Hierarchical Deep Reinforcement Learning

Tejas D Kulkarni
Karthik Narasimhan
Ardavan Saeedi
Joshua Tenenbaum

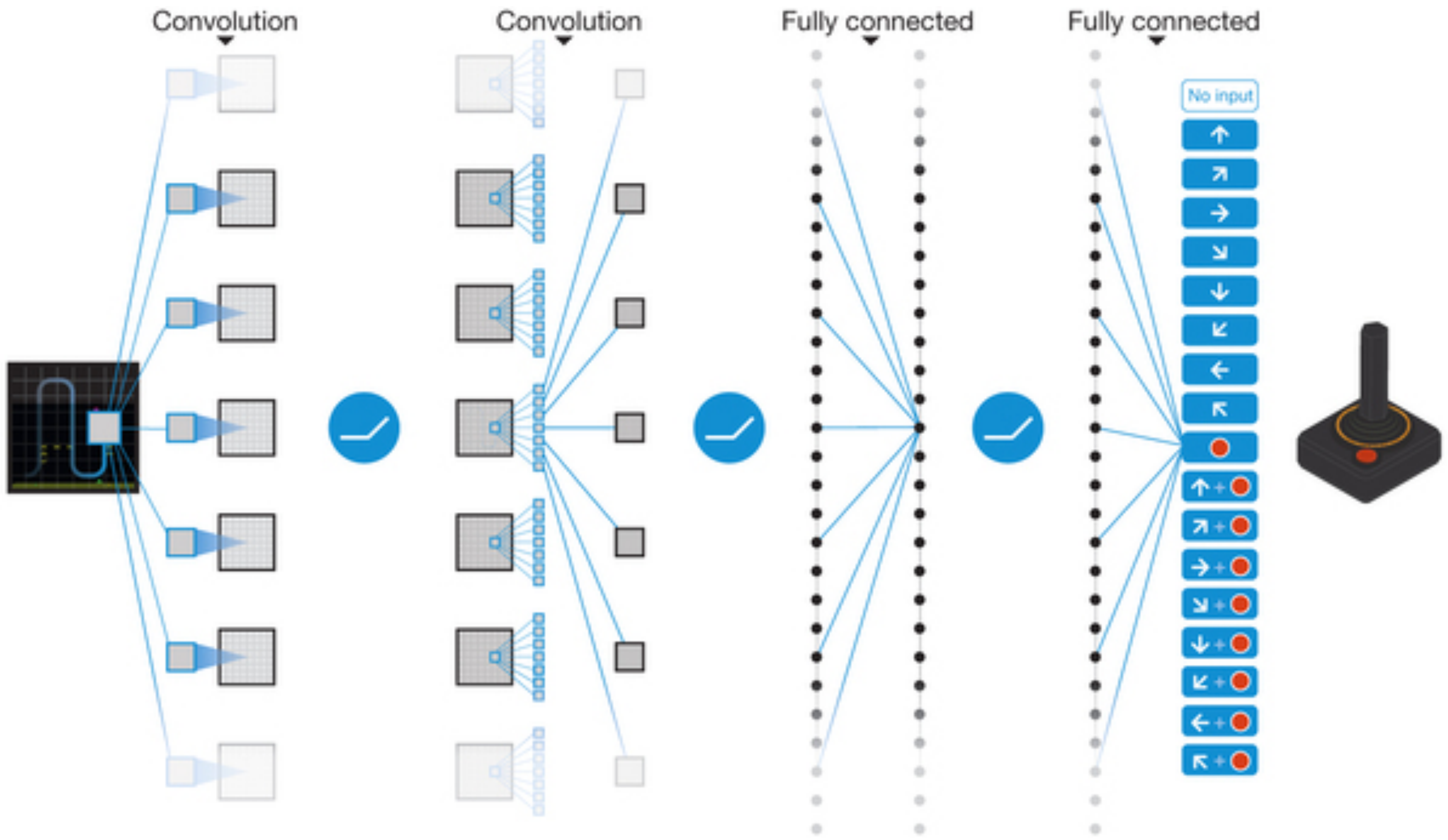
Brain and Cognitive Sciences, CSAIL
Massachusetts Institute of Technology

ICML 2016

Deep Reinforcement Learning = DL + RL

Deep Reinforcement Learning = DL + RL

Input: Raw Pixels **Output:** Actions



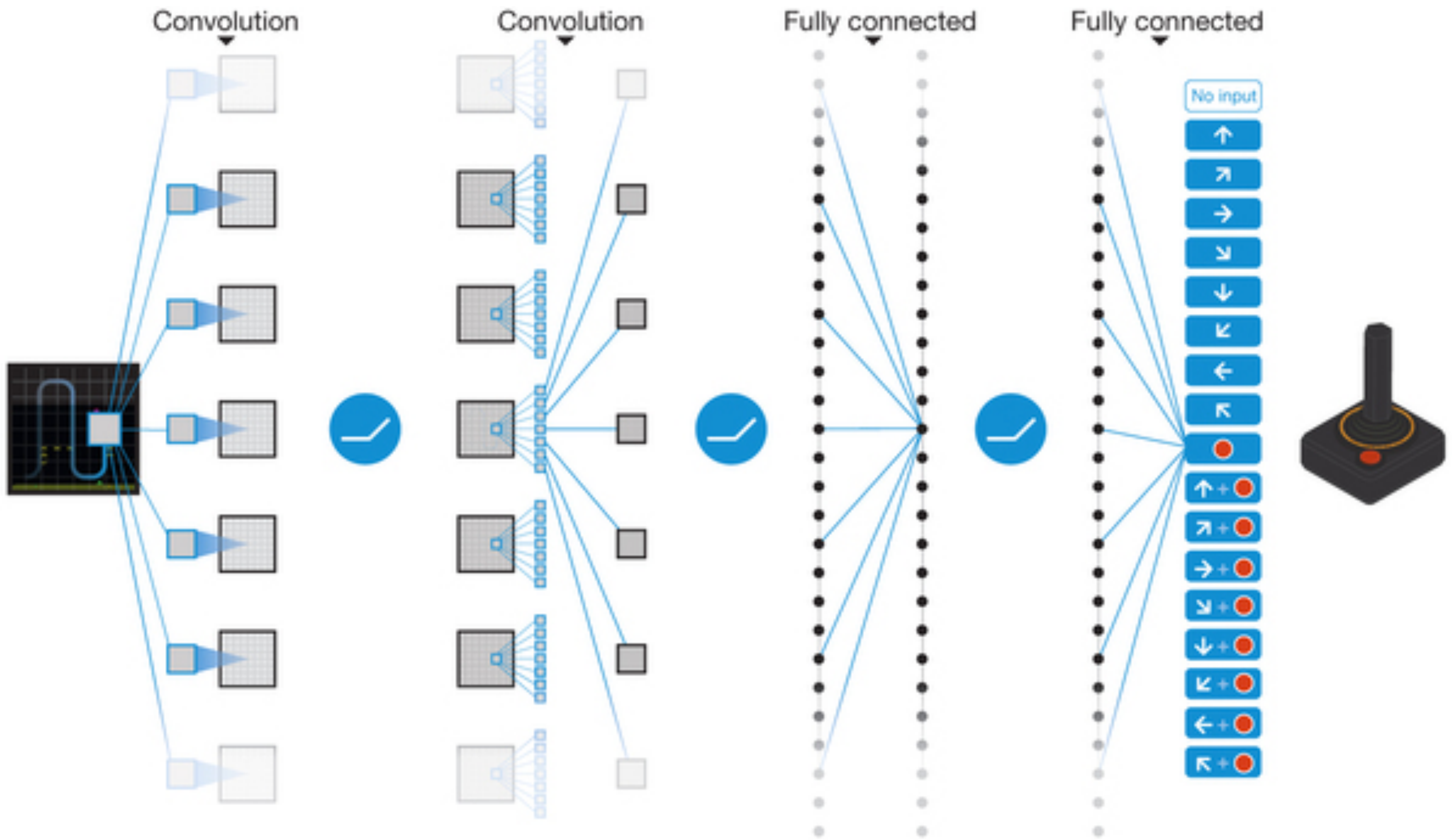
Mnih et al. Nature '15

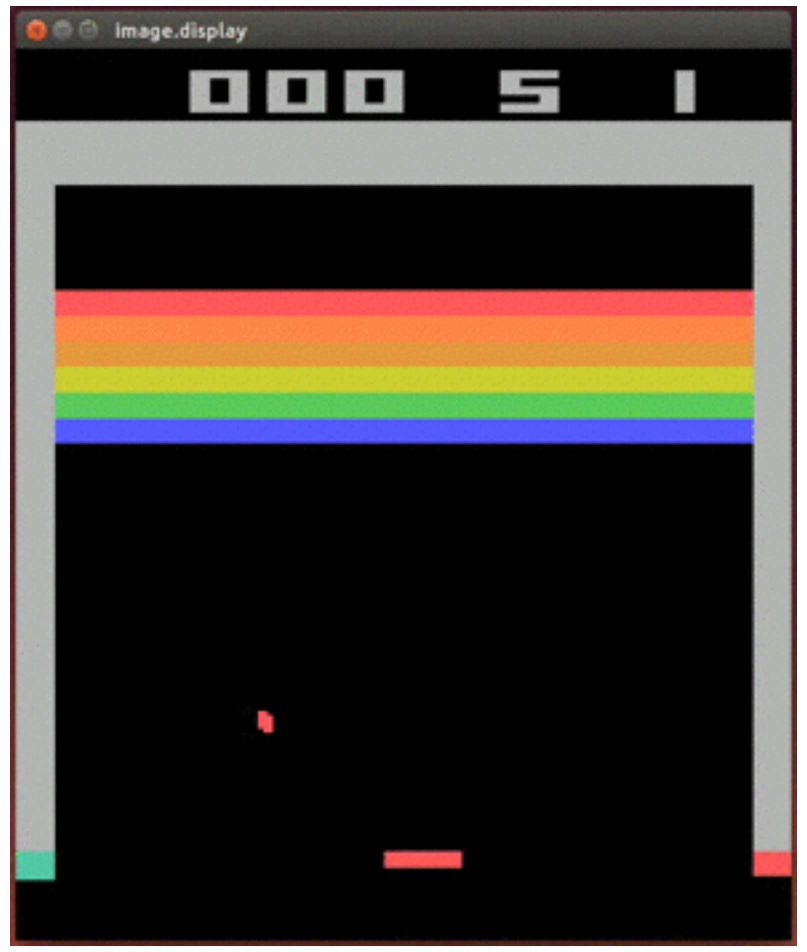
Earlier deep RL variants:
Koutnik et al., Online Evolution of Deep Convolutional Network for Vision-Based Reinforcement Learning. 2013

Hausknecht et al., A Neuroevolution Approach to General Atari Game Playing. 2013

Deep Reinforcement Learning = DL + RL

Input: Raw Pixels **Output:** Actions





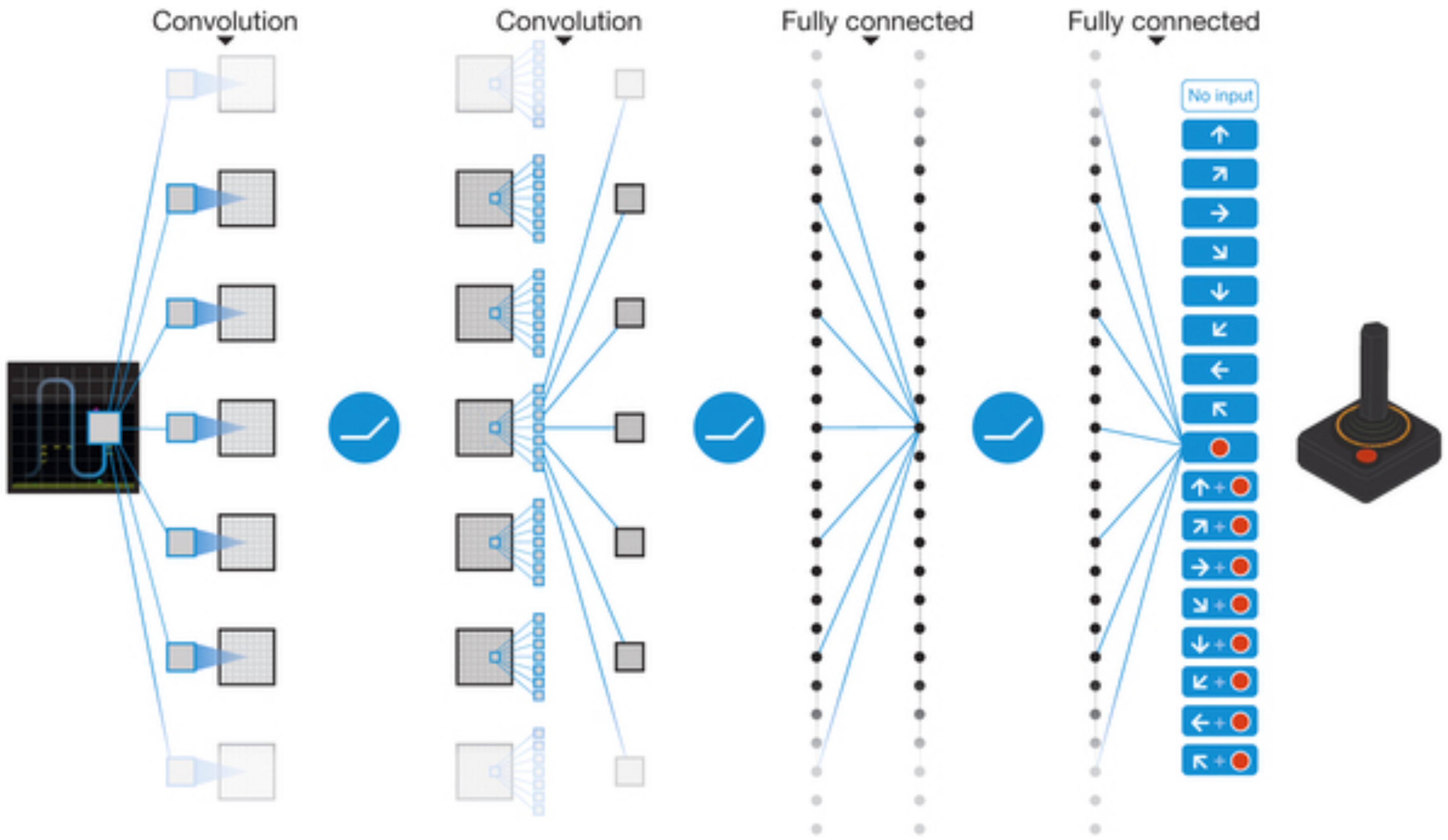
Mnih et al. Nature '15

Earlier deep RL variants:
Koutnik et al., Online Evolution of Deep Convolutional Network for Vision-Based Reinforcement Learning. 2013

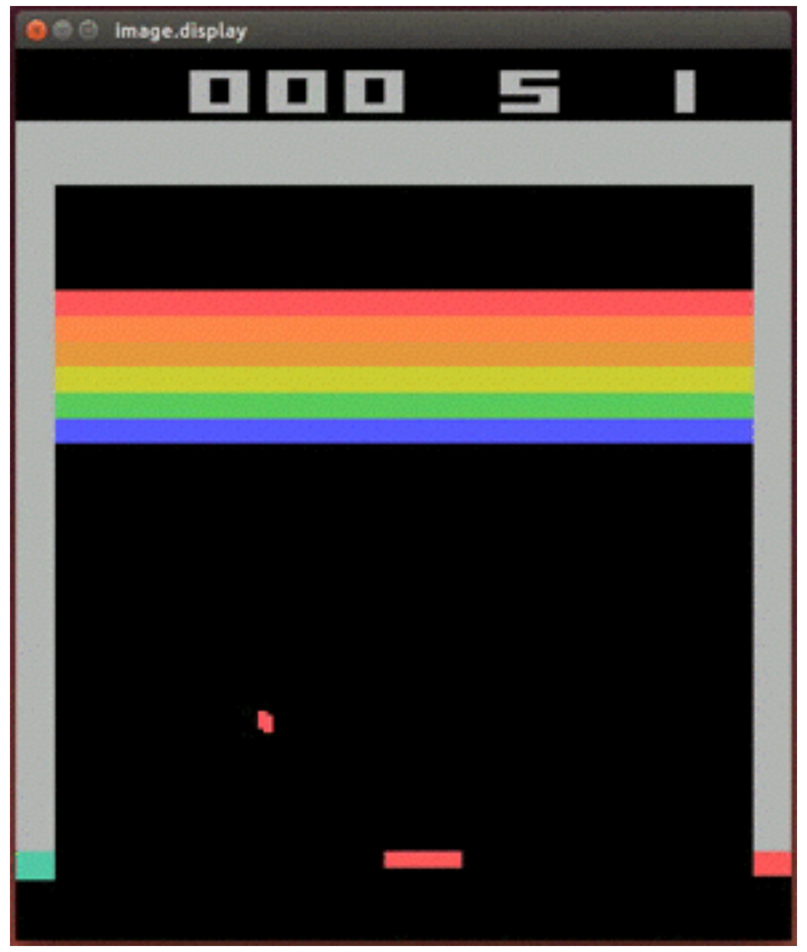
Hausknecht et al., A Neuroevolution Approach to General Atari Game Playing. 2013

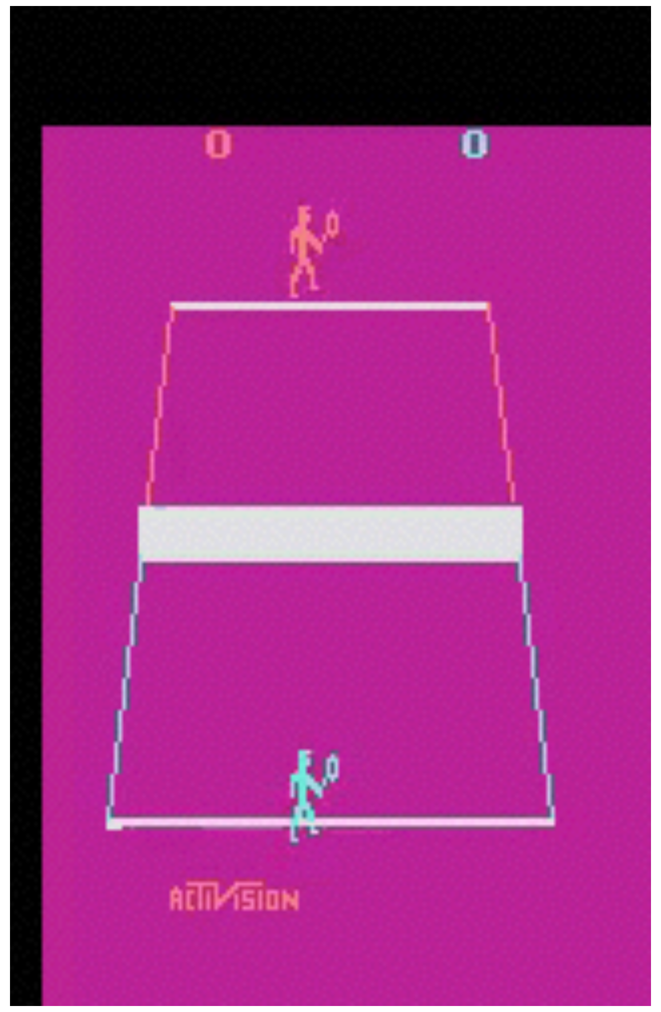
Deep Reinforcement Learning = DL + RL

Input: Raw Pixels **Output:** Actions



Mnih et al. Nature '15



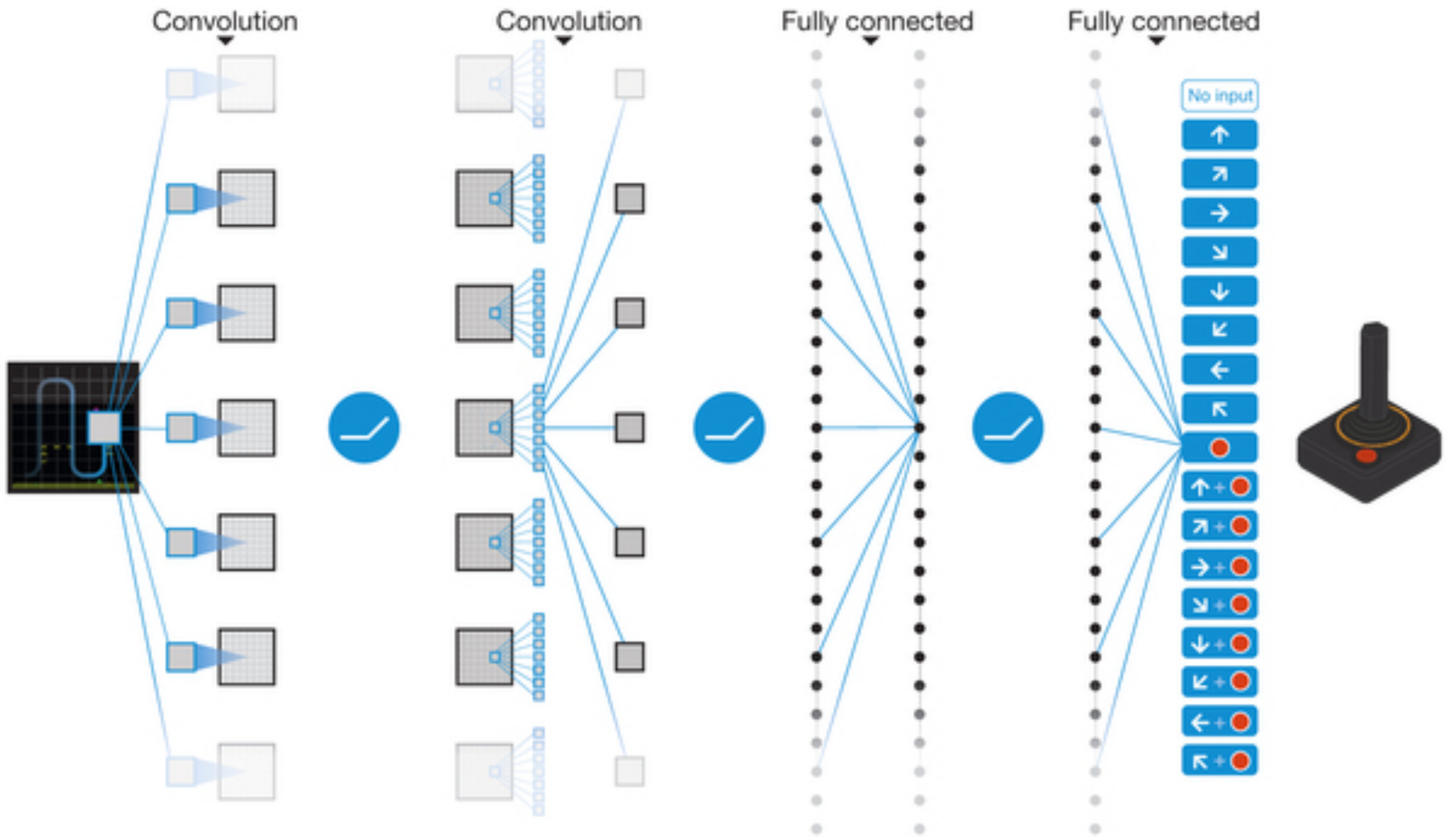


Earlier deep RL variants:
Koutnik et al., Online Evolution of Deep Convolutional Network for Vision-Based Reinforcement Learning. 2013

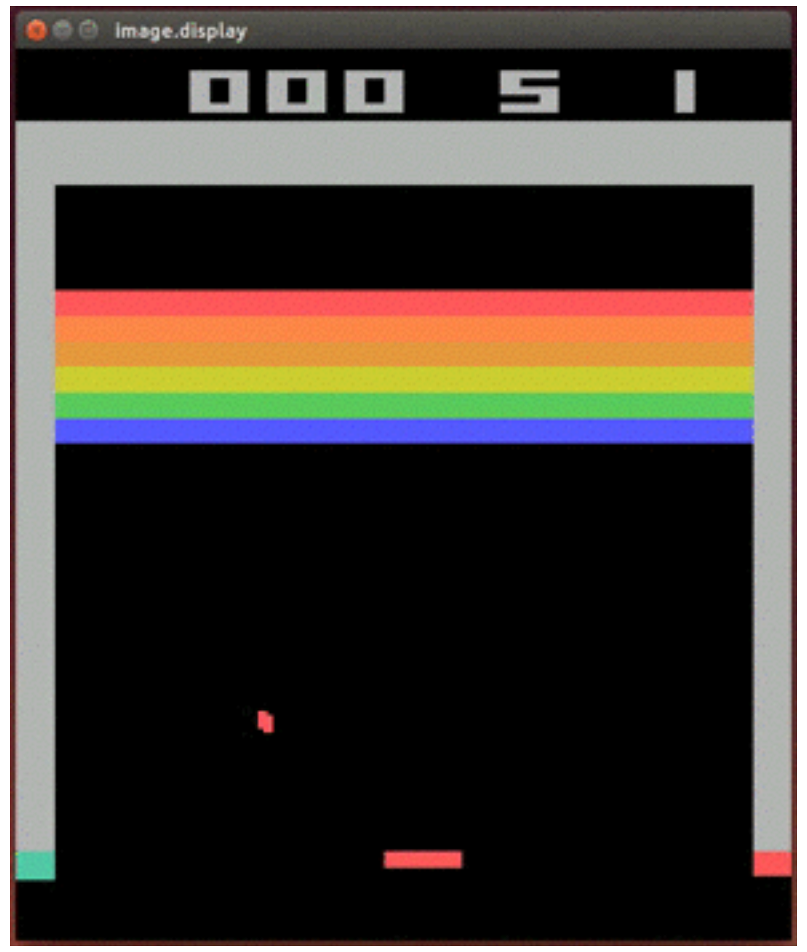
Hausknecht et al., A Neuroevolution Approach to General Atari Game Playing. 2013

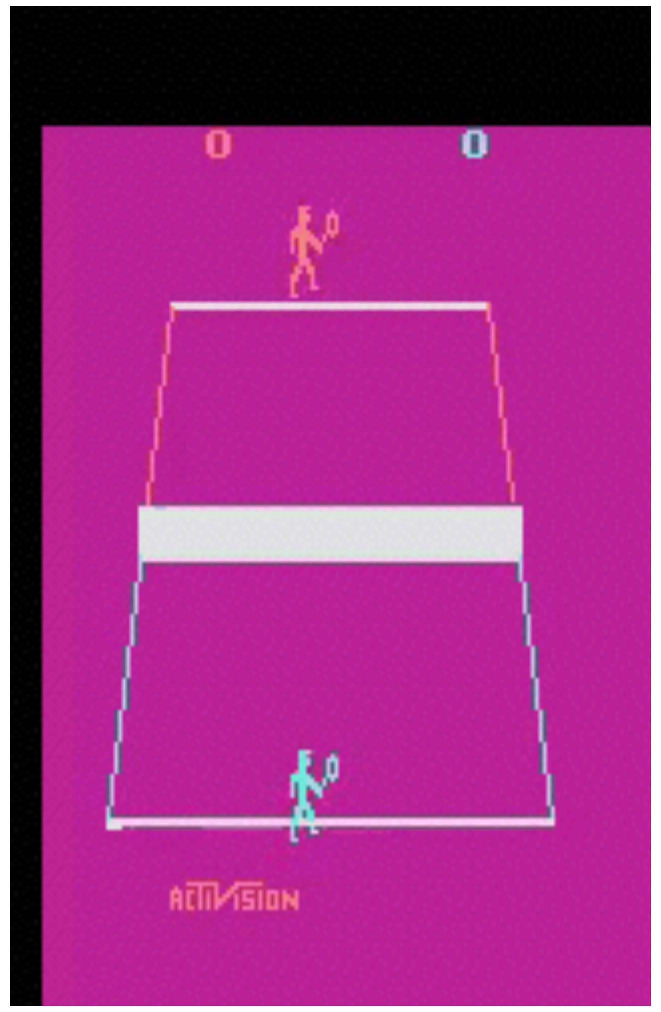
Deep Reinforcement Learning = DL + RL

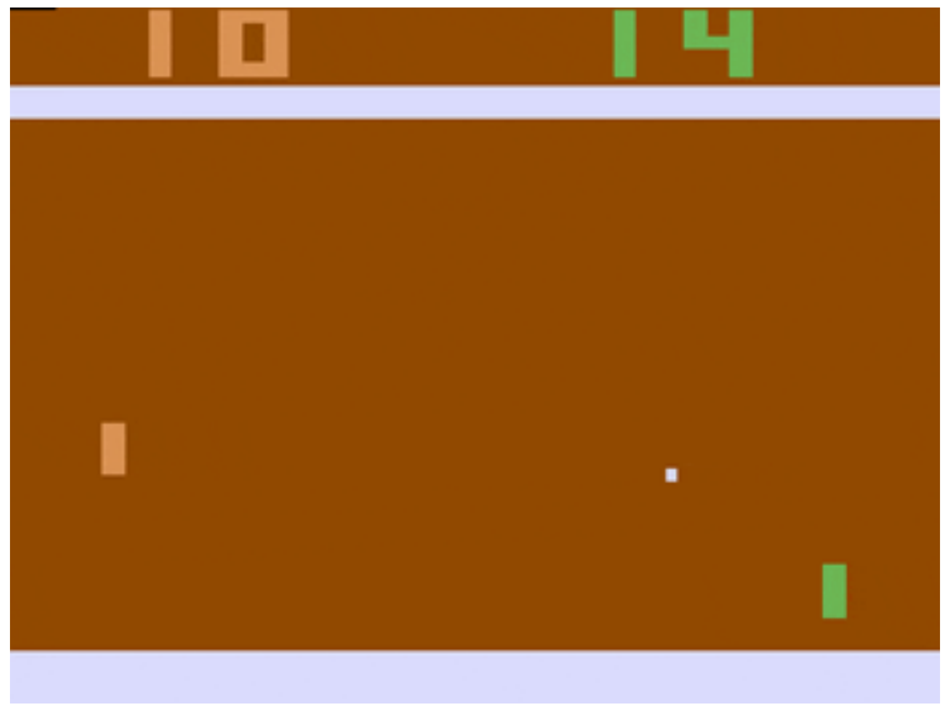
Input: Raw Pixels **Output:** Actions



Mnih et al. Nature '15





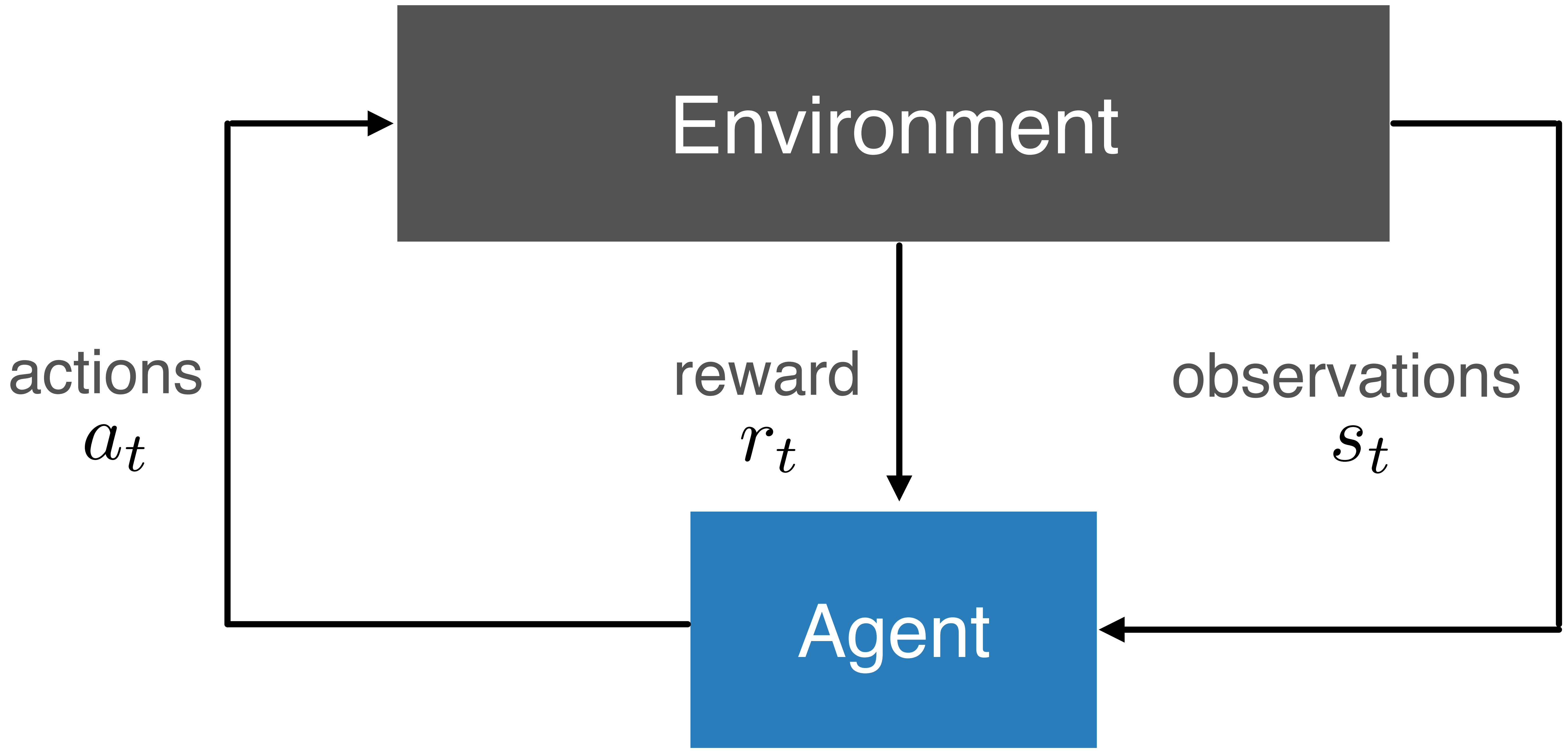


Earlier deep RL variants:

Koutnik et al., Online Evolution of Deep Convolutional Network for Vision-Based Reinforcement Learning. 2013

