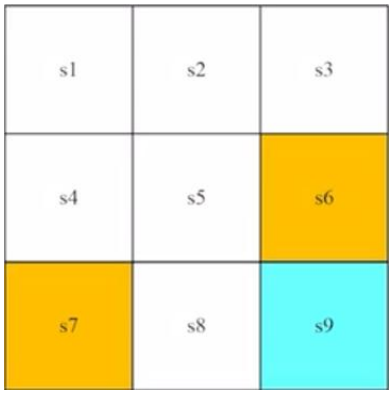
1. State & State space

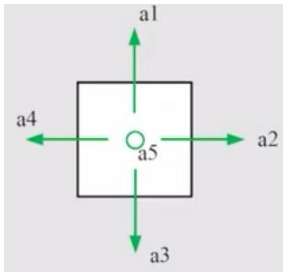
State：表示当前的状态。例如在以下网格中，共有s_1，s_2，…，s_9九个状态

State space：状态空间，表示所有状态的集合，$S = \{s_i\}_{i=1}^{9}$



2. Action & Action space of a state

Action：每一个 state 接下来要采取的行动，比如：



Action space of a state：在一个状态下可能采取的所有行动的集合，表示为$A(s_i) = \{a_i\}_{i=1}^{5}$

3. State transition：状态转移

例如我们在s_1，分别采取a_1和a_2两种动作，分别表示为：

$$s_1 \xrightarrow{a_1} s_1$$

$$s_1 \xrightarrow{a_2} s_2$$

4. Tabular representation：表格表示

将以上所有状态，采取不同行动的可能结果用表格表示为：

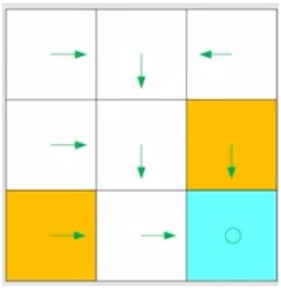
	a_1 (upwards)	a_2 (rightwards)	a_3 (downwards)	a_4 (leftwards)	a_5 (unchanged)
s_1	s_1	s_2	s_4	s_1	s_1
s_2	s_2	s_3	s_5	s_1	s_2
s_3	s_3	s_3	s_6	s_2	s_3
s_4	s_1	s_5	s_7	s_4	s_4
s_5	s_2	s_6	s_8	s_4	s_5
s_6	s_3	s_6	s_9	s_5	s_6
s_7	s_4	s_8	s_7	s_7	s_7
s_8	s_5	s_9	s_8	s_7	s_8
s_9	s_6	s_9	s_9	s_8	s_9

5. State transition probability：状态转移的概率

$p(s_2|s_1, a_2) = 1$ 表示在s_1的状态下采取a_2的行动，到达s_2的概率为 1

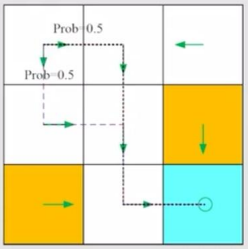
$p(s_i|s_1, a_2) = 0 \ \ \forall i \neq 2$ 表示在s_1的状态下采取a_2的行动，到达任何非s_2的概率为 0

6. Policy：策略，表示在某个 state 该采取的 action，如图所示：



7. Mathematical representation：用数学概率来描述每个状态采取某种策略的概率，如图所示：

For example:



In this policy, for s_1:

$$\pi(a_1|s_1) = 0$$
$$\pi(a_2|s_1) = 0.5$$
$$\pi(a_3|s_1) = 0.5$$
$$\pi(a_4|s_1) = 0$$
$$\pi(a_5|s_1) = 0$$

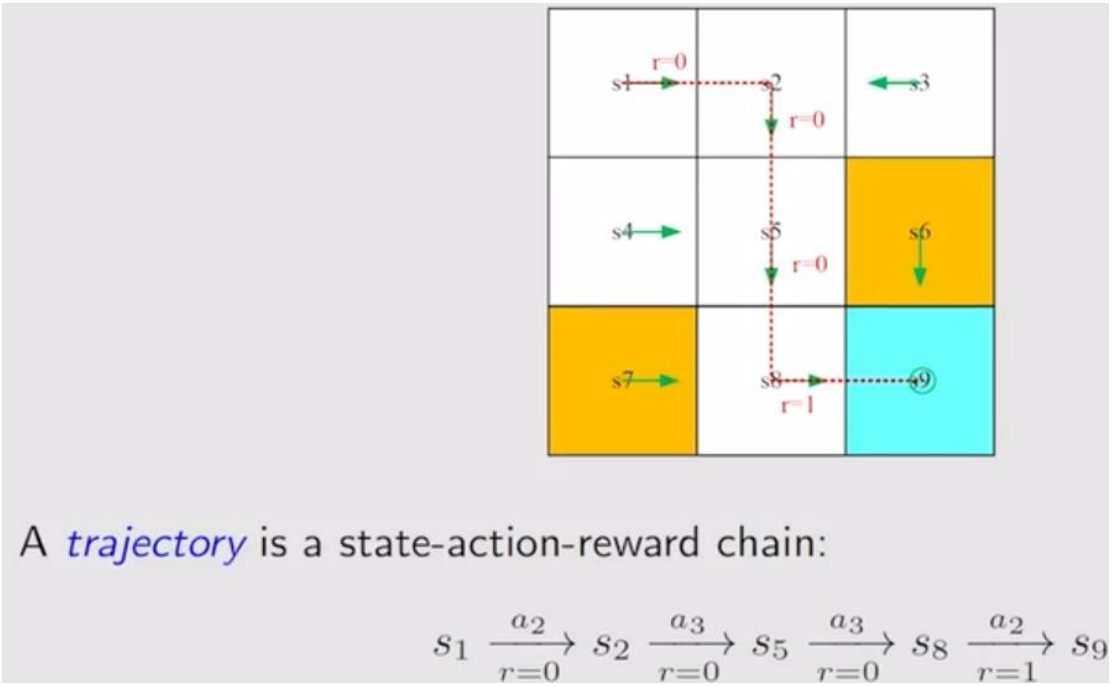
8. Tabular representation of a policy：采取表格形式来表示每个状态采取某种策略的概率，如下所示：

