# Behavioral Cloning

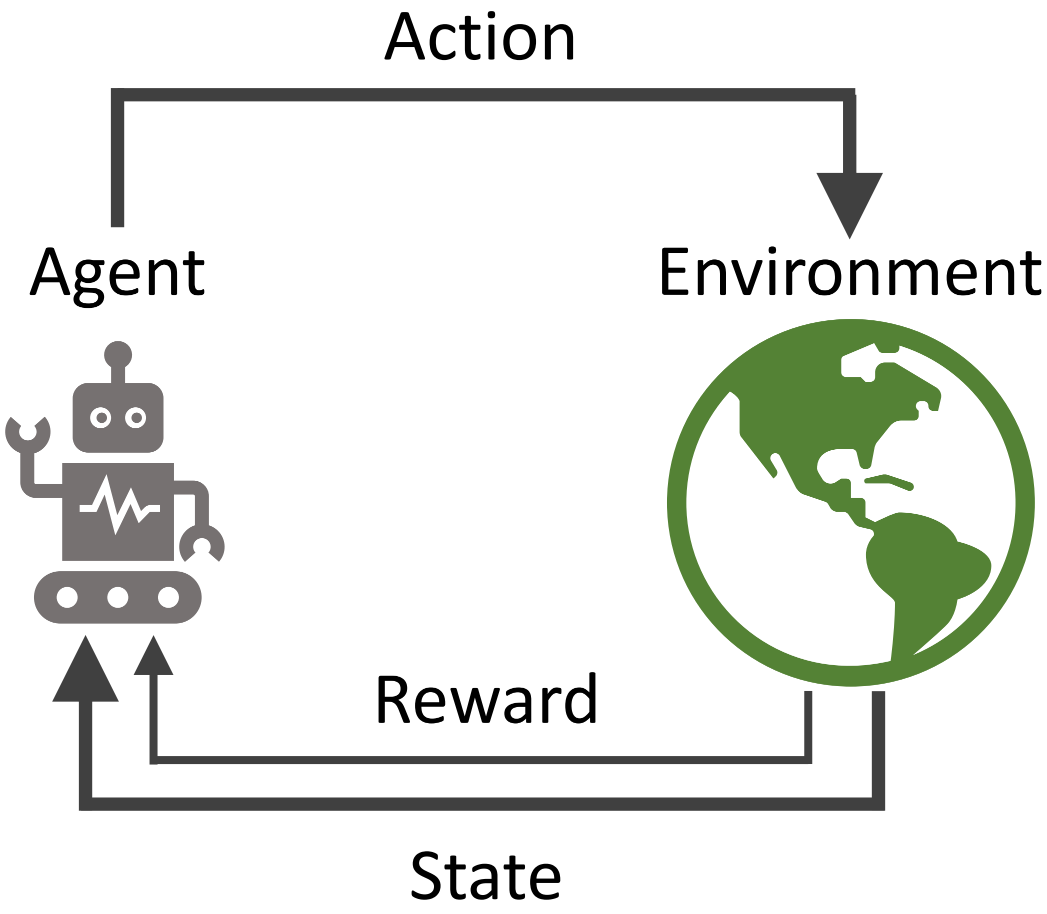
CMPT 729 G100

Jason Peng

# Overview

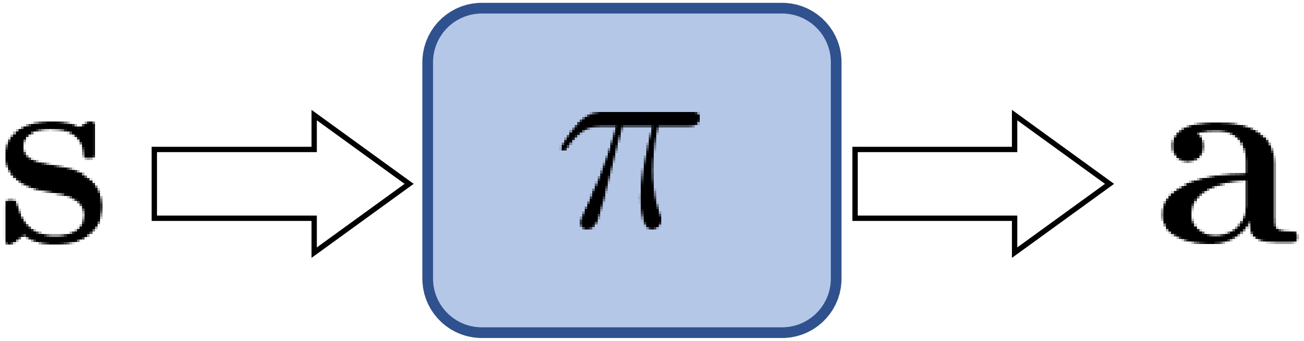
- Behavioral Cloning
- Drift
- Theoretical Analysis
- DAgger
- Applications

# Agent-Environment Interface



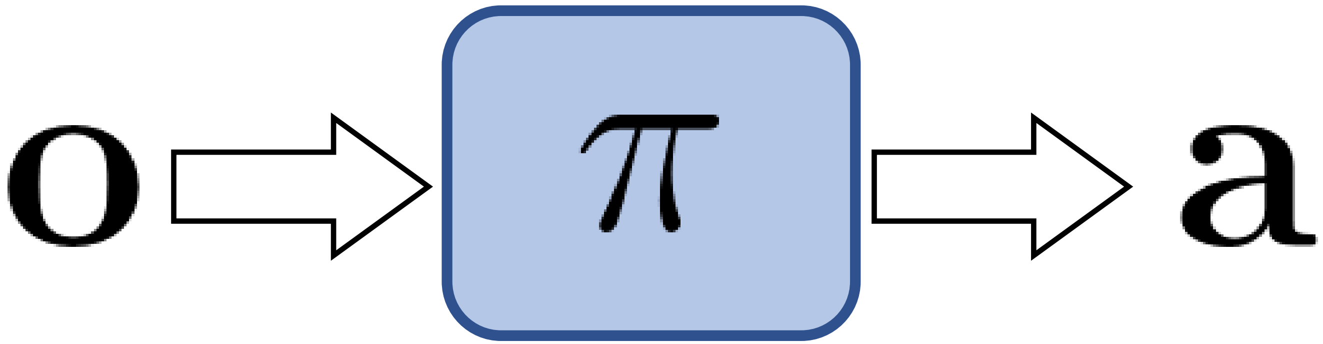
# Policy

$$\pi(\mathbf{a}|\mathbf{s})$$



# Policy

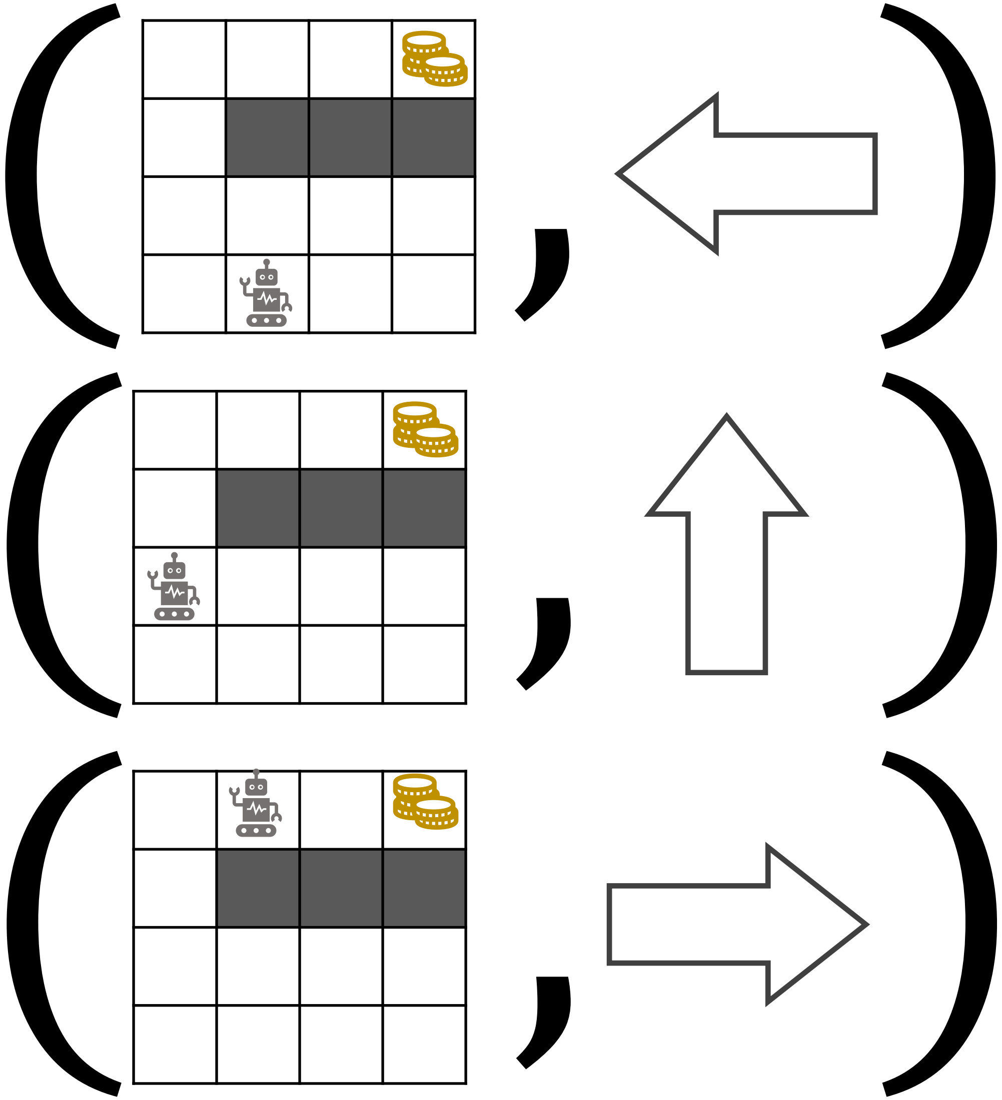
$$\pi(\mathbf{a}|\mathbf{o})$$



# Supervised Learning

$$\{(\mathbf{o}_0, \mathbf{a}_0), (\mathbf{o}_1, \mathbf{a}_1), \ldots\}$$

Dataset



# Supervised Learning

$$\left\{ \left(\mathbf{o}_0, \mathbf{a}_0\right), \left(\mathbf{o}_1, \mathbf{a}_1\right), \ldots \right\}$$

Dataset


Nvidia Automotive Simulation
[NVIDIA]

$\mathbf{o}$ $\mathbf{a}$

(  ,  )

# Supervised Learning

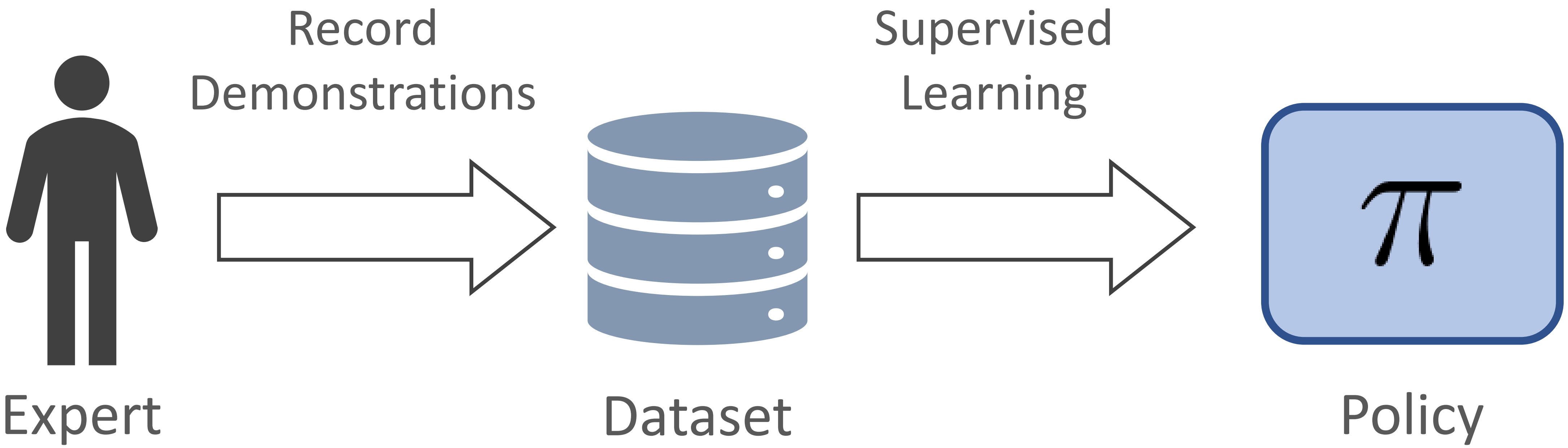
$$\{(\mathbf{o}_0, \mathbf{a}_0), (\mathbf{o}_1, \mathbf{a}_1), \ldots\}$$

$$\min_{\pi} \; \mathbb{E}_{(\mathbf{o},\mathbf{a})\sim\mathcal{D}} \left[-\log \pi(\mathbf{a}|\mathbf{o})\right]$$

Dataset

Behavioral Cloning

# Behavioral Cloning



# Behavioral Cloning

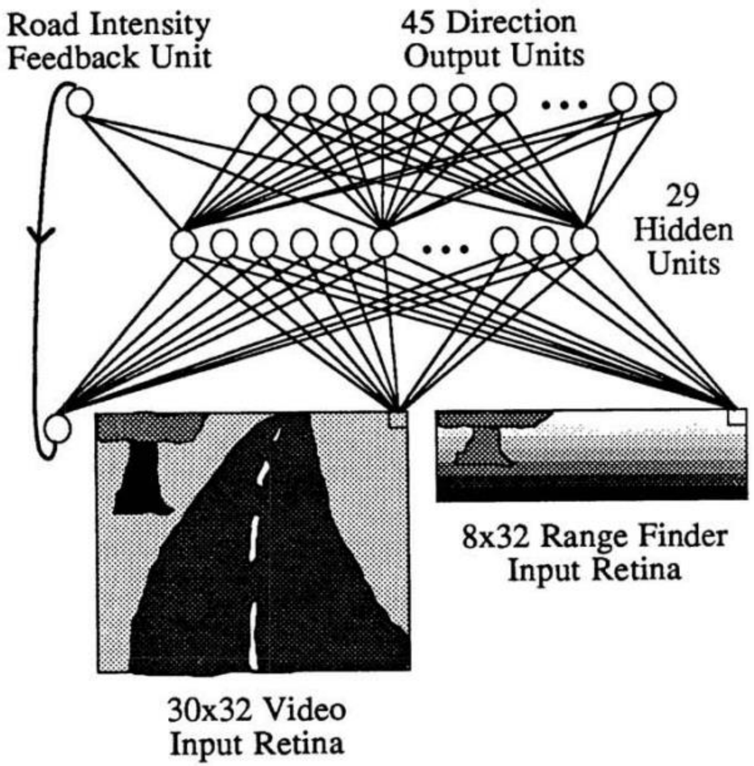
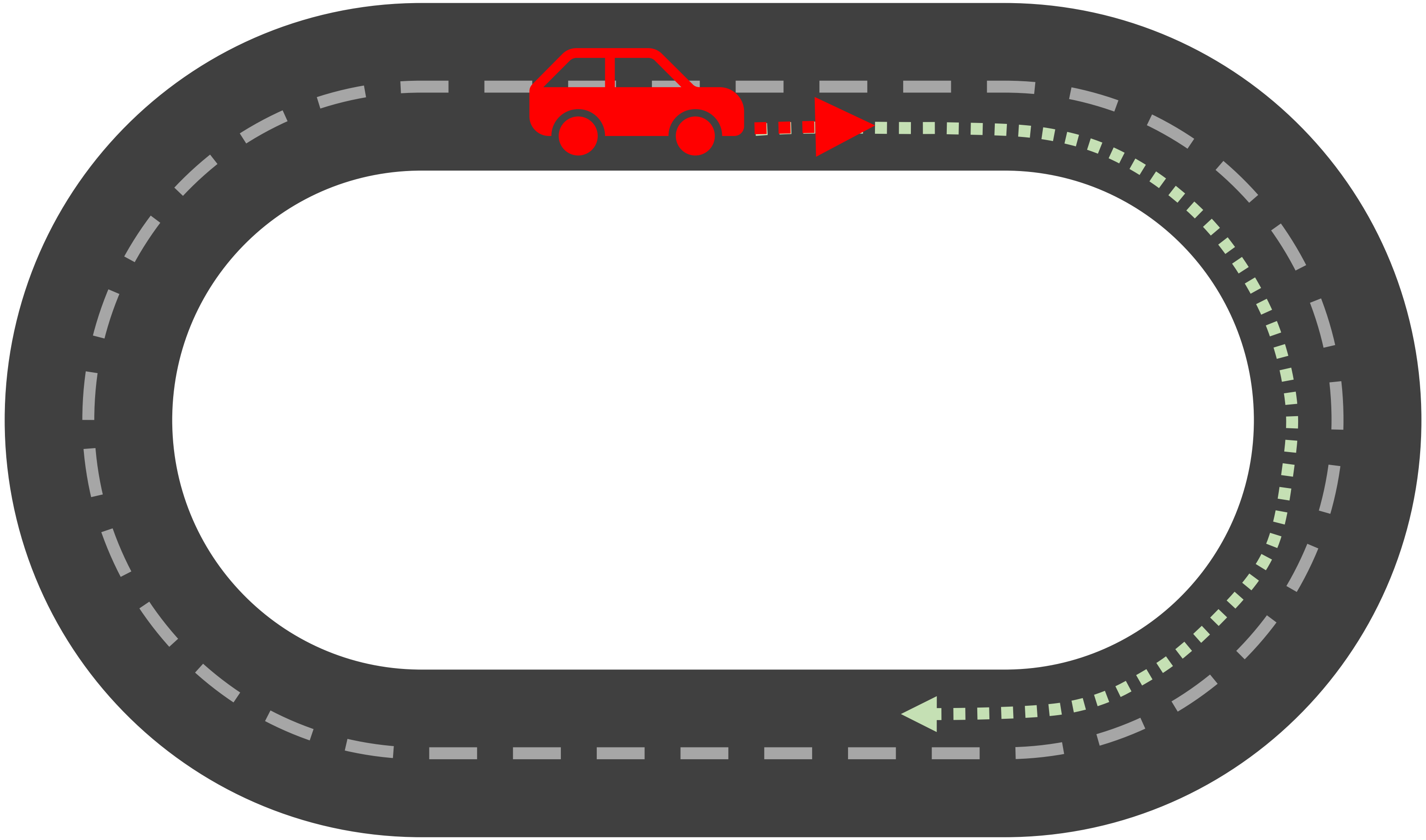