Hausknecht et al., A Neuroevolution Approach to General Atari Game Playing. 2013

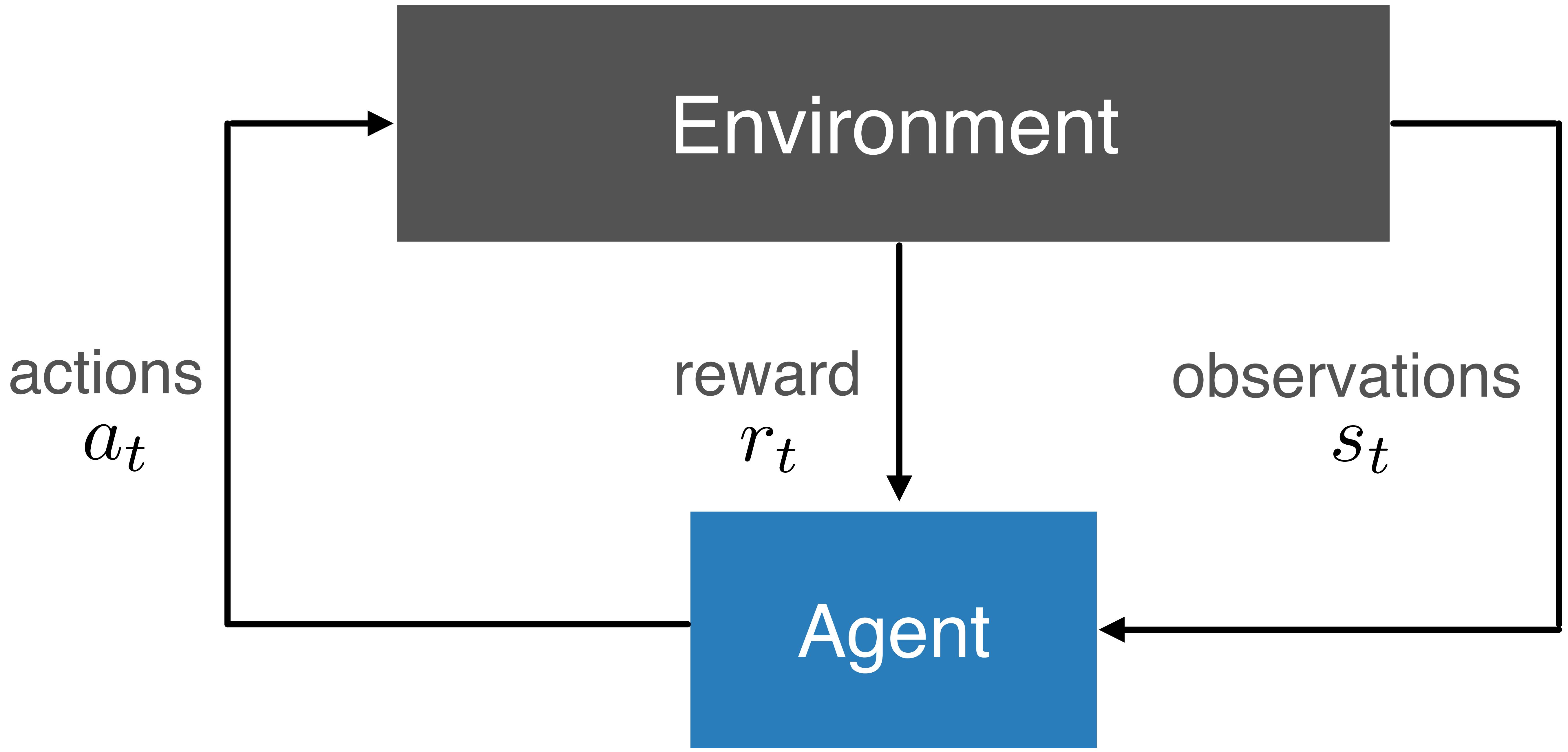
Reinforcement Learning



Reinforcement Learning

Agent's policy

$$\pi(a_t | s_0, r_0, a_0, s_1, r_1, a_1, ...)$$



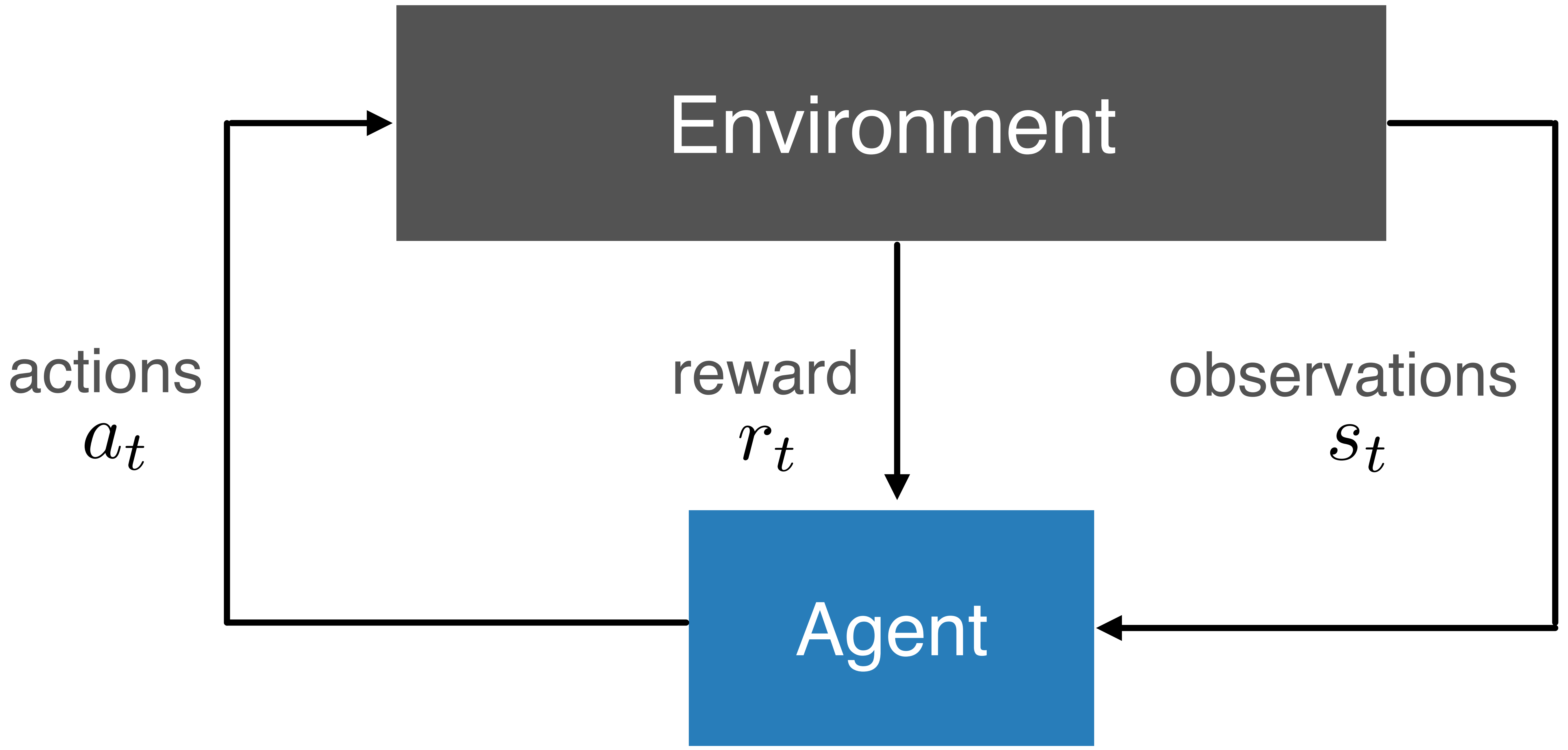
Reinforcement Learning

Agent's policy

$$\pi(a_t | s_0, r_0, a_0, s_1, r_1, a_1, ...)$$

Agent's Life

$$\{s_0, r_0, a_0, s_1, r_1, a_1, s_2, r_2, a_2, ...\}$$



Reinforcement Learning

Agent's policy

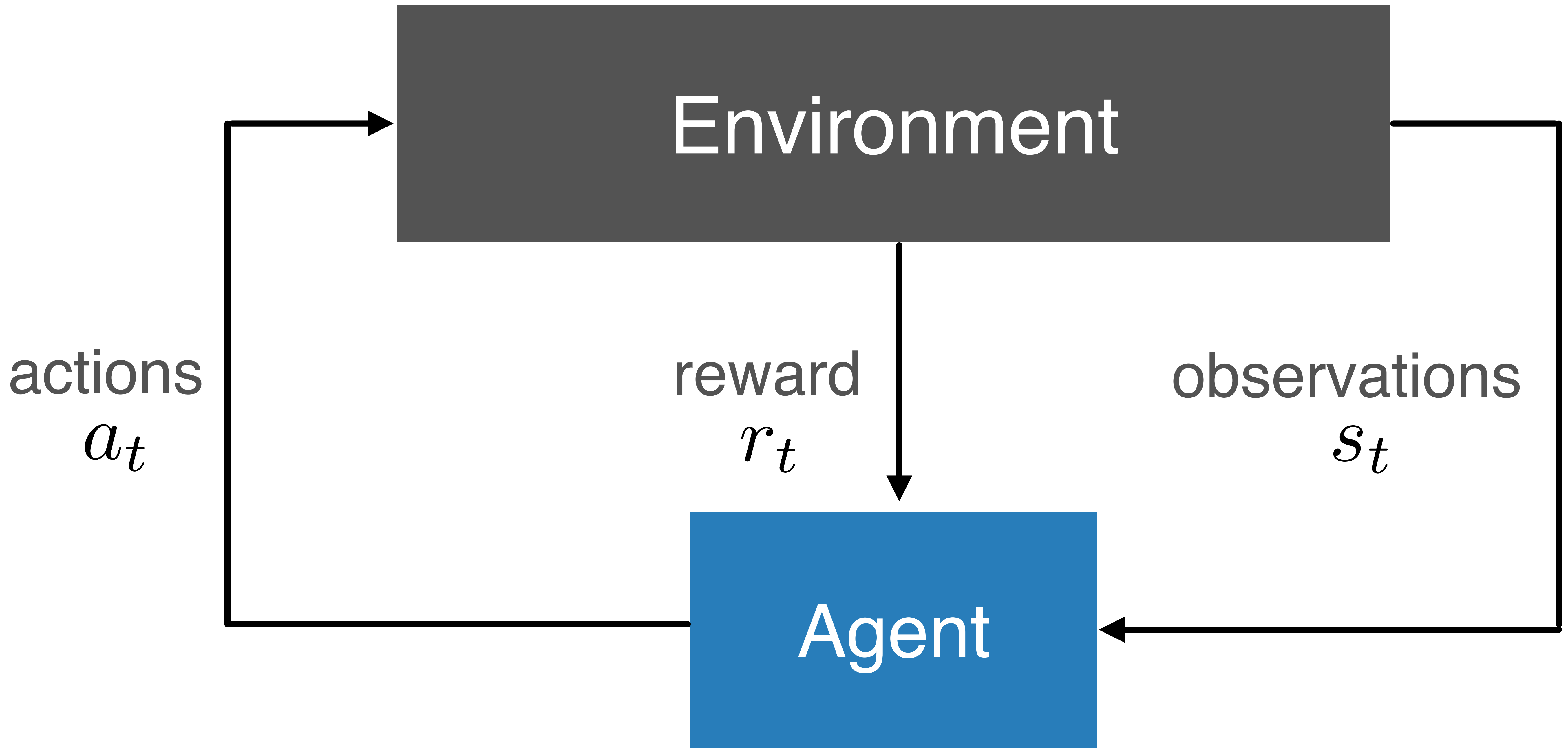
$$\pi(a_t|s_0, r_0, a_0, s_1, r_1, a_1, ...)$$

Agent's Life

$$\{s_0, r_0, a_0, s_1, r_1, a_1, s_2, r_2, a_2, ...\}$$

Objective

$$V_\mu^\pi = \mathbb{E}(\sum_{i=0}^{\infty} r_i)$$



Reinforcement Learning

Agent's policy

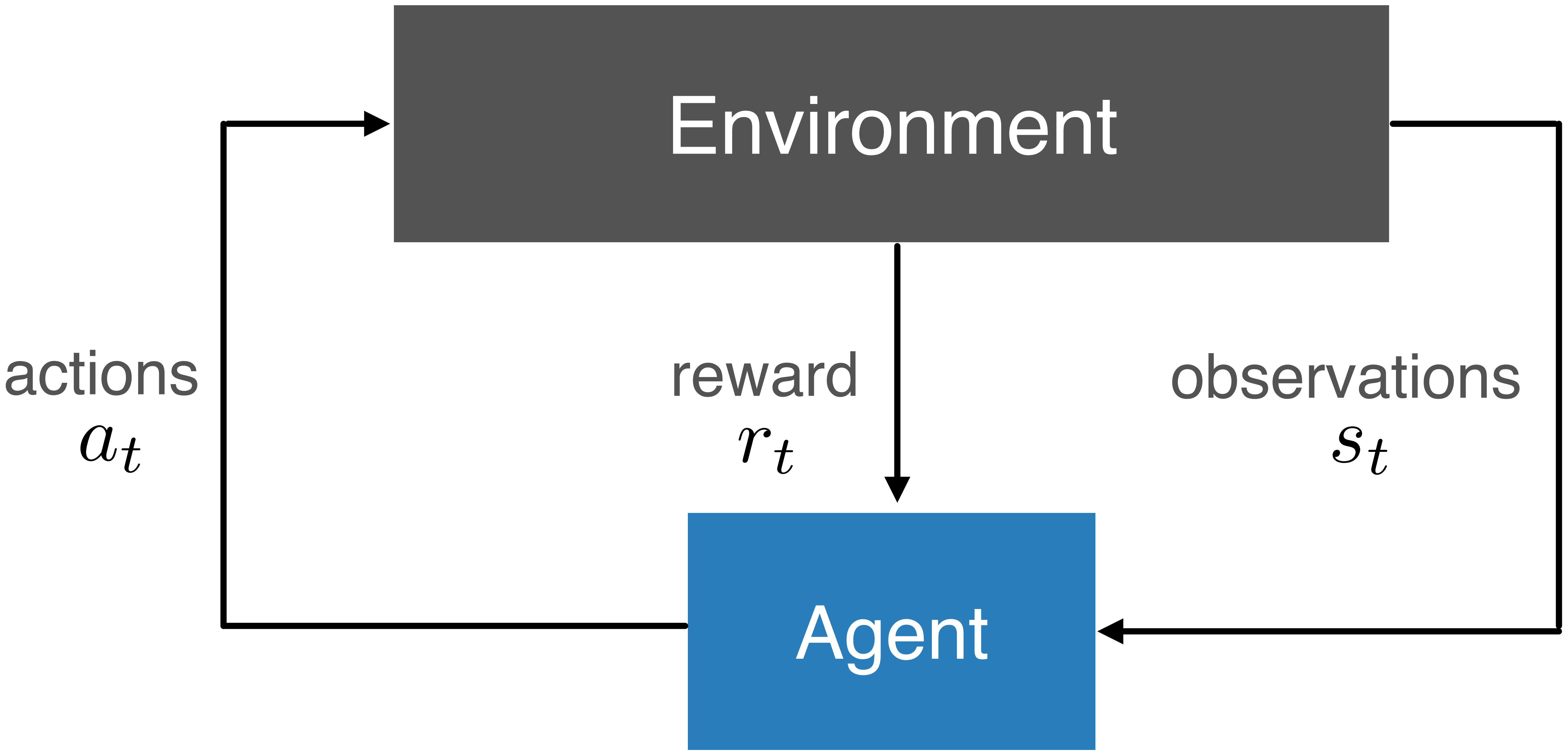
$$\pi(a_t | s_0, r_0, a_0, s_1, r_1, a_1, ...)$$

Agent's Life

$$\{s_0, r_0, a_0, s_1, r_1, a_1, s_2, r_2, a_2, ...\}$$

Objective

$$V_\mu^\pi = \mathbb{E}(\sum_{i=0}^{\infty} r_i)$$



- Where do rewards come from?

Reinforcement Learning

Agent's policy

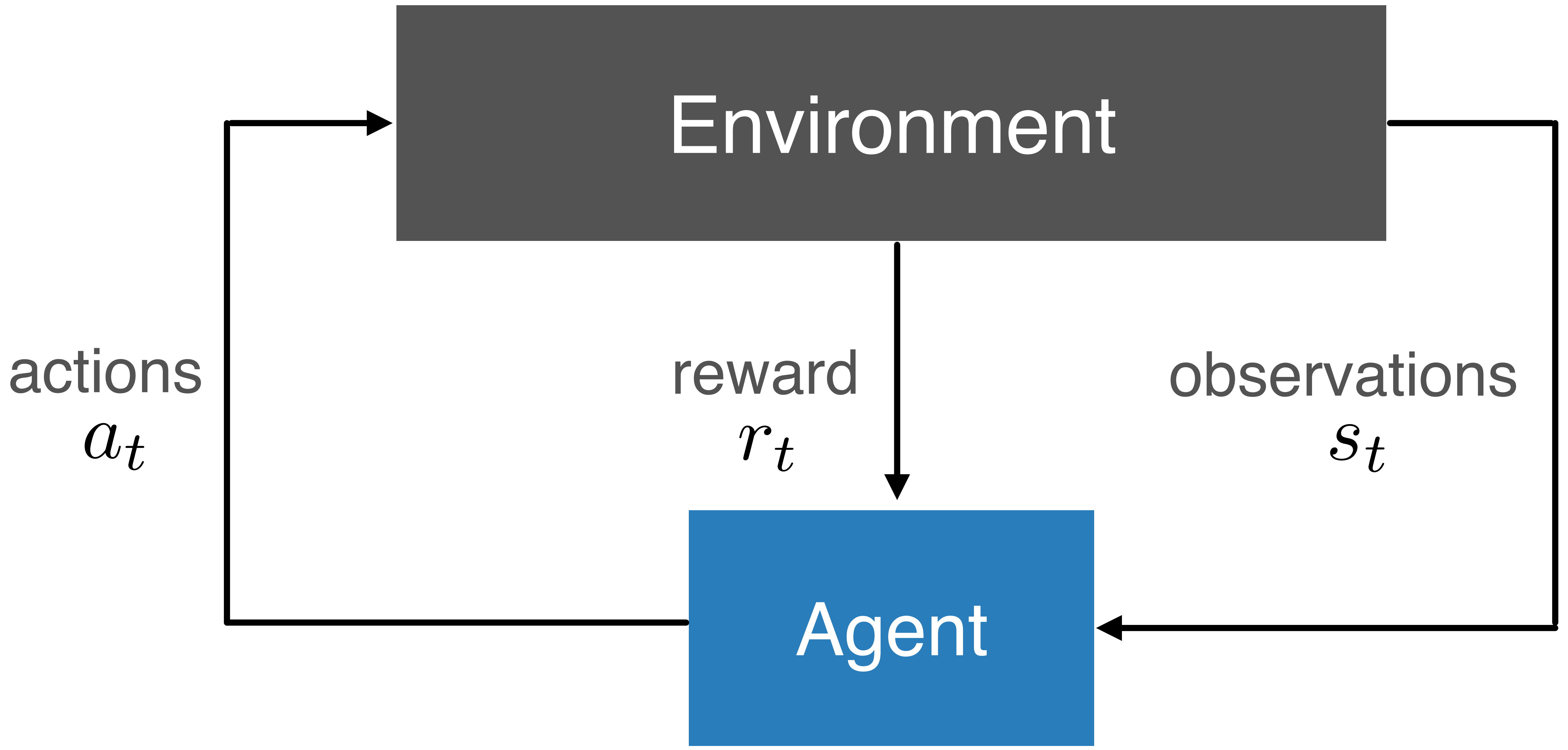
$$\pi(a_t|s_0, r_0, a_0, s_1, r_1, a_1, ...)$$

Agent's Life

$$\{s_0, r_0, a_0, s_1, r_1, a_1, s_2, r_2, a_2, ...\}$$

Objective

$$V_\mu^\pi = \mathbb{E}(\sum_{i=0}^{\infty} r_i)$$



- Where do rewards come from?
- What are effective exploration strategies with and without extrinsic rewards?

Reinforcement Learning

Agent's policy

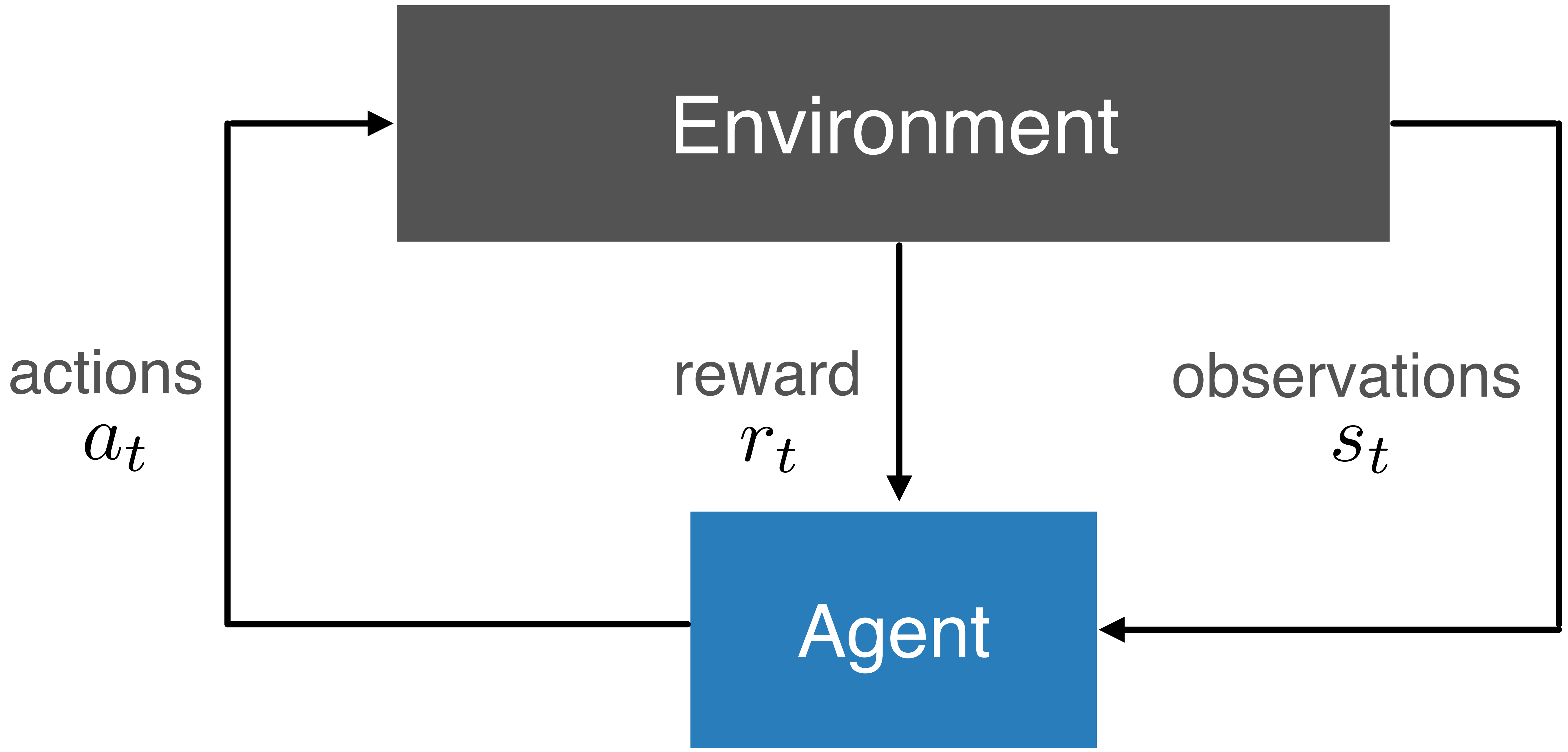
$$\pi(a_t | s_0, r_0, a_0, s_1, r_1, a_1, ...)$$

Agent's Life

$$\{s_0, r_0, a_0, s_1, r_1, a_1, s_2, r_2, a_2, ...\}$$

Objective

$$V_\mu^\pi = \mathbb{E}(\sum_{i=0}^{\infty} r_i)$$



- Where do rewards come from?
- What are effective exploration strategies with and without extrinsic rewards?
- There is rich structure in the space of actions that can be exploited

Deep RL + Intrinsic Motivation + Options

Deep RL + Intrinsic Motivation + Options

Extrinsic Motivation: Activities done in order to attain some separable outcome

Deep RL + Intrinsic Motivation + Options

Extrinsic Motivation: Activities done in order to attain some separable outcome

Intrinsic Motivation: Activities pursued for their own sake, not instrumental value.
Taxonomy *(Oudeyer & Kaplan, 2008)* :

Deep RL + Intrinsic Motivation + Options

Extrinsic Motivation: Activities done in order to attain some separable outcome

Intrinsic Motivation: Activities pursued for their own sake, not instrumental value.
Taxonomy *(Oudeyer & Kaplan, 2008)* :

Knowledge Based Models

Learning Progress. *(Berlyne, 1965; Schmidhuber, 1991; Oudeyer et al., 2007; Lopes et al.,2012)*

Predictive novelty motivation. *(Thrun, 1995; Barto et al., 2004)*

Novelty via Prediction Error. *(Singh et al, 2004; Stadie et al., 2015)*

Novelty via Value Error. *(Simsek and Barto. 2006)*

Mutual Information. *(Rezende, 2015; Houthhooft et al., 2016)*

Visitation free via pseudo-counts. *(Bellemare, 2016)*

Deep RL + Intrinsic Motivation + Options

Extrinsic Motivation: Activities done in order to attain some separable outcome

Intrinsic Motivation: Activities pursued for their own sake, not instrumental value.
Taxonomy *(Oudeyer & Kaplan, 2008)* :

Knowledge Based Models

Learning Progress. *(Berlyne, 1965; Schmidhuber, 1991; Oudeyer et al., 2007; Lopes et al.,2012)*

Predictive novelty motivation. *(Thrun, 1995; Barto et al., 2004)*

Novelty via Prediction Error. *(Singh et al, 2004; Stadie et al., 2015)*

Novelty via Value Error. *(Simsek and Barto. 2006)*

Mutual Information. *(Rezende, 2015; Houthhooft et al., 2016)*

Visitation free via pseudo-counts. *(Bellemare, 2016)*

Competence Based Models

Effectance Motivation (White, 1959)

Competence and self-determination (Deci & Ryan, 1985)

Goal driven exploration (Oudeyer et al. 2013)

Deep RL + Intrinsic Motivation + Options

Extrinsic Motivation: Activities done in order to attain some separable outcome

Intrinsic Motivation: Activities pursued for their own sake, not instrumental value.
Taxonomy *(Oudeyer & Kaplan, 2008)* :

Knowledge Based Models

Learning Progress. *(Berlyne, 1965; Schmidhuber, 1991; Oudeyer et al., 2007; Lopes et al.,2012)*

Predictive novelty motivation. *(Thrun, 1995; Barto et al., 2004)*

Novelty via Prediction Error. *(Singh et al, 2004; Stadie et al., 2015)*

Novelty via Value Error. *(Simsek and Barto. 2006)*

Mutual Information. *(Rezende, 2015; Houthhooft et al., 2016)*

Visitation free via pseudo-counts. *(Bellemare, 2016)*

Competence Based Models

Effectance Motivation (White, 1959)

Competence and self-determination (Deci & Ryan, 1985)

Goal driven exploration (Oudeyer et al. 2013)

Deep RL + Intrinsic Motivation + Options

Deep RL + Intrinsic Motivation + Options

Options: Hierarchies of Behavior. Temporal abstractions over actions. Subgoals

Deep RL + Intrinsic Motivation + Options

Options: Hierarchies of Behavior. Temporal abstractions over actions. Subgoals

Options framework.
(Sutton et al., 1999)

Universal option model.
(Szepesvari et al., 2004)

Universal Value Func Appx.
(Schaul et al., 2015)

Deep RL + Intrinsic Motivation + Options

Options: Hierarchies of Behavior. Temporal abstractions over actions. Subgoals

Options framework.
(Sutton et al., 1999)

Universal option model.
(Szepesvari et al., 2004)

Universal Value Func Appx.
(Schaul et al., 2015)

Option discovery (subgoals/macro-actions)

Visit frequency. *(McGovern & Barto, 2001; Digney,1998)*

Salience. *(Singh et al. 2004)*

Graph partitioning. *(Simsek et al., 2005)*

Purposefulness. *(Machado et al., 2016*

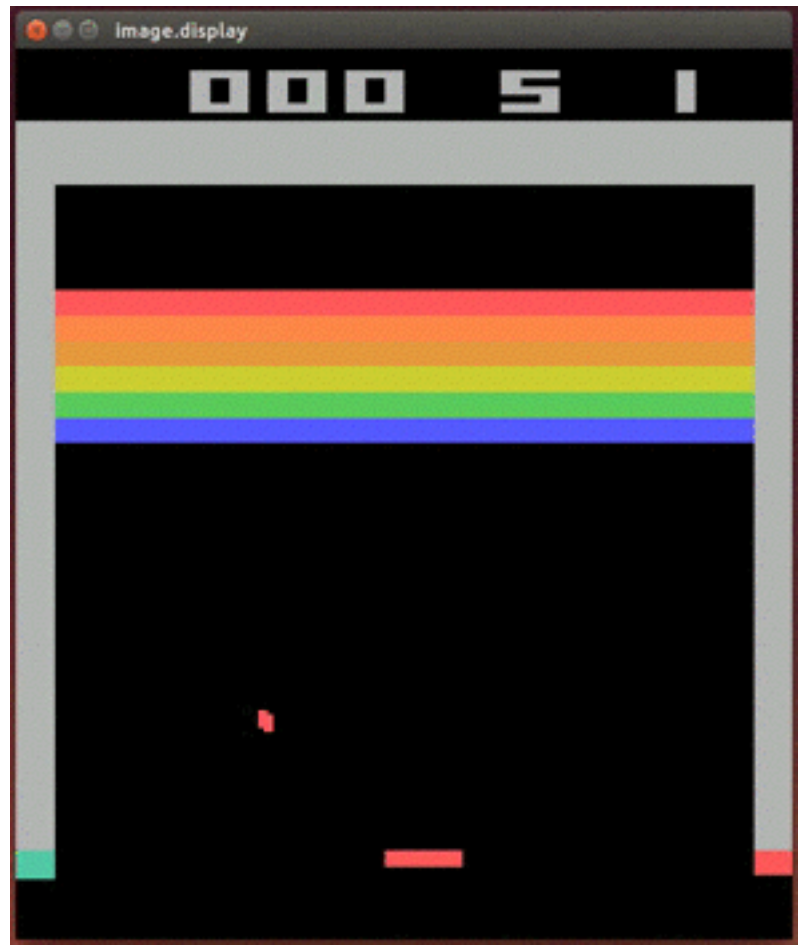
Structure in collection of policies. *(Thrun 1995, Bernstein 1999, Perkins 1999, Pickett 2002)*

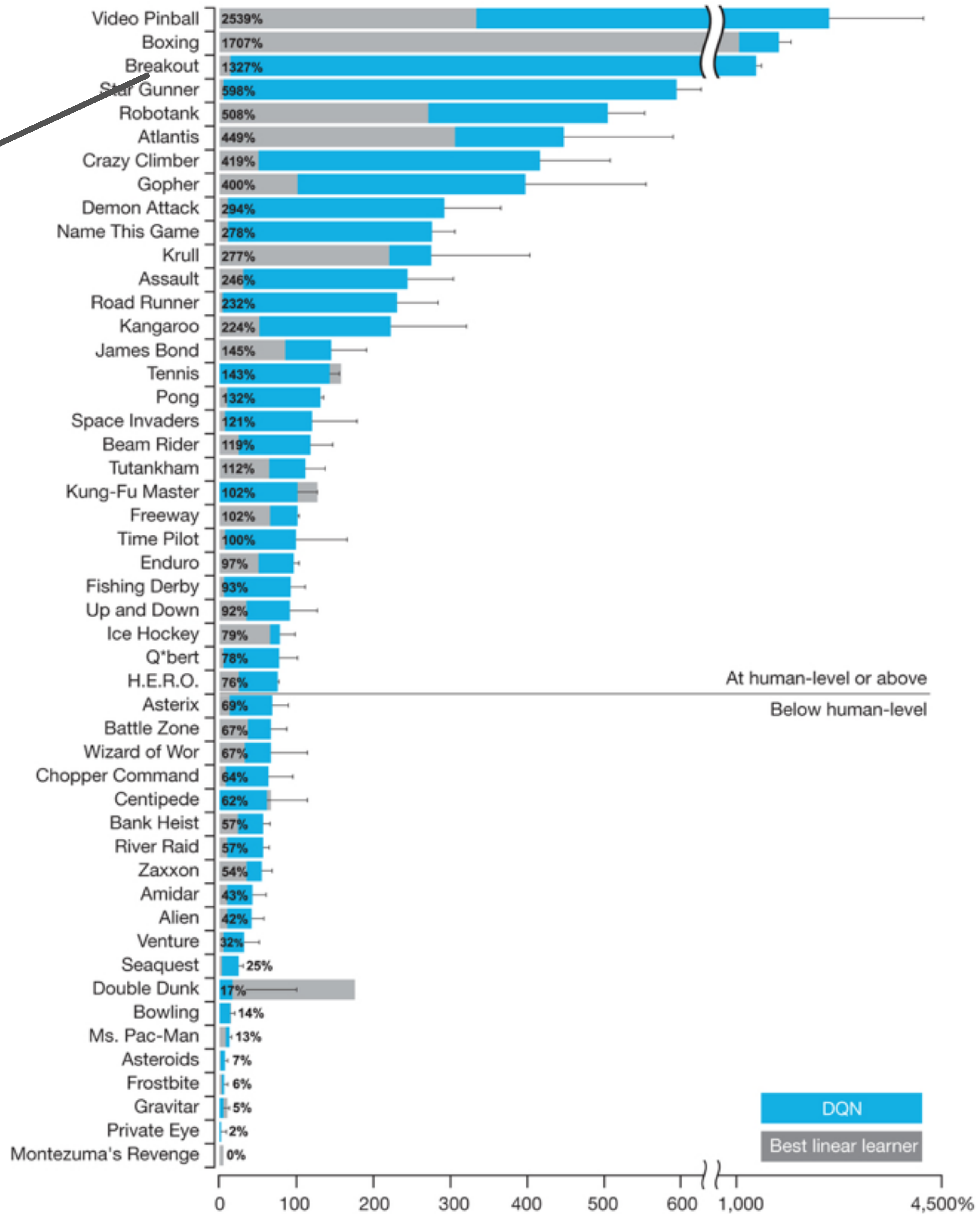
Clustering algorithms and value gradients. *(Mannor et al. 2004)*

Deep sucessor representations. *(Kulkarni et al., 2016)*

Strategic attn writer for macro-actions. *(Vezhnevets et al. 2016)*

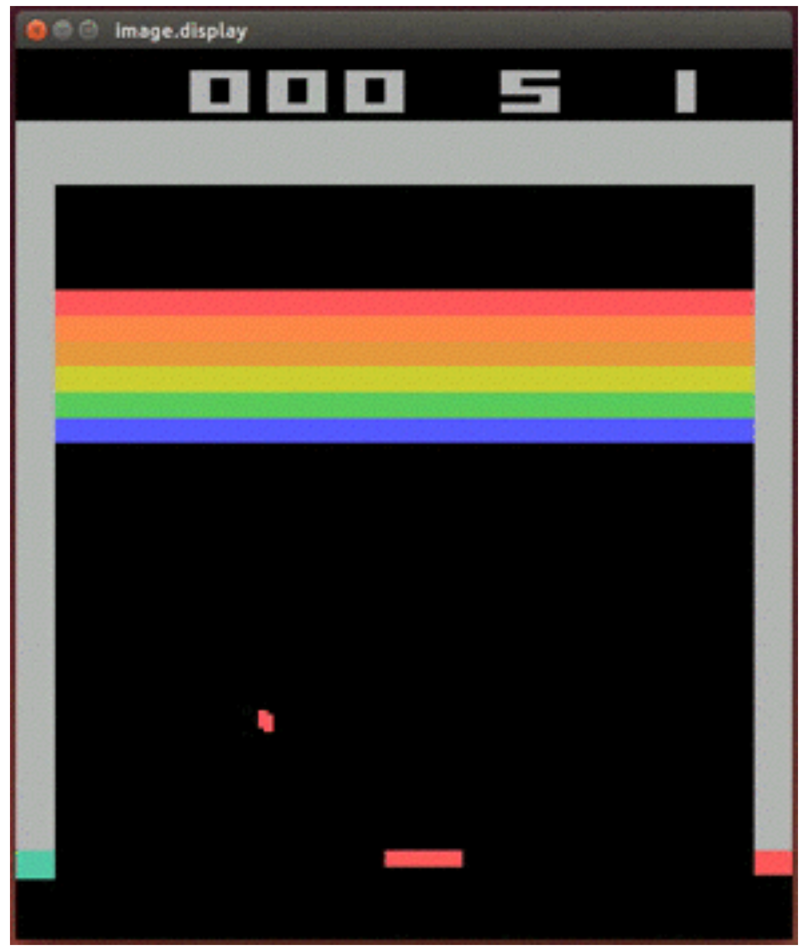
Deep RL + Intrinsic Motivation + Options

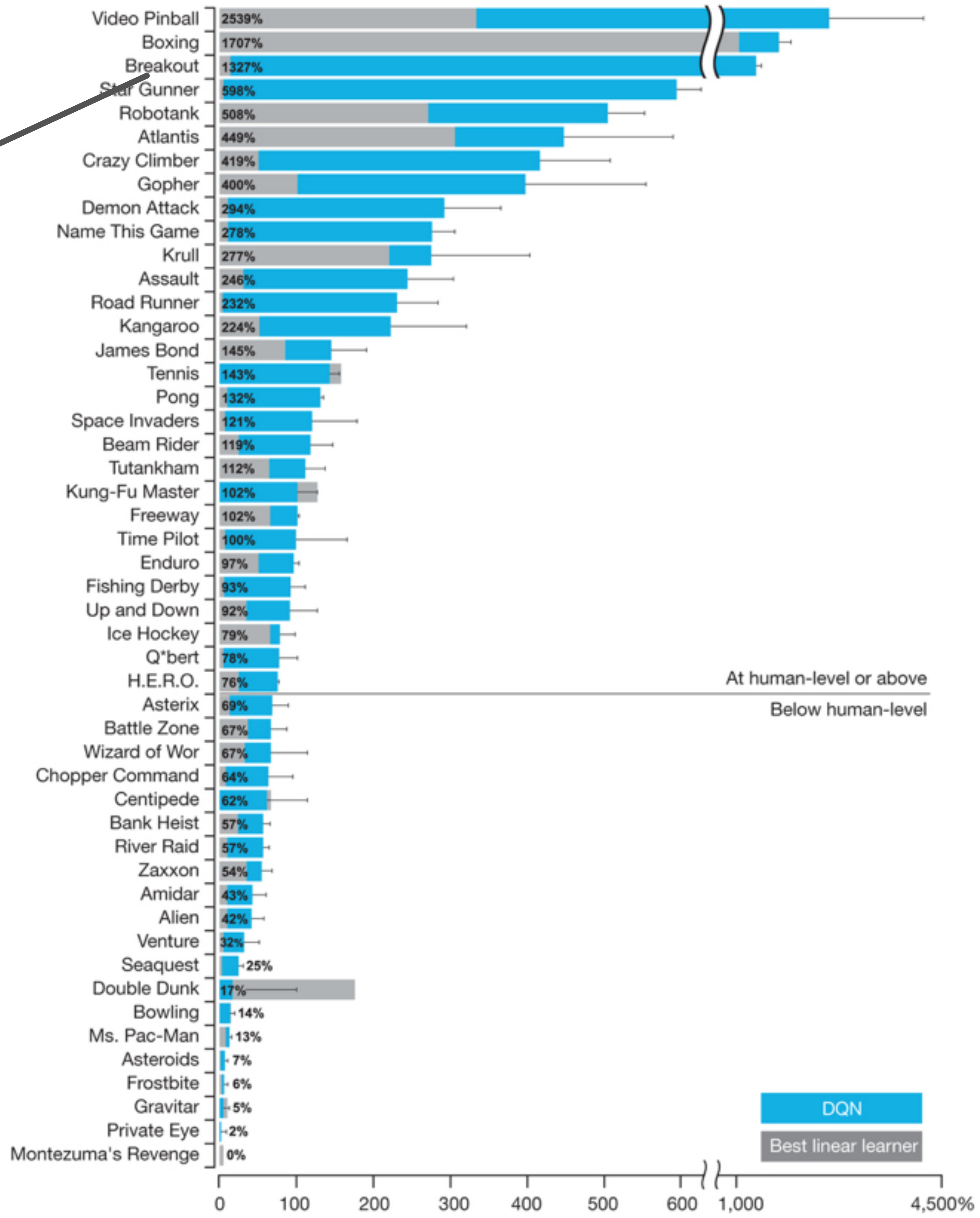




Mnih et al., Nature 2015

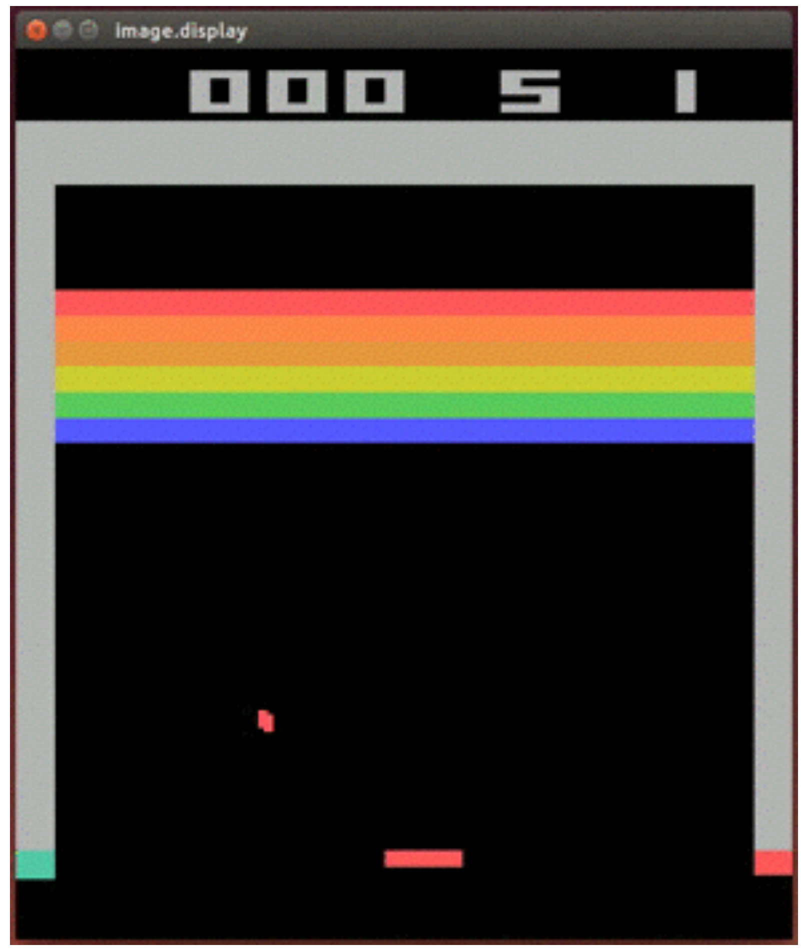
Deep RL + Intrinsic Motivation + Options

