6.036/6.862: Introduction to Machine Learning

Lecture: starts Tuesdays 9:35am (Boston time zone)

Course website: introml.odl.mit.edu

Who's talking? Prof. Tamara Broderick

vote.mit.edu

Questions? discourse.odl.mit.edu ("Lecture 10" category)

Materials: Will all be available at course website

Last Time

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

II. Choosing "best" actions

Today's Plan

I. Don't know reward function or transition function in advance

II. How to choose "best" actions

Recall

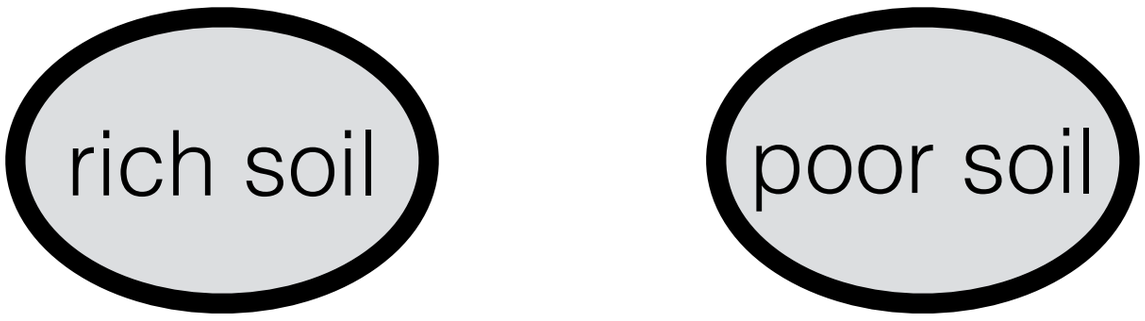
Recall

- Markov decision process

Recall

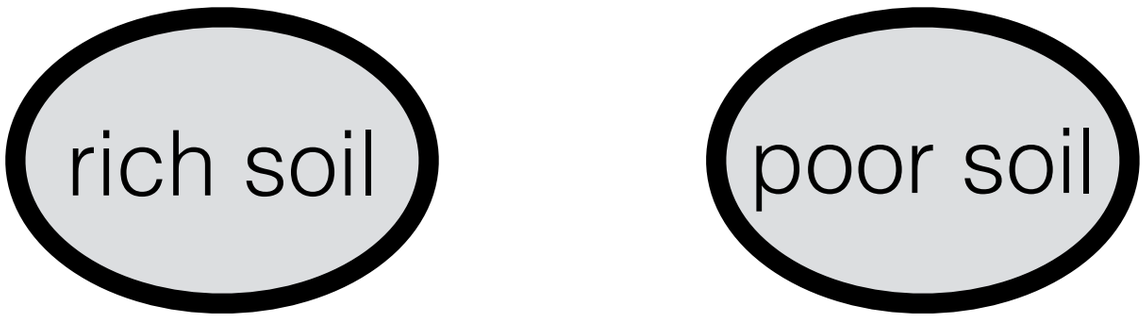
- Markov decision process: states $\mathcal{S}$

Recall



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

Recall



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

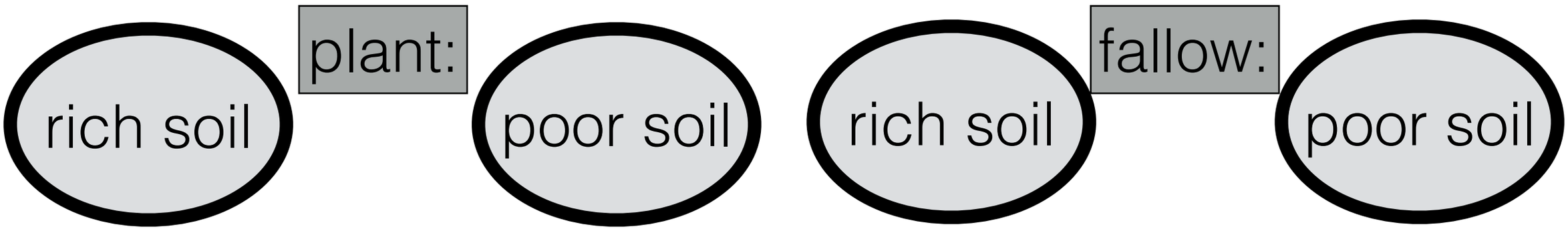
Recall

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

Recall

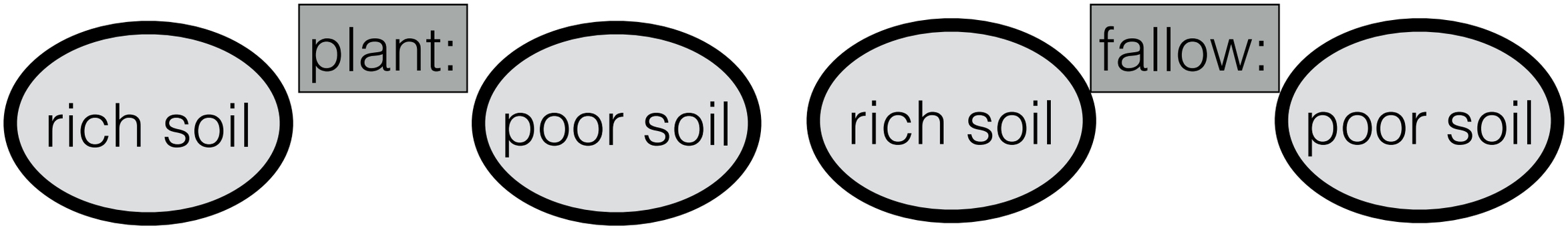
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$

Recall



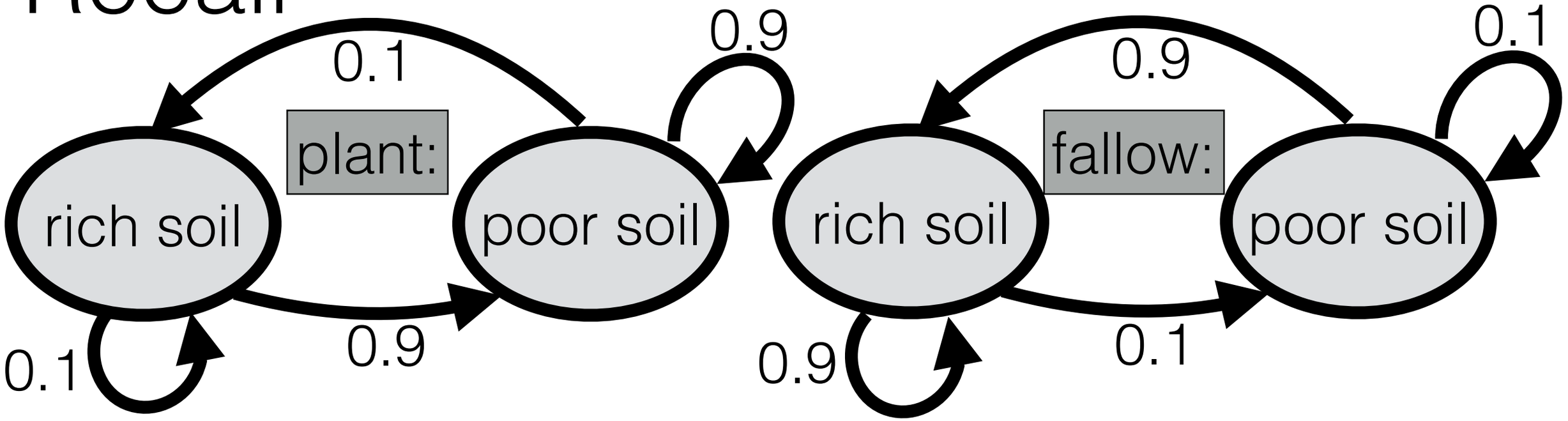
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$

Recall



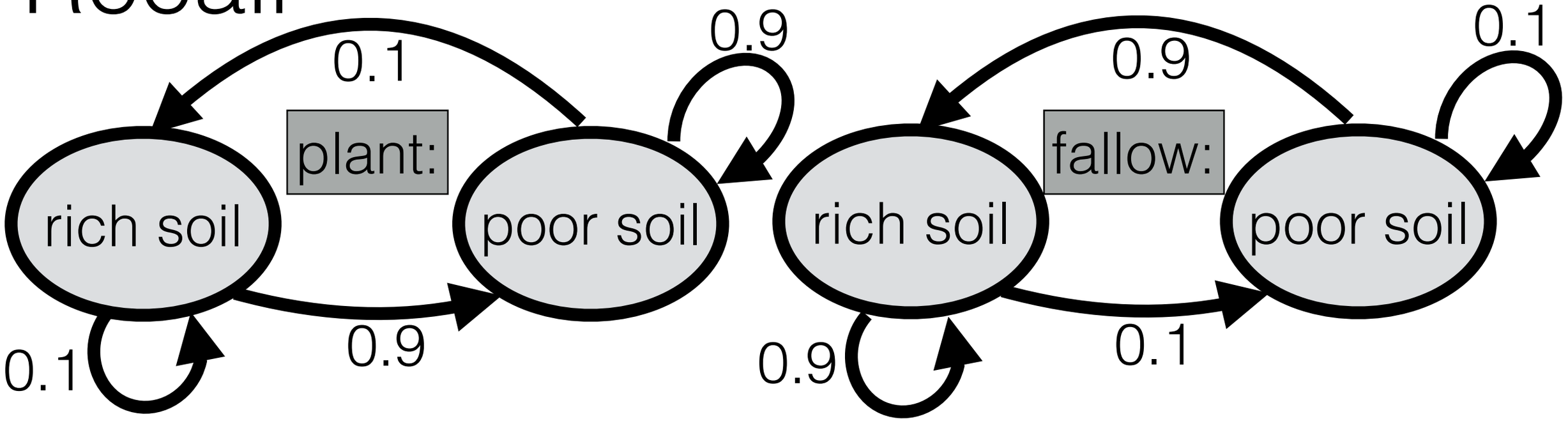
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



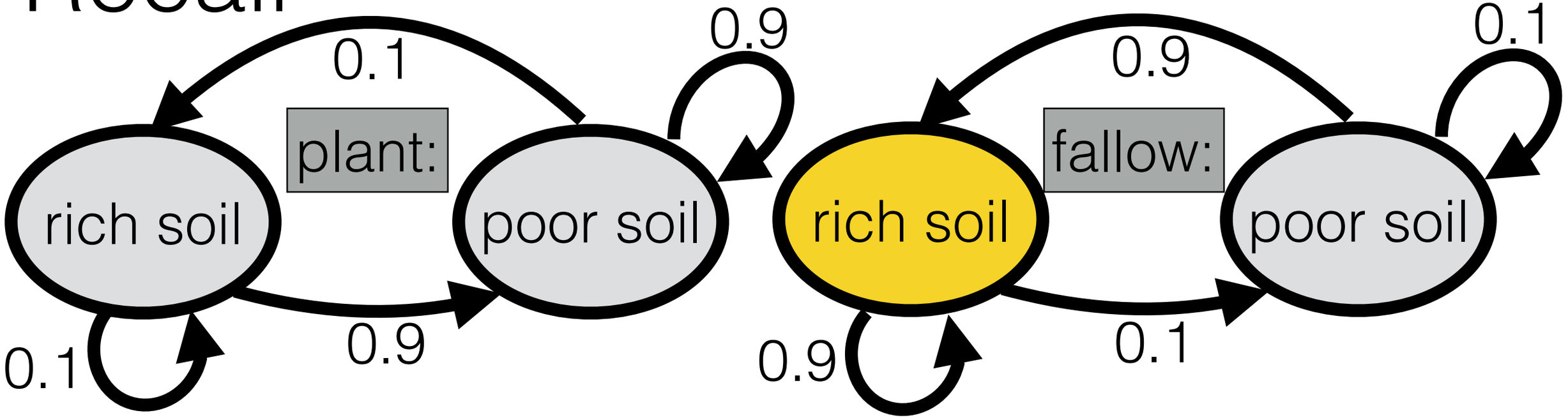
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



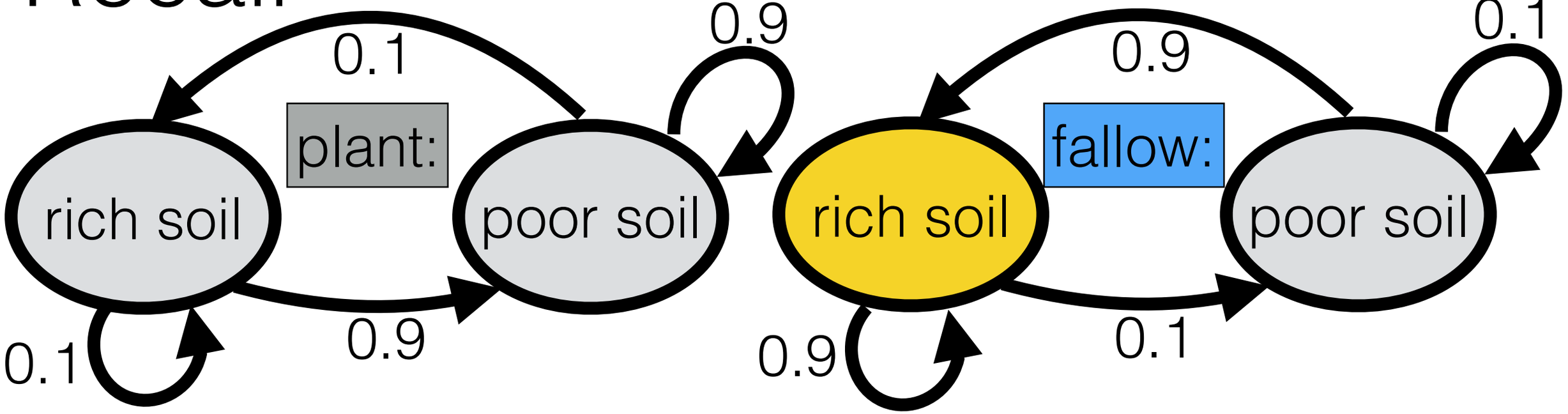
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



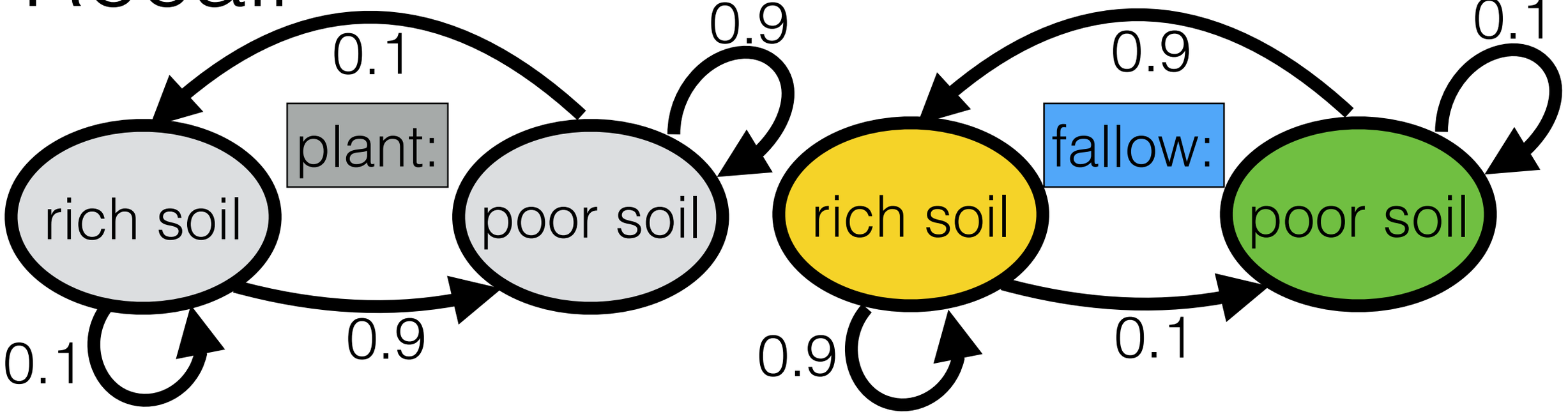
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



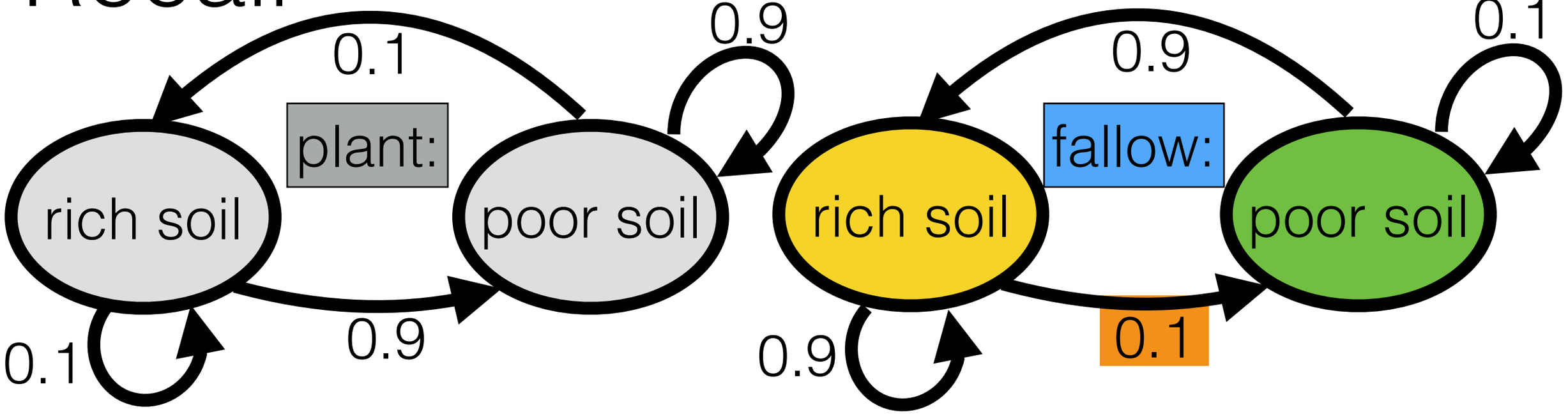
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



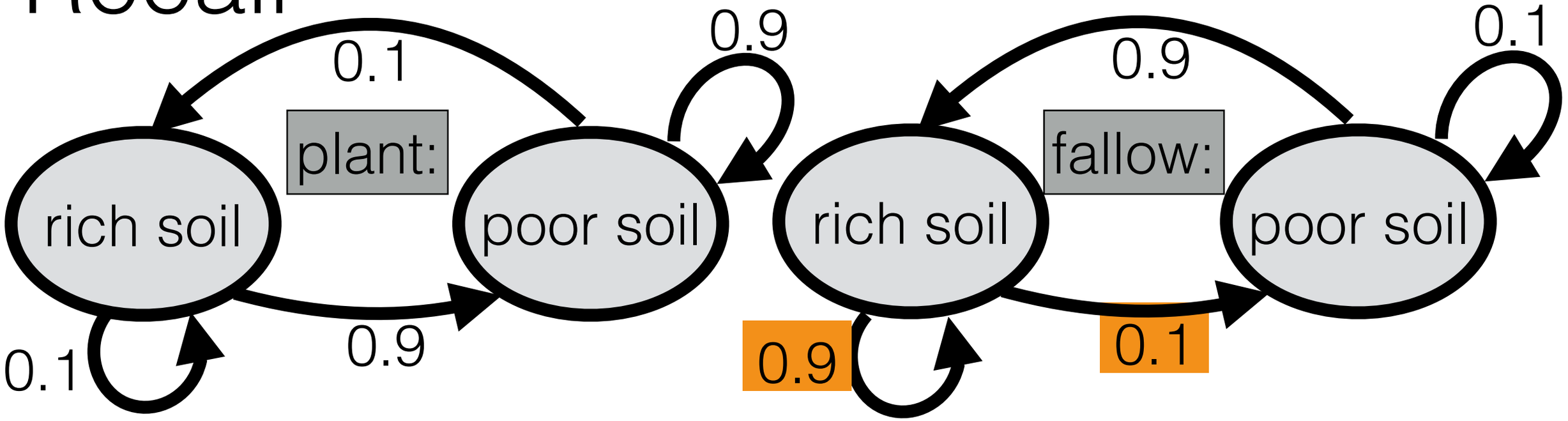
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



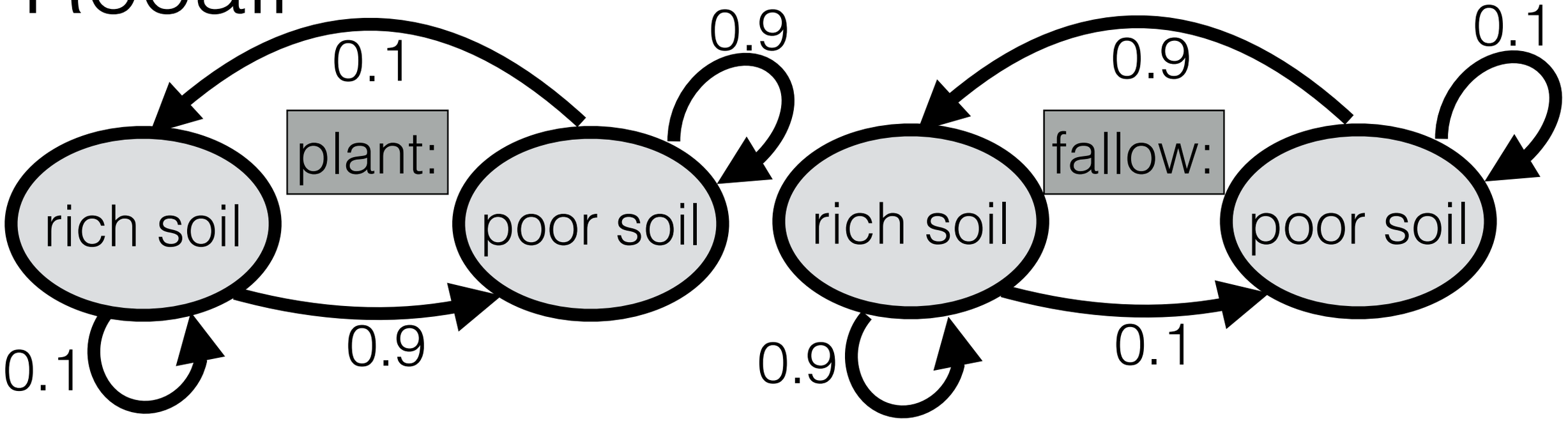
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



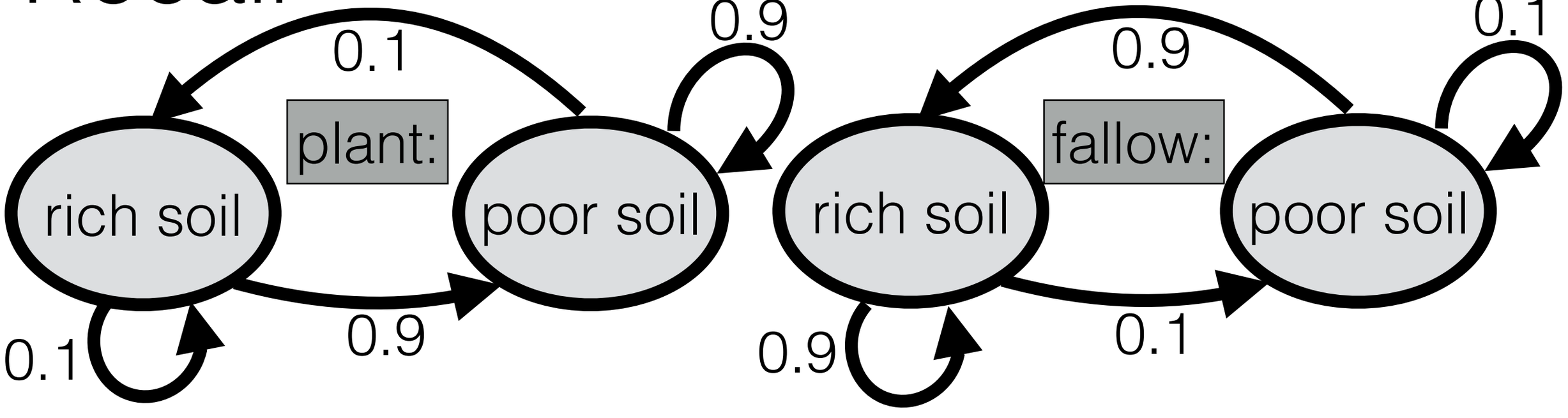
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



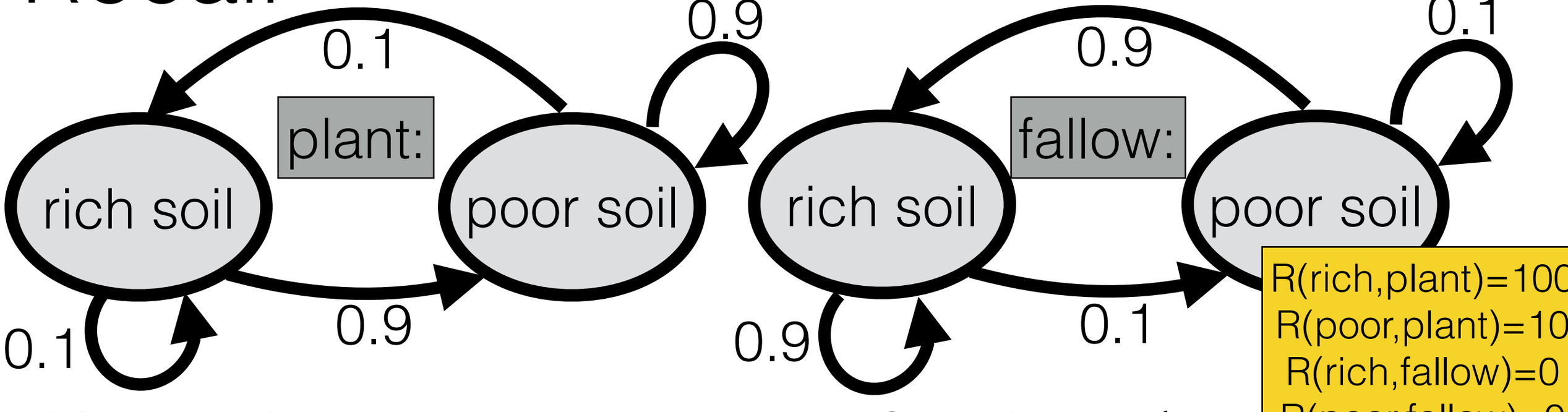
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

Recall



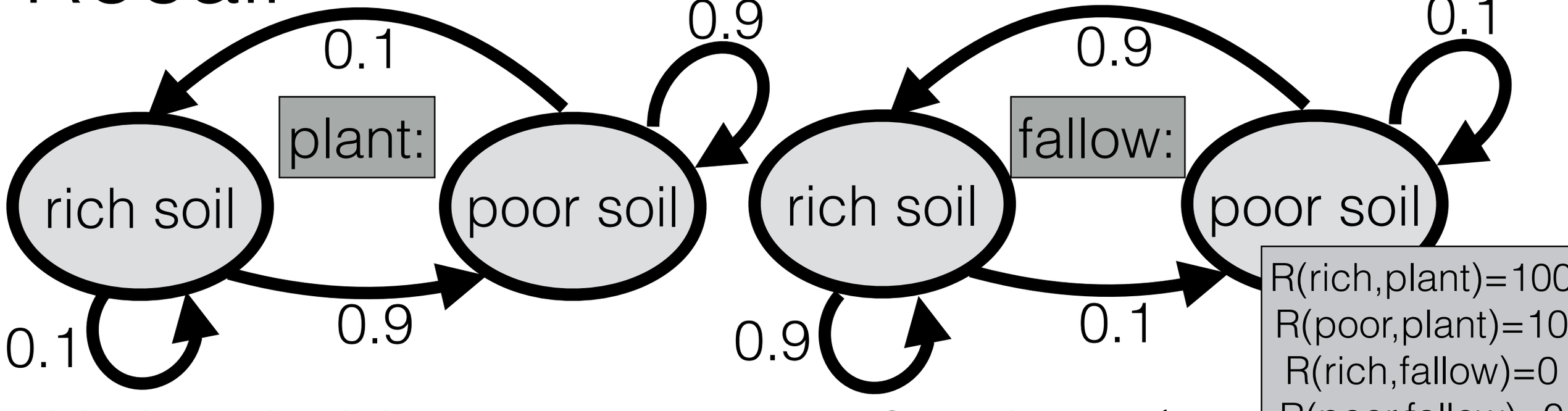
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

Recall



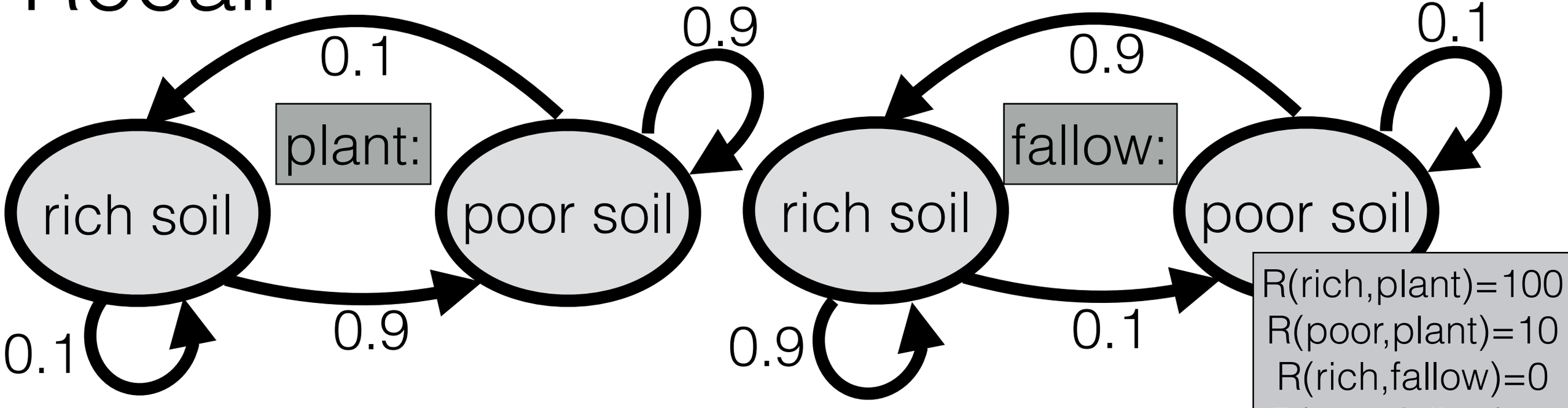
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

Recall



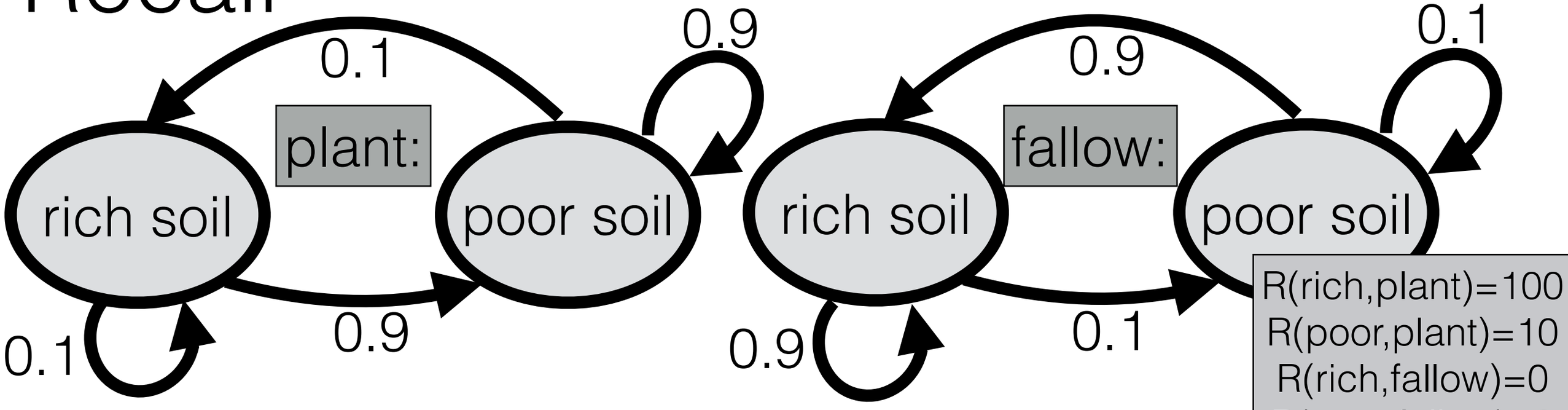
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ

Recall



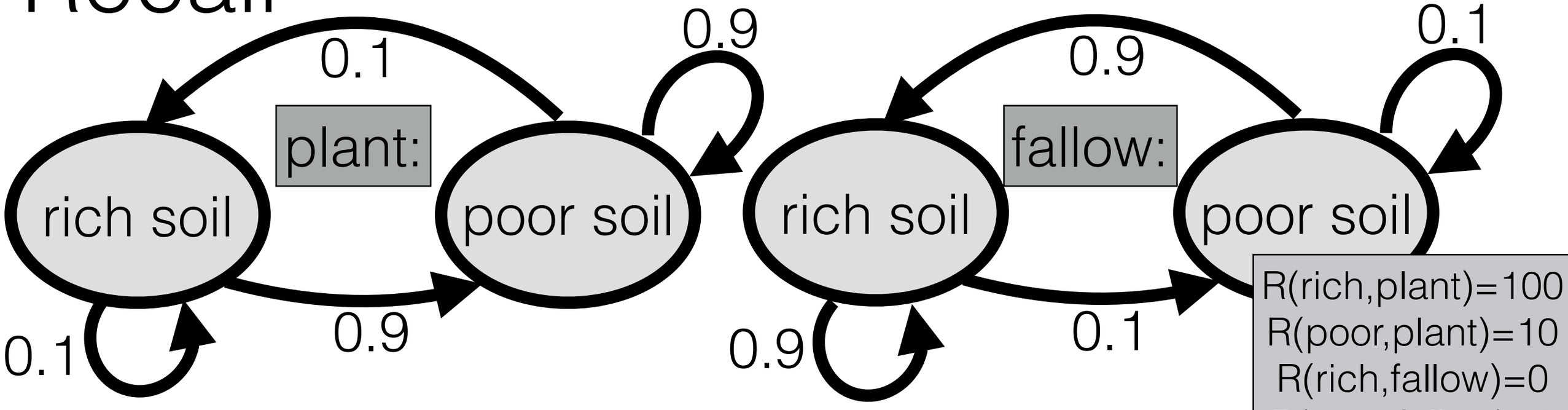
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action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

Recall



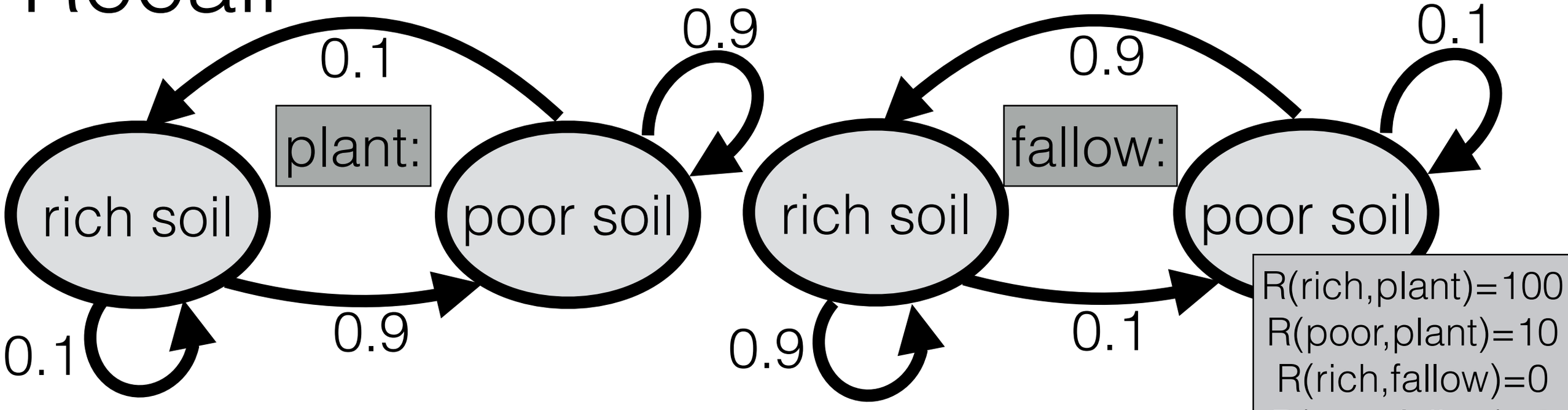
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action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

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
- Infinite horizon: Value of policy if start in state *s*

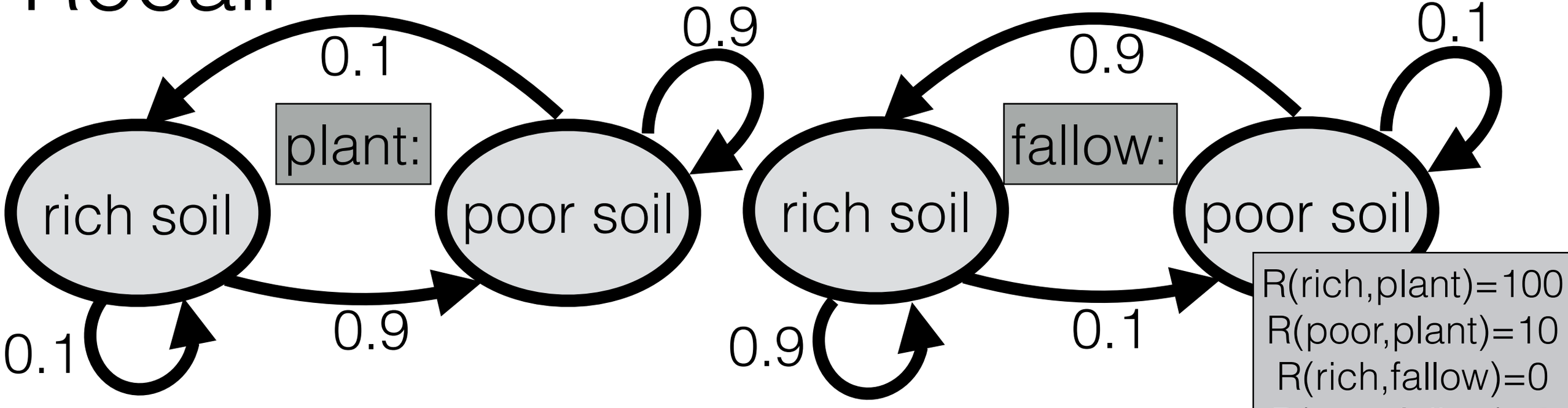
Recall



- 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
- Infinite horizon: Value of policy if start in state *s*

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

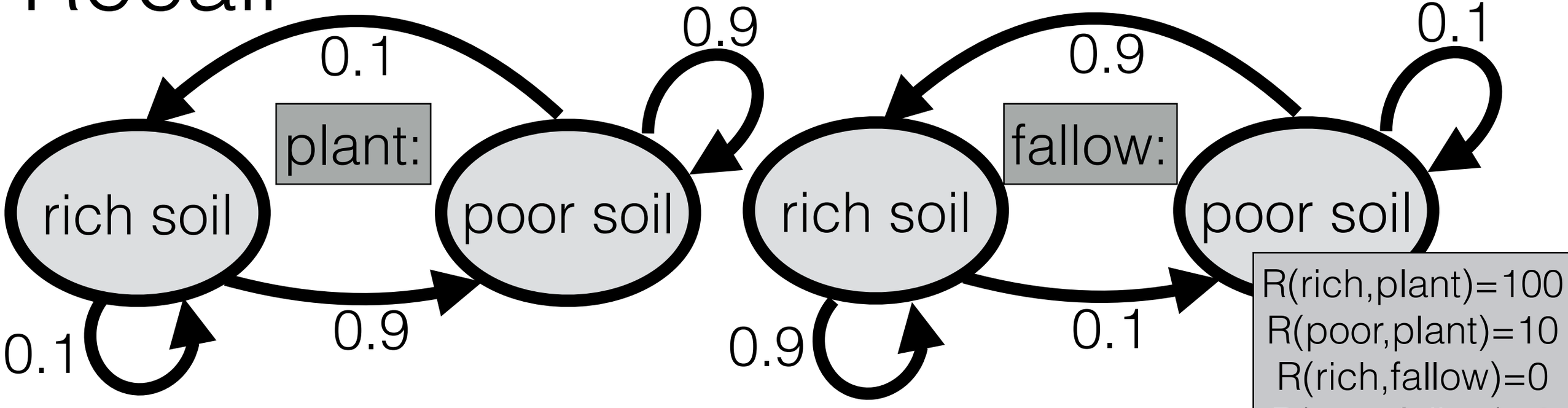
Recall



- 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
- Infinite horizon: Value of policy if start in state s

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

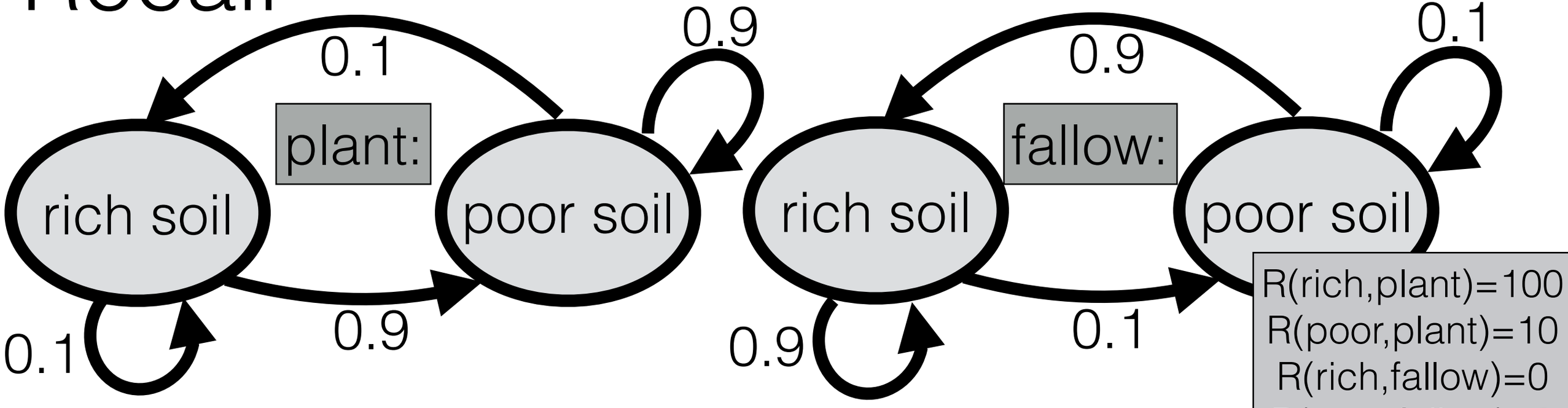
Recall



- 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
- Infinite horizon: Value of policy if start in state s

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

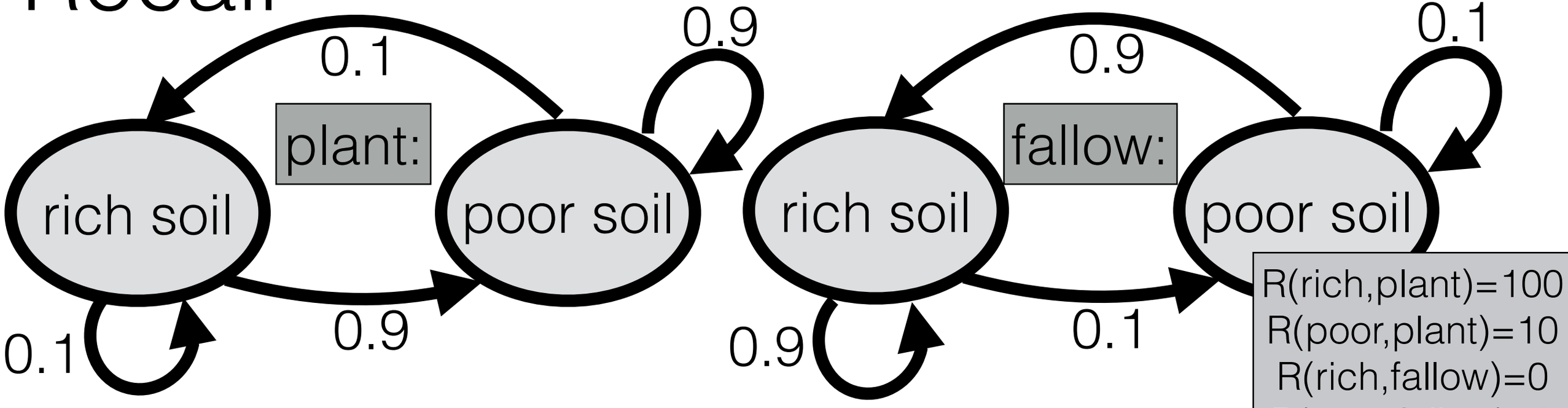
Recall



- 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
- Infinite horizon: Value of policy if start in state s

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

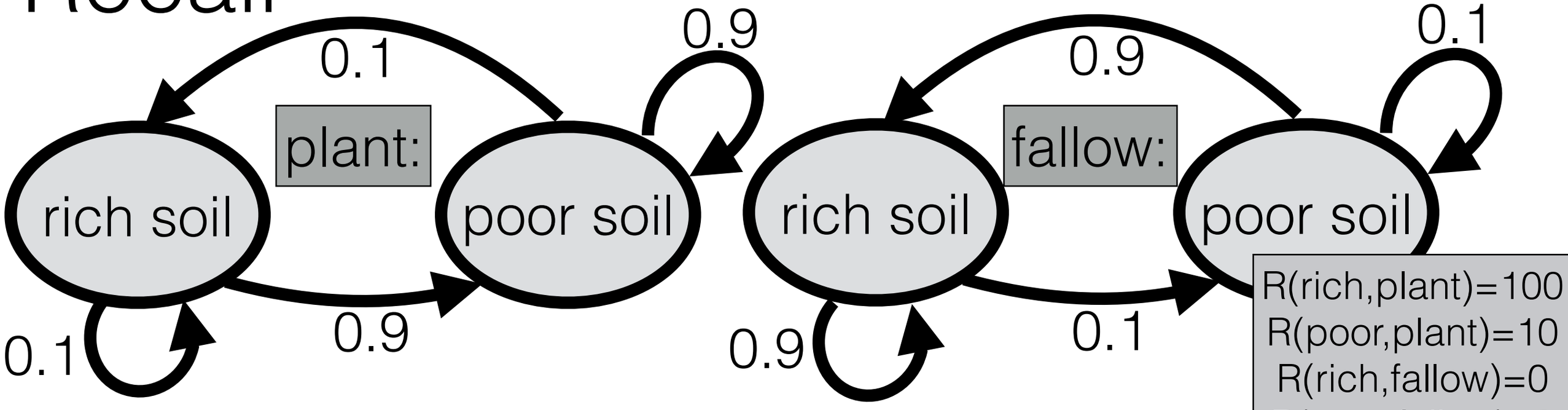
Recall



- 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
- Infinite horizon: Value of policy if start in state s

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

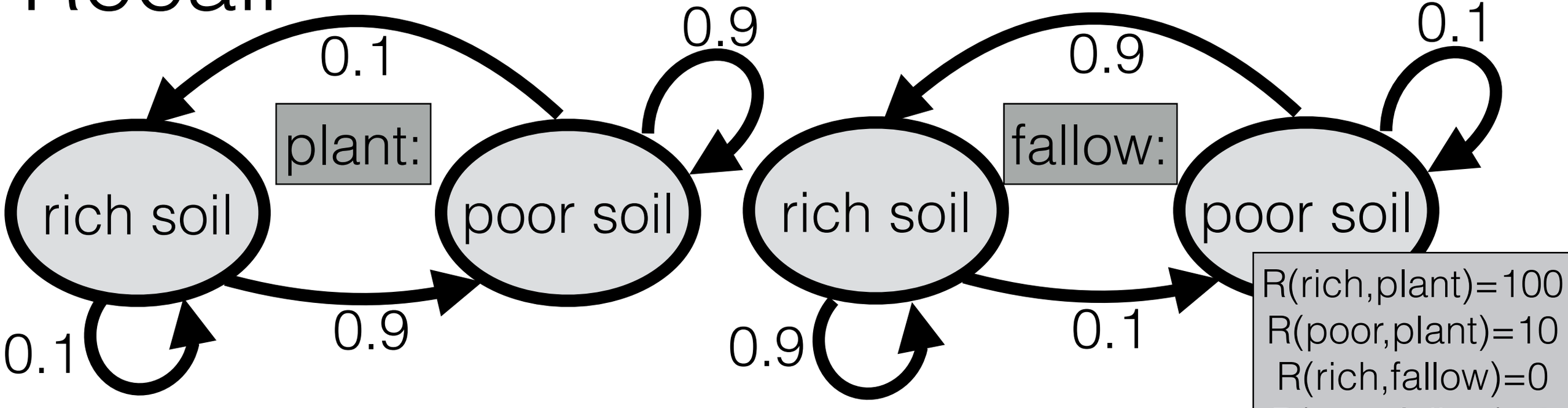
Recall



- 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
- Infinite horizon: Value of policy if start in state s

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

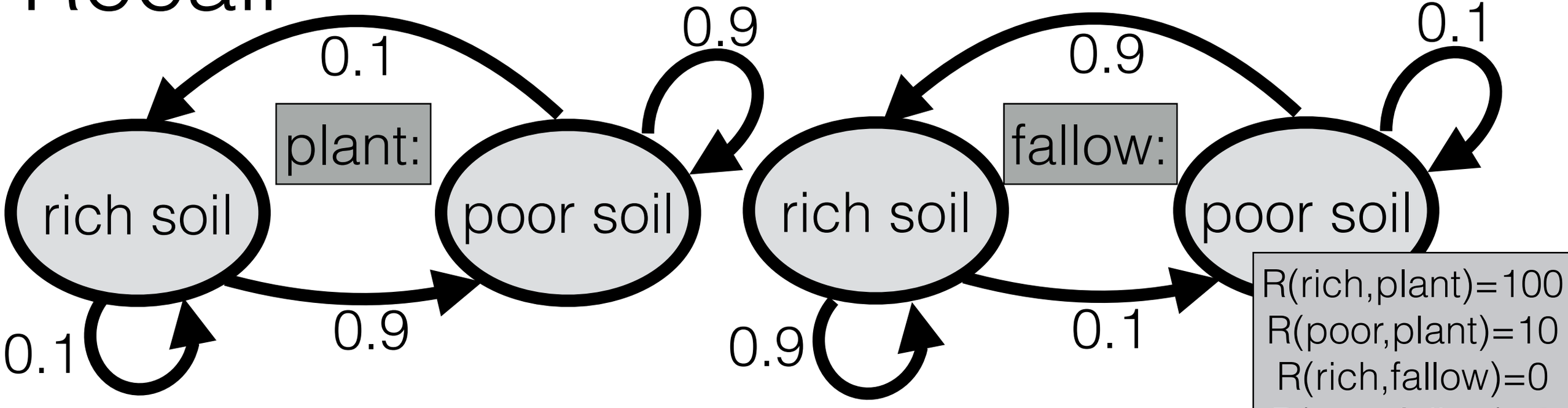
Recall



- 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
- Infinite horizon: Value of policy if start in state s

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

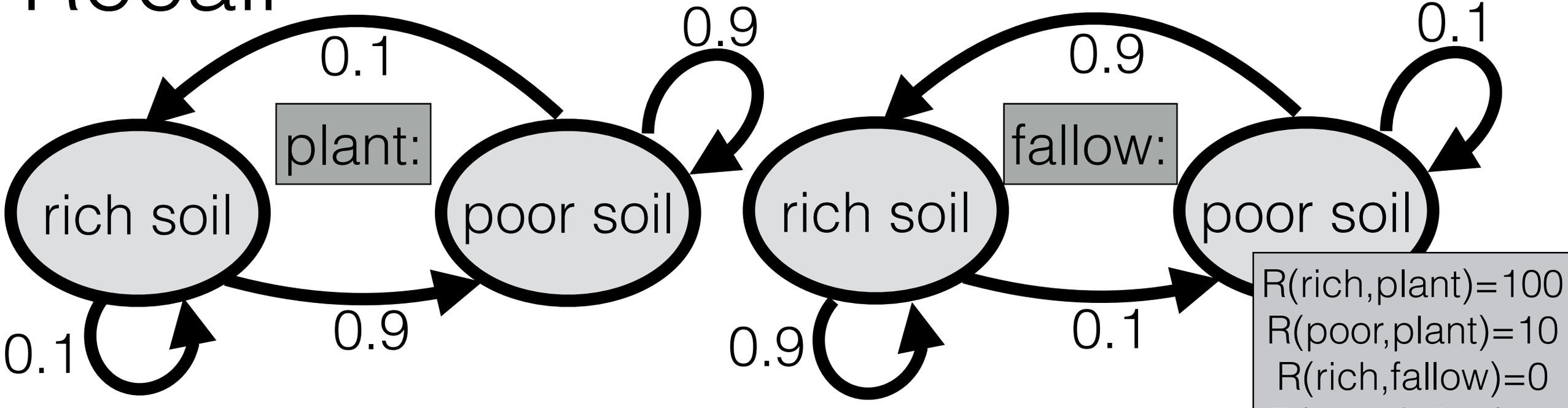
Recall



- 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
- Infinite horizon: Value of policy if start in state s

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

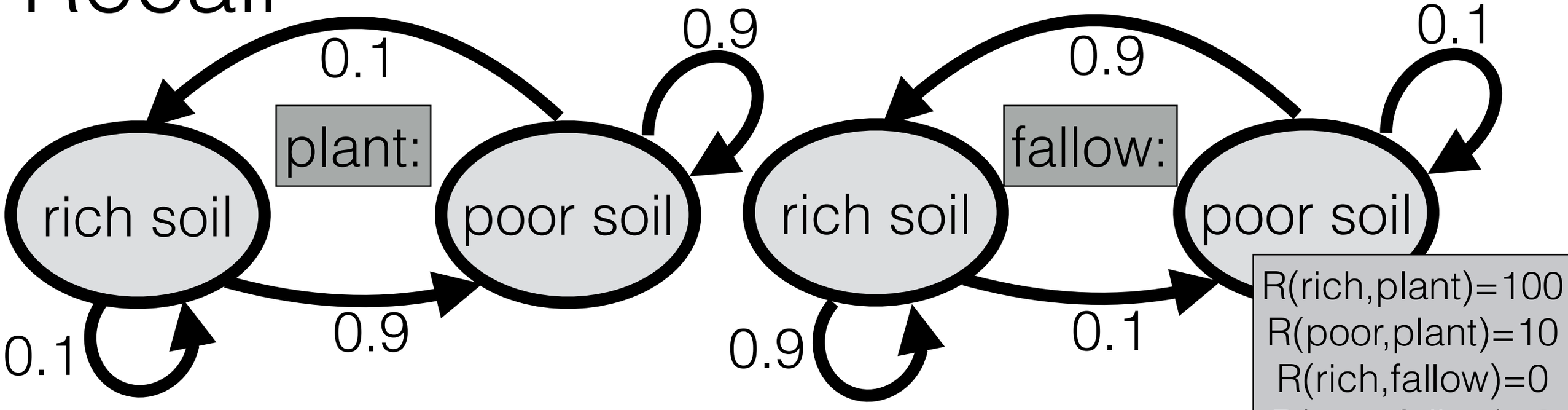
Recall



- 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
- Infinite horizon: Value of policy if start in state s

