6.036: Introduction to Machine Learning

Final exam: Thurs 12/16, 1:30pm. See Canvas for full info.

Lecture start: Tuesdays 9:35am

Who's talking? Prof. Tamara Broderick

Questions? Ask on Piazza: "lecture (week) 12" folder

Materials: slides, video will all be available on Canvas

Live Zoom feed: https://mit.zoom.us/j/94238622313

Last Time

I. Actions change the state of the world and gain reward: Markov decision processes (MDPs)

II. Value of a policy

Today's Plan

I. How to choose the best policy?

II. What if we don't know the transition model or reward function?

Recall

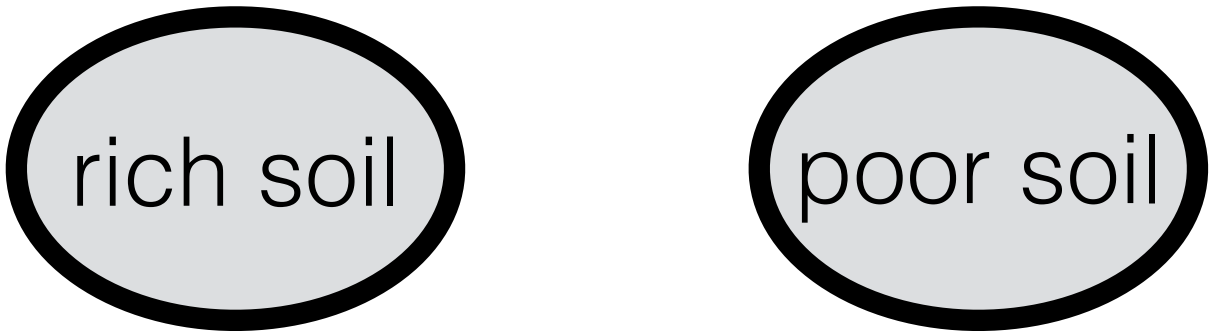
Recall

- Markov decision process

Recall

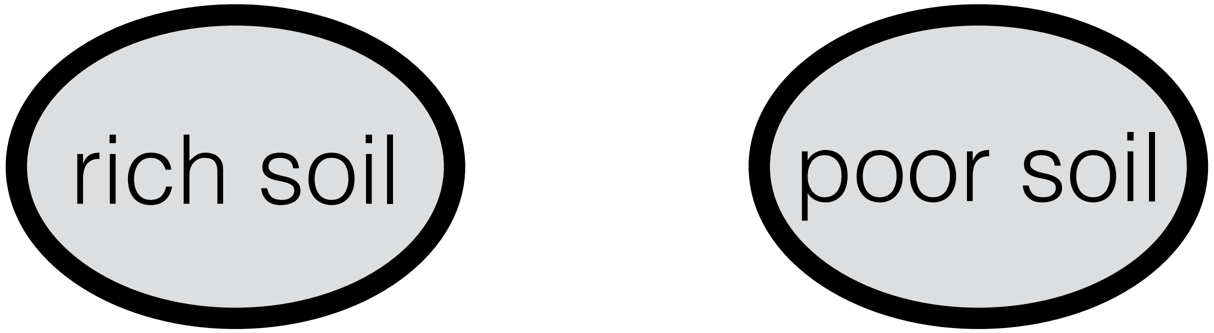
- Markov decision process: states $\mathcal{S}$

Recall



- Markov decision process: states $\mathcal{S}$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$

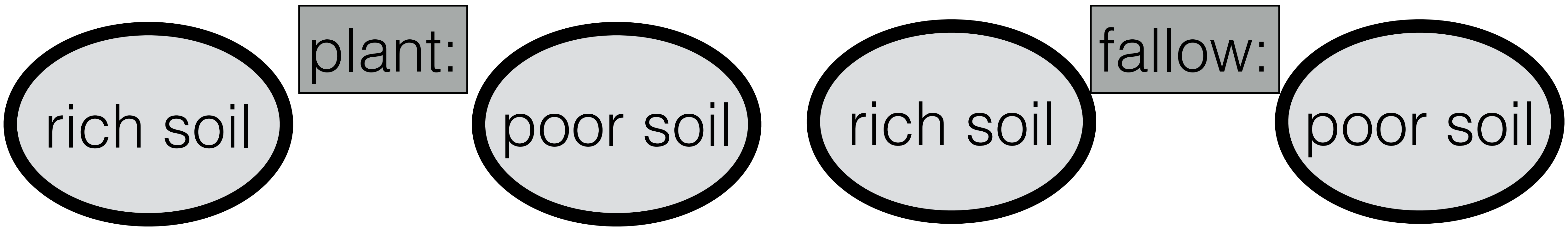
Recall

- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$

Recall

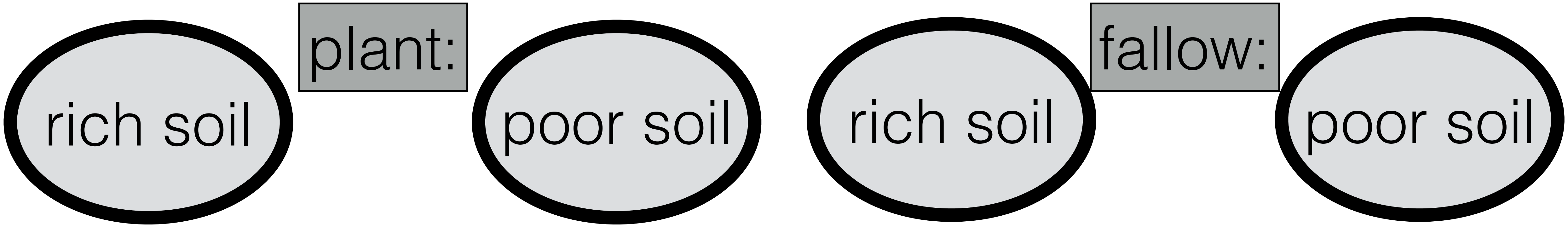
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$

Recall



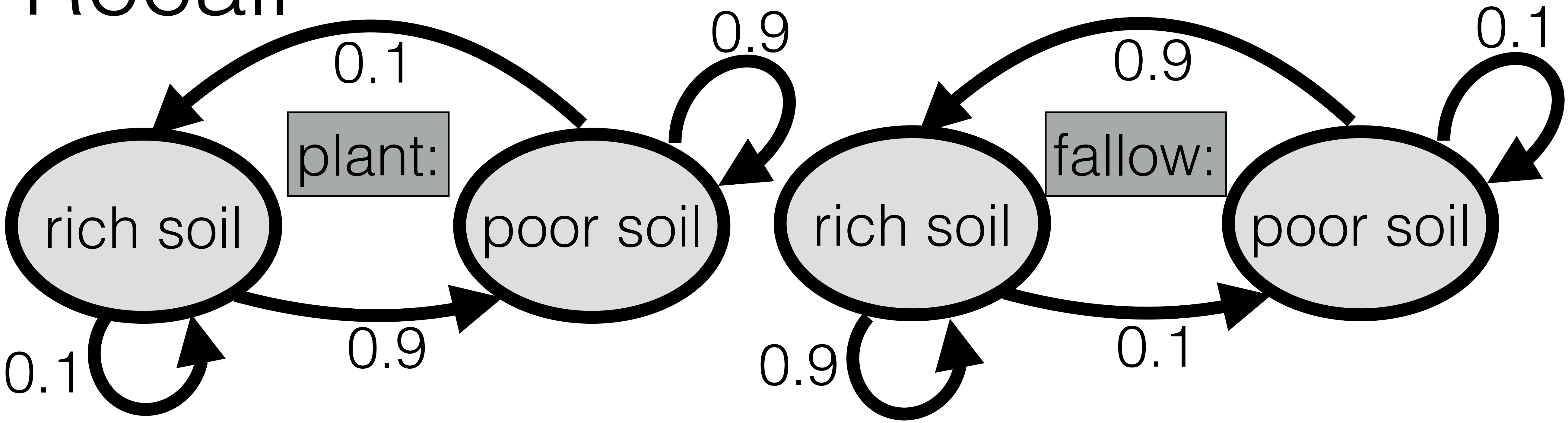
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$

Recall



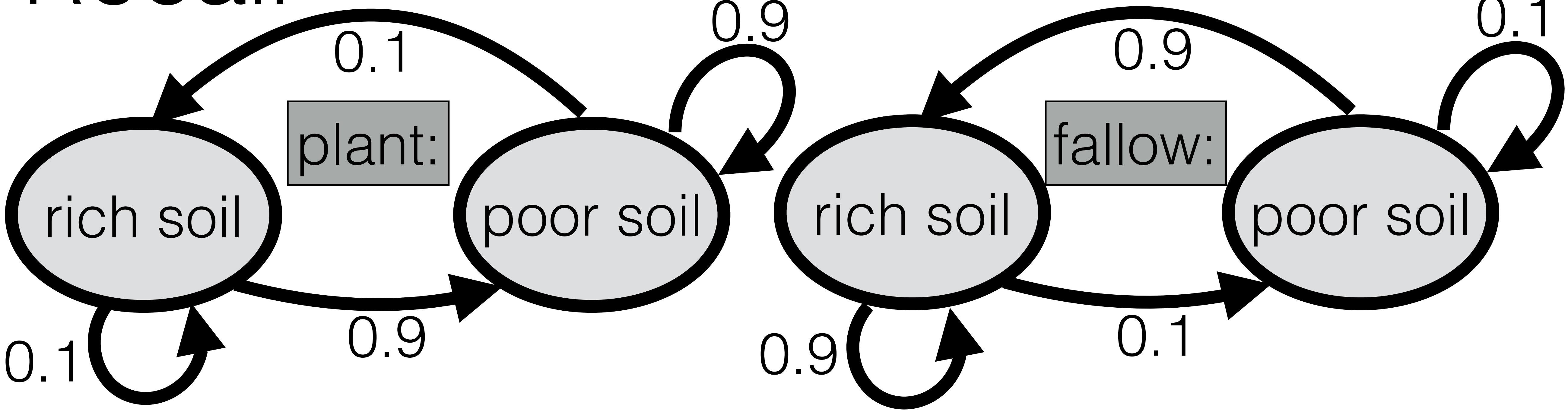
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



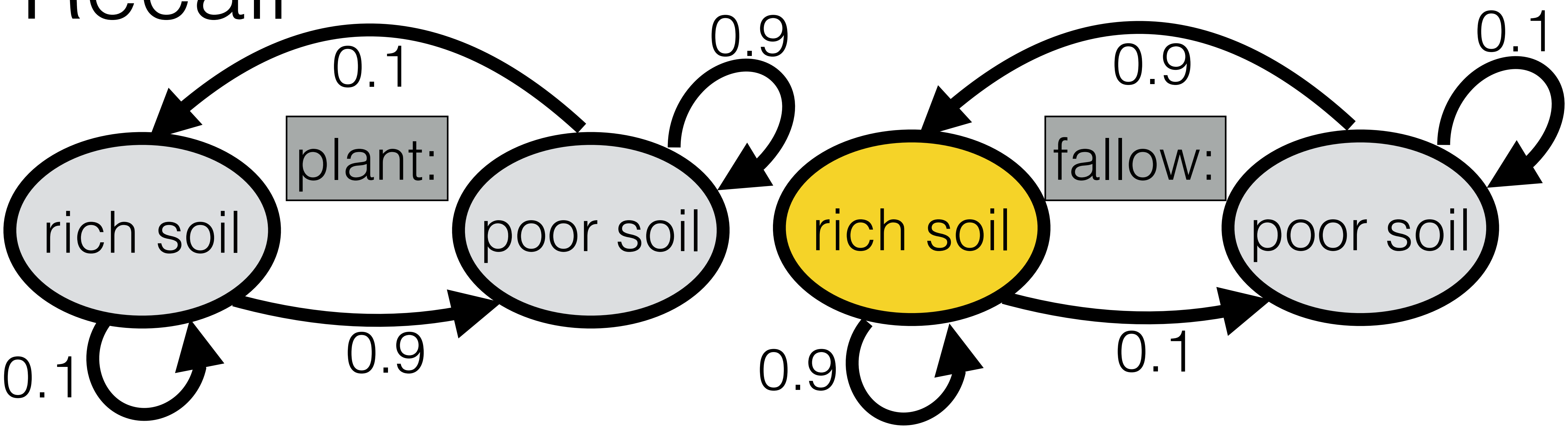
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



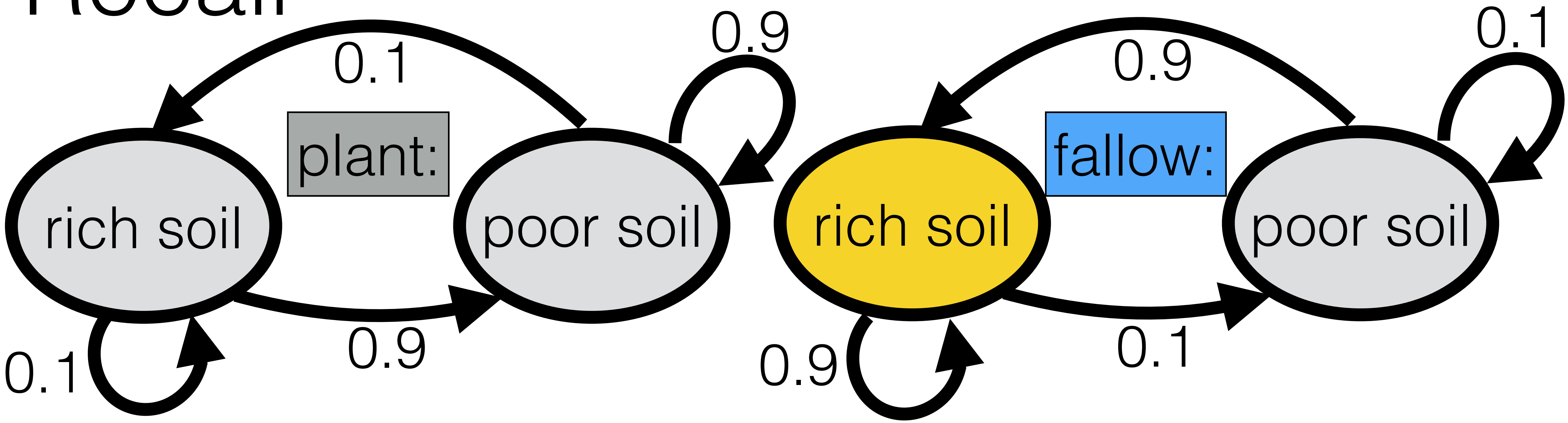
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



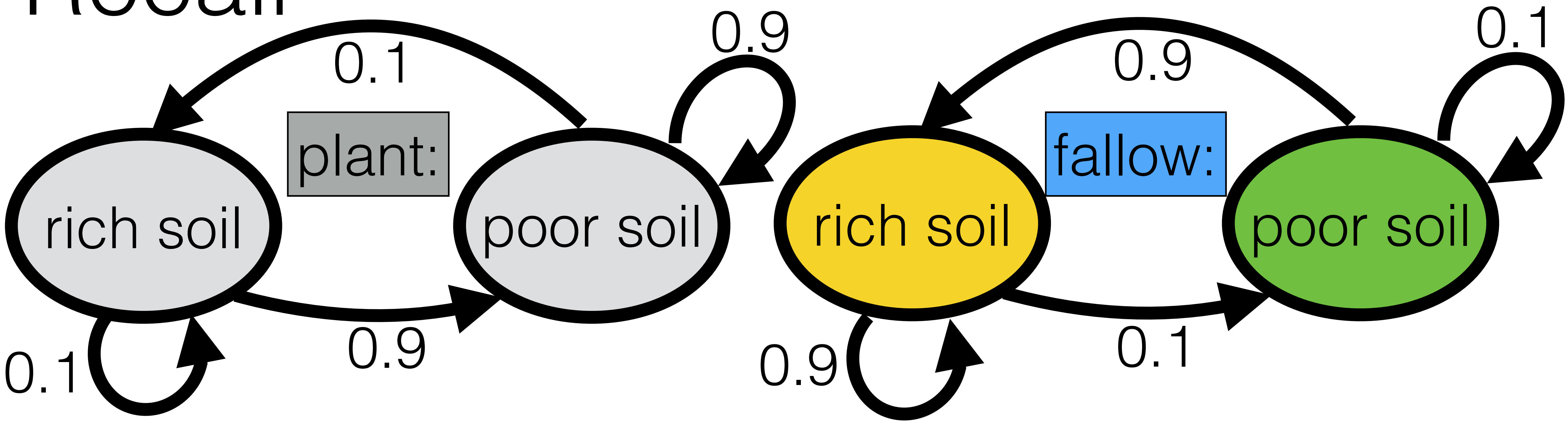
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



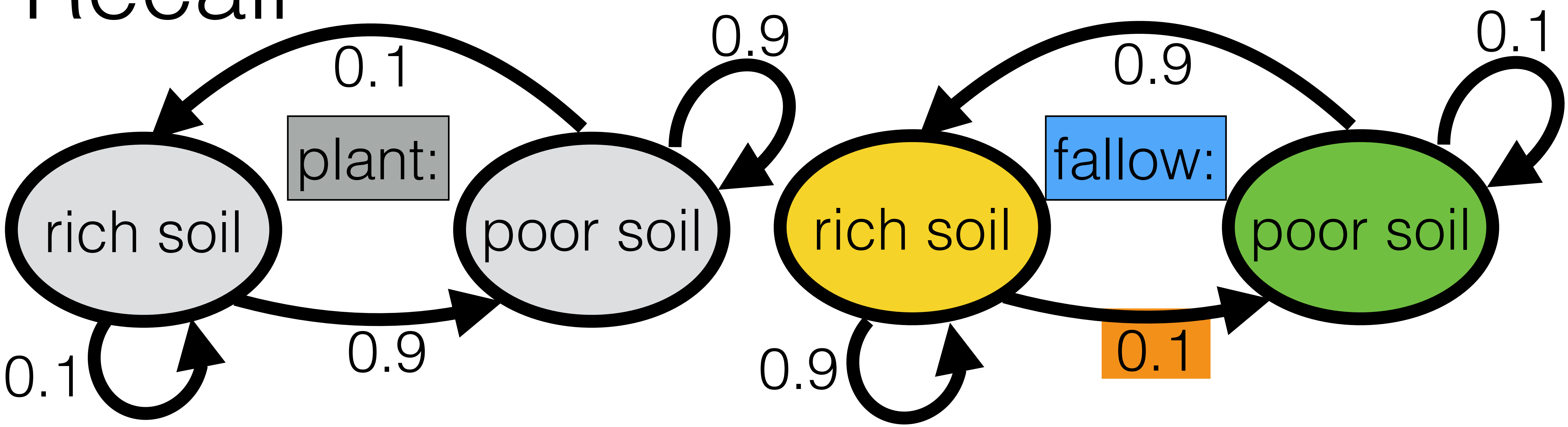
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



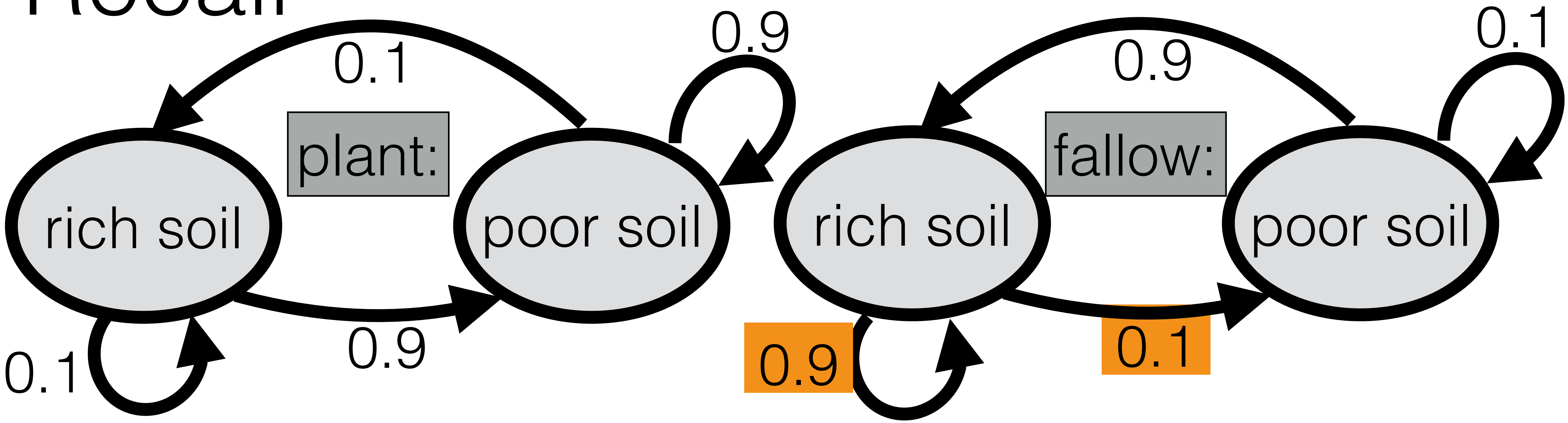
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



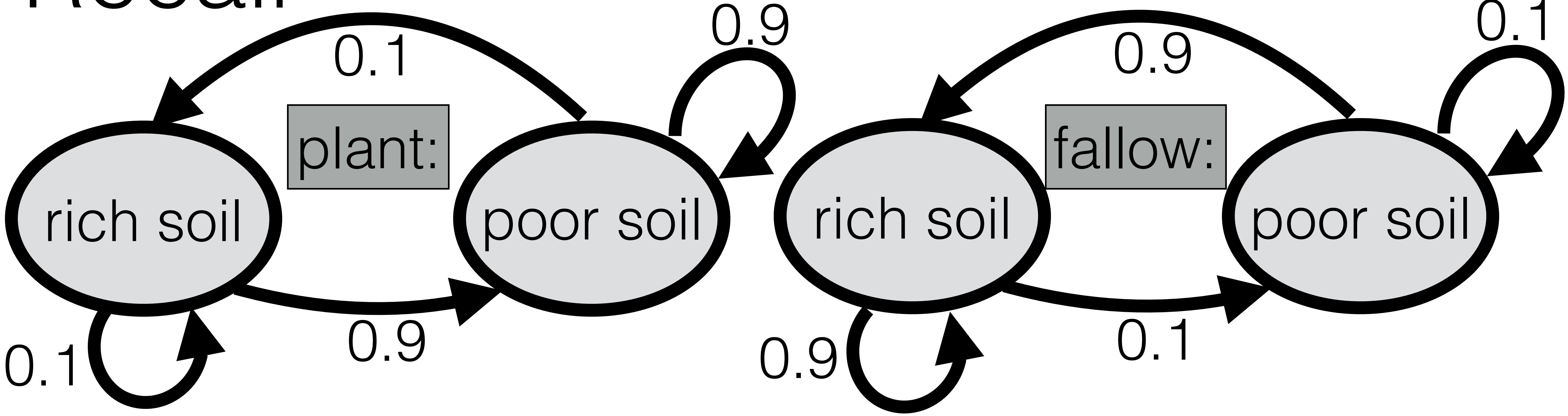
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



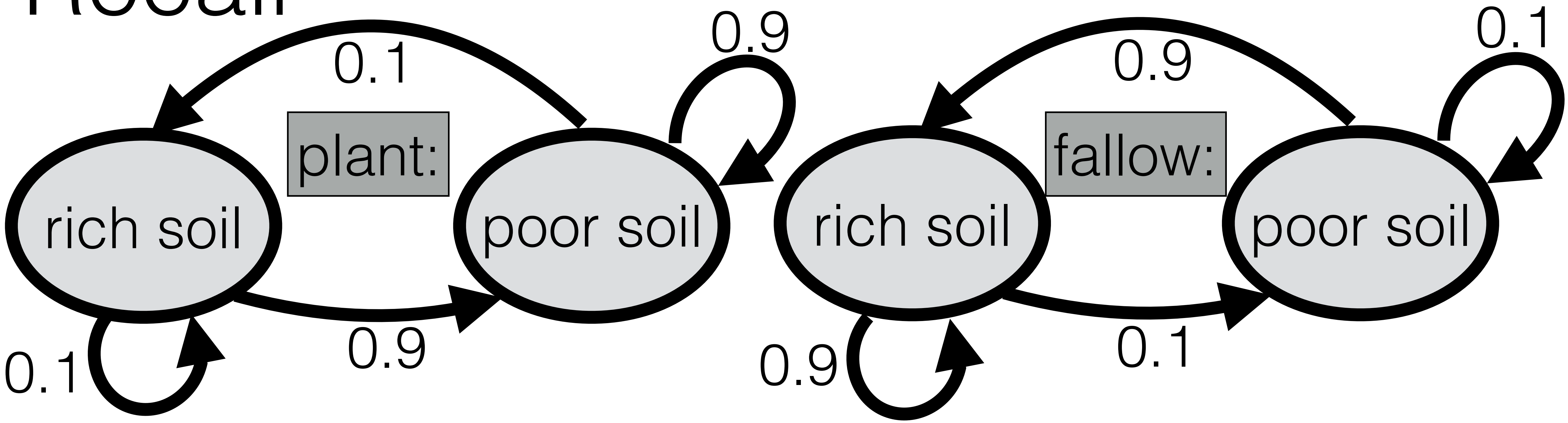
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



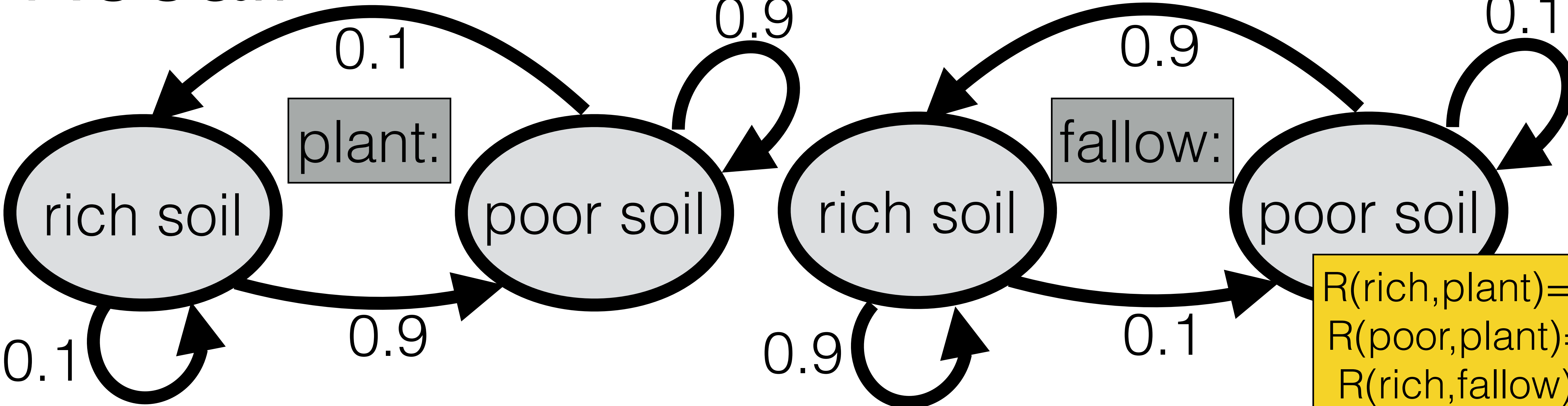
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

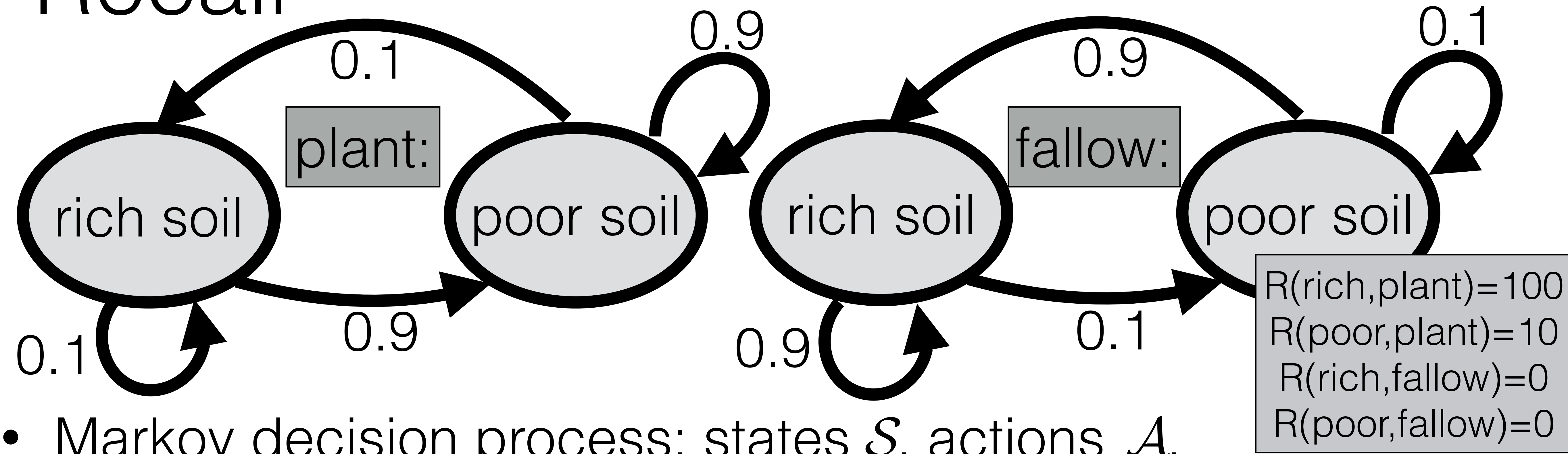
Recall



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

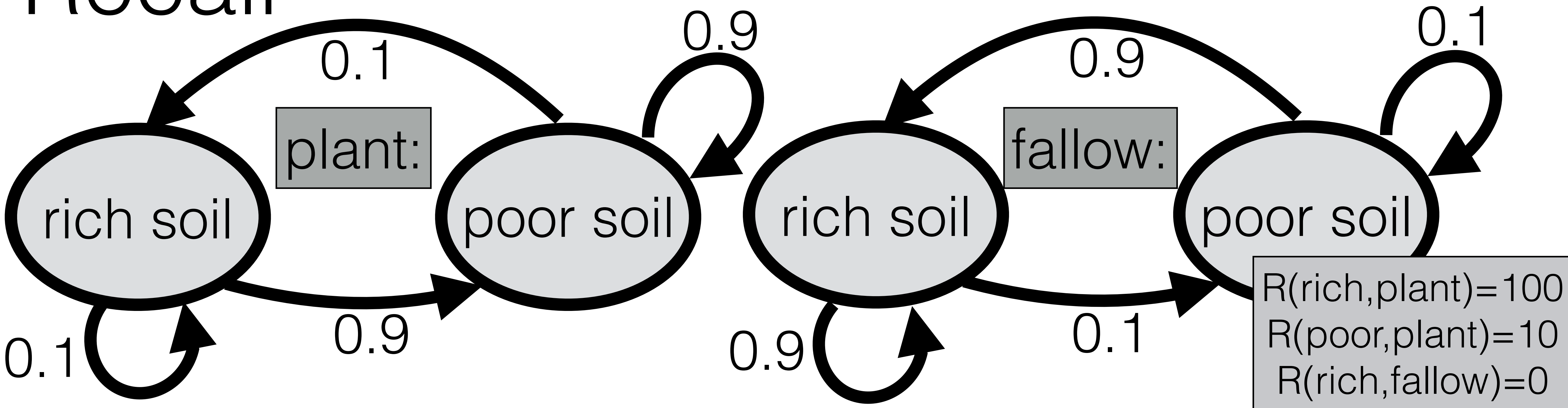
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

Recall



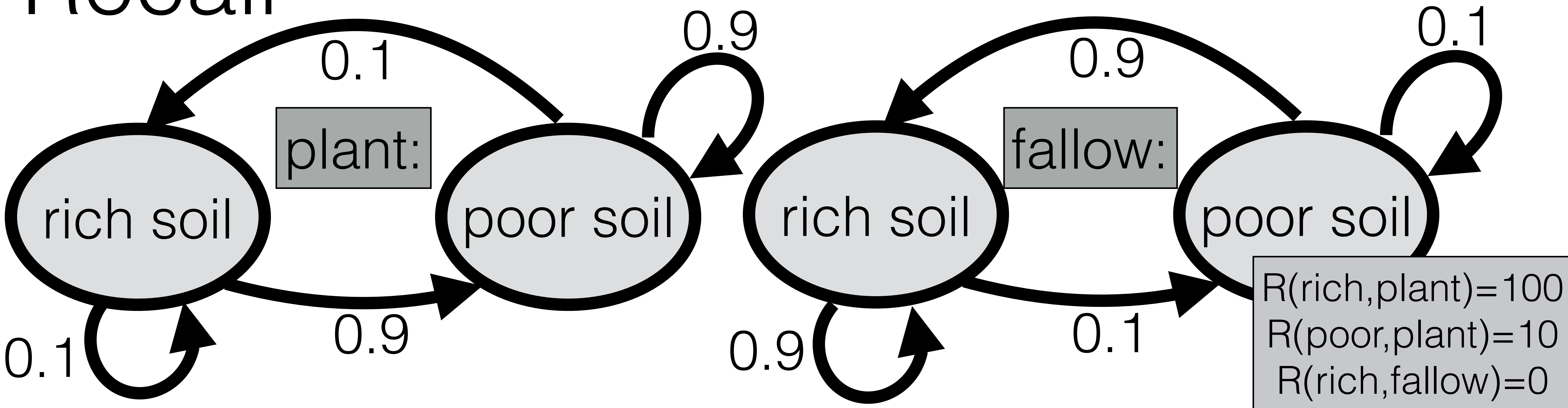
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ

Recall



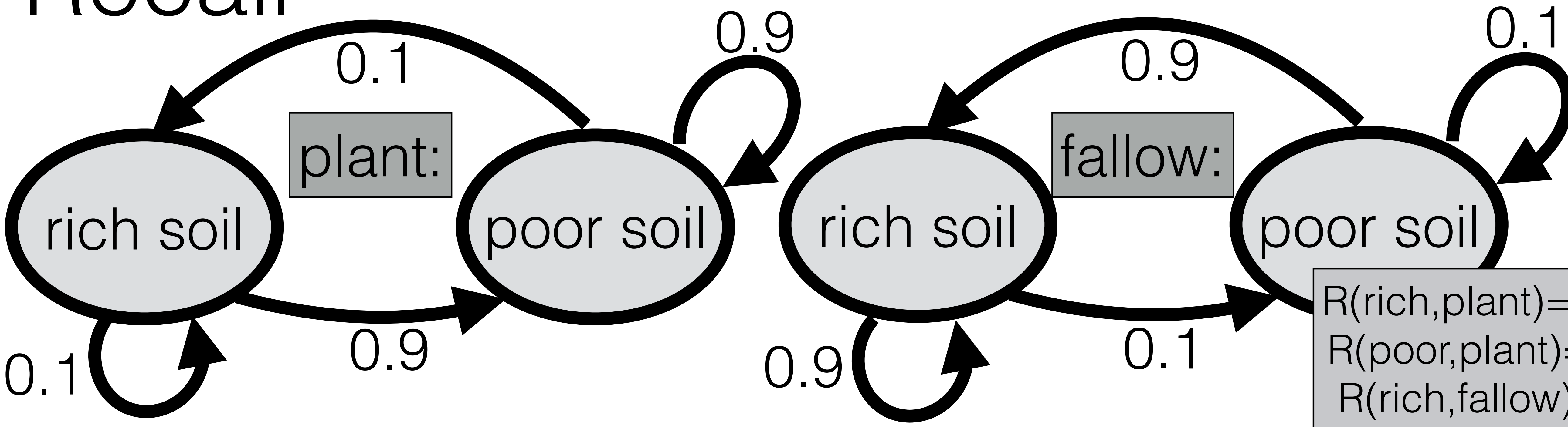
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon *h* (e.g. # planting seasons left)

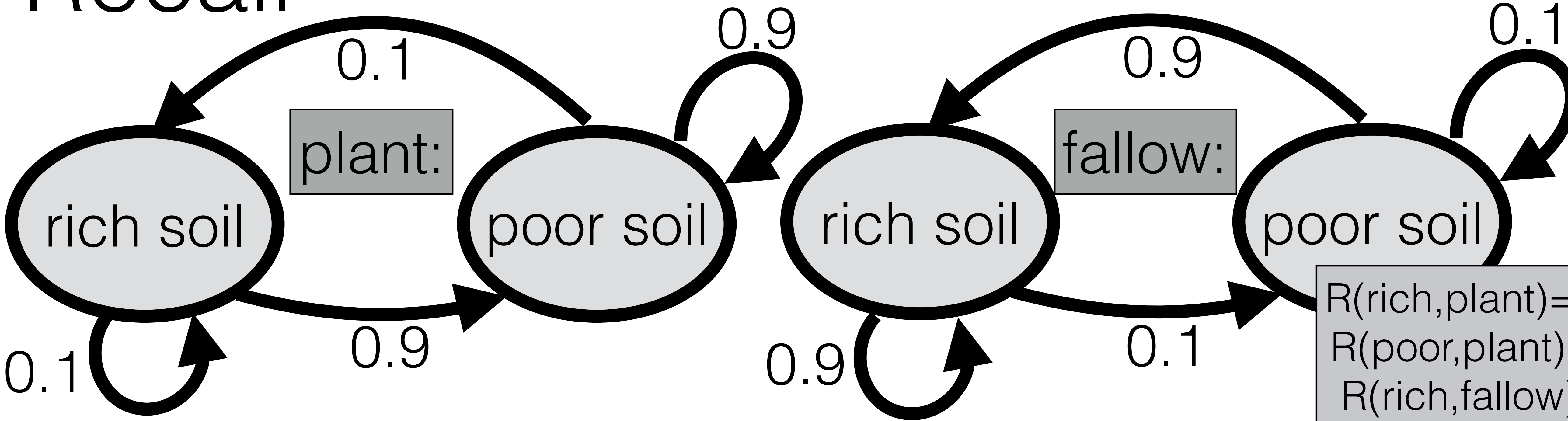
Recall



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon h (e.g. # planting seasons left)
- Value of a policy π if we start in state s

