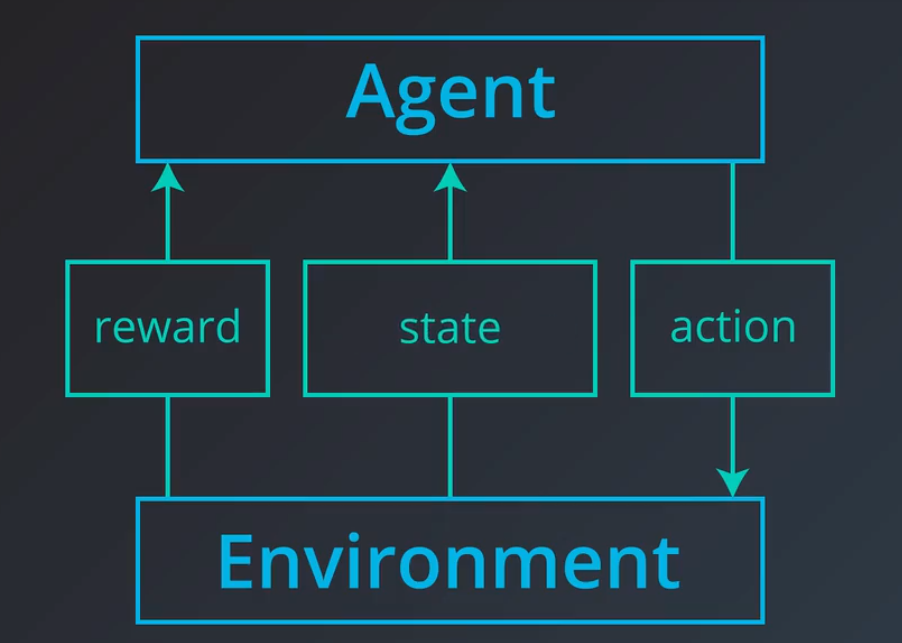
# 设置

## 强化学习

智能体通过试错 学习如何在环境中完成各种动作并最大化奖励。例如小狗，我们用它来比喻智能体

RL（强化学习）框架是指 智能体 学习如何 与环境互动。

## 原理



1、agent 接收 state，然后 agent 给出 相应 action。

2、environment接收 action,然后对 agent做出如下响应：

1、environment 向 agent 呈现 新的state

2、environment 向 agent 反馈一个reward

3、作为回应，Agent体必须选择一个Action.

(action表示Agent是否对环境 做出了正确的响应)