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

Recall

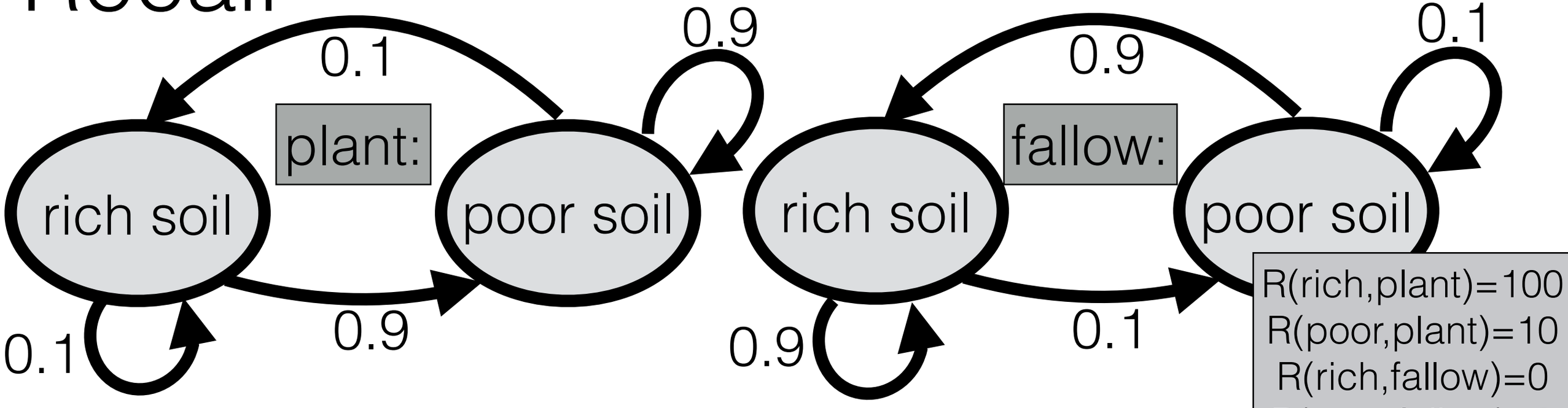


- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state
- Infinite horizon: Value of policy if start in state *s*

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

  - Value of action *a* in state *s* if make the best actions in future

Recall



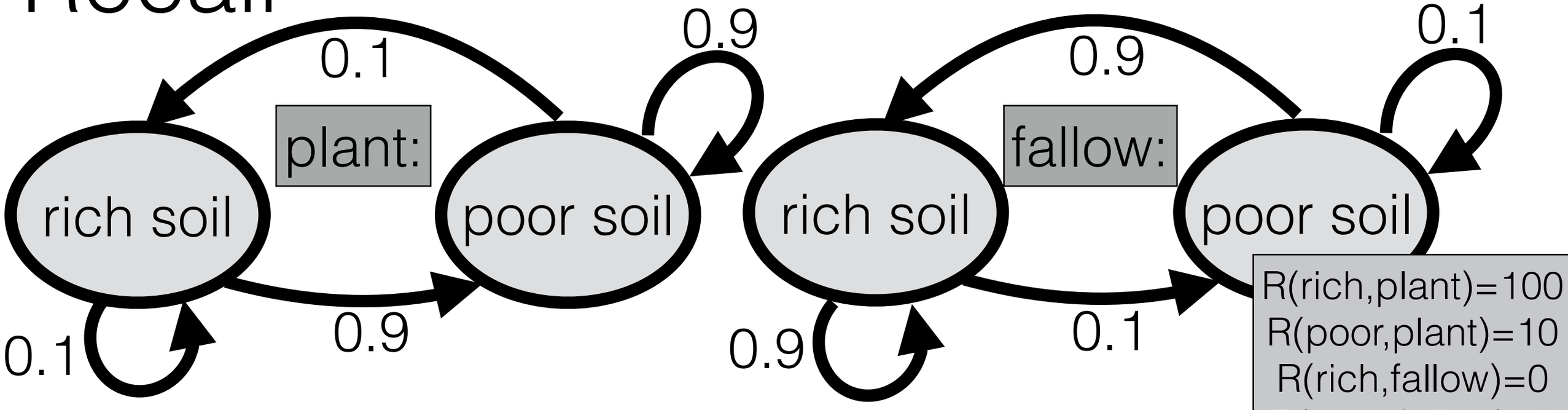
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state
- Infinite horizon: Value of policy if start in state *s*

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

  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

Recall



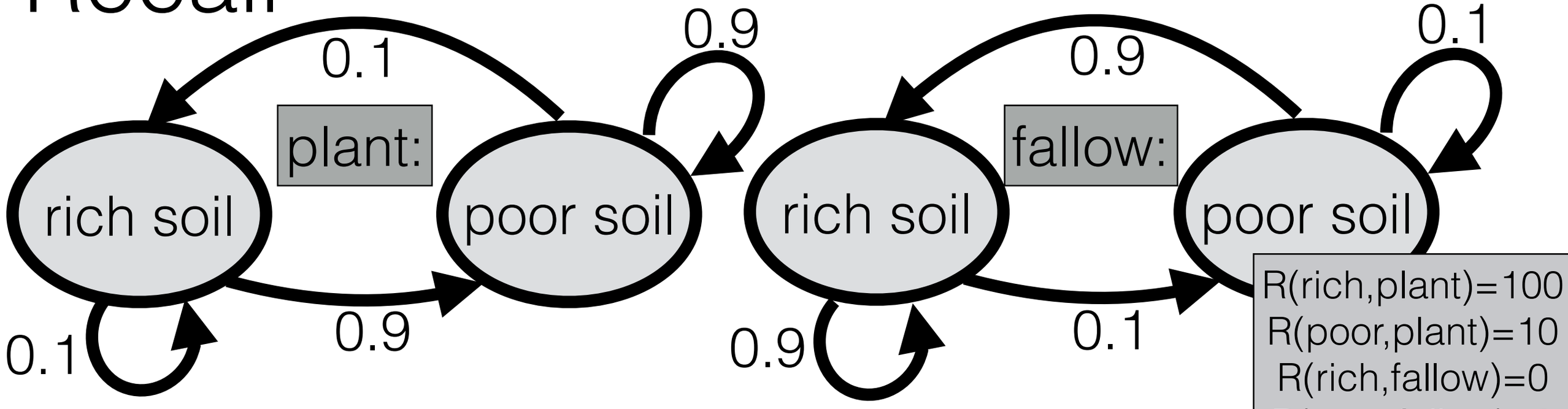
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state
- Infinite horizon: Value of policy if start in state *s*

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

  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

Recall



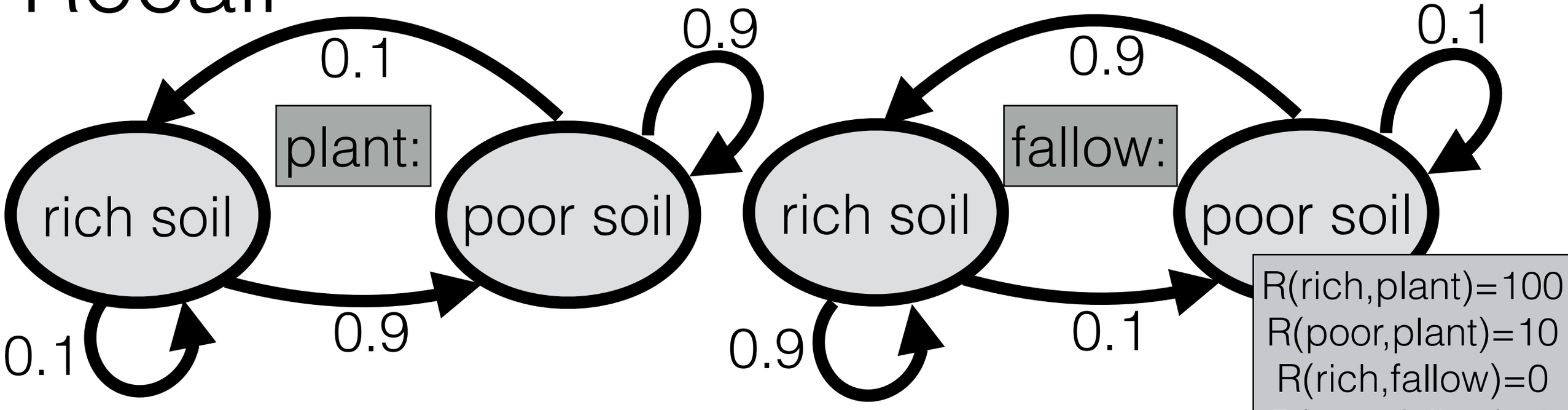
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state
- Infinite horizon: Value of policy if start in state *s*

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

  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

Recall



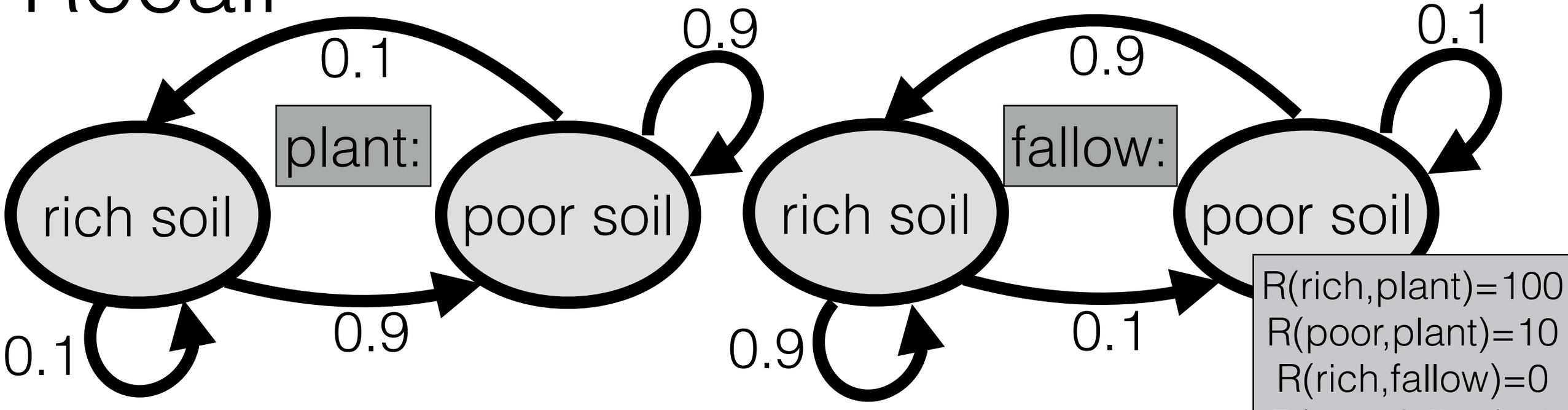
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state
- Infinite horizon: Value of policy if start in state *s*

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

  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

Recall



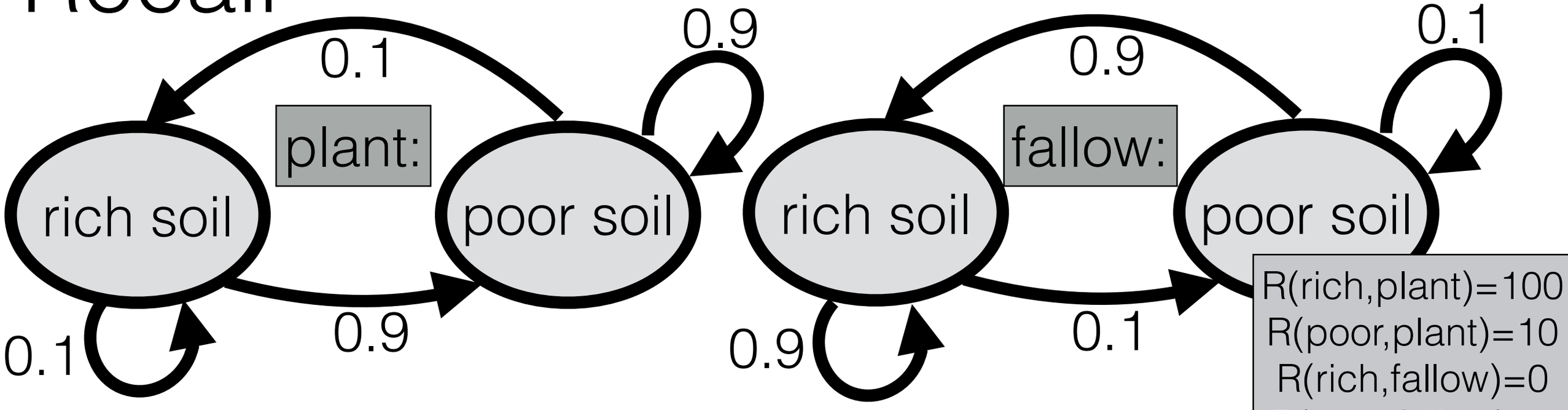
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state
- Infinite horizon: Value of policy if start in state *s*

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

  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

Recall



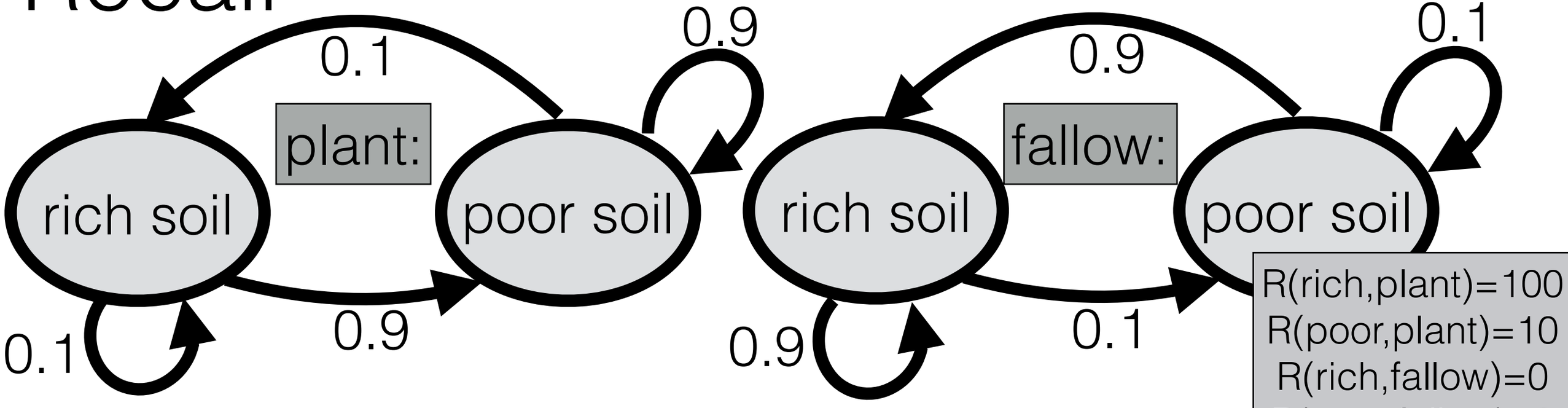
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state
- Infinite horizon: Value of policy if start in state *s*

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

  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

Recall



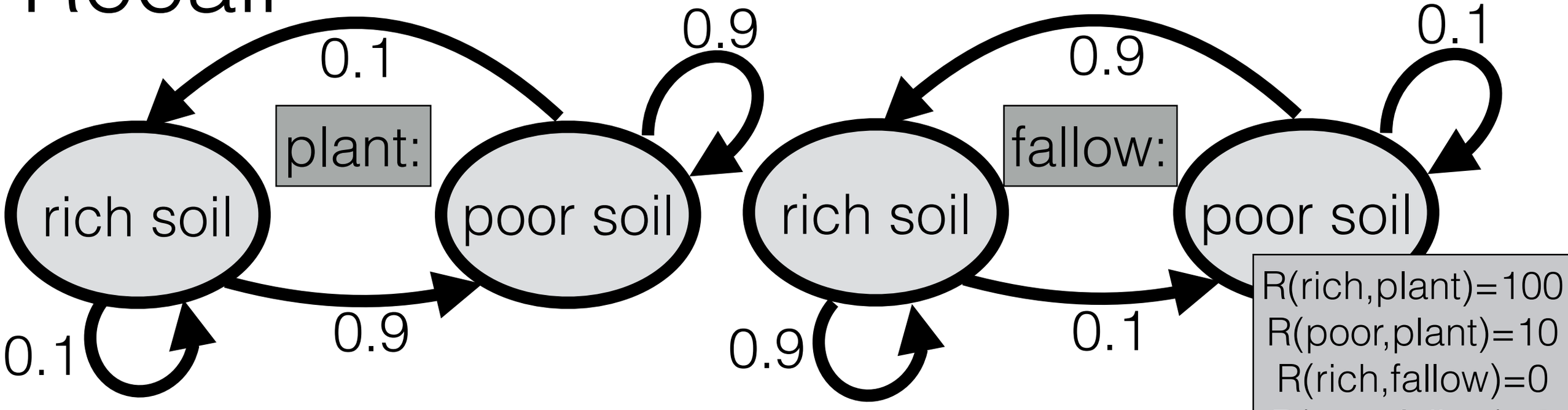
- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, discount factor γ
- Policy $\pi : \mathcal{S} \to \mathcal{A}$: action to take in a state
- Infinite horizon: Value of policy if start in state *s*

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

  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, 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
- Infinite horizon: Value of policy if start in state *s*

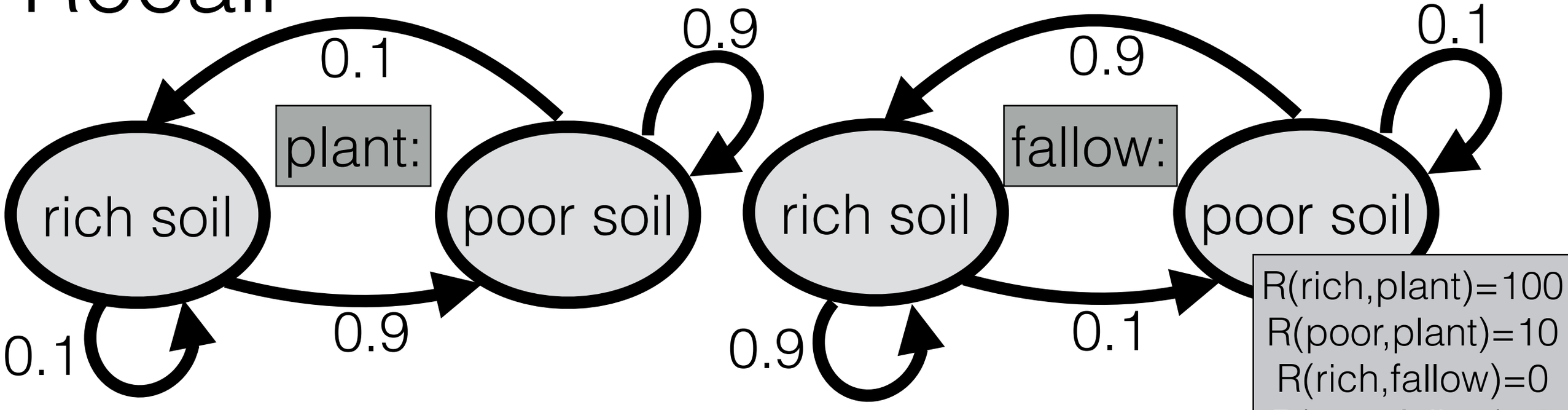
$$V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$$

  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

  - What's the best policy?

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
- Infinite horizon: Value of policy if start in state *s*

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

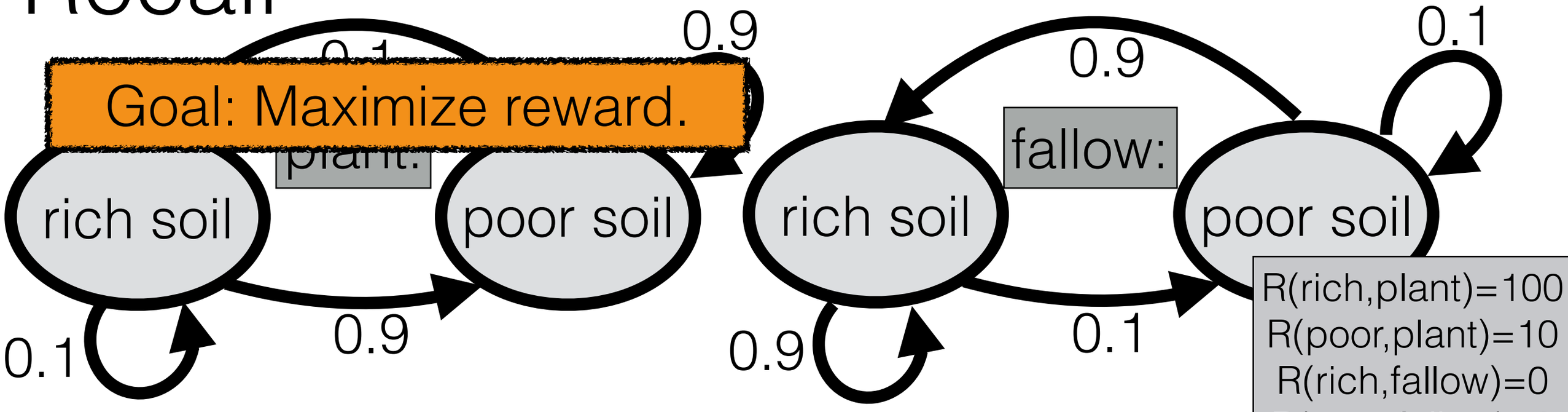
  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

  - What's the best policy?

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, 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
- Infinite horizon: Value of policy if start in state *s*

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

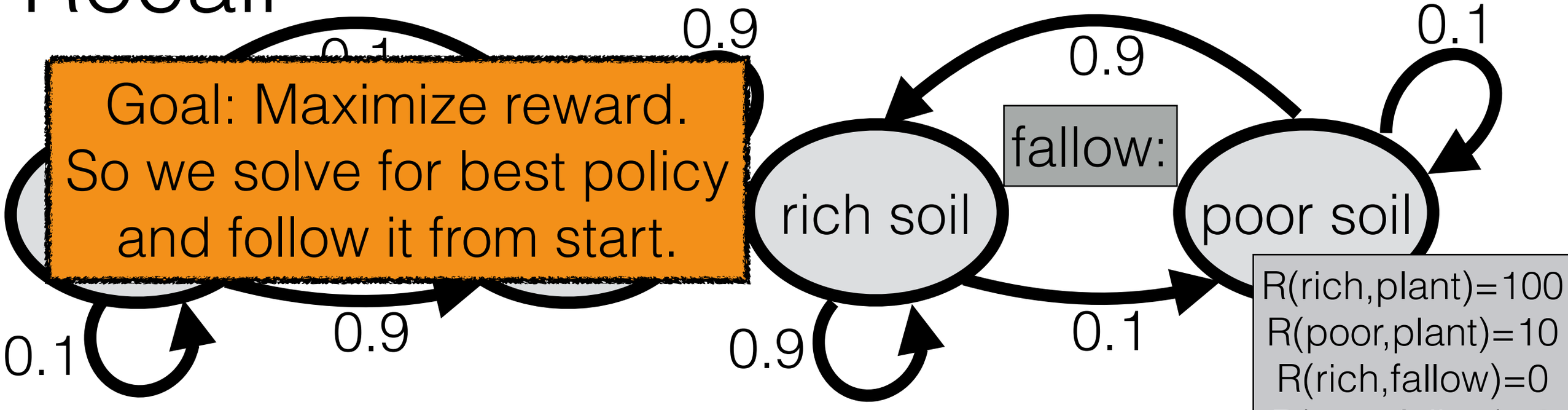
  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

  - What's the best policy?

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, 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
- Infinite horizon: Value of policy if start in state *s*

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

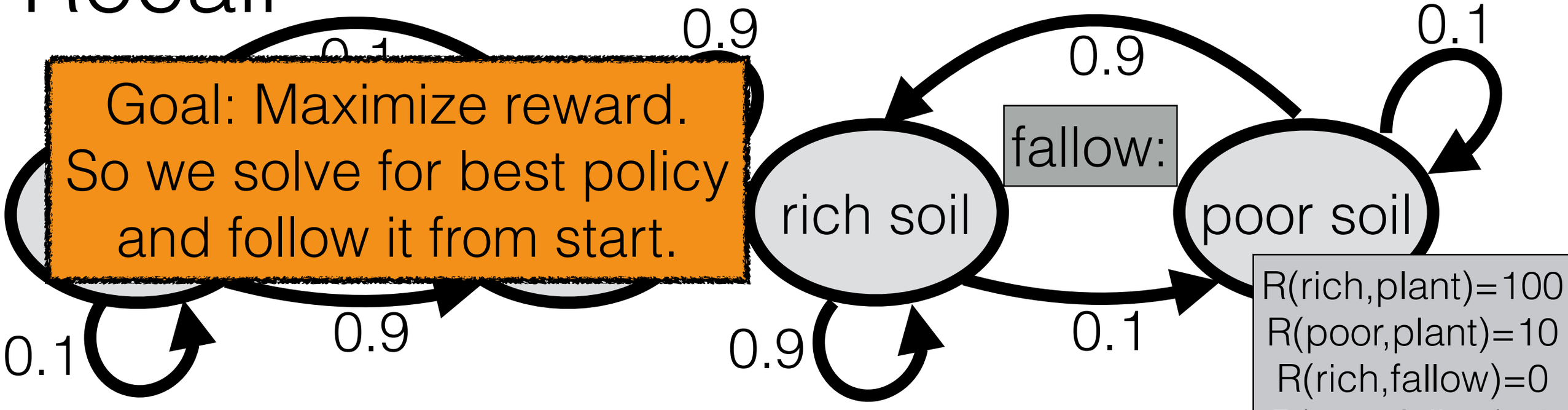
  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

  - What's the best policy?

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, 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
- Infinite horizon: Value of policy if start in state *s*

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

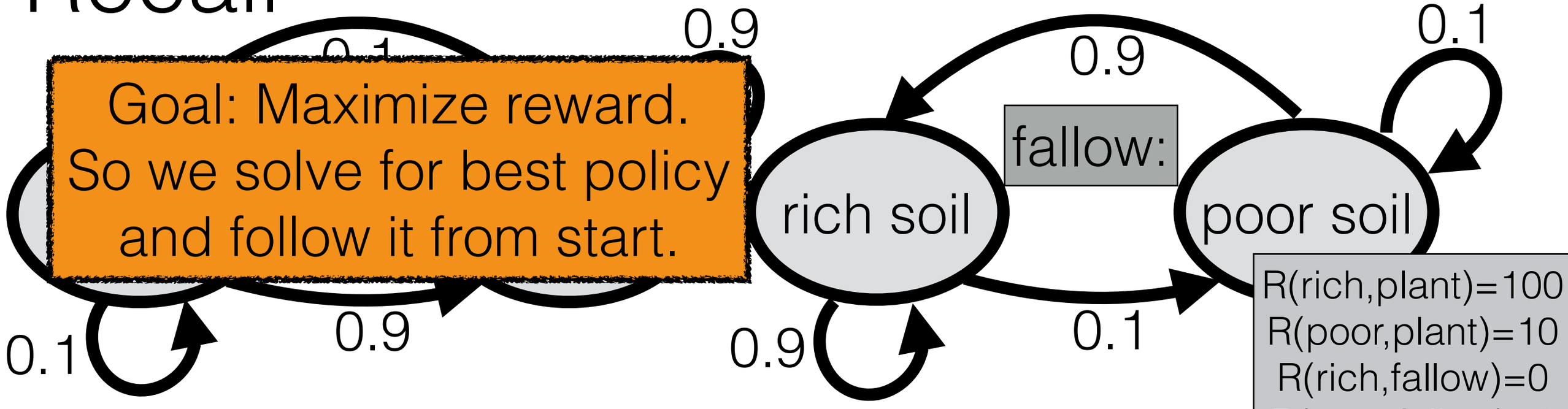
  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

  - What's the best policy?

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, 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
- Infinite horizon: Value of policy if start in state *s*

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

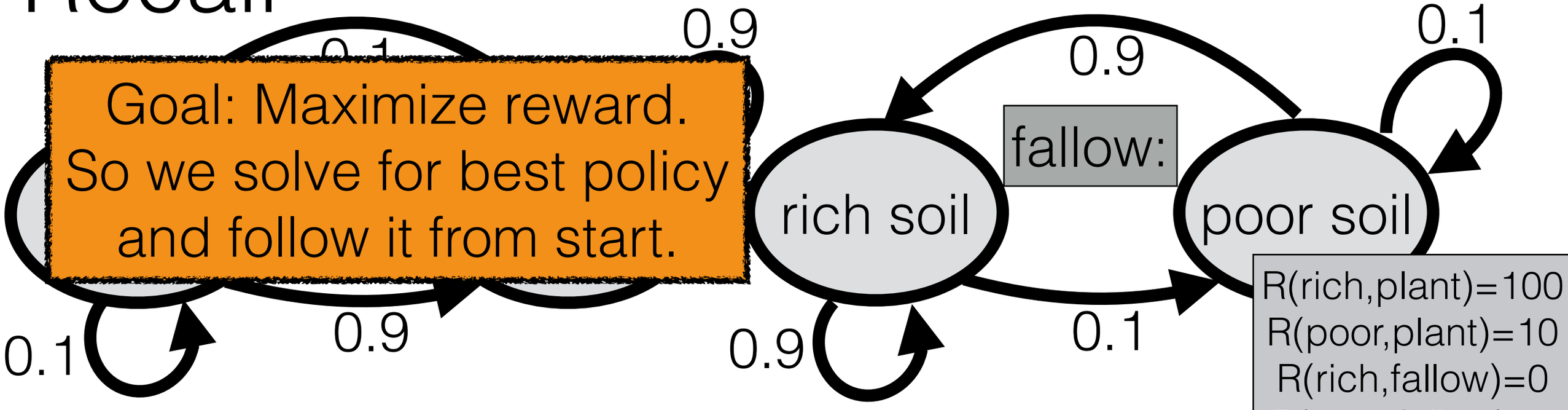
  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

  - What's the best policy?

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, 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
- Infinite horizon: Value of policy if start in state *s*

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

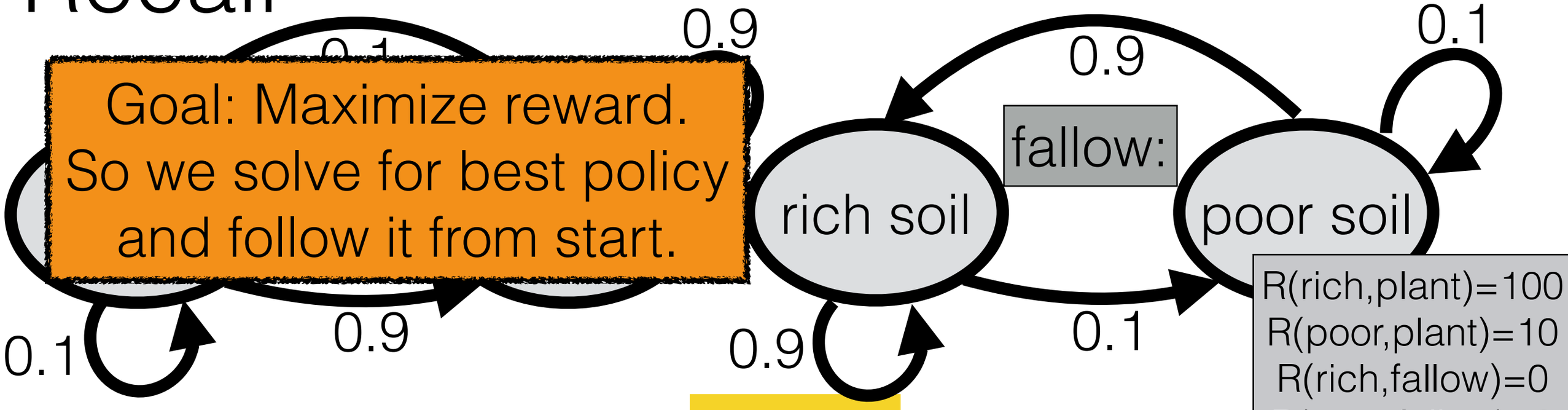
  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

  - What's the best policy?

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, 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
- Infinite horizon: Value of policy if start in state *s*

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

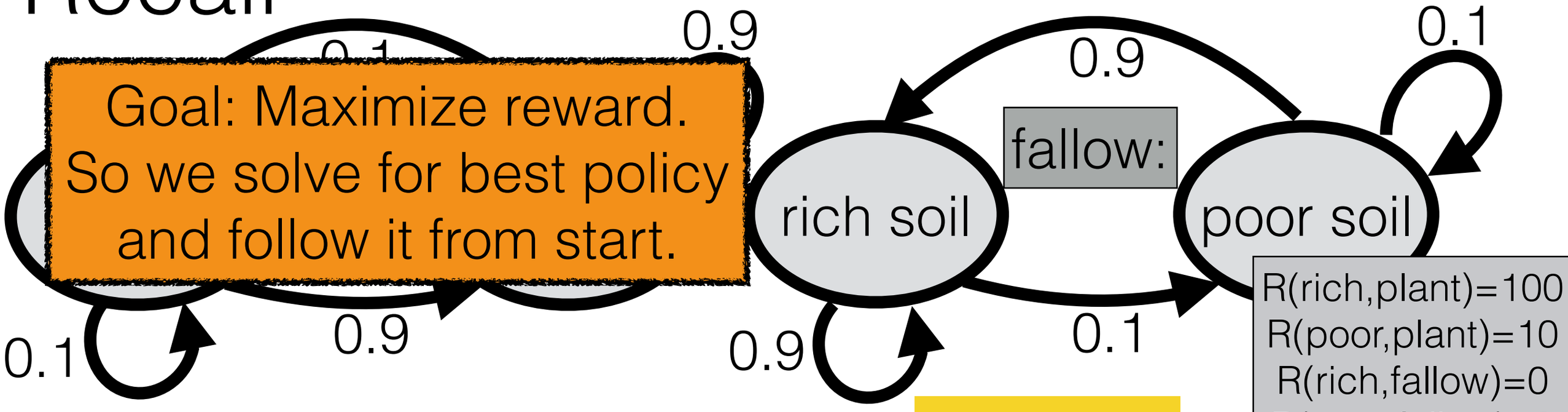
  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

  - What's the best policy?

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, 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
- Infinite horizon: Value of policy if start in state *s*

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

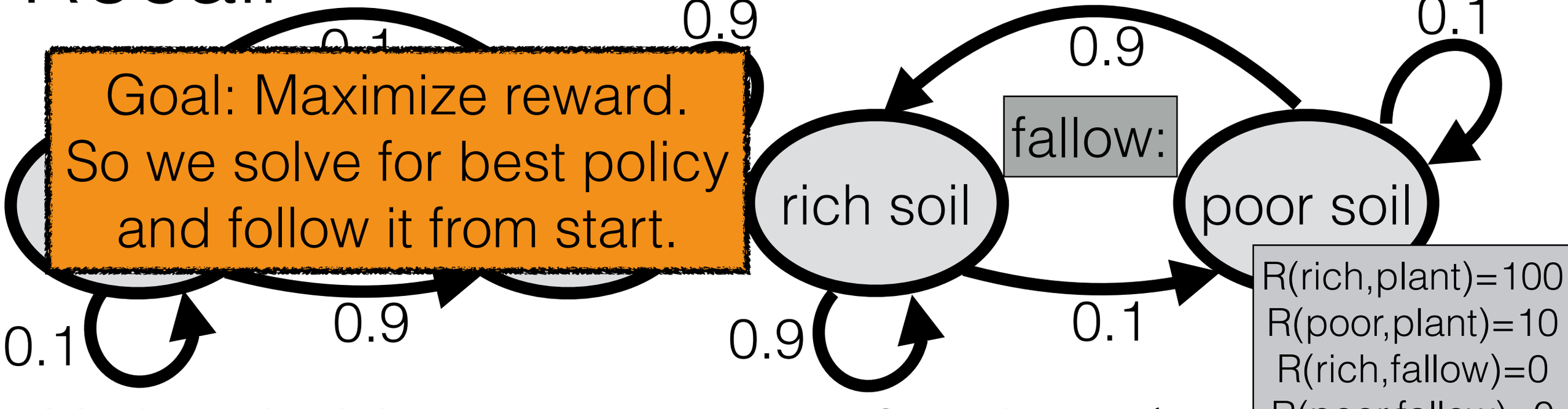
  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

  - What's the best policy?

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, 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
- Infinite horizon: Value of policy if start in state *s*

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

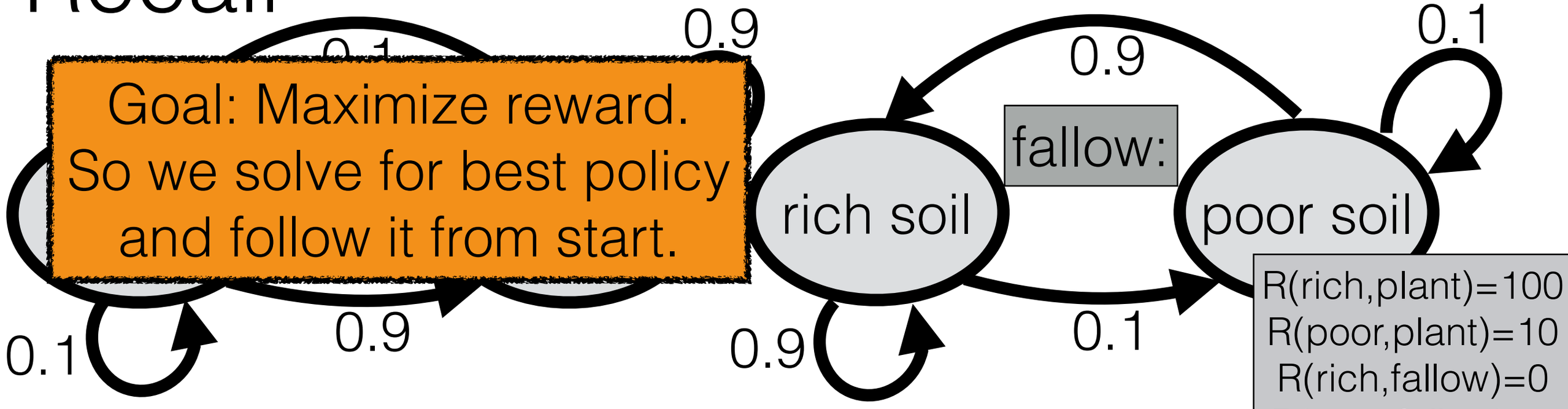
  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

  - What's the best policy?

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, 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
- Infinite horizon: Value of policy if start in state *s*

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

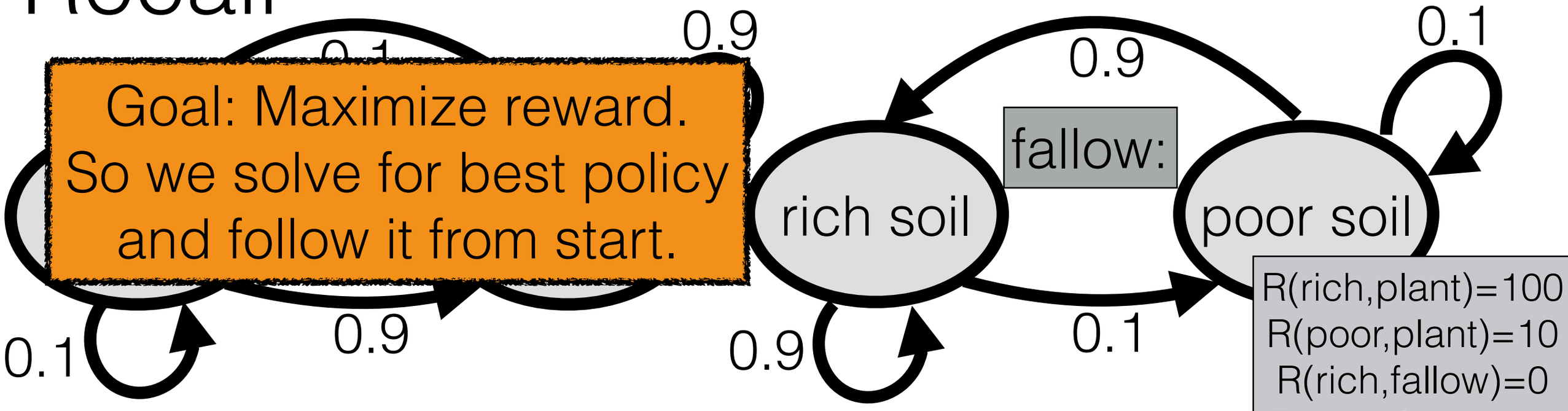
  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

  - What's the best policy?

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, 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
- Infinite horizon: Value of policy if start in state *s*

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

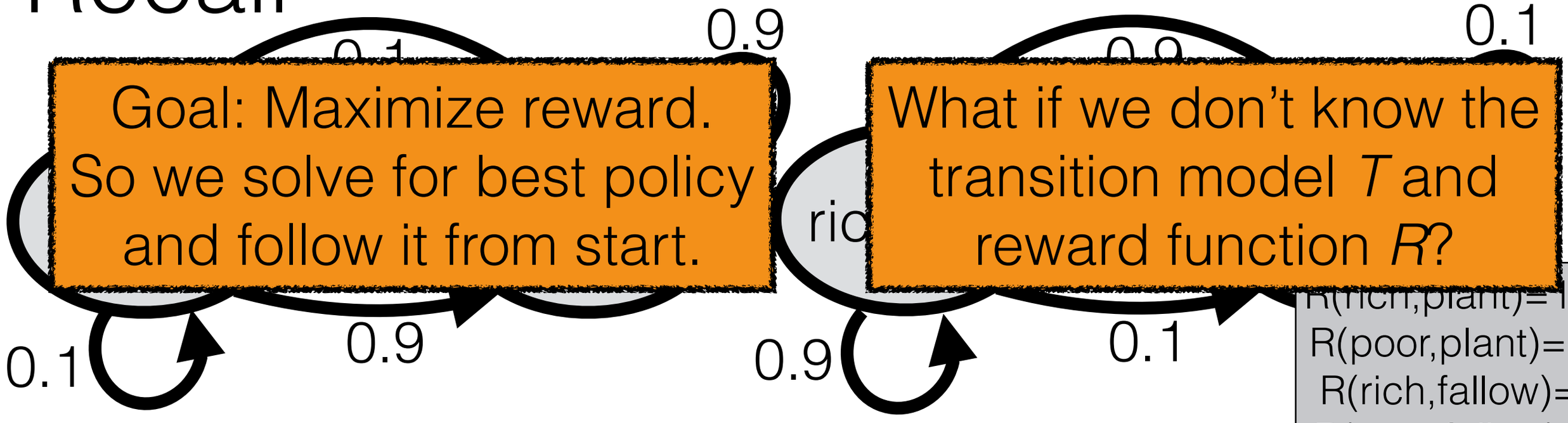
  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

  - What's the best policy?

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

Recall



Goal: Maximize reward. So we solve for best policy and follow it from start.

What if we don't know the transition model *T* and reward function *R*?

- 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
- Infinite horizon: Value of policy if start in state *s*

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

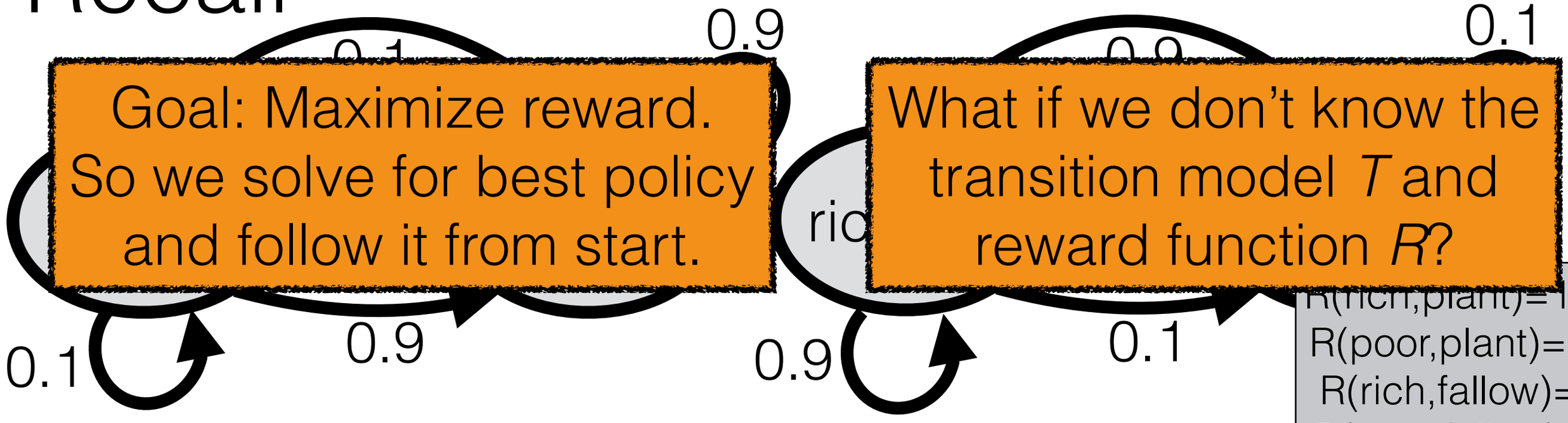
  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

  - What's the best policy?

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

Recall



- Markov decision process: states $\mathcal{S}$, actions $\mathcal{A}$, transition model $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, 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
- Infinite horizon: Value of policy if start in state *s*

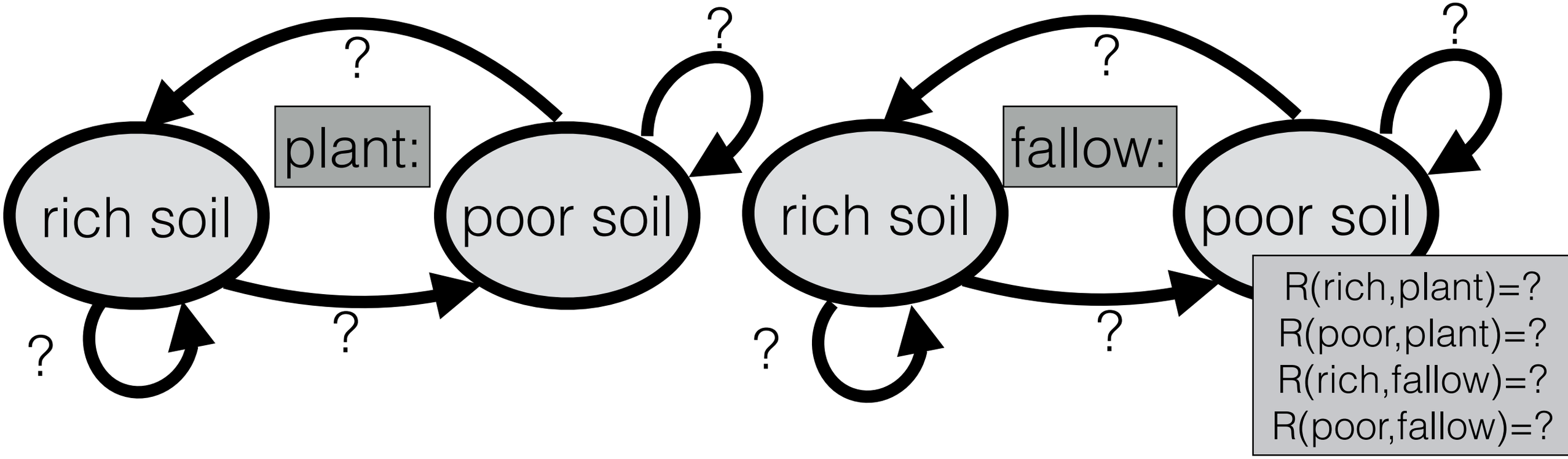
$$V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$$

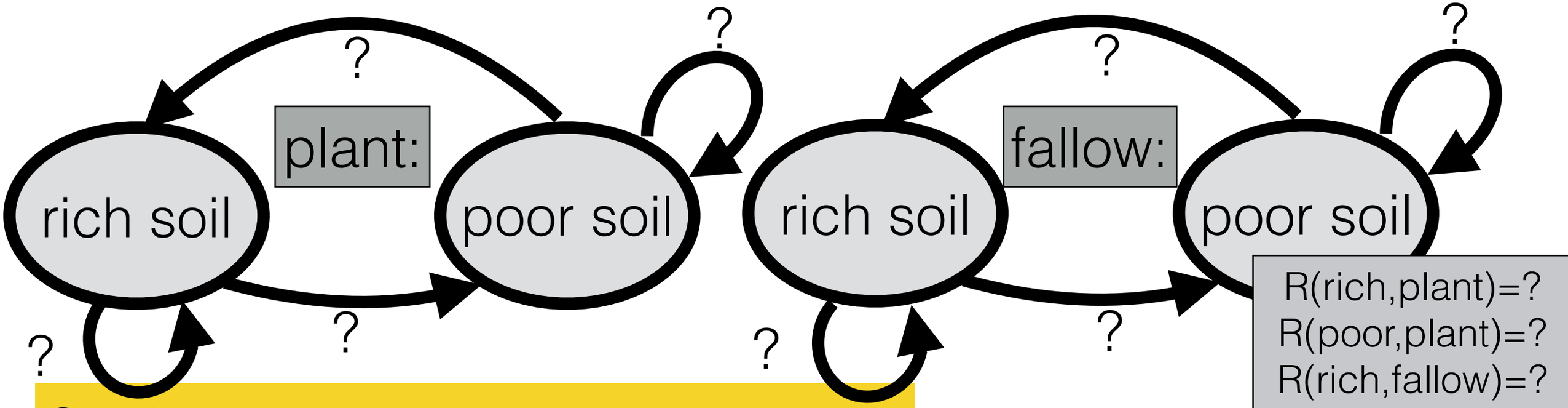
  - Value of action *a* in state *s* if make the best actions in future

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} T(s, a, s') \max_{a'} Q^*(s', a')$$

  - What's the best policy?

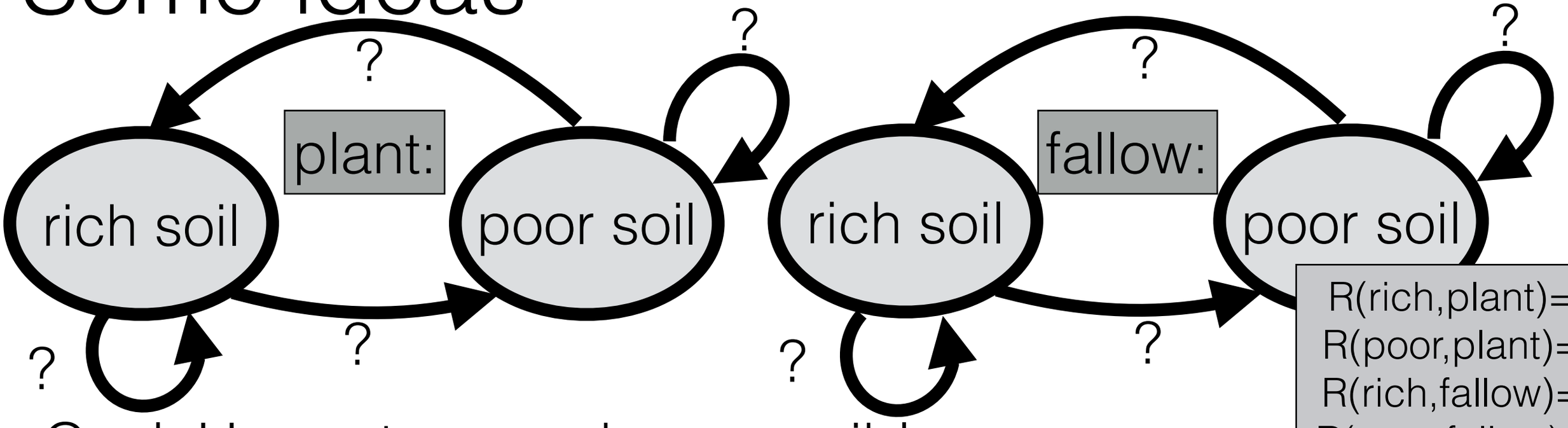
$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$


?
?
?
plant:
rich soil
poor soil
?
?
?
?
?
fallow:
rich soil
poor soil
?
R(rich,plant)=?
R(poor,plant)=?
R(rich,fallow)=?
R(poor,fallow)=?



