Deep Reinforcement Learning

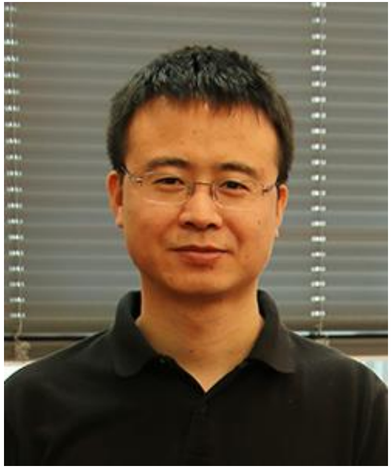
Lecture 1: Introduction

Instructor: Chongjie Zhang

Tsinghua University

Course Staff

Instructor



Chongjie Zhang

TAs

Tonghan Wang

Jianhao Wang

Logistics

- Communication
  - Announcements on WebLearning or WebChat
  - Discussions on WebLearning or WebChat
- Weekly recitations
  - Recitations are optional, for questions about lectures or problem sets.
  - Times: poll posted on the first homework
- Office hours:
  - Instructor: 5-6pm Wed, MMW S-221

DRL-2020

该二维码7天内(2月22日前)有效，重新进入将更新

Course Information

- Textbook: Not required, but for students who want to read more we recommend:
  - Sutton & Barto, Introduction to RL, 2nd Ed. (online)
  - Russell & Norvig, AI: A Modern Approach, 3rd Ed.
  - Tutorial: OpenAI Spinning Up in Deep RL

Requirements and Grading

- All enrolled students must have taken a course of artificial intelligence, machine learning, or an equivalent course
  - Please contact me if you haven’t taken one of those courses

- Grading
  - Homework (40%)
  - Final project (60%)

Homework

- Four Problem Sets (10% each)

- Policies
  - Late work will not be accepted.
  - Please don't ask us for extensions. If you have a medical excuse, no problem, have deans send note to TAs.
  - Collaboration is fine, if acknowledged and your write-up is yours.
  - Don't copy (from each other, or online). We'll probably find out, and it will make you and us very unhappy.

Final Project

- Research-level project of your choice
  - Improving an existing approach
  - Focus on an unsolved task / benchmark
  - Create a new task / problem that hasn’t been addressed by RL
- Projects should be done by a group of 2 students
  - Contact me if you want to do it independently or with a bigger group
- Milestones:
  - Proposal (max 2 pages)
  - Progress report with survey (max 4 pages)
  - Presentation session
  - Final report (7 – 10 pages with NIPS format)

Important This Week

- Follow 荷塘雨课堂 on WeChat
- Join the course WeChat Group
- Start to form your final project group
- Check out the TensorFlow MNIST tutorial, unless you’re a TensorFlow pro

What’s reinforcement learning and why should we care about it?

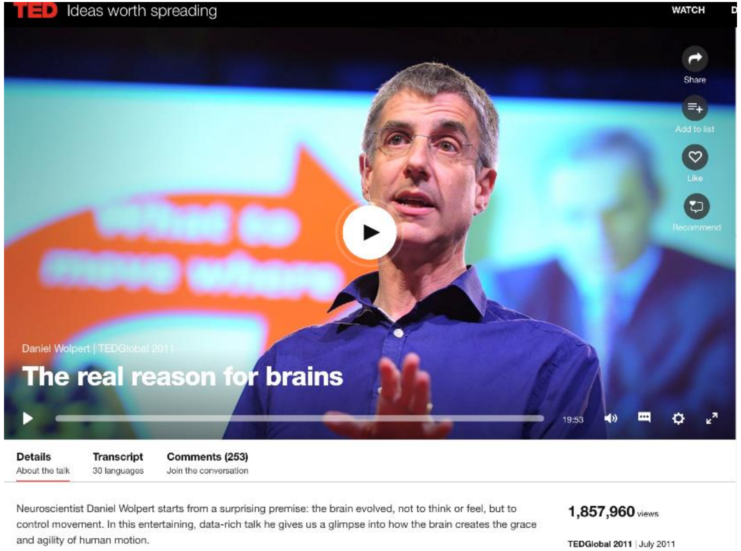
Goals of This Course

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

Acting to achieve goals is key part of intelligence

The only reason for us and animals to have brains is to produce adaptable and complex movements.

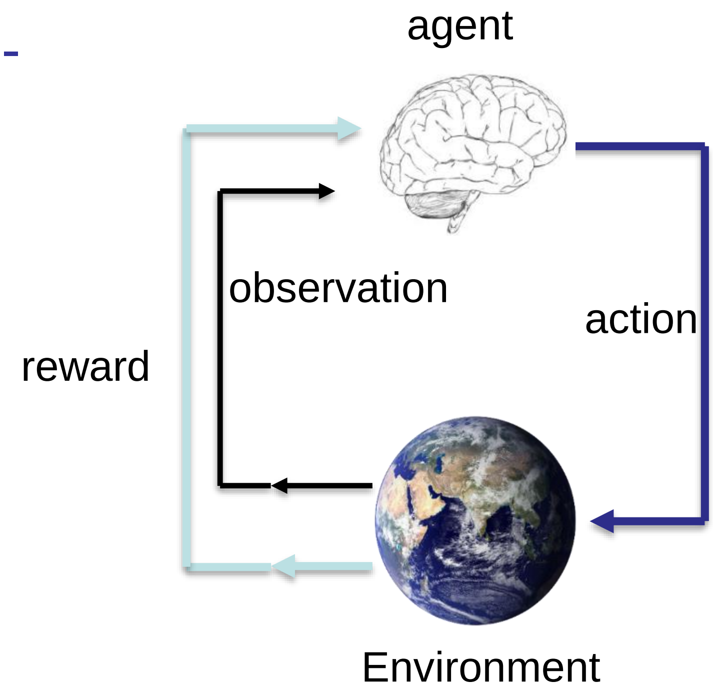
-- Daniel Wolpert



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

Reinforcement Learning (RL)

- A general-purpose framework for decision-making/behavior learning
  - RL is for an agent with the capacity to act
  - Each action influences the agent's future observation
  - Success is measured by a scalar reward signal
  - Goal: find a policy that maximizes expected total rewards