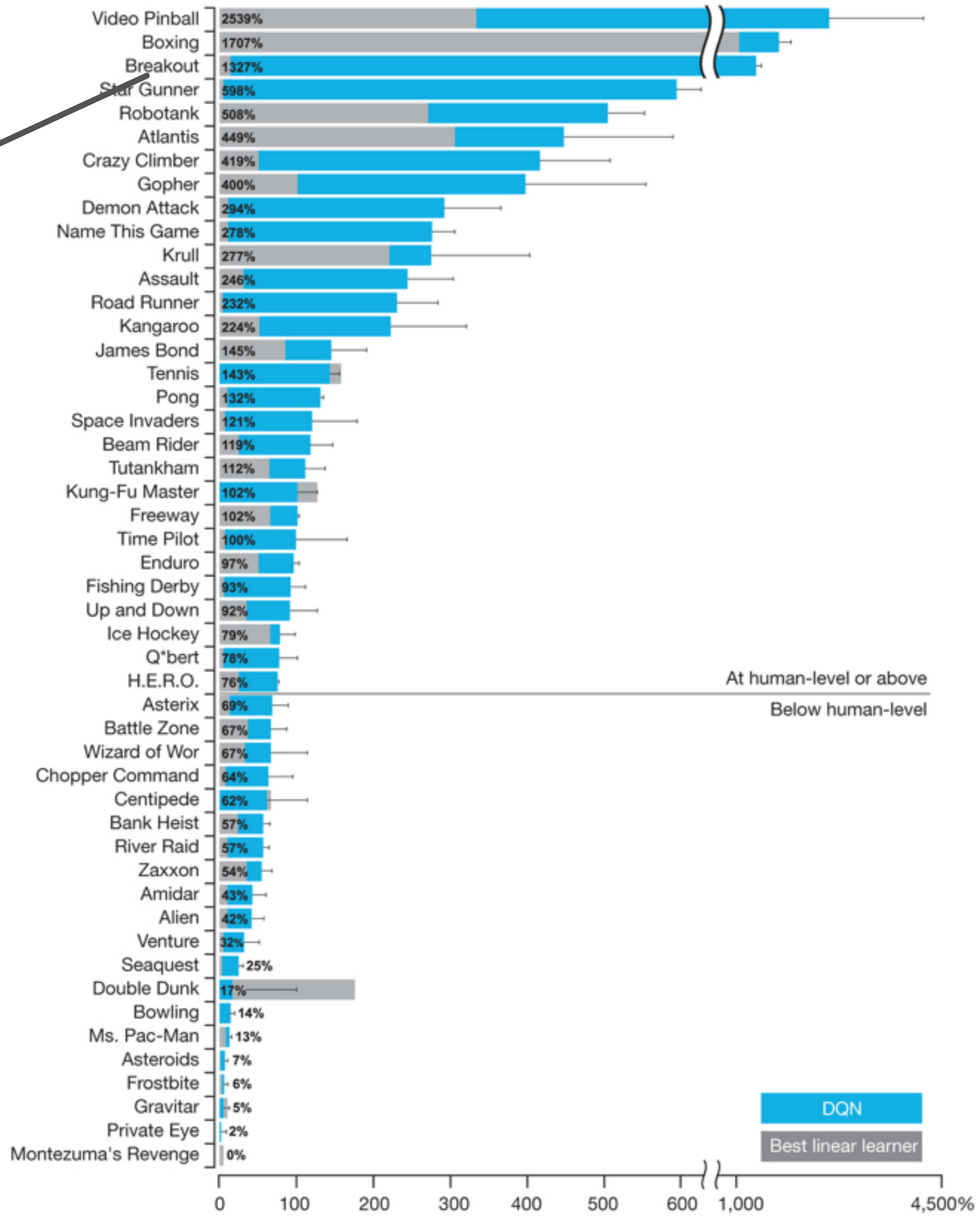




Mnih et al., Nature 2015

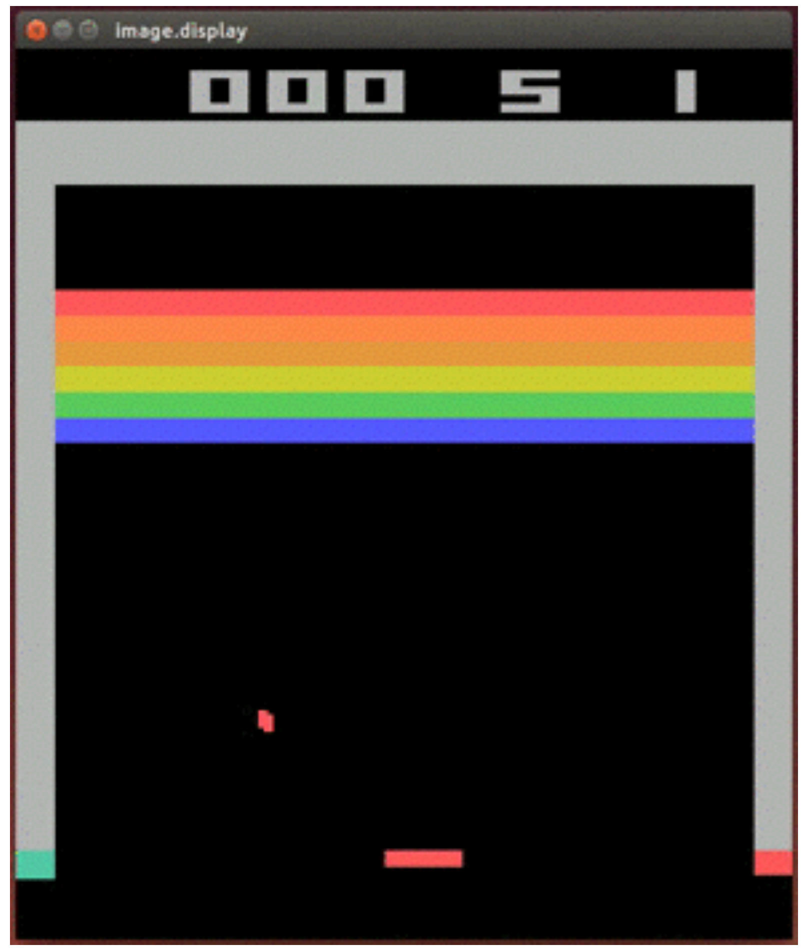
Deep RL + Intrinsic Motivation + Options

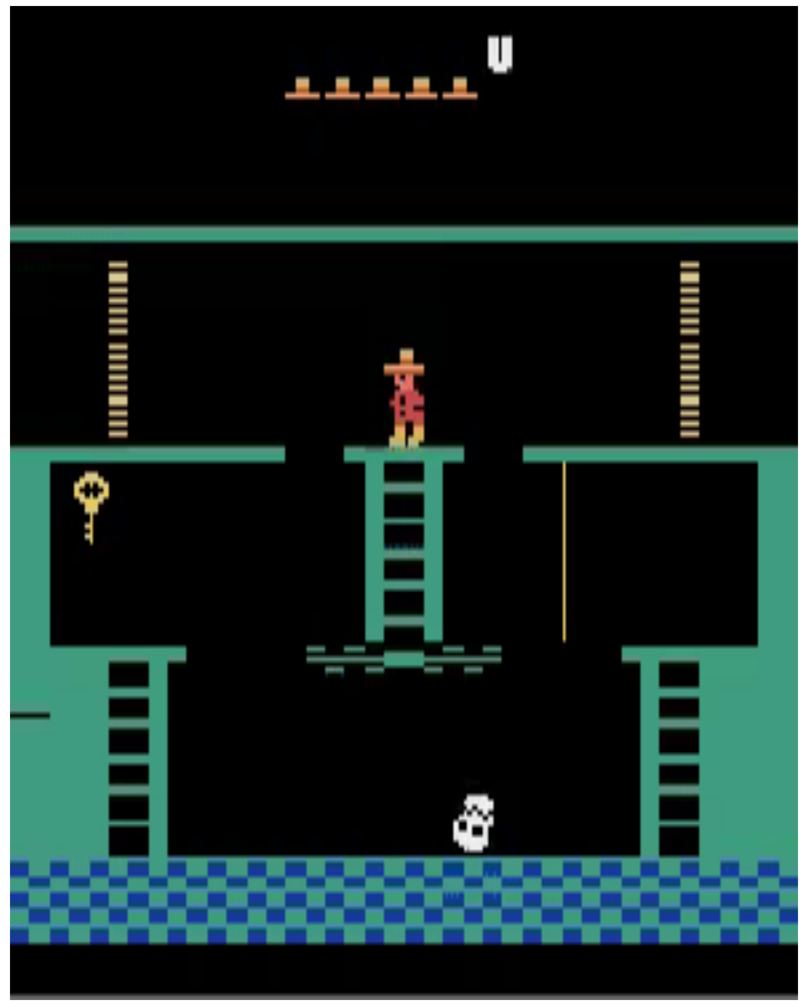




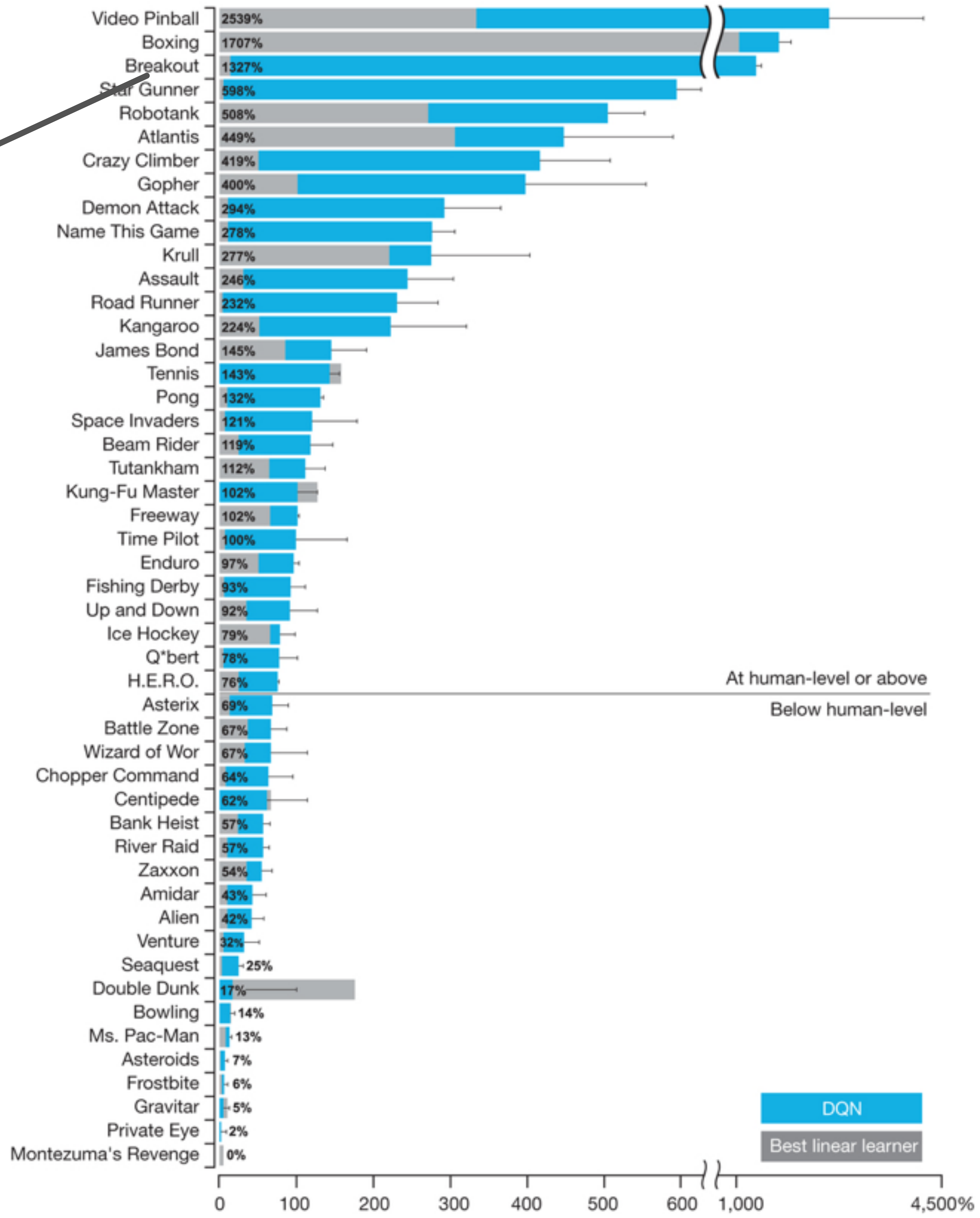
Mnih et al., Nature 2015

Deep RL + Intrinsic Motivation + Options



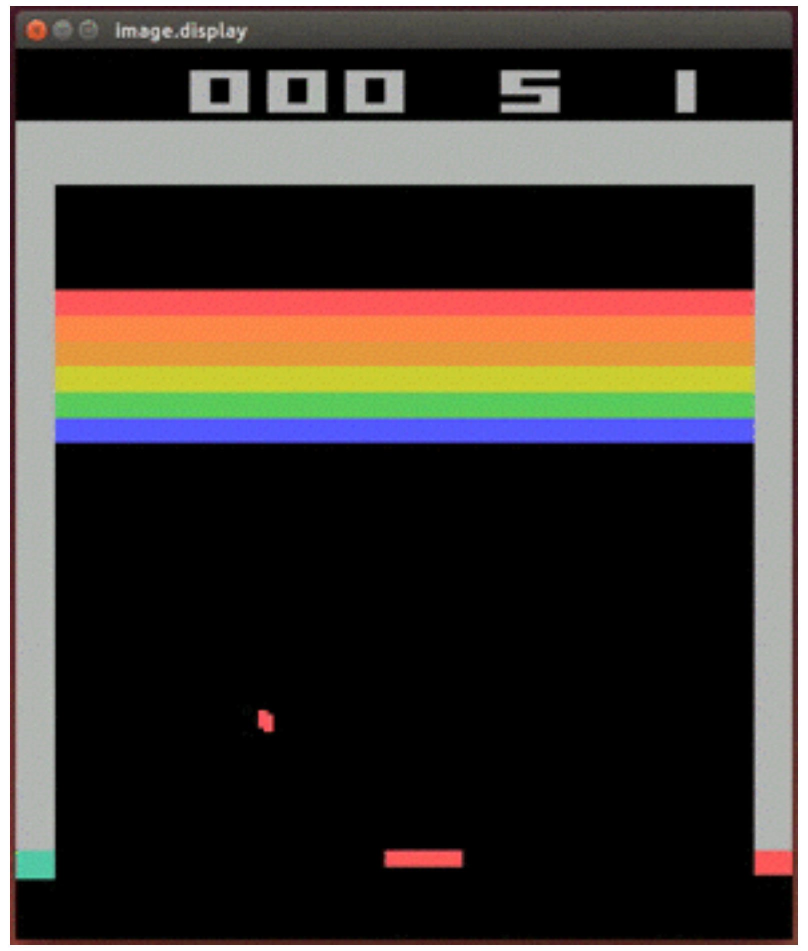


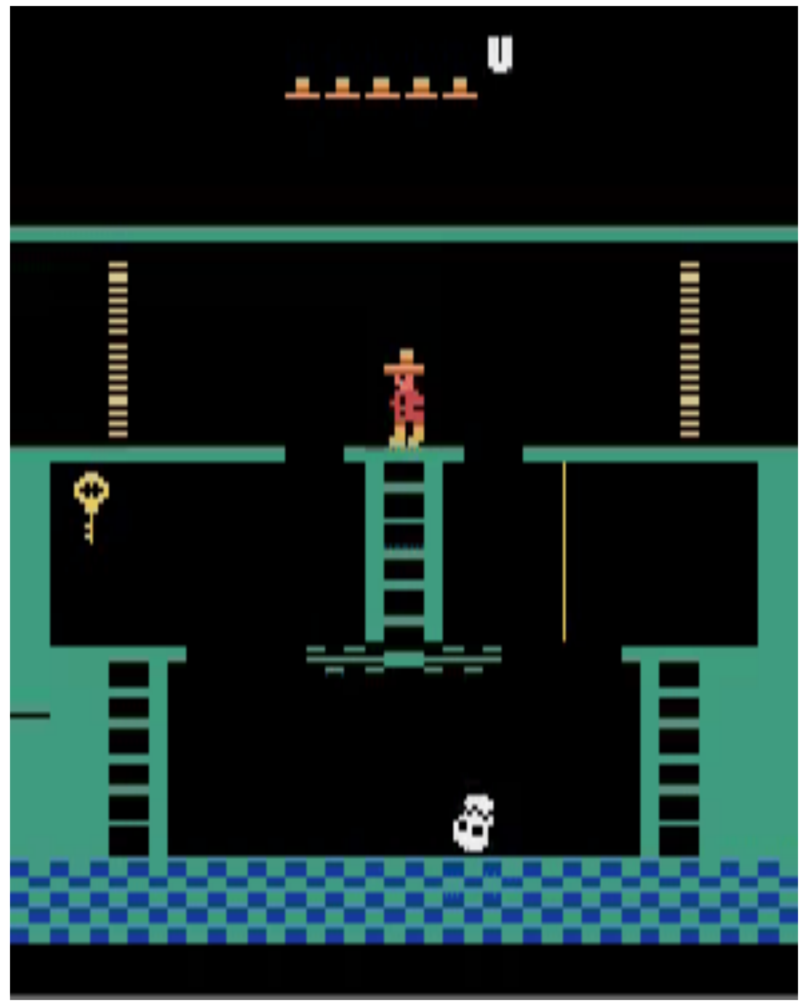
epsilon greedy



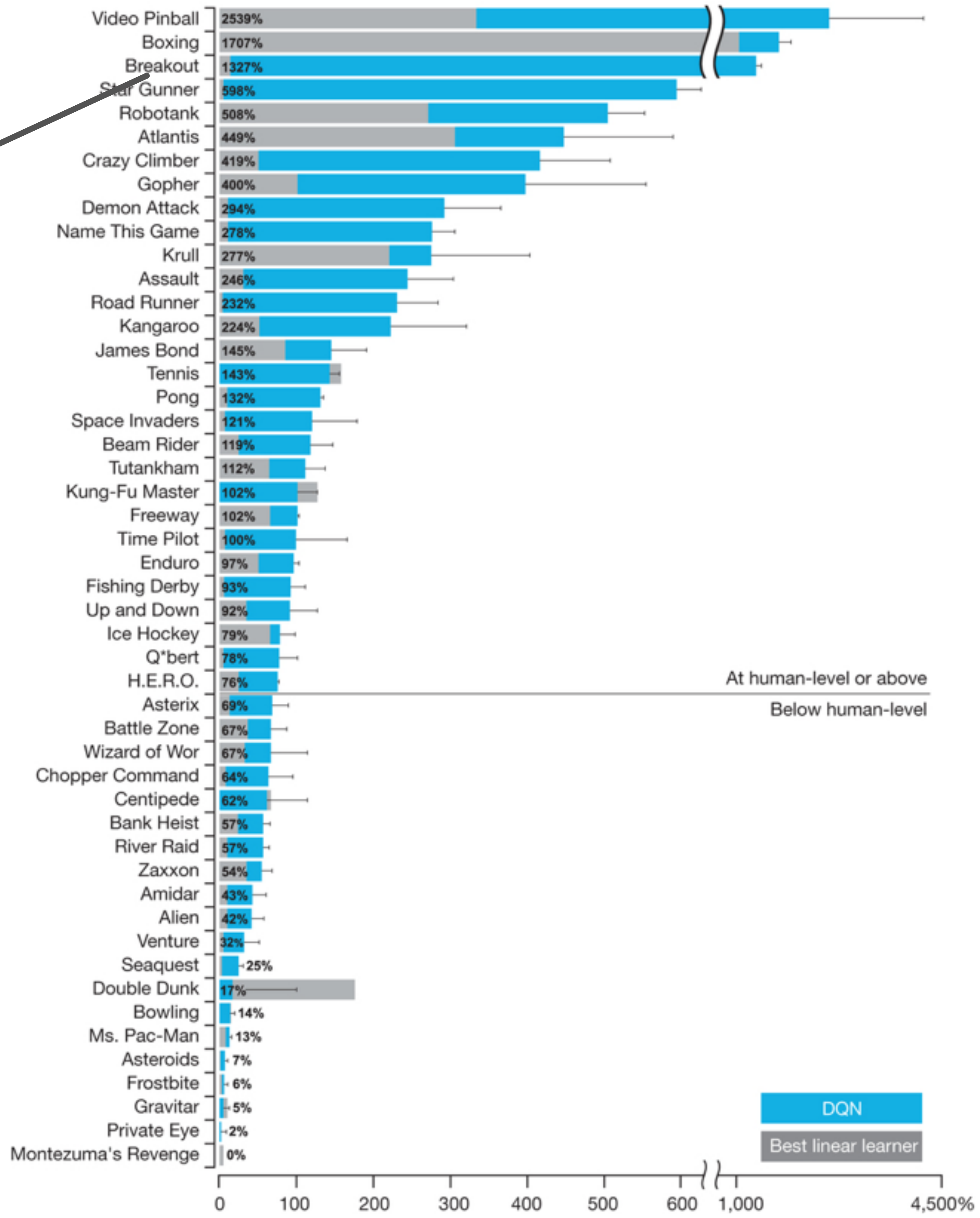
Mnih et al., Nature 2015

Deep RL + Intrinsic Motivation + Options



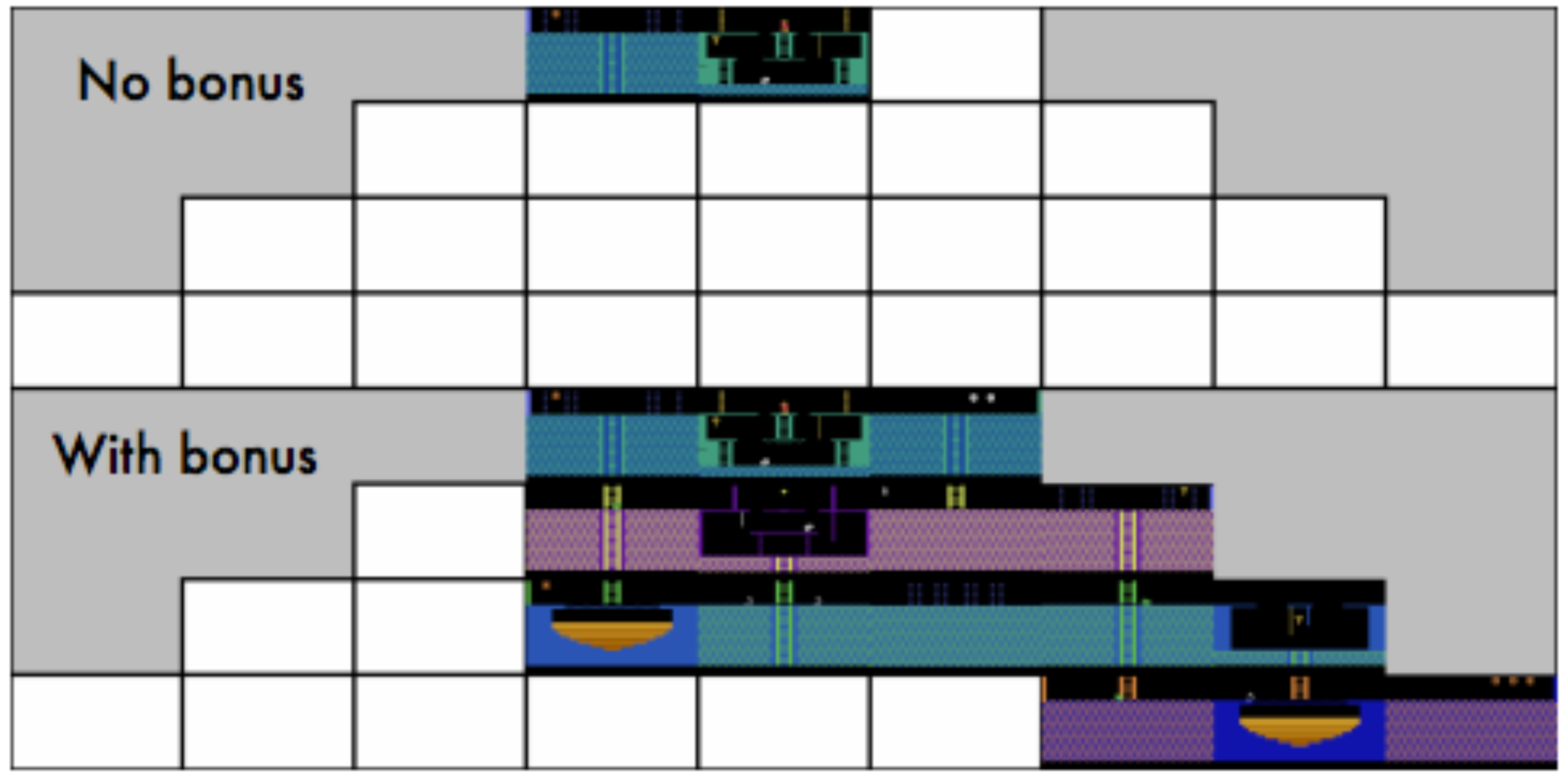


epsilon greedy



Mnih et al., Nature 2015

50 Million frames



Bellemare et al., 2016

Hierarchical Deep Reinforcement Learning (h-DQN)

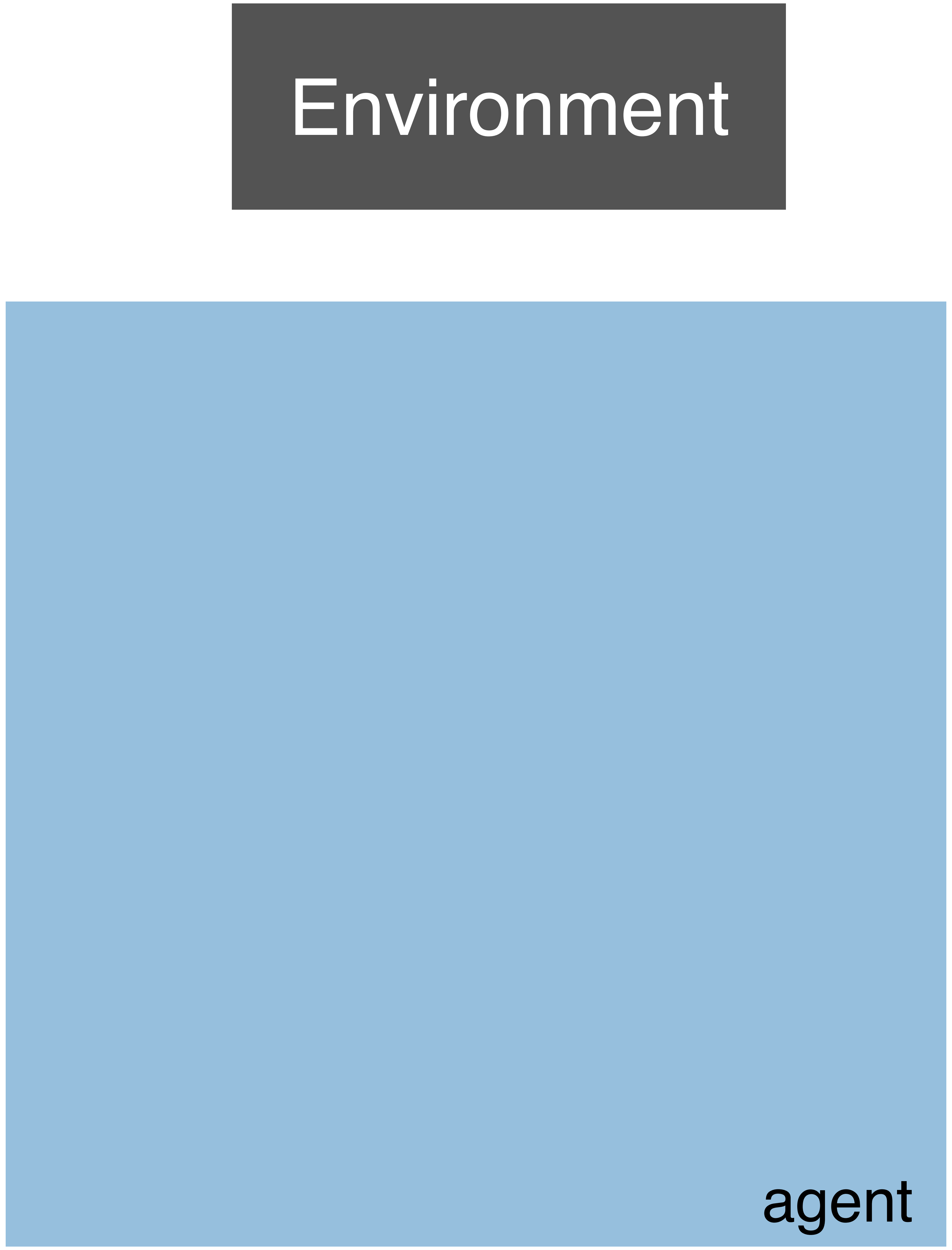
Goal-driven exploration (Oudeyer et al., ‘13)
Intrinsically Motivated RL.*(Singh et al., ’04)*

Hierarchical Deep Reinforcement Learning (h-DQN)

Environment

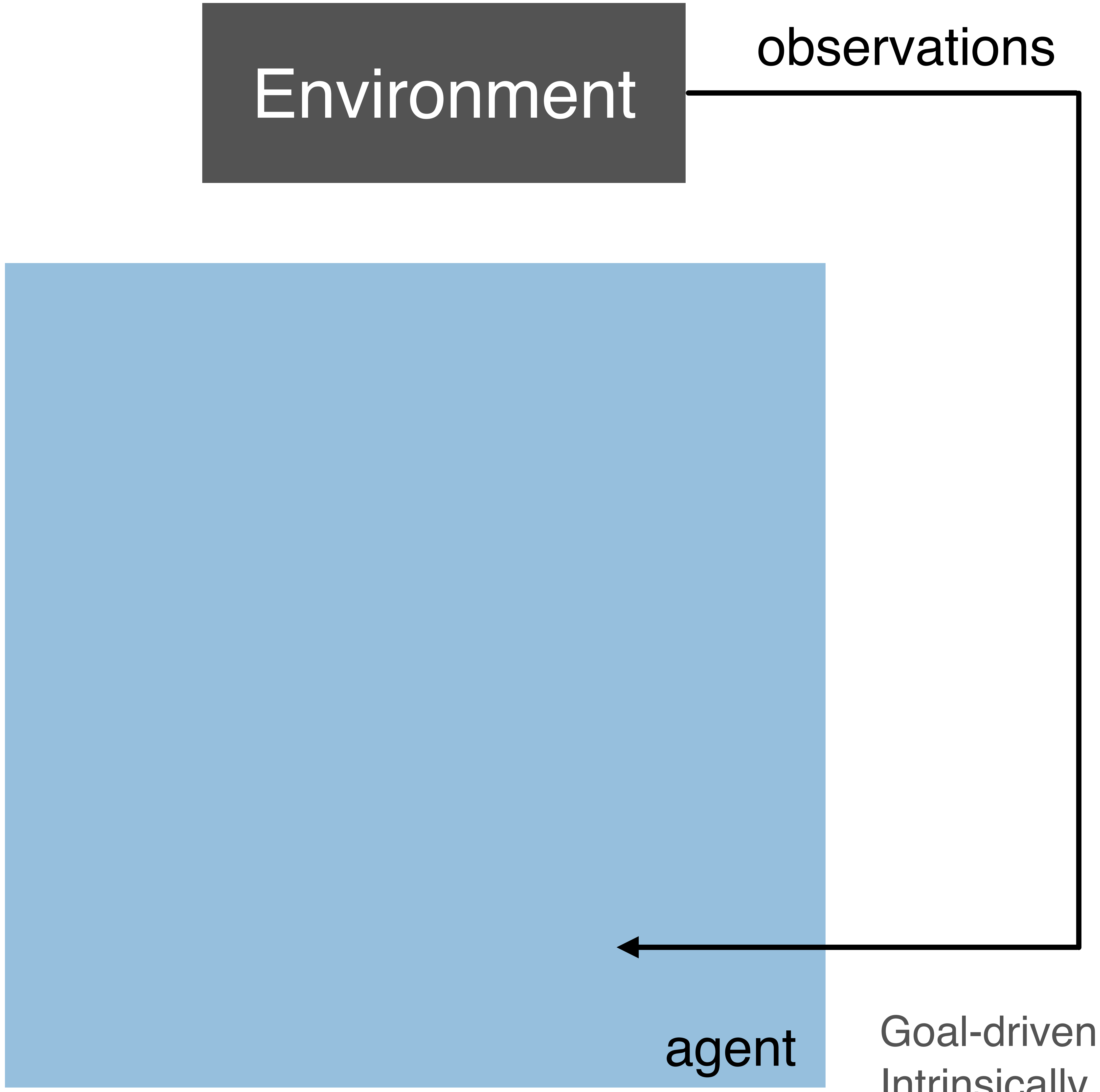
Goal-driven exploration (Oudeyer et al., ‘13)
Intrinsically Motivated RL.*(Singh et al., ’04)*

Hierarchical Deep Reinforcement Learning (h-DQN)



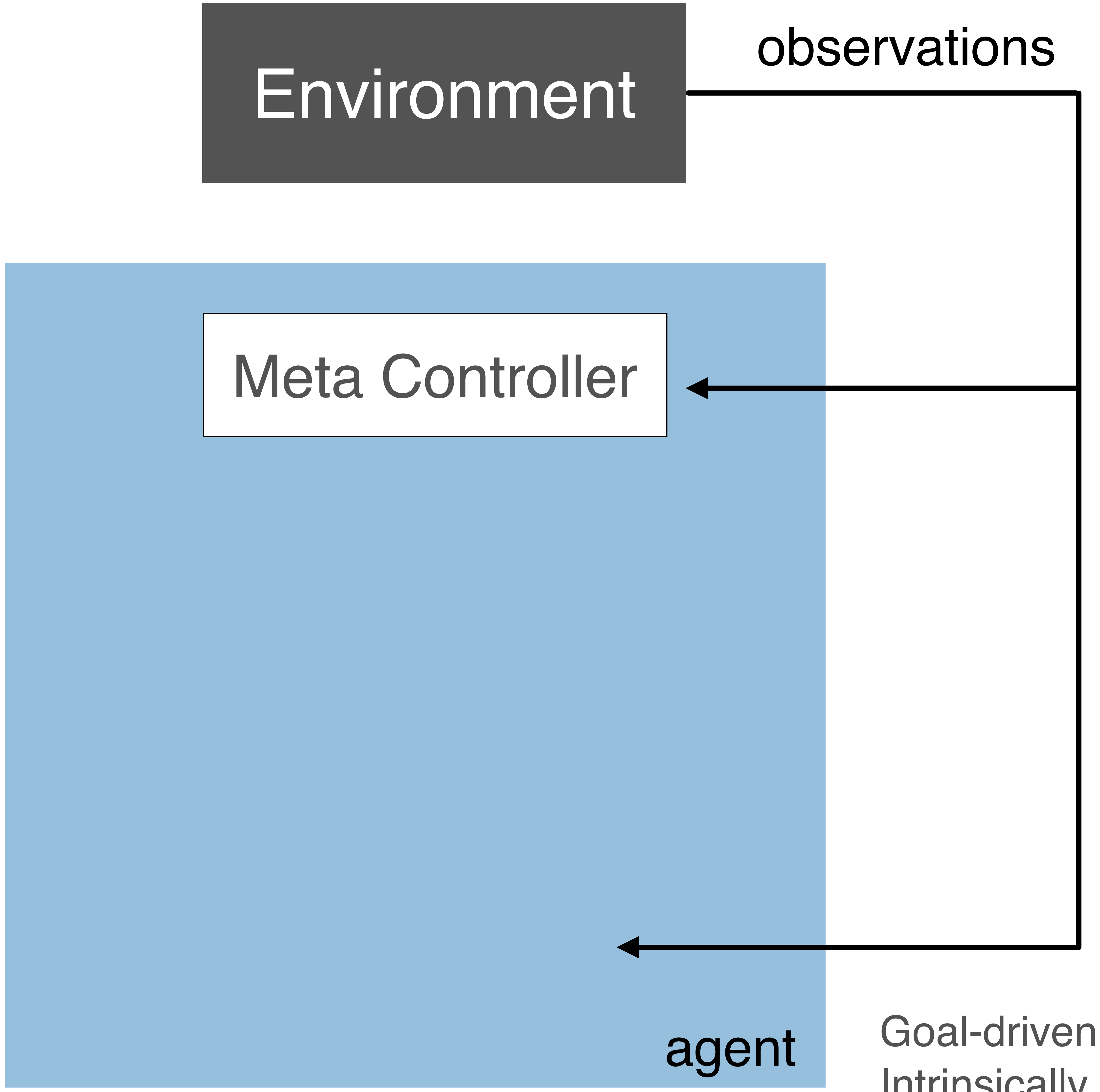
Goal-driven exploration (Oudeyer et al., ‘13)
Intrinsically Motivated RL.*(Singh et al., ’04)*