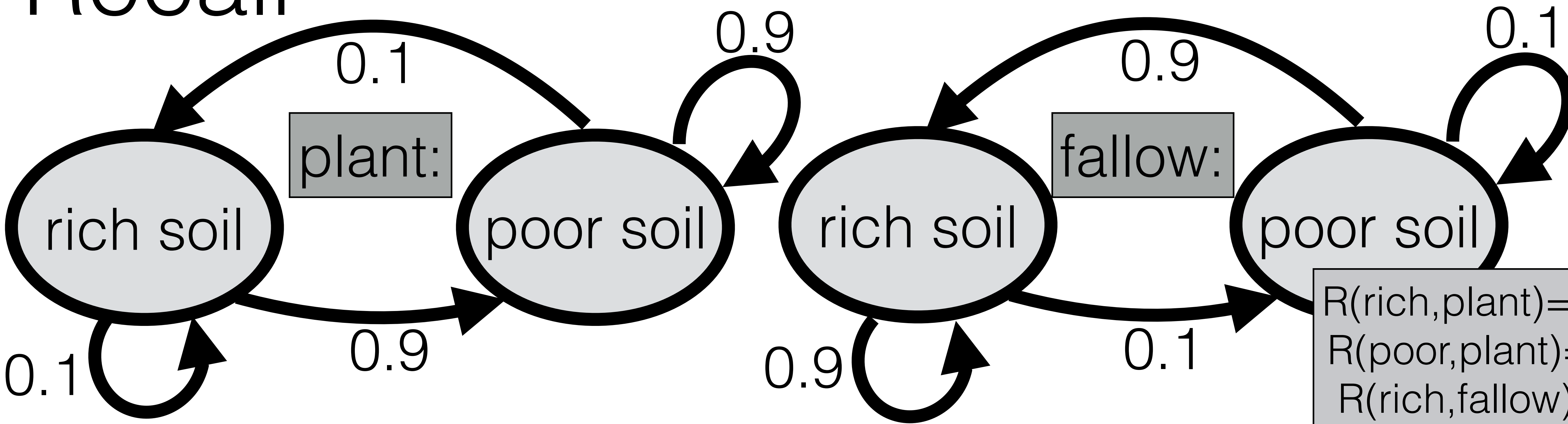
Recall



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon *h* (e.g. # planting seasons left)
- Value of a policy π if we start in state *s*
  - Finite horizon

Recall

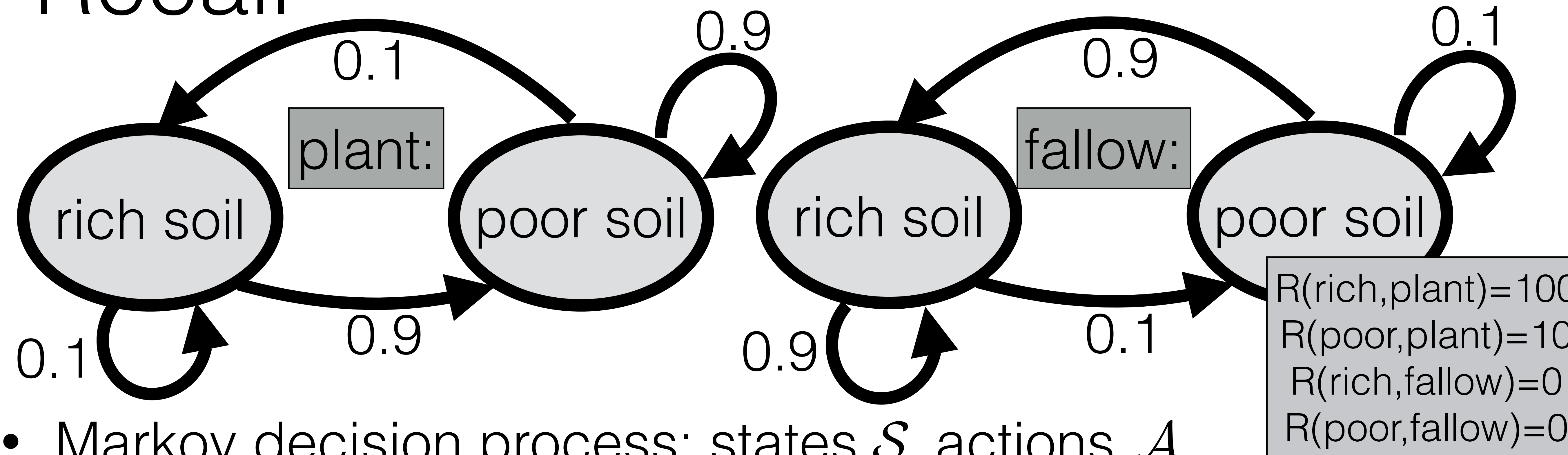


R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon *h* (e.g. # planting seasons left)
- Value of a policy π if we start in state *s*
  - Finite horizon

$$V_\pi^0(s) = 0$$

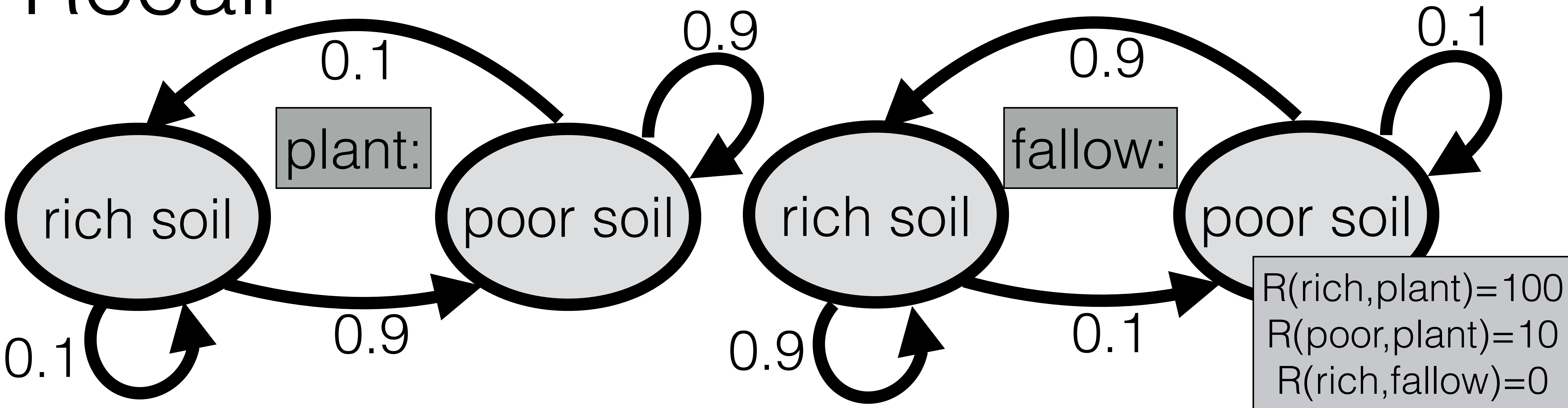
Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
- Value of a policy π if we start in state *s*
  - horizon *h* (e.g. # planting seasons left)
  - Finite horizon

$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

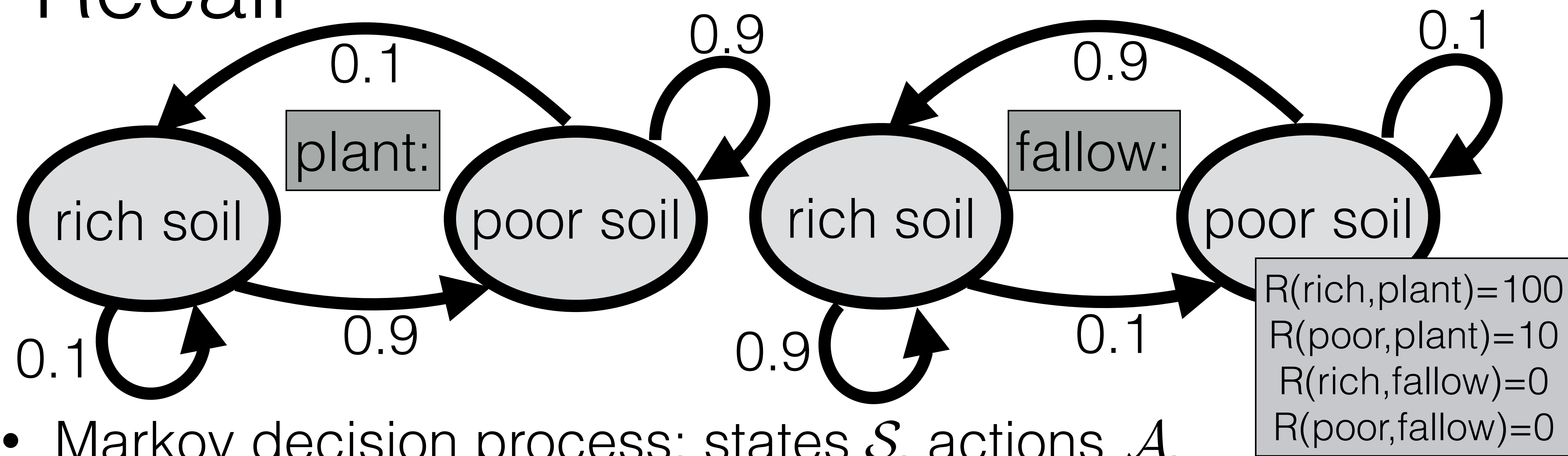
Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
- Value of a policy π if we start in state *s*
  - horizon *h* (e.g. # planting seasons left)
  - Finite horizon

$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

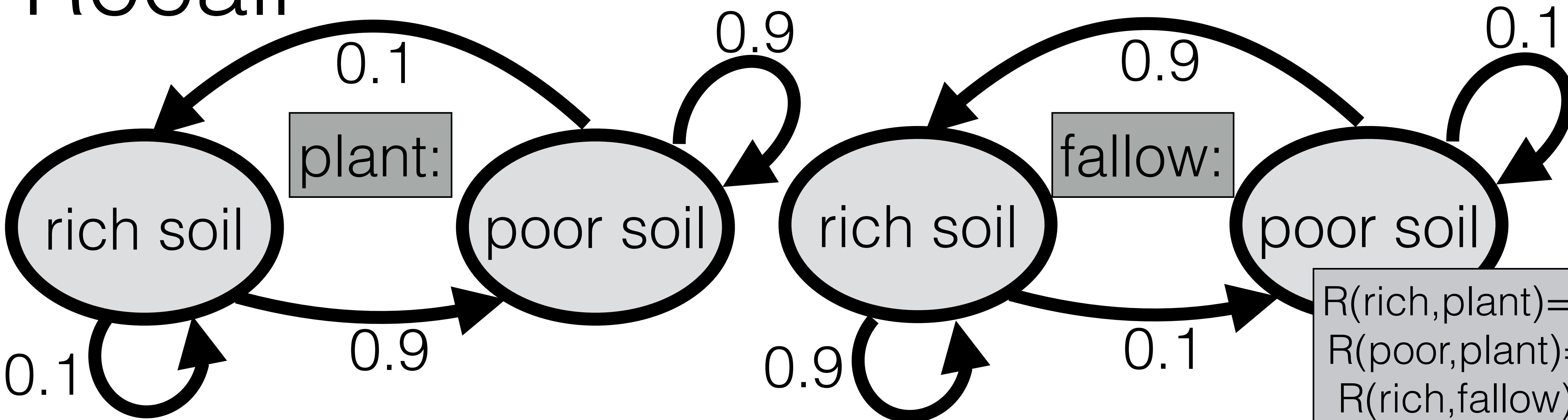
Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
- Value of a policy π if we start in state *s*
  - horizon *h* (e.g. # planting seasons left)
  - Finite horizon

$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

Recall

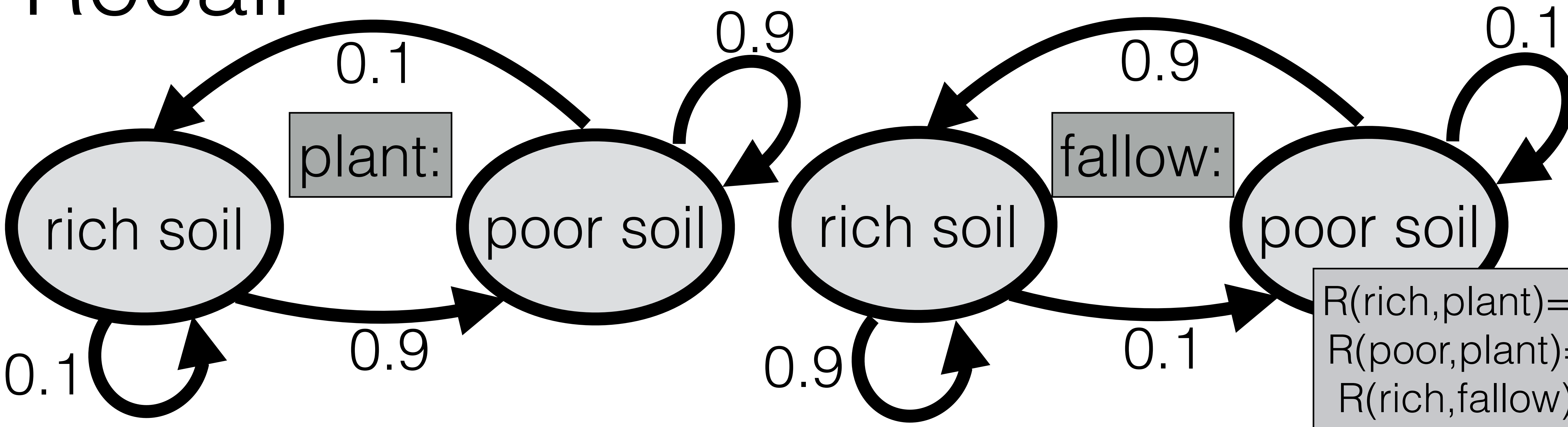


R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$: action to take in a state (nonstationary π_h)
- Value of a policy π if we start in state *s*
  - horizon *h* (e.g. # planting seasons left)
  - Finite horizon

$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

Recall

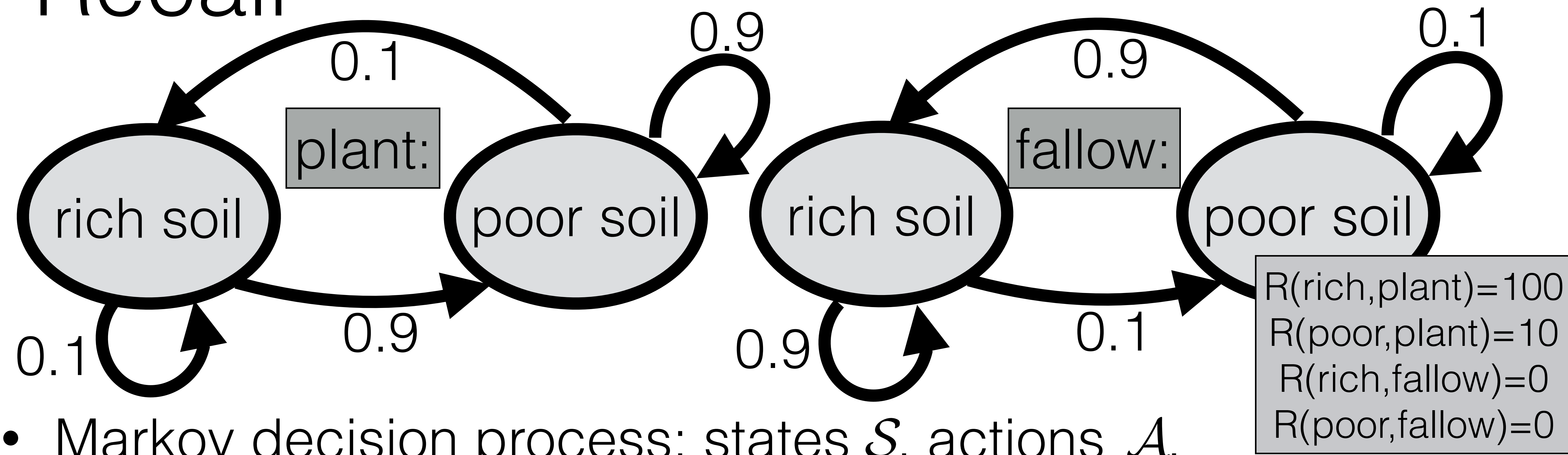


R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
- Value of a policy π if we start in state *s*
  - horizon *h* (e.g. # planting seasons left)
  - Finite horizon

$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

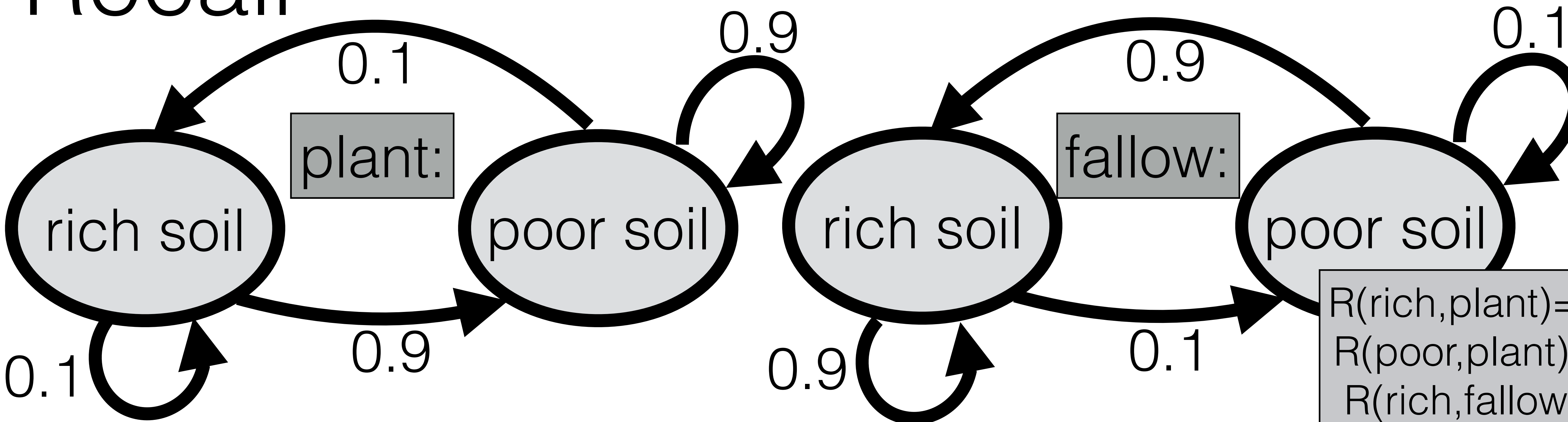
Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$: action to take in a state (nonstationary π_h)
- Value of a policy π if we start in state *s*
  - horizon *h* (e.g. # planting seasons left)
  - Finite horizon

$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

Recall

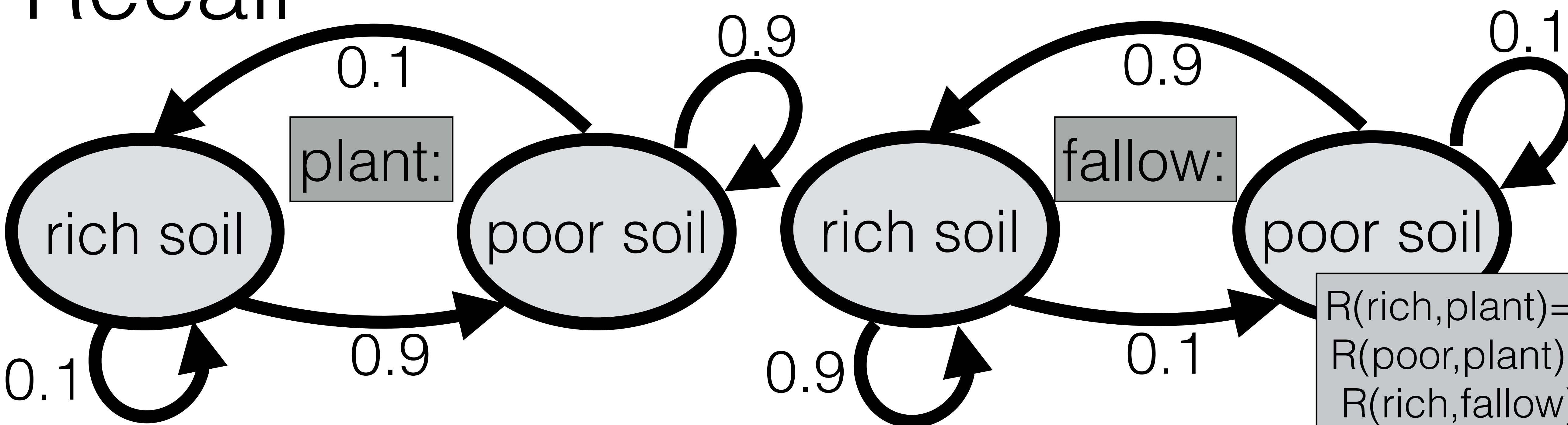


R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon h (e.g. # planting seasons left)
- Value of a policy π if we start in state s
  - Finite horizon

$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

Recall

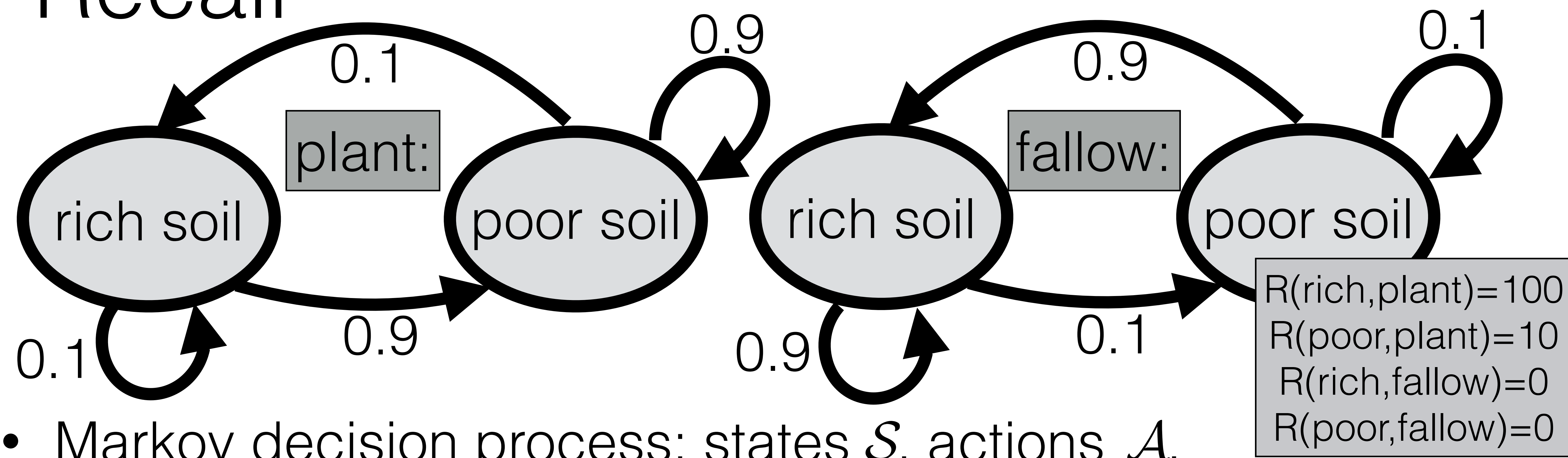


R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon *h* (e.g. # planting seasons left)
- Value of a policy π if we start in state *s*
  - Finite horizon (often assume discount factor γ equals 1)

$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

Recall

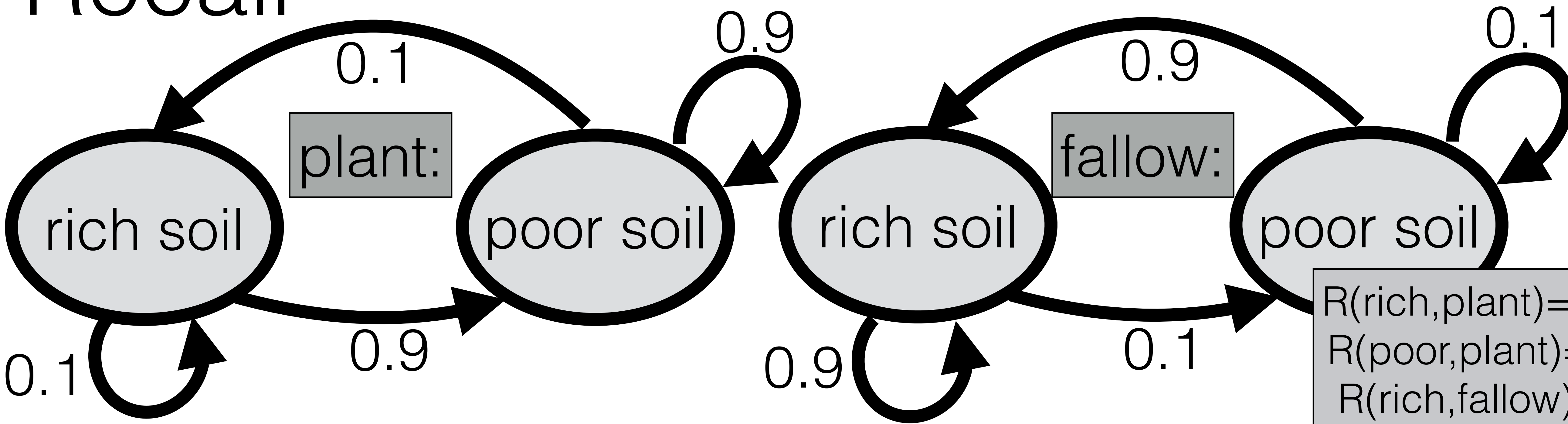


- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon *h* (e.g. # planting seasons left)
- Value of a policy π if we start in state *s*
  - Finite horizon (often assume discount factor γ equals 1)

$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

  - Infinite horizon

Recall



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

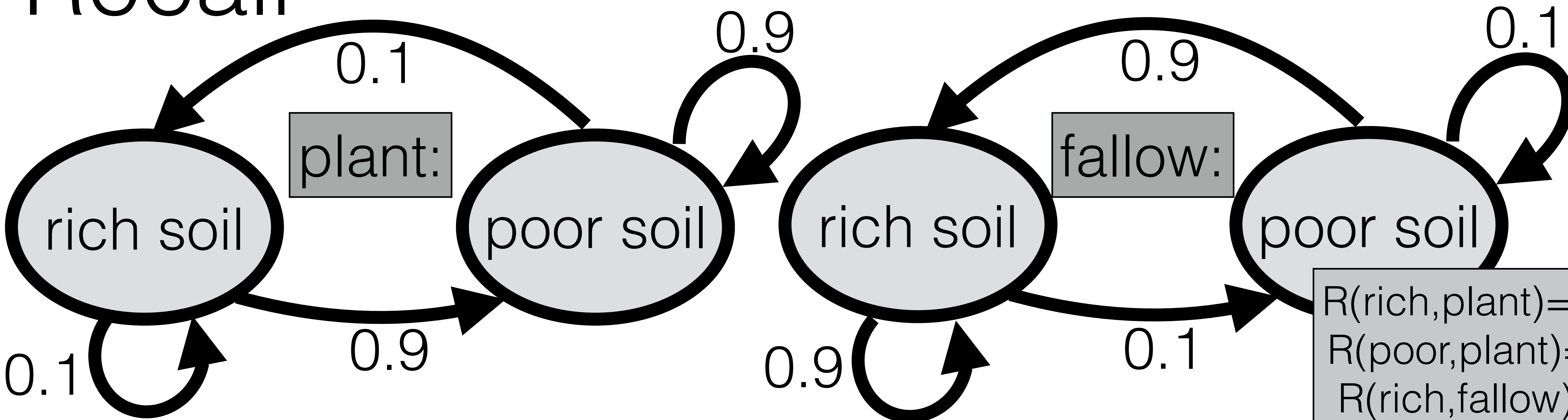
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon *h* (e.g. # planting seasons left)
- Value of a policy π if we start in state *s*
  - Finite horizon (often assume discount factor γ equals 1)

$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

  - Infinite horizon

$$V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$$

Recall



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

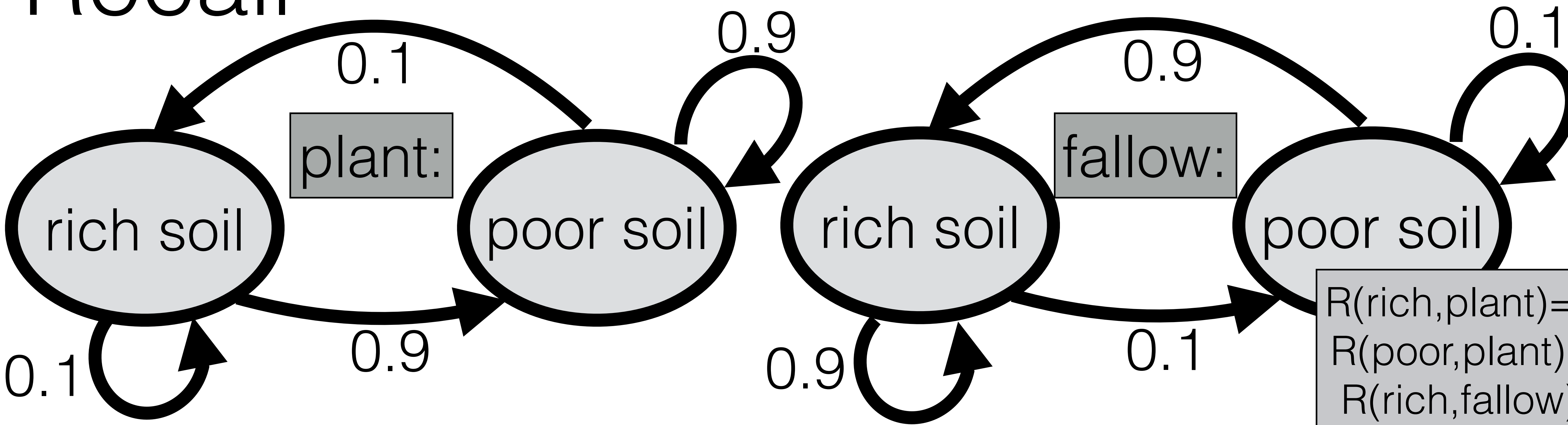
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon *h* (e.g. # planting seasons left)
- Value of a policy π if we start in state *s*
  - Finite horizon (often assume discount factor γ equals 1)

$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

  - Infinite horizon

$$V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$$

Recall



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

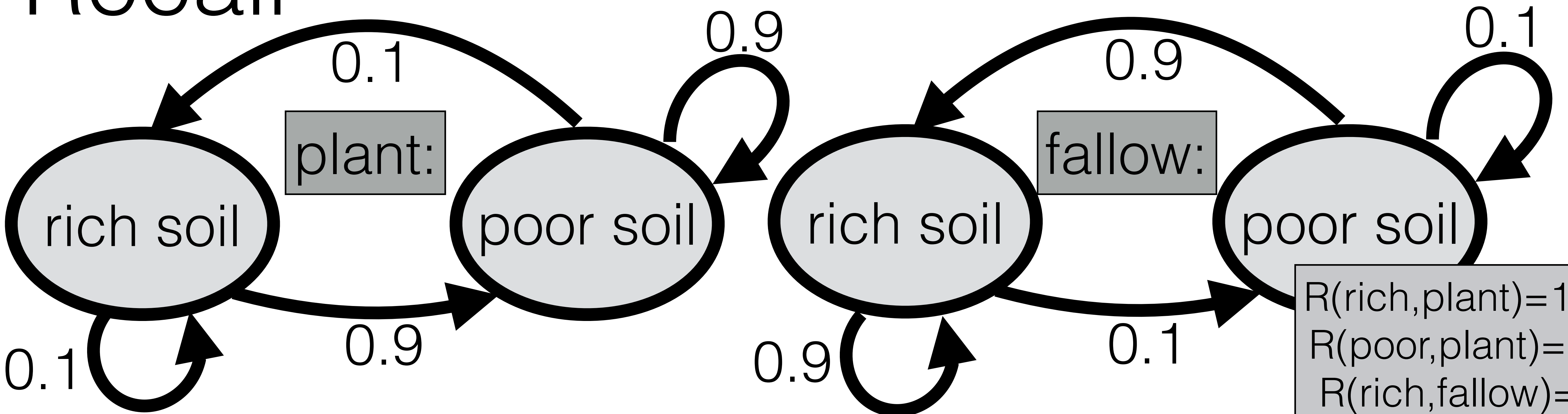
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
    - horizon *h* (e.g. # planting seasons left)
- Value of a policy π if we start in state *s*
    - Finite horizon (often assume discount factor γ equals 1)

      $$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

    - Infinite horizon

      $$V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$$

Recall



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

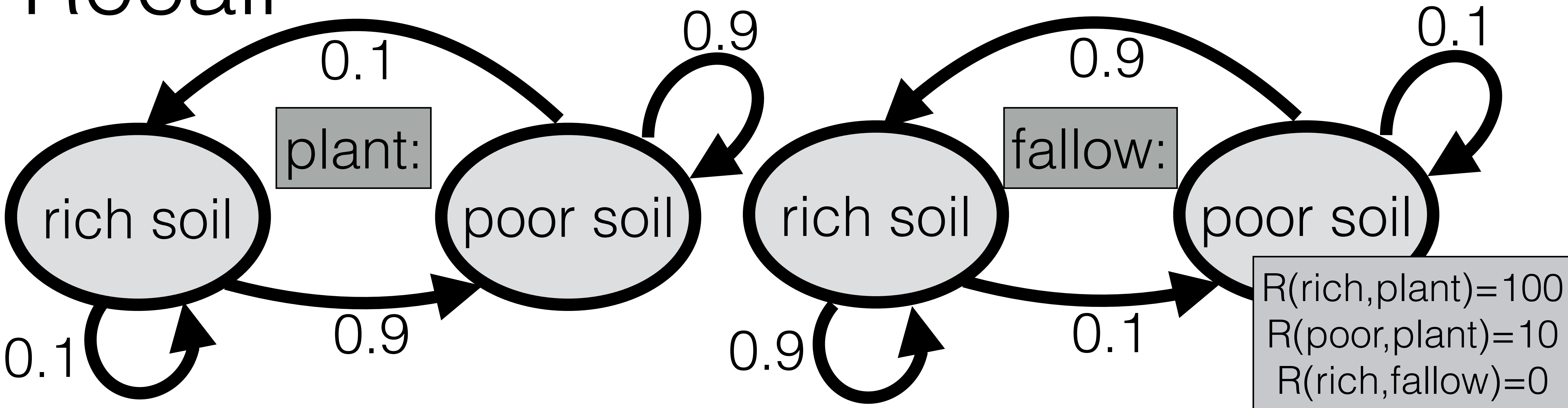
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon *h* (e.g. # planting seasons left)
- Value of a policy π if we start in state *s*
  - Finite horizon (often assume discount factor γ equals 1)

$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

  - Infinite horizon

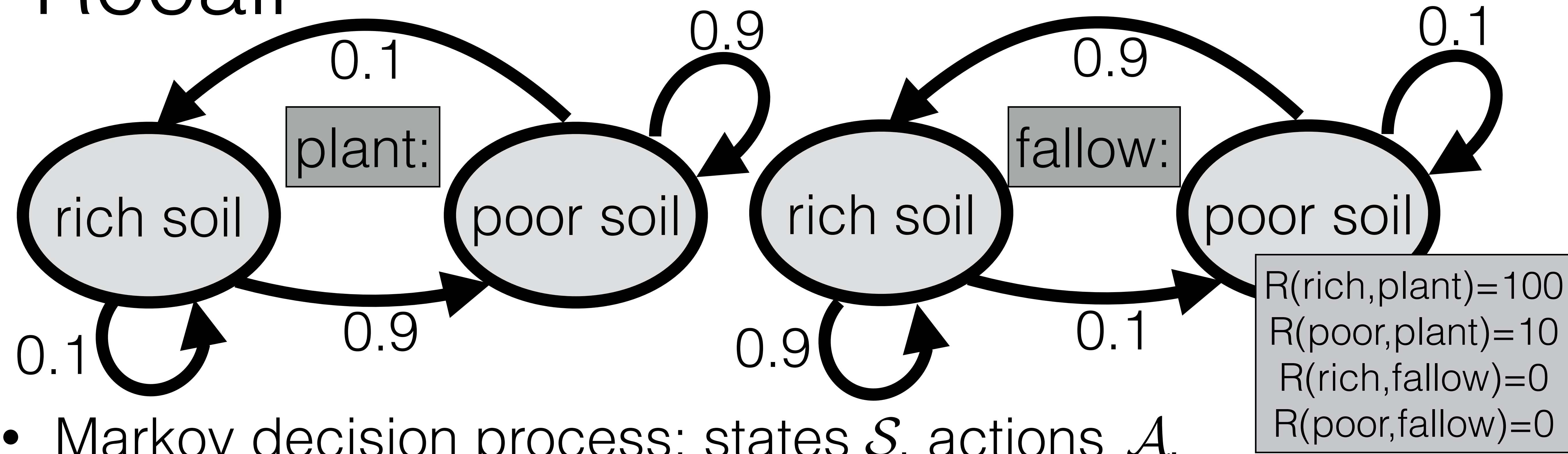
$$V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon *h* (e.g. # planting seasons left)
- Value of a policy π if we start in state *s*
  - Finite horizon (often assume discount factor γ equals 1)
    $V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$
  - Infinite horizon
    $V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$