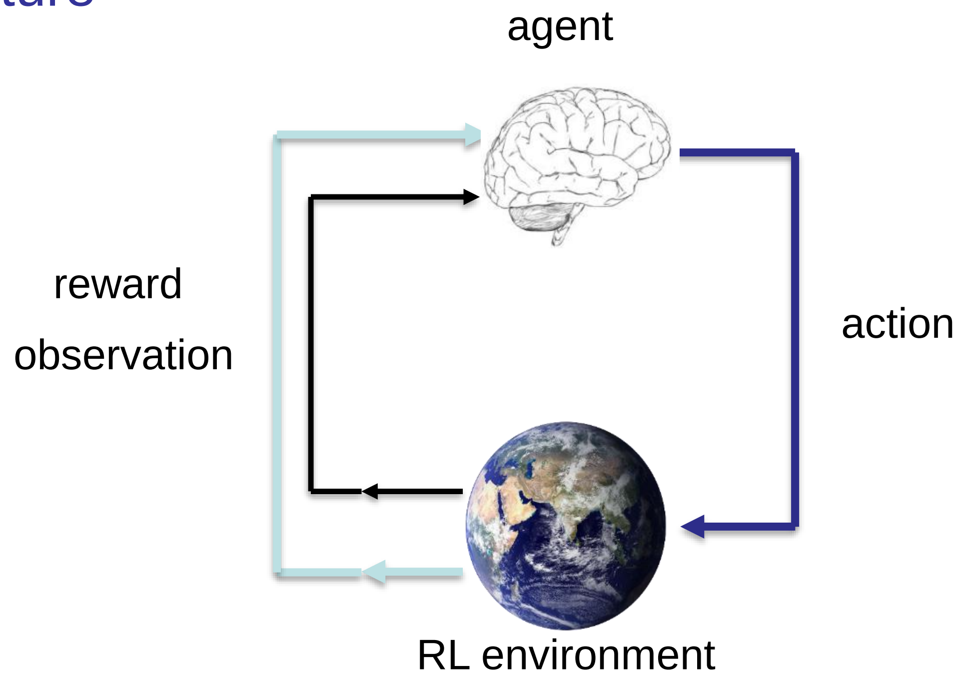


Reinforcement learning might be considered to encompass all of AI: an agent is placed in an environment and must learn to behave successfully therein.

-- Artificial Intelligence: a Modern Approach

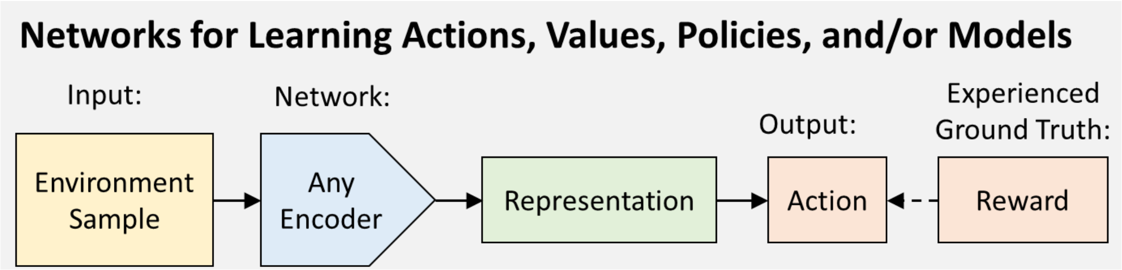
RL algorithms in a nutshell

- Exploration: add randomness to your actions
- If the result was better than expected, do more of the same in the future



What is deep reinforcement learning?

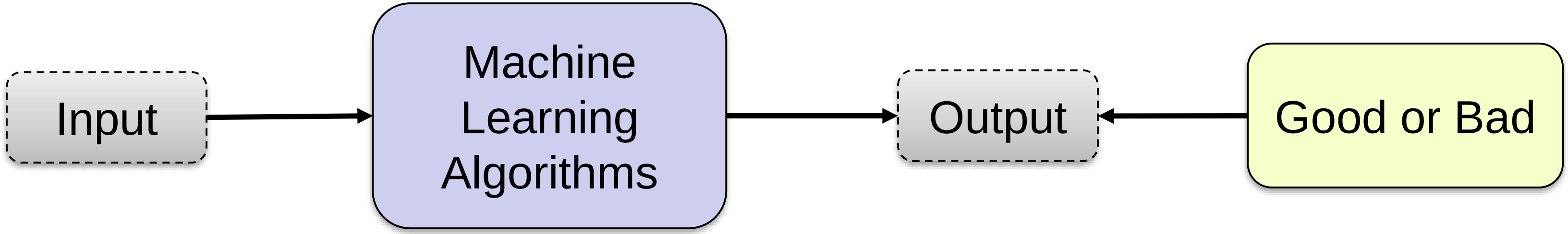
- Deep RL = RL + DL (deep learning)
- DL is a general-purpose framework for representation learning
  - Given an objective
  - Learn representation that is required to achieve objective
  - Directly from raw inputs
  - Using minimal domain knowledge
- Deep learning enables RL algorithms to solve complex problems in a end-to-end manner



Machine Learning Paradigms

- Supervised learning: learning from examples
- Unsupervised learning: learning structures in data
- Reinforcement learning: learning from experiences

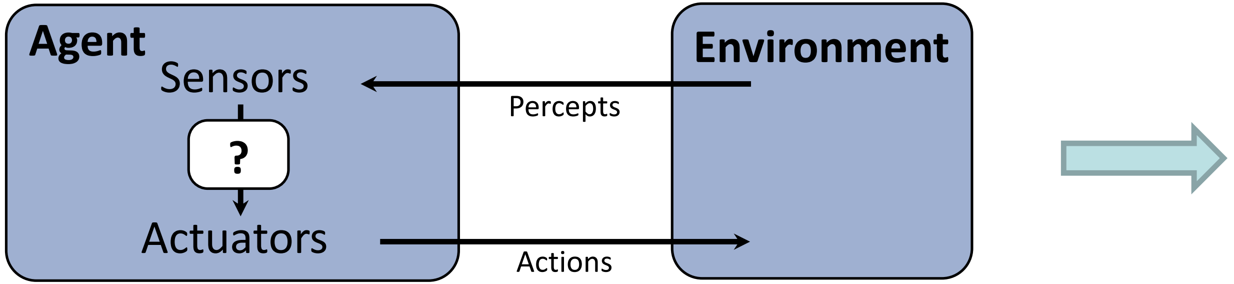
It's all "supervised" by a loss function!

