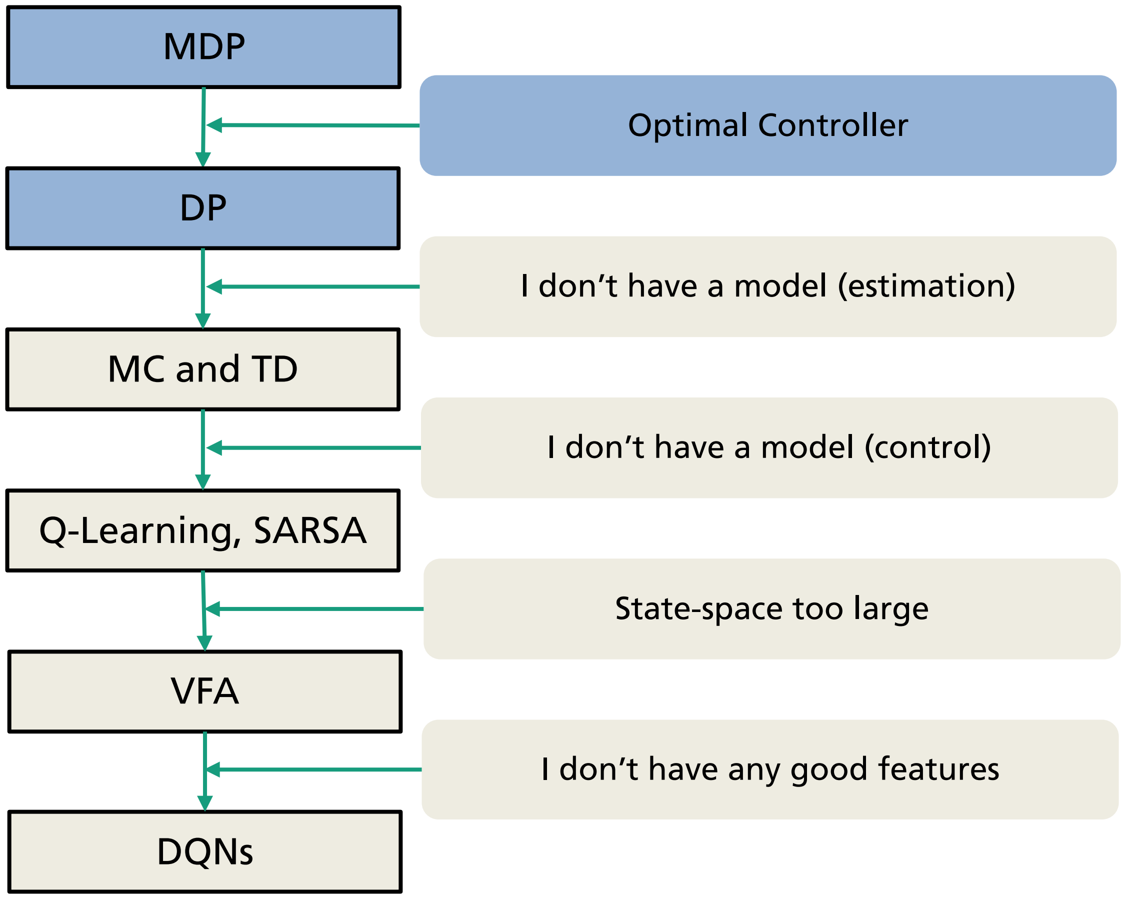
Introduction to Dynamic Programming

Christopher Mutschler

Overview

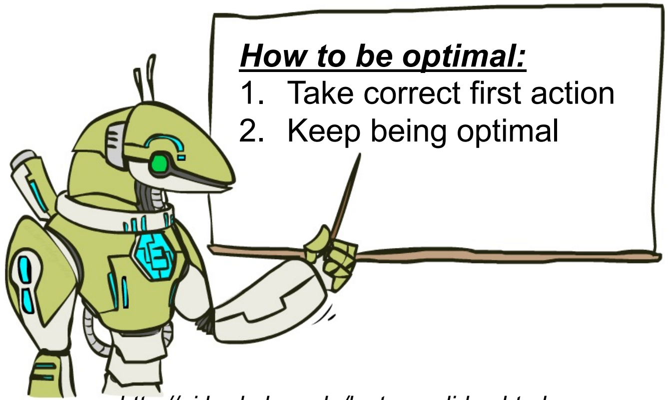

MDP
Optimal Controller
DP
I don’t have a model (estimation)
MC and TD
I don’t have a model (control)
Q-Learning, SARSA
State-space too large
VFA
I don’t have any good features
DQNs

Dynamic Programming

- How do we find **optimal** controllers for given (known) MDPs?
- Bellman equation & Bellman’s principle of optimality

Principle of Optimality:
„An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.“

(see Bellman, 1957, Chap. III.3.)

