



FROM MAB TO RL AND BEYOND

Francesco Trovò and Alberto Maria Metelli

Machine Learning Modena Meetup
27th January 2021

SHORT BIOS



POLITECNICO
MILANO 1863

DIPARTIMENTO DI ELETTRONICA
INFORMAZIONE E BIOINGEGNERIA



Assistant professor
Research interest in online learning
Course in AI @uniBG



Research assistant
Research interest in reinforcement learning
Oral @NeurIPS 2018

MACHINE LEARNING AND ALGORITHMIC GAME THEORY GROUP

Expertise in:

- Reinforcement learning
- Algorithmic game theory
- Online learning



341 +

Articoli scientifici
pubblicati



5,307 +

Citazioni
scientifiche



30 +

Progetti industriali

INTESA 
SANPAOLO



www.mlcube.com

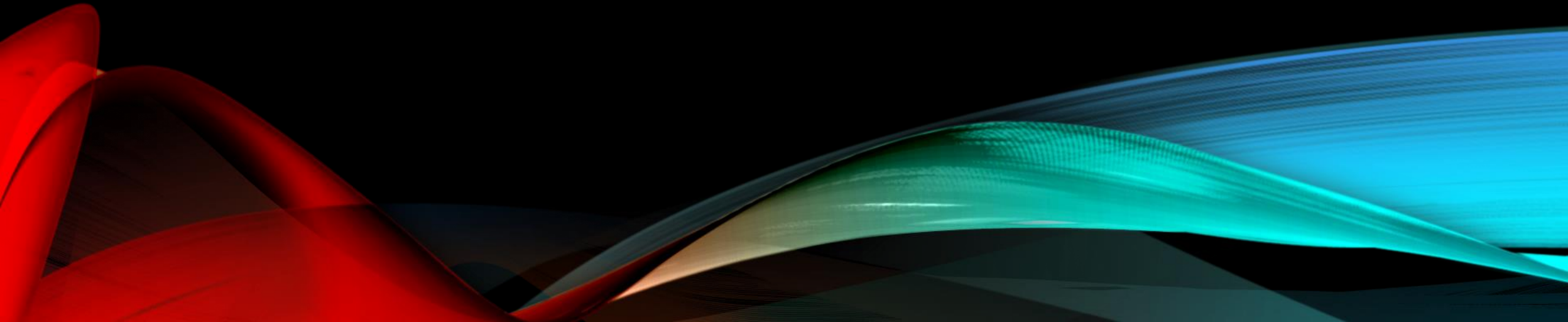
- Innovative startup (founded in 2020)
- Bridging the expertise of academics to the industrial world
- Developing AI projects for the industrial and web world
- Developing ML platform



PART I

MULTI-ARMED BANDITS

Francesco Trovò



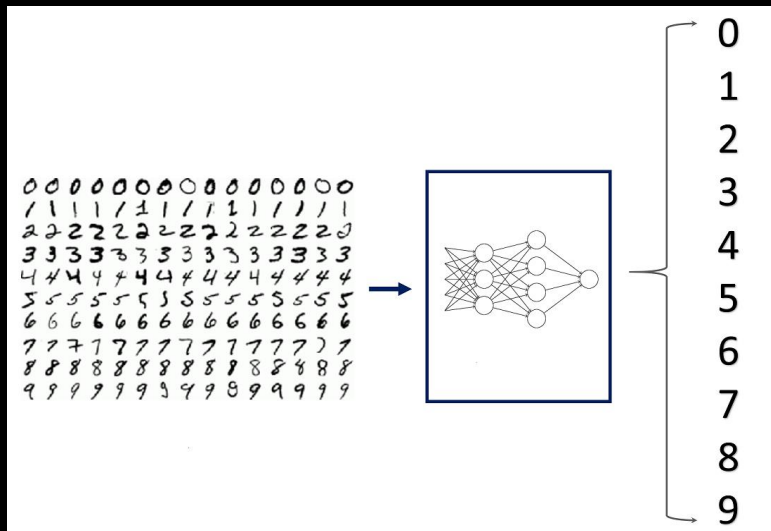
CLASSICAL ML

- “A computer program is said to learn from **experience E** with respect to some **class of tasks T** and **performance measure P** if its performance at tasks in T , as measured by P , improves with experience E ”



