IASD M2 at Paris Dauphine

Deep Reinforcement Learning

1: Introduction and Course overview

Eric Benhamou Thérèse Des Escotais

PSL UNIVERSITÉ PARIS

Dauphine UNIVERSITÉ PARIS

ENS ÉCOLE NORMALE SUPÉRIEURE

MINES ParisTech

Organisation

Agenda – Grading

Lectures posted on

https://www.lamsade.dauphine.fr/~ebenhamou/deep_reinforcement_learning/index.html

Grading

3 Homeworks (after week 2, 2 weeks) to be filled on

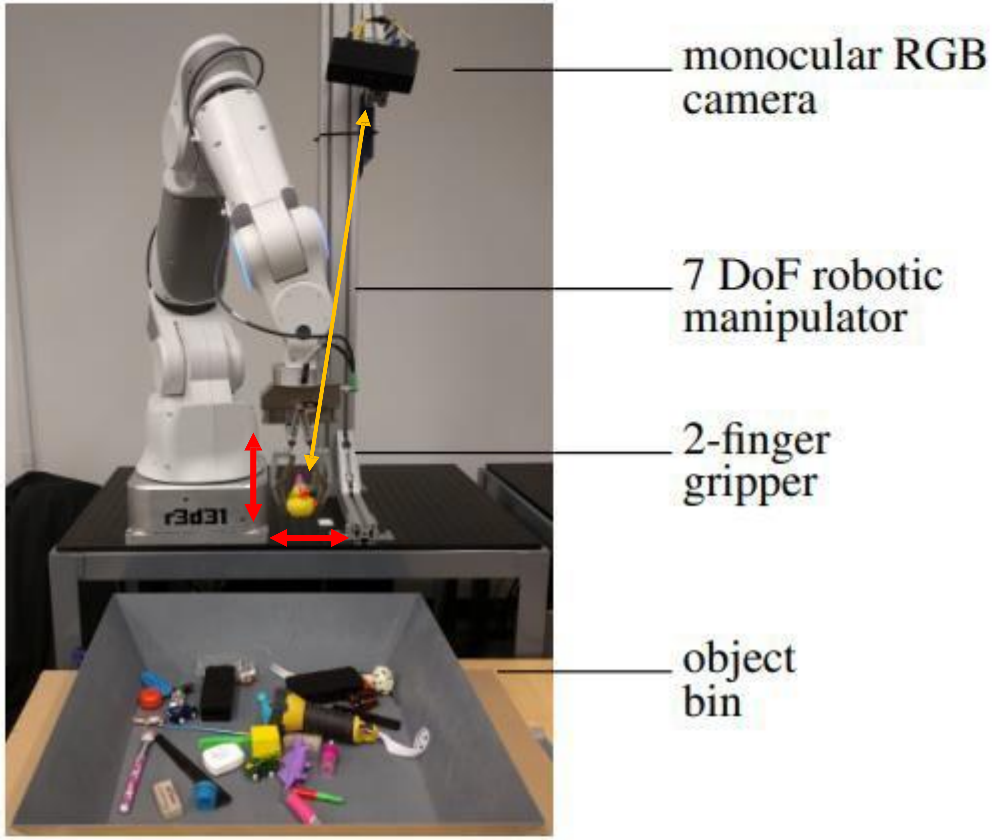
Dauphine | PSL Université Paris Moodle

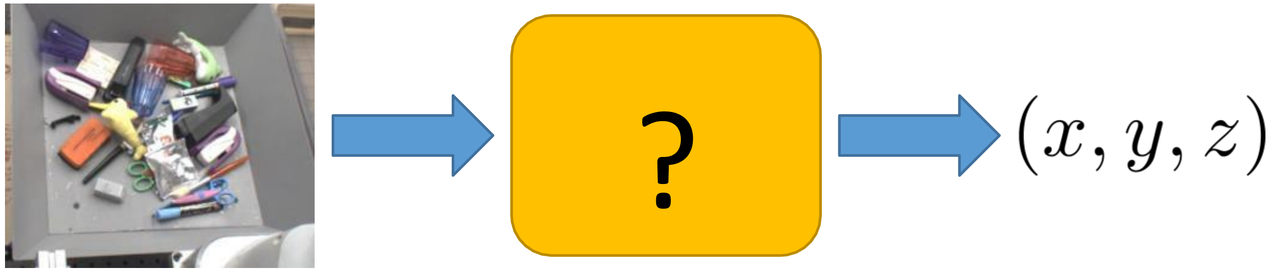
Final project consisting of a kaggle competition

Score: 20% x 3 homeworks + 40% project

Acknowledgement

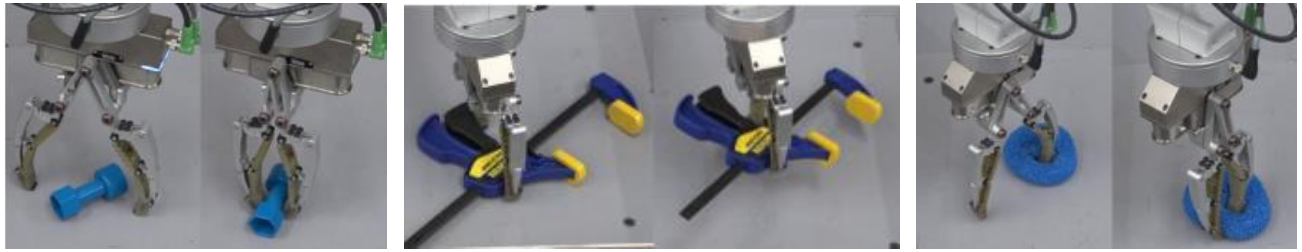
The materials of this course is entirely based on the seminal course of Sergey Levine CS285





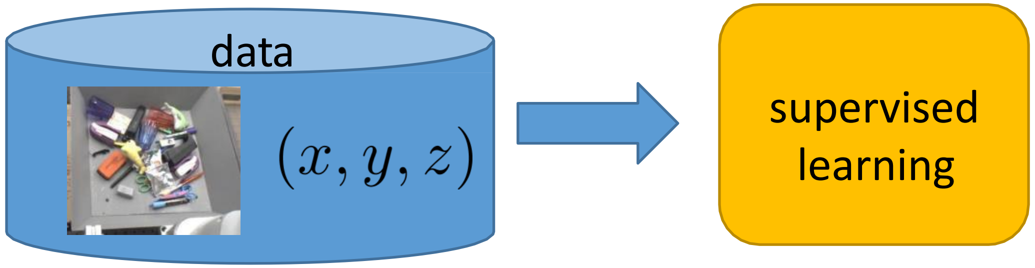
Option 1:

Understand the problem, design a solution



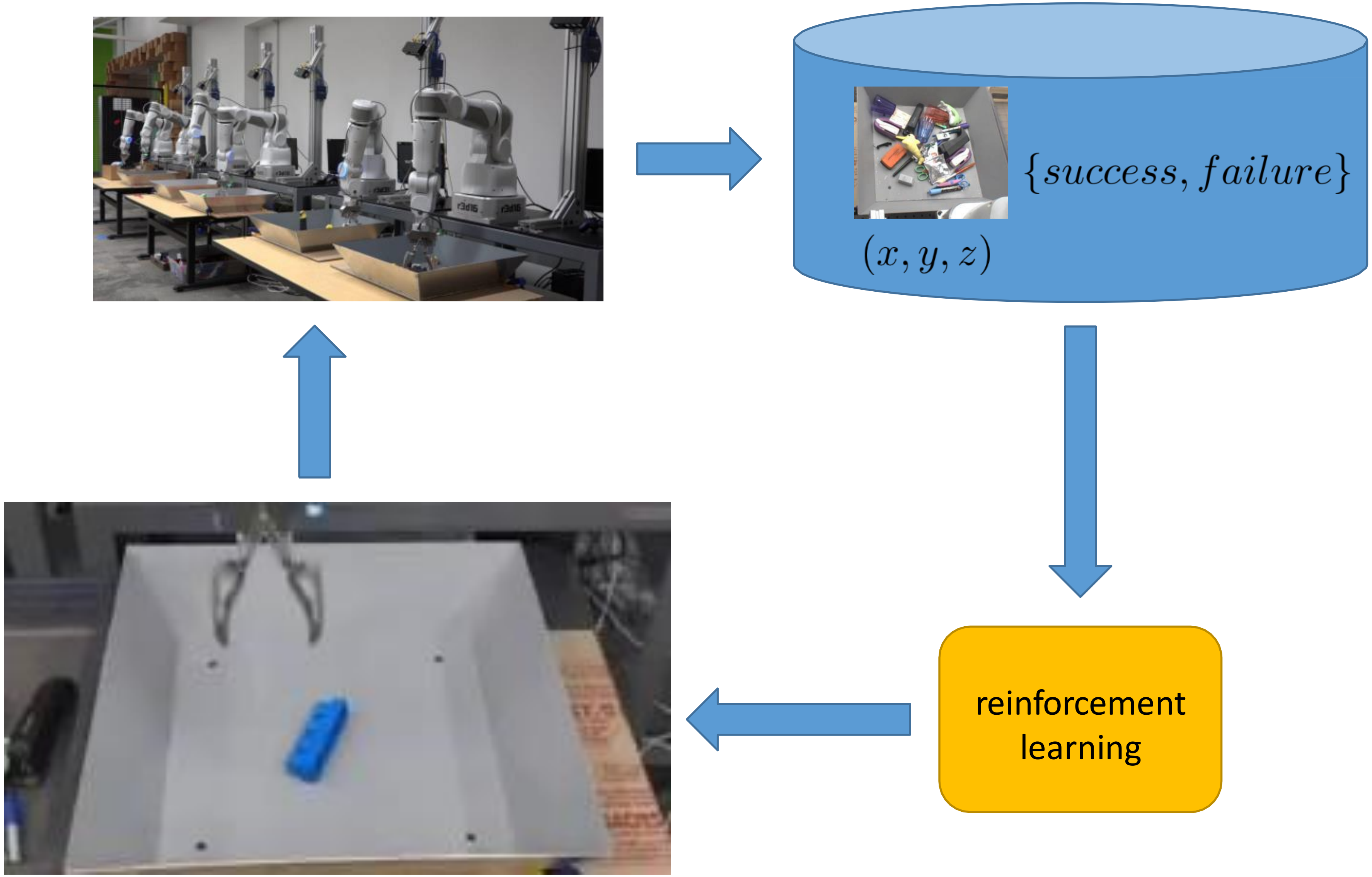
Option 2:

Set it up as a machine learning problem



Objective of the class

r3d24
r3d32



$\{success, failure\}$

(x, y, z)

reinforcement learning

What is reinforcement learning?

What is reinforcement learning?

Mathematical formalism for learning-based decision making

Approach for learning decision making and control from experience

How is this different from other machine learning topics?

Standard (supervised) machine learning:

given $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$

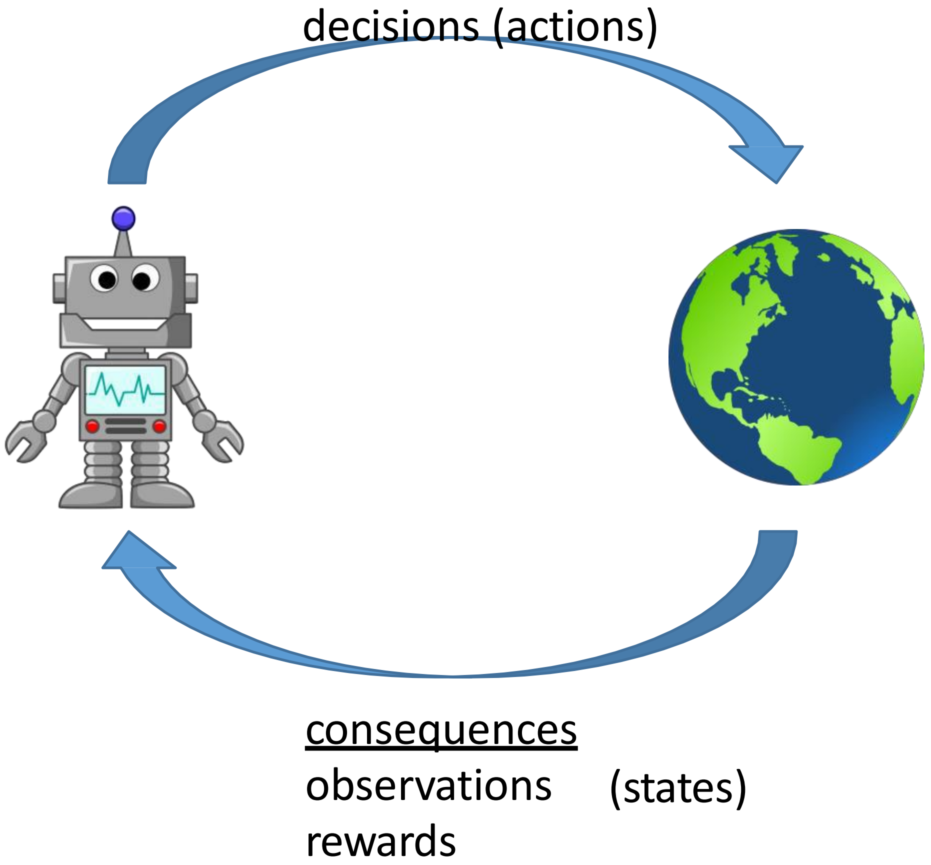
learn to predict y from $\mathbf{x}$ $\qquad f(\mathbf{x}) \approx y$

