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

- 3 assignments and a project
- Russ will announce those in the next lecture!

Goal of the Course

How to build agents that **learn** behaviors in a **dynamic** world?

as opposed to agents that execute **preprogrammed** behavior in a **static** world…

Behavior: a sequence of actions with a particular goal

Behaviors are 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

Behaviors are Important

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

Daniel Wolpert, nice TED talk

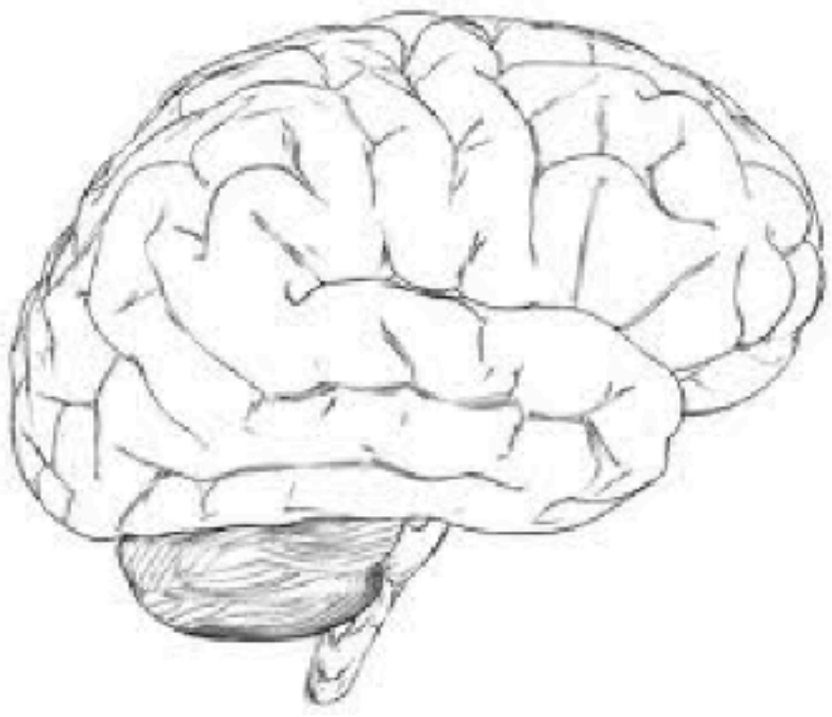
Sea squirts digest their own brain when they decide not to move anymore

Behaviors are Important

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

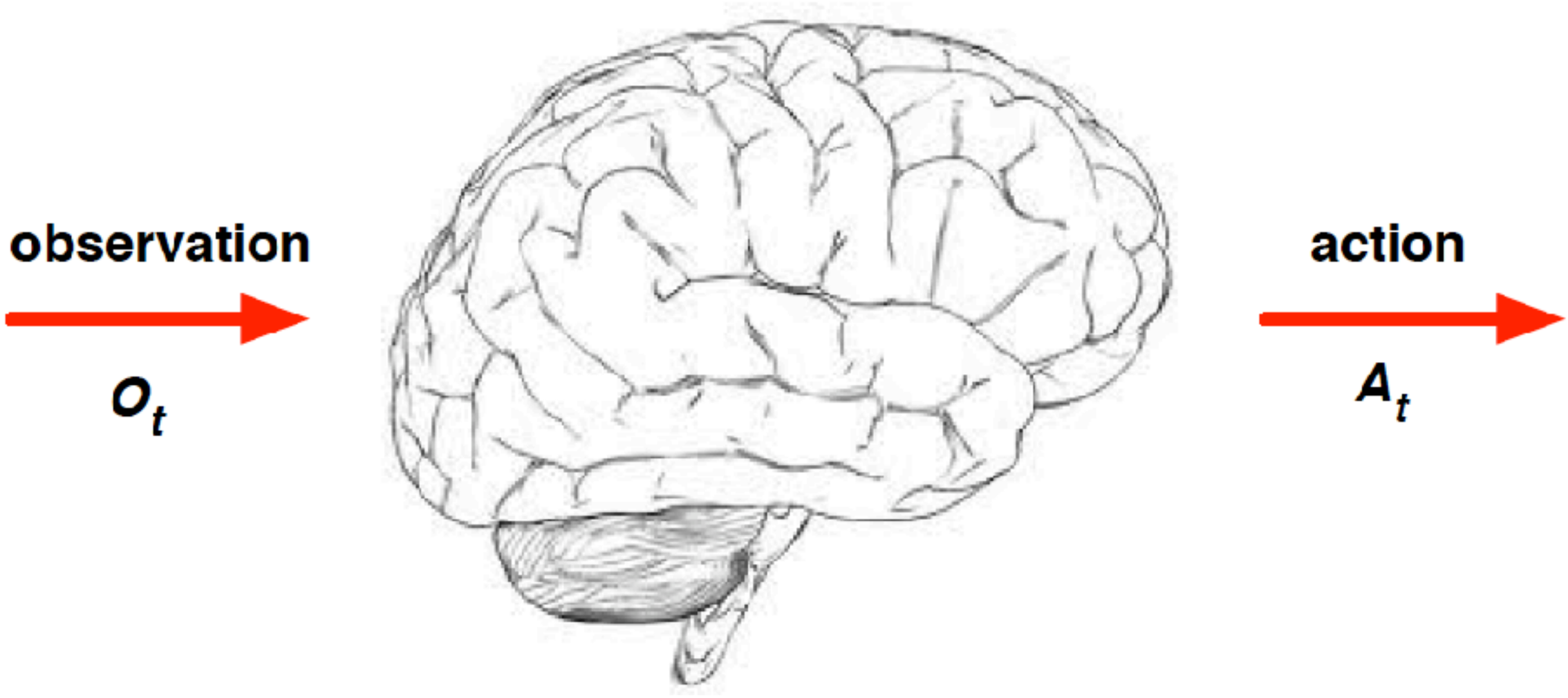
Daniel Wolpert, nice TED talk

Learning behaviors that adapt to a changing environment is considered the hallmark of human intelligence
(though definitions of intelligence are not easy)



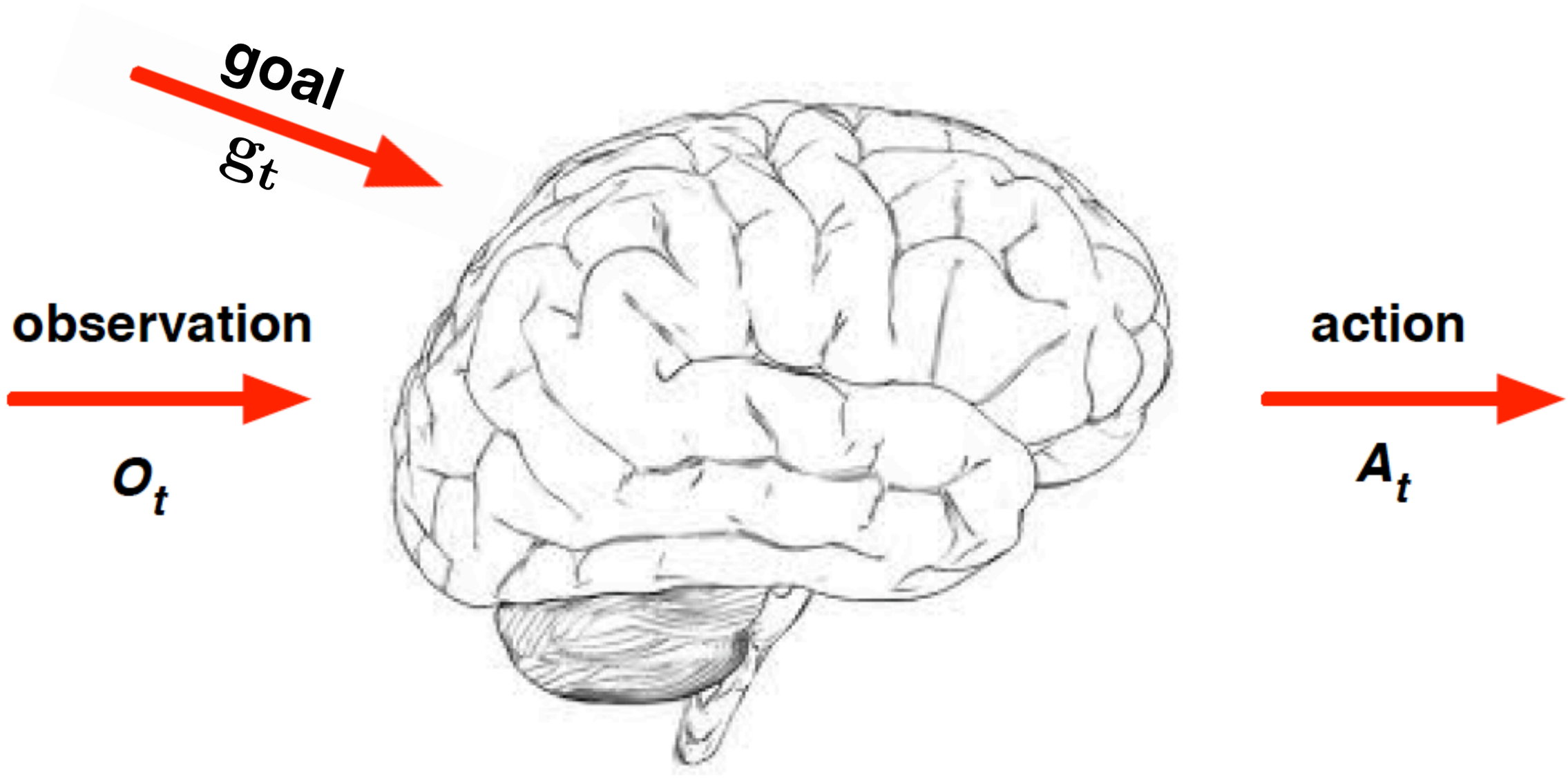
Learning Behaviors

Learning to map sequences of observations to actions



Learning Behaviors

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



Supervision

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

- Rewards: sparse feedback from the environment whether the desired behavior is achieved e.g., game is won, car has not crashed, agent is out of the maze etc.
- Demonstrations: experts demonstrate the desired behavior, e.g. by kinesthetic teaching, teleoperation, or through visual imitation (e.g., instructional youtube videos)
- Specifications/Attributes of good behavior: e.g., for driving such attributes would be respect the lane, keep adequate distance from the front car, etc. *DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving*, Chen at al.

