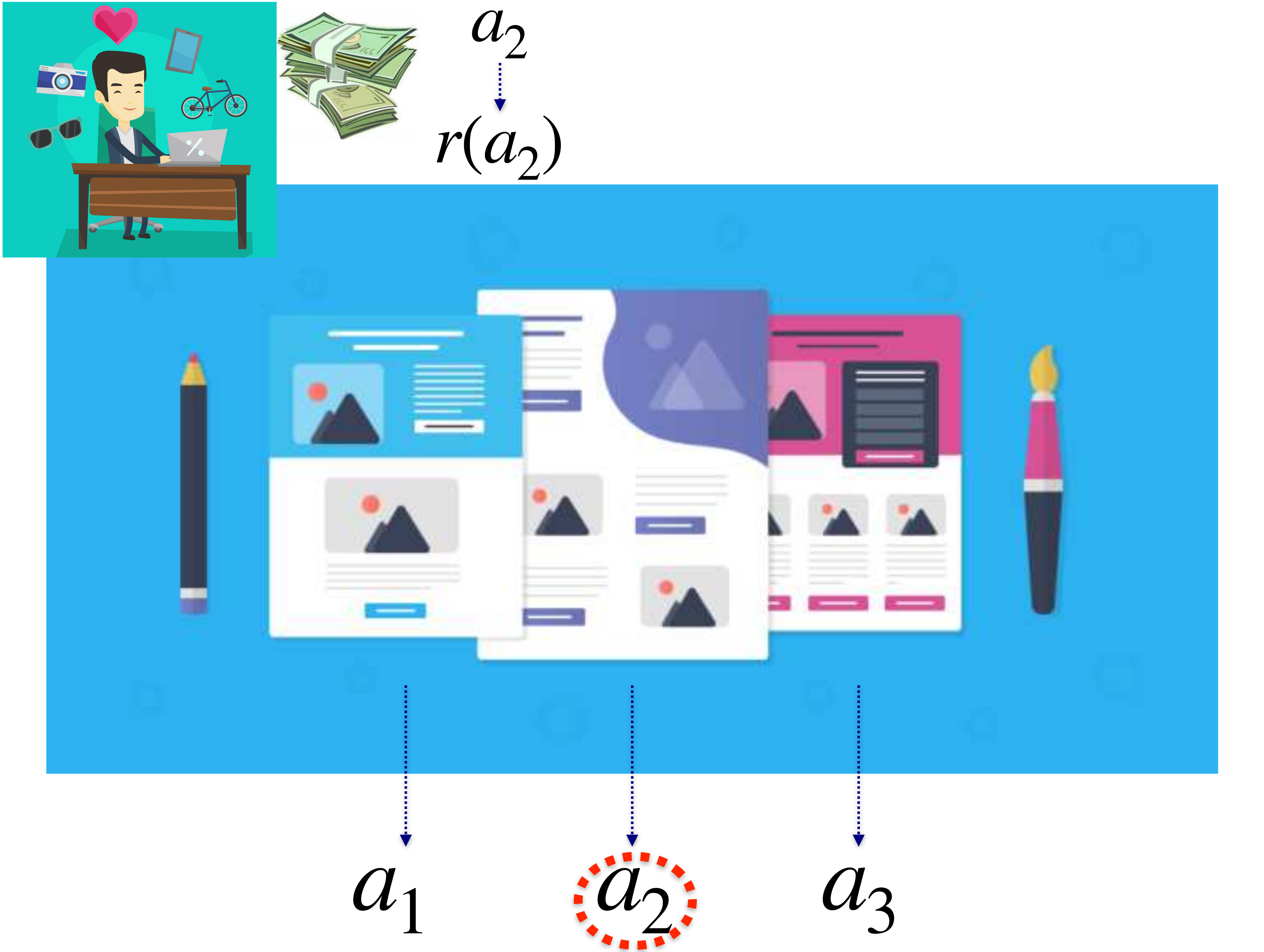
Challenges in Real World RL

Improbable AI Lab

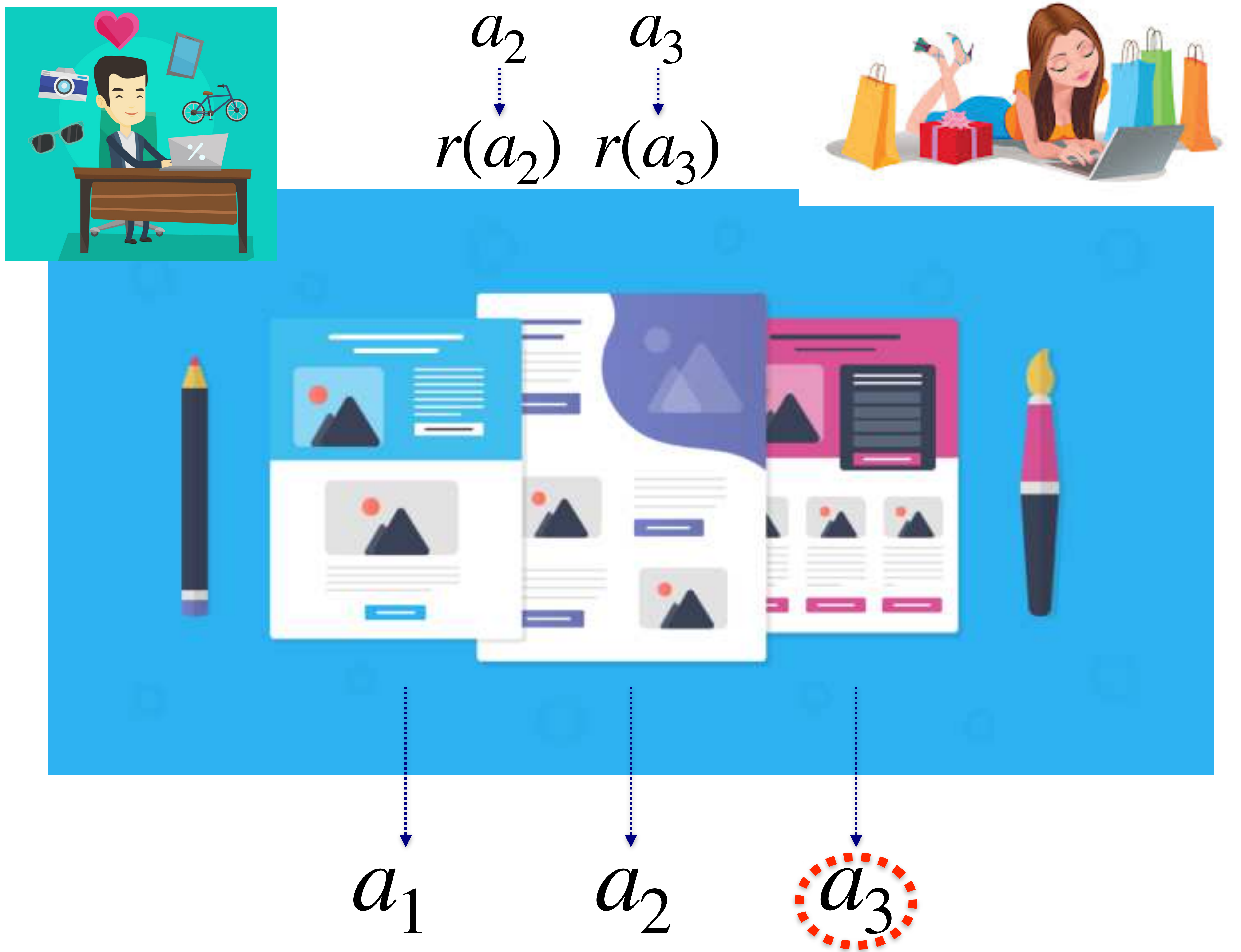
Pulkit Agrawal

When to use RL?

a_1 a_2 a_3



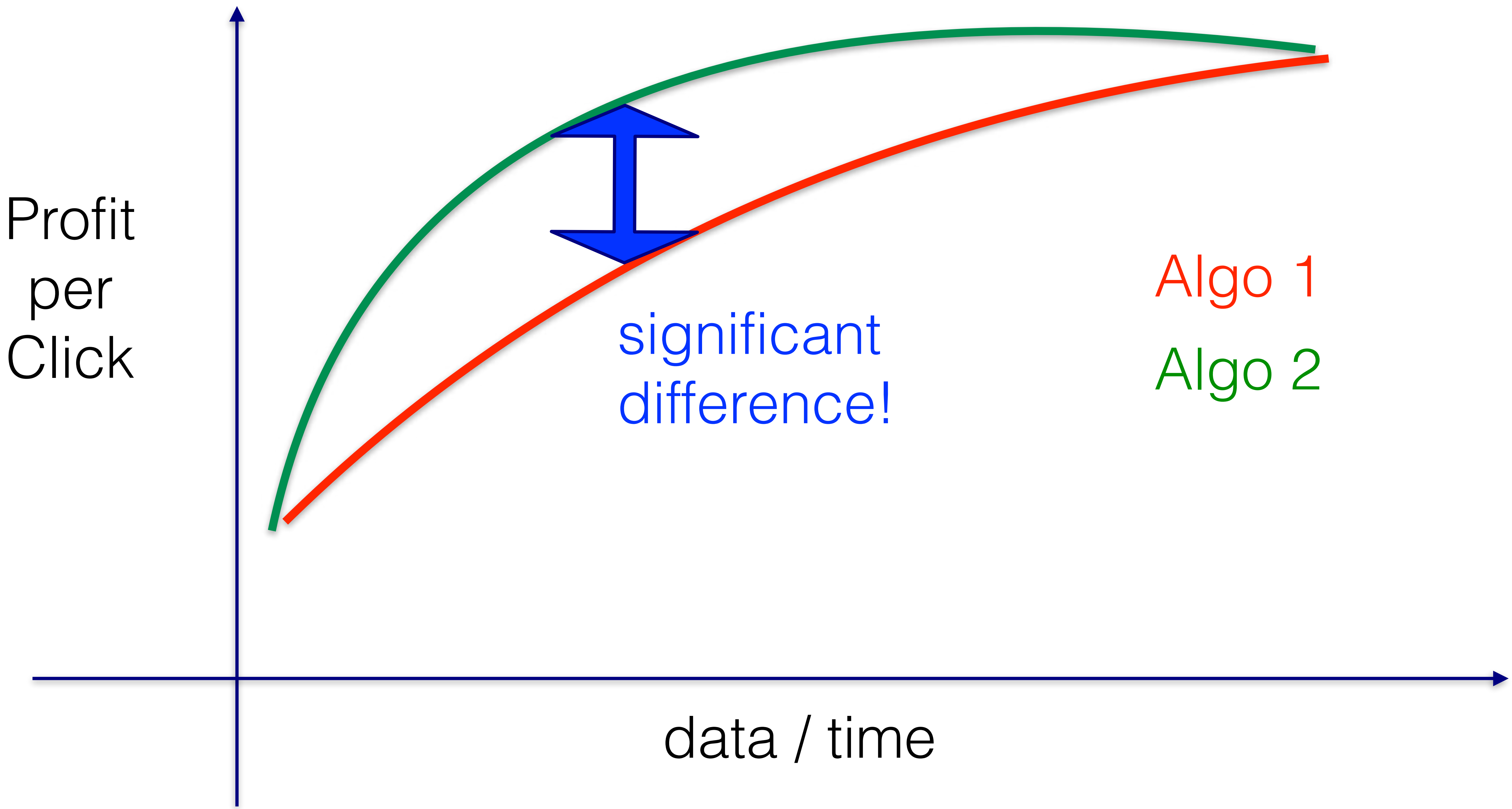
a_2
$r(a_2)$
a_1
a_2
a_3


a_2
a_3
$r(a_2)$
$r(a_3)$
a_1
a_2
a_3

When to use RL?

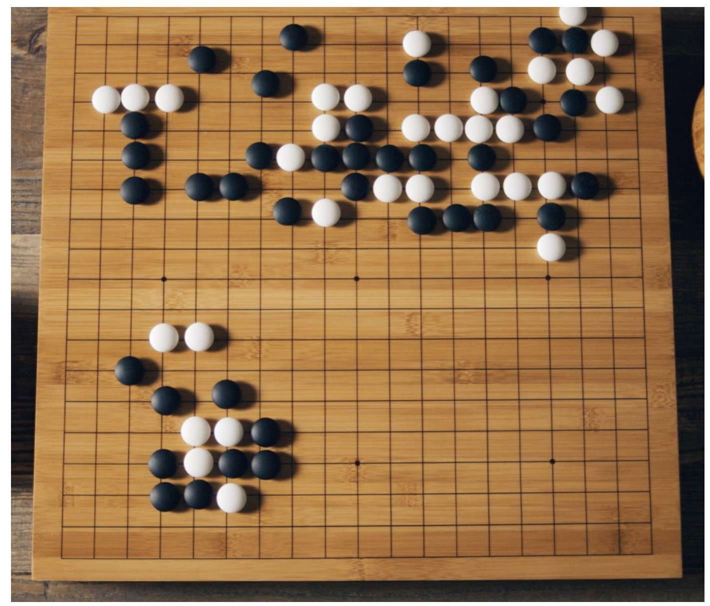
Online Decisions!

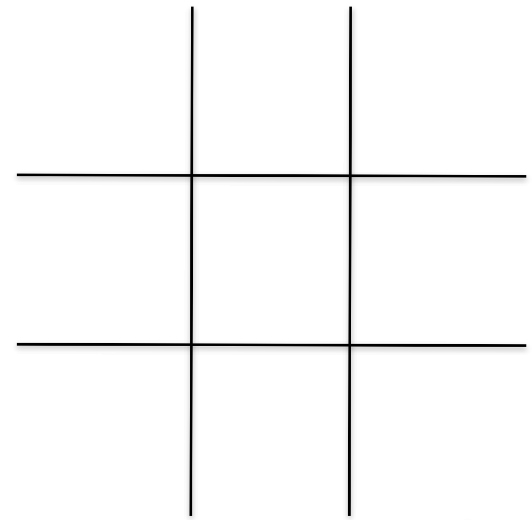
(i.e., can’t wait to collect data first)

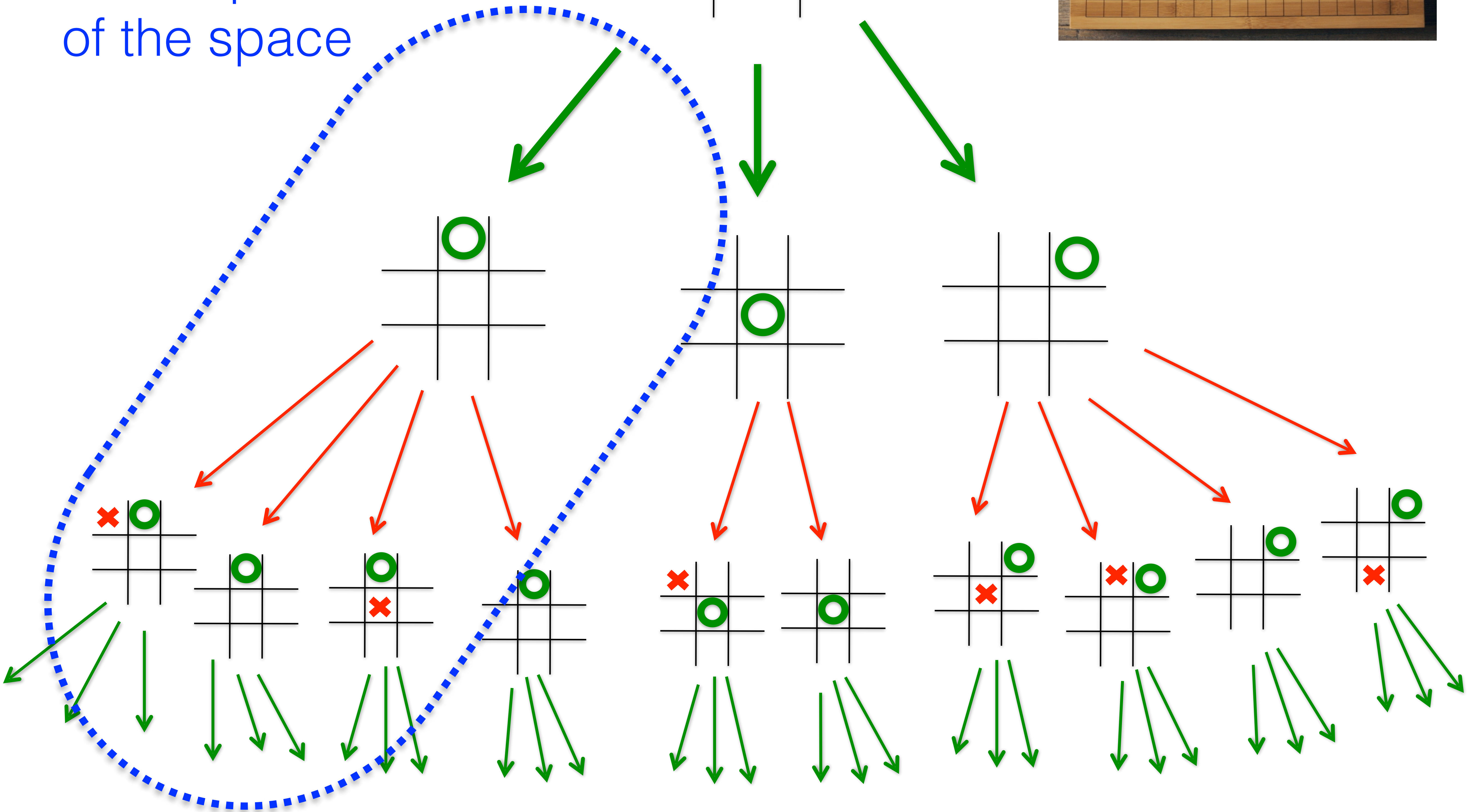


When to use RL?

can only explore
small part
of the space

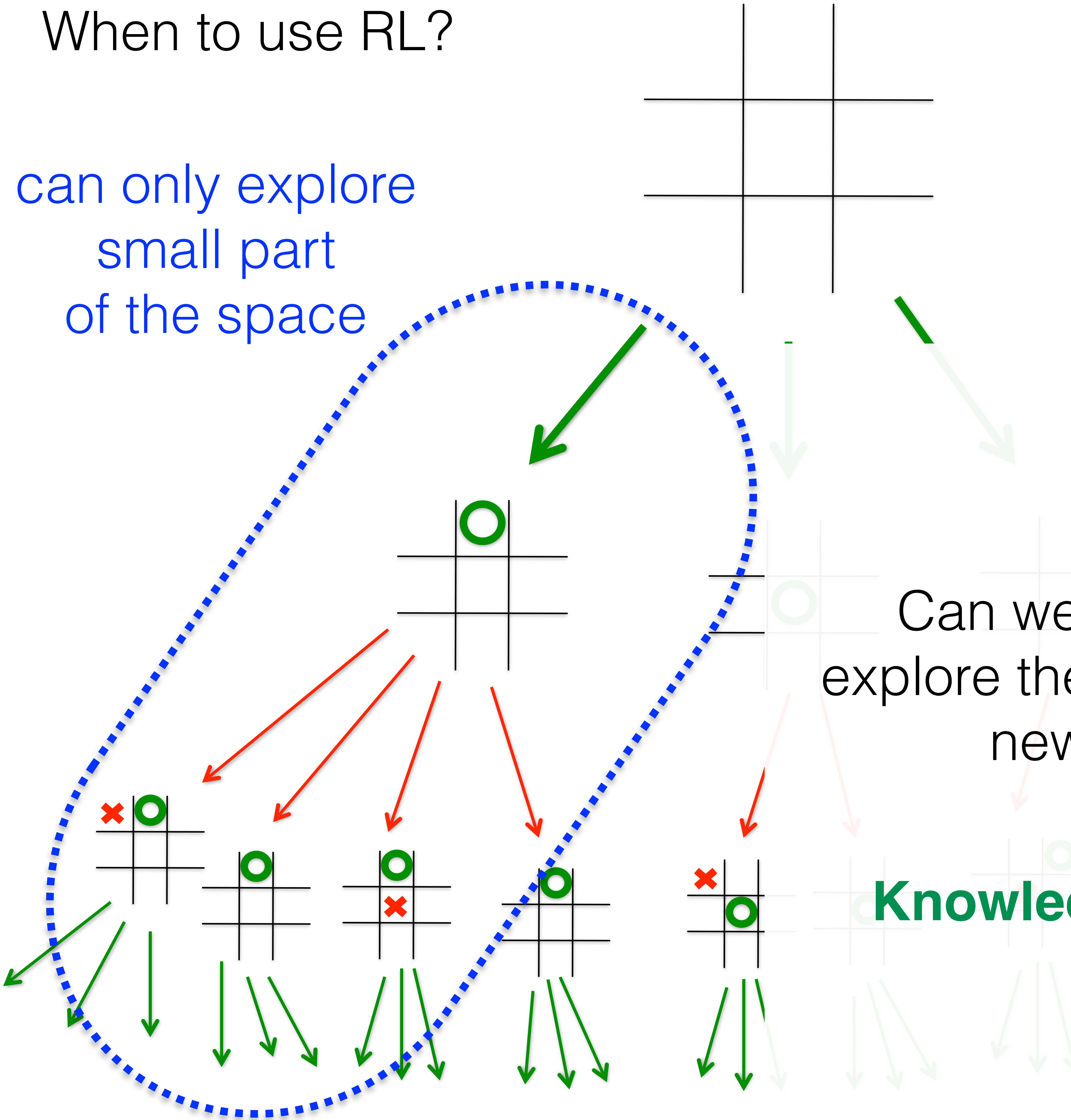


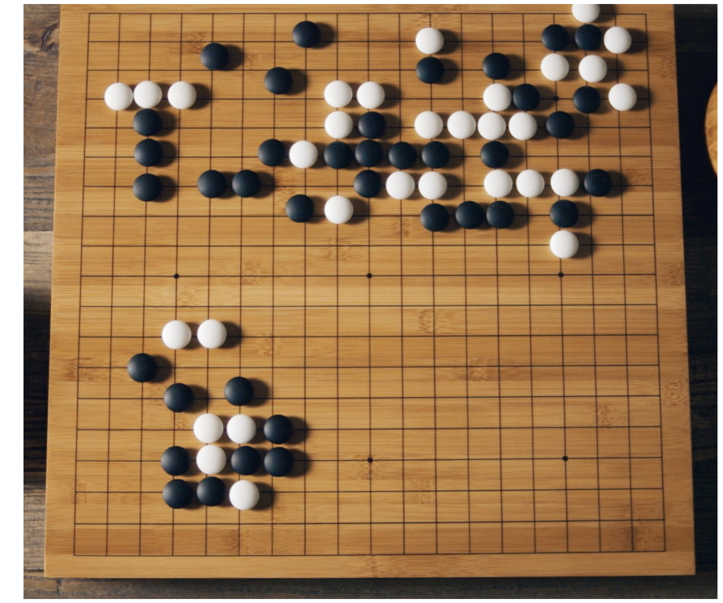




When to use RL?

can only explore
small part
of the space

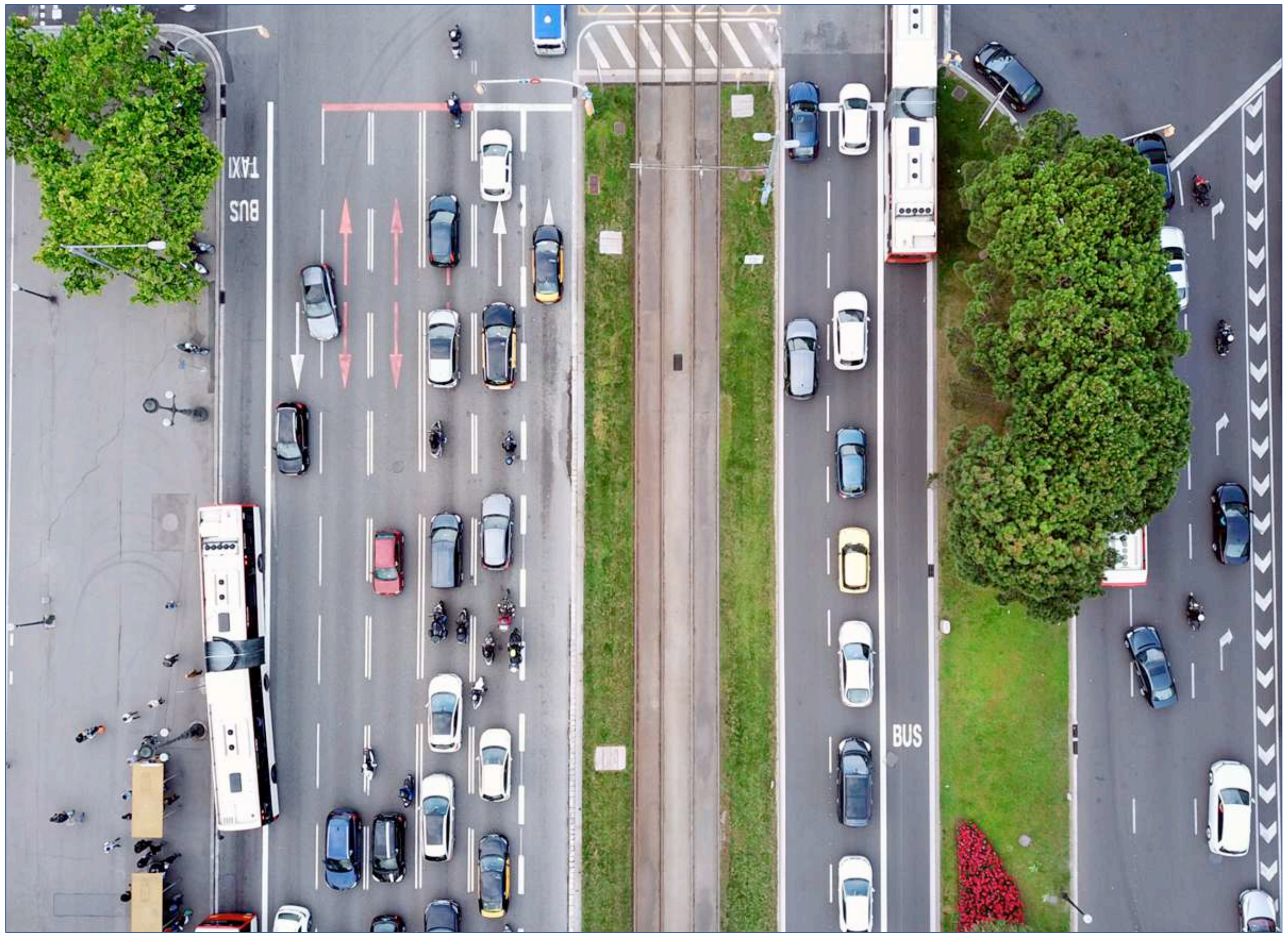


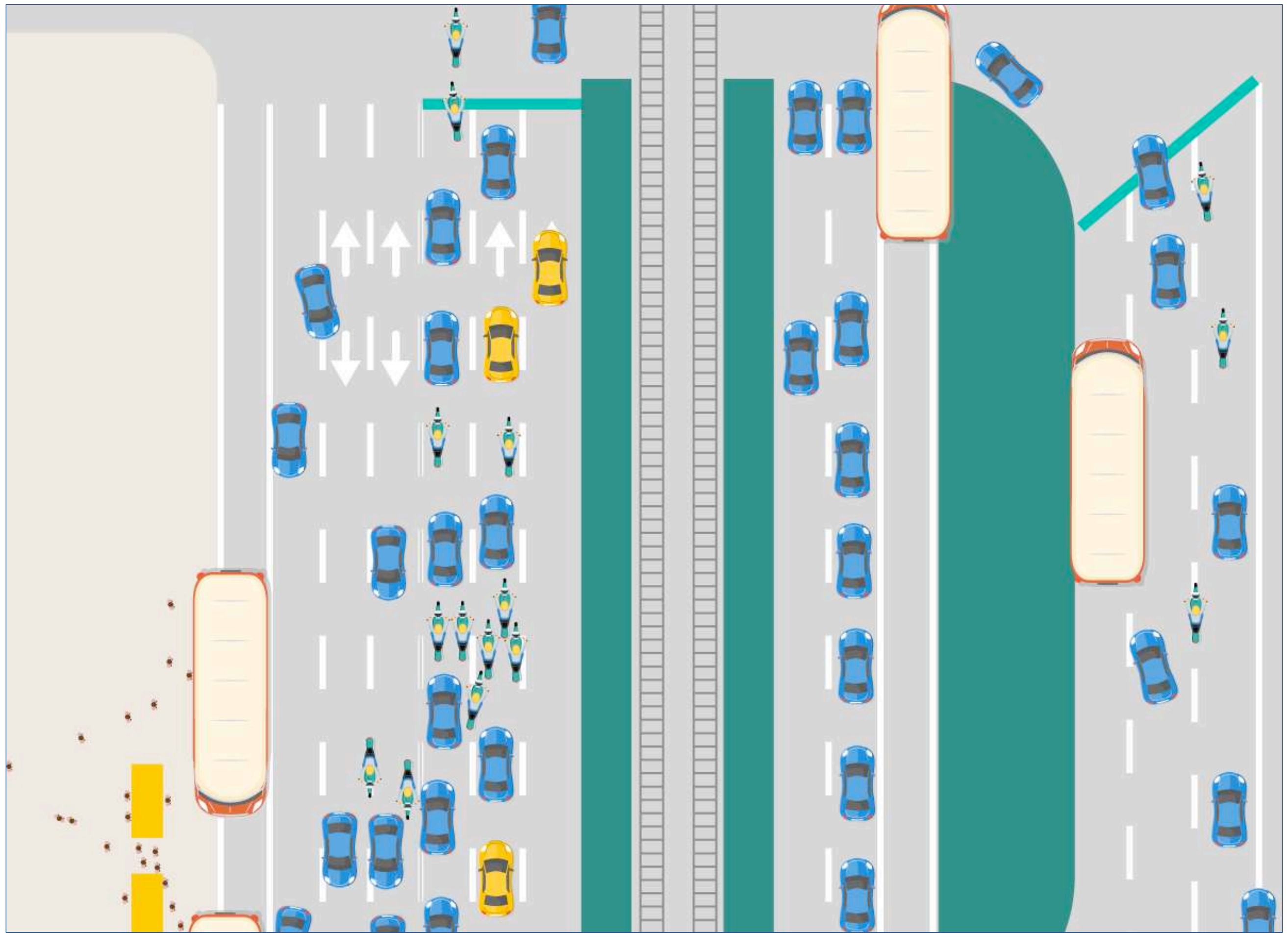


Can we “intelligently”
explore the space to create
new insights?

Knowledge Synthesis!

Example of Knowledge Synthesis: Urban Planning



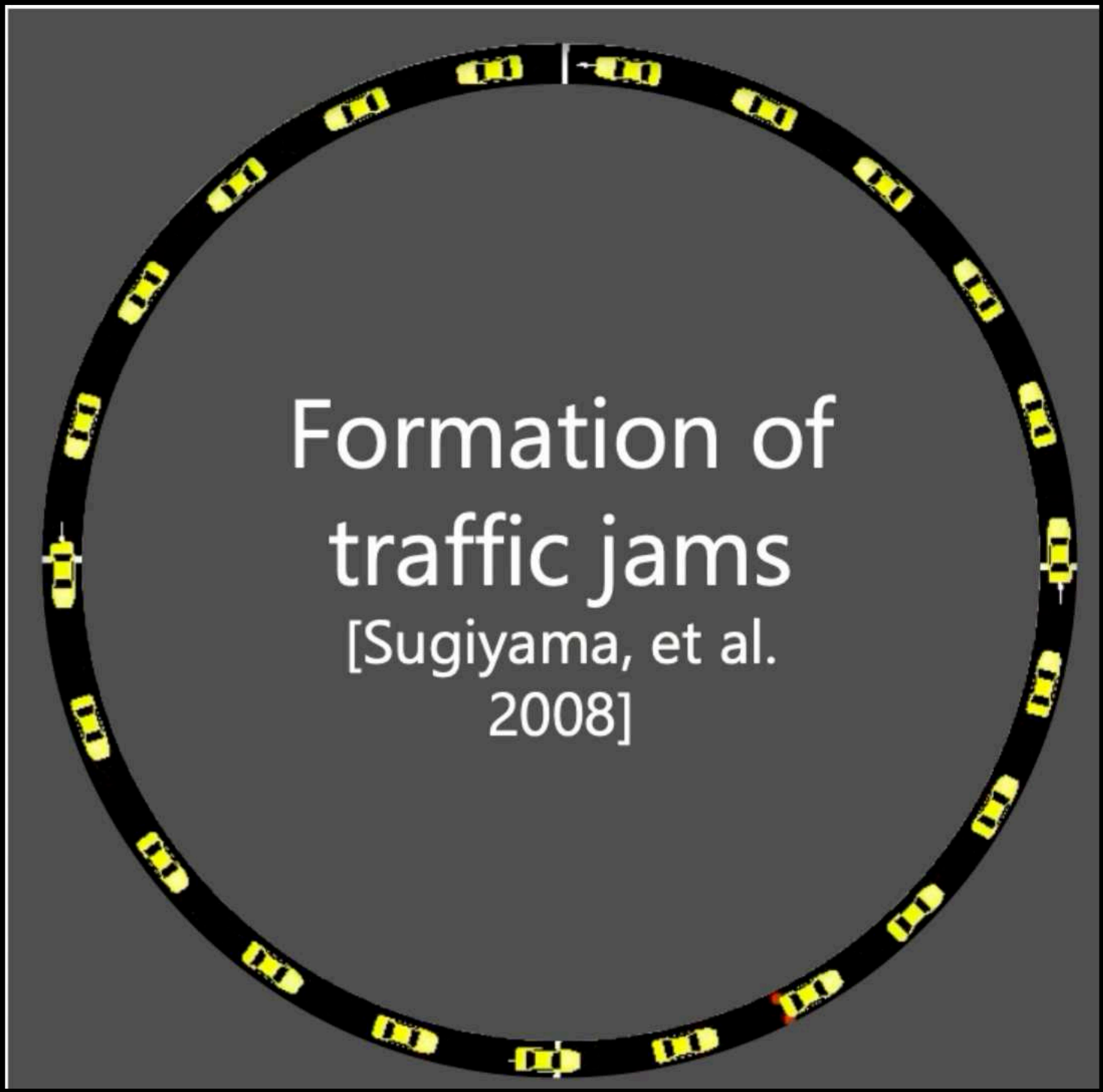


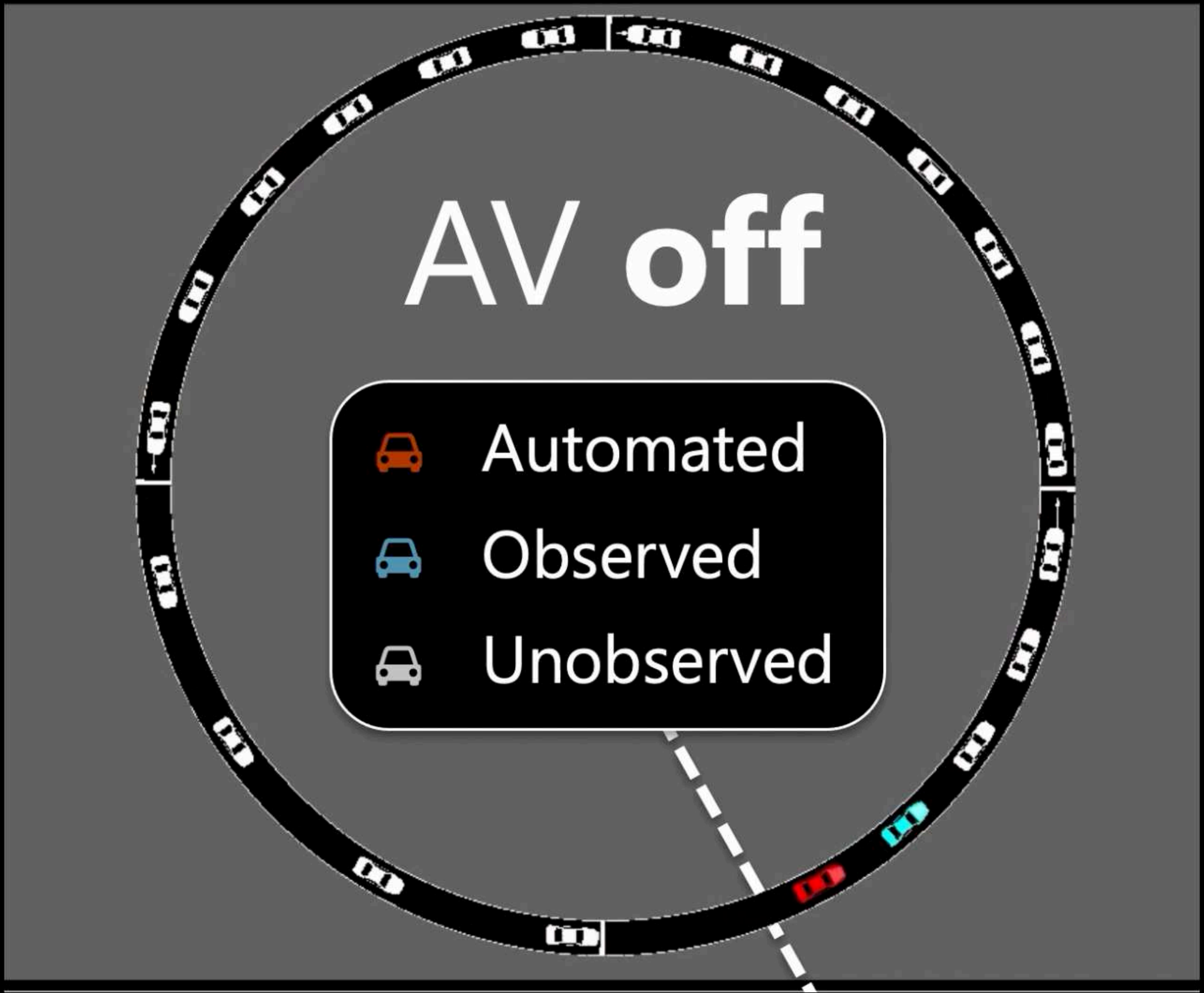
We have simulators

But how to use them for the desired purpose?

Traffic Jam Problem

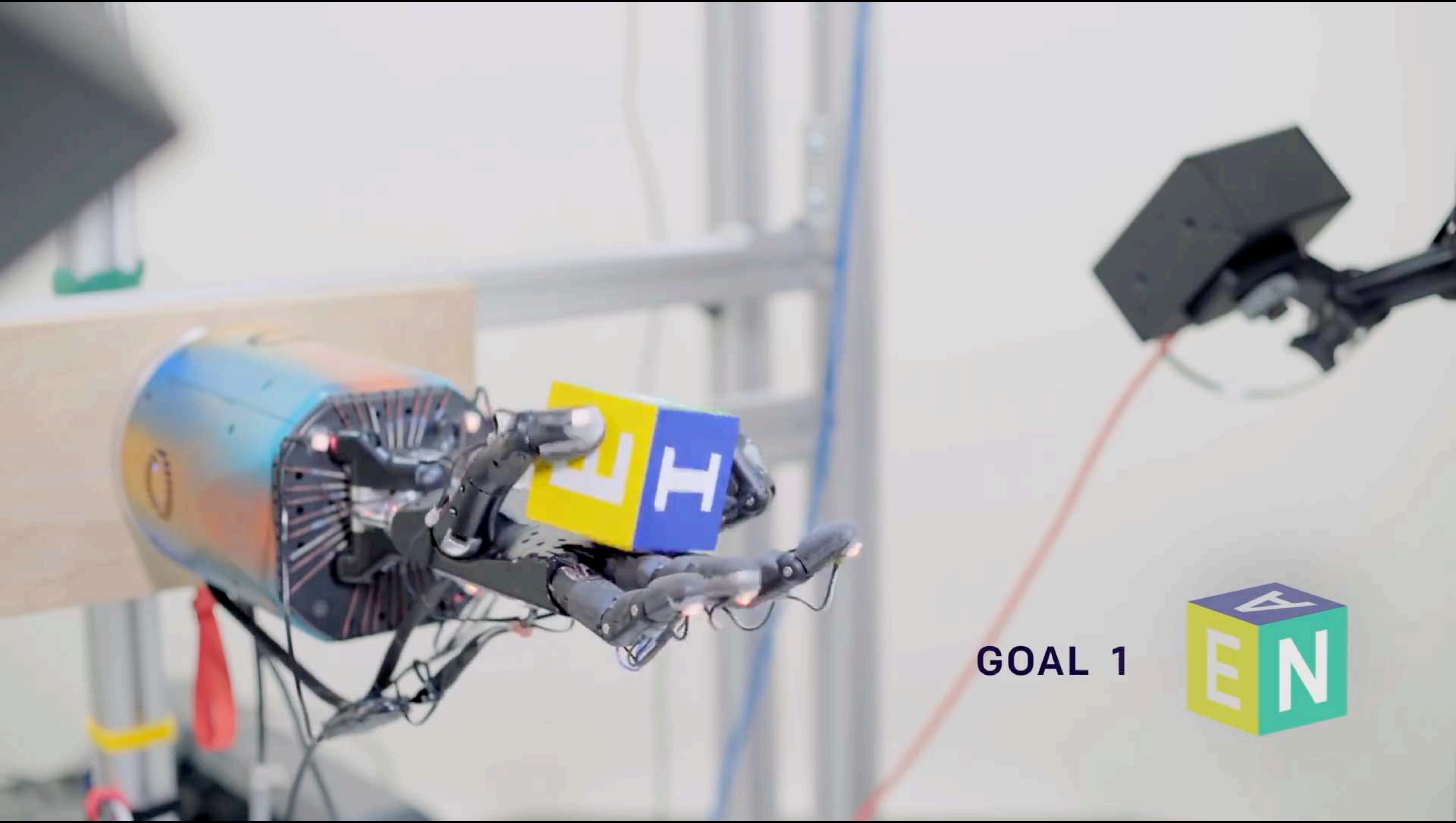
Simulated Illustration





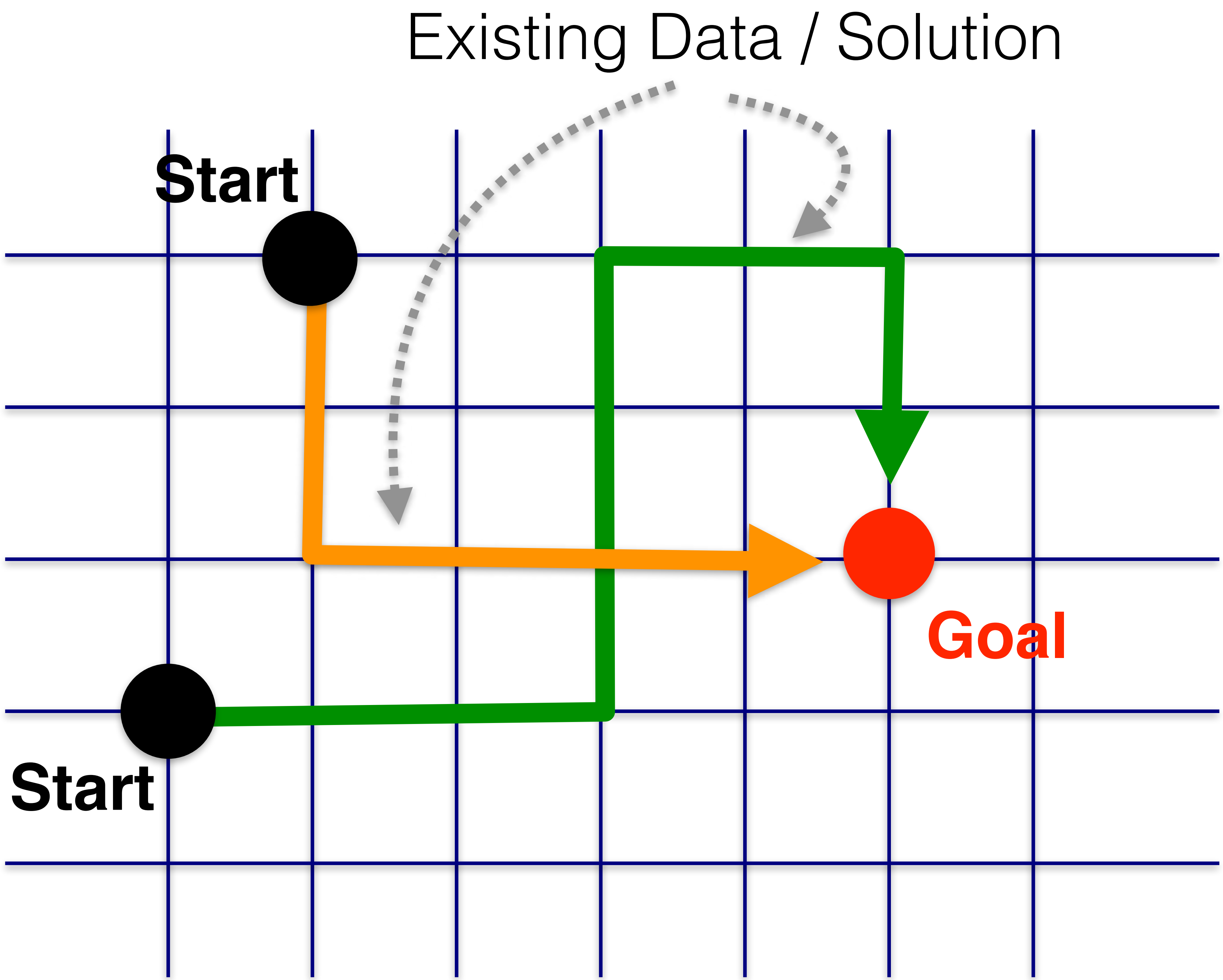
Wu et al. 2018

Learning Dexterity

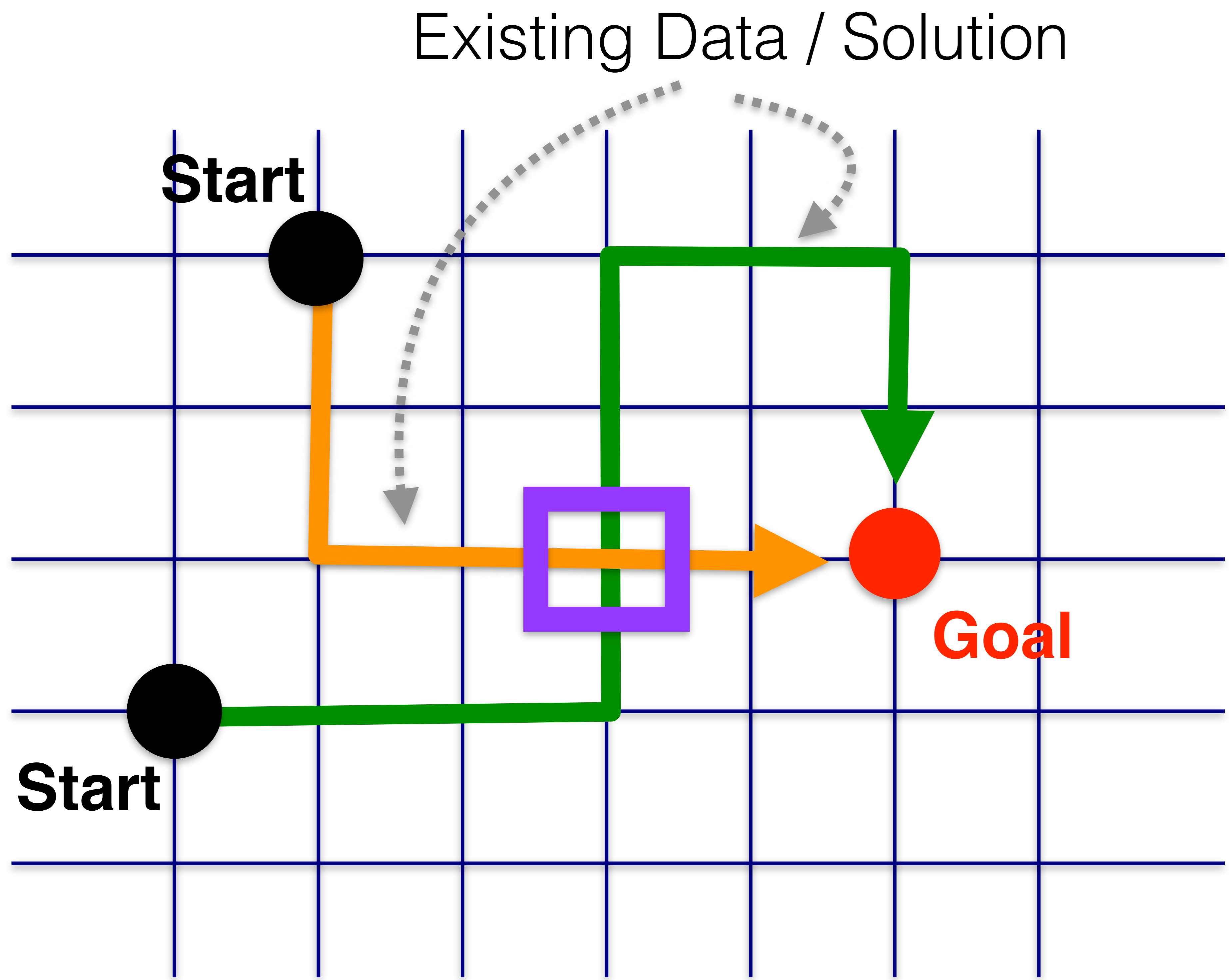


Known Simulator, 24 degrees of freedom!

When to use RL?

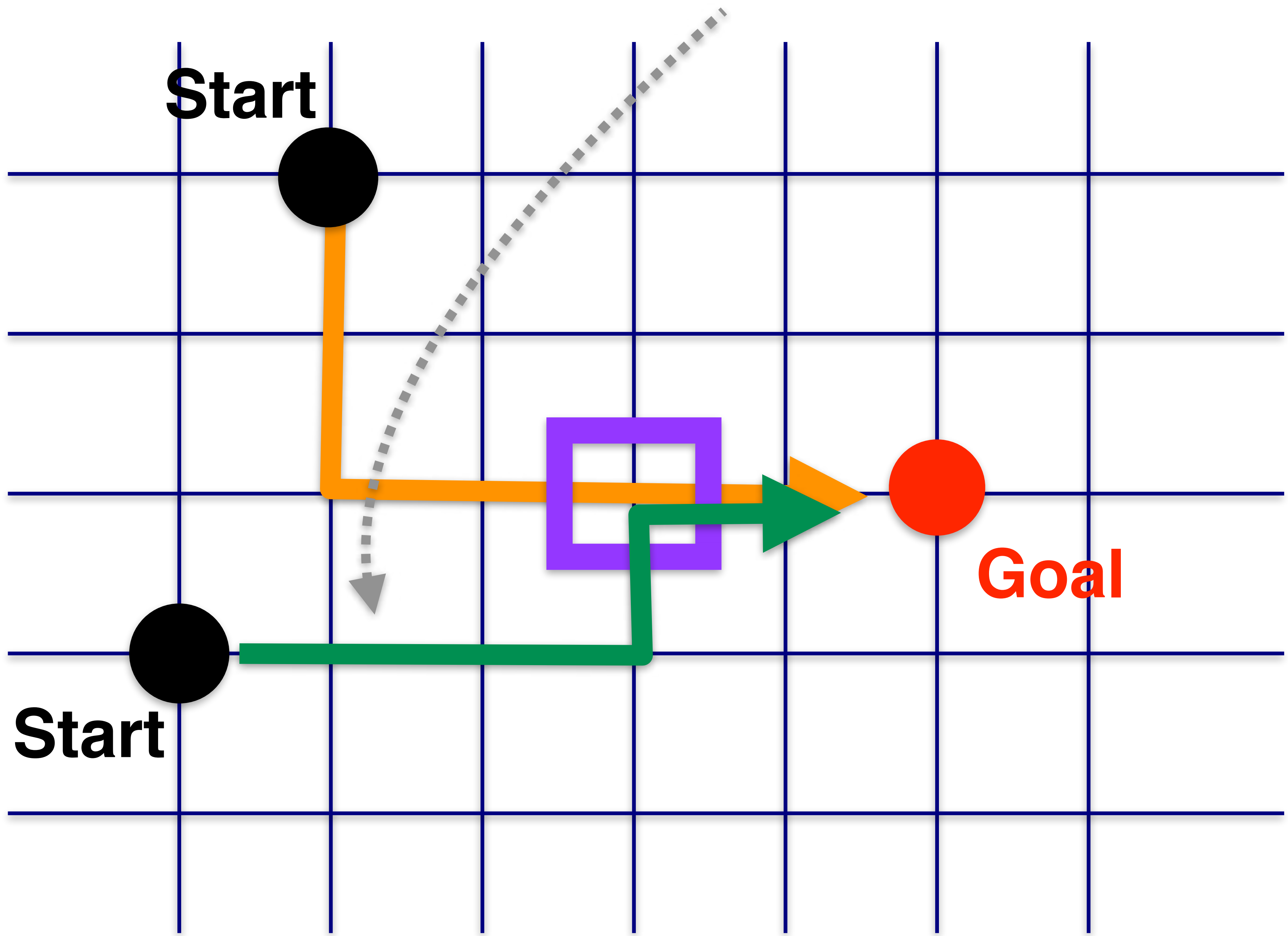


When to use RL?



When to use RL?

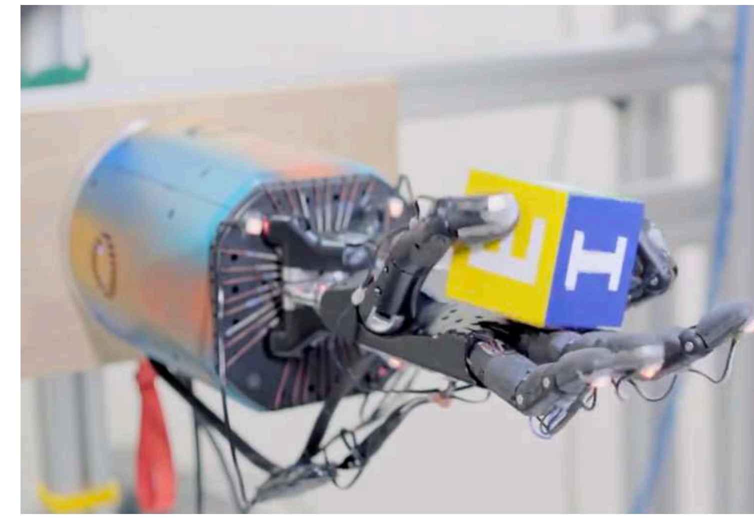
Improve the solution!



