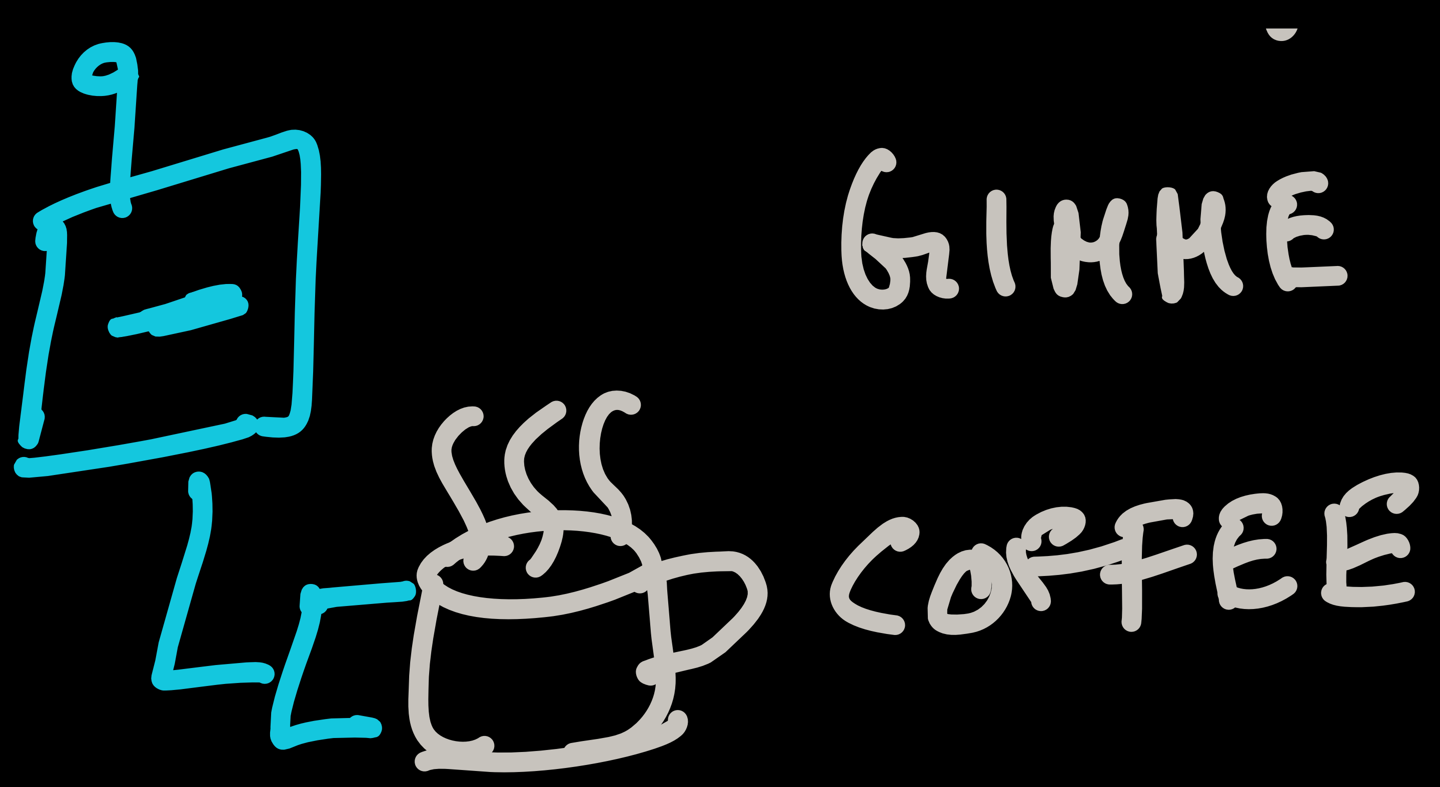
Imitation Learning: Feedback and Covariate Shift

Sanjiban Choudhury

Cornell Bowers CIS
Computer Science

Programming a task ...

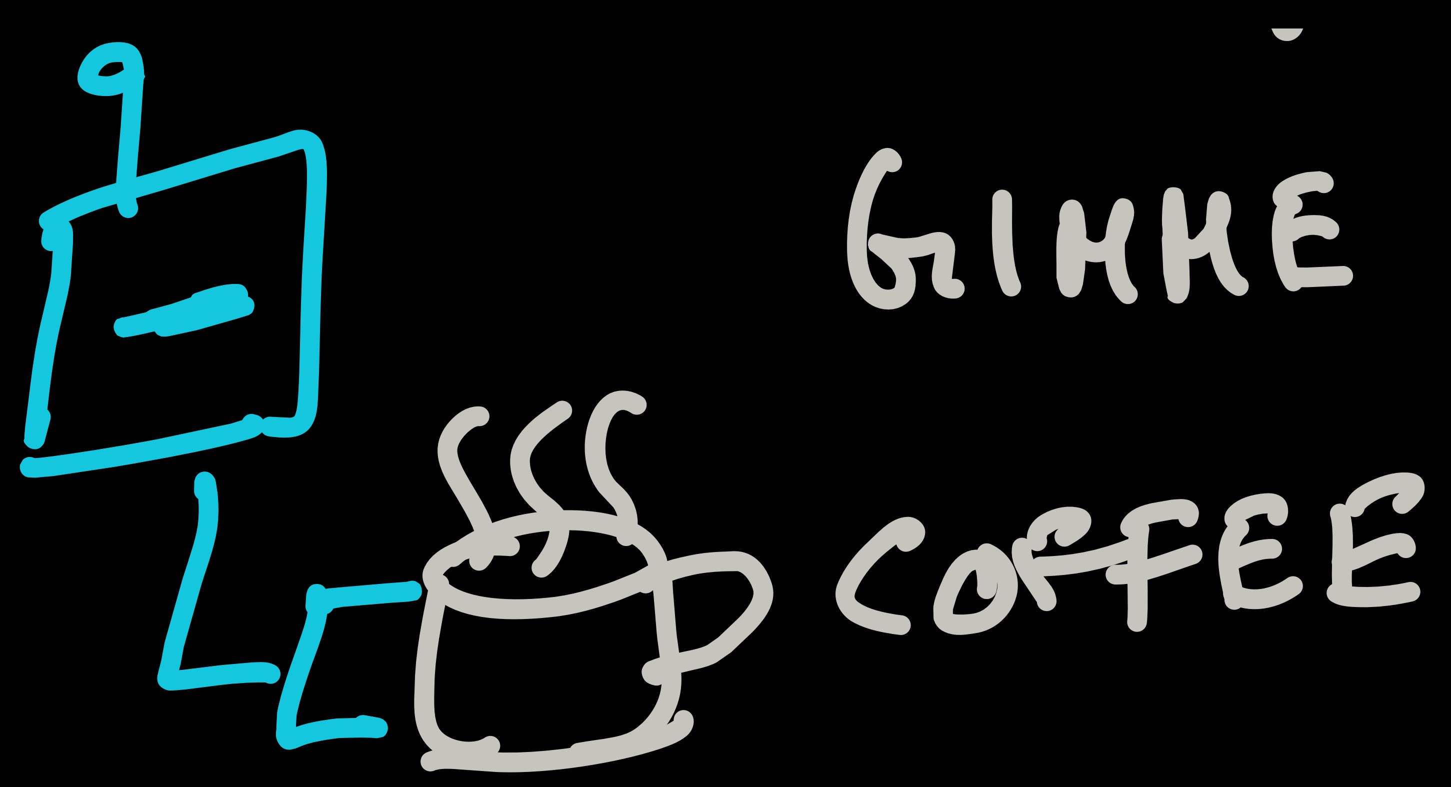
tell the robot to make coffee ..



robot burns down the house!

Programming a task ...

tell the robot to make coffee ..



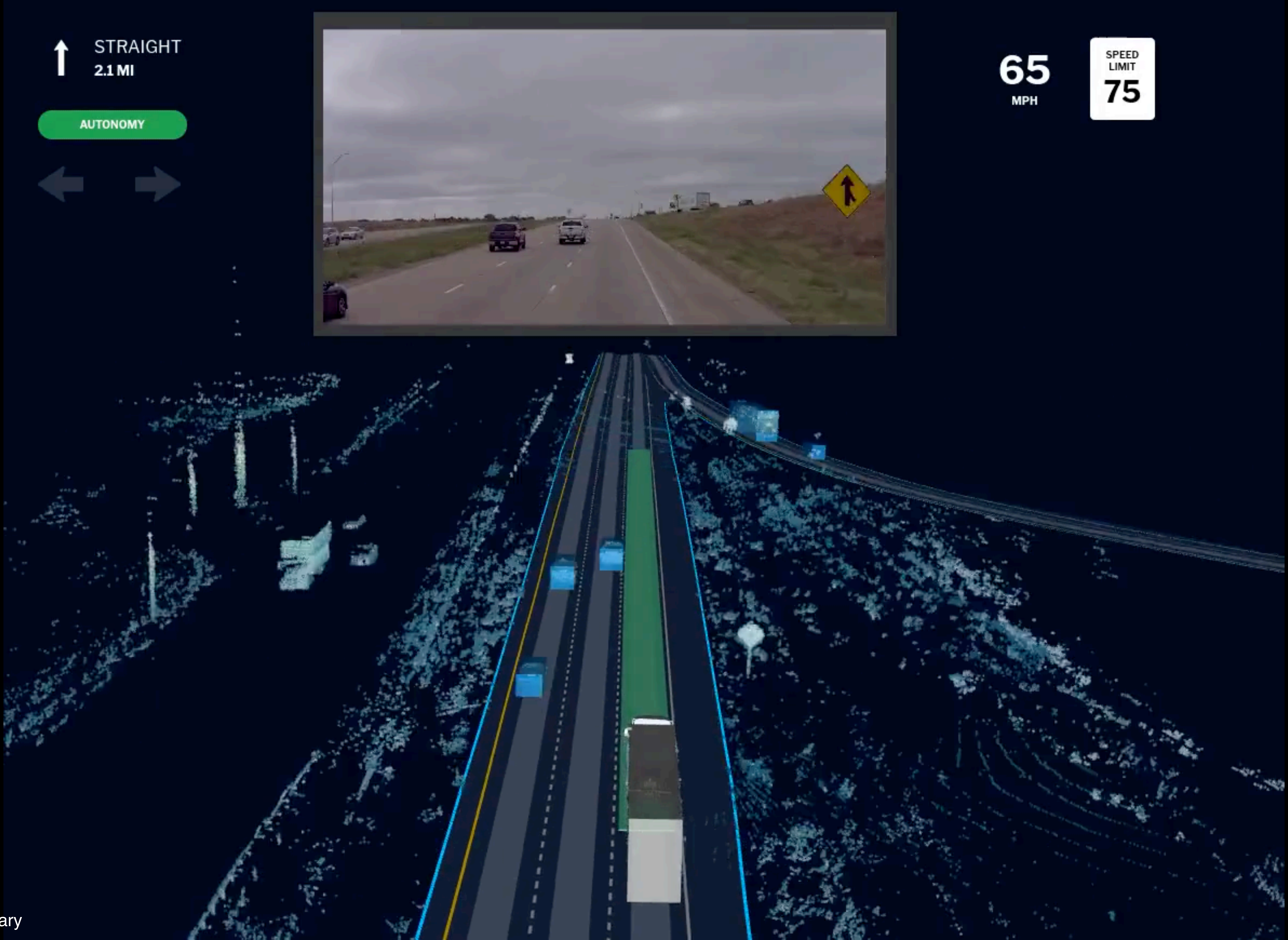
DON'T ...

