Reinforcement Learning & Autonomous systems

MBD Sept 24

UNIVERSITY

SCHOOL OF SCIENCE & TECHNOLOGY

Lecture 1
Introduction and Objectives

jmanero@faculty.ie.edu

Jaume Manero

Short Bio

- PhD in Artificial Intelligence Universitat Politècnica de Catalunya (UPC)
- Adjunct Faculty Dalhousie University (Canada)
- Researcher at Knowledge Management and Machine Learning Group (UPC)
- I research in AI application to the renewable industry
- I am Intelligent Enterprise Director at T4S with clients mainly in Europe
- I teach at IE University in different Master Programs (MBA, and technical programs)

UPC

DALHOUSIE UNIVERSITY
FACULTY OF COMPUTER SCIENCE

ie UNIVERSITY
Driving Innovation

Lecture 1

Contents

- Preamble
- Objectives and Evaluation
- Course Contents
- Introduction to RL
- Some RL Applications

Preamble

JiDU
集度汽车

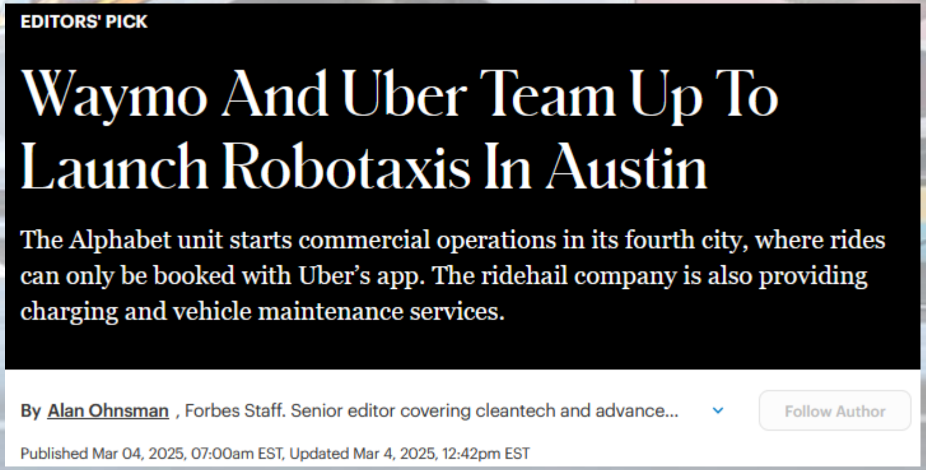
Somewhere in China – 2024

Jidu Auto (Geely+Baidu)

Baidu 百度

吉利汽车
GEELY AUTO

Are AV taking off?



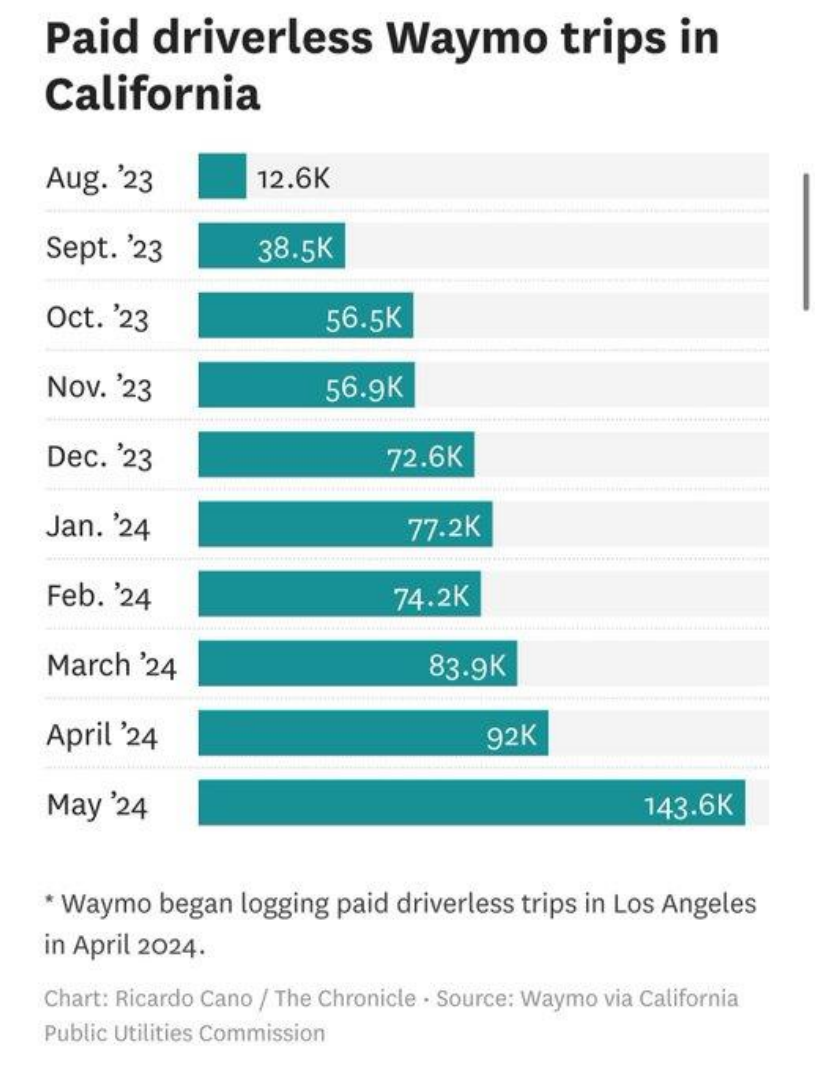
EDITORS' PICK

Waymo And Uber Team Up To Launch Robotaxis In Austin

The Alphabet unit starts commercial operations in its fourth city, where rides can only be booked with Uber's app. The ridehail company is also providing charging and vehicle maintenance services.

By Alan Ohnsman , Forbes Staff. Senior editor covering cleantech and advance...

Published Mar 04, 2025, 07:00am EST, Updated Mar 4, 2025, 12:42pm EST

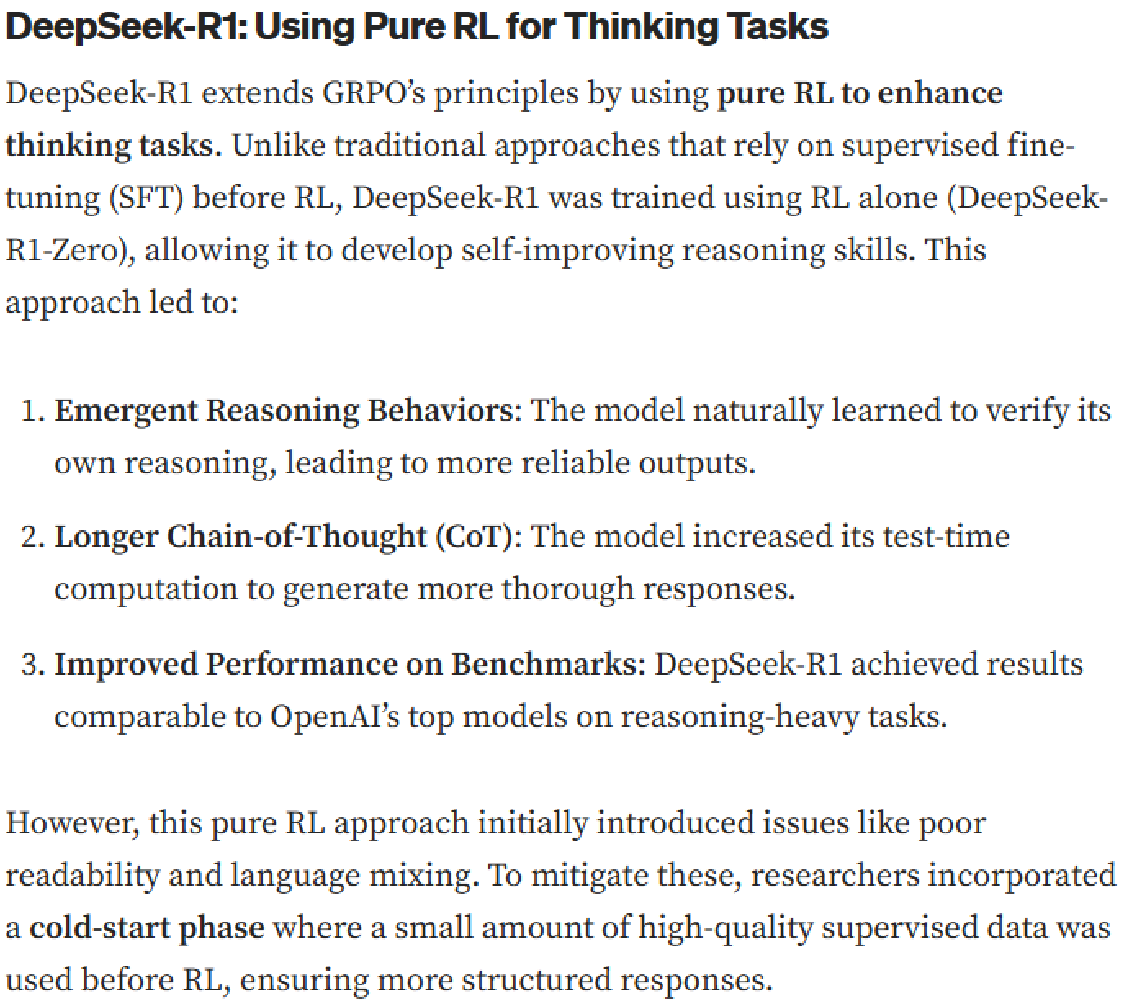


Paid driverless Waymo trips in California

* Waymo began logging paid driverless trips in Los Angeles in April 2024.

Chart: Ricardo Cano / The Chronicle · Source: Waymo via California Public Utilities Commission

Use of RL in LLM



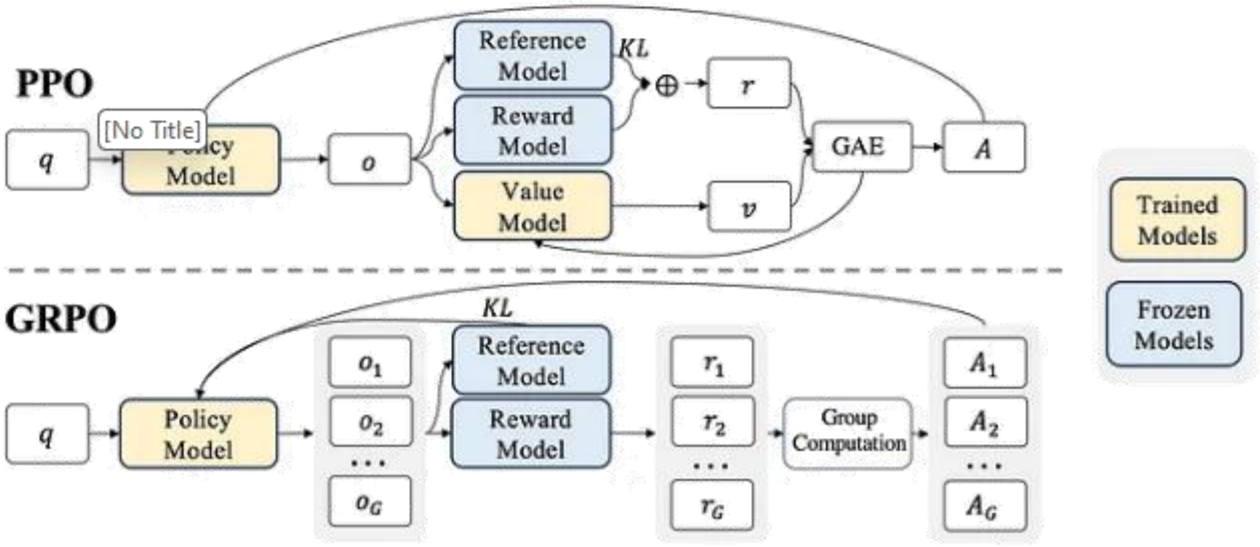
DeepSeek-R1: Using Pure RL for Thinking Tasks

DeepSeek-R1 extends GRPO's principles by using **pure RL to enhance thinking tasks.** Unlike traditional approaches that rely on supervised fine-tuning (SFT) before RL, DeepSeek-R1 was trained using RL alone (DeepSeek-R1-Zero), allowing it to develop self-improving reasoning skills. This approach led to:

1. **Emergent Reasoning Behaviors:** The model naturally learned to verify its own reasoning, leading to more reliable outputs.
2. **Longer Chain-of-Thought (CoT):** The model increased its test-time computation to generate more thorough responses.
3. **Improved Performance on Benchmarks:** DeepSeek-R1 achieved results comparable to OpenAI's top models on reasoning-heavy tasks.