4、循环此流程

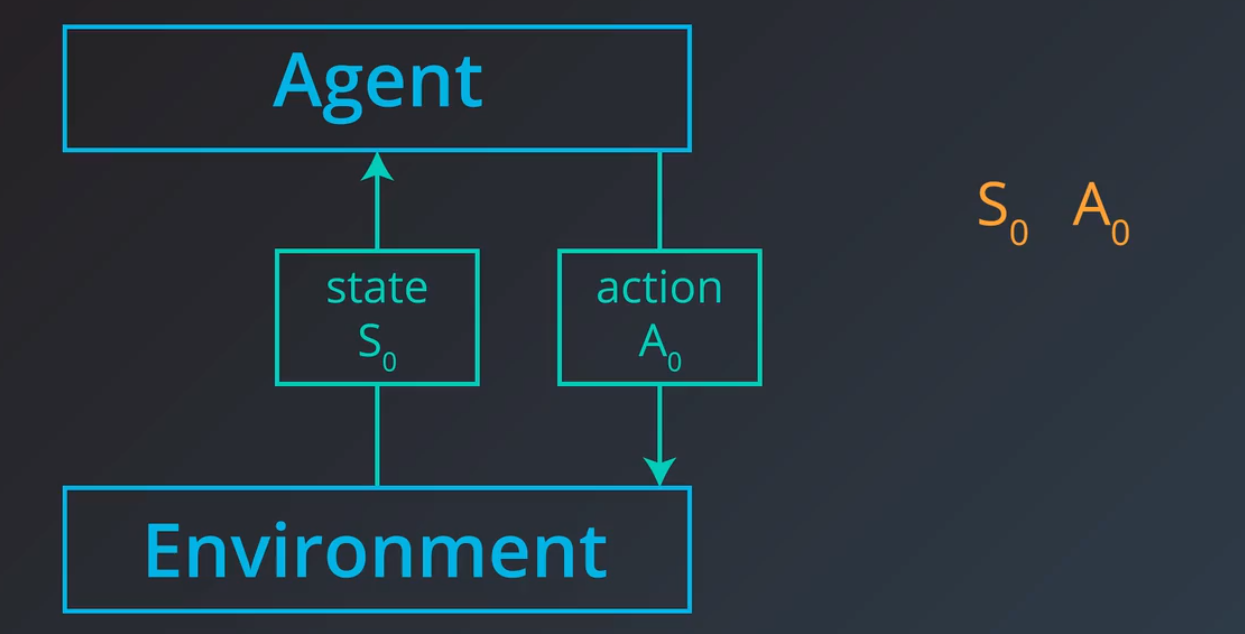
注意：

在每个时间步 环境 **都向** 智能体发送一个 观察结果和奖励

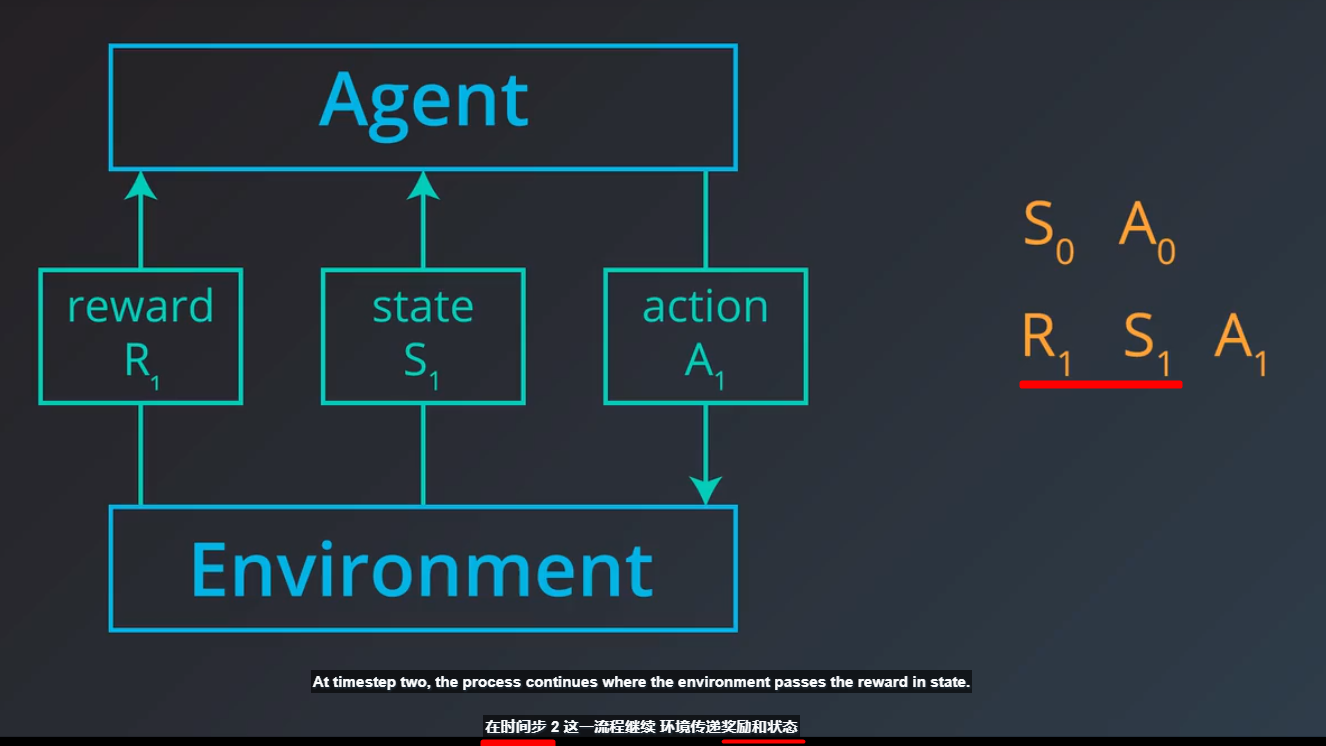
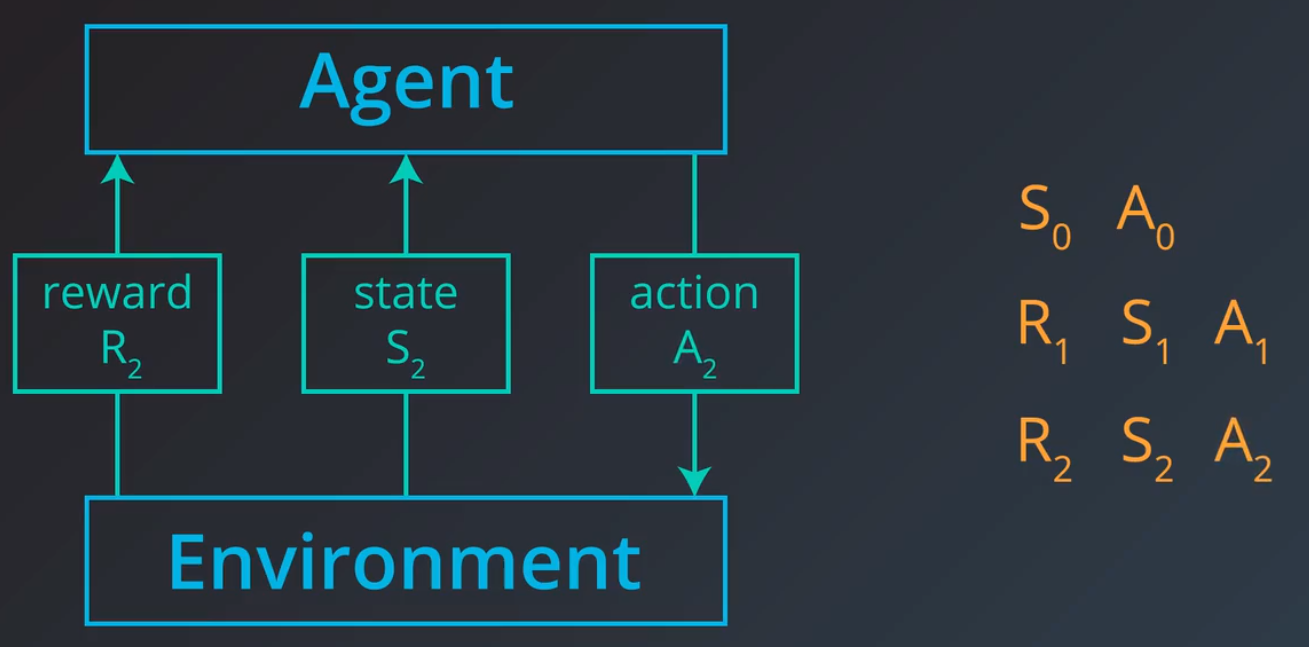
作为回应 智能体 必须选择 一个action

