Behavior Cloning, Feedback and Covariate Shift

Sanjiban Choudhury

Cornell Bowers CIS
Computer Science

Today's class

- ☐ What is imitation learning?
- ☐ Behavior Cloning
- ☐ Feedback drives Covariate Shift

What have we learnt so far?

1. How do define a MDP

2. How to solve a MDP given I know S, A, C, T

But there are challenges in applying this

Q1. What if I can write down my costs, but my transitions are unknown?

Reinforcement Learning! (Later in the course)

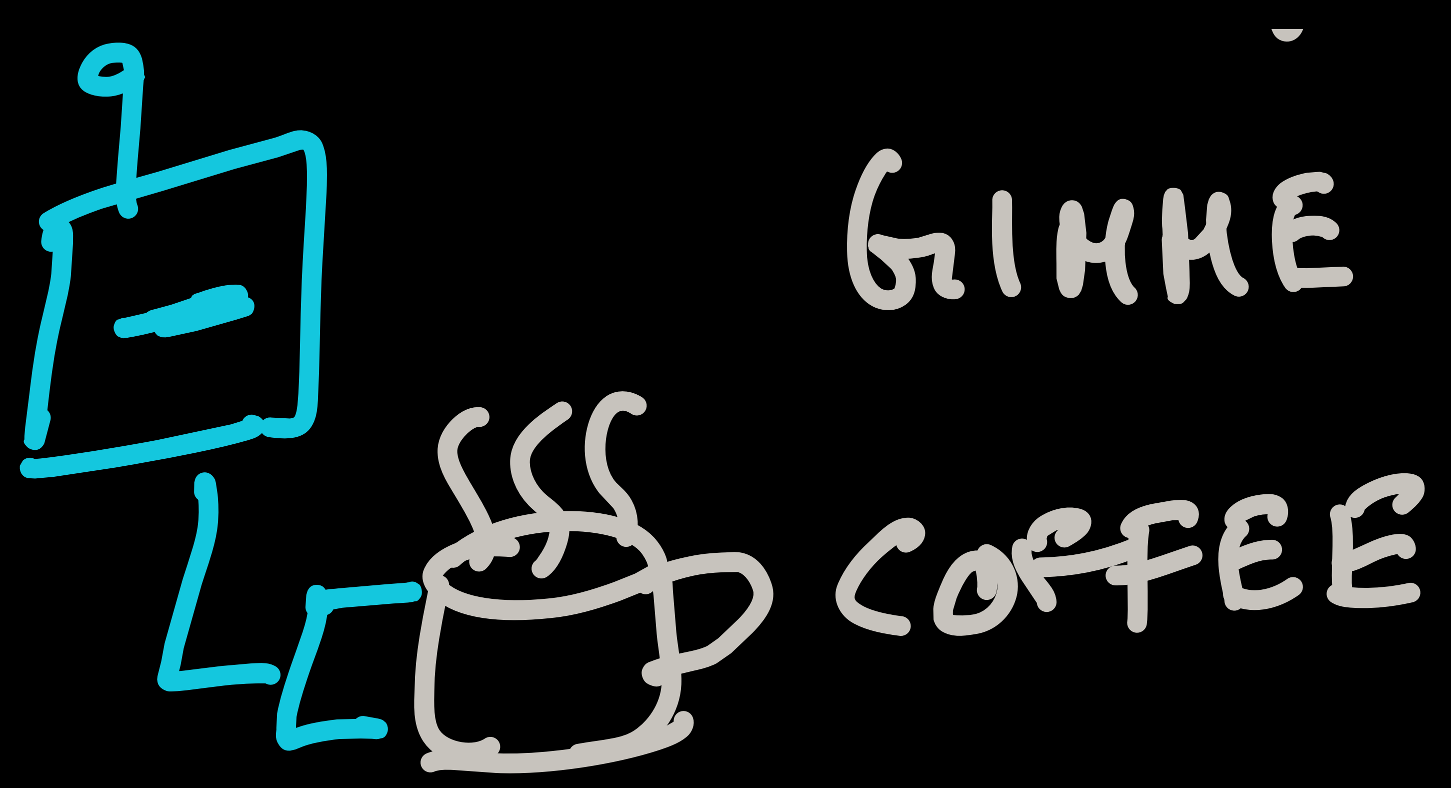
Q2. But what if even writing down costs is hard?

Imitation Learning! (Today)

How do we program robots to do tasks?

Programming a task ...

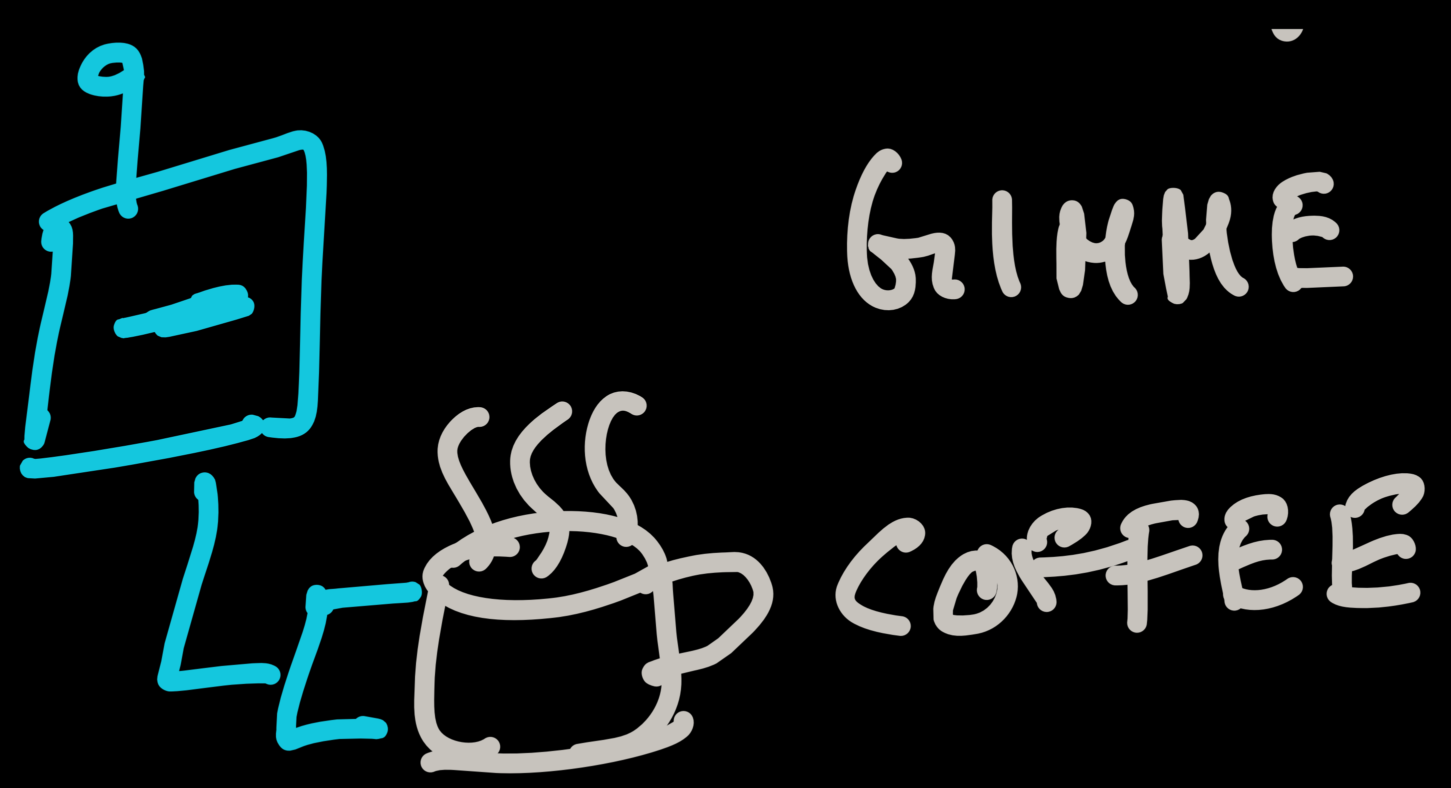
tell the robot to make coffee ..



robot burns down
the house!

