Deep Reinforcement Learning
Introduction and State-of-the-art

Arjun Chandra
Research Scientist
Telenor Research / Telenor-NTNU AI Lab
arjun.chandra@telenor.com
@boelger

24 October 2017

https://join.slack.com/t/deep-rl-tutorial/signup

The Plan

- Some history
- RL and Deep RL in a nutshell
- Deep RL Toolbox
- Challenges and State-of-the-art
  - Data Efficiency
  - Exploration
  - Temporal Abstractions
  - Generalisation

Robot Motor Skill Coordination with EM-based Reinforcement Learning

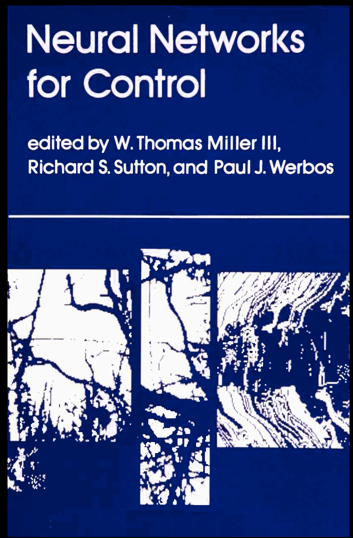
Petar Kormushev, Sylvain Calinon, and Darwin G. Caldwell

Italian Institute of Technology

https://vimeo.com/20042665

Brief History

Rich Sutton et al.



late 1980s

RL for robots using NNs, L-J Lin. **PhD 1993, CMU**

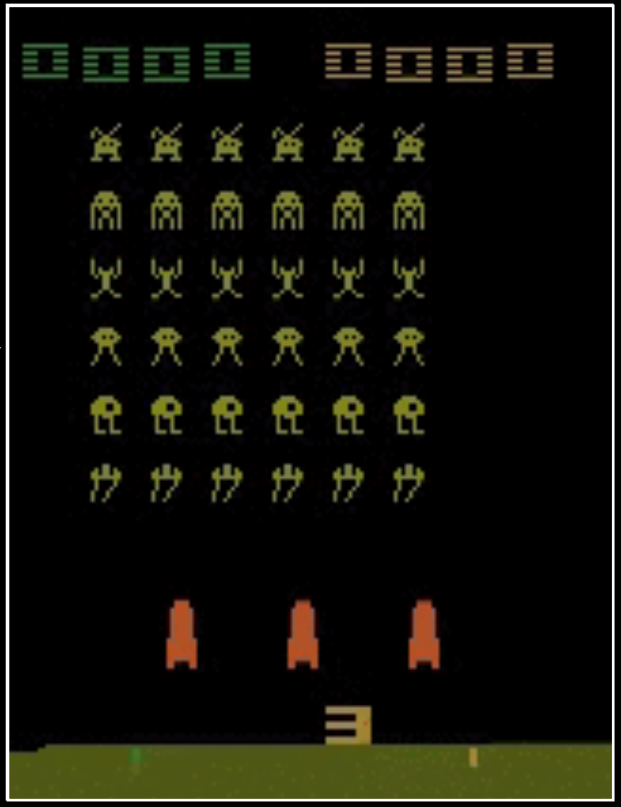
Gerald Tesauro

1995

Stanford

2004

Vlad Mnih et. al.



2013 —

Google DeepMind

David Silver et. al.

2015 —

Problem Characteristics

dynamic

uncertainty/volatility

uncharted/**unimagined**/
exception laden

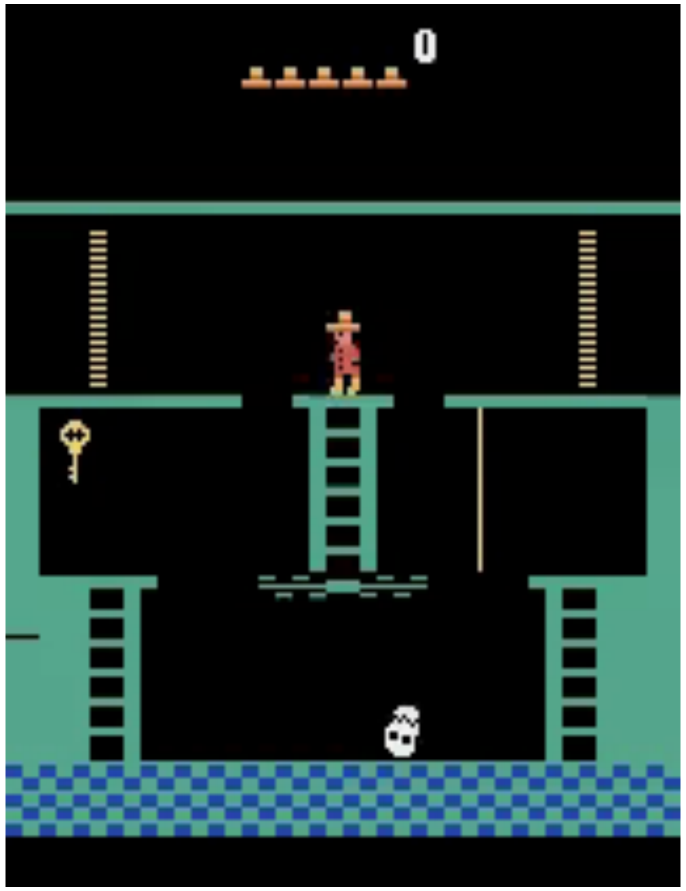
delayed consequences

requires **strategy**

Image credit: http://wonderfulengineering.com/inside-the-data-center-where-google-stores-all-its-data-pictures/

Solution

machine with **agency** which **learn**, **plan**, and **act** to find a strategy for solving the problem



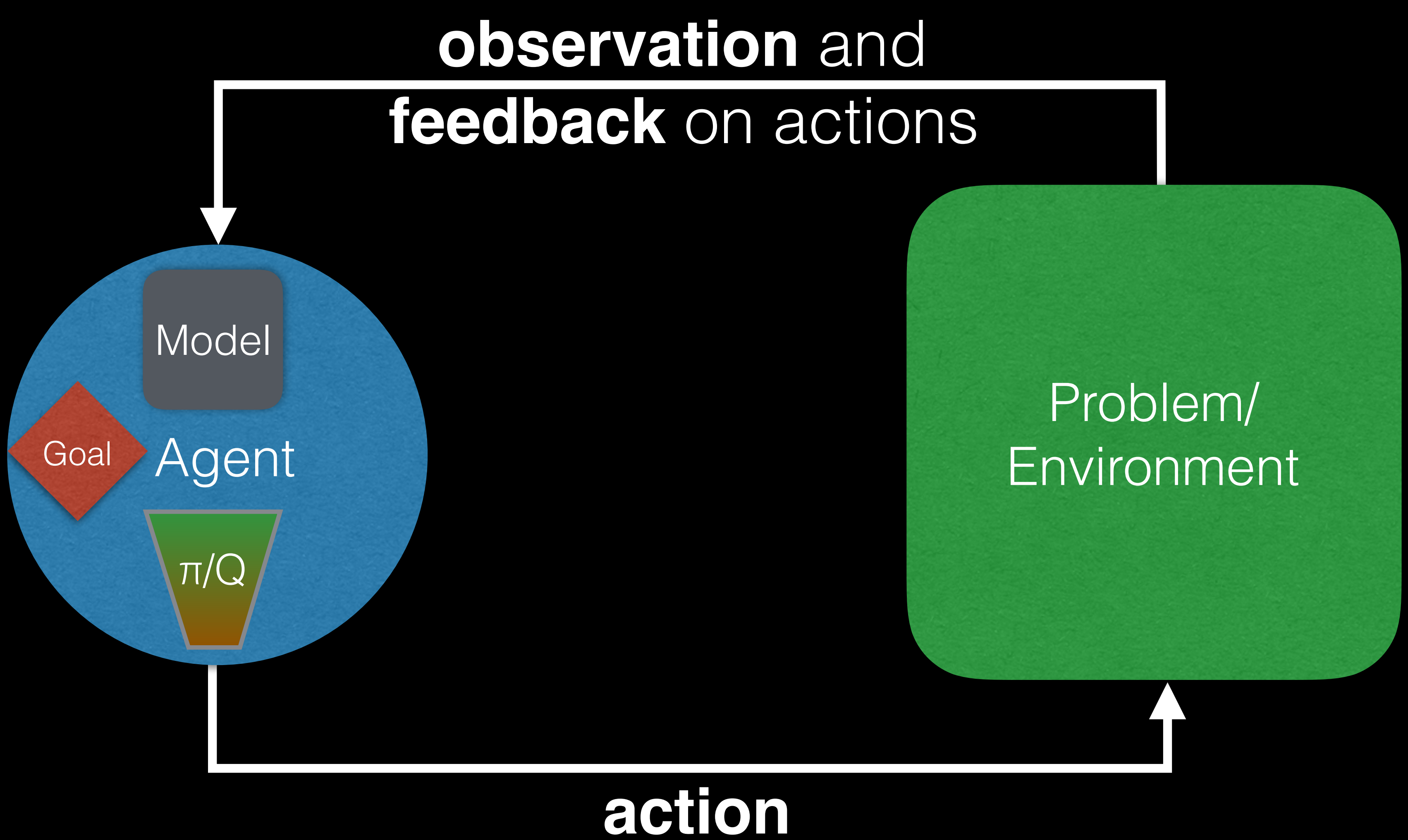
autonomous to some extent

probe and **learn from feedback**

focus on the **long-term objective**

explore and **exploit**

Reinforcement Learning

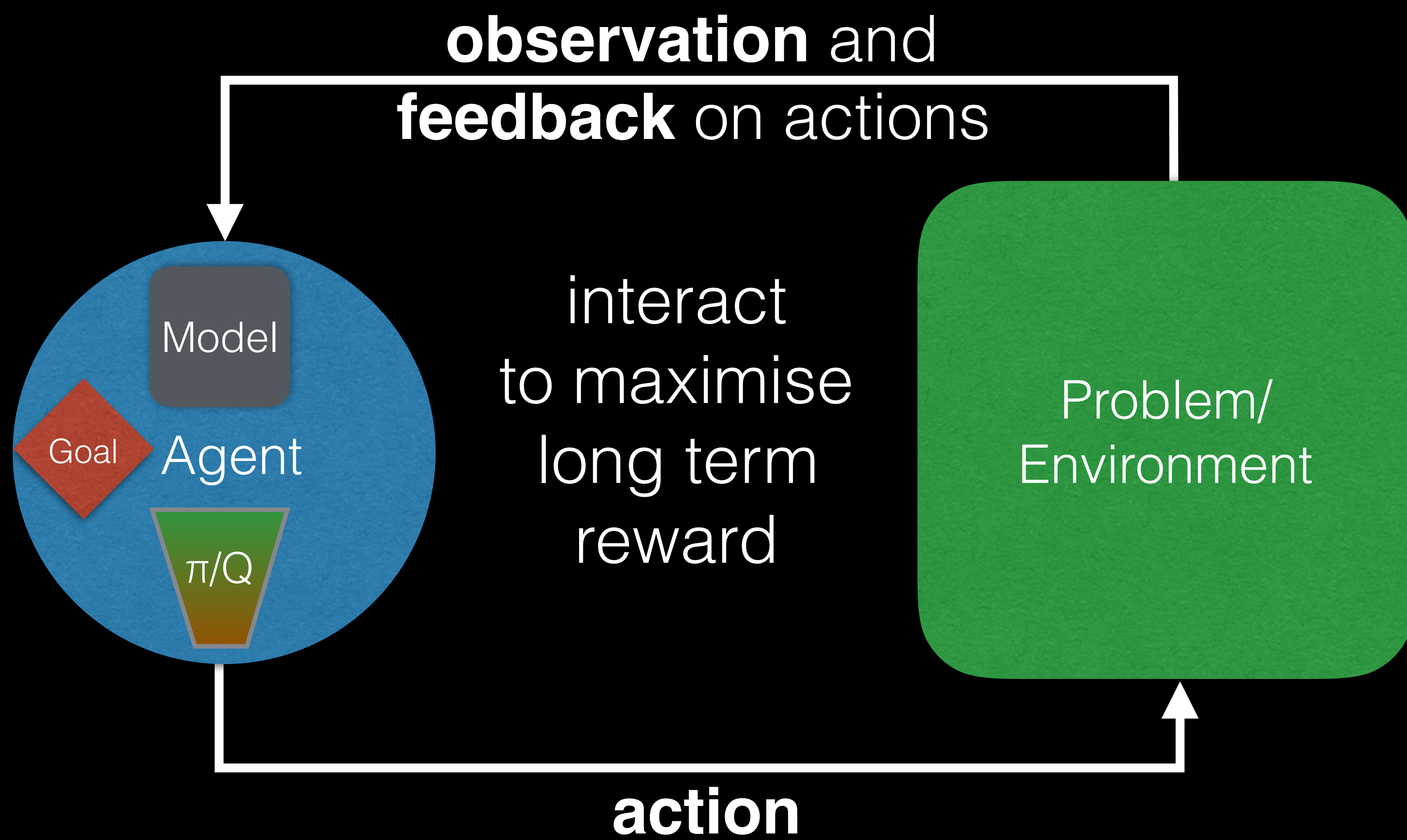


maximise return **E{R}**

dynamics model

policy/value function

The MDP game!

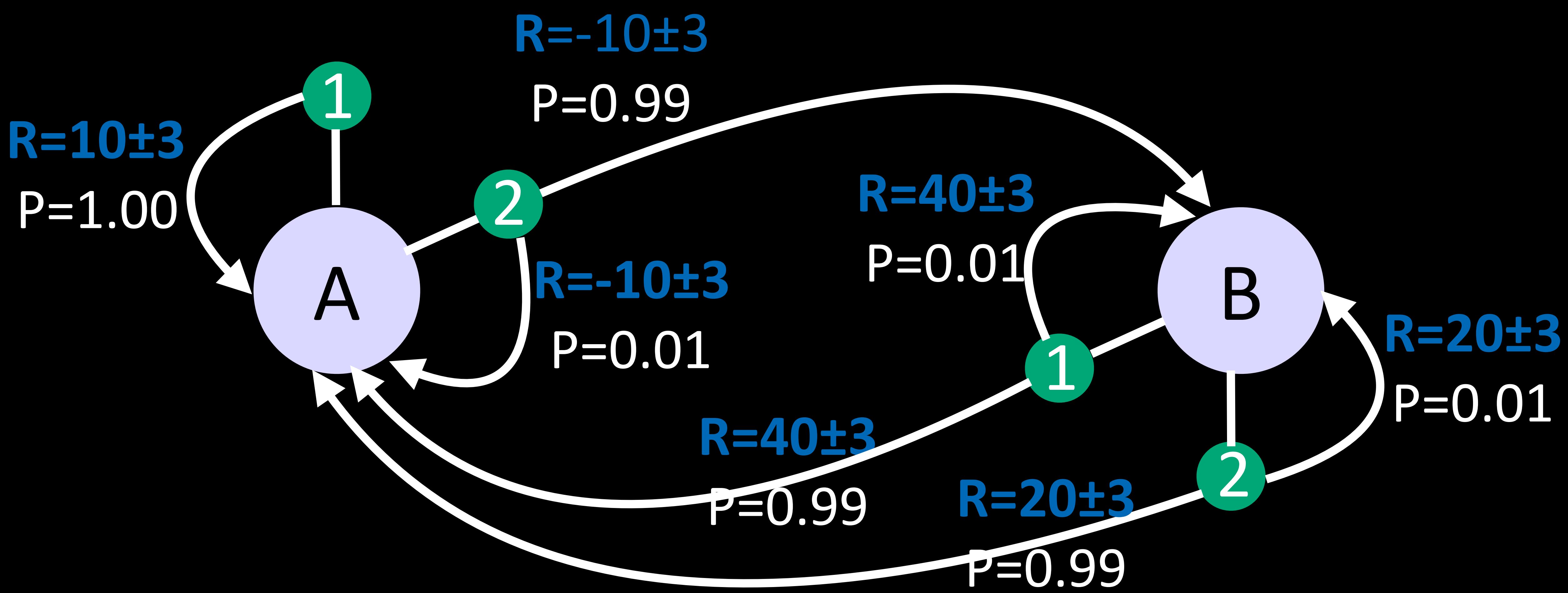


Goal maximise return **E{R}**

Inspired by Prof. Rich Sutton's tutorial:
https://www.youtube.com/watch?v=ggqnxyjaKe4

The MDP (S,A,P,R,ϒ)

R: immediate reward function R(s, a)
P: state transition probability P(s'|s, a)



https://github.com/traai/basic-rl

Terminology

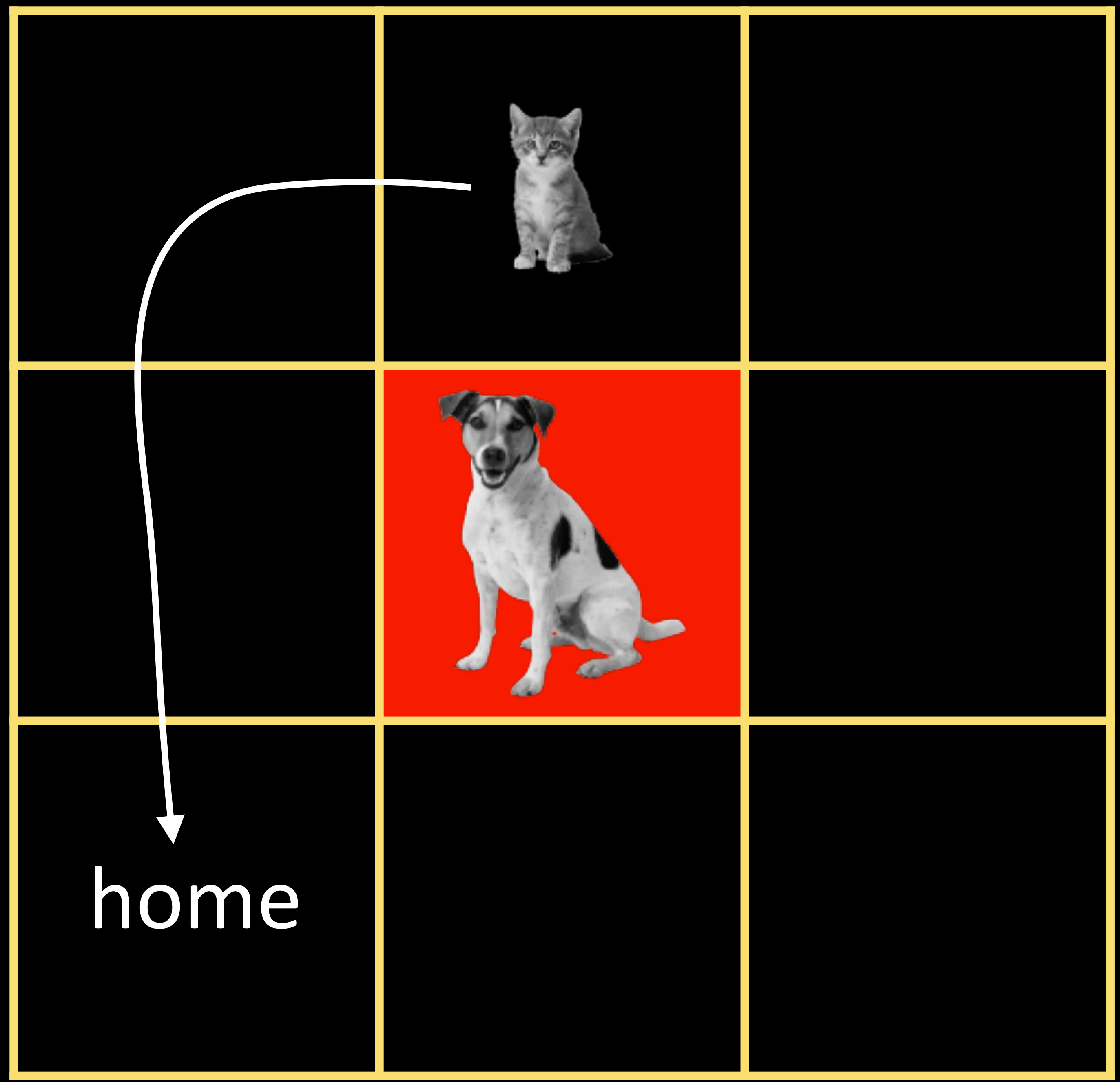
state or action
value function

policy

dynamics model

reward

goal



Terminology

state or action value function

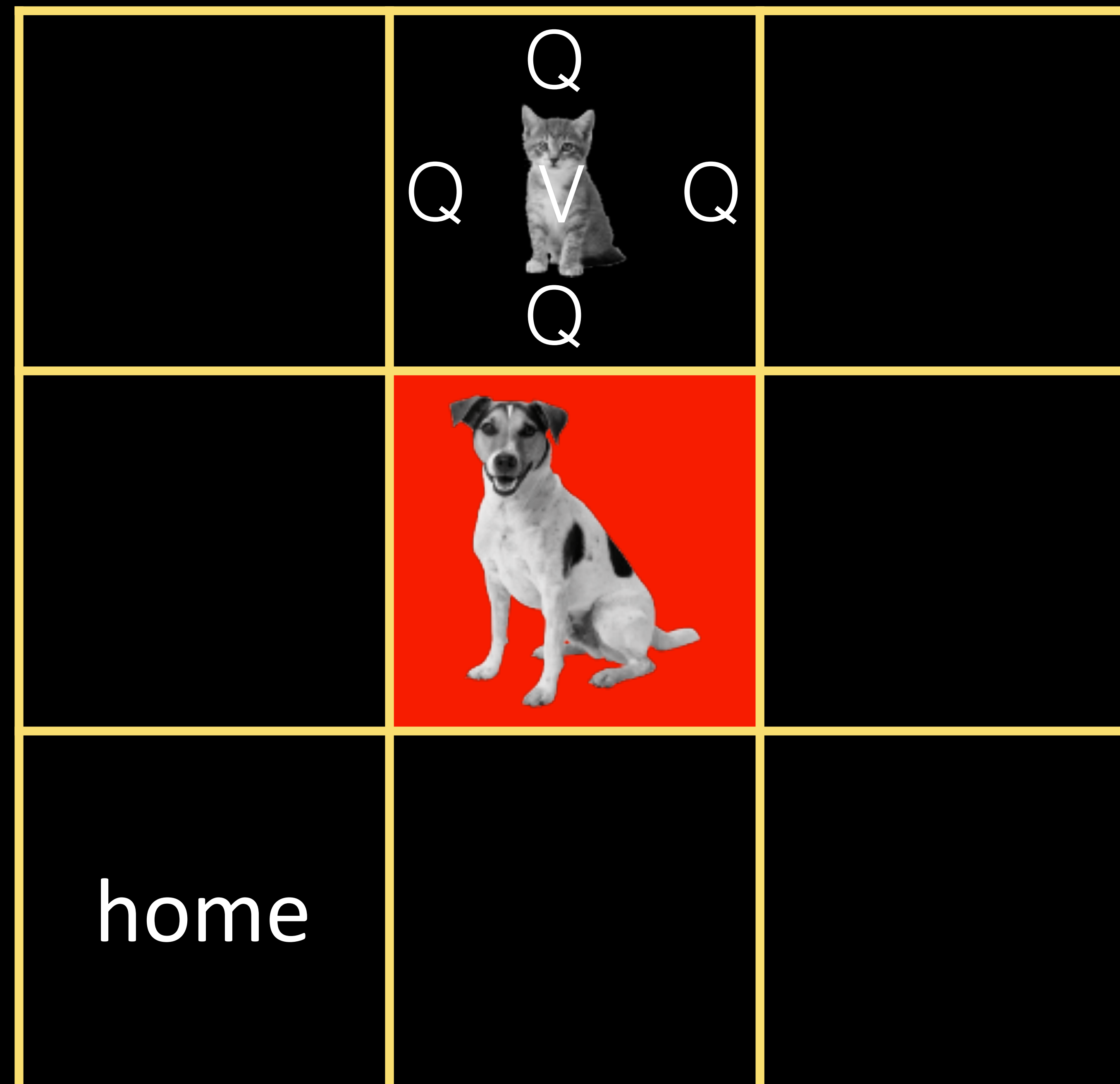
Q(s,a) V(s)

