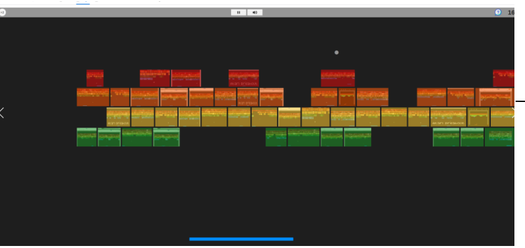
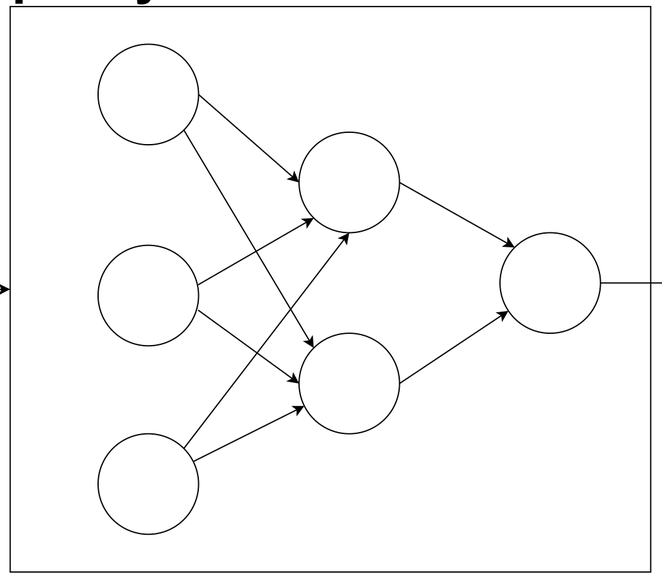
Policy Gradient: Directly Optimize the policy

$\pi(s)$

$$\sum_{a_i \in \mathcal{A}} P(a_i|s) = 1$$



State, s



$P(a_1|s) = 0.9$

$P(a_2|s) = 0.1$

$P(a_3|s) = 0.0$

$\pi(s) \sim P(a|s)$

Input

Agent

Output