

# 强化学习之q-learning

杨耀

## 什么是强化学习

强化学习是机器学习大家族中的一大类, 使用强化学习能够让机器学着如何在环境中拿到高分, 表现出优秀的成绩. 而这些成绩背后却是他所付出的辛苦劳动, 不断的试错, 不断地尝试, 累积经验, 学习经验.

## 从无到有













强化学习是一类算法, 是让计算机实现从一开始什么都不懂, 脑袋里没有一点想法, 通过不断地尝试, 从错误中学习, 最后找到规律, 学会了达到目的的方法. 这就是一个完整的强化学习过程. 实际中的强化学习例子有很多. 比如最有名的 Alpha go, 机器头一次在围棋场上战胜人类高手, 让计算机自己学着玩经典游戏 Atari, 这些都是让计算机在不断的尝试中更新自己的行为准则, 从而一步步学会如何下好围棋, 如何操控游戏得到高分. 既然要让计算机自己学, 那计算机通过什么来学习呢?

## 虚拟老师

计算机需要一位虚拟的老师, 这个老师比较吝啬, 他不会告诉你如何移动, 如何做决定, 他为你做的事只有给你的行为打分, 那我们应该以什么形式学习这些现有的资源, 或者说怎么样只从分数中学习到我应该怎样做决定呢? 很简单, 我只需要记住那些高分, 低分对应的行为, 下次用同样的行为拿高分, 并避免低分的行为.

比如老师会根据我的开心程度来打分, 我开心时, 可以得到高分, 我不开心时得到低分. 有了这些被打分的经验, 我就能判断为了拿到高分, 我应该选择一张开心的脸, 避免选到伤心的脸. 这也是强化学习的核心思想. 可以看出在强化学习中, 一种行为的分数是十分重要的. 所以强化学习具有分数导向性. 我们换一个角度来思考. 这种分数导向性好比我们在监督学习中的正确标签.

# 对比监督学习

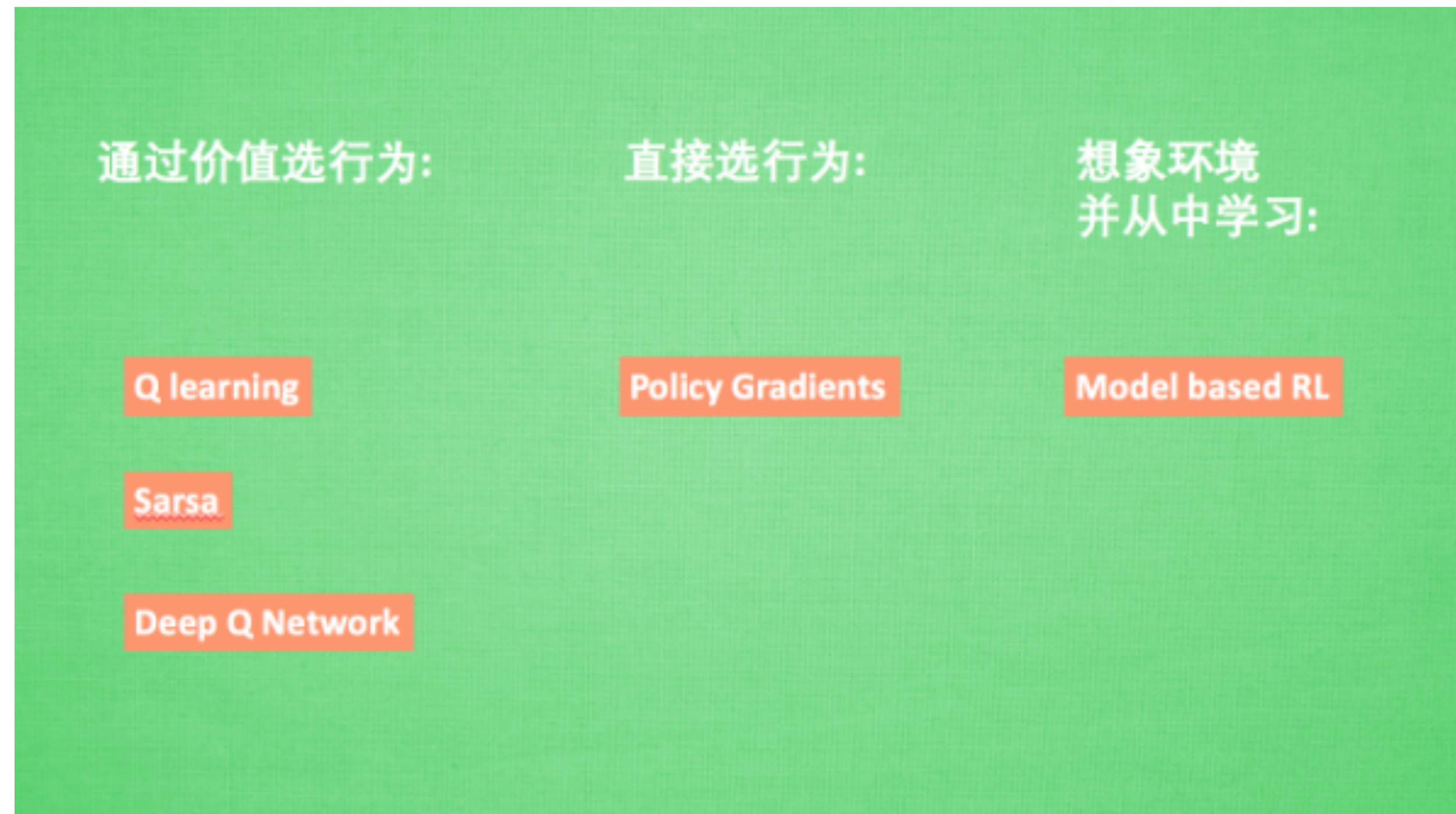
数据			标签		
			高分	高分	高分
			高分	高分	高分
			低分	低分	低分
			低分	低分	低分