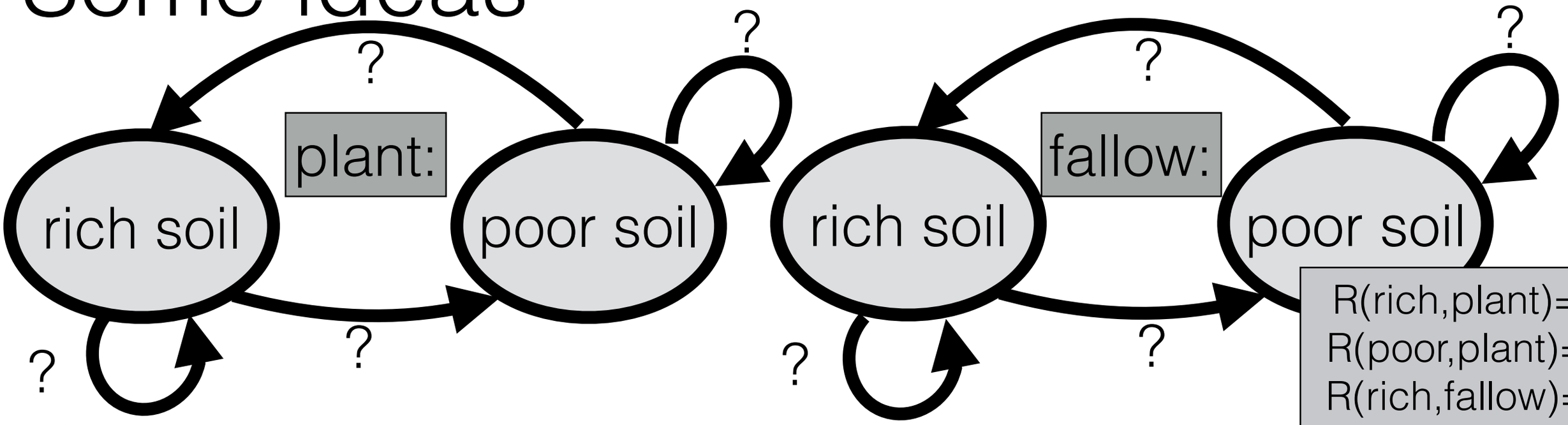
- Goal: Harvest as much as possible

Some ideas



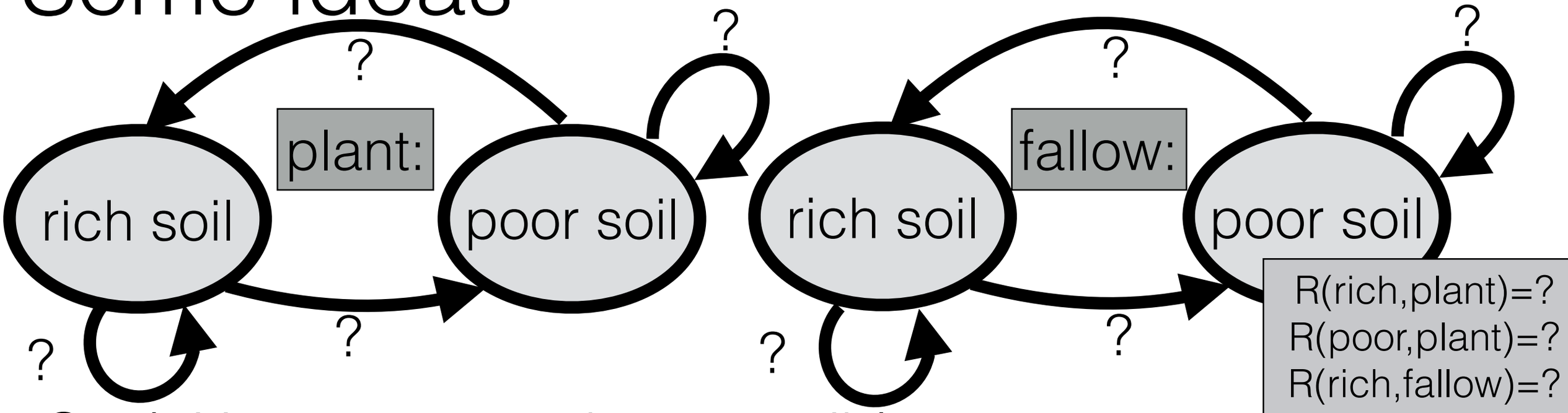
- Goal: Harvest as much as possible

Some ideas



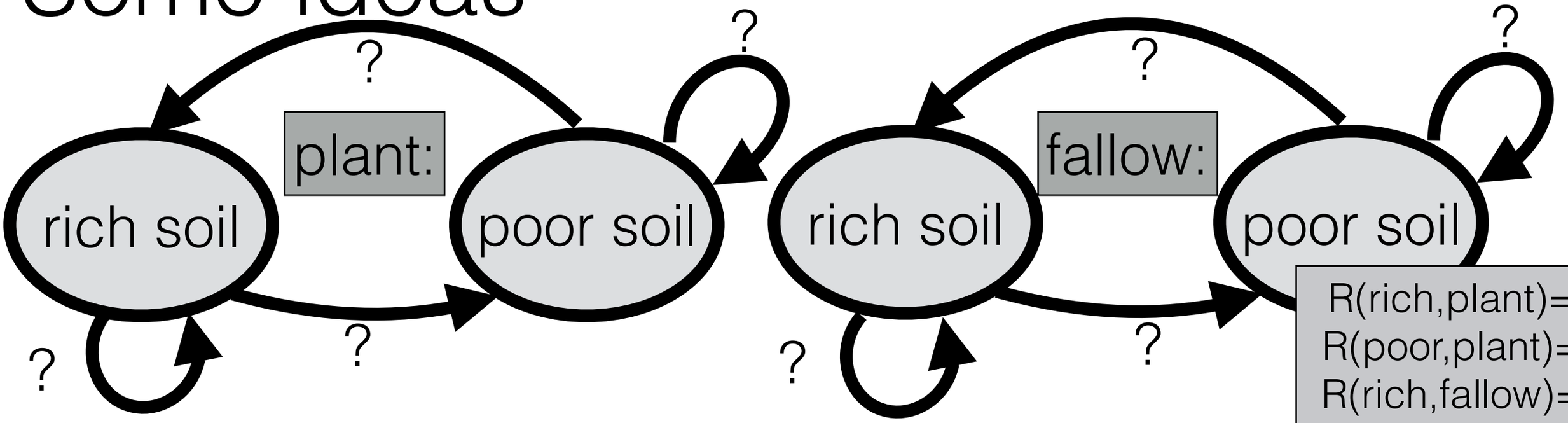
- Goal: Harvest as much as possible
- Strategy A

Some ideas



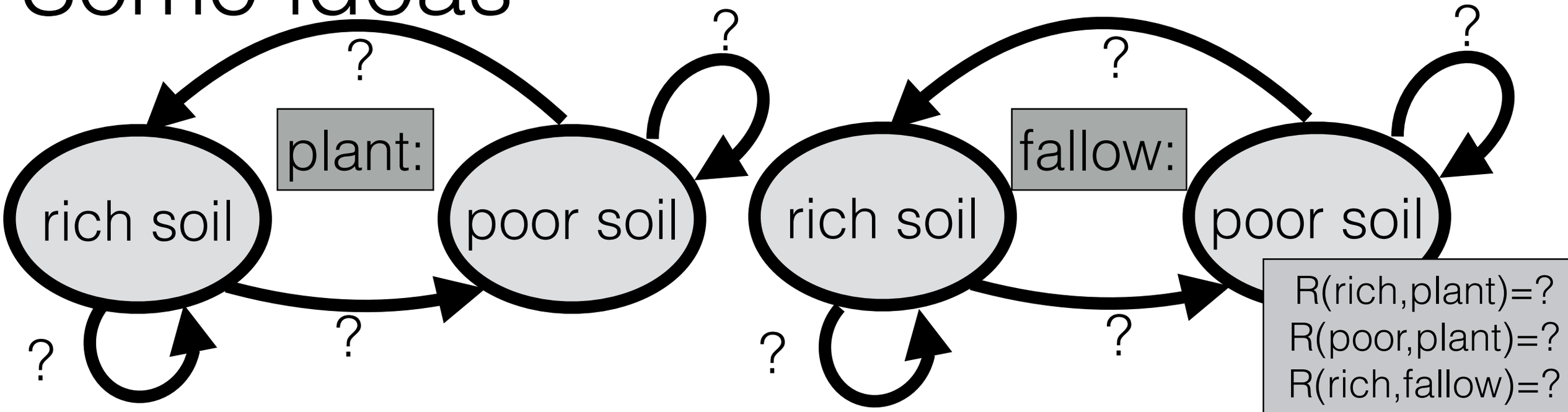
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best

Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
- Example of this strategy in action:

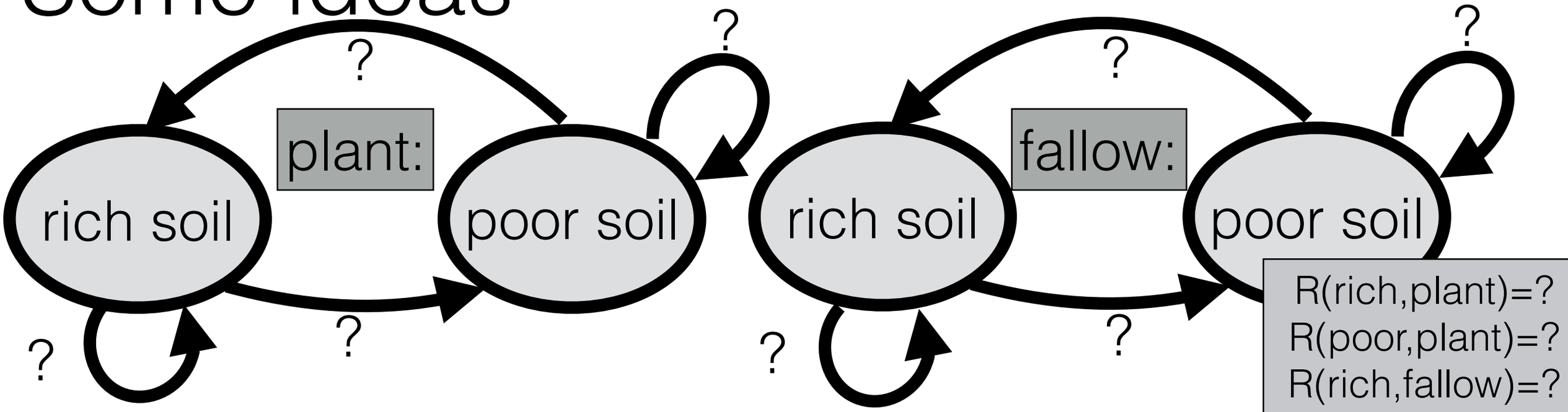
Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
- Example of this strategy in action:

$s^{(1)} = \text{rich}$

Some ideas

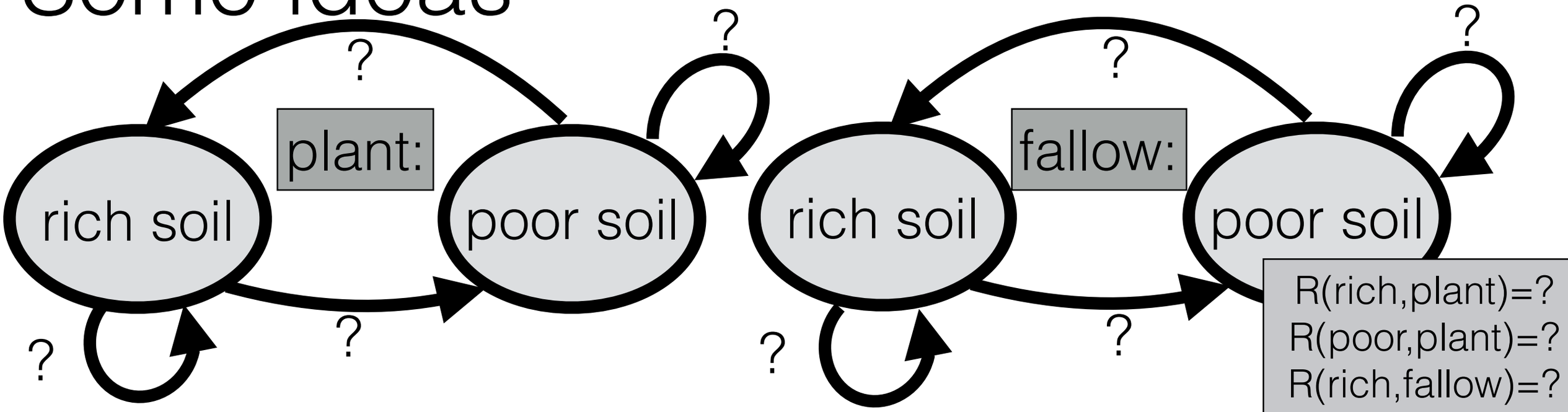


- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
- Example of this strategy in action:

$s^{(1)} = \text{rich}$

$a^{(1)} =$

Some ideas

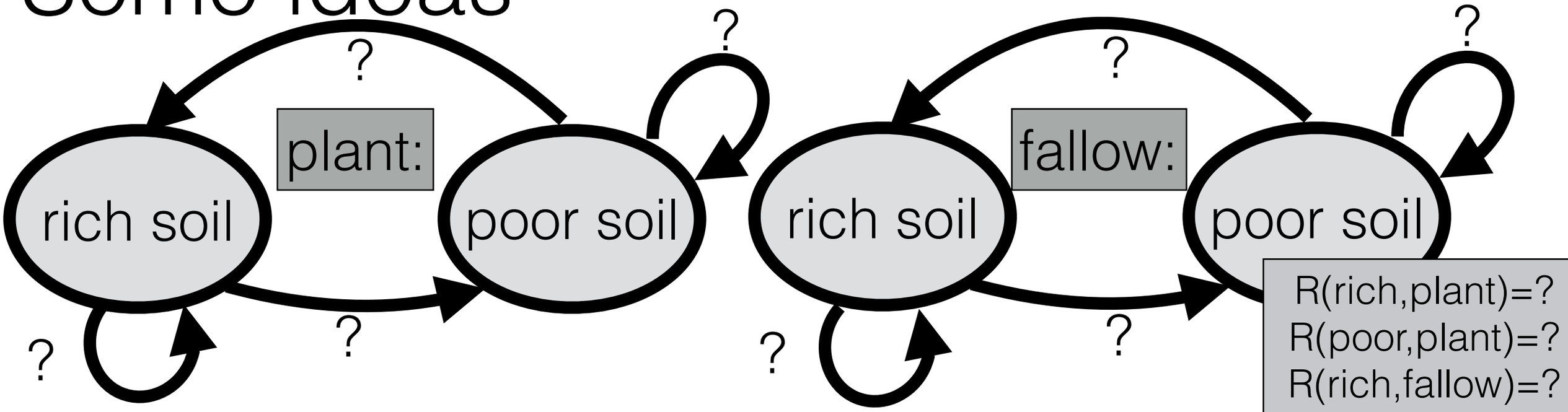


- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
- Example of this strategy in action:

$s^{(1)} = \text{rich}$

$a^{(1)} = \text{plant}$

Some ideas

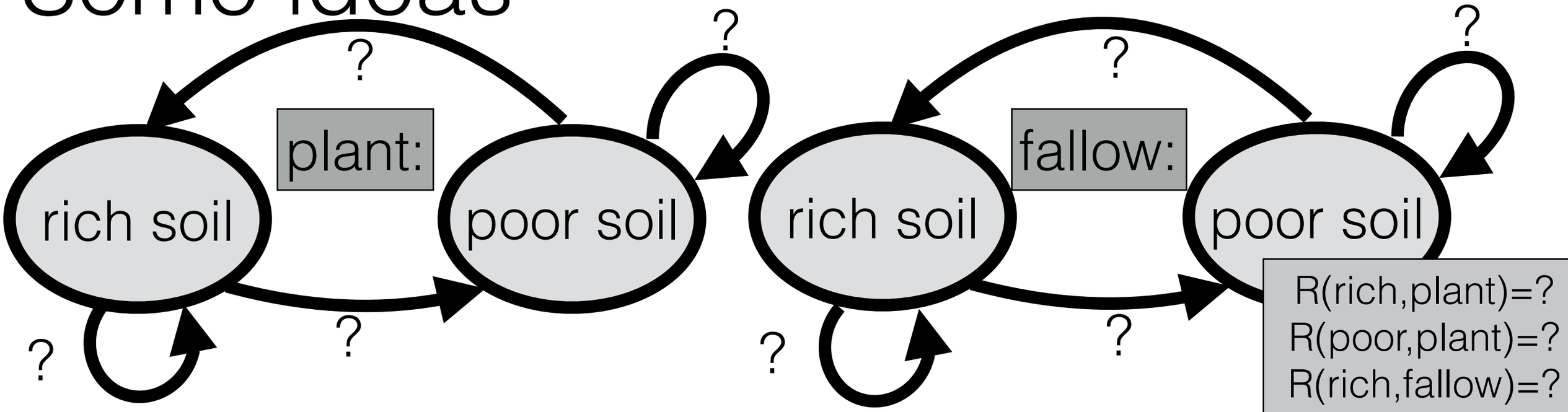


- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
- Example of this strategy in action:

$s^{(1)} = \text{rich}$

$a^{(1)} = \text{plant}; s^{(2)} =$

Some ideas

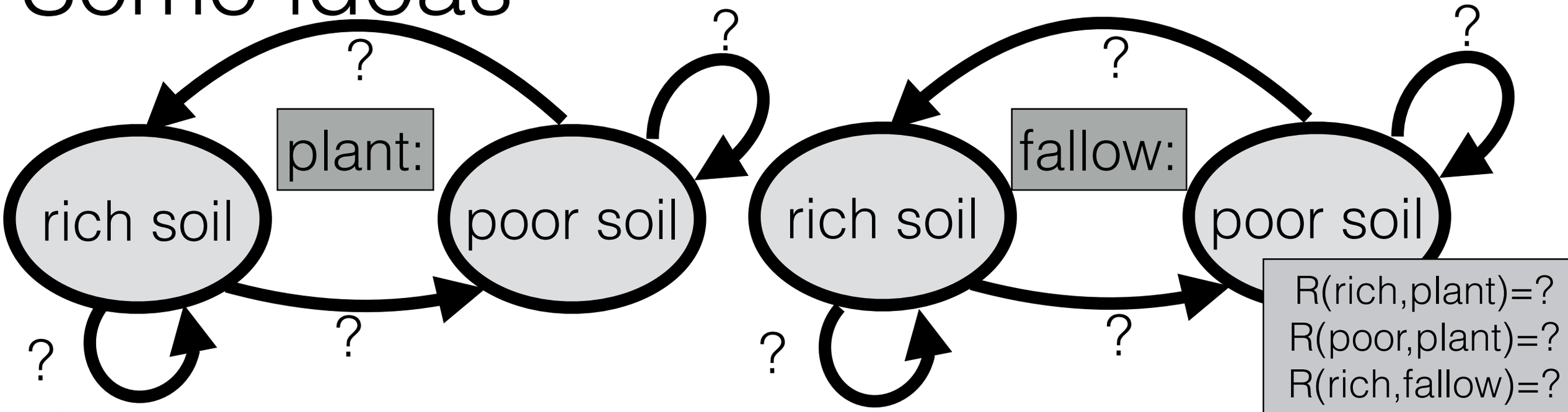


- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
- Example of this strategy in action:

$s^{(1)} = \text{rich}$

$a^{(1)} = \text{plant}; s^{(2)} = \text{rich}$

Some ideas

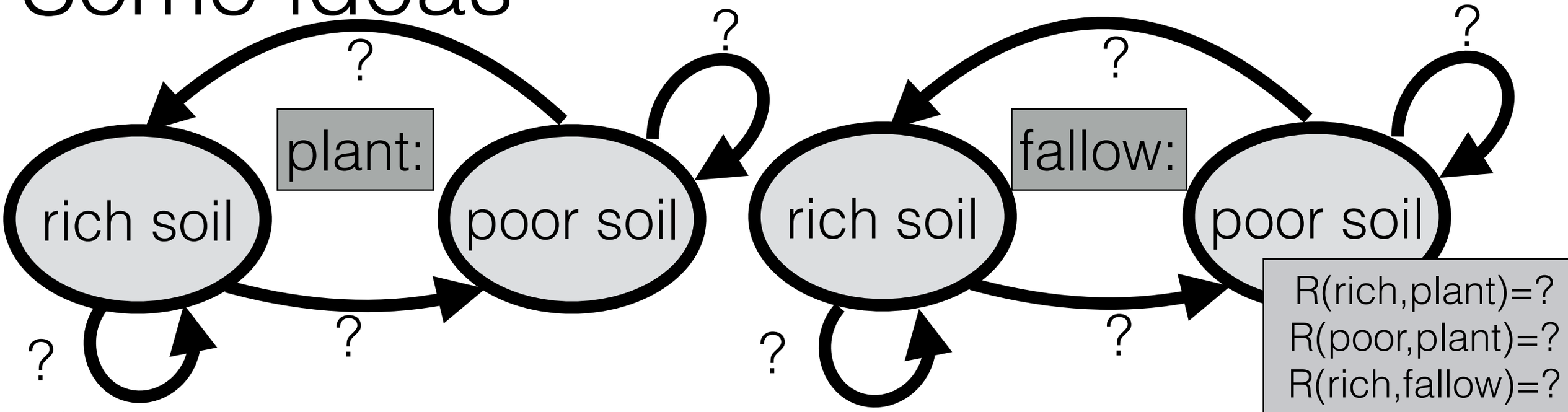


- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
- Example of this strategy in action:

$s^{(1)} = \text{rich}$

$a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} =$

Some ideas

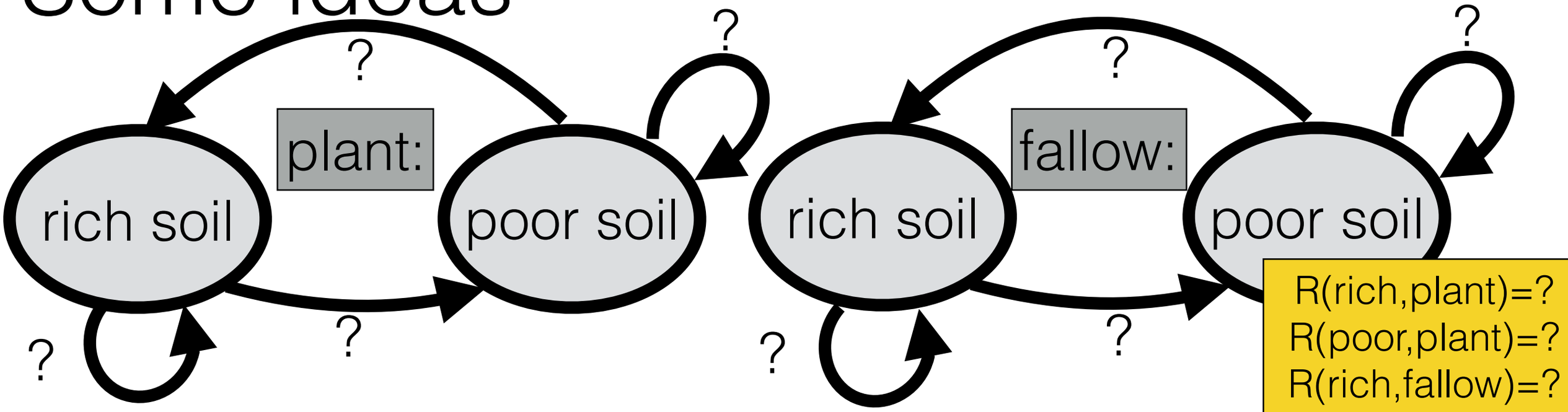


- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
- Example of this strategy in action:

$s^{(1)} = \text{rich}$

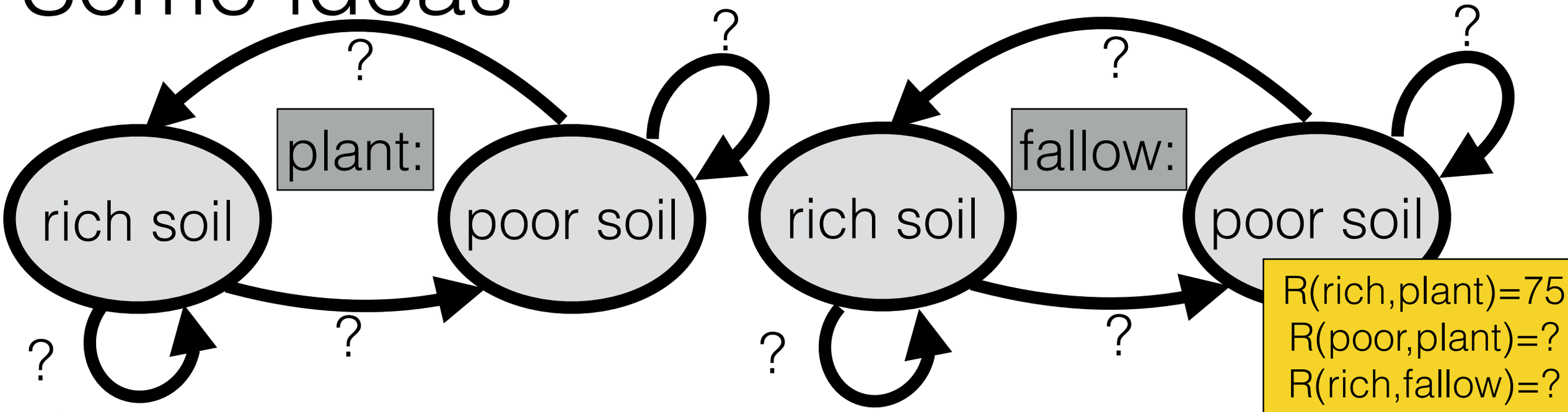
$a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

Some ideas



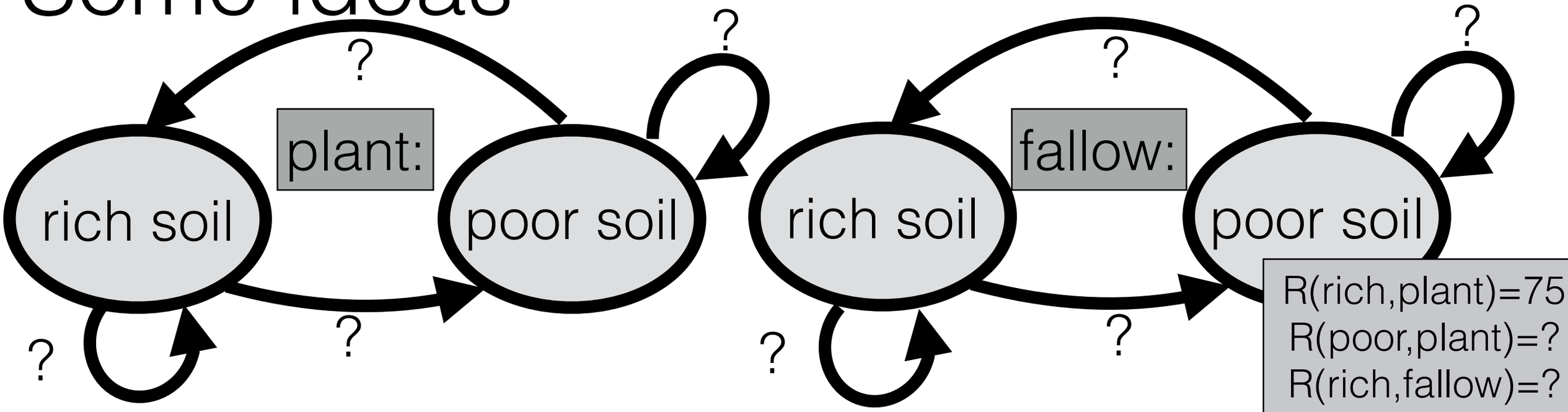
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
- Example of this strategy in action:
  $s^{(1)} = \text{rich}$
  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
- Example of this strategy in action:
  $s^{(1)} = \text{rich}$
  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

Some ideas



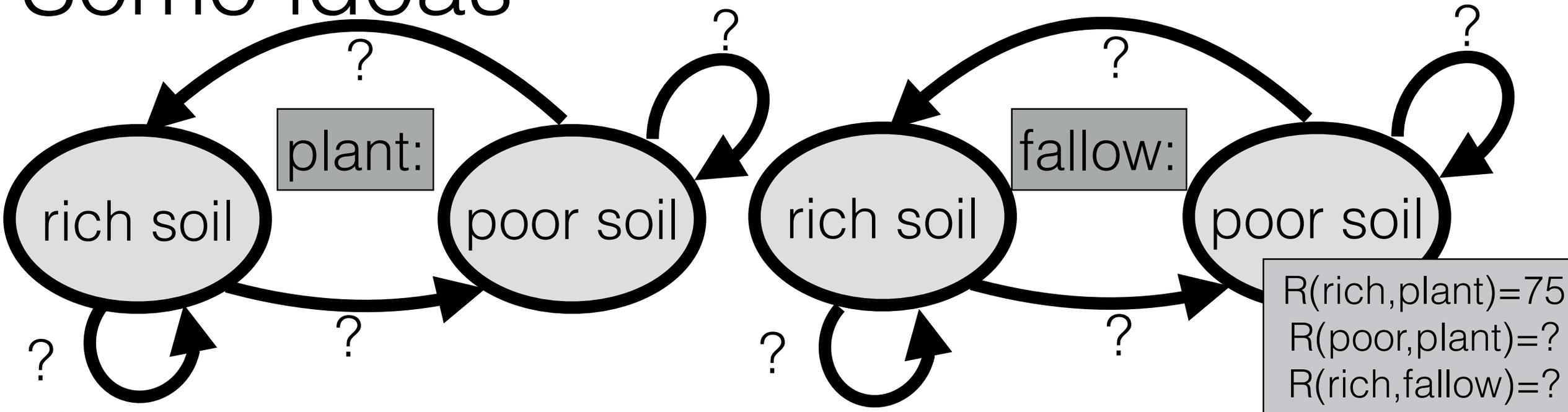
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
- Example of this strategy in action:

$s^{(1)} = \text{rich}$

$a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

- Strategy from here: always plant!

Some ideas



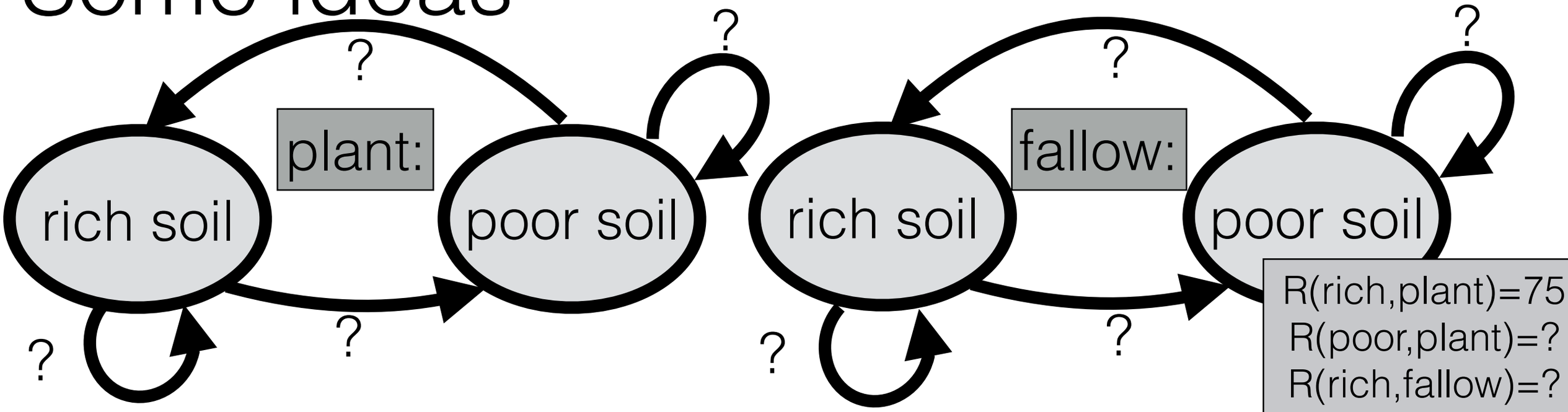
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
- Example of this strategy in action:

  $s^{(1)} = \text{rich}$

  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
  - Strategy from here: always plant!

Is this a good strategy?

Some ideas



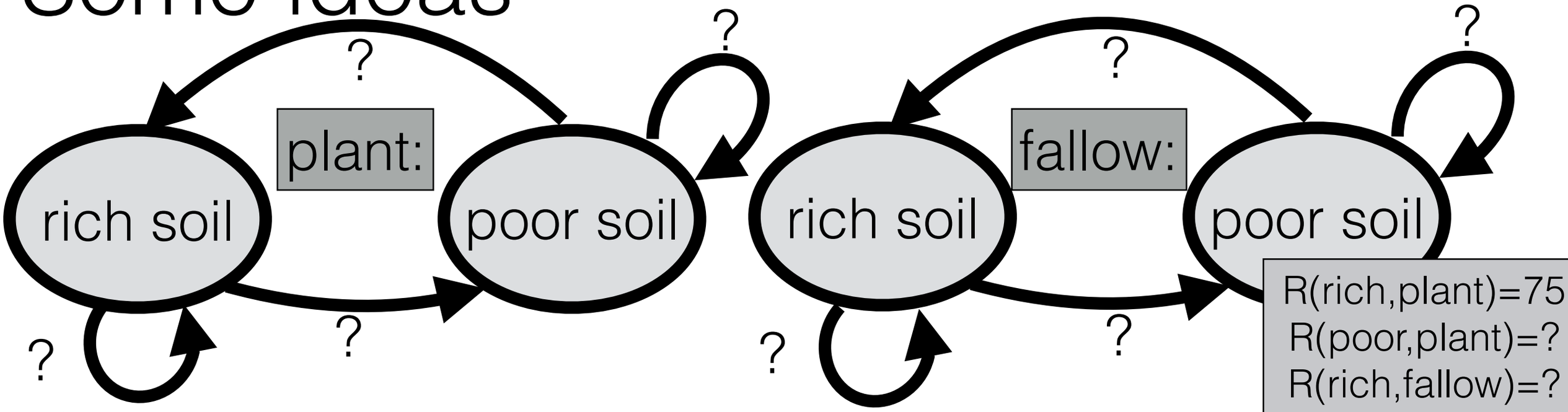
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:

    $s^{(1)} = \text{rich}$

    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!

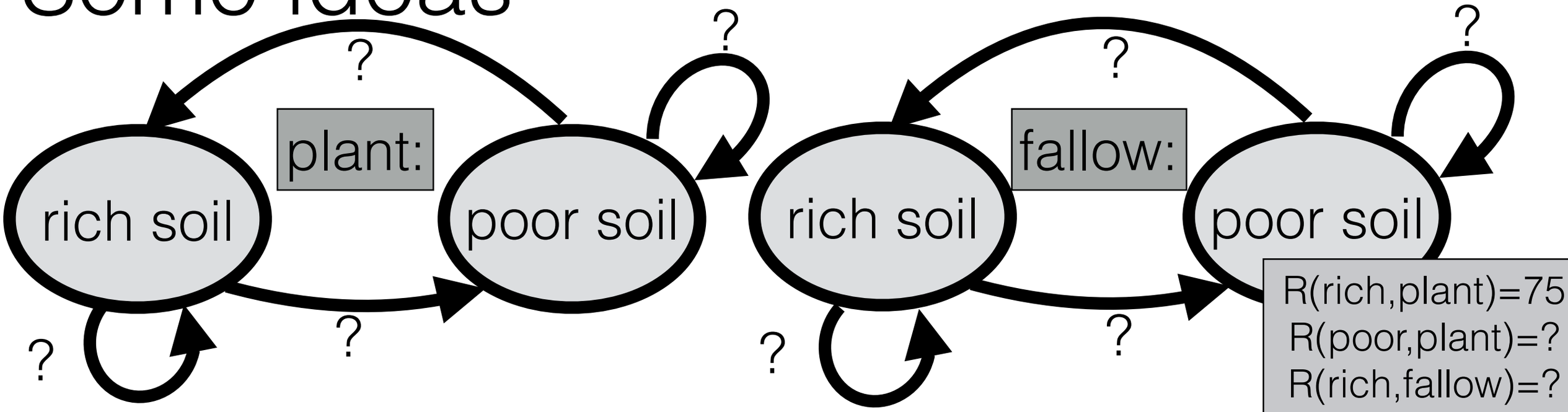
Could anything be improved about this strategy?

Some ideas



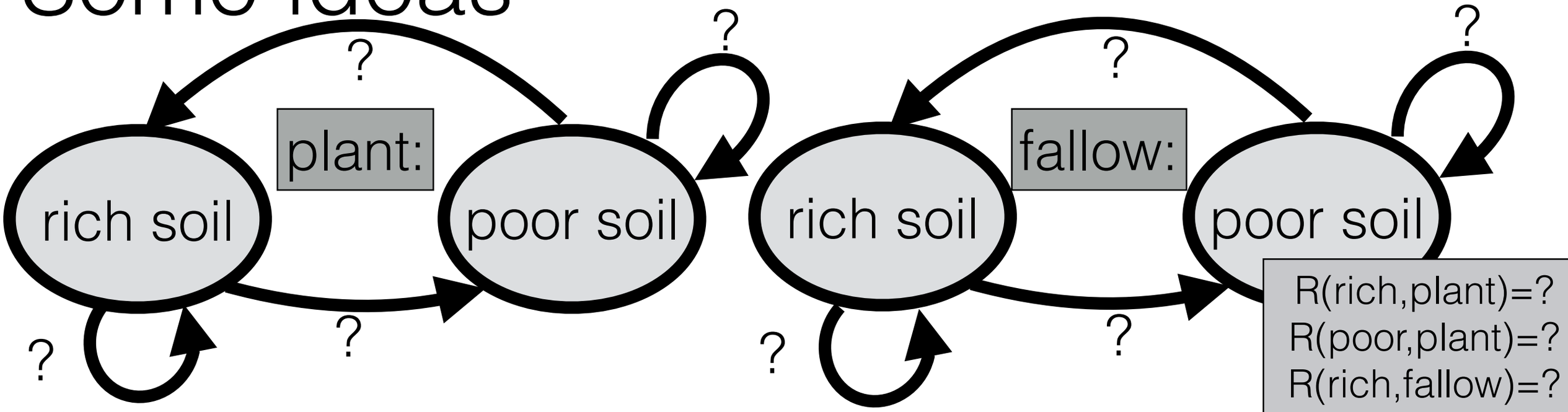
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B

Some ideas



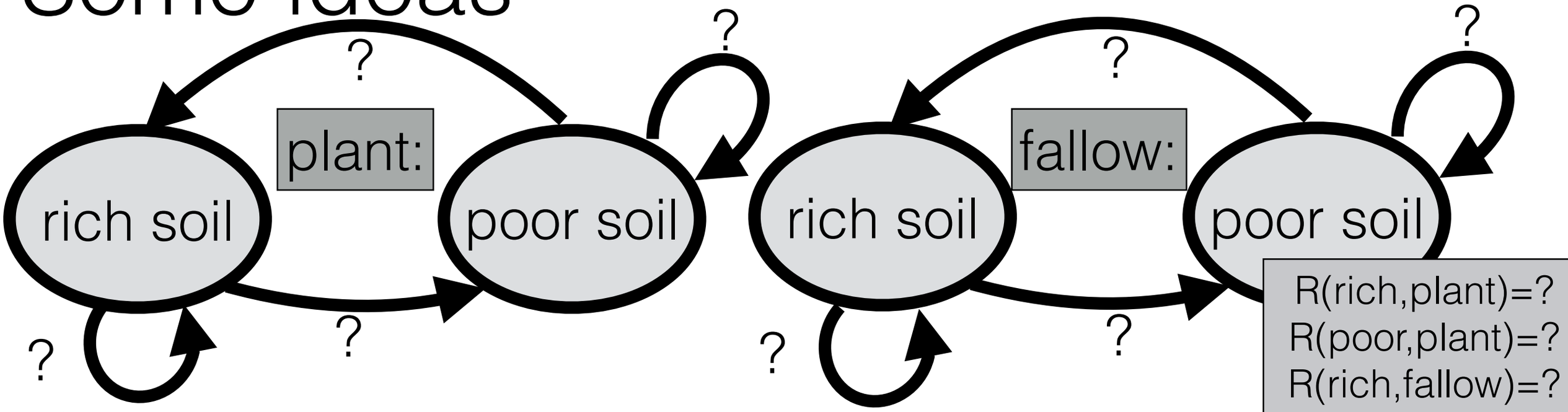
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random

Some ideas



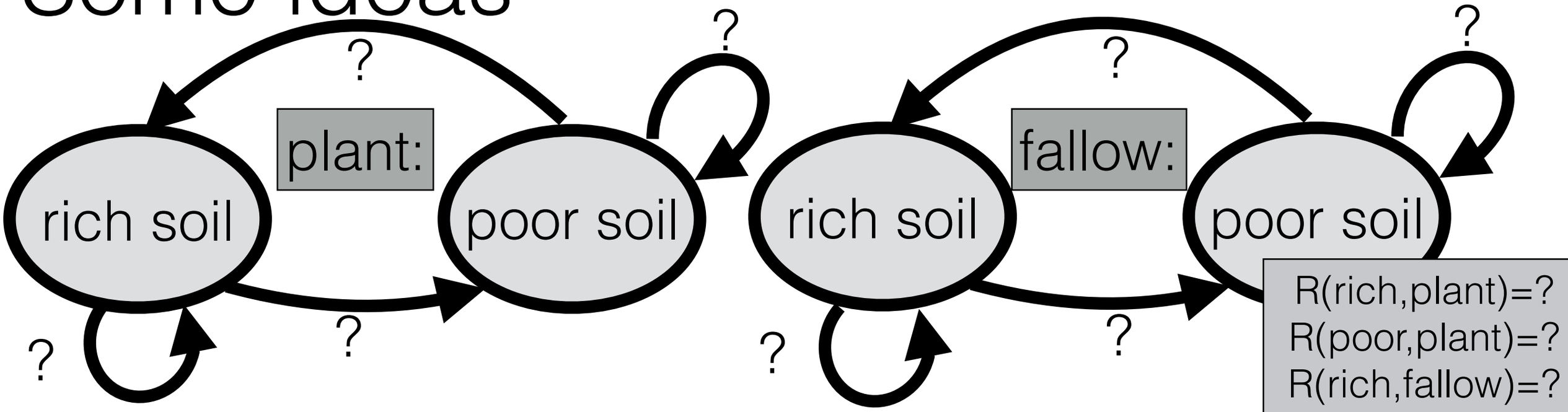
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random

Some ideas



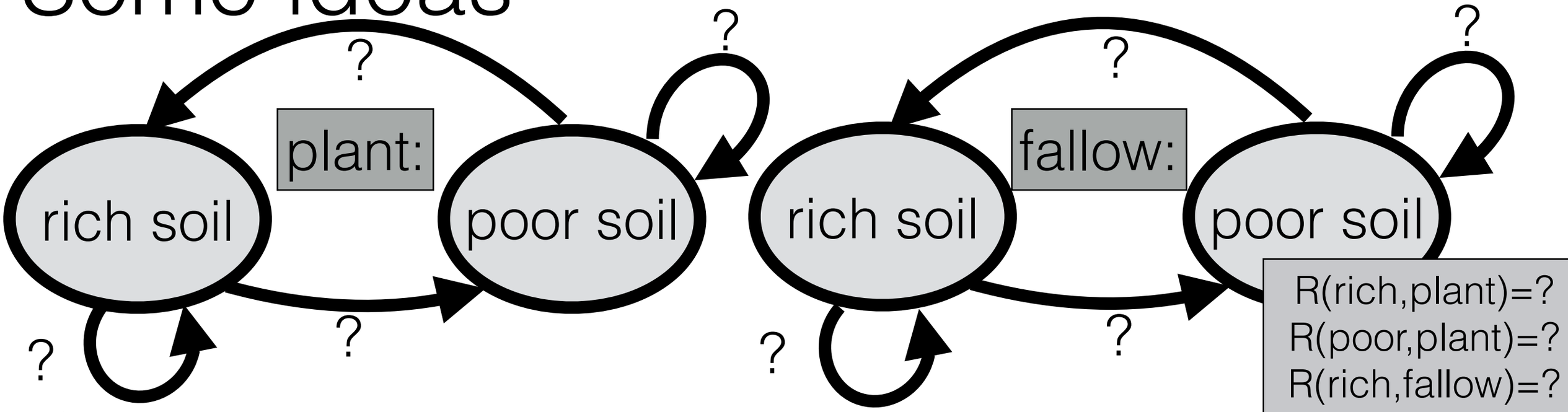
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random
  $s^{(1)} = \text{rich}$

Some ideas



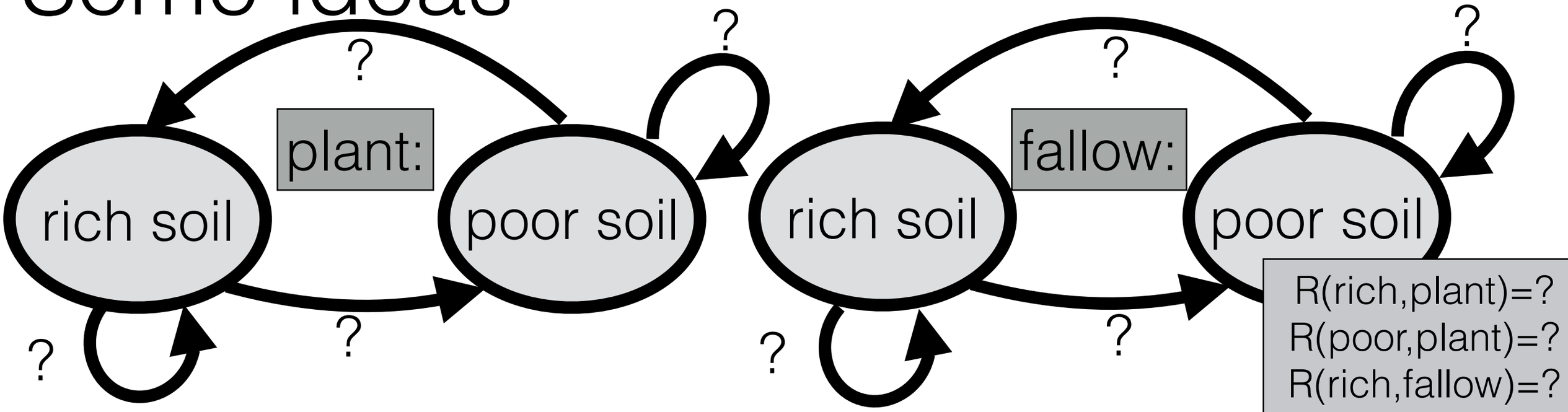
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random
  $s^{(1)} = \text{rich}$
  $a^{(1)} = \text{plant}$

Some ideas



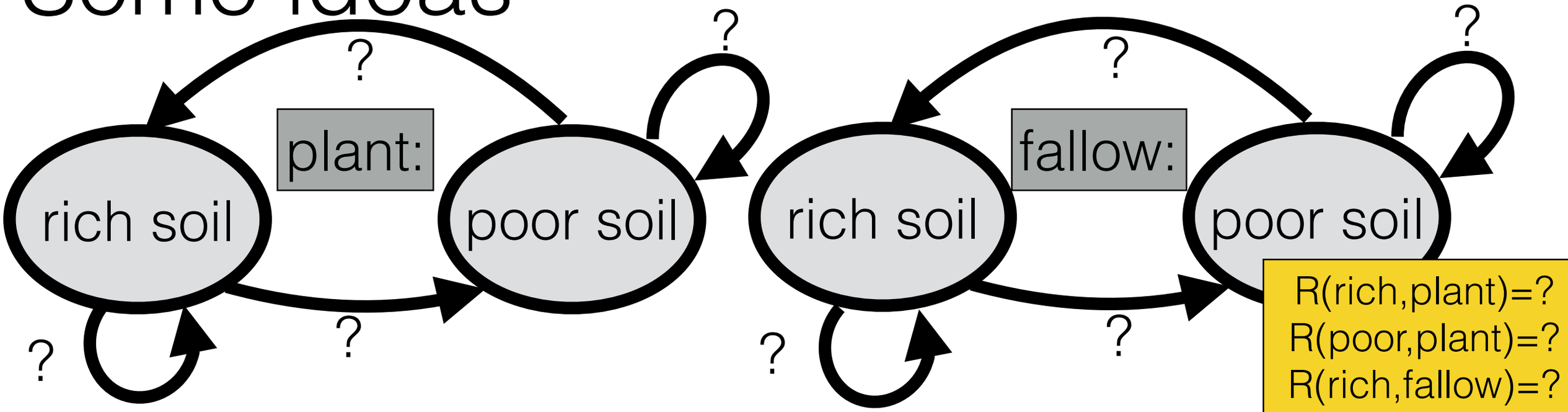
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}$

Some ideas



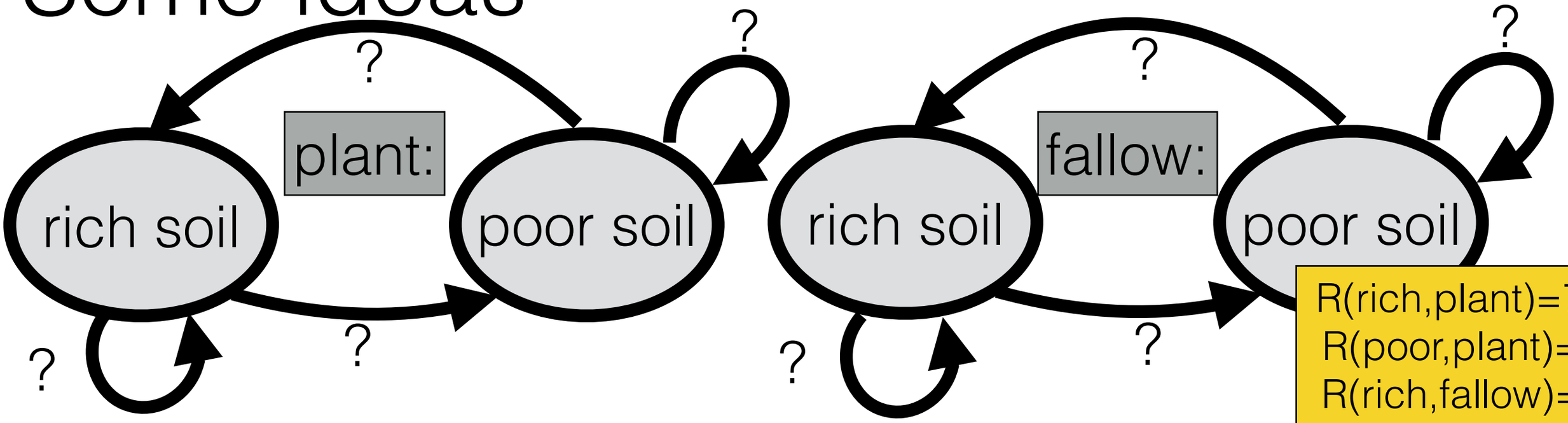
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

Some ideas



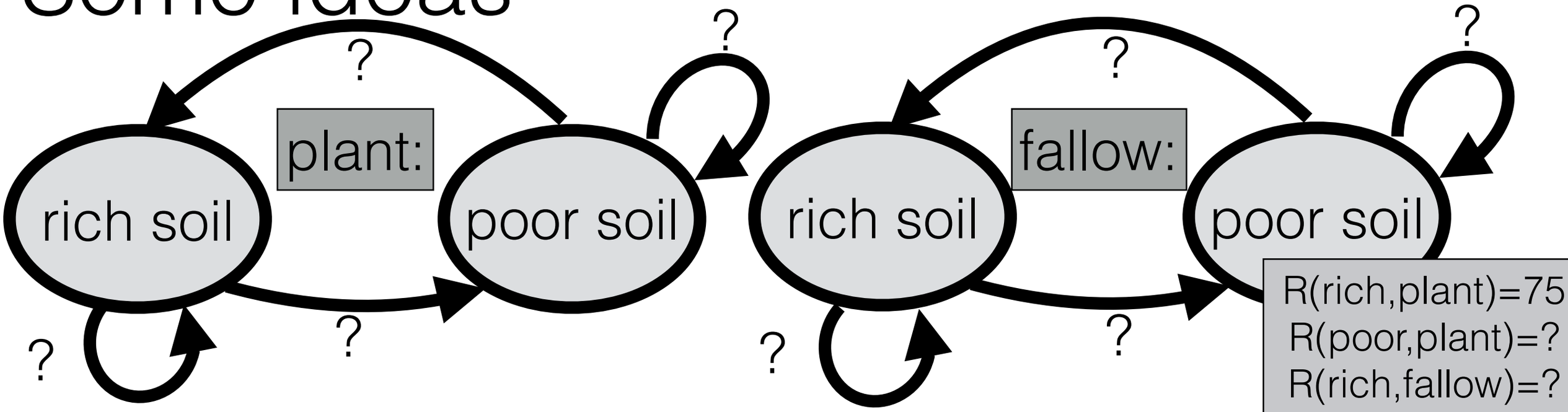
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:

    $s^{(1)} = \text{rich}$

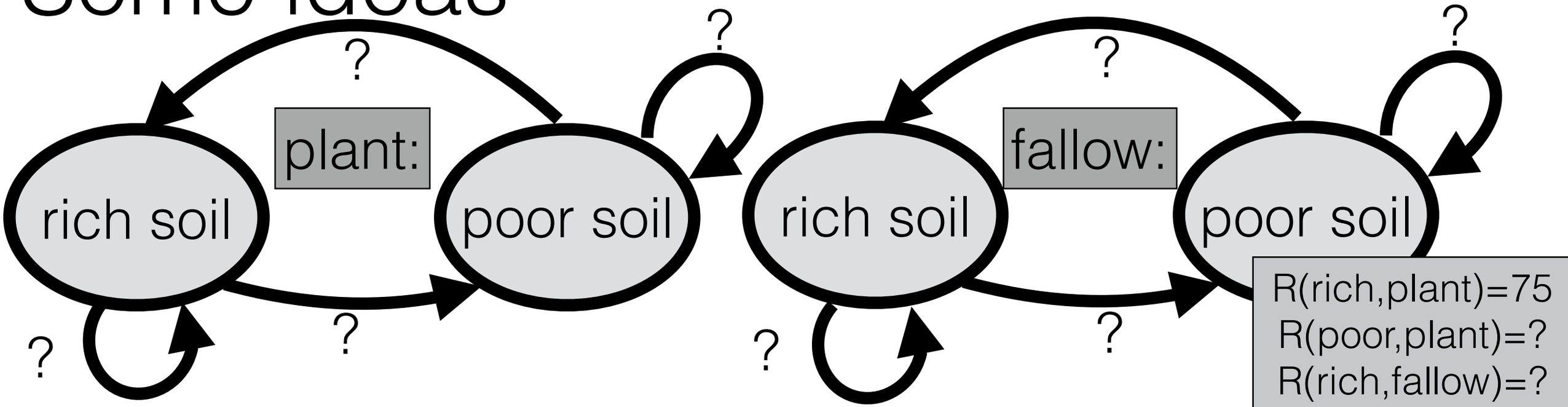
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random

  $s^{(1)} = \text{rich}$

  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

  $a^{(2)} = \text{plant}$

Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:

    $s^{(1)} = \text{rich}$

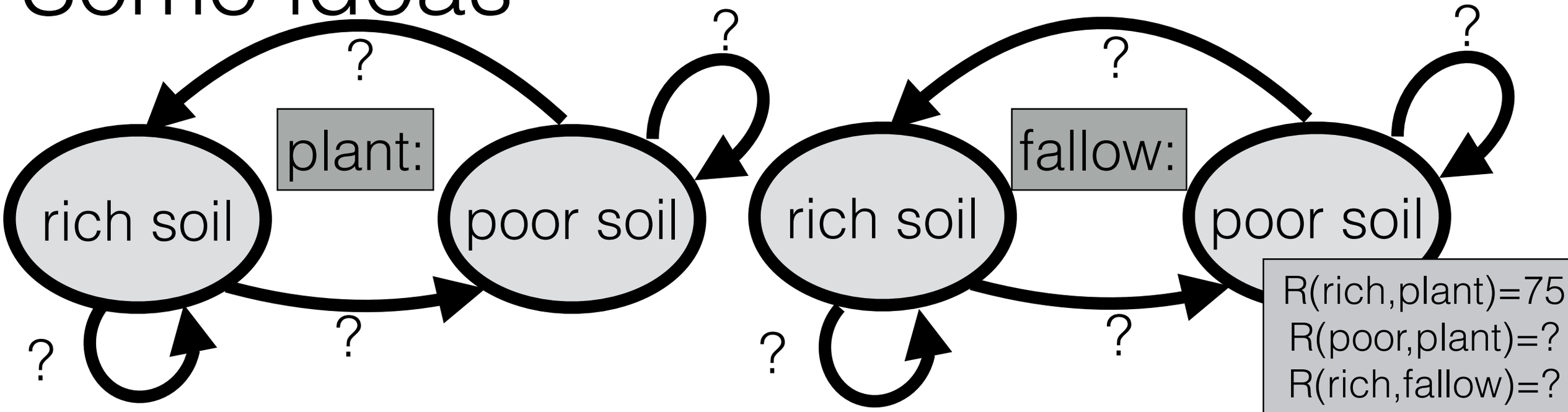
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random

  $s^{(1)} = \text{rich}$

  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

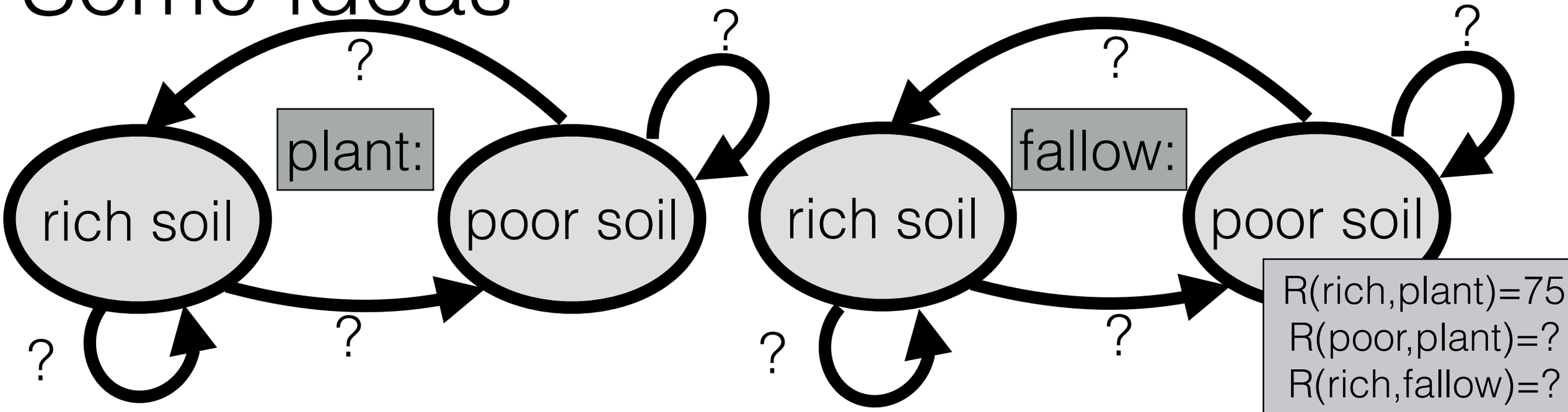
  $a^{(2)} = \text{plant}; s^{(3)} = \text{poor}$

Some ideas



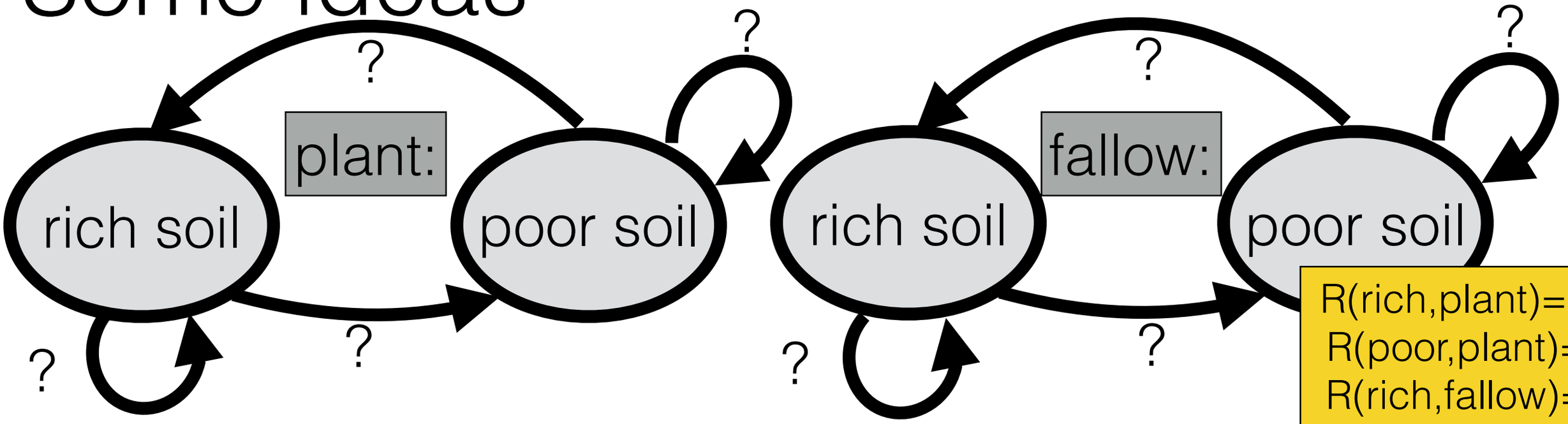
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    $a^{(2)} = \text{plant}; s^{(3)} = \text{poor}; r^{(2)} =$

