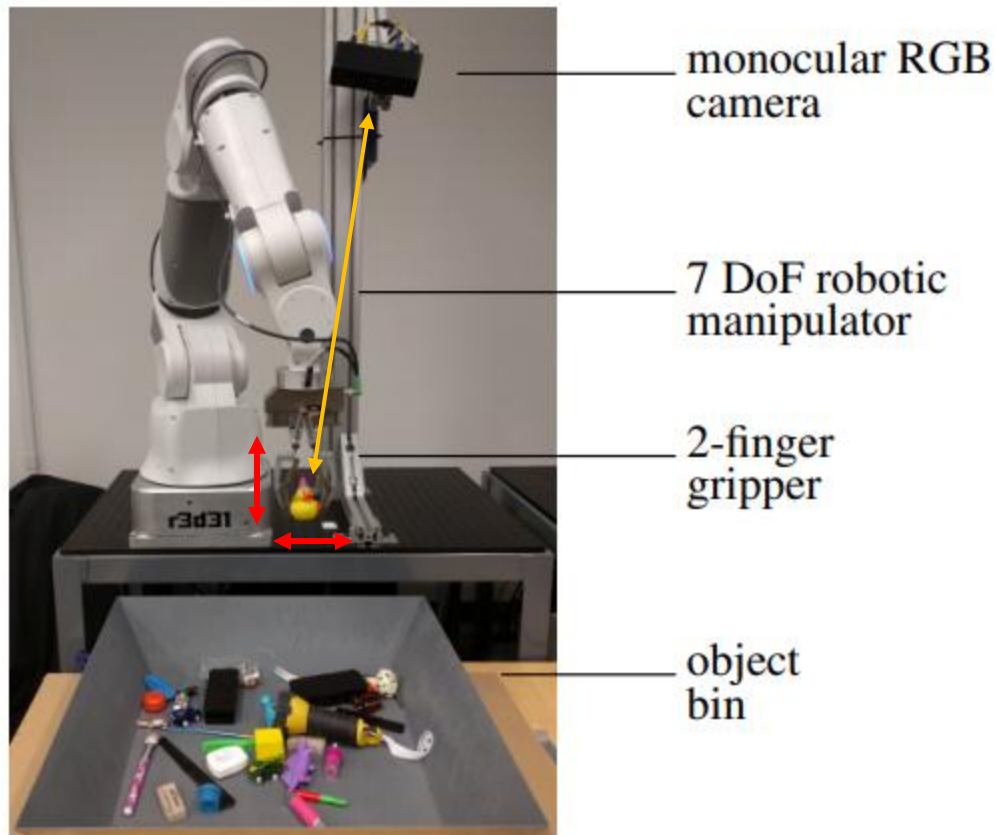


Deep Reinforcement Learning

CS 185/285

Instructor: Sergey Levine
UC Berkeley





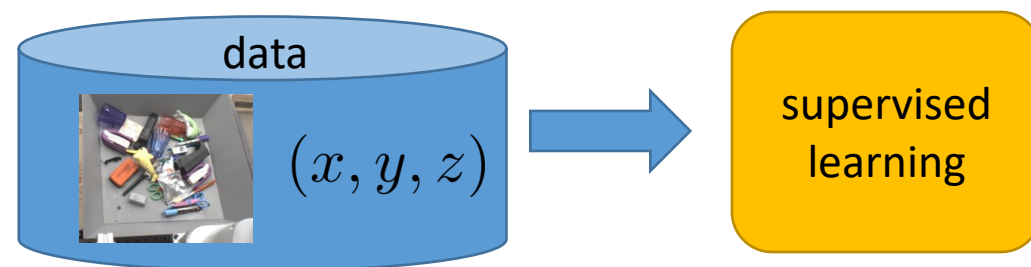
Option 1:

Understand the problem, design a solution

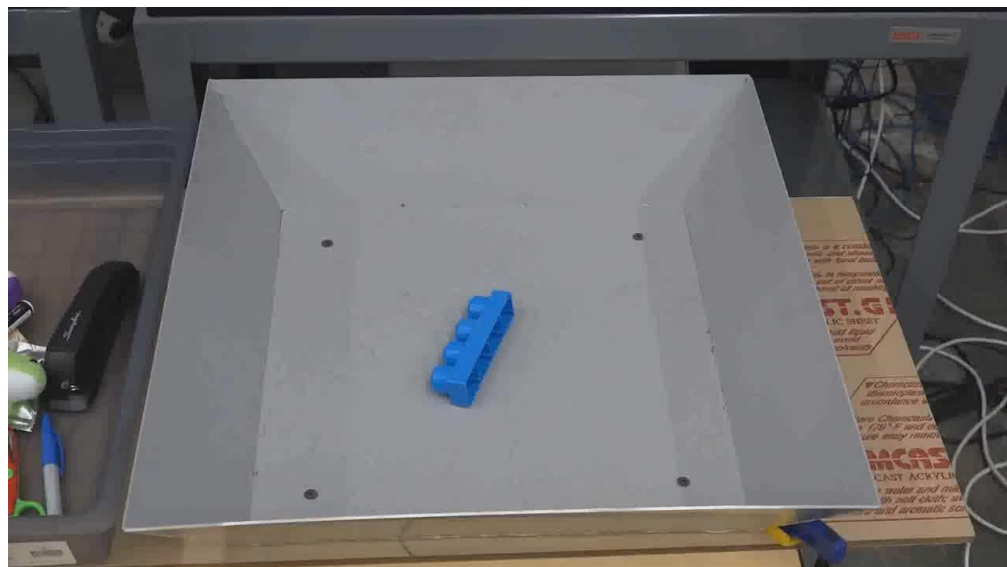
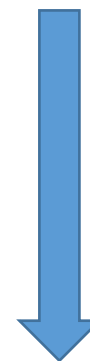


Option 2:

Set it up as a machine learning problem







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



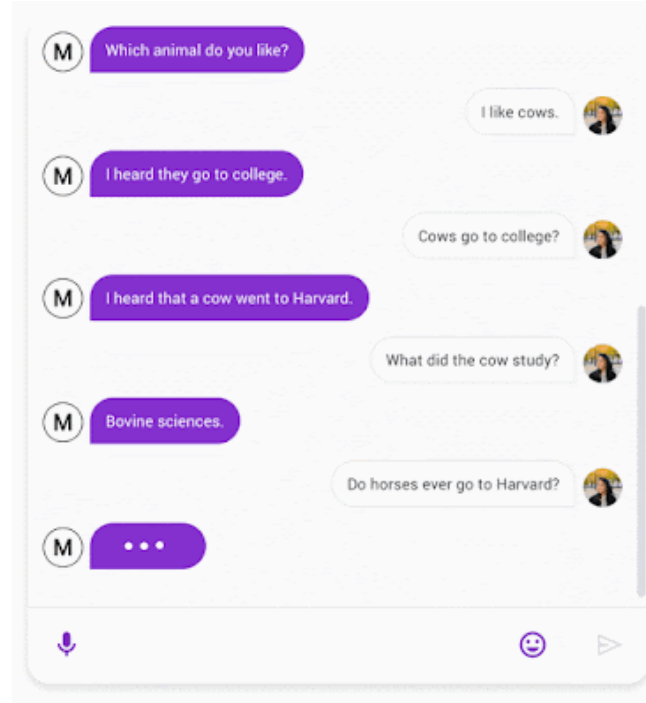
an espresso machine that makes coffee from human souls, artstation



panda mad scientist mixing sparkling chemicals, artstation



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



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.

