

monocular RGB camera

7 DoF robotic manipulator

2-finger gripper

object bin



Option 1:

Understand the problem, design a solution



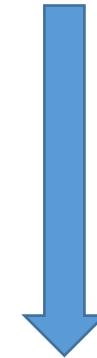
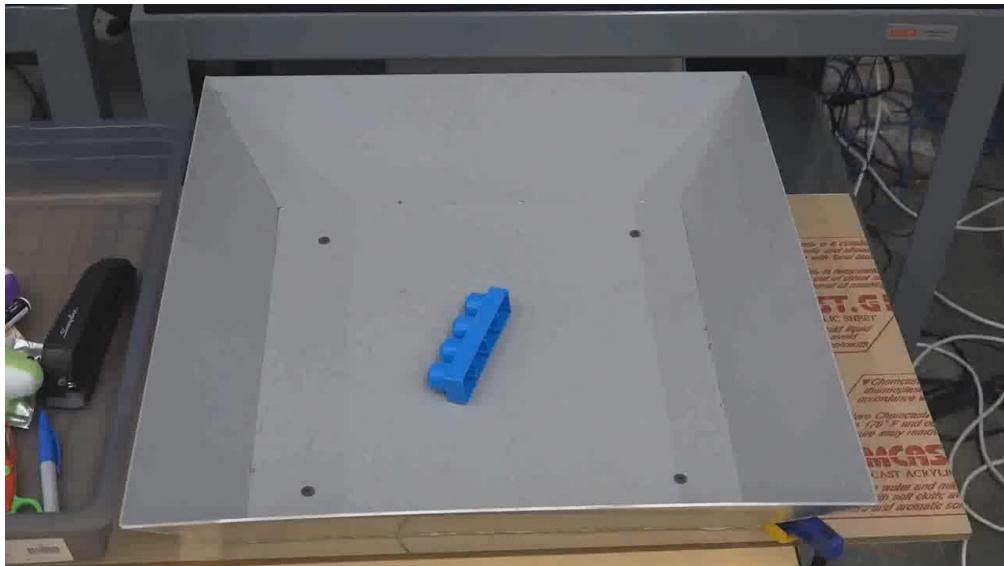
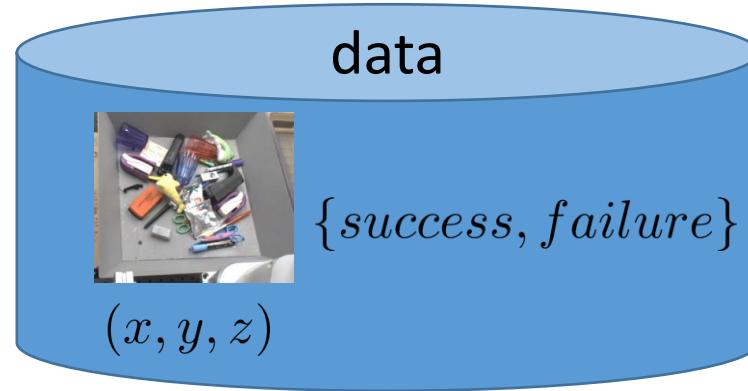
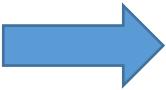
Option 2:

Set it up as a machine learning problem



supervised learning





What are some recent advances we've seen in AI?



vibrant portrait painting of Salvador Dalí with a robotic half face



a shiba inu wearing a beret and black turtleneck



a close up of a handpalm with leaves growing from it



an espresso machine that makes coffee from human souls, artstation



panda mad scientist mixing sparkling chemicals, artstation



a corgi's head depicted as an explosion of a nebula

A screenshot of a messaging interface. The conversation starts with the user asking "Which animal do you like?". The AI responds with "I like cows.". The user then asks "I heard they go to college.". The AI replies with "Cows go to college?". The user asks "I heard that a cow went to Harvard.". The AI responds with "What did the cow study?". The user asks "Bovine sciences.". The AI replies with "Do horses ever go to Harvard?". The user ends the message with an ellipsis (...).

M Which animal do you like?
I like cows.
M I heard they go to college.
Cows go to college?
M I heard that a cow went to Harvard.
What did the cow study?
M Bovine sciences.
Do horses ever go to Harvard?
M ...

Explaining a joke

Prompt

Explain this joke:

Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

Model Response

Prediction: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

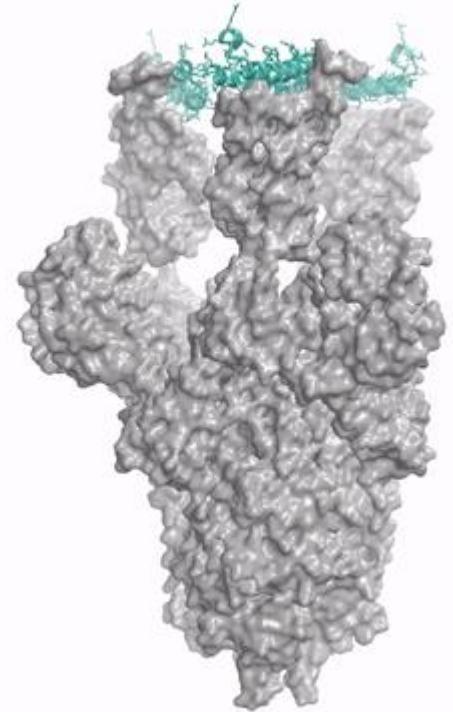
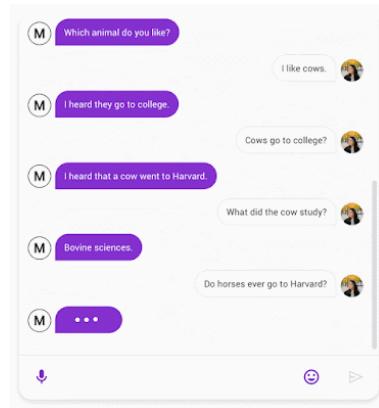


image credit: UW IPD

What's the main idea behind this?



Explaining a joke

Prompt

Explain this joke:

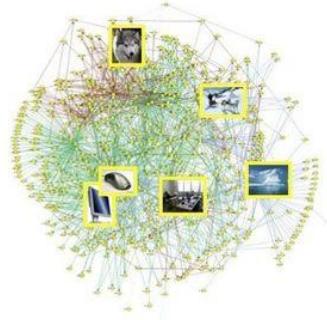
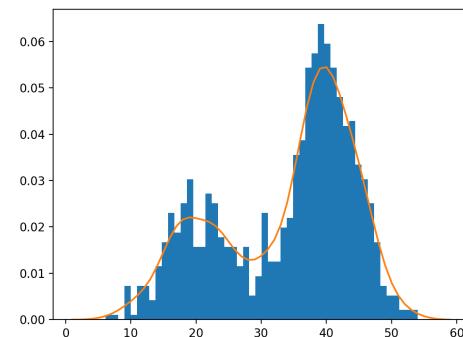
Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

Model Response

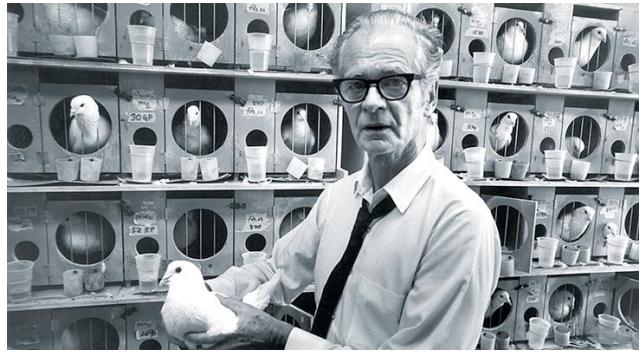
Prediction: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

$$p_{\theta}(\mathbf{x})$$

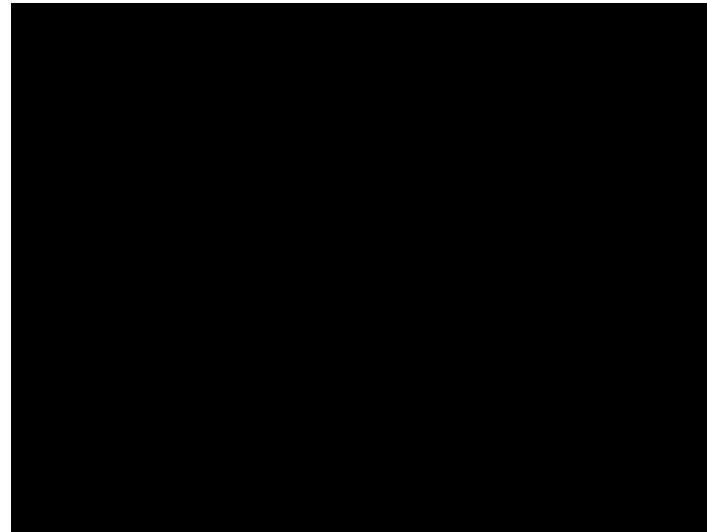
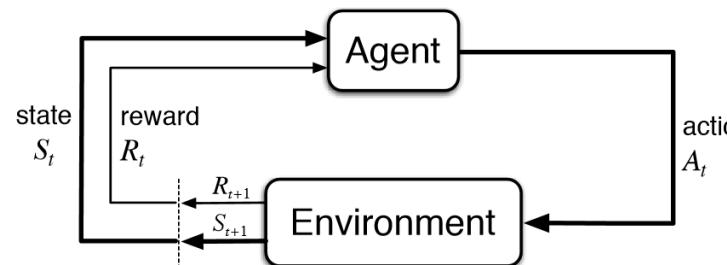
$$p_{\theta}(\mathbf{y}|\mathbf{x})$$



What does reinforcement learning do differently?



classical reinforcement learning



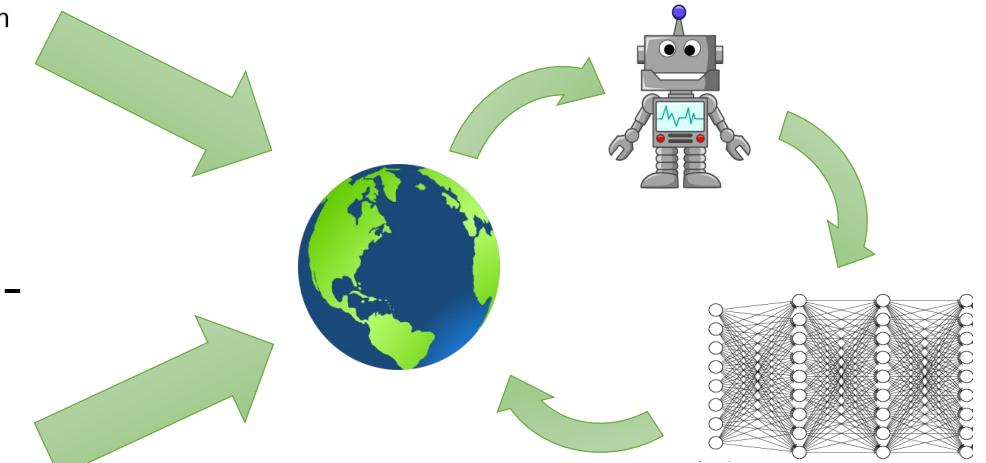
Evolved Virtual Creatures.
Karl Sims, 1994

Synthesis of Complex Behaviors
with
Online Trajectory Optimization

Yuval Tassa, Tom Erez & Emo Todorov

IEEE International Conference
on Intelligent Robots and Systems
2012

Model-Predictive Control with iLQG.
Yuval Tassa, 2012

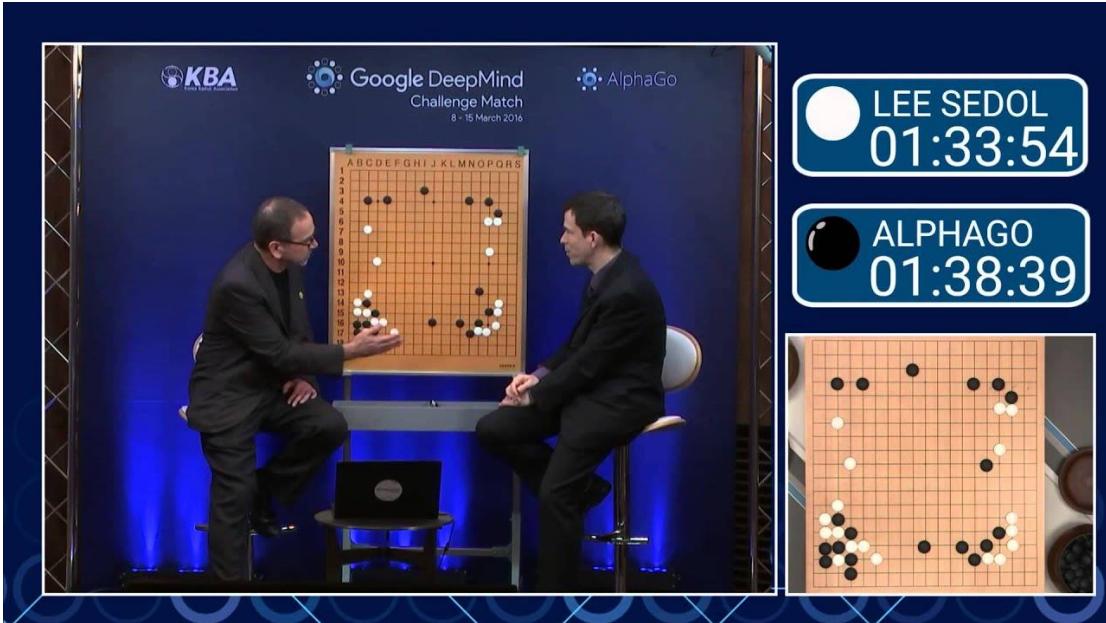


deep reinforcement learning



Reinforcement learning can discover new solutions

Impressive because no person had thought of it!