Usually assumes:

- i.i.d. data
- known ground truth outputs in training

Reinforcement learning:

- Data is **not** i.i.d.: previous outputs influence future inputs!
- Ground truth answer is not known, only know if we succeeded or failed
  - more generally, we know the reward



decisions (actions)

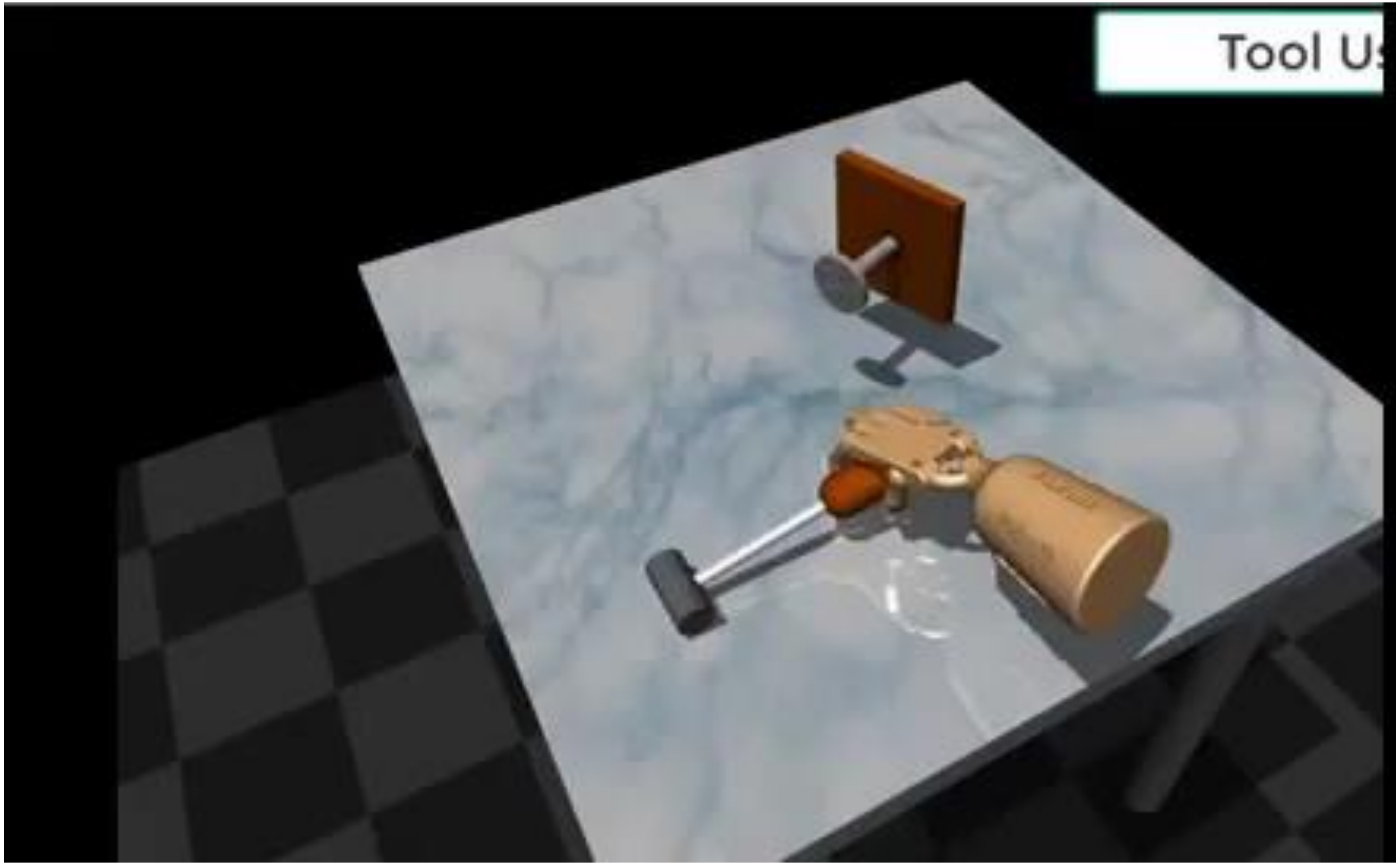
consequences
observations (states)
rewards

Actions: muscle contractions
Observations: sight, smell
Rewards: food

Actions: motor current or torque
Observations: camera images
Rewards: task success measure (e.g.,
ıg speed)

Actions: what to purchase
Observations: inventory levels
Rewards: profit

Complex physical tasks…

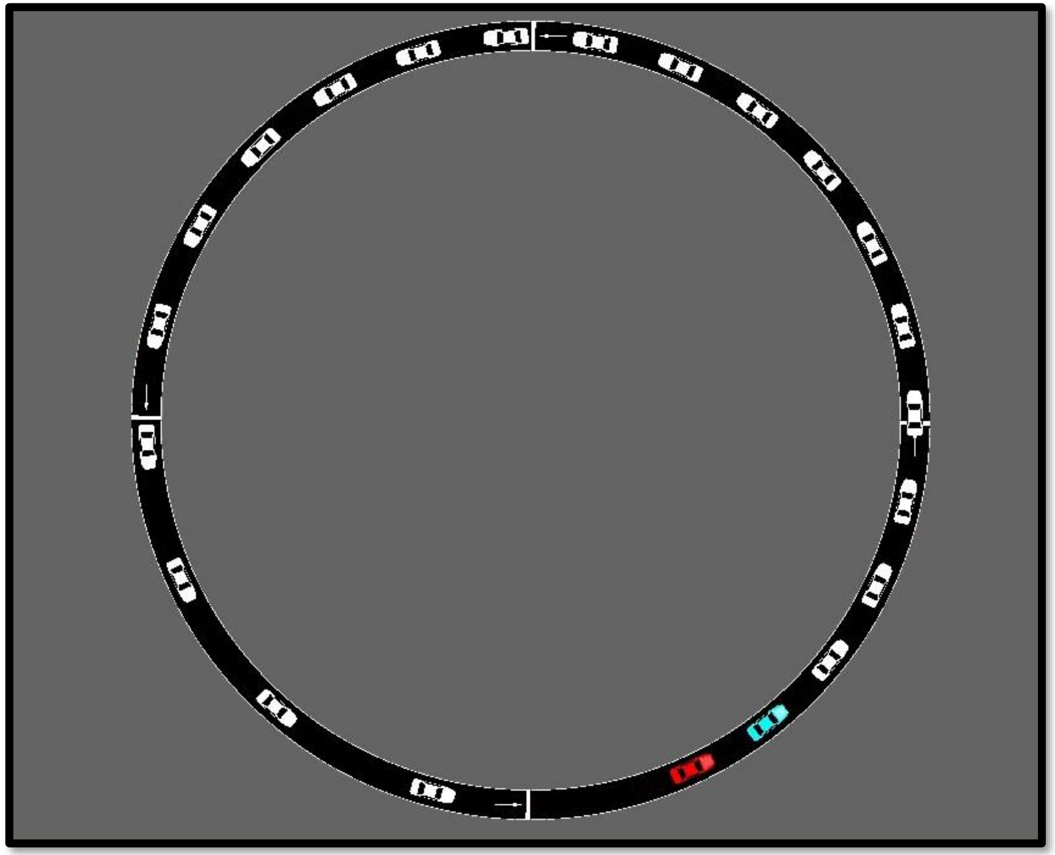


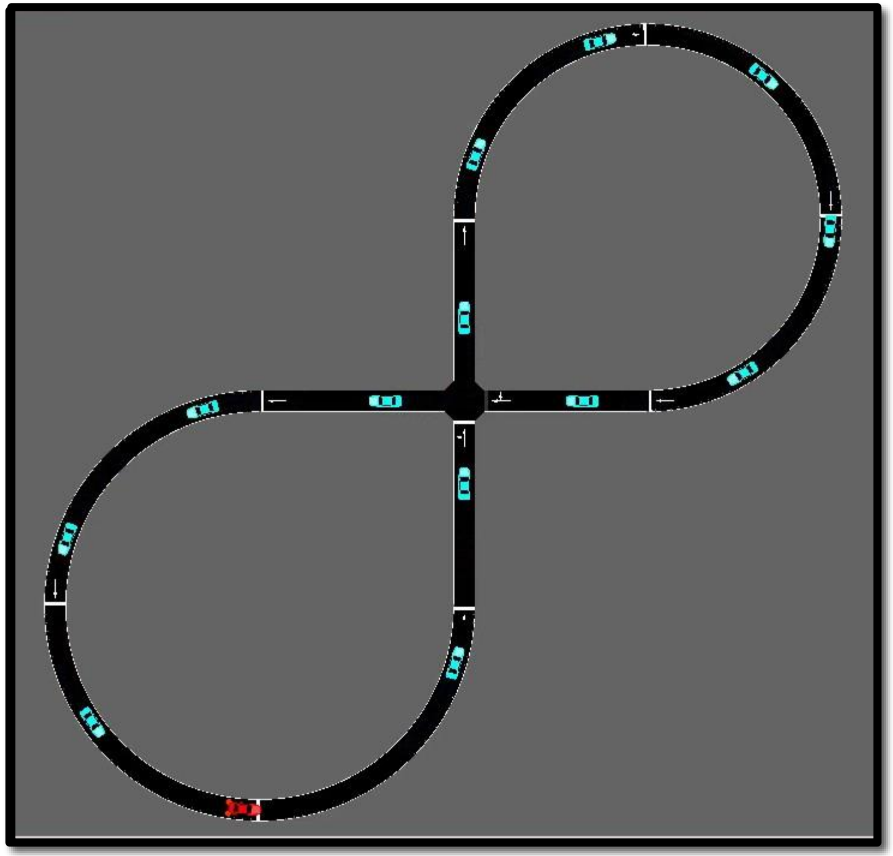
Rajeswaran, et al. 2018

In the real world…

Kalashnikov et al. ‘18

Not just games and robots!

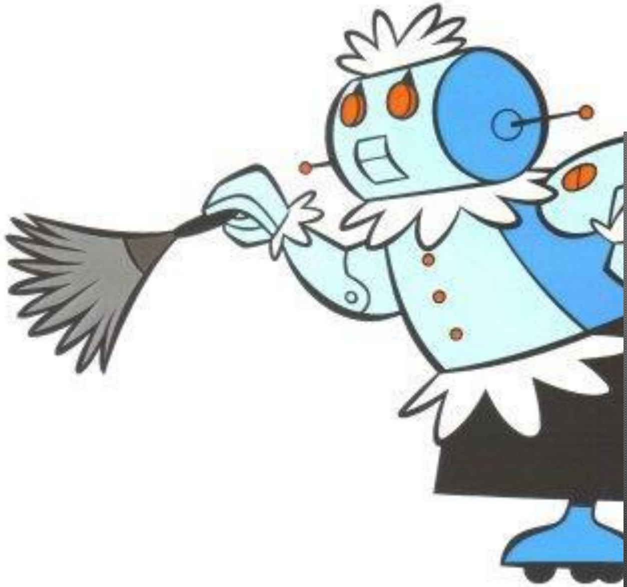


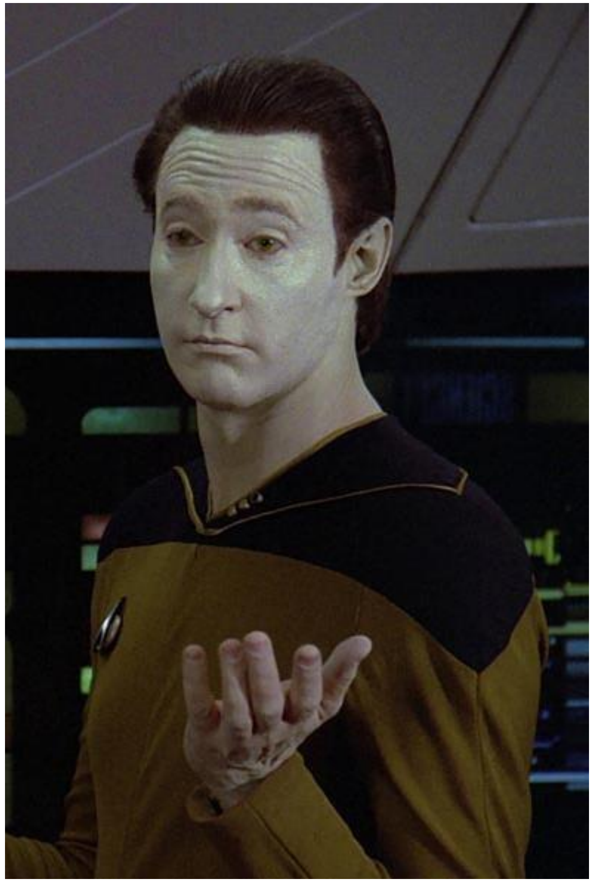


Cathy Wu

Why should we care about deep reinforcement learning?