Some ideas



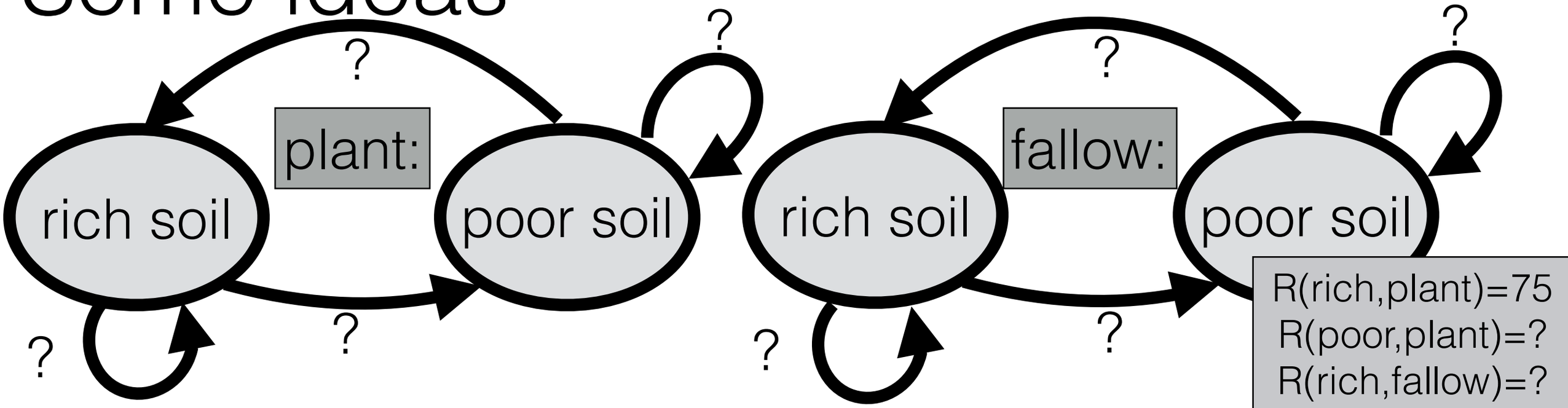
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random
  $s^{(1)} = \text{rich}$
  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
  $a^{(2)} = \text{plant}; s^{(3)} = \text{poor}; r^{(2)} = 75 \text{ bushels}$

Some ideas



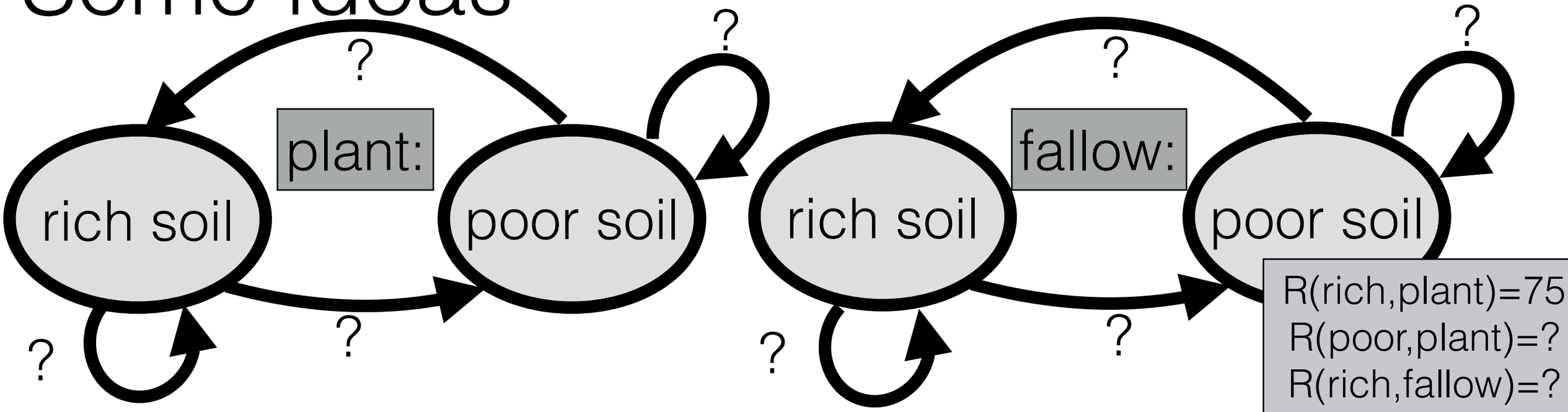
- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random
  $s^{(1)} = \text{rich}$
  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
  $a^{(2)} = \text{plant}; s^{(3)} = \text{poor}; r^{(2)} = 75 \text{ bushels}$

Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:
    $s^{(1)} = \text{rich}$
    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random
  $s^{(1)} = \text{rich}$
  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
  $a^{(2)} = \text{plant}; s^{(3)} = \text{poor}; r^{(2)} = 75 \text{ bushels}$

Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:

    $s^{(1)} = \text{rich}$

    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random

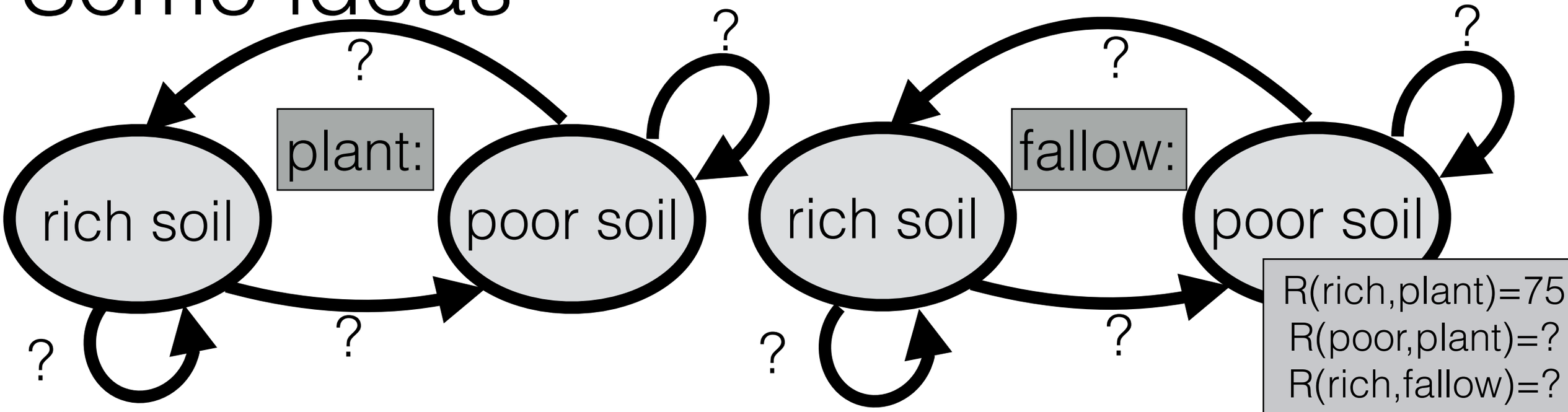
  $s^{(1)} = \text{rich}$

  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

  $a^{(2)} = \text{plant}; s^{(3)} = \text{poor}; r^{(2)} = 75 \text{ bushels}$

  $a^{(3)} = \text{plant}$

Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:

    $s^{(1)} = \text{rich}$

    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random

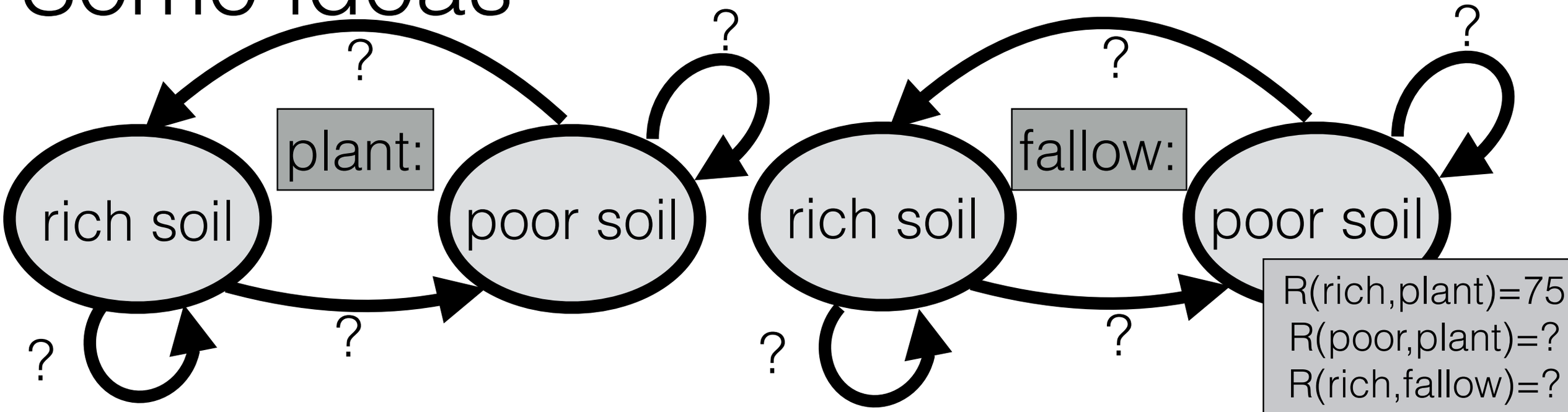
  $s^{(1)} = \text{rich}$

  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

  $a^{(2)} = \text{plant}; s^{(3)} = \text{poor}; r^{(2)} = 75 \text{ bushels}$

  $a^{(3)} = \text{plant}; s^{(4)} = \text{poor}$

Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:

    $s^{(1)} = \text{rich}$

    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random

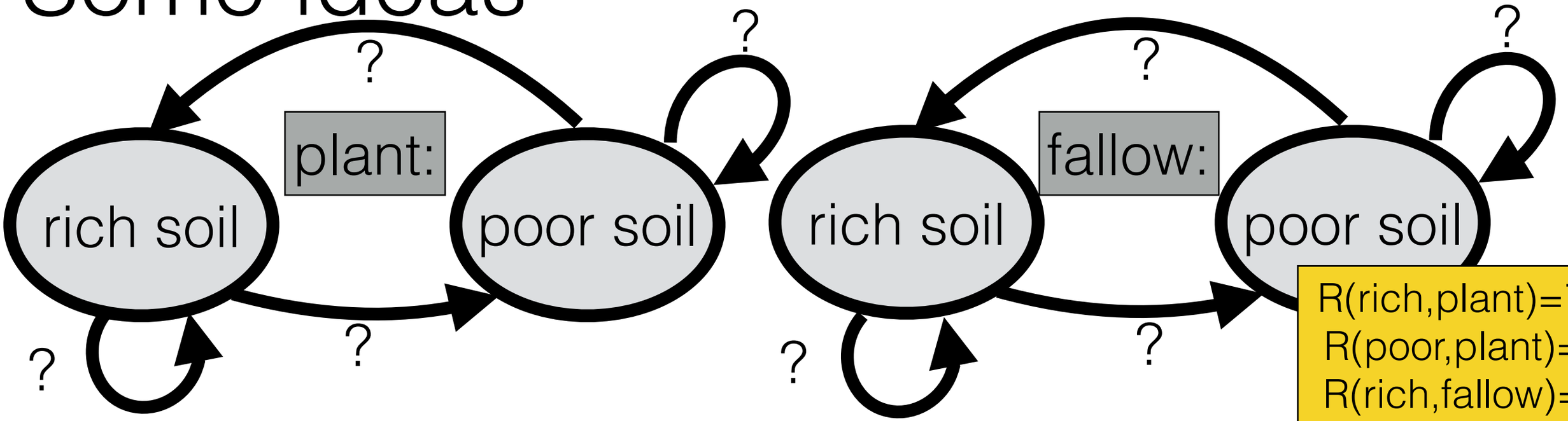
  $s^{(1)} = \text{rich}$

  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

  $a^{(2)} = \text{plant}; s^{(3)} = \text{poor}; r^{(2)} = 75 \text{ bushels}$

  $a^{(3)} = \text{plant}; s^{(4)} = \text{poor}; r^{(3)} = 2 \text{ bushels}$

Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:

    $s^{(1)} = \text{rich}$

    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random

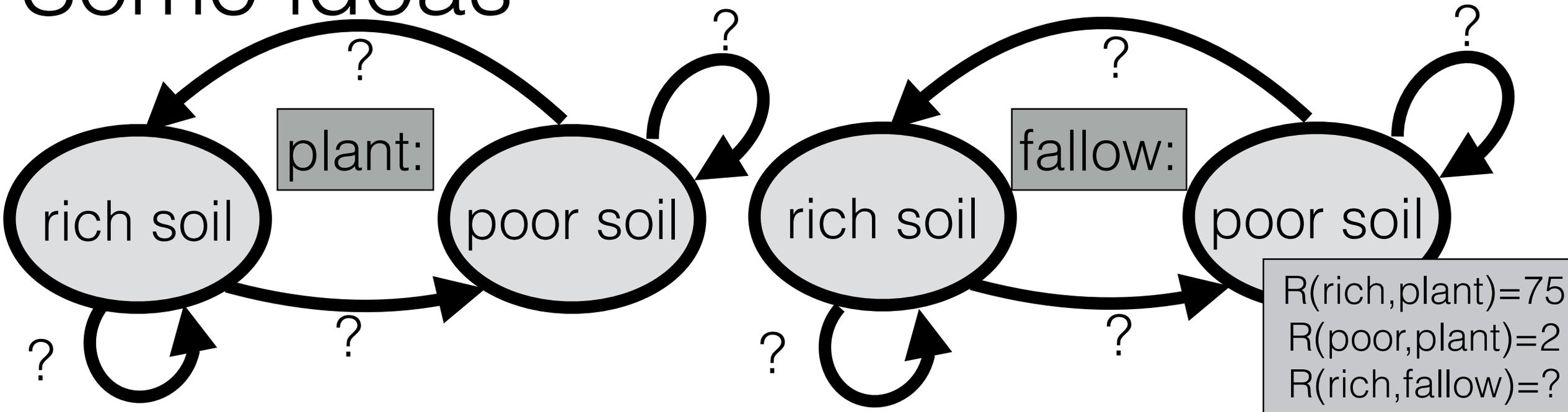
  $s^{(1)} = \text{rich}$

  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

  $a^{(2)} = \text{plant}; s^{(3)} = \text{poor}; r^{(2)} = 75 \text{ bushels}$

  $a^{(3)} = \text{plant}; s^{(4)} = \text{poor}; r^{(3)} = 2 \text{ bushels}$

Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:

    $s^{(1)} = \text{rich}$

    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random

  $s^{(1)} = \text{rich}$

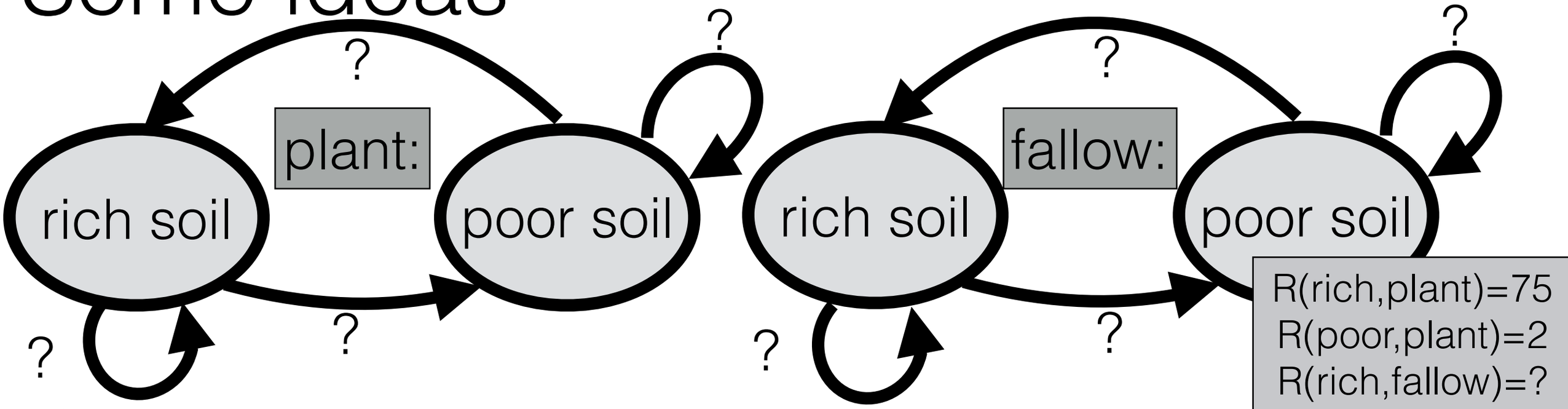
  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

  $a^{(2)} = \text{plant}; s^{(3)} = \text{poor}; r^{(2)} = 75 \text{ bushels}$

  $a^{(3)} = \text{plant}; s^{(4)} = \text{poor}; r^{(3)} = 2 \text{ bushels}$

Is this a good strategy?

Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:

    $s^{(1)} = \text{rich}$

    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random

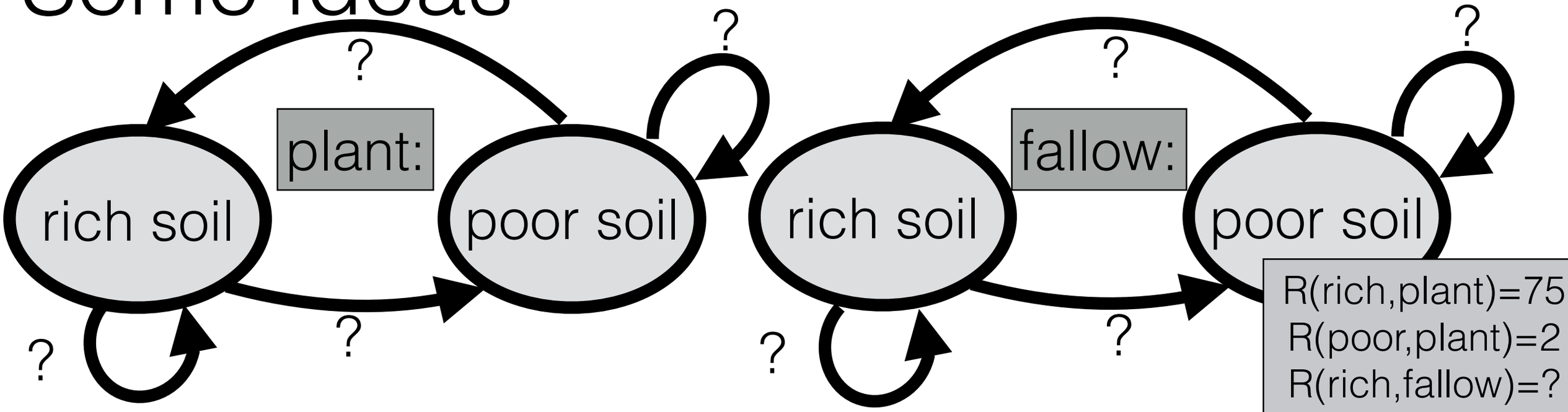
  $s^{(1)} = \text{rich}$

  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

  $a^{(2)} = \text{plant}; s^{(3)} = \text{poor}; r^{(2)} = 75 \text{ bushels}$

  $a^{(3)} = \text{plant}; s^{(4)} = \text{poor}; r^{(3)} = 2 \text{ bushels}$

Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:

    $s^{(1)} = \text{rich}$

    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!
- Strategy B: always try actions uniformly at random

  $s^{(1)} = \text{rich}$

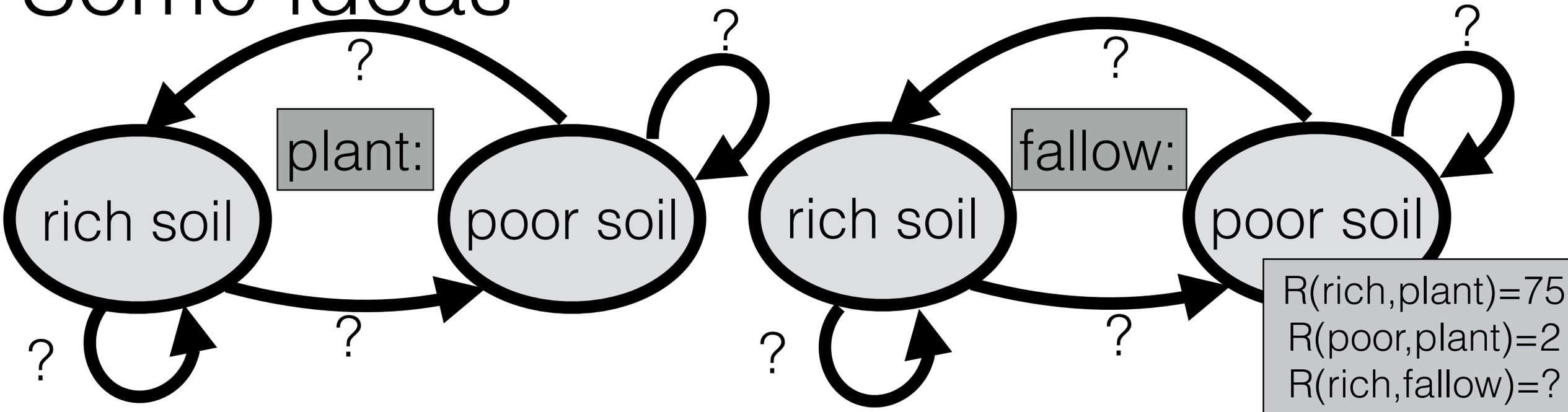
  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

  $a^{(2)} = \text{plant}; s^{(3)} = \text{poor}; r^{(2)} = 75 \text{ bushels}$

  $a^{(3)} = \text{plant}; s^{(4)} = \text{poor}; r^{(3)} = 2 \text{ bushels}$

Focused on exploring

Some ideas



- Goal: Harvest as much as possible
- Strategy A: try a random action then do what seems best
  - Example of this strategy in action:

    $s^{(1)} = \text{rich}$

    $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$
    - Strategy from here: always plant!

Focused on exploiting

- Strategy B: always try actions uniformly at random

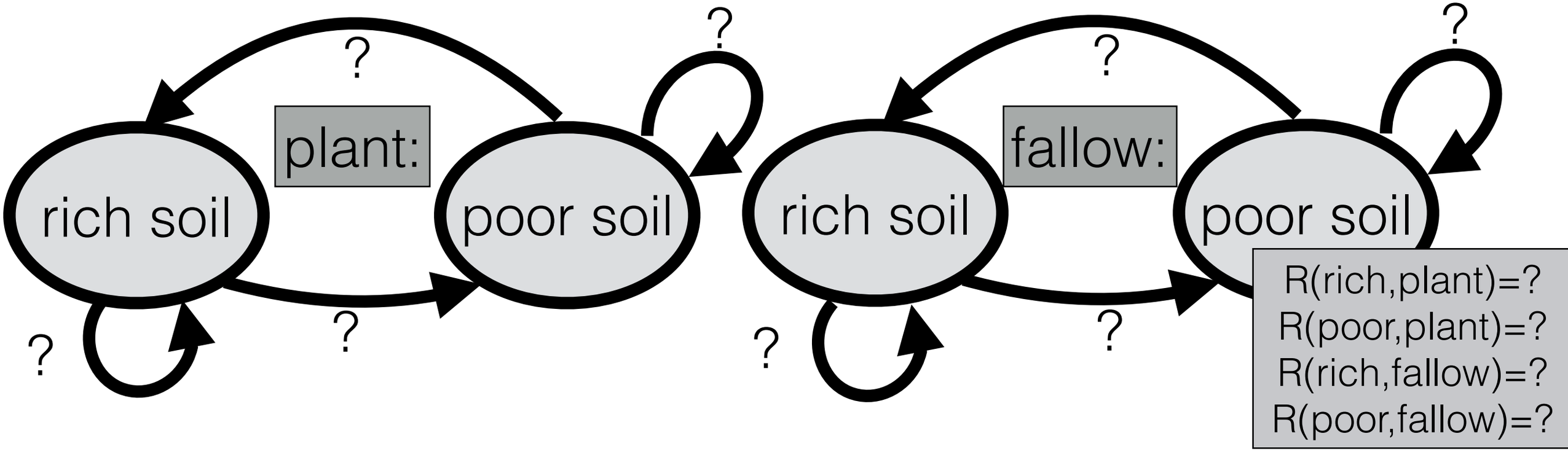
  $s^{(1)} = \text{rich}$

  $a^{(1)} = \text{plant}; s^{(2)} = \text{rich}; r^{(1)} = 75 \text{ bushels}$

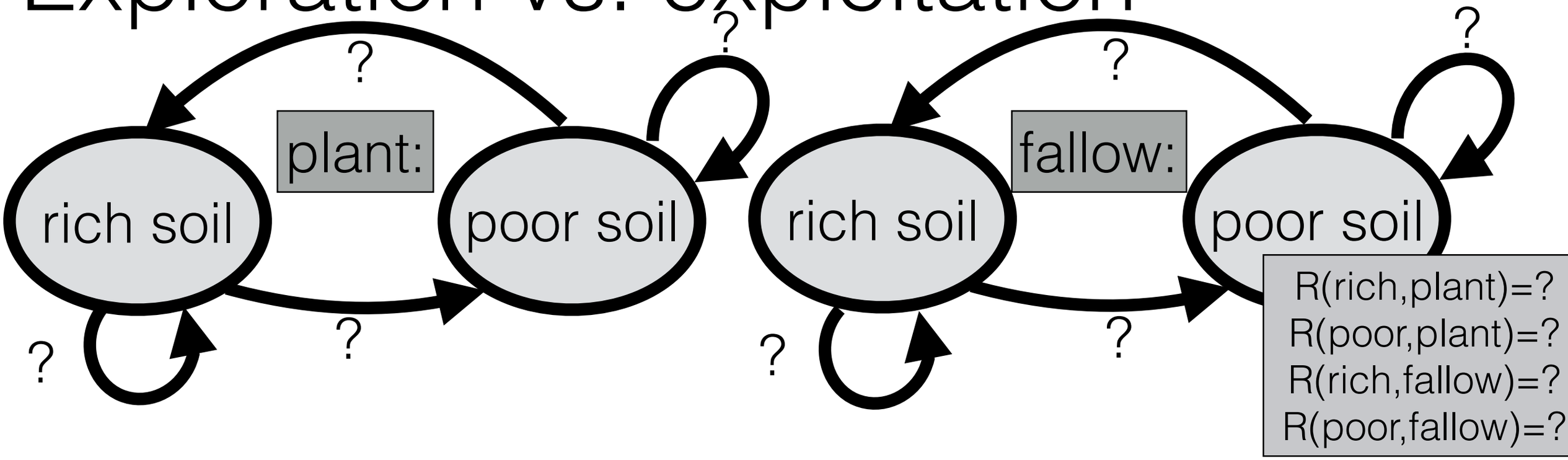
  $a^{(2)} = \text{plant}; s^{(3)} = \text{poor}; r^{(2)} = 75 \text{ bushels}$

  $a^{(3)} = \text{plant}; s^{(4)} = \text{poor}; r^{(3)} = 2 \text{ bushels}$

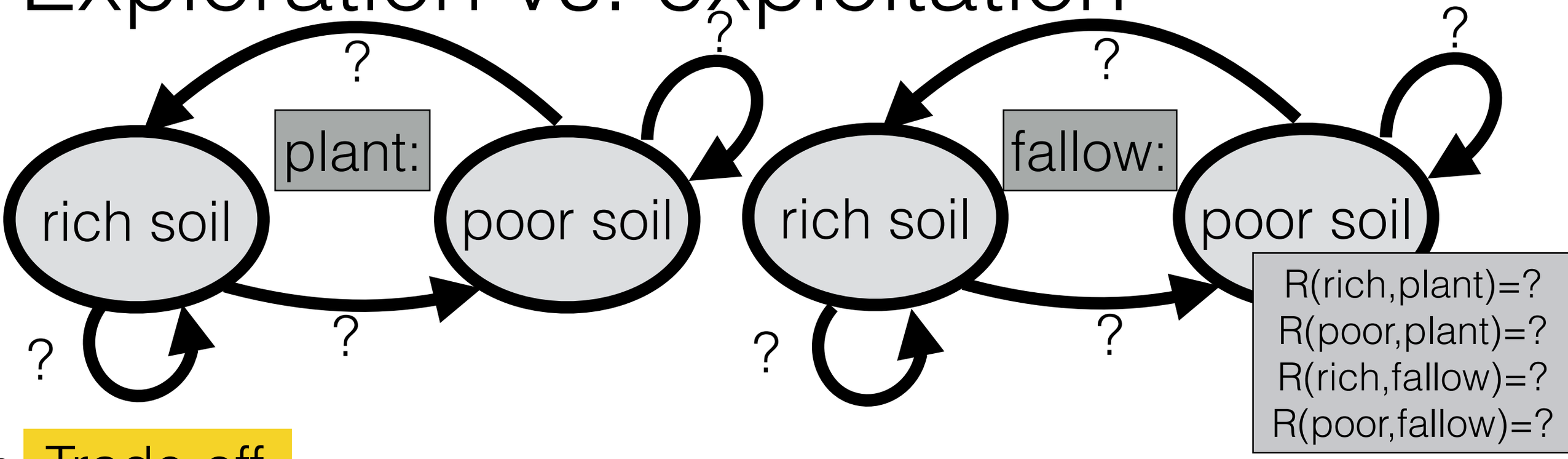
Focused on exploring


?
?
?
plant:
rich soil
poor soil
?
?
?
?
fallow:
rich soil
poor soil
?
?
R(rich,plant)=?
R(poor,plant)=?
R(rich,fallow)=?
R(poor,fallow)=?

Exploration vs. exploitation

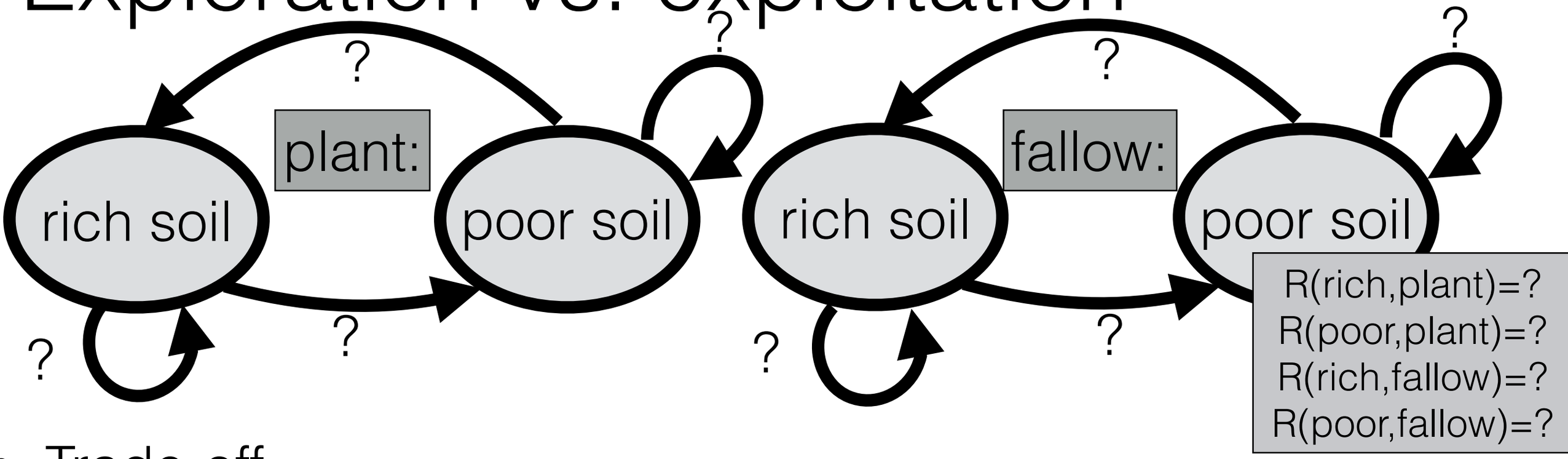


Exploration vs. exploitation



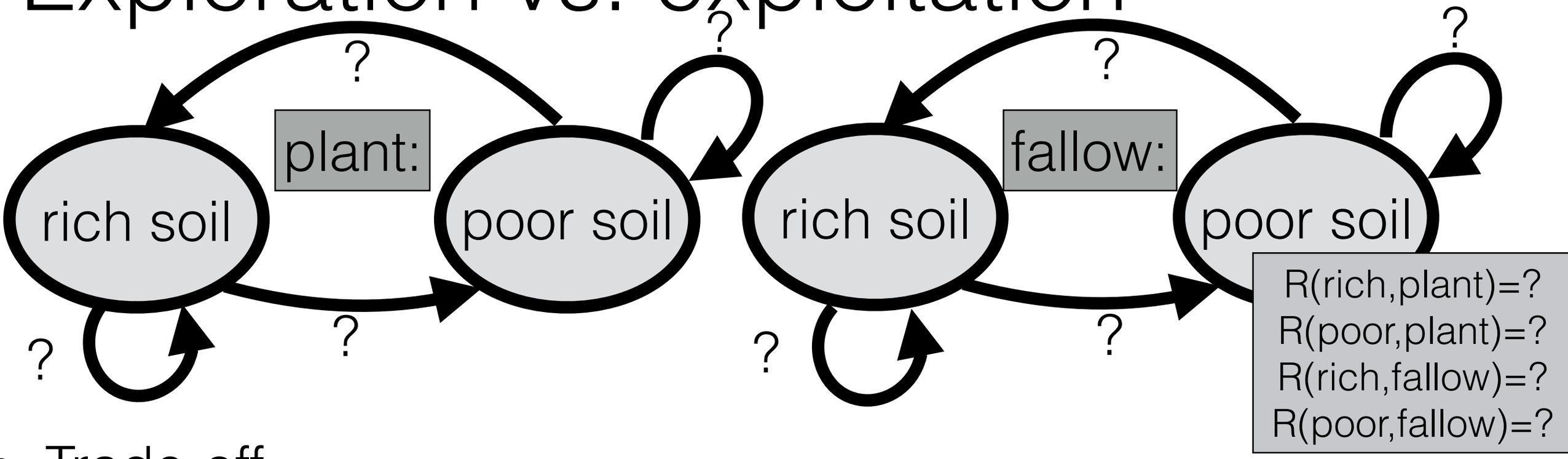
- Trade-off

Exploration vs. exploitation



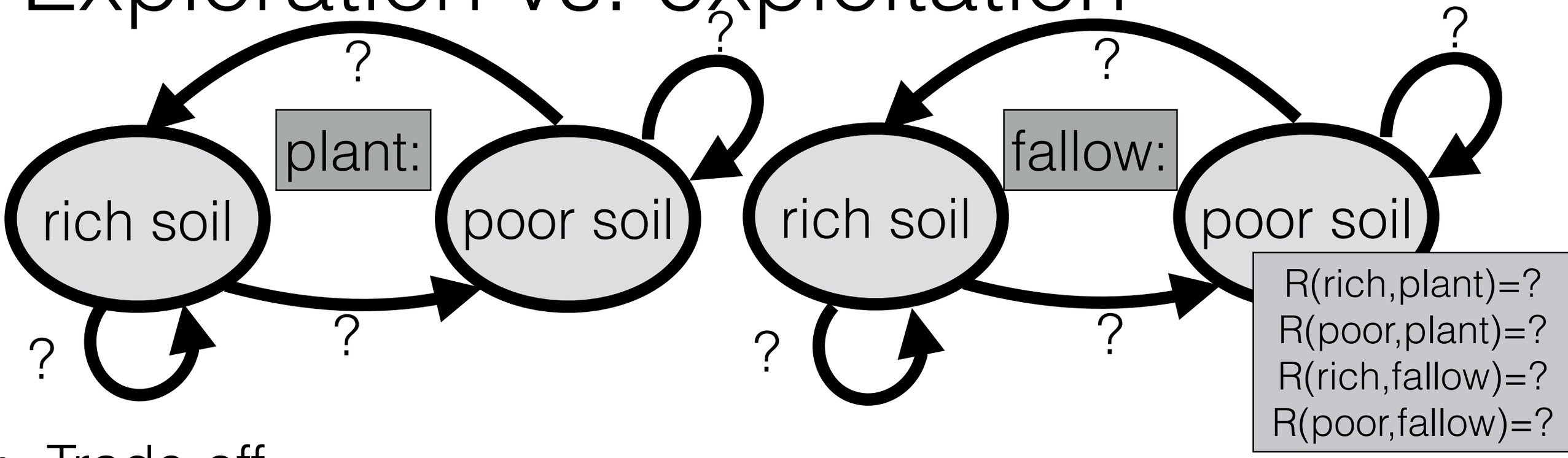
- Trade-off
- **Exploration**: the more we explore, the better we understand the world (e.g. *T* and *R*)

Exploration vs. exploitation



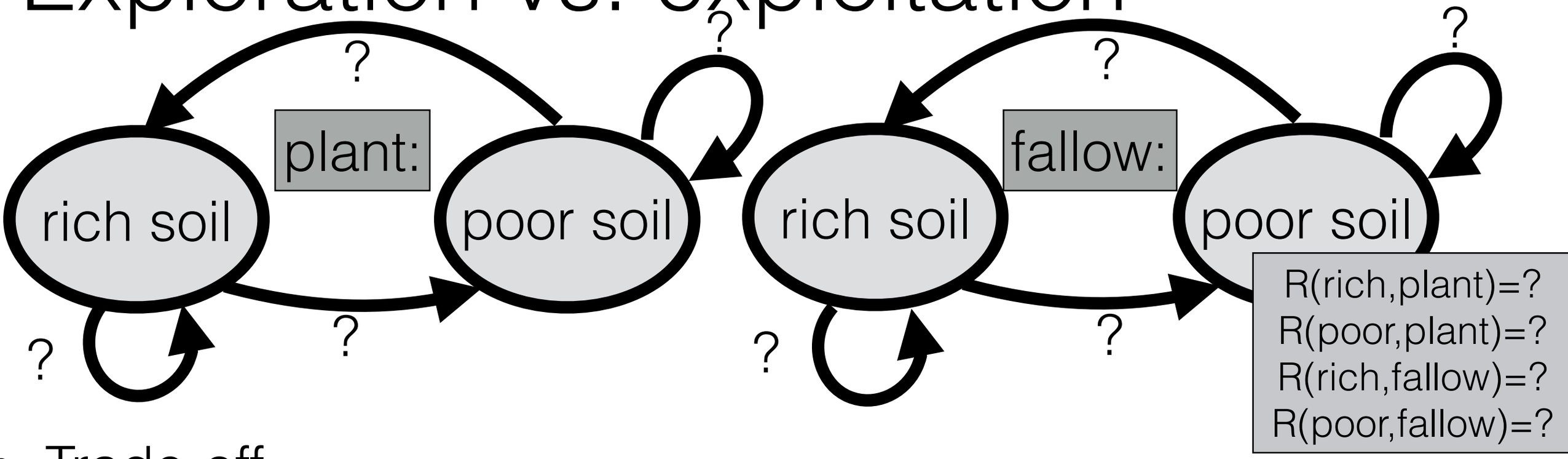
- Trade-off
- **Exploration**: the more we explore, the better we understand the world (e.g. *T* and *R*)
- **Exploitation**: based on what we know about the world, we can take actions with the aim to get highest reward

Exploration vs. exploitation



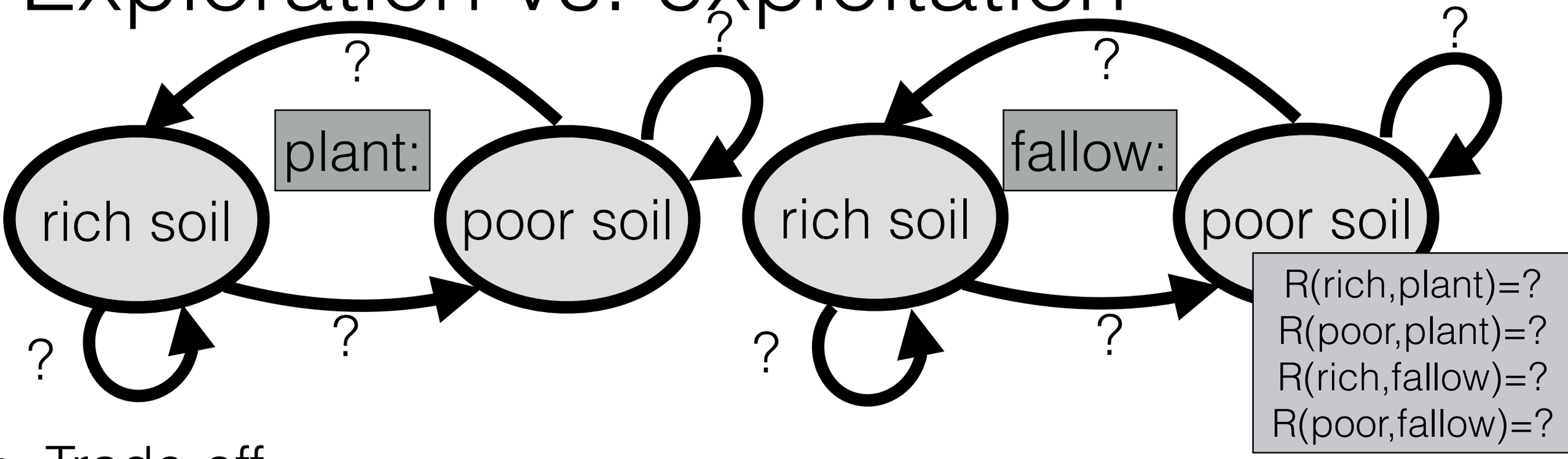
- Trade-off
  - **Exploration**: the more we explore, the better we understand the world (e.g. *T* and *R*)
  - **Exploitation**: based on what we know about the world, we can take actions with the aim to get highest reward
- One option

Exploration vs. exploitation



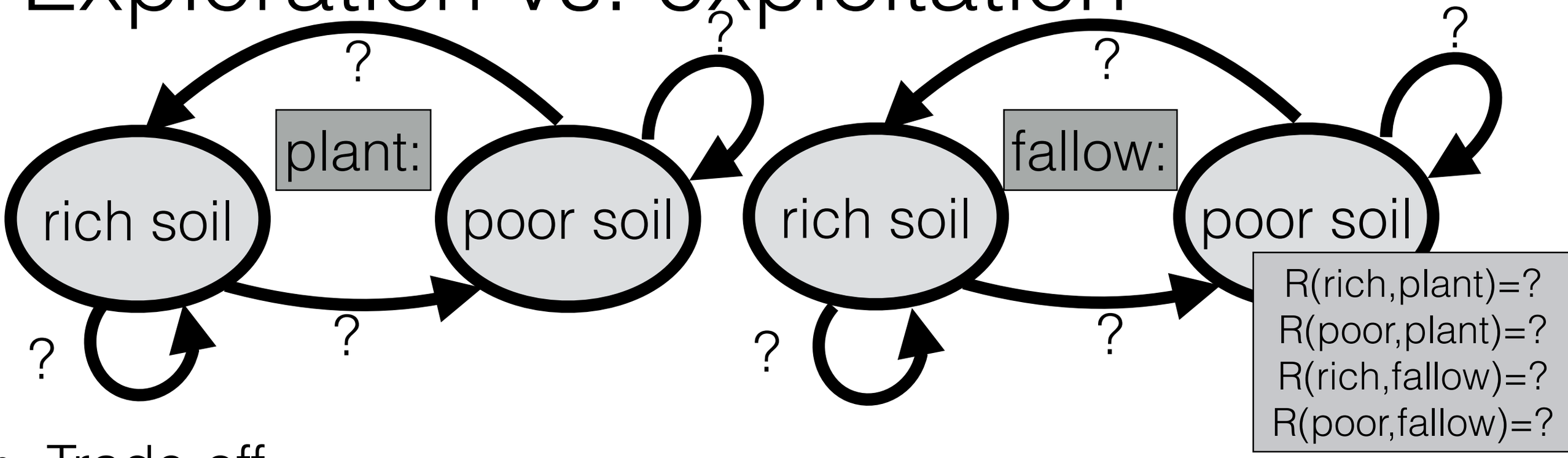
- Trade-off
  - **Exploration**: the more we explore, the better we understand the world (e.g. *T* and *R*)
  - **Exploitation**: based on what we know about the world, we can take actions with the aim to get highest reward
- One option (not the only one!)

Exploration vs. exploitation



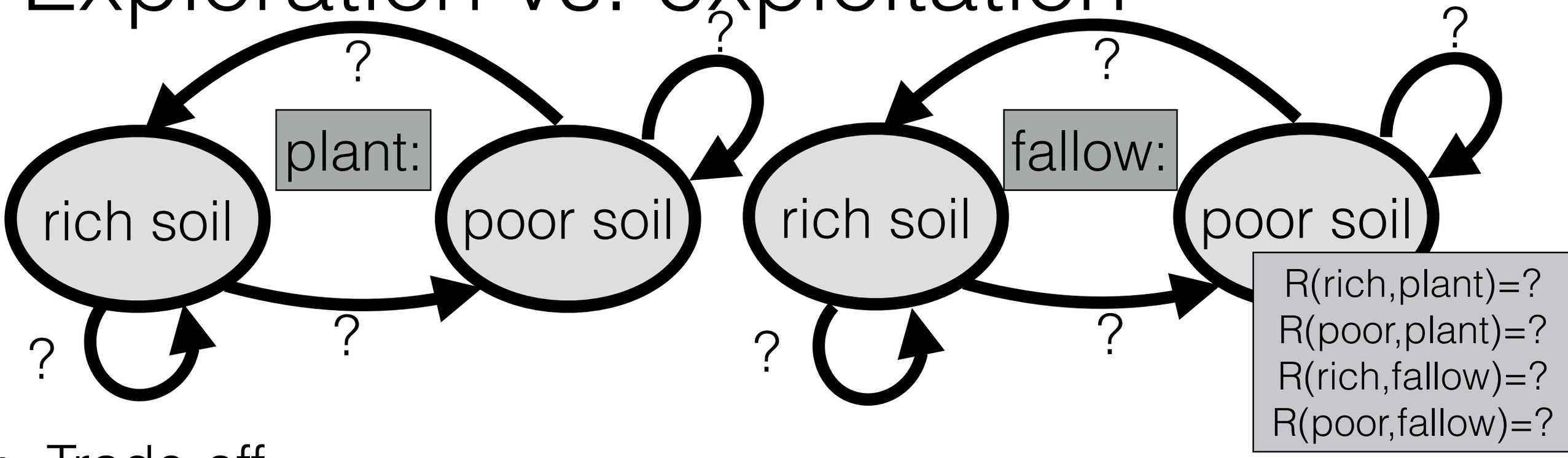
- Trade-off
  - **Exploration**: the more we explore, the better we understand the world (e.g. *T* and *R*)
  - **Exploitation**: based on what we know about the world, we can take actions with the aim to get highest reward
- One option (not the only one!): **ϵ-greedy strategy**

Exploration vs. exploitation



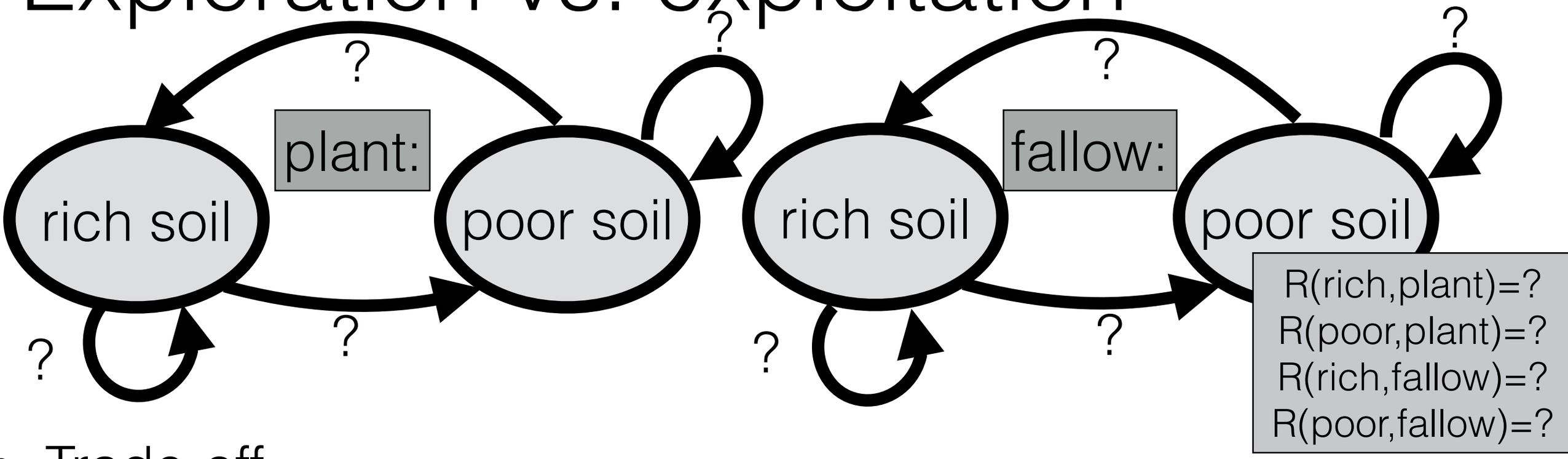
- Trade-off
  - **Exploration**: the more we explore, the better we understand the world (e.g. *T* and *R*)
  - **Exploitation**: based on what we know about the world, we can take actions with the aim to get highest reward
- One option (not the only one!): **ϵ-greedy strategy**
  - With probability 1-ϵ , exploit

Exploration vs. exploitation



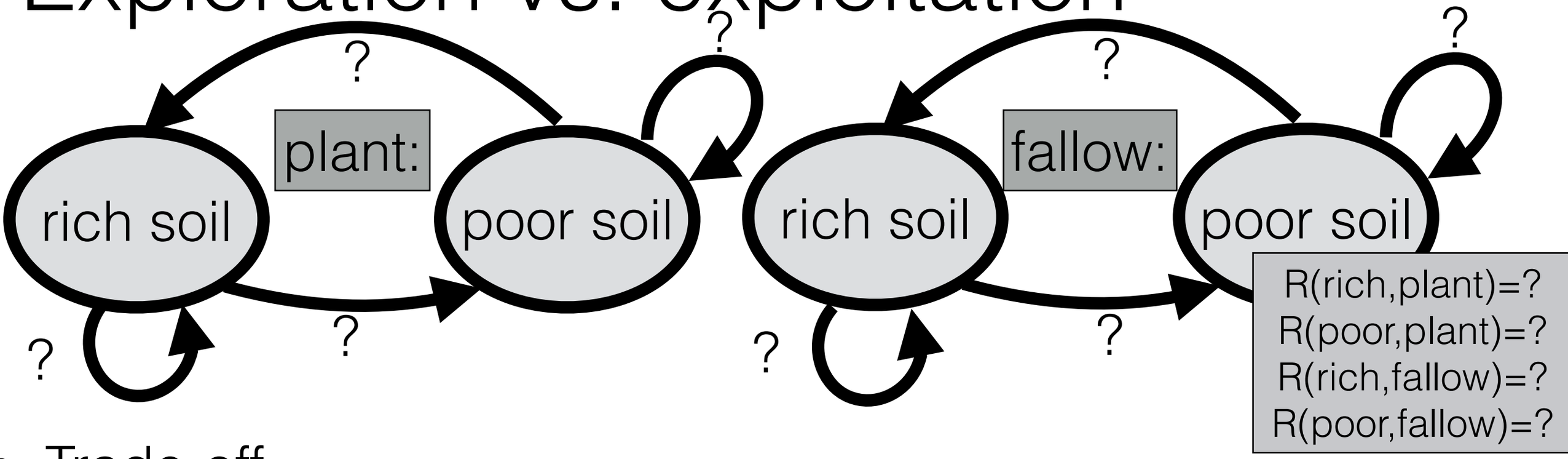
- Trade-off
  - **Exploration**: the more we explore, the better we understand the world (e.g. *T* and *R*)
  - **Exploitation**: based on what we know about the world, we can take actions with the aim to get highest reward
- One option (not the only one!): **ϵ-greedy strategy**
  - With probability 1-ϵ , exploit
  - With probability ϵ , explore

Exploration vs. exploitation



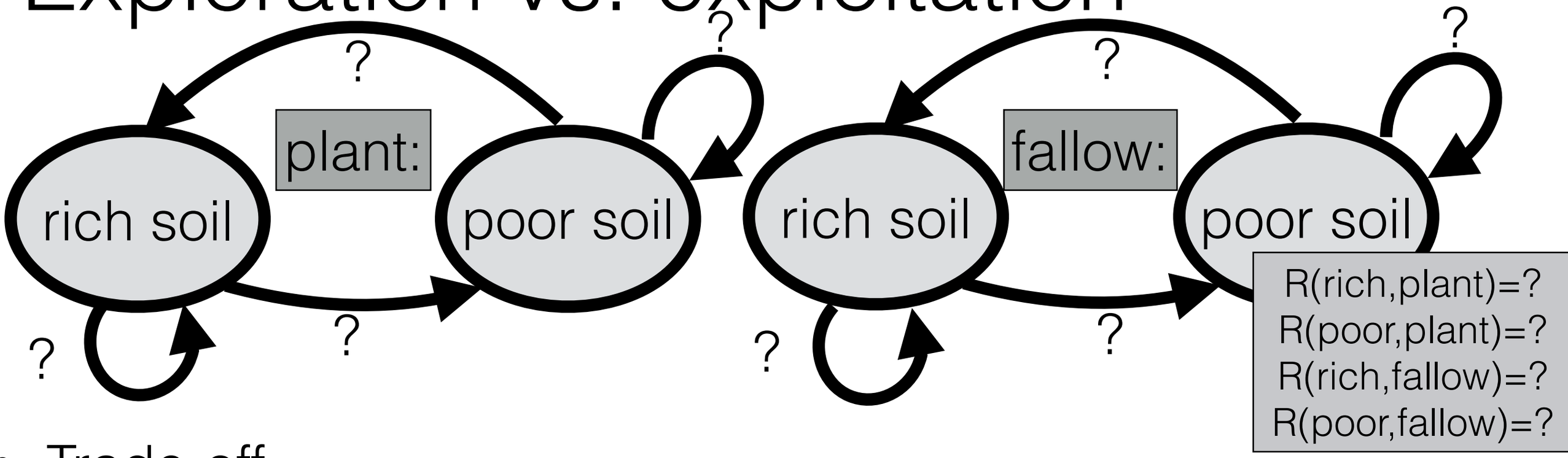
- Trade-off
  - **Exploration**: the more we explore, the better we understand the world (e.g. *T* and *R*)
  - **Exploitation**: based on what we know about the world, we can take actions with the aim to get highest reward
- One option (not the only one!): **ϵ-greedy strategy**
  - With probability 1-ϵ , exploit
  - With probability ϵ , explore

Exploration vs. exploitation



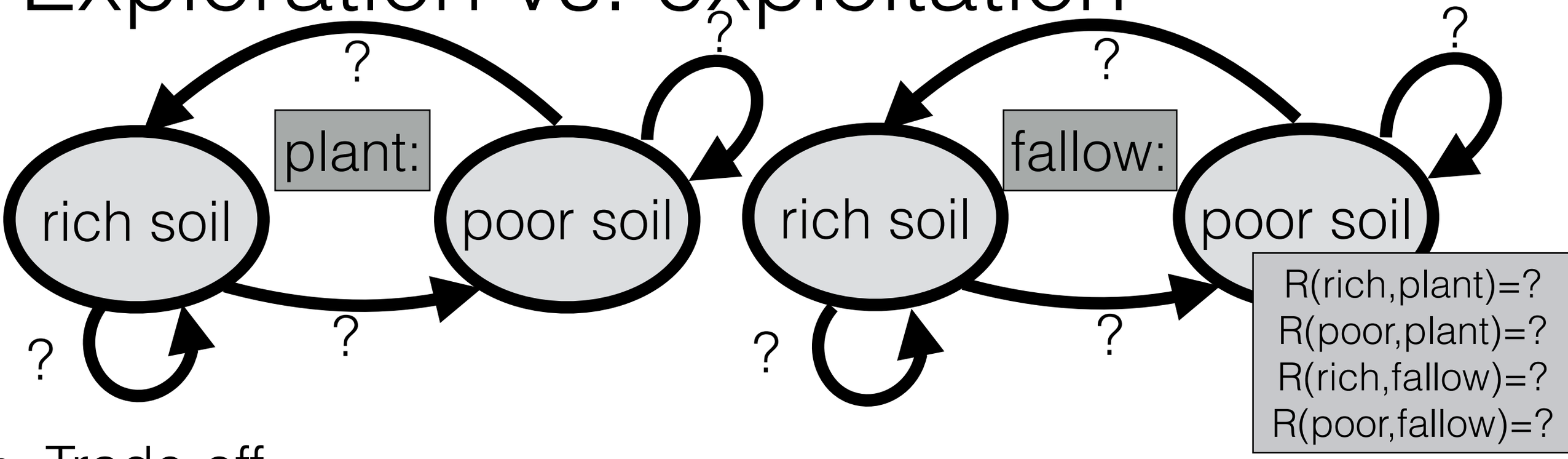
- Trade-off
  - **Exploration**: the more we explore, the better we understand the world (e.g. *T* and *R*)
  - **Exploitation**: based on what we know about the world, we can take actions with the aim to get highest reward
- One option (not the only one!): **ϵ-greedy strategy**
  - With probability 1-ϵ , exploit
  - With probability ϵ , choose an action uniformly at random