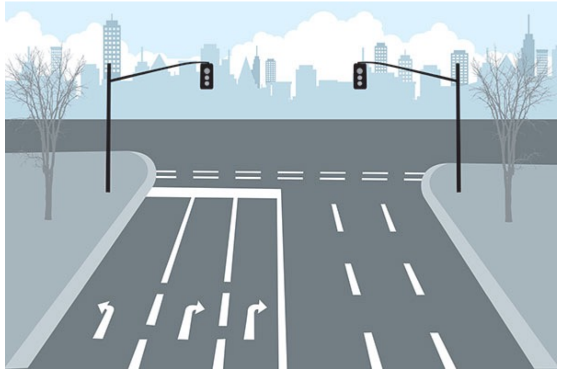


http://ai.berkeley.edu/lecture_slides.html

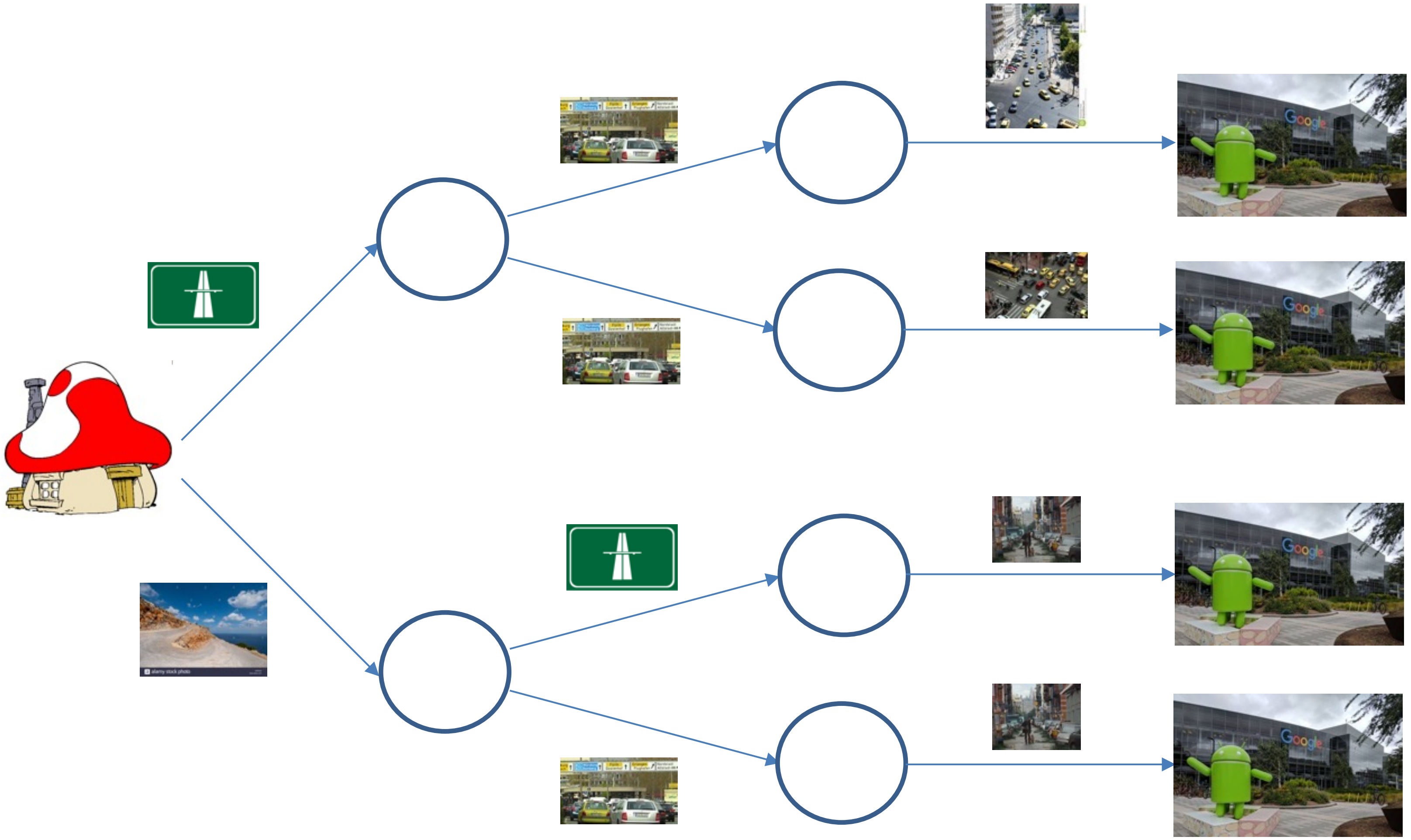
Dynamic Programming

Example # 1 (Simplistic)