Programming a task ...

tell the robot to make coffee ..



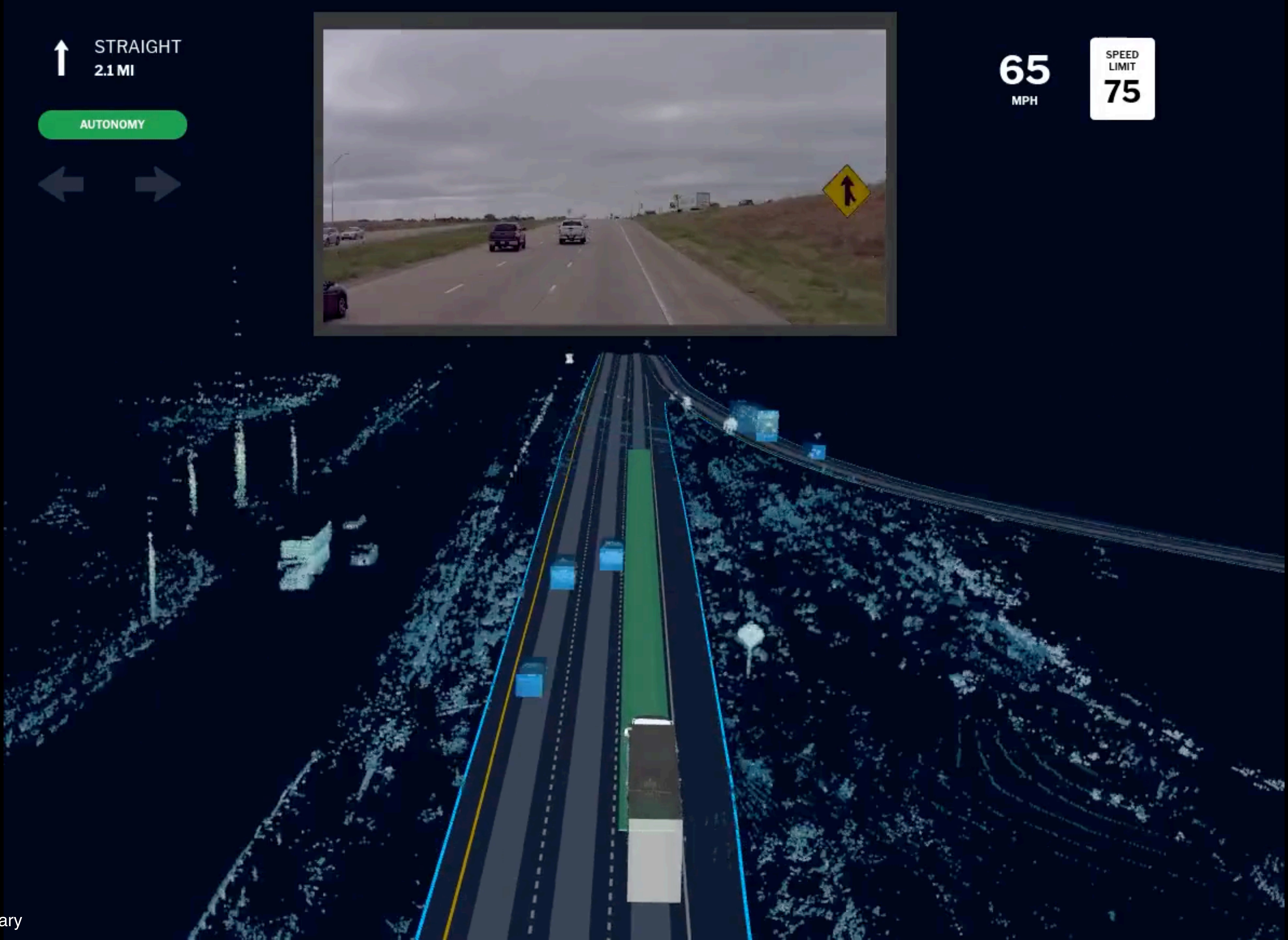
DON'T ...

