Carnegie Mellon

School of Computer Science

Deep Reinforcement Learning and Control

Introduction to Deep Reinforcement Learning and Control

Lecture 1, CMU 10703

Katerina Fragkiadaki

Logistics

- Three homework assignments and a final project, 60%/40%
- Final project: making progress on manipulating novel objects or navigating simple mazes.
- Resources: AWS for those that do not have access to GPUs
- Lectures will be recorded and will be available inside CMU
- Prerequisites: email us if you have not taken the official prerequisites but you equivalent of those
- Time conflicts
- People can audit the course, unless there are no seats left in class

Goal of the Course

How to build agents that **learn** to act and accomplish specific **goals** in **dynamic** environments?

as opposed to agents that execute **preprogrammed** behaviors in a **static** environment…

Motor control is Important

The brain evolved, not to think or feel, but to control movement.

Daniel Wolpert, nice TED talk

Daniel Wolpert: The real reason for brains | TED Talk | TED.com

https://www.ted.com/talks/daniel_wolpert_the_real_reason_for_brains

Motor control is Important

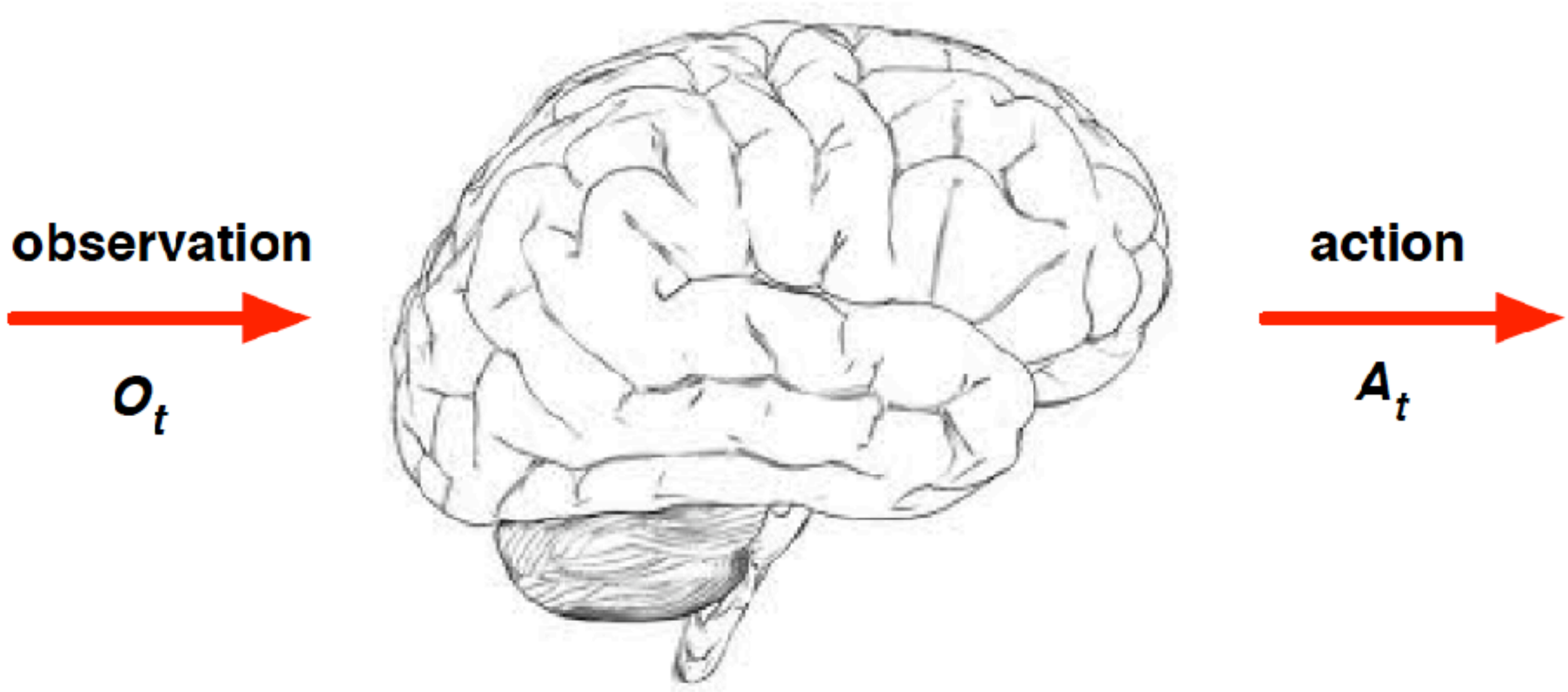
The brain evolved, not to think or feel, but to control movement.

Daniel Wolpert, nice TED talk

Sea squirts digest their own brain when they decide not to move anymore

Learning to Act

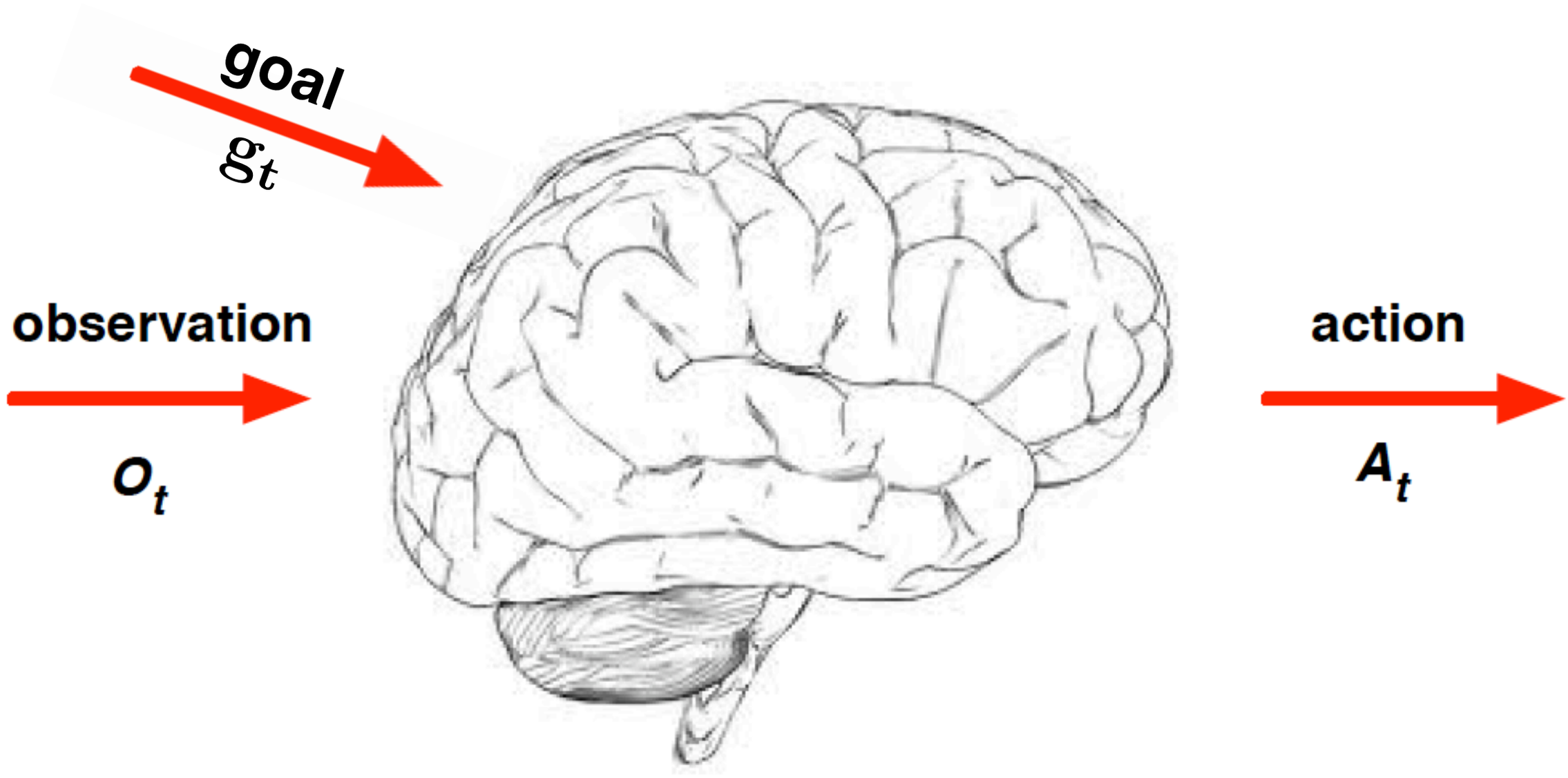
Learning to map sequences of observations to actions



observations: inputs from our sensor

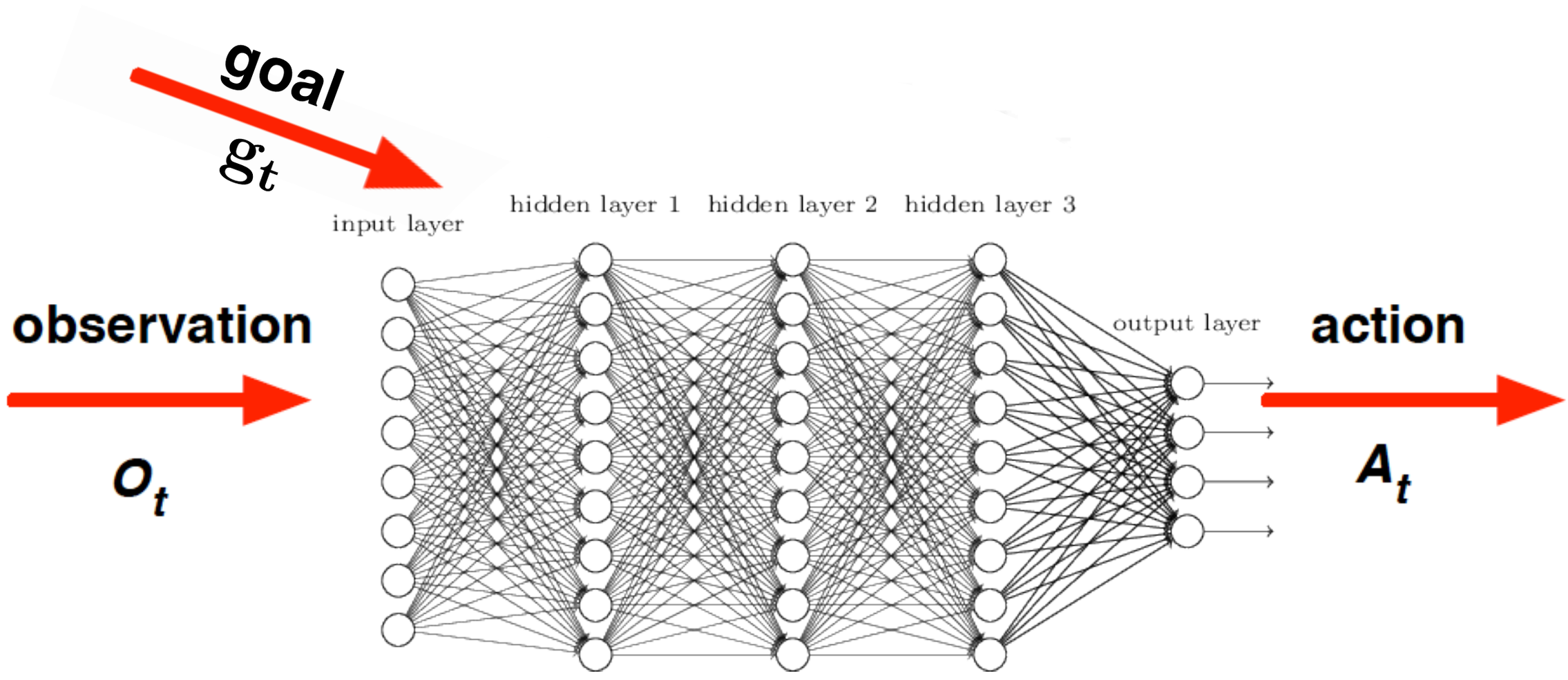
Learning to Act

Learning to map sequences of observations to actions, for a particular goal



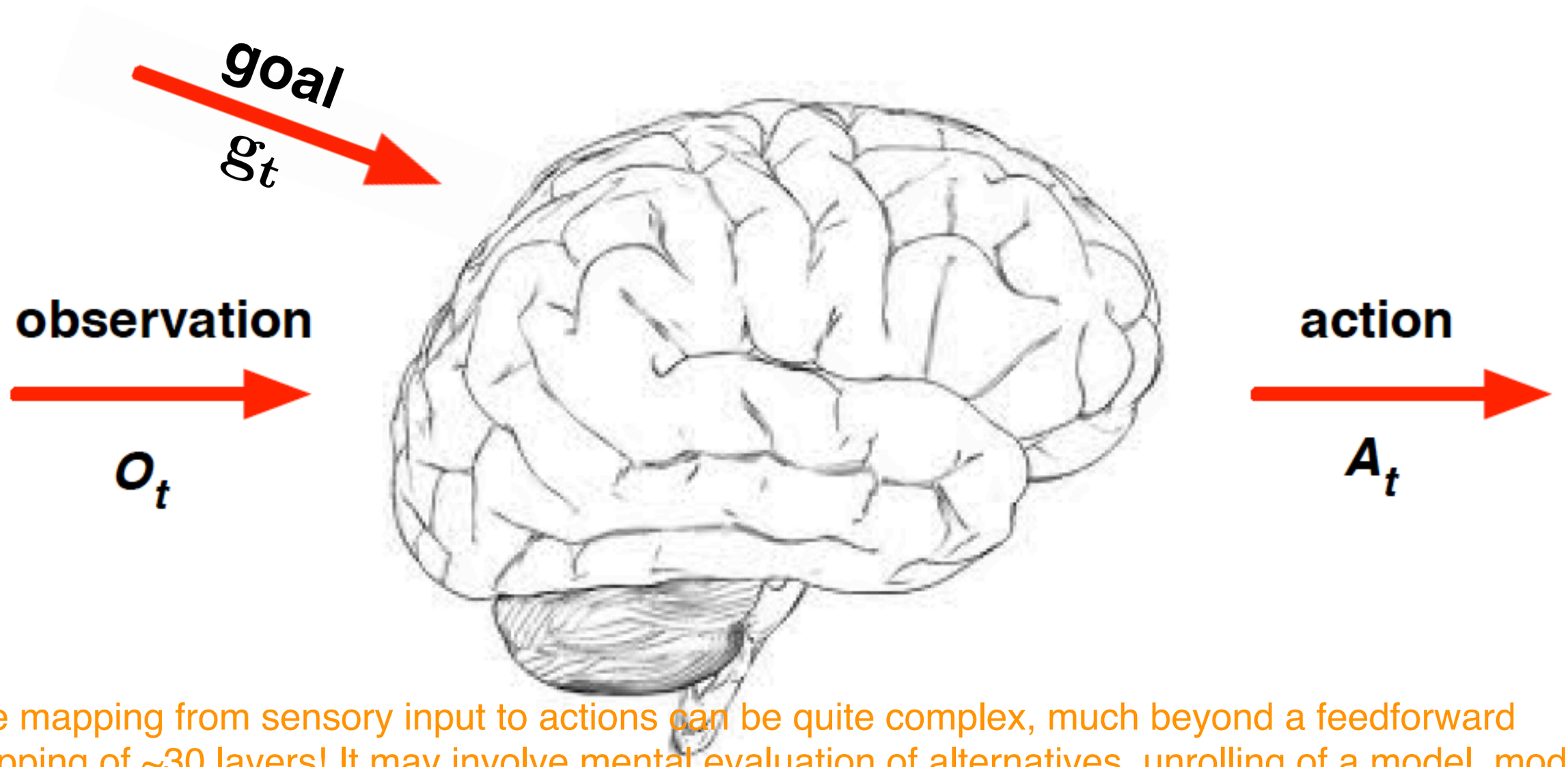
Learning to Act

Learning to map sequences of observations to actions, for a particular goal



Learning to Act

Learning to map sequences of observations to actions, for a particular goal



The mapping from sensory input to actions can be quite complex, much beyond a feedforward mapping of ~30 layers! It may involve mental evaluation of alternatives, unrolling of a model, model updates, closed loop feedback, retrieval of relevant memories, hypothesis generation, etc. .