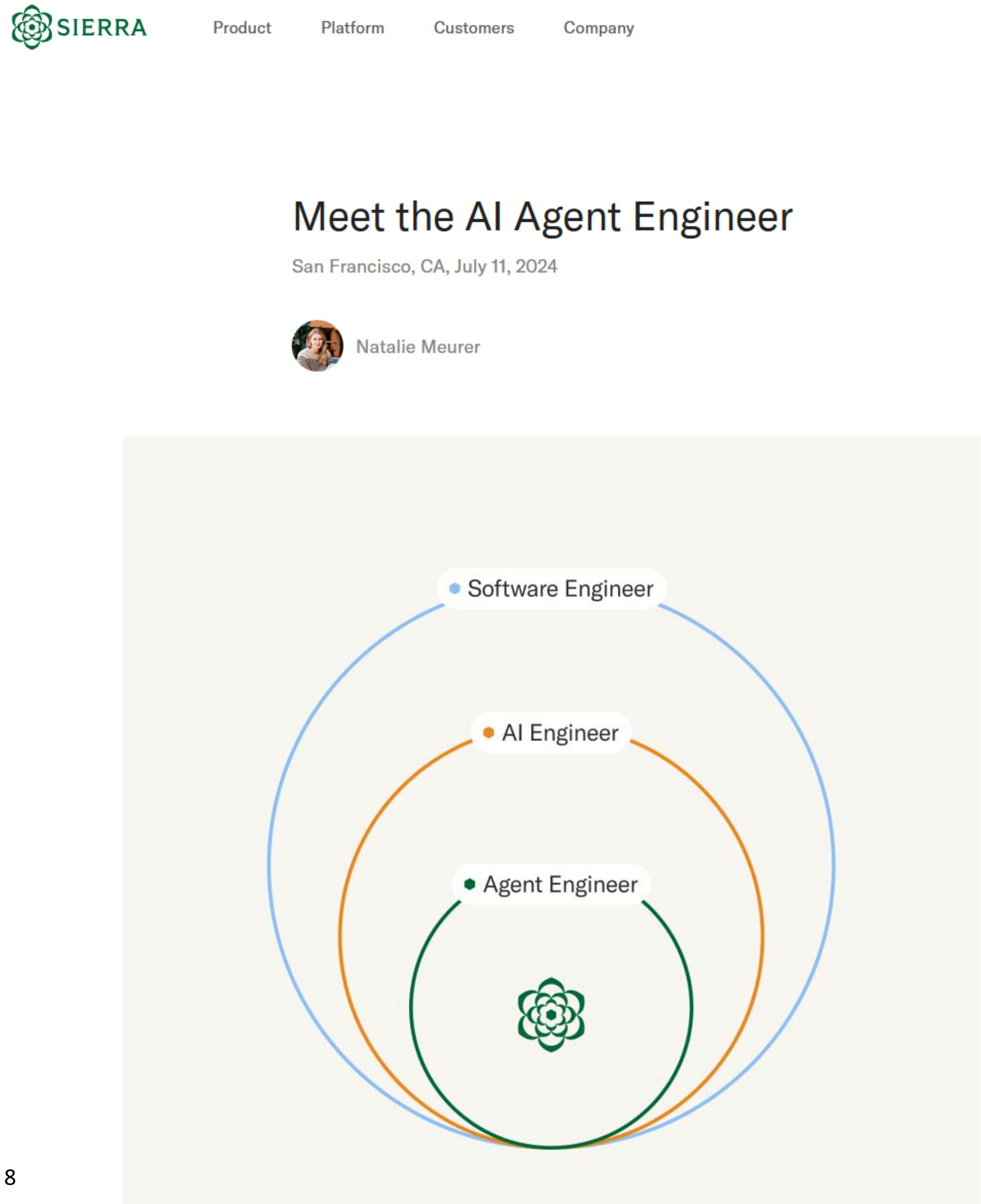
However, this pure RL approach initially introduced issues like poor readability and language mixing. To mitigate these, researchers incorporated a **cold-start phase** where a small amount of high-quality supervised data was used before RL, ensuring more structured responses.

deepseek



Comparaison between PPO and GRPO in the Deepseek math paper

The next big thing?



SIERRA Product Platform Customers Company

Meet the AI Agent Engineer

San Francisco, CA, July 11, 2024

Natalie Meurer

- Sierra, a new startup closely related to OpenAI is working in agents.
- Agents are autonomous programs that can work independently with objectives that guide their evolution
- ¿Can they become the next killer application of generative AI after ChatGPT?

Sierra sees a new kind of engineer which is the Agent Engineer, who develop agents working closely with customers and understand how to ALIGN the agent capabilities with the customer expectation to be allowed to roam freely out there. Read the article and have a look at the Sierra web site.

https://sierra.ai/blog/meet-the-ai-agent-engineer?utm_source=tldrai

Objectives and Evaluation

Objectives

- Understand the basics of Reinforcement Learning and its importance in the Artificial Intelligence discipline
- Review the most important algorithms used in Reinforcement Learning, the concept of reward, policy and optimization
- From Dynamic Programming to Deep Reinforcement Learning
- Perform some practical LABS to obtain hands-on experience in RL Challenges
- Learn some of the real use cases and applications for RL today
- Have a view on where the research for RL is heading to in the future

Practices / Labs

- Mac / Windows / Linux work
- You need to have installed a Python environment on your PC 3.x If you don't have it installed use Anaconda as is the easiest to install
- You need to have some understanding of Deep Learning, specifically with KERAS
- You can execute the practices in Google Colab. Some exercises will benefit from this as they may challenge your laptop capacity
- You'll receive detailed instructions for each practice. Don't worry, and all make available to you resources to jump start in this field

I am more familiar with Linux and Windows; I am not a user expert with MAC. My environment is based in windows but I use a mixture of Linux and Windows using the WSL. If you are interested have a look at
https://ubuntu.com/blog/wsl-for-data-scientist
However, as I said the easiest way to have a Python environment in your laptop is Anaconda

Evaluation

- Quiz 1 – After Class 7 (30%)
- Quiz 2 – Final Class (15%) – From class 8 to 14
- Group Practice (30%):
  - Group Exercise
  - Presentation in Class (Class 13-14)
- Assignments (15%)
  - 4 Assignments in Labs+home
  - Python
- Participation (10%)
  - Participation in class, in Forums

Course Contents and Structure

Introduction to Reinforcement Learning

15 Sessions, one Group Project and 5 Assignments

Session	Summary	Examples/Exercises
Lecture 1 Introduction and Objectives	• Course description, objectives and evaluation • Tools and Evaluation • Bibliography • What is RL? • Why RL is relevant in the Actual AI revolution? • RL Real Applications	
Lecture 2 Basic Concepts of Reinforcement Learning	• The Taxonomy of RL Approaches • Basic Concepts • Model • Policy • Reward Functions • Short Term and Long Term • Dynamic Programming • Value Iteration • Policy iteration • https://towardsdatascience.com/reinforcement-learning-with-openai-d445c2c687d2	
Lecture 3 Dynamic Programming	Dynamic Programming • Value Iteration • Policy Iteration	
Lecture 4 LAB 1 Using GYMNASIUM	• Setting up Python environment • Introducing Gymnasium • See the different environments (Taxi, Cart Pole, Russells Grid) • Understanding the API • Showing the results	Assignment 1 Russells Grid & introduction to GYMNASIUM

Introduction to Reinforcement Learning

15 Sessions, one Group Project and 5 Assignments

	Summary	Examples/Exercises
Lecture 5 **Model Free** Monte-Carlo, TD	• MonteCarlo Methods • Temporal Difference Methods TD(0) • A quick review of the TD Backgammon pioneer example	Frozen Lake MC Frozen Lake TD(0)
Lecture 6 LAB 2 Group Practice	• Dynamic Programming • Frozen Lake MC example • Presentation of Group Practice • Assignment 2	Assignment 2 Frozen Lake Taxi
Lecture 7 SARSA	• SARSA Strategies / Q-Learning	Frozen Lake Taxi Cartpole

15 Sessions, one Group Project and 5 Assignments

Sessiones	Summary	Challenges / Exercises
Lecture 8 LAB3 Q-Learning	• **Quiz 1** • Assignment 3 – Q-Learning in CARTPOLE	Assignment 3 Model-Free Discrete SARSA and Q-learning
Lecture 9 Deep Learning review	• Review of Neural Networks using KERAS • MLP / CNN Networks • Properties, issues, capabilities of Neural Networks • Using KERAS in local and COLAB environments • GPU's and other nuances	
Lecture 10 Non-Linear function approximation DQN	• What happens is the learning function is non-linear? • Using Deep Learning to approximate functions • Using Agents learning with deep learning • The Structure of DQN	
Lecture 11 LAB 4 DQN	• Applying DQN to CARTPOLE • The problem of convergence • When does it start learning?	Assignment 5 Applying DQN to Cartpole

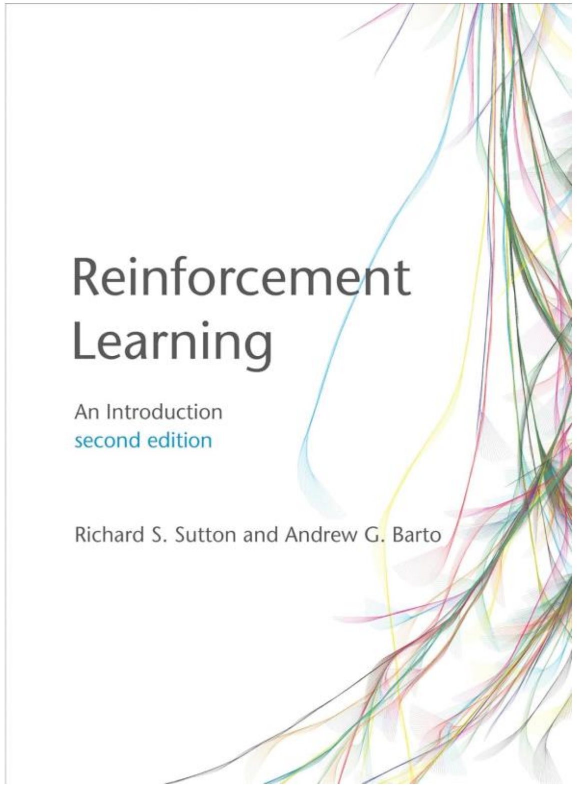
15 Sessions, one Group Project and 5 Assignments

Session	Summary	Examples/Exercises
Lecture 12 Deep RL	• After DQN – DQN sophistications • DDQN • Actor Critic, Gradient Policy, Duelling	DDQN Soft_update Duelling Networks
Lecture 13/14 **Group Practice** **Presentations**	• Group Practice Presentations	
Lecture 15 Wrap-up	**Quiz 2** **What is Next?** The use of Reinforcement Learning in Generative AI The problem of Alignment Why RL introduces biases in Gen AI?	

4 Activities/Assignments and 1 Group Project

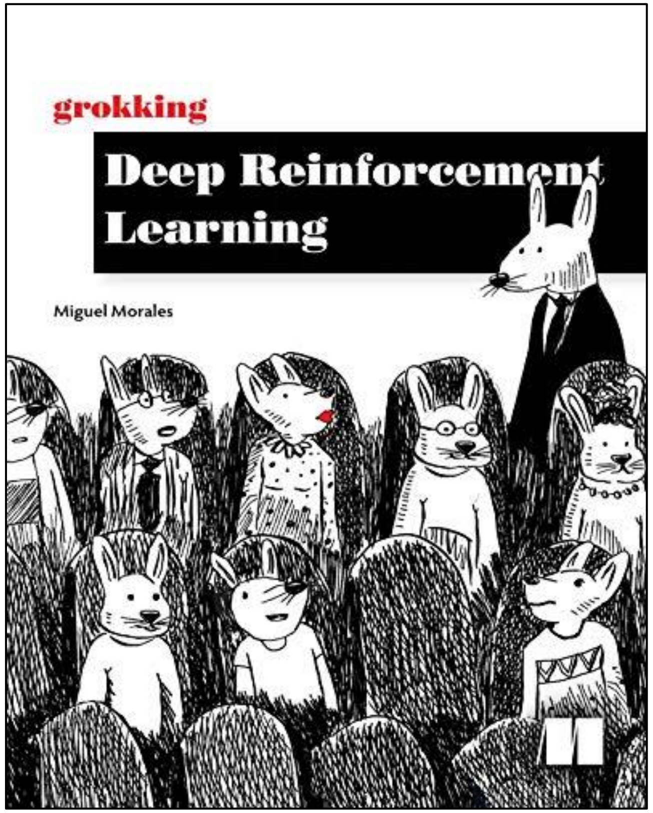
LABS	Activity	Name
Lab1	1. Generating New environments with Gymnasium	Environments, Rewards and Policies
Lab2	2. Dynamic Programming Value and Policy Iterations	Dynamic Programming, MC
Lab3	3. Model Free – Monte Carlo algorithm	Q-Learning
Lab4	4. Model Free – Q-learning and Sarsa	DQN

Support Bibliography



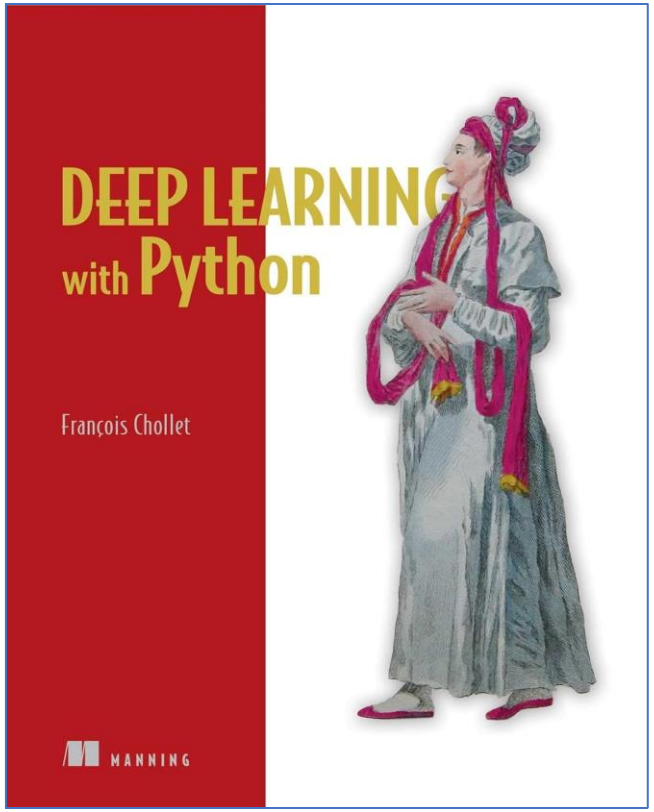
The Sutton-Barto

This is the Main book of Reinforcement Learning we'll use it as reference book. can be used for an introductory course



