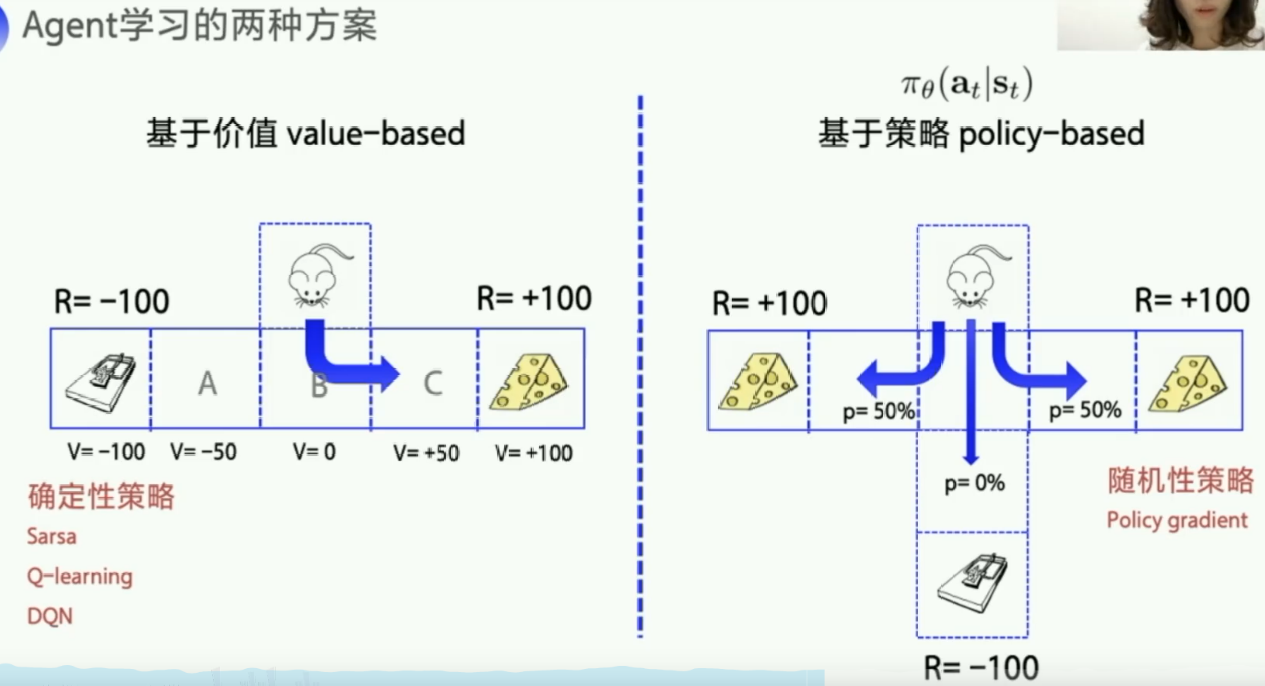
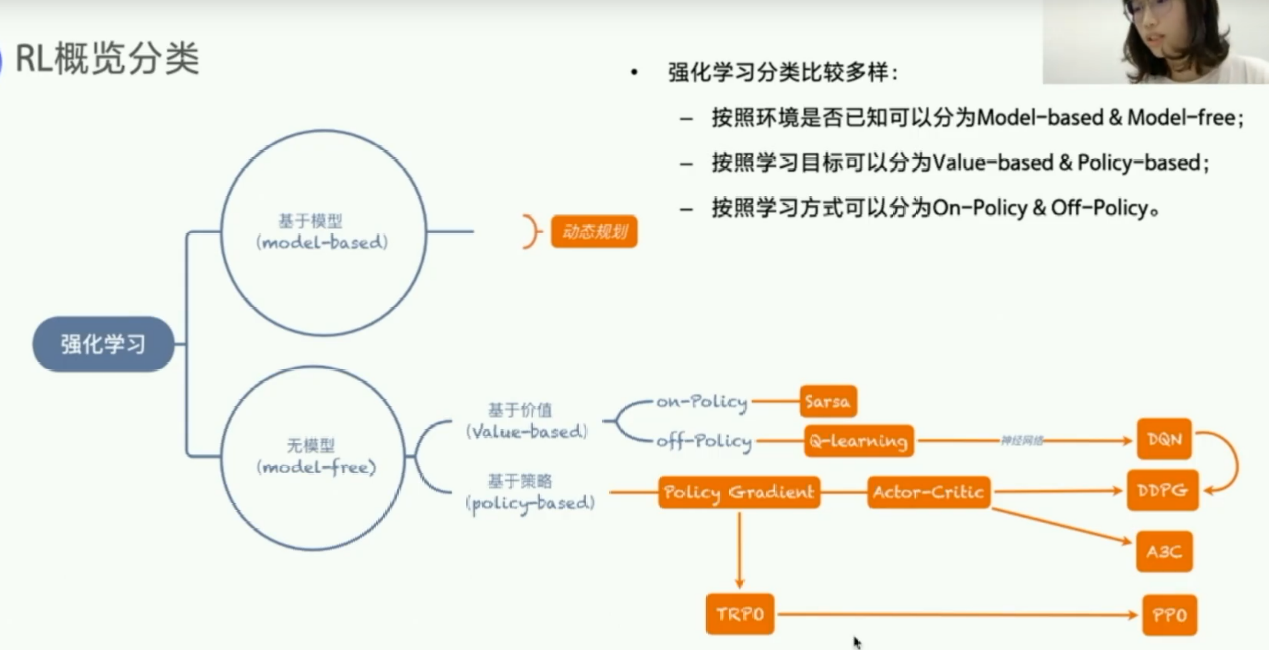


