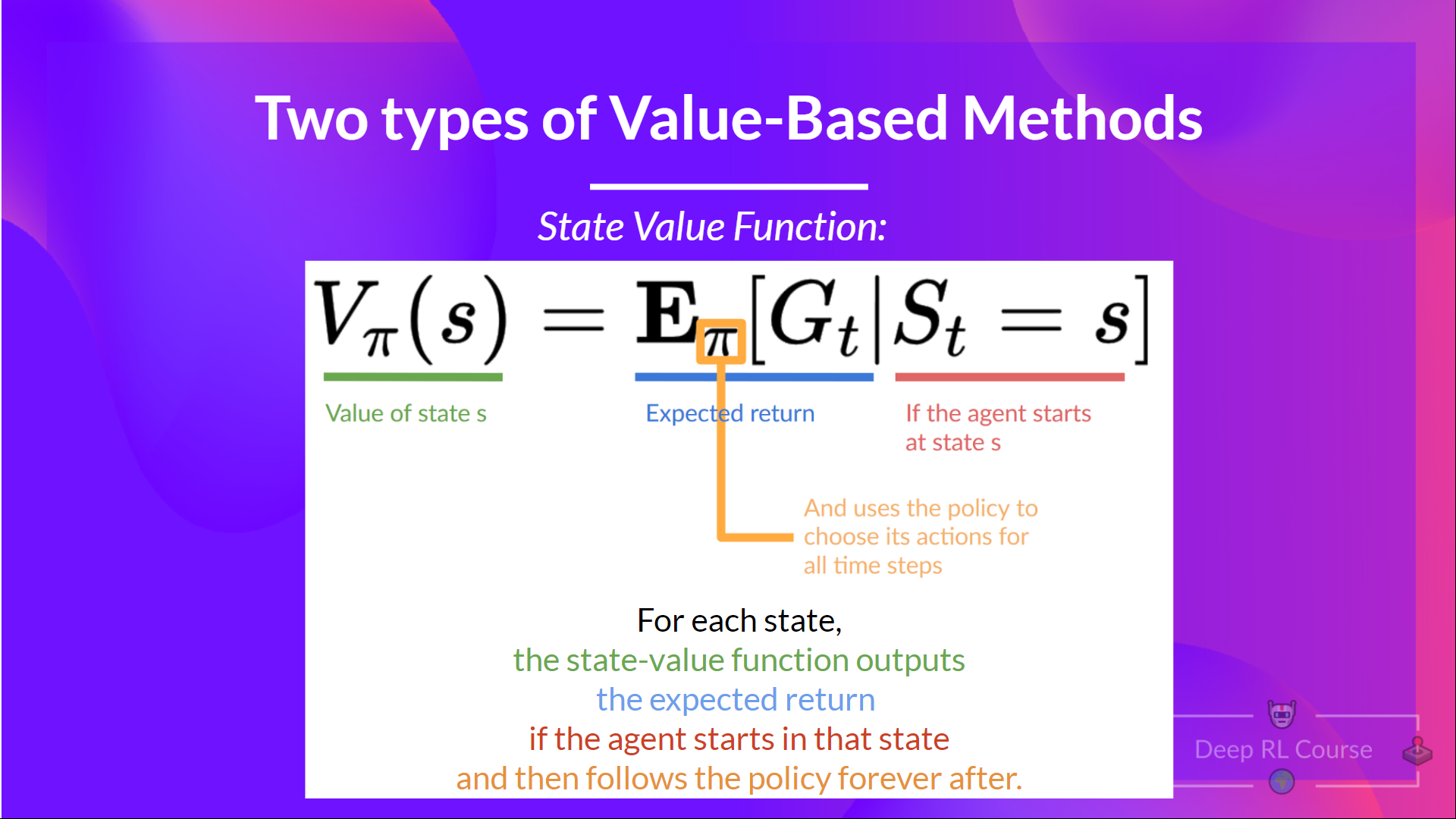