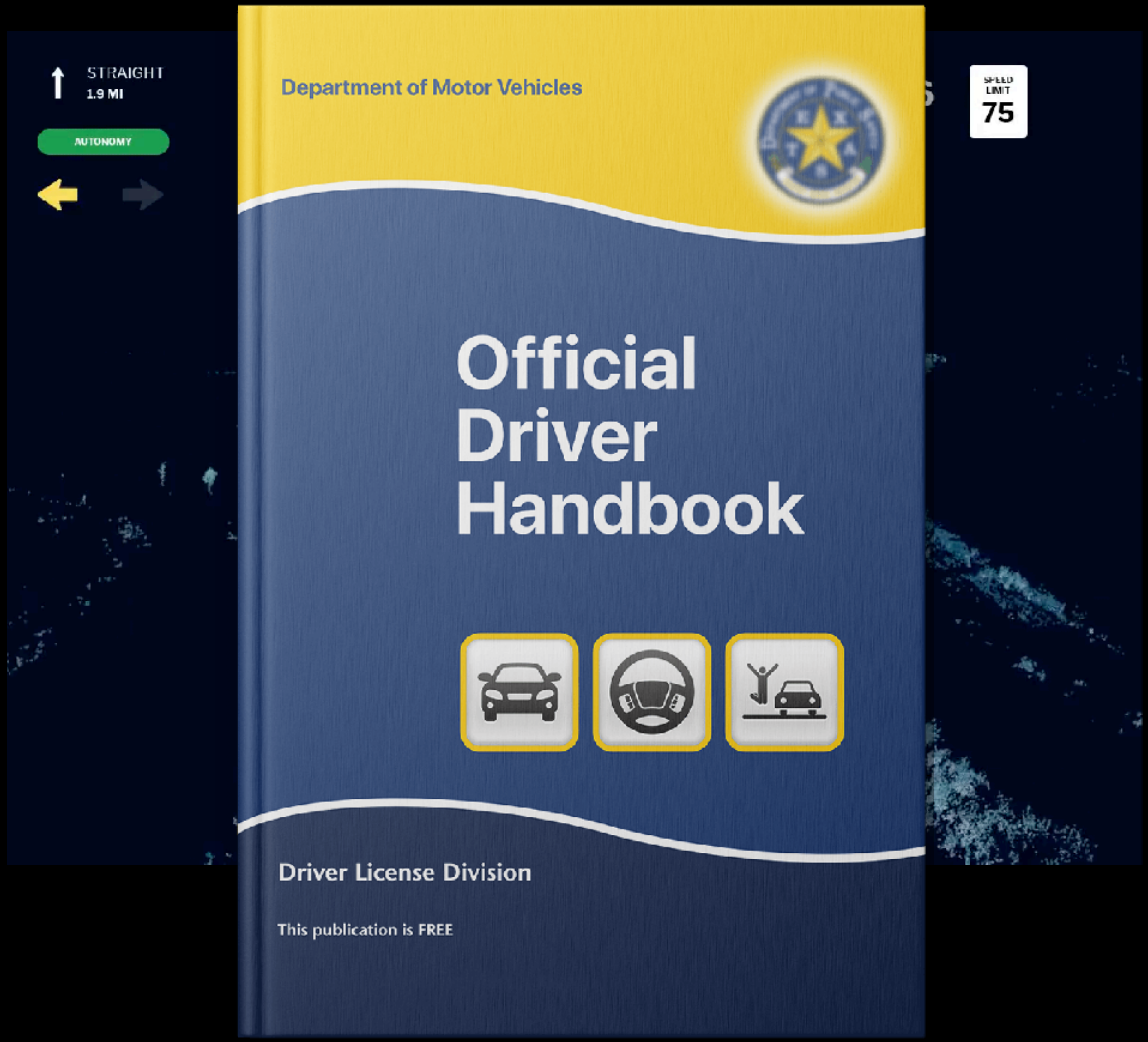
burn down the house

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!

Imitation Learning

Implicitly program robots

Activity!

Think-Pair-Share!

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

Pair: Find a partner

Share (45 sec): Partners exchange ideas

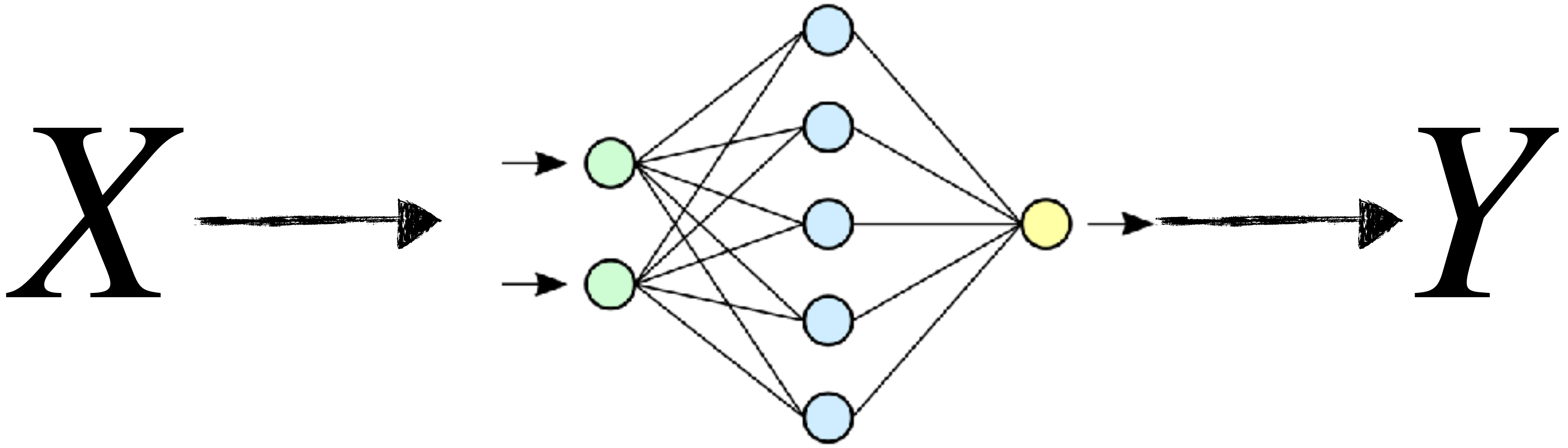
Today's class

- [x] What is imitation learning?
- [ ] Behavior Cloning
- [ ] Feedback drives Covariate Shift

How do we solve imitation learning?

Treat robotics as a “simple” ML problem ...

Ultimately, we just need to learn a function



Input
(State)

Model
(Policy)

Output
(Action)

Behavior Cloning

(aka Supervised Learning)

Behavior Cloning

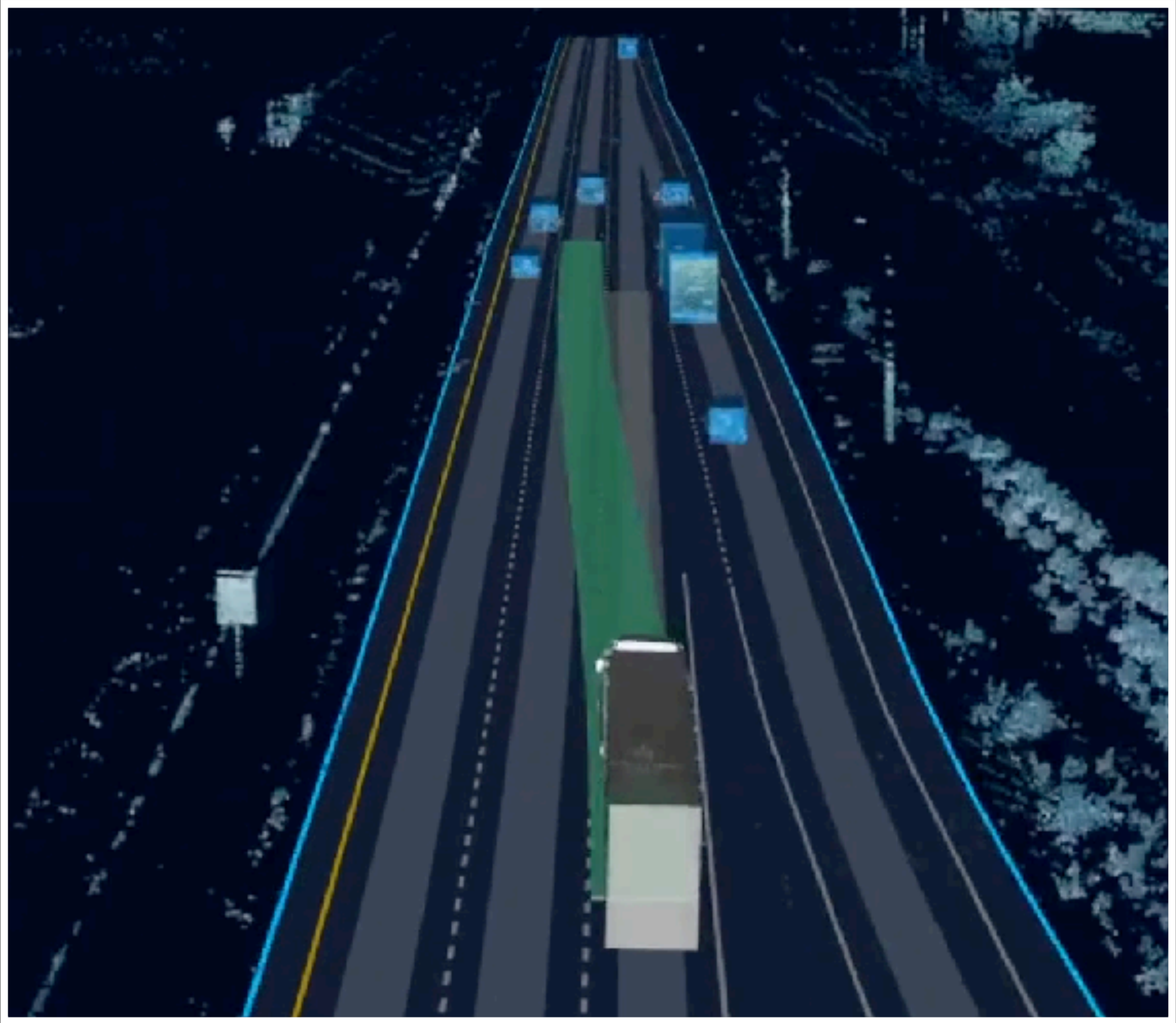
1. Collect data from a human demonstrator

$$[\quad (s_1, a_1^*), \quad (s_2, a_2^*), \quad (s_3, a_3^*), \quad \ldots \quad]$$