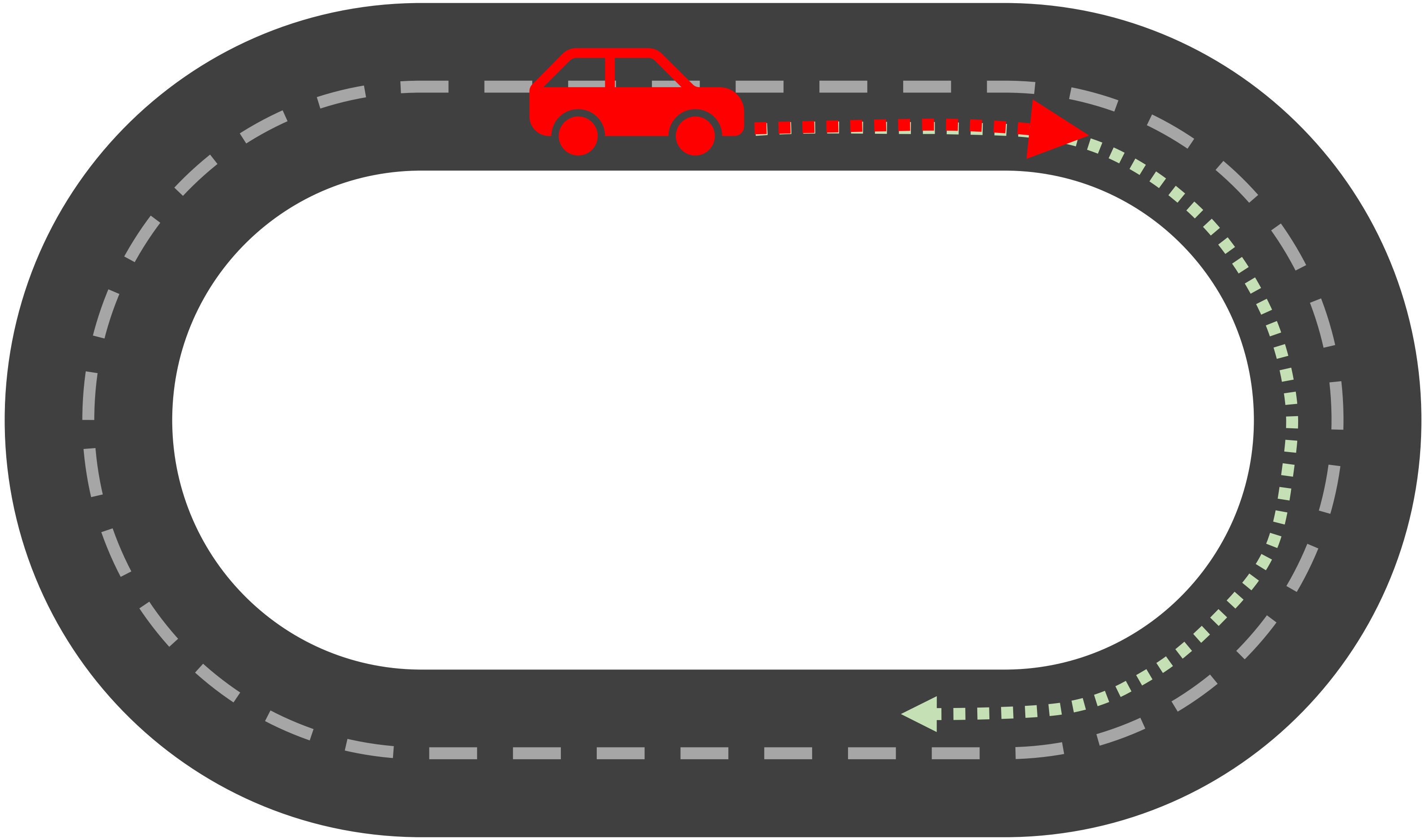
Figure 1: ALVINN Architecture

ALVINN: An Autonomous Land Vehicle in a Neural Network
[Pomerleau 1989]

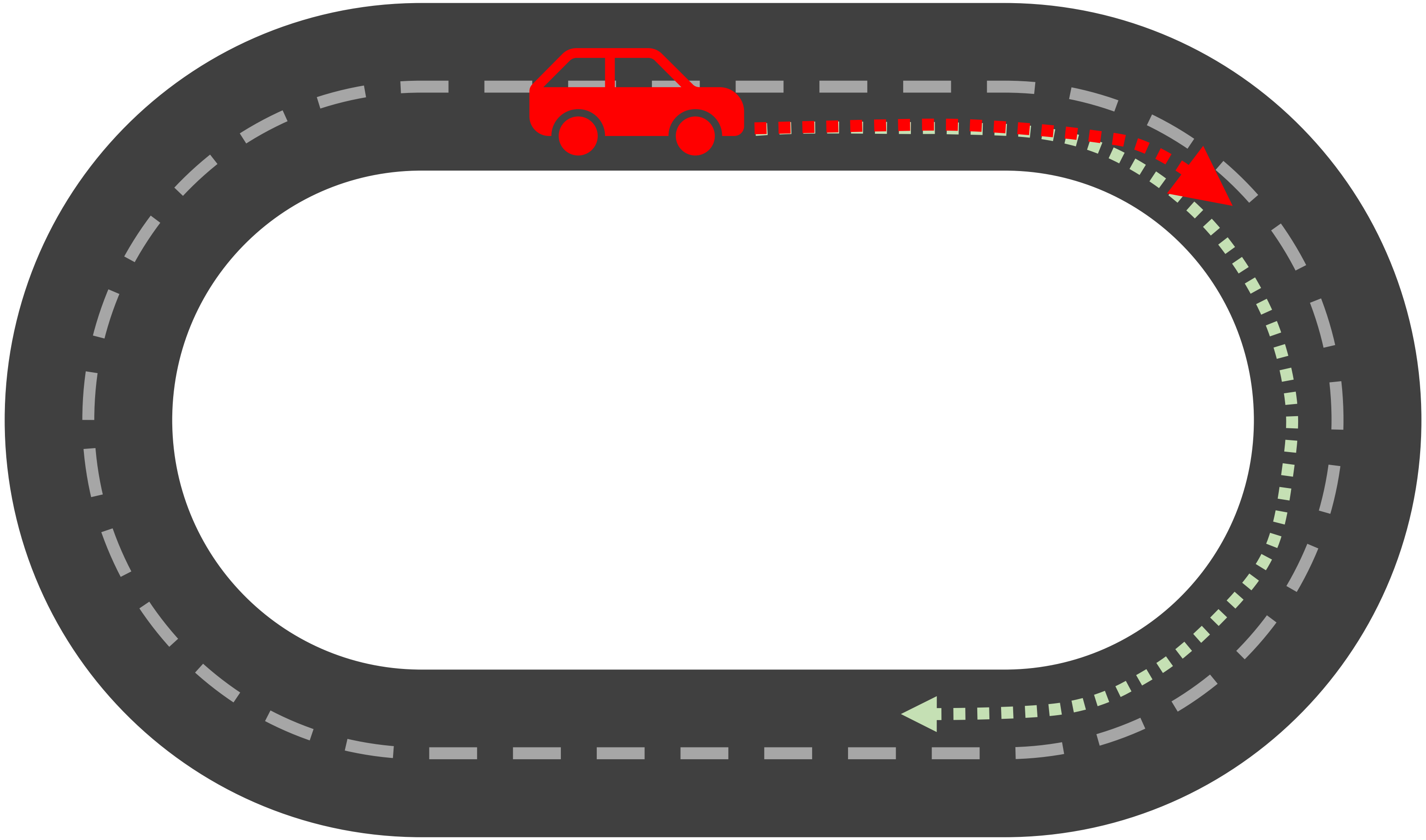
# Does it work?



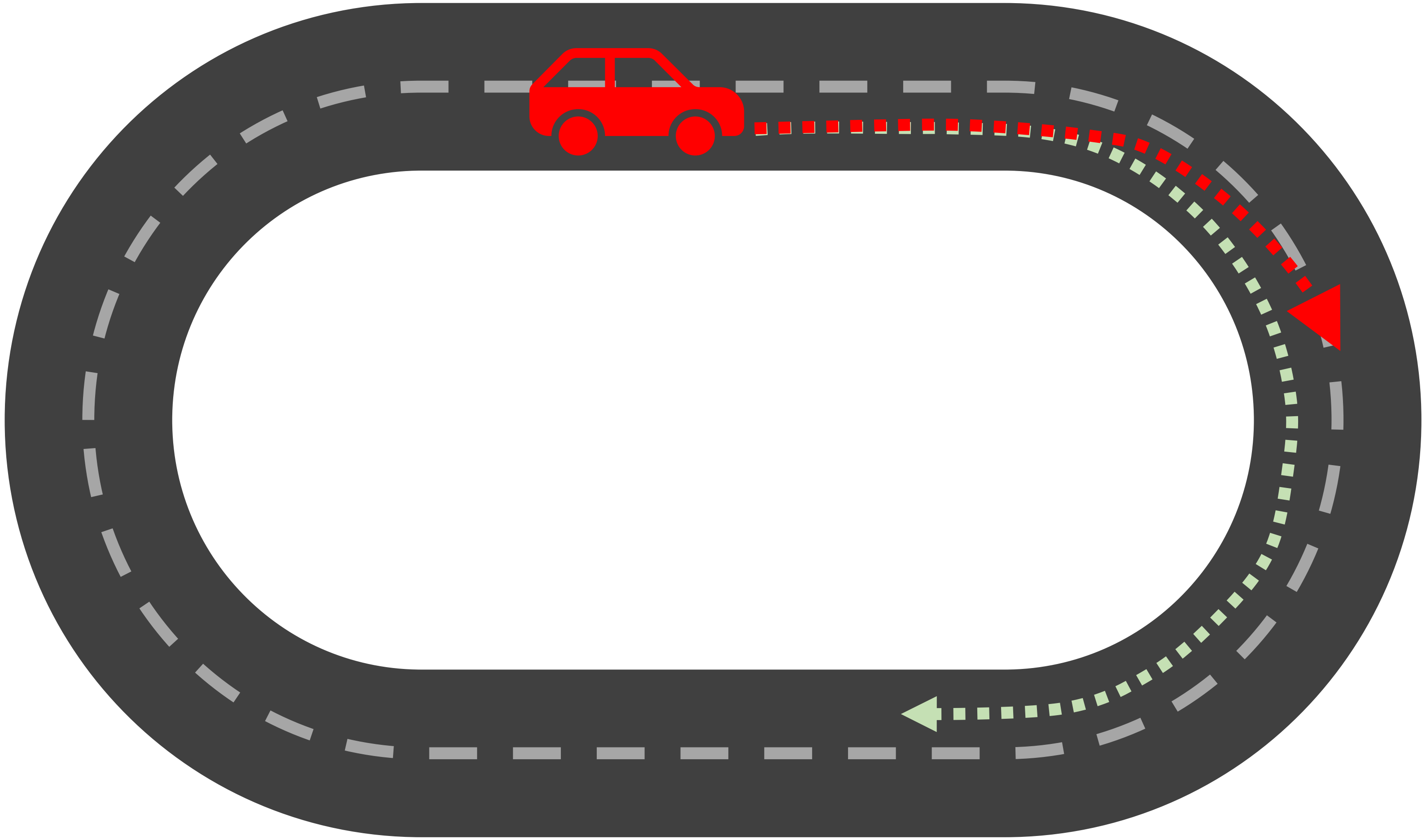
# Does it work?



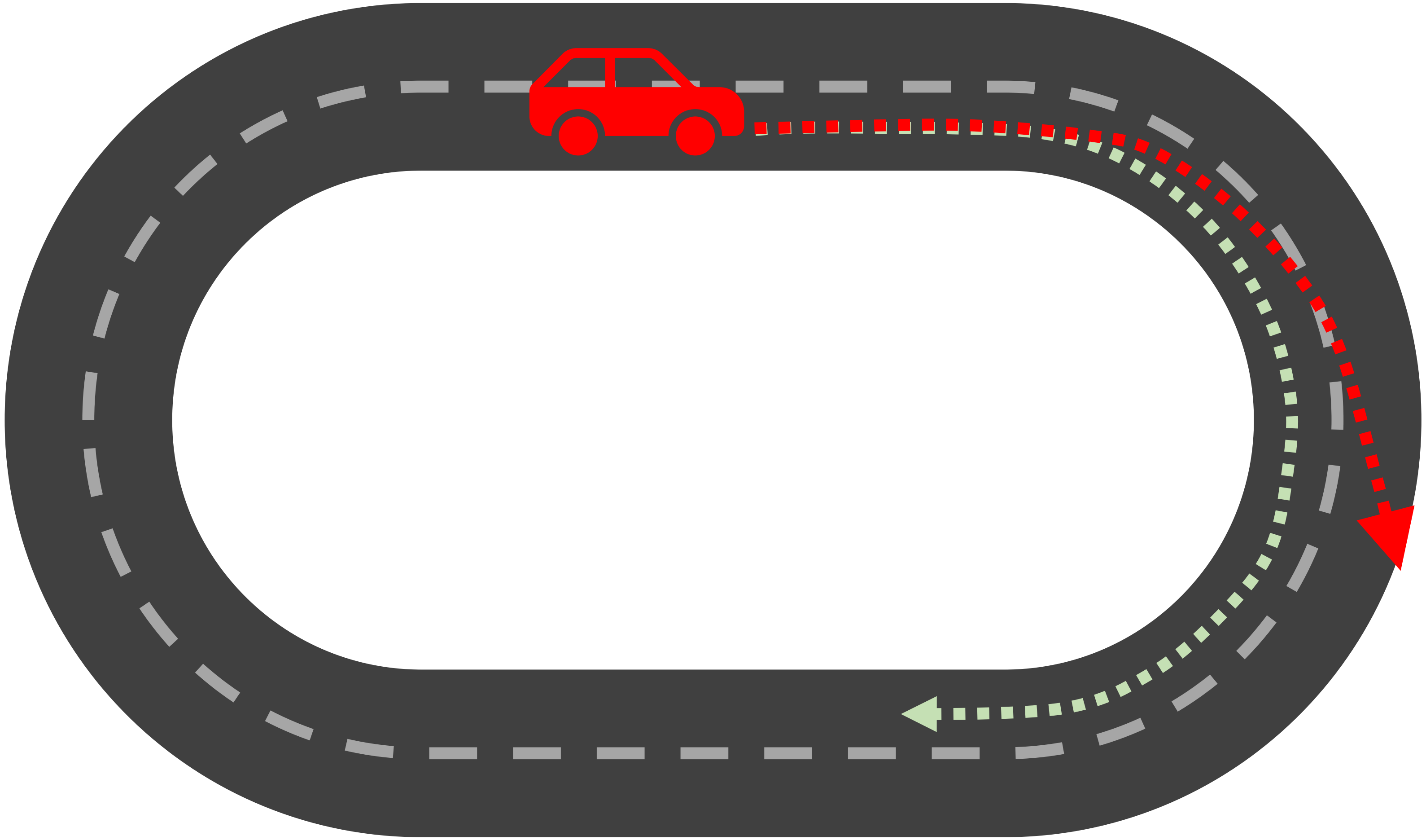
# Does it work?



