EXPLORATION IN RL

Scott Niekum

Assistant Professor, Department of Computer Science
The University of Texas at Austin

TEXAS
The University of Texas at Austin

PeARL
Personal Autonomous Robotics Lab

Count-based exploration (Bellemare et al. 2016)

$$\mu_n(x) := \mu(x\,;\,x_{1:n}) := \frac{N_n(x)}{n}.$$

$$\rho'_n(x) = \Pr_\rho(X_{n+2} = x \mid X_1 \ldots X_n = x_{1:n}, X_{n+1} = x).$$

$$\rho_n(x) = \frac{\hat{N}_n(x)}{\hat{n}} \qquad \rho'_n(x) = \frac{\hat{N}_n(x) + 1}{\hat{n} + 1}.$$

$$\hat{N}_n(x) = \frac{\rho_n(x)(1 - \rho'_n(x))}{\rho'_n(x) - \rho_n(x)} = \hat{n}\rho_n(x).$$

Count-based exploration (Bellemare et al. 2016)

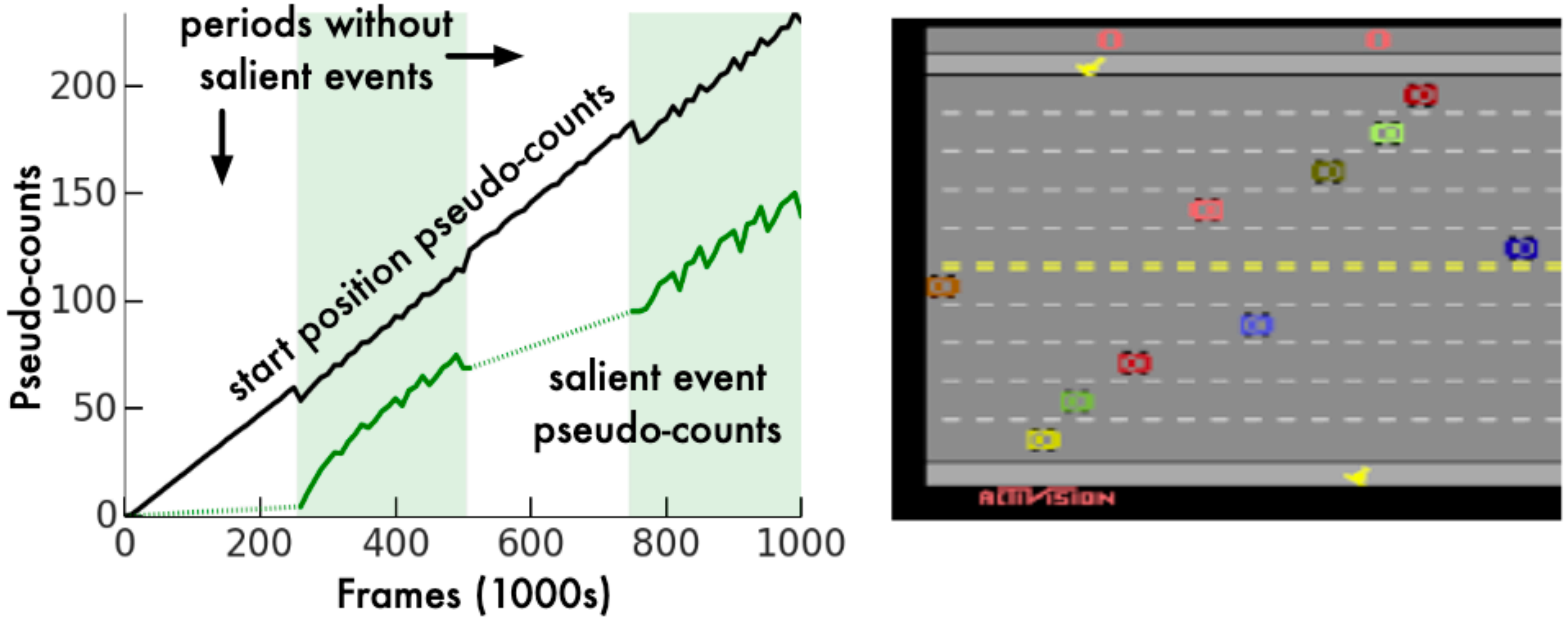
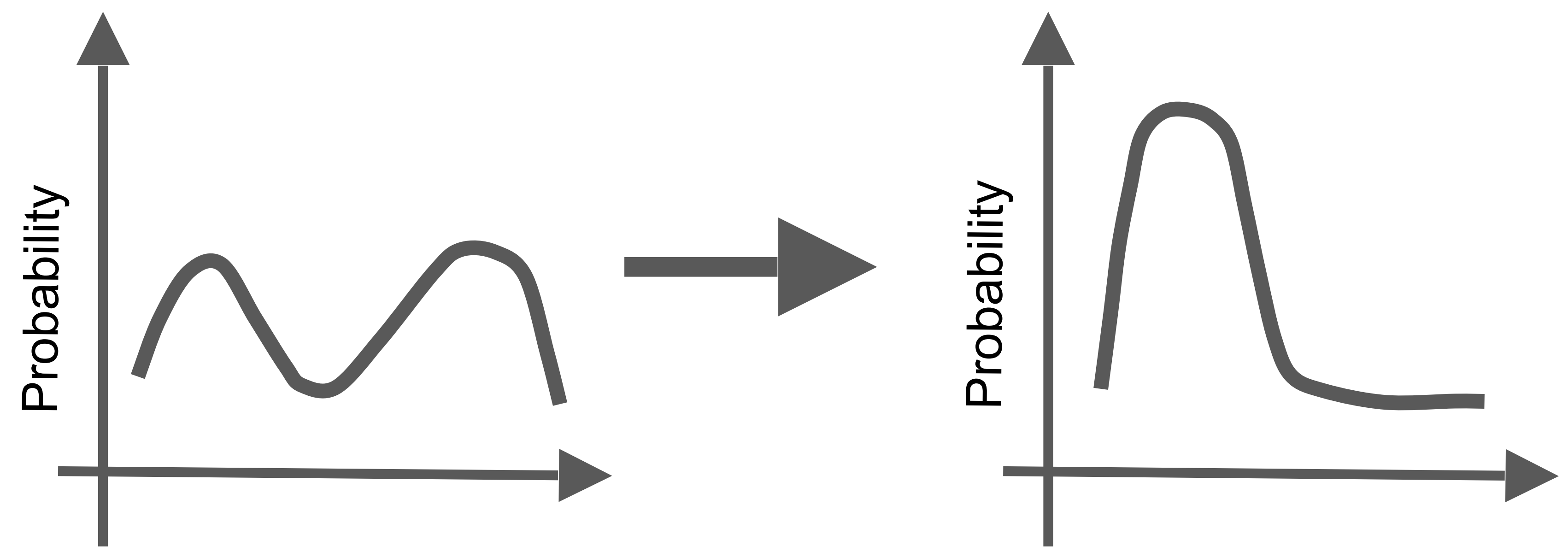