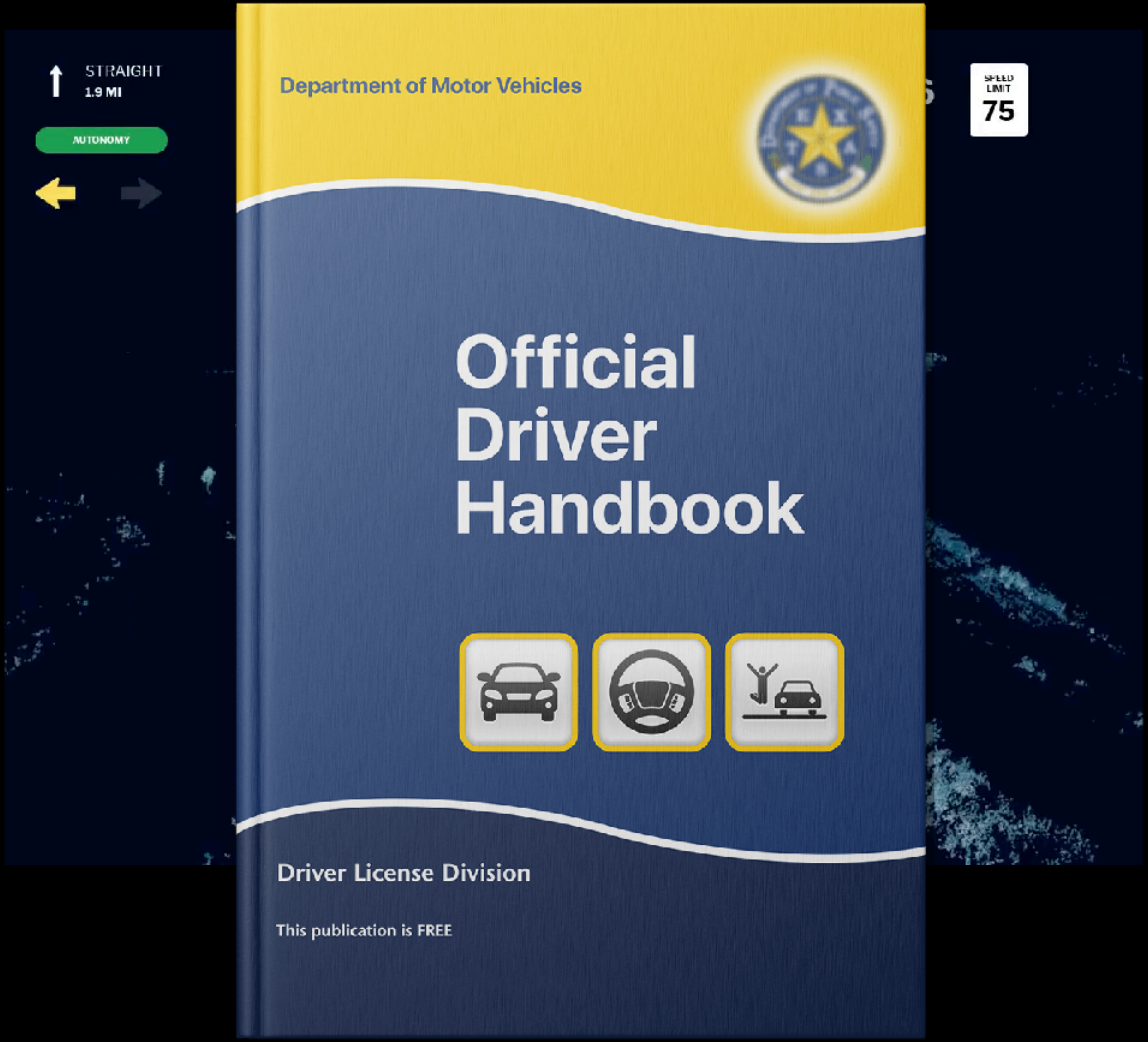
burn down the house

steal the neighbors coffee

don't make a mess

⋮


STRAIGHT
2.1 MI
AUTONOMY
65
MPH
SPEED LIMIT
75


STRAIGHT
1.9 MI
AUTONOMY
SPEED LIMIT
75
Department of Motor Vehicles
Official Driver Handbook
Driver License Division
This publication is FREE

The implicit rules of human driving

Implicit rules in a gridlocked intersection

Explicitly programming rules may be tedious …

… but rules are implicit in how we drive everyday!

Implicitly program robots

via

imitation learning

Imitation learning is *everywhere*

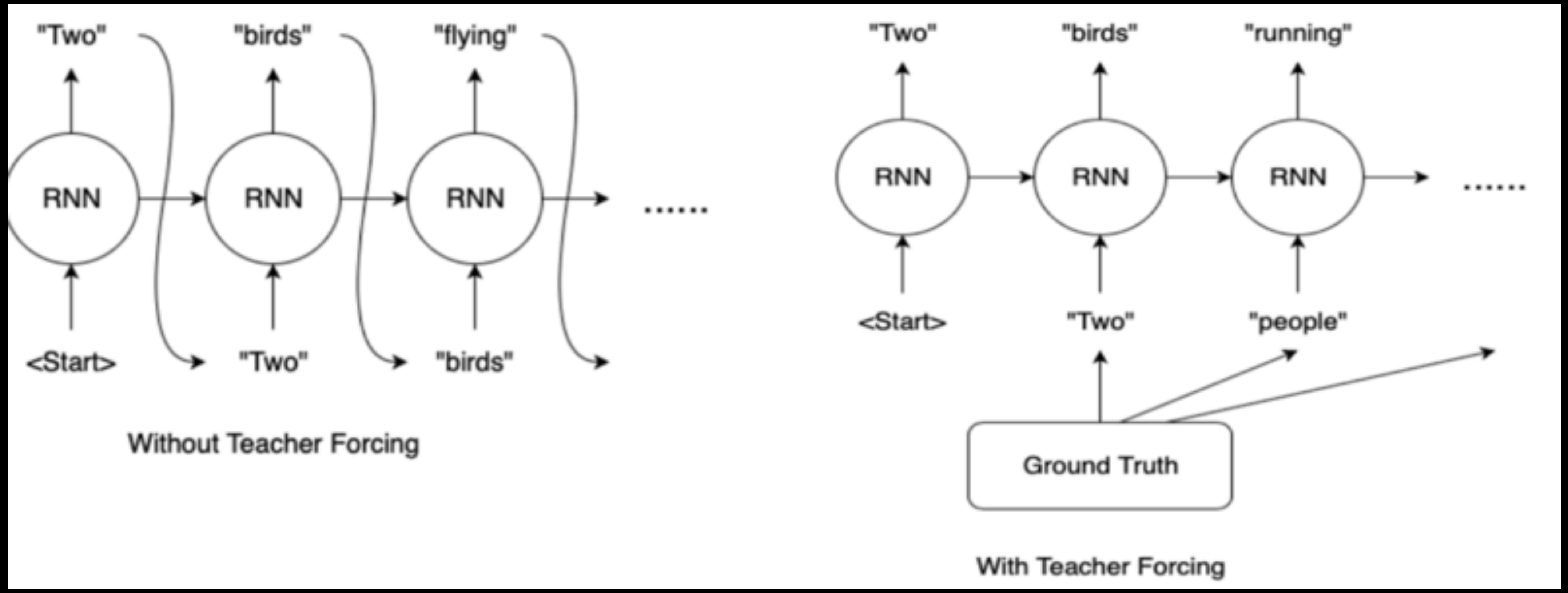
Helicopter Aerobatics

Abbeel et al. 2009

Game AI

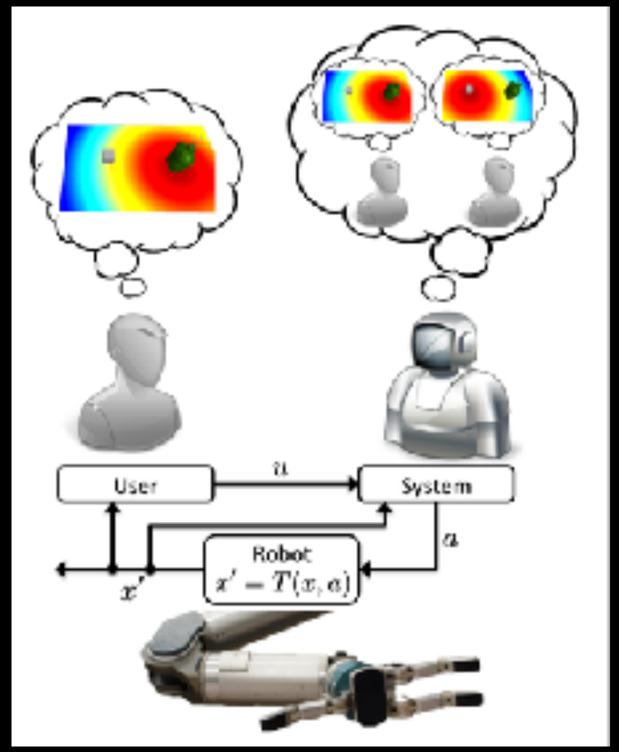
Kozik et al. 2021

Sequence models in NLP



Daume et al. 2009

Shared autonomy



Javdani et al. 2015

Activity!

Think-Pair-Share!

Think (30 sec): What are the various ways to give feedback to a robot to teach it a new task?

Pair: Find a partner

Share (45 sec): Partners exchange ideas

Myths about Imitation Learning

✘ Imitation learning: Do exactly what the human will do

✘ Imitation learning requires humans to demonstrate actions

✘ Imitation learning is a way to warm start reinforcement learning

✘ Imitation learning means you can’t do better than the human

Two Core Ideas

Data

“What is the distribution of states?”

Loss

“What is the metric to match to human?”

Two Core Ideas

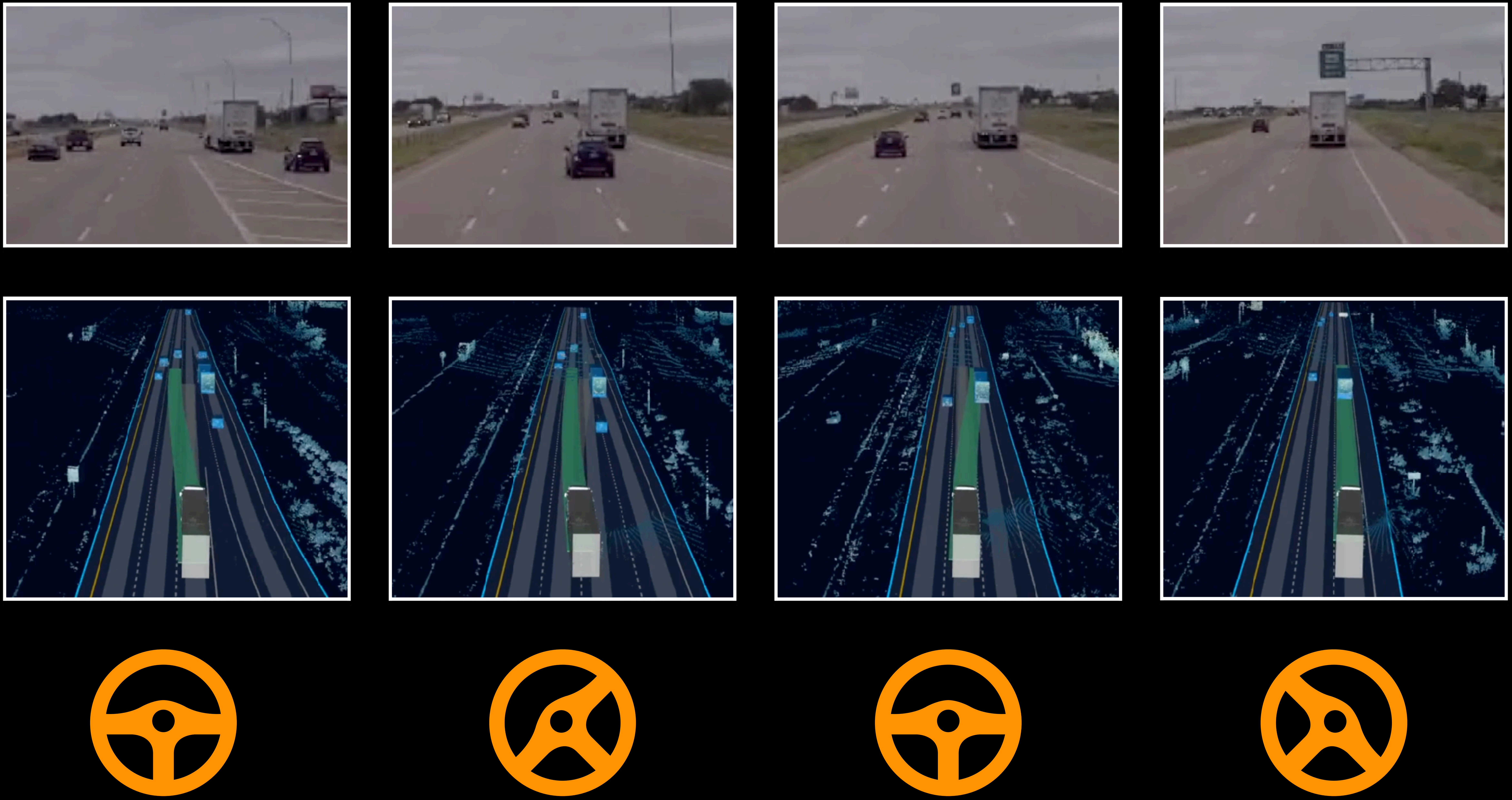
Data

"What is the distribution of states?"