Learning to Act

Different mappings as we go from Novice to Expert

https://youtu.be/H6Ah-Fa_R9c?t=17

Supervision

What supervision does an agent need to learn purposeful behaviors in dynamic environments?

- **Rewards:** sparse feedback from the environment whether the desired goal is achieved e.g., game is won, car has not crashed, agent is out of the maze etc.

Supervision

What supervision does an agent need to learn purposeful behaviors in dynamic environments?

- **Rewards:** sparse feedback from the environment whether the desired goal is achieved e.g., game is won, car has not crashed, agent is out of the maze etc.

Rewards can be intrinsic, i.e., generated by the agent and guided by its curiosity as opposed to an external task

Behavior: High Jump

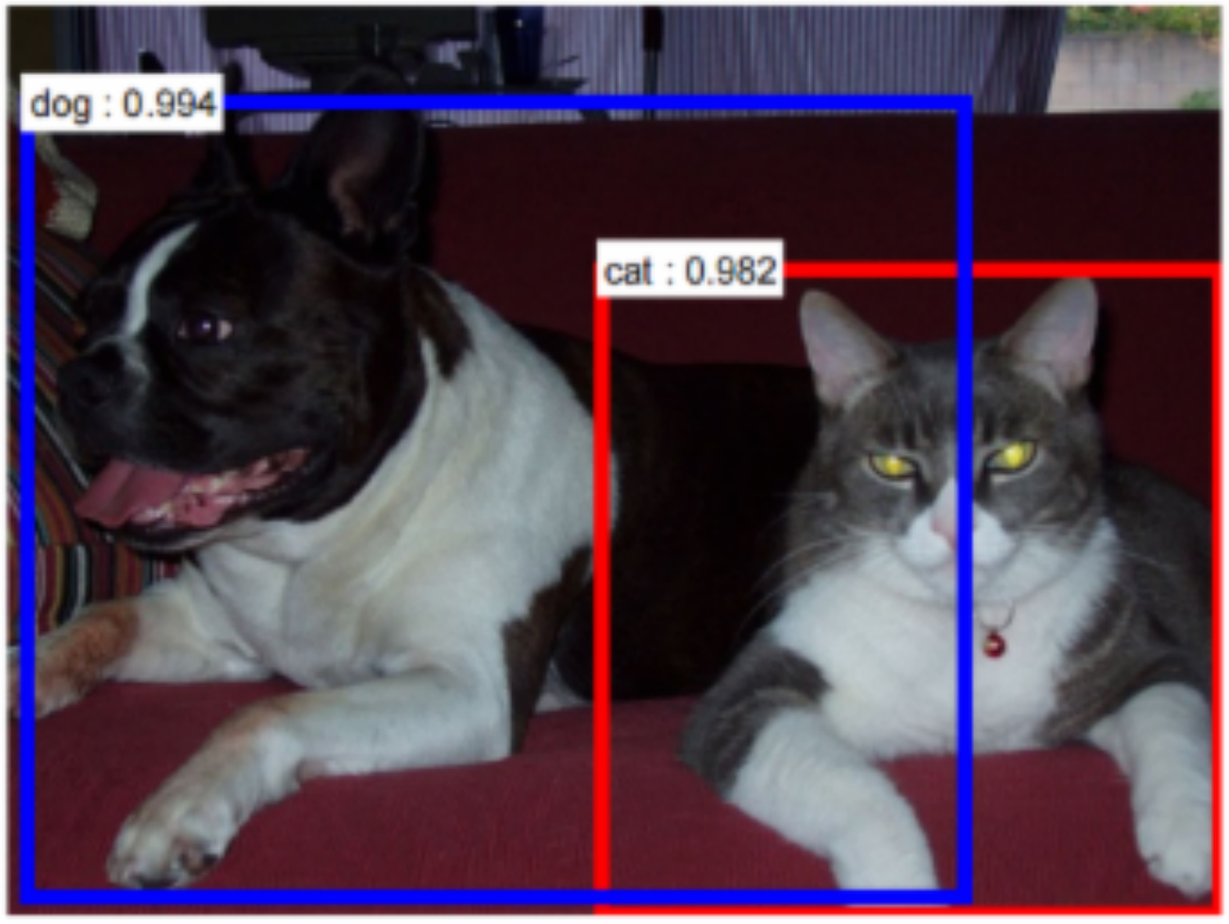
scissors

Fosbury flop

1. Learning from rewards
   Reward: jump as high as possible: It took years for athletes to find the right behavior to achieve this
2. Learns from demonstrations
   It was way easier for athletes to perfection the jump, once someone showed the right general trajectory
3. Learns from specifications of optimal behavior
   For novices, it is much easier to replicate this behavior if additional guidance is provided based on specifications: where to place the foot, how to time yourself etc.

Learning to Act

How learning to act is different than other machine learning paradigms, e.g., object detection?



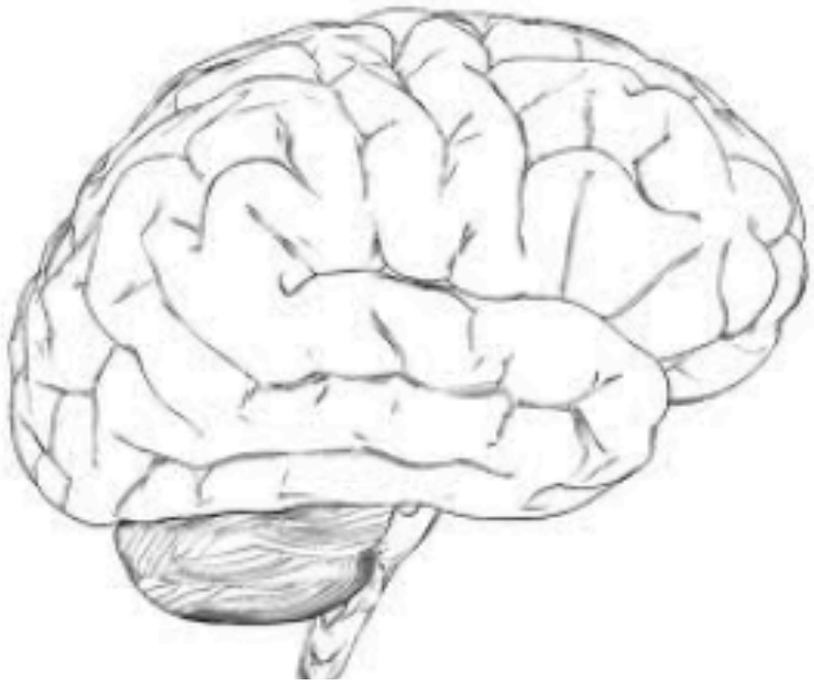
Learning to Act

How learning to act is different than other machine learning paradigms?

- The agent's actions affect the data she will receive in the future

observation

O_t



action

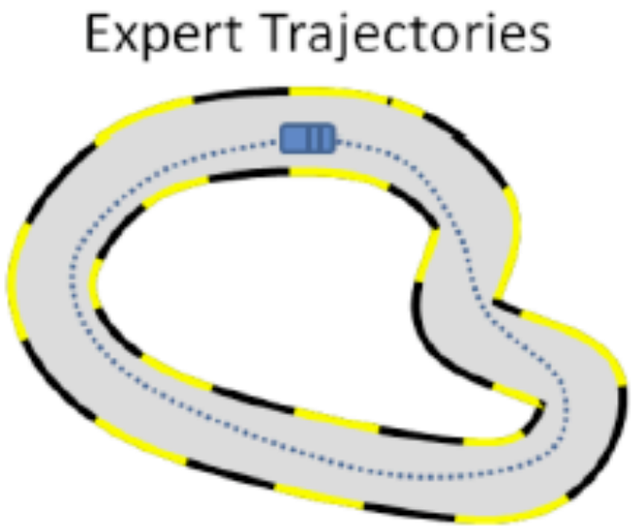
A_t

Learning to Act

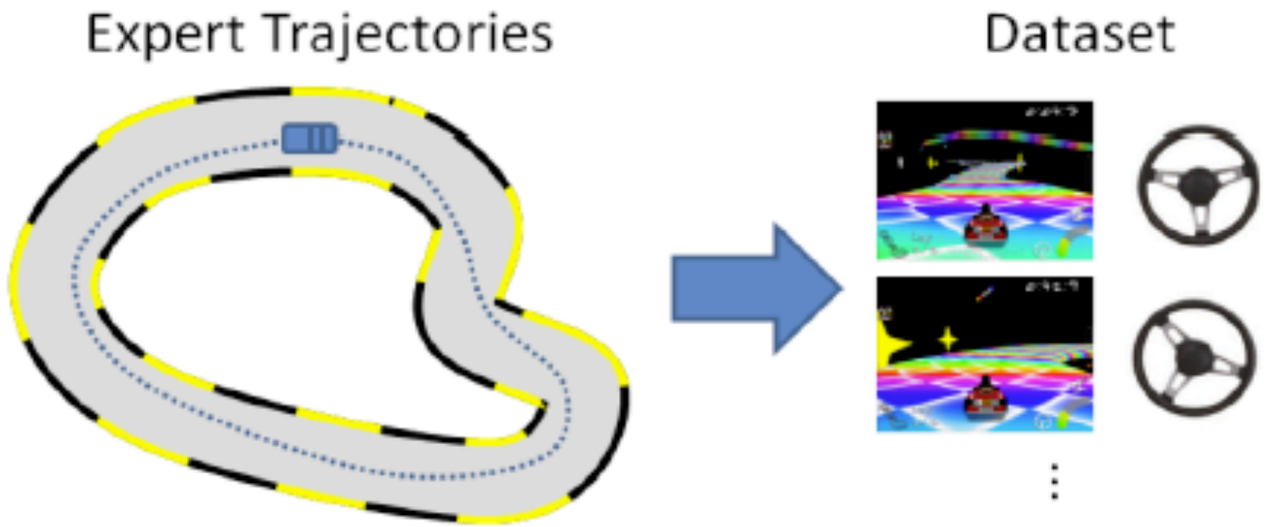
How learning behaviors is different than other machine learning paradigms?

- The agent's actions affect the data she will receive in the future:
- The data the agent receives are sequential in nature, not i.i.d.
- Standard supervised learning approaches lead to compounding errors, *An invitation to imitation*, Drew Bagnell

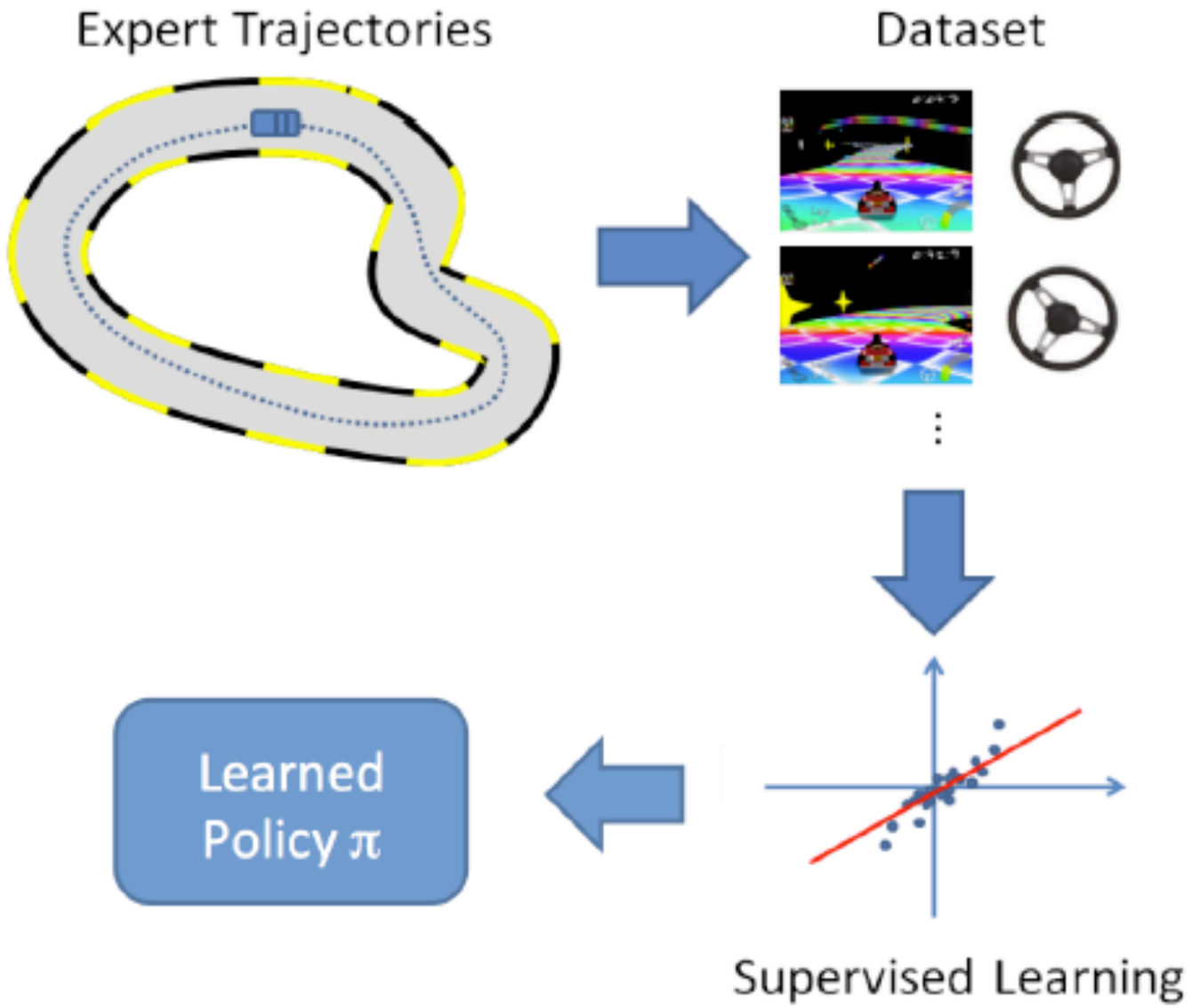
Learning to Drive a Car: Supervised Learning