Grokking-Morales

Good book as it has exercises. Focused on Deep Reinforcement Learning (pytorch)



Chollet ed.2

Keras focused Deep Learning. Some RL apps there

Motivation

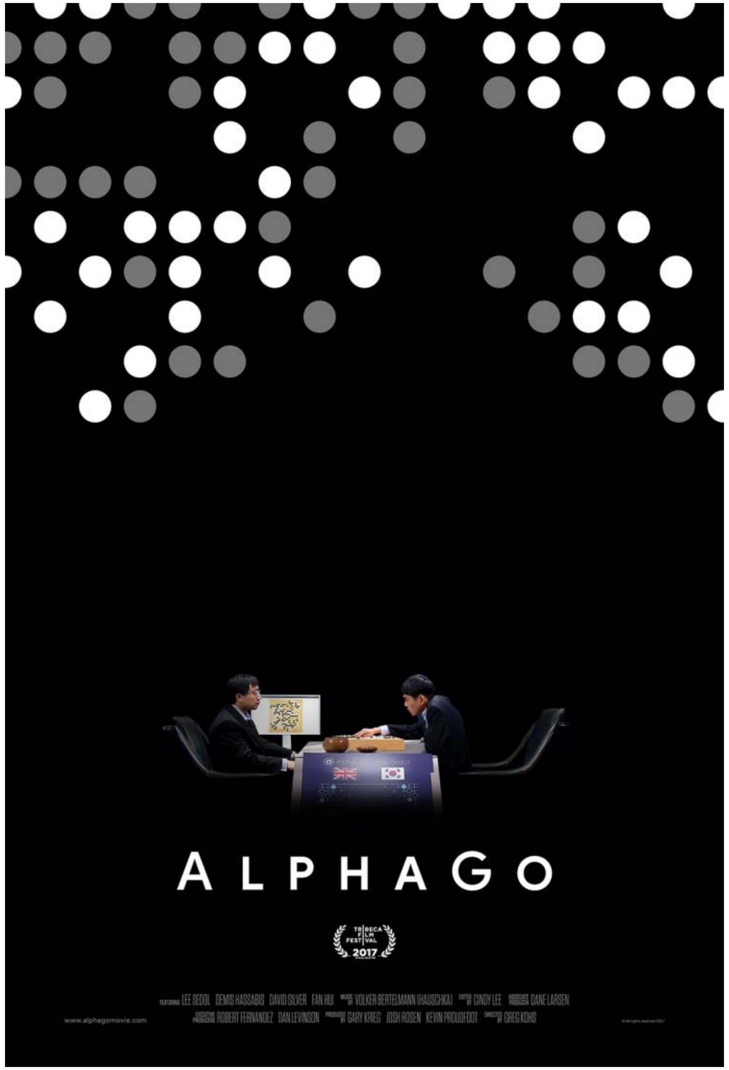
- The idea of programs (agents) that roam freely and perform activities it really captures our imagination (See “I Robot” from Asimov and its 3 laws of Robotics)
- The work in this kind of programs (agents that learn and perform actions in a universe) is not new and Dr. Sutton has been working on this since the ‘80s.
- However, the explosion of, ML first and then DL has given new tools to this discipline, tools that have allowed to accomplish impossible feats like the Autonomous Vehicle or the shocking defeat of the world champion of GO in the hands of AlphaGo, an agent from Deepmind in 2017
- There are many applications waiting for RL and the irruption of Generative AI is the last push that may revolutionize this area

Lee Sedol was the world champion of GO, a game tknown to be impossible to solve using ‘brute force’ algorithms. However, using a Deep Learning approach the Deepmind team managed to put together an impressive artificial intelligence player.

To show the power of its algorithm, Lee Sedol naively thinking he was unbeatable by a machine, accepted the Challenge. In this challenge DeepMind won 4 games and Lee Sedol won 1. Lee Sedol retired a couple of years after, while Deepmind has found new and amazing applications for their RL algorithms.

Reinforcement Learning

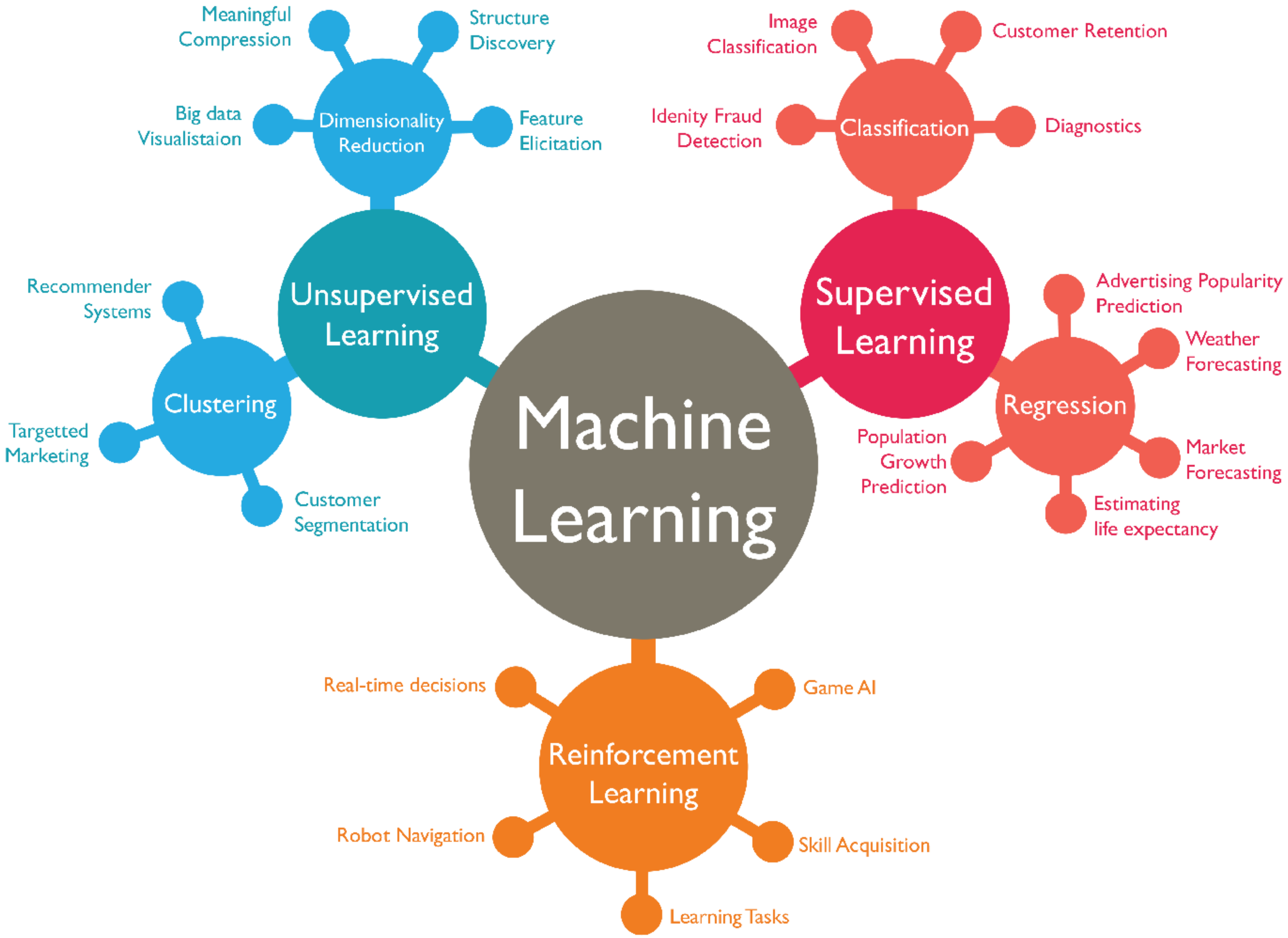
A must watch movie


ALPHAGO
TRIBECA FILM FESTIVAL 2017
FEATURING LEE SEDOL DEMIS HASSABIS DAVID SILVER FAN HUI MUSIC BY VOLKER BERTELMANN (HAUSCHKA) EDITED BY CINDY LEE ASSOCIATE PRODUCER DANE LARSEN
EXECUTIVE PRODUCERS ROBERT FERNANDEZ DAN LEVINSON PRODUCED BY GARY KRIEG JOSH ROSEN KEVIN PROUDFOOT DIRECTED BY GREG KOHS
www.alphagomovie.com

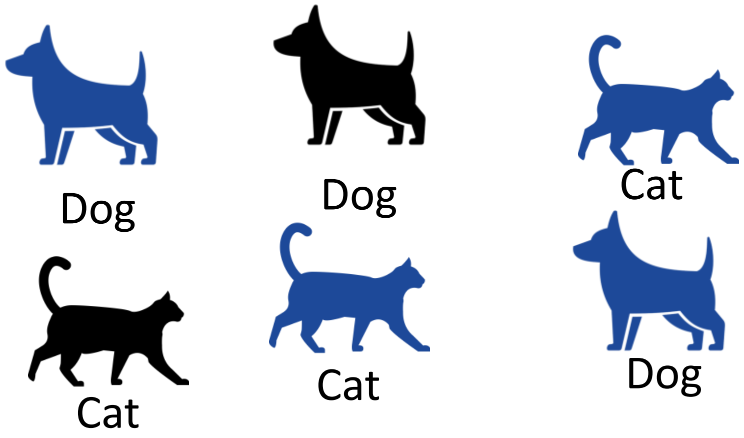
https://www.youtube.com/watch?v=WXuK6gekU1Y

Quick introduction to Reinforcement Learning

3 types of Learning

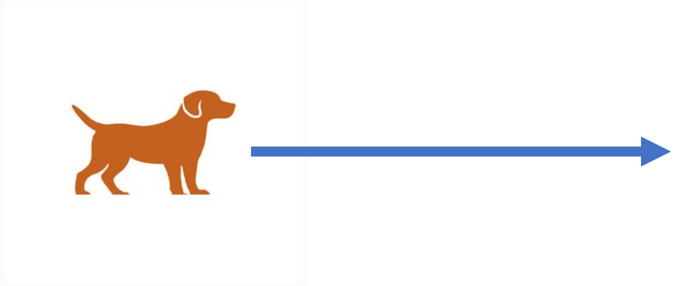


Supervised Learning



"Learn the Features of each image, Cat or Dog"

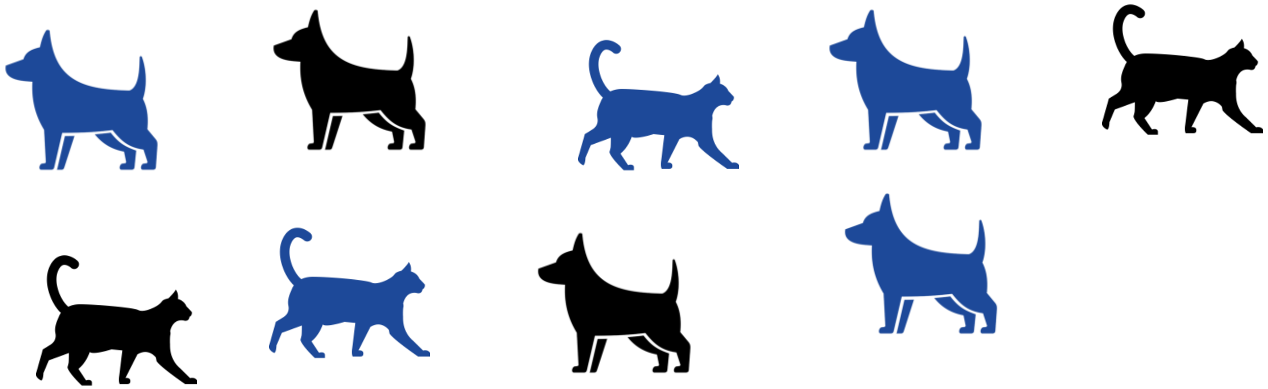
"Having learned the features, tell
me if this image is a Cat or a Dog"



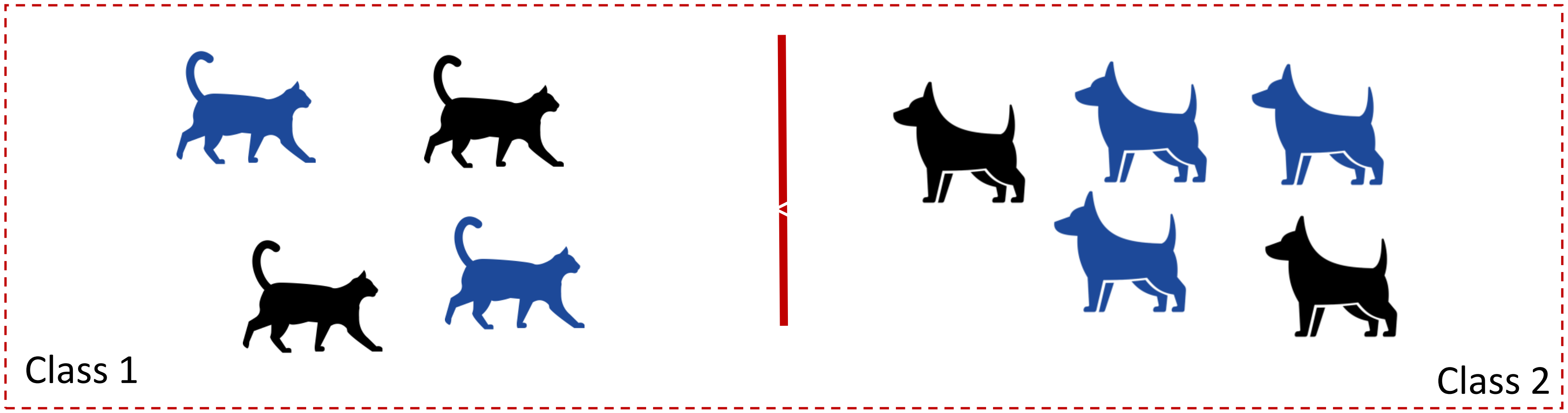
"This is a Dog"