Exploration vs. exploitation



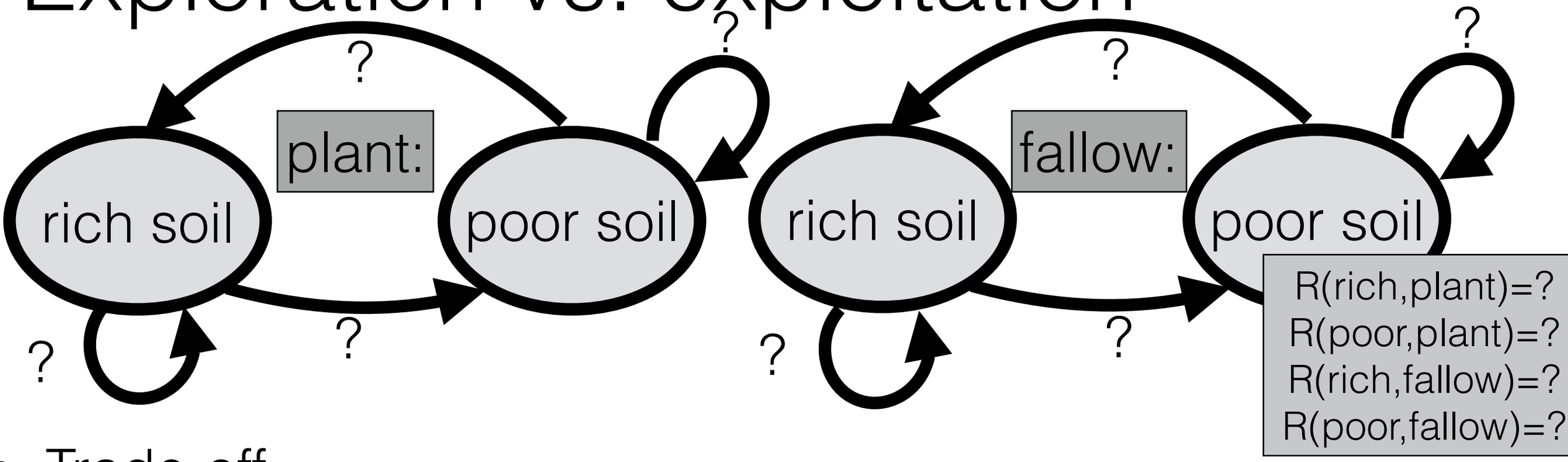
- Trade-off
  - **Exploration**: the more we explore, the better we understand the world (e.g. *T* and *R*)
  - **Exploitation**: based on what we know about the world, we can take actions with the aim to get highest reward
- One option (not the only one!): **ϵ-greedy strategy**
  - With probability 1-ϵ , exploit
  - With probability ϵ , choose an action uniformly at random

Exploration vs. exploitation



- Trade-off
  - **Exploration**: the more we explore, the better we understand the world (e.g. *T* and *R*)
  - **Exploitation**: based on what we know about the world, we can take actions with the aim to get highest reward
- One option (not the only one!): **ϵ-greedy strategy**
  - With probability 1-ϵ , exploit
  - With probability ϵ , choose an action uniformly at random

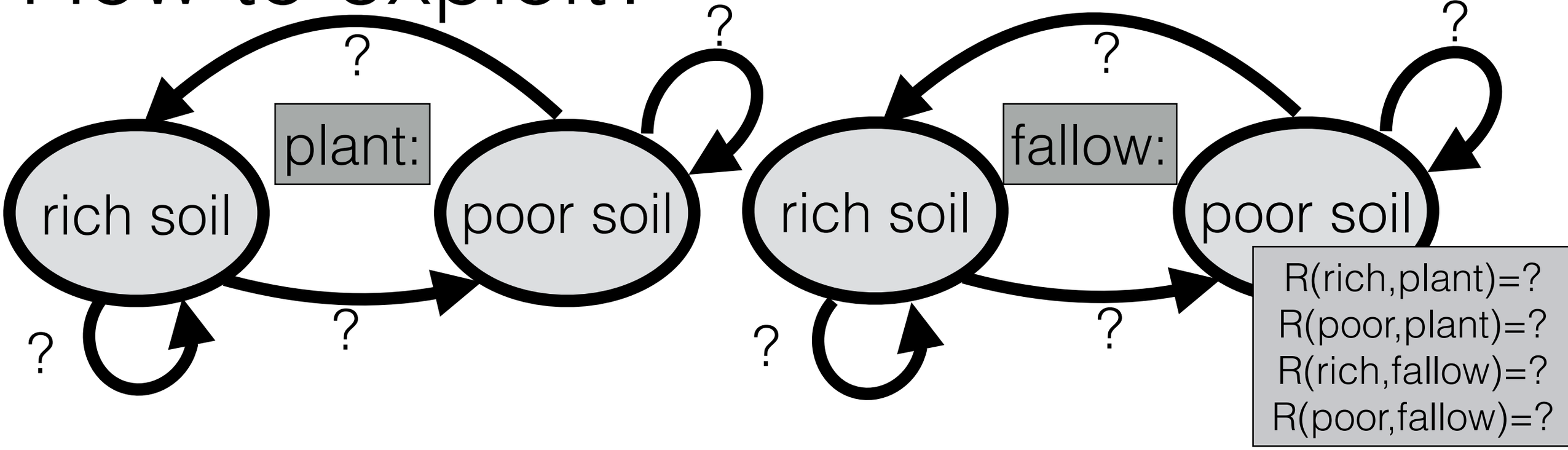
Exploration vs. exploitation



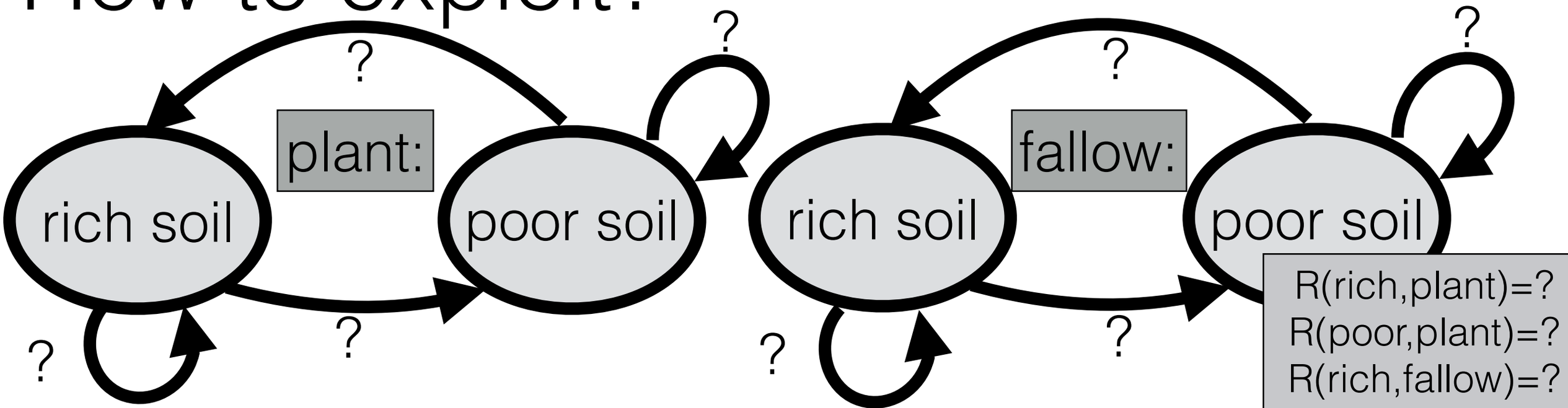
- Trade-off
  - **Exploration**: the more we explore, the better we understand the world (e.g. *T* and *R*)
  - **Exploitation**: based on what we know about the world, we can take actions with the aim to get highest reward
- One option (not the only one!): **ϵ-greedy strategy**
  - With probability 1-ϵ , exploit
  - With probability ϵ , choose an action uniformly at random

Need to specify how!

How to exploit?

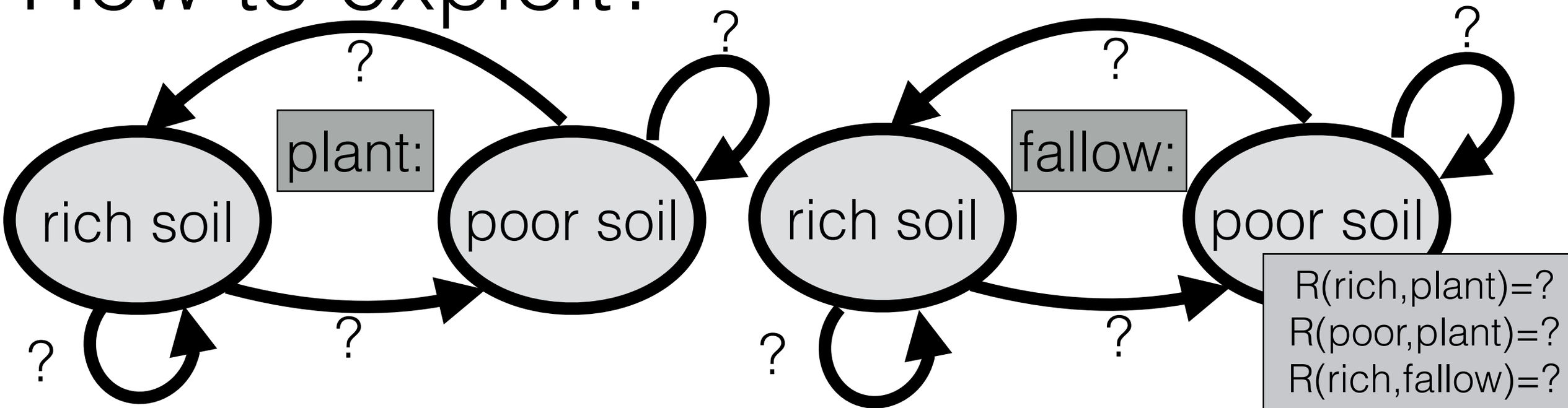


How to exploit?



- If we knew *Q** ("value" of taking action *a* at state *s* if we make all the best actions after this step):

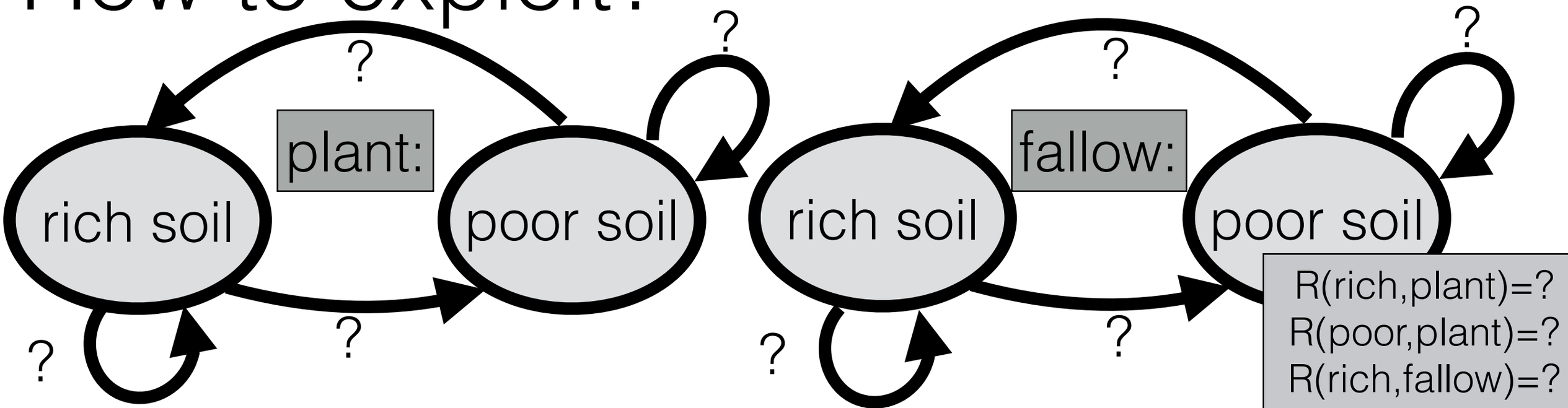
How to exploit?



- If we knew *Q** ("value" of taking action *a* at state *s* if we make all the best actions after this step):

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

How to exploit?

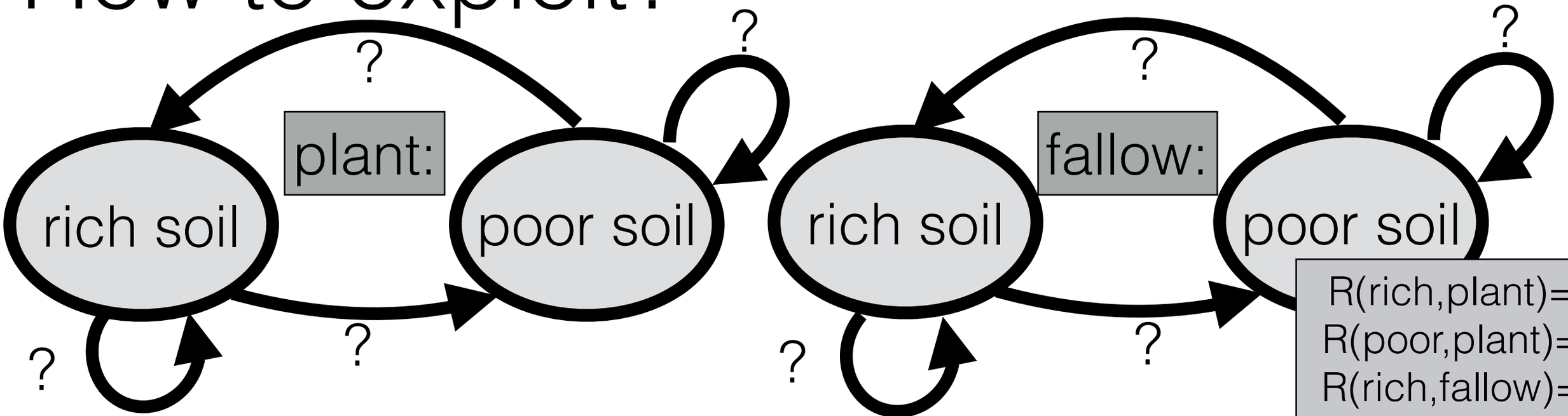


- If we knew *Q** ("value" of taking action *a* at state *s* if we make all the best actions after this step):

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

- Problem: we don't know *Q**

How to exploit?

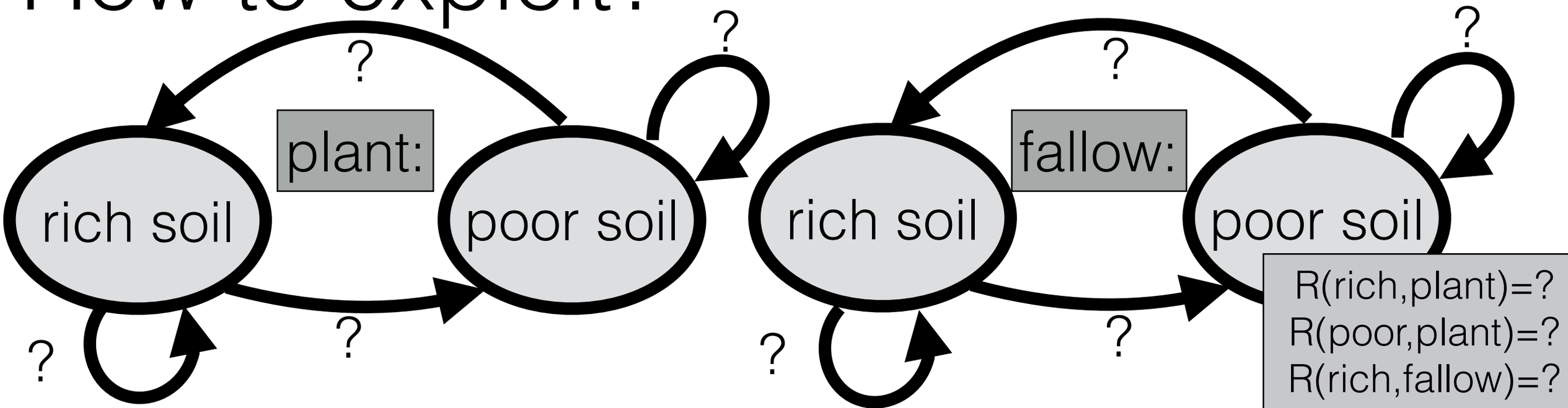


- If we knew *Q** ("value" of taking action *a* at state *s* if we make all the best actions after this step):

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

- Problem: we don't know *Q**
- Idea: use a guess/estimate *Q* for *Q**

How to exploit?



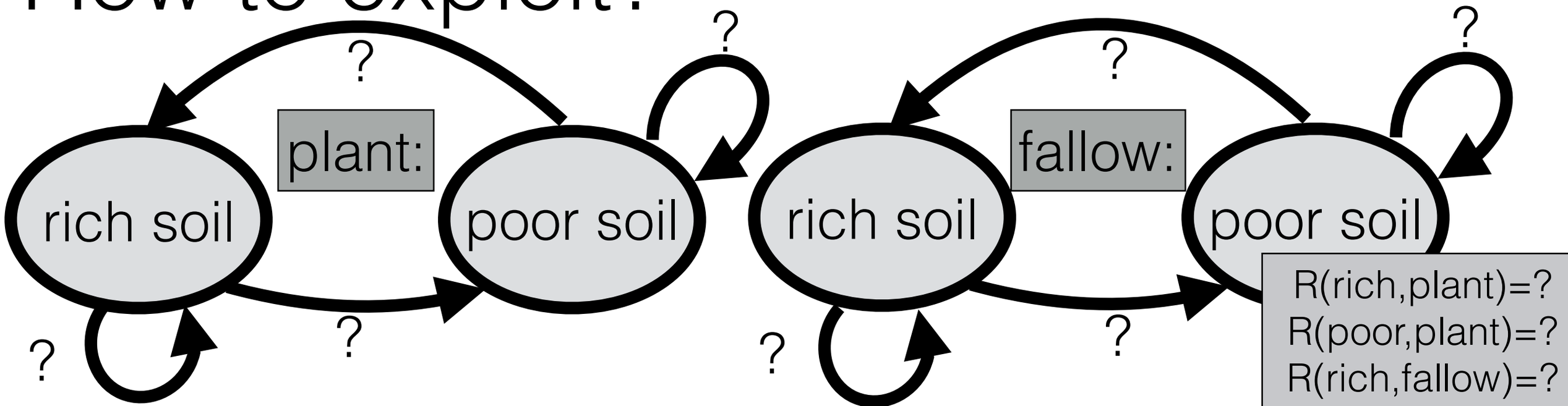
- If we knew *Q** ("value" of taking action *a* at state *s* if we make all the best actions after this step):

$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

- Problem: we don't know *Q**
- Idea: use a guess/estimate *Q* for *Q**

$$\pi_Q(s) = \arg\max_a Q(s, a)$$

How to exploit?



- If we knew *Q** ("value" of taking action *a* at state *s* if we make all the best actions after this step):

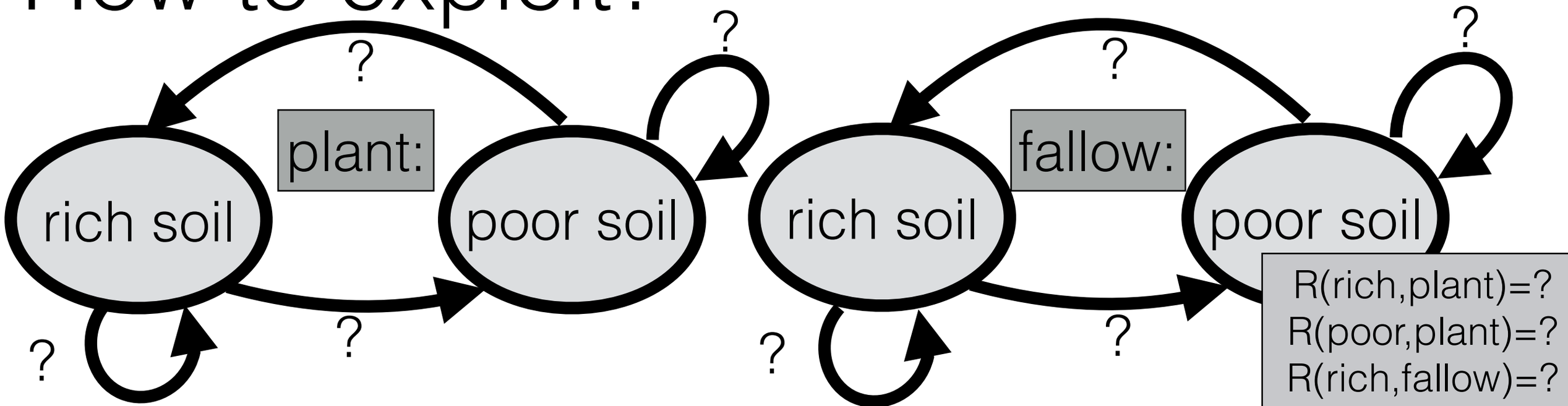
$$\pi_{Q^*}(s) = \arg\max_a Q^*(s,a)$$

- Problem: we don't know *Q**
- Idea: use a guess/estimate *Q* for *Q**

$$\pi_Q(s) = \arg\max_a Q(s,a)$$

- Question: how to get a good guess/estimate *Q* for *Q**?

How to exploit?



- If we knew *Q** ("value" of taking action *a* at state *s* if we make all the best actions after this step):

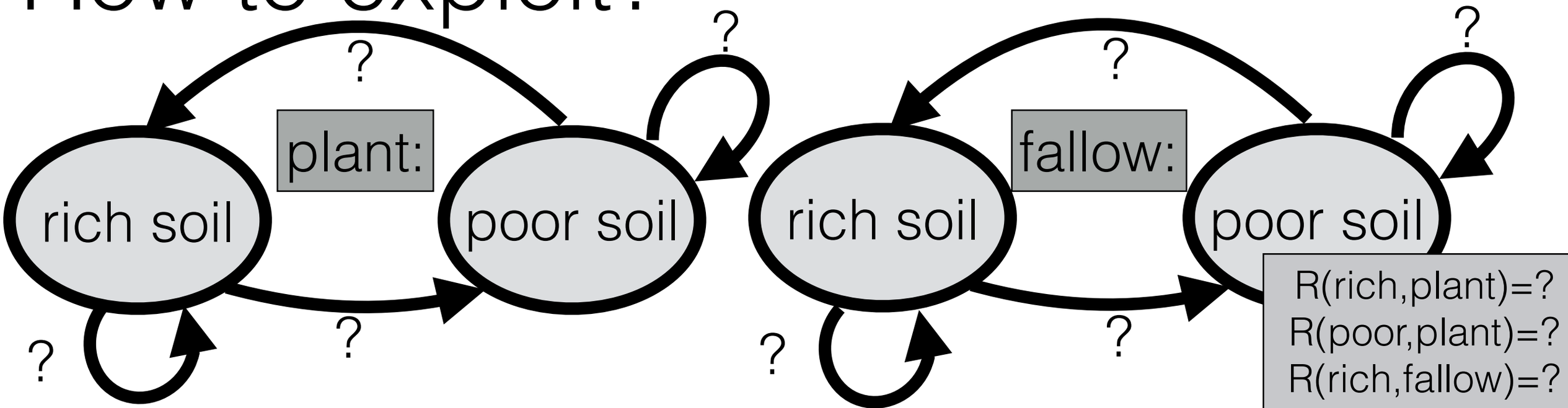
$$\pi_{Q^*}(s) = \arg\max_a Q^*(s,a)$$

- Problem: we don't know *Q**
- Idea: use a guess/estimate *Q* for *Q**

$$\pi_Q(s) = \arg\max_a Q(s,a)$$

- Question: how to get a good guess/estimate *Q* for *Q**?
- Idea: Learn *Q* from the data we've seen so far

How to exploit?



- If we knew Q^* ("value" of taking action a at state s if we make all the best actions after this step):

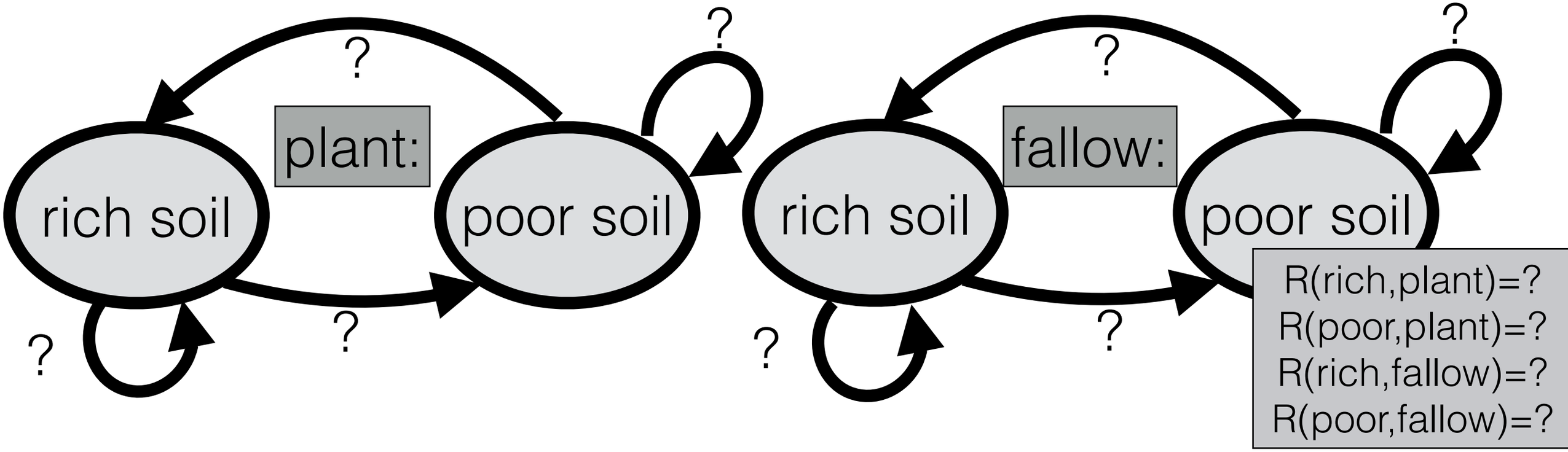
$$\pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$$

- Problem: we don't know Q^*
- Idea: use a guess/estimate Q for Q^*

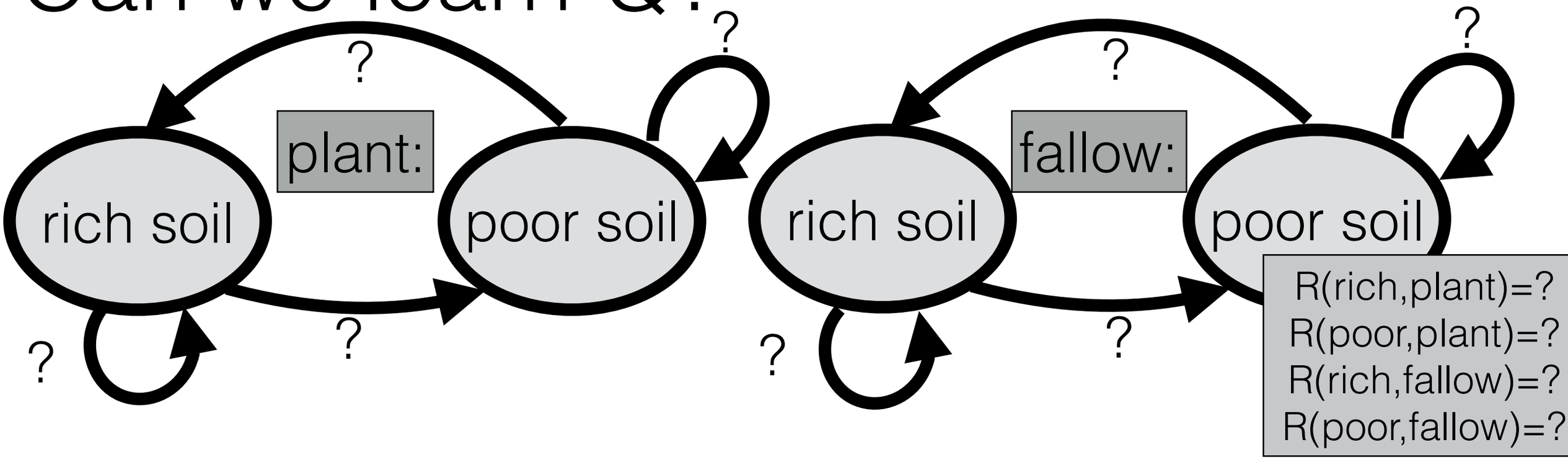
$$\pi_Q(s) = \arg\max_a Q(s, a)$$

- Question: how to get a good guess/estimate Q for Q^*?
- Idea: Learn Q from the data we've seen so far

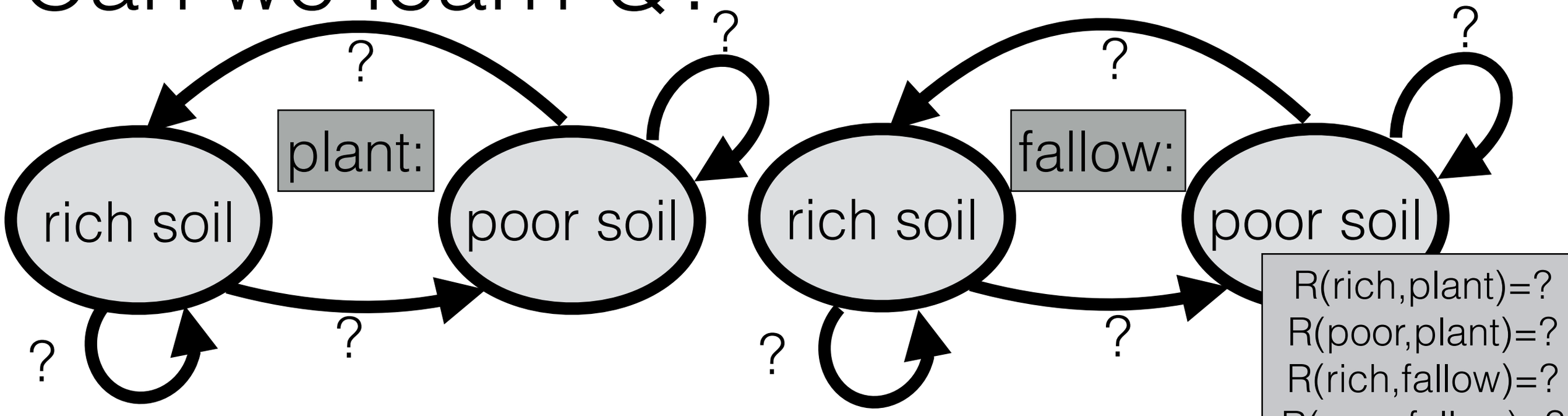
Data at step t: $s^{(t)}, a^{(t)}, r^{(t)}, s^{(t+1)}$


?
?
?
plant:
rich soil
poor soil
?
?
?
?
?
fallow:
rich soil
poor soil
?
R(rich,plant)=?
R(poor,plant)=?
R(rich,fallow)=?
R(poor,fallow)=?

Can we learn *Q*?

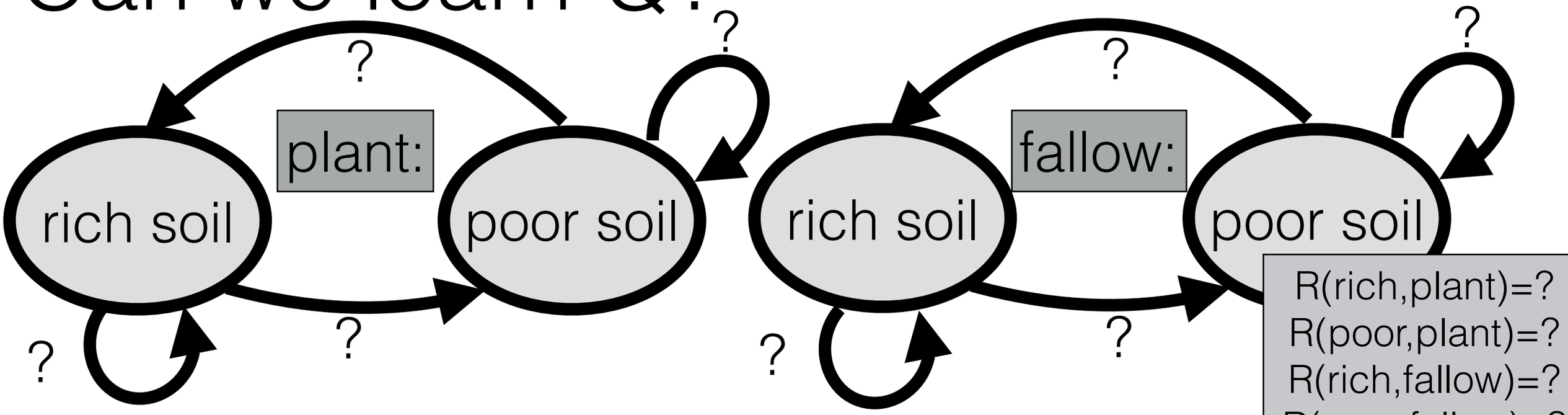


Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

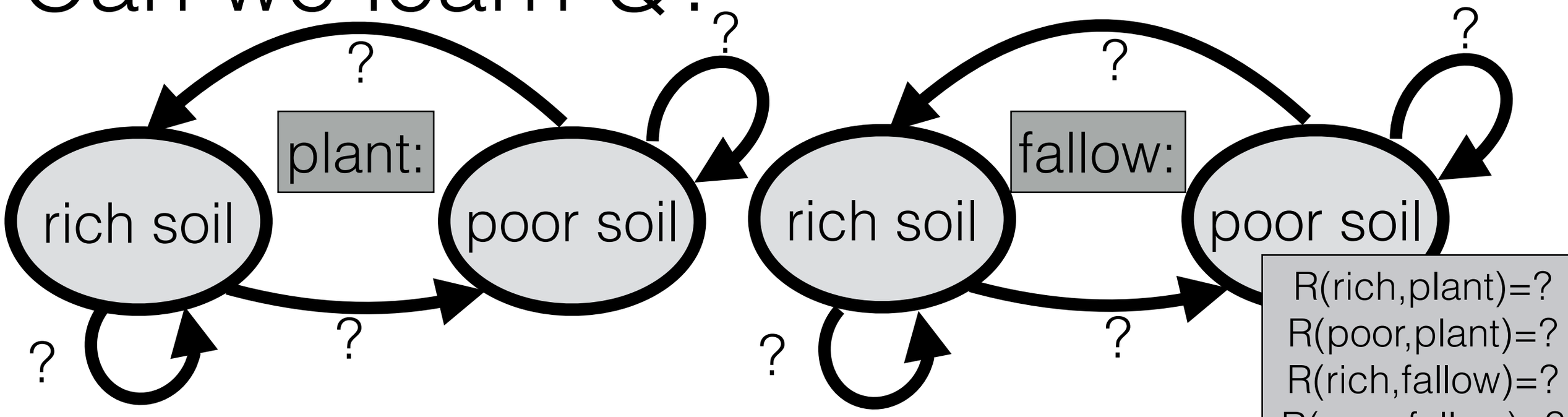
Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

`Initialize` $s^{(1)} = s_0$

Can we learn *Q*?

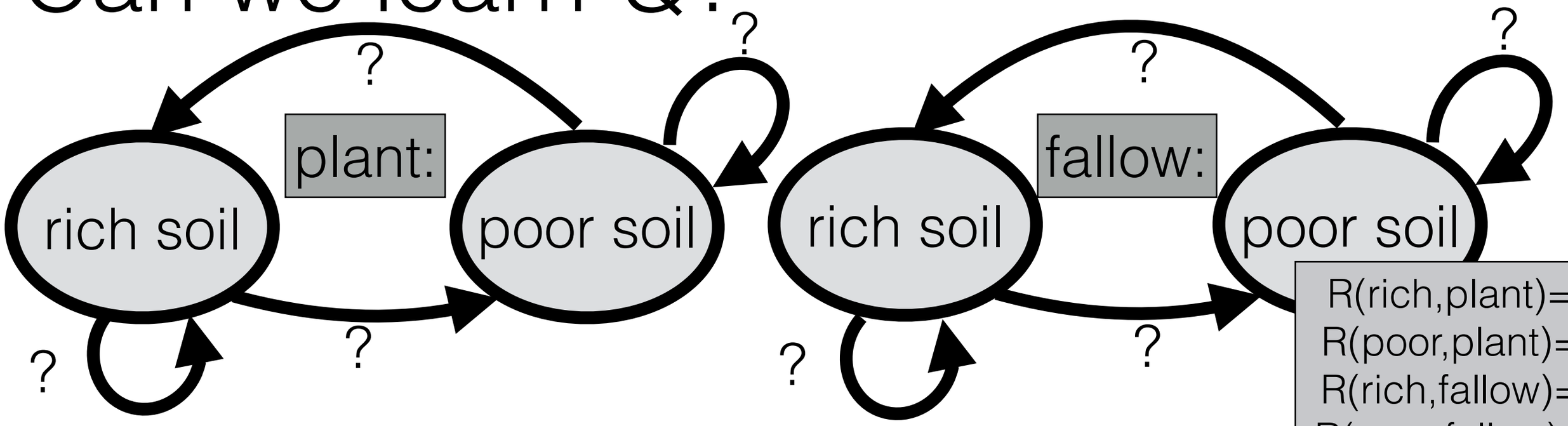


- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0$

Can we learn *Q*?

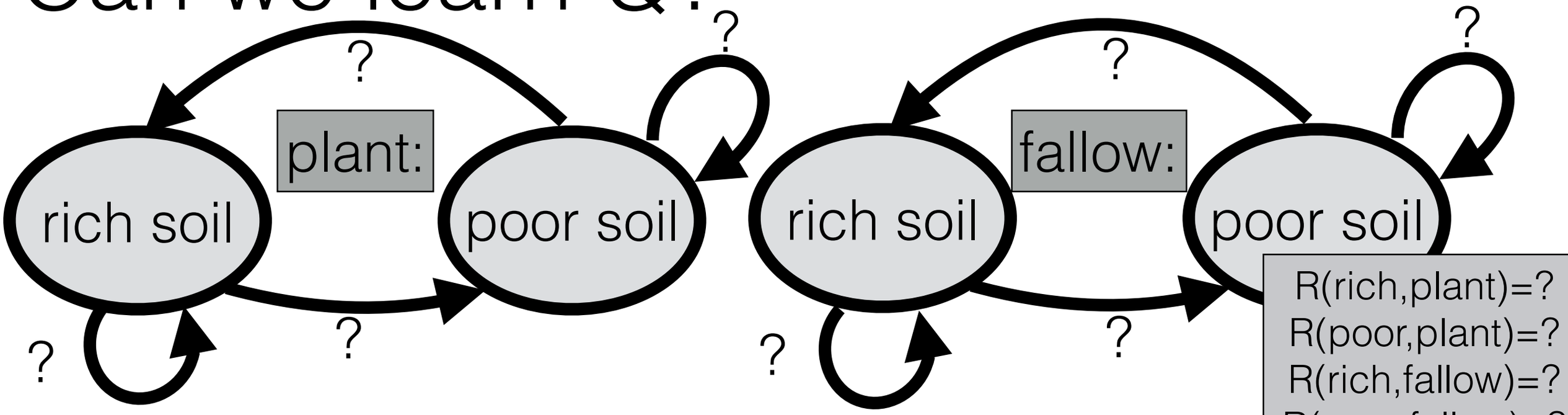


- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0; Q$

Can we learn *Q*?



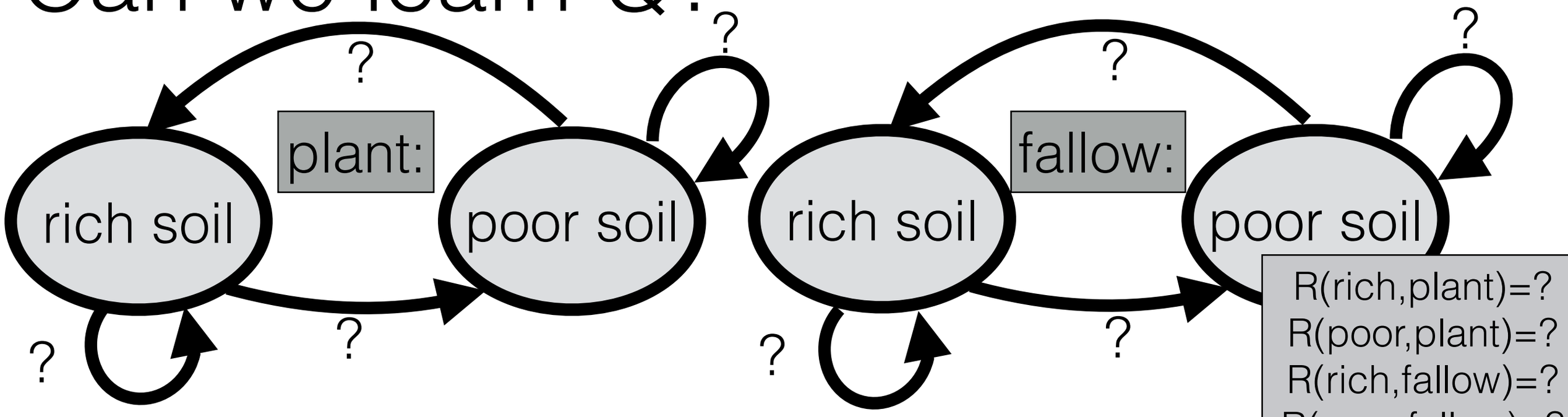
- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|S|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

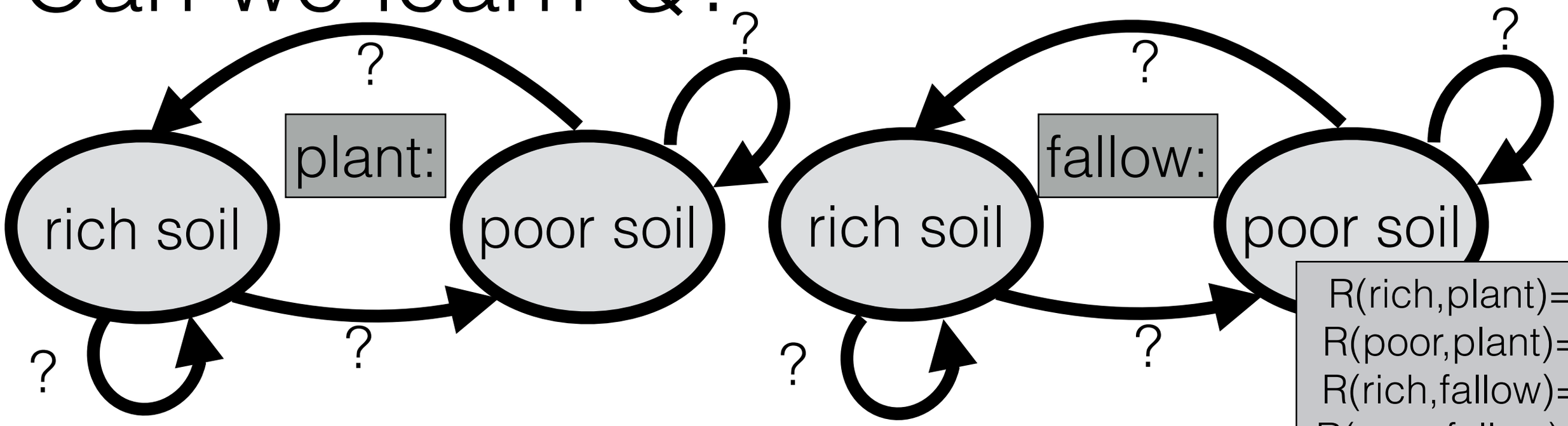
Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|S|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

  $a^{(t)}$ = select_action($s^{(t)}, Q$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

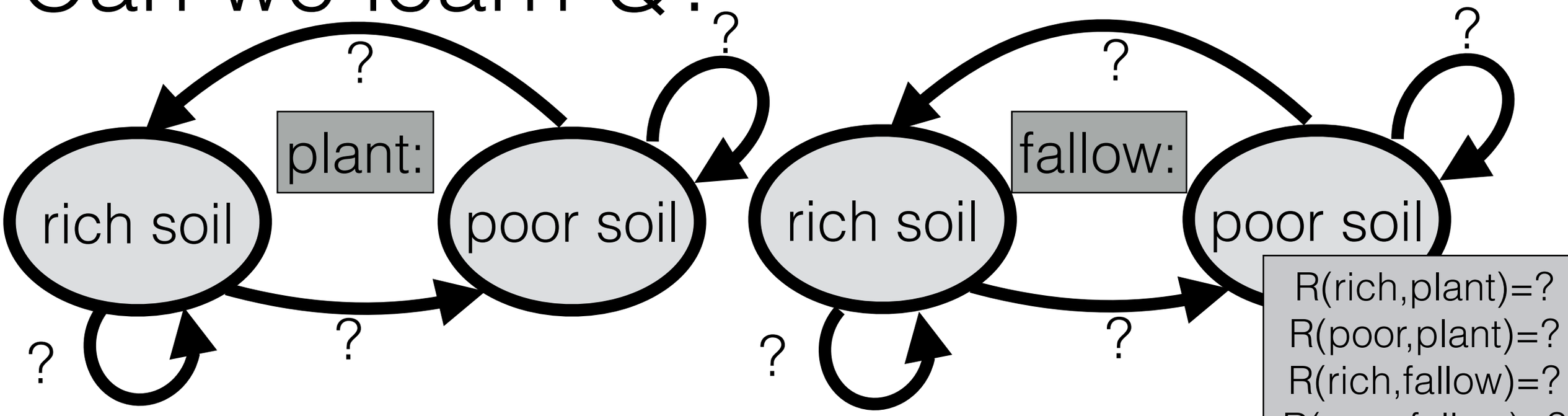
Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|S|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

  $a^{(t)}$ = select_action($s^{(t)}, Q$)

E.g. ϵ**-greedy**

Can we learn *Q*?

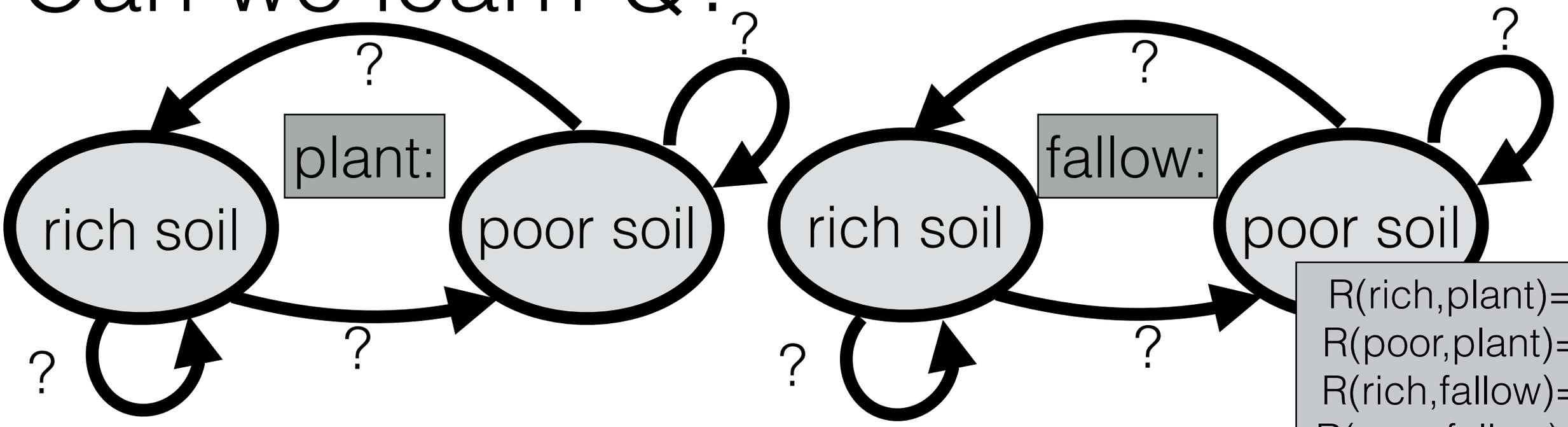


- Option 1: Estimate transition model *T* and reward function *R*

```
Initialize s^(1) = s_0
Initialize: any s, a, s': T̂(s,a,s') = 1/|S|; R̂(s,a) = 0; Q
for t = 1, 2, 3, …
  a^(t) = select_action(s^(t), Q)
  r^(t), s^(t+1) = execute(a^(t))
```

E.g. ϵ-greedy

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|S|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

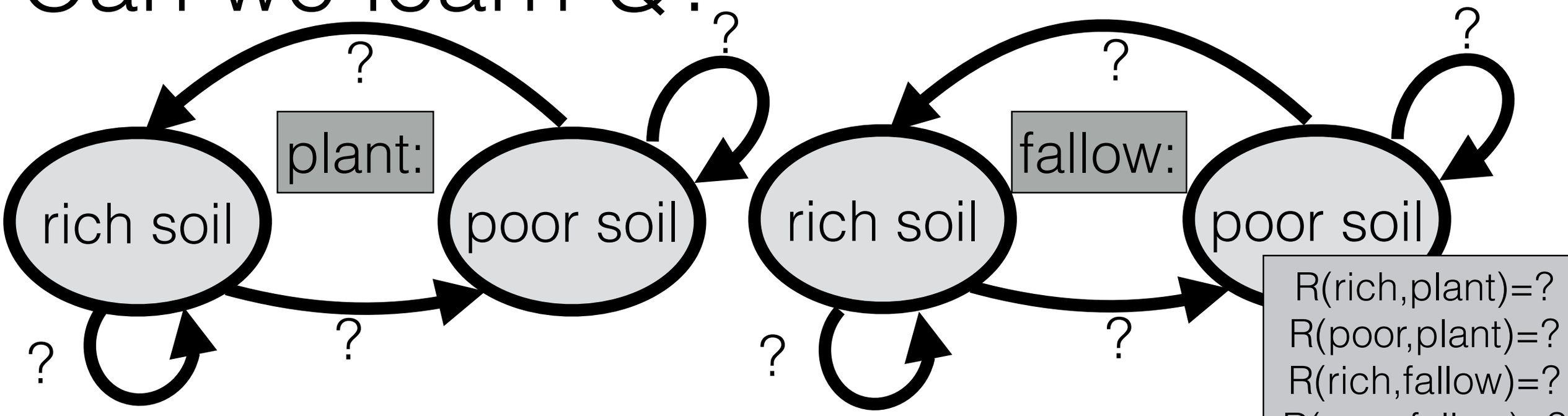
  $a^{(t)}$ = select_action($s^{(t)}, Q$)

  $r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

  $\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|S|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

  $a^{(t)}$ = select_action($s^{(t)}, Q$)

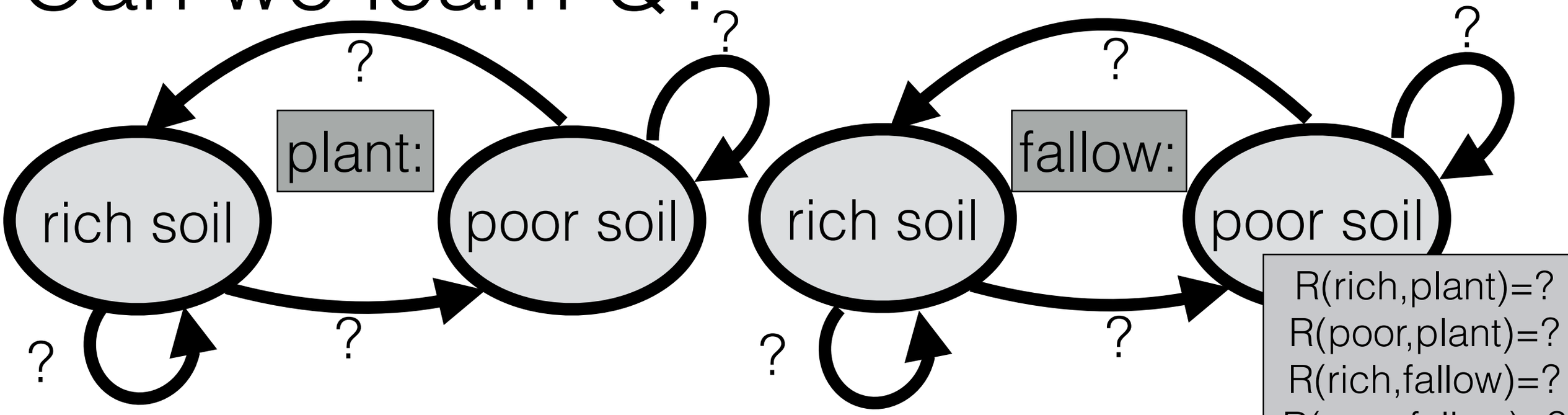
  $r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

  $\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

  Any s,a,s′: $\hat{T}(s,a,s') =$

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

  $a^{(t)}$ = select_action($s^{(t)}, Q$)

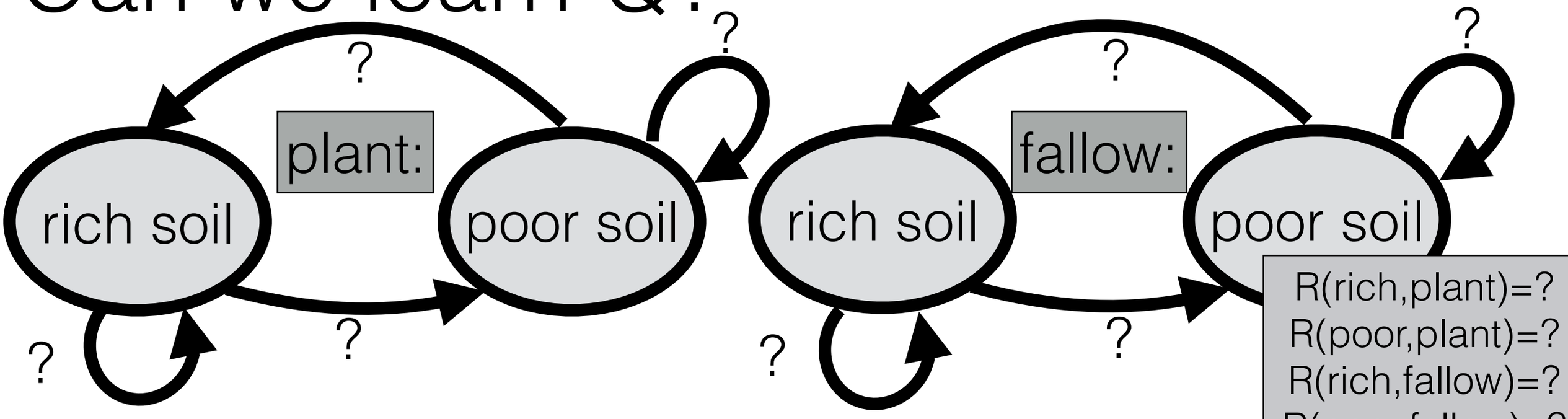
  $r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

  $\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

  Any s,a,s′: $\hat{T}(s,a,s') = \frac{\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|S|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