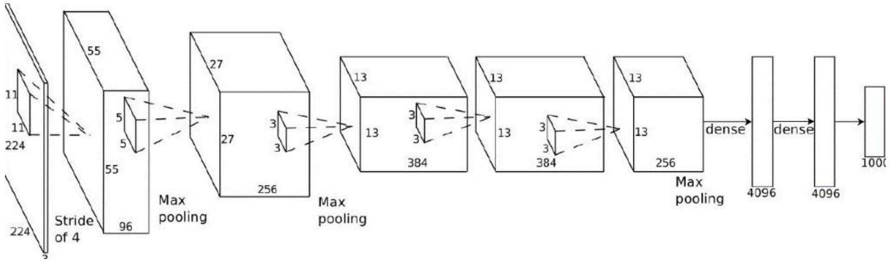
How do we build intelligent machines?

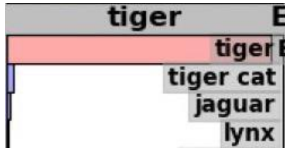


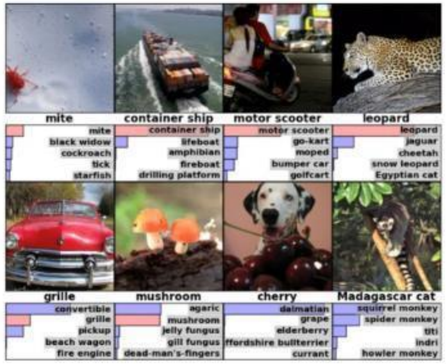


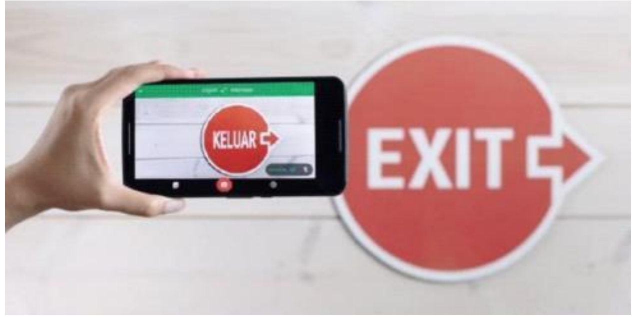
Intelligent machines must be able to adapt

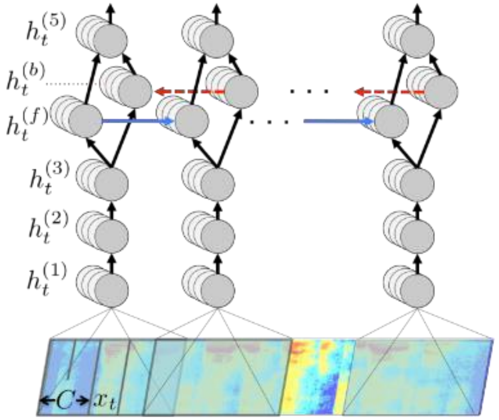
Deep learning helps us handle *unstructured environments*



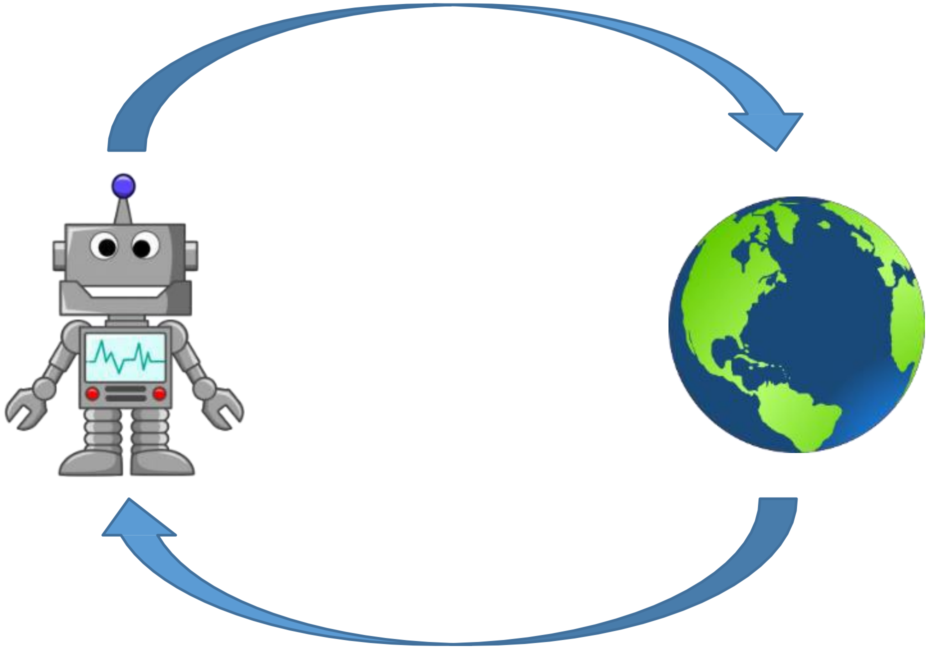








Reinforcement learning provides a formalism for *behavior*



decisions (actions)

Consequences observations rewards

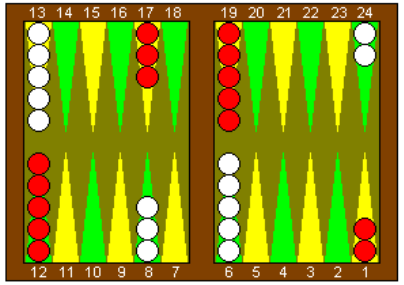
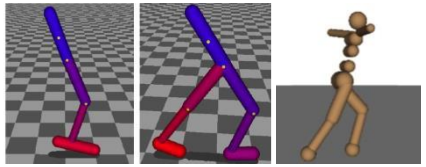
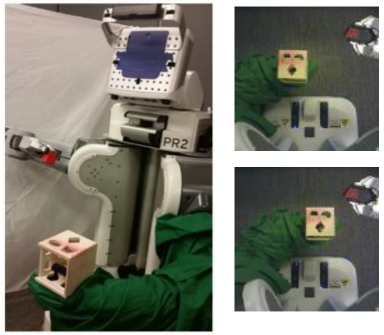


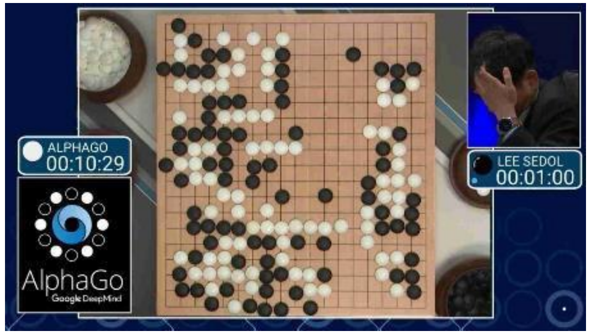
Figure 2. An illustration of the normal opening position in backgammon. TD-Gammon has sparked a near-universal conversion in the way experts play certain opening rolls. For example, with an opening roll of 4-1, most players have now switched from the traditional move of 13-9, 6-5, to TD-Gammon's preference, 13-9, 24-23. TD-Gammon's analysis is given in Table 2.



Schulman et al. '14 & '15

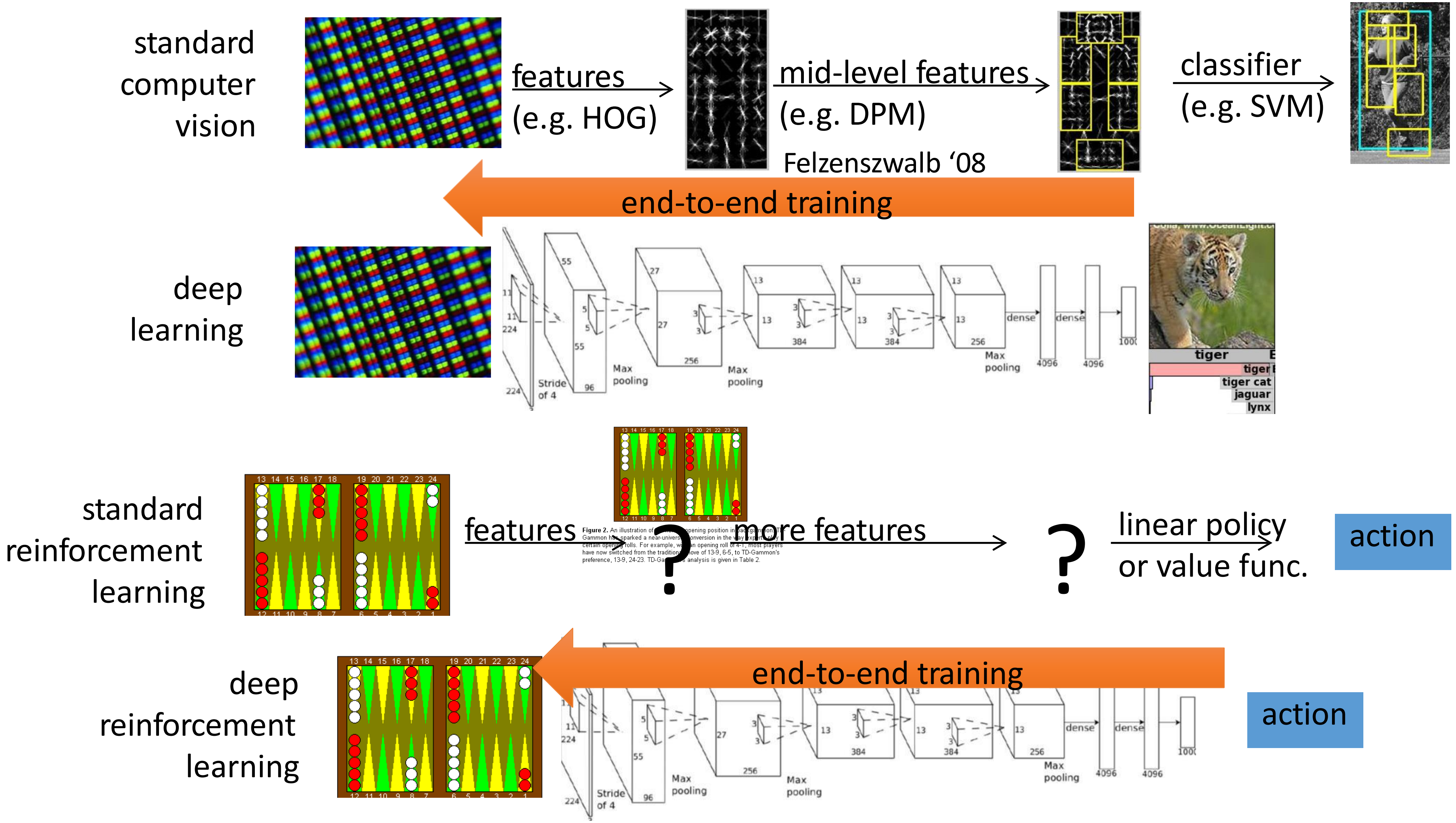


Levine*, Finn*, et al. '16



Mnih et al. '13

What is deep RL, and why should we care?



standard computer vision

features (e.g. HOG)

mid-level features (e.g. DPM)

Felzenszwalb ‘08

classifier (e.g. SVM)

end-to-end training

deep learning

standard reinforcement learning

features

Figure 2. An illustration of [illegible] opening position in backgammon. TD-Gammon has sparked a near-universal conversion in the way experts play certain opening rolls. For example, with an opening roll of 4-1, most players have now switched from the traditional move of 13-9, 6-5, to TD-Gammon's preference, 13-9, 24-23. TD-Gammon's analysis is given in Table 2.