2. Train a policy $\pi : s_t \rightarrow a_t$ on the data

3. Check validation error on held out dataset   Why?

Let's apply Behavior Cloning!



1. Collect data from a human demonstrator

$$[\quad (s_1, a_1^*), \quad (s_2, a_2^*), \quad (s_3, a_3^*), \quad \dots \quad]$$

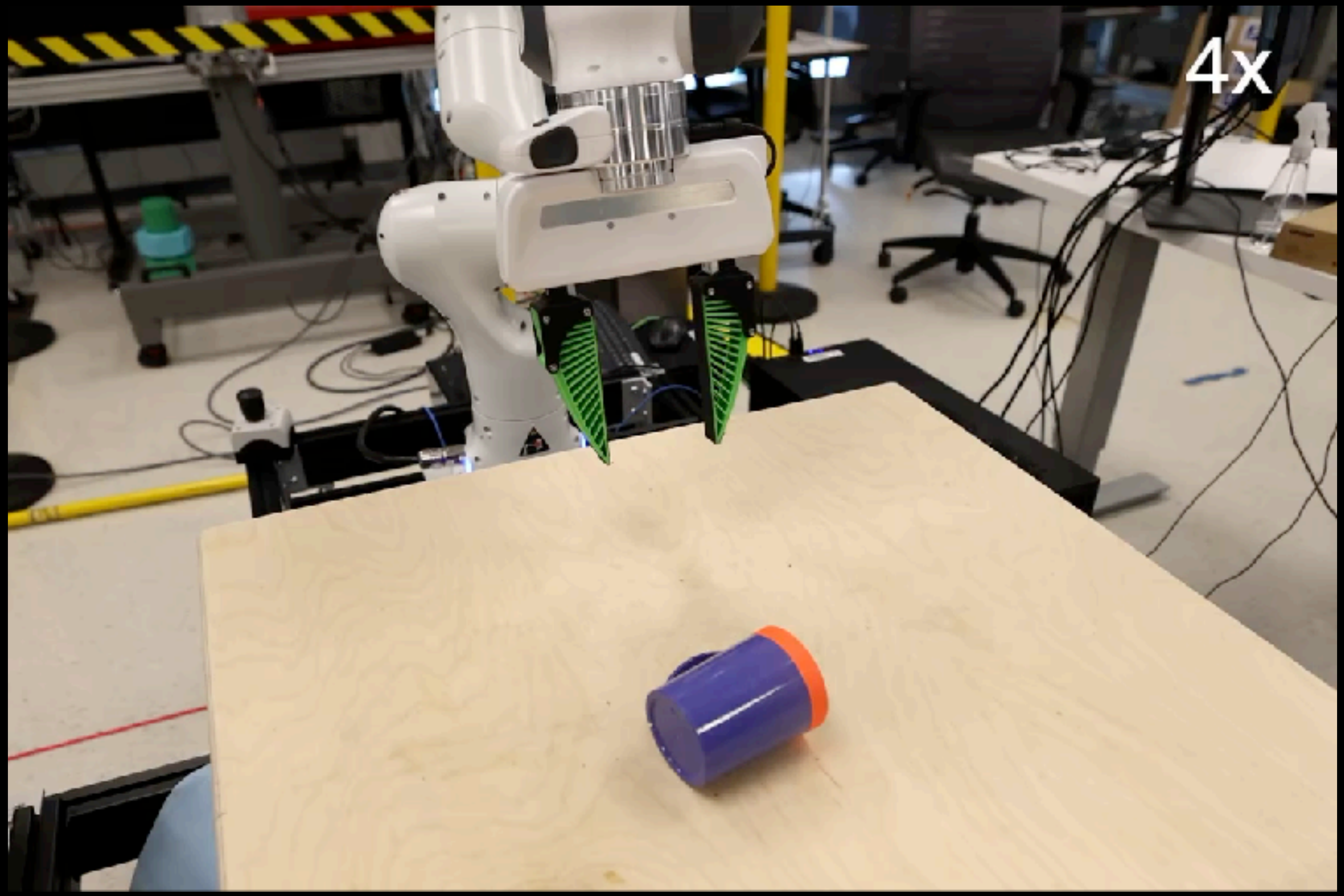
How do I collect demonstrations?

2. Train a policy $\pi : s_t \rightarrow a_t$ on the data

What is my state? Action? Loss?

3. Check validation error on held out dataset

Let's apply Behavior Cloning!



1. Collect data from a human demonstrator

$$[\quad (s_1, a_1^*), \quad (s_2, a_2^*), \quad (s_3, a_3^*), \quad \ldots \quad]$$

How do I collect demonstrations?

2. Train a policy $\pi : s_t \to a_t$ on the data

What is my state? Action? Loss?

3. Check validation error on held out dataset

Why we love Behavior Cloning

It's EASY!

If you can drive down validation error perfectly to 0,
it is *guaranteed* to do what the expert does

It may work often in practice, but ...

Quiz!

Which assumption of supervised learning is most likely to be violated in behavior cloning?

When poll is active respond at **PollEv.com/sc2582**

Send **sc2582** to **22333**

How things go wrong with BC

Today's class

- [x] What is imitation learning?
- [x] Behavior Cloning
- [ ] Feedback drives Covariate Shift

