RL presentation:

1. State space S: the height of grid multiply the width of grid.
2. Action space A: UP DOWN LEFT RIGHT STAY
3. Hyperparameters settings:
   1. Not learning: alpha = 0
   2. High variance but fast learning: 0.7 < alpha < 1
   3. low variance and high long-term return: 0 < alpha < 0.3, 0.7 < gamma < 1
   4. high variance and high long-term return: 0.7 < alpha < 1, 0.7 < gamma < 1
4. Given the 3\_2\_3.yml reward structure and initial position of jelly fish/king fish/diver, what is the value of the long term return of the optimal policy?
   1. Gamma \* 50
   2. Gamma \* (-2 + gamma \* 50) = gamma \* (a - 2)
   3. Gamma \* (-2 + gamma \* (-2 + gamma \* 50)) = gamma \* (b – 2)
   4. ……
5. Other RL algorithms:
   1. MDP assumption but don’t know T(environment transition model) or R(reward)
   2. Bellman equation
   3. Offline solution: known T and R, offline learning: need to estimate T and R
   4. Model based learning(approximate T and R from data value and policy iteration) and model free learning (passive (temporal difference learning)and active RL(Q learning UCB))
   5. MODEL-BASED Advantage: make good use of data, dis: requires building the actual MDP model, intractable if state space is too large
   6. MODEL-FREE: learn while optimizing policy
   7. Passive reinforcement learning
   8. Value iteration: computing utility of states, iterate until policy has converged
   9. Policy iteration: computing the value of policy evaluation V(s), iterate until policy does not changed any more