# Advances in machine learning???

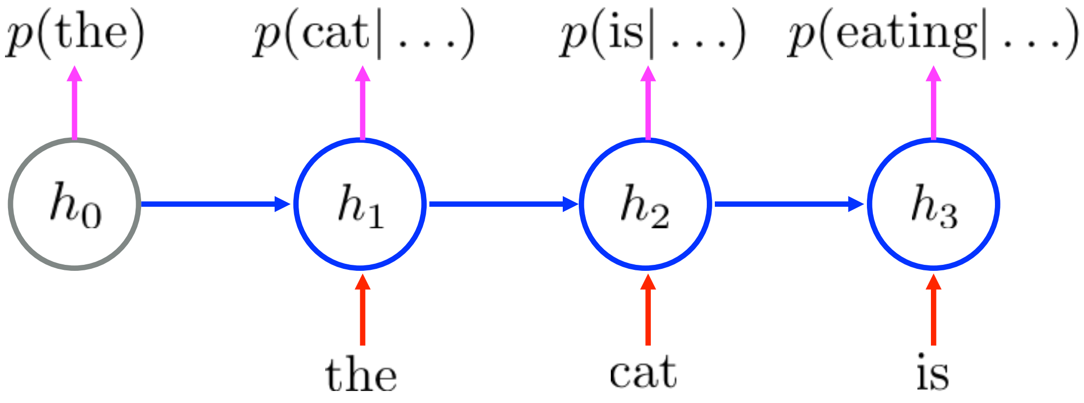
## Beyond maximum likelihood estimation and supervised learning

Kyunghyun Cho

# Supervised learning for sequence modelling

- Given a ground-truth trajectory, maximize the predictability of a next action: $\max \log p(x_t | x_{<t})$
- Maximum (log-)likelihood estimation
- Two issues
  1. Weak correlation with a true reward
  2. Mismatch between training and inference

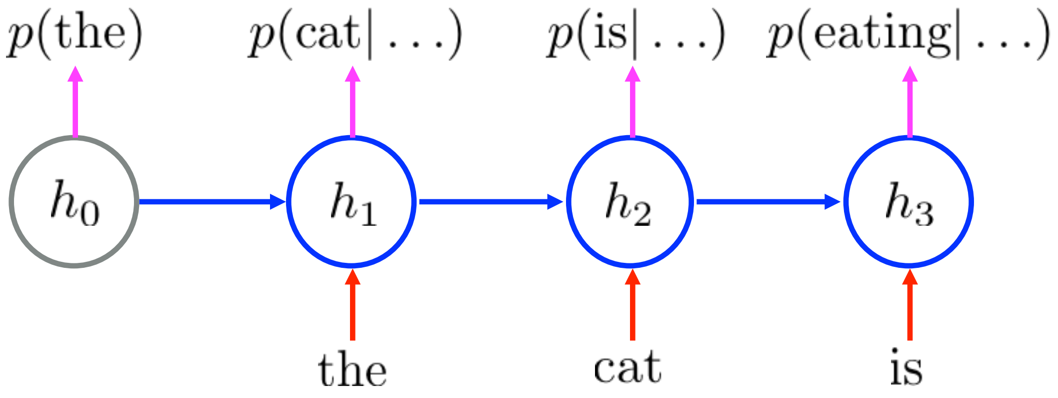
*Example)* $p(\text{the}, \text{cat}, \text{is}, \text{eating})$