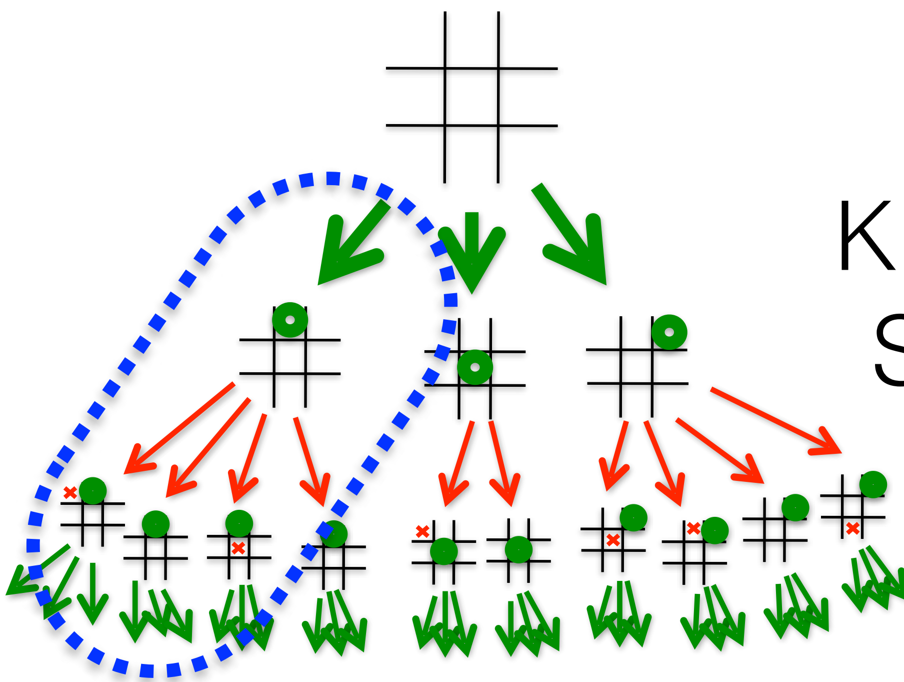
RL can be thought of as finding the “shortest path”

When to use RL?

online decisions

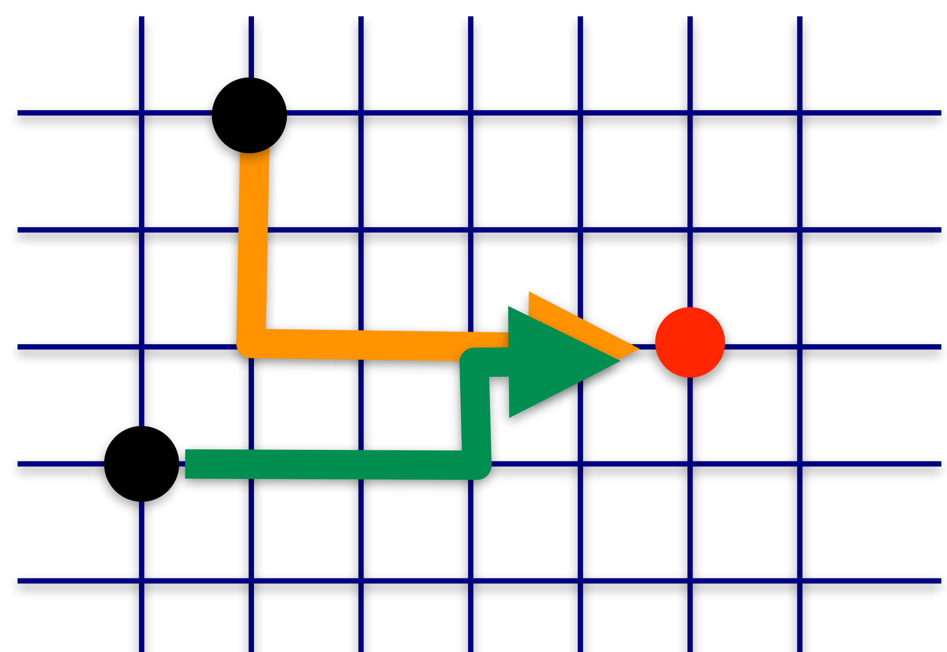


Control non-linear systems



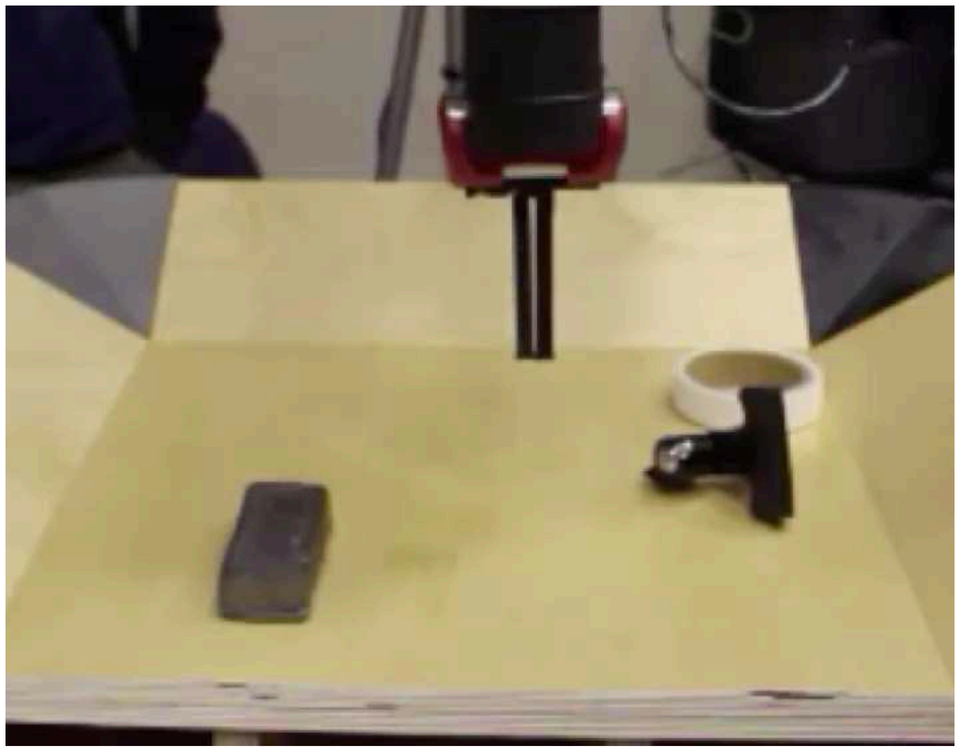
Knowledge Synthesis

Improve existing solutions

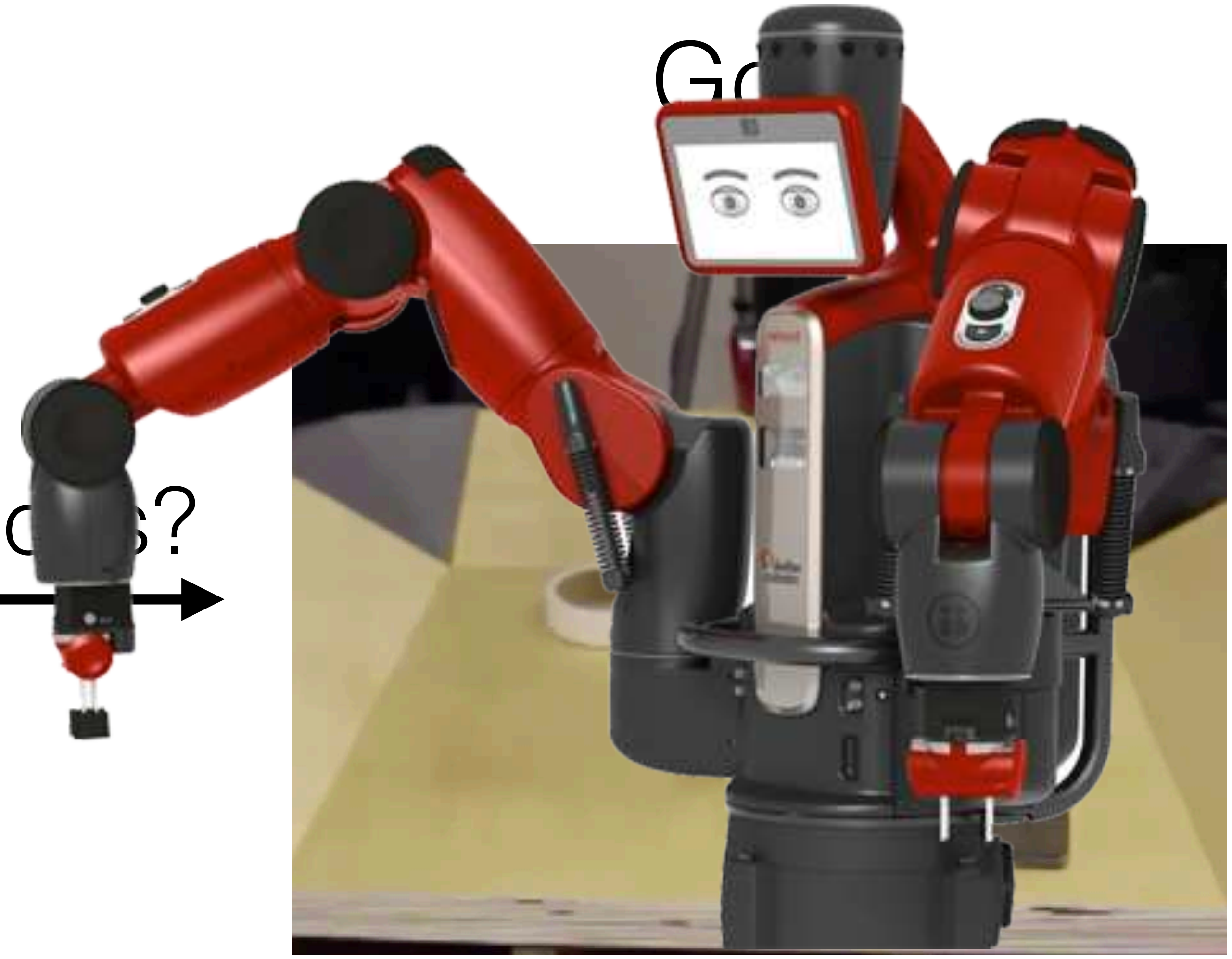


Current Observation

Go



Acti s?



Actions?

One Approach: Reinforcement Learning

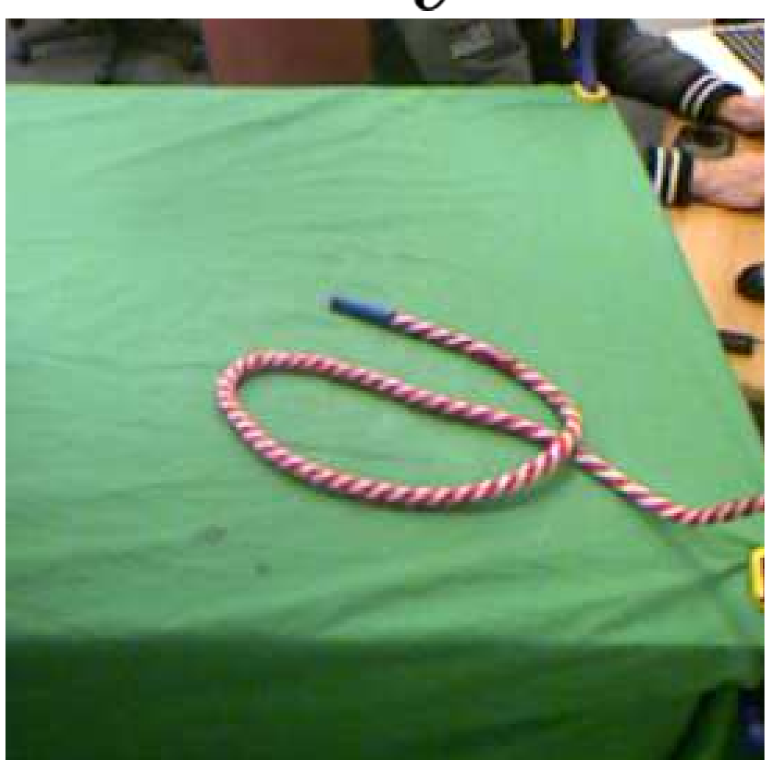
$x_t \rightarrow$ **Policy** $\pi(x_t;\theta) \rightarrow a_t \dashrightarrow x_{t+1}$

initially random

$$\max_\theta \mathbb{E}\left(\sum_{t=1}^{T} r_t\right)$$

r_{t+1}

x_t



Actions? $\longrightarrow$

x_G

One Approach: Reinforcement Learning

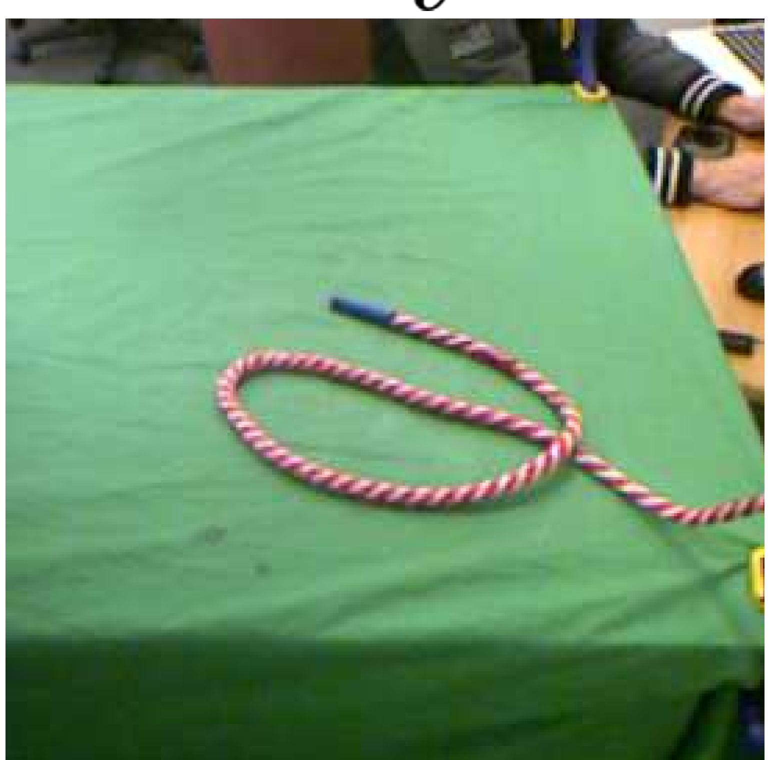
$x_t \rightarrow$ Policy $\pi(x_t;\theta)$ $\rightarrow a_t \dashrightarrow x_{t+1}$

$\max_\theta \mathbb{E}\left(\sum_{t=1}^{T} r_t\right)$ r_{t+1}

initially random

How likely?

x_t



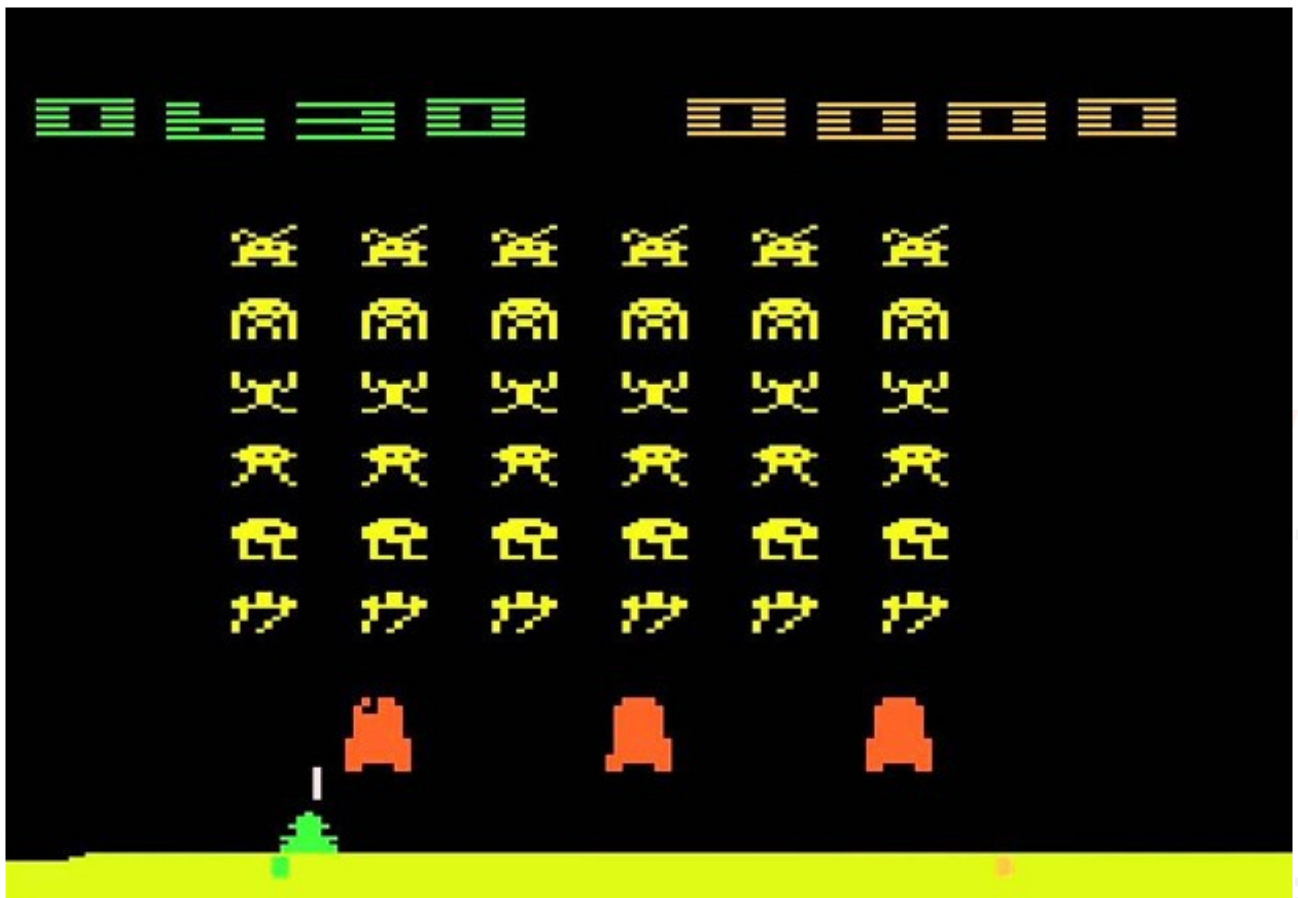
Actions?

x_G

One Approach: Reinforcement Learning

ATARI Games

AlphaGo



~10-50 million interactions!

21 million games!

Simulation: Ginormous number of interactions!

One Approach: Reinforcement Learning

$x_t \rightarrow$ Policy $\pi(x_t;\theta) \rightarrow a_t \rightarrow x_{t+1}$

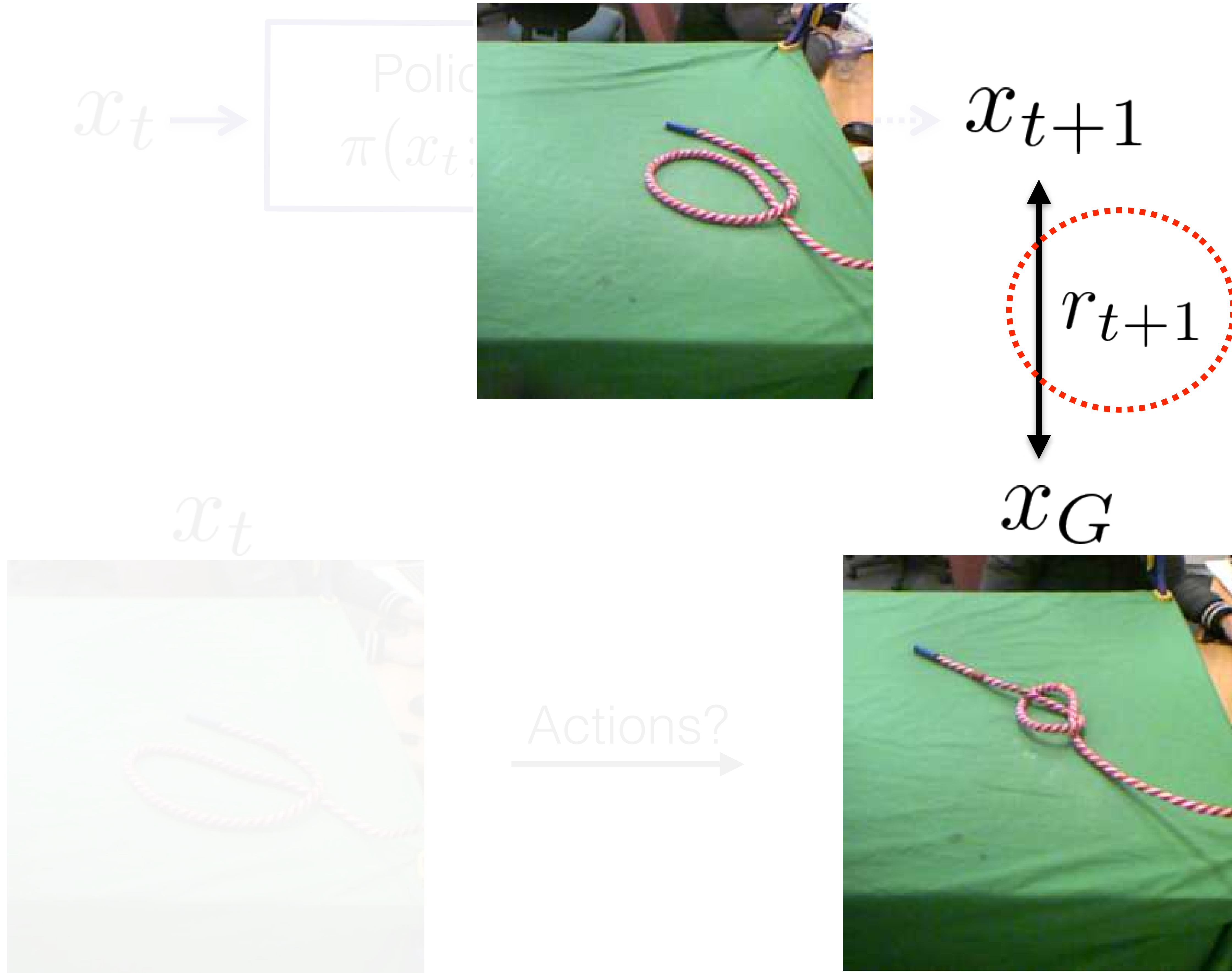
r_{t+1}

x_G

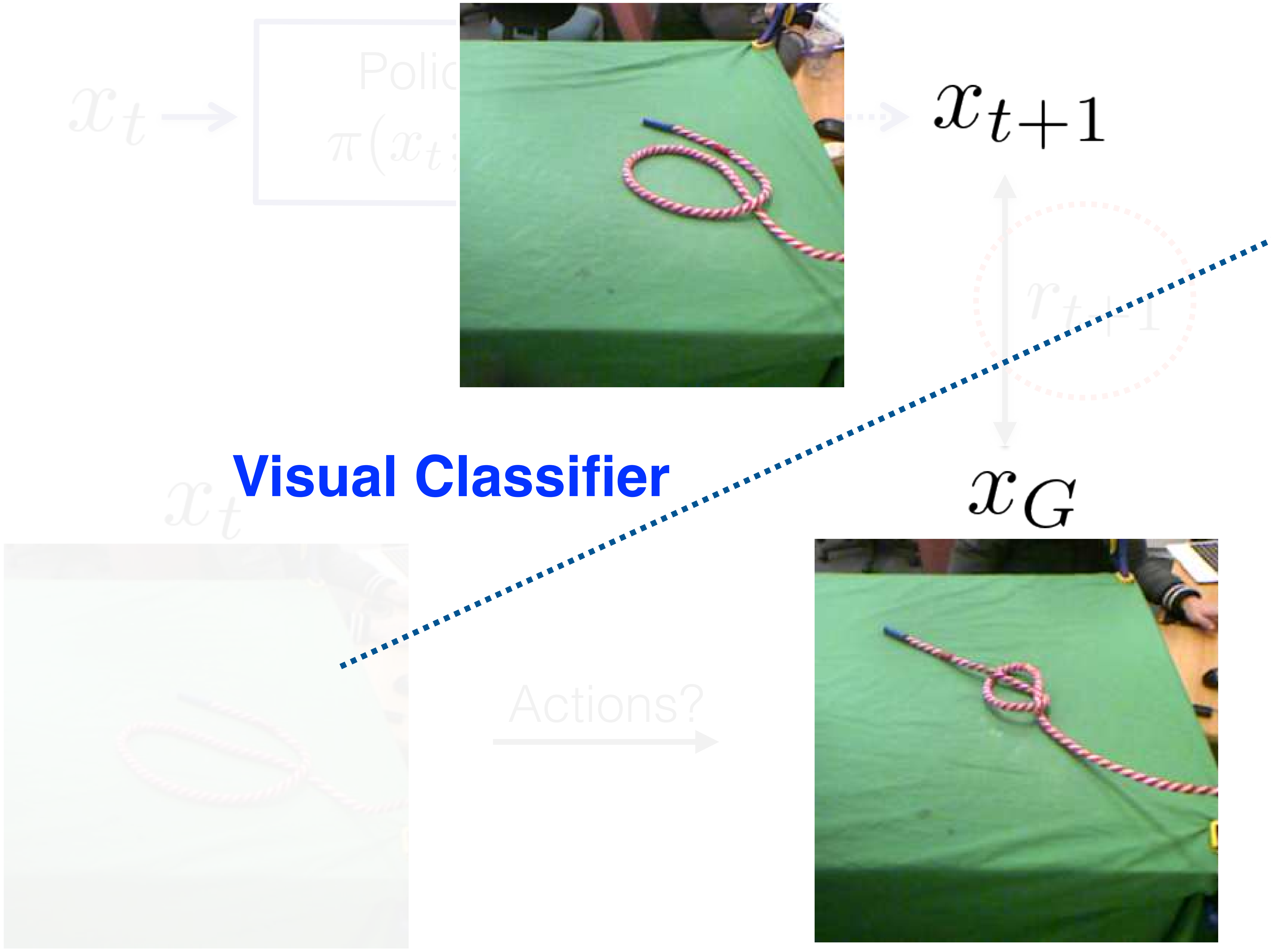
Actions?

Game score is the reward!

One Approach: Reinforcement Learning



One Approach: Reinforcement Learning

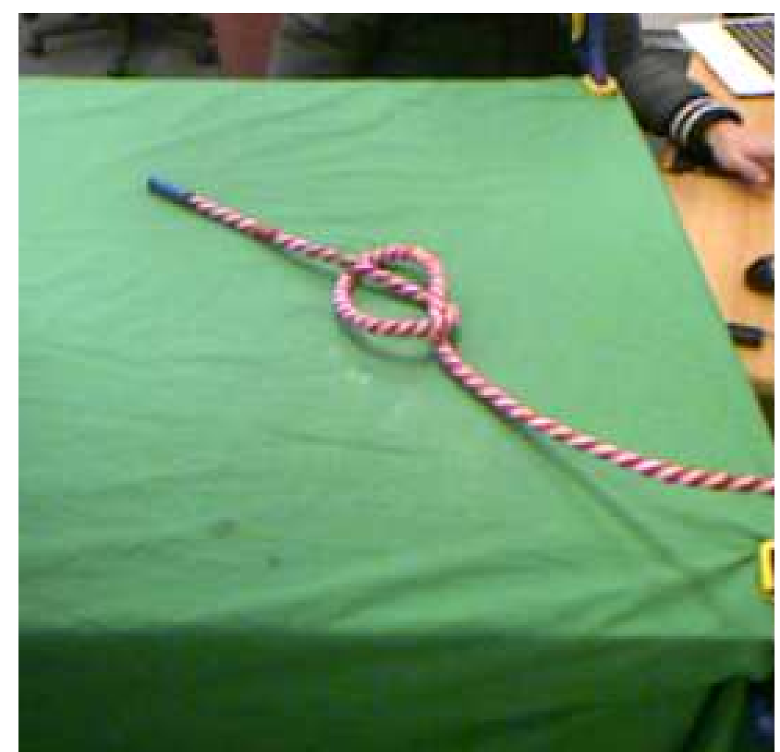


One Approach: Reinforcement Learning

x_{t+1}

Visual Classifier

x_G



One Approach: Reinforcement Learning

x_{t+1}

repeat for every goal!

Visual Classifier

x_G

Issues with Reinforcement Learning

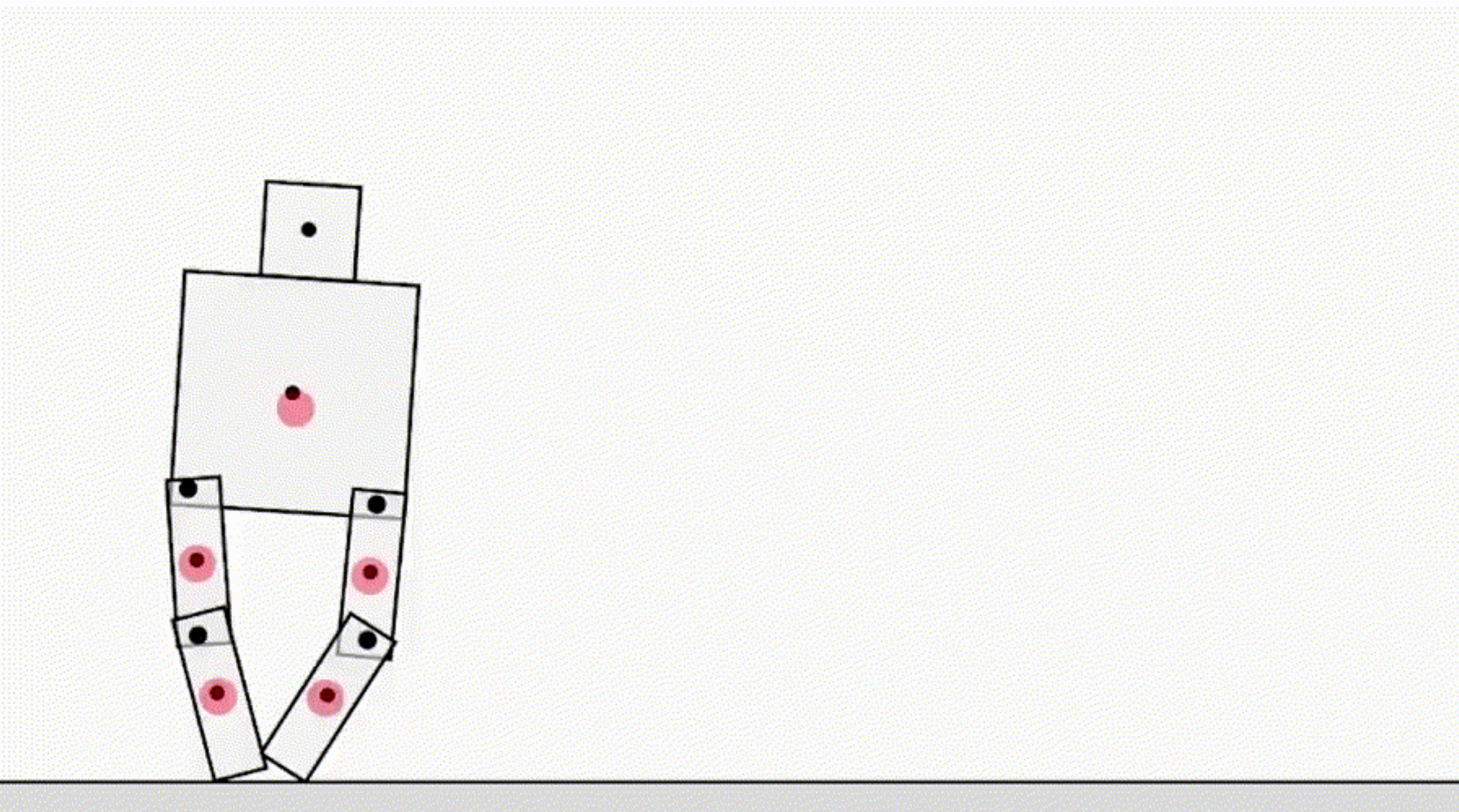
Lots of data

Where do rewards come from?

Task Specific

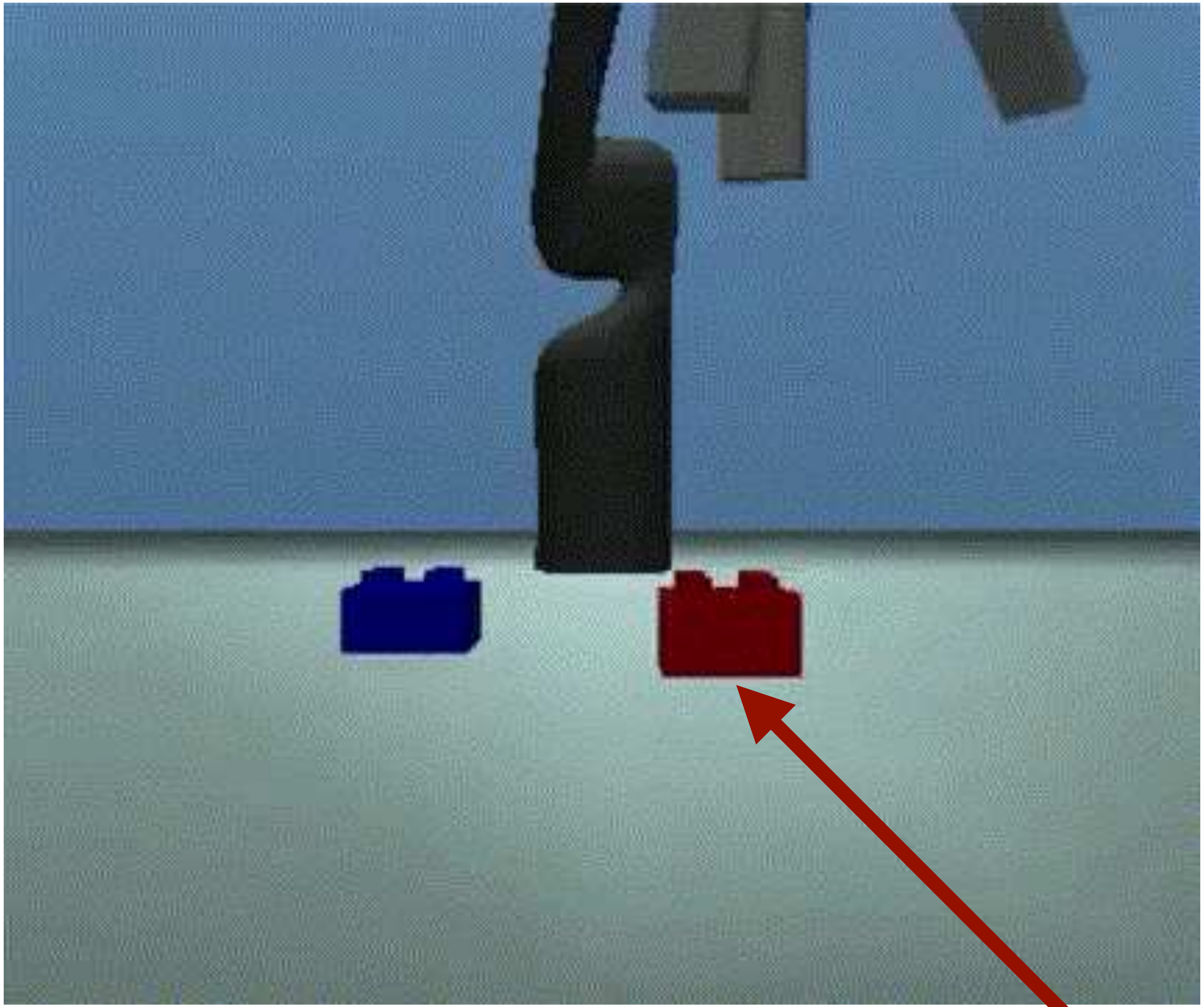
actions?

Learning to walk using RL

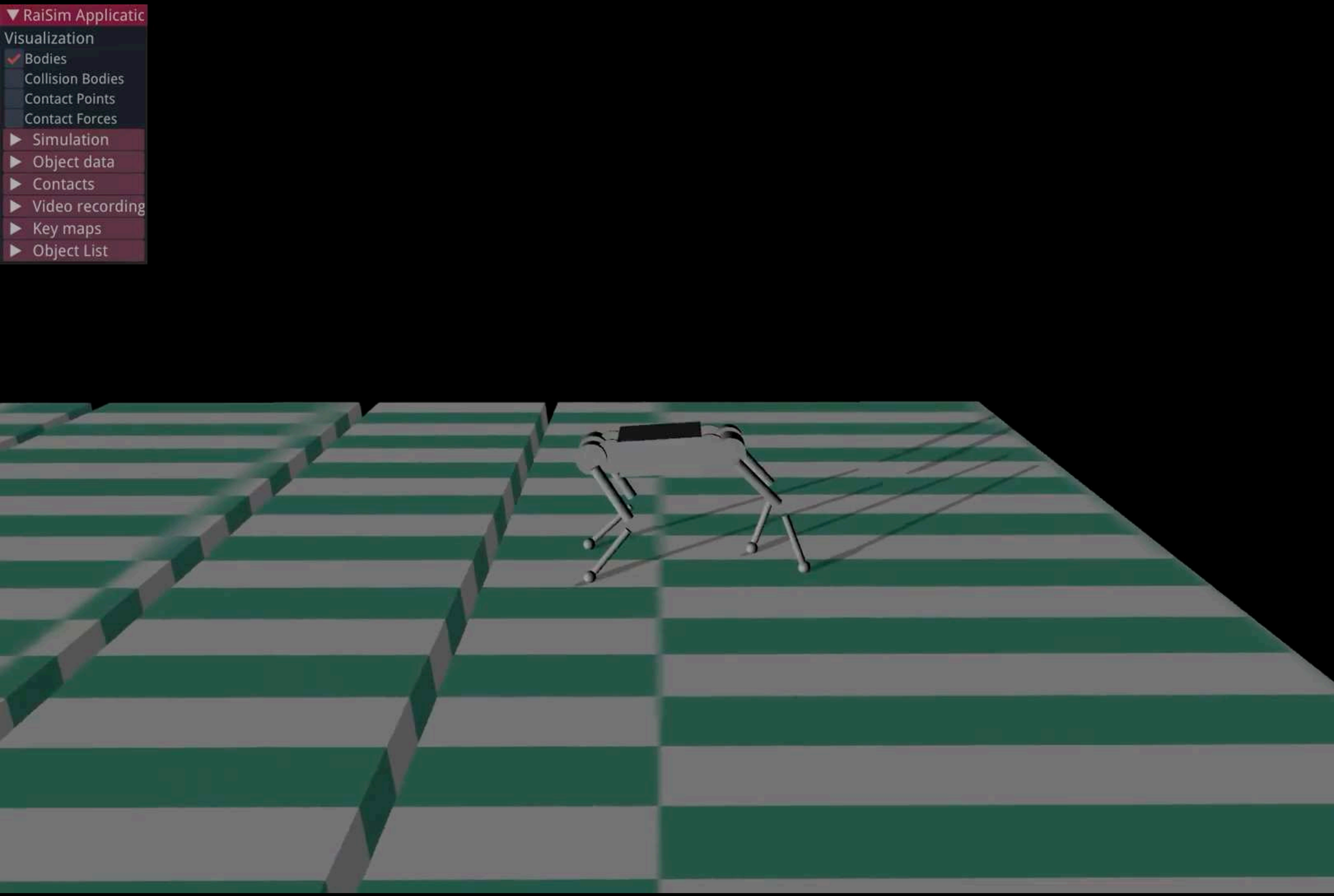


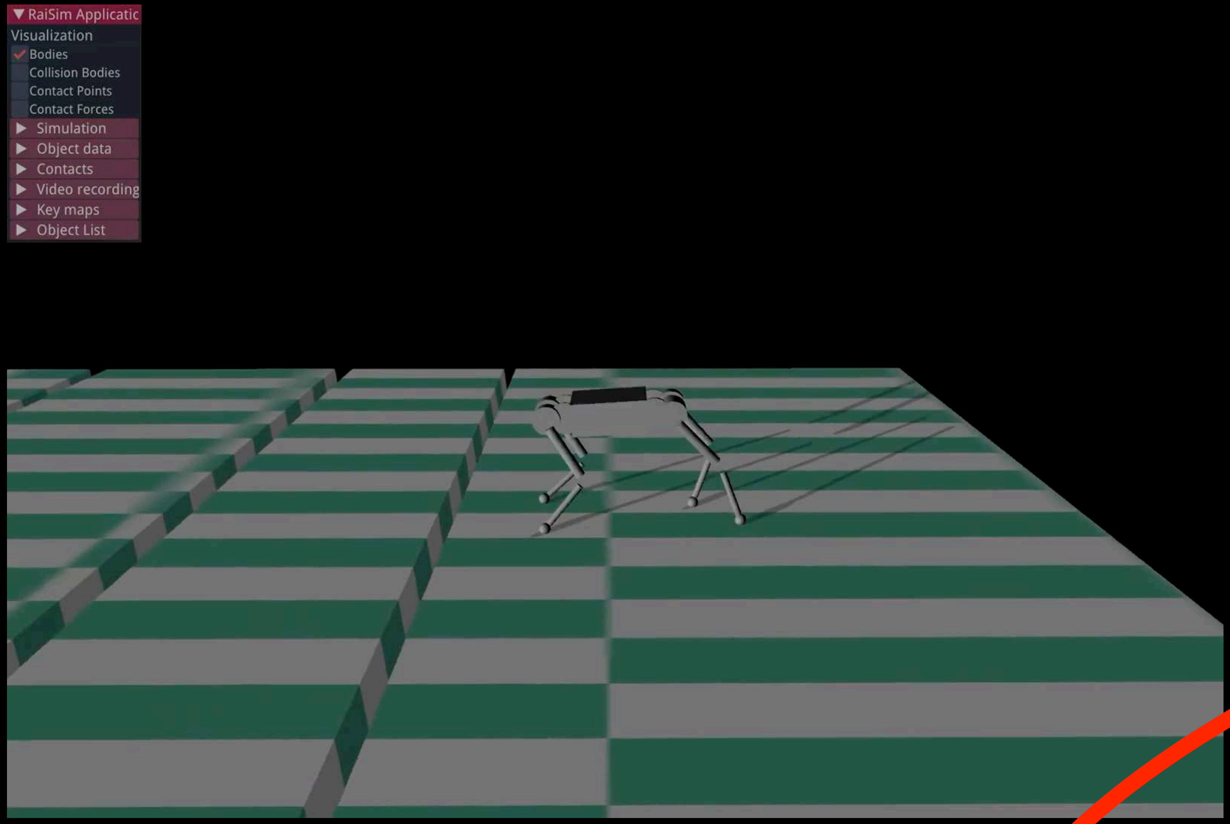
Reward to move right

Learning to stack using RL

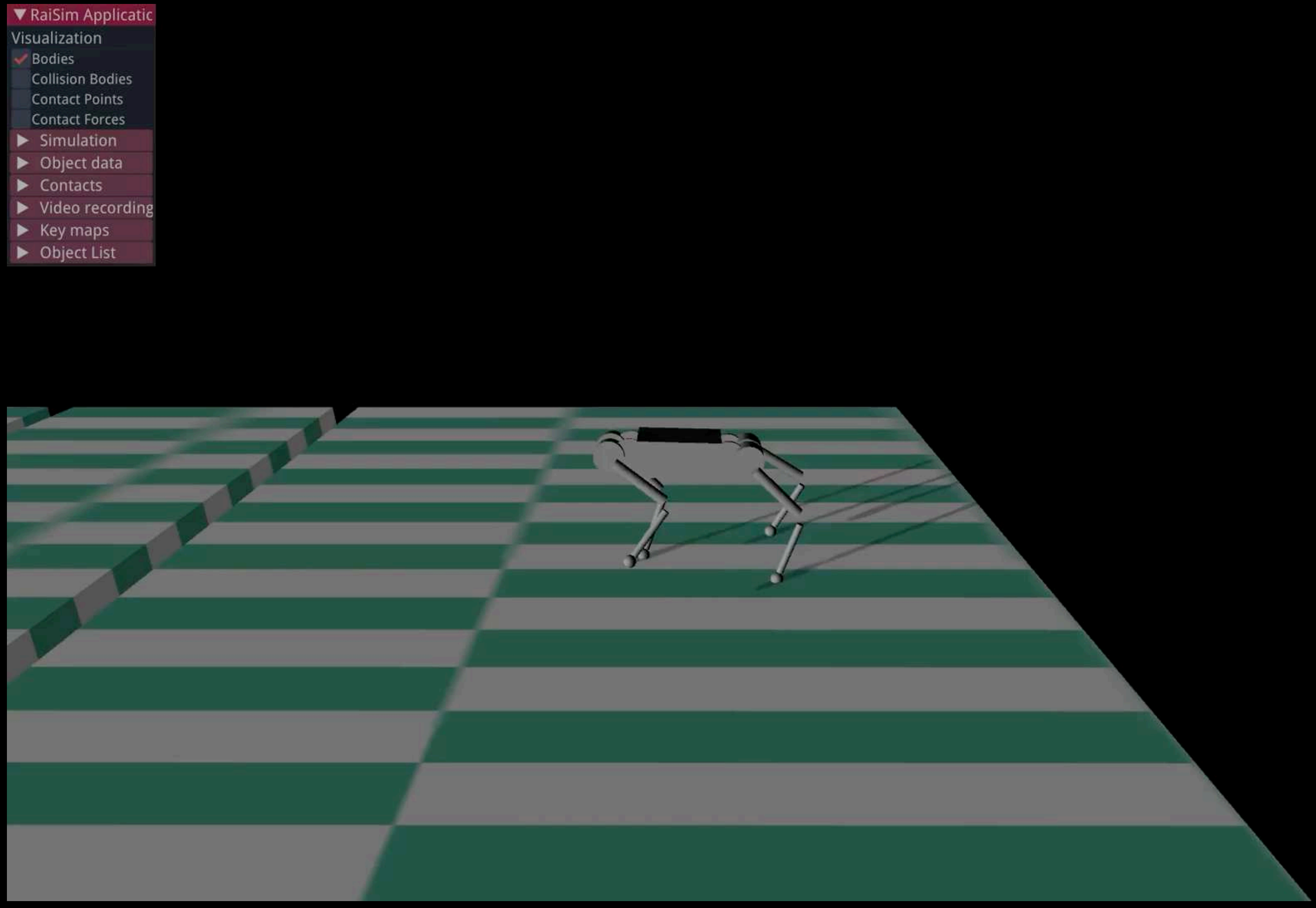


Reward for the bottom of the block
to be raised

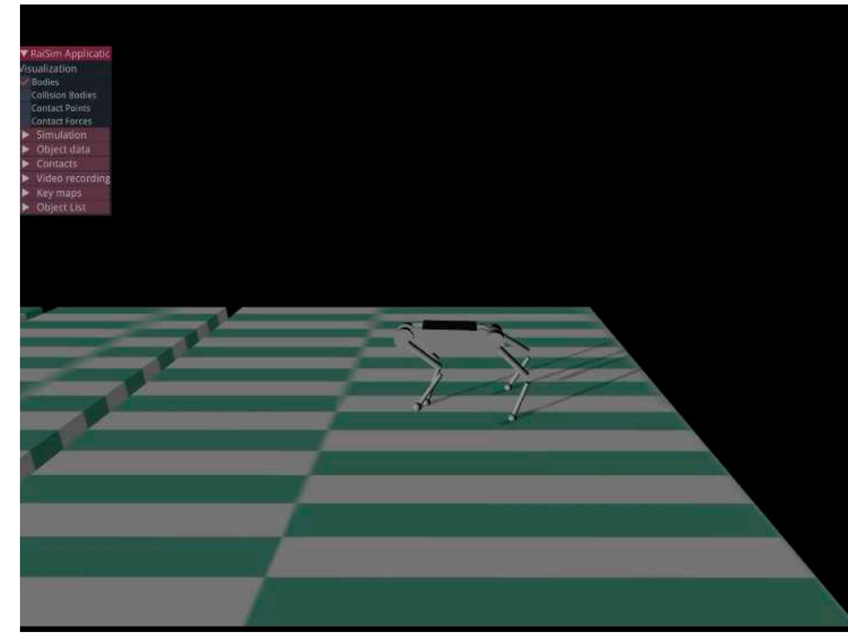

RaiSim Applicatic
Visualization
Bodies
Collision Bodies
Contact Points
Contact Forces
Simulation
Object data
Contacts
Video recording
Key maps
Object List



modify reward

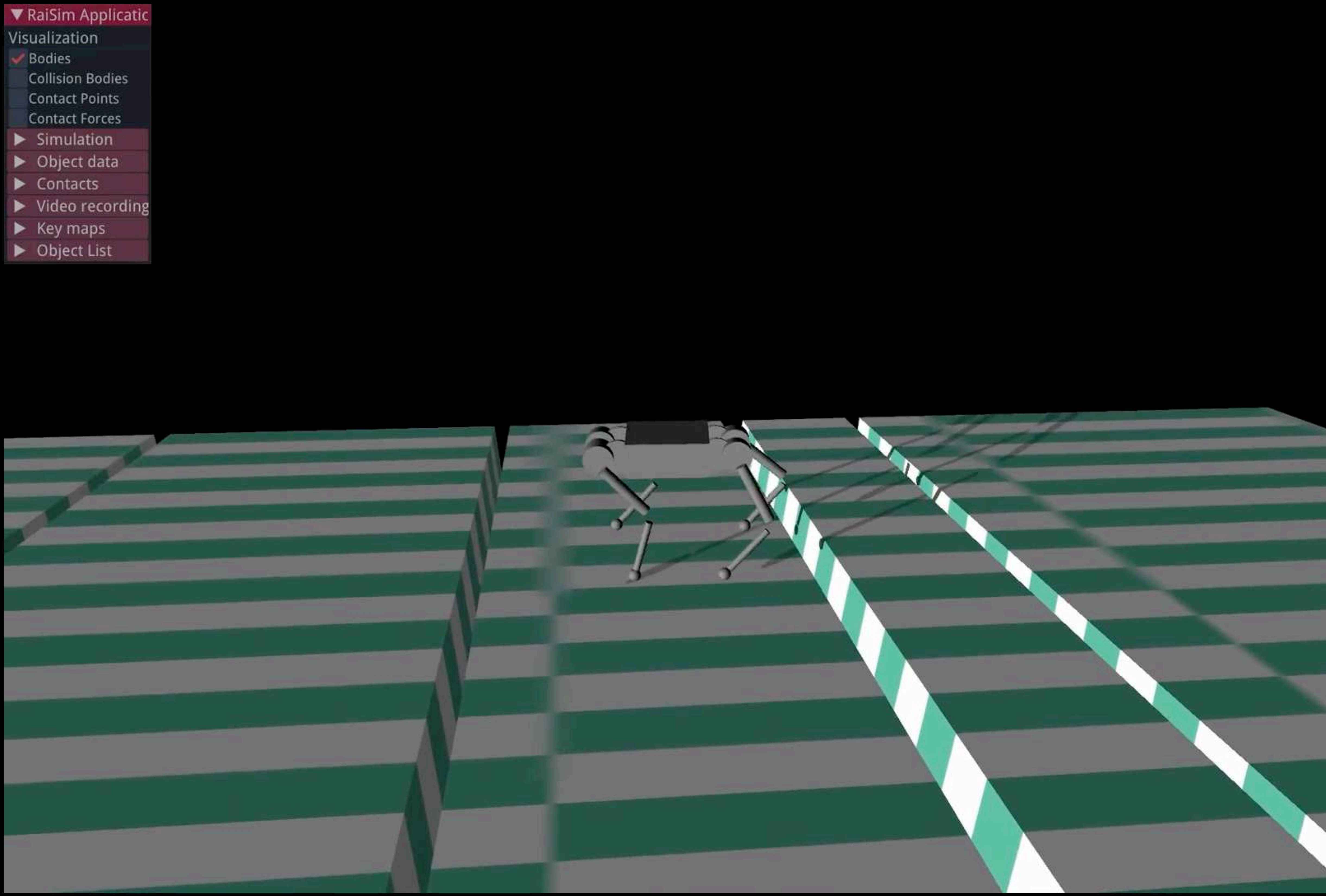


modify reward

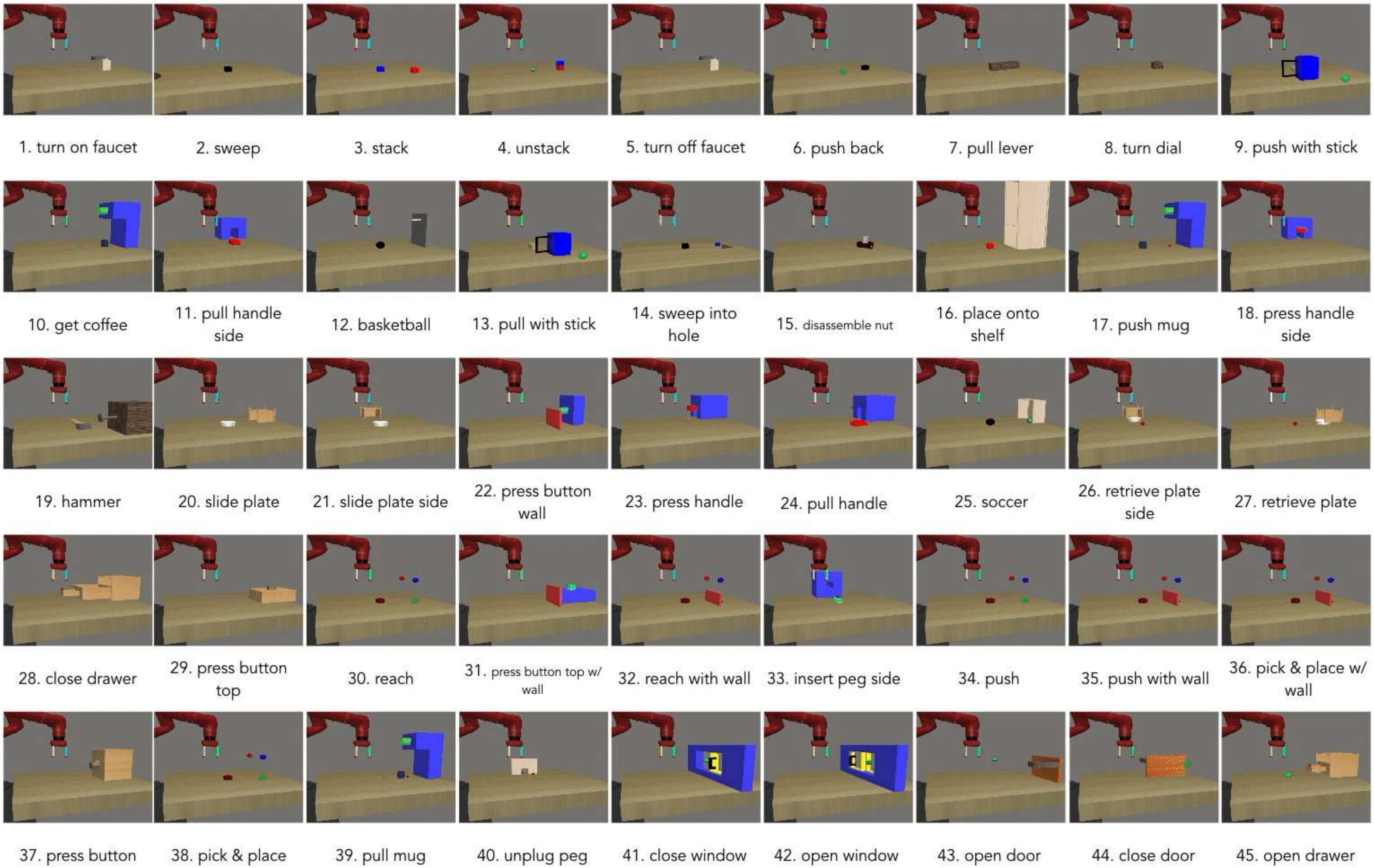


grad student reward descent

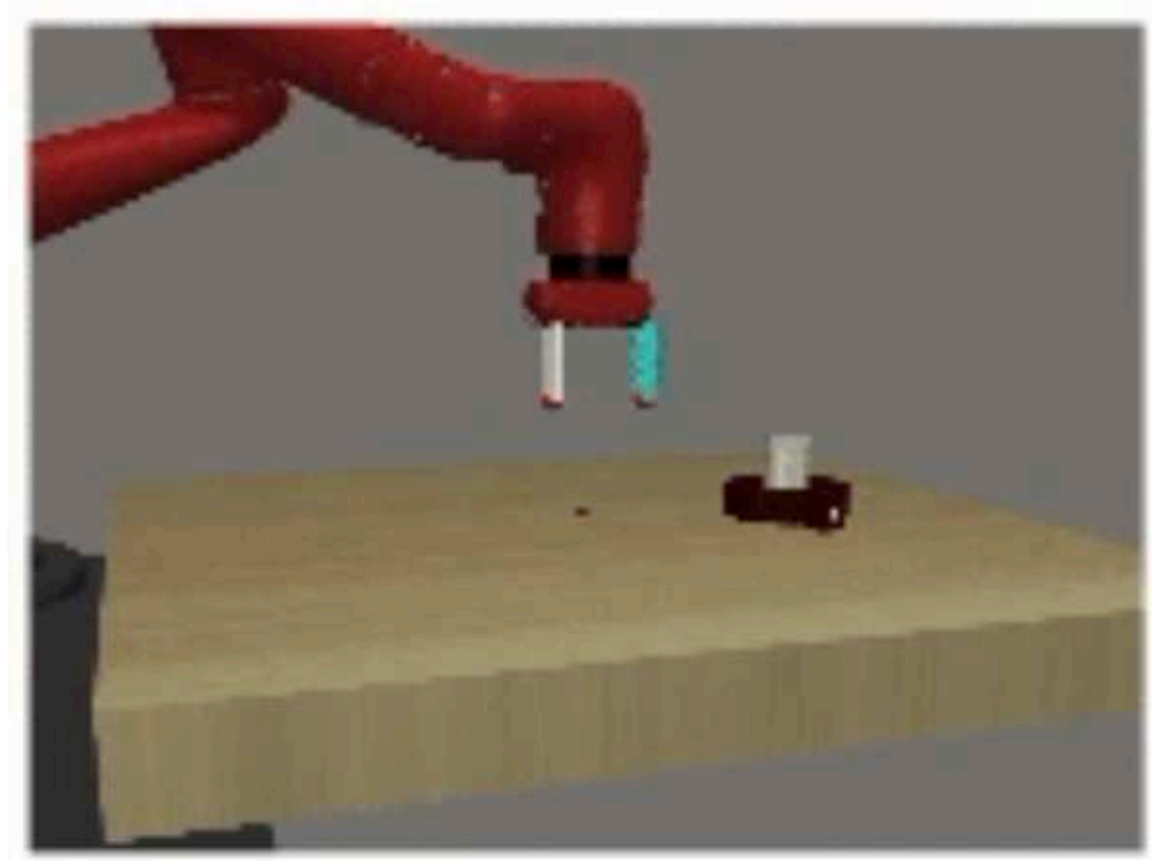
Eventually works .. but not the desired way