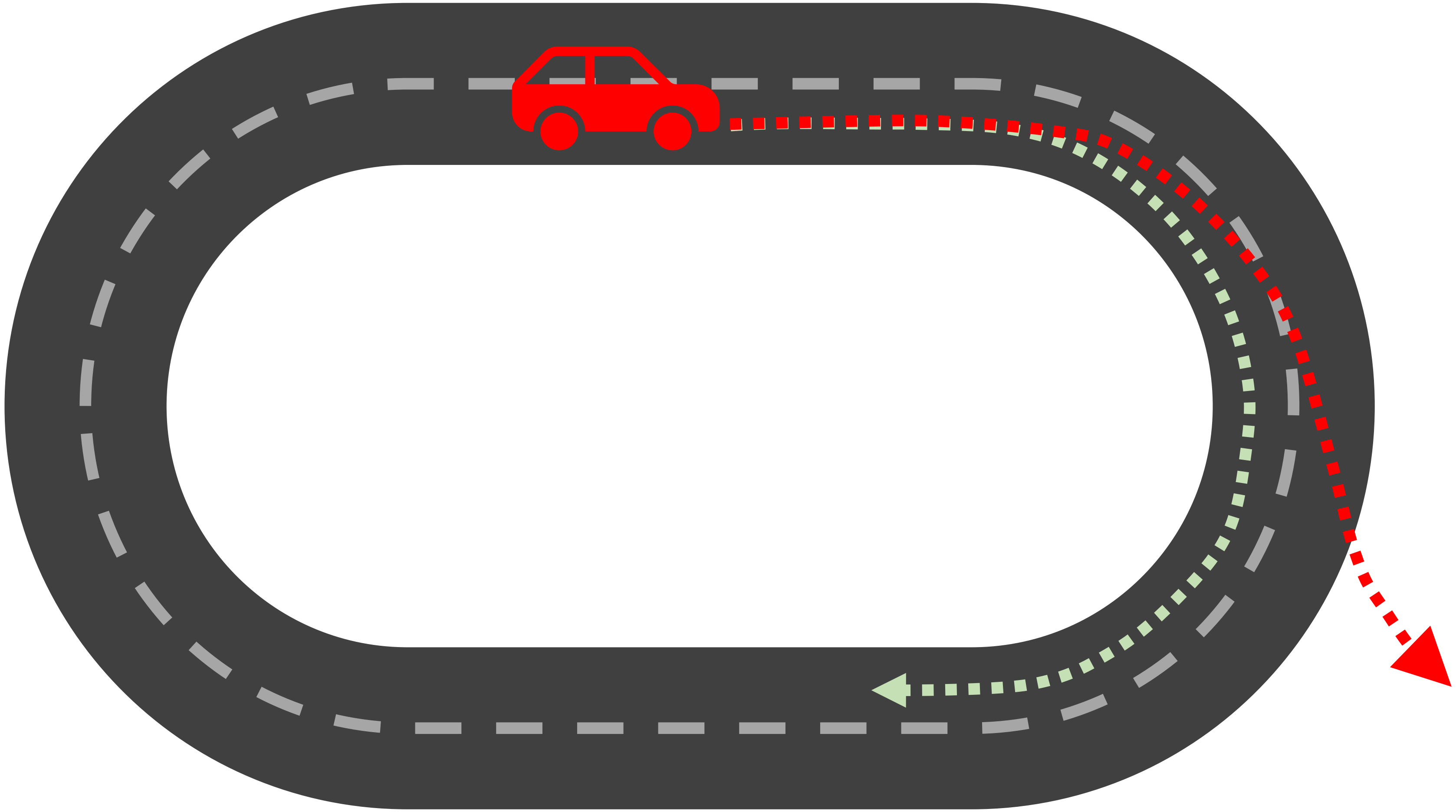
# Does it work?



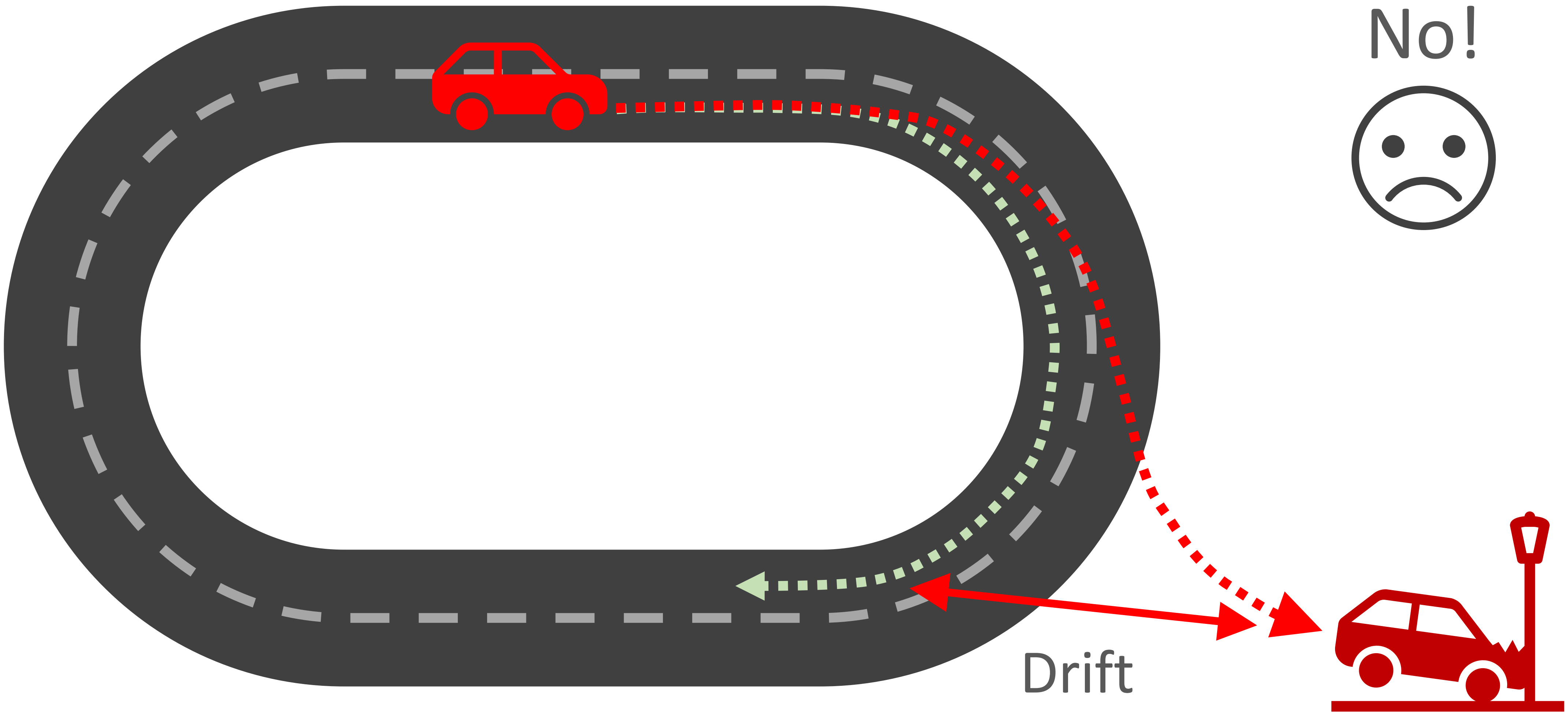
# Does it work?



# Does it work?



# Does it work?



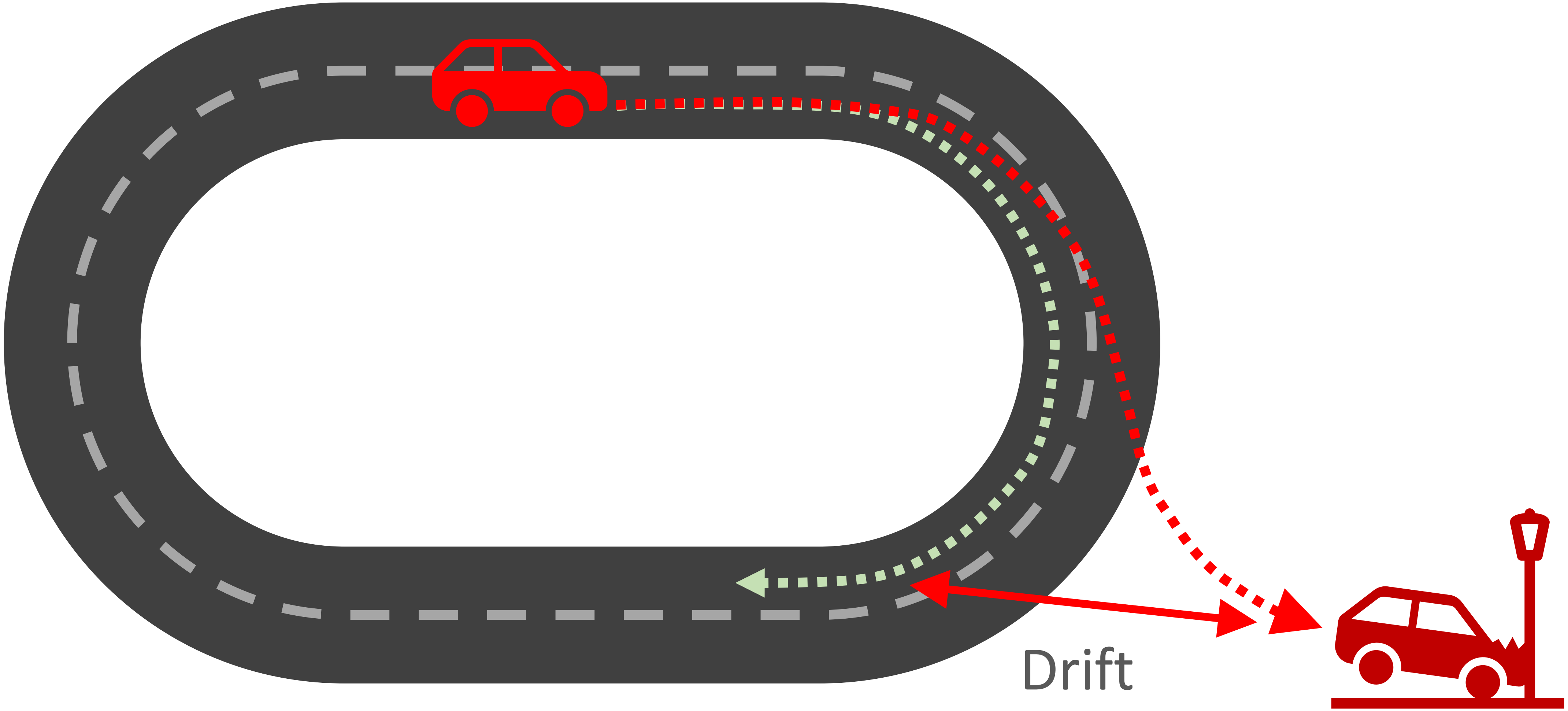
# Drift

- Expert is too good
- Lack of corrective feedback
- Policy inaccuracies
- Errors compound over time

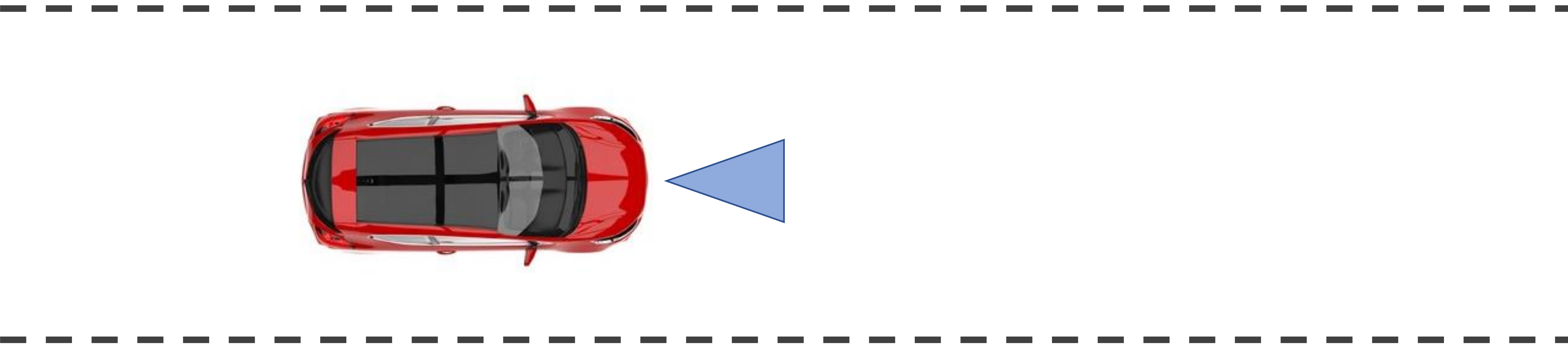
# Drift

- Expert is too good
- Lack of corrective feedback
- Policy inaccuracies
- Errors compound over time

