Lecture will start at 8:40

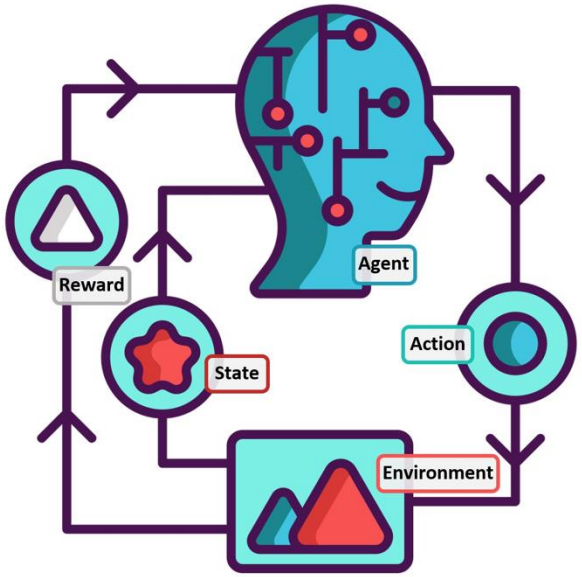
AMCO
ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND CONTROL RESEARCH GROUP

Reinforcement Learning
2025/2026

RL2
Reinforcement Learning Research Lab @ UniPD

Lecture #01
Organization & Intro

Gian Antonio Susto



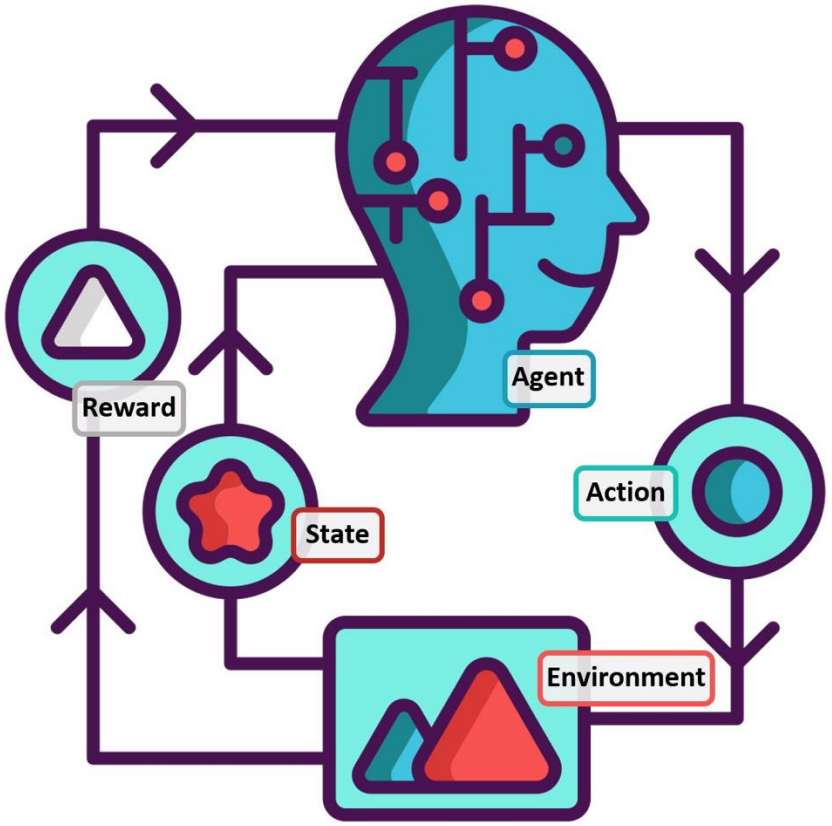
Let’s start with 2 questions…

1. (The difficult one) What is Reinforcement Learning?

Reinforcement Learning

A Machine Learning paradigm which is unique under many aspects:

- We can combine learning and long-term decision making
- We don’t leverage historical data
- We learn ‘on the job’, ie. by interacting with a system/environment



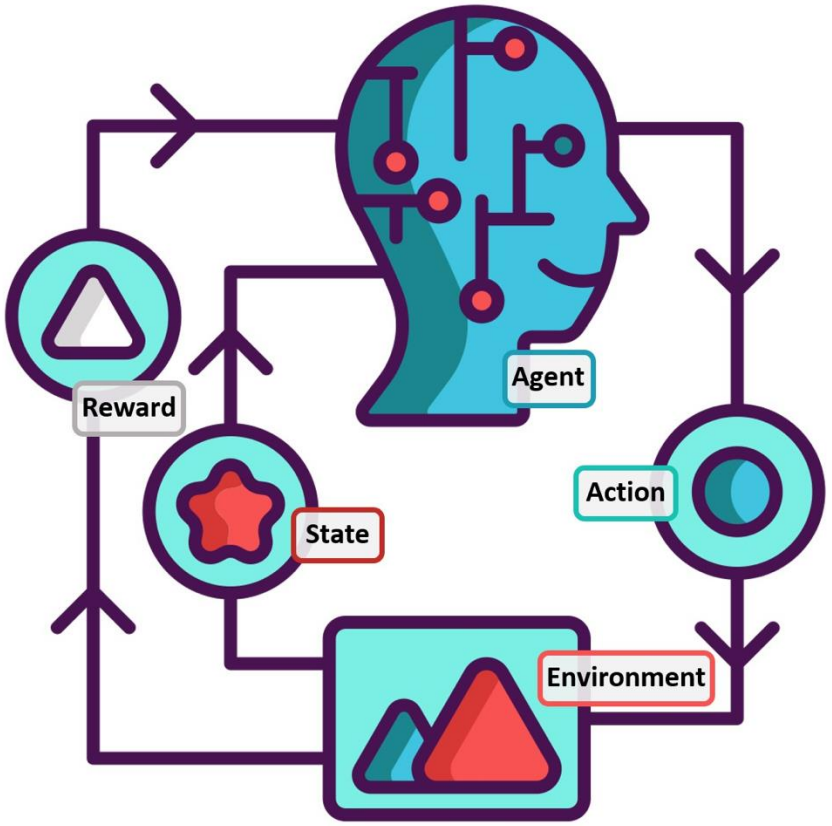
Reinforcement Learning

A Machine Learning paradigm which is unique under many aspects:

- We can combine learning and long-term decision making
- We don't leverage historical data
- We learn 'on the job', ie. by interacting with a system/environment

More on this later!

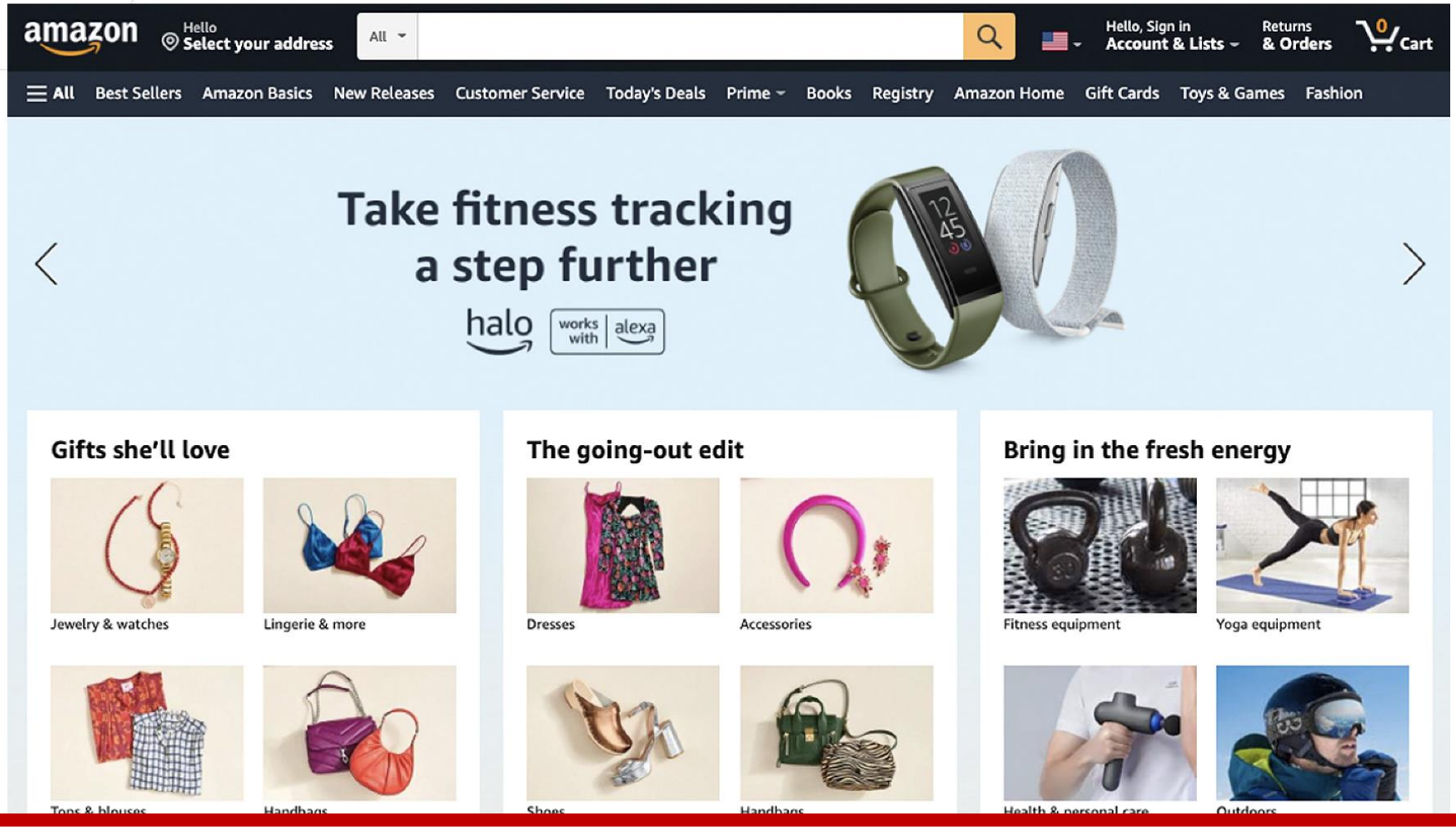
Keep in mind that some problems in Machine Learning can be taken only with Reinforcement Learning approaches!



2. (The easy one) Have you ever interacted with a Reinforcement Learning-based technology?


NETFLIX ORIGINAL
STRANGER THINGS
Popular on Netflix
Recently Watched
okja
13
Master of None
CLUB DE CUERVOS
CROWN
NETFLIX


amazon
TOM CLANCY'S
WITHOUT REMORSE
#1 movie on Prime Video
Personal care best sellers
Shop snack best sellers
Deal of the Day
Sign in for the best experience



How does the Amazon Store know what products and offers to display? Part of the answer involves reinforcement learning. Lihong Li, a senior principal applied scientist, develops reinforcement learning techniques to improve outcomes for customers.

https://www.amazon.science/working-at-amazon/amazon-advertising-lihong-li-using-reinforcement-learning-algorithms

Reinforcement Learning for Budget Constrained Recommendations

Netflix Technology Blog · Follow
Published in Netflix TechBlog · 12 min read · Aug 15, 2022

by Ehtsham Elahi
with James McInerney, Nathan Kallus, Dario Garcia Garcia and Justin Basilico

Introduction

This writeup is about using reinforcement learning to construct an optimal list of recommendations when the user has a finite time budget to make a decision from the list of recommendations. Working within the time budget introduces an extra resource constraint for the recommender system. It is similar to many other decision problems (for e.g. in economics and operations research) where the entity making the decision has to find tradeoffs in the face of finite resources and multiple (possibly conflicting) objectives. Although time is the most important and finite resource, we think that it is an often ignored aspect of recommendation problems.

Netflix Research

Search Research Site

InTune: Reinforcement Learning-based Data Pipeline Optimization for Deep Recommendation Models

Machine Learning | Machine Learning Platform

Publication | Paper

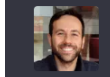
August 13, 2023

Abstract

Deep learning-based recommender models (DLRMs) have become an essential component of many modern recommender systems. Several companies are now building large compute clusters reserved only for DLRM training, driving new interest in cost- & time- saving optimizations. The systems challenges faced in this setting are unique; while typical deep learning (DL) training jobs are dominated by model execution times, the most important factor in DLRM training performance is often online data ingestion. In this paper, we explore the unique characteristics of this data ingestion problem and provide insights into the specific bottlenecks and challenges of the DLRM training pipeline at scale. We study real-world DLRM data processing pipelines taken from our compute cluster at Netflix to both observe the performance impacts of online ingestion and to identify

https://netflixtechblog.com/reinforcement-learning-for-budget-constrained-recommendations-6cbc5263a32a
https://research.netflix.com/publication/intune-reinforcement-learning-based-data-pipeline-optimization-for-deep

Reinforcement Learning-based Technologies



i'm an associate professor: why do I have to teach Reinforcement Learning at 8:30 in the morning?

As an associate professor, you may not have complete control over the scheduling of your courses, including the time slots in which they are offered. Course scheduling decisions are typically made by the academic institution or department based on a variety of factors, including classroom availability, the availability of students and faculty, and other logistical considerations.

There are several reasons why a course, such as Reinforcement Learning, might be scheduled for 8:30 in the morning:

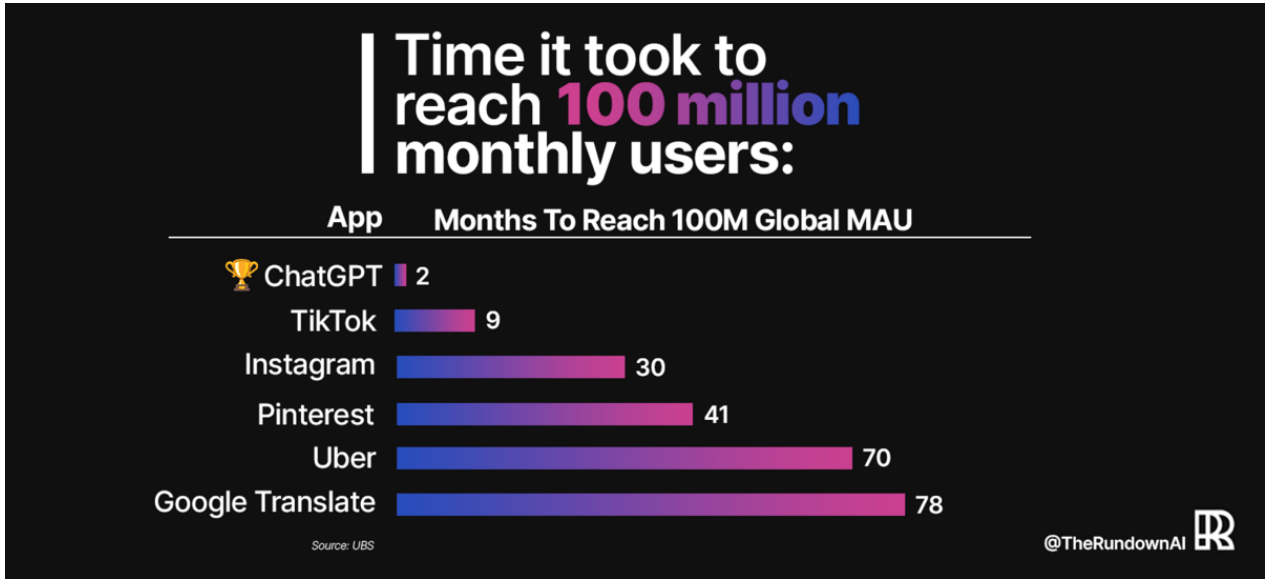
1. Classroom availability: Academic institutions often have limited classroom space, and scheduling courses at various times throughout the day can help maximize the use of available classrooms.
2. Faculty availability: The availability of faculty members to teach at specific times can also influence the scheduling of courses. Some professors may have other teaching commitments or research responsibilities that make them more available in the morning.
3. Student preferences: The scheduling of courses may also take into account student preferences and demand. Some students may prefer morning classes, while others may prefer afternoon or evening classes.
4. Balancing course offerings: Departments may need to balance the scheduling of various courses to ensure that students have access to a diverse set of classes and that prerequisites are met.

Regenerate

Send a message

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT September 25 Version

Reinforcement Learning-based Technologies

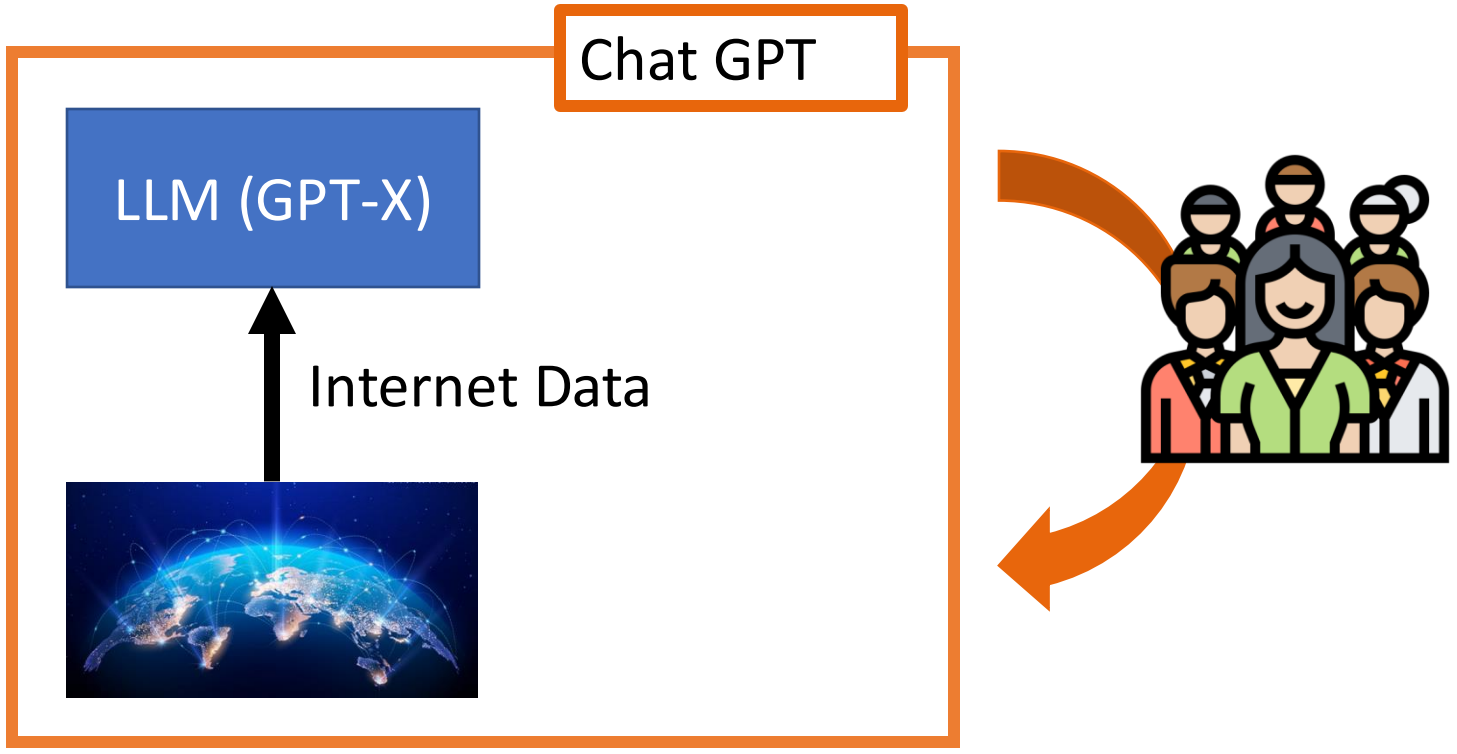


Data is compiled from Similarweb and Semrush as of August 2024. This list does not factor subpages that use the same domain as the parent site.[1][2]

Website	Domain name	Ranking		Type	Company	Country
		Similarweb (August-24)	Semrush (August-24)			
Google Search	google.com	1 (–)	1 (–)	Search Engine	Google	United States
YouTube	youtube.com	2 (–)	2 (–)	Video-sharing platform	Google	United States
Facebook	facebook.com	3 (–)	3 (–)	Social Media Networks	Meta	United States
Instagram	instagram.com	4 (–)	5 (▼)	Social Media Networks	Meta	United States
X	x.com	5 (–)	9 (▲)	Social network	X Corp.	United States
WhatsApp	whatsapp.com	6 (–)	10 (▼)	Social Media Networks	Meta	United States
Wikipedia	wikipedia.org	7 (–)	5 (–)	Dictionaries and Encyclopedias	Wikipedia	United States
Yahoo	yahoo.com	8 (–)	12 (▲)	News & Media Publishers	Yahoo!	United States
Reddit	reddit.com	9 (–)	6 (–)	Social Media Networks	Reddit	United States
Yahoo Japan	yahoo.co.jp	10 (–)	19 (▲)	News & Media Publishers	LY Corporation	Japan
Yandex	yandex.ru	11 (▲)1	18 (▼)	Search Engines	Yandex	Russia
Amazon	amazon.com	12 (▼)1	13 (▲)	Marketplace	Amazon	United States
ChatGPT	chatgpt.com	13 (▲)1	16 (–)	Programming and Developer Software	OpenAI	United States

Reinforcement Learning-based Technologies

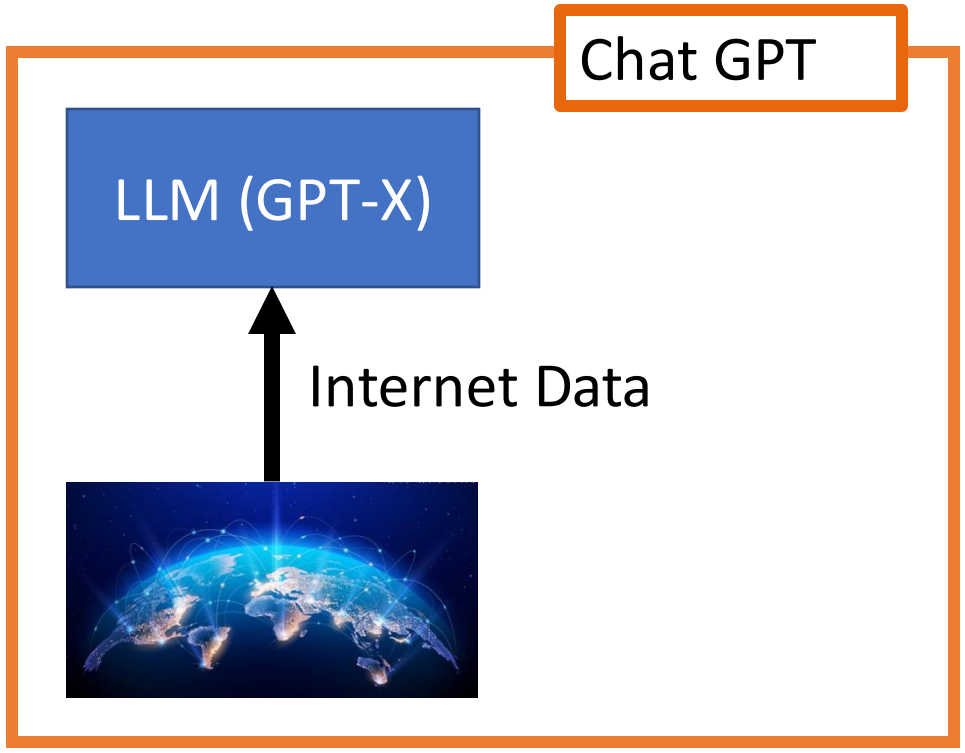
ChatGPT is based on a Large Language Model (LLM) and human feedback (exploited thanks to a Reinforcement Learning procedure)



Reinforcement Learning from human Feedback (RLHF)

Reinforcement Learning-based Technologies

ChatGPT is based on a Large Language Model (LLM) and human feedback (exploited thanks to a Reinforcement Learning procedure)



Reinforcement Learning from human Feedback (RLHF)

New version o1: https://openai.com/index/learning-to-reason-with-llms/

New interdisciplinary ‘AI path’ @ DEI (from 2026-2027)

Common path on AI between M.Sc. Bioengineering, Computer Engineering, Control Systems Engineering, ICT on AI

Mandatory courses:

- Foundations of AI
- Machine Learning
- Deep Learning and Neural Networks
- Reinforcement Learning

19/09/25 Presentation to the stakeholder of the teaching initiative ‘AI@DEI’

Outline

- Organizational Aspects
- Intro to Reinforcement Learning
- Course Outline
- Reinforcement Learning Elements

Organizational Aspects

Lecturer

Gian Antonio Susto

gianantonio.susto@unipd.it

- Member of the Automatica Group
- Coordinator of the AMCO (Artificial intelligence, Machine learning and COntrol) Lab
- Co-founder of the Reinforcement Learning Research Lab
- Heavily involved with companies/industries

Lecturer

AMCO
ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND CONTROL RESEARCH GROUP

- Member of the Automatica Group
- Coordinator of the AMCO (Artificial intelligence, Machine learning and COntrol) Lab
- Co-founder of the Reinforcement Learning Research Lab
- Heavily involved with companies/industries

➔ Research on: Reinforcement Learning, Anomaly Detection, Continual Learning, Explainability and Fairness, Deep Learning, Machine Learning Applications

➔ More info @:

https://amco.dei.unipd.it/ (work-in-progress)

https://scholar.google.it/citations?user=7bgABaoAAAAJ&hl=it&oi=ao

https://www.linkedin.com/company/amco-lab-unipd/

Lecturer

Gian Antonio Susto

gianantonio.susto@unipd.it

- Member of the Automatica Group
- Coordinator of the AMCO (Artificial intelligence, Machine learning and COntrol) Lab
- Co-founder of the Reinforcement Learning Research Lab
- Heavily involved with companies/industries

➔ Interdisciplinary lab.