“Move 37” in Lee Sedol AlphaGo match: reinforcement learning “discovers” a move that surprises everyone

Impressive because it looks like something a person might draw!



Deep Reinforcement Learning, Decision Making, and Control

CS 285

Instructor: Sergey Levine
UC Berkeley



Course logistics

Class Information & Resources



Sergey Levine
Instructor



Kyle Stachowicz
Head GSI



Joey Hong
GSI



Vivek Myers
GSI



Kevin Black
GSI

- Course website: <http://rail.eecs.berkeley.edu/deeprlcourse>
- Ed: CS285 Deep Reinforcement Learning: <https://edstem.org/us/join/c9XPxK>
- Gradescope: CS285 Deep Reinforcement Learning: <https://www.gradescope.com/courses/571673>
- Office hours: check course website, mine are after class right here (starting next week)
- Links to lecture videos, etc.: always posted on Ed (pinned post)

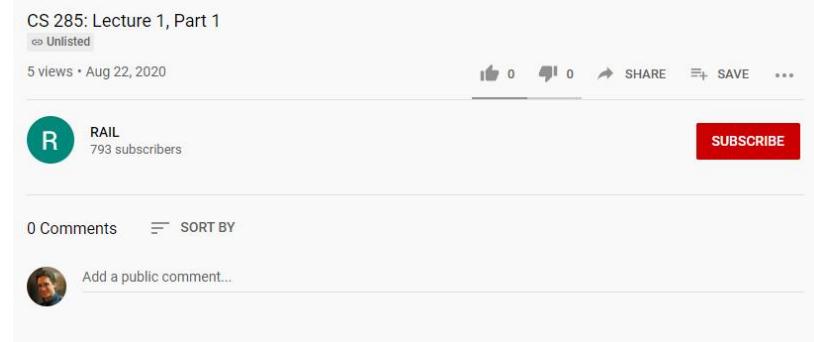
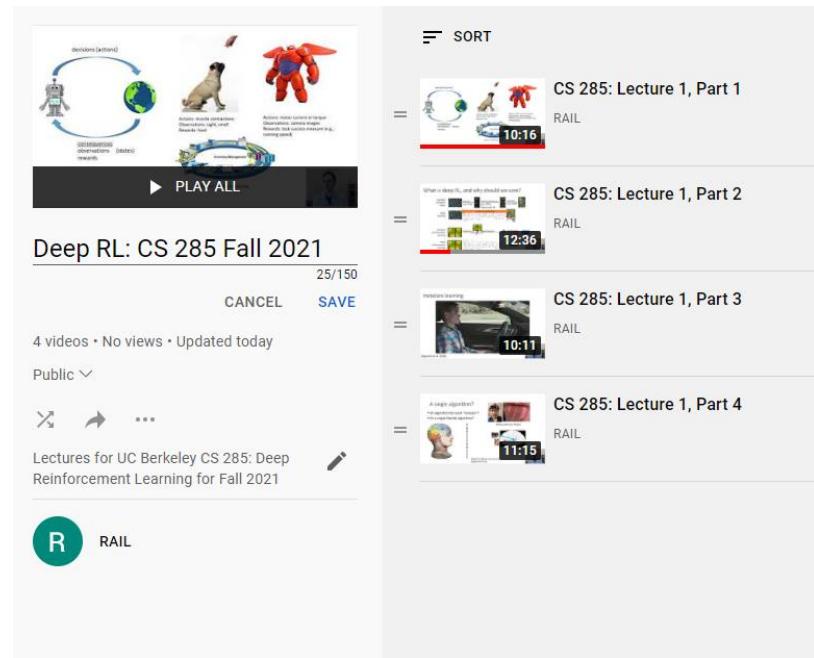
Prerequisites & Enrollment

- All enrolled students must have taken CS182, CS189, CS289, CS281A, CS282, or an equivalent course at your home institution
 - Please contact Sergey Levine if you haven't
- If you are not eligible to enroll directly into the class, **fill out the enrollment application form (do this today):**
<http://rail.eecs.berkeley.edu/deeprlcourse/>
 - We will enroll subject to availability based on responses to this form
 - We will not use the official CalCentral wait list!
 - **Fill out an application before the end of this week (8/25/23)!**

Class format

- **Lectures**

- Lectures will be posted over the weekend on **YouTube**
- The YouTube playlist link will be posted on **Ed** and on the course website:
 - https://youtube.com/playlist?list=PL_iWQOsE6TfVYGEGiAOMaOzzv41Jfm_Ps
- Some lectures from past years, some new
- **Watch the lectures in advance**
- **Post questions on Ed thread**
- **Upvote questions on Ed thread**
- **Every lecture has a quiz**



Class format

- **Every lecture has a quiz**
 - Complete on Gradescope prior to class
 - Due at 5 pm on the day of the corresponding lecture
 - Except today's quiz, which is due 5 pm next wk Mon
 - Should never take more than 5-10 min if you watched the lecture and understood it
 - Check your understanding, lightweight & quick
 - Also serves to summarize the key concepts you should know
- **If you don't like your grade you can actually take it again (second time)!**
 - We release your grades automatically on the due date at 5 pm
 - We release a “second try” quiz that you can take if you don't like your grade (within 48 hours)
 - Answers will be released with “second try” quiz

Lecture 1 Quiz

STUDENT NAME
Search students by name or email...

Q1
1 Point
Which of the following are true for RL, but not supervised learning? Check all that apply

In RL, sequential inputs are not independent and identically distributed

In RL, we cannot use deep neural networks

In RL, the task is not simply to match the ground truth label

In RL, we don't have any access to a loss, reward, or objective signal of any sort

Save Answer

Q2
1 Point
Which of the following could be framed as reinforcement learning problems? Check all that apply

buying and selling inventory for a warehousing application

controlling a robot to run at a desired speed

controlling a vehicle to regulate the flow of traffic

Class format

- **In class**
 - **Attend class live:** meant to be interactive
 - **Discuss questions:** submit & upvote Ed questions to have a fun class!
 - **More in-depth derivations:** we'll go through some whiteboard derivations
 - **Participate and ask questions:** plenty of time for Q&A, bring your questions!
 - **Length:** Class will be about 50 min, followed by about 30 min of OH, since you are already watching 50-70 minutes of online lecture
- **Office hours:** My office hours will take place immediately after class
 - As soon as we cover all discussion, we will switch to “office hours” – use this time to ask questions about final projects, etc.
 - Exception is today, since we have a “classic” lecture, and presumably you don’t have many OH questions on first day of class :)

What you should know

- Assignments will require training neural networks with standard automatic differentiation packages (PyTorch)
- Review Section
 - We will have a review section to cover PyTorch next week on Wednesday
 - Optional, no quiz, but very useful if you are unfamiliar with it!
 - If you are unsure about background, try HW1 as soon as it comes out (next Monday) and if you're having trouble, come to the review section

Course content

What we'll cover

1. From supervised learning to decision making
2. Model-free algorithms: Q-learning, policy gradients, actor-critic
3. Model-based algorithms: planning, sequence models, etc.
4. Exploration
5. Offline reinforcement learning
6. Inverse reinforcement learning
7. Advanced topics, research talks, and invited lectures

Assignments

1. Homework 1: Imitation learning (control via supervised learning)
2. Homework 2: Policy gradients
3. Homework 3: Q-learning and actor-critic algorithms
4. Homework 4: Model-based reinforcement learning
5. Homework 5: Offline reinforcement learning
6. Final project: Research-level project of your choice (form a group of up to 2-3 students, you're welcome to start early!)

Grading: 50% homework (10% each), 40% project, 10% quizzes

5 late days total for homeworks

(does **not** apply to quizzes, proposal, milestone report, or final project report)

Your “Homework” Today

1. Make sure you are signed up for Ed (UC Berkeley CS285)
2. Start forming your final project groups, unless you want to work alone, which is fine
3. Take the lecture 1 quiz
 - it should be super quick if you watched lecture 1, mostly to familiarize yourself with Gradescope interface

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?

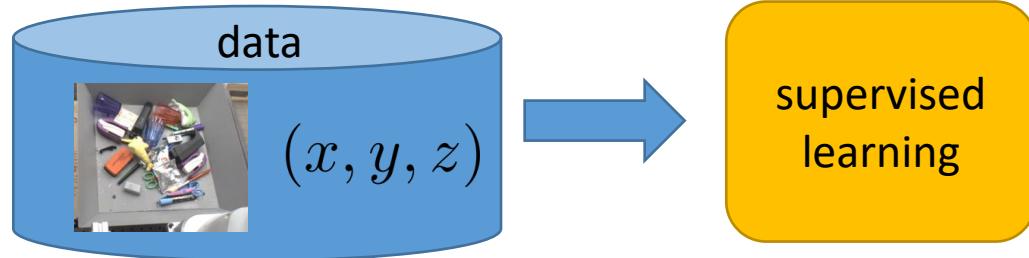
Standard (supervised) machine learning:

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

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

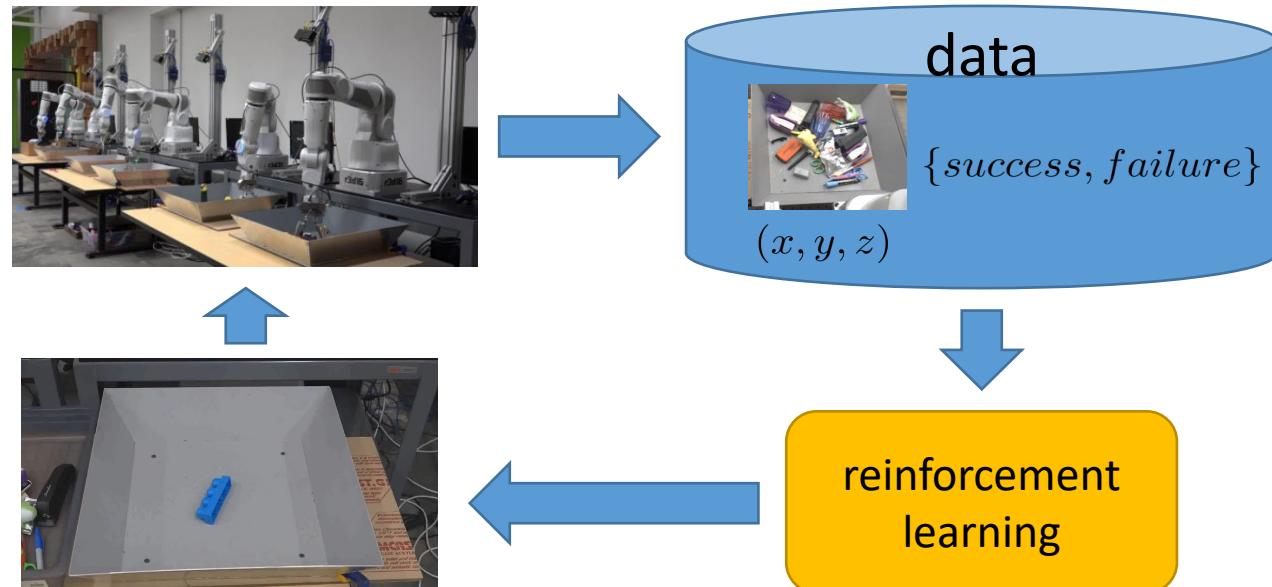
Usually assumes:

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



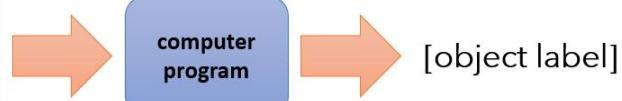
Reinforcement learning:

- 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



What is reinforcement learning?