Train tasks



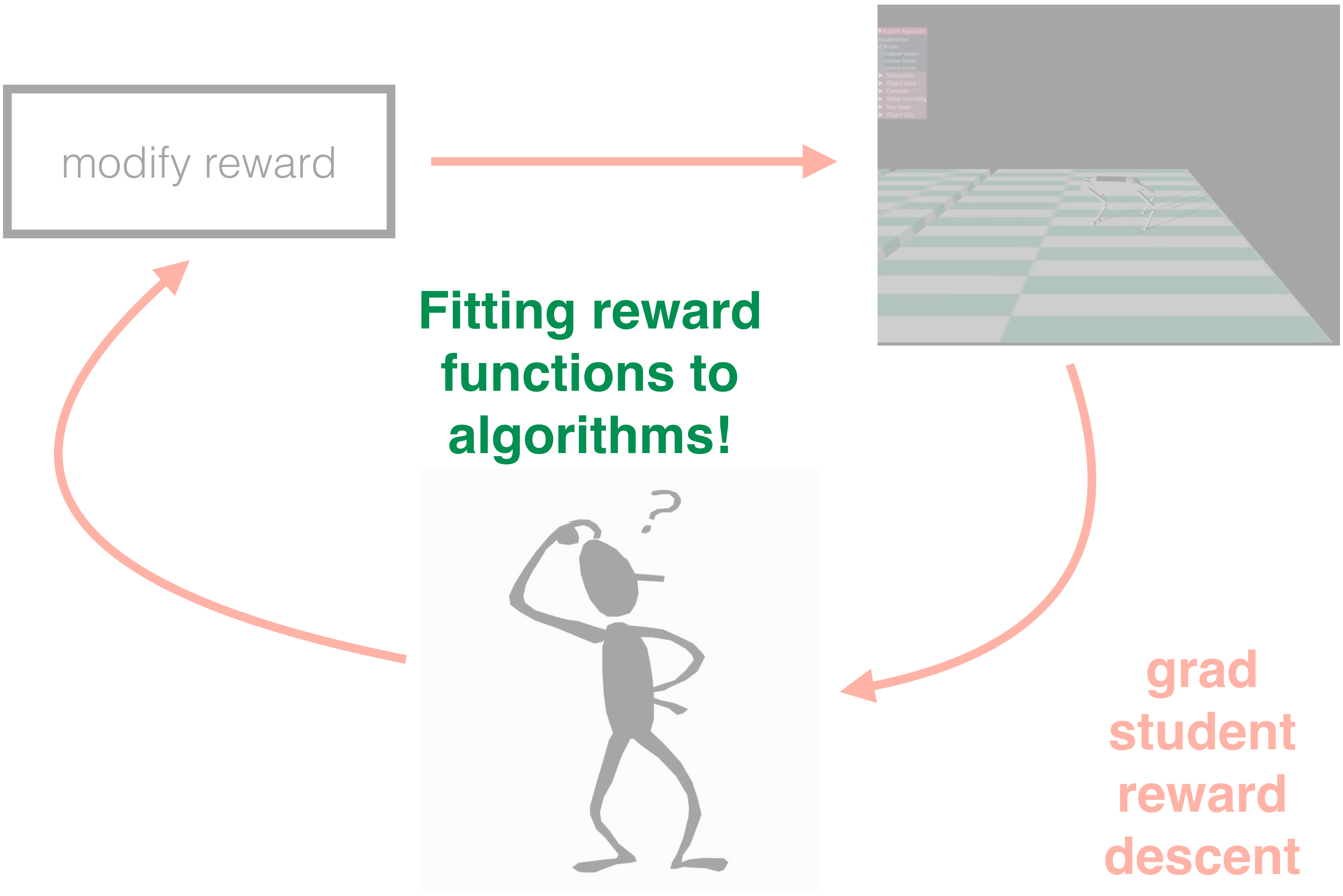
Meta-World, Yu et al., 2019



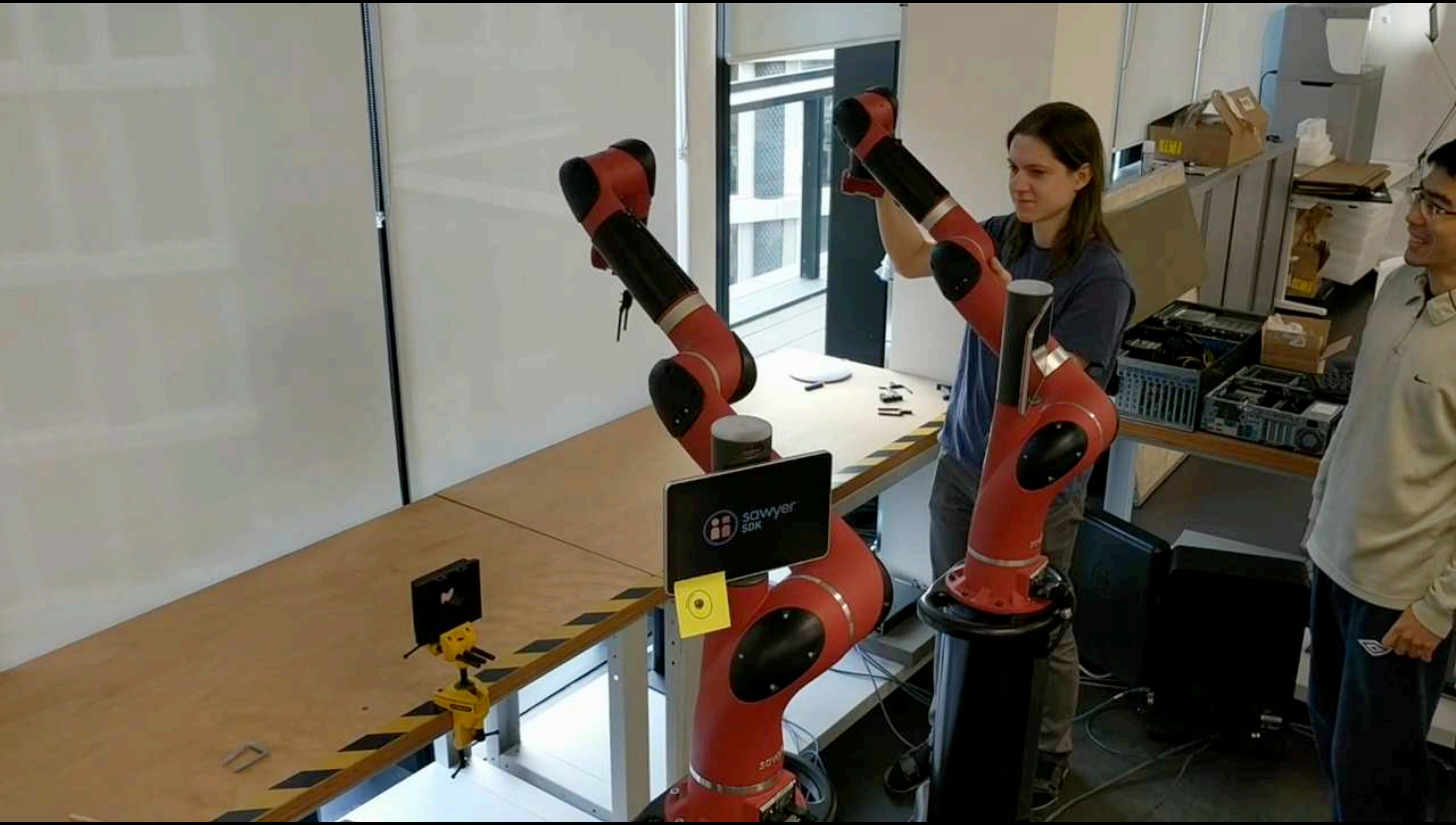
```
def reachReward():
    reachRew = -reachDist
    if reachDistxy < 0.04:
        reachRew = -reachDist
    else:
        reachRew =  -reachDistxy - 2*zDist

    # incentive to close fingers when reachDist is small
    if reachDist < 0.04:
        reachRew = -reachDist + max(actions[-1],0)/50
    return reachRew, reachDist
```

```
:**2)/c2) + np.exp(-(placingDist**2)/c3))
```

Meta-World, Yu et al., 2019


modify reward
Fitting reward functions to algorithms!
grad student reward descent

Overcoming Reward Specification: Provide Demonstrations



Leveraging Demonstrations for Deep Reinforcement Learning on Robotics Problems with Sparse Rewards, 2018

Issues with Reinforcement Learning

Lots of data

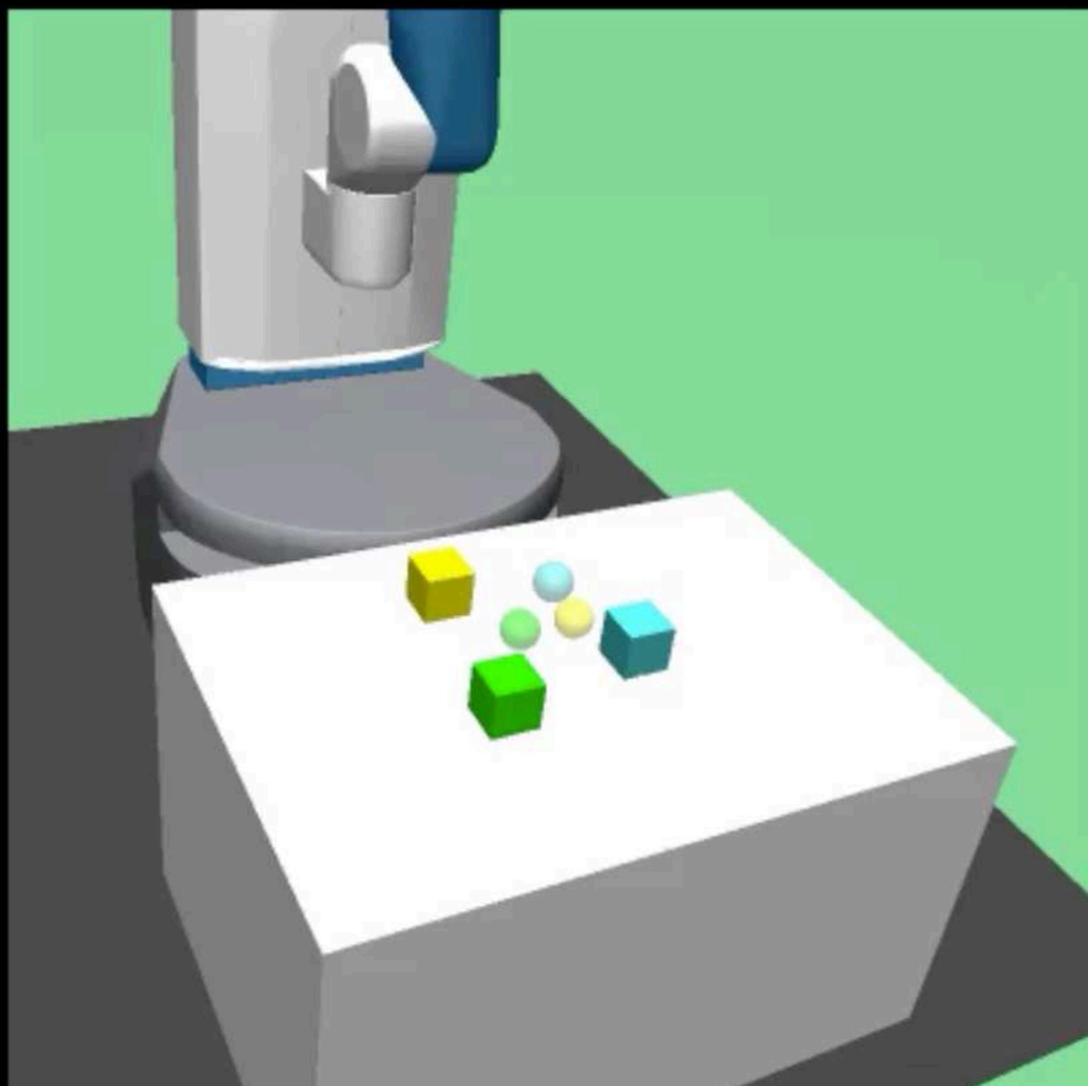
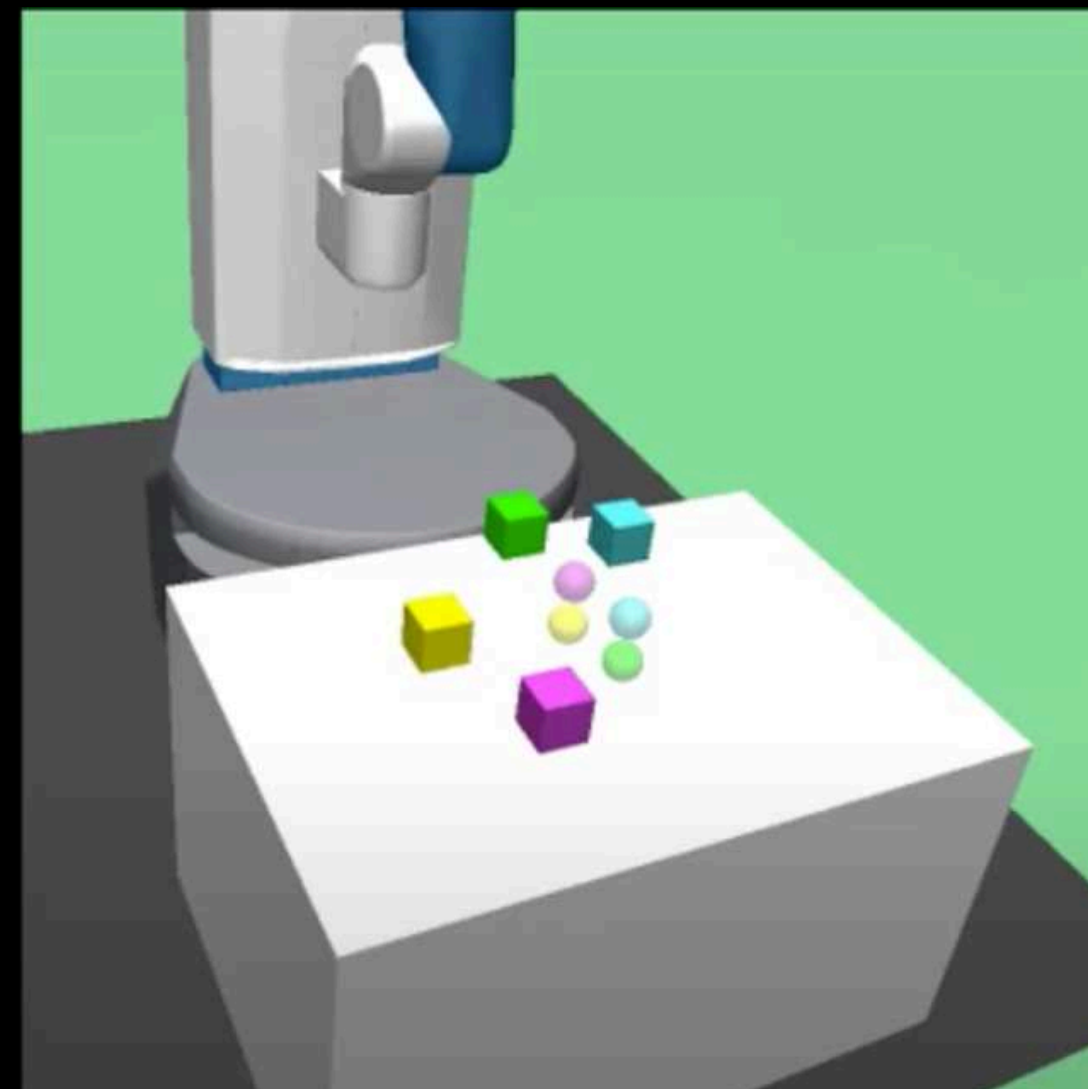
Where do rewards come from?

Task Specific

Demonstrations

(tedious to collect)

Consider Block Stacking

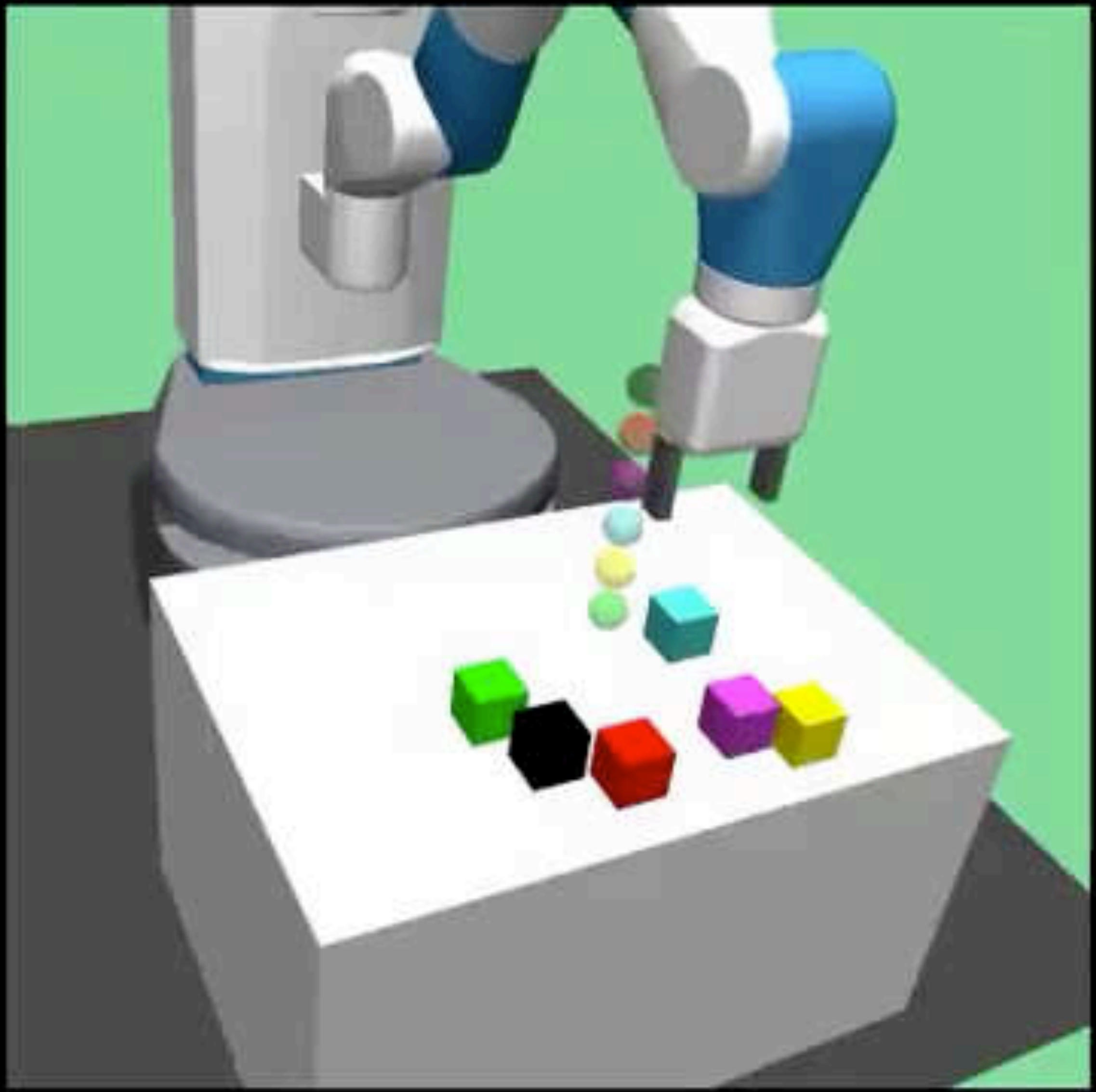
 

State Space: Position/Orientation of Blocks

Action Space: Position of end effector + open/close gripper

Pure RL on this Task

Standard RL + No Curriculum



30 mil steps

The case of "sparse" reward

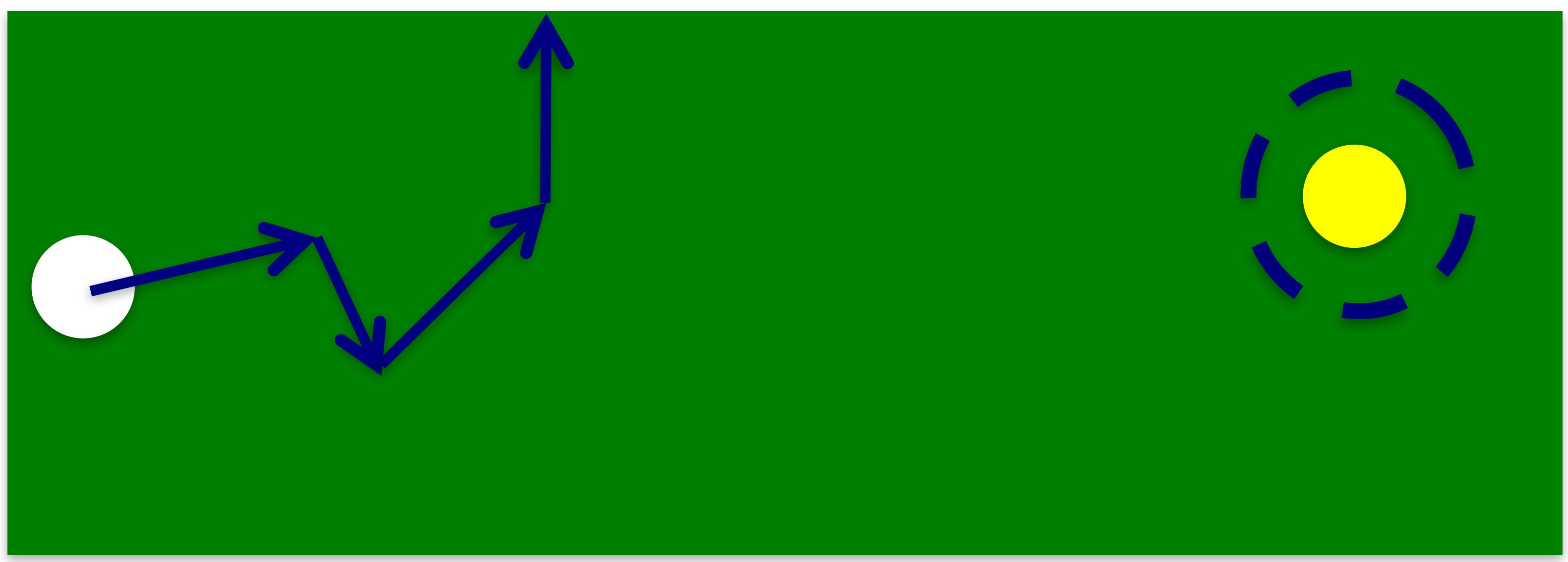
Sparse Rewards: Typically easy to define

The Sparse Reward Problem

The Sparse Reward Problem

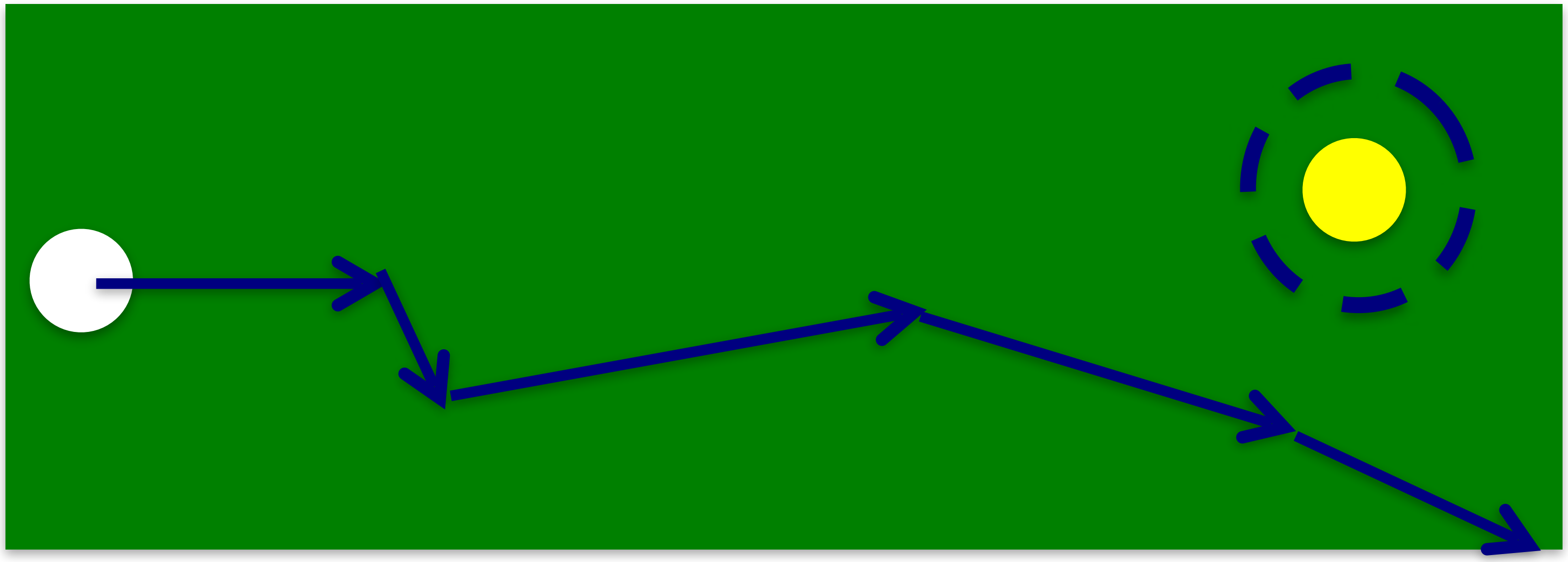
"0" Reward

The Sparse Reward Problem



“0” Reward

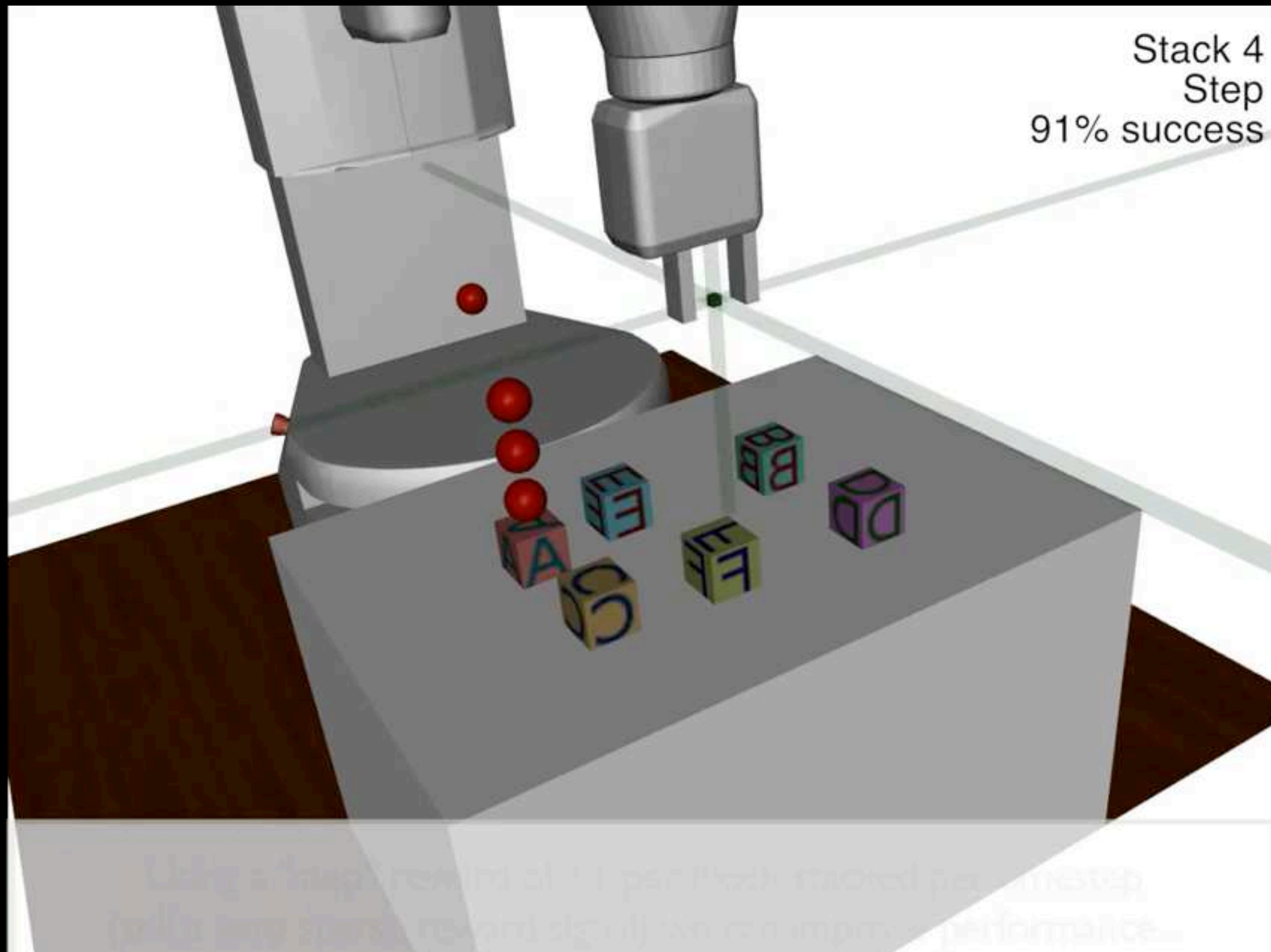
The Sparse Reward Problem



“0” Reward

Exploration Problem

Using Demonstrations to Overcome Exploration



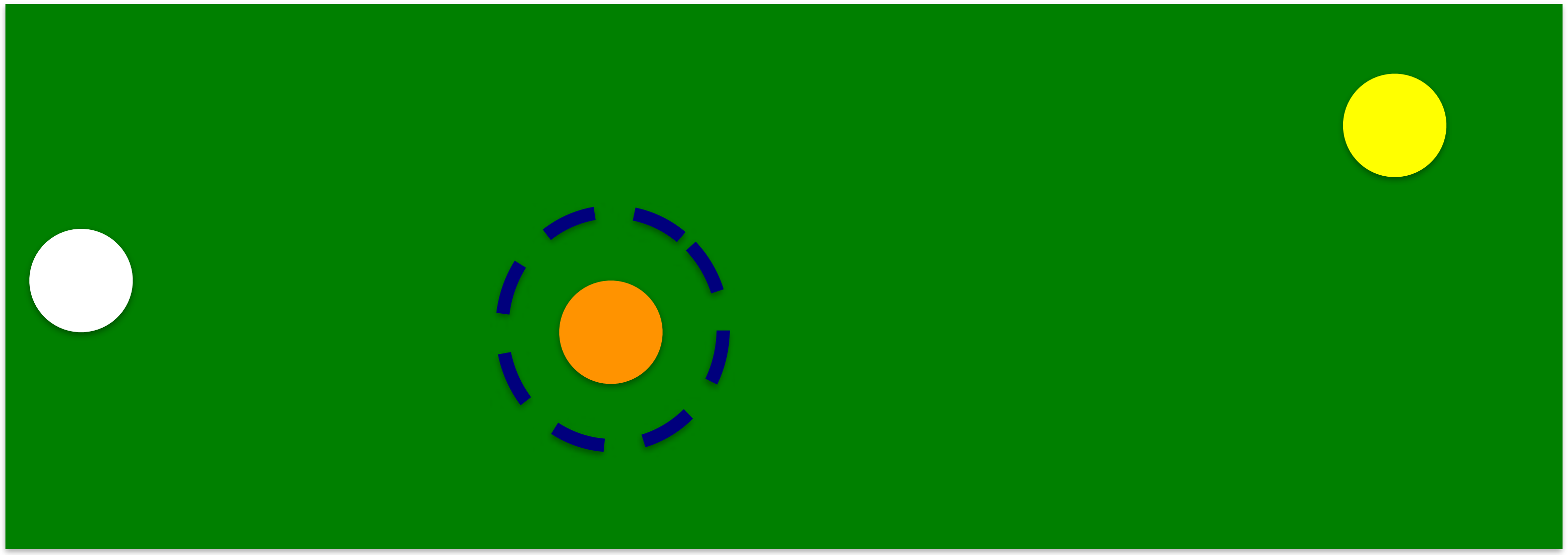
Overcoming Exploration in Reinforcement Learning with Demonstrations, Nair et al., 2018

Using Task Curriculum

Start with goal close to
initial state

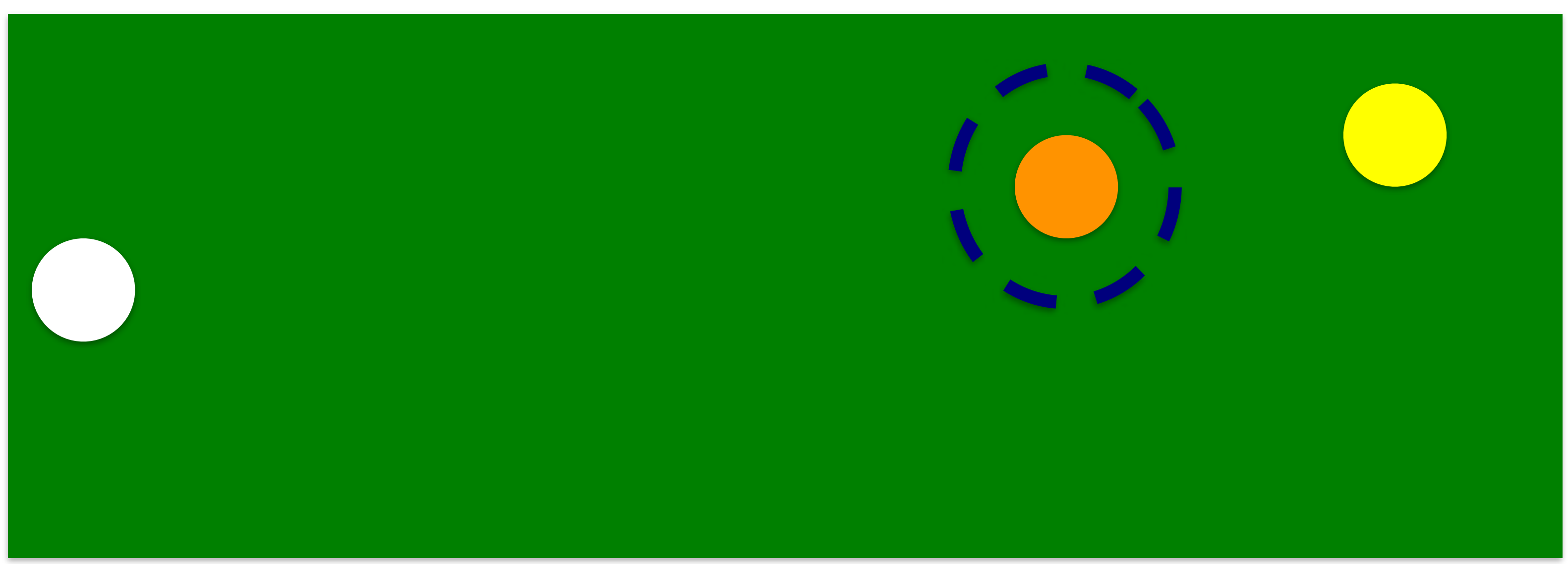
Using Task Curriculum

Slowly move the goal
farther



Using Task Curriculum

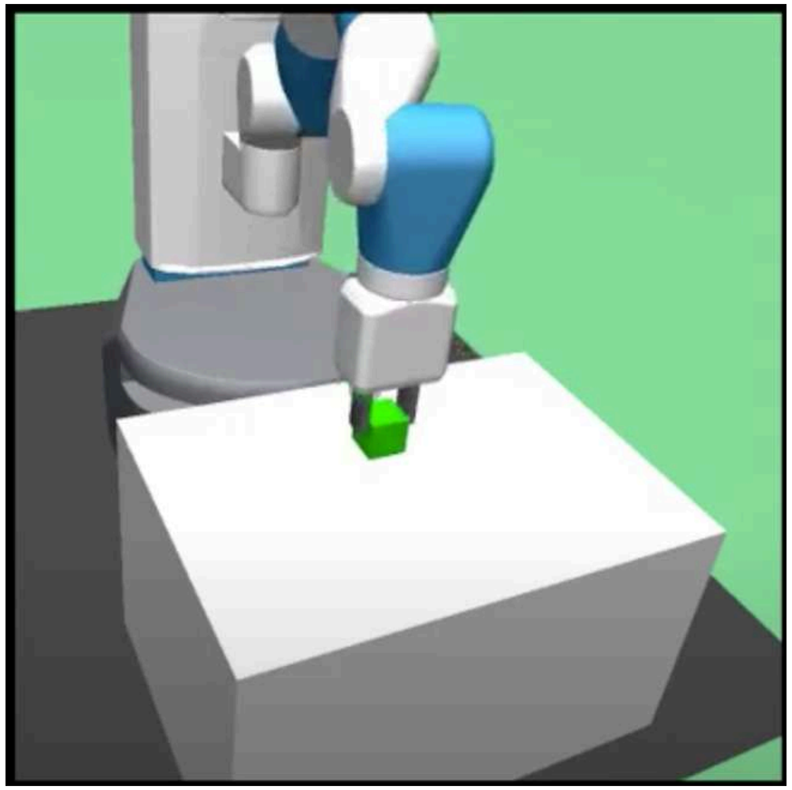
Slowly move the goal farther

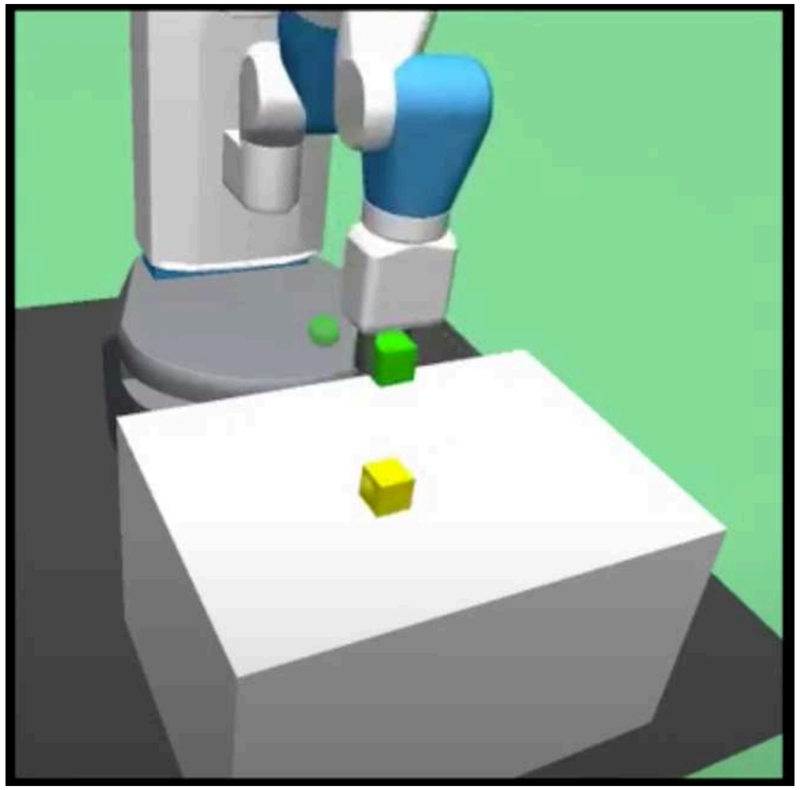


Using Task Curriculum

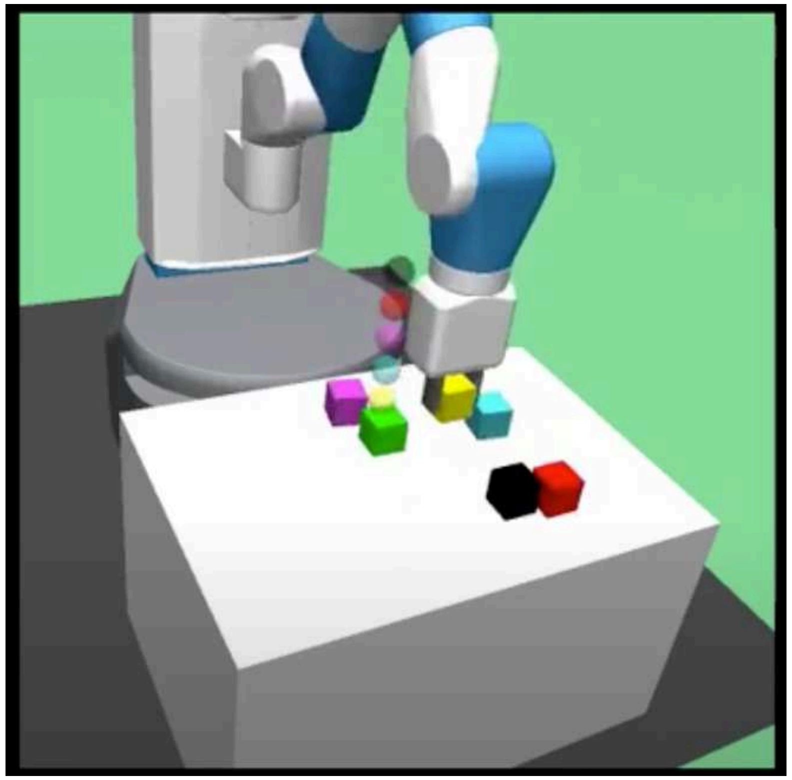
Slowly move the goal farther

Creating a Task Curriculum



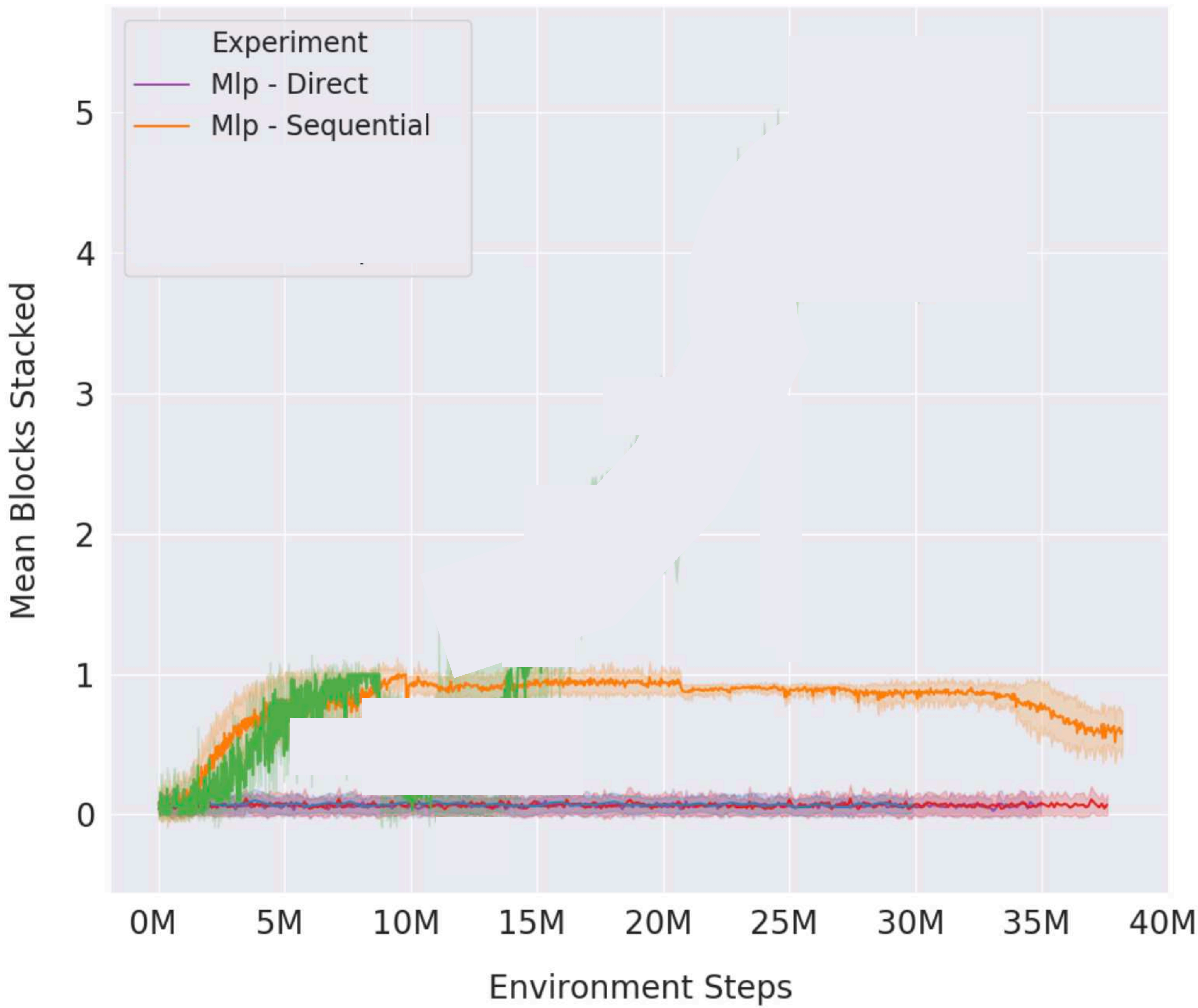


. . .

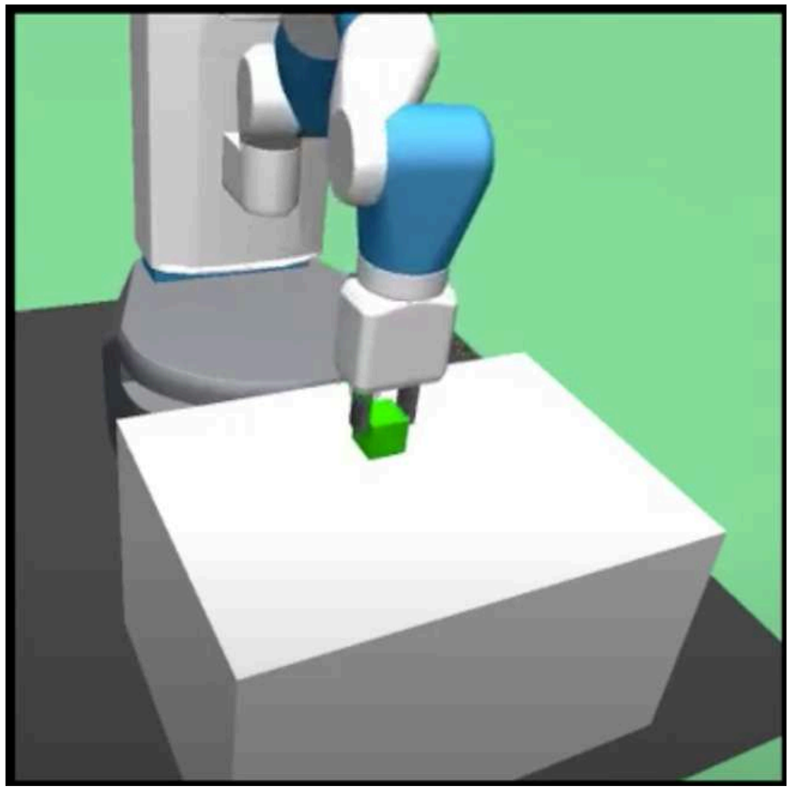


Towards Practical Robot Manipulation Using Relational Reinforcement Learning, ICRA 2020

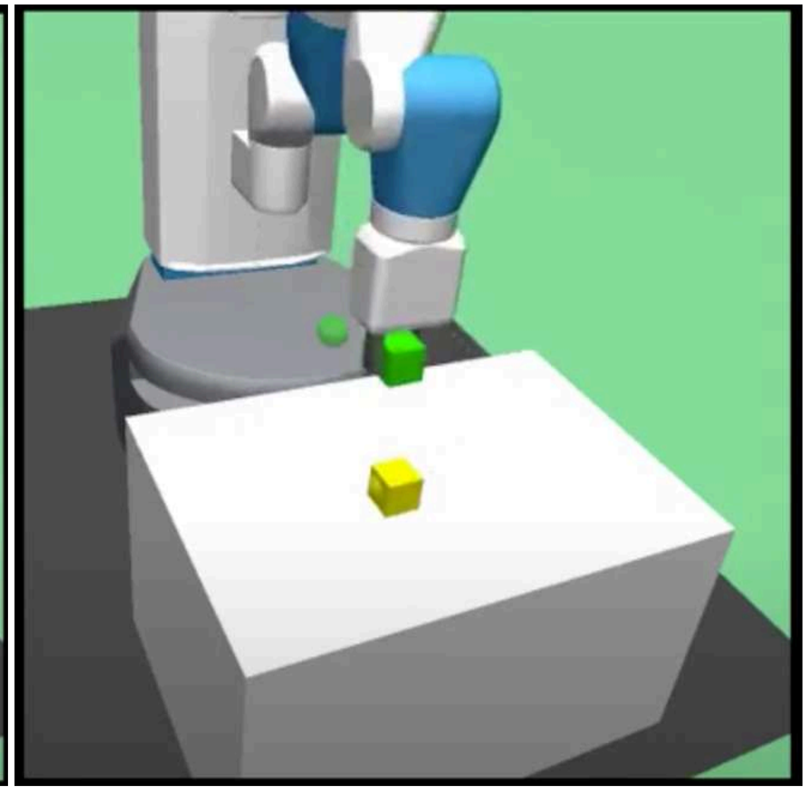
Using Curriculum Stacks Only 1 Block



Why did the curriculum fail?