Feedback drives covariate shift

An old problem

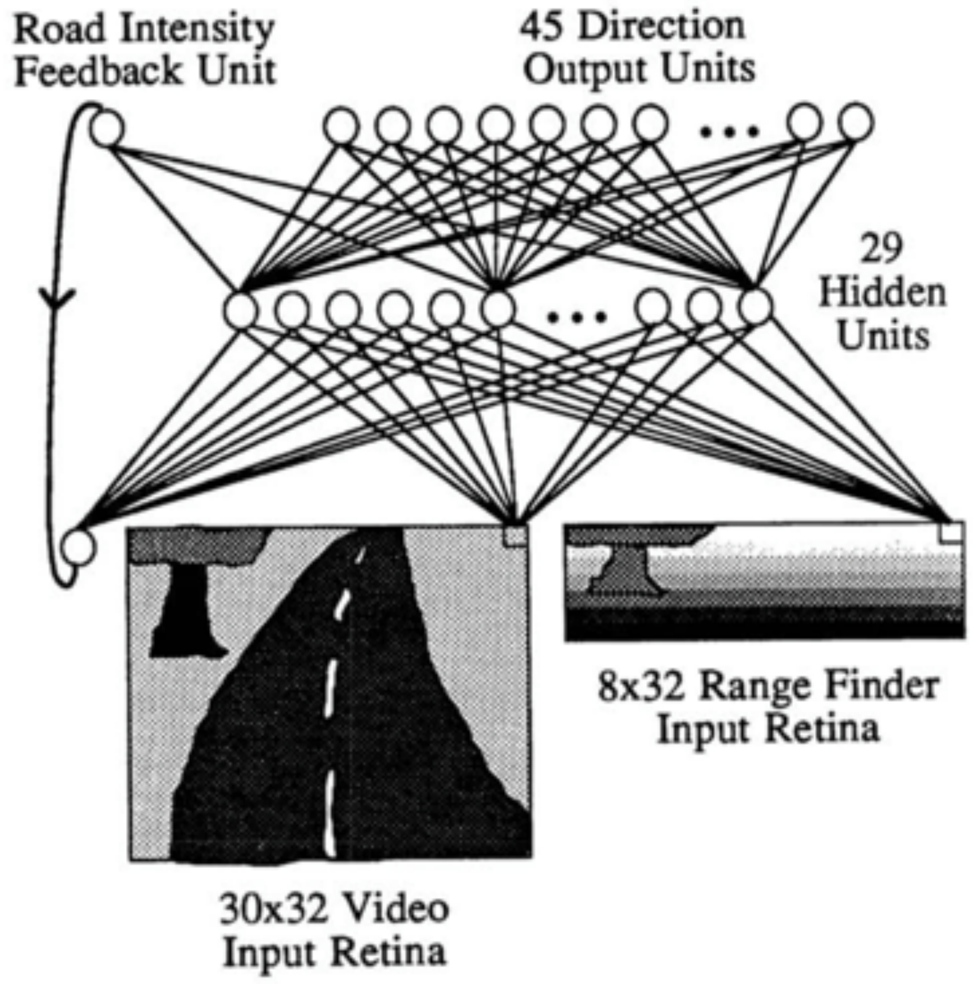


Figure 1: ALVINN Architecture

"...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 ..."

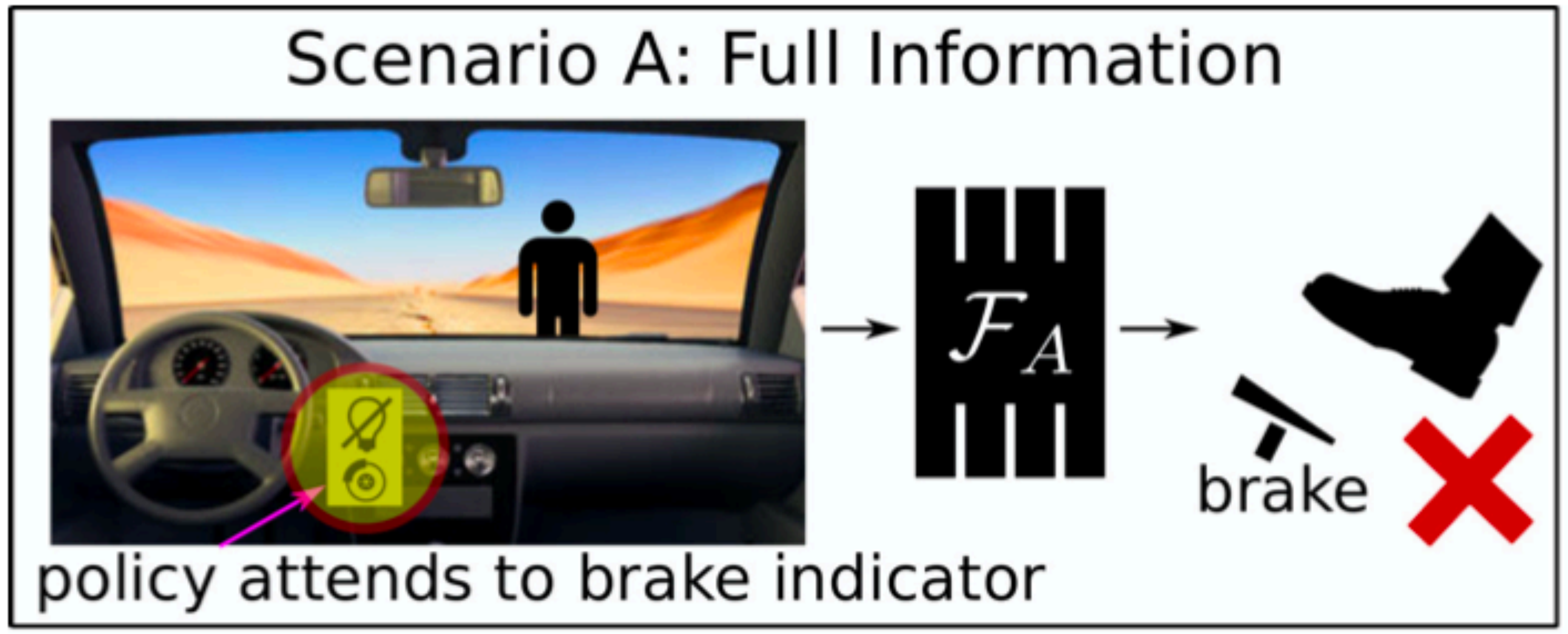
"Exploring the Limitations of Behavior Cloning for Autonomous Driving." F. Codevilla, E. Santana, A. M. Lopez, A. Gaidon. ICCV 2019

"... 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



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

Feedback is a problem for LLMs

Beam Search

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

"The probability of a repeated phrase increases with each repetition, creating a positive feedback loop"

The curious case of neural text de-generation
Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019).

"The main problem is that mistakes made early in the sequence generation process are fed as input to the model and can be quickly amplified because the model might be in a part of the state space it has never seen at training time."

"Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks." Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015).

Thus, the model trained with teacher forcing may over-rely on previously predicted words, which would exacerbate error propagation

"On exposure bias, hallucination and domain shift in neural machine translation." Wang, C., & Sennrich, R. (2020).