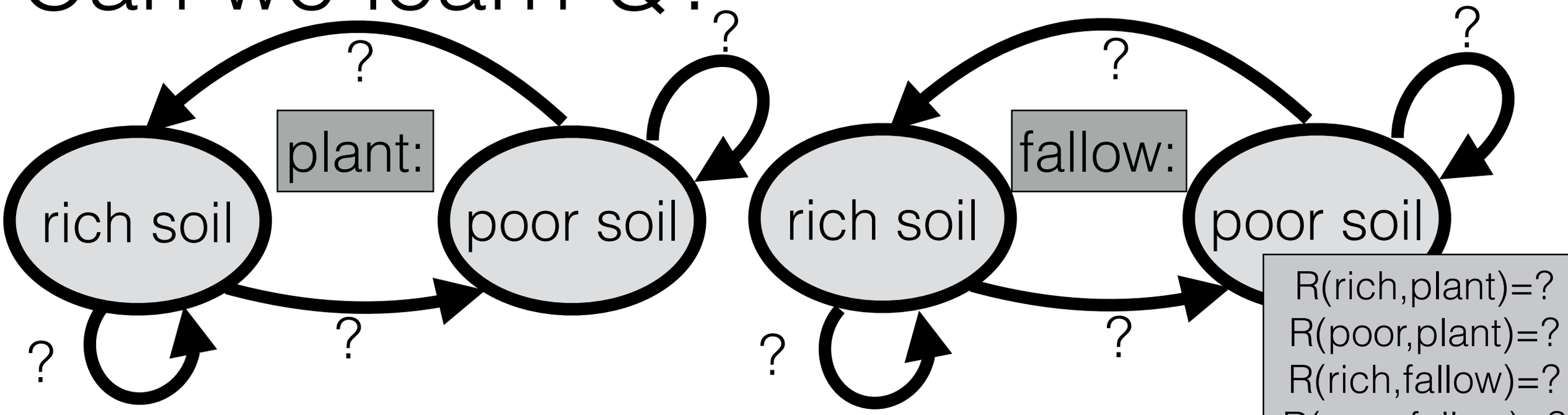
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|S|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

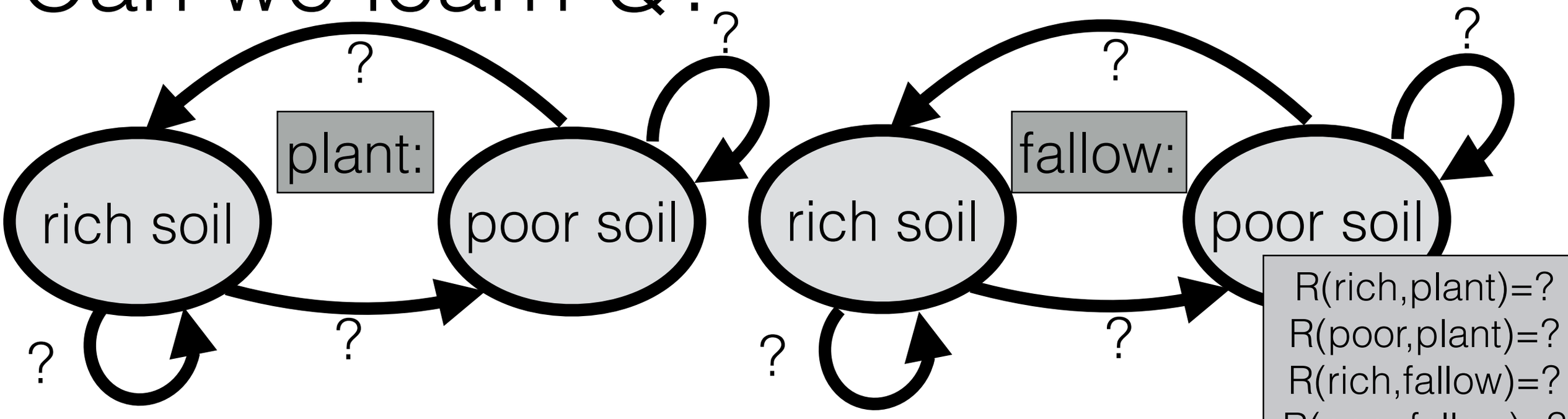
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|S|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

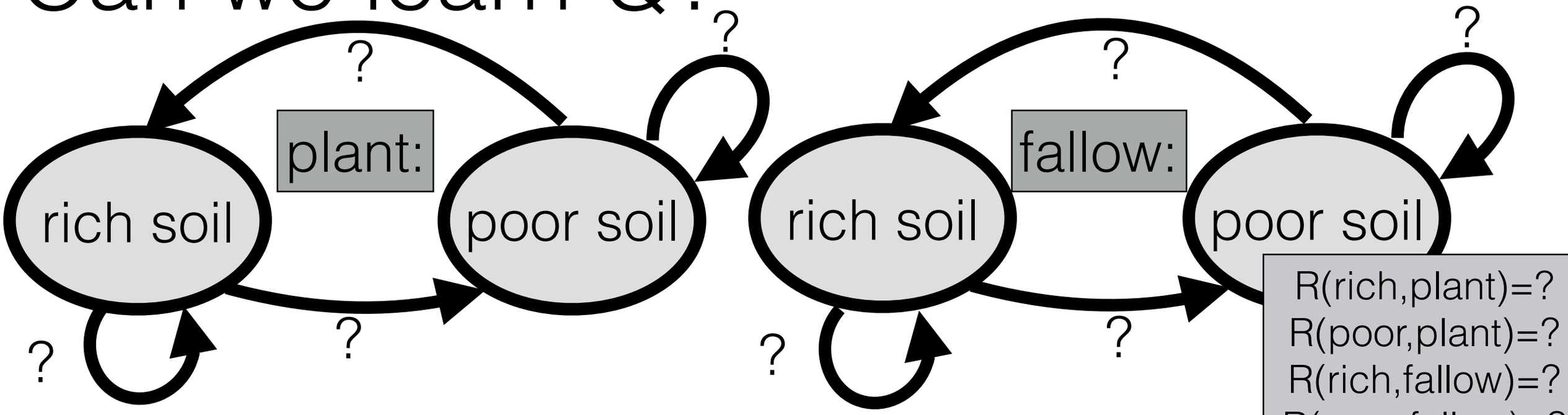
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|S|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

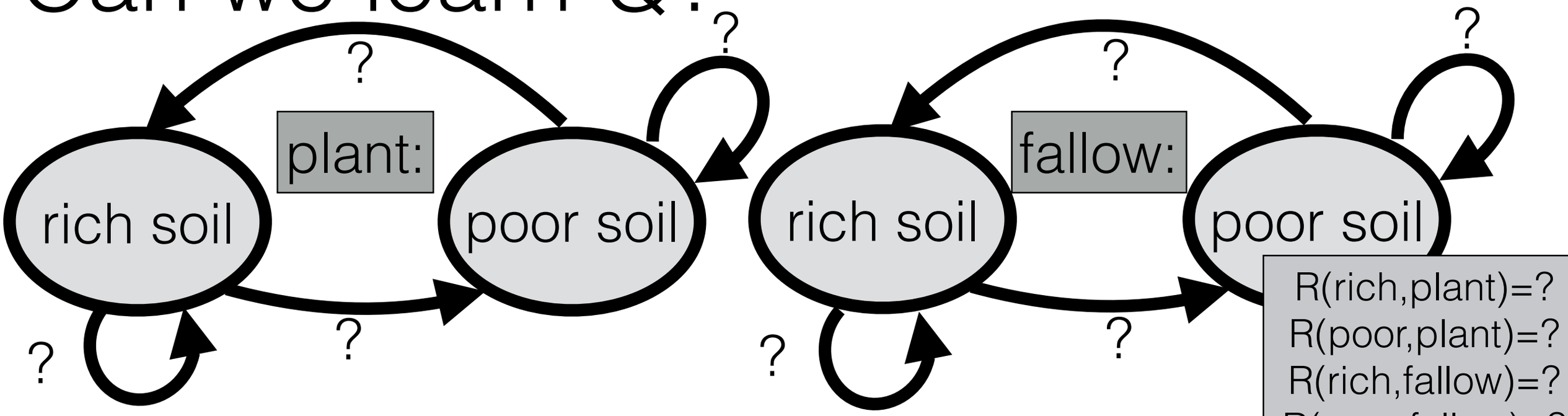
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

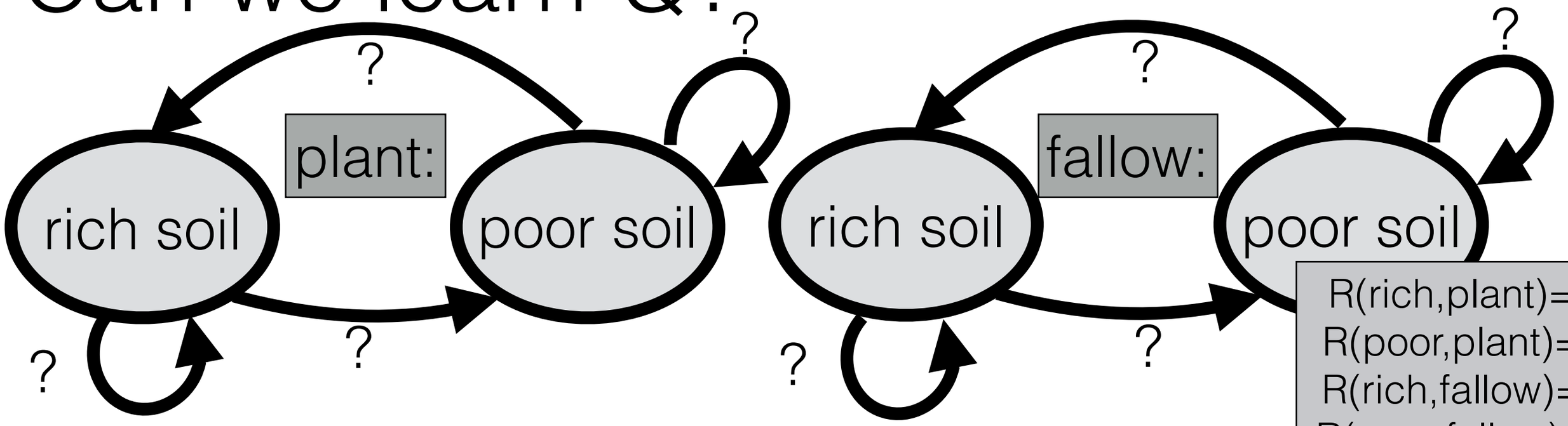
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

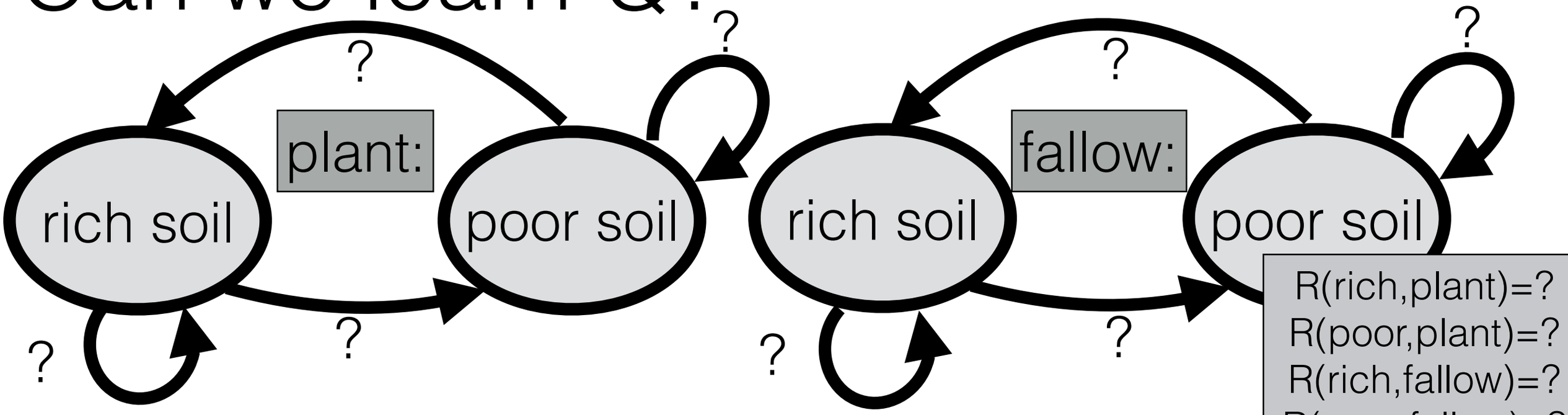
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

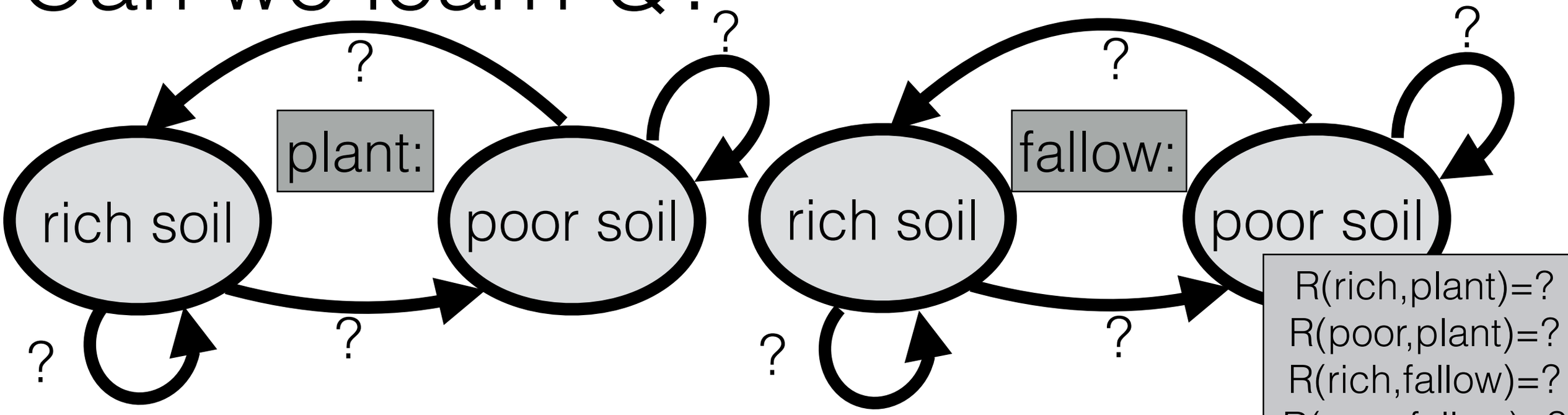
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data at step *t*: $s^{(t)}, a^{(t)}, r^{(t)}, s^{(t+1)}$

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

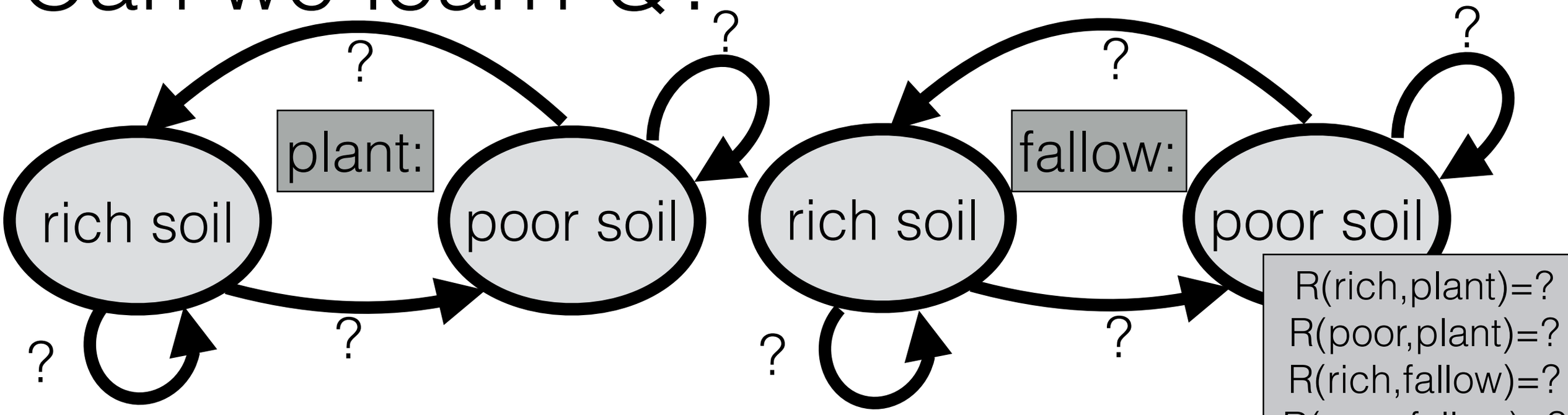
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data at step *t*: $s^{(t)}, a^{(t)}, r^{(t)}, s^{(t+1)}$

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

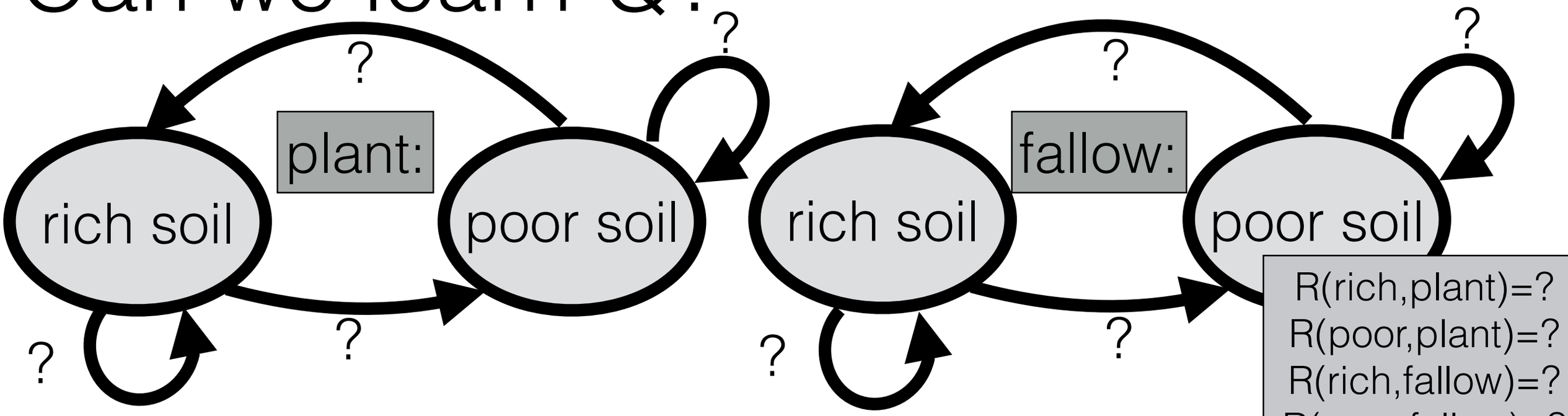
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

E.g. **ϵ-greedy**

Can we learn *Q*?



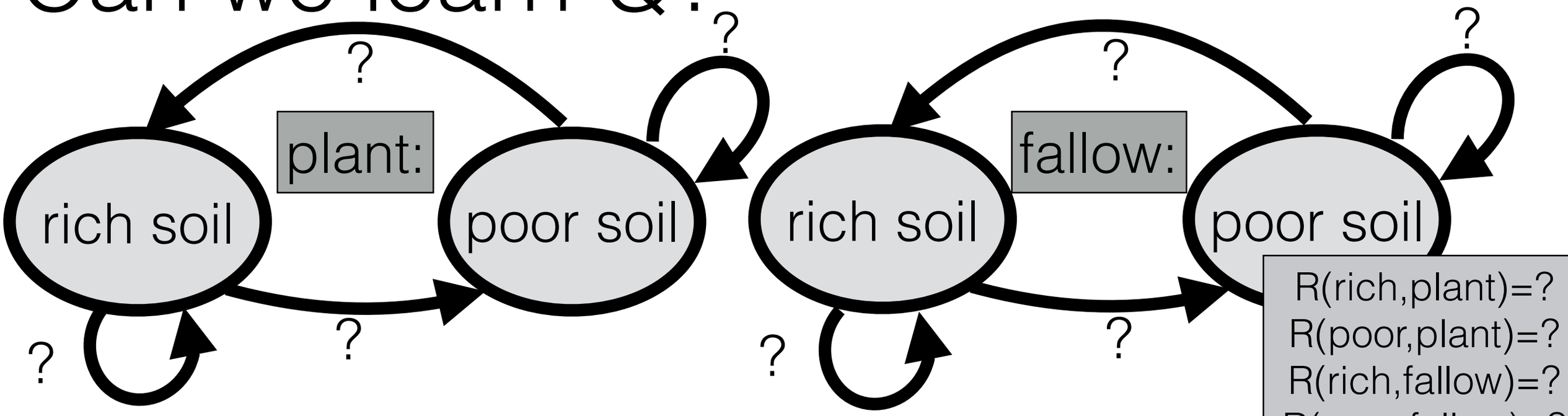
- Option 1: Estimate transition model *T* and reward function *R*

Data at step *t*: $s^{(t)}, a^{(t)}, r^{(t)}, s^{(t+1)}$

```
Initialize s^(1) = s_0
Initialize: any s, a, s': T̂(s,a,s') = 1/|S|; R̂(s,a) = 0; Q
for t = 1, 2, 3, …
    a^(t) = select_action(s^(t), Q)
    r^(t), s^(t+1) = execute(a^(t))
    R̂(s^(t), a^(t)) = r^(t)
    Any s,a,s': T̂(s,a,s') = (1 + Σ_{i=1}^{t} 1{s^(i)=s, a^(i)=a, s^(i+1)=s'}) / (|S| + Σ_{i=1}^{t} 1{s^(i)=s, a^(i)=a})
    Q = infinite-horizon-value-iteration(R̂, T̂)
```

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data Example:

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

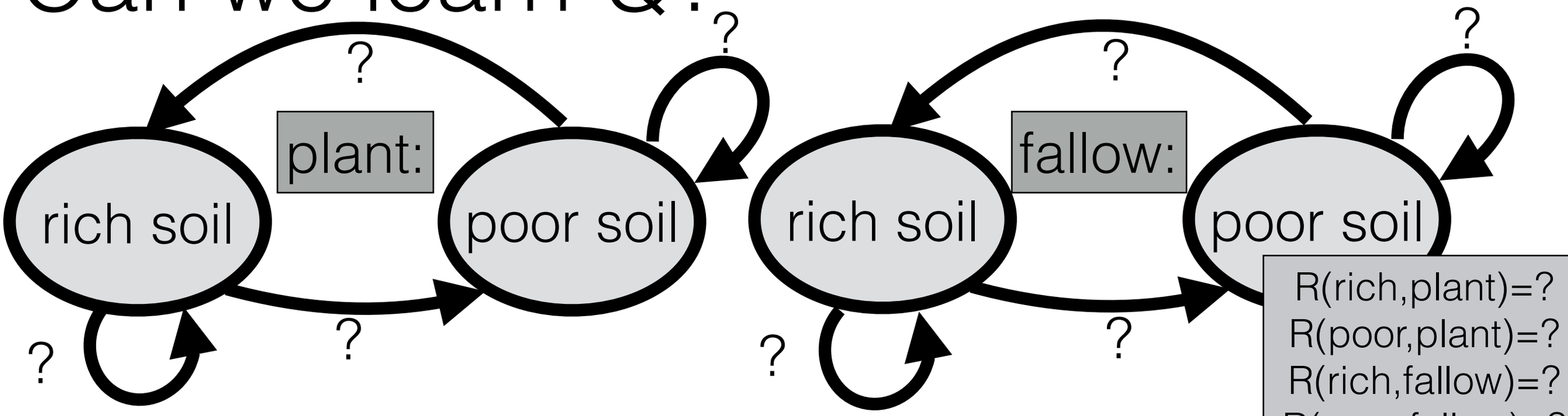
$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data Example: $\epsilon = 0.3$

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

E.g. **ϵ-greedy**

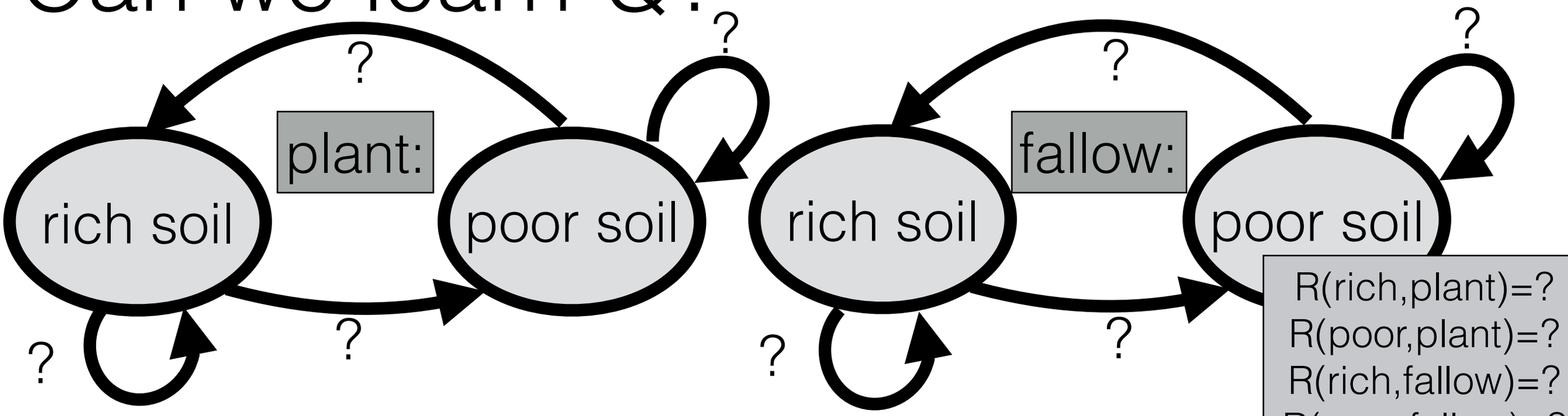
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data Example: $\epsilon = 0.3; \gamma = 0.99$

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

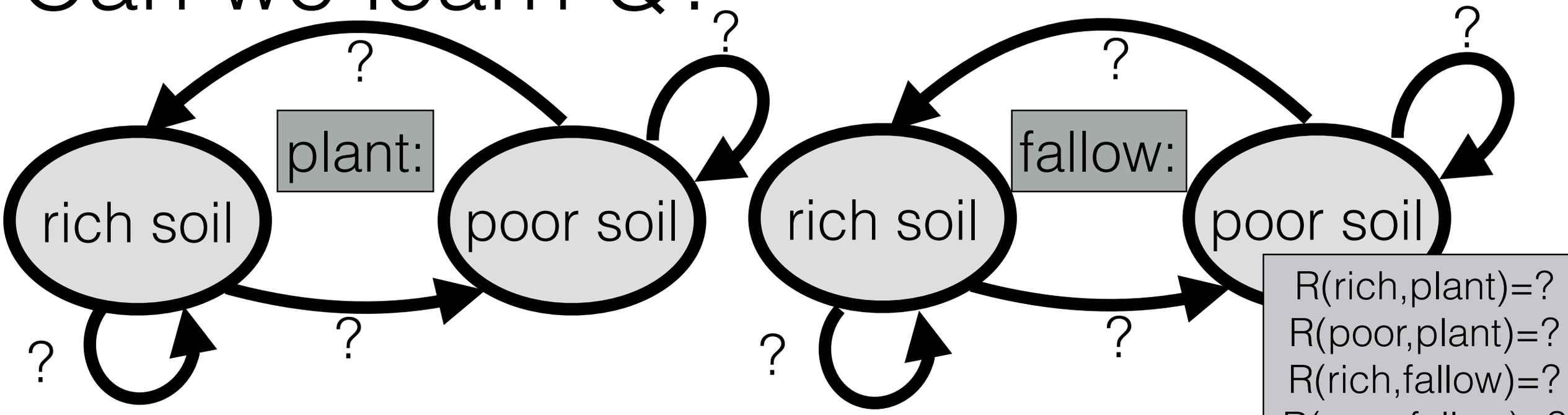
$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data Example: $\epsilon = 0.3; \gamma = 0.99$

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

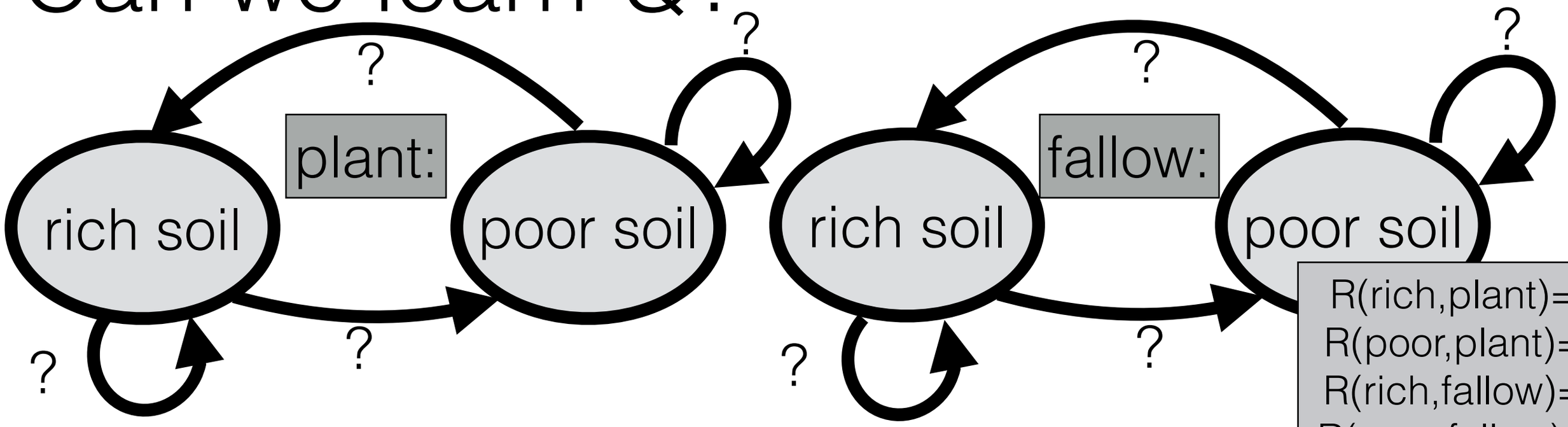
$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}$

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; R(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

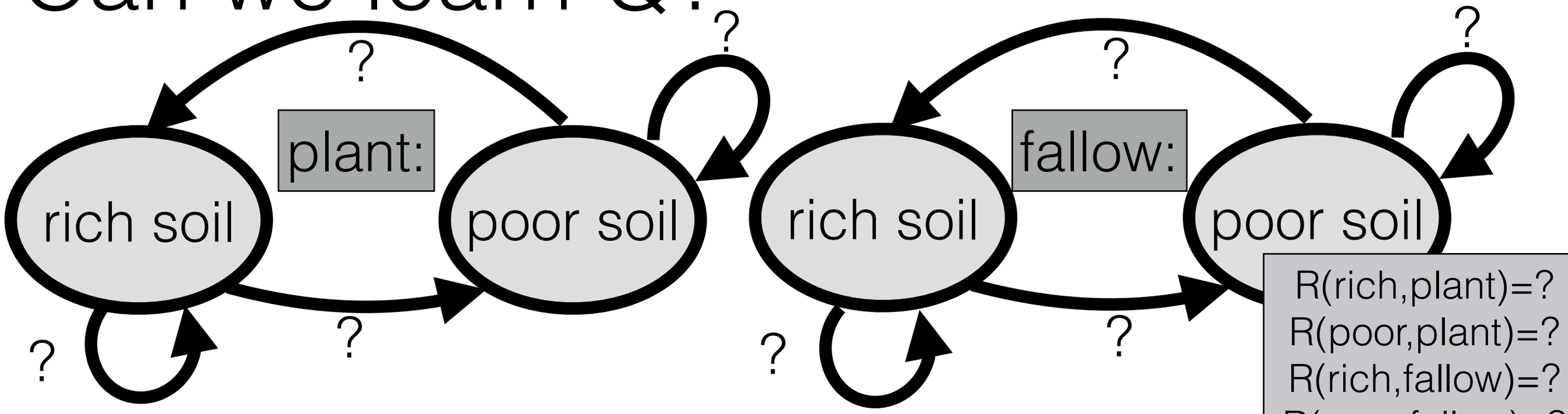
$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}$

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s, a, s') = \frac{1}{|S|}; \hat{R}(s, a) = 0; Q$

for t = 1, 2, 3, …

E.g. **ϵ-greedy**

$a^{(t)}$ = select_action($s^{(t)}, Q$)

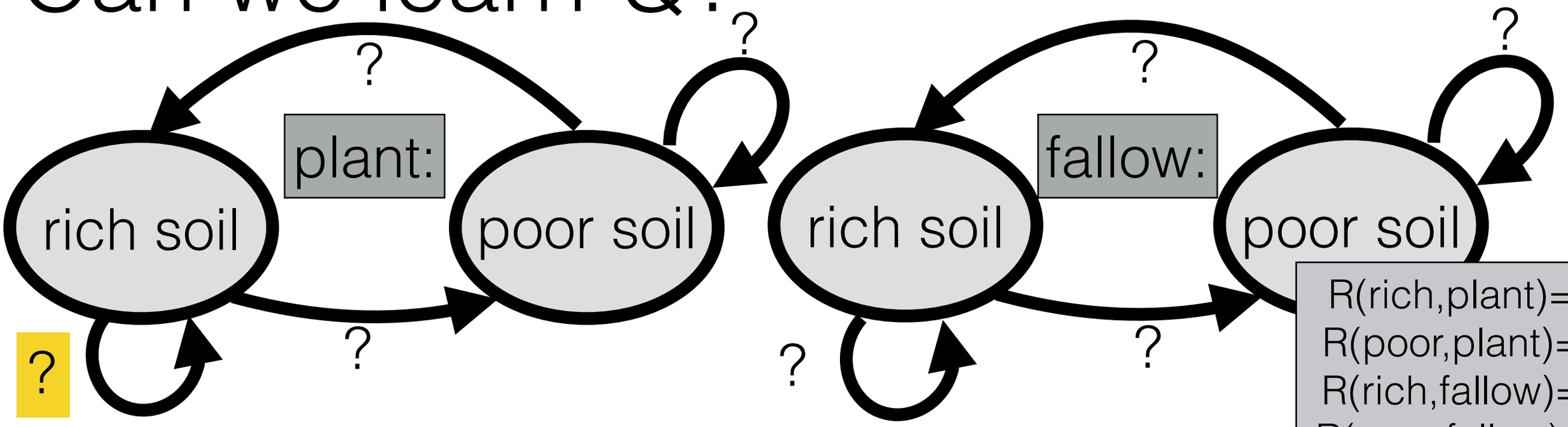
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

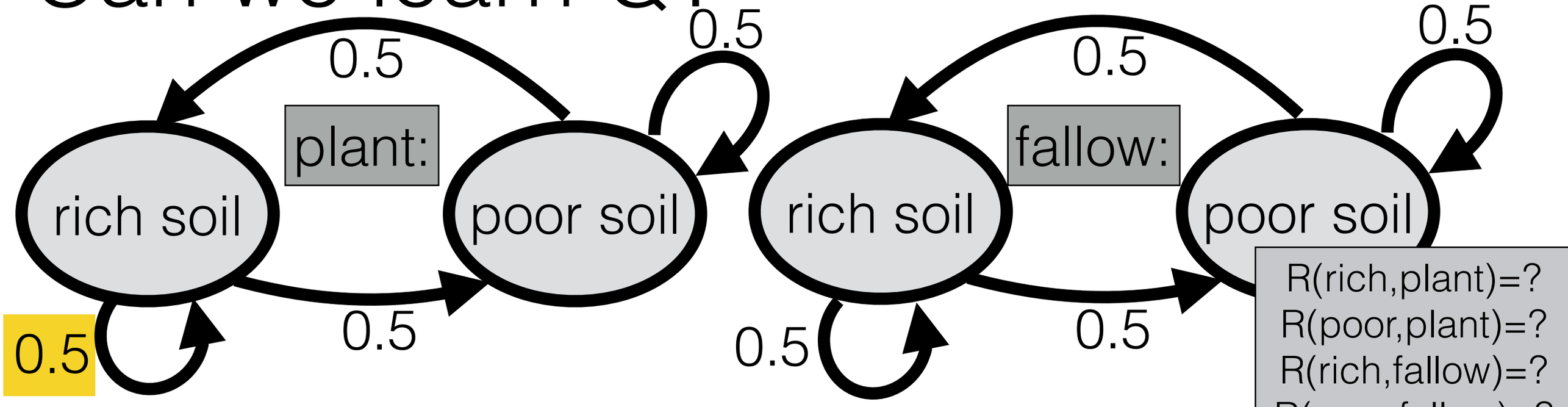
Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}$

```
Initialize s^(1) = s_0
Initialize: any s, a, s': T̂(s,a,s') = 1/|S|; R̂(s,a) = 0; Q
for t = 1, 2, 3, …
    a^(t) = select_action(s^(t), Q)
    r^(t), s^(t+1) = execute(a^(t))
    R̂(s^(t), a^(t)) = r^(t)
    Any s,a,s': T̂(s,a,s') = (1 + Σ_{i=1}^{t} 1{s^(i)=s, a^(i)=a, s^(i+1)=s'}) / (|S| + Σ_{i=1}^{t} 1{s^(i)=s, a^(i)=a})
    Q = infinite-horizon-value-iteration(R̂, T̂)
```

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

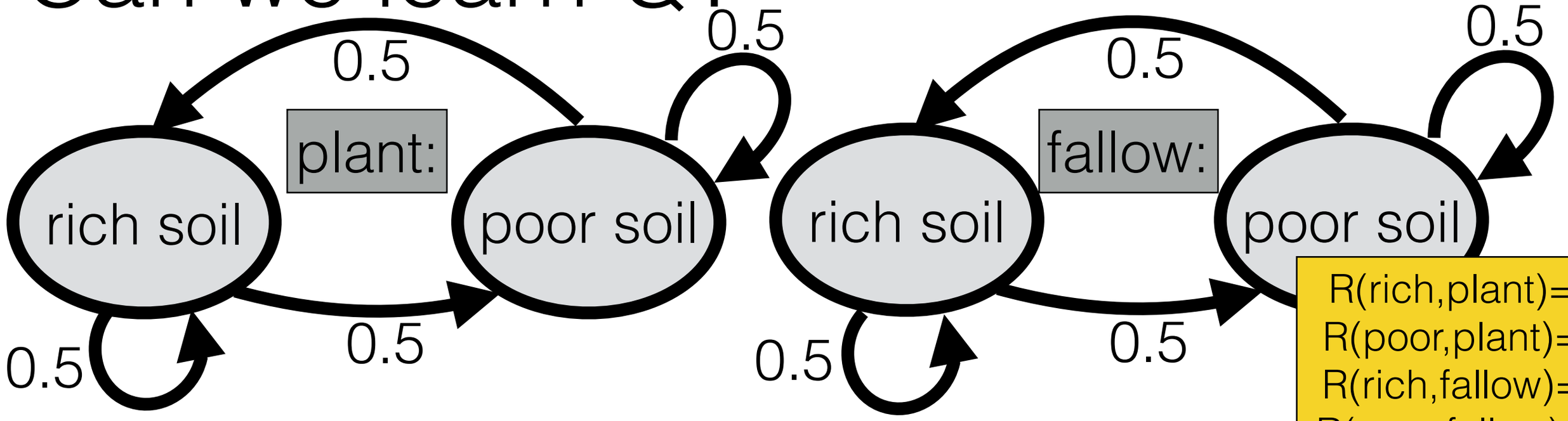
Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}$

```
Initialize s^(1) = s_0
Initialize: any s, a, s': T̂(s, a, s') = 1/|S|; R̂(s, a) = 0; Q
for t = 1, 2, 3, …
    a^(t) = select_action(s^(t), Q)
    r^(t), s^(t+1) = execute(a^(t))
    R̂(s^(t), a^(t)) = r^(t)
    Any s,a,s': T̂(s, a, s') = (1 + Σ_{i=1}^{t} 1{s^(i)=s, a^(i)=a, s^(i+1)=s'}) / (|S| + Σ_{i=1}^{t} 1{s^(i)=s, a^(i)=a})
    Q = infinite-horizon-value-iteration(R̂, T̂)
```

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}$

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s, a, s') = \frac{1}{|S|}; R(s, a) = 0; Q$

for t = 1, 2, 3, …

E.g. **ϵ-greedy**

$a^{(t)}$ = select_action($s^{(t)}, Q$)

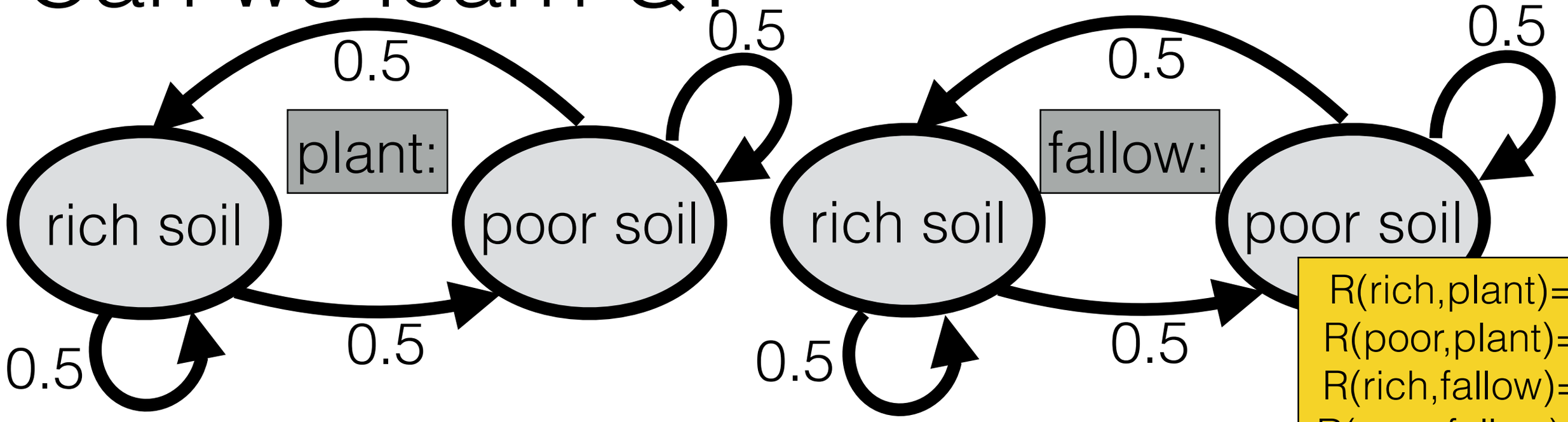
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}$

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s, a, s') = \frac{1}{|S|}; \hat{R}(s, a) = 0; Q$

for t = 1, 2, 3, …

E.g. **ϵ-greedy**

$a^{(t)}$ = select_action($s^{(t)}, Q$)

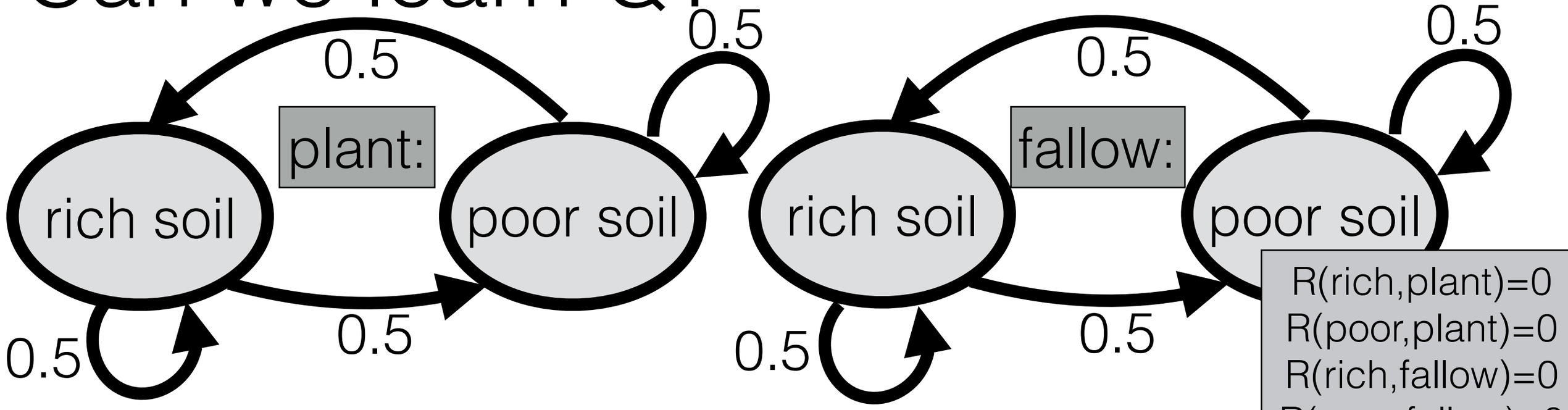
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}$

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

E.g. **ϵ-greedy**

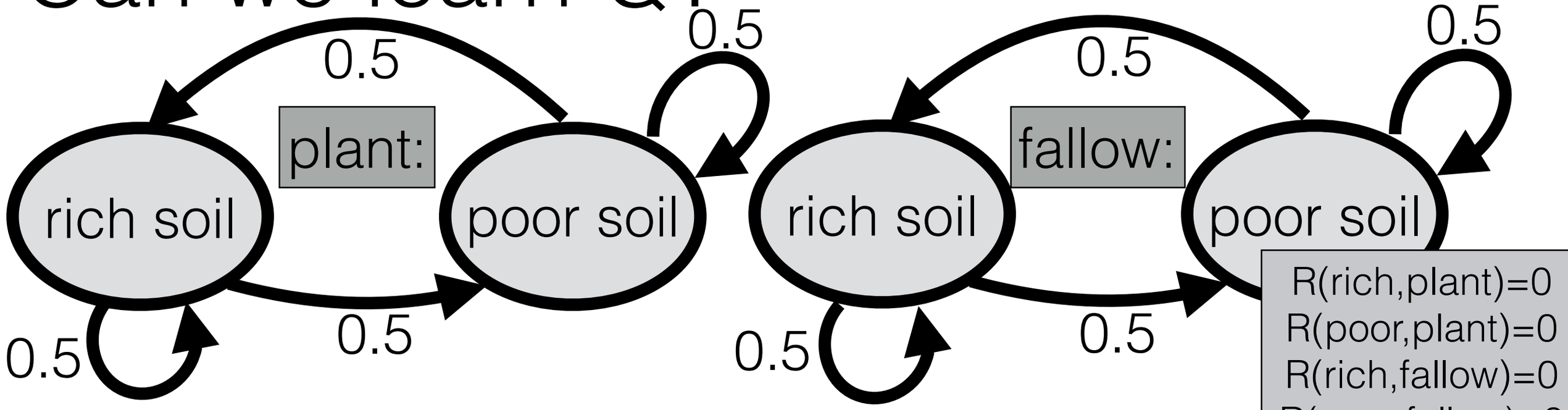
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich; explore}$

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s, a, s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s, a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

E.g. **ϵ-greedy**

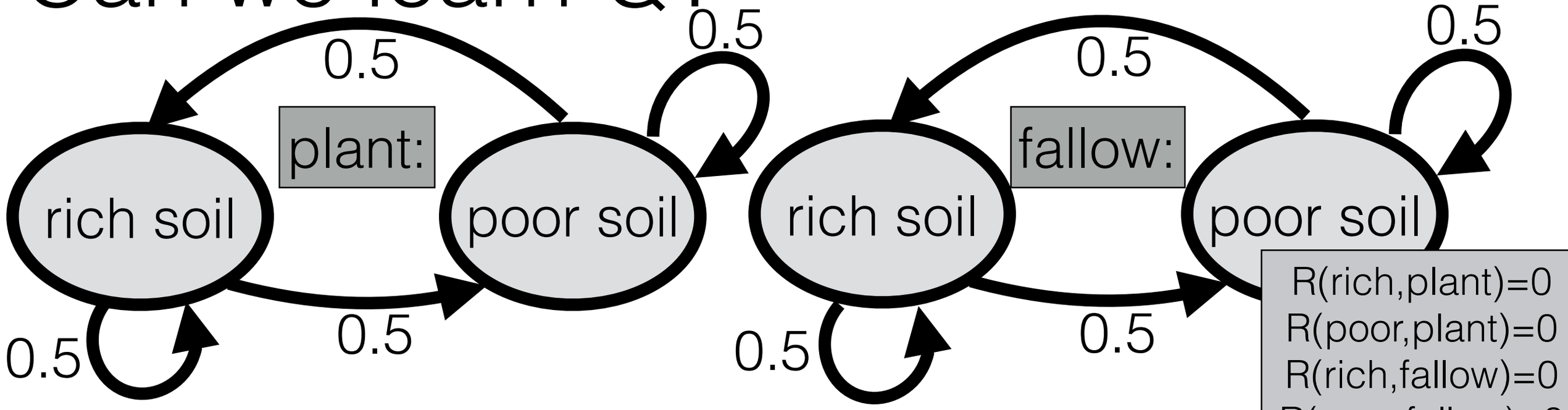
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

E.g. **ϵ-greedy**

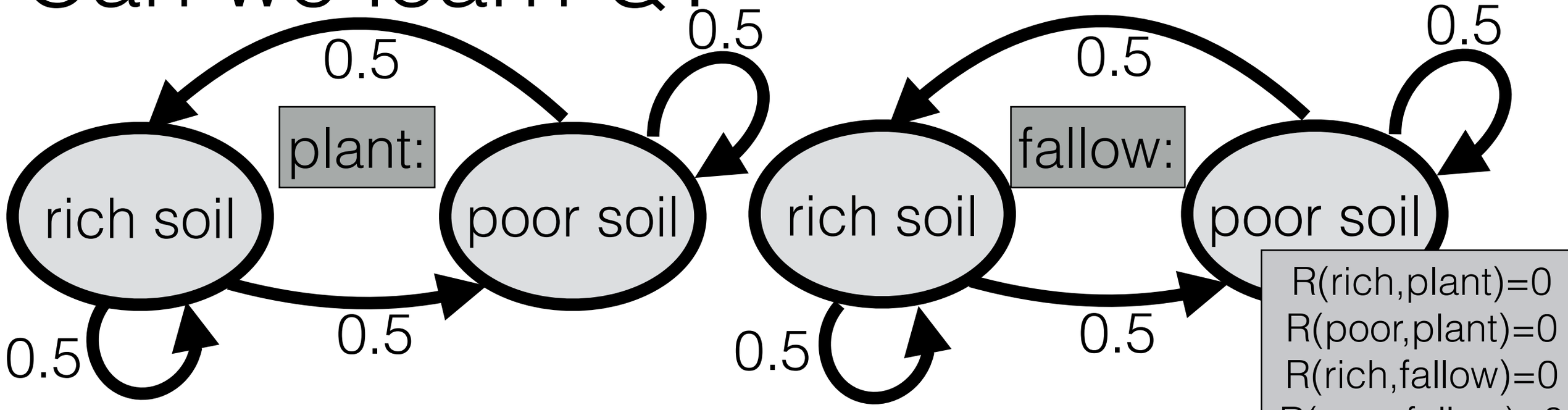
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}(s,a,s') = \frac{1}{|\mathcal{S}|}; \hat{R}(s,a) = 0; Q$

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

E.g. **ϵ-greedy**

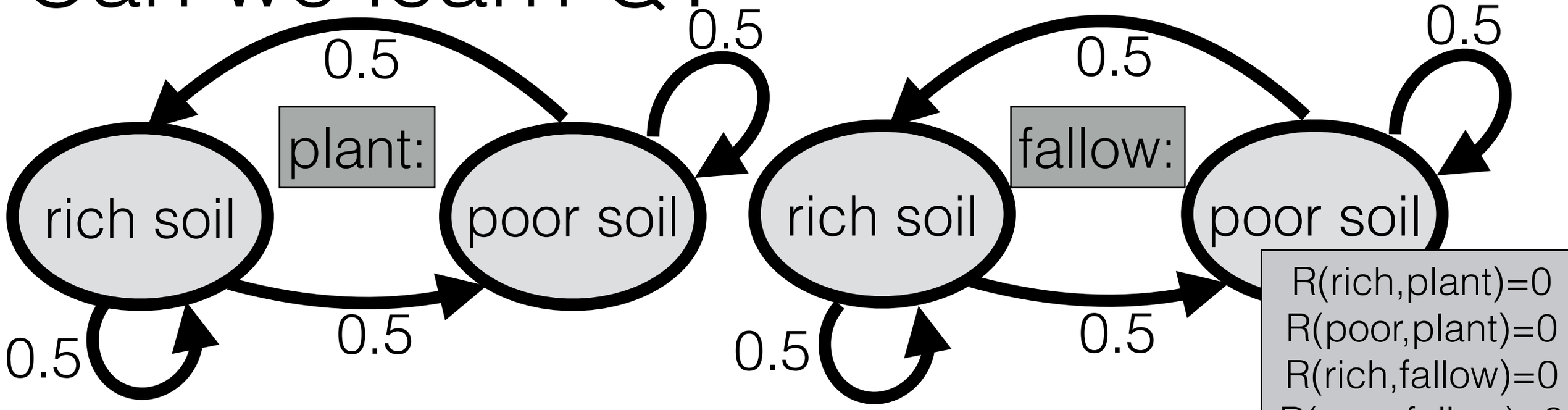
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s,a,s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$
$s^{(2)} = \text{poor}$

E.g. **ϵ-greedy**

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}($

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

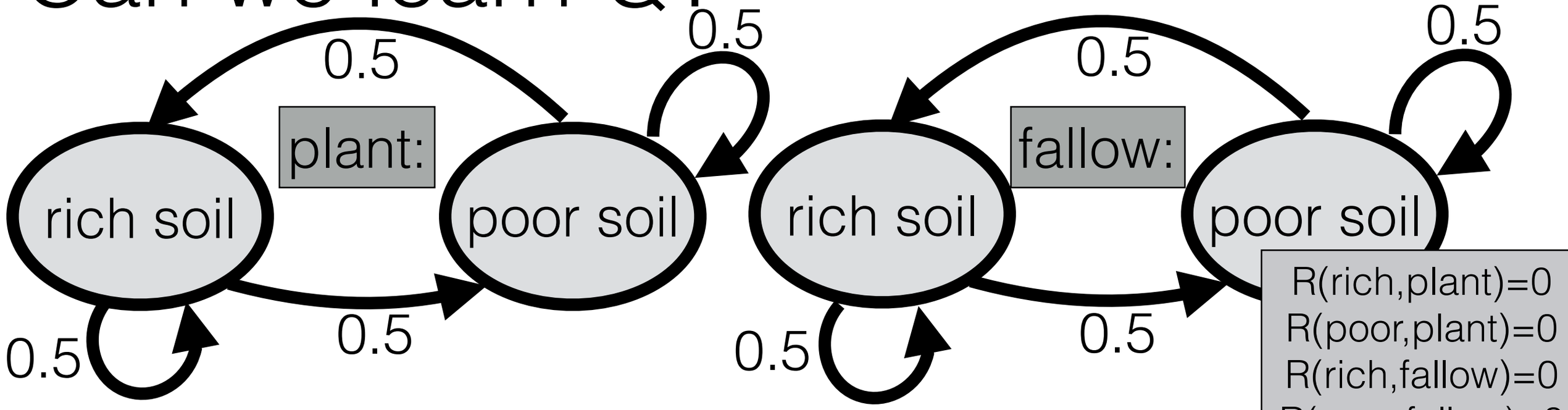
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$
$s^{(2)} = \text{poor}$

E.g. **ϵ-greedy**

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}($

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

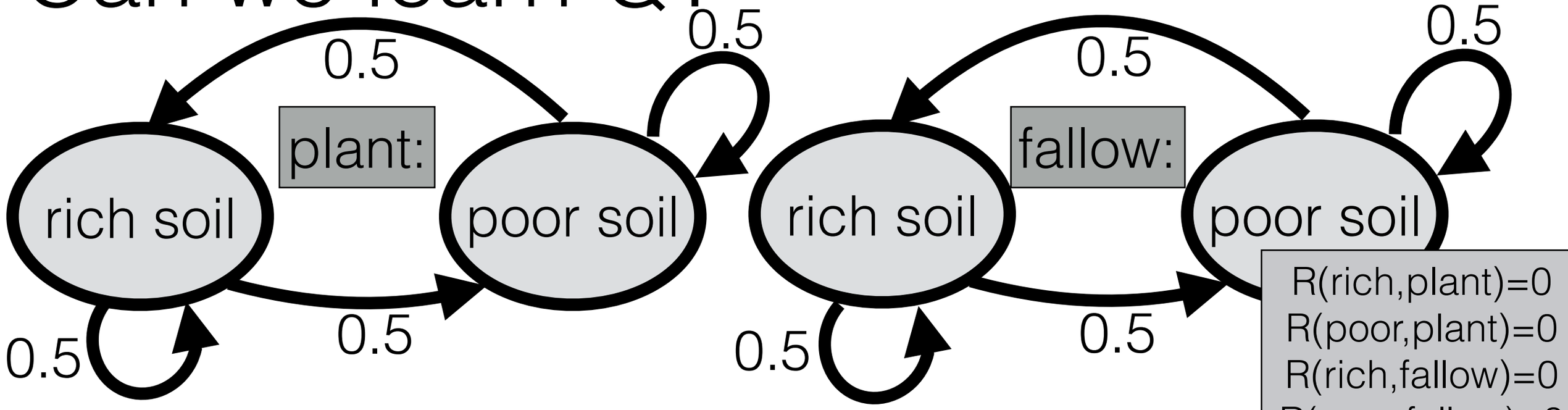
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$
$s^{(2)} = \text{poor}$

