第五章
行为学派

智能优化、强化学习

行为学派

- 行为学派简介
- 智能优化的产生与发展
- 智能优化典型思想与方法
- 强化学习基本概念与发展历史
- 经典强化学习基本思想与算法

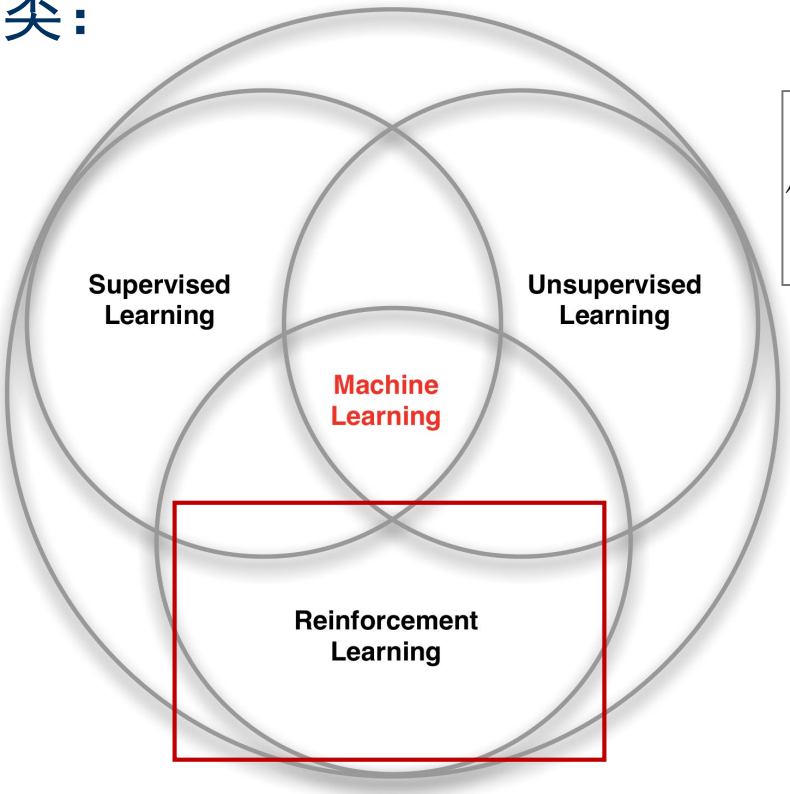
强化学习

- 强化学习基本概念与发展历史
  - 强化学习的基本概念
  - 强化学习的机制由来
  - 强化学习的发展历史
- 经典强化学习基本思想与算法
  - 基本模型和核心概念
  - 强化学习的若干关键问题
  - 经典强化学习：Q学习

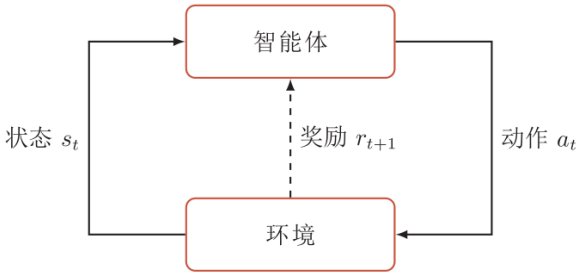
机器学习与强化学习

- 机器学习的分类：

监督学习
例：分类器，预测，回归



无监督学习
例：主元分析，特征学习



强化学习是机器学习三大范式之一

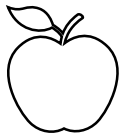
机器学习与强化学习

监督学习

输入：(x, y)

x 是数据，y 是标签

目的：建立 x 到 y 的映射 $y = f(x)$



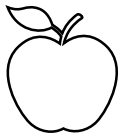
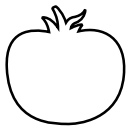
这是苹果

无监督学习

输入：x

x 是数据，无标签

目的：学习 x 的底层结构/特征

这些都是水果

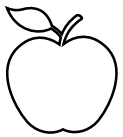
强化学习

输入：状态-动作空间

$available\ action$:

$= f(state)$

目的：通过选择动作来最大化预期未来奖励



吃掉苹果来让自己(更好地)生存

机器学习与强化学习

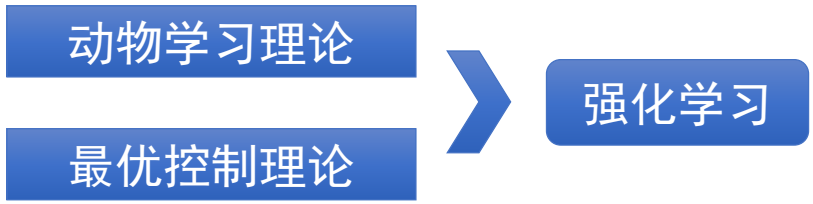
- 强化学习（Reinforcement Learning，RL）的主体：智能体（Agent）
- 强化学习与其他机器学习范式的区别：
  - 没有监督，只有奖励信号
  - 延迟反馈，而非瞬时结果
  - 时序的重要作用（使用序列训练数据，而非独立同分布数据）
  - 智能体与环境的互动（动态特性）
    - 机器的动作影响了它接下来获取的数据

参考书目

- An Introduction to Reinforcement Learning, Sutton and Barto, 1998
  - MIT Press, 1998
  - http://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf
- Algorithms for Reinforcement Learning, Szepesvari, 2009
  - Morgan and Claypool, 2010
  - https://www.semanticscholar.org/paper/Algorithms-for-Reinforcement-Learning-Szepesvari/e60f3c1cb857daa3233f2c5b17b6f111ff86698c

强化学习

- 强化学习基本概念与发展历史
  - 强化学习的基本概念
  - 强化学习的机制由来
  - 强化学习的发展历史
- 经典强化学习基本思想与算法
  - 基本模型和核心概念
  - 强化学习的若干关键问题
  - 经典强化学习：Q学习



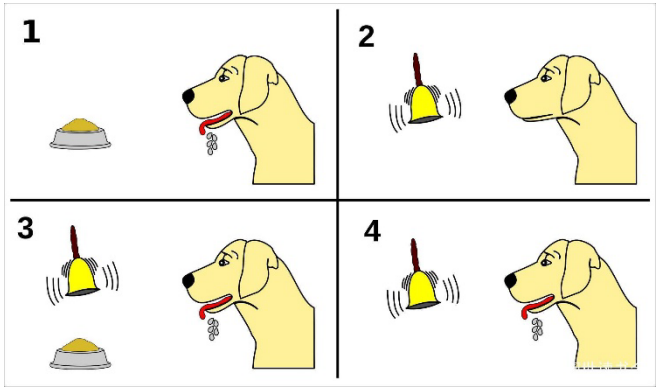
动物学习、实验心理学和神经科学

- 效果法则（Law of Effect）
  - “在其他条件相同的情况下，对同一情况作出的几种反应中，如果伴随着动物意志的满足感发生在其中某些反应的期间，或者紧随其后，那么这些反应将与该情况建立更牢固的联系，因此，当这种情况再次发生时，这些反应发生的可能性会提高；如果在其他条件相同的情况下，动物意志的不适感发生在某些反应期间，或者紧随其后，那么这些反应与该情况的联系将被削弱，因此，当这种情况再次发生时，这些反应发生的可能性会降低。满足感或不适感的程度越大，联系加强或削弱的程度就越大。”[**桑代克，1911，《动物智慧》**]

行为主义心理学的主要原理

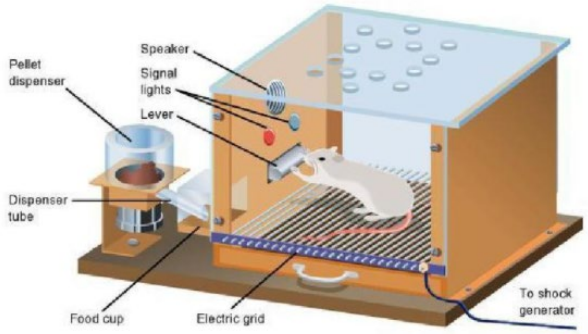
实验心理学启发

- 经典的动物（包含人）条件反射。“条件响应(conditioned response)的幅度和时效，会随着条件刺激和非条件刺激之间的偶然性(contingency)发生变化” [巴甫洛夫，1927年]



- 操作条件反射（或工具性条件反射）：人类和动物学习行为以获得奖励(obtain rewards)和避免惩罚(avoid punishments)的过程 [斯金纳，1938]。