Figure 1: Pseudo-counts obtained from a CTS density model applied to FREEWAY, along with a frame representative of the salient event (crossing the road). Shaded areas depict periods during which the agent observes the salient event, dotted lines interpolate across periods during which the salient event is not observed. The reported values are 10,000-frame averages.

Count-based exploration (Bellemare et al. 2016)

Info gain: KL divergence between prior and posterior (in this case, of the density model) when observing new data

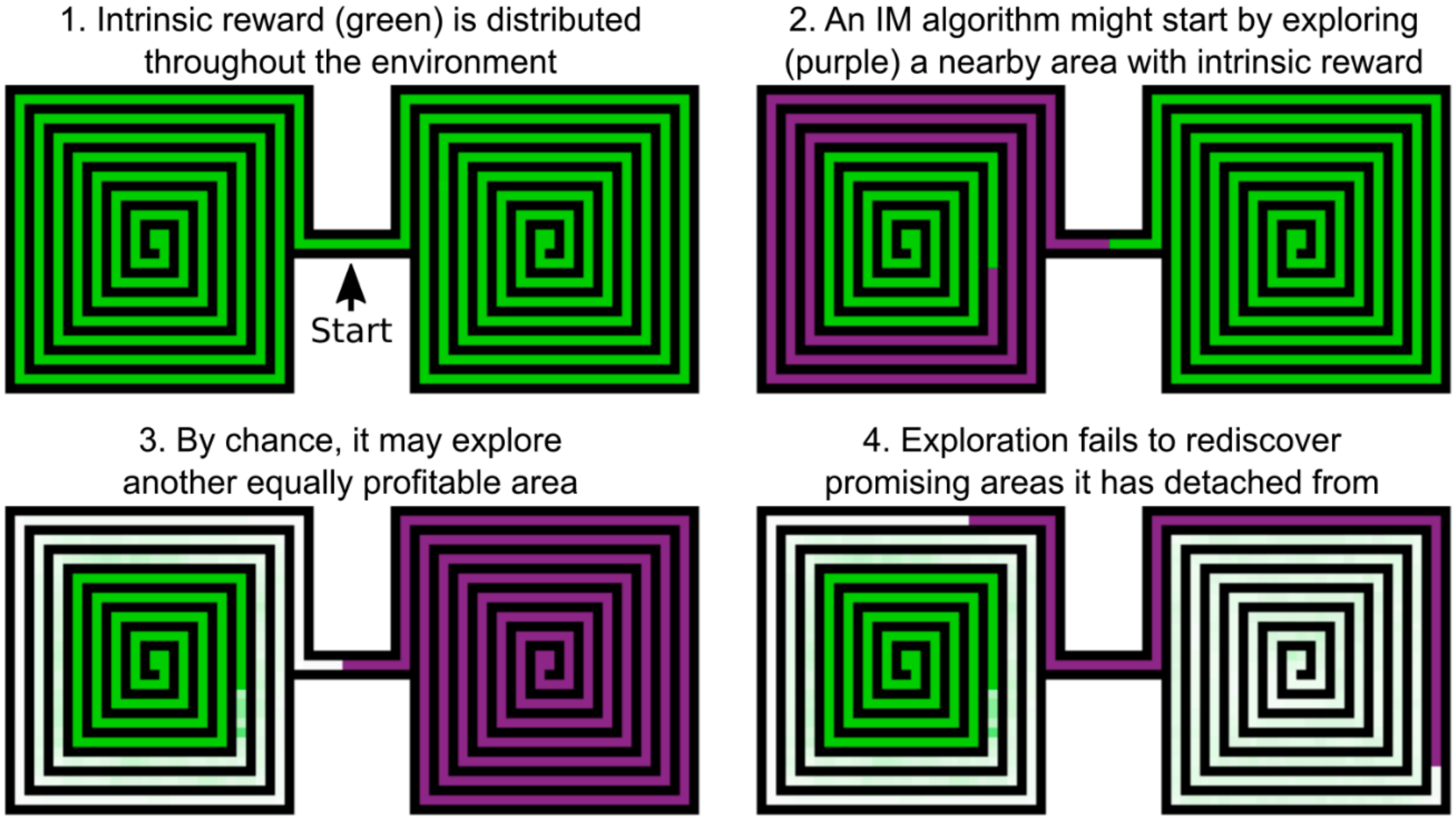
Intuitively: how much does the data change your beliefs?



$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \ln \left(\frac{P(x)}{Q(x)} \right)$$

A particular choice of pseudo count-based exploration bonus is at least as exploratory as computing a (usually intractable) information gain bonus!

Go-Explore (Ecoffet et al. 2019)



Go-Explore (Ecoffet et al. 2019)

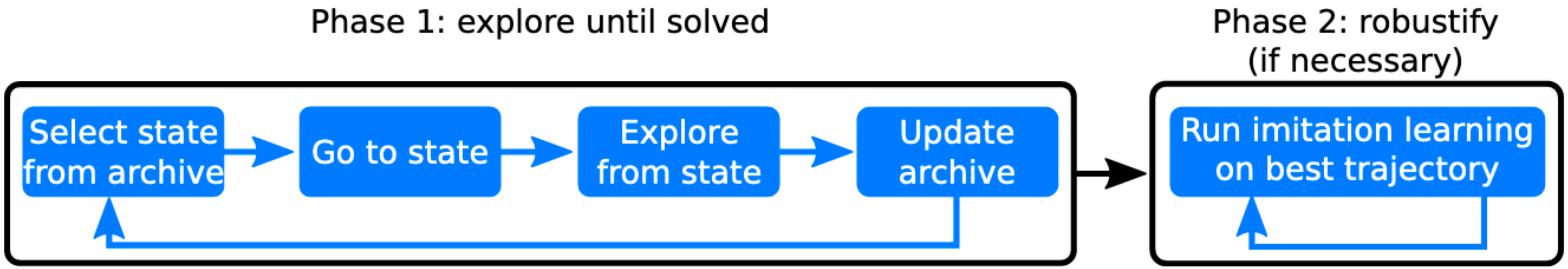


Figure 2: **A high-level overview of the Go-Explore algorithm.**

Where do rewards come from? (Singh et al.)