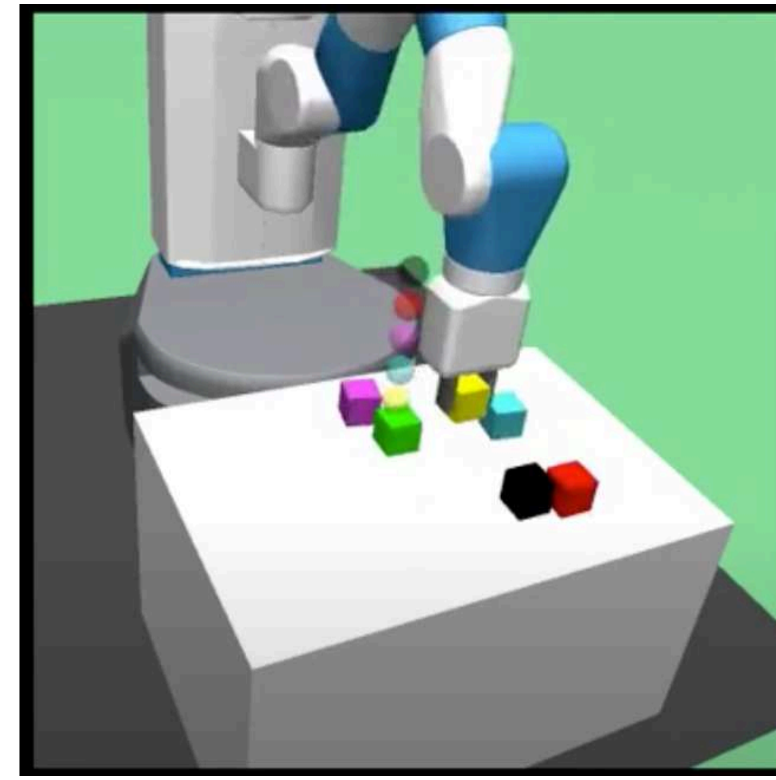


$s : (x_1, y_1)$



$s : (x_1, y_1, x_2, y_2)$

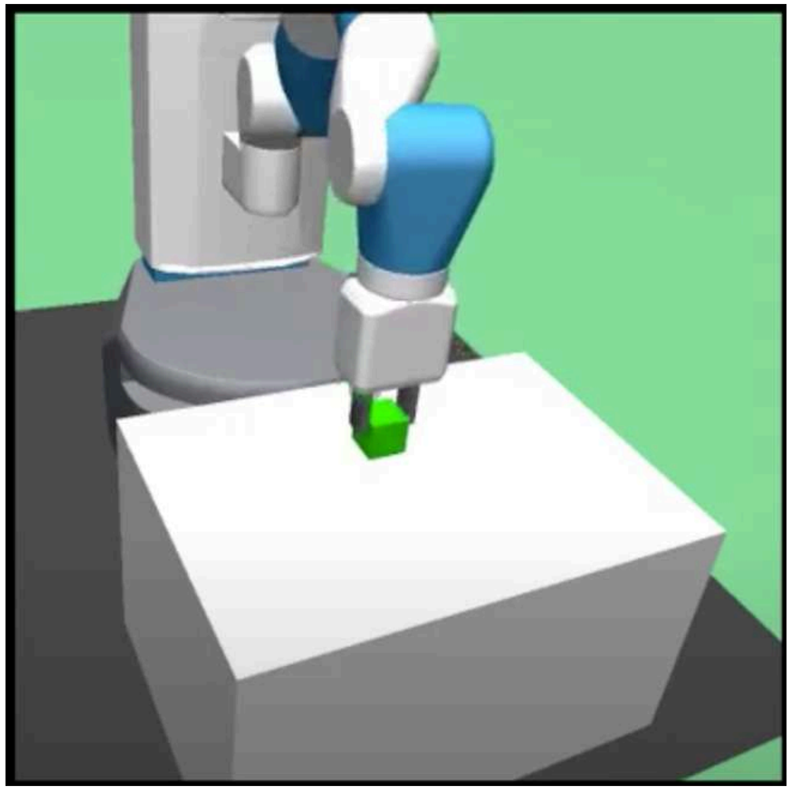
. . .

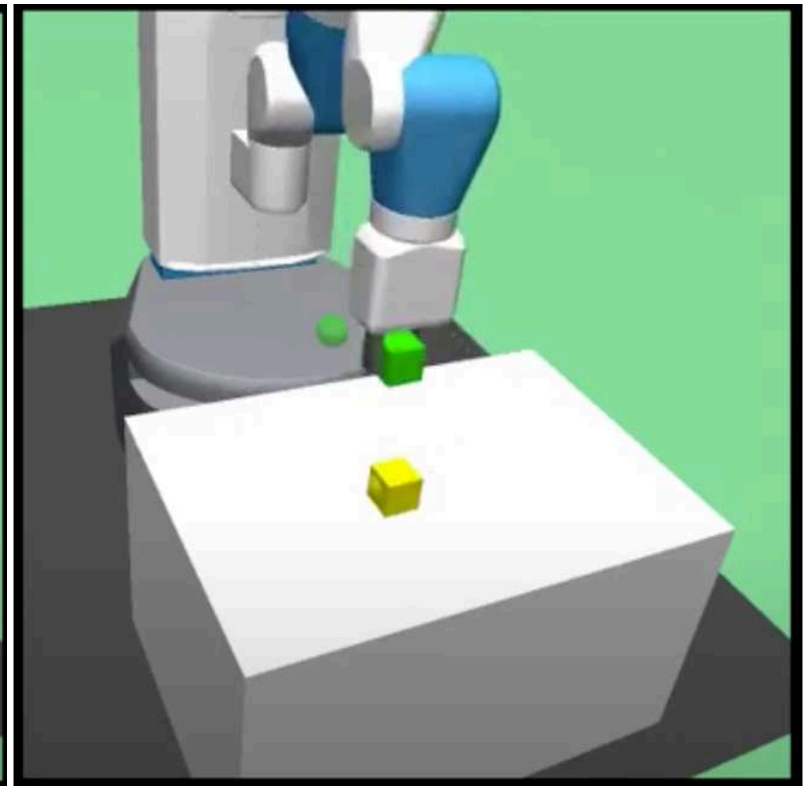


$s : (x_1, y_1, \ldots, x_N, y_N)$

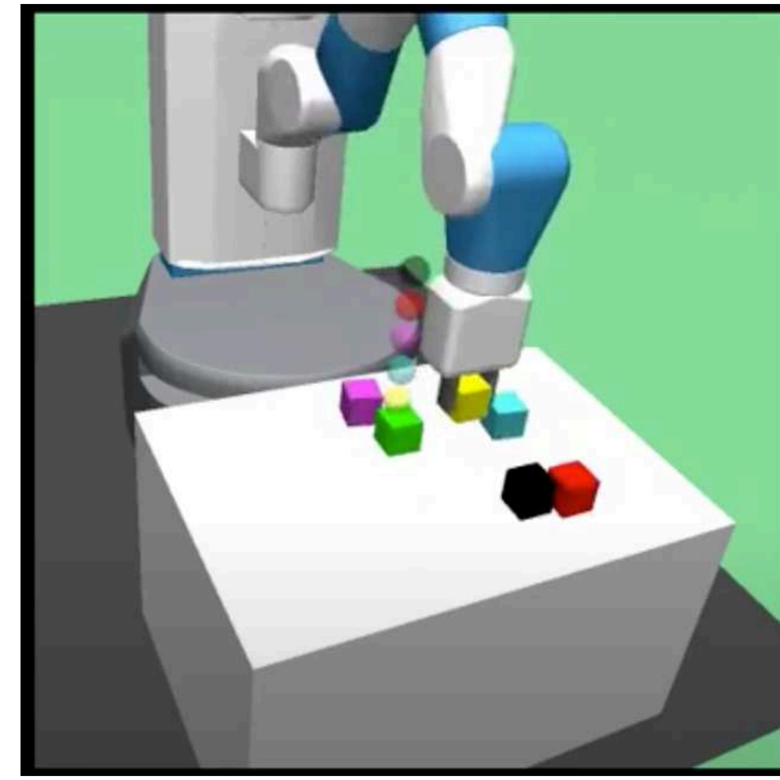
$$s \downarrow \boxed{\pi(s;\theta)} \downarrow a$$

Why did the curriculum fail?





...

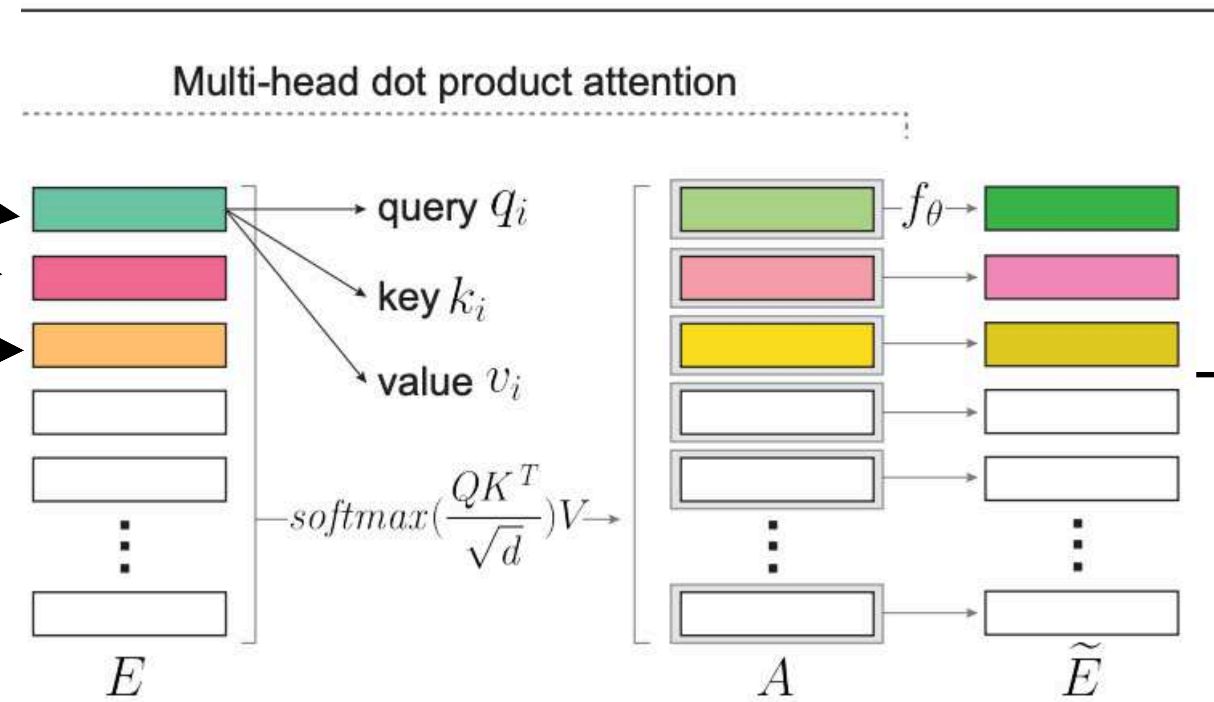


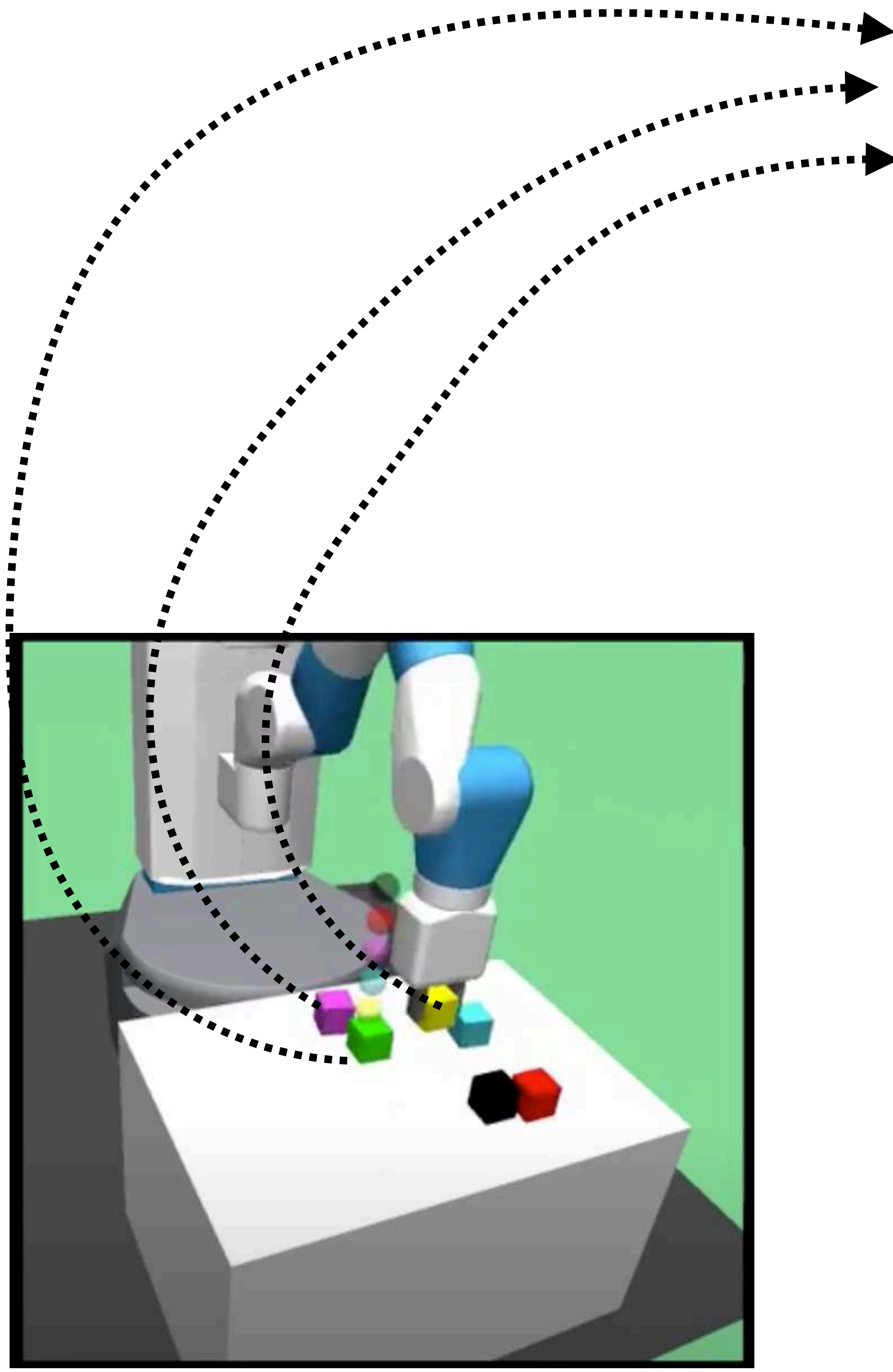
$s:(x_1, y_1)$ $\quad$ $s:(x_1, y_1, x_2, y_2)$ $\quad$ $s:(x_1, y_1, \ldots, x_N, y_N)$

$$\downarrow$$
$$\boxed{\pi(s;\theta)}$$
$$\downarrow$$
$$a$$

May not generalize to the new state space!

Graph Neural Network for Generalizable State Representation



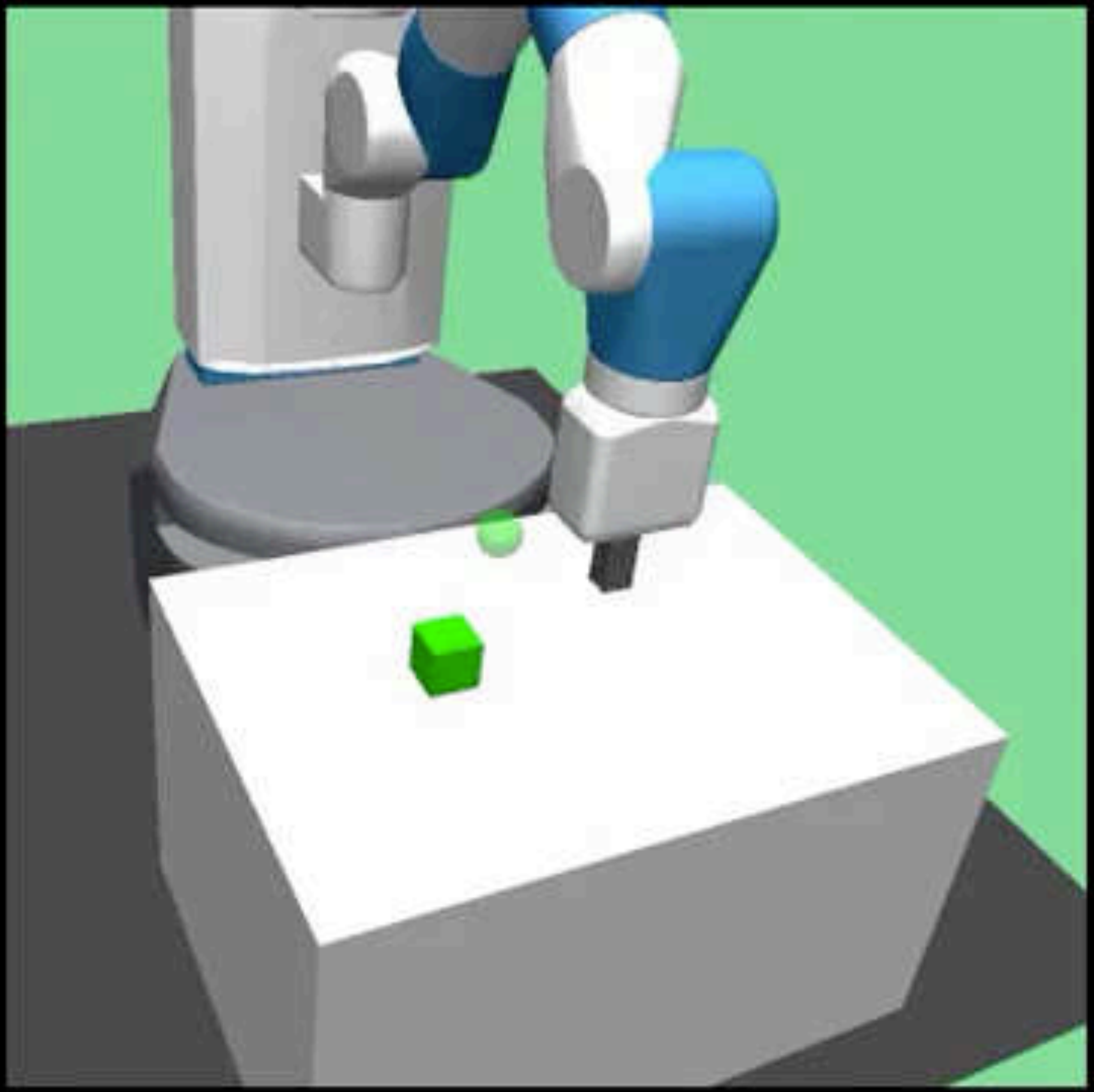


reduce

(max pool, avg pool, etc)

Figure taken from Zambaldi et al., 2018

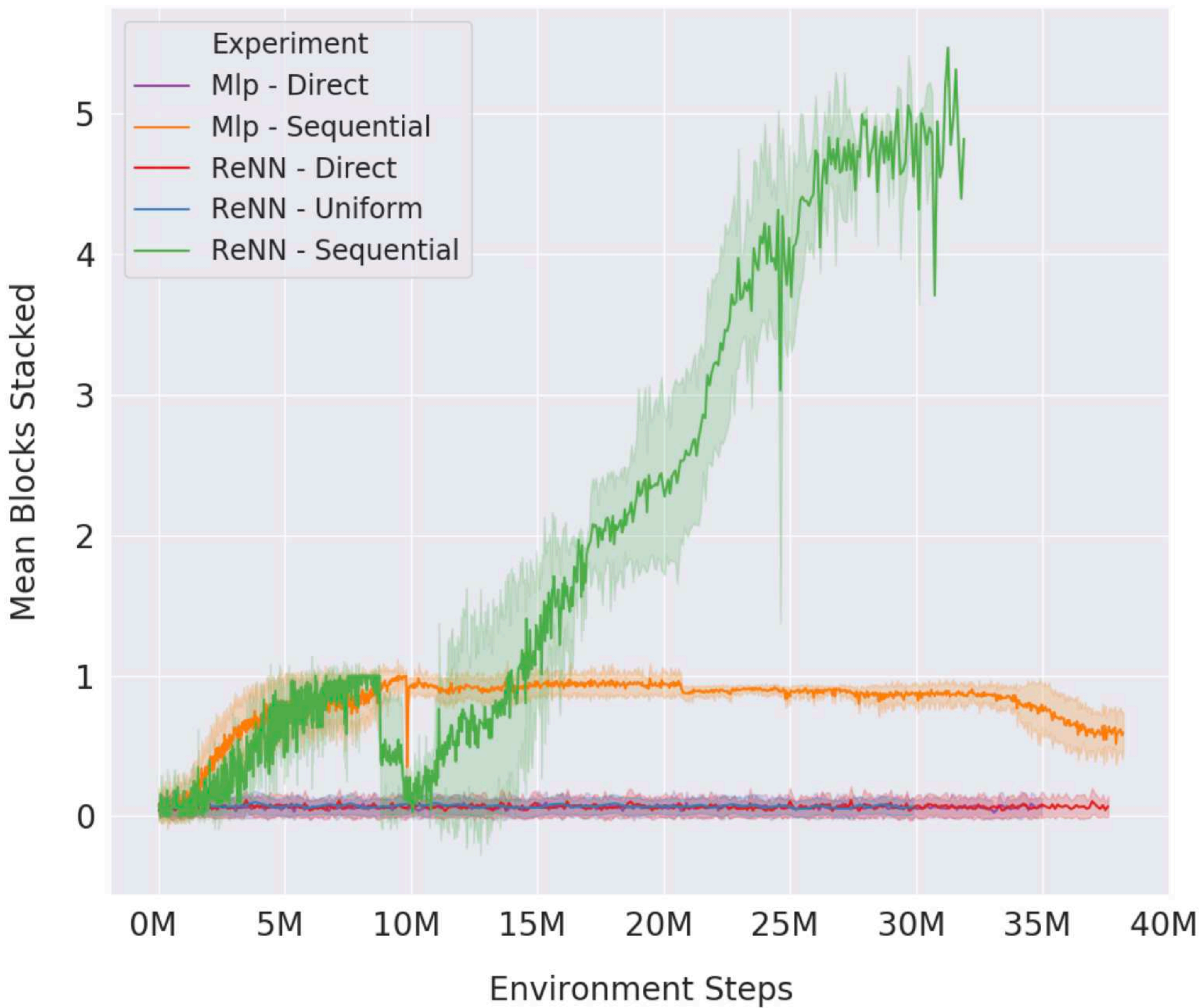
Our Method



9 mil steps

Towards Practical Robot Manipulation Using Relational Reinforcement Learning, ICRA 2020

Both Graph Network (ReNN) + Curriculum are Important



Prior Work

Nair et al.: Human Demonstrations

Task	Single Tower 4	Single Tower 5	Single Tower 6
Nair'17 [4]	91% (850M)	50% (1000M)	32% (2300M)

Towards Practical Robot Manipulation Using Relational Reinforcement Learning, ICRA 2020

Zero Shot Generalization

Generalization w/o Fine-tuning

training

Emergent Behaviors

Emergent Behaviors

Issues with Reinforcement Learning

Lots of data

Where do rewards come from?

Task Specific

Task Curriculum

(less human effort than demonstrations)

Take away

State representations must have inductive biases to generalize to more complex tasks

Issues with Reinforcement Learning

Lots of data

Where do rewards come from?

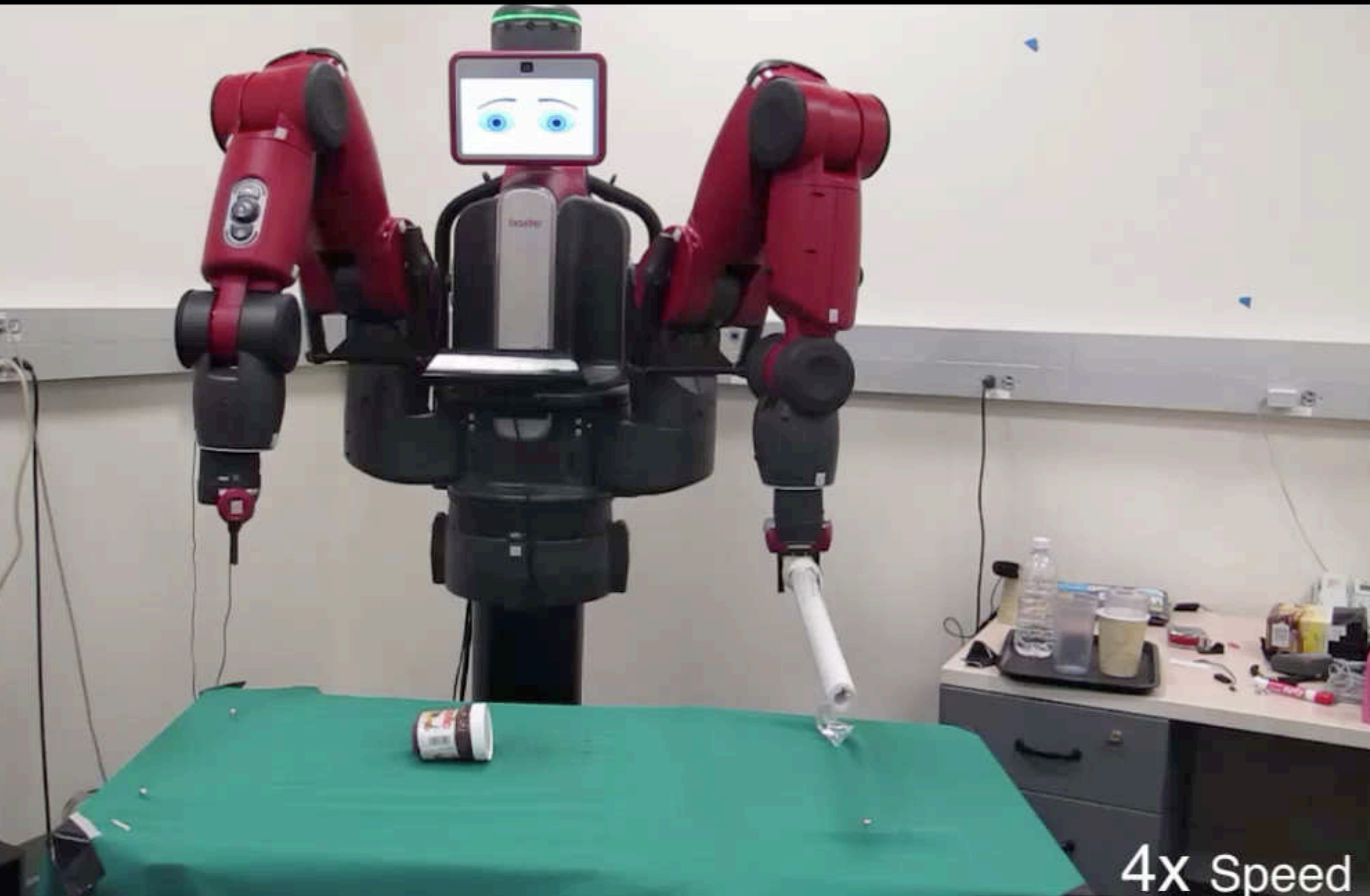
Task Specific

Demonstrations

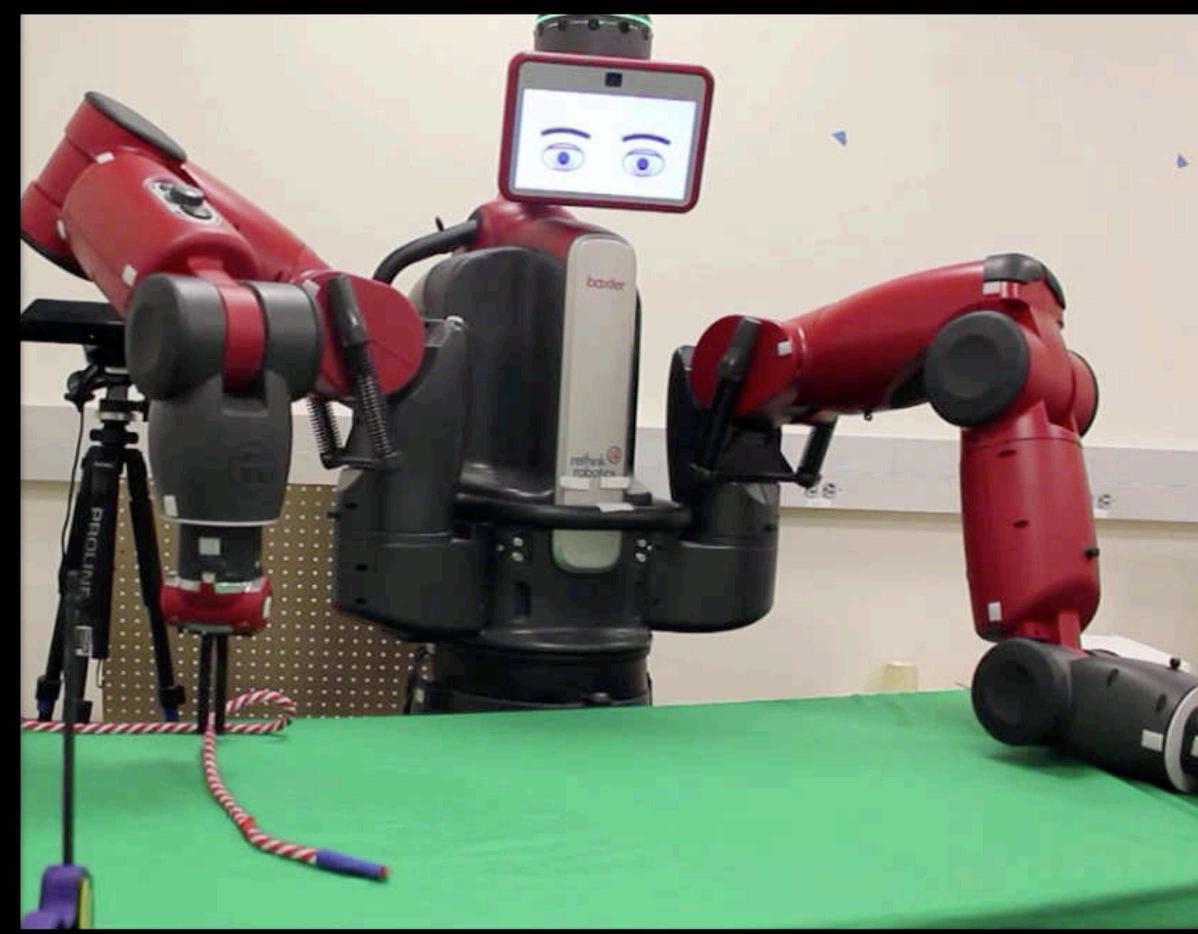
Task Curriculum

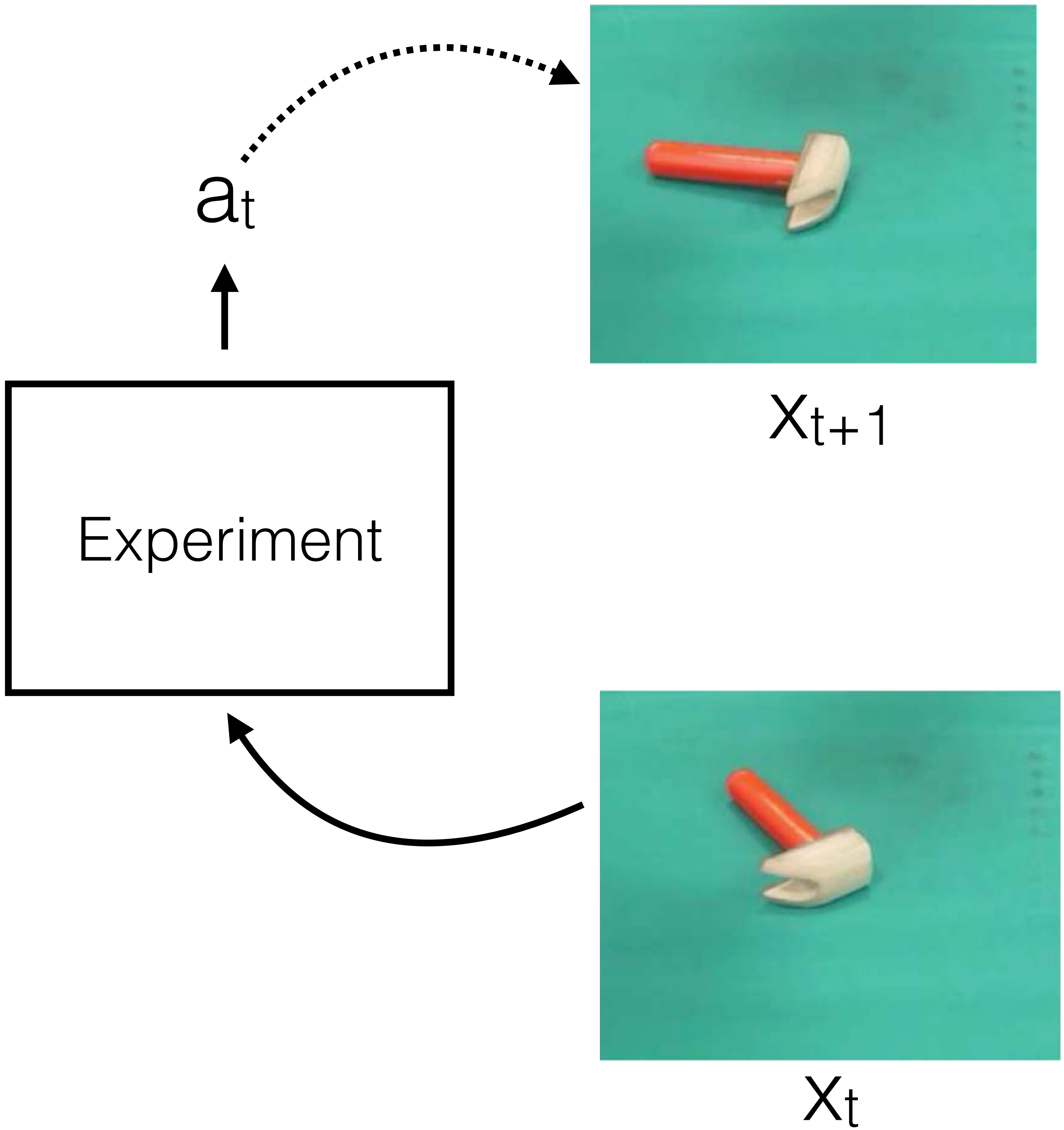
Lets not wait to find rewards

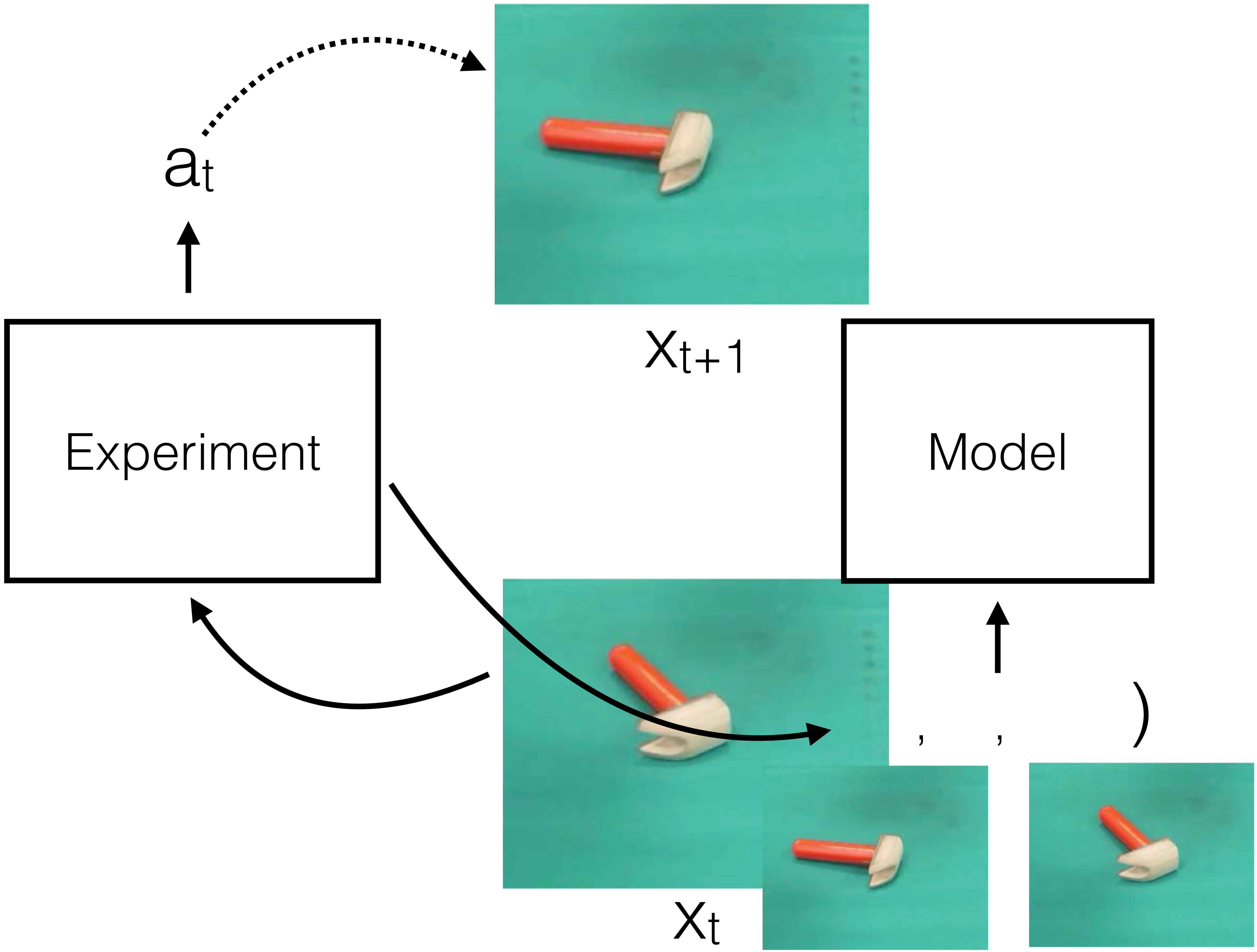
Learn skills in anticipation of future tasks!



Robots
Exploring




a_t
Experiment
X_{t+1}
X_t



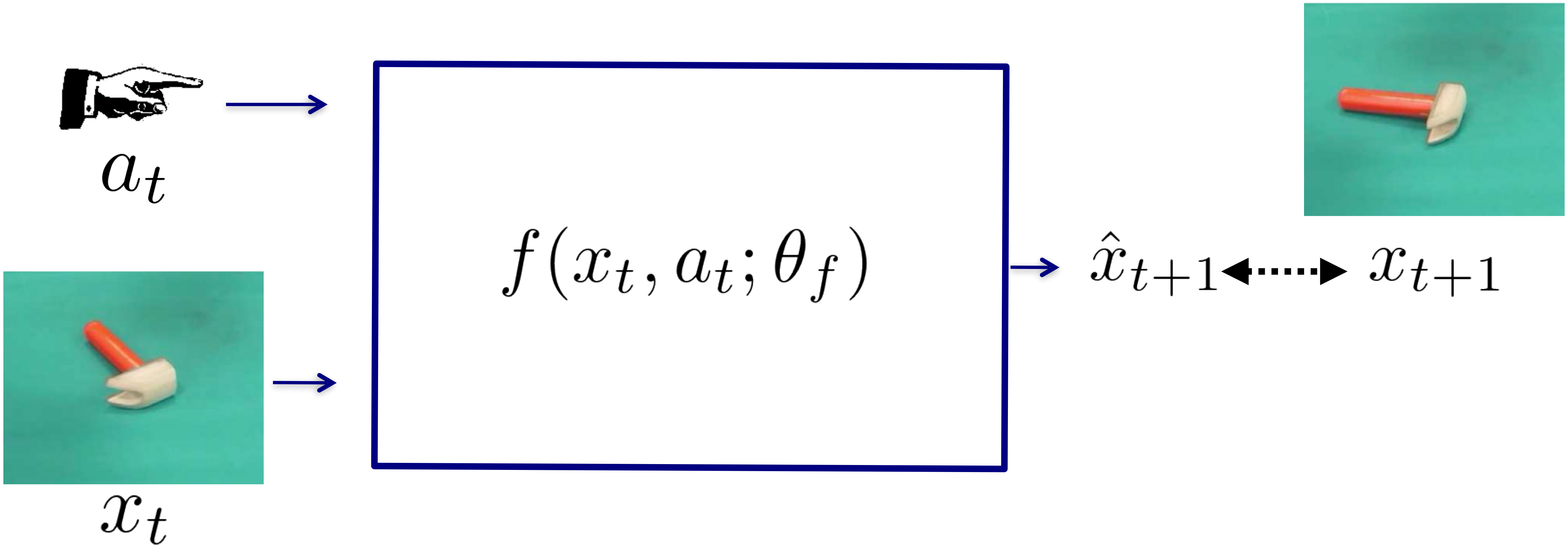
a_t
x_{t+1}
Experiment
Model
x_t

Useful Model: Predict what will happen next

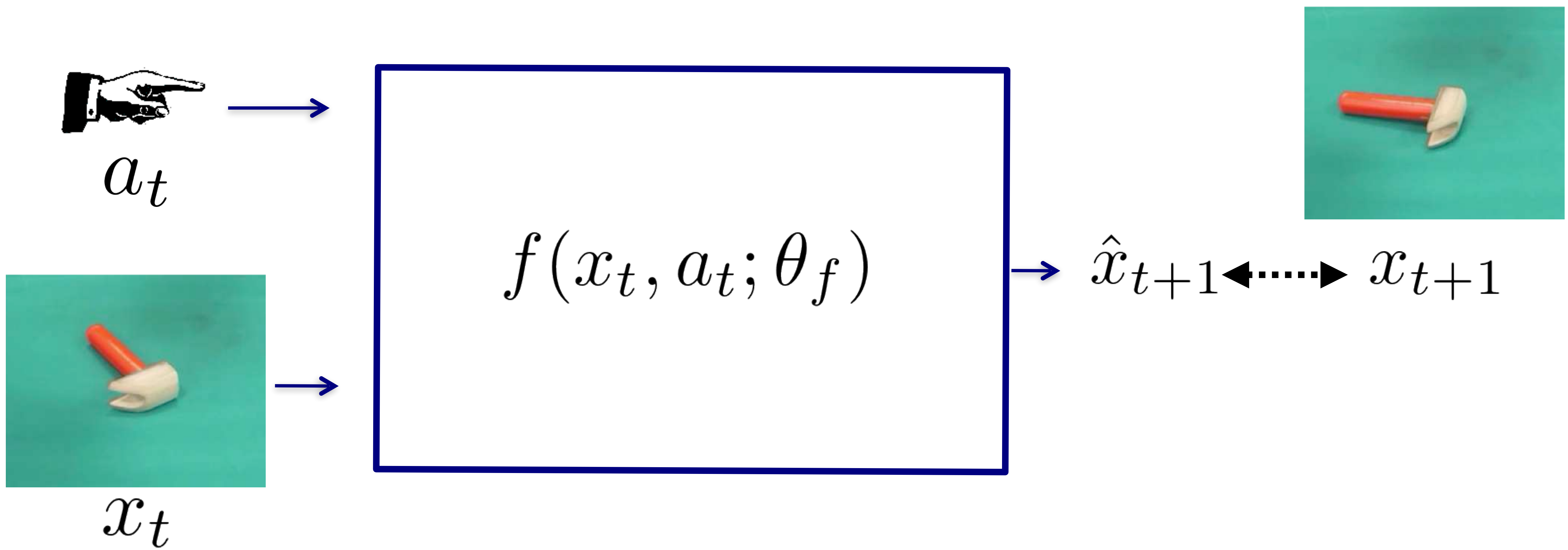
a_t

x_t

Useful Model: Predict what will happen next



Useful Model: Predict what will happen next



Forward model in pixel space

Petrovic et al., 2006

Goodfellow et al., 2014

Ranzato et al., 2014

Oh et al., 2015

Mathieu et al., 2015

Vondrick et al., 2015

Xue et al., 2016

Vondrick et al., 2016

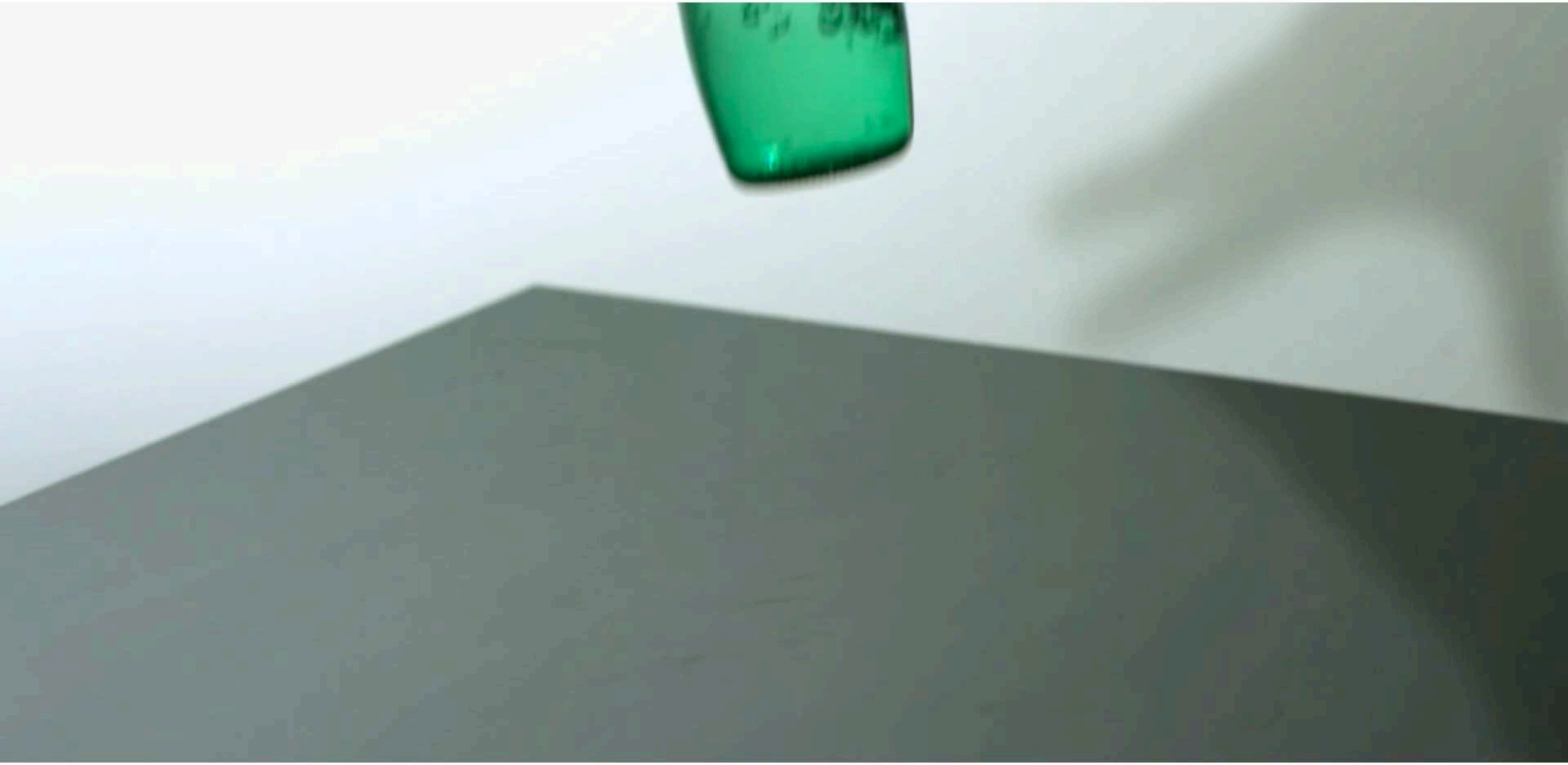
Finn et al., 2017

Not only hard,

but is this the right model to build ???

Consider a glass bottle

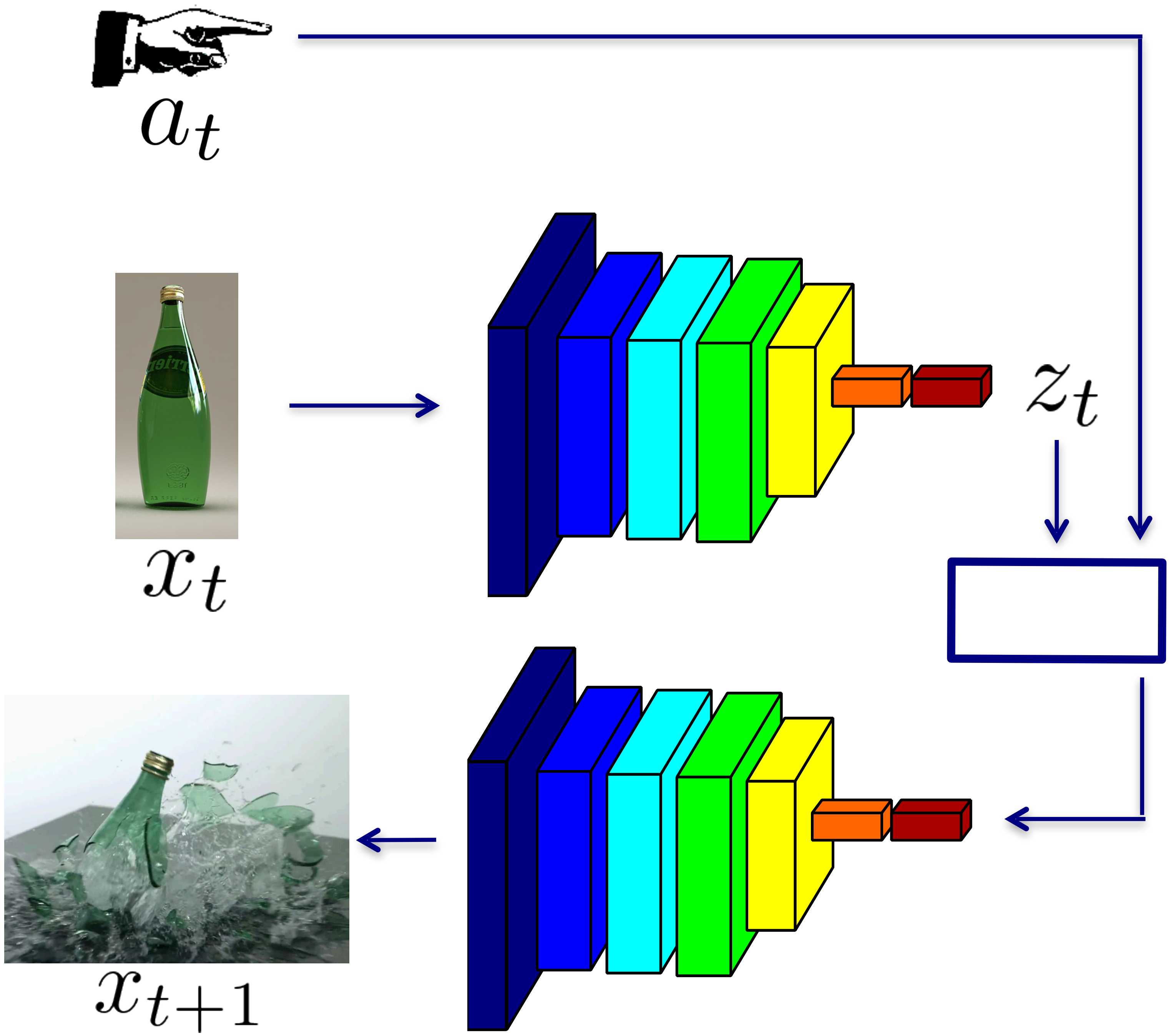
What will happen on dropping the bottle?