GOAL: Learn from the labels in the training data and then perform pattern recognition to classify unseen data

Unsupervised Learning



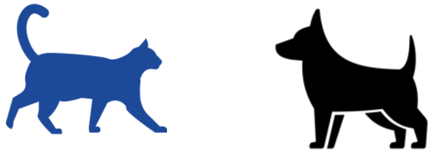
"Classify the data in groups of animals"



GOAL: Find underlying patterns in data structure that help to understand complex data

Reinforcement Learning

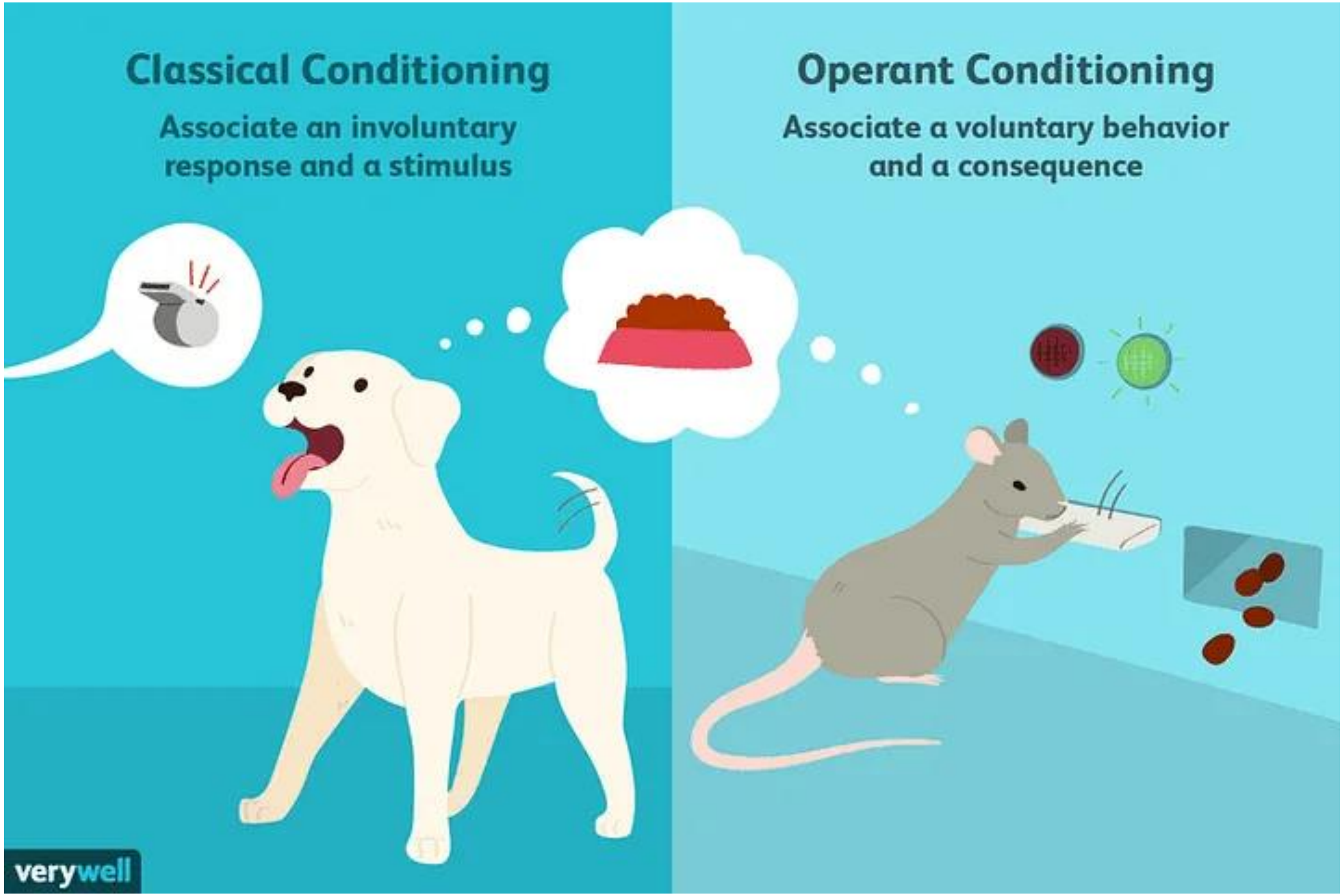
Reinforcement Learning



“After interacting with both of them I’ve learned that dogs are dangerous and bigger than cats”

GOAL: Maximize rewards in a universe over many time steps to achieve an objective (for instance to stay alive in a game). Through the process the agent will learn about the universe around and how it impacts (Rewards) in the actions he must perform to achieve the objective

The basis for reward and action



Classical Conditioning
Associate an involuntary response and a stimulus
Operant Conditioning
Associate a voluntary behavior and a consequence
verywell

Positive or negative reward

NEGATIVE
REINFORCEMENT
UNWANTED
STIMULUS
REMOVED BY
BEHAVIOR
POSITIVE
REINFORCEMENT
REWARDING
STIMULUS
PRESENTED BY
BEHAVIOR
Great Work!
Bart Simpson
Grade 2

https://towardsdatascience.com/reinforcement-learning-with-openai-d445c2c687d2

Reinforcement Learning
Key Concepts: Agent

Agent

AGENT DEFINITION

An Agent is an autonomous computer program

It takes structured actions that can be defined

Key Concepts: Environment

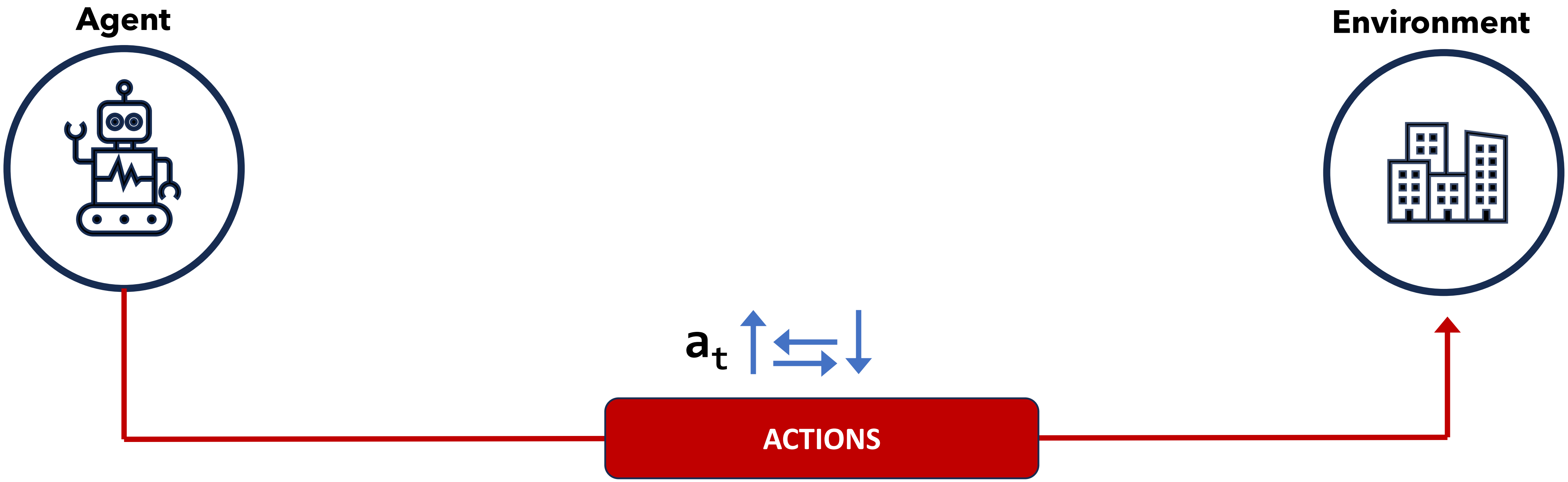
ENVIRONMENT DEFINITION

Is the world where the Agent exists and operates or interacts

Environment

Reinforcement Learning

Key Concepts: Actions



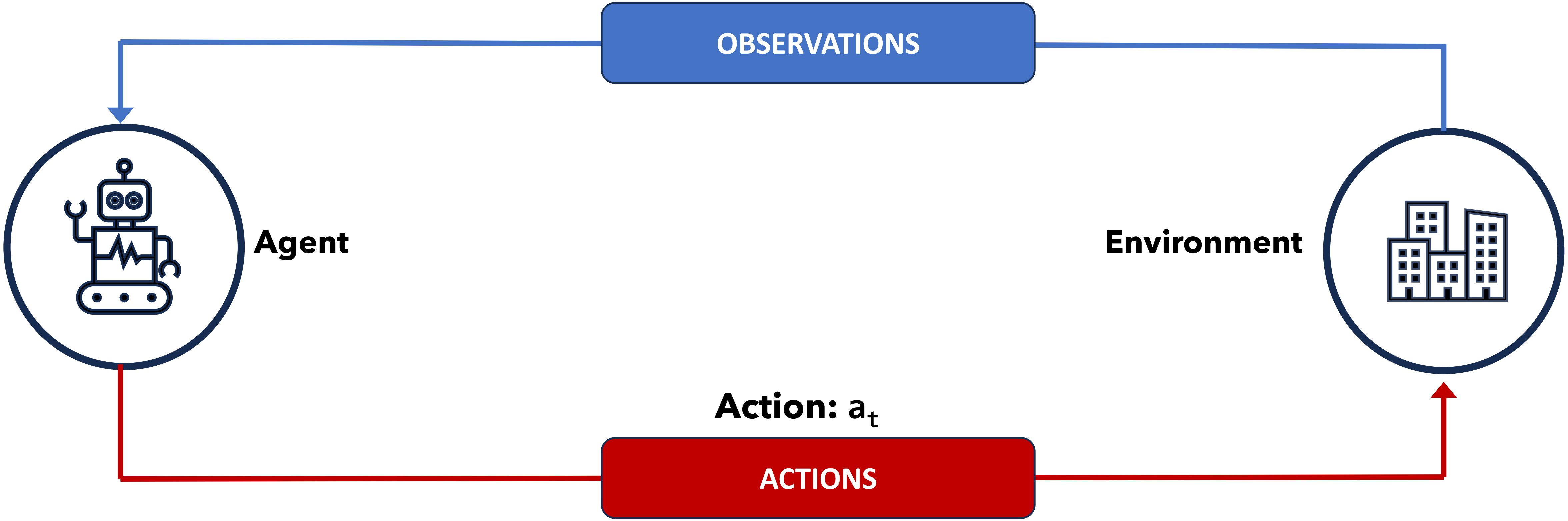
ACTION DEFINITION

A move the agent can make in the Environment

Action Space ***A***: The set of possible actions an agent can perform in the environment

$$A = \{a_1, a_2, \ldots, a_n\}$$

Key Concepts: Observations



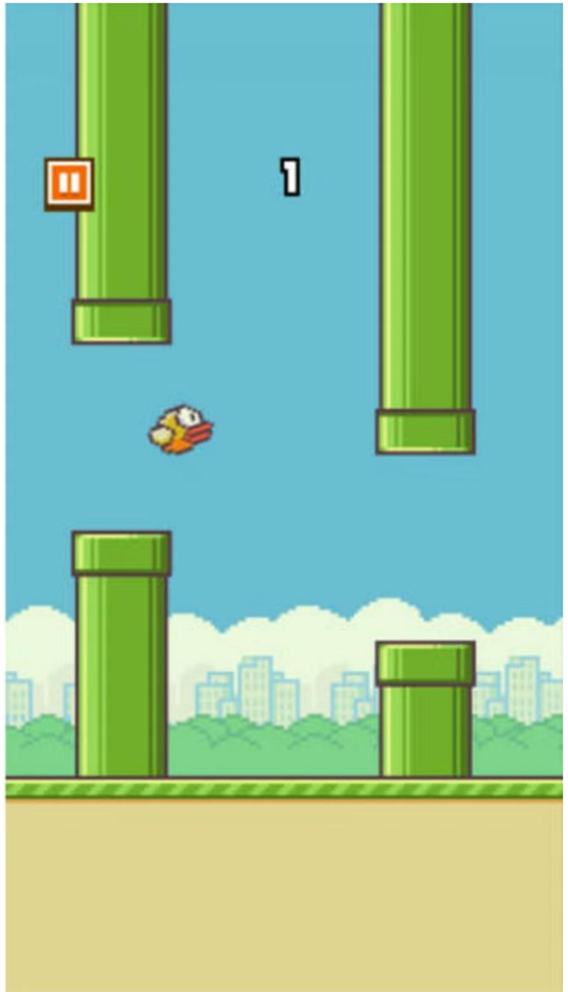
OBSERVATION DEFINITION

Understand the environment after taking actions

(What has changed from last observation, what is new, …)