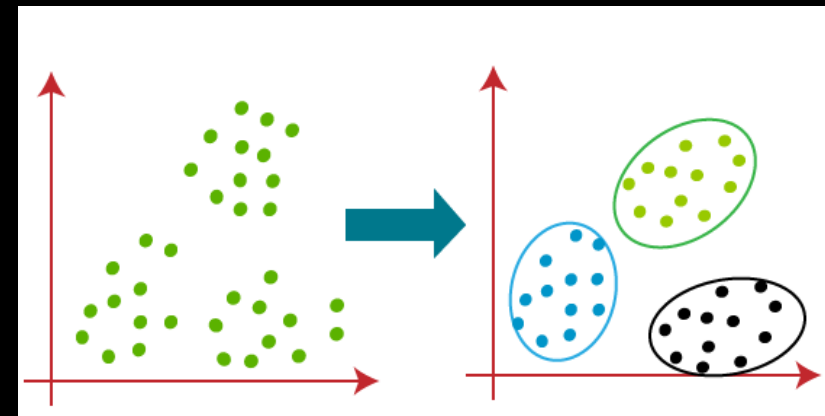
OUTPUT

Prediction

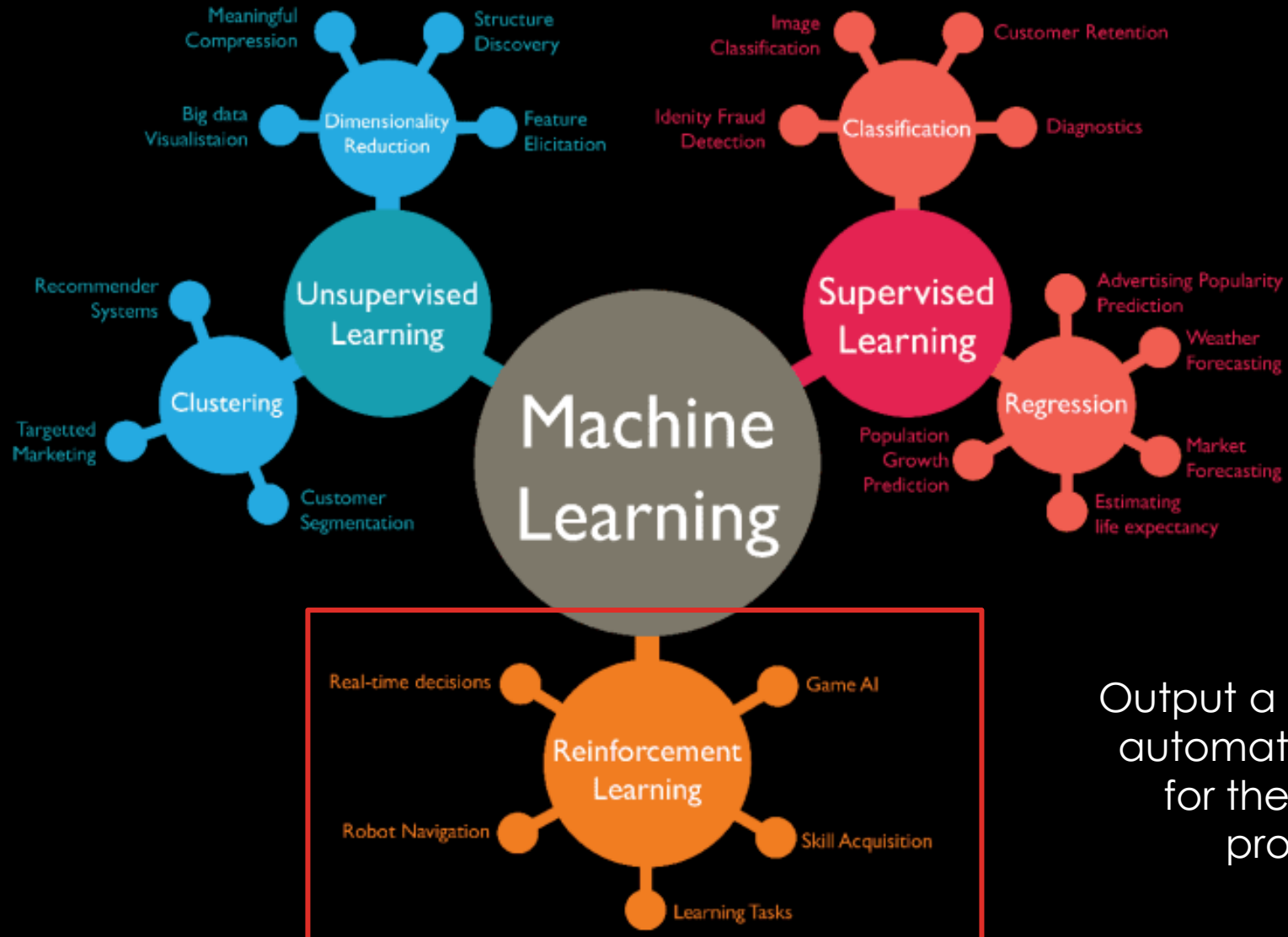


Supervised

or Structure in the data



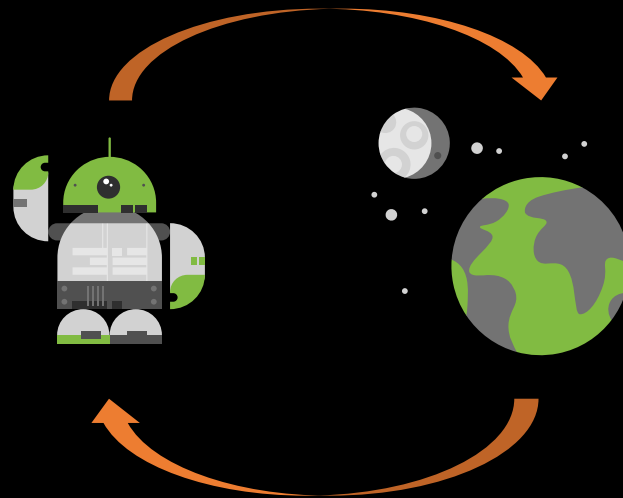
vs Unsupervised



Output a policy or an automatic decision for the specific problem

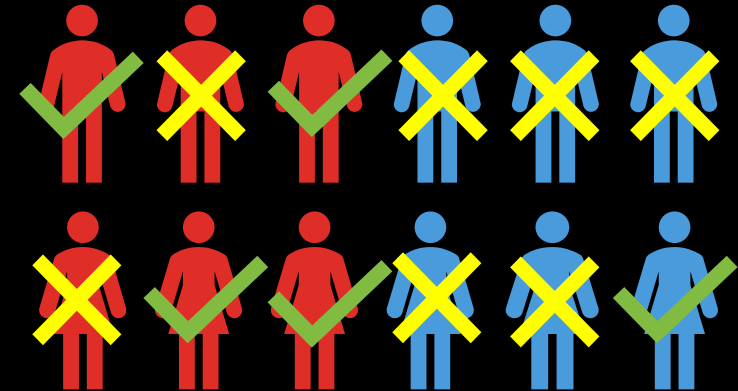
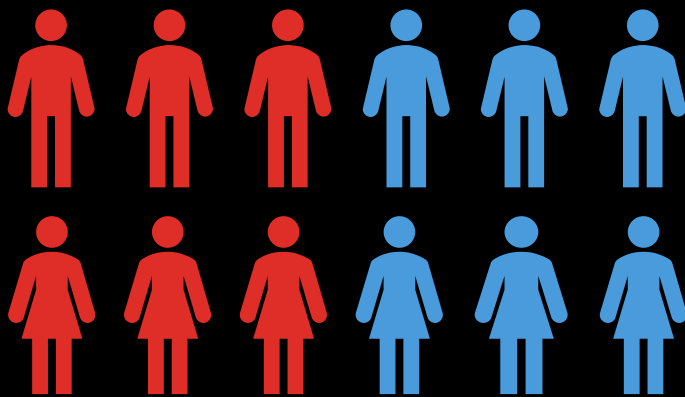
SEQUENTIAL DECISION PROBLEM

- In some setting we are required to take a **decision**
- Therefore, the task is to learn a policy or strategy
- As a new data is coming we want to update our knowledge and act accordingly



HOW MAB ARE BORN

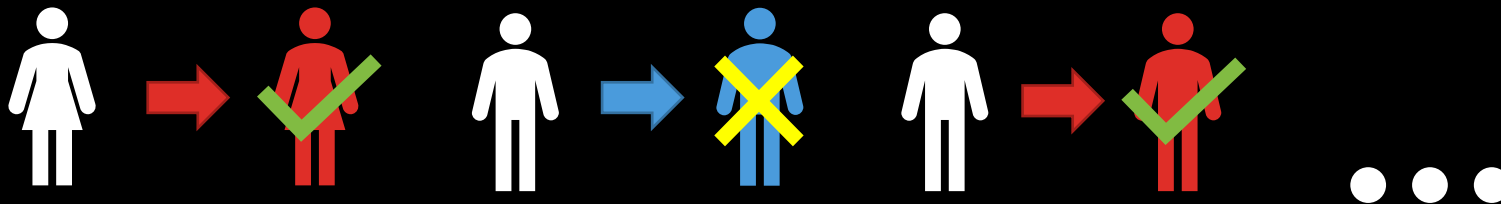
Classical Clinical Trial: red and blue pills



Use statistics to infer which treatment is the most promising one

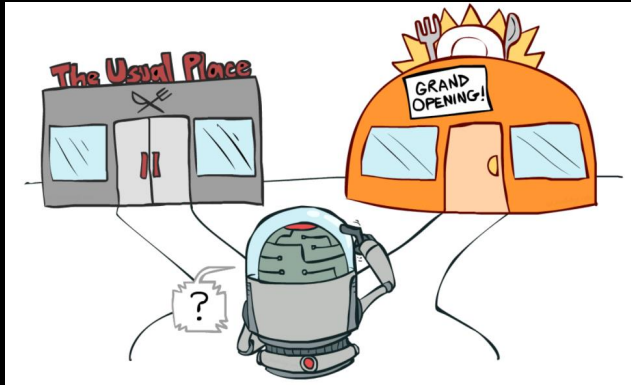
SEQUENTIAL PROCESS

- Design a clinical trial to minimize the number of suboptimal treatments provided to patients



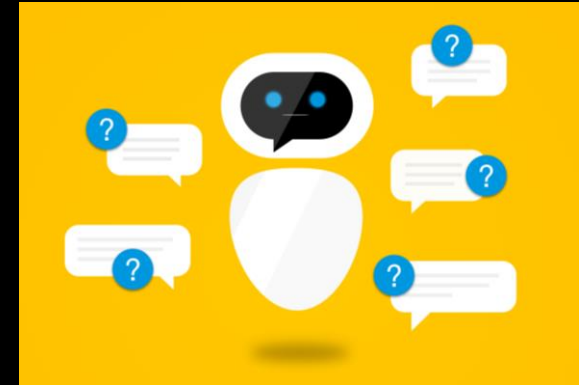
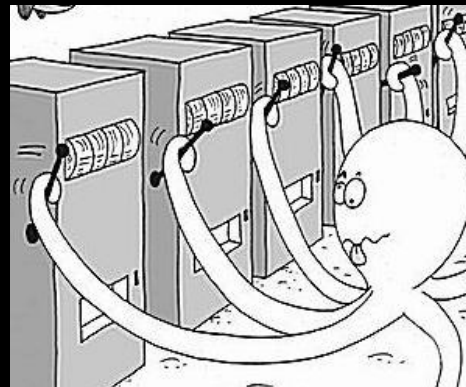
- We call the loss incurred due to lack of information **regret**

MULTI ARMED BANDIT: A COMMON SETTING TO MANY PROBLEMS



Restaurant
selection
problem

Slot selection
problem



Dialogue
model
selection

MEAN IS NOT ENOUGH

- Assume value 1 for success and 0 for failure
- The empirical mean of the blue treatment is zero



- On the following patients we would only select the red treatment

WANDERING AROUND IS NOT HEALTHY

- Using this approach, we would have the same results in terms of regret as the classical clinical trial
- We need to solve the so called:

exploration vs. exploitation (dilemma)



Evaluate and refine the
currently unexplored
options



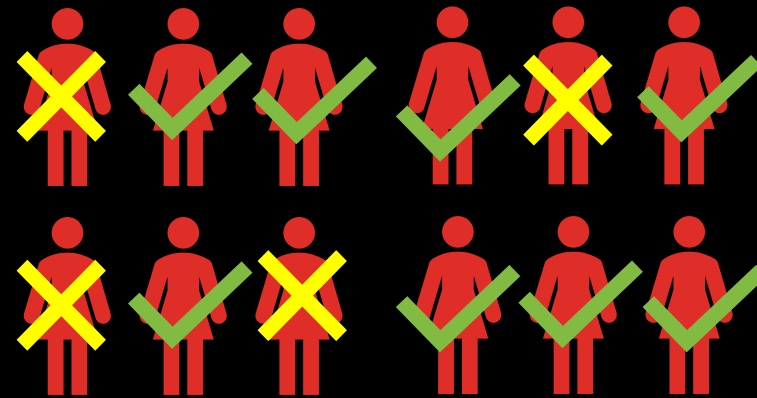
Using the currently
available information
to gain a profit