Hierarchical Deep Reinforcement Learning (h-DQN)



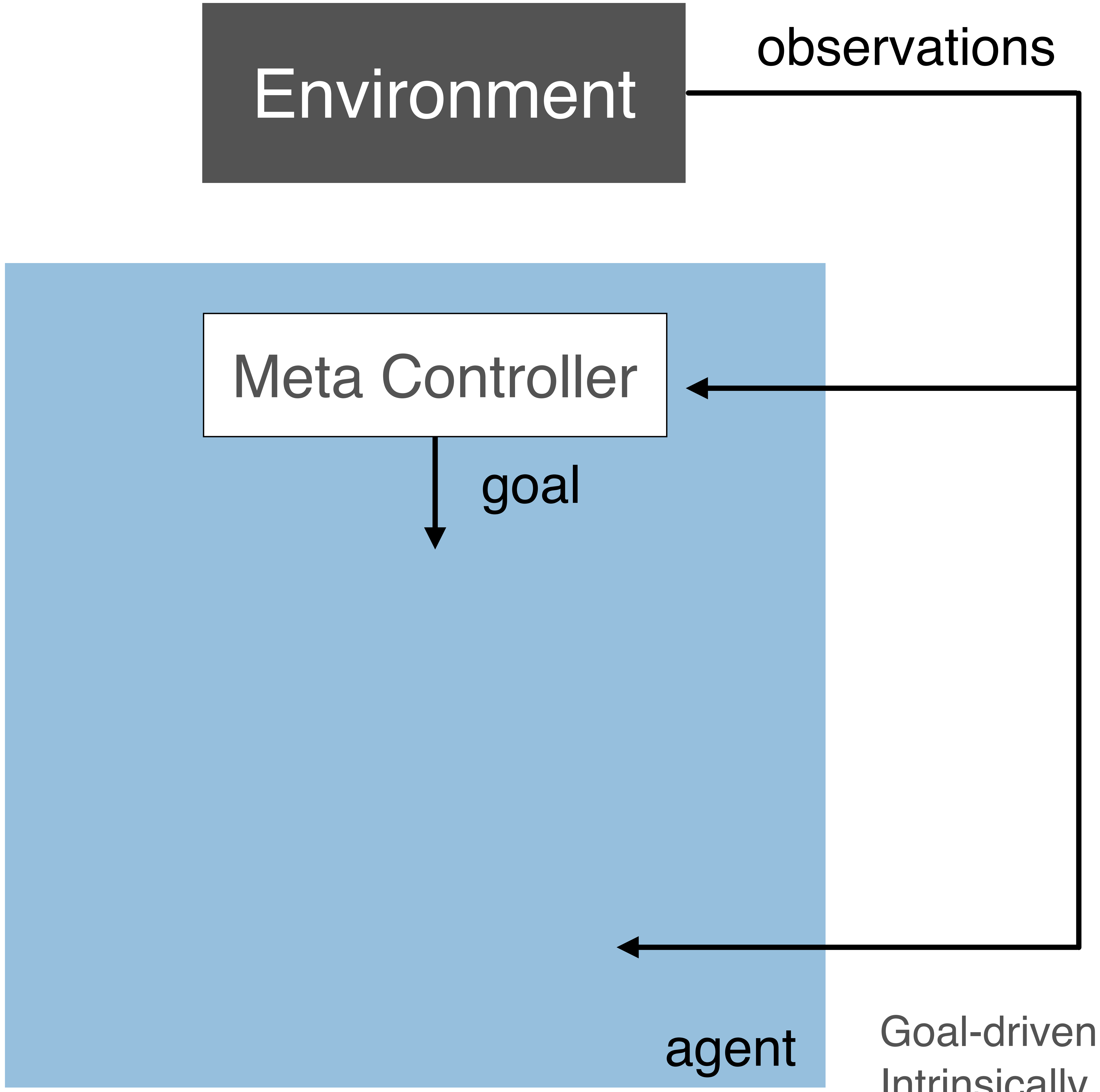
Goal-driven exploration (Oudeyer et al., ‘13)
Intrinsically Motivated RL.*(Singh et al., ’04)*

Hierarchical Deep Reinforcement Learning (h-DQN)



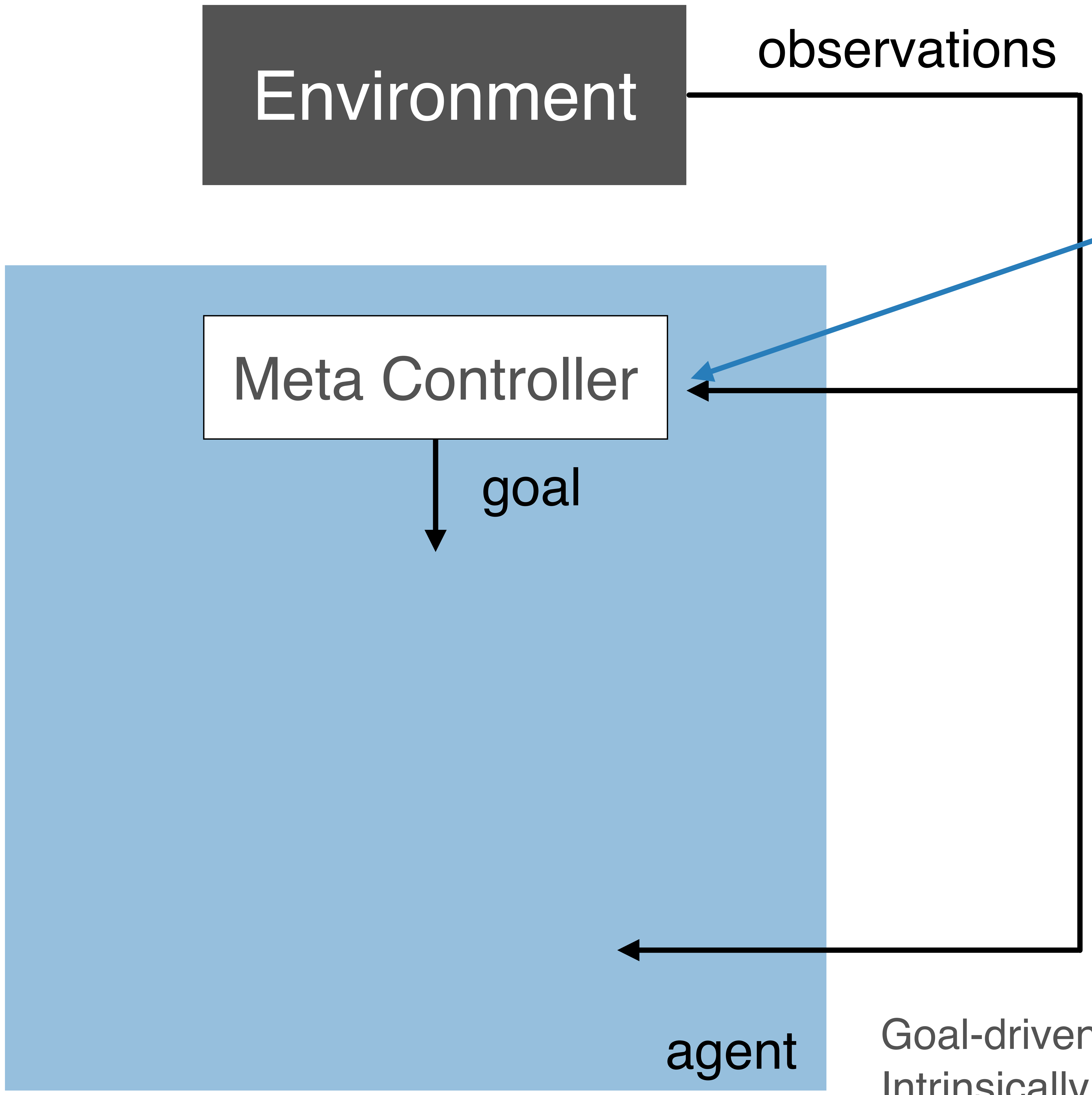
Goal-driven exploration (Oudeyer et al., ‘13)
Intrinsically Motivated RL.*(Singh et al., ’04)*

Hierarchical Deep Reinforcement Learning (h-DQN)



Goal-driven exploration (Oudeyer et al., ‘13)
Intrinsically Motivated RL.*(Singh et al., ’04)*

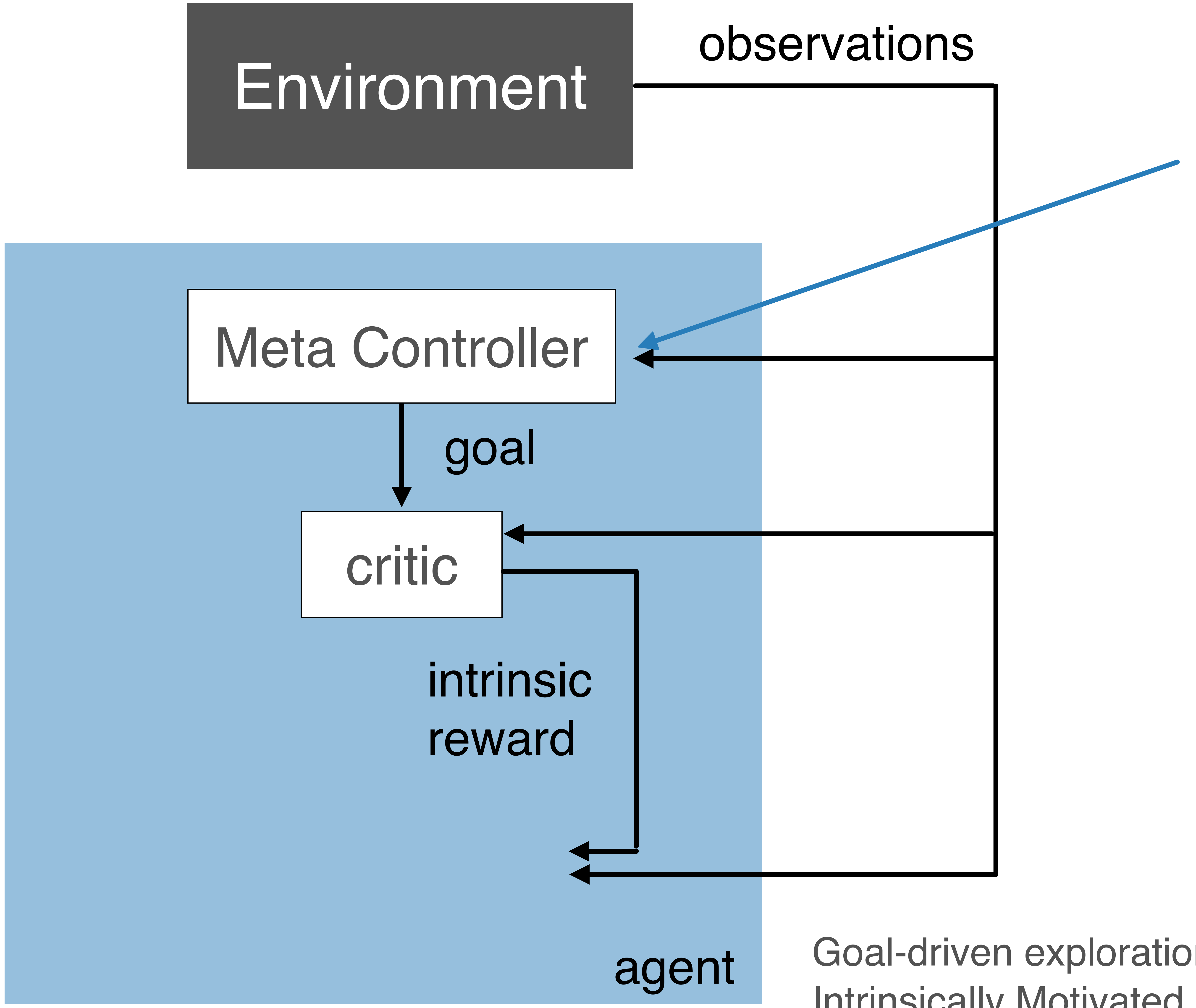
Hierarchical Deep Reinforcement Learning (h-DQN)



Meta controller uses a DNN to learn policy over goals

Goal-driven exploration (Oudeyer et al., ‘13)
Intrinsically Motivated RL.*(Singh et al., ’04)*

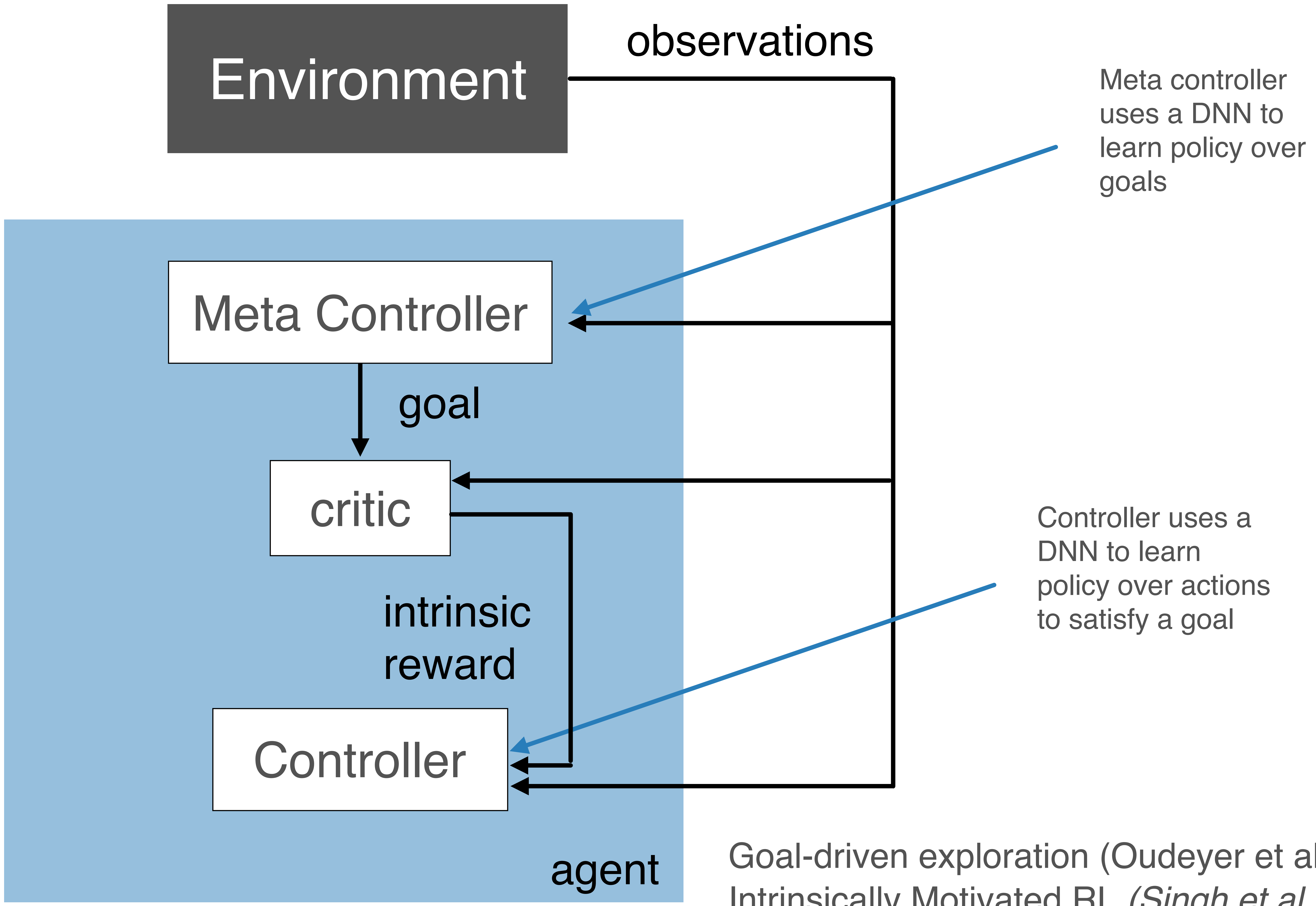
Hierarchical Deep Reinforcement Learning (h-DQN)



Meta controller uses a DNN to learn policy over goals

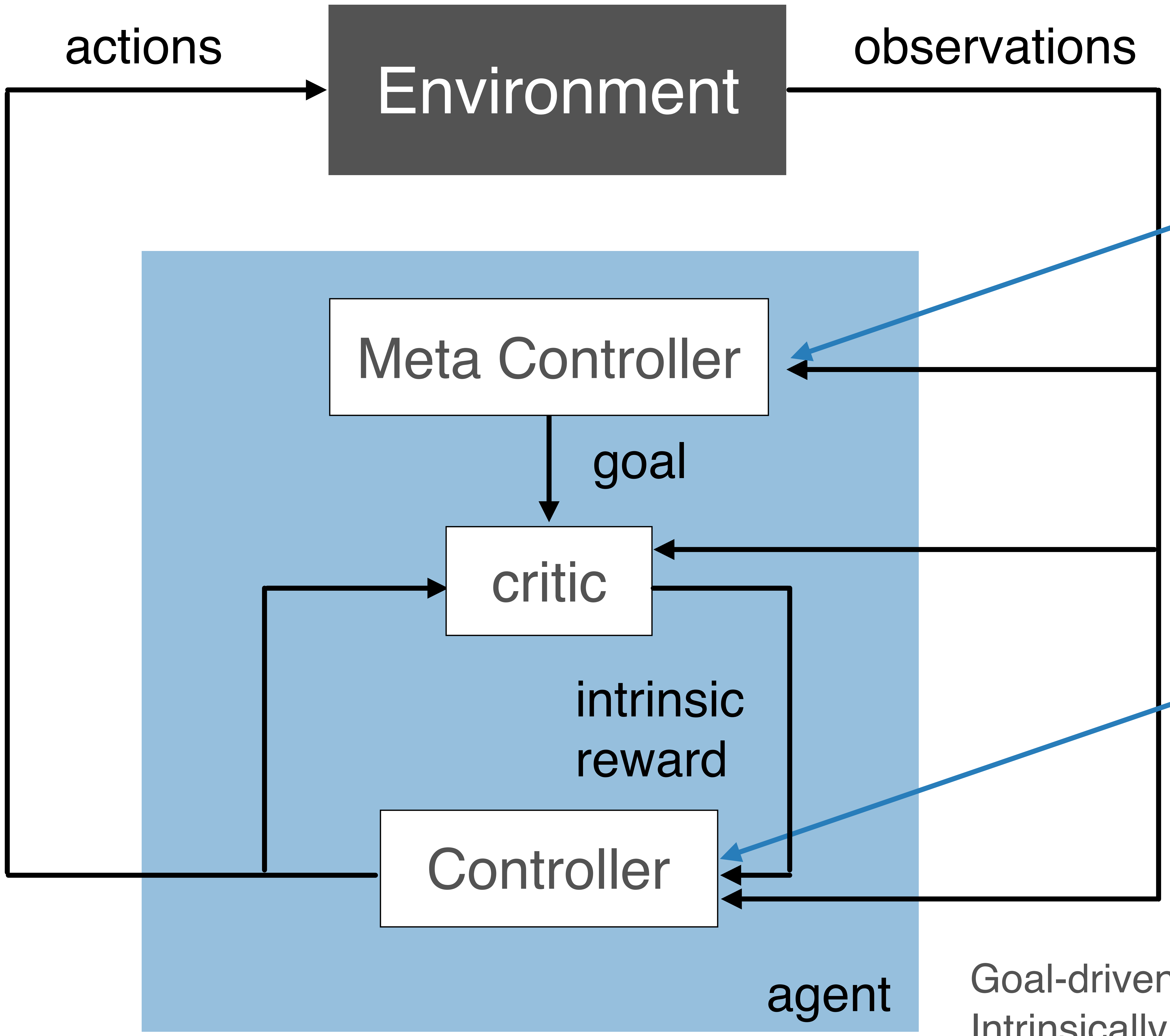
Goal-driven exploration (Oudeyer et al., ‘13)
Intrinsically Motivated RL.*(Singh et al., ’04)*

Hierarchical Deep Reinforcement Learning (h-DQN)



Goal-driven exploration (Oudeyer et al., ‘13)
Intrinsically Motivated RL.*(Singh et al., ’04)*

Hierarchical Deep Reinforcement Learning (h-DQN)

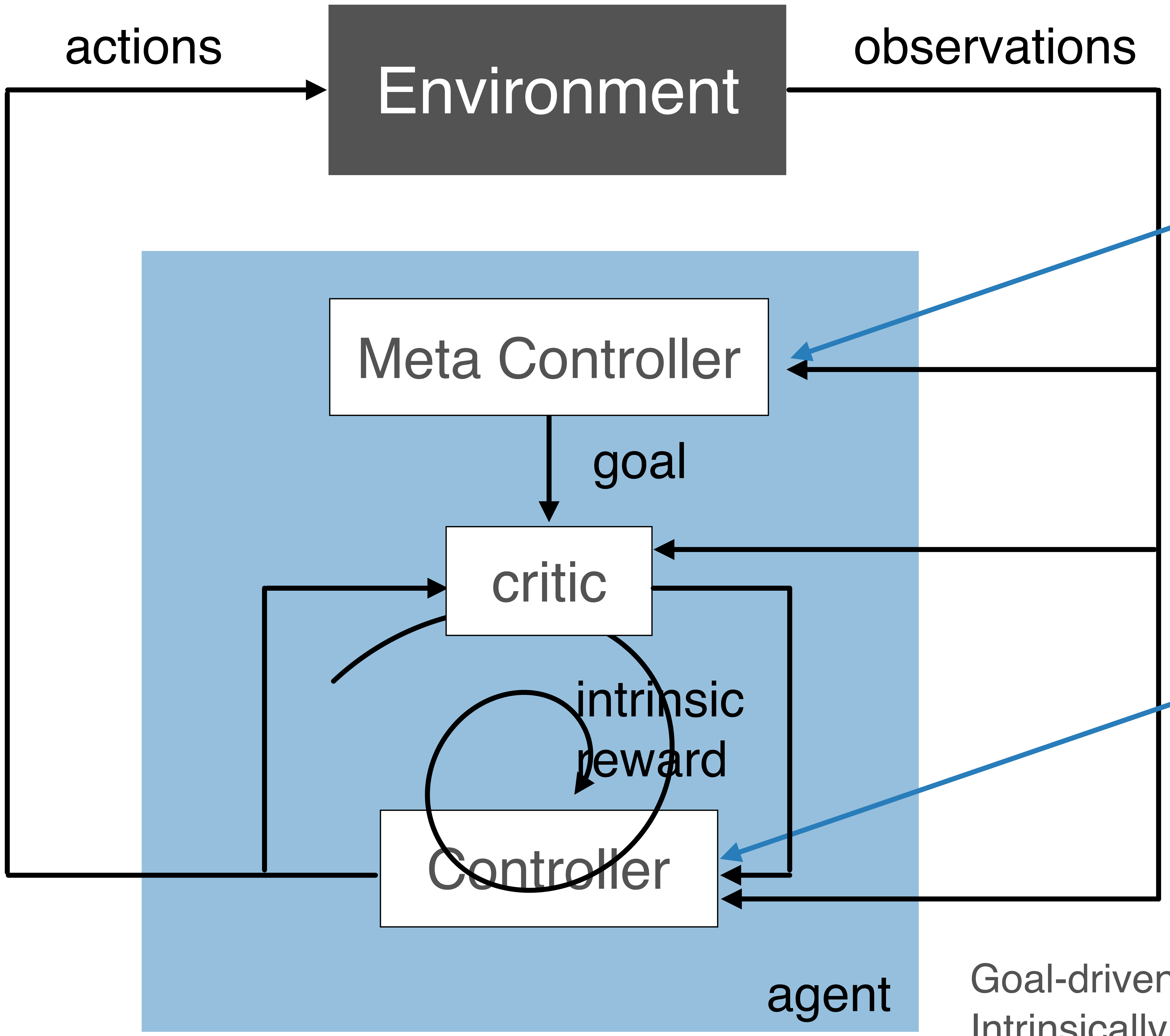


Meta controller uses a DNN to learn policy over goals

Controller uses a DNN to learn policy over actions to satisfy a goal

Goal-driven exploration (Oudeyer et al., ‘13)
Intrinsically Motivated RL.*(Singh et al., ’04)*

Hierarchical Deep Reinforcement Learning (h-DQN)

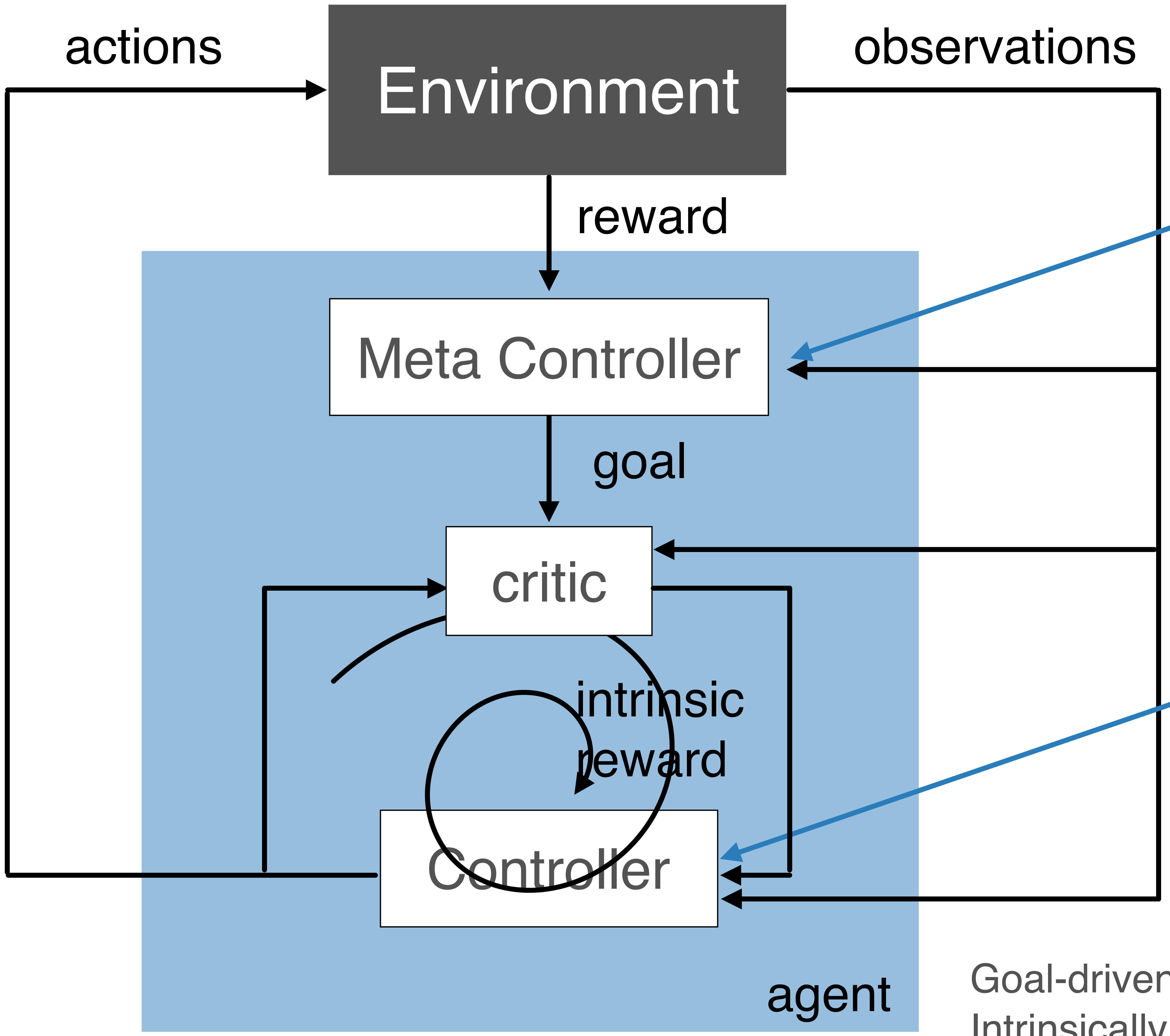


Meta controller uses a DNN to learn policy over goals

Controller uses a DNN to learn policy over actions to satisfy a goal

Goal-driven exploration (Oudeyer et al., '13)
Intrinsically Motivated RL.*(Singh et al., '04)*

Hierarchical Deep Reinforcement Learning (h-DQN)

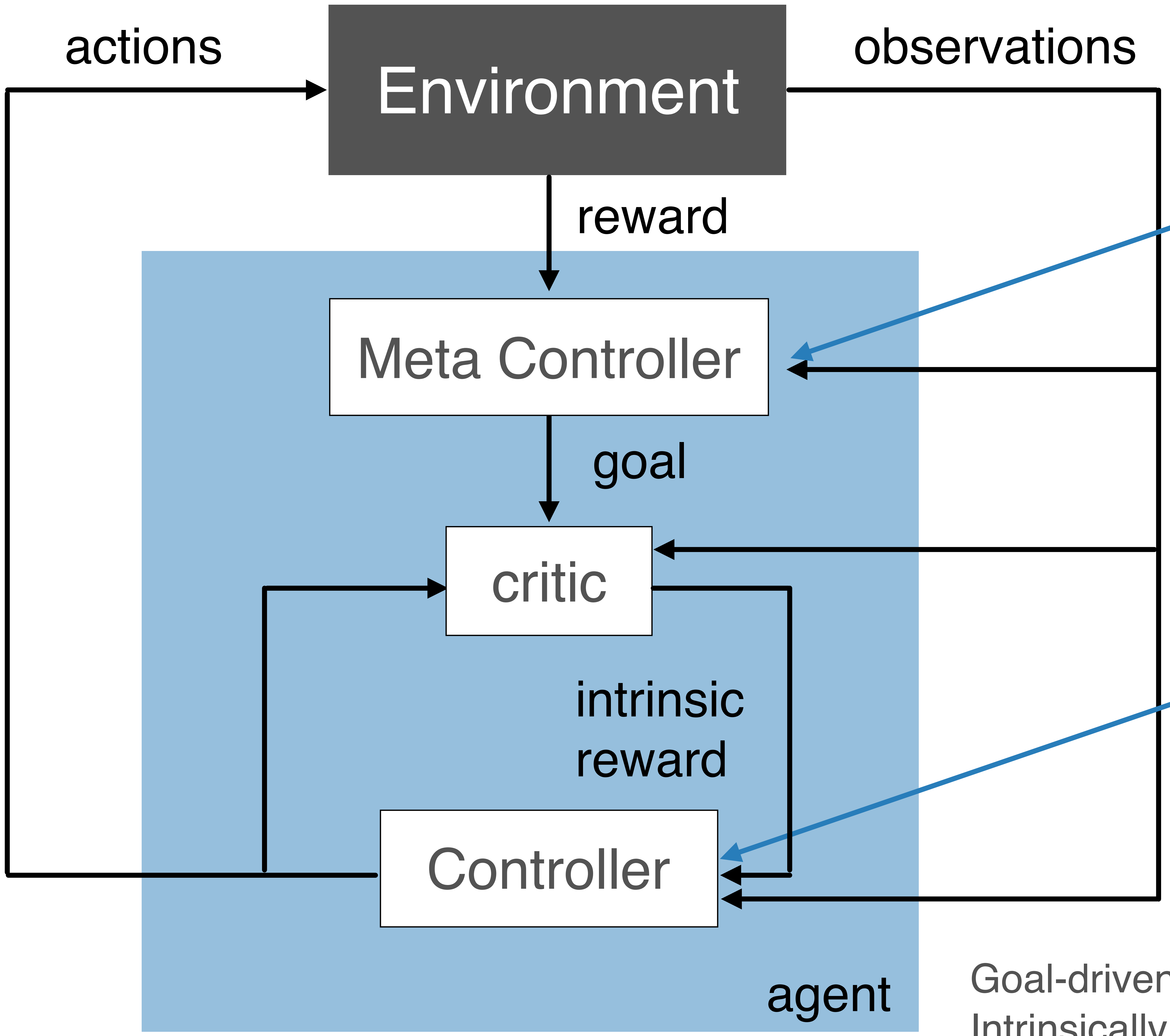


Meta controller uses a DNN to learn policy over goals

Controller uses a DNN to learn policy over actions to satisfy a goal

Goal-driven exploration (Oudeyer et al., '13)
Intrinsically Motivated RL.*(Singh et al., '04)*

Hierarchical Deep Reinforcement Learning (h-DQN)



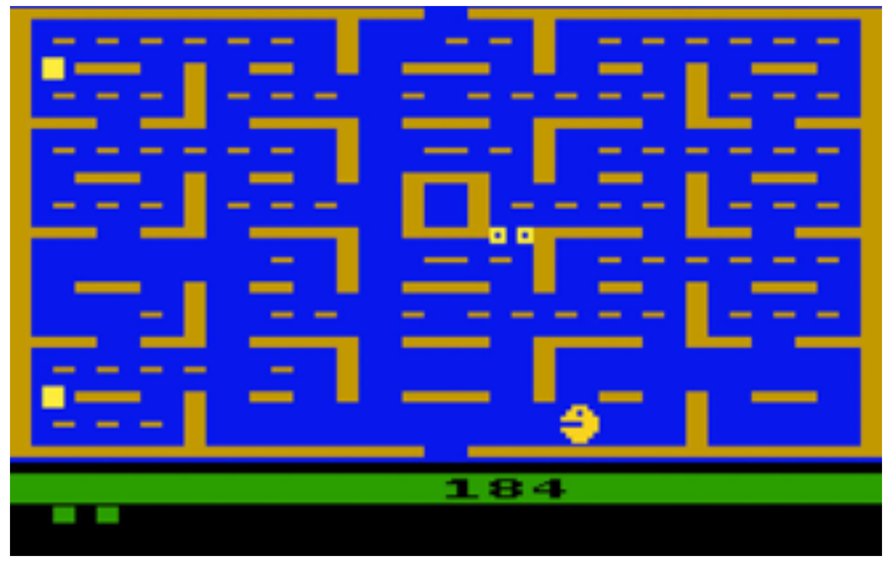
Meta controller uses a DNN to learn policy over goals

Controller uses a DNN to learn policy over actions to satisfy a goal

Goal-driven exploration (Oudeyer et al., ‘13)
Intrinsically Motivated RL.*(Singh et al., ’04)*

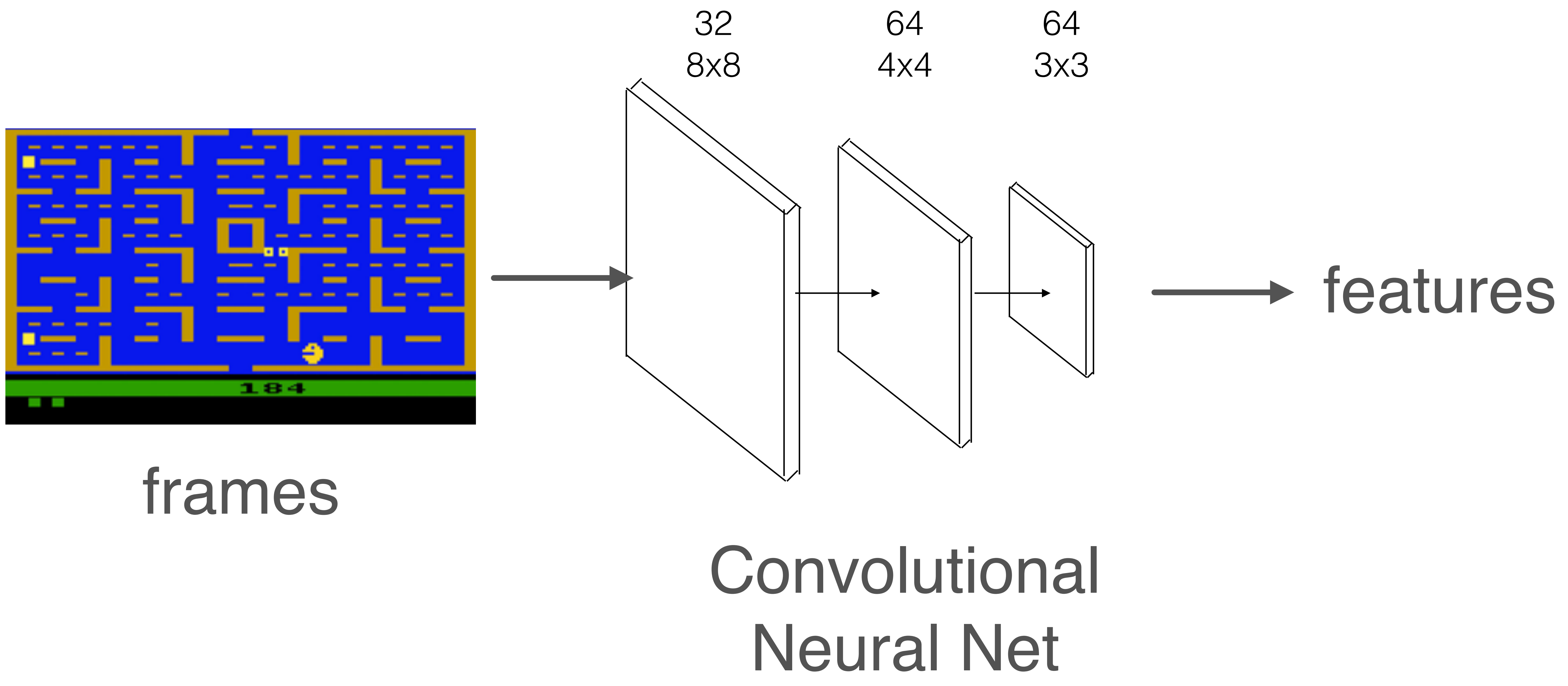
Hierarchical Deep Reinforcement Learning (h-DQN)