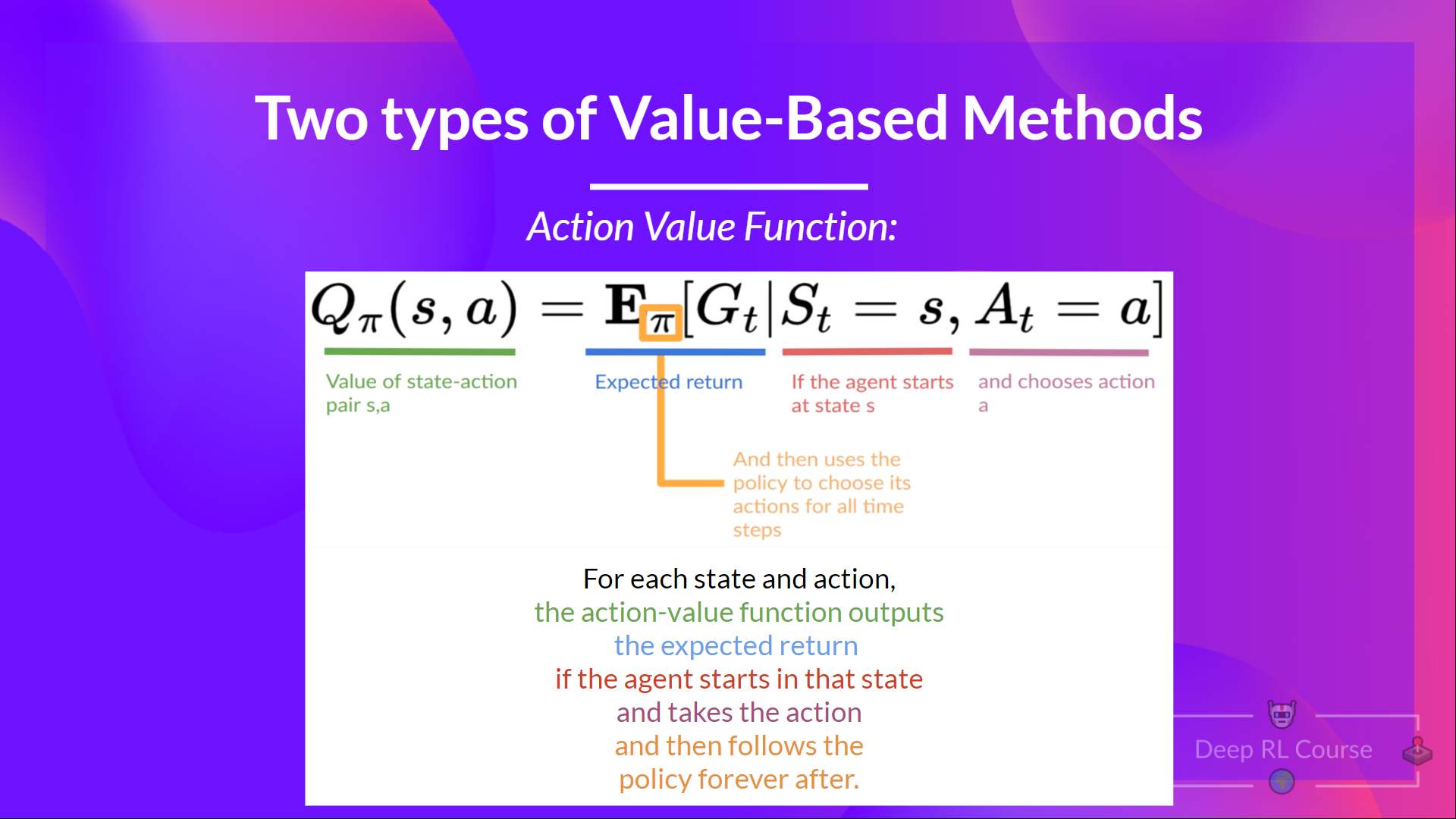
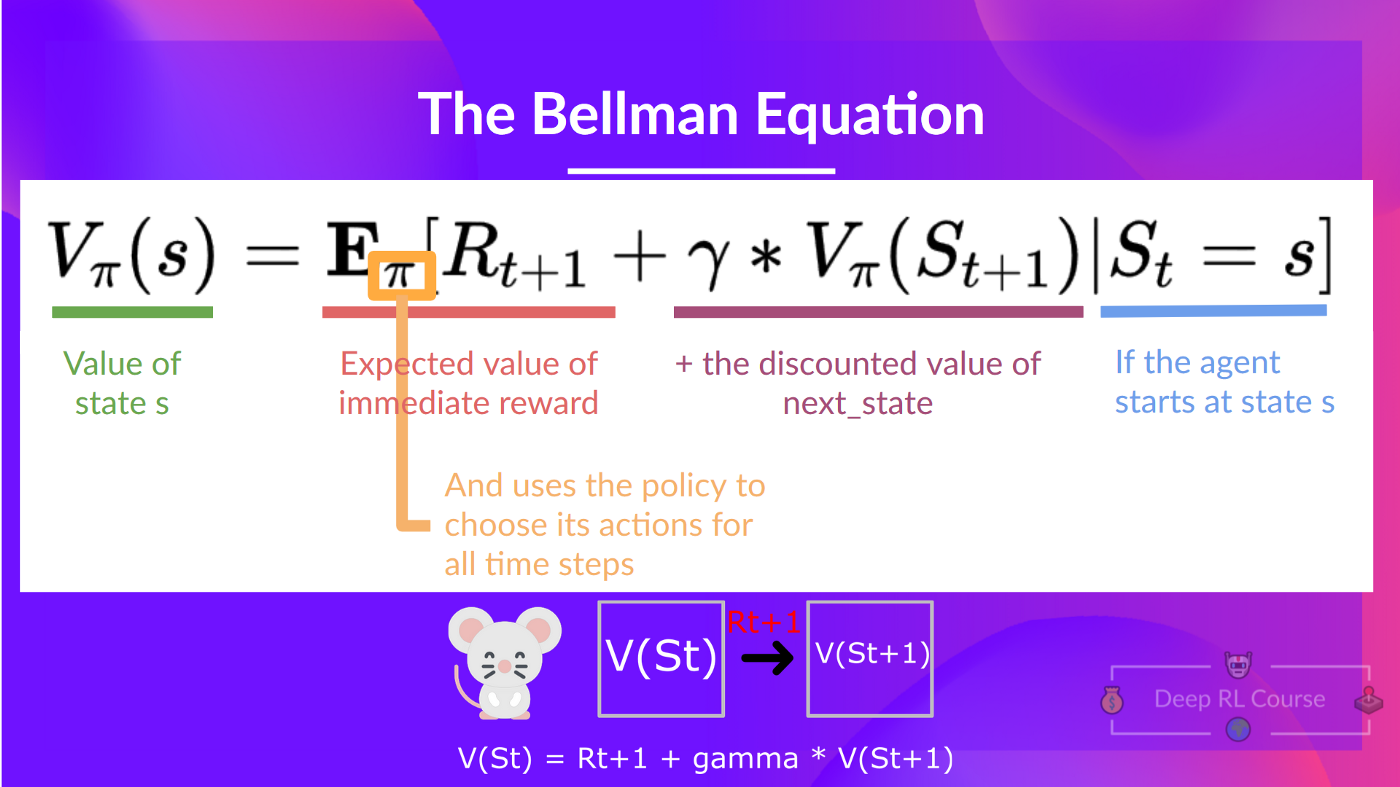
More On Policy and Value Functions