Recall

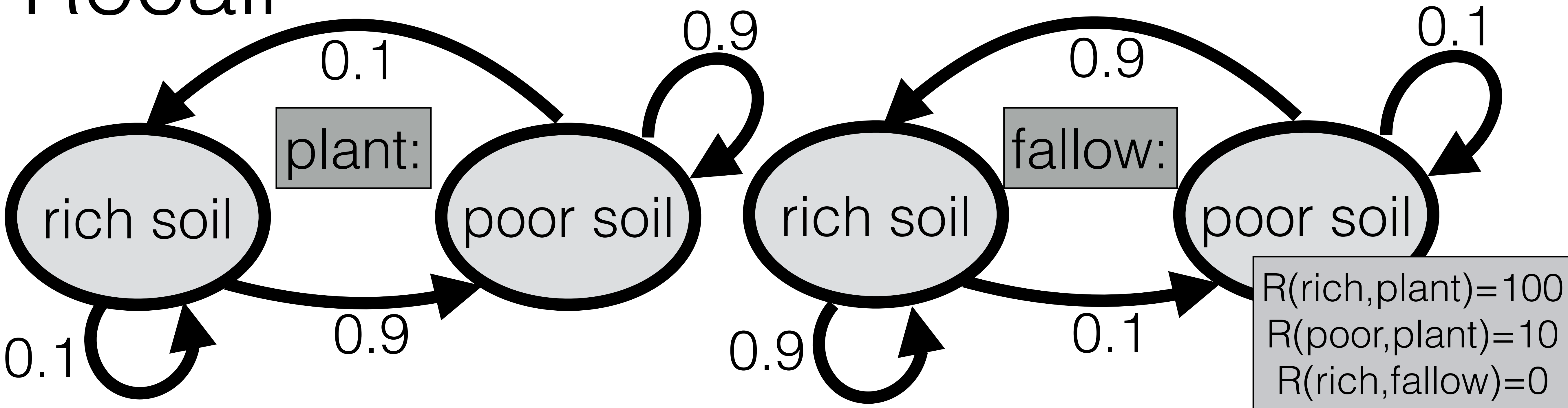


- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon *h* (e.g. # planting seasons left)
- Value of a policy π if we start in state *s*
  - Finite horizon (often assume discount factor γ equals 1)

    $V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$
  - Infinite horizon

    $V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$

Recall

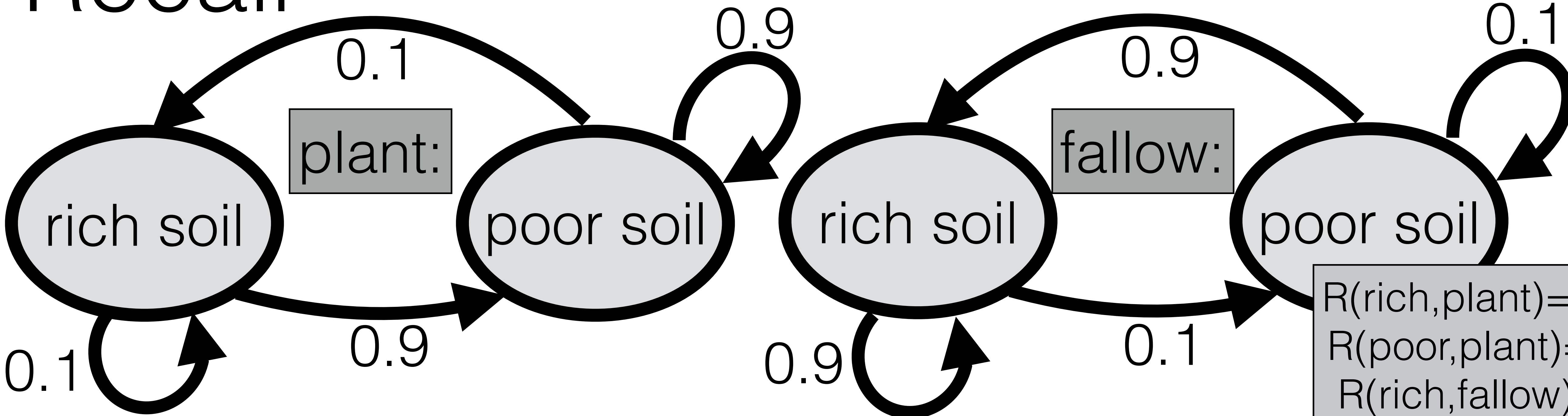


- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon *h* (e.g. # planting seasons left)
- Value of a policy π if we start in state *s*
  - Finite horizon (often assume discount factor γ equals 1)

    $V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$
  - Infinite horizon

    $V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$

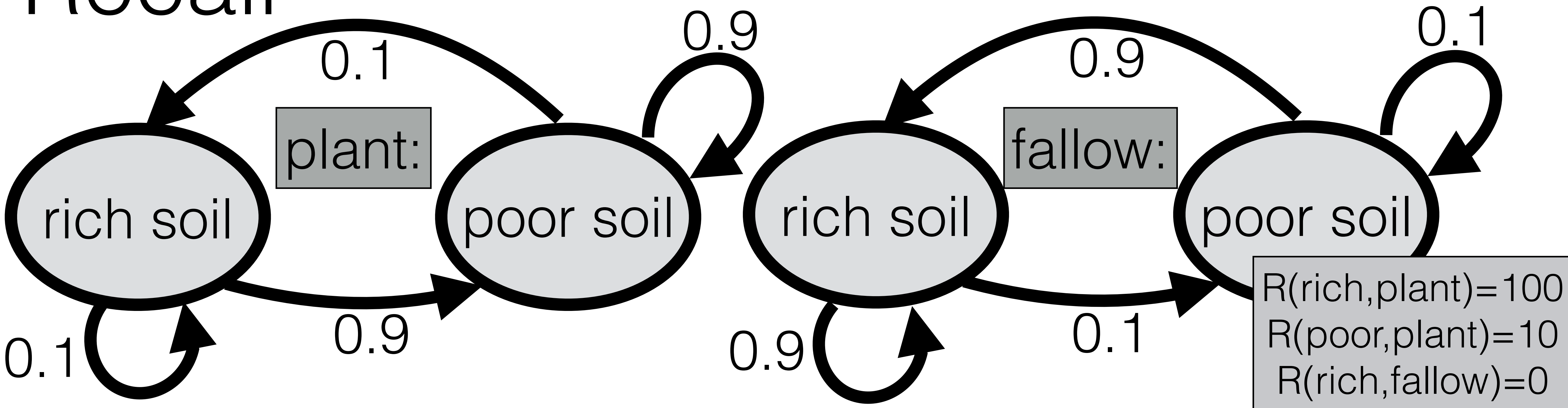
Recall



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon h (e.g. # planting seasons left)
- Value of a policy π if we start in state s
  - Finite horizon (often assume discount factor γ equals 1)
    $V_\pi^0(s) = 0$; $V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$
  - Infinite horizon
    $V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$

Recall

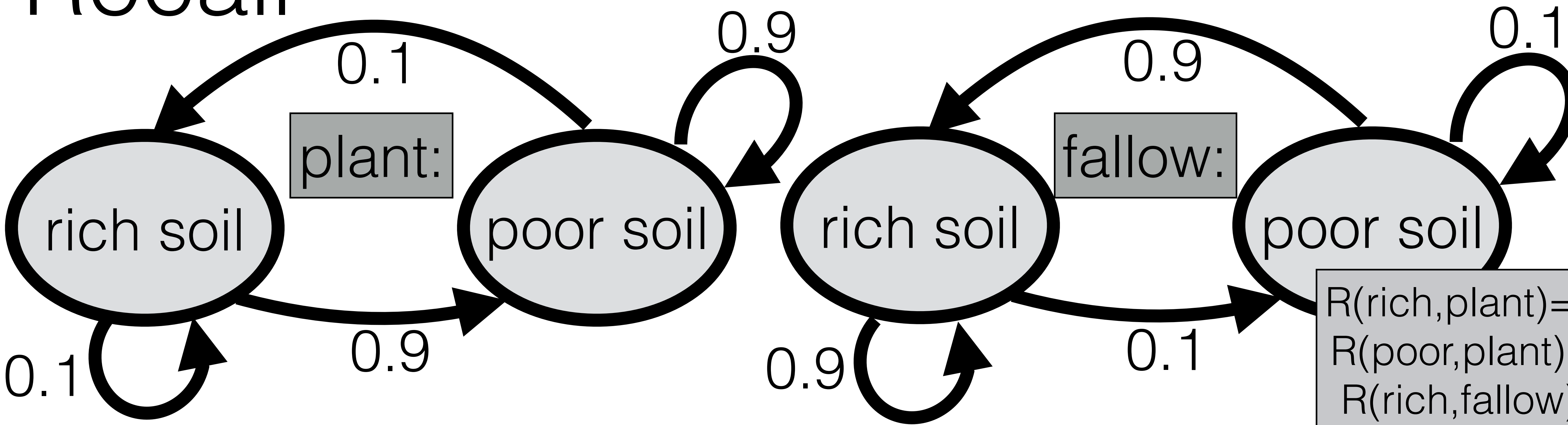


- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon *h* (e.g. # planting seasons left)
- Value of a policy π if we start in state *s*
  - Finite horizon (often assume discount factor γ equals 1)

    $V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$
  - Infinite horizon

    $V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$

Recall



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

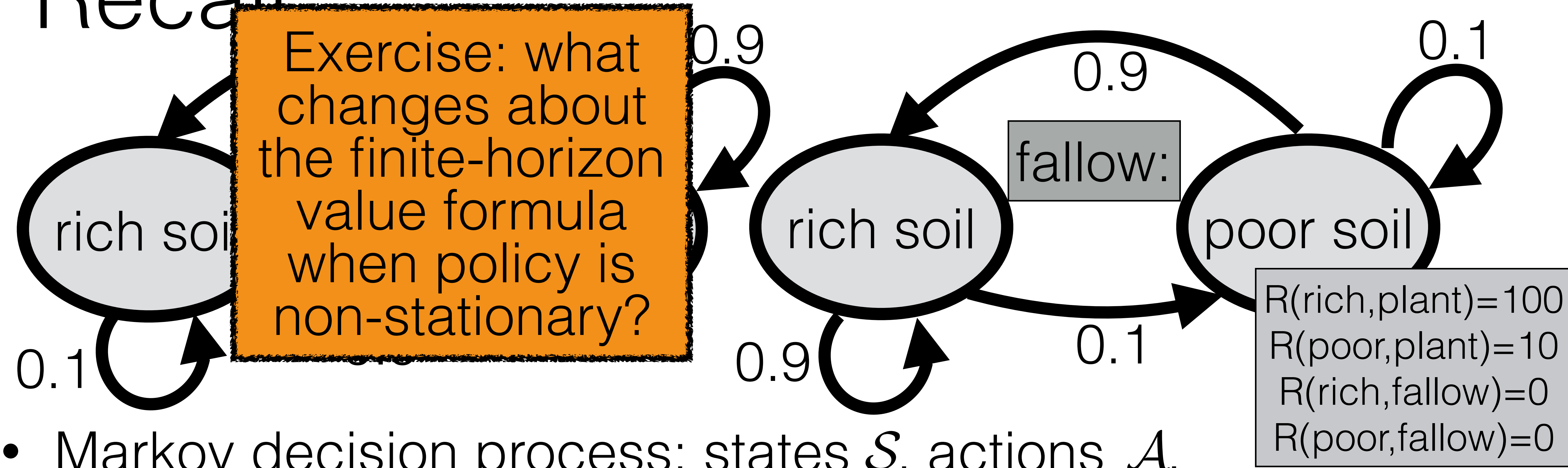
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon *h* (e.g. # planting seasons left)
- Value of a policy π if we start in state *s*
  - Finite horizon (often assume discount factor γ equals 1)

    $V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$
  - Infinite horizon (typically *need* to assume $0 < \gamma < 1$)

    $V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$

Recall

Exercise: what changes about the finite-horizon value formula when policy is non-stationary?

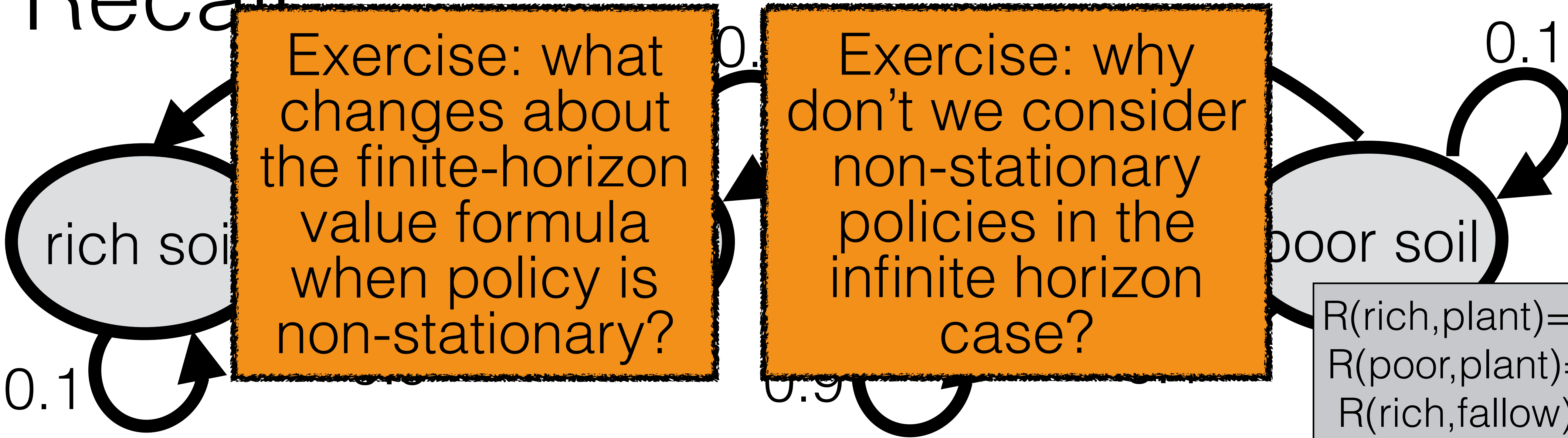


- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon h (e.g. # planting seasons left)
- Value of a policy π if we start in state s
  - Finite horizon (often assume discount factor γ equals 1)

    $V_\pi^0(s) = 0; V_\pi^h(s) = R(s,\pi(s)) + \gamma \sum_{s'} T(s,\pi(s),s') V_\pi^{h-1}(s')$
  - Infinite horizon (typically *need* to assume $0 < \gamma < 1$)

    $V_\pi(s) = R(s,\pi(s)) + \gamma \sum_{s'} T(s,\pi(s),s') V_\pi(s')$

Recall



Exercise: what changes about the finite-horizon value formula when policy is non-stationary?

Exercise: why don't we consider non-stationary policies in the infinite horizon case?

R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

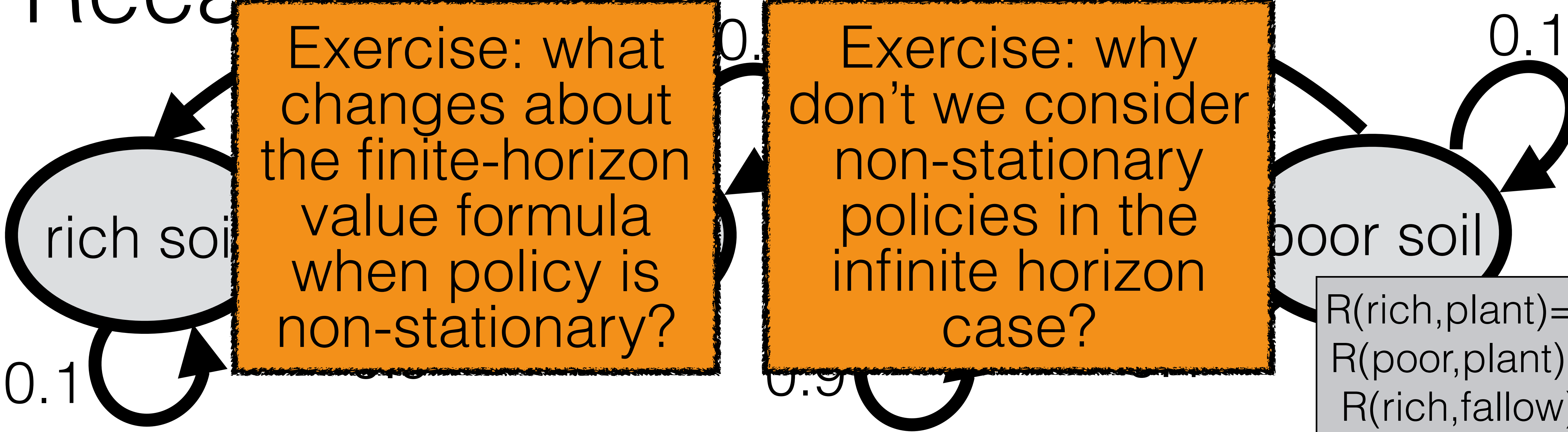
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
  - horizon *h* (e.g. # planting seasons left)
- Value of a policy π if we start in state *s*
  - Finite horizon (often assume discount factor γ equals 1)

$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

  - Infinite horizon (typically *need* to assume $0 < \gamma < 1$)

$$V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$$

Recall



Exercise: what changes about the finite-horizon value formula when policy is non-stationary?

Exercise: why don't we consider non-stationary policies in the infinite horizon case?

R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state (nonstationary π_h)
- Value of a policy π if we start in state *s*
  - horizon *h* (e.g. # planting seasons left)
  - Finite horizon (often assume discount factor γ equals 1)

$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

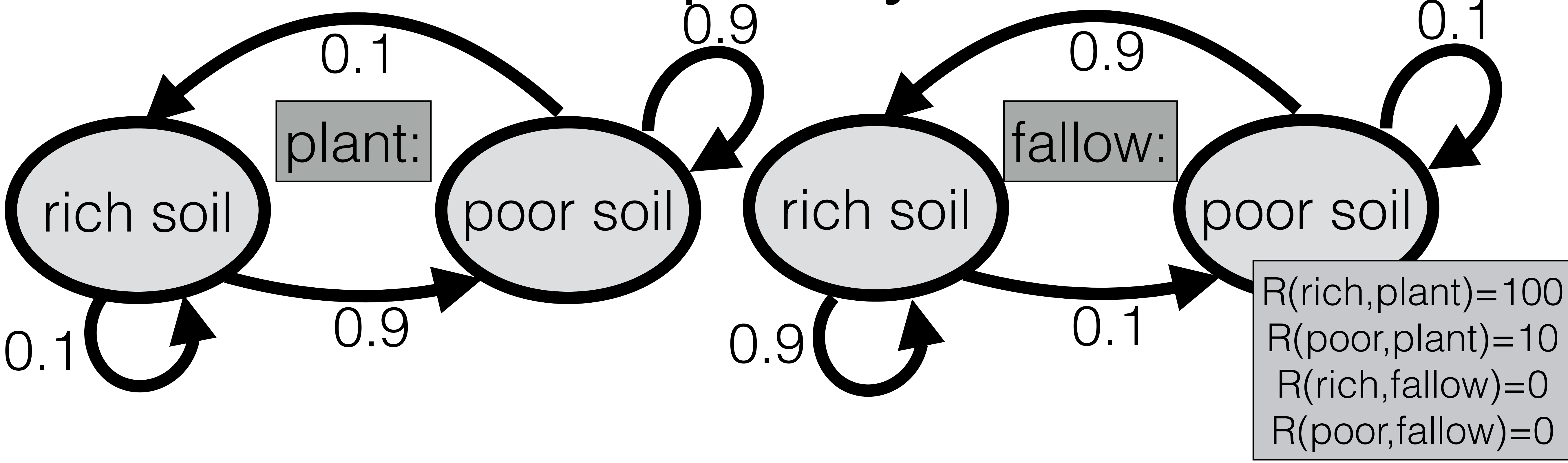
  - Infinite horizon (typically *need* to assume $0 < \gamma < 1$)

$$V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$$

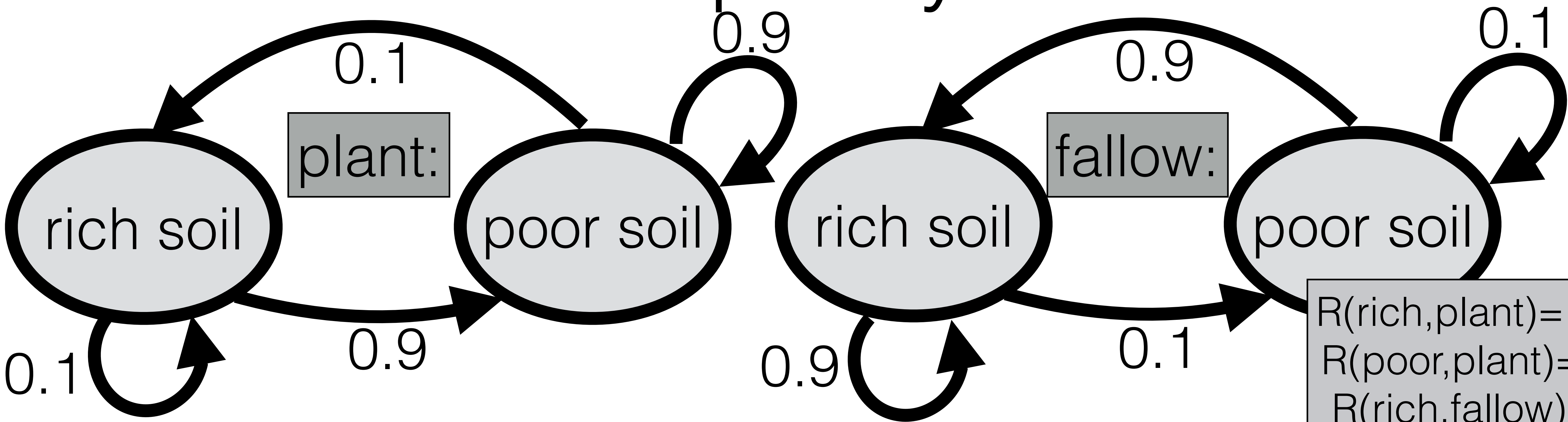
- Next question: What's the best policy?

What's the best policy? Finite horizon

What’s the best policy? Finite horizon



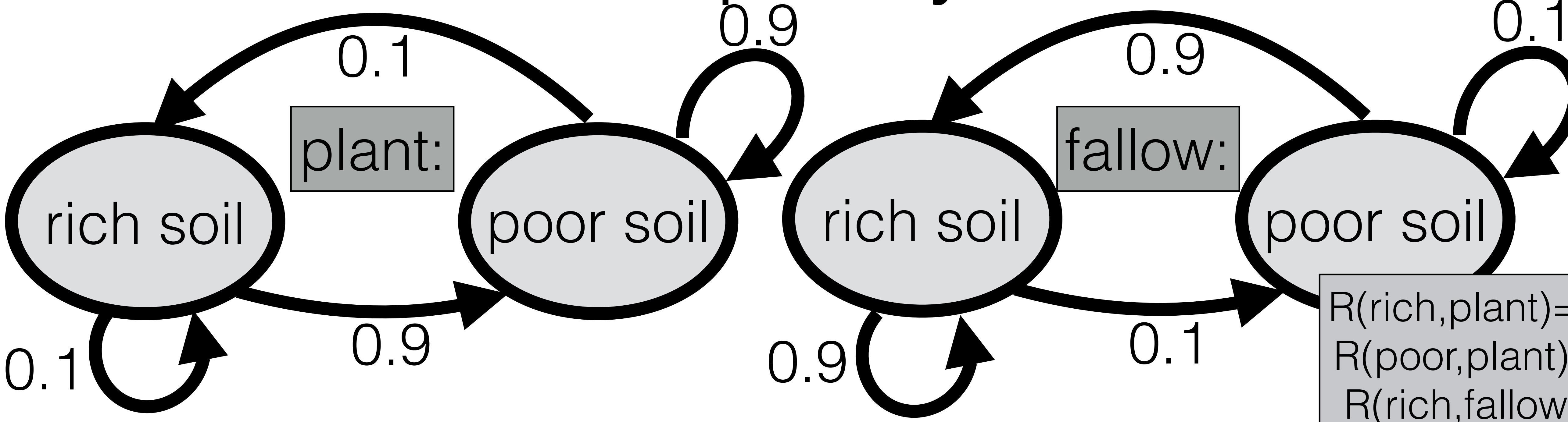
What’s the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)

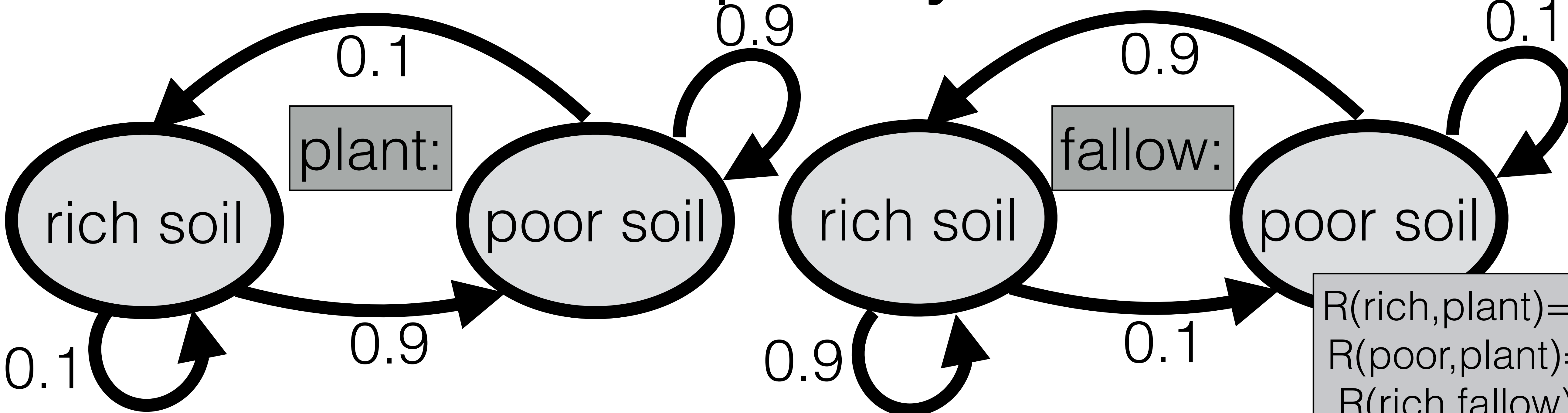
What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left

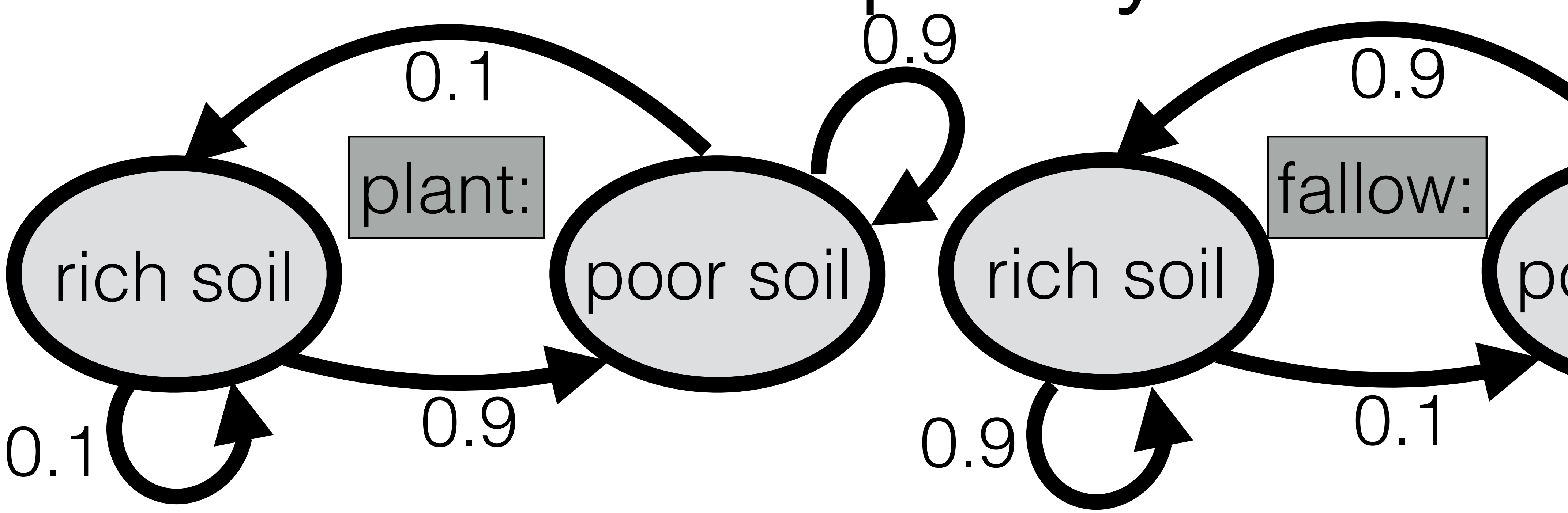
What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- h: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s, a)$

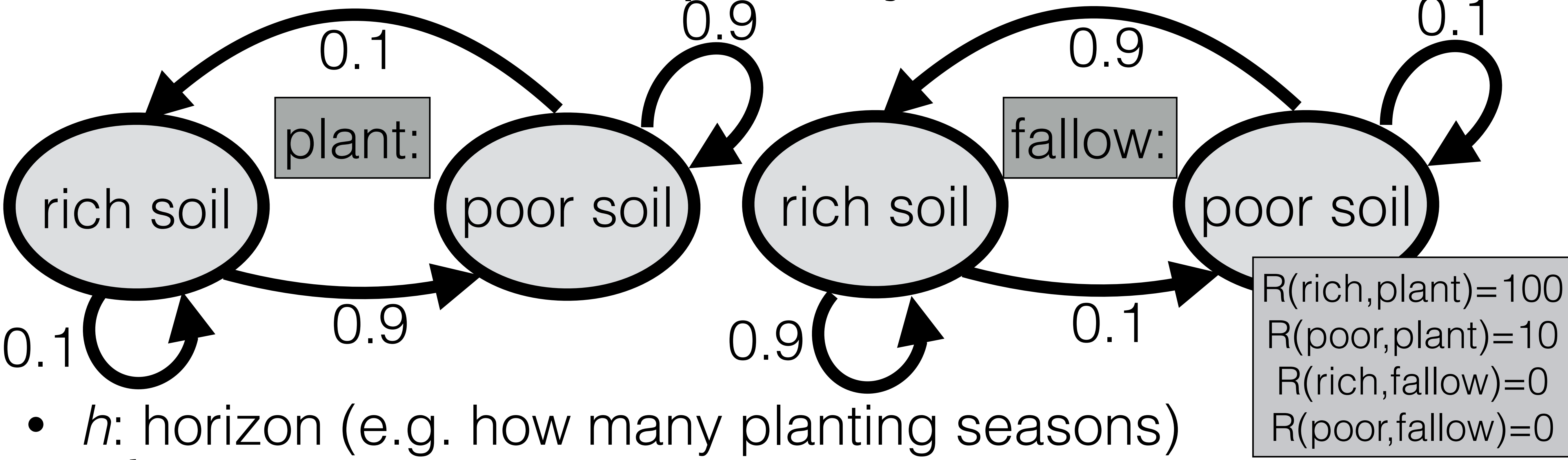
What’s the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

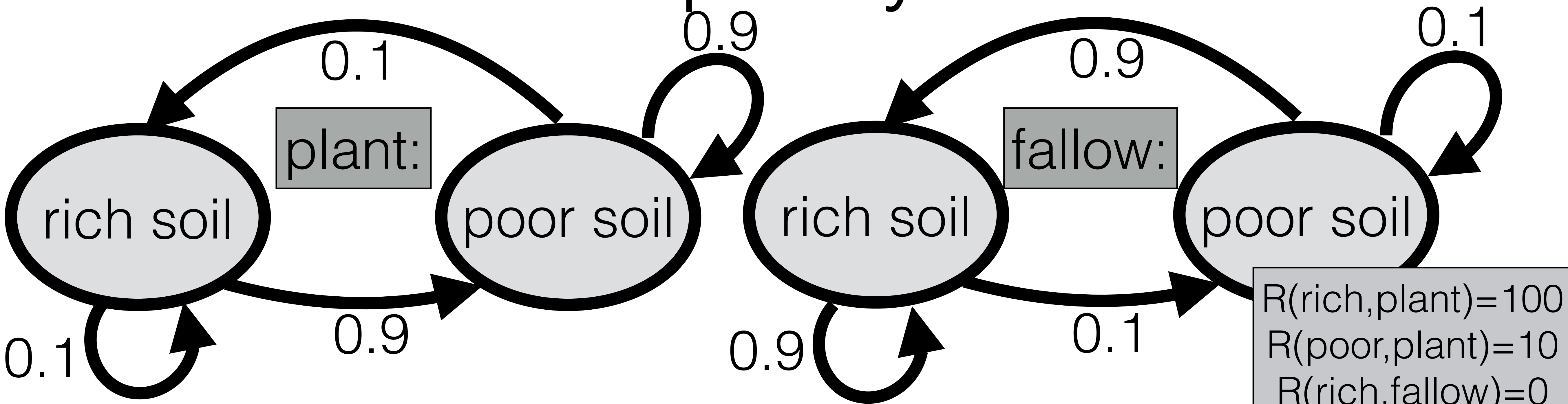
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the “best” action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

What's the best policy? Finite horizon



- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

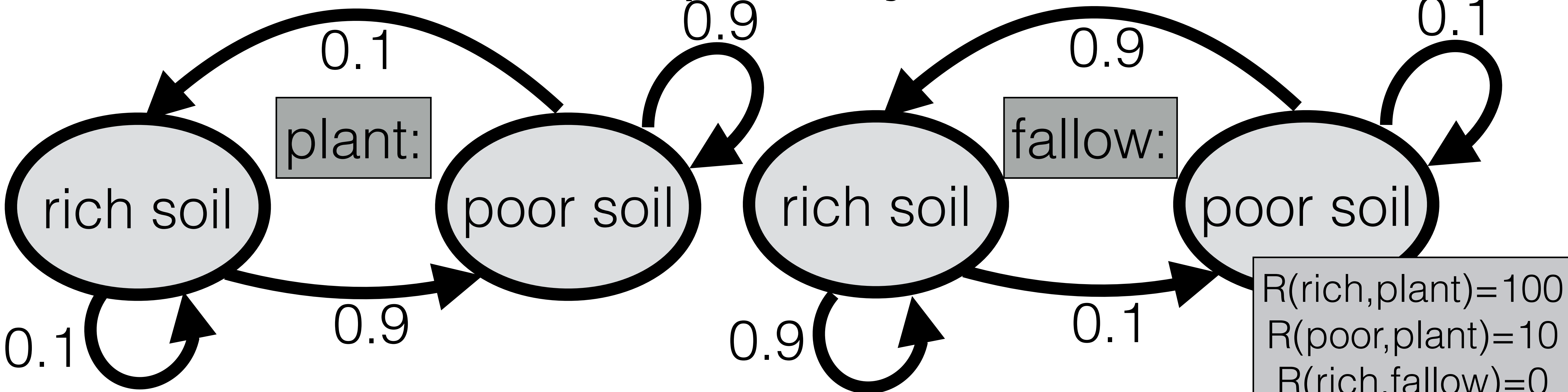
What's the best policy? Finite horizon



- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$Q^0(s,a) =$

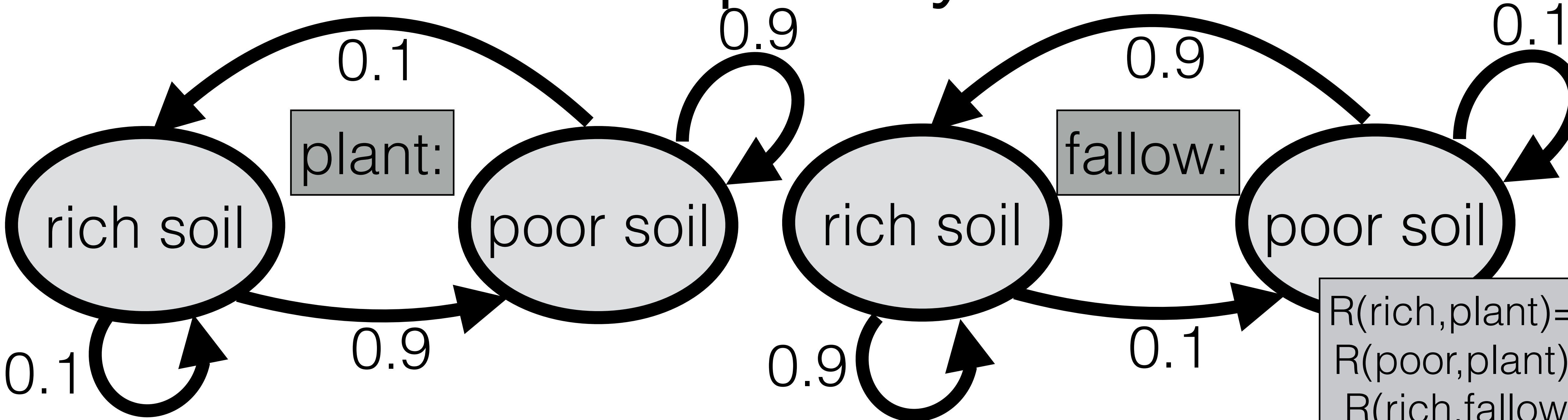
What's the best policy? Finite horizon



- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$Q^0(s,a) = 0$

What's the best policy? Finite horizon

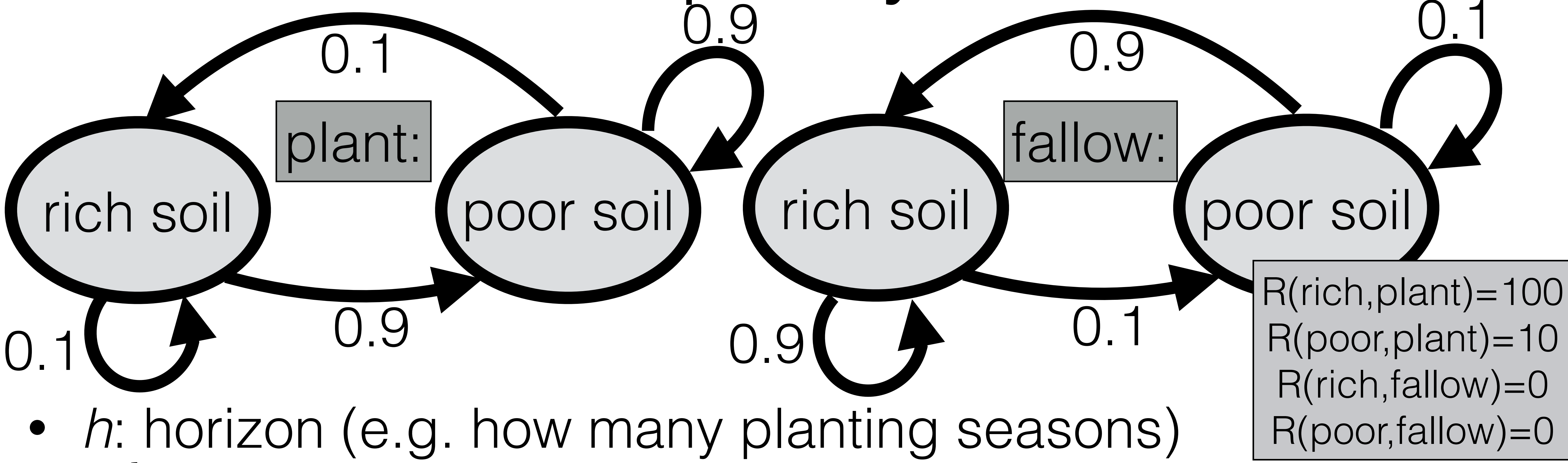


R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi^*_h(s) = \arg\max_a Q^h(s,a)$