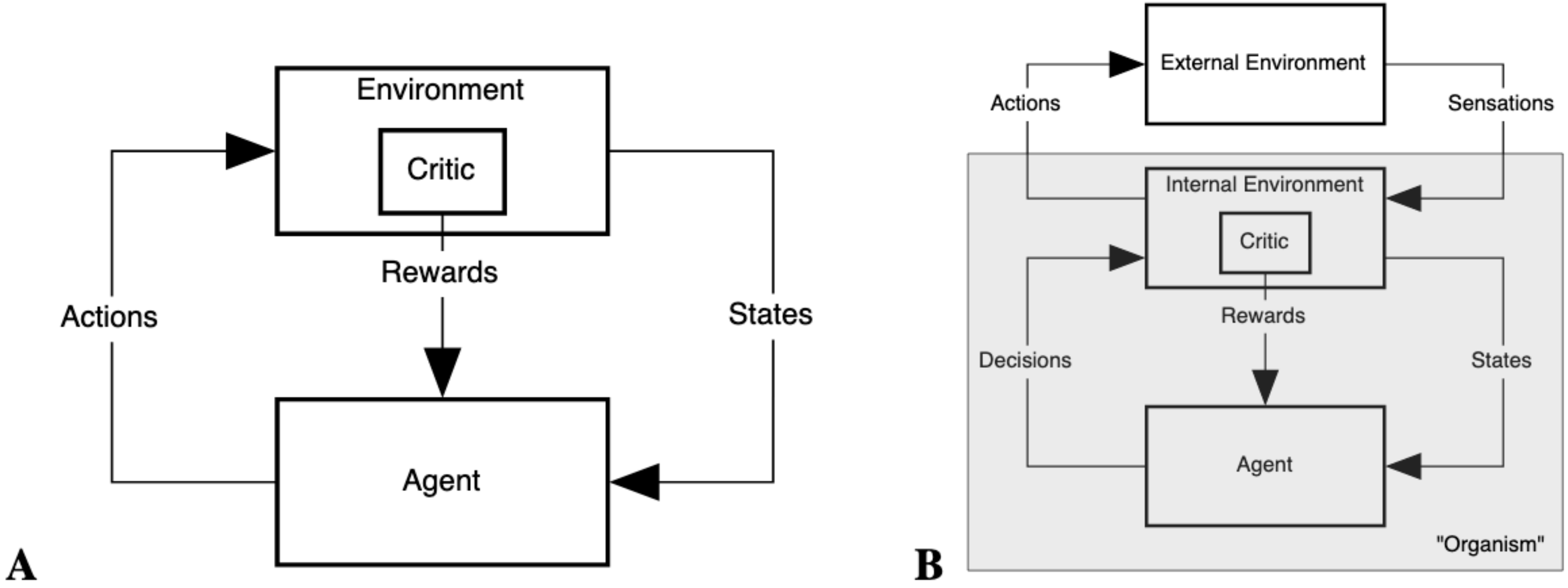


Figure 1: *Agent-Environment Interaction in RL*. **A**: *The usual view*. **B**: *An elaboration*.

What about evolved intrinsic rewards? (Niekum el al. 2010)