supervised learning



input: \mathbf{x}

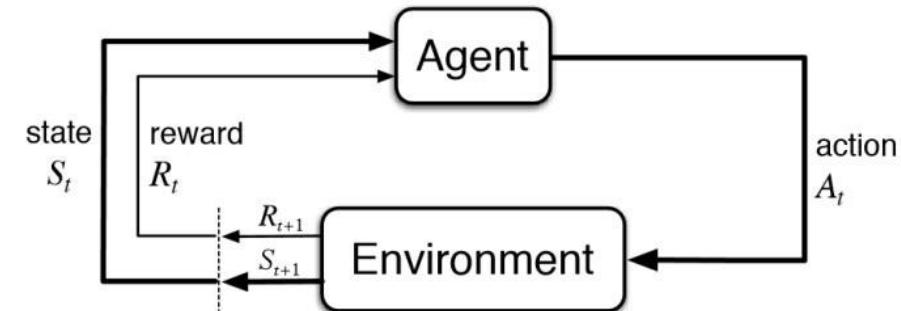
output: \mathbf{y}

data: $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}$

goal: $f_\theta(\mathbf{x}_i) \approx \mathbf{y}_i$

someone gives
this to you

reinforcement learning



pick your
own actions

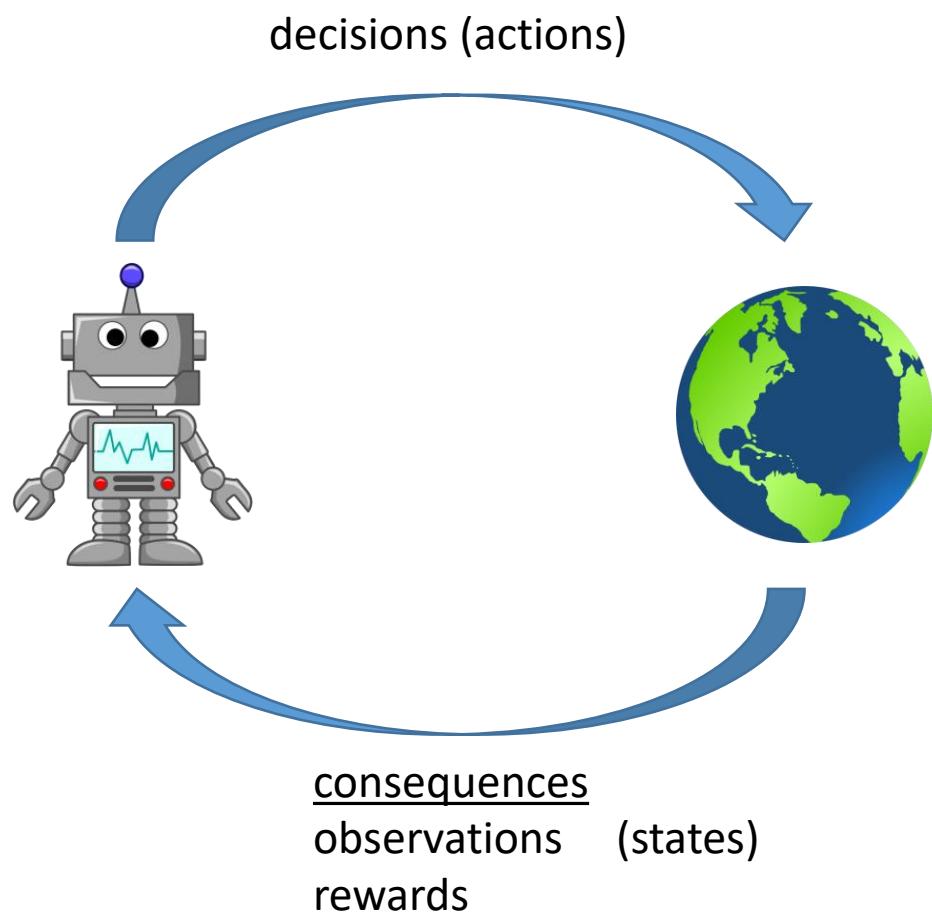
input: s_t at each time step

output: a_t at each time step

data: $(s_1, a_1, r_1, \dots, s_T, a_T, r_T)$

goal: learn $\pi_\theta : s_t \rightarrow a_t$

to maximize $\sum_t r_t$



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

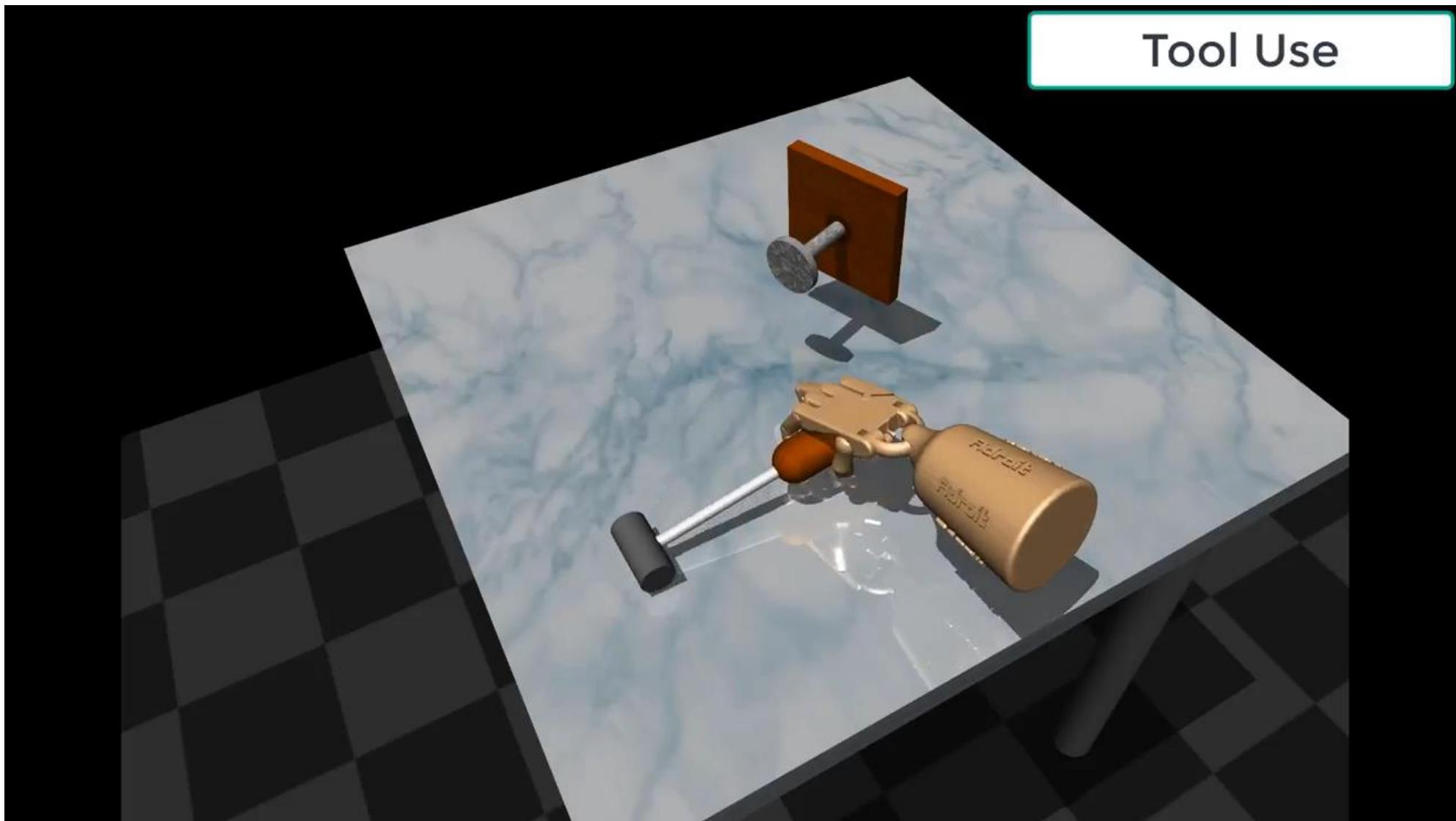


Actions: motor current or torque
 Observations: camera images
 Rewards: task success measure (e.g., running speed)



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

Complex physical tasks



Really complex physical tasks!



Learning and Adapting Agile Locomotion Skills



[lauramsmith.github.io](https://sites.google.com/berkeley.edu/twirl)
<https://sites.google.com/berkeley.edu/twirl>

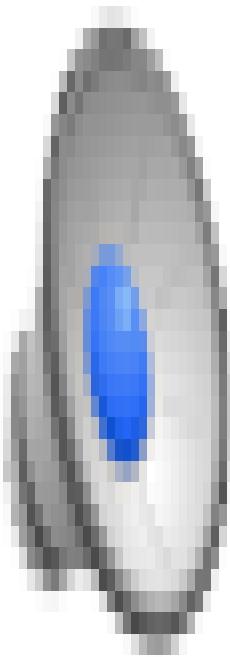
Really really complex physical tasks!

Advanced Skills by Learning Locomotion and Local Navigation End-to-End

Nikita Rudin, David Hoeller,
Marko Bjelonic and Marco Hutter



Unexpected solutions



At scale in the real world

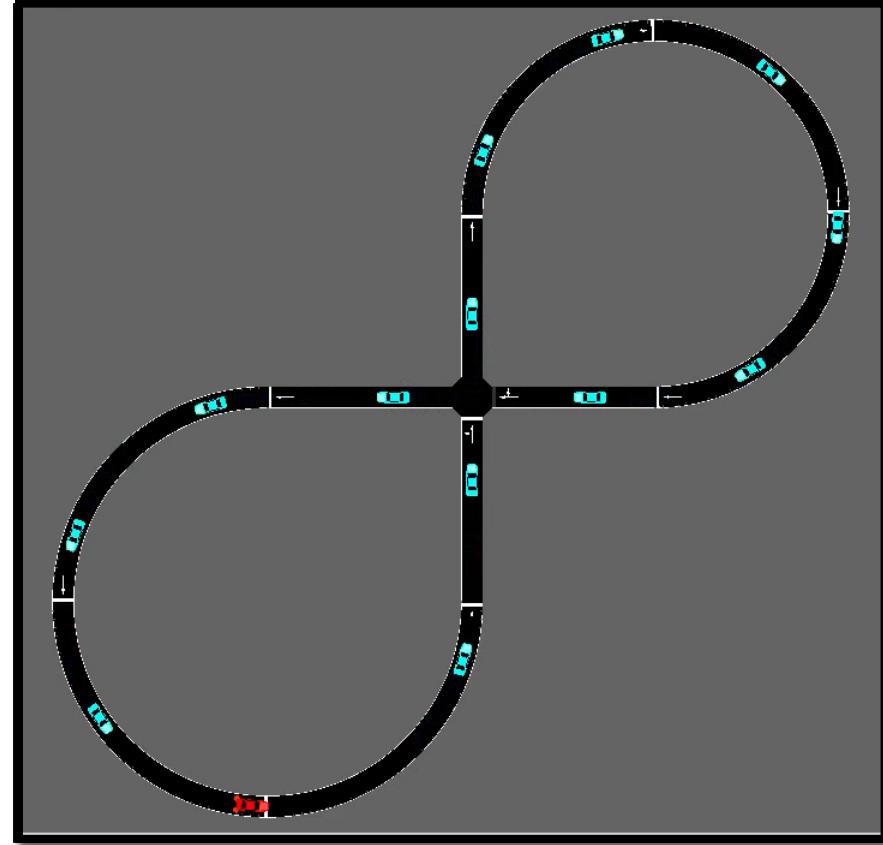
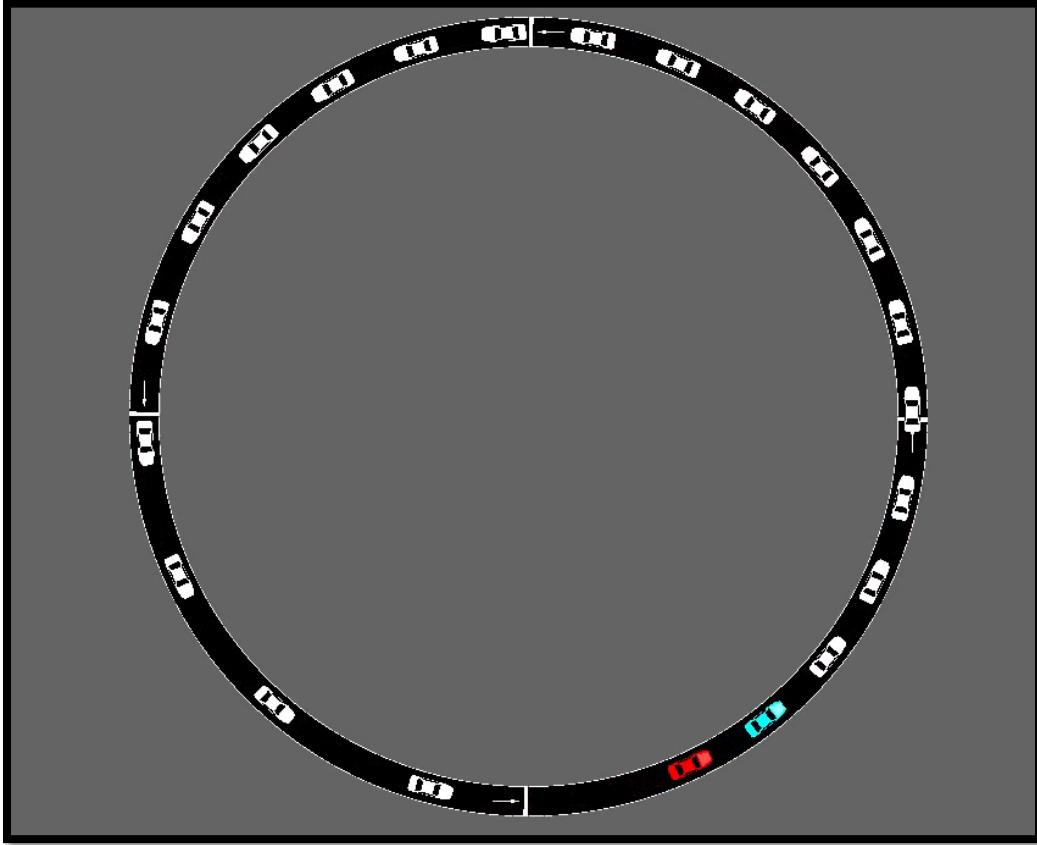