policy

dynamics model

reward

goal



Terminology

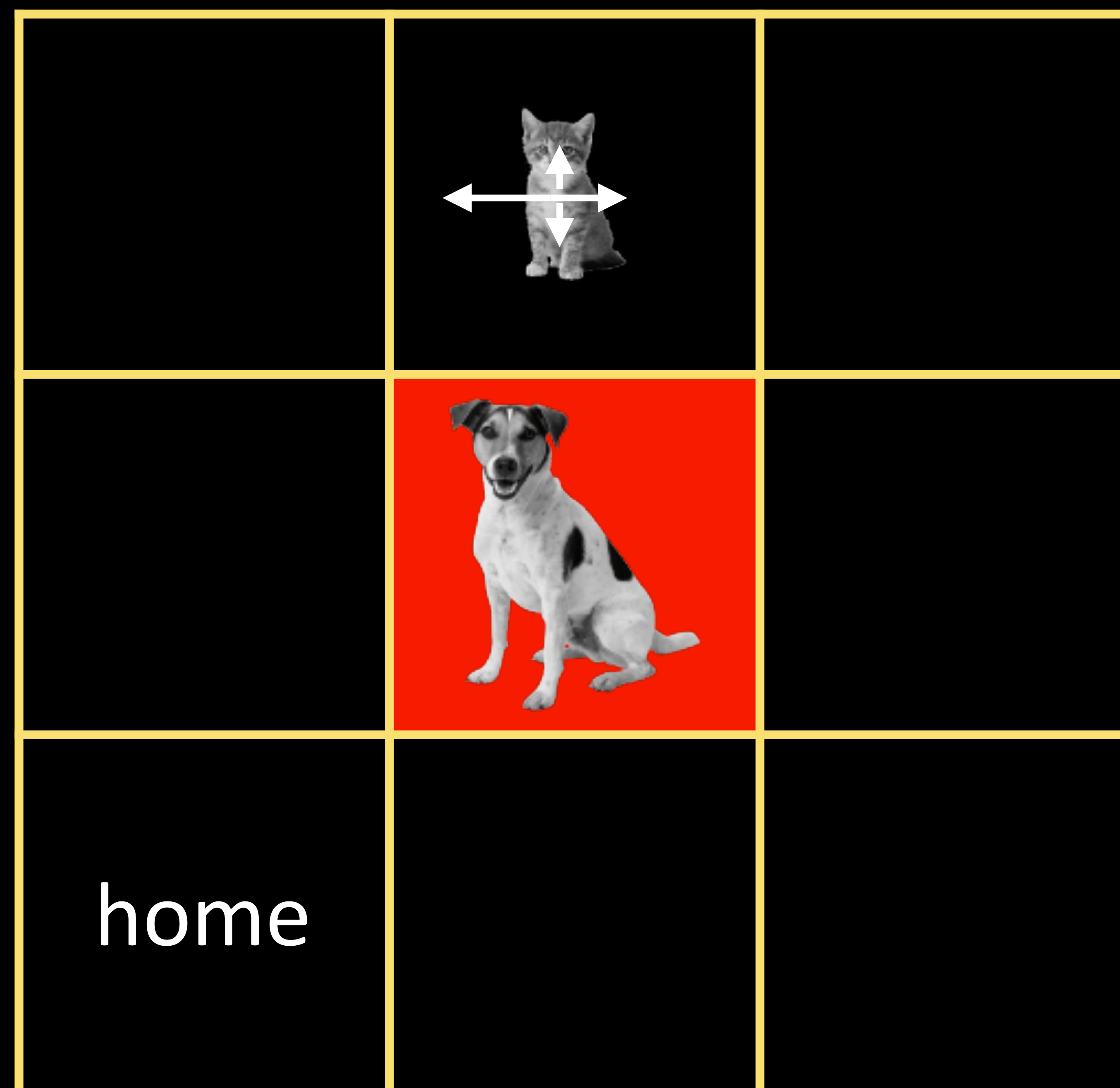
state or action
value function

policy $\pi(s|a)$
$\pi(s)$

dynamics model

reward

goal



Terminology

state or action
value function

policy

dynamics model

reward

goal

Terminology

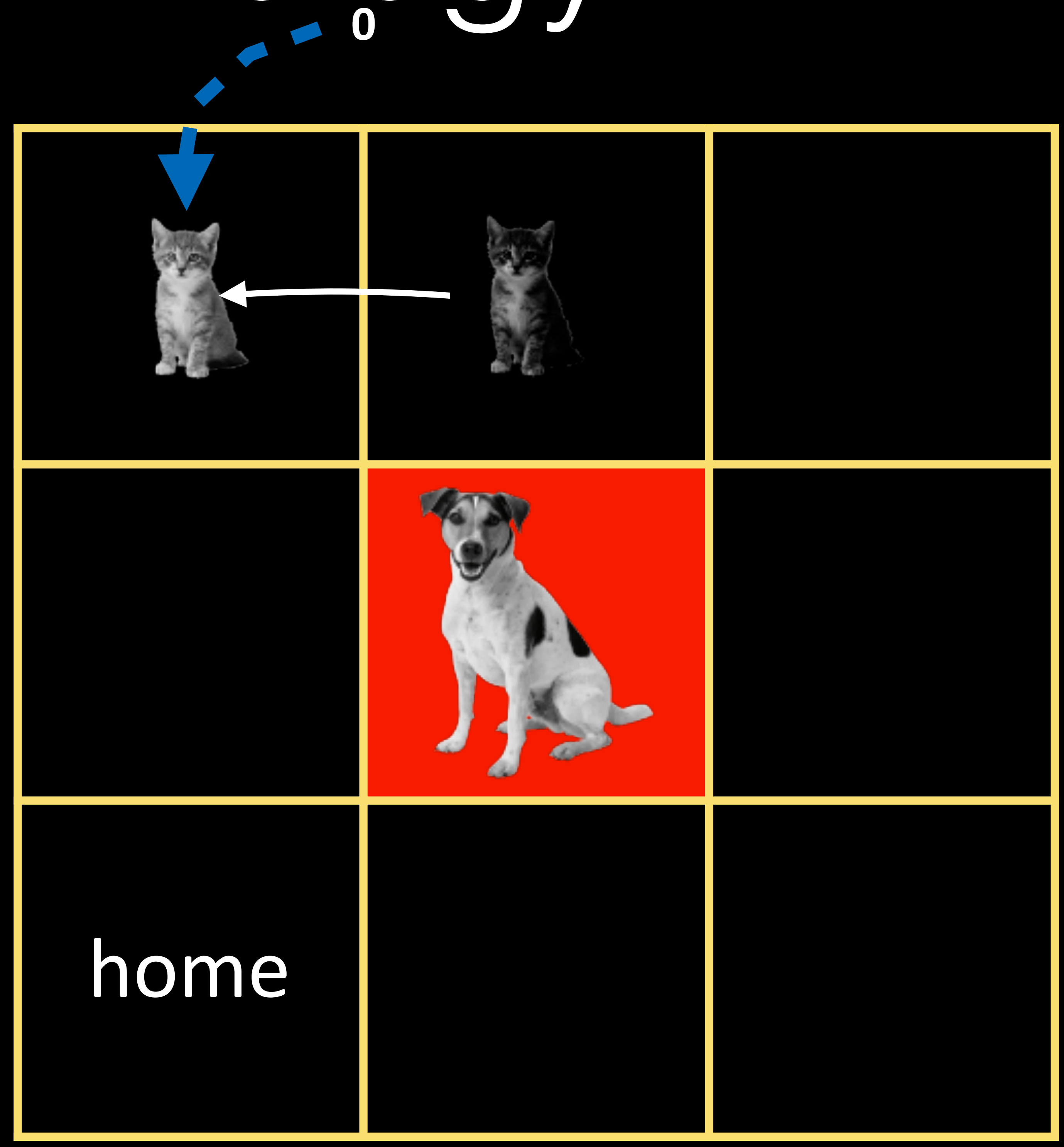
state or action
value function

policy

dynamics model

reward

goal



Terminology

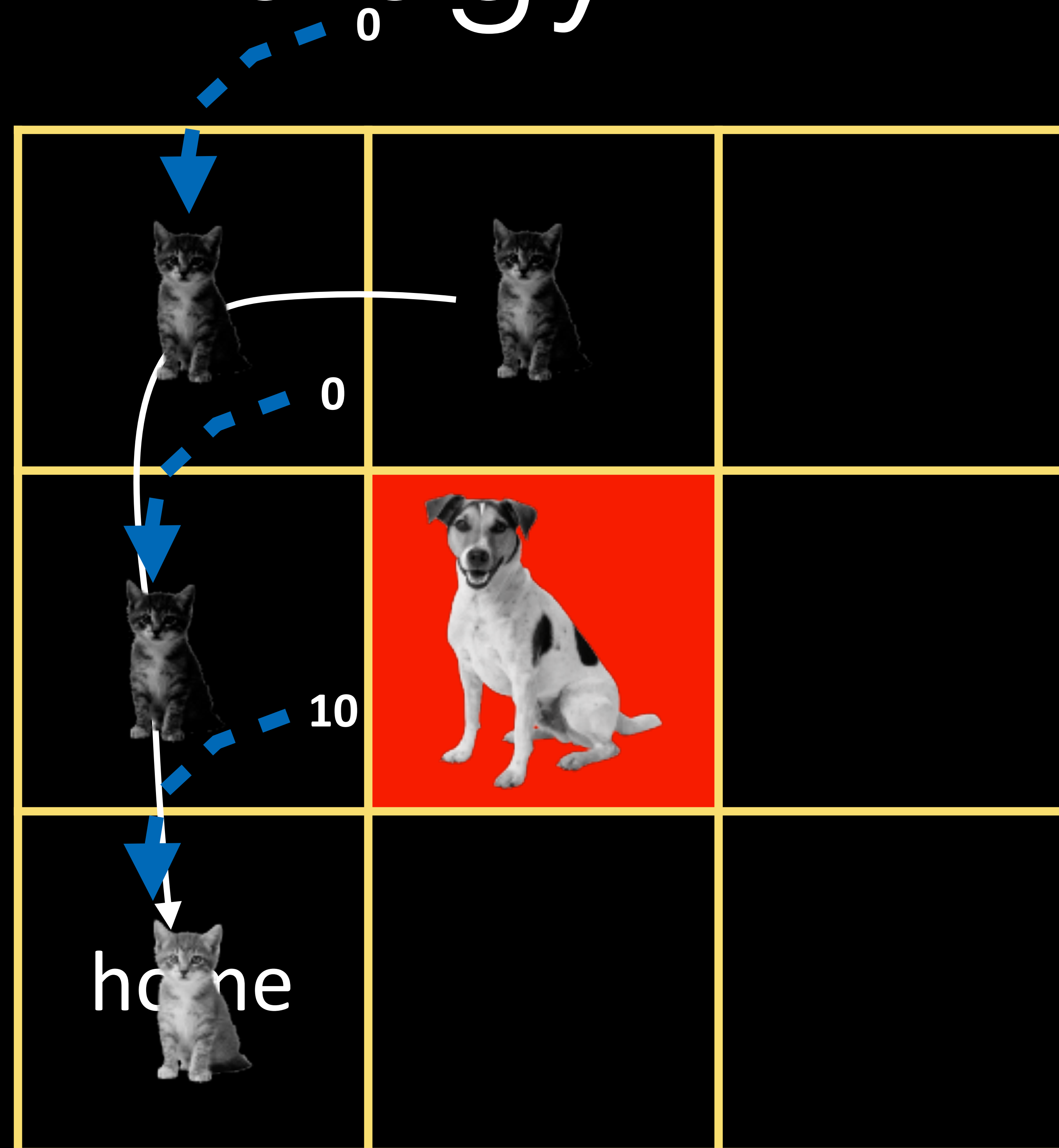
state or action value function

policy

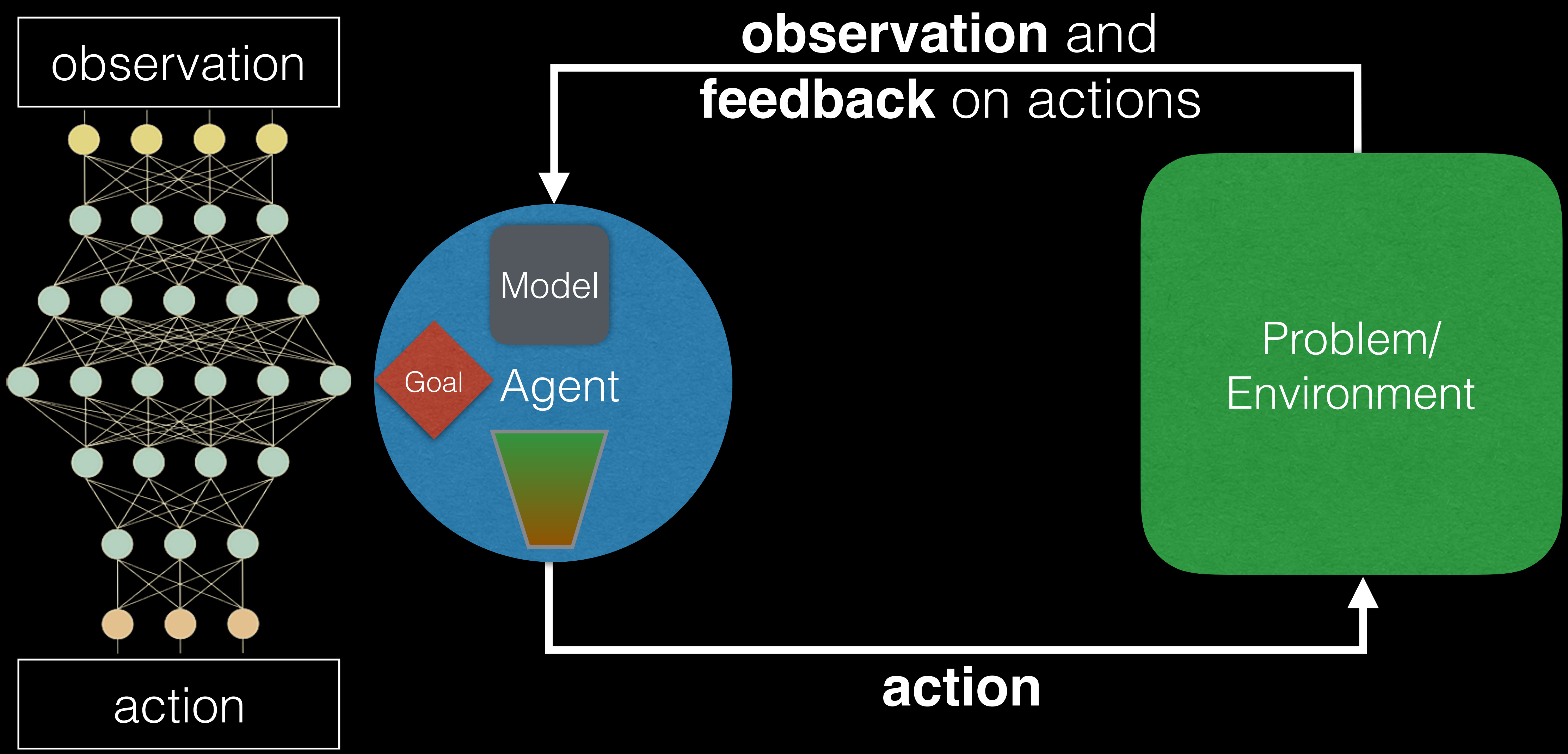
dynamics model

reward

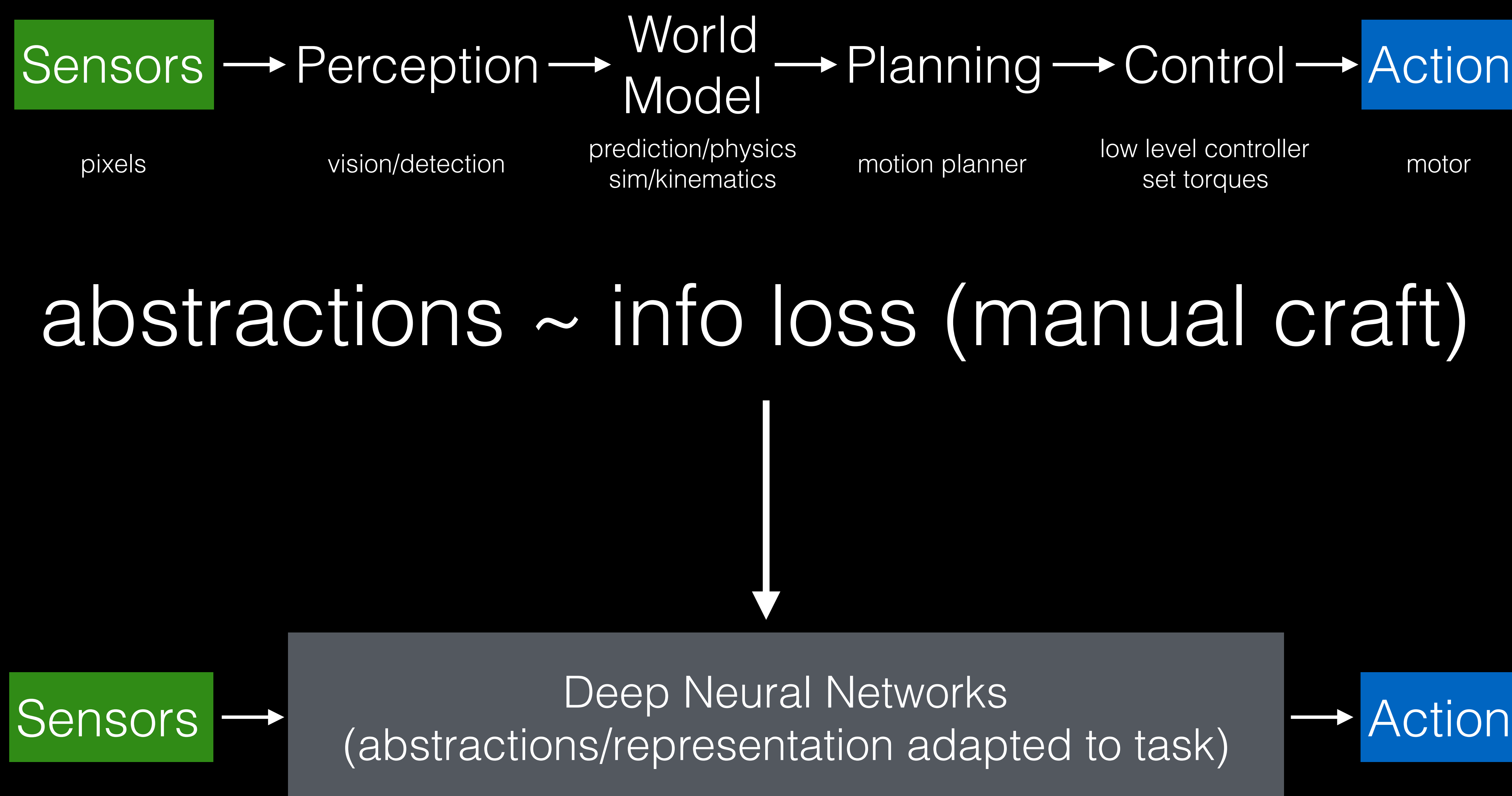
goal



Deep Reinforcement Learning

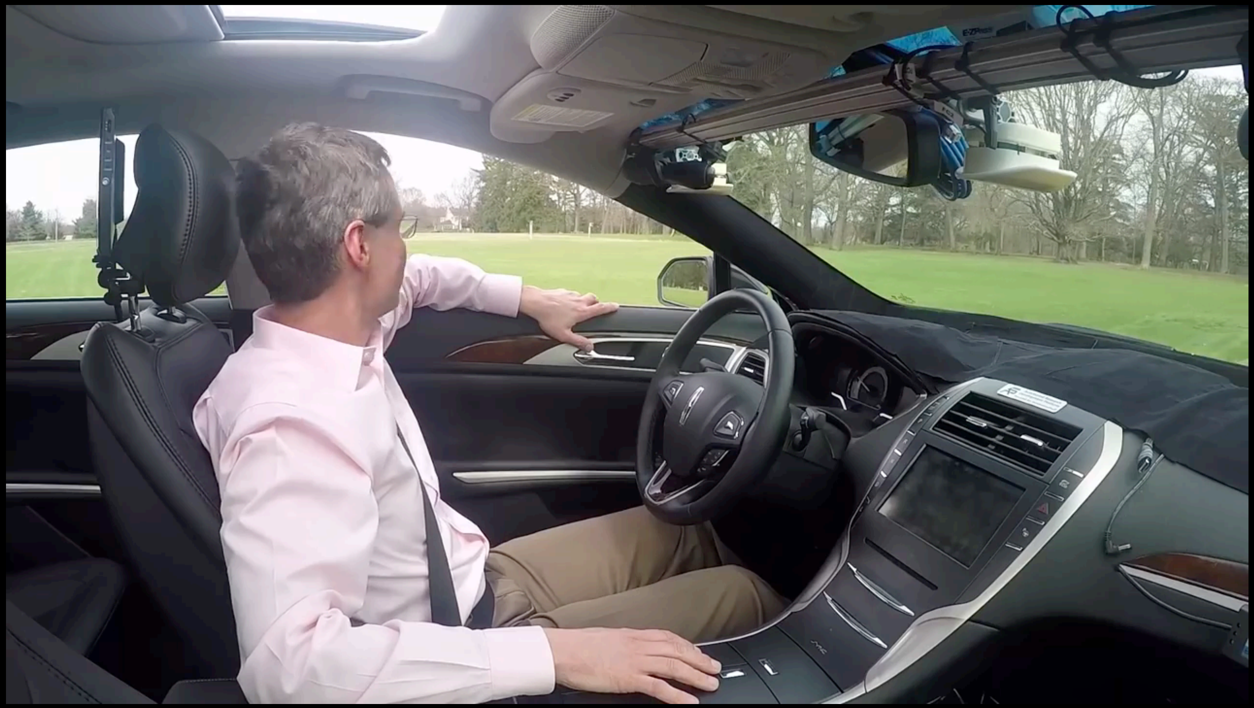


Deep Reinforcement Learning



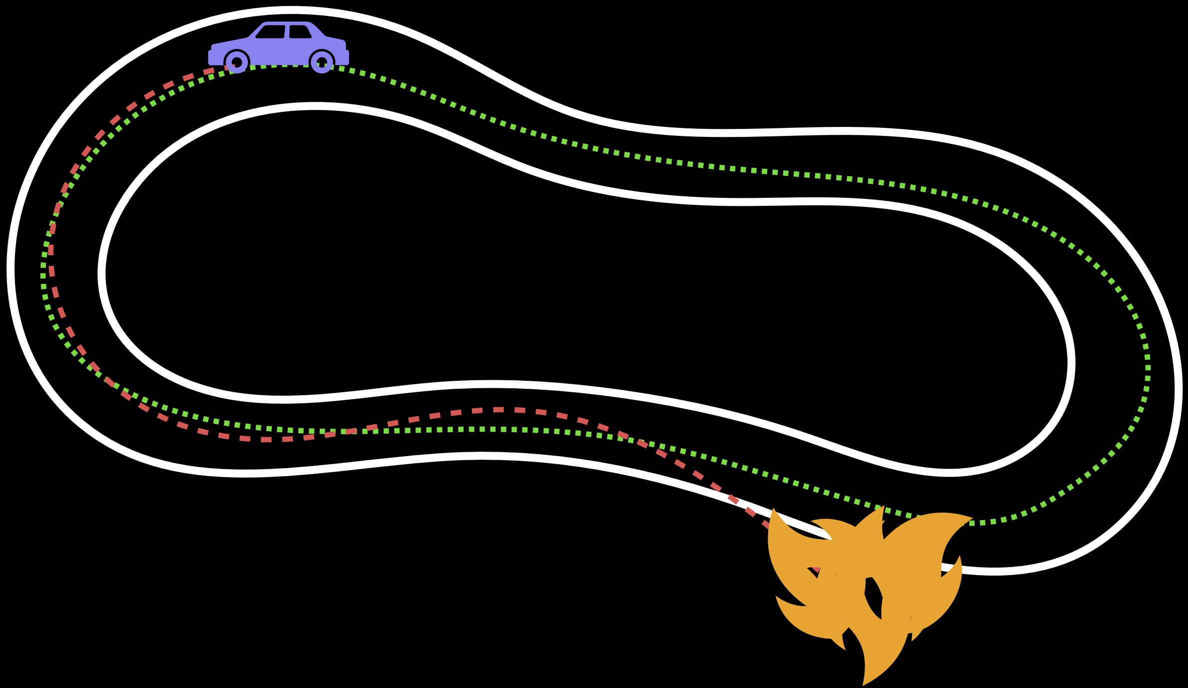
Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car, Bojarski et. al., **https://arxiv.org/pdf/1704.07911.pdf** 2017