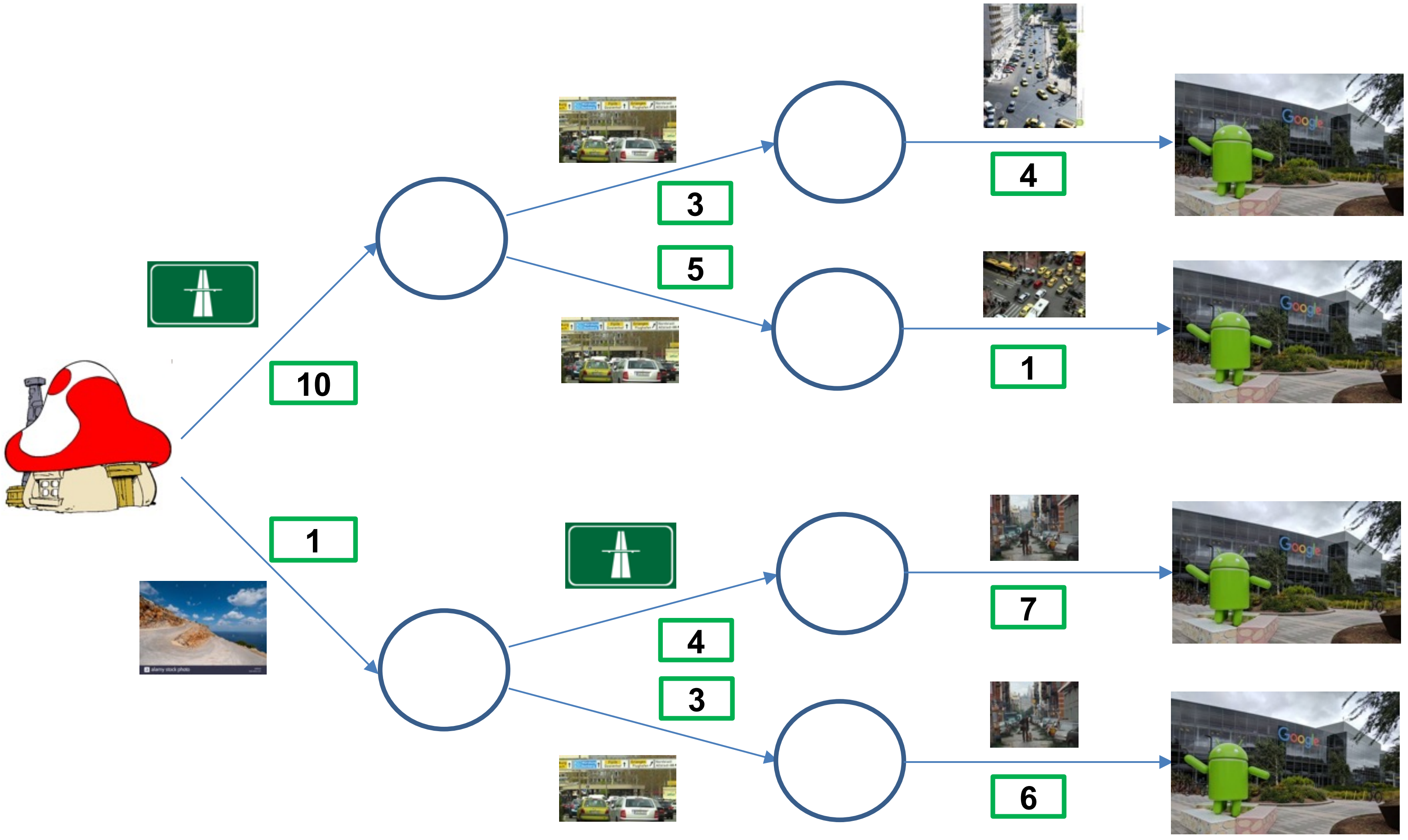
- We need to go from our house to our work as fast as possible
- Actions: Left, Right
- Different actions lead to different road segments (e.g., highway, country road, etc.)
- Reward: $-t$, where t is the time needed for each road segment



