CS 295: Optimal Control and Reinforcement Learning

Winter 2020

Lecture 2: Imitation Learning

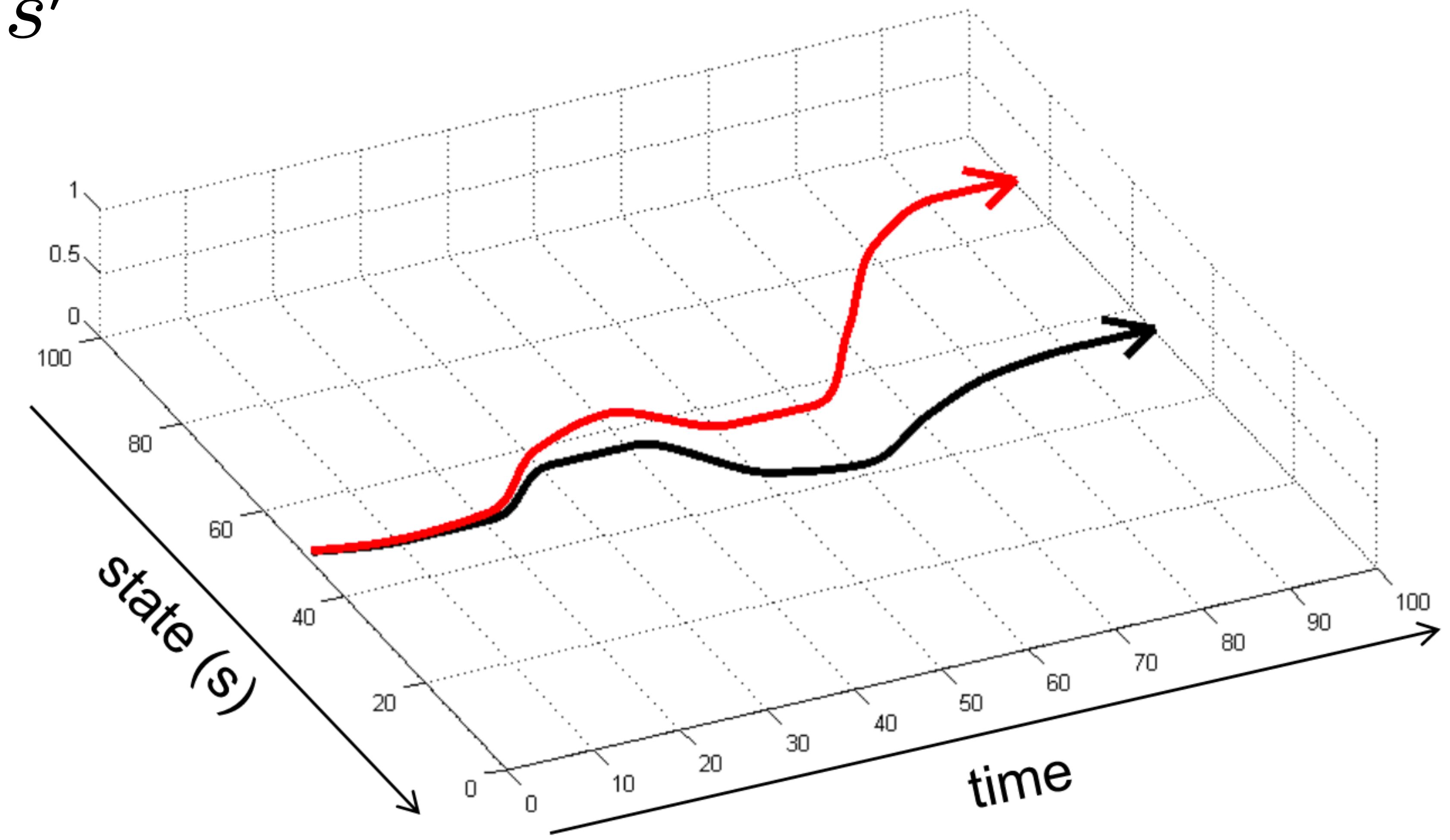
Roy Fox
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine

Today's lecture

- Behavior Cloning
- Modeling humans
- DAgger
- DART
- HVIL

The impact of inaccurate dynamics

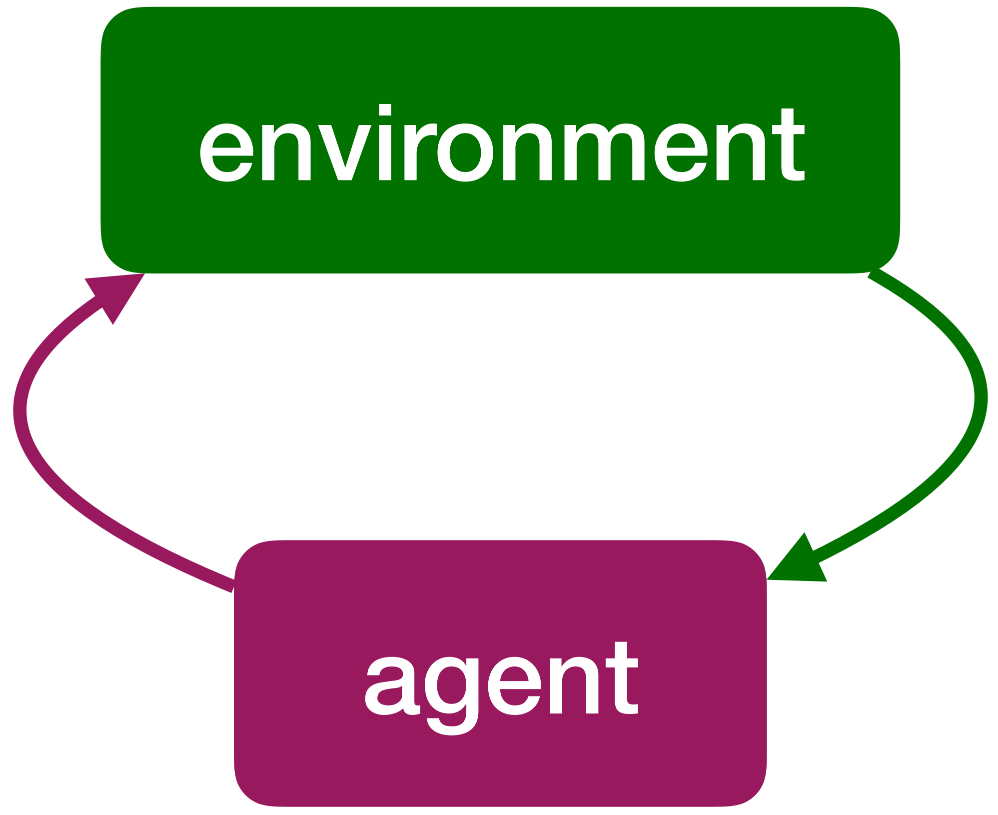
$$\sum_{s'} |p^1(s'|s) - p^2(s'|s)| \leqslant \epsilon$$