Hierarchical Deep Reinforcement Learning (h-DQN)

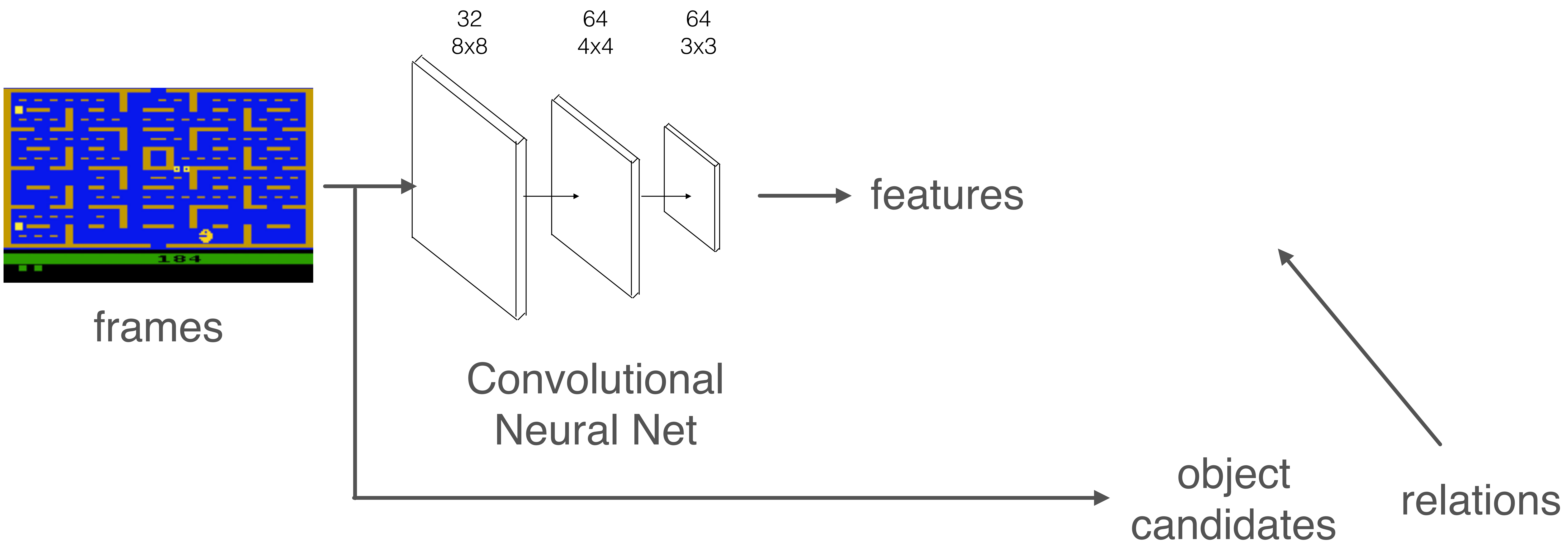


frames

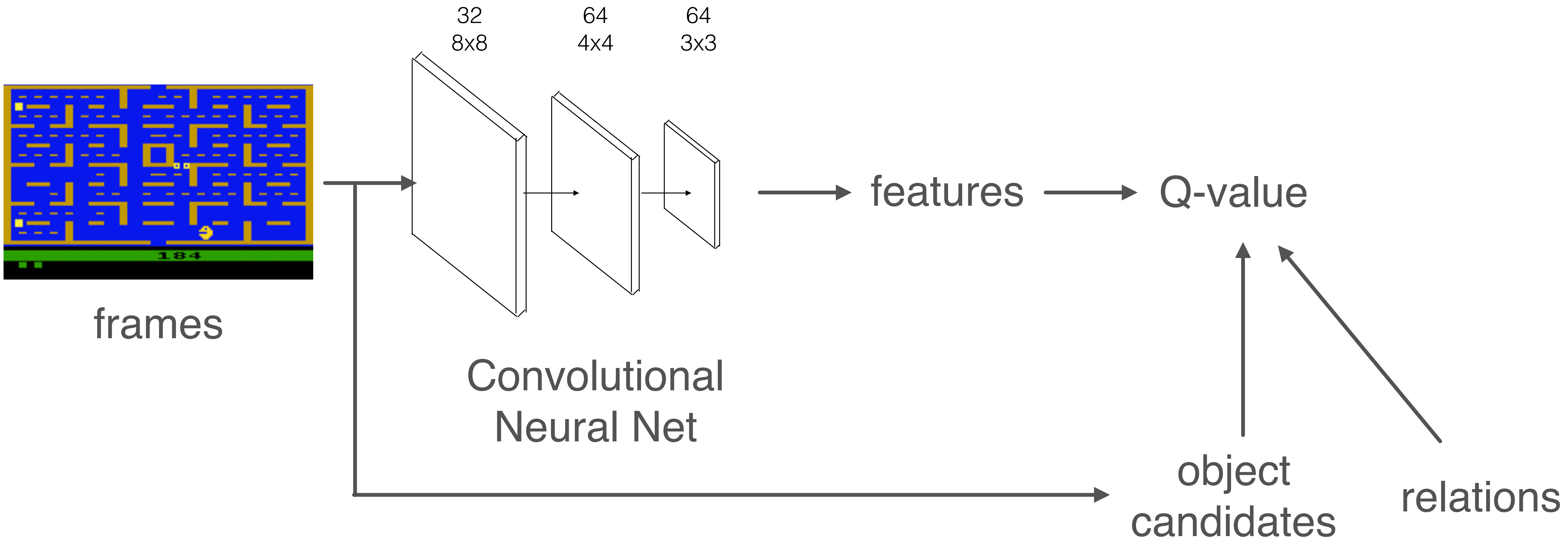
Hierarchical Deep Reinforcement Learning (h-DQN)



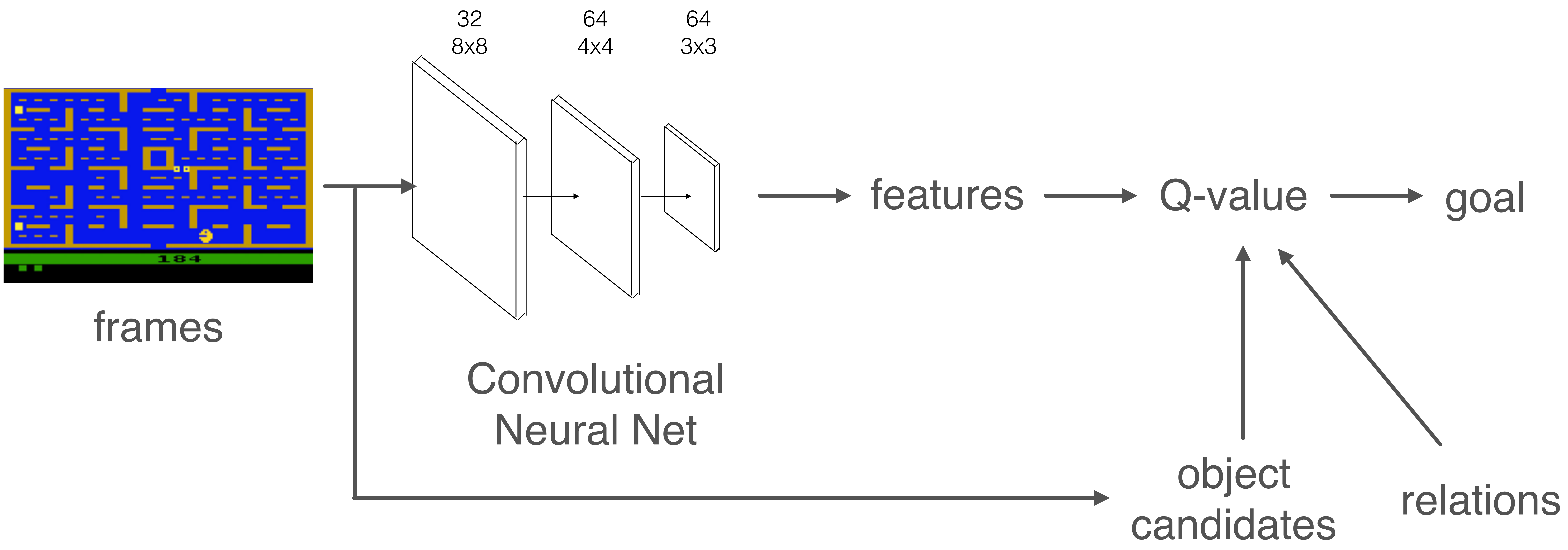
Hierarchical Deep Reinforcement Learning (h-DQN)



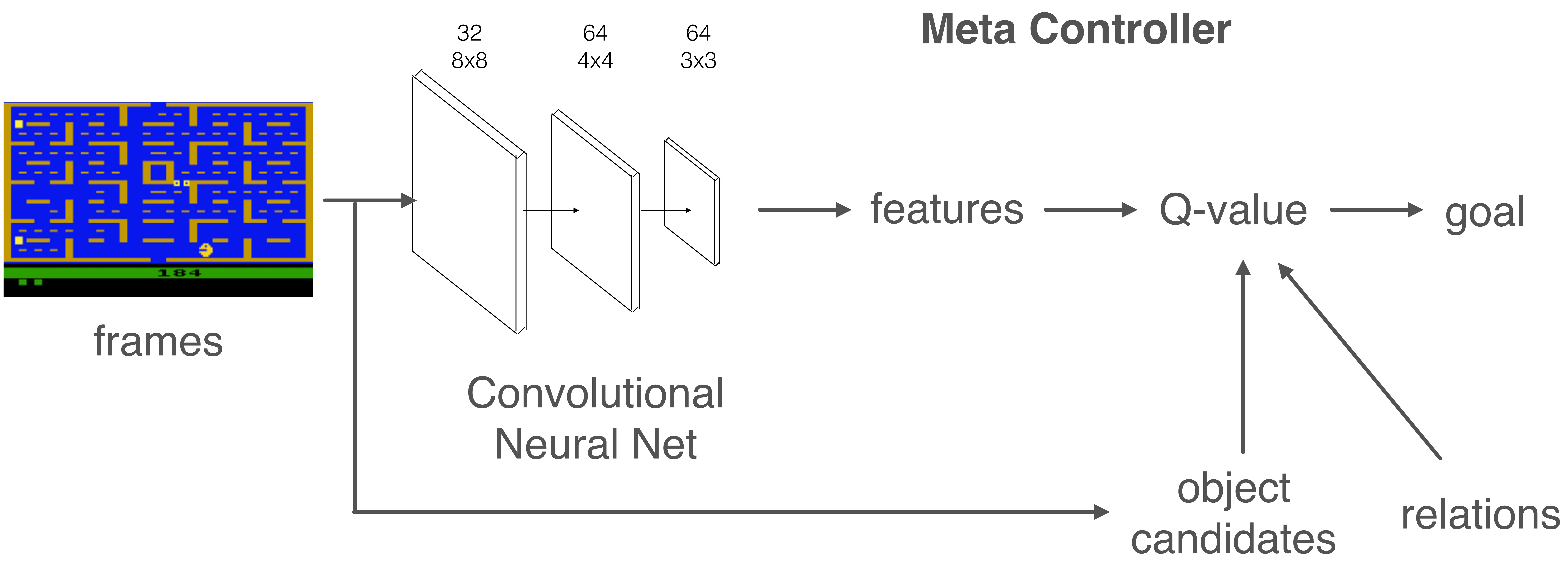
Hierarchical Deep Reinforcement Learning (h-DQN)



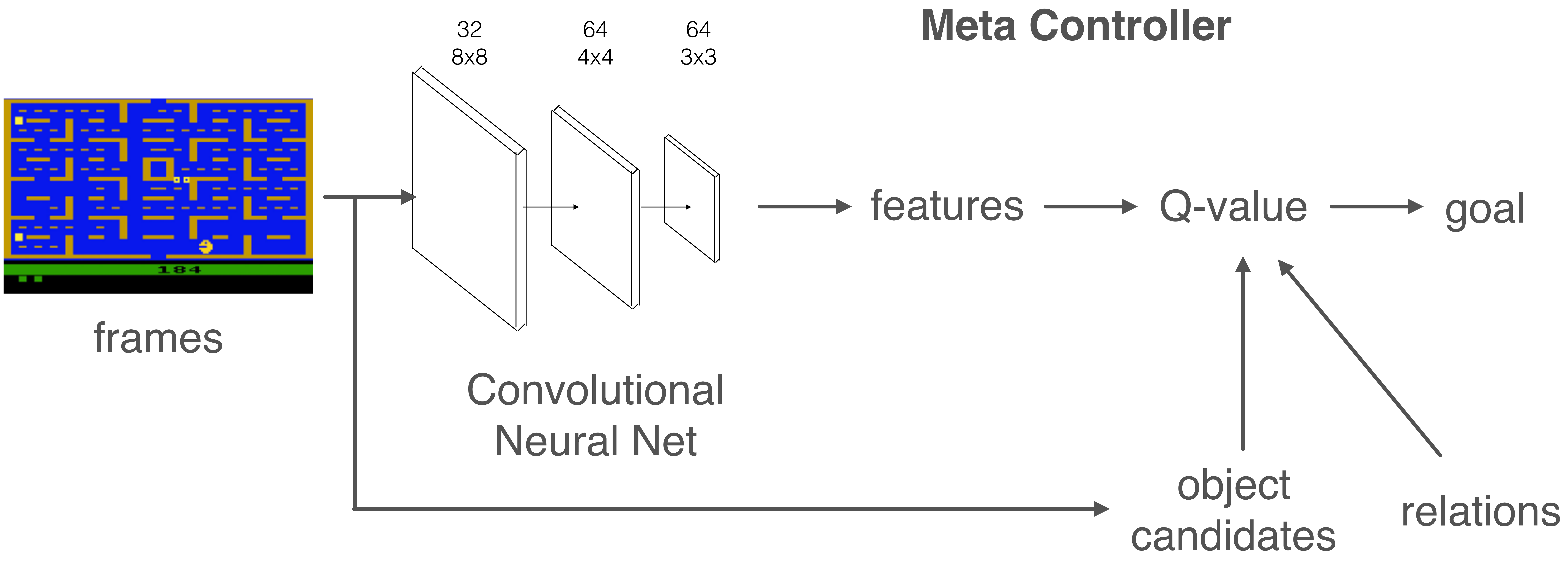
Hierarchical Deep Reinforcement Learning (h-DQN)



Hierarchical Deep Reinforcement Learning (h-DQN)

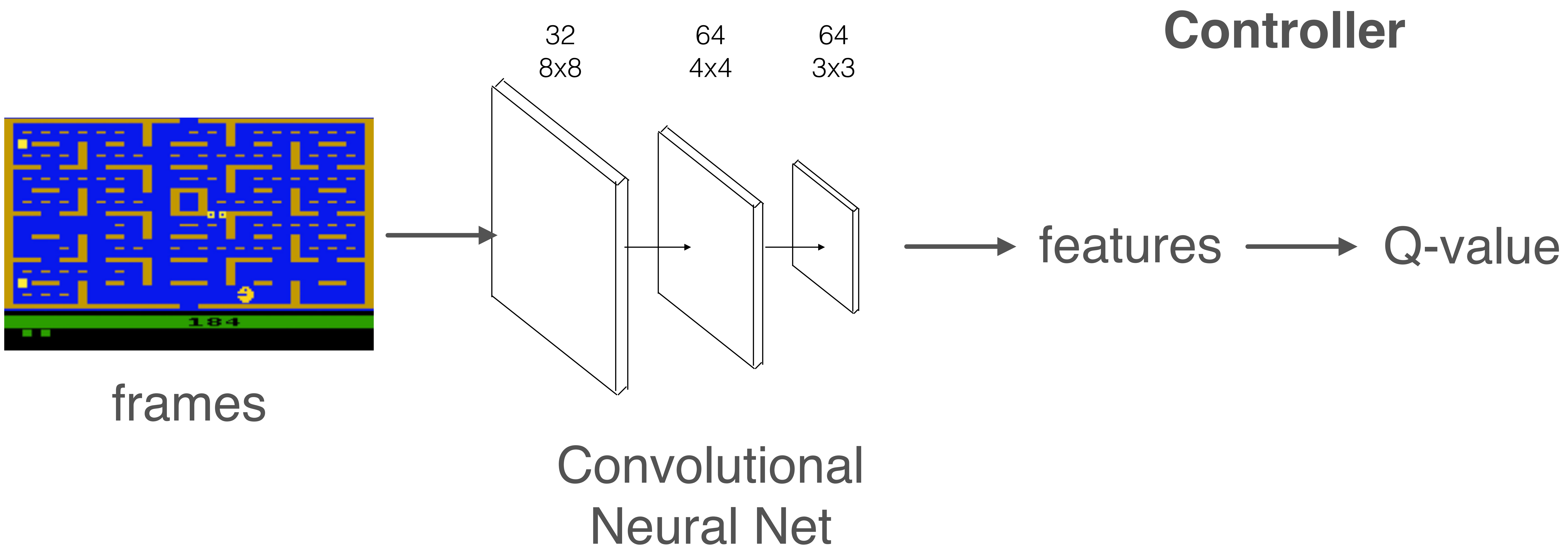


Hierarchical Deep Reinforcement Learning (h-DQN)

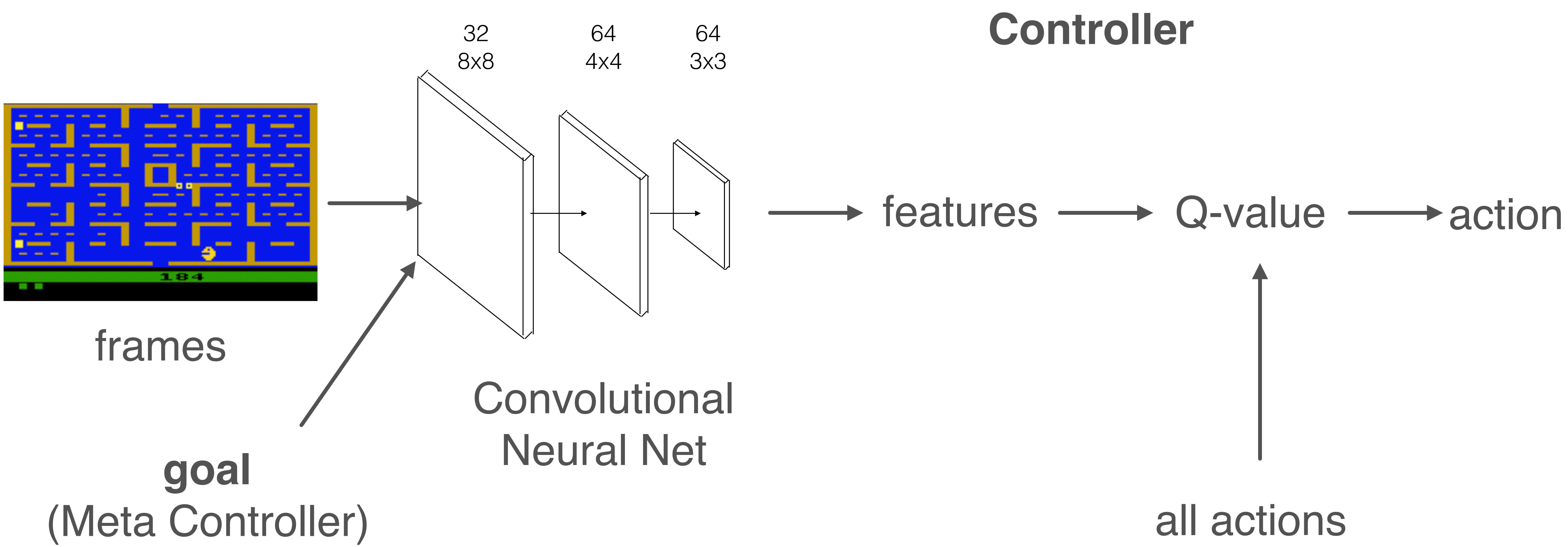


symbolic goals: *<object1, relation, object2>* *<agent, go-near, pellet>*

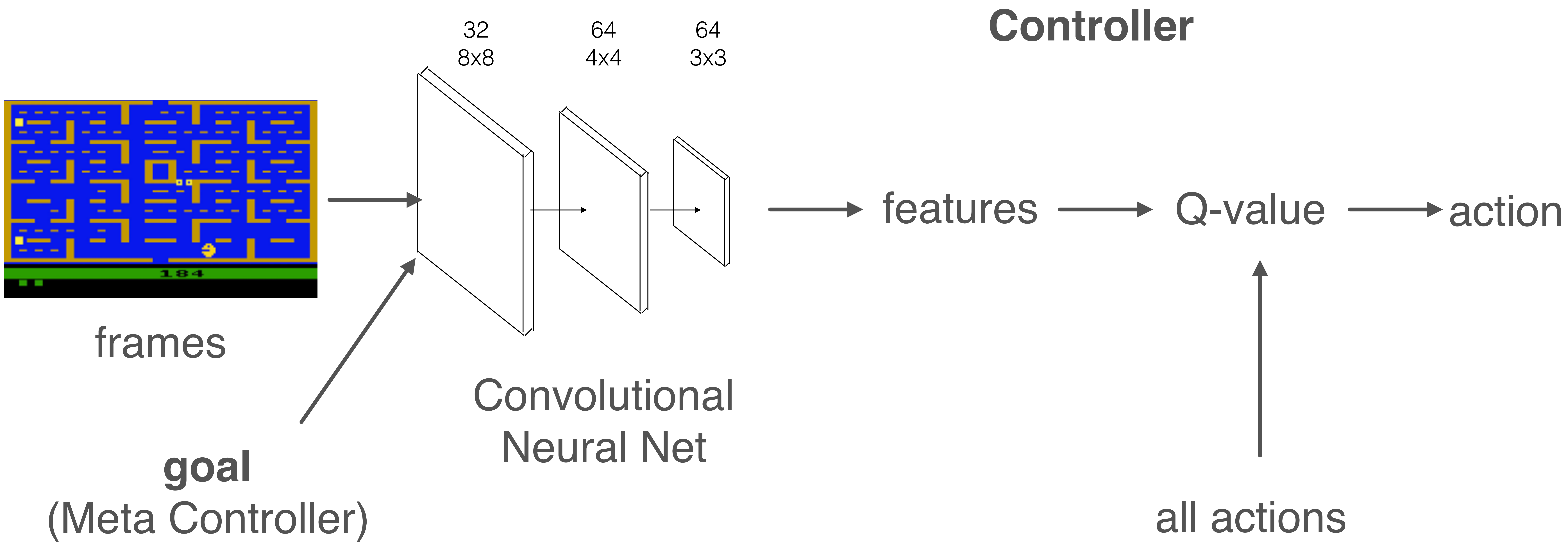
Hierarchical Deep Reinforcement Learning (h-DQN)



Hierarchical Deep Reinforcement Learning (h-DQN)



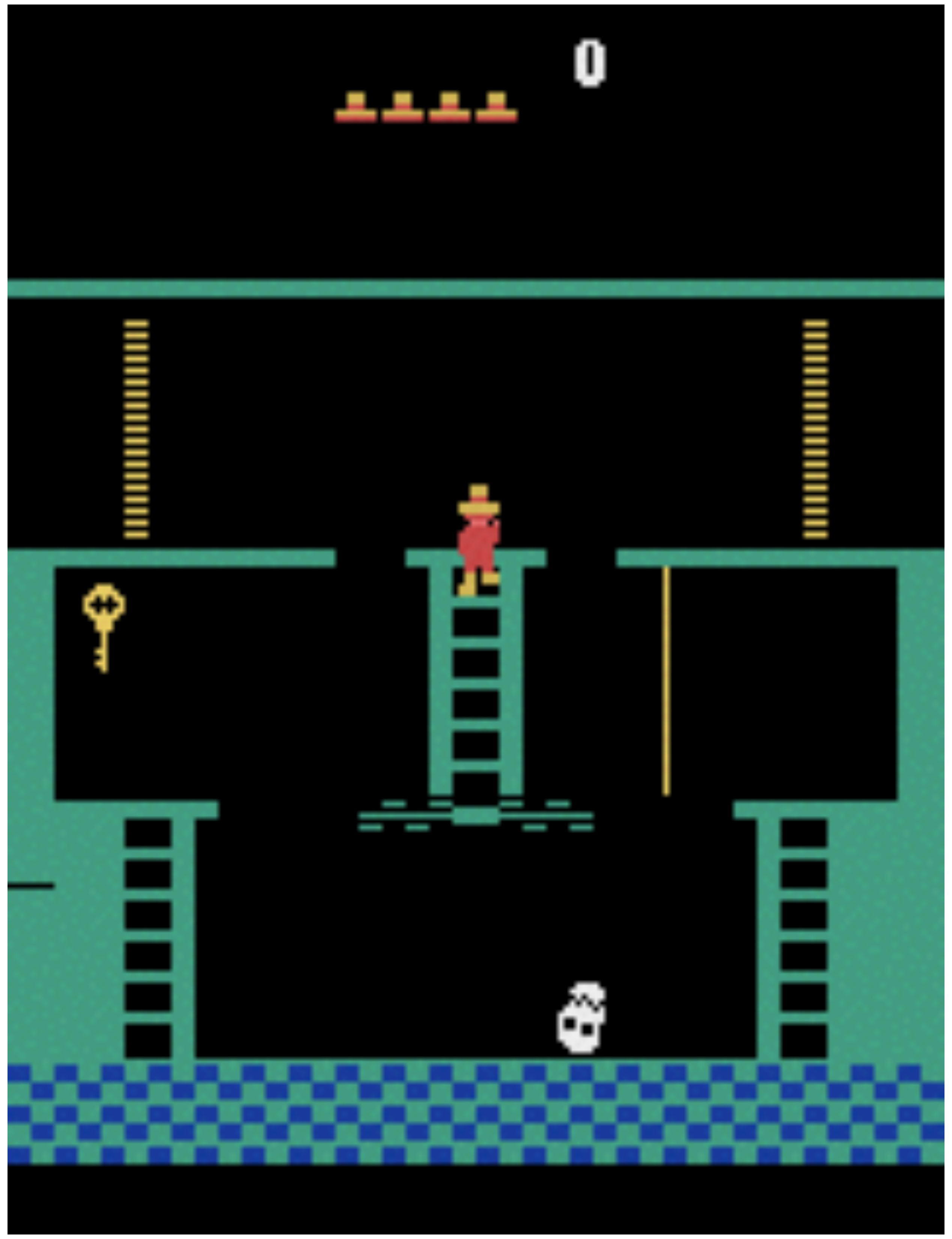
Hierarchical Deep Reinforcement Learning (h-DQN)



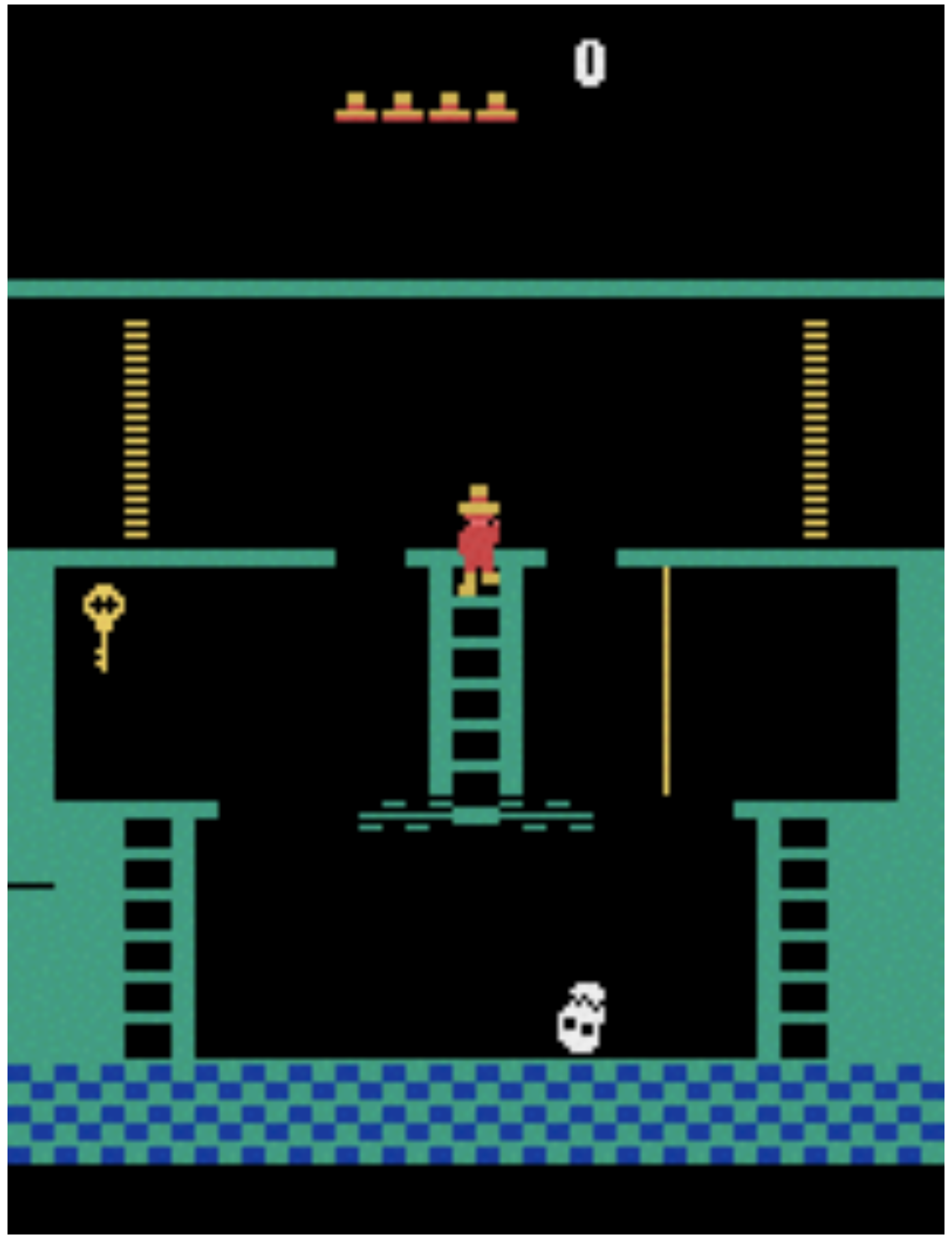
actions: *left, right, up, down, ….*

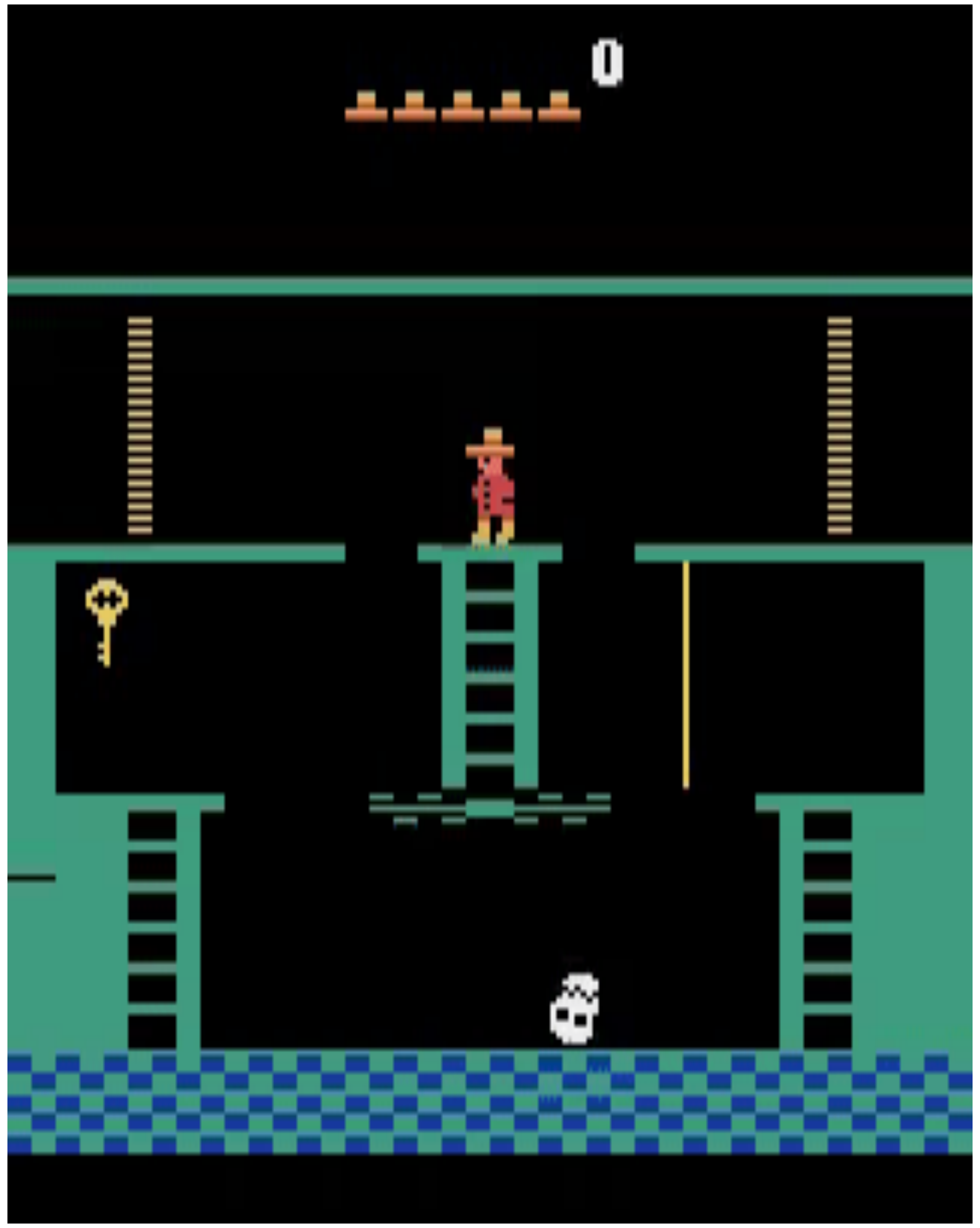
Intrinsically motivated exploration and learning in task space

Intrinsically motivated exploration and learning in task space



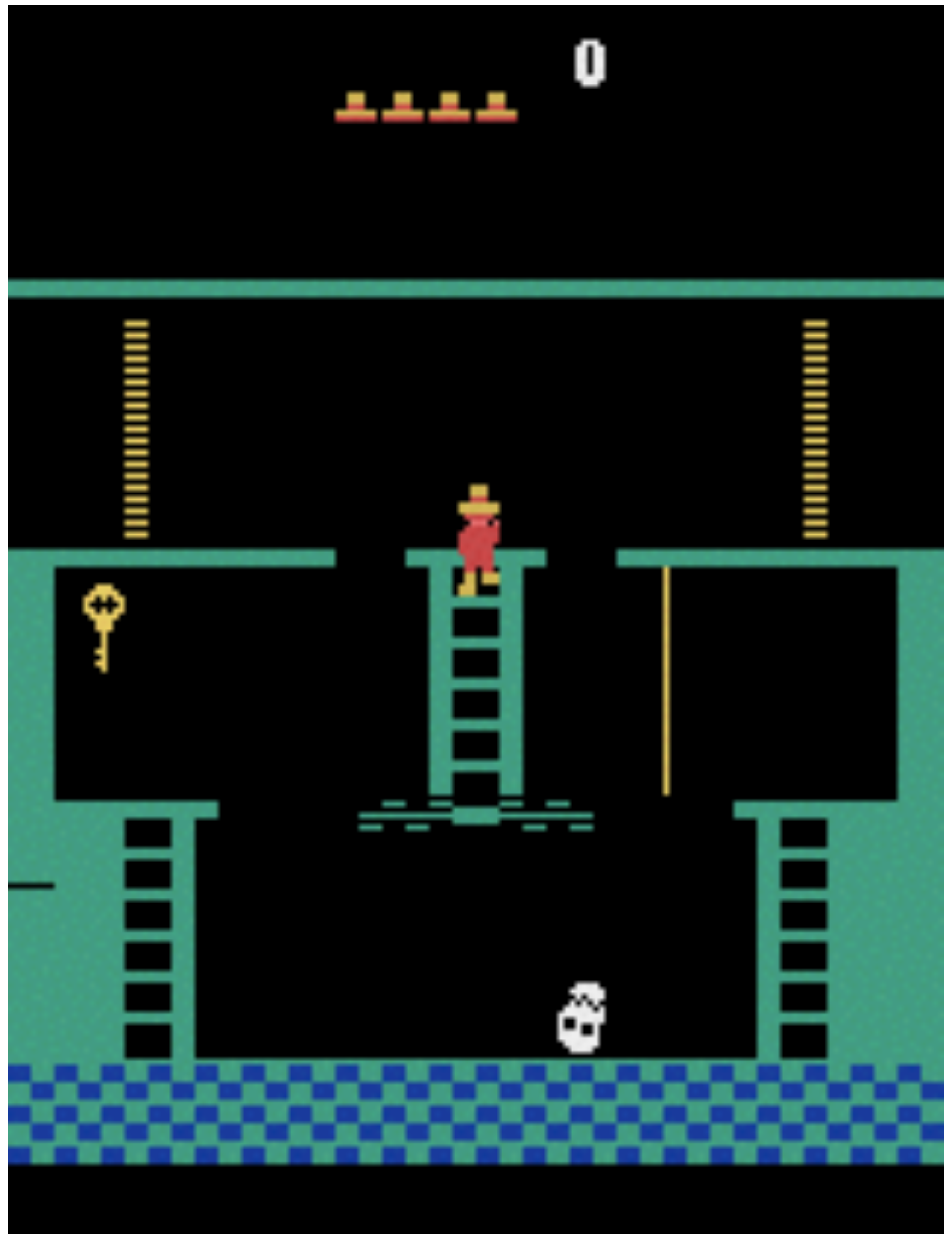
Intrinsically motivated exploration and learning in task space

