Reinforcement Learning

Team: **Buggy Bunnies**
Catherine Nguyen
Chengjing Li
Hengchuan Zhang
Gayathri Talasila

Approach

- For the passive part of learning, we are using Q-learning.
  - Iterating through Q-values instead of just Values (value iteration).
  - Eliminating the need for a Transition and Reward function that we don't have yet, since we're learning Q values as we go

Data Structures

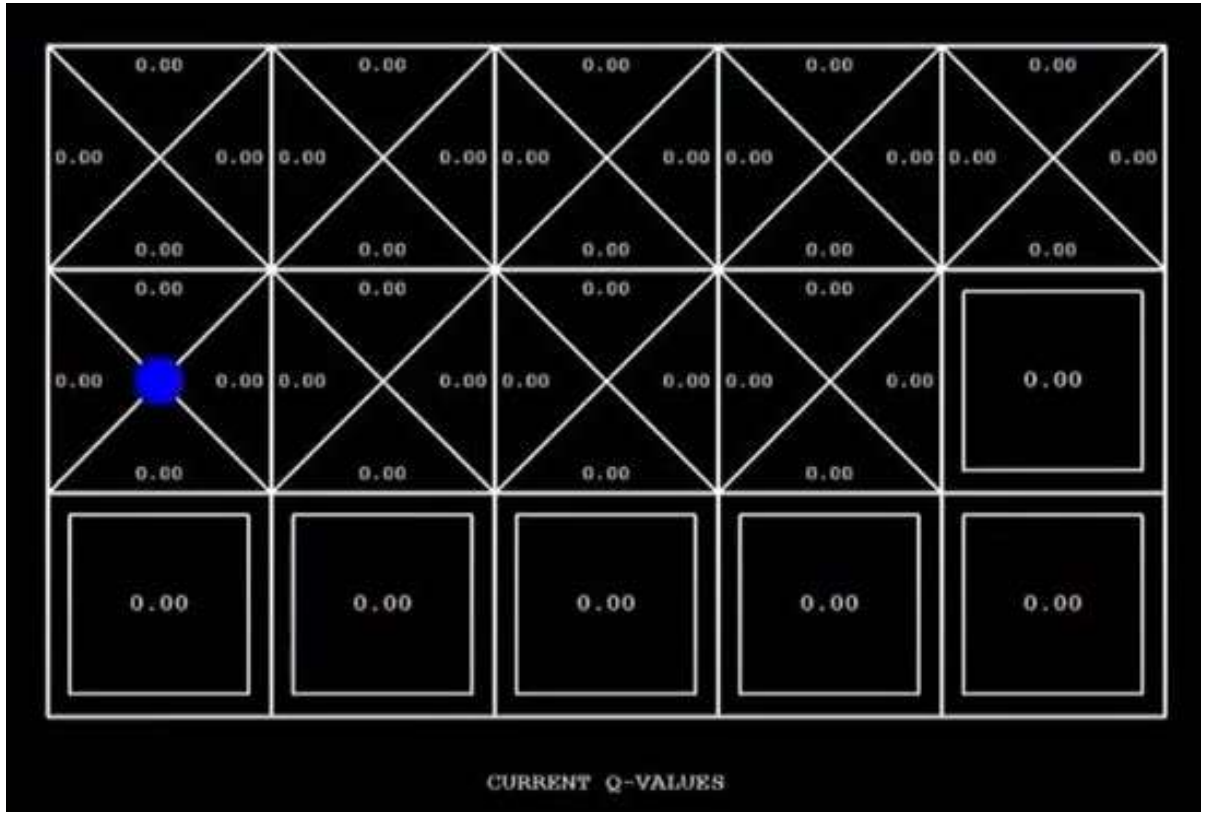
- Data Structure: [40x40x4] numpy arrays to store Q values.
  - Reflecting the N,S,E,W choices for the 40x40 grid worlds.
  - Since we are using numpy, the data will be saved locally as a .npy file.

```
def tableInitiate():
    return (np.zeros((40, 40, 4)))


def numToMove(num):
    if num == 0:
        return 'N'
    elif num == 1:
        return 'S'
    elif num == 2:
        return 'E'
    elif num == 3:
        return 'W'

    return False
```

Data Structures



What are the constants/parameters

- Discount factor γ (gamma) is used to calculate the penalty for staying alive. As of right now, we decided to use $\gamma = 0.95$
  - However, that number could change as we explore more of the worlds. Since the number will be dependent on the horizon, which will require more learning before we settle on the value.

What are the constants/parameters

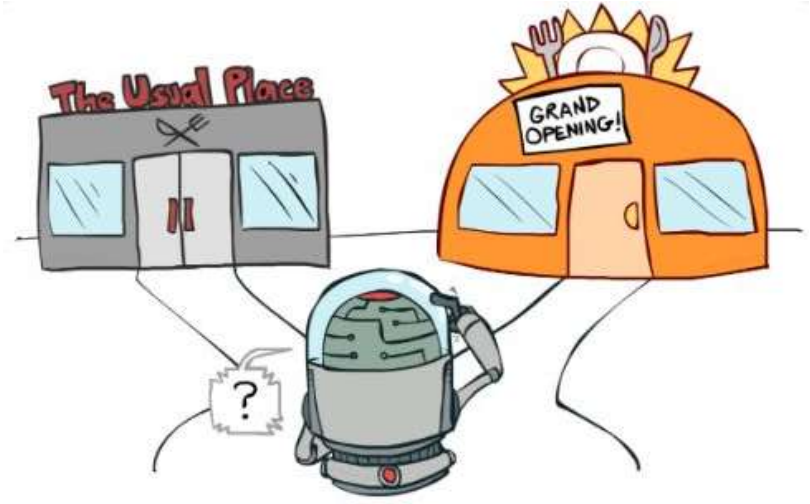
- Discount factor γ (gamma) = 0.95
- Learning/Decay rate α is used for updating our Q-values after each iteration. This number is needed since we don't have all the samples like MDP, our average is an *exponential moving average*.
  - Typically, α is decayed over time to help with convergence. Right now, we are arbitrarily setting the value of α = 0.5 for the first episode and decrease it from there.

What are the constants/parameters

- Discount factor γ (gamma) = 0.95
- Learning/Decay rate α = 0.5
- Epsilon ε for forcing exploration.

Forcing Exploration

- Once we have a scheme, do we want to continue to explore and risk penalty but potentially high reward? Or do we want to keep doing what we're doing well already?



Source: Berkeley AI Course

- ε dictates the probability that it will either explore or exploit.
- But it will decrease over time.

What are the constants/parameters

- Discount factor γ (gamma) = 0.95
- Learning/Decay rate α = 0.5
- Starting epsilon ε = 0.9