➔ Guest lectures later in the course: Alberto Dalla Libera, Ruggero Carli, Federico Chiariotti

➔ More info @:

https://reinforcementlearning.dei.unipd.it/ (work-in-progress)

RL2
Reinforcement Learning
Research Lab @ UniPD

Lecturer

Gian Antonio Susto

gianantonio.susto@unipd.it

- Member of the Automatica Group
- Coordinator of the AMCO (Artificial intelligence, Machine learning and COntrol) Lab
- Co-founder of the Reinforcement Learning Research Lab
- Heavily involved with companies/industries

➔ Collaborations with several companies: Statwolf (co-founder), Infineon, Technogym, Diasorin, LFoundry, Seagate, Breton, Swegon, Galdi, Santex, GoldenGoose, Electrolux, Zoppas Industries, Maschio Gaspardo, Luxottica, Pietro Fiorentini, …

Lecturer

Gian Antonio Susto

gianantonio.susto@unipd.it

- Member of the Automatica Group
- Coordinator of the AMCO (Artificial intelligence, Machine learning and COntrol) Lab
- Co-founder of the Reinforcement Learning Research Lab
- Heavily involved with companies/industries

➔ Collaborations with several companies: Statwolf (co-founder), Infineon, Technogym, Diasorin, LFoundry, Seagate, Breton, Swegon, Galdi, Santex, GoldenGoose, Electrolux, Zoppas Industries, Maschio Gaspardo, Luxottica, Pietro Fiorentini, …

➔ Yes, I do have thesis topics in collaboration with industries

Main Teaching Assistants

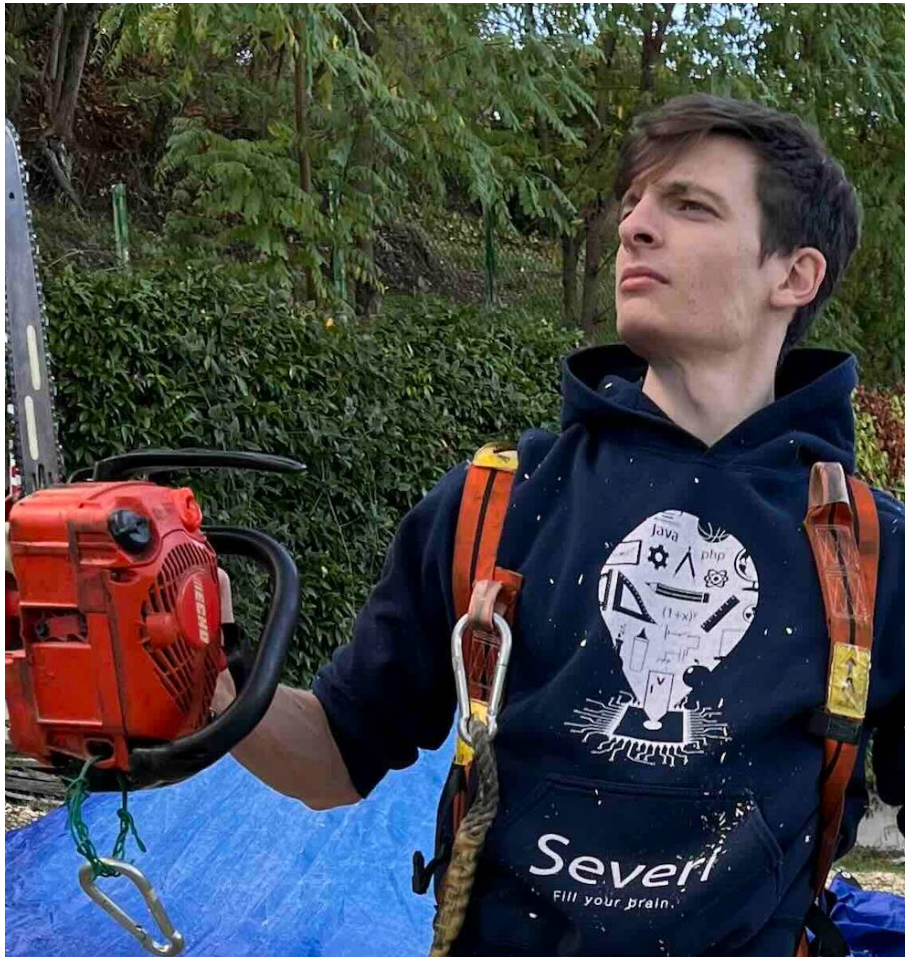
Alberto Sinigaglia, PhD Student

He’ll be involved:

- Laboratories
- Games Project

Data Scientist, attended RL2022-2023 class

alberto.sinigaglia@phd.unipd.it



Main Teaching Assistants

Alberto Sinigaglia, PhD Student

He’ll be involved:

- Laboratories
- Games Project

Data Scientist, attended RL2022-2023 class

alberto.sinigaglia@phd.unipd.it

Main Teaching Assistants

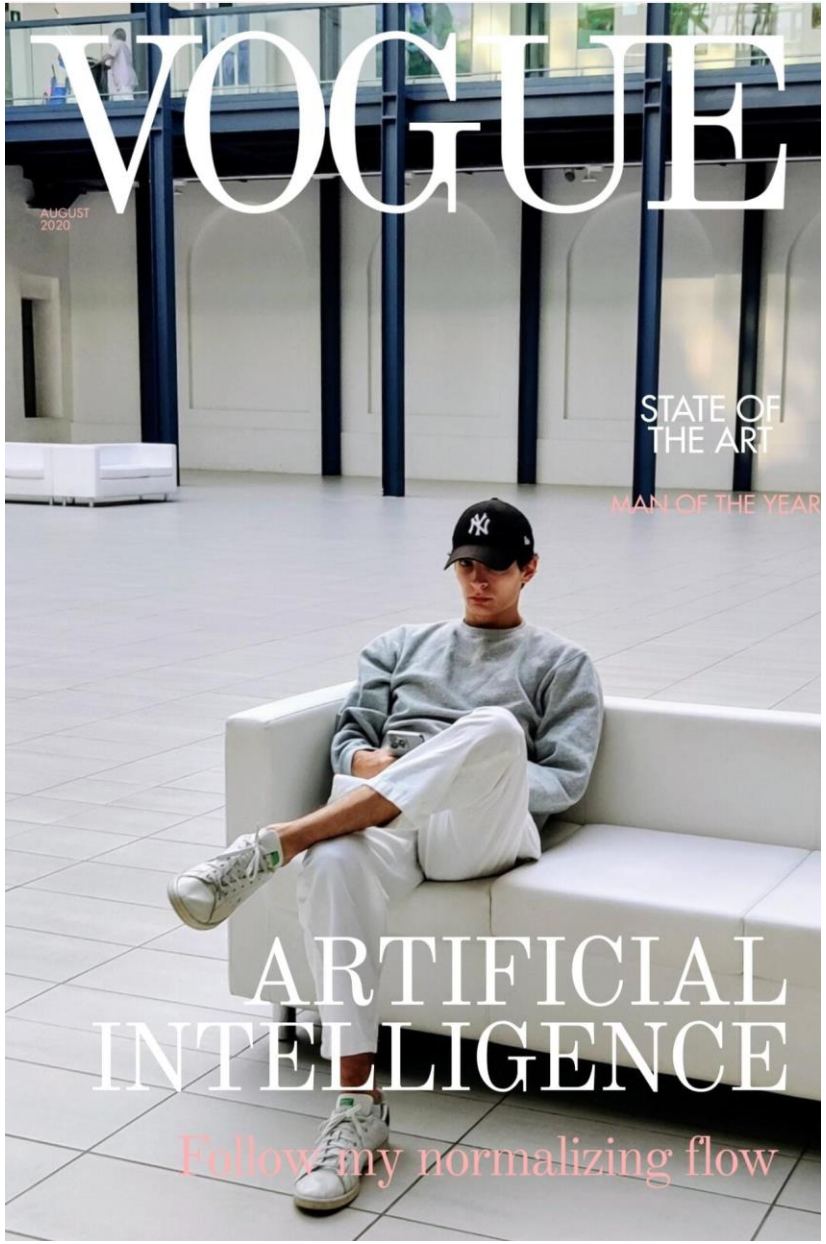
Alberto Sinigaglia, PhD Student

He’ll be involved:

- Laboratories
- Games Project

Data Scientist, attended RL2022-2023 class

alberto.sinigaglia@phd.unipd.it



Main Teaching Assistants

Riccardo De Monte, PhD Student

He'll be involved:

- Laboratories
- Deep Learning Lectures

Control Systems Engineer, attended RL2022-2023 class

riccardo.demonte@phd.unipd.it

Other Teaching Assistants

Matteo Cederle, PhD Student
(Driving Project)
Control Systems Engineer, attended RL2022-2023 class
matteo.cederle@phd.unipd.it

Alessandro Adami, PhD Student
(PyTorch Laboratory + Robotics Project)
Control Systems Engineer
alessandro.adami.4@studenti.unipd.it

Who are you

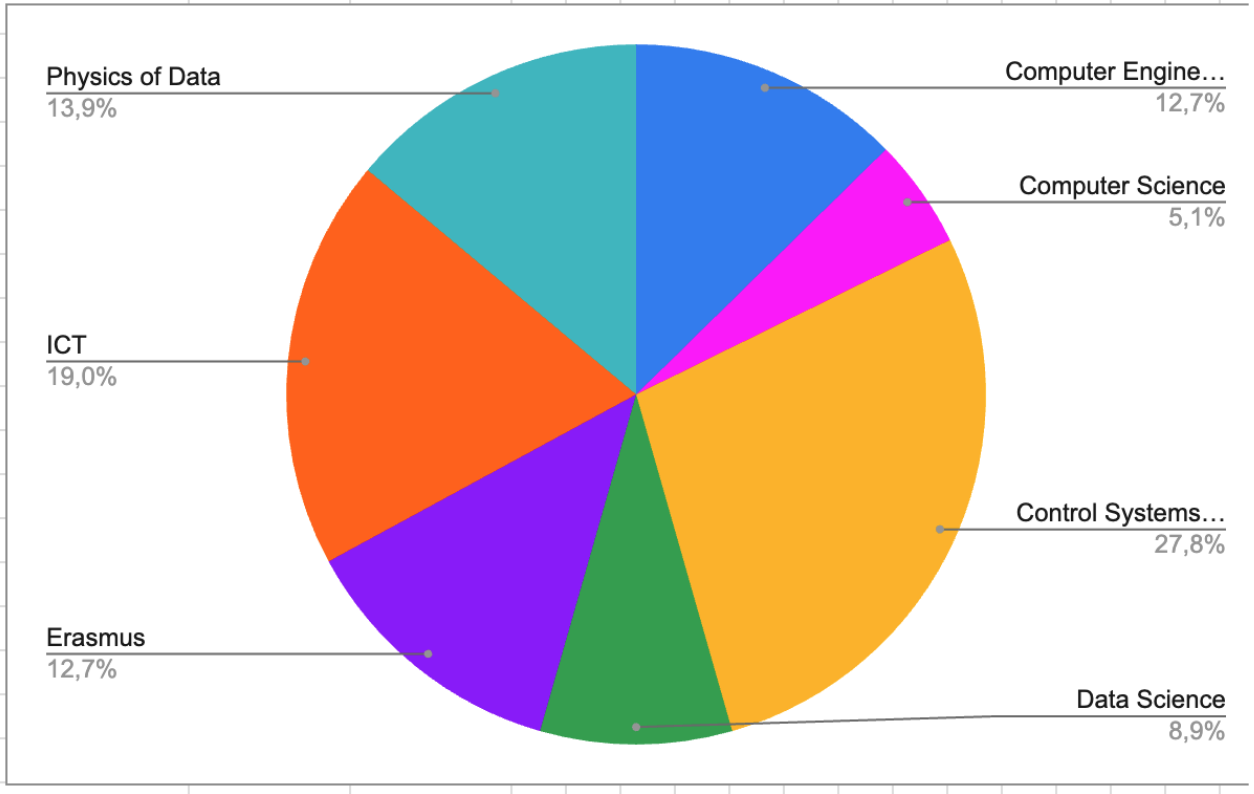
Course offered for: Control Systems Eng., ICT, Data Science, Computer Eng.

Who are you

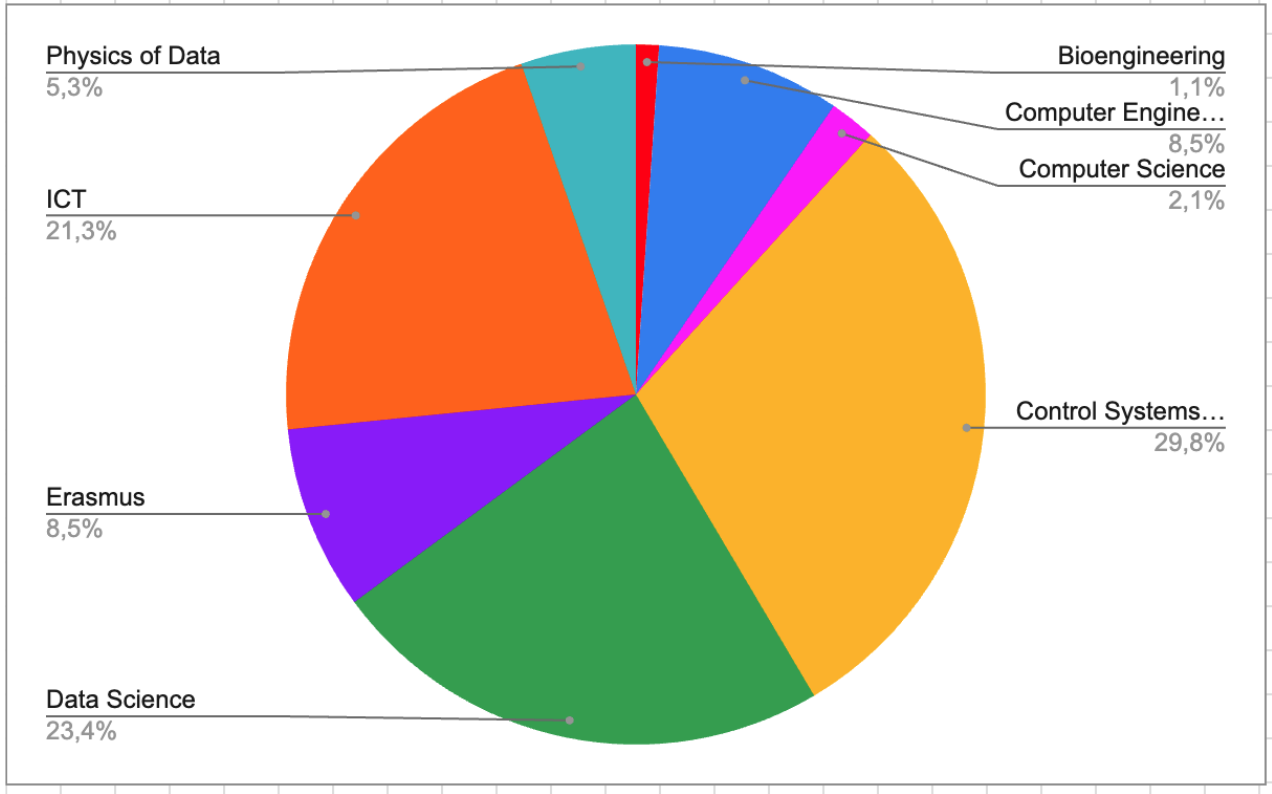
Course offered for: Control Systems Eng., ICT, Data Science, Computer Eng.

Also suggested for: Physics of Data, Computer Engineering, Computer Science, Erasmus, …

2023



2024



Lectures and Course Page

Lectures on:

Wednesday @ 8:40-10:10 – Room Le

Thursday @ 10:30-12:00 – Room De

We are planning on finishing all the content before the end of the year, however things may go differently…

Moodle Page: https://stem.elearning.unipd.it/course/view.php?id=14042

Lectures and Course Page

Lectures on:

Wednesday @ 8:40-10:10 – Room Le

Thursday @ 10:30-12:00 – Room De

We are planning on finishing all the content before the end of the year, however things may go differently…

Moodle Page: https://stem.elearning.unipd.it/course/view.php?id=14042

Programming Laboratory: (rooms to be confirmed)

- Lab 1a ‘Deep Learning frameworks - Tensorflow’ Thu 2 Oct 2025 8:40-10:10 Room Te + Ue
- Lab 1b ‘Deep Learning frameworks - Pytorch’ Fri 3 Oct 2025 14:30-16:00 Room De

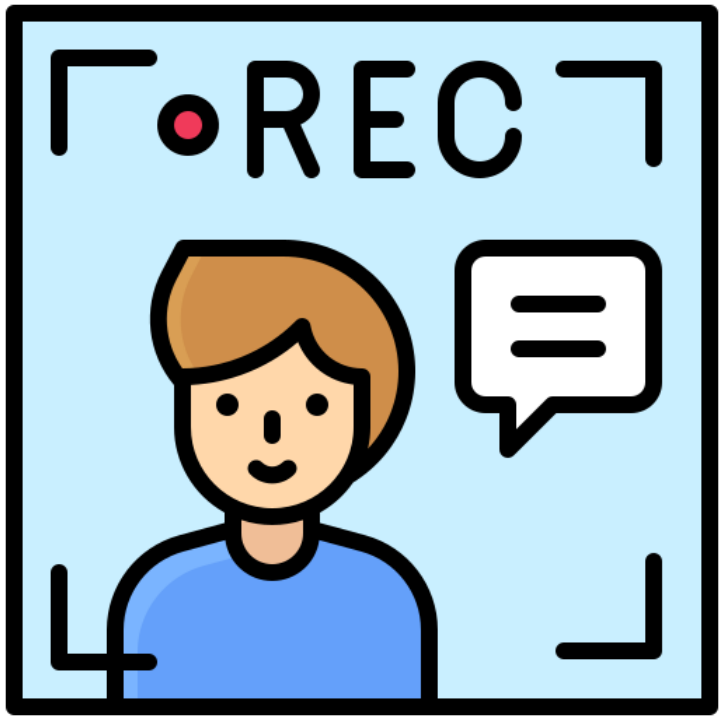
(Lab 1a and 1b can be considered as alternative and we recommend to attend the one with the related project you'd like to do)

- Lab 2 ‘k-armed Bandits’ Fri 10 Oct 2025 14:30-16:00 Room De
- Lab 3 ‘Dynamic Programming’ Fri 17 Oct 2025 14:30-16:00 Room De
- Lab 4 ‘Monte Carlo’ Fri 24 Oct 2025 14:30-16:00 Room De
- Lab 5 ‘TD Learning’ Fri 31 Oct 2025 14:30-16:00 Room De
- Lab 6 ‘Deep RL’ Early Jan 2026 (dates to be decided)

Course Recordings

University of Padova imposes synchronous, in presence participation.

Lecture and laboratories recordings will be made available shortly after the lecture.



Course Material

- Main references:

1. Course Slides

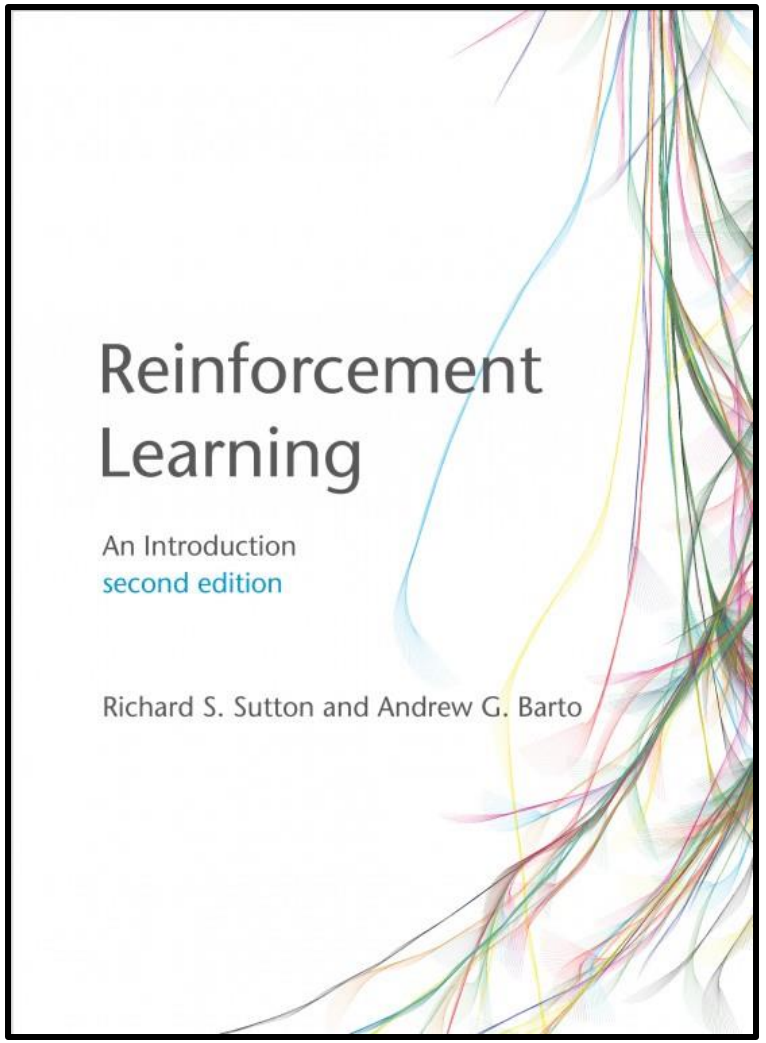
2. Richard Sutton, Andrew Barto ‘Reinforcement Learning. An Introduction’ 2nd Edition 2018

Available for free (but consider buying the book if you can!): http://incompleteideas.net/book/the-book.html

At the same links you’ll find code for the book example

- Other references:

We will take some contents from other papers and materials (for example D. Bertsekas ‘Reinforcement Learning and Optimal Control’), but everything ‘extra’ – relevant to the exam – will be in the slides



Exams 1/2

Written Part (Mandatory, max 28!)

Written exam: on content treated during the lectures, with special focus on RL 'principles' (concepts and scenarios) and algorithms.

How to train for the written exam:

- follow the lectures;
- study the slides and the associated book chapters in the book;
- answer the exercises in the book;
- take a look at previous years exams on the moodle page.

Grades up to 32/33, but without the 'project' you cannot register more than 28

Exams 1/2

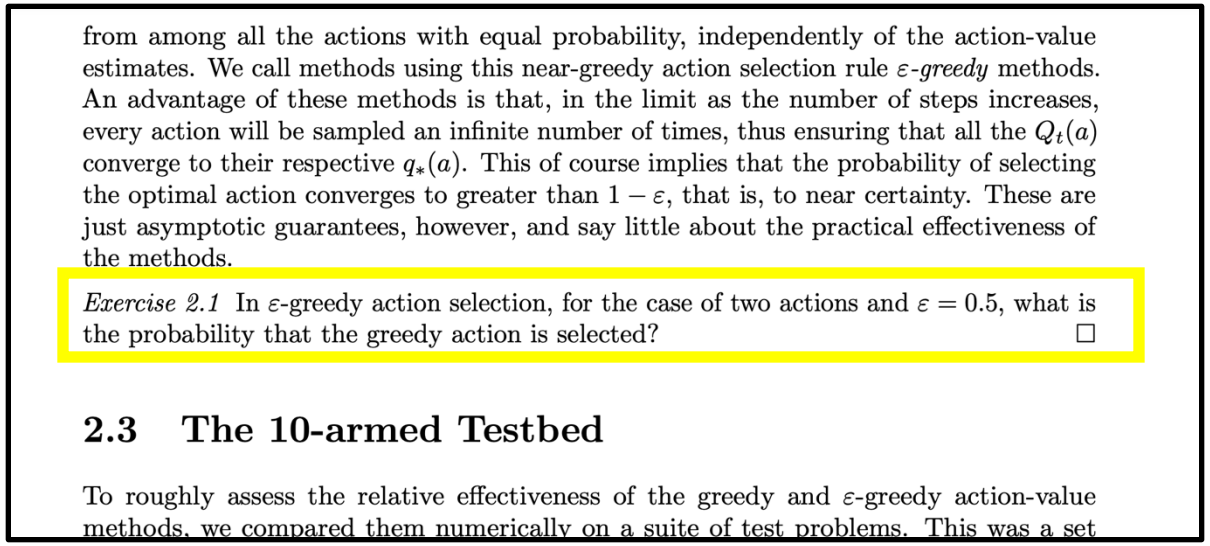
Written Part (Mandatory, max 28!)

Written exam: on content treated during the lectures, with special focus on RL 'principles' (concepts and scenarios) and algorithms.

How to train for the written exam:

- follow the lectures;
- study the slides and the associated book chapters in the book;
- answer the exercises in the book;
- take a look at previous years exams on the moodle page.

Grades up to 32/33, but without the 'project' you cannot register more than 28



from among all the actions with equal probability, independently of the action-value estimates. We call methods using this near-greedy action selection rule *ε-greedy* methods. An advantage of these methods is that, in the limit as the number of steps increases, every action will be sampled an infinite number of times, thus ensuring that all the $Q_t(a)$ converge to their respective $q_*(a)$. This of course implies that the probability of selecting the optimal action converges to greater than $1 - \varepsilon$, that is, to near certainty. These are just asymptotic guarantees, however, and say little about the practical effectiveness of the methods.