Behavior: High Jump

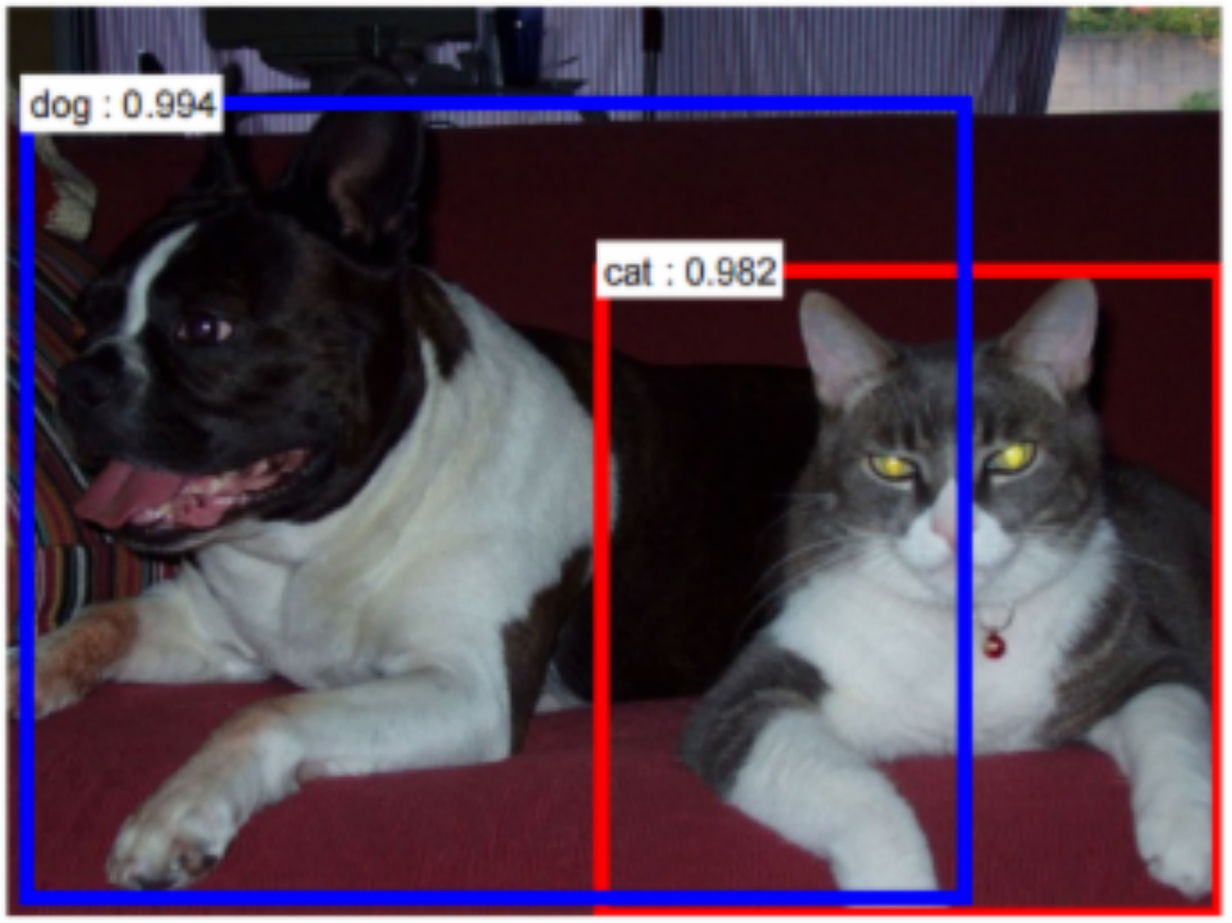
scissors

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 Behaviors

How learning behaviors is different than other machine learning paradigms, e.g., learning to detect objects in images?



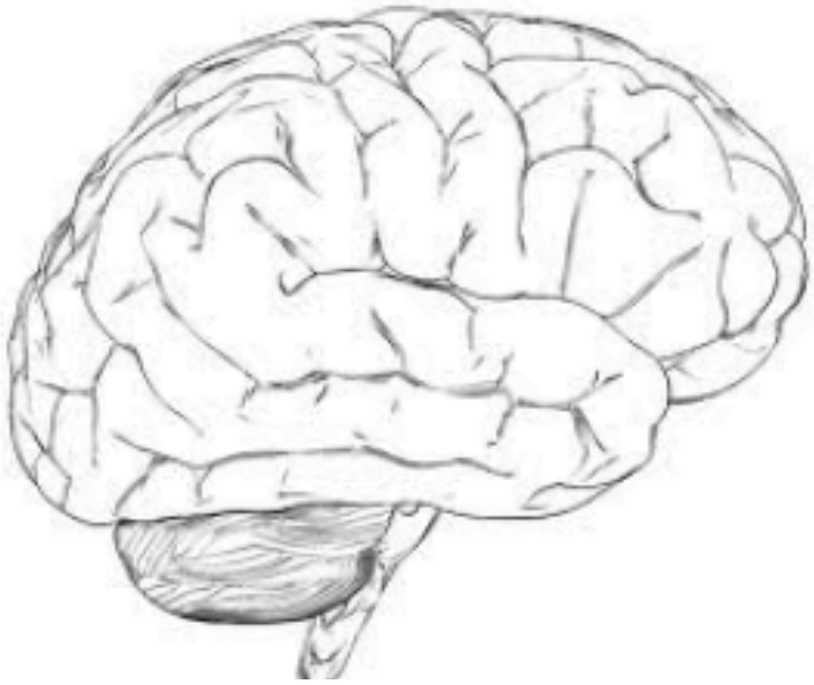
Learning Behaviors

How learning behaviors is different than other machine learning paradigms?

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

observation

O_t



action

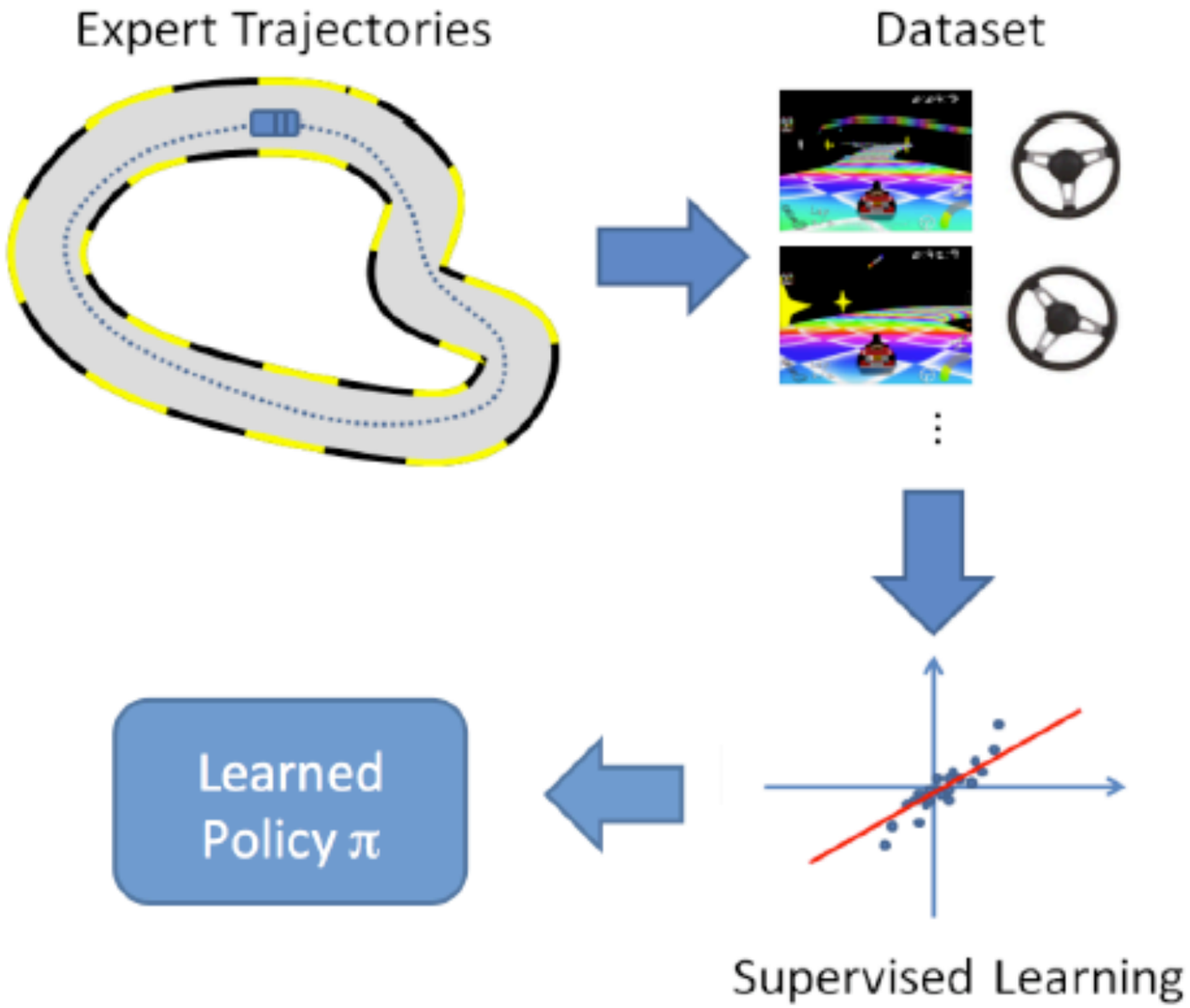
A_t

Learning Behaviors

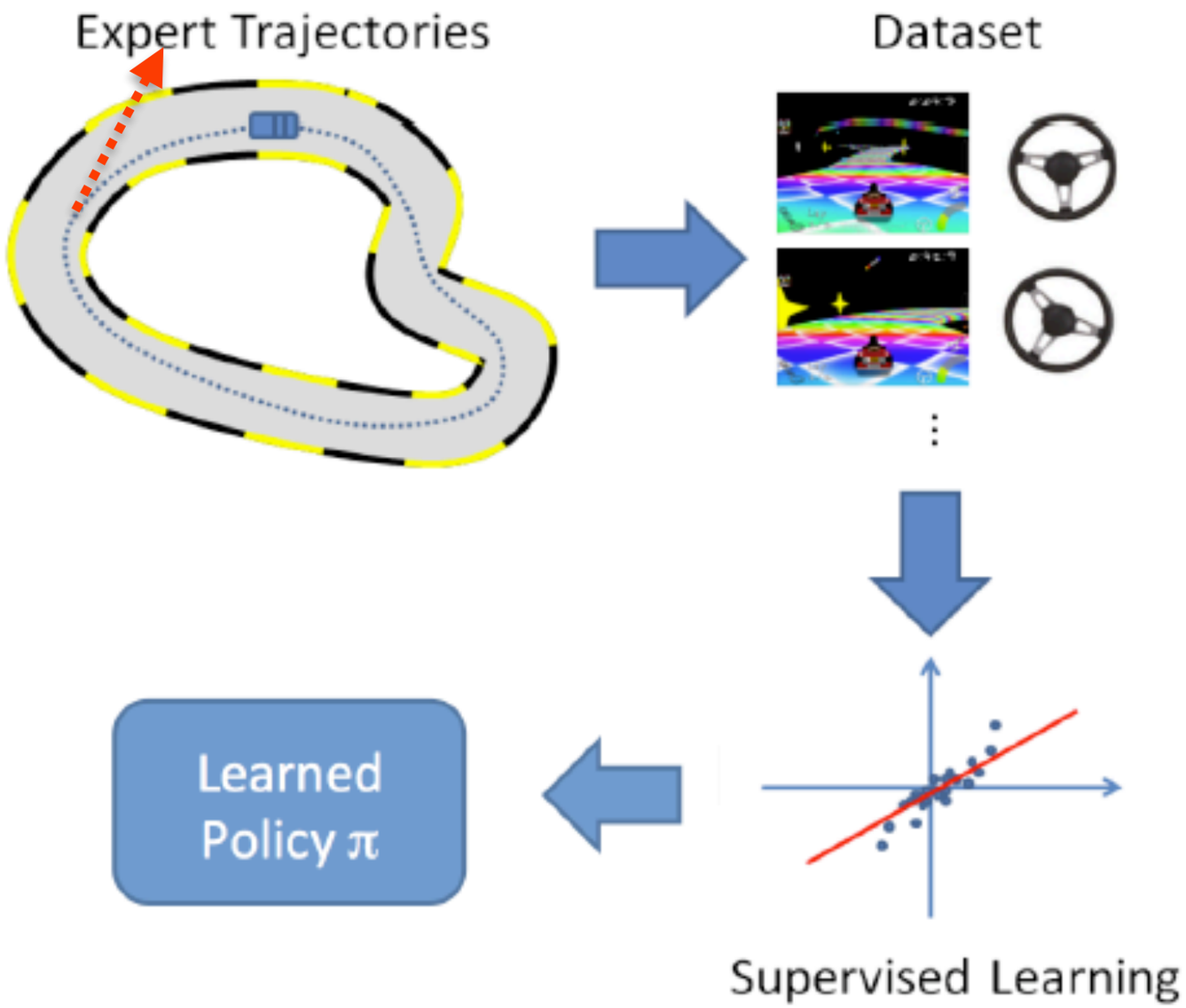
How learning behaviors is different than other machine learning paradigms?