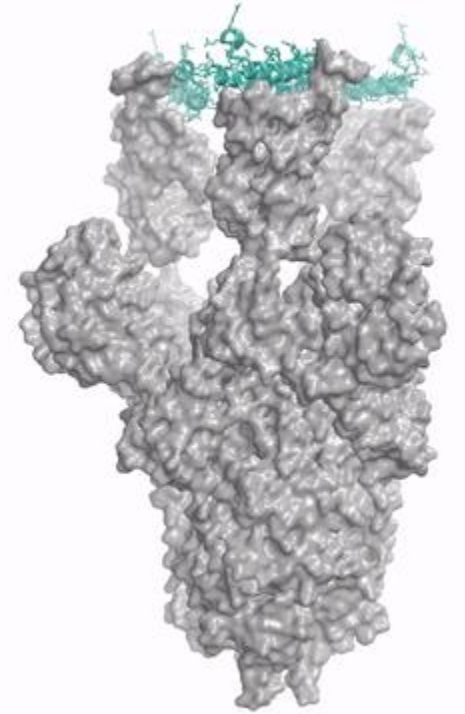
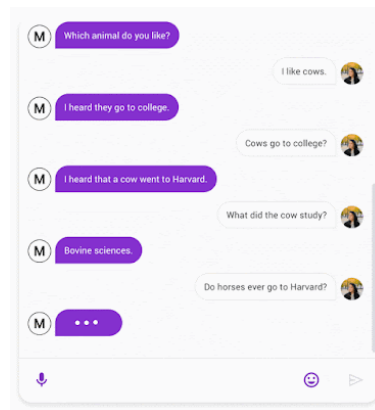


image credit: UW IPD

What's the main idea behind this?



Explaining a joke

Prompt

Explain this joke:

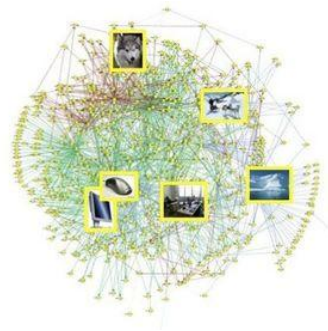
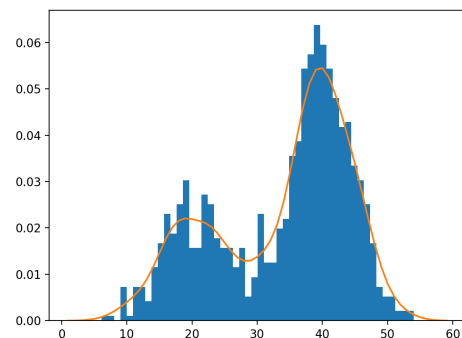
Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

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.

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

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

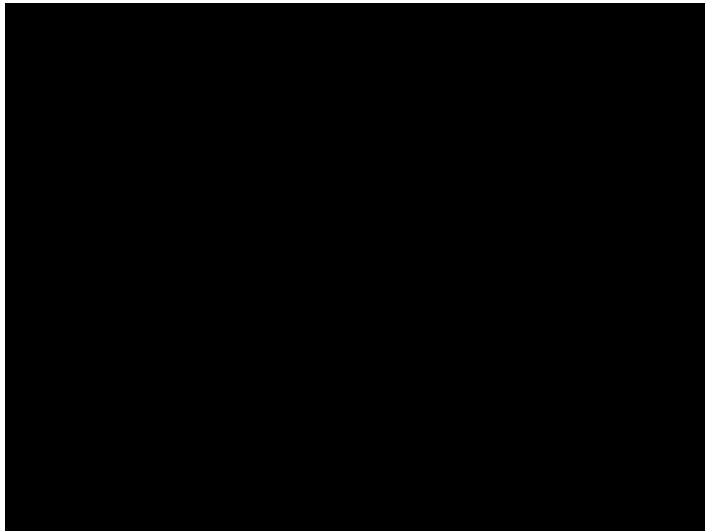


What does reinforcement learning do differently?

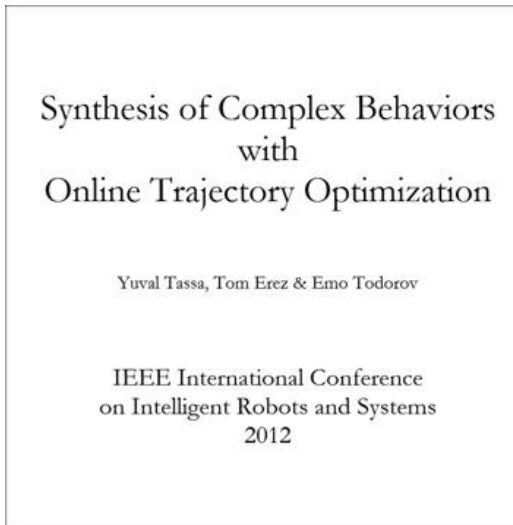
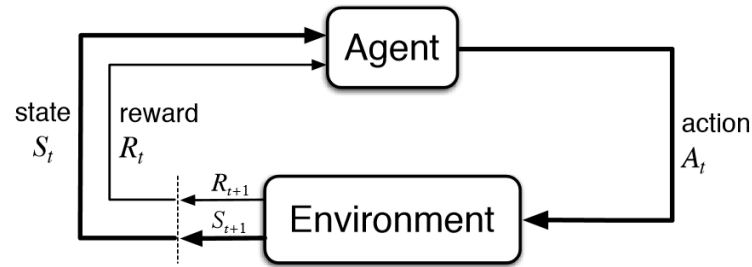


classical reinforcement learning

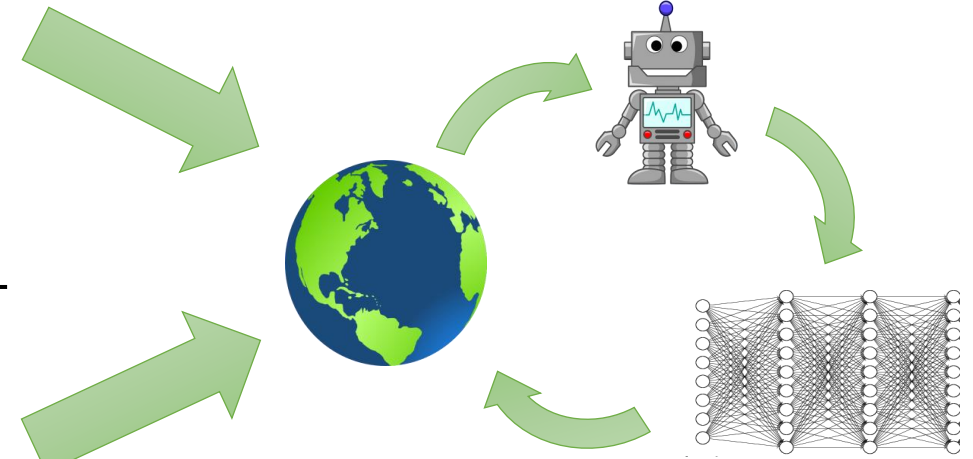
evolutionary algorithms, controls, optimization



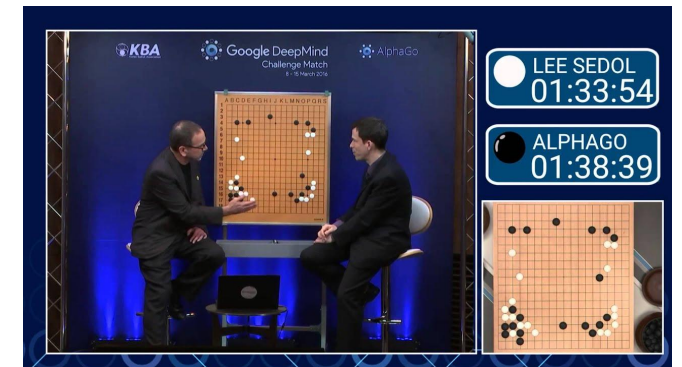
Evolved Virtual Creatures.
Karl Sims, 1994