DeepMind

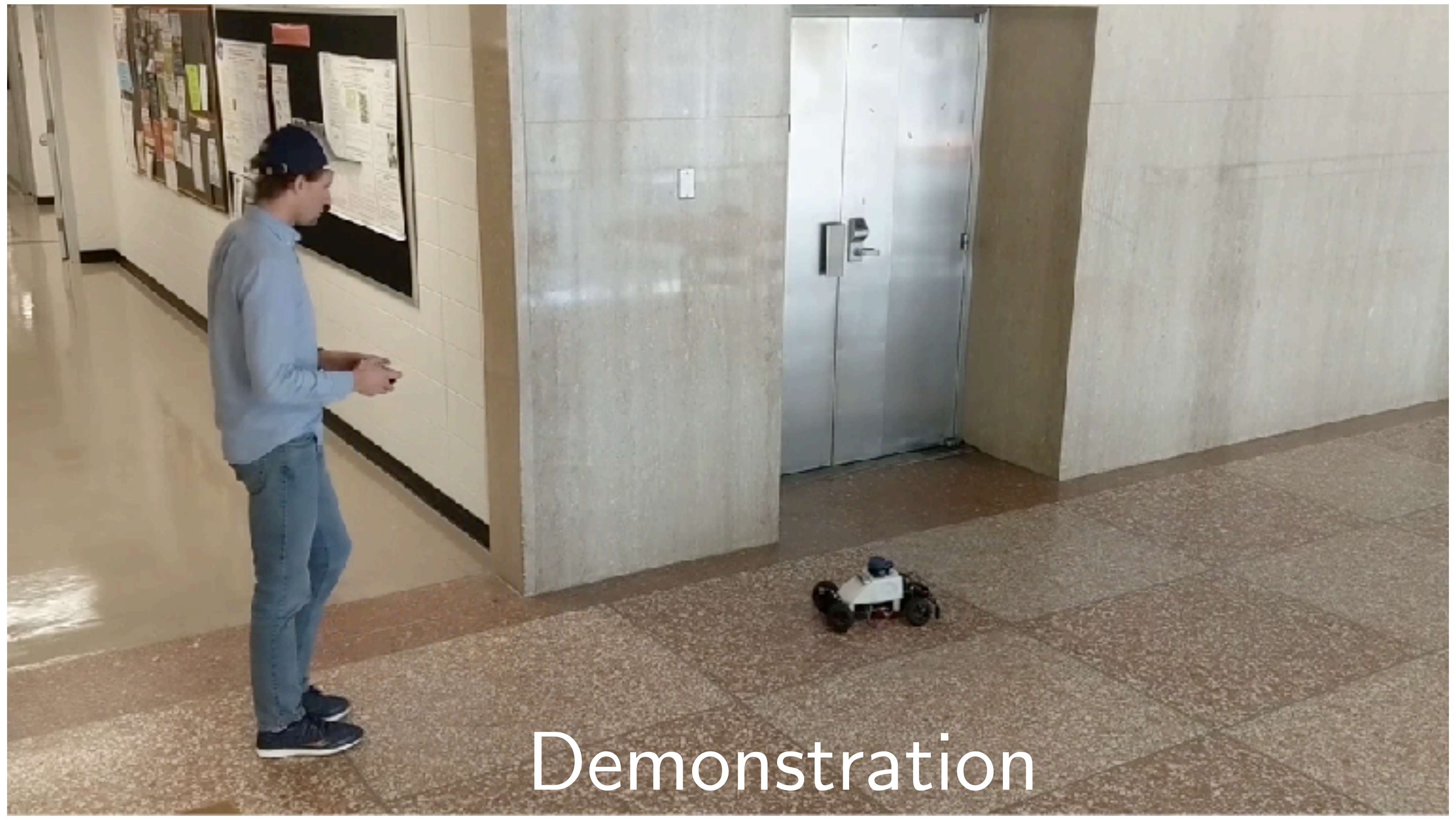
Technical Report
2021-10-22

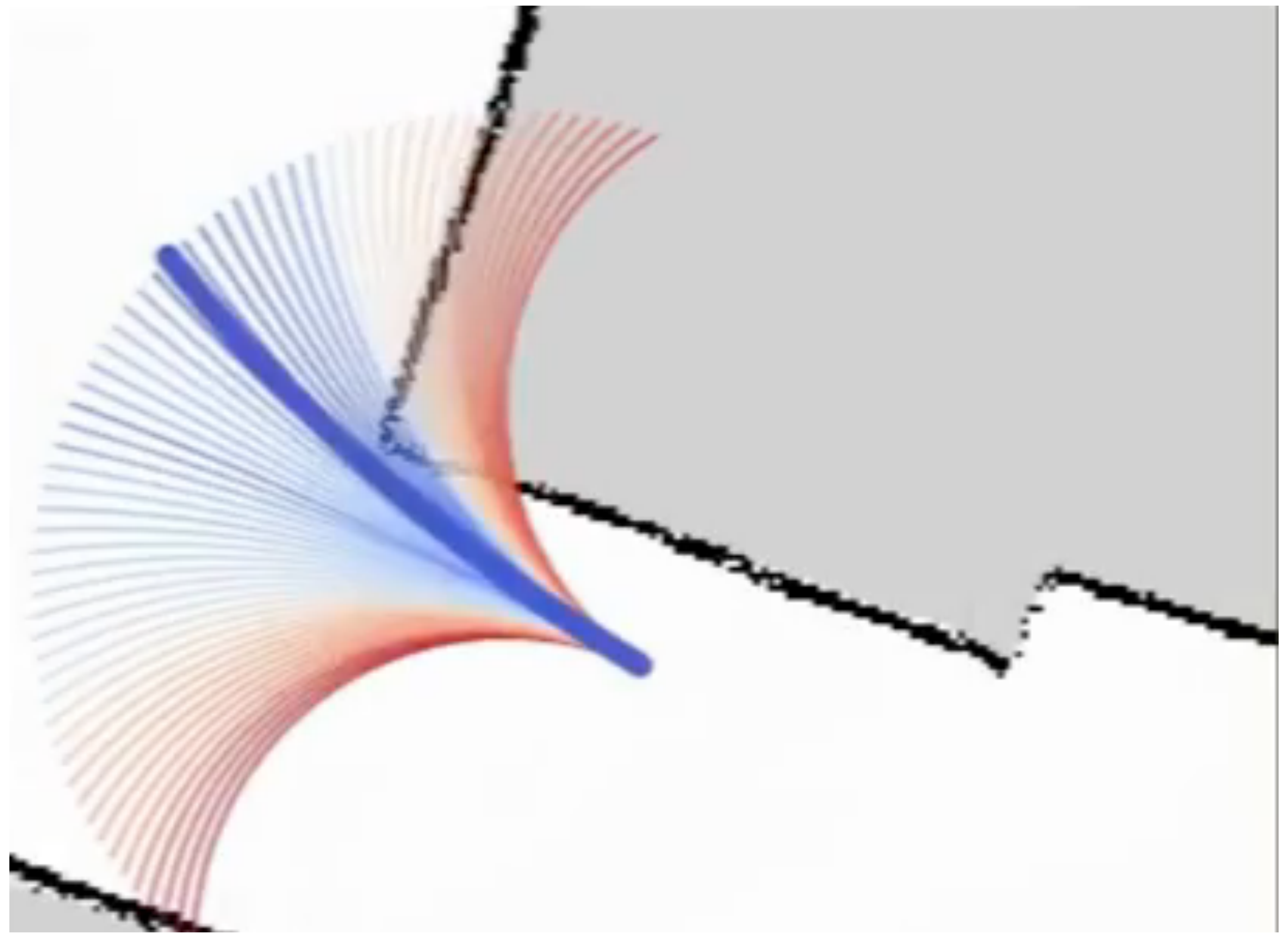
Shaking the foundations: delusions in sequence models for interaction and control

Pedro A. Ortega[*], Markus Kunesch[*], Grégoire Delétang[*], Tim Genewein[*], Jordi Grau-Moya[*], Joel Veness[1], Jonas Buchli[1], Jonas Degrave[1], Bilal Piot[1], Julien Perolat[1], Tom Everitt[1], Corentin Tallec[1], Emilio Parisotto[1], Tom Erez[1], Yutian Chen[1], Scott Reed[1], Marcus Hutter[1], Nando de Freitas[1] and Shane Legg[1]
[*]Deepmind Safety Analysis, [1]DeepMind

Feedback is an old adversary!