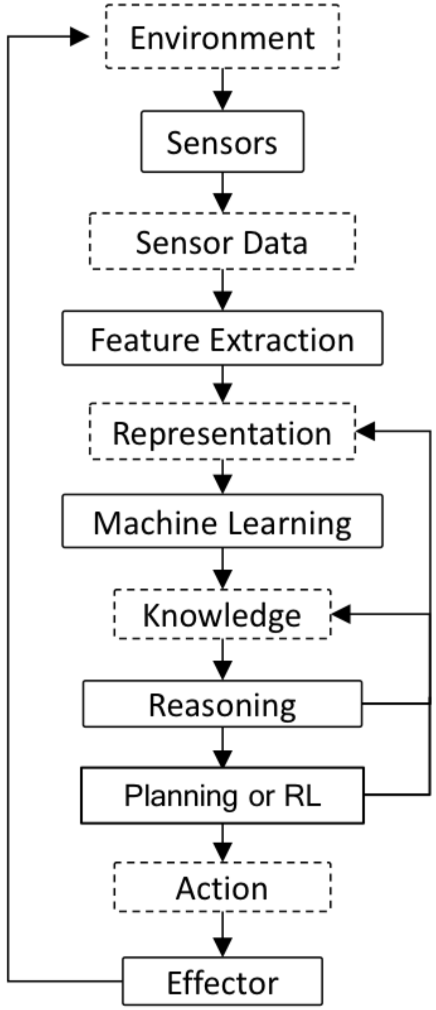


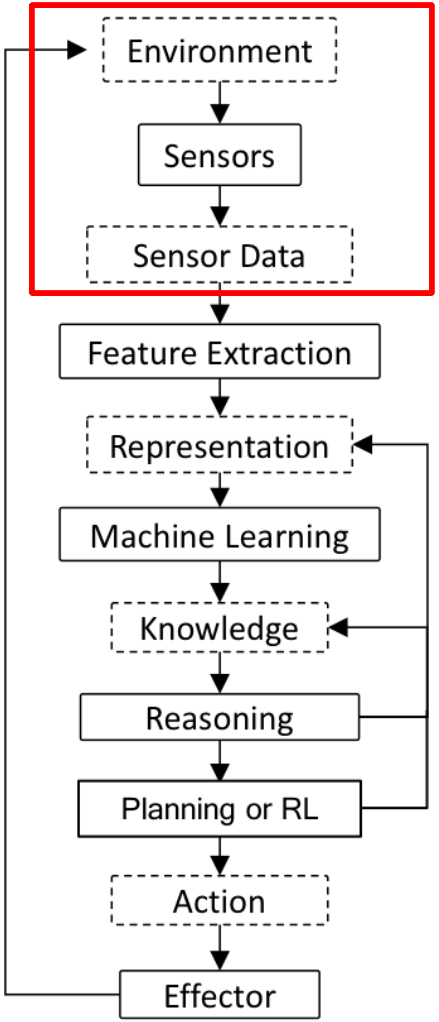
How RL is different from other machine learning paradigm

- Exploration: the agent does not have prior data knowing what actions is good and bad
- Non-stationarity: the agent’s actions affect the data it will receive in the future
- Credit assignment: the supervision signal (i.e., reward) is often delayed or far in the future
- Limited samples: actions take time to execute in the real world, which may limit the amount of experiences

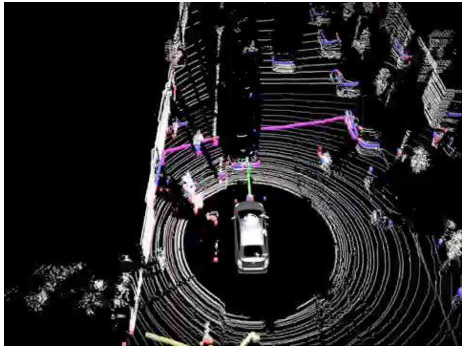
Elaborating the Agent Model







Sensing




Lidar


Camera
(Visible, Infrared)