We have two types of value-based functions:

1. **State-value function**: outputs the expected return if the agent starts at a given state and acts according to the policy forever after.
2. **Action-value function**: outputs the expected return if the agent starts in each state, takes a given action at that state and then acts accordingly to the policy forever after.





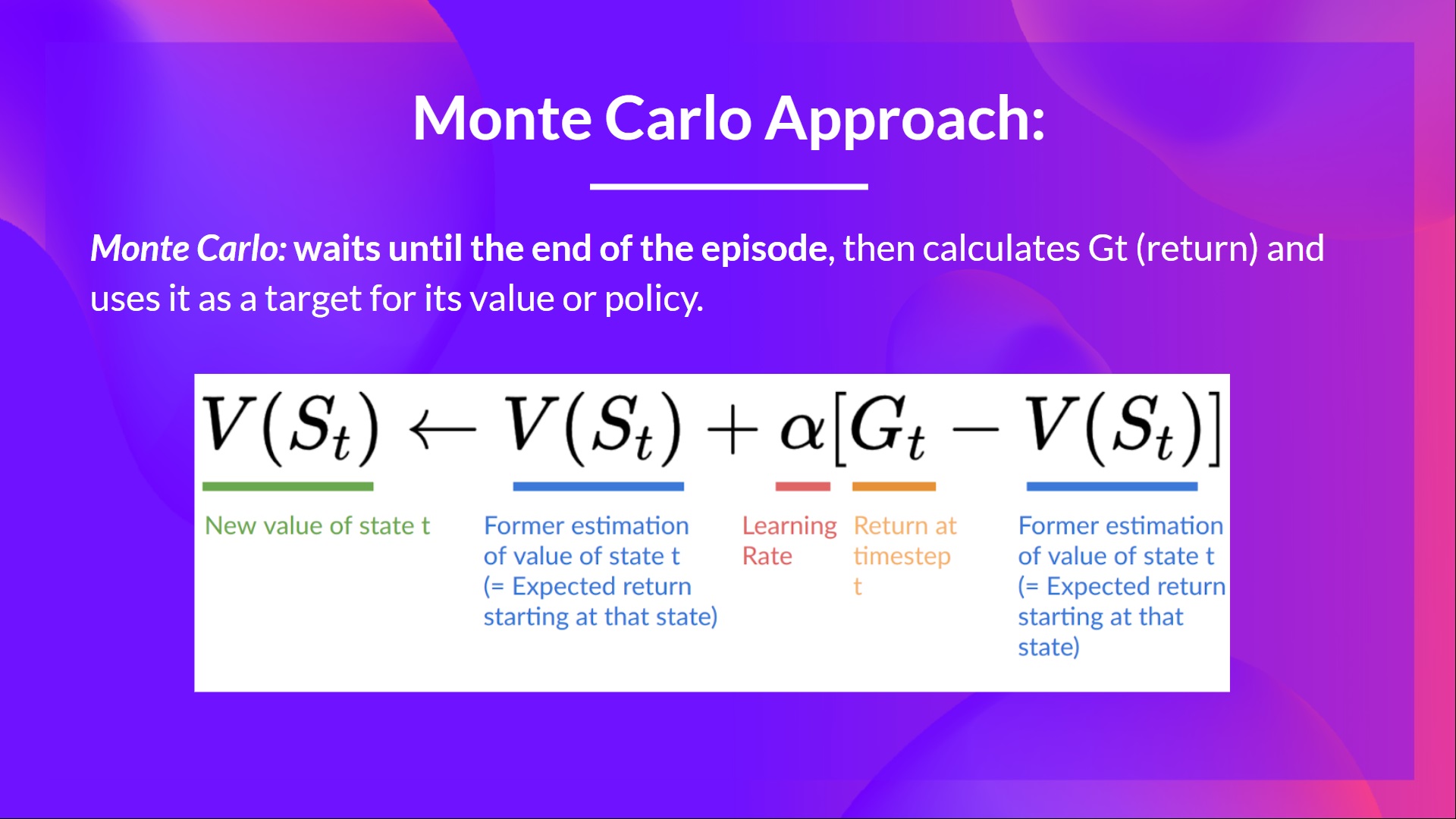


**Monte Carlo vs Temporal Difference Learning**

Monte Carlo and Temporal Difference Learning are two different strategies on how to train our value function or our policy function. Both use experience to solve the RL problem.

