In the previous study, we talked about function approximation that optimize the value of a parameterized vector, w, that predicts the value function which is then compared with a target value from the output of incremental mean methods, Monte Carlo or TD Learning. Of course, the target is partially fixed, to avoid hitting a moving target. The output of the value function from the approximation was then used to determine a policy, either greedily or E-greedily. That was the extent of Deep Q Learning.

Now we come to talk about Policy based learning. Direct improvement of policy rather than the sequence of value functions and control. By parameterize-sing the policy distribution itself i.e. a the function approximator will be used to approximate the policy, preferred stochastic policy over deterministic because of state aliasing and equilibrium, and the gradient will be followed up towards the maximum, hopefully global.