$Q^0(s,a) = 0; Q^1(s,a) =$

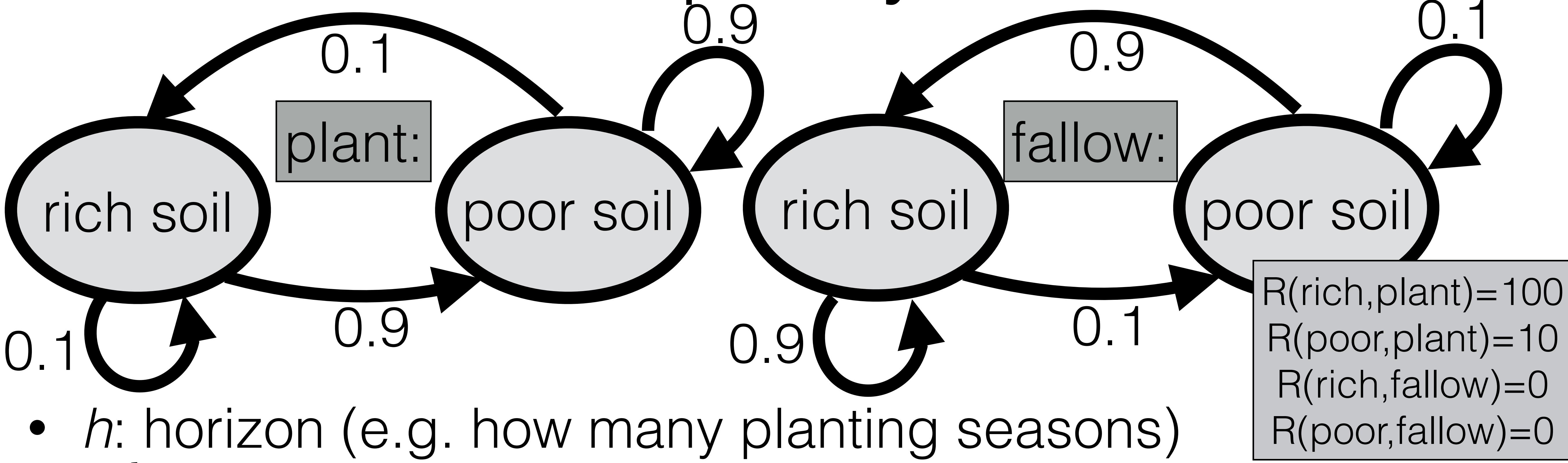
What's the best policy? Finite horizon



- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With Q, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s, a)$

$Q^0(s, a) = 0; Q^1(s, a) = R(s, a)$

What's the best policy? Finite horizon

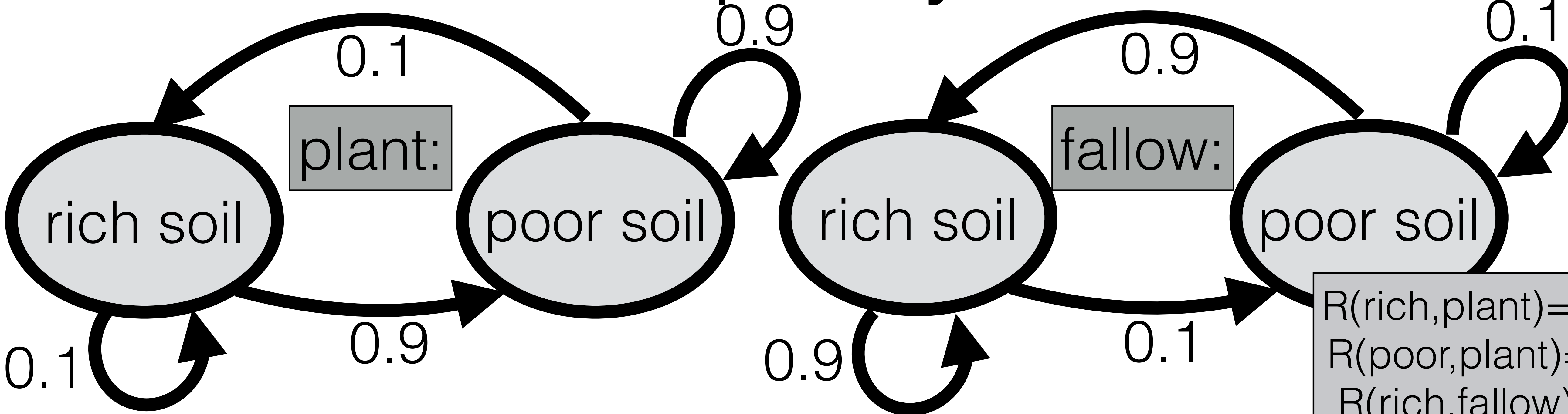


- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^1(s,a) = R(s,a)$$

$$Q^1(\text{rich}, \text{plant}) =$$

What's the best policy? Finite horizon



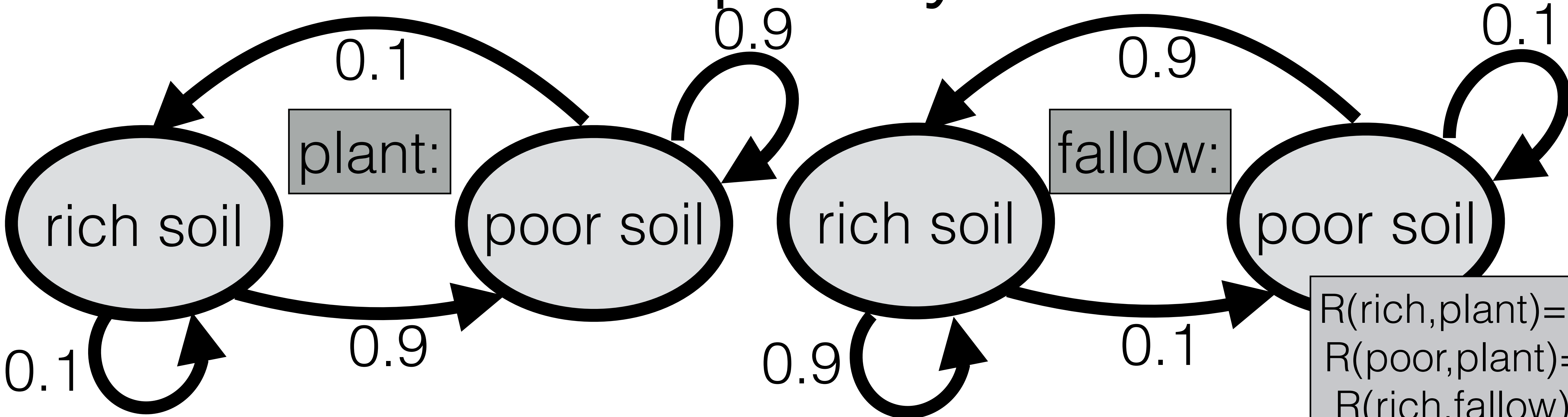
R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s, a)$

$$Q^0(s, a) = 0; Q^1(s, a) = R(s, a)$$

$$Q^1(\text{rich}, \text{plant}) = 100$$

What's the best policy? Finite horizon



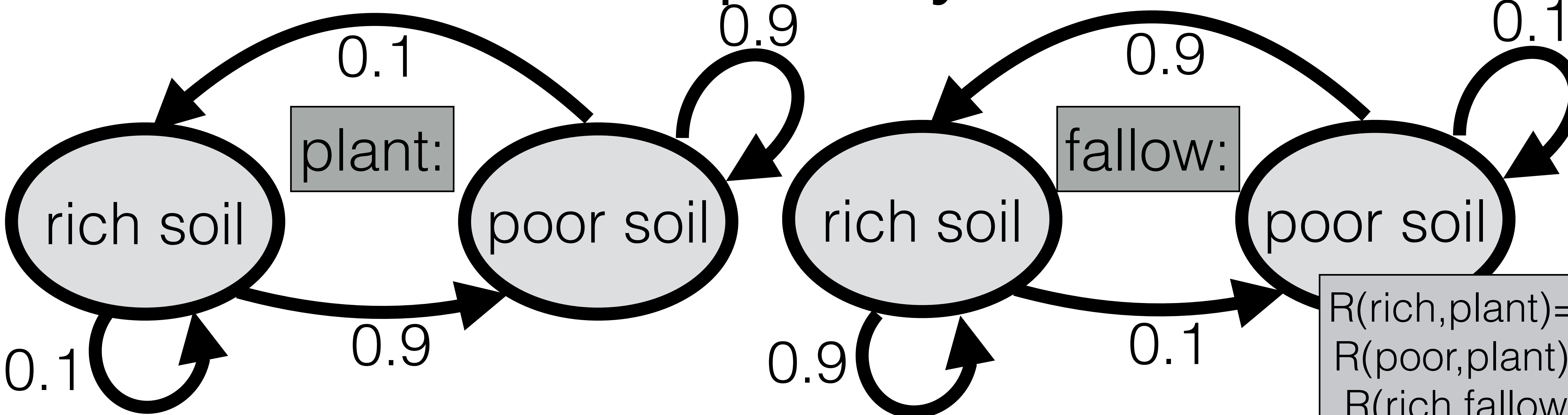
R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^1(s,a) = R(s,a)$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

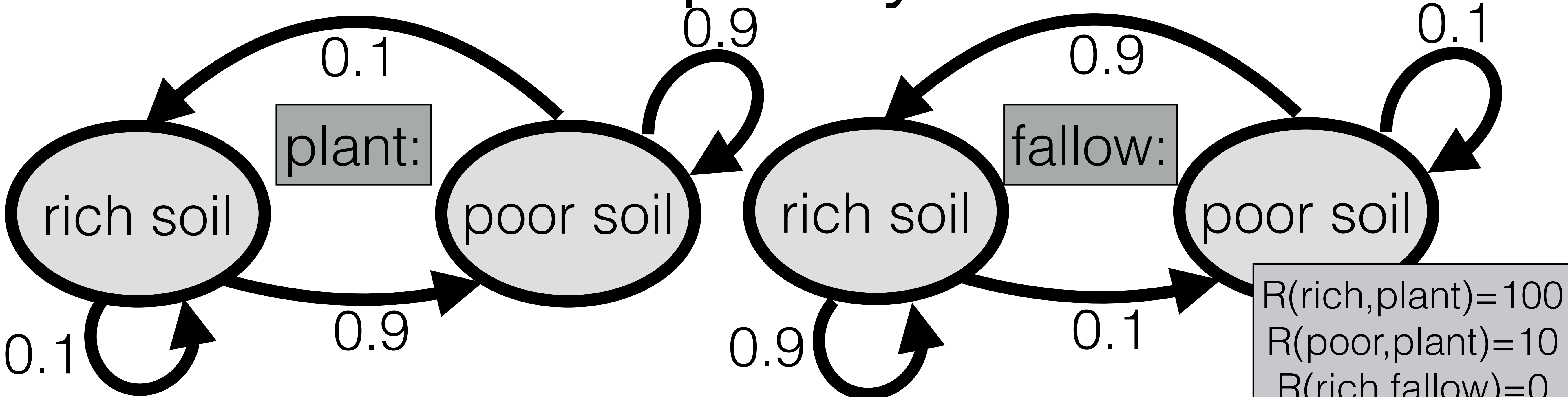
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^1(s,a) = R(s,a)$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

What's best?

What's the best policy? Finite horizon



- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

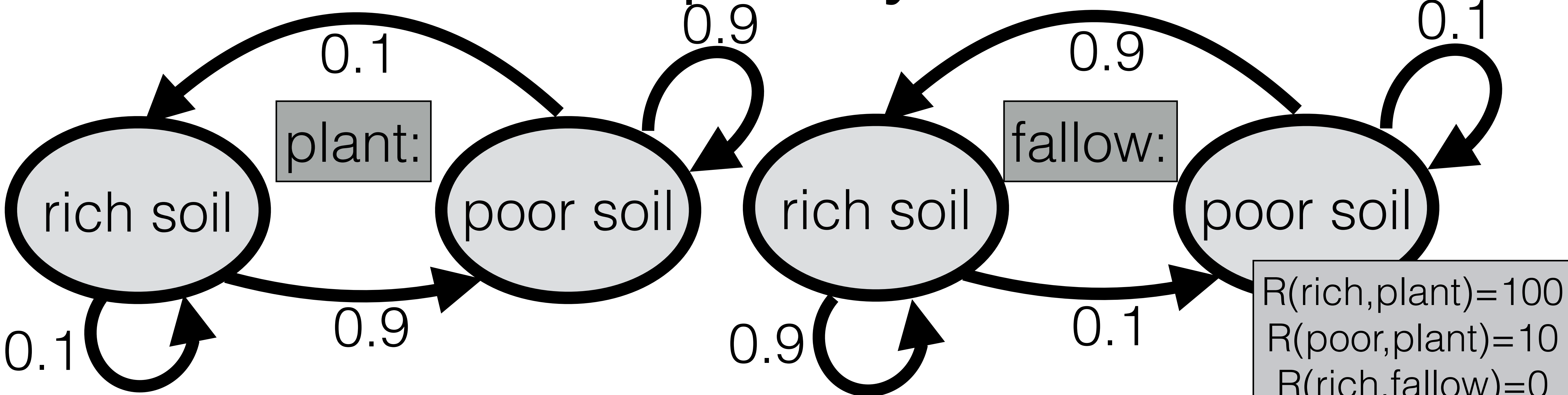
$$Q^0(s,a) = 0; Q^1(s,a) = R(s,a)$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

What's best?

π_1^*

What's the best policy? Finite horizon



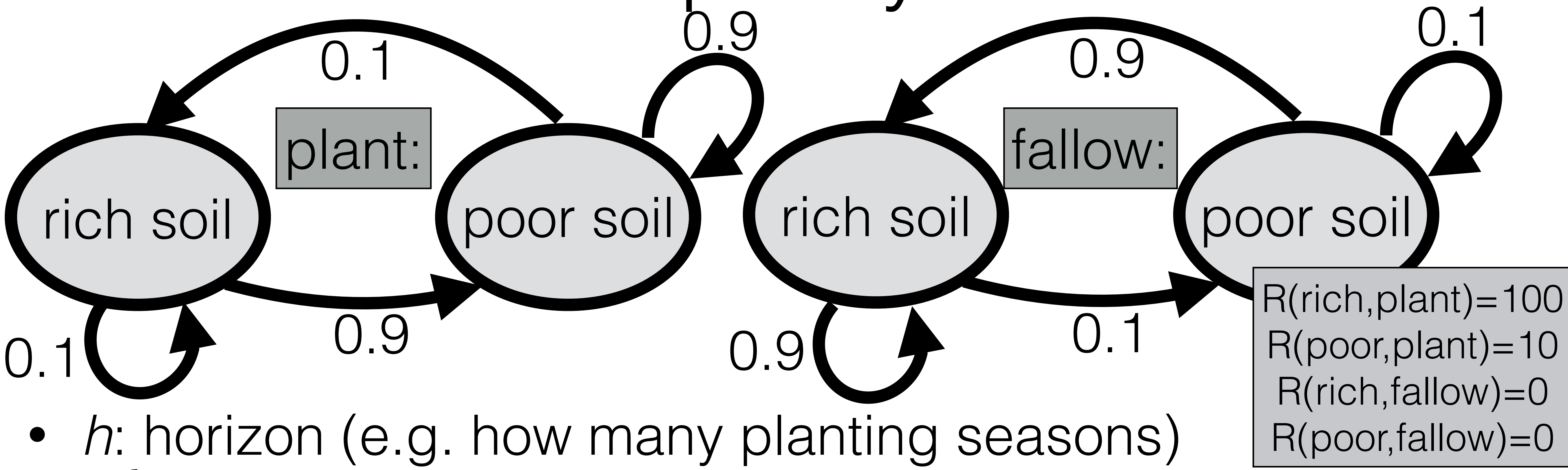
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^1(s,a) = R(s,a)$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



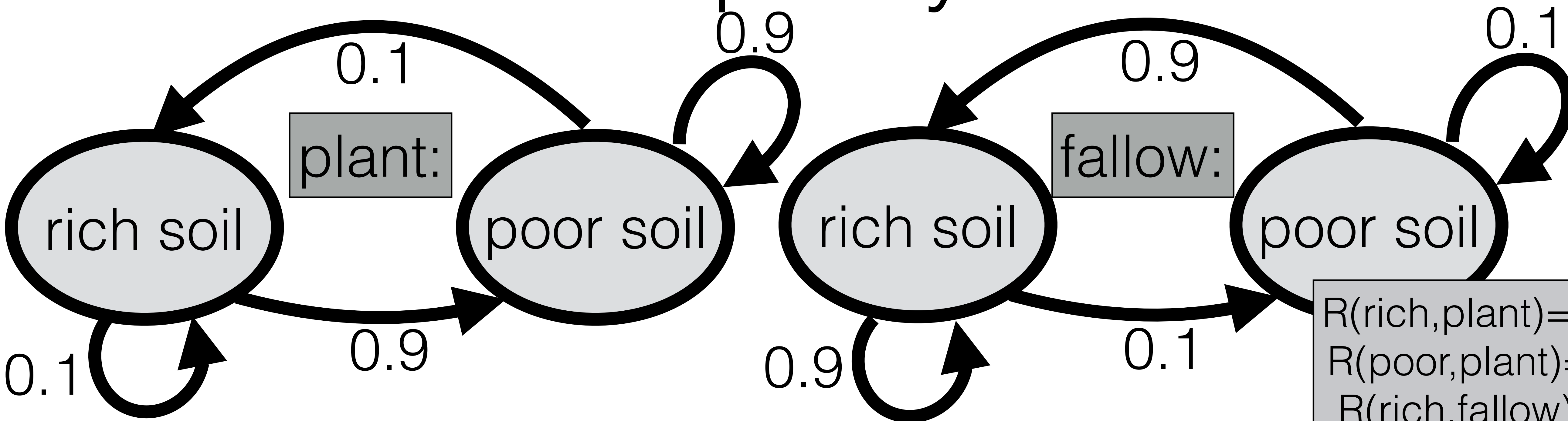
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^1(s,a) = R(s,a)$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

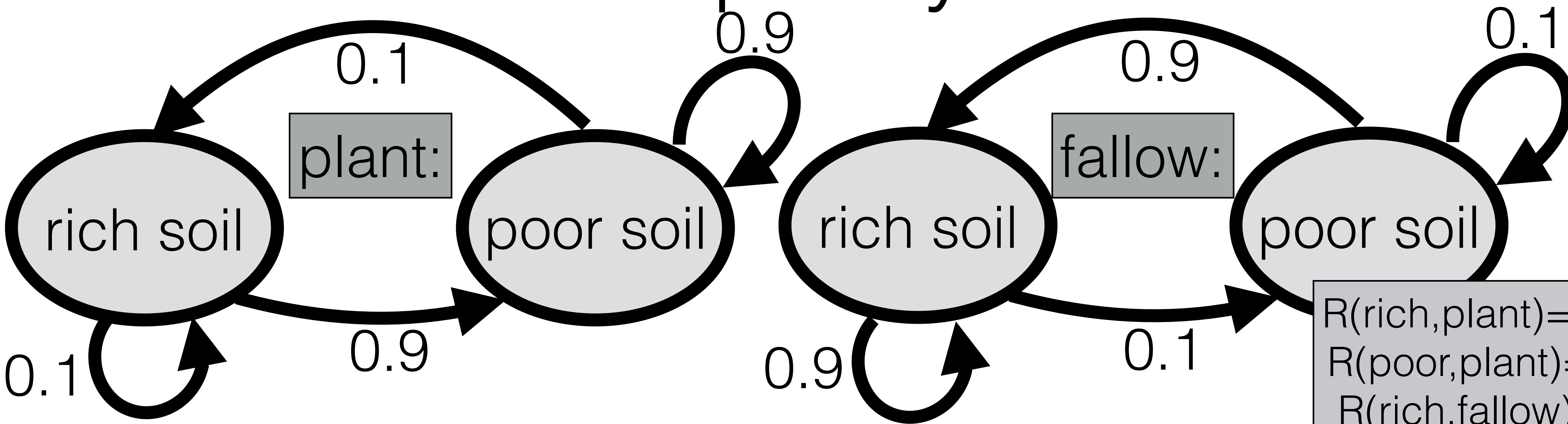
$Q^0(s,a) = 0; Q^1(s,a) = R(s,a)$

$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$

$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

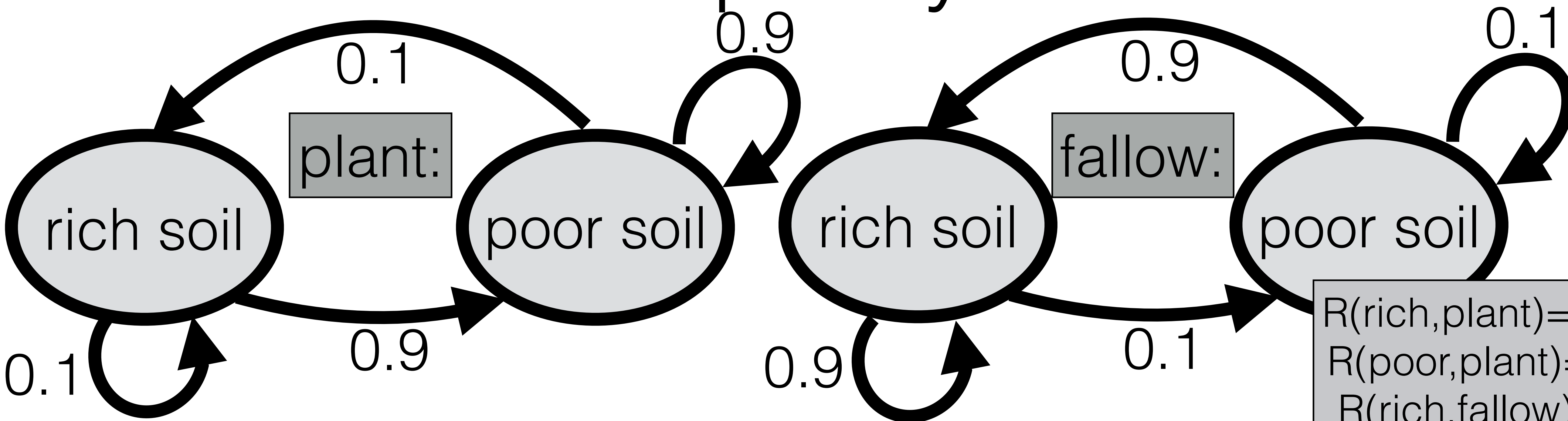
$Q^0(s,a) = 0; Q^1(s,a) = R(s,a)$

$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$

$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s, a)$

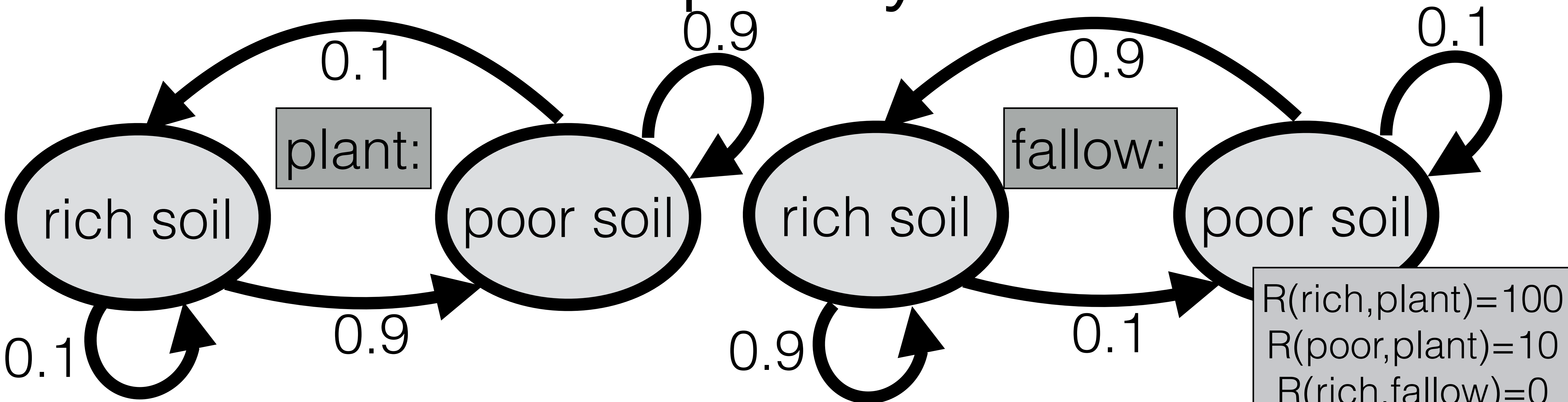
$Q^0(s, a) = 0; \; Q^h(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \max_{a'} Q^{h-1}(s', a')$

$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$

$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



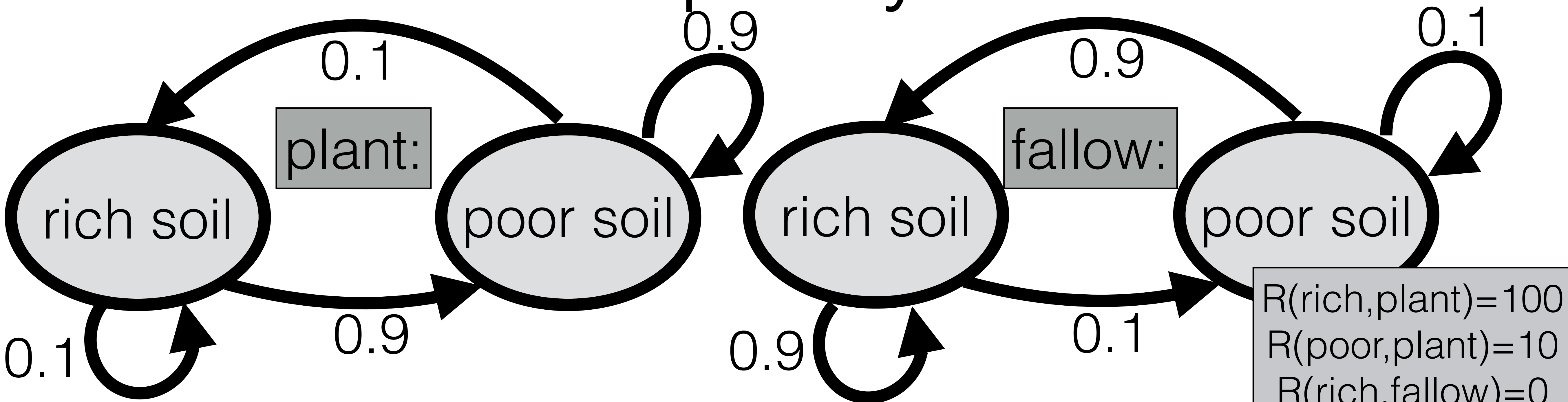
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; \; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



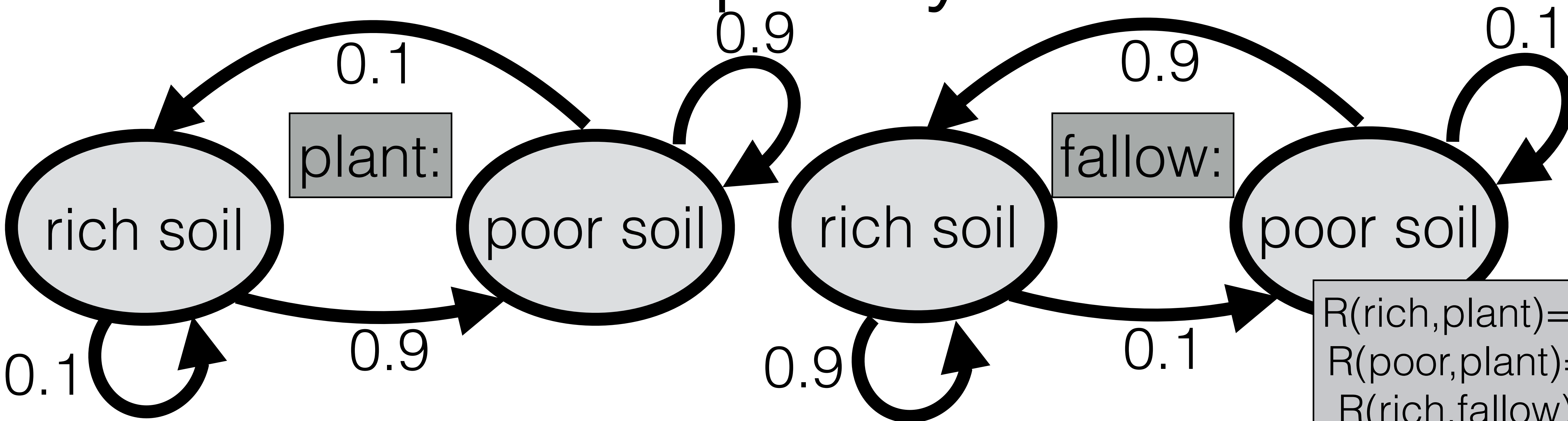
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

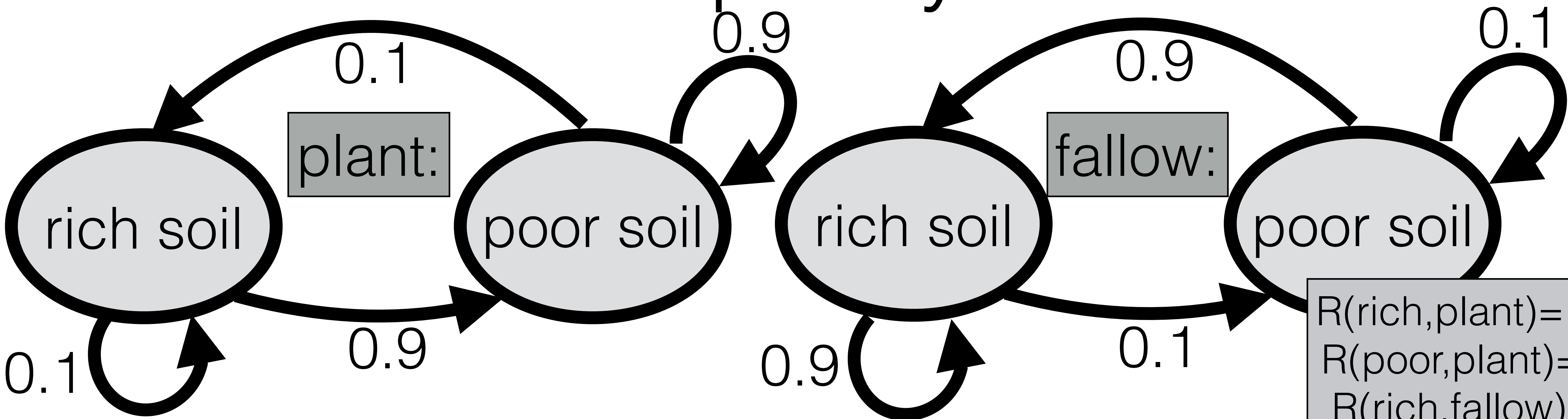
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

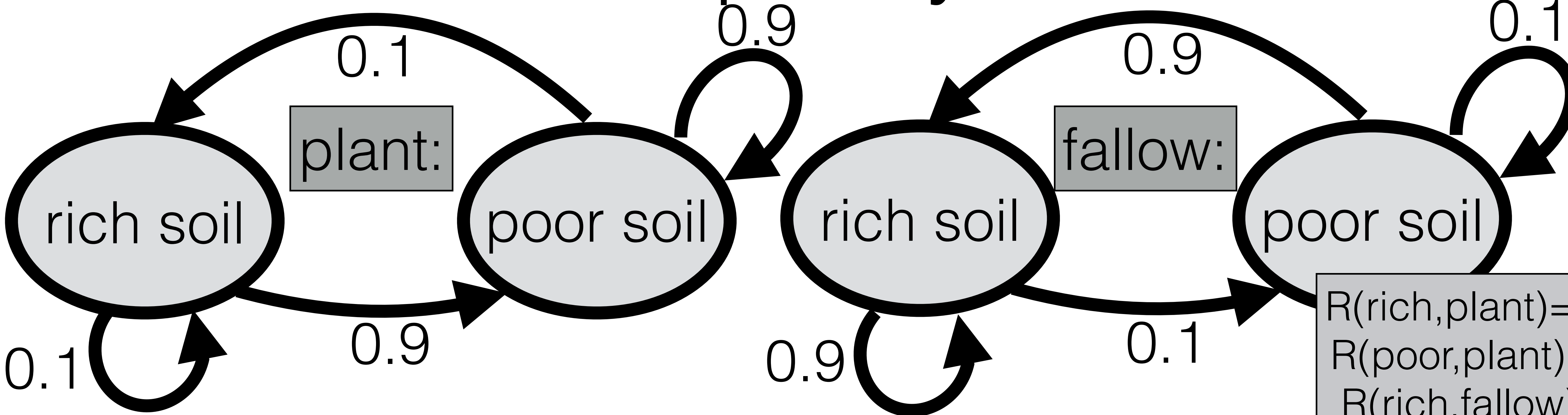
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

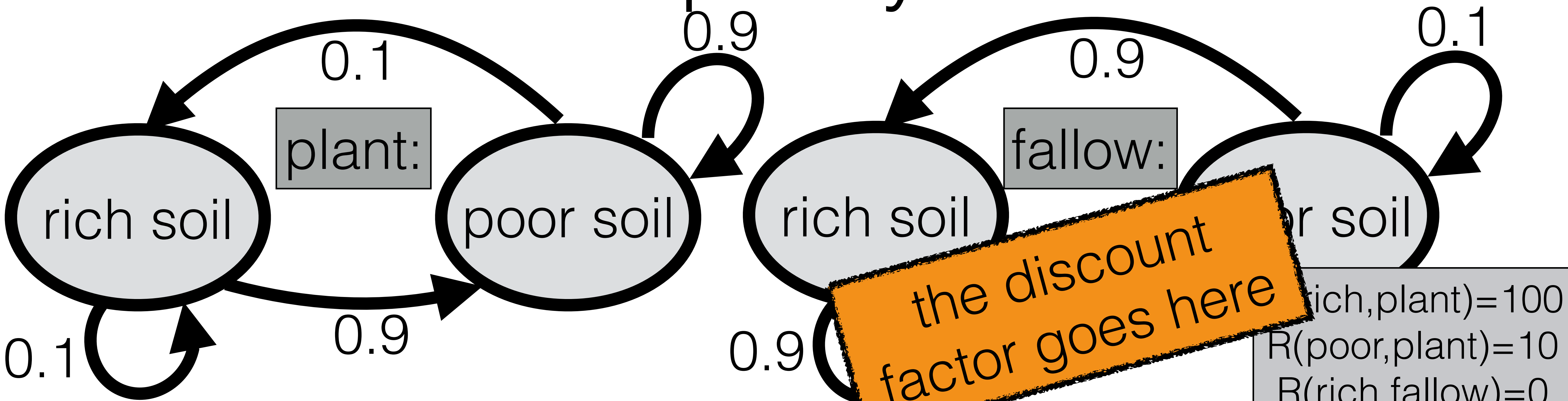
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