SL + RL



https://www.youtube.com/watch?v=NJU9ULQUwng

https://www.youtube.com/watch?v=KnPiP9PkLAs

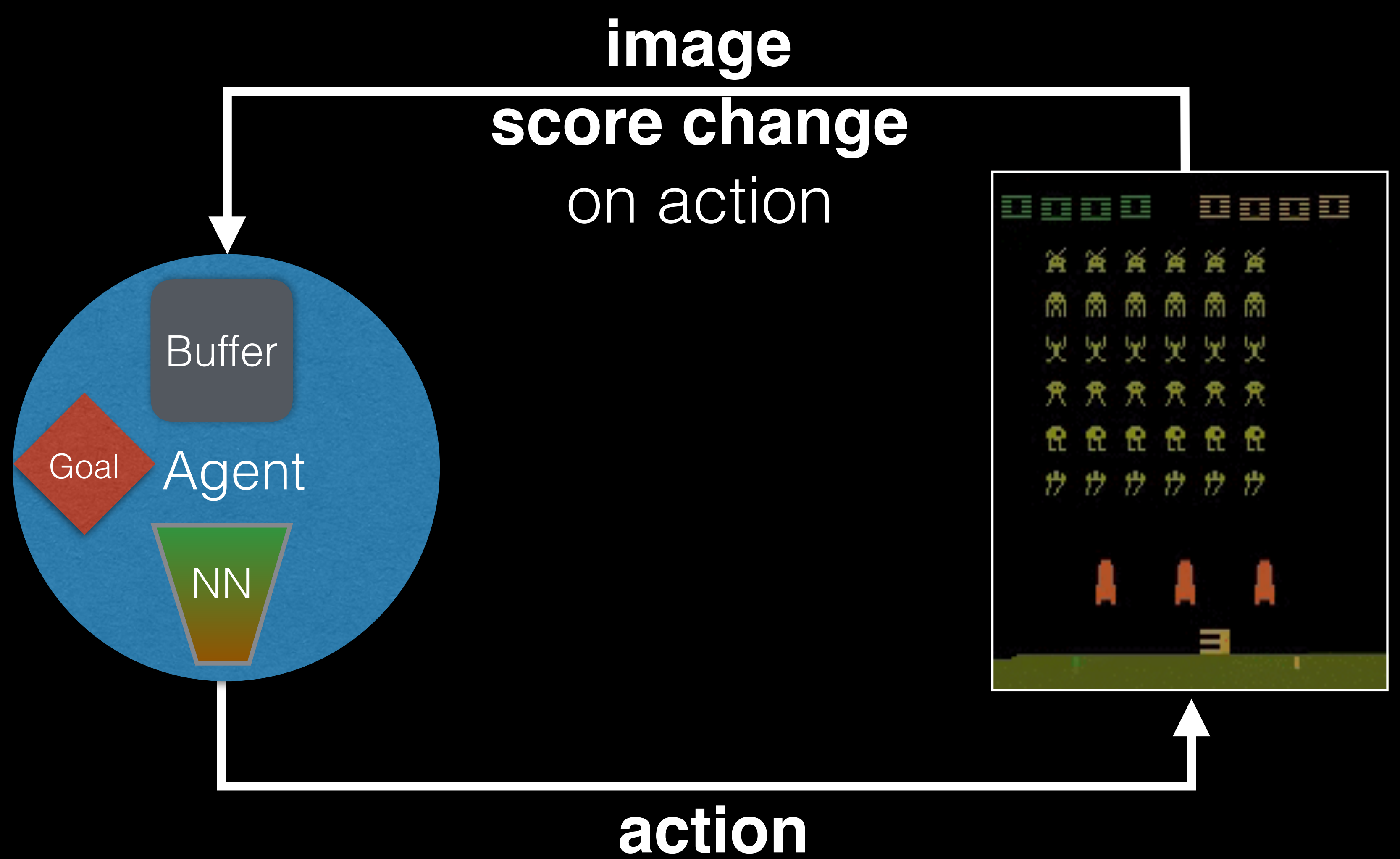


data mismatch

Toolbox

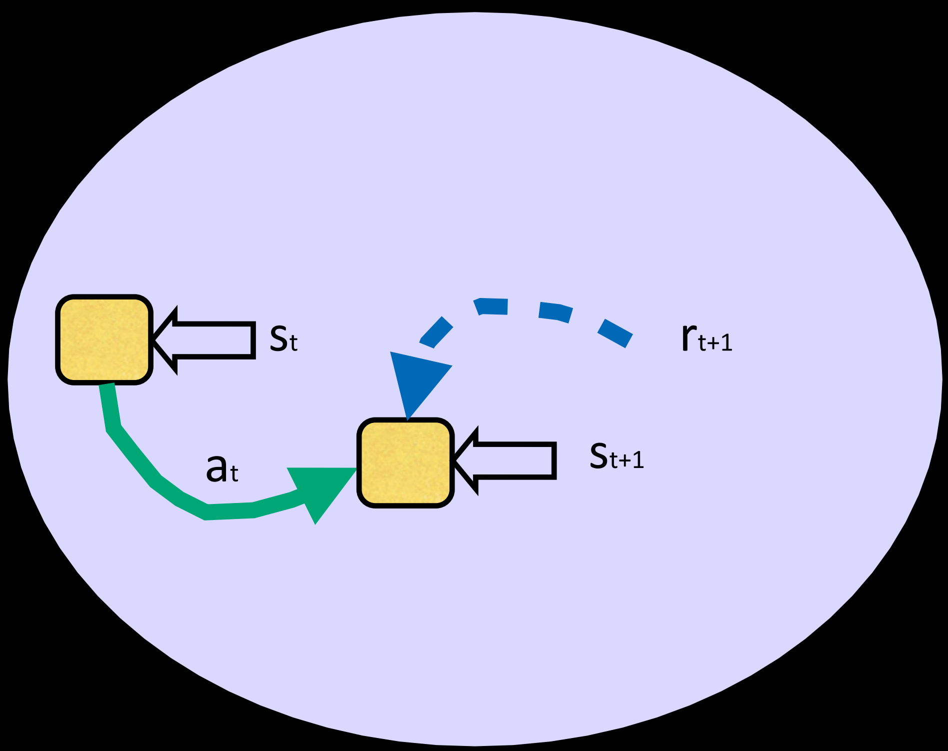
Standard algorithms to give you a **flavour of the norm**!

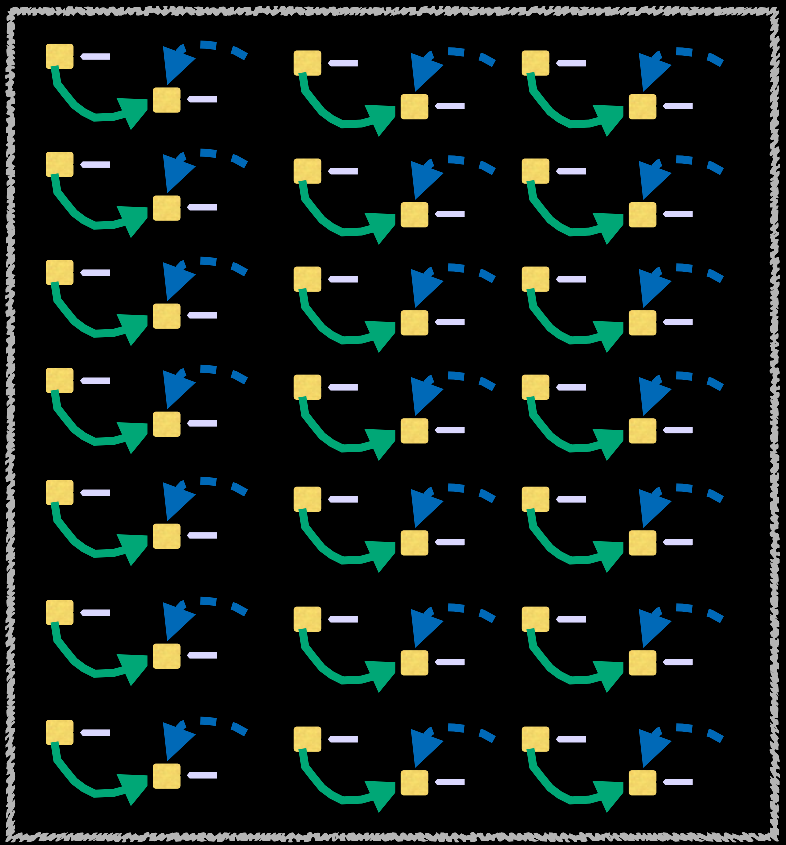
DQN



Human-level control through deep reinforcement learning,
Mnih et. al., Nature 518, Feb 2015

experience replay buffer





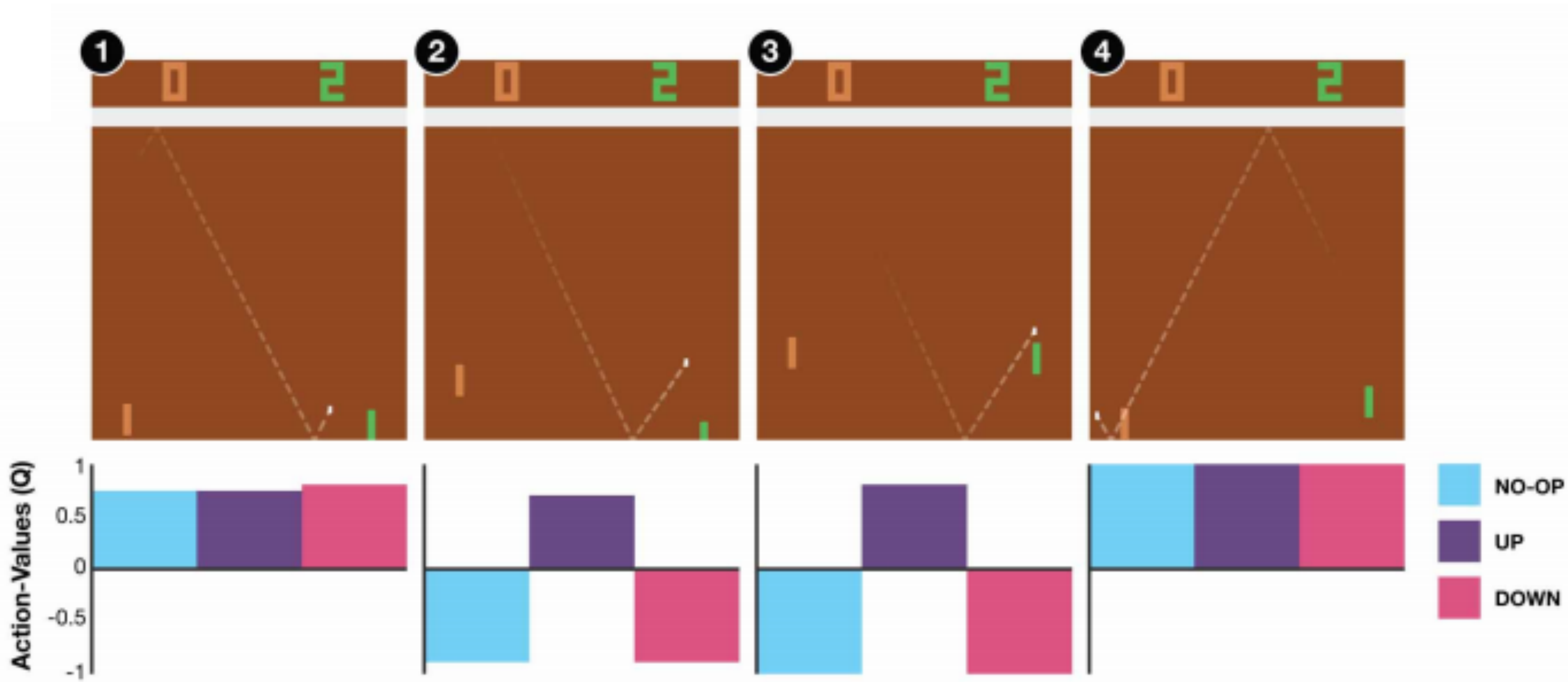
save transition in memory

randomly **sample** from memory for training = i.i.d

freeze target

freeze

$$\left(r + \gamma \max_{a'} Q(s', a', \mathbf{w}^{-}) - Q(s, a, \mathbf{w}) \right)^2$$



https://storage.googleapis.com/deepmind-media/dqn/DQNNaturePaper.pdf

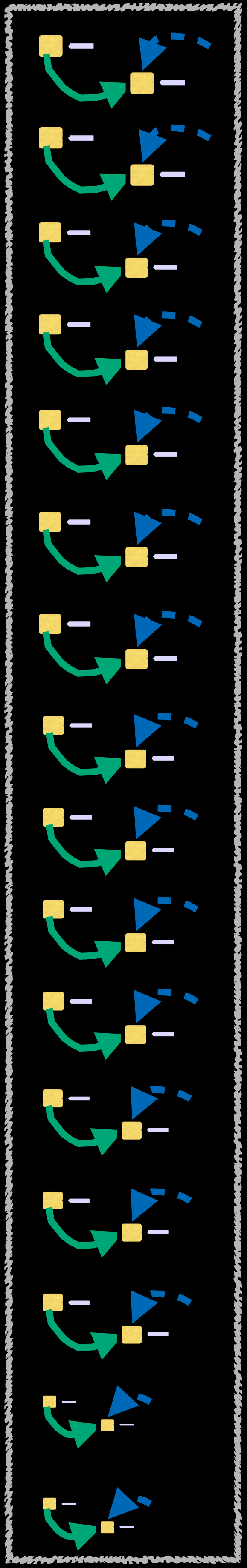
Human-level control through deep reinforcement learning, Mnih et. al., Nature 518, Feb 2015

prioritised experience replay

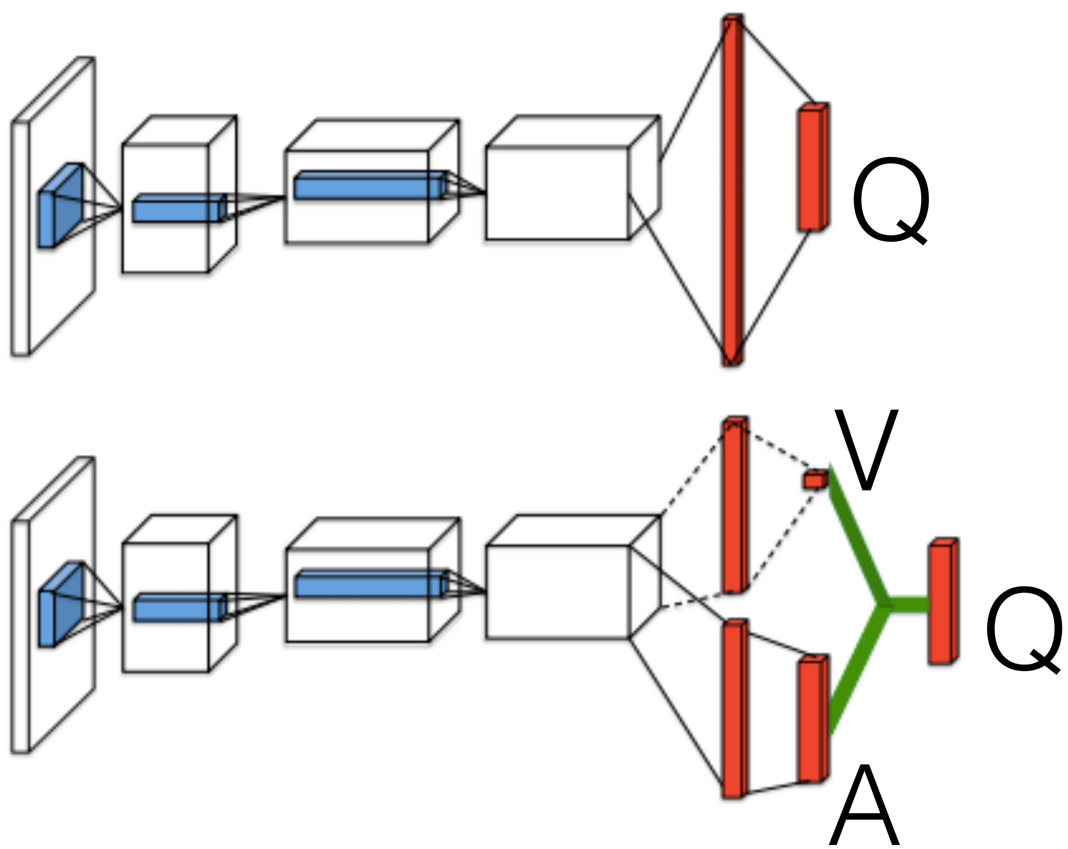
sample
from memory
based on surprise

$$\left| r + \gamma \max_{a'} Q(s', a', \mathbf{w}^-) - Q(s, a, w) \right|$$

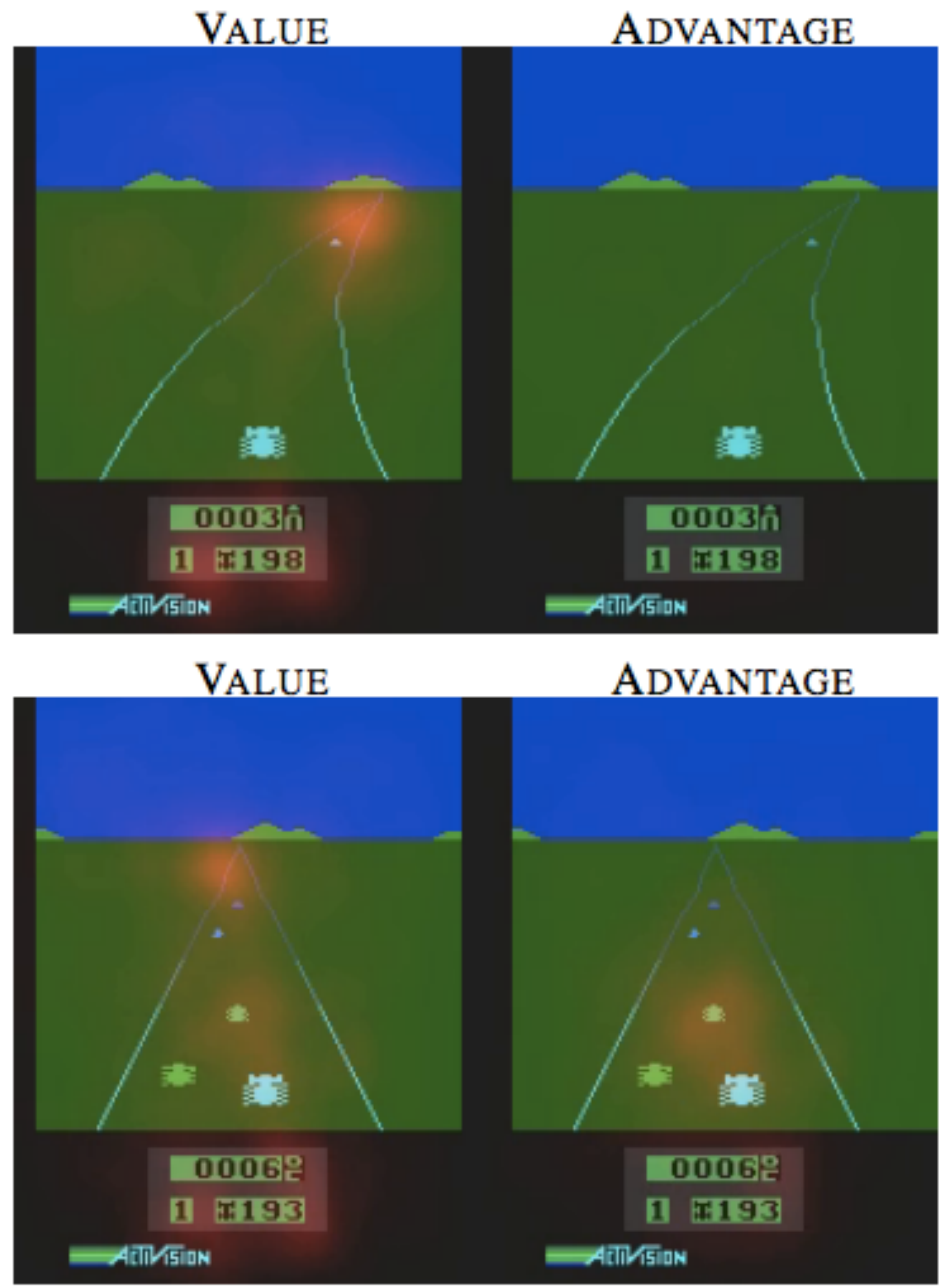
Prioritised Experience Replay, Schaul et. al., **ICLR 2016**



dueling architecture



$$Q(s, a) = V(s) + A(s, a)$$



Dueling Network Architectures for Deep RL Wang et. al., **ICML 2016**

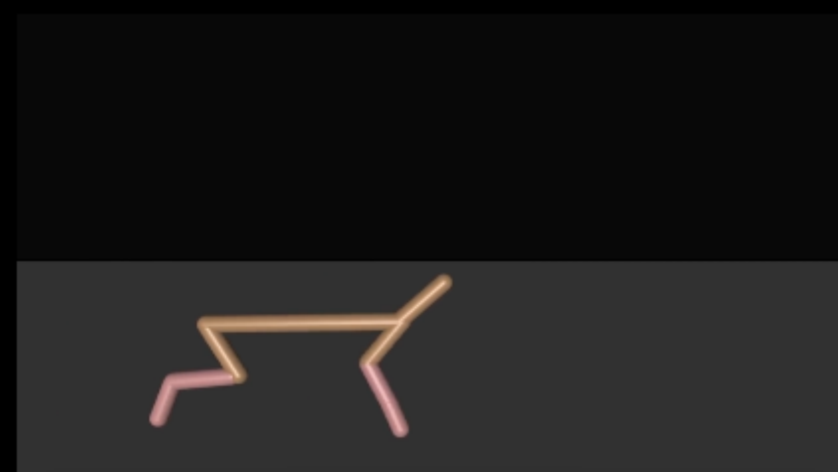
however
training is

SLOOOOo....W

Parallel Asynchronous Training

value and **policy** based methods

https://youtu.be/0xo1Ldx3L5Q



https://youtu.be/Ajjc08-iPx8

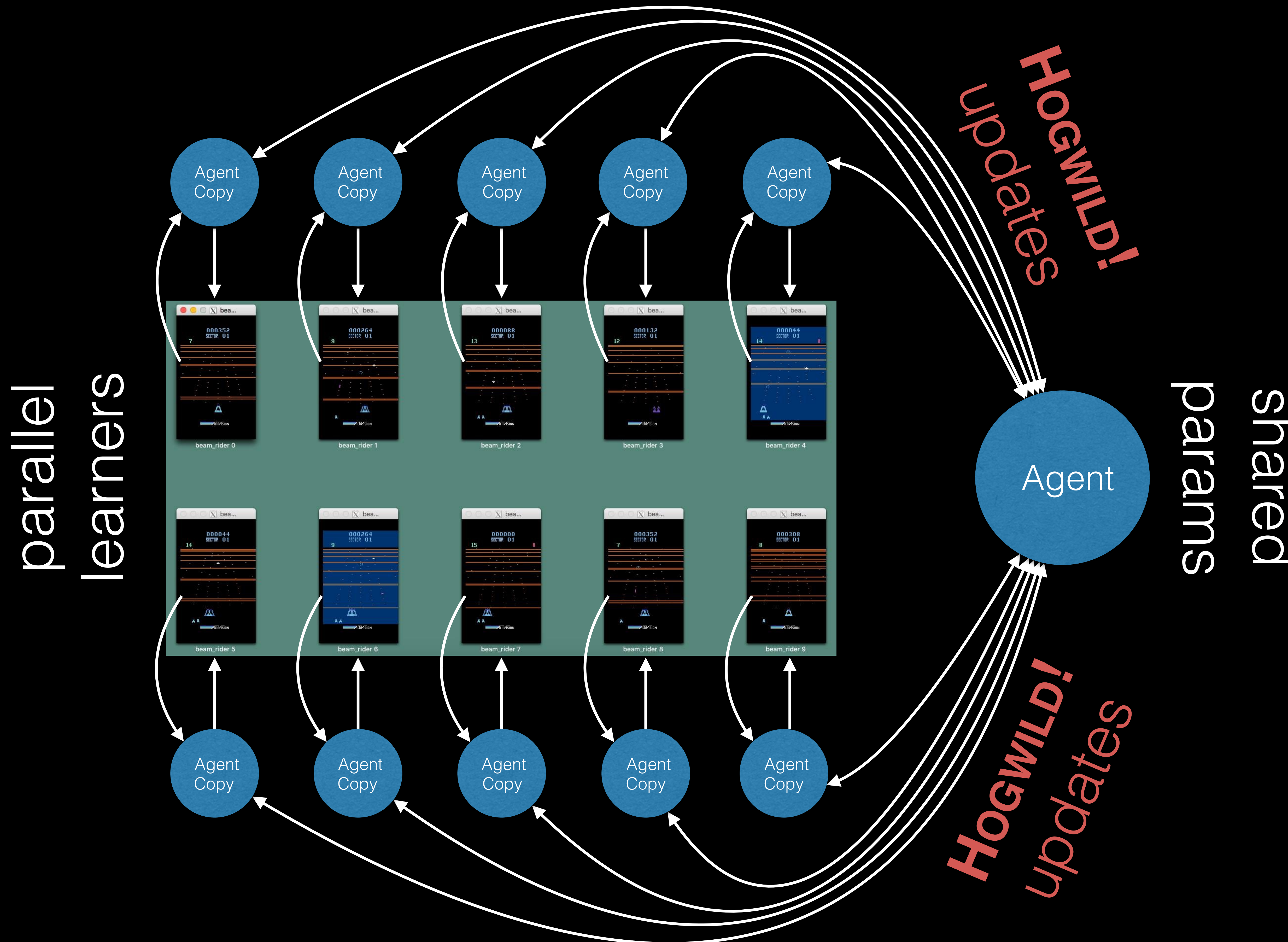
https://youtu.be/nMR5mjCFZCw

parallel agents

shared parameters

lock-free updates

Asynchronous Methods for Deep Reinforcement Learning, Mnih et. al., **ICML 2016**