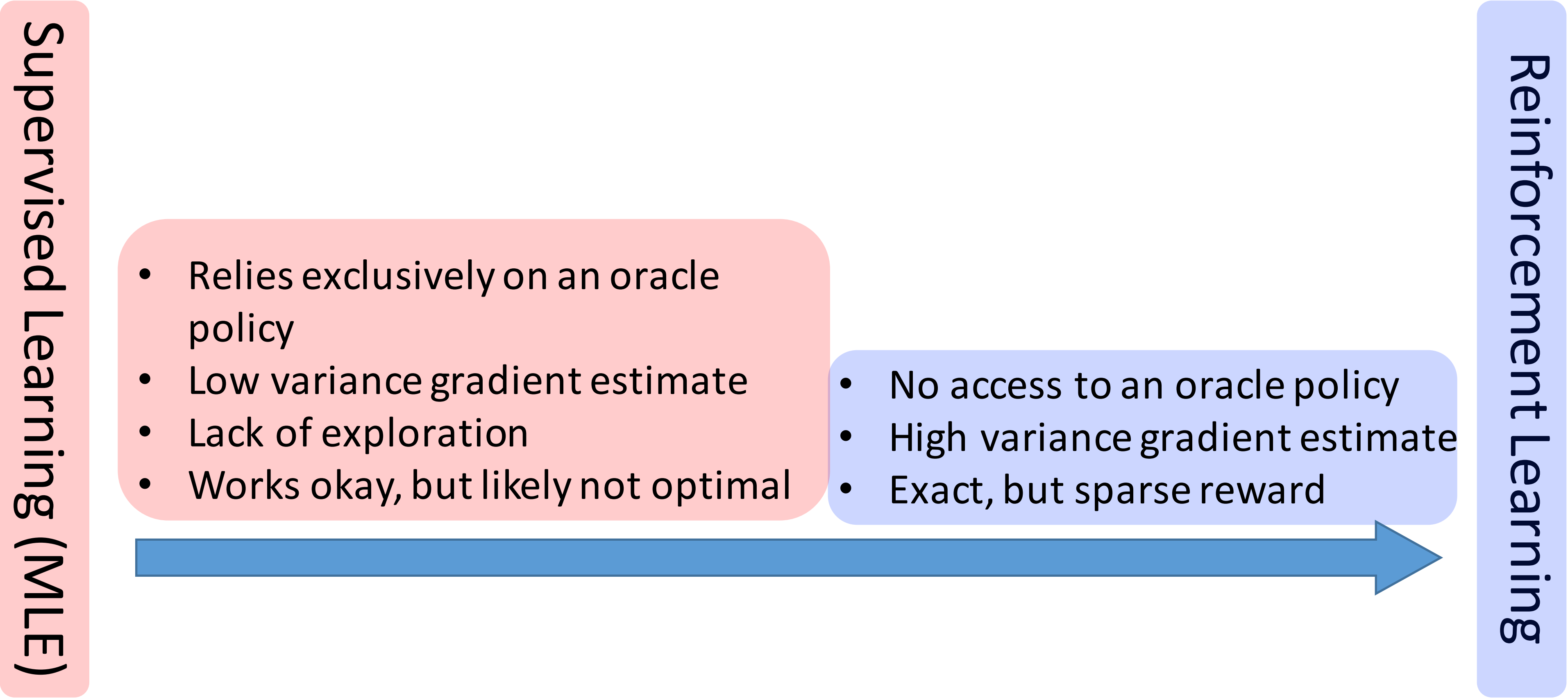
# Reinforcement learning

- Maximize a true reward instead of probabilities
- Inference is a part of training: better match between these two
- Q-learning, REINFORCE, actor-critic, …
- Great, except that
  1. Sparse reward
  2. High variance of gradient estimate
  3. Difficult balance between exploration and exploitation

*Example)* $p(\text{the}, \text{cat}, \text{is}, \text{eating})$



# Active Imitation learning as an intermediate step



Supervised Learning (MLE)

- Relies exclusively on an oracle policy
- Low variance gradient estimate
- Lack of exploration
- Works okay, but likely not optimal

- No access to an oracle policy
- High variance gradient estimate
- Exact, but sparse reward

Reinforcement Learning

# Active Imitation learning as an intermediate step

## DAgger

1. Initialize/pre-train a policy with supervised learning
2. Let the policy drive, while collecting the oracle's decisions
3. Retrain a policy with the aggregated data
4. Iterate 2 – 3 until convergence
5. [Finetune with reinforcement learning]

**Supervised learning**

**Exploration**

**Easier, because most action sequences end up with some positive reward**

[Ross et al., 2011; and others]