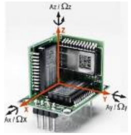

Radar


GPS

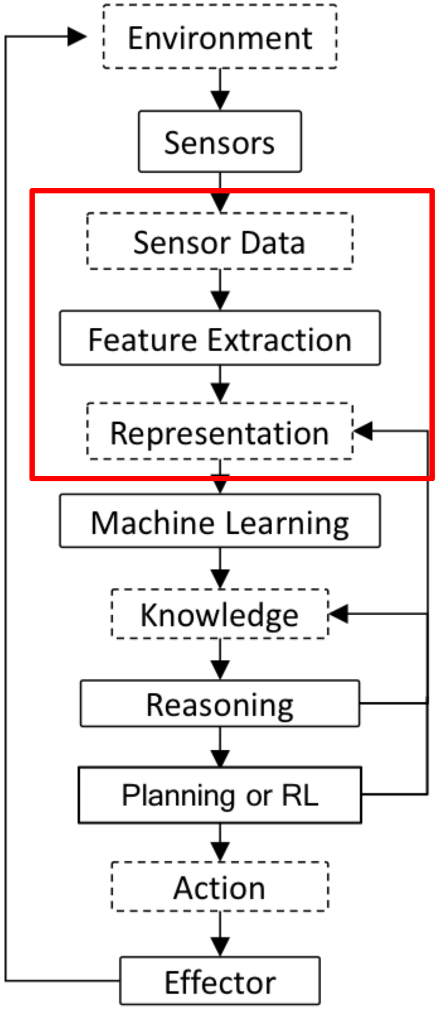

Stereo Camera

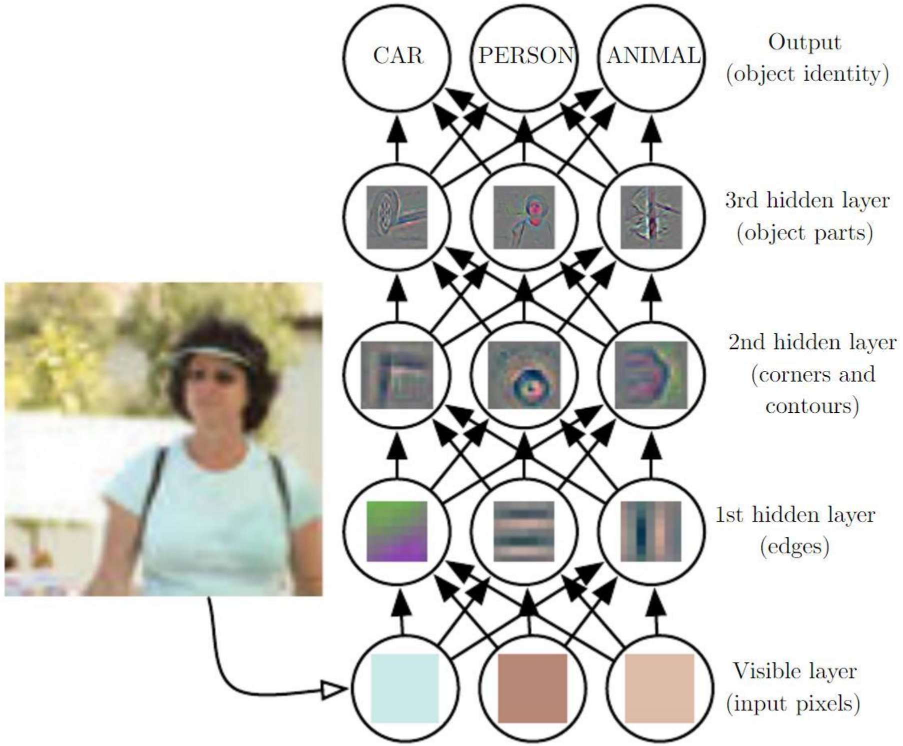

Microphone

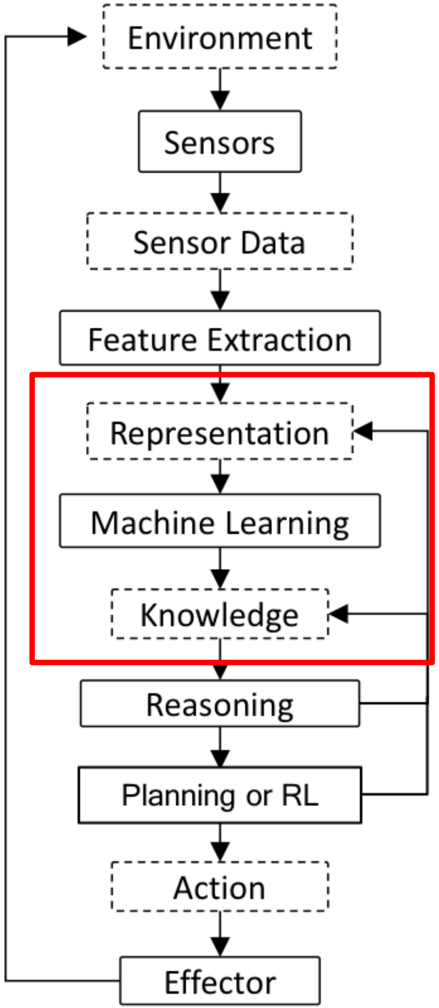

Networking
(Wired, Wireless)


IMU

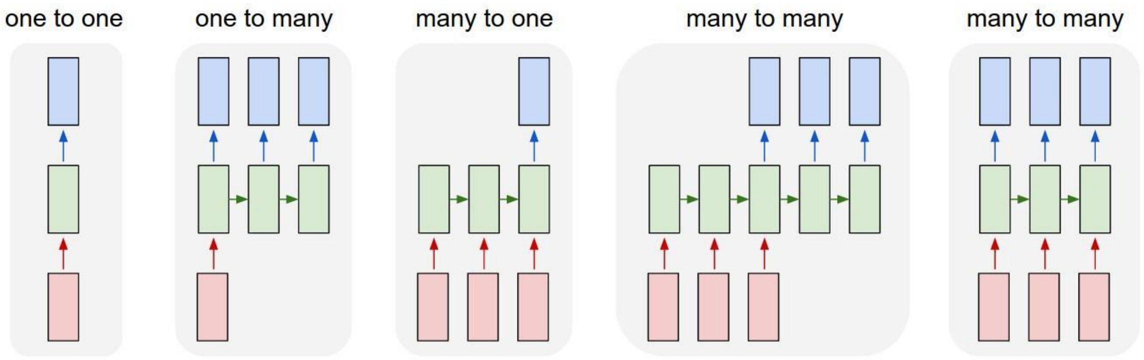
Representation Engineering or Learning

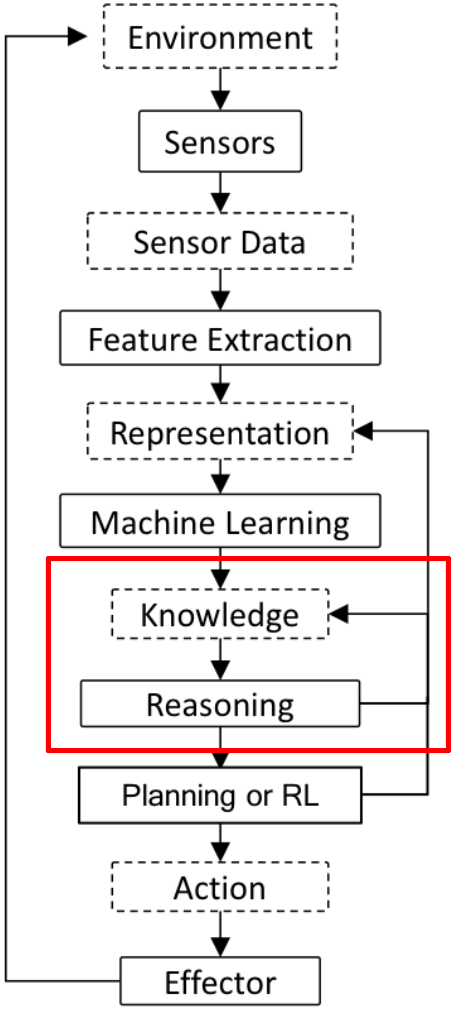






Machine Learning: Classification, Regression, or Clustering





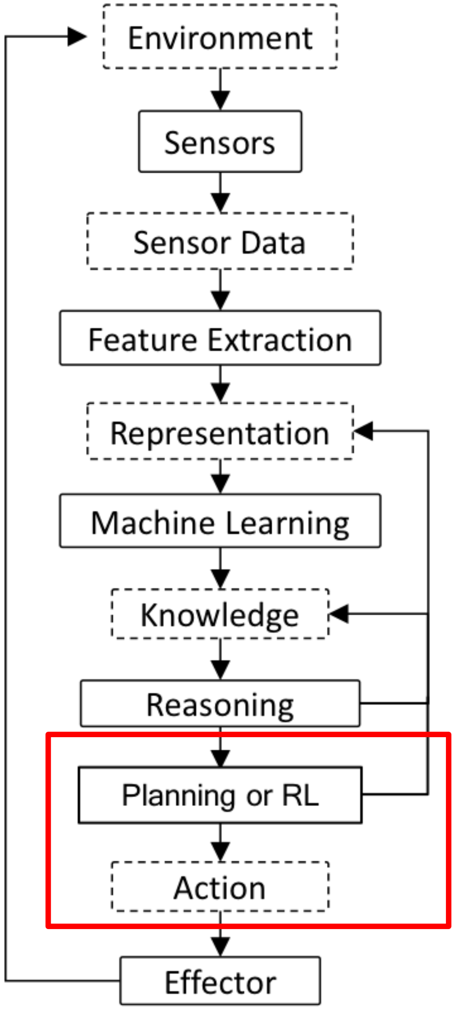
Reasoning or Inference

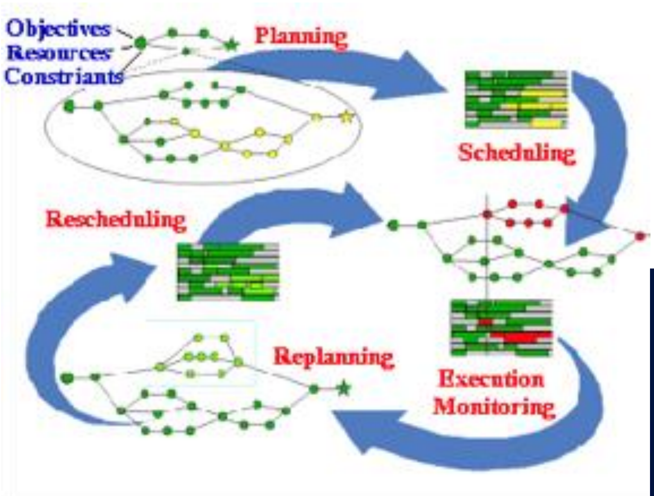
Image Recognition:
If it looks like a duck

