Interactive Imitation Learning

Sham Kakade and Wen Sun

CS 6789: Foundations of Reinforcement Learning

Announcements

Presentation: Dec 2, 7, 9

Final report: NeurIPS format
Maximum 9 pages for main tex (not including references and appendix)

Recap

Offline IL and Hybrid Setting:

Recap

Offline IL and Hybrid Setting:

Ground truth reward $r(s, a) \in [0,1]$ is unknown;
assume expert is a near optimal policy $\pi^\star$

Recap

Offline IL and Hybrid Setting:

Ground truth reward $r(s, a) \in [0,1]$ is unknown;
assume expert is a near optimal policy $\pi^\star$

We have a dataset $\mathscr{D} = (s_i^\star, a_i^\star)_{i=1}^M \sim d^{\pi^\star}$

Today:

Interactive Imitation Learning Setting

Today:

Interactive Imitation Learning Setting

Key assumption:
we can query expert $\pi^\star$ at any time and any state during training

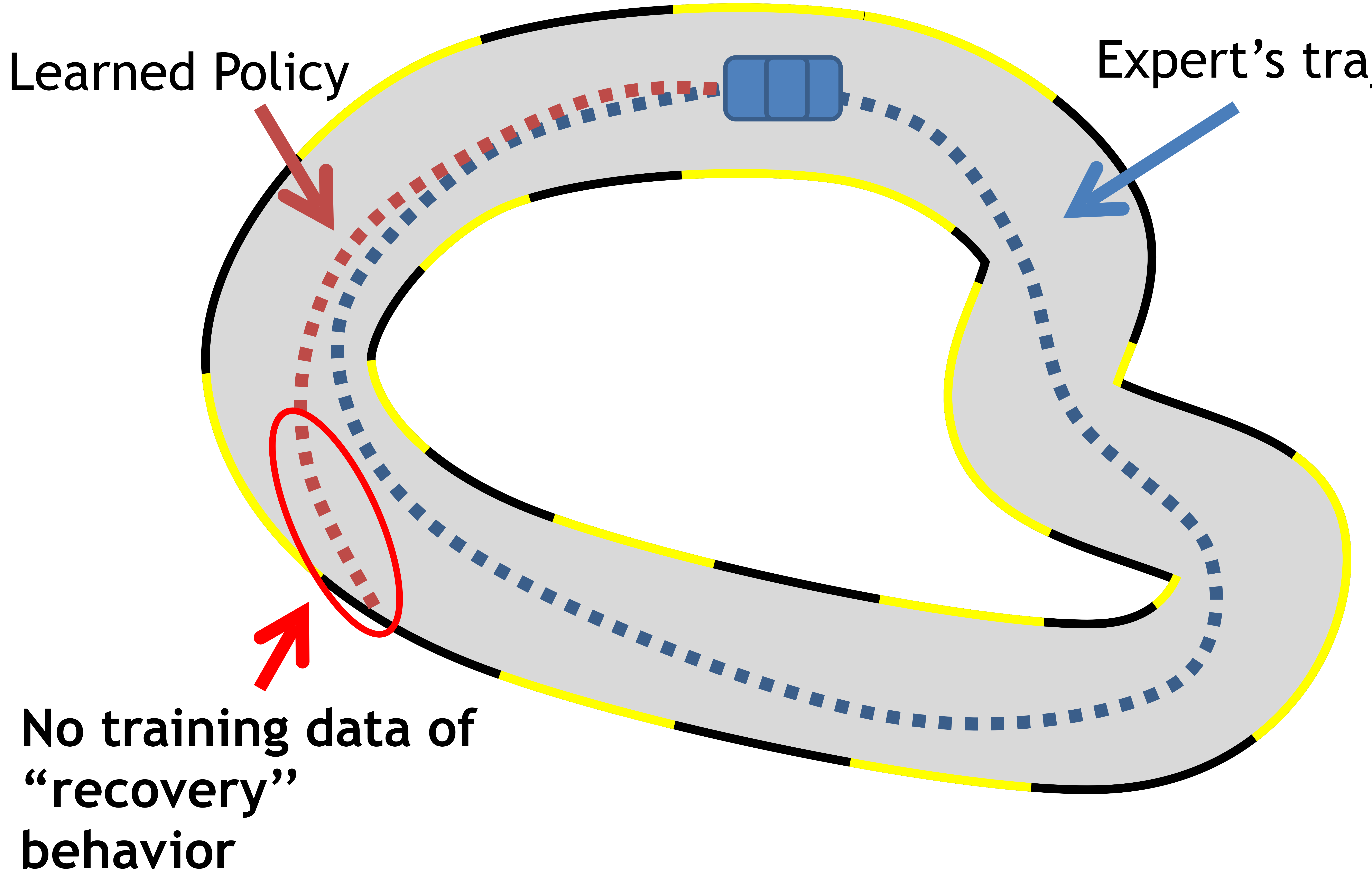
Today:

Interactive Imitation Learning Setting

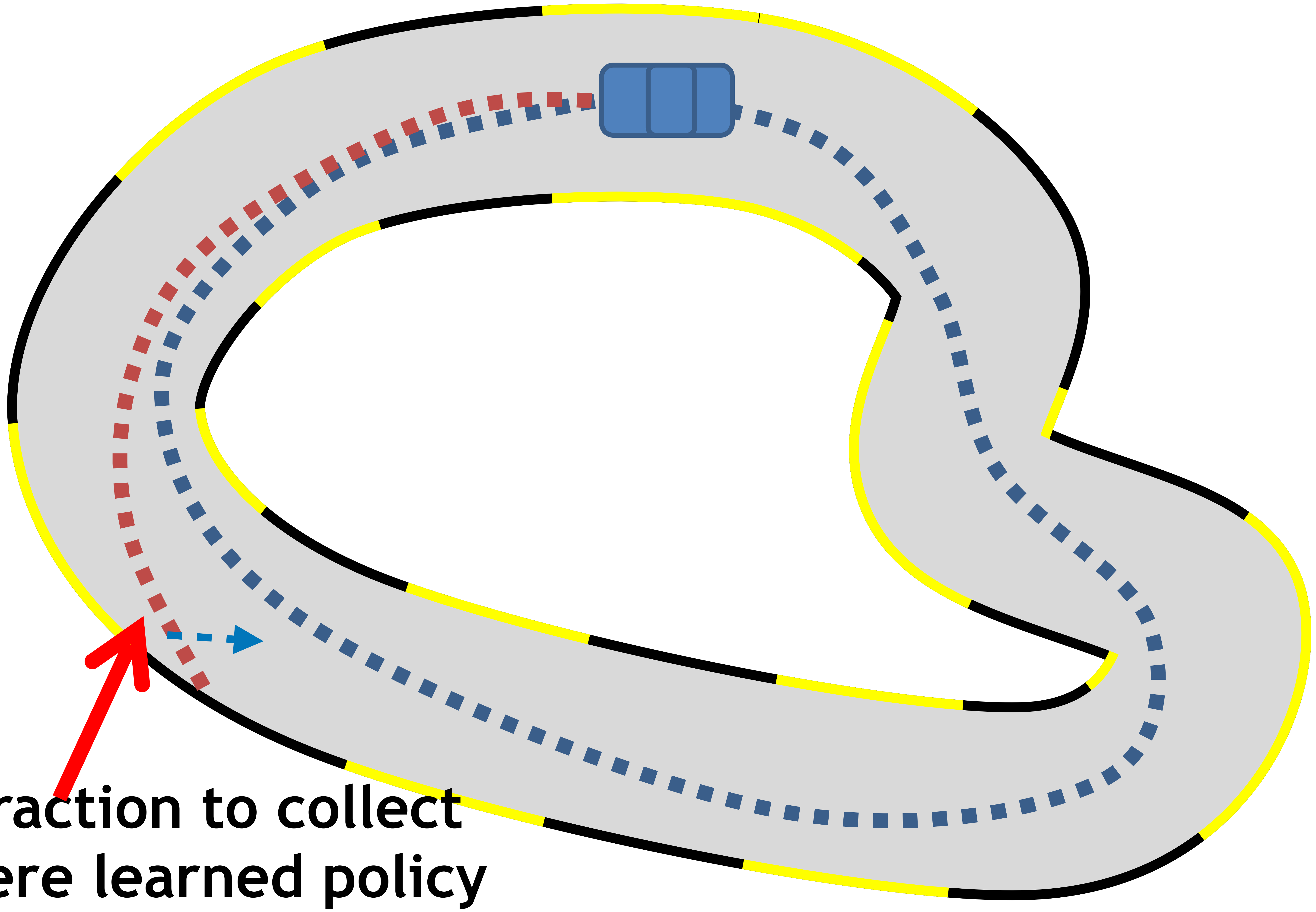
Key assumption:
we can query expert $\pi^\star$ at any time and any state during training

(Recall that previously we only had an offline dataset $\mathscr{D} = (s_i^\star, a_i^\star)_{i=1}^M \sim d^{\pi^\star}$)

Recall the Main Problem from Behavior Cloning:

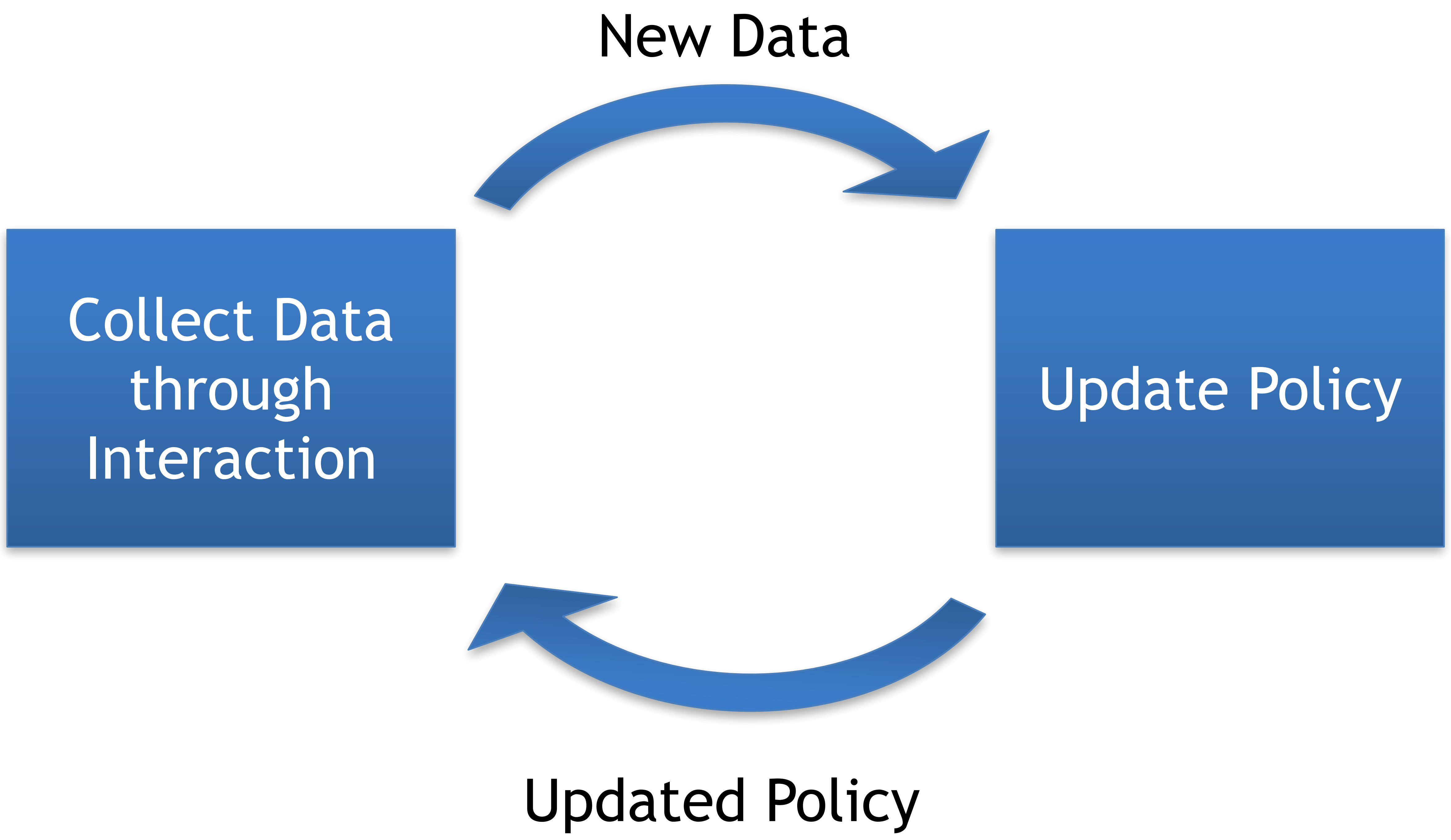


Intuitive solution: **Interaction**



Use interaction to collect data where learned policy goes

General Idea: Iterative Interactive Approach

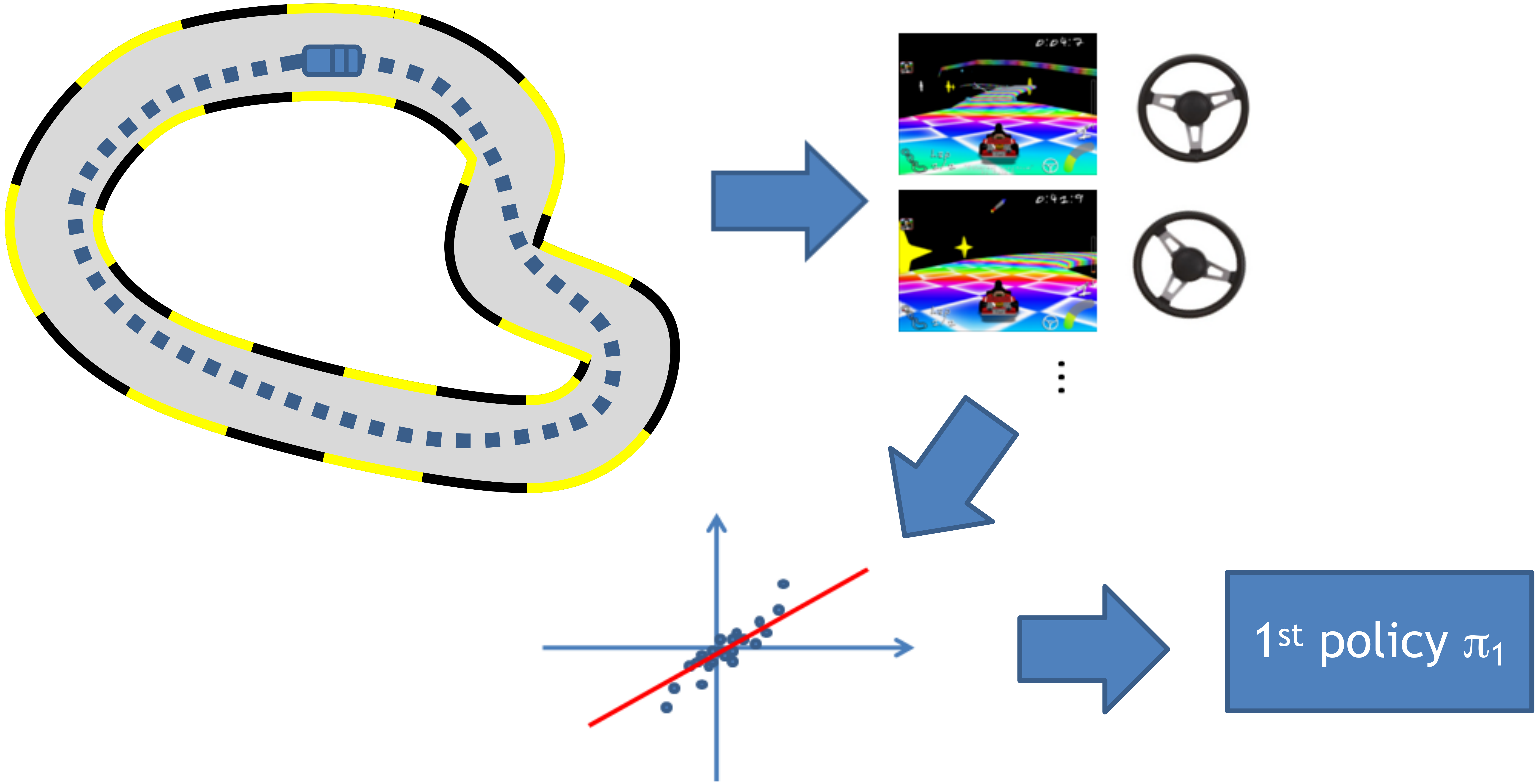


[Ross11a]