# Active Imitation learning as an intermediate step

DAgger

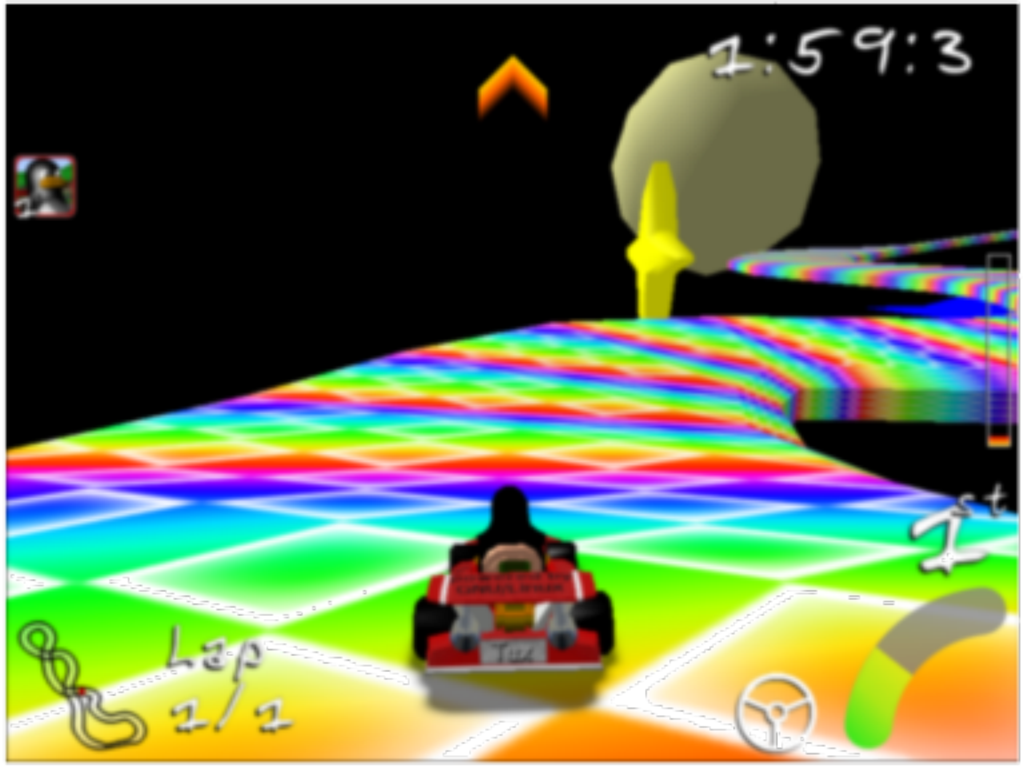
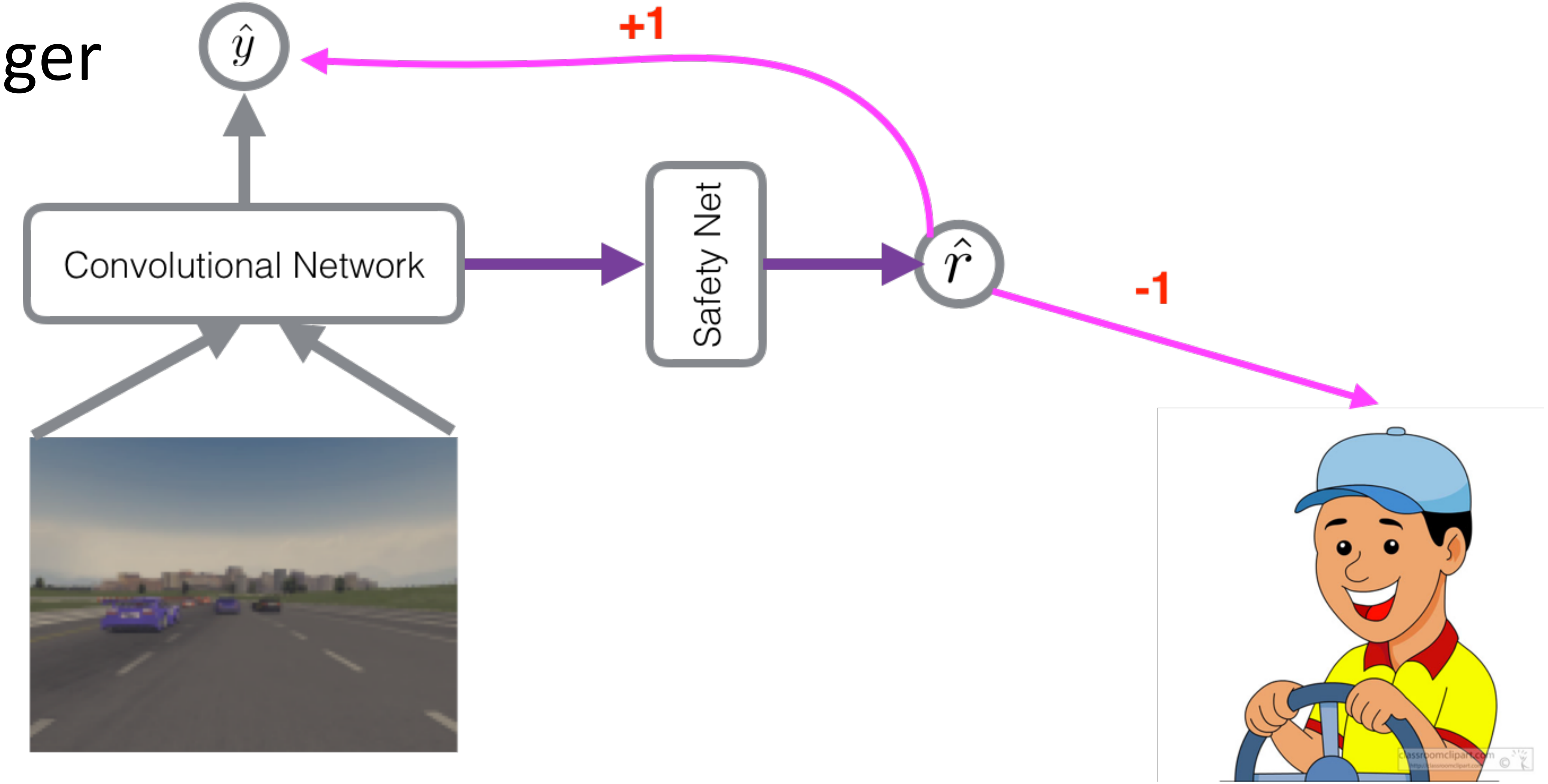