Exercise 2.1 In ε-greedy action selection, for the case of two actions and $\varepsilon = 0.5$, what is the probability that the greedy action is selected? □

2.3 The 10-armed Testbed

To roughly assess the relative effectiveness of the greedy and ε-greedy action-value methods, we compared them numerically on a suite of test problems. This was a set

Exams 1/2

Written Part (Mandatory, max 28!)

Written exam: on content treated during the lectures, with special focus on RL 'principles' (concepts and scenarios) and algorithms.

How to train for the written exam:

- follow the lectures;
- study the slides and the associated book chapters in the book;
- answer the exercises in the book;
- take a look at previous years exams on the moodle page.

Grades up to 32/33, but without the 'project' you cannot register more than 28

Two options for the written exam:

'Partial' Exams:

Part A (50% weight, materials from Lec.1 to Lec.10) – Friday November 7th, 2025 16:45-18:15

Part B (50% weight, material from Lec. 11 to Lec. 20) – Friday December 19th, 2025 16:45-18:15

**Part B exam can be taken only if a sufficient mark is taken in Part A*

'Regular' Exams:

Friday January 23th, 2026 14:30-18:00
Friday February 13th, 2026 14:30-18:00
Friday June 19th, 2026 9:30-12:00
Friday September 4th, 2026 14:30-18:00

Exams 2/2

Projects (Optional, you can reach above 28)

Practical exam (individual programming project) based on applying RL ‘principles.

- The practical exam is optional, you can stop at the written part, but performing a project could allow you to get a better mark.
- You can do the practical exam only if you are sufficient in the written part
- You can finish (and present) the project until the end of September, if you pass the written part on September, until the end of the year

Exams 2/2

Projects (Optional, you can reach above 28)

Practical exam (individual programming project) based on applying RL 'principles.

- The practical exam is optional, you can stop at the written part, but performing a project could allow you to get a better mark.
- You can do the practical exam only if you are sufficient in the written part
- You can finish (and present) the project until the end of September, if you pass the written part on the August exam, until the end of the year

Grading algorithm:

Was the project performed?

➔ YES ->

Was your project sufficient?

YES -> Final mark = average(written part*, project**)****

NO -> Final mark = min(28, written part)****

➔ NO -> Final mark = min(28, written part)

* The written part will be graded in a scale of at least 32

** The project will be graded in a scale of 32

*** In case of grades with decimals, they will be rounded up / In case

**** If you perform the project, you cannot decrease your final score

Exams 2/2

Projects (Optional, you can reach above 28)

Practical exam (individual programming project) based on applying RL 'principles.

- The practical exam is optional, you can stop at the written part, but performing a project could allow you to get a better mark.
- You can do the practical exam only if you are sufficient in the written part
- You can finish (and present) the project until the end of September, if you pass the written part on the August exam, until the end of the year

Grading algorithm:

Was the project performed?

➔ YES ->

Was your project sufficient?

YES -> Final mark = average(written part*, project**)****

NO -> Final mark = min(28, written part)****

➔ NO -> Final mark = min(28, written part)

Let's see some examples!

* The written part will be graded in a scale of at least 32

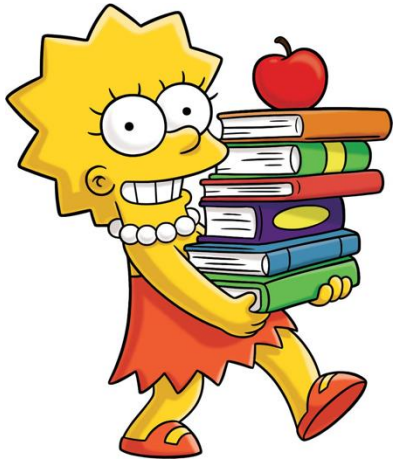
** The project will be graded in a scale of 32

*** In case of grades with decimals, they will be rounded up / In case

**** If you perform the project, you cannot decrease your final score

Exams 2/2

Projects (Optional, you can reach above 28)



Student #01

- Partial exams graded: first 33 over 33, second 32 over 32
- Project done before the second semester stars (so she can start her thesis on Reinforcement Learning as soon as possible): graded 32 over 32
- Final grade: 30 with laude

Student #02

- Partial exams graded: first 29 over 33, second 28 over 32 (average 28.5)
- Project done during summer: graded 30 over 32
- Final grade: 29 (average 29.25)

Exams 2/2

Projects (Optional, you can reach above 28)

Student #03

- Passed the January exam: 32 over 32
- Did not took the exam (too busy, upgrading the grade ended up not being necessary)
- Waited until the 30th of September to register the grade
- Final grade: 28

Student #04

- Passed the exam on the August exams graded: graded 23 over 32
- Project submitted in December: graded 26 over 32
- Final grade: 25 (average 24.5 rounded up)

Exams 2/2

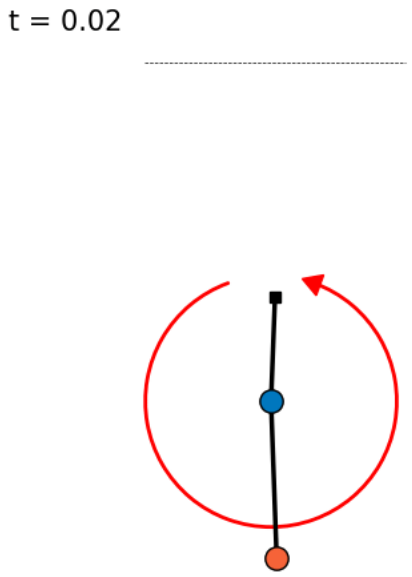
Projects (Optional, you can reach above 28)

Past year ‘Standard’ projects

(5-10 days of work)

- Robotics
- Snake
- Briscola
- Autonomous

Driving



Exams 2/2

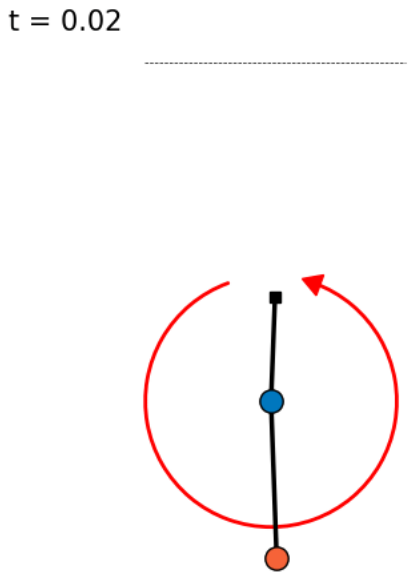
Projects (Optional, you can reach above 28)

Past year ‘Standard’ projects

(5-10 days of work)

- Robotics
- Snake
- Briscola
- Autonomous

Driving



Exams 2/2

Projects (Optional, you can reach above 28)

Past year ‘Standard’ projects

(5-10 days of work)

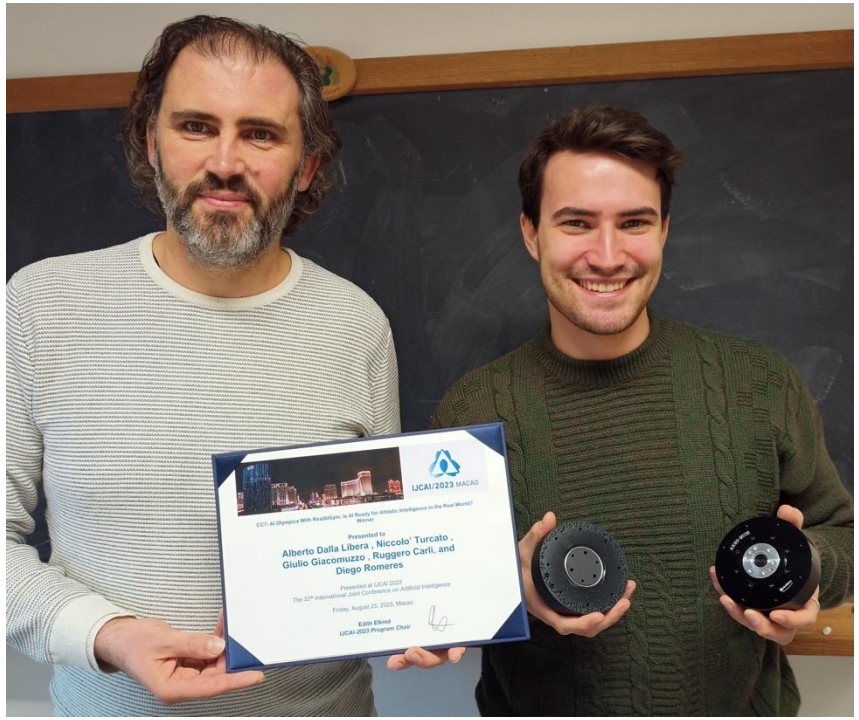
- Robotics
- Snake
- Briscola
- Autonomous

Driving

- Good solutions may be eligible to partecipate to the international **AI Olympics with RealAI Gym competition**

- Our group has a strong record of good results in the competition:

- **We won the first edition**
- Good results in the second edition (ongoing)



https://doi.org/10.24963/ijcai.2024/1043

Exams 2/2

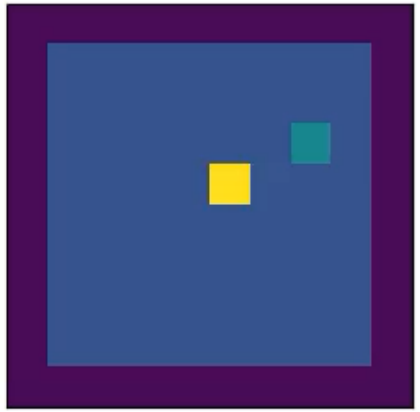
Projects (Optional, you can reach above 28)

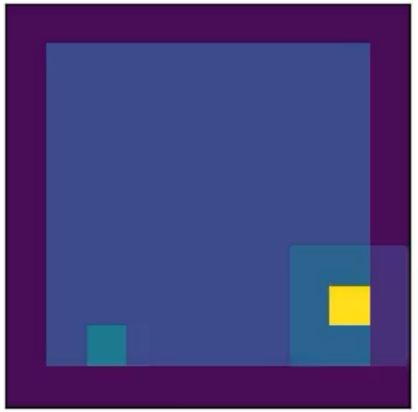
Past year ‘Standard’ projects

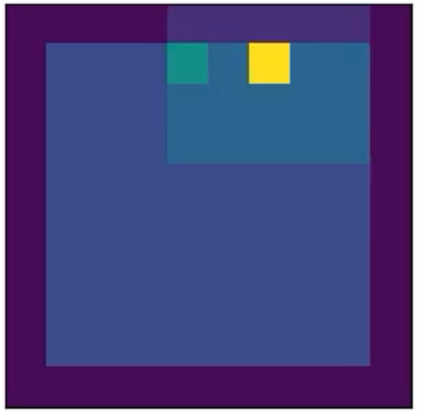
(5-10 days of work)

- Robotics
- Snake
- Briscola
- Autonomous

Driving







Exams 2/2

Projects (Optional, you can reach above 28)

Past year ‘Standard’ projects

(5-10 days of work)

- Robotics
- Snake
- Briscola
- Autonomous

Driving

Exams 2/2

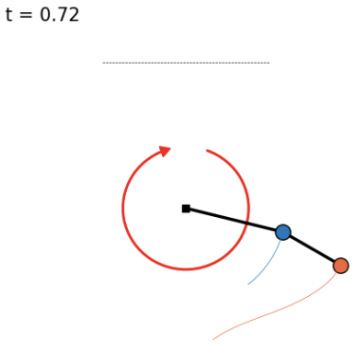
Projects (Optional, you can reach above 28)

Past year ‘Standard’ projects

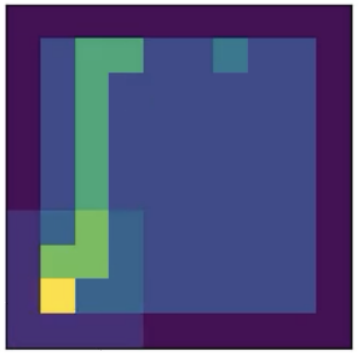
(5-10 days of work)

- Robotics
- Snake
- Briscola
- Autonomous

Driving

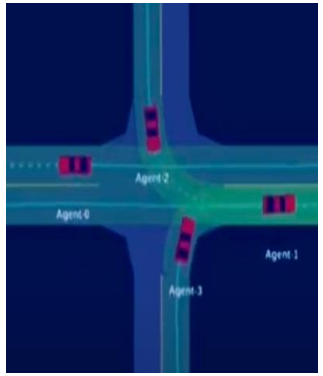


PyTorch



TensorFlow

TensorFlow



PyTorch

Exams 2/2

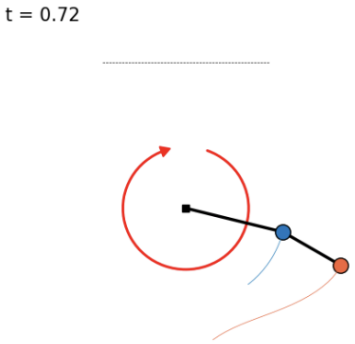
Projects (Optional, you can reach above 28)

Past year ‘Standard’ projects

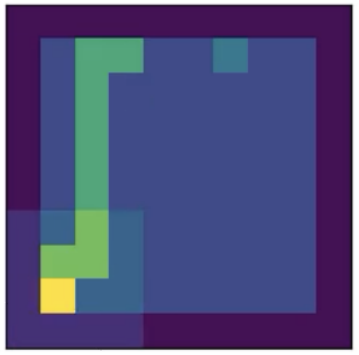
(5-10 days of work)

- Robotics
- Snake
- Briscola
- Autonomous

Driving

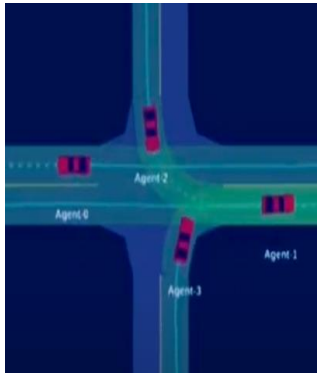


PyTorch



TensorFlow

TensorFlow



PyTorch

Advanced projects

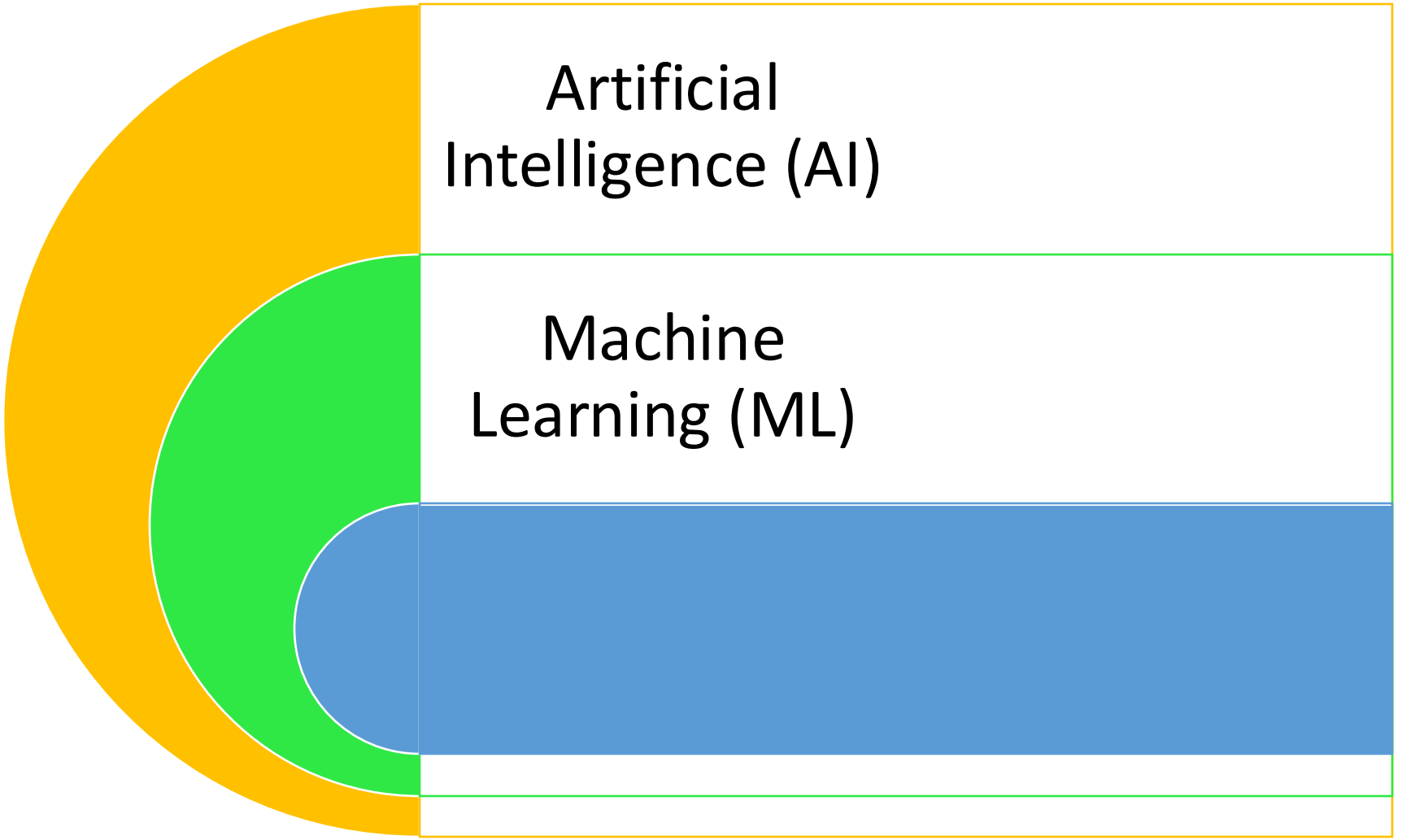
- More challenging tasks, presented by us on the last lecture of the course or students’ ideas.
- Designed for students who want to do a thesis on Reinforcement Learning / intermediate step of the thesis

Communications

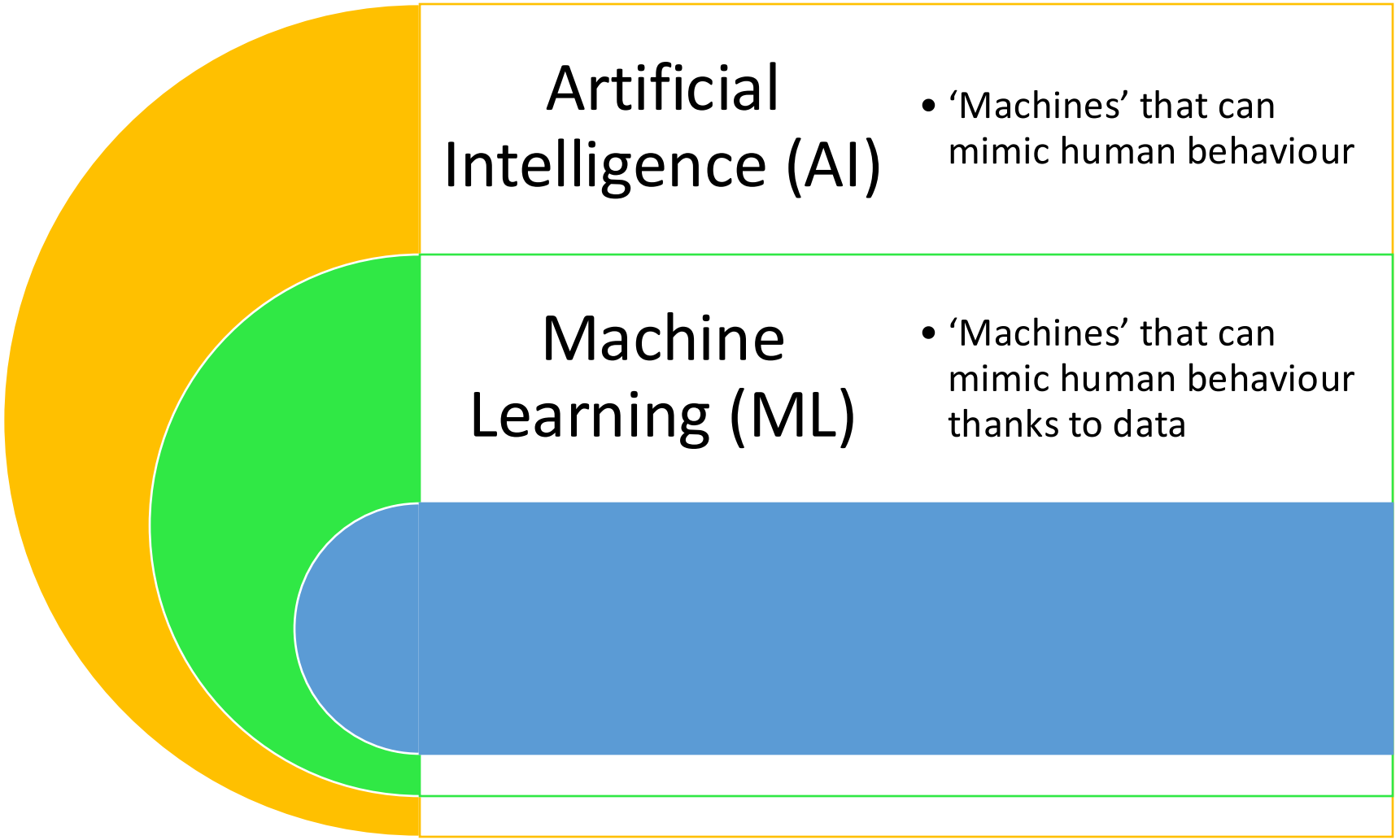
- In email communications, please use object ‘[RL2025-26] <your title>’
- If you have a question, probably your colleagues has the same/similar one: we will try to answer during the lectures

Introduction to RL

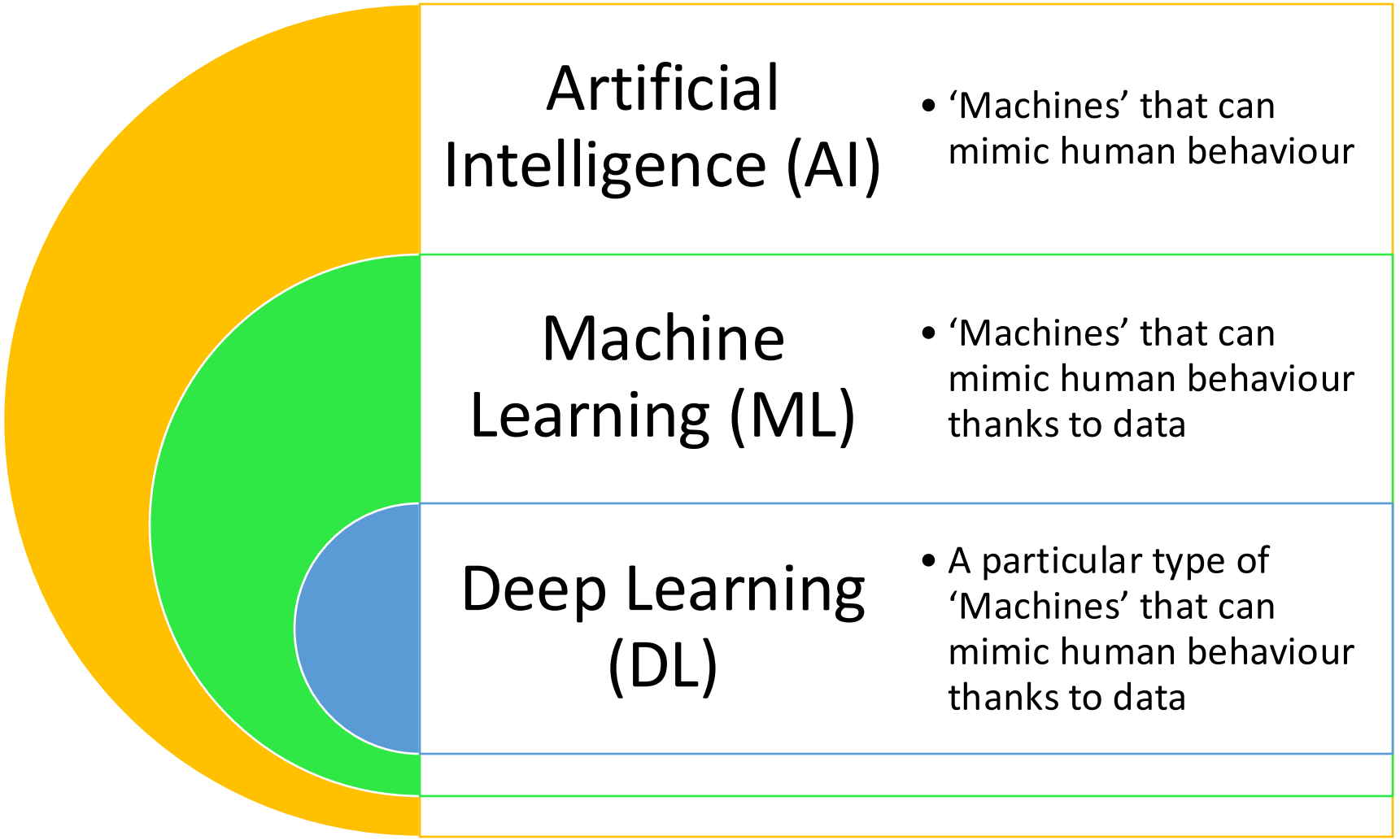
What is ~~Reinforcement~~ Learning?