**Deep RL at Scale: Sorting Waste in Office Buildings with
a Fleet of Mobile Manipulators**



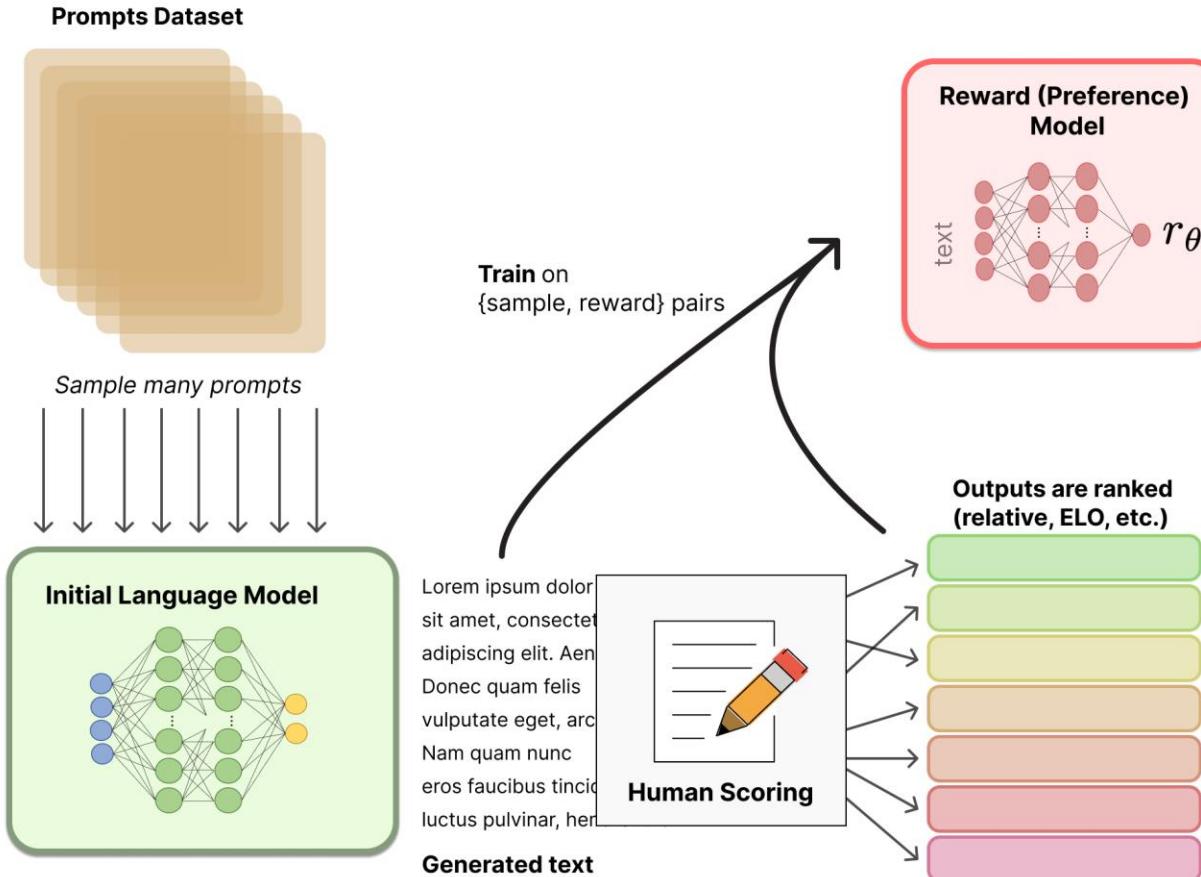
**Deep RL at Scale: Sorting Waste in Office Buildings with
a Fleet of Mobile Manipulators**

Not just games and robots!



Cathy Wu

Reinforcement learning with language models



LE

Can you explain to me how RL with human feedback works for language models like I am five?



Sure! So, imagine you have a robot named Rufus who wants to learn how to talk like a human. Rufus has a language model that helps him understand words and sentences.

First, Rufus will say something using his language model. For example, he might say "I am a robot."

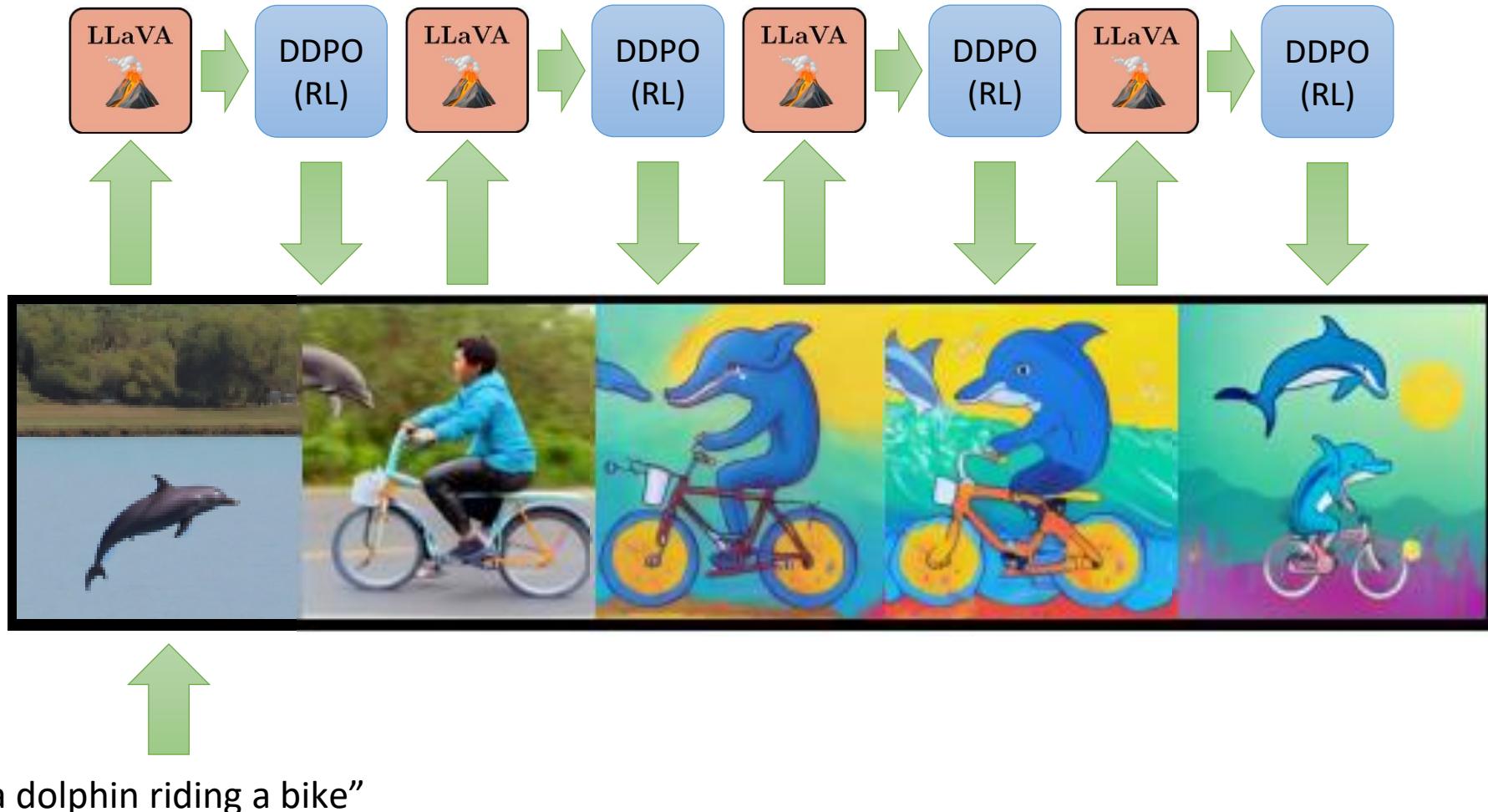
Then, a human will listen to what Rufus said and give him feedback on whether it sounded like a natural sentence a human would say. The human might say, "That's not quite right, Rufus. Humans don't usually say 'I am a robot.' They might say 'I'm a robot' or 'I am a machine.'"

Rufus will take this feedback and use it to update his language model. He will try to say the sentence again, using the new information he received from the human. This time, he might say "I'm a robot."

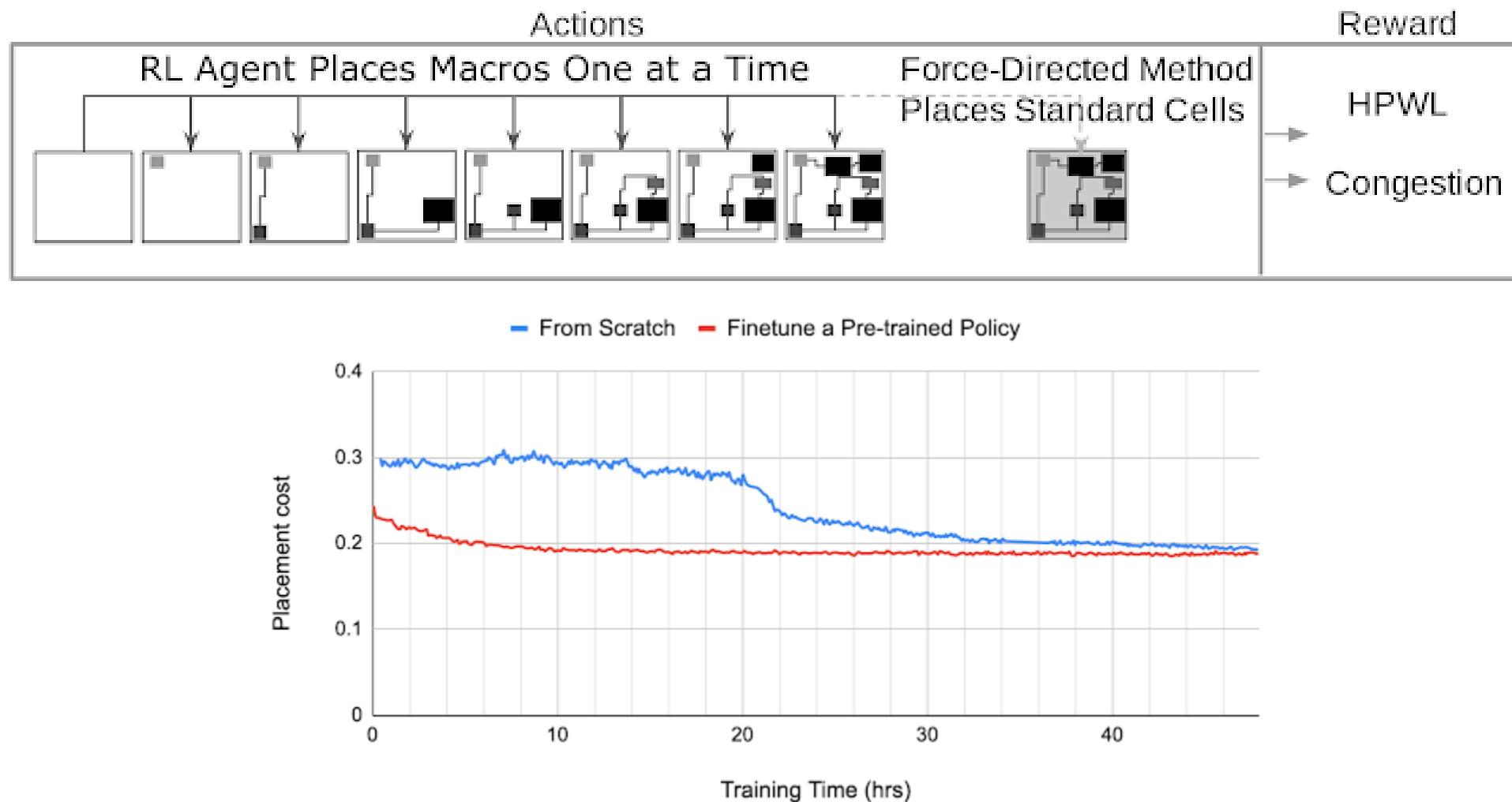
The human will listen again and give Rufus more feedback. This process will continue until Rufus can say sentences that sound natural to a human.