E.g. **ϵ-greedy**

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}($

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

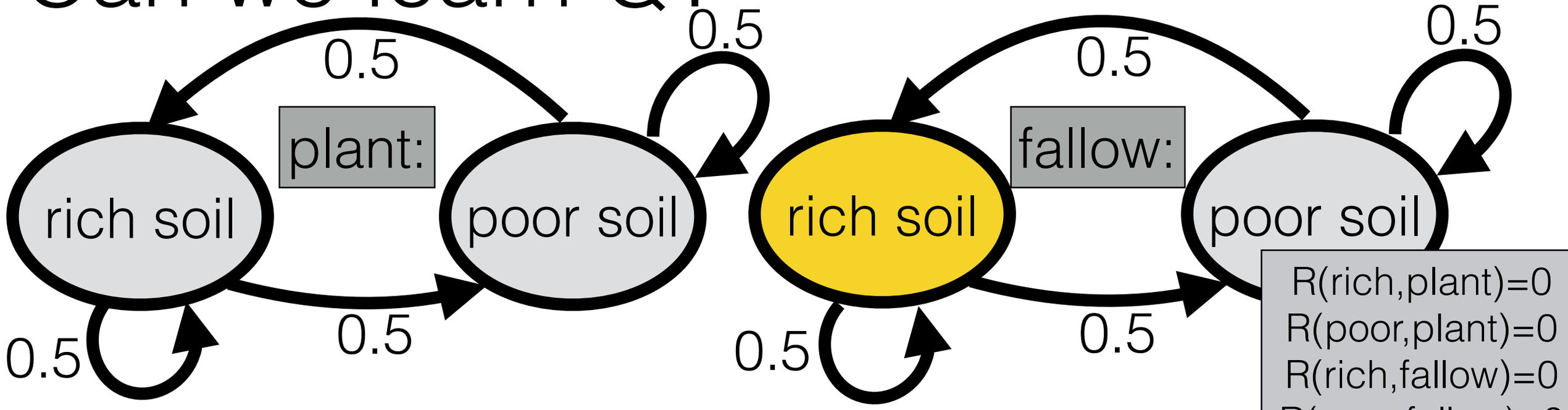
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$
$s^{(2)} = \text{poor}$

E.g. **ϵ-greedy**

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}($

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

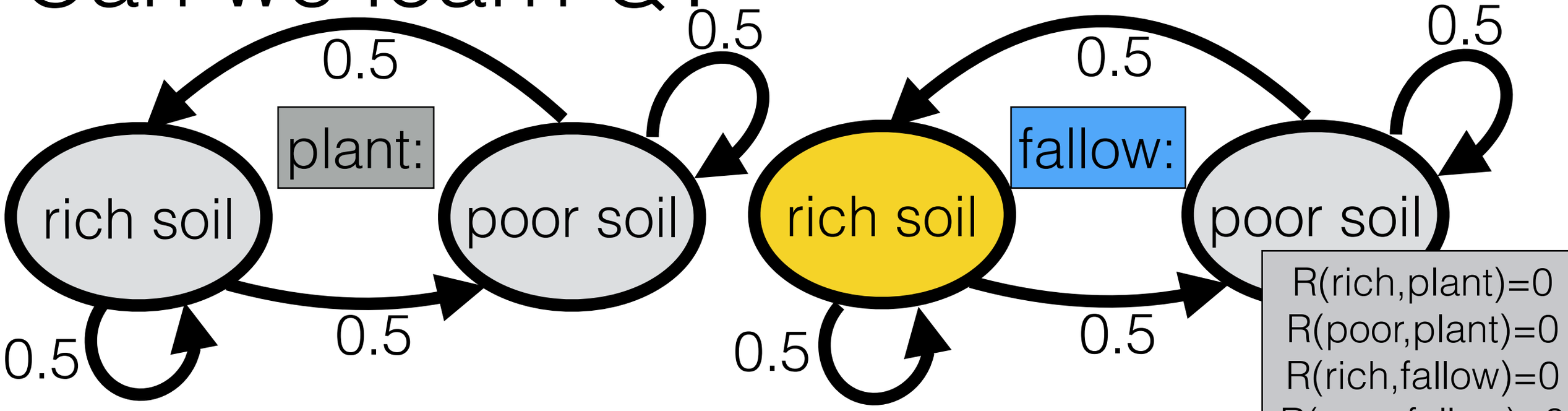
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$
$s^{(2)} = \text{poor}$

E.g. **ϵ-greedy**

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}($

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

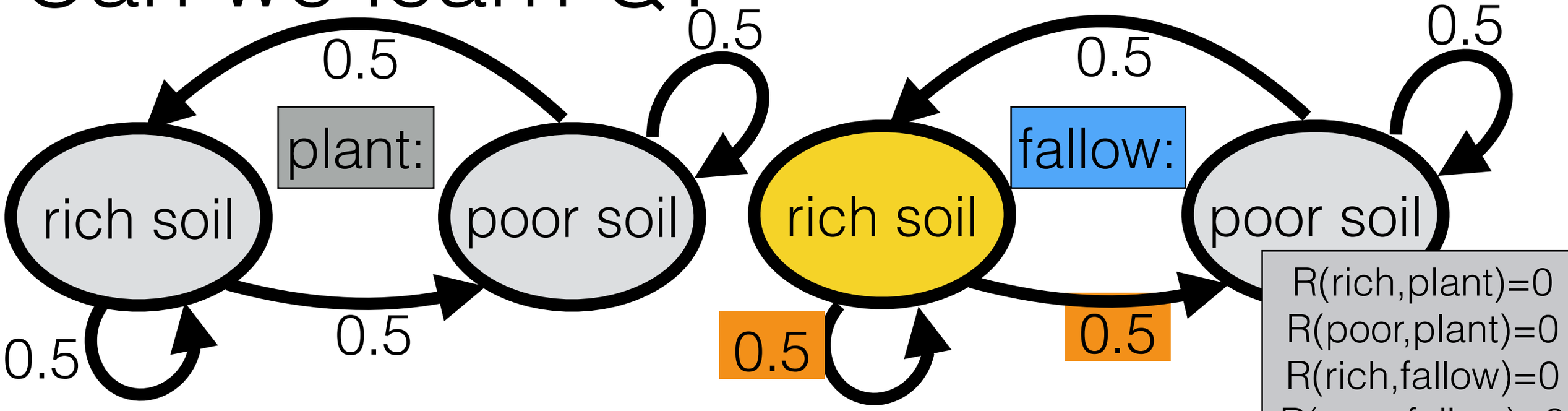
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$
$s^{(2)} = \text{poor}$

E.g. **ϵ-greedy**

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}($

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

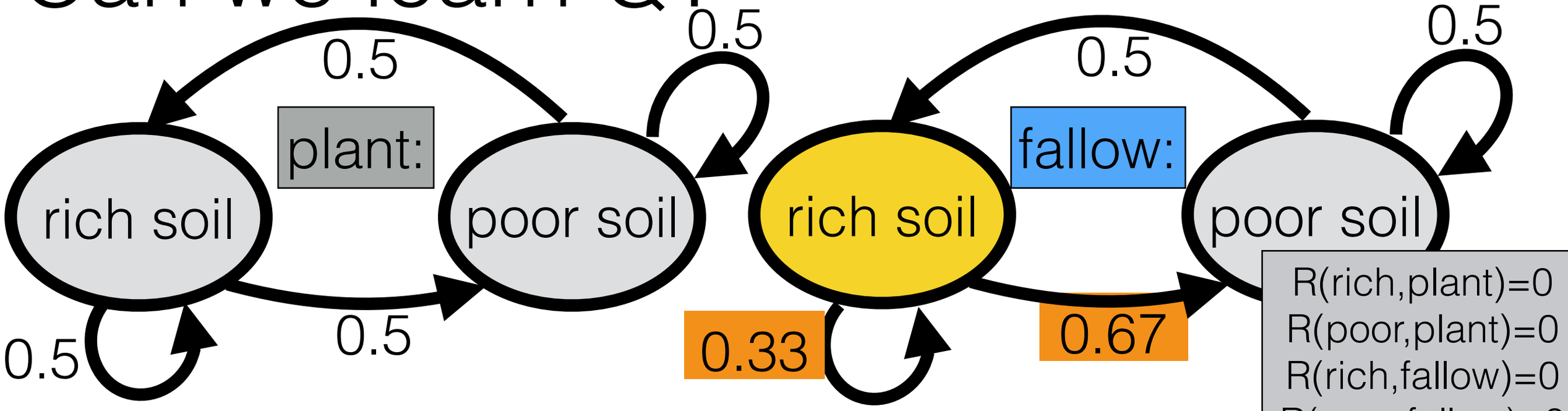
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$
$s^{(2)} = \text{poor}$

E.g. **ϵ-greedy**

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}($

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

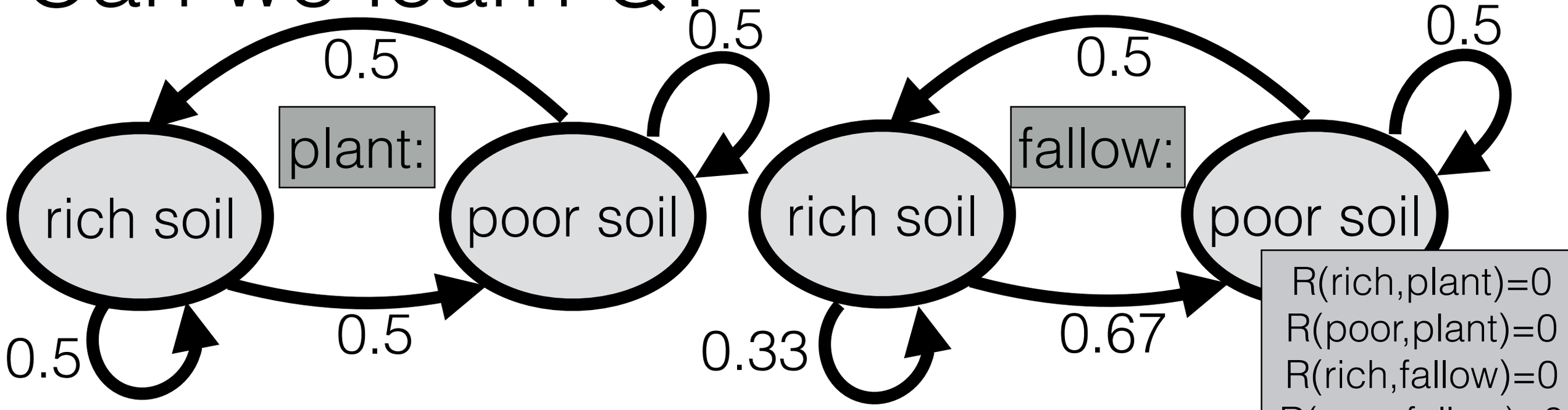
$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

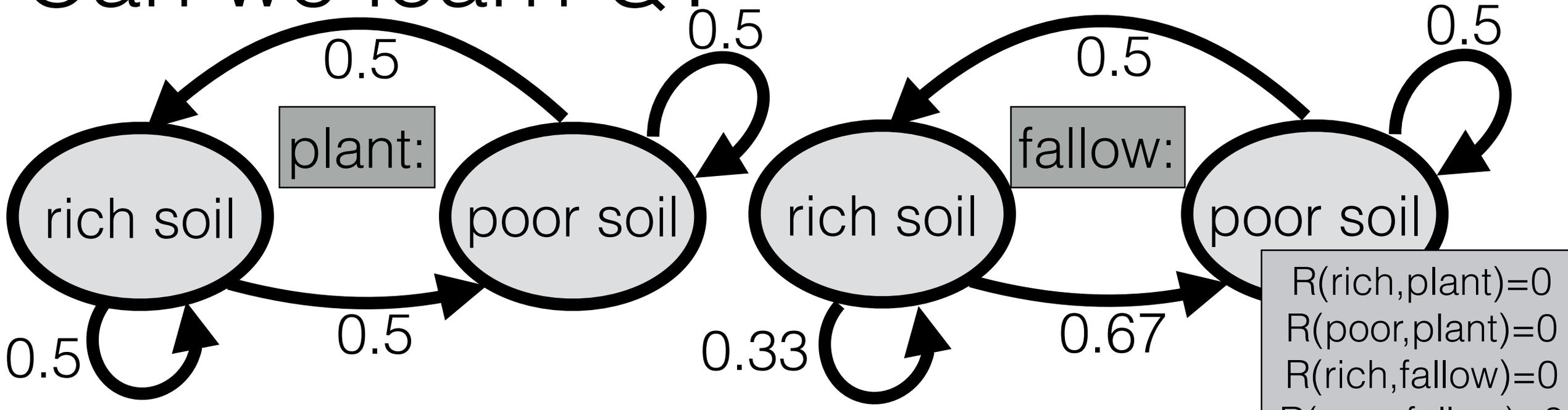
Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$
$s^{(2)} = \text{poor}$

E.g. **ϵ-greedy**

```
Initialize s^(1) = s_0
Initialize: any s, a, s': T̂(
for t = 1, 2, 3, …
    a^(t) = select_action(s^(t), Q)
    r^(t), s^(t+1) = execute(a^(t))
    R̂(s^(t), a^(t)) = r^(t)
    Any s, a, s': T̂(s,a,s') = (1 + Σ_{i=1}^{t} 1{s^(i)=s, a^(i)=a, s^(i+1)=s'}) / (|S| + Σ_{i=1}^{t} 1{s^(i)=s, a^(i)=a})
    Q = infinite-horizon-value-iteration(R̂, T̂)
```

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

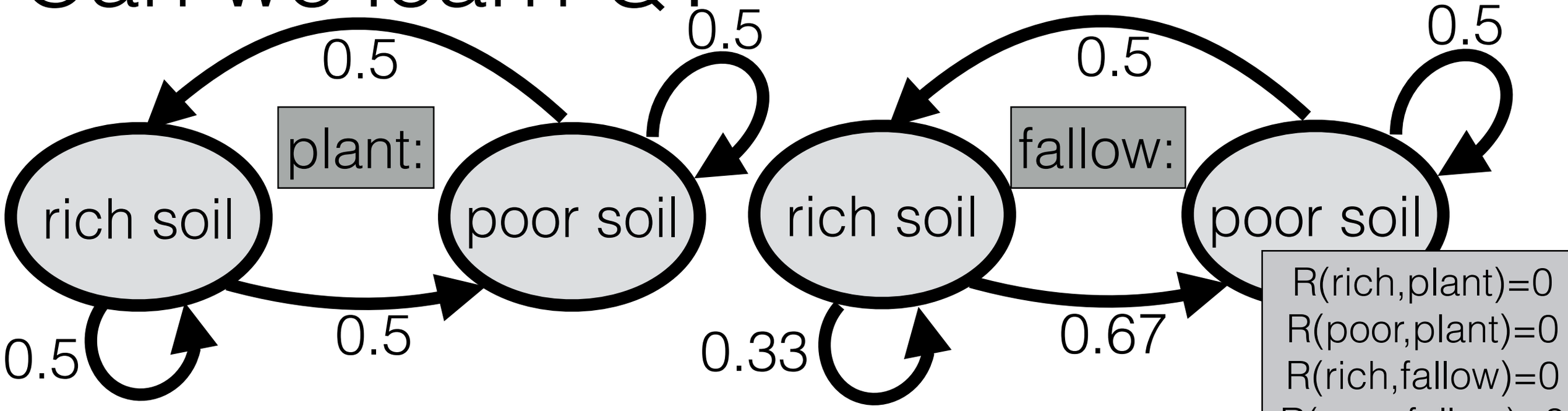
Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$
$s^{(2)} = \text{poor}$

E.g. **ϵ-greedy**

```
Initialize s^(1) = s_0
Initialize: any s, a, s': T̂(
for t = 1, 2, 3, …
  a^(t) = select_action(s^(t), Q)
  r^(t), s^(t+1) = execute(a^(t))
  R̂(s^(t), a^(t)) = r^(t)
  Any s,a,s': T̂(s,a,s') = (1+Σ_{i=1}^{t} 1{s^(i)=s,a^(i)=a,s^(i+1)=s'}) / (|S|+Σ_{i=1}^{t} 1{s^(i)=s,a^(i)=a})
  Q = infinite-horizon-value-iteration(R̂, T̂)
```

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

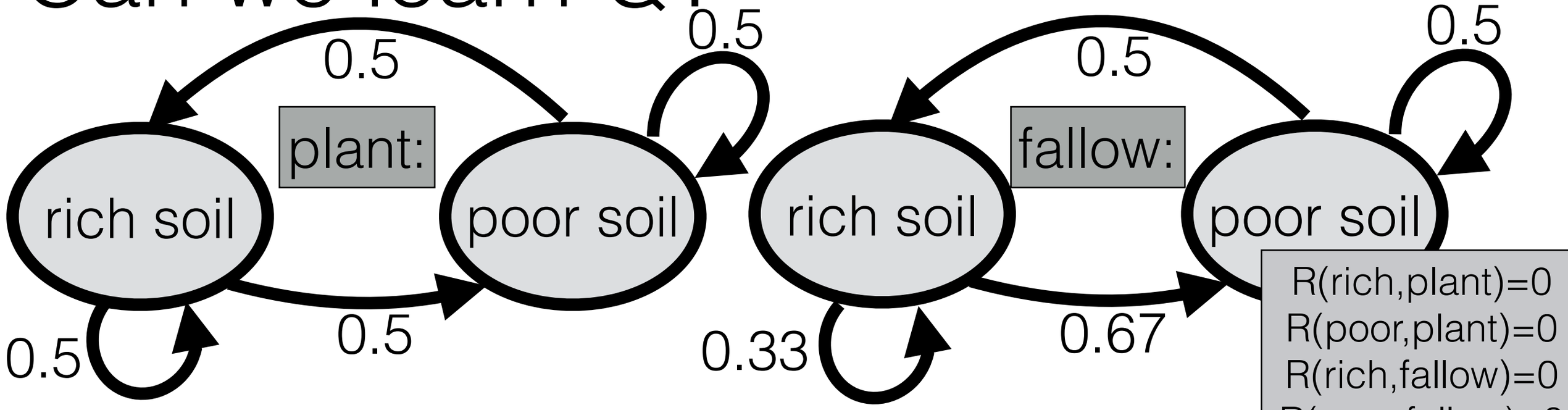
Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$
$s^{(2)} = \text{poor}$

E.g. **ϵ-greedy**

```
Initialize s^(1) = s_0
Initialize: any s, a, s': T̂(
for t = 1, 2, 3, …
  a^(t) = select_action(s^(t), Q)
  r^(t), s^(t+1) = execute(a^(t))
  R̂(s^(t), a^(t)) = r^(t)
  Any s,a,s': T̂(s,a,s') = (1+Σ_{i=1}^{t} 1{s^(i)=s,a^(i)=a,s^(i+1)=s'}) / (|S|+Σ_{i=1}^{t} 1{s^(i)=s,a^(i)=a})
  Q = infinite-horizon-value-iteration(R̂, T̂)
```

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

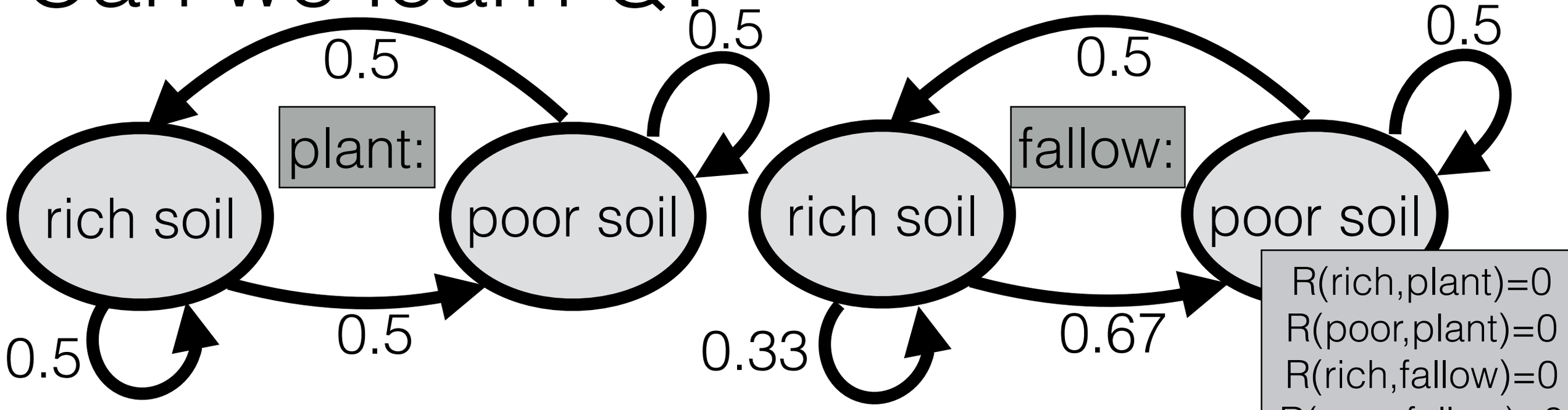
Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$
$s^{(2)} = \text{poor}; \text{exploit}$

E.g. **ϵ-greedy**

```
Initialize s^(1) = s_0
Initialize: any s, a, s': T̂(
for t = 1, 2, 3, …
    a^(t) = select_action(s^(t), Q)
    r^(t), s^(t+1) = execute(a^(t))
    R̂(s^(t), a^(t)) = r^(t)
    Any s,a,s': T̂(s,a,s') = (1+Σ_{i=1}^{t} 1{s^(i)=s,a^(i)=a,s^(i+1)=s'}) / (|S|+Σ_{i=1}^{t} 1{s^(i)=s,a^(i)=a})
    Q = infinite-horizon-value-iteration(R̂, T̂)
```

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

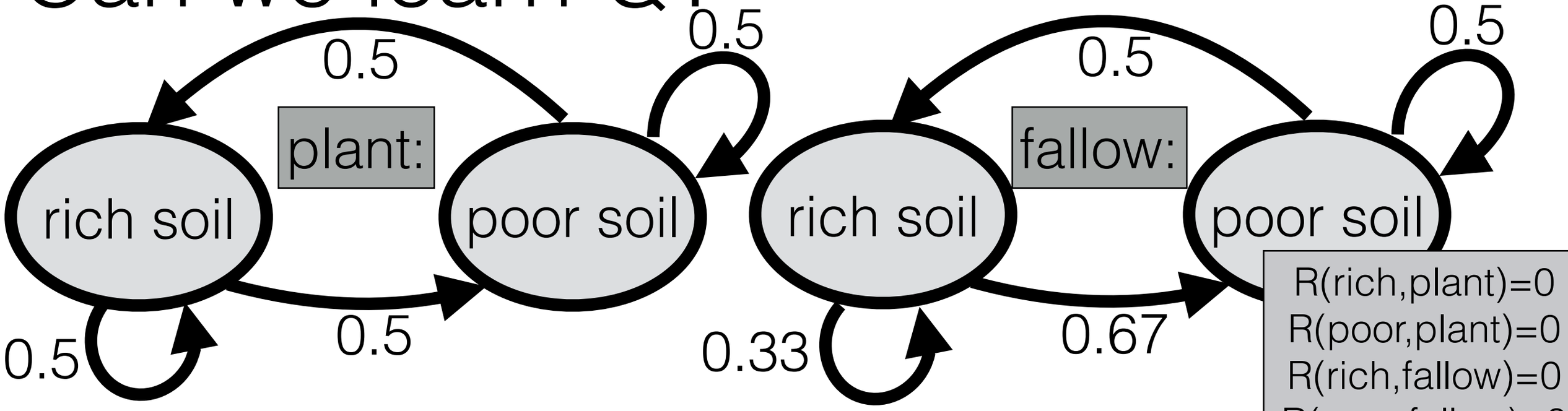
Data

Example: $\epsilon = 0.3; \gamma = 0.99$
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$
$s^{(2)} = \text{poor}; \text{exploit}; a^{(2)} = \text{fallow}$

E.g. **ϵ-greedy**

```
Initialize s^(1) = s_0
Initialize: any s, a, s': T̂(
for t = 1, 2, 3, …
    a^(t) = select_action(s^(t), Q)
    r^(t), s^(t+1) = execute(a^(t))
    R̂(s^(t), a^(t)) = r^(t)
    Any s,a,s': T̂(s, a, s') = (1 + Σ_{i=1}^{t} 1{s^(i)=s, a^(i)=a, s^(i+1)=s'}) / (|S| + Σ_{i=1}^{t} 1{s^(i)=s, a^(i)=a})
    Q = infinite-horizon-value-iteration(R̂, T̂)
```

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}($

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Data

Example: $\epsilon = 0.3; \gamma = 0.99$

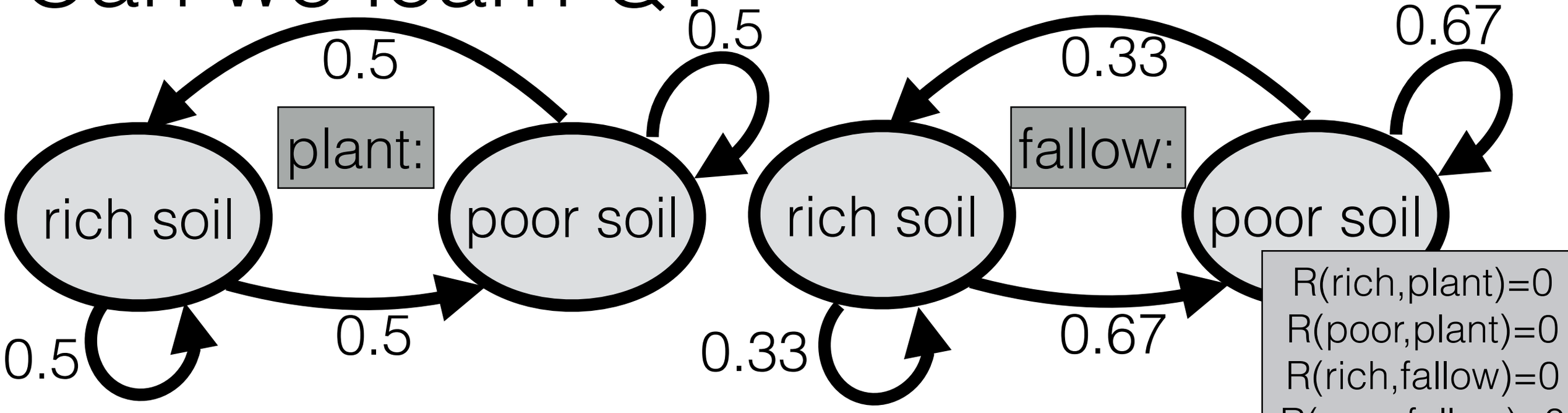
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$

$s^{(2)} = \text{poor}; \text{exploit}; a^{(2)} = \text{fallow}$

$s^{(3)} = \text{poor}$

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}($

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}, Q$)

$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Data

Example: $\epsilon = 0.3; \gamma = 0.99$

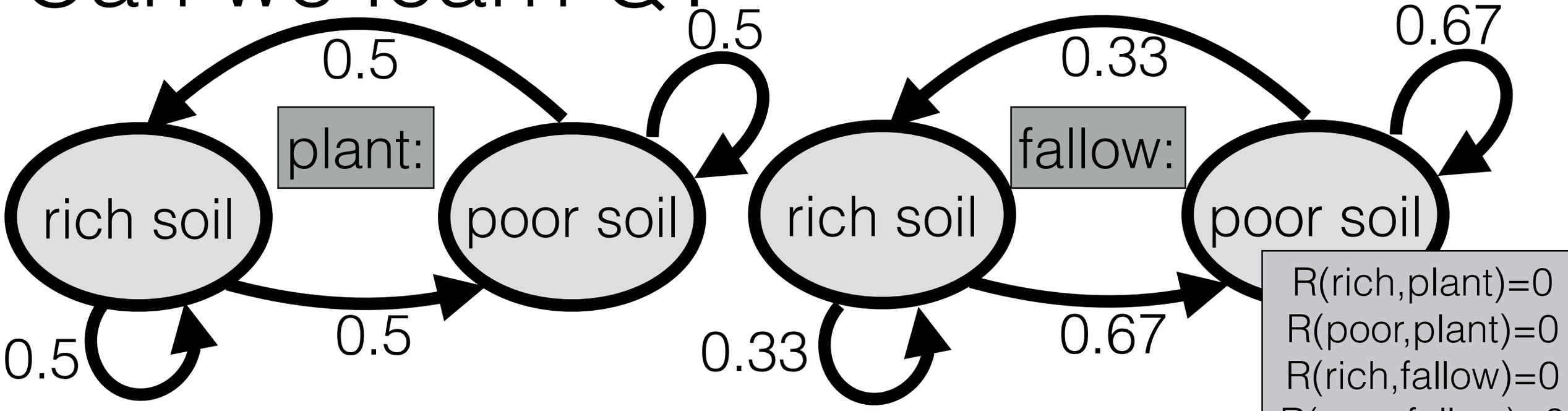
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$

$s^{(2)} = \text{poor}; \text{exploit}; a^{(2)} = \text{fallow}$

$s^{(3)} = \text{poor}$

E.g. **ϵ-greedy**

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$

$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$

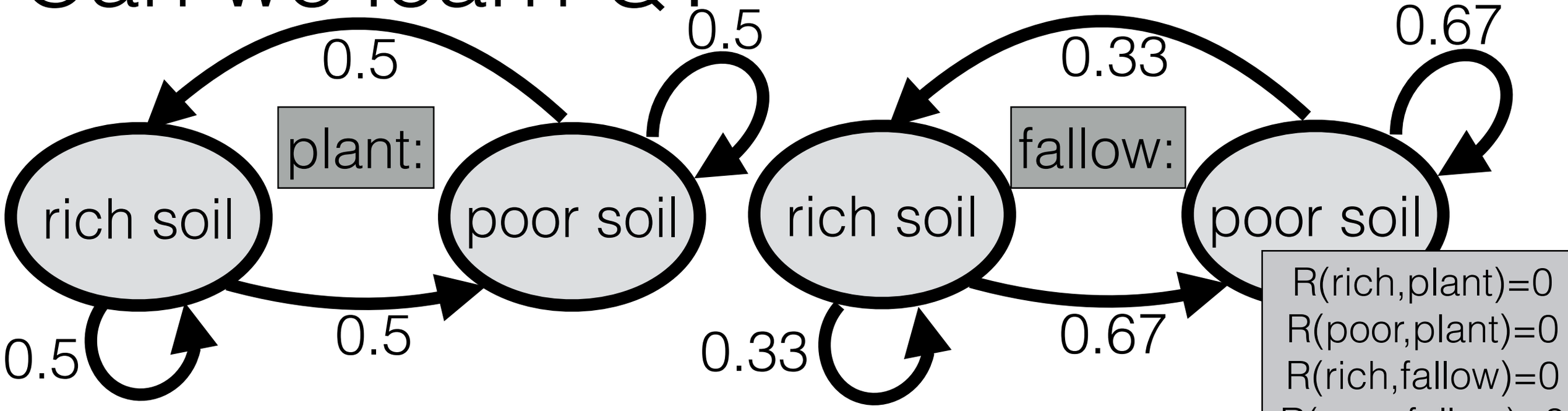
$s^{(2)} = \text{poor}; \text{exploit}; a^{(2)} = \text{fallow}$

$s^{(3)} = \text{poor}; \text{exploit}; a^{(3)} = \text{fallow}$

E.g. **ϵ-greedy**

```
Initialize s^(1) = s_0
Initialize: any s, a, s': T̂(
for t = 1, 2, 3, …
    a^(t) = select_action(s^(t), Q)
    r^(t), s^(t+1) = execute(a^(t))
    R̂(s^(t), a^(t)) = r^(t)
    Any s,a,s': T̂(s,a,s') = (1+Σ_{i=1}^{t} 1{s^(i)=s, a^(i)=a, s^(i+1)=s'}) / (|S|+Σ_{i=1}^{t} 1{s^(i)=s, a^(i)=a})
    Q = infinite-horizon-value-iteration(R̂, T̂)
```

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}($

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}$

$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Data

Example: $\epsilon = 0.3; \gamma = 0.99$

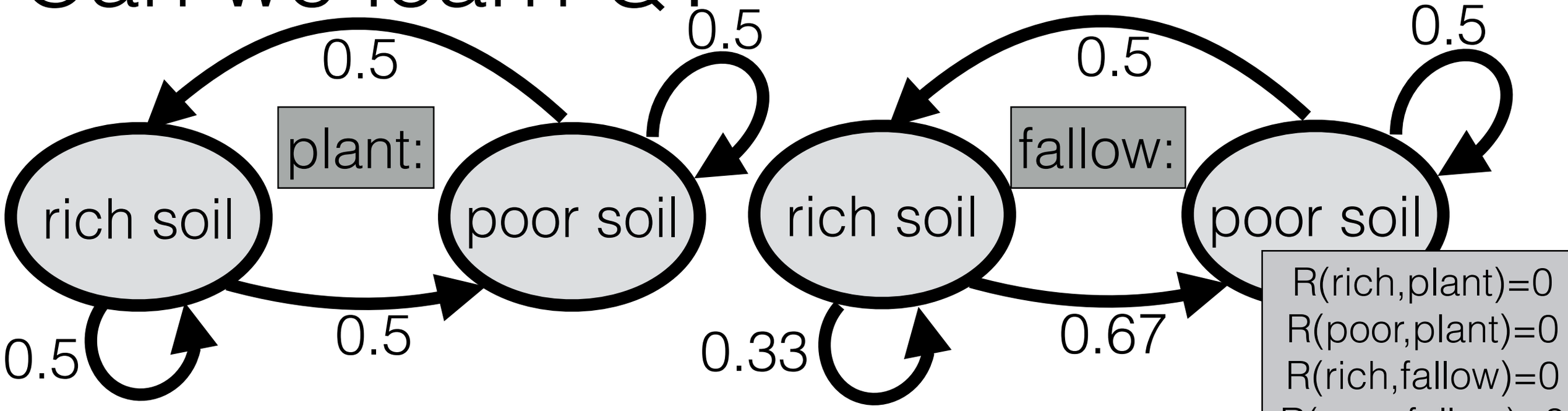
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$

$s^{(2)} = \text{poor}; \text{exploit}; a^{(2)} = \text{fallow}$

$s^{(3)} = \text{poor}; \text{exploit}; a^{(3)} = \text{fallow}$

$s^{(4)} = \text{rich}$

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}($

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}$

$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Data

Example: $\epsilon = 0.3; \gamma = 0.99$

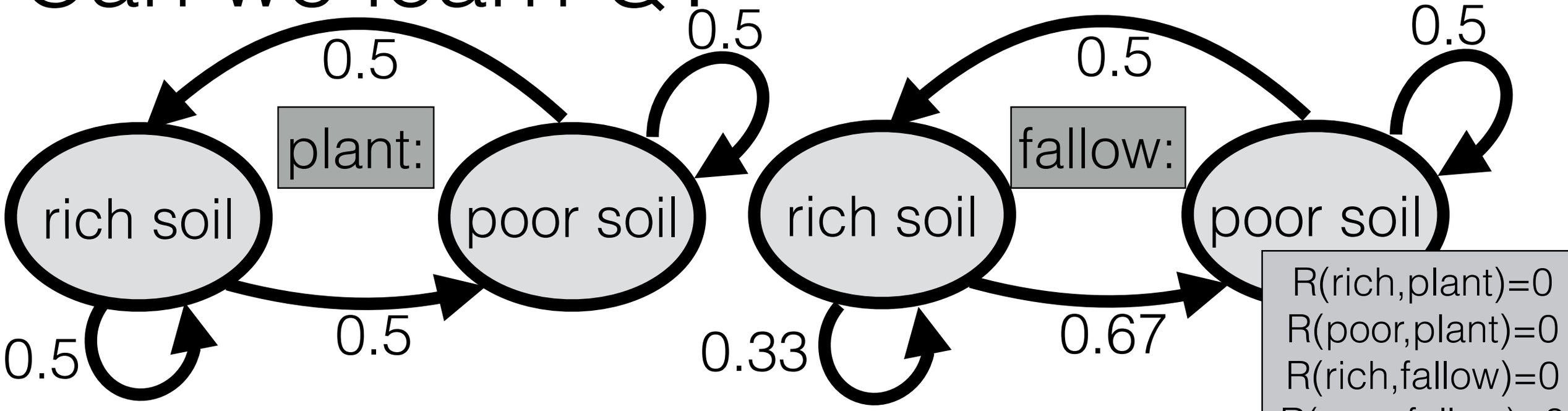
$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$

$s^{(2)} = \text{poor}; \text{exploit}; a^{(2)} = \text{fallow}$

$s^{(3)} = \text{poor}; \text{exploit}; a^{(3)} = \text{fallow}$

$s^{(4)} = \text{rich}$

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Data

Example: $\epsilon = 0.3; \gamma = 0.99$

$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$

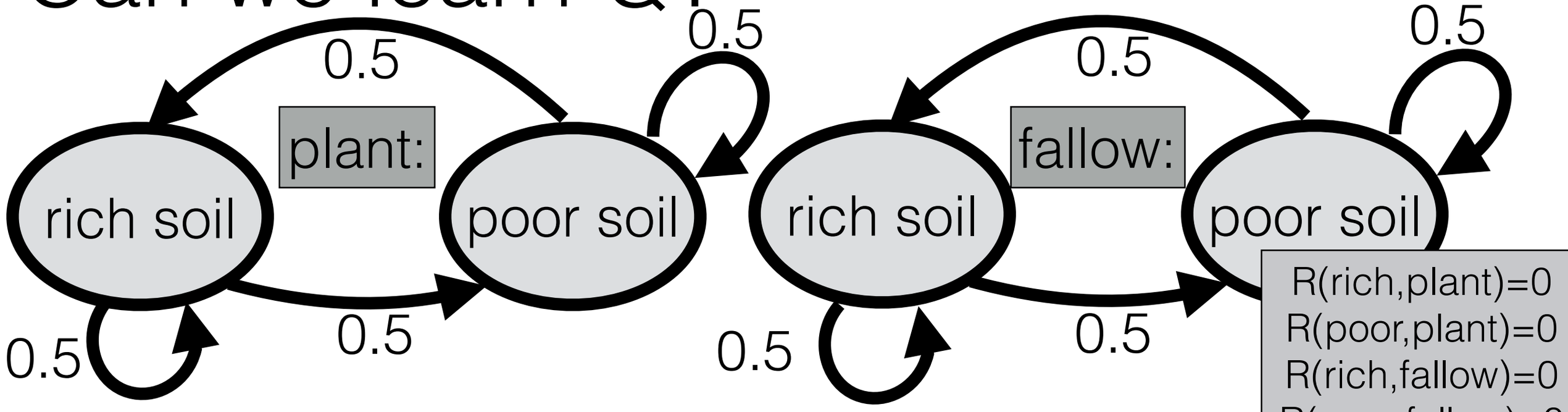
$s^{(2)} = \text{poor}; \text{exploit}; a^{(2)} = \text{fallow}$

$s^{(3)} = \text{poor}; \text{exploit}; a^{(3)} = \text{fallow}$

$s^{(4)} = \text{rich}; \text{exploit}; a^{(4)} = \text{fallow}$

```
Initialize s^(1) = s_0
Initialize: any s, a, s': T̂(
for t = 1, 2, 3, …
    a^(t) = select_action(s^(t)
    r^(t), s^(t+1) = execute(a^(t))
    R̂(s^(t), a^(t)) = r^(t)
    Any s,a,s': T̂(s,a,s') = (1+Σ_{i=1}^{t} 1{s^(i)=s,a^(i)=a,s^(i+1)=s'}) / (|S|+Σ_{i=1}^{t} 1{s^(i)=s,a^(i)=a})
    Q = infinite-horizon-value-iteration(R̂, T̂)
```

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}($

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}$

$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Data

Example: $\epsilon = 0.3; \gamma = 0.99$

$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$

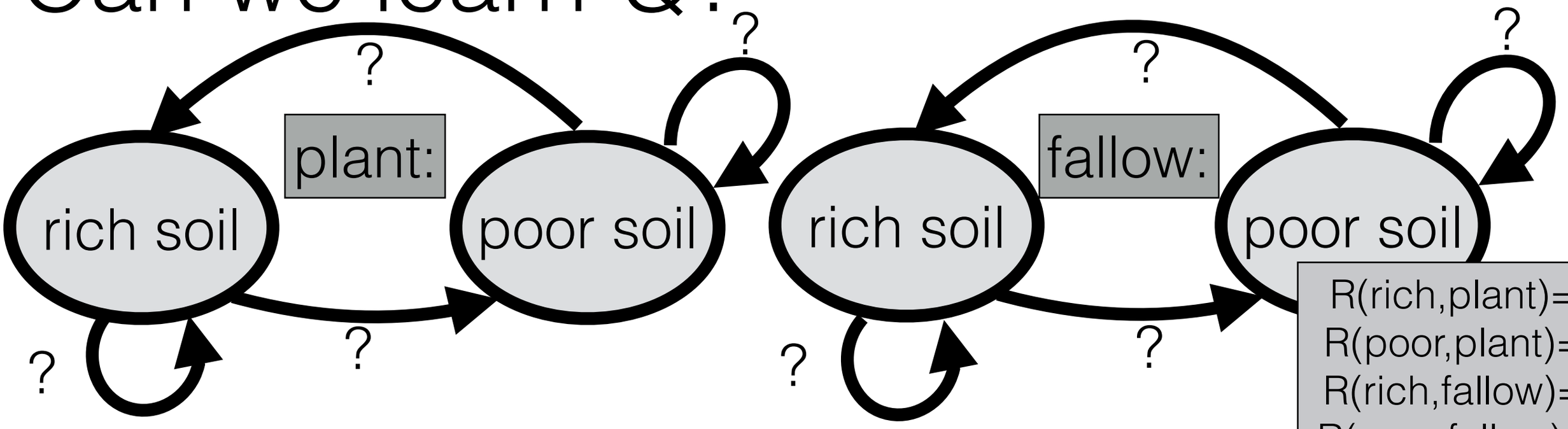
$s^{(2)} = \text{poor}; \text{exploit}; a^{(2)} = \text{fallow}$

$s^{(3)} = \text{poor}; \text{exploit}; a^{(3)} = \text{fallow}$

$s^{(4)} = \text{rich}; \text{exploit}; a^{(4)} = \text{fallow}$

$s^{(5)} = \text{rich}; \text{explore}; a^{(5)} = \text{plant}$

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

Initialize $s^{(1)} = s_0$

Initialize: any s, a, s': $\hat{T}($

for t = 1, 2, 3, …

$a^{(t)}$ = select_action($s^{(t)}$,

$r^{(t)}, s^{(t+1)}$ = execute($a^{(t)}$)

$\hat{R}(s^{(t)}, a^{(t)}) = r^{(t)}$

Any s,a,s′: $\hat{T}(s, a, s') = \frac{1+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a, s^{(i+1)}=s'\}}{|\mathcal{S}|+\sum_{i=1}^{t} \mathbf{1}\{s^{(i)}=s, a^{(i)}=a\}}$

Q = infinite-horizon-value-iteration($\hat{R}, \hat{T}$)

Data

Example: $\epsilon = 0.3; \gamma = 0.99$

$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$

$s^{(2)} = \text{poor}; \text{exploit}; a^{(2)} = \text{fallow}$

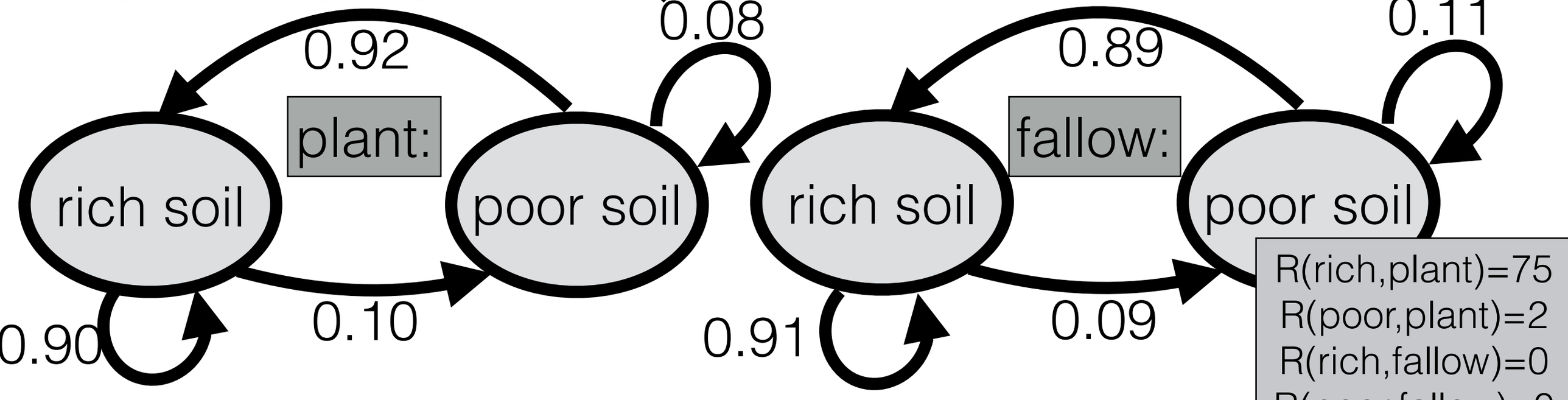
$s^{(3)} = \text{poor}; \text{exploit}; a^{(3)} = \text{fallow}$

$s^{(4)} = \text{rich}; \text{exploit}; a^{(4)} = \text{fallow}$

$s^{(5)} = \text{rich}; \text{explore}; a^{(5)} = \text{plant}$

…

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

```
Initialize s^(1) = s_0
Initialize: any s, a, s': T̂(
for t = 1, 2, 3, …
  a^(t) = select_action(s^(t)
  r^(t), s^(t+1) = execute(a^(t))
  R̂(s^(t), a^(t)) = r^(t)
  Any s,a,s': T̂(s,a,s') = 1
  Q = infinite-horizon-value-iteration(R̂, T̂)
```

Data

Example: $\epsilon = 0.3; \gamma = 0.99$

$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$

$s^{(2)} = \text{poor}; \text{exploit}; a^{(2)} = \text{fallow}$

$s^{(3)} = \text{poor}; \text{exploit}; a^{(3)} = \text{fallow}$

$s^{(4)} = \text{rich}; \text{exploit}; a^{(4)} = \text{fallow}$

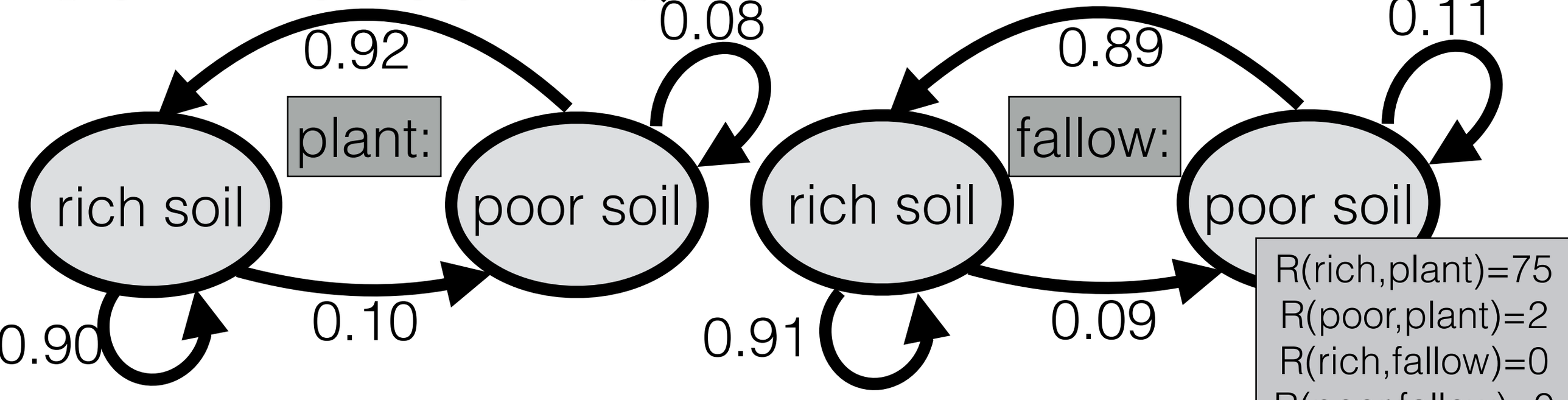
$s^{(5)} = \text{rich}; \text{explore}; a^{(5)} = \text{plant}$

…

$s^{(2718)} = \text{rich}; \text{exploit}; a^{(2718)} = \text{plant}$

$$\frac{|S| + \sum_{i=1}^{t} \mathbf{1}\{s^{(t)}=s, a^{(t)}=a\}}{}$$

Can we learn Q?



- Option 1: Estimate transition model T and reward function R

```
Initialize s^(1) = s_0
Initialize: any s, a, s': T̂(
for t = 1, 2, 3, …
  a^(t) = select_action(s^(t)
  r^(t), s^(t+1) = execute(a^(t))
  R̂(s^(t), a^(t)) = r^(t)
  Any s,a,s': T̂(s,a,s') = 1
  Q = infinite-horizon-value-iteration(R̂, T̂)
```

Data

Example: $\epsilon = 0.3; \gamma = 0.99$

$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$

$s^{(2)} = \text{poor}; \text{exploit}; a^{(2)} = \text{fallow}$

$s^{(3)} = \text{poor}; \text{exploit}; a^{(3)} = \text{fallow}$

$s^{(4)} = \text{rich}; \text{exploit}; a^{(4)} = \text{fallow}$

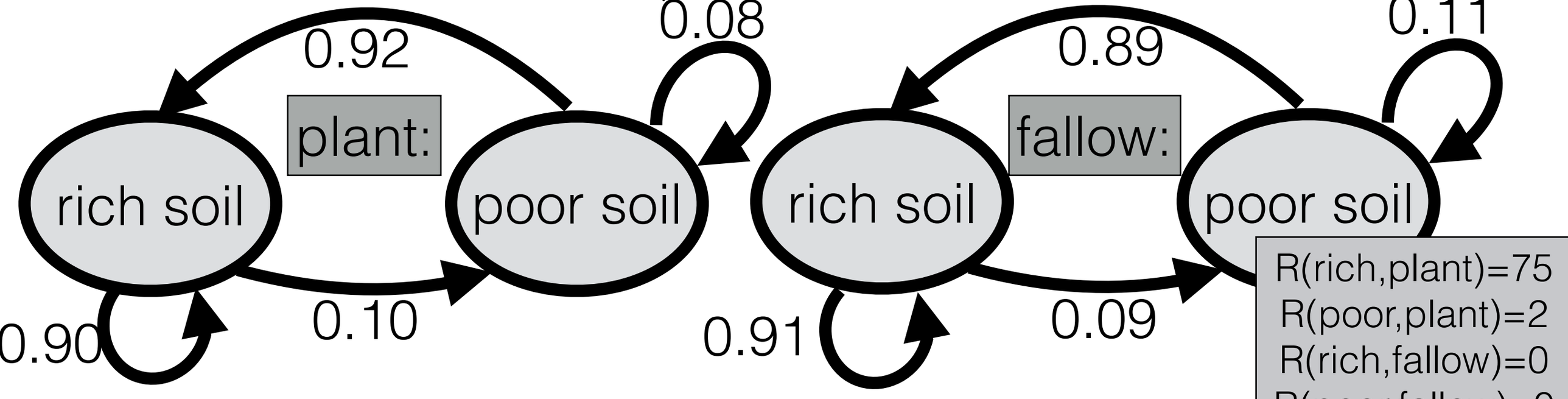
$s^{(5)} = \text{rich}; \text{explore}; a^{(5)} = \text{plant}$

…

$s^{(2718)} = \text{rich}; \text{exploit}; a^{(2718)} = \text{plant}$

$s^{(2719)} = \text{poor}; \text{exploit}; a^{(2719)} = \text{fallow}$

Can we learn *Q*?



- Option 1: Estimate transition model *T* and reward function *R*

```
Initialize s^(1) = s_0
Initialize: any s, a, s': T̂(
for t = 1, 2, 3, …
  a^(t) = select_action(s^(t)
  r^(t), s^(t+1) = execute(a^(t))
  R̂(s^(t), a^(t)) = r^(t)
  Any s,a,s': T̂(s, a, s') = 1
  Q = infinite-horizon-v
```

Data

Example: $\epsilon = 0.3; \gamma = 0.99$

$s^{(1)} = \text{rich}; \text{explore}; a^{(1)} = \text{fallow}$

$s^{(2)} = \text{poor}; \text{exploit}; a^{(2)} = \text{fallow}$

$s^{(3)} = \text{poor}; \text{exploit}; a^{(3)} = \text{fallow}$

$s^{(4)} = \text{rich}; \text{exploit}; a^{(4)} = \text{fallow}$

$s^{(5)} = \text{rich}; \text{explore}; a^{(5)} = \text{plant}$

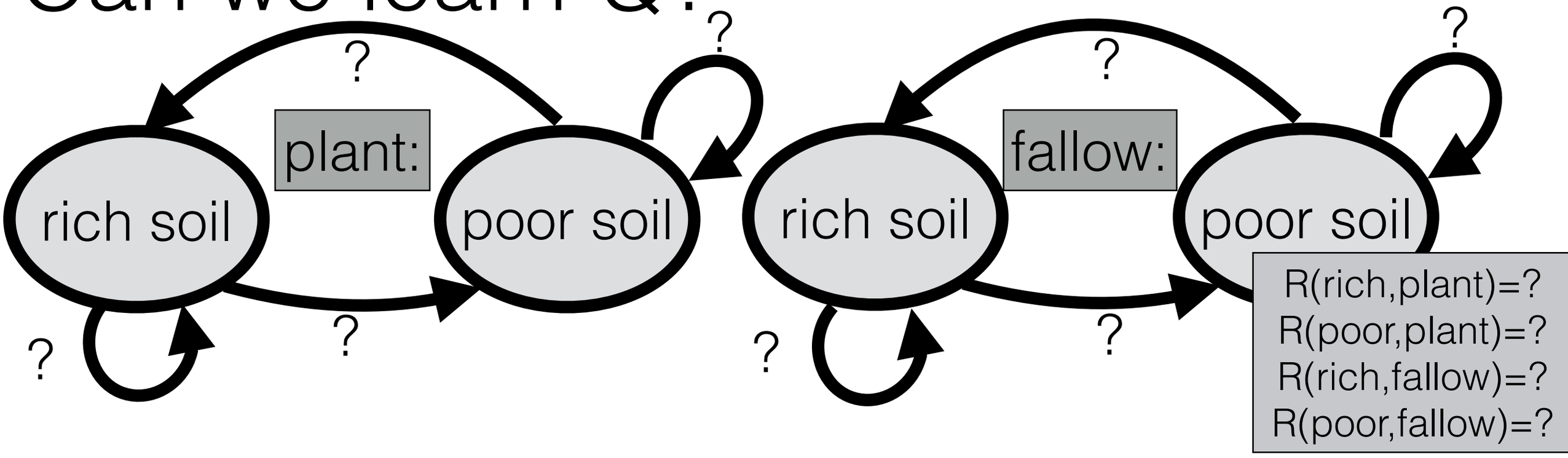
…

$s^{(2718)} = \text{rich}; \text{exploit}; a^{(2718)} = \text{plant}$

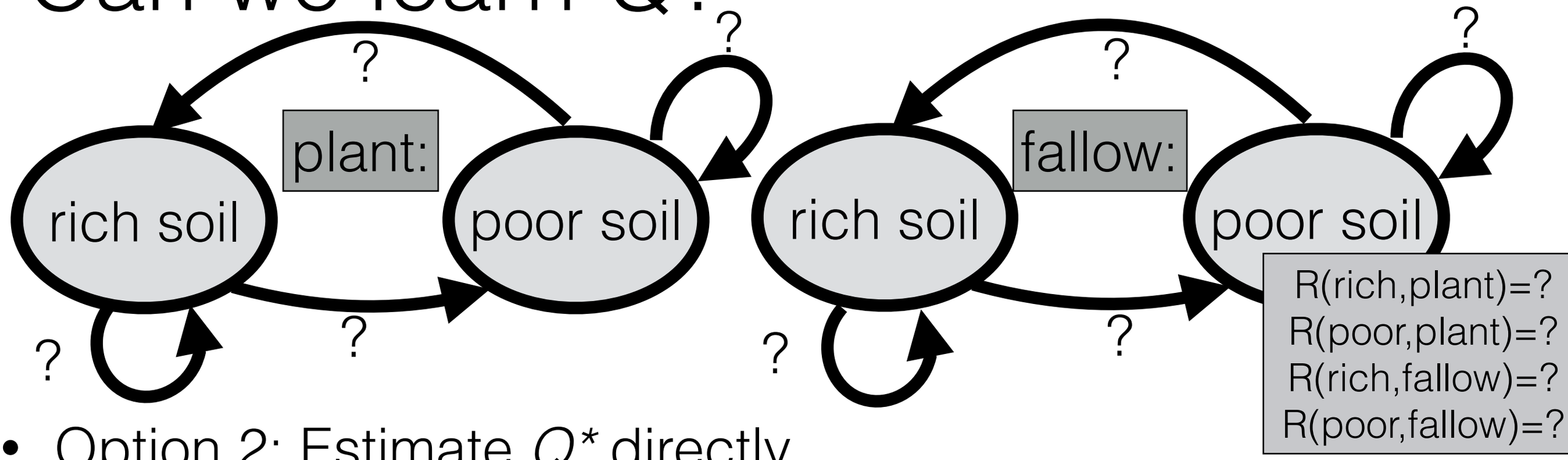
$s^{(2719)} = \text{poor}; \text{exploit}; a^{(2719)} = \text{fallow}$

$s^{(2720)} = \text{rich}; \text{explore}; a^{(2720)} = \text{fallow}$

Can we learn *Q*?