[SCB+ RSS'20]

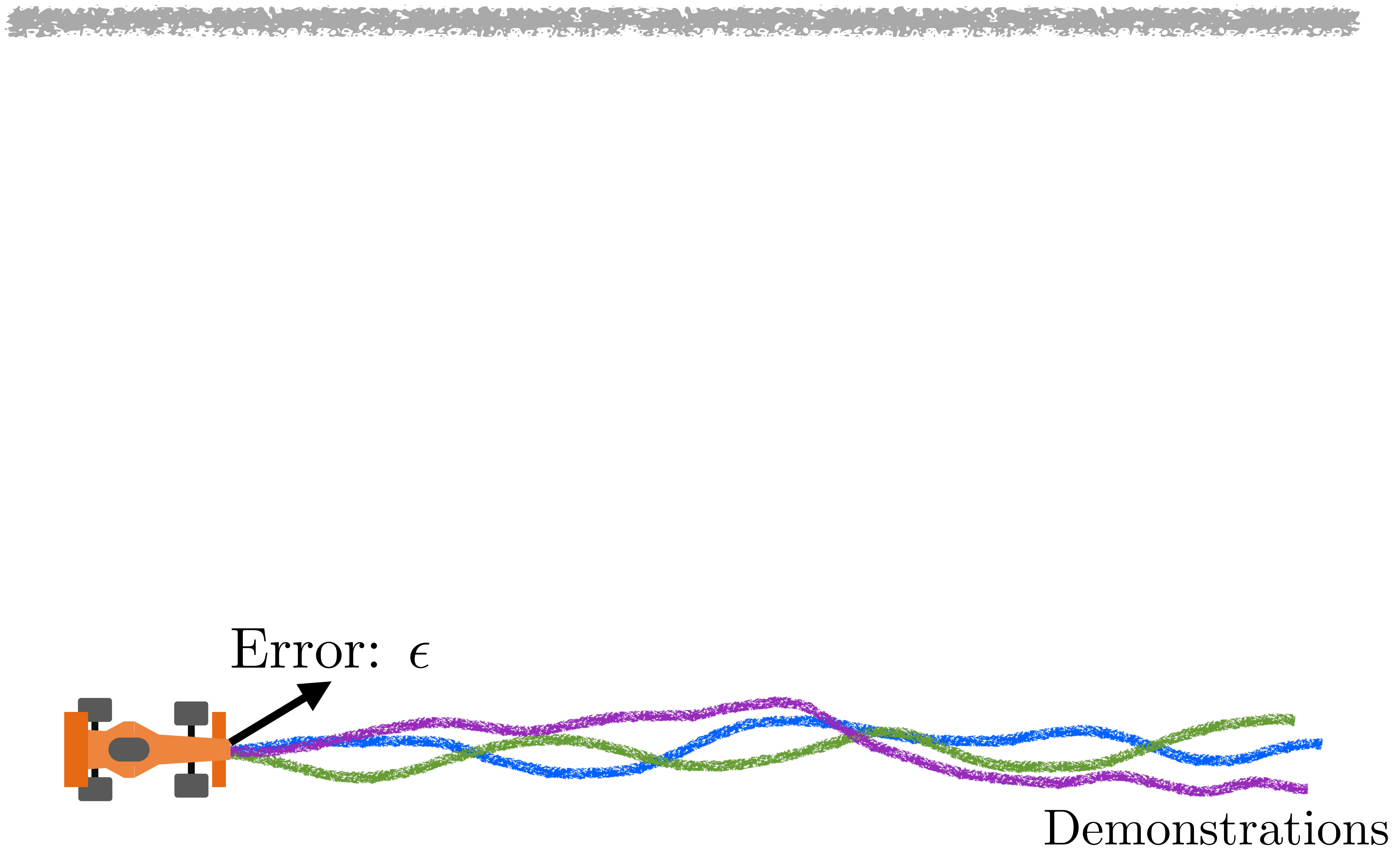




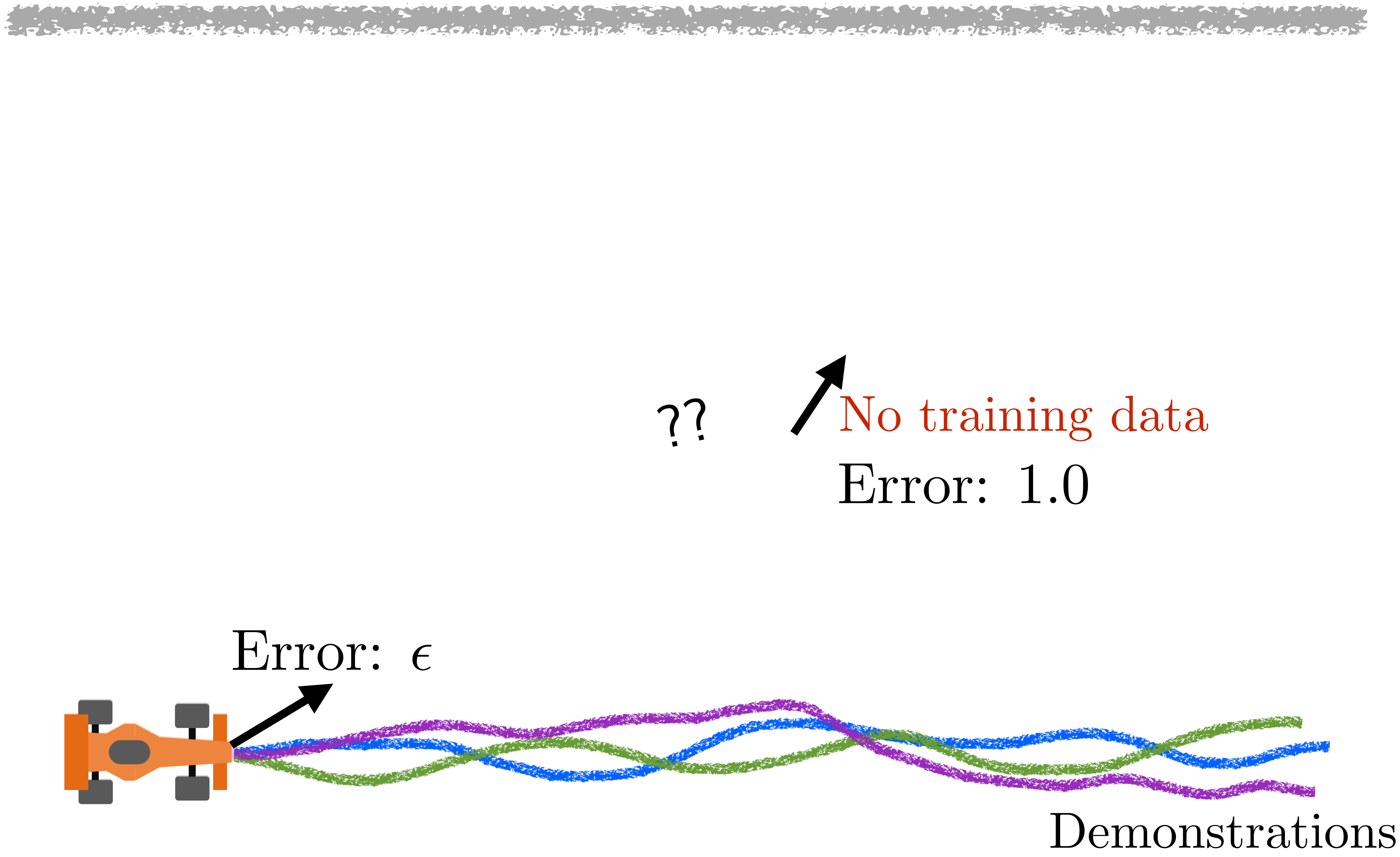
Learnt policy

Behavior Cloning crashes into a wall

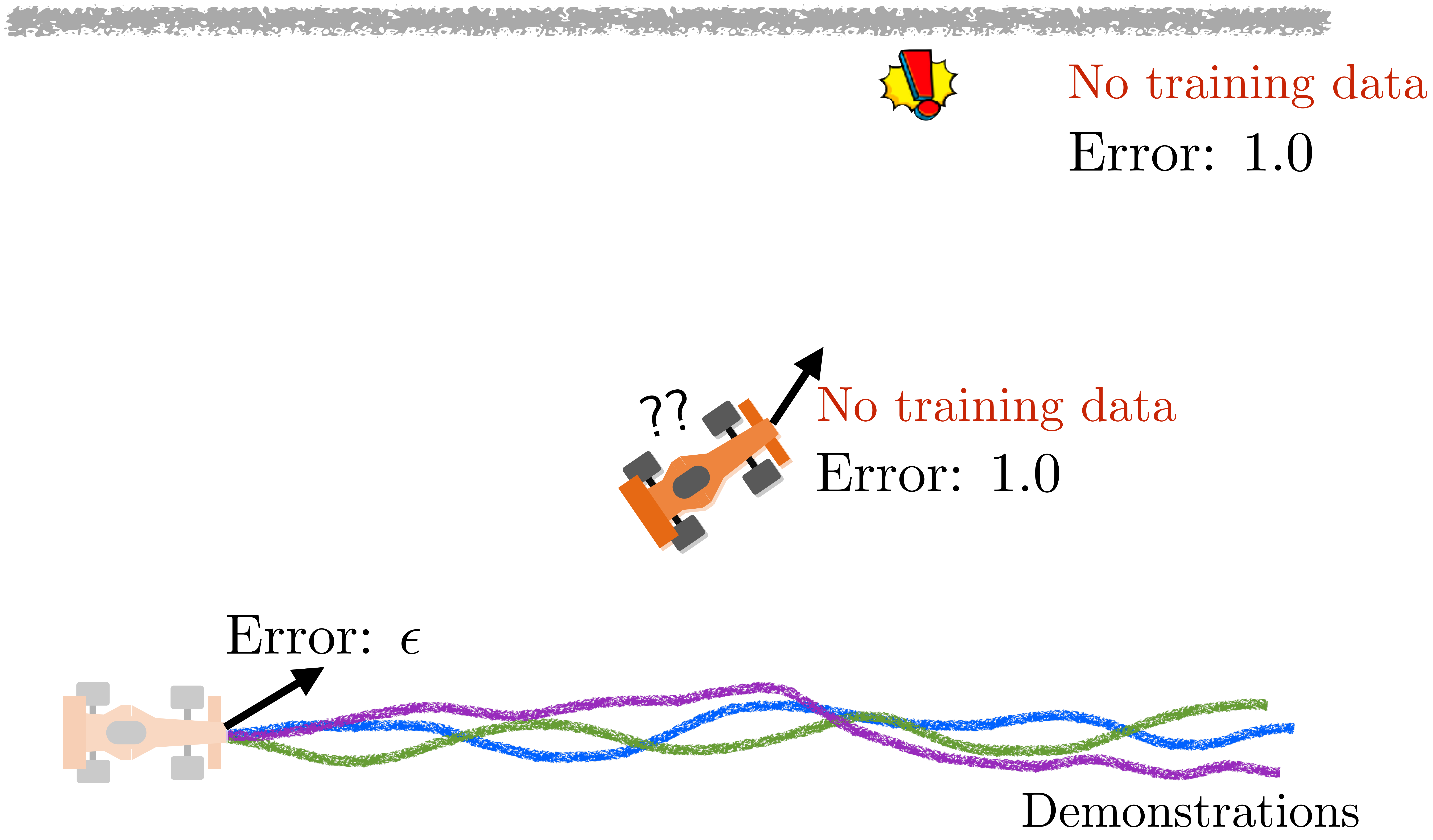
Why did the robot crash?



Why did the robot crash?



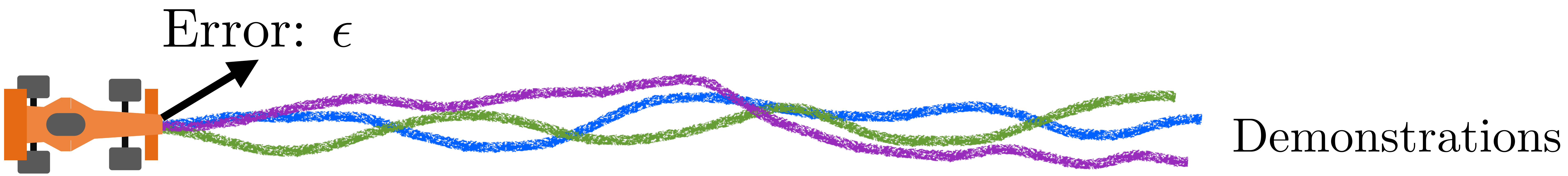
Why did the robot crash?



Train Test Mismatch

Training / Validation Loss

$$\sum_{t=0}^{T-1} \mathbb{E}_{s_t \sim d_t^{\pi^\star}}[\ell(s_t, \pi(s_t))]$$



$$\neq$$

Test Loss

$$\sum_{t=0}^{T-1} \mathbb{E}_{s_t \sim d_t^{\pi}}[\ell(s_t, \pi(s_t))]$$

Today's class

- ☑ What is imitation learning?
- ☑ Behavior Cloning
- ☑ Feedback drives Covariate Shift

Can we mathematically quantify how much worse BC is compared to the demonstrator?

First, let's define performance of a policy

$$J(\pi) = \mathbb{E}_{\substack{a_t \sim \pi(s_t) \\ s_{t+1} \sim \mathcal{T}(s_t, a_t)}} \left[\sum_{t=0}^{T-1} c(s_t, a_t) \right]$$

(Performance)

Second, let's define performance difference

$$J(\pi) - J(\pi^*)$$

(Performance of my learner) (Performance of my demonstrator)

We want to *minimize* the performance difference

How low can we drive performance difference?

Let's say my learner is not perfect and can only drive down training / validation error to be ε

😄 The best we can hope for is that error grows **linearly** in time

$$J(\pi) - J(\pi^*) \leq O(\epsilon T)$$

😱 The worst case is if error **compounds quadratically** in time

$$J(\pi) - J(\pi^*) \leq O(\epsilon T^2)$$

Behavior cloning hits the worst case!

There exists an MDP where BC
has a performance difference of $O(\epsilon T^2)$

We are going to such a MDP next week,
and you will see more in A1!