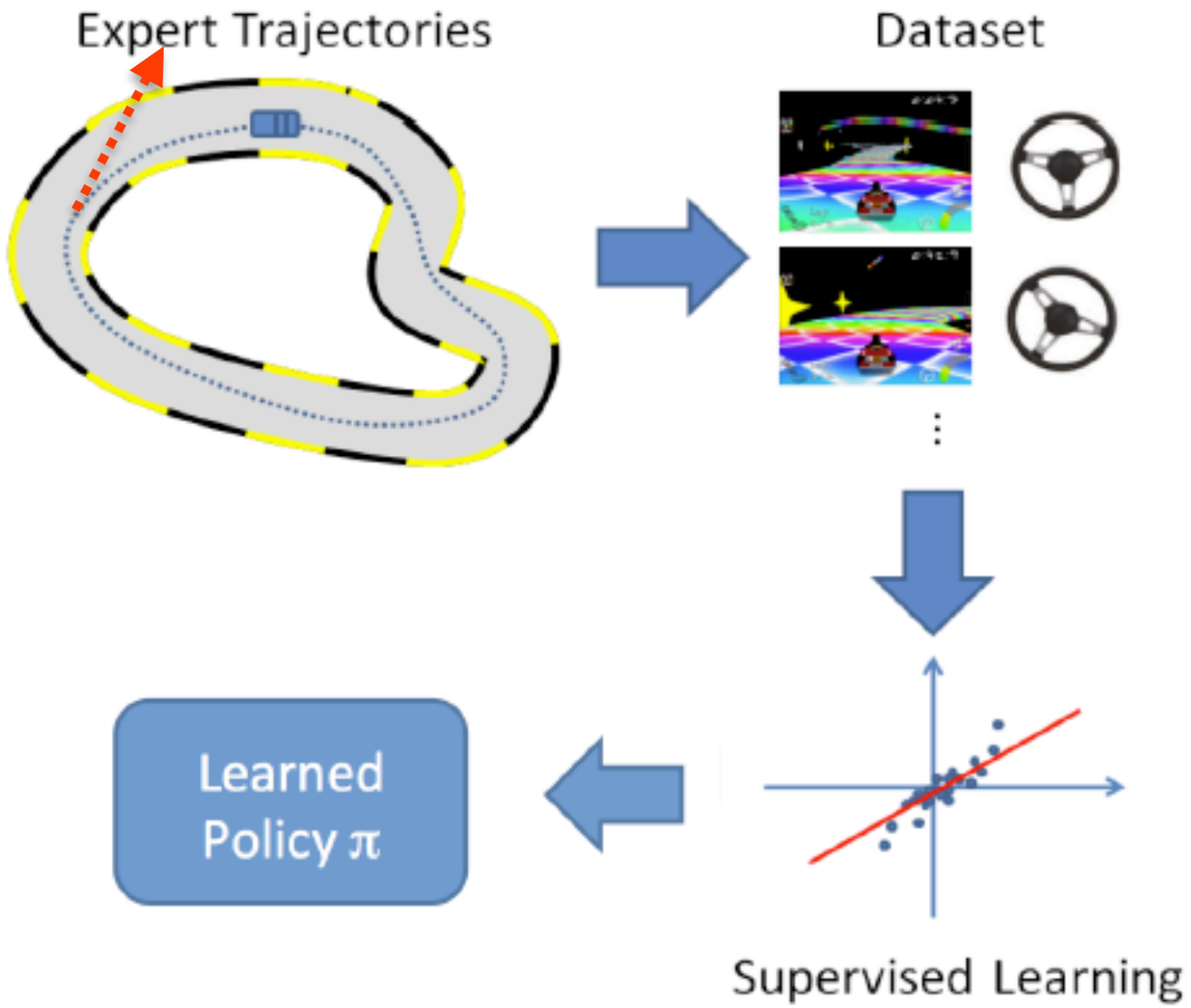
Learning to Drive a Car: Supervised Learning



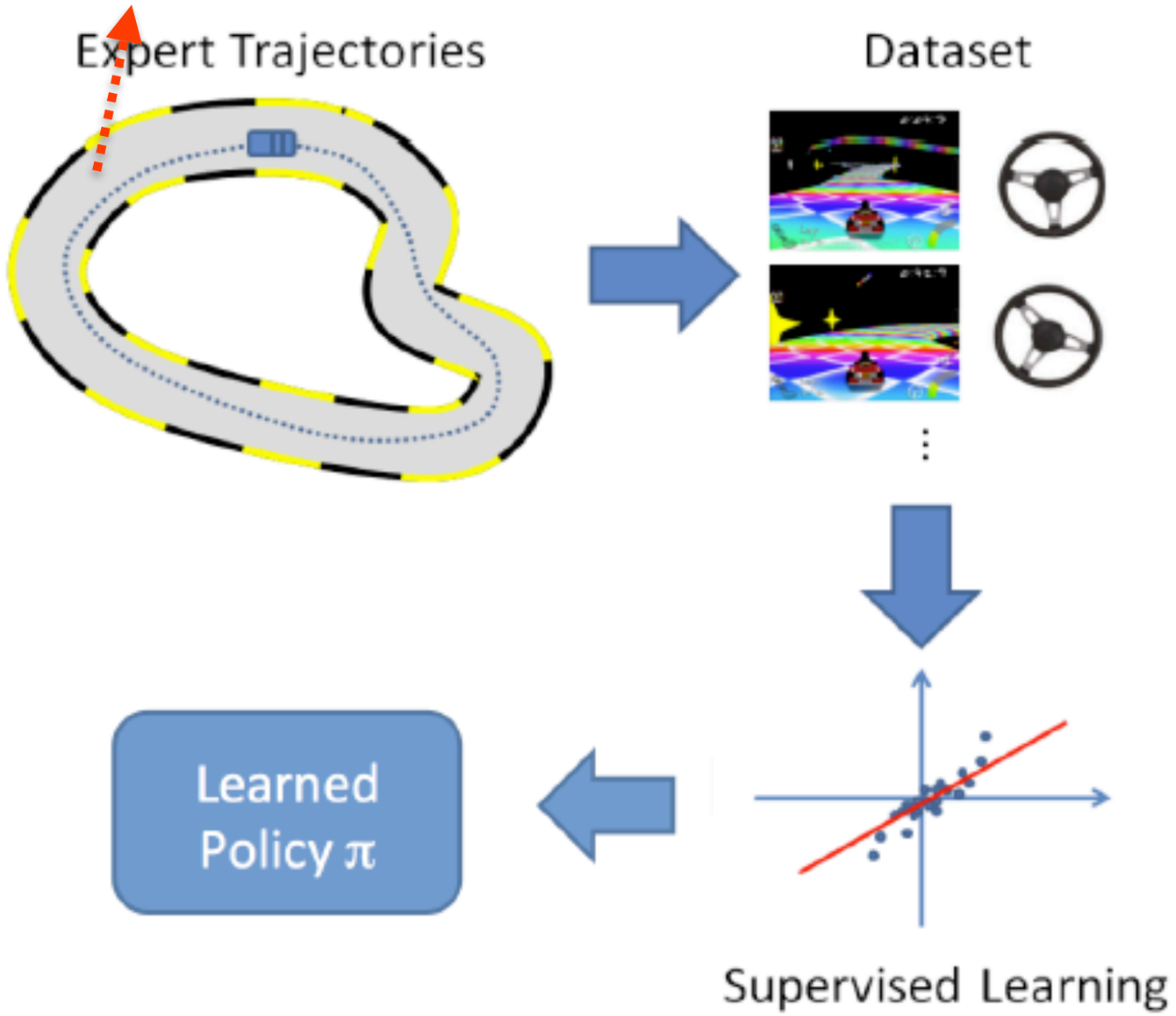
Learning to Drive a Car: Supervised Learning



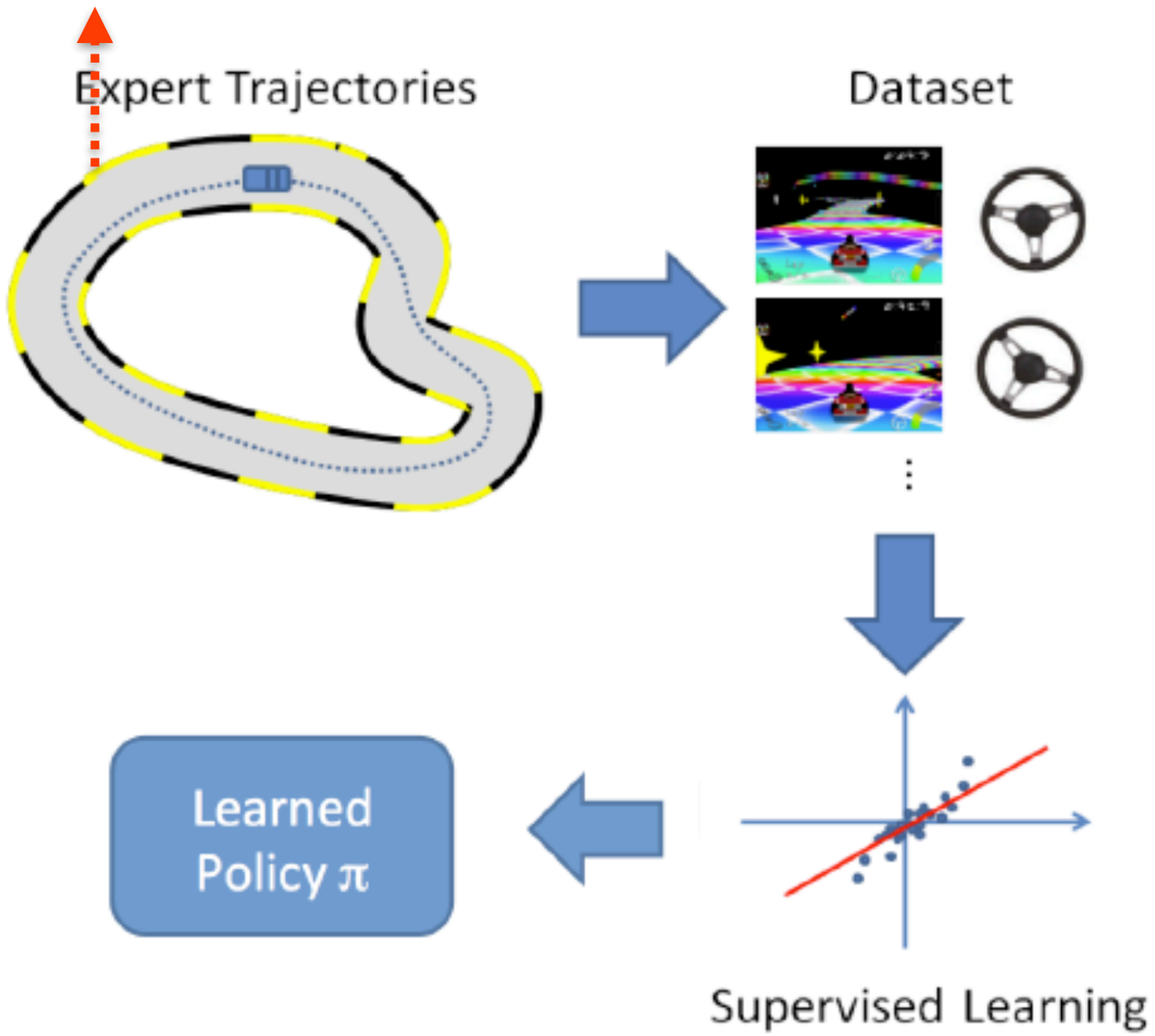
Learning to Drive a Car: Supervised Learning



Learning to Drive a Car: Supervised Learning



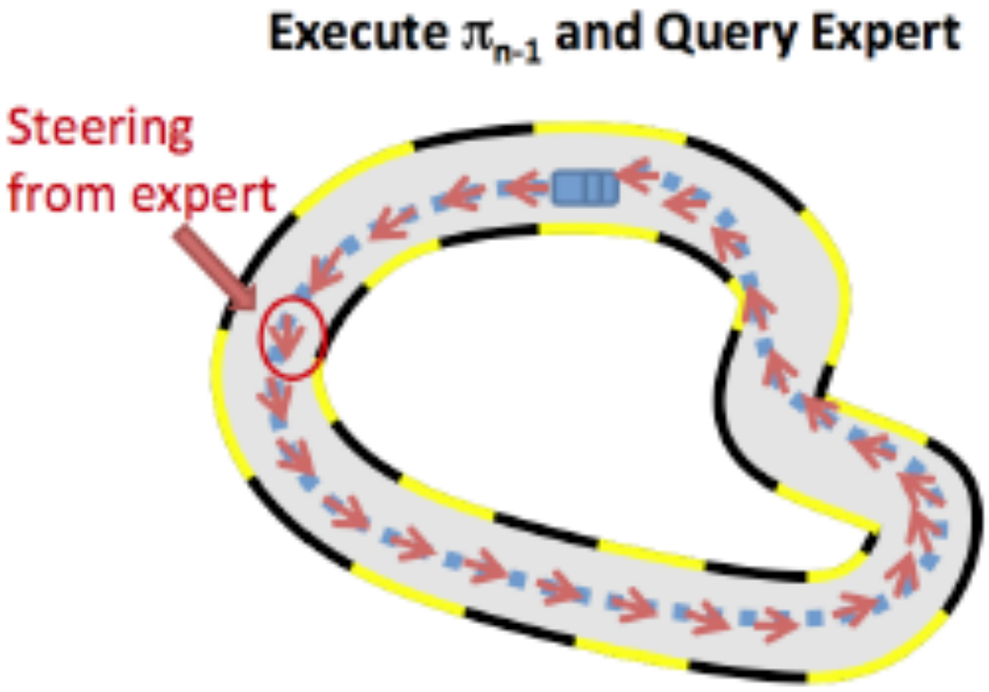
Learning to Drive a Car: Supervised Learning