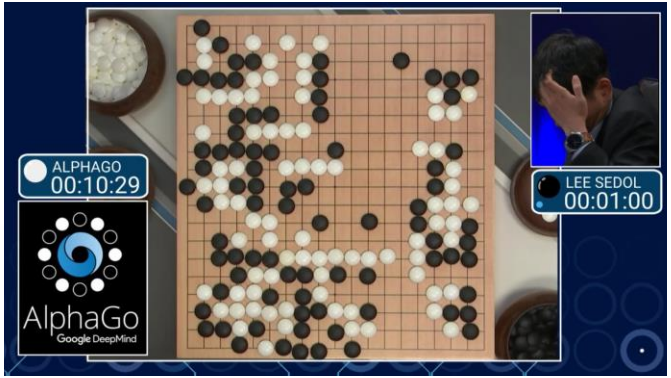
Audio Recognition:
Quacks like a duck

Activity Recognition:
Swims like a duck

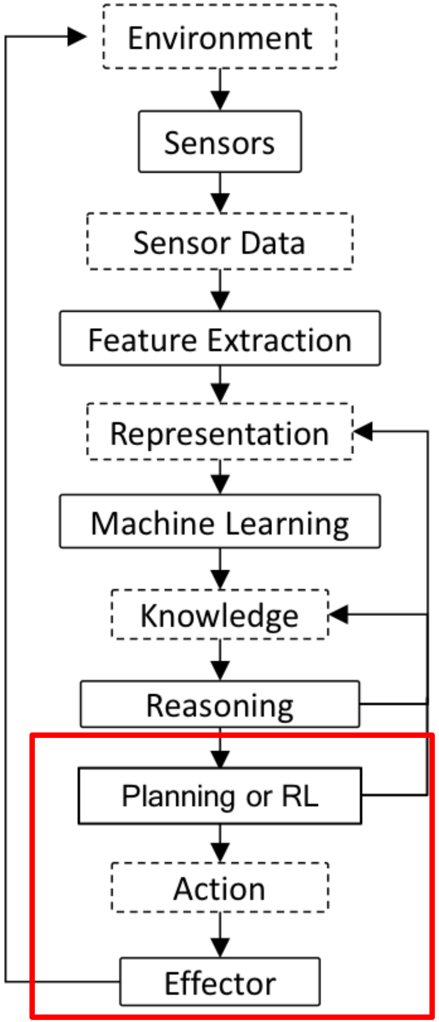
Planning and Reinforcement Learning

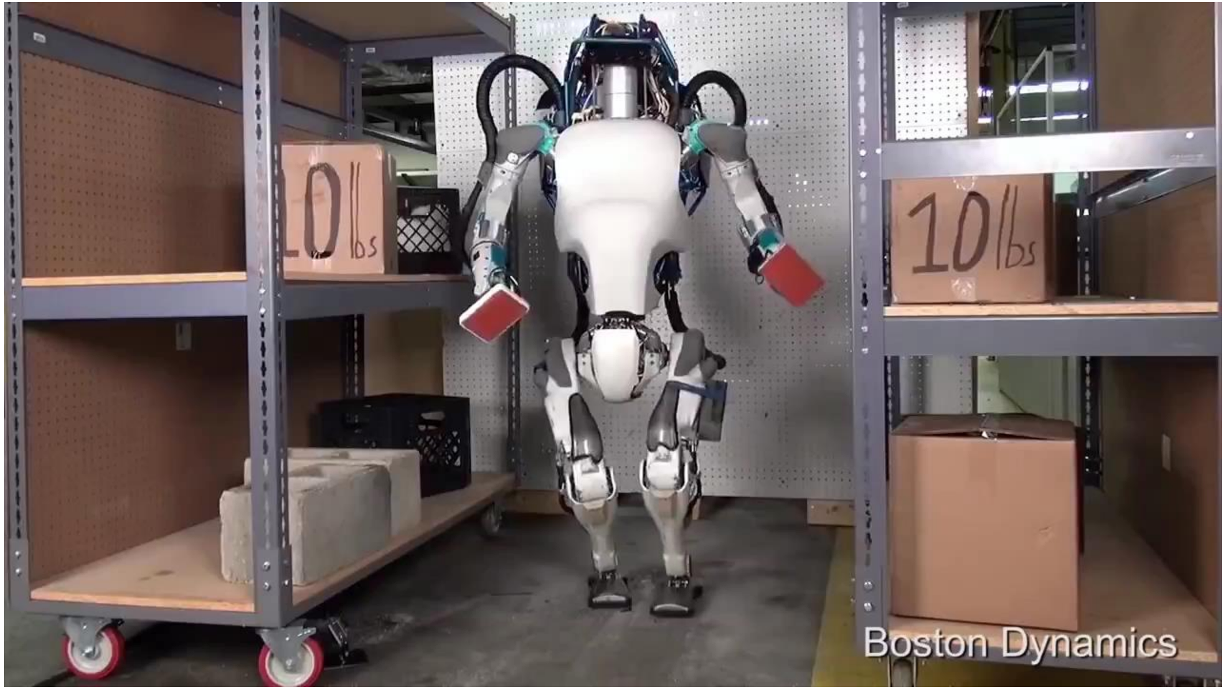


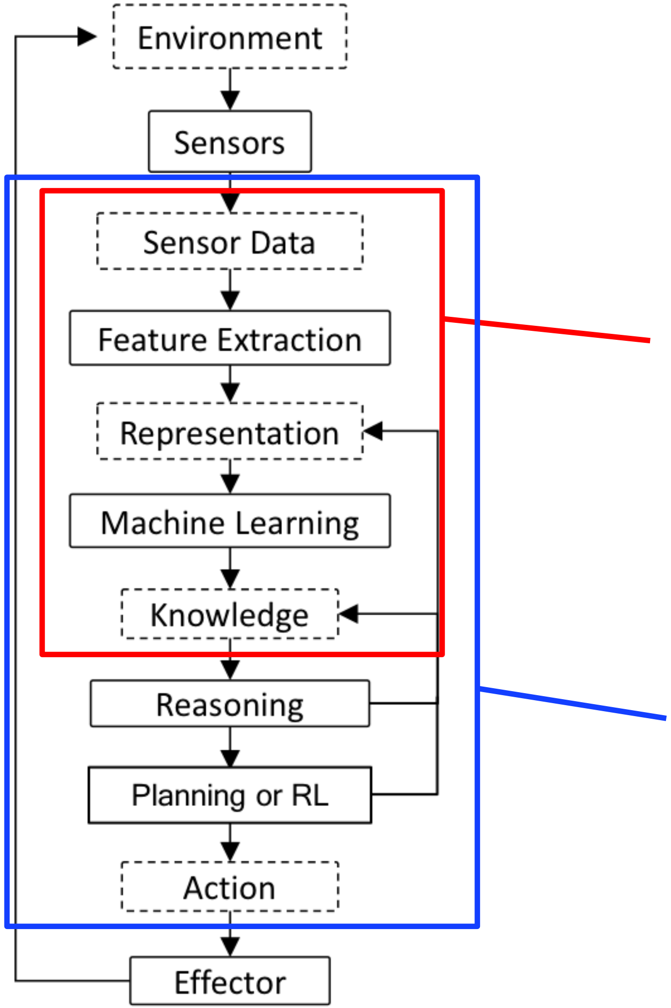




Robotics


Environment
Sensors
Sensor Data
Feature Extraction
Representation
Machine Learning
Knowledge
Reasoning
Planning or RL
Action
Effector