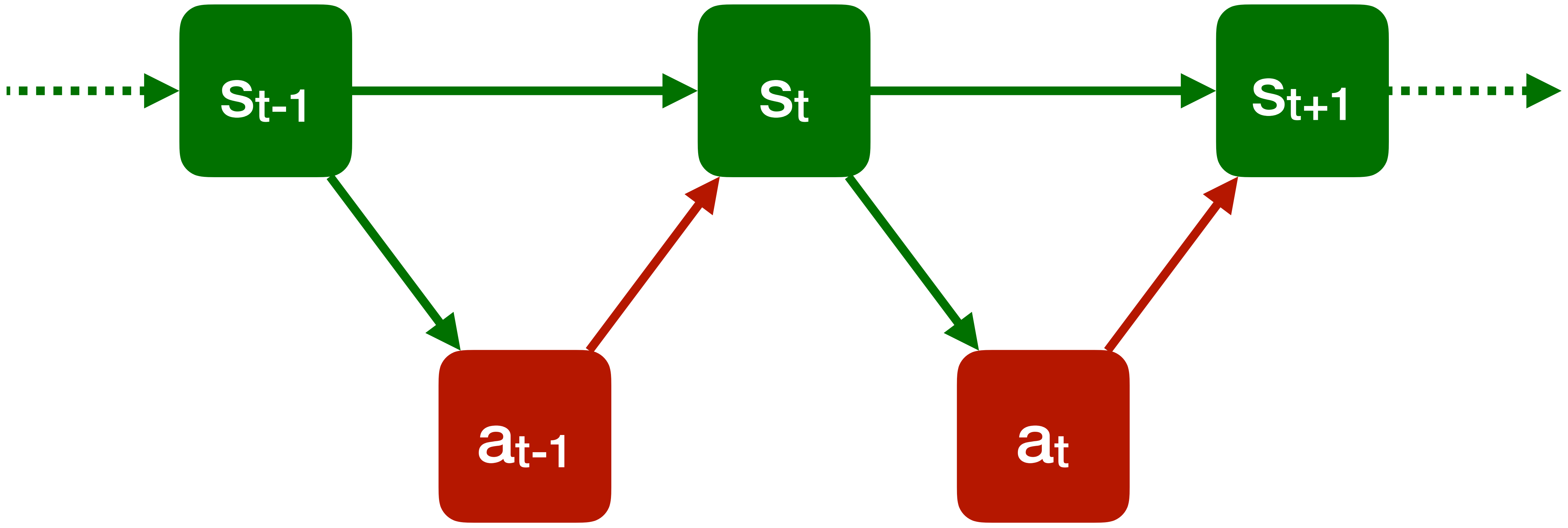


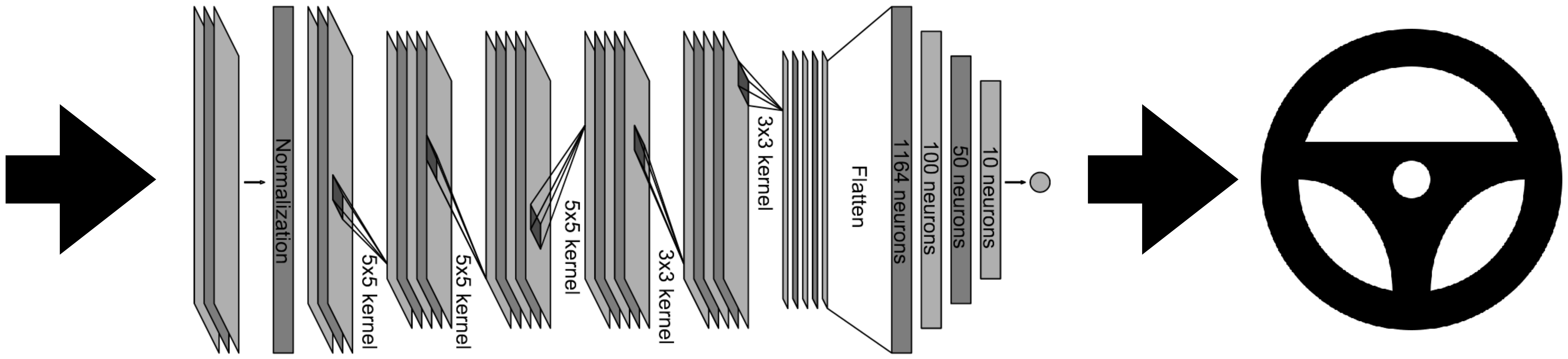
$$\sum_{s_t} |p^1(s_t) - p^2(s_t)| \leqslant \epsilon t$$