## 流程图

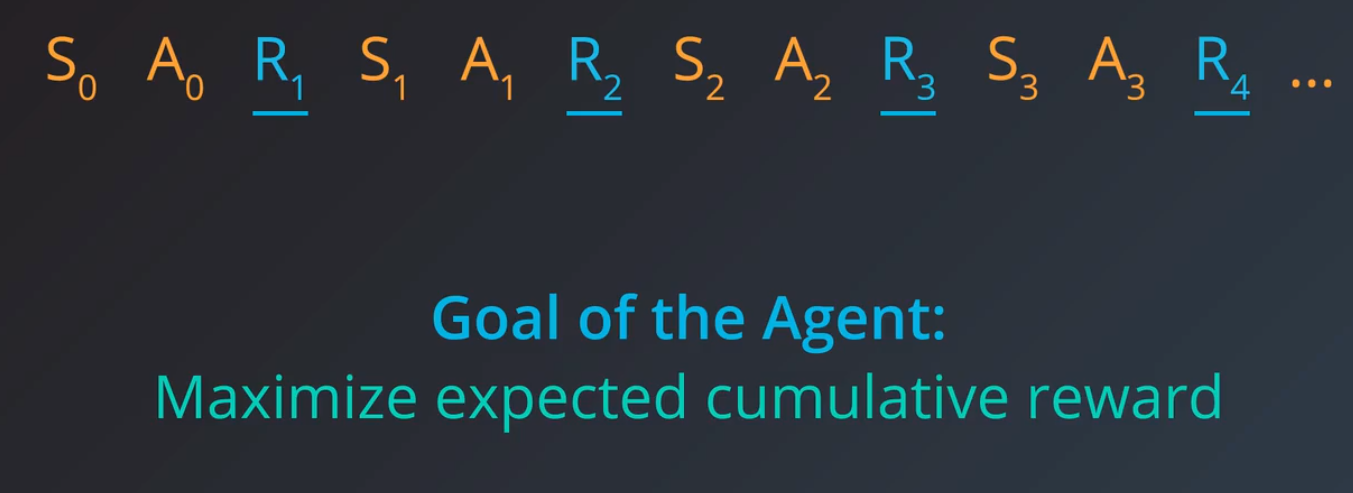
S0



S1、S2

## 目标



***最大化 期望累积奖励***。也就是选择使 累积奖励 最高的Action

智能体只有通过与环境互动 才能实现这一目标，因为在每个时间步，环境都决定智能体会接收多少奖励，因此智能体必须遵守环境规则。

## 总结

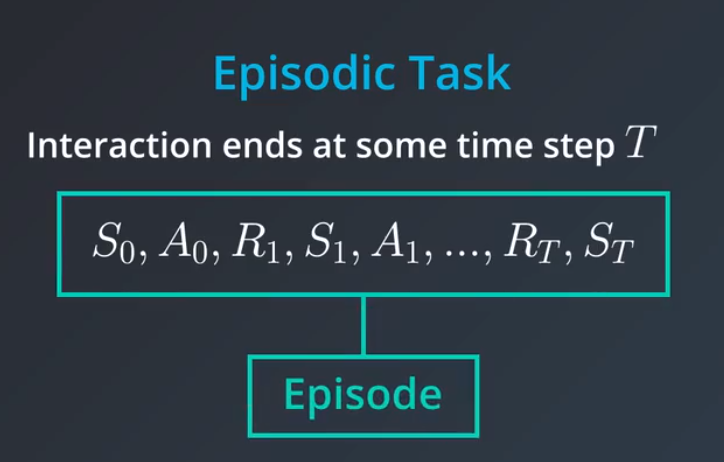
这些只是解决现实问题的数学模型，你需要**指定状态 动作和奖励** 并且需要**制定环境规则**

# 阶段任务与连续任务

## 阶段性任务

具有清晰结束点的任务称之为 **阶段性任务（Episodic Task）**

这里我们将 从头到尾的 一系列完整互动称之为 **阶段(Episodic)**

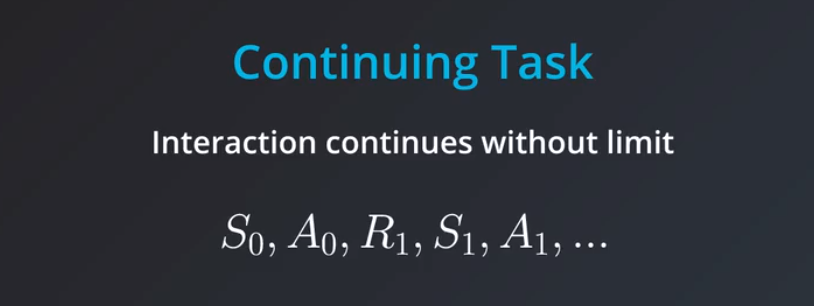


当某个阶段结束时,智能体看看它收到的奖励总量 并判断自己的表现如何,然后能够从头开始 就好像完全第一次接触同一环境.

但是掌握了前生了解的一些知识,这样 随着它重生多次之后,智能体能够做出越来越好的决策

## 连续性任务

一直持续下去的任务称之为 连续性任务(Continuing Task)



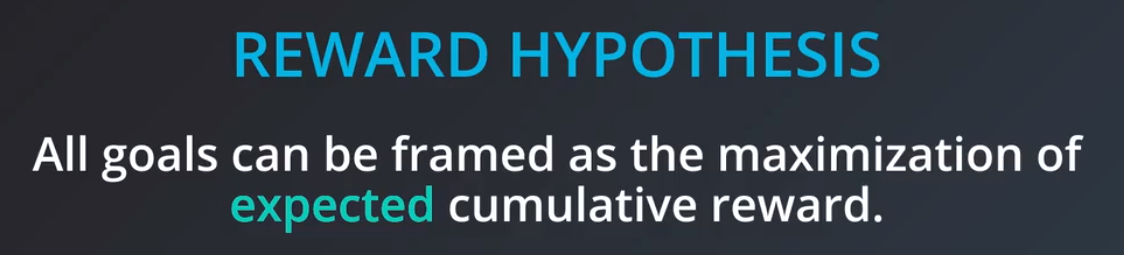
例如 根据金融市场买入和卖出股票的算法 最适合建模为连续性任务智能体

此时智能体会一直存活下去.