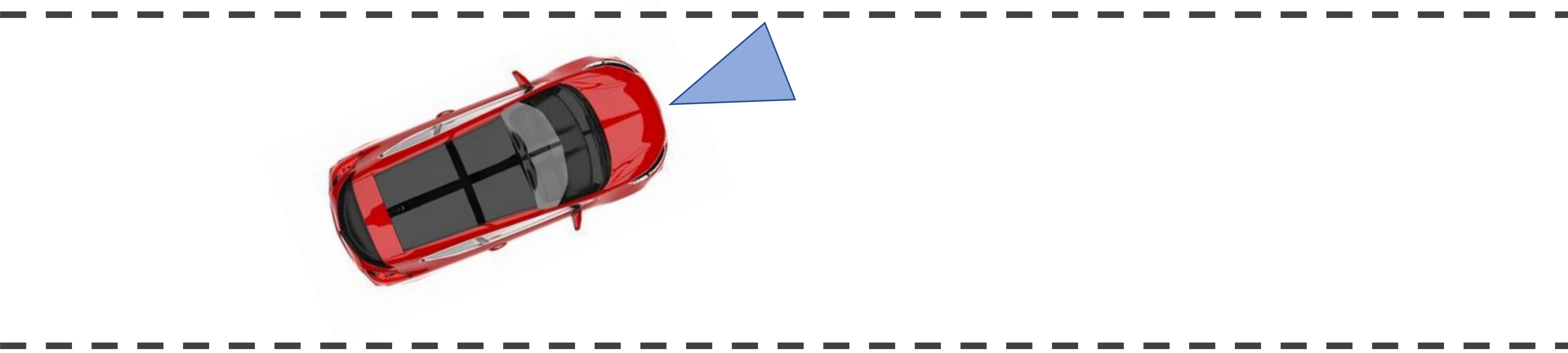
# Feedback



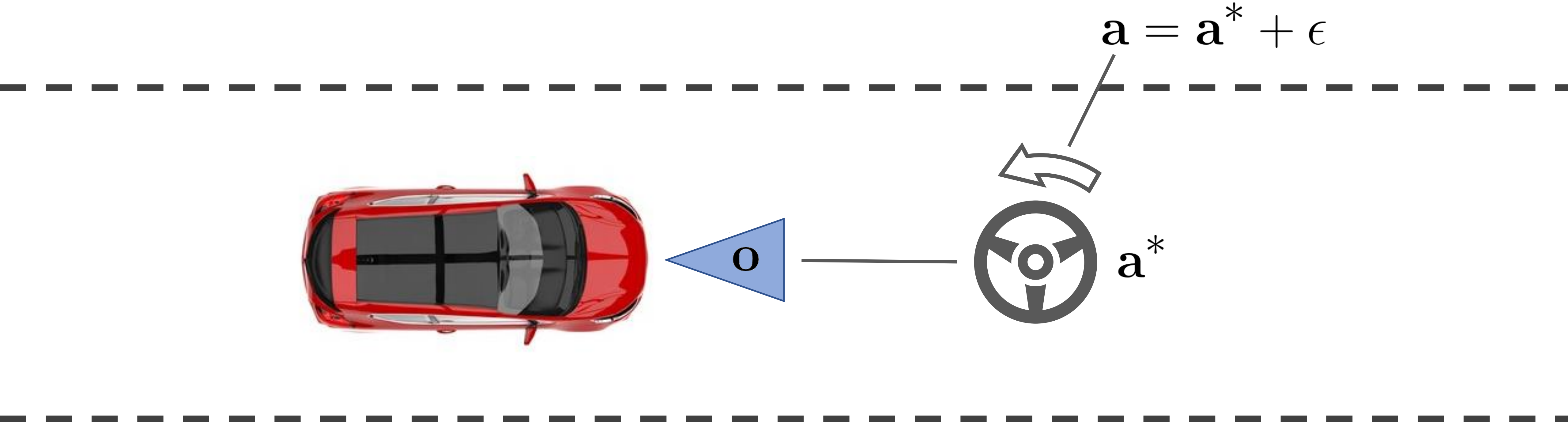
# Feedback



# Feedback

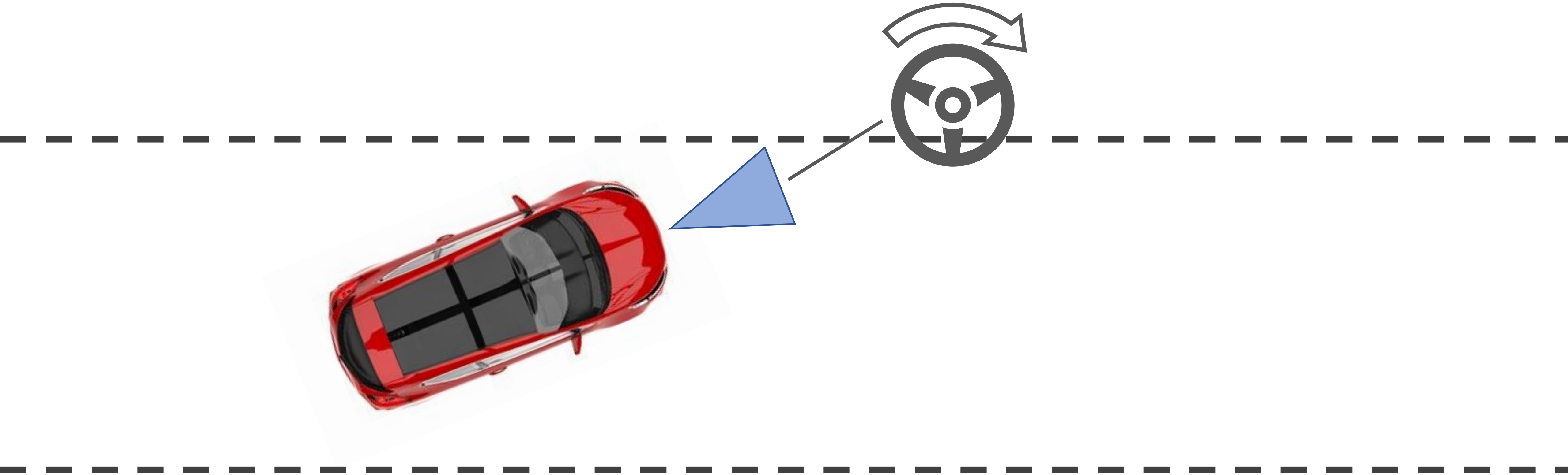


# Noise Injection




DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

# Noise Injection



DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

# Noise Injection

**ALGORITHM 2:** BC with Noise Injection

$\mathcal{D} \leftarrow \emptyset$ initialize dataset

**for** timestep $t$ **do**
  $\mathbf{o_t} \leftarrow$ record observation
  $\mathbf{a_t^*} \leftarrow$ query expert for an action

  $\epsilon_t \leftarrow$ sample noise
  $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
  Apply $\mathbf{a_t}$ to environment

  Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset $\mathcal{D}$
**end for**

$\pi^{\mathrm{BC}} = \arg\min_\pi \mathbb{E}_{(\mathbf{o}_i,\mathbf{a}_i)\sim\mathcal{D}}\left[-\log\pi(\mathbf{a}_i|\mathbf{o}_i)\right]$
return $\pi^{\mathrm{BC}}$

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

# Noise Injection

**ALGORITHM 2:** BC with Noise Injection

$\mathcal{D} \leftarrow \emptyset$ initialize dataset

**for** timestep $t$ **do**

 $\mathbf{o_t} \leftarrow$ record observation

 $\mathbf{a_t^*} \leftarrow$ query expert for an action

 $\epsilon_t \leftarrow$ sample noise

 $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$

 Apply $\mathbf{a_t}$ to environment

 Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset $\mathcal{D}$

**end for**

$\pi^{\mathrm{BC}} = \arg\min_\pi \mathbb{E}_{(\mathbf{o}_i,\mathbf{a}_i)\sim\mathcal{D}} [-\log \pi(\mathbf{a}_i|\mathbf{o}_i)]$

return $\pi^{\mathrm{BC}}$

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

# Noise Injection

**ALGORITHM 2:** BC with Noise Injection

$\mathcal{D} \leftarrow \emptyset$ initialize dataset

**for** timestep $t$ **do**
  $\mathbf{o_t} \leftarrow$ record observation
  $\mathbf{a_t^*} \leftarrow$ query expert for an action

  $\epsilon_t \leftarrow$ sample noise
  $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
  Apply $\mathbf{a_t}$ to environment

  Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset $\mathcal{D}$
**end for**

$\pi^{\mathrm{BC}} = \arg\min_\pi \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} [-\log \pi(\mathbf{a}_i | \mathbf{o}_i)]$
return $\pi^{\mathrm{BC}}$

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

# Noise Injection

**ALGORITHM 2:** BC with Noise Injection

$\mathcal{D} \leftarrow \emptyset$ initialize dataset

**for** timestep $t$ **do**
  $\mathbf{o_t} \leftarrow$ record observation
  $\mathbf{a_t^*} \leftarrow$ query expert for an action

  $\epsilon_t \leftarrow$ sample noise
  $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
  Apply $\mathbf{a_t}$ to environment

  Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset $\mathcal{D}$
**end for**

$\pi^{\mathrm{BC}} = \arg\min_\pi \mathbb{E}_{(\mathbf{o}_i,\mathbf{a}_i)\sim\mathcal{D}}\left[-\log\pi(\mathbf{a}_i|\mathbf{o}_i)\right]$
return $\pi^{\mathrm{BC}}$

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

# Noise Injection

**ALGORITHM 2:** BC with Noise Injection

1: $\mathcal{D} \leftarrow \emptyset$ initialize dataset

2: **for** timestep $t$ **do**
3: $\quad \mathbf{o_t} \leftarrow$ record observation
4: $\quad \mathbf{a_t^*} \leftarrow$ query expert for an action

5: $\quad \epsilon_t \leftarrow$ sample noise
6: $\quad \mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
7: $\quad$ Apply $\mathbf{a_t}$ to environment

8: $\quad$ Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset $\mathcal{D}$
9: **end for**

10: $\pi^{\mathrm{BC}} = \arg\min_\pi \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} [-\log \pi(\mathbf{a}_i | \mathbf{o}_i)]$
11: return $\pi^{\mathrm{BC}}$

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

# Noise Injection

**ALGORITHM 2:** BC with Noise Injection

1: $\mathcal{D} \leftarrow \emptyset$ initialize dataset

2: **for** timestep $t$ **do**
3: $\quad \mathbf{o_t} \leftarrow$ record observation
4: $\quad \mathbf{a_t^*} \leftarrow$ query expert for an action

5: $\quad \epsilon_t \leftarrow$ sample noise
6: $\quad \mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
7: $\quad$ Apply $\mathbf{a_t}$ to environment

8: $\quad$ Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset $\mathcal{D}$
9: **end for**

10: $\pi^{\text{BC}} = \arg\min_{\pi} \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} [-\log \pi(\mathbf{a}_i | \mathbf{o}_i)]$
11: return $\pi^{\text{BC}}$

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

# Noise Injection

**ALGORITHM 2:** BC with Noise Injection

1: $\mathcal{D} \leftarrow \emptyset$ initialize dataset

2: **for** timestep $t$ **do**
3: $\mathbf{o_t} \leftarrow$ record observation
4: $\mathbf{a_t^*} \leftarrow$ query expert for an action

5: $\epsilon_t \leftarrow$ sample noise
6: $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
7: Apply $\mathbf{a_t}$ to environment

8: Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset $\mathcal{D}$
9: **end for**

10: $\pi^{\mathrm{BC}} = \arg\min_\pi \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} \left[ -\log \pi(\mathbf{a}_i | \mathbf{o}_i) \right]$
11: return $\pi^{\mathrm{BC}}$

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

# Noise Injection

**ALGORITHM 2:** BC with Noise Injection

1: $\mathcal{D} \leftarrow \emptyset$ initialize dataset

2: **for** timestep $t$ **do**
3: $\quad \mathbf{o_t} \leftarrow$ record observation
4: $\quad \mathbf{a_t^*} \leftarrow$ query expert for an action

5: $\quad \epsilon_t \leftarrow$ sample noise
6: $\quad \mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$
7: $\quad$ Apply $\mathbf{a_t}$ to environment

8: $\quad$ Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset $\mathcal{D}$
9: **end for**

10: $\pi^{\mathrm{BC}} = \arg\min_\pi \mathbb{E}_{(\mathbf{o}_i, \mathbf{a}_i) \sim \mathcal{D}} \left[ -\log \pi(\mathbf{a}_i | \mathbf{o}_i) \right]$
11: return $\pi^{\mathrm{BC}}$

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

# Noise Injection

**ALGORITHM 2:** BC with Noise Injection

1: $\mathcal{D} \leftarrow \emptyset$ initialize dataset

2: **for** timestep $t$ **do**

3: $\quad \mathbf{o_t} \leftarrow$ record observation

4: $\quad \mathbf{a_t^*} \leftarrow$ query expert for an action

5: $\quad \epsilon_t \leftarrow$ sample noise

6: $\quad \mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$

7: $\quad$ Apply $\mathbf{a_t}$ to environment

8: $\quad$ Store $(\mathbf{o}_t, \underline{\mathbf{a}_t^*})$ in dataset $\mathcal{D}$

9: **end for**

10: $\pi^{\mathrm{BC}} = \arg\min_\pi \mathbb{E}_{(\mathbf{o}_i,\mathbf{a}_i)\sim\mathcal{D}} [-\log\pi(\mathbf{a}_i|\mathbf{o}_i)]$

11: return $\pi^{\mathrm{BC}}$

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

# Noise Injection

**ALGORITHM 2:** BC with Noise Injection

$\mathcal{D} \leftarrow \emptyset$ initialize dataset

**for** timestep $t$ **do**

  $\mathbf{o_t} \leftarrow$ record observation

  $\mathbf{a_t^*} \leftarrow$ query expert for an action

  $\epsilon_t \leftarrow$ sample noise

  $\mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$

  Apply $\mathbf{a_t}$ to environment

  Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset $\mathcal{D}$

**end for**

$\pi^{BC} = \arg\min_\pi \mathbb{E}_{(\mathbf{o}_i,\mathbf{a}_i)\sim\mathcal{D}}[-\log\pi(\mathbf{a}_i|\mathbf{o}_i)]$

return $\pi^{BC}$

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

# Noise Injection

**ALGORITHM 2:** BC with Noise Injection

1: $\mathcal{D} \leftarrow \emptyset$ initialize dataset

2: **for** timestep $t$ **do**

3: $\quad \mathbf{o_t} \leftarrow$ record observation

4: $\quad \mathbf{a_t^*} \leftarrow$ query expert for an action

5: $\quad \epsilon_t \leftarrow$ sample noise

6: $\quad \mathbf{a}_t \leftarrow \mathbf{a}_t^* + \epsilon_t$

7: $\quad$ Apply $\mathbf{a_t}$ to environment

8: $\quad$ Store $(\mathbf{o}_t, \mathbf{a}_t^*)$ in dataset $\mathcal{D}$

9: **end for**

10: $\pi^{\mathrm{BC}} = \arg\min_\pi \mathbb{E}_{(\mathbf{o}_i,\mathbf{a}_i)\sim\mathcal{D}}\left[-\log\pi(\mathbf{a}_i|\mathbf{o}_i)\right]$

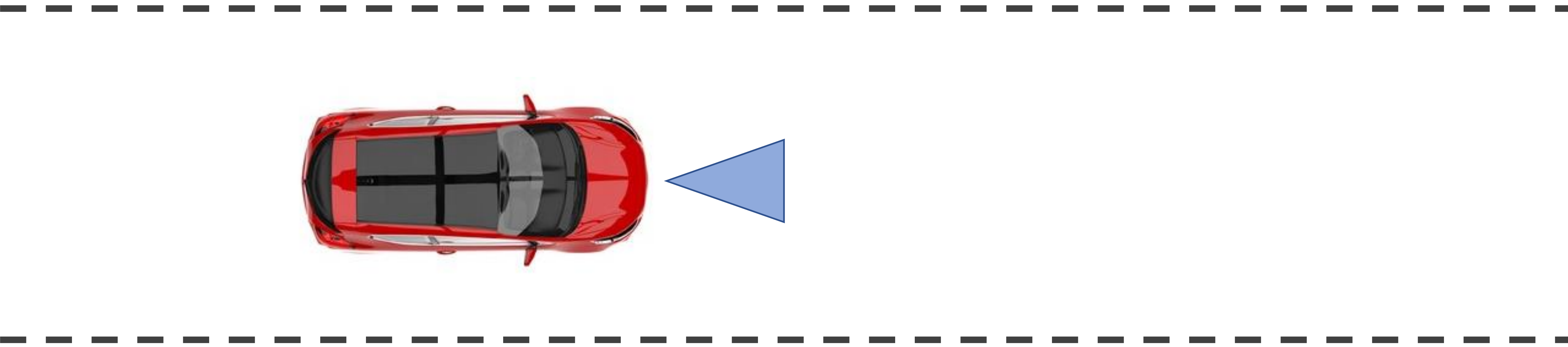
11: return $\pi^{\mathrm{BC}}$

DART: Noise Injection for Robust Imitation Learning
[Laskey et al. 2017]