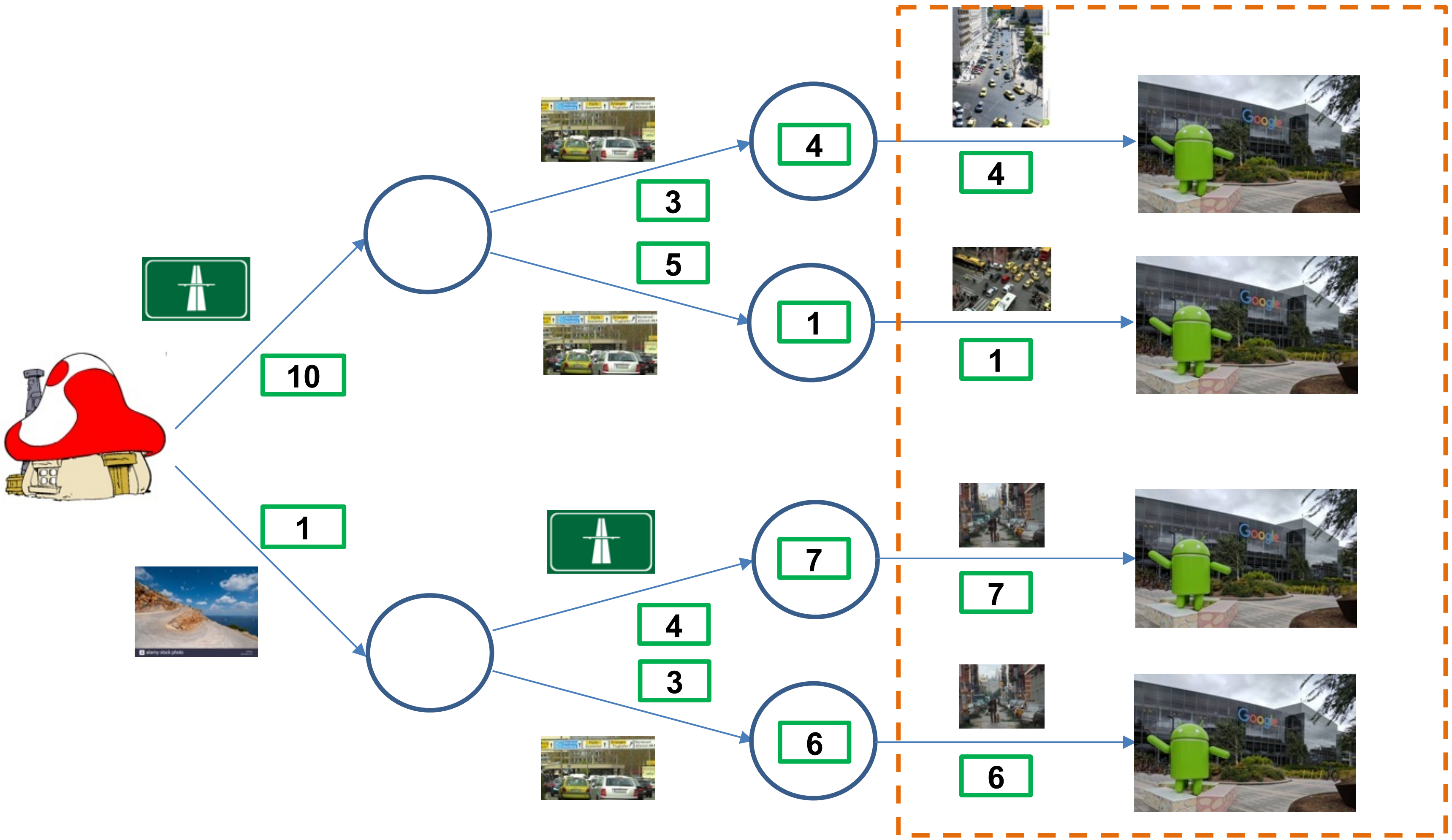
Dynamic Programming



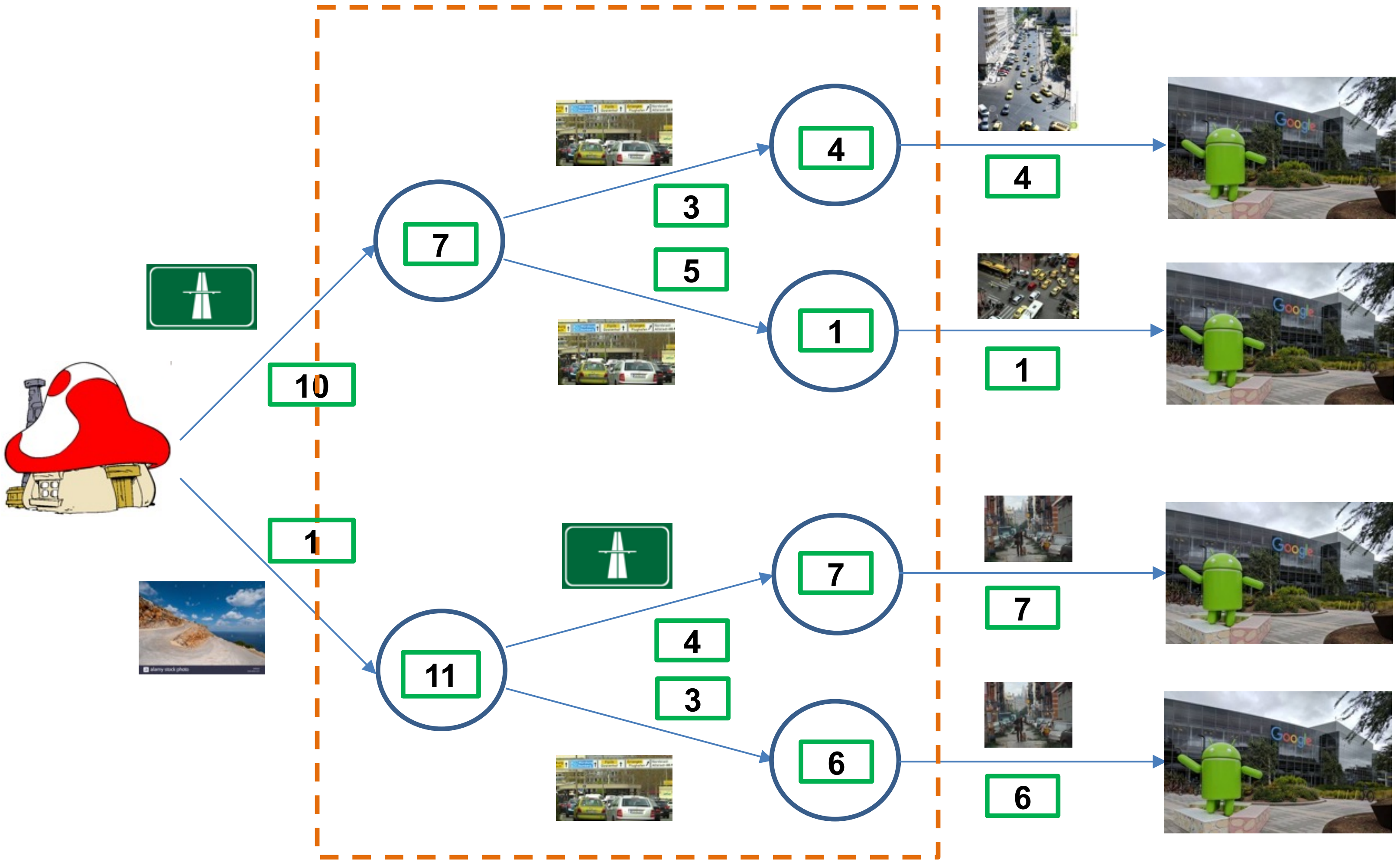
Dynamic Programming



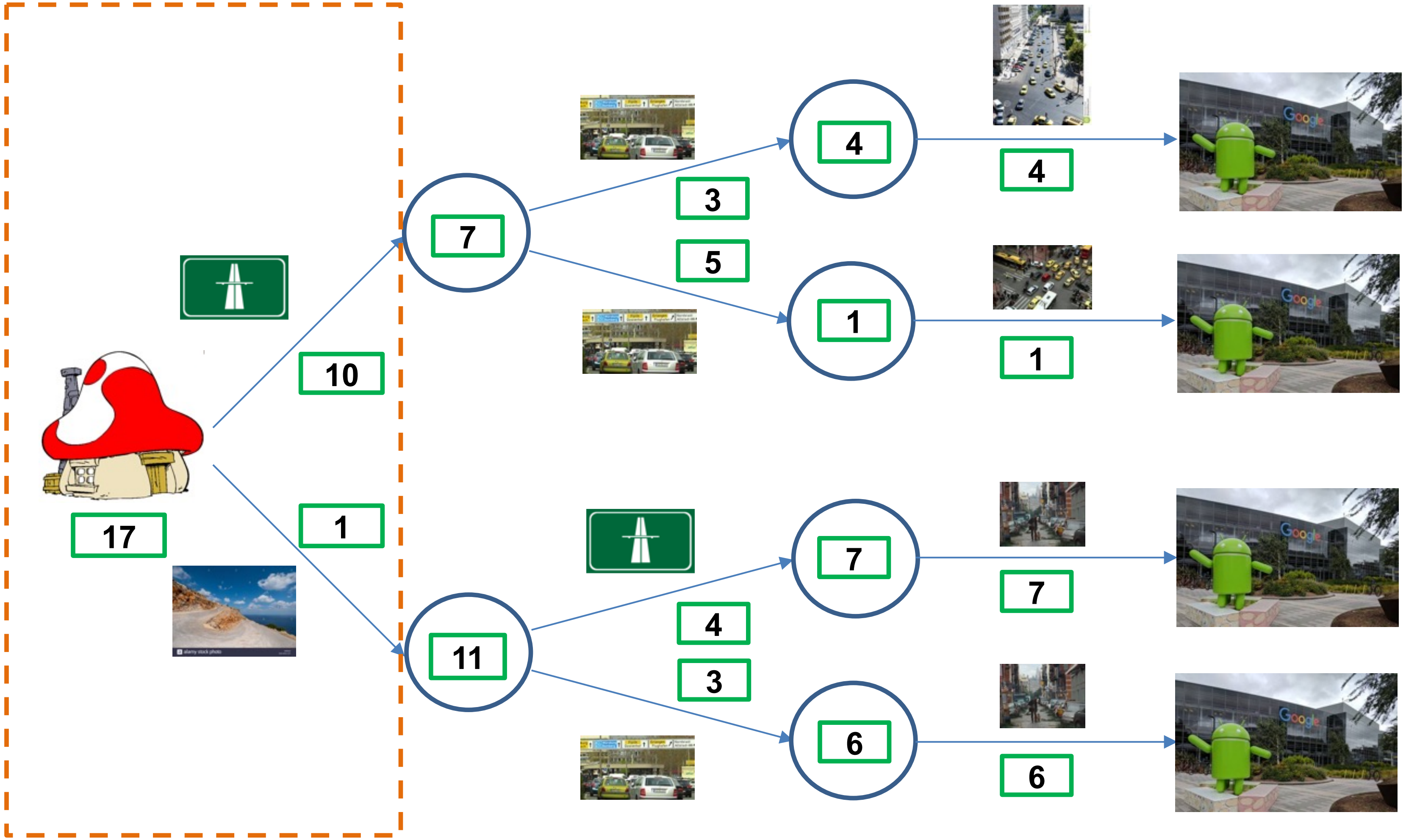
Dynamic Programming



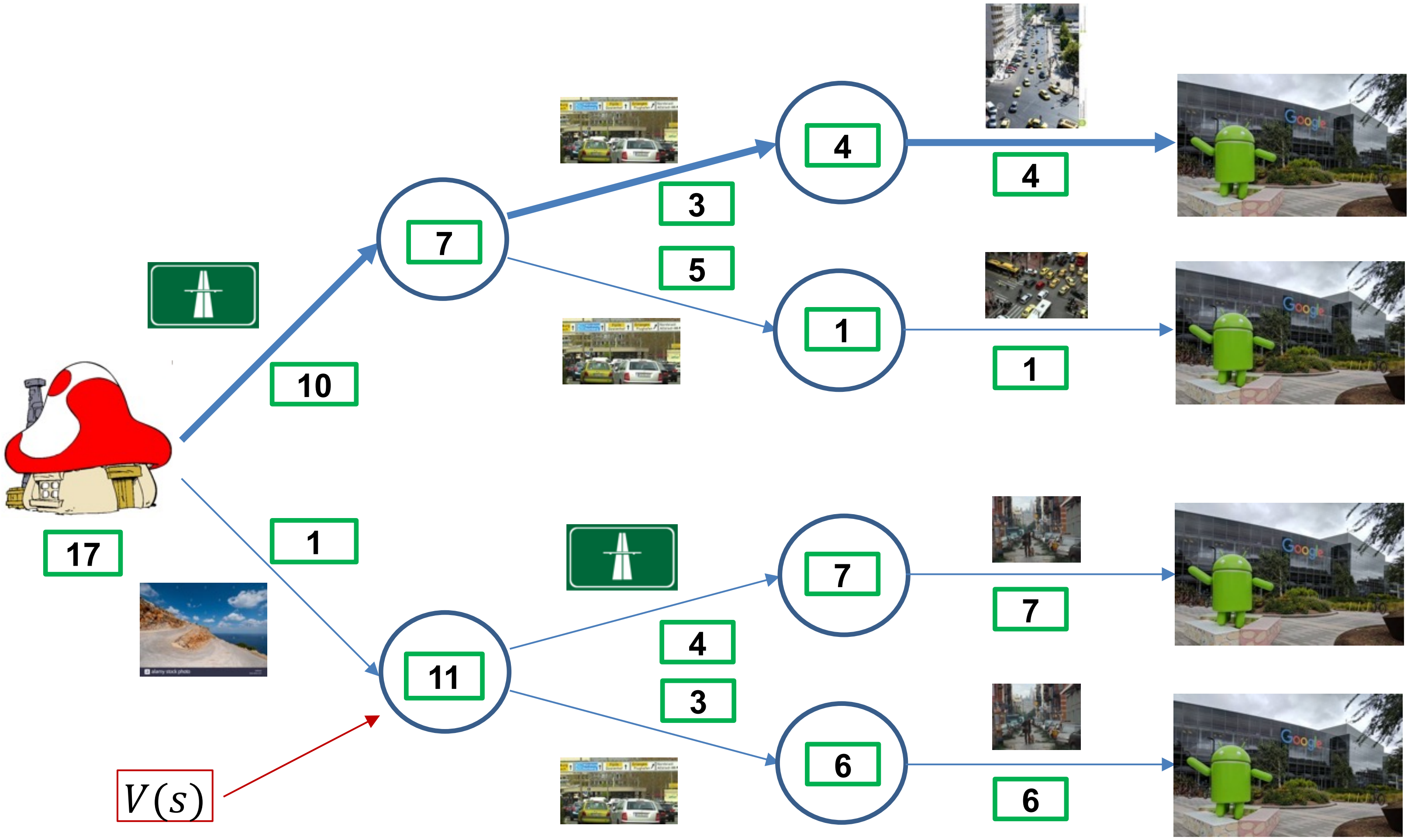
Dynamic Programming



Dynamic Programming

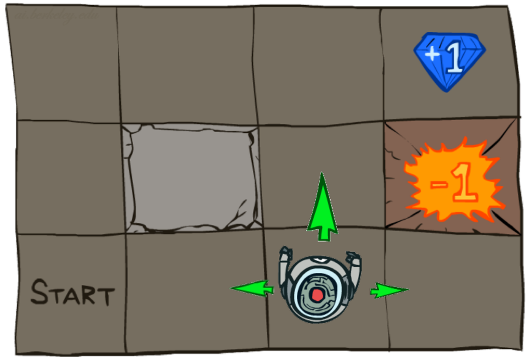


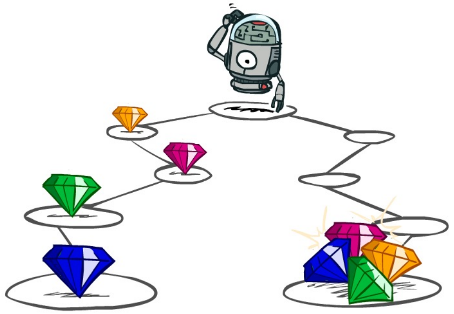