?

more features

?

linear policy or value func.

action

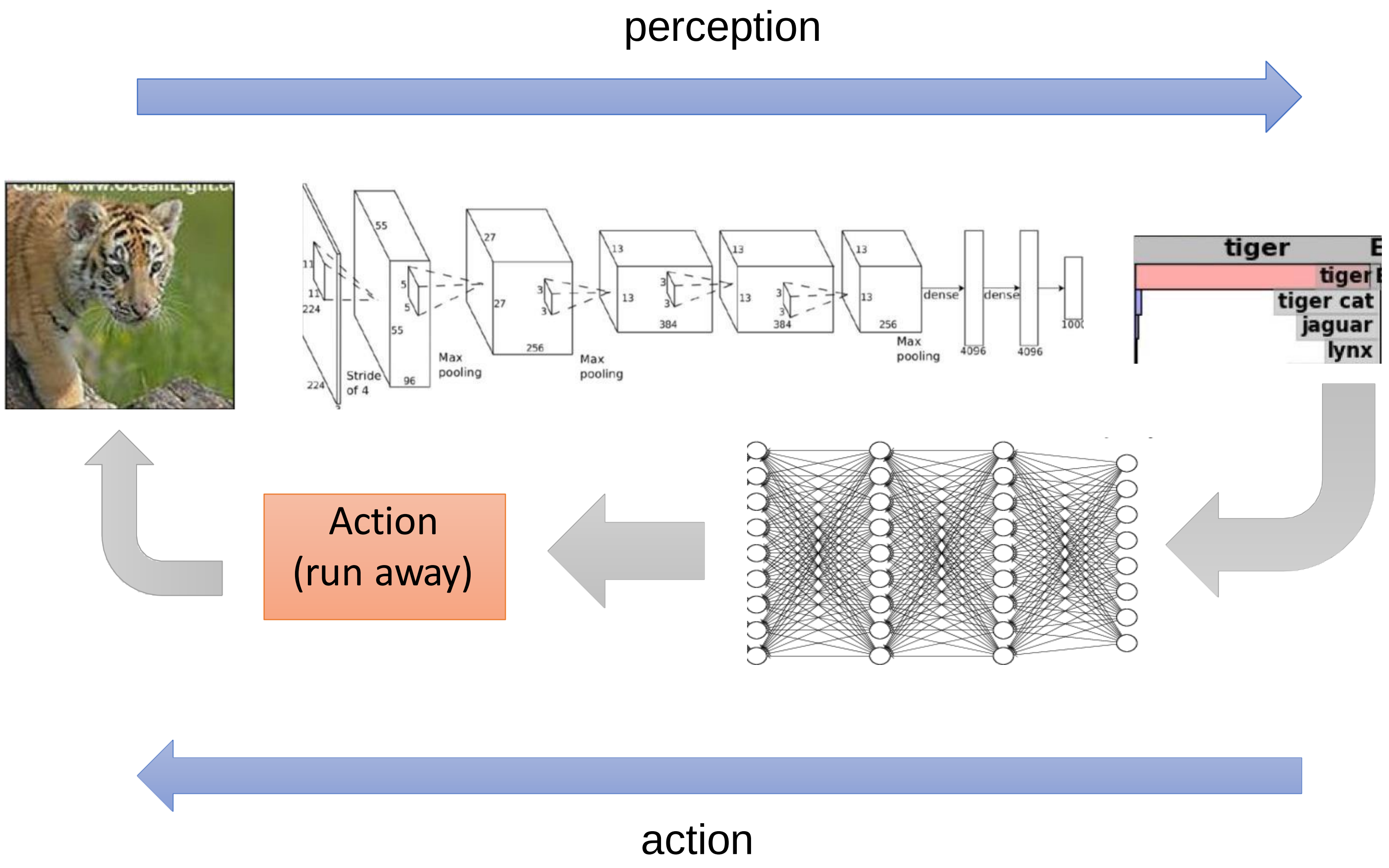
deep reinforcement learning

end-to-end training

action

What does end-to-end learning mean for sequential decision making?

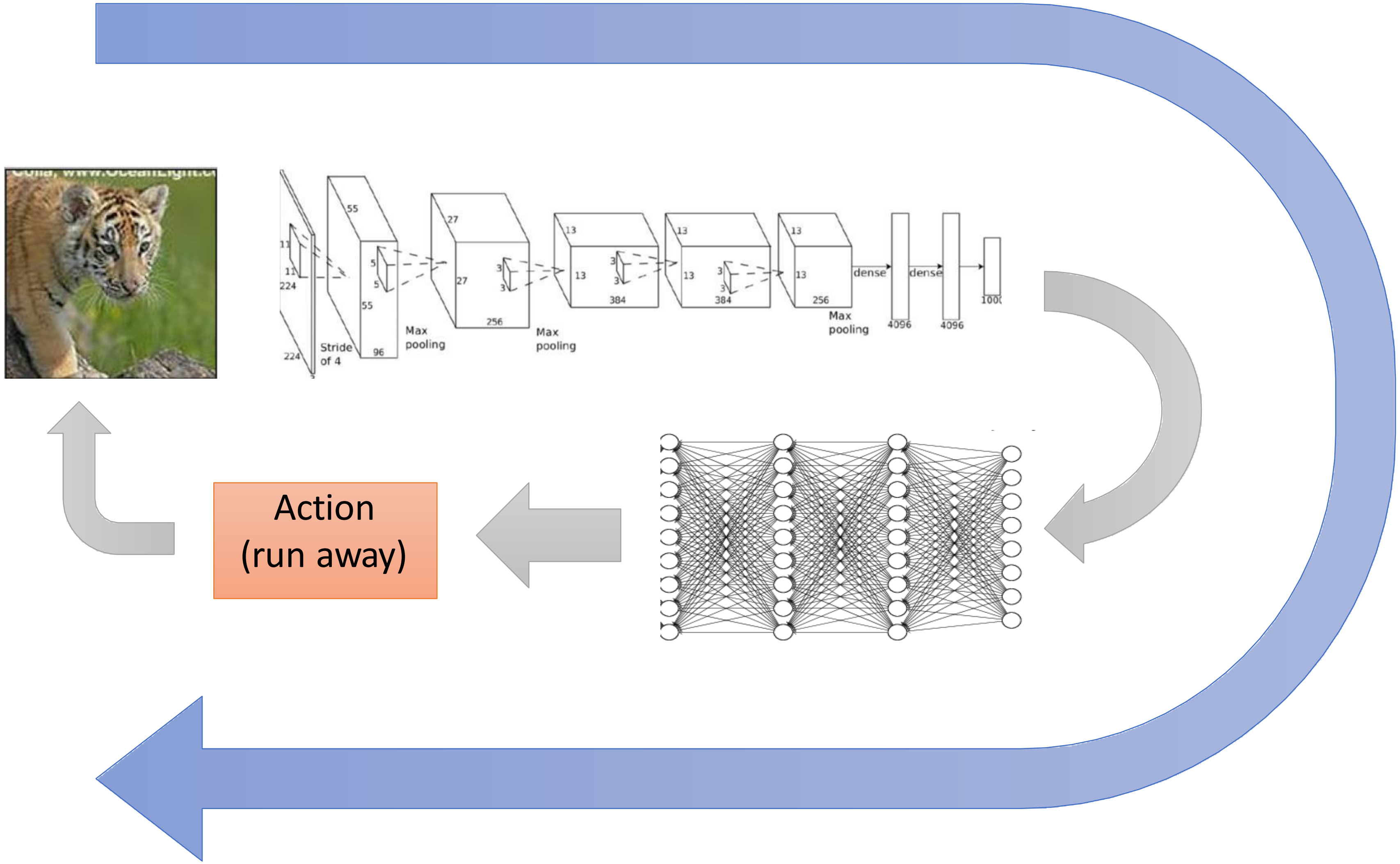
perception



Action
(run away)

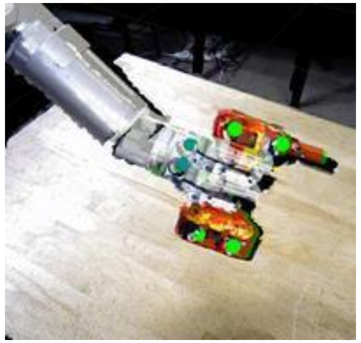
action

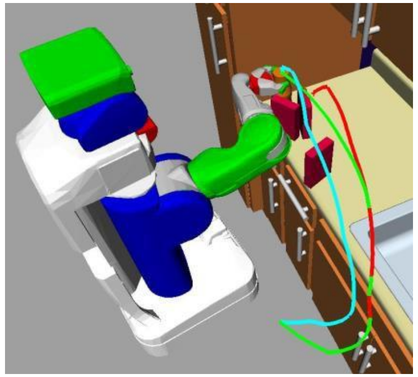
sensorimotor loop



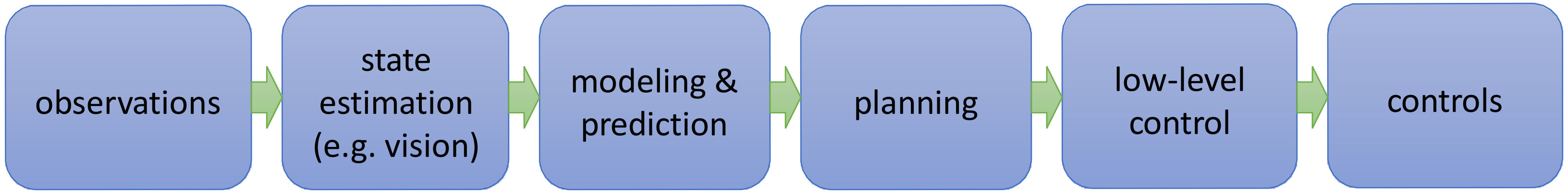
Action
(run away)

Example: robotics

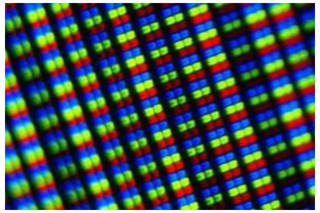


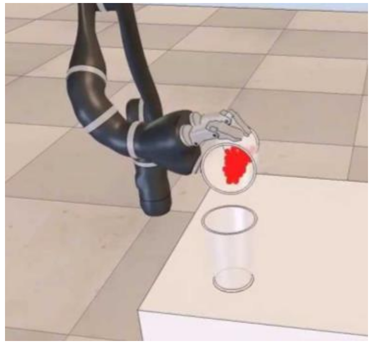


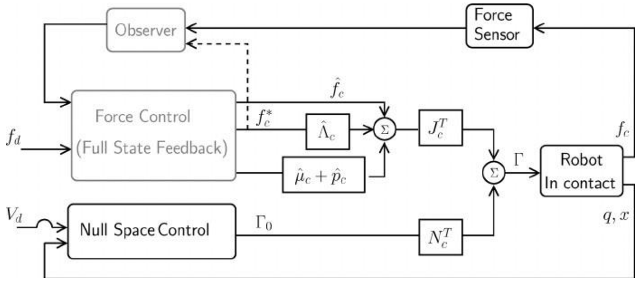
robotic
control
pipeline



observations → state estimation (e.g. vision) → modeling & prediction → planning → low-level control → controls