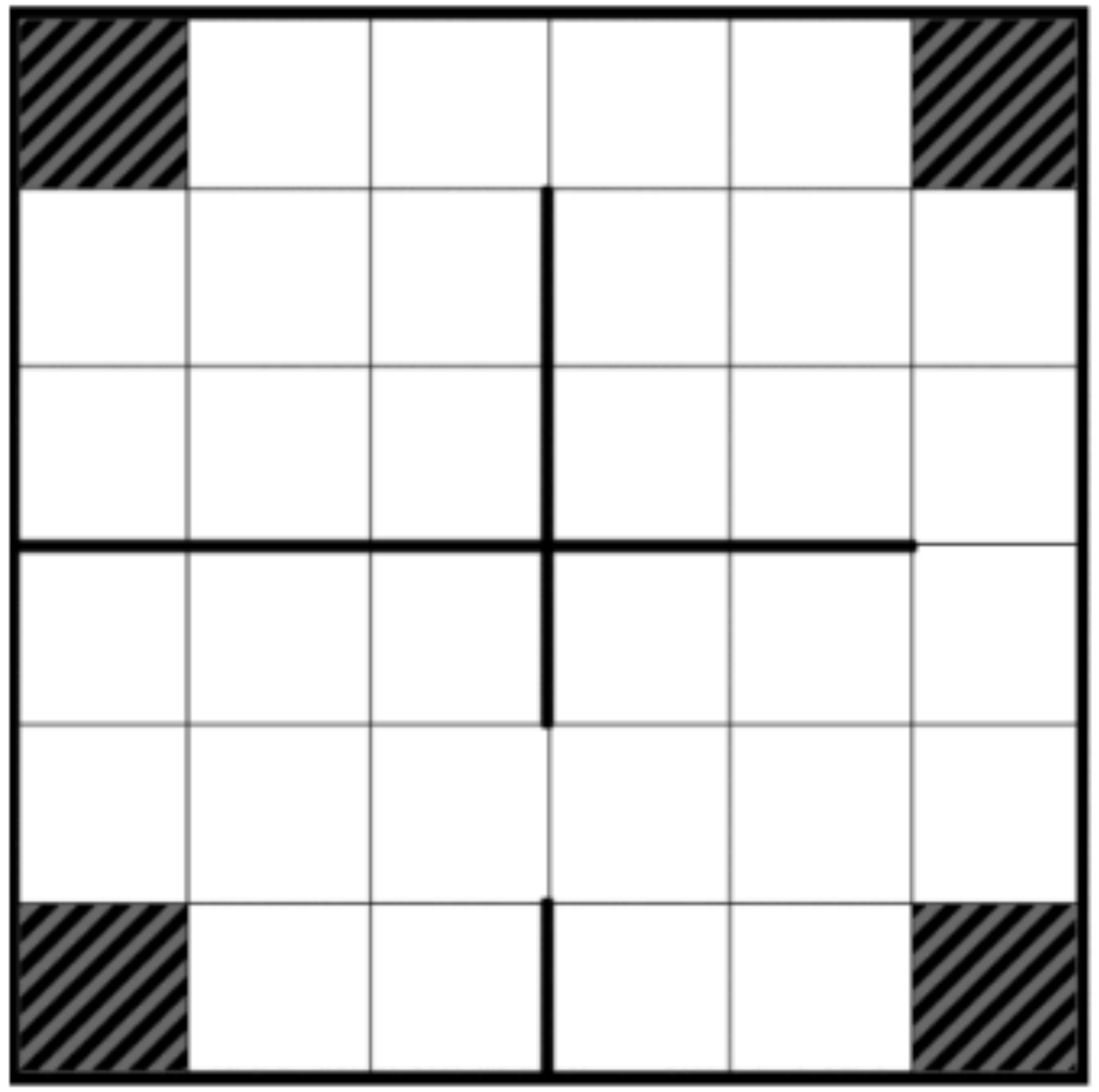


Fig. 1. Hungry–Thirsty domain. Thick lines are walls, striped squares denote possible food or water sites.

What about evolved intrinsic rewards? (Niekum el al. 2010)

Fig. 3. Evolved reward function from the Hungry–Thirsty domain.

STATE2 4.5 FLOAT.% STATE2 FLOAT.% FLOAT.DUP STATE1 STATE2 FLOAT.% FLOAT.+ FLOAT.* STATE1 FLOAT.- 0.5 FLOAT.+

What about evolved intrinsic rewards? (Niekum el al. 2010)

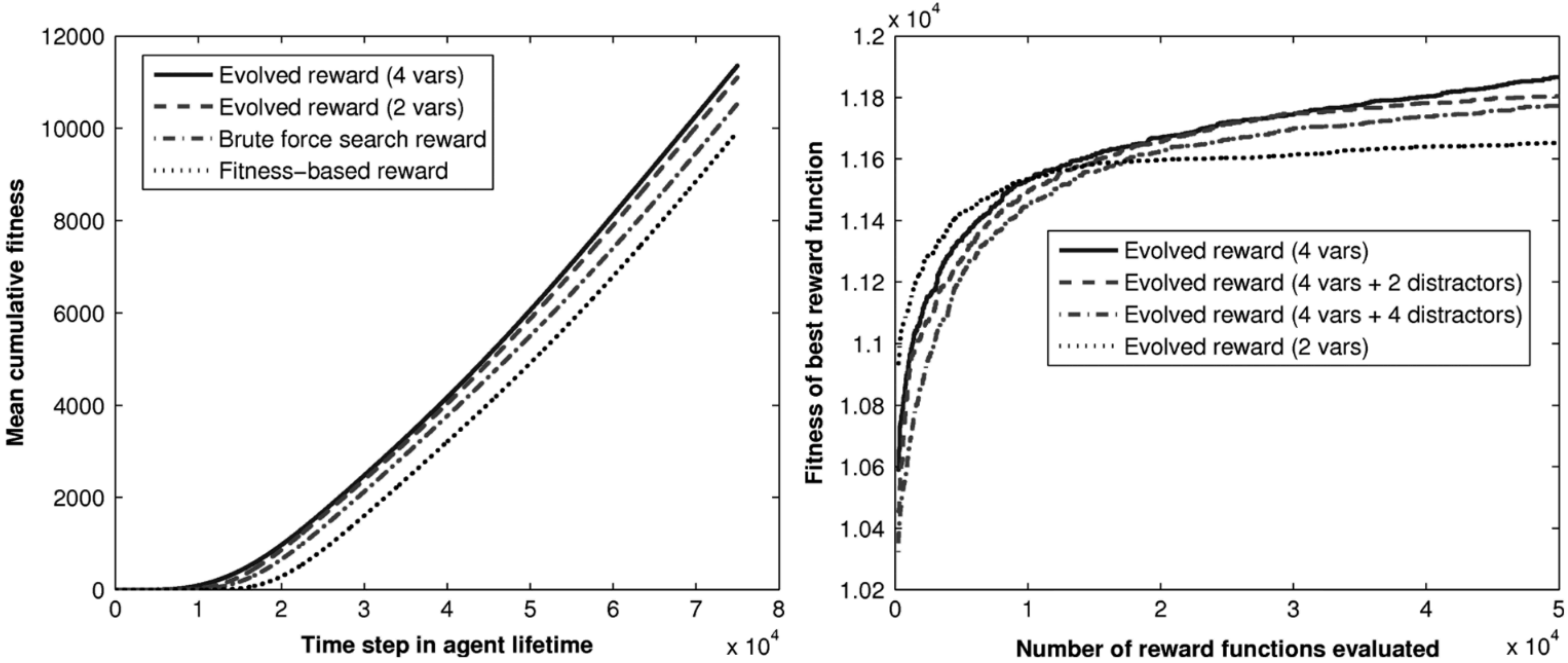


Fig. 2. Agent fitness (left) and evolutionary progress (right) over a distribution of environments.

What about evolved intrinsic rewards? (Niekum el al. 2010)