- 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

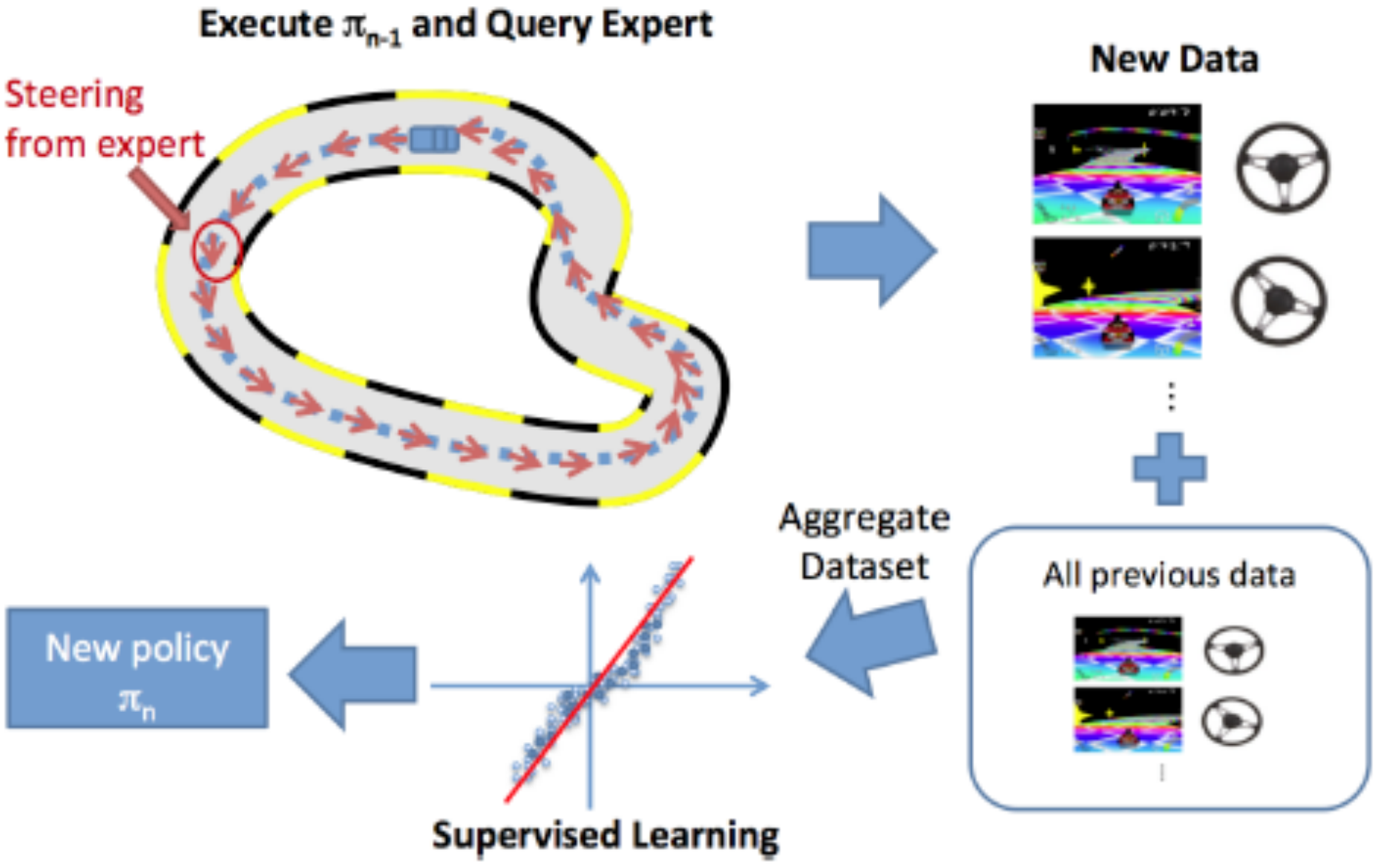
Learning to Drive a Car: Supervised Learning



Learning to Drive a Car: Supervised Learning

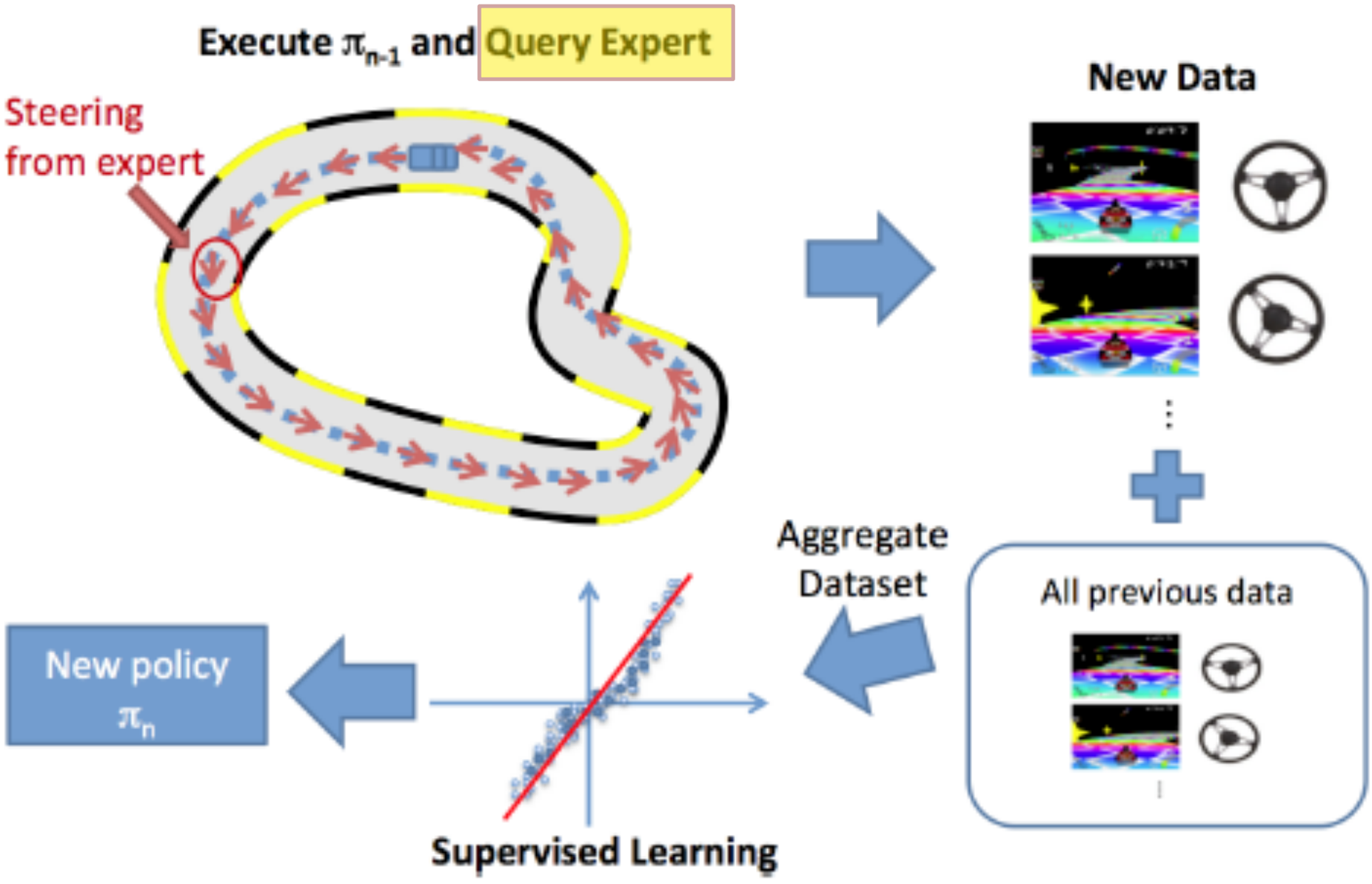


Learning to Race a Car : Interactive learning-DAGGer



Learning to Race a Car : Interactive learning-DAGGer

This assumes you can actively access an expert during training!

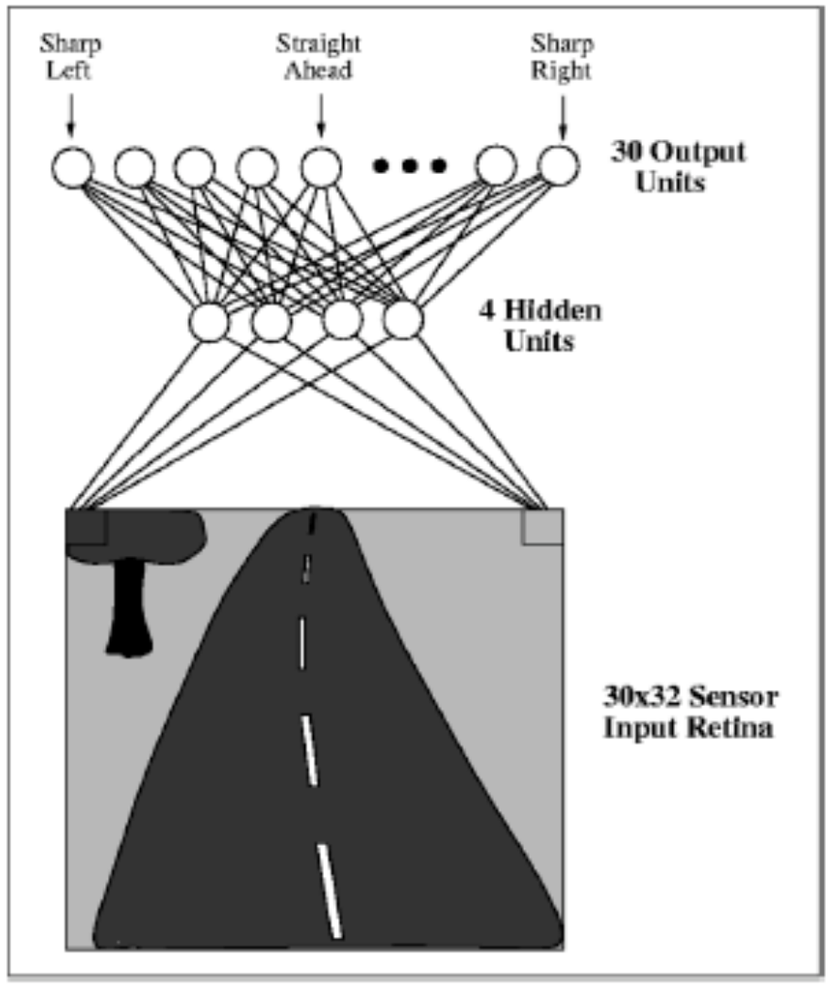


A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning
Stephane Ross, Geoffrey J. Gordon, J. Andrew Bagnell