OPTIMISM IN THE FACE OF UNCERTAINTY

Setting 1



Setting 2



- We need to take into account also the uncertainty of the estimates
- "L'ottimismo è il profumo della vita!" (Tonino Guerra)

UCB1 ALGORITHM

Given a set of arms (options) $A = \{a_1, \dots, a_K\}$

Compute the empiric mean

$$\hat{R}_t(a_i) = \frac{\sum_{i=1}^t r_{i,t} \mathbb{1}\{a_i = a_{i_t}\}}{N_t(a_i)} \quad \forall a_i$$

Compute the uncertainty bound

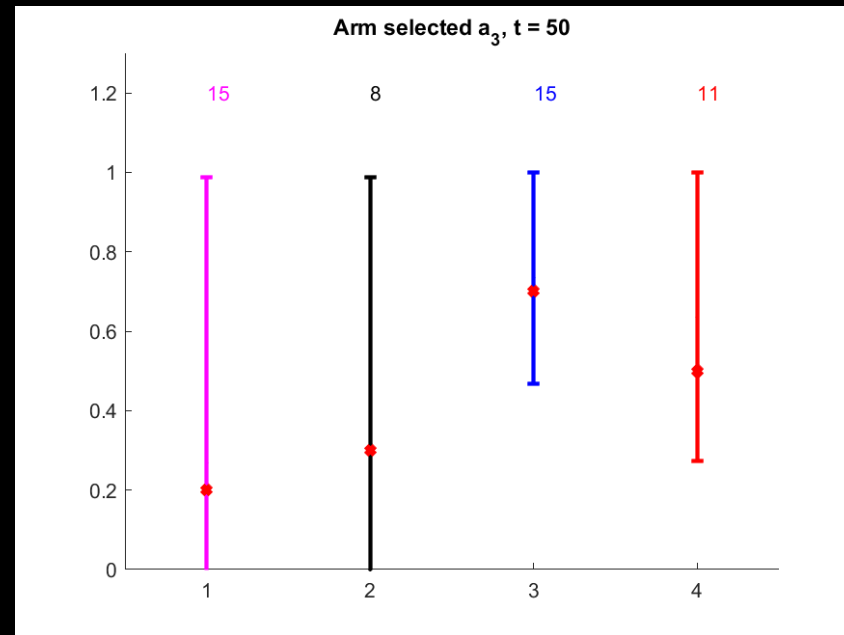
$$B_t(a_i) = \sqrt{\frac{2 \log t}{N_t(a_i)}} \quad \forall a_i$$

Select the arm with the largest UCB

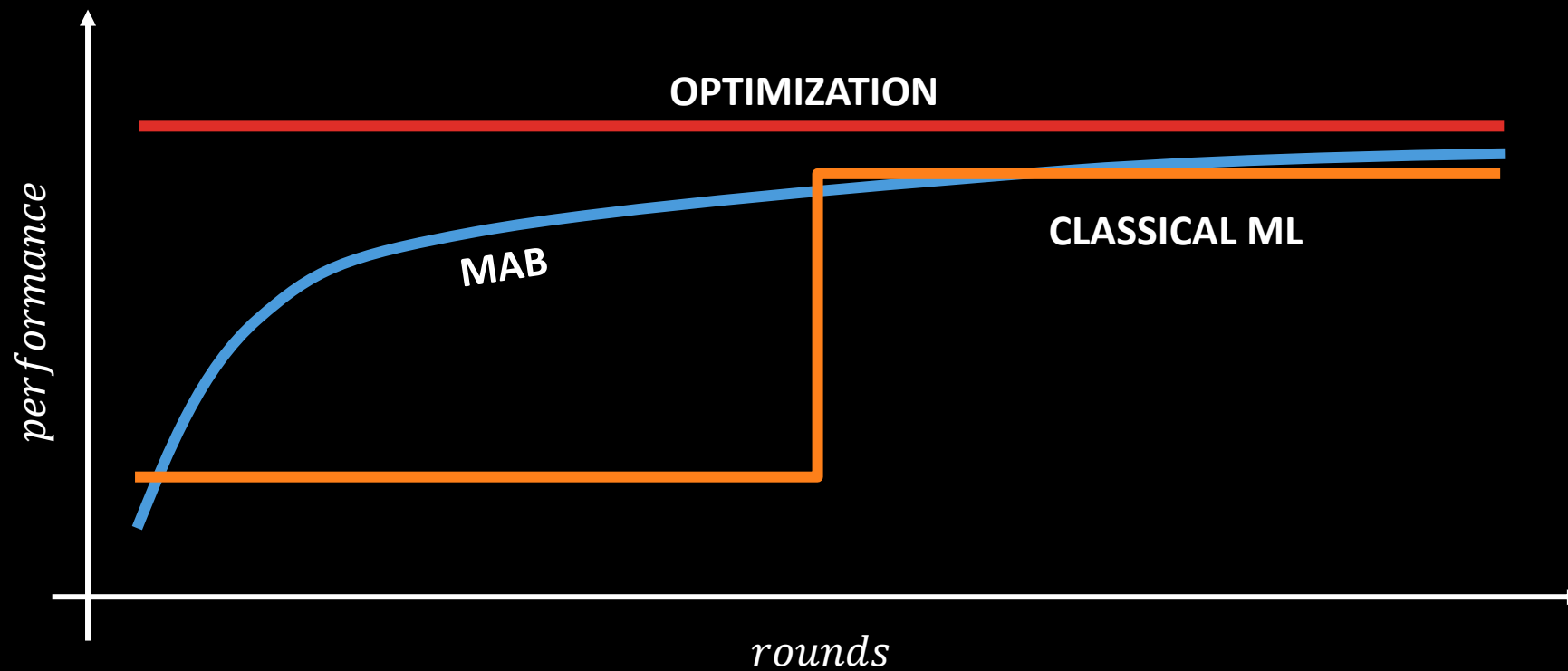
$$a_{i_t} = \arg \max_{a_i \in \mathcal{A}} \left(\hat{R}_t(a_i) + B_t(a_i) \right)$$

EXECUTION EXAMPLE

- $A = \{a_1, a_2, a_3, a_4\}$



SEQUENTIAL APPROACH ADVANTAGES



THEORETICAL GUARANTEE

- Lower Bound

$$\lim_{T \rightarrow \infty} L_T \geq \boxed{\log T} \sum_{a_i | \Delta_i > 0} \frac{\Delta_i}{KL(\mathcal{R}(a_i), \mathcal{R}(a^*))}$$

- Upper bound UCB1

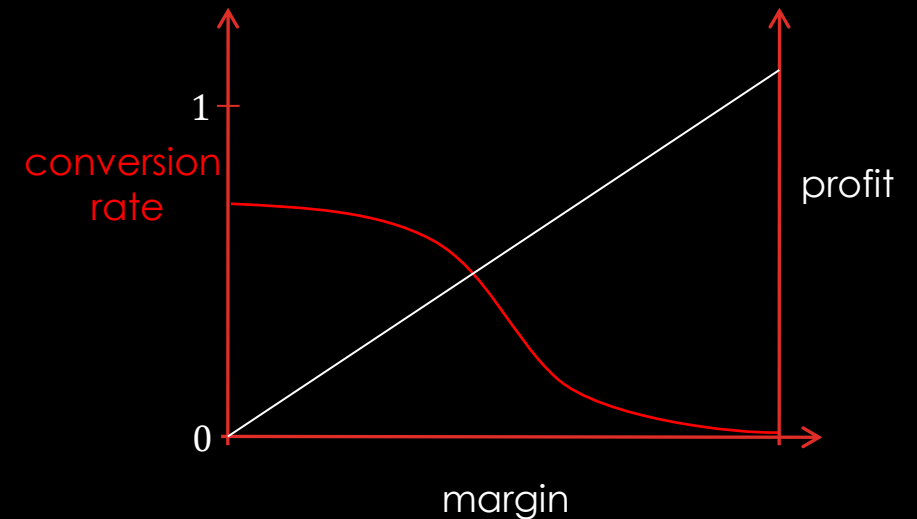
$$L_T \leq 8 \boxed{\log T} \sum_{i | \Delta_i > 0} \frac{1}{\Delta_i} + \left(1 + \frac{\pi^2}{3}\right)$$