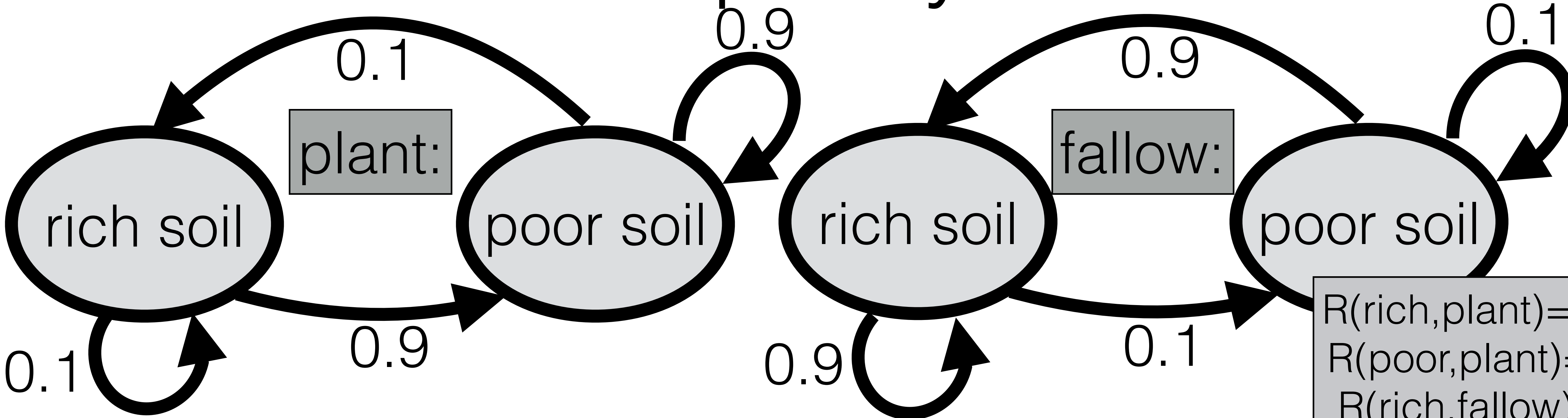
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

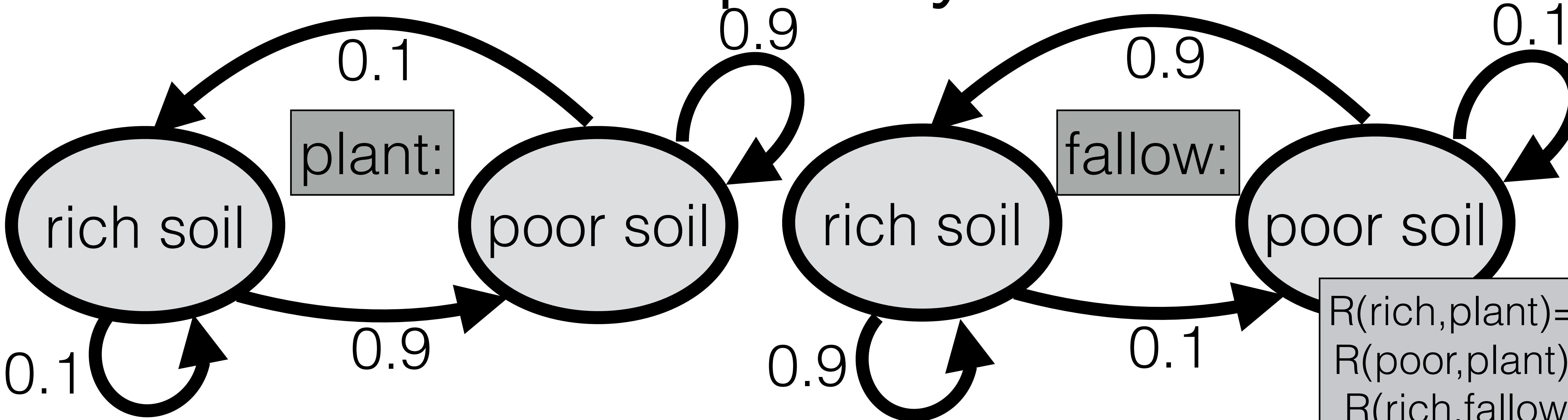
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

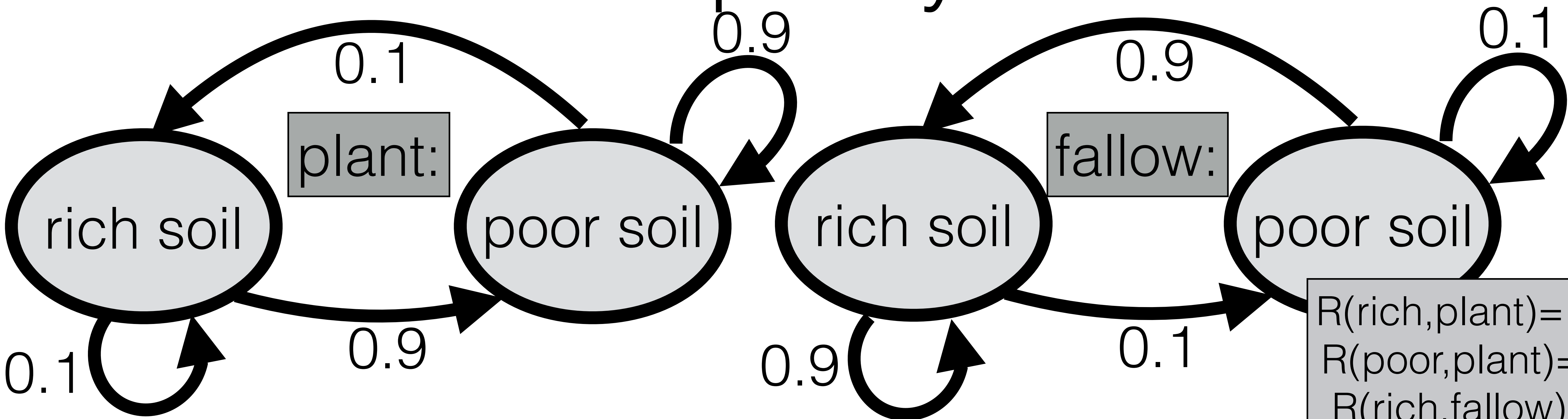
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

$Q^2(\text{rich}, \text{plant}) =$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi^*_h(s) = \arg\max_a Q^h(s, a)$

$Q^0(s, a) = 0; Q^h(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \max_{a'} Q^{h-1}(s', a')$

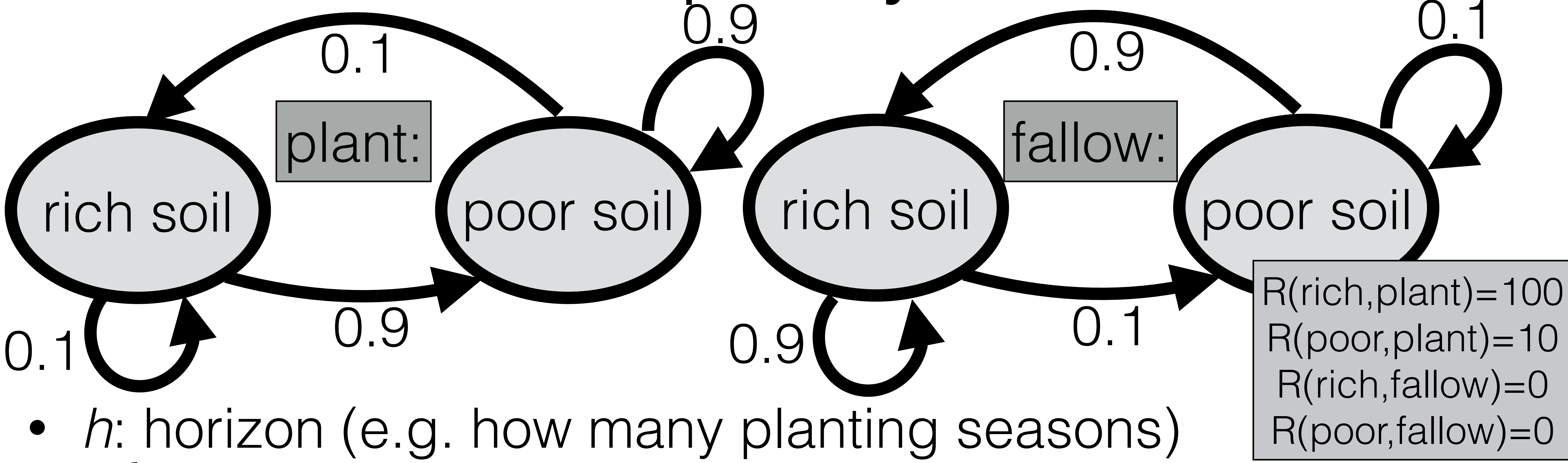
$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$

$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$

$Q^2(\text{rich}, \text{plant}) = R(\text{rich}, \text{plant}) +$

What's best? Any *s*, $\pi^*_1(s) = \text{plant}$

What's the best policy? Finite horizon



- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s, a)$

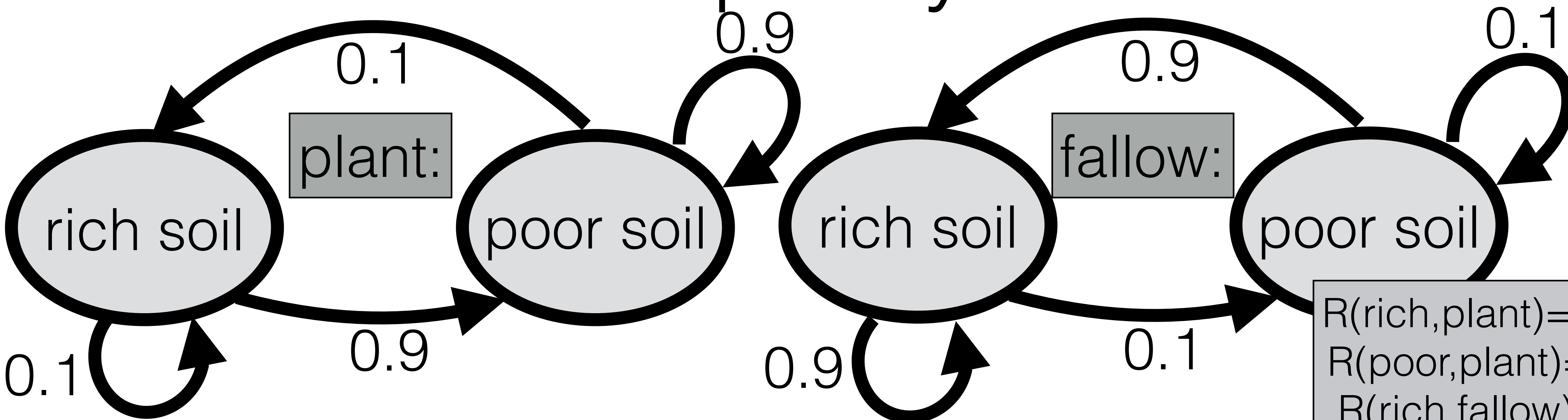
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

$$Q^2(\text{rich}, \text{plant}) = R(\text{rich}, \text{plant}) + T(\text{rich}, \text{plant}, \text{rich}) \max_{a'} Q^1(\text{rich}, a') + T(\text{rich}, \text{plant}, \text{poor}) \max_{a'} Q^1(\text{poor}, a')$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi^*_h(s) = \arg\max_a Q^h(s, a)$

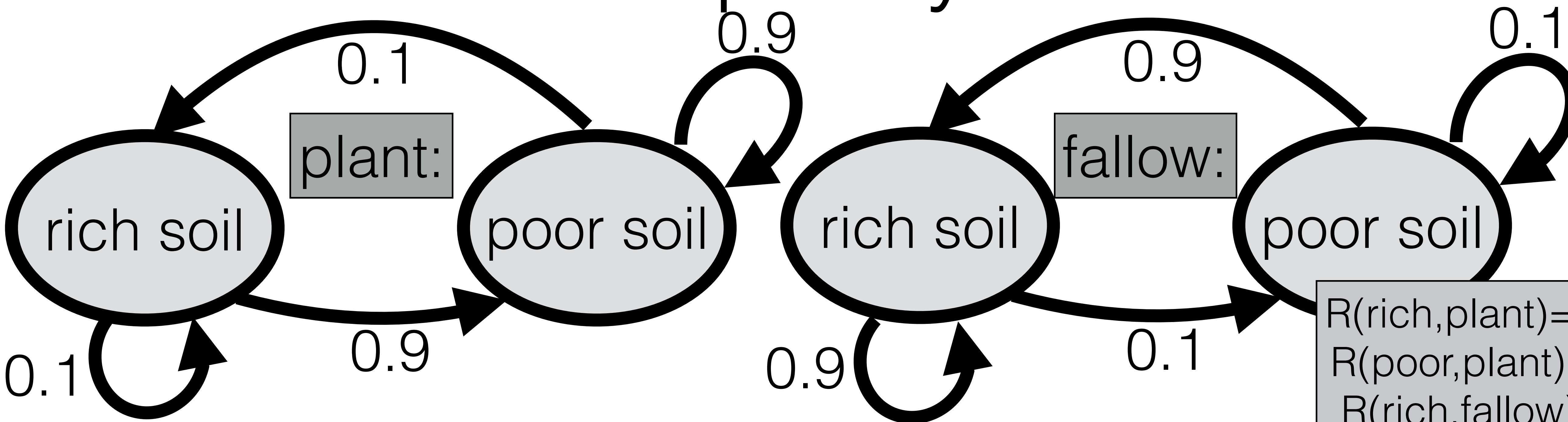
$$Q^0(s, a) = 0; Q^h(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \max_{a'} Q^{h-1}(s', a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

$$Q^2(\text{rich}, \text{plant}) = R(\text{rich}, \text{plant}) + T(\text{rich}, \text{plant}, \text{rich}) \max_{a'} Q^1(\text{rich}, a') + T(\text{rich}, \text{plant}, \text{poor}) \max_{a'} Q^1(\text{poor}, a')$$

What's best? Any *s*, $\pi^*_1(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi^*_h(s) = \arg\max_a Q^h(s,a)$

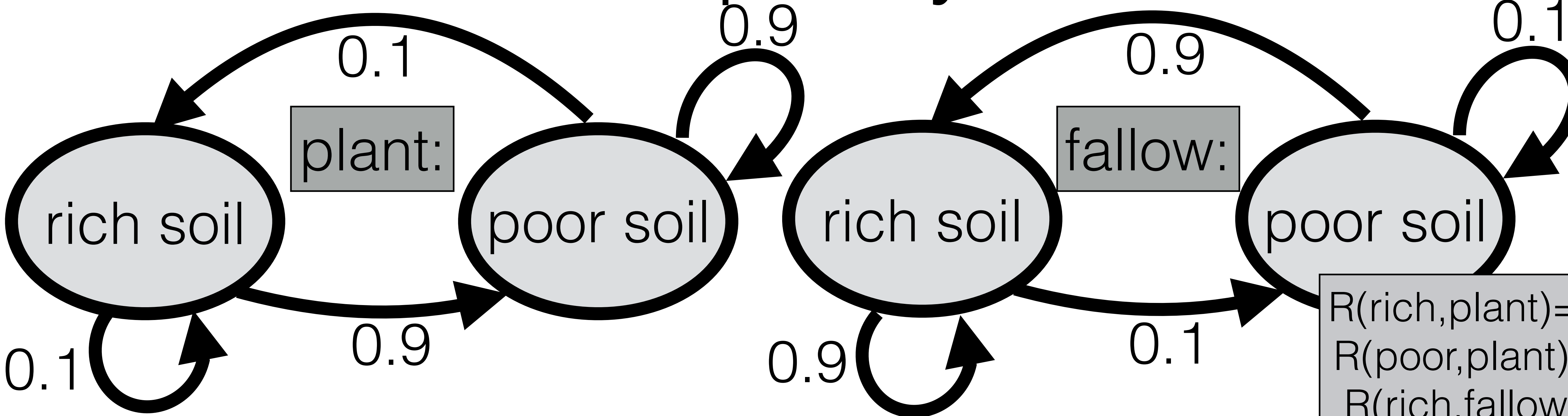
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

$$Q^2(\text{rich},\text{plant}) = R(\text{rich},\text{plant}) + T(\text{rich},\text{plant},\text{rich}) \max_{a'} Q^1(\text{rich},a') + T(\text{rich},\text{plant},\text{poor}) \max_{a'} Q^1(\text{poor},a')$$

What's best? Any *s*, $\pi^*_1(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi^*_h(s) = \arg\max_a Q^h(s,a)$

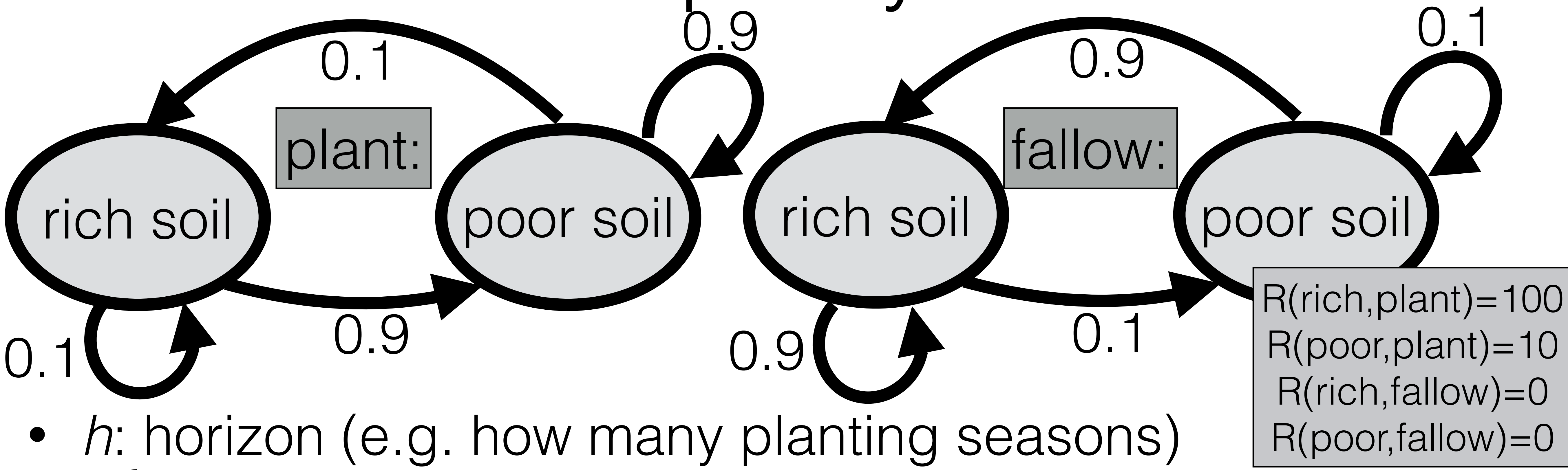
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

$$Q^2(\text{rich}, \text{plant}) = R(\text{rich}, \text{plant}) + T(\text{rich}, \text{plant}, \text{rich}) \max_{a'} Q^1(\text{rich}, a') + T(\text{rich}, \text{plant}, \text{poor}) \max_{a'} Q^1(\text{poor}, a')$$

What's best? Any *s*, $\pi^*_1(s) = \text{plant}$

What's the best policy? Finite horizon



- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

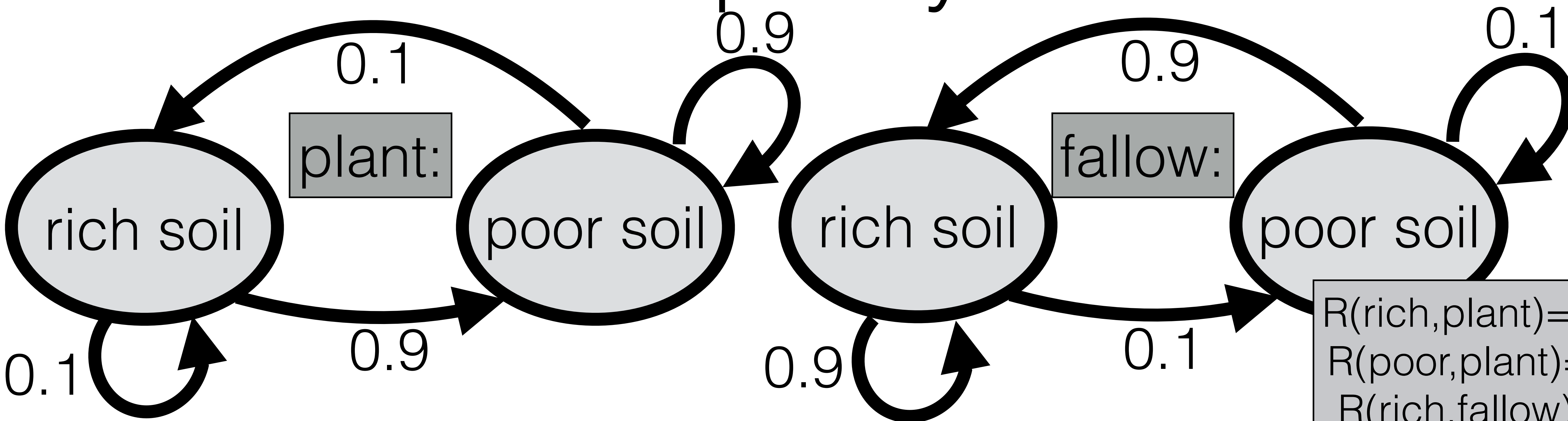
$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$

$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

$$Q^2(\text{rich},\text{plant}) = R(\text{rich},\text{plant}) + T(\text{rich},\text{plant},\text{rich}) \max_{a'} Q^1(\text{rich},a') + T(\text{rich},\text{plant},\text{poor}) \max_{a'} Q^1(\text{poor},a')$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s, a)$

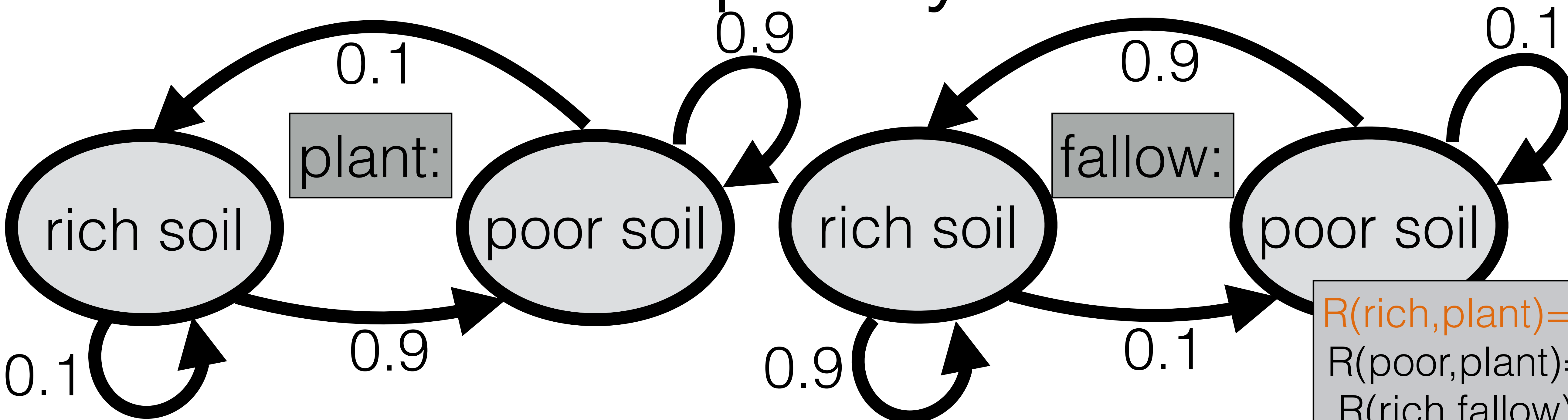
$$Q^0(s, a) = 0; Q^h(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \max_{a'} Q^{h-1}(s', a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

$$Q^2(\text{rich}, \text{plant}) = R(\text{rich}, \text{plant}) + T(\text{rich}, \text{plant}, \text{rich}) \max_{a'} Q^1(\text{rich}, a') + T(\text{rich}, \text{plant}, \text{poor}) \max_{a'} Q^1(\text{poor}, a')$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi^*_h(s) = \arg\max_a Q^h(s,a)$

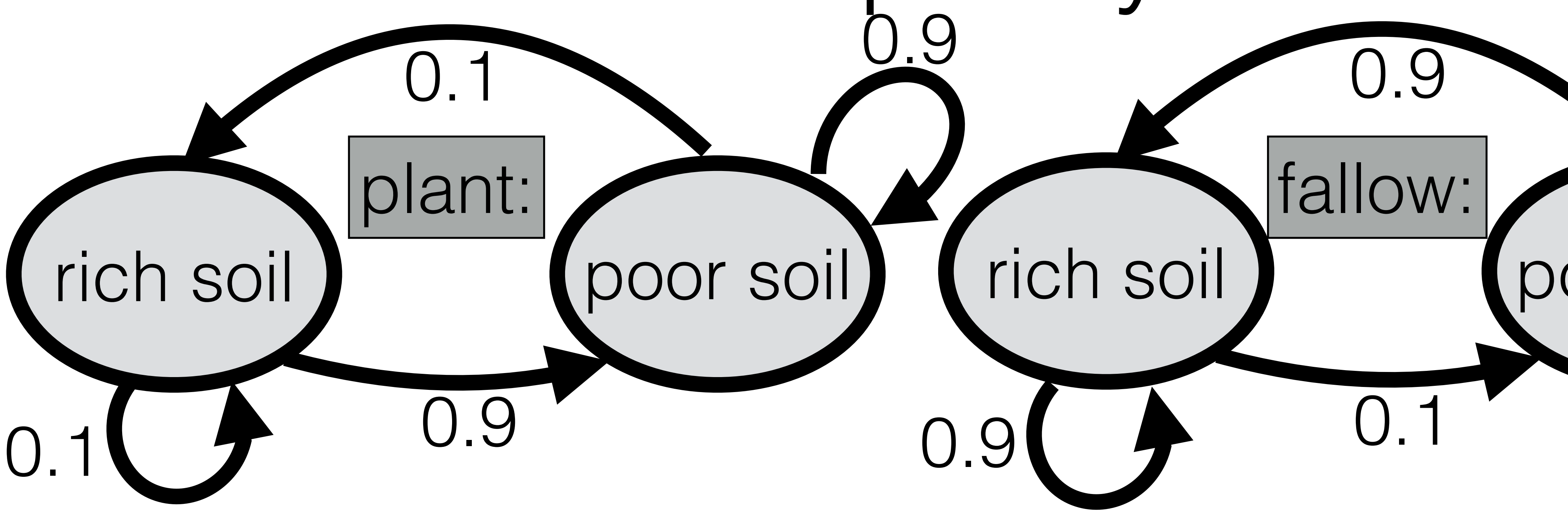
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

$$Q^2(\text{rich},\text{plant}) = R(\text{rich},\text{plant}) + T(\text{rich},\text{plant},\text{rich}) \max_{a'} Q^1(\text{rich},a')$$
$$+ T(\text{rich},\text{plant},\text{poor}) \max_{a'} Q^1(\text{poor},a')$$

What's best? Any *s*, $\pi^*_1(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

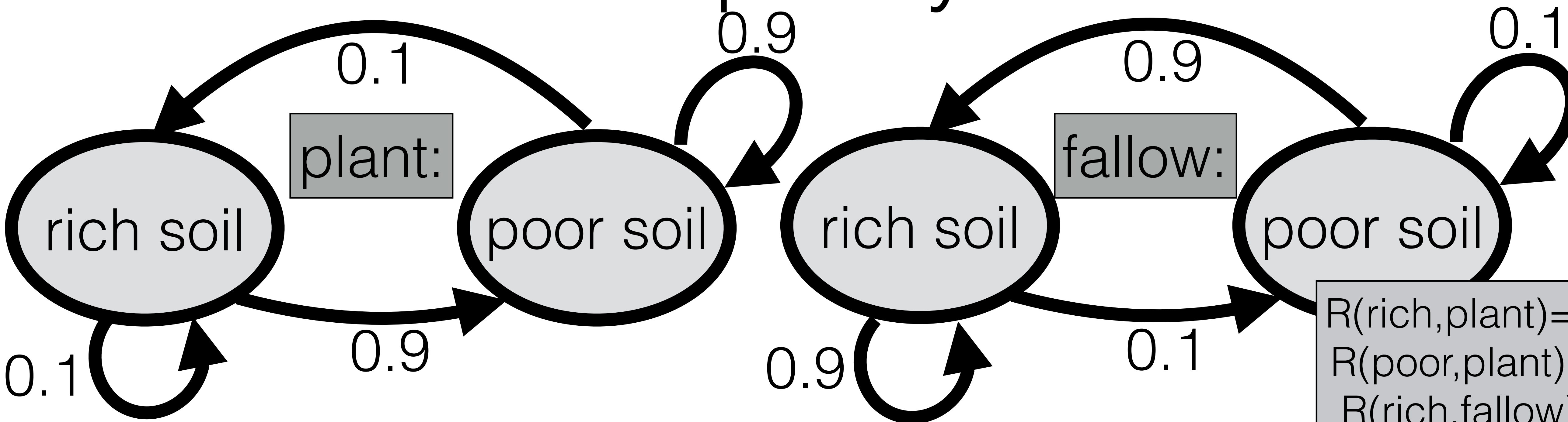
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

$$Q^2(\text{rich},\text{plant}) = 100 + T(\text{rich},\text{plant},\text{rich}) \max_{a'} Q^1(\text{rich},a') + T(\text{rich},\text{plant},\text{poor}) \max_{a'} Q^1(\text{poor},a')$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi^*_h(s) = \arg\max_a Q^h(s,a)$

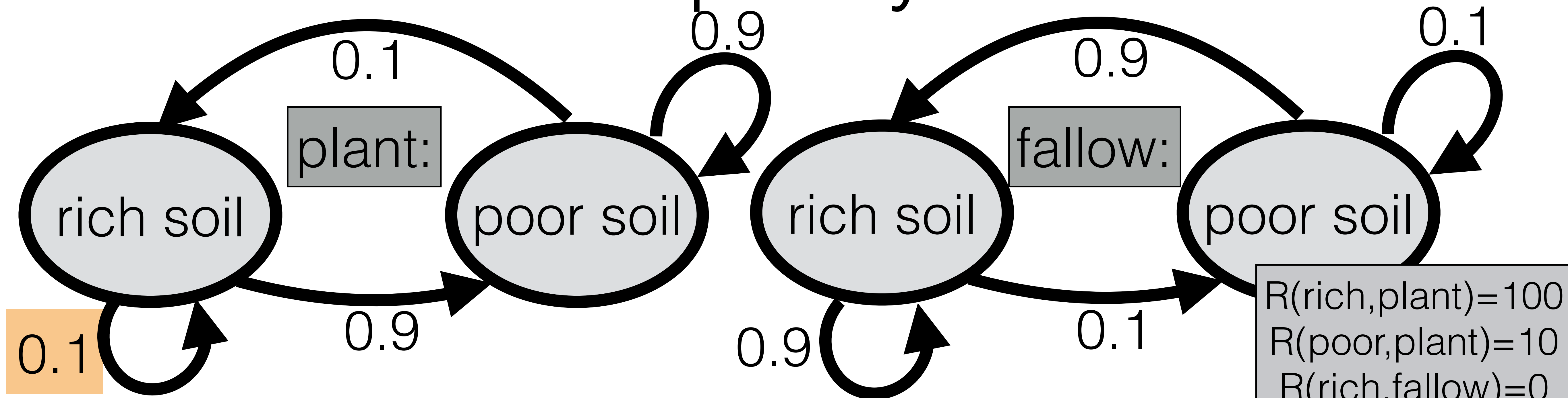
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

$$Q^2(\text{rich},\text{plant}) = 100 + T(\text{rich},\text{plant},\text{rich}) \max_{a'} Q^1(\text{rich},a') + T(\text{rich},\text{plant},\text{poor}) \max_{a'} Q^1(\text{poor},a')$$

What's best? Any *s*, $\pi^*_1(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

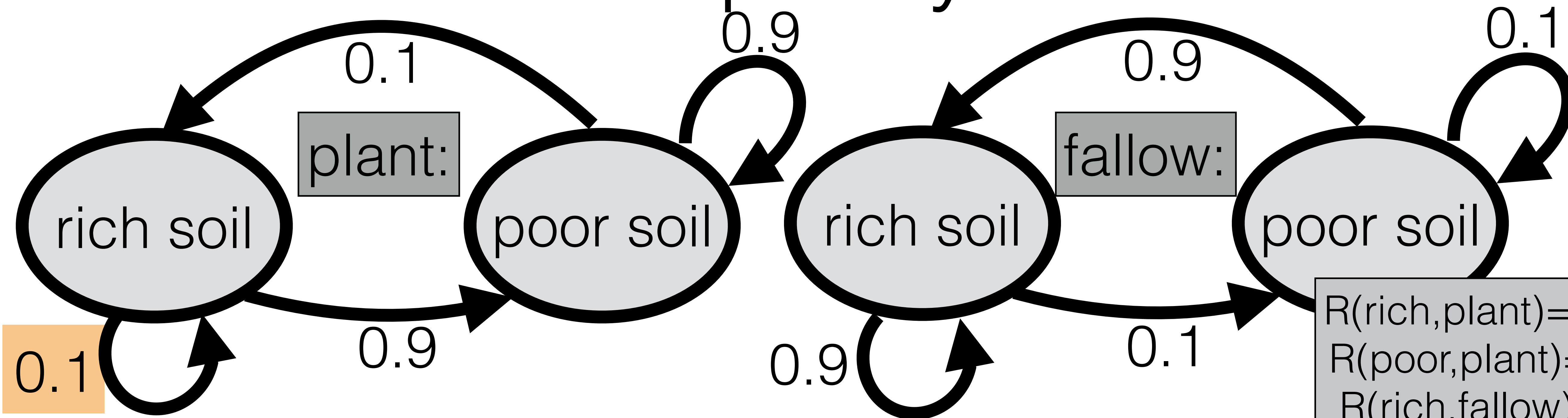
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

$$Q^2(\text{rich}, \text{plant}) = 100 + T(\text{rich}, \text{plant}, \text{rich}) \max_{a'} Q^1(\text{rich}, a') + T(\text{rich}, \text{plant}, \text{poor}) \max_{a'} Q^1(\text{poor}, a')$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s, a)$

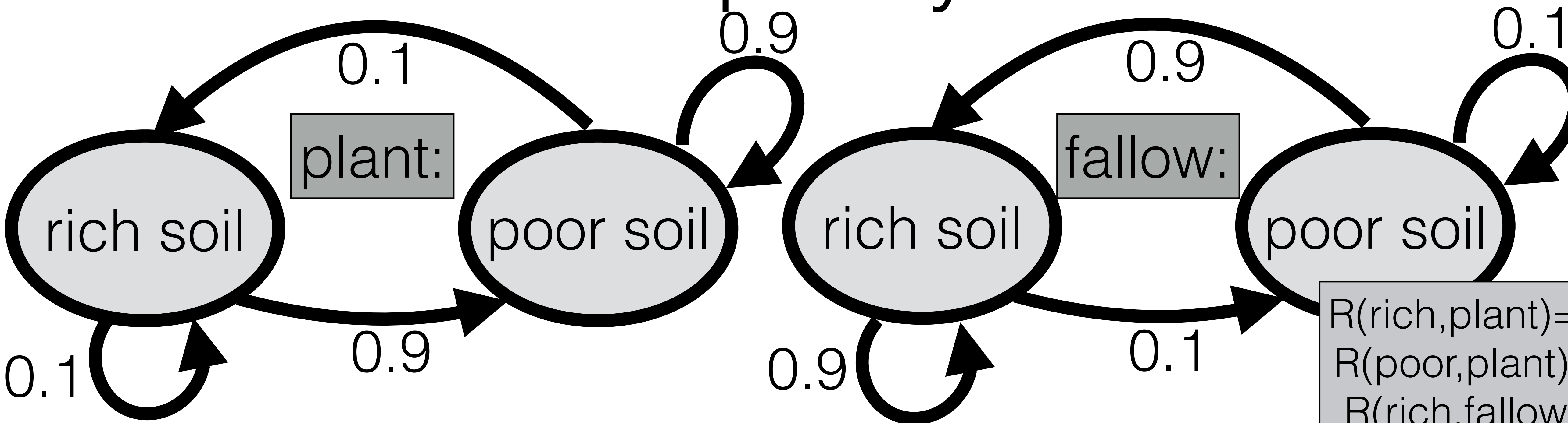
$$Q^0(s, a) = 0; Q^h(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \max_{a'} Q^{h-1}(s', a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

$$Q^2(\text{rich}, \text{plant}) = 100 + (0.1) \max_{a'} Q^1(\text{rich}, a') + T(\text{rich}, \text{plant}, \text{poor}) \max_{a'} Q^1(\text{poor}, a')$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

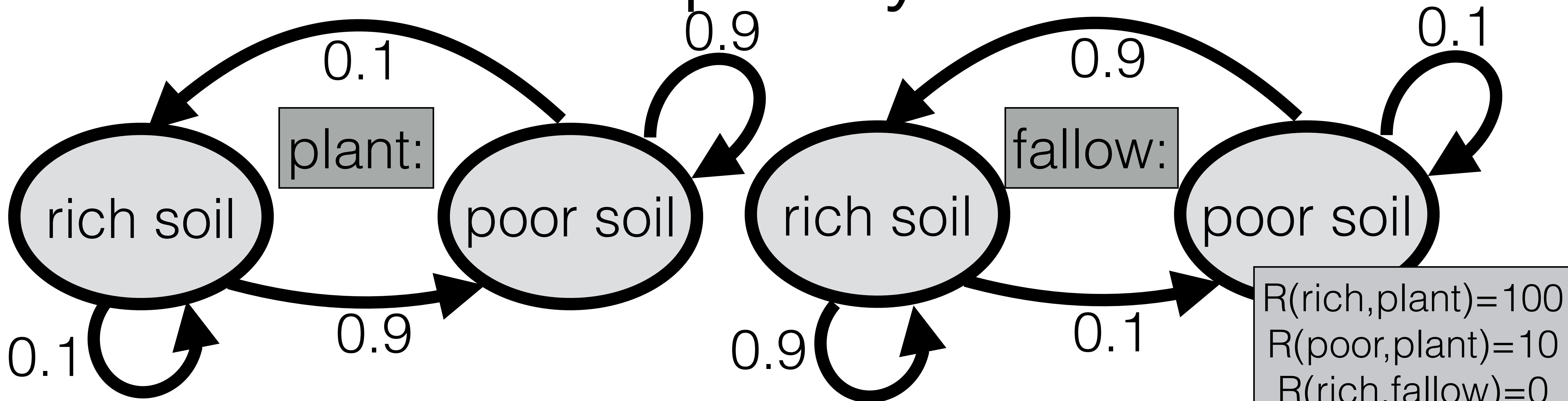
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

$$Q^2(\text{rich},\text{plant}) = 100 + (0.1) \max_{a'} Q^1(\text{rich},a') + T(\text{rich},\text{plant},\text{poor}) \max_{a'} Q^1(\text{poor},a')$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