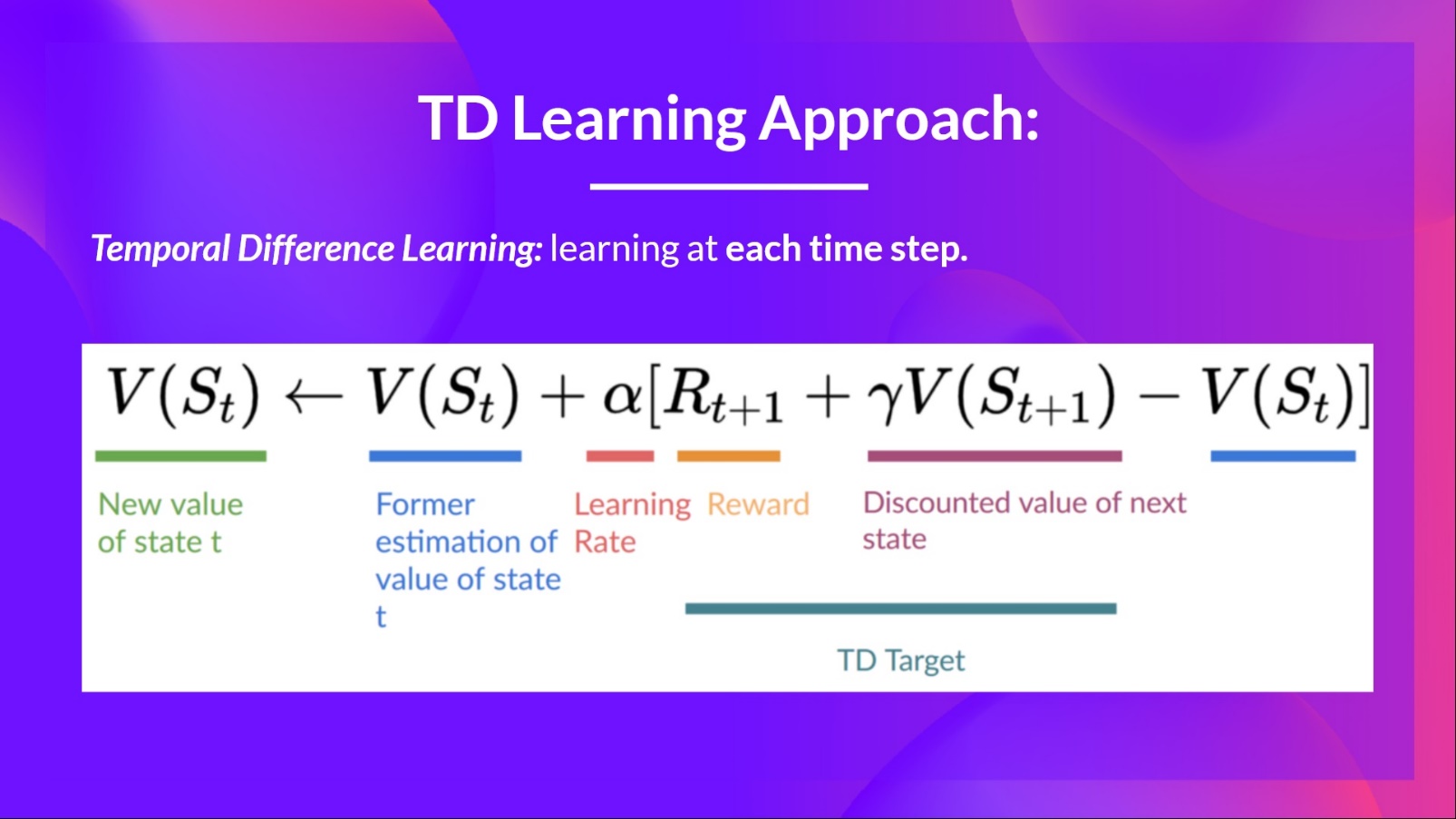
**Monte Carlo:**

* Monte Carlo waits until the end of the episode, calculates Gt (return) and uses it as a target for updating V(St).
* It requires a complete episode of interaction before updating our value function.



**Temporal Difference:**

* Temporal Difference, on the other hand, waits for only one interaction (one step) St+1 to form a TD target and update V(St) using Rt+1 and gamma\*V(St+1).
* The idea with TD is to update the V(St) at each step.
* Because we did not experience an entire episode, we do not have Gt (expected return). Instead, we estimate Gt by adding Rt+1 and the discounted value of the following state.



**To summarize:**

* With Monte Carlo, we update the value function from a complete episode, and so we use the actual accurate discounted return of this episode.
* With TD Learning, we update the value function from a step, and we replace Gt, which we do not know, with an estimated return called the TD target.