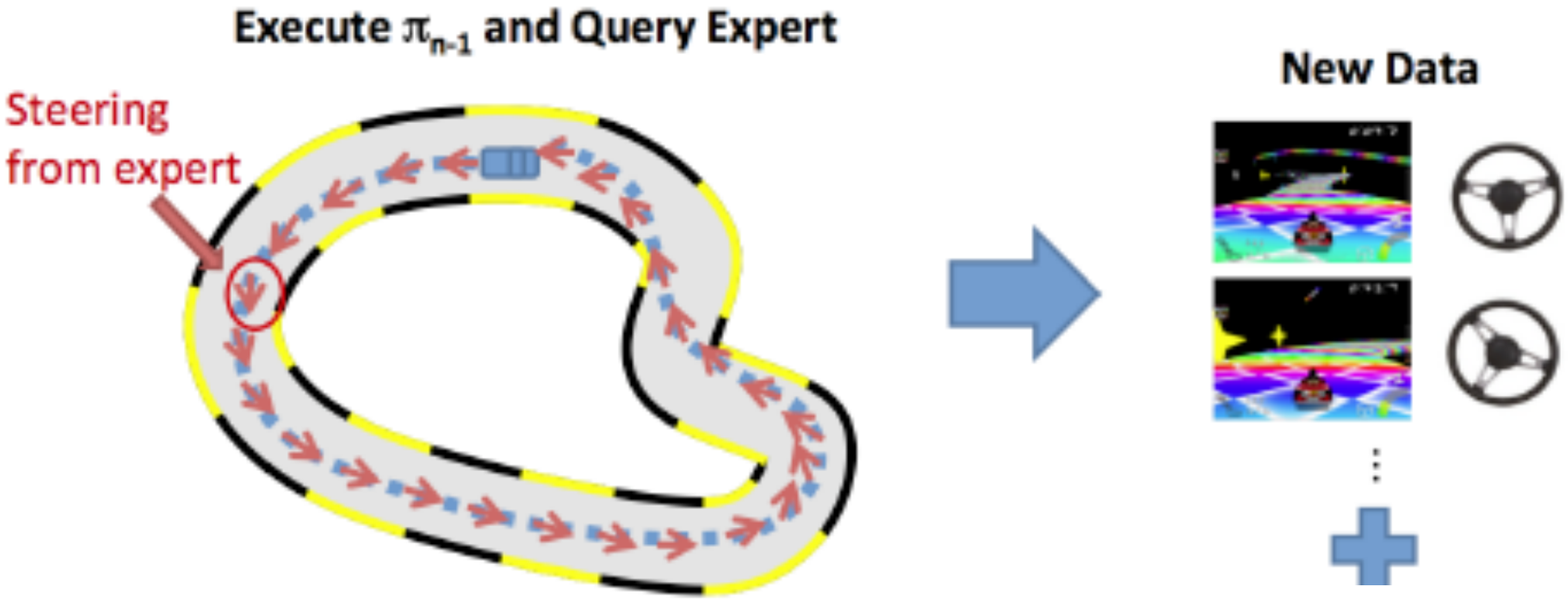
Learning to Drive a Car: Supervised Learning

Compounding errors

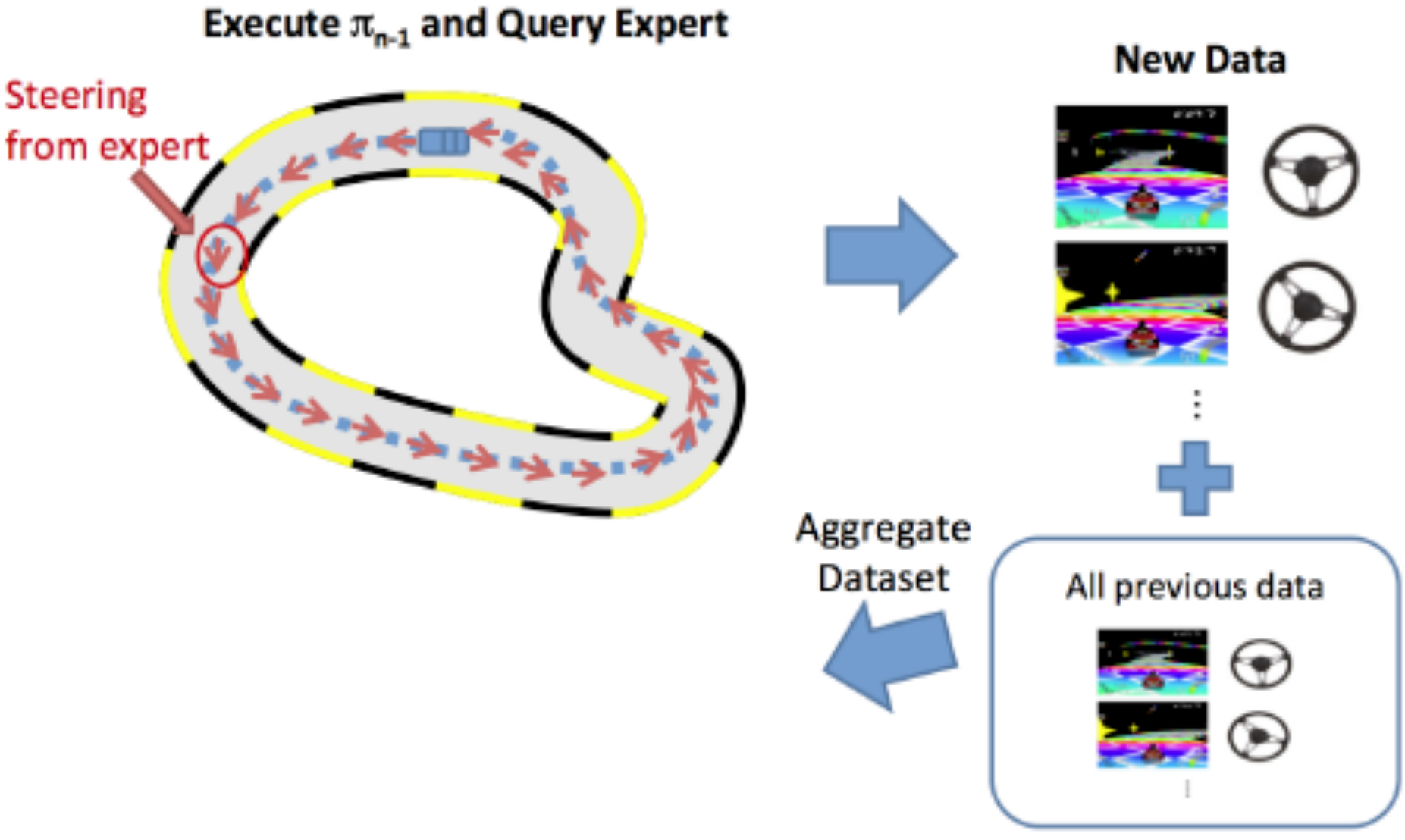
Learning to Race a Car : Interactive learning-DAGGer



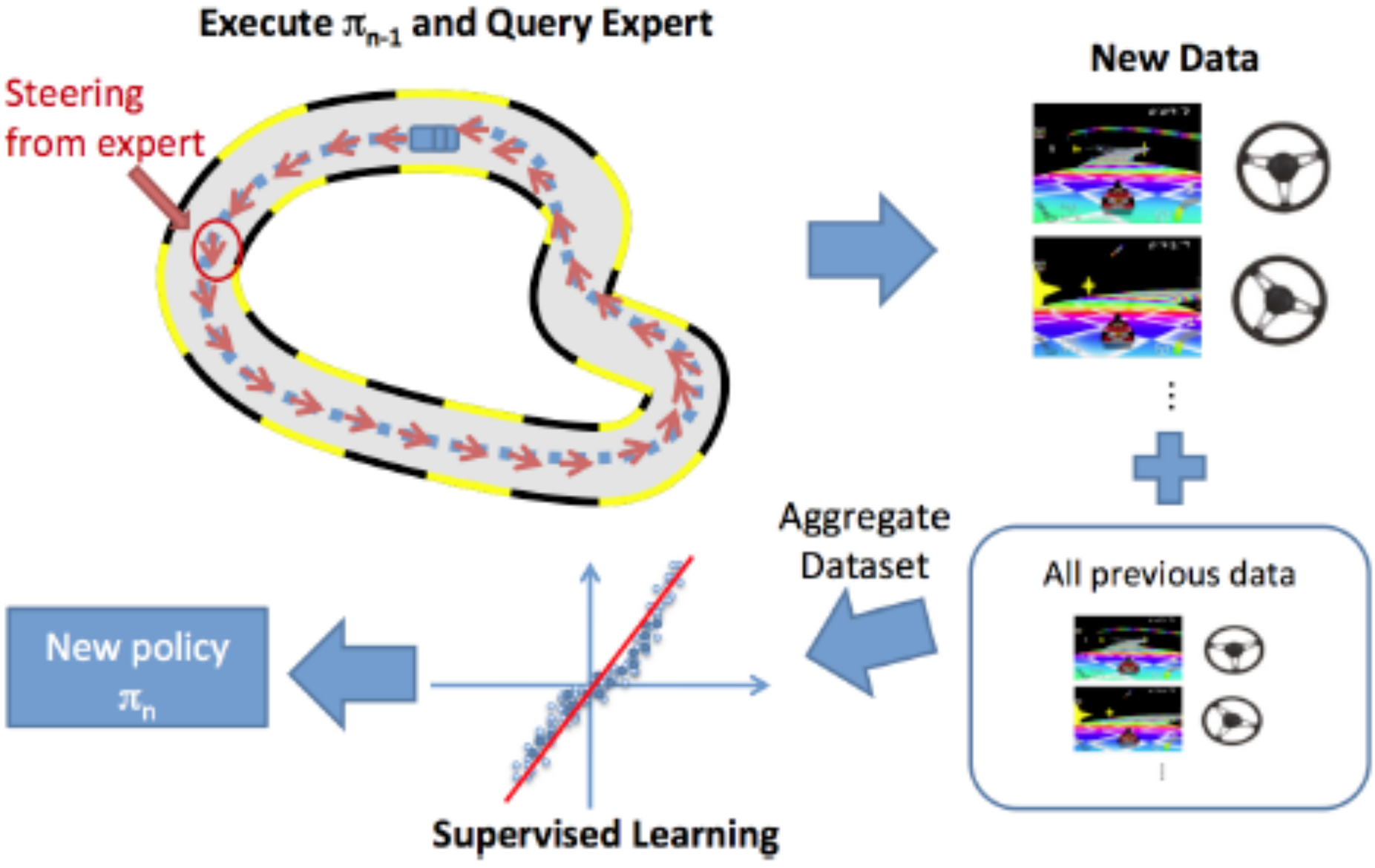
Learning to Race a Car : Interactive learning-DAGGer



