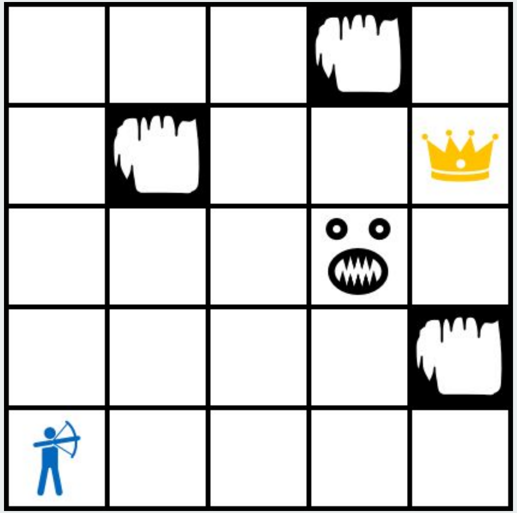
Hunt The Wumpus



Green Team

Marco Di Panfilo, Alessandra Lorefice, Denis Mugisha, Gianluigi Pellè

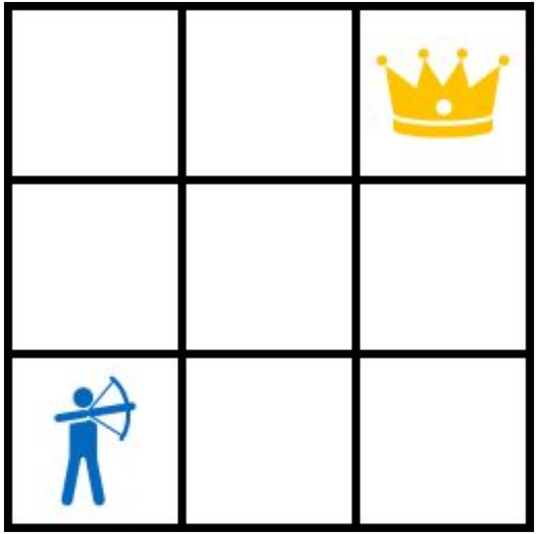
Hunt Wumpus State Safe

Simple approach only for “safe” worlds.

Minimal information needed for state:

- agent _location
- agent_orientation
- has_agent_grabbed_gold
- has_agent_climbed _out

Advantage: a very small q-table (1024 different states)



Hunt Wumpus State Custom

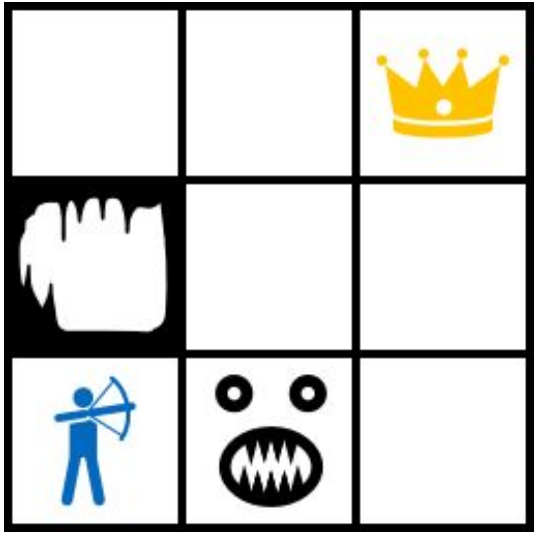
General solution for all worlds.

Minimal information needed for state:

- agent _location
- agent_orientation
- has_agent_grabbed_gold
- has_agent_climbed _out

Wumpus:

- Is_arrow_available
- is_wumpus_alive
- has_agent_perceived _scream



So, how did we solve the game?

Q-Learning Agent

We implemented a q-learning agent, building a q-table based on this update formula:

$$q^{new}(s, a) = (1 - \alpha)\ \underbrace{q(s, a)}_{\text{old value}} + \alpha\ \overbrace{\left(R_{t+1} + \gamma\ max\ q(s^{'}, a^{'})\right)}^{\text{learned value}}$$

*We chose alpha=0.20 and gamma=0.80

Rewards strategy

We first tried only using the final reward, but this lead to a huge training in order to solve the game.