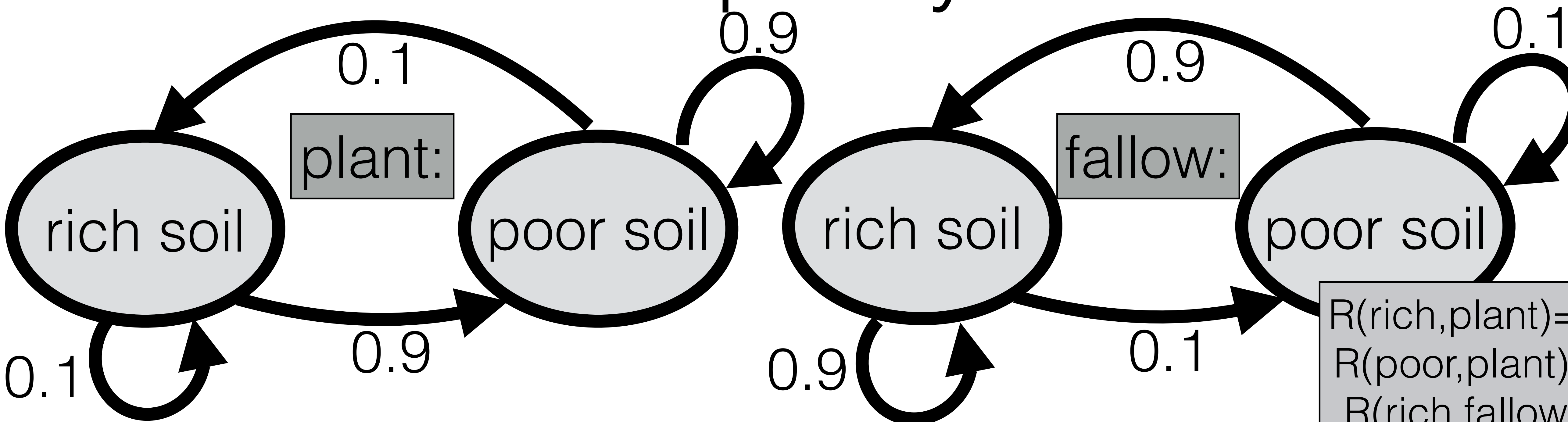
$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$

$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

$$Q^2(\text{rich},\text{plant}) = 100 + (0.1)\max_{a'} Q^1(\text{rich},a') + T(\text{rich},\text{plant},\text{poor}) \max_{a'} Q^1(\text{poor},a')$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

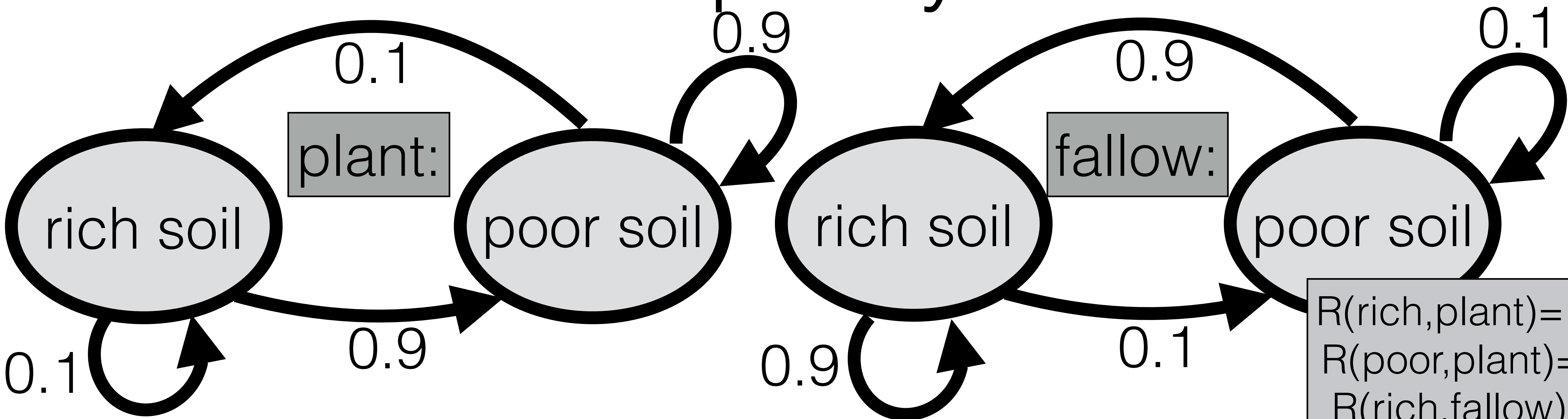
$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$

$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

$$Q^2(\text{rich},\text{plant}) = 100 + (0.1)(100) + T(\text{rich},\text{plant},\text{poor}) \max_{a'} Q^1(\text{poor},a')$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

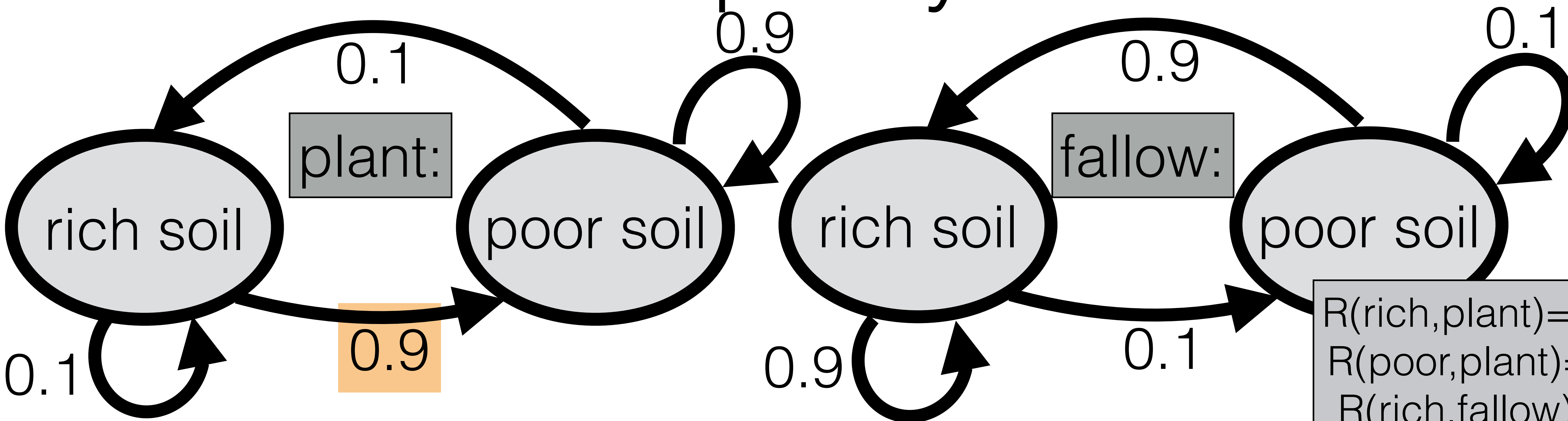
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

$$Q^2(\text{rich},\text{plant}) = 100 + (0.1)(100) + T(\text{rich},\text{plant},\text{poor}) \max_{a'} Q^1(\text{poor},a')$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

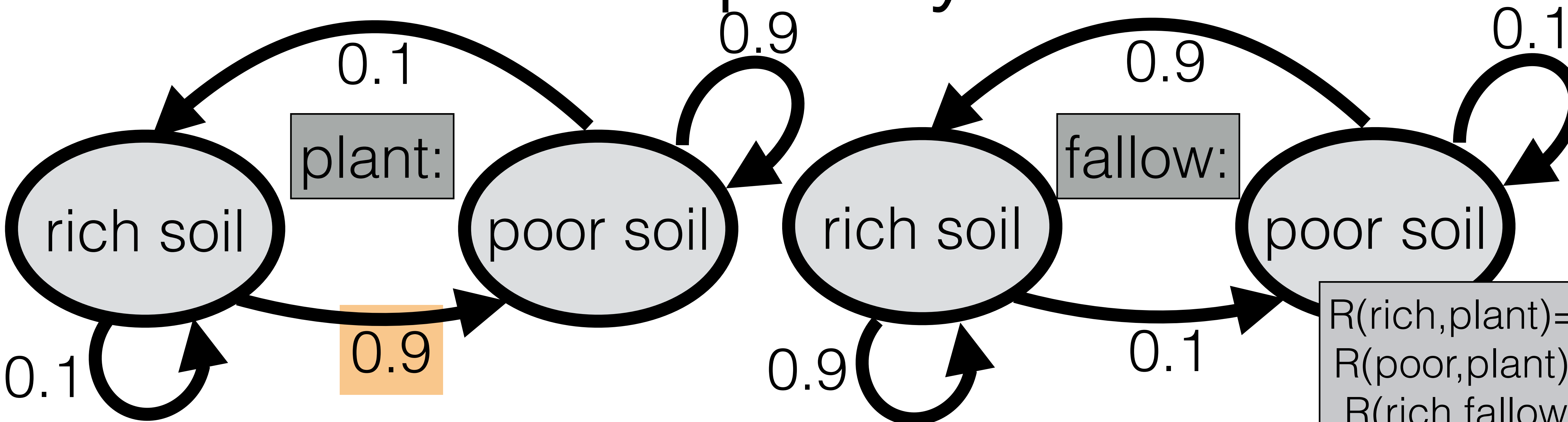
$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$

$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

$$Q^2(\text{rich},\text{plant}) = 100 + (0.1)(100) + T(\text{rich},\text{plant},\text{poor}) \max_{a'} Q^1(\text{poor},a')$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi^*_h(s) = \arg\max_a Q^h(s,a)$

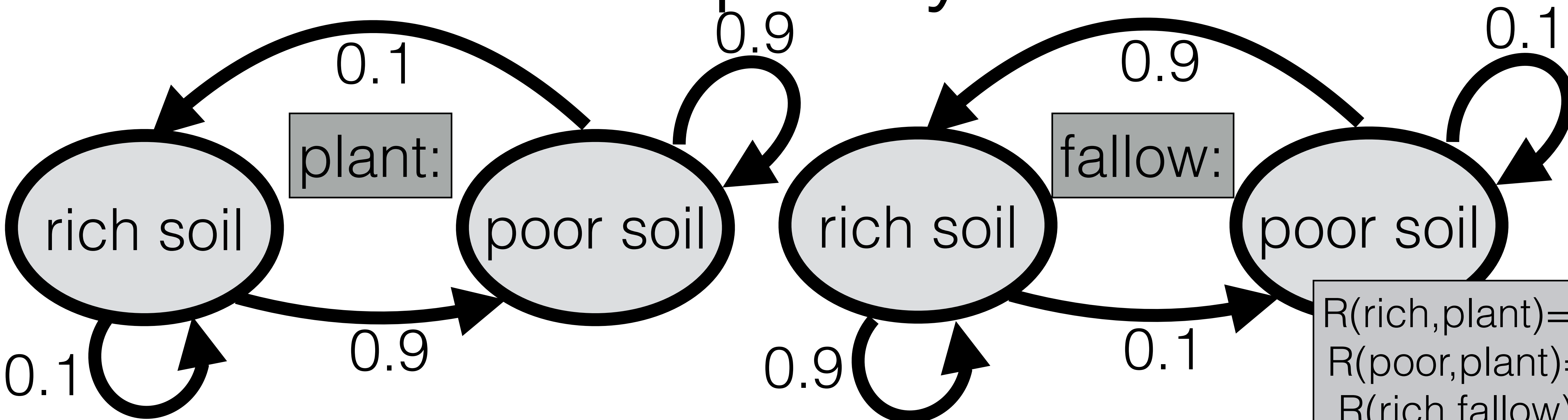
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

$$Q^2(\text{rich},\text{plant}) = 100 + (0.1)(100) + (0.9)\max_{a'} Q^1(\text{poor},a')$$

What's best? Any *s*, $\pi^*_1(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

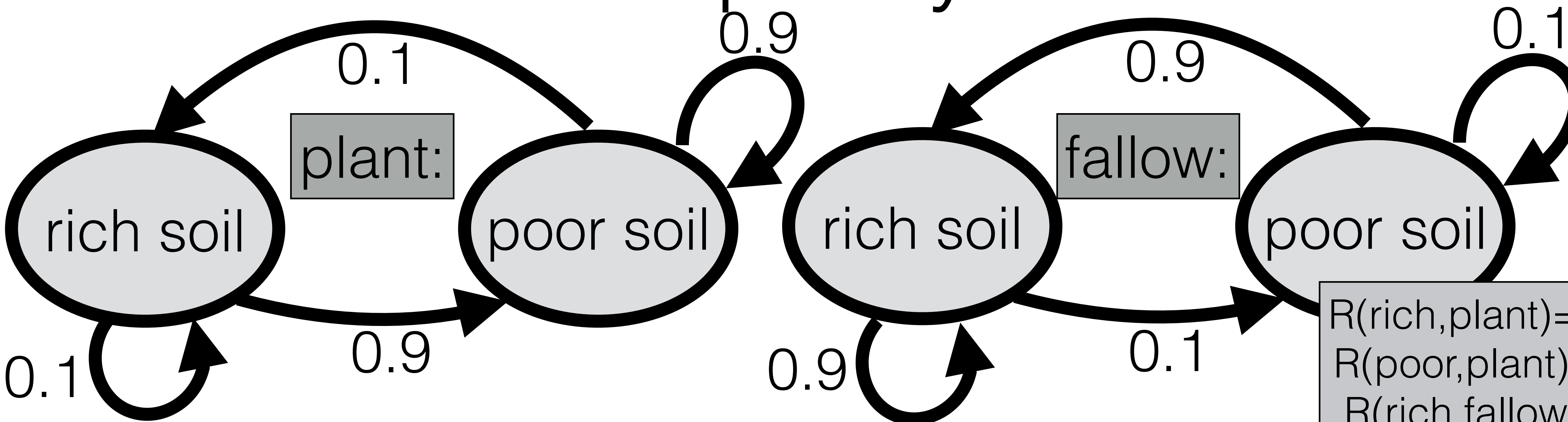
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

$$Q^2(\text{rich},\text{plant}) = 100 + (0.1)(100) + (0.9)\max_{a'} Q^1(\text{poor},a')$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s, a)$

$$Q^0(s, a) = 0; Q^h(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \max_{a'} Q^{h-1}(s', a')$$

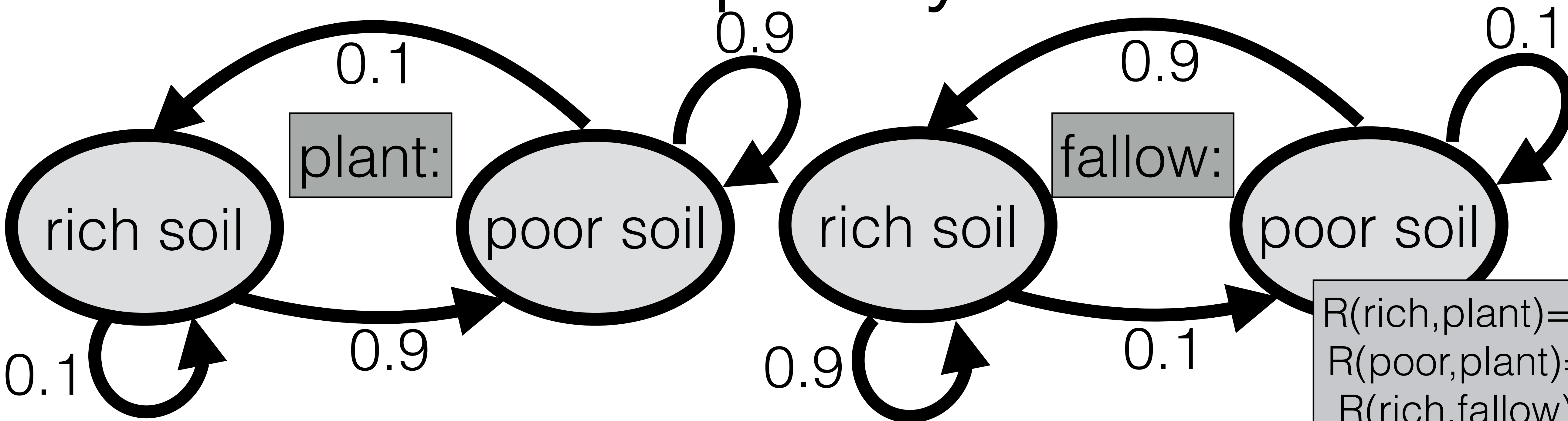
$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$

$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

$$Q^2(\text{rich}, \text{plant}) = 100 + (0.1)(100) + (0.9) \max_{a'} Q^1(\text{poor}, a')$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s, a)$

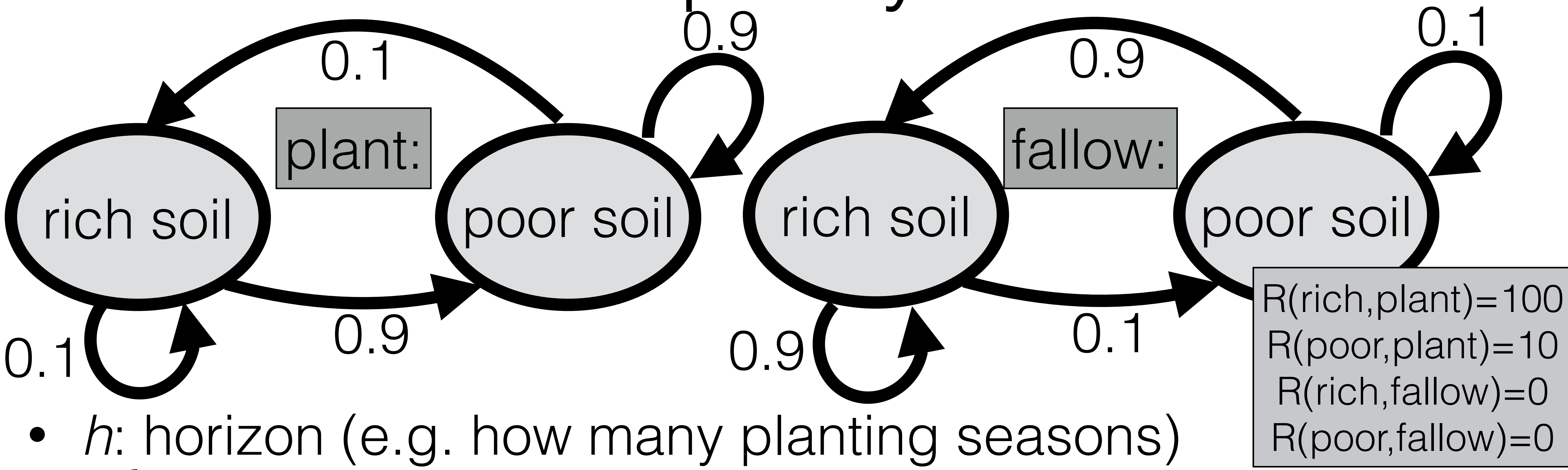
$$Q^0(s, a) = 0; Q^h(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \max_{a'} Q^{h-1}(s', a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

$$Q^2(\text{rich}, \text{plant}) = 100 + (0.1)(100) + (0.9)(10)$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi^*_h(s) = \arg\max_a Q^h(s, a)$

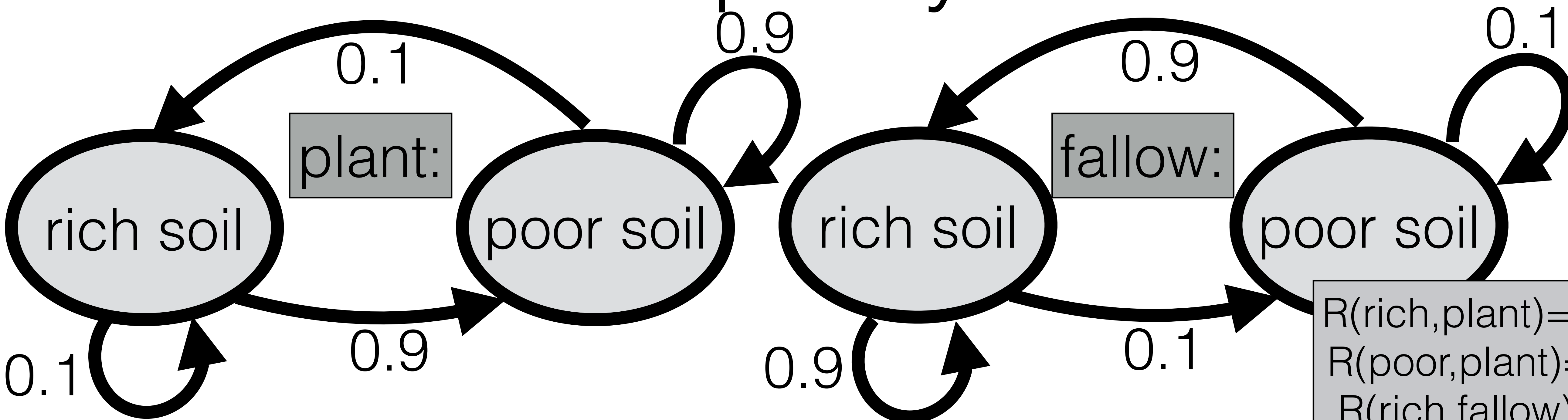
$$Q^0(s, a) = 0; Q^h(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \max_{a'} Q^{h-1}(s', a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

$$Q^2(\text{rich}, \text{plant}) = 100 + (0.1)(100) + (0.9)(10) = 119$$

What's best? Any *s*, $\pi^*_1(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s, a)$

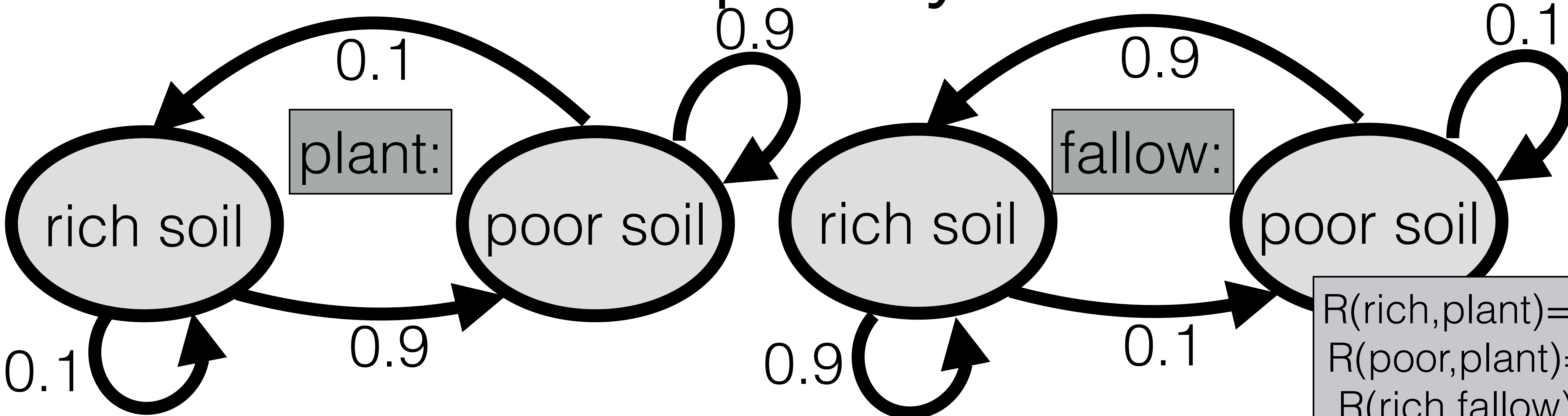
$$Q^0(s, a) = 0; Q^h(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \max_{a'} Q^{h-1}(s', a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

$$Q^2(\text{rich}, \text{plant}) = 119$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

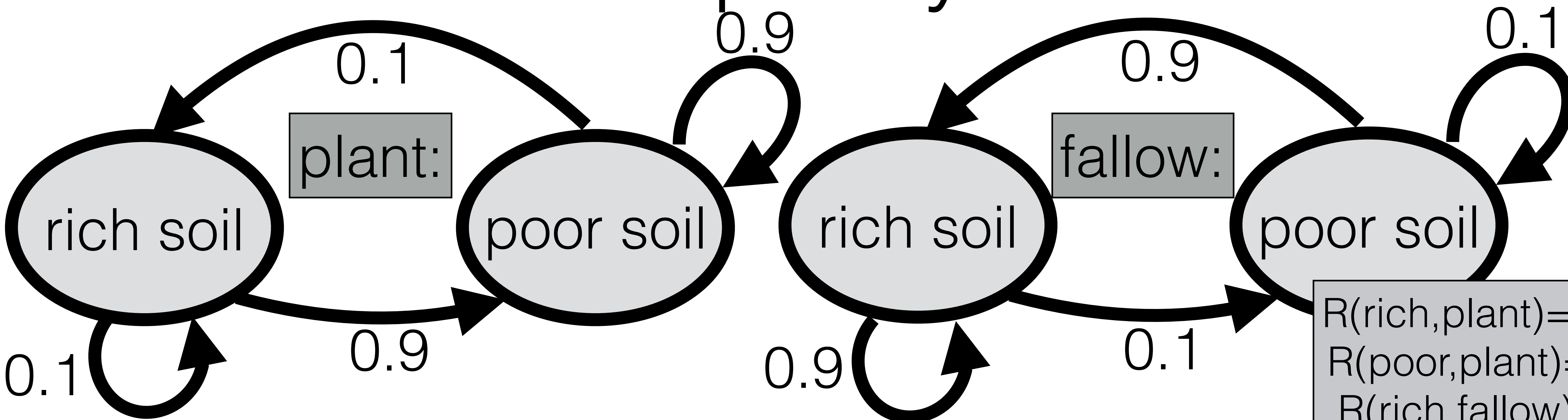
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$
$$Q^2(\text{rich},\text{plant}) = 119; Q^2(\text{rich},\text{fallow}) = 91;$$
$$Q^2(\text{poor},\text{plant}) = 29; Q^2(\text{poor},\text{fallow}) = 91$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

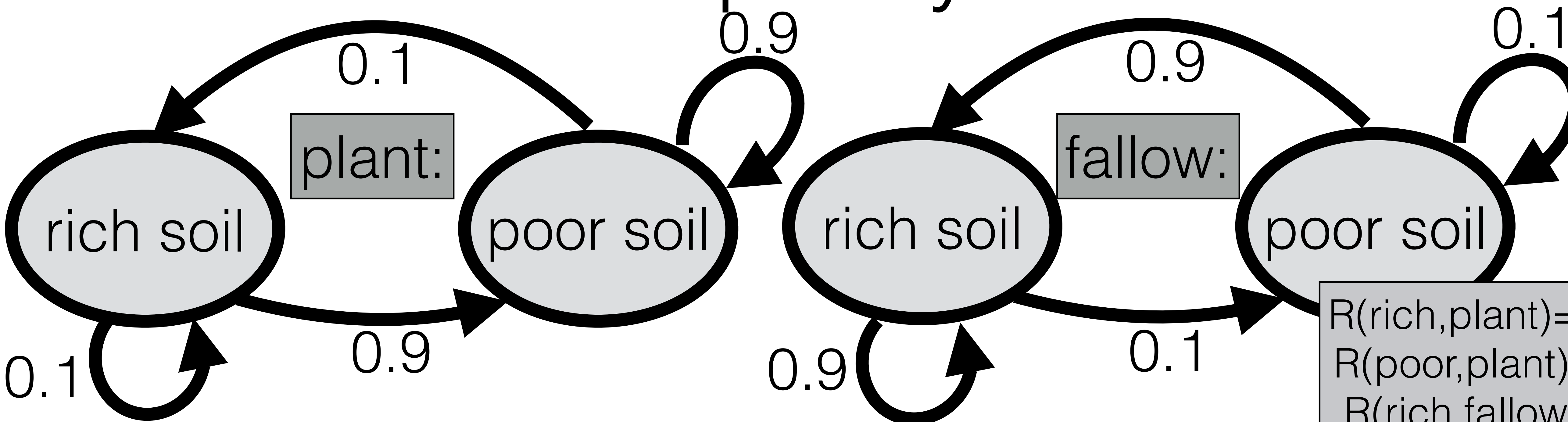
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$
$$Q^2(\text{rich},\text{plant}) = 119; Q^2(\text{rich},\text{fallow}) = 91;$$
$$Q^2(\text{poor},\text{plant}) = 29; Q^2(\text{poor},\text{fallow}) = 91$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$

π_2^*

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

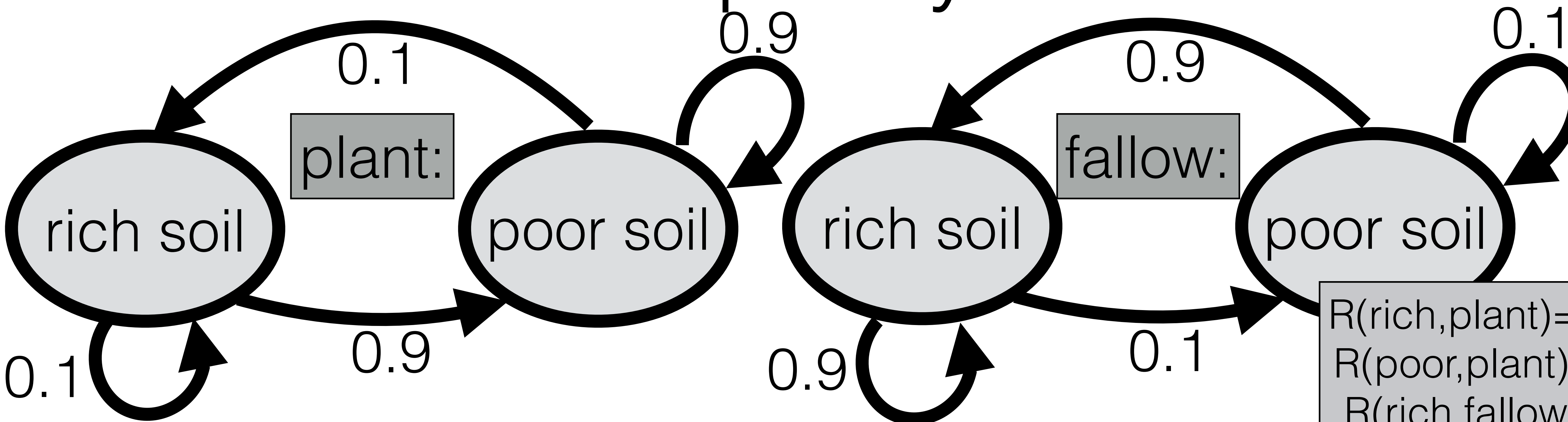
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$

$$Q^2(\text{rich}, \text{plant}) = 119; Q^2(\text{rich}, \text{fallow}) = 91;$$
$$Q^2(\text{poor}, \text{plant}) = 29; Q^2(\text{poor}, \text{fallow}) = 91$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$; $\pi_2^*(\text{rich})$ $\pi_2^*(\text{poor})$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

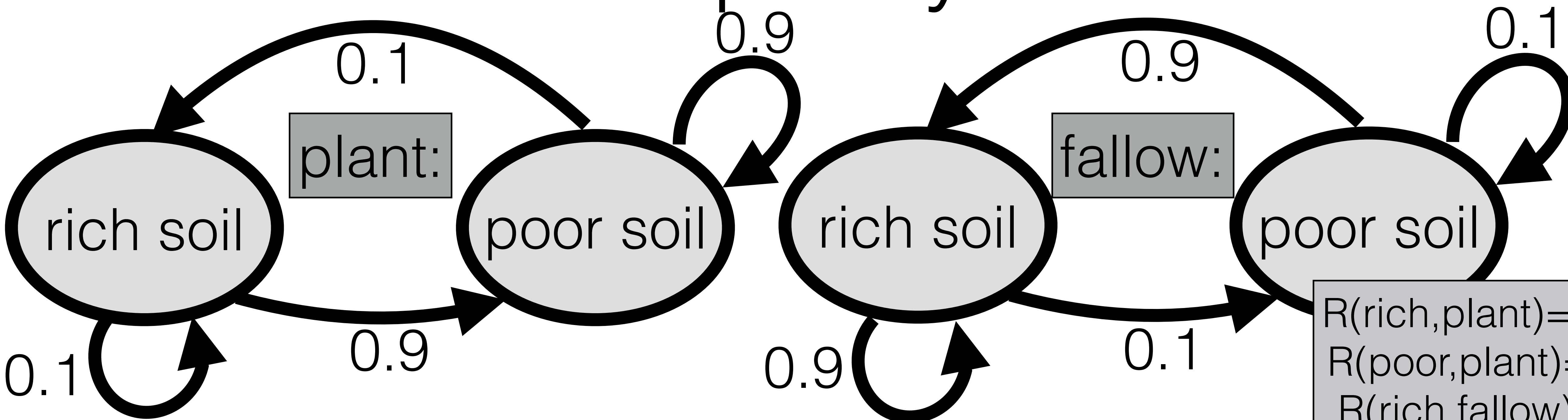
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$
$$Q^2(\text{rich},\text{plant}) = 119; Q^2(\text{rich},\text{fallow}) = 91;$$
$$Q^2(\text{poor},\text{plant}) = 29; Q^2(\text{poor},\text{fallow}) = 91$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}$; $\pi_2^*(\text{rich}) = \text{plant}, \pi_2^*(\text{poor}) = \text{fallow}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

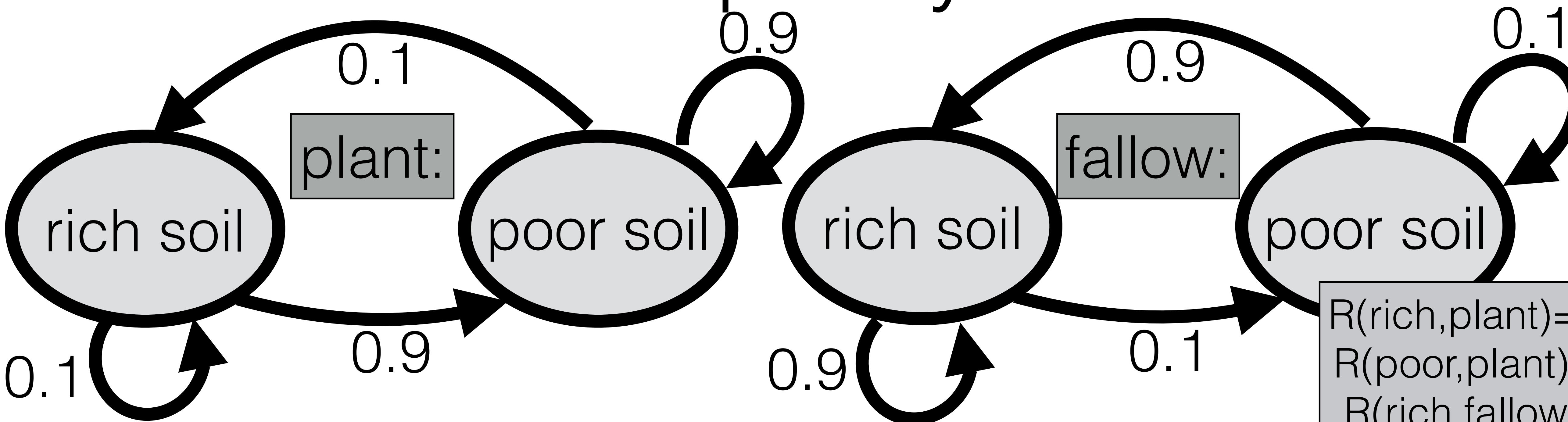
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$
$$Q^2(\text{rich}, \text{plant}) = 119; Q^2(\text{rich}, \text{fallow}) = 91;$$
$$Q^2(\text{poor}, \text{plant}) = 29; Q^2(\text{poor}, \text{fallow}) = 91$$

What's best? Any *s*, $\pi_1^*(s) = \text{plant}; \pi_2^*(\text{rich}) = \text{plant}, \pi_2^*(\text{poor}) = \text{fallow}$

What's the best policy? Finite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$

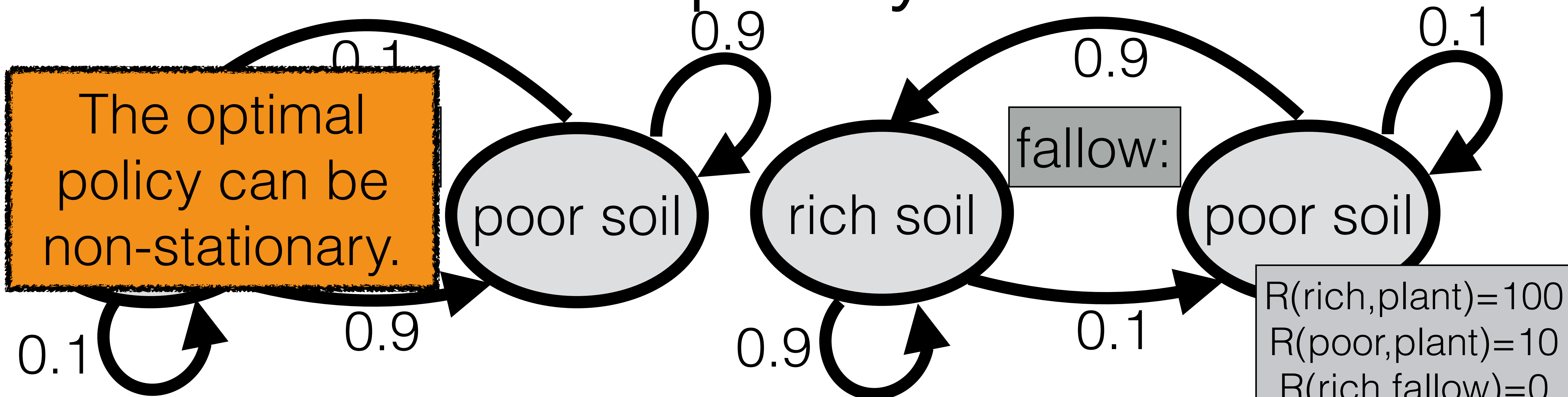
$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$
$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$

$Q^2(\text{rich}, \text{plant}) = 119; Q^2(\text{rich}, \text{fallow}) = 91;$
$Q^2(\text{poor}, \text{plant}) = 29; Q^2(\text{poor}, \text{fallow}) = 91$

"finite-horizon value iteration"

What's best? Any *s*, $\pi_1^*(s) = \text{plant}; \pi_2^*(\text{rich}) = \text{plant}, \pi_2^*(\text{poor}) = \text{fallow}$

What's the best policy? Finite horizon



- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

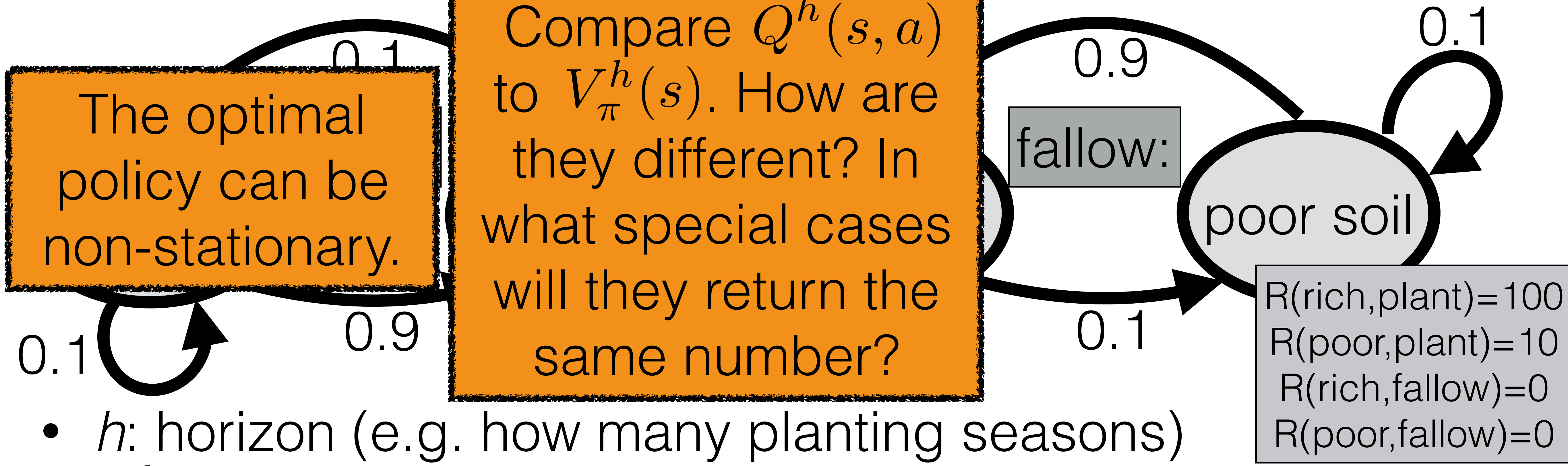
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich, plant}) = 100; Q^1(\text{rich, fallow}) = 0;$$
$$Q^1(\text{poor, plant}) = 10; Q^1(\text{poor, fallow}) = 0$$
$$Q^2(\text{rich, plant}) = 119; Q^2(\text{rich, fallow}) = 91;$$
$$Q^2(\text{poor, plant}) = 29; Q^2(\text{poor, fallow}) = 91$$

"finite-horizon value iteration"

What's best? Any *s*, $\pi_1^*(s) = \text{plant}; \pi_2^*(\text{rich}) = \text{plant}, \pi_2^*(\text{poor}) = \text{fallow}$

What's the best policy? Finite horizon



The optimal policy can be non-stationary.