Learning to Race a Car : Interactive learning-DAGGer

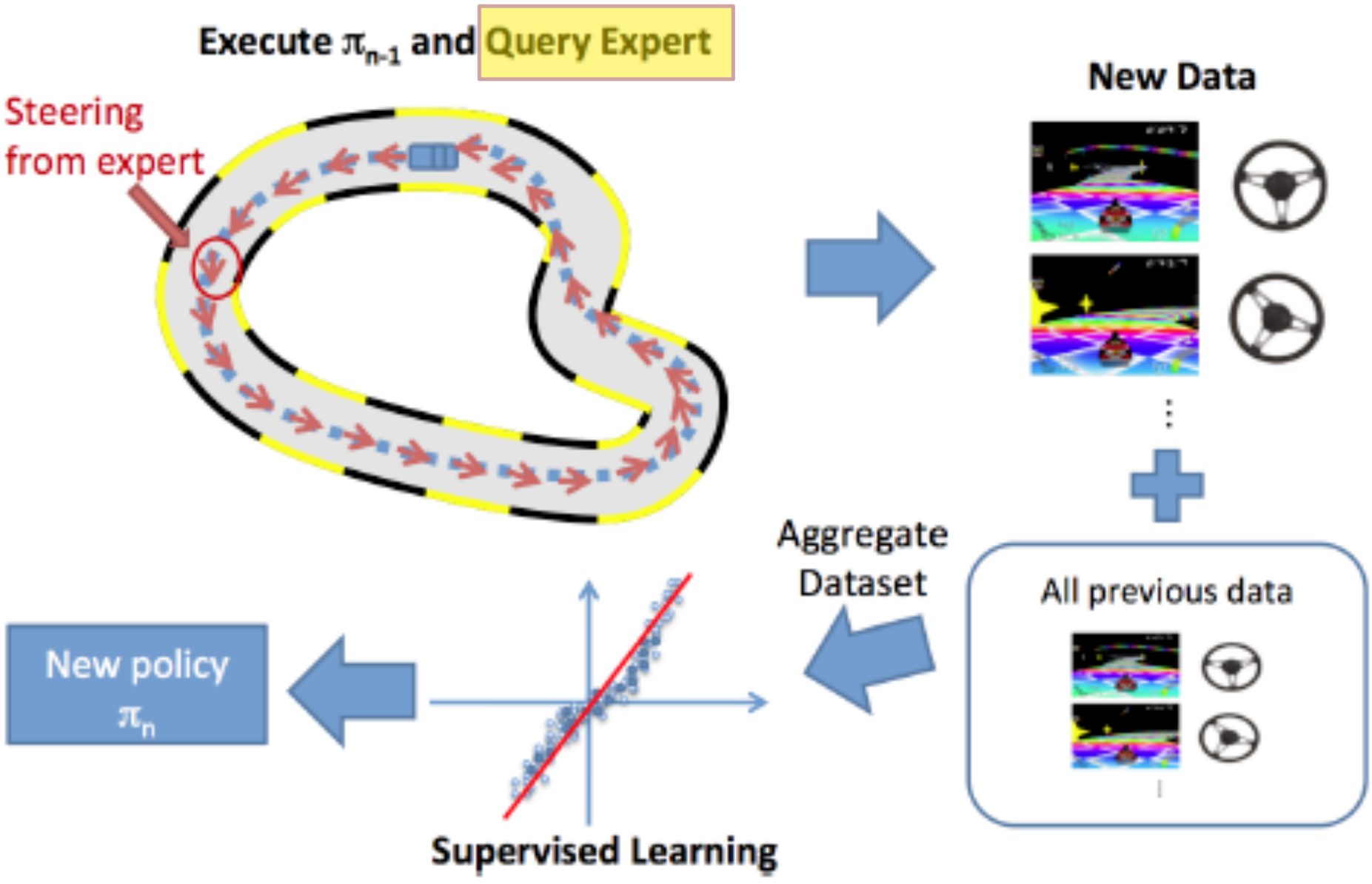


Learning to Race a Car : Interactive learning-DAGGer



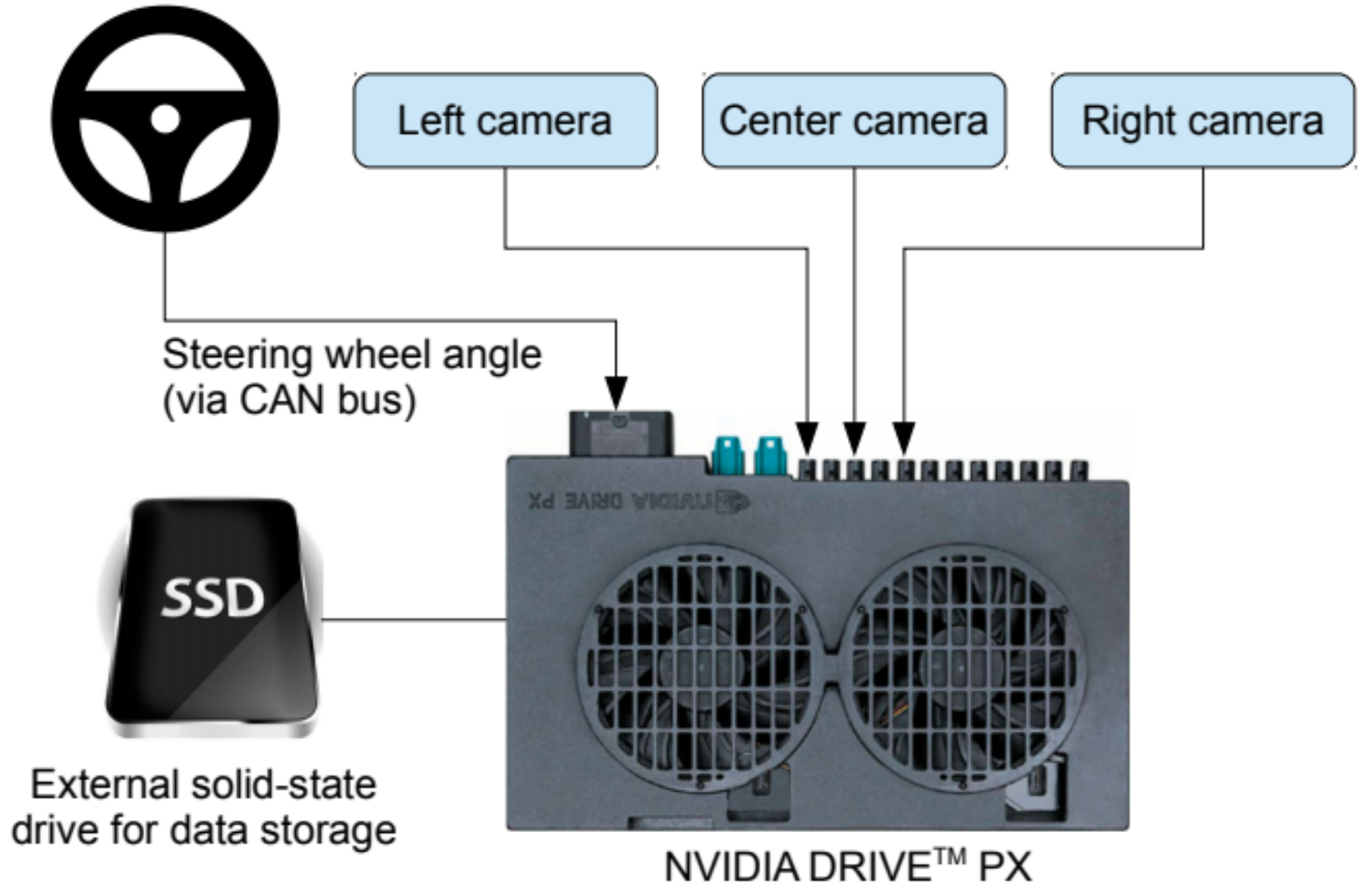
Learning to Race a Car : Interactive learning-DAGGer

This assumes you can actively access an expert during training!



A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
Stephane Ross, Geoffrey J. Gordon, J. Andrew Bagnell

Learning to Drive a Car: Supervised Learning



End to End Learning for Self-Driving Cars, NVIDIA, 2016

Learning to Act

How learning behaviors is different than other machine learning paradigms?

1) The agent's actions affect the data she will receive in the future

2) The reward (whether the goal of the behavior is achieved) is far in the future:

   - Temporal credit assignment: which actions were important and which were not, is hard to know

Learning to Act

How learning behaviors is different than other machine learning paradigms?

1) The agent's actions affect the data she will receive in the future

2) The reward (whether the goal of the behavior is achieved) is far in the future:

- Temporal credit assignment: which actions were important and which were not, is hard to know

But wait! isn't it the same with object detection??

Reward (loss) only at the top layer!

Learning to Act

How learning behaviors is different than other machine learning paradigms?

1) The agent's actions affect the data she will receive in the future

2) The reward (whether the goal of the behavior is achieved) is far in the future:

- Temporal credit assignment: which actions were important and which were not, is hard to know

But wait! isn't it the same with object detection??
No: here the horizon involves acting in the environment, rather than going from one neural layer to the next, we cannot apply chain rule to propagate the rewards backwards..

Learning to Act

How learning behaviors is different than other machine learning paradigms?

1) The agent's actions affect the data she will receive in the future
2) The reward (whether the goal of the behavior is achieved) is far in the future:
3) Actions take time to carry out in the real world, and thus this may limit the amount of experience

Learning to Act

How learning behaviors is different than other machine learning paradigms?

1) The agent's actions affect the data she will receive in the future
2) The reward (whether the goal of the behavior is achieved) is far in the future:
3) Actions take time to carry out in the real world, and thus this may limit the amount of experience

- We can use simulated experience and tackle the sim2real transfer

Learning to Act

How learning behaviors is different than other machine learning paradigms?

1) The agent's actions affect the data she will receive in the future
2) The reward (whether the goal of the behavior is achieved) is far in the future:
3) Actions take time to carry out in the real world, and thus this may limit the amount of experience

- We can use simulated experience and tackle the sim2real transfer
- We can have robots working 24/7

Supersizing Self-Supervision

Supersizing Self-supervision: Learning to Grasp from 50K Tries and 700 Robot Hours, Pinto and Gupta

Learning to Act

How learning behaviors is different than other machine learning paradigms?

1) The agent's actions affect the data she will receive in the future
2) The reward (whether the goal of the behavior is achieved) is far in the future:
3) Actions take time to carry out in the real world, and thus this may limit the amount of experience

- We can use simulated experience and tackle the sim2real transfer
- We can have robots working 24/7
- We can buy many robots

Google's Robot Farm

What if

We had fantastic simulators, with realistic Physics and realistic visuals and tactile sensing, and we could crowdsource tons of demonstrations, would we solve the problem them?

Successes so far

Backgammon

Backgammon

How is it different than chess?

Chess

Brute force manual development of a broad evaluation function

Backgammon

High branching factor due to dice roll prohibits brute force deep searches such as in chess

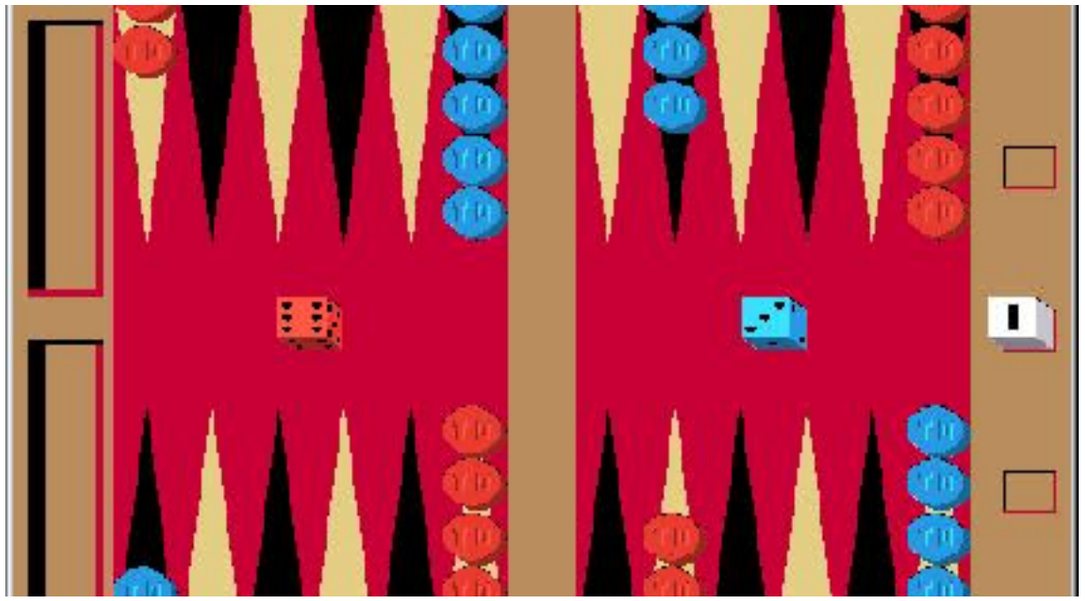
Backgammon

Neuro-Gammon

- Developed by Gerald Tesauro in 1989 in IBM’s research center
- Trained to mimic expert demonstrations using supervised learning
- Achieved intermediate-level human player

Backgammon

TD-Gammon

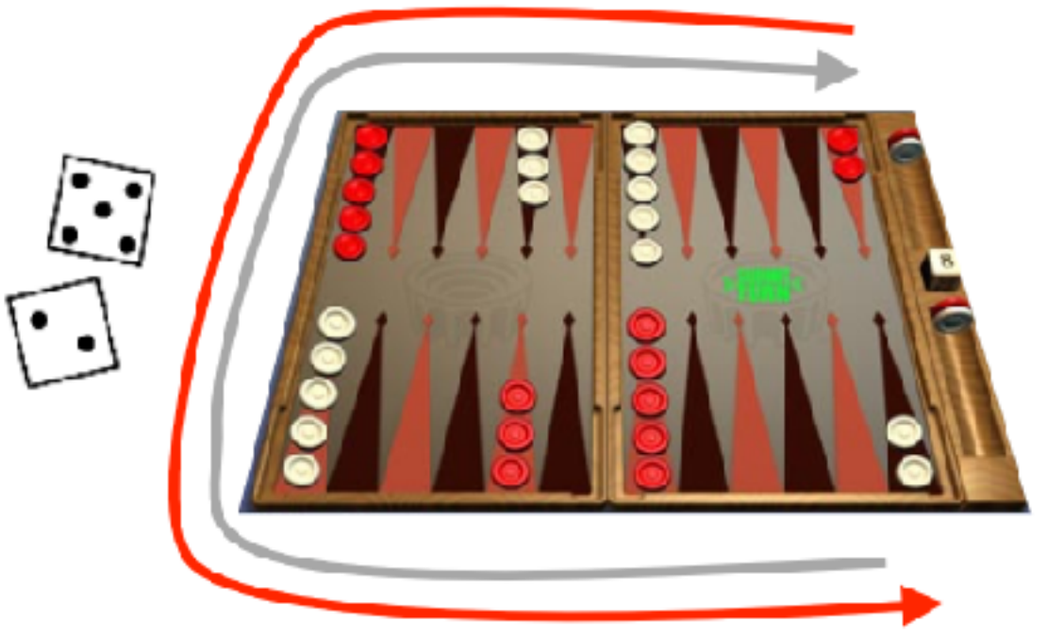


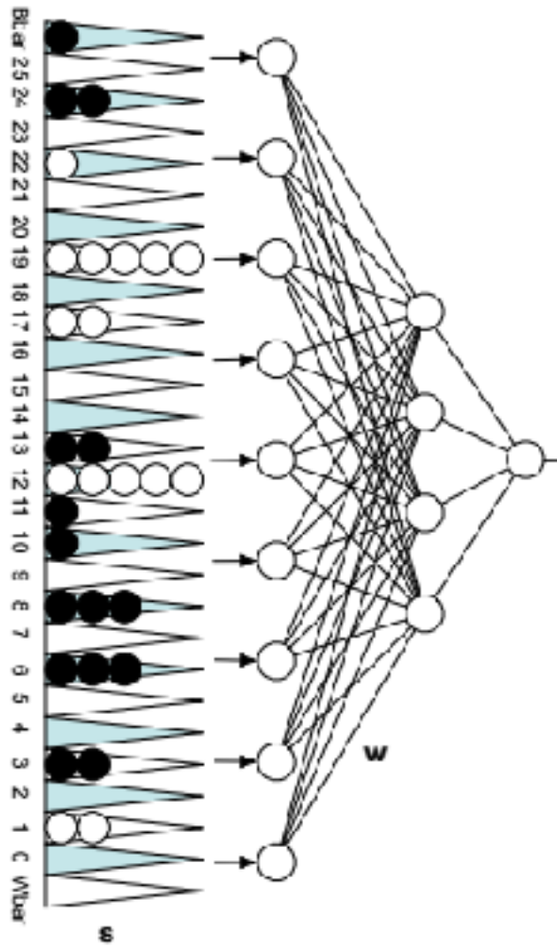
- Developed by Gerald Tesauro in 1992 in IBM's research center
- A neural network that trains itself to be an **evaluation function** by playing against itself starting from random weights
- Achieved performance close to top human players of its time

Neuro-Gammon

- Developed by Gerald Tesauro in 1989 in IBM's research center
- Trained to mimic expert demonstrations using supervised learning
- Achieved intermediate-level human player

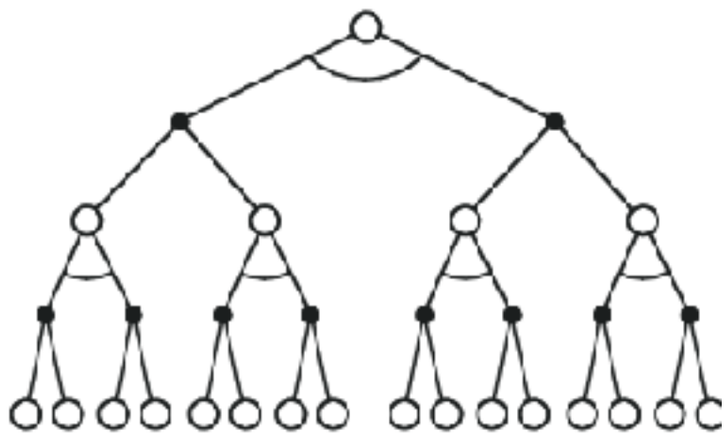
Evaluation function





estimated state value
(≈ prob of winning)