Loss

"What is the metric to match to human?"

Behavior Cloning



State

Action

Behavior Cloning

1. Collect data from a human demonstrator

$$s_1, a_1^*, s_2, a_2^*, s_3, a_3^*, \ldots$$

2. Train a policy $\pi : s_t - > a_t$

3. Validate on held out dataset

What could possibly go wrong?

Feedback drives covariate shift

An old problem

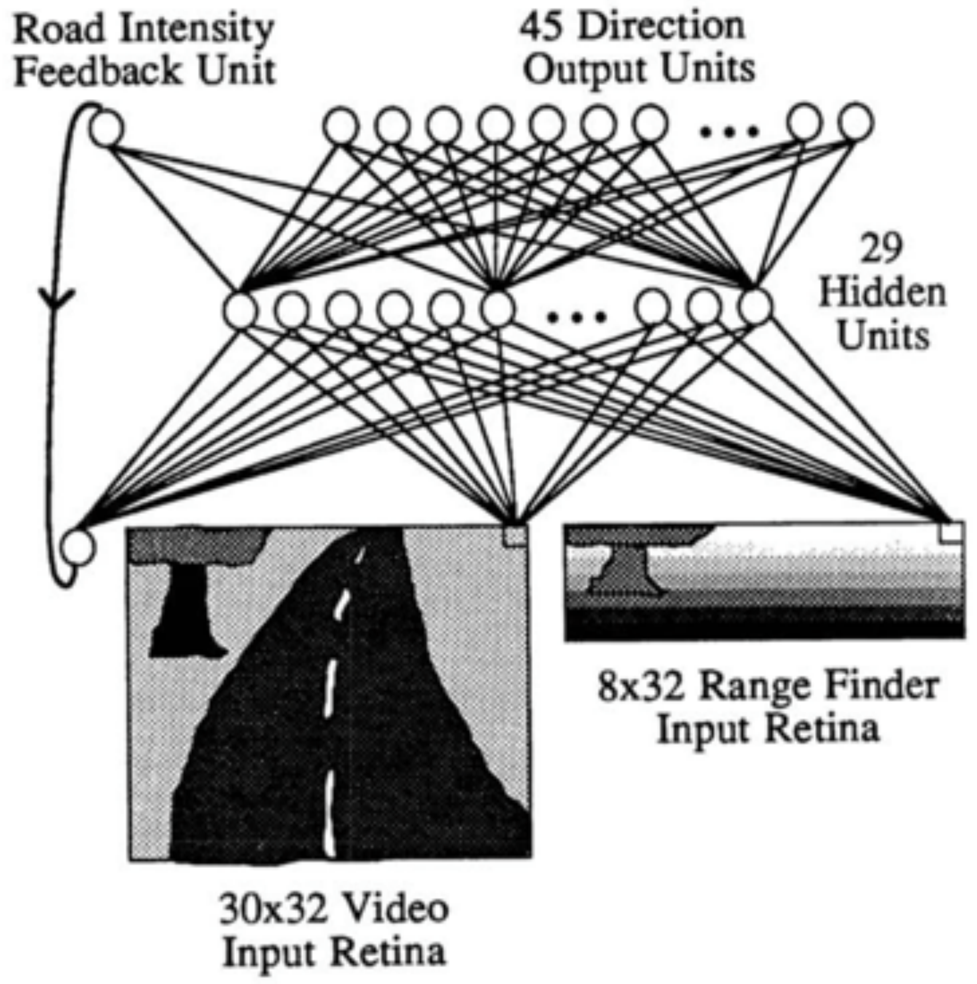


Figure 1: ALVINN Architecture

"...the network must not solely be shown examples of accurate driving, but also how to recover (i.e. return to the road center) once a mistake has been made."

D. Pomerleau
ALVINN: An Autonomous Land Vehicle In A Neural Network, NeurIPS'89

Also observed by [LeCun'05]

Feedback is a pervasive problem in self-driving

"... the inertia problem. When the ego vehicle is stopped (e.g., at a red traffic light), the probability it stays static is indeed overwhelming in the training data. This creates a spurious correlation between low speed and no acceleration, inducing excessive stopping and difficult restarting in the imitative policy ..."

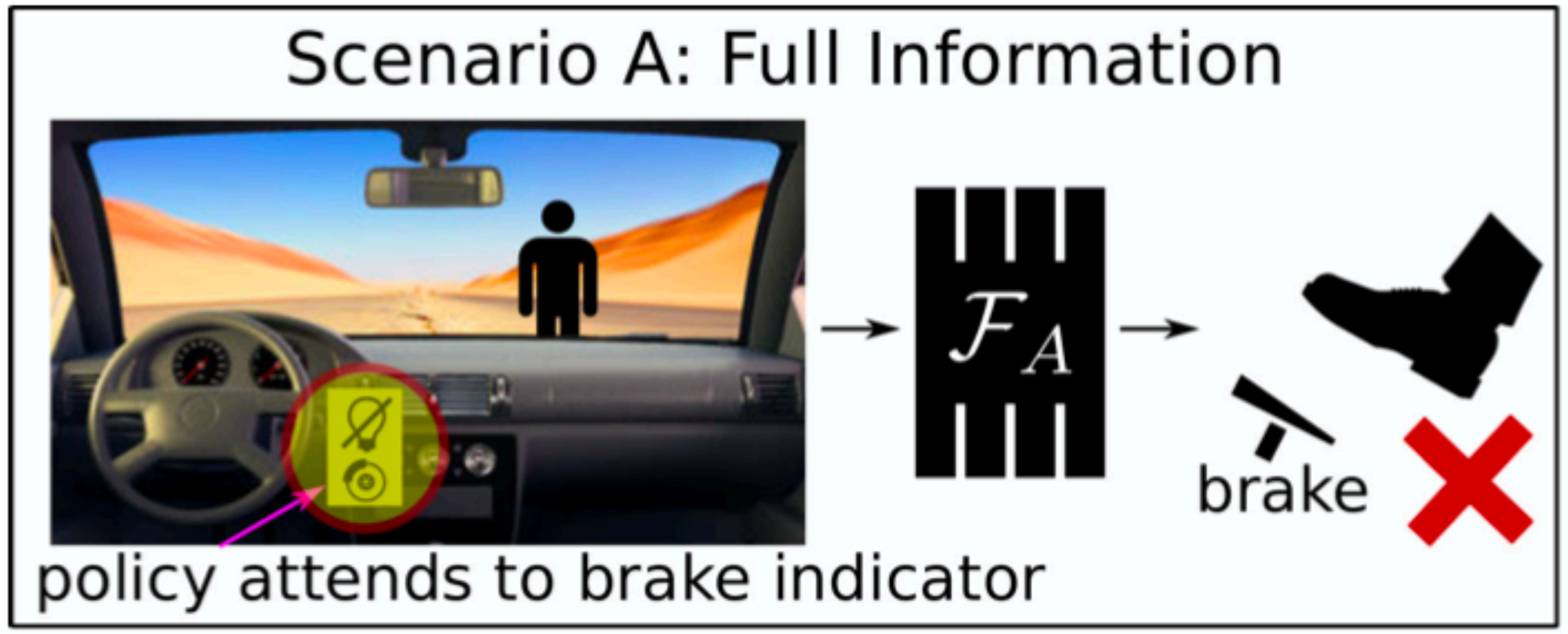
"Exploring the Limitations of Behavior Cloning for Autonomous Driving." F. Codevilla, E. Santana, A. M. Lopez, A. Gaidon. ICCV 2019

"... small errors in action predictions to compound over time, eventually leading to states that human drivers infrequently visit and are not adequately covered by the training data. Poorer predictions can cause a feedback cycle known as cascading errors ..."

"Imitating Driver Behavior with Generative Adversarial Networks". A. Kuefler, J. Morton, T. Wheeler, M. Kochenderfer, IV 2017

"... During closed-loop inference, this breaks down because the past history is from the net's own past predictions. For example, such a trained net may learn to only stop for a stop sign if it sees a deceleration in the past history, and will therefore never stop for a stop sign during closed-loop inference ..."

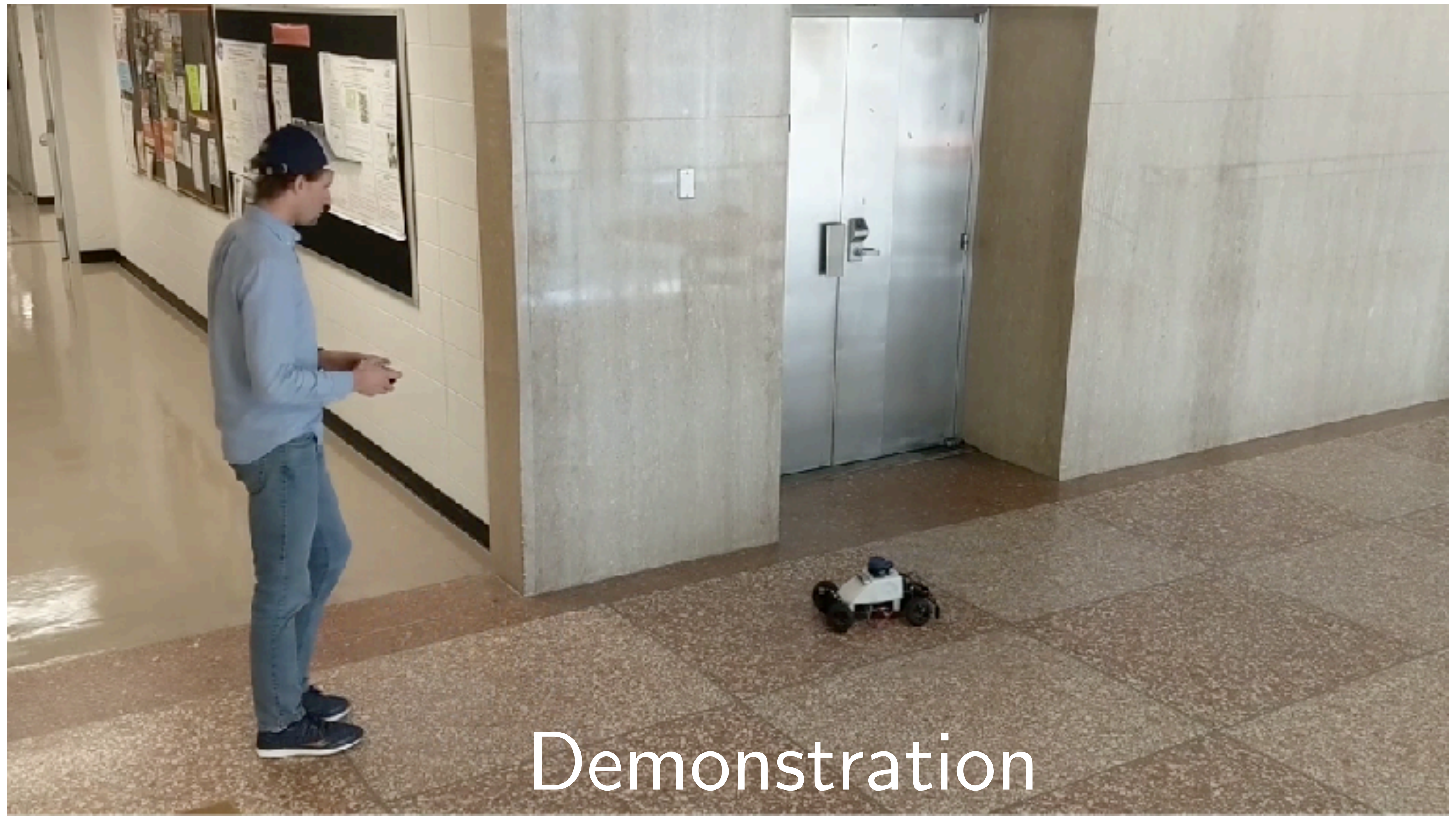
"ChauffeurNet: Learning to Drive by Imitating the Best and Synthesizing the Worst". M. Bansal, A. Krizhevsky, A. Ogale, Waymo 2018