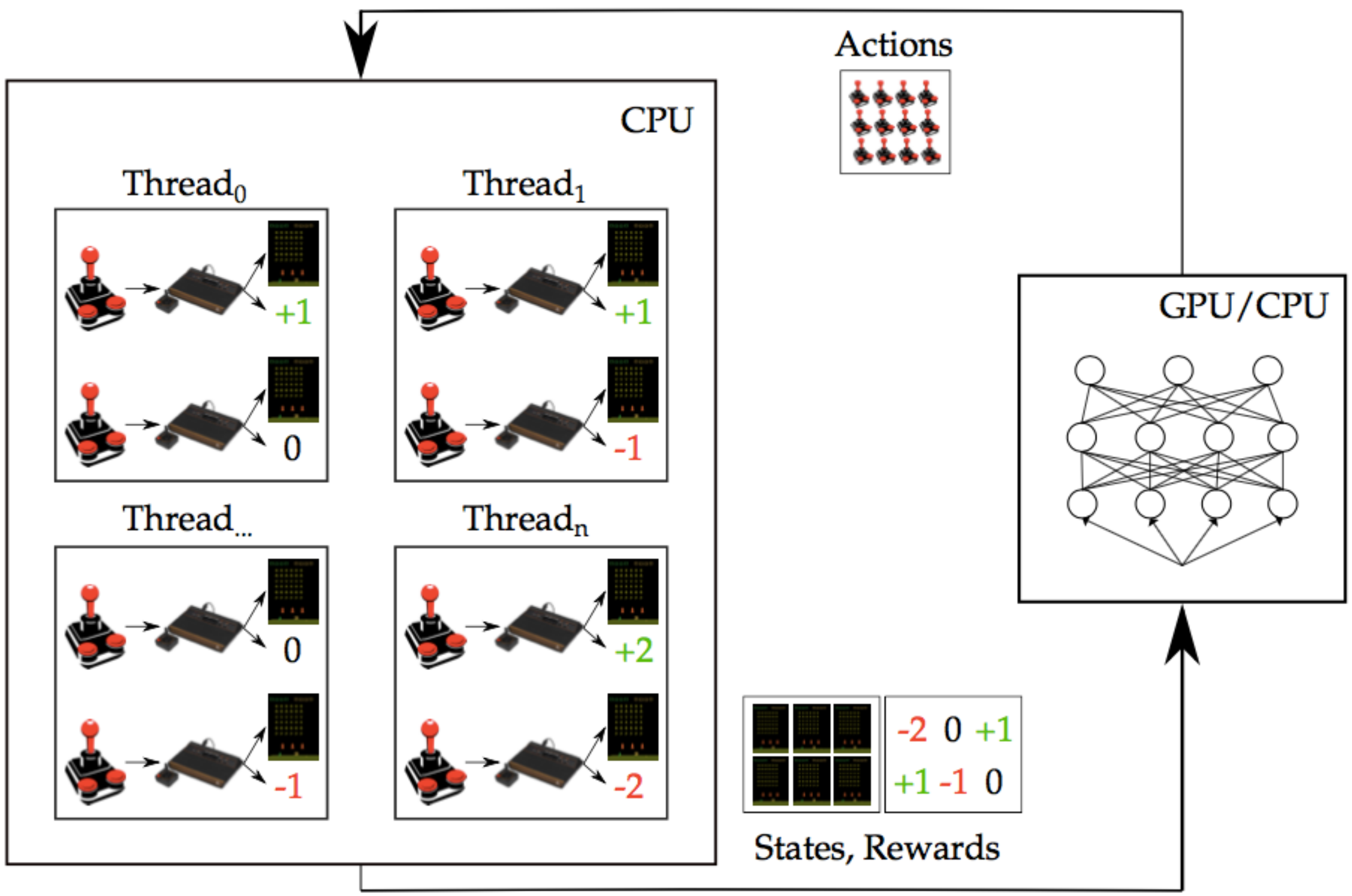


https://github.com/traai/async-deep-rl

So 2016…

Can we train even faster?

PAAC
(**P**arallel **A**dvantage **A**ctor-**C**ritic)



1 GPU/CPU

Reduced training time

SOTA performance

https://github.com/alfredvc/paac

Efficient Parallel Methods for Deep Reinforcement Learning, A. V. Clemente, H. N. Castejón, and A. Chandra, **RLDM 2017**

Alfredo Clemente

Challenges and SOTA

Data Efficiency

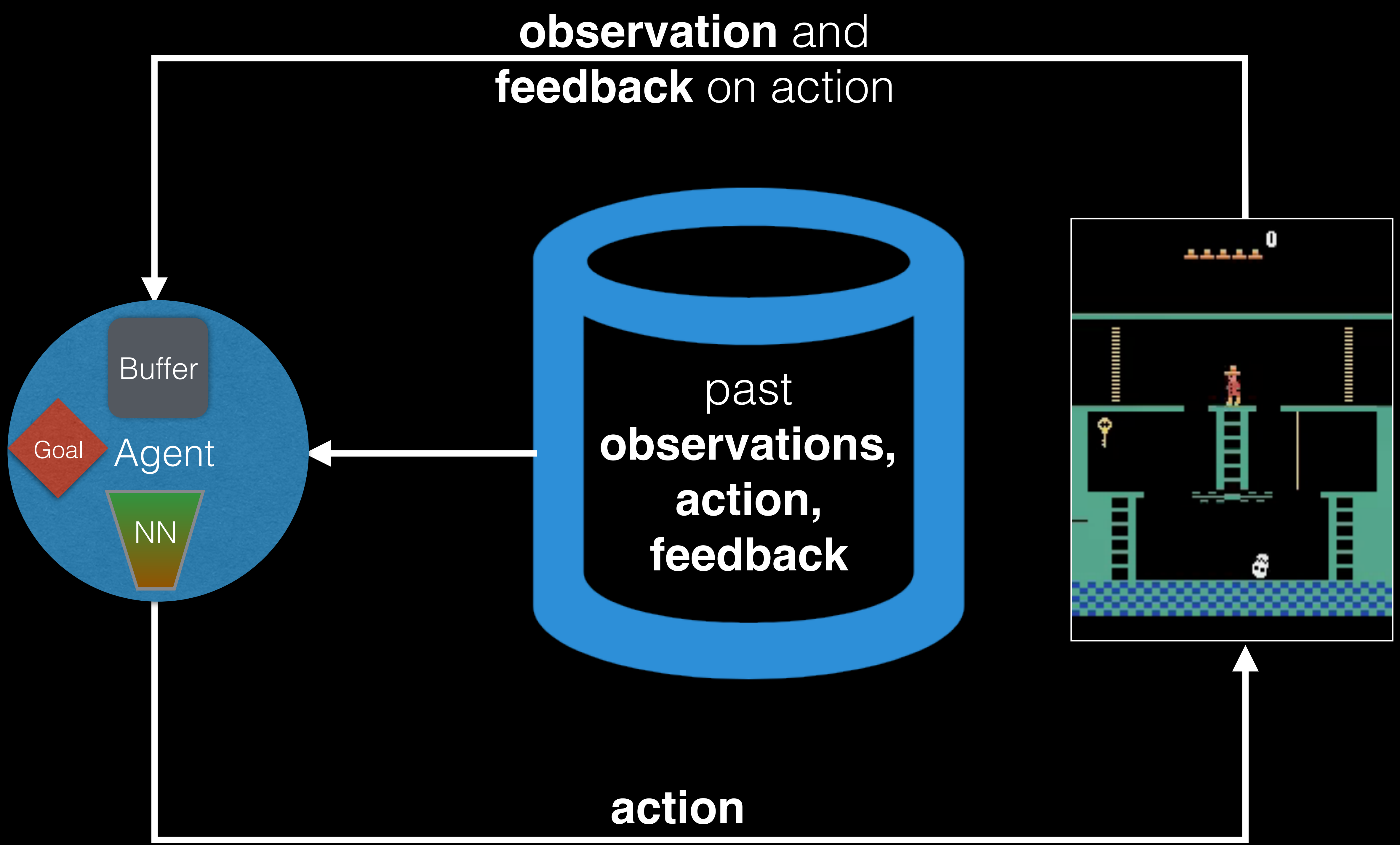
Exploration

Temporal Abstractions

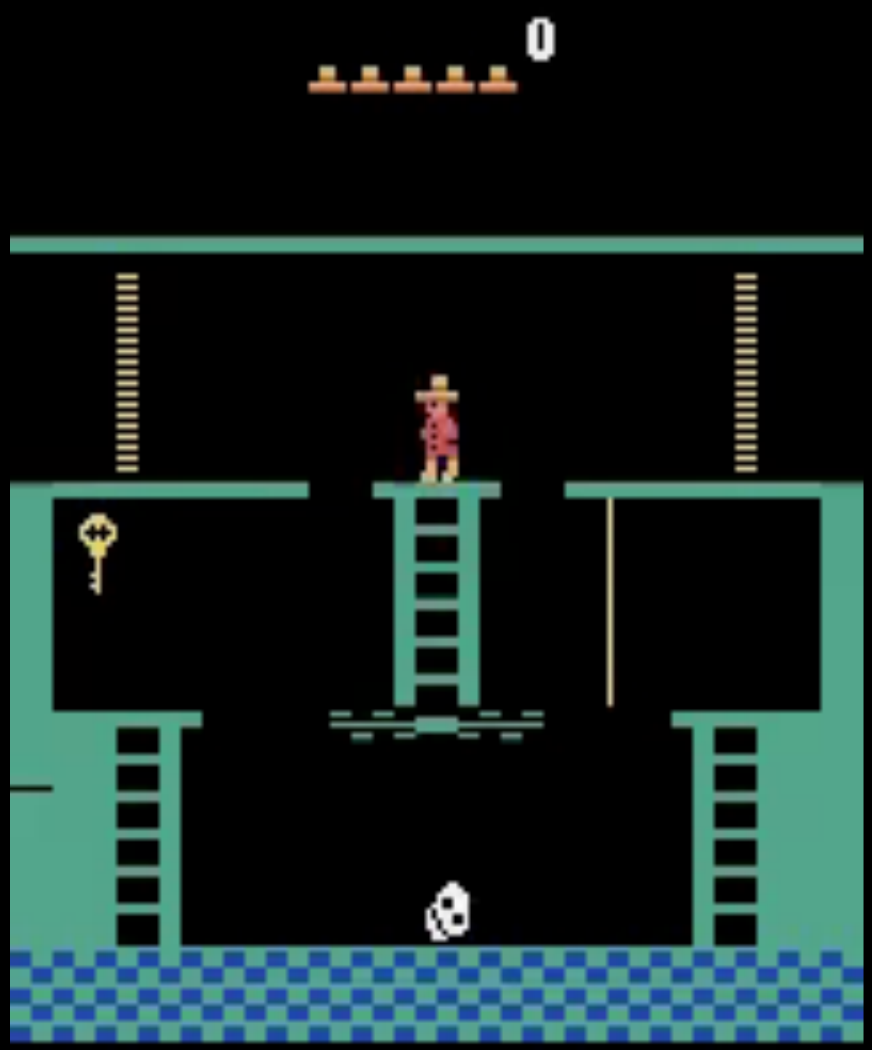
Generalisation

Data Efficiency

Demonstrations



Learning from Demonstrations for Real World Reinforcement Learning, Hester et. al., arXiv e-print, Jul 2017

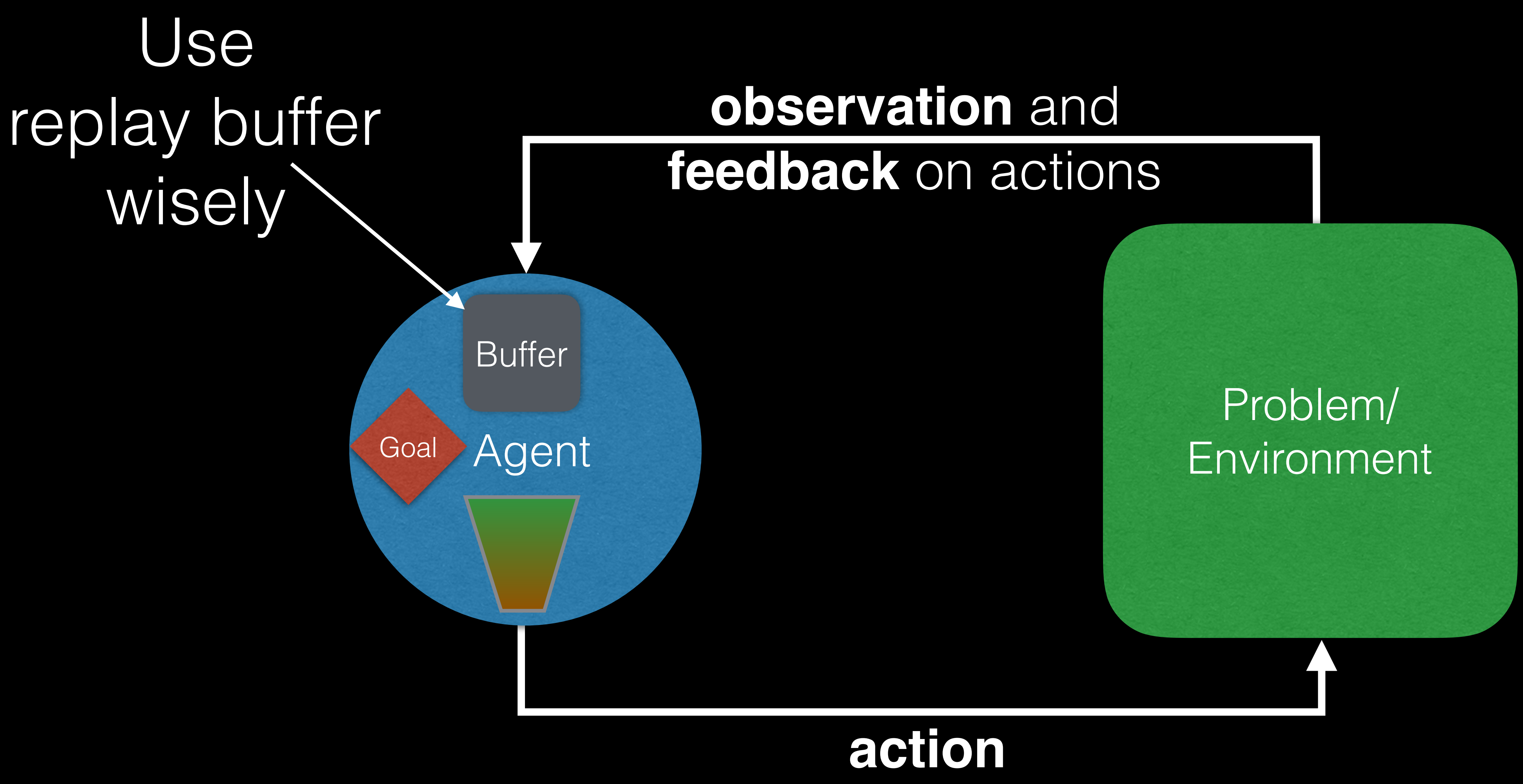


https://www.youtube.com/watch?v=JR6wmLaYuu4

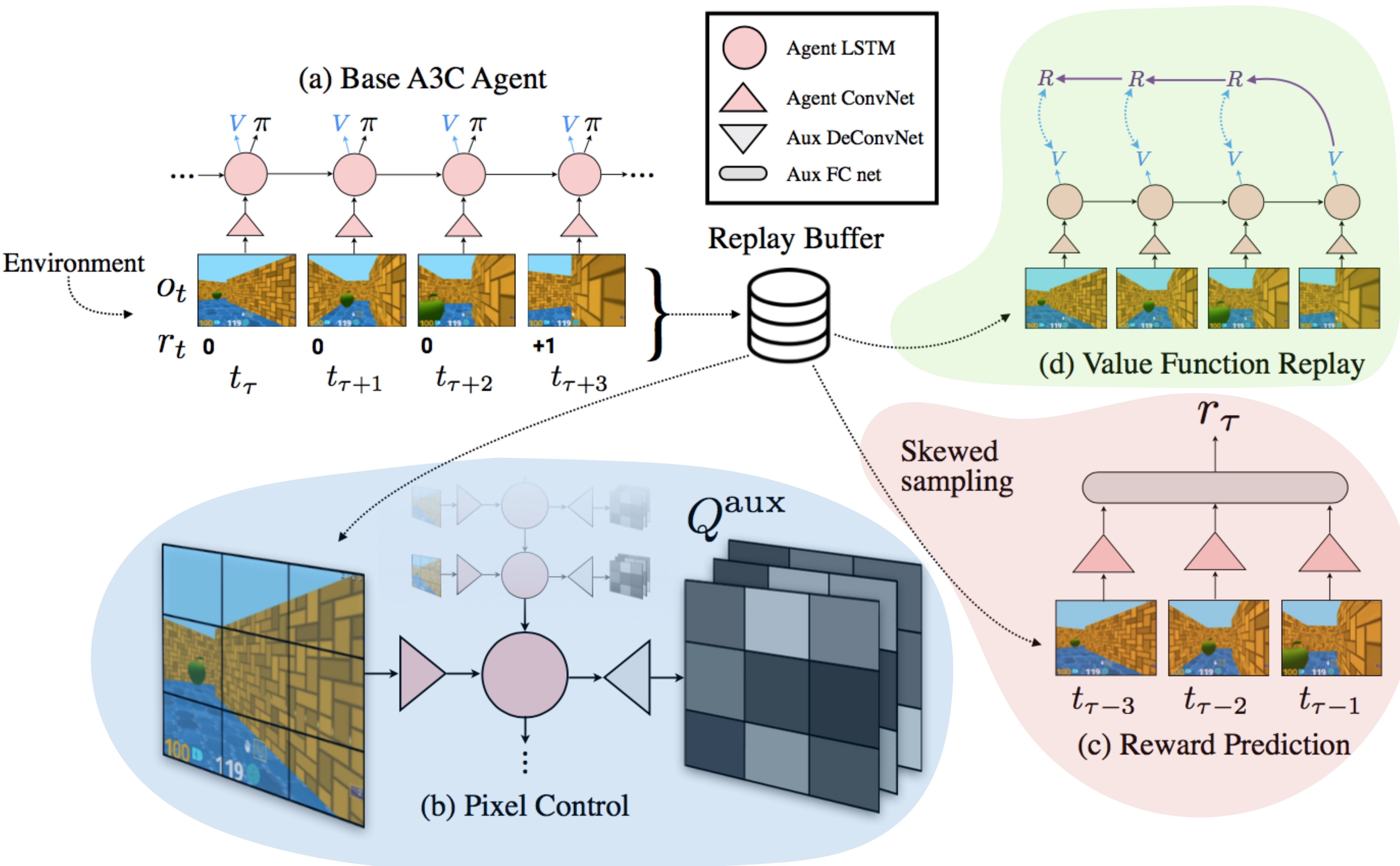
https://www.youtube.com/watch?v=1wsCZk0Im54

https://www.youtube.com/watch?v=B3pf7NJFtHE

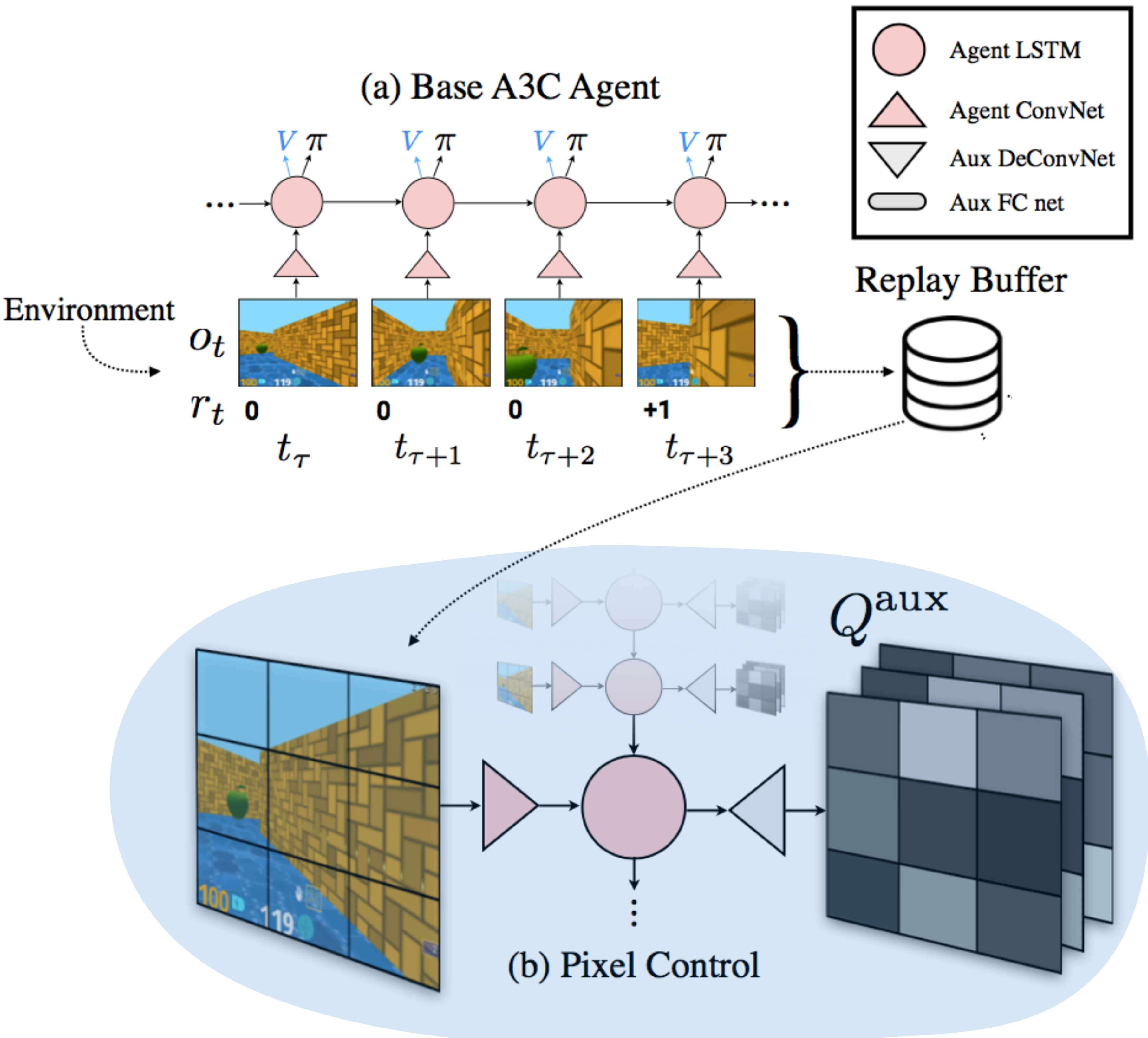
Deep RL with Unsupervised Auxiliary Tasks