Dynamic Programming



Dynamic Programming

Example #2 (Simplistic)



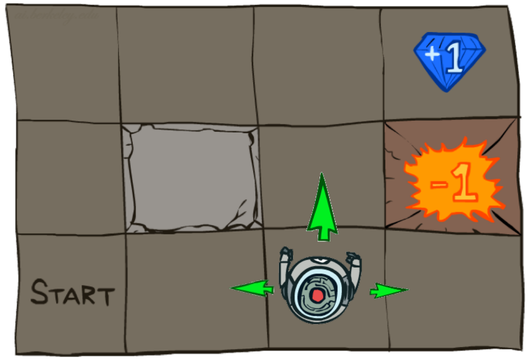


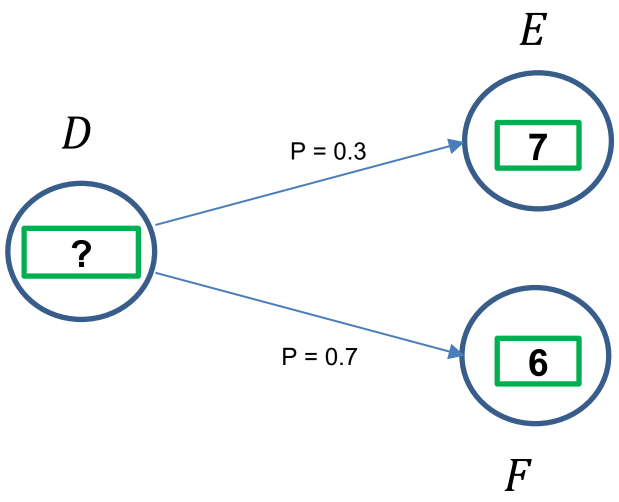
http://ai.berkeley.edu/lecture_slides.html

Dynamic Programming

Example #2 (Simplistic)

$$P(s' = E, | a = 0, s = \mathrm{D}) = 0.3$$
$$P(s' = F, | a = 0, s = \mathrm{D}) = 0.7$$