Model-Predictive Control with iLQG.
Yuval Tassa, 2012



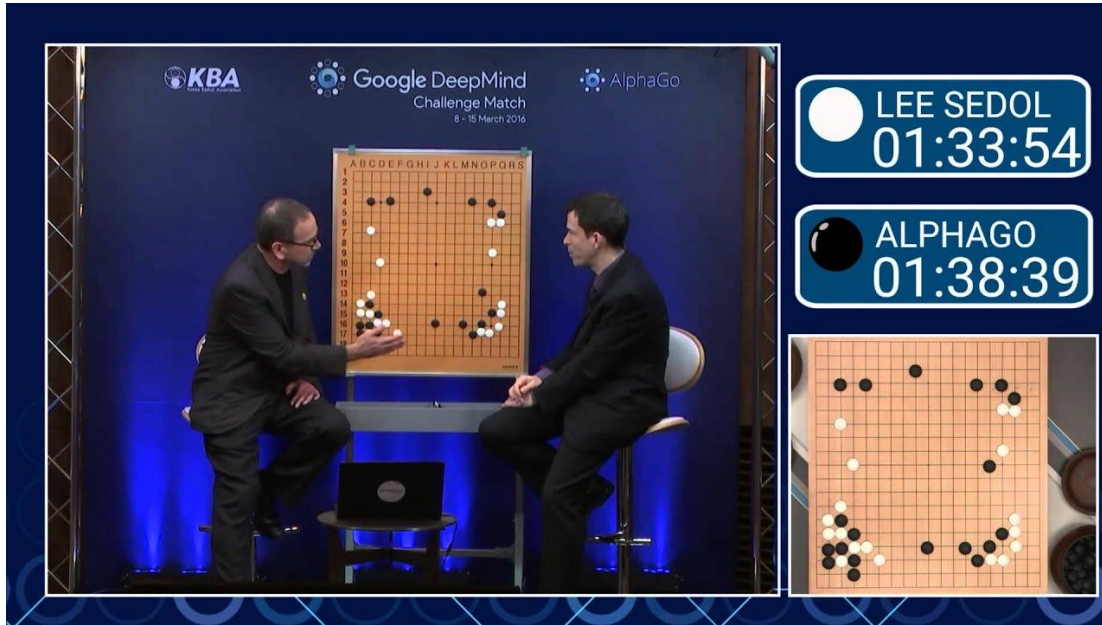
deep reinforcement learning



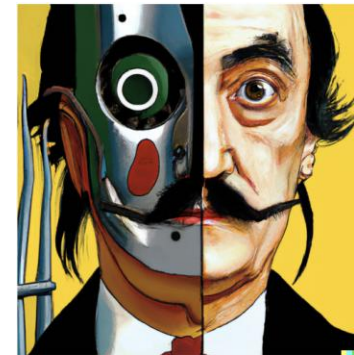
Reinforcement learning can discover new solutions

Impressive because no person had thought of it!

Impressive because it looks like something a person might draw!



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



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

Part 1:

Course logistics

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
- Assignments will require training neural networks with standard automatic differentiation packages (PyTorch)
 - We'll review this in section
- Concurrent enrollment students, please **fill out enrollment form:**
<http://rail.eecs.berkeley.edu/deeprlcourse/>
- For everyone else, if you cannot enroll directly, please add yourself to the waitlist in CalCentral

Class format

- **Lectures**

- Lectures will be posted on bCourses (“Media Gallery”)
 - <https://bcourses.berkeley.edu/courses/1550524>
- I’ll also link a YouTube playlist on **Ed** and on the **course website**
- **Every lecture has a quiz**, to be completed within a week of the lecture (10% of grade)
 - The quiz is not supposed to be hard
 - Check your understanding, lightweight and quick
 - Also serves to summarize the key concepts you should know
 - You can try it twice, we’ll take the higher of the two scores

- **Sections**

- TAs will organize sections most weeks starting next week (see Ed for schedule)
- Sections will review material in lecture, rather than covering new materials

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

- **Homeworks (50% of the grade)**

- There are 5 total programming homeworks
- About two weeks per homework
- Don't use AI coding tools (I know...)
- Come to office hours for help if you need it
- Should be possible for everyone to get full score on each homework!

- **Exam (20% of the grade)**

- Held in week 12 of the course (toward the end)
- We'll have a practice exam
- Quasi-final, but before final exam period

- **Final project (20% of the grade)**

- Programming project with a few options
- More info soon!

Part 2:

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?

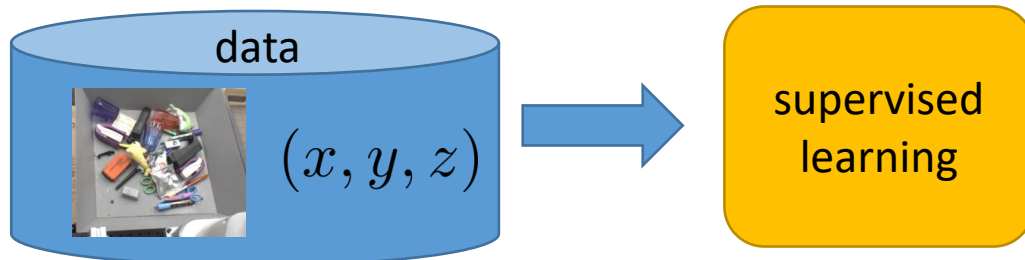
Standard (supervised) machine learning:

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

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

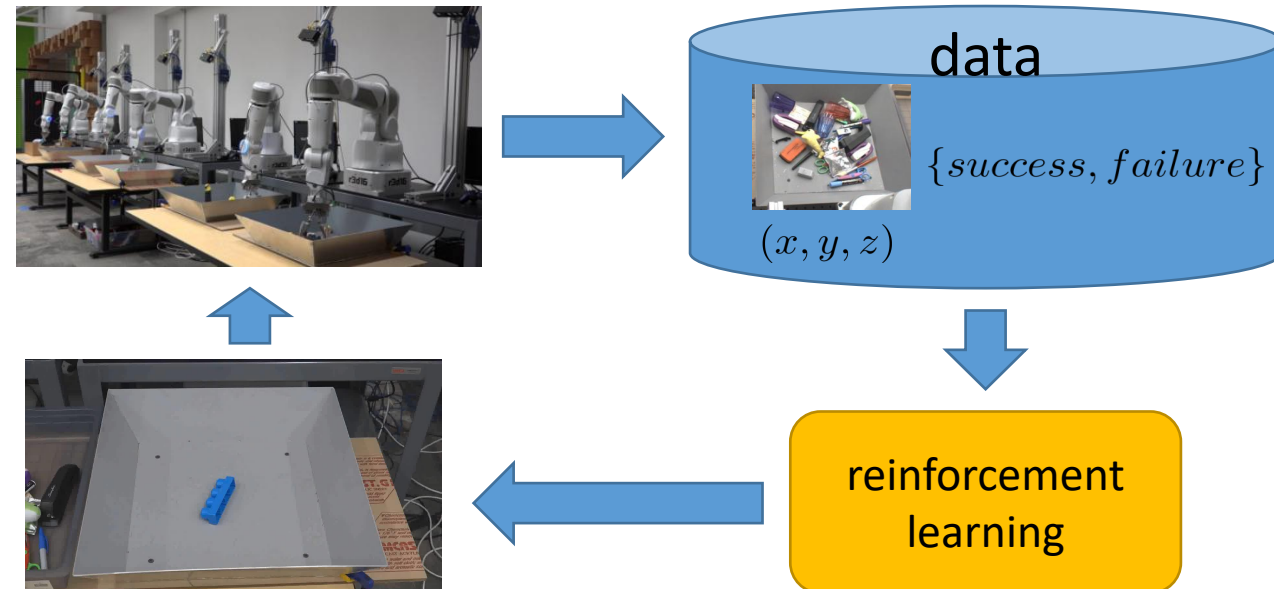
Usually assumes:

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



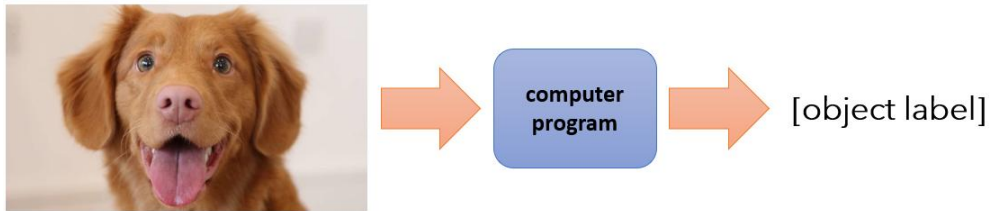
Reinforcement learning:

- 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?