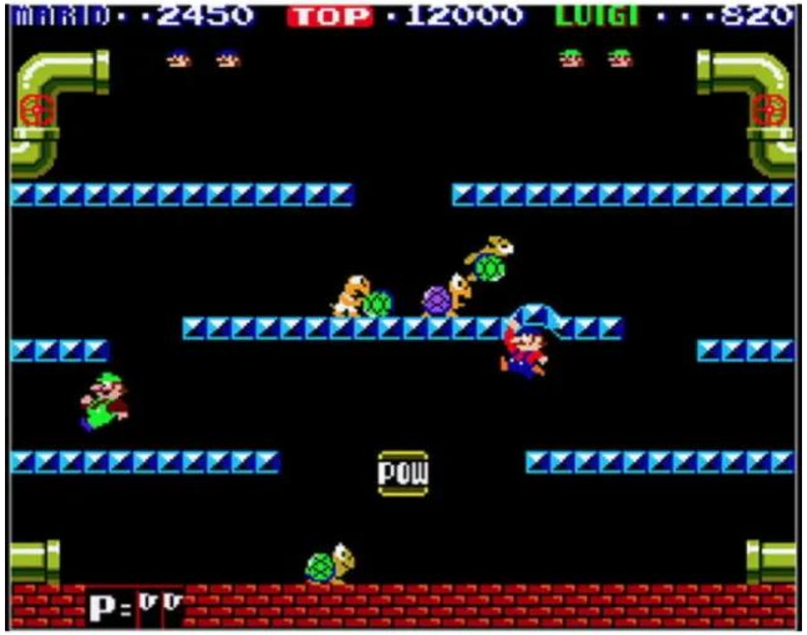
Some Reinforcement Learning applications

Reinforcement Learning
Examples



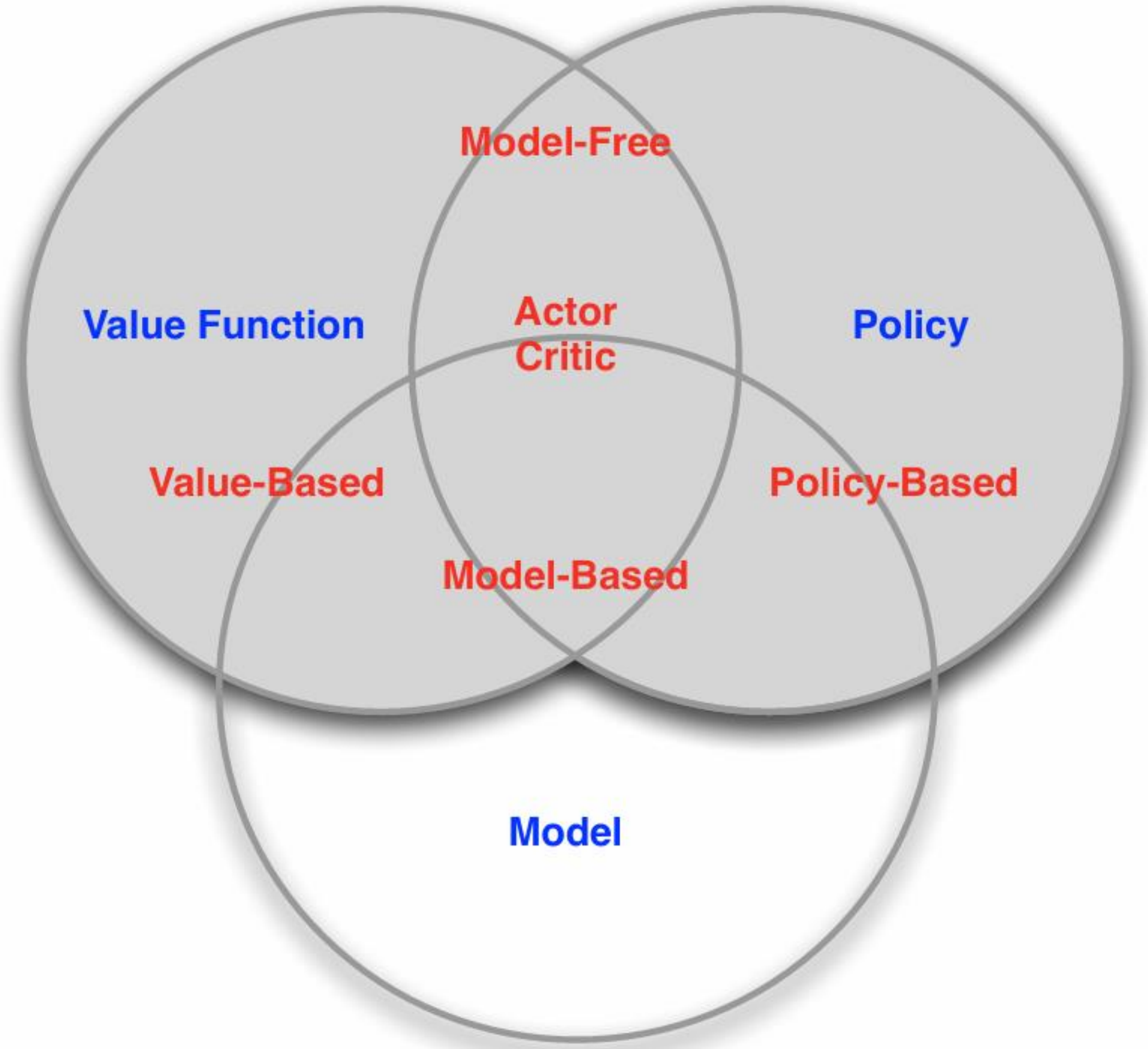
Flappy Bird

Classic Mario Bros



Who is the agent, the universe and the Action Space in these two games?

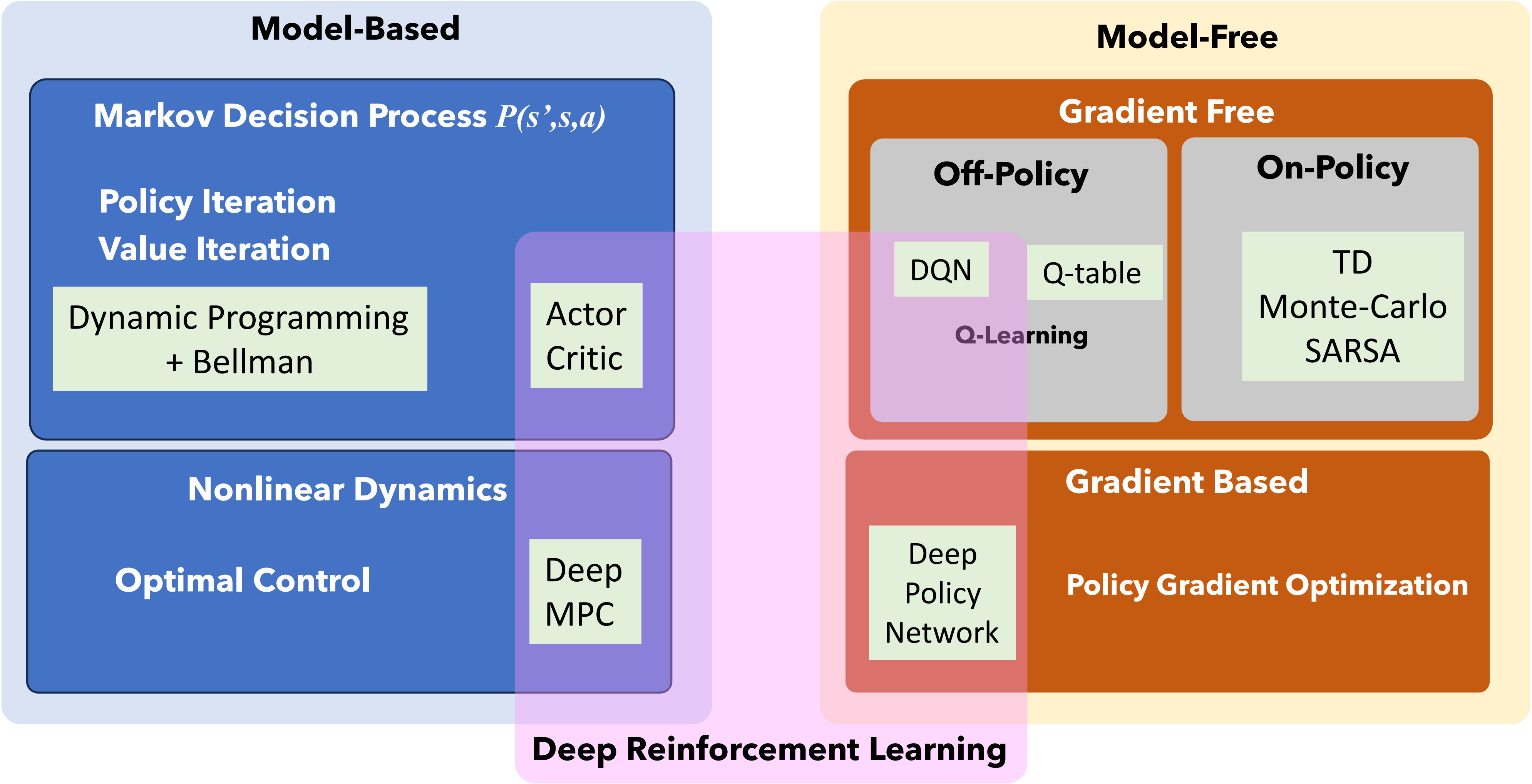
Taxonomy of RL Agents



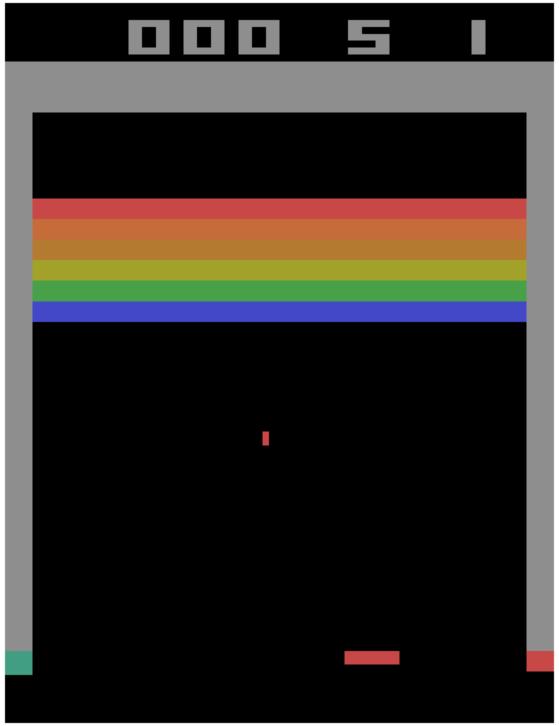
Taxonomy of RL Agents

(*) From David Silver RL Course UCL

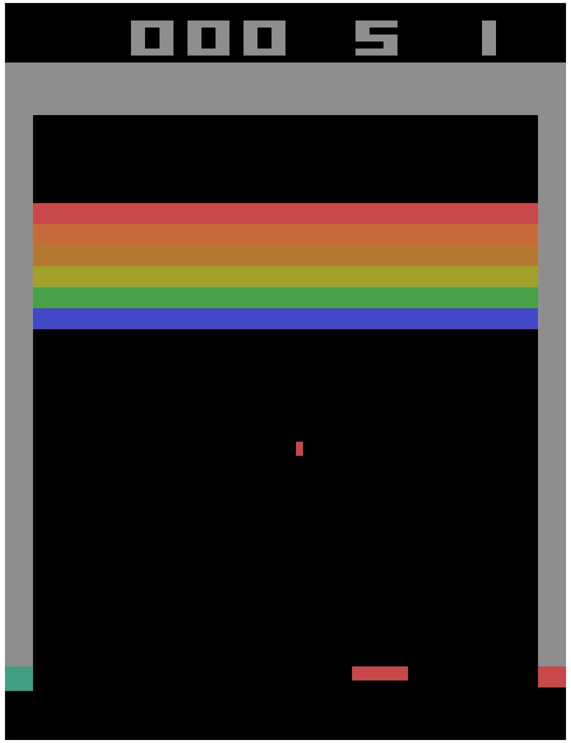
A Taxonomy of RL Methods
Classical classification by David Silver