Reinforcement Learning with Unsupervised Auxiliary Tasks,
Jaderberg et. al. ICML 2017

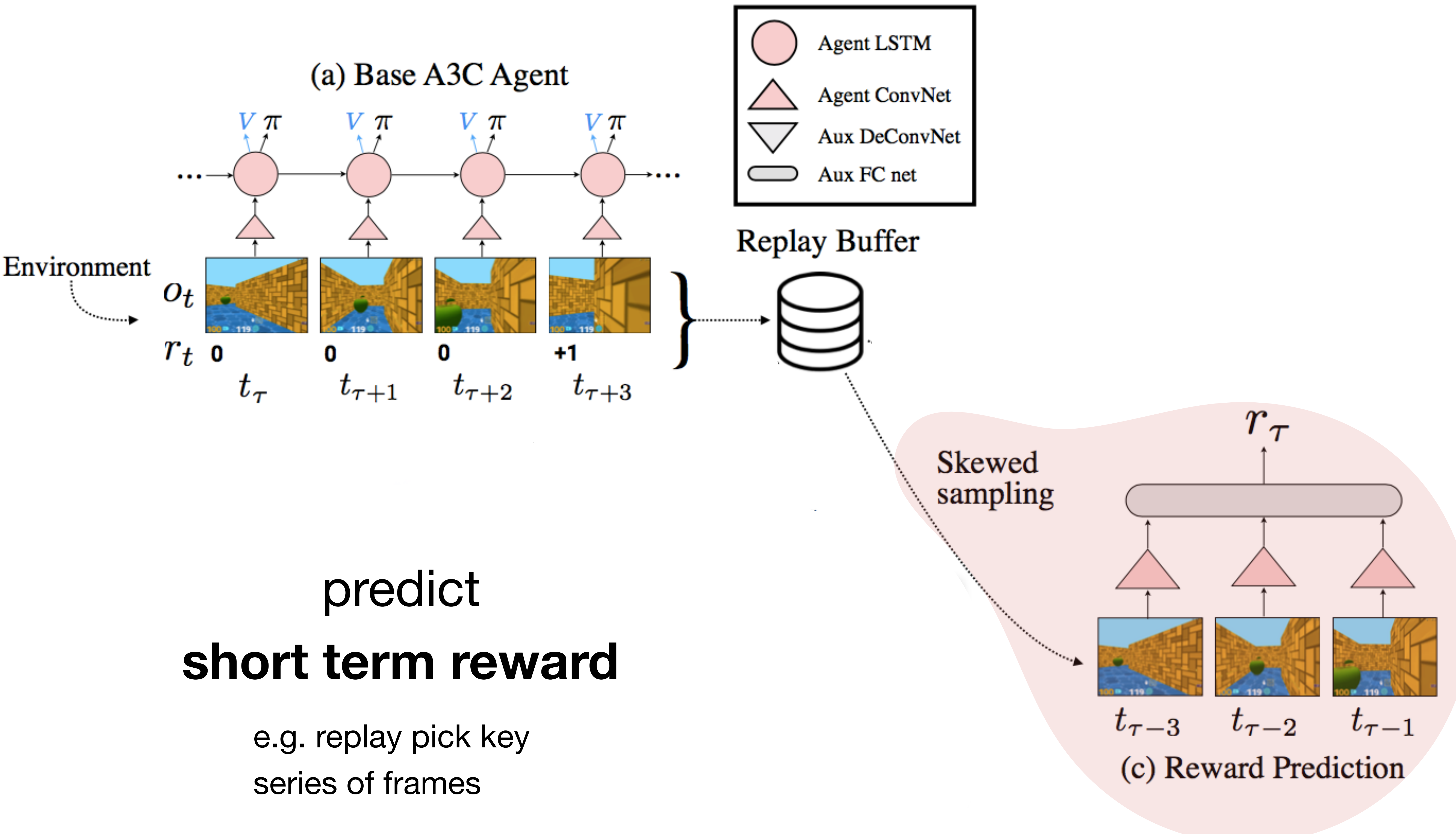


Reinforcement Learning with Unsupervised Auxiliary Tasks, Jaderberg et. al. ICML 2017



learn to **act to affect pixels**

e.g. if grabbing fruit makes it disappear, agent would do it

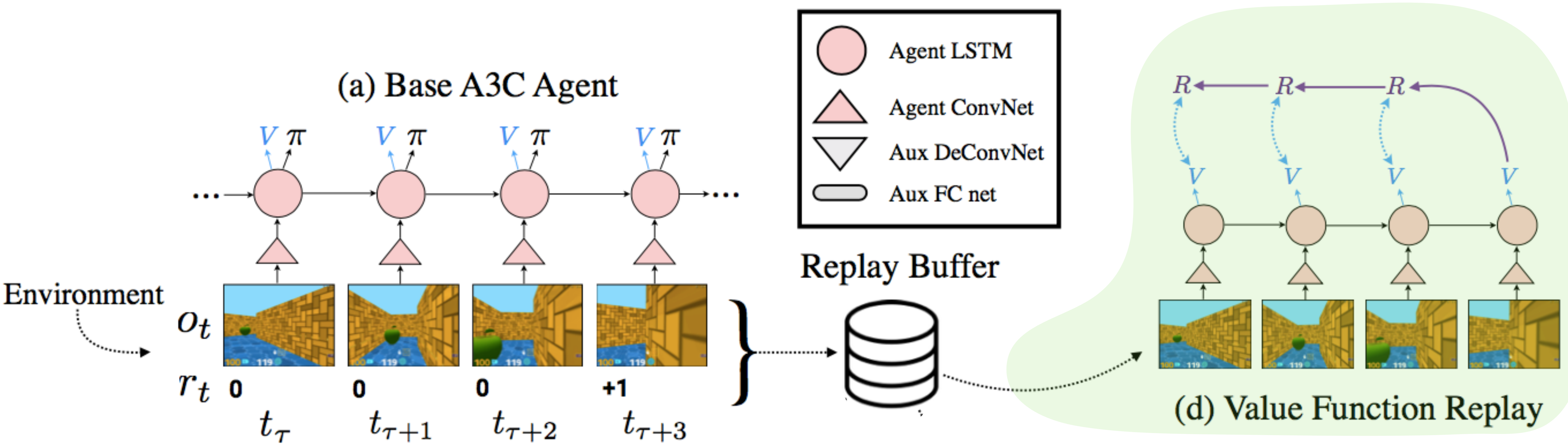


predict

short term reward

e.g. replay pick key

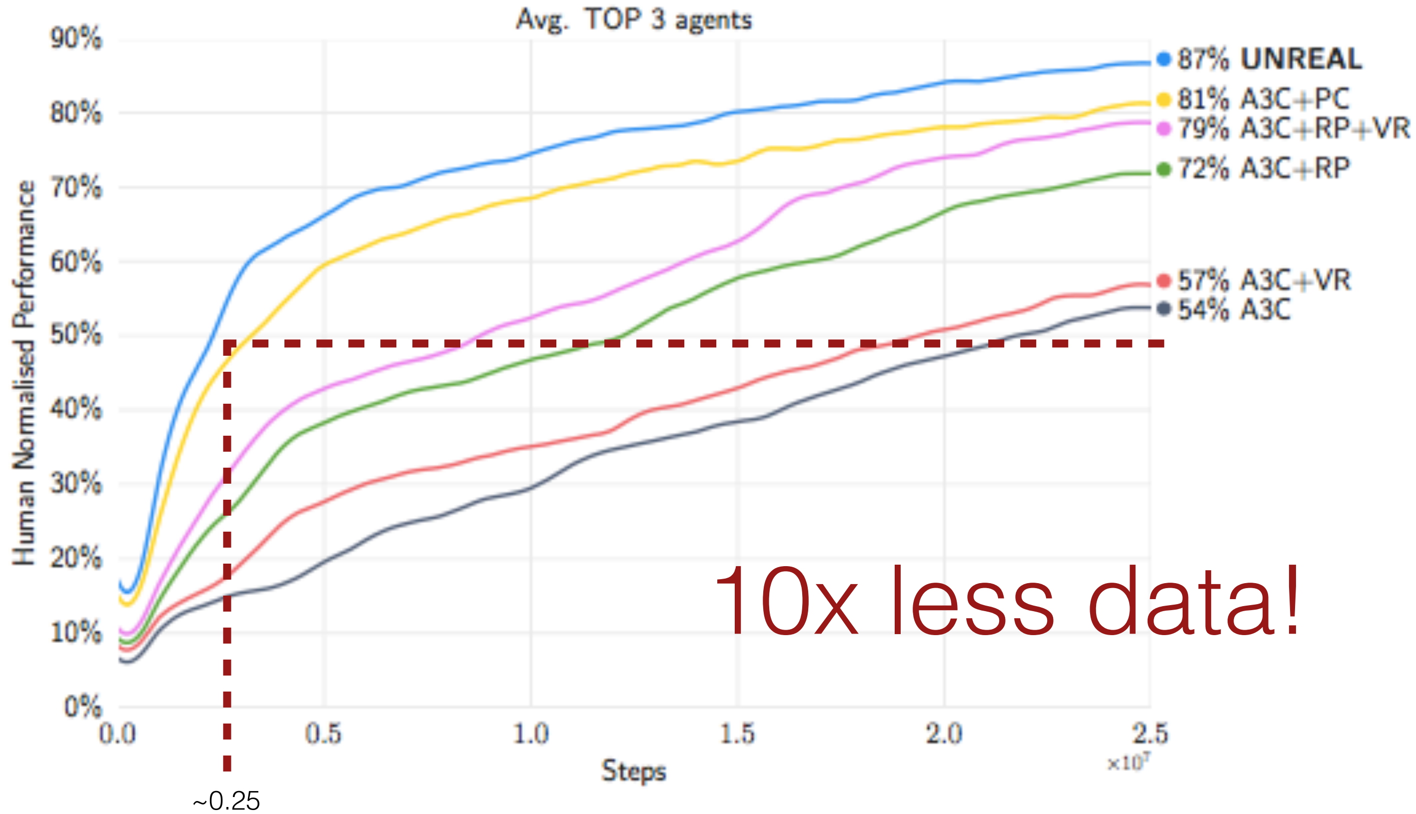
series of frames



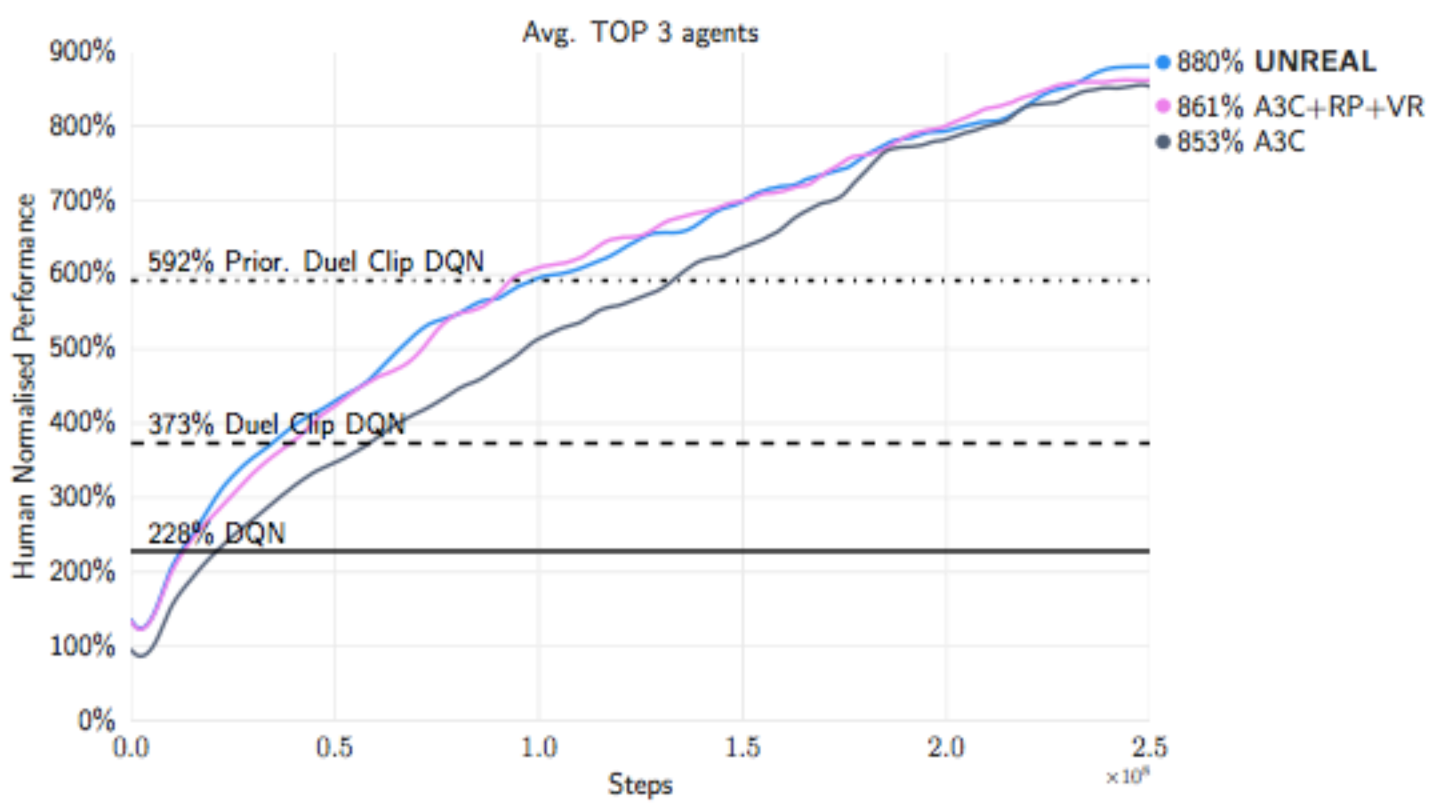
predict

long term reward

Labyrinth Performance



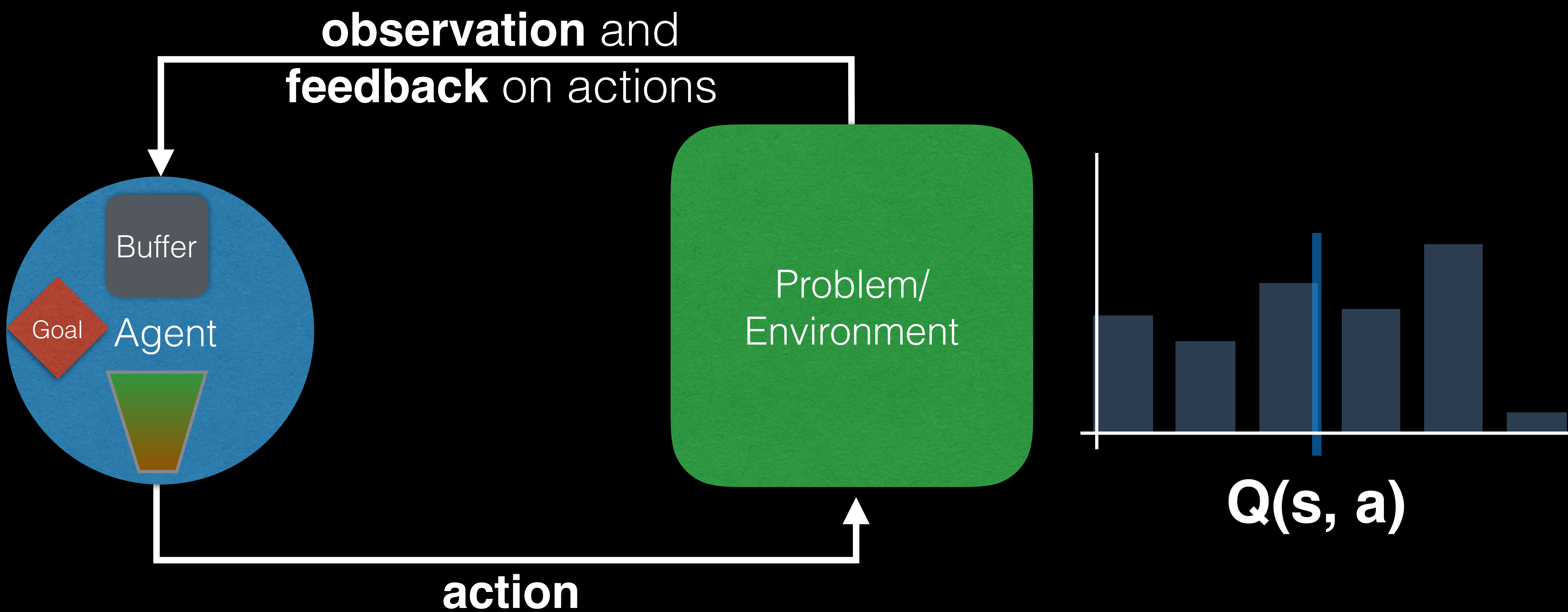
Atari Performance



Reinforcement Learning with Unsupervised Auxiliary Tasks, Jaderberg et. al. ICML 2017

https://deepmind.com/blog/reinforcement-learning-unsupervised-auxiliary-tasks/

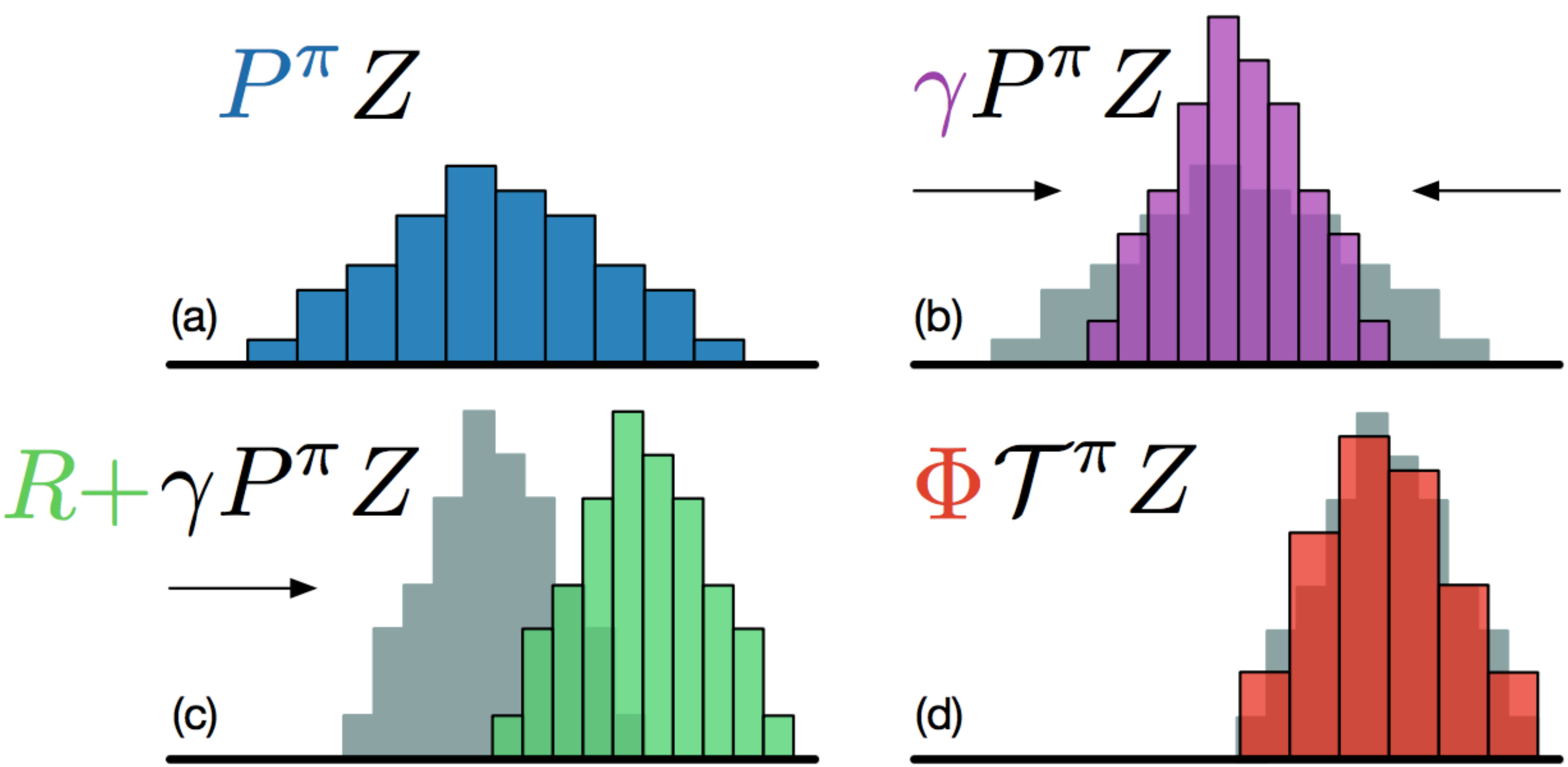
Distributional RL



A Distributional Perspective on Reinforcement Learning,
Bellemare et. al., ICML 2017

Normal DQN **target**:
[sample **reward** after step + **discounted** previous **return** estimate from then on]

BUT this:
[fuse **R** with **discounted** previous **return distribution]**



A Distributional Perspective on Reinforcement Learning, Bellemare et. al., ICML 2017