DAgger: Dataset Aggregation

0th iteration

Expert Demonstrates Task **Dataset**

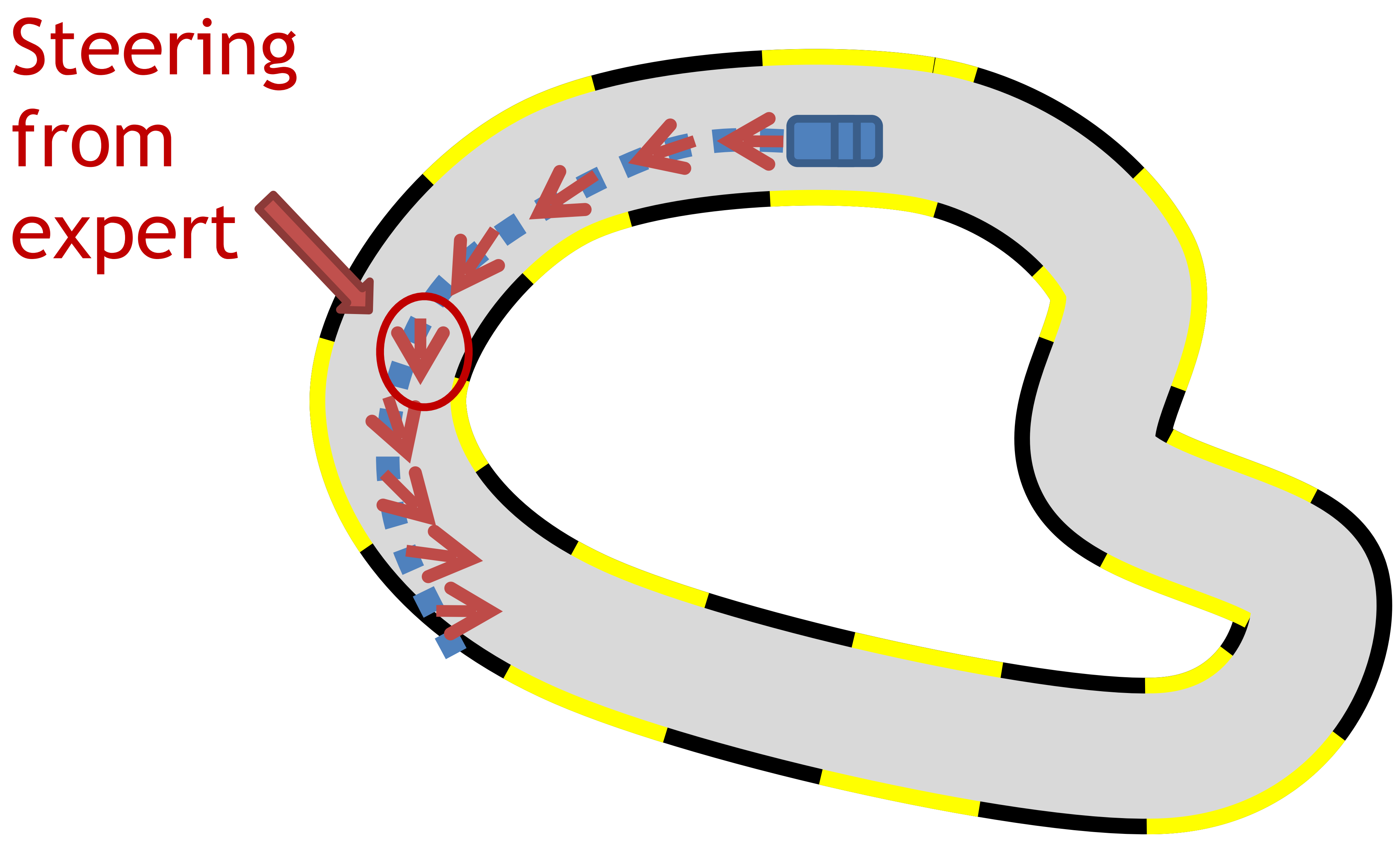


Supervised Learning

DAgger: Dataset Aggregation [Ross11a]

1st iteration

Execute π_1 and Query Expert

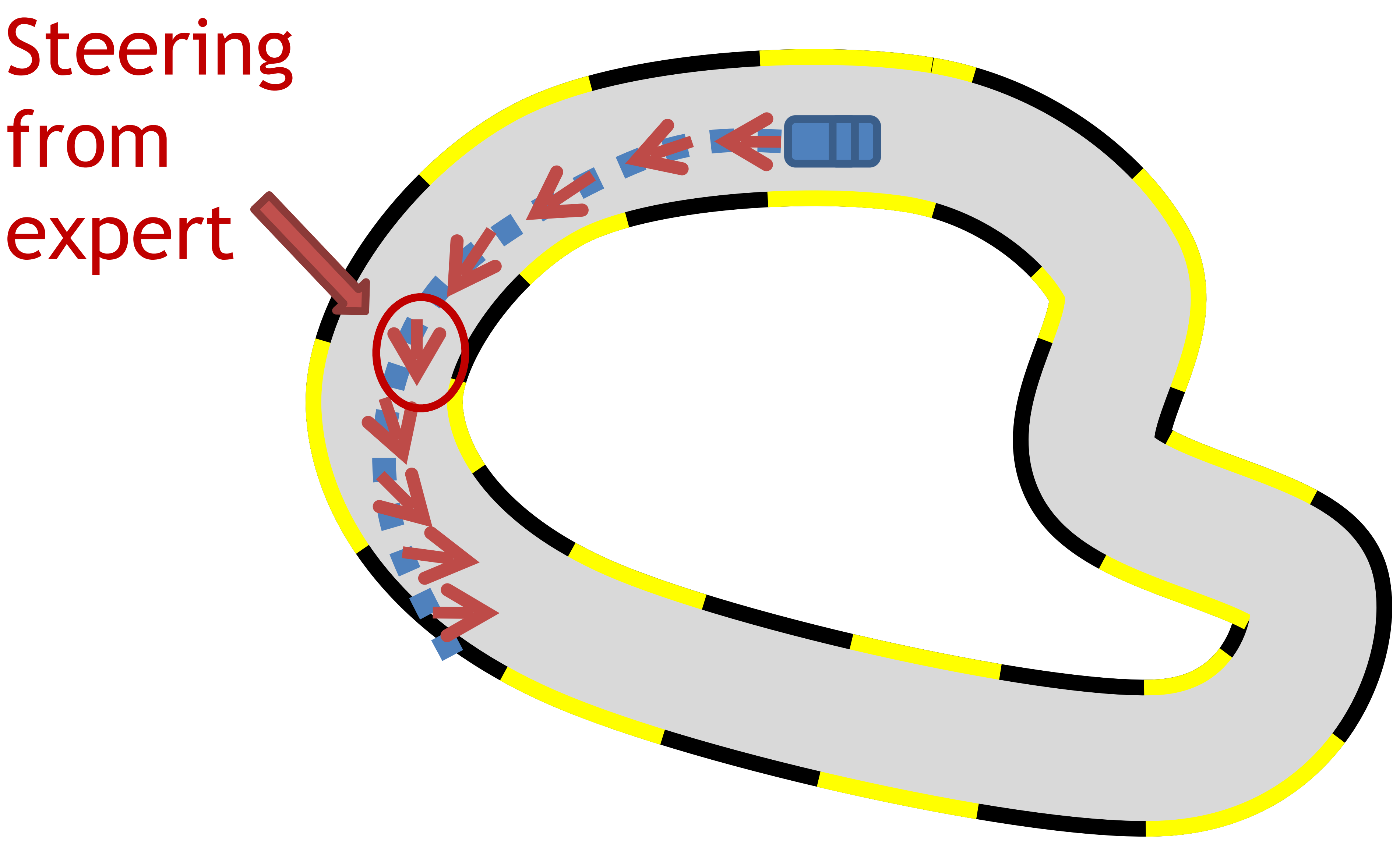


DAgger: Dataset Aggregation

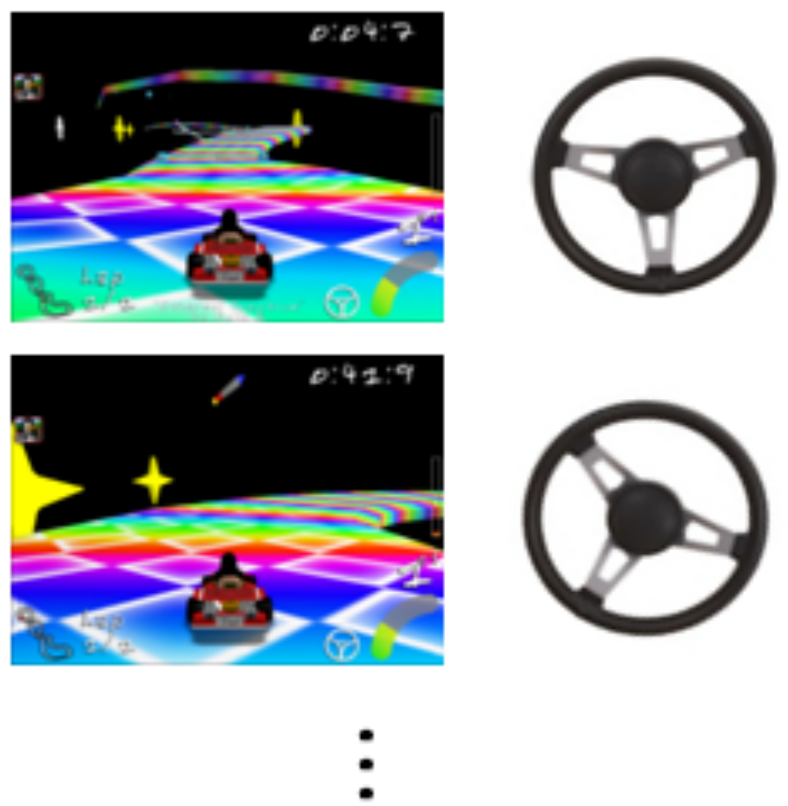
[Ross11a]

1st iteration

Execute π_1 and Query Expert



New Data



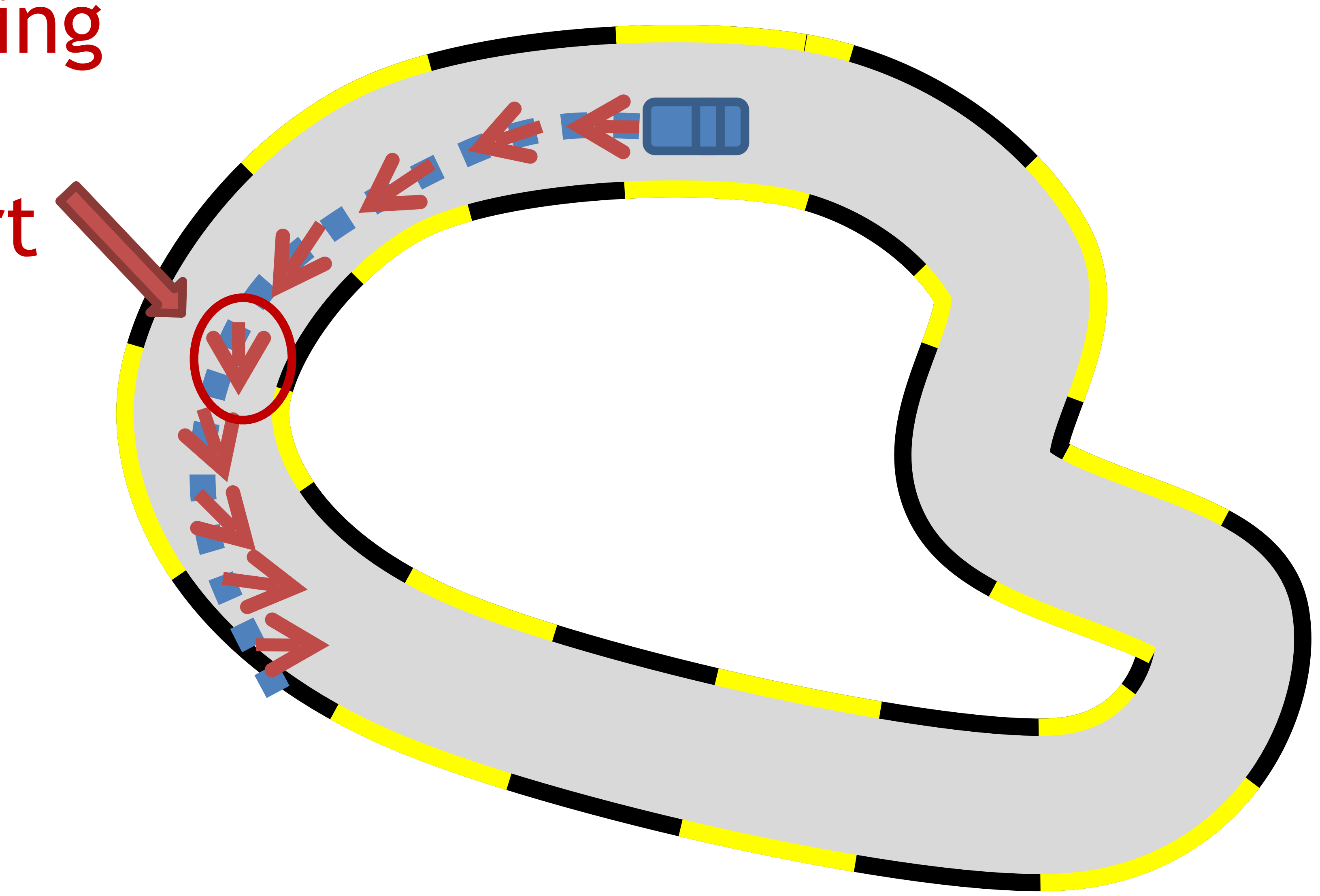
DAgger: Dataset Aggregation

[Ross11a]