The Atari Paper from David Silver



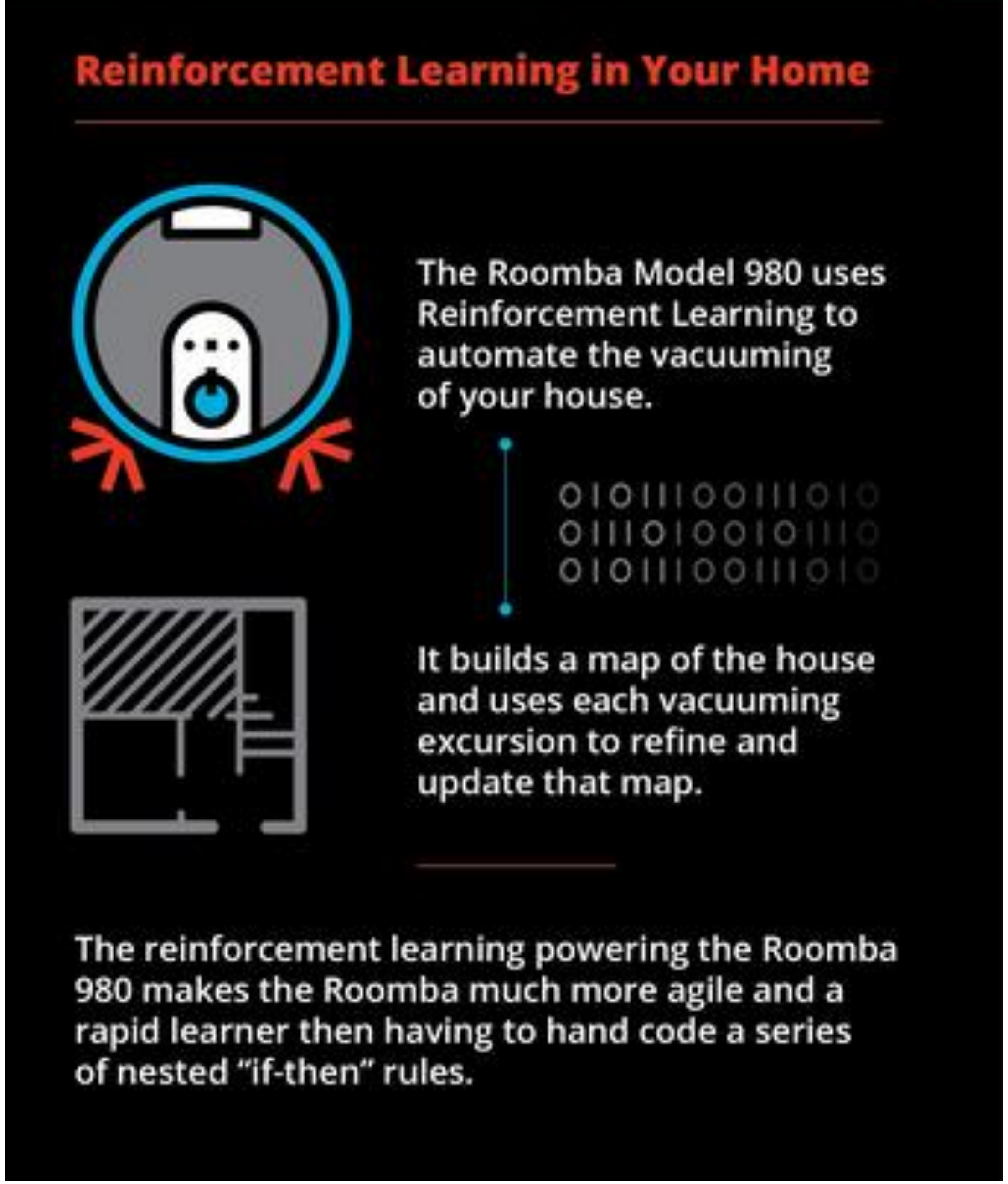
After 1,000 episodes



After 11,000 episodes

Roomba 980 RL

https://youtu.be/oj3Vawn-kRE



Reinforcement Learning in Your Home

The Roomba Model 980 uses Reinforcement Learning to automate the vacuuming of your house.

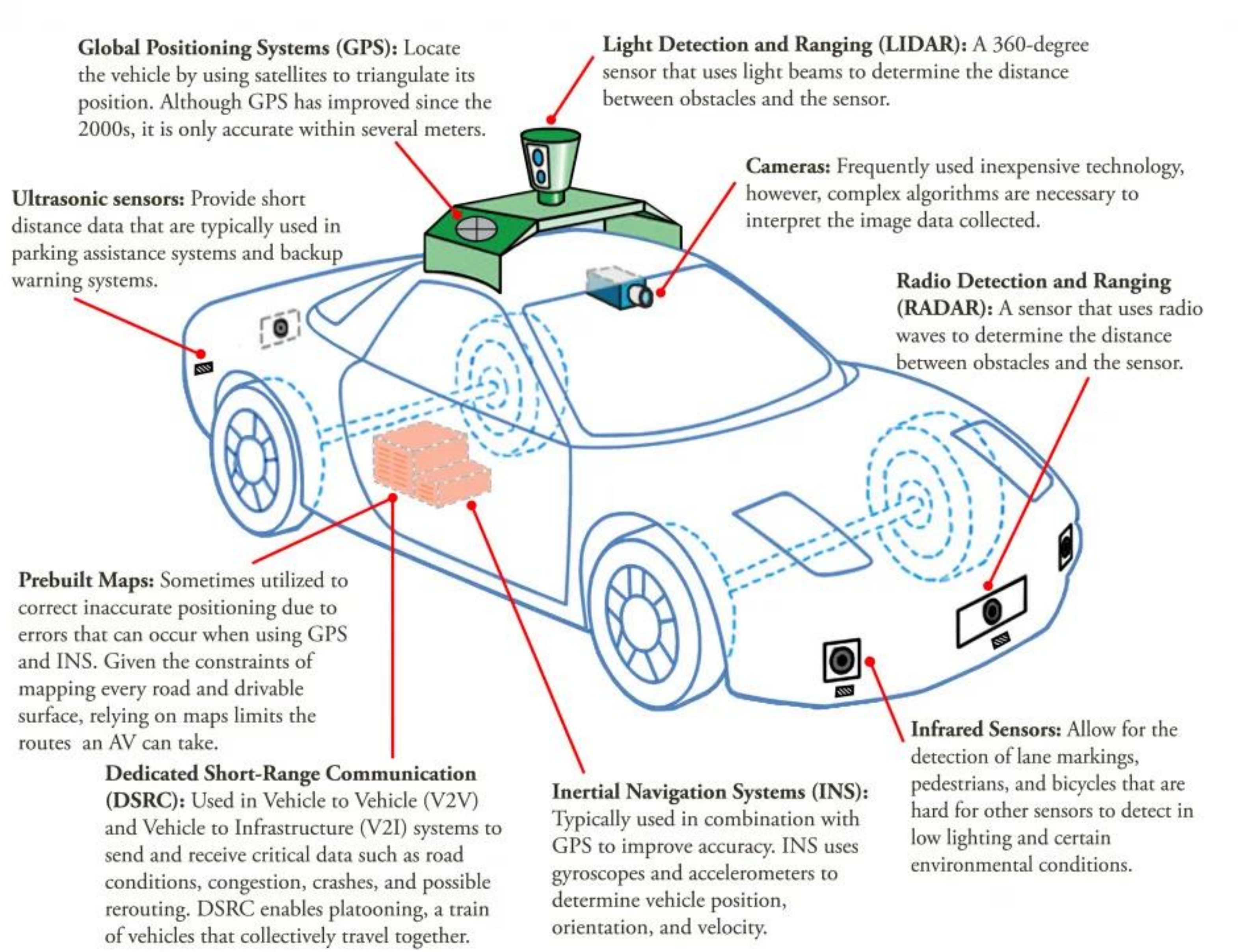
0101110011101
0111010010111
0101110011101

It builds a map of the house and uses each vacuuming excursion to refine and update that map.

The reinforcement learning powering the Roomba 980 makes the Roomba much more agile and a rapid learner then having to hand code a series of nested “if-then“ rules.

Waymo car

Autonomous Vehicle : Sensors + Reinforcement learning



Global Positioning Systems (GPS): Locate the vehicle by using satellites to triangulate its position. Although GPS has improved since the 2000s, it is only accurate within several meters.

Light Detection and Ranging (LIDAR): A 360-degree sensor that uses light beams to determine the distance between obstacles and the sensor.

Ultrasonic sensors: Provide short distance data that are typically used in parking assistance systems and backup warning systems.

Cameras: Frequently used inexpensive technology, however, complex algorithms are necessary to interpret the image data collected.

Radio Detection and Ranging (RADAR): A sensor that uses radio waves to determine the distance between obstacles and the sensor.

Prebuilt Maps: Sometimes utilized to correct inaccurate positioning due to errors that can occur when using GPS and INS. Given the constraints of mapping every road and drivable surface, relying on maps limits the routes an AV can take.

Dedicated Short-Range Communication (DSRC): Used in Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I) systems to send and receive critical data such as road conditions, congestion, crashes, and possible rerouting. DSRC enables platooning, a train of vehicles that collectively travel together.

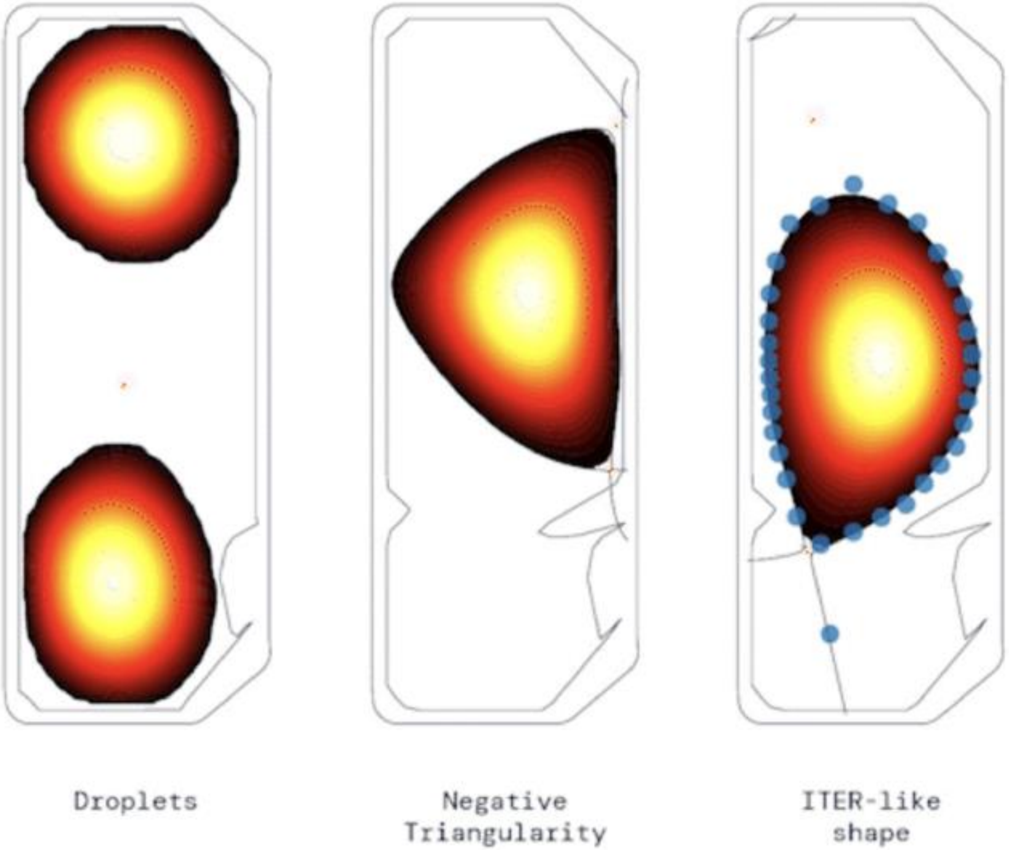
Inertial Navigation Systems (INS): Typically used in combination with GPS to improve accuracy. INS uses gyroscopes and accelerometers to determine vehicle position, orientation, and velocity.

Infrared Sensors: Allow for the detection of lane markings, pedestrians, and bicycles that are hard for other sensors to detect in low lighting and certain environmental conditions.

More Examples

- Fly stunt manoeuvres in a helicopter
- Defeat the world champion at Backgammon
- Manage an investment portfolio
- Control a power station
- Make a humanoid robot walk

Learning Plasma Control for Fusion Science



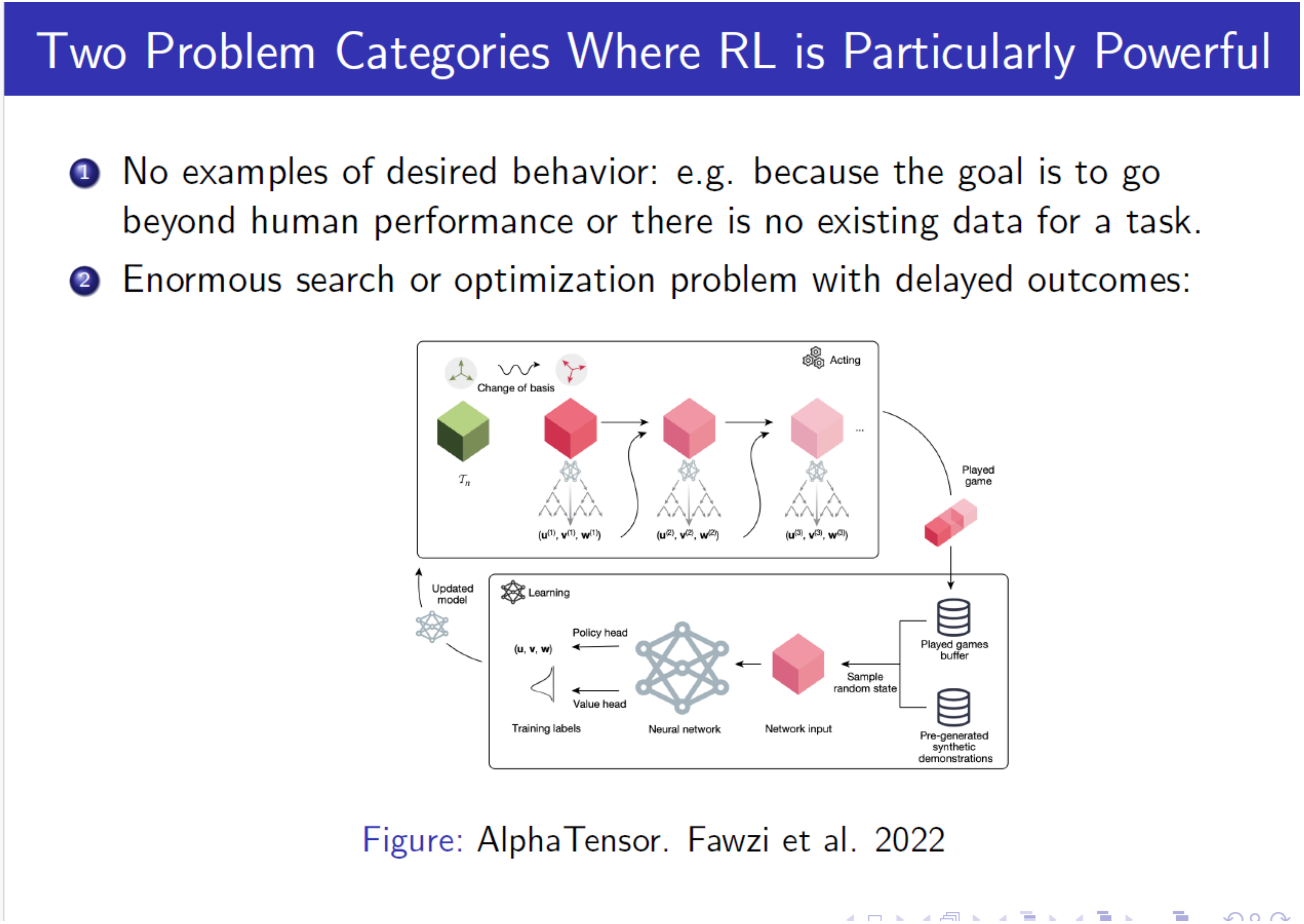
Droplets

Negative Triangularity

ITER-like shape

[2] Image credits: left Alain Herzog / EPFL, right DeepMind & SPC/EPFL. Degrave et al. Nature 2022 https://www.nature.com/articles/s41586-021-04301-9

When is RL powerful?