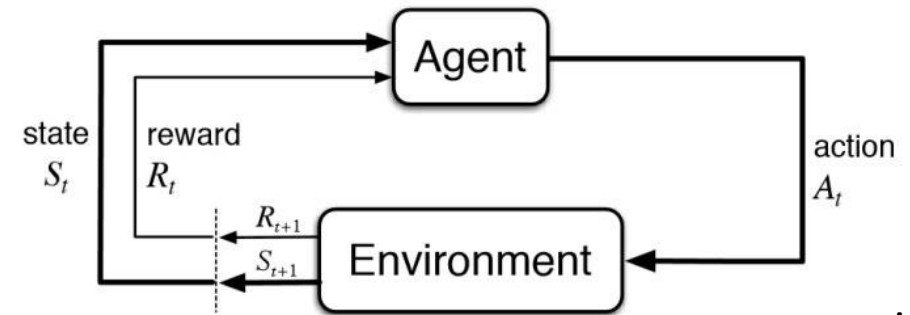
supervised learning



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

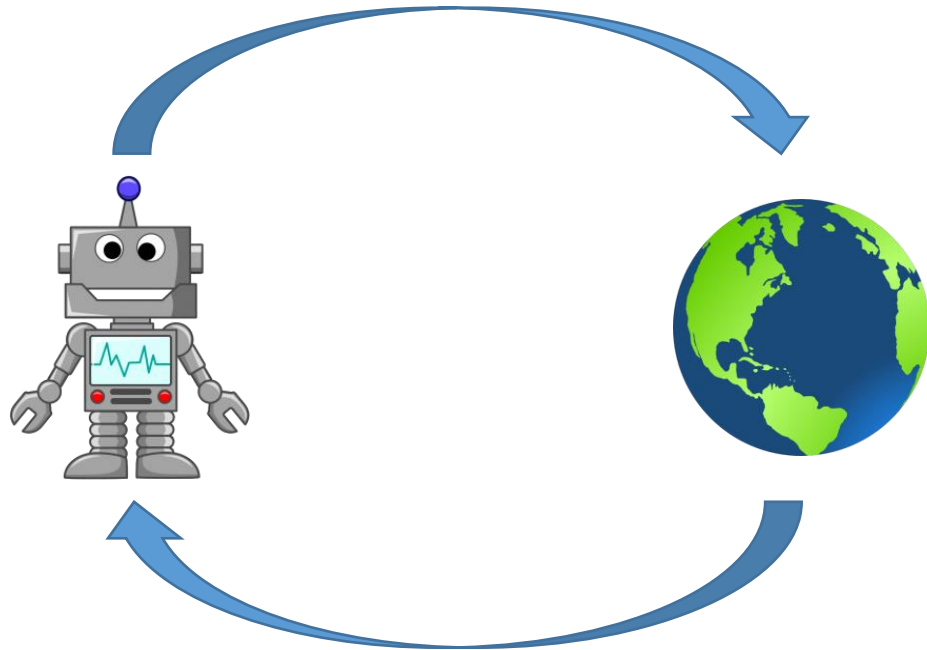
reinforcement learning



input: \mathbf{s}_t at each time step
output: \mathbf{a}_t at each time step
data: $(\mathbf{s}_1, \mathbf{a}_1, r_1, \dots, \mathbf{s}_T, \mathbf{a}_T, r_T)$
goal: learn $\pi_{\theta} : \mathbf{s}_t \rightarrow \mathbf{a}_t$
to maximize $\sum_t r_t$

pick your own actions

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., running speed)



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

Complex physical tasks!

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

Nikita Rudin, David Hoeller,
Marko Bjelonic and Marco Hutter



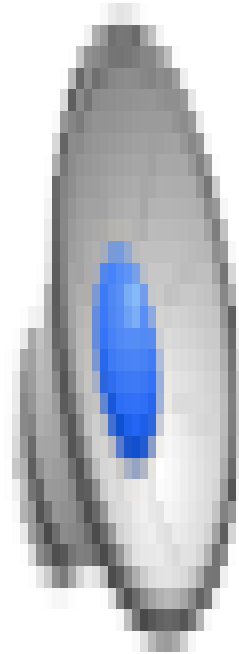
Really complex physical tasks!