Image: Sergey Levine

A policy is a (stochastic) function



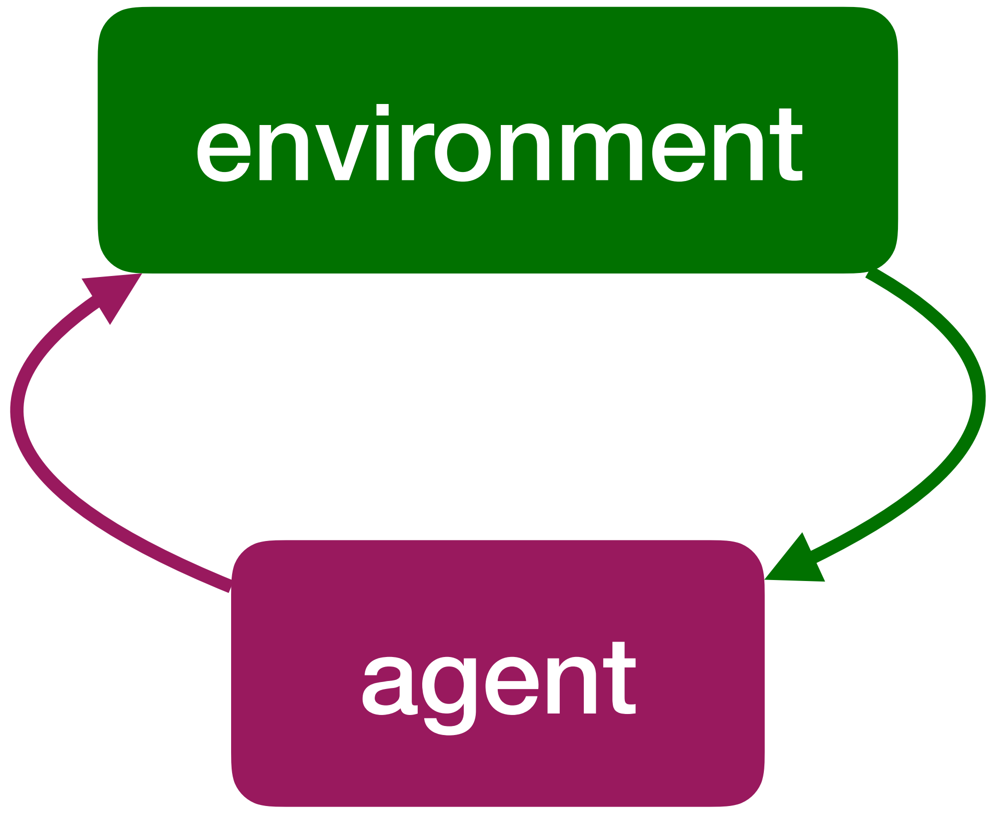


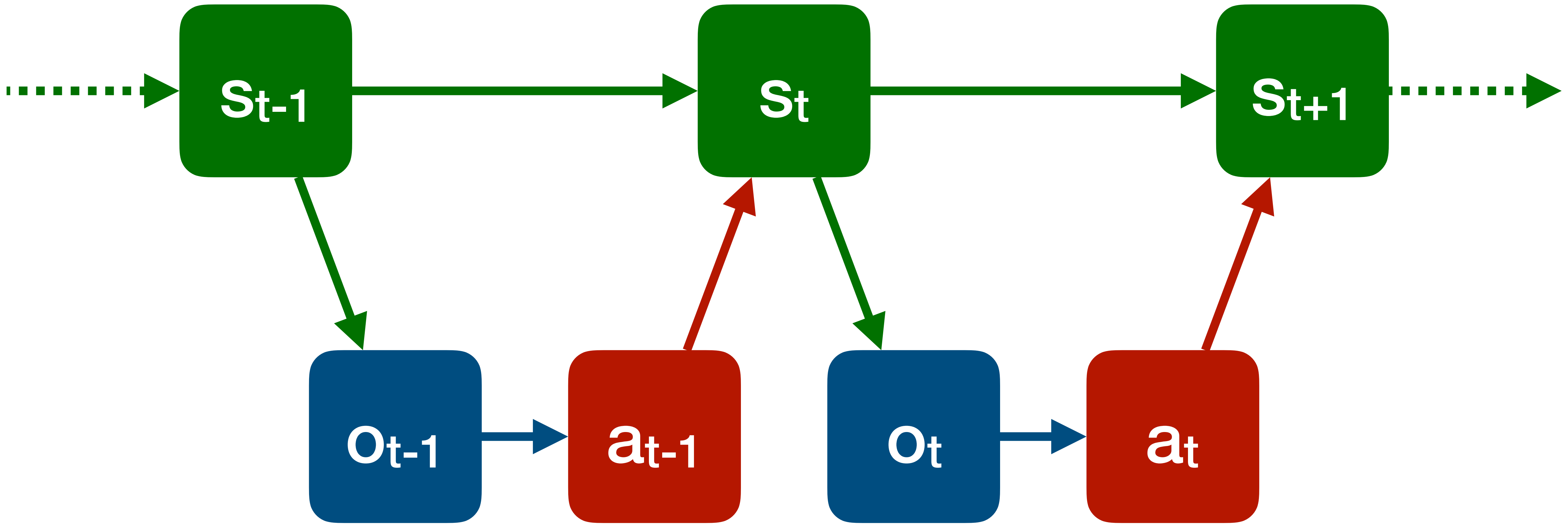


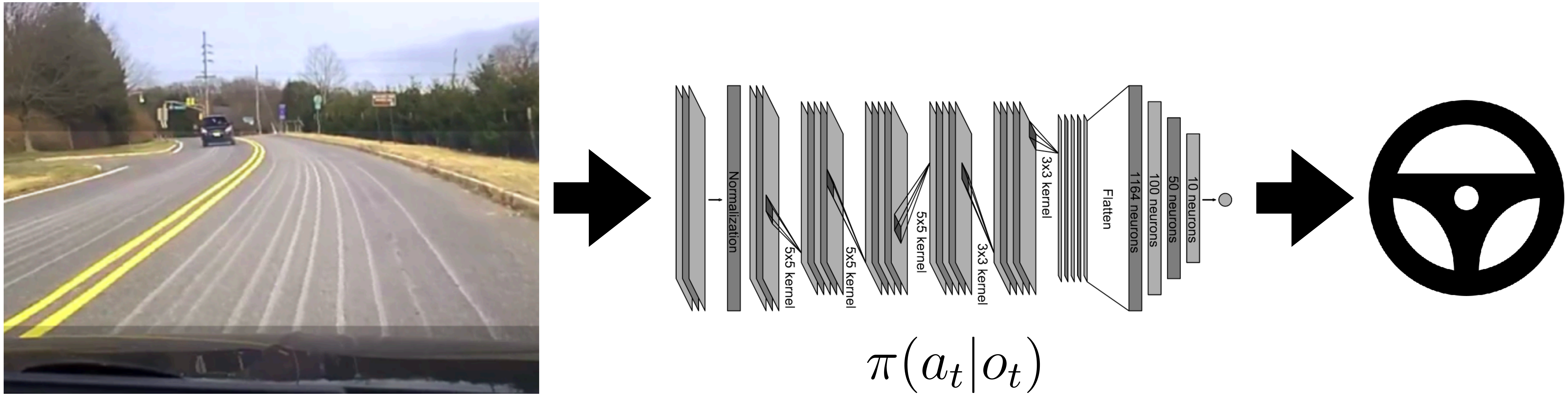
$$\pi(a_t|s_t)$$

Image: Bojarski et al. 2016

A policy is a (stochastic) function



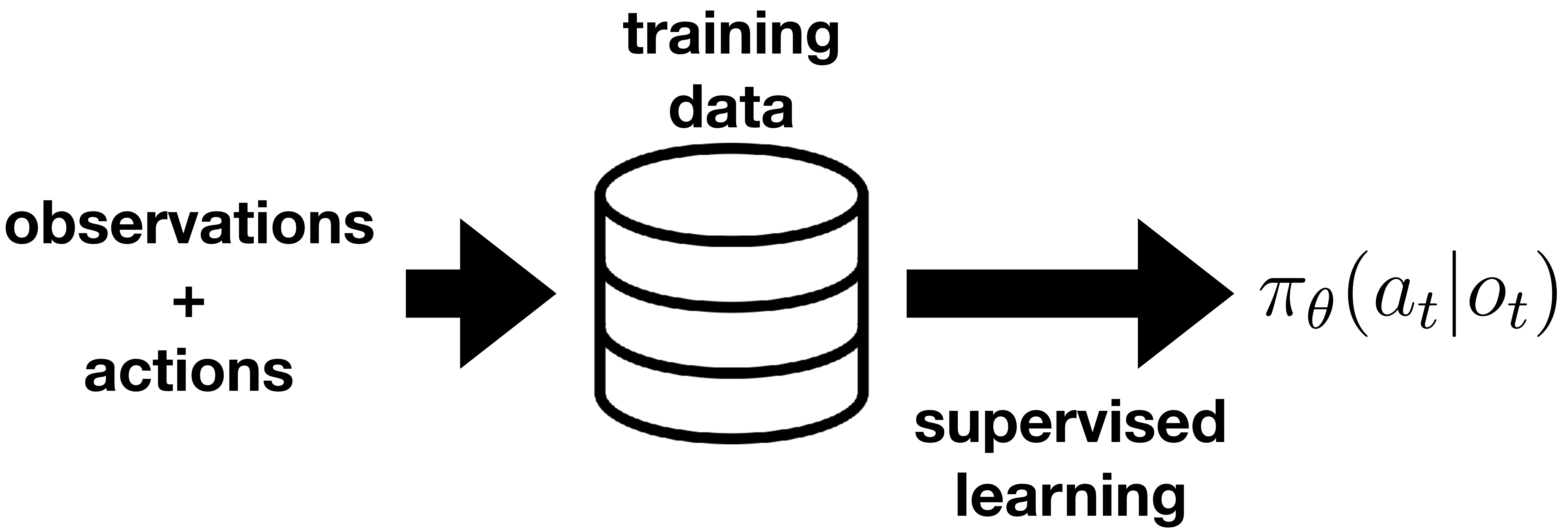




$$\pi(a_t|o_t)$$

observation **action**

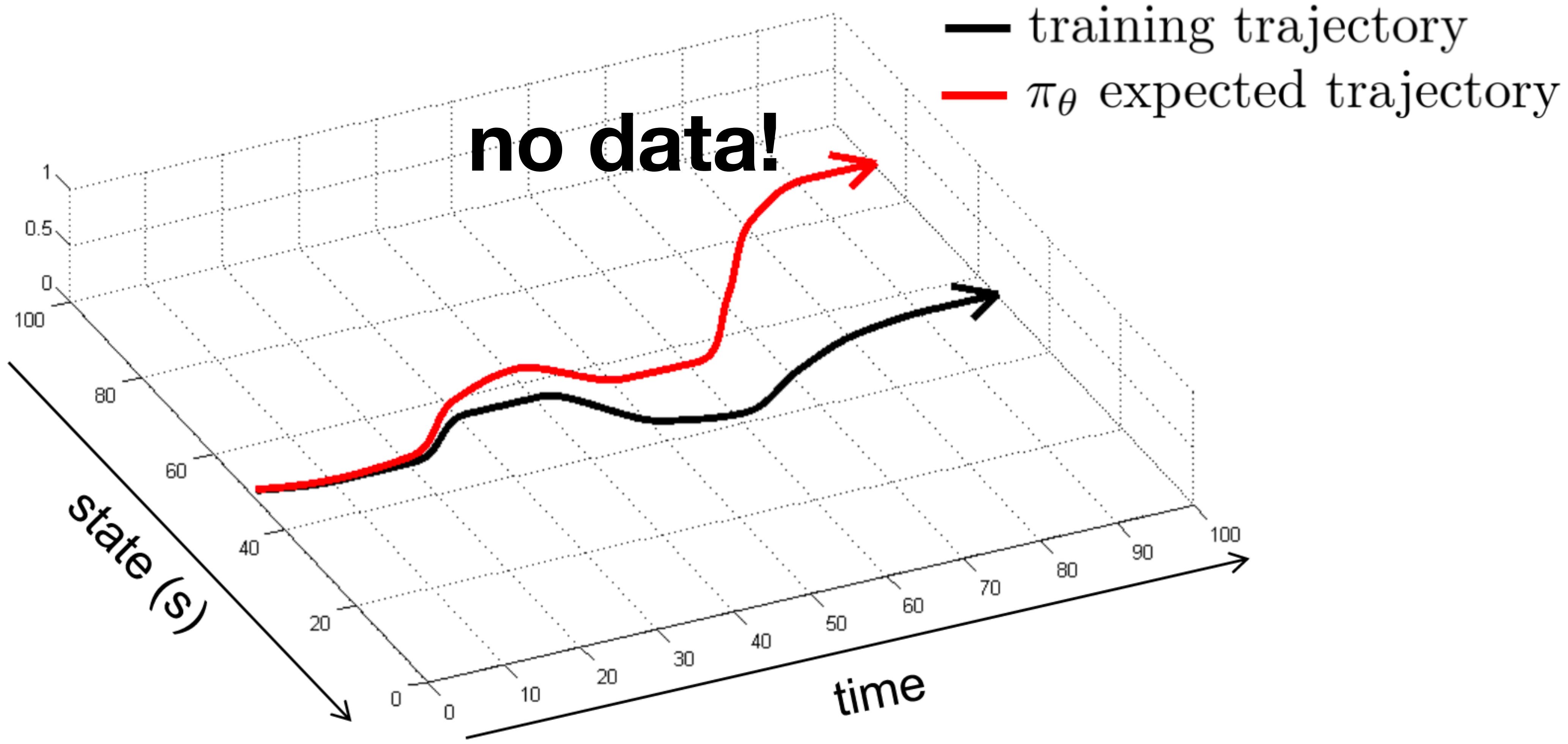
Behavior Cloning



$$p_\pi(s_{t+1}|s_t) = \sum_{o_t,a_t} p(o_t|s_t)\pi(a_t|o_t)p(s_{t+1}|s_t,a_t)$$

$$\pi_\theta(a_t|o_t) \approx \pi^*(a_t|o_t)$$

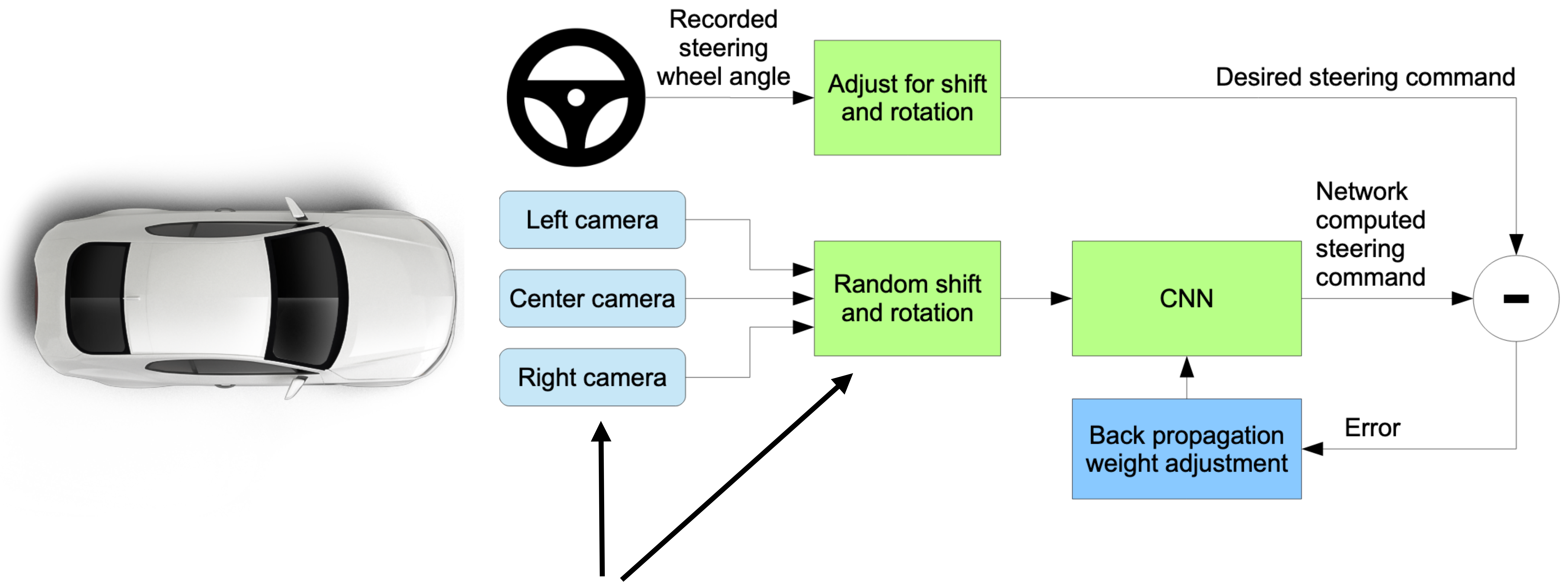
$$p_{\pi_\theta}(s_{t+1}|s_t) \approx p_{\pi^*}(s_{t+1}|s_t)$$



But wait...