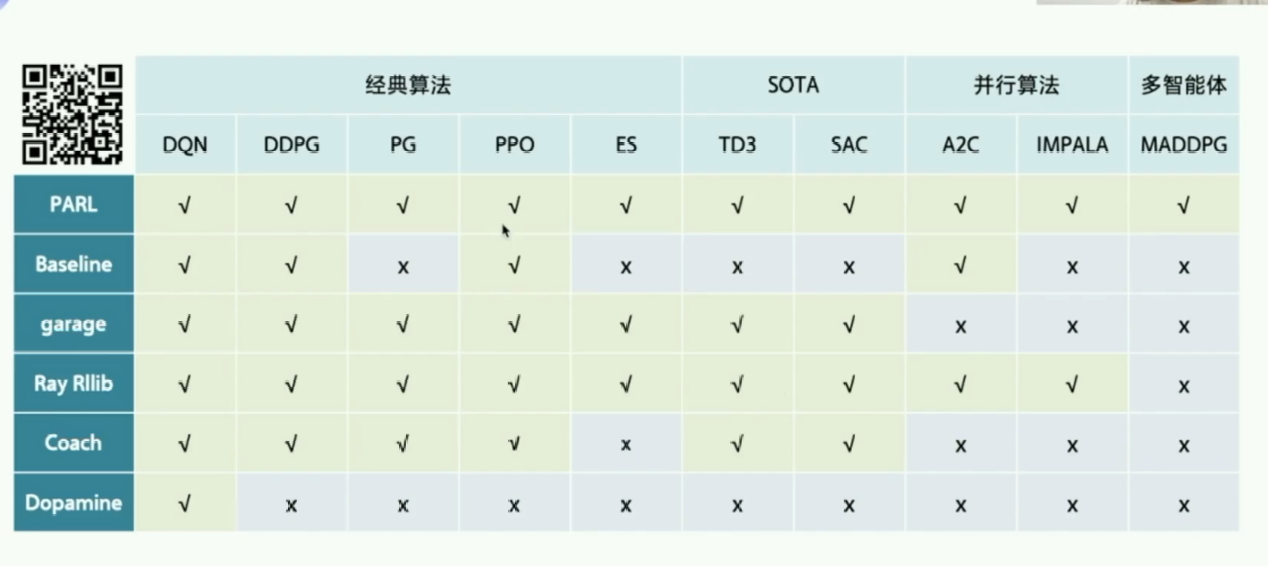
监督学习有样本，也有结果，用结果来校验训练精度，相当于向老师学习，老师会的我才会，而强化学习是从自己的经验中学习，自己去尝试，然后更具奖励去学习；