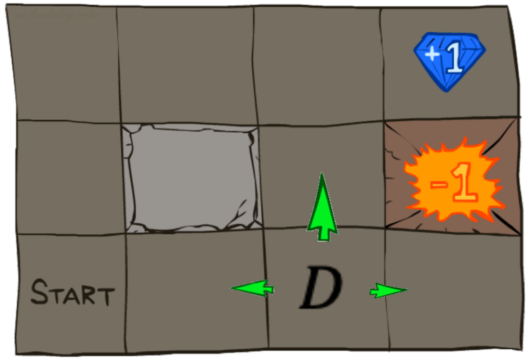


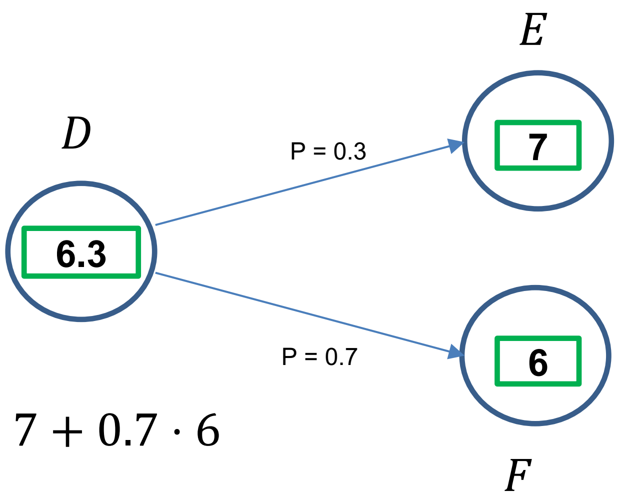


Dynamic Programming

Example #2 (Simplistic)

$$P(s' = E, | a = 0, s = \mathrm{D}) = 0.3$$
$$P(s' = F, | a = 0, s = \mathrm{D}) = 0.7$$





$$0.3 \cdot 7 + 0.7 \cdot 6$$

Dynamic Programming

- How do we find **optimal** controllers for given (known) MDPs?
- Unfortunately, we need some definitions:
  - state-value function V for policy π

$$s_0 \xrightarrow{\pi(s_0), r(s,\pi(s_0))} s_1 \xrightarrow{\pi(s_1), r_1} s_2 \xrightarrow{\pi(s_2), r_2} s_3 \dots \; s_{h-1} \xrightarrow{\pi(s_{h-1}), r_{h-1}} s_h$$

$$V^\pi(s) \triangleq Q^\pi\big(s, \pi(s)\big) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s\right]$$