Over time, Rufus will learn how to talk like a human thanks to the feedback he receives from humans. This is how language models can be improved using RL with human feedback.

Reinforcement learning with image generation



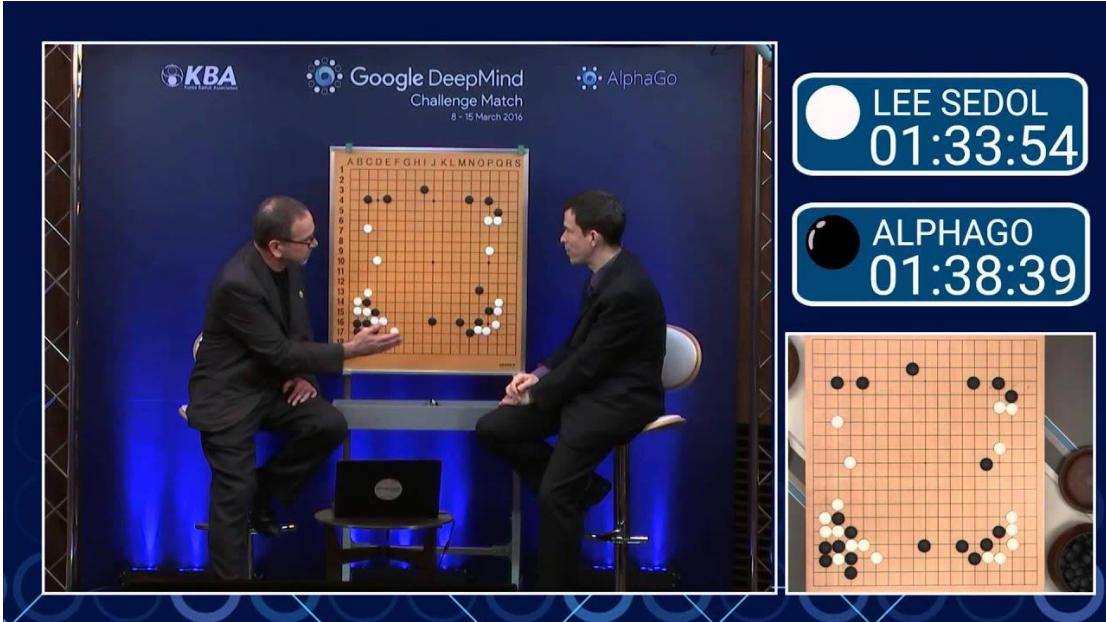
Reinforcement learning for chip design



Why should we study
deep reinforcement learning?

Reinforcement learning can discover new solutions

Impressive because no person had thought of it!

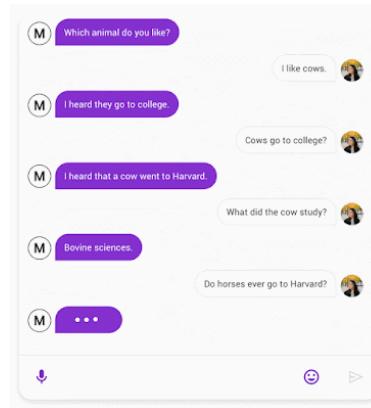


“Move 37” in Lee Sedol AlphaGo match: reinforcement learning “discovers” a move that surprises everyone

Impressive because it looks like something a person might draw!



How does data-driven AI work?



Explaining a joke

Prompt

Explain this joke:

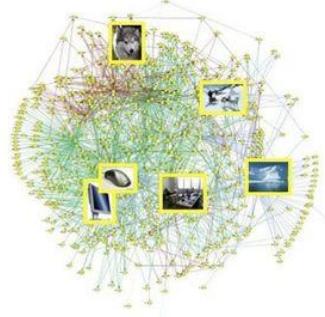
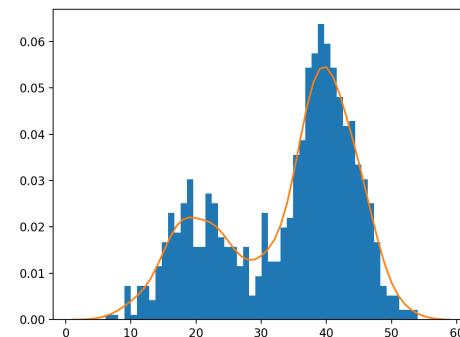
Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

Model Response

Prediction: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

$$p_{\theta}(\mathbf{x})$$

$$p_{\theta}(\mathbf{y}|\mathbf{x})$$



So where does that leave us?

Data-Driven AI



Reinforcement Learning



- + learns about the real world from data
- doesn't try to do **better** than the data

- + optimizes a goal with emergent behavior
- but need to figure out how to use at scale!

Data without optimization
doesn't allow us to solve new problems in new ways

A bitter but misunderstood lesson



Richard Sutton

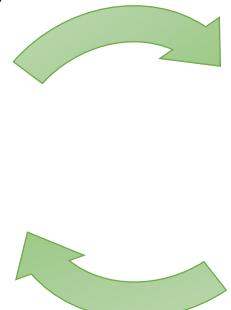
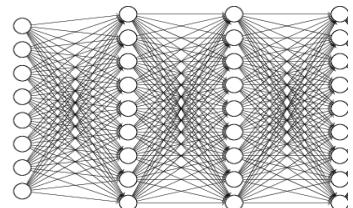
We have to learn the bitter lesson that building in how we think we think does not work in the long run.

The two methods that seem to scale arbitrarily are *learning* and *search*

Data without optimization
doesn't allow us to solve new problems in new ways

Learning

use **data** to extract **patterns**



allows us to **understand** the world

Search

use **computation** to extract **inferences**

optimization

leverages that **understanding** for **emergence**



xkcd

A bit of philosophy

Why do we need **machine learning** anyway?

A postulate:

We need machine learning for one reason and one reason only – that's to produce **adaptable and complex decisions**.



Decision: how do I move my joints?



Decision: how do I steer the car?



What is the decision? The image label?

What happens with that label afterwards?

Aside: why do we need **brains** anyway?



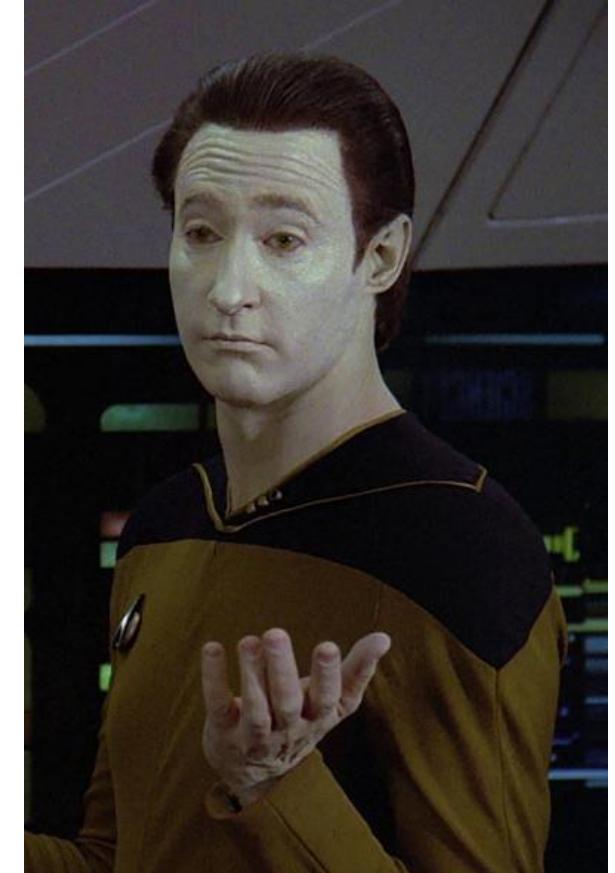
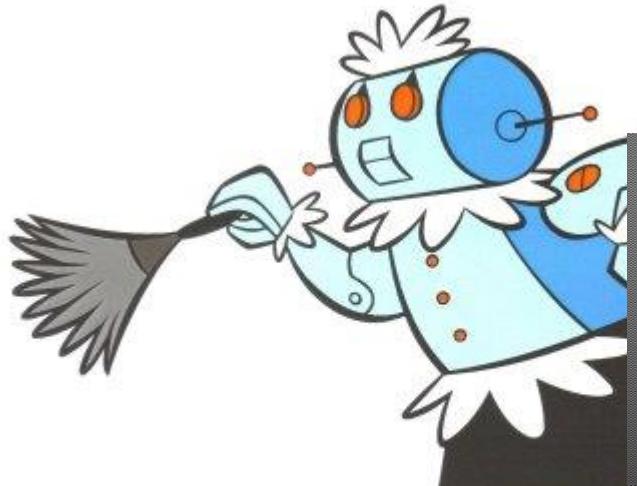
Daniel Wolpert

(knows quite a lot about brains)



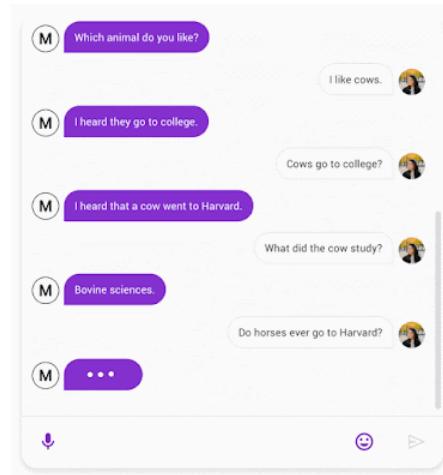
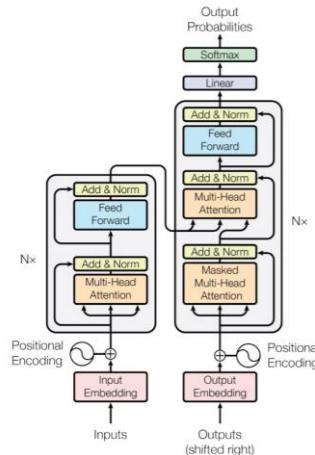
"We have a brain for one reason and one reason only -- that's to produce adaptable and complex movements. Movement is the only way we have affecting the world around us... I believe that to understand movement is to understand the whole brain."

How do we build intelligent machines?

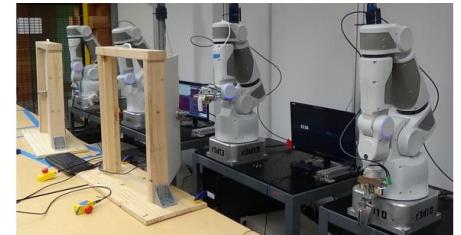
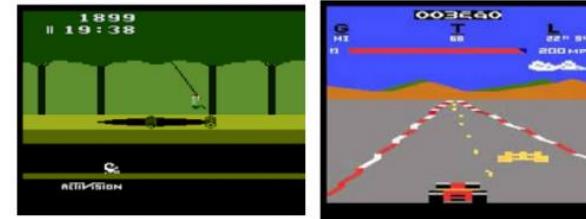


Why should we study this now?

big end-to-end trained models work quite well!



we have RL algorithms that we can feasibly combine with deep networks



and yet learning-based control in truly real-world settings remains a major open problem!