非监督学习的话，

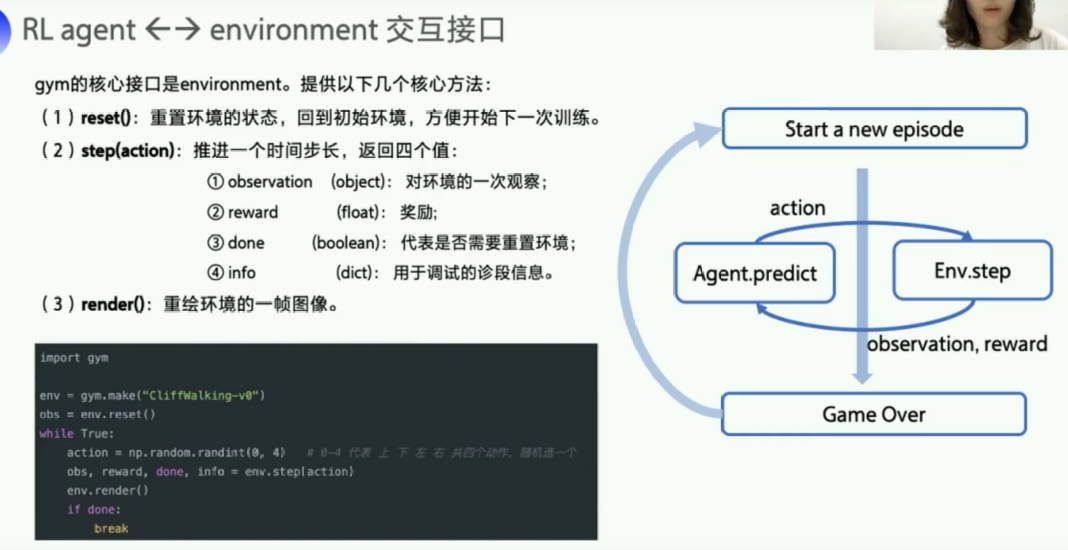


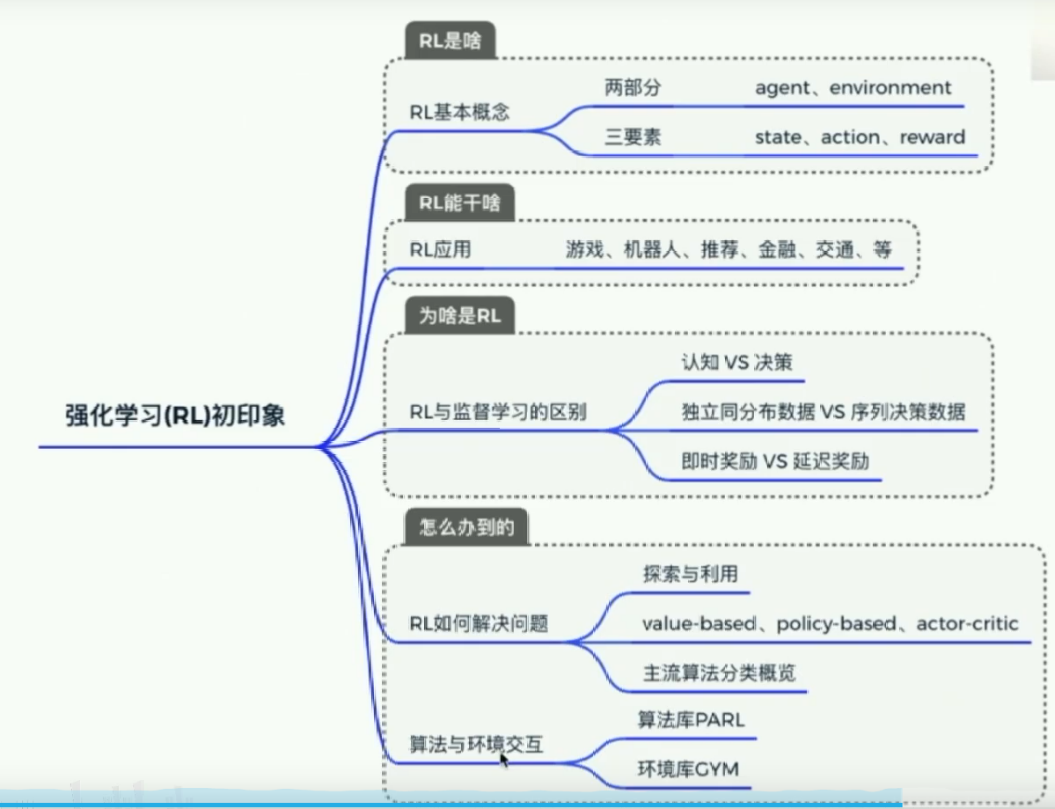
随机性策略就是输出的最终策略是有概率的，例如执行某个动作的概率是多少

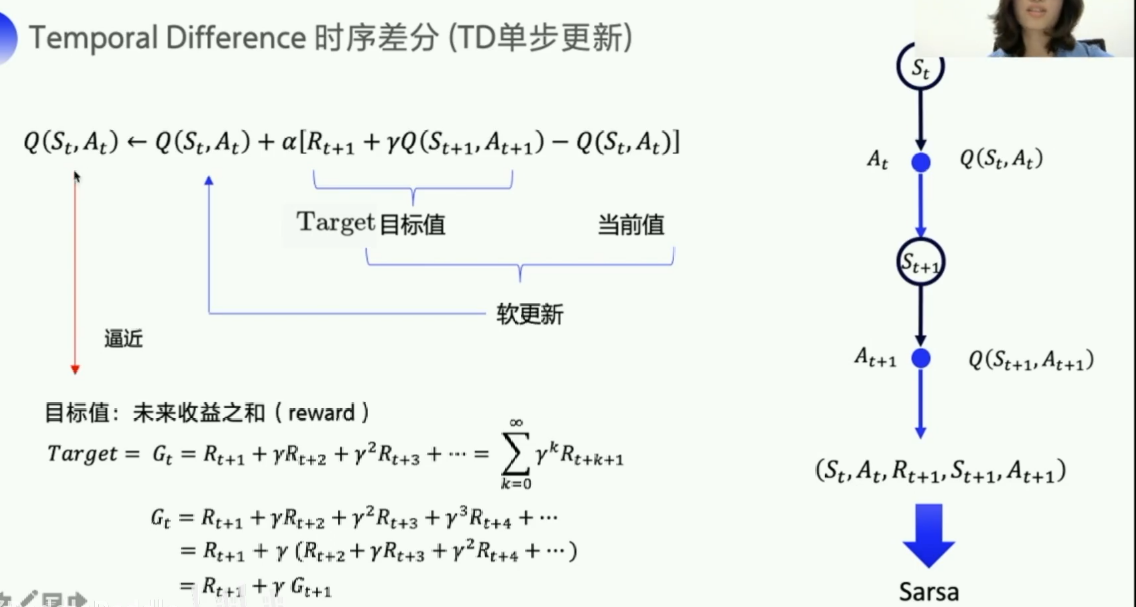






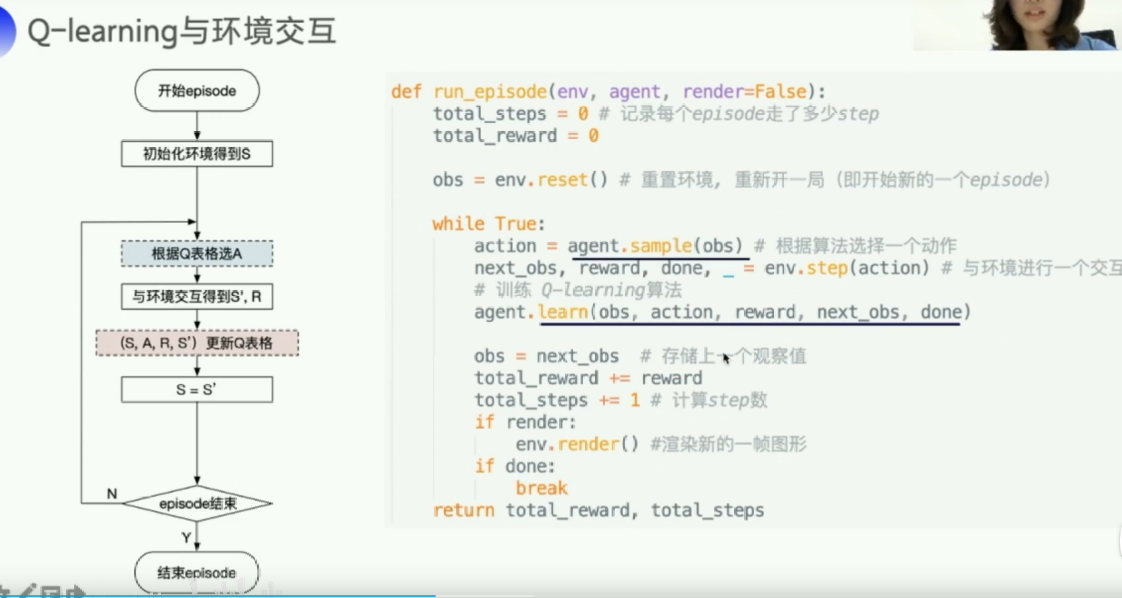