PRACTICAL PROBLEMS

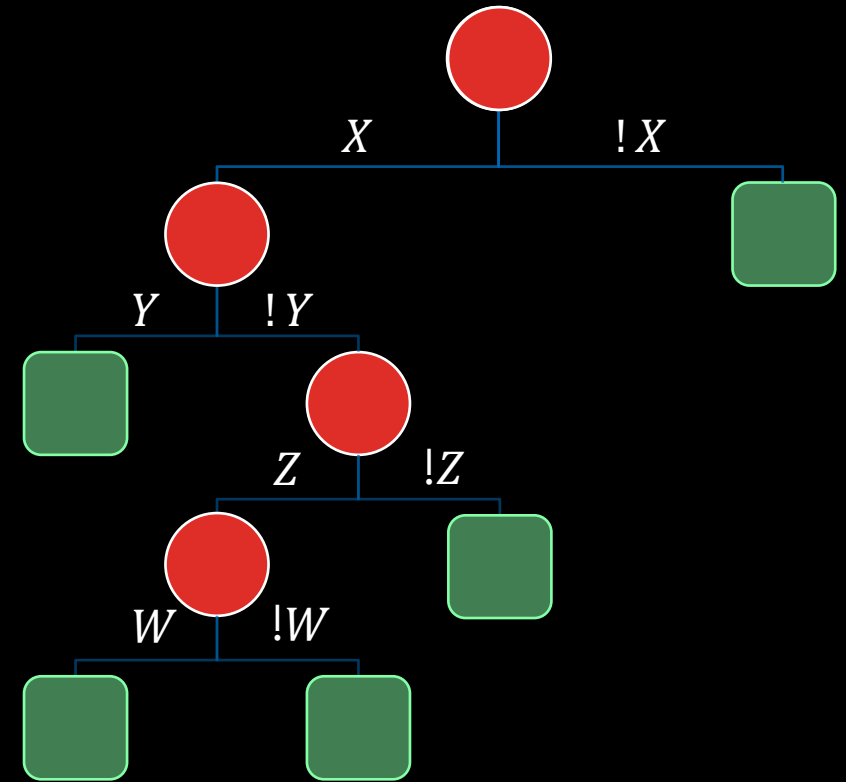
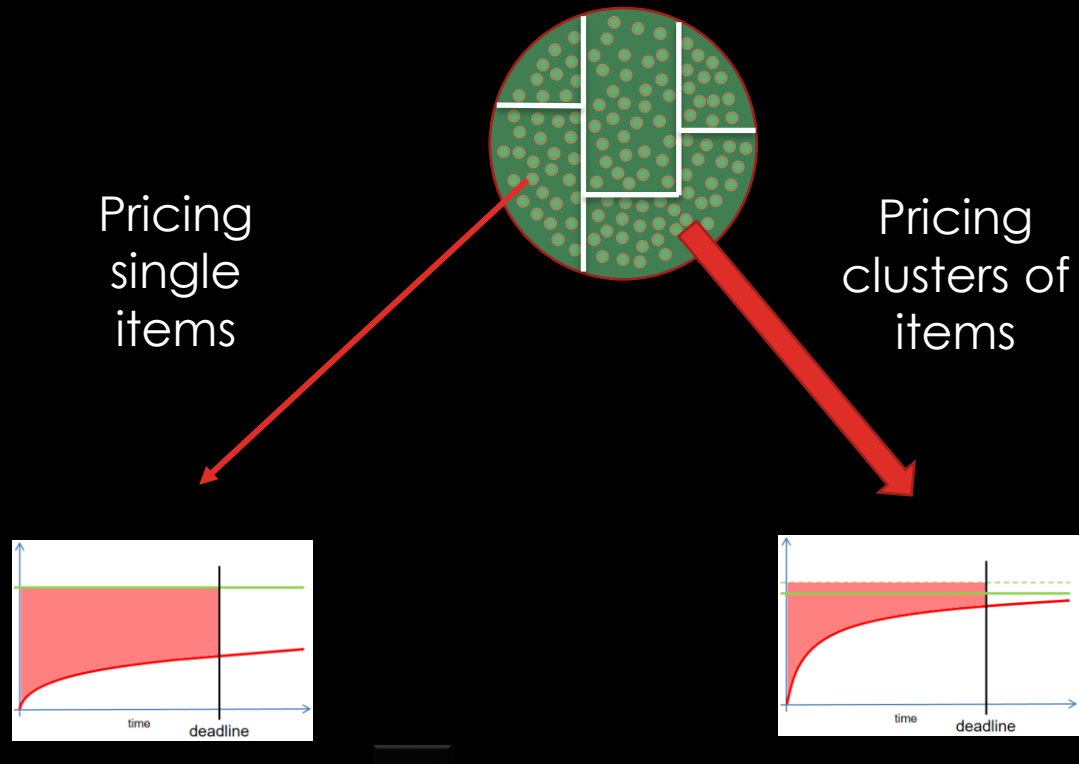
- It is a good alternative to A/B testing in general
 - Allows multiple options
 - Minimize the loss
 - Does not exclude completely the use of any option
- Even if this framework is formal and elegant, it hardly generalizes to real problems (too simple)

PRICING PROBLEM

- Problem formulation:
 - Given an inventory of products
 - Select the most profitable price for each product
- Characteristics:
 1. Large inventory
 2. Continuous choice
 3. Non-stationarity

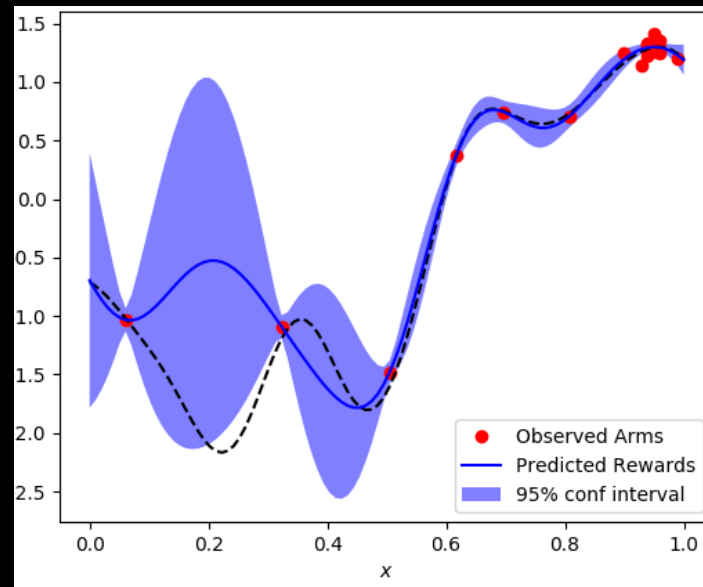


1. LARGE INVENTORY



2. CONTINUOUS CHOICE

- Gaussian Process
- Selection of the next arm to play according to UCB provided by GPs



Srinivas, N., Krause, A., Kakade, S., & Seeger, M. (2010). Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. In *Proceedings of the 27th International Conference on Machine Learning*.

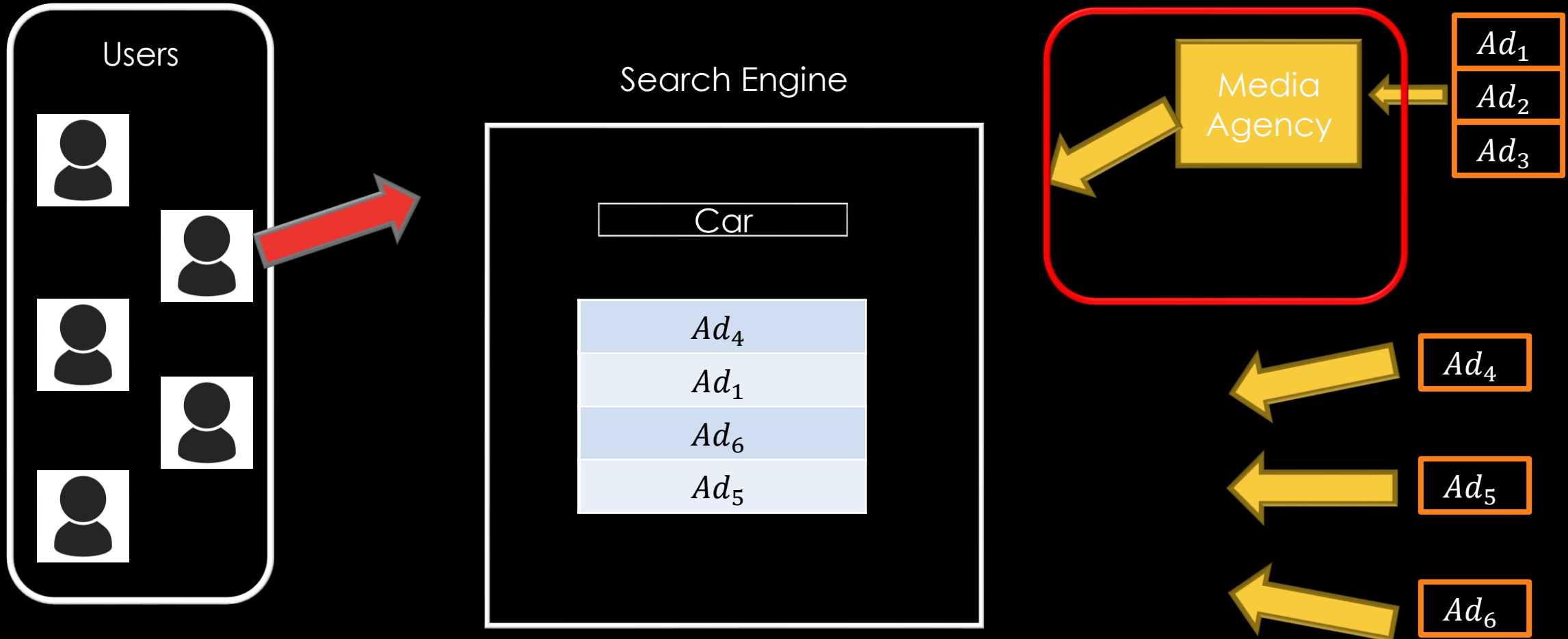
3. NON-STATIONARITY

- Apply a sliding window to the system



- Passive approach
- Active approaches aim at identifying a change in the distribution of the rewards