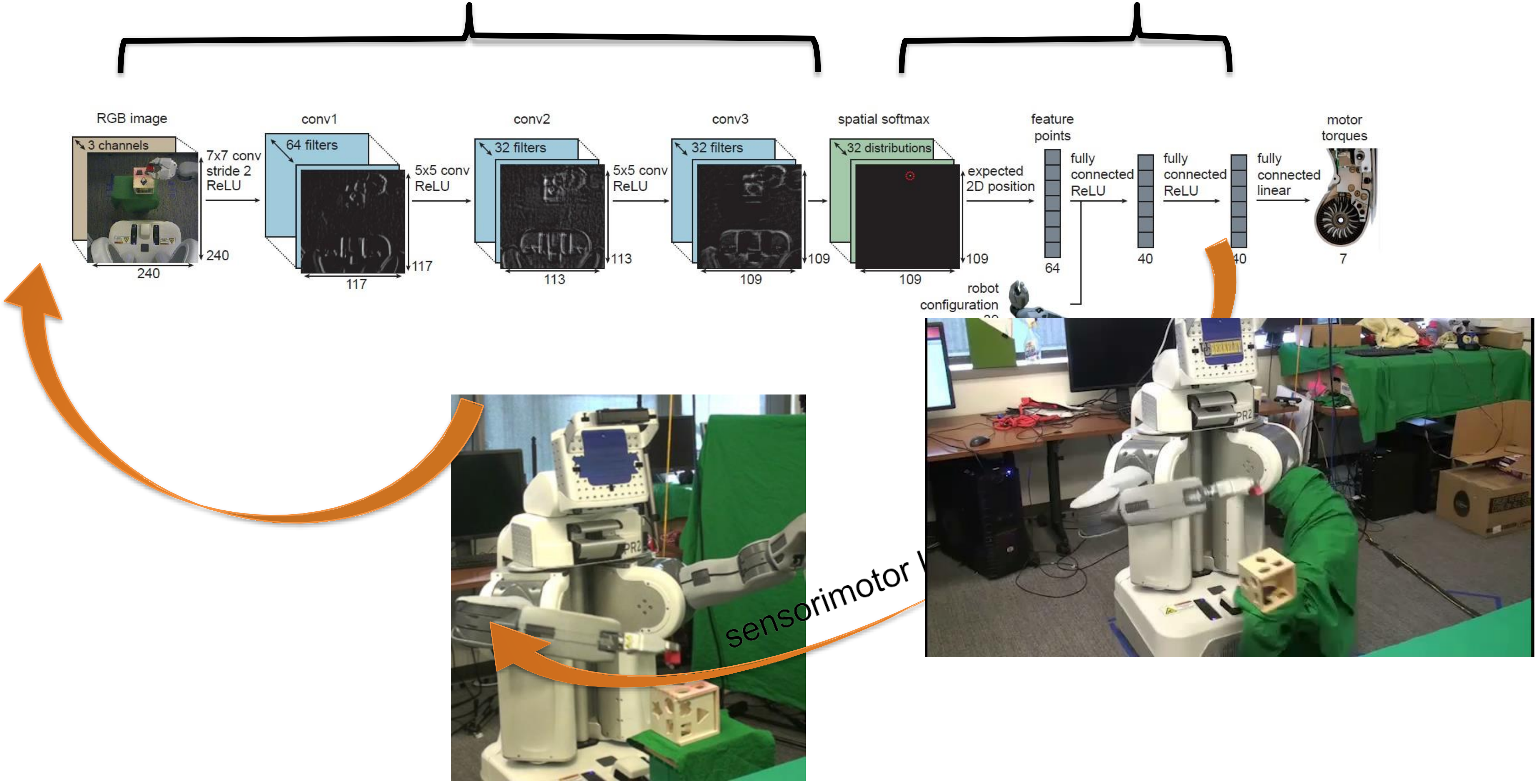


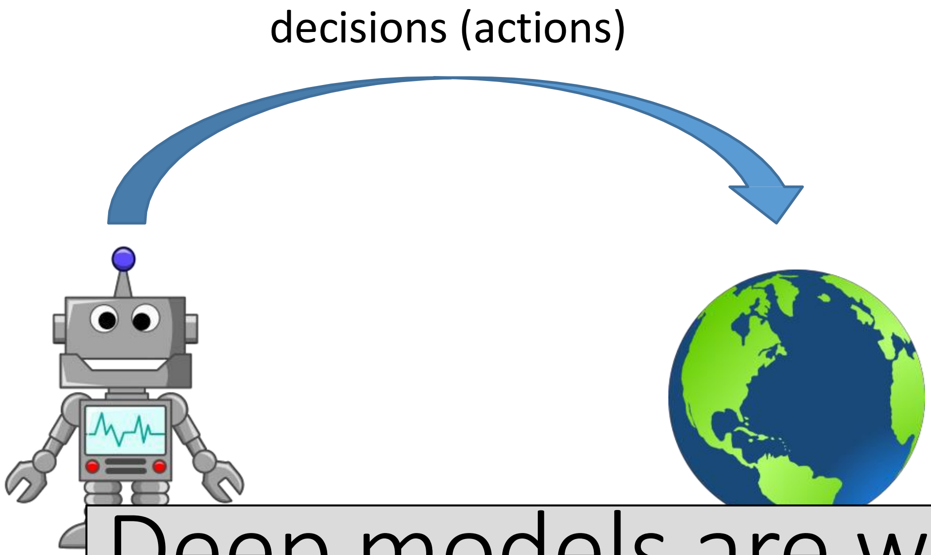




tiny, highly specialized
“visual cortex”

tiny, highly specialized
“motor cortex”





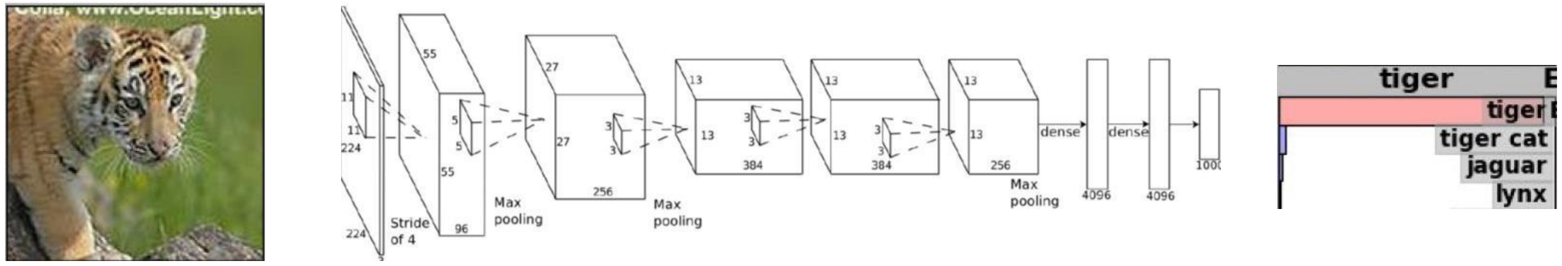
Deep models are what allow reinforcement learning algorithms to solve complex problems end to end!

g.,

The reinforcement learning problem **is** the AI problem!

Actions: what to purchase
Observations: inventory levels
Rewards: profit

Why should we study this now?



1. Advances in deep learning

2. Advances in reinforcement learning

3. Advances in computational capability

Why should we study this now?

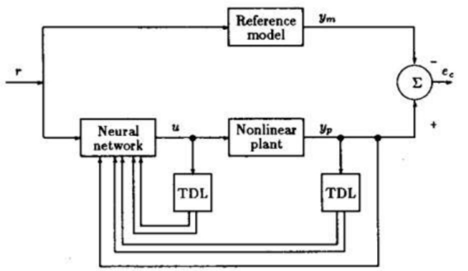
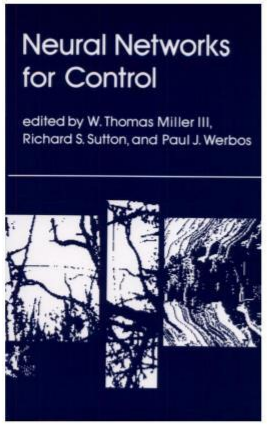
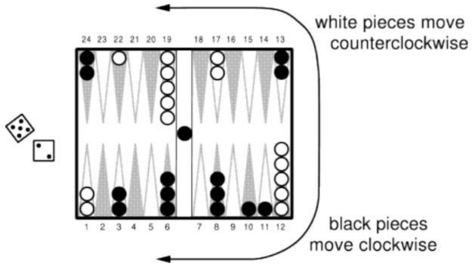
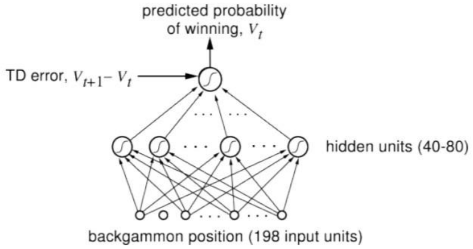


Fig. 21. Direct adaptive control of nonlinear plants using neural networks.


Neural Networks for Control
edited by W. Thomas Miller III, Richard S. Sutton, and Paul J. Werbos


24 23 22 21 20 19
18 17 16 15 14 13
1 2 3 4 5 6
7 8 9 10 11 12
white pieces move counterclockwise
black pieces move clockwise


predicted probability of winning, V_t
TD error, $V_{t+1} - V_t$
hidden units (40-80)
backgammon position (198 input units)

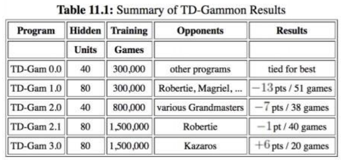
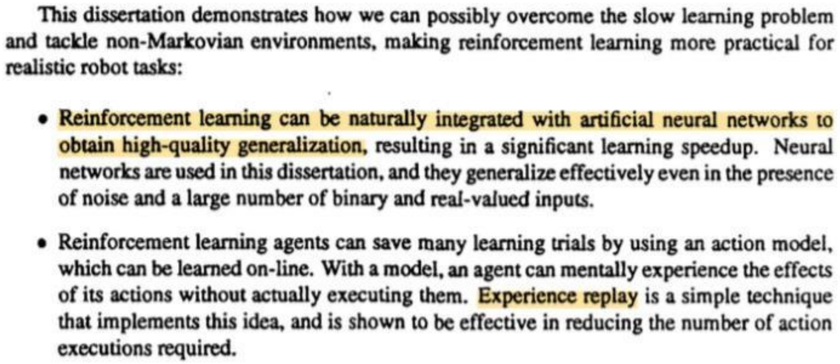


Table 11.1: Summary of TD-Gammon Results

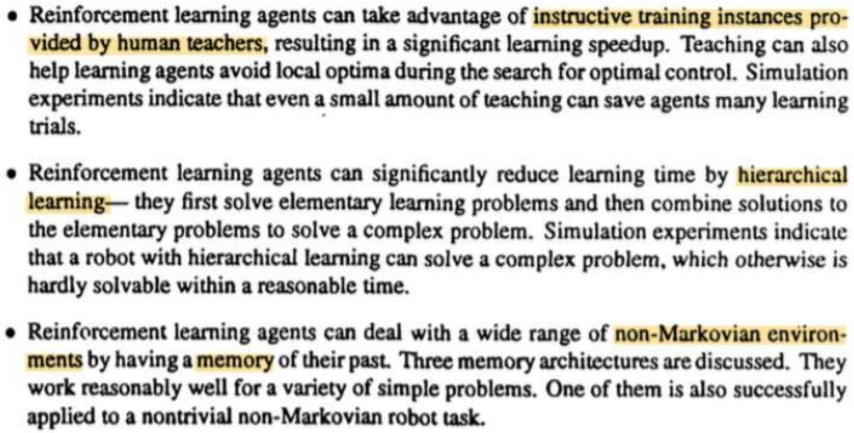
Program	Hidden	Training	Opponents	Results
	Units	Games		
TD-Gam 0.0	40	300,000	other programs	tied for best
TD-Gam 1.0	80	300,000	Robertie, Magriel, ...	−13 pts / 51 games
TD-Gam 2.0	40	800,000	various Grandmasters	−7 pts / 38 games
TD-Gam 2.1	80	1,500,000	Robertie	−1 pt / 40 games
TD-Gam 3.0	80	1,500,000	Kazaros	+6 pts / 20 games

Tesauro, 1995



This dissertation demonstrates how we can possibly overcome the slow learning problem and tackle non-Markovian environments, making reinforcement learning more practical for realistic robot tasks:

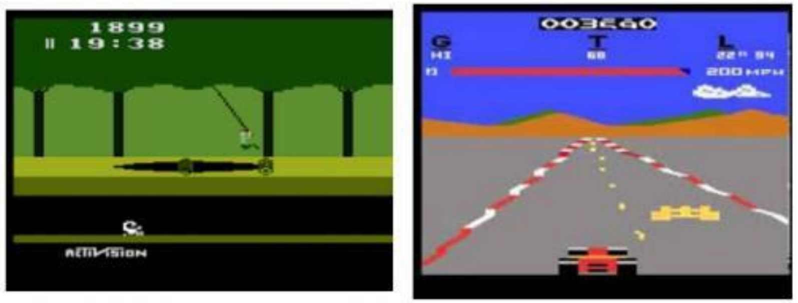
- Reinforcement learning can be naturally integrated with artificial neural networks to obtain high-quality generalization, resulting in a significant learning speedup. Neural networks are used in this dissertation, and they generalize effectively even in the presence of noise and a large number of binary and real-valued inputs.
- Reinforcement learning agents can save many learning trials by using an action model, which can be learned on-line. With a model, an agent can mentally experience the effects of its actions without actually executing them. Experience replay is a simple technique that implements this idea, and is shown to be effective in reducing the number of action executions required.



- Reinforcement learning agents can take advantage of instructive training instances provided by human teachers, resulting in a significant learning speedup. Teaching can also help learning agents avoid local optima during the search for optimal control. Simulation experiments indicate that even a small amount of teaching can save agents many learning trials.
- Reinforcement learning agents can significantly reduce learning time by hierarchical learning— they first solve elementary learning problems and then combine solutions to the elementary problems to solve a complex problem. Simulation experiments indicate that a robot with hierarchical learning can solve a complex problem, which otherwise is hardly solvable within a reasonable time.
- Reinforcement learning agents can deal with a wide range of non-Markovian environments by having a memory of their past. Three memory architectures are discussed. They work reasonably well for a variety of simple problems. One of them is also successfully applied to a nontrivial non-Markovian robot task.

L.-J. Lin, “Reinforcement learning for robots using neural networks.” 1993

Why should we study this now?



Atari games:

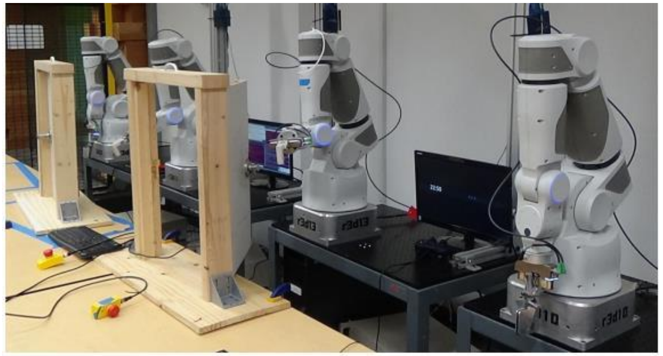
Q-learning:

I.Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, et al. “Playing Atari with Deep Reinforcement Learning”. (2013).

Policy gradients:

J.Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel. “Trust Region Policy Optimization”. (2015).

V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, et al. “Asynchronous methods for deep reinforcement learning”. (2016).