Video: NVIDIA

How did they do it?



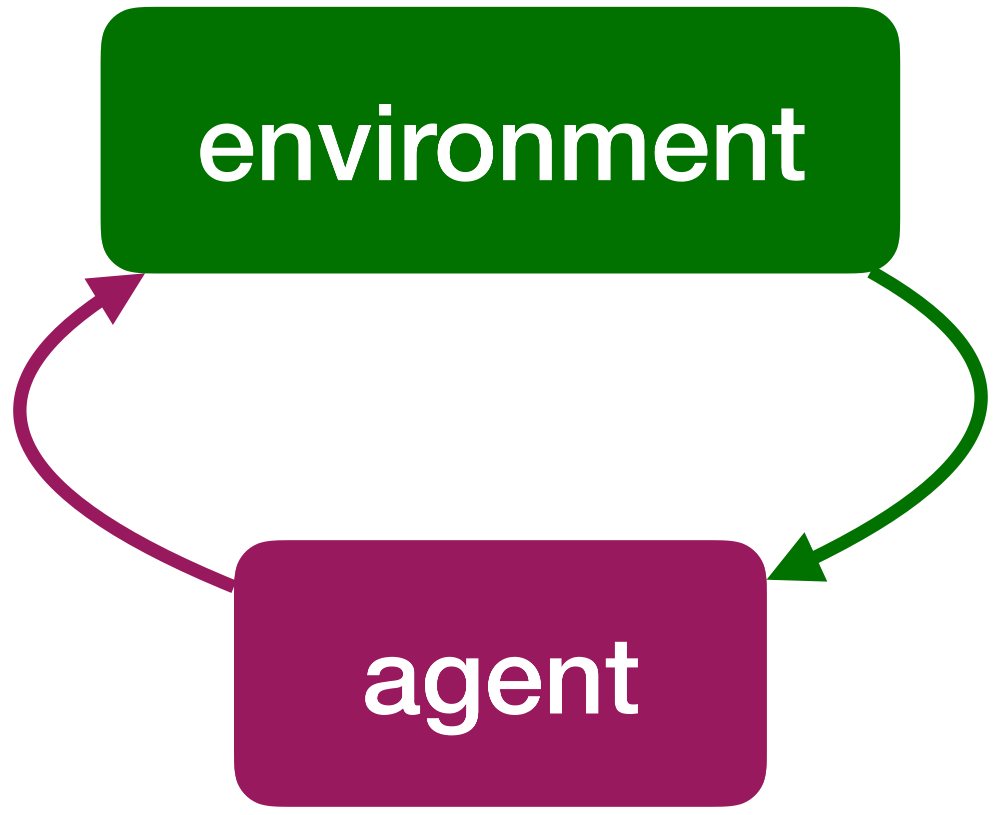
augmented data to better cover test distribution

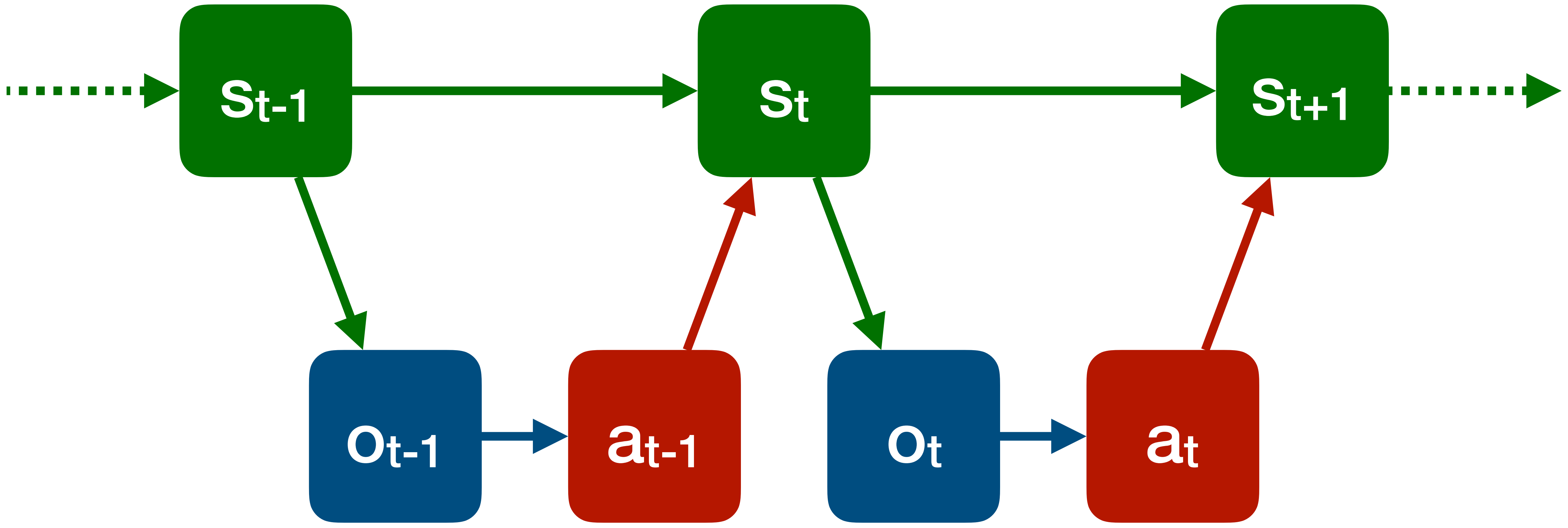
Image: Bojarski et al. 2016

Modeling humans is hard

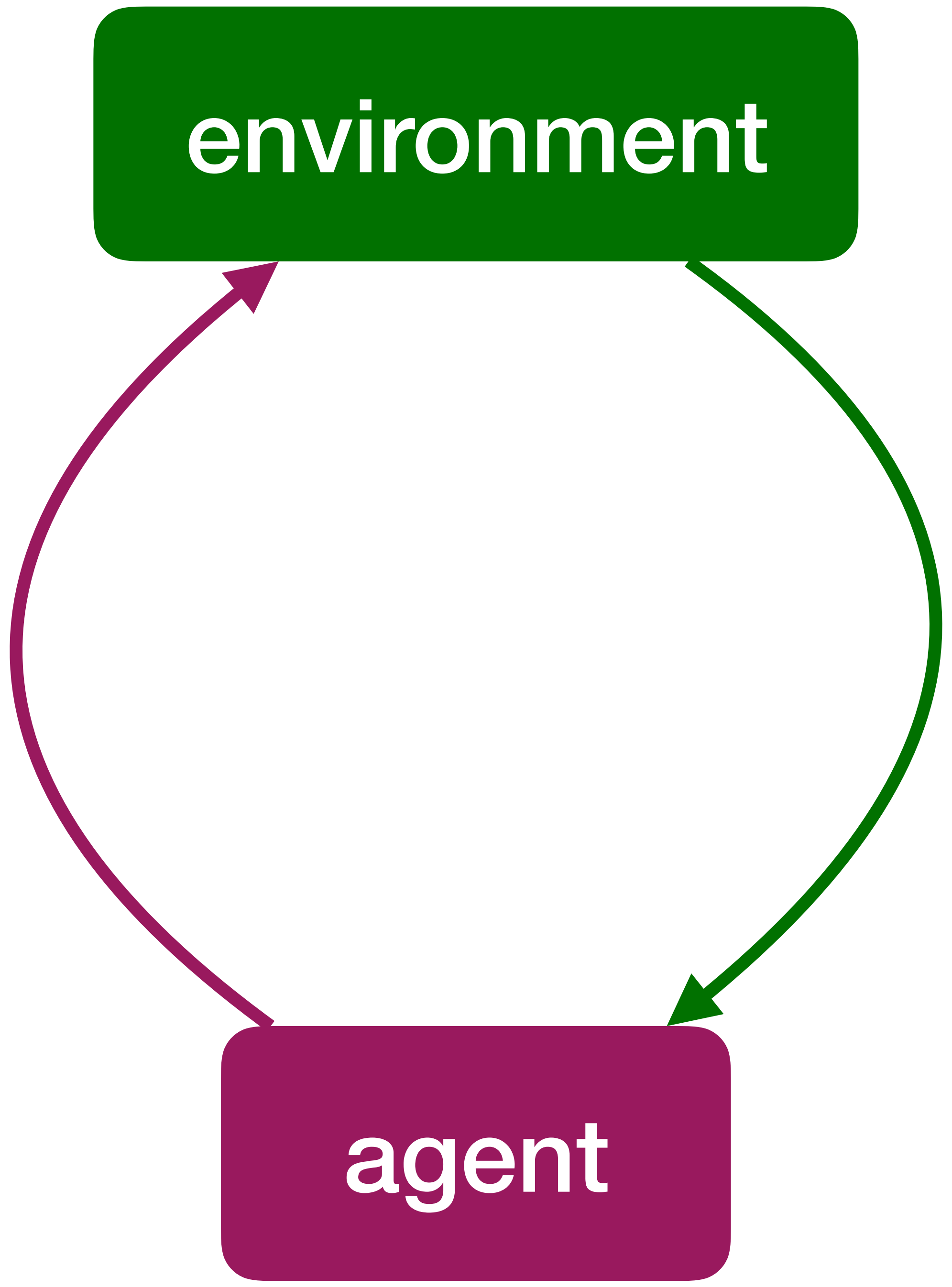
- Perhaps $o_t \neq o^*t$
- Perhaps $o_t \neq s_t$, so $p(o_{t+1}|o_t, a_t) \neq p(o_{t+1}|o_0, a_0, \ldots, o_t, a_t)$
  - Generally, this requires $\pi_\theta(a_t|o_0, a_0, \ldots, o_t)$
  - Can use RNN, other models
  - Modeling memory is hard → prior structure may help
- Perhaps there is insufficient data
  - Demonstrating is a burden!
- Perhaps demonstrations are inconsistent
  - Humans are fallible
  - Some supervision is hard to give