Learning to Drive a Car: Supervised Learning

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

Learning Behaviors

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

Learning Behaviors

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 Behaviors

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 Behaviors

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 buy many robots

Supersizing Self-Supervision

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

Google's Robot Farm

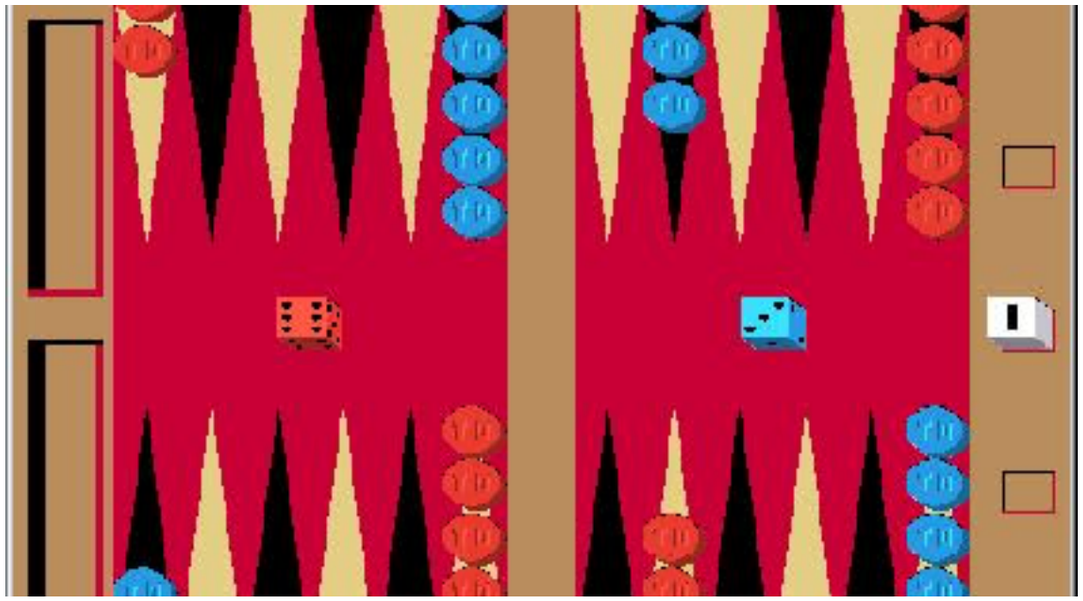
Successes of behavior learning

Backgammon

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

Backgammon

TD-Gammon

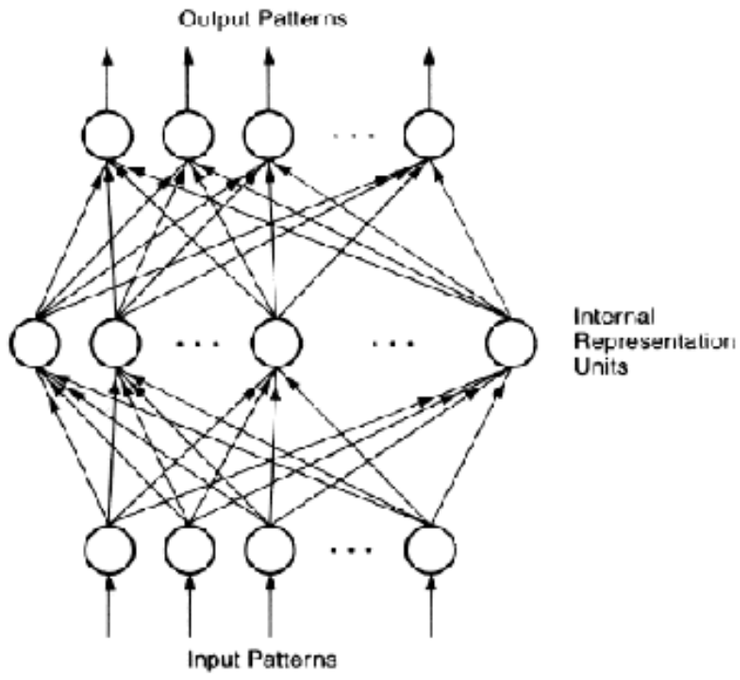


Developed by Gerarl Tesauro in 1992 in IBM’s research center

Neuro-Gammon

Backgammon

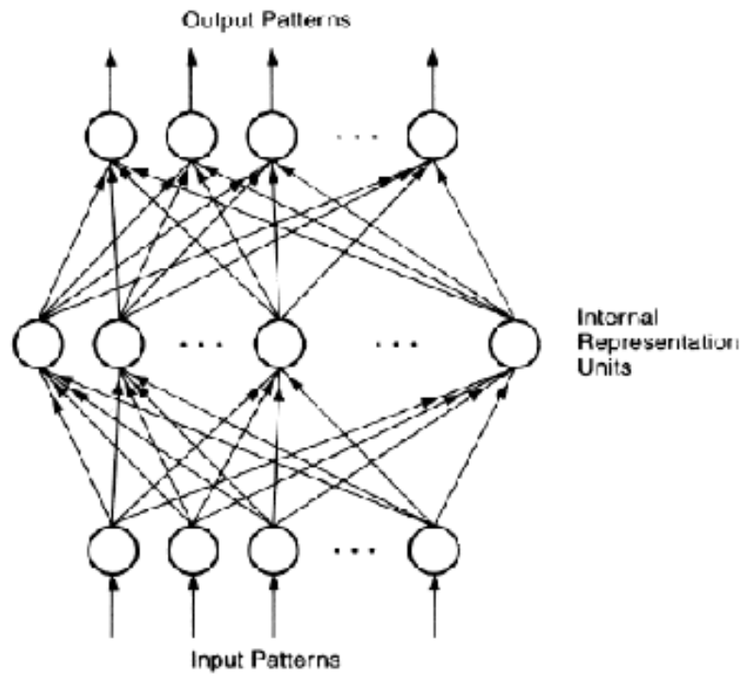
TD-Gammon



Temporal Difference learning

Developed by Gerarl 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
Using features from Neuro-gammon it beat the world's champions

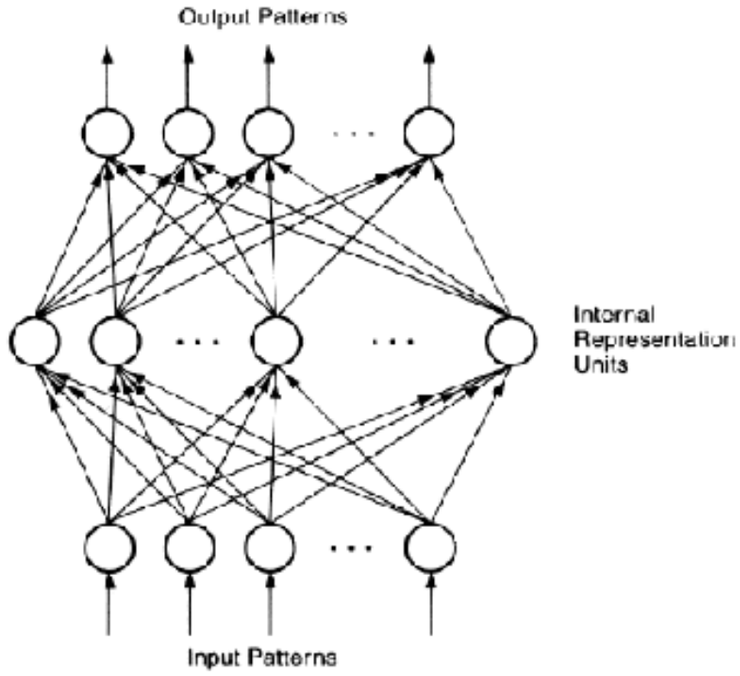
Neuro-Gammon



Learning from human experts, supervised learning

Backgammon

TD-Gammon



Temporal Difference learning

Developed by Gerarl 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
Using features from Neuro-gammon it beat the world's champions

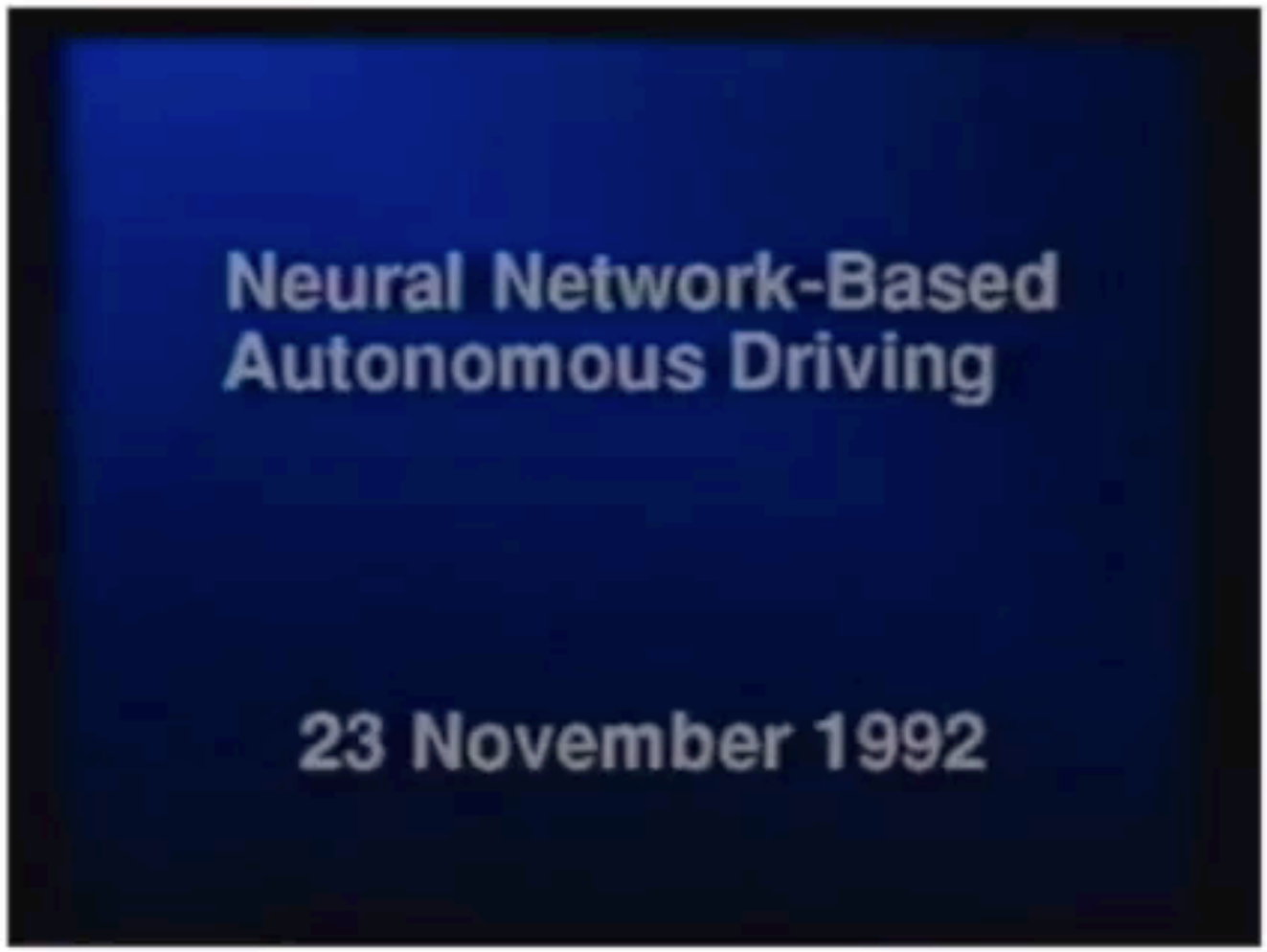
There is no question that its positional judgement is far better than mine. Its technique is less than perfect is such things as building up a board without opposing contact when the human can often come up with a better play by calculating it out.
Kit Woolsey

Locomotion

Optimization and learning for rough terrain legged locomotion, Zucker et al.

Self-Driving Cars

Self-Driving Cars



[Courtesy of Dean Pomerleau]

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

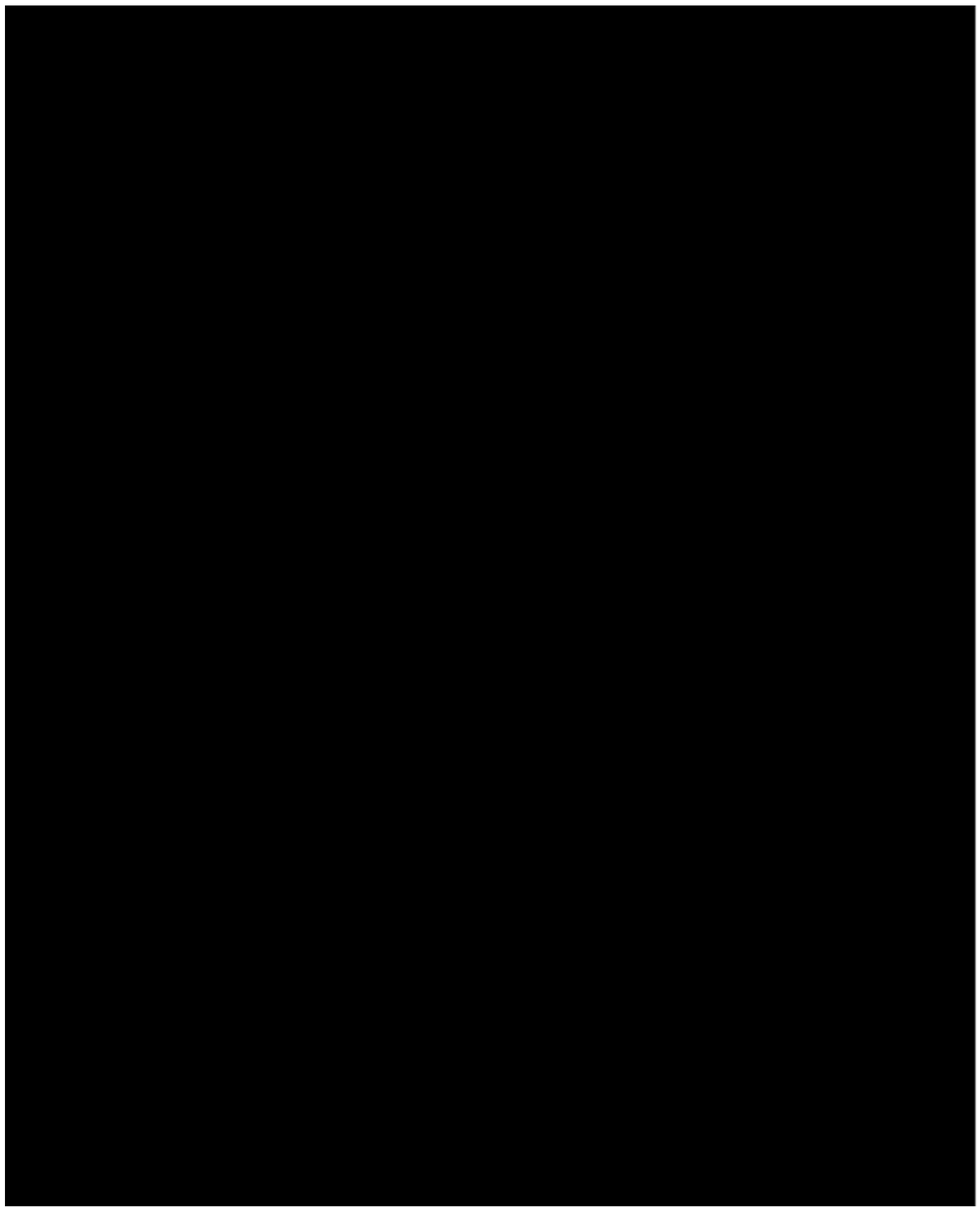
ALVINN (Autonomous Land Vehicle In a Neural Network), *Efficient Training of Artificial Neural Networks for Autonomous Navigation*, Pomerleau 1991