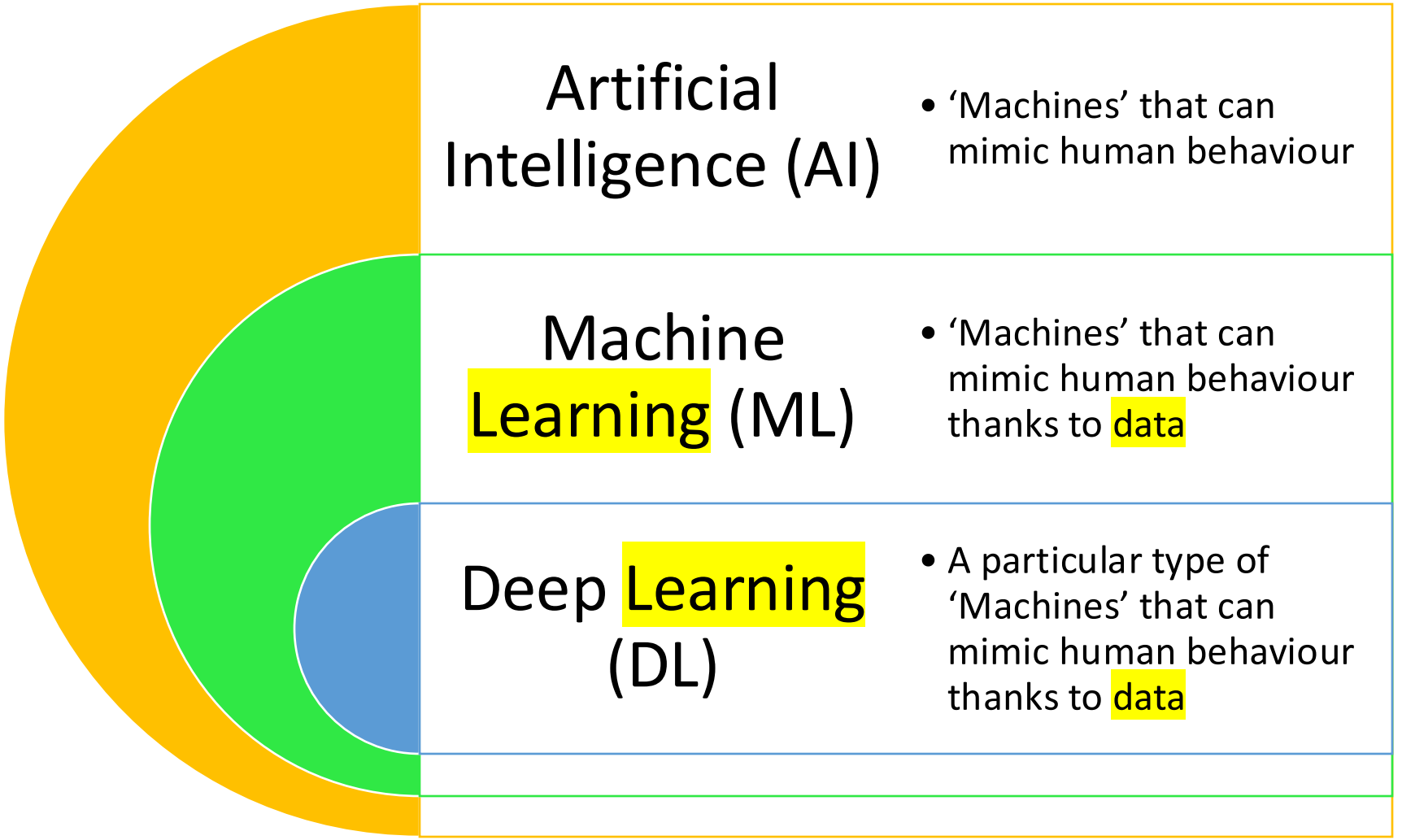
What is ~~Reinforcement~~ Learning?



What is ~~Reinforcement~~ Learning?

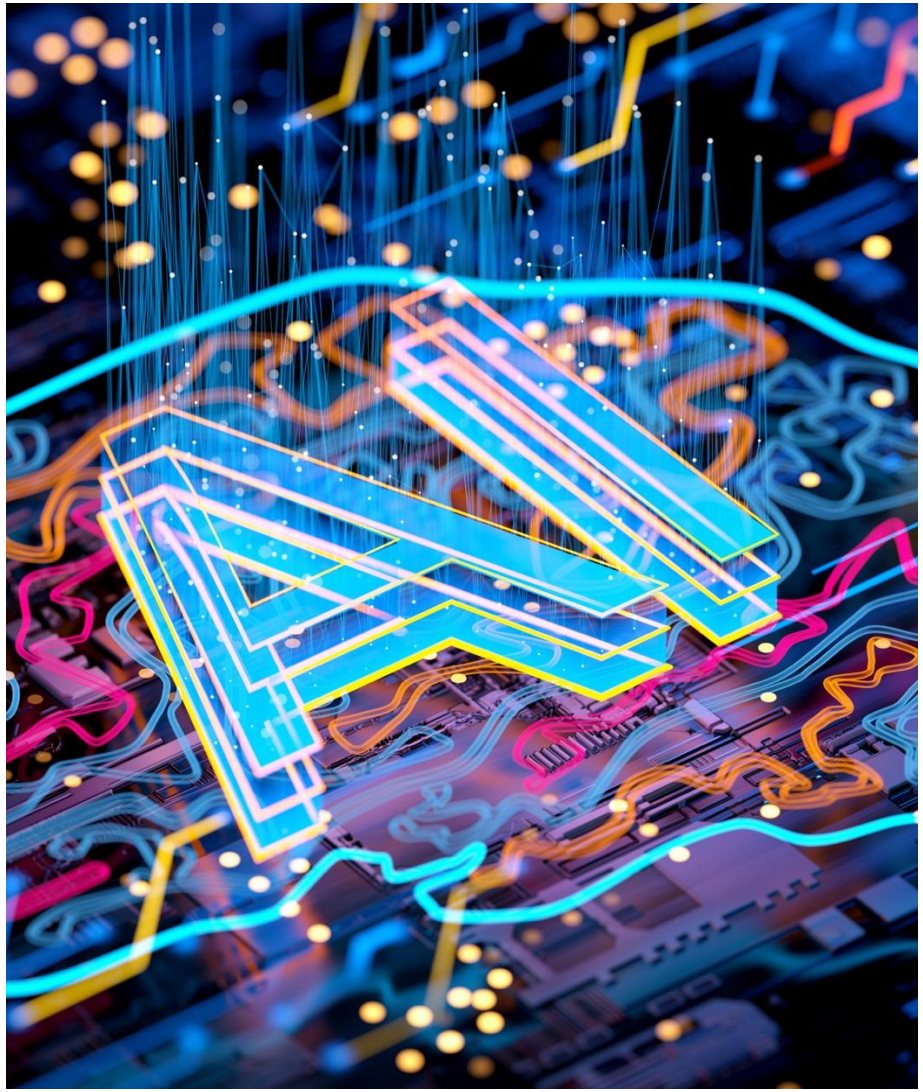


What is ~~Reinforcement~~ Learning?

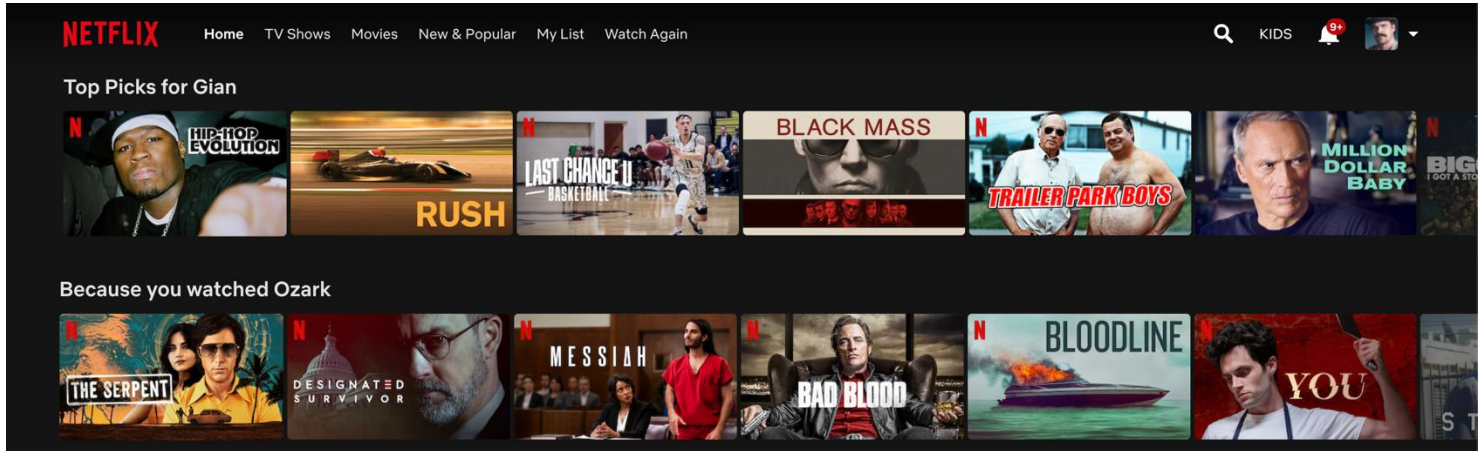


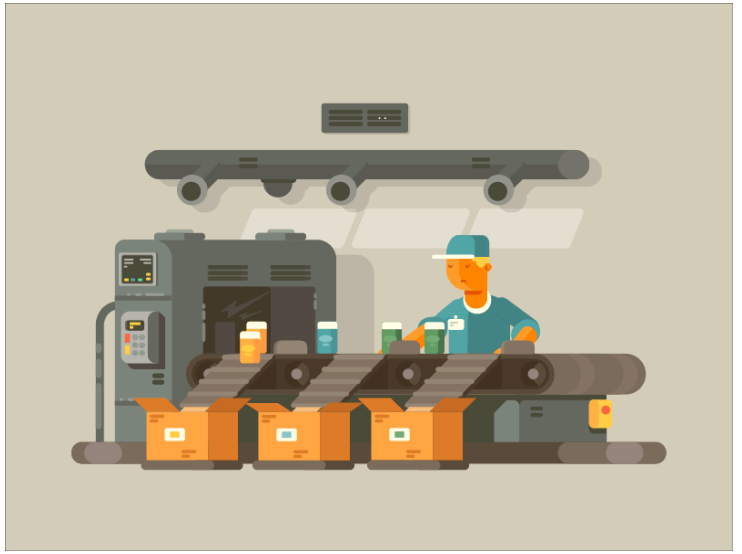
Machine Learning performs really well in specific tasks

- While there are some AI tools (like ChatGPT) whose capabilities are quite broad and general, ML-based technologies solves well specific tasks
- And ML-based technologies are everywhere: we have countless ML-based tools for specific tasks



Machine Learning performs really well in specific tasks… this is why is pervasive!



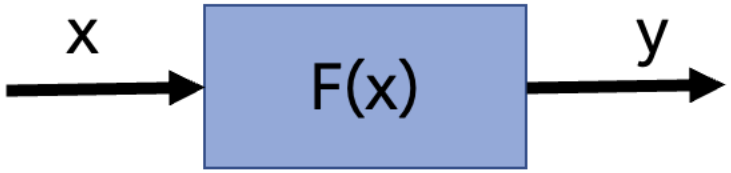


(‘Classic’) Machine Learning by itself is not enough

- ‘Classic’ = not RL-based (think about supervised learning for example)
- For example, ML algorithms may sense the environment in a self-driving car, but you need control and planning
- RL can be seen as the intersection of control and planning (more on this later)

Machine Learning without a supervisor

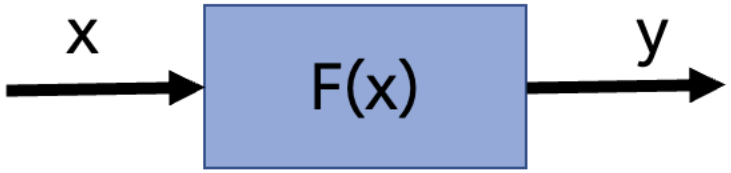
- ML solutions typically require a ‘supervisor’ that provides meaningful data of the phenomenon we want to characterize



Supervised learning settings: we have historical data of (x,y) and we look for a relationship F

Machine Learning without a supervisor

- ML solutions typically require a ‘supervisor’ that provides meaningful data of the phenomenon we want to characterize

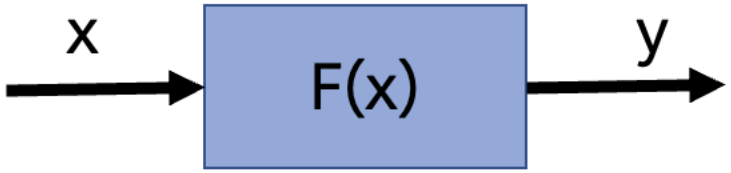


Supervised learning settings: we have historical data of (x,y) and we look for a relationship F

The Matrix (1999)

Machine Learning without a supervisor

- ML solutions typically require a ‘supervisor’ that provides meaningful data of the phenomenon we want to characterize

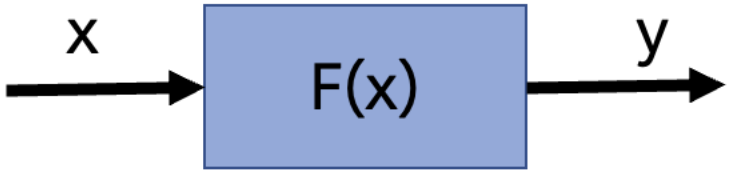


Supervised learning settings: we have historical data of (x,y) and we look for a relationship F

The Matrix (1999)

Machine Learning without a supervisor

- ML solutions typically require a ‘supervisor’ that provides meaningful data of the phenomenon we want to characterize
- This is not always feasible in many scenarios
- RL aims at solving the case where data must be collected ’in the wild’: the agent must interact with the world, possibly without having ‘precise’ instructions and without knowing a thing about the world, and learn how to maximize some objectives…



Supervised learning settings: we have historical data of (x,y) and we look for a relationship F

Machine Learning without a supervisor

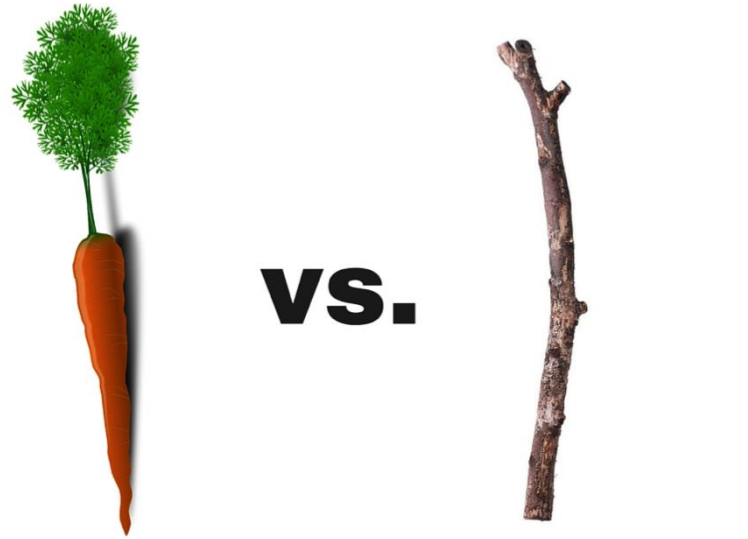
- ML solutions typically require a ‘supervisor’ that provides meaningful data of the phenomenon we want to characterize
- This is not always feasible in many scenarios
- RL aims at solving the case where data must be collected ’in the wild’: the agent must interact with the world, possibly without having ‘precise’ instructions and without knowing a thing about the world, and learn how to maximize some objectives…

'Limits’ of Supervised Learning

- Data are not available without interacting with the world

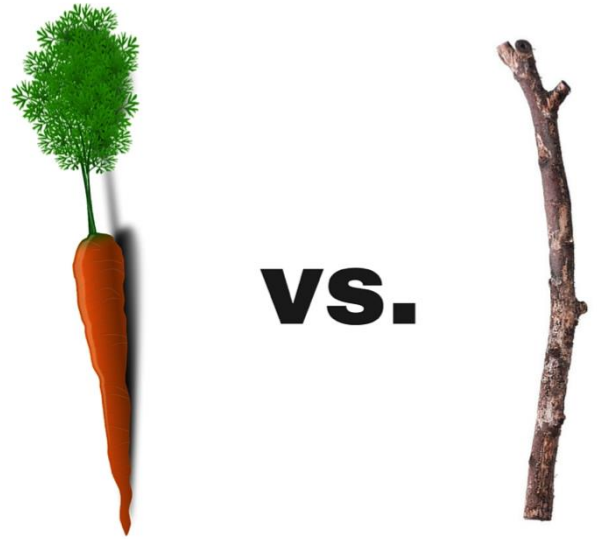
'Limits' of Supervised Learning

- Data are not available without interacting with the world
- Collected data may not be explicitly associated with actions: a lot of scenarios we get 'partial' information of how good a 'task' was executed, in the form of rewards (more on this later)
- In many cases, time matters: we don't always have



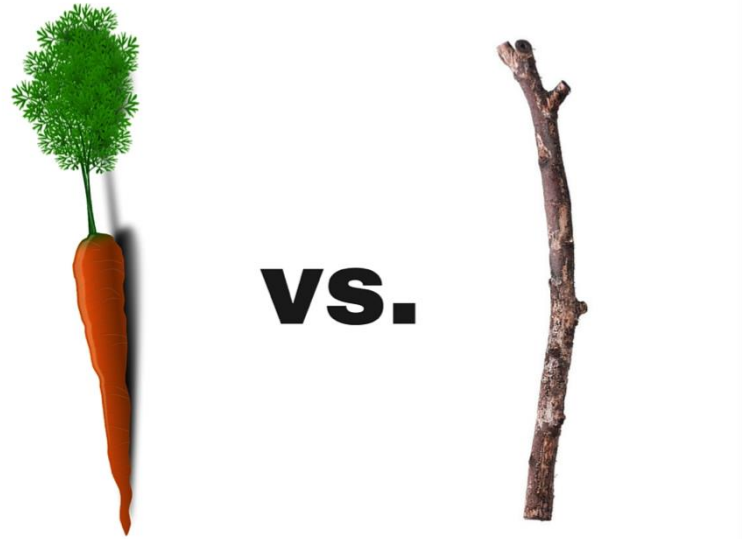
'Limits' of Supervised Learning

- Data are not available without interacting with the world
- Collected data may not be explicitly associated with actions: a lot of scenarios we get 'partial' information of how good a 'task' was executed, in the form of rewards (more on this later)
- In many cases, time matters: we don't always have independent identical distributed (i.i.d.) data (where rows can be swapped)
- In many settings, supervised learning do not envision



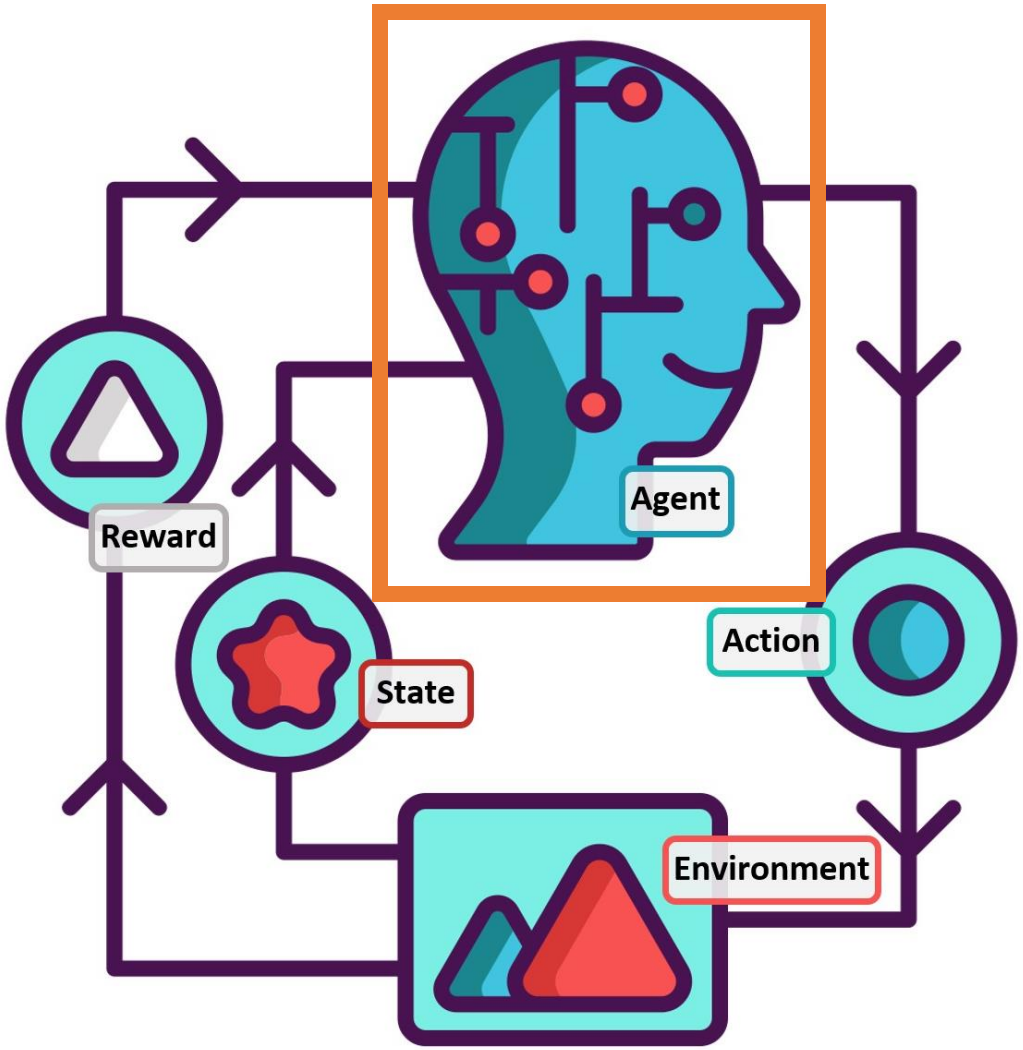
'Limits' of Supervised Learning

- Data are not available without interacting with the world
- Collected data may not be explicitly associated with actions: a lot of scenarios we get 'partial' information of how good a 'task' was executed, in the form of rewards (more on this later)
- In many cases, time matters: we don't always have independent identical distributed (i.i.d.) data (where rows can be swapped)
- In many settings, supervised learning do not envision long-term scenarios



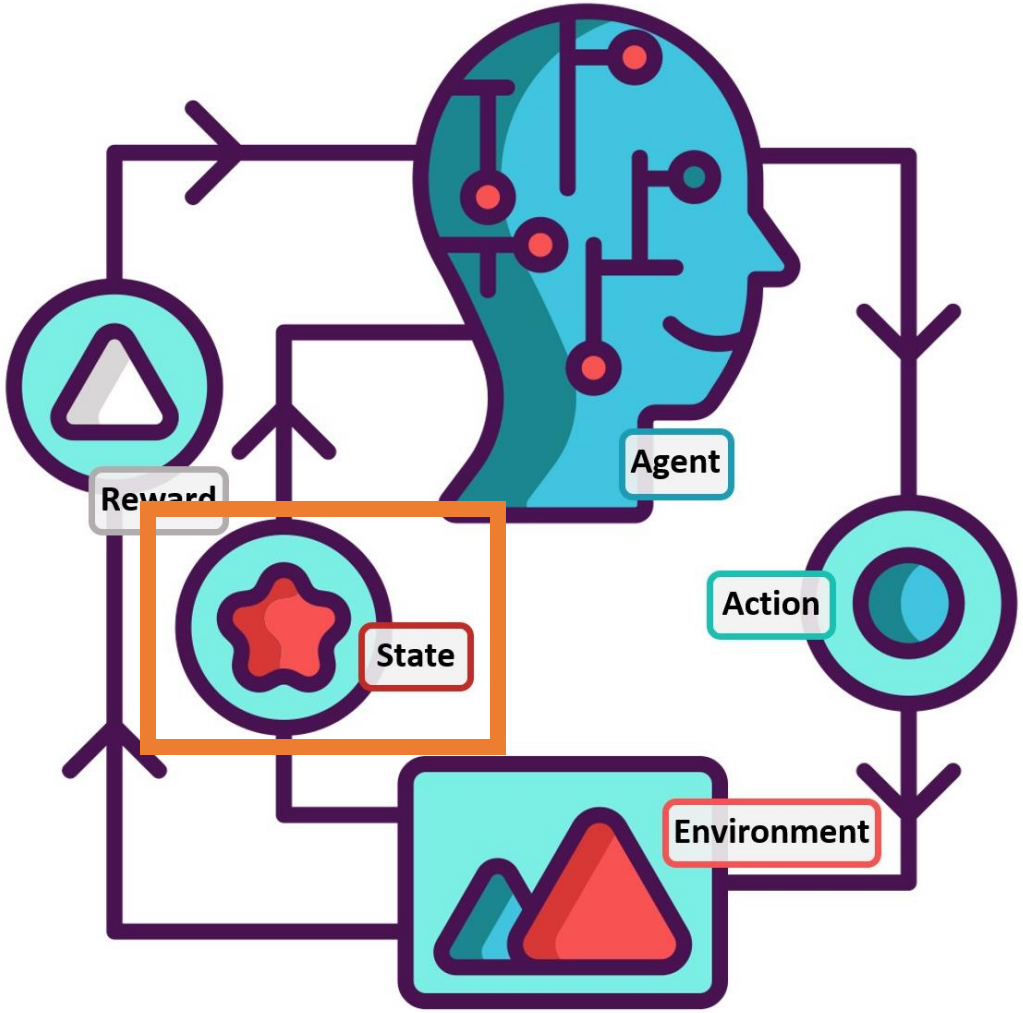
RL Formalism: The elements…

- An agent: the entity aiming at ‘solving a task’