Two Problem Categories Where RL is Particularly Powerful

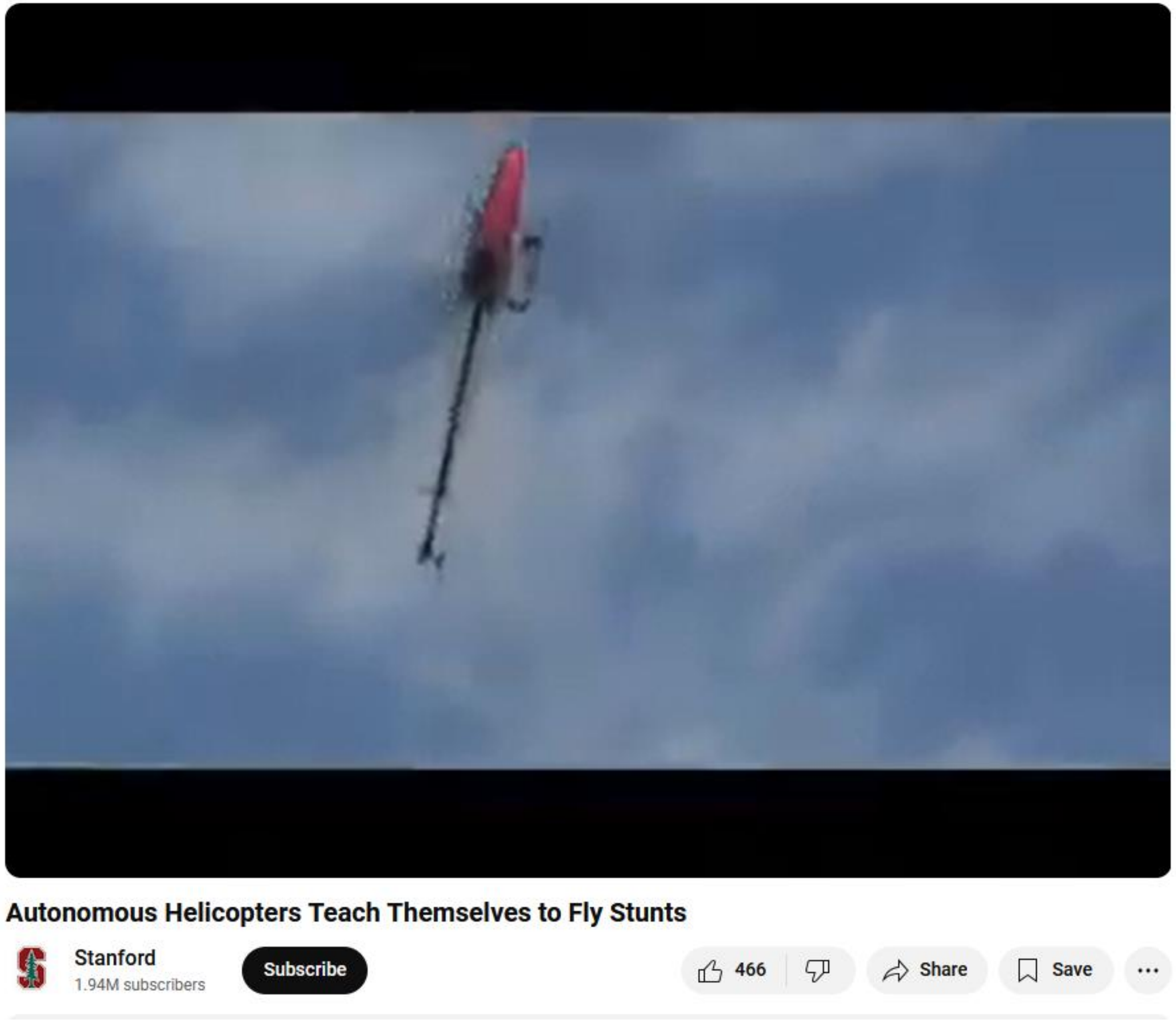
1. No examples of desired behavior: e.g. because the goal is to go beyond human performance or there is no existing data for a task.
2. Enormous search or optimization problem with delayed outcomes:

Change of basis
Acting
$\mathcal{T}_n$
...
$(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}, \mathbf{w}^{(1)})$
$(\mathbf{u}^{(2)}, \mathbf{v}^{(2)}, \mathbf{w}^{(2)})$
$(\mathbf{u}^{(3)}, \mathbf{v}^{(3)}, \mathbf{w}^{(3)})$
Played game
Updated model
Learning
Policy head
(u, v, w)
Value head
Training labels
Neural network
Network input
Sample random state
Played games buffer
Pre-generated synthetic demonstrations

Figure: AlphaTensor. Fawzi et al. 2022

Helicopter Stunts

- Andrew Ng is a well known AI researcher
- His PhD work was based on the application of RL to teaching helicopers to perfom difficult stunts

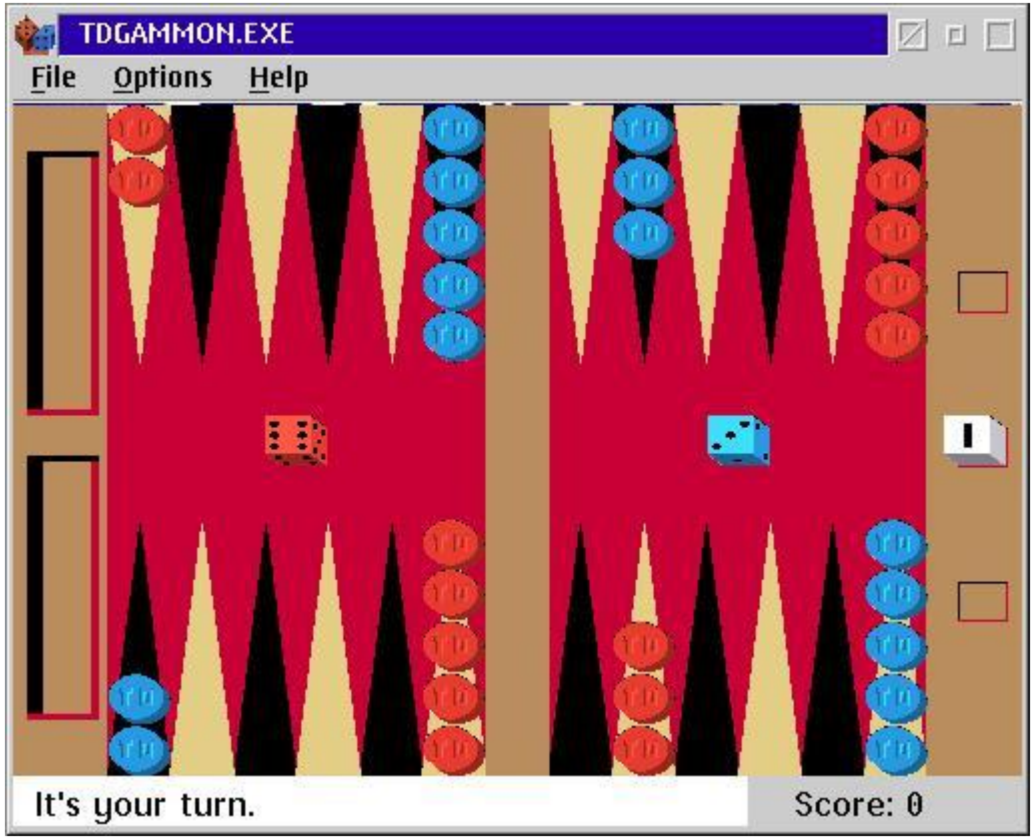


Autonomous Helicopters Teach Themselves to Fly Stunts

https://www.youtube.com/watch?v=M-QUkgk3HyE

TD-Gamon

- In 1992 Gerald Tesauro (IBM) developed a RL algorithm (using temporal difference approach) and was able to play at human level
- In 1998 was defeated by the world champion by a mere margin of 8 points



It's your turn.

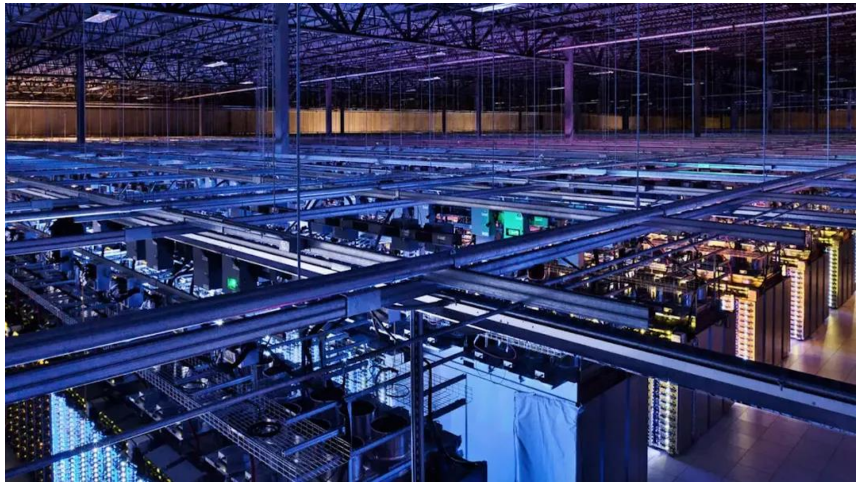
Score: 0

Industry automation – Google Data Centers

A great example is the use of AI agents by Deepmind to cool Google Data Centers. This led to a 40% reduction in **energy spending**. The centers are now fully controlled with the AI system without the need for human intervention. There is obviously still supervision from data center experts. The system works in the following way:

- Taking snapshots of data from the data centers every five minutes and feeding this to deep neural networks
- It then predicts how different combinations will affect future energy consumptions
- Identifying actions that will lead to minimal power consumption while maintaining a set standard of safety criteria
- Sending and implement these actions at the data center

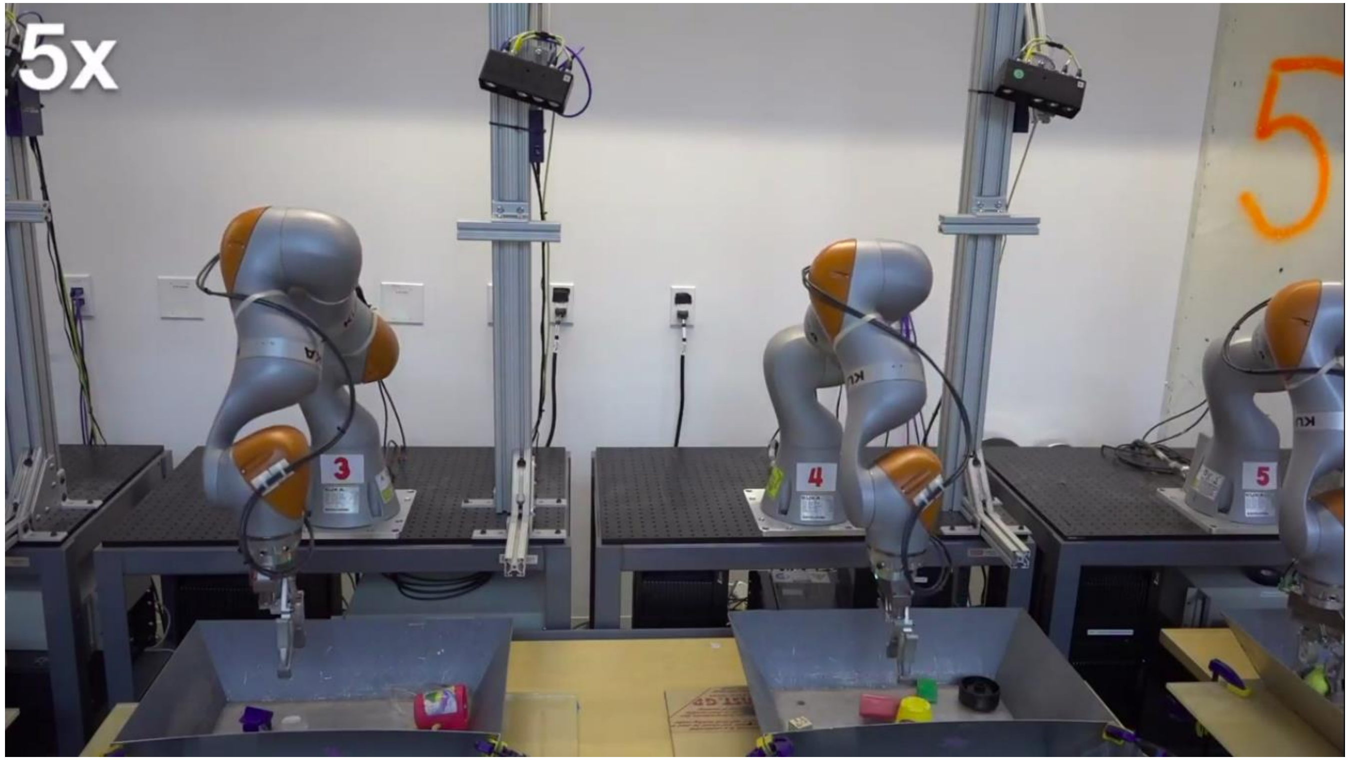
The actions are verified by the local control system.