Self-Driving Cars

- P

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

Atari



Deep Mind 2014+

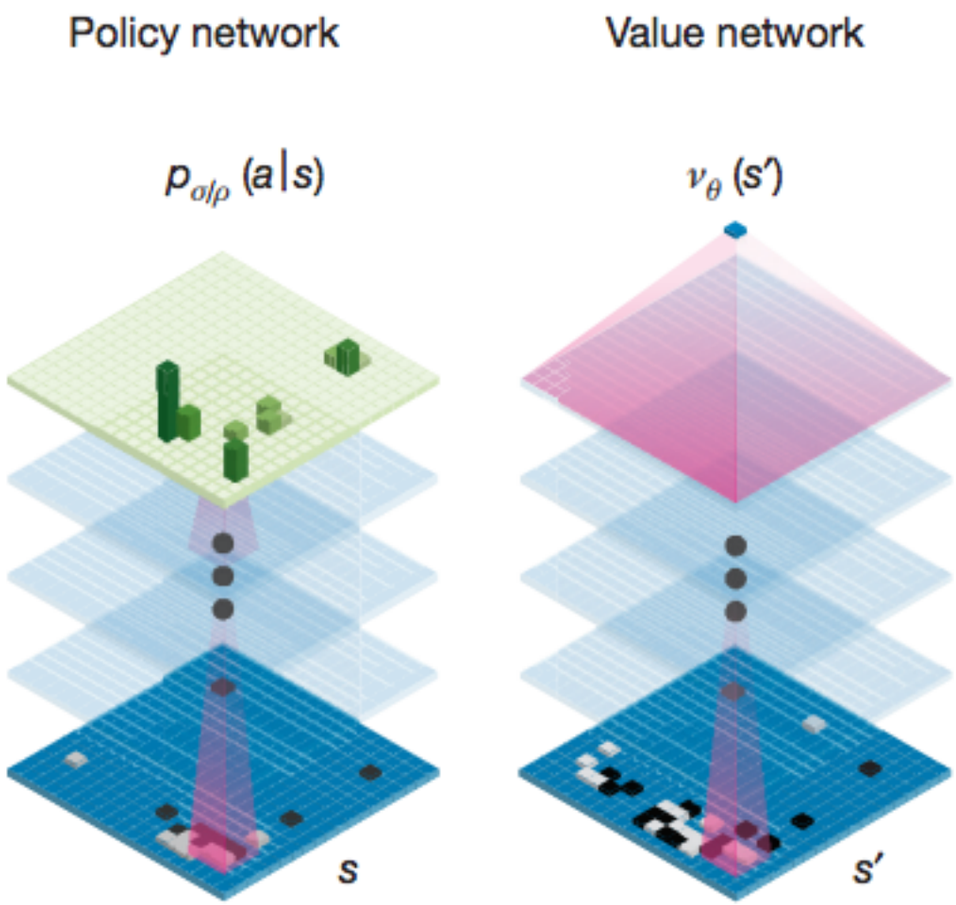
Deep Q learning

GO

AlphaGo

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

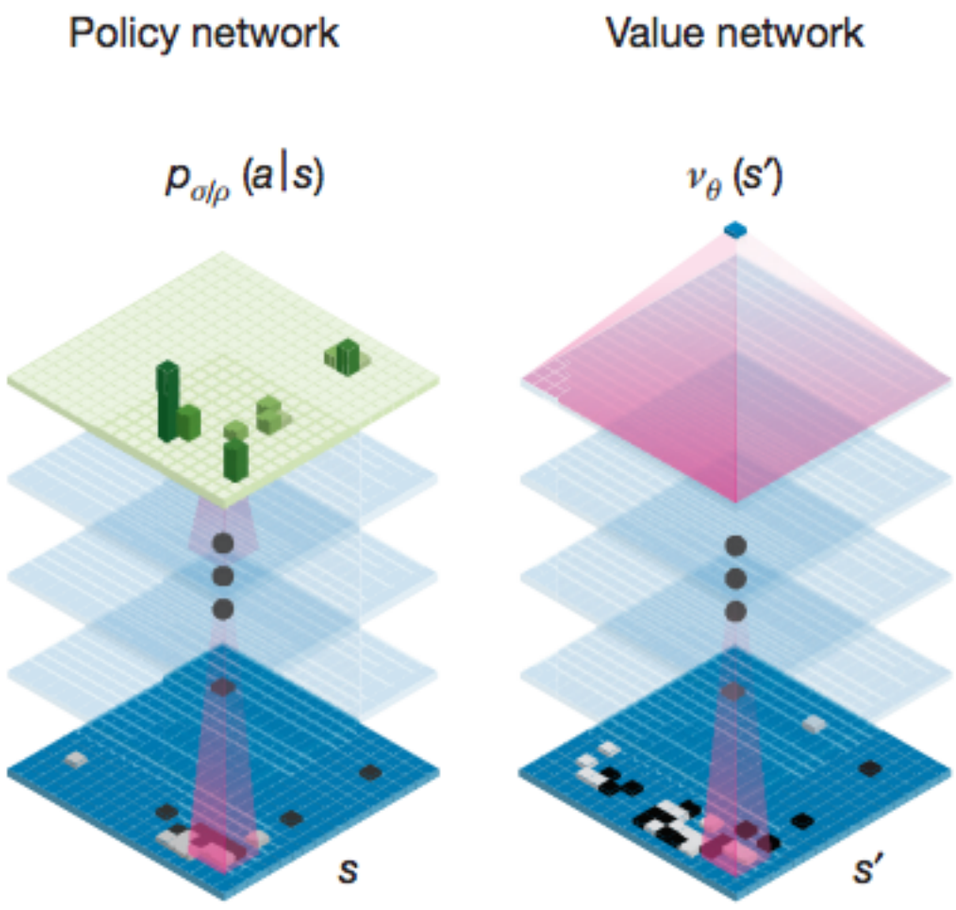
AlphaGo



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

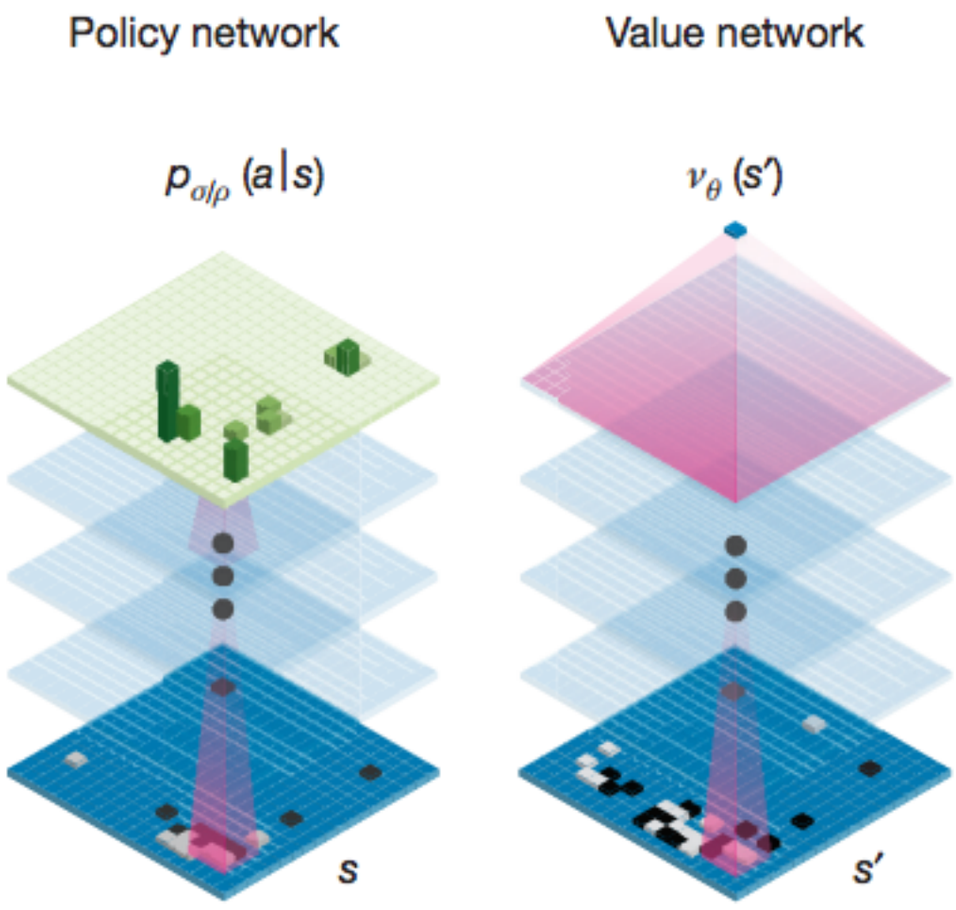


Policy net trained to mimic expert moves, and then fine-tuned using self-play

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.

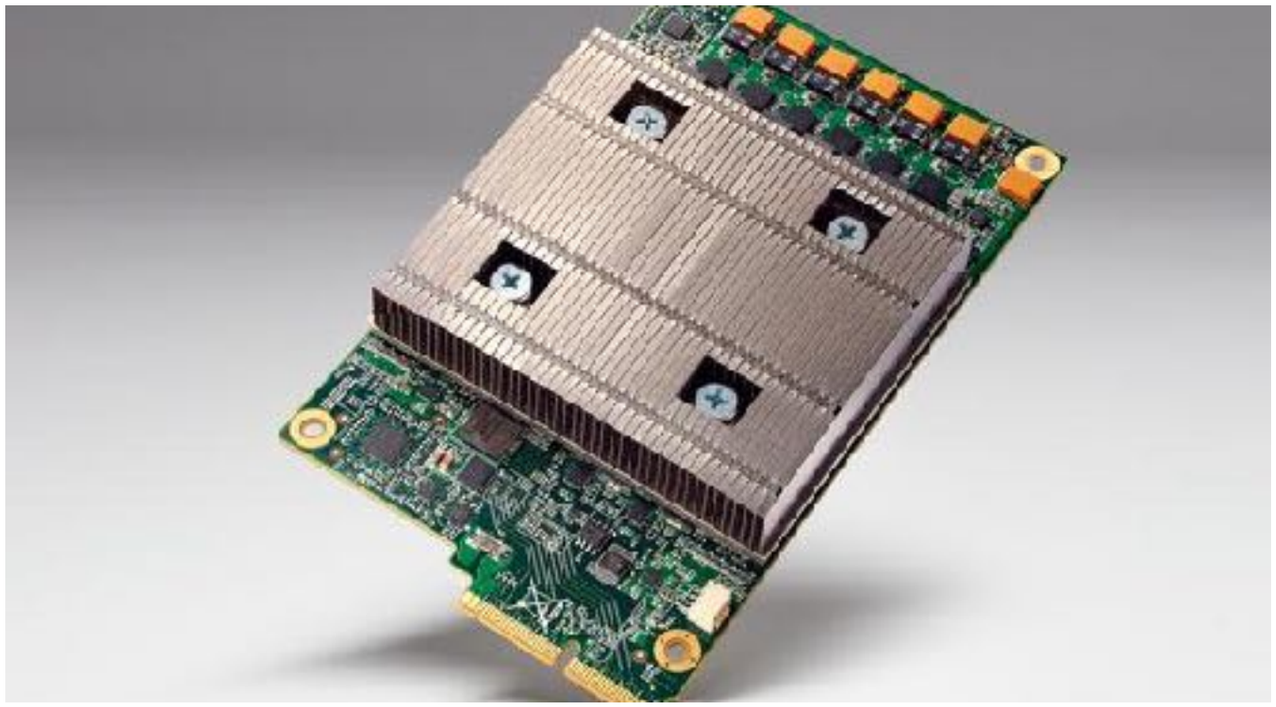
At test time, policy and value nets guide a MCTS to select stronger moves by deep look ahead.