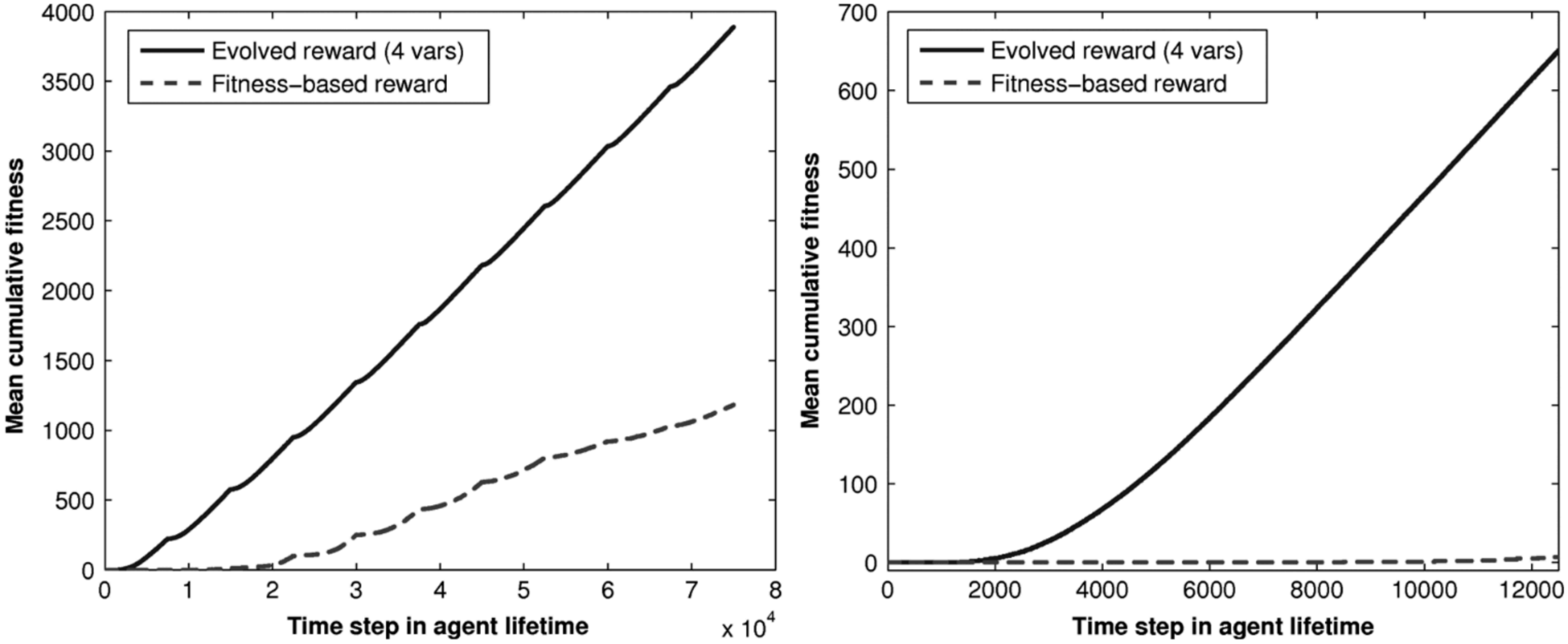
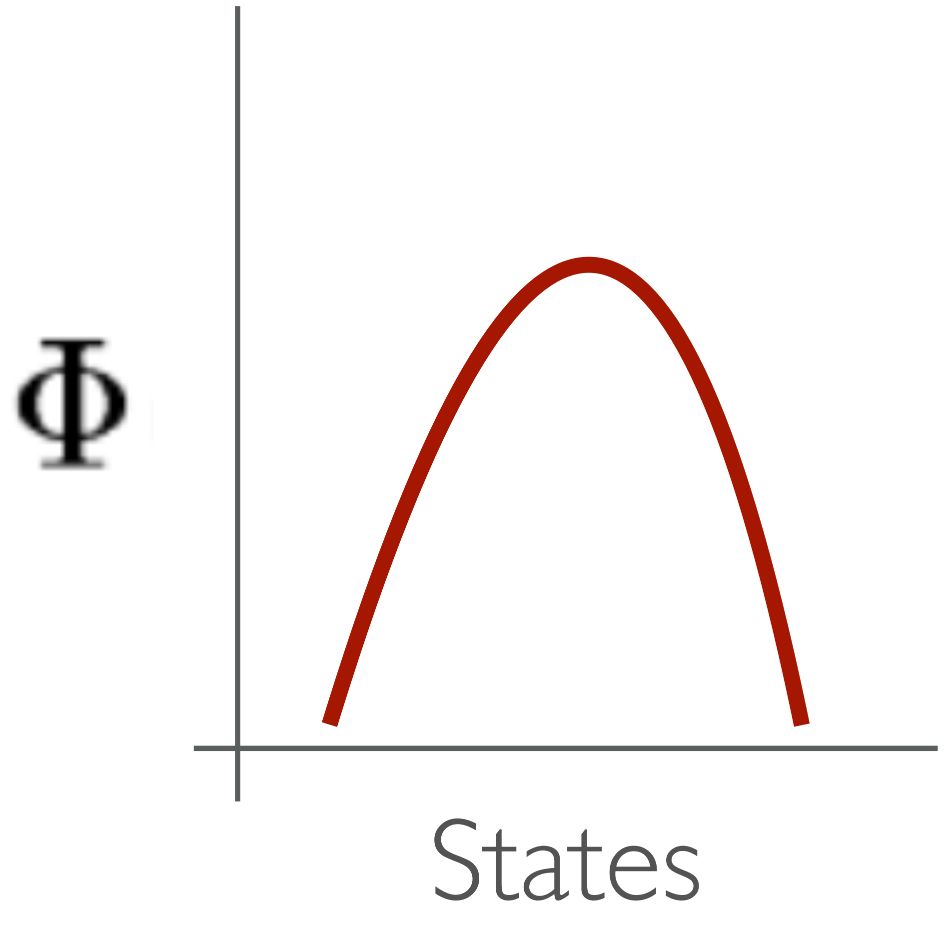


Fig. 4. Agent fitness on nonstationary (left) and short lifetime (right) problems.

Potential-based shaping rewards (Ng 1999)

$$F(s, a, s') = \gamma\Phi(s') - \Phi(s)$$



Alternate idea: is it possible to explore by generating and testing causal hypotheses about the world?