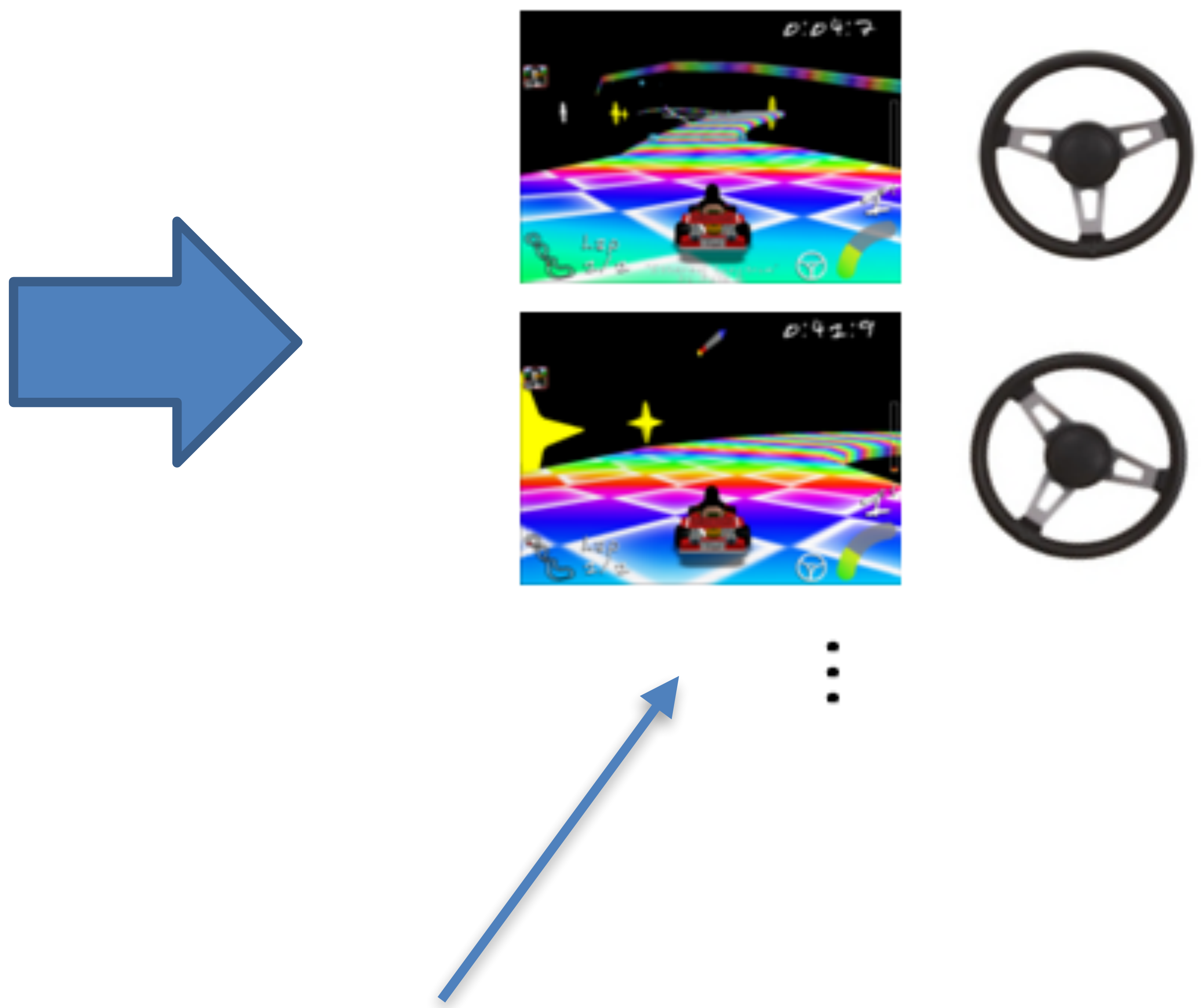
1st iteration

Execute π_1 and Query Expert

Steering
from
expert



New Data



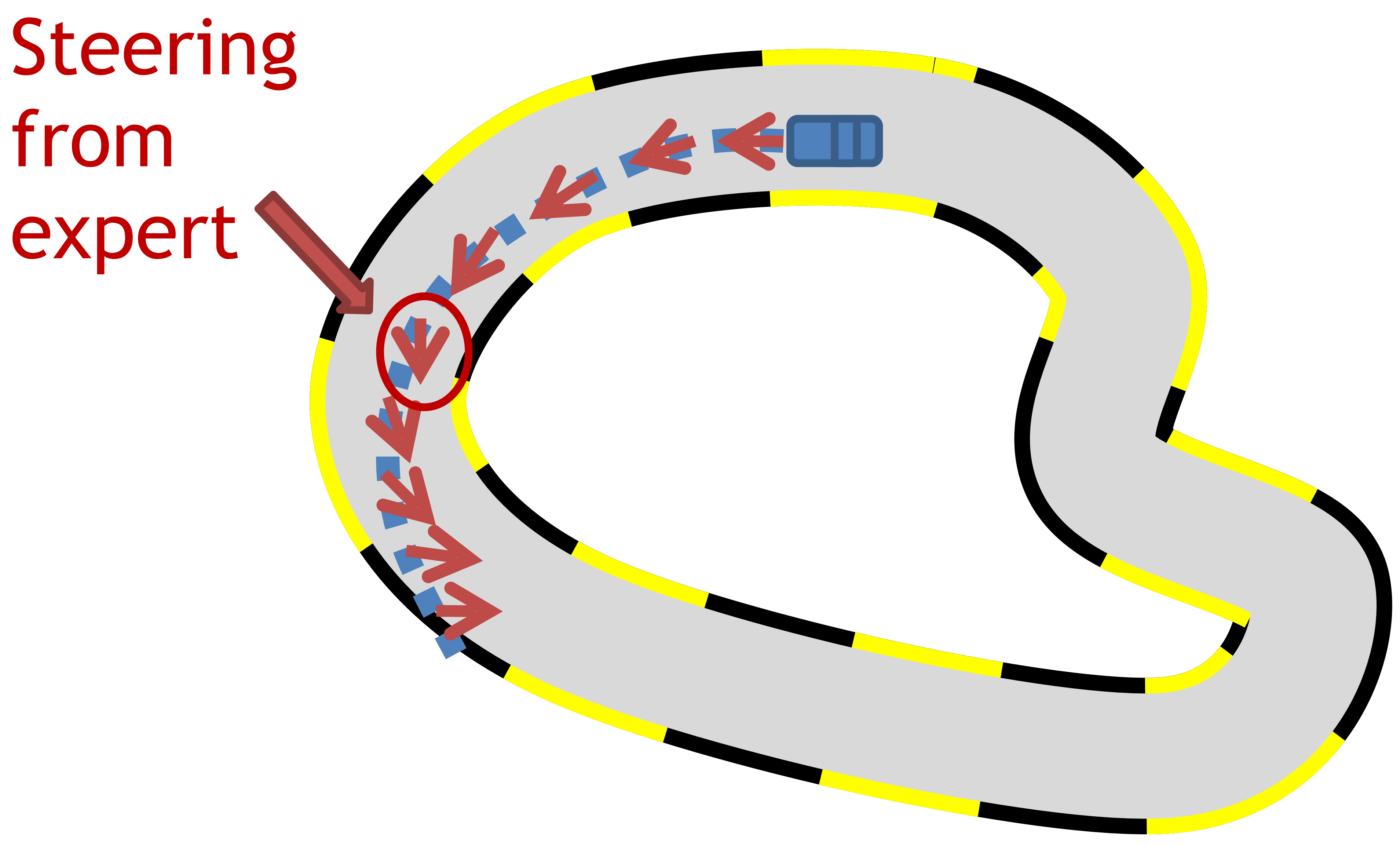
States from
the learned policy

[Ross11a]

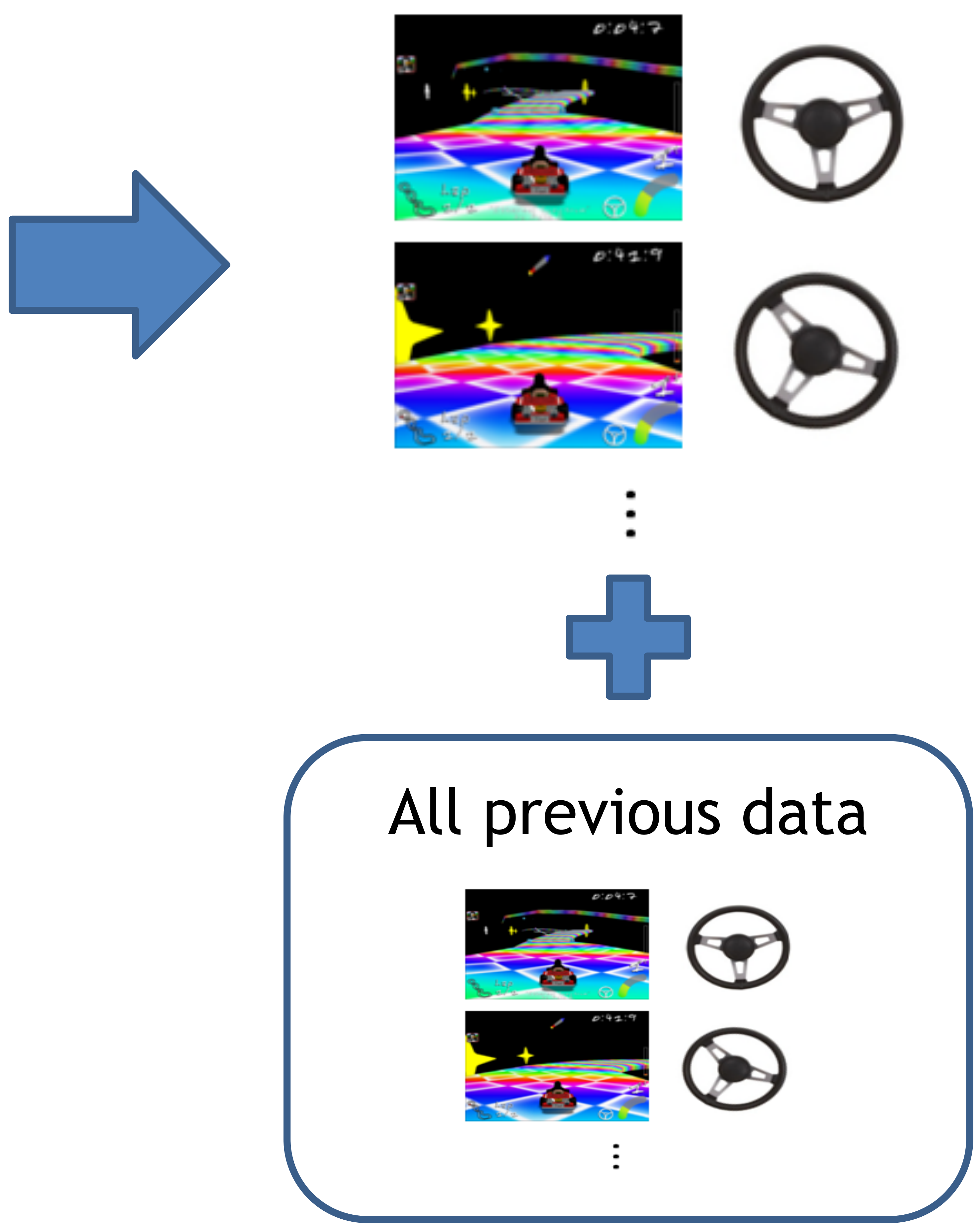
DAgger: Dataset Aggregation

1st iteration

Execute π_1 and Query Expert



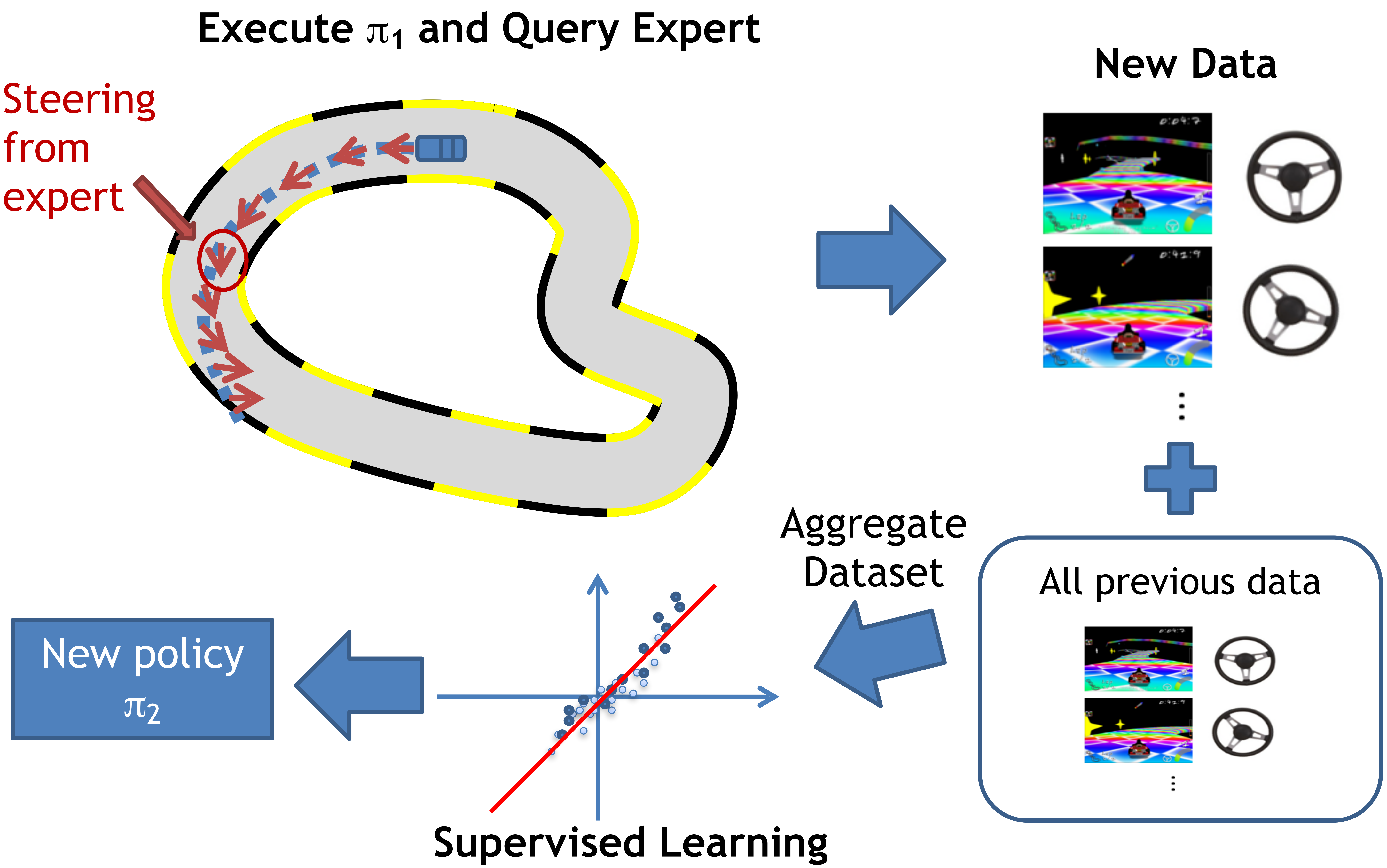
New Data



DAgger: Dataset Aggregation

[Ross11a]

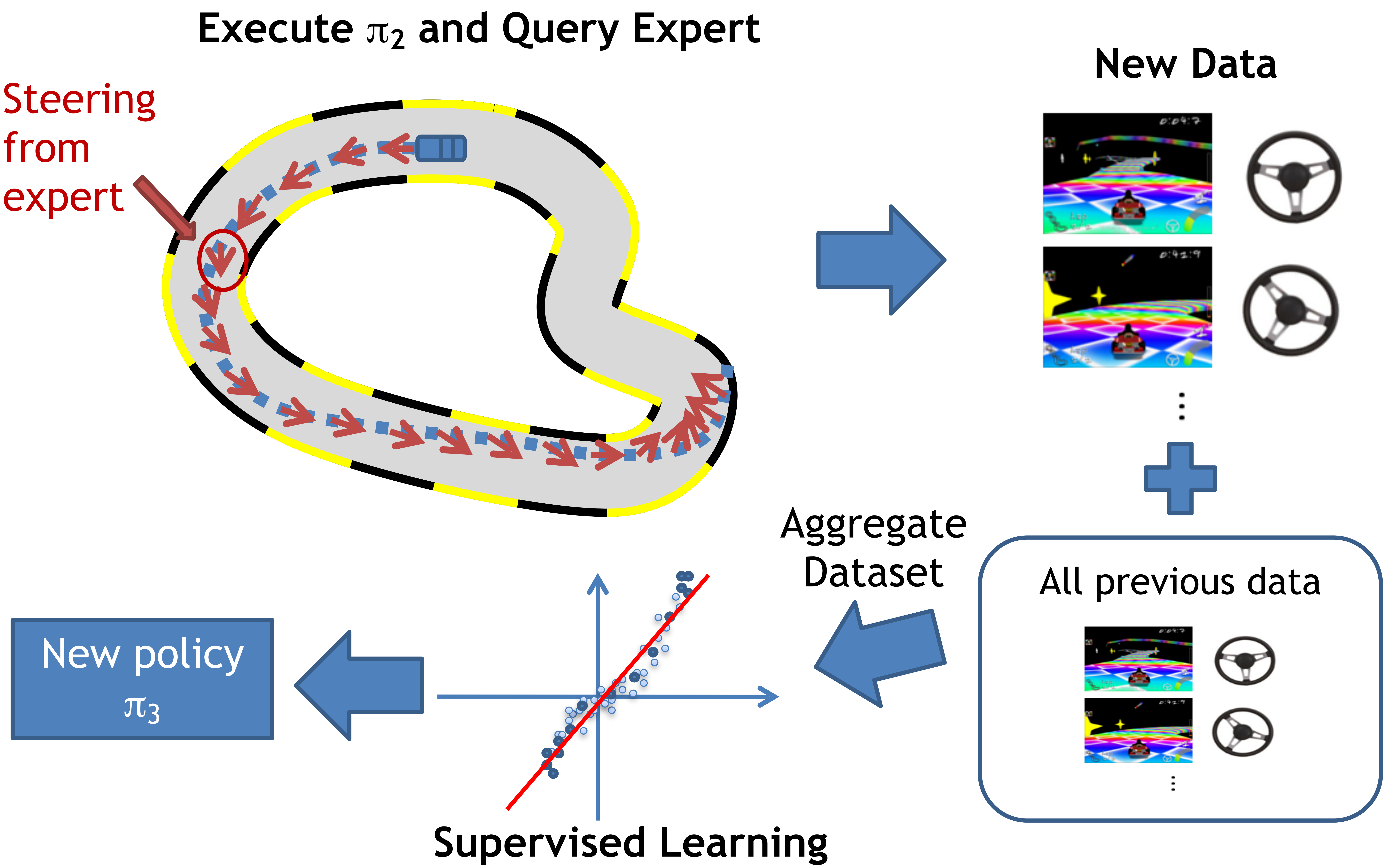
1st iteration



DAgger: Dataset Aggregation

[Ross11a]