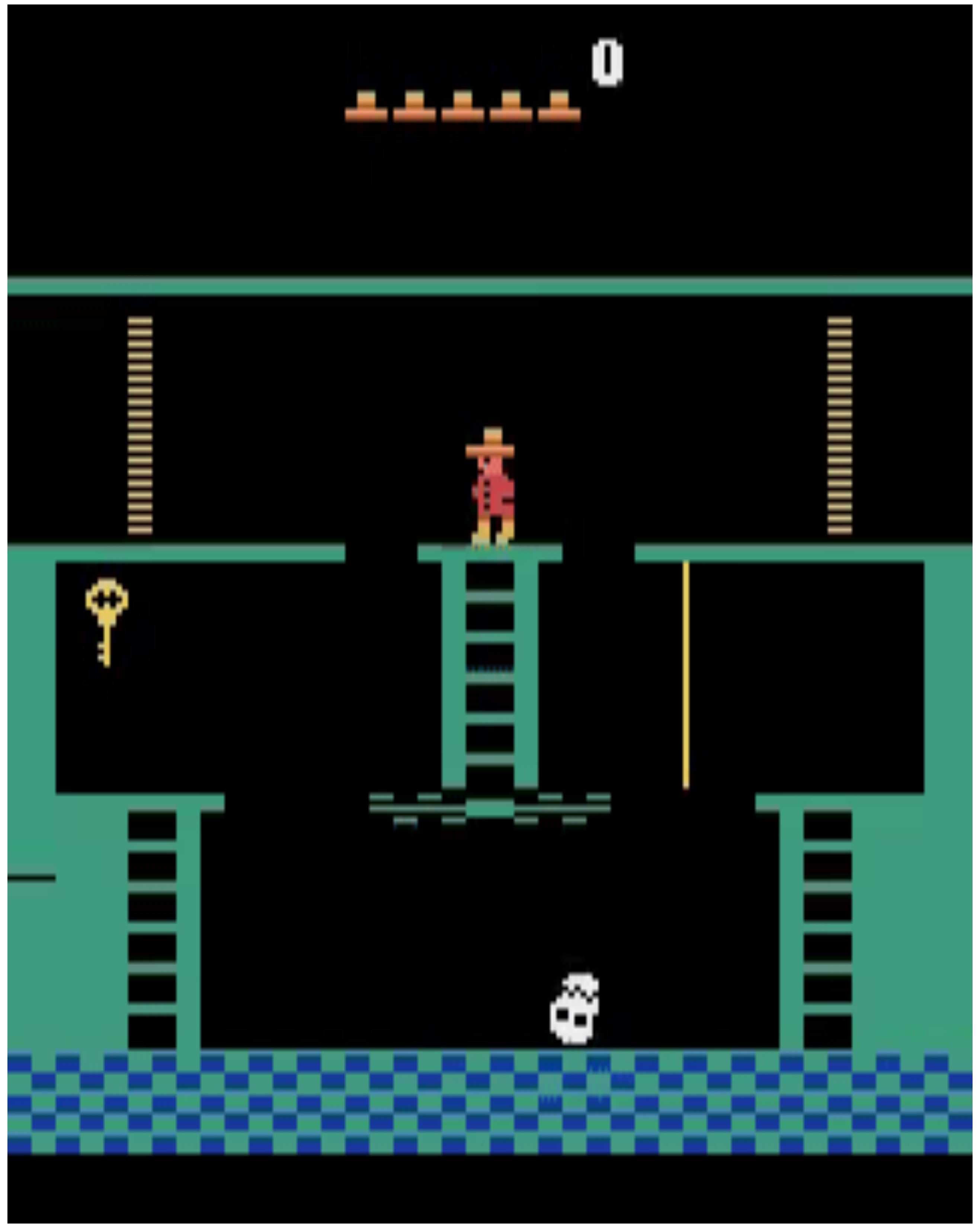




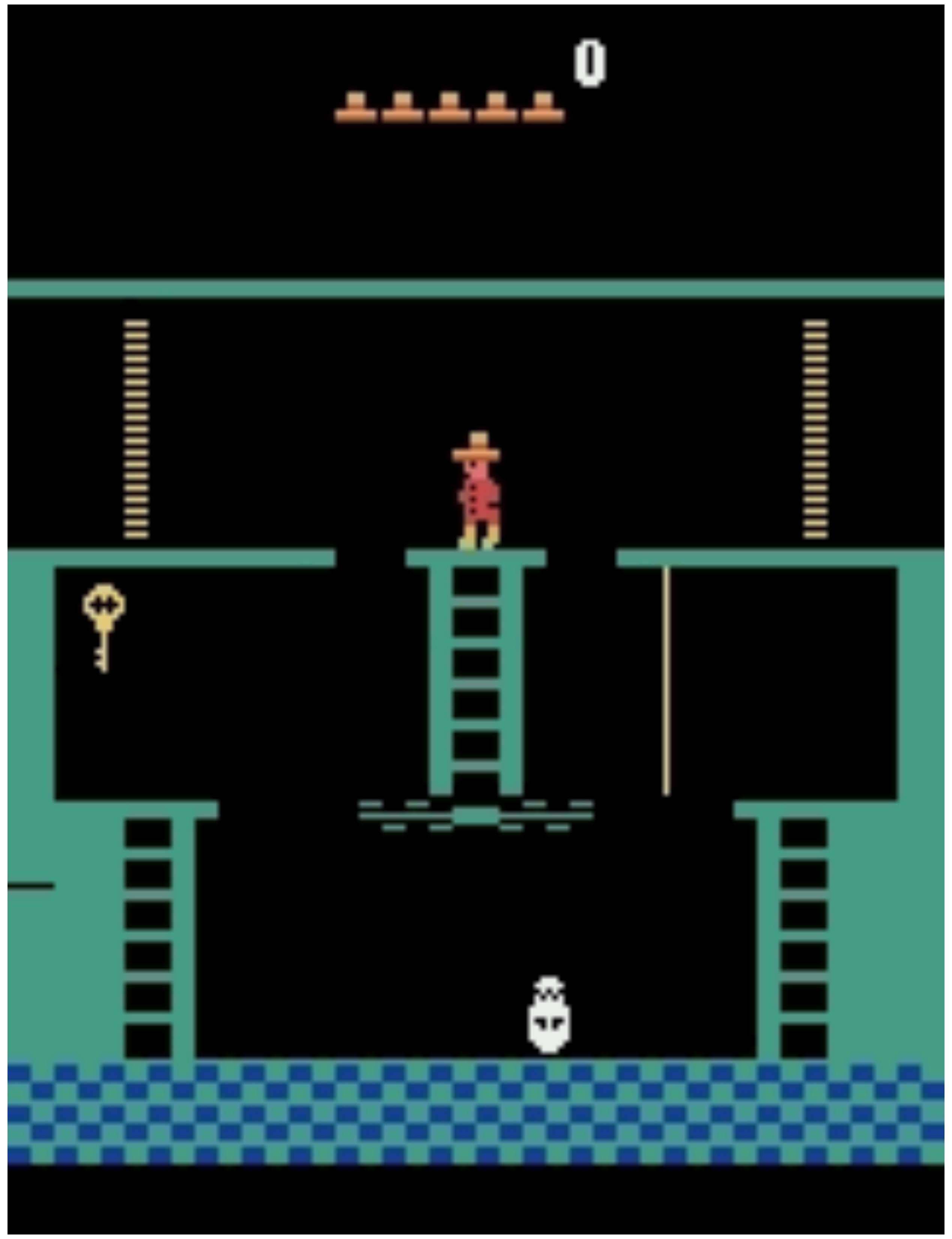
DQN (epsilon greedy exploration)

Intrinsically motivated exploration and learning in task space





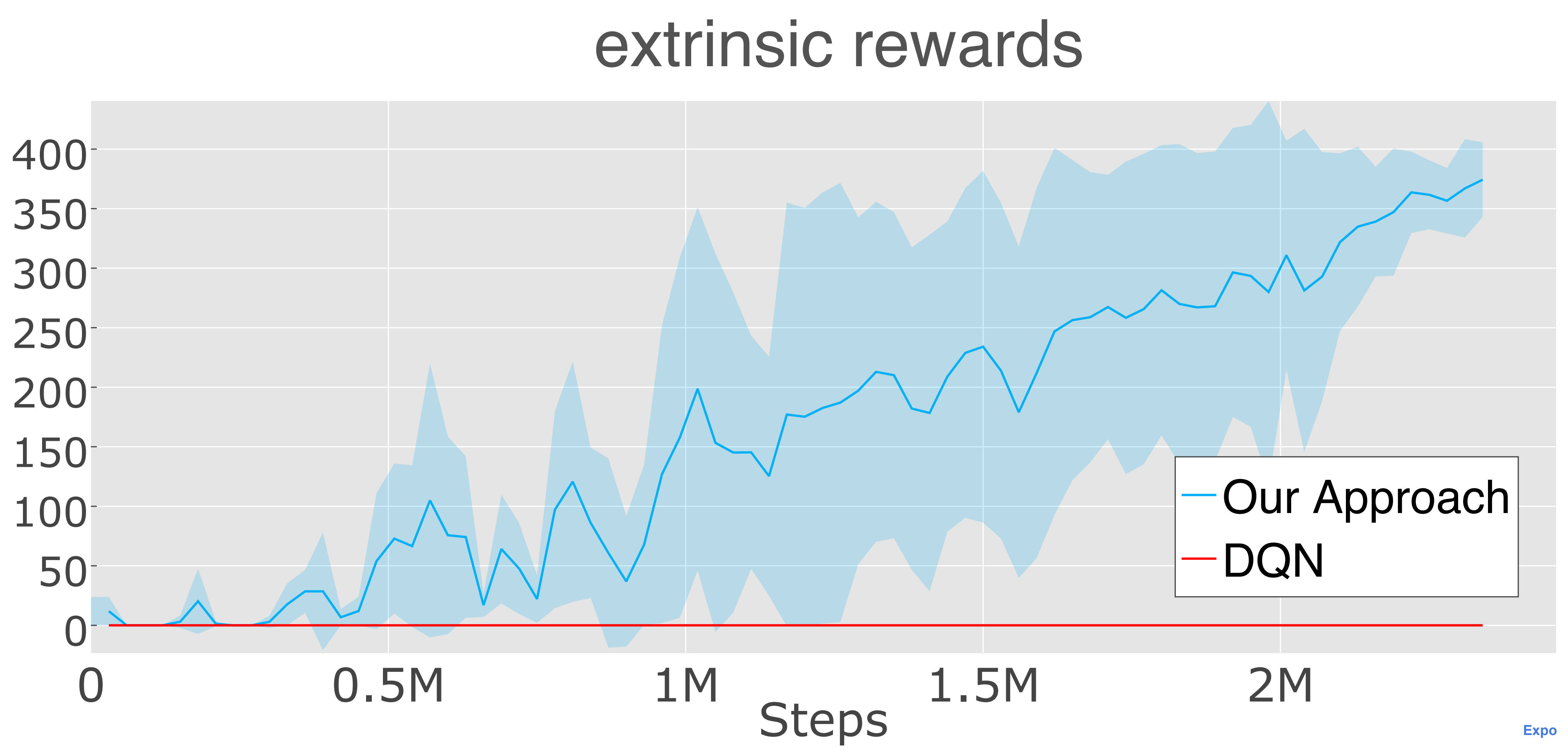
DQN (epsilon greedy exploration)



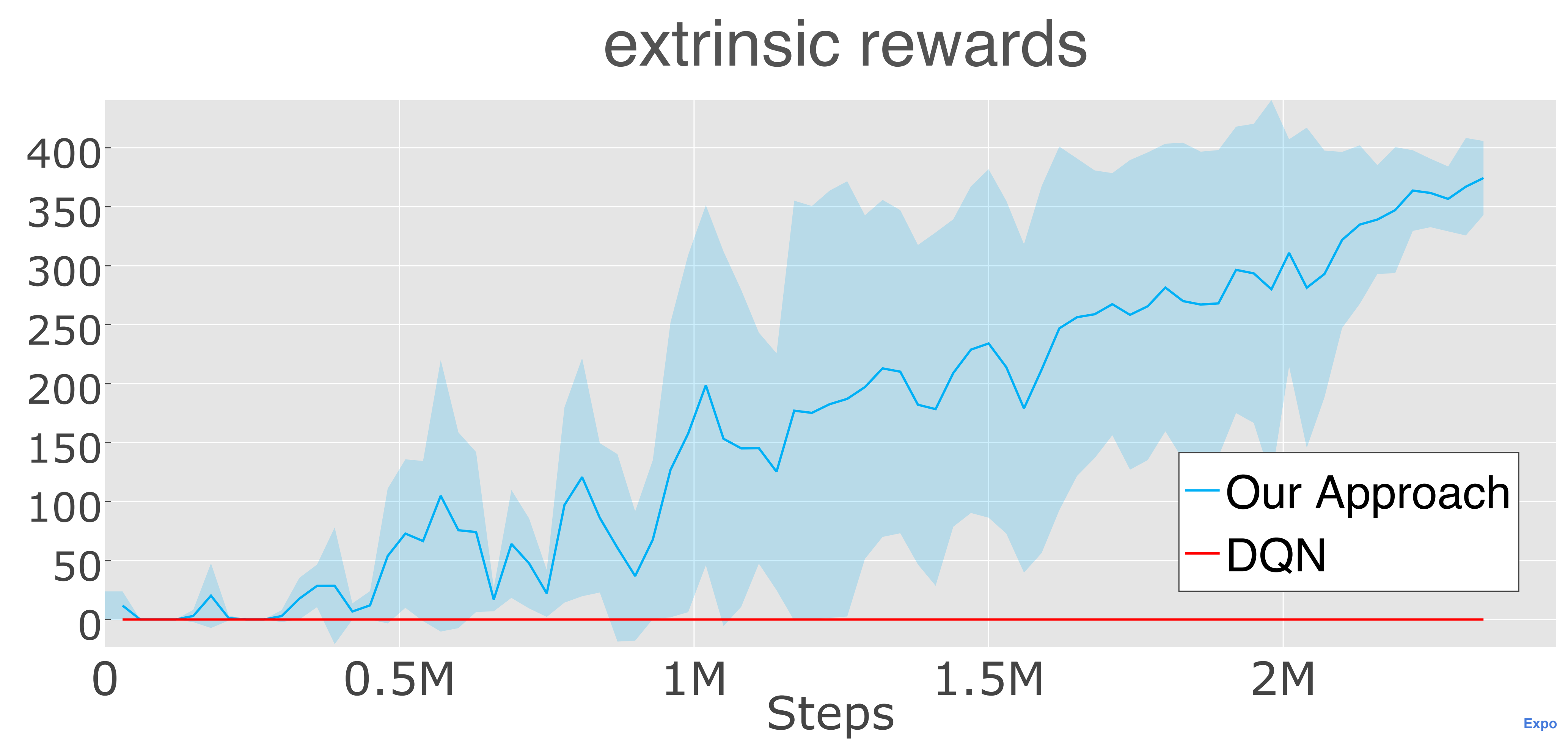
h-DQN

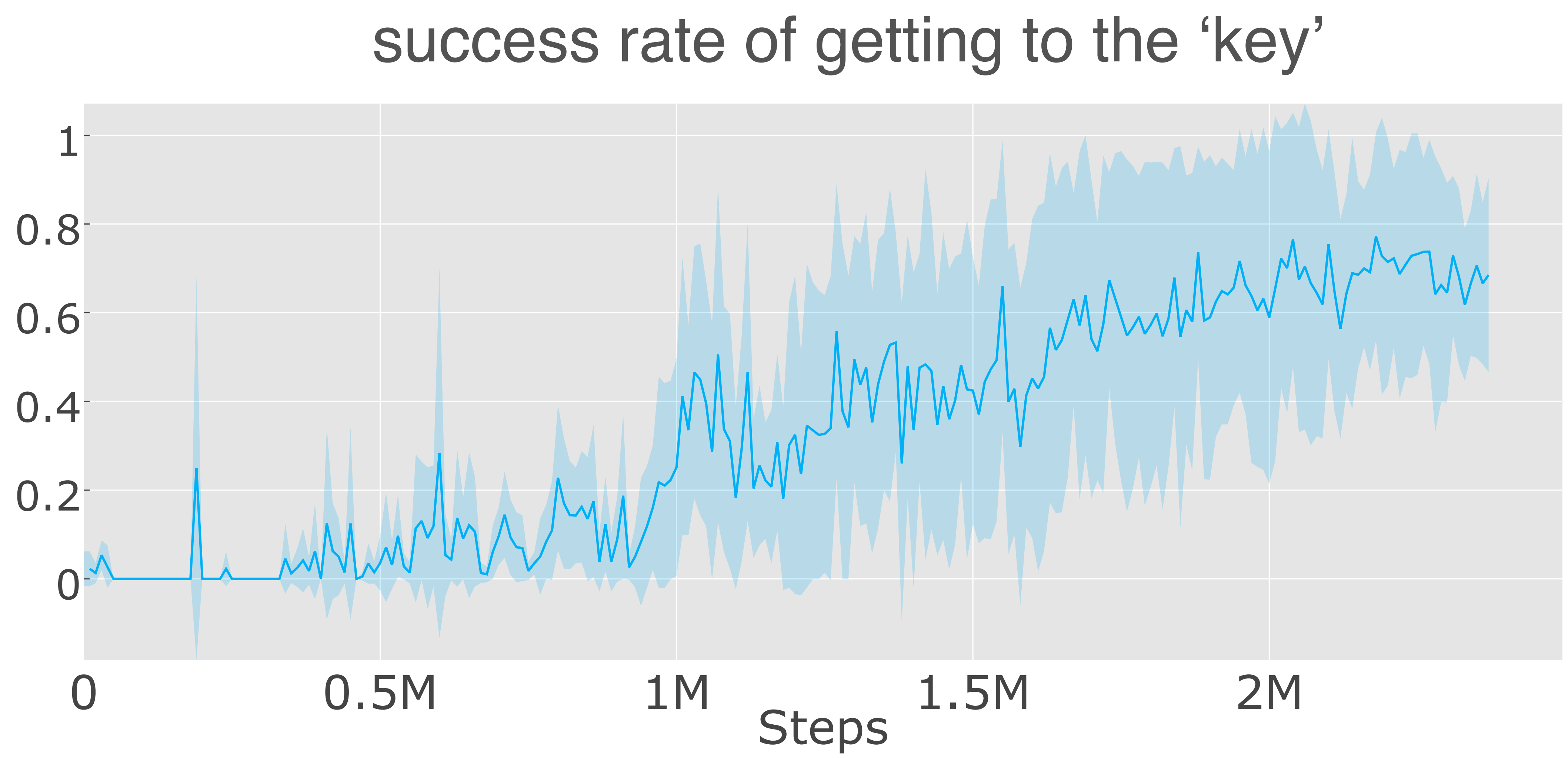
Example

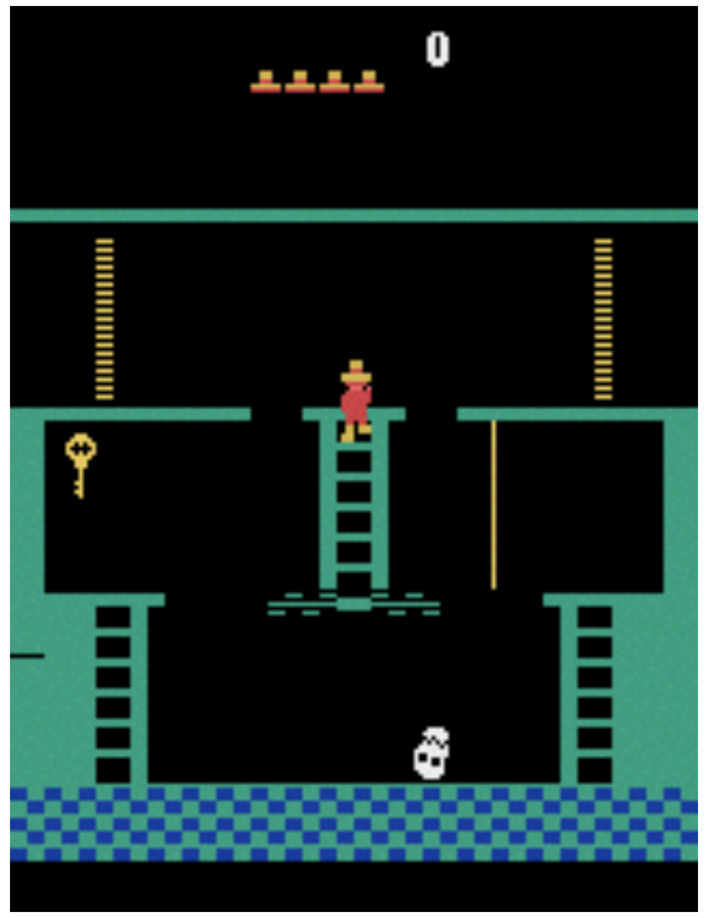
Example


extrinsic rewards
400
350
300
250
200
150
100
50
0
0
0.5M
1M
1.5M
2M
Steps
Our Approach
DQN
Expo

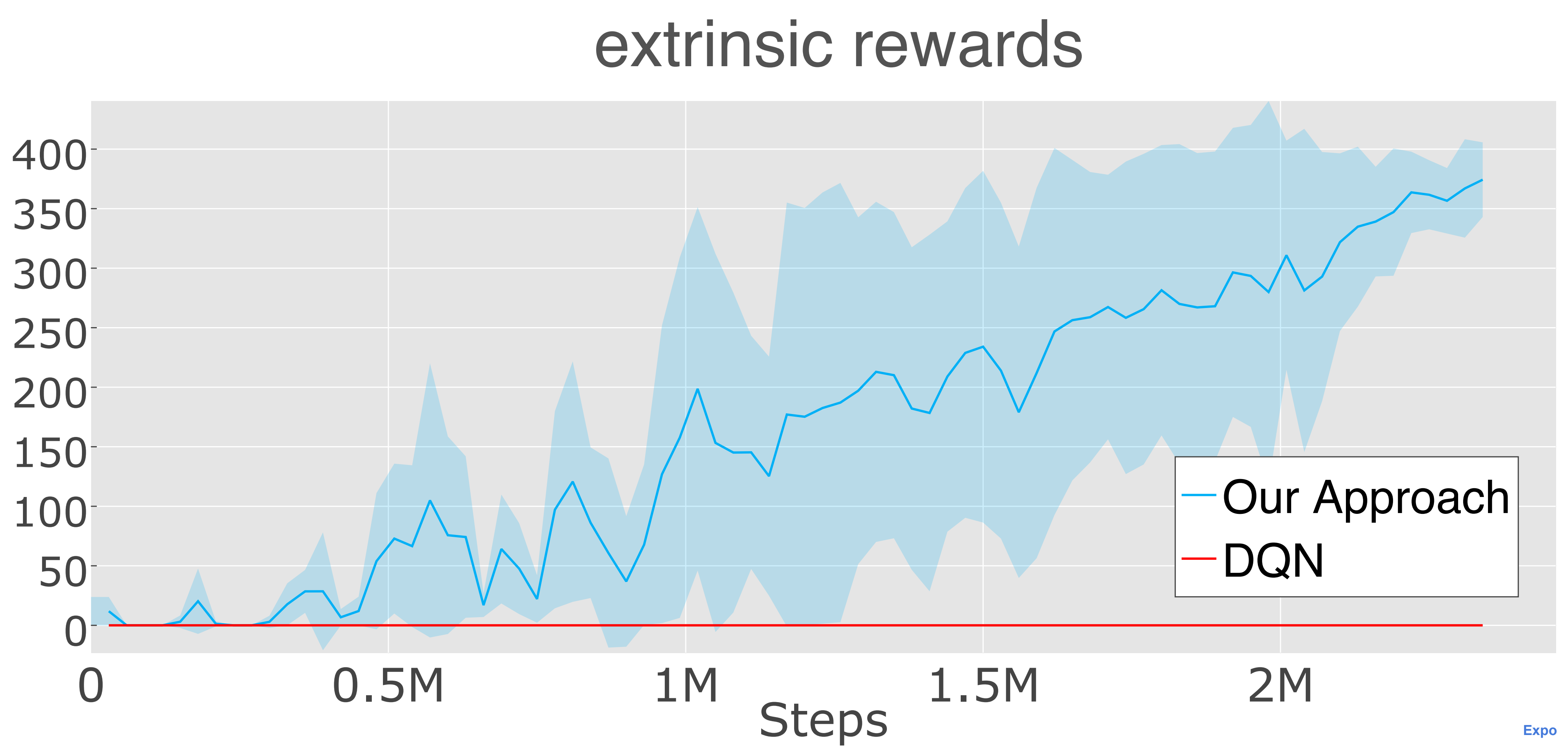
Example

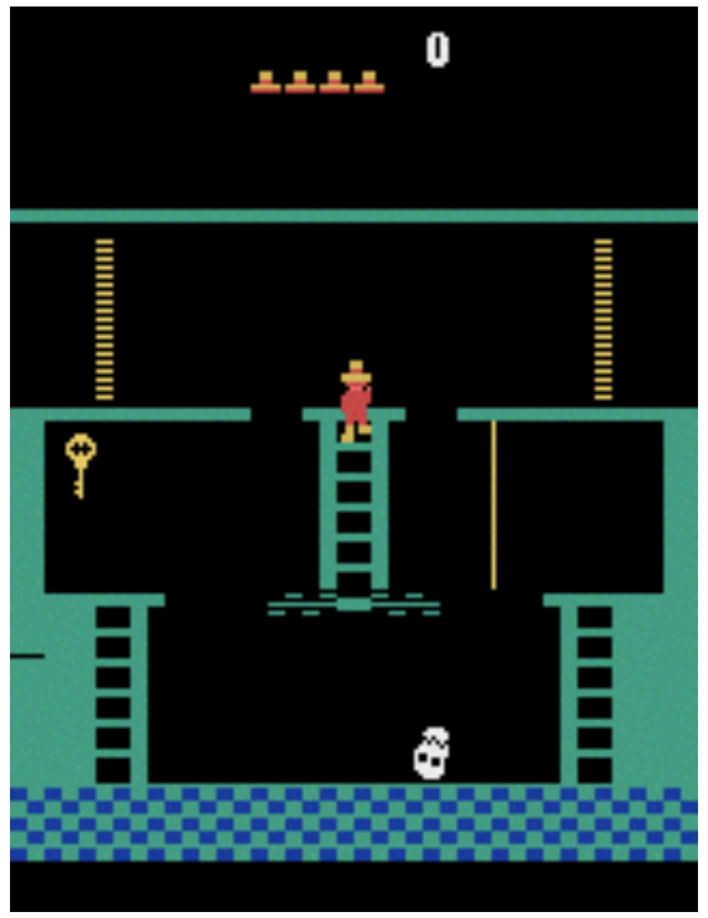

extrinsic rewards
400
350
300
250
200
150
100
50
0
0
0.5M
1M
1.5M
2M
Steps
Our Approach
DQN
Expo

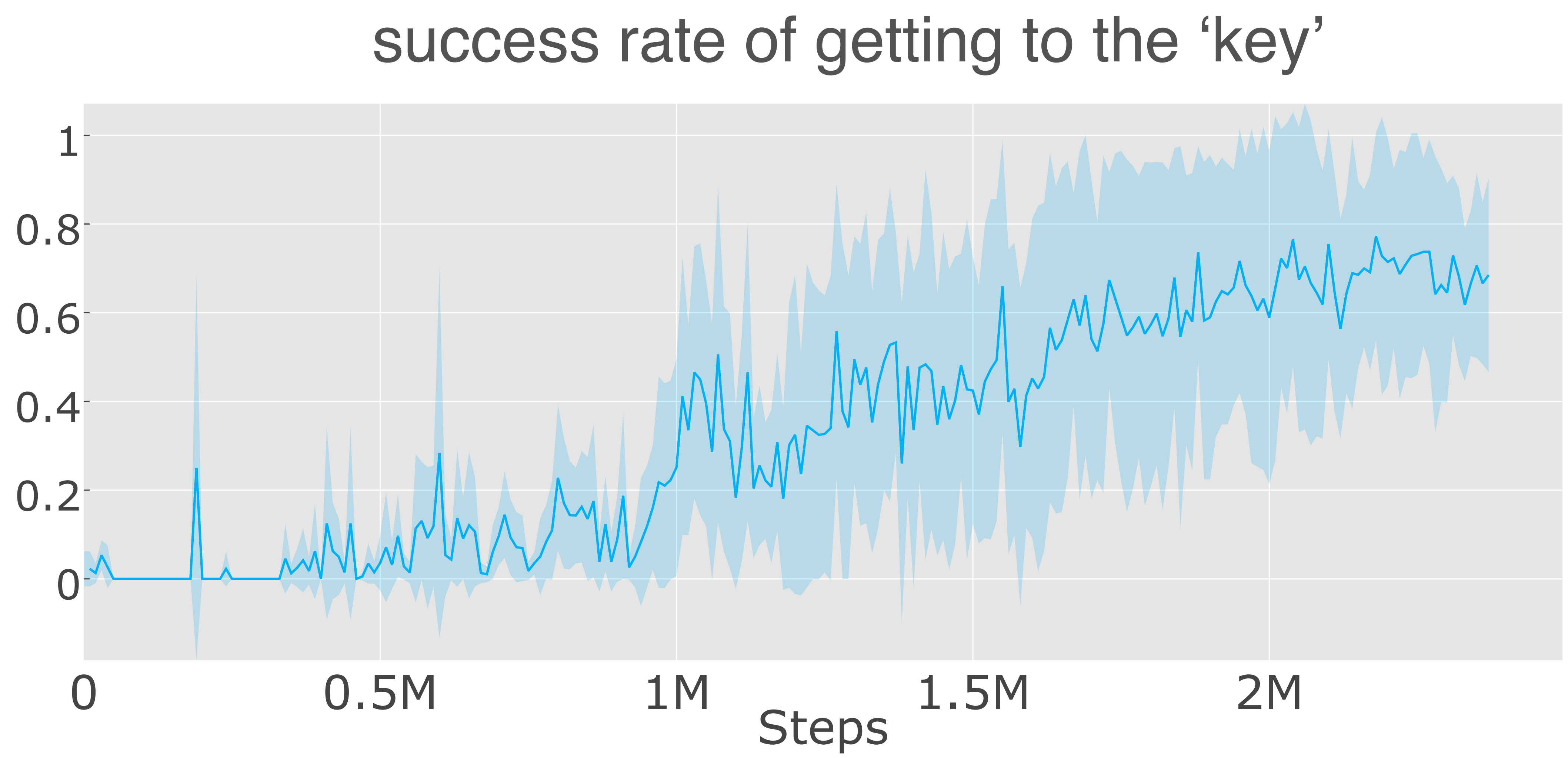

success rate of getting to the ‘key’
1
0.8
0.6
0.4
0.2
0
0
0.5M
1M
1.5M
2M
Steps

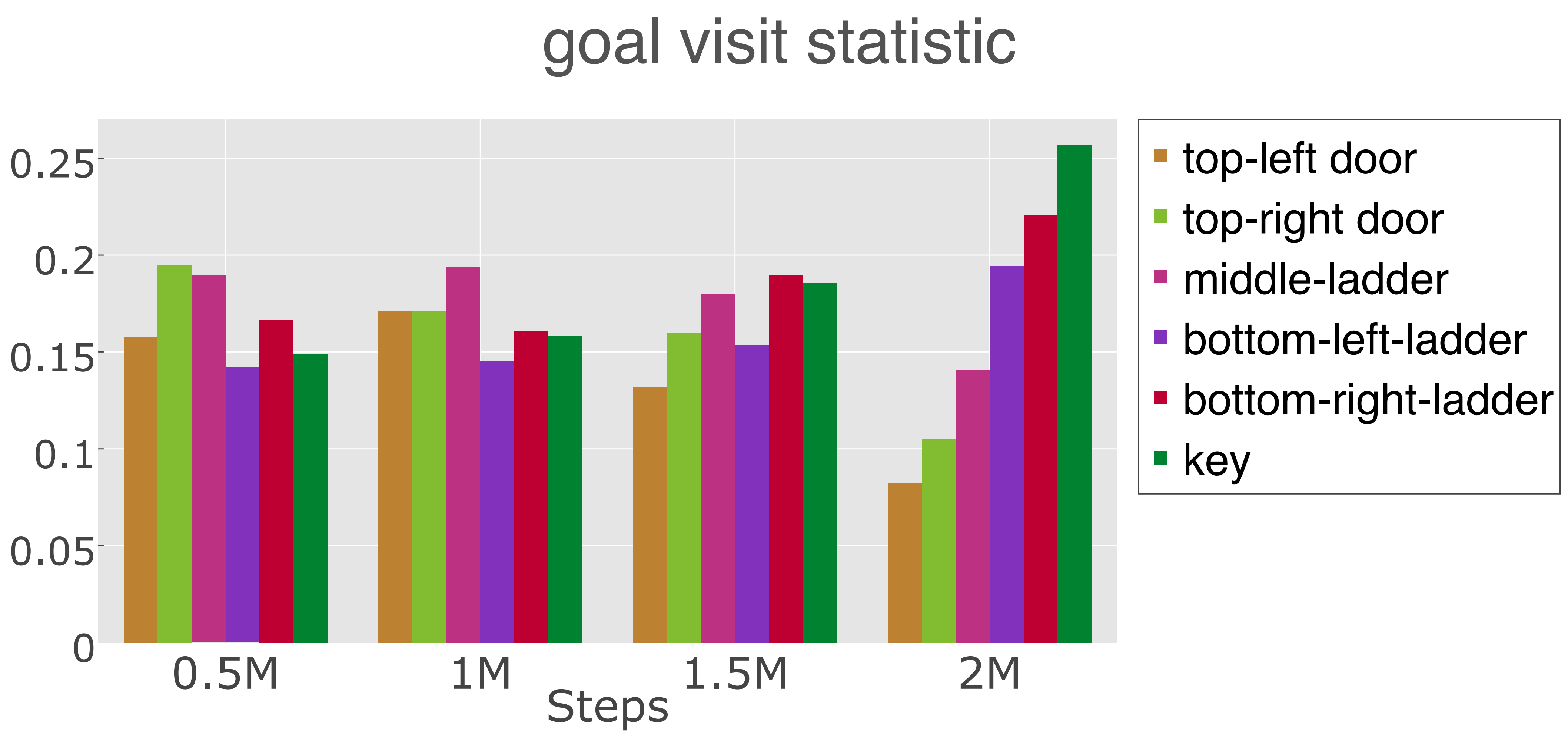

0

Example

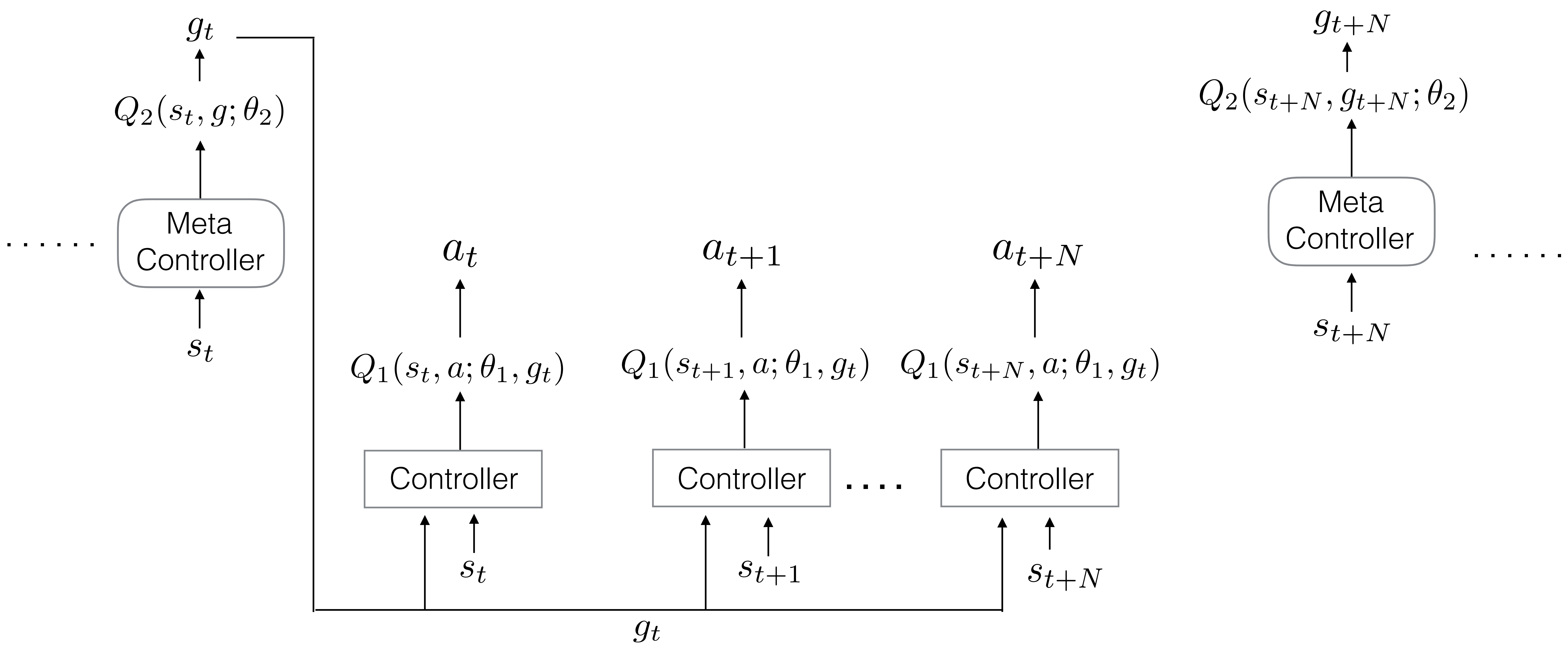

extrinsic rewards
400
350
300
250
200
150
100
50
0
0
0.5M
1M
1.5M
2M
Steps
Our Approach
DQN
Expo