# 目标和奖励

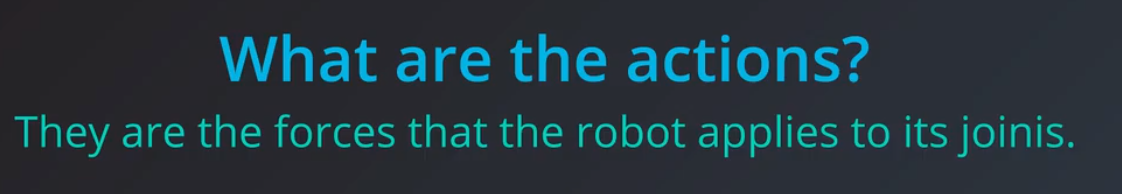
## 奖励假设

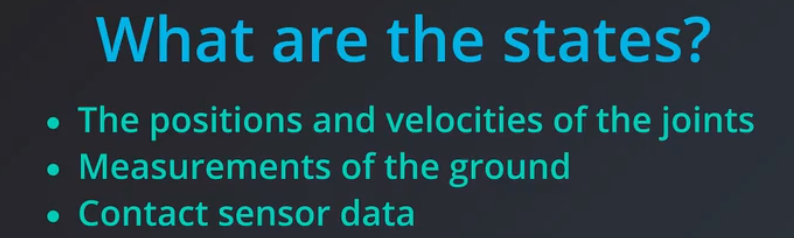
即智能体的目标始终可以描述为 最大化期望累积奖励

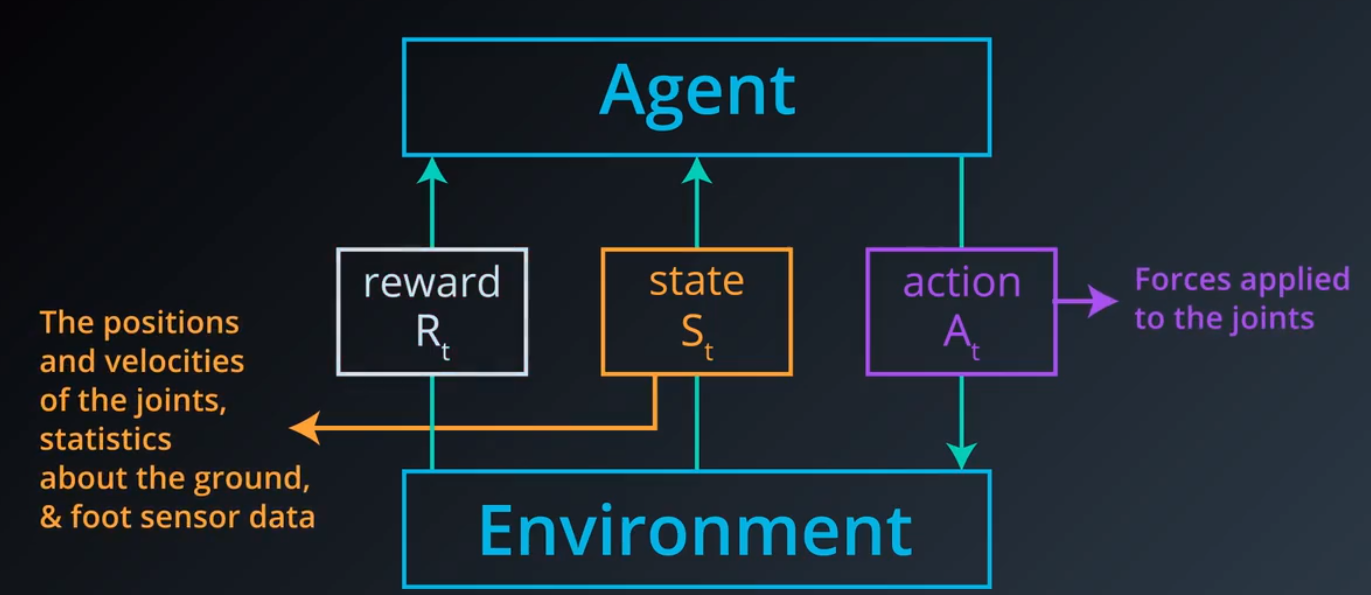


机器人行走

## 目标和奖励

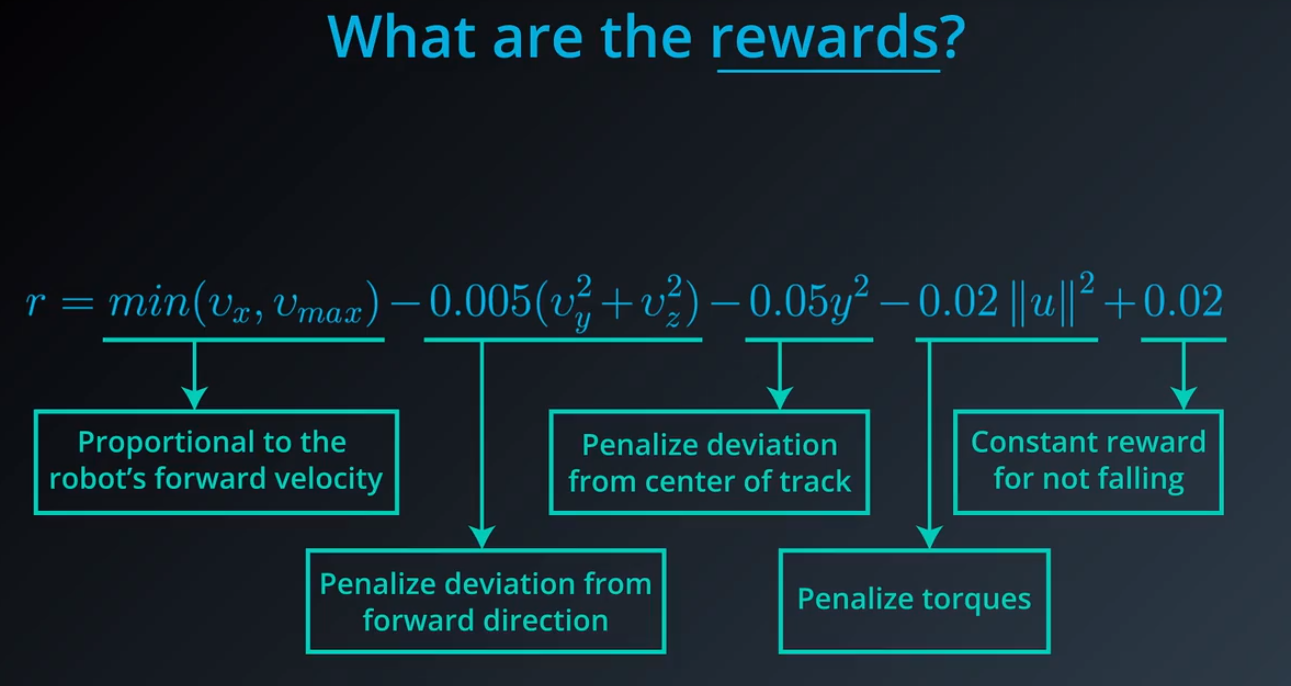




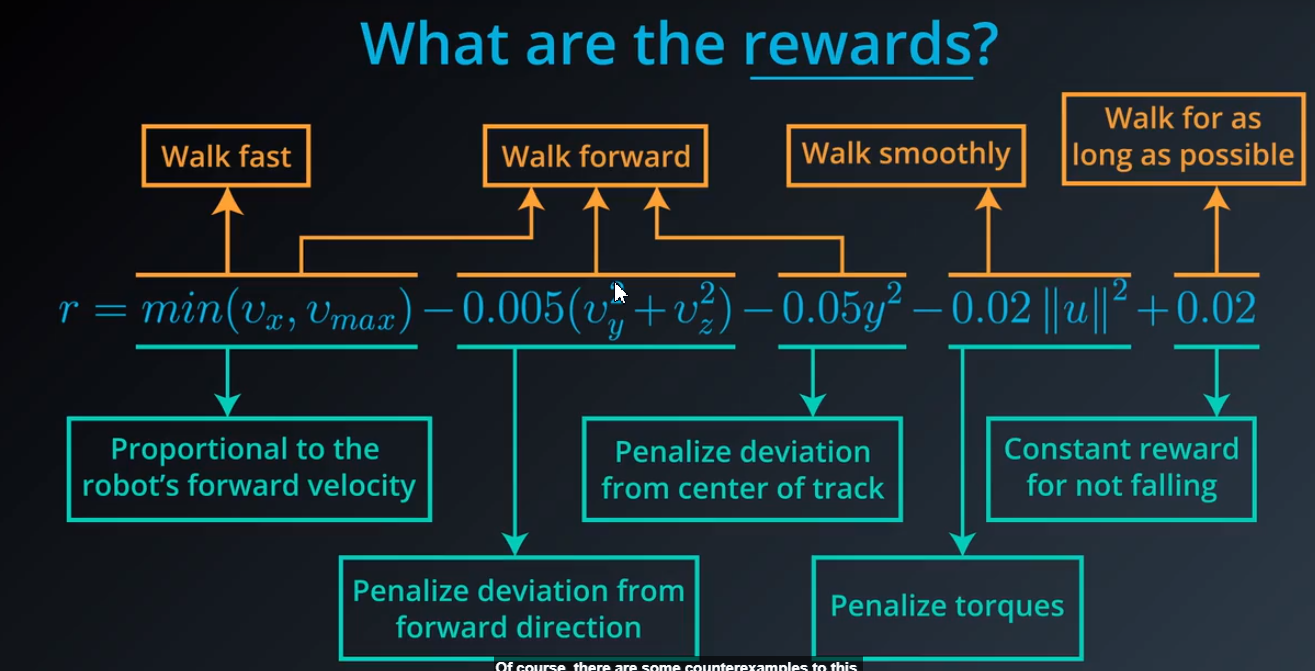


**动作是 运用到机器人关节，以便行走的力**

**状态是 提供给智能体的环境信息 使其能够选择合理的动作**

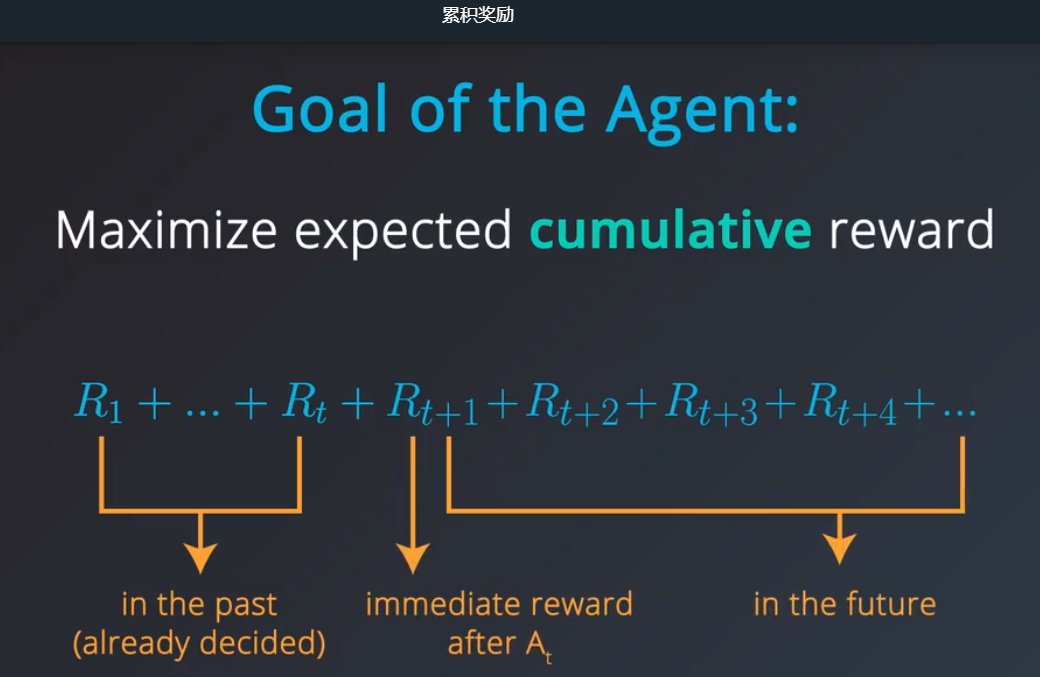