	a_1 (upwards)	a_2 (rightwards)	a_3 (downwards)	a_4 (leftwards)	a_5 (unchanged)
s_1	0	0.5	0.5	0	0
s_2	0	0	1	0	0
s_3	0	0	0	1	0
s_4	0	1	0	0	0
s_5	0	0	1	0	0
s_6	0	0	1	0	0
s_7	0	1	0	0	0
s_8	0	1	0	0	0
s_9	0	0	0	0	1

9. Reward：反馈，当我们采取某个行动之后得到的一个数值

- 正数：奖励，代表对该行为进行鼓励
- 负数：惩罚，代表不希望该行为发生
- 整数 0 表示不惩罚
- 可以反过来进行表示，即正数惩罚，负数奖励

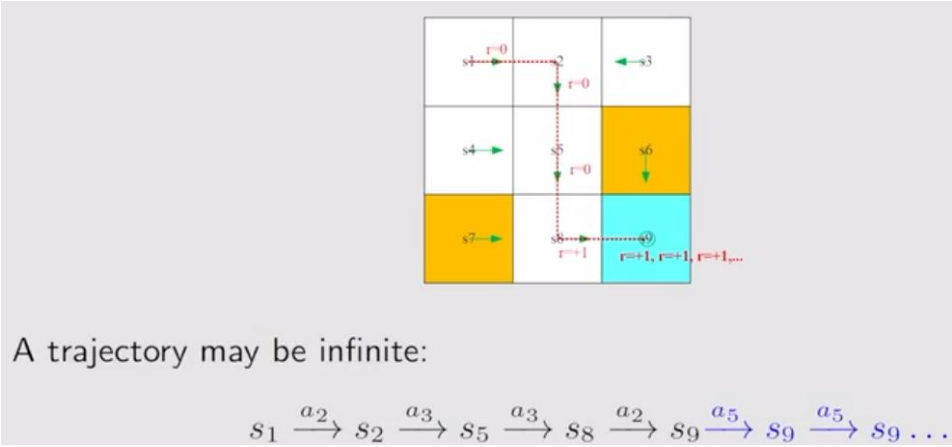
10. Trajectory：轨迹，用来表示 state-action-reward 链，如图所示：



11. Return：针对一个 trajectory 而言，将该轨迹上的所有 reward 统计起来的结果，例如上述 trajectory 的$return = 0 + 0 + 0 + 1 = 1$

12. Discount rate：$\gamma \in [0,1)$

若在上述例子中，有一条轨迹是这样无穷的：



那该 trajectory 的$return = 0 + 0 + 0 + 1 + 1 + \ldots + 1 = \infty$

此时 return 发散，该结果无意义。为了解决这个问题，引入γ，得到

$discounted\ return = 0 + \gamma 0 + \gamma^2 0 + \gamma^3 1 + \ldots = \gamma^3(1 + \gamma + \gamma^2 + \ldots) = \frac{\gamma^3}{1-\gamma}$

γ的作用：

- 得到的 discounted return 收敛，结果有意义
- 可以平衡更远或更近未来的 reward：
  - 若γ更接近于 0，则更在意前面的 action，考虑更近的未来
  - 若γ更接近于 1，则更在意后面的 action，考虑更远的未来

13. Episode：有限的 trajectory

14. Markov property：马尔可夫性质，即与历史无关性质

- $p(s_{t+1}|a_{t+1}, s_t, \ldots, a_1, s_0) = p(s_{t+1}|a_{t+1}, s_t)$，从$s_0$采取行动一直到$s_t$后，采取

行动a_{t+1}到达s_{t+1}的概率与直接从s_t采取行动a_{t+1}到达s_{t+1}的概率相等，即与历史路径无关

- $p(r_{t+1}|a_{t+1}, s_t, \ldots, a_1, s_0) = p(r_{t+1}|a_{t+1}, s_t)$，从$s_0$采取行动一直到$s_t$后，采取行动$a_{t+1}$到达$s_{t+1}$所获得 reward 的概率与直接从$s_t$采取行动$a_{t+1}$到达$s_{t+1}$所获得 reward 的概率相等，即与历史路径无关