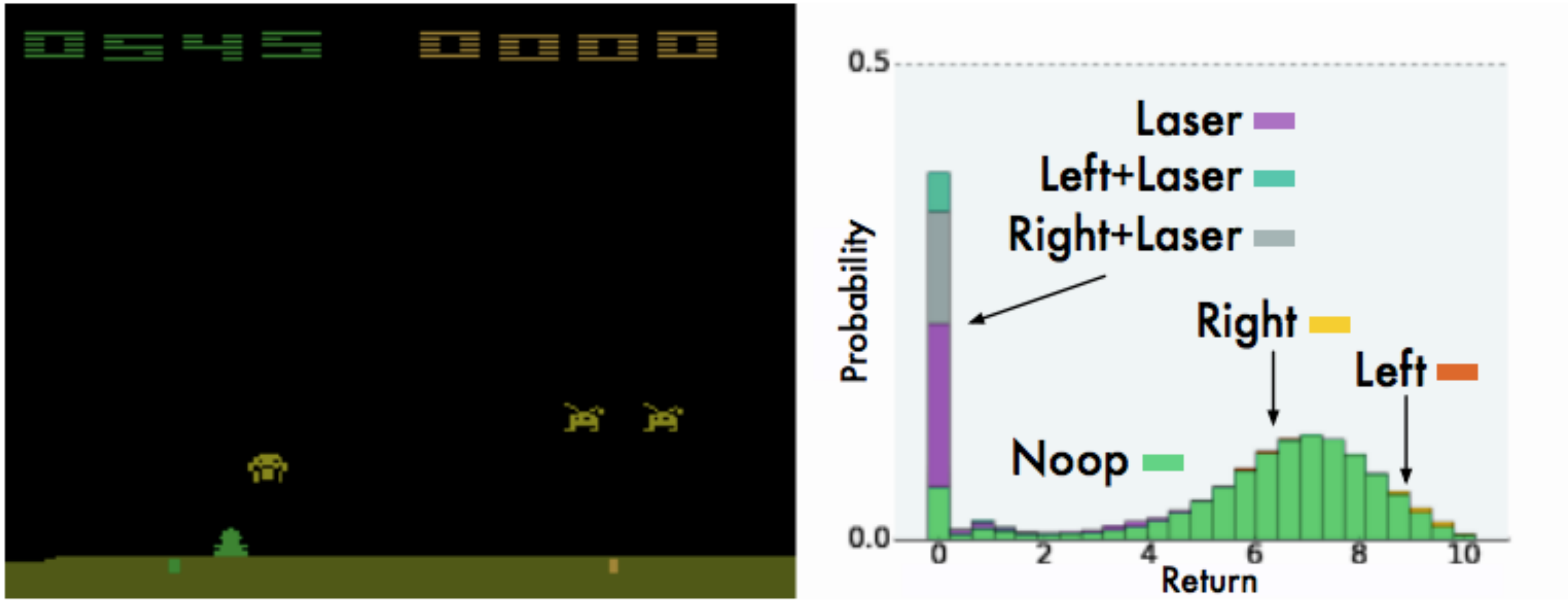
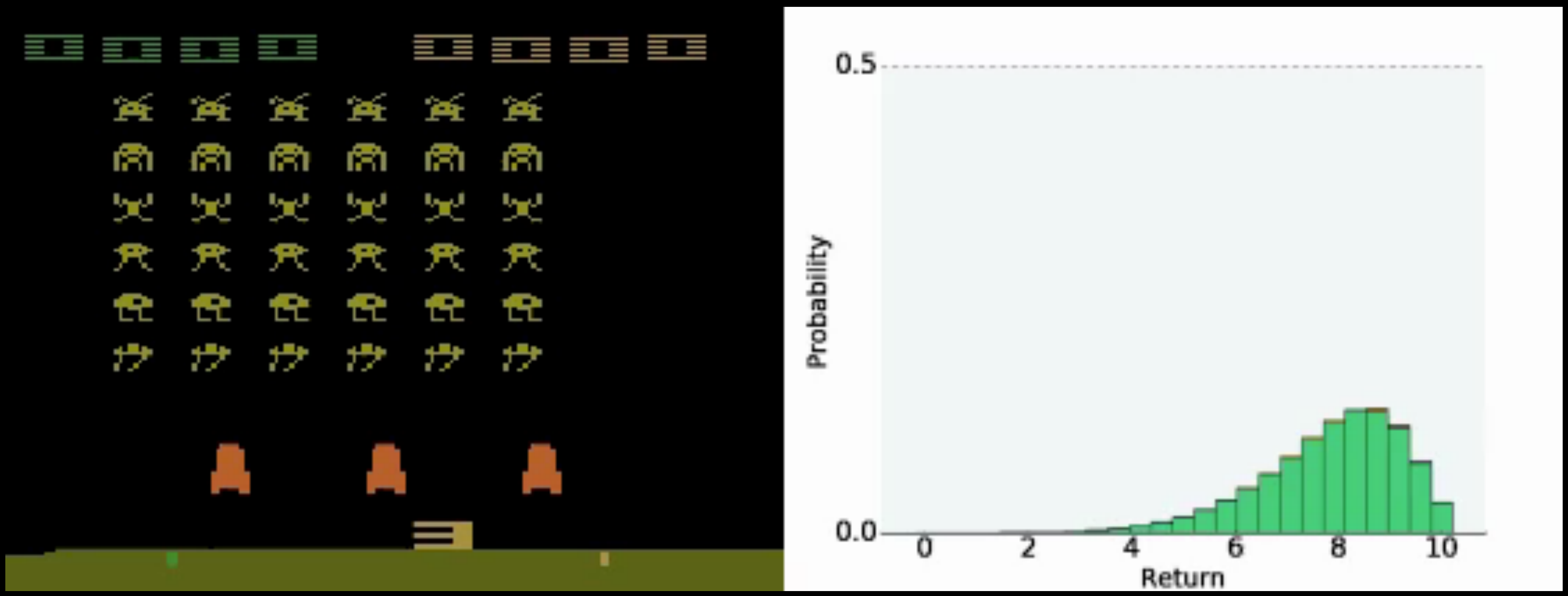


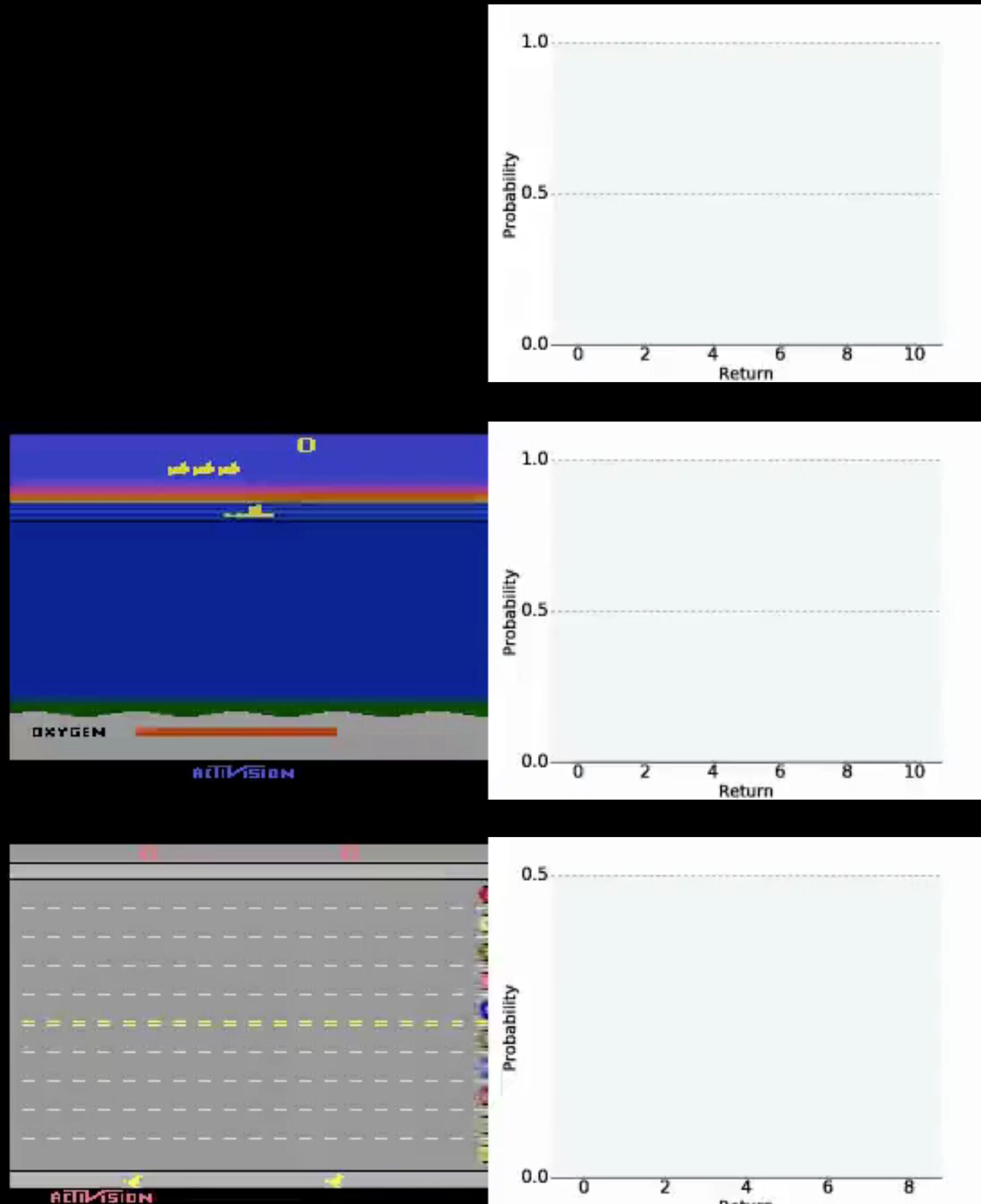
Figure 4. Learned value distribution during an episode of SPACE INVADERS. Different actions are shaded different colours. Returns below 0 (which do not occur in SPACE INVADERS) are not shown here as the agent assigns virtually no probability to them.

"If I shoot now, it is game over for me"

A Distributional Perspective on Reinforcement Learning, Bellemare et. al., ICML 2017



A Distributional Perspective on Reinforcement Learning, Bellemare et. al., ICML 2017



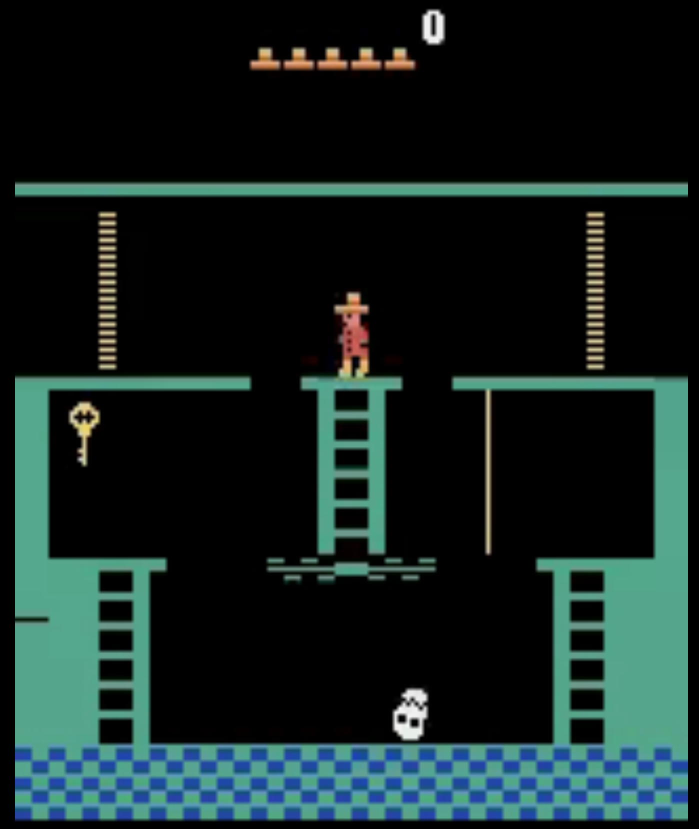
wrong/fatal actions

bimodal

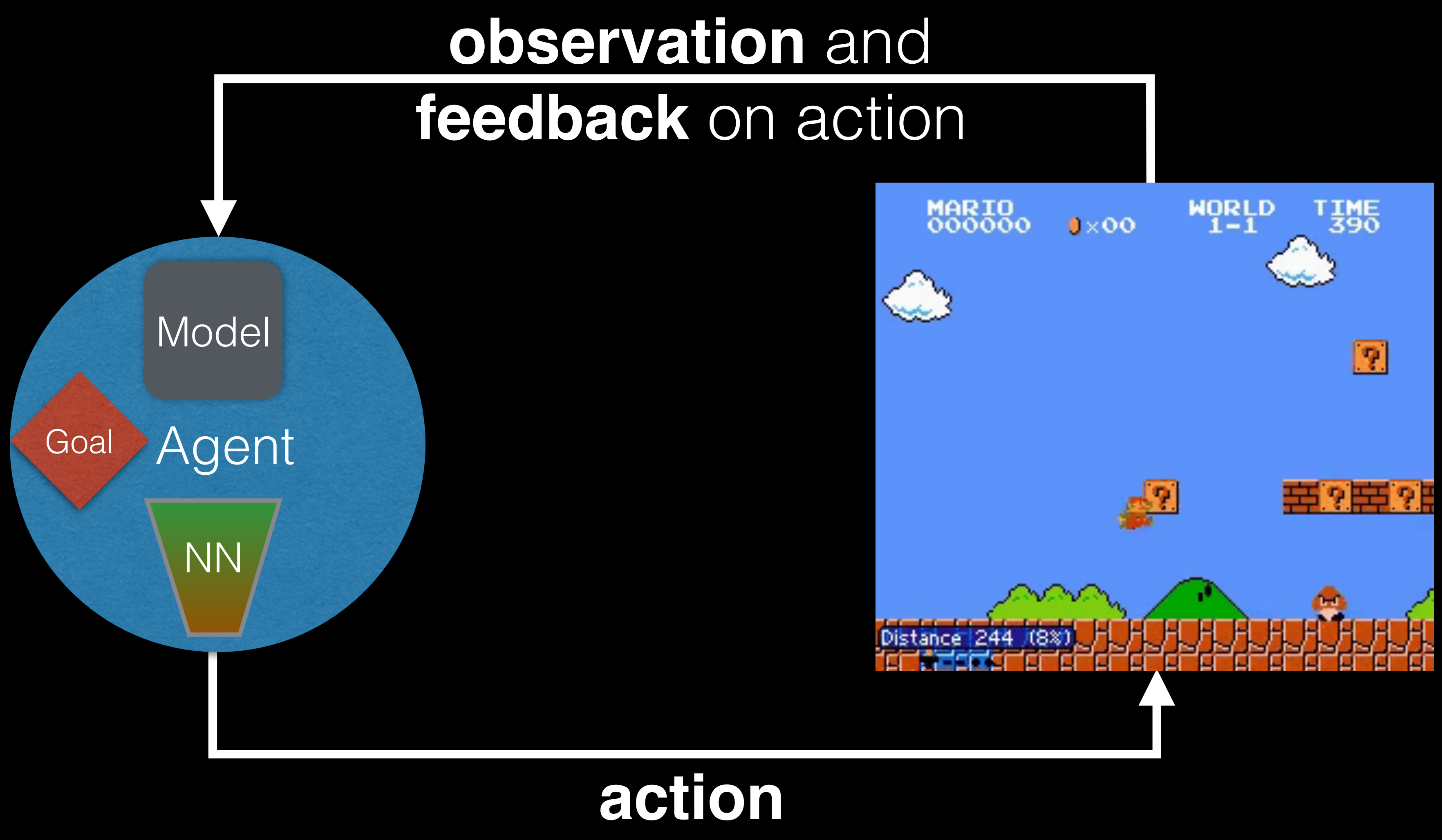
under pressure

A Distributional Perspective on Reinforcement Learning, Bellemare et. al., ICML 2017

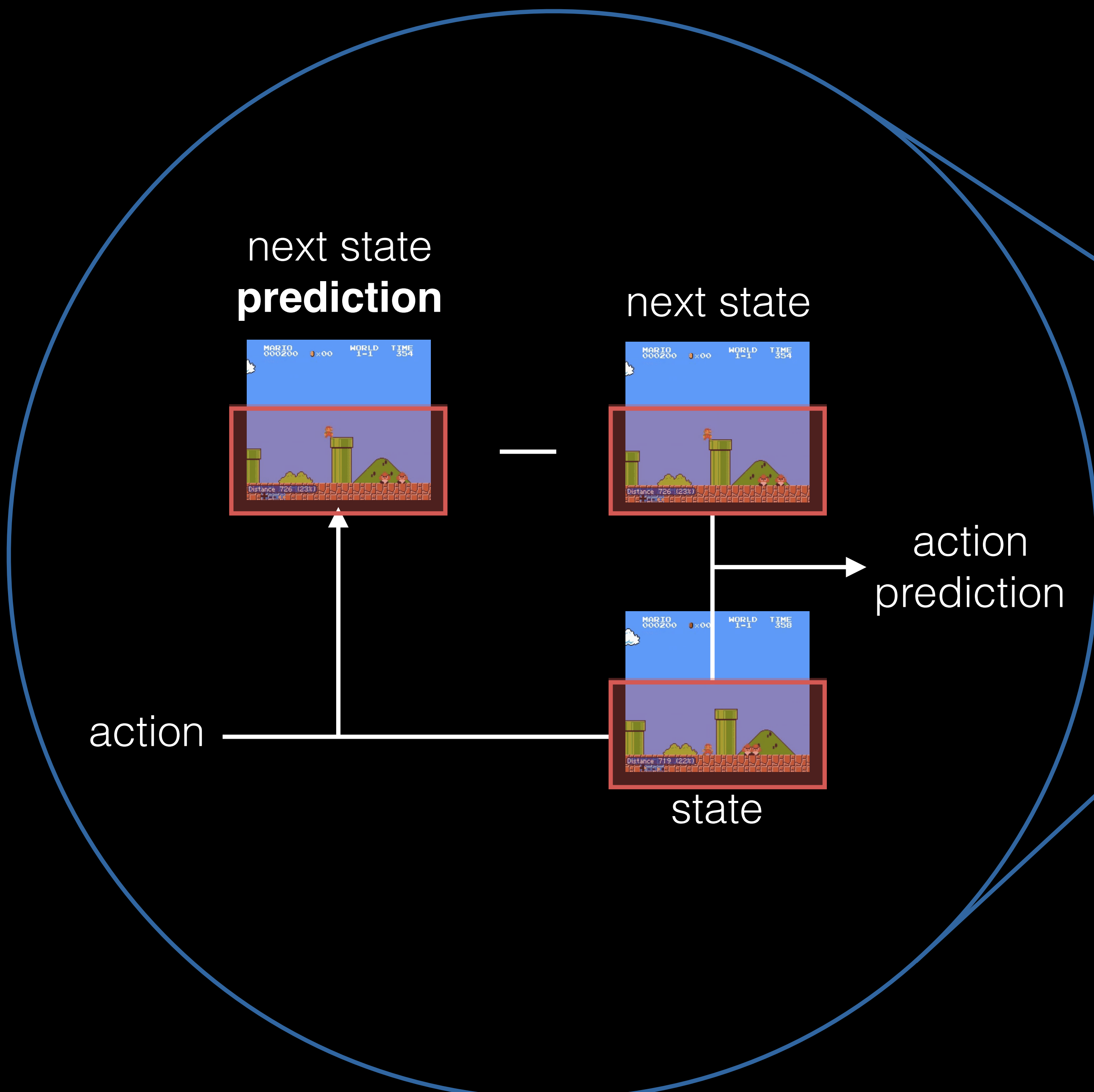
Exploration



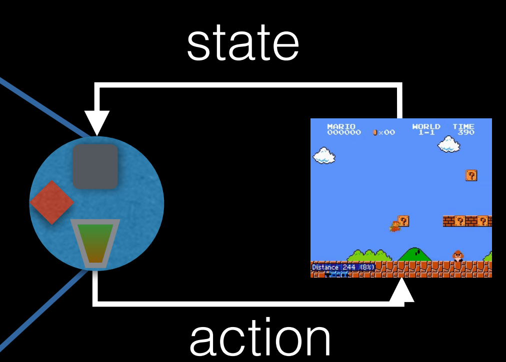
Curiosity Driven Exploration



Curiosity Driven Exploration



curiosity as
next state
prediction error



… only focus on
relevant
parts of state

Curiosity-driven Exploration by Self-supervised Prediction,
Pathak, Agrawal et al., ICML 2017.

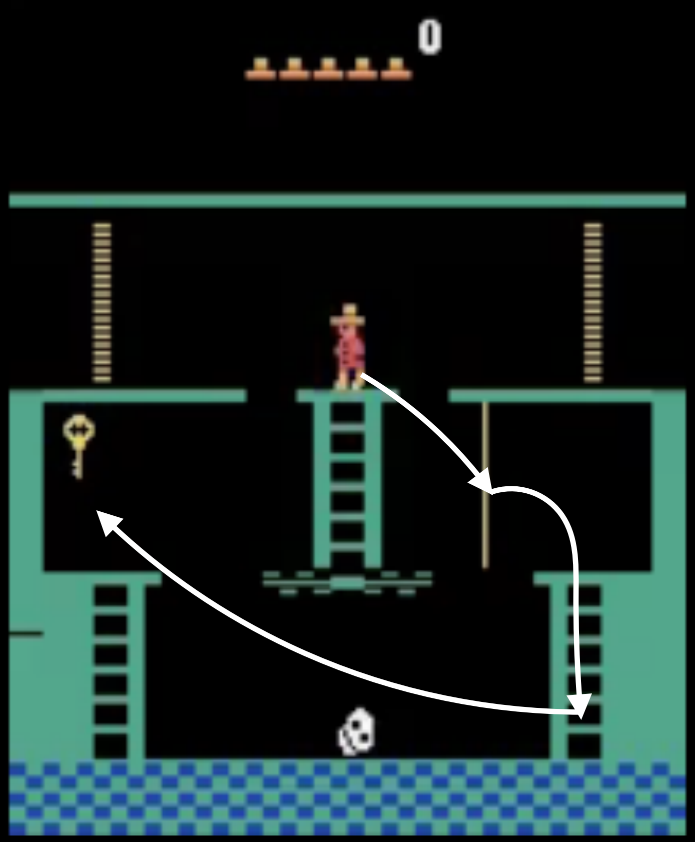
Curiosity Driven Exploration by Self-Supervised Prediction

ICML 2017

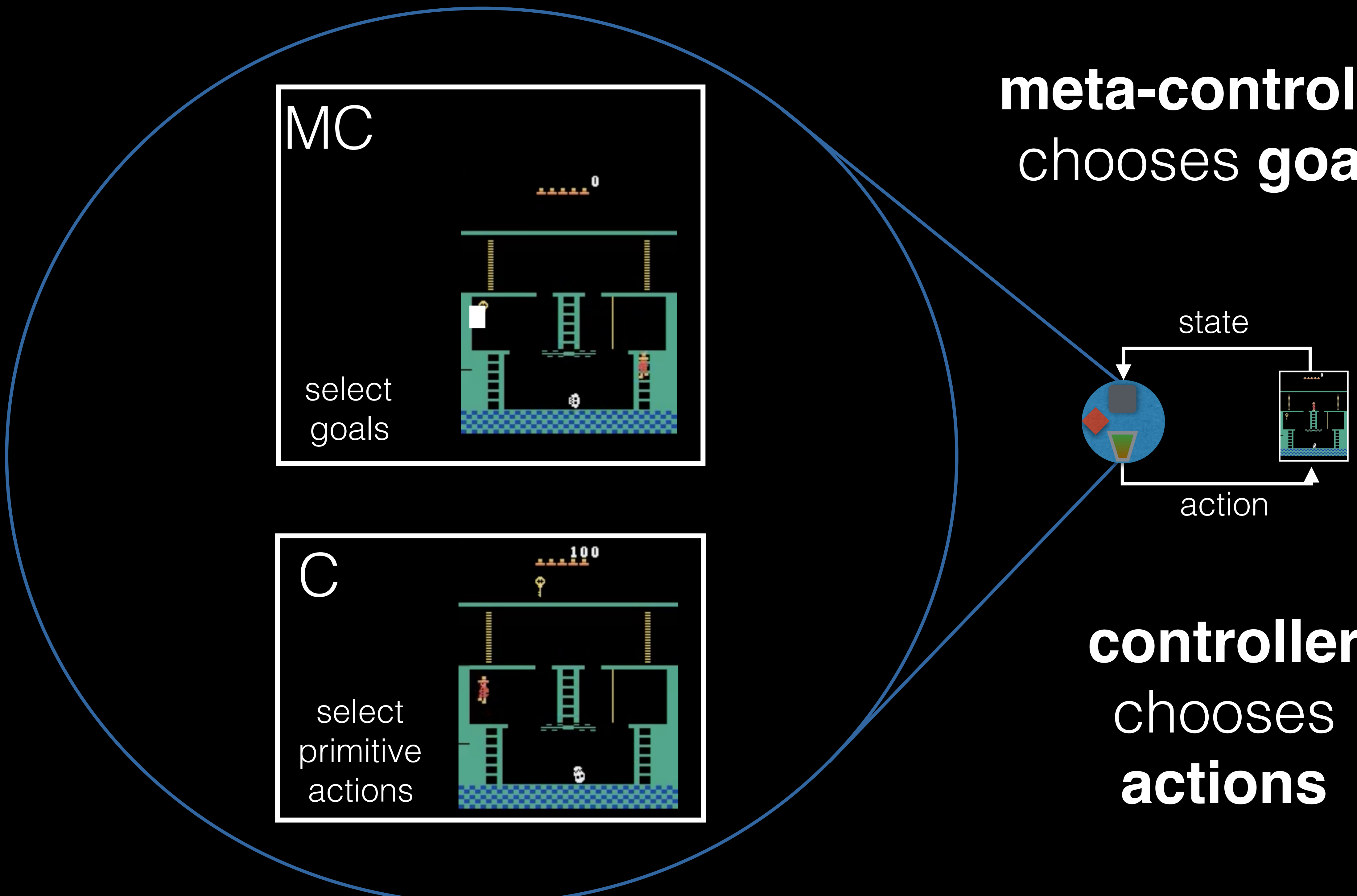
Deepak Pathak, Pulkit Agrawal, Alexei Efros, Trevor Darrell
UC Berkeley

https://github.com/pathak22/noreward-rl
https://pathak22.github.io/noreward-rl/

Temporal Abstractions



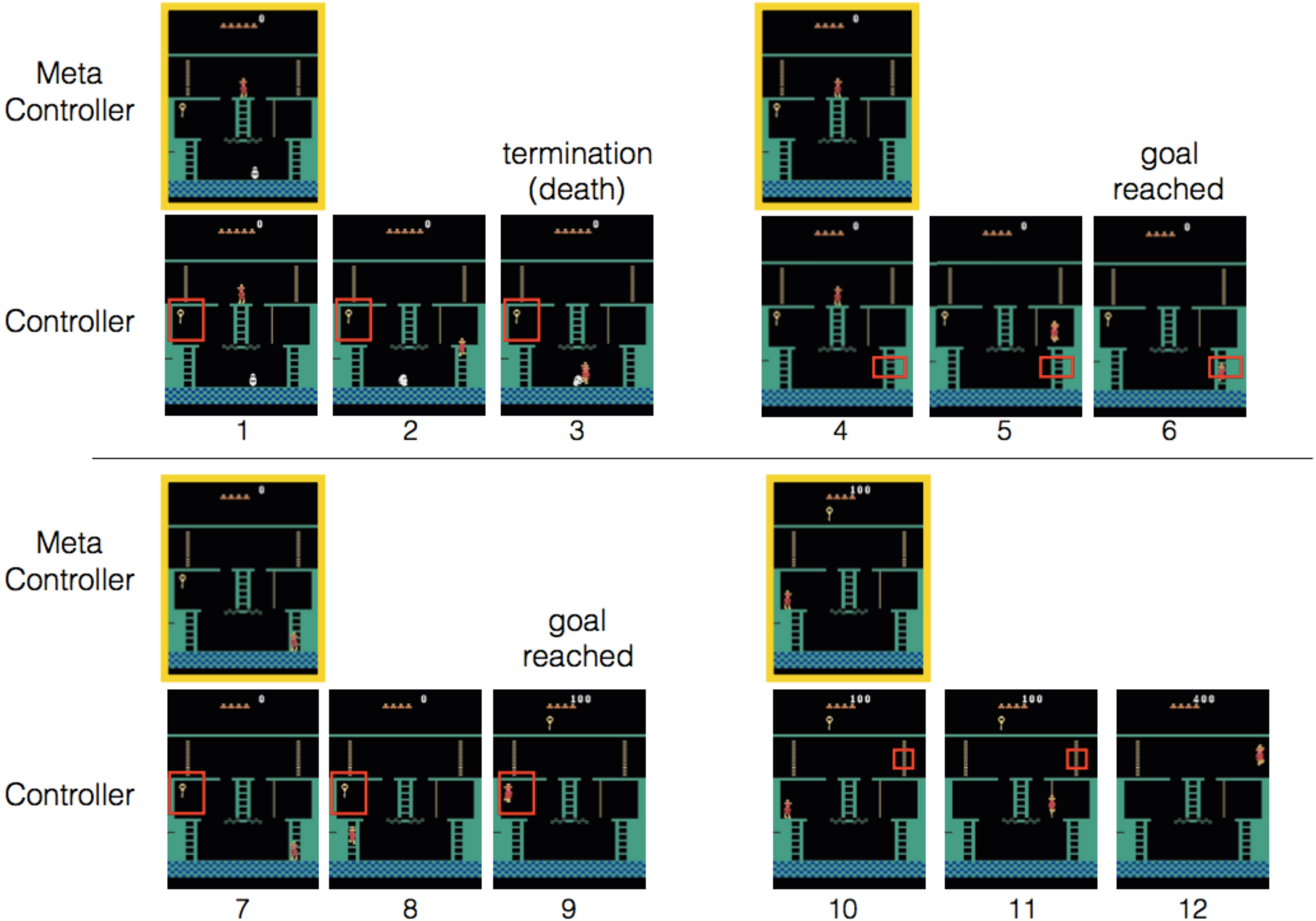
HRL with pre-set Goals



meta-controller chooses **goals**

controller chooses **actions**

Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation, T. D. Kulkarni, K. R. Narasimhan et. al. NIPS 2016