0


success rate of getting to the 'key'
1
0.8
0.6
0.4
0.2
0
0
0.5M
1M
1.5M
2M
Steps

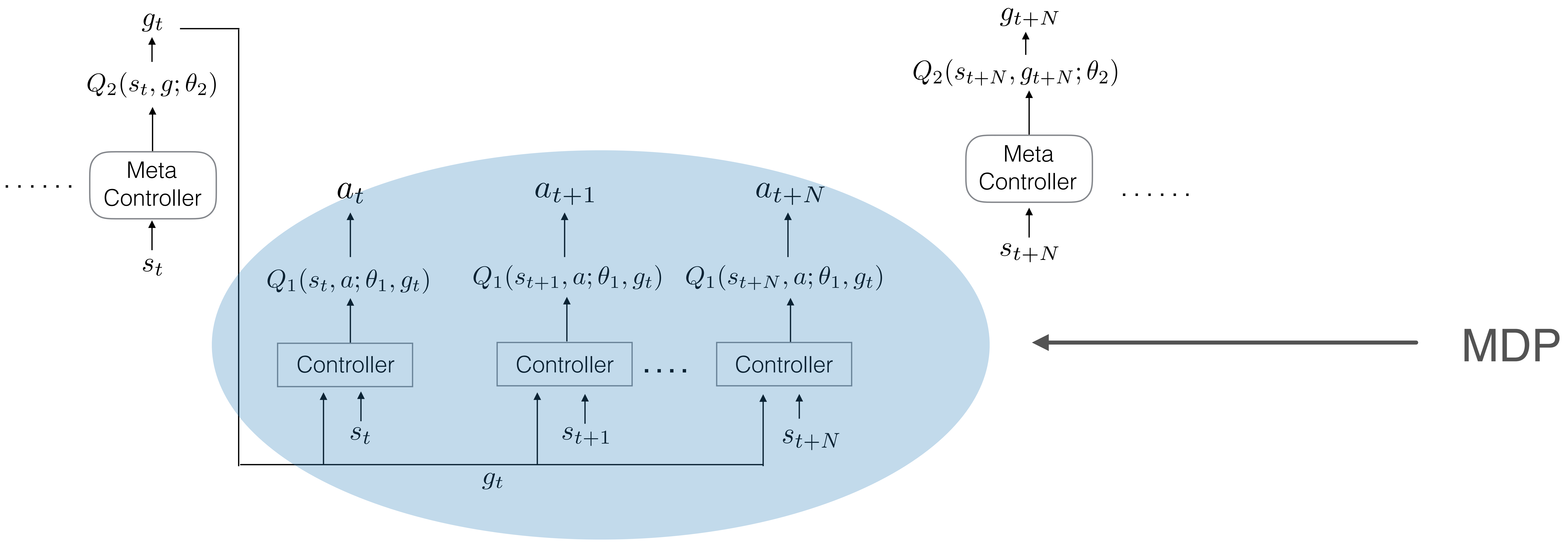

goal visit statistic
0.25
0.2
0.15
0.1
0.05
0
0.5M
1M
1.5M
2M
Steps
top-left door
top-right door
middle-ladder
bottom-left-ladder
bottom-right-ladder
key

Markov Decision Processes (MDPs) and Semi-MDPs



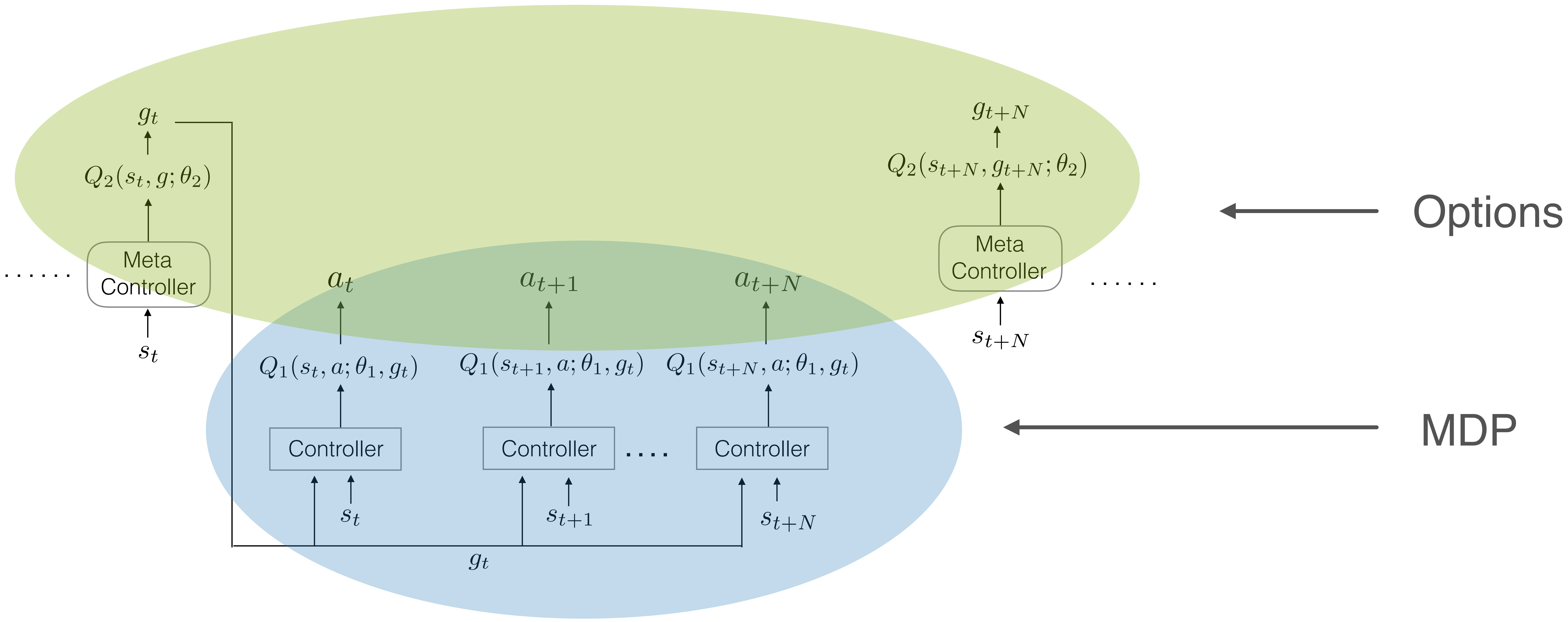
Sutton et al., 1999

Markov Decision Processes (MDPs) and Semi-MDPs



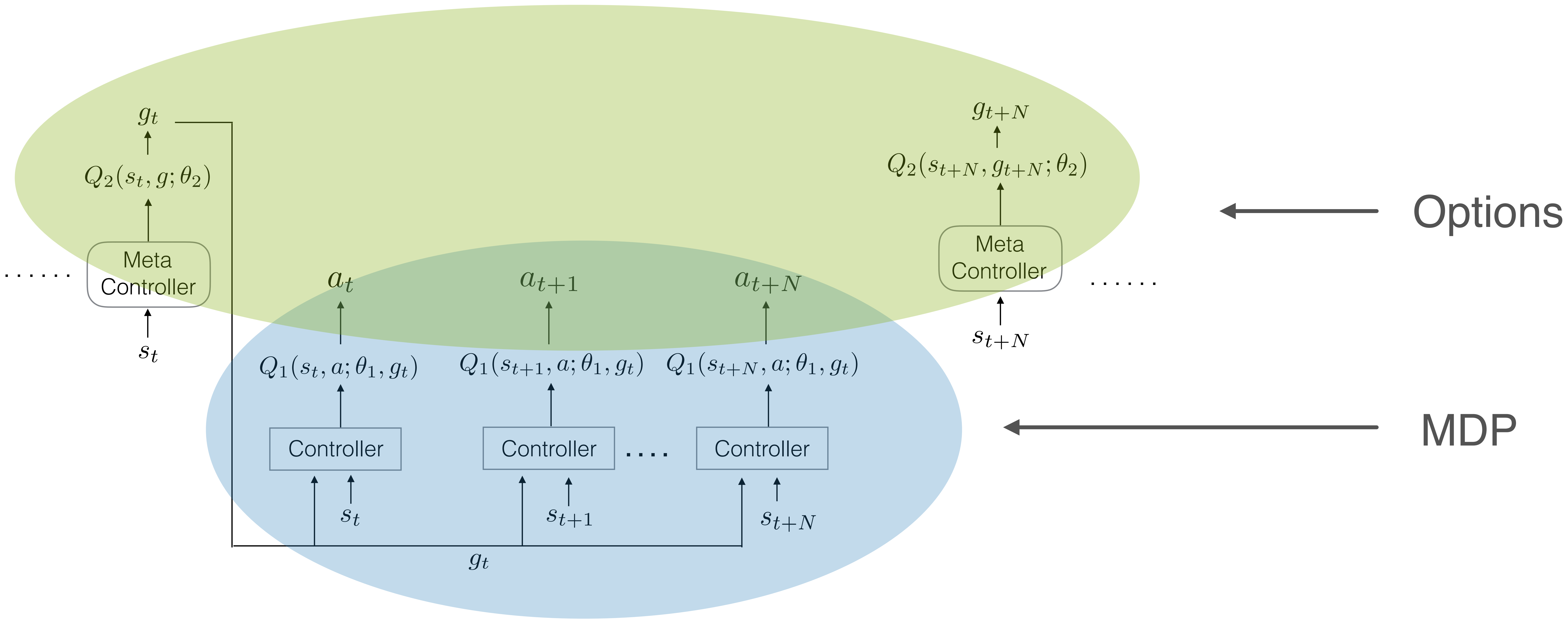
Sutton et al., 1999

Markov Decision Processes (MDPs) and Semi-MDPs



Sutton et al., 1999

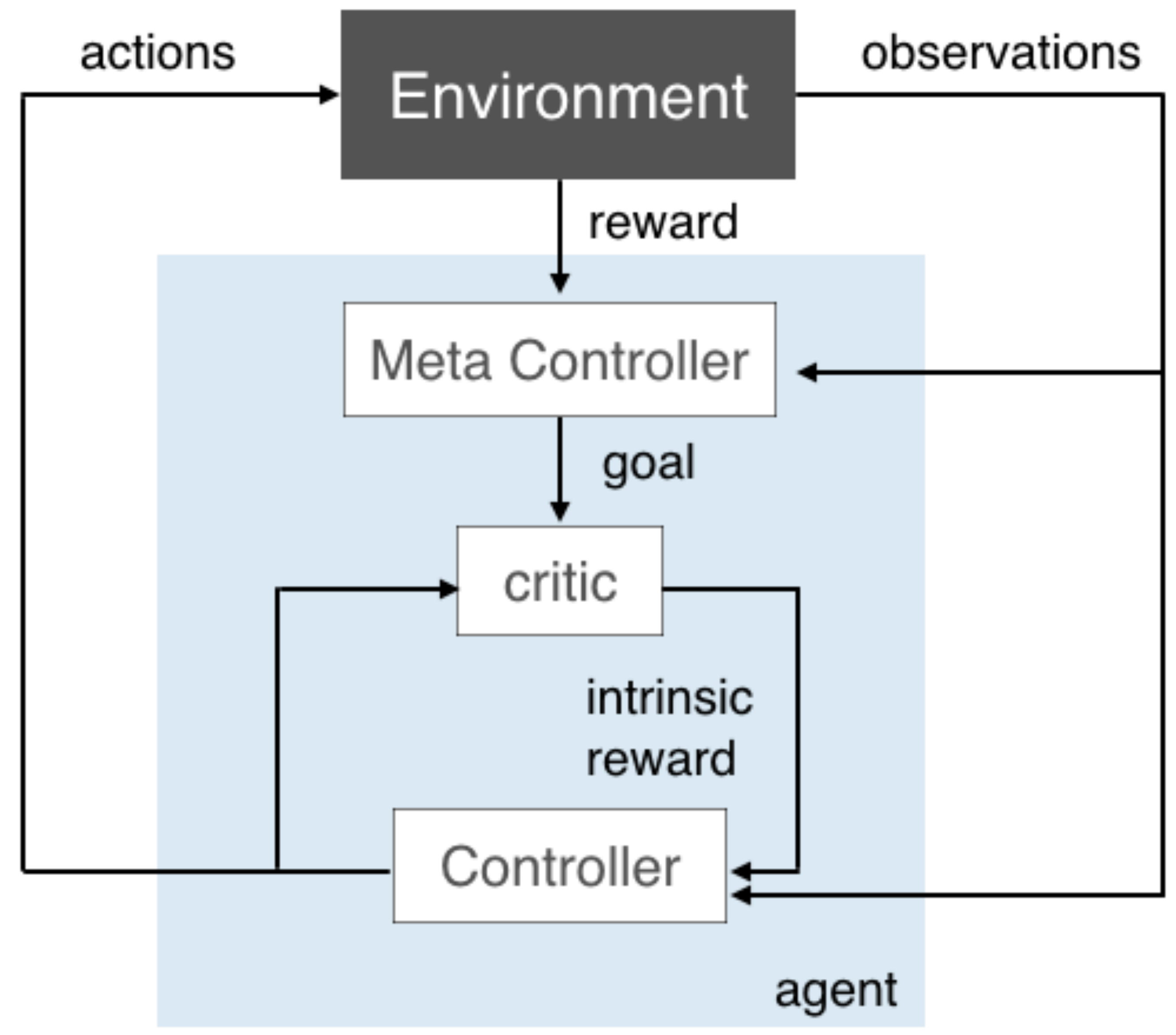
Markov Decision Processes (MDPs) and Semi-MDPs



Options + MDP = Semi MDP

Sutton et al., 1999

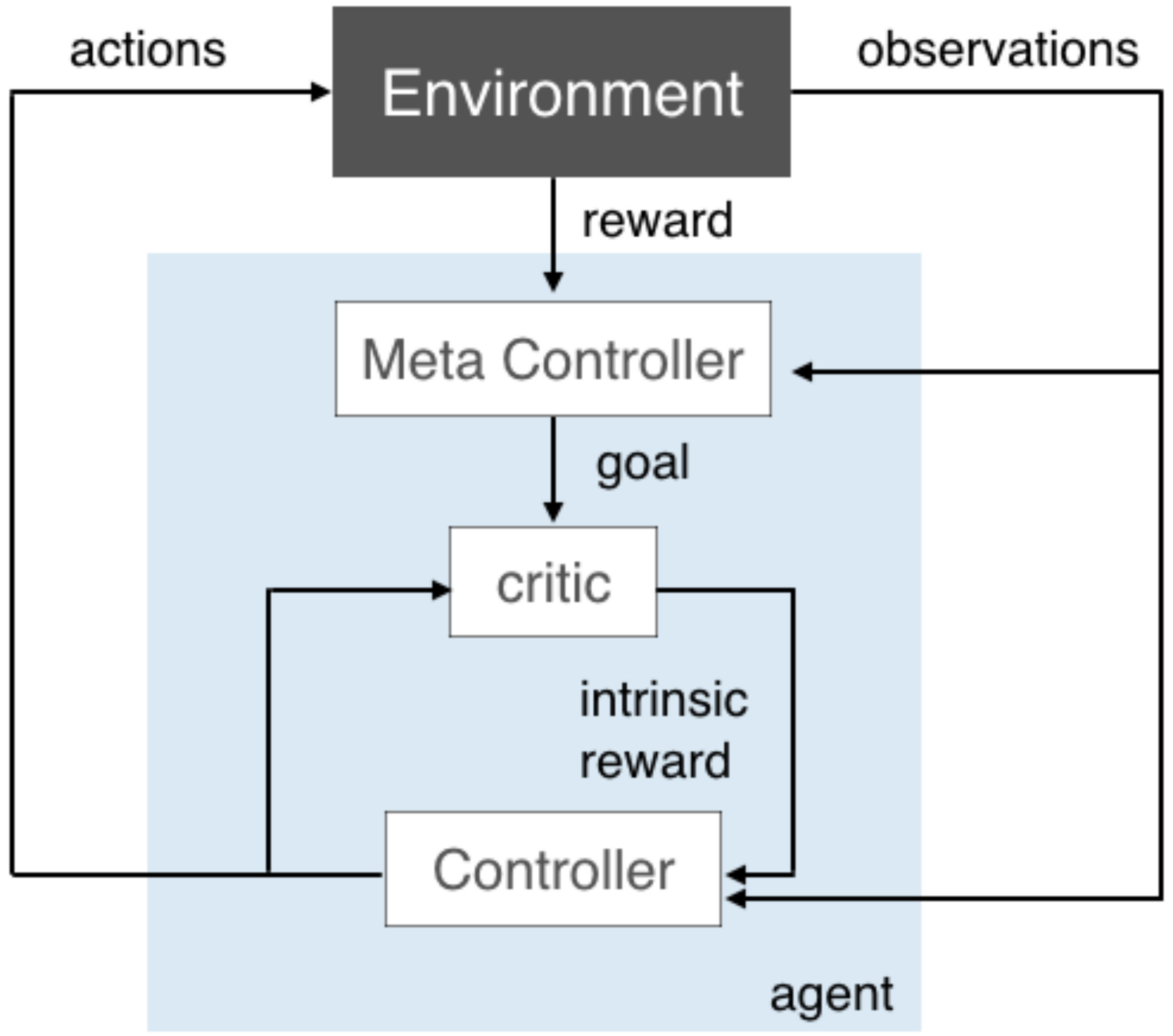
Semi Markov Decision Processes



Semi Markov Decision Processes

Metacontroller

$$Q_2^*(s,g) = \max_{\pi_g} \mathrm{E}[\sum_{t'=t}^{t+N} f_{t'} + \gamma \max_{g'} Q_2^*(s_{t+N}, g') \mid s_t = s, g_t = g, \pi_g]$$



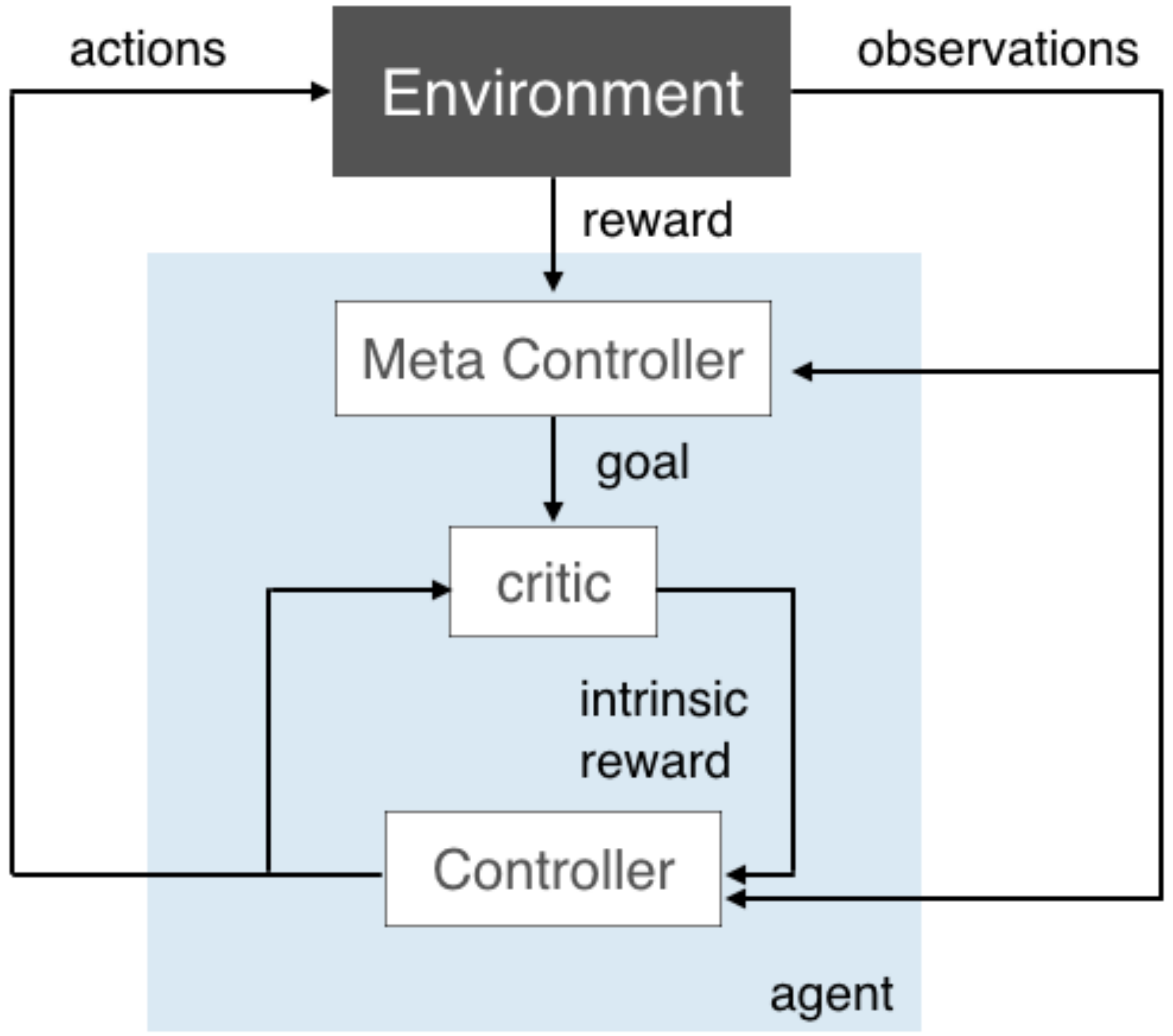
Semi Markov Decision Processes

Metacontroller

$$Q_2^*(s,g) = \max_{\pi_g} \mathrm{E}[\sum_{t'=t}^{t+N} f_{t'} + \gamma \max_{g'} Q_2^*(s_{t+N}, g') \mid s_t = s, g_t = g, \pi_g]$$

Controller

$$Q_1^*(s,a;g) = \max_{\pi_{ag}} \mathrm{E}[\sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'} \mid s_t = s, a_t = a, g_t = g, \pi_{ag}]$$
$$= \max_{\pi_{ag}} \mathrm{E}[r_t + \gamma \max_{a_{t+1}} Q_1^*(s_{t+1}, a_{t+1}; g) \mid s_t = s, a_t = a, g_t = g, \pi_{ag}]$$



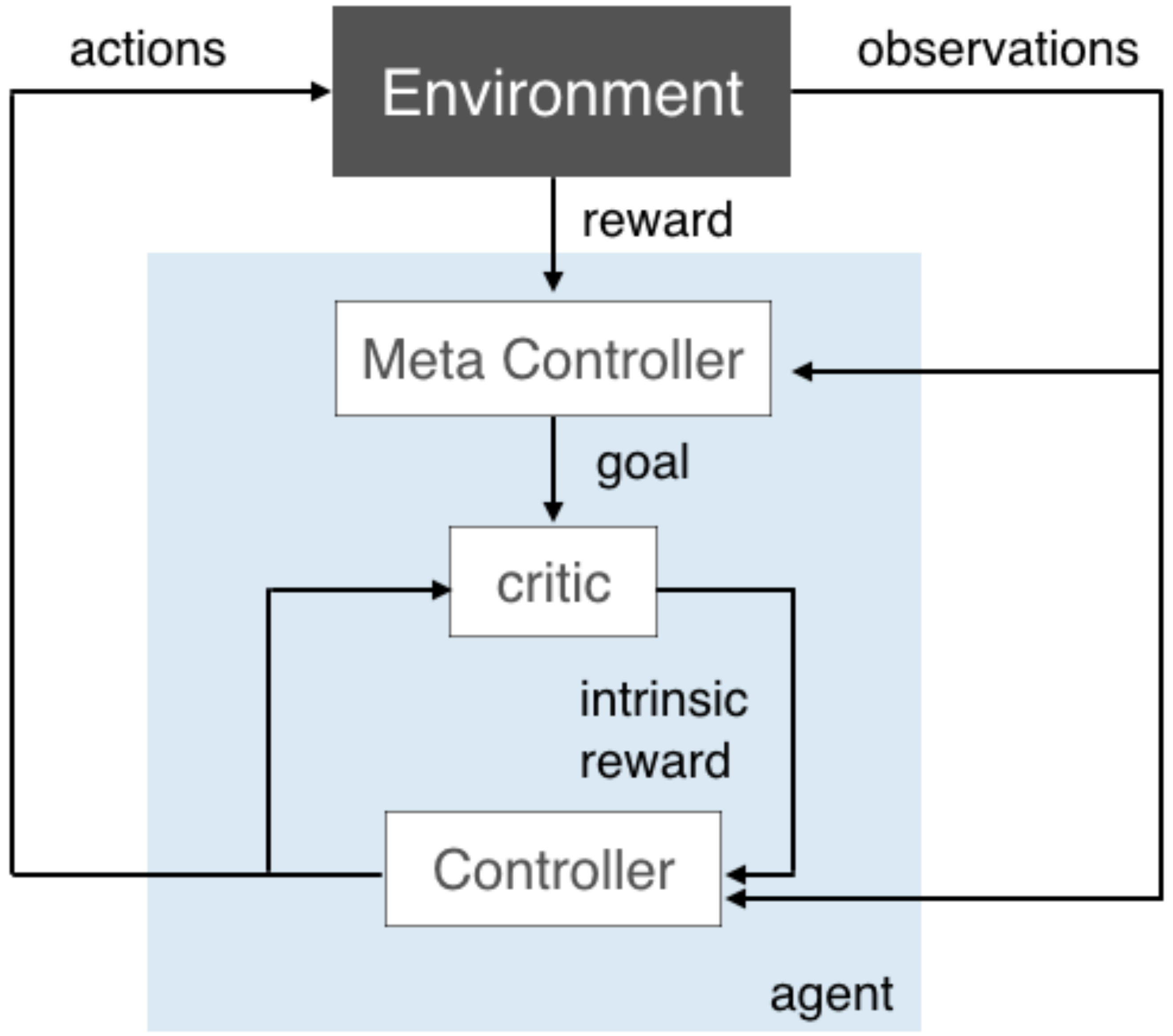
Semi Markov Decision Processes

Metacontroller

$$Q_2^*(s,g) = \max_{\pi_g} \mathrm{E}[\sum_{t'=t}^{t+N} f_{t'} + \gamma \max_{g'} Q_2^*(s_{t+N}, g') \mid s_t = s, g_t = g, \pi_g]$$

Controller

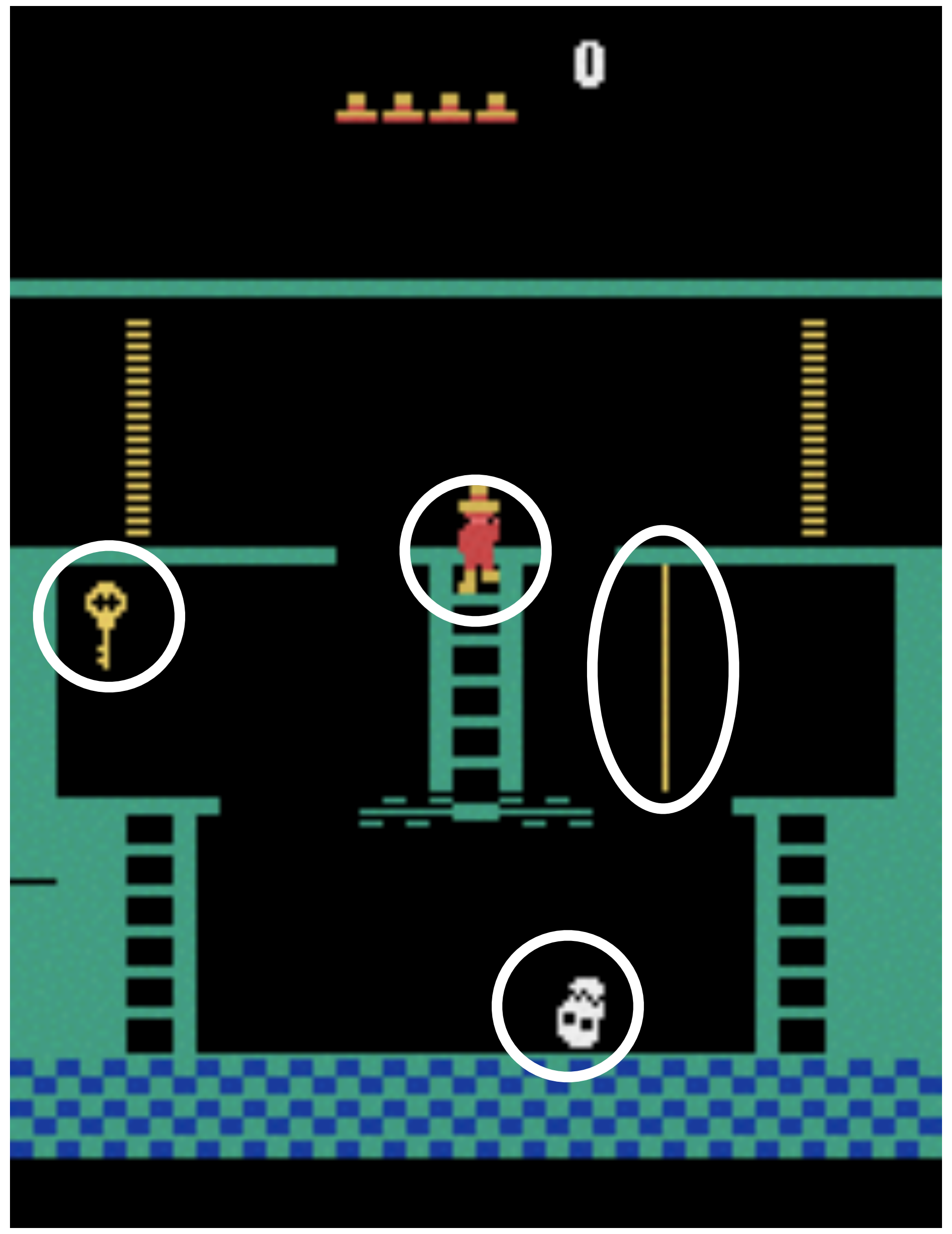
$$\begin{aligned} Q_1^*(s,a;g) &= \max_{\pi_{ag}} \mathrm{E}[\sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'} \mid s_t = s, a_t = a, g_t = g, \pi_{ag}] \\ &= \max_{\pi_{ag}} \mathrm{E}[r_t + \gamma \max_{a_{t+1}} Q_1^*(s_{t+1}, a_{t+1}; g) \mid s_t = s, a_t = a, g_t = g, \pi_{ag}] \end{aligned}$$



Solve for Q1 and Q2 using separate Deep Q-Networks and replay memories and using **stochastic gradient descent** at **different time scales**