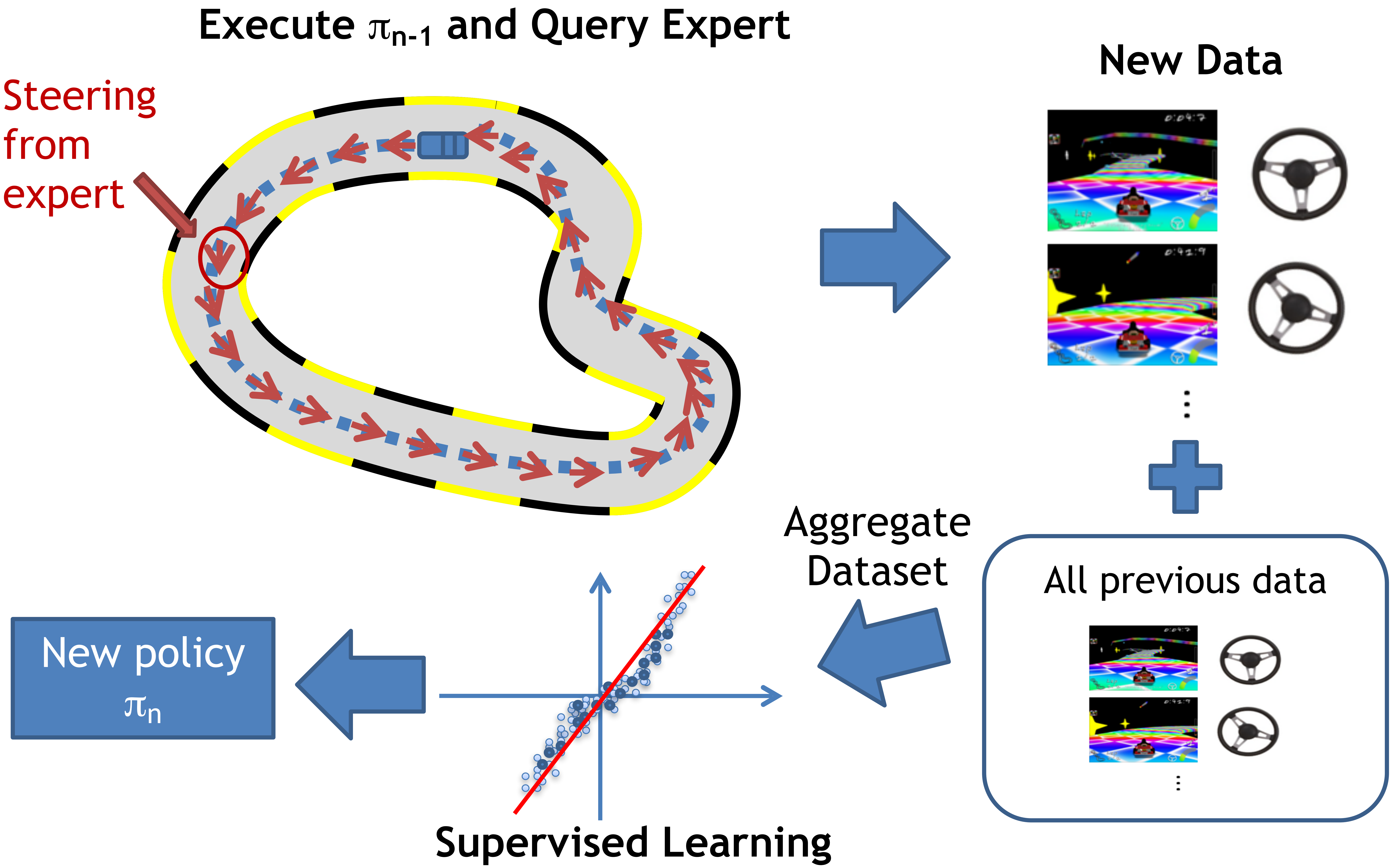
2nd iteration



DAgger: Dataset Aggregation

[Ross11a]

n^{th} iteration



Success!

[Ross AISTATS 2011]

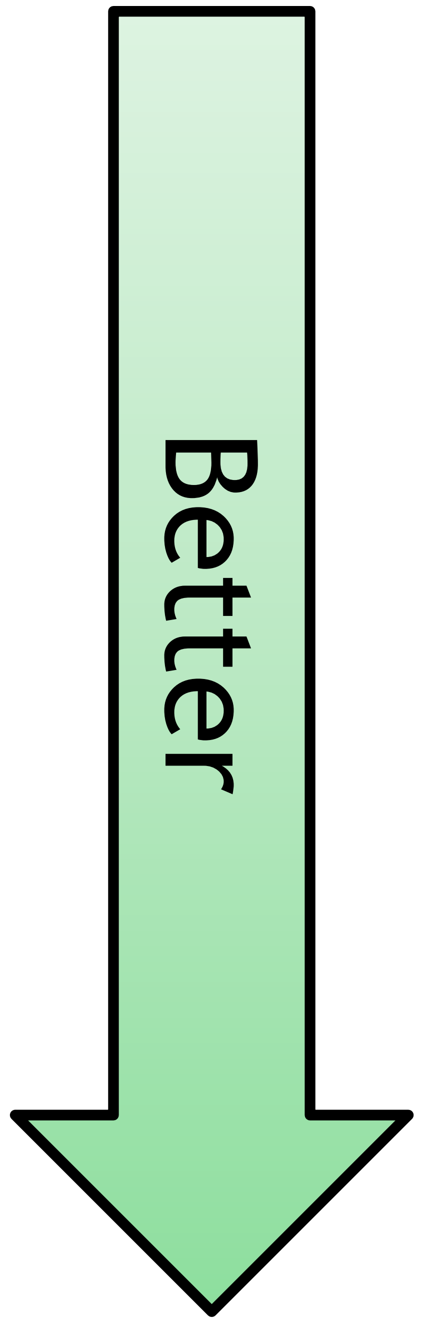
Success!

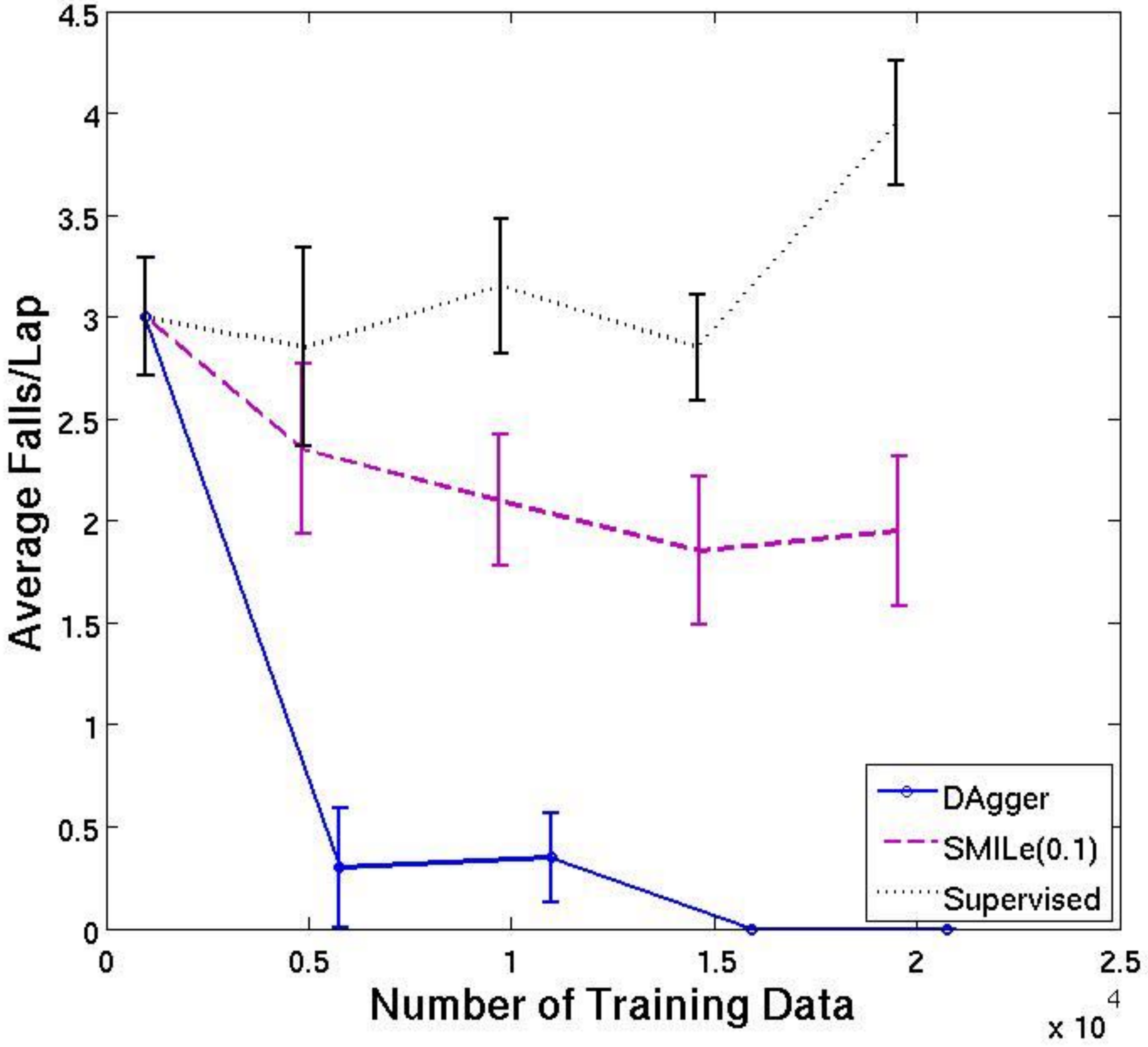
[Ross AISTATS 2011]

Success!

[Ross AISTATS 2011]

Average Falls/Lap





More fun than Video Games…

[Ross ICRA 2013]

More fun than Video Games…

[Ross ICRA 2013]

More fun than Video Games…

[Ross ICRA 2013]

Forms of the Interactive Experts

Interactive Expert is expensive, especially when the expert is human…

But expert does not have to be human…

Forms of the Interactive Experts

Interactive Expert is expensive, especially when the expert is human…

But expert does not have to be human…

Example: high-speed off-road driving
[Pan et al, RSS 18, Best System Paper]

Fig. 4: The AutoRally car and the test track.

Forms of the Interactive Experts

Interactive Expert is expensive, especially when the expert is human…

But expert does not have to be human…

Example: high-speed off-road driving
[Pan et al, RSS 18, Best System Paper]

Fig. 4: The AutoRally car and the test track.

Goal: learn a racing control policy that maps from data on cheap on-board sensors (raw-pixel imagine) to low-level control (steer and throttle)

Forms of the Interactive Experts

Interactive Expert is expensive, especially when the expert is human…

But expert does not have to be human…

Example: high-speed off-road driving
[Pan et al, RSS 18, Best System Paper]

Fig. 4: The AutoRally car and the test track.

Goal: learn a racing control policy that maps from data on cheap on-board sensors (raw-pixel imagine) to low-level control (steer and throttle)

→ Steering + throttle

(a) raw image

Forms of the Interactive Experts

Example: high-speed off-road driving
[Pan et al, RSS 18, Best System Paper]

Forms of the Interactive Experts

Example: high-speed off-road driving

[Pan et al, RSS 18, Best System Paper]

Their Setup:

At Training, we have expensive sensors for accurate state estimation

and we have computation resources for **MPC** (i.e., high-frequency replanning)

Forms of the Interactive Experts

Example: high-speed off-road driving
[Pan et al, RSS 18, Best System Paper]

Their Setup:
At Training, we have expensive sensors for accurate state estimation
and we have computation resources for **MPC** (i.e., high-frequency replanning)

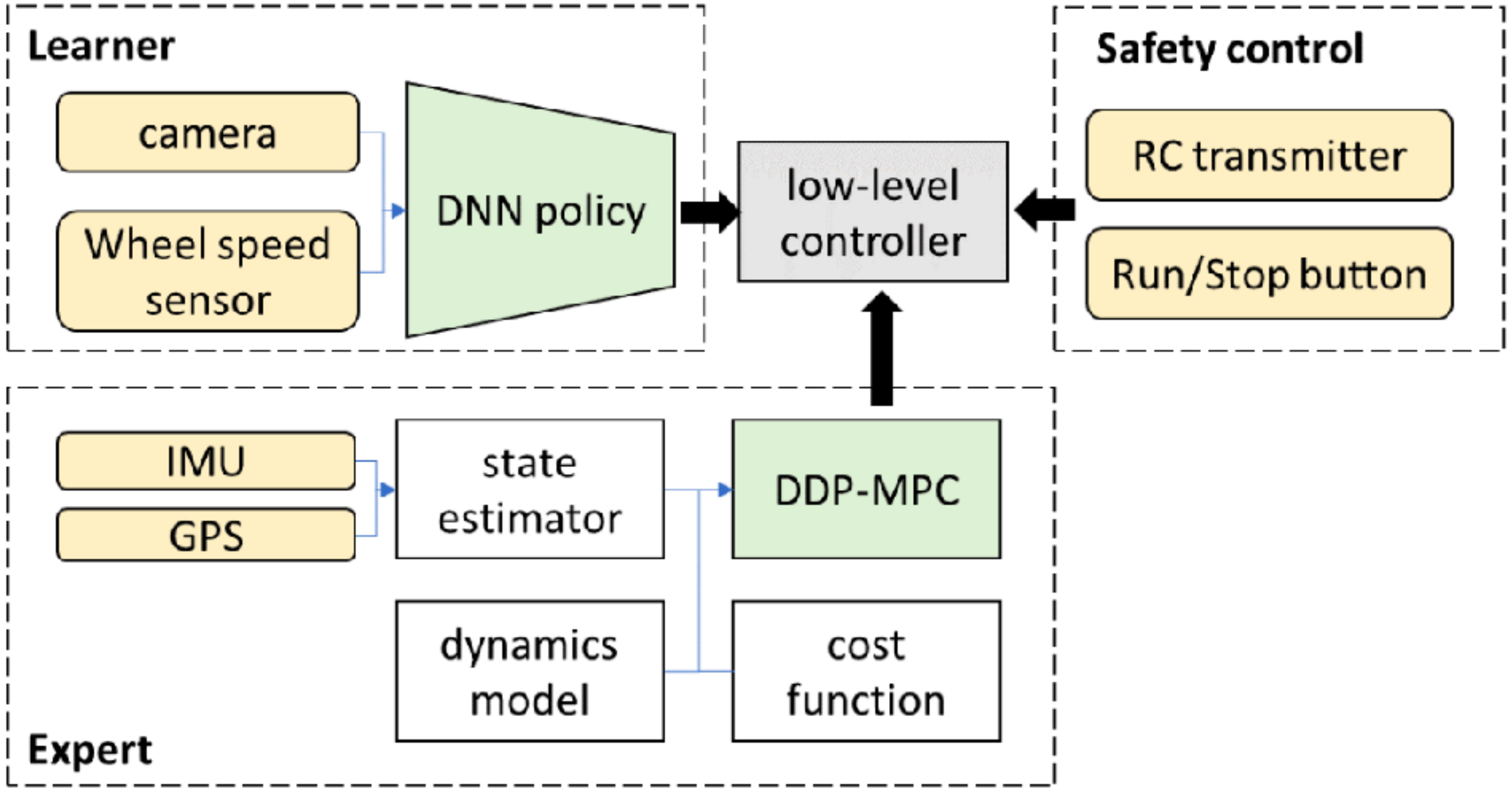
The MPC is the expert in this case!

Forms of the Interactive Experts

Example: high-speed off-road driving
[Pan et al, RSS 18, Best System Paper]

Their Setup:
At Training, we have expensive sensors for accurate state estimation and we have computation resources for **MPC** (i.e., high-frequency replanning)

The MPC is the expert in this case!