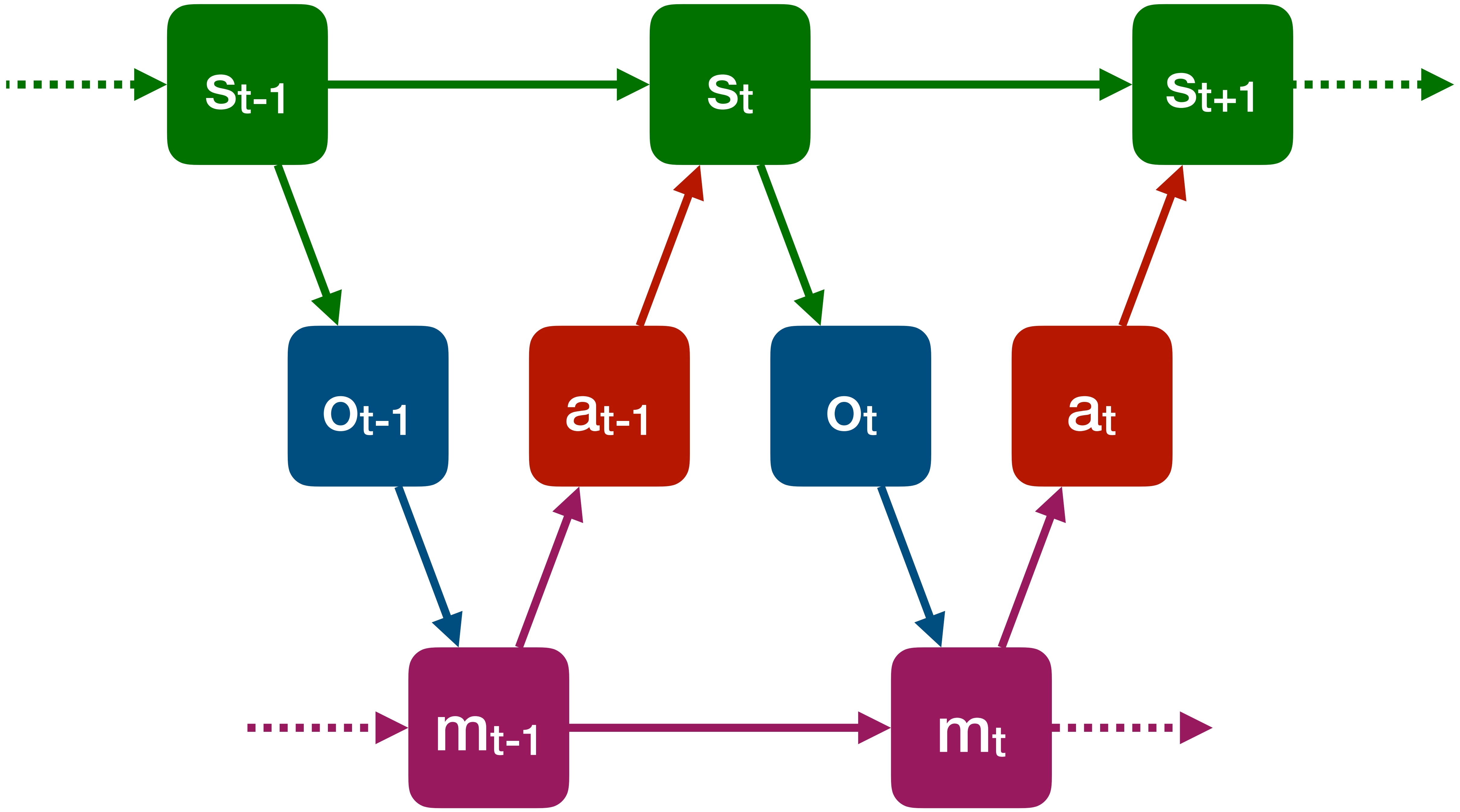
Modeling memory





Modeling memory





$$\pi_\theta(m_t, a_t | m_{t-1}, o_t)$$

DAgger: Dataset Aggregation

Can we collect demonstration data for $p_{\pi_\theta}(o_t)$?

Algorithm 1 DAgger

Collect dataset $\mathcal{D}$ of teacher demonstrations

$(o_0, a_0^*, o_1, a_1^*, \ldots) \sim p_{\pi^*}$

Train π_θ on $\mathcal{D}$

Execute π_θ to get $(o_0, a_0, \ldots) \sim p_{\pi_\theta}$

Ask teacher to label $a_t^* | o_t \sim \pi^*$

Aggregate $(o_0, a_0^*, o_1, a_1^*, \ldots)$ into $\mathcal{D}$

Repeat!

DAgger demo

Video: Stéphane Ross

DAgger: Dataset Aggregation

Can we collect demonstration data for $p_{\pi_\theta}(o_t)$?

Algorithm 1 DAgger

Collect dataset $\mathcal{D}$ of teacher demonstrations

$(o_0, a_0^*, o_1, a_1^*, \ldots) \sim p_{\pi^*}$

Train π_θ on $\mathcal{D}$

Execute π_θ to get $(o_0, a_0, \ldots) \sim p_{\pi_\theta}$

$\boxed{\text{Ask teacher to label } a_t^* | o_t \sim \pi^*}$ **but how?**

Aggregate $(o_0, a_0^*, o_1, a_1^*, \ldots)$ into $\mathcal{D}$

Repeat!

DAgger: Dataset Aggregation

Can we collect demonstration data for $p_{\pi_\theta}(o_t)$?

Algorithm 1 DAgger

Collect dataset $\mathcal{D}$ of teacher demonstrations

$(o_0, a_0^*, o_1, a_1^*, \ldots) \sim p_{\pi^*}$

Train π_θ on $\mathcal{D}$

Execute π_θ to get $(o_0, a_0, \ldots) \sim p_{\pi_\theta}$

$\boxed{\text{Ask teacher to label } a_t^* | o_t \sim \pi^*}$ **but how?**

Aggregate $(o_0, a_0^*, o_1, a_1^*, \ldots)$ into $\mathcal{D}$

Repeat!

DAgger can reduce the imitation loss from $O(\epsilon T^2)$ to $O(\epsilon T)$

Goal-conditioned Behavior Cloning

- Can we train one policy to reach multiple goals? $\pi_\theta(a_t | s_t, g_t)$
- How can we know the goal?
  - E.g., what is the goal in a demonstration $(s_0, a_0, s_1, a_1, \ldots)$?

Goal-conditioned Behavior Cloning

- Can we train one policy to reach multiple goals? $\pi_\theta(a_t | s_t, g_t)$
- How can we know the goal?
  - E.g., what is the goal in a demonstration $(s_0, a_0, s_1, a_1, \ldots)$?
- Idea: take each s_t as the goal of the demonstration $(s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t)$