Figure 1: Image from Super Tux Kart's Star Track.

[Ross et al., 2011; and others]

# **Safer** Active Imitation learning as an intermediate step

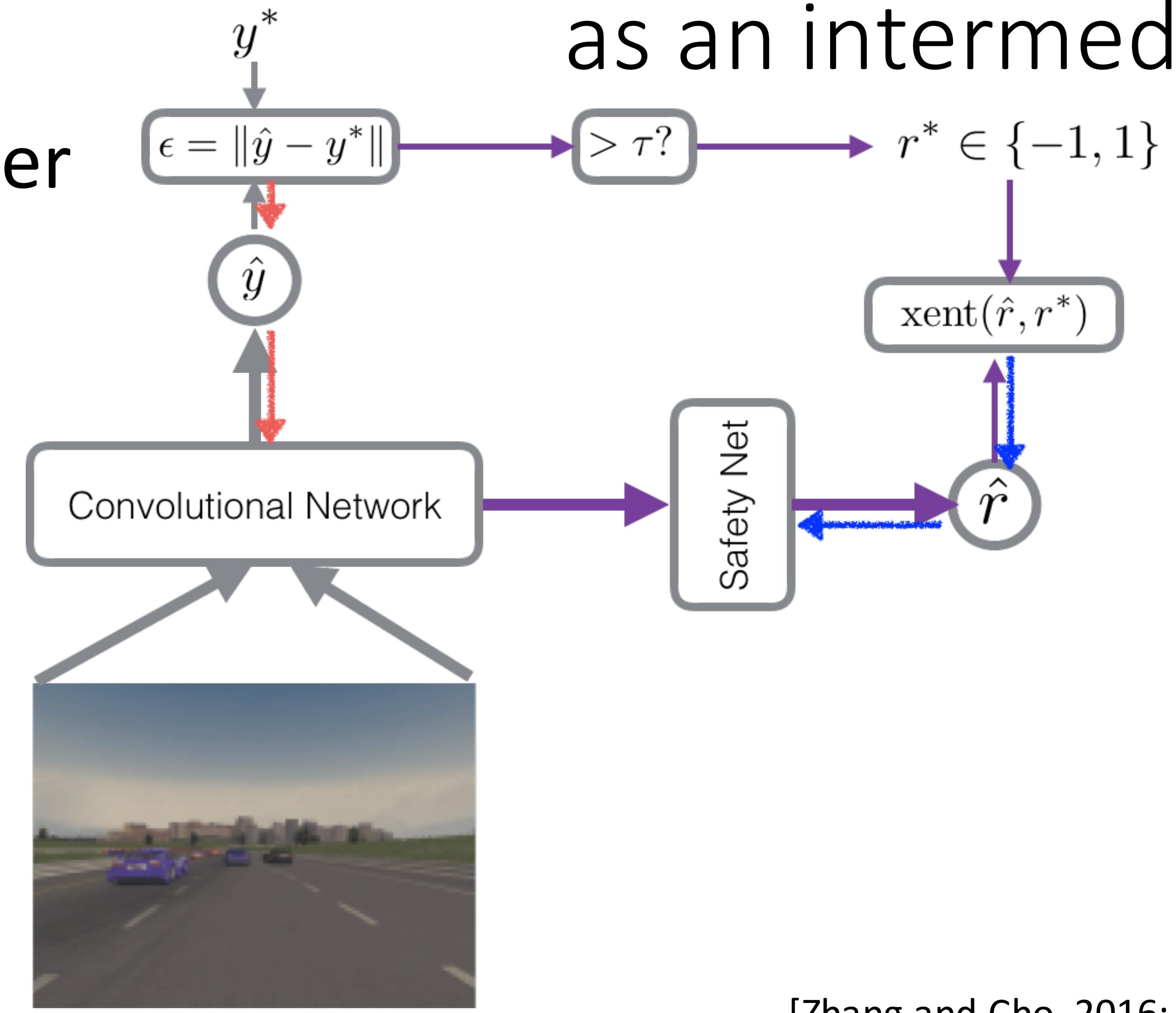
SafeDAgger



[Zhang and Cho, 2016; Laskey et al., 2016]

# **Safer** Active Imitation learning as an intermediate step

SafeDAgger



[Zhang and Cho, 2016; Laskey et al., 2016]