So, we added an intermediate reward, which assigns 1’000 points for grabbing the gold, which helped us to reduce the required training.

Exploration strategy

- First attempt:

Epsilon-greedy exploration strategy

Problem: it converges to a local minimum: immediate climb out

Exploration strategy

- Second attempt:

Soft-max action selection strategy

Problem: it converges to a local minimum: immediate climb out

Exploration strategy

- Third attempt:

Exponential decay action selection strategy

Problem: it works well on **huge training** since the exploration rate decreases in an **exponential way**. For small trainings it decreases too fast or too slow.

Exploration strategy

4

- Fourth attempt:

Custom made strategy

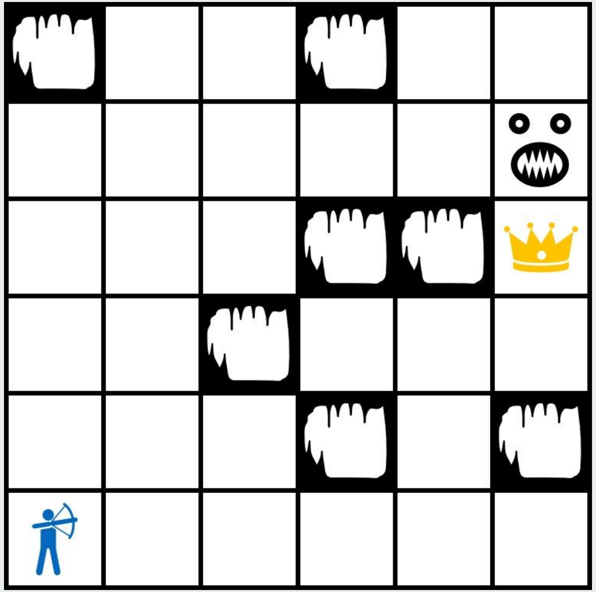
Training: 75% 25%

Exploration: 70%
Exploitation: 30%

Exploration: 15%
Exploitation: 85%

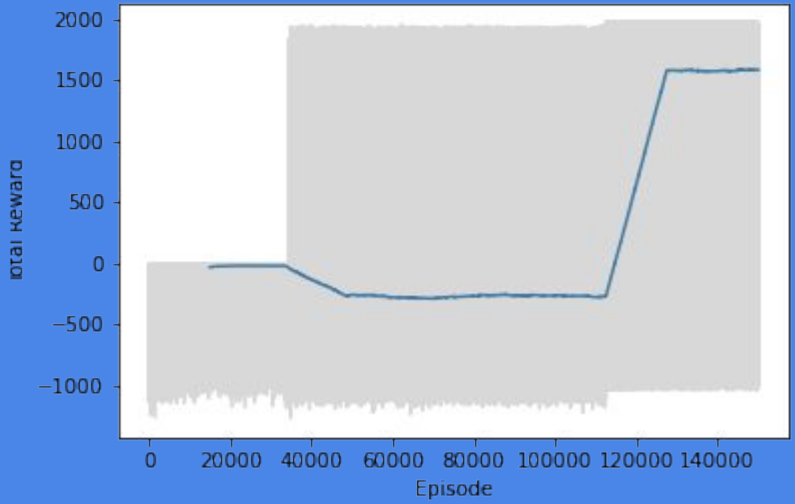
Results

1° scenario (safe)