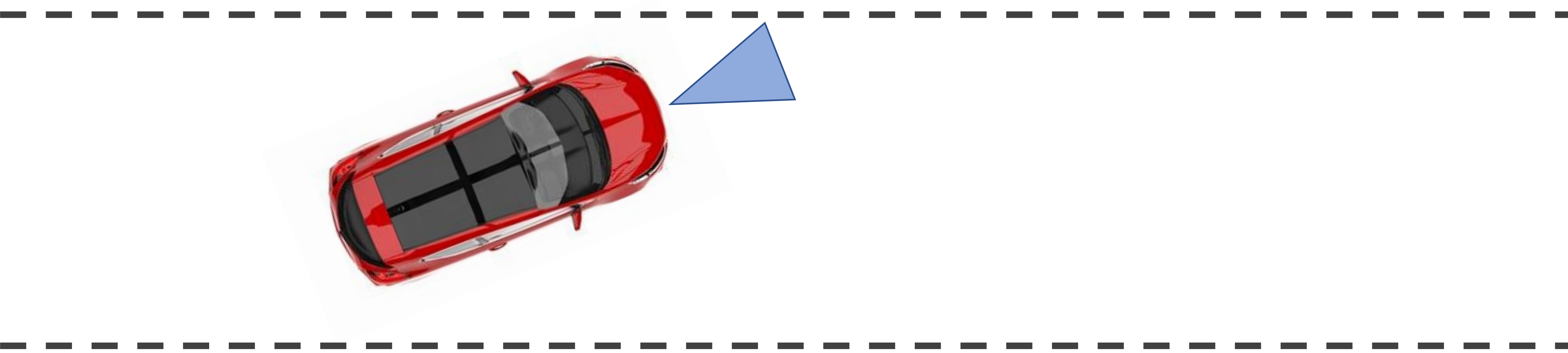
# Noise Injection

- ✓ Simple method to get corrective feedback
- ✓ Can work well in practice
- ✗ Dangerous for expert!
- ✗ Difficult to pick effective perturbations

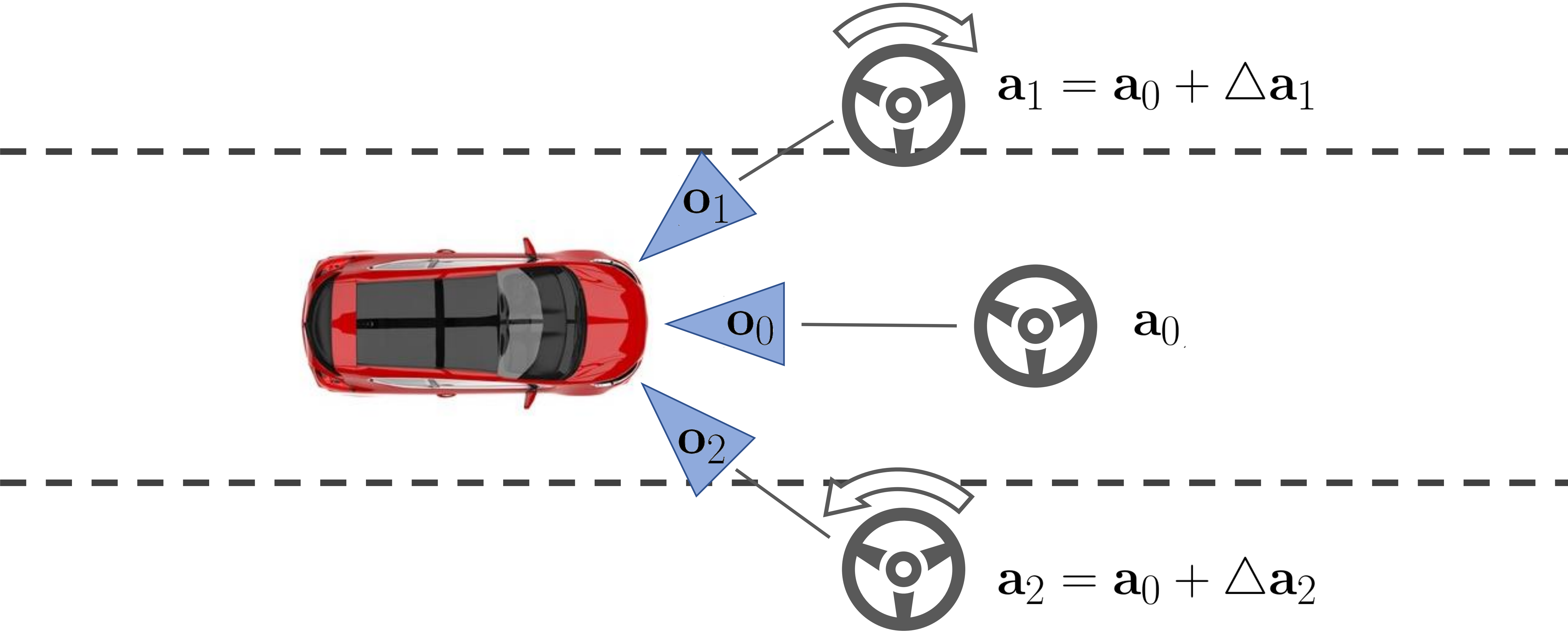
# Data Augmentation



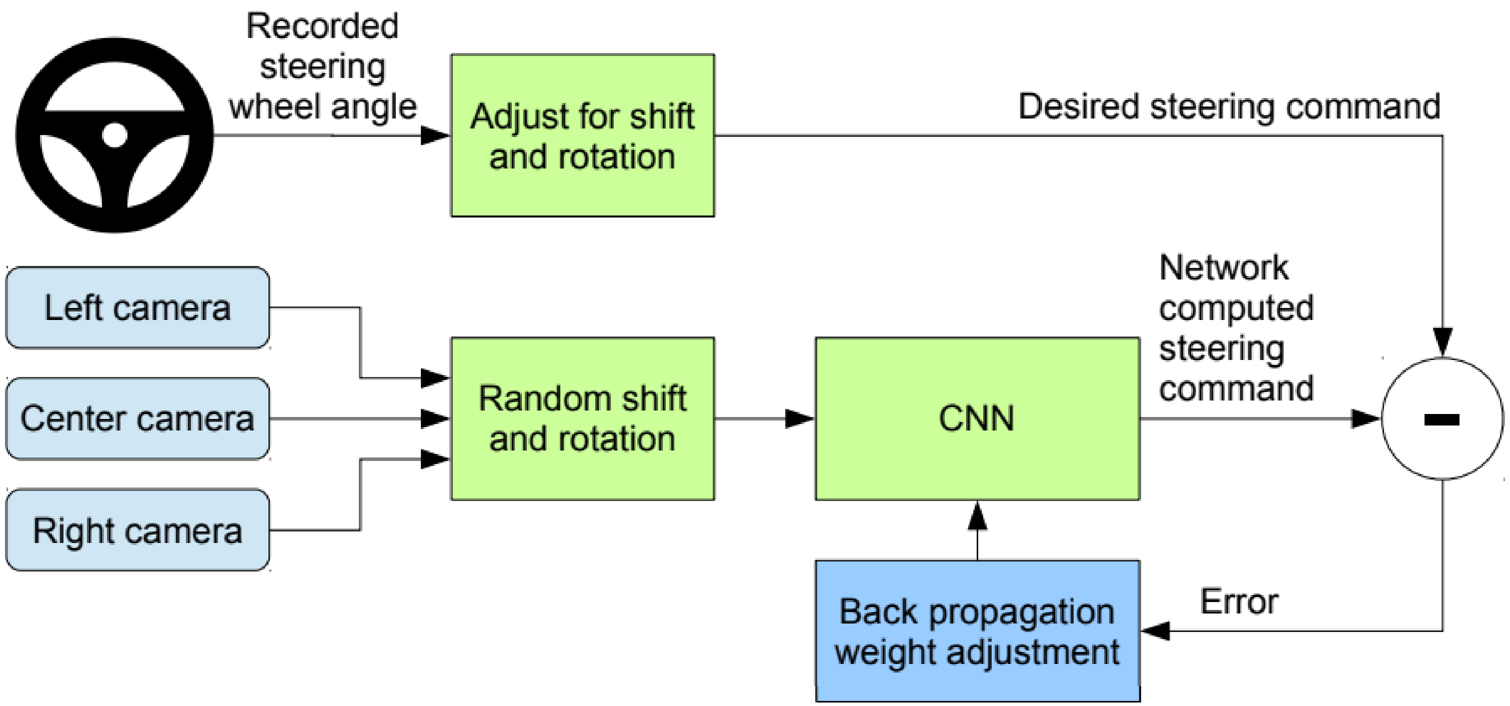
# Data Augmentation



# Data Augmentation



# Data Augmentation

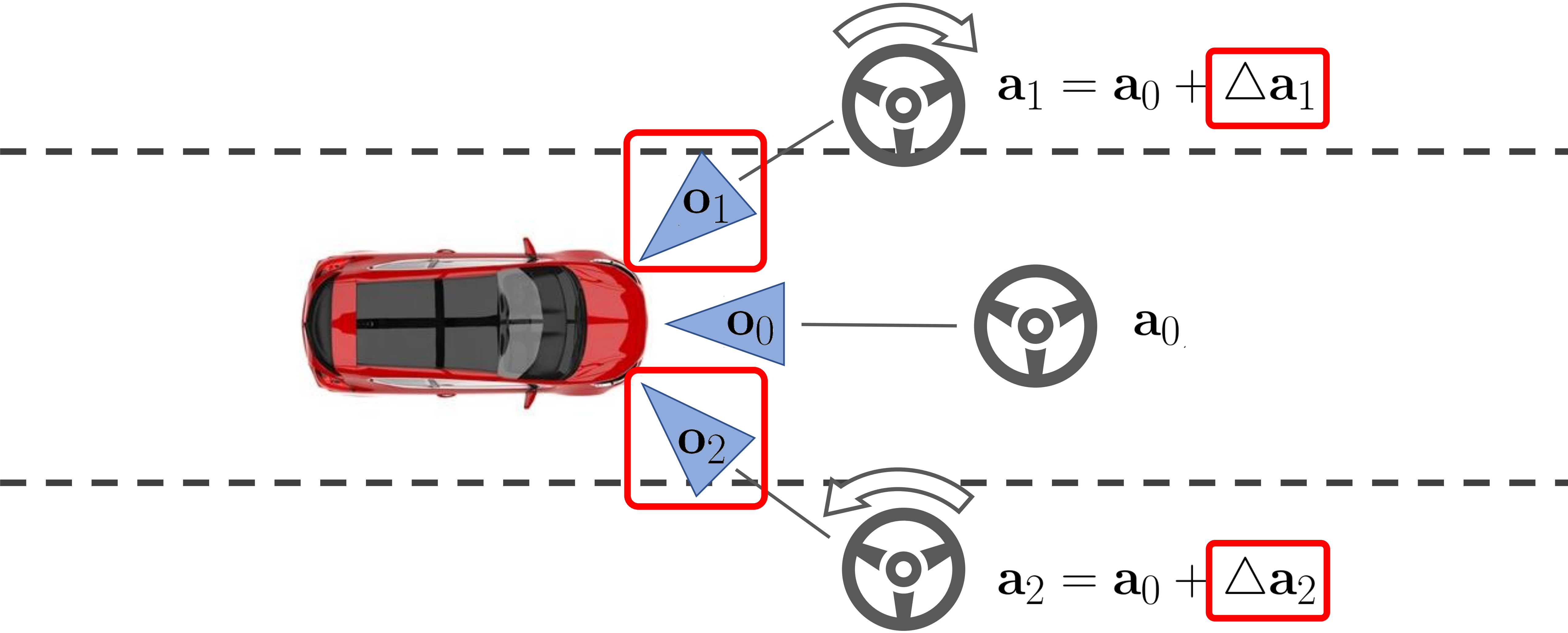



End to End Learning for Self-Driving Cars
[Bojarski et al. 2016]

# Data Augmentation


End to End Learning for Self-Driving Cars
[Bojarski et al. 2016]

# Data Augmentation



# Drift

- Expert is too good
- Lack of corrective feedback
- Policy inaccuracies
- Errors compound over time

# Theoretical Analysis

Analyze the number of mistakes $\pi$ makes over time

**Theorem 1.** The number of mistakes grow $O(\epsilon T^2)$

# Theoretical Analysis

Given dataset sampled from $p_{\text{data}}(\mathbf{s},\mathbf{a})$

$$\min_{\pi} \ \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim p_{\text{data}}(\mathbf{s},\mathbf{a})} \left[-\log\pi(\mathbf{a}|\mathbf{s})\right]$$

Such that

$$\pi\left(\mathbf{a} \neq \pi^{*}(\mathbf{s})|\mathbf{s}\right) \leq \epsilon \ \text{ for all } \ \mathbf{s} \sim p_{\text{data}}(\mathbf{s})$$

i.e. the probability of $\pi$ making a mistake is bounded.

Cost: $c(\mathbf{s}, \mathbf{a}) = \begin{cases} 0 & \text{if } \mathbf{a} = \pi^{*}(\mathbf{s}) \\ 1 & \text{otherwise} \end{cases}$

# Theoretical Analysis

Assume: $\pi\left(\mathbf{a} \neq \pi^{*}(\mathbf{s}) | \mathbf{s}\right) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$\underline{p_{\pi}^{t}(\mathbf{s})} = (1-\epsilon)^{t} p_{\text{data}}^{t}(\mathbf{s}) + (1-(1-\epsilon)^{t}) p_{\text{mistake}}^{t}(\mathbf{s})$$

probability of being in $\mathbf{s}$ after following $\pi$ for $t$ timesteps

# Theoretical Analysis

Assume: $\pi\left(\mathbf{a} \neq \pi^*(\mathbf{s}) | \mathbf{s}\right) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_\pi^t(\mathbf{s}) = \underline{\boxed{(1-\epsilon)^t} p_{\text{data}}^t(\mathbf{s})} + (1-(1-\epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

no mistakes in $t$ timsteps

# Theoretical Analysis

Assume: $\pi\left(\mathbf{a} \neq \pi^*(\mathbf{s}) | \mathbf{s}\right) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_\pi^t(\mathbf{s}) = (1-\epsilon)^t \underbrace{p_{\text{data}}^t(\mathbf{s})}_{\text{no mistakes in } t \text{ timsteps}} + \boxed{(1-(1-\epsilon)^t)} \underbrace{p_{\text{mistake}}^t(\mathbf{s})}_{\text{at least 1 mistakes in } t \text{ timsteps}}$$

# Theoretical Analysis

Assume: $\pi\left(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}\right) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_\pi^t(\mathbf{s}) = (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1-(1-\epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\underbrace{\sum_t E_{p_\pi^t(\mathbf{s})} E_{\pi(\mathbf{a}|\mathbf{s})} \left[c(\mathbf{s}, \mathbf{a})\right]}_{\text{expected cost}}$$

# Theoretical Analysis

Assume: $\pi\left(\mathbf{a} \neq \pi^*(\mathbf{s}) | \mathbf{s}\right) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_\pi^t(\mathbf{s}) = (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1-(1-\epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[c(\mathbf{s}, \mathbf{a})\right]$$

# Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_\pi^t(\mathbf{s}) = (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1-(1-\epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t \mathbb{E}_{\underline{p_\pi^t(\mathbf{s})}} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s},\mathbf{a})]$$

# Theoretical Analysis

Assume: $\pi\left(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}\right) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_\pi^t(\mathbf{s}) = (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1-(1-\epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[c(\mathbf{s}, \mathbf{a})\right]$$

# Theoretical Analysis

Assume: $\pi\left(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}\right) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_\pi^t(\mathbf{s}) = (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1-(1-\epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t E_{p_\pi^t(\mathbf{s})} E_{\pi(\mathbf{a}|\mathbf{s})} \left[ c(\mathbf{s}, \mathbf{a}) \right]$$

$$c(\mathbf{s},\mathbf{a}) = \begin{cases} 0 & \text{if } \mathbf{a} = \pi^*(\mathbf{s}) \\ 1 & \text{otherwise} \end{cases}$$

# Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_\pi^t(\mathbf{s}) = (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1-(1-\epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t \underline{E_{p_\pi^t(\mathbf{s})}} E_{\pi(\mathbf{a}|\mathbf{s})}[c(\mathbf{s},\mathbf{a})] = \sum_t \underline{\sum_{\mathbf{s}} p_\pi^t(\mathbf{s})} E_{\pi(\mathbf{a}|\mathbf{s})}[c(\mathbf{s},\mathbf{a})]$$

# Theoretical Analysis

Assume: $\pi\left(\mathbf{a} \neq \pi^*(\mathbf{s}) | \mathbf{s}\right) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_\pi^t(\mathbf{s}) = (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1-(1-\epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right] = \sum_t \sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right]$$

$$= \sum_t \sum_{\mathbf{s}} \left(p_\pi^t(\mathbf{s}) \underbrace{- p_{\text{data}}^t(\mathbf{s}) + p_{\text{data}}^t(\mathbf{s})}_{=0}\right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right]$$