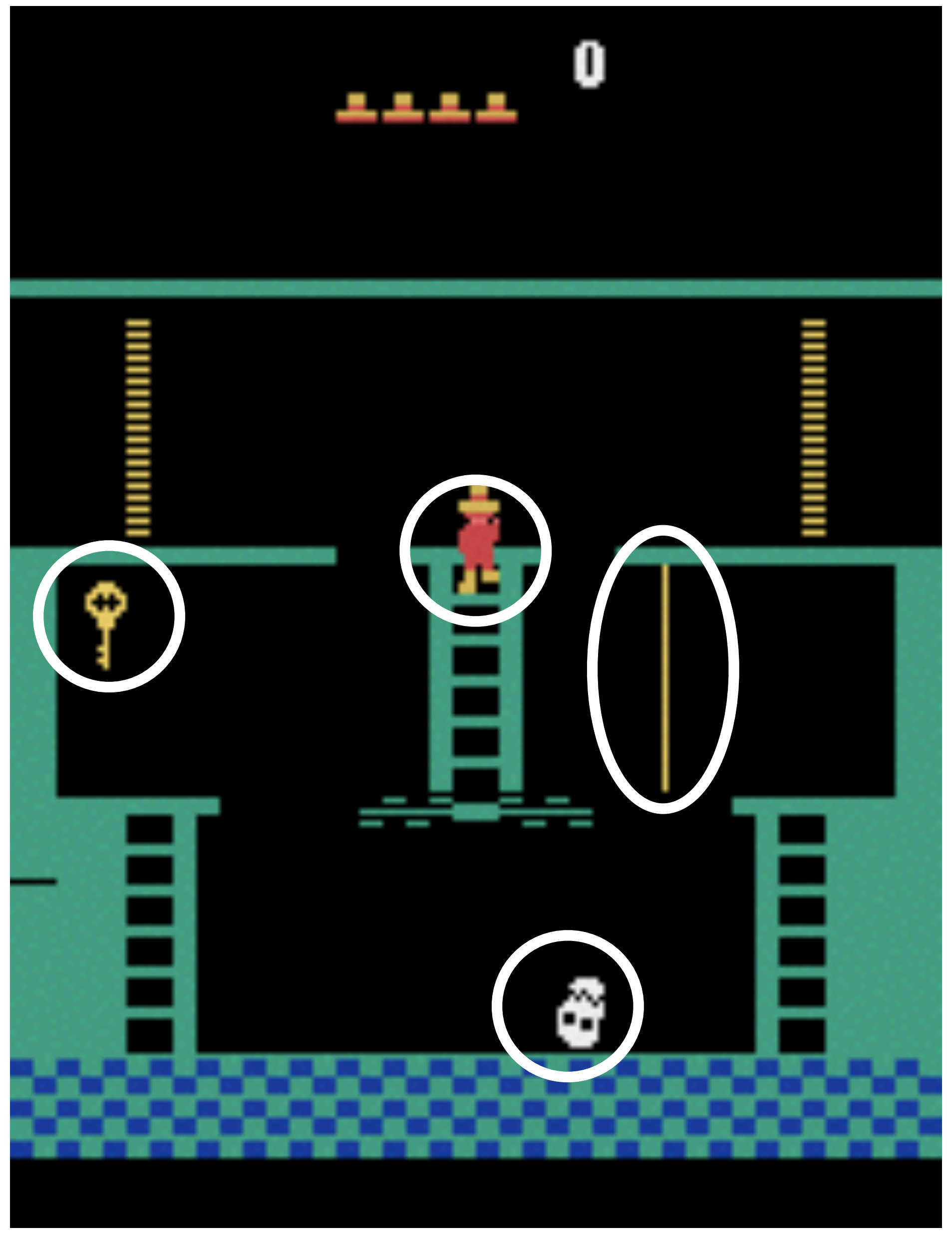
Path for scaling Deep HRL

Path for scaling Deep HRL

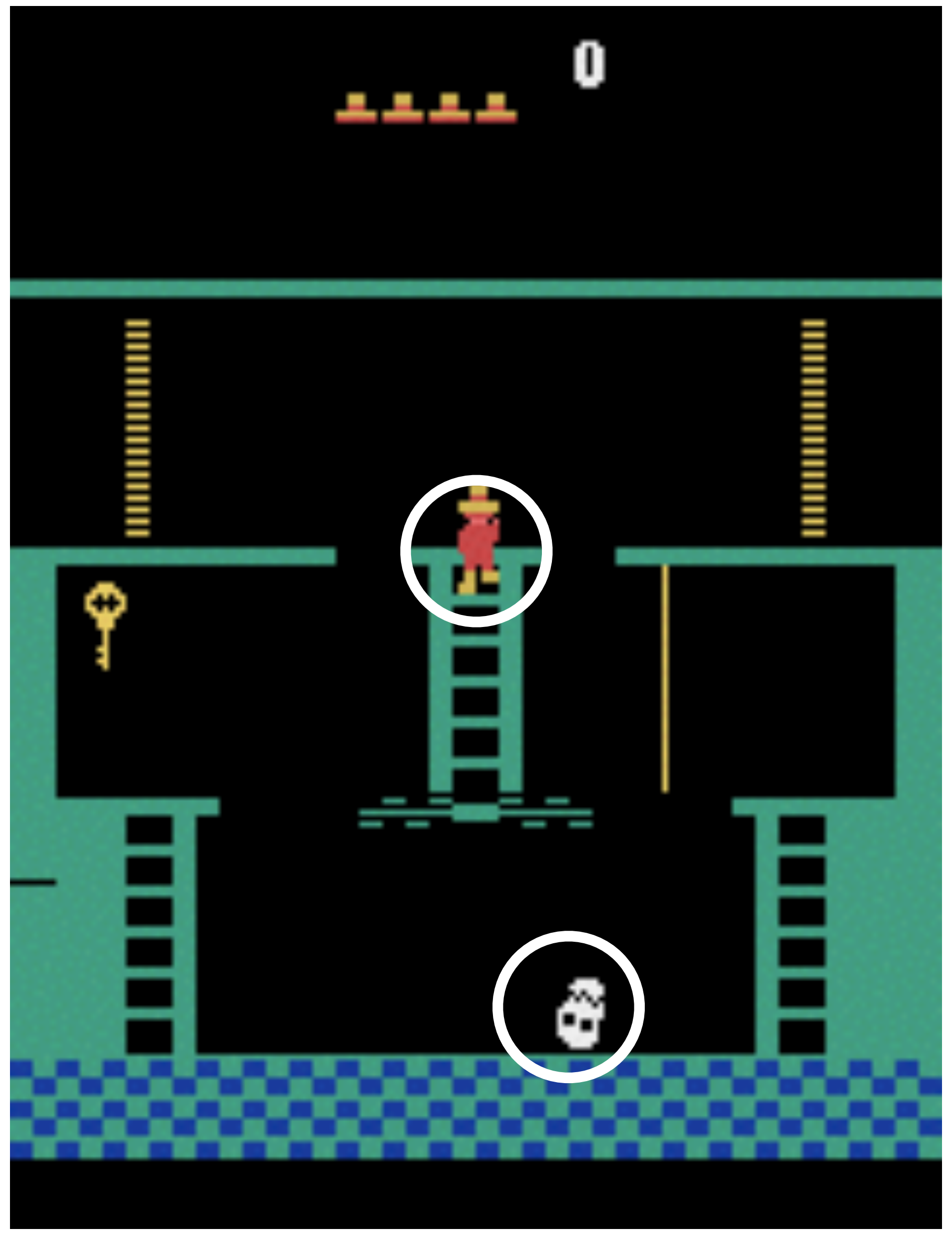


Features + Concepts

Path for scaling Deep HRL

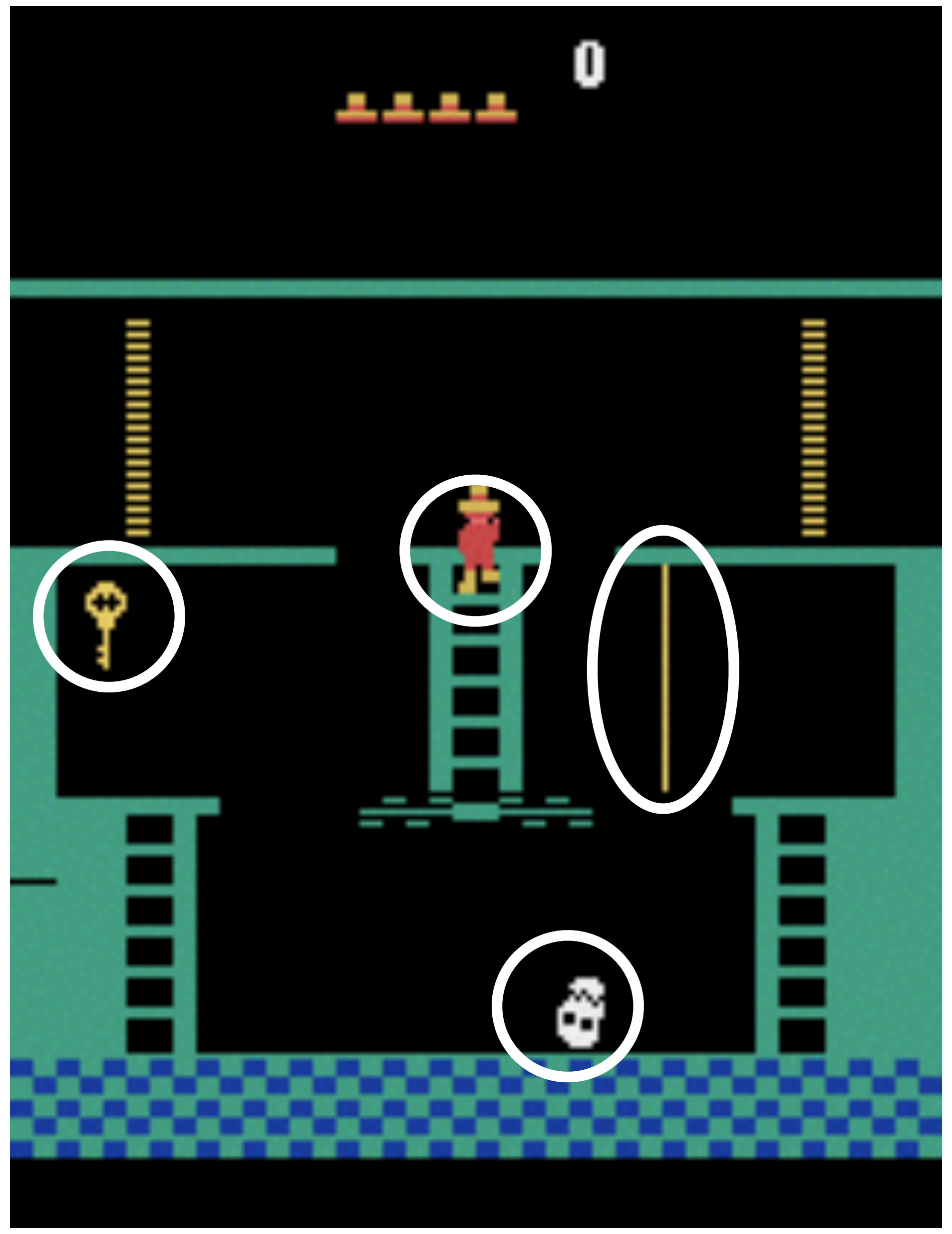


Features + Concepts

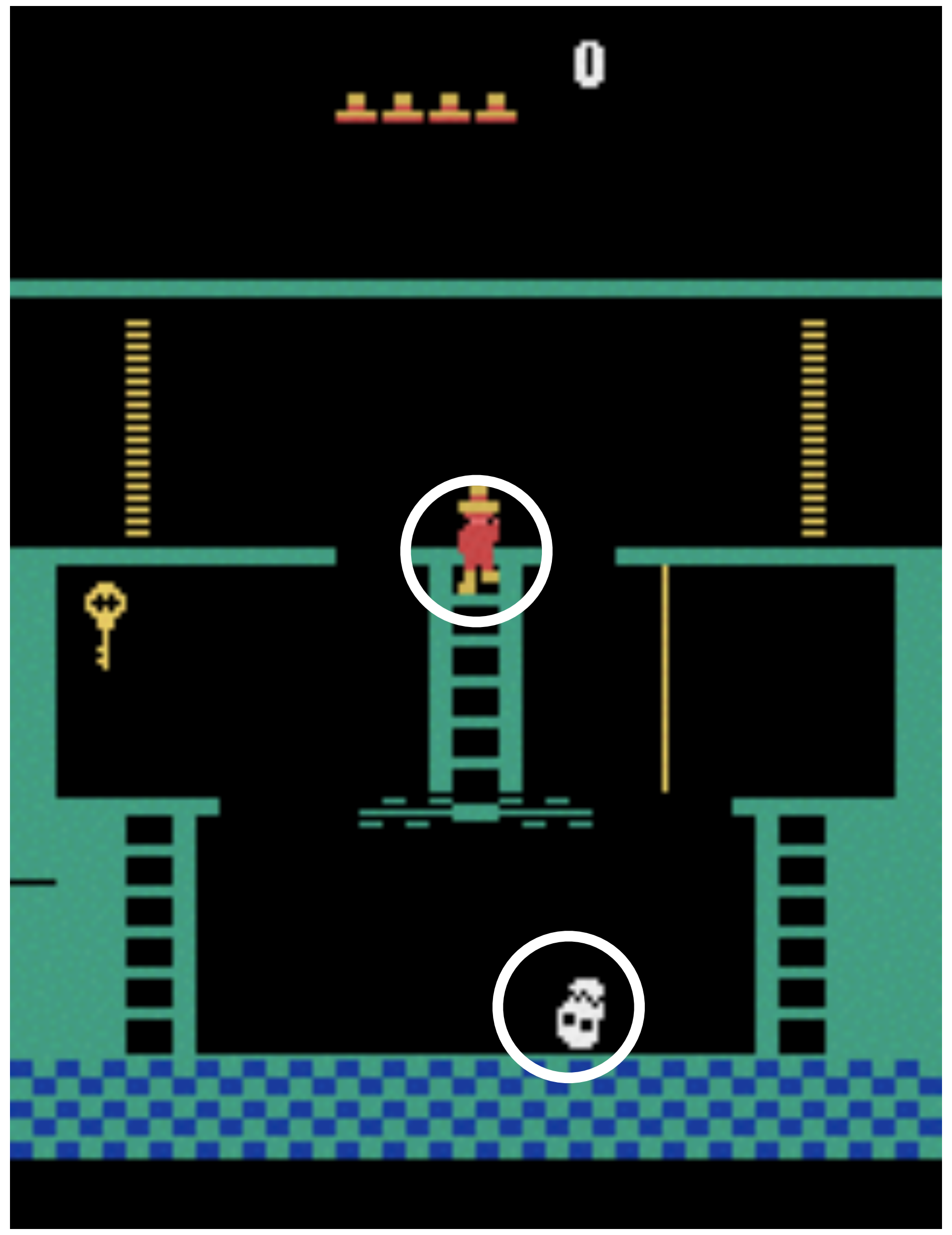


Level of agency.
Important for estimating
unlearnability

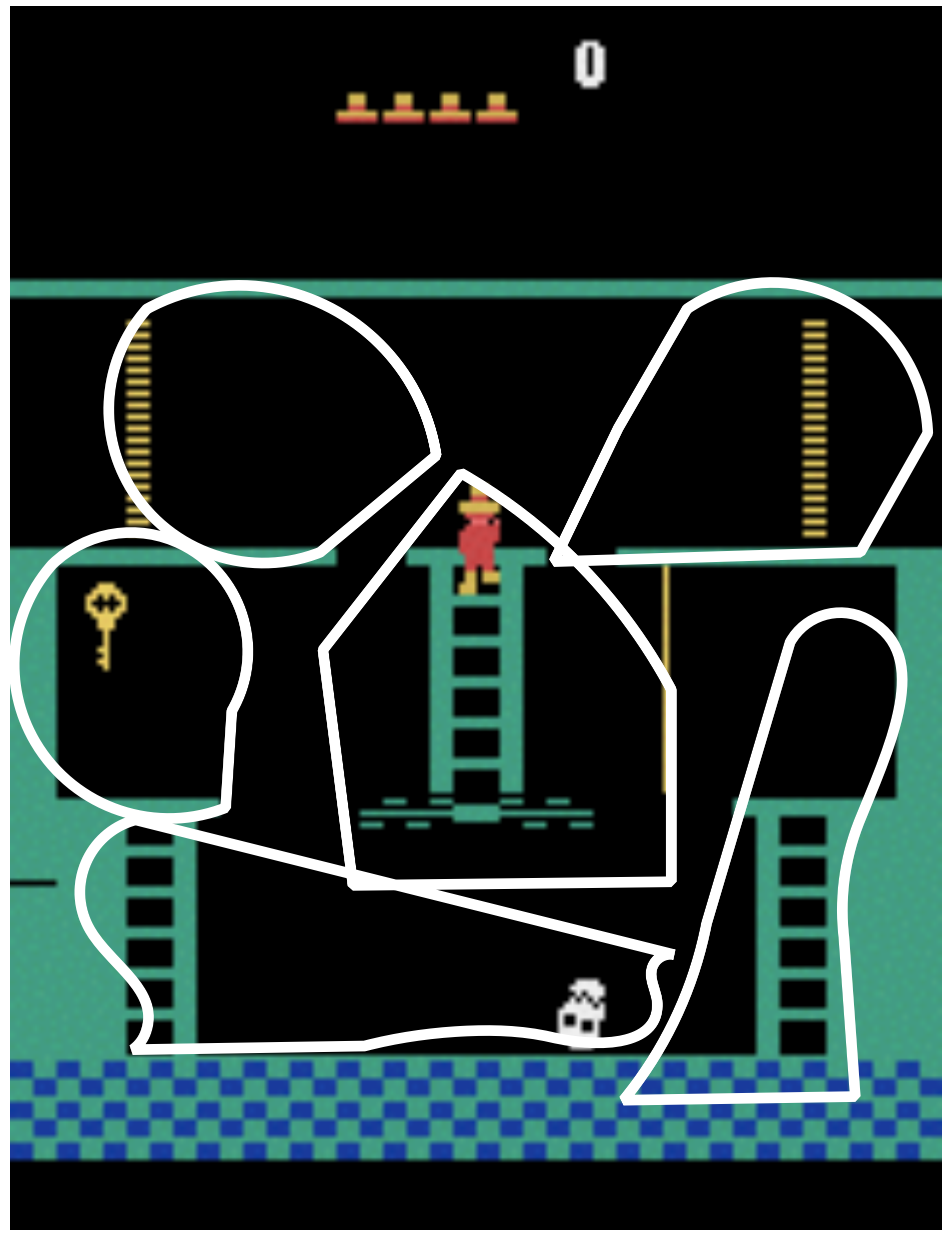
Path for scaling Deep HRL



Features + Concepts

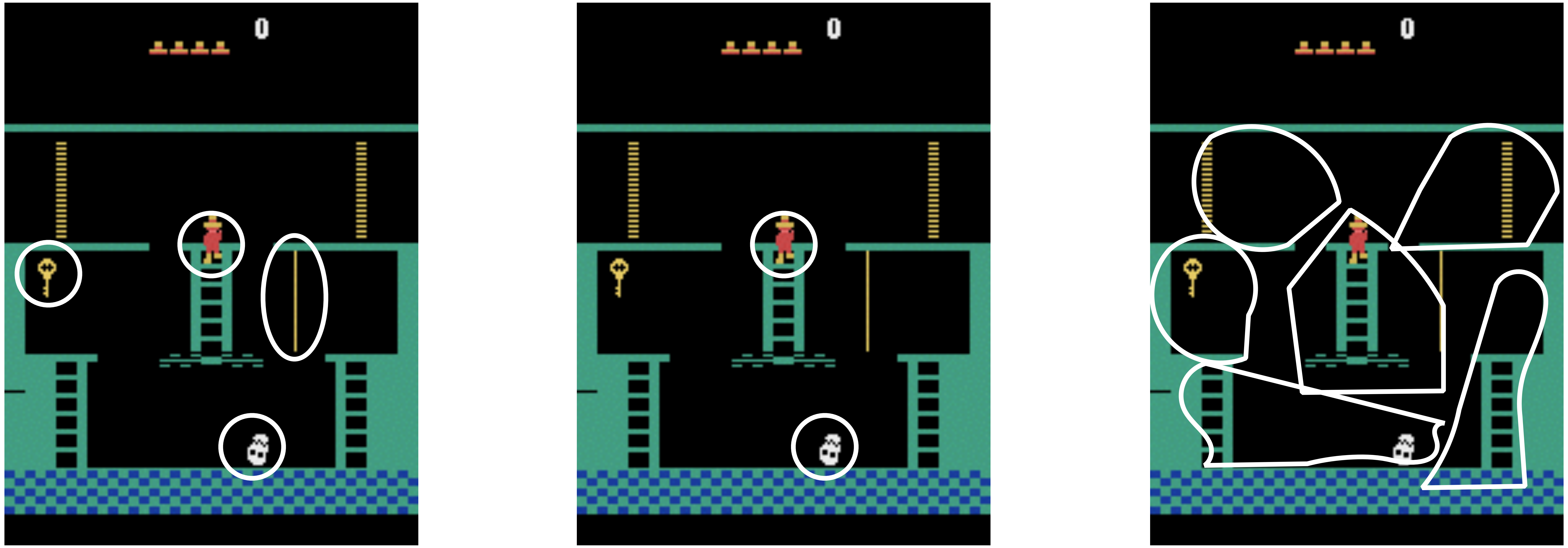


Level of agency.
Important for estimating *unlearnability*



Environment decomposition via exploration

Path for scaling Deep HRL



Features + Concepts

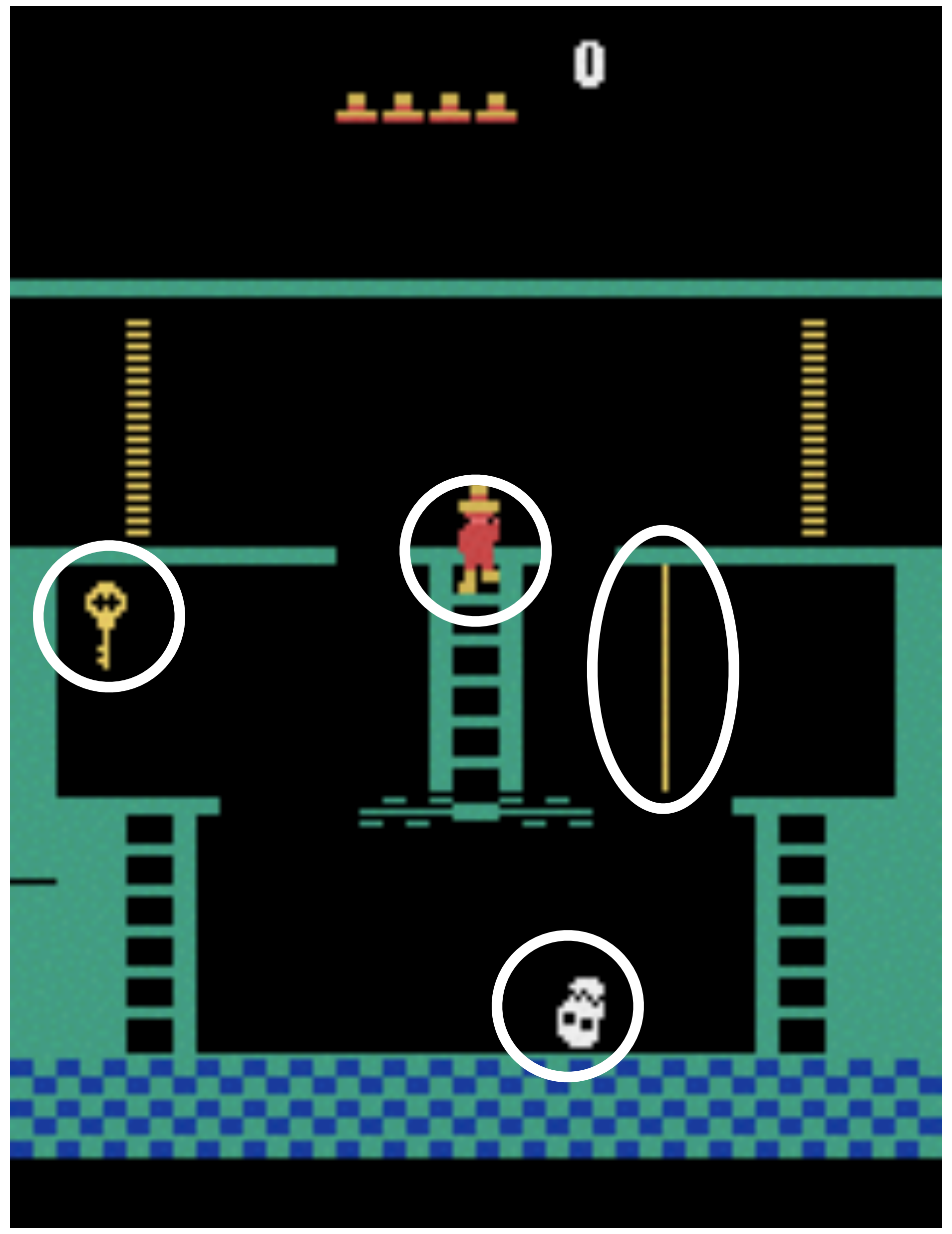
Level of agency.
Important for estimating *unlearnability*

Environment decomposition via exploration

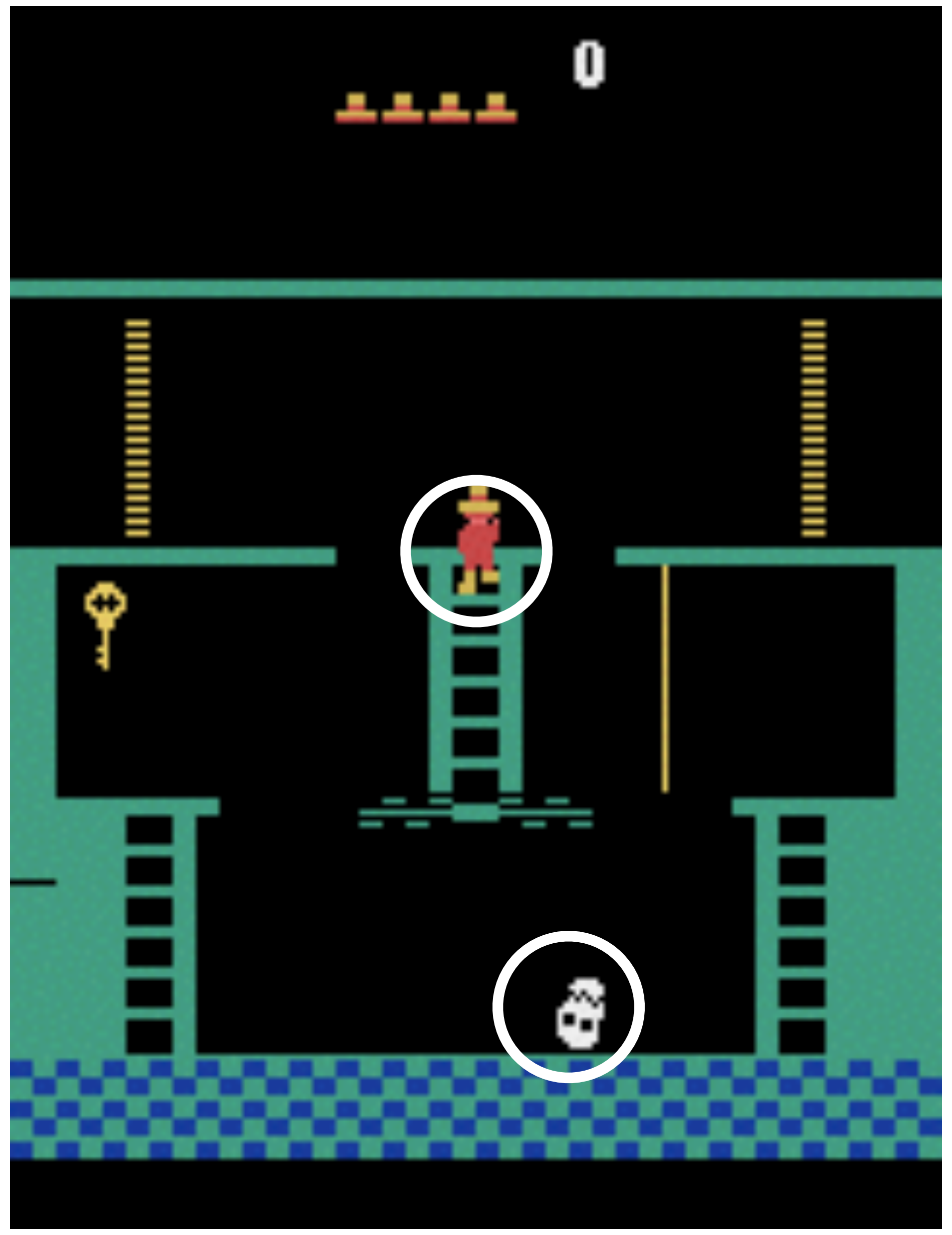
bootstrapping option discovery

Path for scaling Deep HRL

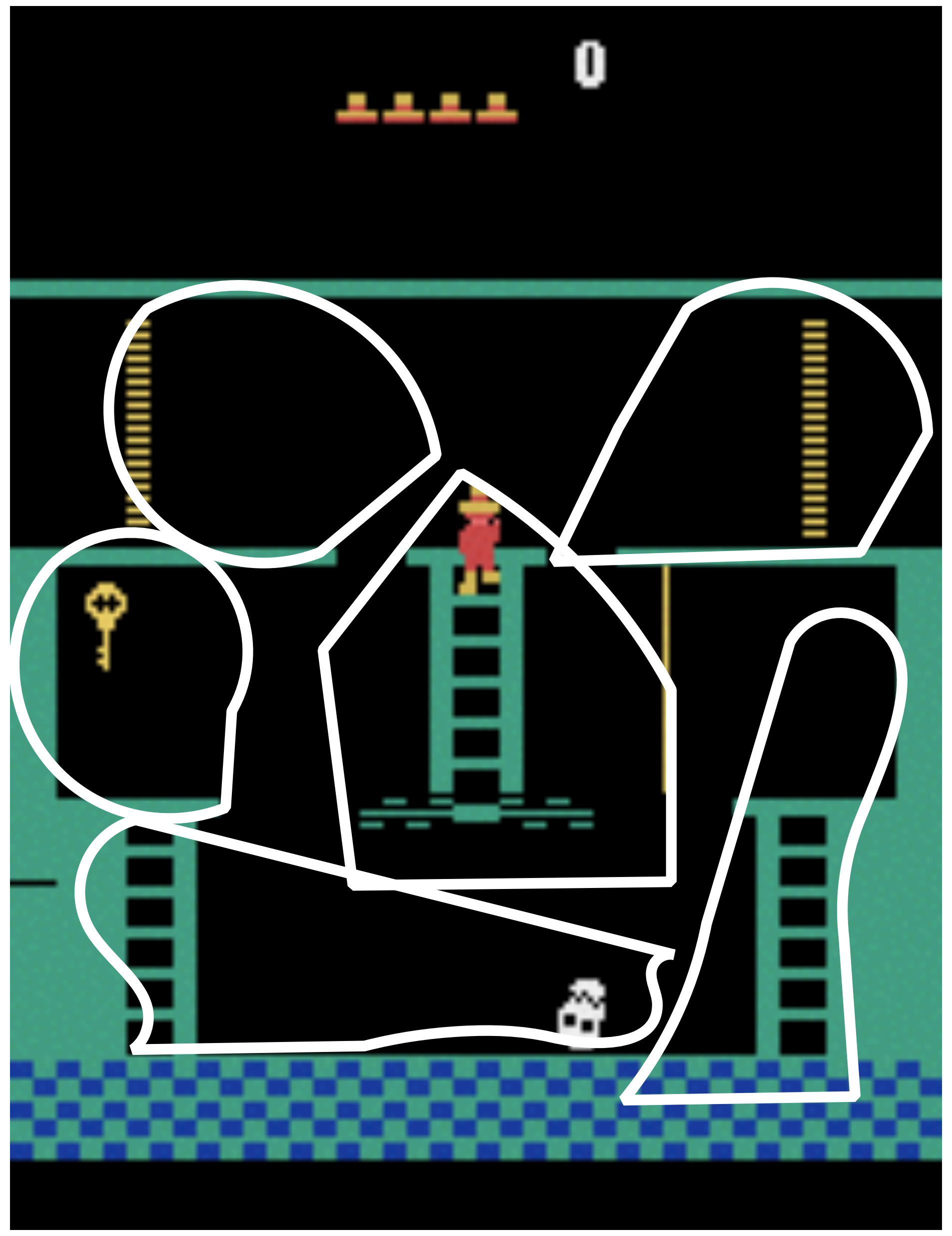
Knowledge-based intrinsically motivation exploration
(e.g. learning progress)



Features + Concepts



Level of agency.
Important for estimating *unlearnability*

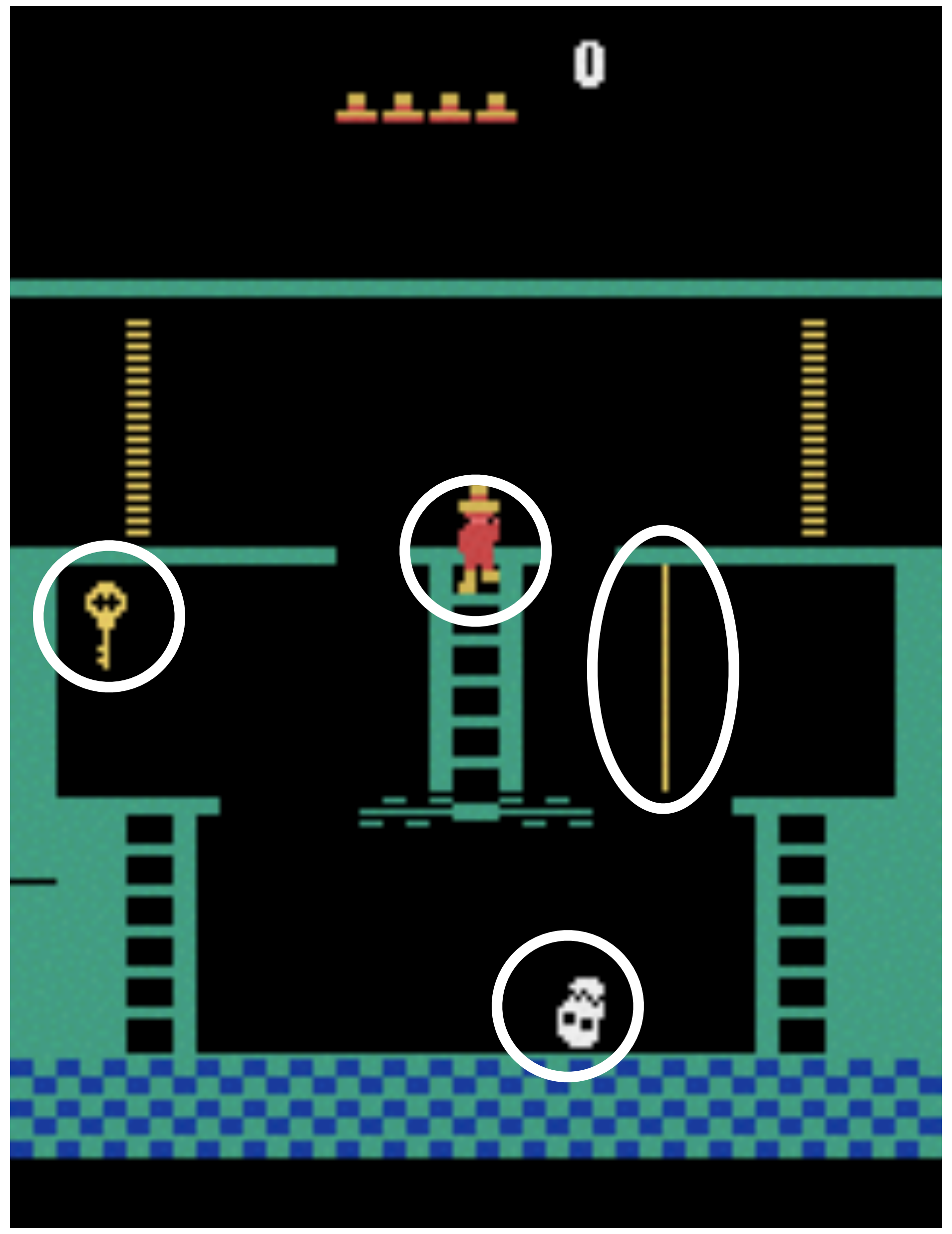


Environment decomposition via exploration

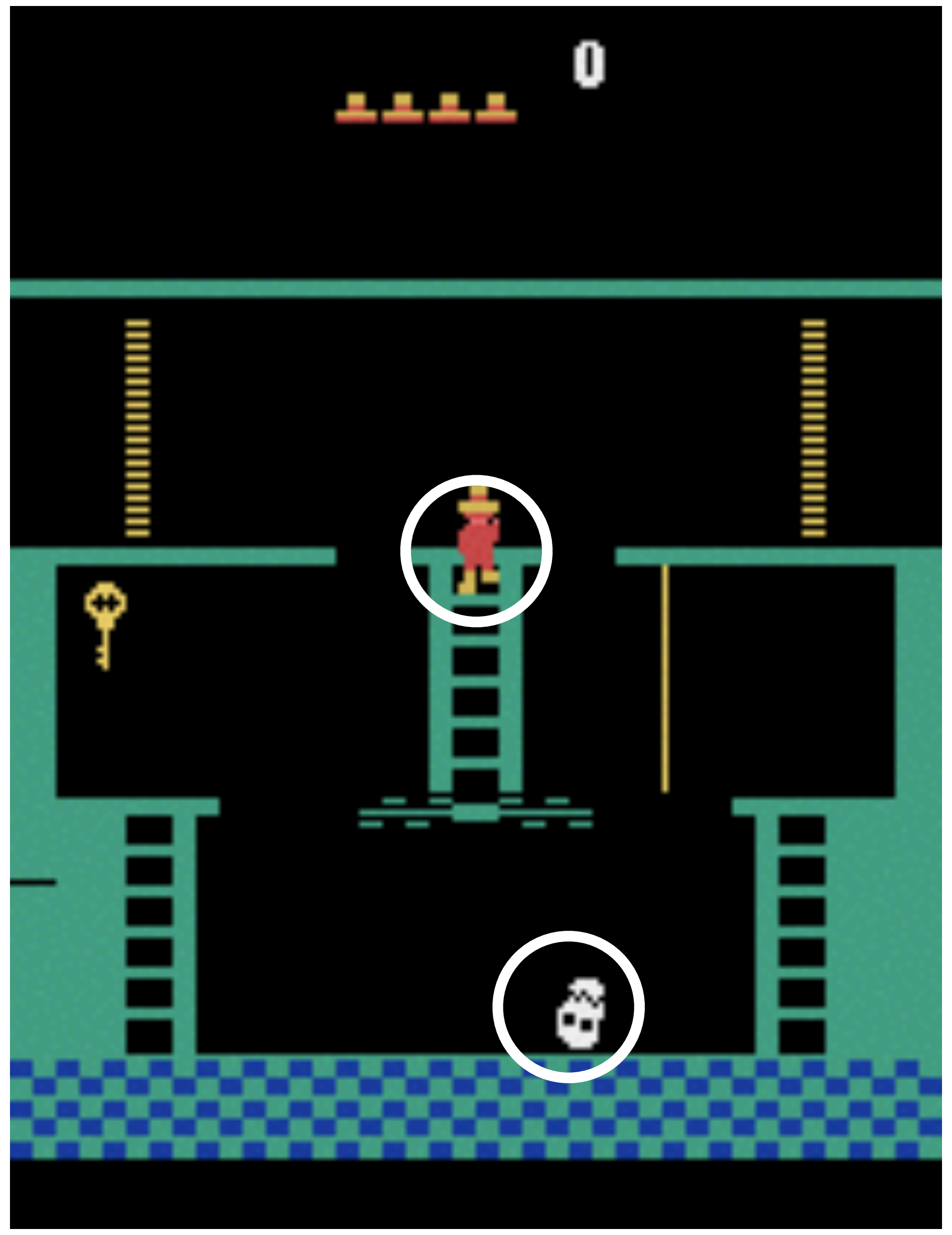
bootstrapping option discovery

Path for scaling Deep HRL

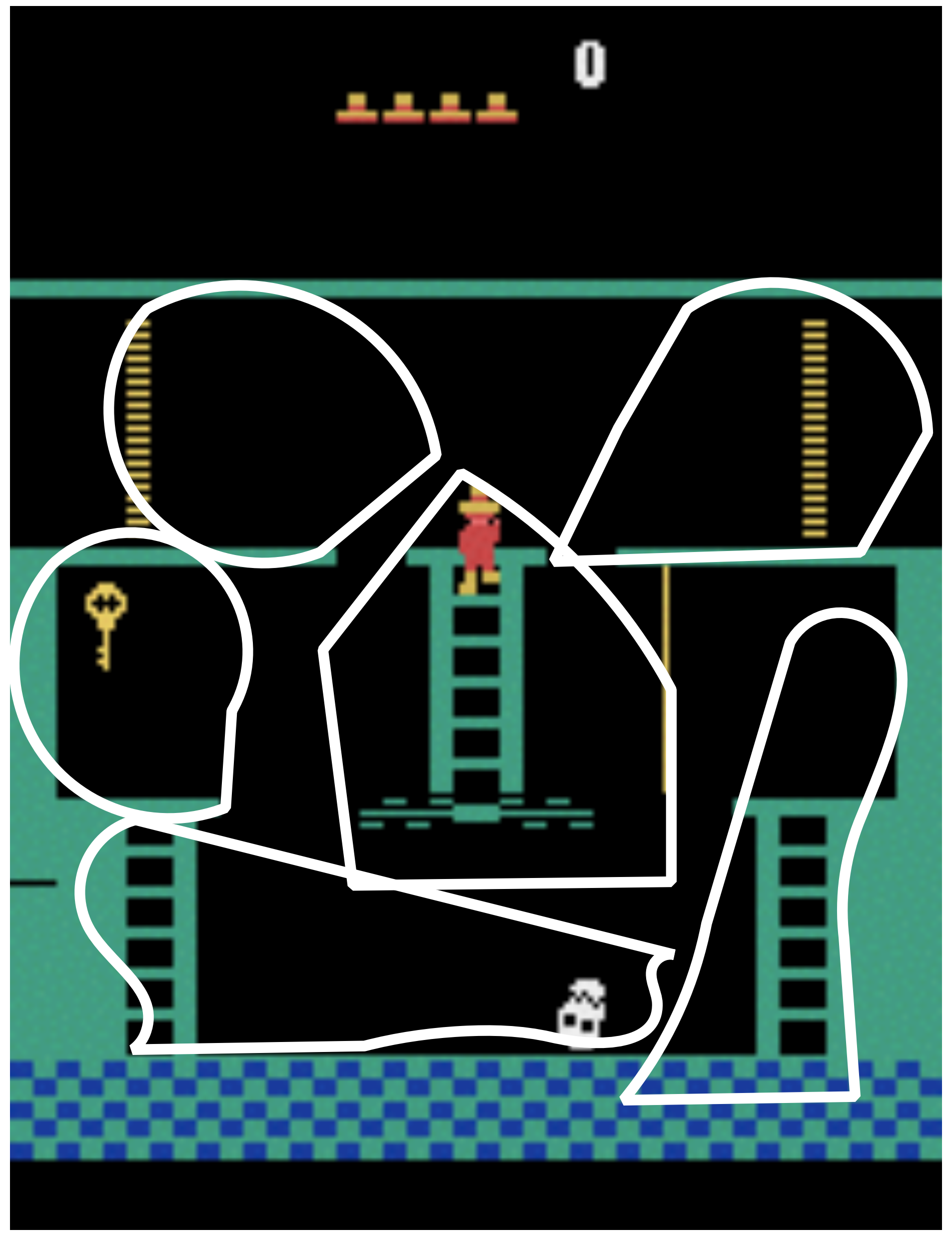
Knowledge-based intrinsically motivation exploration (e.g. learning progress) **+**



Features + Concepts



Level of agency. Important for estimating *unlearnability*



Environment decomposition via exploration

bootstrapping option discovery

Thank you