Carnegie Mellon

School of Computer Science

Deep Reinforcement Learning and Control

Introduction

Spring 2023, CMU 10-403

Katerina Fragkiadaki

Course Logistics

- Course website: https://cmudeeprl.github.io/403website_s23/ all you need to know
- Grading:
  - 5 homework assignments: implementation and question/answering 55%
  - 3 quizzes - 45%
- Resources: AWS for those that do not have access to GPUs
- People can unofficially audit the course
- The readings on the schedule are **required** unless noted otherwise

Overview for today

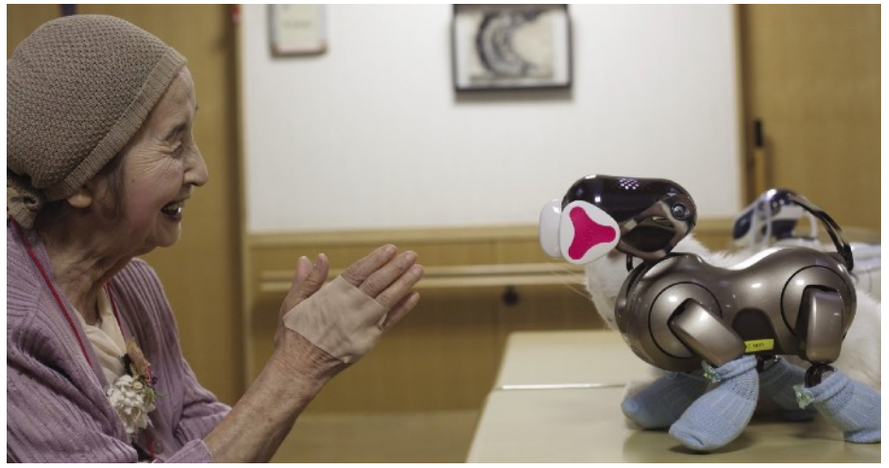
- Goal of the course / why it is important
- What is reinforcement learning
- What is representation learning; how it relates to reinforcement learning
- Reinforcement learning versus supervised learning
- AI’s paradox: what is hard and what is easy for machines to learn

Overview for today

- Goal of the course / why it is important
- What is reinforcement learning
- What is representation learning (and how it helps reinforcement learning and behavior learning in general)
- Reinforcement learning versus supervised learning
- AI's paradox: what is hard and what is easy in behavior learning

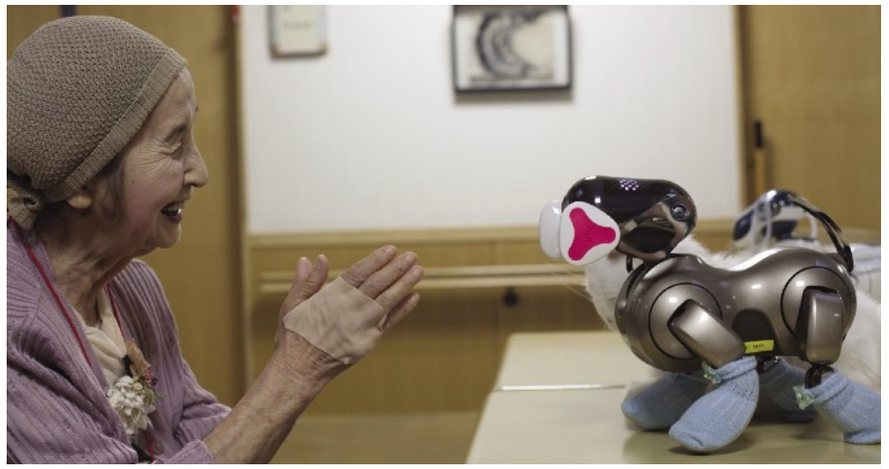
Goal of the course: Learning to act

Building agents that **learn** to act and accomplish **goals** in **dynamic** environments



Goal of the course: Learning to act

Building agents that **learn** to act and accomplish **goals** in **dynamic** environments



…as opposed to agents that execute **pre-programmed** behaviors in **static** environments…

Motion and Action are important

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

Daniel Wolpert

https://www.ted.com/talks/daniel_wolpert_the_real_reason_for_brains?language=en

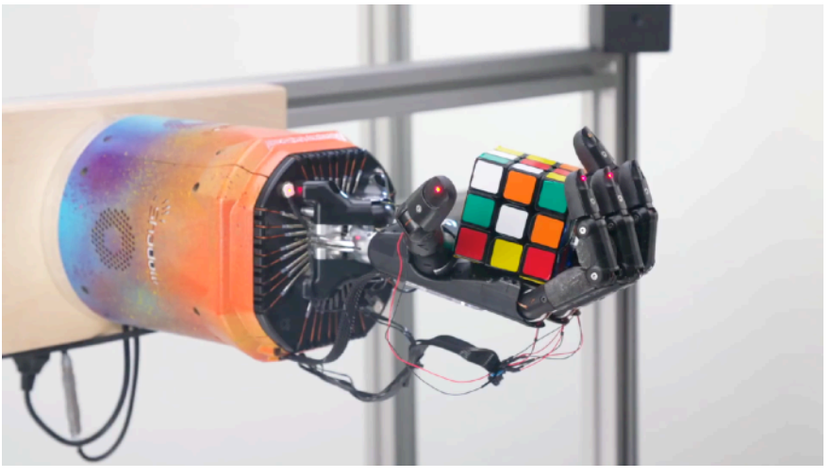
Motion and Action are important

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

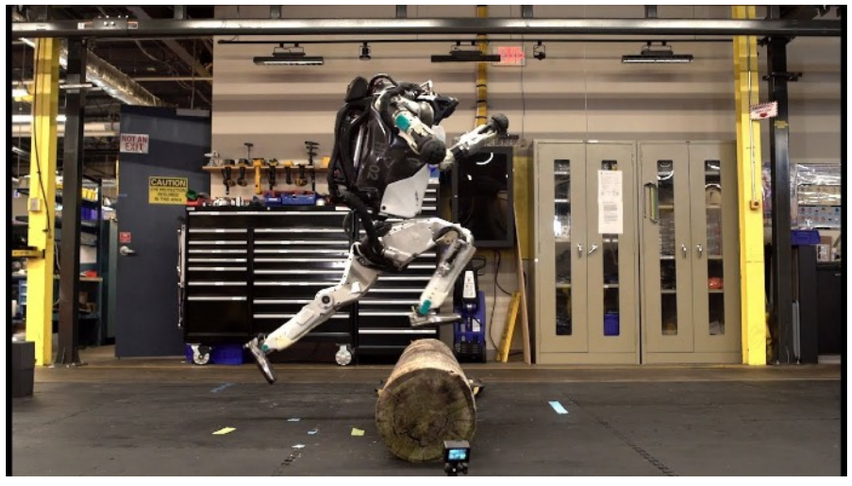
Daniel Wolpert

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

Where are we today?

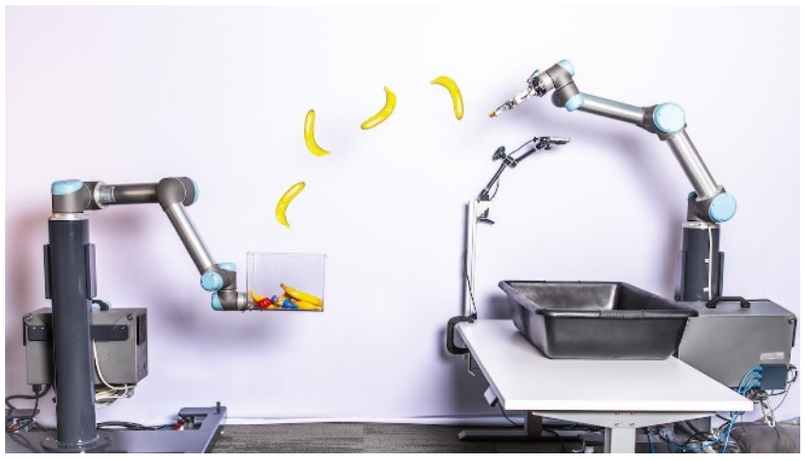


openAI



Atlas, BostonDynamics

Amazon



TossingBot, google

Tesla self-driving car

DeepMind, AlphaGo, muzero

Overview for today

- Goal of the course / why it is important
- What is reinforcement learning
- What is representation learning (and how it helps reinforcement learning and behavior learning in general)
- Reinforcement learning versus supervised learning
- AI's paradox: what is hard and what is easy in behavior learning

Reinforcement Learning (RL): How behaviors are shaped

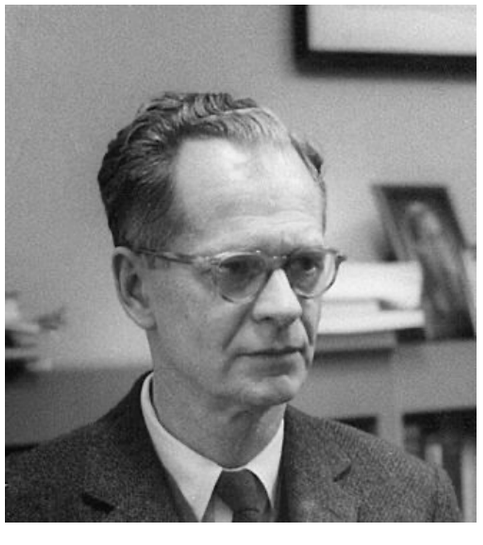
Behavior is primarily shaped by reinforcement rather than free-will.

B.F. Skinner
1904-1990
Harvard psychology

- behaviors that result in praise/pleasure tend to repeat,
- behaviors that result in punishment/pain tend to become extinct.

How are behaviors shaped?

Behavior is primarily shaped by reinforcement rather than free-will.



B.F. Skinner
1904-1990
Harvard psychology

- behaviors that result in praise/pleasure tend to repeat,
- behaviors that result in punishment/pain tend to become extinct.