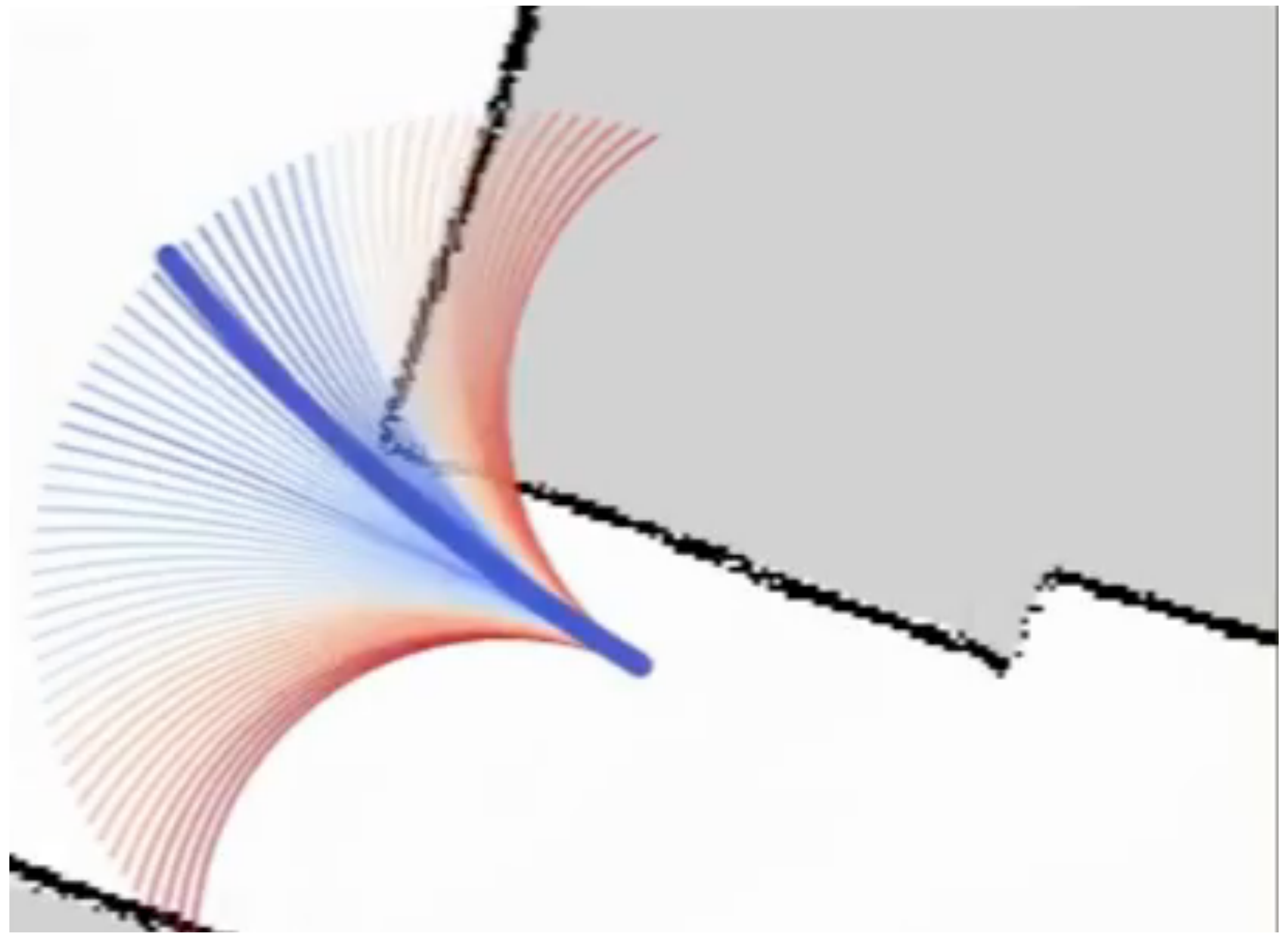

Scenario A: Full Information
$\mathcal{F}_A$
brake
policy attends to brake indicator

"Causal Confusion in Imitation Learning". P. de Haan, D. Jayaraman, S. Levine, NeurIPS '19

Feedback is an old adversary!

[SCB+ RSS'20]

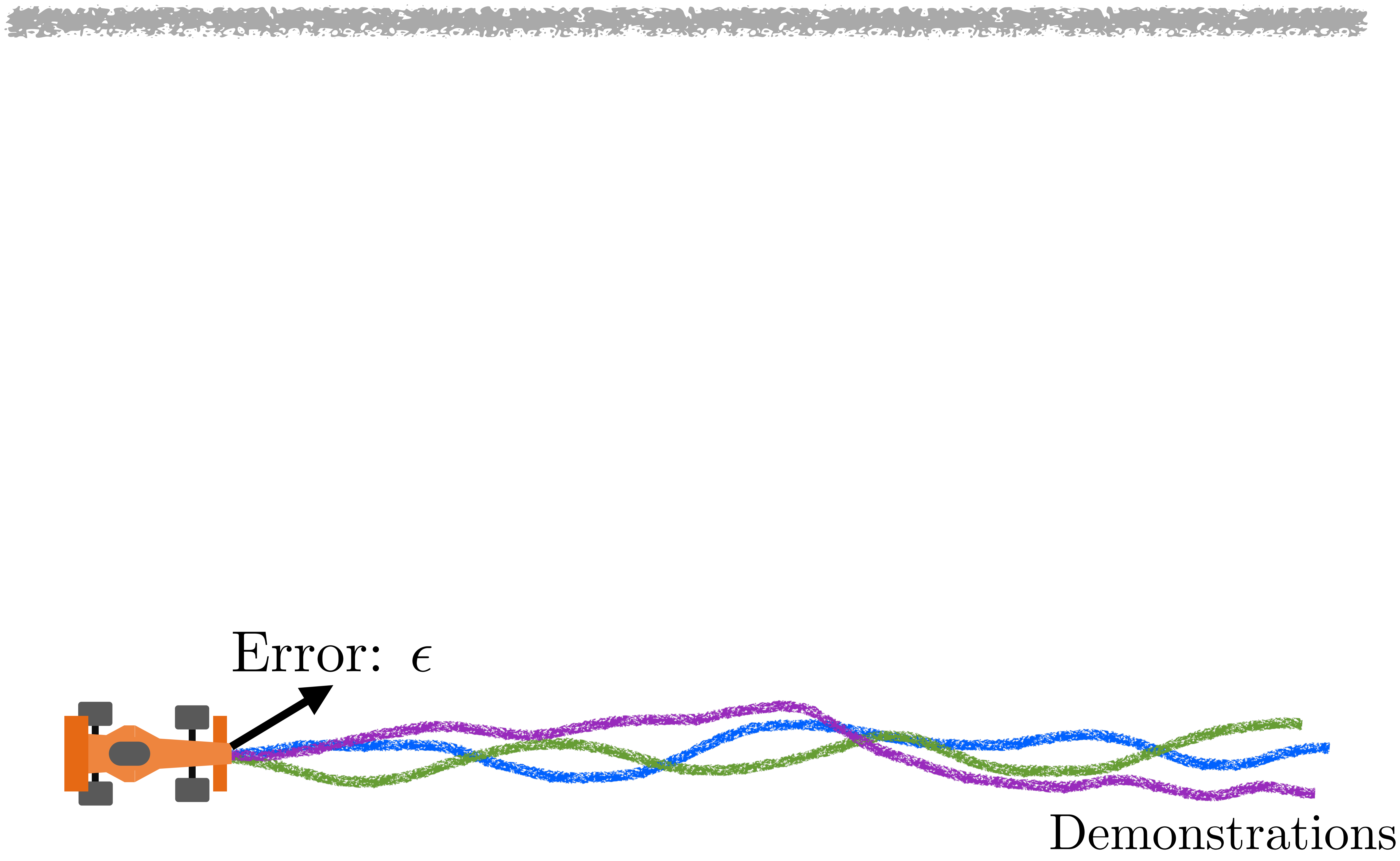




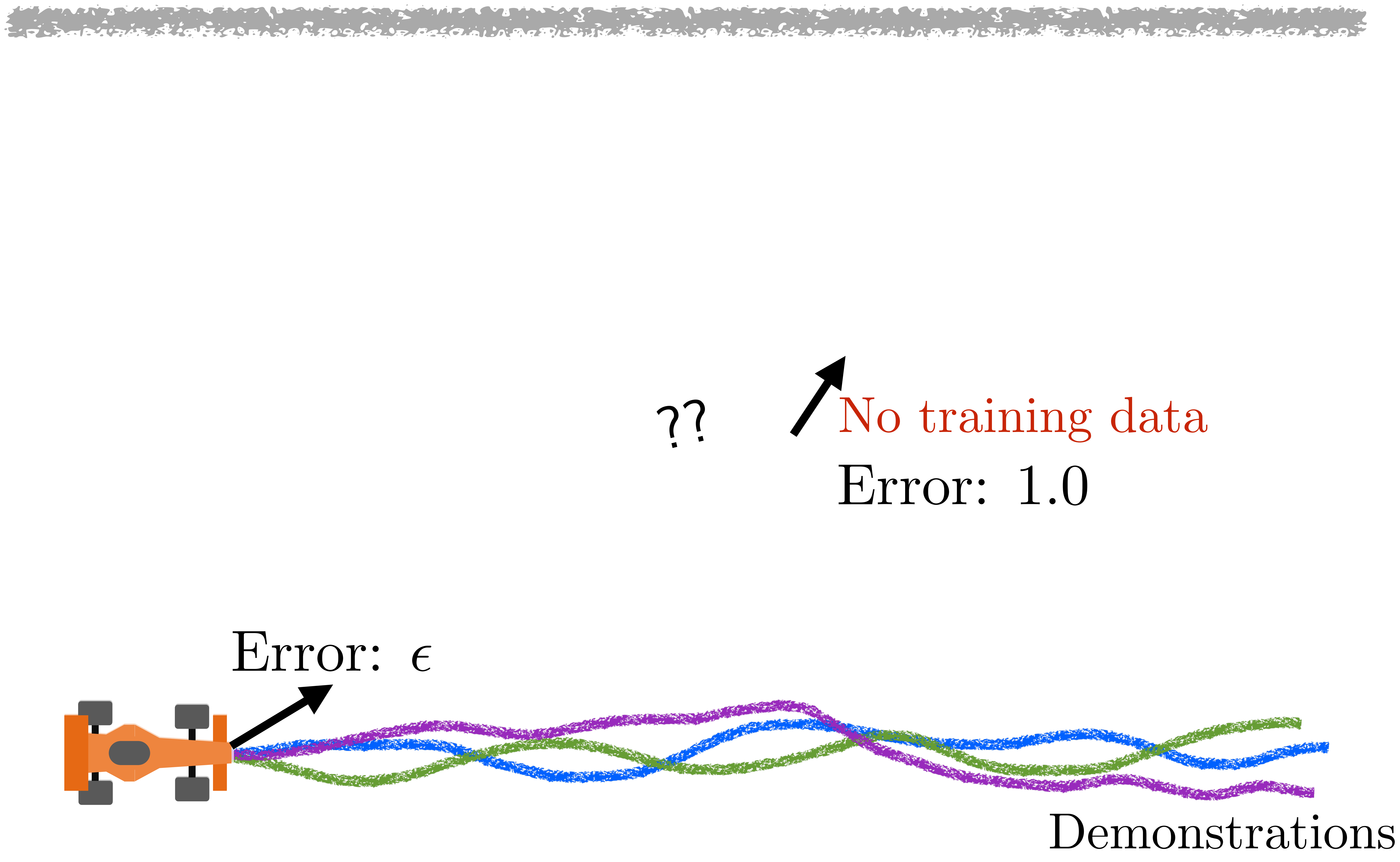
Learnt policy

Behavior Cloning crashes into a wall

Why did the robot crash?