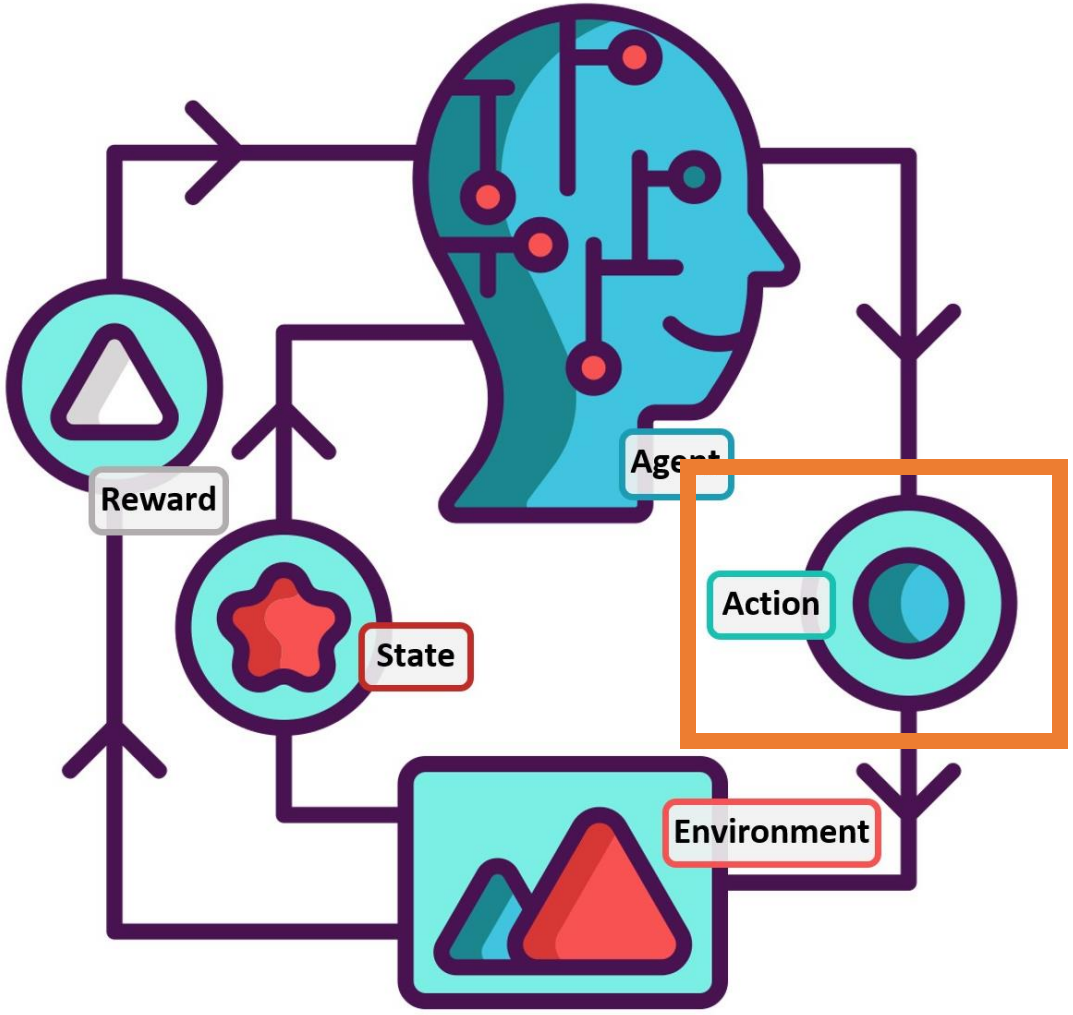
RL Formalism: The elements…

- An agent: the entity aiming at ‘solving a task’
- A set of states in which the agent can be



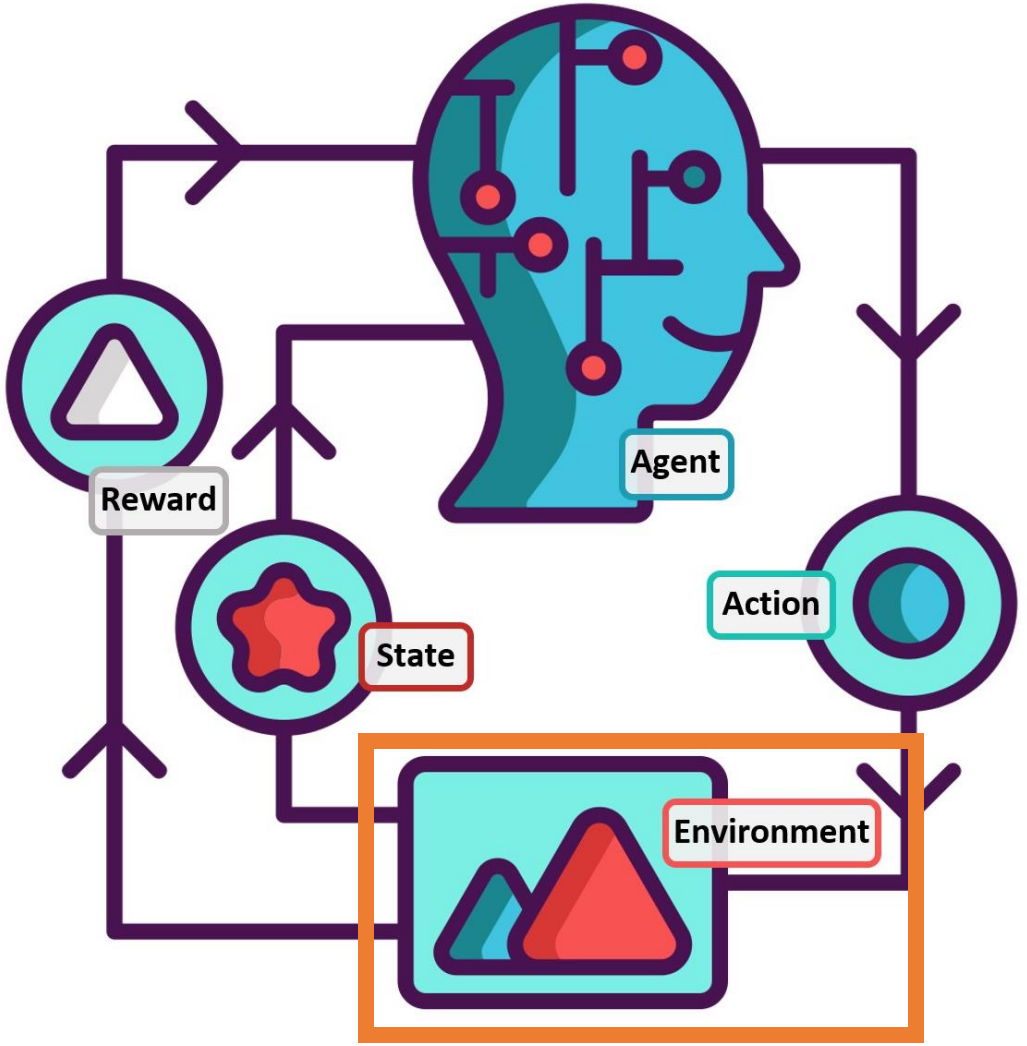
RL Formalism: The elements…

- An agent: the entity aiming at ‘solving a task’
- A set of states in which the agent can be
- A set of actions (that could depend on the state) that can be taken by the agent



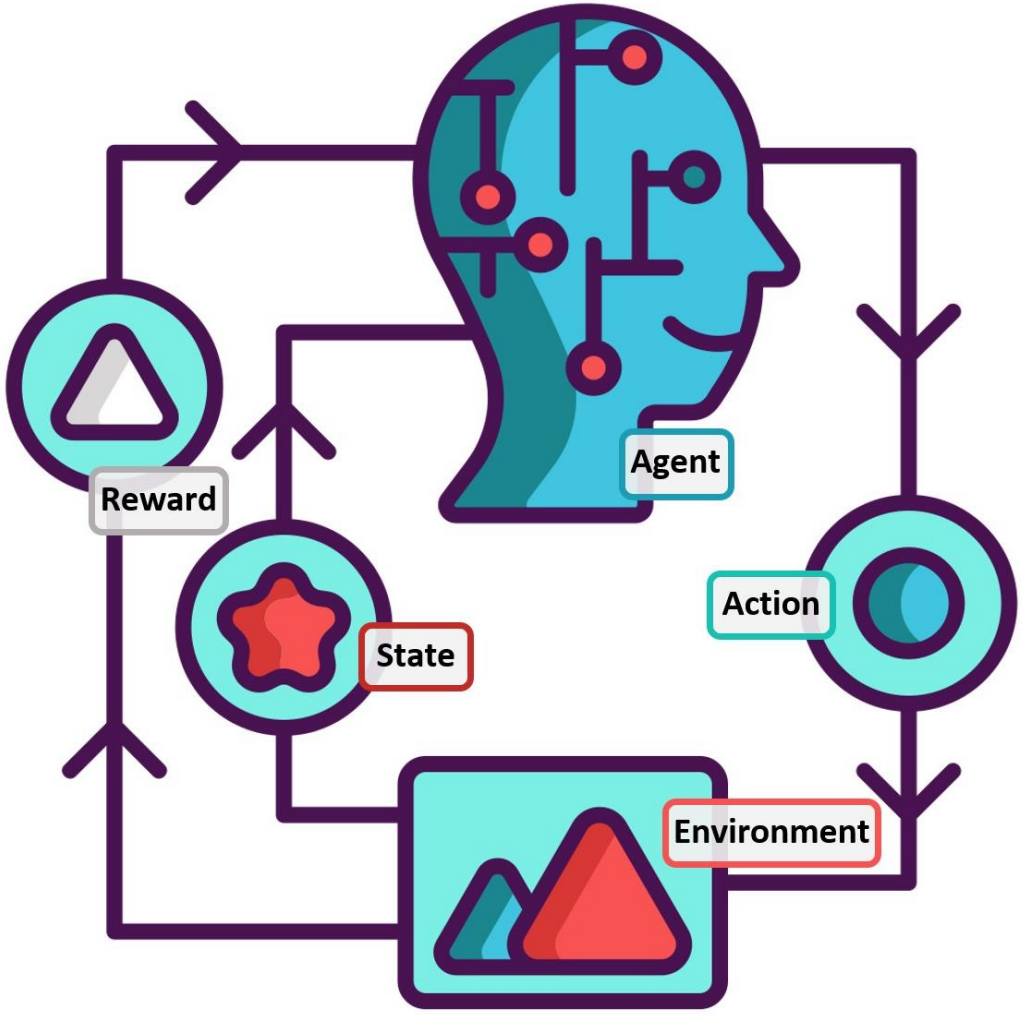
RL Formalism: The elements…

- An agent: the entity aiming at ‘solving a task’
- A set of states in which the agent can be
- A set of actions (that could depend on the state) that can be taken by the agent
- An environment with which the agent interacts and that could provide, at each time, rewards for state/state-action pair

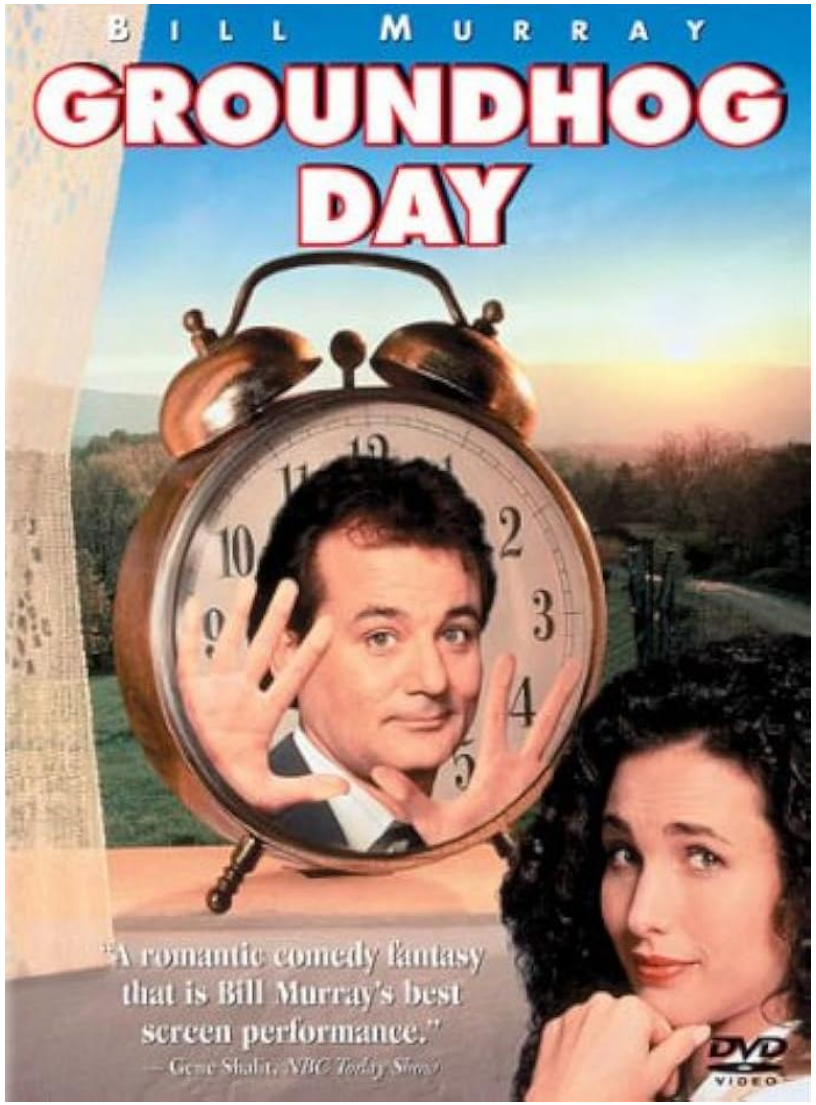


RL Formalism: The elements…

- An agent: the entity aiming at ‘solving a task’
- A set of states in which the agent can be
- A set of actions (that could depend on the state) that can be taken by the agent
- An environment with which the agent interacts and that could provide, at each time, rewards for state/state-action pair
- An agent could be more interested in long-term rewards, more than short-term…



… in movies terms:



Groundhog Day (1993)

… in movies terms:

- In Groundhog Day, Phil (played by Bill Murray) enters a time-loop and revive the same day over and over again. During these ‘days’ he tries to make Rita (played by Andie MacDowell) to fall in love with him
- Phil is the agent: he plays many ‘episodes’

Groundhog Day (1993)

… in movies terms:

- In Groundhog Day, Phil (played by Bill Murray) enters a time-loop and revive the same day over and over again. During these ‘days’ he tries to make Rita (played by Andie MacDowell) to fall in love with him
- Phil is the agent: he plays many ‘episodes’ from which he can learn
- He tries different actions (be nice, be funny, …) from different states (at the restaurant, the park, …)
- The environment is composed by Lisa, the town, the people in the town…
- During his experience Phil collects data and

Groundhog Day (1993)

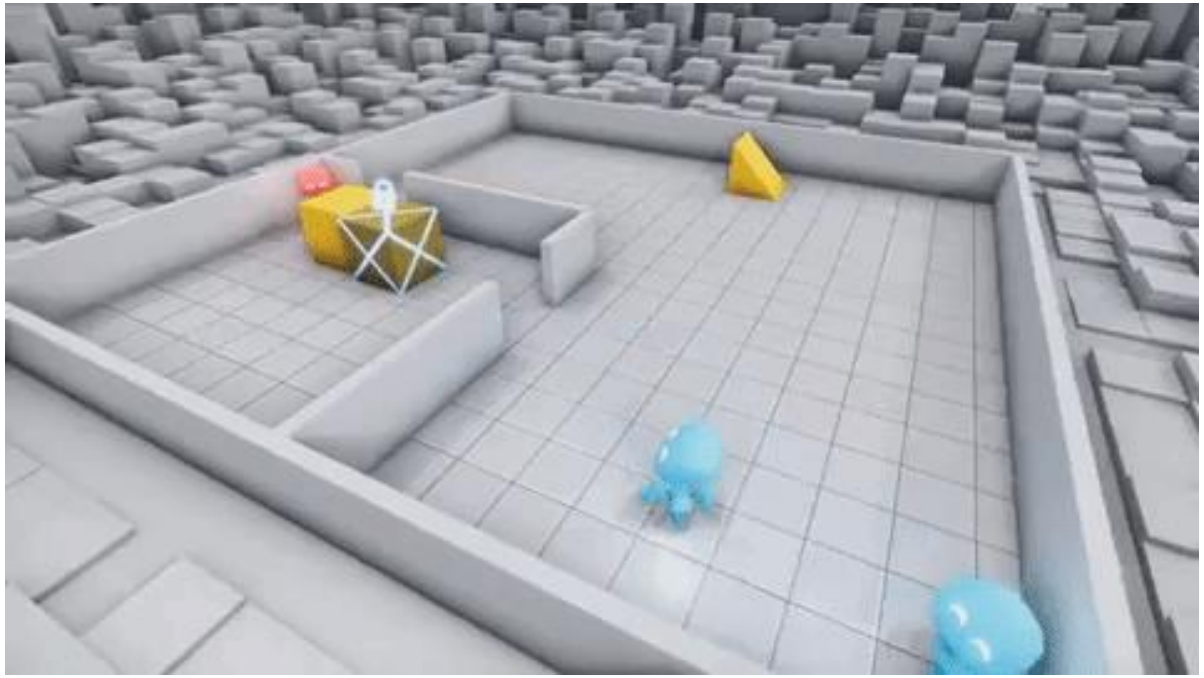
… in movies terms:

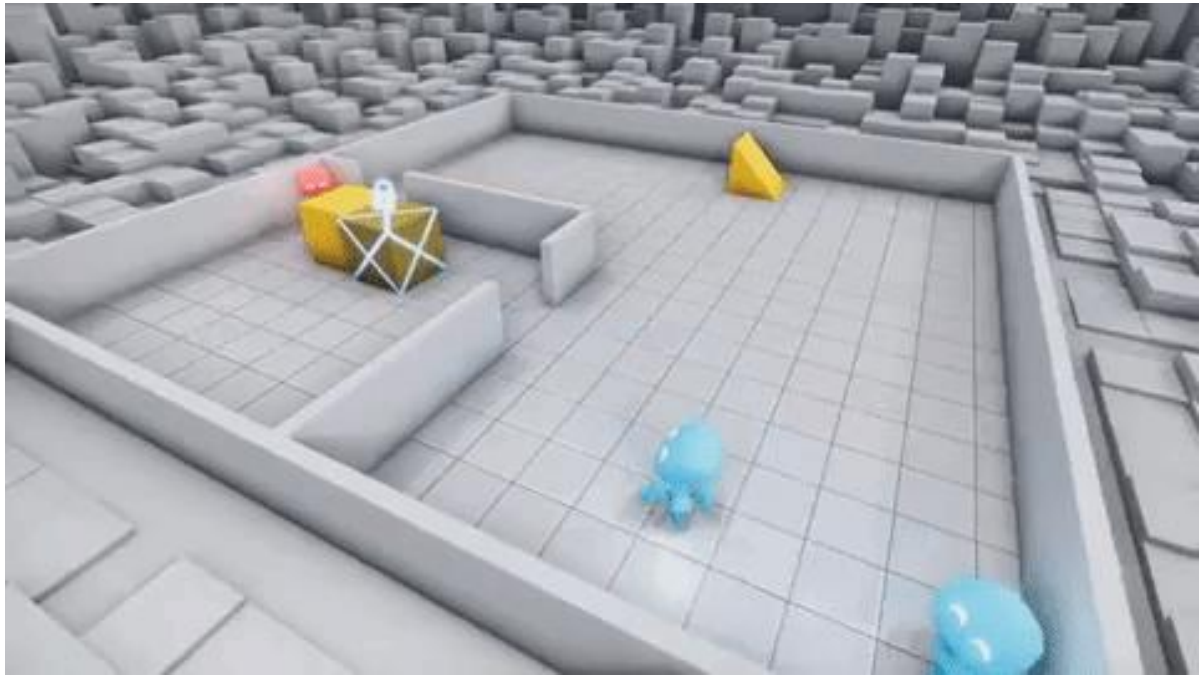
- In Groundhog Day, Phil (played by Bill Murray) enters a time-loop and revive the same day over and over again. During these ‘days’ he tries to make Rita (played by Andie MacDowell) to fall in love with him
- Phil is the agent: he plays many ‘episodes’ from which he can learn
- He tries different actions (be nice, be funny, …) from different states (at the restaurant, the park, …)
- The environment is composed by Lisa, the town, the people in the town…
- During his experience Phil collects data and learn how to maximize rewards (’love’ from Rita) from the environment

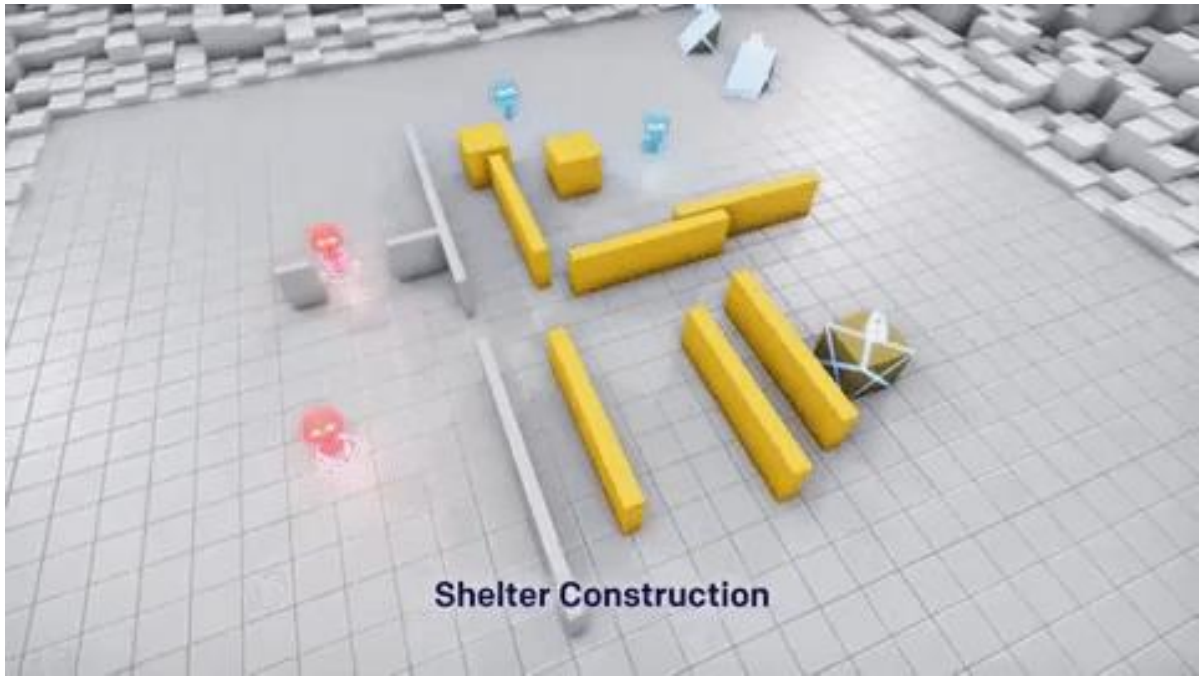
Groundhog Day (1993)

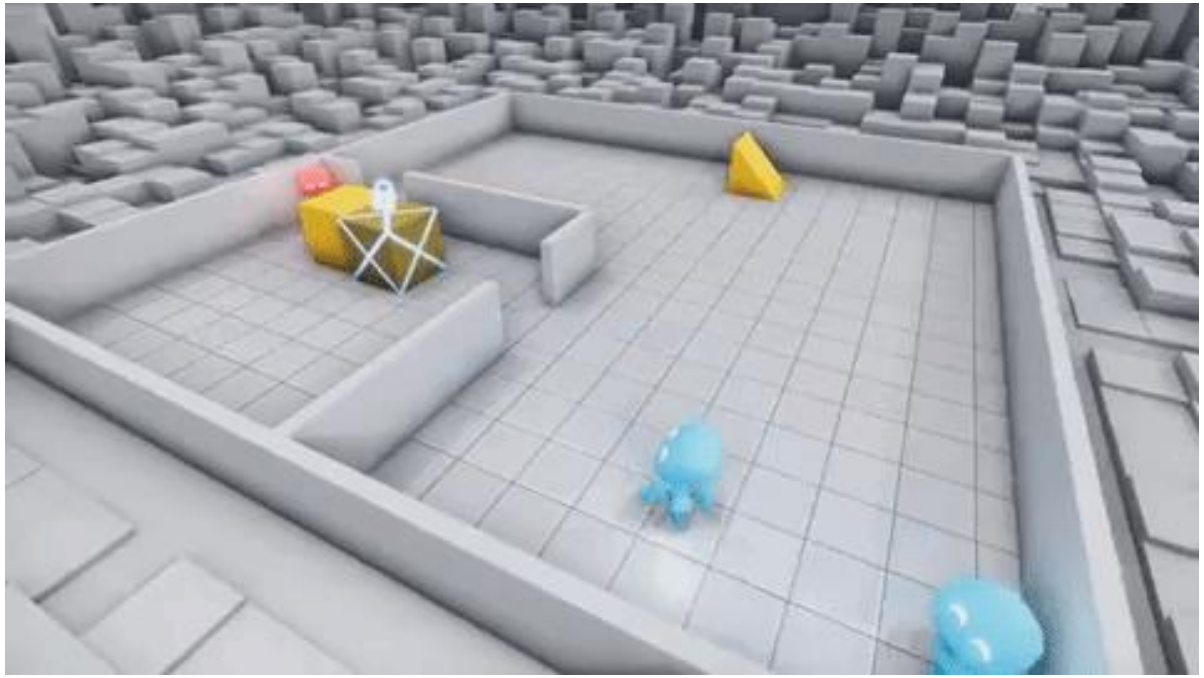
Life is a continuous collection of experience and, hopefully, learning