斯金纳的研究

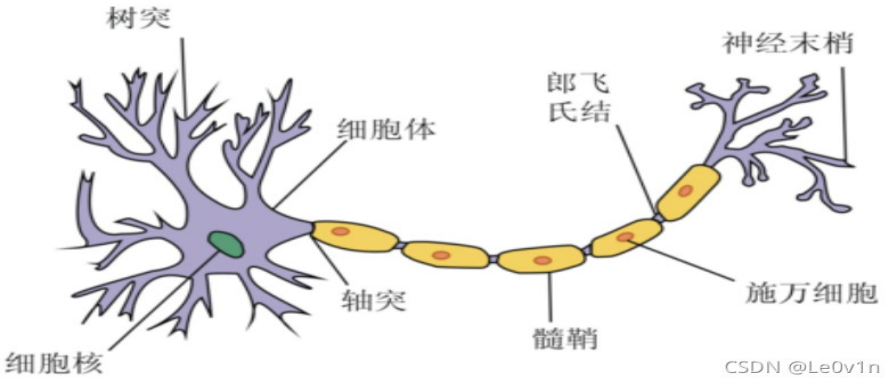


Remark: 强化指任何形式的条件反射，既可以是正面的（奖励）也可以是负面的（惩罚）

行为主义心理学的主要原理

计算神经科学启发

- 赫布（Hebbian）学习：通过共同激活神经元，来强化它们之间的突触权重，从而发展模型的形式。"如果先激活一个神经元，然后马上激活另一个神经元，它们就会连在一起"。[赫布，1961]。



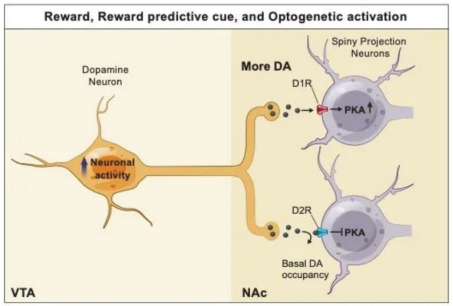
Neurons that fire together, wire together.
— Donald Hebb

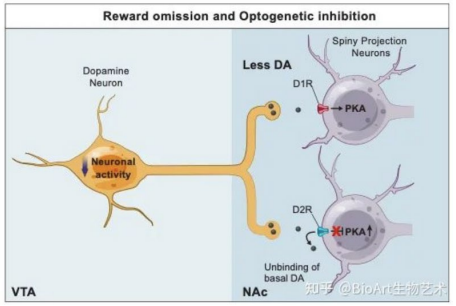
例：肌肉记忆完成下意识动作

计算神经科学启发

- 多巴胺和基底核模型：与运动控制和决策有直接联系 [铜谷贤治，1999]
  - *Remark*：强化代表了多巴胺（和惊喜）的作用。

- 首先由瑞典药理学家Arvid Carlsson （2000诺贝尔奖获得者） 1957年发现。作为神经递质，帮助细胞传送脉冲化学物质，调控中枢神经系统的多种生理功能。
- 在学习过程中，Dopamine作为一种奖励预测误差信号来提高对未来奖励的预测精度，从而可以指导人/动物改进动作。

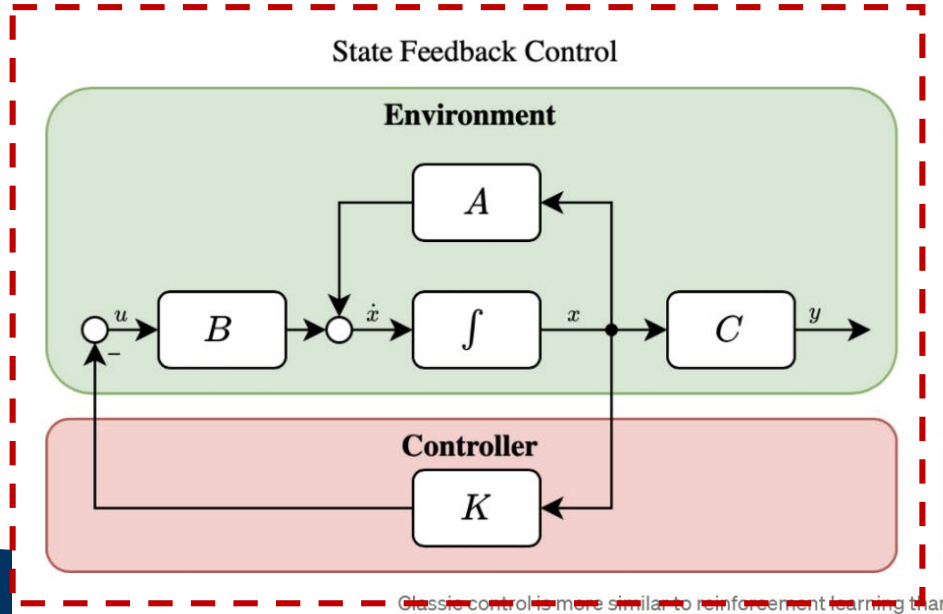


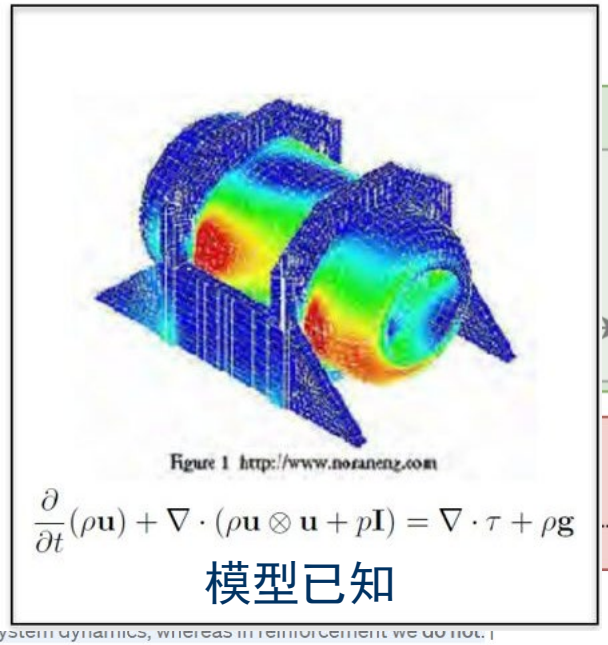


如果可以精准预测某一动作对未来奖励的影响，智能体就可以对全部允许采取的动作能够带来的奖励进行预测，进而选择带来最大奖励的动作。相当于RL中的预测算法。

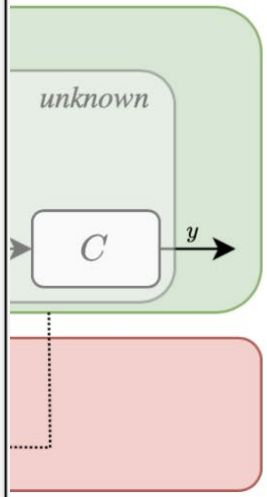
最优控制与强化学习

- **最优控制**：以优化方法的形式框架，求取连续时间控制问题中的最优控制策略。
  - 假设模型已知（Model-based，如左图）
  - **动态规划**是求解最优控制问题的一种经典方法





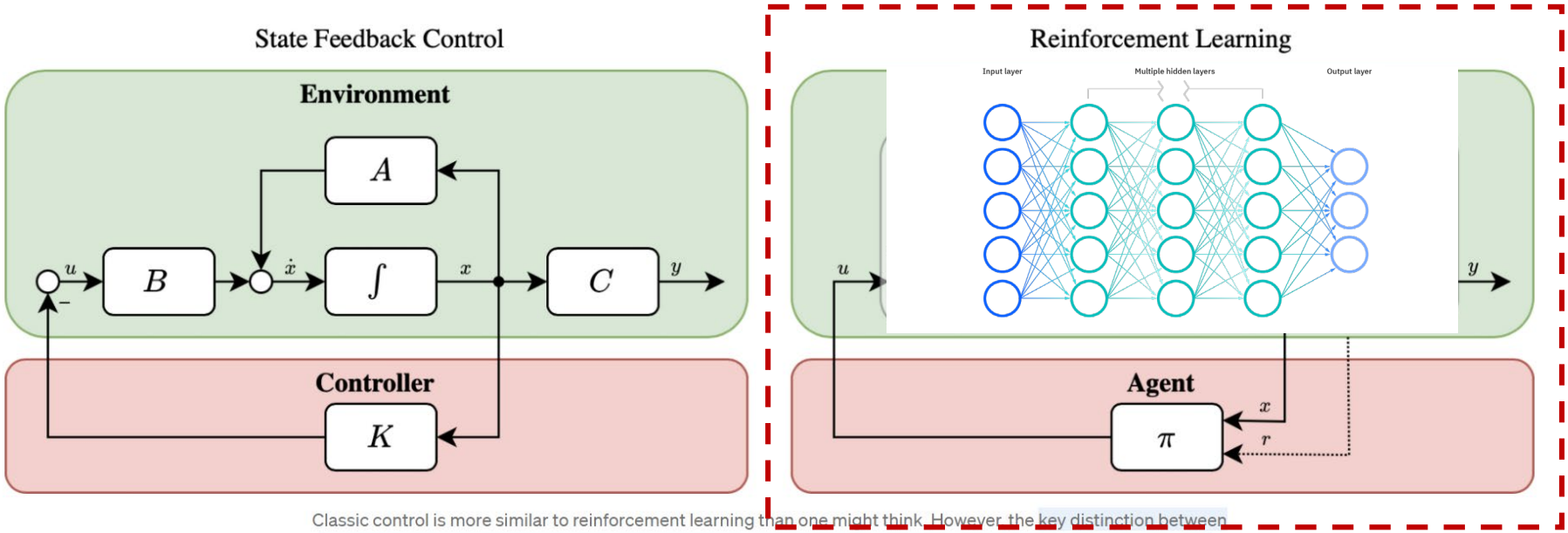
模型已知



Classic control is more similar to reinforcement learning than them is, that in control we assume to know the underlying system dynamics, whereas in reinforcement we do not.
(Image by author)