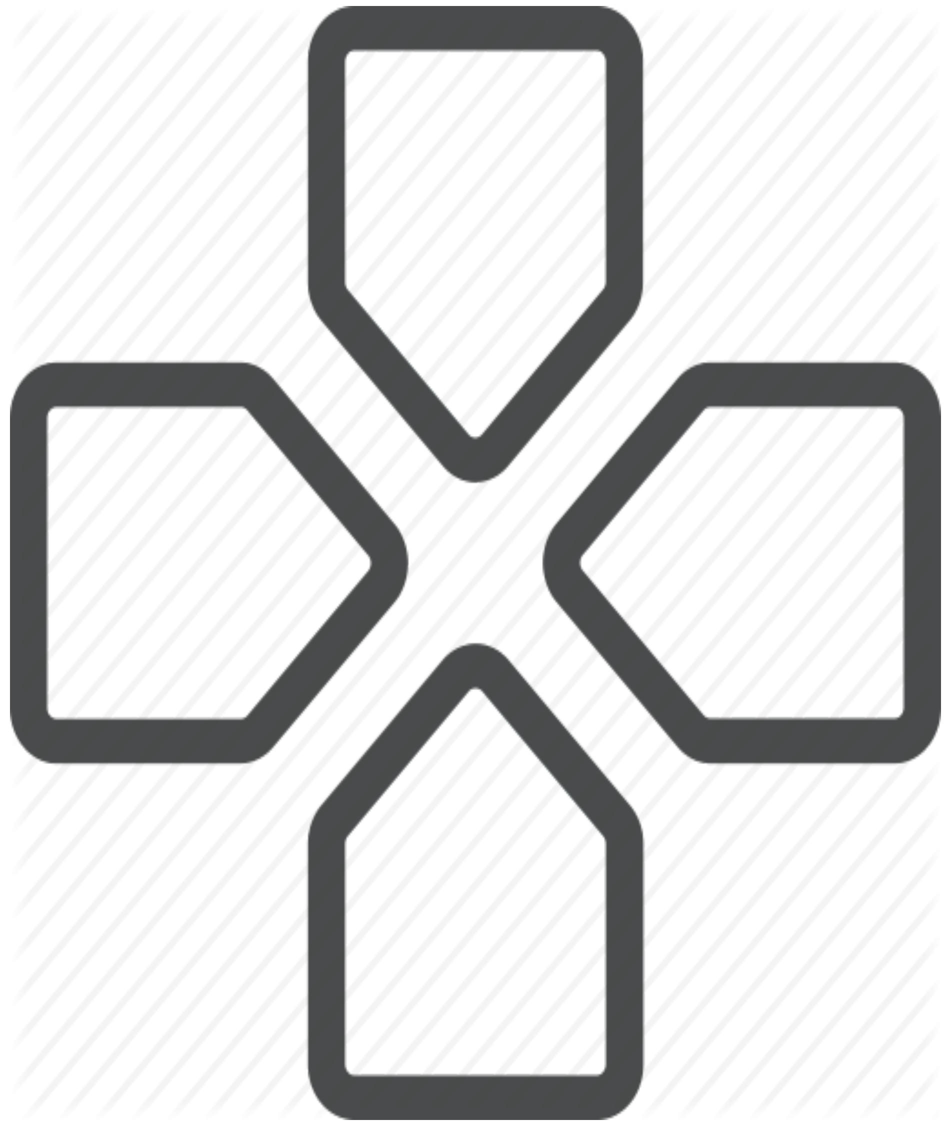
Simplifying assumptions:

- Objects: spatially localized, permanence, temporal coherence, static appearances and properties
- Quasi-static assumption: Objects have properties such as position or velocity that do not change unless acted upon
- Proximity assumption: Objects cannot interact unless they make contact

Two Ways to improve Exploration:

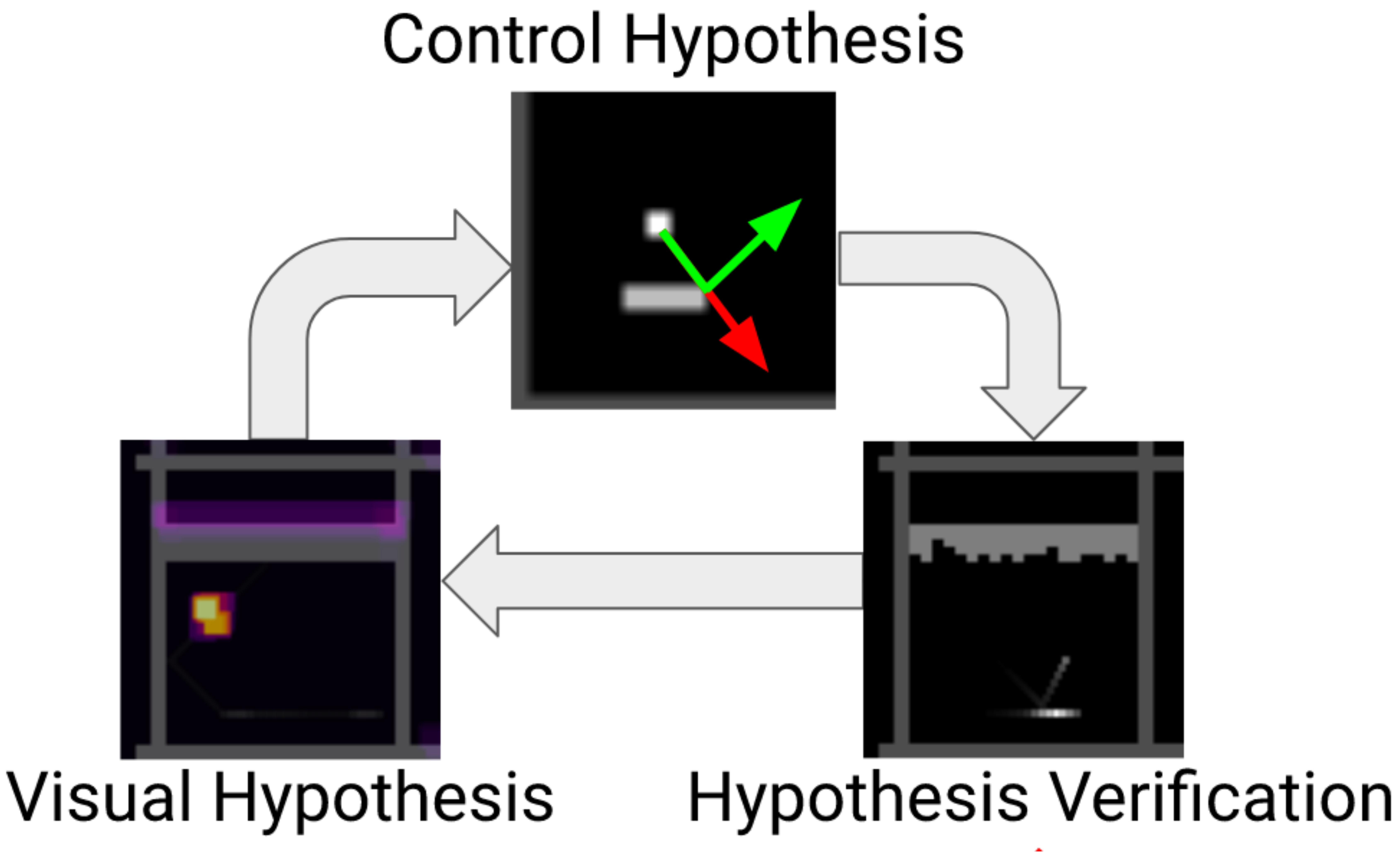
Hypothesis driven

Control directed



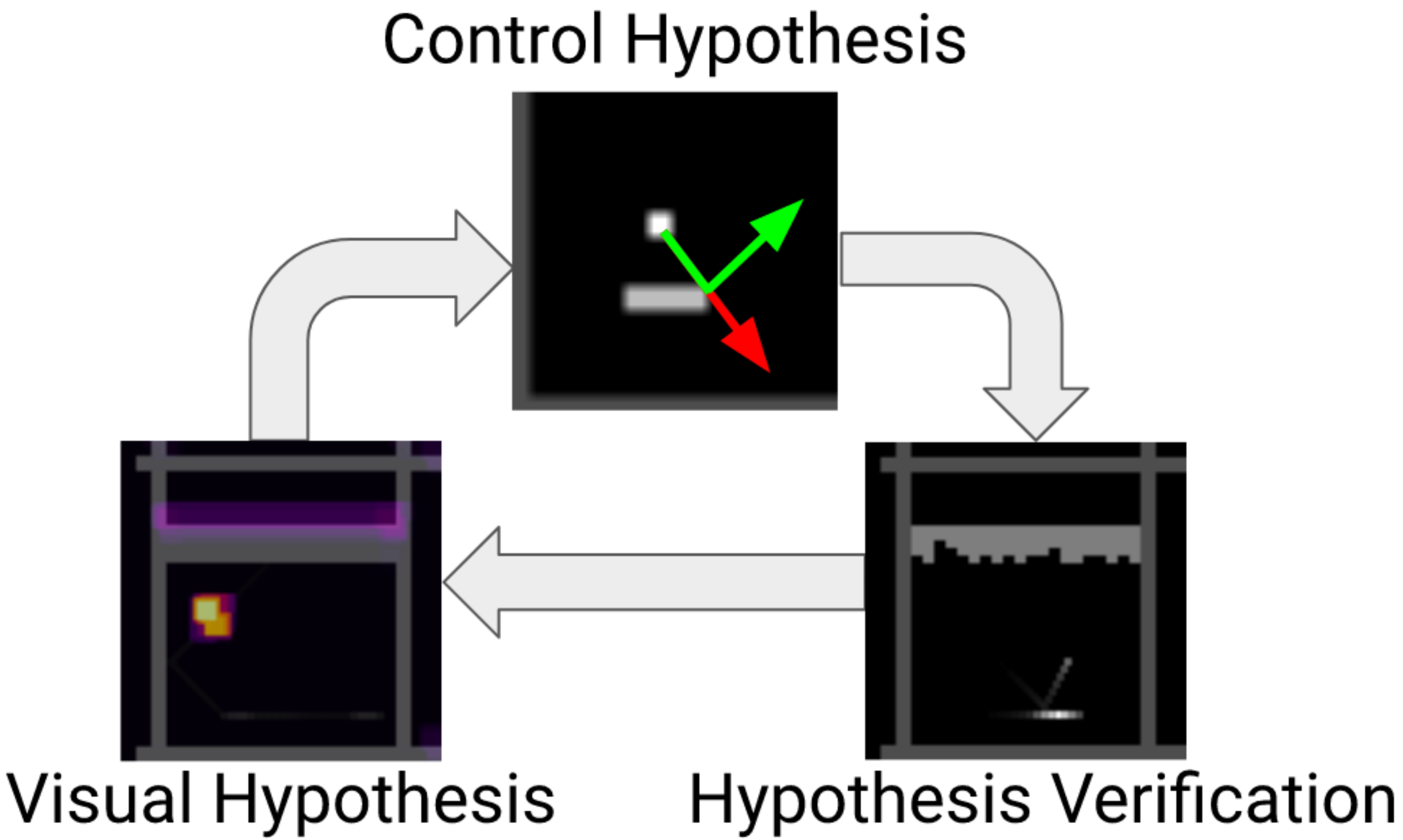
Core idea:

Given a hypothesized interaction between two objects, verify if a relation exists by learning to control that interaction