考虑因素

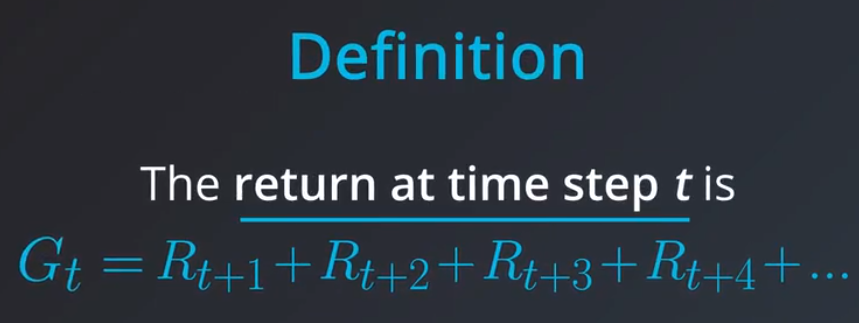


# 累计奖励（G）

时间步之前的奖励已经确定了，因为它们是过去的奖励，只有**未来的奖励**受到智能体的控制。



我们将后续时间步的奖励之和称之为 ***回报 表示为 G***



智能体无法完全肯定地 预测未来奖励会怎样，因此它**必须依赖于预测或估算**。

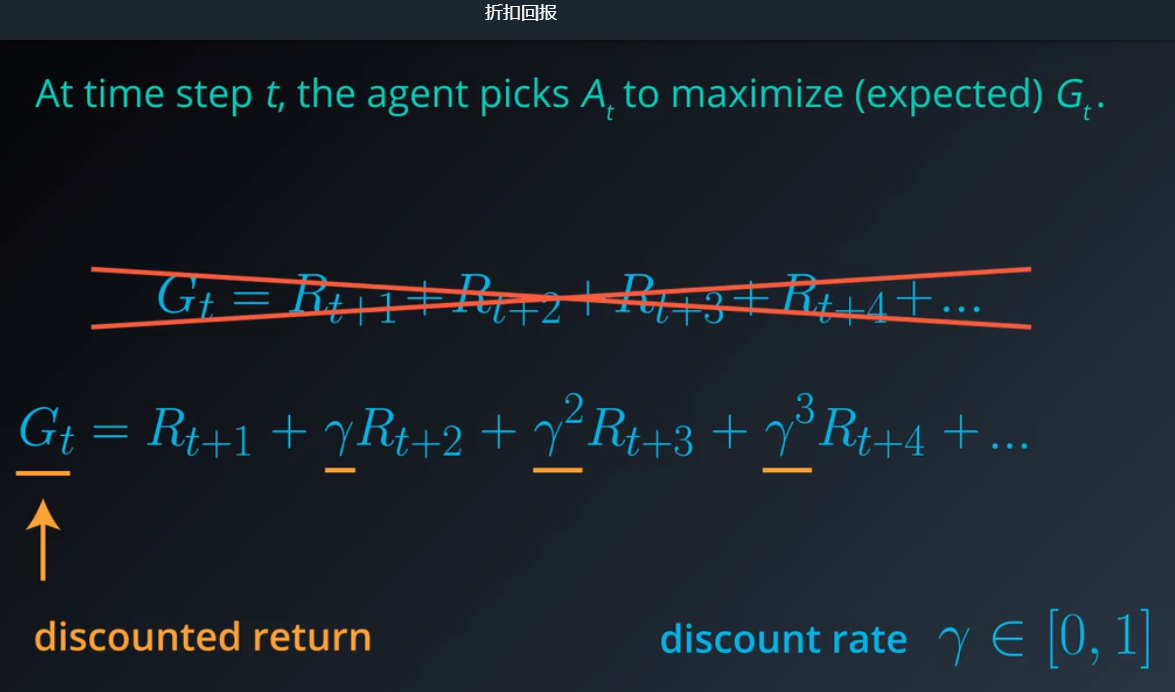
# 折扣回报

通常 与很远的未来相比，我们更渴望知道不久将来的状况。那么应该将不久将来的奖励权重设置更高，因为这些奖励更容易预测

折扣回报（discounted return）

**折扣率γ**，它始终是 0 到 1 之间的值

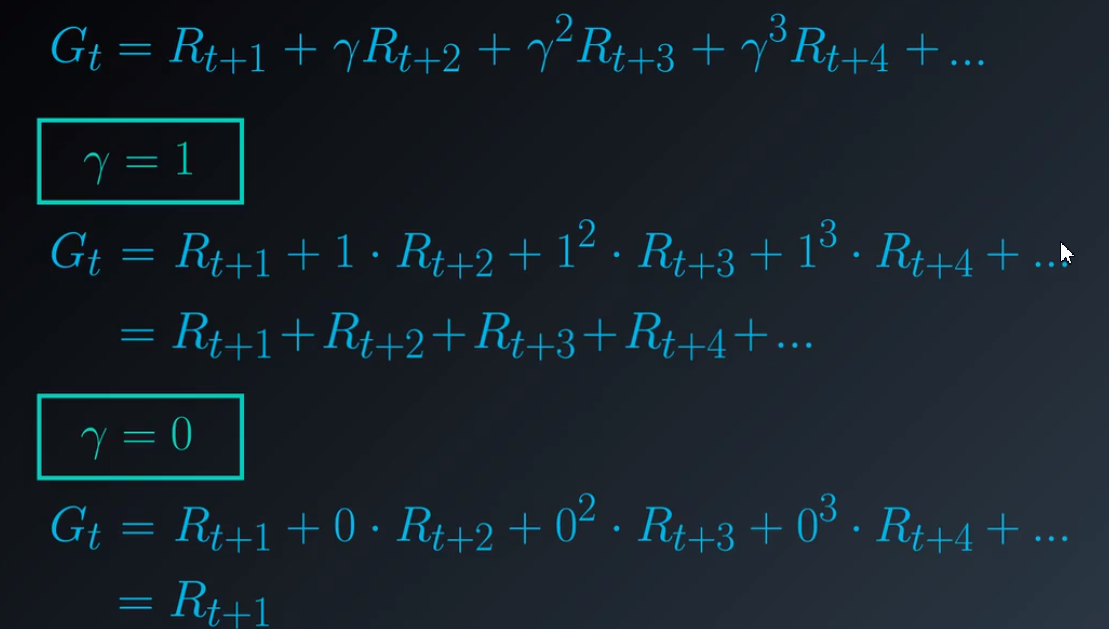