ADVERTISING PROBLEM

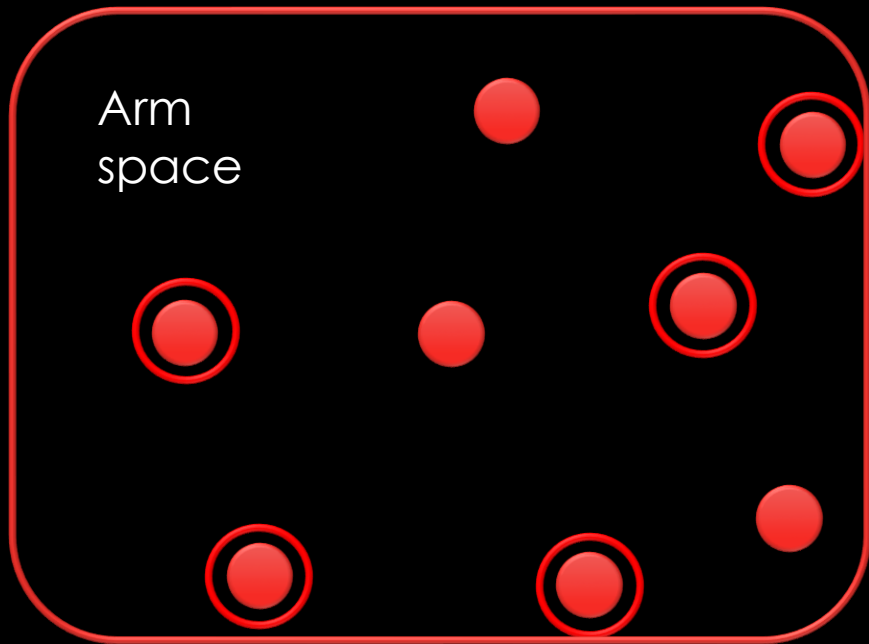


ADVERTISING MODEL

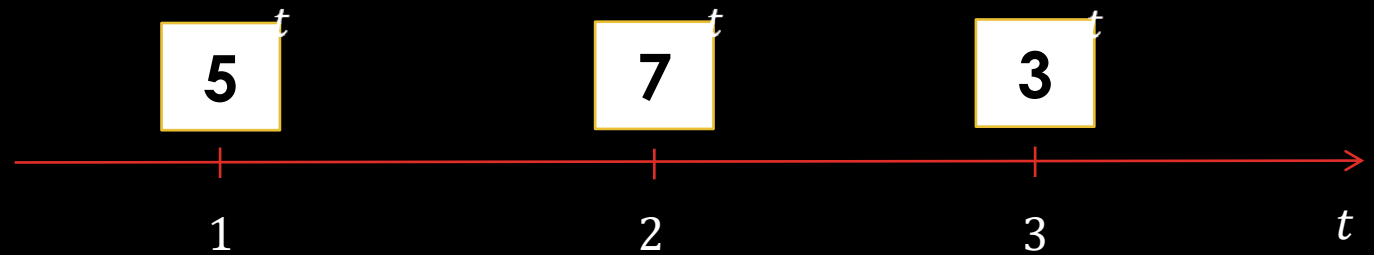
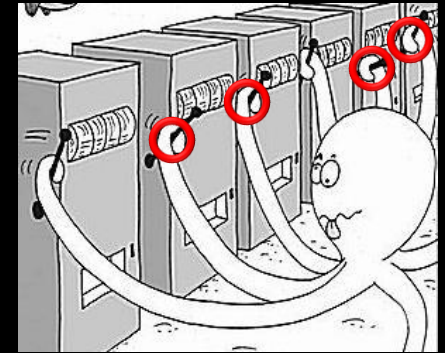
- Problem formulation:
 - Given a set of ads, select bid and budget
 - Assuring that the overall budget is no more than a given daily one



COMBINATORIAL BANDITS

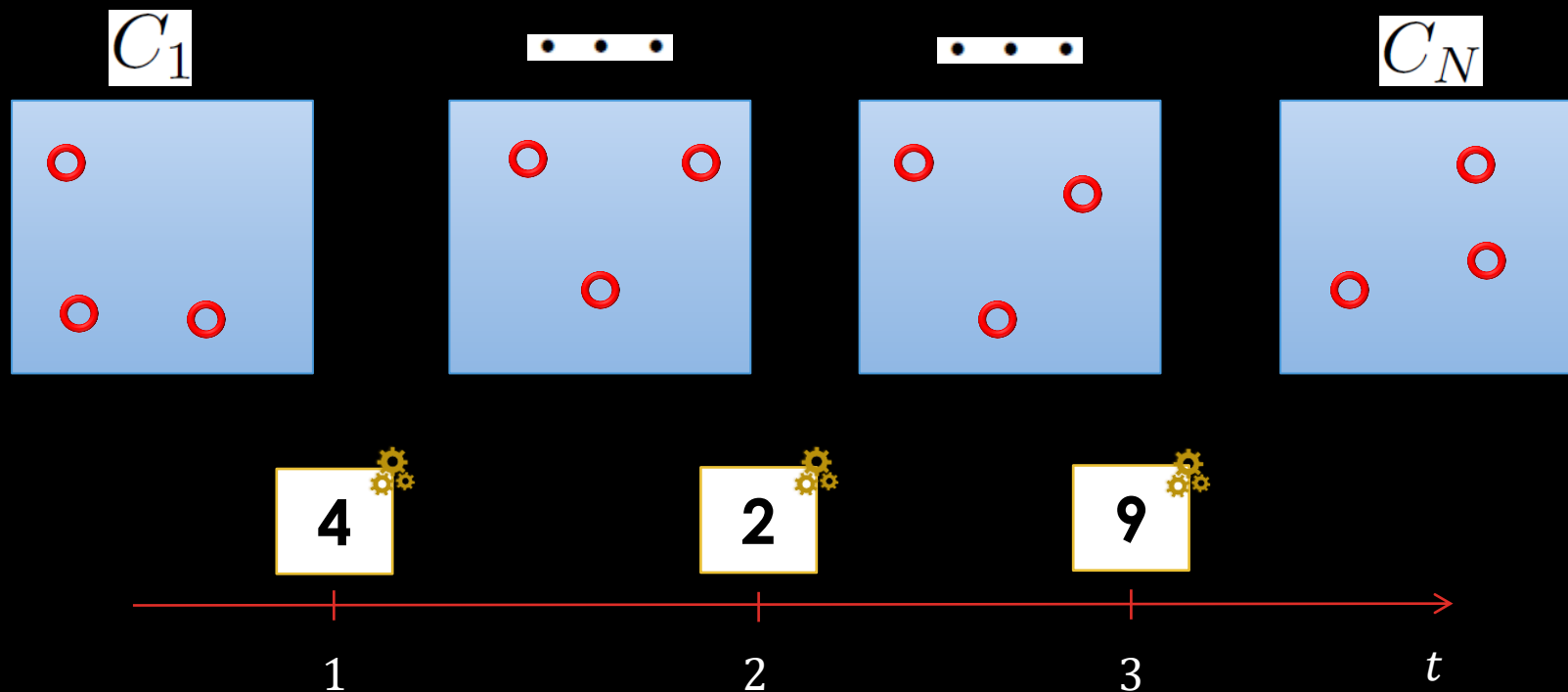


Optimization
Oracle

A small icon of three interlocking gears, one yellow and two grey, positioned to the right of the "Optimization Oracle" text.