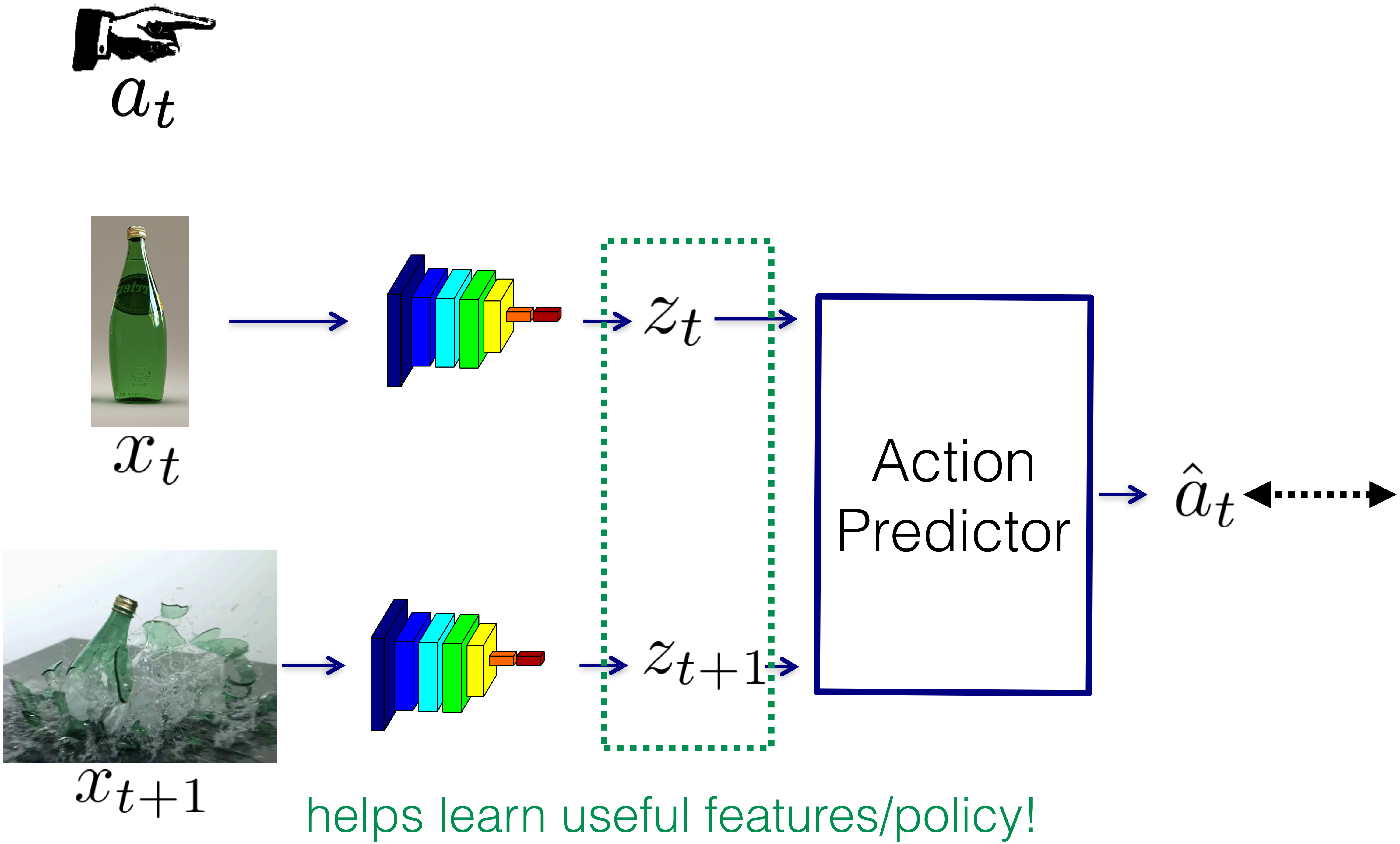
Different Feature Abstractions afford Different Predictions

Easy to predict: bottle breaks
but
Hard to predict: exact location of glass pieces

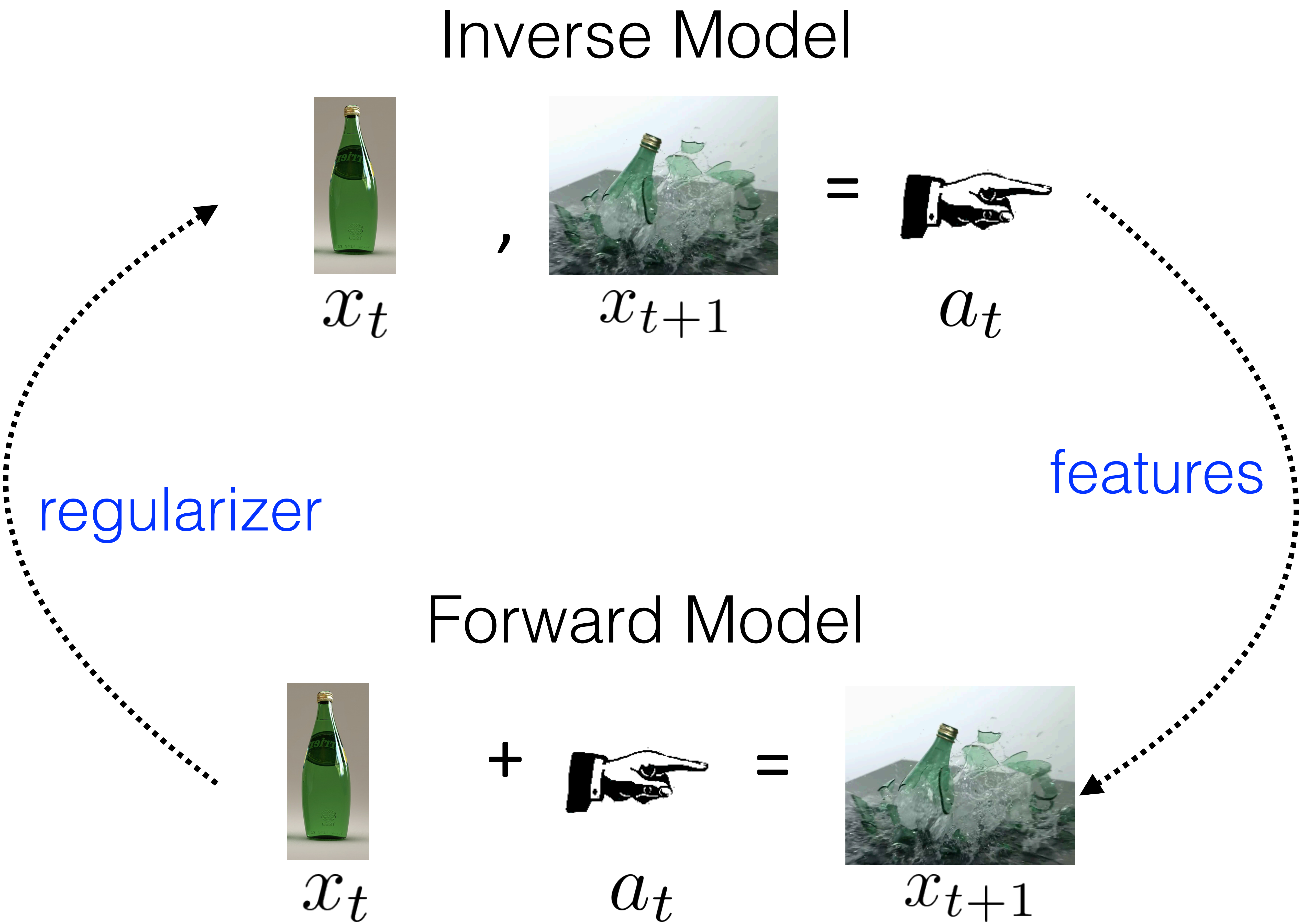
Instead of predicting pixels,

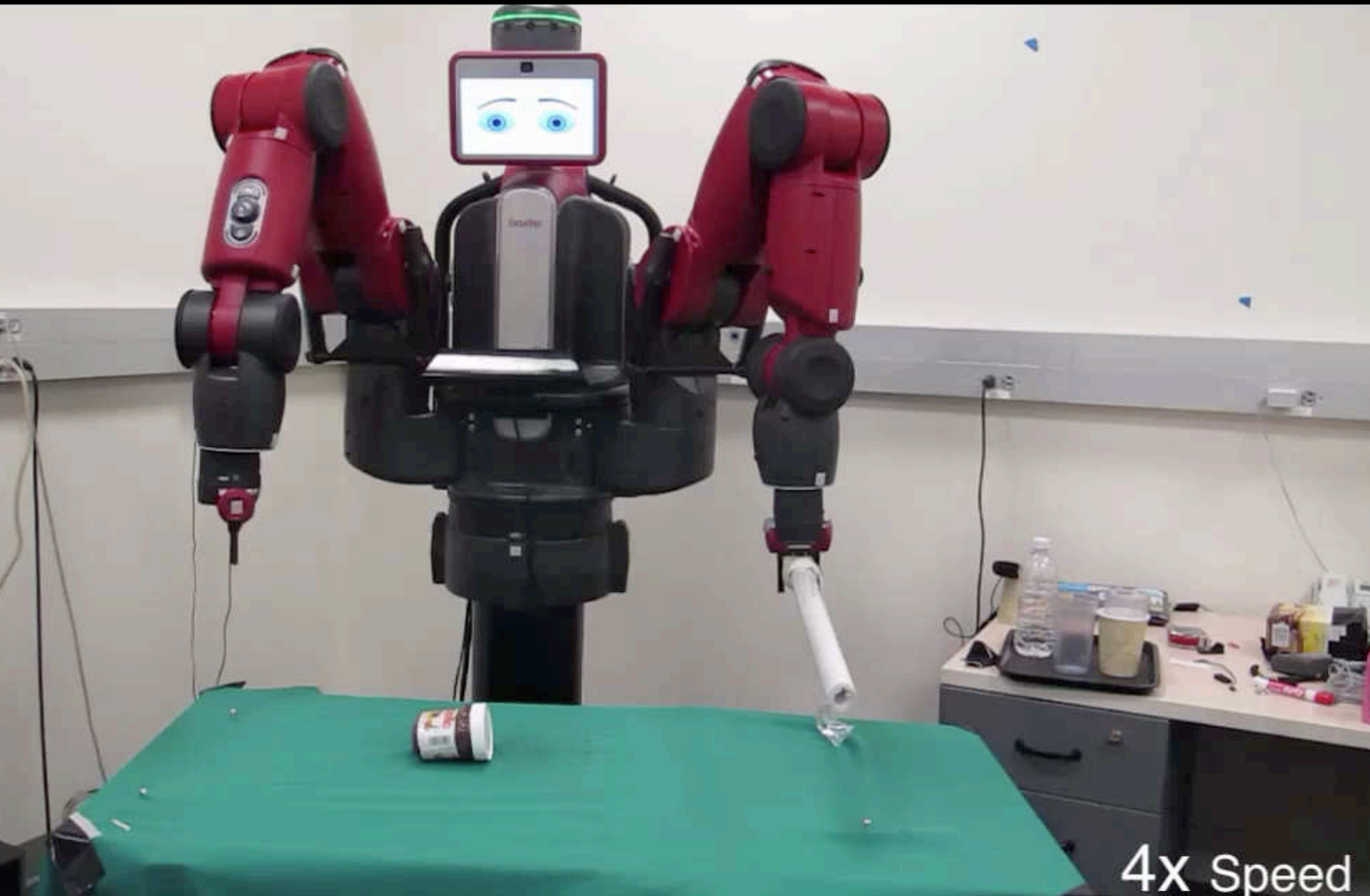


How about a different task?

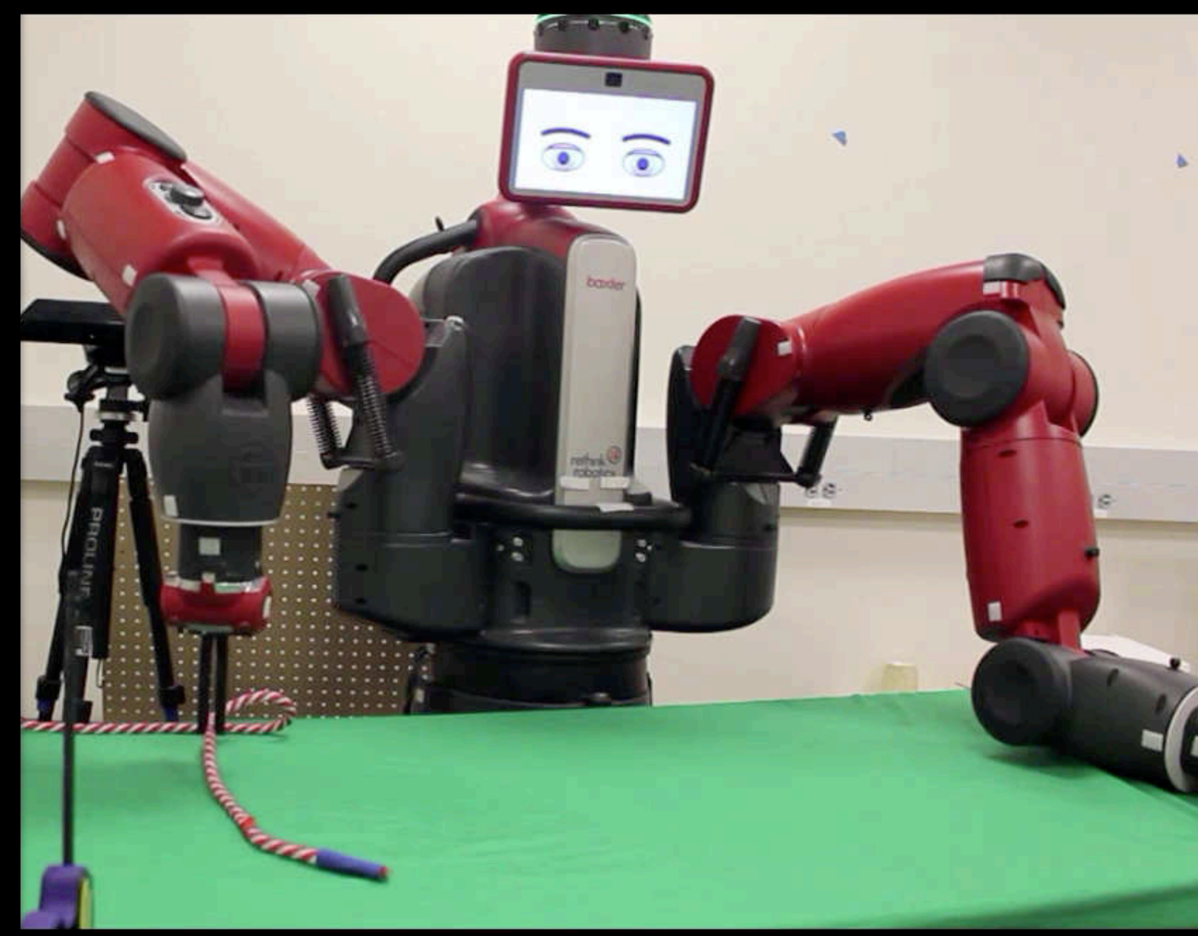


Inverse Model


Inverse Model
x_t , x_{t+1} = a_t
features
regularizer
Forward Model
x_t + a_t = x_{t+1}

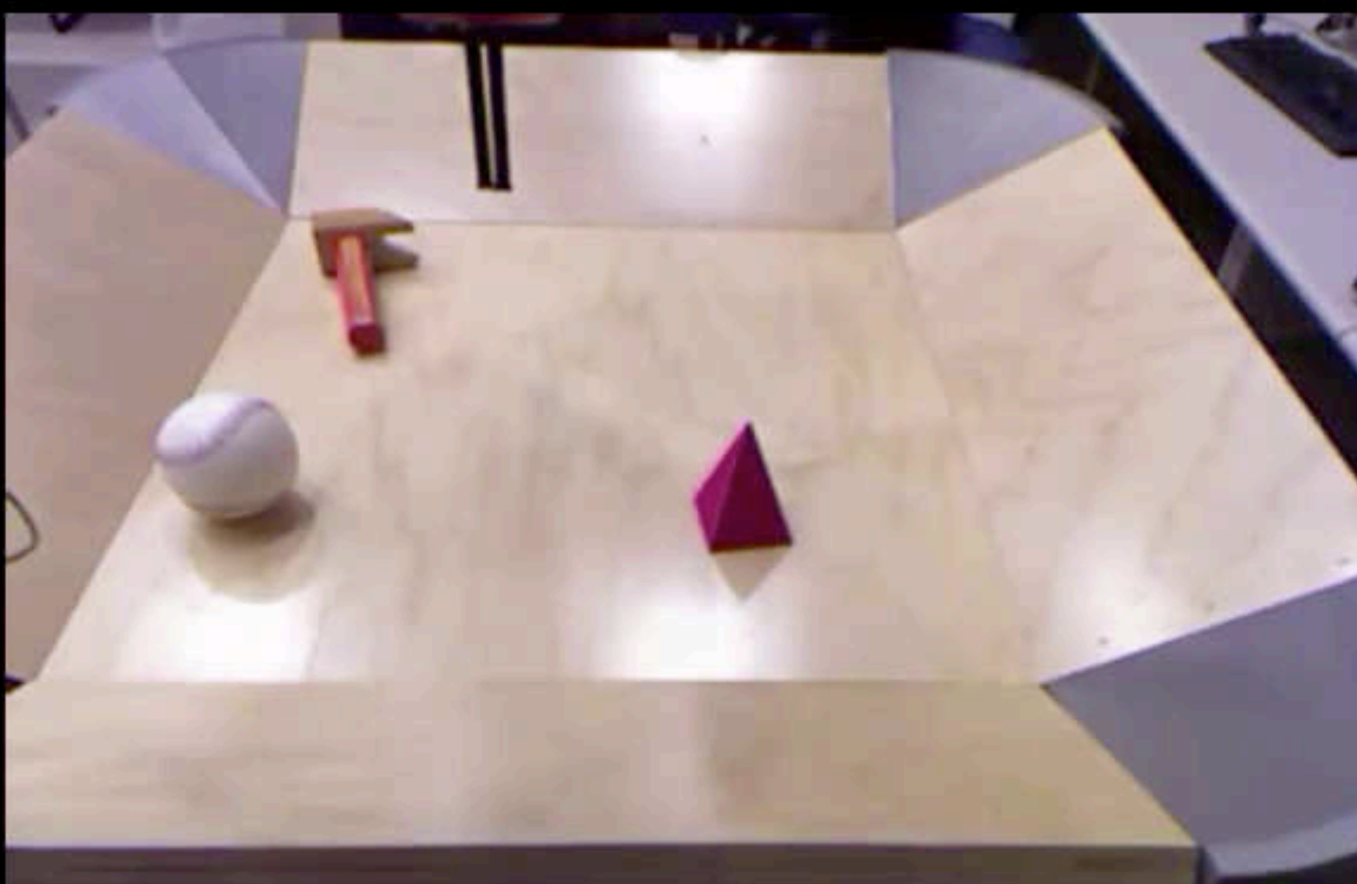


Robots Exploring

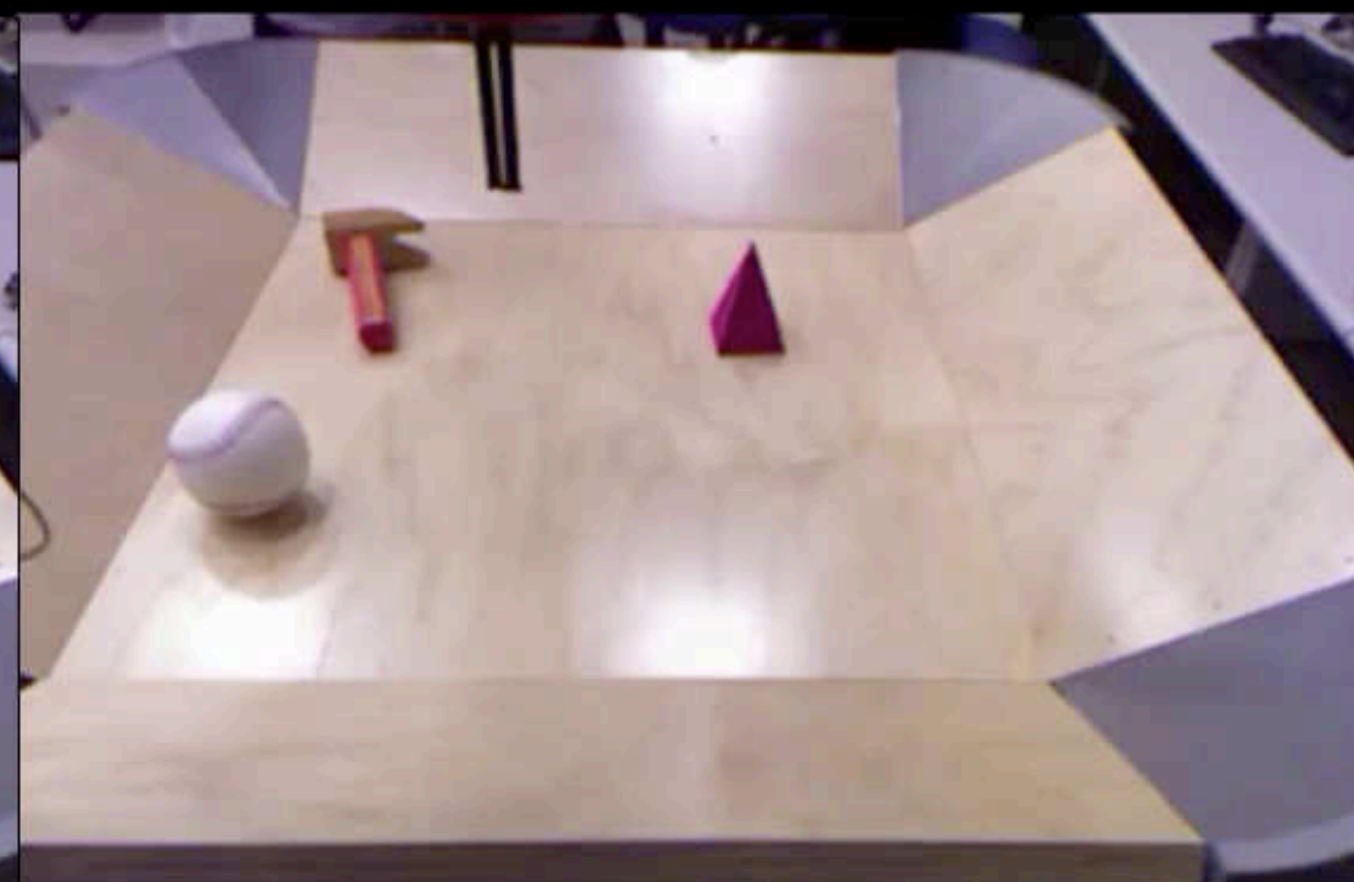


Pushing Objects

Current State



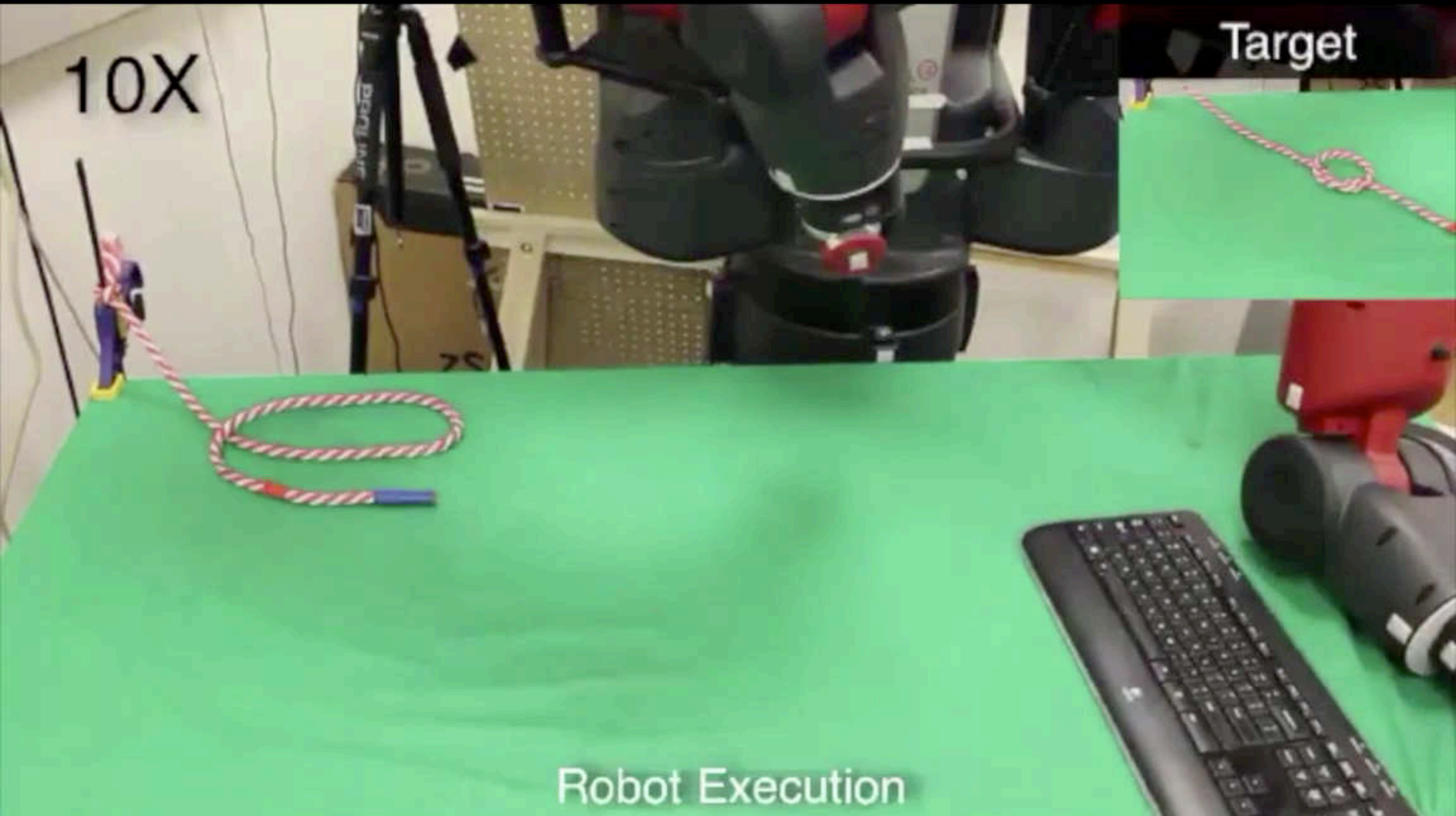
Goal State



Robot did not see any pyramids during training

Learning to Poke by Poking: Experiential Learning of Intuitive Physics, Agrawal et al., NIPS 2016

Rope Manipulation



robot gets only RGB images as input!

Combining Self-Supervision and Imitation for Vision Based Rope Manipulation, Ashvin Nair*, Dian Chen*, **Pulkit Agrawal***, Phillip Isola, Pieter Abbeel , Jitendra Malik, Sergey Levine, ICRA 2017 (*equal contribution)

Robot's Emergent Behavior

Current Image

Goal Image

Zero Shot Visual Imitation, Pathak D.*, Mahmoudieh P*., Luo M.*, **Agrawal P. *,**
Shentu Y., Chen D., Shelhamer E., Malik J., Darrell, T. (ICLR 2018, *equal contribution)

What experiment to run?
(exploration policy)

Model of how things work
(intuitive physics, behavior)

Issues with Reinforcement Learning

Lots of data

Where do rewards come from?

Task Specific

Demonstrations

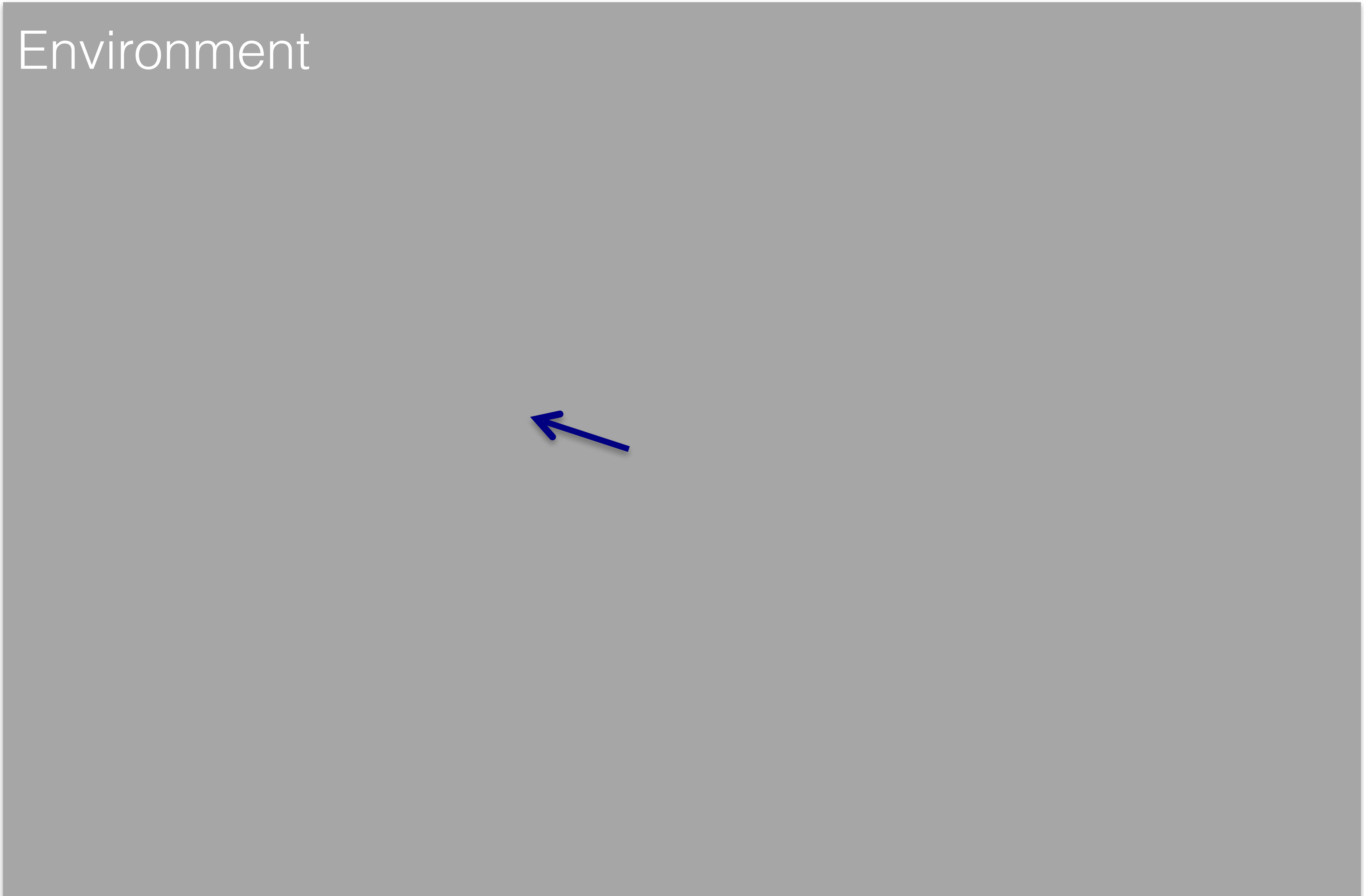
Task Curriculum

Self-Supervised Model Learning

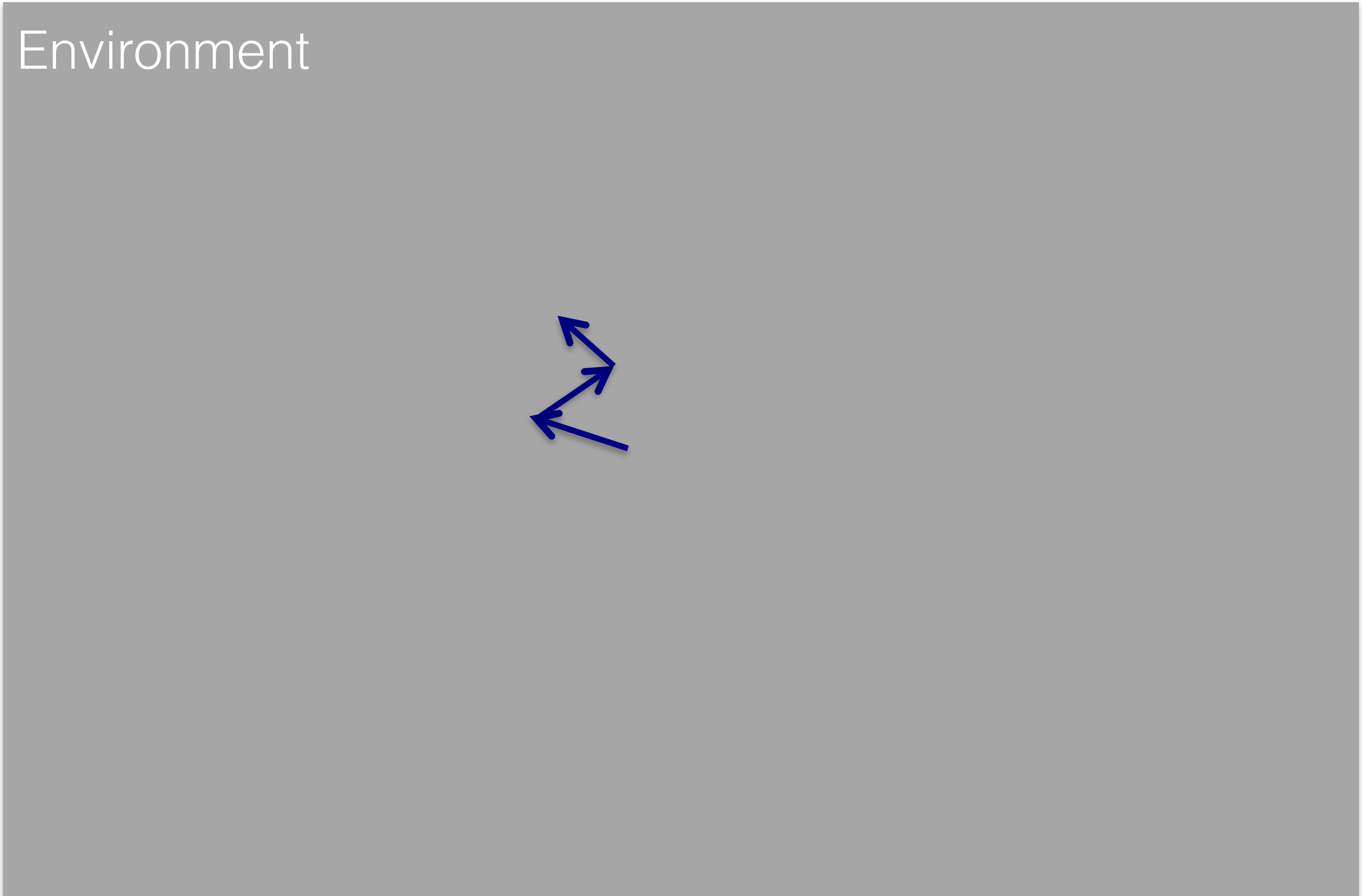
What experiment to run?
(exploration policy)

Model of how things work
(intuitive physics, behavior)