Why did the robot crash?



??

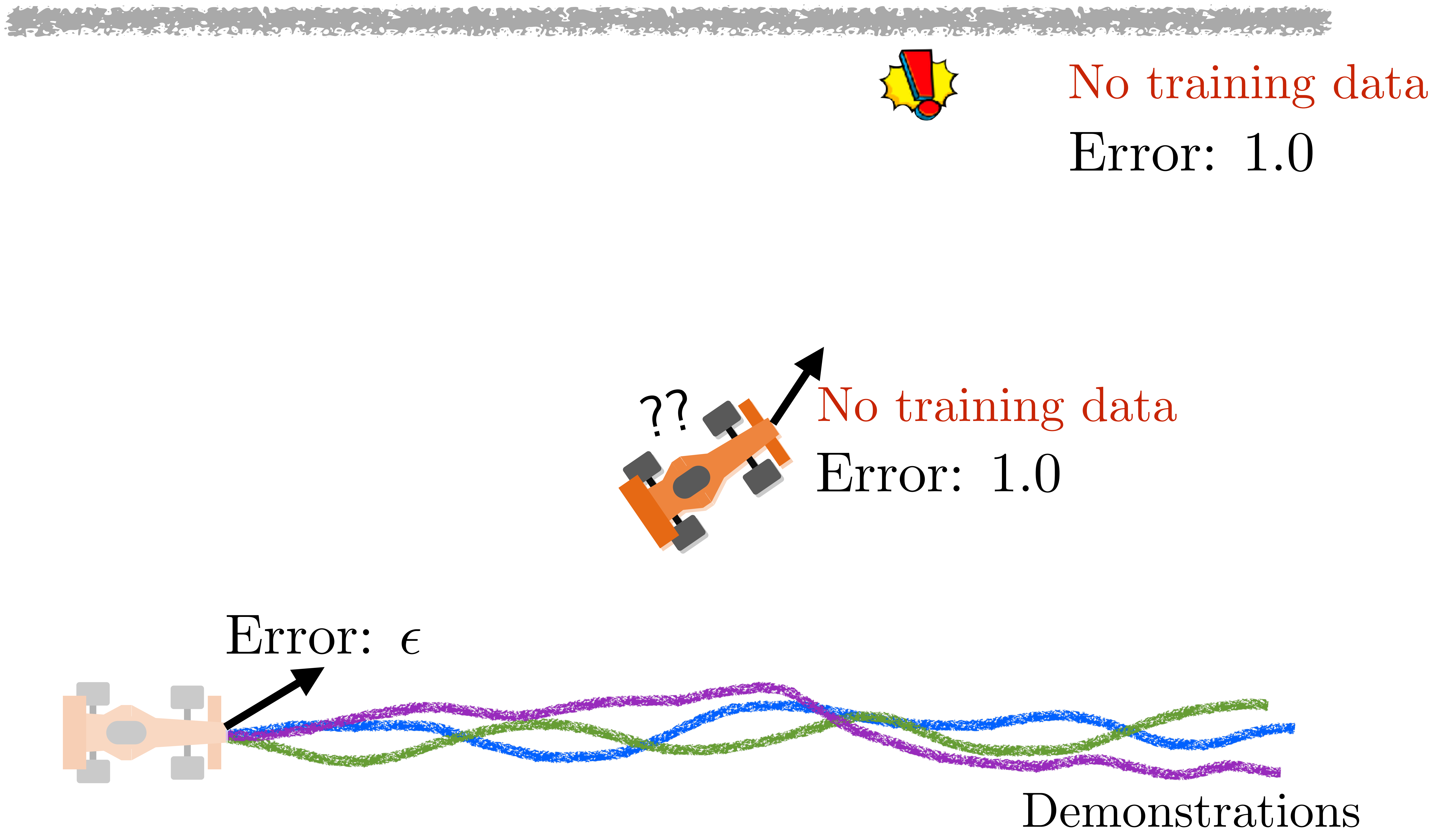
No training data

Error: 1.0

Error: ϵ

Demonstrations

Why did the robot crash?



No training data

Error: 1.0

??

No training data

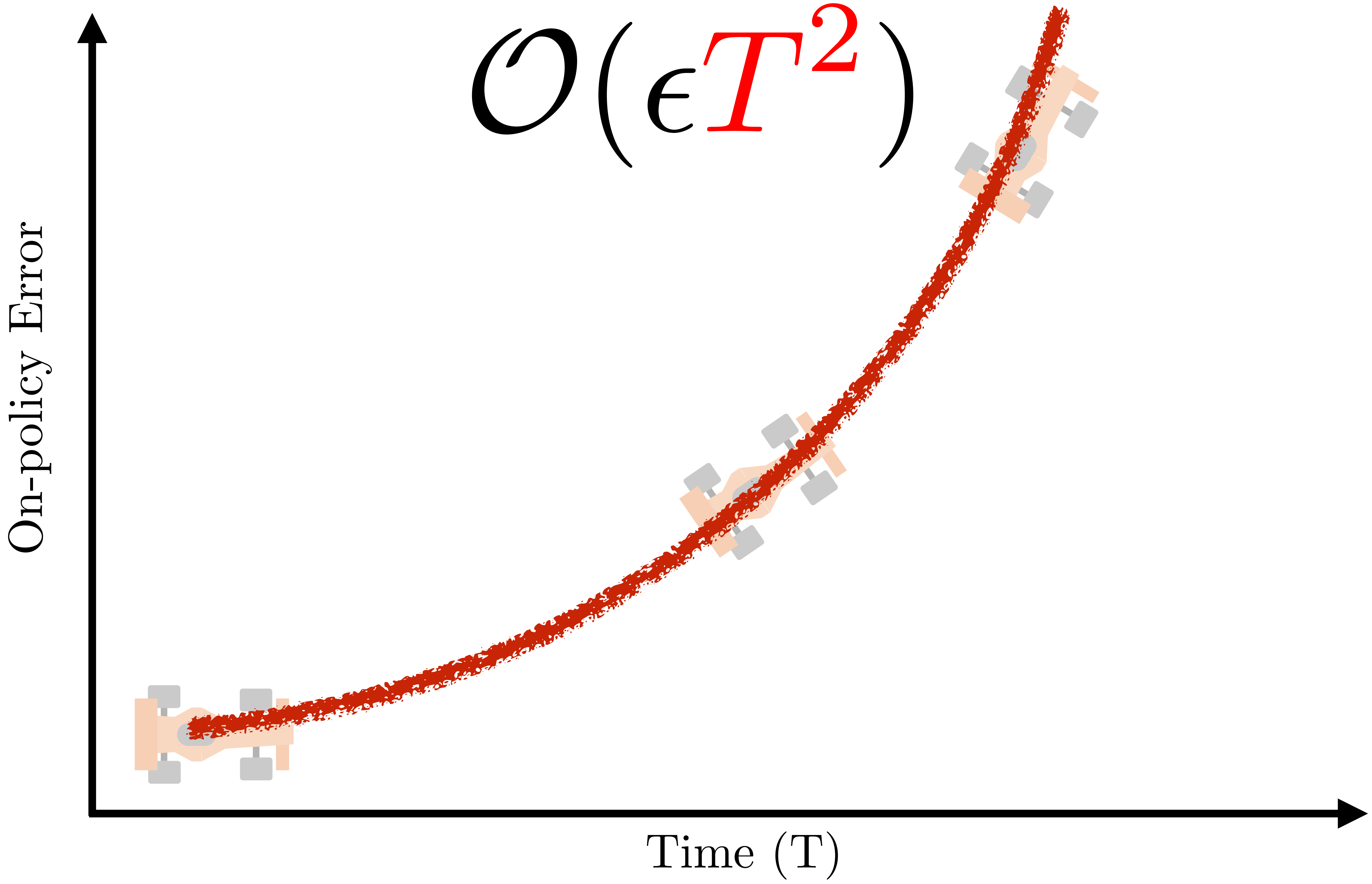
Error: 1.0

Error: ϵ

Demonstrations

Errors feedback and compound

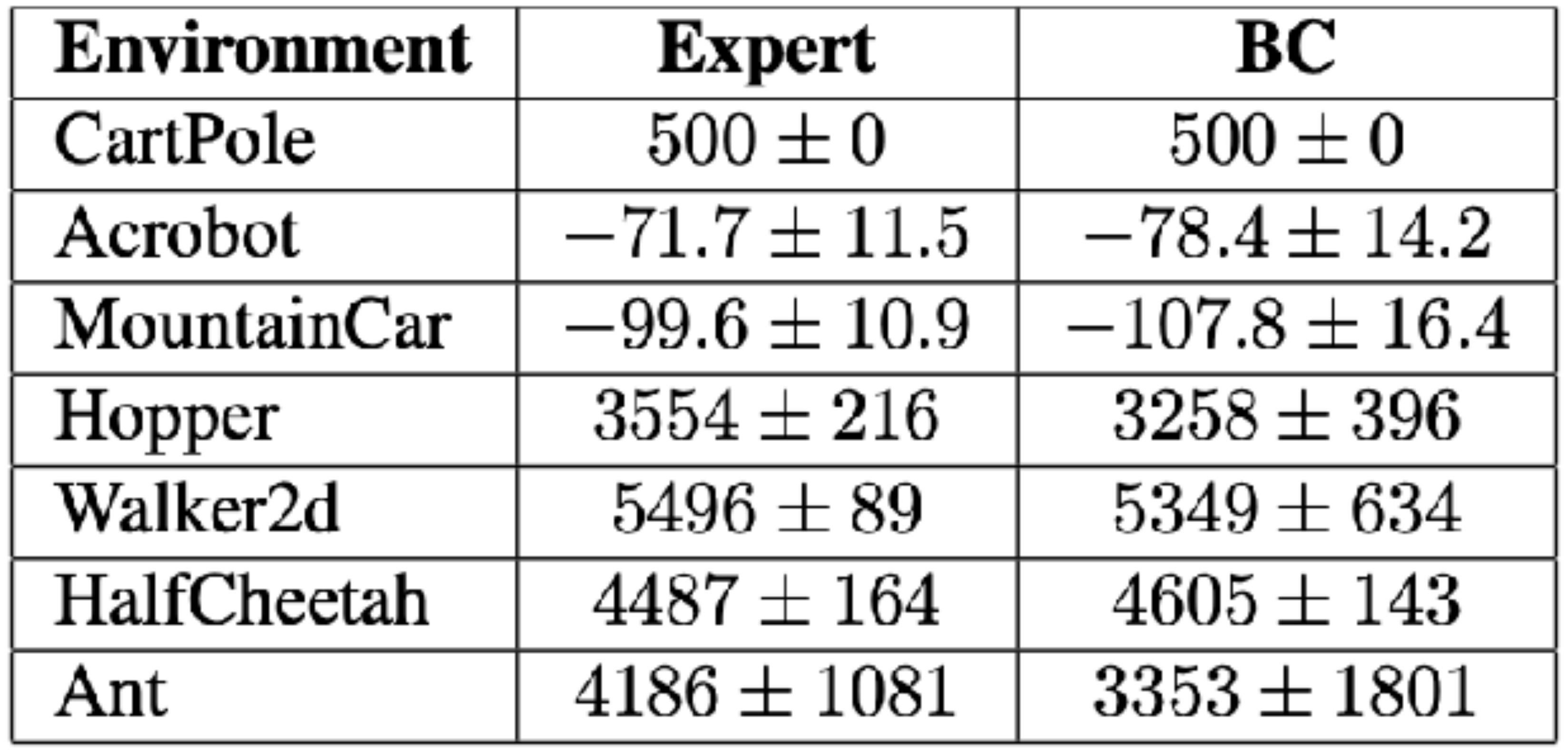
[Ross and Bagnell'10]



Prove it!