CMAB FOR ADVERTISING

- Choose using Upper Confidence Bounds





CONCLUSION

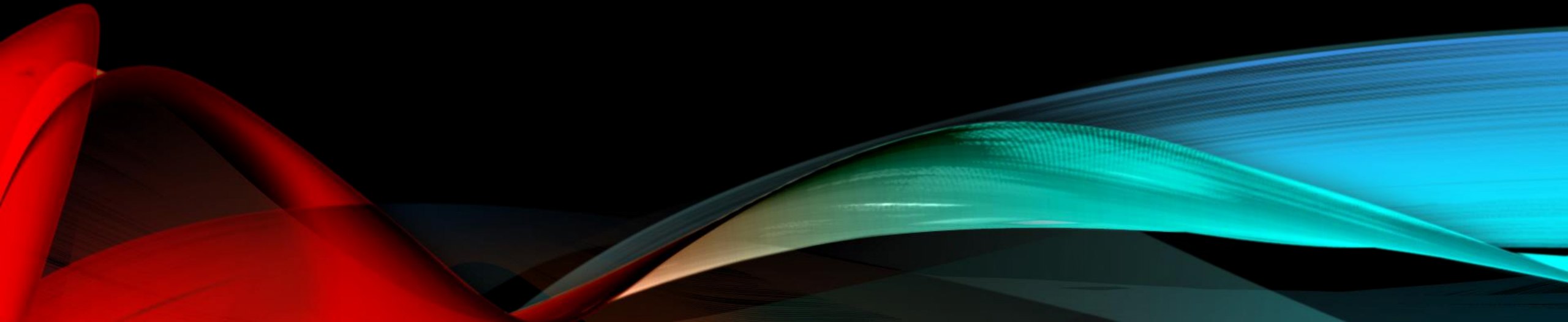
- MAB algorithms are only nice and elegant tools to study theoretically
- Its extensions are used in many applicative fields

Future directions:

- Fairness in MAB
- Dynamics in MAB
- Domain specific MAB

PART II REINFORCEMENT LEARNING

Alberto Maria Metelli

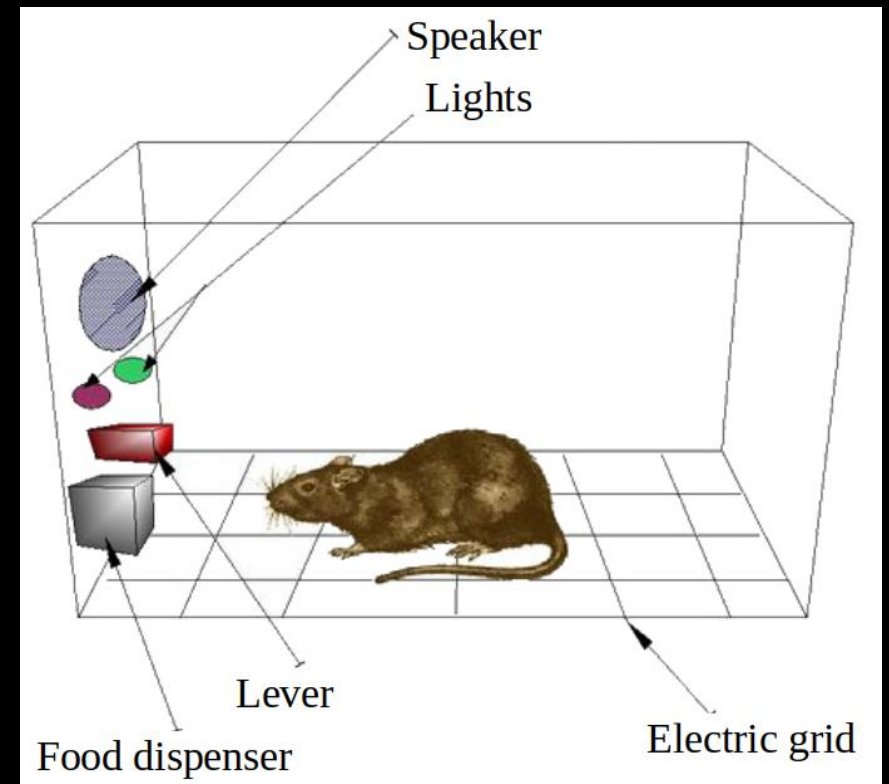


ORIGINS OF REINFORCEMENT LEARNING

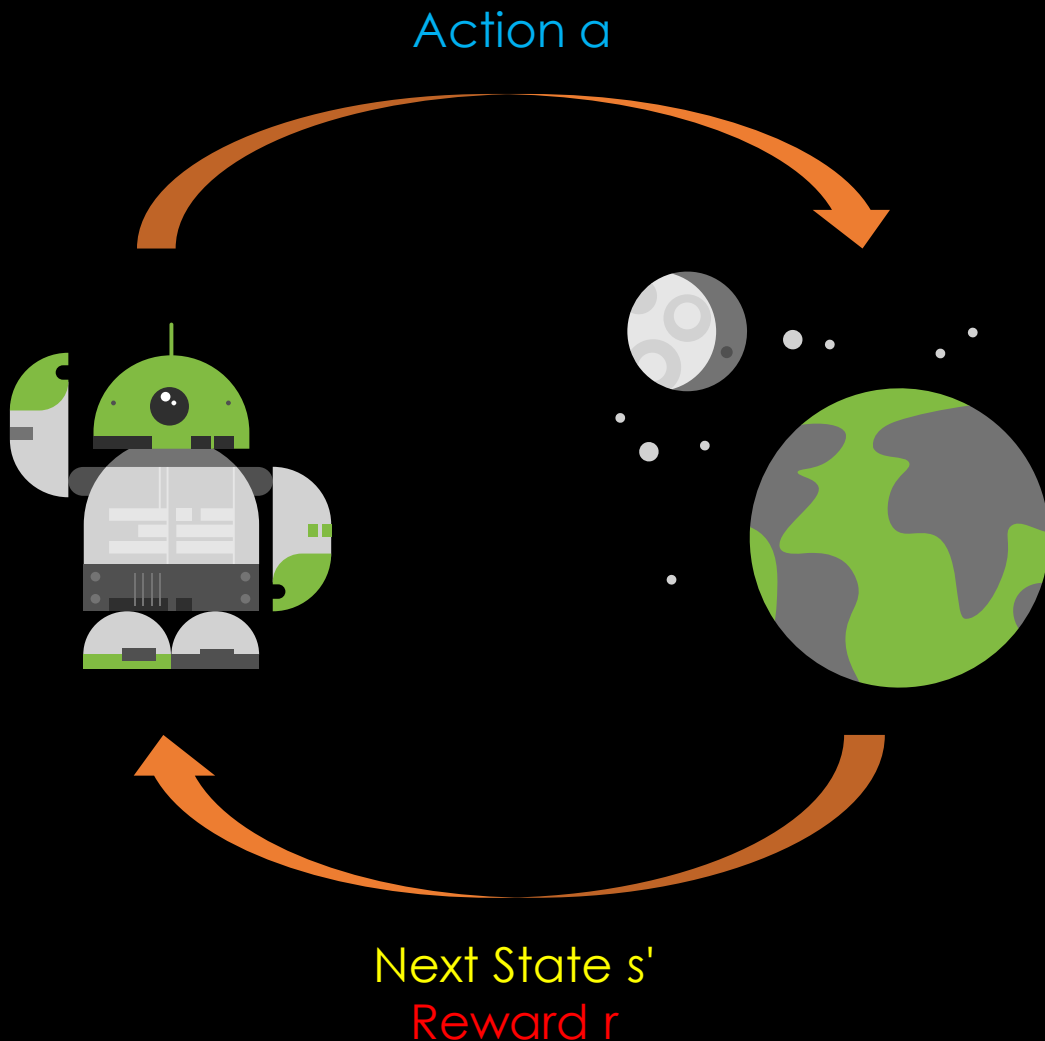
- RL originates in **behavioral psychology**

«a consequence applied that will strengthen an organism's future behavior whenever that behavior is preceded by a specific antecedent stimulus»

- Skinner box → Operant conditioning



AGENT AND ENVIRONMENT



At each step:

- The agent observes the **state** s
- The agent plays **action** $a \sim \pi(\bullet | s)$
- The environment transitions to the **next state** $s' \sim P(\bullet | s, a)$
- The environment emits a scalar **reward** $r = R(s, a)$

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

HOW DOES RL DIFFERS FROM MAB?

- **Same goal:** select actions to maximize cumulative rewards but ...
- ... actions may have **long-term** consequences
- ... reward may be **delayed**
- ... it may be better to **sacrifice** immediate reward to gain more long-term reward



WHEN TO USE RL INSTEAD OF AUTOMATIC CONTROL?

- When the environment dynamics is **unknown**
- When the environment dynamics is known but **too complex** to be effectively used



OPTIMALITY CRITERIA

Goal of an RL agent: maximize the (expected discounted) cumulative reward

$$\sum_{t \geq 0} \gamma^t r_t$$

$\gamma \approx 0$



«meglio un uovo
oggi che una
gallina domani?»

