Reinforcement learning is all about the interaction between agent and the environment. Agent is the unit who makes decision, usually be considered as in a form of robot. The environment is made up with rewards motivating the agent to make its decisions on its actions, it can be anywhere in our real world.

In bandit problems, the main focus is greedy selecting each action for immediate reward, which is not the best selection in many cases. Finite Markov Decision Process (MDP) differs from bandit by evaluating each action at a specific state, and the reward from a sequence of actions to determine a final decision, sacrificing immediate rewards. The rewards in bandit problems are fixed at a mean following a normal distribution, while the rewards in MDP are different according to a state. An example for MDP could be like entering a building from front door allows you to be on time but back door will make you late.

An agent's ultimate goal is to maximize the reward in long-term. To determine the optimal set of actions, one might need to conduct numerous experiments (episodes). Reinforcement learning algorithms are mainly about estimating value functions, the performance of the agent under current state. The probability of making each decision under a state is defined as policy. Bellman's Equation summarizes the relationship between policy and the value function.

However, for real world problems that people are interested today, optimal solution rarely exists. One factor is because of extensive computational costs, simply solving Bellman's Equation still does not output the optimal policy for most cases. Another factor is because it requires large amount of memories to store the data generated from each episode.