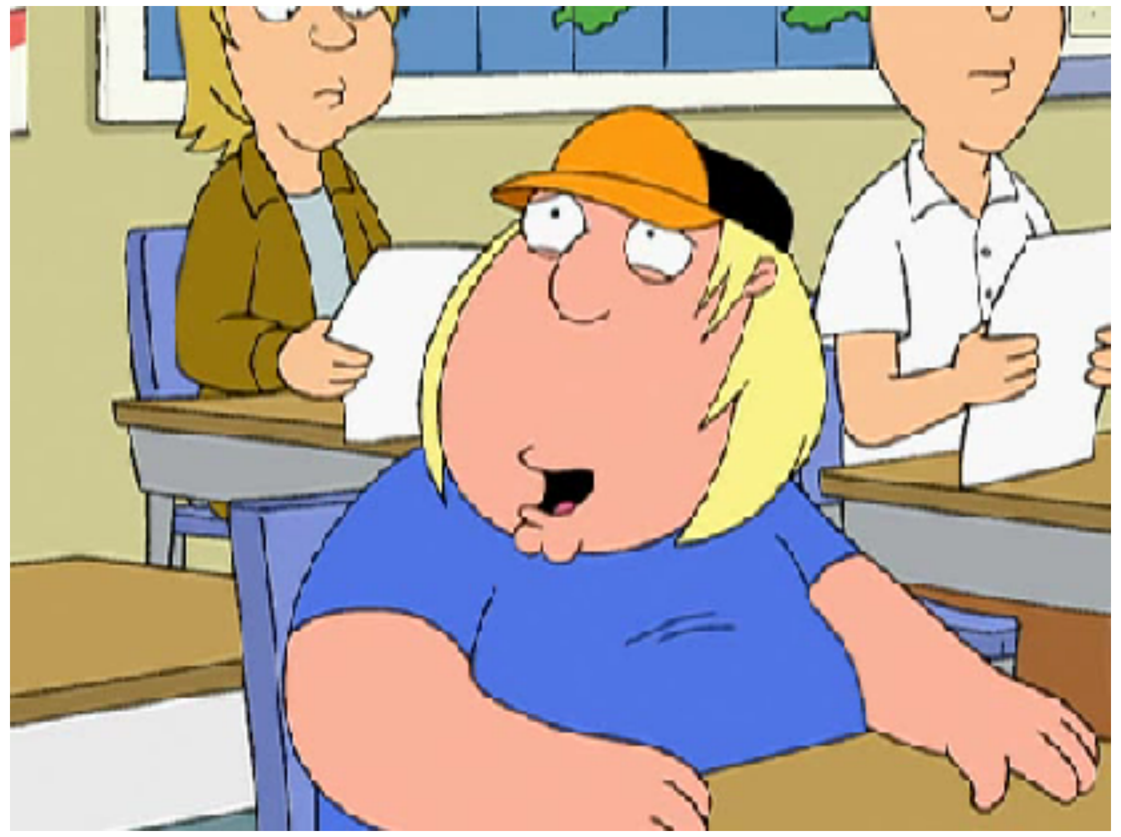
Analysis of DAgger

First let's do a quick introduction of online no-regret learning

[Vovk92,Warmuth94,Freund97,Zinkevich03,Kalai05,Hazan06,Kakade08]

Online Learning

Learner



convex Decision set $\mathcal{X}$

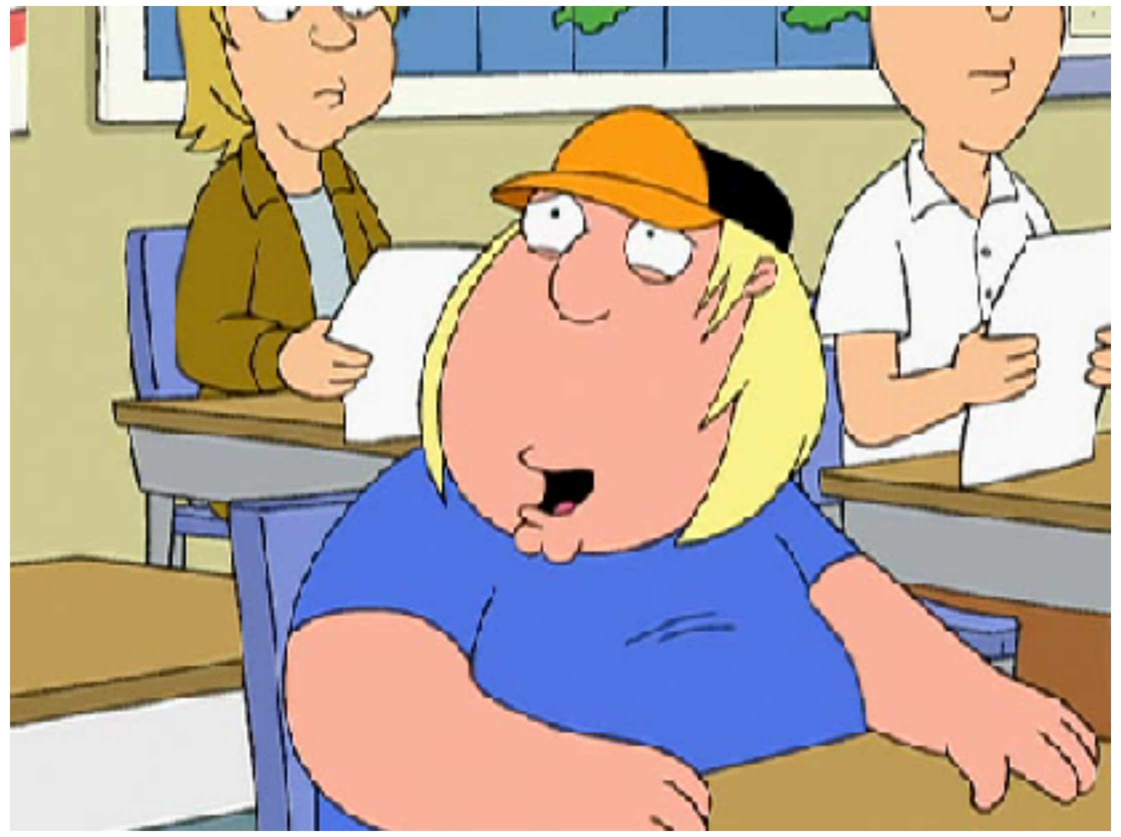
Adversary

...

[Vovk92,Warmuth94,Freund97,Zinkevich03,Kalai05,Hazan06,Kakade08]

Online Learning

Learner



convex Decision set $\mathscr{X}$

Learner picks a decision x_0 →

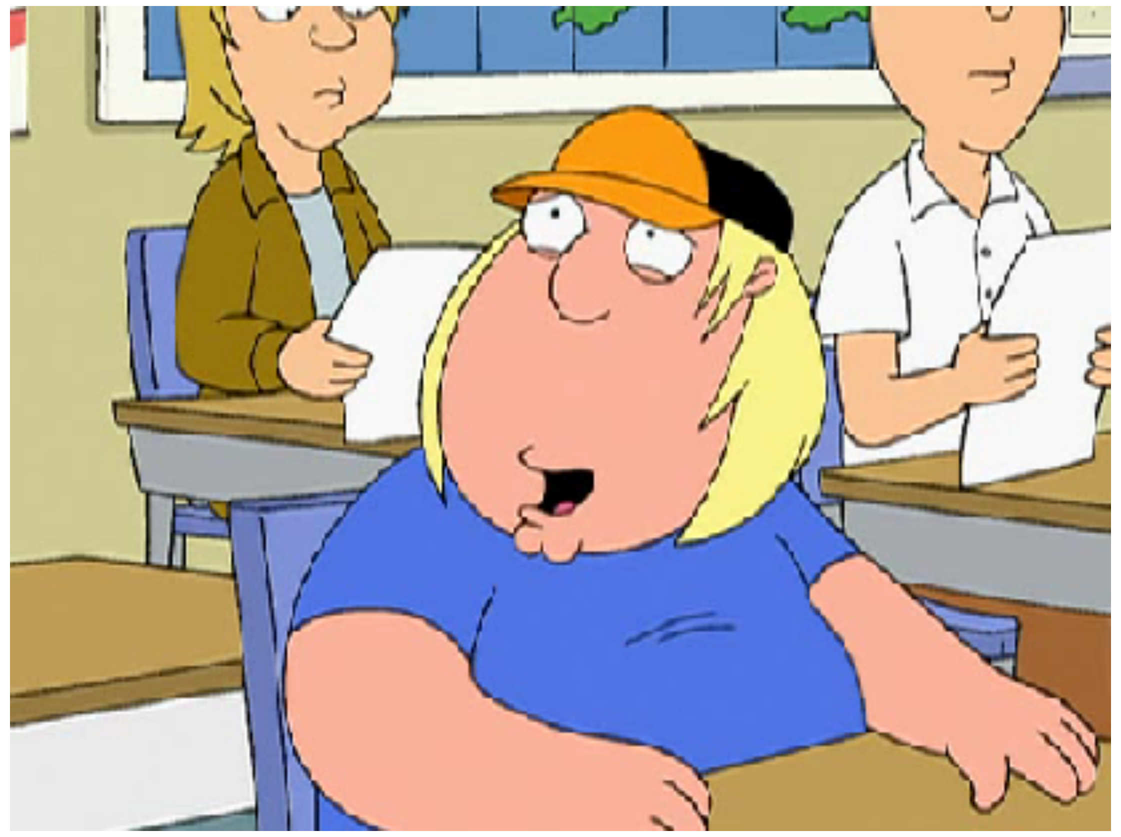
Adversary

...

[Vovk92,Warmuth94,Freund97,Zinkevich03,Kalai05,Hazan06,Kakade08]

Online Learning

Learner



convex Decision set $\mathscr{X}$

Learner picks a decision x_0

Adversary picks a loss $\ell_0 : \mathscr{X} \to \mathbb{R}$

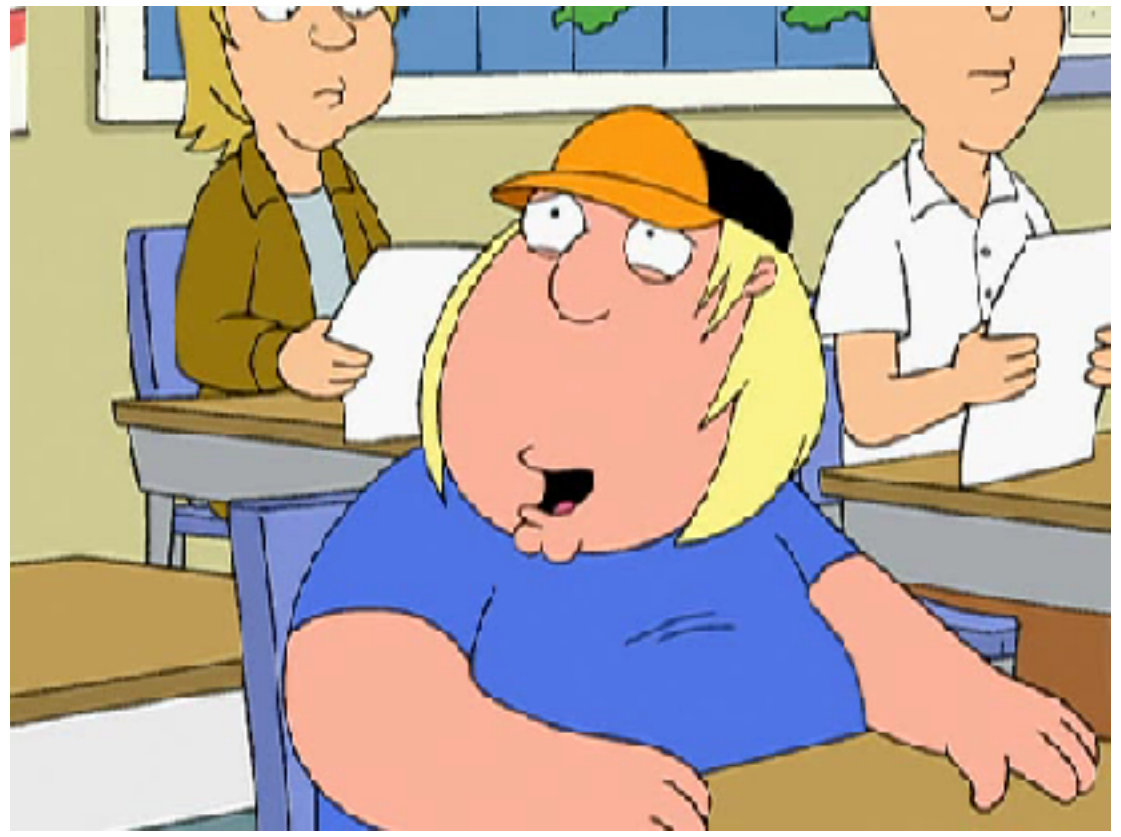
...

Adversary

[Vovk92,Warmuth94,Freund97,Zinkevich03,Kalai05,Hazan06,Kakade08]

Online Learning

Learner



convex Decision set $\mathscr{X}$

Learner picks a decision x_0

Adversary picks a loss $\ell_0 : \mathscr{X} \to \mathbb{R}$

Learner picks a new decision x_1

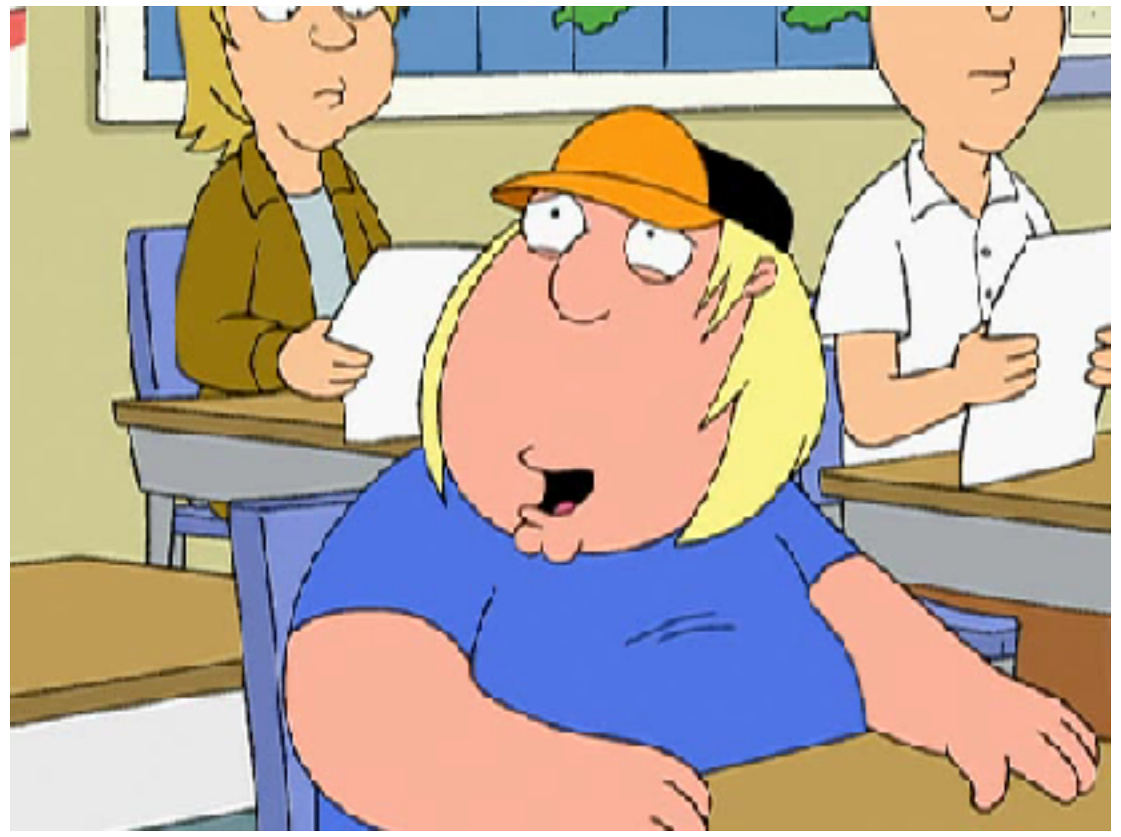
…

Adversary

[Vovk92,Warmuth94,Freund97,Zinkevich03,Kalai05,Hazan06,Kakade08]

Online Learning

Learner



convex Decision set $\mathcal{X}$

Learner picks a decision x_0

Adversary picks a loss $\ell_0 : \mathcal{X} \to \mathbb{R}$

Learner picks a new decision x_1

Adversary picks a loss $\ell_1 : \mathcal{X} \to \mathbb{R}$

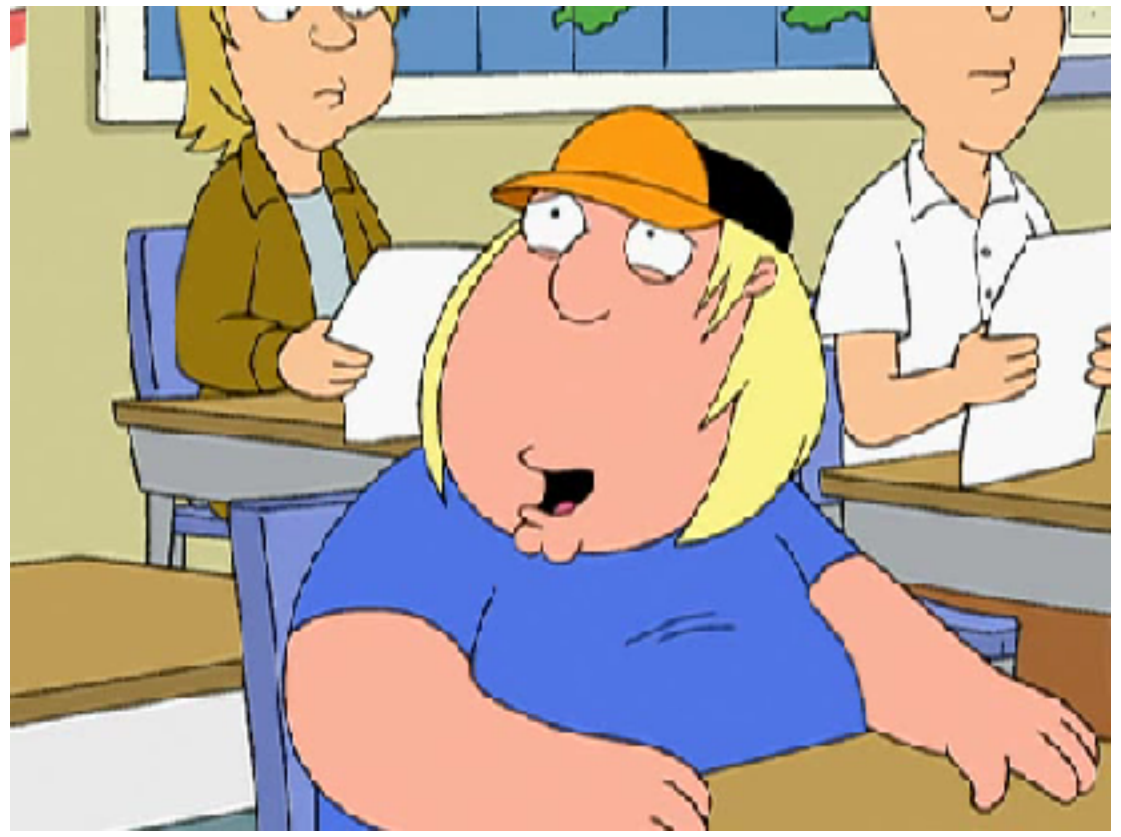
...