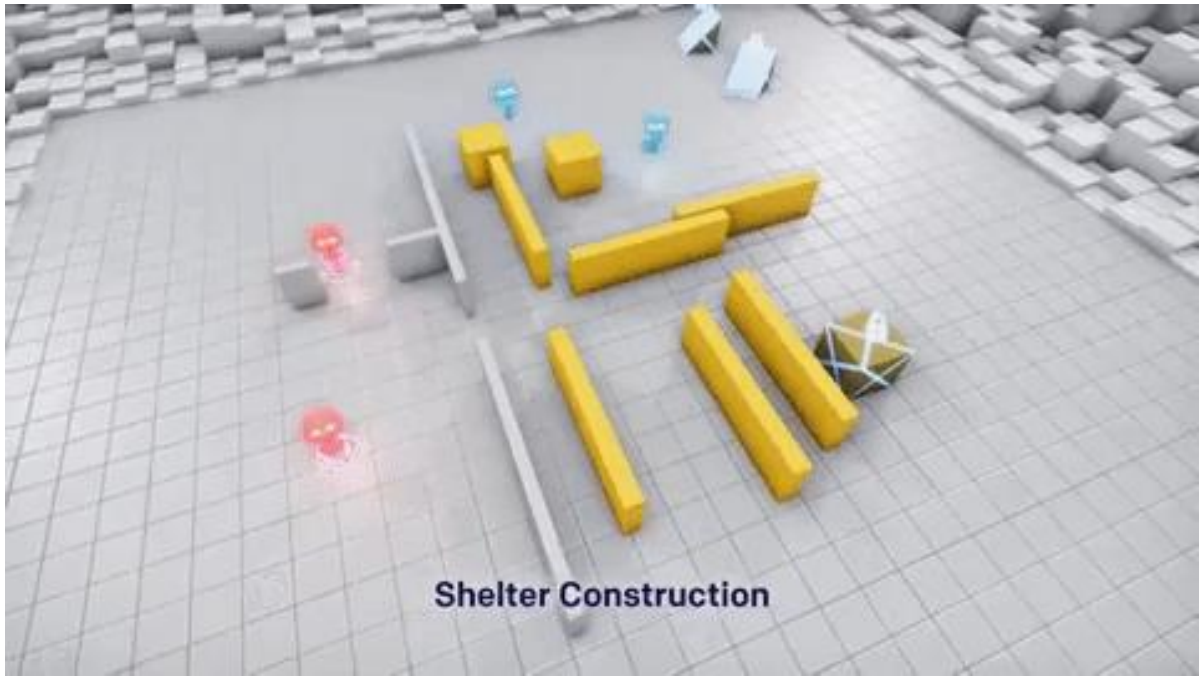
D. Silver 'Life is a continuous stream of data'

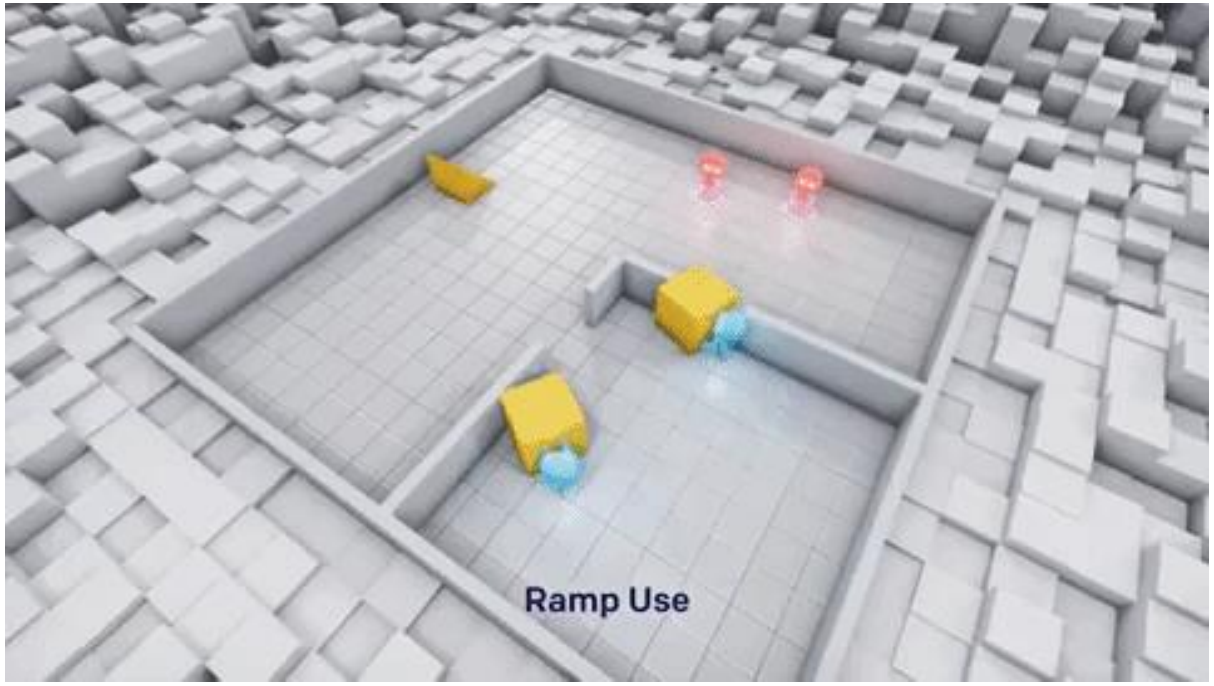


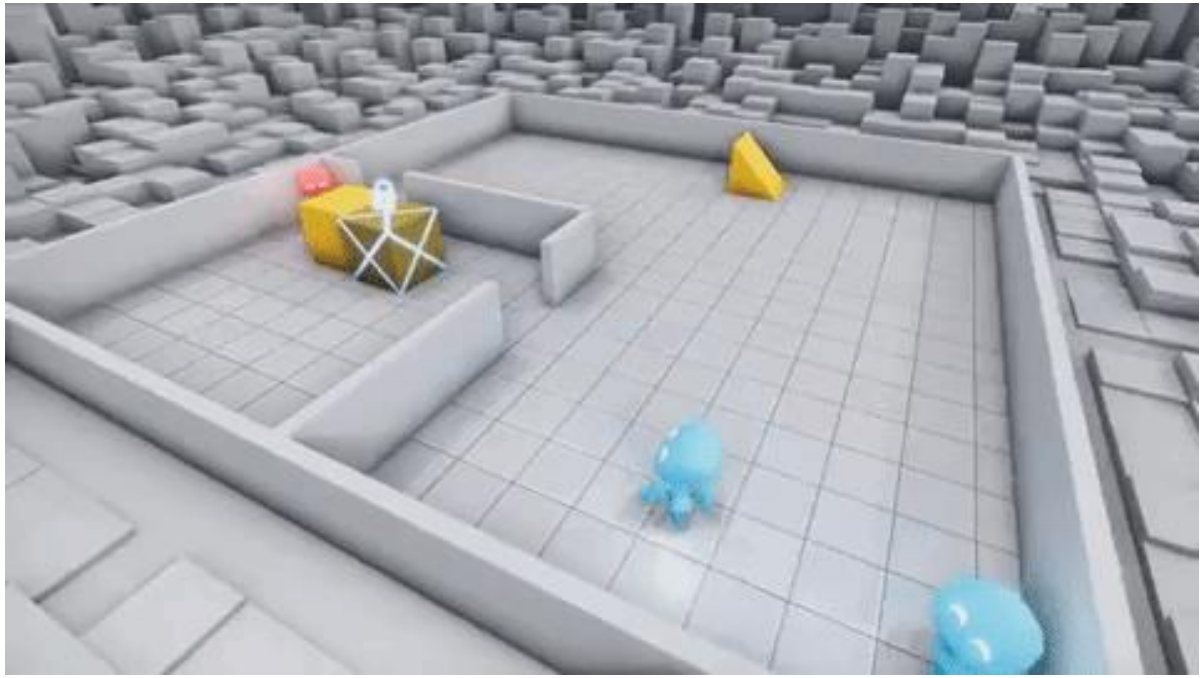


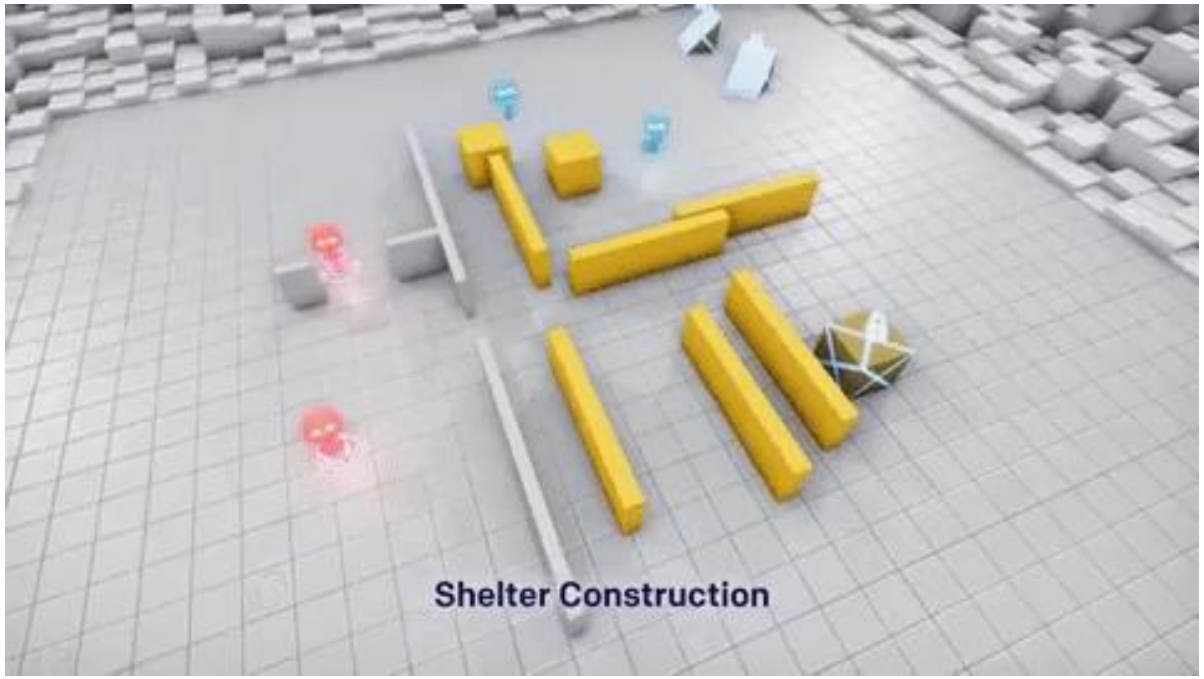

Shelter Construction

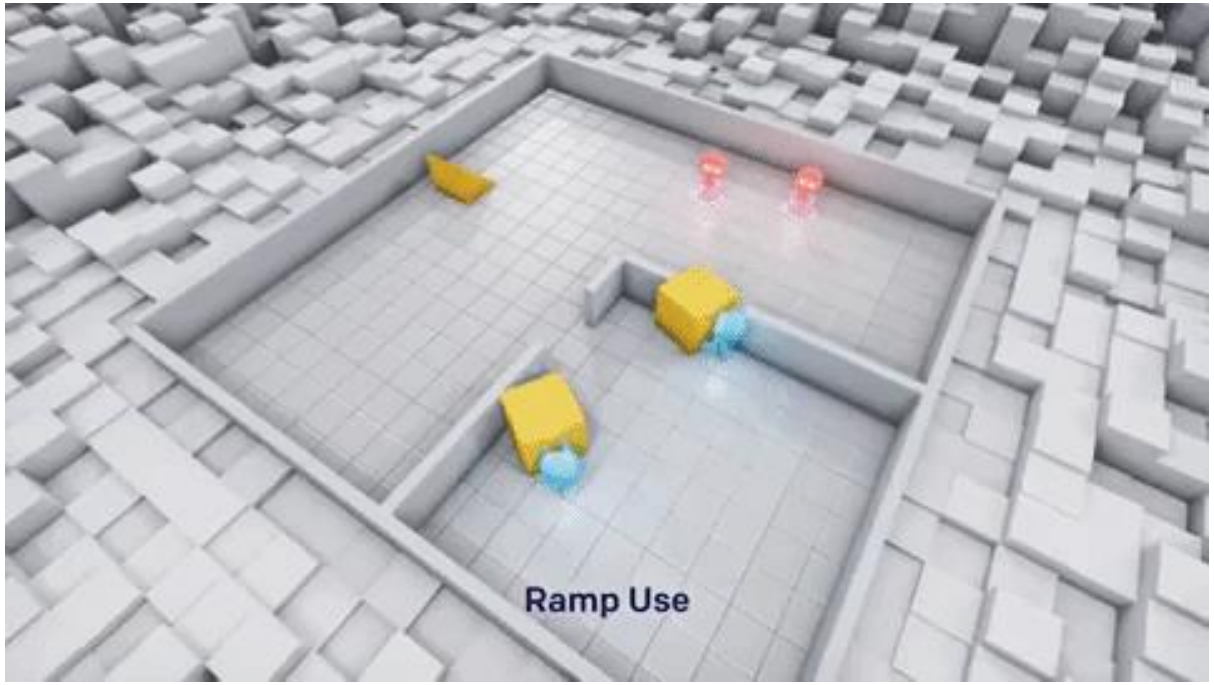


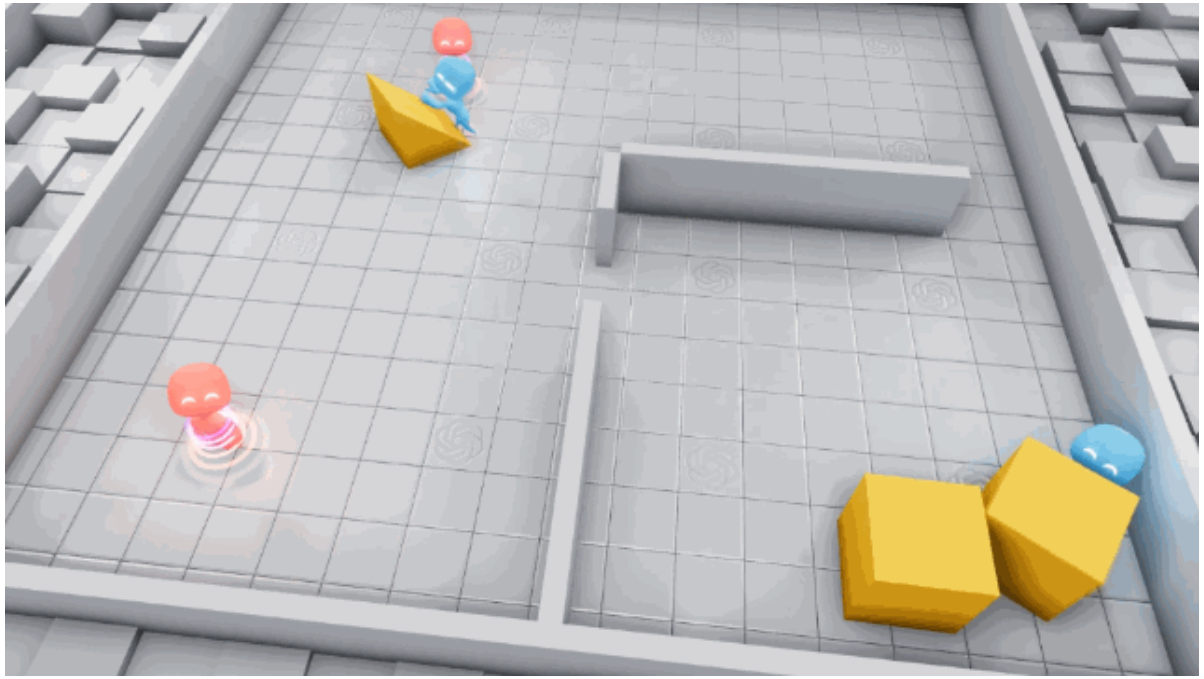

Shelter Construction


Ramp Use



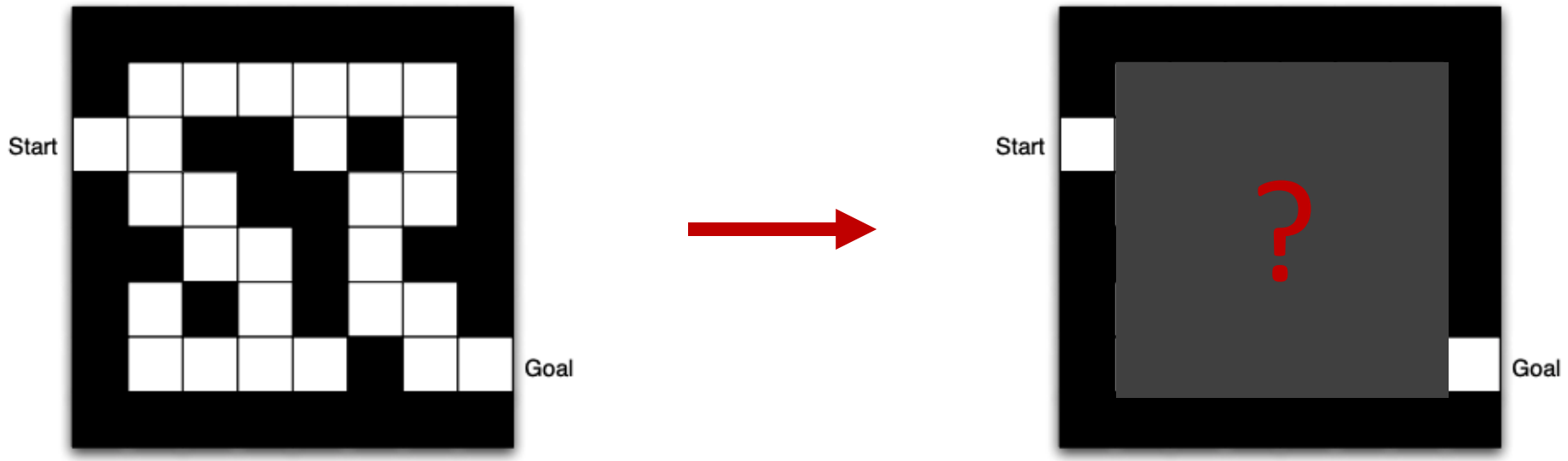

Shelter Construction


Ramp Use



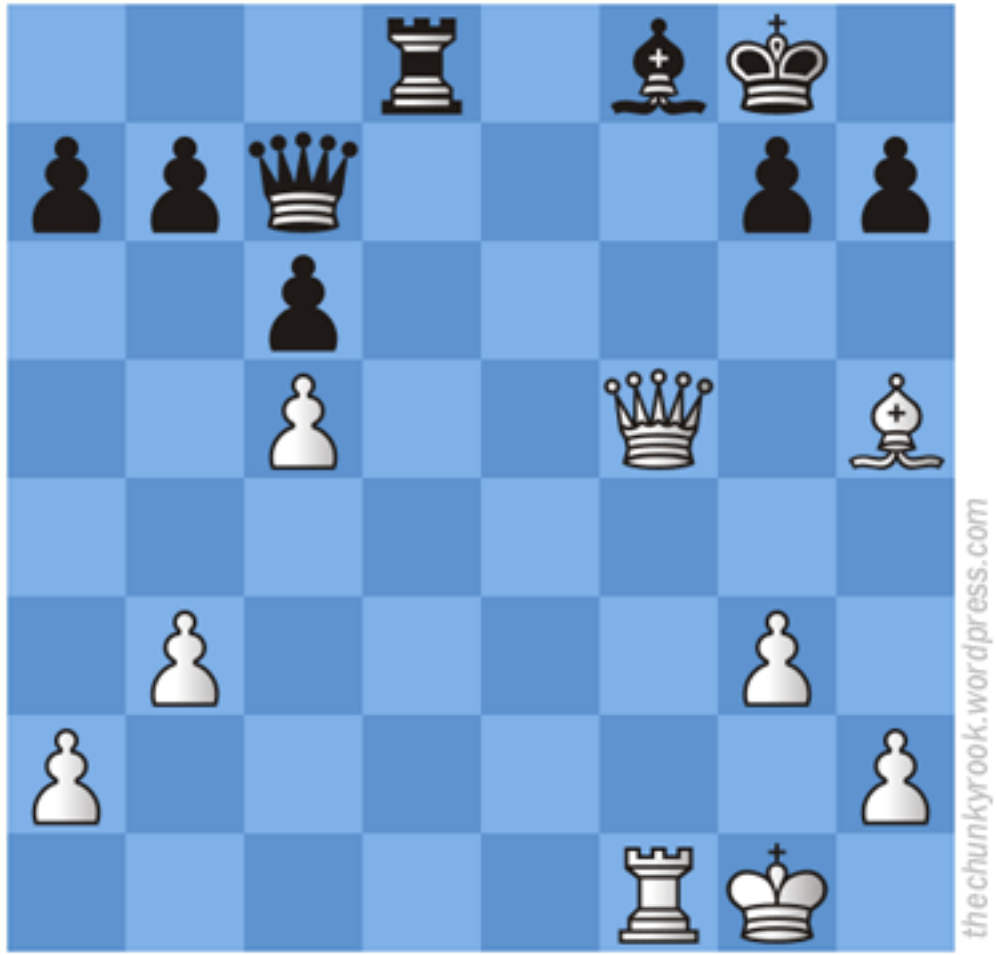
The RL Problem: data/information are gathered by Interacting with the Environment

- The agent could not know in advance a thing about the environment and which rewards and next state a given action from a state could lead to
- In the RL settings a model of the environment is not known!
- Such things will be learned by interacting with the environment!



The RL Problem: Planning and Long-term Rewards

- The goal of the agent is to maximize long term rewards, not necessarily at all steps
- Planning is fundamental in RL
- Ex. In chess the ‘true’ reward came when winning a game: at each steps of the game, event the losing of a piece can be ‘optimal’ in order to get a better state for reaching the target



Again, what is Reinforcement Learning (RL)?

Reinforcement Learning (RL) is both:

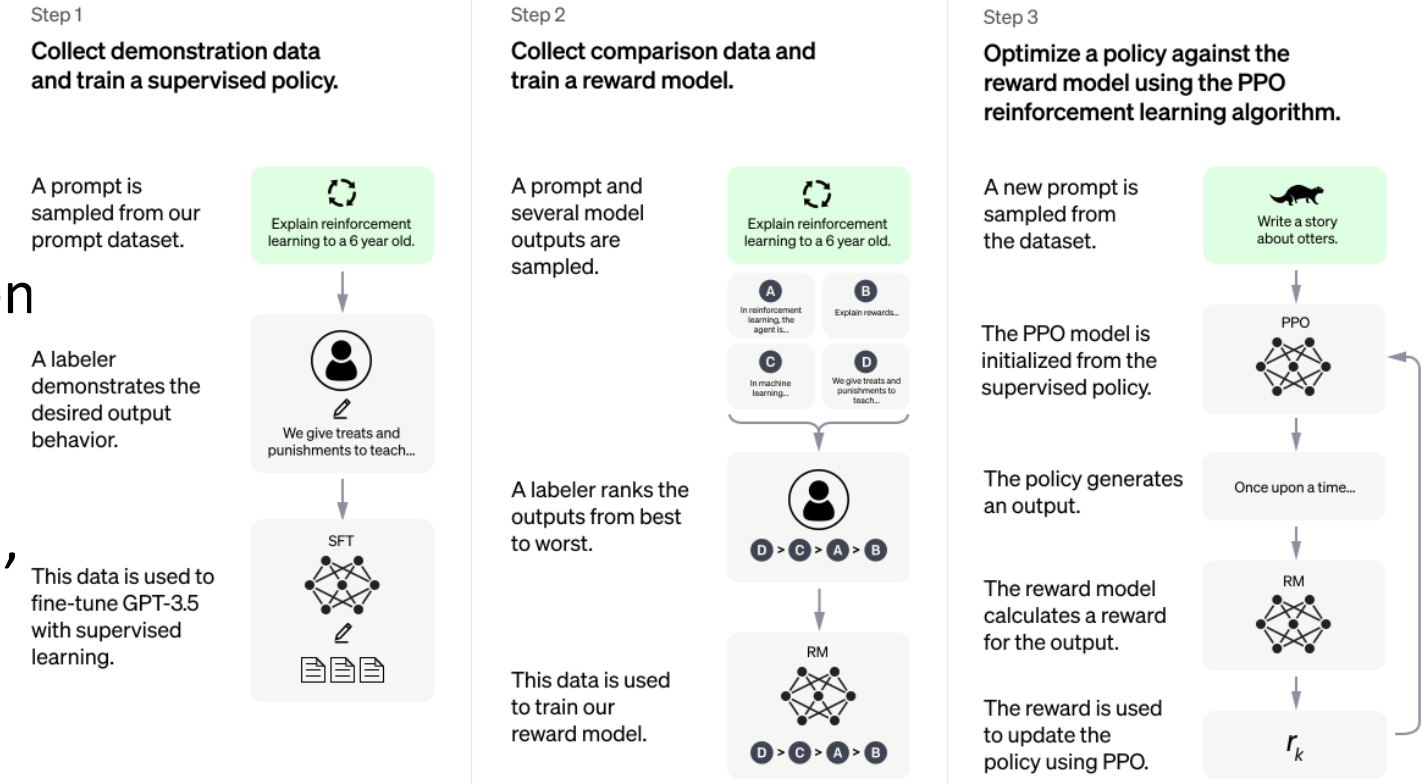
- a research area (sub-field of Machine Learning -> more in the following)
- a learning problem/paradigm concerned with learning to control a system (with many unknown elements) so as to maximize a numerical performance measure

Many real-world problems are better formalized in the RL fashion instead of the classical control or supervised learning formalization…

Some problems that can be formalized in a RL fashion

- Chatbots!
- Autonomous agents (self-driving cars, drones, robots…)
- Games
- HVAC (Heating, Ventilating, Air Conditioning) energy optimization
- Trading and Portfolio management
- Online advertising & Recommendation systems (news, items, …)
- Healthcare, biology

….



https://medium.com/aiguys/reinforcement-learning-from-human-feedback-instructgpt-and-chatgpt-693d00cb9c58

Some problems that can be formalized in a RL fashion

- Chatbots!
- Autonomous agents (self-driving cars, drones, robots…)
- Games
- HVAC (Heating, Ventilating, Air Conditioning) energy optimization
- Trading and Portfolio management
- Online advertising & Recommendation systems (news, items, …)
- Healthcare, biology

….