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

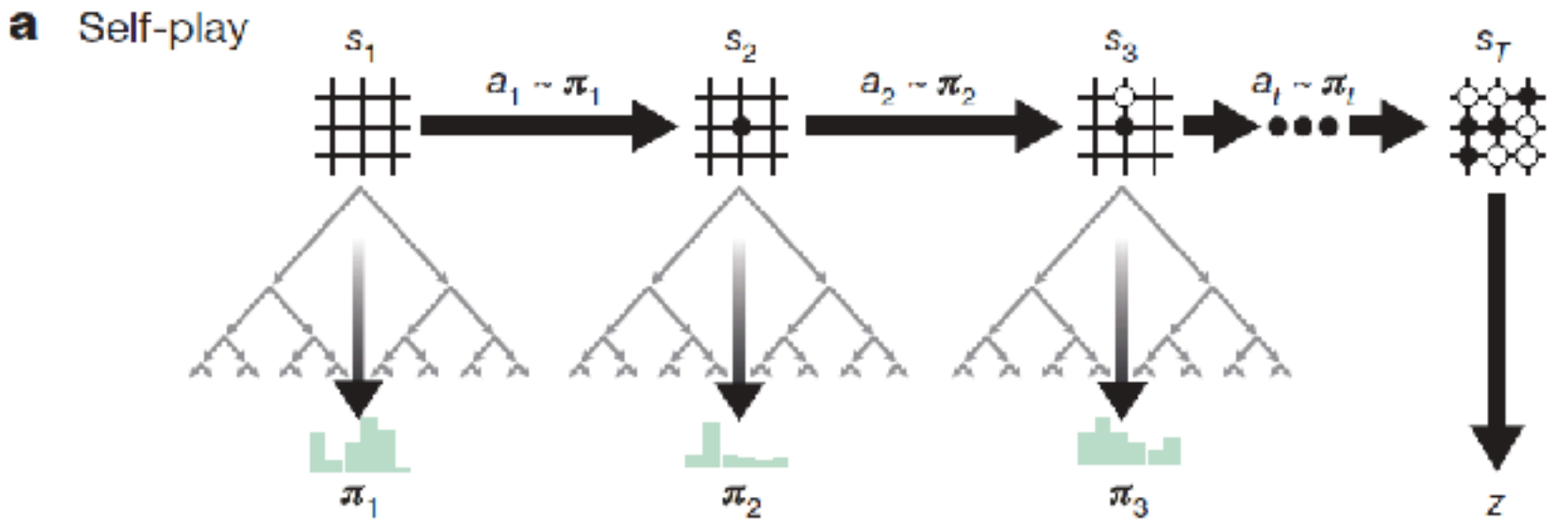


Tensor Processing Unit from Google

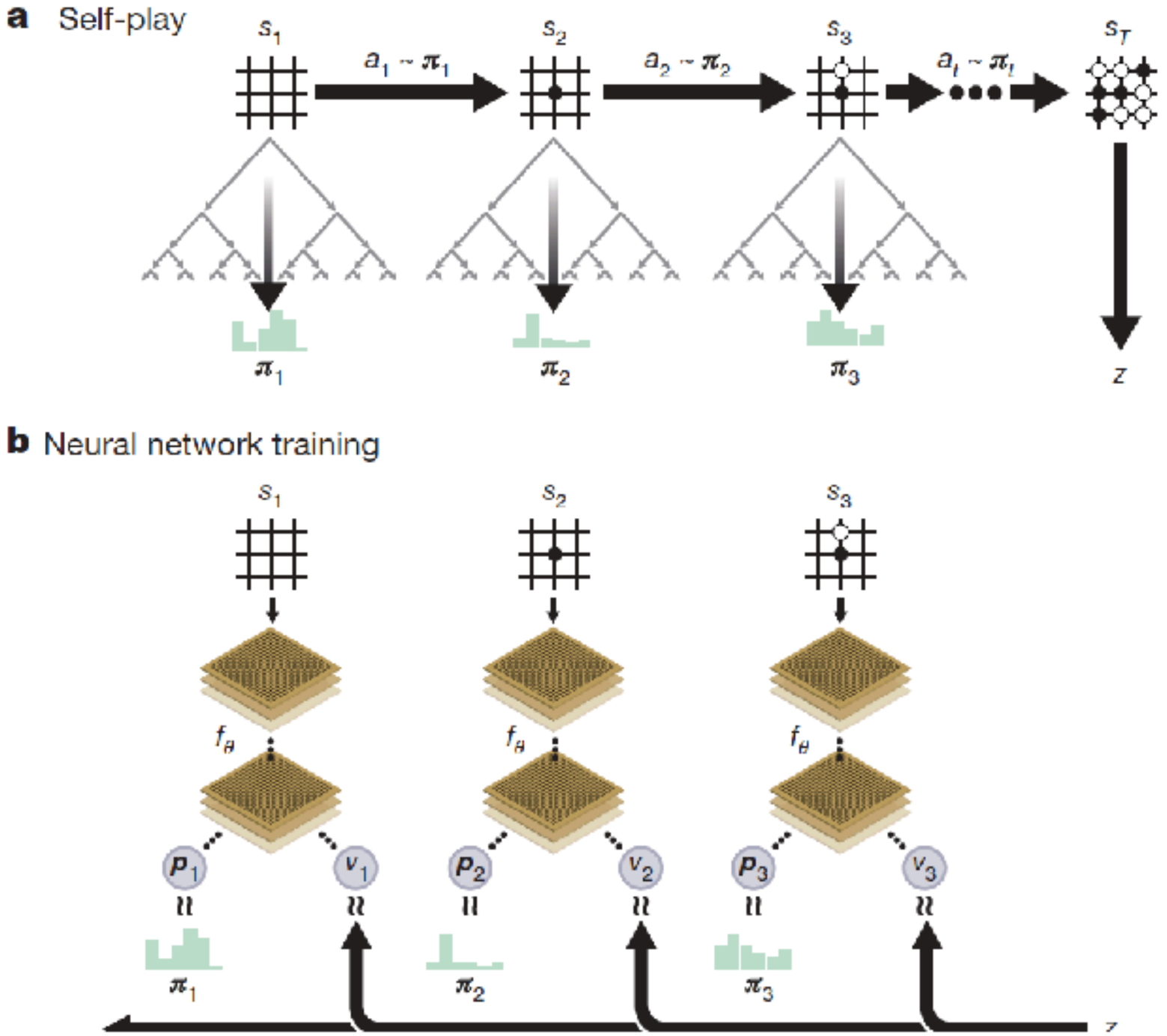
AlphaGoZero

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

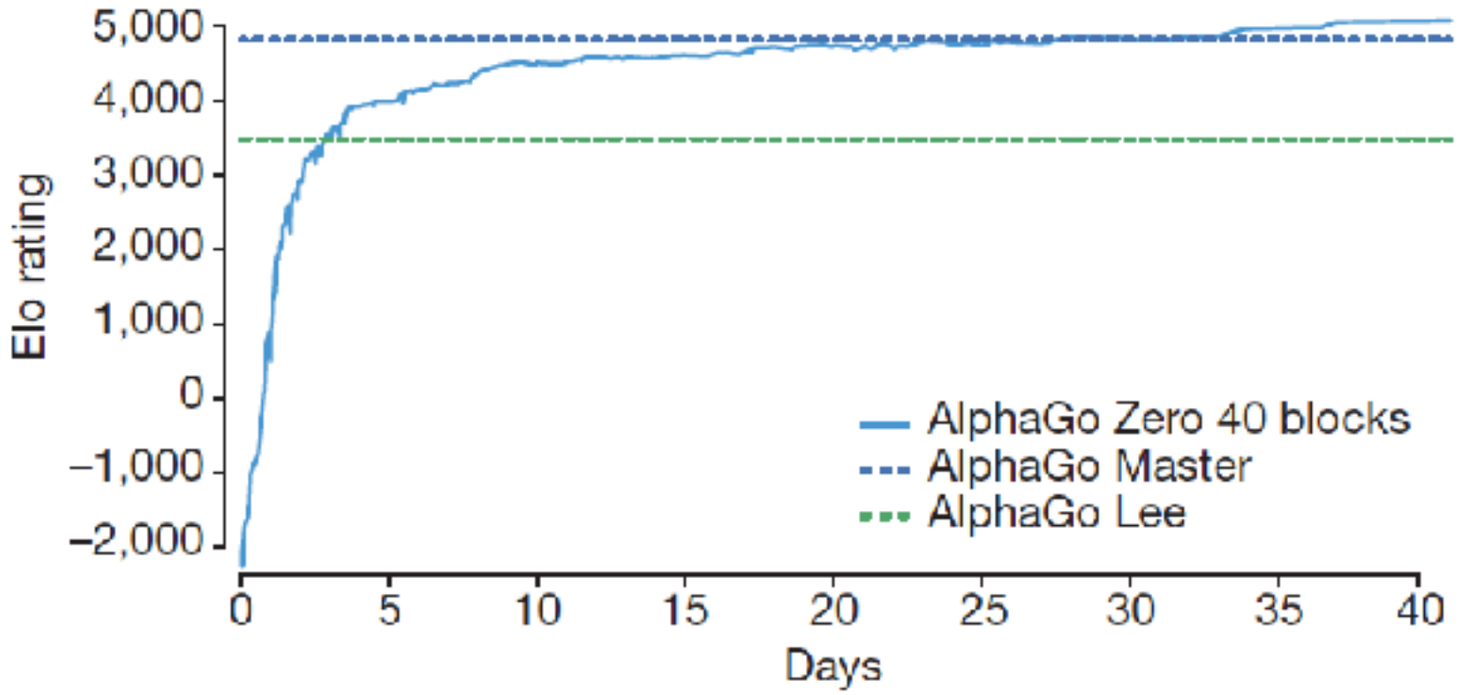
AlphaGoZero

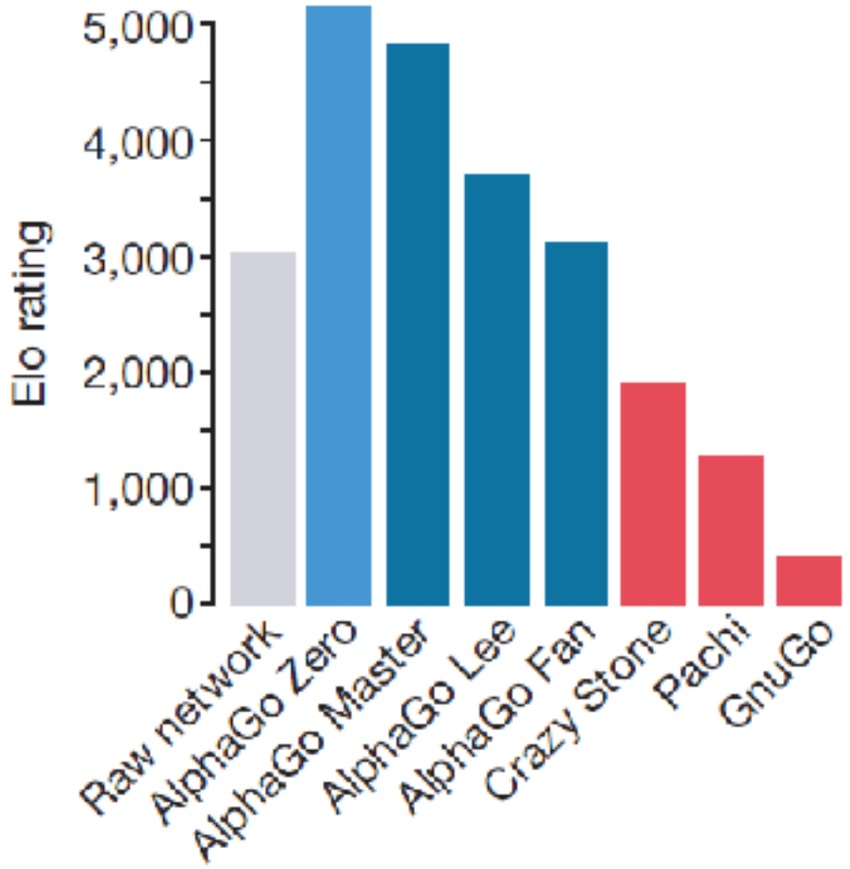


AlphaGoZero

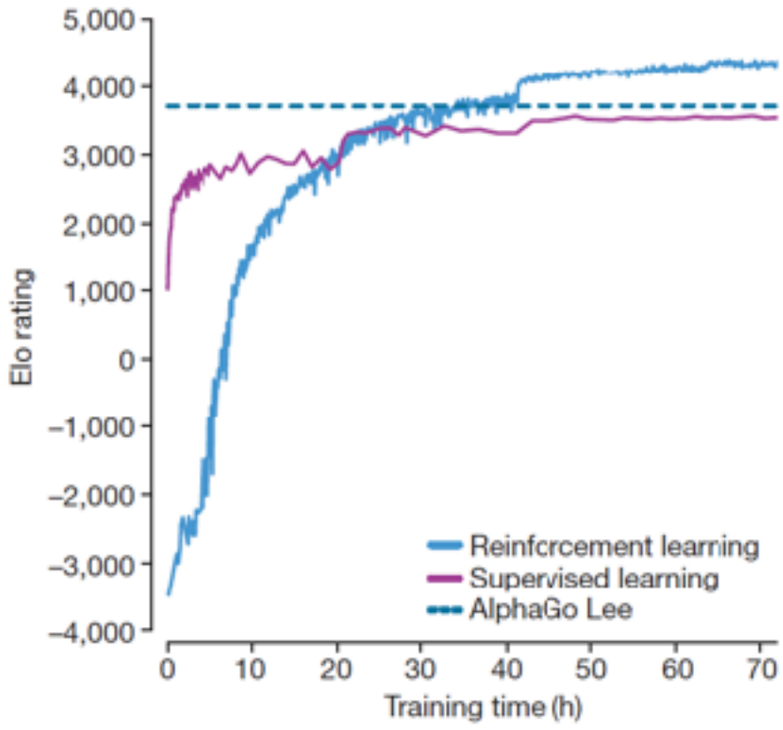


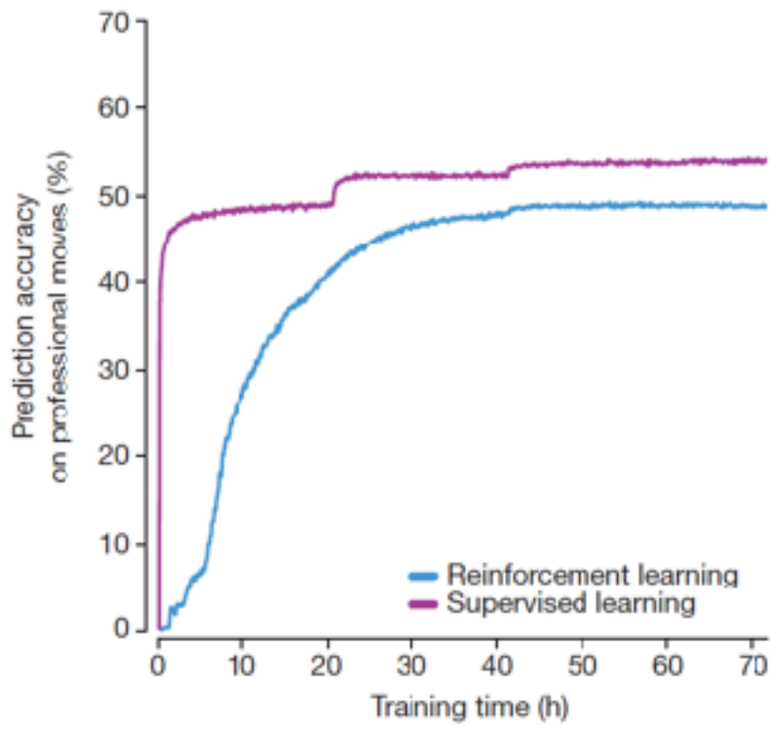
AlphaGoZero

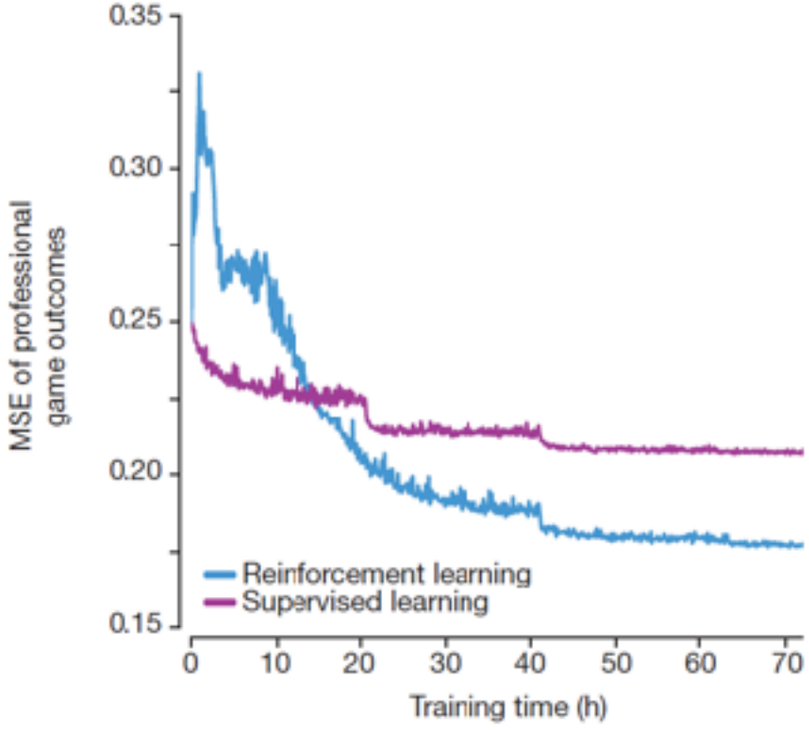




AlphaGoZero







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?

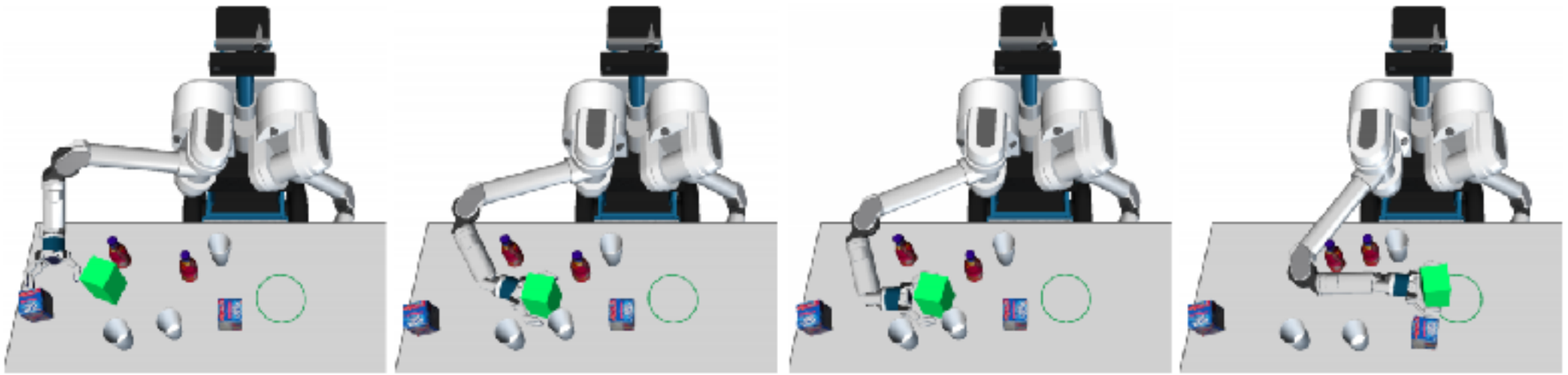
1. **Known environment** (known entities and dynamics) **Vs Unknown environment** (unknown entities and dynamics).
2. Need for behaviors to **transfer** across environmental variations since the real world is very diverse

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.



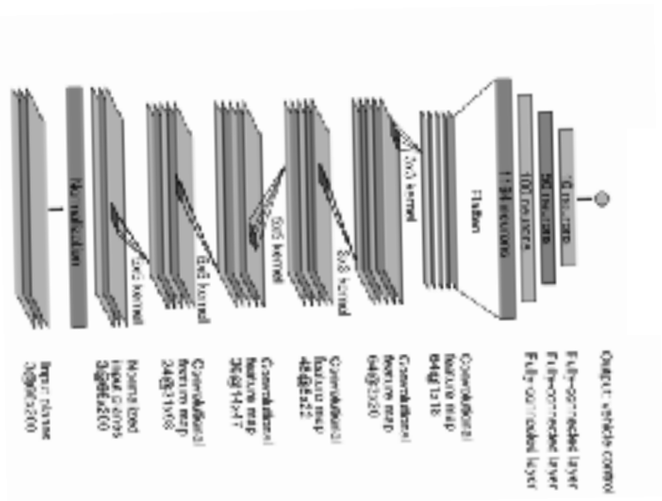