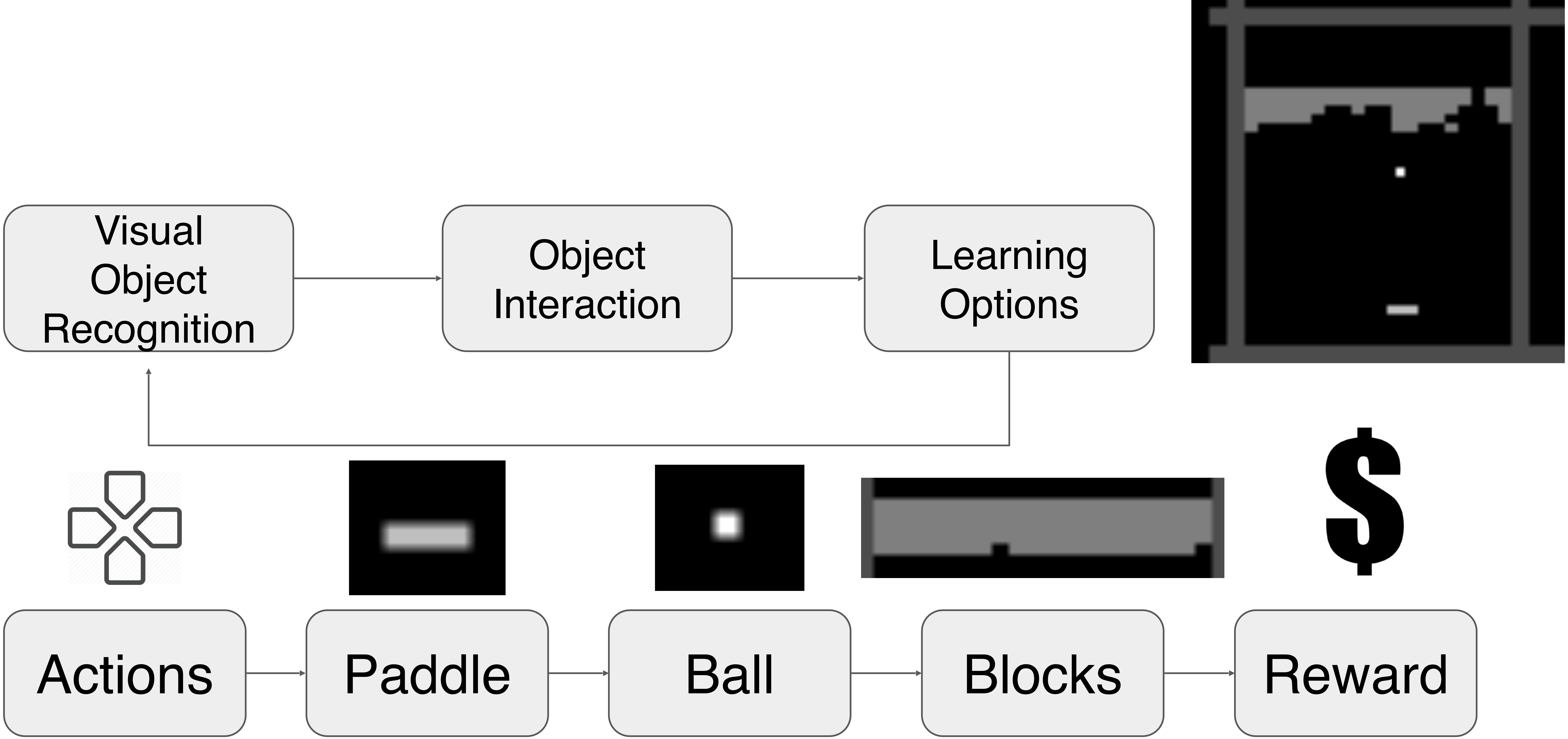
Control Hypothesis
Visual Hypothesis
Hypothesis Verification

Detect changepoints in dynamics

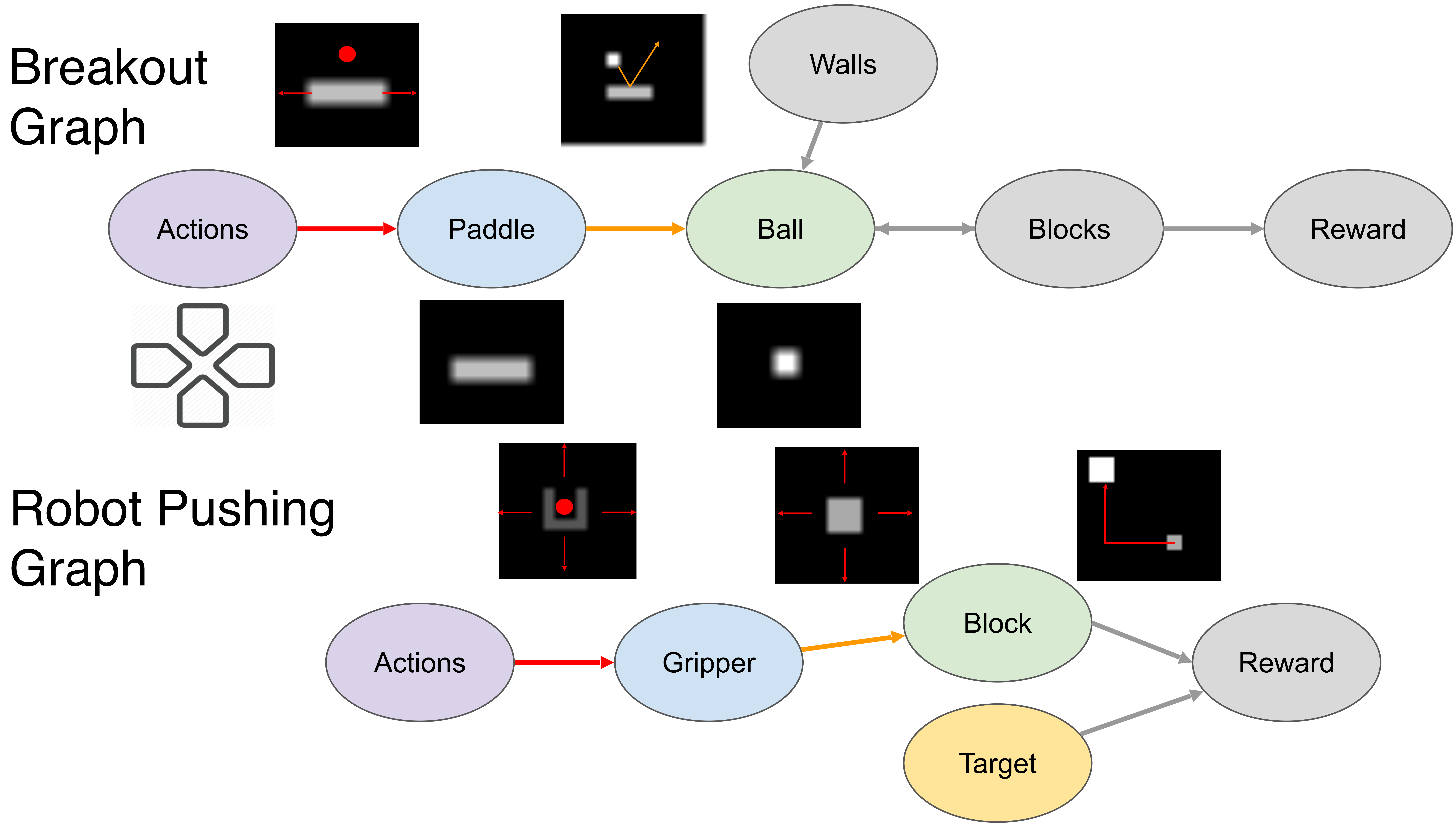


Discover convolutional filters that behave like objects

Learn options with goal of creating particular changepoints



Visual Object Recognition
Object Interaction
Learning Options
Actions
Paddle
Ball
Blocks
Reward



Breakout Graph
Walls
Actions
Paddle
Ball
Blocks
Reward
Robot Pushing Graph
Actions
Gripper
Block
Reward
Target

Breakout Training Curves