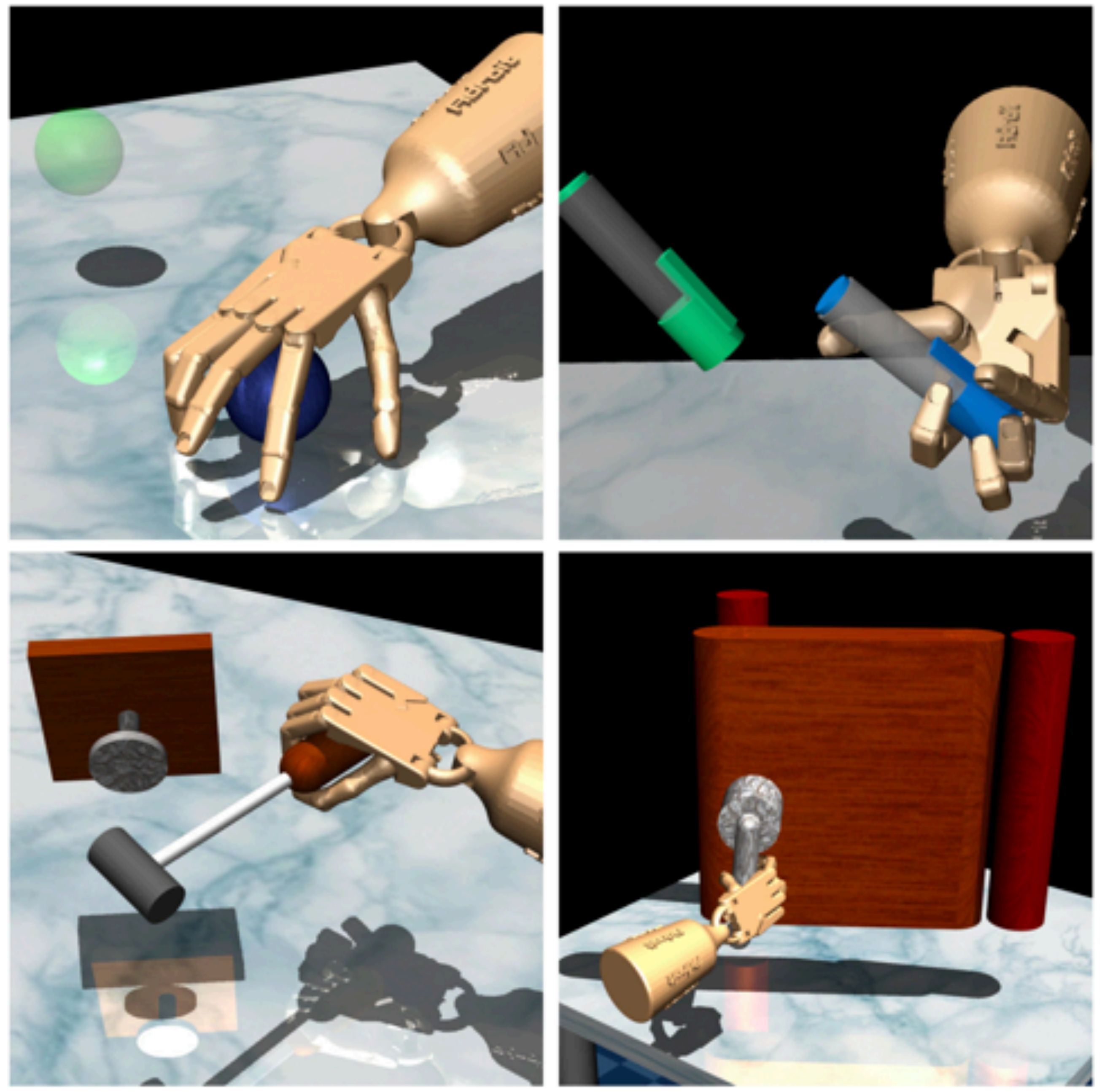
Feedback drives covariate shift

But … Behavior Cloning works just fine on benchmark datasets!

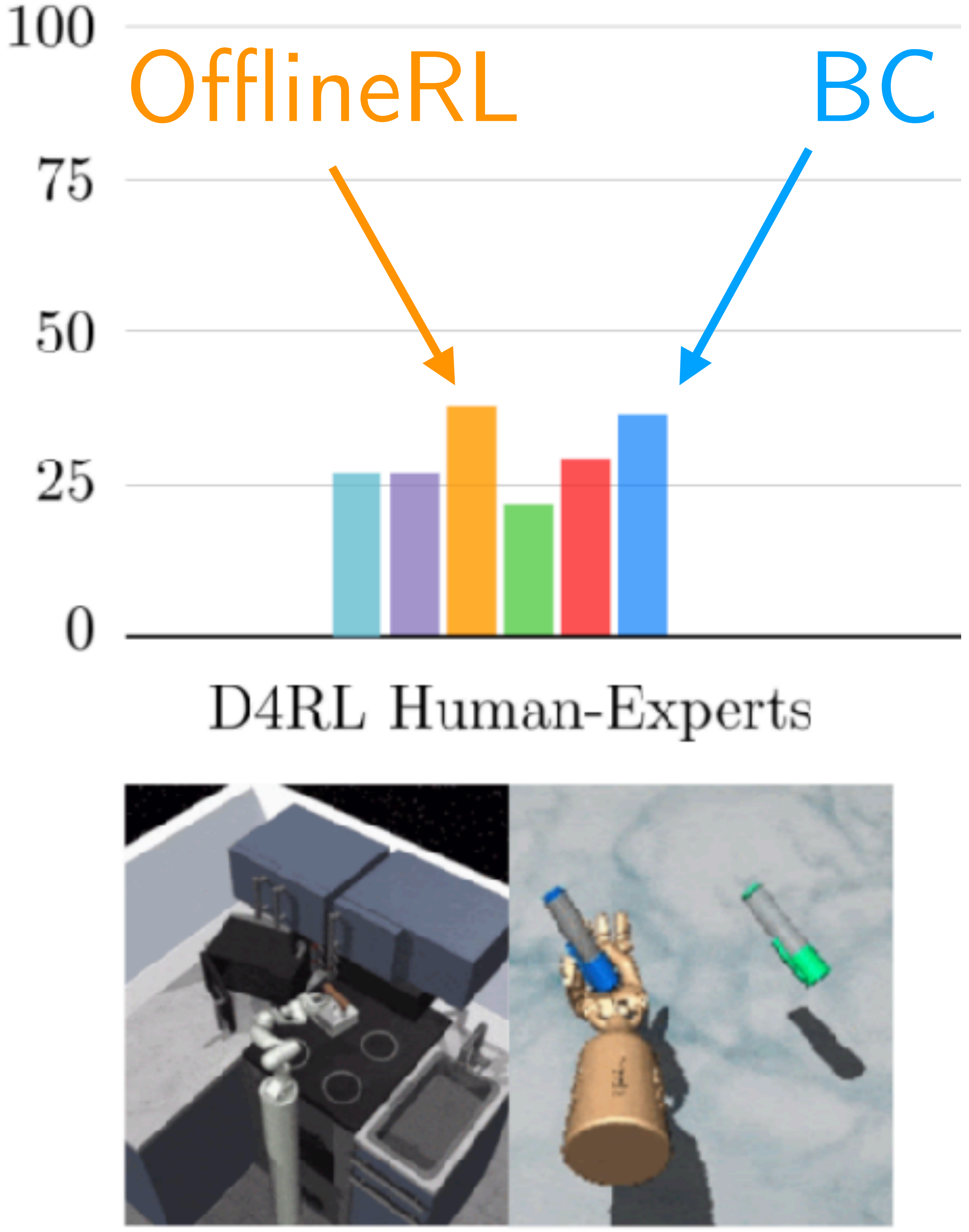


Environment	Expert	BC
CartPole	500 ± 0	500 ± 0
Acrobot	-71.7 ± 11.5	-78.4 ± 14.2
MountainCar	-99.6 ± 10.9	-107.8 ± 16.4
Hopper	3554 ± 216	3258 ± 396
Walker2d	5496 ± 89	5349 ± 634
HalfCheetah	4487 ± 164	4605 ± 143
Ant	4186 ± 1081	3353 ± 1801

[SCV+ arXiv ’21]



[Rajeswaran et al. ’17]



[Florence et al. ’21]

🤔 What explains this mismatch?

Real-world self-driving

Feedback drives covariate shift, Behavior Cloning compounds in error

vs

Benchmark datasets

Behavior Cloning does just fine!

Let’s travel to the INFINITE data limit!

The
Three Regimes
of
Covariate
Shift

Easy 😄

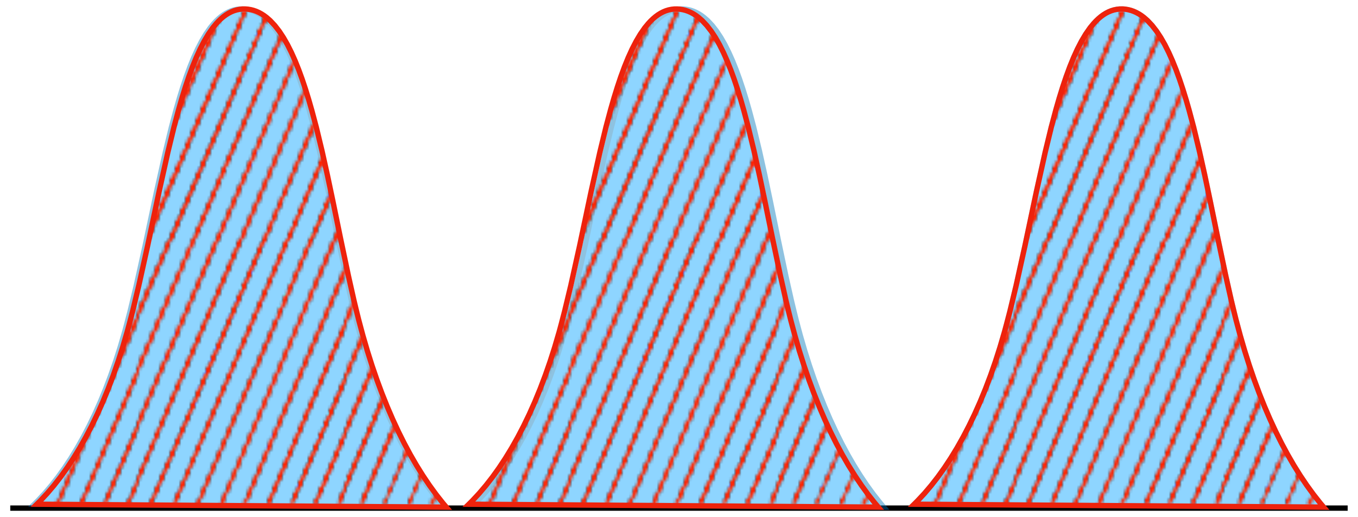
Setting

Expert is realizable

$\pi^E \in \Pi$

As $N \to \infty$, drive down

$\epsilon = 0$ (or Bayes error)