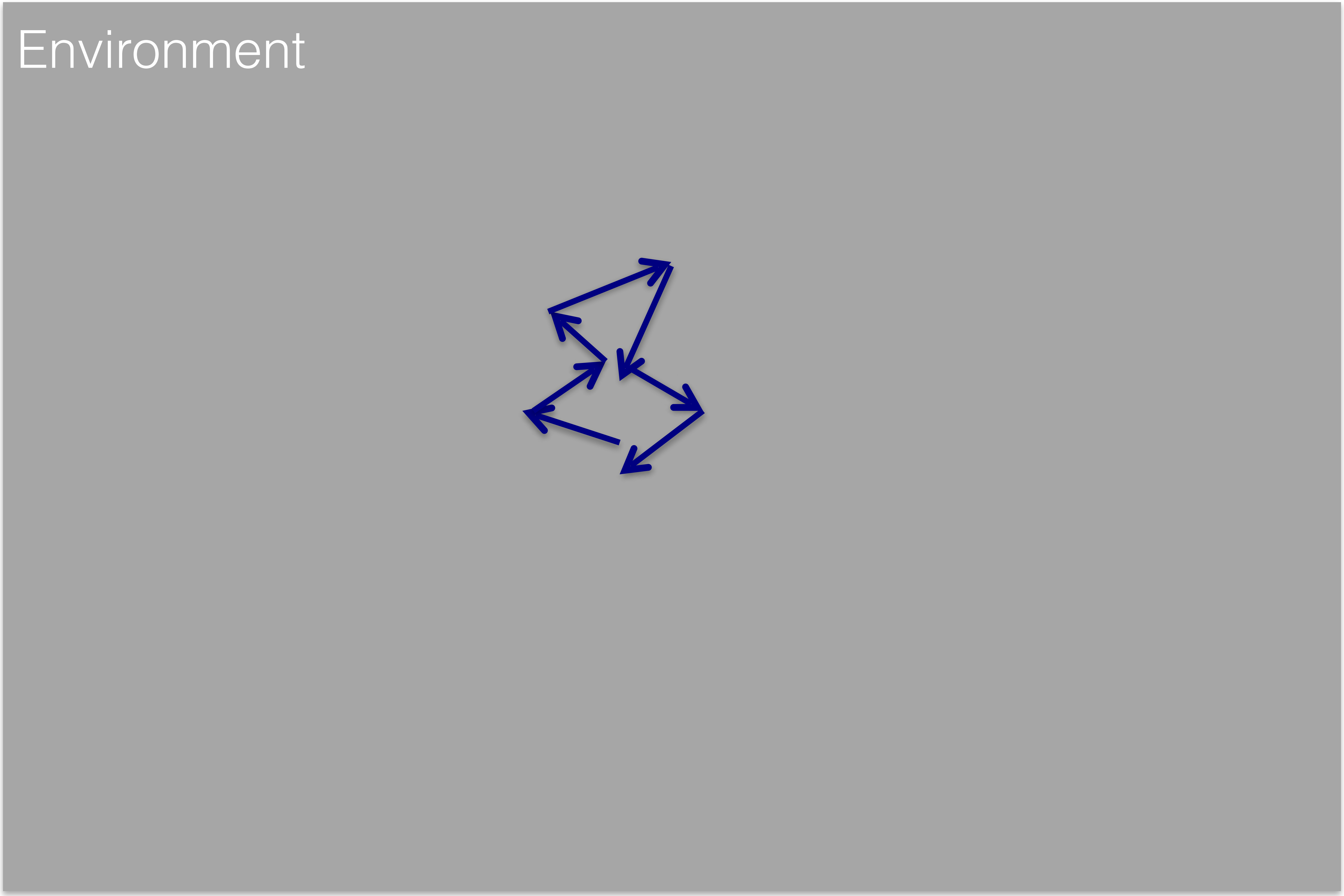
Random Exploration is Limiting



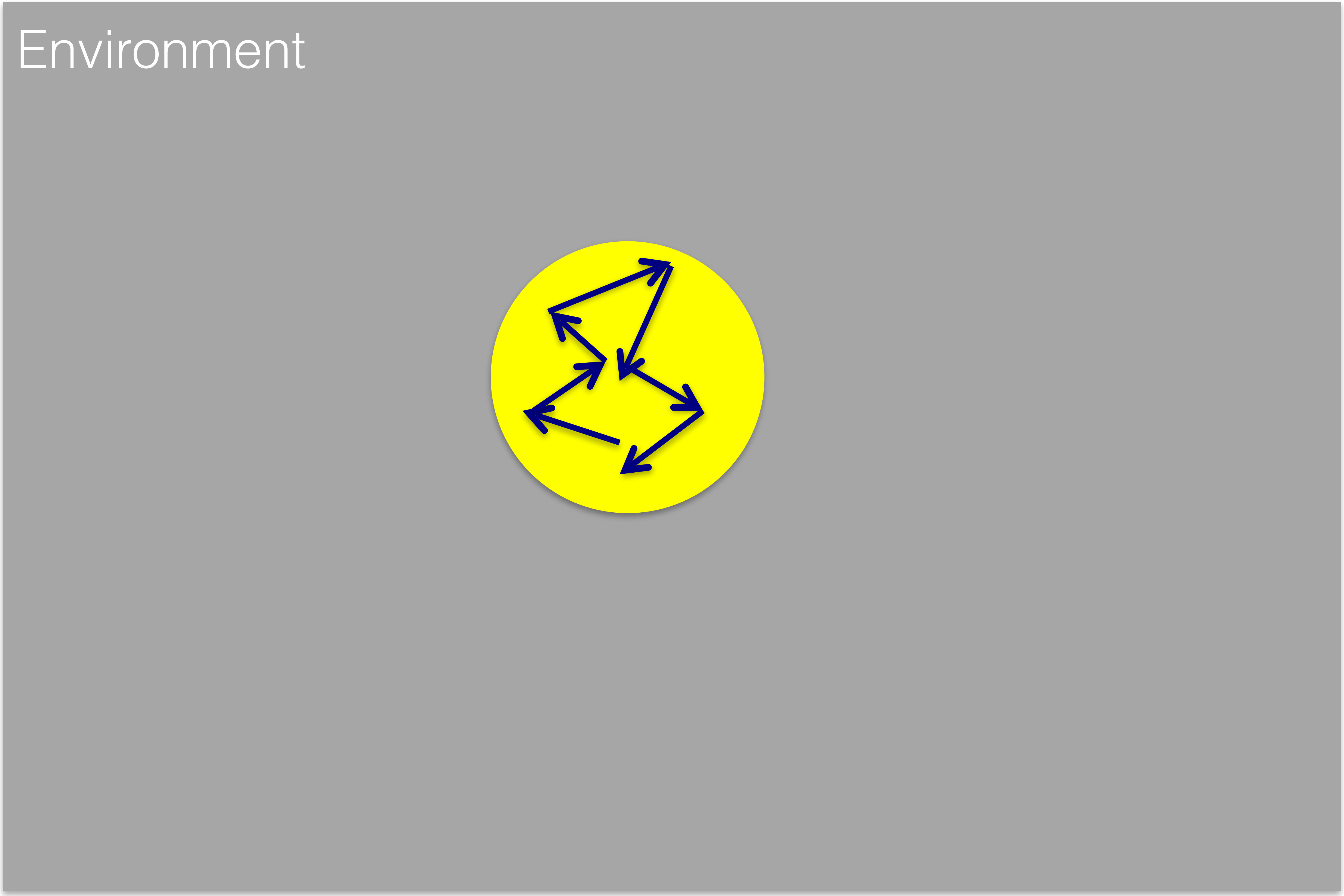
Random Exploration is Limiting



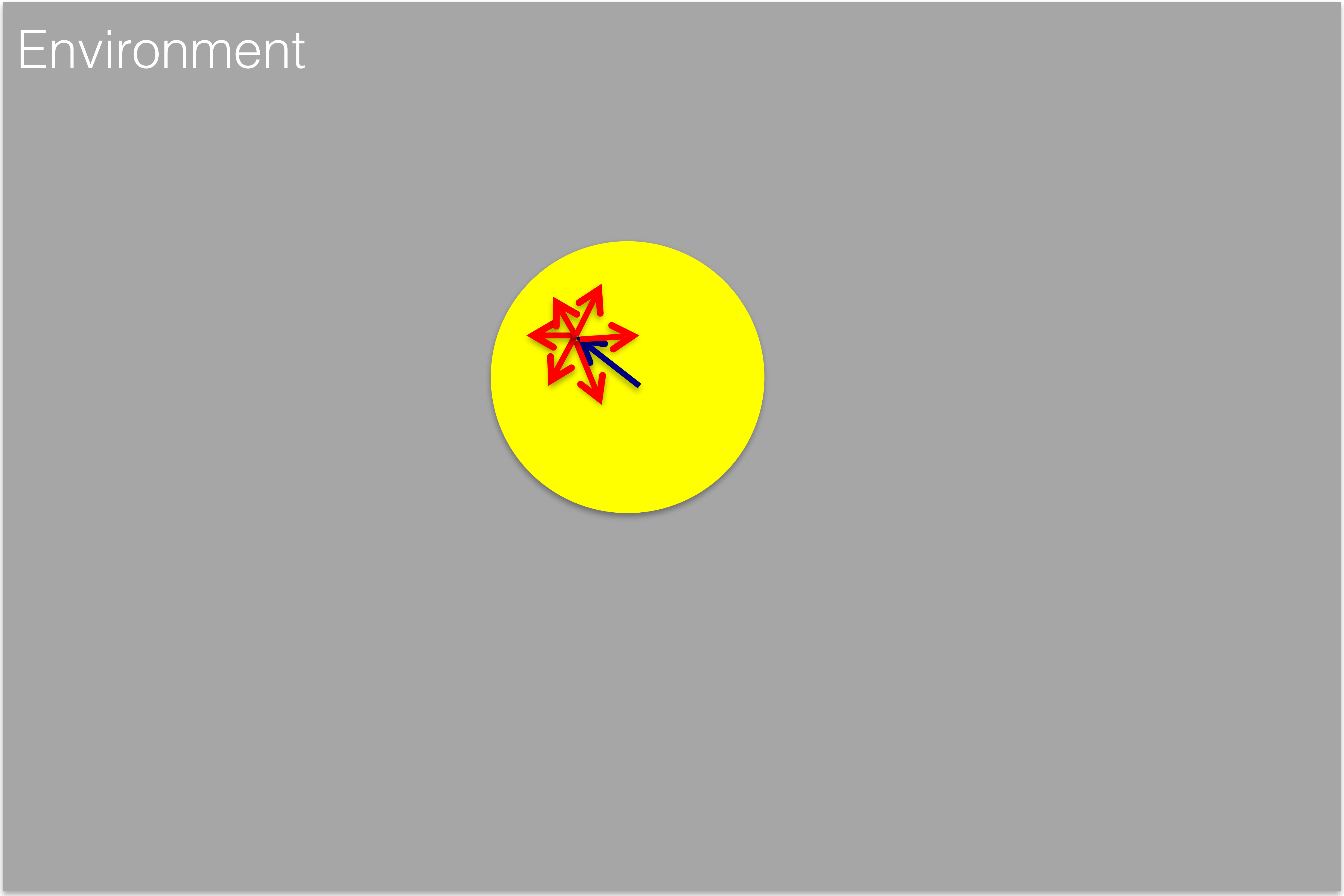
Random Exploration is Limiting



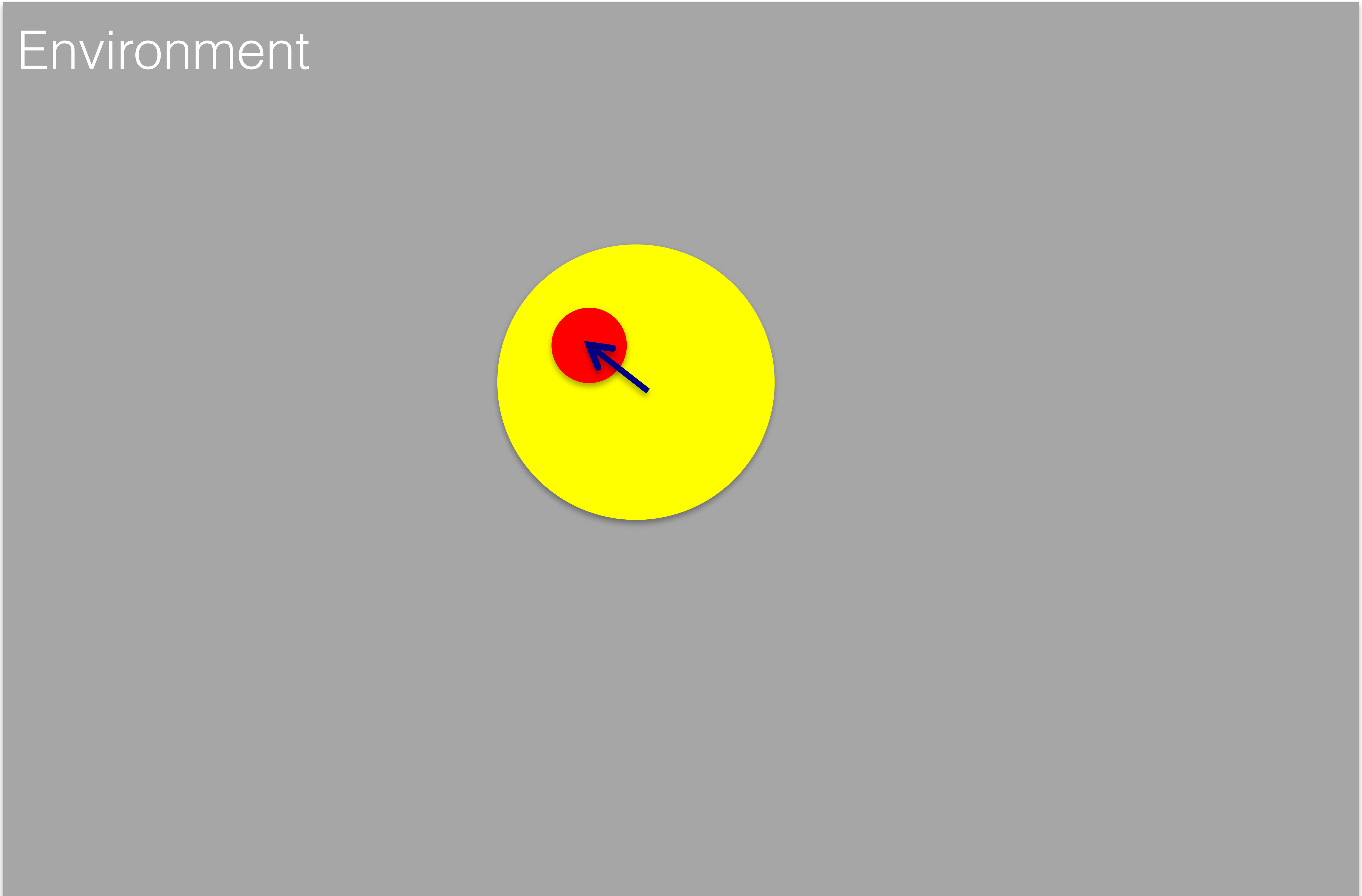
Random Exploration is Limiting



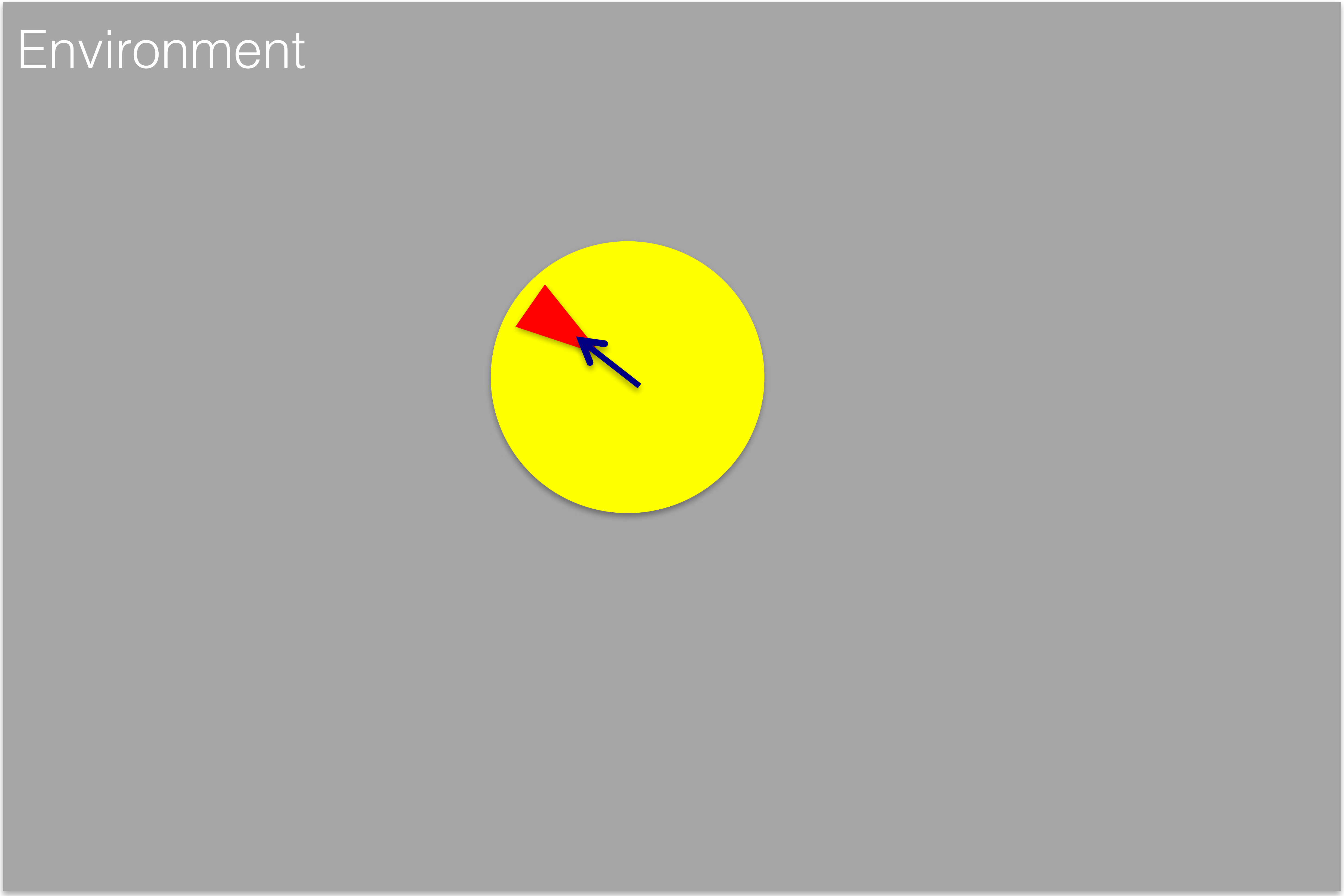
Random Exploration is Limiting



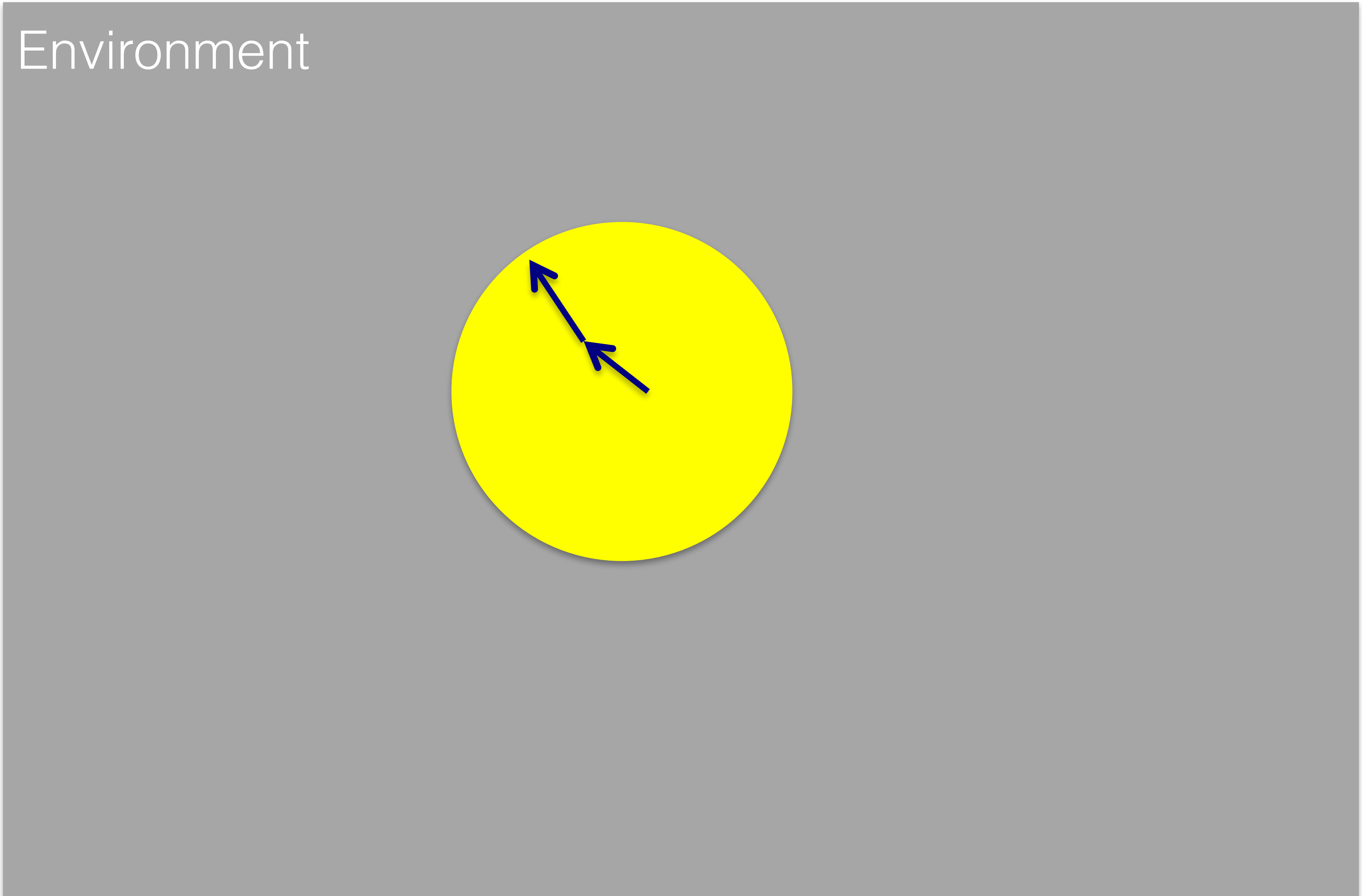
Random Exploration is Limiting



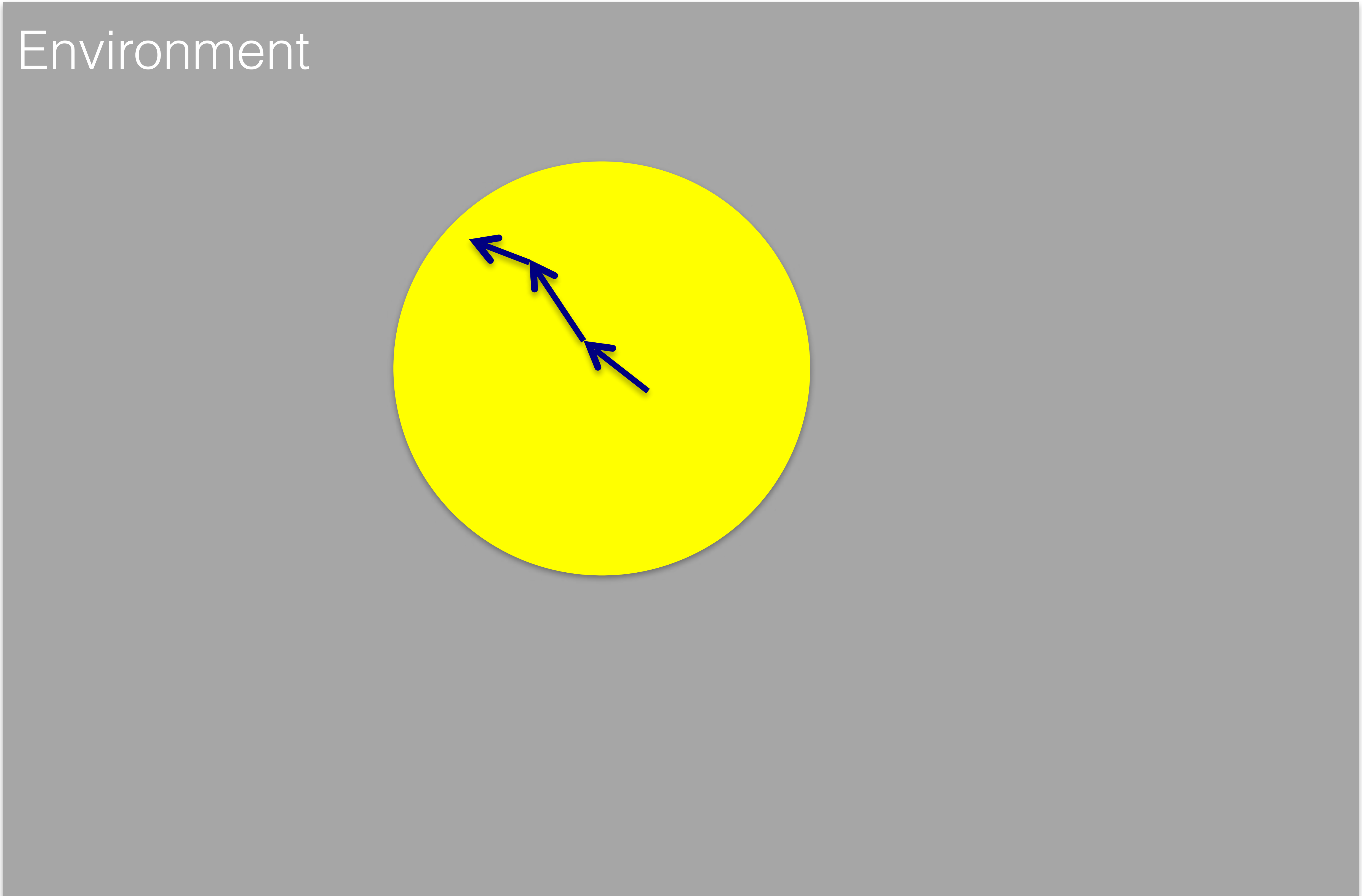
Novelty Seeking Exploration



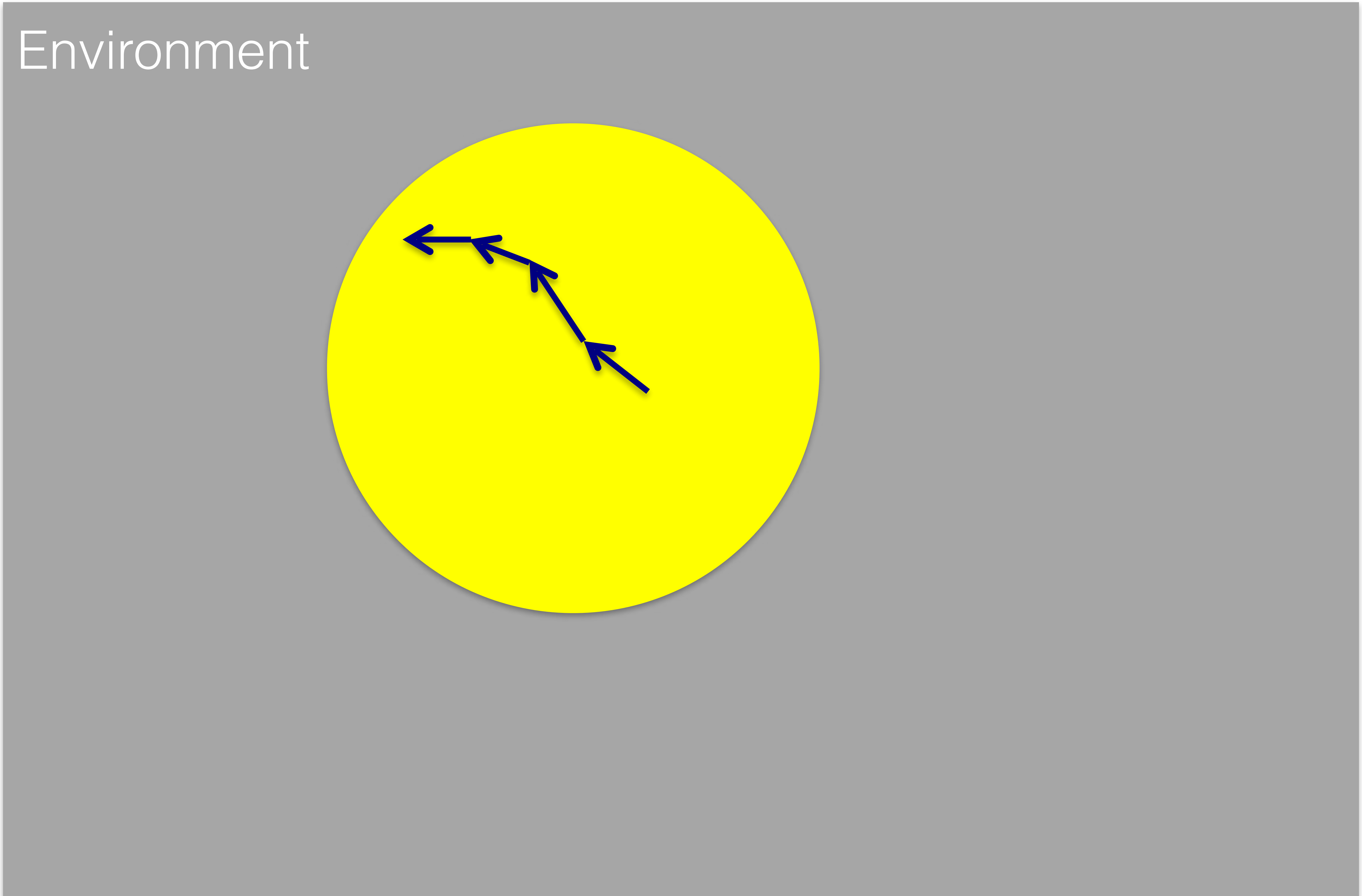
Novelty Seeking Exploration

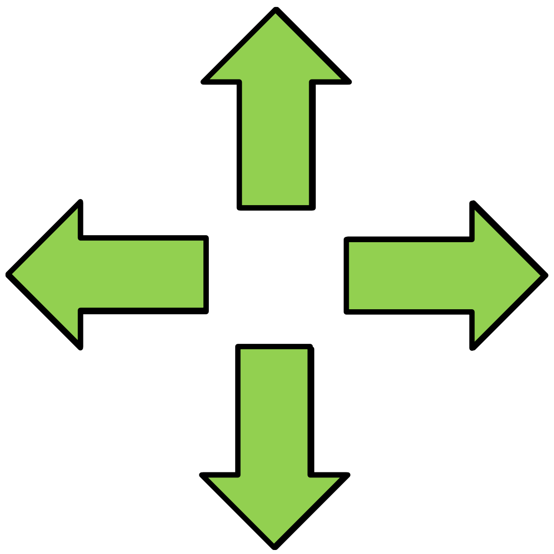


Novelty Seeking Exploration



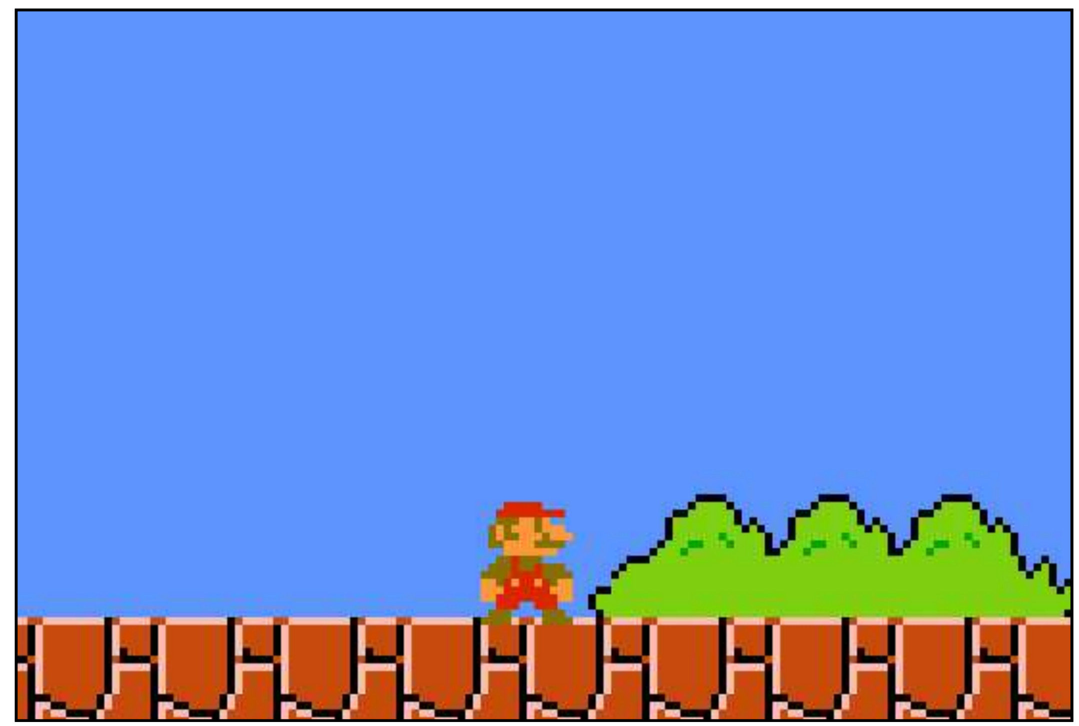
Novelty Seeking Exploration





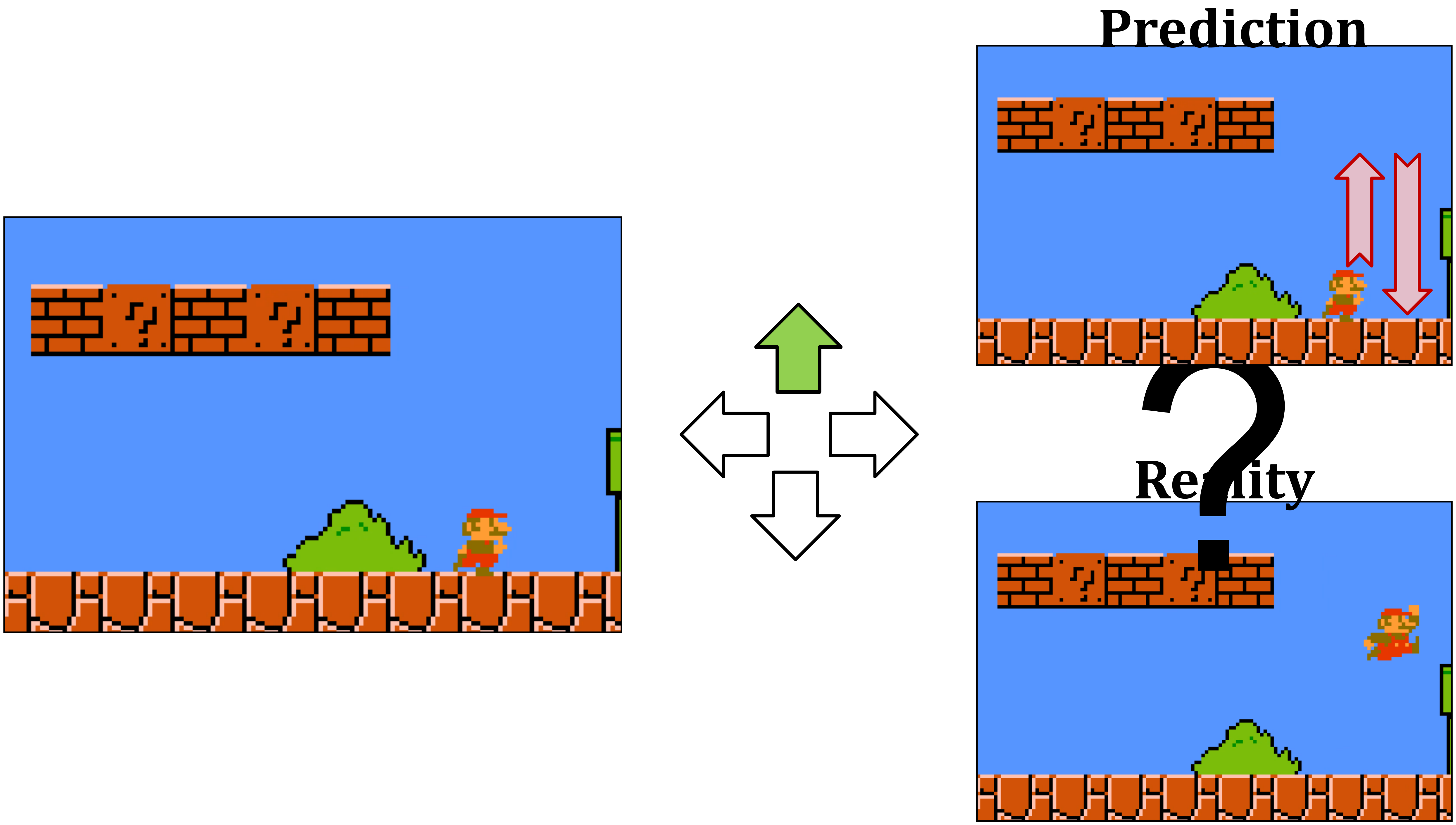
“**Down**” has no effect

Action

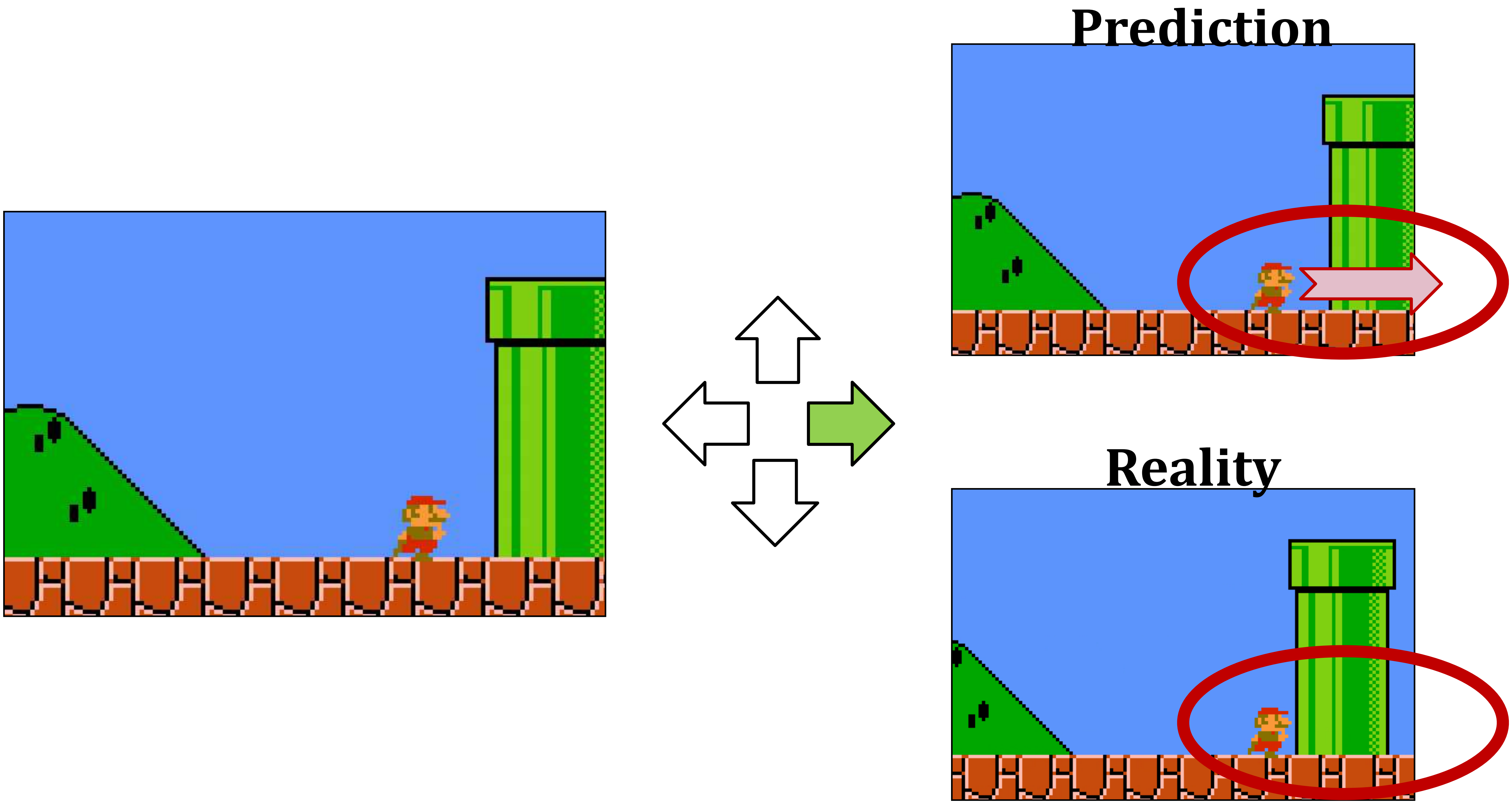


Observation

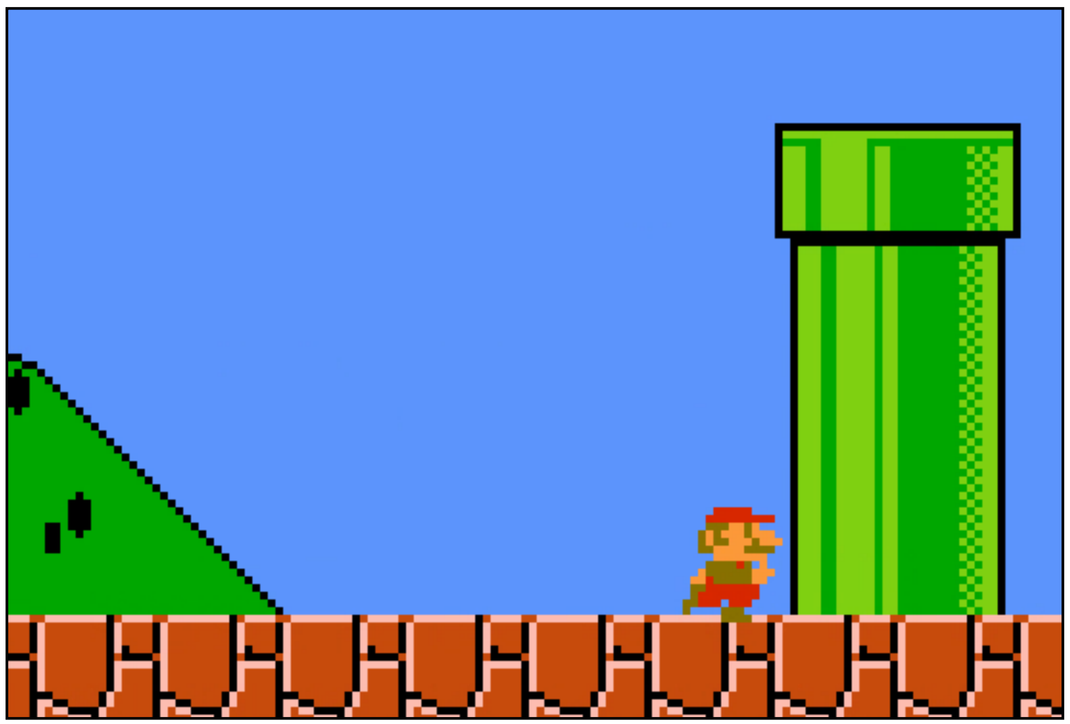
Curiosity driven Exploration by Self-Supervised Prediction
Pathak D., **Agrawal P.**, Efros A., Darrell T. , ICML 2017

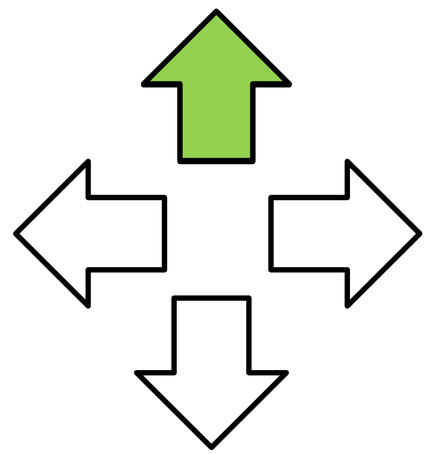


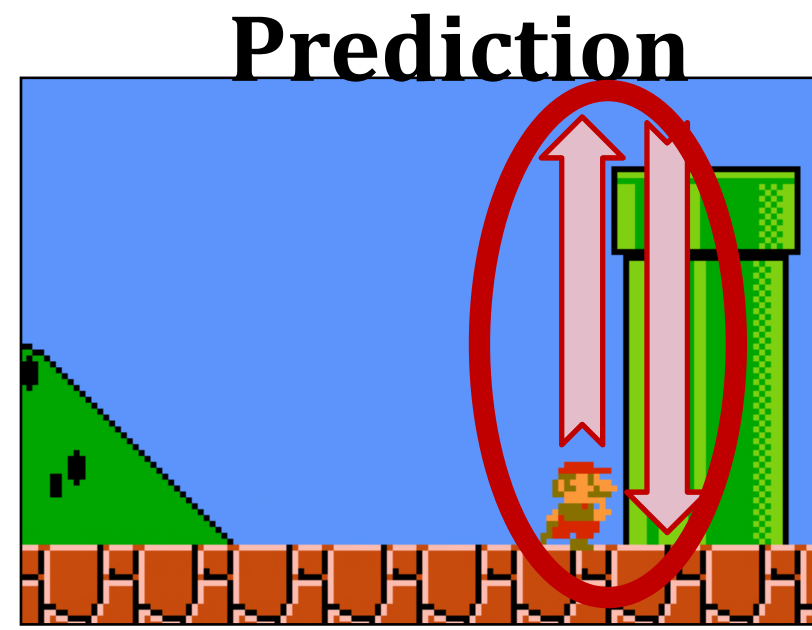
Curiosity driven Exploration by Self-Supervised Prediction
Pathak D., **Agrawal P.**, Efros A., Darrell T. , ICML 2017

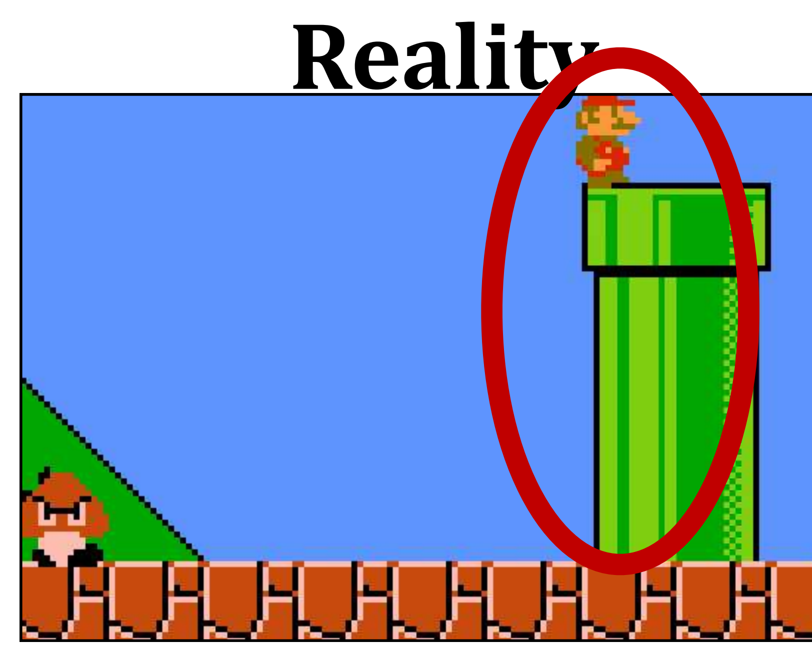


Curiosity driven Exploration by Self-Supervised Prediction
Pathak D., **Agrawal P.**, Efros A., Darrell T. , ICML 2017



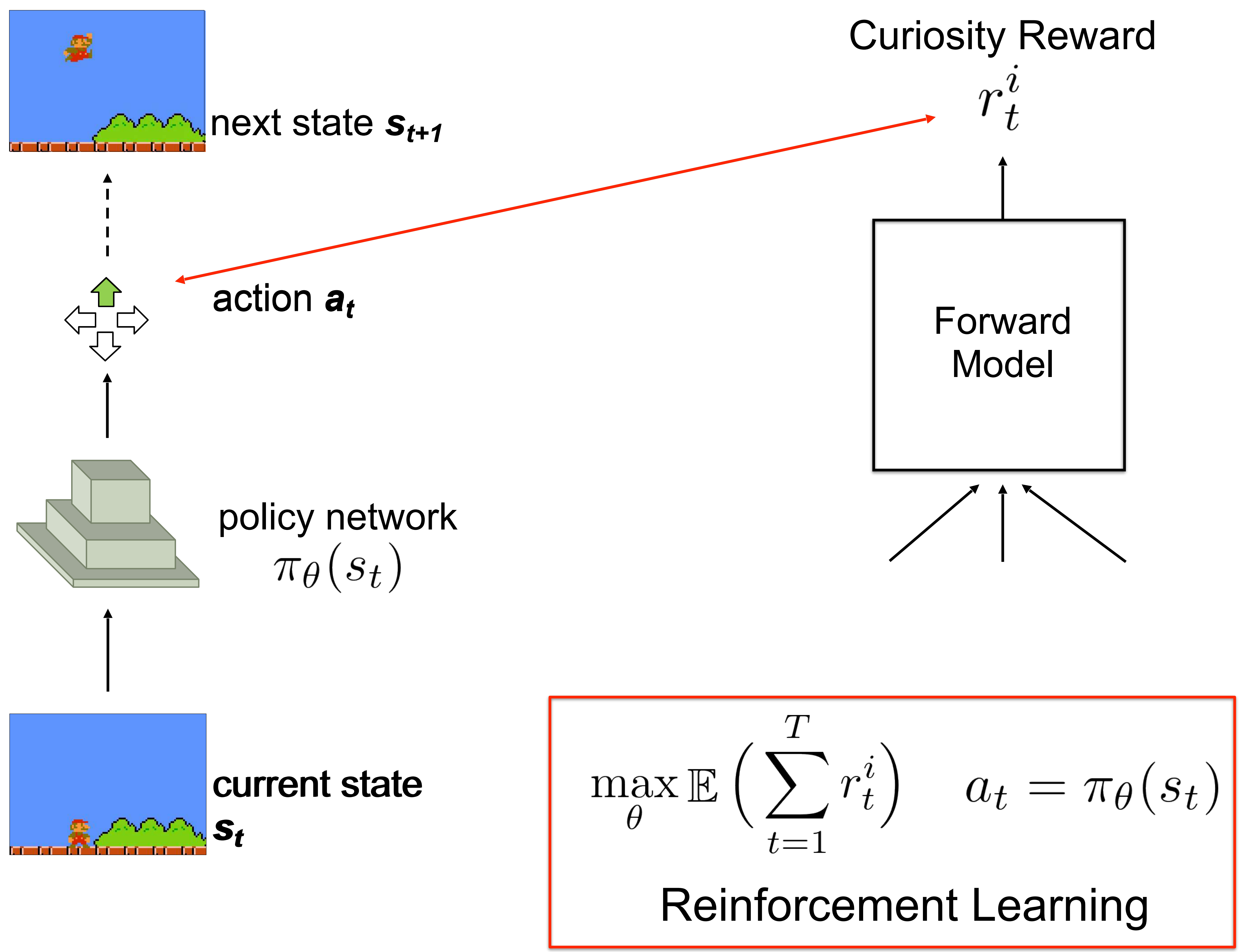






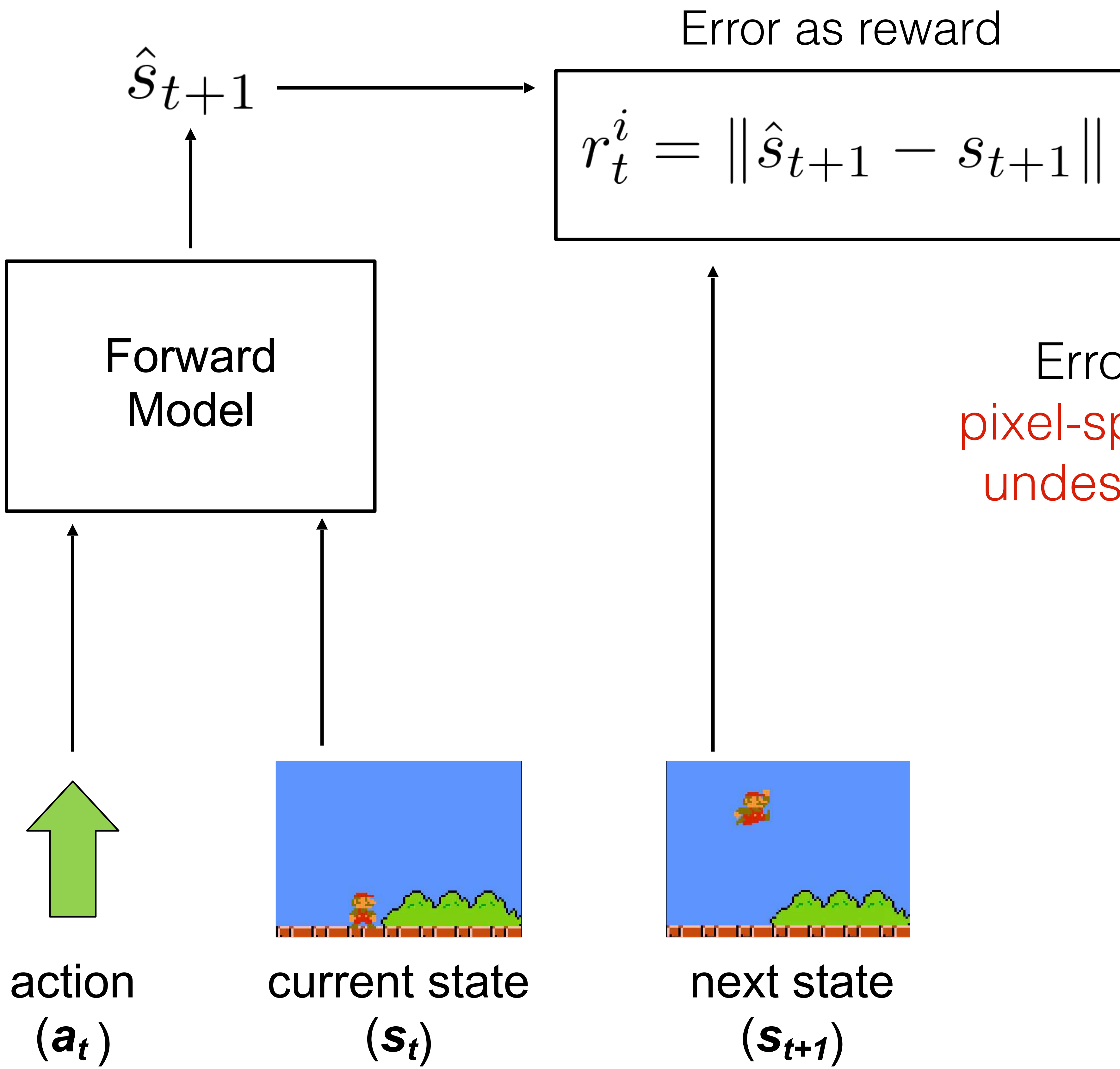
Curiosity $\triangleq$ Prediction Error

Curiosity driven Exploration by Self-Supervised Prediction
Pathak D., **Agrawal P.**, Efros A., Darrell T. , ICML 2017



Curiosity Reward
r_t^i
next state s_{t+1}
action a_t
Forward Model
policy network
$\pi_\theta(s_t)$
current state
s_t
$\max_\theta \mathbb{E}\left(\sum_{t=1}^{T} r_t^i\right) \quad a_t = \pi_\theta(s_t)$
Reinforcement Learning

Prediction Error Reward



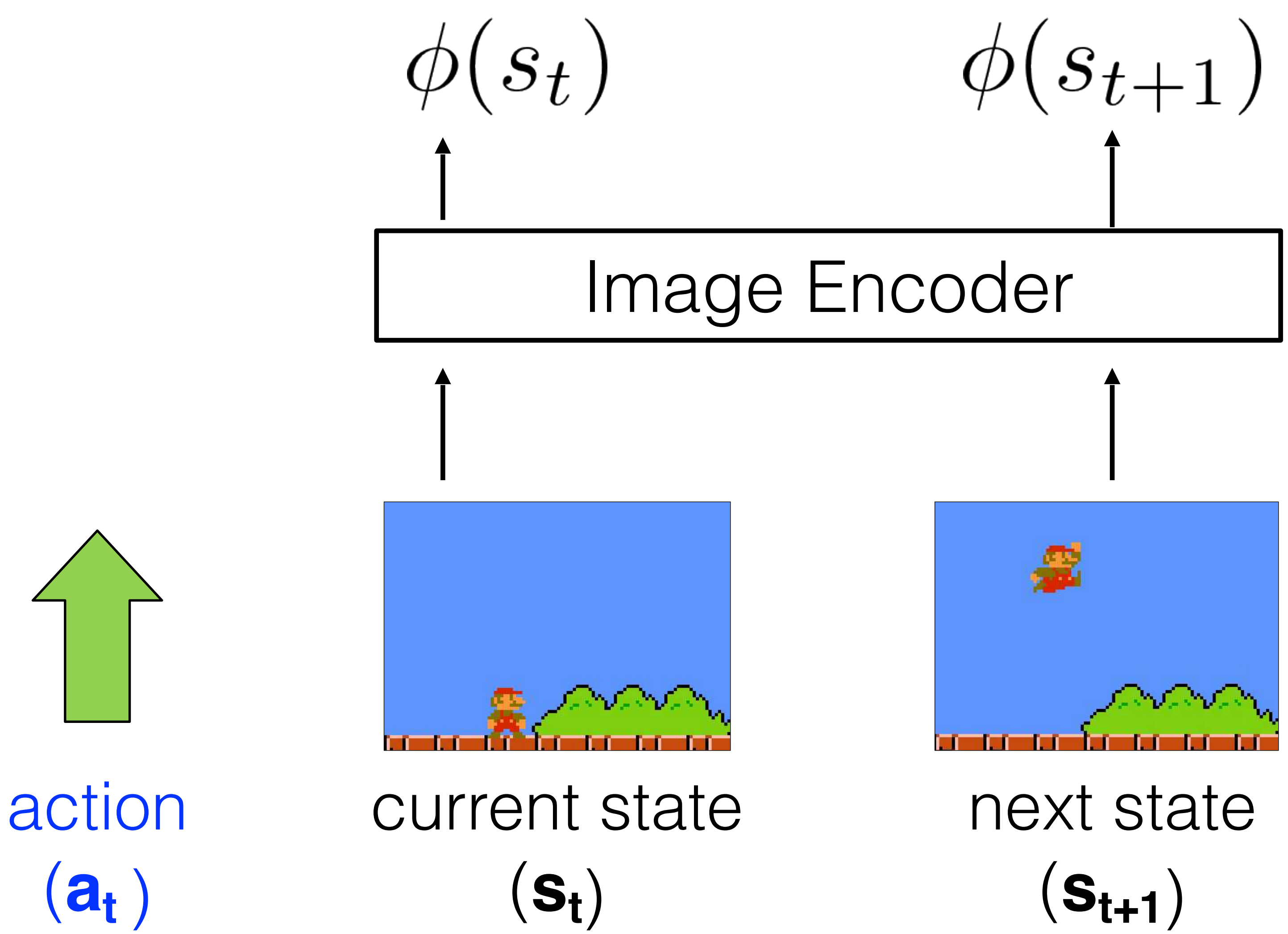
Error in pixel-space is undesirable

Only be curious about

things that can

affect the agent

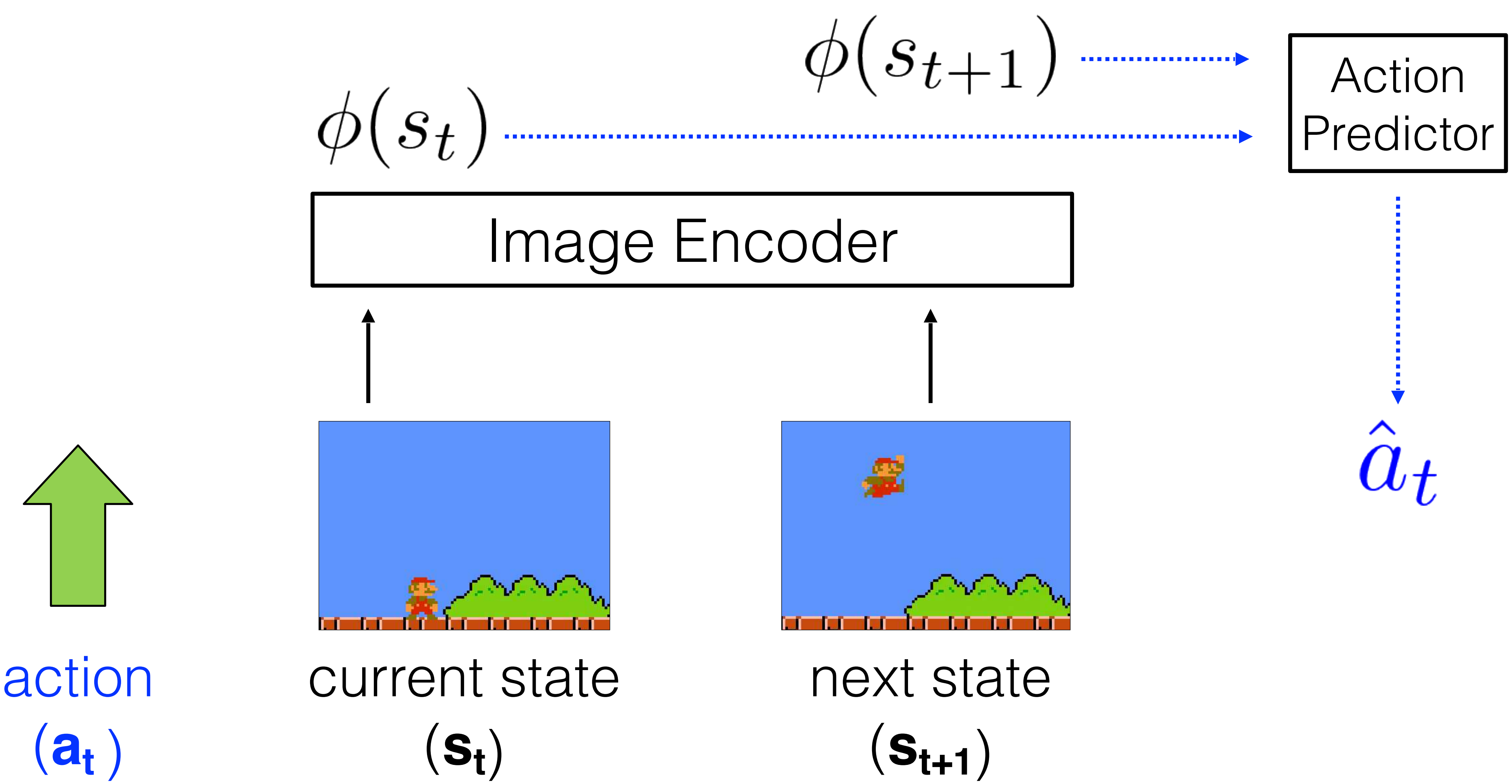
Prediction Error in Feature Space



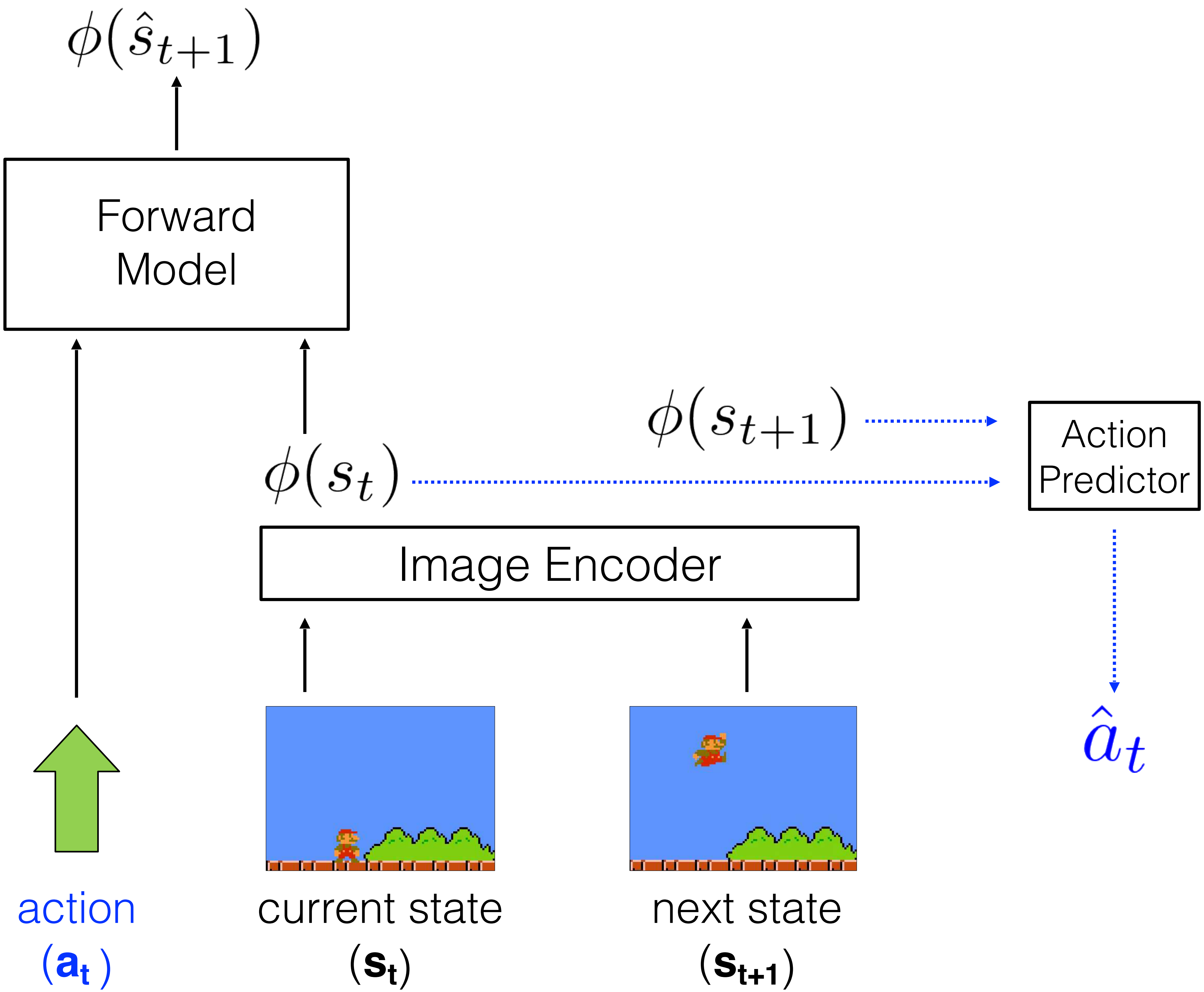
Prediction Error in Feature Space

Inverse Model

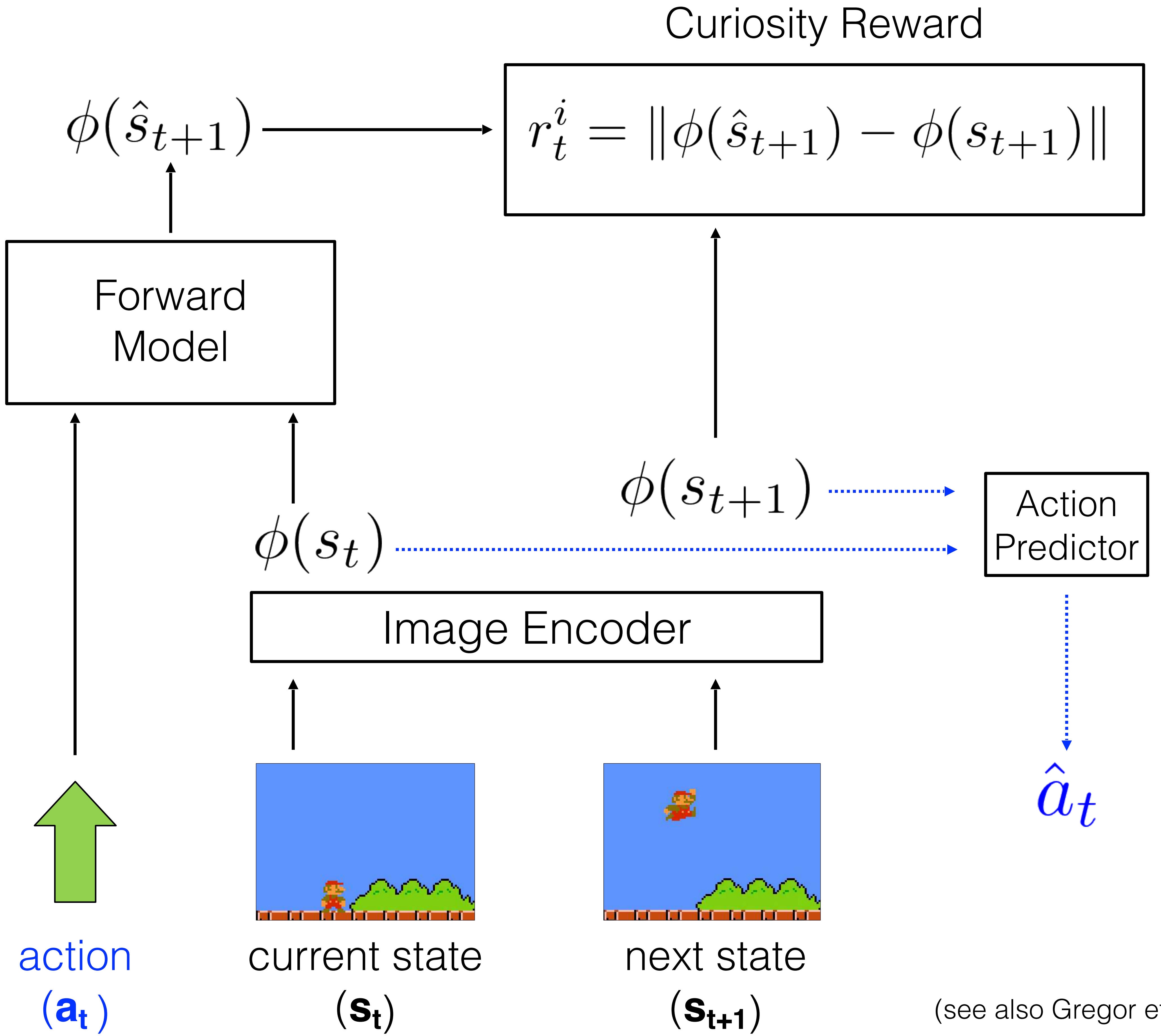
for learning feature representation



Prediction Error in Feature Space

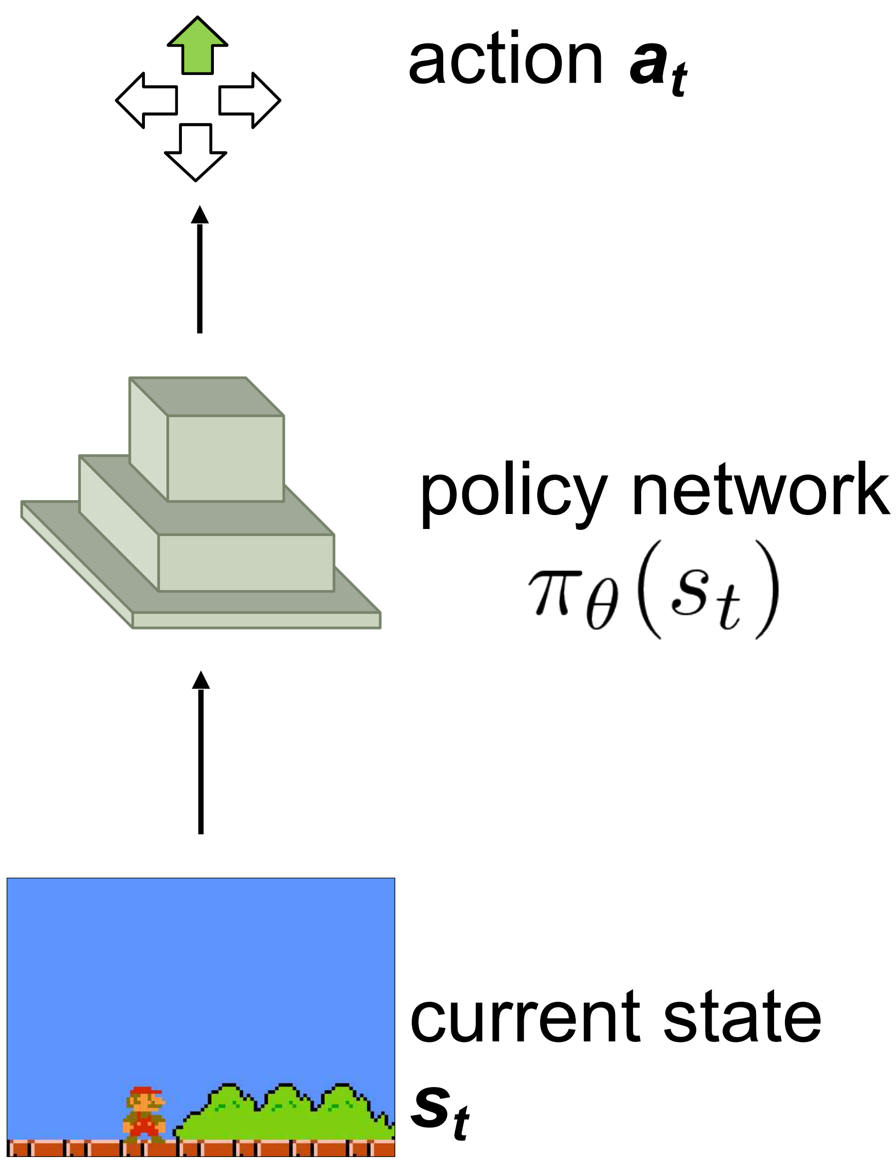


Prediction Error in Feature Space



(see also Gregor et al. 2017)

Is this a good exploration policy?



