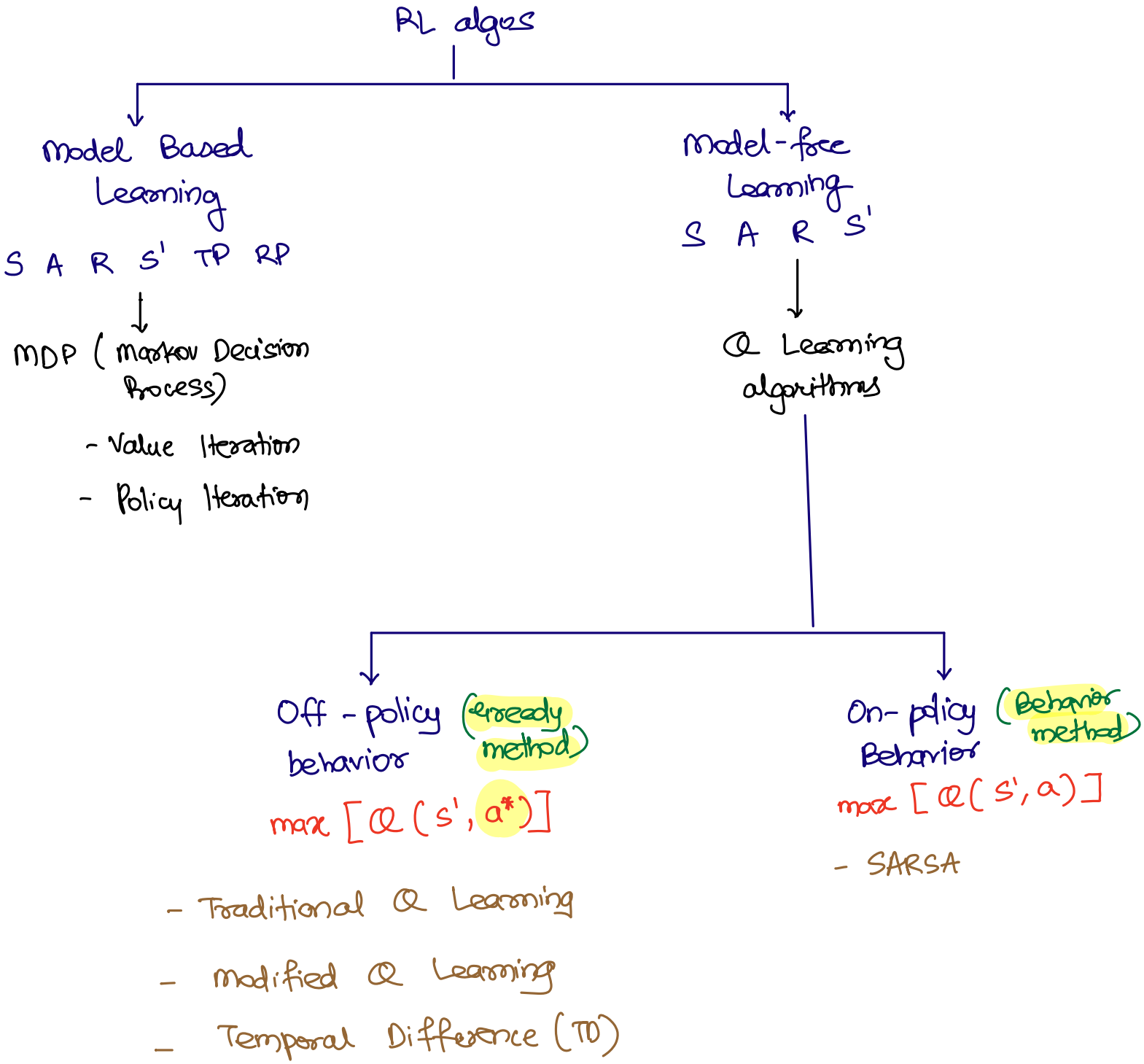
Reinforcement Learning

Trainer : Prashant N



Traditional Q learning*

$$Q(S, A) = IR + (gamma * max(S', a^*))$$

- $Q(S, A)$: S = current state, A = current action
- IR: immediate reward
- gamma: discount factor (hyperparameter)
- $max(S', a^*)$: S' = next state, a^* = all valid actions

Q value of next state for all possible action and extract the max q value.

Modified Q learning

$$Q(S, A) = Q(S, A) + \alpha \left[\underbrace{R(S,A) + \gamma \max Q'(S', a^*)}_{\text{q target}} - \underbrace{Q(S,A)}_{\text{q-predict}} \right]$$

- $Q(S, A)$ (left): current Q value from Q table
- α: learning rate
- $R(S,A)$: IR
- γ: discount factor
- $\max Q'(S', a^*)$: max q value of next state from all possible actions
- $Q(S,A)$ (right): current Q value from Q table.

Temporal Difference (TD)

$$Q(s, a) = Q(s,a) * (1 - \alpha) + \alpha * \left[\underbrace{R(s,a)}_{IR} + \gamma \max [Q(s', a^*)] \right]$$

- $Q(s, a)$ (left): current Q value from initialized Q Table
- α: learning rate
- γ: discount factor