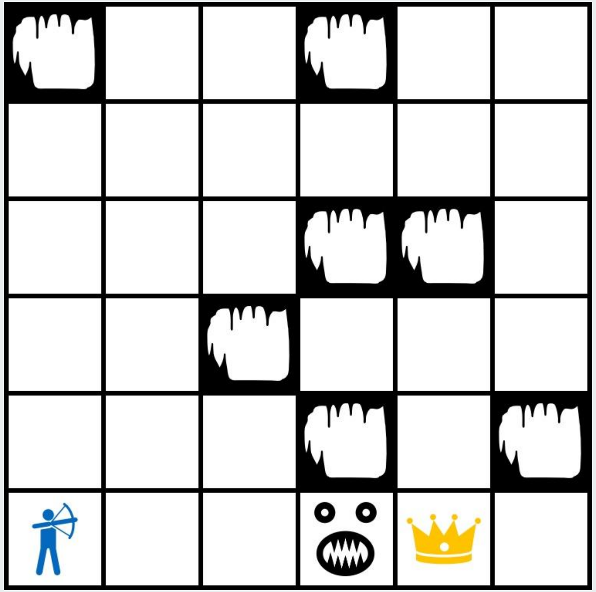


Training plot

(150’000 episodes)

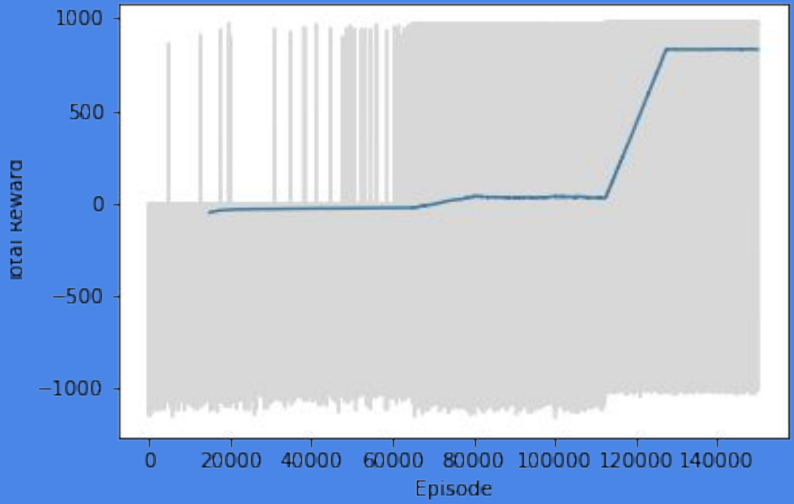


2° scenario (wumpus)

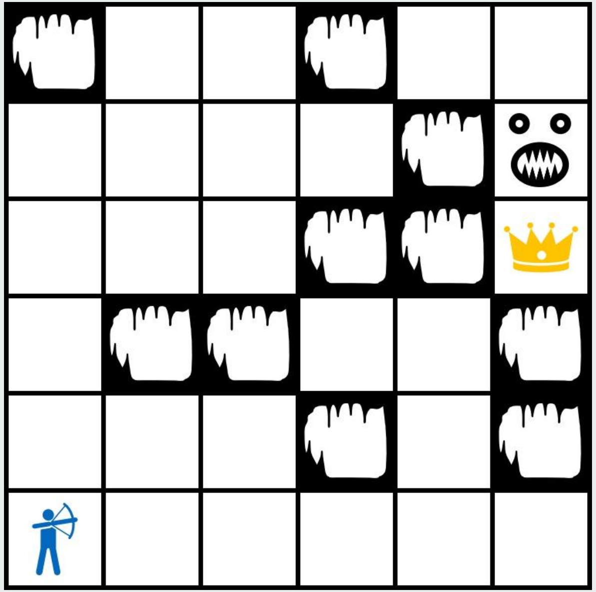


Training plot

(150’000 episodes)