Action selection
by a shallow search

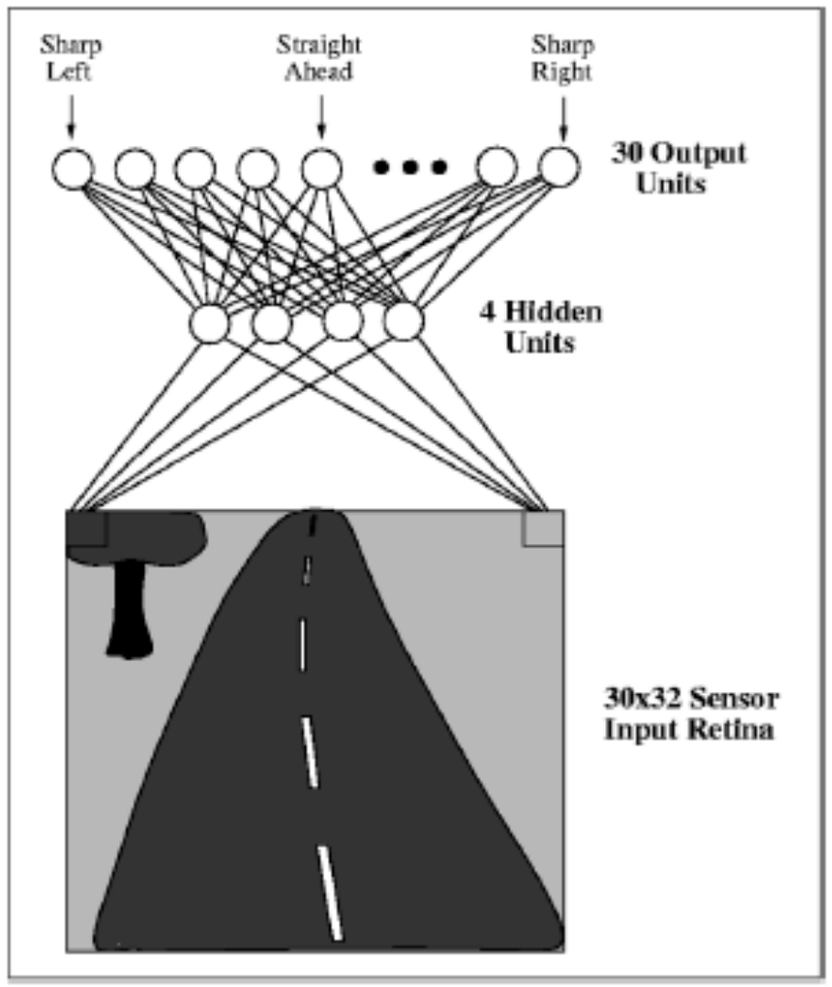


Self-Driving Cars

Self-Driving Cars

Policy network π:
mapping of
observations to actions

A_t



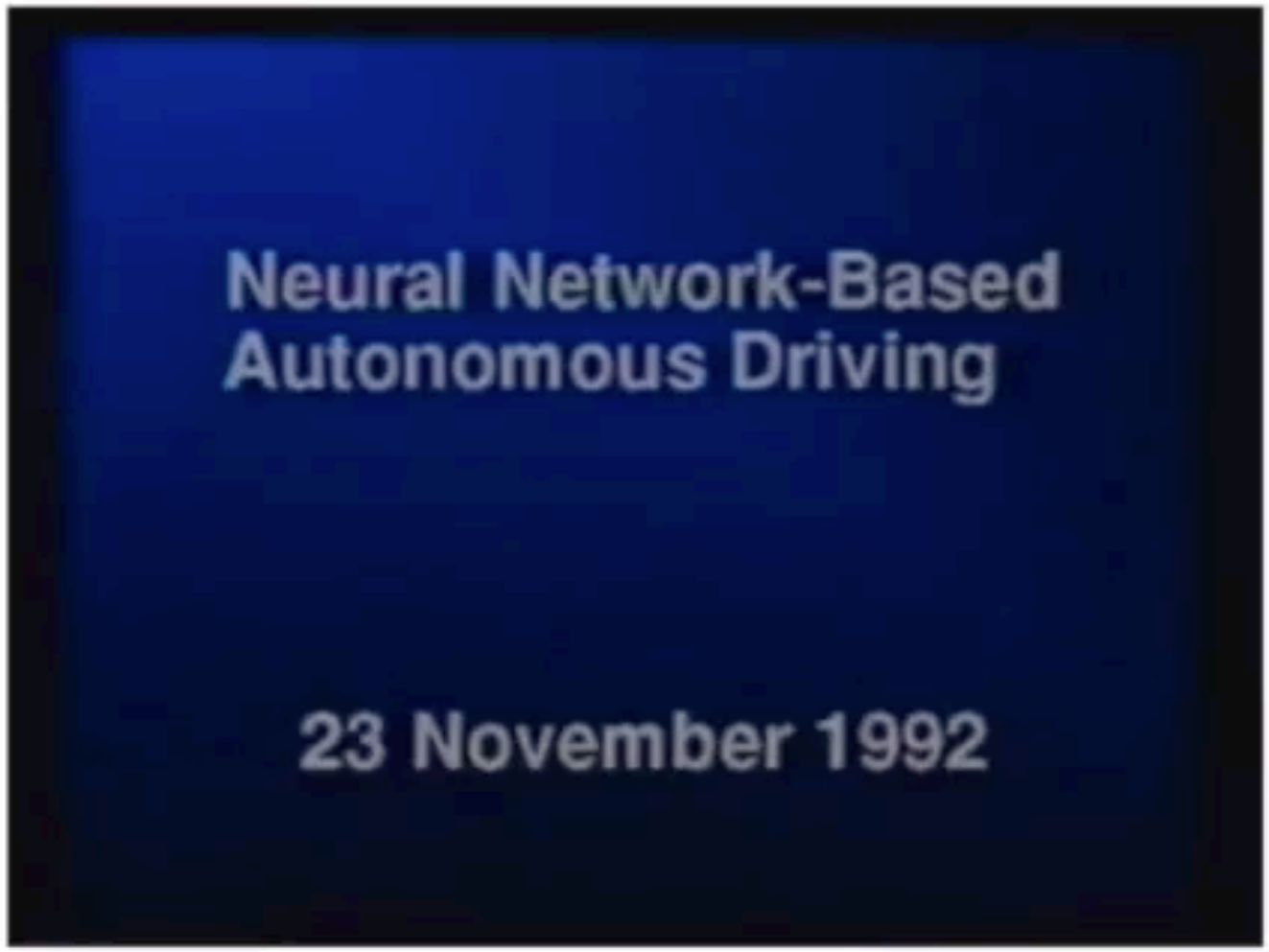
O_t

1989

ALVINN, an autonomous land vehicle in a neural network

Dean A. Pomerleau
Carnegie Mellon University

Self-Driving Cars



[Courtesy of Dean Pomerleau]

Behavior Cloning: data augmentation to deal with compounding errors, online adaptation (interactive learning)

ALVINN (Autonomous Land Vehicle In a Neural Network), *Efficient Training of Artificial Neural Networks for Autonomous Navigation*, Pomerleau 1991

Self-Driving Cars

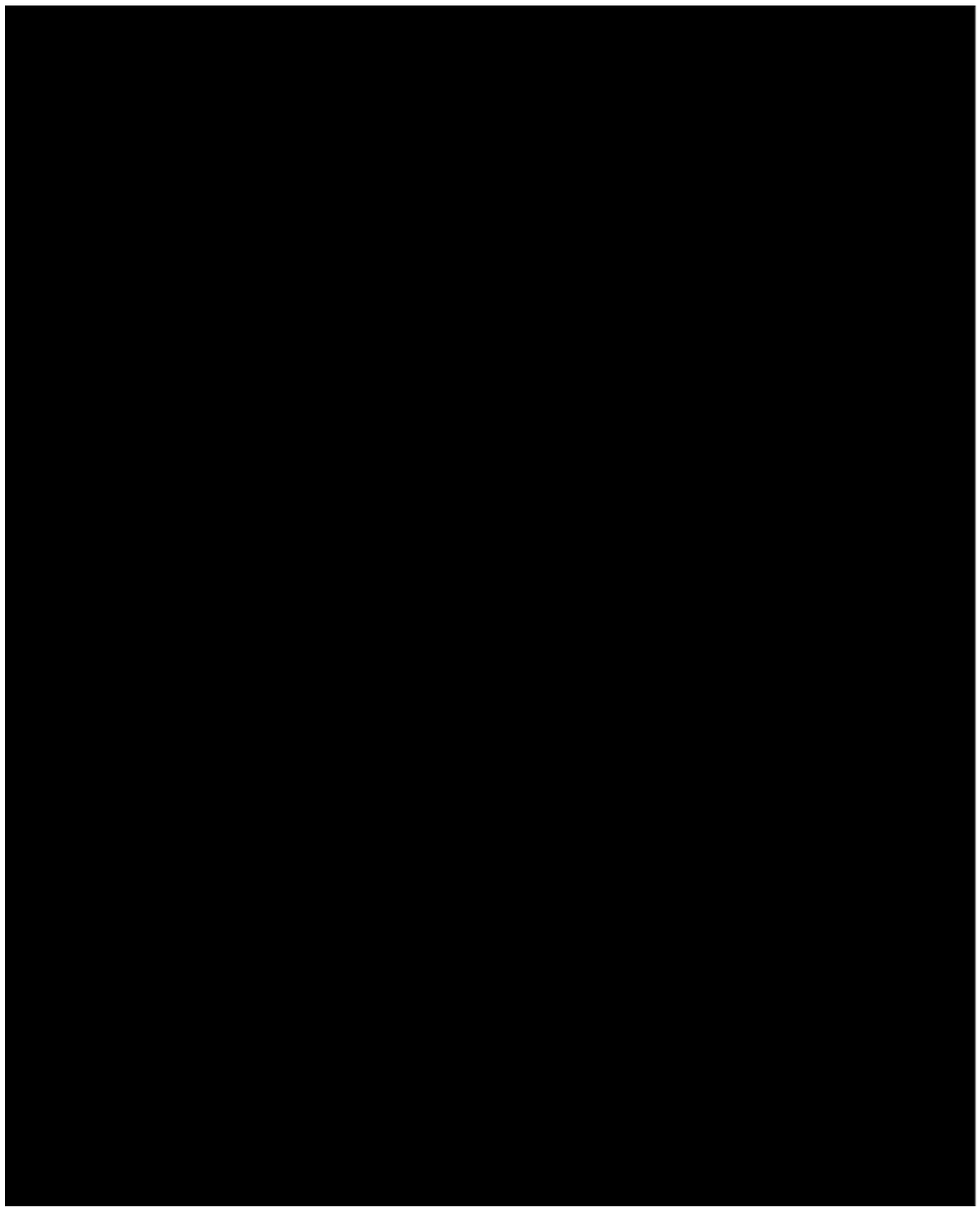
- P

Computer Vision, Velodyne sensors, object detection, 3D pose estimation, trajectory prediction

Self-Driving Cars

- Highway driving: solved problem
- Traffic jams, crowded intersection, complicated decision making, rare situations

Atari



Deep Mind 2014+

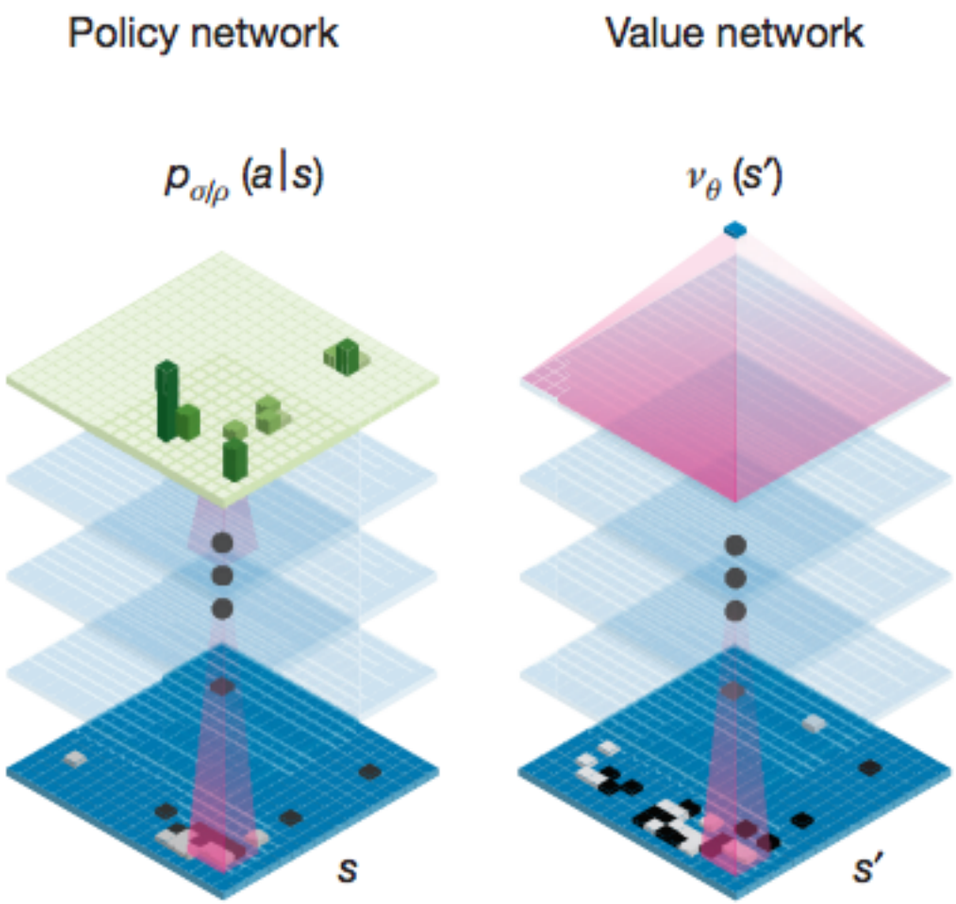
Deep Q learning

GO

AlphaGo

Monte Carlo Tree Search, learning policy and value function networks for pruning the search tree, trained from expert demonstrations, self play