https://deepmind.google/discover/blog/safety-first-ai-for-autonomous-data-centre-cooling-and-industrial-control/

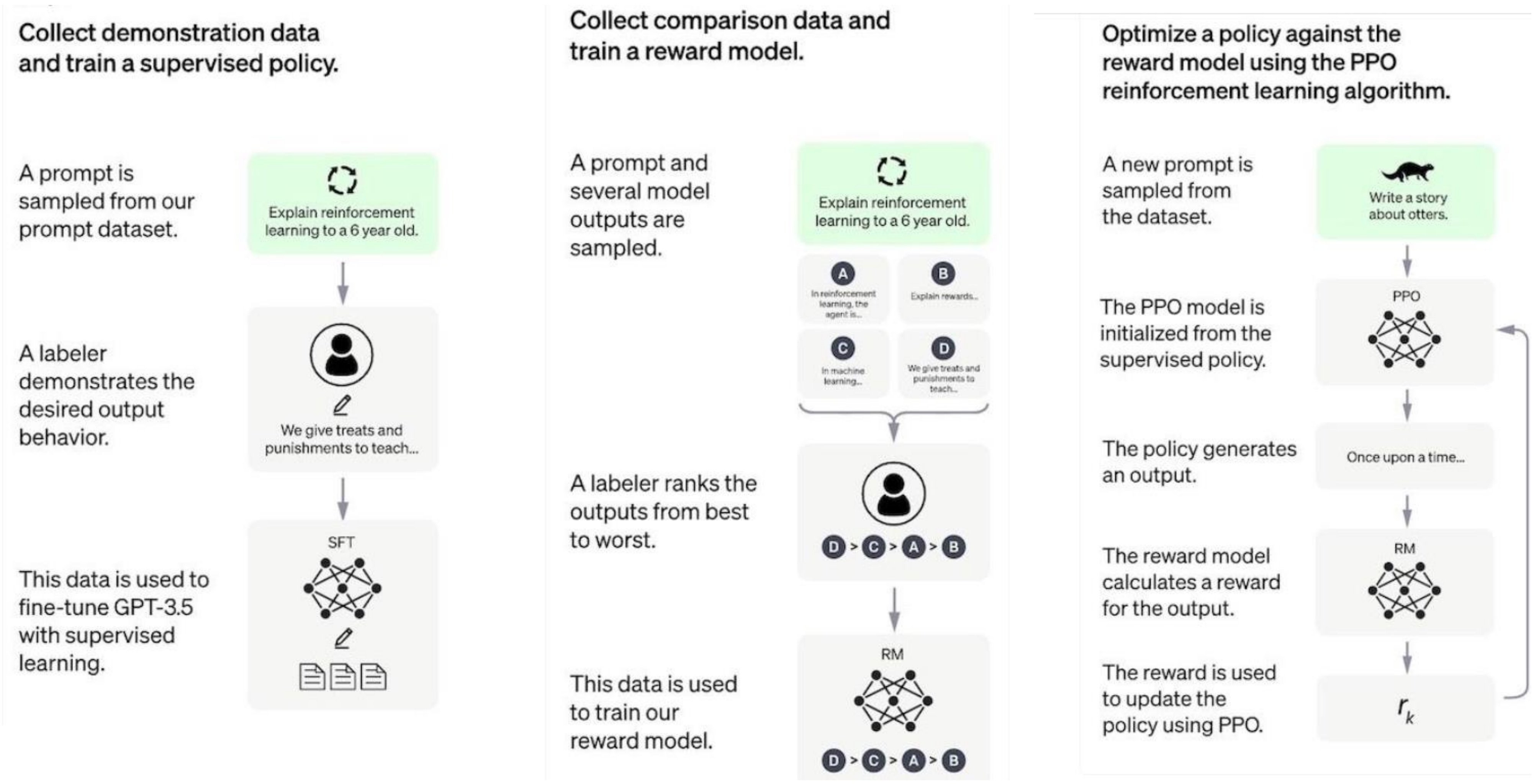
Industry automation – Robotics

Deel and reinforcement learning can train robots that have the ability to grasp various objects—even those unseen during training. This can, for example, be used in building products in an assembly line.
This is achieved by combining large-scale distributed optimization and a variant of deep Q-Learning called QT-Opt. QT-Opt support for continuous action spaces makes it suitable for robotics problems. A model is first trained offline and then deployed and fine-tuned on the real robot.
Google AI applied this approach to **robotics grasping** where 7 real-world robots ran for 800 robot hours in a 4-month period.



https://www.youtube.com/watch?v=W4joe3zzglU

Training the Large Language Models RLHF



Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

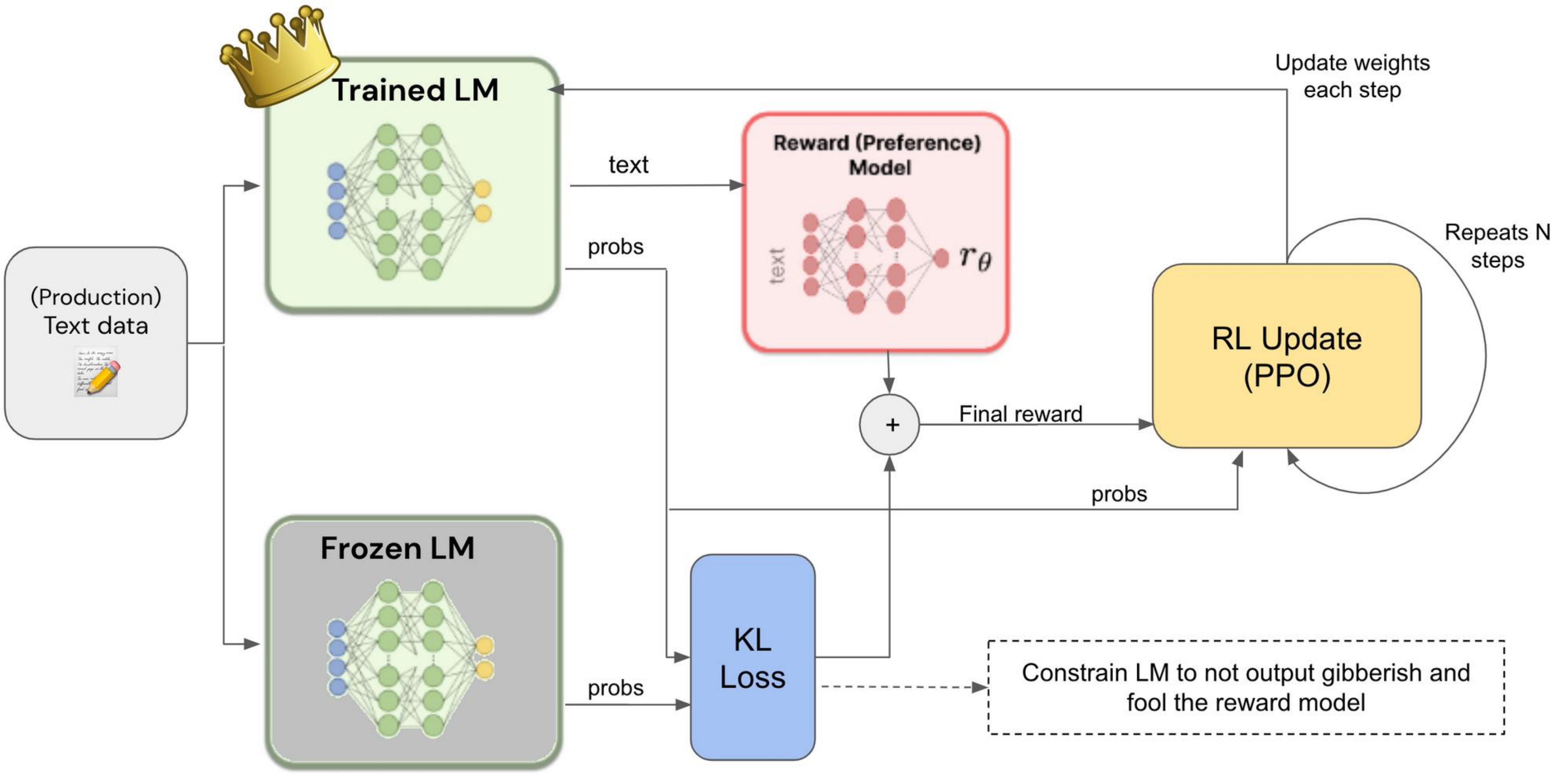
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

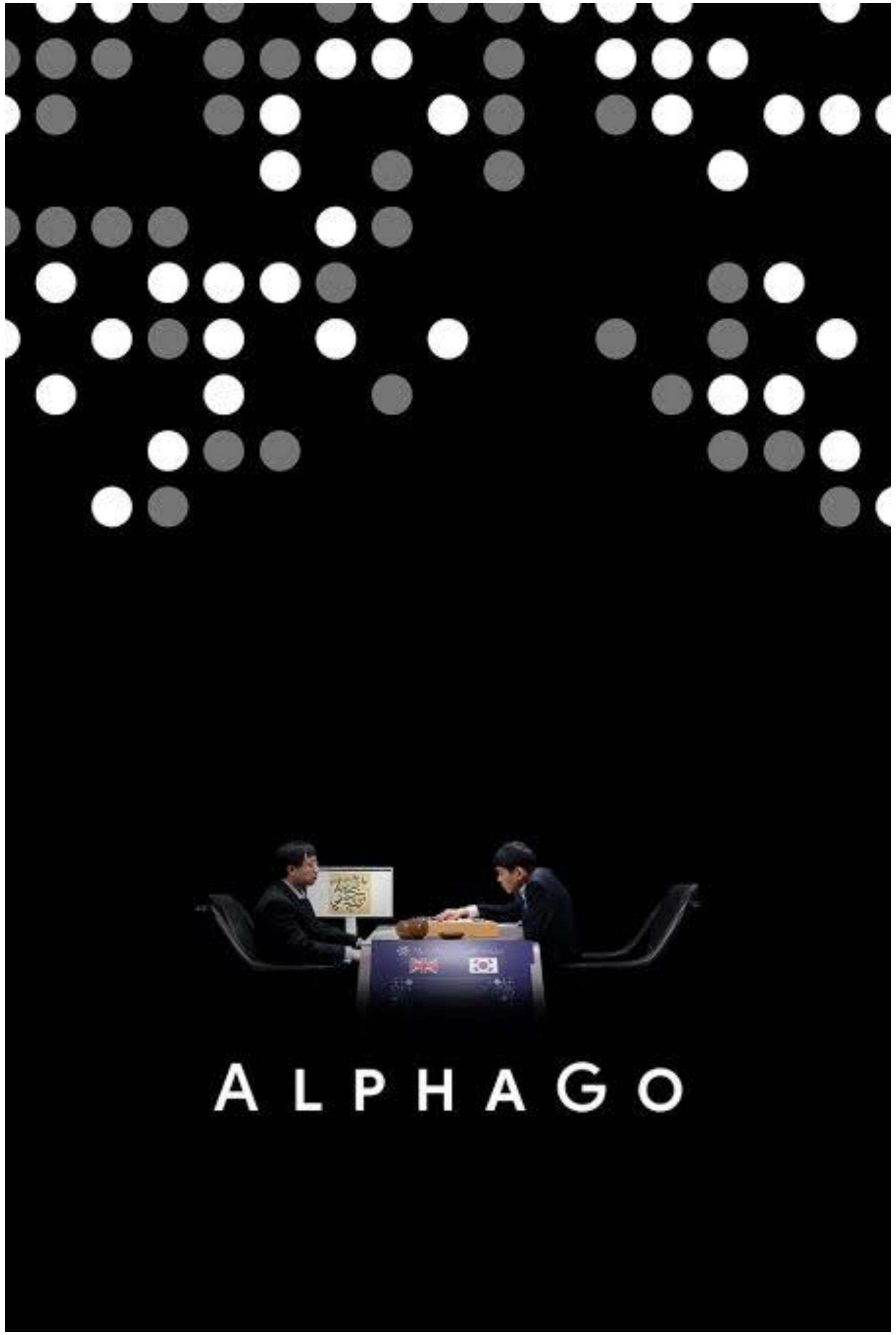
Explain reinforcement learning to a 6 year old.
We give treats and punishments to teach...
SFT
Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B
RM
D > C > A > B
Write a story about otters.
PPO
Once upon a time...
RM
r_k

RLHF



https://www.labellerr.com/blog/reinforcement-learning-with-human-feedback-for-llms/

Must Watch



Wrap-up

Wrap-up
Lecture 1

- **Reinforcement Learning:** Machine Learning Algorithms have 3 learning strategies
  - **Supervised Learning:** Algorithms are trained using labeled examples
  - **Unsupervised Learning:** Algorithms are not trained and find patterns in data
  - **Reinforcement Learning:** The algorithm (agent) learns by performing actions in an universe
- **RL Main components**
  - **Agents**: Is the program or algorithm
  - **Universe**: Is the environment where the Agent leaves and must learn
  - **Action**: Are the possible actions that the Agent can take in the environment
- **Applications**
  - There are many applications and growing
  - LLM are trained using RL as one additional training strategy
  - Autonomous vehicles, autonomous robots, …

END
Lecture 1

IE UNIVERSITY

SCHOOL OF SCIENCE & TECHNOLOGY