**折扣回报等于**奖励 乘以 折扣率（随时间越远，值越小），并将他们相加。





折扣回报规律

γ 的值越大，智能体对遥远未来的关注就越高



连续性任务是指智能体与环境的互动一直持续下去 不会结束。

值得注意的是 **折扣与连续性任务的关系最大。**

由于未来是无限的 智能体 想要 最大化累积奖励目标 是就很困难。

因此我们使用折扣 避免无限未来的奖励 产生不良的影响

但是注意 无论是否有折扣 目标始终不变 一直都是最大化累积奖励

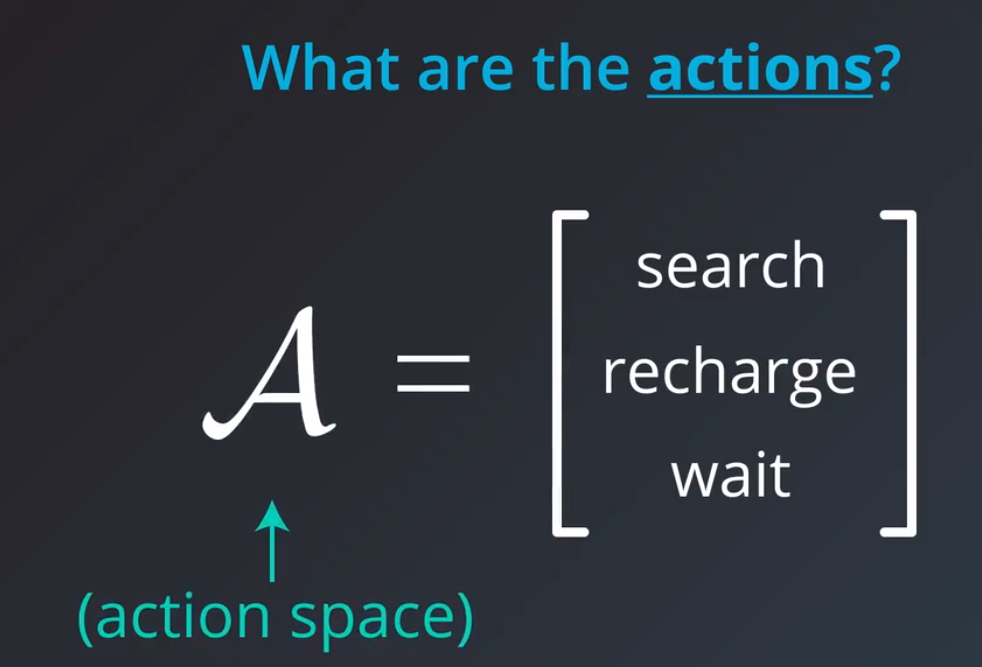
# MDP

## 动作空间（Action space）

1、能够在房间中搜索易拉罐,

2、能够前往充电基站充电,

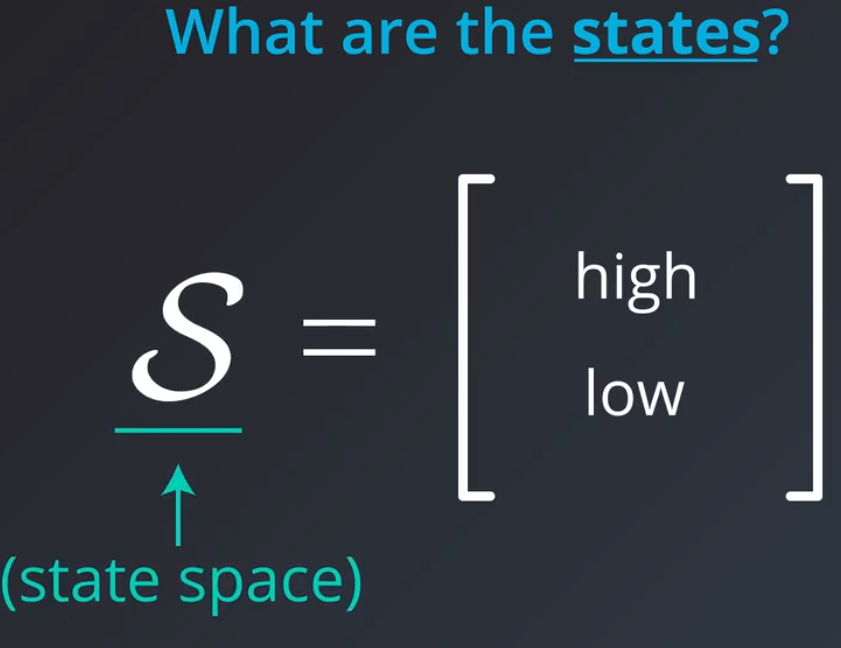