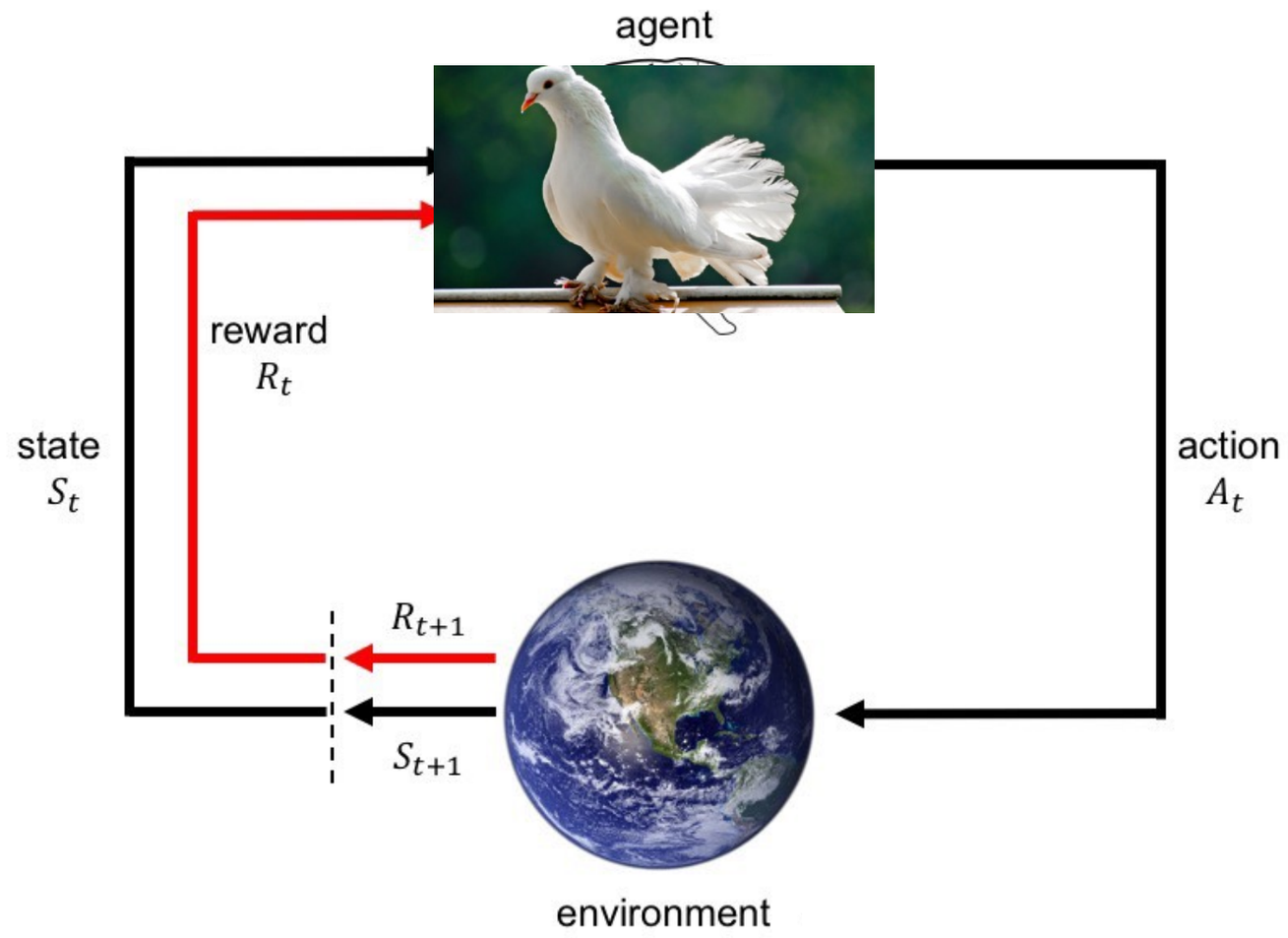
Reinforcement Learning (RL): How behaviors are shaped

B.F. Skinner
1904-1990
Harvard psychology

https://www.youtube.com/watch?v=yhvaSEJtOV8

Reinforcement learning = trial-and-error learning

Learning policies that maximize a reward function by interacting with the world



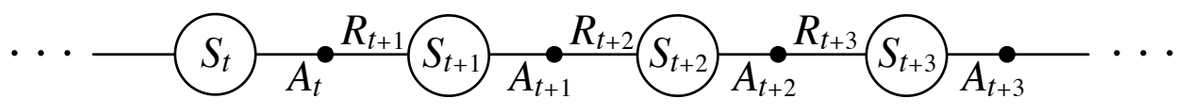
Agent and environment interact at discrete time steps: $t = 0, 1, 2, \mathrm{K}$

Agent observes state at step t: $S_t \in \mathcal{S}$

produces action at step t : $A_t \in \mathcal{A}(S_t)$

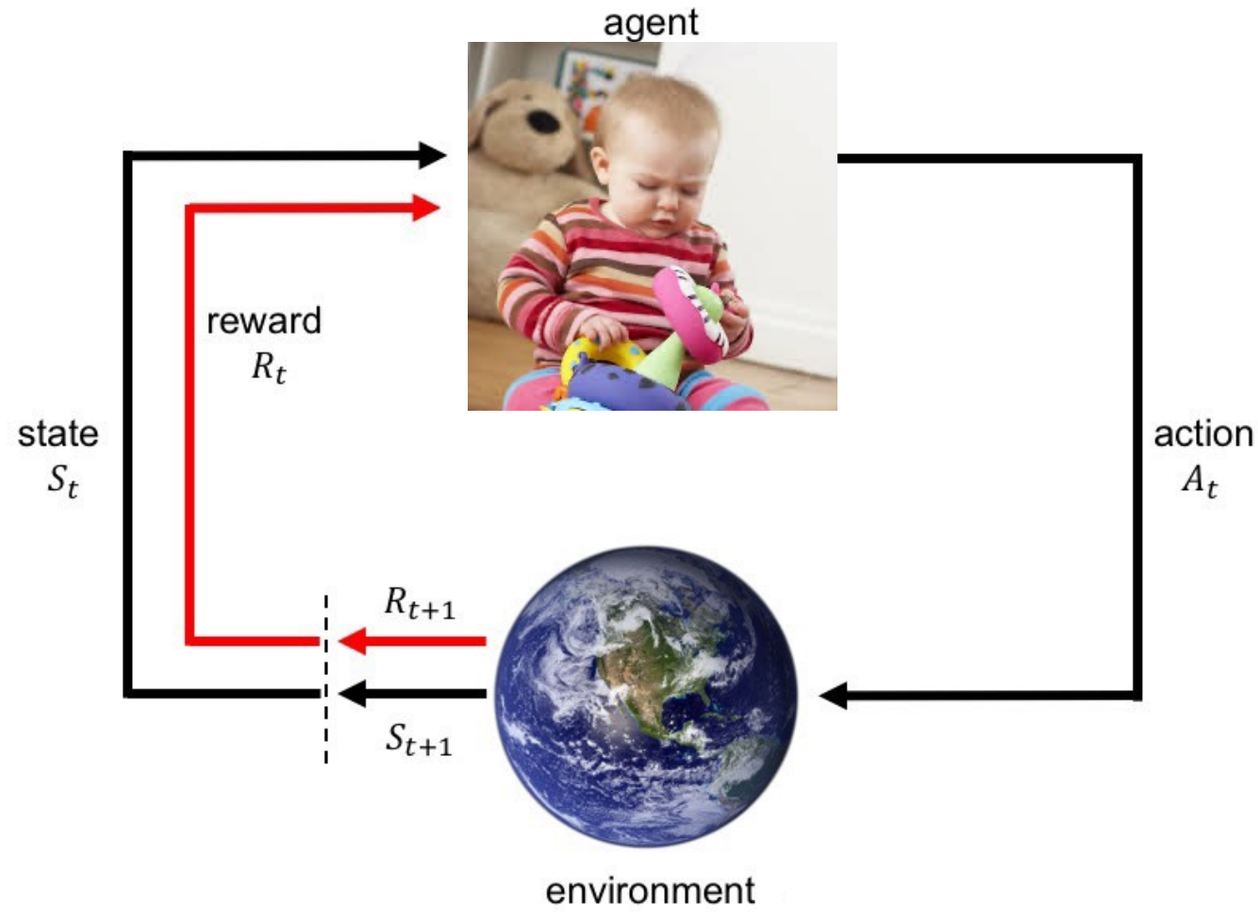
gets resulting reward: $R_{t+1} \in \mathbb{R}$

and resulting next state: $S_{t+1} \in \mathcal{S}^+$



Reinforcement learning

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



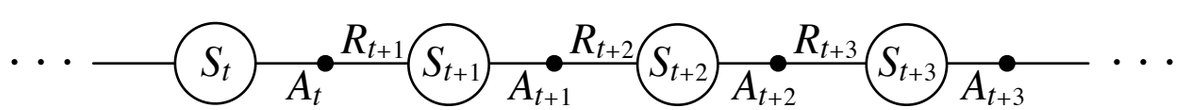
Agent and environment interact at discrete time steps: $t = 0, 1, 2, \mathrm{K}$

Agent observes state at step t: $S_t \in \mathcal{S}$

produces action at step t : $A_t \in \mathcal{A}(S_t)$

gets resulting reward: $R_{t+1} \in \mathbb{R}$

and resulting next state: $S_{t+1} \in \mathcal{S}^+$



Reinforcement learning

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

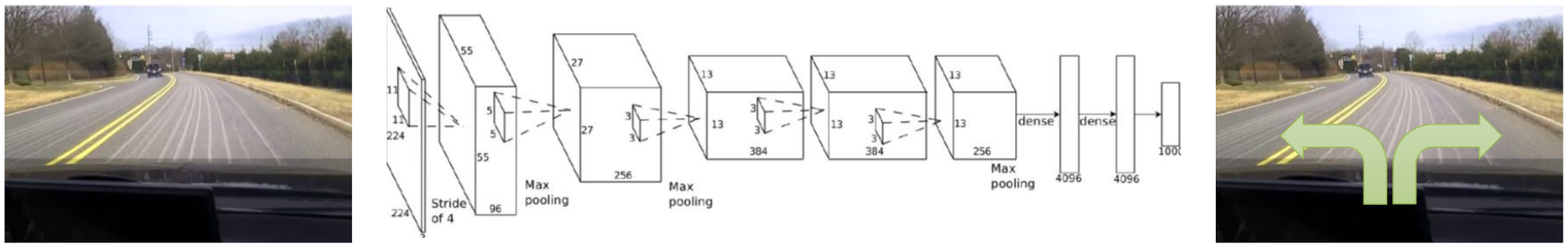
https://youtu.be/8vNxjwt2AqY

No food shows up but the baby keeps exploring

Policy

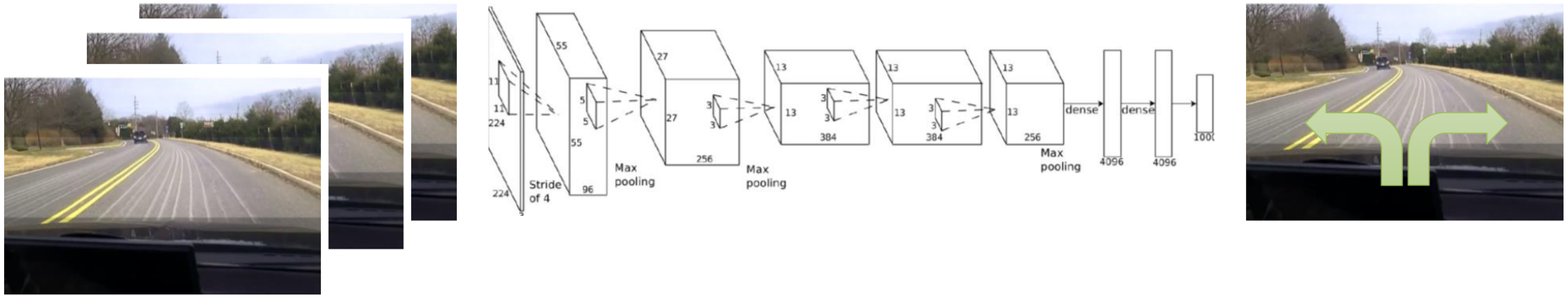
A mapping function from states and goal to actions.

$$\pi(a \mid s, g) = \mathbb{P}[A_t = a \mid S_t = s, G = g]$$



Why deep reinforcement learning?