10 lbs
10 lbs
Boston Dynamics



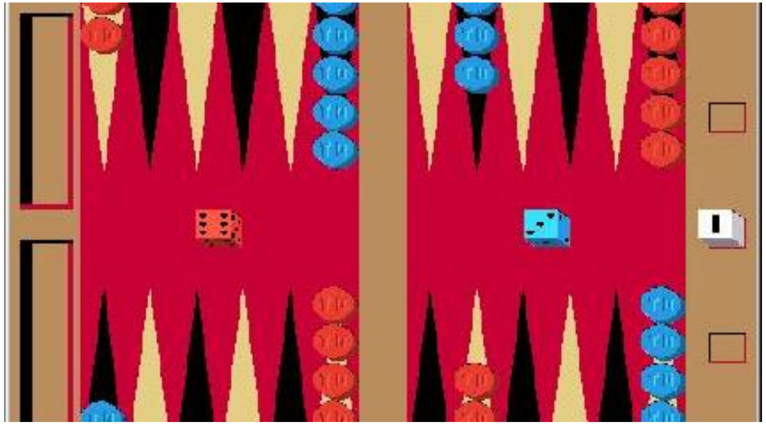
The Promise of **Deep Learning**

The Promise of **Deep Reinforcement Learning**

Successes so far

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

DeepMind Atari (©Two Minute Lectures)

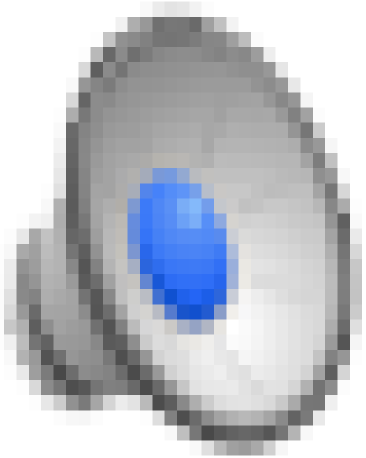
[Minh et al., Nature 2015]

AlphaGo

[Silver et al., Nature 2017]

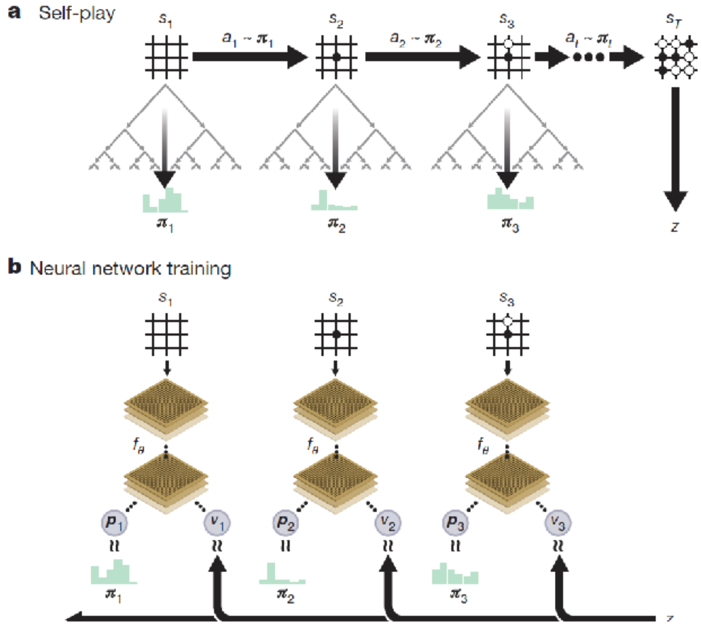
Monte Carlo Tree Search, learning policy and value function networks for pruning the search tree, expert demonstrations, self play, Tensor Processing Unit

AlphaGo



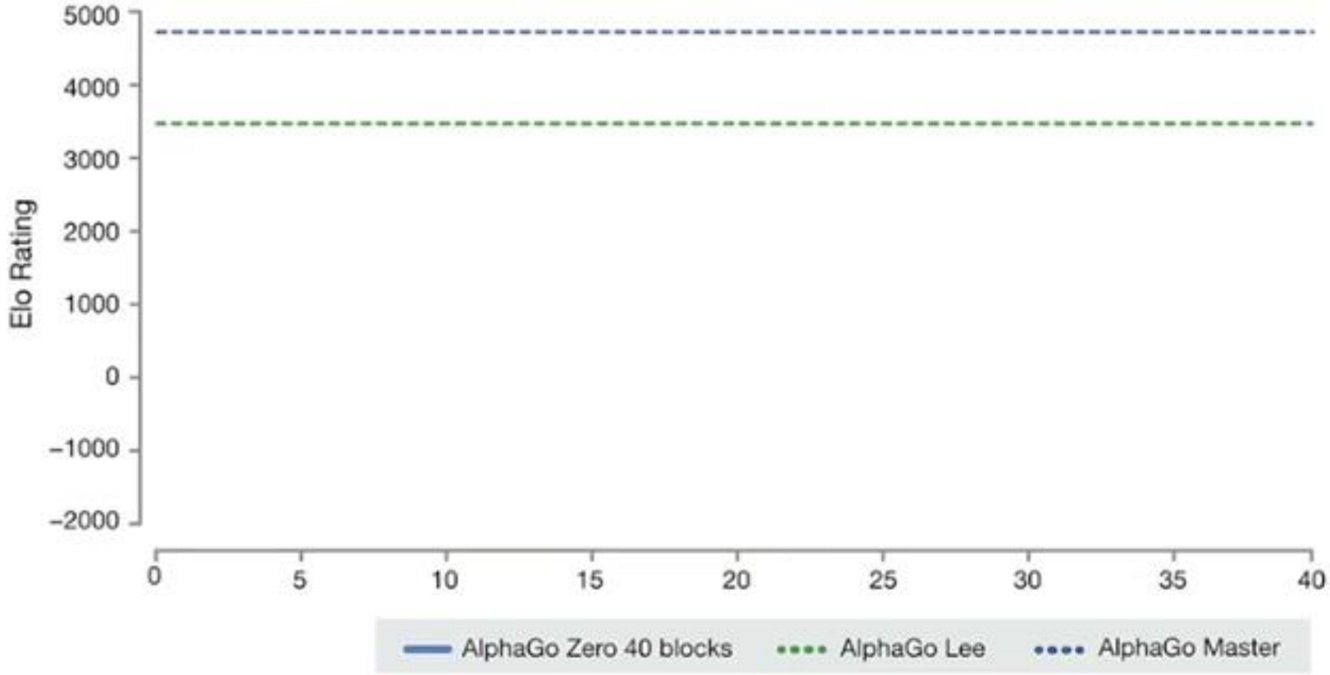
AlphaGo Zero

- Learning from scratch by self-play
- No human expert data
- MCTS to select moves during training and testing!