DART: Disturbances Augmenting Robot Training

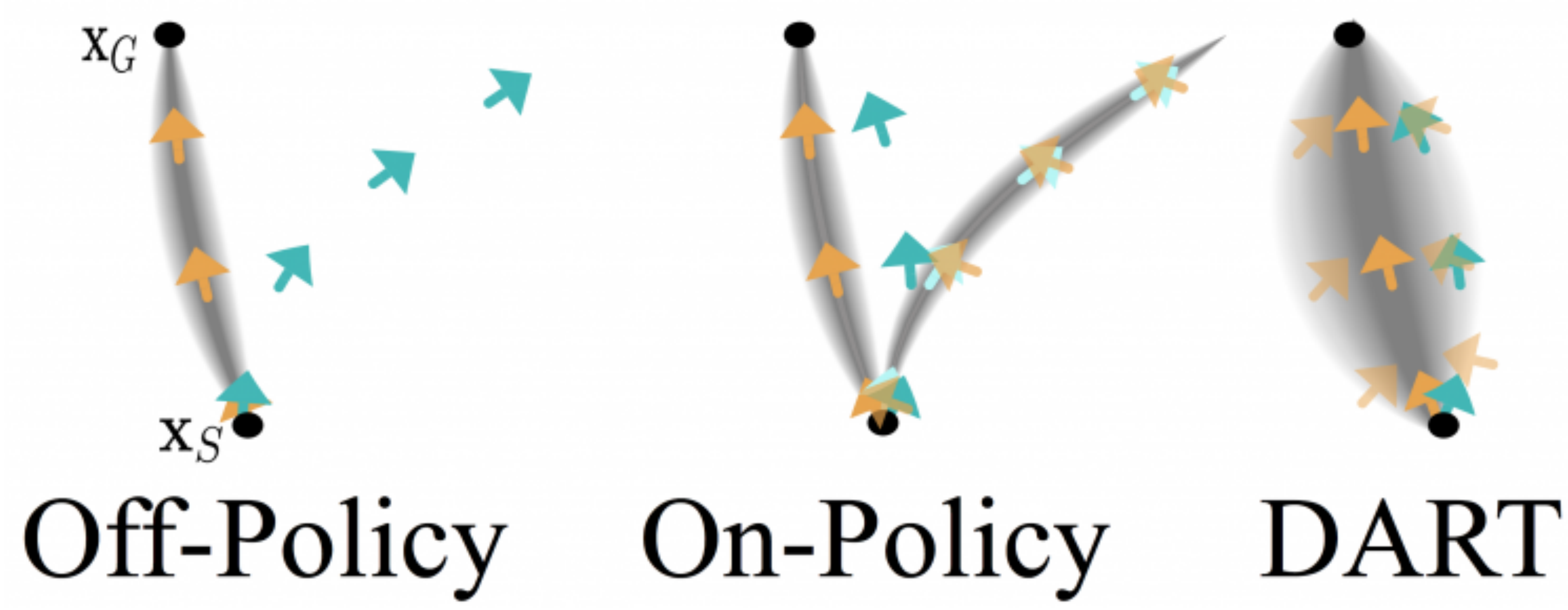


Image: Laskey et al. 2017

DART

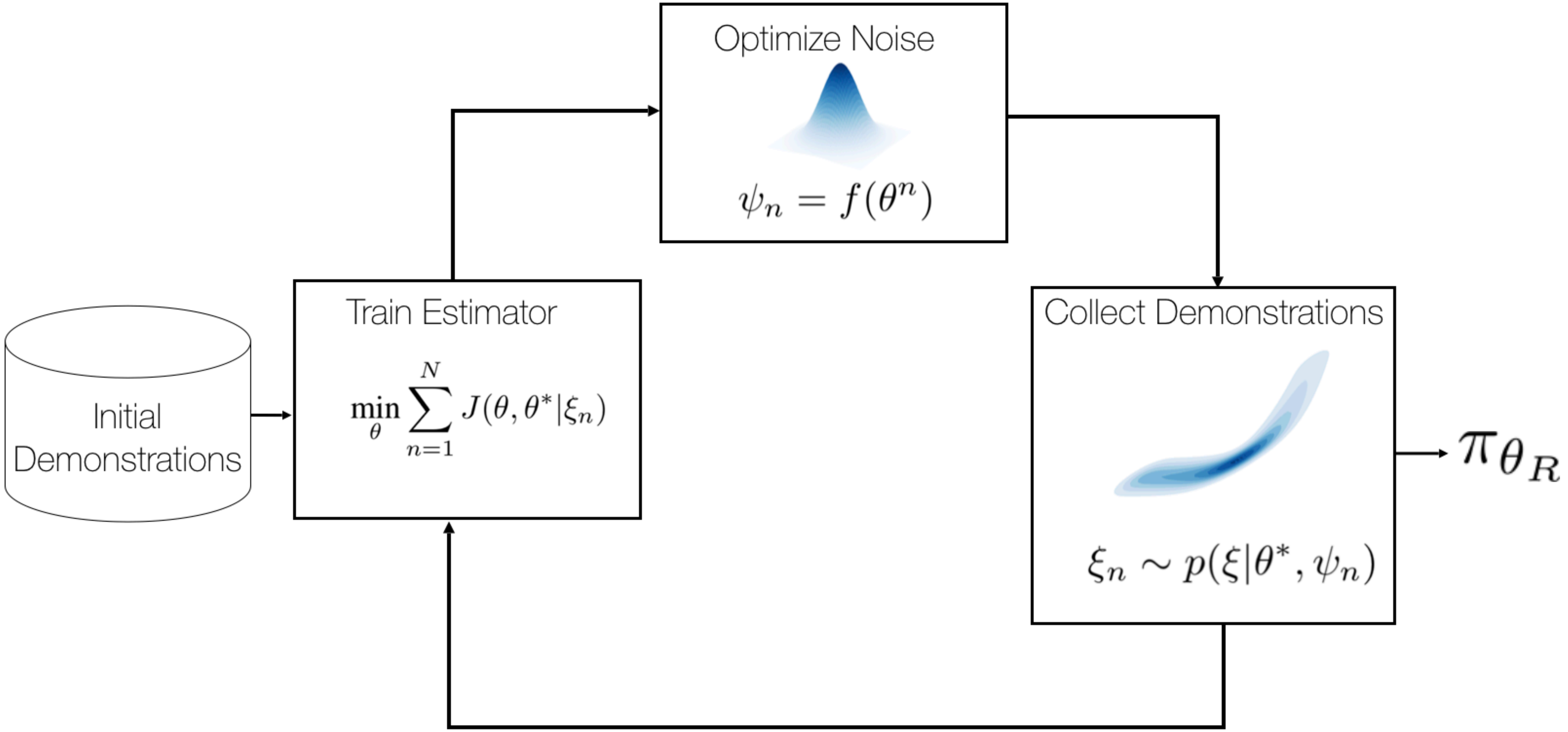


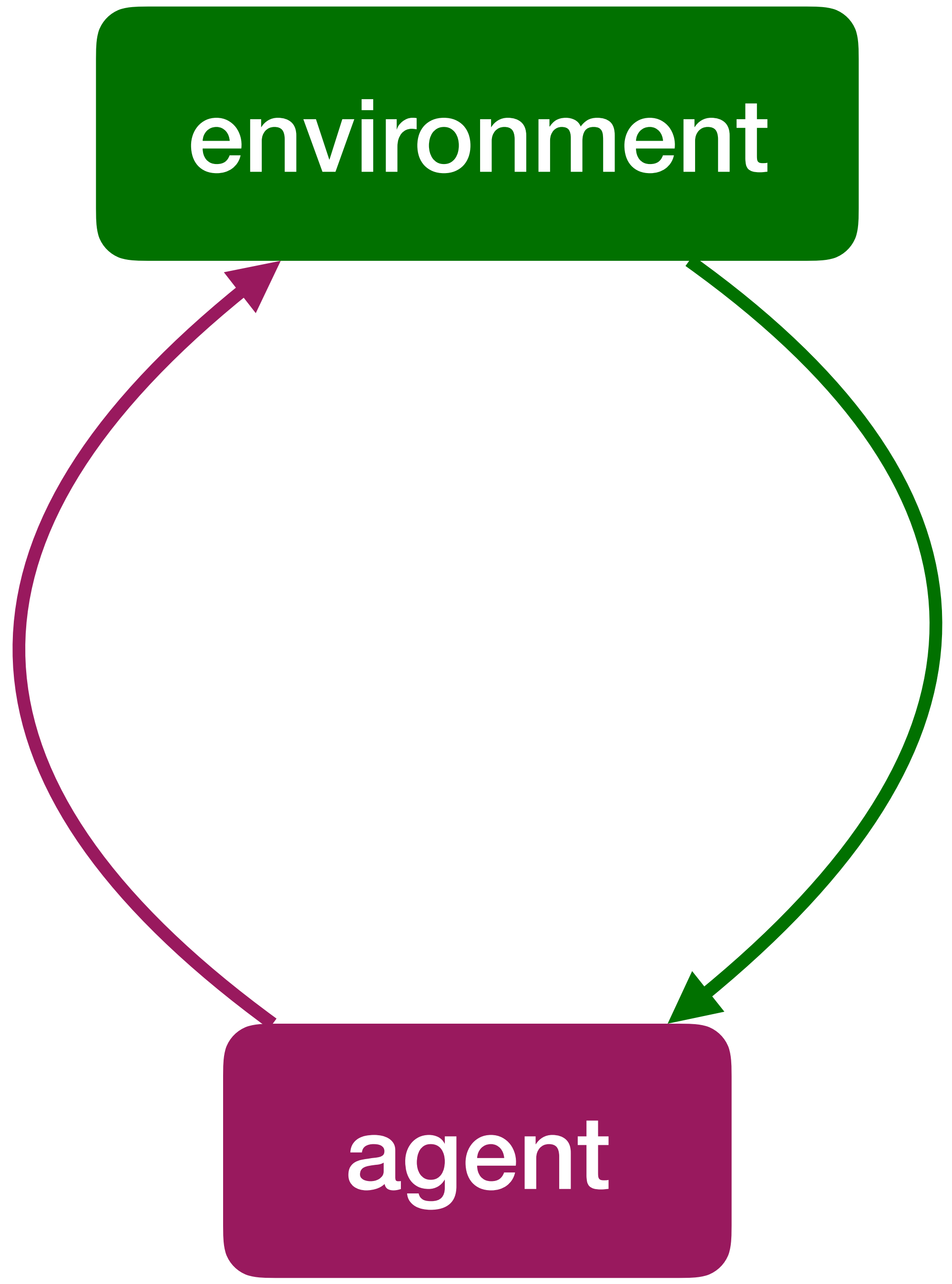
Image: Michael Laskey

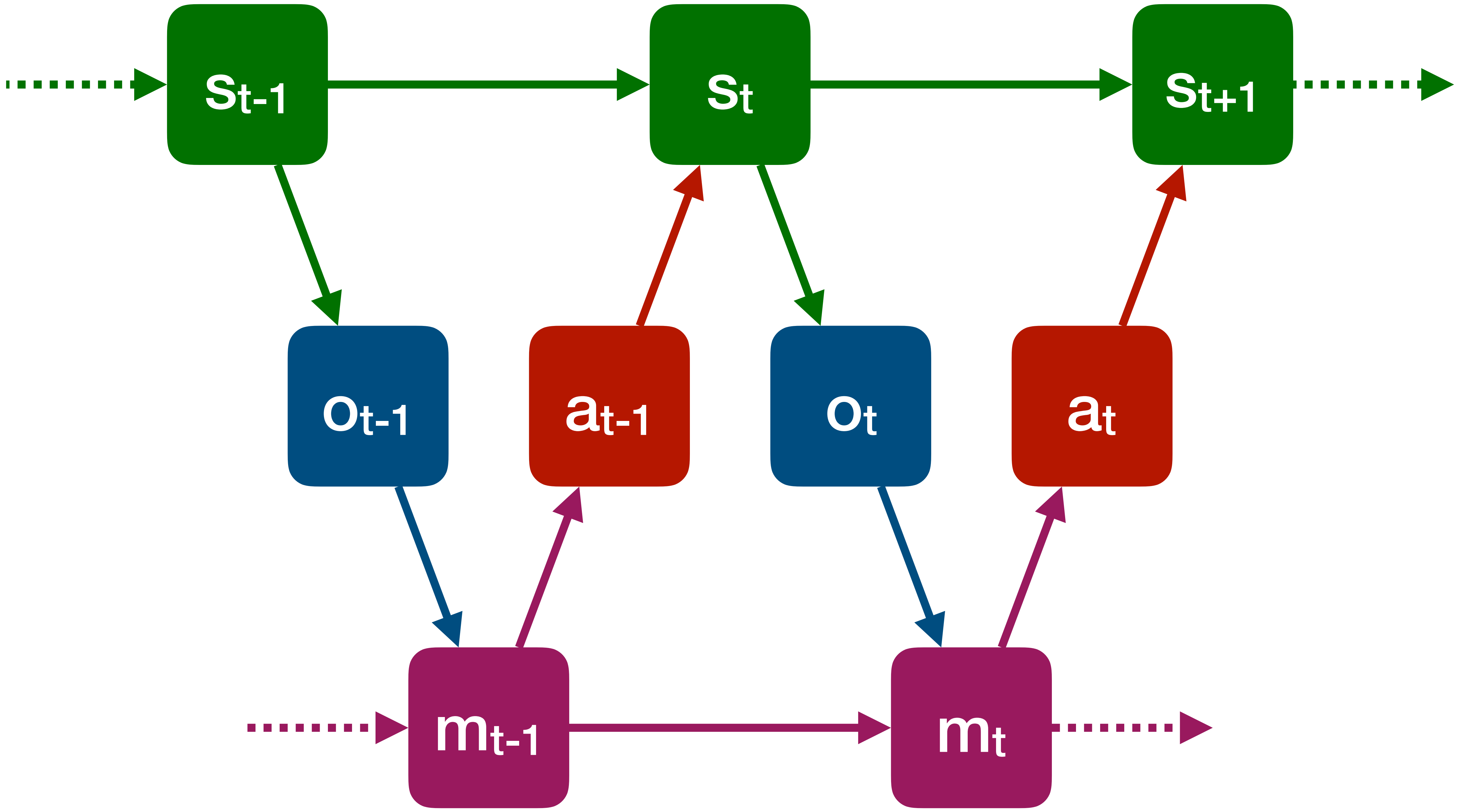
Grasping task

Behavior Cloning

DART

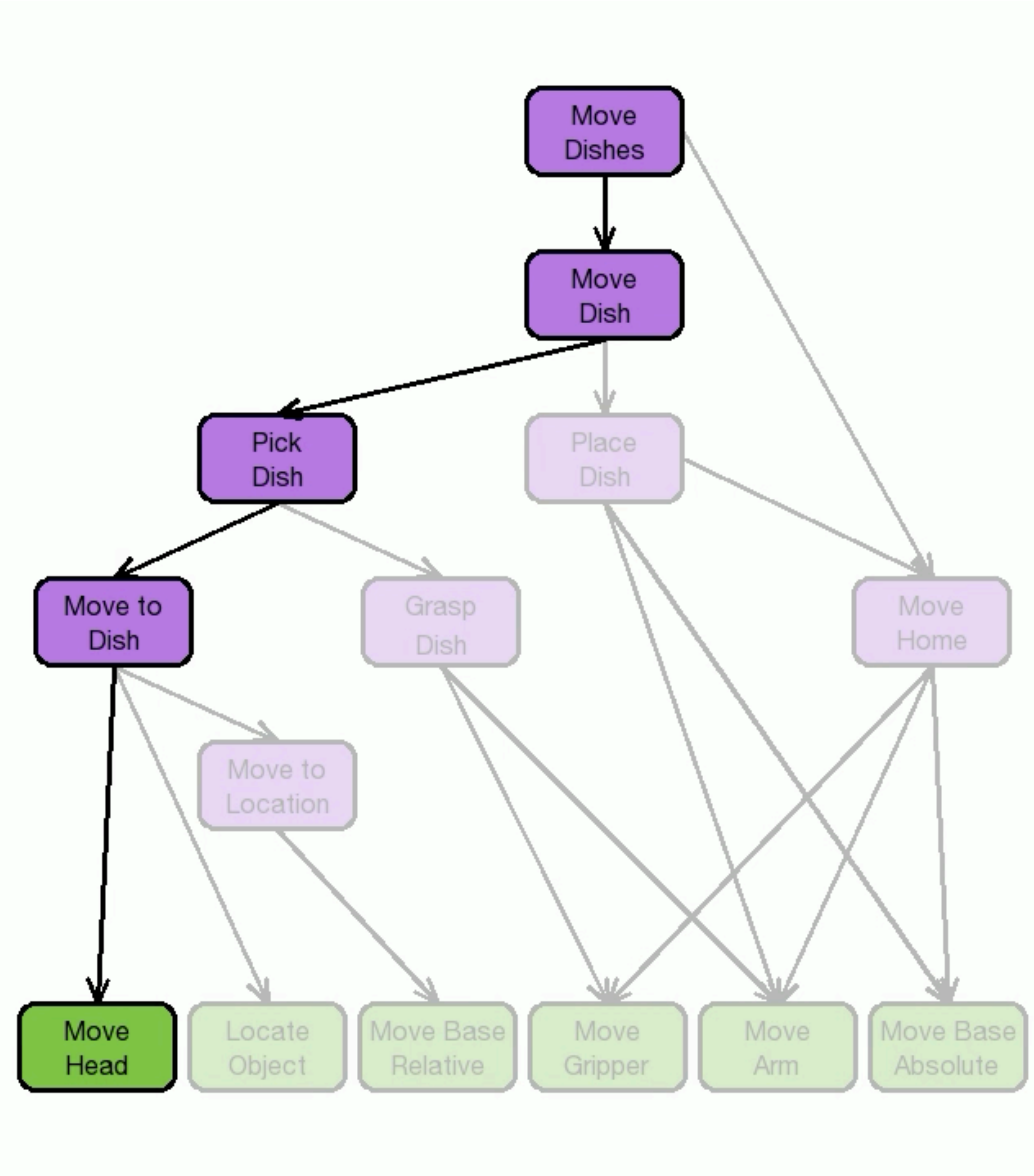
Modeling memory

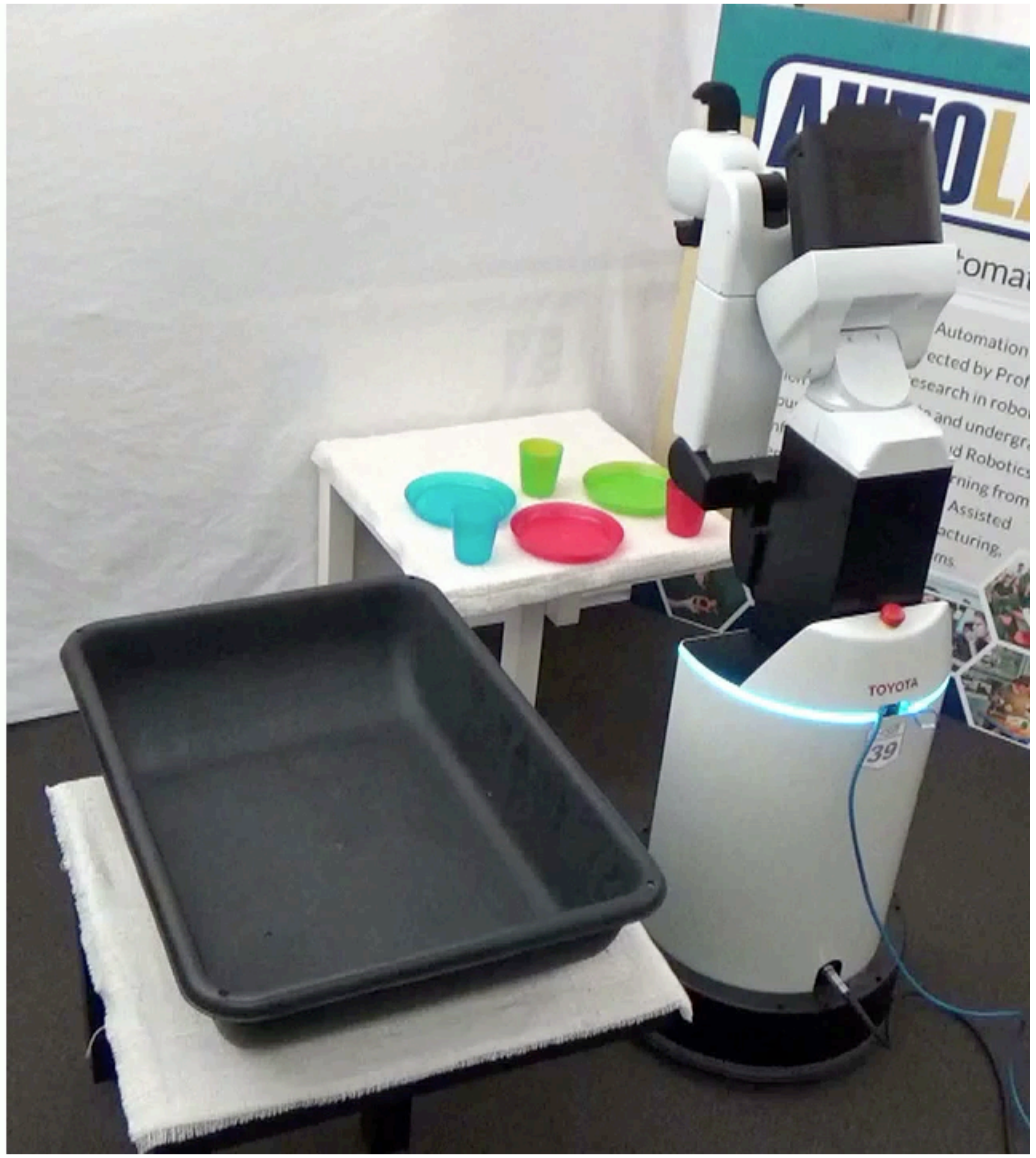




$$\pi_\theta(m_t, a_t | m_{t-1}, o_t)$$

HVIL: Hierarchical Variational Imitation Learning