Testing Exploration on the game of Mario

Emergent behaviors:

- Jumping enemies, pipes and pits
- Killing enemies

Curiosity driven Exploration by Self-Supervised Prediction
Pathak D., Agrawal P., Efros A., Darrell T. , ICML 2017

Does the exploration generalize?

Trained on Level-1

Testing on Level-3

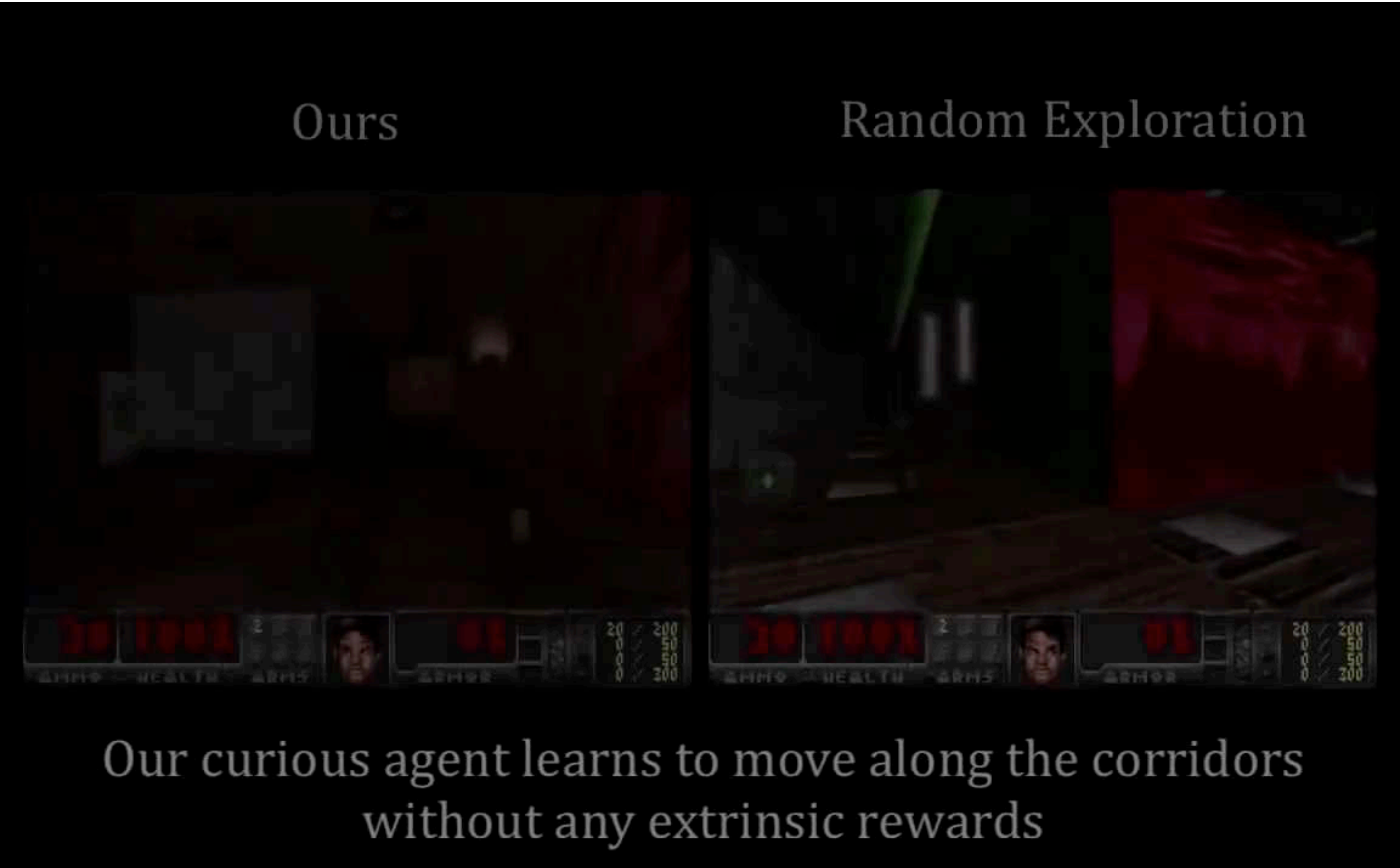
Curiosity driven Exploration by Self-Supervised Prediction
Pathak D., **Agrawal P.**, Efros A., Darrell T. , ICML 2017

Curious Agent in 3D Maze

Agent's Observation

Curiosity driven Exploration by Self-Supervised Prediction, Pathak D., **Agrawal P.**, Efros A., Darrell T. , ICML 2017

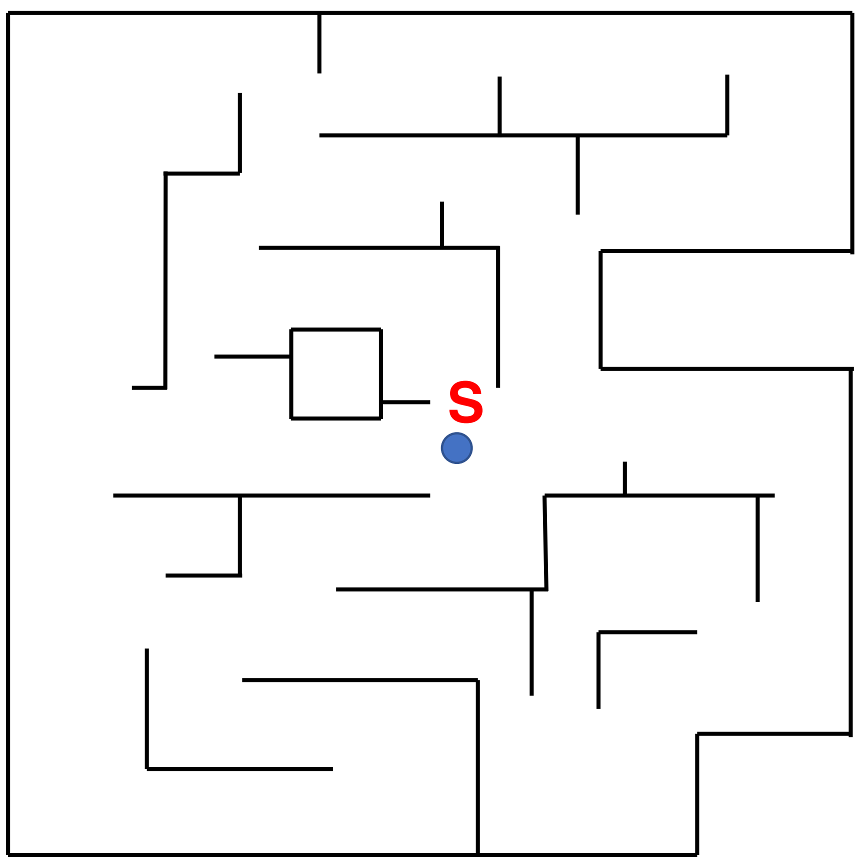
Curious Agent in 3D Maze



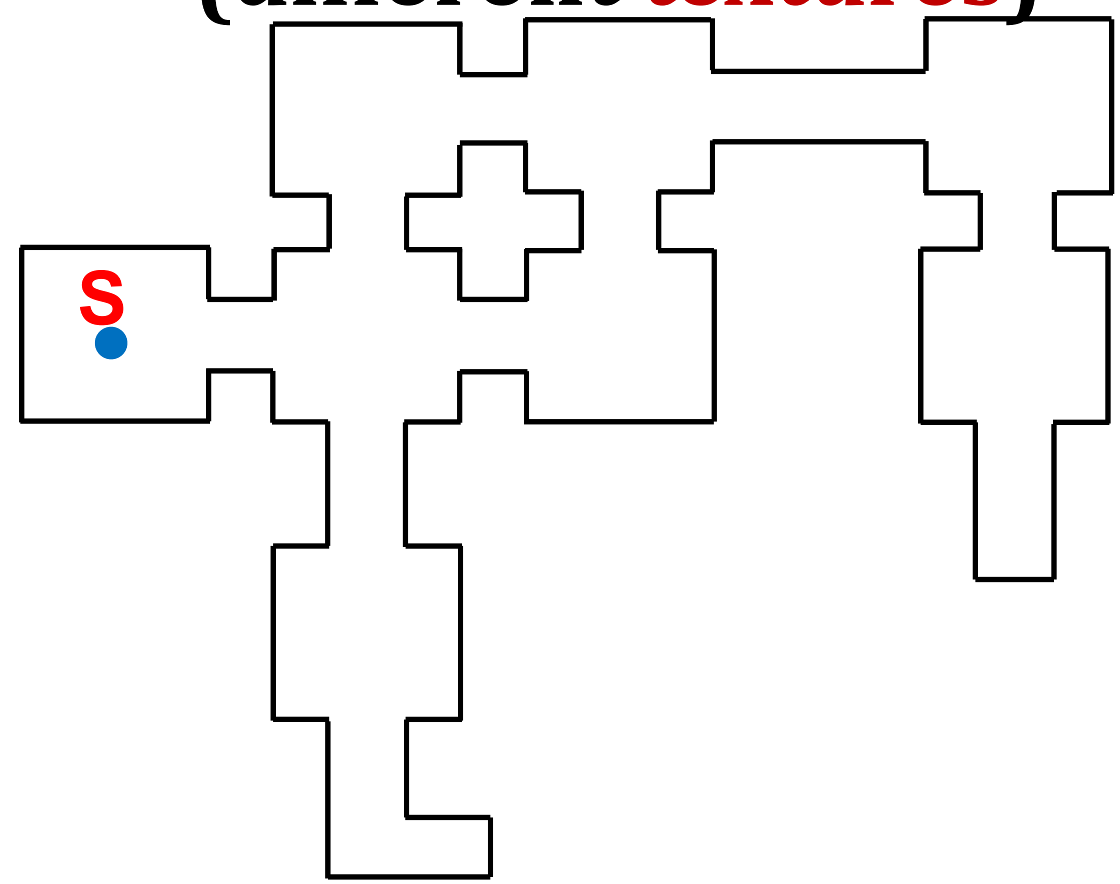
Curiosity driven Exploration by Self-Supervised Prediction, Pathak D., **Agrawal P.**, Efros A., Darrell T. , ICML 2017

Does the exploration policy generalize?

Train Map



Test Map
(different textures)



Note: Agent does not have access to Map

Curiosity driven Exploration by Self-Supervised Prediction, Pathak D., **Agrawal P.**, Efros A., Darrell T. , ICML 2017

Does the exploration policy generalize?

Train Map

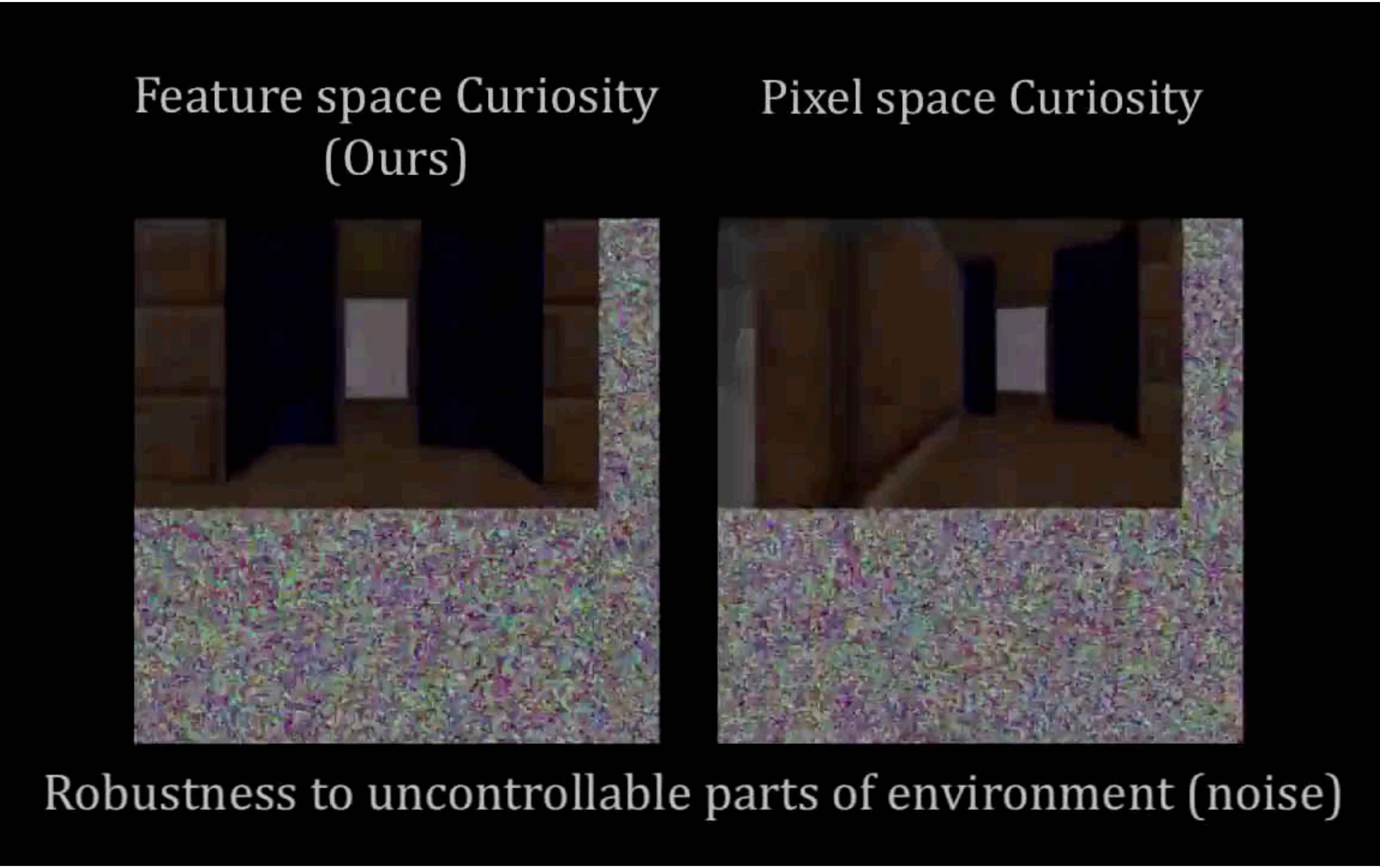
No Finetuning

Test Map
(different textures)

Note: Agent does not have access to Map

Curiosity driven Exploration by Self-Supervised Prediction, Pathak D., **Agrawal P.**, Efros A., Darrell T. , ICML 2017

Robustness to irrelevant parts



Curiosity driven Exploration by Self-Supervised Prediction, Pathak D., **Agrawal P.**, Efros A., Darrell T. , ICML 2017

Issues with Reinforcement Learning

Lots of data

Where do rewards come from?

Task Specific

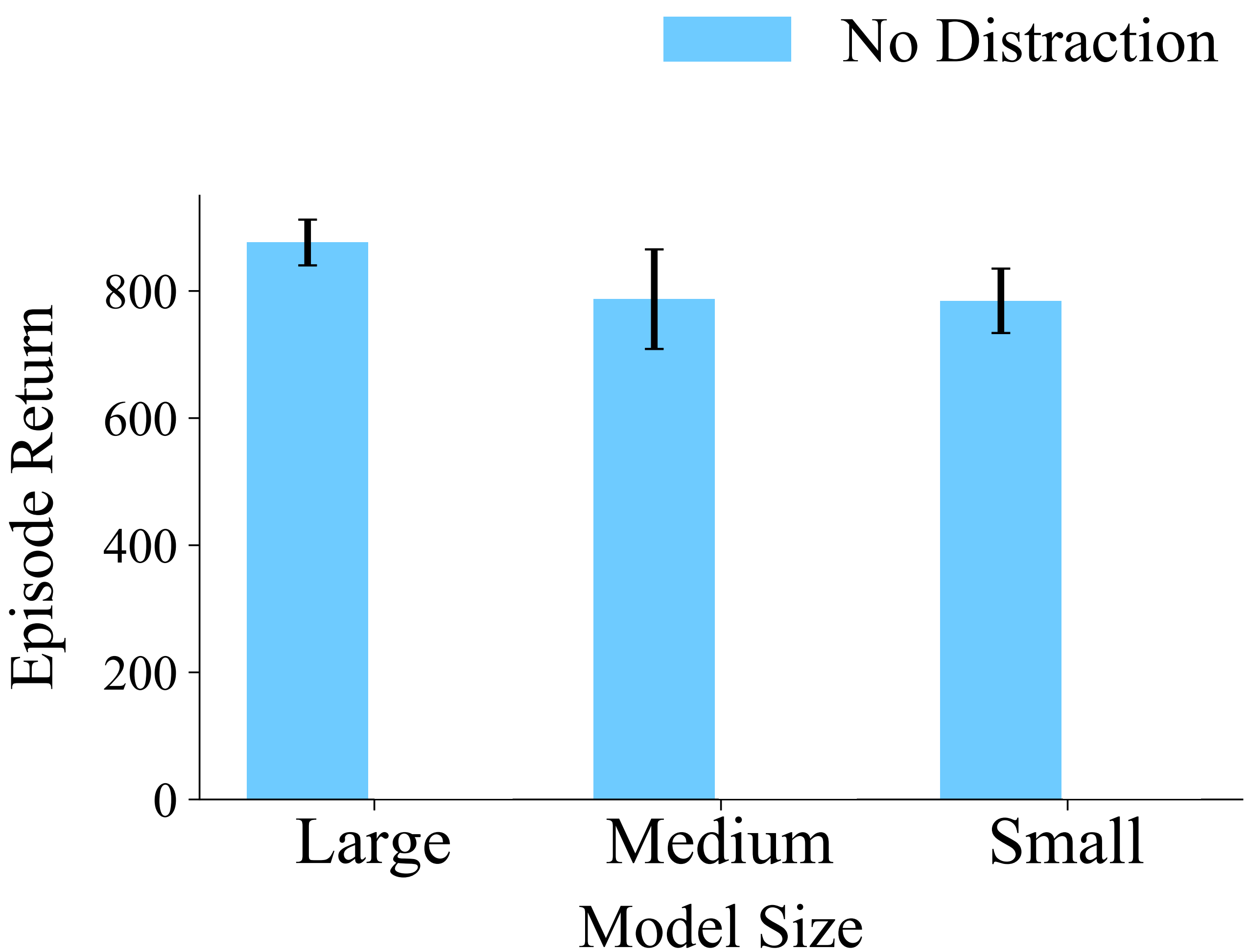
Demonstrations

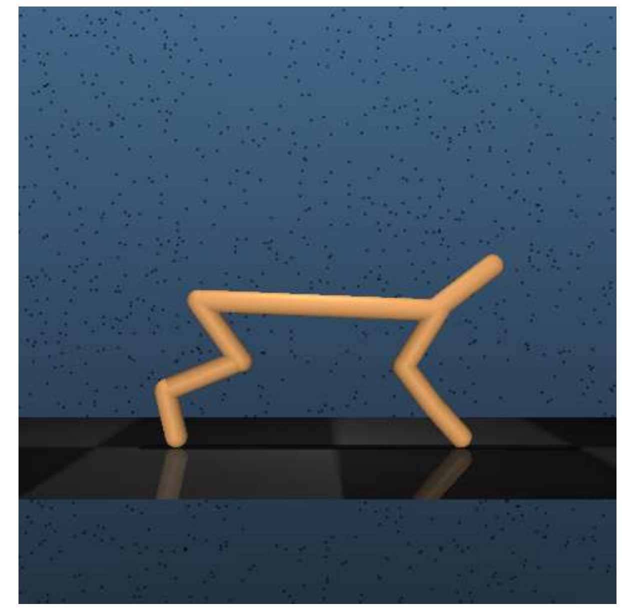
Task Curriculum

Exploration

Self-Supervised Model Learning

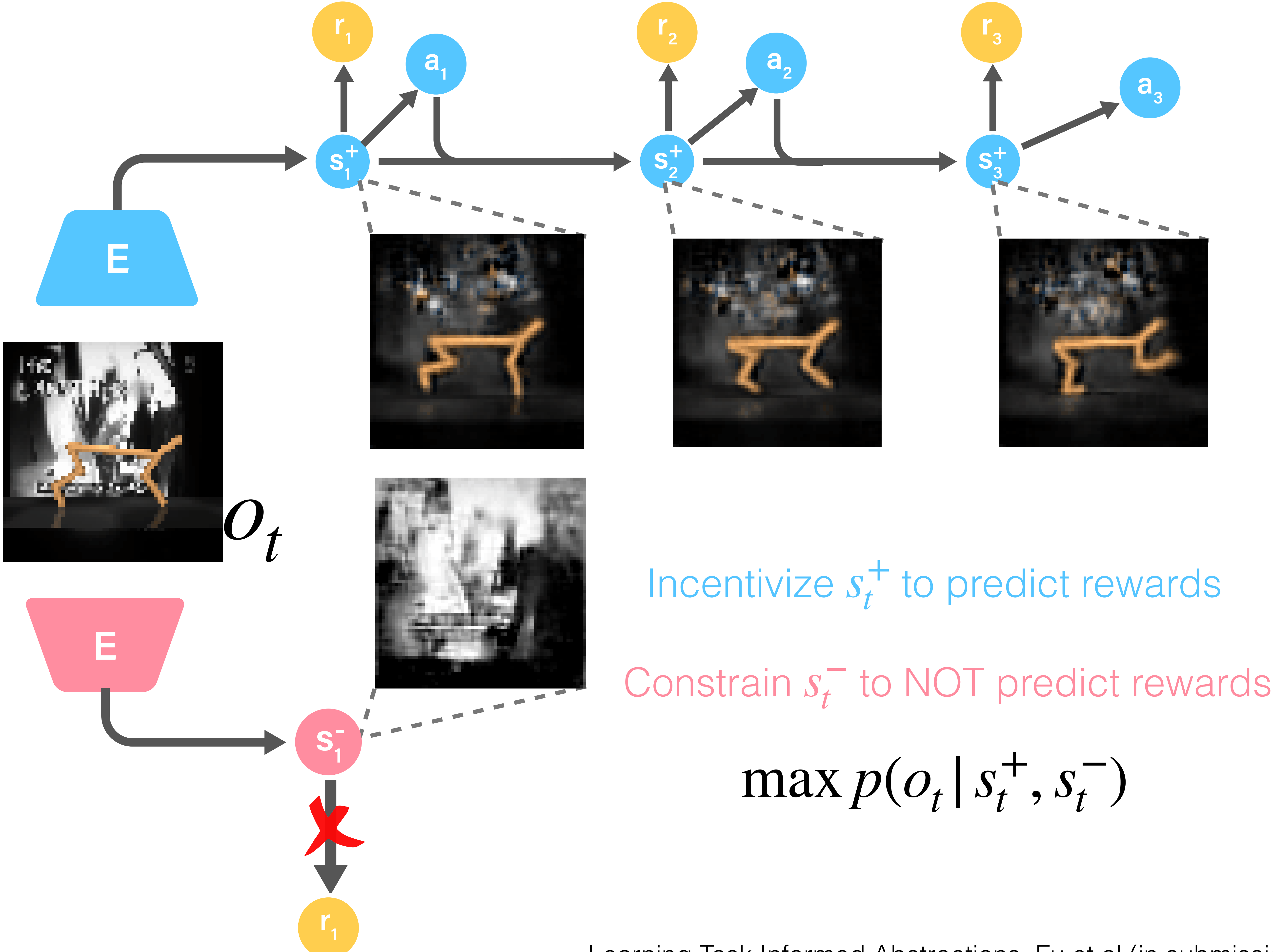
Learning Models from Natural Visual Data is Hard





Most model capacity
is consumed by distractors

Learning Task Informed Abstractions, Fu et al (in submission)



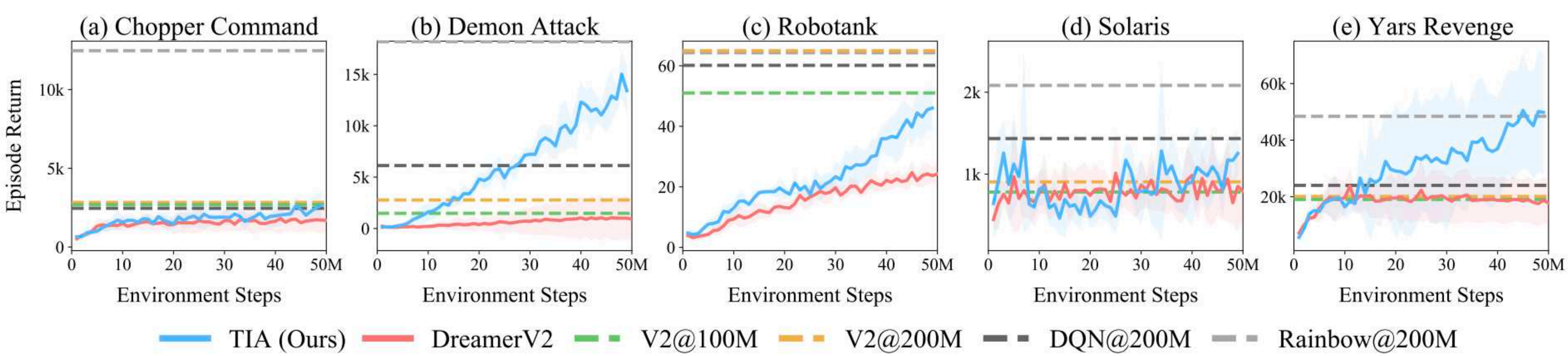
Learning Task Informed Abstractions, Fu et al (in submission)

Results Teaser

Raw Observation

Dreamer
(best prior work)

Ours



Learning Task Informed Abstractions, Fu et al (in submission)

Issues with Reinforcement Learning

Lots of data

Where do rewards come from?

Task Specific

Demonstrations

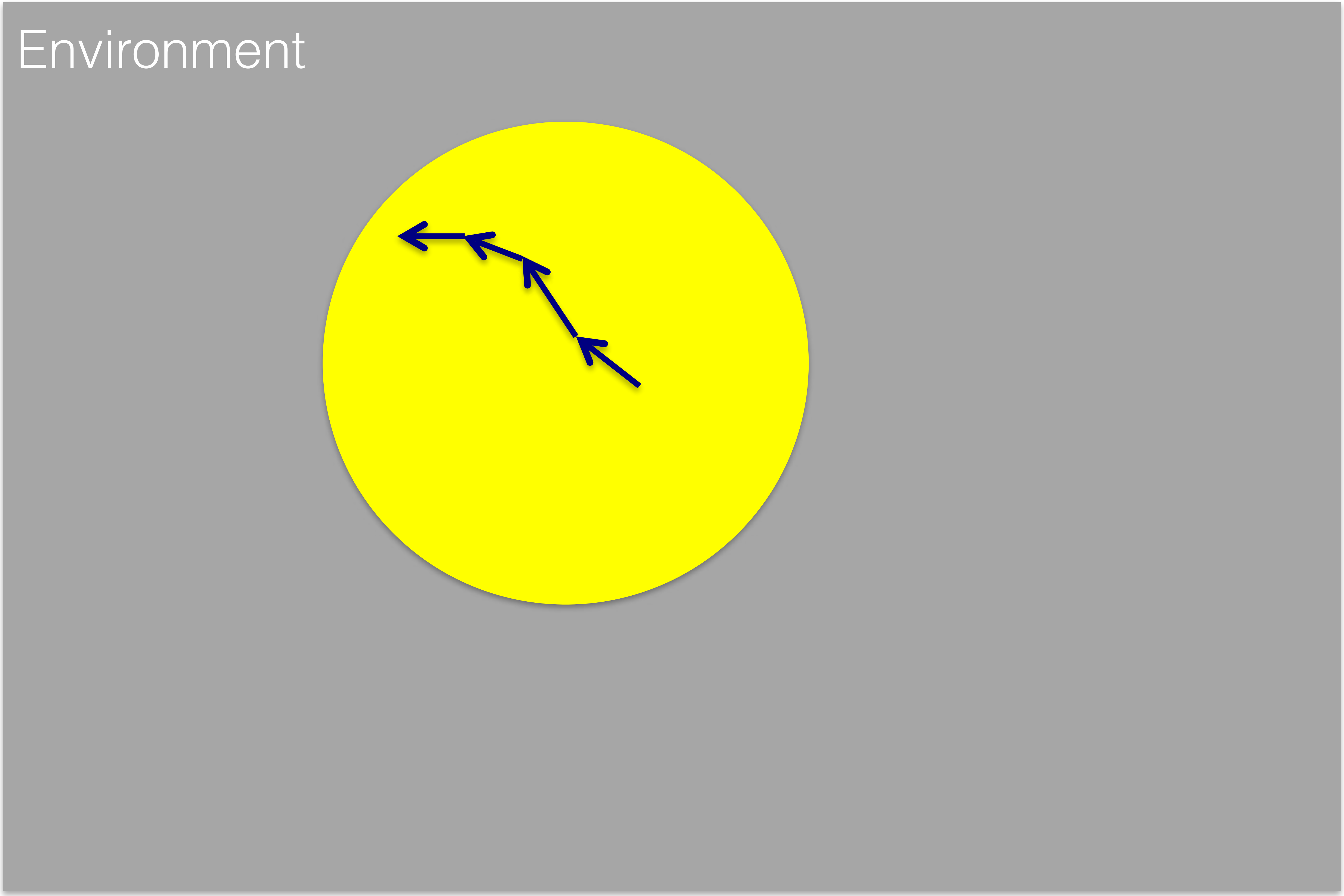
Task Curriculum

Self-Supervised Model Learning

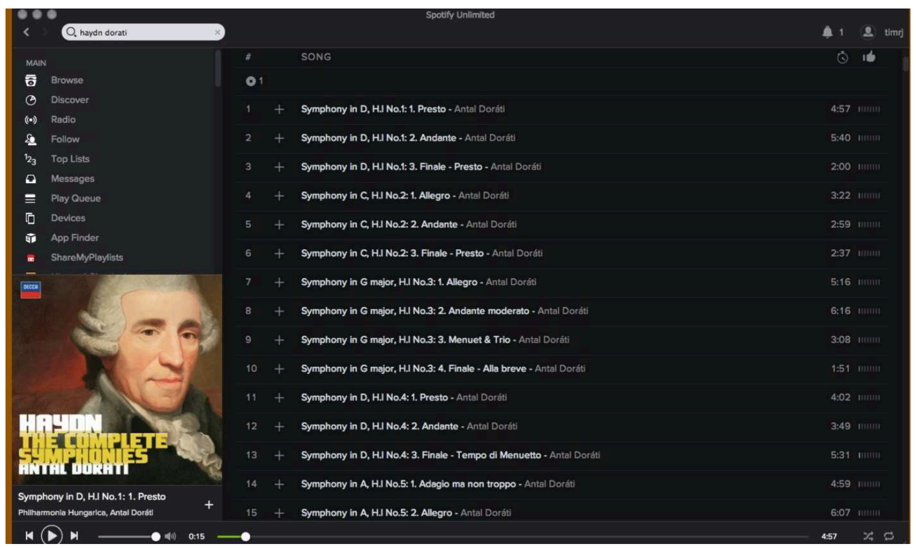
Exploration

Learning
Task-Relevant Models

Exploration has benefits, but is undesirable at times!

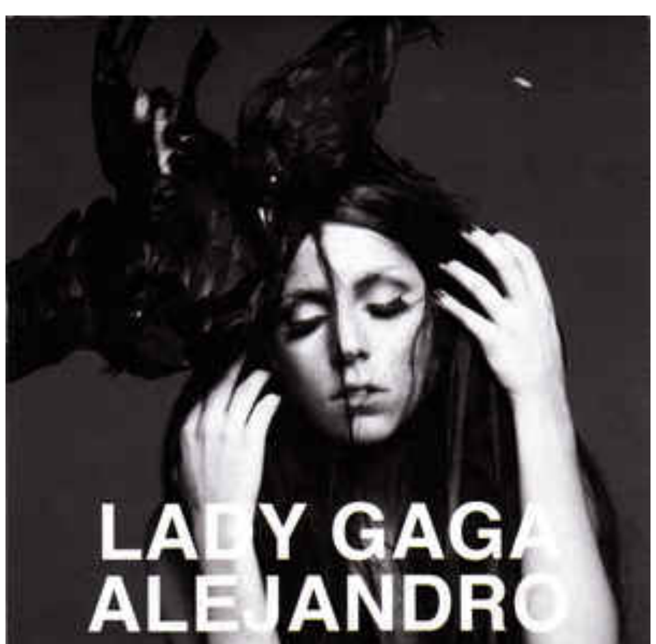


Imagine your favorite playlist



Spotify

(they want you hooked)



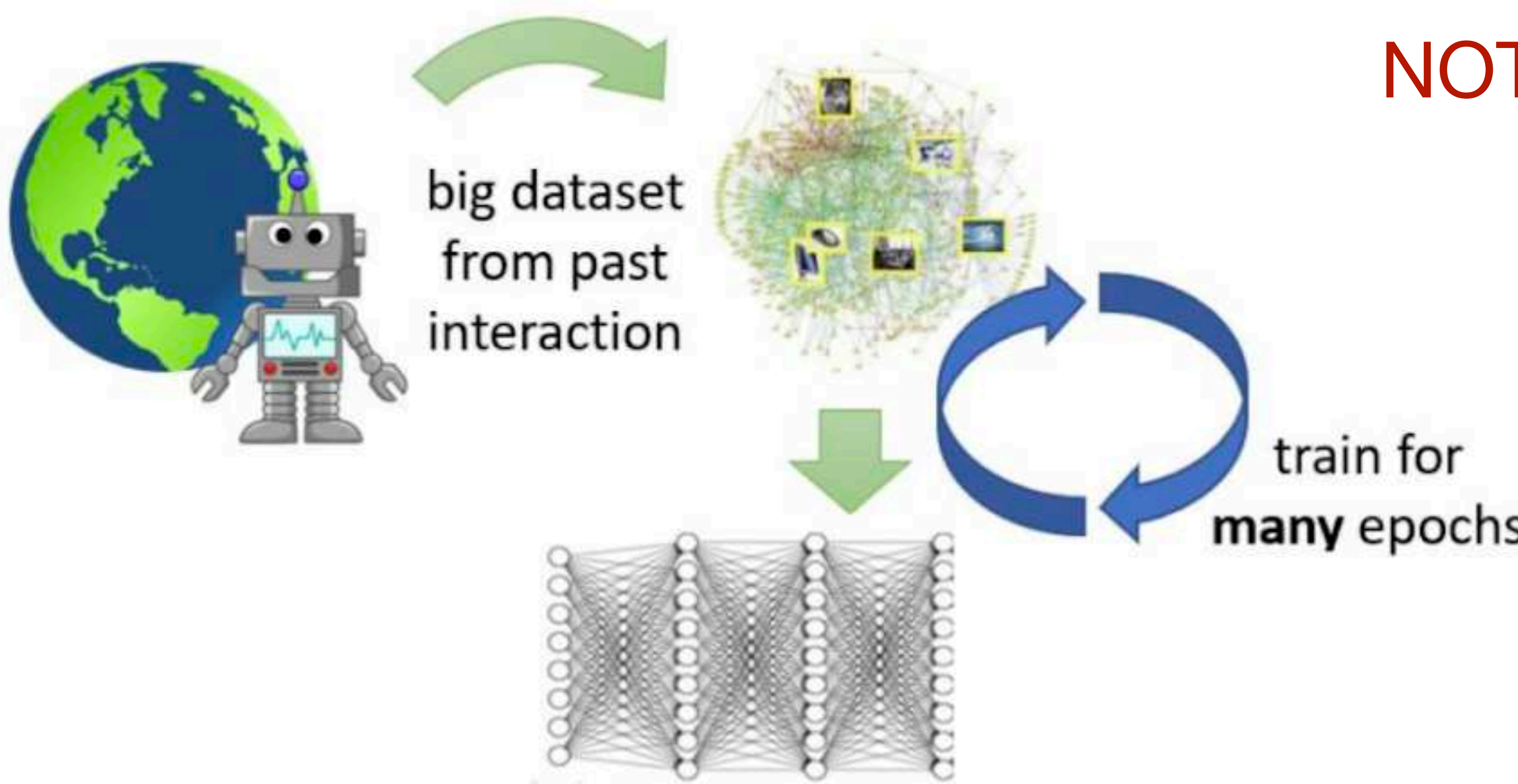
Explore by
Suggesting other music

Imagine your favorite playlist

Sometimes Exploration can be very costly!

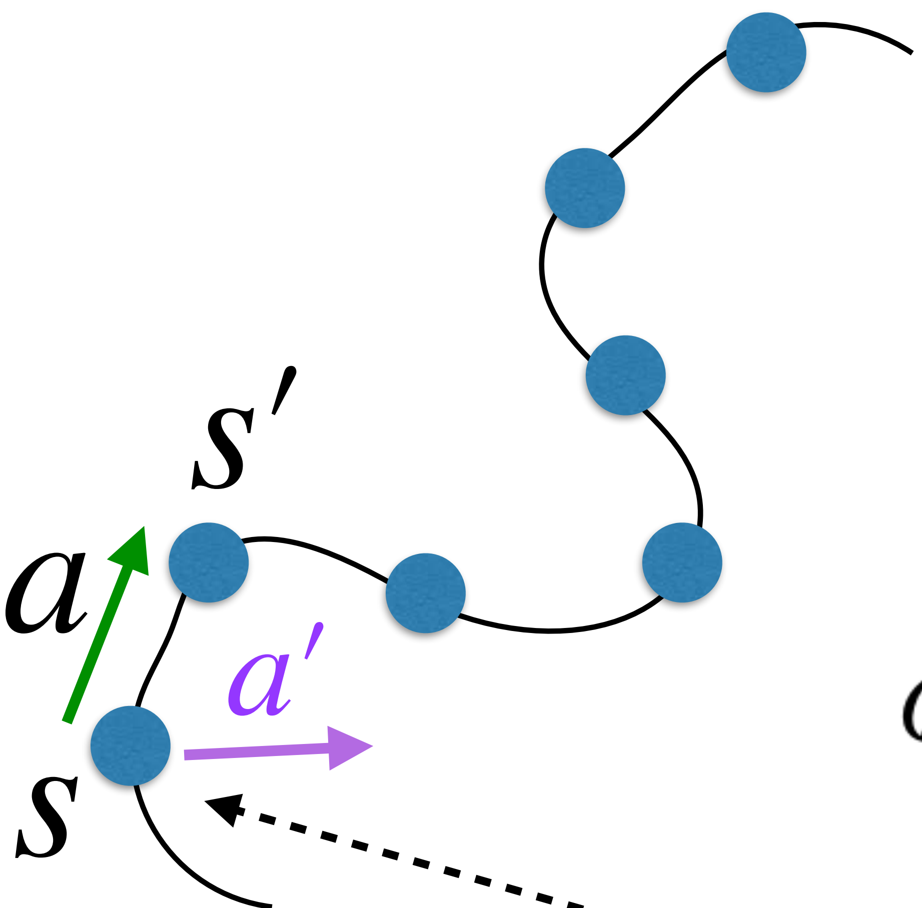
Suggesting other music

Online Data Collection
is
NOT always practical



Offline RL

Conservative Q-Learning for Offline RL



Existing Dataset

$$D : \{s_i, a_i, s_i'\} \quad i \in [1,N]$$

Q-Learning

$$\hat{Q}^{k+1} \leftarrow \frac{1}{2}\mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{s}'\sim\mathcal{D}}\left[\left(Q(\mathbf{s},\mathbf{a}) - \hat{\mathcal{B}}^{\pi}\hat{Q}^{k}(\mathbf{s},\mathbf{a})\right)^{2}\right]$$

$\hat{\mathcal{B}}^{\pi}$:Bellman Operator

$$a \sim \mu(a \mid s)$$

Learnt Policy

Conservative Q-Learning

$$\hat{Q}^{k+1} \leftarrow \arg\min_{Q} \; \alpha\, \mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\mu(\mathbf{a}|\mathbf{s})}\left[Q(\mathbf{s},\mathbf{a})\right] + \frac{1}{2}\mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}}\left[\left(Q(\mathbf{s},\mathbf{a}) - \hat{\mathcal{B}}^{\pi}\hat{Q}^{k}(\mathbf{s},\mathbf{a})\right)^{2}\right]$$

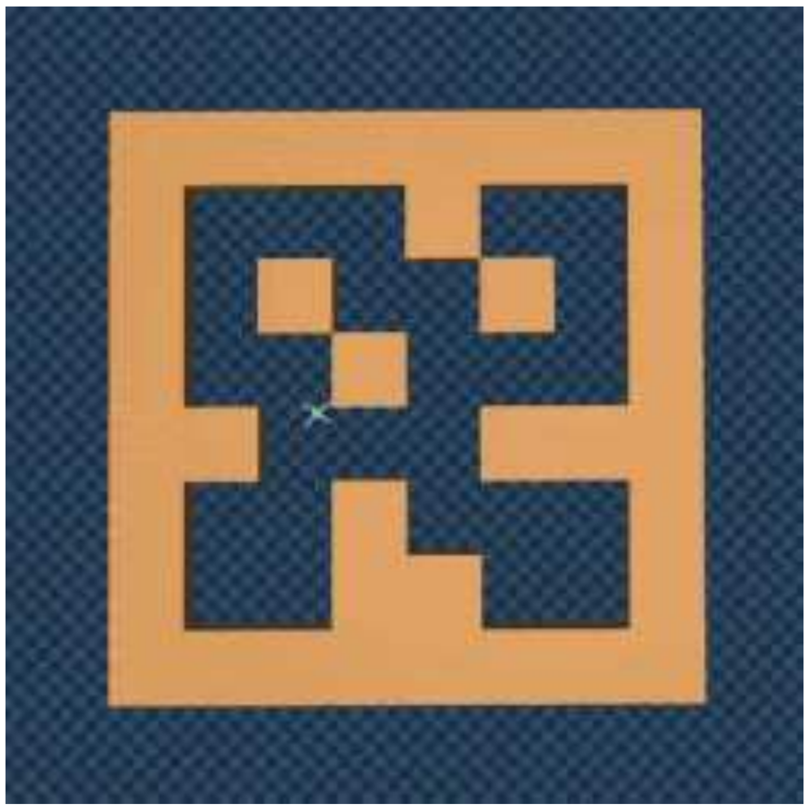
Overestimate

$$Q(s, a')$$

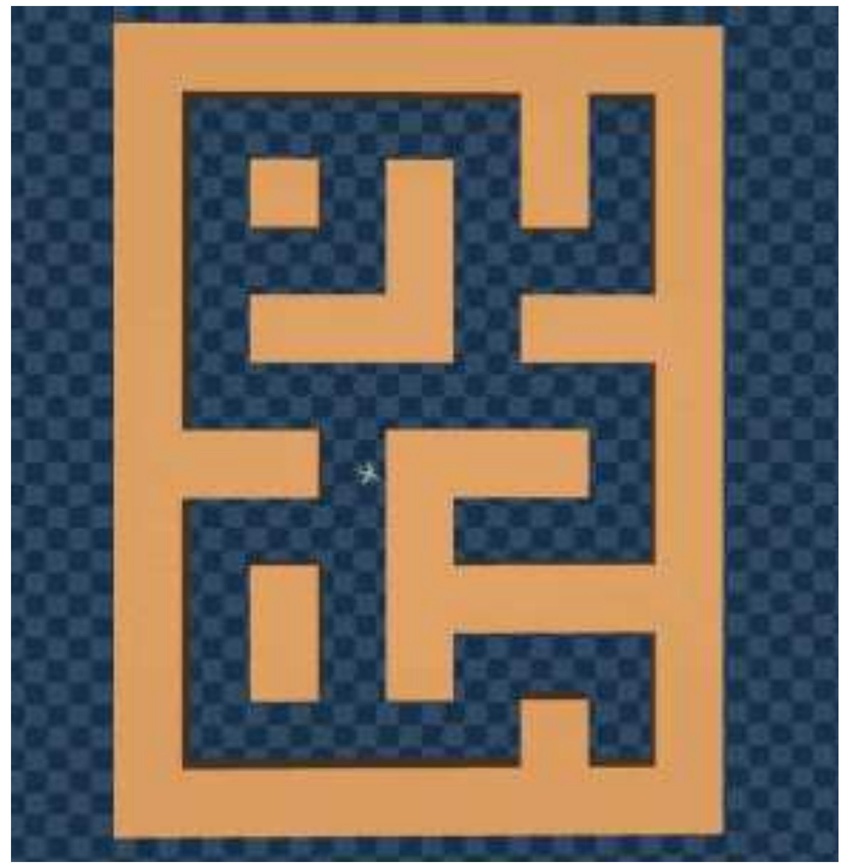
Kumar et al., 2020

Improving Offline Learning with Action Primitives

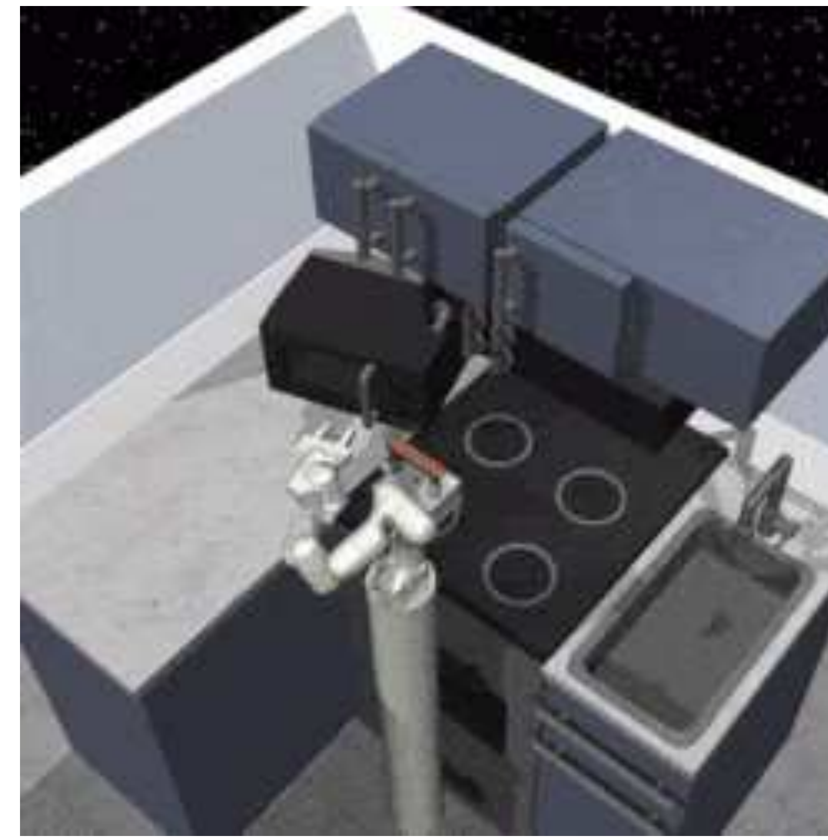
Improving Offline Learning with Action Primitives



Antmaze medium



Antmaze large



kitchen

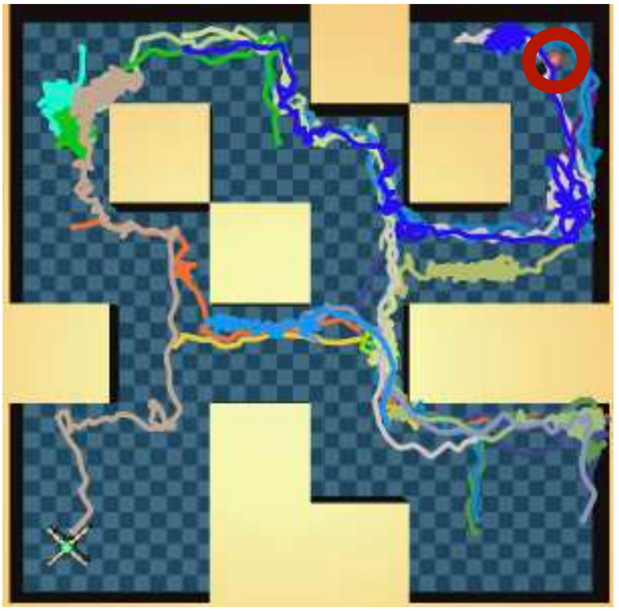
State (joint angles + xy pose): 29 dim

Action (joint torques): 8 dim

State (joint angles + xy pose): 60 dim

Action (joint torques): 9 dim

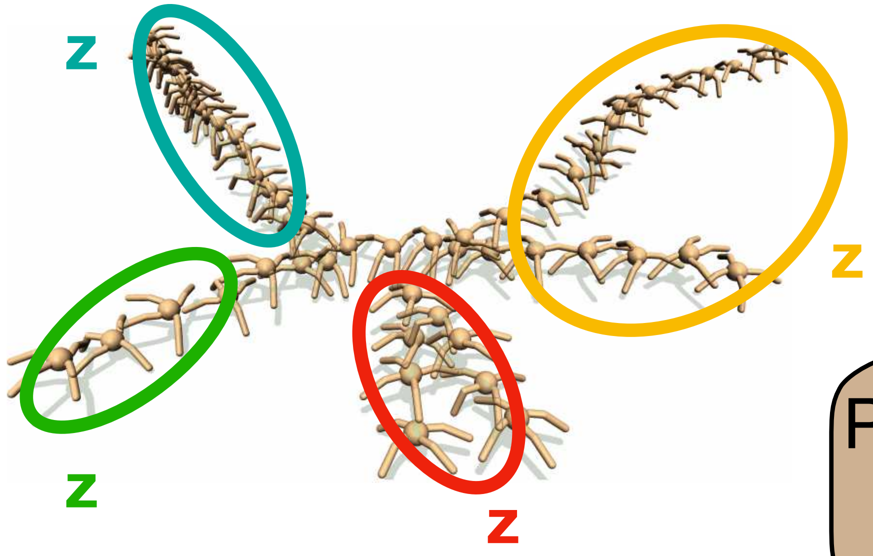
OPAL: **O**ffline **P**rimitive Discovery for **A**ccelerating R**L**



+1 for reaching goal

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

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



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

Easier to optimize
(reduction in plan length)

Cluster actions to learn "skills" (or action primitives)

Ajay et al., ICLR 2021

Results

Environment	BC	BEAR	EMAQ	CQL	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: Conservative Q Learning (Kumar et al, 2020)

BC: Behavioral Cloning

BEAR: Bootstrapping error accumulation reduction (Kumar et al, 2019)

EMAQ: Expected Max-Q Learning (Ghasemipour et al, 2020)

Ajay et al., ICLR 2021

Issues with Reinforcement Learning

Lots of data

Where do rewards come from?