Rearrangement Planning via Heuristic Search, Jennifer E. King, Siddhartha S. Srinivasa

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



End to End Learning for Self-Driving Cars, NVIDIA

State estimation

Recent works have shown that to be able to transfer behaviors across environment variations, factorization of the world state in terms of entities, their attributes and their dynamics are important!

Schema Networks: Zero-shot Transfer with a Generative Causal Model of Intuitive Physics

Ken Kansky Tom Silver David A. Mély Mohamed Eldawy Miguel Lázaro-Gredilla Xinghua Lou
Nimrod Dorfman Szymon Sidor Scott Phoenix Dileep George

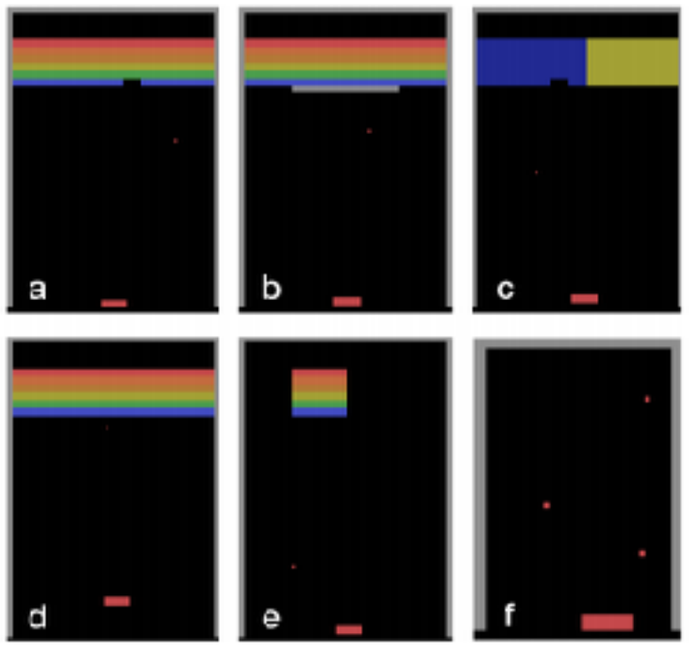


Figure 1. Variations of Breakout. From top left: standard version, middle wall, half negative bricks, offset paddle, random target, and juggling. After training on the standard version, Schema Networks are able to generalize to the other variations without any additional training.