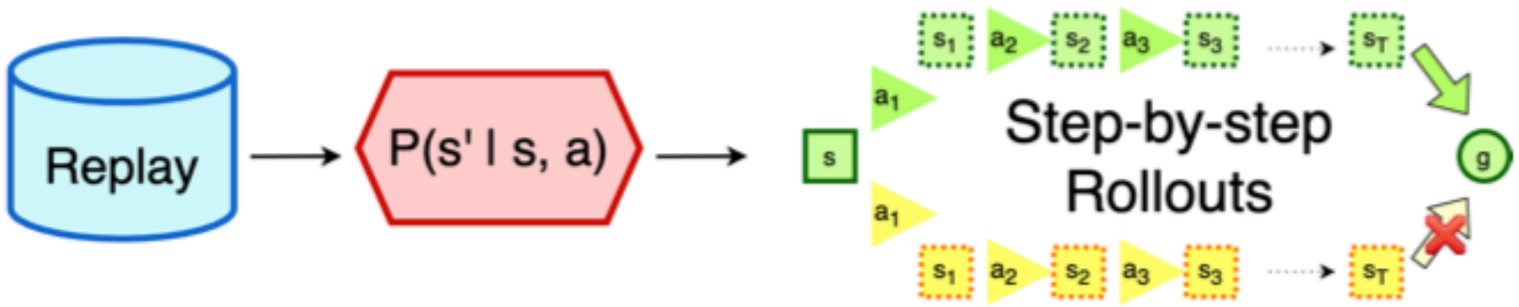
Because the policy, the model and the value functions (expected returns) will often be represented by some form of a deep neural network.



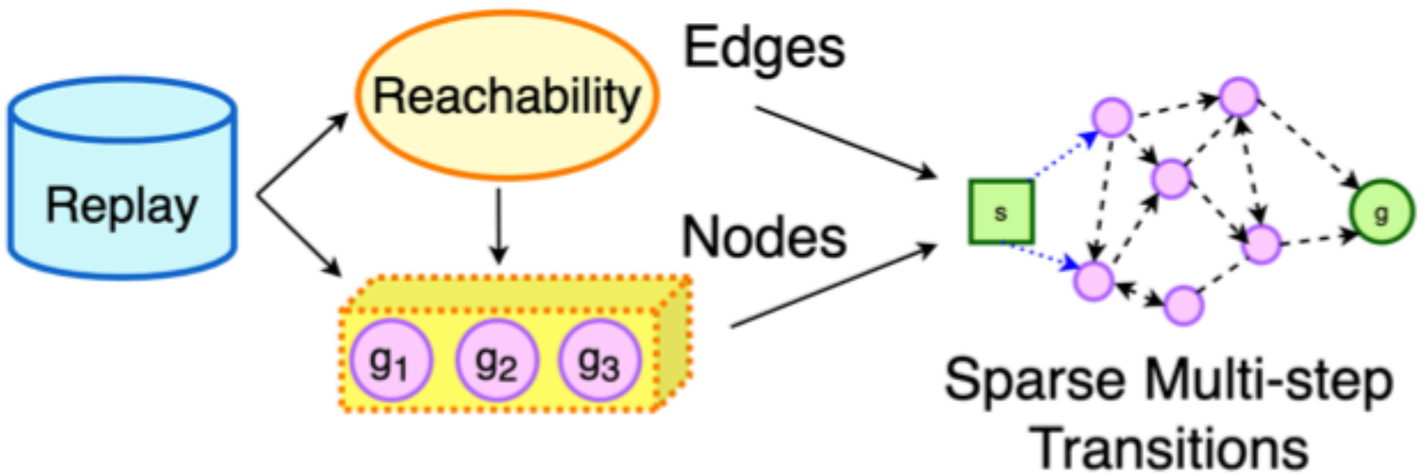
Dynamics a.k.a. the World Model

world model /
dynamics model

planning over actions



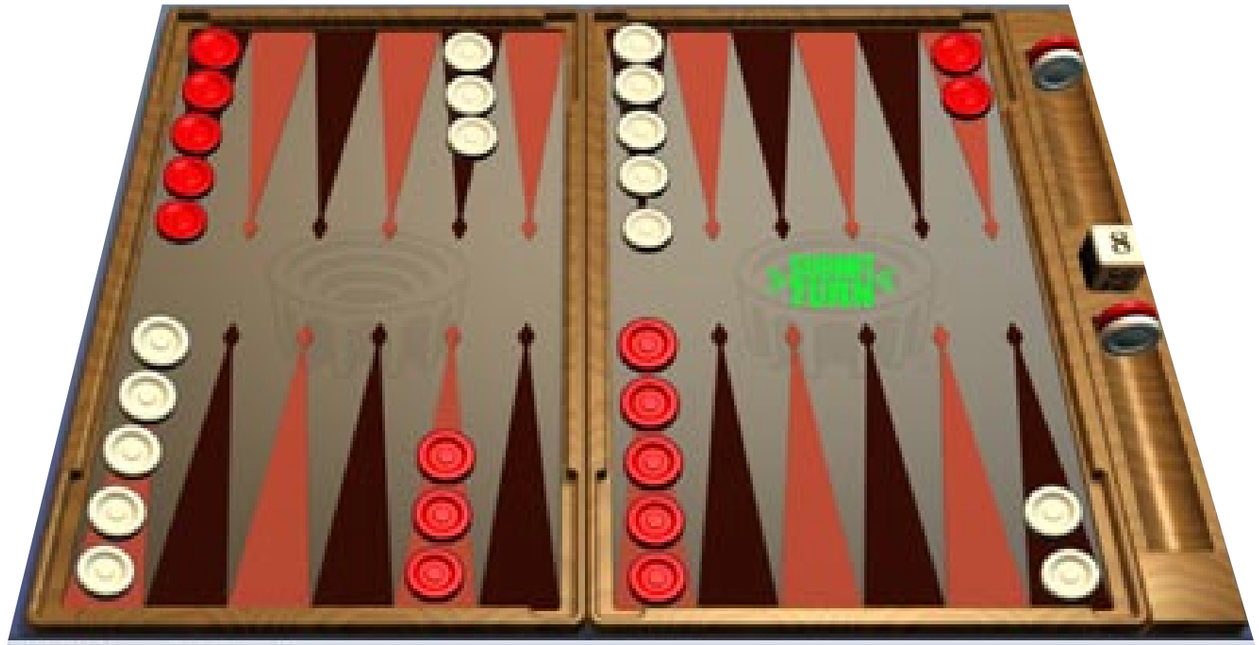
planning over sub-goals



Learning world models is a central open research topic.

Backgammon

- States: Configurations of the playing board (≈1020)
- Actions: Moves
- Rewards:
  - win: +1
  - lose: −1
  - else: 0

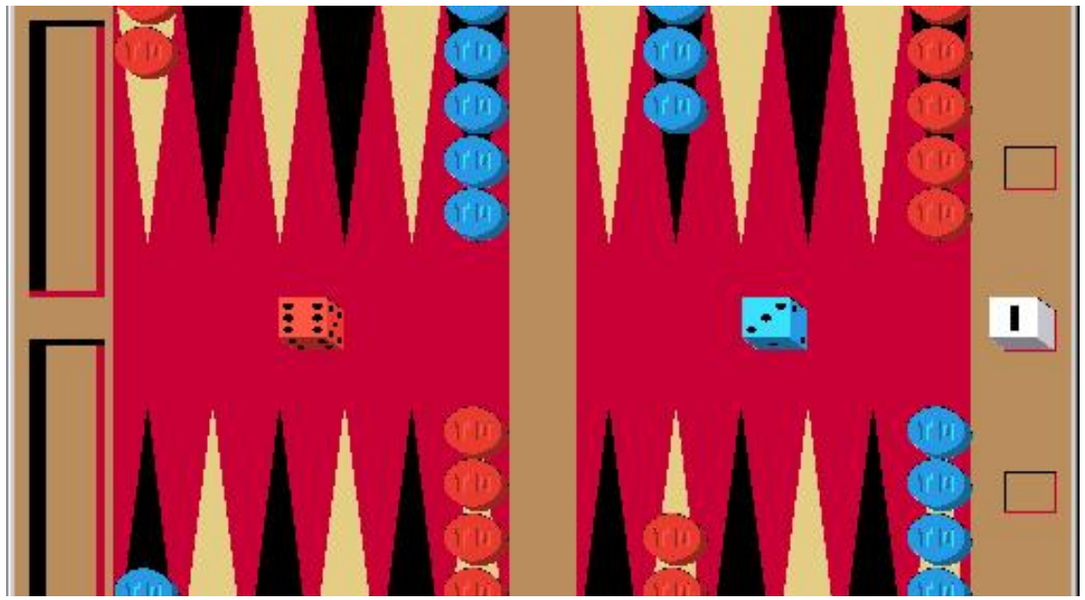


Backgammon

Neuro-Gammon: Learning to Play Backgammon by imitating human experts

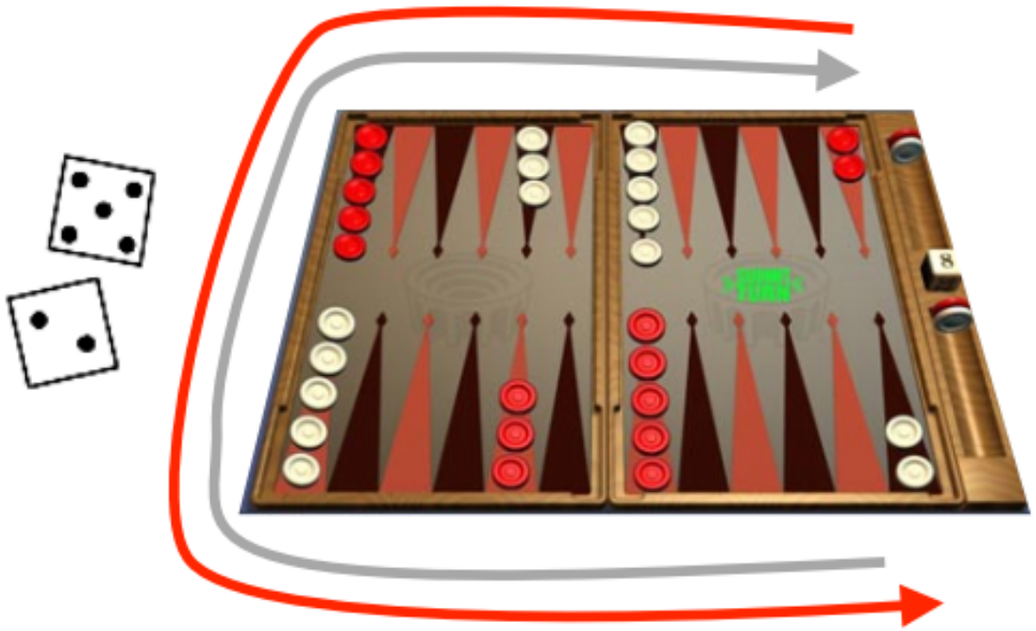
- Developed by Gerald Tesauro in 1989 in IBM’s research center
- Trained to mimic expert demonstrations using supervised learning
- Achieved intermediate-level human player