我们知道监督学习, 是已经有了数据和数据对应的正确标签, 比如这样. 监督学习就能学习出那些脸对应哪种标签. 不过强化学习还要更进一步, 一开始它并没有数据和标签.

他要通过一次次在环境中的尝试, 获取这些数据和标签, 然后再学习通过哪些数据能够对应哪些标签, 通过学习到的这些规律, 尽可能地选择带来高分的行为 (比如这里的开心脸). 这也就证明了在强化学习中, 分数标签就是他的老师, 他和监督学习中的老师也差不多.



## RL 算法们



强化学习是一个大家族, 他包含了很多种算法, 比如有通过行为的价值来选取特定行为的方法, 包括使用表格学习的 q learning, sarsa, 使用神经网络学习的 deep q network, 还有直接输出行为的 policy gradients, 又或者了解所处的环境, 想象出一个虚拟的环境并从虚拟的环境中学习 等等.

- 强化学习分类（基于价值、基于策略）

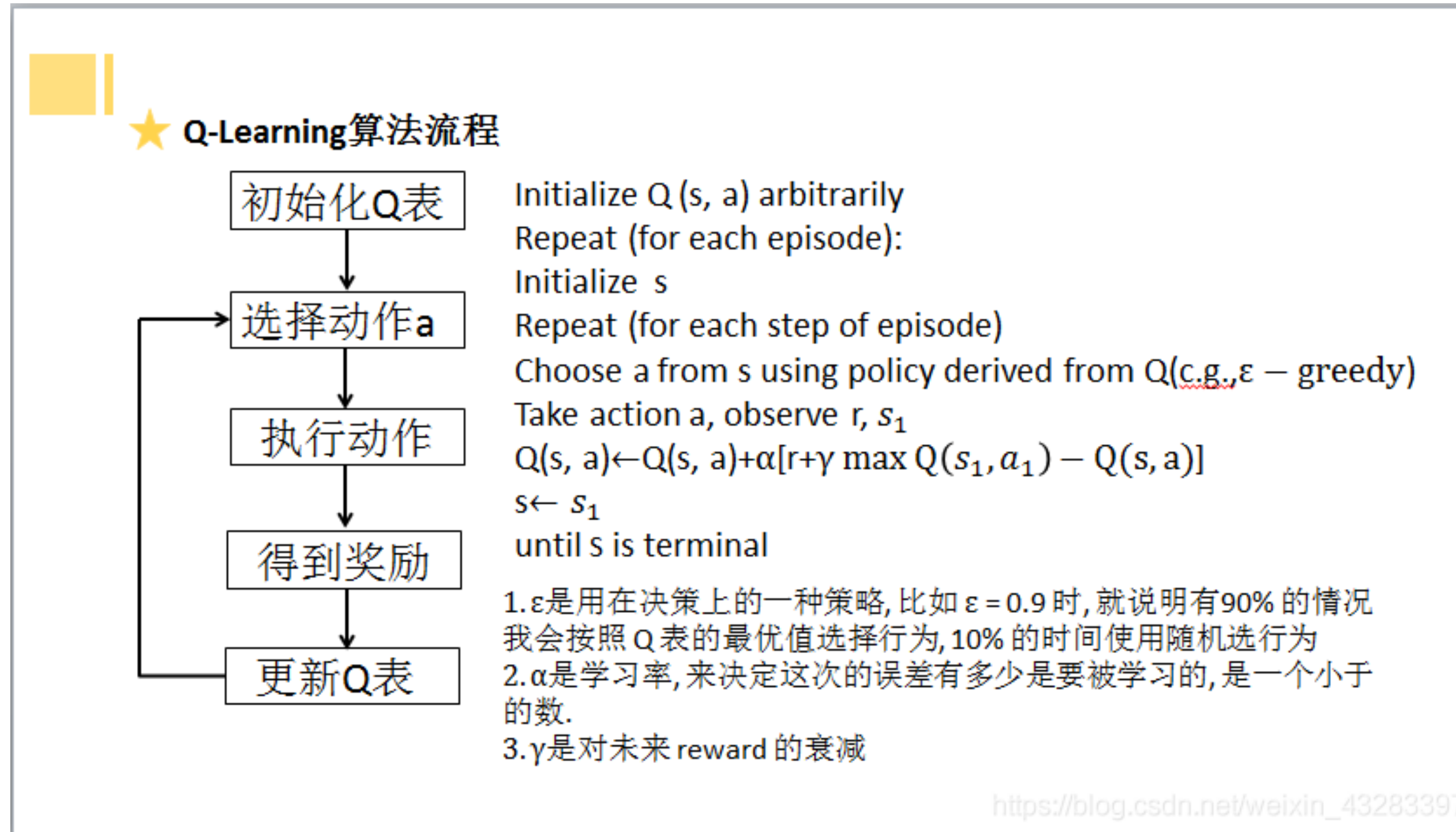
基于价值（Value-Based）的强化学习：智能体通过学习价值函数，隐式的策略，如 $\epsilon$ -greedy。Value-Based 算法的缺点：1）对连续动作的处理能力不足；2）对受限状态下的问题处理能力不足；3）无法解决随机策略问题。包括Q-Learning、SARSA、Deep-Q-network算法。

基于策略（Policy-Based）的强化学习：没有价值函数，直接学习策略。基于策略（Policy-Based）适用于随机策略、连续动作。包括Policy Gradient算法、TRPO、PPO。