$\gamma \approx 1$



OPTIMAL VALUE FUNCTION AND OPTIMAL POLICY

- Maximum cumulative reward from (s,a)
- **Bellman** equation

$$Q^*(s, a) = \underset{\text{instantaneous reward}}{r(s, a)} + \underset{\text{transition model}}{\gamma \sum_{s' \in \mathcal{S}}} \underset{\text{cumulative reward from next state on}}{\max_{a' \in A}} Q^*(s', a')$$

- **Optimal** policy

$$\pi^*(s) = \operatorname{argmax}_{a \in A} Q^*(s, a)$$

EXAMPLE

Reward

0	0	0	1.00
0		0	- 0.88
0	0	0	0

Reflexion

→	→	→	
↑		↑	
↑	→	↑	←

LEARNING IN TABULAR PROBLEMS: Q-LEARNING

- **Problem:** learn the optimal value function from samples

Initialize Q

Observe the initial state s_0

For each step $t=0, 1, \dots$

 Select action a_t with an **exploration policy**

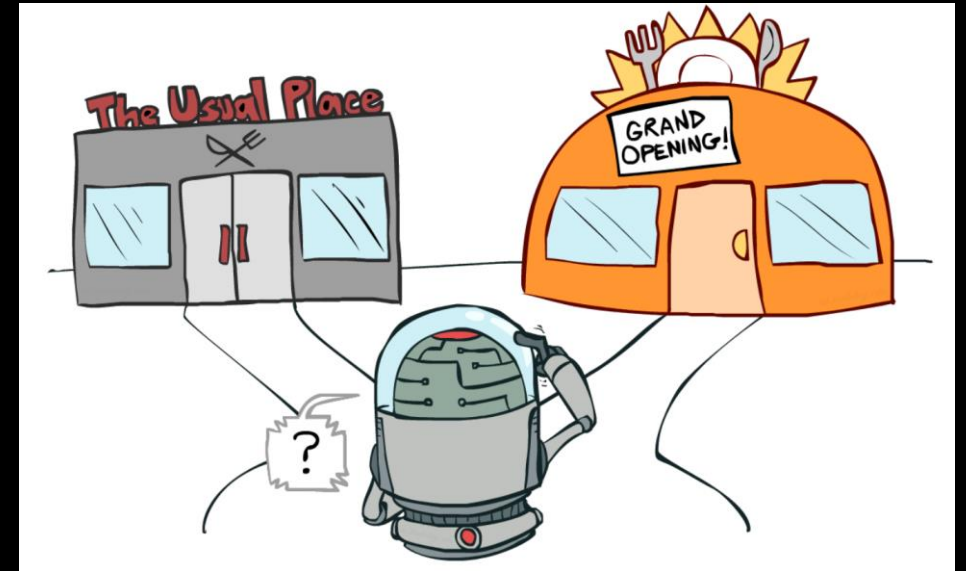
 Take action a_t and observe reward r_t and next state s_{t+1}

 Update $Q(s_t, a_t) \leftarrow (1-\beta) Q(s_t, a_t) + \beta[r_t + \gamma \max_a Q(s_{t+1}, a)]$

- How to select the **exploration policy**?

EXPLORATION VS EXPLOITATION

- All actions should be tried sufficiently often!
- **exploration-exploitation** dilemma
- Cost of exploration (simulation vs real system)
- **Examples:** epsilon greedy, Boltzmann, UCB



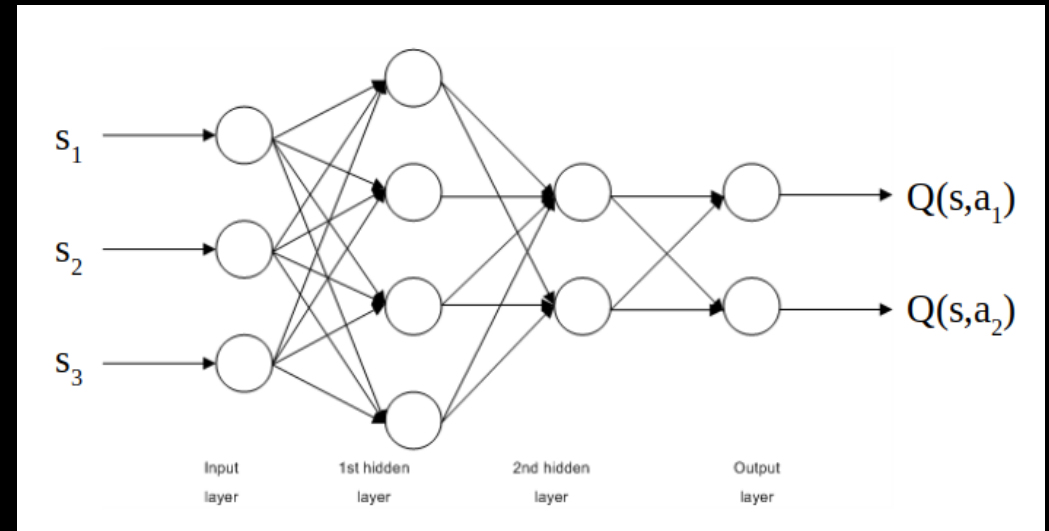
LEARNING WITH CONTINUOUS STATES

- What if the state space is infinite?
- **Function approximation**

$$Q(s_t, a_t; \theta)$$

- Minimize the loss via gradient descent over θ (not working well...)

$$\min_{\theta} \left(Q(s_t, a_t; \theta) - \left(r_t + \gamma \max_{a \in A} Q(s_{t+1}, a; \theta) \right) \right)^2$$



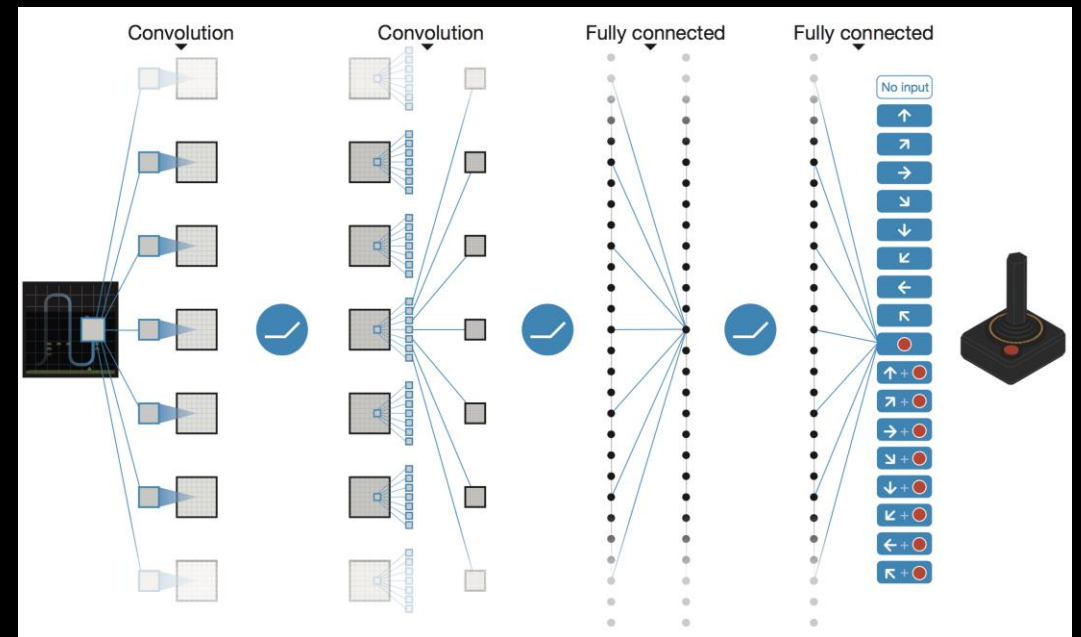
STEPS TOWARDS DEEP RL

$$\min_{\theta} \left(Q(s_t, a_t; \theta) - \left(r_t + \gamma \max_{a \in A} Q(s_{t+1}, a; \theta) \right) \right)^2$$