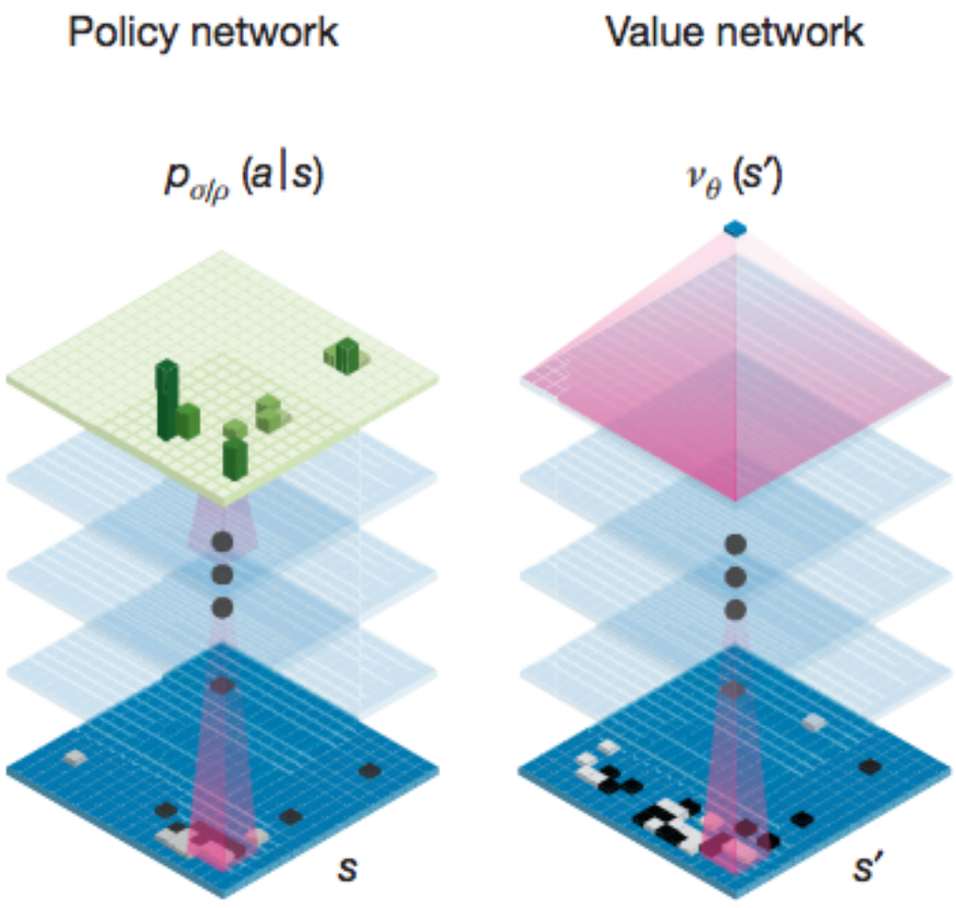
AlphaGo



Policy net trained to mimic expert moves, and then fine-tuned using self-play

Monte Carlo Tree Search, learning policy and value function networks for pruning the search tree, trained from expert demonstrations, self play

AlphaGo

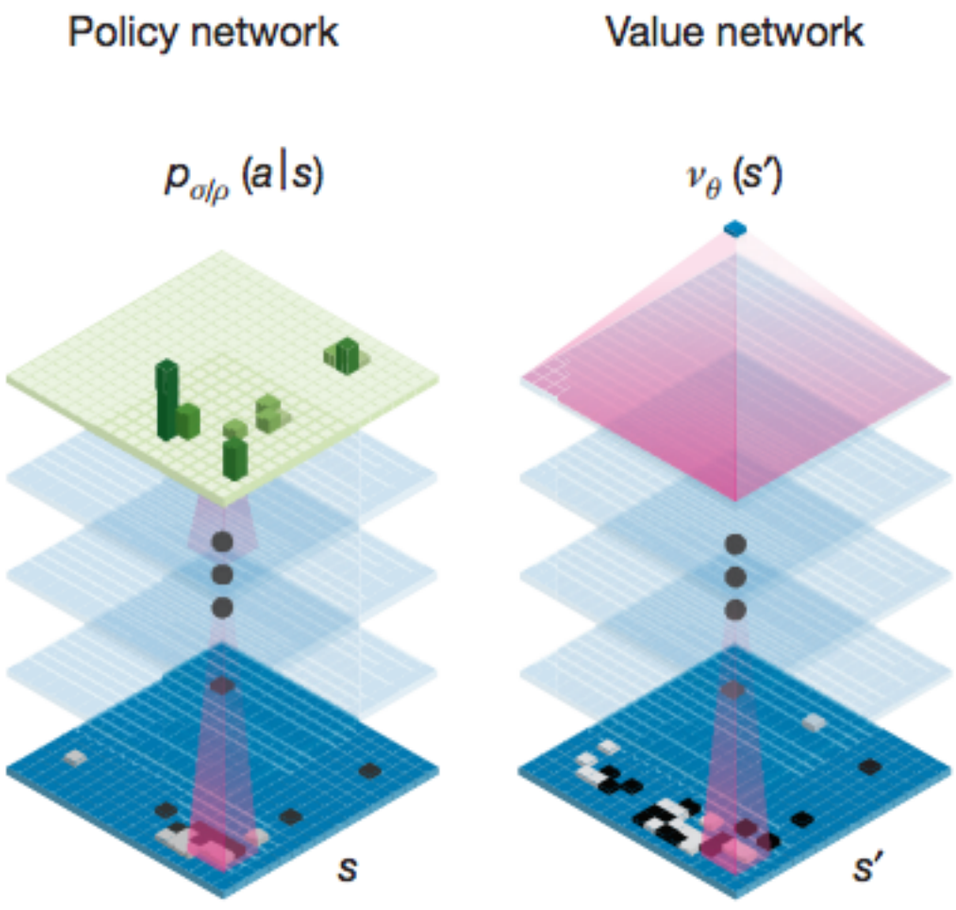


Policy net trained to mimic expert moves, and then fine-tuned using self-play

Value network trained with regression to predict the outcome, using self play data of the best policy.

Monte Carlo Tree Search, learning policy and value function networks for pruning the search tree, trained from expert demonstrations, self play

AlphaGo



Policy net trained to mimic expert moves, and then fine-tuned using self-play

Value network trained with regression to predict the outcome, using self play data of the best policy.

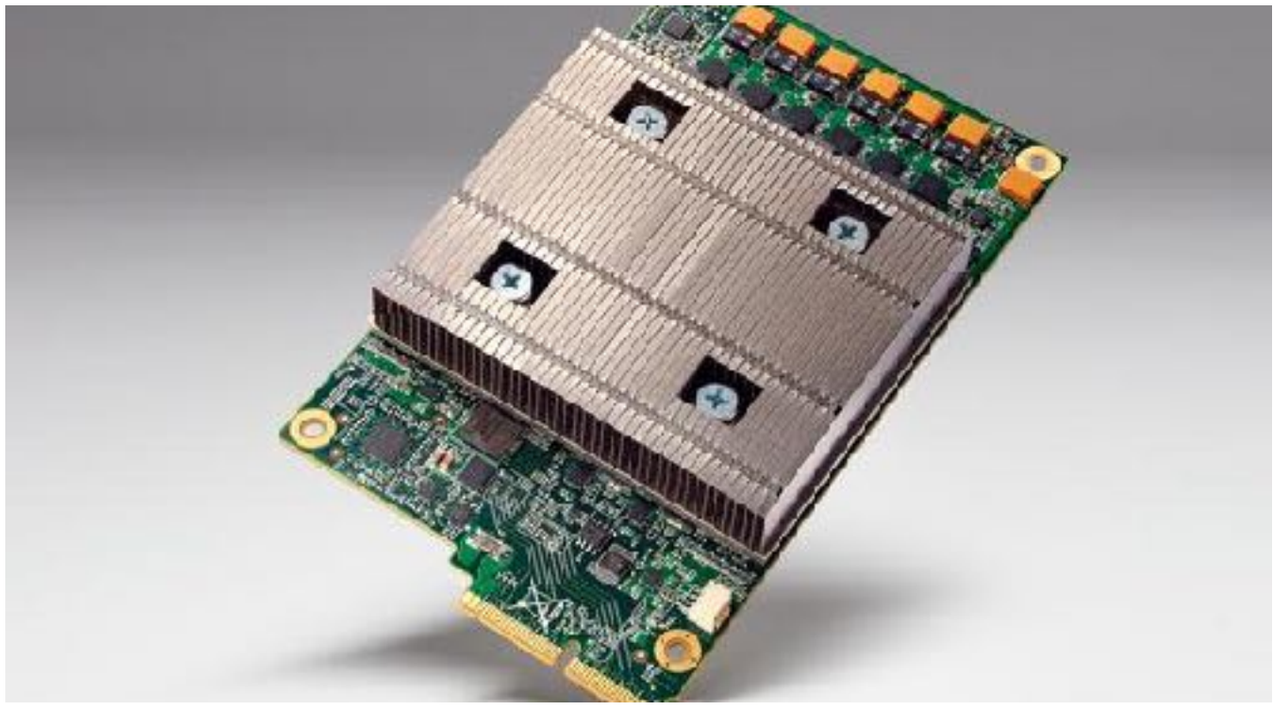
At test time, policy and value nets guide a MCTS to select stronger moves by deep look ahead.

Monte Carlo Tree Search, learning policy and value function networks for pruning the search tree, trained from expert demonstrations, self play

AlphaGo

Monte Carlo Tree Search, learning policy and value function networks for pruning the search tree, expert demonstrations, self play, Tensor Processing Unit

AlphaGo

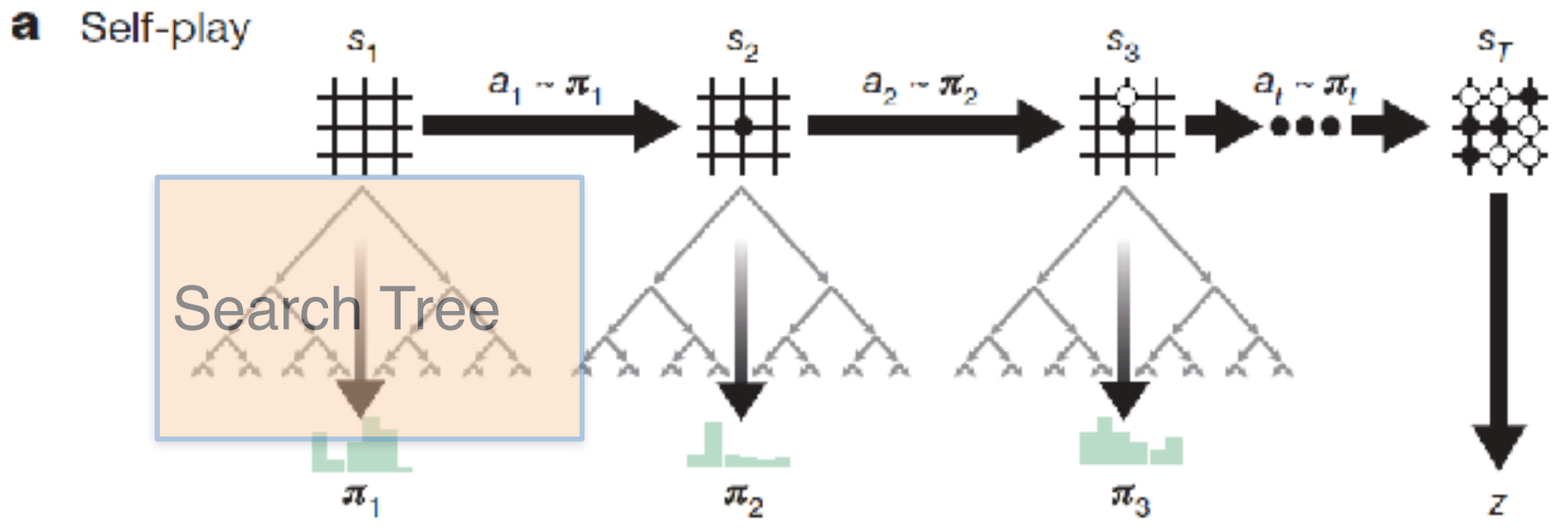


Tensor Processing Unit from Google

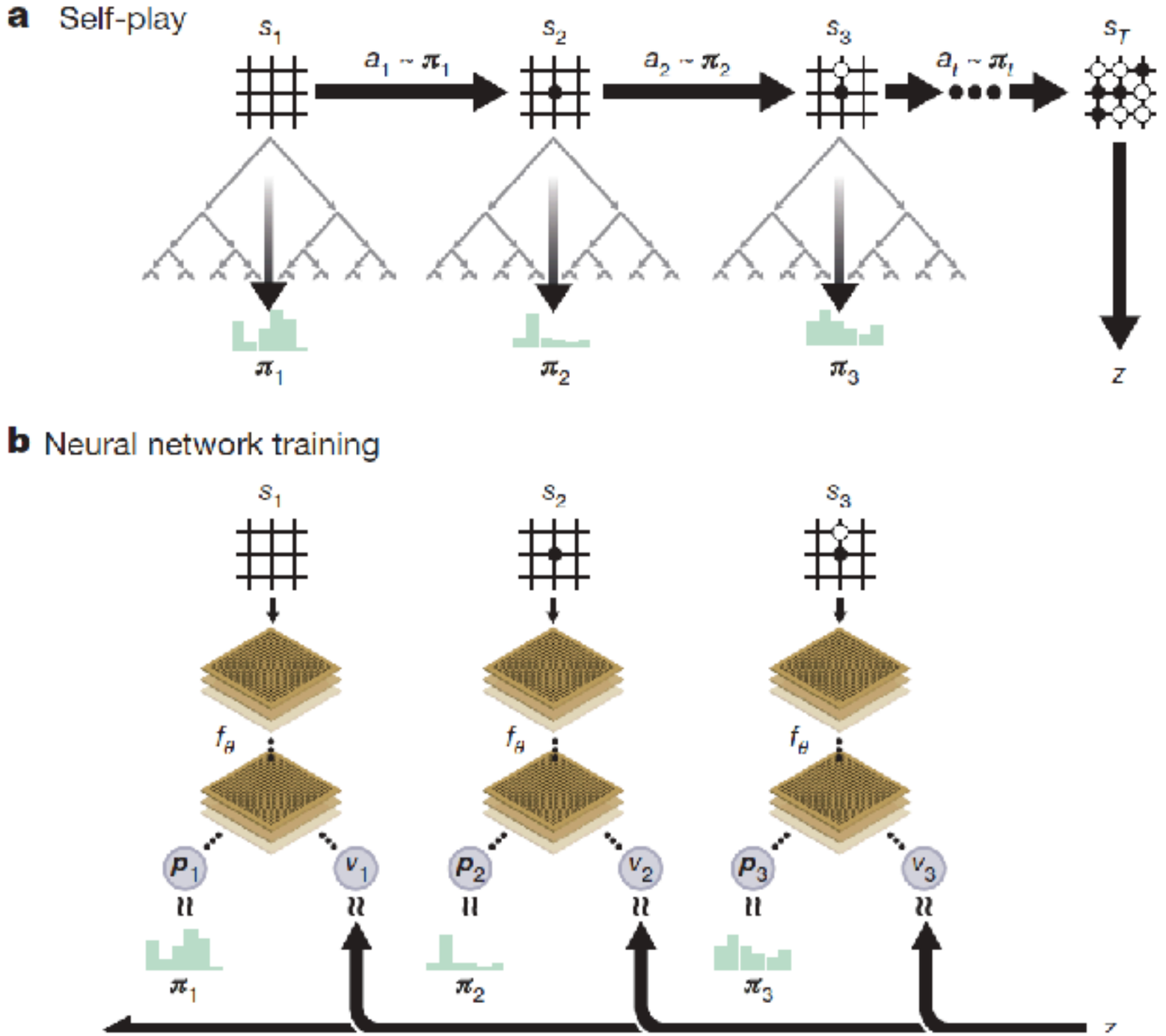
AlphaGoZero

- No human supervision!
- MCTS to select great moves during training and testing!

