1 MDP

A reward R_t is a scalar feedback signal: indicates how well agent is doing at step t.
The agents job is to select actions to maximize (expected) cumulative reward!!

- Actions may have long term consequences;
- Rewards may be delayed;
- It may be better to sacrifice immediate reward to gain more long-term reward.

The agent state S_t^a is the agent's internal representation, i.e. whatever information it uses to pick the next action. Formally, state is a function of the history $S_t^a = f(H_t)$, where $H_t = O_1, R_1, A_1, ..., A_{t-1}, O_t, R_t$, i.e. all observable variables up to time t.

An agent state (a.k.a. Markov state) contains all useful information from the history. S_t is Markov iff

$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_t, S_{t-1}, ...S_1]$$

- The future is independent of the past, given present.
- once the state is known the history can be thrown away.
- The state is sufficient statistics of the history.

(???) POMDP (partially observable) vs MDP

An RL agent may include one or more of these components:

- Policy: agents behavior function
- Value function: how good is each state and/or action
- Model: agents representation of the environment