# **Safer** Active Imitation learning as an intermediate step

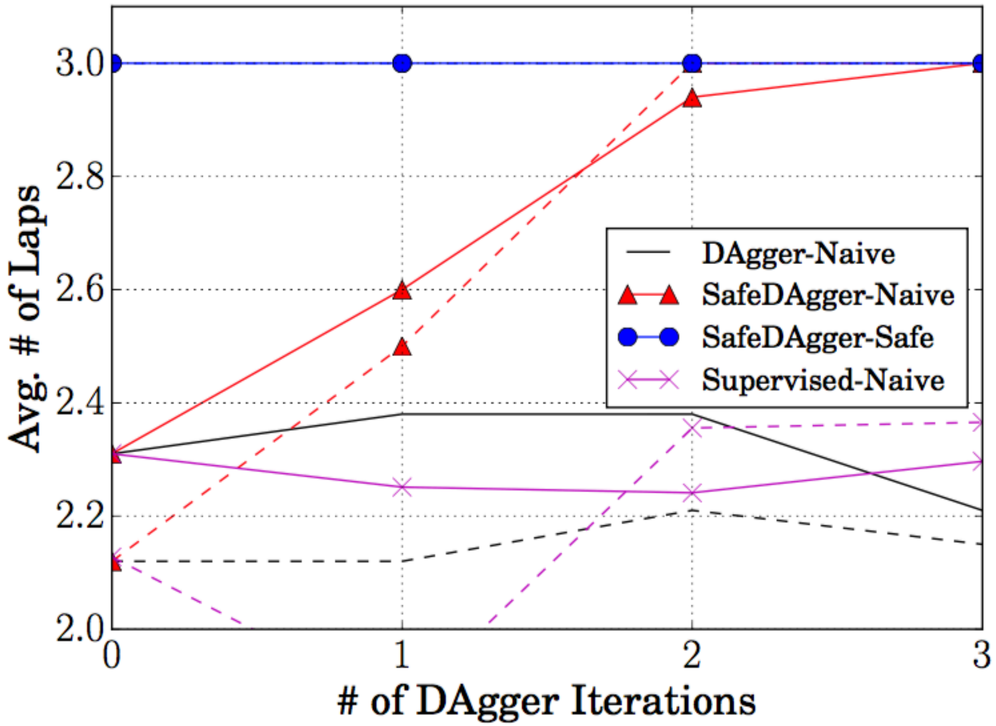
## SafeDAgger

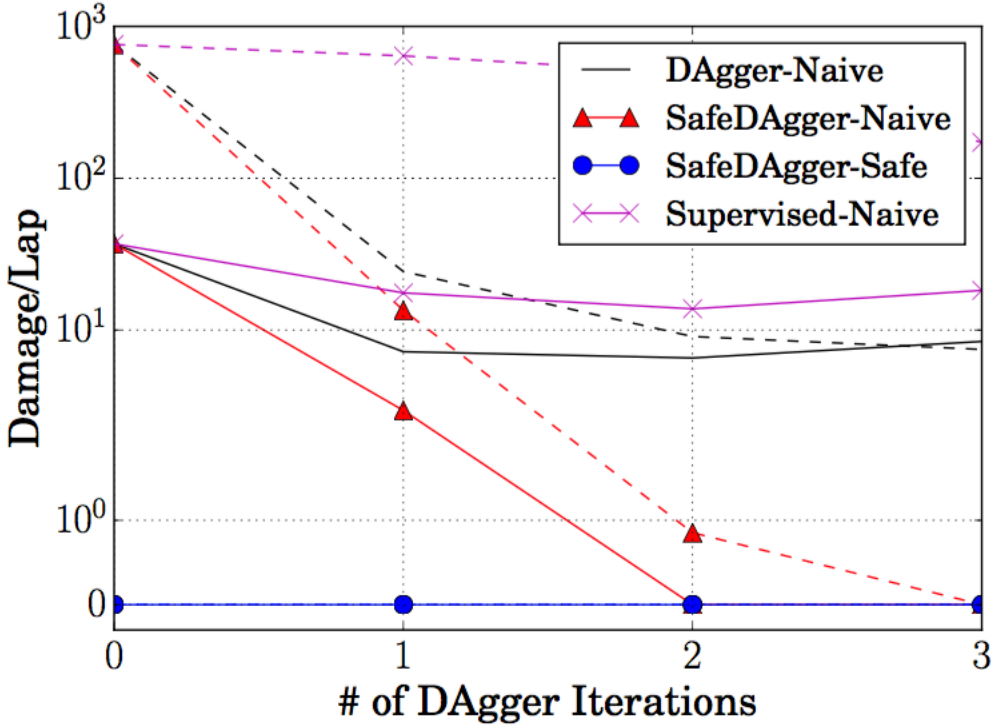
1. Initialize/pre-train a policy with supervised learning
2. Let the policy drive
3. Collect a data point only when it's *not safe*
4. Retrain a policy with the aggregated data
5. Iterate 2 – 3 until convergence
6. [Finetune with reinforcement learning]

[Zhang and Cho, 2016; Laskey et al., 2016]