State estimation

Recent works have shown that to be able to transfer behaviors across environment variations, factorization of the world state in terms of entities, their attributes and their dynamics are important!

Schema Networks: Zero-shot Transfer with a Generative Causal Model of Intuitive Physics

Ken Kansky Tom Silver David A. Mély Mohamed Eldawy Miguel Lázaro-Gredilla Xinghua Lou
Nimrod Dorfman Szymon Sidor Scott Phoenix Dileep George

They assume perfect entity/attribute detection and attempt to learn the dynamics of the game.

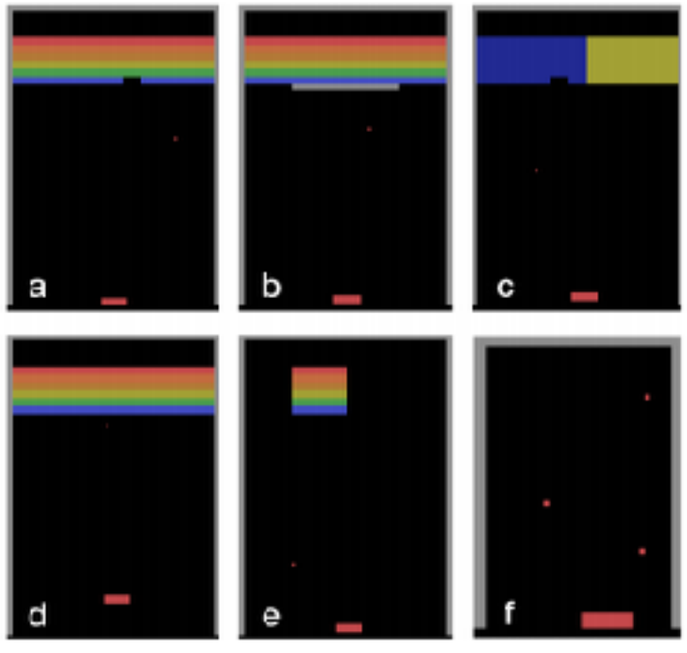


Figure 1. Variations of Breakout. From top left: standard version, middle wall, half negative bricks, offset paddle, random target, and juggling. After training on the standard version, Schema Networks are able to generalize to the other variations without any additional training.

State estimation

Recent works have shown that to be able to transfer behaviors across environment variations, factorization of the world state in terms of entities, their attributes and their dynamics are important!

Schema Networks: Zero-shot Transfer with a Generative Causal Model of Intuitive Physics

Ken Kansky Tom Silver David A. Mély Mohamed Eldawy Miguel Lázaro-Gredilla Xinghua Lou
Nimrod Dorfman Szymon Sidor Scott Phoenix Dileep George

They assume perfect entity/attribute detection and attempt to learn the dynamics of the game. They infer actions by chaining dynamics forward to achieve desired goals.

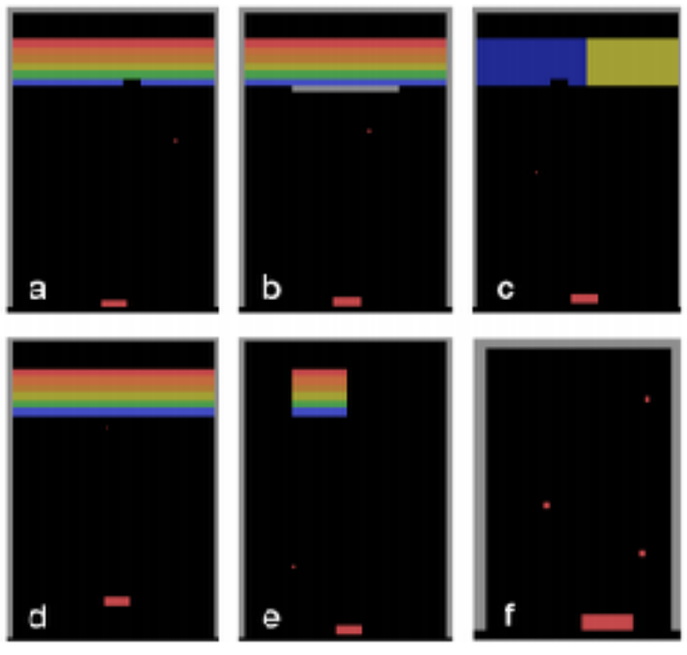
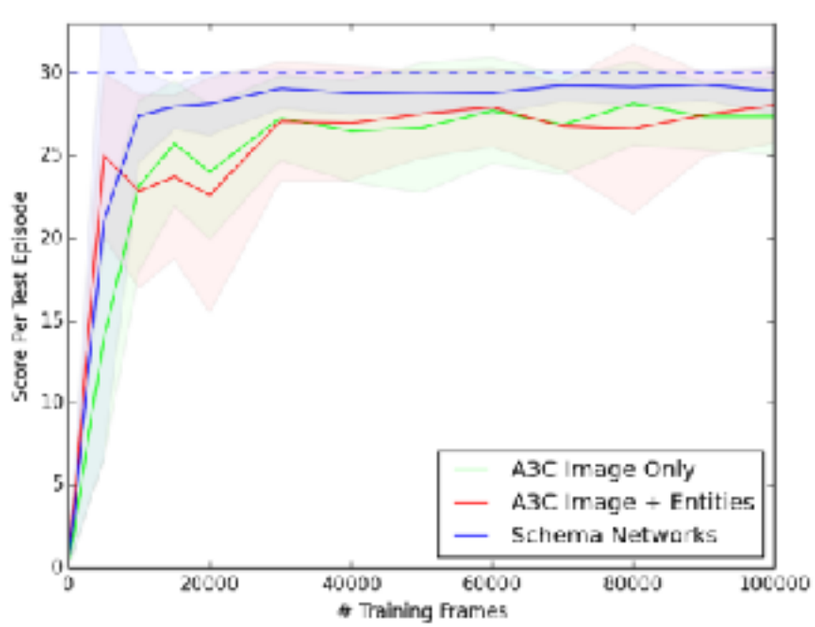
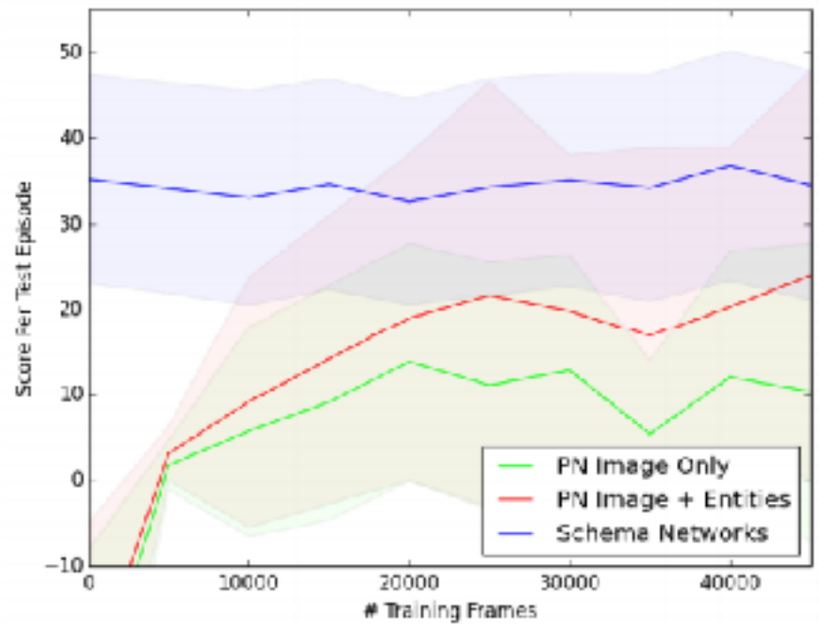


Figure 1. Variations of Breakout. From top left: standard version, middle wall, half negative bricks, offset paddle, random target, and juggling. After training on the standard version, Schema Networks are able to generalize to the other variations without any additional training.

State estimation

Recent works have shown that to be able to transfer behaviors across environment variations, factorization of the world state in terms of entities, their attributes and their dynamics are important!

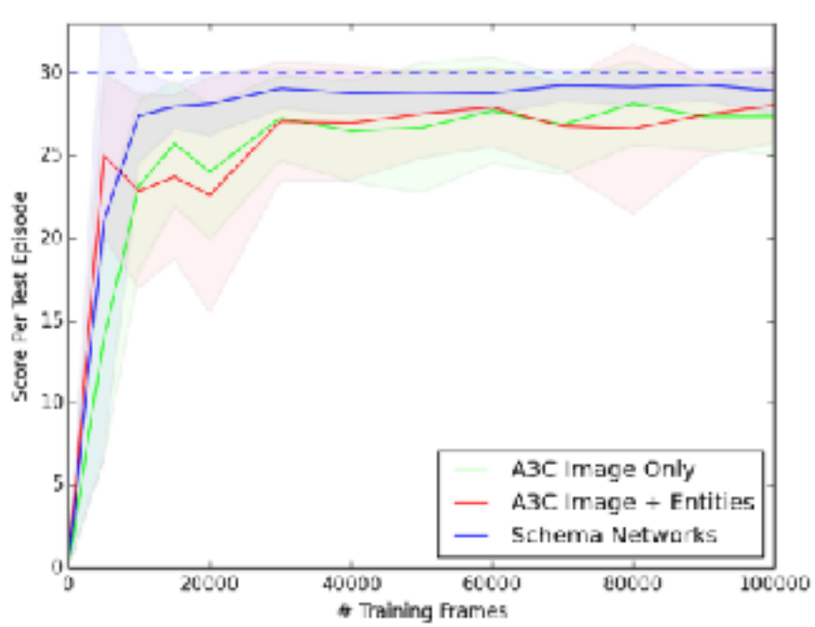


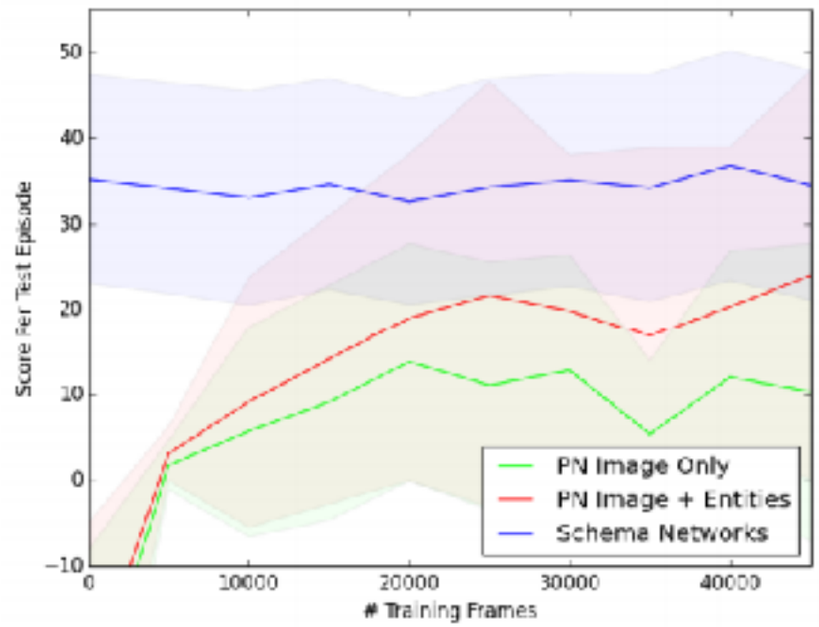


Schema Nets can handle the environmental variation, since dynamics do not change (only the entities). Thus, structured state representations in terms of objects, attributes are important.

State estimation

Recent works have shown that to be able to transfer behaviors across environment variations, factorization of the world state in terms of entities, their attributes and their dynamics are important!





Schema Nets can handle the environmental variation, since dynamics do not change (only the entities). Thus, structured state representations in terms of objects, attributes are important.

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

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)

Learning Behaviors in the Real World

- *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