# Theoretical Analysis

Assume: $\pi\left(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}\right) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_\pi^t(\mathbf{s}) = (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1-(1-\epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t E_{p_\pi^t(\mathbf{s})} E_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right] = \sum_t \sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) E_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right]$$

$$= \sum_t \sum_{\mathbf{s}} \left(p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) + p_{\text{data}}^t(\mathbf{s})\right) E_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right]$$

$$= \sum_t \sum_{\mathbf{s}} \underline{p_{\text{data}}^t(\mathbf{s}) E_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right]} + \sum_t \sum_{\mathbf{s}} \left(p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s})\right) E_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right]$$

# Theoretical Analysis

Assume: $\pi\left(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}\right) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_\pi^t(\mathbf{s}) = (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1-(1-\epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right] = \sum_t \sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right]$$

$$= \sum_t \sum_{\mathbf{s}} \left(p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) + p_{\text{data}}^t(\mathbf{s})\right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right]$$

$$= \sum_t \sum_{\mathbf{s}} p_{\text{data}}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right] + \sum_t \sum_{\mathbf{s}} \underline{\left(p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s})\right)} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right]$$

# Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_\pi^t(\mathbf{s}) = (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1-(1-\epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s},\mathbf{a})] = \sum_t \sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s},\mathbf{a})]$$

$$= \sum_t \sum_{\mathbf{s}} \left( p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) + p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s},\mathbf{a})]$$

$$= \sum_t \sum_{\mathbf{s}} p_{\text{data}}^t(\mathbf{s}) \underbrace{\mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s},\mathbf{a})]}_{\leq \epsilon} + \sum_t \sum_{\mathbf{s}} \left( p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} [c(\mathbf{s},\mathbf{a})]$$

# Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_\pi^t(\mathbf{s}) = (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1-(1-\epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t E_{p_\pi^t(\mathbf{s})} E_{\pi(\mathbf{a}|\mathbf{s})}[c(\mathbf{s},\mathbf{a})] = \sum_t \sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) E_{\pi(\mathbf{a}|\mathbf{s})}[c(\mathbf{s},\mathbf{a})]$$

$$= \sum_t \sum_{\mathbf{s}} \left(p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) + p_{\text{data}}^t(\mathbf{s})\right) E_{\pi(\mathbf{a}|\mathbf{s})}[c(\mathbf{s},\mathbf{a})]$$

$$= \sum_t \underbrace{\sum_{\mathbf{s}} p_{\text{data}}^t(\mathbf{s}) E_{\pi(\mathbf{a}|\mathbf{s})}[c(\mathbf{s},\mathbf{a})]}_{\leq \epsilon} + \sum_t \sum_{\mathbf{s}} \left(p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s})\right) E_{\pi(\mathbf{a}|\mathbf{s})}[c(\mathbf{s},\mathbf{a})]$$

# Theoretical Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$p_\pi^t(\mathbf{s}) = (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1-(1-\epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}[c(\mathbf{s},\mathbf{a})] = \sum_t \sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}[c(\mathbf{s},\mathbf{a})]$$

$$= \sum_t \sum_{\mathbf{s}} \left(p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) + p_{\text{data}}^t(\mathbf{s})\right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}[c(\mathbf{s},\mathbf{a})]$$

$$= \underbrace{\sum_t \sum_{\mathbf{s}} p_{\text{data}}^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}[c(\mathbf{s},\mathbf{a})]}_{\leq \epsilon T} + \sum_t \sum_{\mathbf{s}} \left(p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s})\right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}[c(\mathbf{s},\mathbf{a})]$$

# Theoretical Analysis

Assume: $\pi\left(\mathbf{a} \neq \pi^{*}(\mathbf{s}) | \mathbf{s}\right) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$$\boxed{p_{\pi}^{t}(\mathbf{s}) = (1-\epsilon)^{t} p_{\text{data}}^{t}(\mathbf{s}) + (1-(1-\epsilon)^{t}) p_{\text{mistake}}^{t}(\mathbf{s})}$$

$$\sum_{t} \mathbb{E}_{p_{\pi}^{t}(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right] = \sum_{t}\sum_{\mathbf{s}} p_{\pi}^{t}(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right]$$

$$= \sum_{t}\sum_{\mathbf{s}} \left(p_{\pi}^{t}(\mathbf{s}) - p_{\text{data}}^{t}(\mathbf{s}) + p_{\text{data}}^{t}(\mathbf{s})\right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right]$$

$$= \sum_{t}\sum_{\mathbf{s}} p_{\text{data}}^{t}(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right] + \sum_{t}\sum_{\mathbf{s}} \left(p_{\pi}^{t}(\mathbf{s}) - p_{\text{data}}^{t}(\mathbf{s})\right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right]$$

$$\leq \epsilon T + \underbrace{\sum_{t}\sum_{\mathbf{s}} \left(p_{\pi}^{t}(\mathbf{s}) - p_{\text{data}}^{t}(\mathbf{s})\right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right]}_{?}$$

# Theoretical Analysis

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) = \sum_{\mathbf{s}} (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + \left(1-(1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - \underline{(1-\epsilon)^t p_{\text{data}}^t(\mathbf{s})} = \sum_{\mathbf{s}} \left(1-(1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

# Theoretical Analysis

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) = \sum_{\mathbf{s}} (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + \left(1-(1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1-(1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) - \underline{\left(1-(1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s})} = \sum_{\mathbf{s}} \left(1-(1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s}) - \underline{\left(1-(1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s})}$$

# Theoretical Analysis

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) = \sum_{\mathbf{s}} (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - \underline{(1-\epsilon)^t p_{\text{data}}^t(\mathbf{s})} - \underline{\left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s})} = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s}) - \left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s})$$

# Theoretical Analysis

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) = \sum_{\mathbf{s}} (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) - \left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \underline{\left(1 - (1-\epsilon)^t\right)} p_{\text{mistake}}^t(\mathbf{s}) - \underline{\left(1 - (1-\epsilon)^t\right)} p_{\text{data}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s})$$

# Theoretical Analysis

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) = \sum_{\mathbf{s}} (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) - \left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1 - (1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s}) - \left(1 - (1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s})$$

$$\begin{aligned} \sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) &= \left(1 - (1-\epsilon)^t\right) \sum_{\mathbf{s}} p_{\text{mistake}}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \\ &\leq \left(1 - (1-\epsilon)^t\right) \underbrace{\sum_{\mathbf{s}} \left| p_{\text{mistake}}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right|}_{\text{total variation distance} \leq 2} \end{aligned}$$

# Theoretical Analysis

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) = \sum_{\mathbf{s}} (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + \left(1-(1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1-(1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s})$$

$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - (1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) - \left(1-(1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s}) = \sum_{\mathbf{s}} \left(1-(1-\epsilon)^t\right) p_{\text{mistake}}^t(\mathbf{s}) - \left(1-(1-\epsilon)^t\right) p_{\text{data}}^t(\mathbf{s})$$

$$\begin{aligned}
\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) &= \left(1-(1-\epsilon)^t\right) \sum_{\mathbf{s}} p_{\text{mistake}}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \\
&\leq \left(1-(1-\epsilon)^t\right) \sum_{\mathbf{s}} \left| p_{\text{mistake}}^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right| \\
&\leq 2\left(1-(1-\epsilon)^t\right) \\
&\leq 2\epsilon t
\end{aligned}$$

Note: $(1-\epsilon)^t \geq 1-\epsilon t \qquad \text{for } \epsilon \in [0,1]$

# Theoretical Analysis

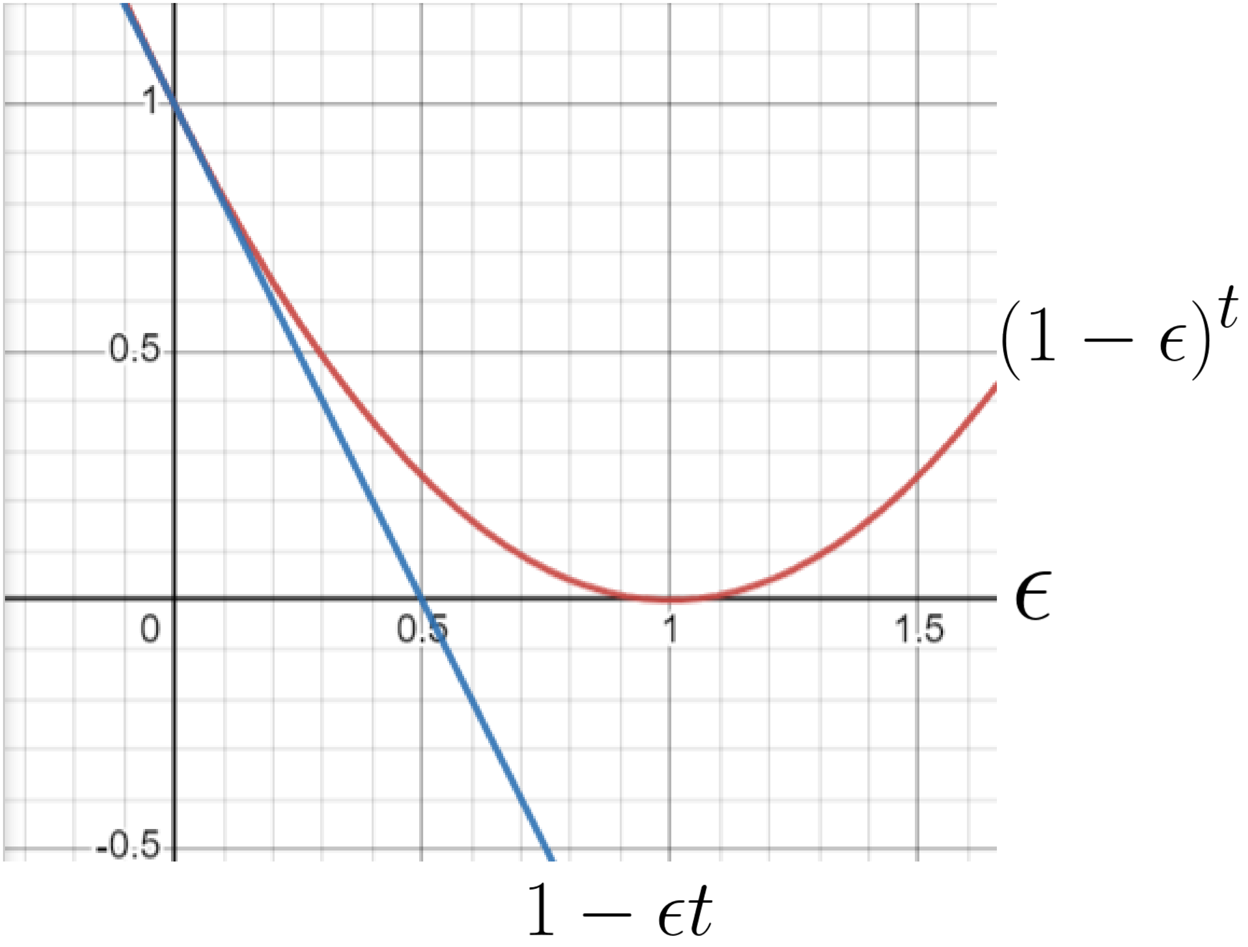
$$\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \leq 2\left(1-(1-\epsilon)^t\right)$$

$$\leq 2\left(1-(1-\epsilon t)\right)$$

$$\leq 2\epsilon t$$

$$\boxed{\sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \leq 2\epsilon t}$$

Note: $(1-\epsilon)^t \geq 1-\epsilon t \qquad \text{for } \epsilon \in [0,1]$



# Theoretical Analysis

$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[ c(\mathbf{s}, \mathbf{a}) \right] = \sum_t \sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[ c(\mathbf{s}, \mathbf{a}) \right]$$

$$\leq \epsilon T + \sum_t \underbrace{\sum_{\mathbf{s}} \left( p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right)}_{\leq 2\epsilon t} \underbrace{\mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[ c(\mathbf{s}, \mathbf{a}) \right]}_{\leq 1}$$

# Theoretical Analysis

$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[ c(\mathbf{s}, \mathbf{a}) \right] = \sum_t \sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[ c(\mathbf{s}, \mathbf{a}) \right]$$

$$\leq \epsilon T + \sum_t \sum_{\mathbf{s}} \left( p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[ c(\mathbf{s}, \mathbf{a}) \right]$$

$$\leq \epsilon T + \underline{\sum_t 2\epsilon t}$$

# Theoretical Analysis

$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[ c(\mathbf{s}, \mathbf{a}) \right] = \sum_t \sum_{\mathbf{s}} p_\pi^t(\mathbf{s}) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[ c(\mathbf{s}, \mathbf{a}) \right]$$

$$\leq \epsilon T + \sum_t \sum_{\mathbf{s}} \left( p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right) \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[ c(\mathbf{s}, \mathbf{a}) \right]$$

$$\leq \epsilon T + \sum_t 2\epsilon t$$

$$\leq \epsilon T + 2\epsilon T^2 \in \boxed{O(\epsilon T^2)}$$

# Worst Case

$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[ c(\mathbf{s}, \mathbf{a}) \right] \leq \epsilon T + \boxed{2\epsilon T^2}$$

# Worst Case

$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})}\mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right] \leq \epsilon T + \boxed{2\epsilon T^2}$$

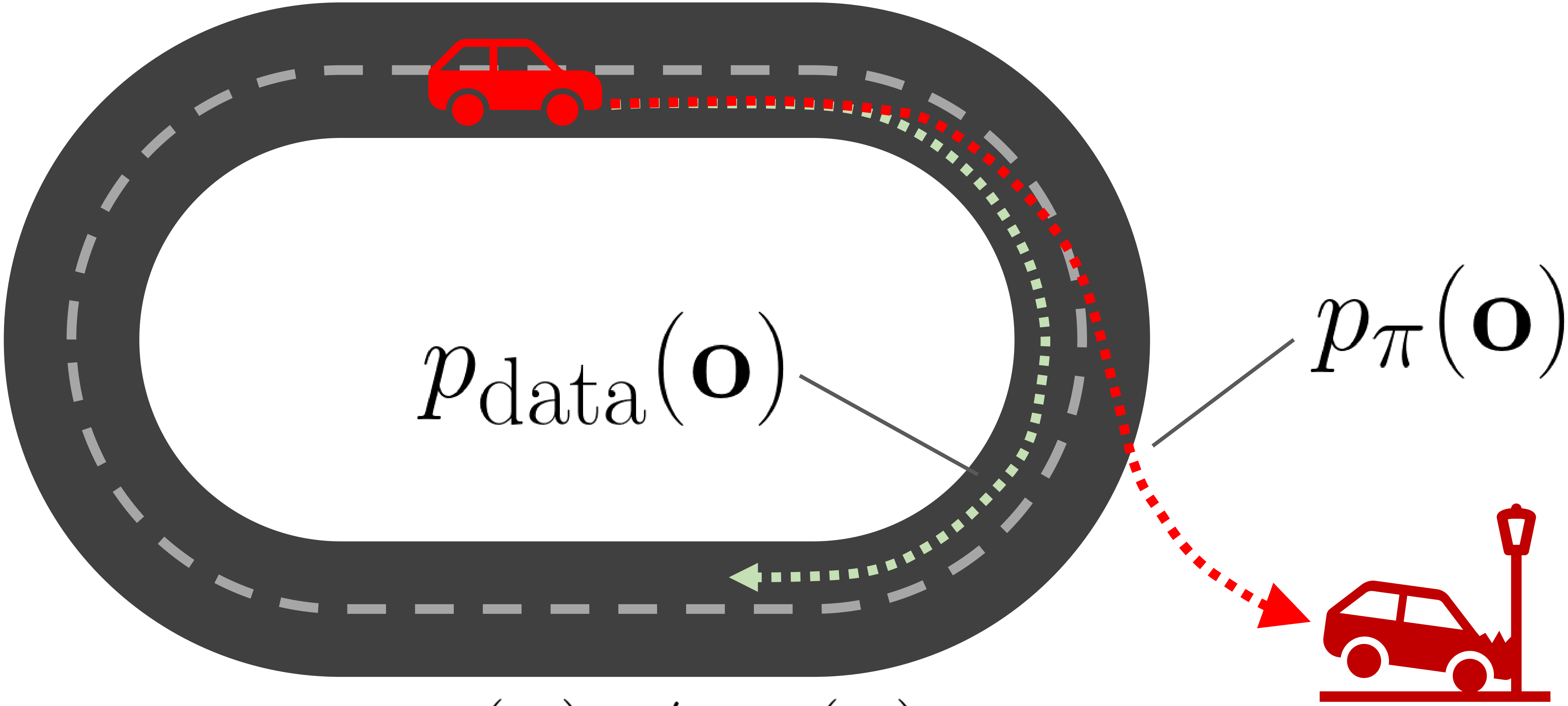
# Distribution Shift

$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[ c(\mathbf{s}, \mathbf{a}) \right] \leq \epsilon T + \boxed{2\epsilon T^2}$$