Video: Fox et al. 2019

Imitation Learning as inference

- Behavior Cloning with cross-entropy loss maximizes

$$\log p_{\pi_\theta}(\mathcal{D}) = \sum_i \log \pi_\theta(a_i|o_i) + \text{const} = \log \pi_\theta(a|o) + \text{const}$$

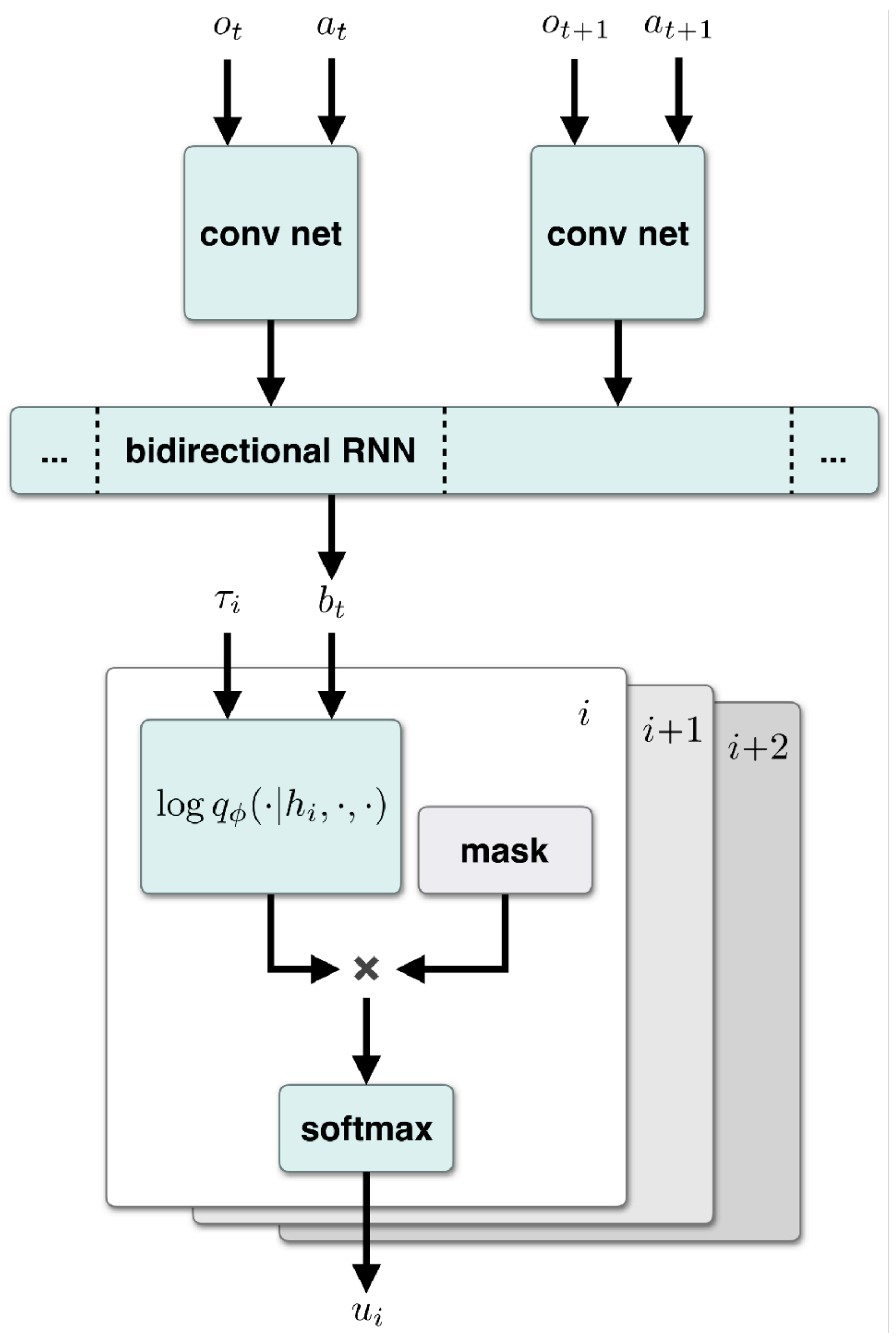
- With latent execution structure m we have $\log \pi_\theta(a|o) = \log \sum_m \pi_\theta(m, a|o)$

- Evidence Lower Bound (ELBO):