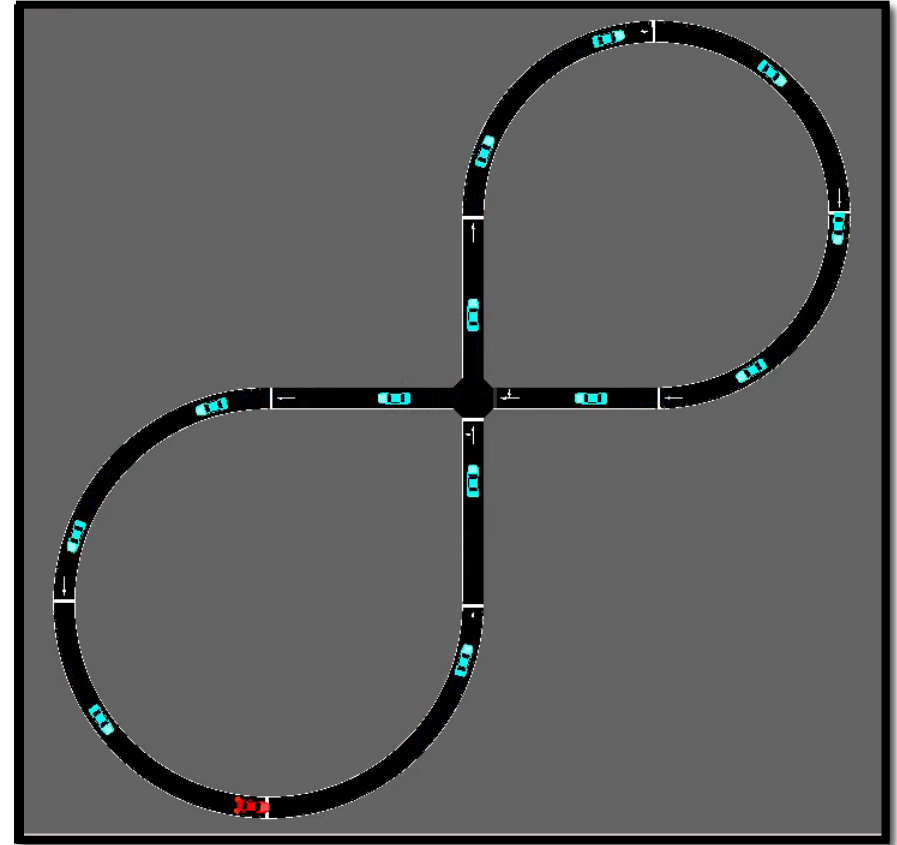
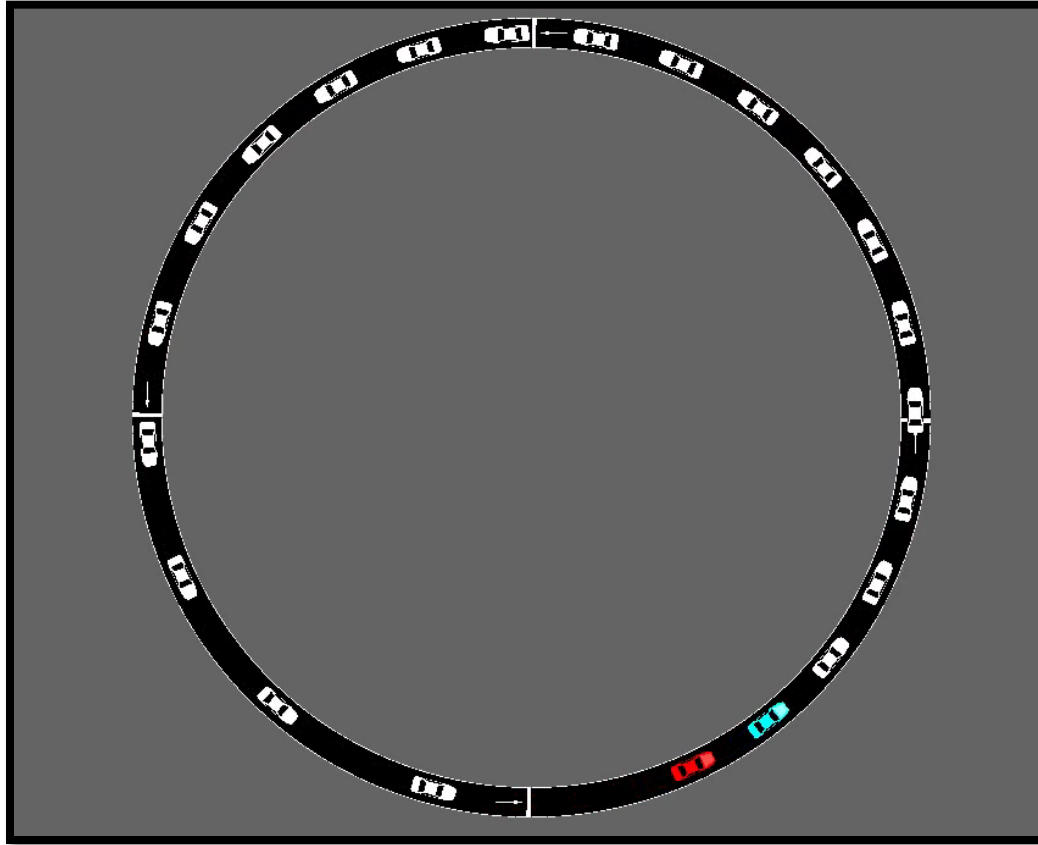
Really really complex physical tasks!



Unexpected solutions

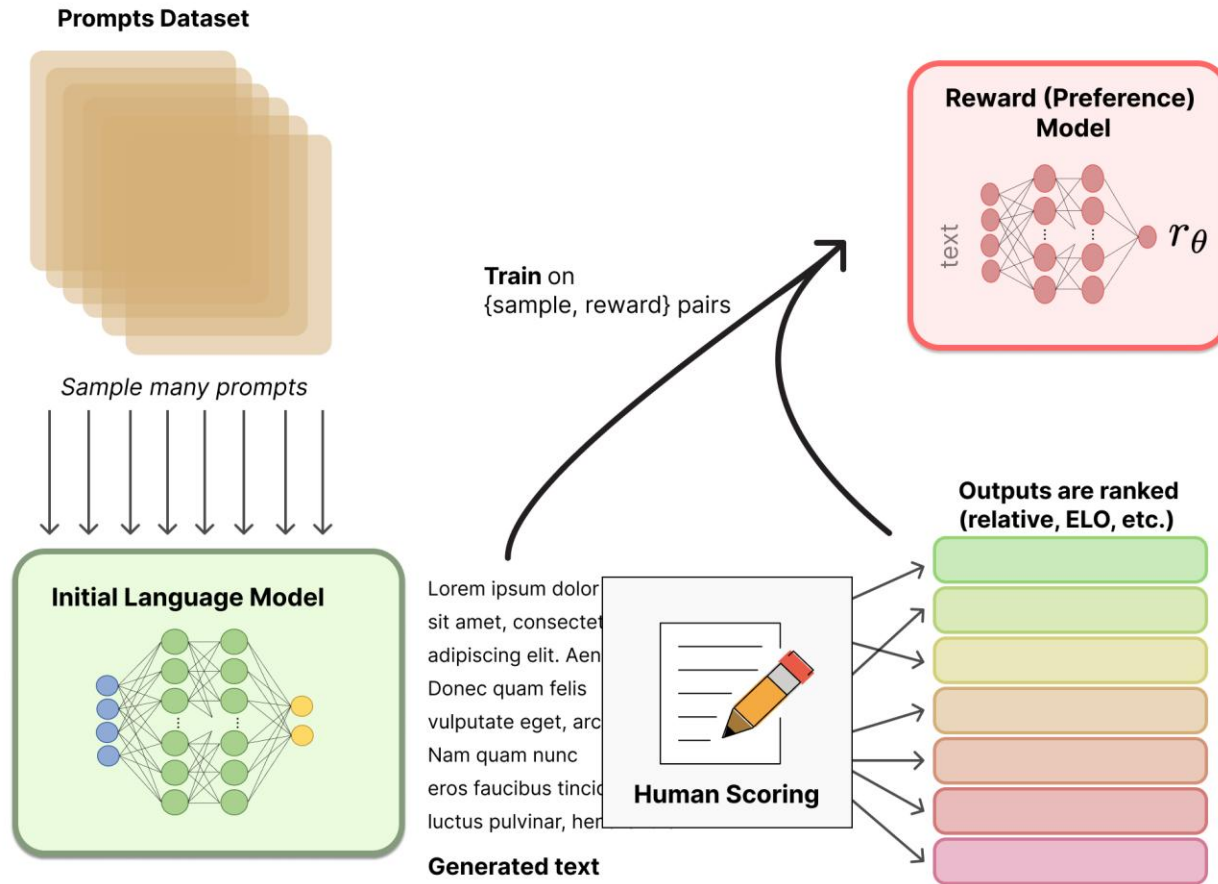


Not just games and robots!



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."