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

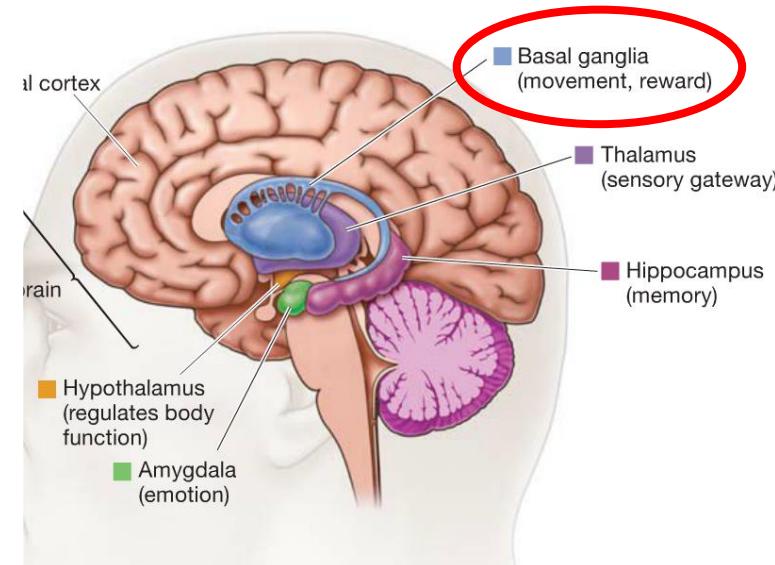
Where do rewards come from?



reinforcement learning agent

[-] 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

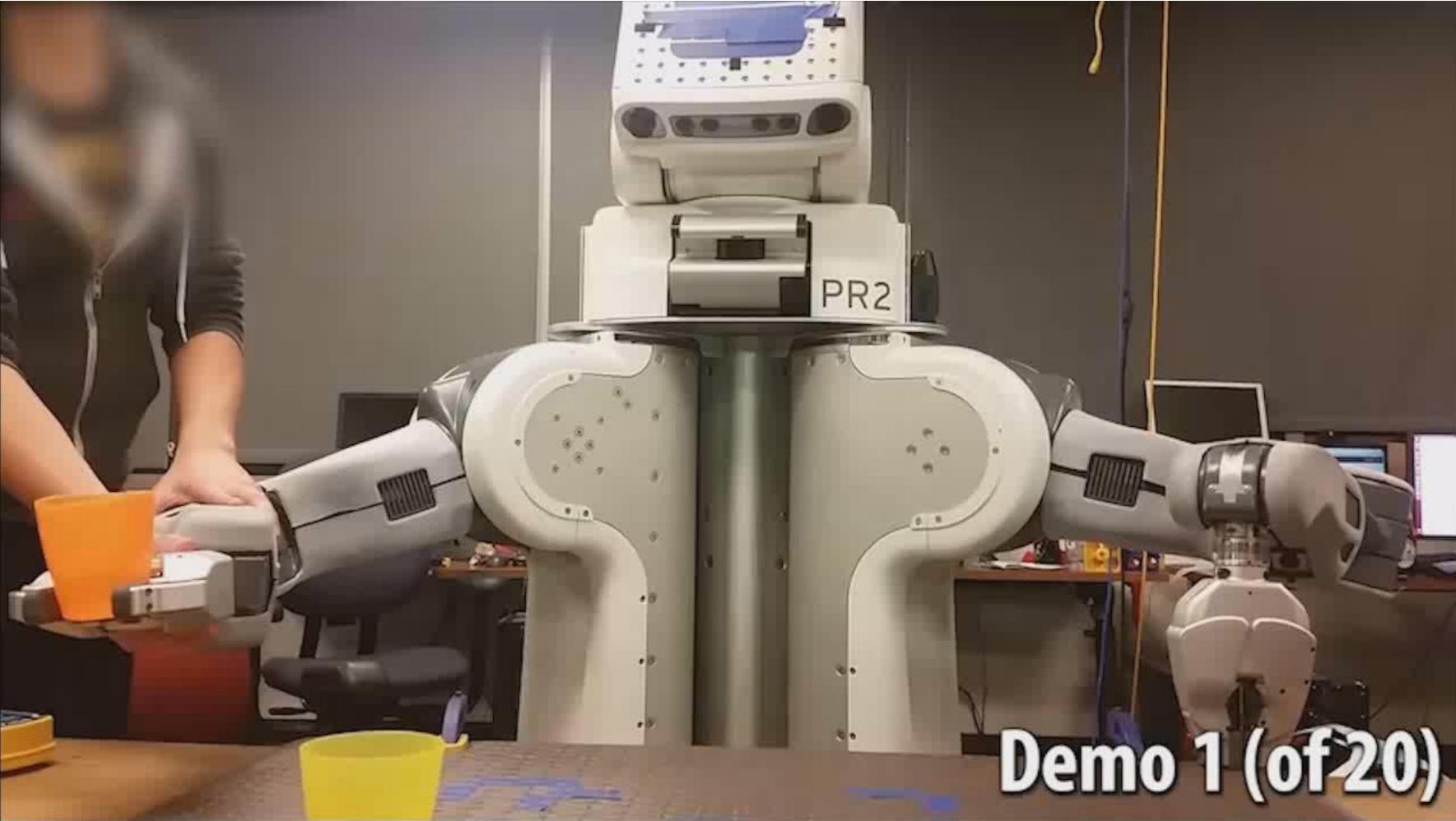
Imitation learning



More than imitation: inferring intentions



Inverse RL examples



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,² and Daniel M. Wolpert²

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

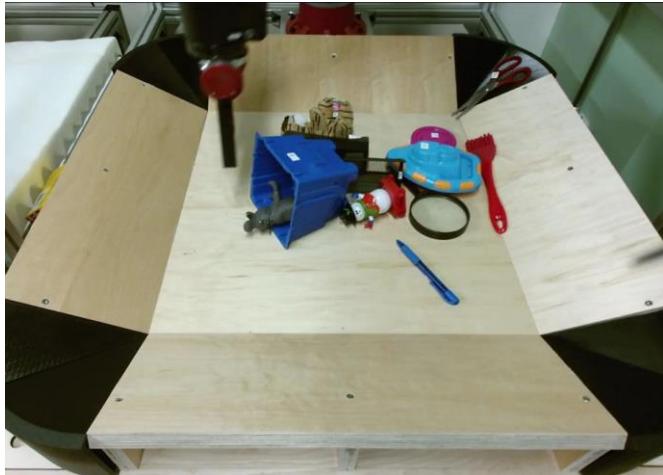
Sarah J. Blakemore, Susan J. Goodbody, and Daniel M. Wolpert

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

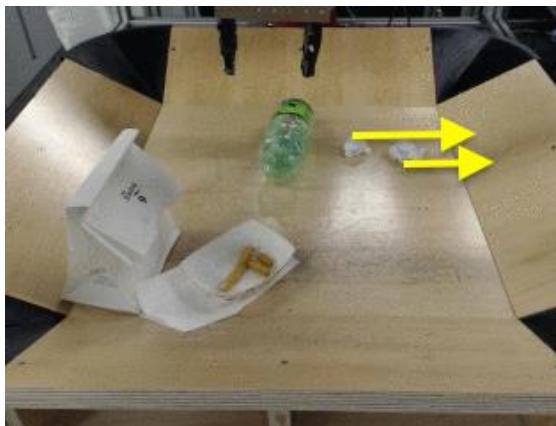
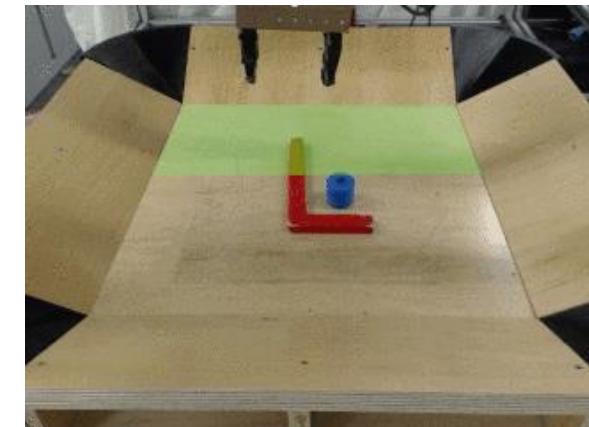
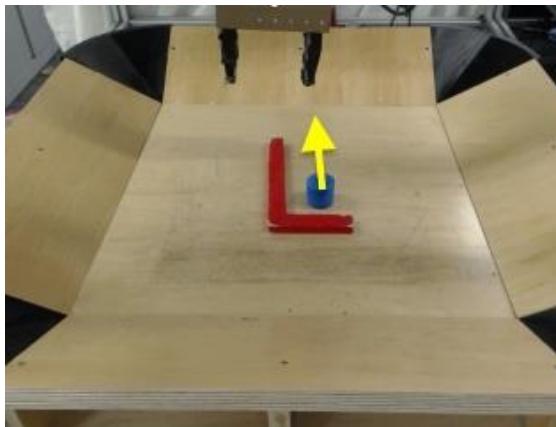
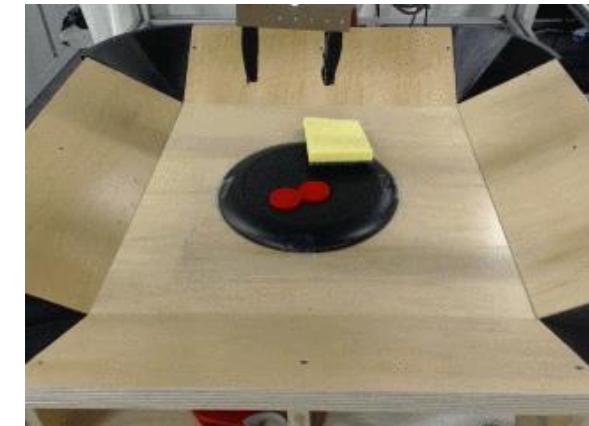
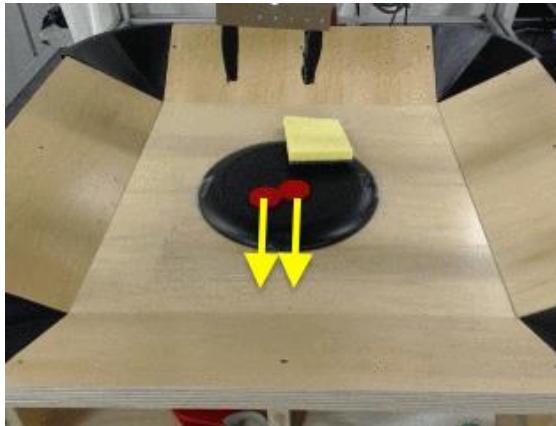
Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects