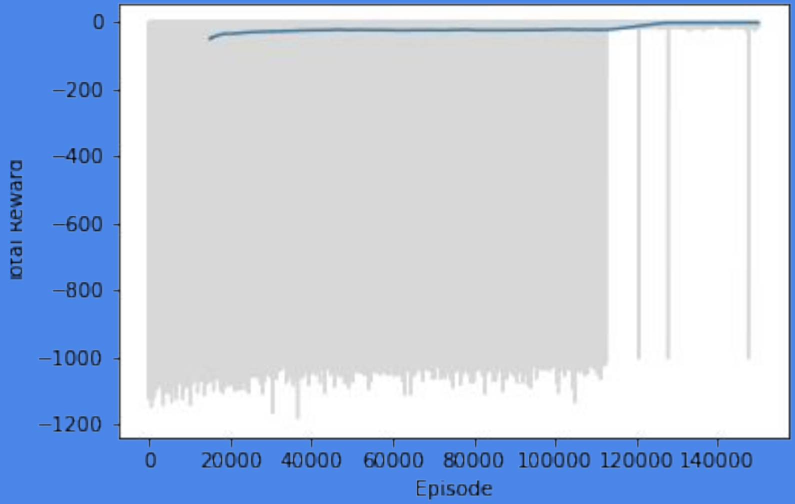


3° scenario (no way)

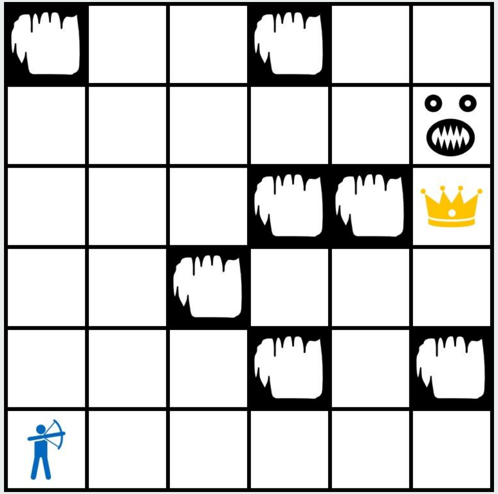


Training plot

(150’000 episodes)

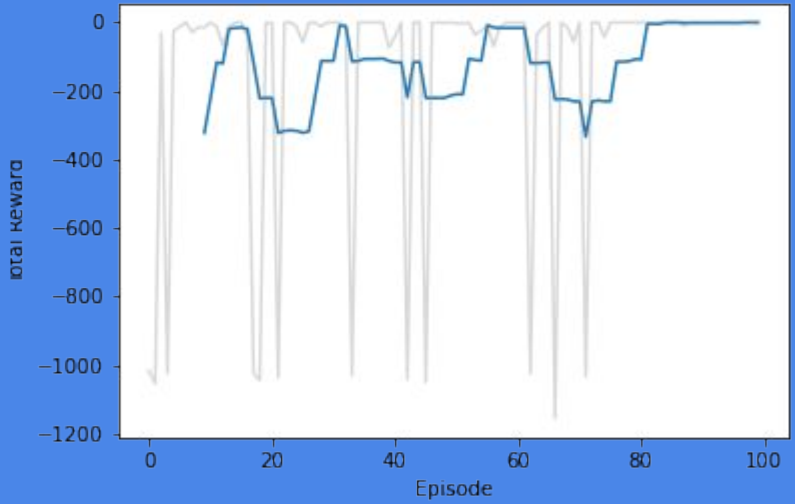


4° scenario
(very small training)

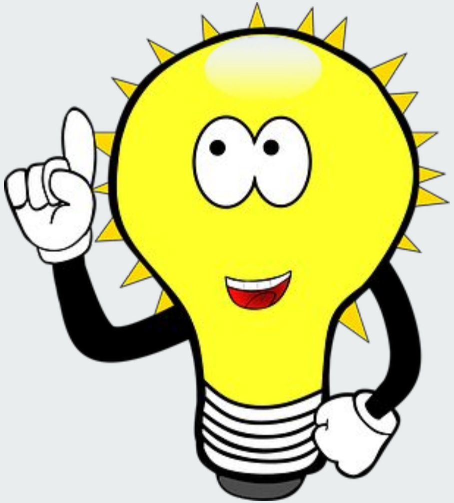


Training plot

(100 episodes)