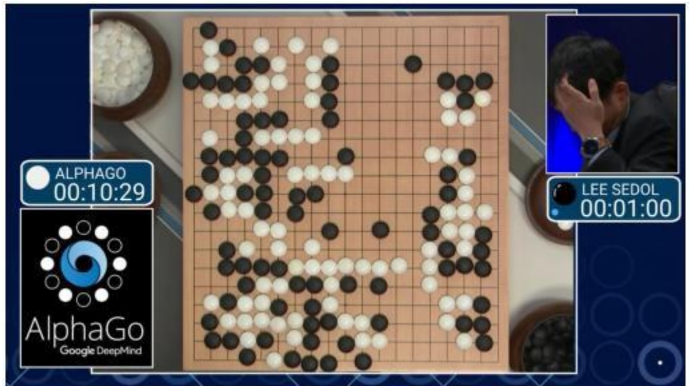
Real-world robots:

Guided policy search:

S. Levine*, C. Finn*, T. Darrell, P. Abbeel. “End-to-end training of deep visuomotor policies”. (2015).

Q-learning:

D. Kalashnikov et al. “QT-Opt: Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation”. (2018).



Beating Go champions:

Supervised learning + policy gradients + value functions + Monte Carlo tree search:

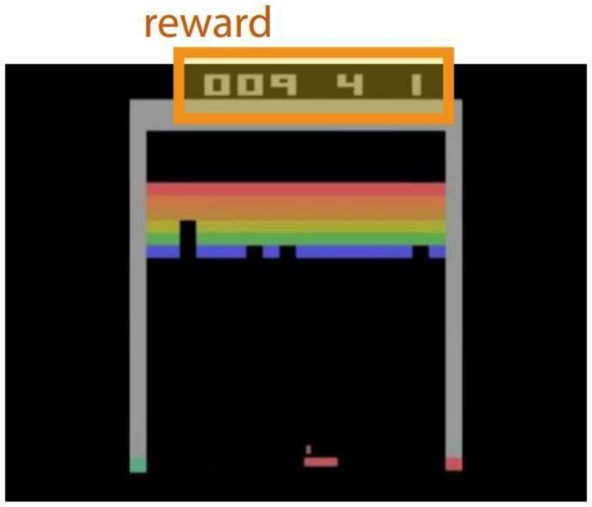
D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, et al. “Mastering the game of Go with deep neural networks and tree search”. Nature (2016).

What other problems do we need to solve to enable real-world sequential decision making?

Beyond learning from reward

- Basic reinforcement learning deals with maximizing rewards
- This is not the only problem that matters for sequential decision making!
- We will cover more advanced topics
    - Learning reward functions from example (inverse reinforcement learning)
    - Transferring knowledge between domains (transfer learning, meta-learning)
    - Learning to predict and using prediction to act

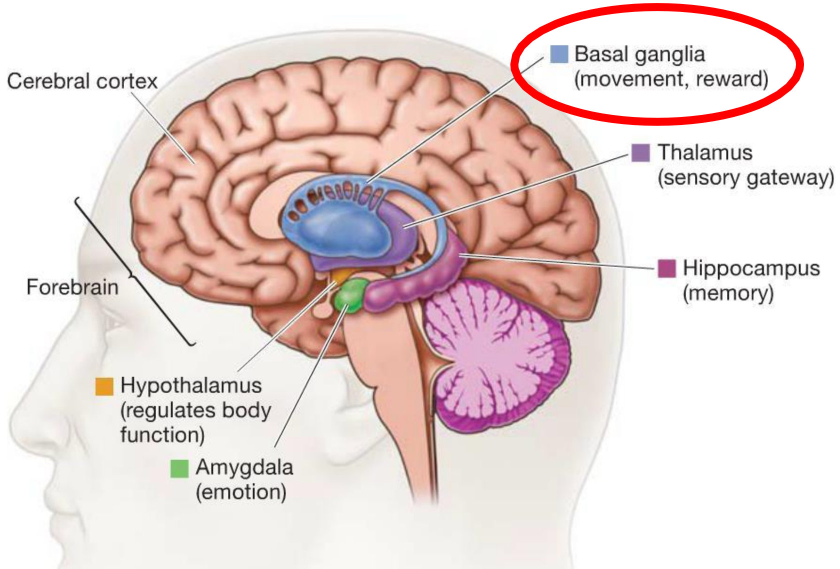
Where do rewards come from?

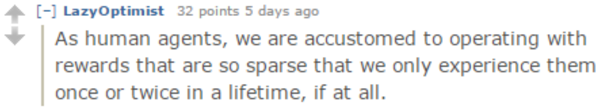


Mnih et al. '15

reinforcement learning agent

what is the reward?


Cerebral cortex
Basal ganglia (movement, reward)
Thalamus (sensory gateway)
Hippocampus (memory)
Forebrain
Hypothalamus (regulates body function)
Amygdala (emotion)



[–] LazyOptimist 32 points 5 days ago

As human agents, we are accustomed to operating with rewards that are so sparse that we only experience them once or twice in a lifetime, if at all.

Are there other forms of supervision?

- Learning from demonstrations
  - Directly copying observed behavior
  - Inferring rewards from observed behavior (inverse reinforcement learning)
- Learning from observing the world
  - Learning to predict
  - Unsupervised learning
- Learning from other tasks
  - Transfer learning
  - Meta-learning: learning to learn

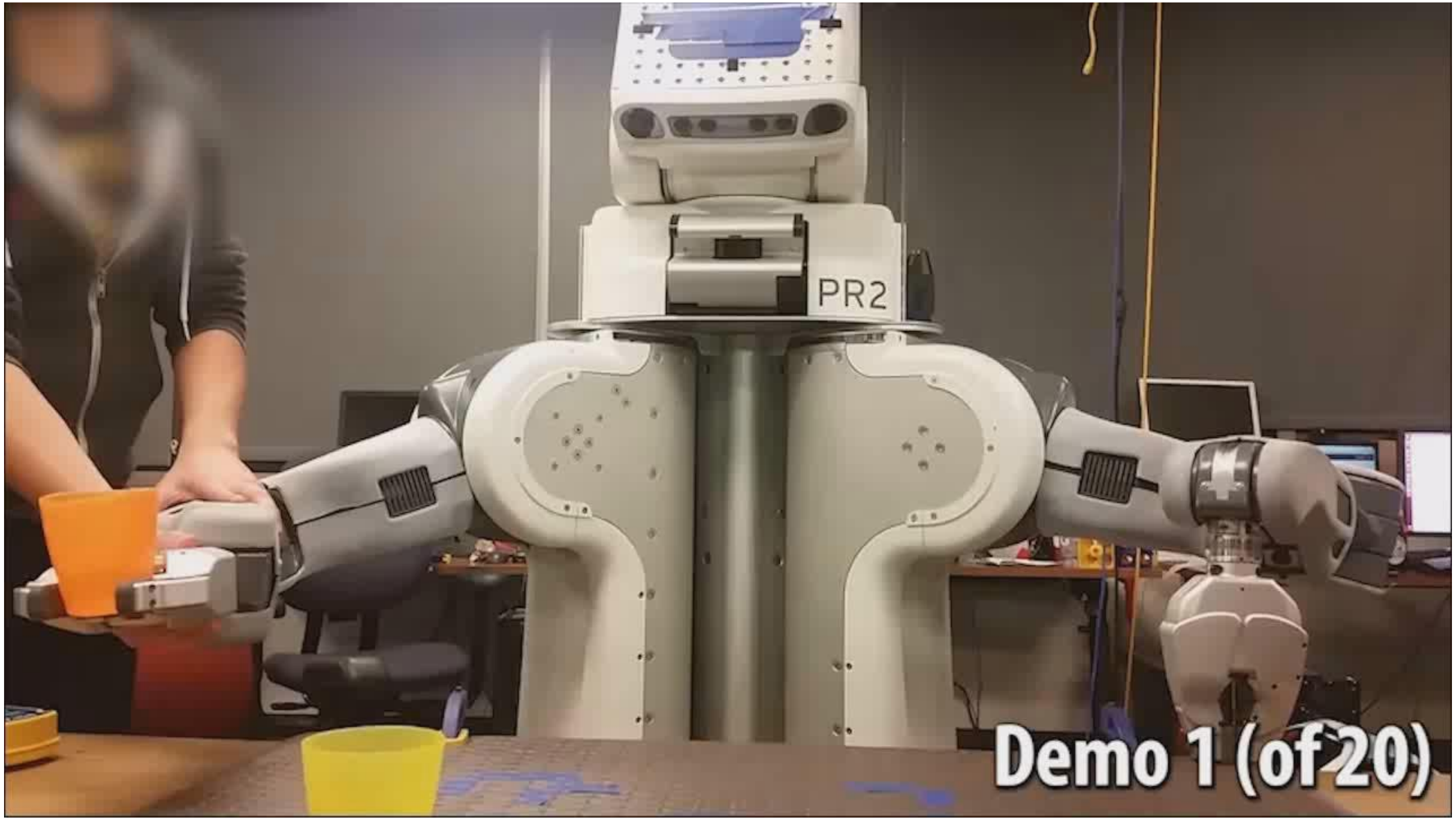
Imitation learning

Bojarski et al. 2016

More than imitation: inferring intentions

Warneken & Tomasello

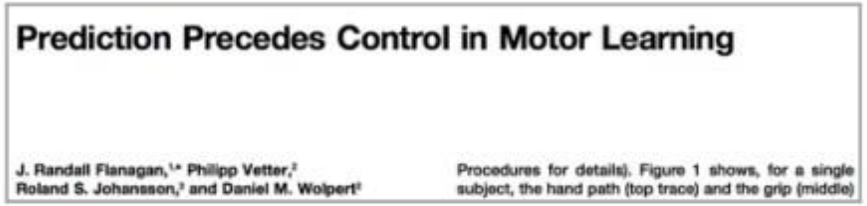
Inverse RL examples



Finn et al. 2016

Prediction

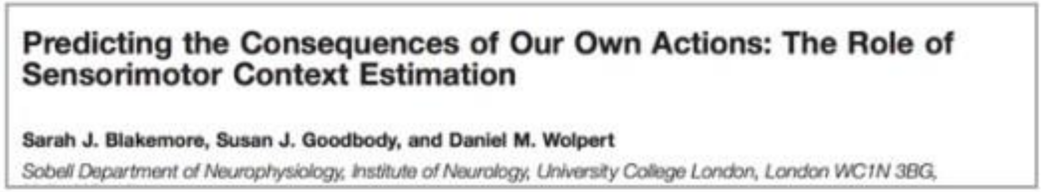
"the idea that we **predict the consequences of our motor commands** has emerged as an important theoretical concept in all aspects of sensorimotor control"



Prediction Precedes Control in Motor Learning

J. Randall Flanagan,[1,*] Philipp Vetter,[2] Roland S. Johansson,[3] and Daniel M. Wolpert[2]

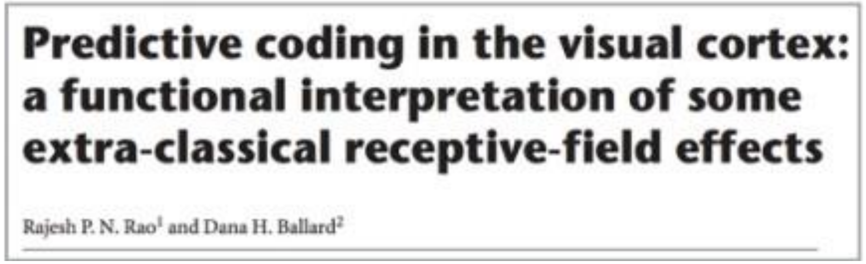
Procedures for details). Figure 1 shows, for a single subject, the hand path (top trace) and the grip (middle)



Predicting the Consequences of Our Own Actions: The Role of Sensorimotor Context Estimation

Sarah J. Blakemore, Susan J. Goodbody, and Daniel M. Wolpert

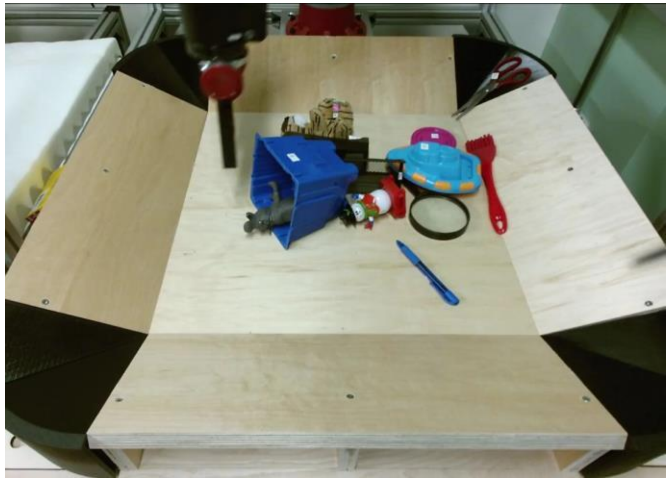
Sobell Department of Neurophysiology, Institute of Neurology, University College London, London WC1N 3BG,

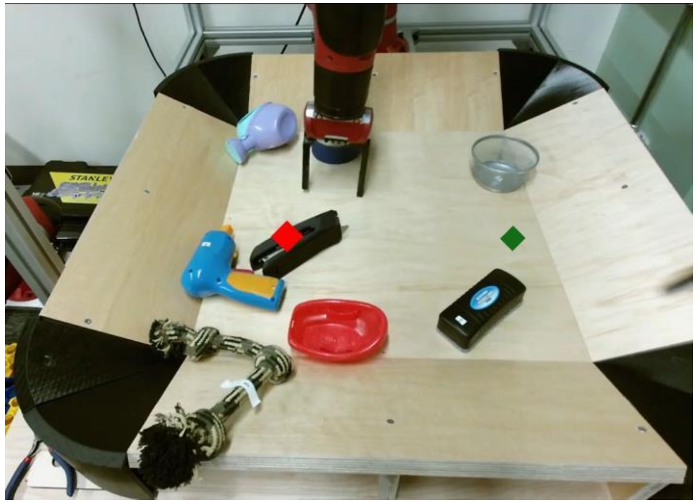


Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects

Rajesh P. N. Rao[1] and Dana H. Ballard[2]