Rajesh P. N. Rao³ and Dana H. Ballard²

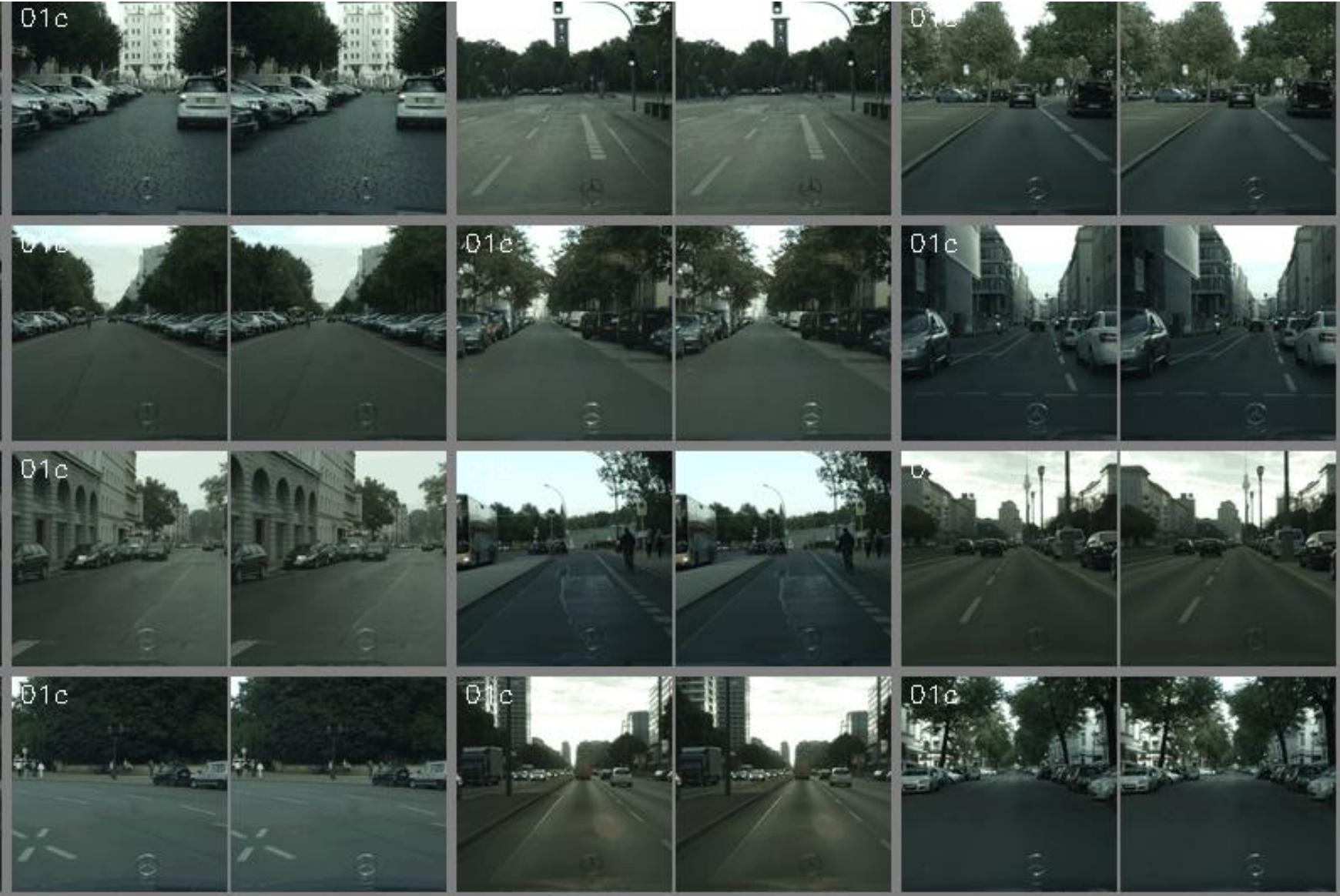
Prediction for real-world control



Using tools with predictive models



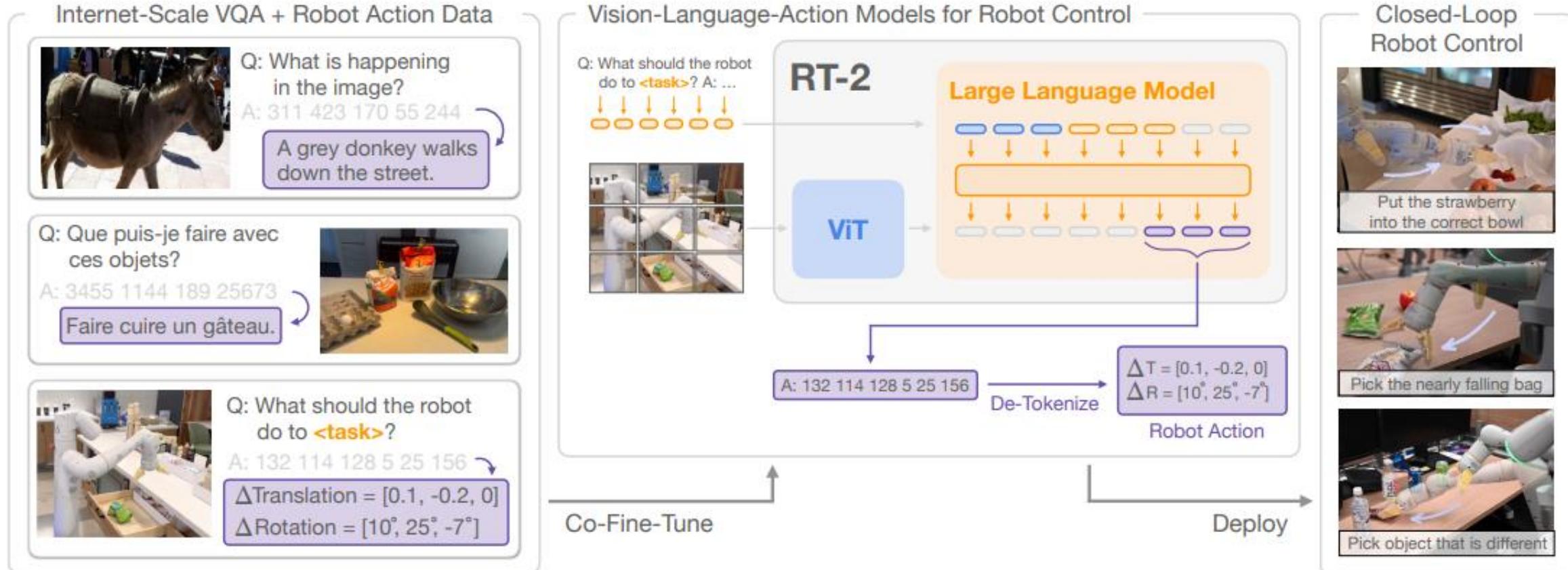
Predictive models have come a long way!



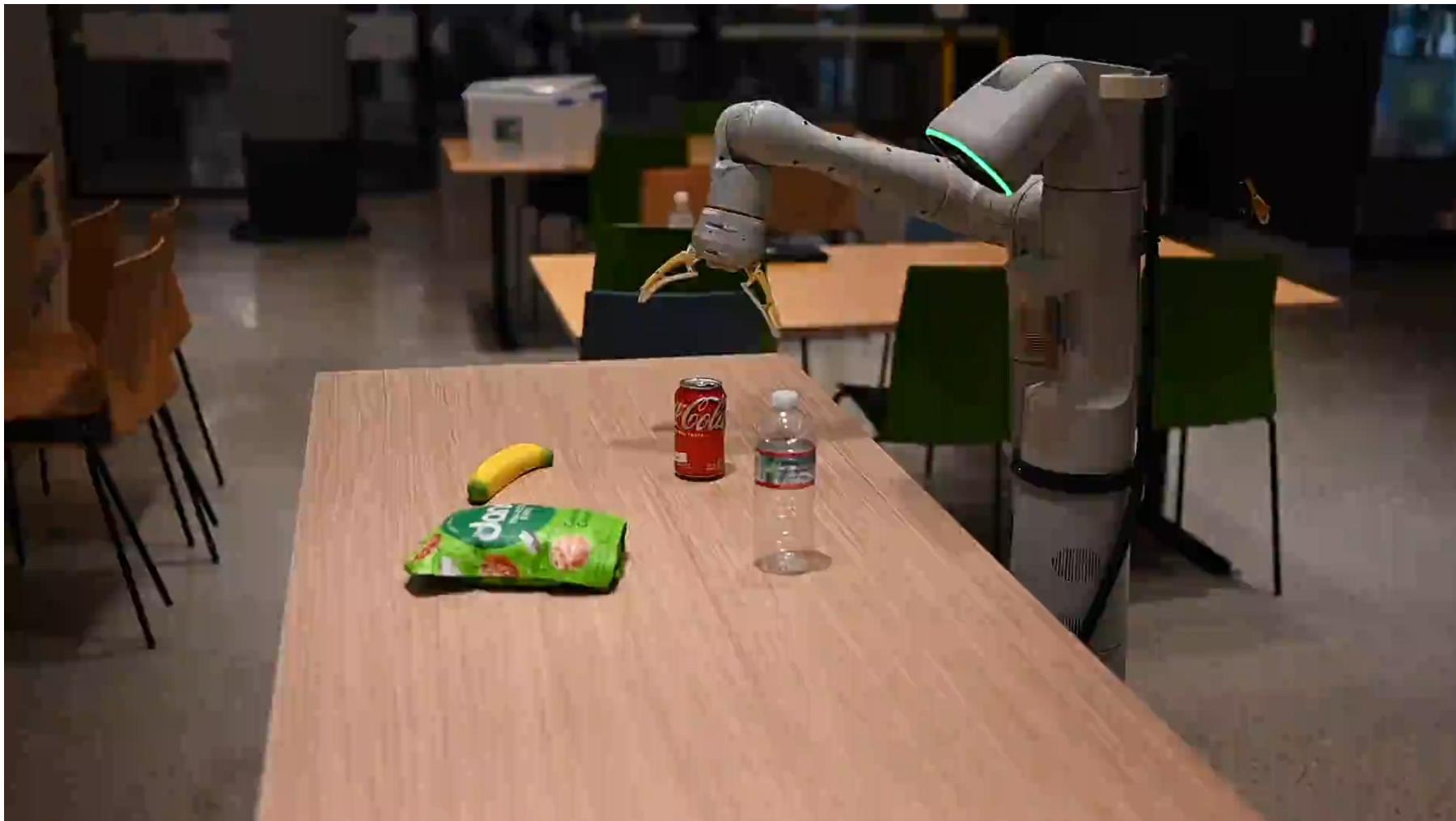
Predictive models have come a long way!



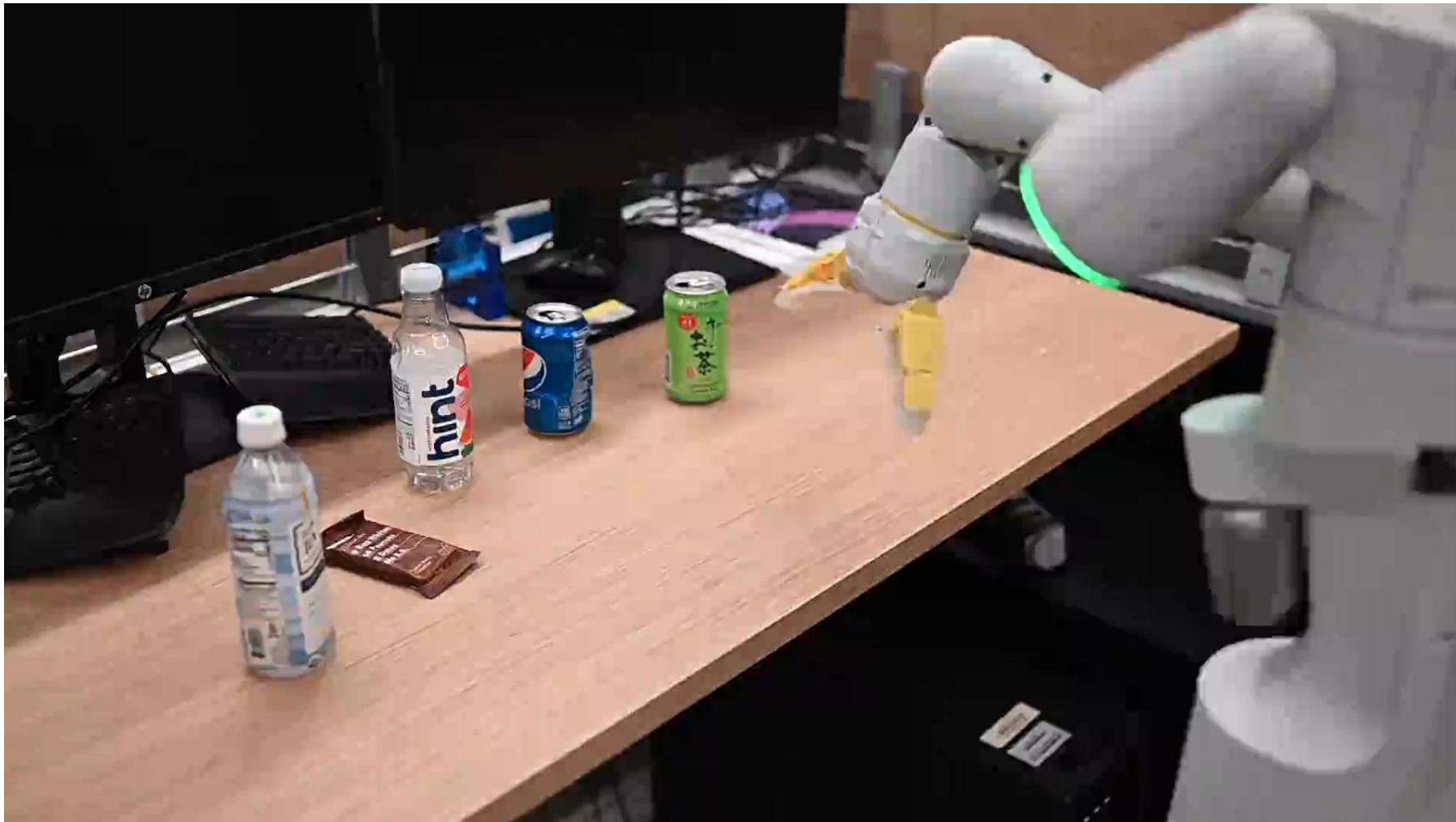
Leveraging advances in pretrained models