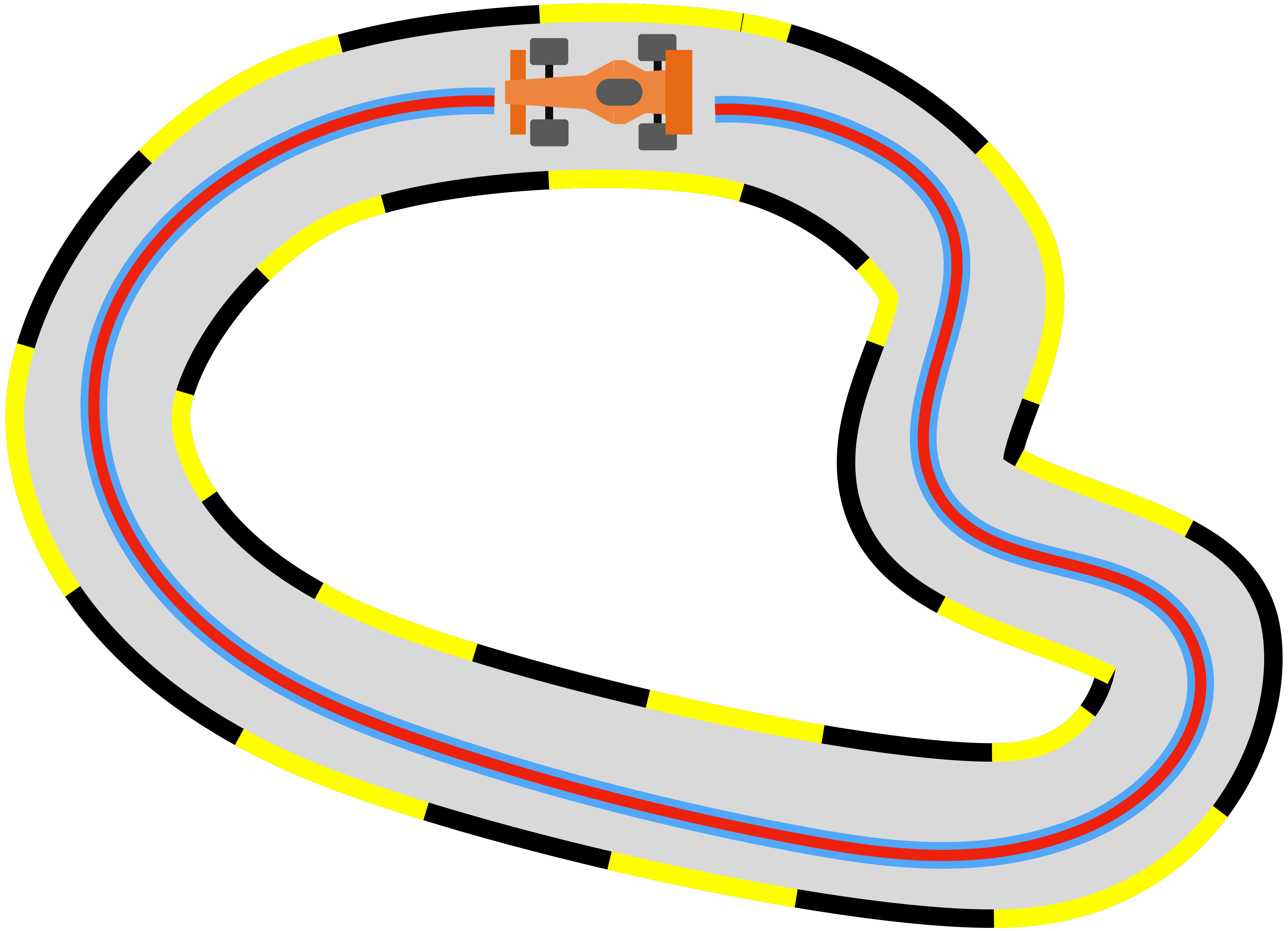
Solution

Nothing special.

Collect lots of data and

do Behavior Cloning

Expert $\rho^{\pi^E}(s)$ $\approx$ Learner $\rho^{\pi}(s)$



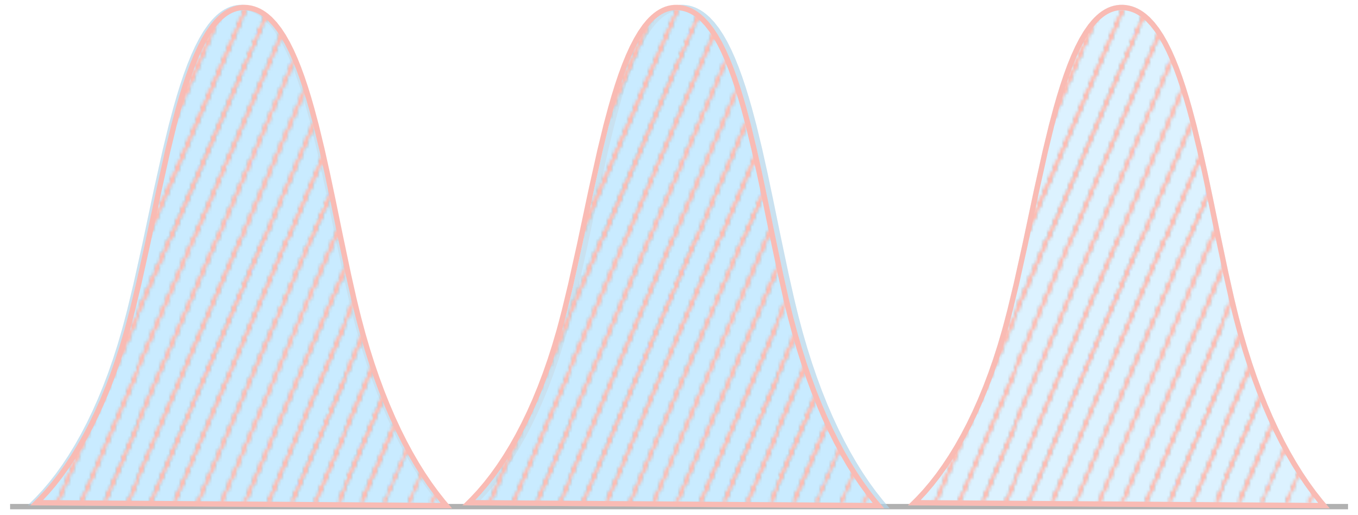
Setting

Solution

Easy 😄

Expert is realizable
$\pi^E \in \Pi$

As $N \to \infty$, drive down
$\epsilon = 0$ (or Bayes error)

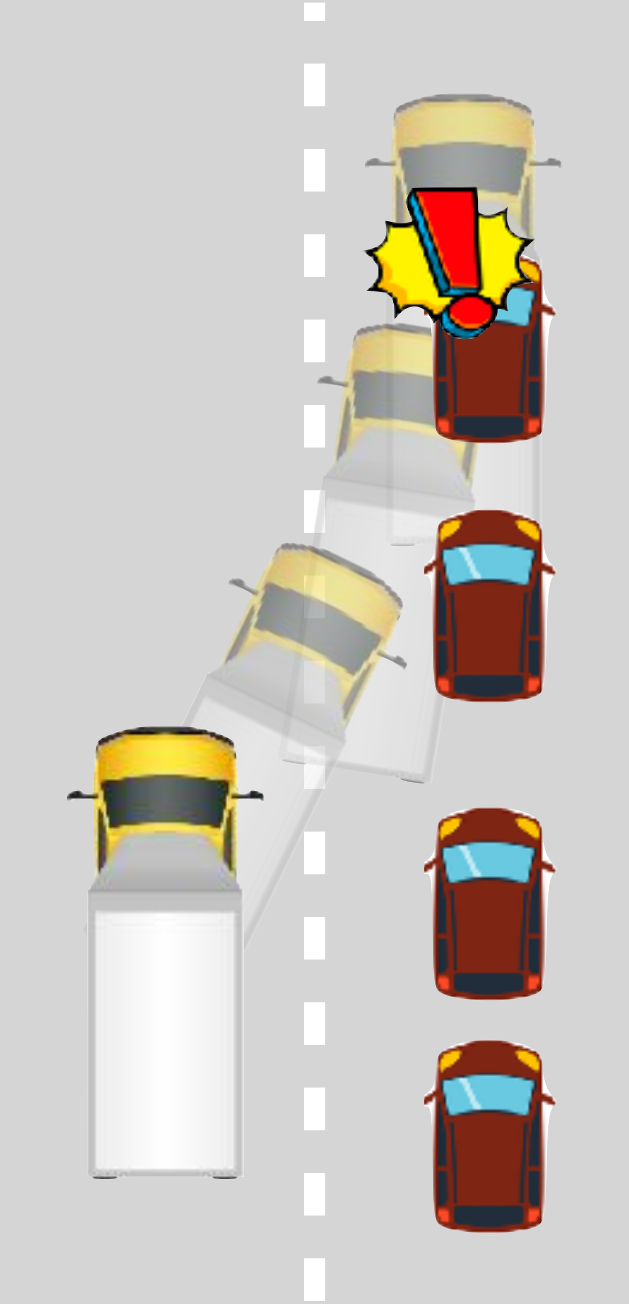


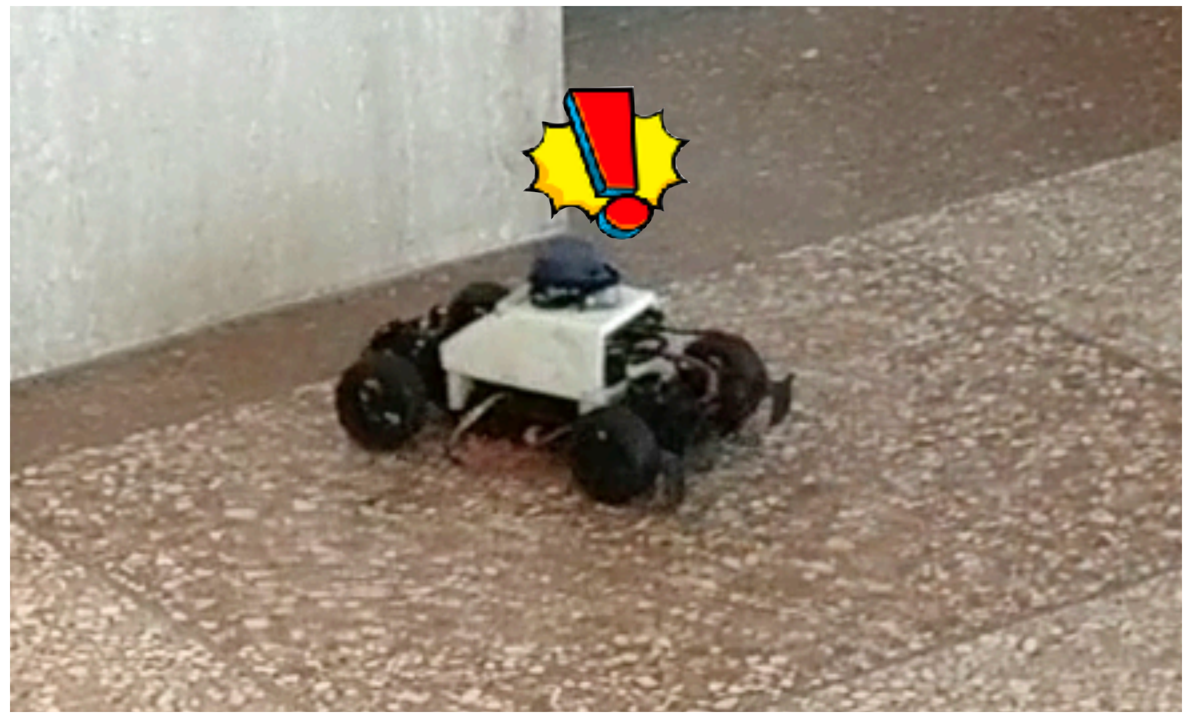
Nothing special.
Collect lots of data and
do Behavior Cloning