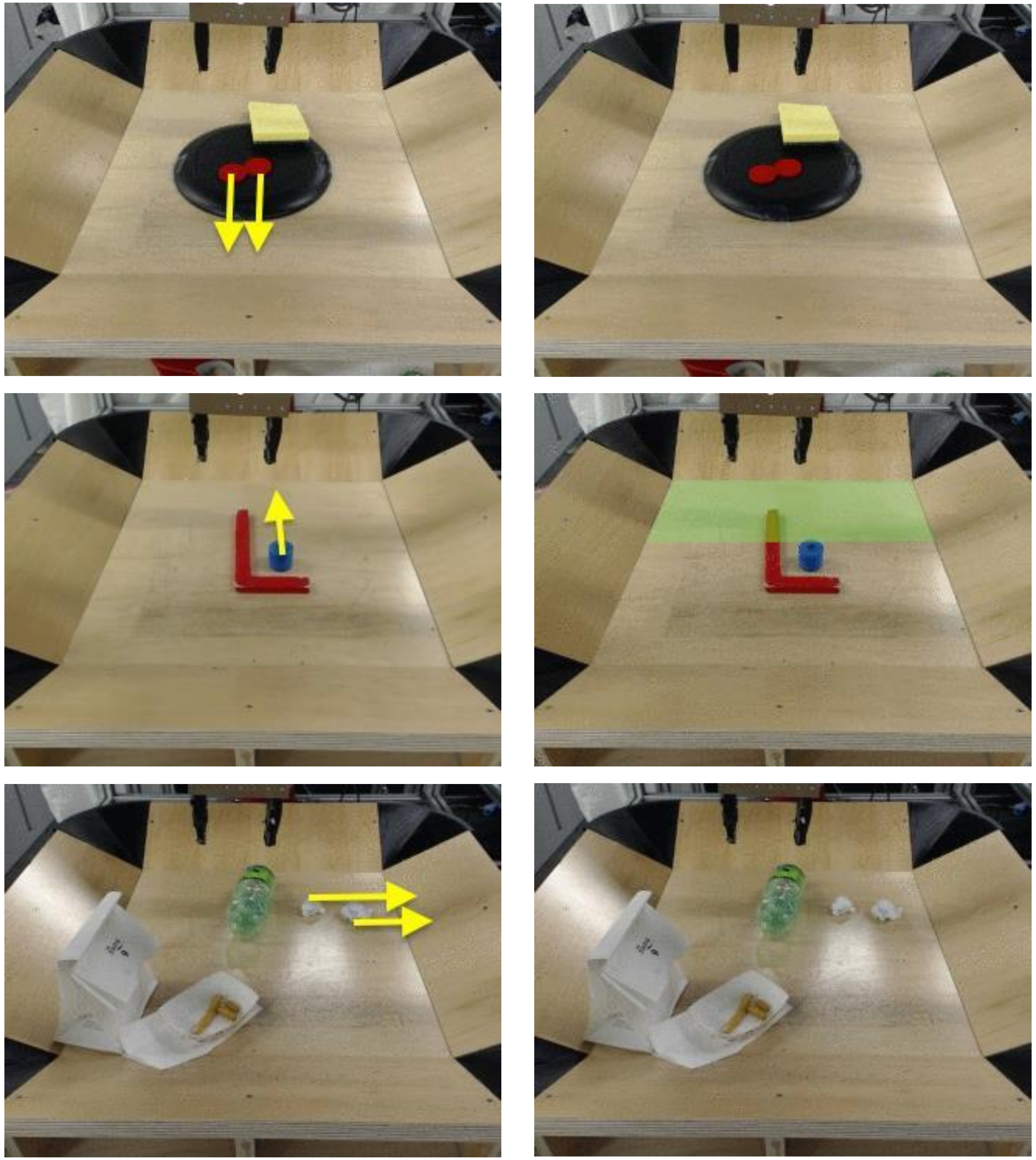
Prediction for real-world control





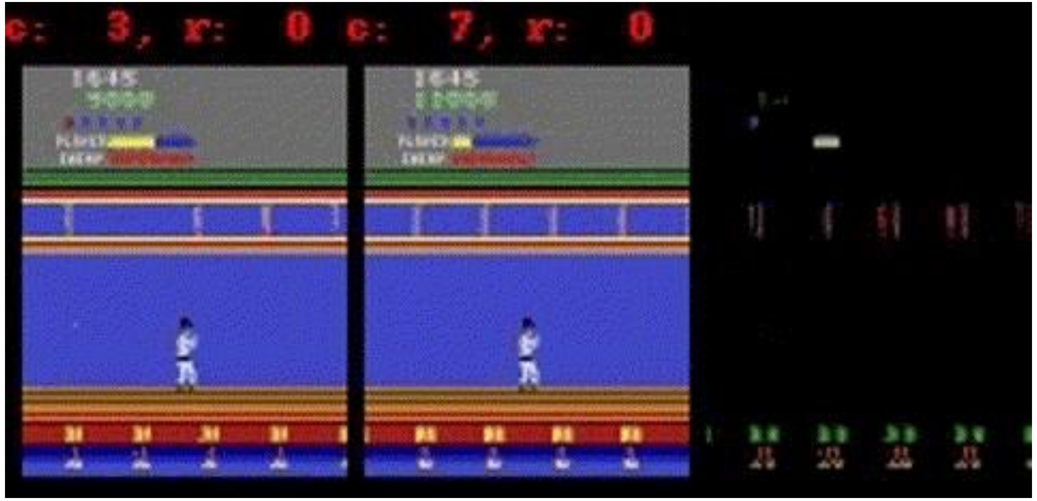
Ebert et al. 2017

Using tools with predictive models



Xie et al. 2019

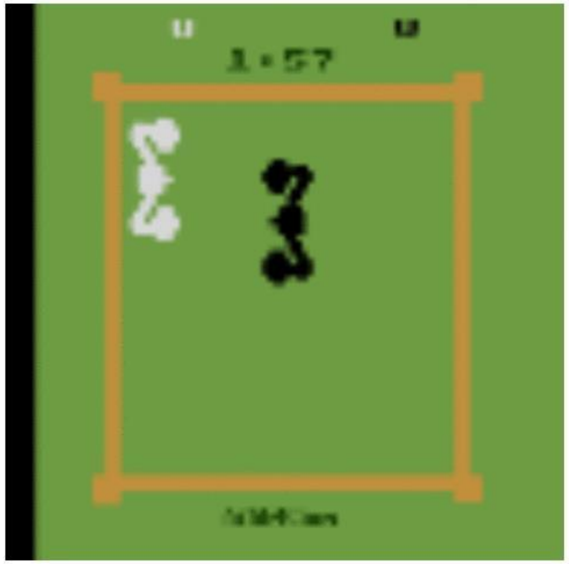
Playing games with predictive models



predicted

real

But sometimes there are issues…

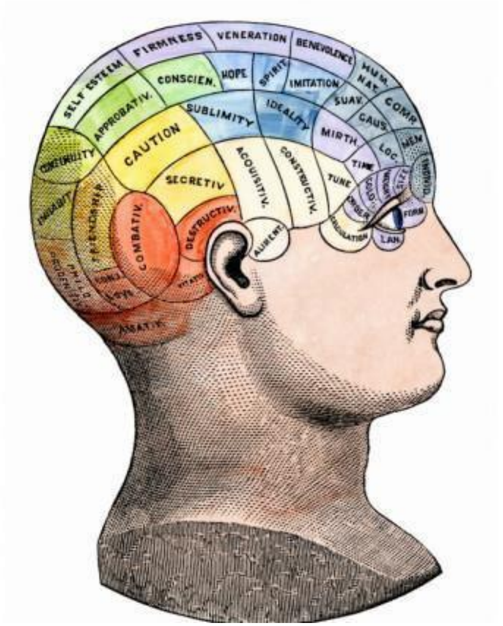


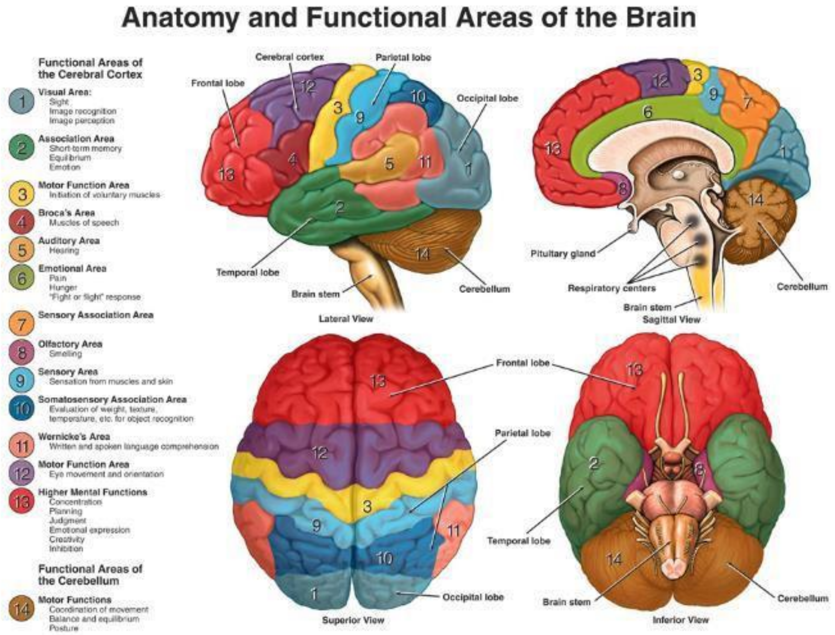
Kaiser et al. 2019

How do we build intelligent machines?

How do we build intelligent machines?

- Imagine you have to build an intelligent machine, where do you start?



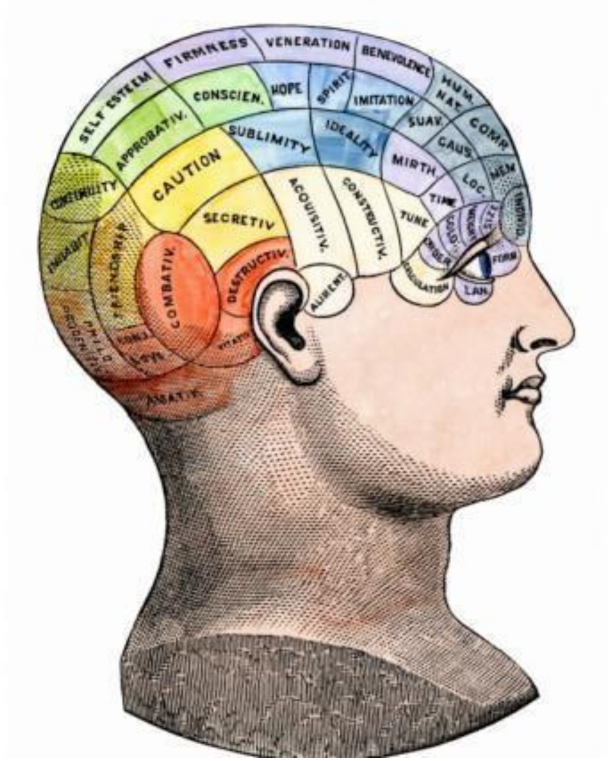

Anatomy and Functional Areas of the Brain
Functional Areas of the Cerebral Cortex
1 Visual Area: Sight, Image recognition, Image perception
2 Association Area: Short-term memory, Equilibrium, Emotion
3 Motor Function Area: Initiation of voluntary muscles
4 Broca's Area: Muscles of speech
5 Auditory Area: Hearing
6 Emotional Area: Pain, Hunger, "Fight or flight" response
7 Sensory Association Area
8 Olfactory Area: Smelling
9 Sensory Area: Sensation from muscles and skin
10 Somatosensory Association Area: Evaluation of weight, texture, temperature, etc. for object recognition
11 Wernicke's Area: Written and spoken language comprehension
12 Motor Function Area: Eye movement and orientation
13 Higher Mental Functions: Concentration, Planning, Judgment, Emotional expression, Creativity, Inhibition
Functional Areas of the Cerebellum
14 Motor Functions: Coordination of movement, Balance and equilibrium, Posture
Cerebral cortex
Parietal lobe
Frontal lobe
Occipital lobe
Temporal lobe
Brain stem
Cerebellum
Lateral View
Pituitary gland
Respiratory centers
Brain stem
Cerebellum
Sagittal View
Frontal lobe
Parietal lobe
Temporal lobe
Occipital lobe
Superior View
Brain stem
Cerebellum
Inferior View

Learning as the basis of intelligence

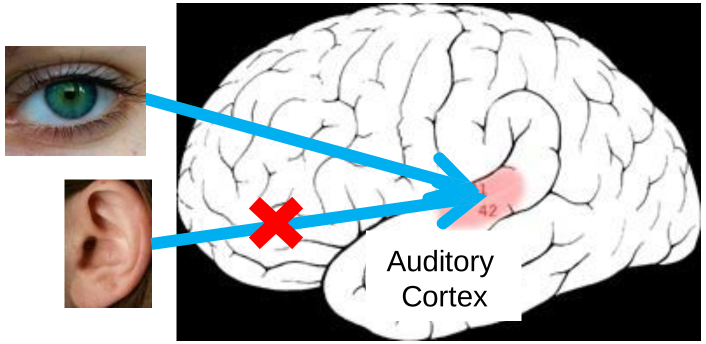
- Some things we can all do (e.g. walking)
- Some things we can only learn (e.g. driving a car)
- We can learn a huge variety of things, including very difficult things
- Therefore our learning mechanism(s) are likely powerful enough to do everything we associate with intelligence
    - But it may still be very convenient to “hard-code” a few really important bits

A single algorithm?