$$\log \pi_\theta(a|o) \geqslant \mathbb{E}_{m|o,a\sim q_\phi}[\log \pi_\theta(m, a|o) - \log q_\phi(m|a, o)]$$

- Inference network $q_\phi(m|a, o)$ samples execution structure m
  - which guides training of the agent $\pi_\theta(m, a|o)$

HVIL

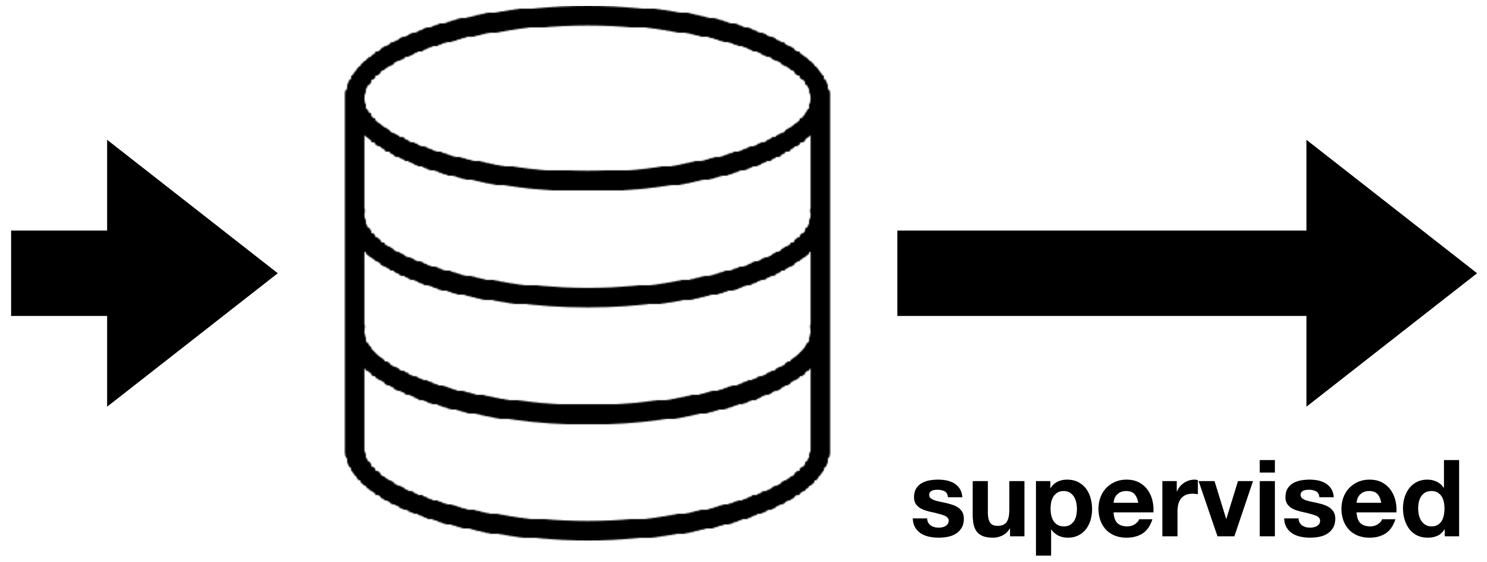


o_t
a_t
o_{t+1}
a_{t+1}
conv net
conv net
...
bidirectional RNN
...
τ_i
b_t
i
$i+1$
$i+2$
$\log q_\phi(\cdot|h_i, \cdot, \cdot)$
mask
×
softmax
u_i

Recap

observations
+
actions

training
data



supervised
learning

$\pi_\theta(a_t|o_t)$