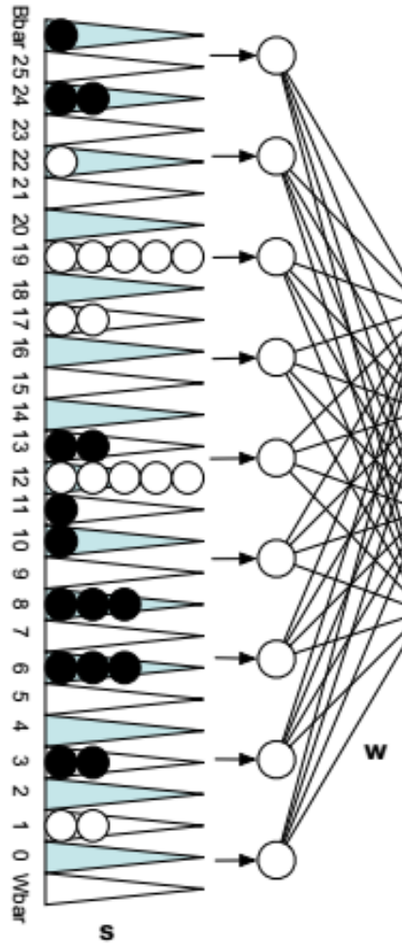
TD-Gammon: Learning to Play Backgammon by Reinforcement Learning



- 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

TD-Gammon's evaluation function

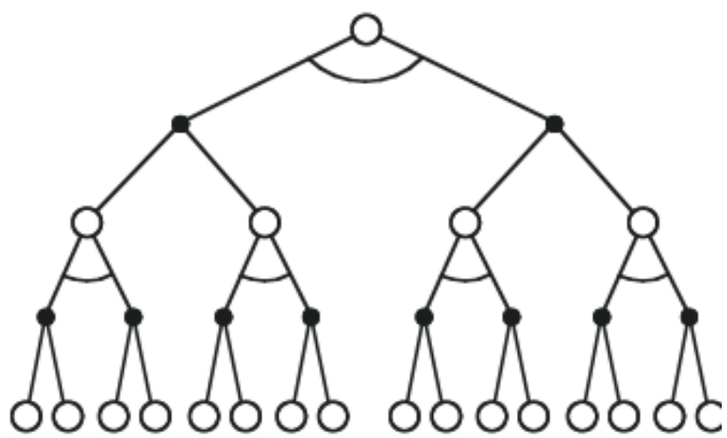




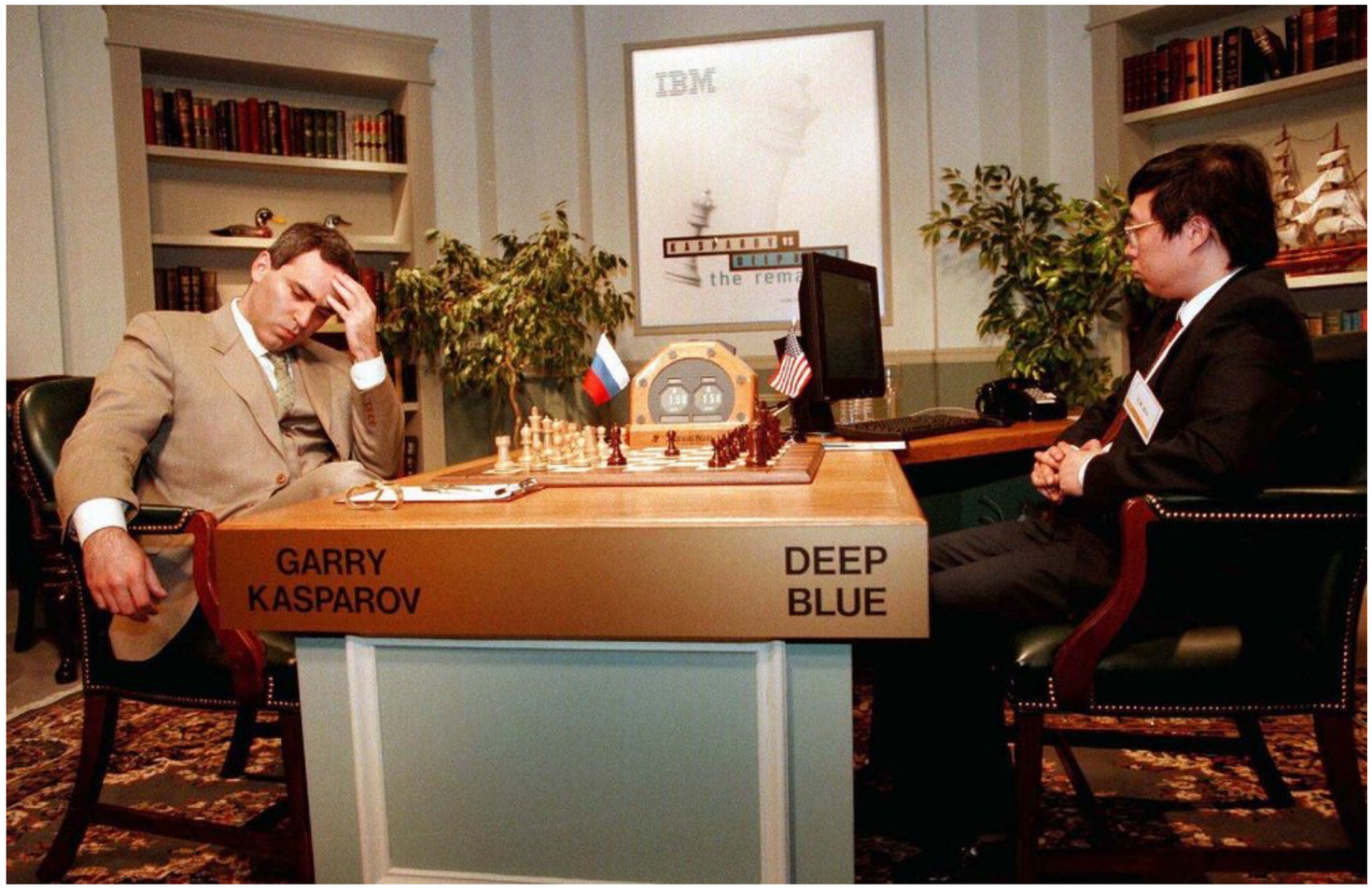
A neural net with only 80 hidden units..

estimated state value
(≈ prob of winning)

Action selection
by a shallow search



Deep blue's evaluation function has largely hand-designed



Limitations of Reinforcement Learning

- The agent should have the chance to try (and fail) enough times
- This is impossible if episode takes too long, e.g., reward="obtain a great Ph.D."
- This is impossible when safety is a concern: we can't learn to drive via reinforcement learning in the real world, failure cannot be tolerated

Behavior: High Jump

scissors

Fosbury flop

- Learning from rewards
  - Reward: jump *as high as possible: It took years to discover the present Fosbury flop*
- *Learning from demonstrations*
  - It was way easier for athletes to perfection the flop, once Fosbury showed the general trajectory
- Learning from specifications of optimal behavior
  - For novices, it is easier to replicate a behavior if additional guidance is provided in natural language: where to place the foot, how to time yourself, etc. .

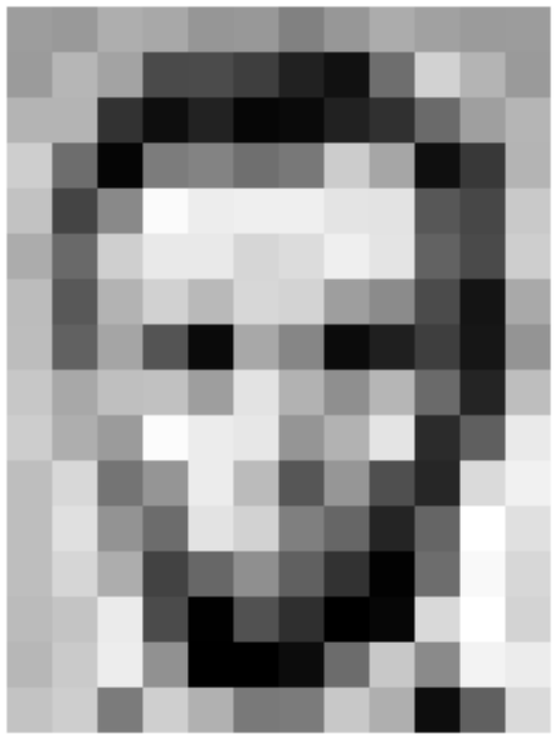
Overview for today

- Goal of the course / why it is important
- What is reinforcement learning
- What is representation learning (and how it helps reinforcement learning and behavior learning in general)
- Reinforcement learning versus supervised learning
- AI's paradox: what is hard and what is easy in behavior learning

Representation learning

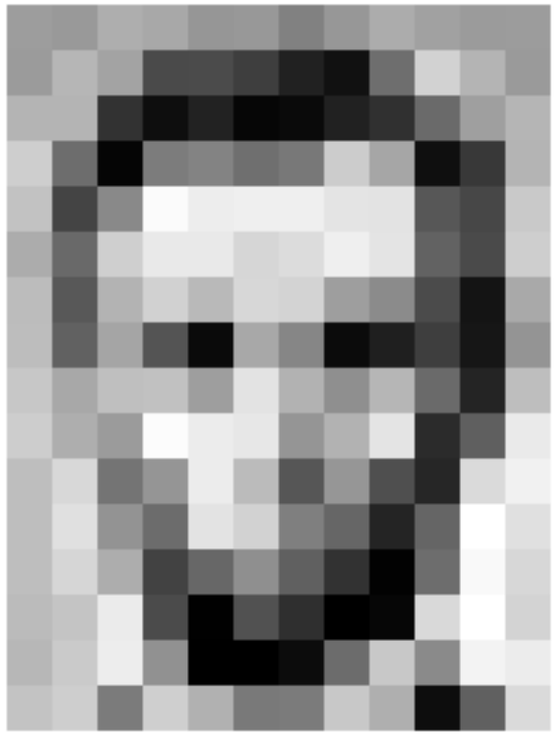
- Representation learning: mapping raw observations to features and structures from which the mapping to actions or to semantic labels is easier to infer.