Compare $Q^h(s,a)$ to $V_\pi^h(s)$. How are they different? In what special cases will they return the same number?

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

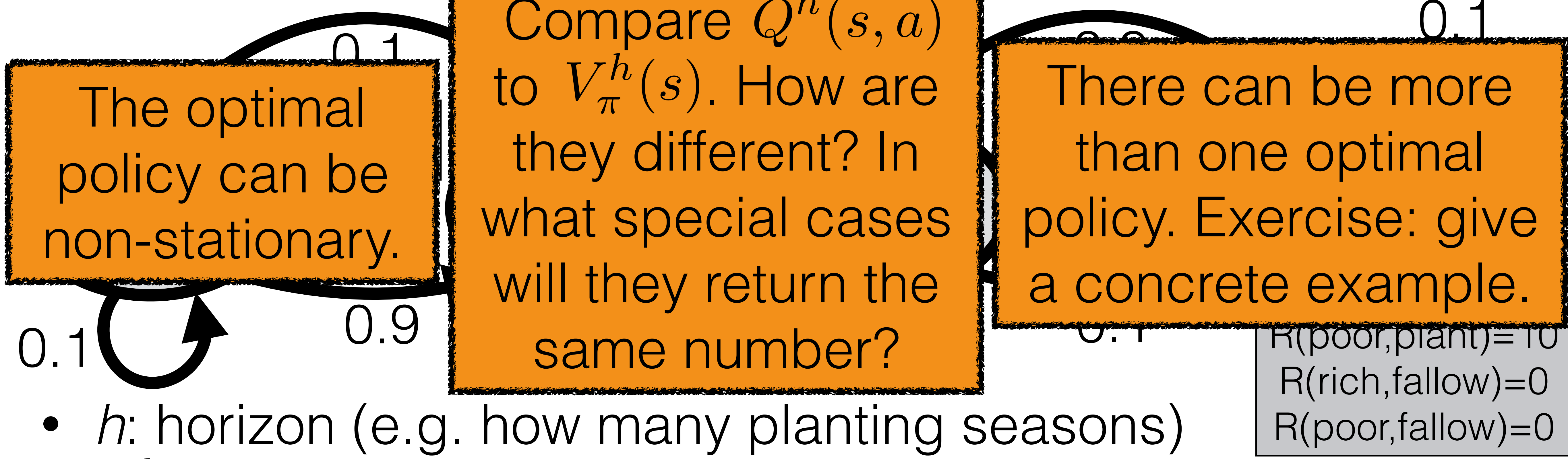
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$
$$Q^1(\text{poor}, \text{plant}) = 10; Q^1(\text{poor}, \text{fallow}) = 0$$
$$Q^2(\text{rich}, \text{plant}) = 119; Q^2(\text{rich}, \text{fallow}) = 91;$$
$$Q^2(\text{poor}, \text{plant}) = 29; Q^2(\text{poor}, \text{fallow}) = 91$$

"finite-horizon value iteration"

What's best? Any *s*, $\pi_1^*(s) = \text{plant}; \pi_2^*(\text{rich}) = \text{plant}, \pi_2^*(\text{poor}) = \text{fallow}$

What's the best policy? Finite horizon



The optimal policy can be non-stationary.

Compare $Q^h(s,a)$ to $V_\pi^h(s)$. How are they different? In what special cases will they return the same number?

There can be more than one optimal policy. Exercise: give a concrete example.

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

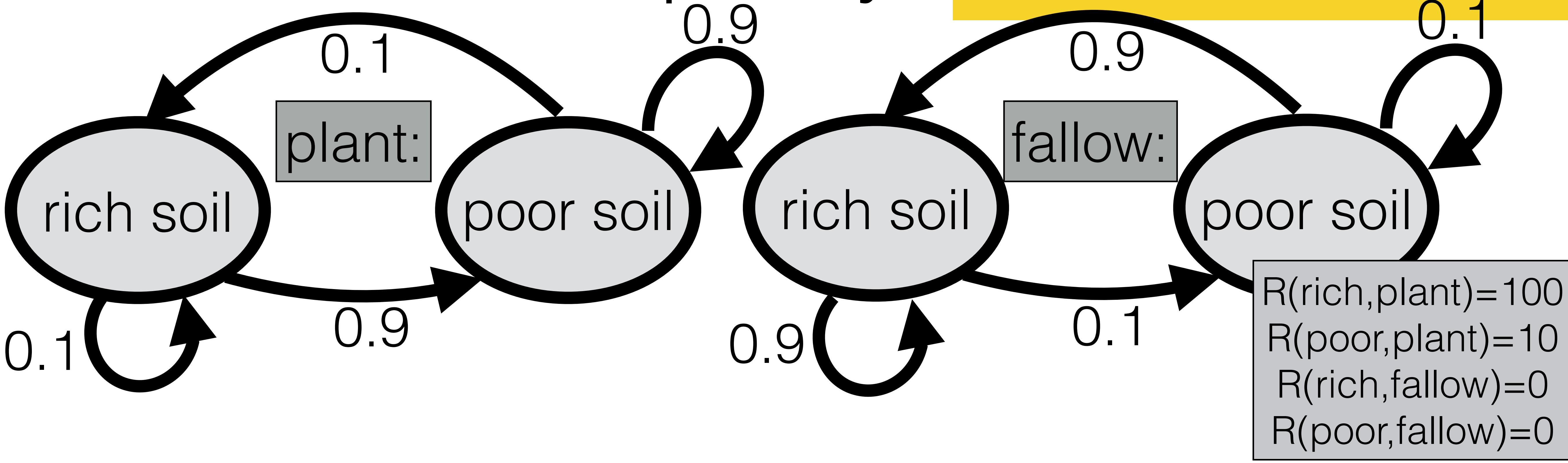
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$
$$Q^2(\text{rich},\text{plant}) = 119; Q^2(\text{rich},\text{fallow}) = 91;$$
$$Q^2(\text{poor},\text{plant}) = 29; Q^2(\text{poor},\text{fallow}) = 91$$

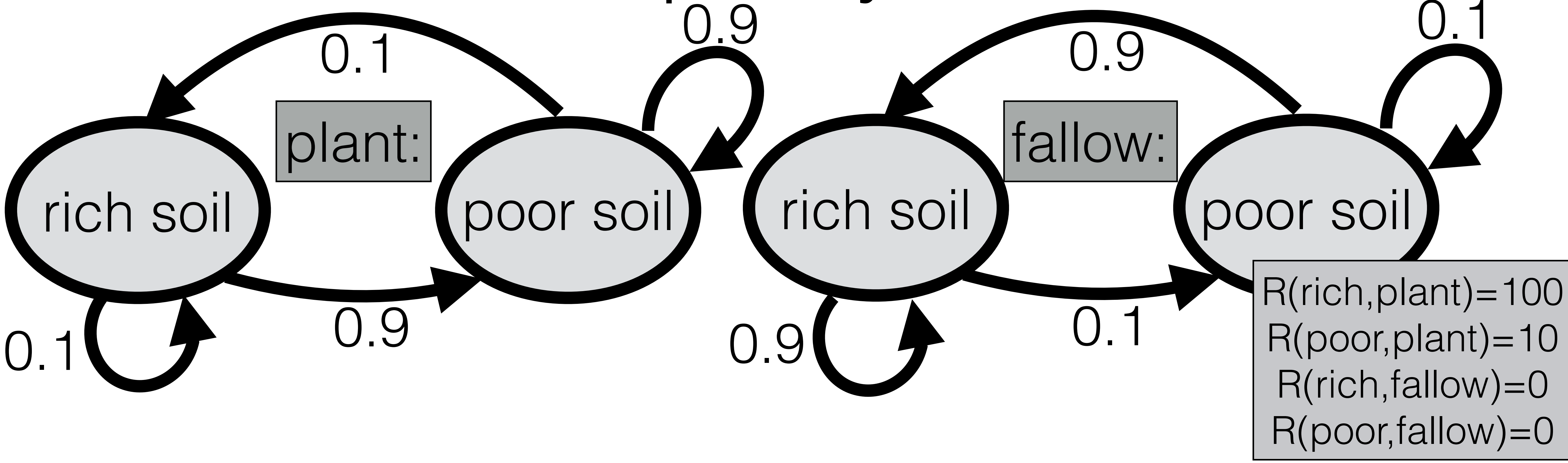
"finite-horizon value iteration"

What's best? Any *s*, $\pi_1^*(s) = \text{plant}; \pi_2^*(\text{rich}) = \text{plant}, \pi_2^*(\text{poor}) = \text{fallow}$

What's the best policy? Infinite horizon

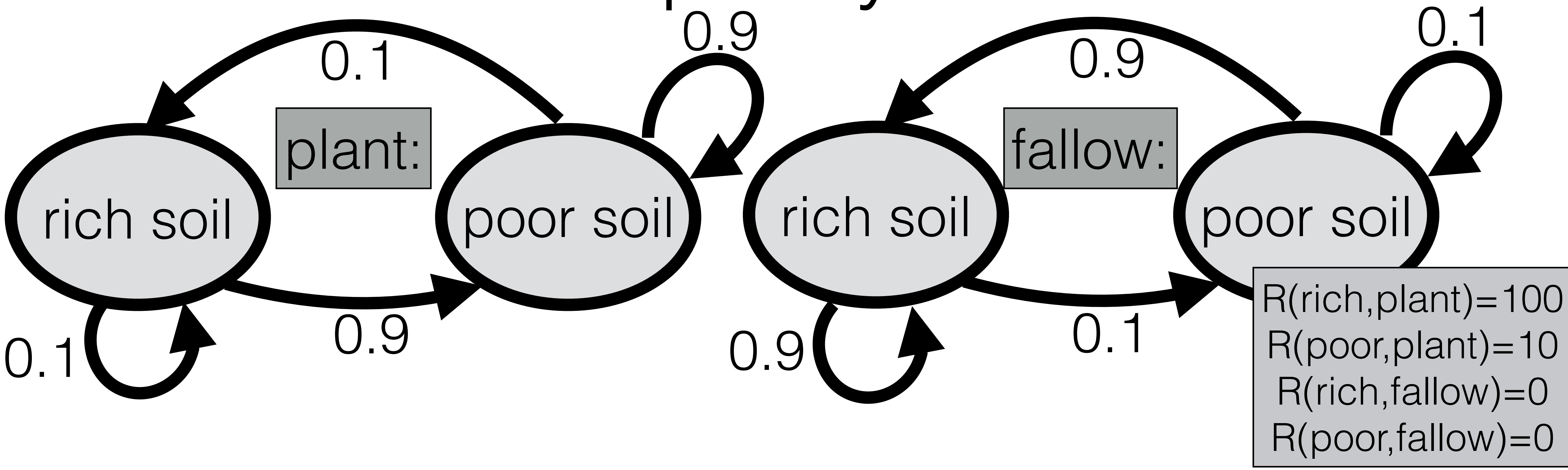


What’s the best policy? Infinite horizon



- What if I don’t stop farming? Is there any optimal policy?

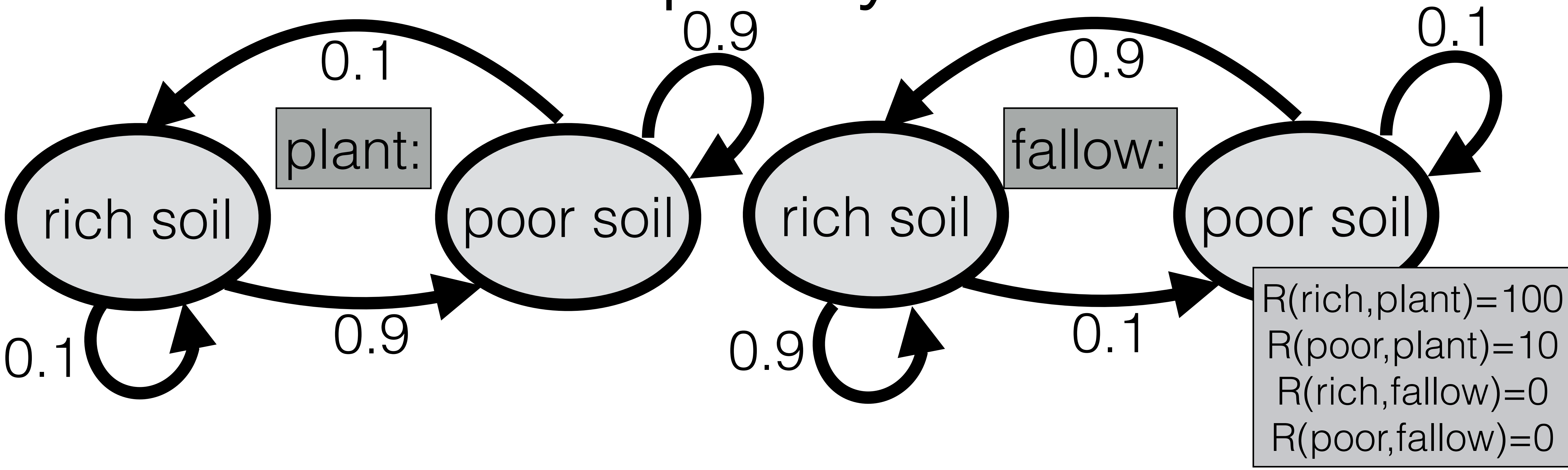
What's the best policy? Infinite horizon



- What if I don't stop farming? Is there any optimal policy?

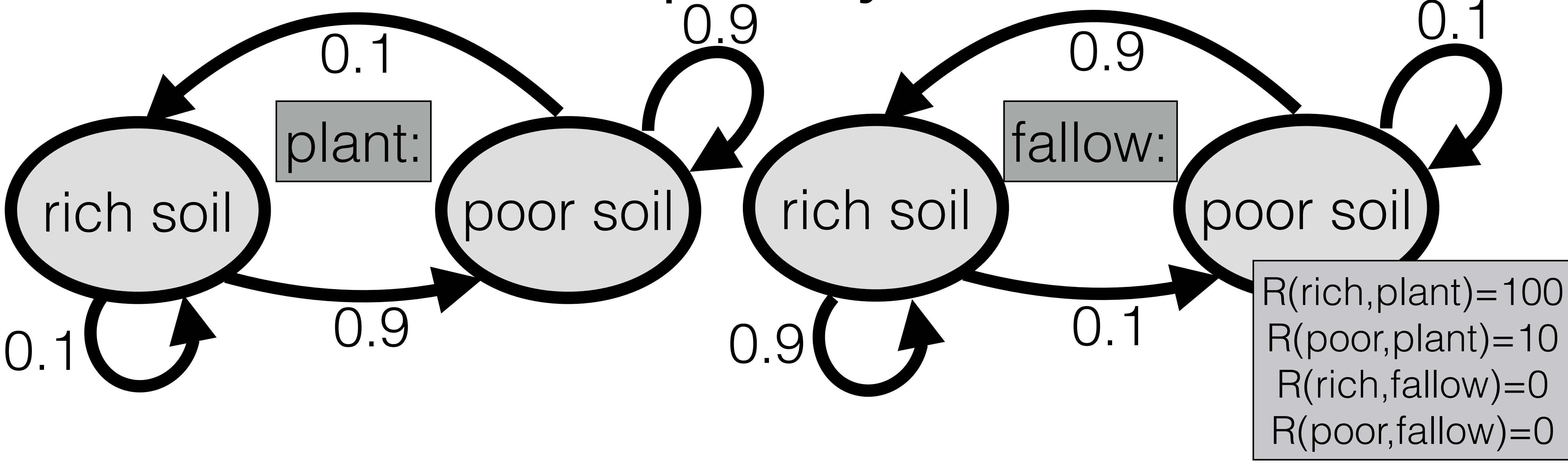
Recall farmer A and farmer B from last time

What's the best policy? Infinite horizon



- What if I don't stop farming? Is there any optimal policy?
- **Theorem**. There exists a (stationary) optimal policy π^*. I.e., for every policy π and for every state $s \in \mathcal{S}$, $V_{\pi^*}(s) \geq V_\pi(s)$

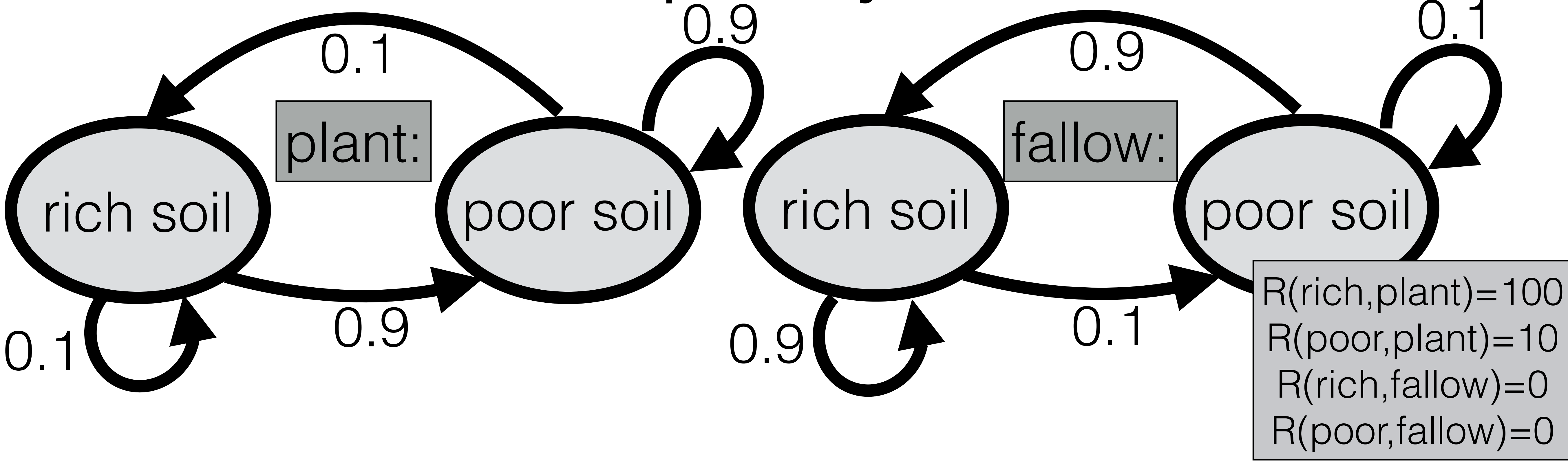
What's the best policy? Infinite horizon



- What if I don't stop farming? Is there any optimal policy?
- **Theorem**. There exists a (stationary) optimal policy π^*. I.e., for every policy π and for every state $s \in \mathcal{S}$, $V_{\pi^*}(s) \geq V_{\pi}(s)$

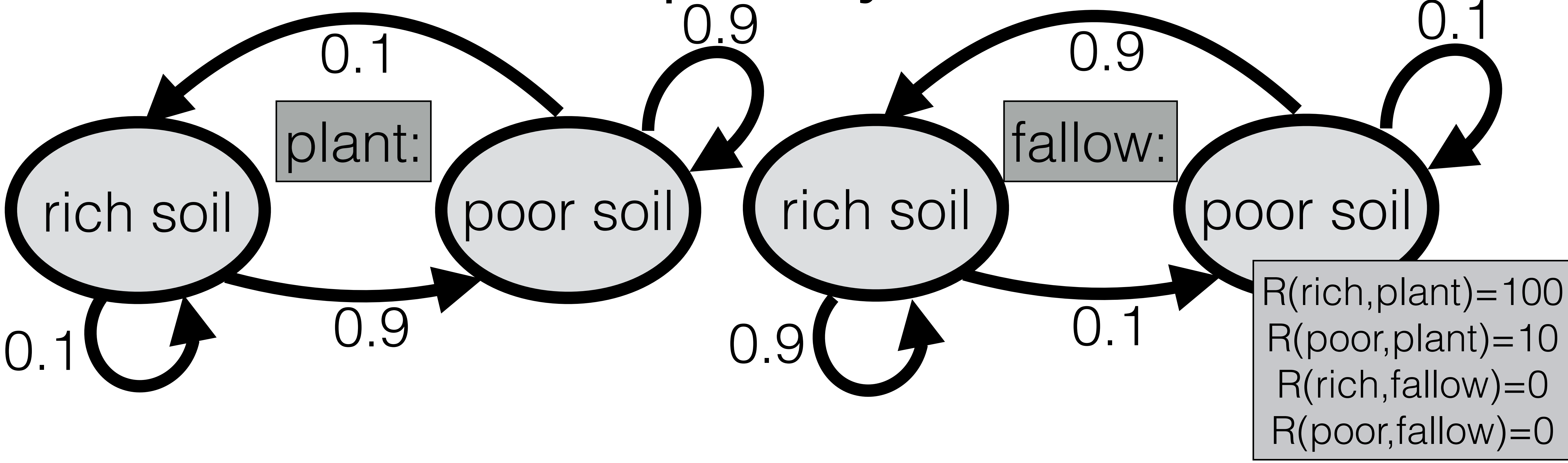
Two (or more) policies can have the same (best) value for all states and all be optimal

What's the best policy? Infinite horizon



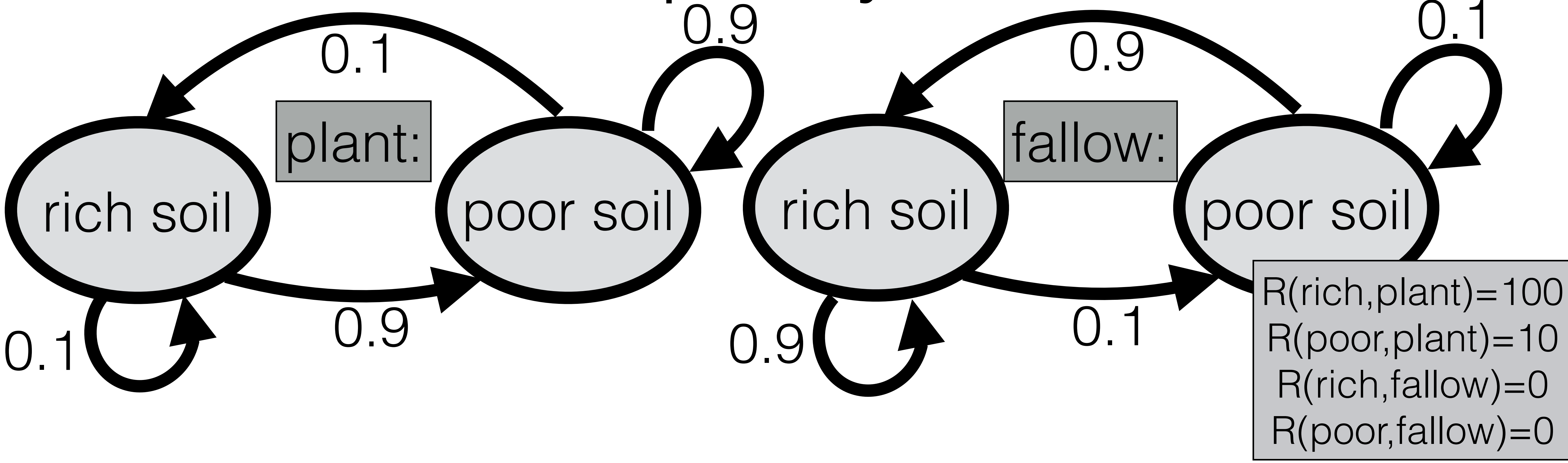
- What if I don't stop farming? Is there any optimal policy?
- **Theorem**. There exists a (stationary) optimal policy π^*. I.e., for every policy π and for every state $s \in \mathcal{S}$, $V_{\pi^*}(s) \geq V_{\pi}(s)$
- $Q^*(s, a)$: expected reward if we make best actions in future

What's the best policy? Infinite horizon



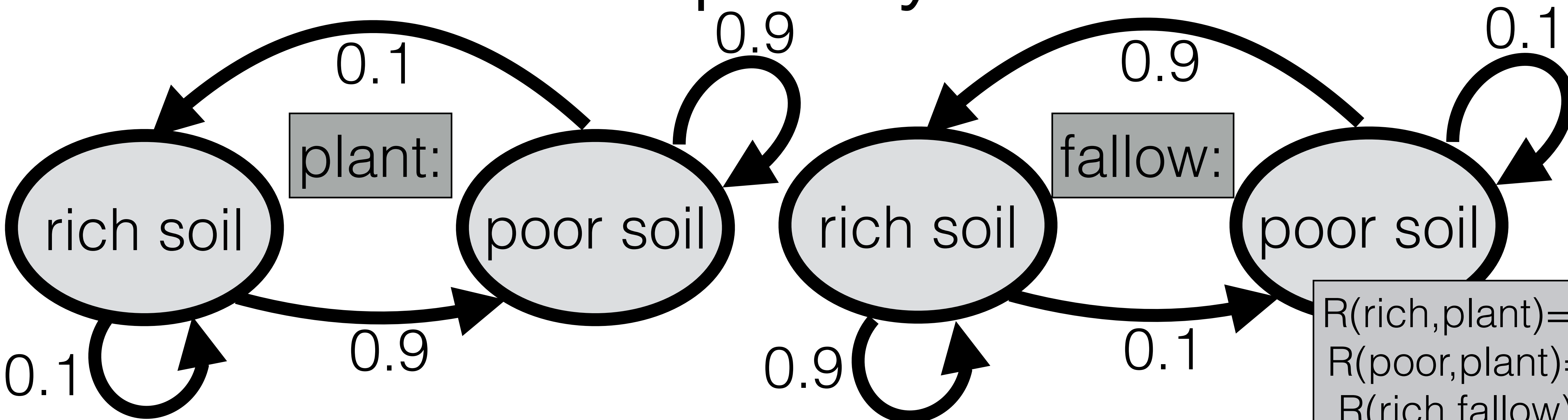
- What if I don't stop farming? Is there any optimal policy?
- **Theorem**. There exists a (stationary) optimal policy π^*. I.e., for every policy π and for every state $s \in \mathcal{S}$, $V_{\pi^*}(s) \geq V_\pi(s)$
- $Q^*(s,a)$: expected reward if we make best actions in future
  - If we knew $Q^*(s,a)$, then: $\pi^*(s) = \arg\max_a Q^*(s,a)$

What's the best policy? Infinite horizon



- What if I don't stop farming? Is there any optimal policy?
- **Theorem**. There exists a (stationary) optimal policy π^*. I.e., for every policy π and for every state $s \in \mathcal{S}$, $V_{\pi^*}(s) \geq V_\pi(s)$
- $Q^*(s, a)$: expected reward if we make best actions in future
  - If we knew $Q^*(s, a)$, then: $\pi^*(s) = \arg\max_a Q^*(s, a)$
- Note: $Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$

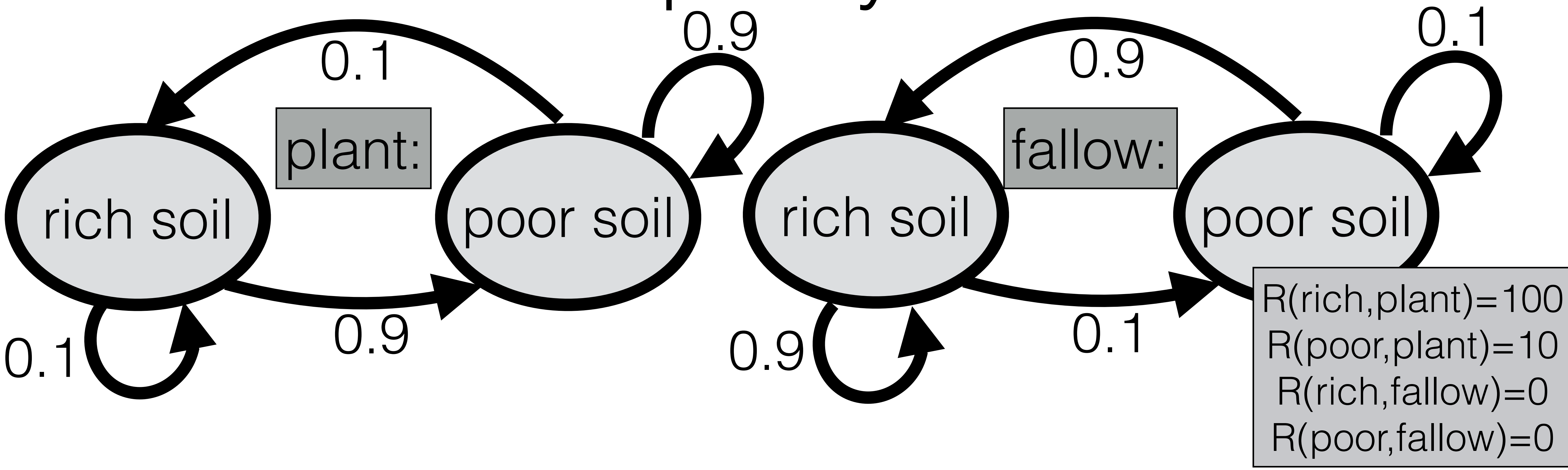
What's the best policy? Infinite horizon



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

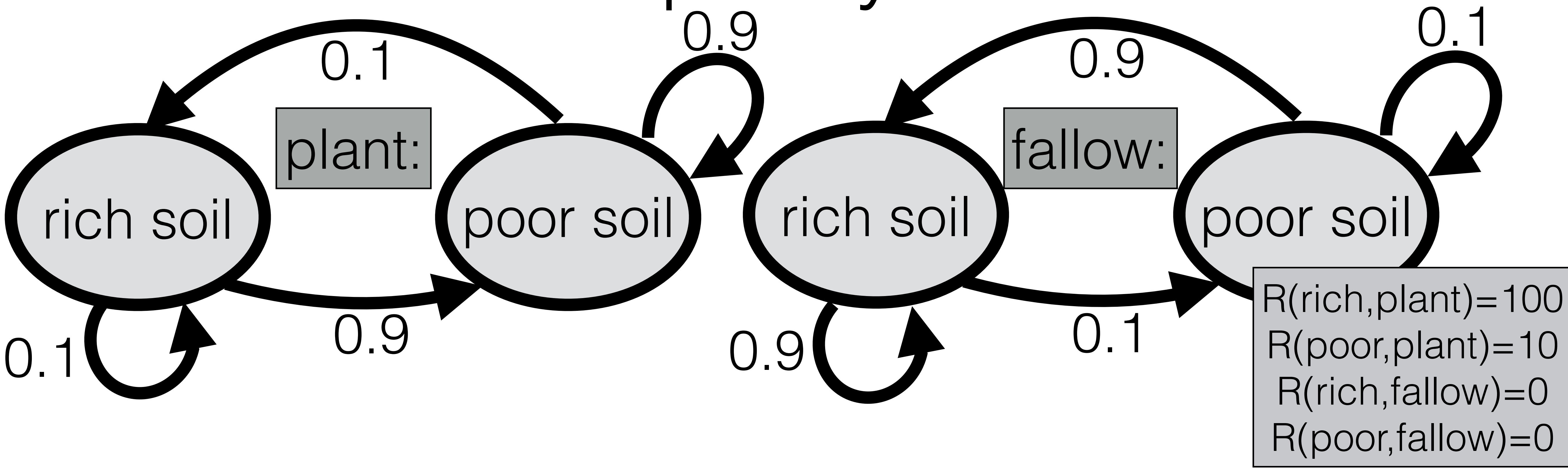
- What if I don't stop farming? Is there any optimal policy?
- **Theorem**. There exists a (stationary) optimal policy π^*. I.e., for every policy π and for every state $s \in \mathcal{S}$, $V_{\pi^*}(s) \geq V_\pi(s)$
- $Q^*(s,a)$: expected reward if we make best actions in future
  - If we knew $Q^*(s,a)$, then: $\pi^*(s) = \arg\max_a Q^*(s,a)$
- Note: $Q^*(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^*(s',a')$

What's the best policy? Infinite horizon



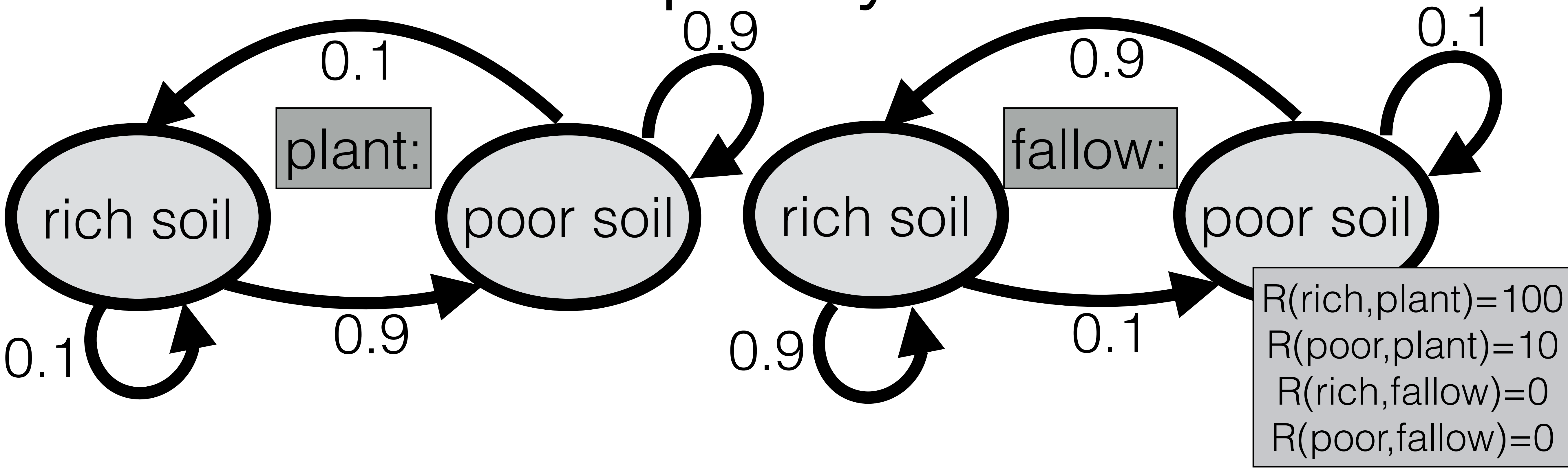
- What if I don't stop farming? Is there any optimal policy?
- **Theorem**. There exists a (stationary) optimal policy π^*. I.e., for every policy π and for every state $s \in \mathcal{S}$, $V_{\pi^*}(s) \geq V_\pi(s)$
- $Q^*(s,a)$: expected reward if we make best actions in future
  - If we knew $Q^*(s,a)$, then: $\pi^*(s) = \arg\max_a Q^*(s,a)$
- Note: $Q^*(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^*(s',a')$