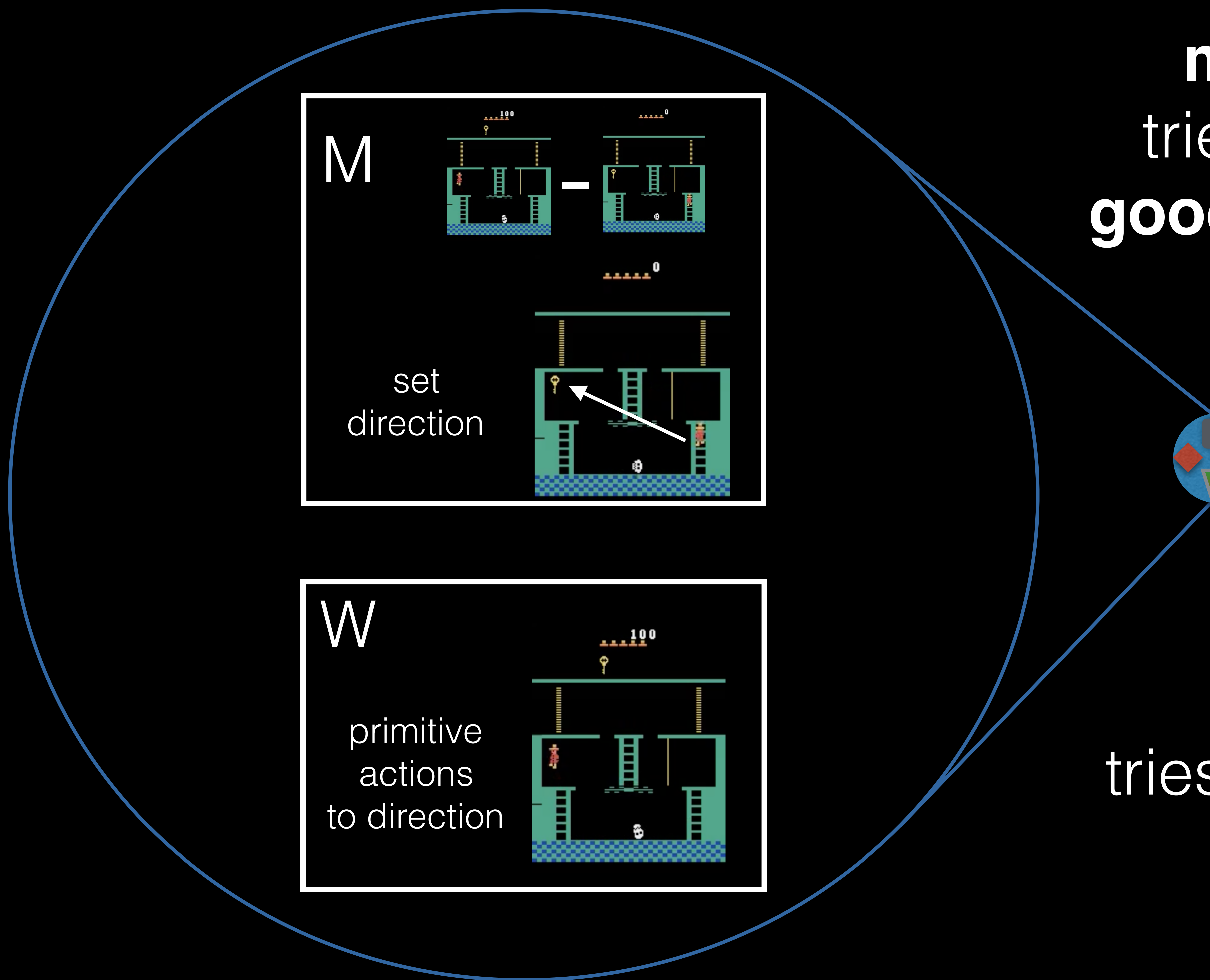


Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation, T. D. Kulkarni, K. R. Narasimhan et. al. NIPS 2016

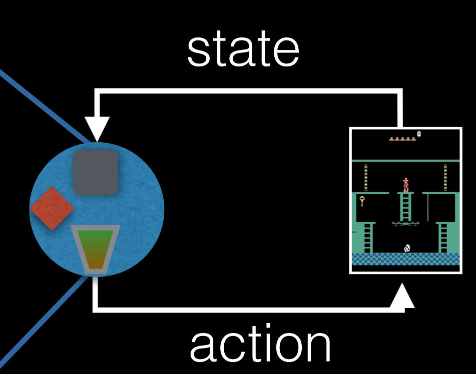
pre-defined goal
selected by
meta-controller

Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation, T. D. Kulkarni, K. R. Narasimhan et. al. NIPS 2016

Fe**U**dal **N**etworks for HRL

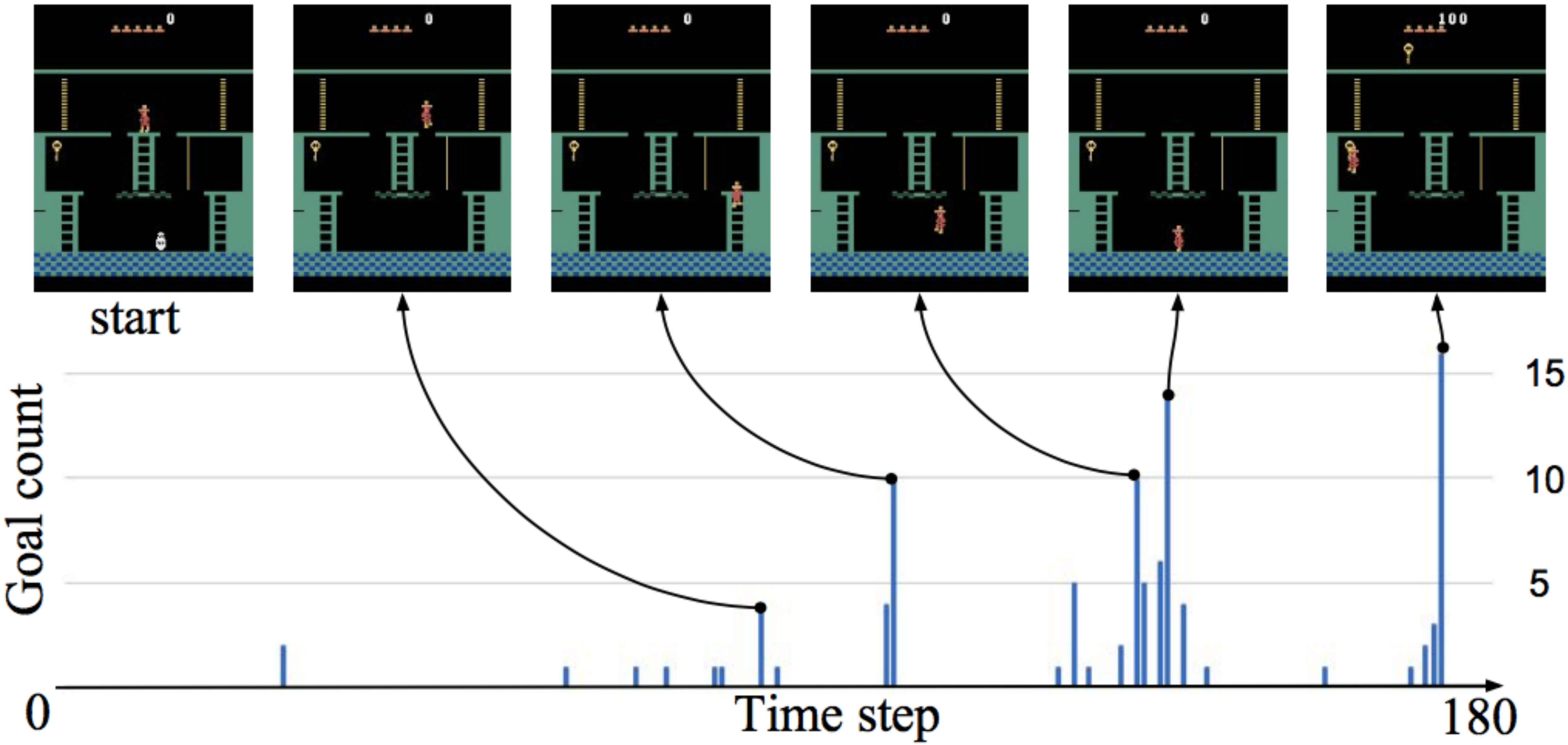


manager
tries to finds
good directions



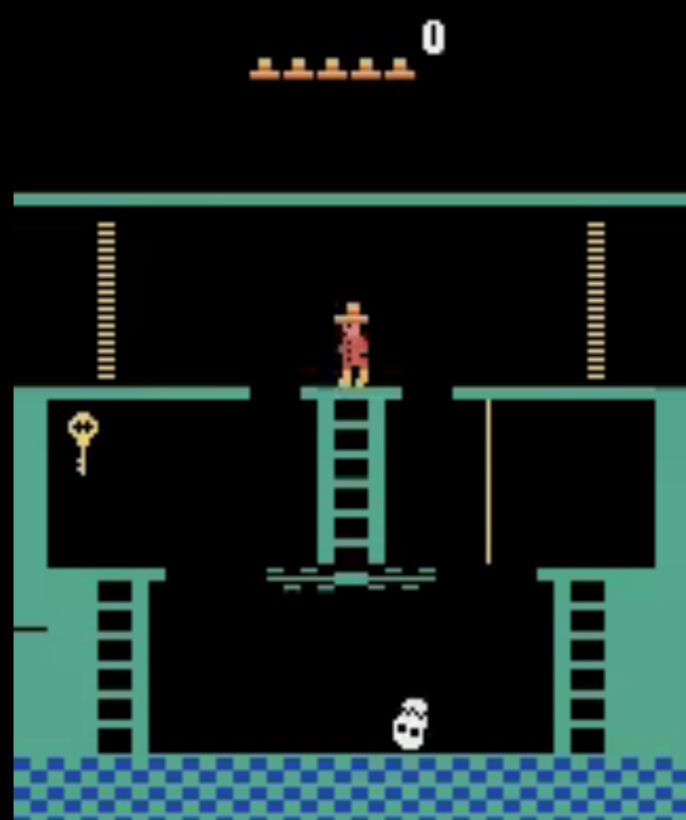
worker
tries to **achieve**
them

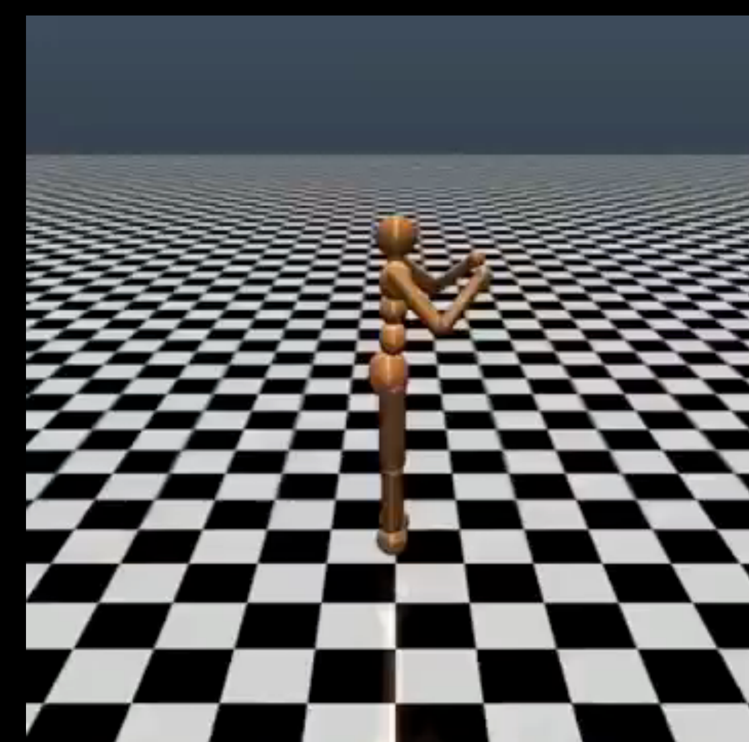
FeUdal Networks for Hierarchical Reinforcement Learning, Vezhnevets et. al. ICML 2017



FeUdal Networks for Hierarchical Reinforcement Learning, Vezhnevets et. al. ICML 2017

Generalisation





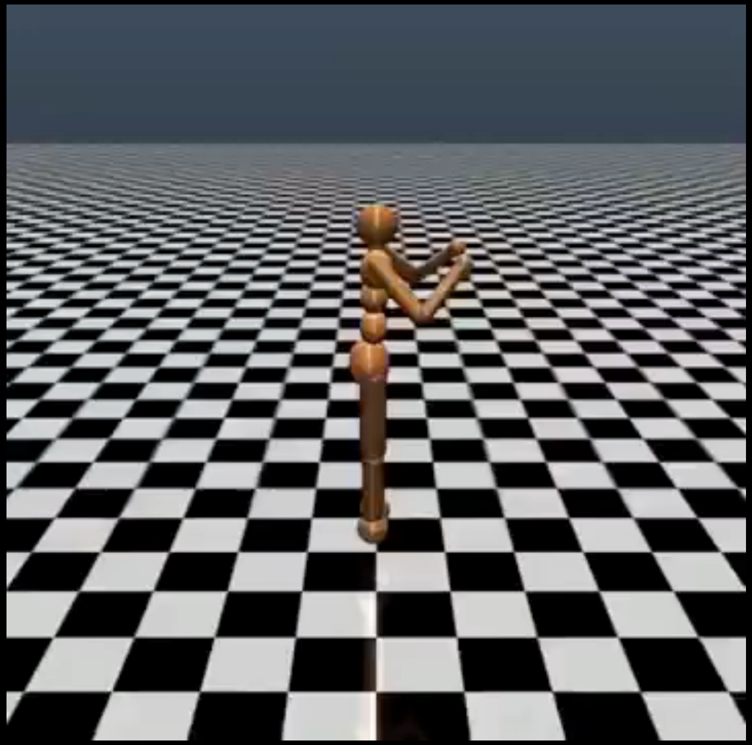
Meta-learning (Learn to Learn)

Versatile agents!

Transfer learning works with images

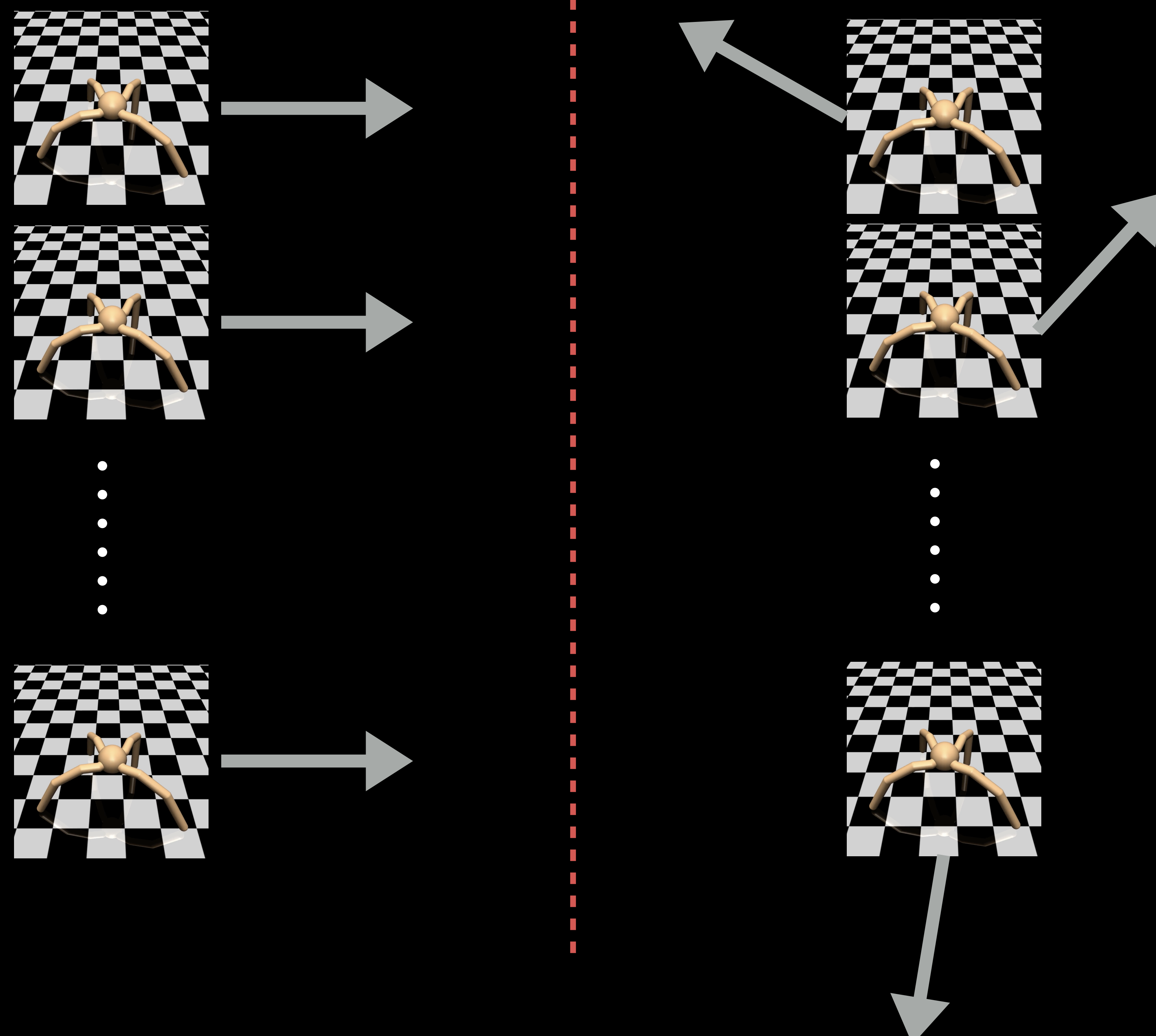
http://www.derinogrenme.com/2015/07/29/makale-imagenet-large-scale-visual-recognition-challenge/

Good **features for decision making**?

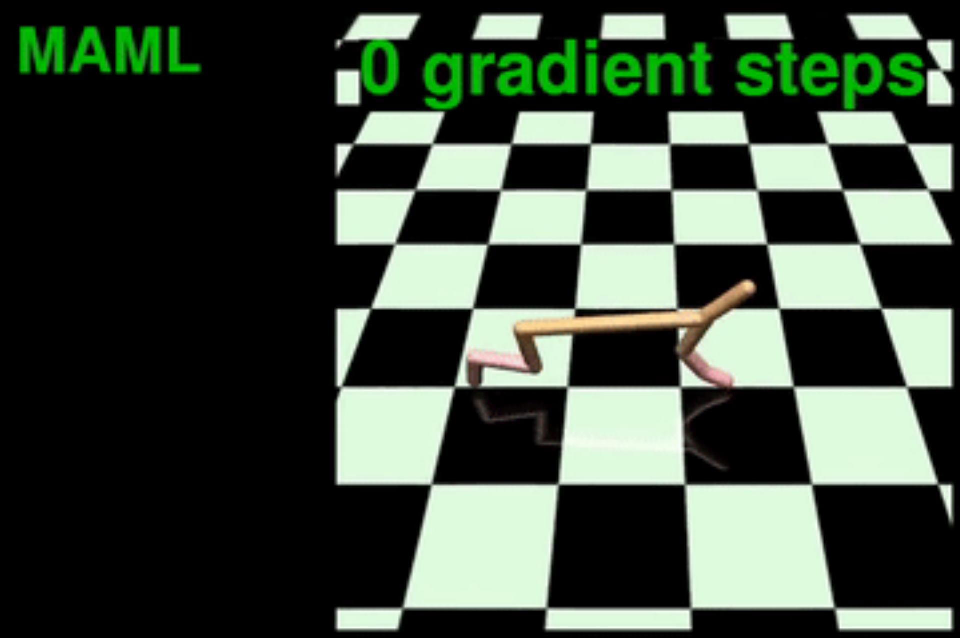


learn
to go East

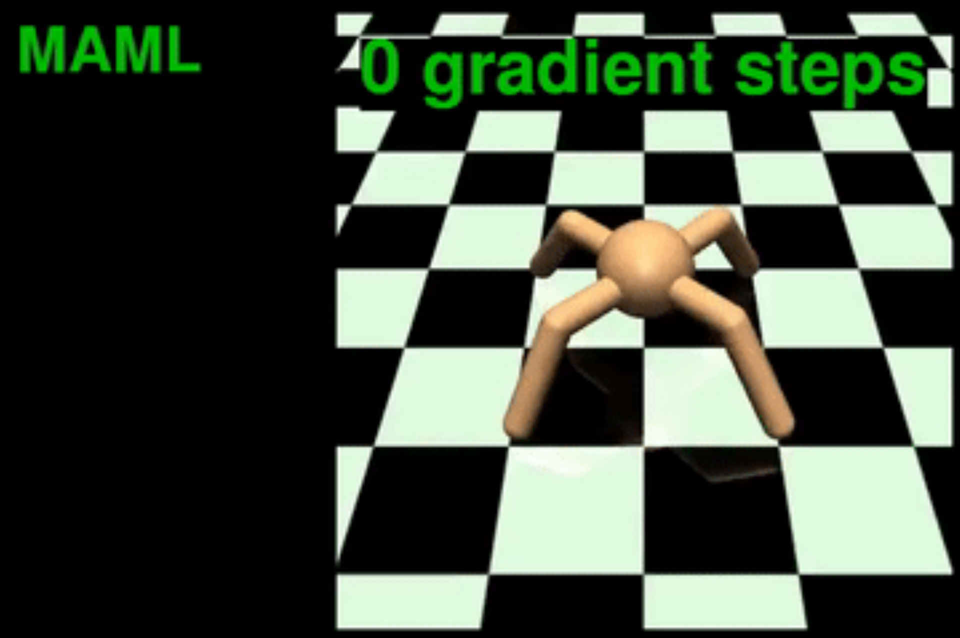
learn to
reduce learning
time to go to X



Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks.
C. Finn, P. Abbeel, S. Levine. ICML 2017.