Aerospace Control Labs @ MIT (2015)

https://www.youtube.com/watch?v=opsmd5yuBF0

Some problems that can be formalized in a RL fashion

- Chatbots!
- Autonomous agents (self-driving cars, drones, robots…)
- Games
- HVAC (Heating, Ventilating, Air Conditioning) energy optimization
- Trading and Portfolio management
- Online advertising & Recommendation systems (news, items, …)
- Healthcare, biology

….

UZH Robotics and Perception Group @ ETH (2021)

https://www.youtube.com/watch?v=Hebpmadjqn8

Some problems that can be formalized in a RL fashion

- Chatbots!
- Autonomous agents (self-driving cars, drones, robots…)
- Games
- HVAC (Heating, Ventilating, Air Conditioning) energy optimization
- Trading and Portfolio management
- Online advertising & Recommendation systems (news, items, …)
- Healthcare, biology

….

Hybrid Robotics @ Berkeley (2021)

https://www.youtube.com/watch?v=6tn-owW6iME

https://arxiv.org/abs/2103.14295

Some problems that can be formalized in a RL fashion

- Chatbots!
- Autonomous agents (self-driving cars, drones, robots…)
- Games
- HVAC (Heating, Ventilating, Air Conditioning) energy optimization
- Trading and Portfolio management
- Online advertising & Recommendation systems (news, items, …)
- Healthcare, biology

….

DeepMind (2015)

https://www.youtube.com/watch?v=V1eYniJ0Rnk

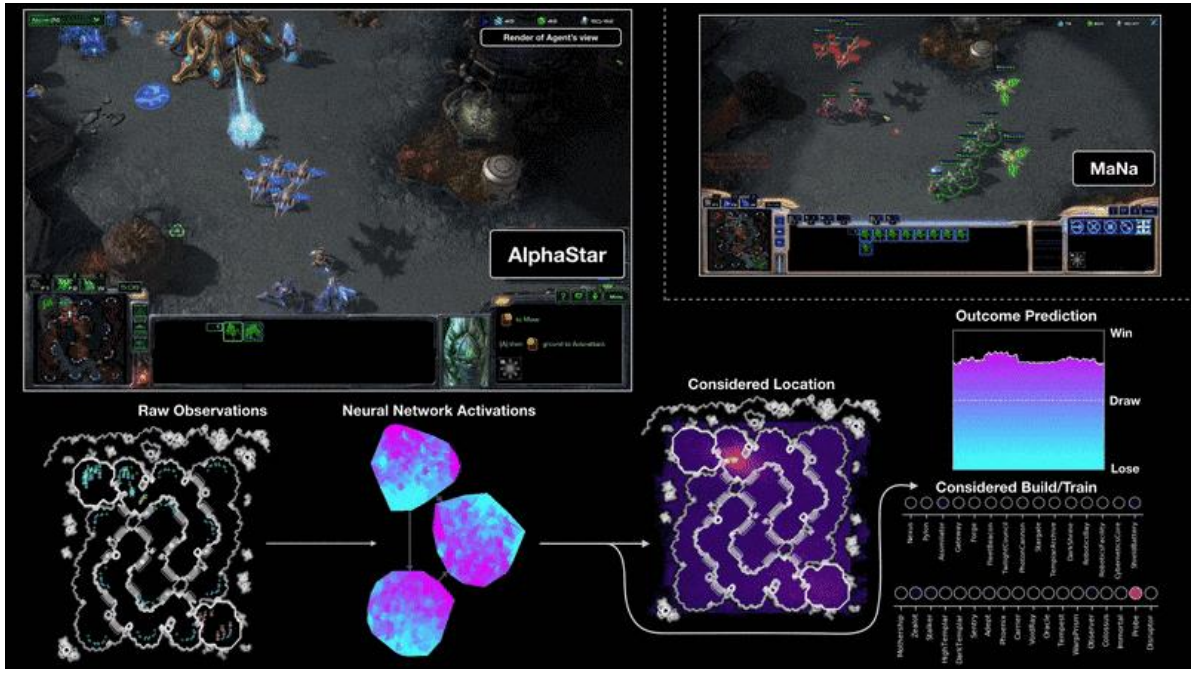
Human-level control through Deep Reinforcement Learning, Nature 518, 529-533 (2105)

https://www.nature.com/articles/nature14236?wm=book_wap_0005

Some problems that can be formalized in a RL fashion

- Chatbots!
- Autonomous agents (self-driving cars, drones, robots…)
- Games
- HVAC (Heating, Ventilating, Air Conditioning) energy optimization
- Trading and Portfolio management
- Online advertising & Recommendation systems (news, items, …)
- Healthcare, biology

….



DeepMind (2019)

https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

https://www.youtube.com/watch?v=jtlrWblOyP4

Some problems that can be formalized in a RL fashion

- Chatbots!
- Autonomous agents (self-driving cars, drones, robots…)
- Games
- HVAC (Heating, Ventilating, Air Conditioning) energy optimization
- Trading and Portfolio management
- Online advertising & Recommendation systems (news, items, …)
- Healthcare, biology

….

DeepMind (2016-2018)

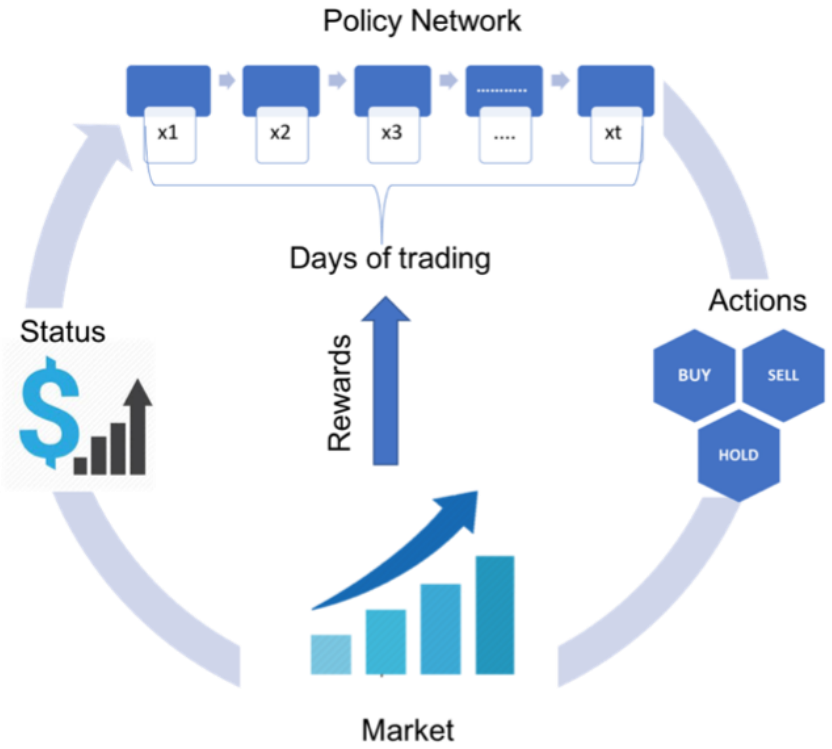
https://deepmind.com/blog/article/deepmind-ai-reduces-google-data-centre-cooling-bill-40

https://deepmind.com/blog/article/safety-first-ai-autonomous-data-centre-cooling-and-industrial-control

Some problems that can be formalized in a RL fashion

- Chatbots!
- Autonomous agents (self-driving cars, drones, robots…)
- Games
- HVAC (Heating, Ventilating, Air Conditioning) energy optimization
- Trading and Portfolio management
- Online advertising & Recommendation systems (news, items, …)
- Healthcare, biology

….



IBM (2018)

https://medium.com/ibm-data-ai/reinforcement-learning-the-business-use-case-part-2-c175740999

https://arxiv.org/abs/1706.10059

Some problems that can be formalized in a RL fashion

- Chatbots!
- Autonomous agents (self-driving cars, drones, robots…)
- Games
- HVAC (Heating, Ventilating, Air Conditioning) energy optimization
- Trading and Portfolio management
- Online advertising & Recommendation systems (news, items, …)
- Healthcare, biology

….

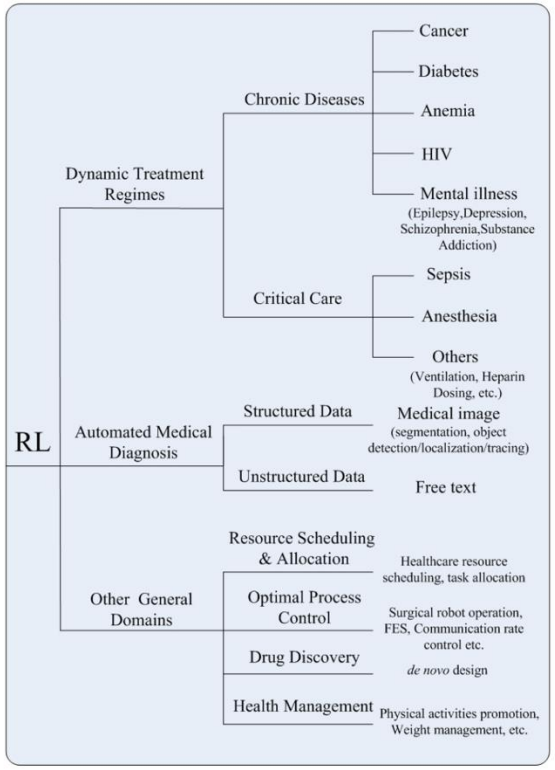
Michigan State University (2021)

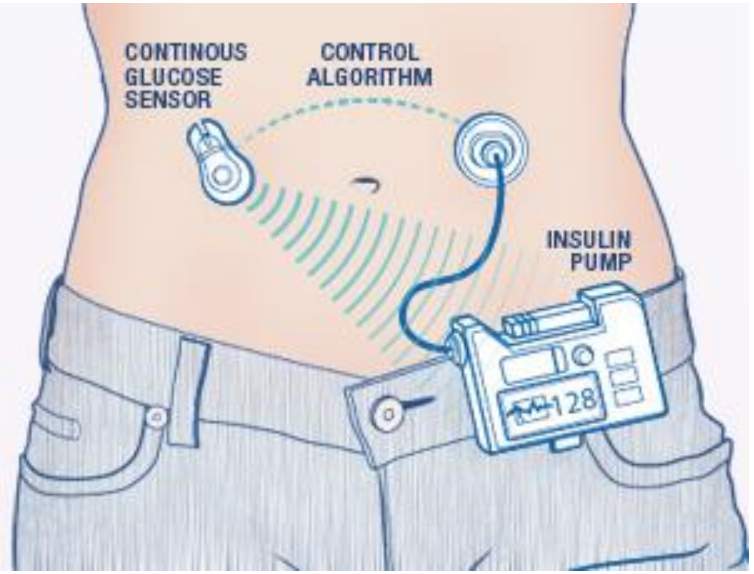
https://arxiv.org/abs/1909.03602

Some problems that can be formalized in a RL fashion

- Chatbots!
- Autonomous agents (self-driving cars, drones, robots…)
- Games
- HVAC (Heating, Ventilating, Air Conditioning) energy optimization
- Trading and Portfolio management
- Online advertising & Recommendation systems (news, items, …)
- Healthcare, biology

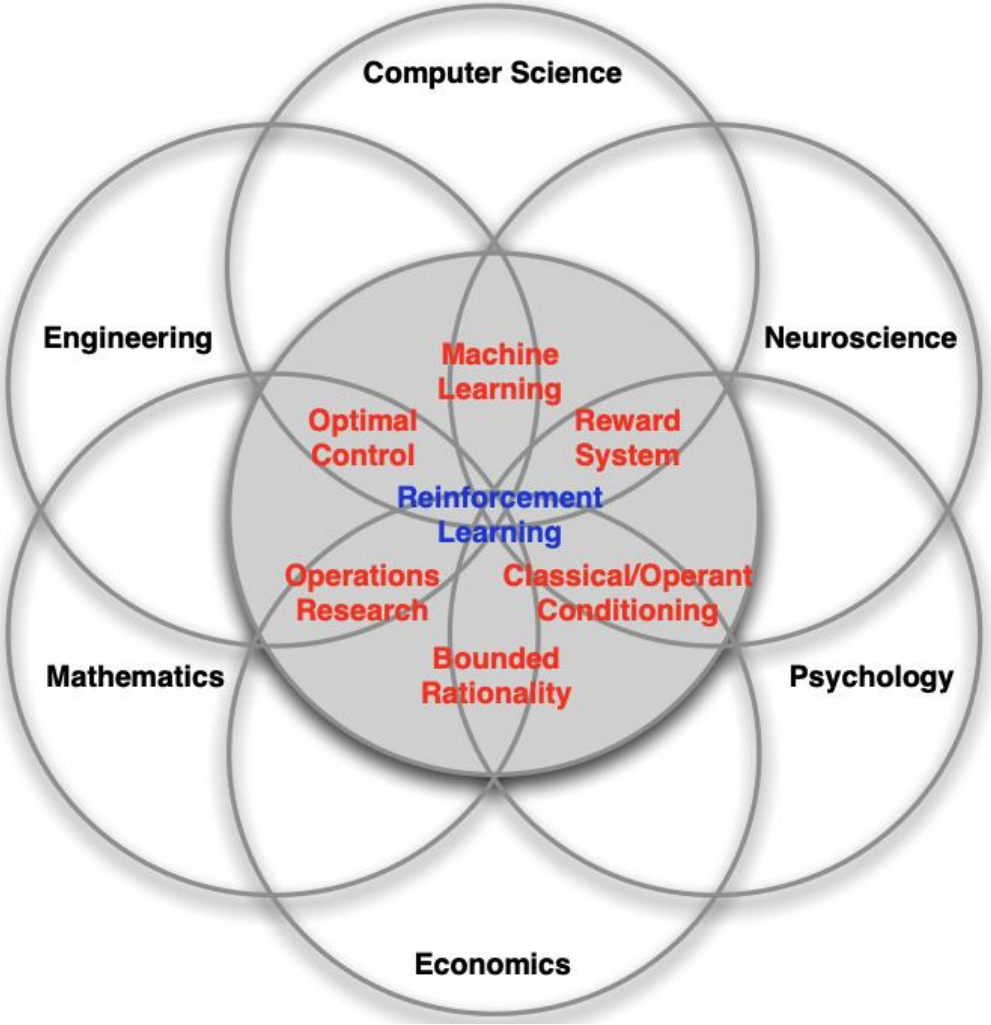
….