最优控制与强化学习

- 强化学习：通过与未知和不确定（如随机）环境的直接交互（试错）学习一种行为策略，使长期的奖励总和（延迟奖励）最大化。
  - 模型通常未知（Model-free，如右图）



Classic control is more similar to reinforcement learning than one might think. However, the key distinction between them is, that in control we assume to know the underlying system dynamics, whereas in reinforcement we do not. | (Image by author)

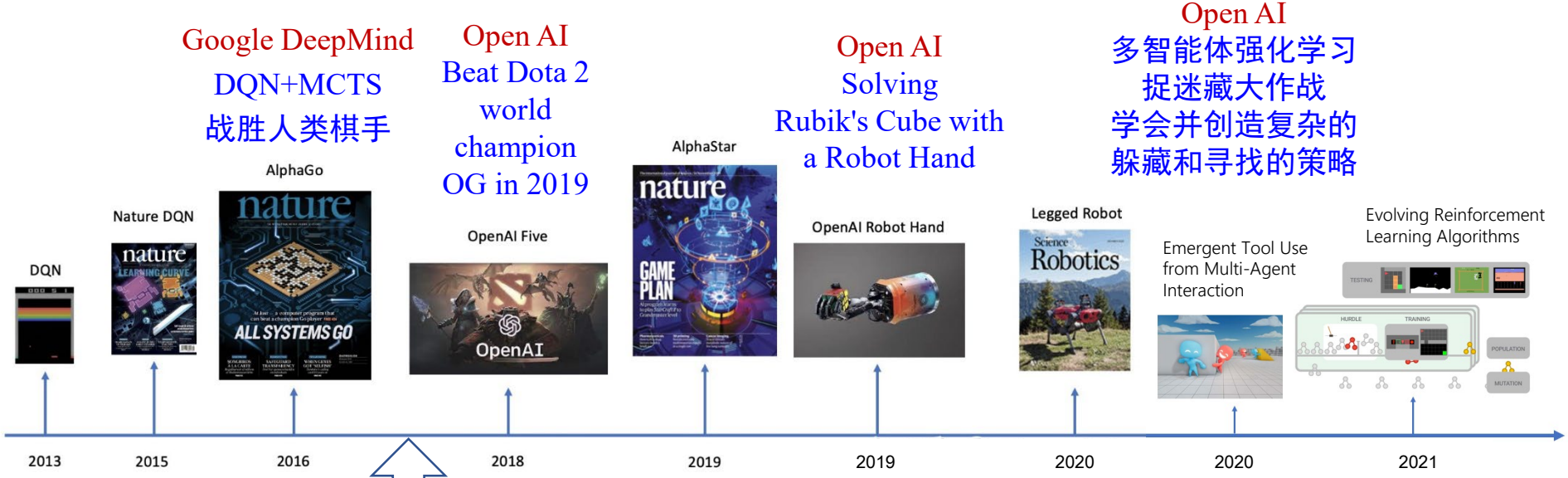
强化学习

- 强化学习基本概念与发展历史
  - 强化学习的基本概念
  - 强化学习的机制由来
  - 强化学习的发展历史
- 经典强化学习基本思想与算法
  - 基本模型和核心概念
  - 强化学习的若干关键问题
  - 经典强化学习：Q学习

强化学习的过去*

- 计算机下棋程序 [香农，1950（论文1988）].
- 神经模拟强化系统理论 [明斯基，1954].
- 利用跳棋游戏进行机器学习的研究 [Samuel，1959]。
- 试错（井字棋）[米基，1961]。
- 自适应控制实验（单连杆倒立摆）[米基 和 钱伯斯，1968]。
- 惩罚/奖励：在自适应阈值系统中与批判者一起学习（神经网络）[威德罗 等，1973].
- 联想搜索网络。强化学习联想记忆 [Barto等，1981].
- 强化学习中的时间分数分配（时间差分学习）[Sutton，1984]。
- 延迟奖励的学习（Q学习）[Watkins，1989]。
- 时间差分法与TD-Gammon [Tesauro，1995], 第一个利用神经网络的RL

现代强化学习研究里程碑



Google
Playing Atari
深度强化学习
鼻祖

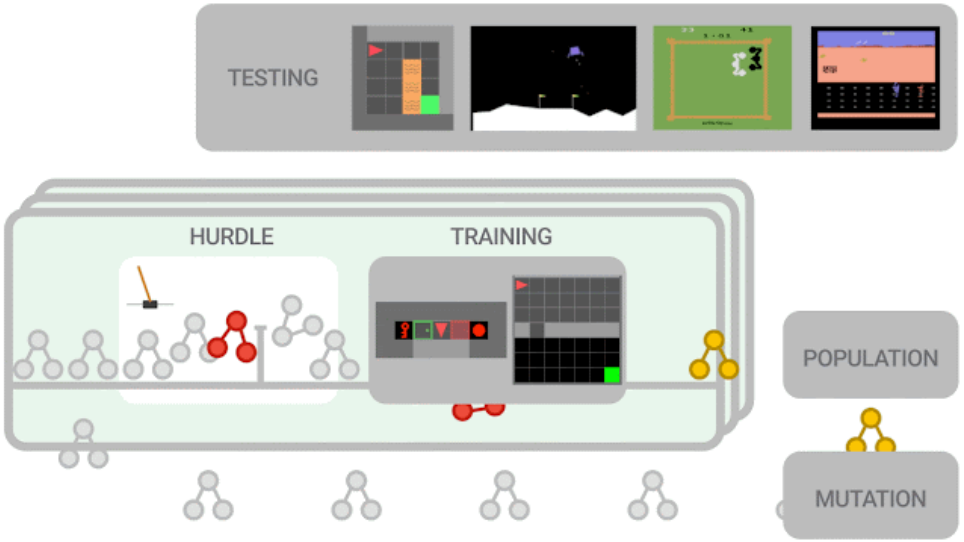
Alpha Zero
自我对弈
不用棋谱
碾压AlphaGo

Google
DeepMind
星际争霸II
击败99.8%
的人类玩家

Robotic Systems
Lab / ANYmal
Zurich,
Switzerland
多足机器人运动
控制

Google Research
用进化算法自动
设计RL算法

参考文献



1. **DQN: Playing Atari with Deep Reinforcement Learning** https://www.deepmind.com/publications/playing-atari-with-deep-reinforcement-learning
2. **Nature DQN: Human-level control through deep reinforcement learning** https://www.semanticscholar.org/paper/Human-level-control-through-deep-reinforcement-Mnih-Kavukcuoglu/e0e9a94c4a6ba219e768b4e59f72c18f0a22e23d
3. **AlphaGo: Mastering the game of Go with deep neural networks and tree search** https://www.nature.com/articles/nature16961
4. **OpenAI Five: Dota 2 with Large Scale Deep Reinforcement Learning** https://arxiv.org/abs/1912.06680
5. **AlphaStar: Grandmaster level in StarCraft II using multi-agent reinforcement learning** https://www.nature.com/articles/s41586-019-1724-z
6. **OpenAI robot hand: Solving Rubik's Cube with a Robot Hand** https://arxiv.org/abs/1910.07113
7. **Legged robot: Learning agile and dynamic motor skills for legged robots** https://www.science.org/doi/full/10.1126/scirobotics.aau5872?casa_token=sGRQXxPeIVYAAAAA%3AVl9WJaYBXvjbtT46KbLBWpv9N1X023dAMxEkWMwYLqp7hcIgJv7u-tCHacv1RgpuKYwdMsga5V98NPA
8. **Learning agile and dynamic motor skills for legged robots** https://www.science.org/doi/full/10.1126/scirobotics.aau5872?casa_token=sGRQXxPeIVYAAAAA%3AVl9WJaYBXvjbtT46KbLBWpv9N1X023dAMxEkWMwYLqp7hcIgJv7u-tCHacv1RgpuKYwdMsga5V98NPA
9. **Emergent Tool Use from Multi-Agent Interaction** https://arxiv.org/abs/1909.07528
10. **Evolving Reinforcement Learning Algorithms** https://ai.googleblog.com/2021/04/evolving-reinforcement-learning.html

强化学习

- 强化学习基本概念与发展历史
  - 强化学习的基本概念
  - 强化学习的机制由来
  - 强化学习的发展历史
- 经典强化学习基本思想与算法
  - 基本模型和核心概念
  - 强化学习的若干关键问题
  - 经典强化学习：Q学习