# **Safer** Active Imitation learning as an intermediate step

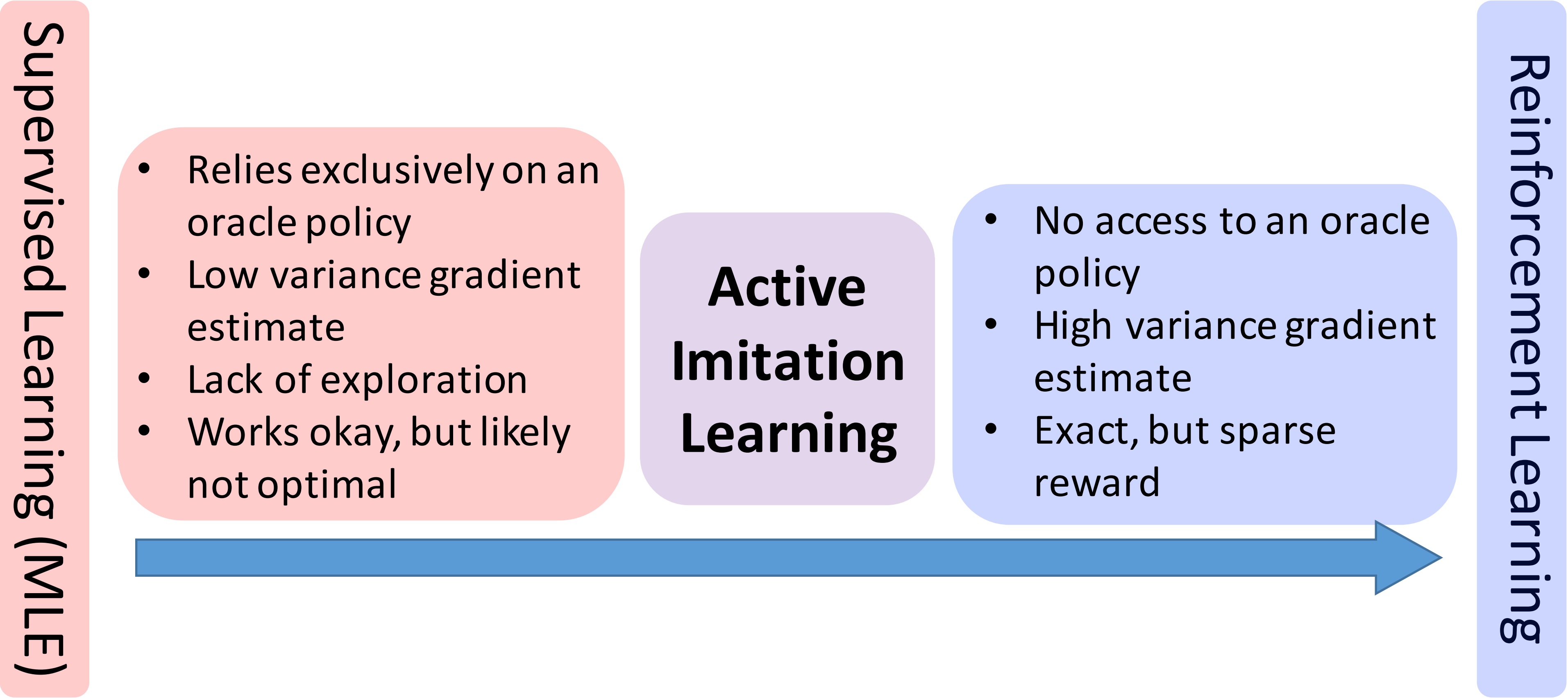
## SafeDAgger





[Zhang and Cho, 2016]