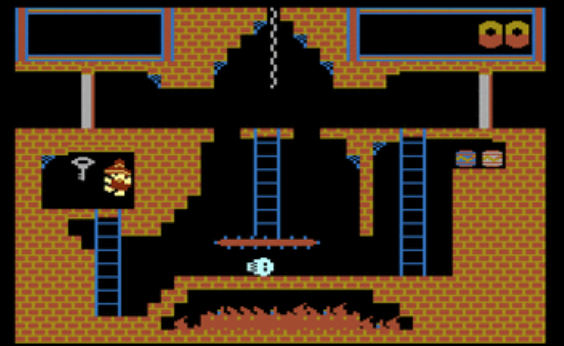
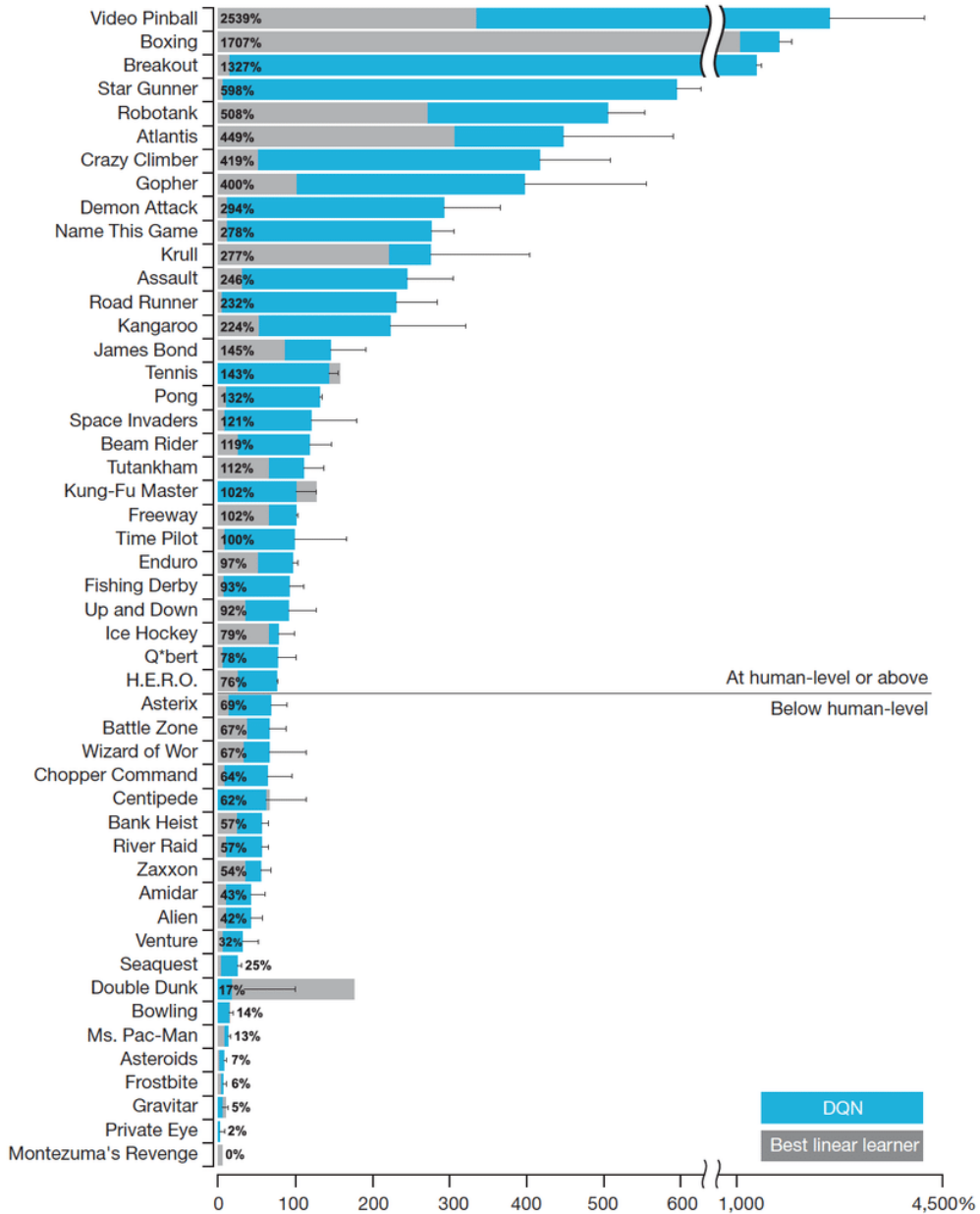
Not exactly **supervised learning**...

- Samples are dependent
- Learning stability

And some other tricks... → **DQN** (Deep Q-Networks)



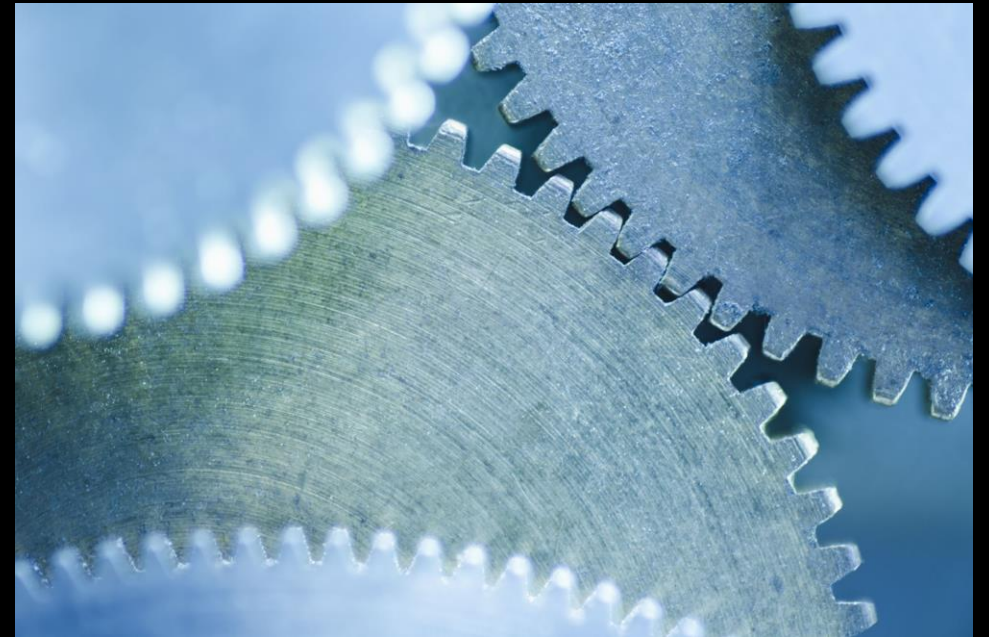
DQN ON ATARI



Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540), 529-533.

TOWARDS REAL-WORLD APPLICATIONS

- Safe Behavior and Safe Learning
- Multi-objective tasks
- Interpretability
- Learn by imitation



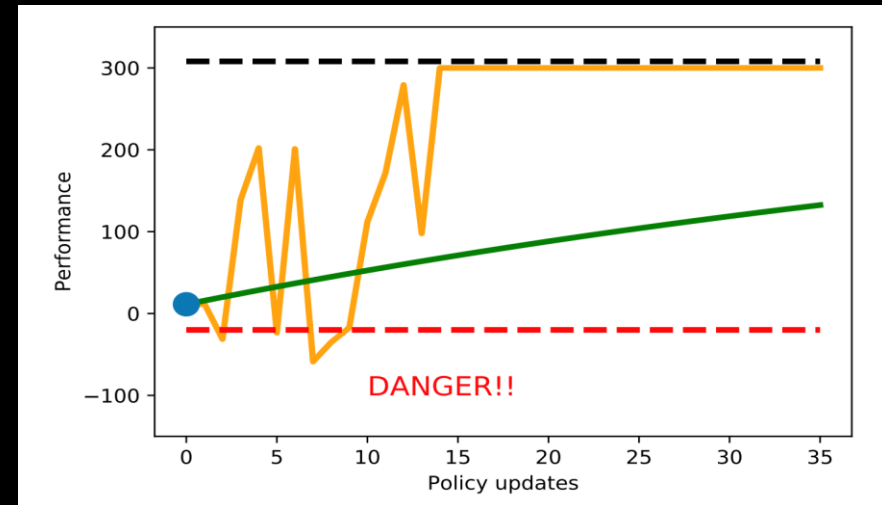
SAFE REINFORCEMENT LEARNING

Learn a "safe" behavior



Garcia, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1), 1437-1480.

Learn/explore "safely"



Papini, Matteo, Andrea Battistello, and Marcello Restelli. "Balancing learning speed and stability in policy gradient via adaptive exploration." *AISTATS 2020*.

RL FOR AUTONOMOUS DRIVING

- **Goal:** display human-like behavior
- Two driving scenarios
 - Highway driving → **multiobjective**
 - Urban (intersection, roundabout)
- Sensor inputs, discrete actions
- **Interpretability** → parametric rule-based policy



RL FOR DRIVING ON A TRACK

- **Goal:** minimize the lap time
- Human expert demonstration collected on a simulator
- **Objectives**
 - mimic the expert → **imitation learning**
 - improve the expert's policy → planning



... AND BEYOND

- Lifelong/Continual RL
- Meta RL
- Multi-Agent RL





POLITECNICO
MILANO 1863

**DIPARTIMENTO DI ELETTRONICA
INFORMAZIONE E BIOINGEGNERIA**



THANK YOU!

Francesco Trovò

francesco1.trovo@polimi.it

Alberto Maria Metelli

albertomaria.metelli@polimi.it