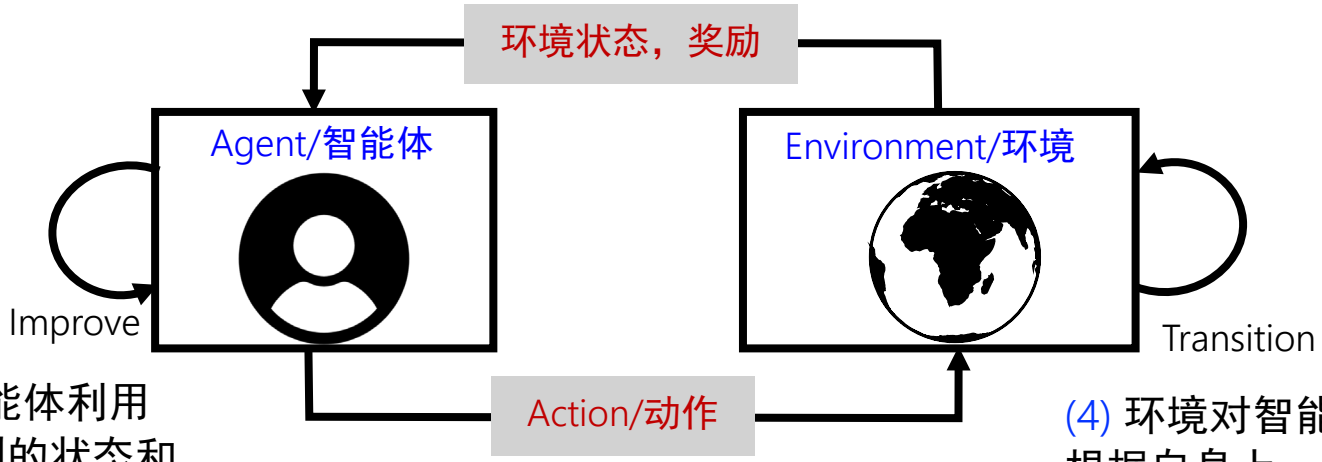
Ref

[1] 【强化学习】Reinforcement Learning Course by David Silver_哔哩哔哩_bilibili
[2] An Introduction to Reinforcement Learning, Sutton and Barto, 1998

强化学习模型

- 基本模型：

循环：
(1) 智能体观察环境



(2) 智能体利用观测到的状态和奖励改进自己的policy/策略

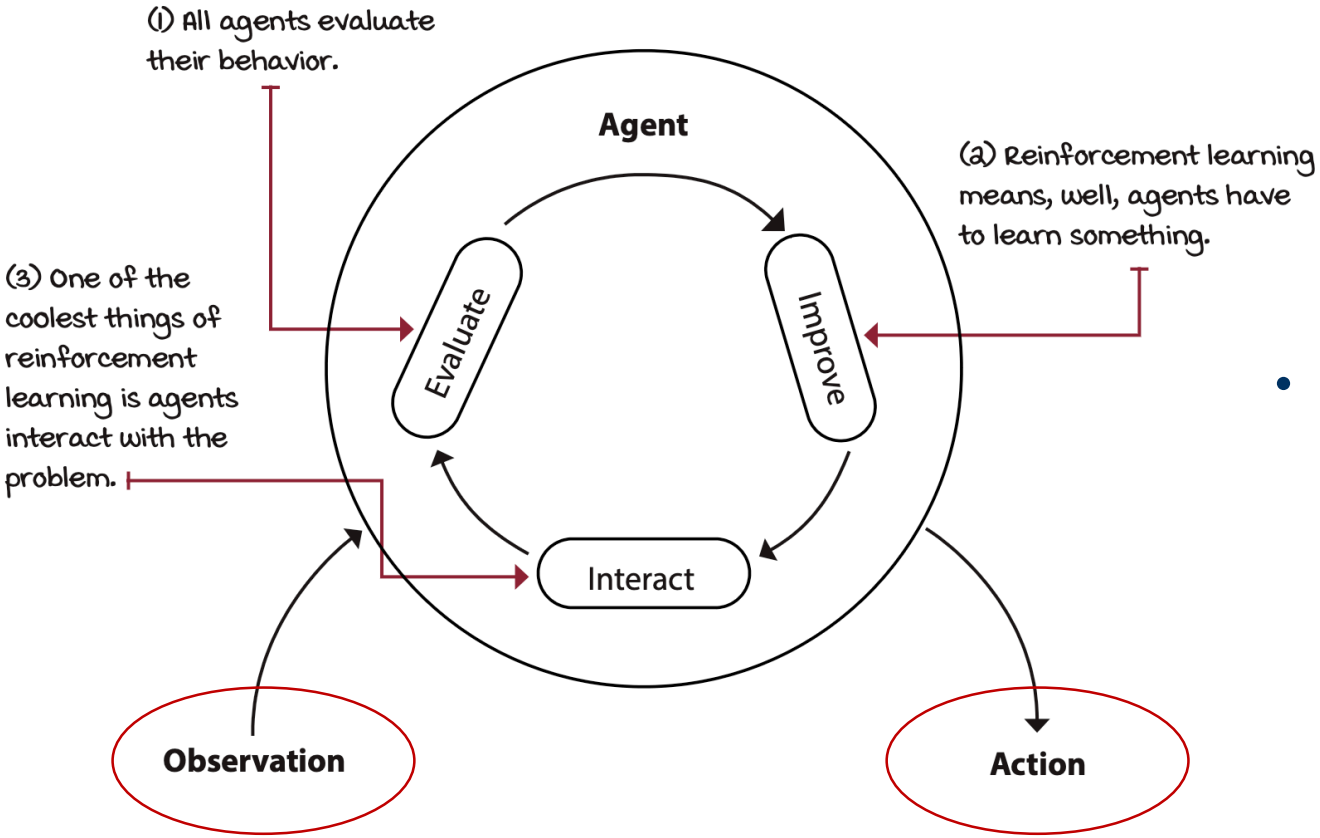
(3) 智能体选择并执行相应动作

(4) 环境对智能体动作做出响应，根据自身上一时刻的状态和智能体动作（可能）转移到下一个状态。回到 (1)