# Active Imitation learning as an intermediate step



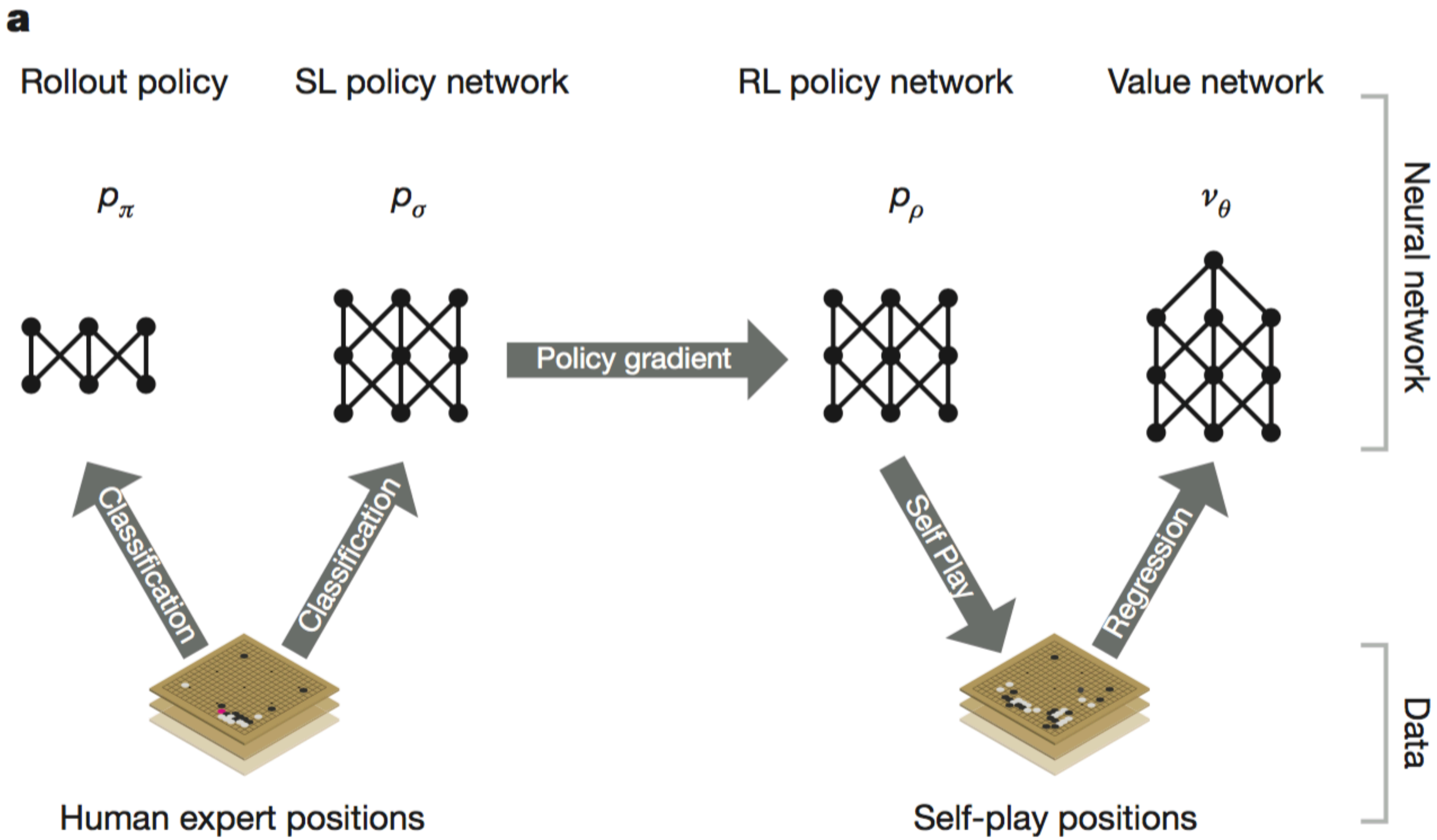
Supervised Learning (MLE)

- Relies exclusively on an oracle policy
- Low variance gradient estimate
- Lack of exploration
- Works okay, but likely not optimal

**Active Imitation Learning**

- No access to an oracle policy
- High variance gradient estimate
- Exact, but sparse reward

Reinforcement Learning



[Silver et al., Nature 2016]

# Strong learning systems are expected to be

Patchwork of many learning algorithms

- **Unsupervised learning**:
  Efficient learning of state representation
- **Supervised learning**:
  Efficient learning of action repre [illegible]
  Stable, focused learning o [illegible] between state and action
- **Active learning**:
  Making sup [illegible] ning more robust to mistakes
- **Reinfo [illegible] arning**:
  [illegible] ased on a true reward and test time inference algorithm.

**Is this how our brain learns?**