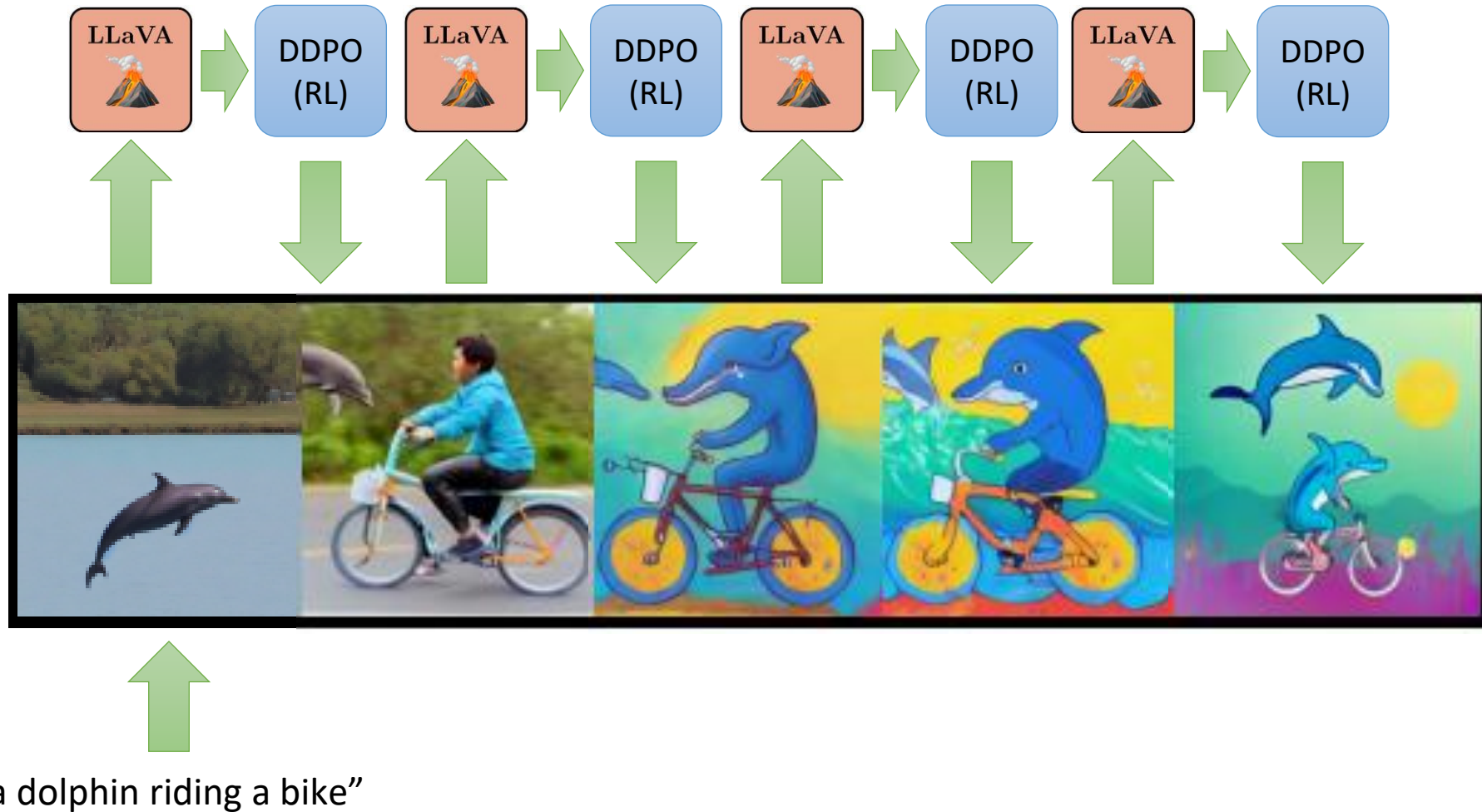
Then, a human will listen to what Rufus said and give him feedback on whether it sounded like a natural sentence a human would say. The human might say, "That's not quite right, Rufus. Humans don't usually say 'I am a robot.' They might say 'I'm a robot' or 'I am a machine.'"

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."

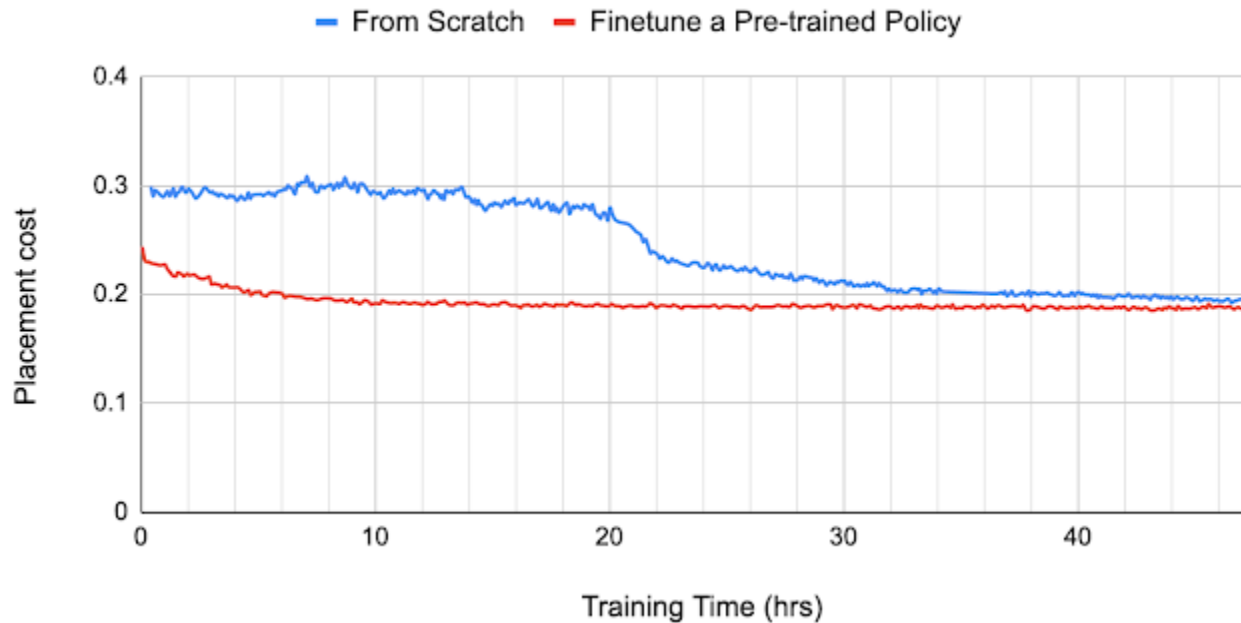
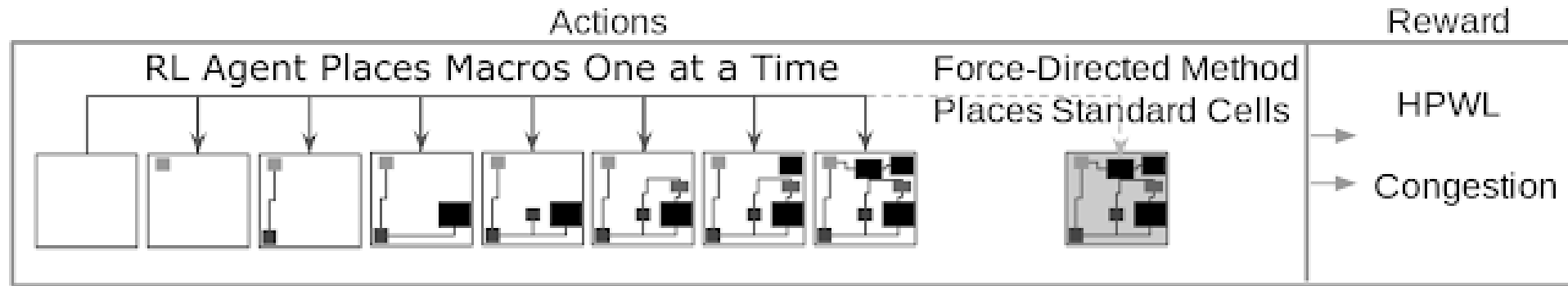
The human will listen again and give Rufus more feedback. This process will continue until Rufus can say sentences that sound natural to a human.