Representation learning



- Remember what the computer sees

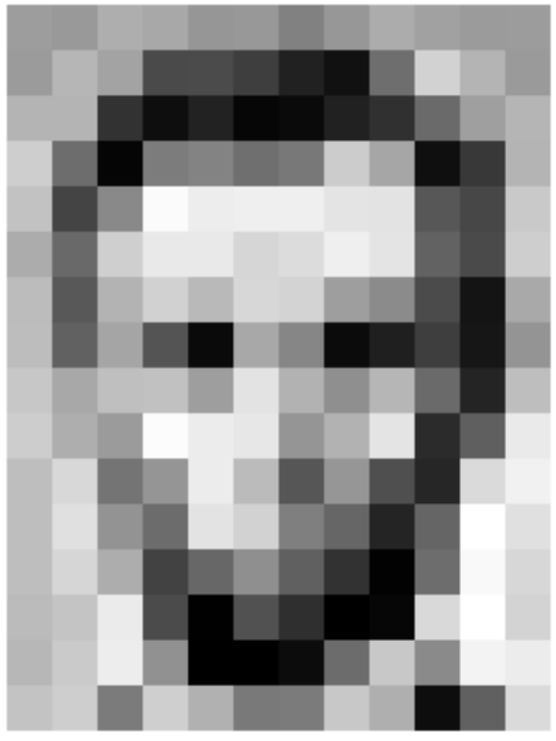
Representation learning



157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	95	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

- Remember what the computer sees

Representation learning

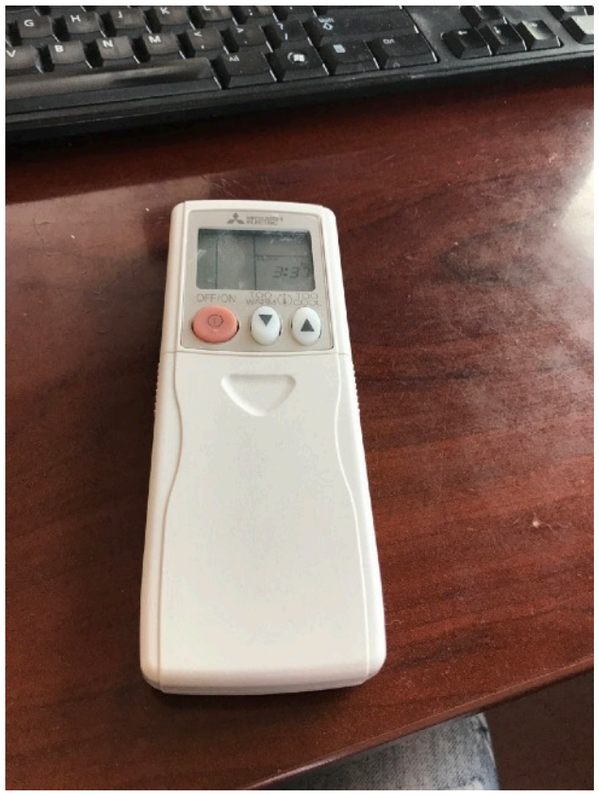
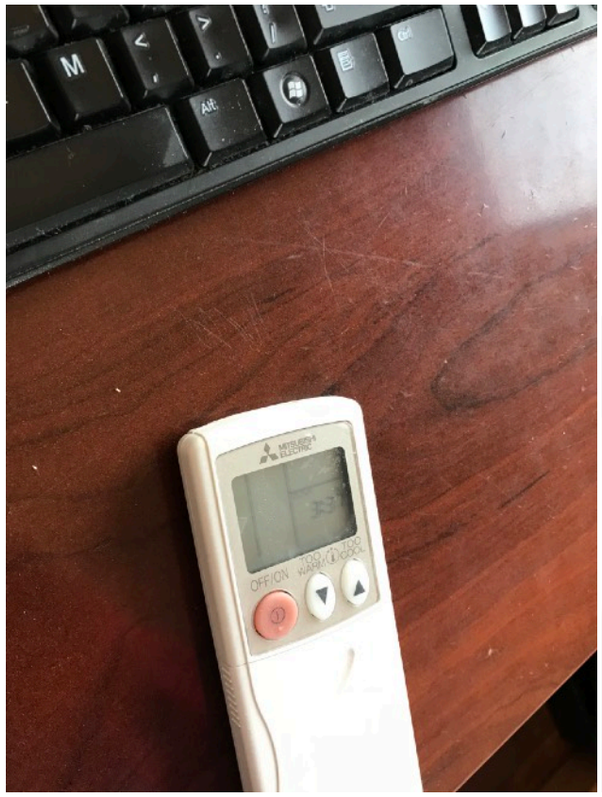


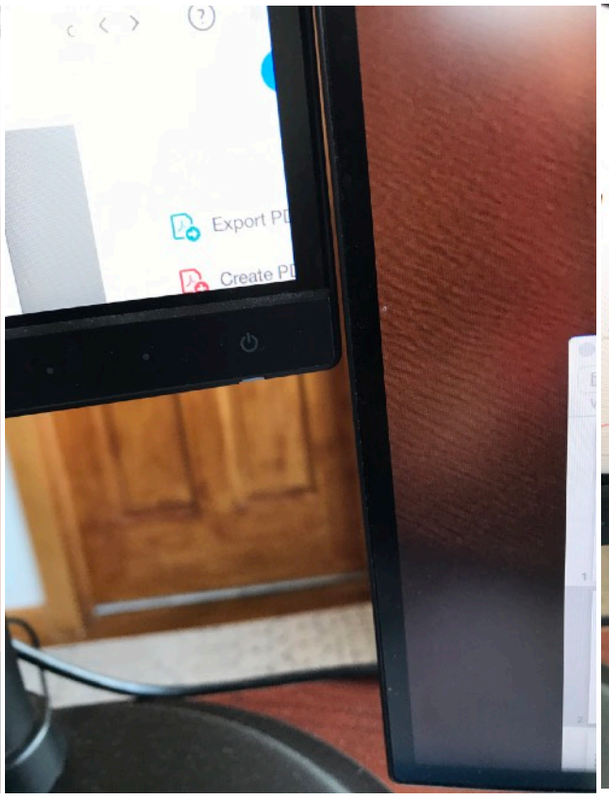
157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

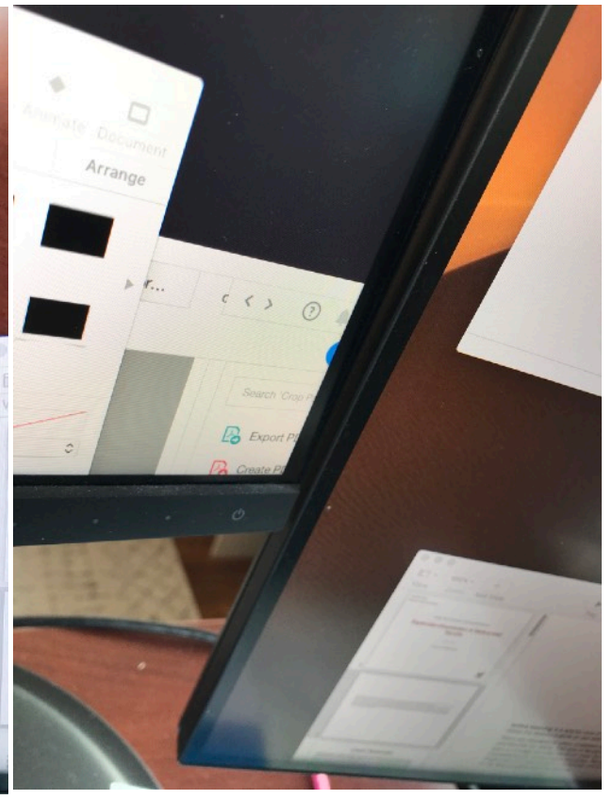
157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