- An algorithm for each “module”?
- Or a single flexible algorithm?



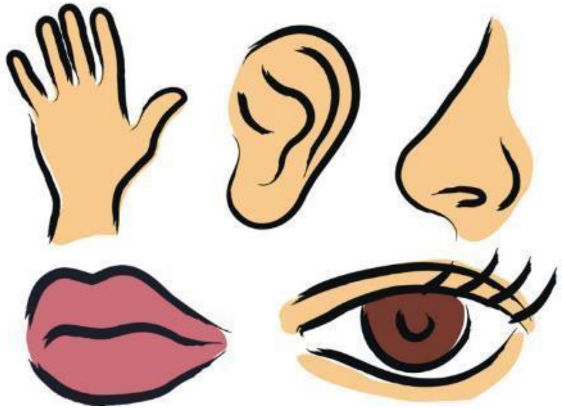
Seeing with your tongue



[BrainPort; Martinez et al; Roe et al.]
adapted from A. Ng

What must that single algorithm do?

- Interpret rich sensory inputs

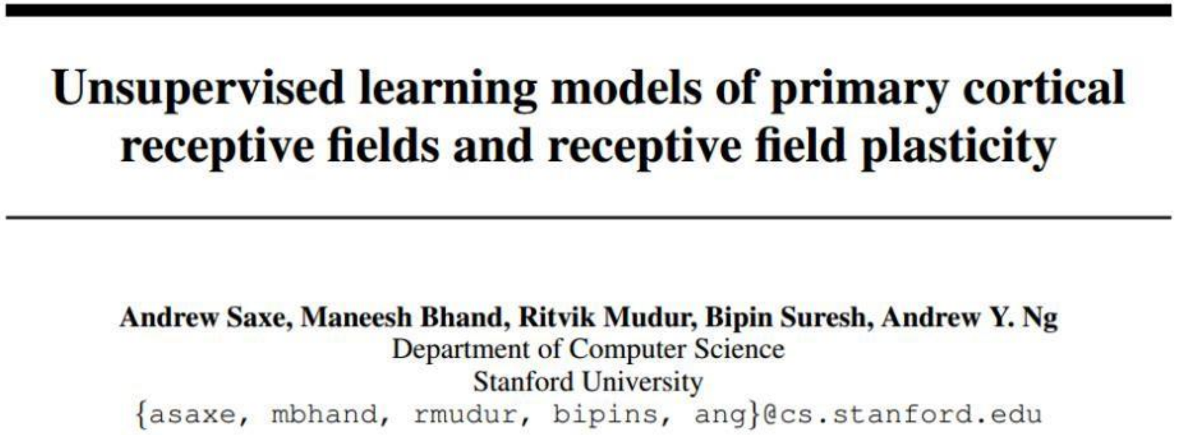


- Choose complex actions

Why deep reinforcement learning?

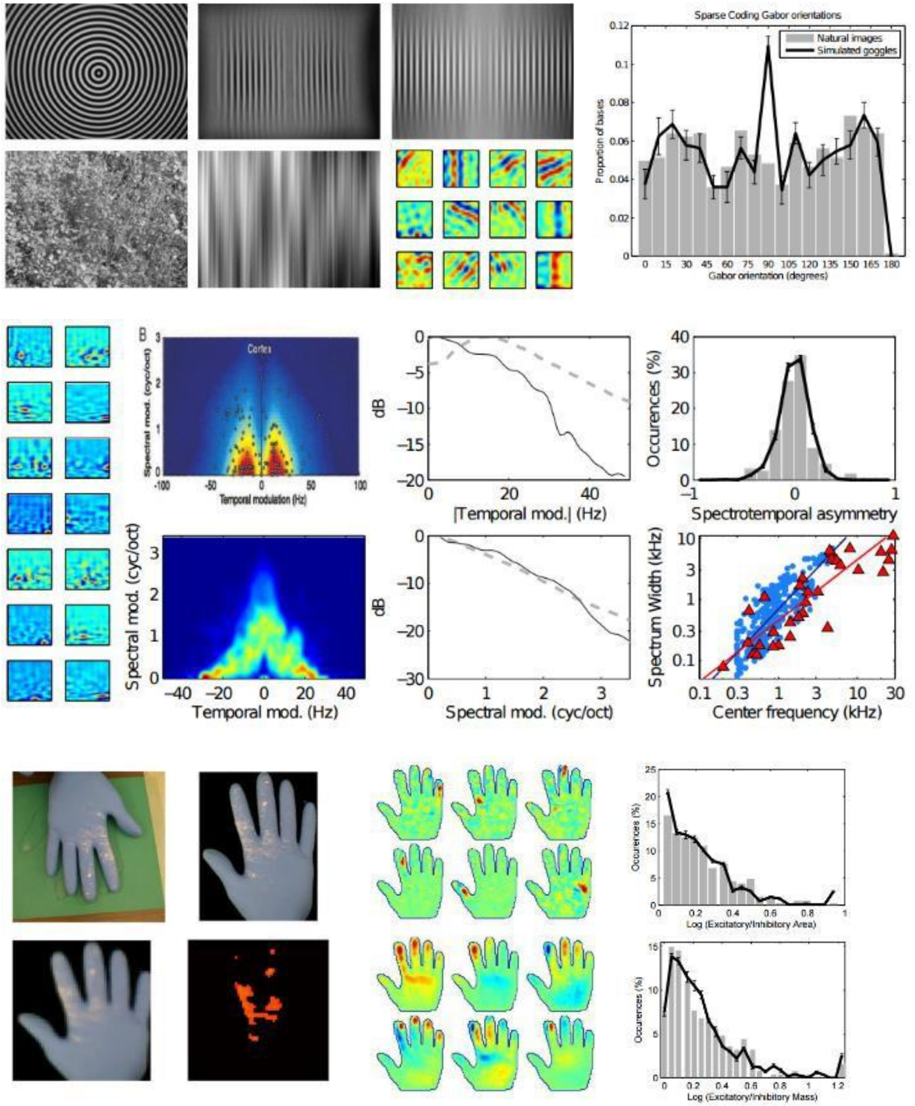
- Deep = can process complex sensory input
    - …and also compute really complex functions
- Reinforcement learning = can choose complex actions

Some evidence in favor of deep learning



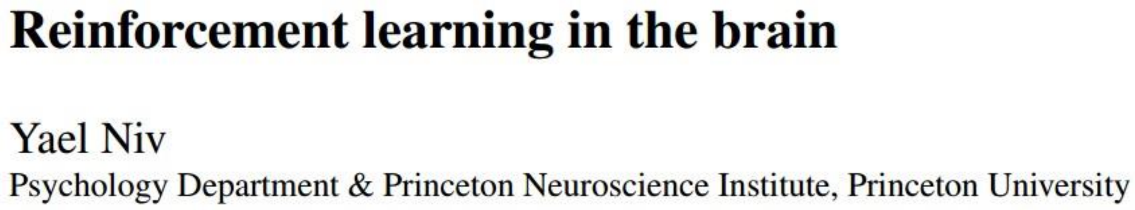
Unsupervised learning models of primary cortical receptive fields and receptive field plasticity

Andrew Saxe, Maneesh Bhand, Ritvik Mudur, Bipin Suresh, Andrew Y. Ng
Department of Computer Science
Stanford University
`{asaxe, mbhand, rmudur, bipins, ang}@cs.stanford.edu`



Some evidence for reinforcement learning

- Percepts that anticipate reward become associated with similar firing patterns as the reward itself
- Basal ganglia appears to be related to reward system
- Model-free RL-like adaptation is often a good fit for experimental data of animal adaptation
  - But not always…



Reinforcement learning in the brain

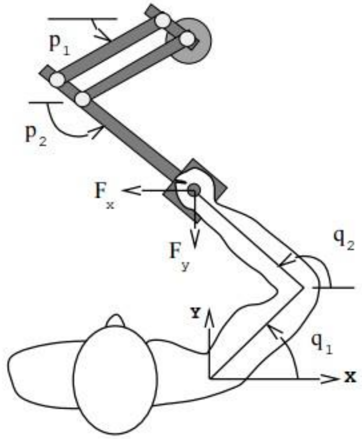
Yael Niv
Psychology Department & Princeton Neuroscience Institute, Princeton University

What can deep learning & RL do well now?

- Acquire high degree of proficiency in domains governed by simple, known rules
- Learn simple skills with raw sensory inputs, given enough experience
- Learn from imitating enough human-provided expert behavior

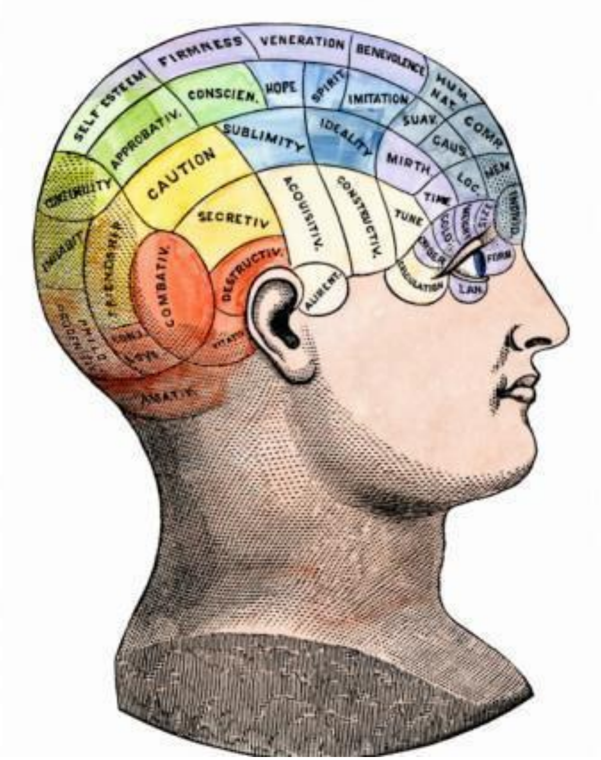
What has proven challenging so far?

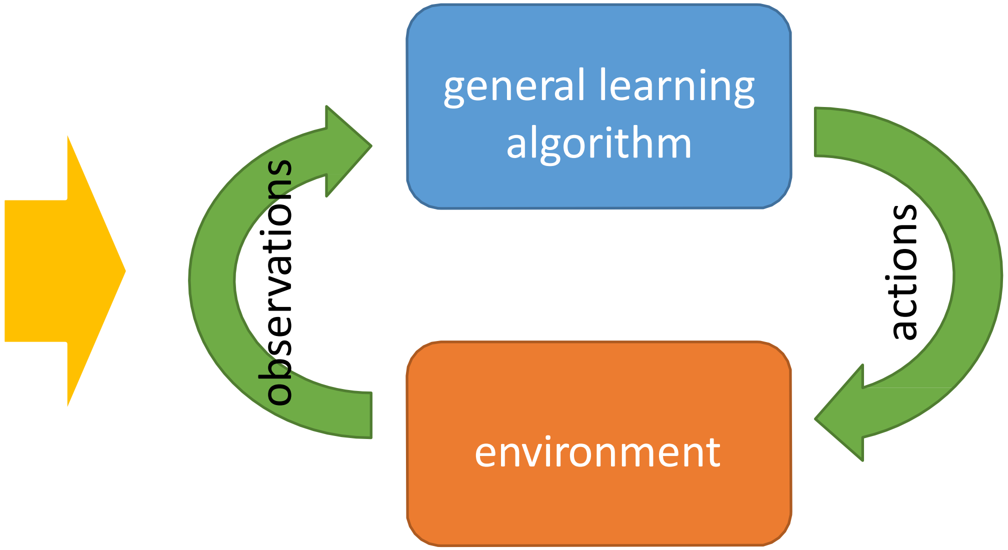
- Humans can learn incredibly quickly
  - Deep RL methods are usually slow
- Humans can reuse past knowledge
  - Transfer learning in deep RL is an open problem
- Not clear what the reward function should be
- Not clear what the role of prediction should be



Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain.

- Alan Turing




general learning algorithm
actions
environment
observations