- 要素：
  - 智能体
  - 环境
  - 状态 s
  - 动作 a
  - 奖励 r

强化学习 (Reinforcement Learning, RL)：通过智能体与环境交互作用的一种试错学习范式

智能体

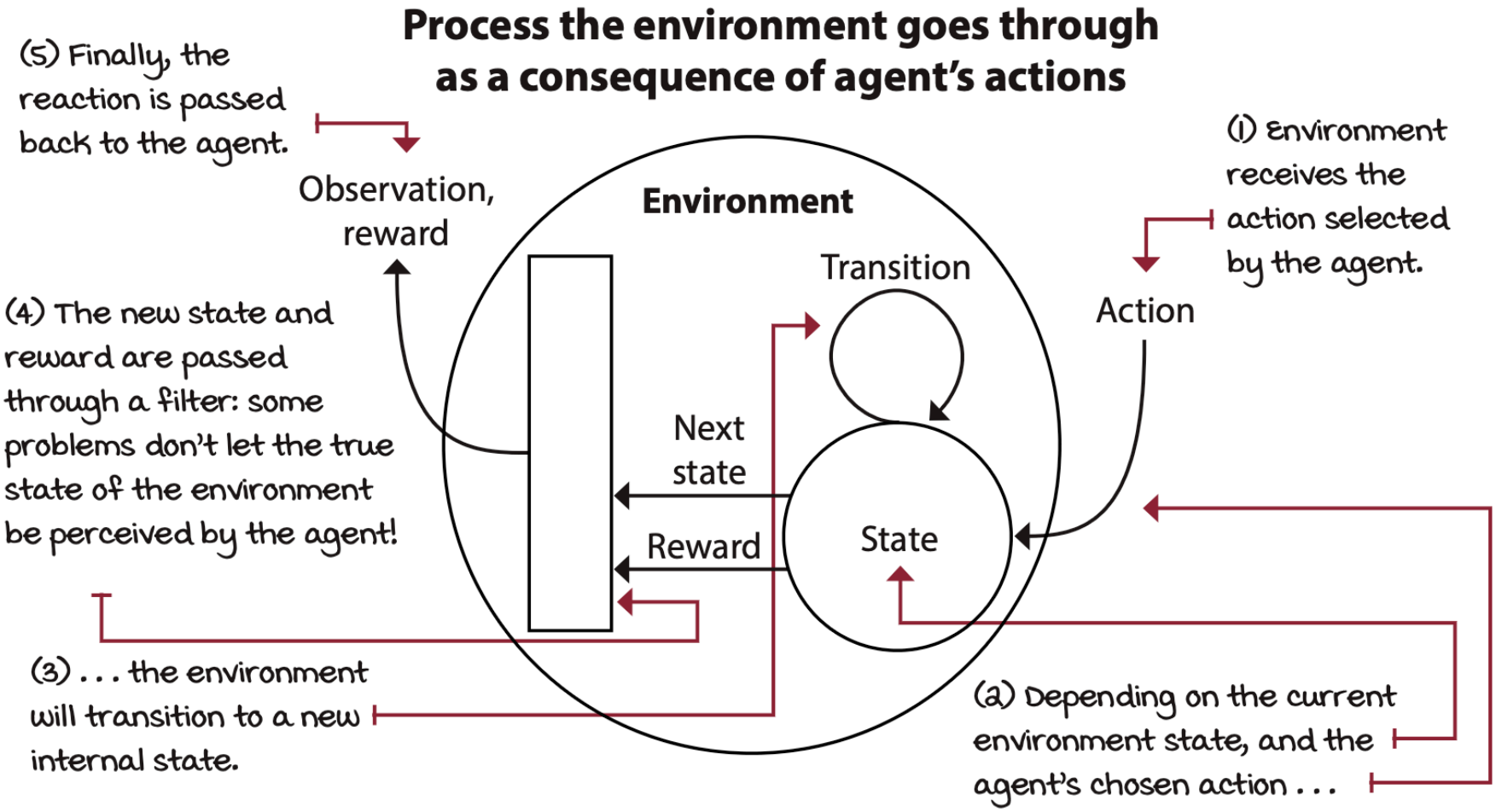


- 智能体：决策者
  - 和环境交互
  - 评估动作的好坏
  - 学习并改进动作选择策略

环境

- 环境：除了智能体之外的一切
- 通常选择会影响决策的变量组合（状态）代表环境，状态所有可能的取值空间称为状态空间 / state space。
- 可能会有一些状态变量是智能体看不到的，智能体可以看到的状态变量称为观测，可观测到的变量的取值空间称为观测空间 / observation space。

环境


Process the environment goes through as a consequence of agent's actions
(5) Finally, the reaction is passed back to the agent.
Observation, reward
Environment
(1) Environment receives the action selected by the agent.
Transition
Action
(4) The new state and reward are passed through a filter: some problems don't let the true state of the environment be perceived by the agent!
Next state
Reward
State
(3) . . . the environment will transition to a new internal state.
(2) Depending on the current environment state, and the agent's chosen action . . .

动作与状态转移

- 环境：除了智能体之外的一切
- 在每一时刻，根据当前状态，环境会决定智能体可选择的动作有哪些，所有可能的动作组成动作空间
- 环境在接收到智能体的动作后，会做出以下响应：
  - 通过其状态转移函数 / transition function 改变自身状态
  - 在完成其自身的状态转移后，环境会通过其奖励函数 / reward function 释放出一个瞬时奖励信号
- 状态转移函数和奖励函数统称为环境模型 / model of the environment

智能体与环境

- 在t时刻，智能体：
  - 执行动作A_t
  - 收到观察O_t
  - 收到标量的奖励R_t
- 环境：
  - 收到动作A_t
  - 释放观察O_{t+1}
  - 释放标量奖励R_{t+1}
- 环境决定了时刻t如何变化

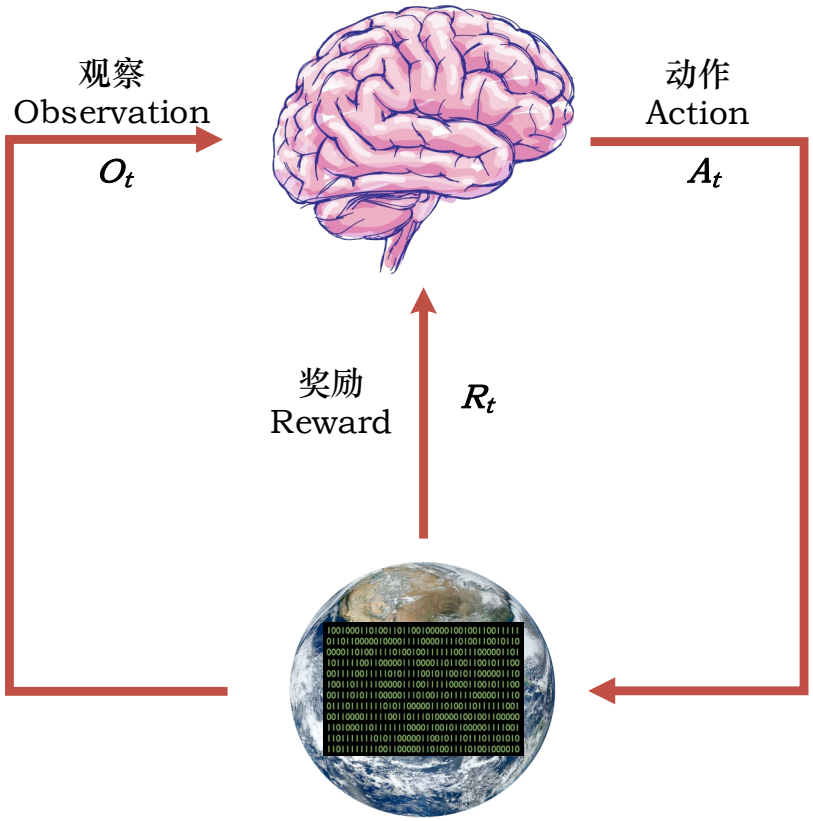
状态

- 状态（State）：决定下一步做什么所需要的信息
  - 状态是历史的函数：$S_t = f(H_t)$
  - （即从历史中提取必要信息）
- 历史（History）：观察（Observation）、动作（Action）和奖励（Reward）的序列

$$H_t = O_1, R_1, A_1, \ldots, A_{t-1}, O_t, R_t$$

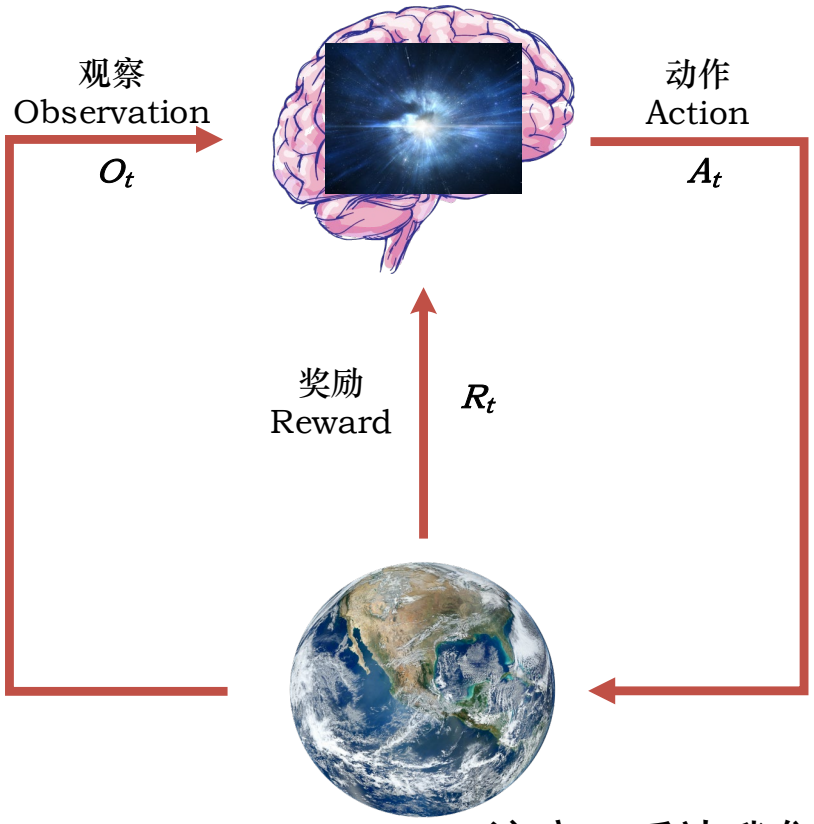
  - 历史包括了截止到t时刻所有能观察到的变量
  - 例如：机器人的全部运动传感器的数据流
- 根据历史做决定：→ 需要所有历史信息么？Usually No

环境状态Environment State



- 环境状态S_t^e是环境的自我刻画private representation：
  - 即，用来决定下一步[观察/奖励]所需的全部信息。
- 环境状态对智能体并不是全部可见的。
- 即使S_t^e可见，它可能含有诸多无关信息。

智能体状态Agent State

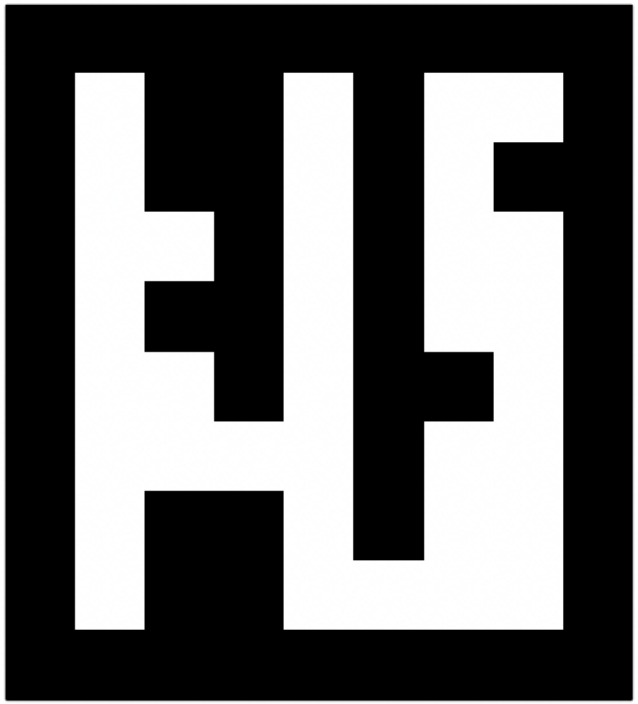


- 智能体状态S_t^a是智能体的内部表达internal representation：
  - 即，用来决定下一步[动作]所需的全部信息。
  - 即，强化学习算法所需要使用的信息。