$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[ c(\mathbf{s}, \mathbf{a}) \right] \leq \epsilon T + \sum_t \sum_{\mathbf{s}} \boxed{\left( p_\pi^t(\mathbf{s}) - p_{\text{data}}^t(\mathbf{s}) \right)} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})} \left[ c(\mathbf{s}, \mathbf{a}) \right]$$

$$p_\pi^t(\mathbf{s}) \neq p_{\text{data}}^t(\mathbf{s})$$

# Distribution Shift




$$p_{\text{data}}(\mathbf{o}) \neq p_\pi(\mathbf{o})$$

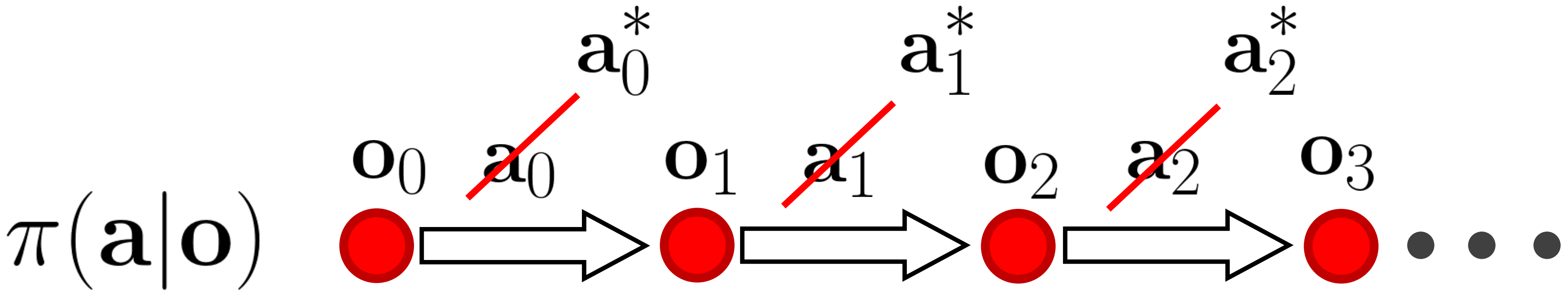
# Dataset Aggregation

Can we make $p_{\mathrm{data}}(\mathbf{o}) = p_\pi(\mathbf{o})$ ?

Key idea:

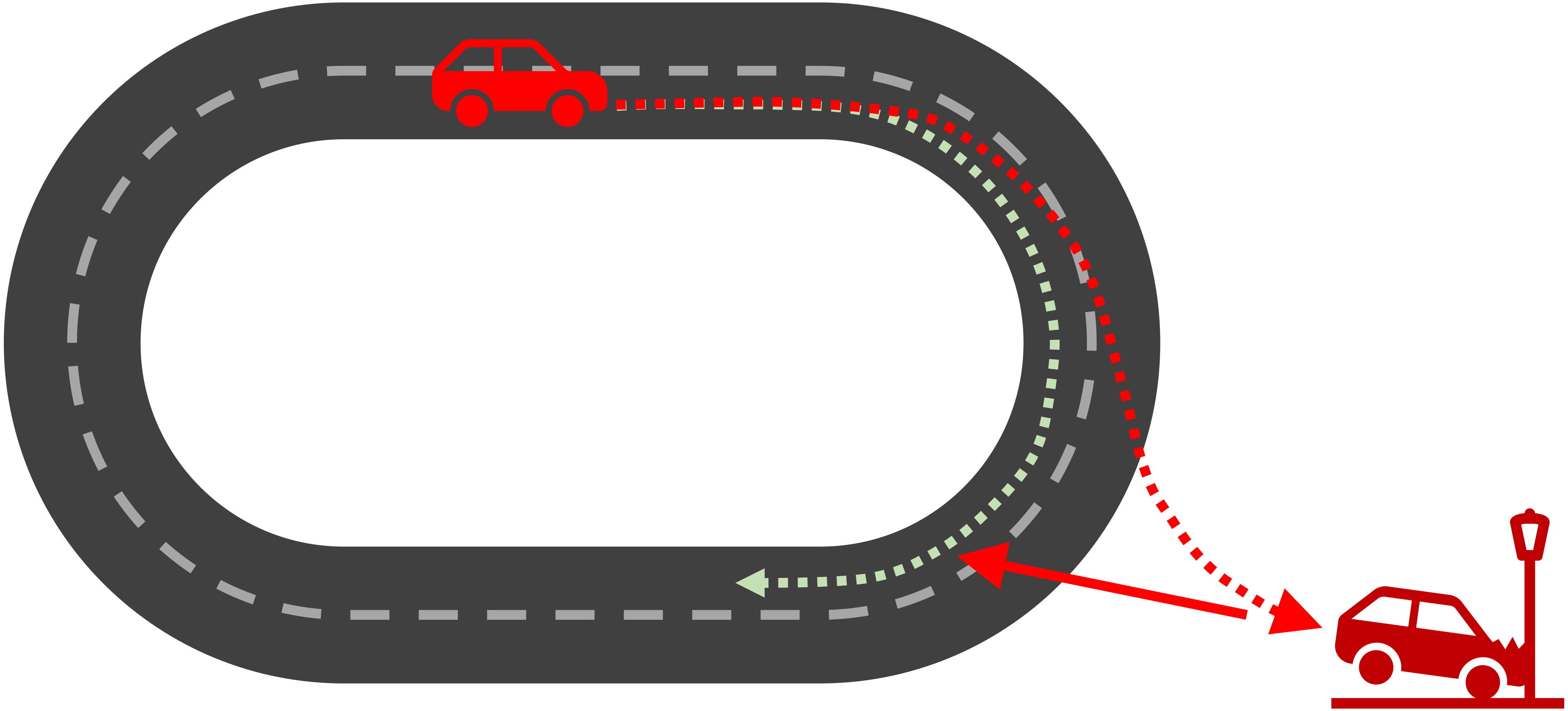
- Collect observations from $p_\pi(\mathbf{o})$ instead of $p_{\mathrm{data}}(\mathbf{o})$
- Label actions with expert
- DAgger: Dataset Aggregation [Ross et al. 2011]

# DAgger




Train with $(\mathbf{o}_i, \mathbf{a}_i^*)$

# DAgger



# DAgger

**ALGORITHM:** DAgger

1: **for** iteration $i = 0, ..., k-1$ **do**
2: train $\pi(\mathbf{a}|\mathbf{o})$ from dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, ...\}$
3: run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, ...\}$
4: Label $\mathcal{D}_\pi$ with actions $\mathbf{a}_i$ from expert
5: Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
6: **end for**

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning [Ross et al. 2011]

# DAgger

**ALGORITHM:** DAgger

1: **for** iteration $i = 0, ..., k-1$ **do**
2: train $\pi(\mathbf{a}|\mathbf{o})$ from dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, ...\}$
3: run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, ...\}$
4: Label $\mathcal{D}_\pi$ with actions $\mathbf{a}_i$ from expert
5: Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
6: **end for**

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning [Ross et al. 2011]

# DAgger

**ALGORITHM:** DAgger

**for** iteration $i = 0, ..., k-1$ **do**
    train $\pi(\mathbf{a}|\mathbf{o})$ from dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, ...\}$
    run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, ...\}$
    Label $\mathcal{D}_\pi$ with actions $\mathbf{a}_i$ from expert
    Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
**end for**

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
[Ross et al. 2011]

# DAgger

**ALGORITHM:** DAgger

1: **for** iteration $i = 0, ..., k-1$ **do**
2: train $\pi(\mathbf{a}|\mathbf{o})$ from dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, ...\}$
3: run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, ...\}$
4: Label $\mathcal{D}_\pi$ with actions $\mathbf{a}_i$ from expert
5: Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
6: **end for**

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning [Ross et al. 2011]

# DAgger

**ALGORITHM:** DAgger

1: **for** iteration $i = 0, ..., k-1$ **do**
2: train $\pi(\mathbf{a}|\mathbf{o})$ from dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, ...\}$
3: run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, ...\}$
4: Label $\mathcal{D}_\pi$ with actions $\mathbf{a}_i$ from expert
5: Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
6: **end for**

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning [Ross et al. 2011]

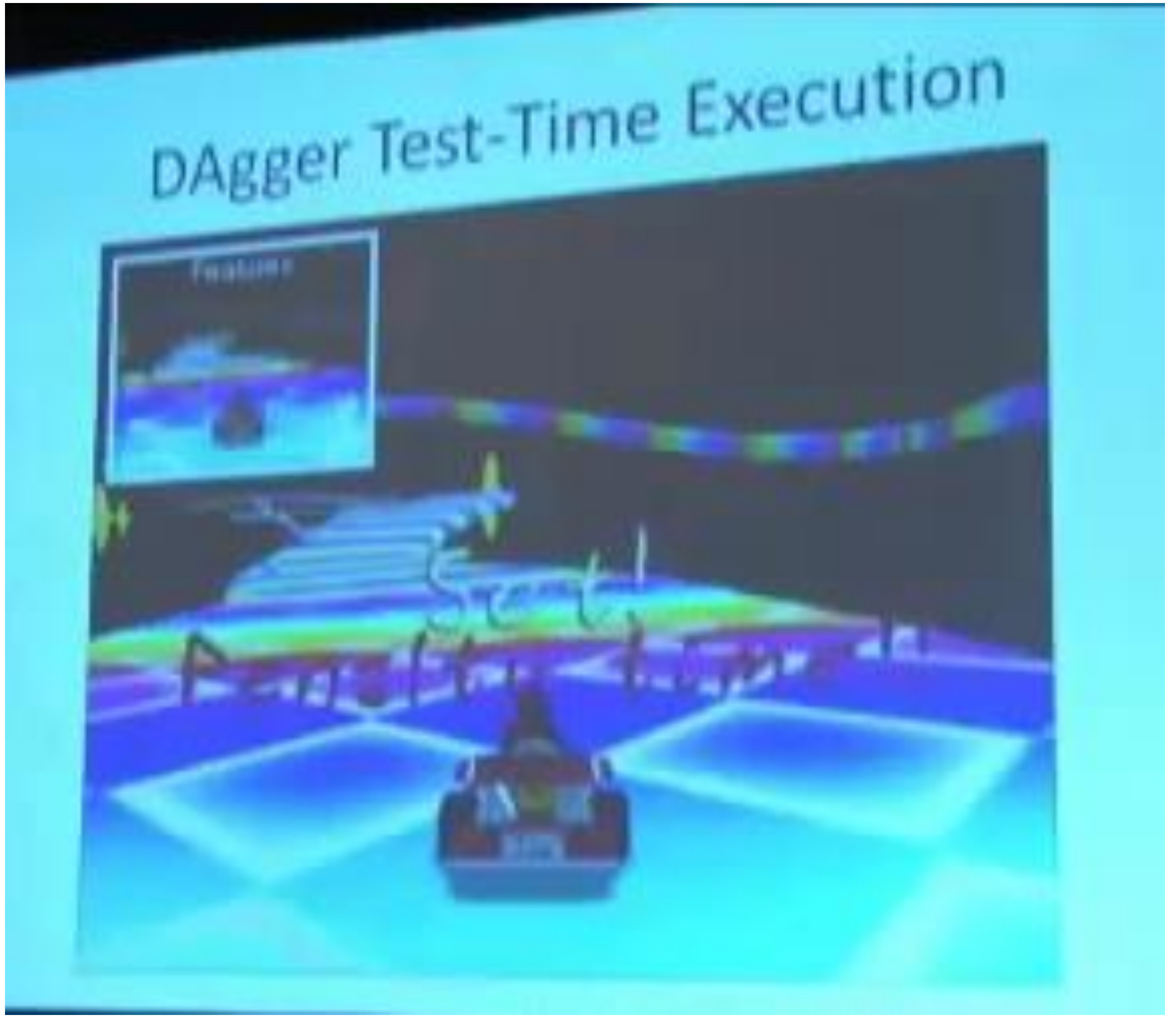
# DAgger

**ALGORITHM:** DAgger

1: **for** iteration $i = 0, ..., k-1$ **do**
2: train $\pi(\mathbf{a}|\mathbf{o})$ from dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, ...\}$
3: run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, ...\}$
4: Label $\mathcal{D}_\pi$ with actions $\mathbf{a}_i$ from expert
5: Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
6: **end for**

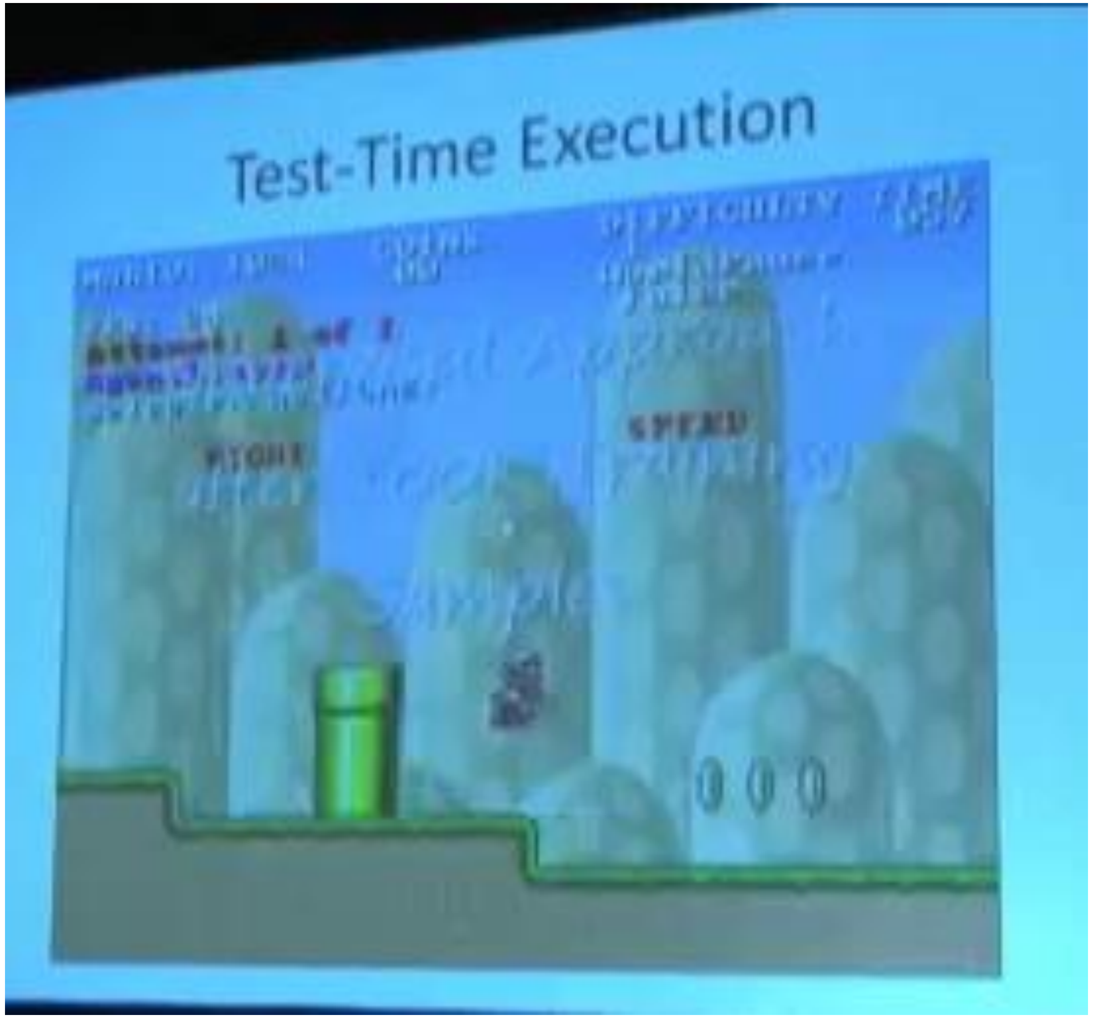
A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning [Ross et al. 2011]