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



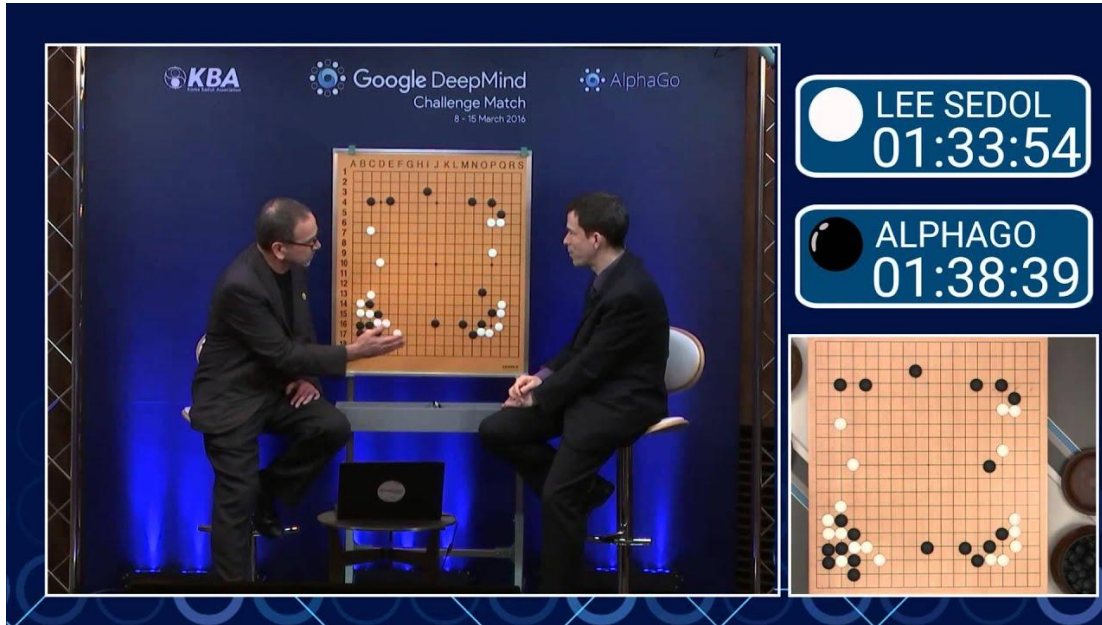
Part 3:

Why learn about reinforcement learning?

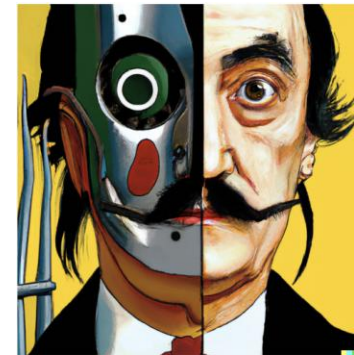
Reinforcement learning can discover new solutions

Impressive because no person had thought of it!

Impressive because it looks like something a person might draw!



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



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

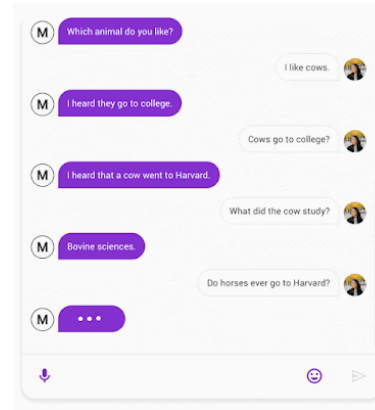


panda mad scientist mixing sparkling chemicals, artstation



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

How does data-driven AI work?



Explaining a joke

Prompt

Explain this joke:

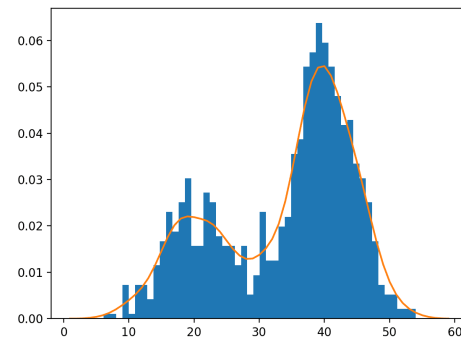
Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

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



+ learns about the real world from data

- doesn't try to do **better** than the data

Reinforcement Learning



+ 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

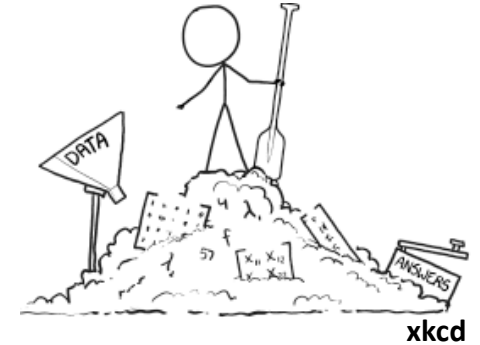
A bitter but misunderstood lesson



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*



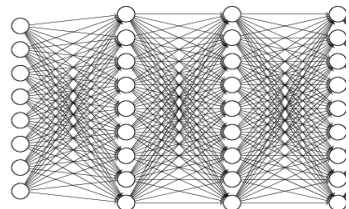
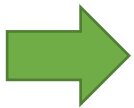
Data without optimization doesn't allow us to solve new problems in new ways

Optimization without data is hard to apply to the real world outside of simulators



Learning

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



allows us to **understand** the world

Search

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

optimization

some optimization process that uses (typically iterative) computation to make rational decisions

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

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?

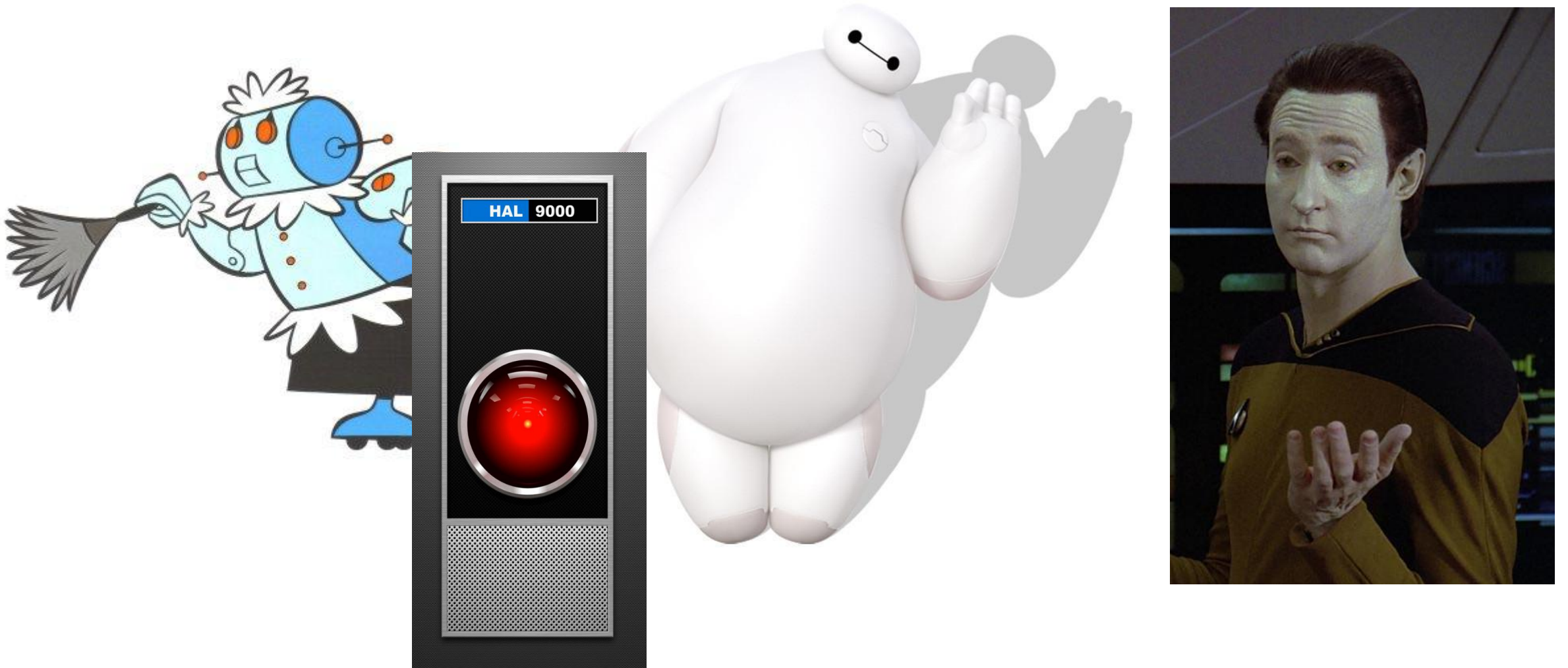


Daniel Wolpert
(knows quite a lot
about brains)