3、保持不动并等待其他人扔进去一个易拉罐

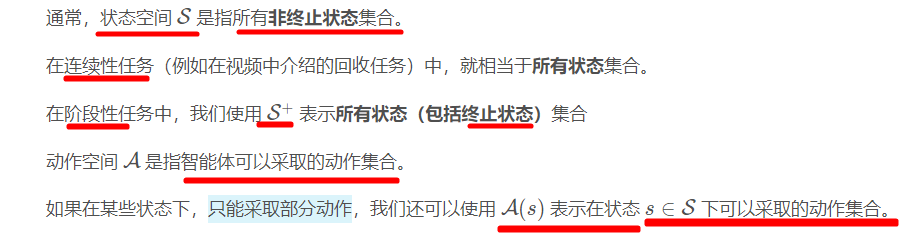


## 状态空间（State space）

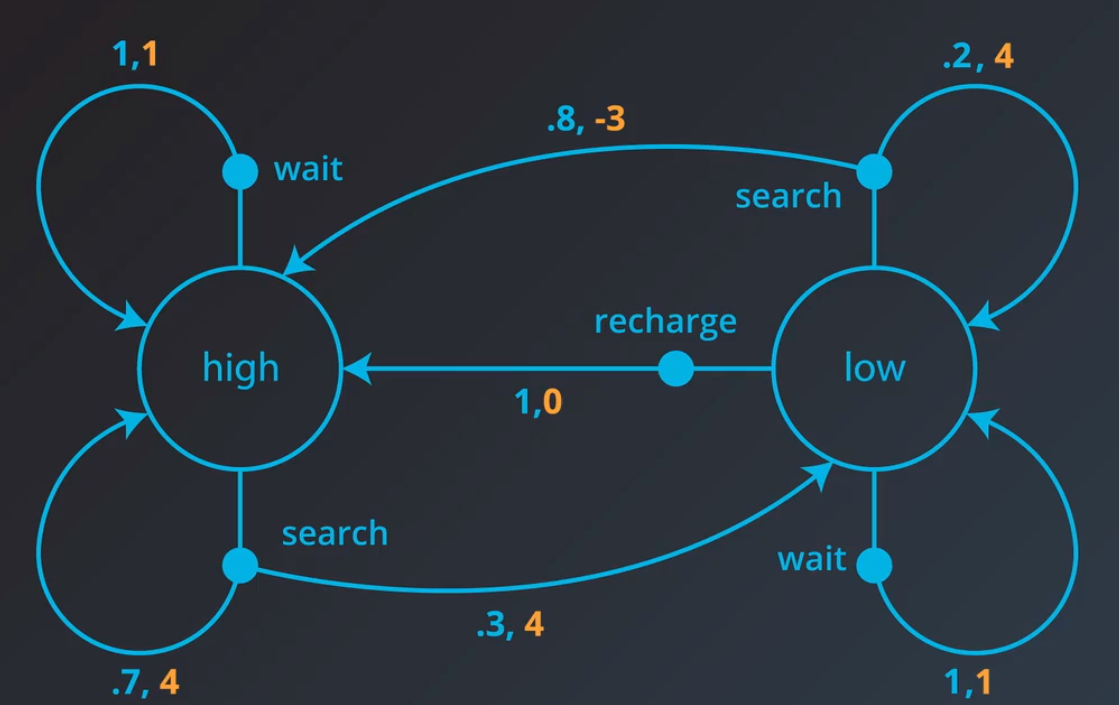
状态是提供给智能体的背景信息，帮助智能体做出合理的动作

状态可能是机器人电池的电量:剩余电量很高，应剩余电量很低





## 机器人收集垃圾MDP

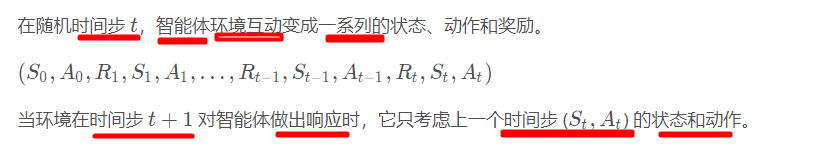


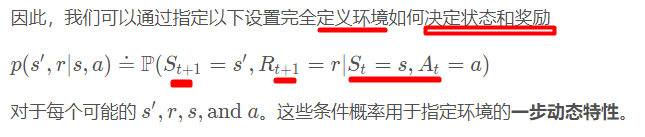
注意：

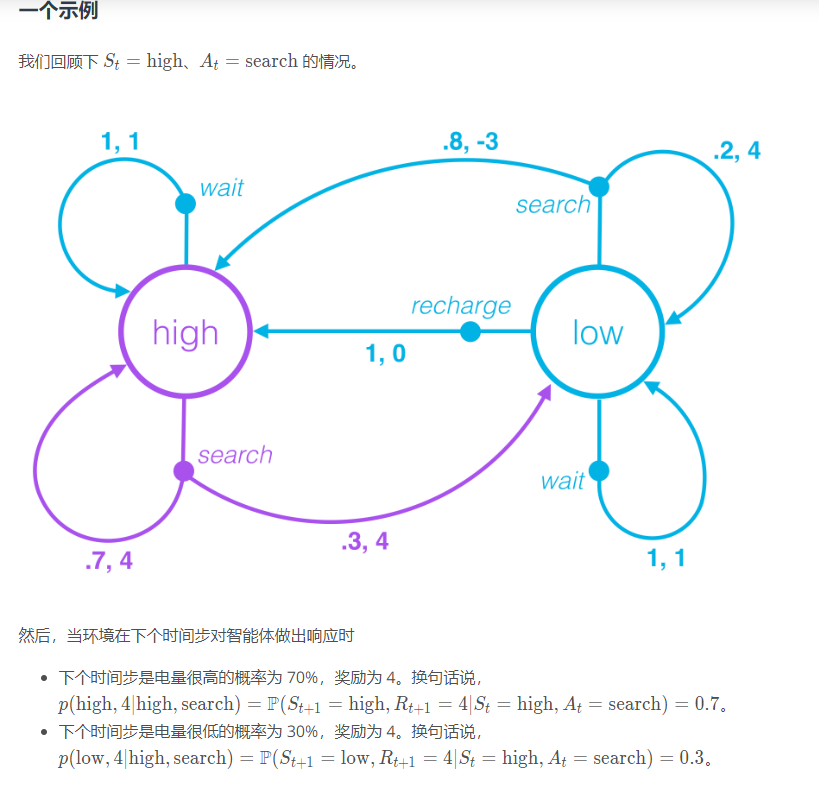