演员-评论家强化学习：学习价值函数(评论家),同时也学习策略(演员)。包括Actor-Critic算法、DDPG。



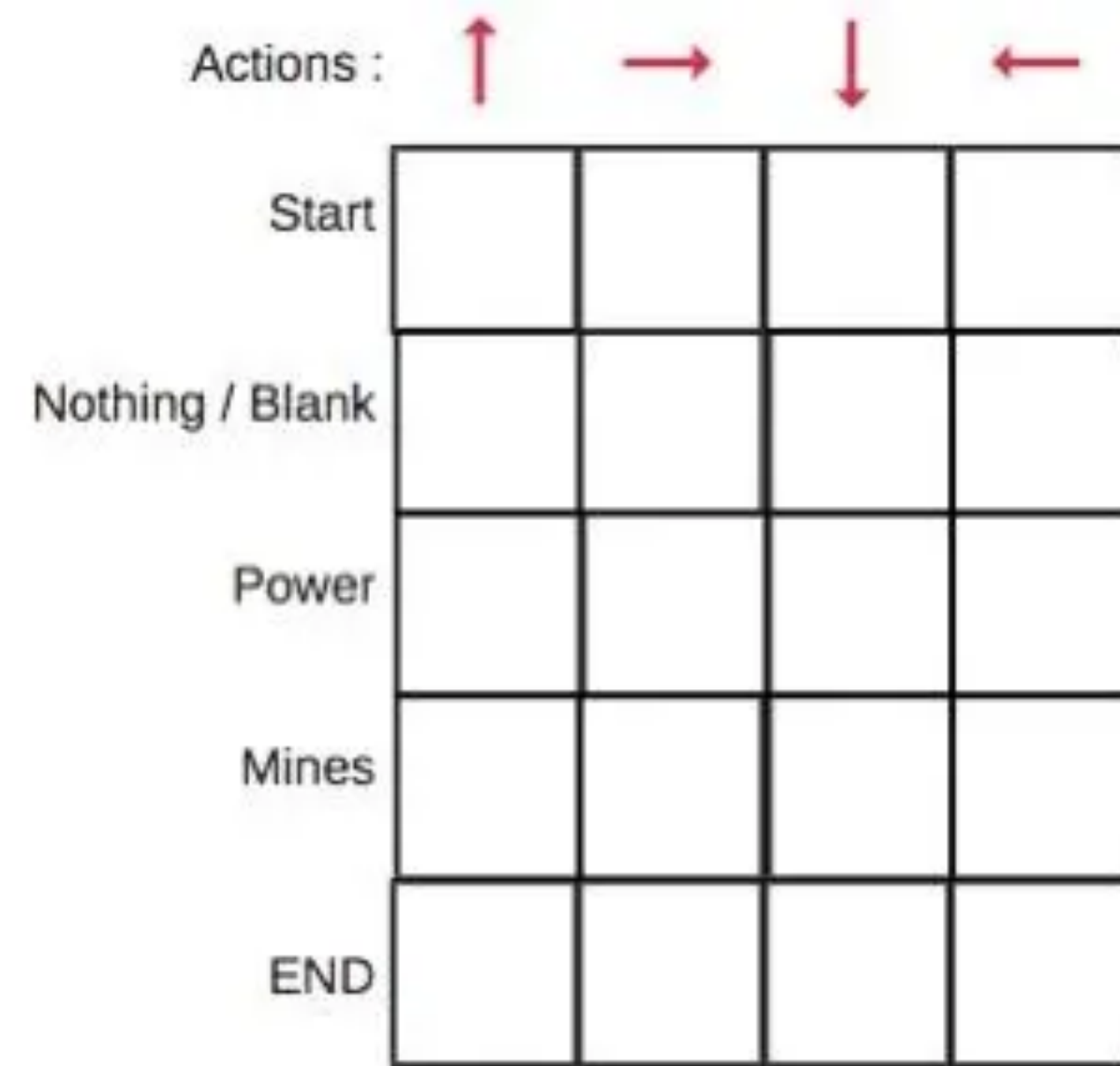
# Q-Learning



Q-Table/状态-价值函数 $Q(s,a)$ : 这个表格的每一行代表每个 state, 每一列代表每个 action, 表格的数值就是在各个 state 下采取各个 action 时能够获得的最大的未来期望奖励。

例如在一个游戏中有下面5种状态和4种行为, 则表格为:





通过 Q table 就可以找到每个状态下的最优行为，进而通过找到所有最优的action得到最大的期望奖励。

Q表更新公式：

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[ R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right].$$

Q-learning在更新Q值时下一步动作是不确定的，它会选取Q值最大的动作来作为下一步的更新。

- 实战
- 寻宝游戏
- 迷宫游戏