0 grad/opt step: policy ready to learn



1 grad/opt step: learnt to achieve goal

http://bair.berkeley.edu/blog/2017/07/18/learning-to-learn/

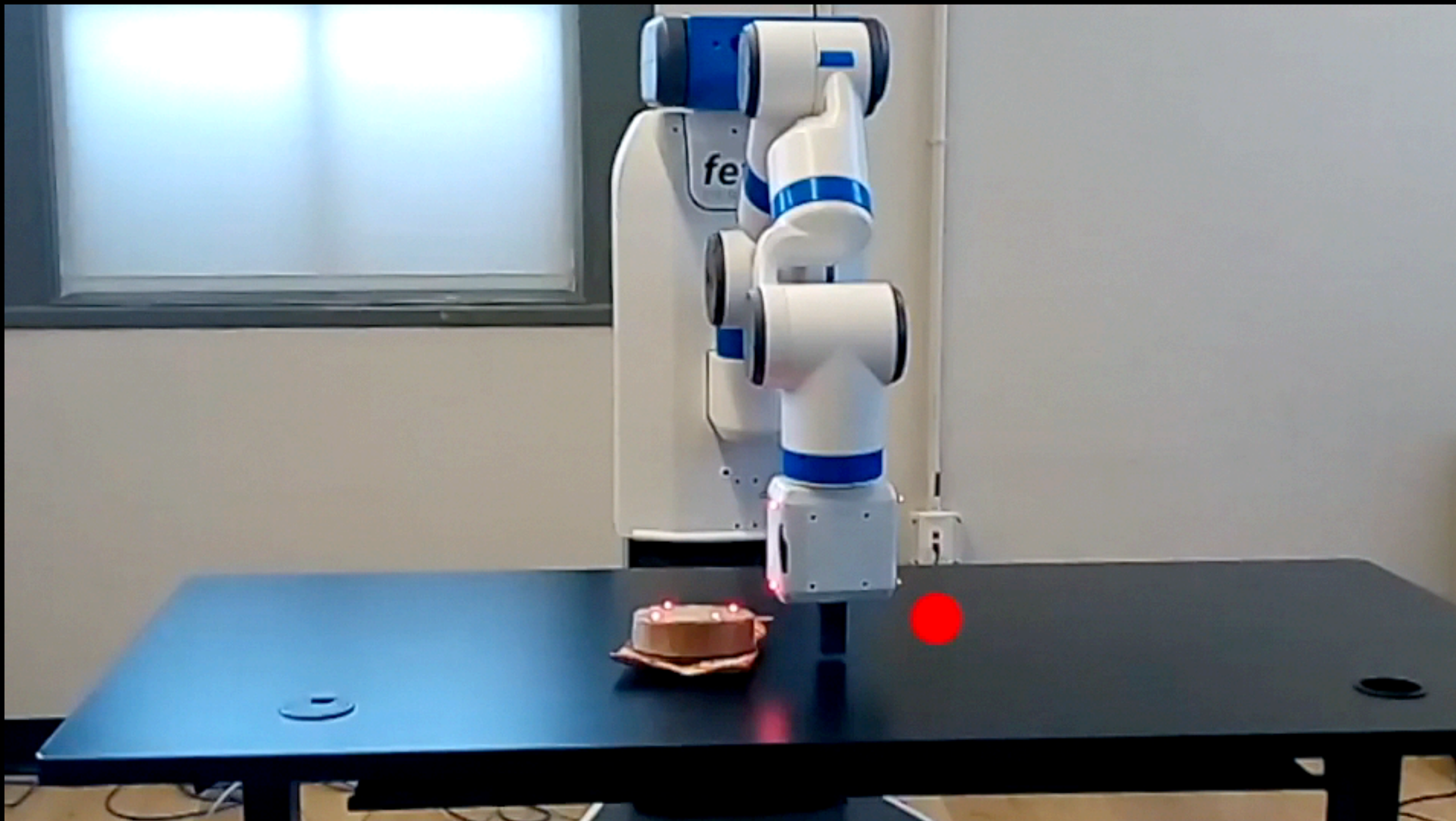
Code: https://github.com/cbfinn/maml_rl

Videos: https://sites.google.com/view/maml

Domain Randomisation

Generalising from Simulation

Sim-to-Real Transfer of Robotic Control with Dynamics Randomization, Peng et al. arXiv preprint, **18 Oct 2017**

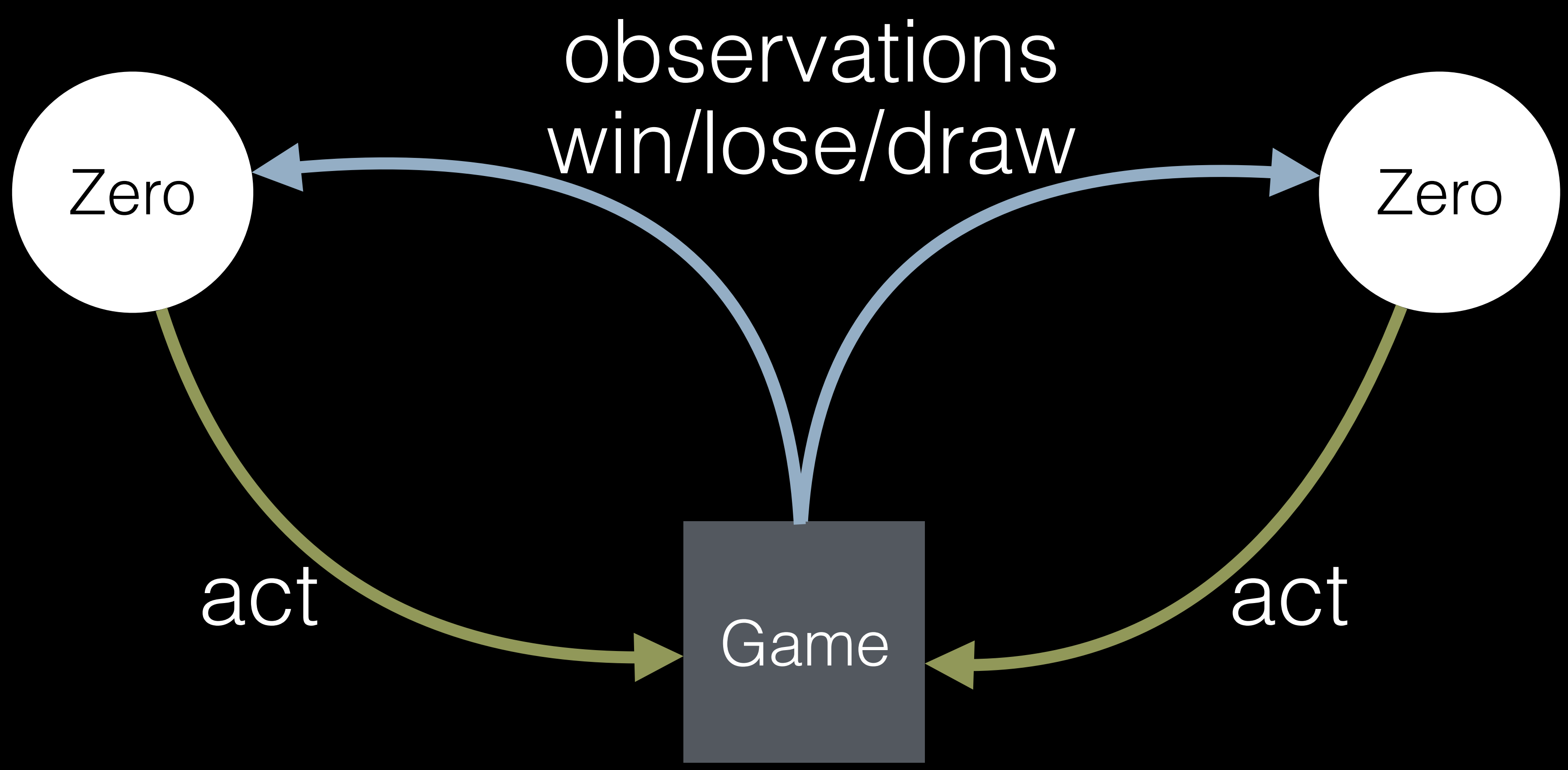


https://blog.openai.com/generalizing-from-simulation/

Generalisation via Self-play

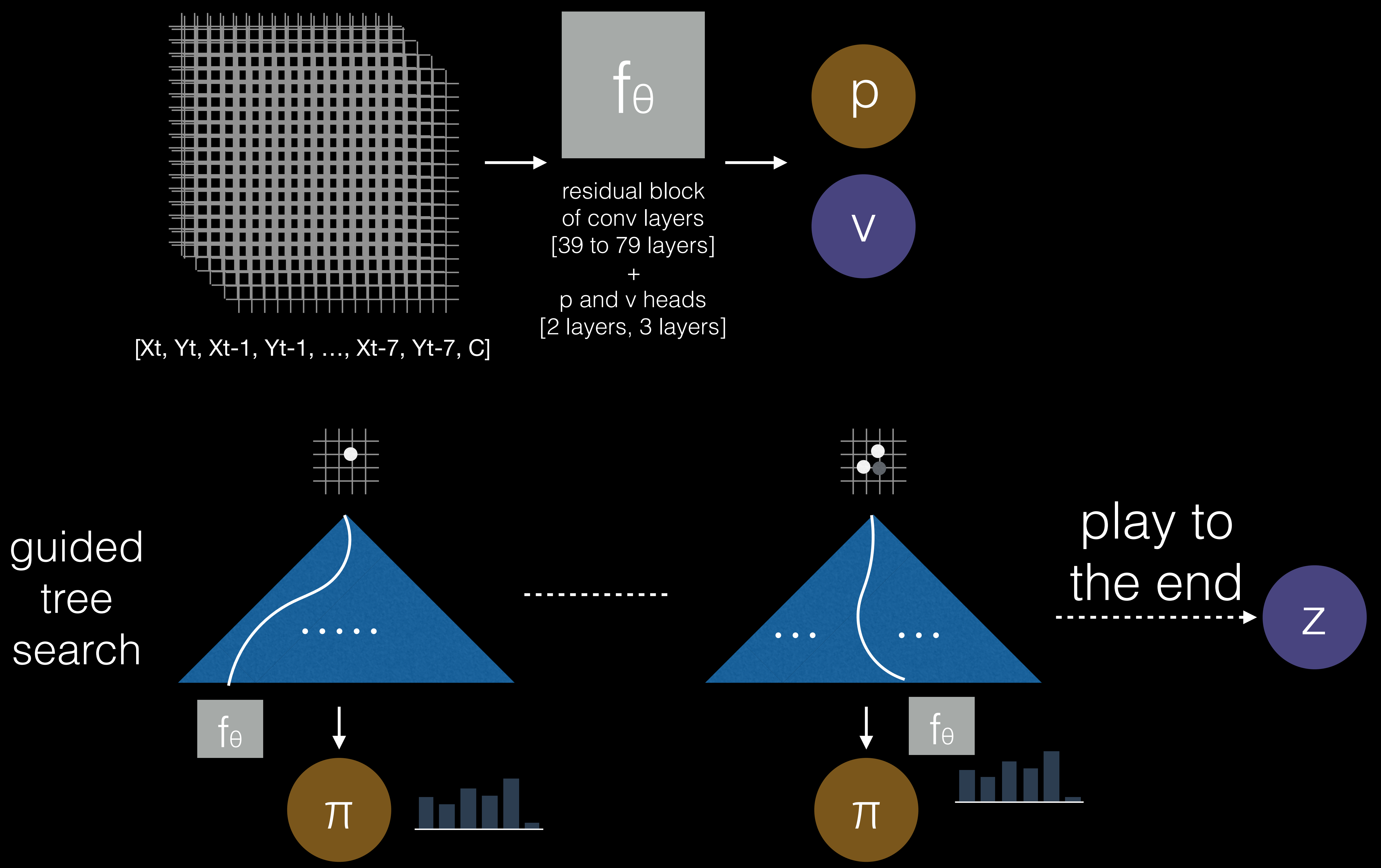
Deep RL in **AlphaGo Zero**

Improve
thinking and **intuition**
with **feedback from self-play**
[**zero** human game data]

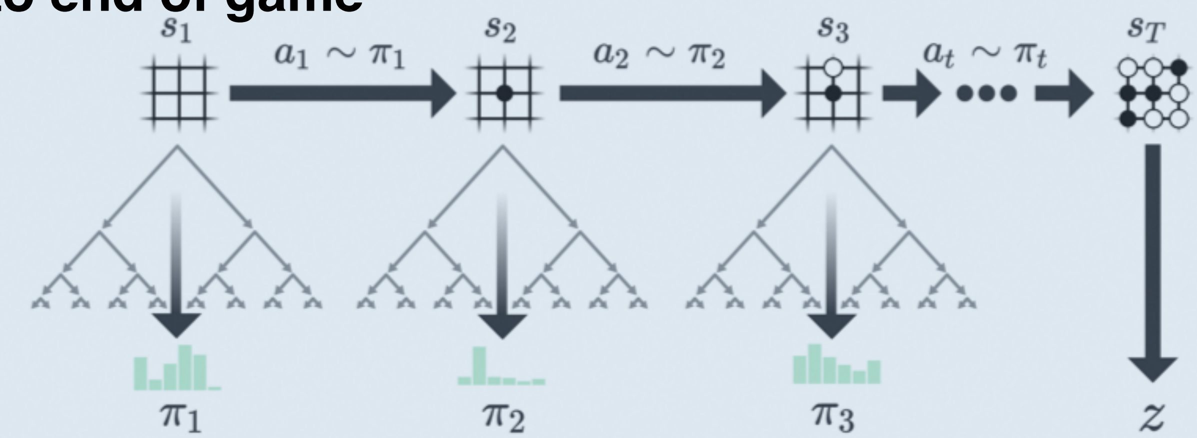


Mastering the game of Go without human knowledge, Silver et.al., Nature, Vol. 550, **October 19, 2017**

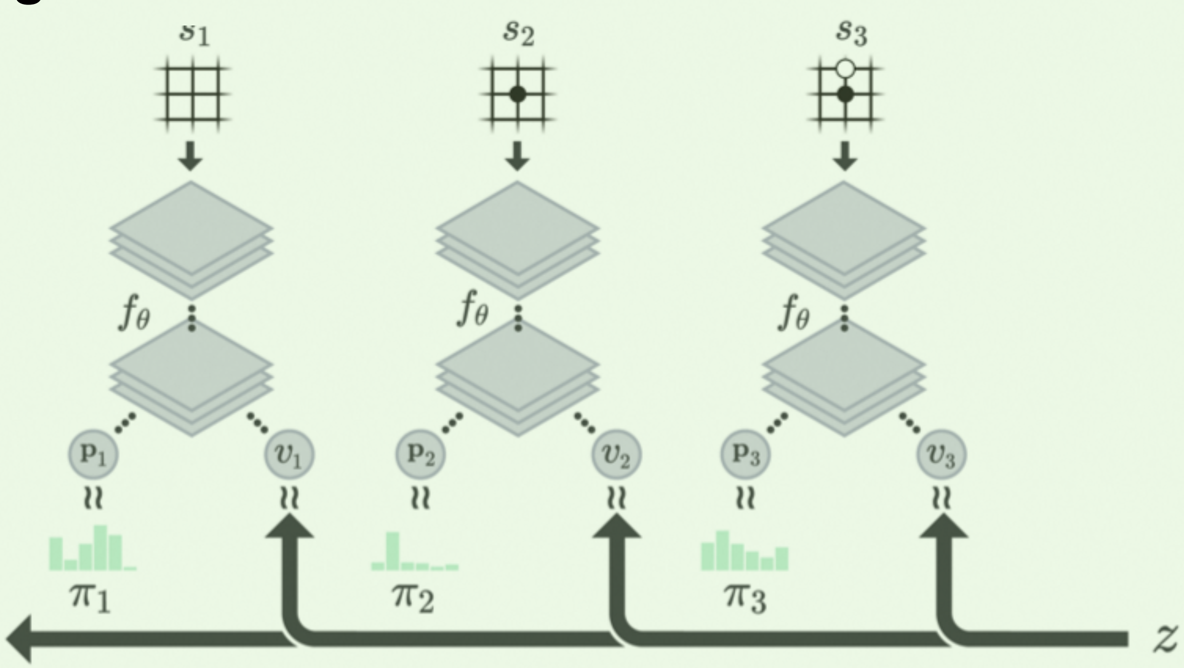
Very High Level Mechanics



Self-play to end of game

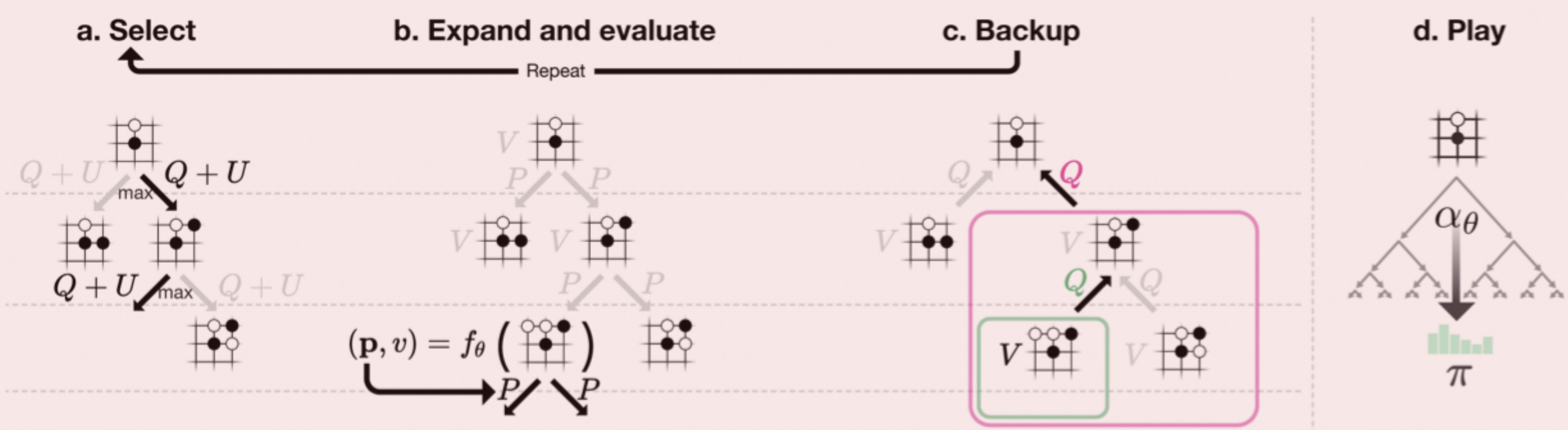


NN training: learn to evaluate

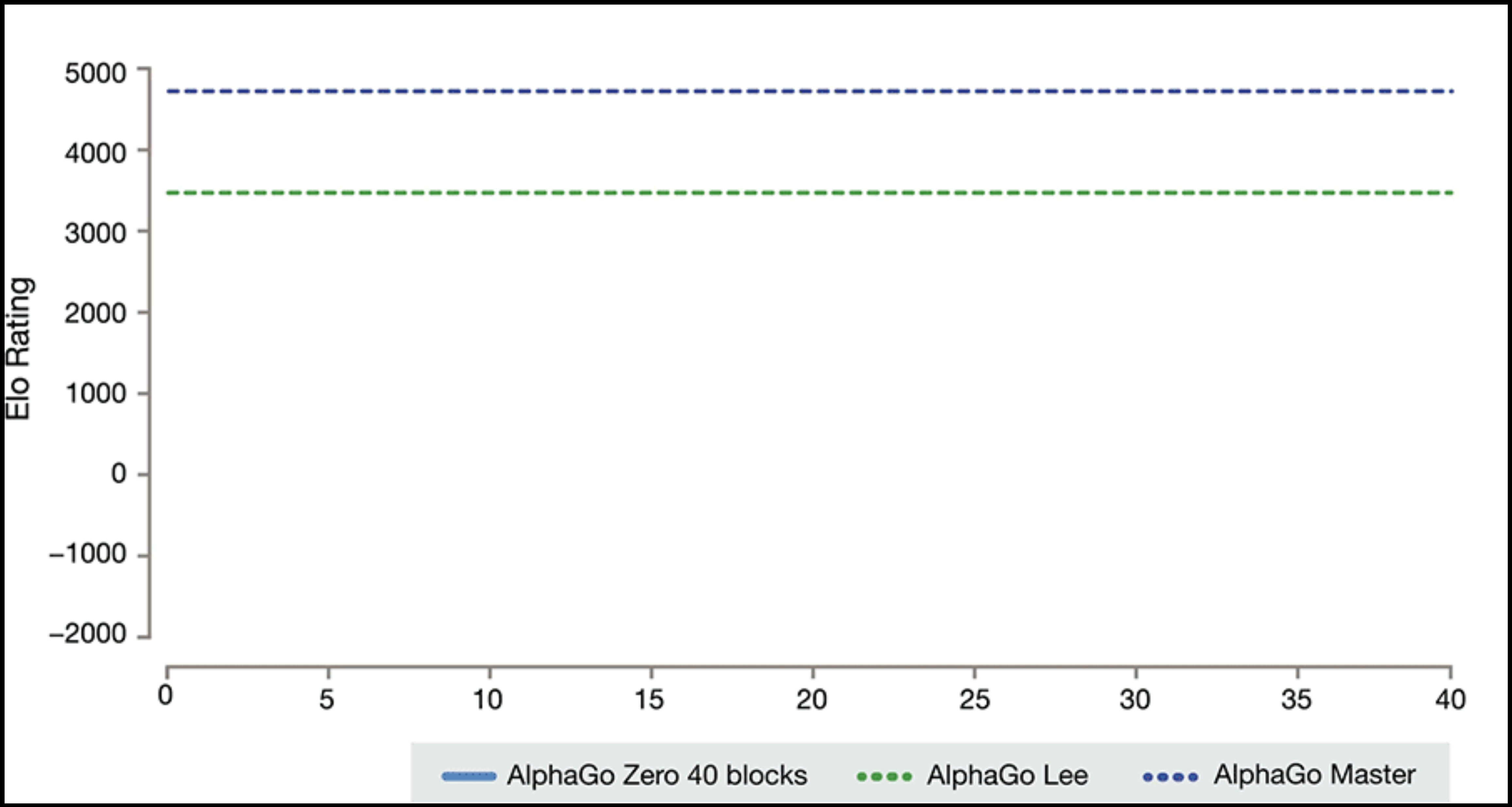


$$l = (z - v)^2 - \boldsymbol{\pi}^{\mathrm{T}} \log \boldsymbol{p} + c\|\theta\|^2$$

Self-play step: select move by simulation + evaluation



Mastering the game of Go without human knowledge, Silver et.al., Nature, Vol. 550, **October 19, 2017**



https://deepmind.com/blog/alphago-zero-learning-scratch/

https://deepmind.com/blog/alphago-zero-learning-scratch/
https://www.youtube.com/watch?v=WXHFqTvfFSw

Inspired to study RL much?

Next lecture:
Building Blocks of (Deep) RL
November 8, 2017

https://join.slack.com/t/deep-rl-tutorial/signup