1.环境做出决策所需的信息很少 ，它不关心在 10 个或 100 个甚至 2 个时间步之前，提供给智能体的情形。并且也不会查看在上个时间步之前智能体采取的动作。

2.智能体的表现或收集的奖励数量，不影响环境如何选择对智能体做出响应。

## 一步动态特性







## MDP定义

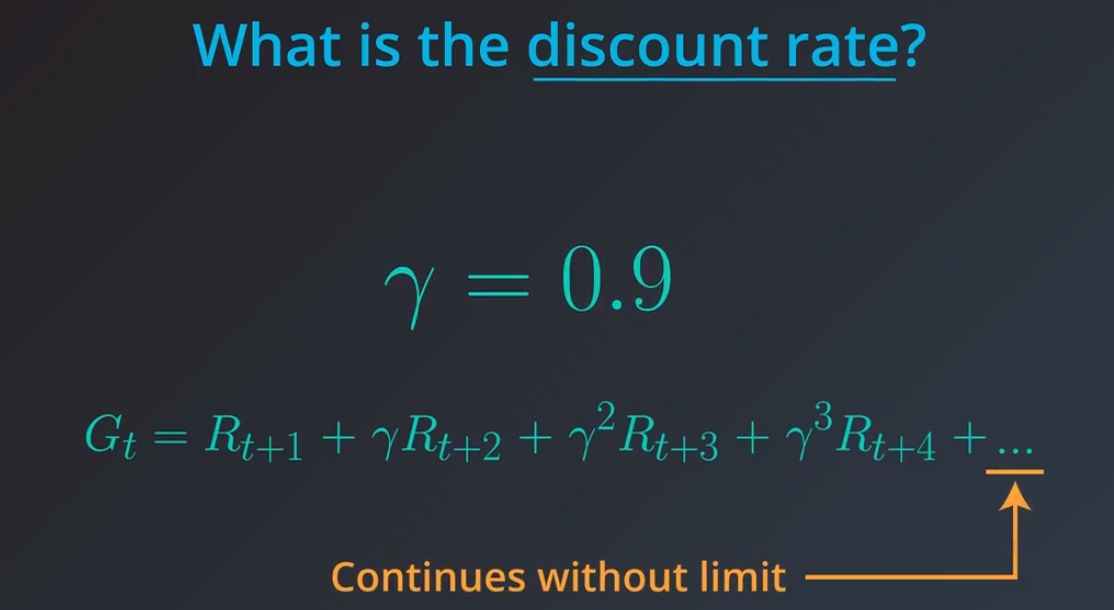
马尔可夫决策流程 (MDP) 由一组状态 一组动作 一组奖励 和一步环境动态特性 及 折扣率决定。



智能体知道：状态 动作 和 折扣因子

智能体不知道：奖励 和 一步动态特性 因为指定了环境的工作原理

# 折扣率



对于连续性任务 有必要将折扣率设为小于 1，否则 智能体需要查看无限的未来，通常我们都将折扣率设为 0.9 。

折扣率始终设为非常接近 1 而不是 0 的某个值，否则 智能体会出现看不清故障的情况。