[Silver et al., Nature 2017]

AlphaGo Zero: Master the game of Go without Human Knowledge



Recent progresses in video games

- OpenAI Five for Dota 2
  - won 5v5 best-of-three match against professional team
  - 256 GPUs, 128k CPUs 180 years of experience per day
- Deepmind AlphaStar for Startcraft
  - defeat a top professional player
  - supervised training followed by a league competition training

AlphaGo vs the Real World

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

RL Challenges: AlphaGo vs the Real World

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

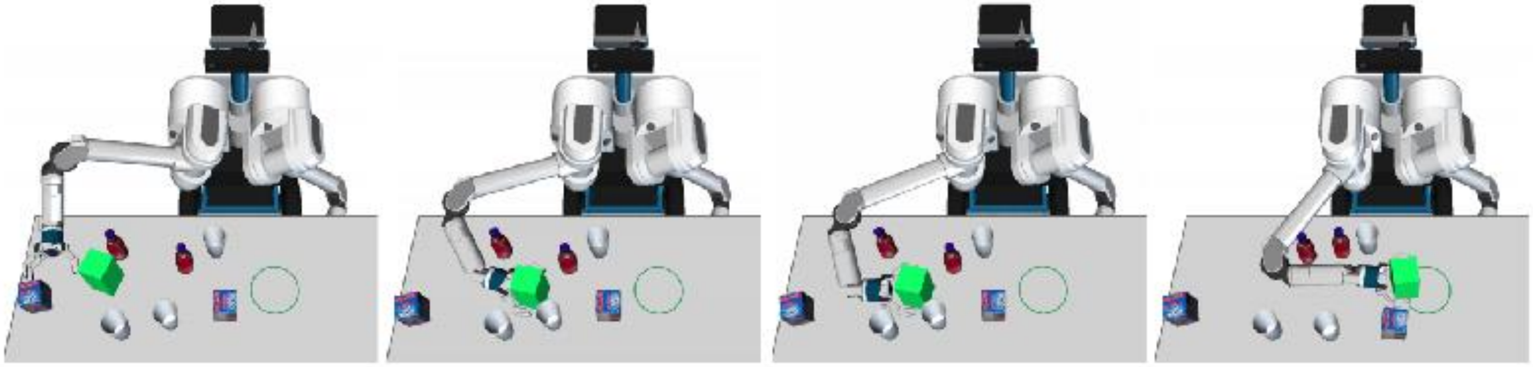
- **Known environment** (known entities and dynamics) vs. **Unknown environment** (unknown entities and dynamics).
- Need for behaviors to **transfer/generalize** across environmental variations since the real world is very diverse

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.

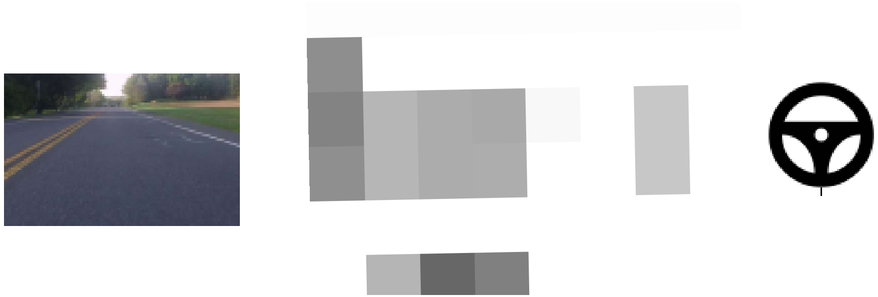


Rearrangement Planning via Heuristic Search, Jennifer E. King, Siddhartha S. Srinivasa

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

Behavior learning is challenging because state estimation is challenging, in other world, because computer vision is challenging.

Interesting direction: leveraging DRL and computer vision

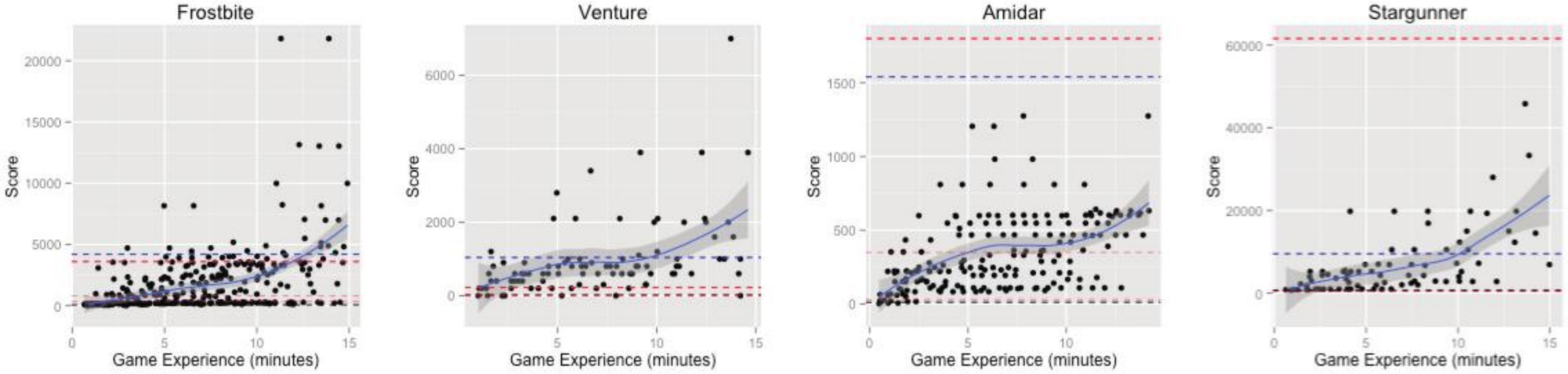
RL Challenges: AlphaGo vs the Real World

How the world of AlphaGo is different than the real world?