  - state-action-value function Q for policy π

$$s_0 \xrightarrow{a, r_0} s_1 \xrightarrow{\pi(S_1), r_1} s_2 \xrightarrow{\pi(s_2), r_2} s_3 \dots \; s_{h-1} \xrightarrow{\pi(s_{h-1}), r_{h-1}} s_h$$

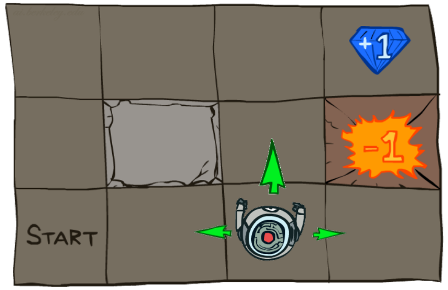
$$Q^\pi(s, a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a\right]$$

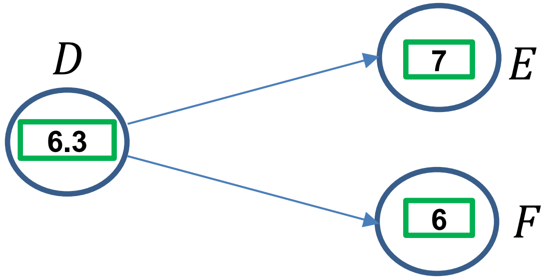
Dynamic Programming

- How do we find **optimal** controllers for given (known) MDPs?
- Unfortunately, we need some definitions:
  - Bellman Equation for V, given policy π

$$s_0 \xrightarrow{\pi(s_0), r(s,\pi(s_0))} s_1 \xrightarrow{\pi(s_1), r_1} s_2 \xrightarrow{\pi(s_2), r_2} s_3 \ldots \; s_{h-1} \xrightarrow{\pi(s_{h-1}), r_{h-1}} s_h$$

$$V^\pi(s) = \underbrace{r(s,\pi(s))}_{\text{first step}} + \underbrace{\gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,\pi(s))\, V^\pi(s')}_{\text{subsequent steps}}$$





Dynamic Programming

- How do we find **optimal** controllers for given (known) MDPs?
- Unfortunately, we need some definitions:
  - Bellman Equation for Q, given policy π

$$s_0 \xrightarrow{a, r(s,a)} s_1 \xrightarrow{\pi(s_1), r_1} s_2 \xrightarrow{\pi(s_2), r_2} s_3 \cdots s_{h-1} \xrightarrow{\pi(s_{h-1}), r_{h-1}} s_h$$

$$Q^\pi(s,a) = \underbrace{r(s,a)}_{\text{first step}} + \underbrace{\gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a)\, Q^\pi(s', \pi(s'))}_{\text{subsequent steps}}$$

Dynamic Programming

- How do we find **optimal** controllers for given (known) MDPs?
- Unfortunately, we need some definitions:
  - Bellman Optimality Equation for V

$$s_0 \xrightarrow{\pi^*(s_0),\, r(s_0,\pi(s_0))} s_1 \xrightarrow{\pi^*(s_1),\, r_1} s_2 \xrightarrow{\pi^*(s_2),\, r_2} s_3 \cdots s_{h-1} \xrightarrow{\pi^*(s_{h-1}),\, r_{h-1}} s_h$$

$$V^{\pi^*}(s) = \max_{a \in \mathcal{A}} \left\{ \underbrace{r(s,a)}_{\text{first step}} + \underbrace{\gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a)\, V^{\pi^*}(s')}_{\text{subsequent steps}} \right\}$$

Dynamic Programming

- How do we find optimal controllers for given (known) MDPs?
- Unfortunately, we need some definitions:
  - Bellman Optimality Equation for Q

$$s \xrightarrow{a,\, r(s,a)} s_1 \xrightarrow{\pi^*(s_1), r_1} s_2 \xrightarrow{\pi^*(s_2), r_2} s_3 \cdots s_{h-1} \xrightarrow{\pi^*(s_{h-1}), r_{h-1}} s_h$$

$$Q^{\pi^*}(s,a) = \underbrace{r(s,a)}_{\text{first step}} + \underbrace{\gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) \max_{a' \in \mathcal{A}} Q^{\pi^*}(s',a')}_{\text{subsequent steps}}$$