# DAgger




A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
[Ross et al. 2011]

# DAgger




A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
[Ross et al. 2011]

# DAgger Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$p_{\text{data}}(\mathbf{s}) = p_\pi(\mathbf{s})!$

$$p_\pi^t(\mathbf{s}) = \underline{(1-\epsilon)^t p_{\text{data}}^t(\mathbf{s}) + (1-(1-\epsilon)^t) p_{\text{mistake}}^t(\mathbf{s})}$$

$$= p_{\text{data}}^t(\mathbf{s})$$

# DAgger Analysis

Assume: $\pi(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$p_{\text{data}}(\mathbf{s}) = p_\pi(\mathbf{s})!$

$p_\pi^t(\mathbf{s}) = p_{\text{data}}^t(\mathbf{s})$

$$\sum_t \mathbb{E}_{p_\pi^t(\mathbf{s})} \mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}[c(\mathbf{s},\mathbf{a})] = \sum_t \mathbb{E}_{p_{\text{data}}^t(\mathbf{s})} \underbrace{\mathbb{E}_{\pi(\mathbf{a}|\mathbf{s})}[c(\mathbf{s},\mathbf{a})]}_{\leq \epsilon}$$

# DAgger Analysis

Assume: $\pi\left(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s}\right) \leq \epsilon$ for all $\mathbf{s} \sim p_{\text{data}}(\mathbf{s})$

$p_{\text{data}}(\mathbf{s}) = p_\pi(\mathbf{s})!$

$p_\pi^t(\mathbf{s}) = p_{\text{data}}^t(\mathbf{s})$

$$\sum_t E_{p_\pi^t(\mathbf{s})} E_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right] = \sum_t E_{p_{\text{data}}^t(\mathbf{s})} E_{\pi(\mathbf{a}|\mathbf{s})}\left[c(\mathbf{s},\mathbf{a})\right]$$

$$\leq \sum_t \epsilon$$

$$\leq \epsilon T \in \boxed{O(\epsilon T)}$$

# DAgger

**ALGORITHM:** DAgger

**for** iteration $i = 0, ..., k-1$ **do**
  train $\pi(\mathbf{a}|\mathbf{o})$ from expert dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, ...\}$
  run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, ...\}$
  Label $\mathcal{D}_\pi$ with actions $\mathbf{a}_i$ from expert
  Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
**end for**

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning [Ross et al. 2011]

# DAgger

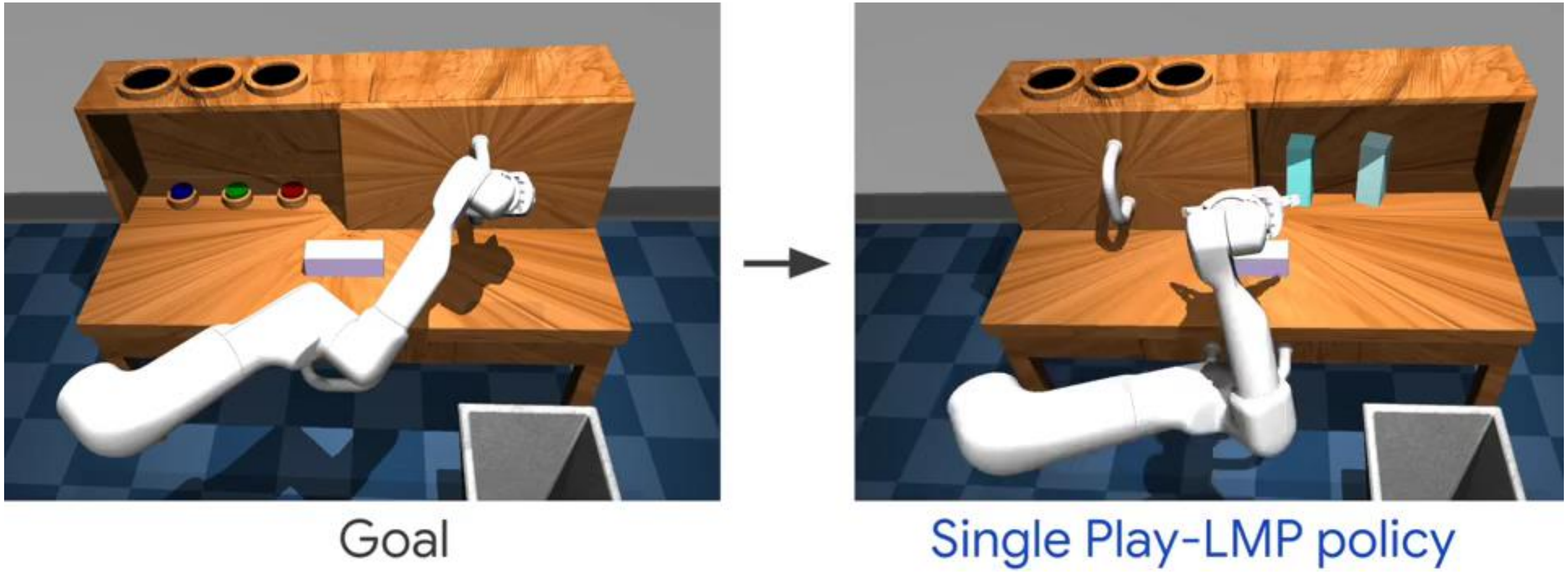
**ALGORITHM:** DAgger

**for** iteration $i = 0, \ldots, k-1$ **do**
  train $\pi(\mathbf{a}|\mathbf{o})$ from expert dataset $\mathcal{D} = \{\mathbf{o}_0, \mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_0, \ldots\}$
  run $\pi(\mathbf{a}|\mathbf{o})$ to collect dataset $\mathcal{D}_\pi = \{\mathbf{o}_0, \mathbf{o}_1, \ldots\}$
  Label $\mathcal{D}_\pi$ with actions $\mathbf{a}_i$ from expert
  Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$
**end for**

A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning [Ross et al. 2011]

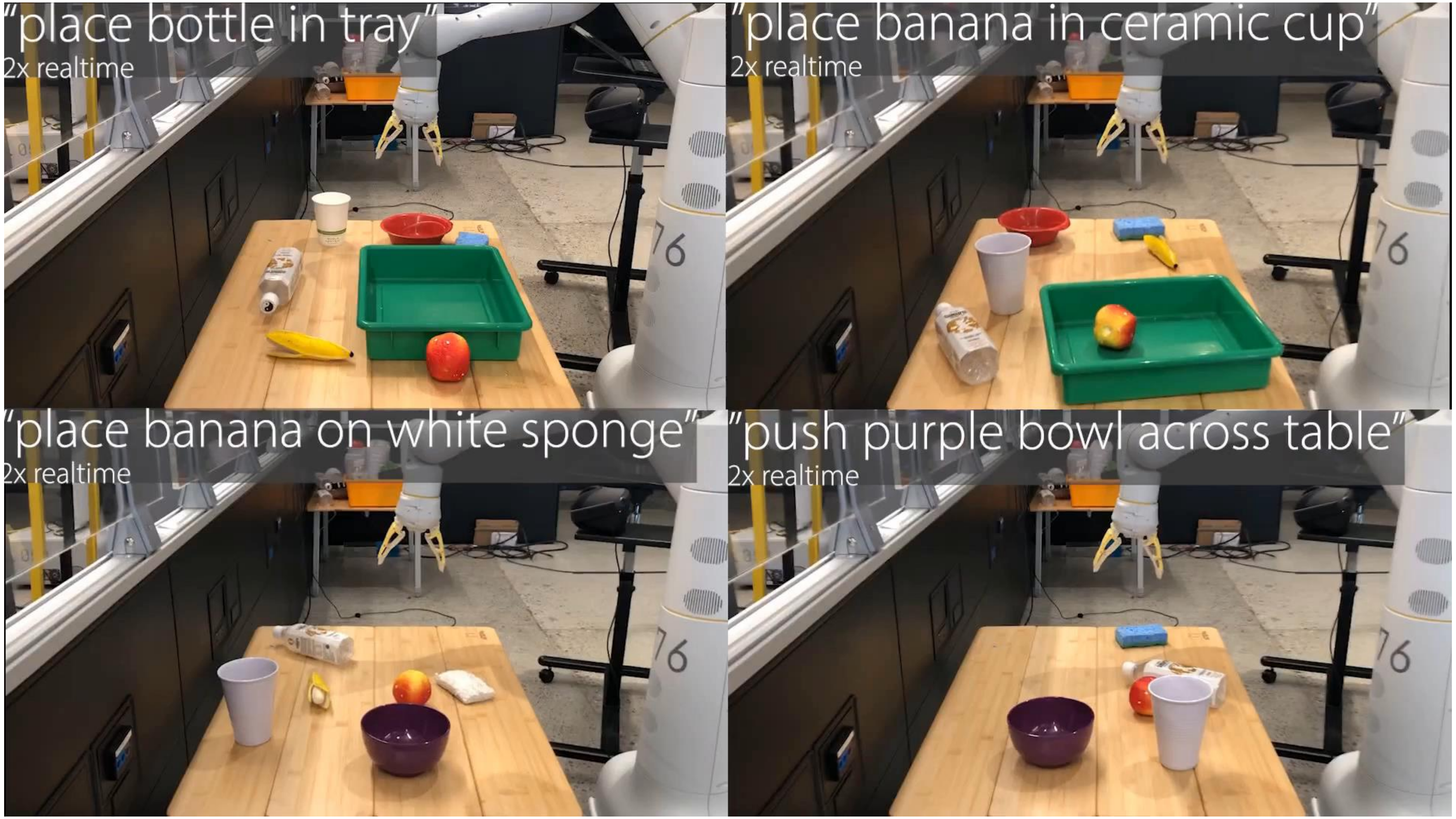
# Applications

# Applications




Learning Latent Plans from Play
[Lynch et al. 2019]

# Applications



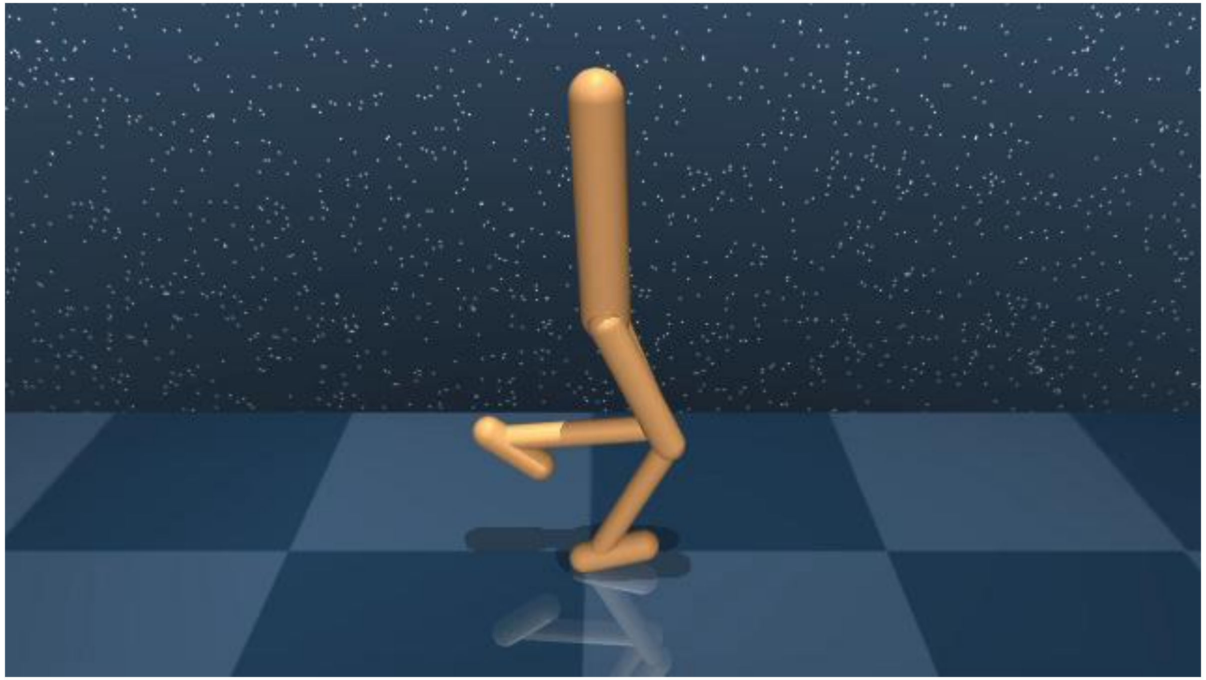

BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning
[Jang et al. 2021]

# Summary

- Behavioral Cloning
- Drift
- Theoretical Analysis
- DAgger
- Applications

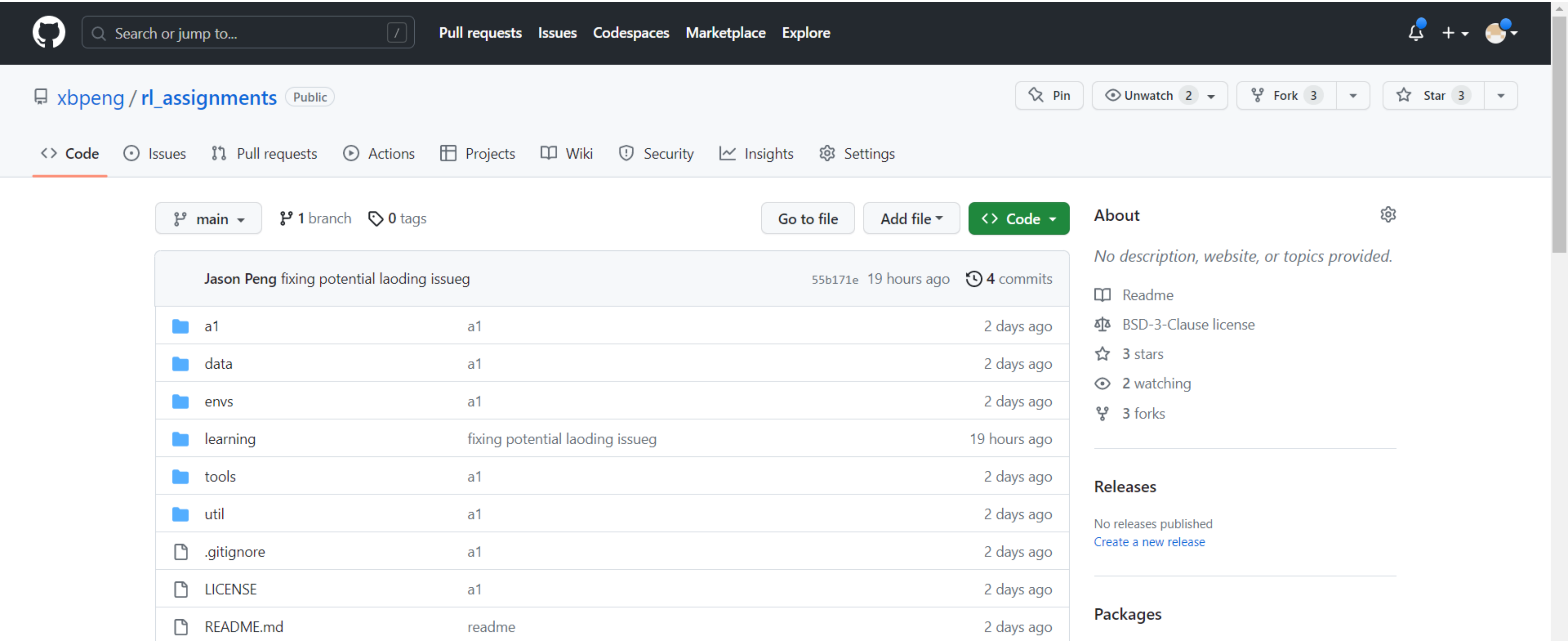
# Assignment 1: Behavioral Cloning

Cheetah



Walker

# Assignment 1: Behavioral Cloning




github.com/xbpeng/rl_assignments