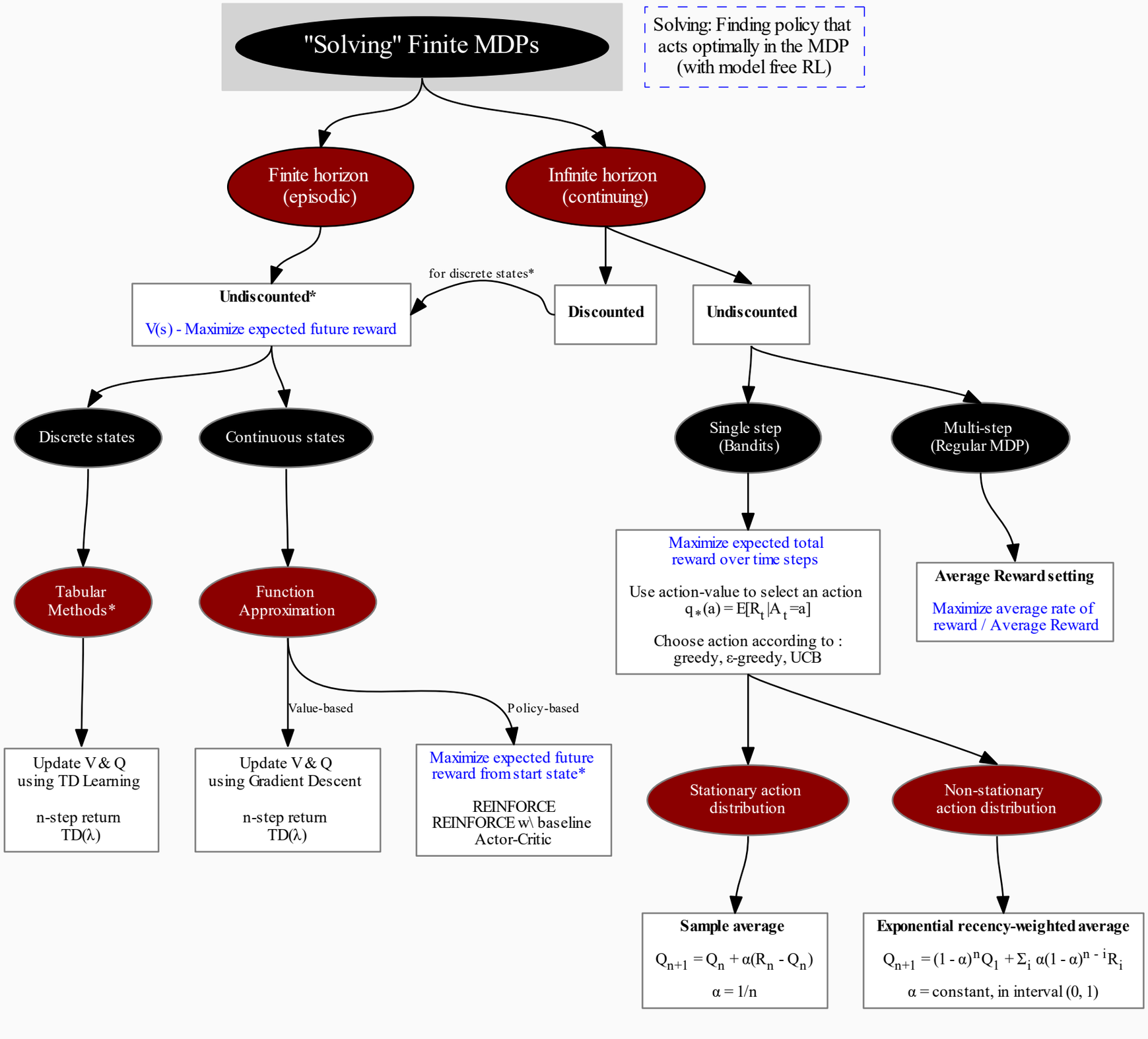
"Solving" Finite MDPs
Solving: Finding policy that acts optimally in the MDP (with model free RL)
Finite horizon (episodic)
Infinite horizon (continuing)
for discrete states*
Undiscounted*
V(s) - Maximize expected future reward
Discounted
Undiscounted
Discrete states
Continuous states
Single step (Bandits)
Multi-step (Regular MDP)
Tabular Methods*
Function Approximation
Maximize expected total reward over time steps
Use action-value to select an action
$q_*(a) = E[R_t | A_t = a]$
Choose action according to : greedy, ε-greedy, UCB
Average Reward setting
Maximize average rate of reward / Average Reward
Value-based
Policy-based
Update V & Q using TD Learning
n-step return
TD(λ)
Update V & Q using Gradient Descent
n-step return
TD(λ)
Maximize expected future reward from start state*
REINFORCE
REINFORCE w\ baseline
Actor-Critic
Stationary action distribution
Non-stationary action distribution
Sample average
$Q_{n+1} = Q_n + \alpha(R_n - Q_n)$
$\alpha = 1/n$
Exponential recency-weighted average
$Q_{n+1} = (1 - \alpha)^n Q_1 + \Sigma_i \alpha(1 - \alpha)^{n-i} R_i$
α = constant, in interval (0, 1)