注意：后边我们用 S_t 指代智能体状态，而不是环境状态

状态示意（以迷宫为例）

环境状态s_t^e



智能体状态s_t^a

环境状态：整个迷宫的布局
智能体状态：只能看到自己当前所处的位置

环境状态：用来决定下一步[观察/奖励]所需的全部信息
智能体状态：用来决定下一步[动作]所需的全部信息

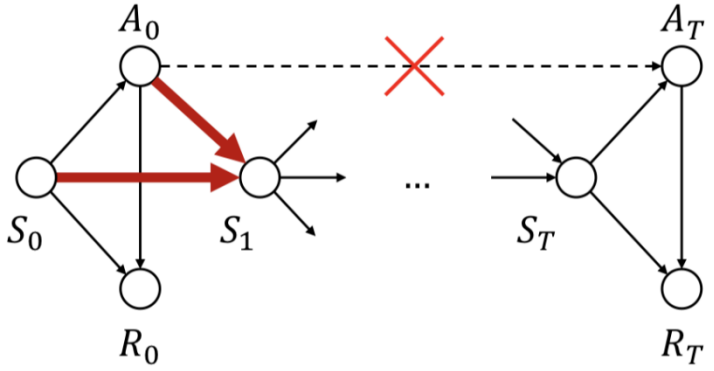
马尔科夫性质

- 信息状态（Information State）：某状态包含了历史中的全部有用信息。→ 马尔科夫状态（Markov State）
- 状态S_t具备马尔科夫性质，当且仅当满足下式：

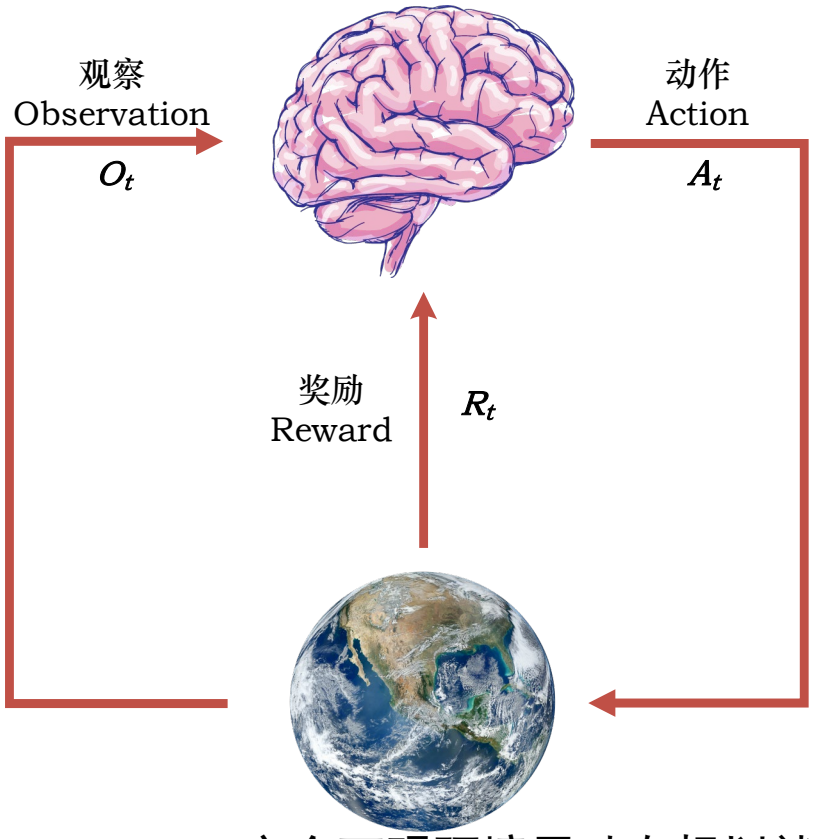
$$P[S_{t+1}|S_t] = P[S_{t+1}|S_t, \dots, S_1]$$

  - 下一时刻的状态仅由当前状态决定，与过去的状态无关；
  - [在给定现在状态时，它与过去状态（即该过程的历史路径）是条件独立的]
  - 状态确定后，可以无需考虑历史：无记忆性/无后效性

强化学习的基石就是马尔可夫决策过程（Markov Decision Processes, MDP）
对于时序决策问题具有很好的建模能力



完全可观环境



完全可观环境用动态规划就可解决
不存在学习的部分

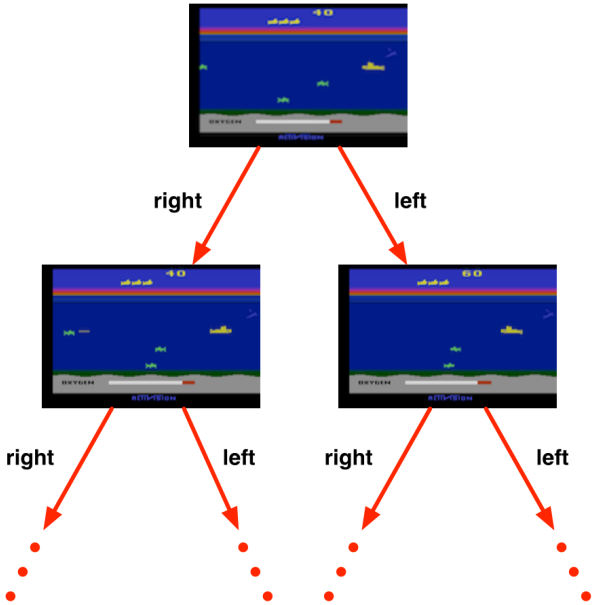
- 完全可观fully observable：智能体可以直接看到所有的环境状态：

$$O_t = S_t^a = S_t^e$$

  - 智能体状态=环境状态
- 问题的马尔科夫性质：马尔科夫决策过程Markov Decision Process (MDP)
  - 运筹学II，动态规划部分将会详细介绍。

雅达利游戏：规划的方式求解

- 游戏运作方式已知（有模型）
  - 相当于可以向一个完美模型可供查询
- 如果在状态s下采取动作a
  - 下一个状态是？
  - 分数会变成？
  - 统统已知！
- 能够在执行前找到最优策略
  - 例：树搜索



部分可观环境

- 部分可观：Partially observable：智能体不能直接观察环境
  - 视觉伺服机器人并不知道自己的准确位置
  - 竞价智能体只能看到当前的出价
  - 打牌机器人只能看台面上的牌
- 这种情况下，智能体状态 ≠ 环境状态
- 部分可观马尔科夫决策过程Partially observable Markov decision process (POMDP)
- 智能体必须自行建立其状态表达S_t^a
  - 例如：全部历史信息$S_t^a = H_t$、循环神经网络、贝叶斯模型等

部分可观，因为有些不知道，所以要通过学习，用强化学习解决部分可观环境

奖励

- 奖励 R_t 是一种标量的反馈信号；
- 反映了在 t 时刻，智能体的Action的好坏；
- 智能体的任务：最大化累积奖励cumulative reward。

回报（Return）$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \ldots$　　吃到的鸡腿数量最大化

强化学习基于奖励假设：

➢所有的目标都可以描述为某种期望的累积奖励的最大化。

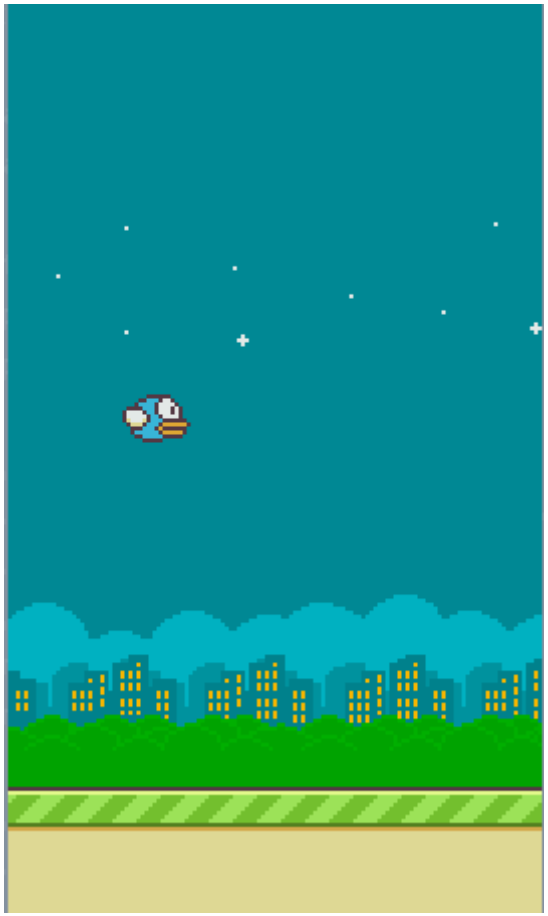
你同意吗?

