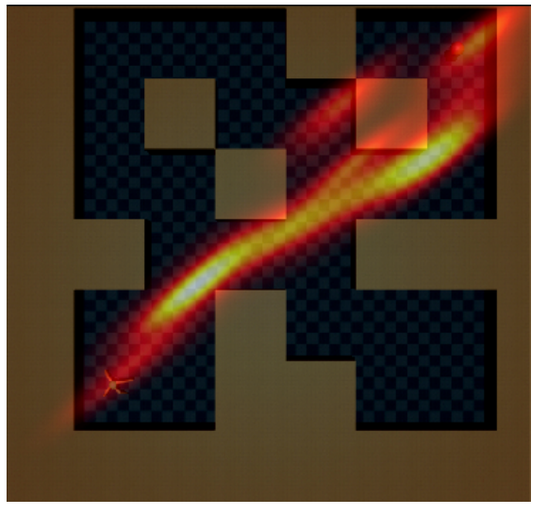
Results

Offline RL

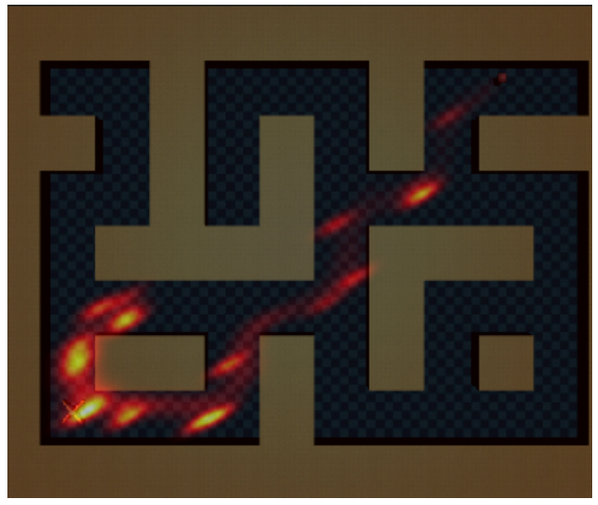
Environment	BC	BEAR (Kumar et al., 2019)	EMAQ (Ghasemipour et al., 2020)	CQL (Kumar et al., 2020)	**CQL+OPAL (ours)**
antmaze medium (diverse)	0.0	8.0	0.0	53.7 ± 6.1	$\mathbf{81.1 \pm 3.1}$
antmaze large (diverse)	0.0	0.0	0.0	14.9 ± 3.2	$\mathbf{70.3 \pm 2.9}$
kitchen mixed	47.5	47.2	$\mathbf{70.8 \pm 2.3}$	52.4 ± 2.5	$\mathbf{69.3 \pm 2.7}$
kitchen partial	33.8	13.1	74.6 ± 0.6	50.1 ± 1.0	$\mathbf{80.2 \pm 2.4}$



CQL (medium)



CQL + OPAL (medium)



CQL (large)

CQL + OPAL (large)

Results

Offline RL

Environment	BC	BEAR (Kumar et al., 2019)	EMAQ (Ghasemipour et al., 2020)	CQL (Kumar et al., 2020)	**CQL+OPAL (ours)**
antmaze medium (diverse)	0.0	8.0	0.0	53.7 ± 6.1	$\mathbf{81.1 \pm 3.1}$
antmaze large (diverse)	0.0	0.0	0.0	14.9 ± 3.2	$\mathbf{70.3 \pm 2.9}$
kitchen mixed	47.5	47.2	$\mathbf{70.8 \pm 2.3}$	52.4 ± 2.5	$\mathbf{69.3 \pm 2.7}$
kitchen partial	33.8	13.1	74.6 ± 0.6	50.1 ± 1.0	$\mathbf{80.2 \pm 2.4}$

Few-shot imitation learning

Online RL

Online multi-task transfer learning

Summary

OPAL discovers a continuous space of *primitives* from a *multi-modal* dataset which can be used to

- Improve offline RL
- Improve few-shot imitation learning
- Accelerate online RL and multi-task learning

Visualization & code:

https://sites.google.com/view/opal-iclr

Check out paper/poster for

- Detailed description of OPAL
- Theoretical guarantees
- Results on few-shot imitation learning, online RL and multi-task learning