Adversary

[Vovk92,Warmuth94,Freund97,Zinkevich03,Kalai05,Hazan06,Kakade08]

Online Learning

Learner



convex Decision set $\mathscr{X}$

Learner picks a decision x_0

Adversary picks a loss $\ell_0 : \mathscr{X} \to \mathbb{R}$

Learner picks a new decision x_1

Adversary picks a loss $\ell_1 : \mathscr{X} \to \mathbb{R}$

...

Adversary

$$\text{Regret} = \sum_{t=0}^{T-1} \ell_t(x_t) - \min_{x \in \mathscr{X}} \sum_{t=0}^{T-1} \ell_t(x)$$

A no-regret algorithm: Follow-the-Leader

At time step t, learner has seen $\ell_0, \ldots \ell_{t-1}$, which new decision she could pick?

$$\textbf{FTL: } x_t = \min_{x \in \mathcal{X}} \sum_{i=0}^{t-1} \ell_i(x)$$

A no-regret algorithm: Follow-the-Leader

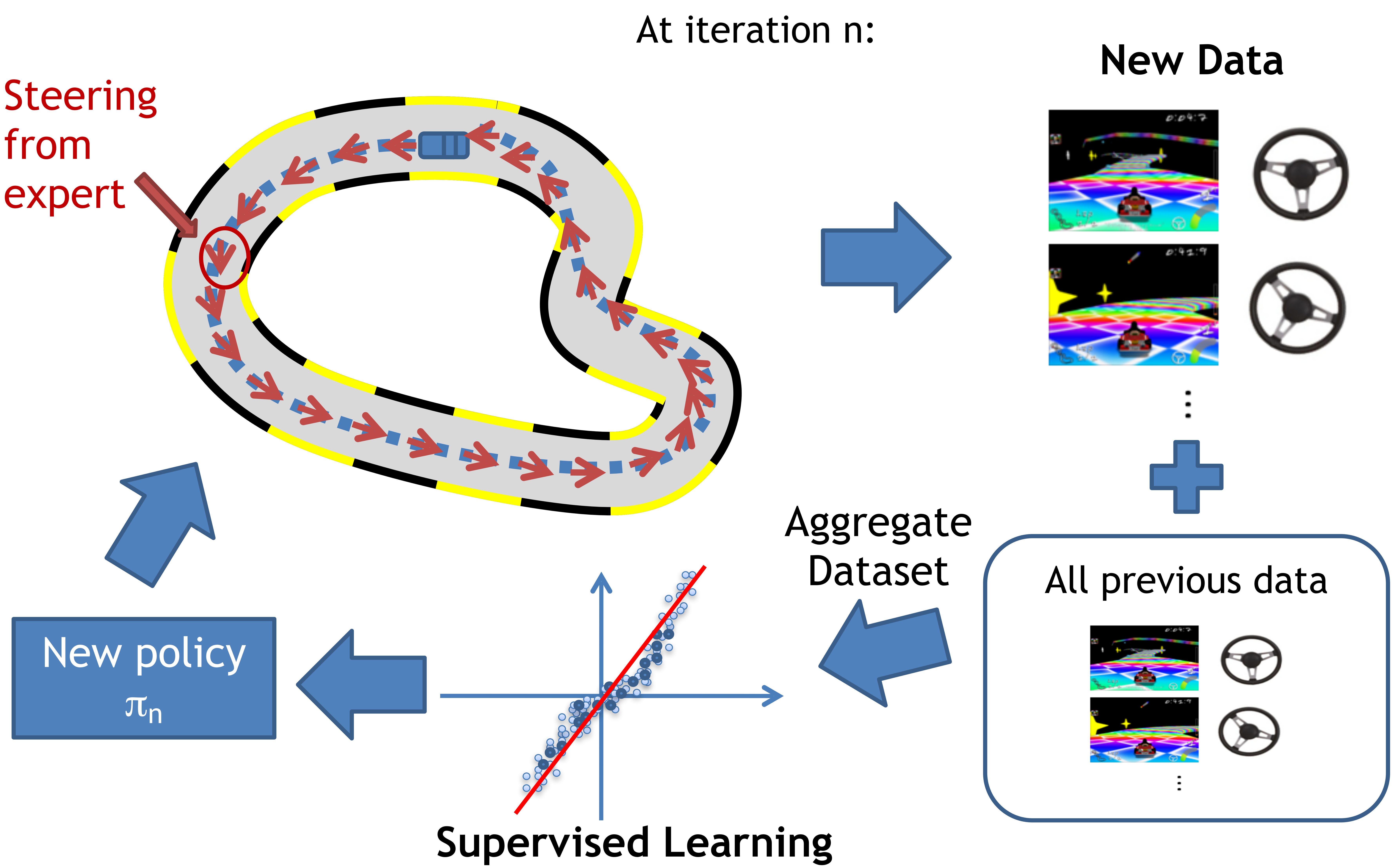
At time step t, learner has seen $\ell_0, \ldots \ell_{t-1}$, which new decision she could pick?

$$\textbf{FTL: } x_t = \min_{x \in \mathcal{X}} \sum_{i=0}^{t-1} \ell_i(x)$$

Theorem (FTL): if $\mathcal{X}$ is convex, and ℓ_t is strongly convex for all t, then for regret of FTL, we have:

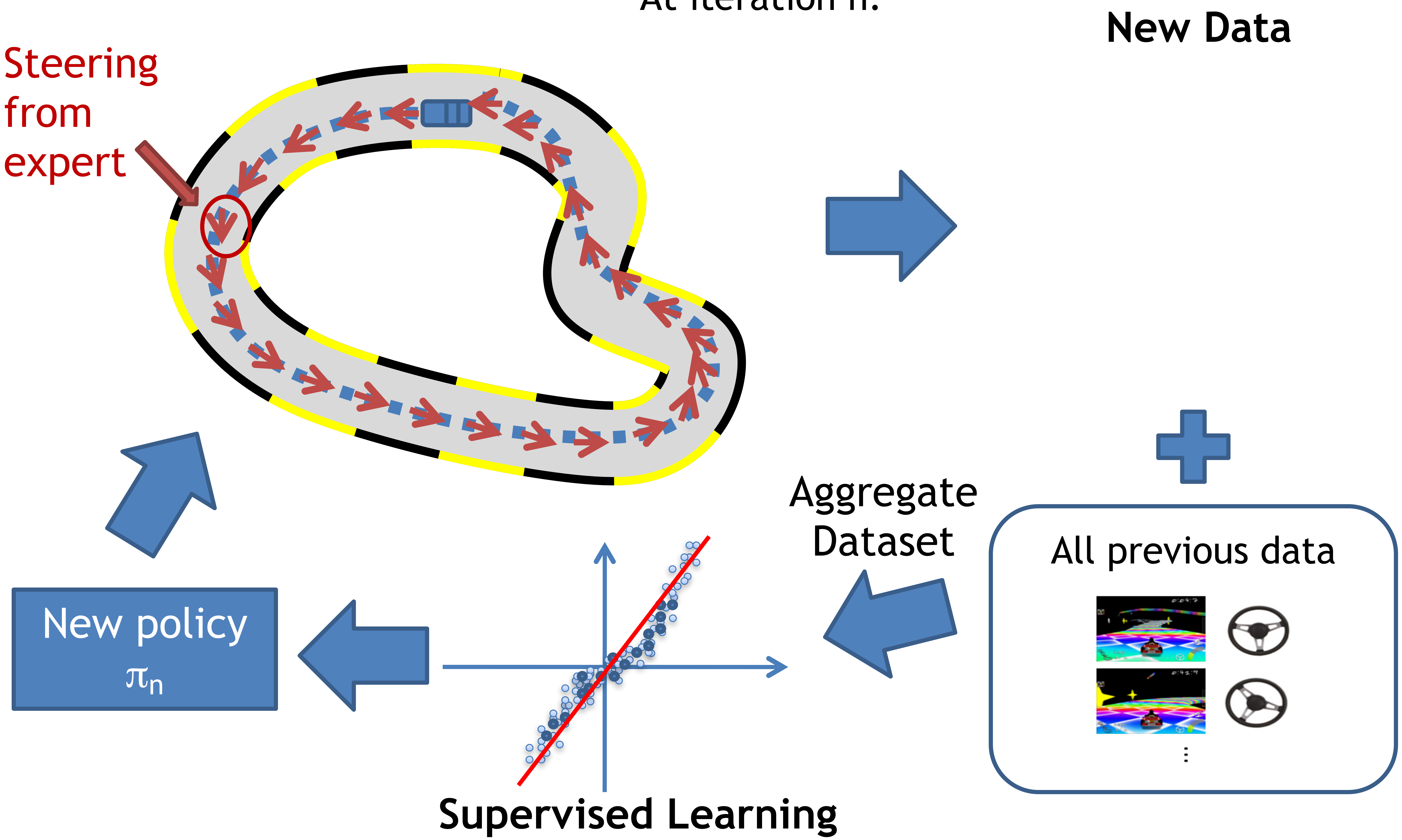
$$\frac{1}{T}\left[\sum_{t=0}^{T-1} \ell_t(x_t) - \min_{x \in \mathcal{X}} \sum_{t=0}^{T-1} \ell_t(x)\right] = O\left(\frac{\log(T)}{T}\right)$$

DAgger Revisit



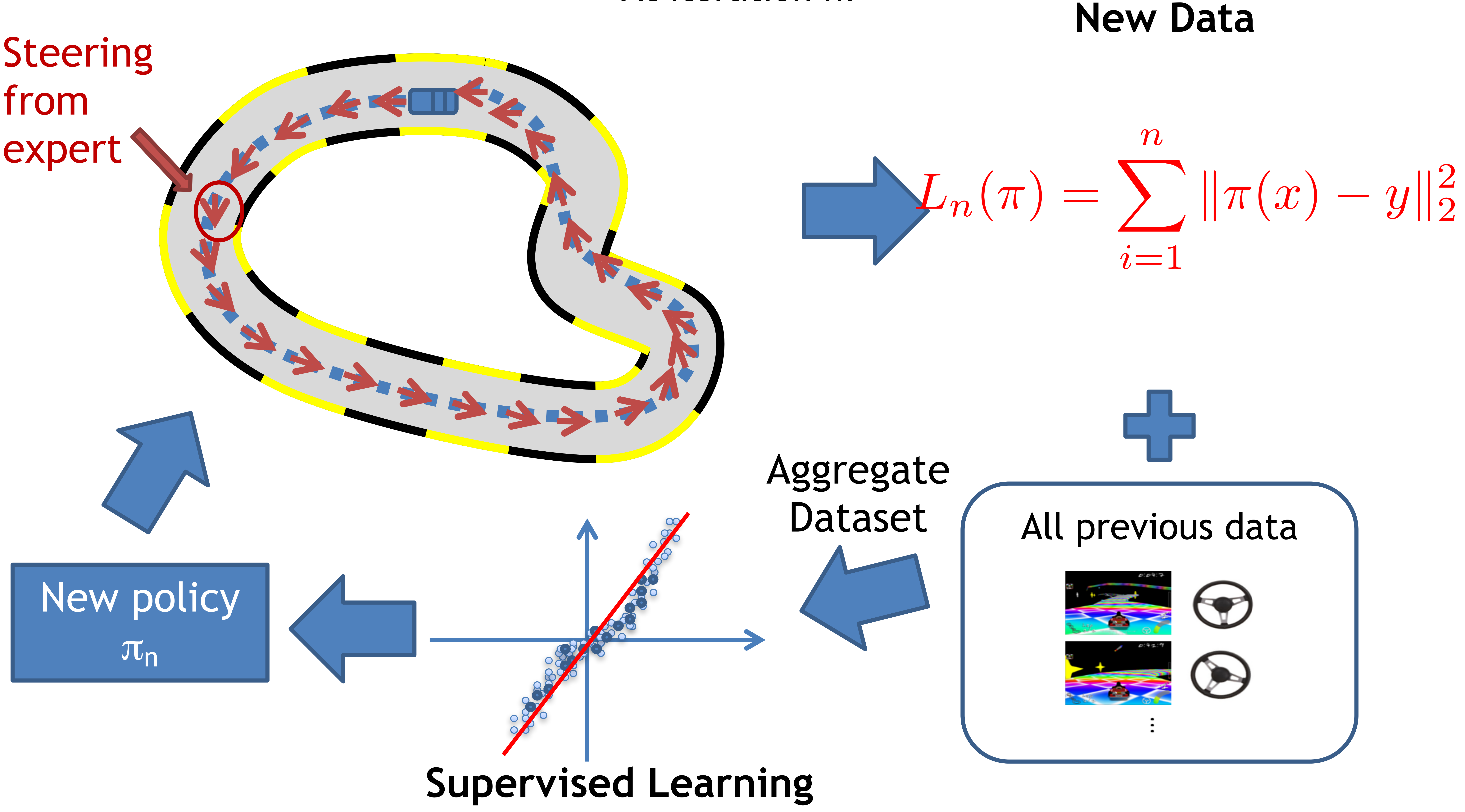
DAgger Revisit

At iteration n:



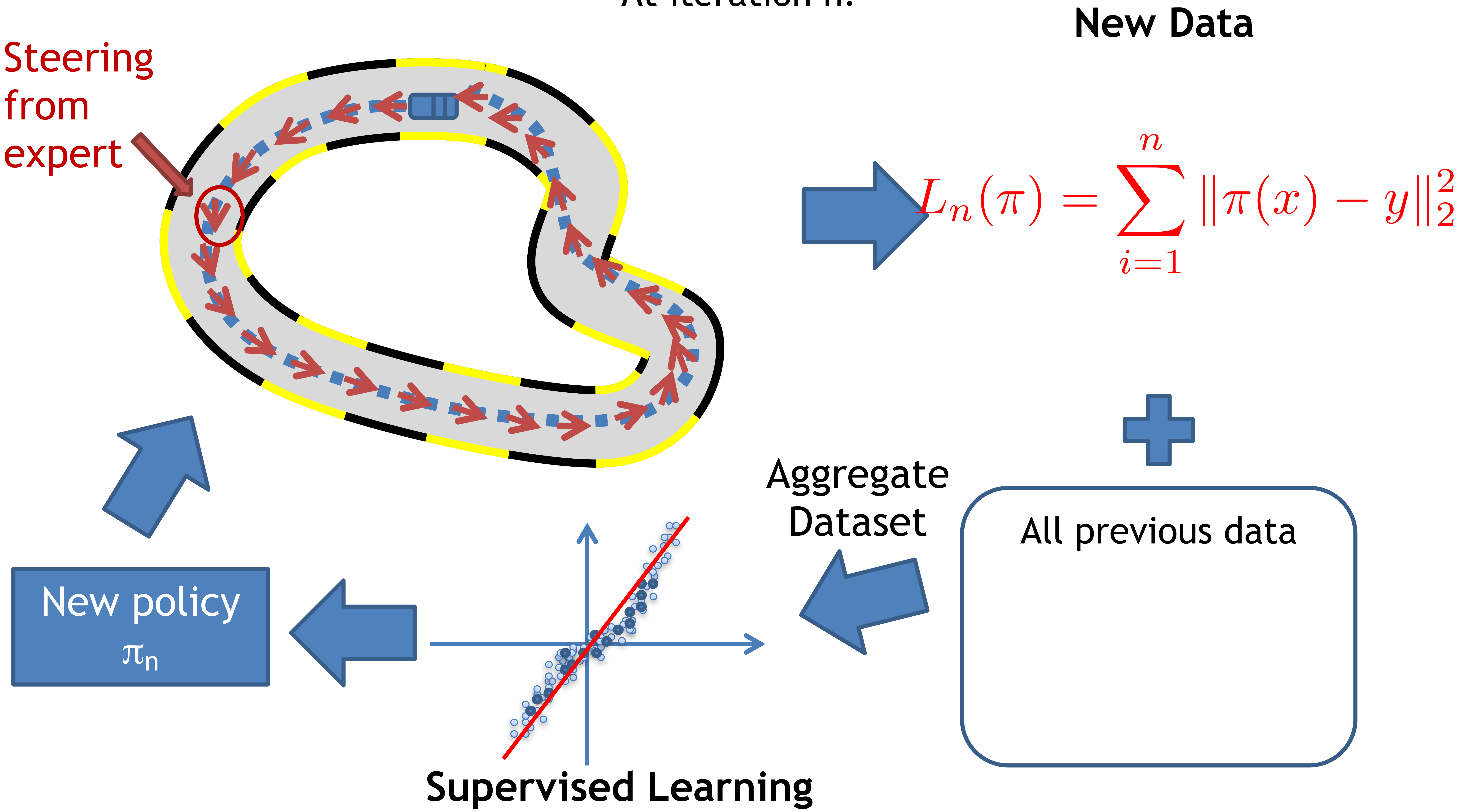
DAgger Revisit

At iteration n:



DAgger Revisit

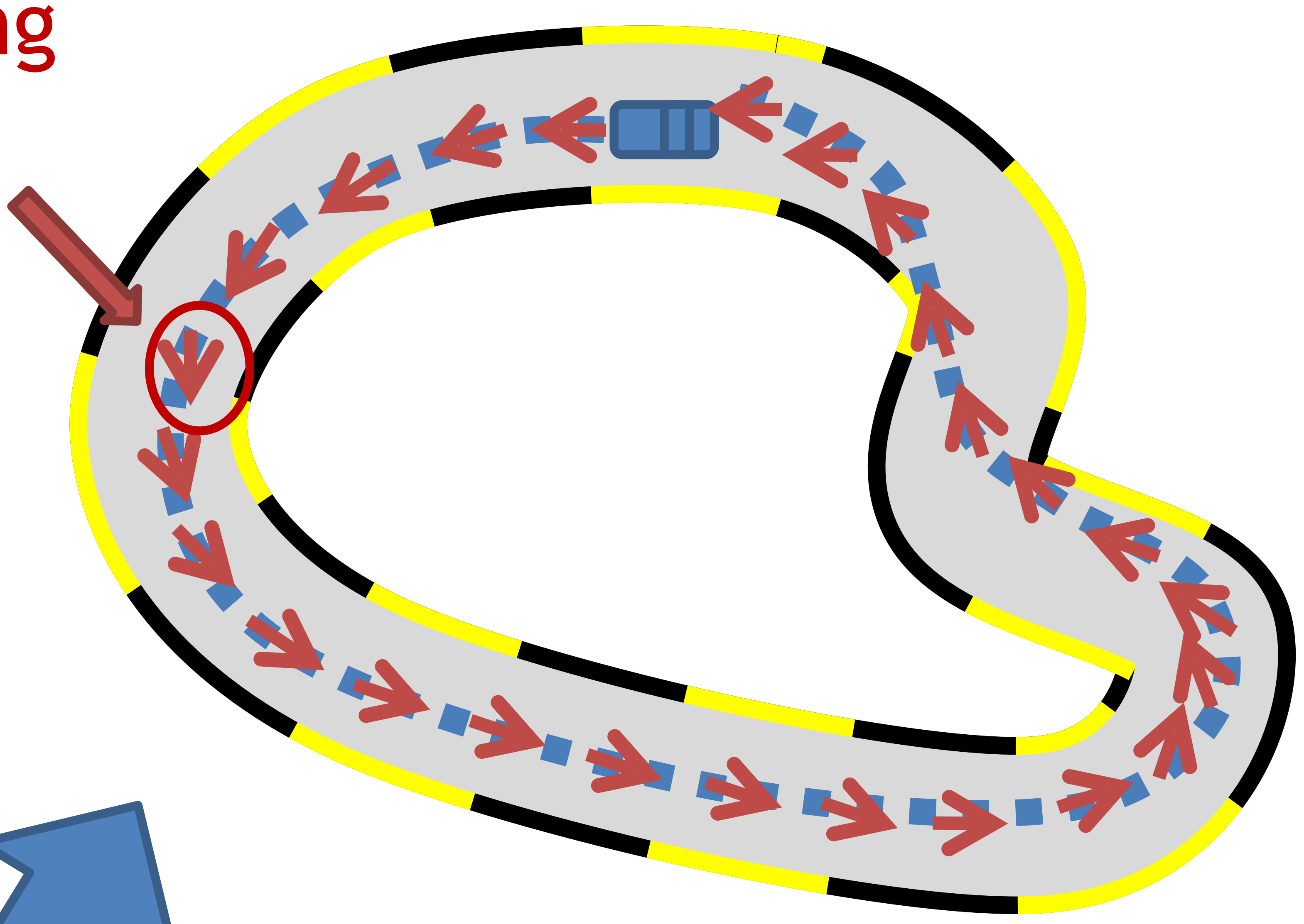
At iteration n:



DAgger Revisit

At iteration n:

Steering from expert



New Data

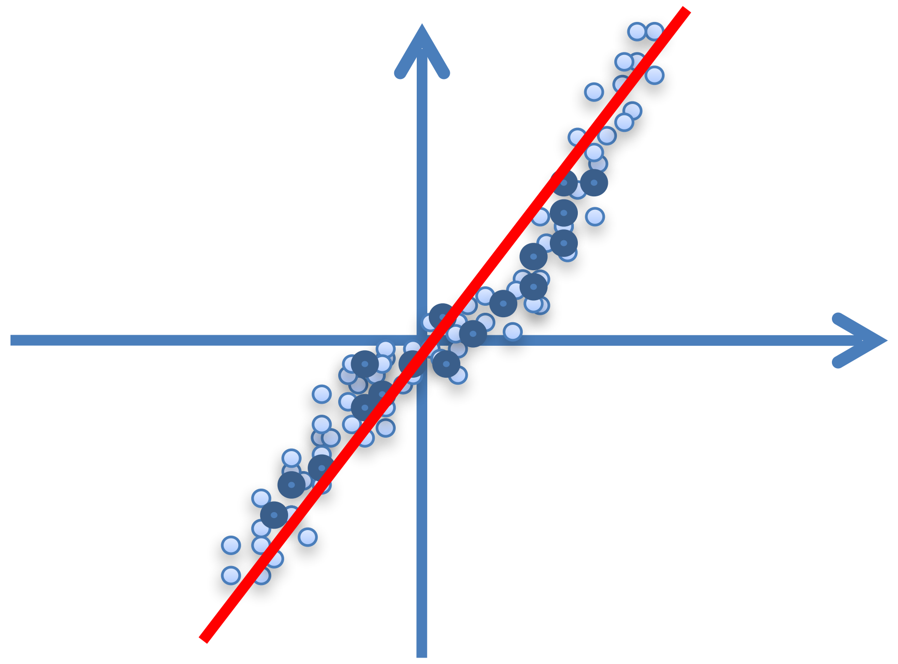
$$L_n(\pi) = \sum_{i=1}^{n} \|\pi(x) - y\|_2^2$$

+

All previous data

$$\sum_{t=1}^{n} L_t(\pi)$$

Aggregate Dataset



Supervised Learning

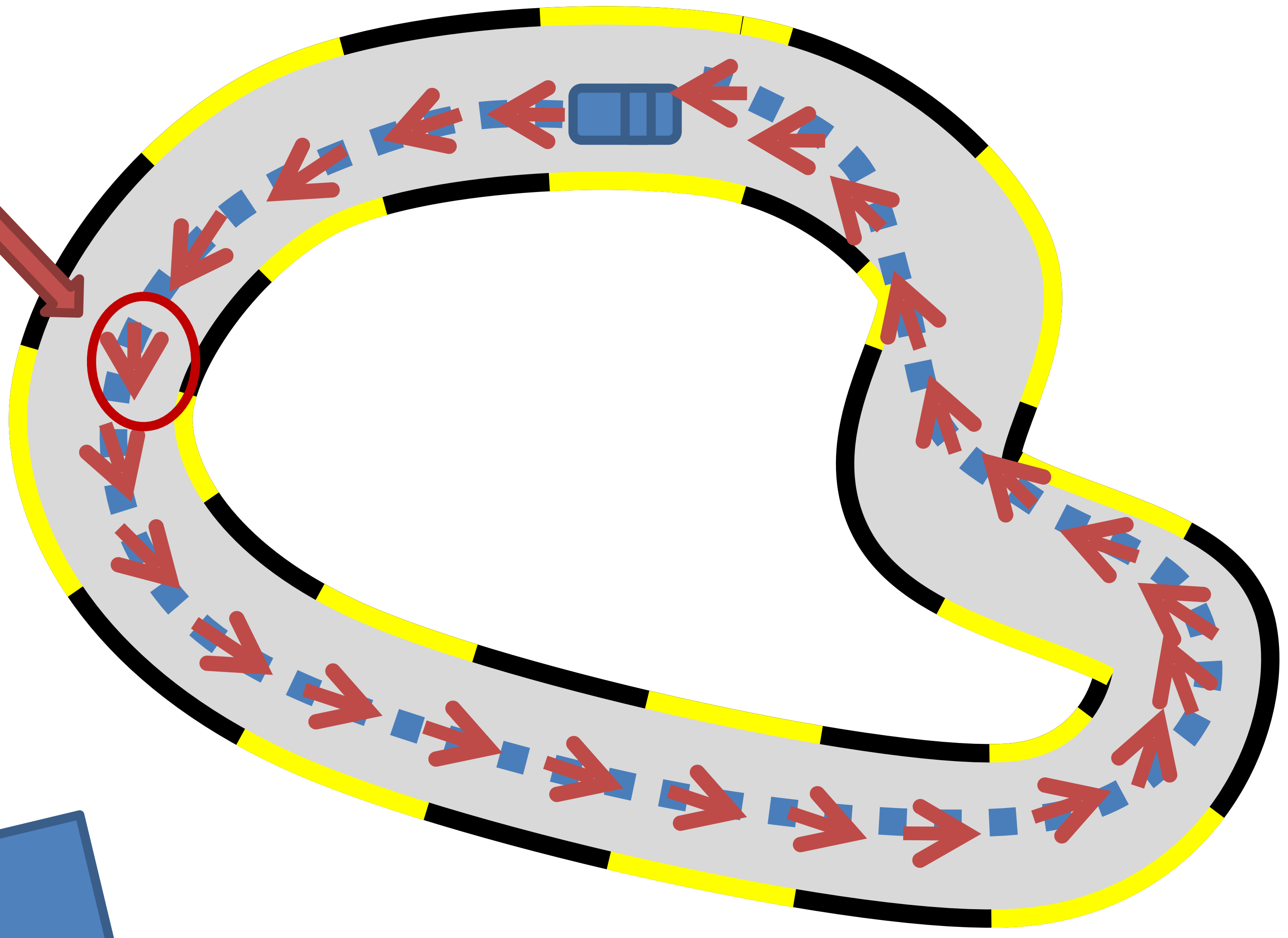
New policy

π_n

DAgger Revisit

At iteration n:

Steering from expert



New Data

$$L_n(\pi) = \sum_{i=1}^{n} \|\pi(x) - y\|_2^2$$

All previous data

$$\sum_{t=1}^{n} L_t(\pi)$$

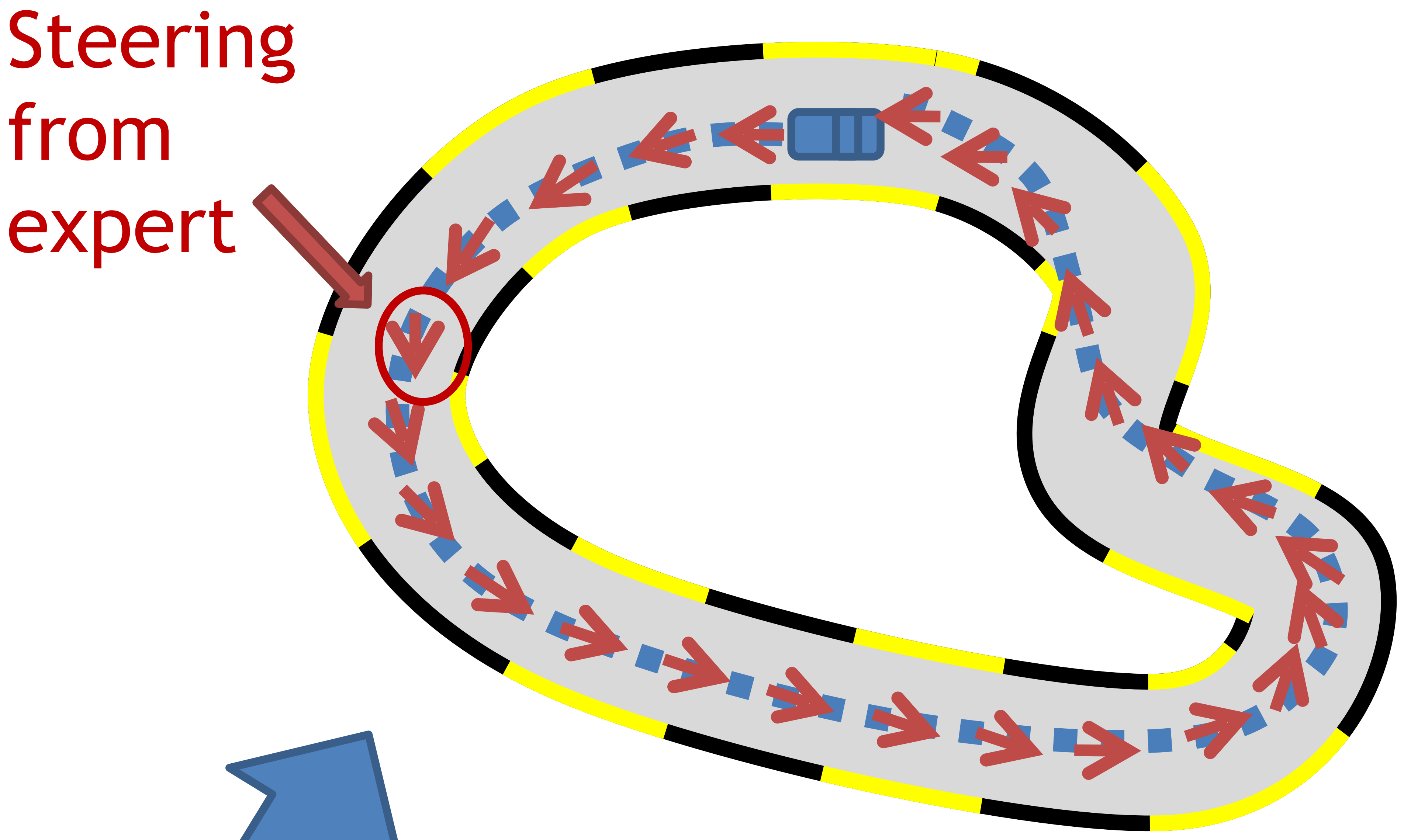
Aggregate Dataset

New policy π_n

Supervised Learning

DAgger Revisit

At iteration n:



New Data

$$L_n(\pi) = \sum_{i=1}^{n} \|\pi(x) - y\|_2^2$$

+

All previous data

$$\sum_{t=1}^{n} L_t(\pi)$$

Aggregate Dataset

$$\arg\min_{\pi} \sum_{t=1}^{n} L_t(\pi)$$

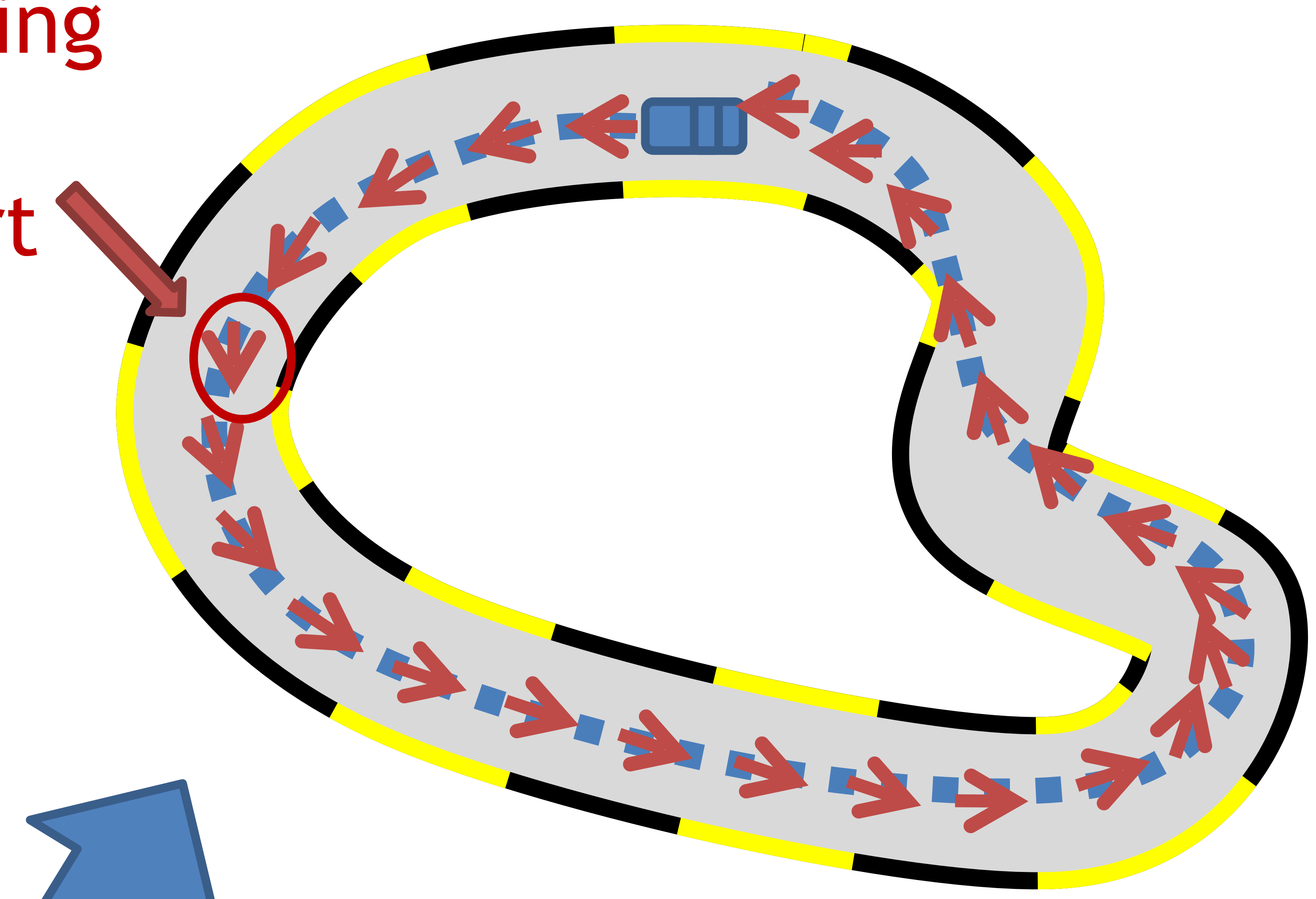
Supervised Learning

New policy π_n

DAgger Revisit

At iteration n:

Steering from expert



New Data

$$L_n(\pi) = \sum_{i=1}^{n} \|\pi(x) - y\|_2^2$$

$+$

All previous data

$$\sum_{t=1}^{n} L_t(\pi)$$

Aggregate Dataset

$$\arg\min_{\pi} \sum_{t=1}^{n} L_t(\pi)$$

Supervised Learning

New policy π_n

Data Aggregation = Follow-the-Leader Online Learner

DAgger Analysis: A reduction to no-regret online learning

Finite horizon episodic MDP, assume discrete action space

Decision set $\Pi := \{\pi : S \mapsto A\}$ (restricted policy class, $\pi^\star$ may not be inside Π)

DAgger Analysis: A reduction to no-regret online learning

Finite horizon episodic MDP, assume discrete action space

Decision set $\Pi := \{\pi : S \mapsto A\}$ (restricted policy class, $\pi^\star$ may not be inside Π)

Online Learning loss at iteration t: $\ell_t(\pi) = \mathbb{E}_{s \sim d^{\pi_t}} \left[\ell(\pi(s), \pi^\star(s)) \right]$

(Here ℓ could be any convex surrogate loss for classification, .e.g, hinge loss)

DAgger Analysis: A reduction to no-regret online learning

Finite horizon episodic MDP, assume discrete action space

Decision set $\Pi := \{\pi : S \mapsto A\}$ (restricted policy class, $\pi^\star$ may not be inside Π)

Online Learning loss at iteration t: $\ell_t(\pi) = \mathbb{E}_{s \sim d^{\pi_t}} \left[\ell(\pi(s), \pi^\star(s)) \right]$

(Here ℓ could be any convex surrogate loss for classification, .e.g, hinge loss)

DAgger is **equivalent to** FTL, i.e., $\pi_{t+1} = \arg\min_{\pi \in \Pi} \sum_{i=0}^{t} \ell_i(\pi)$

DAgger Analysis: A reduction to no-regret online learning

Finite horizon episodic MDP, assume discrete action space

Decision set $\Pi := \{\pi : S \mapsto A\}$ (restricted policy class, $\pi^\star$ may not be inside Π)

Online Learning loss at iteration t: $\ell_t(\pi) = \mathbb{E}_{s\sim d^{\pi_t}}\left[\ell(\pi(s), \pi^\star(s))\right]$

(Here ℓ could be any convex surrogate loss for classification, .e.g, hinge loss)

DAgger **is equivalent to** FTL, i.e., $\pi_{t+1} = \arg\min_{\pi\in\Pi} \sum_{i=0}^{t} \ell_i(\pi)$

If the online learning procedure ensures no-regret, then

$$\frac{1}{T}\left[\sum_{t=0}^{T-1} \ell_t(\pi_t) - \min_{\pi\in\Pi} \sum_{t=0}^{T-1} \ell_t(\pi)\right] = o(T)/T$$

DAgger Analysis: A reduction to no-regret online learning

Online Learning loss at iteration t: $\ell_t(\pi) = \mathbb{E}_{s \sim d^{\pi_t}} \left[\ell(\pi(s), \pi^\star(s)) \right]$

$$\sum_{t=0}^{T-1} \ell_t(\pi_t) - \min_{\pi \in \Pi} \sum_{t=0}^{T-1} \ell_t(\pi) = o(T)$$

DAgger Analysis: A reduction to no-regret online learning

Online Learning loss at iteration t: $\ell_t(\pi) = \mathbb{E}_{s \sim d^{\pi_t}} \left[\ell(\pi(s), \pi^\star(s)) \right]$

$$\sum_{t=0}^{T-1} \ell_t(\pi_t) - \min_{\pi \in \Pi} \sum_{t=0}^{T-1} \ell_t(\pi) = o(T)$$

$$\frac{1}{T} \sum_{t=0}^{T-1} \ell_t(\pi_t) = \underbrace{o(T)/T}_{\epsilon_{avg-reg}} + \underbrace{\min_{\pi \in \Pi} \frac{1}{T} \sum_{t=0}^{T-1} \ell_t(\pi)}_{\epsilon_\Pi}$$

DAgger Analysis: A reduction to no-regret online learning

Online Learning loss at iteration t: $\ell_t(\pi) = \mathbb{E}_{s\sim d^{\pi_t}}\left[\ell(\pi(s), \pi^\star(s))\right]$

$$\sum_{t=0}^{T-1} \ell_t(\pi_t) - \min_{\pi\in\Pi} \sum_{t=0}^{T-1} \ell_t(\pi) = o(T)$$

$$\frac{1}{T}\sum_{t=0}^{T-1} \ell_t(\pi_t) = \underbrace{o(T)/T}_{\epsilon_{avg-reg}} + \underbrace{\min_{\pi\in\Pi} \frac{1}{T}\sum_{t=0}^{T-1} \ell_t(\pi)}_{\epsilon_\Pi}$$

$\exists \hat{t} \in [0,\ldots,T-1]$, such that: $\ell_{\hat{t}}(\pi_{\hat{t}}) \leq \epsilon_{avg-reg} + \epsilon_\Pi$

DAgger Analysis: A reduction to no-regret online learning

Online Learning loss at iteration t: $\ell_t(\pi) = \mathbb{E}_{s \sim d^{\pi_t}} \left[\ell(\pi(s), \pi^\star(s)) \right]$

$$\sum_{t=0}^{T-1} \ell_t(\pi_t) - \min_{\pi \in \Pi} \sum_{t=0}^{T-1} \ell_t(\pi) = o(T)$$

$$\frac{1}{T} \sum_{t=0}^{T-1} \ell_t(\pi_t) = \underbrace{o(T)/T}_{\epsilon_{avg-reg}} + \underbrace{\min_{\pi \in \Pi} \frac{1}{T} \sum_{t=0}^{T-1} \ell_t(\pi)}_{\epsilon_\Pi}$$

$\exists \hat{t} \in [0, \ldots, T-1]$, such that: $\ell_{\hat{t}}(\pi_{\hat{t}}) \leq \epsilon_{avg-reg} + \epsilon_\Pi$

Under the assumption that surrogate loss upper bounds zero-one loss:

$$\mathbb{E}_{s \sim d^{\pi_{\hat{t}}}} \left[\pi_{\hat{t}}(s) \neq \pi^\star(s) \right] \leq \mathbb{E}_{s \sim d^{\pi_{\hat{t}}}} \left[\ell(\pi_{\hat{t}}(s), \pi^\star(s)) \right] \leq \epsilon_{avg-reg} + \epsilon_\Pi$$

DAgger Analysis: A reduction to no-regret online learning

$$\mathbb{E}_{s \sim d^{\pi_{\hat{i}}}}\left[\pi_{\hat{i}}(s) \neq \pi^\star(s)\right] \leq \mathbb{E}_{s \sim d^{\pi_{\hat{i}}}}\left[\ell(\pi_{\hat{i}}(s), \pi^\star(s))\right] \leq \epsilon_{reg} + \epsilon_\Pi$$

$\pi_{\hat{i}}$ can predict $\pi^\star$ well under its own state distribution

DAgger Analysis: A reduction to no-regret online learning

$$\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\pi_{\hat{i}}(s) \neq \pi^\star(s)\right] \leq \mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\ell(\pi_{\hat{i}}(s), \pi^\star(s))\right] \leq \epsilon_{reg} + \epsilon_{\Pi}$$

$\pi_{\hat{i}}$ can predict $\pi^\star$ well under its own state distribution

Let's turn this to the true performance under the cost function $c(s, a)$

DAgger Analysis: A reduction to no-regret online learning

$$\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\pi_{\hat{i}}(s) \neq \pi^\star(s)\right] \leq \mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\ell(\pi_{\hat{i}}(s), \pi^\star(s))\right] \leq \epsilon_{reg} + \epsilon_{\Pi}$$

$\pi_{\hat{i}}$ can predict $\pi^\star$ well under its own state distribution

Let's turn this to the true performance under the cost function $c(s, a)$

$$V^{\pi_{\hat{i}}} - V^{\pi^\star} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[A^\star(s, \pi_{\hat{i}}(s))\right]$$

DAgger Analysis: A reduction to no-regret online learning

$$\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\pi_{\hat{i}}(s) \neq \pi^\star(s)\right] \leq \mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\ell(\pi_{\hat{i}}(s), \pi^\star(s))\right] \leq \epsilon_{reg} + \epsilon_\Pi$$

$\pi_{\hat{i}}$ can predict $\pi^\star$ well under its own state distribution

Let's turn this to the true performance under the cost function $c(s, a)$

$$V^{\pi_{\hat{i}}} - V^{\pi^\star} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[A^\star(s, \pi_{\hat{i}}(s))\right]$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[A^\star(s, \pi_{\hat{i}}(s)) - A^\star(s, \pi^\star(s))\right]$$

DAgger Analysis: A reduction to no-regret online learning

$$\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\pi_{\hat{i}}(s) \neq \pi^\star(s)\right] \leq \mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\ell(\pi_{\hat{i}}(s), \pi^\star(s))\right] \leq \epsilon_{reg} + \epsilon_\Pi$$

$\pi_{\hat{i}}$ can predict $\pi^\star$ well under its own state distribution

Let's turn this to the true performance under the cost function $c(s, a)$

$$V^{\pi_{\hat{i}}} - V^{\pi^\star} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[A^\star(s, \pi_{\hat{i}}(s))\right]$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[A^\star(s, \pi_{\hat{i}}(s)) - A^\star(s, \pi^\star(s))\right]$$

$$\leq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\mathbf{1}\{\pi_{\hat{i}}(s) \neq \pi^\star(s)\} \max_{s,a} \left|A^\star(s, a)\right|\right]$$

DAgger Analysis: A reduction to no-regret online learning

$$\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\pi_{\hat{i}}(s) \neq \pi^\star(s)\right] \leq \mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\ell(\pi_{\hat{i}}(s), \pi^\star(s))\right] \leq \epsilon_{reg} + \epsilon_\Pi$$

$\pi_{\hat{i}}$ can predict $\pi^\star$ well under its own state distribution

Let's turn this to the true performance under the cost function $c(s, a)$

$$V^{\pi_{\hat{i}}} - V^{\pi^\star} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[A^\star(s, \pi_{\hat{i}}(s))\right]$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[A^\star(s, \pi_{\hat{i}}(s)) - A^\star(s, \pi^\star(s))\right]$$

$$\leq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\mathbf{1}\{\pi_{\hat{i}}(s) \neq \pi^\star(s)\} \max_{s,a} \left|A^\star(s, a)\right|\right]$$

$$\leq \frac{\max_{s,a}\left|A^\star(s, a)\right|}{1-\gamma} \cdot (\epsilon_{reg} + \epsilon_\Pi)$$

DAgger Analysis: A reduction to no-regret online learning

$$\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\pi_{\hat{i}}(s) \neq \pi^\star(s)\right] \leq \mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\ell(\pi_{\hat{i}}(s), \pi^\star(s))\right] \leq \epsilon_{reg} + \epsilon_\Pi$$

$\pi_{\hat{i}}$ can predict $\pi^\star$ well under its own state distribution

Let's turn this to the true performance under the cost function $c(s, a)$

$$V^{\pi_{\hat{i}}} - V^{\pi^\star} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[A^\star(s, \pi_{\hat{i}}(s))\right]$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[A^\star(s, \pi_{\hat{i}}(s)) - A^\star(s, \pi^\star(s))\right]$$

$$\leq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\mathbf{1}\{\pi_{\hat{i}}(s) \neq \pi^\star(s)\} \max_{s,a}\left|A^\star(s,a)\right|\right]$$

$$\leq \frac{\max_{s,a}\left|A^\star(s,a)\right|}{1-\gamma} \cdot (\epsilon_{reg} + \epsilon_\Pi)$$

Case study:

1. Worst case: $A^\star(s, a) \approx \frac{1}{1-\gamma}$ (not recoverable from a mistake): quadratic dependence on horizon, i.e., no better than BC;

DAgger Analysis: A reduction to no-regret online learning

$$\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\pi_{\hat{i}}(s) \neq \pi^\star(s)\right] \leq \mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\ell(\pi_{\hat{i}}(s), \pi^\star(s))\right] \leq \epsilon_{reg} + \epsilon_\Pi$$

$\pi_{\hat{i}}$ can predict $\pi^\star$ well under its own state distribution

Let's turn this to the true performance under the cost function $c(s, a)$

$$V^{\pi_{\hat{i}}} - V^{\pi^\star} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[A^\star(s, \pi_{\hat{i}}(s))\right]$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[A^\star(s, \pi_{\hat{i}}(s)) - A^\star(s, \pi^\star(s))\right]$$

$$\leq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_{\hat{i}}}}\left[\mathbf{1}\{\pi_{\hat{i}}(s) \neq \pi^\star(s)\} \max_{s,a} \left|A^\star(s, a)\right|\right]$$

$$\leq \frac{\max_{s,a}\left|A^\star(s, a)\right|}{1-\gamma} \cdot (\epsilon_{reg} + \epsilon_\Pi)$$

Case study:

1. Worst case: $A^\star(s, a) \approx \frac{1}{1-\gamma}$ (not recoverable from a mistake): quadratic dependence on horizon, i.e., no better than BC;

2. **Good case:** $A^\star(s, a) \approx o\left(\frac{1}{1-\gamma}\right)$ (easily recoverable from a one-step mistake): Better than BC;

Summary of Imitation Learning

Summary of Imitation Learning

1. Behavior Cloning (Maximum Likelihood Estimation)

$$\text{Performance-gap} \approx \frac{1}{(1-\gamma)^2}\,(\text{classification error})$$

Summary of Imitation Learning

1. Behavior Cloning (Maximum Likelihood Estimation)

$$\text{Performance-gap } \approx \frac{1}{(1-\gamma)^2}(\text{classification error})$$

2. Hybrid Distribution Matching (w/ IPM or MaxEnt-IRL):

$$\text{Performance-gap } \approx \frac{1}{(1-\gamma)}(\text{classification error})$$

Summary of Imitation Learning

1. Behavior Cloning (Maximum Likelihood Estimation)

$$\text{Performance-gap} \approx \frac{1}{(1-\gamma)^2}\,(\text{classification error})$$

2. Hybrid Distribution Matching (w/ IPM or MaxEnt-IRL):

$$\text{Performance-gap} \approx \frac{1}{(1-\gamma)}\,(\text{classification error})$$

3.DAgger w/ Interactive Experts:

$$\text{Performance-gap} \approx \frac{\sup_{s,a} |A^{\star}(s,a)|}{(1-\gamma)}\,(\text{classification error})$$

Summary of the Course

Basics of MDPs (and LQRs):

Planning: VI, PI, LP formulations, Fitted Q-iteration (under B-complete), Low bounds on linear $Q^\star$

Summary of the Course

Basics of MDPs (and LQRs):

Planning: VI, PI, LP formulations, Fitted Q-iteration (under B-complete), Low bounds on linear $Q^\star$

Exploration in MDPs (bandit / tabular / linear mdp / Bellman rank):

Key intuition: optimism in the face of uncertainty

Summary of the Course

Basics of MDPs (and LQRs):

Planning: VI, PI, LP formulations, Fitted Q-iteration (under B-complete), Low bounds on linear $Q^{\star}$

Exploration in MDPs (bandit / tabular / linear mdp / Bellman rank):

Key intuition: optimism in the face of uncertainty

Policy Gradient methods (tabular, linear, and neural)

Global convergence of PG and NPG (if the reset distribution covers $d^{\star}$)

Summary of the Course

Basics of MDPs (and LQRs):

Planning: VI, PI, LP formulations, Fitted Q-iteration (under B-complete), Low bounds on linear $Q^\star$

Exploration in MDPs (bandit / tabular / linear mdp / Bellman rank):

Key intuition: optimism in the face of uncertainty

Policy Gradient methods (tabular, linear, and neural)

Global convergence of PG and NPG (if the reset distribution covers $d^\star$)

Imitation Learning

Distribution shift in offline IL and how we overcome it in the hybrid and interactive settings