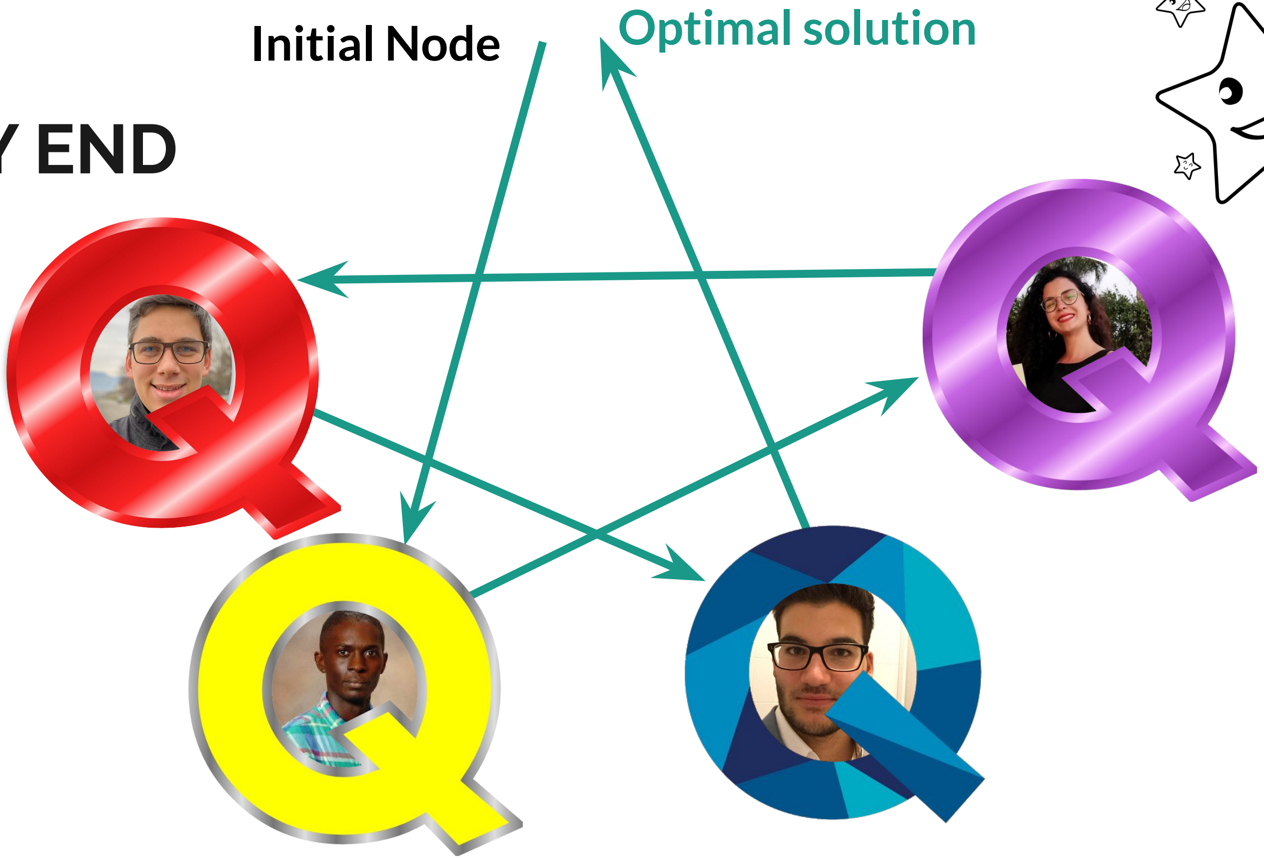
Conclusions

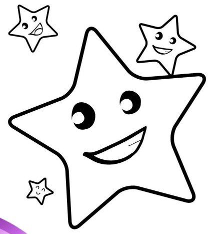


- An intermediate reward was essential to reduce the training
- An extensive exploration was needed to escape the local maximum
- We had to find a trade-off between the execution time and solution outcome:

  1,5 minutes execution for a 80% chance of getting the optimal solution

THE VERY END





Thank Q