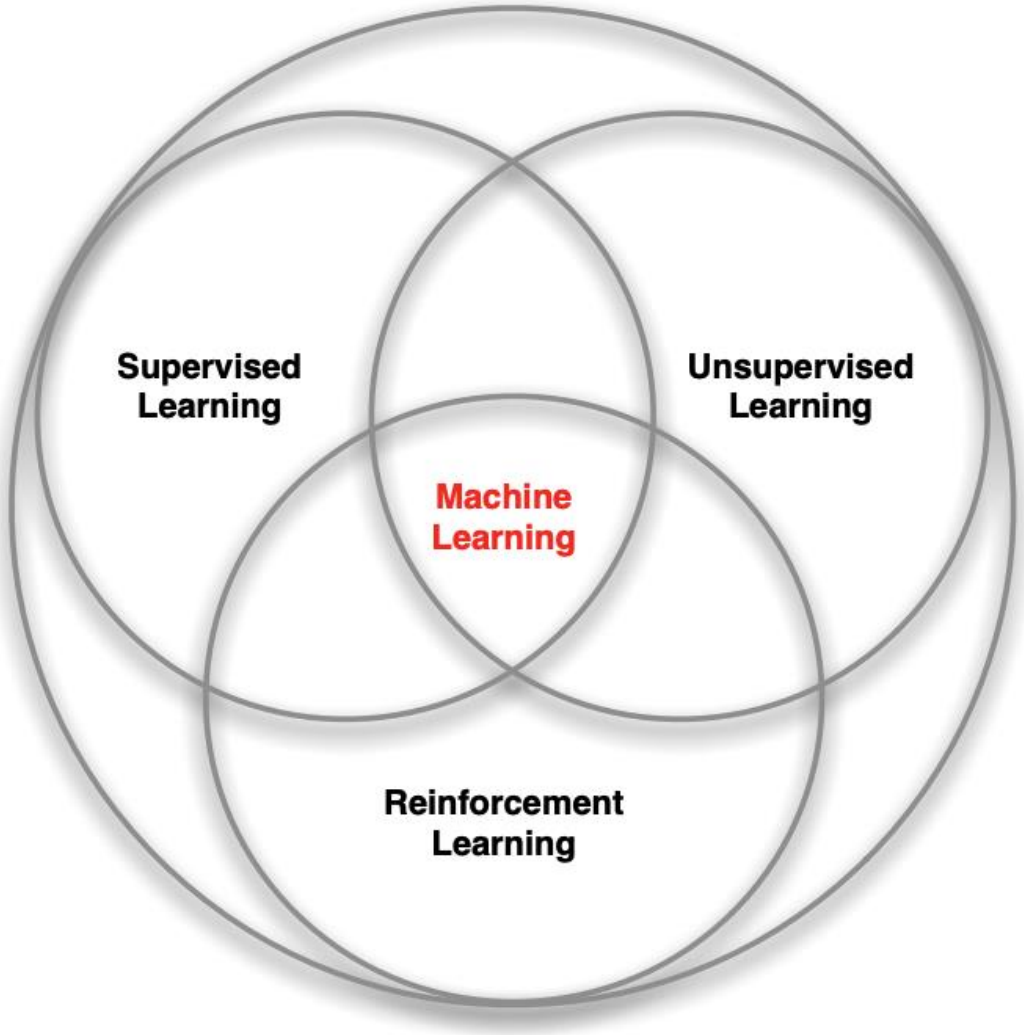




https://arxiv.org/pdf/1908.08796.pdf

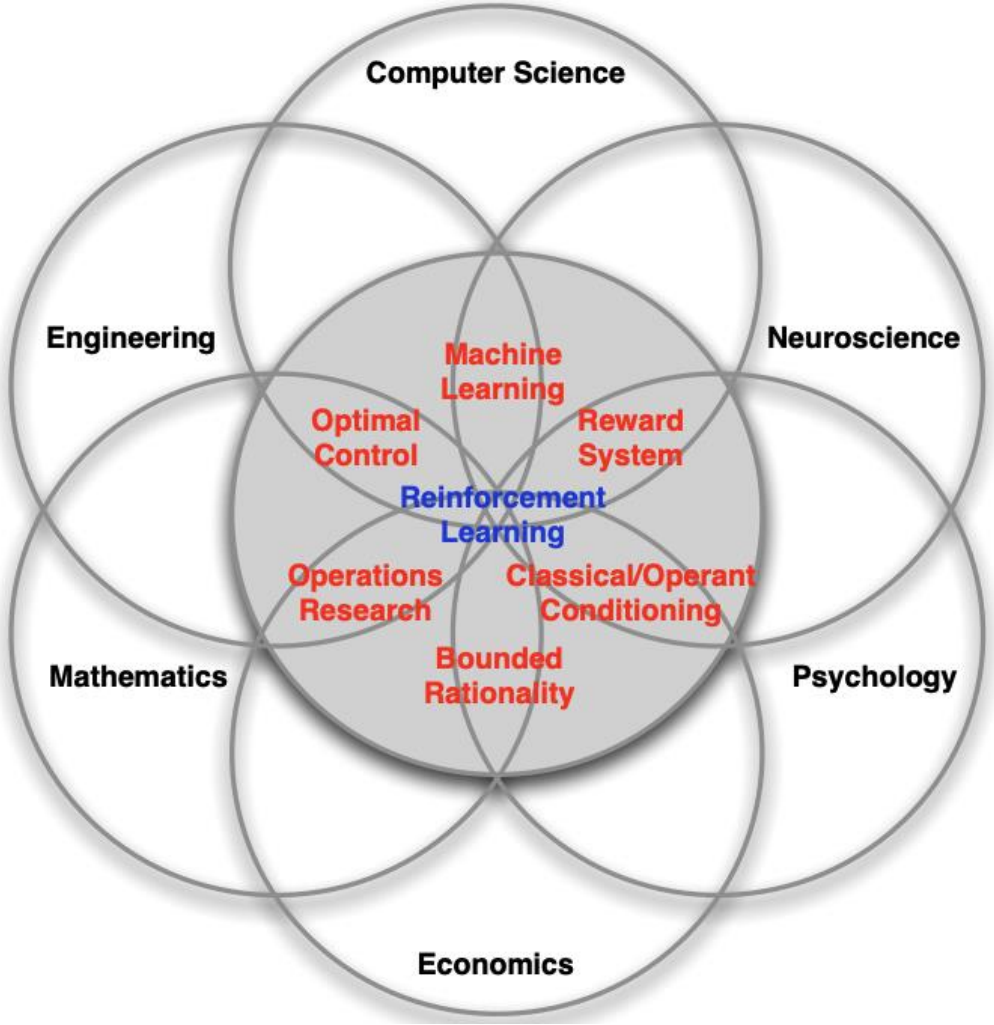
RL: an interdisciplinary field



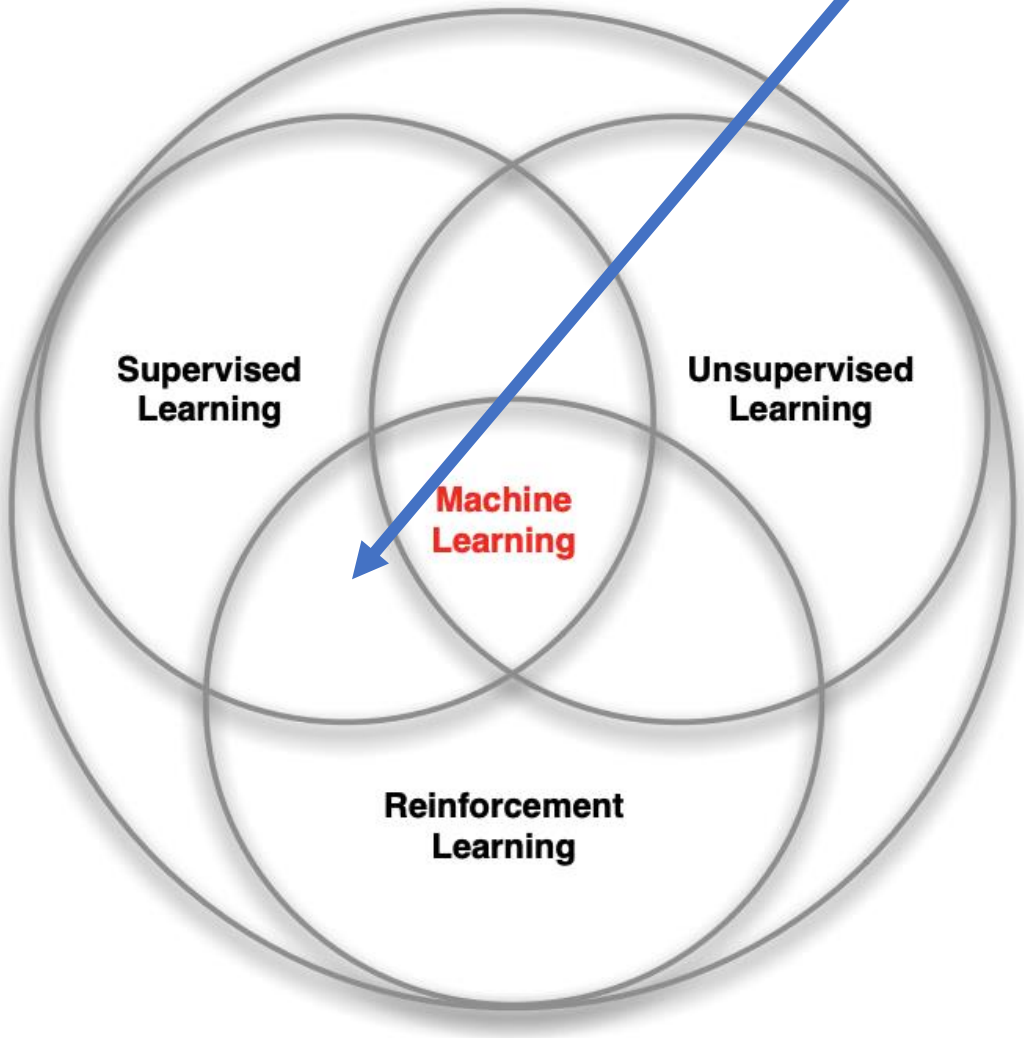


D. Silver 'RL course' @ UCL (2015)

RL: an interdisciplinary field



Supervised Learning approach are used in solving RL problems

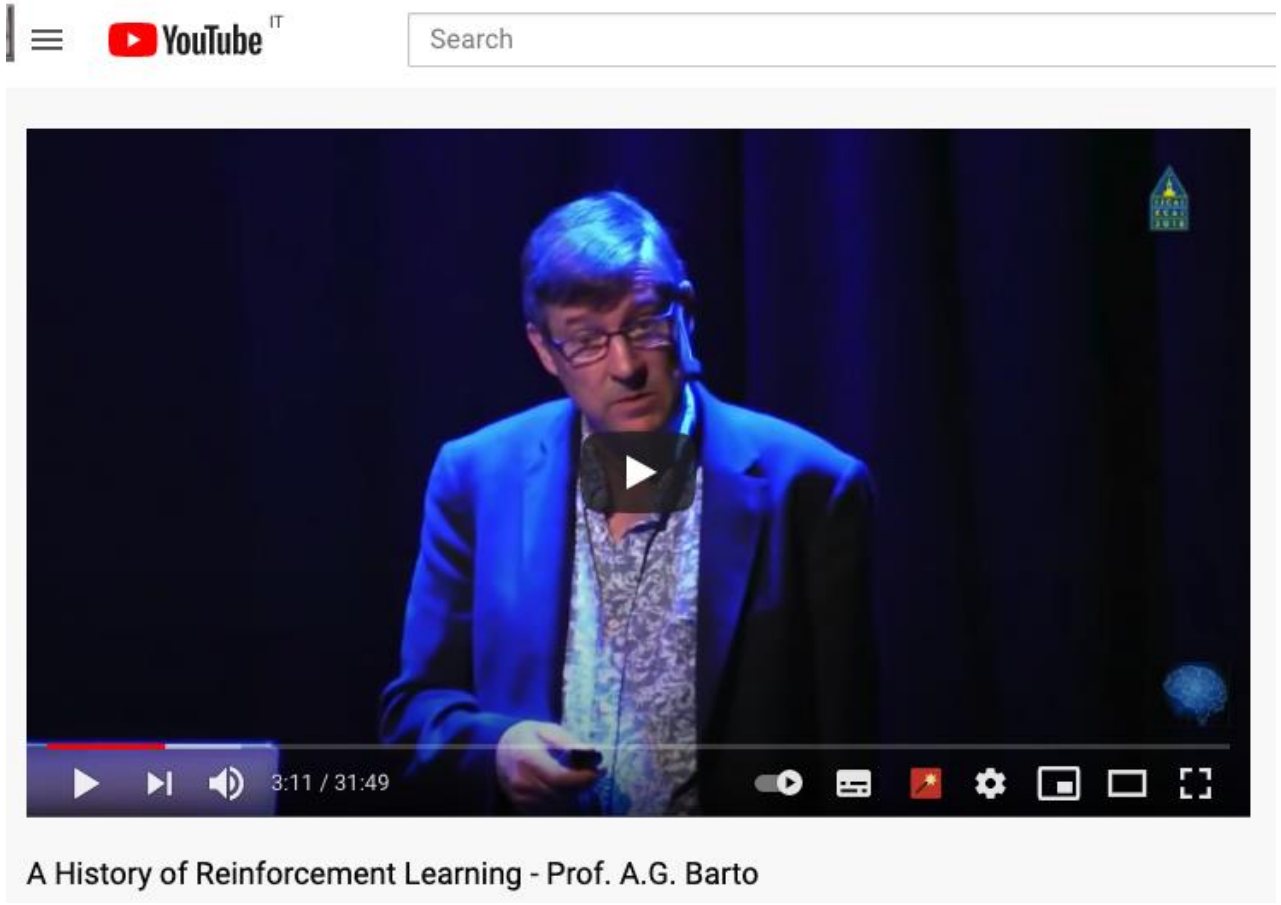


D. Silver 'RL course' @ UCL (2015)

RL Brief History

RL Book: Chapter 1

- Dynamic Programming (Bellman 1957)
- Markov Decision Processes (Bellman, Howard 1960)
- Temporal Difference (Witten 1977)
- TD Lambda (Sutton 1988)
- Q-Learning (Watkins 1989)
- TD-Gammon (Tesauro 1992)
- Deep Learning (Krizhevsky 2012)

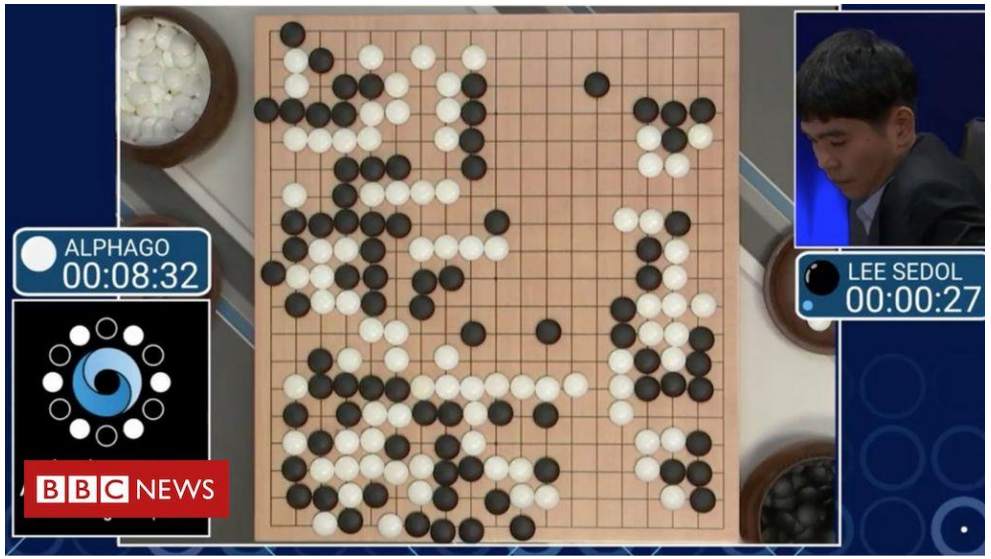


https://www.youtube.com/watch?v=ul6B2oFPNDM

RL Brief History

RL Book: Chapter 1

- Dynamic Programming (Bellman 1957)
- Markov Decision Processes (Bellman, Howard 1960)
- Temporal Difference (Witten 1977)
- TD Lambda (Sutton 1988)
- Q-Learning (Watkins 1989)
- TD-Gammon (Tesauro 1992)
- Deep Learning (Krizhevsky 2012)
- AlphaGo (DeepMind 2016)

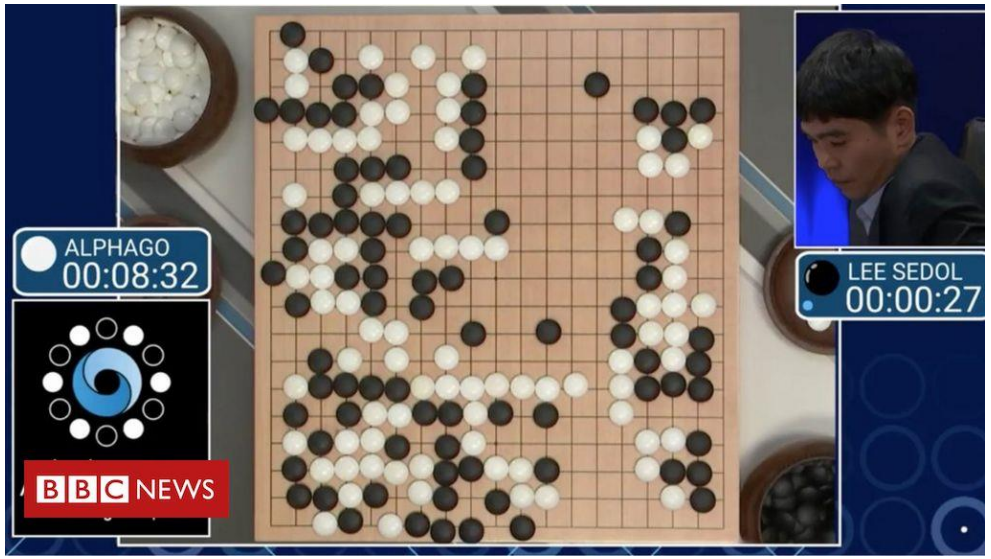


Game	Board size	State space	Game tree size
Go	19 x 19	10^{172}	10^{360}
Chess	8 x 8	10^{50}	10^{123}

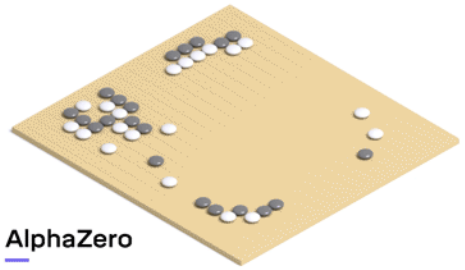
RL Brief History

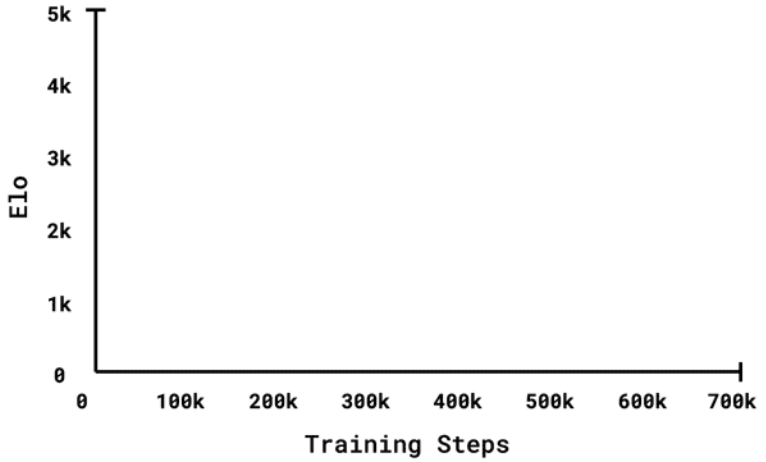
RL Book: Chapter 1

- Dynamic Programming (Bellman 1957)
- Markov Decision Processes (Bellman, Howard 1960)
- Temporal Difference (Witten 1977)
- TD Lambda (Sutton 1988)
- Q-Learning (Watkins 1989)
- TD-Gammon (Tesauro 1992)
- Deep Learning (Krizhevsky 2012)
- AlphaGo (DeepMind 2016)
- AlphaZero (DeepMind 2018)



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

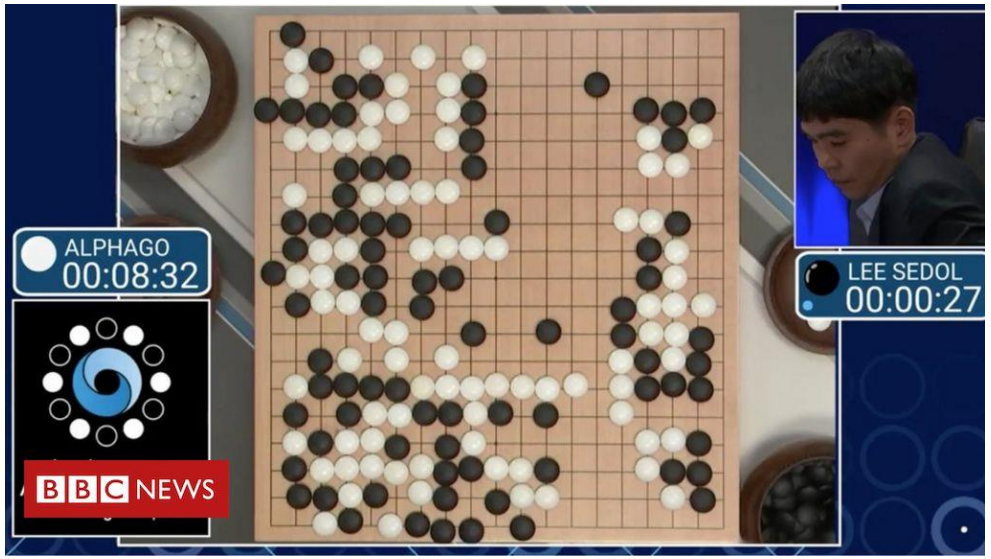




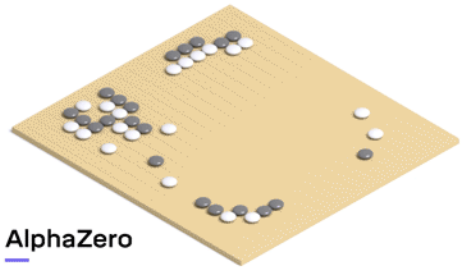
RL Brief History

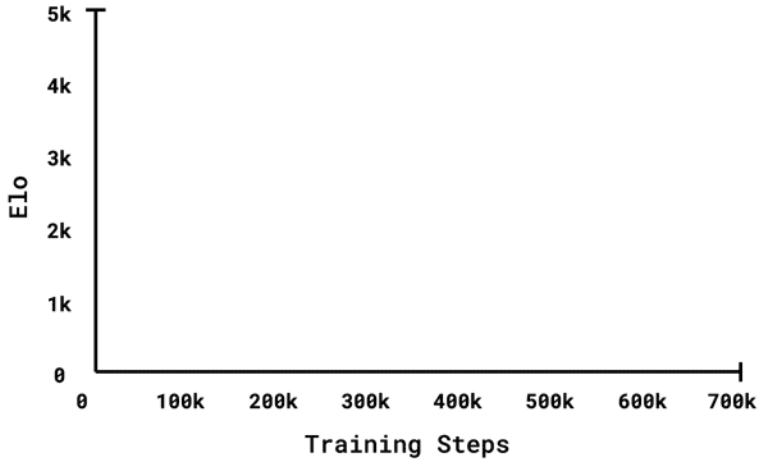
RL Book: Chapter 1

- Dynamic Programming (Bellman 1957)
- Markov Decision Processes (Bellman, Howard 1960)
- Temporal Difference (Witten 1977)
- TD Lambda (Sutton 1988)
- Q-Learning (Watkins 1989)
- TD-Gammon (Tesauro 1992)
- Deep Learning (Krizhevsky 2012)
- AlphaGo (DeepMind 2016)
- AlphaZero (DeepMind 2018)
- First Course @ University of Padova (Susto, Carli 2021)
- ChatGPT (OpenAI, 2022)



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





On-line resources: Gyms and more…

- OpenAI Gyms

https://gym.openai.com/

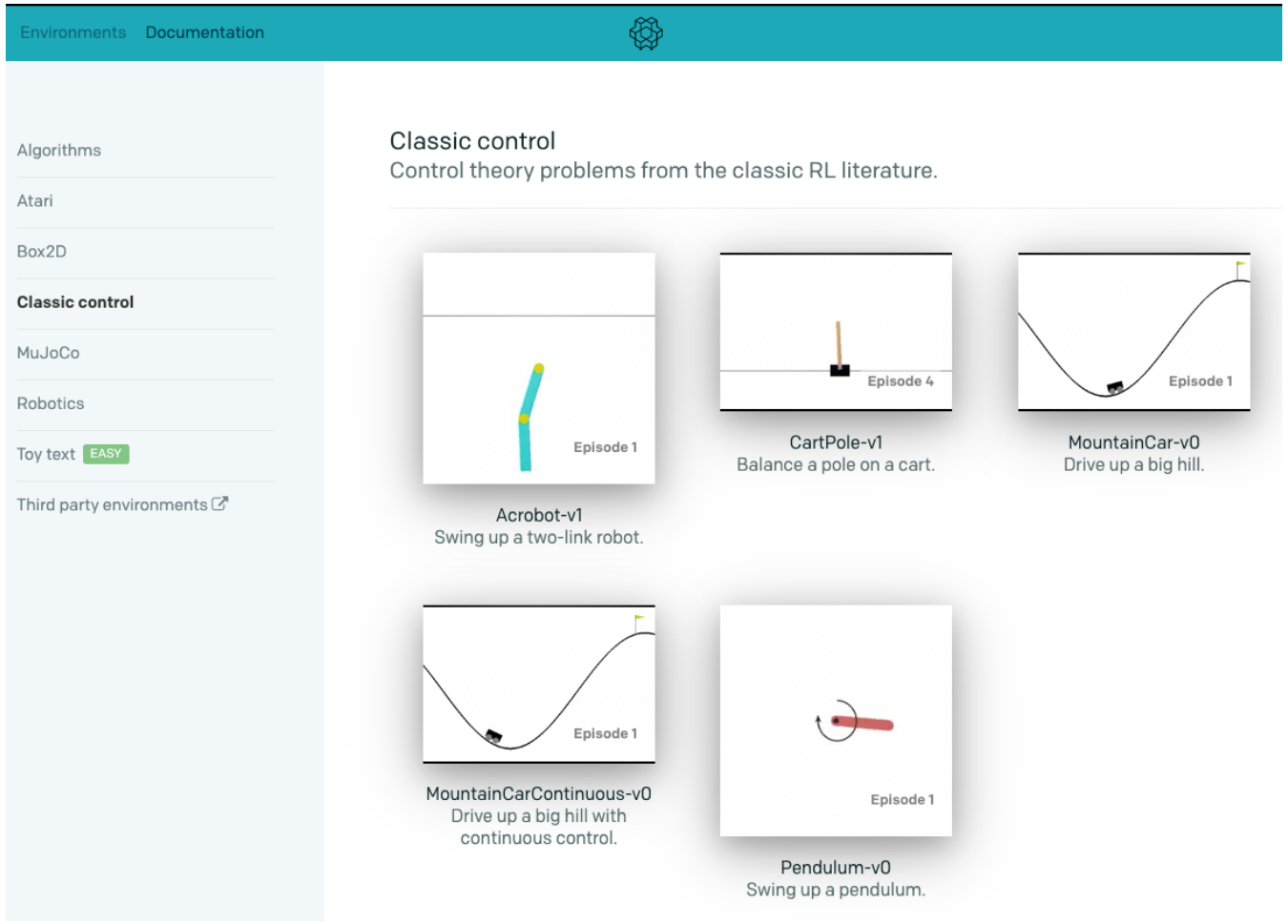
- DeepMind Lab (3D navigation)

https://github.com/deepmind/lab

- Microsoft Malmo (Minecraft)

https://github.com/Microsoft/malmo

- Trading: https://github.com/tensortrade-org/tensortrade



Course Outline

Course Outline

The goal of the course is to provide an overview of the most common algorithms in RL. The course also aims at providing programming experiences on basic RL tasks.

We will cover:

- Tabular Solution methods for RL (Sutton & Barto, Section I): k-armed Bandits, Markov Decision Process, Dynamic Programming, Model-free RL (prediction & control), Planning & Learning
- Approximate Solution methods for RL (Sutton & Barto, Section II): on-policy (prediction & control) with approximation, eligibility traces, policy gradient methods
- Deep Learning approaches for RL

Course Outline

Lec. 01 - Intro
Lec. 02 – k-armed Bandits
Lec. 03 – Markov Decision Processes (MDPs)
Lec. 04 – MDPs and Bellman Equations
Lec. 05 – Dynamic Programming (DP)
Lec. 06 – DP for MDPs and Monte Carlo Methods
Lec. 07 – Monte Carlo Methods
Lec. 08 – Off-policy Monte Carlo & Temporal Difference (TD) Learning
Lec. 09 – TD Learning
Lec. 10 – TD Learning & n-step bootstrapping
Lec. 11, 12 – TD-lambda & eligibility traces
Lec. 13, 14, 15, 16, 17 – Value Function Approximation
Lec. 18 – Value Function Approximation & Policy Gradient
Lec. 19, 20 – Policy Gradient
Lec. 21, 22, 23 – Deep Reinforcement Learning
Lec. 24 – Guest + Advanced Project Proposal

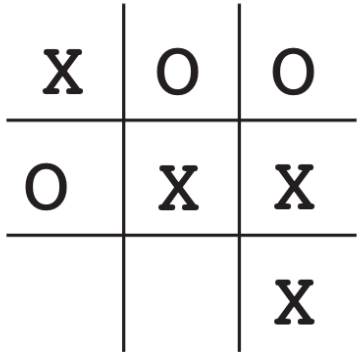
‘Full RL problem’ starts here!
Start scaling up the problem to large systems
Deep Learning-based Approaches

Course Outline (tentative)

Lec. 01 - Intro
Lec. 02 – k-armed Bandits
Lec. 03 – Markov Decision Processes (MDPs)
Lec. 04 – MDPs and Bellman Equations
Lec. 05 – Dynamic Programming (DP)
Lec. 06 – DP for MDPs and Monte Carlo Methods
Lec. 07 – Monte Carlo Methods
Lec. 08 – Off-policy Monte Carlo & Temporal Difference (TD) Learning
Lec. 09 – TD Learning
Lec. 10 – TD Learning & n-step bootstrapping
Lec. 11, 12 – TD-lambda & eligibility traces
Lec. 13, 14, 15 – Value Function Approximation
Lec. 16 – Value Function Approximation & Policy Gradient
Lec. 17, 18 – Policy Gradient
Lec. 19, 20, 21 – Deep Reinforcement Learning
Lec. 22 – MARL
Lec. 23 – LLMs?
Lec. 23 – Model-based RL
Lec. 24 – Guest + Advanced Project Proposal

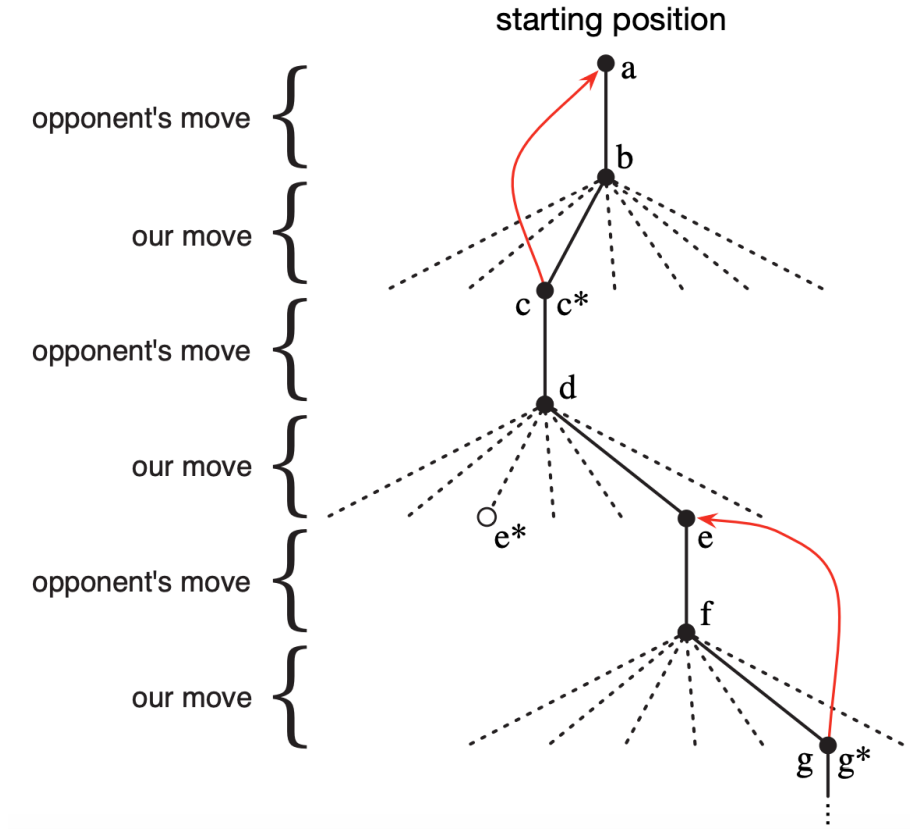
General Knowledge
1st partial written exam
2nd partial written exam
Programming project

Prerequisites



- Calculus
- Supervised Machine Learning (some references in the Moodle page)
- Programming and Python:

1. Python for everybody https://www.coursera.org/specializations/python
2. Anaconda https://www.anaconda.com/products/individual
3. Github https://github.com/ (just to fork/clone https://github.com/ShangtongZhang/reinforcement-learning-an-introduction by Shangtong Zhang); we will go through the code of some examples during the lectures and the lab sessions: you can already take a look at the Tic-Tac-Toe example



Credits

- Image of the course is taken from C. Mahoney ‘Reinforcement Learning’ https://towardsdatascience.com/reinforcement-learning-fda8ff535bb6

- ‘Hide and Seek’ from OpenAI

https://openai.com/blog/emergent-tool-use/

- Examples and applications were inspired by D. Mwiti https://neptune.ai/blog/reinforcement-learning-applications

AMCO
ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND CONTROL RESEARCH GROUP

Reinforcement Learning 2025/2026

RL2
Reinforcement Learning Research Lab @ UniPD

Lecture #01
Organization & Intro

Thank you! Questions?

Gian Antonio Susto