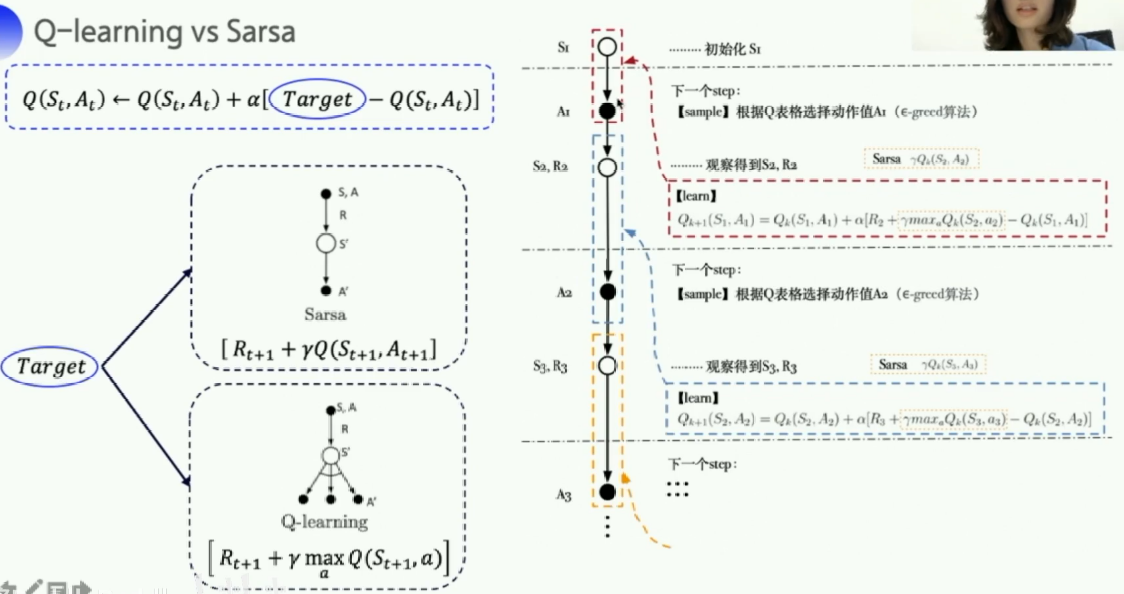




Sarsa就相当于（s,a,r,s,a）它需要获取下一时刻的状态和下一状态执行的动作；

而下面的Q-learning则不需要下一状态执行的动作；





Q-learning优化的时候默认下一个动作就是Q最大的动作，它知道下一步动作有百分之十的探索，但它不担心后面的探索策略，而是按照最优的策略来优化目标策略，就是默认下一个动作就是Q最大的动作；我感觉就是这里把探索和优化分开了，训练过程中的实际执行里面有探索这个过程，但在优化的时候，只利用最优策略（下一状态最大的Q值）来优化；

而sarsa这里就是利用探索策略来优化，它用于优化的下一步的Q值就是实际执行的下一步的Q值；

为什么悬崖问题sarsa和Q-learning训练的结果不一样？

最后的策略是啥？结束训练之后得到的是啥？

我感觉是一个Q表，然后按照这个Q表来执行；这个Q表就是最优策略