- Remember what the computer sees

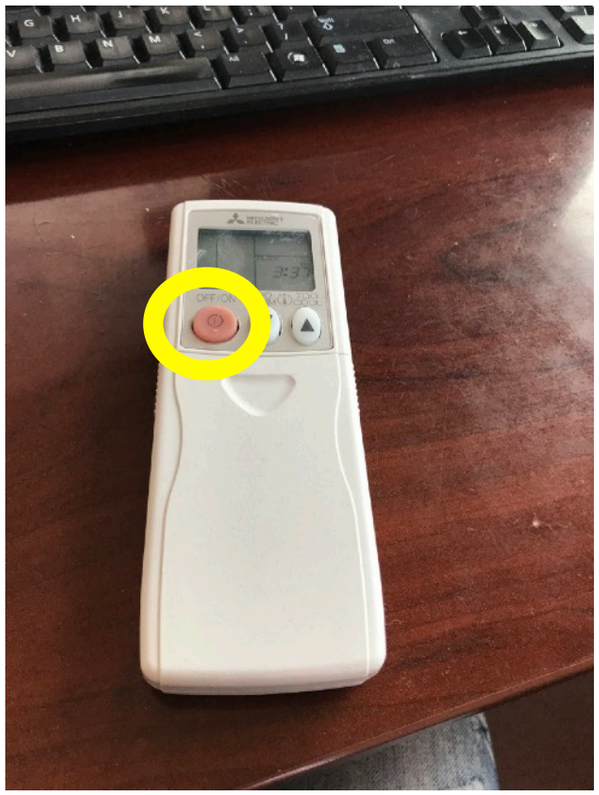
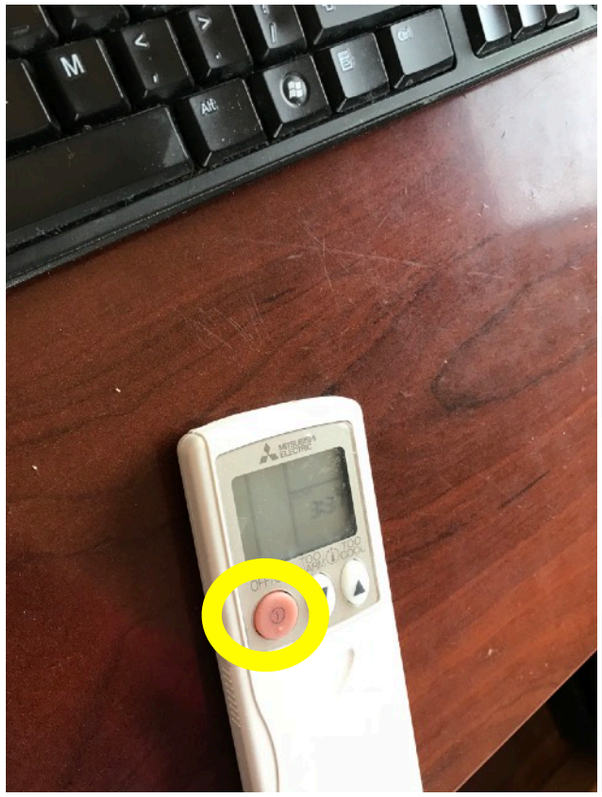
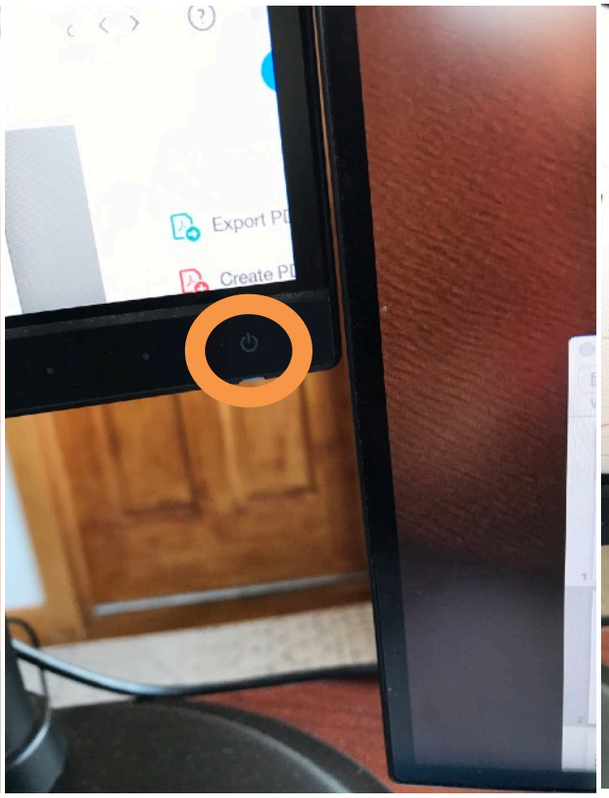
Representation learning

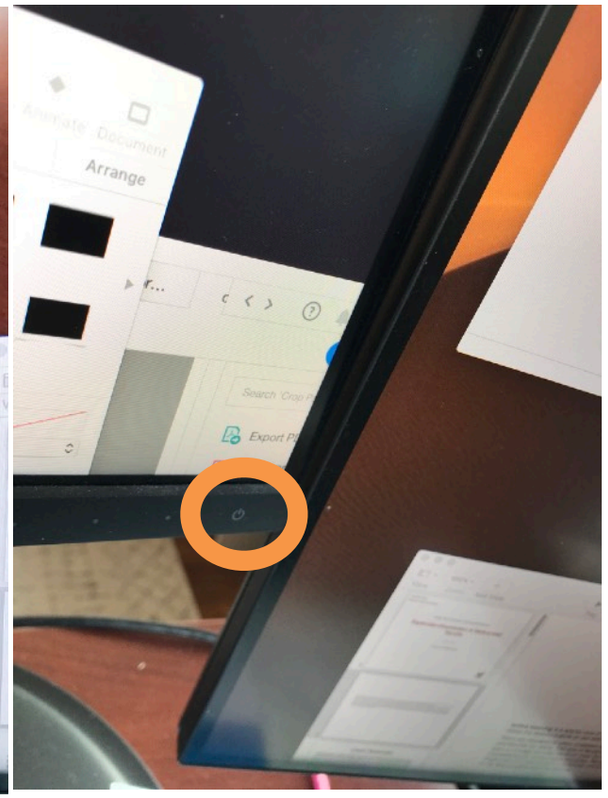







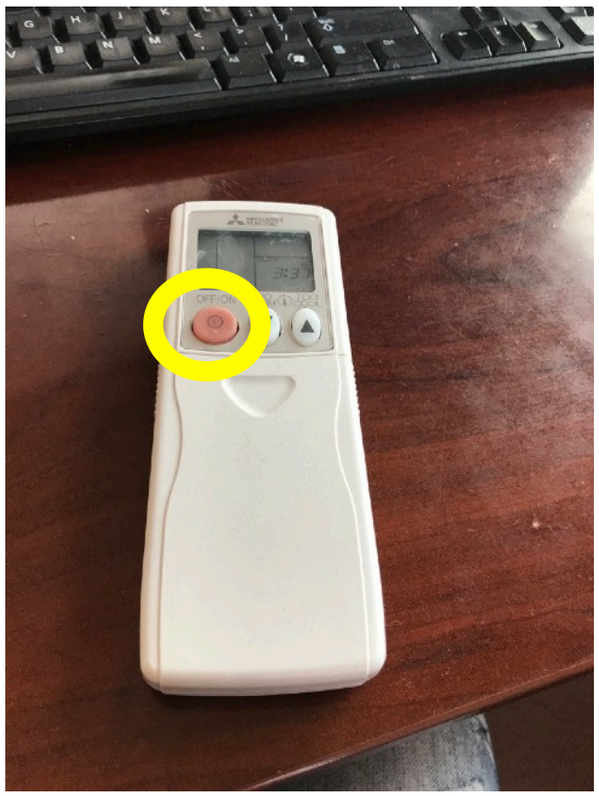
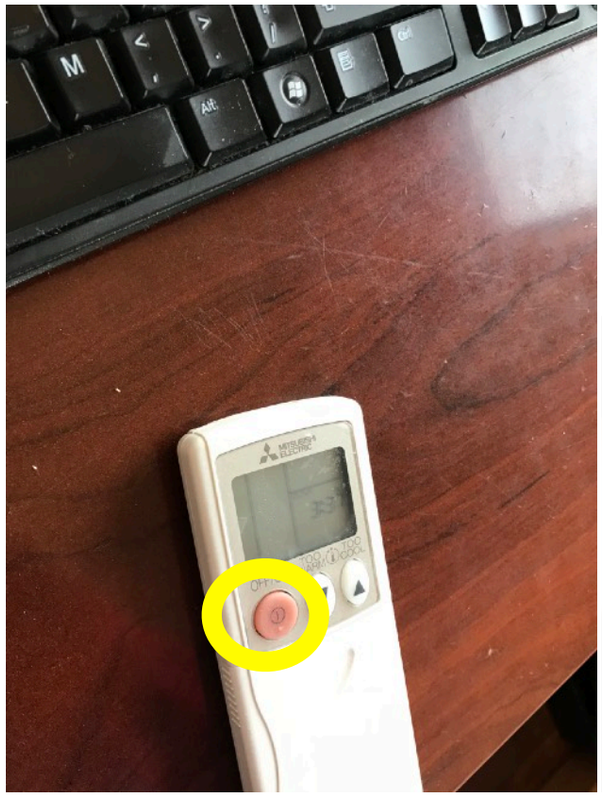
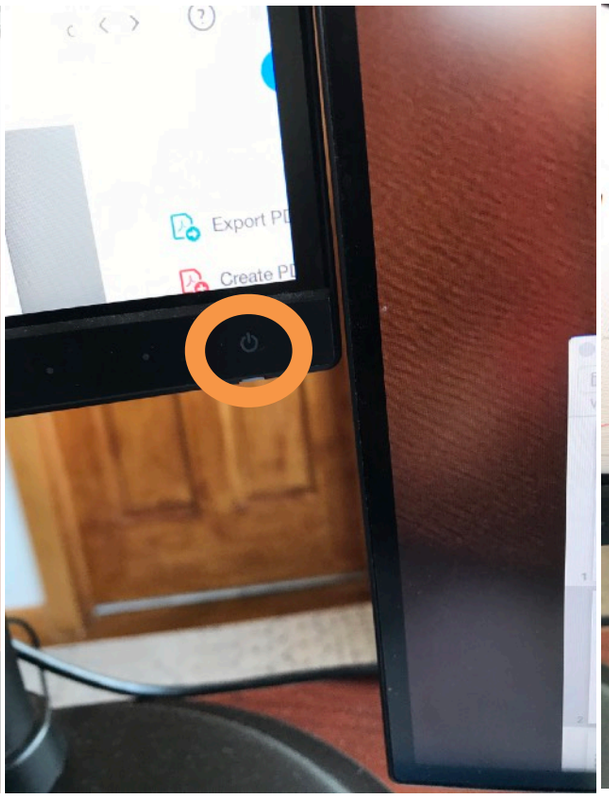
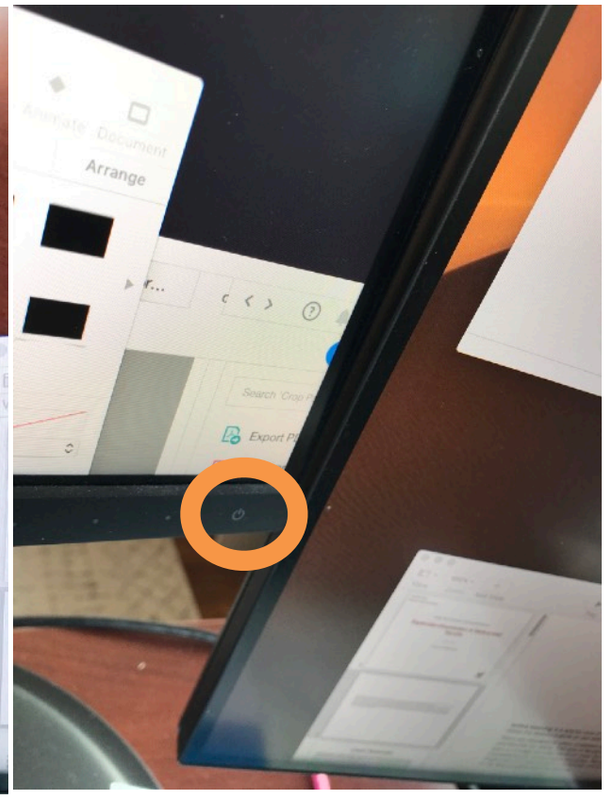
Representation learning