Task Specific

Demonstrations

Task Curriculum

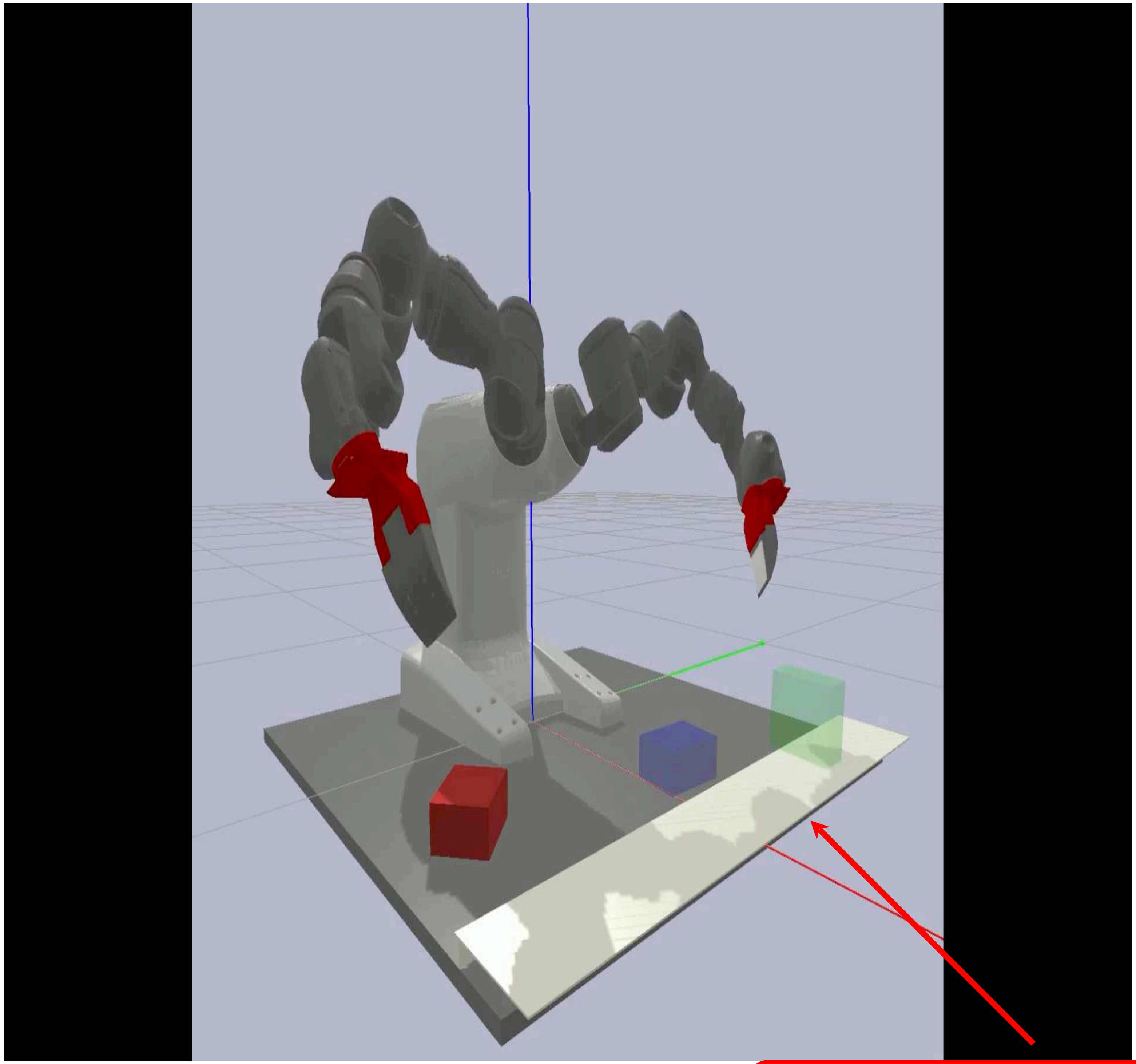
Self-Supervised Model Learning

Exploration

Learning Task-Relevant Models

Safer learning from existing data

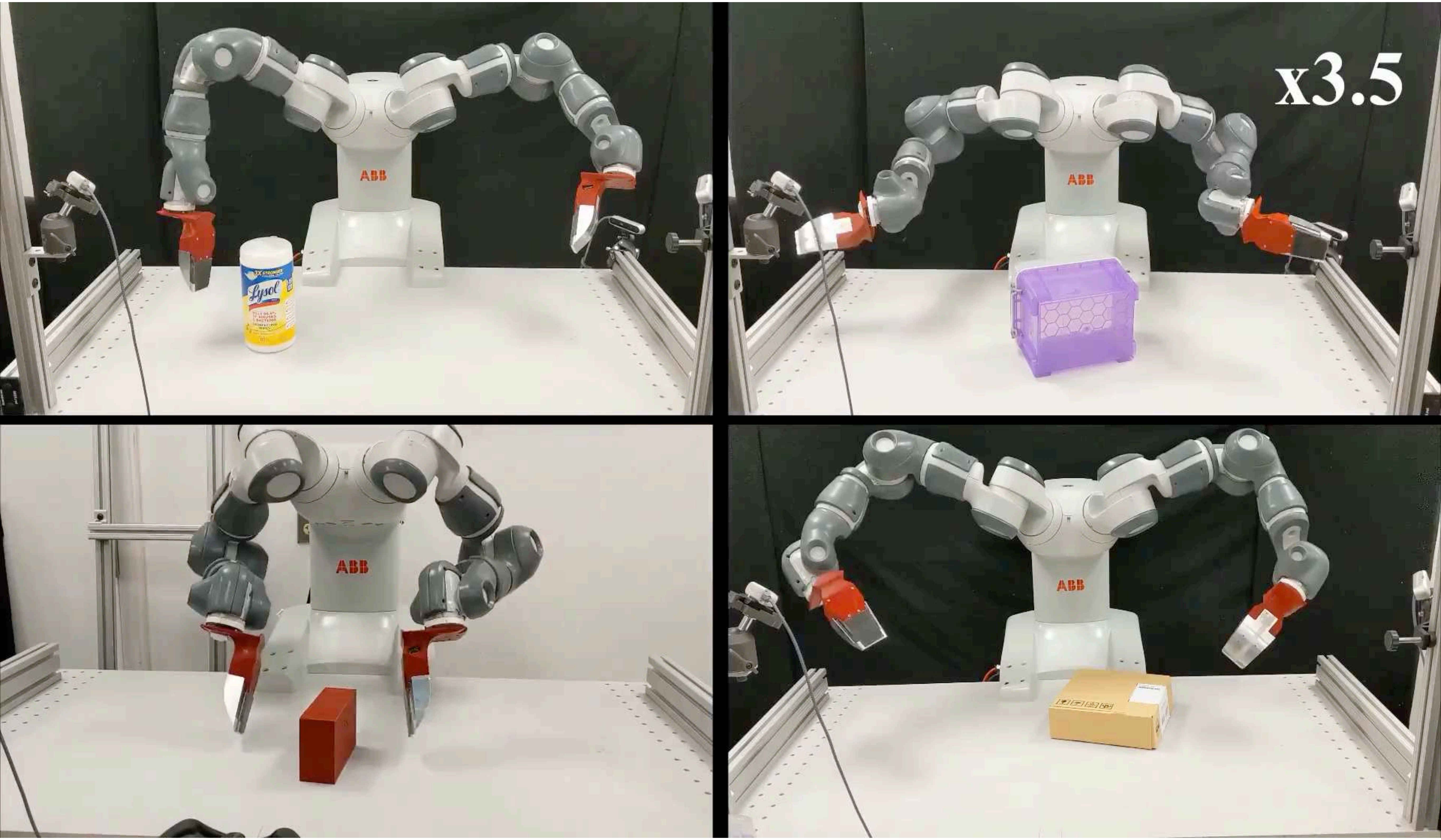
Using Skills for Long-Term Planning from Visual Sensing



$\texttt{PullRight} \rightarrow \texttt{GraspReorient} \rightarrow \texttt{PullRight} \rightarrow \boxed{\texttt{GraspPlaceShelf}} \rightarrow \texttt{PullLeft}$

A Long Horizon Planning Framework For Manipulating Rigid Pointcloud Objects,
A. Simeonov, Y. Du, B. Kim, F. Hogan, J. Tenenbaum, **P. Agrawal,** A. Rodriguez, CoRL 2020

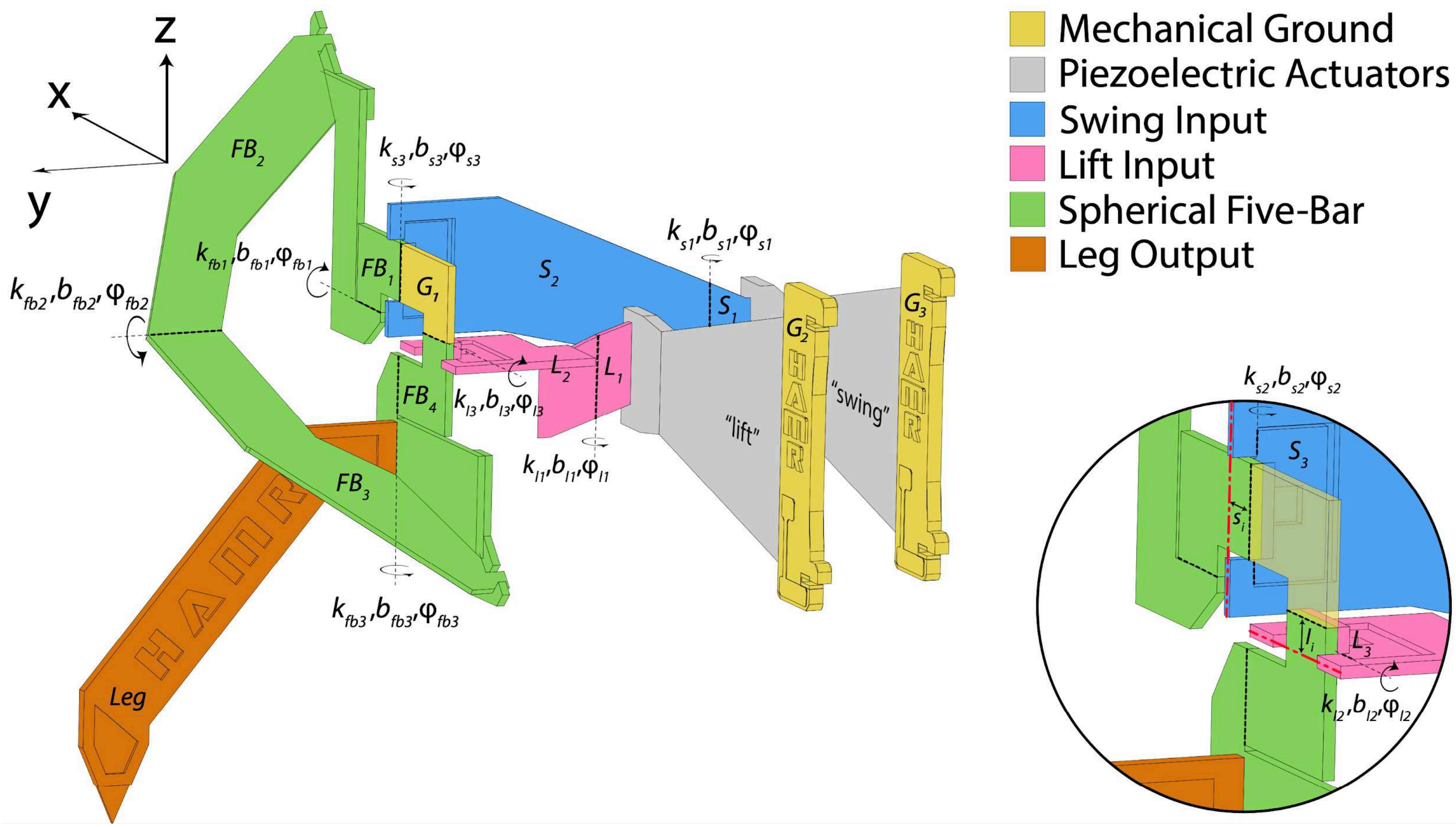
Using Skills for Long-Term Planning from Visual Sensing



A Long Horizon Planning Framework For Manipulating Rigid Pointcloud Objects,
A. Simeonov, Y. Du, B. Kim, F. Hogan, J. Tenenbaum, **P. Agrawal,** A. Rodriguez, CoRL 2020

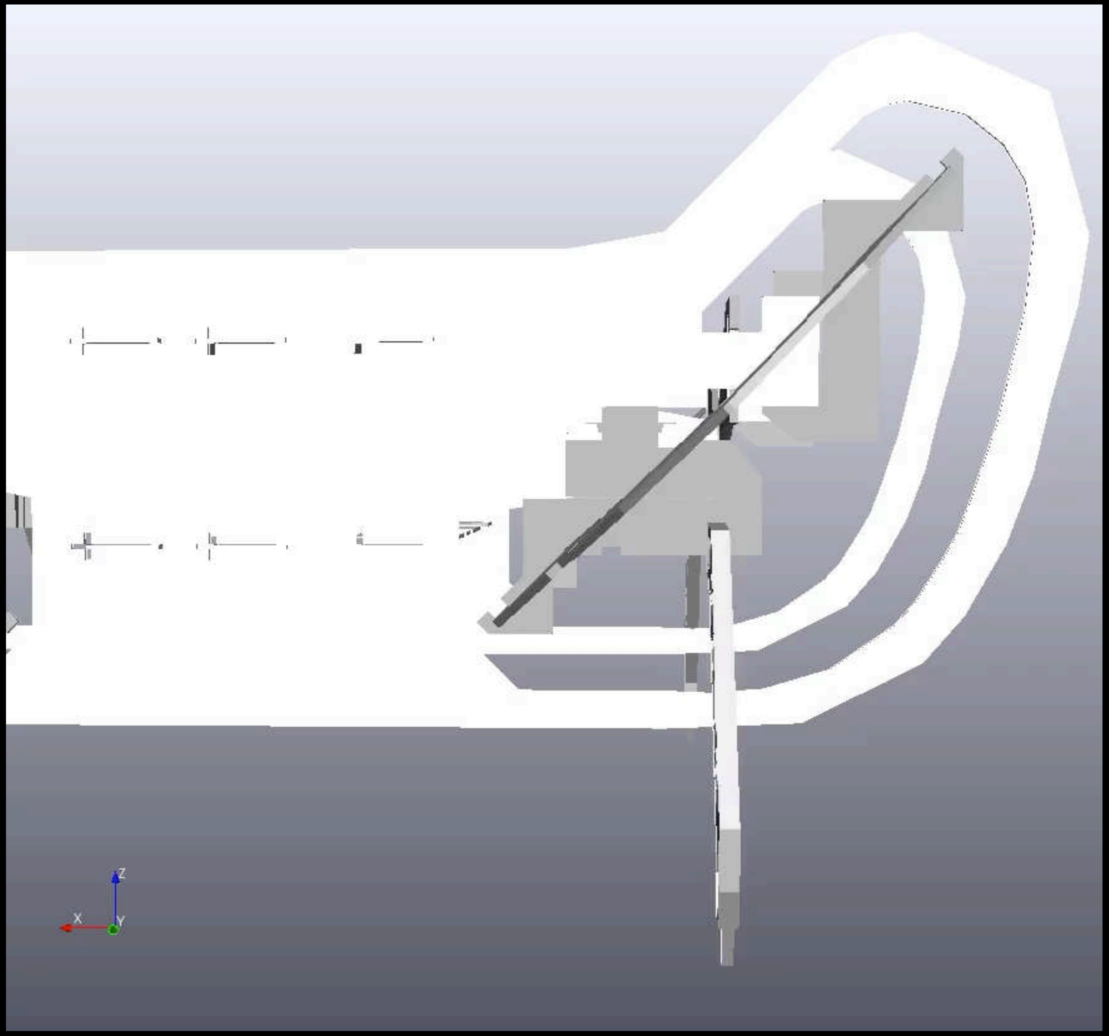
Consider Microrobots

Simulating HAMR is expensive and is inaccurate

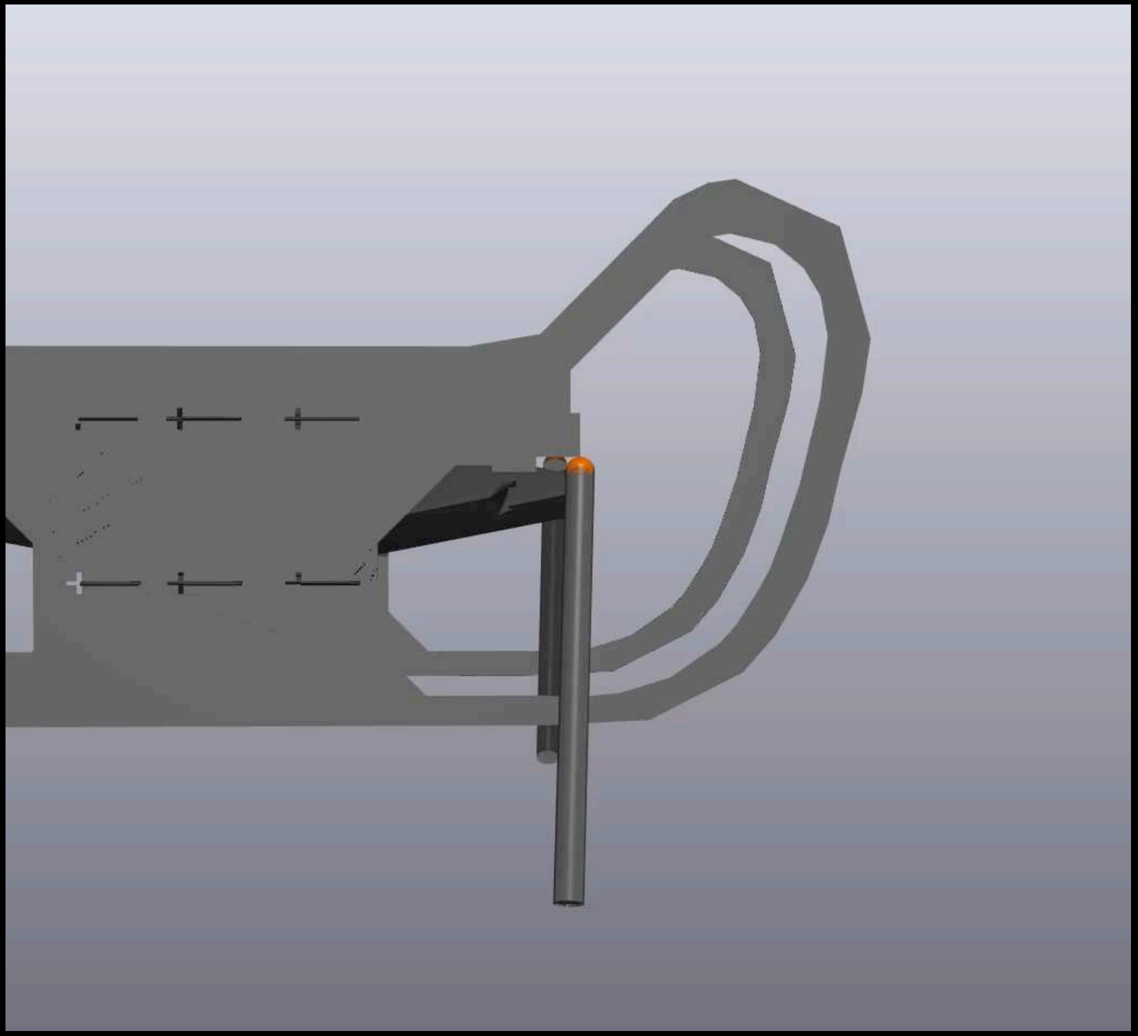


98.96x slower than realtime

Full Model



Simple Model



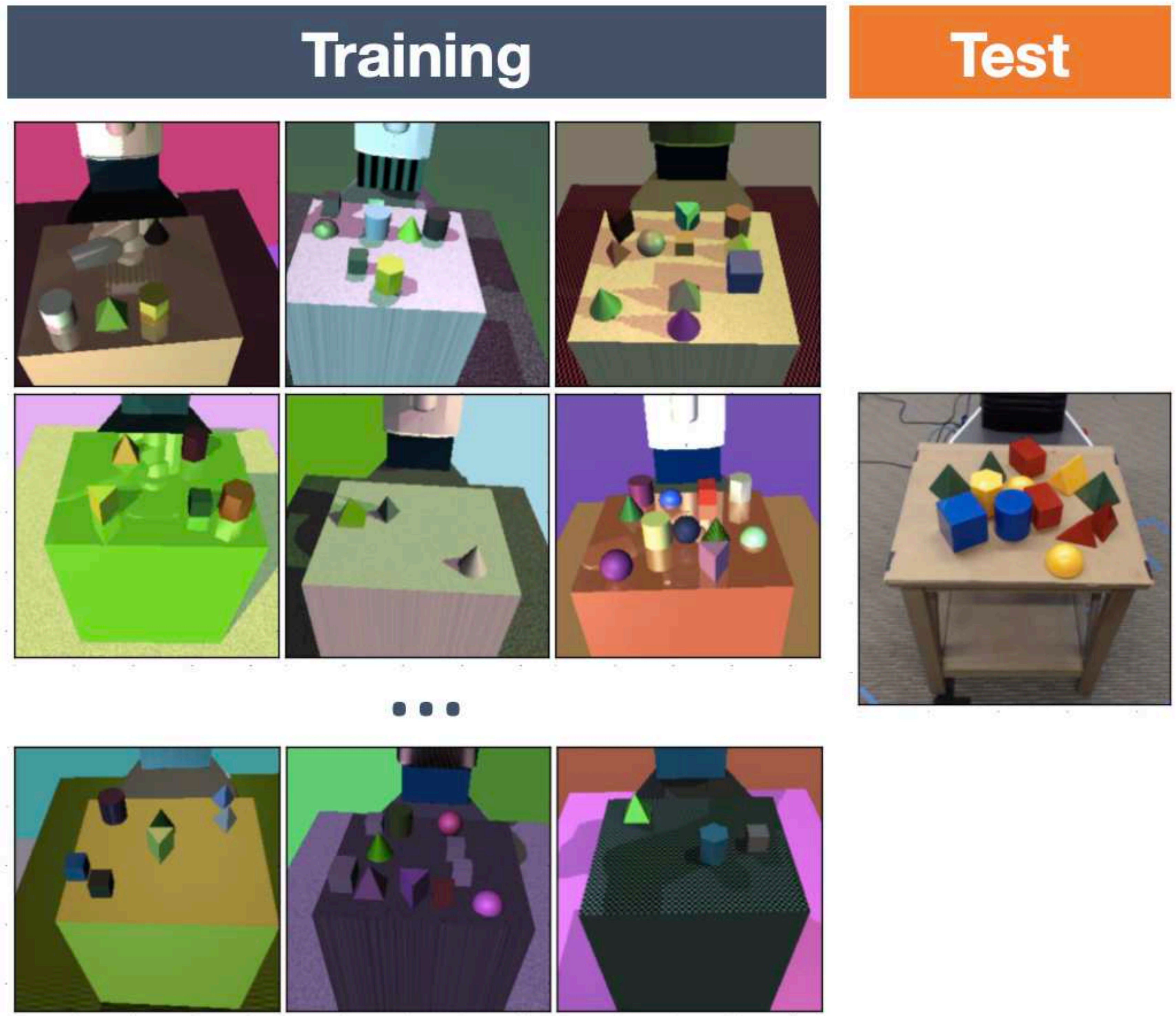
Faster than real-time simulation

(but not accurate enough for learning to control)

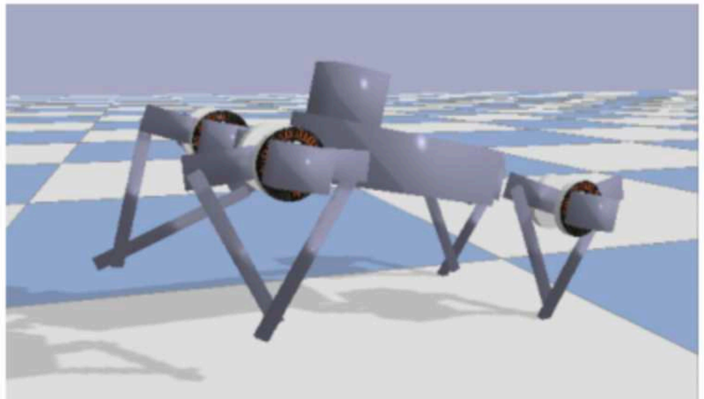
Learn the full model from scratch

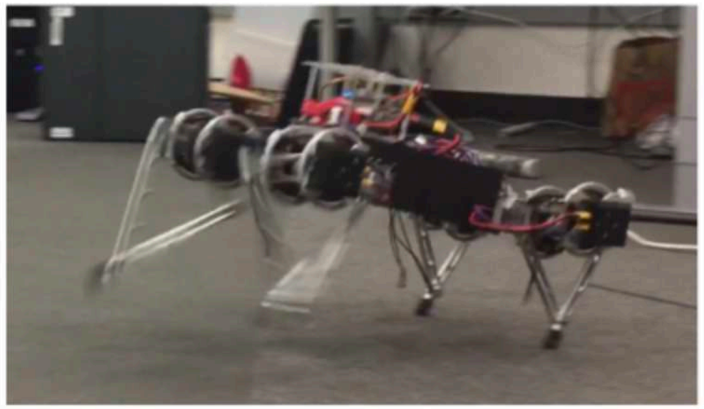
Domain Randomization

Domain Randomization

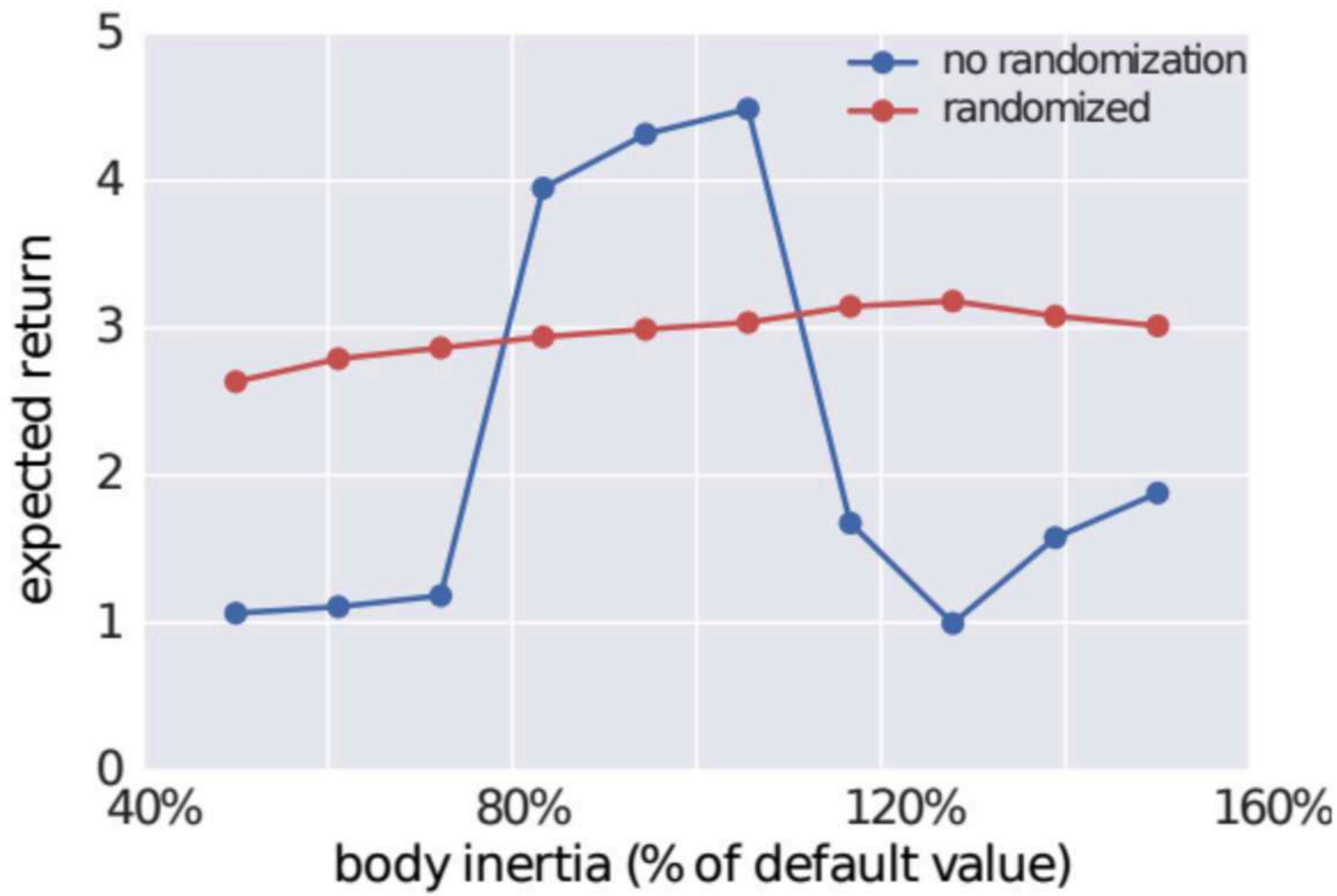


Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World, Tobin et al., 2017





Disadvantage of Domain Randomization



Sim-to-Real: Learning Agile Locomotion For Quadruped Robots, Tan et al., 2018

Residual Model Learning

Don't learn more than we need to!

simulators are reasonably good at hard contacts and rigid-body dynamics.

↓

- Use simulator to simulate a bare-minimum simplified model
- Use learning to compensate for the model difference (residual)

Residual Model Learning for Microrobot Control, Gruenstein et al. (in submission)

Learning the Residual Model

Simulator with Simple Model

$$\ddot{\mathbf{q}} = f_{\text{simple}}(\mathbf{q}, \dot{\mathbf{q}}, \mathbf{u})$$

Simulator with Full Model

$$\ddot{\mathbf{q}} = f_{\text{full}}(\mathbf{q}, \dot{\mathbf{q}}, \mathbf{u})$$

However, $\boxed{f_{\text{simple}} \neq f_{\text{full}}}$

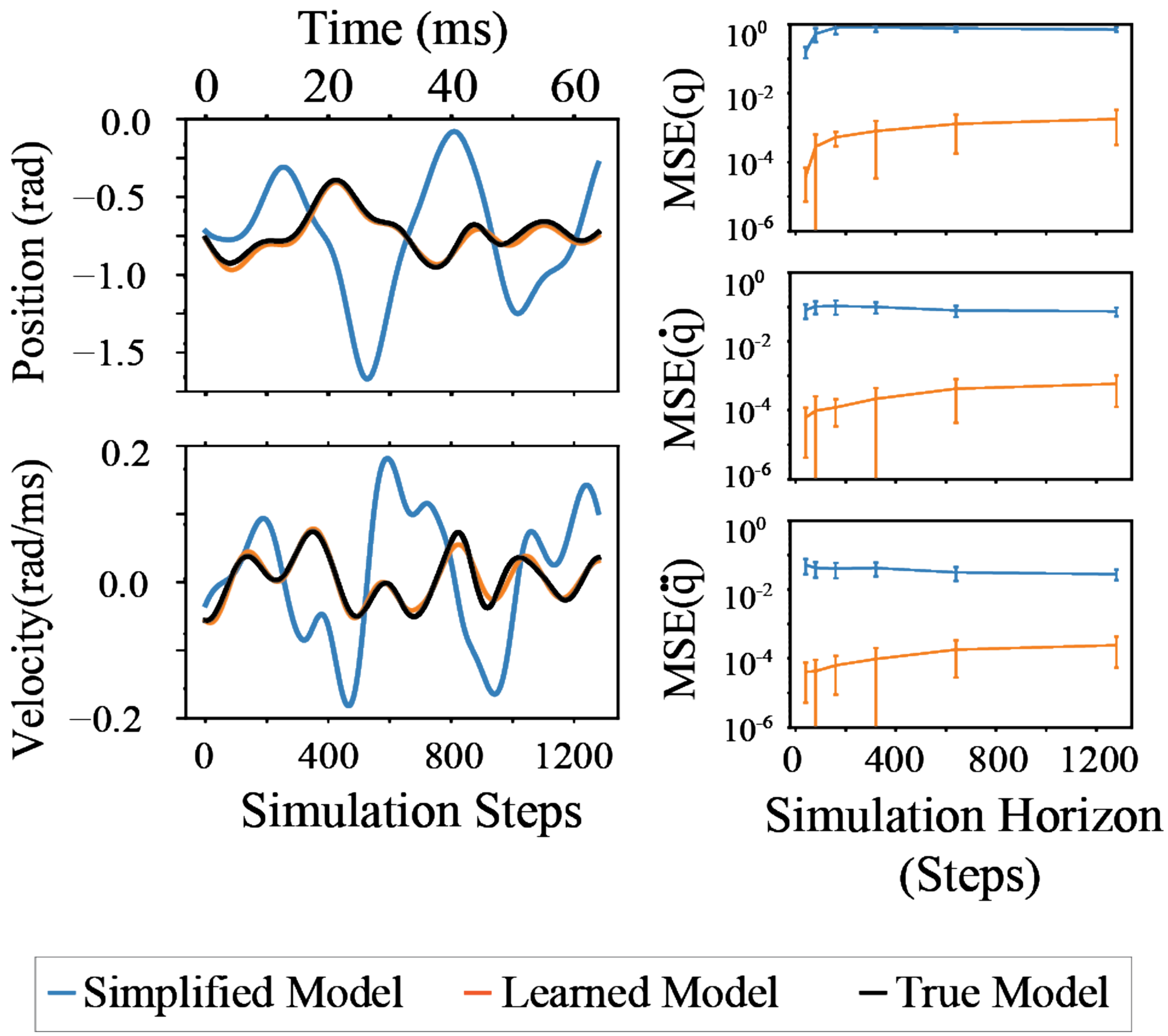
Learn $\mathbf{u}_{\text{simple}} = \hat{k}(\mathbf{q}, \dot{\mathbf{q}}, \mathbf{u}_{\text{full}})$

$$\boxed{f_{\text{simple}}(\mathbf{q}, \dot{\mathbf{q}}, \hat{k}(\mathbf{q}, \dot{\mathbf{q}}, \mathbf{u}_{\text{full}})) \approx f_{\text{full}}(\mathbf{q}, \dot{\mathbf{q}}, \mathbf{u}_{\text{full}}).}$$

Key Idea: Learn to modify inputs to simple model, so it matches the full model

Residual Model Learning for Microrobot Control, Gruenstein et al. (in submission)

How well does this perform?



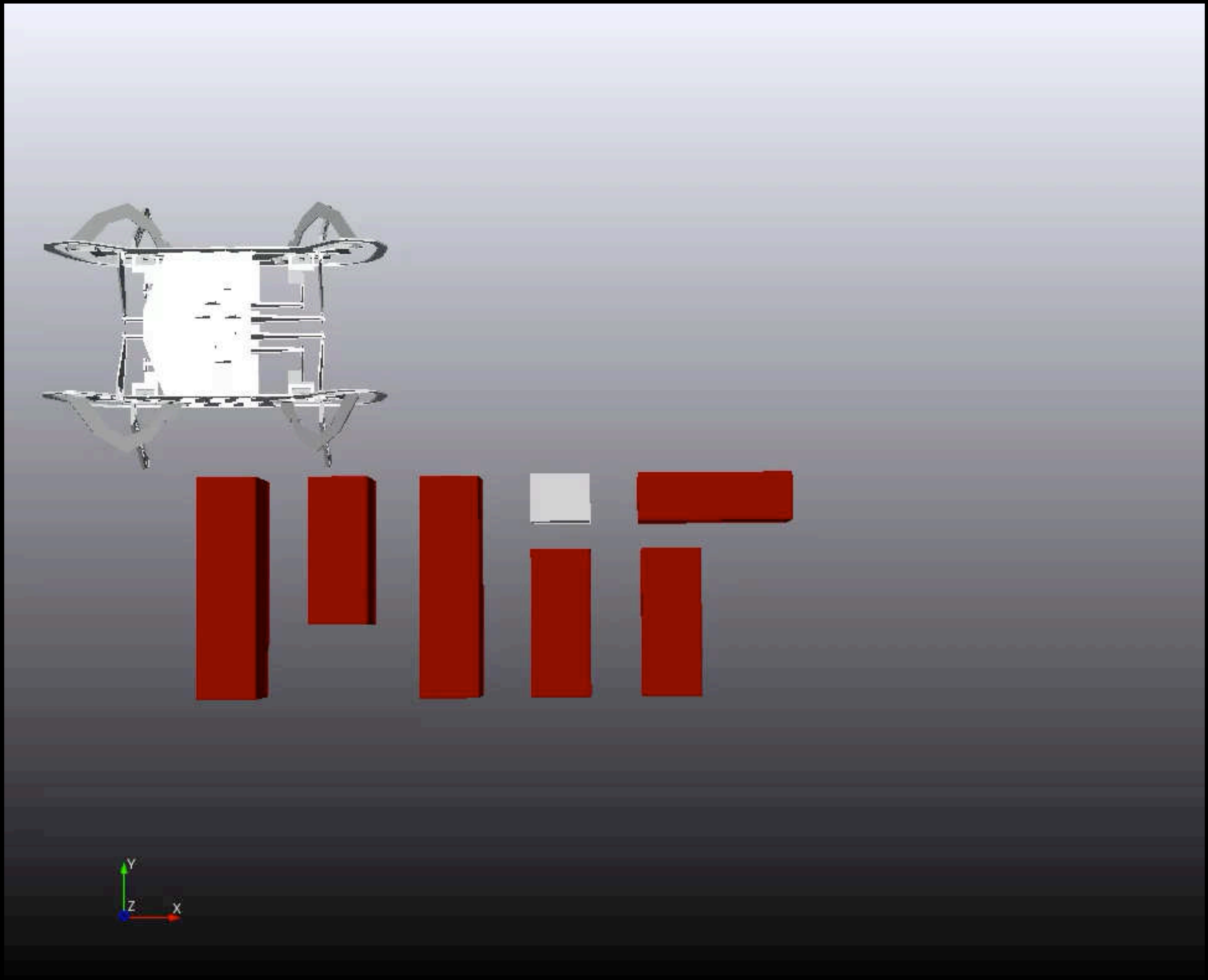
Only requires 12 seconds of data

47x faster

than simulating the full model

Residual Model Learning for Microrobot Control, Gruenstein et al. (in submission)

Residual Model can be used for learning control



Residual Model Learning for Microrobot Control, Gruenstein et al. (in submission)

Issues with Reinforcement Learning

Lots of data

Where do rewards come from?

Task Specific

Demonstrations

Task Curriculum

Exploration

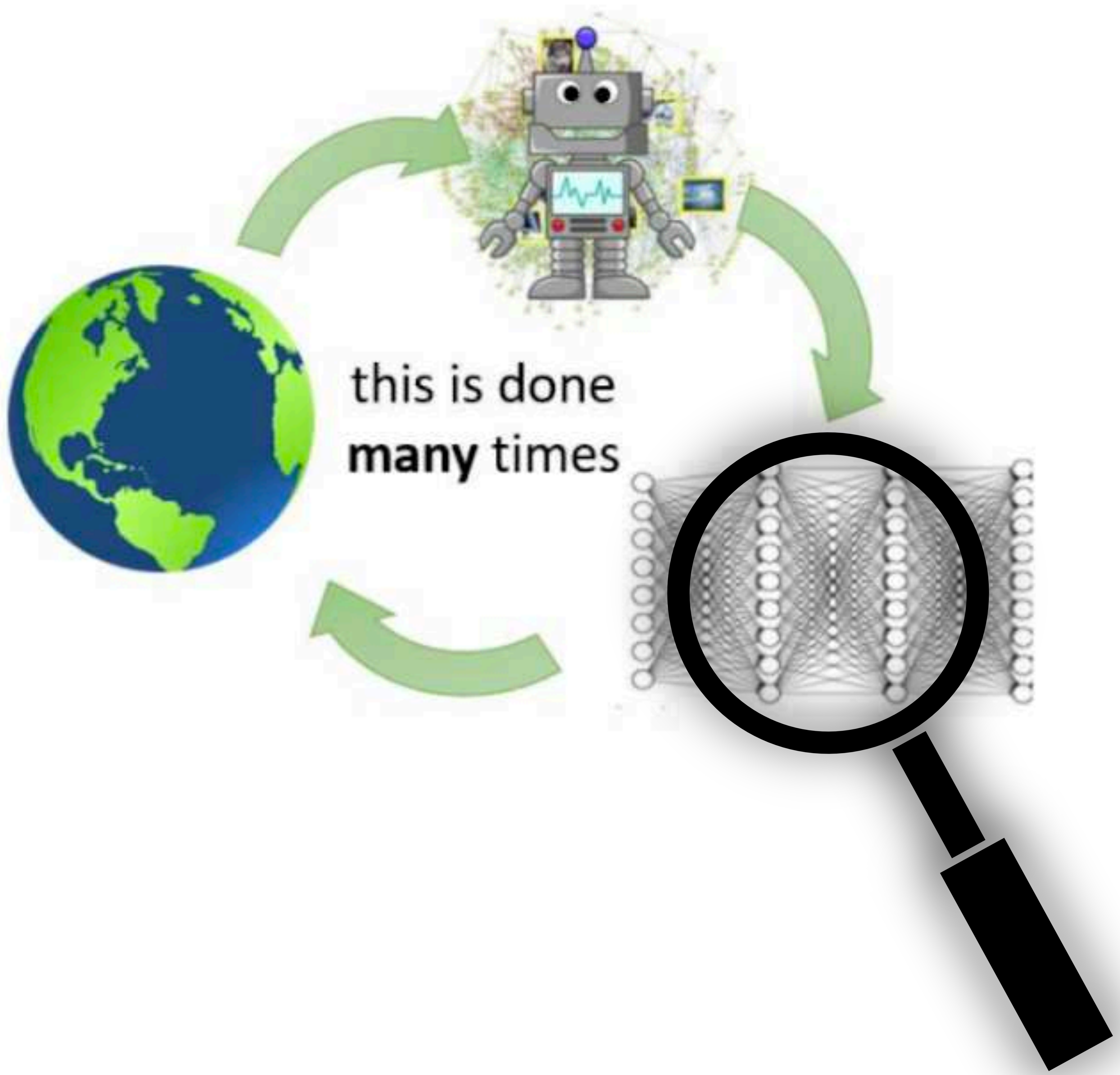
Self-Supervised Model Learning

Learning Task-Relevant Models

Efficient Learning In complex systems

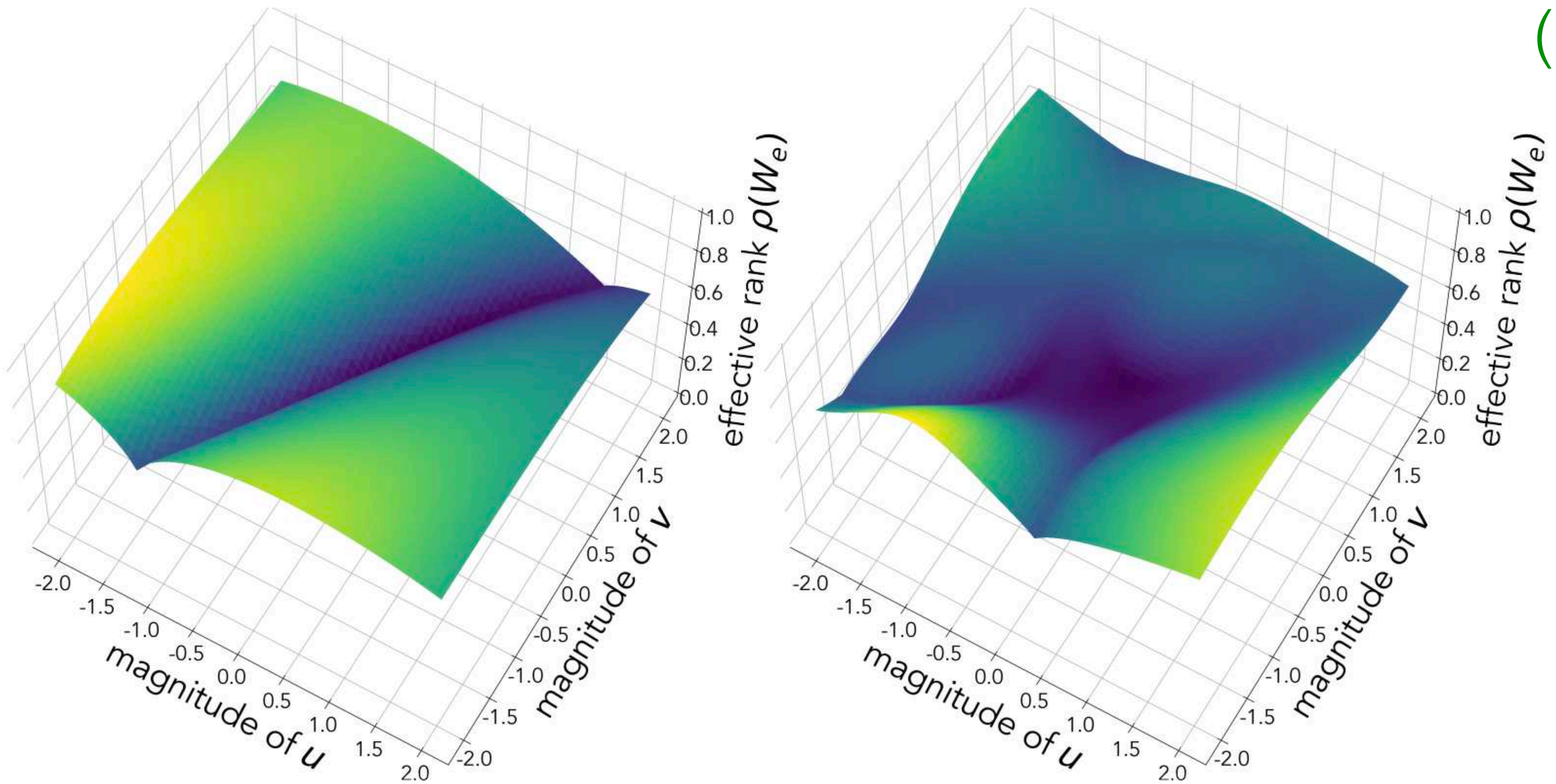
Safer learning from existing data

Use Deep Neural Networks as the Workhose



Deeper Nets → More Parameters → ~~Complex~~ Functions

Simple Functions
(low rank)

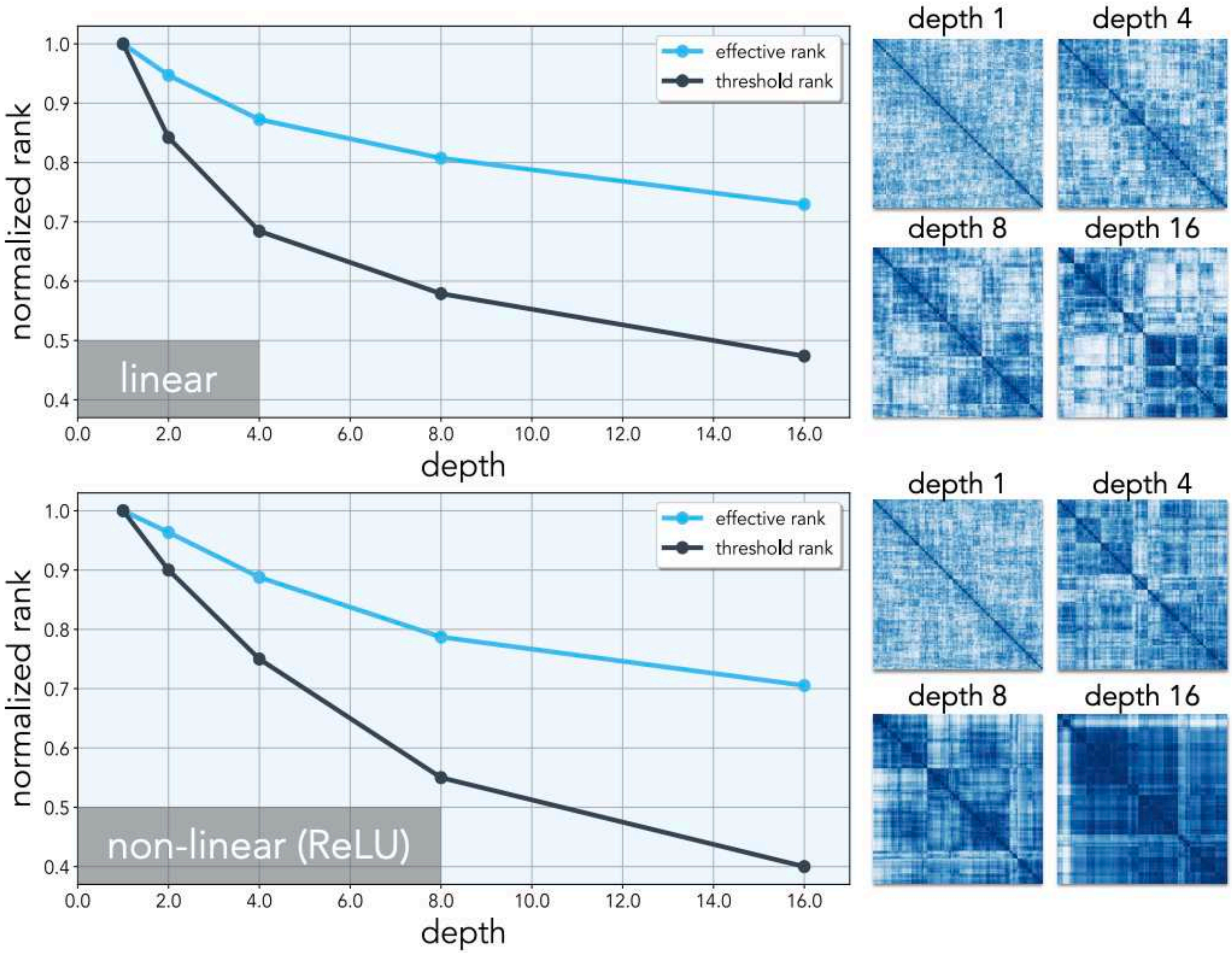


single-layer   two-layer

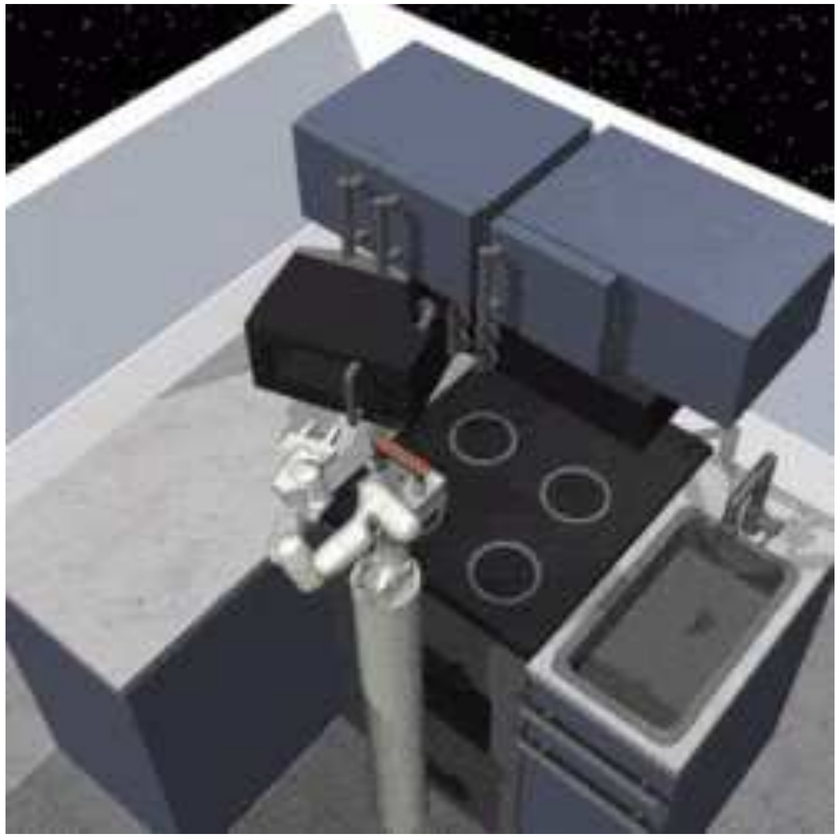
The Low Rank Simplicity Bias in Deep Neural Networks

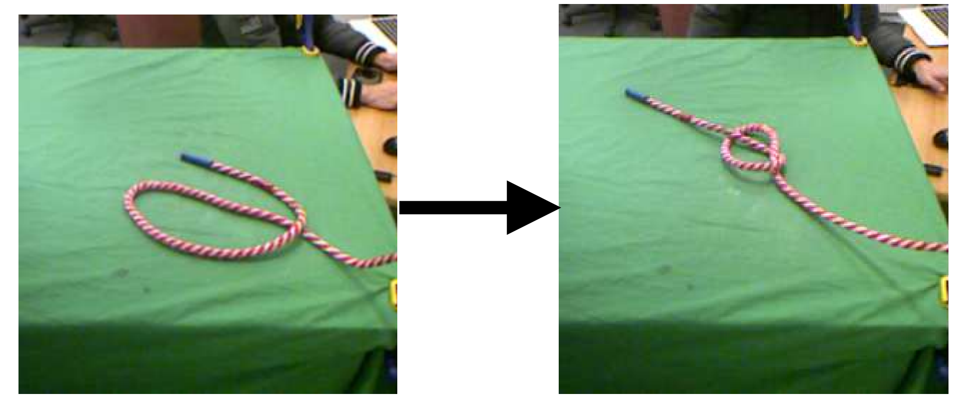
Huh et al. (in submission)

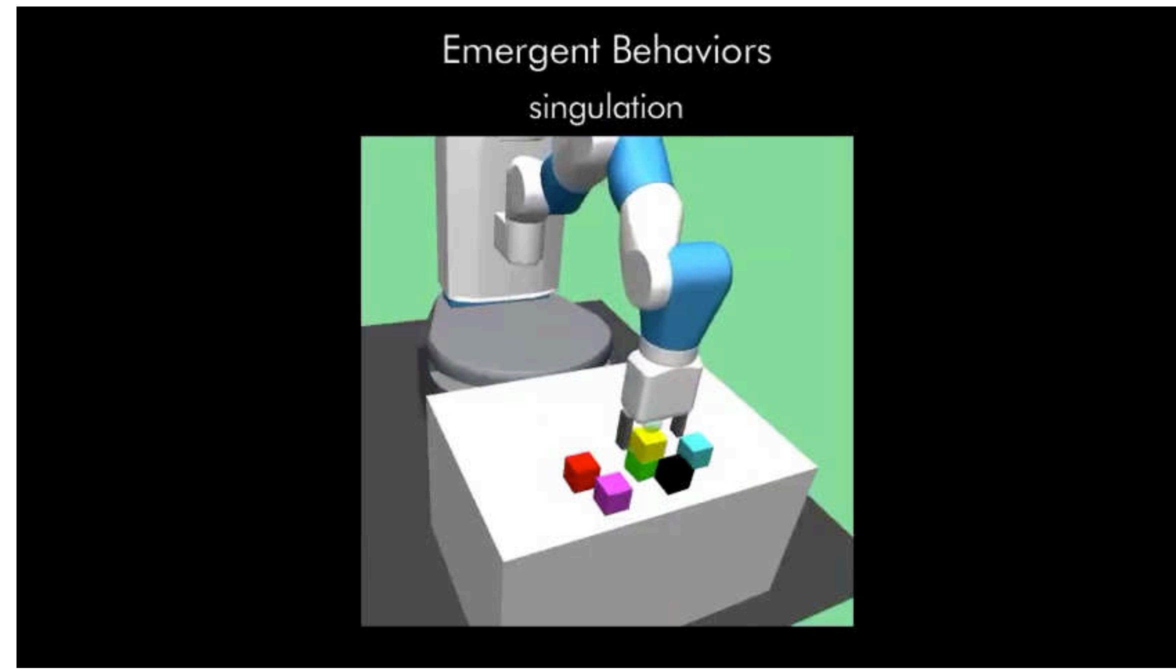
The Low Rank Simplicity Bias in Deep Neural Networks



Huh et al. (in submission)






Emergent Behaviors
singulation


MARIO
000000
×00
WORLD
1-1
TIME
381
Distance 306 (9%)

Questions?