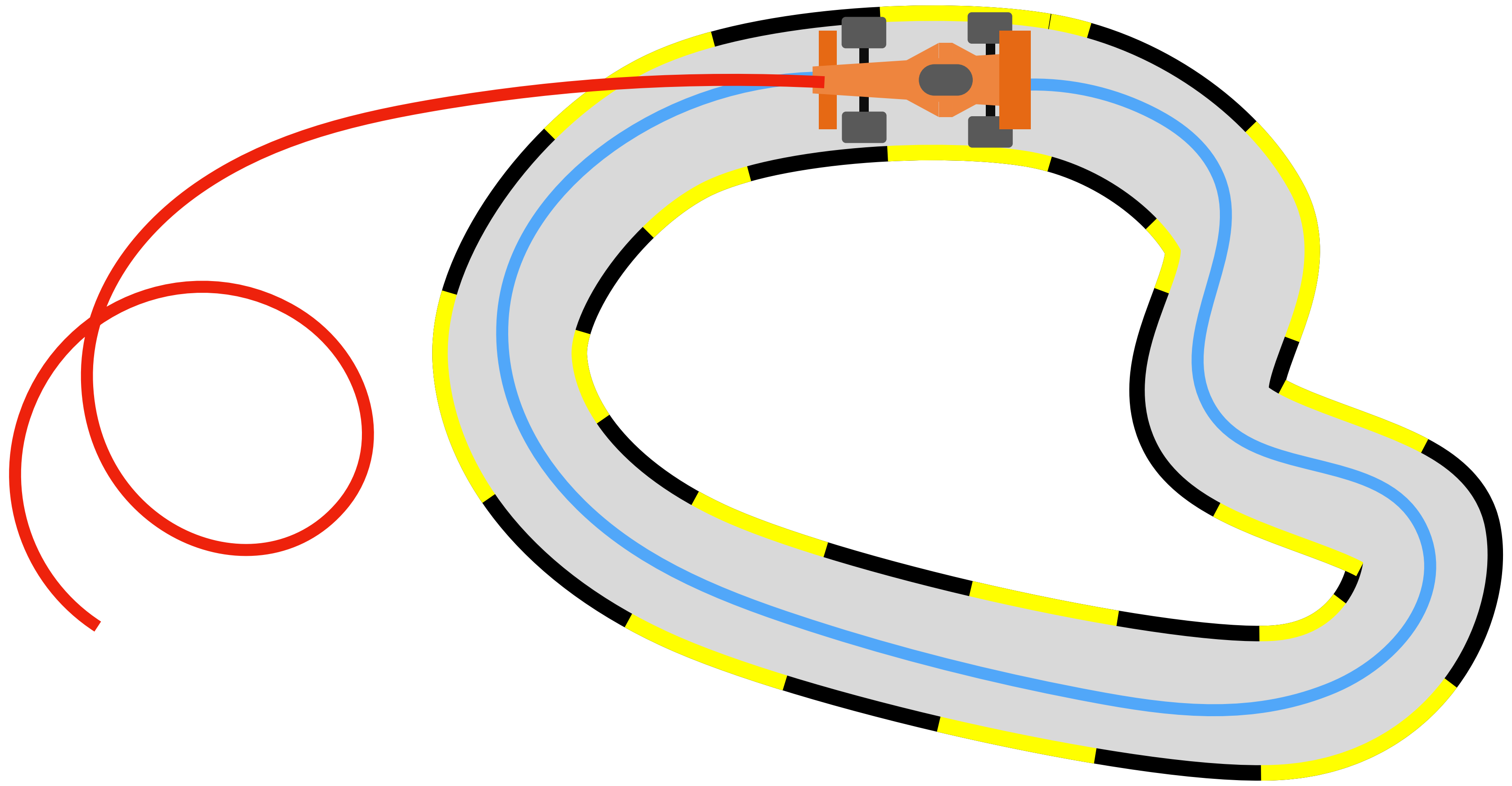
Hard 😱

Non-realizable expert +
limited expert support

Non-realizable expert + limited support?

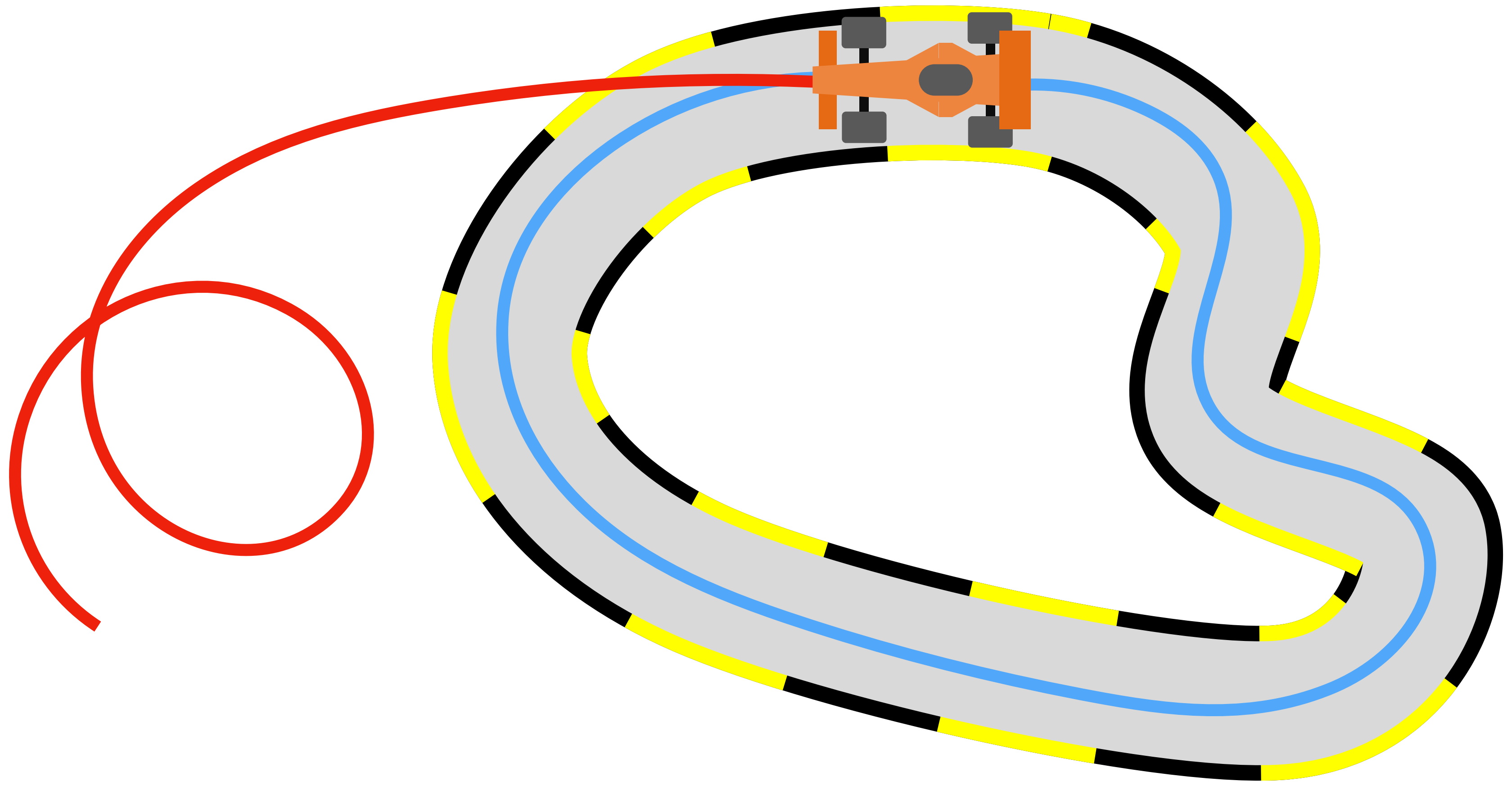






No label for what to do
in this state!

Non-realizable expert + limited support?



Hard 😱

Behavior Cloning
compounds in error $O(\epsilon T^2)$

[Ross & Bagnell ’10]

Setting

Solution

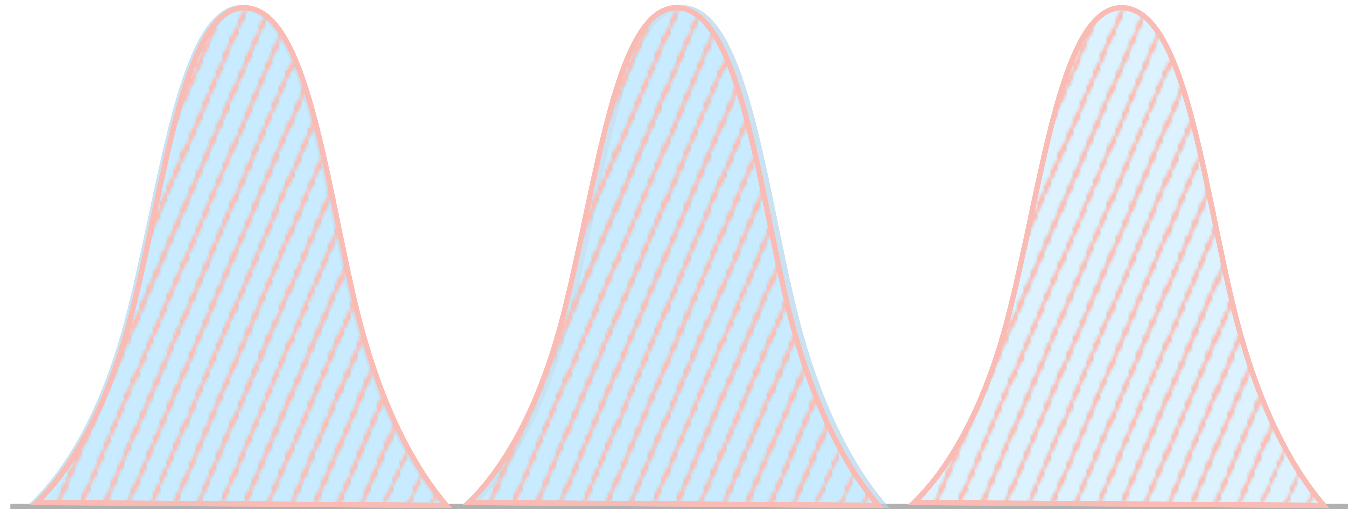
Easy 😄

Expert is realizable

$\pi^E \in \Pi$

As $N \to \infty$, drive down

$\epsilon = 0$ (or Bayes error)



Nothing special.

Collect lots of data and

do Behavior Cloning

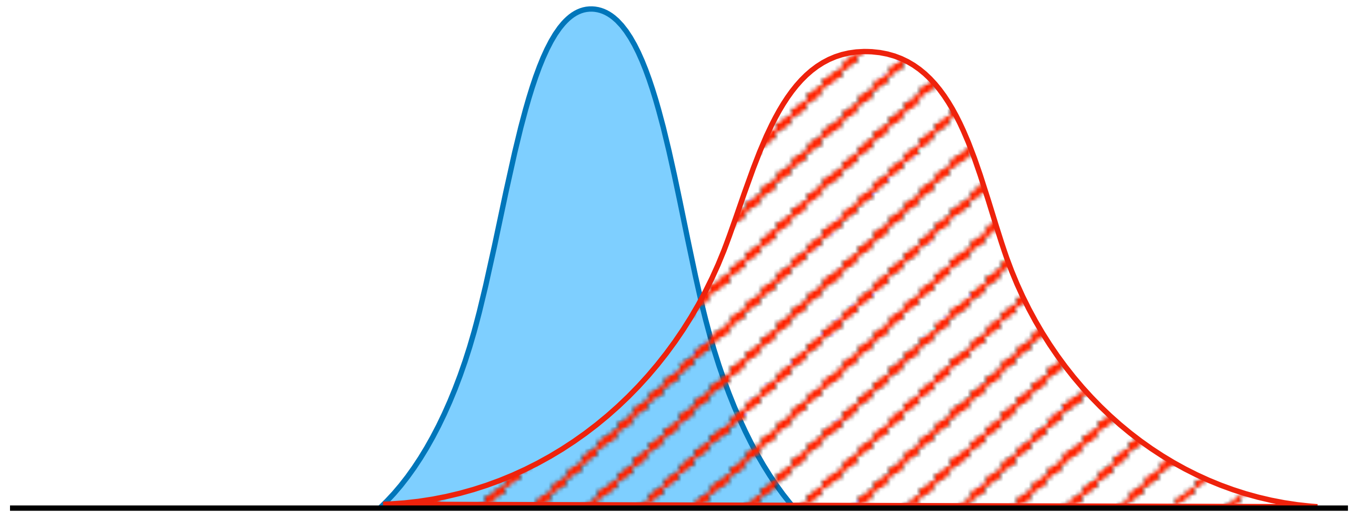
Hard 😱

Non-realizable expert +

limited expert support

Even as $N \to \infty$,

behavior cloning $O(\epsilon T^2)$



?