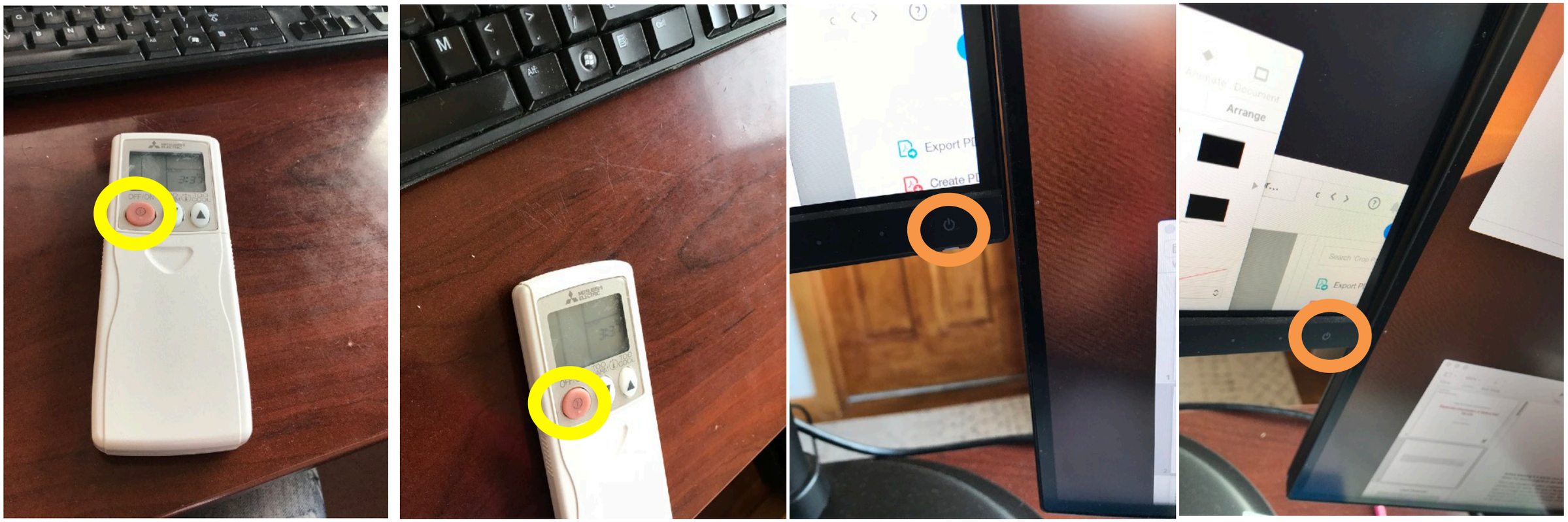
(Visual) Representation learning helps learning to act

- Despite these images have very different pixel values, actions required to achieve the goal of switching on the device are similar.
- Visual perception is instrumental to learning to act, in transforming raw pixels to action-relevant feature vectors and structures.

(Visual) Representation learning helps learning to act

- Training our visual representations with auxiliary tasks is likely to dramatically decrease the number of interactions with the environment we need to learn to press buttons.



- Q: What are reasonable auxiliary tasks?
  - Supervised: object detection, image classification, pixel labelling.
  - Self-supervised: instance discrimination, masked image prediction

In practice: A lot of domain knowledge for going from observations to states

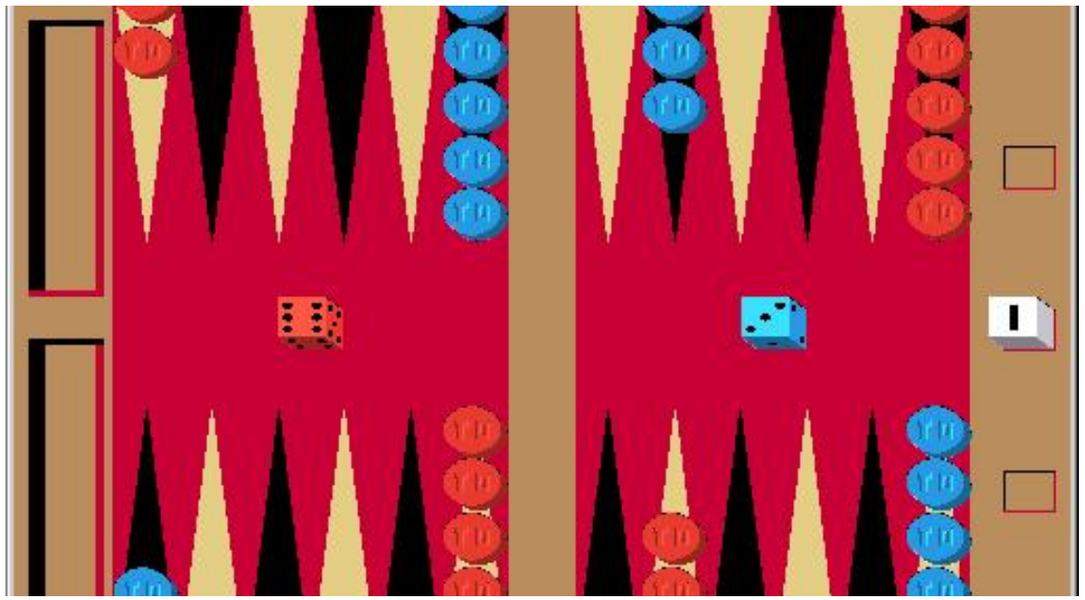
Learning to Act

- *Discovering* a behavior through trial-and-error guided by rewards.
- *Generalizing/transferring* a behavior across different scenarios (camera viewpoints, object identities, objects arrangements) E.g., you show me how to open one door, and I now need to learn how to open other similar doors
- *Generalization is critical for efficient exploration to help discovery*
- *Representation learning helps in cross-environment and cross-task behaviour transfer*

Overview for today

- Goal of the course / why it is important
- What is reinforcement learning
- What is representation learning (and how it helps reinforcement learning and behavior learning in general)
- Reinforcement learning versus supervised learning
- AI's paradox: what is hard and what is easy in behavior learning

RL:TD-Gammon SL:Neuro-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

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

Reinforcement learning Versus supervised learning

- RL is a form of **active learning**:
  - the agent gets the chance **to collect her own data** by acting in the world, querying humans, and so on.
  - the data changes over time, it depends on the policy of the agent.
  - To query the environment effectively, the agent needs to keep track of its uncertainty: what she knows and what she does not, and thus needs to explore next.
- Supervised learning is a form of passive learning:
  - the data does not depend on the agent in anyway, it is provided by external labellers.
  - the data is static throughout learning.

Reinforcement learning Versus supervised learning

- In RL, we ***often*** cannot use gradient-based optimization:
  - e.g., when the agent does not know neither the world model to unroll nor the reward function to maximize.
- In supervised learning, we usually can use gradient-based optimization:
  - E.g., we consider a parametric form for our regressor or classifier and optimize it via stochastic gradient descent (SGD).

Reinforcement learning Versus supervised learning

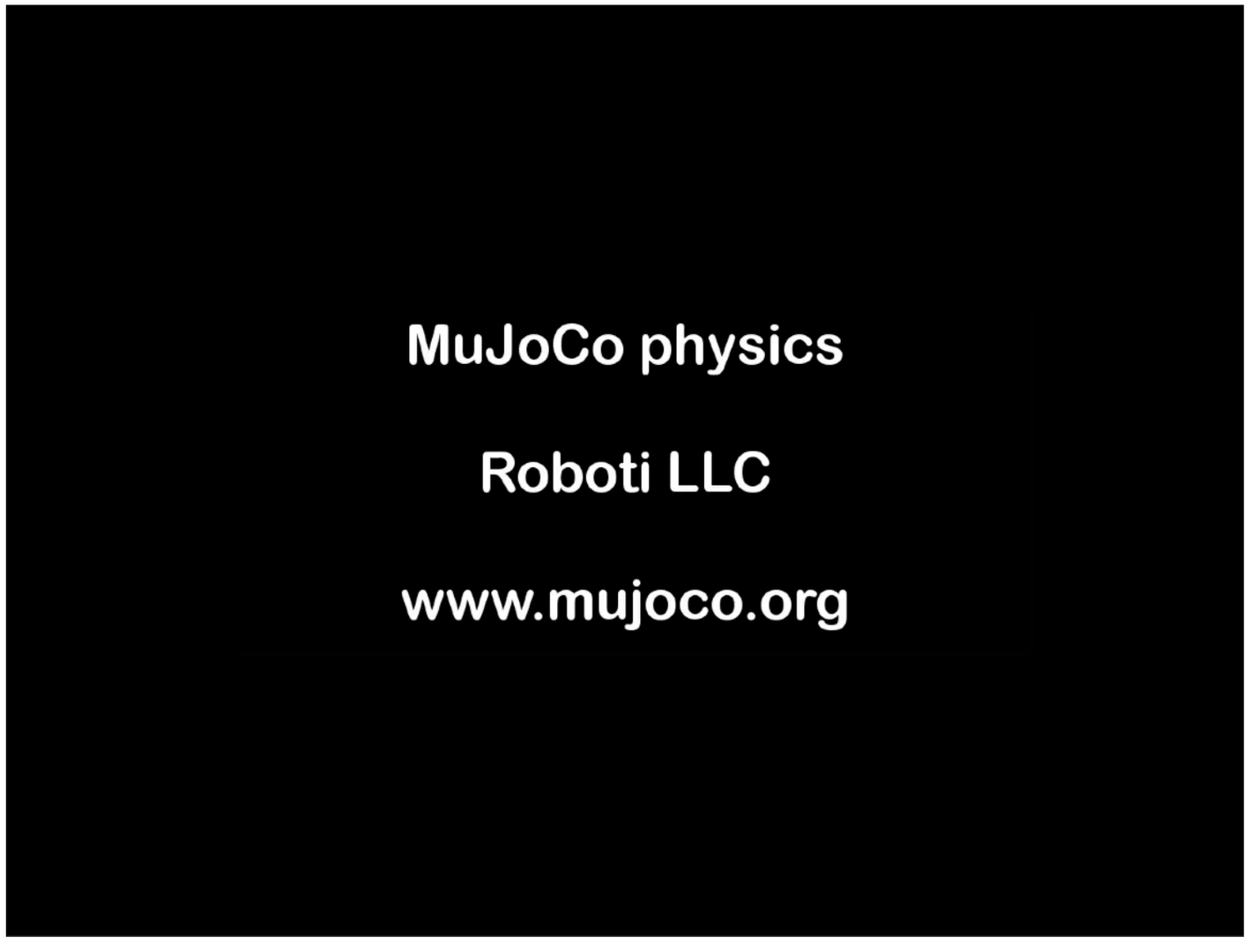
- RL can be time consuming. Actions take time to carry out in the real world, i.e., each interaction has a non-negligible cost. Our goal is the agent to **minimize the amount of interactions with the environment** while succeeding in the task.