- **Known environment** (known entities and dynamics) vs. **Unknown environment** (unknown entities and dynamics).
- Need for behaviors to **transfer/generalize** across environmental variations since the real world is very diverse
- **Cheap** vs. **Expensive to get experience samples**

DRL Sample Efficiency

Humans after 15 minutes tend to outperform DDQN after 115 hours



Black dots: human play
Blue curve: mean of human play
Blue dashed line: “expert”human play

Red dashed line:
DDQN after 40, 115, 920 hours

[Image credit: Tsividis, Pouncy, Xu, Tenenbaum, Gershman, 2017]

Reinforcement Learning in Humans

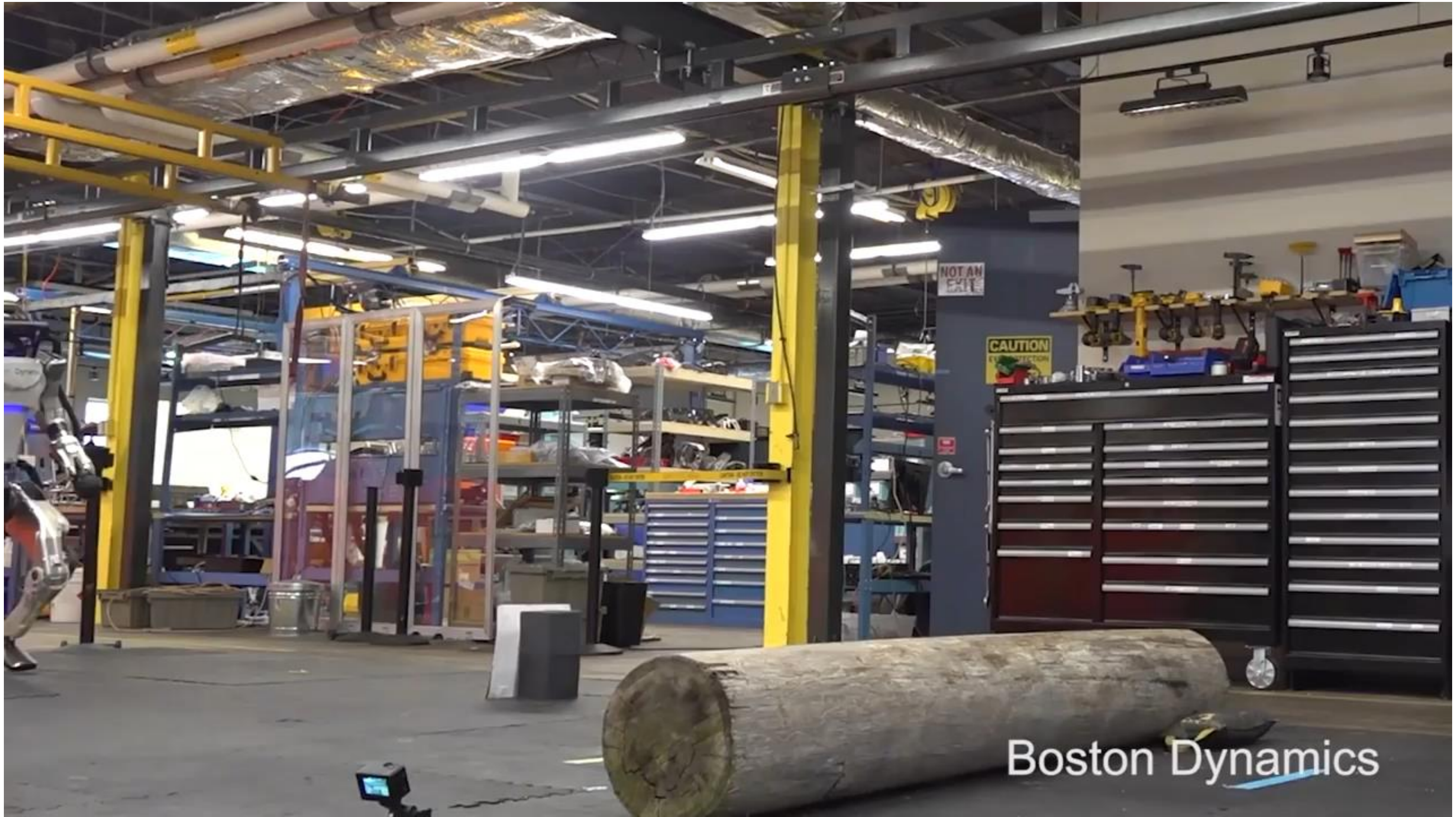
- Human appear to learn to act (e.g., walk) through “very few examples” of trial and error. **How** is an open question...
- Possible answers:
  - **Hardware:** 230 million years of bipedal movement data
  - **Imitation Learning:** Observation of other humans walking
  - **Algorithms:** Better than backpropagation and stochastic gradient descent

RL Challenges: AlphaGo vs the Real World

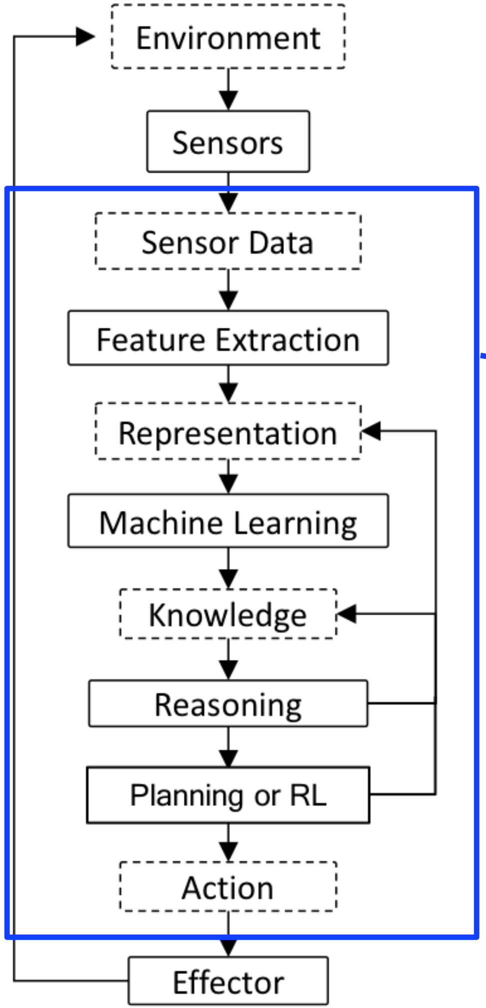
How the world of AlphaGo is different than the real world?

- **Known environment** (known entities and dynamics) vs. **Unknown environment** (unknown entities and dynamics).
- Need for behaviors to **transfer/generalize** across environmental variations since the real world is very diverse
- **Cheap** vs. **Expensive** to get experience samples
- **Discrete** vs. **Continuous** actions
- **One goal** vs. **Many goals**
- **Rewards automatic** vs. rewards need themselves to be **detected**

To date, for most successful robots operating in the real world: Deep RL is not involved



To date, for most successful robots operating in the real world: Deep RL is not involved



The Promise of deep reinforcement learning for artificial general intelligence

But many unsolved challenges for the real-world applications

DRL Challenges

- Sample efficiency
- Transfer learning
- Generalization
- Long horizon reasoning
- Model-based RL
- Sparse reward
- Reward design/learning
- Planning + Learning
- Lifelong learning
- Safe learning
- Interpretability
- …

Reinforcement Learning Algorithms

Better sample efficiency ⟷ Less sample efficiency

Model-Based | Off-policy Q-learning | Actor-critic | On-policy Policy gradient | Evolutionary/ Gradient-free

Model-Based

- Learn the model of the world, then plan using the model
- Update model often
- Re-plan often

Value-Based

- Learn the state or state-action value
- Act by choosing best action in state
- Exploration is a necessary add-on

Policy-based

- Learn the stochastic policy function that maps state to action
- Act by sampling policy
- Exploration is baked in