Leveraging advances in pretrained models



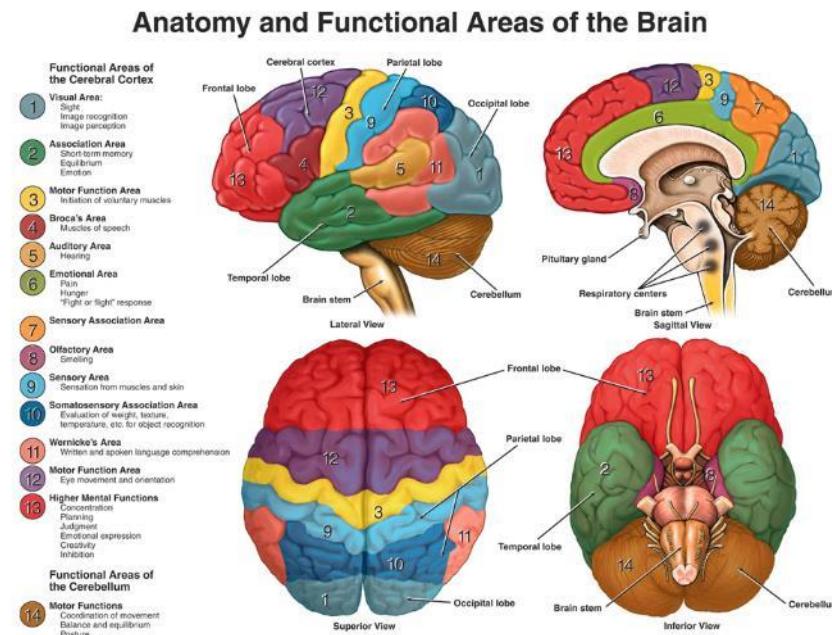
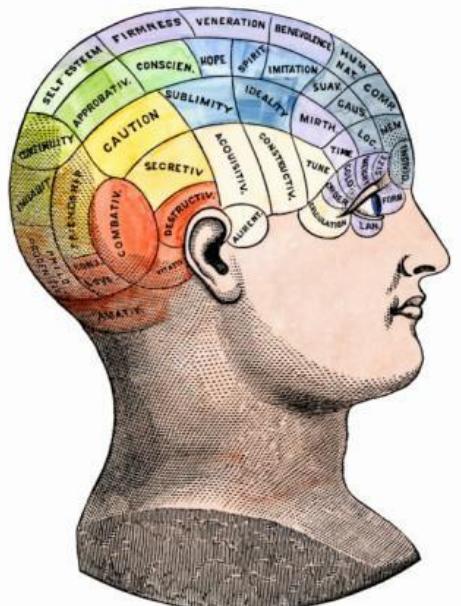
Leveraging advances in pretrained models



How do we build intelligent machines?

How do we build intelligent machines?

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



Learning as the basis of intelligence

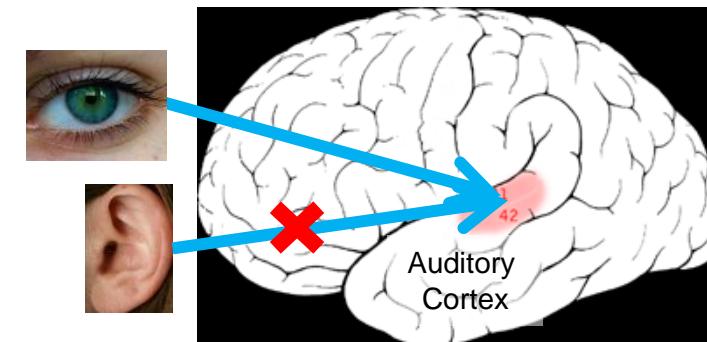
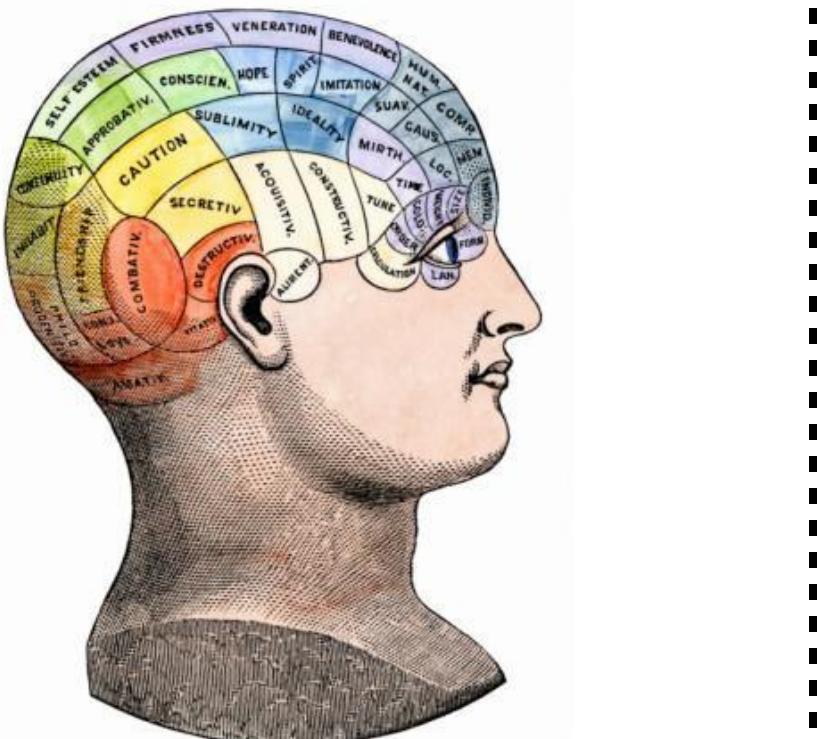
- Some things we can all do (e.g. walking)
- Some things we can only learn (e.g. driving a car)
- We can learn a huge variety of things, including very difficult things
- Therefore our learning mechanism(s) are likely powerful enough to do everything we associate with intelligence
 - But it may still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

A single algorithm?

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



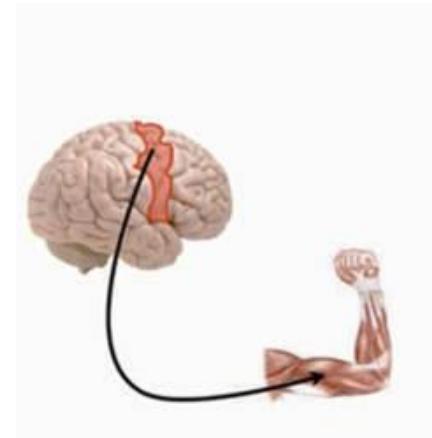
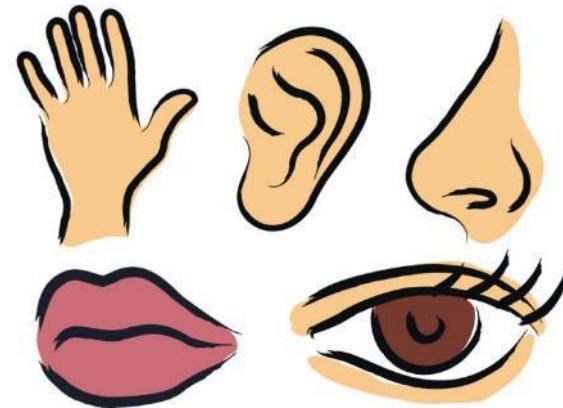
Seeing with your tongue



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

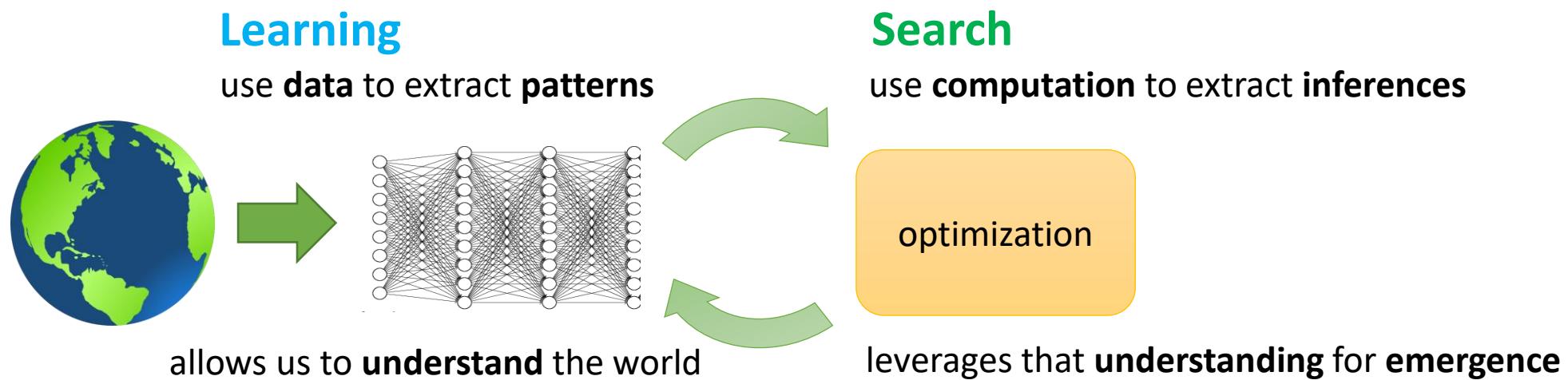
What must that single algorithm do?

- Interpret rich sensory inputs
- Choose complex actions

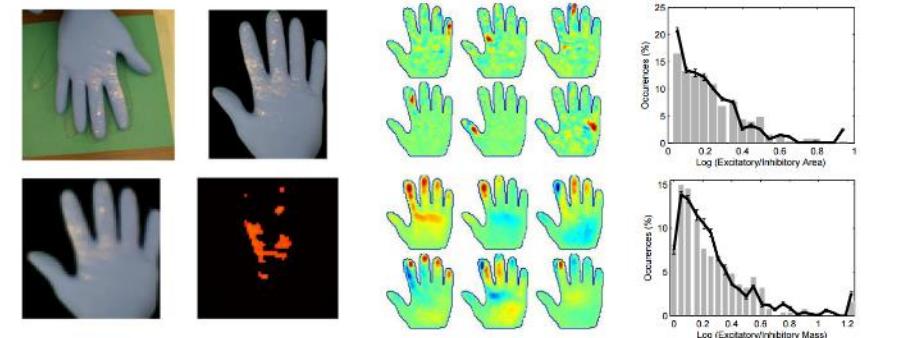
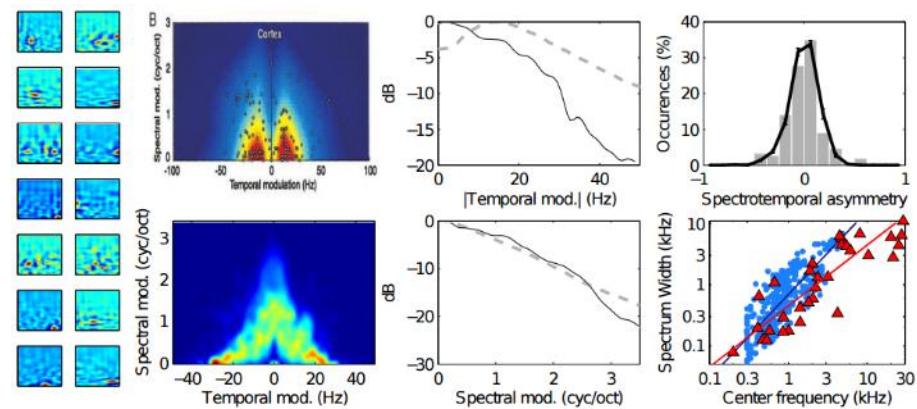
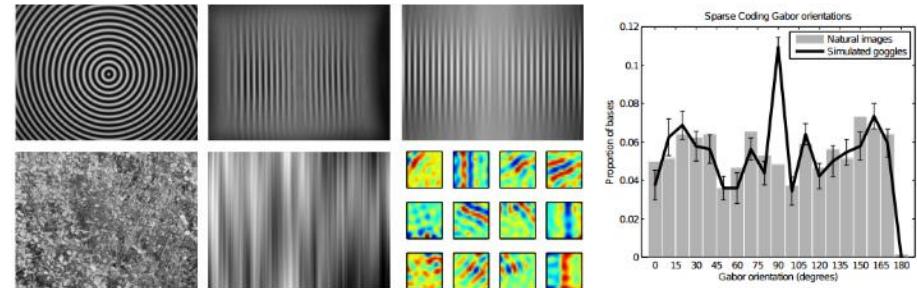


Why deep reinforcement learning?

- Deep = scalable learning from large, complex datasets
- Reinforcement learning = optimization



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

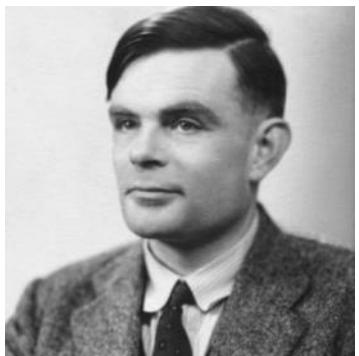
Yael Niv

Psychology Department & Princeton Neuroscience Institute, Princeton University

What challenges still remain?

- We have great methods that can learn from huge amounts of data
- We have great optimization methods for RL
- We don't (yet) have amazing methods that **both** use data and RL
- Humans can learn incredibly quickly, deep RL methods are usually slow
- Humans reuse past knowledge, transfer learning in 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