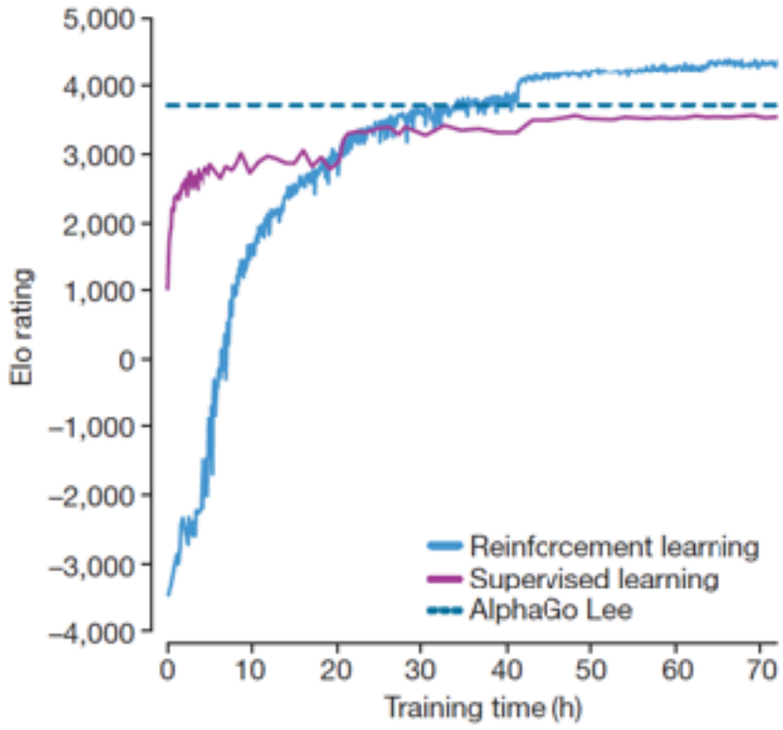
AlphaGoZero

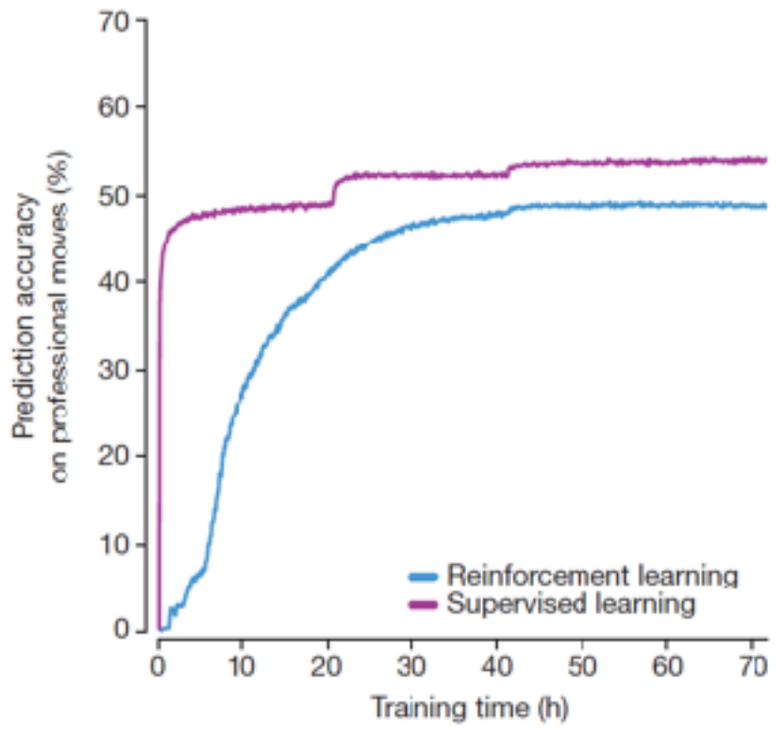


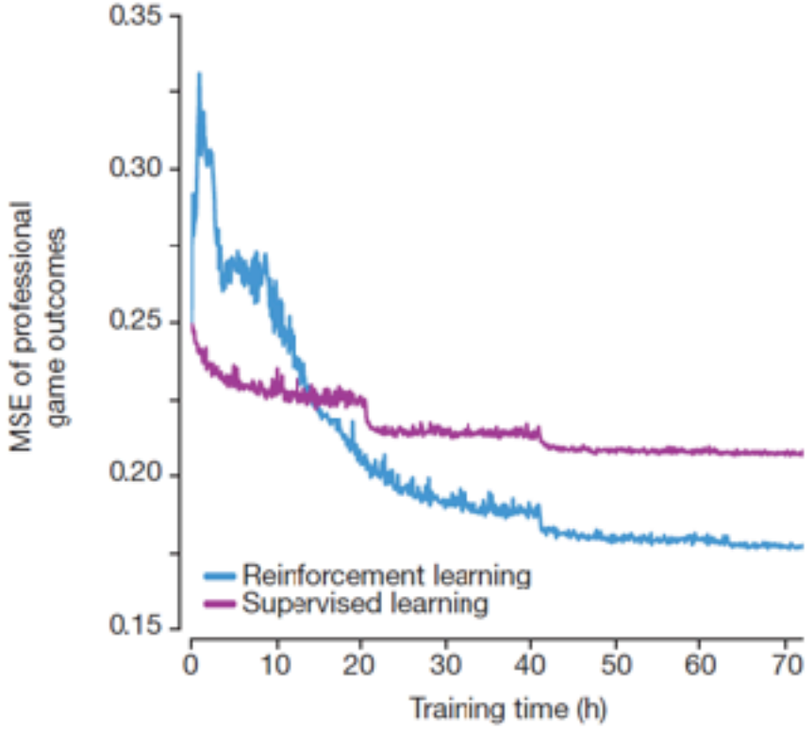
AlphaGoZero



AlphaGoZero







Alpha Go Versus the real world

Beating the world champion is easier than moving the Go stones.

Alpha Go Versus the real world

How the world of Alpha Go is different than the real world?

1. **Known environment** (known entities and dynamics) **Vs Unknown environment** (unknown entities and dynamics).
2. Need for behaviors to **transfer** across environmental variations since the real world is very diverse
3. **Discrete Vs Continuous actions**
4. **One goal Vs many goals**
5. **Rewards automatic VS rewards need themselves to be detected**

Alpha Go Versus the real world

How the world of Alpha Go is different than the real world?

1. **Known environment** (known entities and dynamics) **Vs Unknown environment** (unknown entities and dynamics).
2. Need for behaviors to **transfer** across environmental variations since the real world is very diverse

Alpha Go Versus the real world

How the world of Alpha Go is different than the real world?

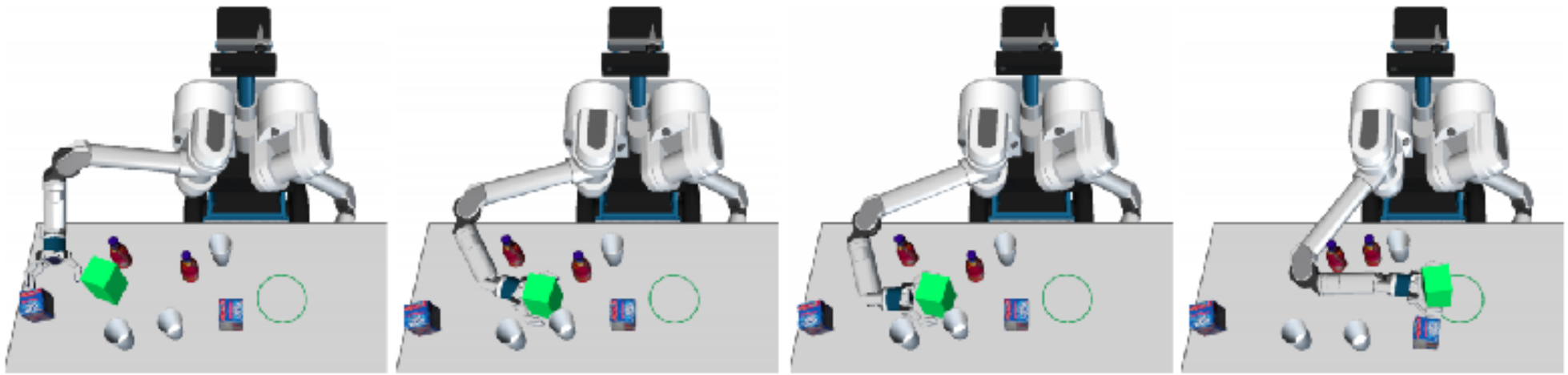
1. **Known environment** (known entities and dynamics) **Vs Unknown environment** (unknown entities and dynamics).
2. Need for behaviors to **transfer** across environmental variations since the real world is very diverse

State estimation: To be able to act you need first to be able to see, detect the objects that you interact with, detect whether you achieved your goal

State estimation

Most works are between two extremes:

- Assuming the world model known (object locations, shapes, physical properties obtain via AR tags or manual tuning), they use planners to search for the action sequence to achieve a desired goal.



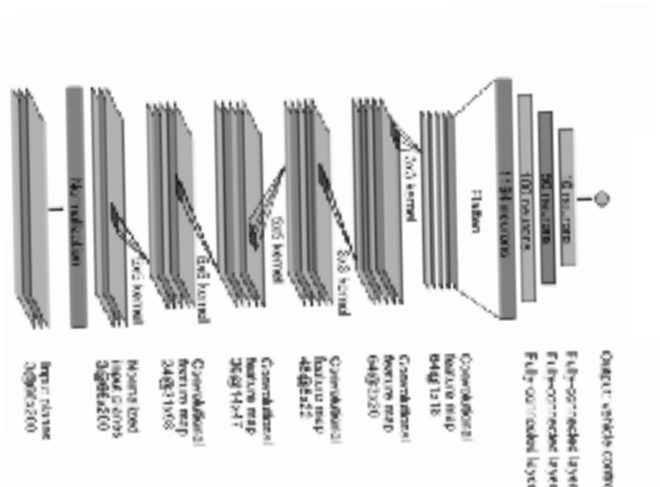
Rearrangement Planning via Heuristic Search, Jennifer E. King, Siddhartha S. Srinivasa

State estimation

Most works are between two extremes:

- Assuming the world model known (object locations, shapes, physical properties obtain via AR tags or manual tuning), they use planners to search for the action sequence to achieve a desired goal.

- Do not attempt to detect any objects and learn to map RGB images directly to actions



Behavior learning is difficult because state estimation is difficult, in other words, because Computer Vision is difficult.

End to End Learning for Self-Driving Cars, NVIDIA

Alpha Go Versus the real world

How the world of Alpha Go is different than the real world?

1. **Known environment** (known entities and dynamics) **Vs Unknown environment** (unknown entities and dynamics).
2. Need for behaviors to **transfer** across environmental variations since the real world is very diverse
3. **Discrete Vs Continuous actions**
4. **One goal Vs many goals**
5. **Rewards automatic VS rewards need themselves to be detected**

Alpha Go Versus the real world

How the world of Alpha Go is different than the real world?

1. **Known environment** (known entities and dynamics) **Vs Unknown environment** (unknown entities and dynamics).
2. Need for behaviors to **transfer** across environmental variations since the real world is very diverse
3. **Discrete Vs Continuous actions** (curriculum learning, progressively add degrees of freedom)
4. **One goal Vs many goals**
5. **Rewards automatic VS rewards need themselves to be detected**

Alpha Go Versus the real world

How the world of Alpha Go is different than the real world?

1. **Known environment** (known entities and dynamics) **Vs Unknown environment** (unknown entities and dynamics).
2. Need for behaviors to **transfer** across environmental variations since the real world is very diverse
3. **Discrete Vs Continuous actions** (curriculum learning, progressively add degrees of freedom)
4. **One goal Vs many goals** (generalized policies parametrized by the goal, Hindsight Experience Replay)
5. **Rewards automatic VS rewards need themselves to be detected**

Alpha Go Versus the real world

How the world of Alpha Go is different than the real world?

1. **Known environment** (known entities and dynamics) **Vs Unknown environment** (unknown entities and dynamics).
2. Need for behaviors to **transfer** across environmental variations since the real world is very diverse
3. **Discrete Vs Continuous actions** (curriculum learning, progressively add degrees of freedom)
4. **One goal Vs many goals** (generalized policies parametrized by the goal, Hindsight Experience Replay)
5. **Rewards automatic VS rewards need themselves to be detected** (learning perceptual rewards, use Computer Vision to detect success)

Alpha Go Versus the real world

Beating the world champion is easier than moving the Go stones.

AI's paradox

Hans Moravec

"it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility"

AI's paradox

Marvin Minsky

"we're more aware of simple processes that don't work well than of complex ones that work flawlessly"

Evolutionary explanation

Hans Moravec

We should expect the difficulty of reverse-engineering any human skill to be roughly proportional to the amount of time that skill has been evolving in animals.
The oldest human skills are largely unconscious and so appear to us to be effortless.
Therefore, we should expect skills that appear effortless to be difficult to reverse-engineer, but skills that require effort may not necessarily be difficult to engineer at all.

What is AI?

Rodney Brooks

intelligence was "best characterized as the things that highly educated male scientists found challenging", such as chess, symbolic integration,
proving mathematical theorems and solving complicated word algebra problems.

What is AI?

Rodney Brooks

intelligence was "best characterized as the things that highly educated male scientists found challenging", such as chess, symbolic integration,
proving mathematical theorems and solving complicated word algebra problems.
"The things that children of four or five years could do effortlessly, such as visually distinguishing between a coffee cup and a chair, or walking around on two legs, or finding their way from their bedroom to the living room were not thought of as activities requiring intelligence.

What is AI?

Rodney Brooks

intelligence was "best characterized as the things that highly educated male scientists found challenging", such as chess, symbolic integration,
proving mathematical theorems and solving complic
"The thi
years could do effortlessly, such as visually distinguishing between a coffee cup and a chair, or walking around on two legs, or finding their way from their bedroom to the living room were not thought of as activities requiring intelligence.

No cognition. Just sensing and action

Learning from Babies

- *Be multi-modal*
- *Be incremental*
- *Be physical*
- *Explore*
- *Be social*
- *Learn a language*

The Development of Embodied Cognition: Six Lessons from Babies
Linda Smith, Michael Gasser