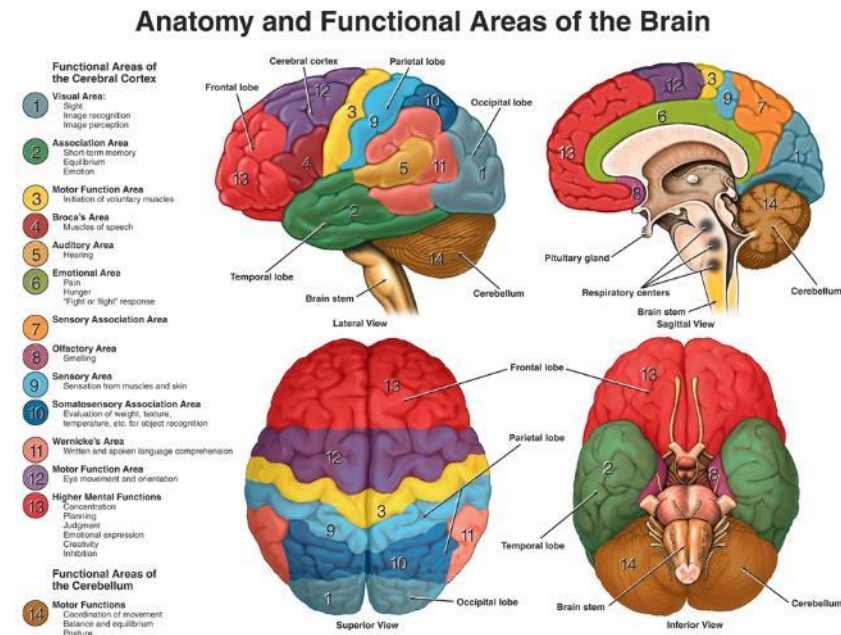
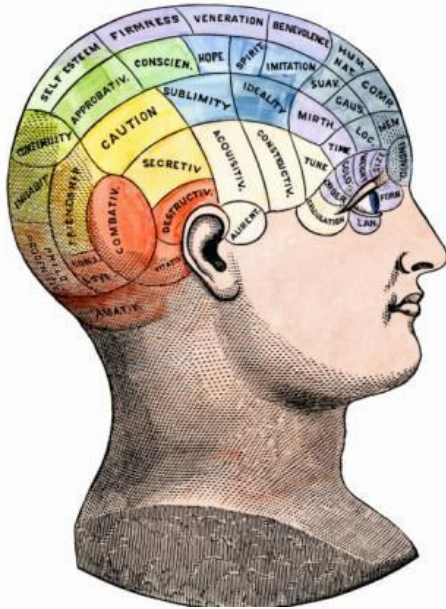
“We have a brain for one reason and one reason only – that's **to produce adaptable and complex movements.** Movement is the only way we have affecting the world around us... I believe that to understand movement is to understand the whole brain.”

How do we build intelligent machines?



How do we build intelligent machines?

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



Learning as the basis of intelligence

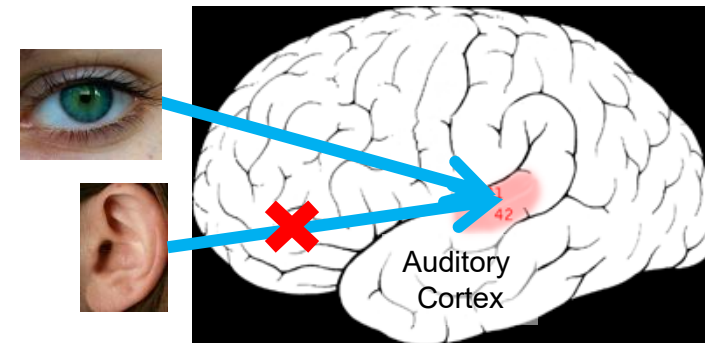
- Some things we can all do (e.g. walking)
- Some things we can only learn (e.g. driving a car)
- We can learn a huge variety of things, including very difficult things
- Therefore our learning mechanism(s) are likely powerful enough to do everything we associate with intelligence
 - But it may still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

A single algorithm?

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



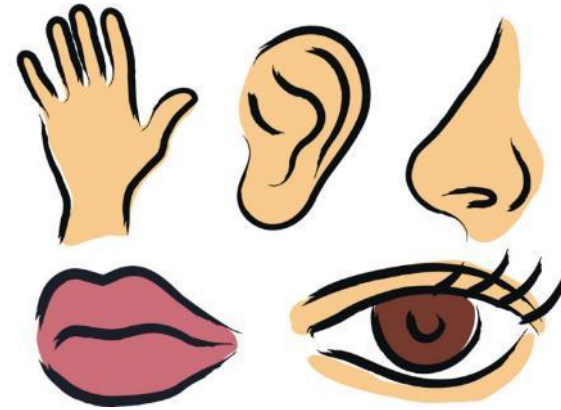
Seeing with your tongue



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

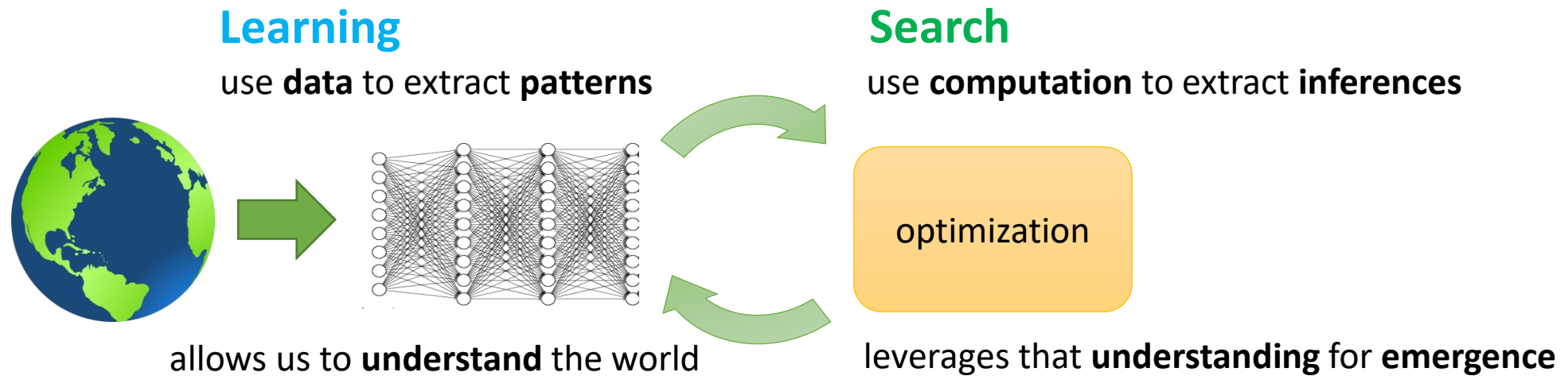
What must that single algorithm do?

- Interpret rich sensory inputs
- Choose complex actions



Why deep reinforcement learning?

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



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