Reinforcement learning Versus supervised learning

- RL can be time consuming. Actions take time to carry out in the real world, i.e., each interaction has a non-negligible cost. Our goal is the agent to **minimize the amount of interactions with the environment** while succeeding in the task.
- We can use simulated experience and tackle the SIM2REAL (simulation to reality) transfer.

Reinforcement learning Versus supervised learning

- RL can be time consuming. Actions take time to carry out in the real world, i.e., each interaction has a non-negligible cost. Our goal is the agent to **minimize the amount of interactions with the environment** while succeeding in the task.
- We can use simulated experience and tackle the SIM2REAL (simulation to reality) transfer.



Reinforcement learning Versus supervised learning

- RL can be time consuming. Actions take time to carry out in the real world, i.e., each query has a non-negligible cost. Our goal is the agent to **minimize the amount of interactions with the environment** while succeeding in the task.
- We can use simulated experience and tackle the SIM2REAL (simulation to reality) transfer.
- We can have robots working 24/7

Reinforcement learning Versus supervised learning

- RL can be time consuming. Actions take time to carry out in the real world, i.e., each query has a non-negligible cost. Our goal is the agent to **minimize the amount of interactions with the environment** while succeeding in the task.
- We can use simulated experience and tackle the SIM2REAL (simulation to reality) transfer.
- We can have robots working 24/7
- We can buy many robots

Google’s Robot Farm

Offline Reinforcement learning : learning from sequence data

- We do not have an agent interacting with the world.
- We have a set of experts contributing experience.
- We want to learn to act better than each of these experts.

Examples?

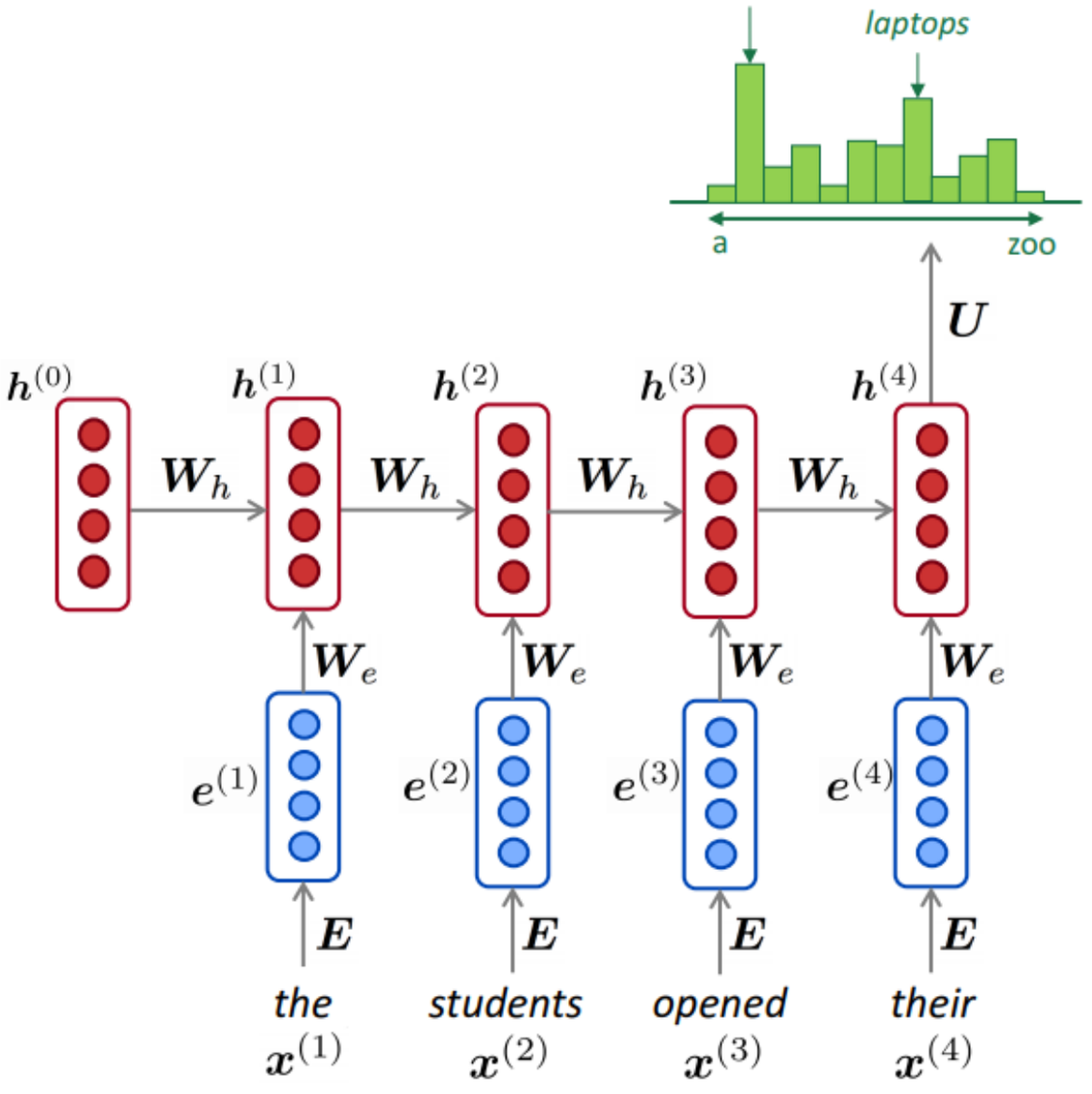
Driving logs!

Offline Reinforcement learning : learning from sequence data

Sequence learning has had tremendous progress in NLP.

Can we cast learning to act as a sequence learning problem?

E.g., subgoal generation and goal conditioned evaluation of the generated sub-goals.



Offline Reinforcement learning : learning from sequence data

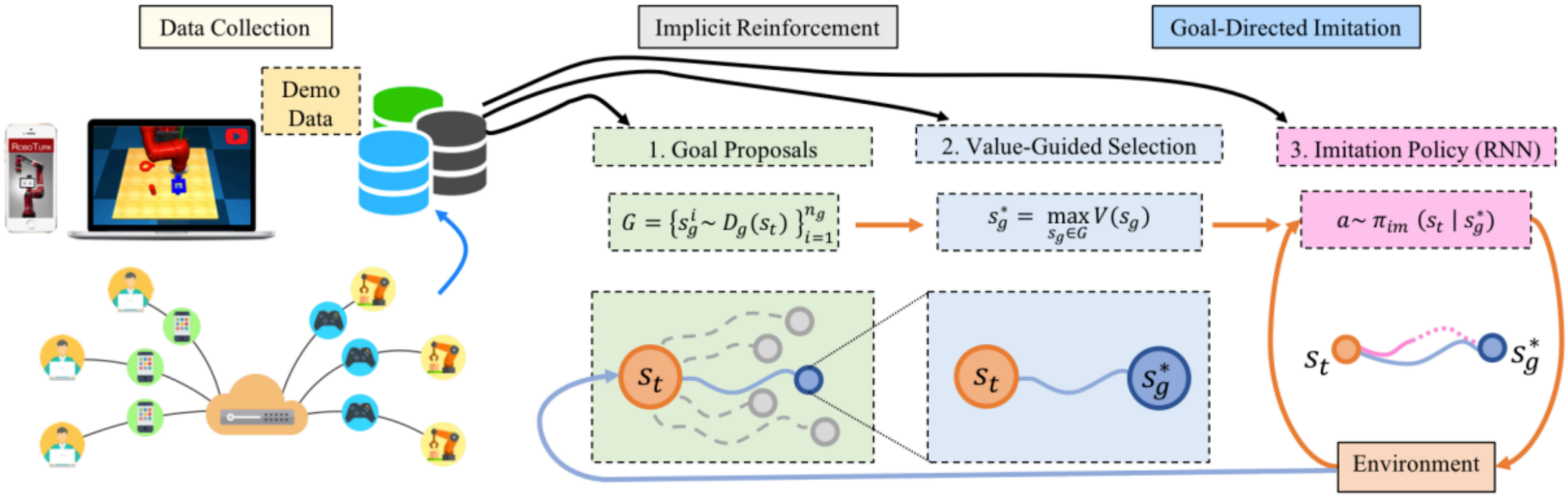


Fig. 1: **Overview** IRIS learns policies from large quantities of demonstration data without environment interaction during learning. It trains a goal-conditioned low-level controller to reproduce short demonstration sequences and a high-level goal selection mechanism consisting of a goal proposal network and a value network. At test-time, a set of goals is proposed by a generative model and selected by the value function, and this is set as the target for low-level imitation. Both high and low levels are run in closed-loop with appropriate rates.

Reinforcement learning in 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. **One goal Vs many goals**

4. **Rewards are provided automatically by an oracle environment VS rewards need themselves to be detected**

5. Interactions take time: we really need **intelligent** exploration

Overview for today

- Goal of the course / why it is important
- What is reinforcement learning
- What is representation learning (and how it helps reinforcement learning and behavior learning in general)
- Reinforcement learning versus supervised learning
- AI’s paradox: what is hard and what is easy in behavior learning

GO

AlphaGoZero the program that beat the world champions with only RL

- Monte Carlo Tree Search with neural nets
- self play

Go Versus the real world

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

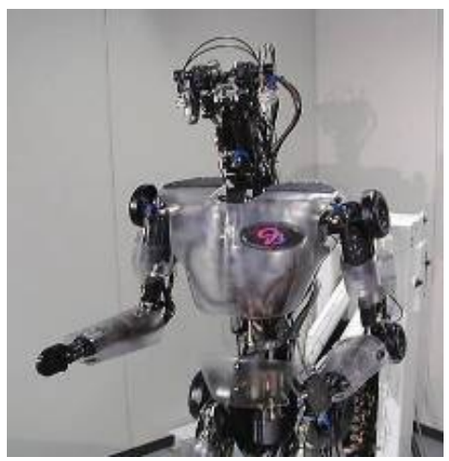
The difficulty of motor control

What to move where

vs.

Moving

vs.



From Dan Wolpert

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.”

AI’s paradox

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

Rodney Brooks

AI’s paradox

Intelligence was "best characterized as the things that highly educated 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.”

Rodney Brooks

AI’s paradox

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

"The thi

effortles

No cognition. Just sensing and action

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.”

Rodney Brooks

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*

Take-aways

- *Forms of supervision for learning to act: mapping observations to actions for a specific goal*
- *The reinforcement learning problem, terminology, basic ingredients*
- *RL vs SL*
- *Learning to search using evaluation functions*
- *AI paradox: is hard to learn the abilities of a 2 year old, and easy to learn to beat GO champions, solve theorems and so on: a big search at a kind of small (compared to the real world) state space at the end of the day.*

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