**Resources:-**<https://towardsdatascience.com/reinforcement-learning-implement-grid-world-from-scratch-c5963765ebff>  
<https://www.mdpi.com/1424-8220/23/7/3625>

**Reinforcement Learning**Area of machine learning where you or an agent learns to maximize the rewards by taking appropriate actions based upon observations from the environments and learn to optimize it.

**Markov Decision Processes (MDP) (sequential decision making)**

Sequential Decision Making: - Some intelligent agent takes an action in the world, gets a reward based on the action and observes the new world/state.

Key goal is to maximize the expected future reward.

Requires balancing between immediate and long-term rewards.

May require strategic behavior to achieve high rewards.

Components:- Agent, Environment, States, Actions and Rewards

*What is?*

1. Decision maker( agent) —--> Interacts with environment.
2. At each time step the agent observes the environment and gets a representation.
3. Agent selects an action to take.
4. Environment transitions to a new state based on action
5. Agent is given reward as a consequence of the action

This process happens sequentially. *The goal is to maximize the reward. (*THE GOAL IS NOT MAXIMIZE THE IMMEDIATE REWARD BUT THE CUMULATIVE REWARD)

*S* set of States,  *A* set of Actions, *R* set of rewards. At each time step *t* = 0,1,2,3. Agent gets representation *S*t from *S.* Based on this he selects *A*t from *A.* Time is then incremented to next time step *t+1.*  Environment is in the new state *S*t +1 from *S* . The agent receives reward *Rt +1*  based upon previous action *At taken from state St.*

Hence we can write as **f(St, At) = Rt+1**

Probability distribution depends on the preceding state and action.

Deterministic: - single observation & reward. Next state is known. Robotics, the world which we live in.

Stochastic: - 1 action has many potential observations and rewards.

**Expected Return**

This is what drives an agent in an MDP.  
If *G* is the Expected Return then *GT = Rt+1 , Rt+2, Rt+3…..RT*Where T is the final time step. Time steps can be looked at as subsequences in an episode.

A new episode is always independent of the previous episode.

BUTTTTTTT in a continuing episodic task there is no final time step T which makes the above definition useless as it the final time step is *infinity.* So now the definition becomes.   
**It is the agent's goal to maximize the expected discounted return of rewards.**

*Discount Factor:-* Denoted by symbol gamma is a value between 0 and 1 which determines the importance of future rewards to immediate reward i.e how much the agent values future reward to immediate reward. Value of 0 means only immediate reward and 1 means both equally important.

*Rt+1 + gamma(Gt+1)*

Where Gt is *Rt+1 + gamma(Rt+2) + gamma^2(Rt+3).....*Gamma is raised to the power of the current timestep value.

**Policies and Value Functions**

*Policies:-* What's the probability that agent will select a specific action from a specific state?

Value:- How good is a specific action or a specific state for the agent?

State Value function:- how good any given state is under given policy. It gives us the value of a state under policy.

Action Value:- how good it is for an agent to take any given action from a given state while following policy. It gives us value of an action