Ref

[1] Silver David, Singh Satinder, Precup Doina, Sutton Richard S.. Reward Is Enough[J]. Artificial Intelligence, 2021(prepublish)

奖励-例子

- 机器人运动控制：
  - 获得+奖励：跟随了预定轨迹
  - 获得-奖励：发生碰撞
- 控制发电站
  - 获得+奖励：正常发出电力
  - 获得-奖励：违反安全约束
- 玩电子游戏
  - 获得+奖励：获得高积分
  - 获得-奖励：死亡



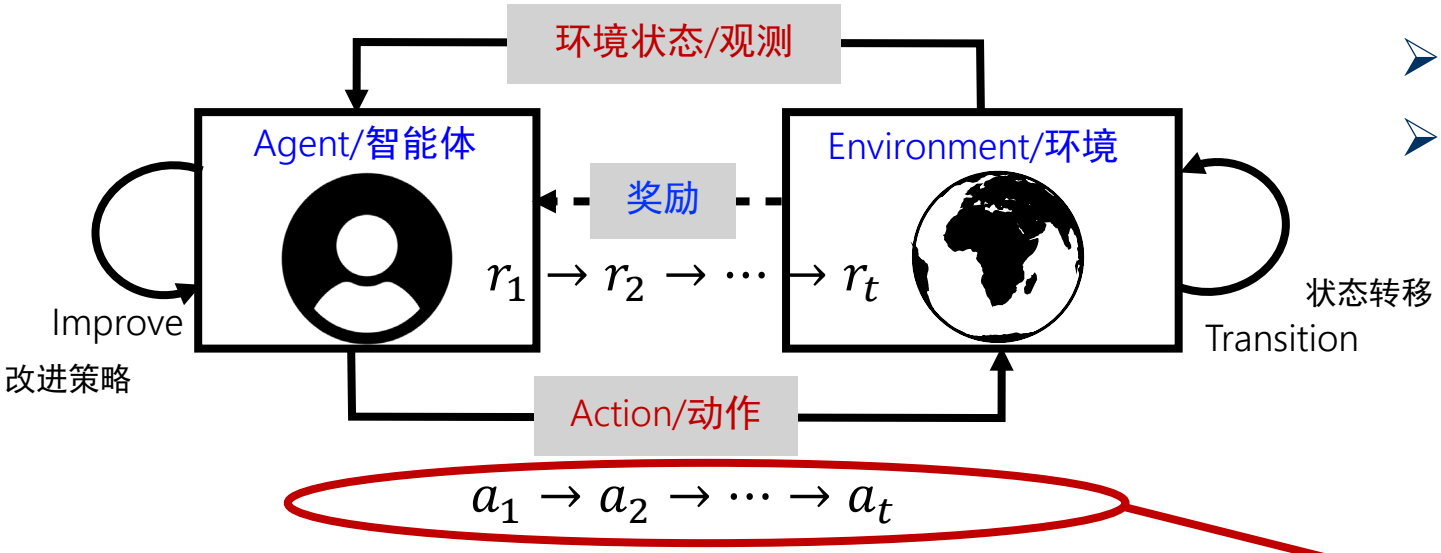
序贯决策

- 目标：选择合适的动作来最大化未来总奖励
- 动作可能造成长期后果
- 延迟奖励
- 可能需要牺牲短期的奖励来获得长期回报
- 例子：
  - 金融投资（需要数月甚至一年才能得到结果）
  - 直升机加油（预防数小时后的坠机）
  - 围棋的一着（可能帮助了许多步后棋局的胜利）

强化学习模型

- 基本模型：

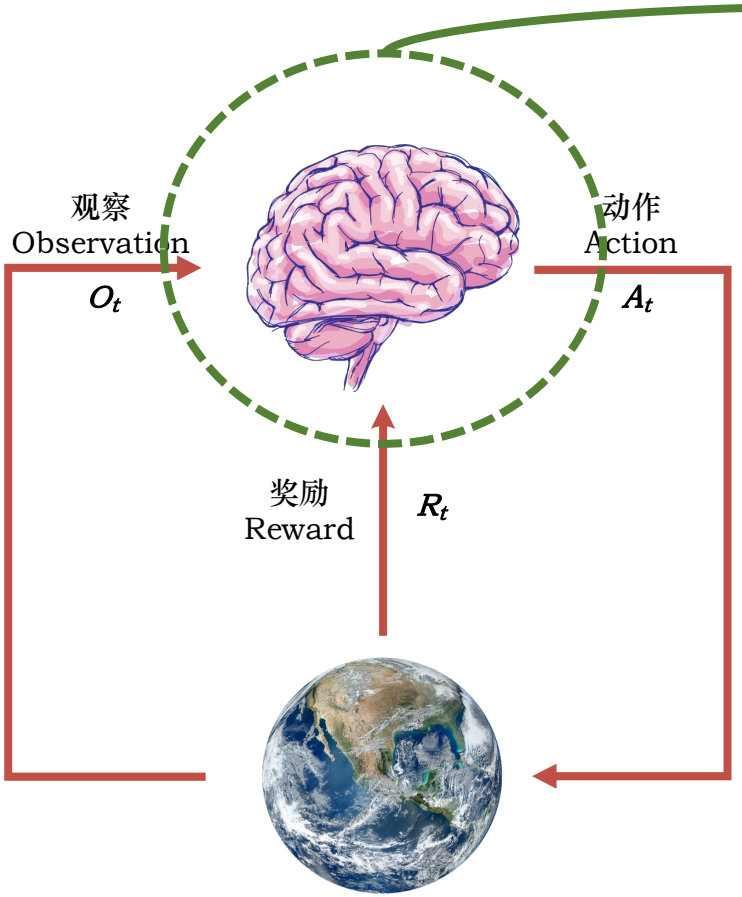
$$s_1 \rightarrow s_2 \rightarrow \cdots \rightarrow s_t$$
$$o_1 \rightarrow o_2 \rightarrow \cdots \rightarrow o_t$$



- 要素：
  - 环境
  - 智能体
  - 状态 s
  - 动作 a
  - 奖励 r

每个时刻，都进行如上图所示的循环，RL是用来解决序贯决策问题的，笼统地讲，RL包括给定状态预测奖励的预测算法和通过试错来学习好的动作的控制算法。

智能体学习的核心要素

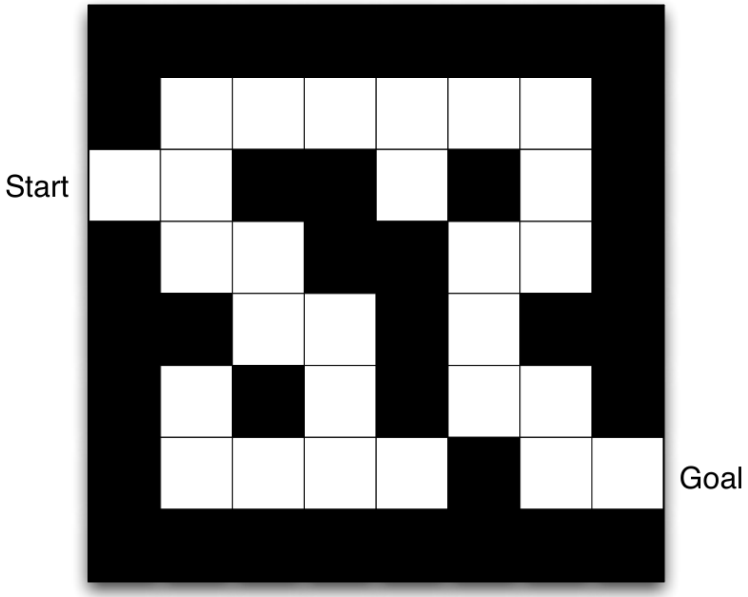


RL智能体的核心构成：

- 策略（Policy）：决定智能体行为的函数
- 值函数（Value function）：评估状态/动作好坏的函数
- 模型（Model）：智能体对环境的表达（主观认识）