What's the best policy? Infinite horizon



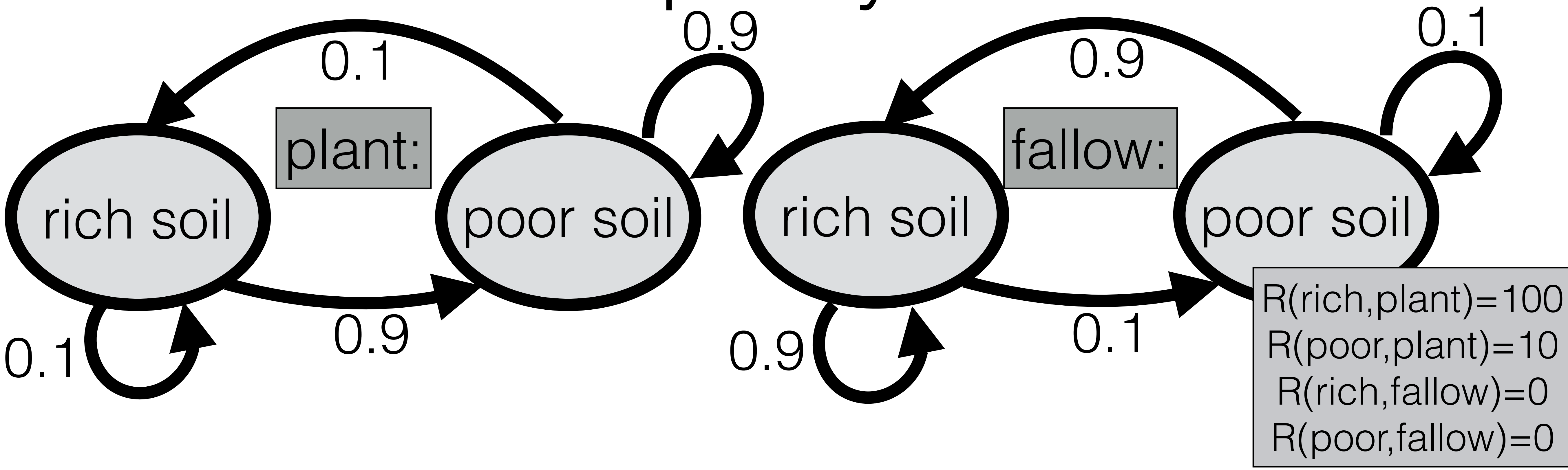
- What if I don't stop farming? Is there any optimal policy?
- **Theorem**. There exists a (stationary) optimal policy π^*. I.e., for every policy π and for every state $s \in \mathcal{S}$, $V_{\pi^*}(s) \geq V_\pi(s)$
- $Q^*(s,a)$: expected reward if we make best actions in future
  - If we knew $Q^*(s,a)$, then: $\pi^*(s) = \arg\max_a Q^*(s,a)$
- Note: $Q^*(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^*(s',a')$

What's the best policy? Infinite horizon



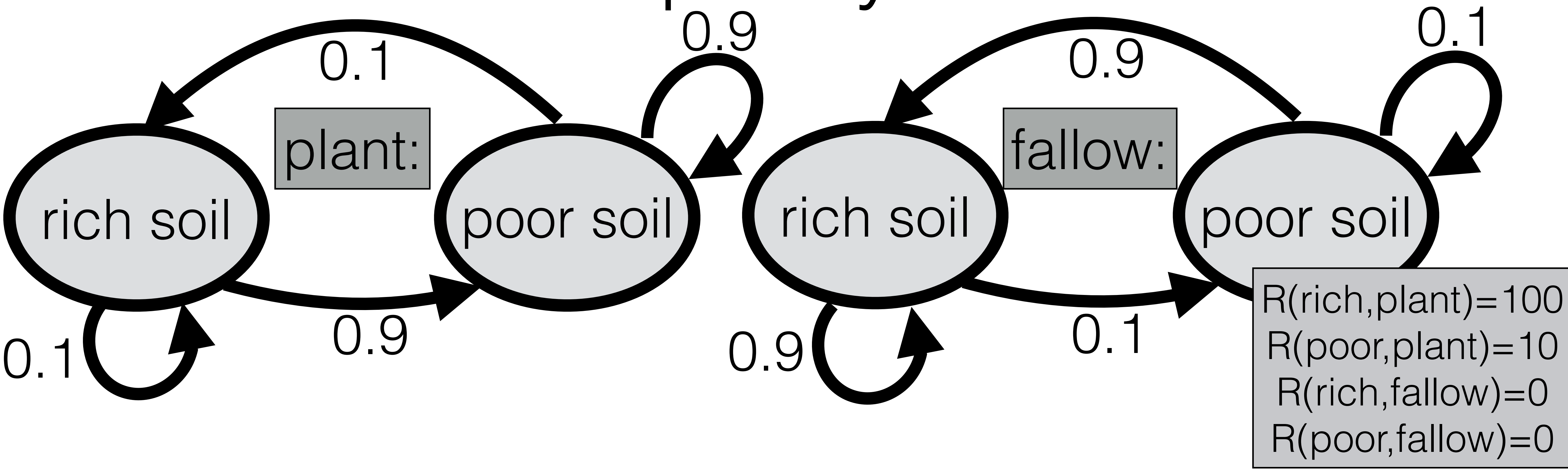
- What if I don't stop farming? Is there any optimal policy?
- **Theorem**. There exists a (stationary) optimal policy π^*. I.e., for every policy π and for every state $s \in \mathcal{S}$, $V_{\pi^*}(s) \geq V_\pi(s)$
- $Q^*(s,a)$: expected reward if we make best actions in future
  - If we knew $Q^*(s,a)$, then: $\pi^*(s) = \arg\max_a Q^*(s,a)$
- Note: $Q^*(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^*(s',a')$

What's the best policy? Infinite horizon



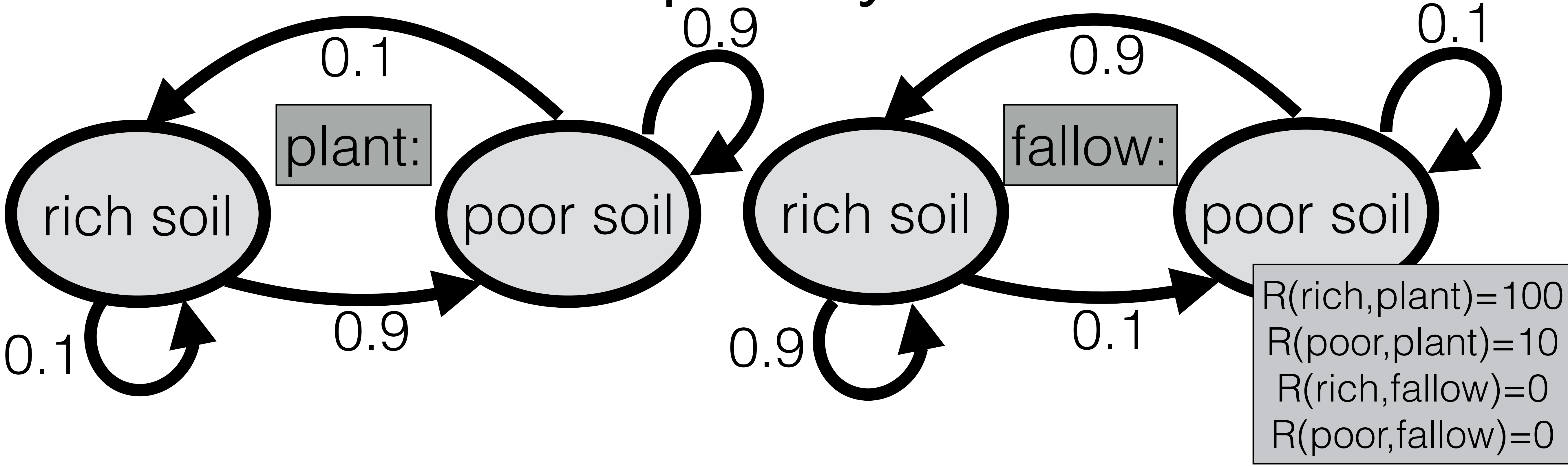
- What if I don't stop farming? Is there any optimal policy?
- **Theorem**. There exists a (stationary) optimal policy π^*. I.e., for every policy π and for every state $s \in \mathcal{S}$, $V_{\pi^*}(s) \geq V_\pi(s)$
- $Q^*(s, a)$: expected reward if we make best actions in future
  - If we knew $Q^*(s, a)$, then: $\pi^*(s) = \arg\max_a Q^*(s, a)$
- Note: $Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$

What's the best policy? Infinite horizon



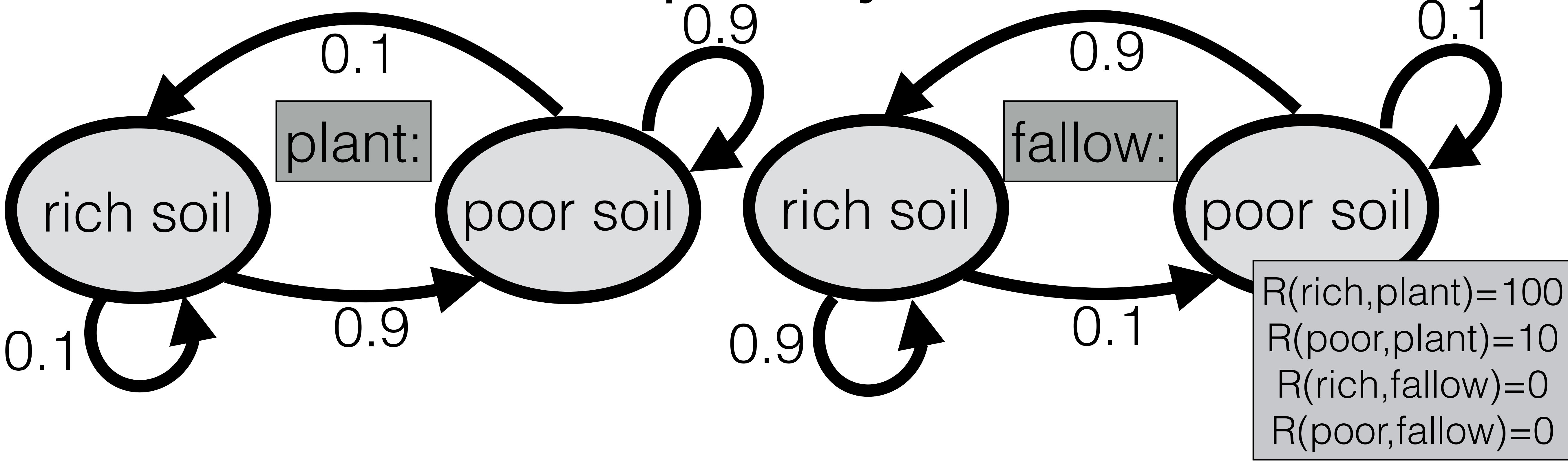
- What if I don't stop farming? Is there any optimal policy?
- **Theorem**. There exists a (stationary) optimal policy π^*. I.e., for every policy π and for every state $s \in \mathcal{S}$, $V_{\pi^*}(s) \geq V_\pi(s)$
- $Q^*(s,a)$: expected reward if we make best actions in future
  - If we knew $Q^*(s,a)$, then: $\pi^*(s) = \arg\max_a Q^*(s,a)$
- Note: $Q^*(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^*(s',a')$

What's the best policy? Infinite horizon



- What if I don't stop farming? Is there any optimal policy?
- **Theorem**. There exists a (stationary) optimal policy π^*. I.e., for every policy π and for every state $s \in \mathcal{S}$, $V_{\pi^*}(s) \geq V_\pi(s)$
- $Q^*(s,a)$: expected reward if we make best actions in future
  - If we knew $Q^*(s,a)$, then: $\pi^*(s) = \arg\max_a Q^*(s,a)$
- Note: $Q^*(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^*(s',a')$
  - Not linear in $Q^*(s,a)$, so not as easy to solve as $V_\pi(s)$

What's the best policy? Infinite horizon



- What if I don't stop farming? Is there any optimal policy?
- **Theorem**. There exists a (stationary) optimal policy π^*. I.e., for every policy π and for every state $s \in \mathcal{S}$, $V_{\pi^*}(s) \geq V_\pi(s)$
- $Q^*(s,a)$: expected reward if we make best actions in future
  - If we knew $Q^*(s,a)$, then: $\pi^*(s) = \arg\max_a Q^*(s,a)$
- Note: $Q^*(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^*(s',a')$
  - Not linear in $Q^*(s,a)$, so not as easy to solve as $V_\pi(s)$

There can be more than one optimal policy.
Exercise: give an infinite-horizon example.

Infinite-Horizon Value Iteration

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s, a) = 0$$

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$
$$Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$
$$Q^h(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$
$$Q^h(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

- A similar flavor for the infinite-horizon case:

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$
$$Q^h(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

- A similar flavor for the infinite-horizon case:

`Infinite-Horizon-Value-Iteration` $(\mathcal{S}, \mathcal{A}, T, R, \gamma, \epsilon)$

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$
$$Q^h(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

- A similar flavor for the infinite-horizon case:

`Infinite-Horizon-Value-Iteration` $(\mathcal{S}, \mathcal{A}, T, R, \gamma, \epsilon)$

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$
$$Q^h(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

- A similar flavor for the infinite-horizon case:

`Infinite-Horizon-Value-Iteration` $(\mathcal{S}, \mathcal{A}, T, R, \gamma, \epsilon)$

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$
$$Q^h(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

- A similar flavor for the infinite-horizon case:

`Infinite-Horizon-Value-Iteration` $(\mathcal{S}, \mathcal{A}, T, R, \gamma, \epsilon)$

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$
$$Q^h(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

- A similar flavor for the infinite-horizon case:

`Infinite-Horizon-Value-Iteration` $(\mathcal{S}, \mathcal{A}, T, R, \gamma, \epsilon)$

  `for` `each state` $s \in \mathcal{S}$ `and each action` $a \in \mathcal{A}$

    `Initialize` $Q_{\text{old}}(s,a) = 0$

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$
$$Q^h(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

- A similar flavor for the infinite-horizon case:

Infinite-Horizon-Value-Iteration $(\mathcal{S}, \mathcal{A}, T, R, \gamma, \epsilon)$

for each state $s \in \mathcal{S}$ and each action $a \in \mathcal{A}$

Initialize $Q_{\text{old}}(s,a) = 0$

while True

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$
$$Q^h(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

- A similar flavor for the infinite-horizon case:

```
Infinite-Horizon-Value-Iteration(𝒮, 𝒜, T, R, γ, ε)
  for each state s ∈ 𝒮 and each action a ∈ 𝒜
    Initialize Q_old(s, a) = 0
  while True
    for each state s ∈ 𝒮 and each action a ∈ 𝒜
```

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$
$$Q^h(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

- A similar flavor for the infinite-horizon case:

`Infinite-Horizon-Value-Iteration` $(\mathcal{S}, \mathcal{A}, T, R, \gamma, \epsilon)$

  for each state $s \in \mathcal{S}$ and each action $a \in \mathcal{A}$

    Initialize $Q_{\text{old}}(s,a) = 0$

  while True

    for each state $s \in \mathcal{S}$ and each action $a \in \mathcal{A}$

      $Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$
$$Q^h(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

- A similar flavor for the infinite-horizon case:

`Infinite-Horizon-Value-Iteration` $(\mathcal{S}, \mathcal{A}, T, R, \gamma, \epsilon)$

  for each state $s \in \mathcal{S}$ and each action $a \in \mathcal{A}$

    Initialize $Q_{\text{old}}(s,a) = 0$

  while True

    for each state $s \in \mathcal{S}$ and each action $a \in \mathcal{A}$

      $Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$

    if $\max_{s,a} |Q_{\text{old}}(s,a) - Q_{\text{new}}(s,a)| < \epsilon$

      return Q_{new}

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$
$$Q^h(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

- A similar flavor for the infinite-horizon case:

Infinite-Horizon-Value-Iteration $(\mathcal{S}, \mathcal{A}, T, R, \gamma, \epsilon)$

for each state $s \in \mathcal{S}$ and each action $a \in \mathcal{A}$

  Initialize $Q_{\text{old}}(s,a) = 0$

while True

  for each state $s \in \mathcal{S}$ and each action $a \in \mathcal{A}$

    $Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$

  if $\max_{s,a} |Q_{\text{old}}(s,a) - Q_{\text{new}}(s,a)| < \epsilon$

    return Q_{new}

  $Q_{\text{old}} = Q_{\text{new}}$

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

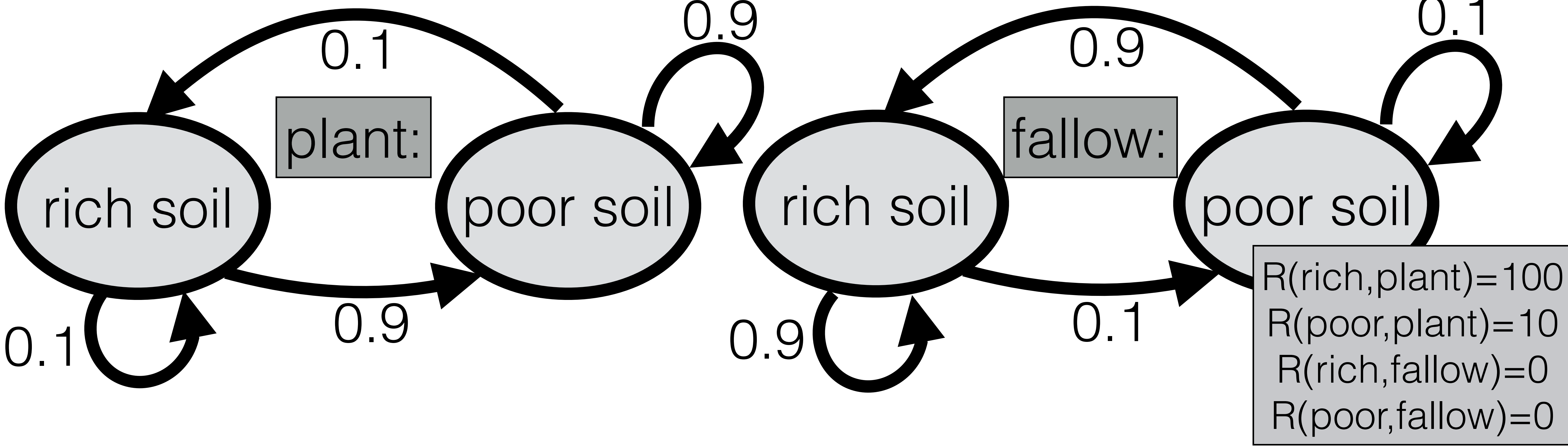
$$Q^0(s,a) = 0$$
$$Q^1(s,a) = R(s,a)$$
$$Q^h(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

- A similar flavor for the infinite-horizon case:

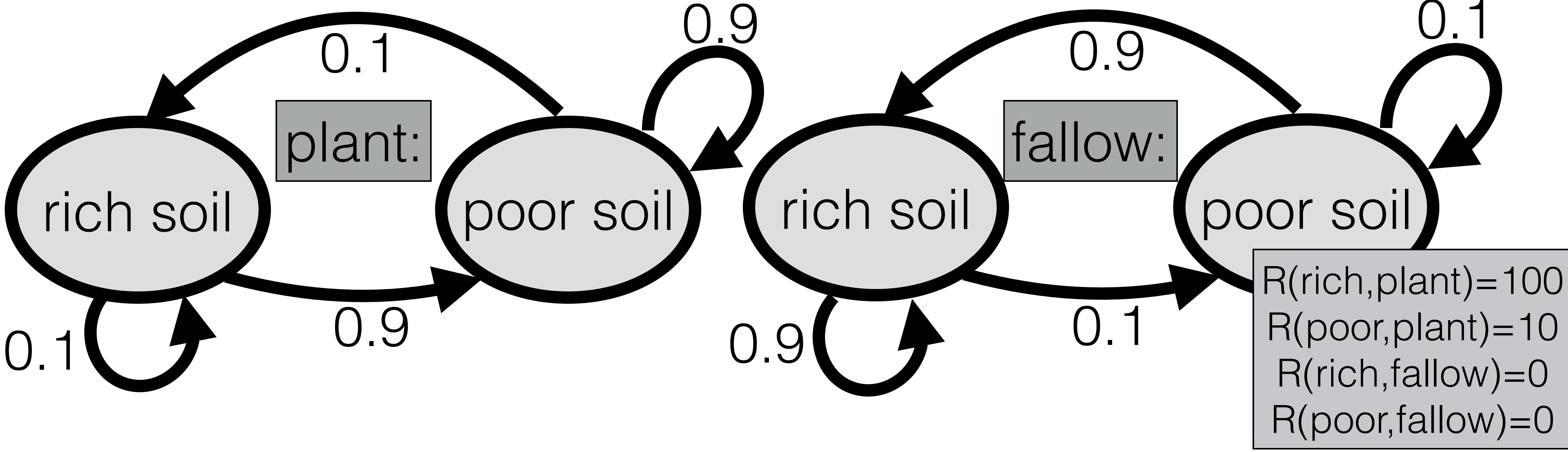
Infinite-Horizon-Value-Iteration $(\mathcal{S}, \mathcal{A}, T, R, \gamma, \epsilon)$
  for each state $s \in \mathcal{S}$ and each action $a \in \mathcal{A}$
    Initialize $Q_{\text{old}}(s,a) = 0$
  while True
    for each state $s \in \mathcal{S}$ and each action $a \in \mathcal{A}$
      $Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$
    if $\max_{s,a} |Q_{\text{old}}(s,a) - Q_{\text{new}}(s,a)| < \epsilon$
      return Q_{new}
    $Q_{\text{old}} = Q_{\text{new}}$

In real code, always cap the # of iterations

But what do we know at the start?

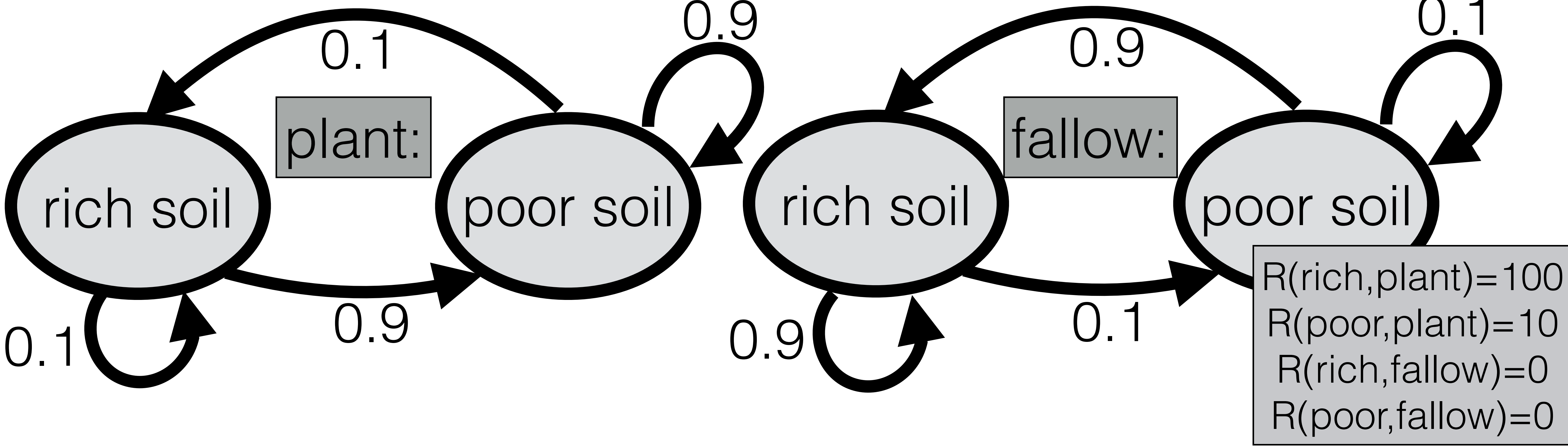


But what do we know at the start?



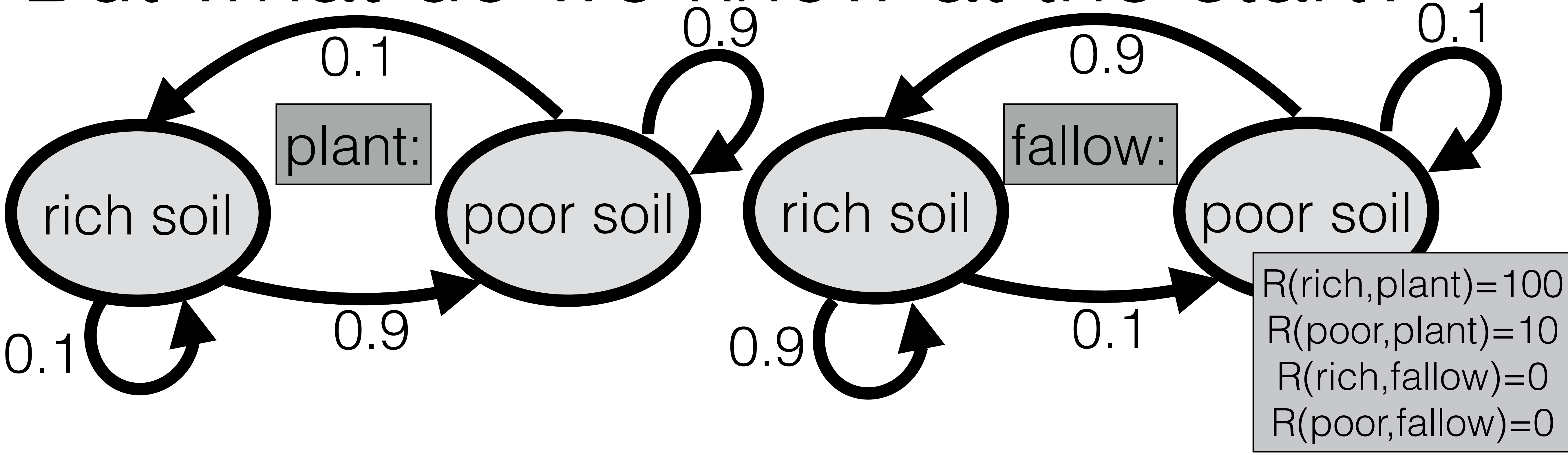
- General goal: Make actions to maximize expected reward.

But what do we know at the start?



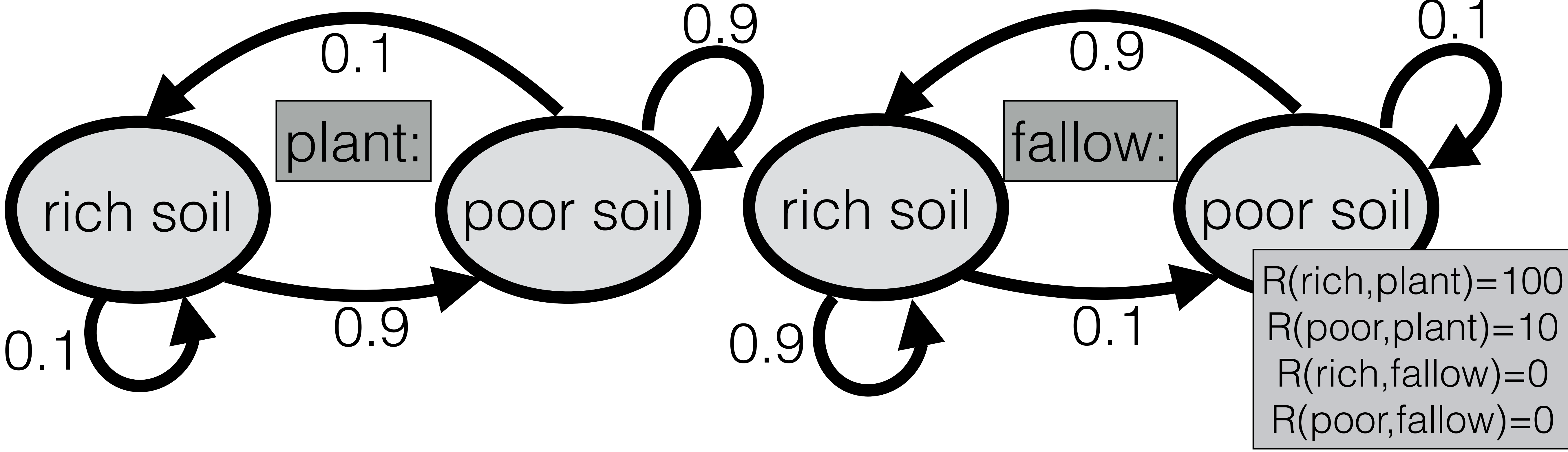
- General goal: Make actions to maximize expected reward.
- Up to this point: Assume we know full Markov decision process (MDP).

But what do we know at the start?



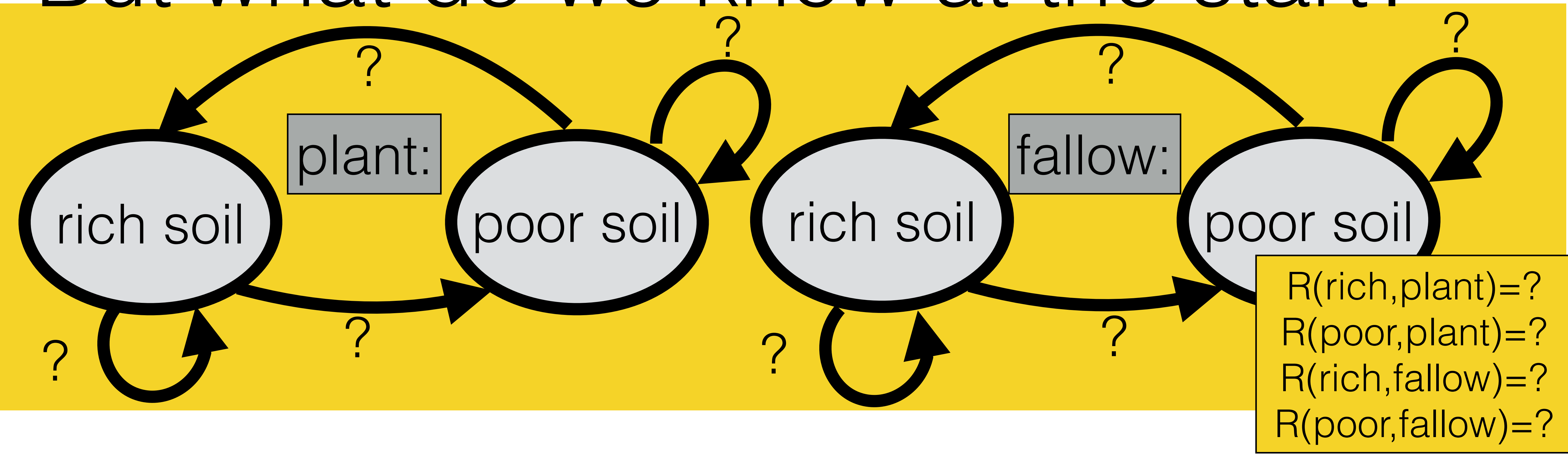
- General goal: Make actions to maximize expected reward.
- Up to this point: Assume we know full Markov decision process (MDP).
  - We figure out best policy and use it from the start.

But what do we know at the start?



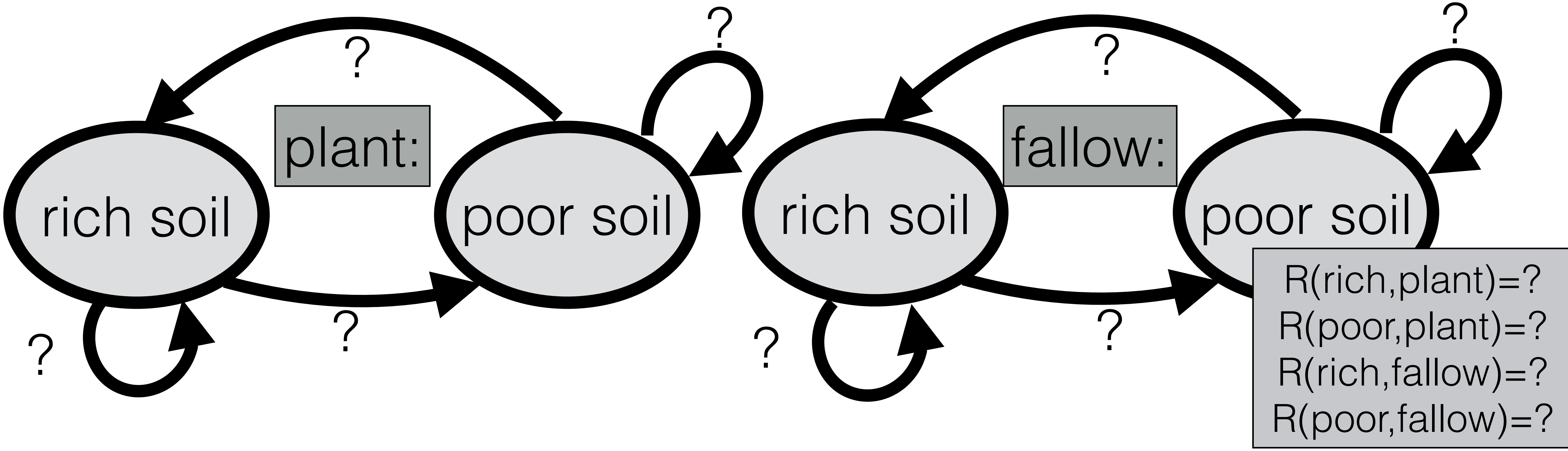
- General goal: Make actions to maximize expected reward.
- Up to this point: Assume we know full Markov decision process (MDP).
  - We figure out best policy and use it from the start.
- But we often *don't* know the transition model *T* or reward function *R* before we start.

But what do we know at the start?



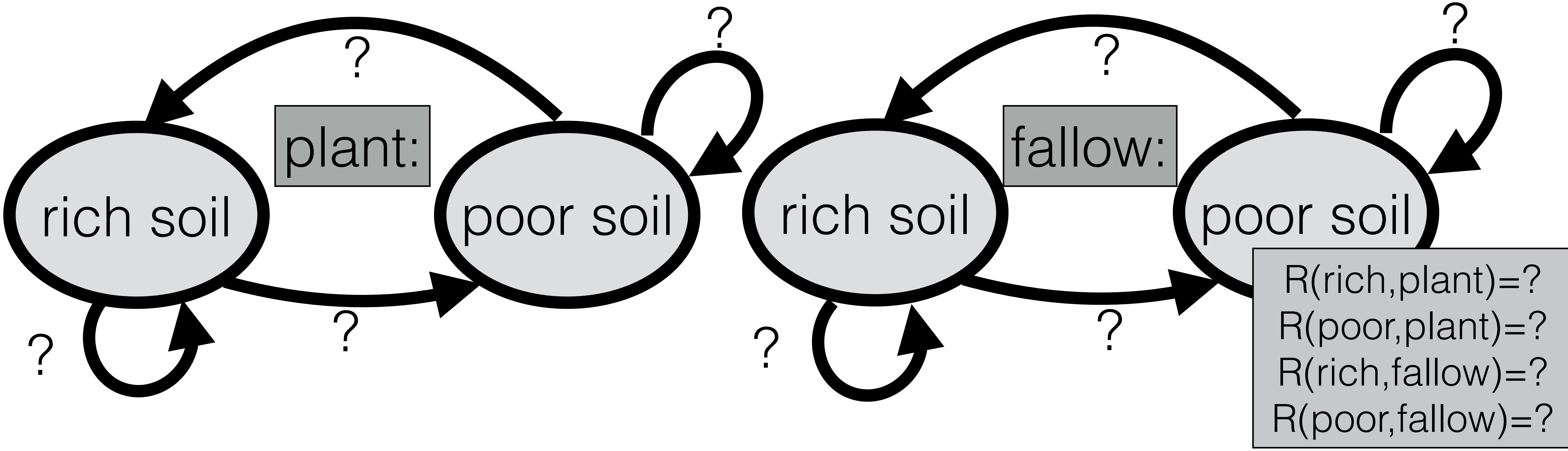
- General goal: Make actions to maximize expected reward.
- Up to this point: Assume we know full Markov decision process (MDP).
  - We figure out best policy and use it from the start.
- But we often *don't* know the transition model *T* or reward function *R* before we start.

But what do we know at the start?



- General goal: Make actions to maximize expected reward.
- Up to this point: Assume we know full Markov decision process (MDP).
  - We figure out best policy and use it from the start.
- But we often *don't* know the transition model *T* or reward function *R* before we start.
- Next: Assume we do know the states, actions, and discount. But we don't know *T* or *R*.

But what do we know at the start?



- General goal: Make actions to maximize expected reward.
- Up to this point: Assume we know full Markov decision process (MDP).
  - We figure out best policy and use it from the start.
- But we often *don't* know the transition model *T* or reward function *R* before we start.
- Next: Assume we do know the states, actions, and discount. But we don't know *T* or *R*.
  - Find a sequence of actions to maximize expected reward.