例子：迷宫

目标是尽可能快地到达终点

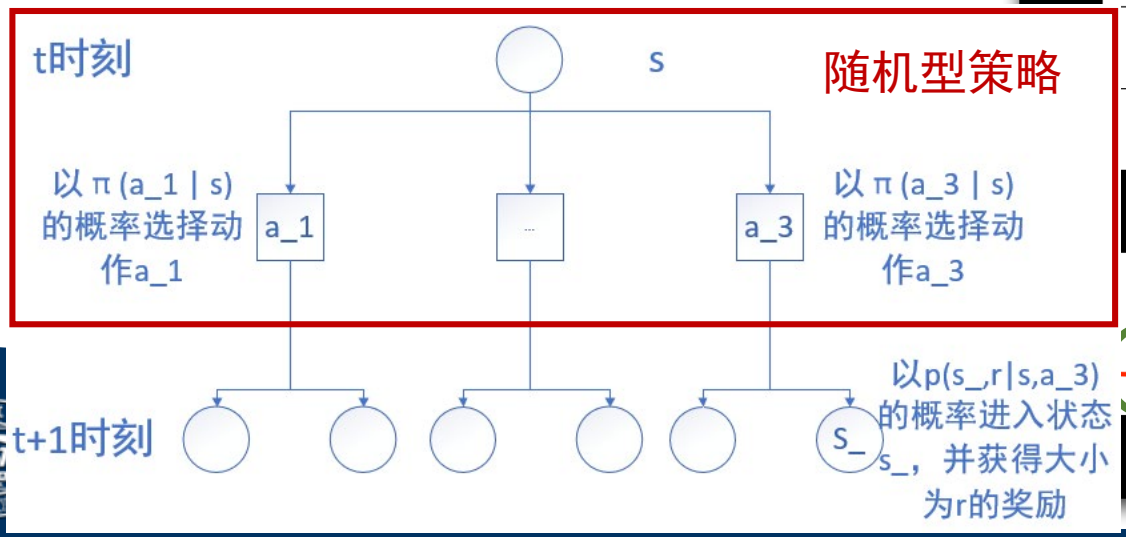


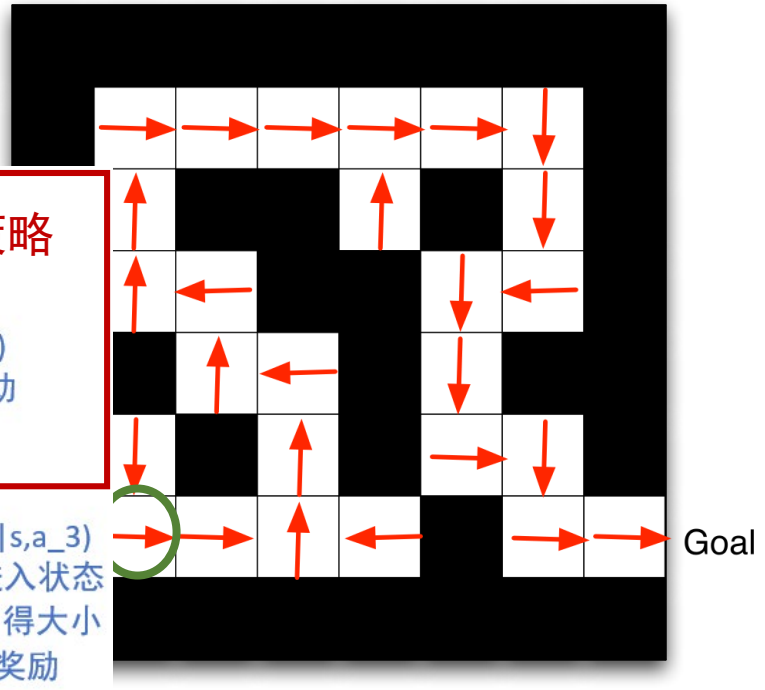
- 奖励：每走一步-1
- 动作：东、南、西、北四方向（用上、下、左、右箭头表示）
- 状态：智能体的位置

策略（Policy）

- 策略（Policy）代表了智能体的行为函数
- 表示从状态到动作的映射，比如：
  - 确定性策略：$a = \pi(s)$
  - 随机型策略（处于状态s时采取动作a的概率）

$$\pi(a|s) = P[A_t = a|S_t = s]$$





值函数（Value Function）

回报：累积奖励

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \dots$$

- 值函数是对未来奖励的预测
- 用来衡量状态的好坏（未来回报的期望）
  - 状态值函数：$V_\pi(s) = E_\pi\left[G_t | S_t = s\right]$
  - 动作值函数：$Q_\pi(s,a) = E_\pi[G_t | S_t = s, A_t = a]$

策略π下agent处于状态s时，未来的回报期望

$$V_\pi(s) = E_\pi\left[G_t | S_t = s\right]$$

依赖于策略

$$= E_\pi\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = s\right]$$

折扣因子

- RL目标：$V^*(s) = \max_\pi V_\pi(s)$或$Q^*(s,a) = \max_\pi Q_\pi(s,a)$

值函数（Value Function）

- 回报（Return）可写作递归形式：

$$G_t = R_{t+1} + \gamma G_{t+1}$$

- 值函数也可写作递归形式：

状态值函数

$$V_\pi(s) = E_\pi\left[R_{t+1} + \gamma G_{t+1} | S_t = s\right]$$

$$= E_\pi\left[R_{t+1} + \gamma V_\pi(S_{t+1}) | S_t = s\right]$$

贝尔曼方程
（动态规划方程）

动作值函数 $Q_\pi(s,a) = E_\pi\left[R_{t+1} + \gamma Q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a\right]$

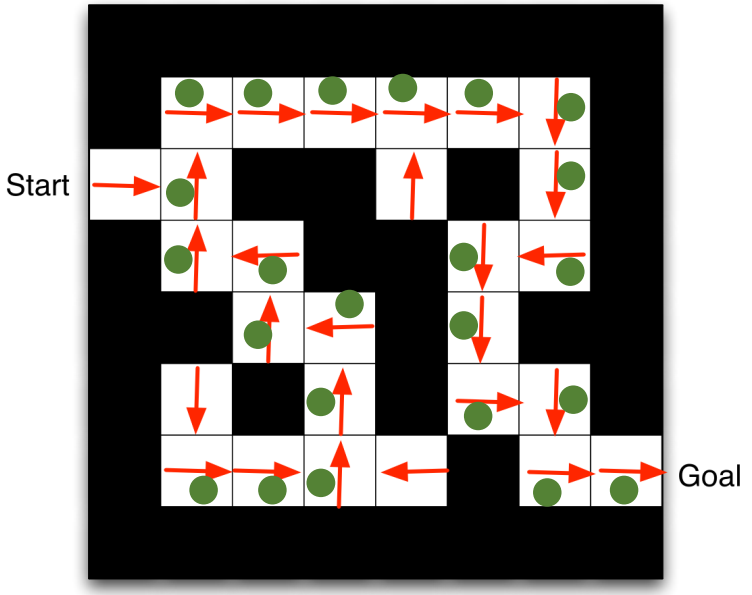
$$V_\pi(s) = \sum_{a \in A} \pi(a|s) Q_\pi(s,a)$$

值函数例子

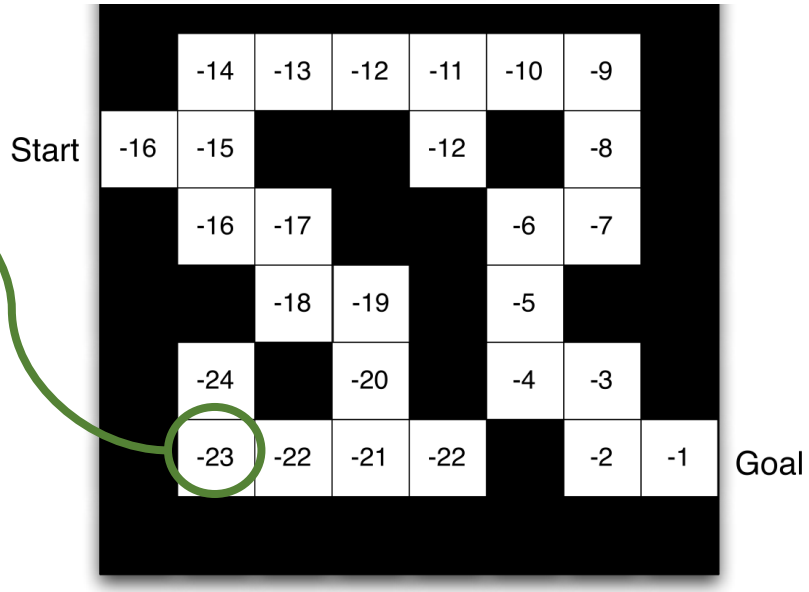
数字表示在每个状态（位置）下的值函数$V_\pi(s)$

本例中：

- 固定策略
- 折扣因子为1
- 每走一步奖励是-1



若agent处在该位置，则根据策略$\pi(s)$，得到该状态下的值函数$V_\pi(s)$为-23



值函数（Value Function）

- 状态值函数：
  - 用来评估给定策略的好坏
  - 需联合MDP模型才能找到最佳策略
  - 解决预测问题
- 动作值函数：
  - 用来改进给定策略，
  - 不需要MDP就可以给出最佳策略
  - 解决控制问题

状态值函数：$V_\pi(s) = E_\pi\left[G_t|S_t = s\right]$

动作值函数：$Q_\pi(s,a) = E_\pi[G_t|S_t = s, A_t = a]$

模型Model

- 模型：预测环境的接下来如何变化
- 转移模型：根据先前的状态以及动作，预测环境所处下一个状态的概率 $\mathcal{P}_{ss'}^{a} = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$
- 奖励模型：根据先前的状态以及动作，预测下一个瞬时奖励的概率 $\mathcal{R}_s^a = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$