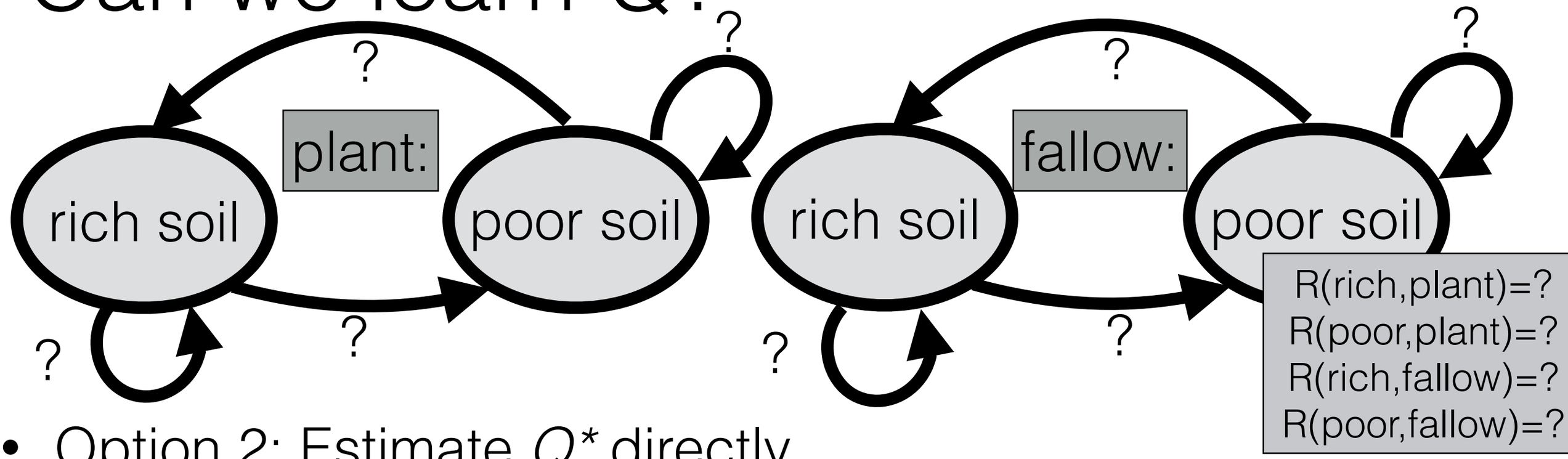


Can we learn *Q*?



- Option 2: Estimate *Q** directly

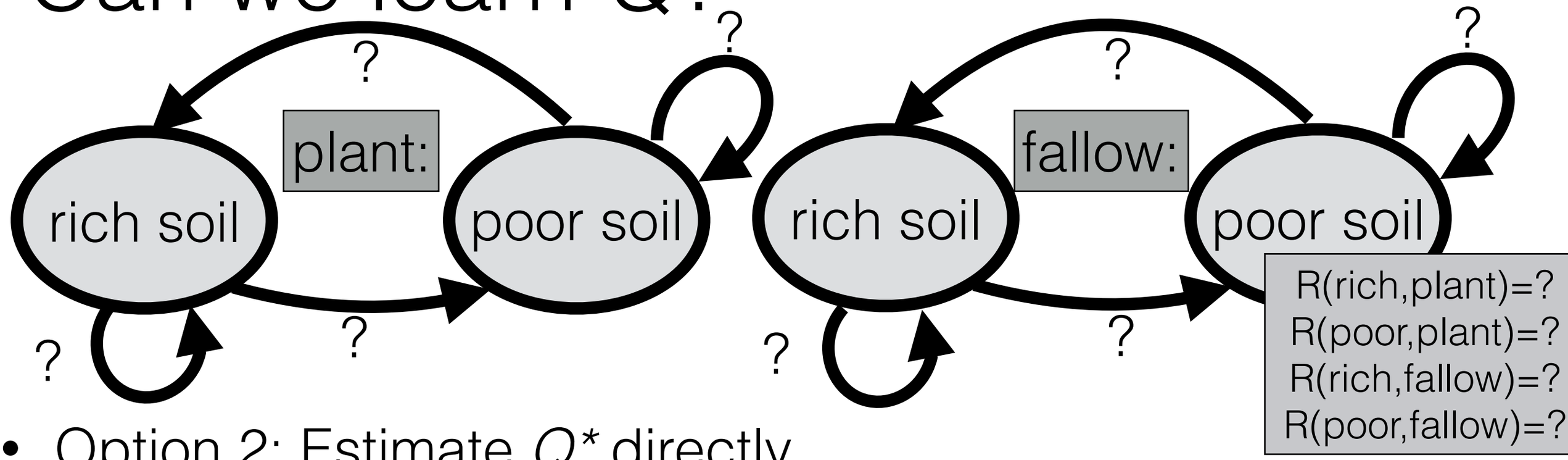
Can we learn *Q*?



- Option 2: Estimate *Q** directly
- Recall infinite horizon value iteration:

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^*(s',a')$$

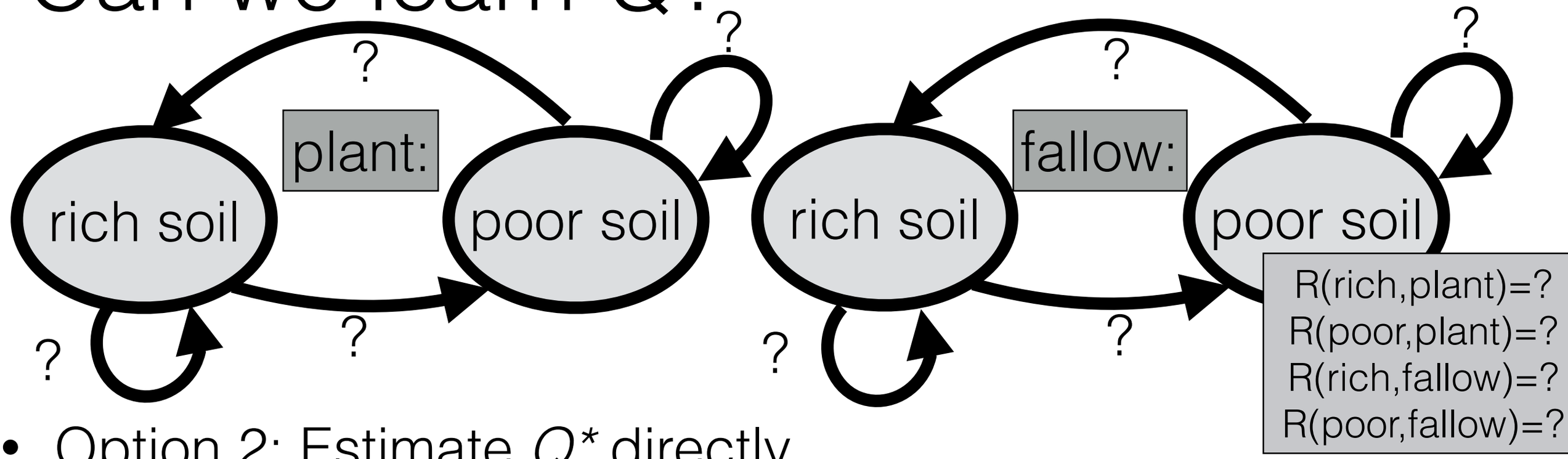
Can we learn *Q*?



- Option 2: Estimate *Q** directly
- Recall infinite horizon value iteration:

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q^*(s',a')$$

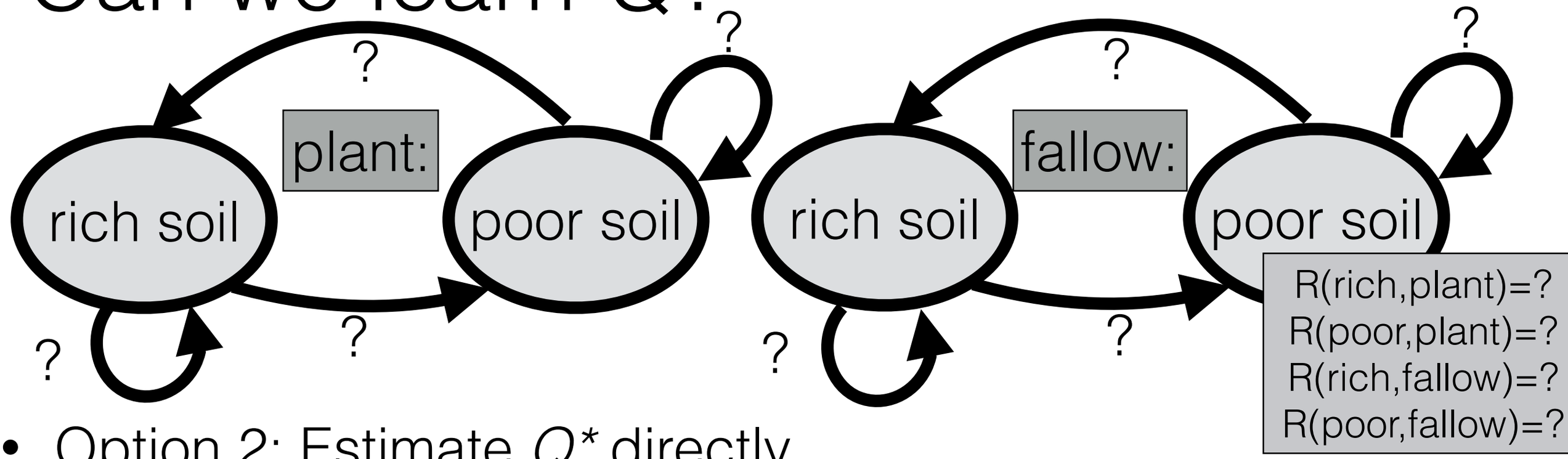
Can we learn *Q*?



- Option 2: Estimate Q^* directly
- Recall infinite horizon value iteration:

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

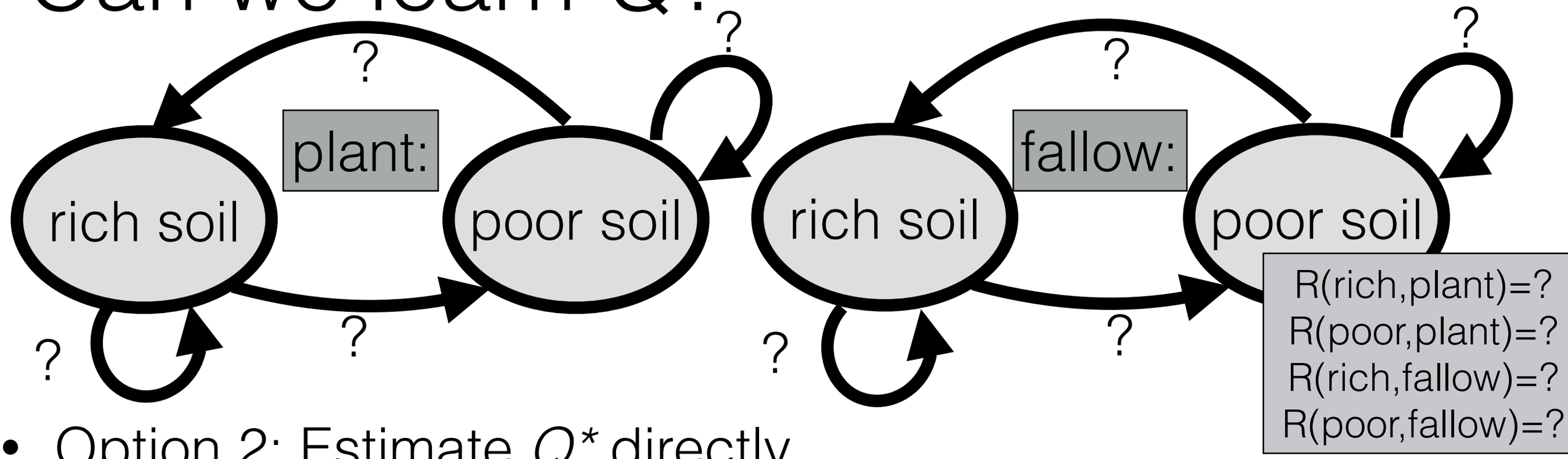
Can we learn *Q*?



- Option 2: Estimate Q^* directly
- Recall infinite horizon value iteration:

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

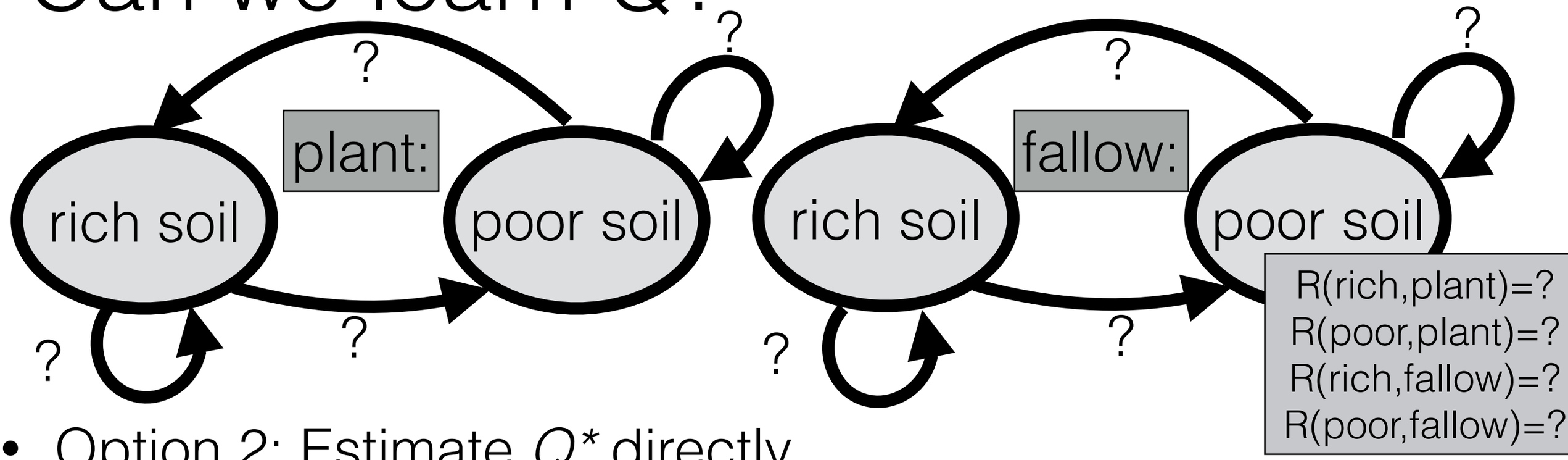
Can we learn *Q*?



- Option 2: Estimate *Q** directly
- Recall infinite horizon value iteration:

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

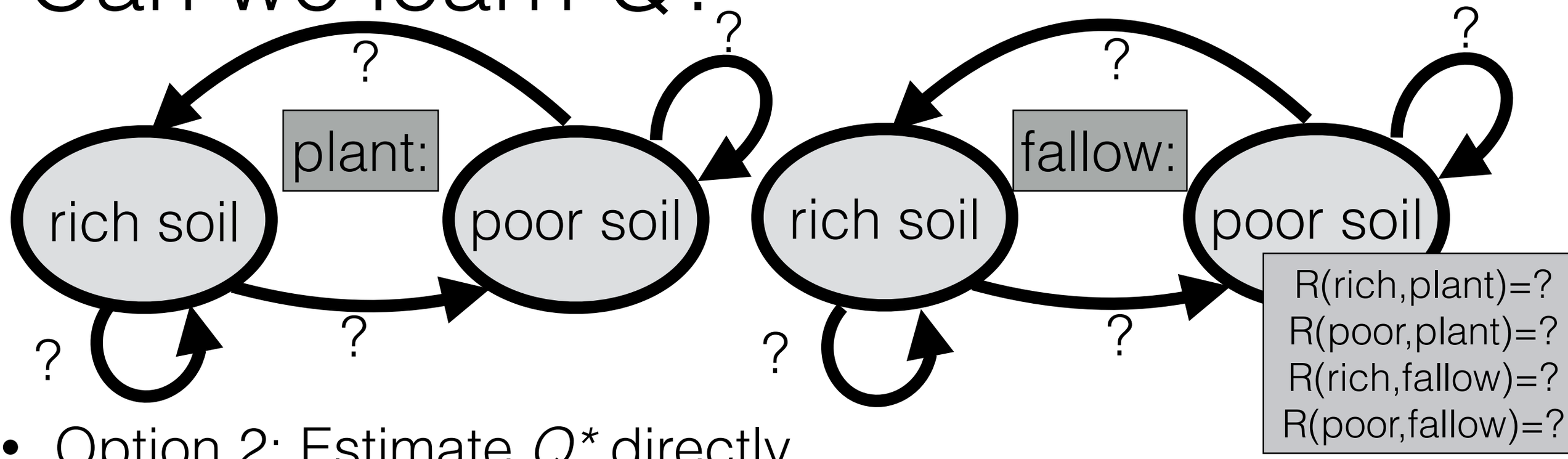
Can we learn *Q*?



- Option 2: Estimate *Q** directly
- Recall infinite horizon value iteration:

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

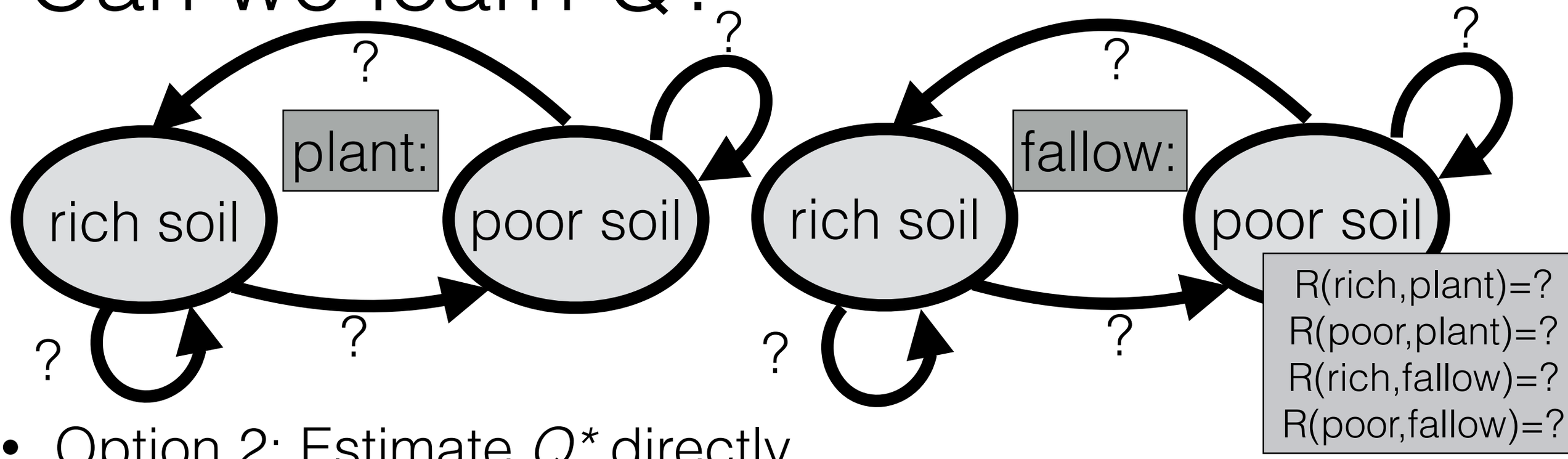
Can we learn *Q*?



- Option 2: Estimate Q^* directly
- Recall infinite horizon value iteration:

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

Can we learn *Q*?

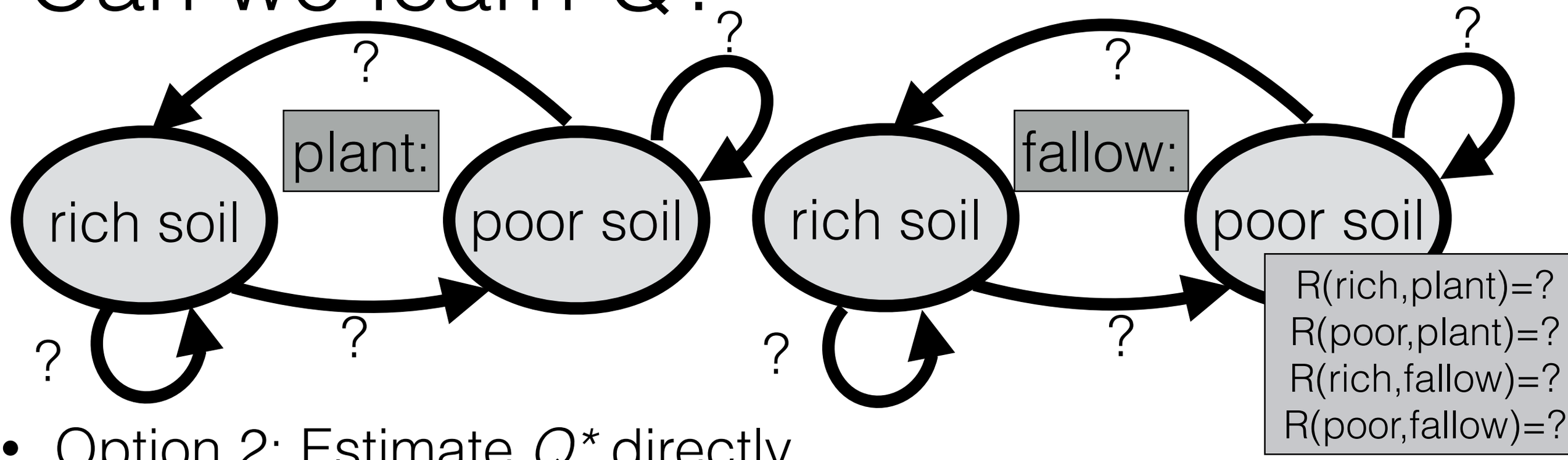


- Option 2: Estimate Q^* directly
- Recall infinite horizon value iteration:

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

- Expected value of a discrete random variable:

Can we learn *Q*?



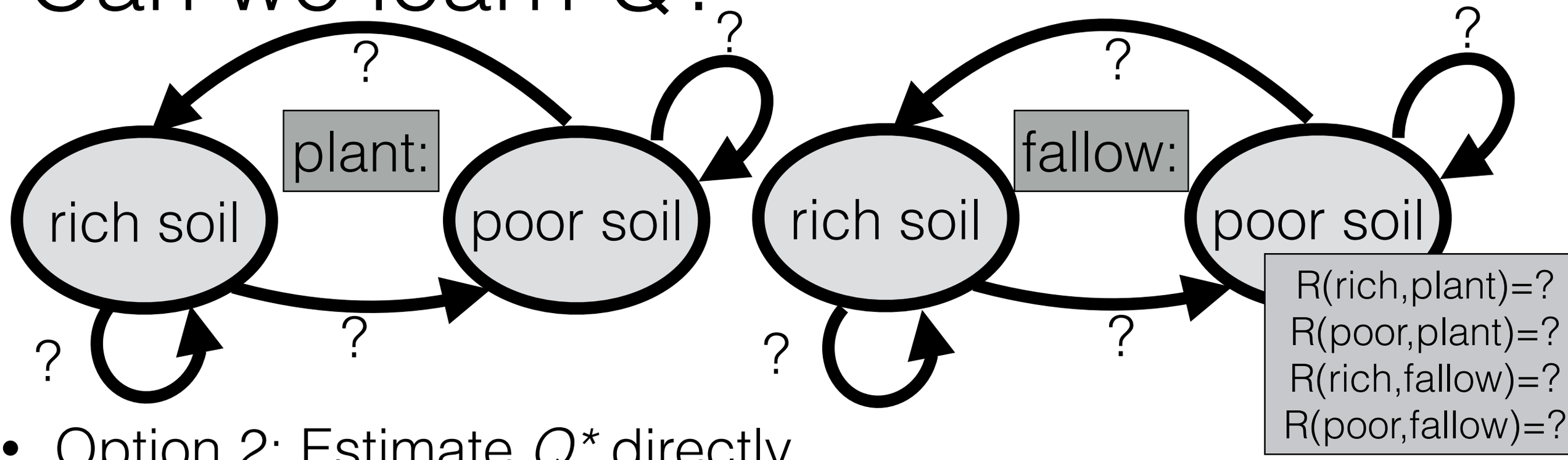
- Option 2: Estimate *Q** directly
- Recall infinite horizon value iteration:

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

- Expected value of a discrete random variable:

$$\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

Can we learn *Q*?



- Option 2: Estimate *Q** directly

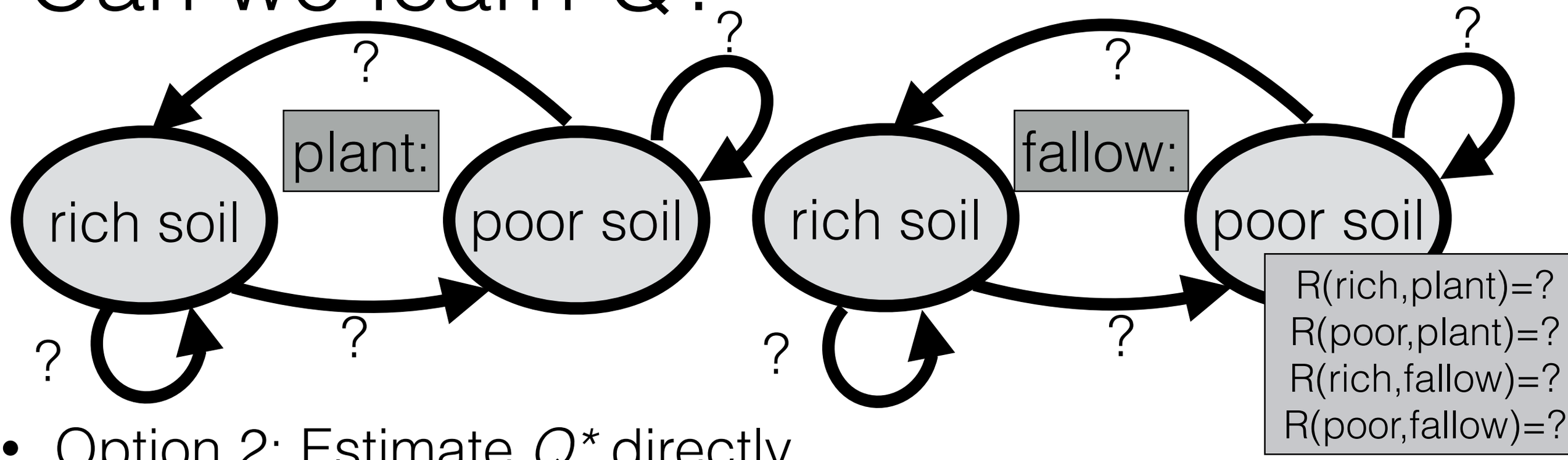
- Recall infinite horizon value iteration:

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

- Expected value of a discrete random variable:

$$\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

Can we learn *Q*?



- Option 2: Estimate *Q** directly

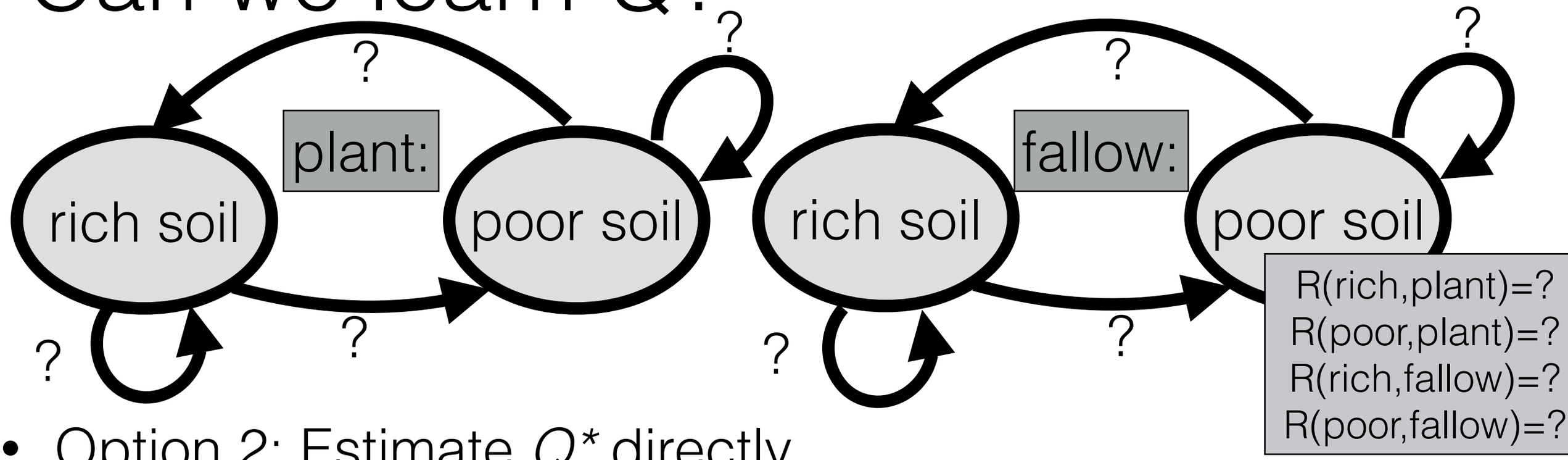
- Recall infinite horizon value iteration:

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

- Expected value of a discrete random variable:

$$\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

Can we learn *Q*?



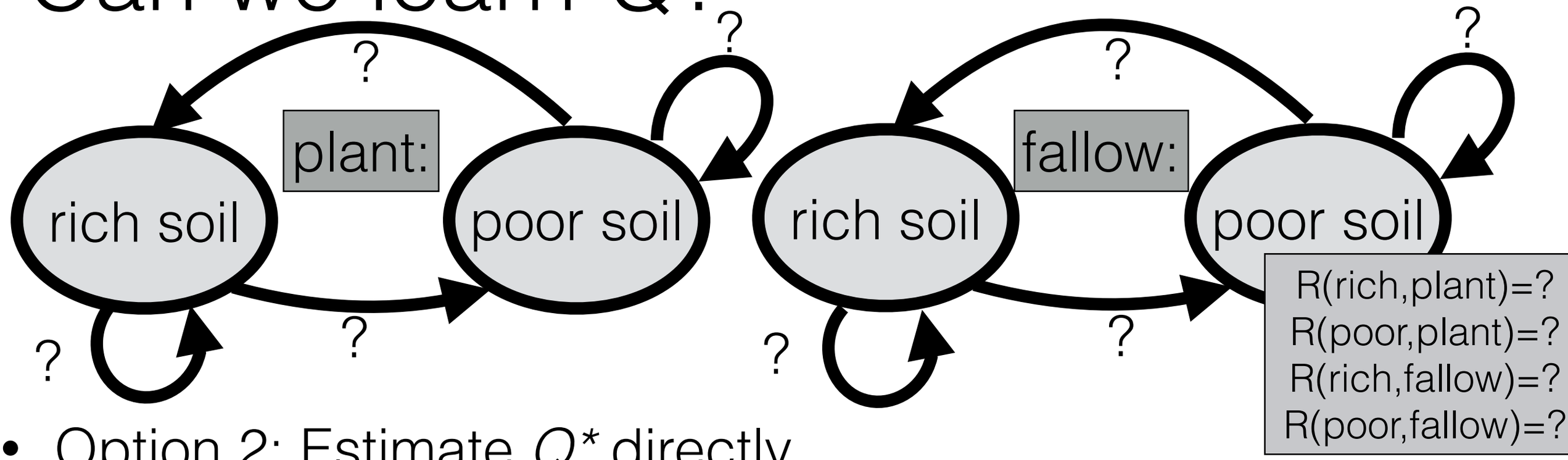
- Option 2: Estimate Q^* directly
- Recall infinite horizon value iteration:

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

- Expected value of a discrete random variable:

$$\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

Can we learn *Q*?



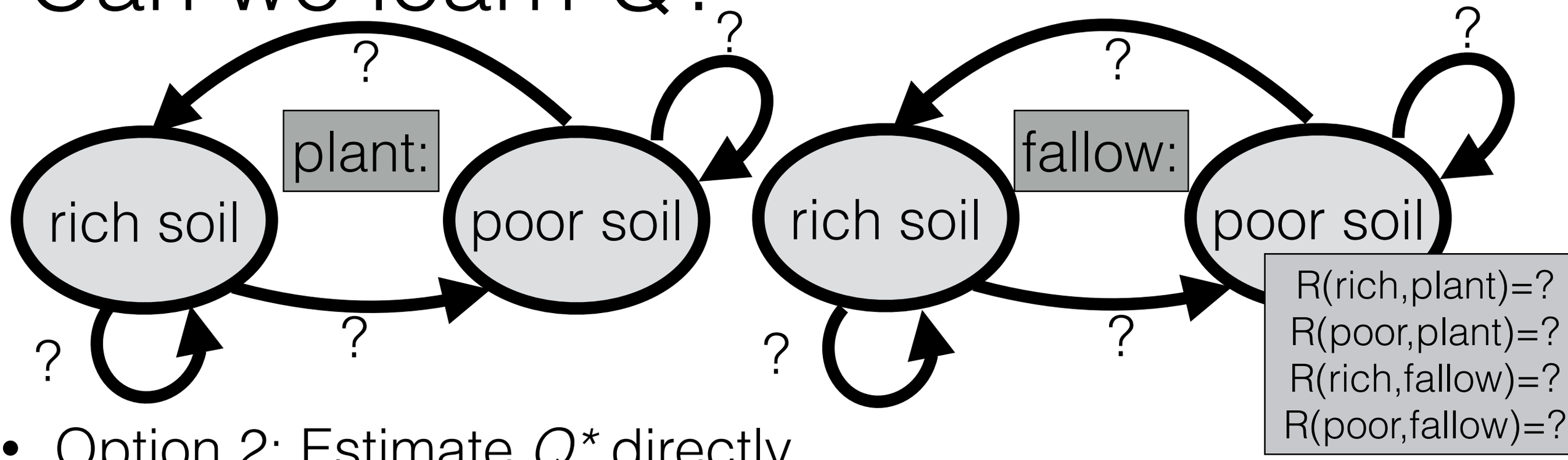
- Option 2: Estimate *Q** directly
- Recall infinite horizon value iteration:

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

- Expected value of a discrete random variable:

$$\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

Can we learn *Q*?



- Option 2: Estimate *Q** directly
- Recall infinite horizon value iteration:

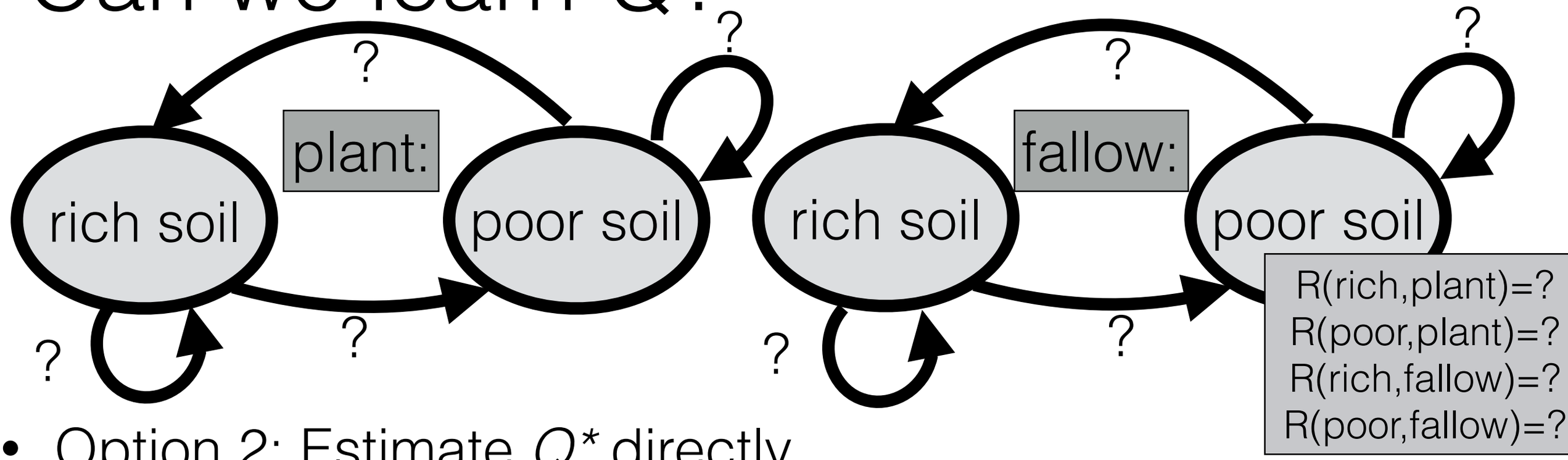
$$Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$$

- Expected value of a discrete random variable:

$$\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: Expected value of an insurance payout

Can we learn *Q*?



- Option 2: Estimate *Q** directly
- Recall infinite horizon value iteration:

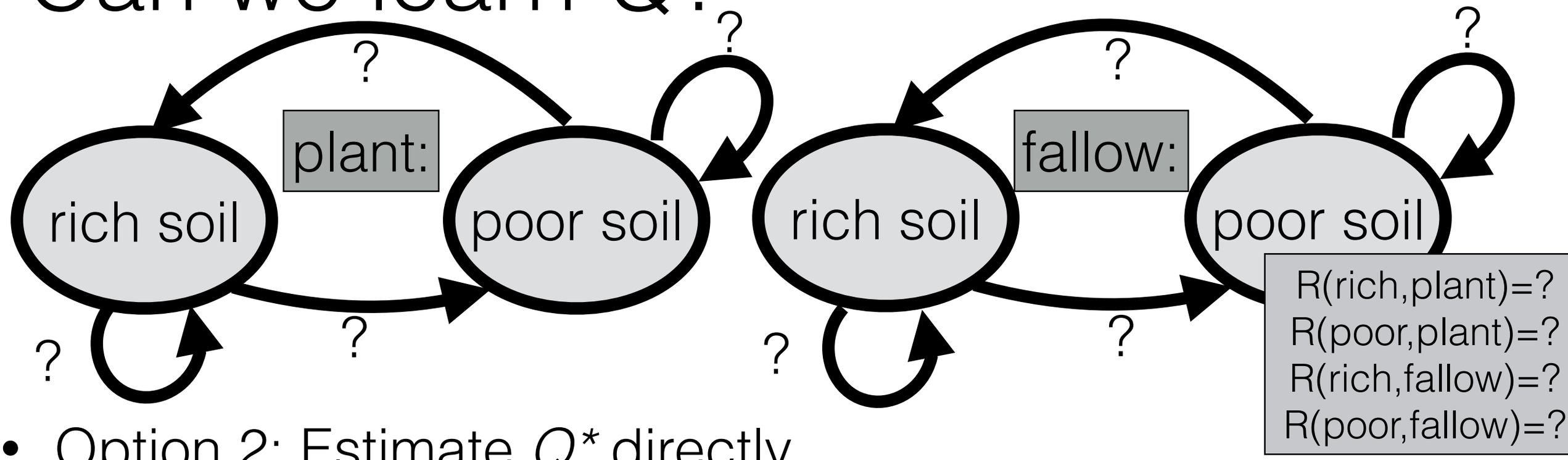
$$Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$$

- Expected value of a discrete random variable:

$$\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: Expected value of an insurance payout
    - $\mathcal{S}$ = {fire, no fire}

Can we learn *Q*?



- Option 2: Estimate Q^* directly

- Recall infinite horizon value iteration:

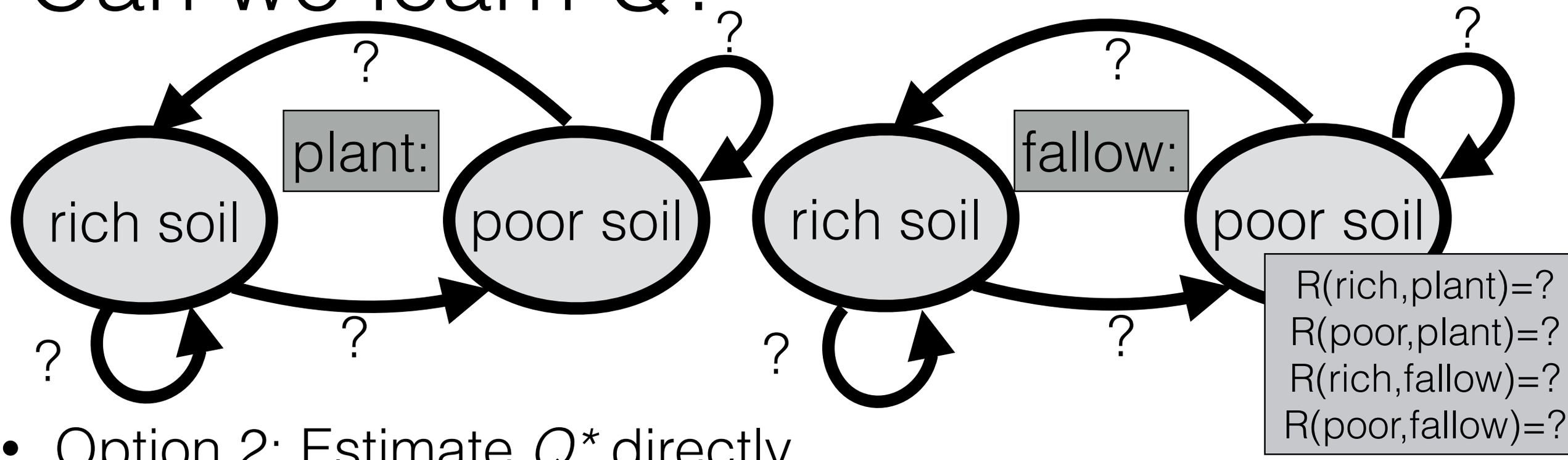
$$Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$$

- Expected value of a discrete random variable:

$$\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: Expected value of an insurance payout
    - $\mathcal{S}$ = {fire, no fire}
    - Expected value: $\frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$

Can we learn *Q*?



- Option 2: Estimate Q^* directly

- Recall infinite horizon value iteration:

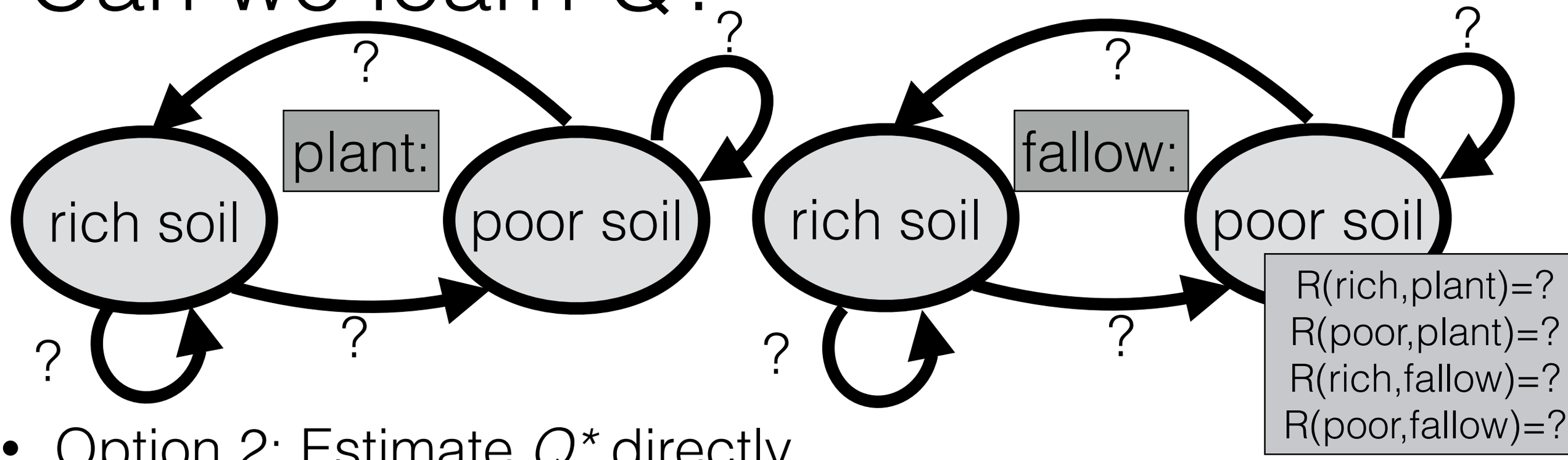
$$Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$$

- Expected value of a discrete random variable:

$$\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: Expected value of an insurance payout
    - $\mathcal{S}$ = {fire, no fire}
    - Expected value: $\frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$

Can we learn *Q*?



- Option 2: Estimate Q^* directly

- Recall infinite horizon value iteration:

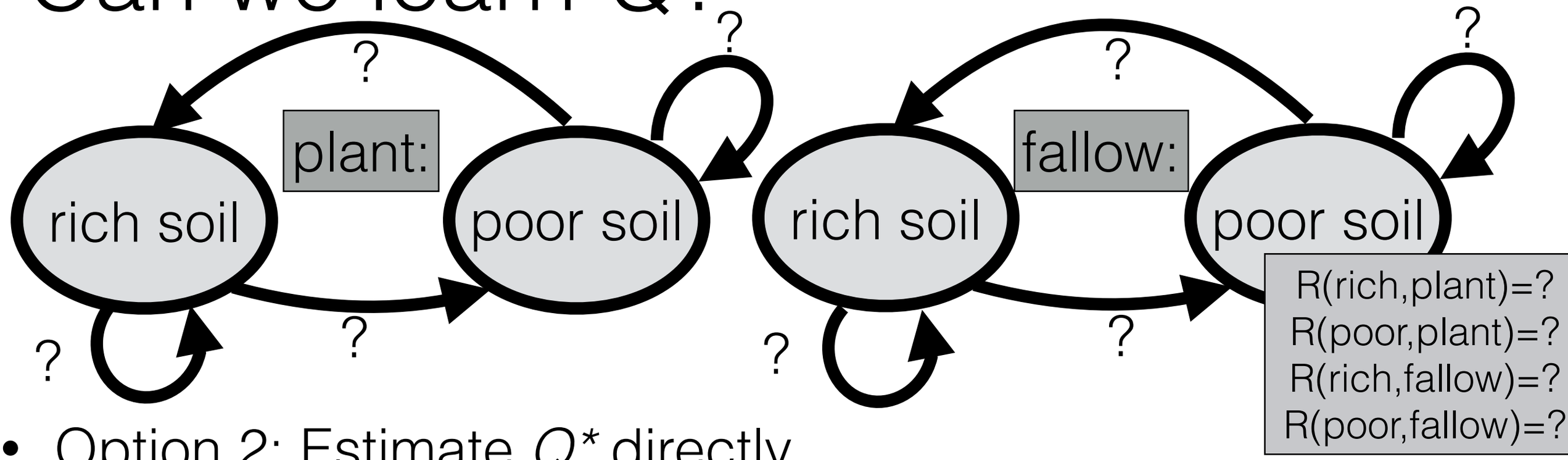
$$Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$$

- Expected value of a discrete random variable:

$$\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: Expected value of an insurance payout
    - $\mathcal{S}$ = {fire, no fire}
    - Expected value: $\frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$

Can we learn *Q*?



- Option 2: Estimate *Q** directly

- Recall infinite horizon value iteration:

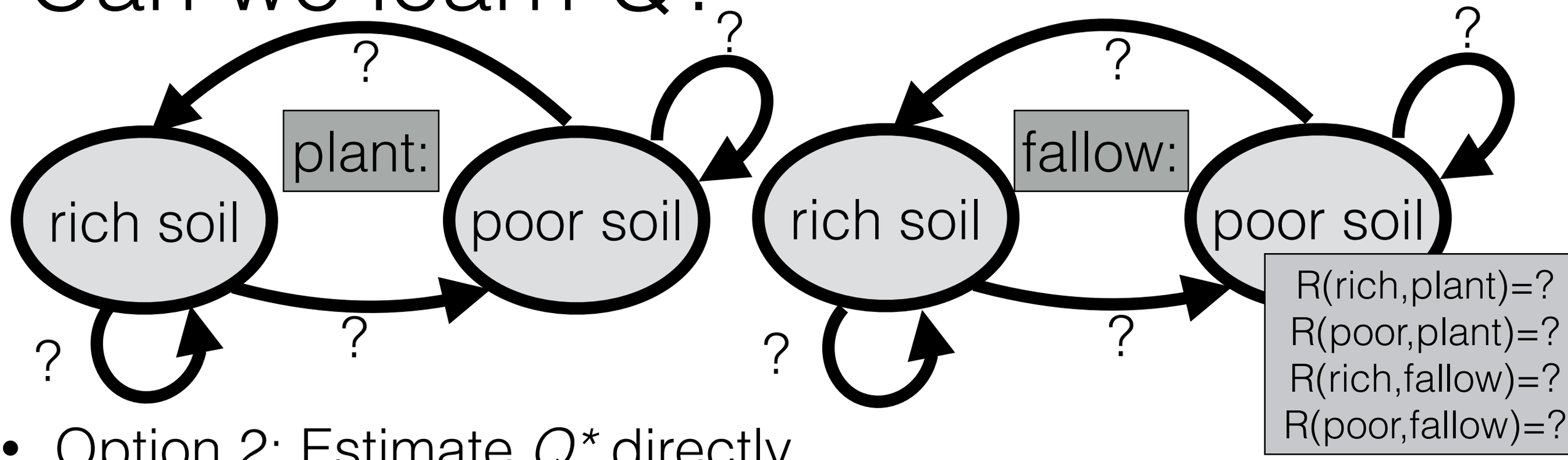
$$Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$$

- Expected value of a discrete random variable:

$$\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: Expected value of an insurance payout
    - $\mathcal{S}$ = {fire, no fire}
    - Expected value: $\frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$

Can we learn *Q*?



- Option 2: Estimate Q^* directly

- Recall infinite horizon value iteration:

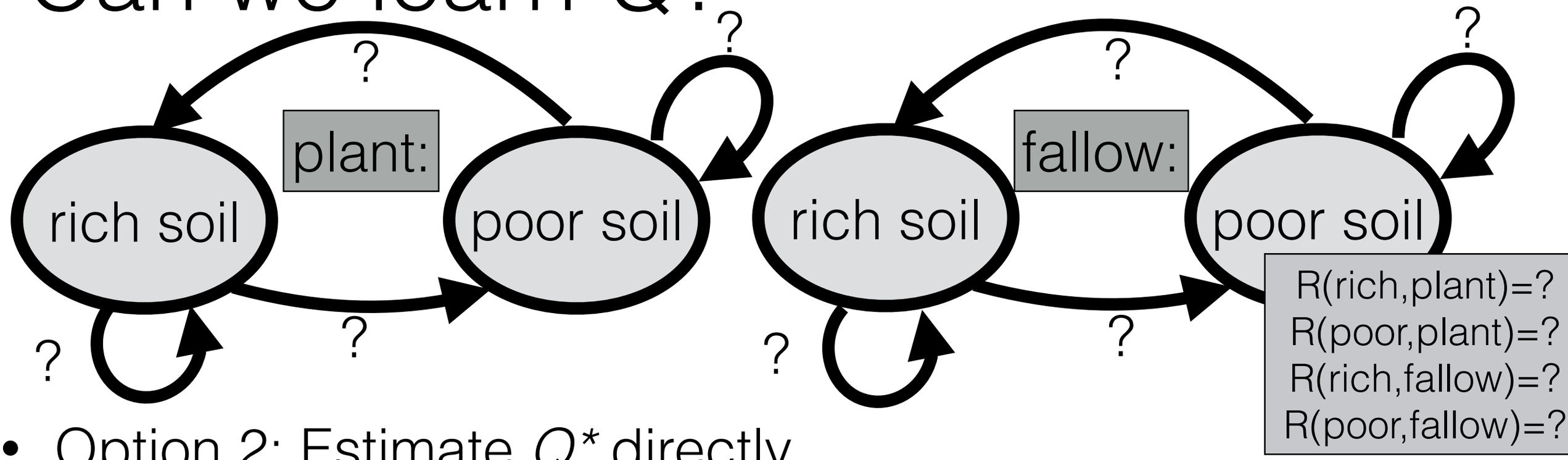
$$Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$$

- Expected value of a discrete random variable:

$$\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: Expected value of an insurance payout
    - $\mathcal{S}$ = {fire, no fire}
    - Expected value: $\frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$

Can we learn *Q*?



- Option 2: Estimate Q^* directly

- Recall infinite horizon value iteration:

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

- Expected value of a discrete random variable:

$$\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: Expected value of an insurance payout
    - $\mathcal{S}$ = {fire, no fire}
    - Expected value: $\frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$

Aside: Estimating an expected value

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

  - Can check: $\hat{E}^{(t)} = \sum_{i=1}^{t}(1 - \alpha)^{t-i}\alpha * \text{value}(s'^{(i)})$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

  - Can check: $\hat{E}^{(t)} = \sum_{i=1}^{t}(1 - \alpha)^{t-i}\alpha * \text{value}(s'^{(i)})$

$$(1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

  - Can check: $\hat{E}^{(t)} = \sum_{i=1}^{t}(1 - \alpha)^{t-i}\alpha * \text{value}(s'^{(i)})$

$$(1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$
$$= (1 - \alpha)\sum_{i=1}^{t-1}(1 - \alpha)^{t-1-i}\alpha\text{value}(s'^{(i)}) + \alpha\text{value}(s'^{(t)})$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

  - Can check: $\hat{E}^{(t)} = \sum_{i=1}^{t}(1 - \alpha)^{t-i}\alpha * \text{value}(s'^{(i)})$

$$\begin{aligned}
&(1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)}) \\
&= (1 - \alpha)\sum_{i=1}^{t-1}(1 - \alpha)^{t-1-i}\alpha\text{value}(s'^{(i)}) + \alpha\text{value}(s'^{(t)}) \\
&= \sum_{i=1}^{t}(1 - \alpha)^{t-i}\alpha\text{value}(s'^{(i)})
\end{aligned}$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

  - Can check: $\hat{E}^{(t)} = \sum_{i=1}^{t}(1 - \alpha)^{t-i}\alpha * \text{value}(s'^{(i)})$

$$\begin{aligned}
&(1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)}) \\
&= (1 - \alpha)\sum_{i=1}^{t-1}(1 - \alpha)^{t-1-i}\alpha\text{value}(s'^{(i)}) + \alpha\text{value}(s'^{(t)}) \\
&= \sum_{i=1}^{t}(1 - \alpha)^{t-i}\alpha\text{value}(s'^{(i)}) = \hat{E}^{(t)}
\end{aligned}$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

  - Can check: $\hat{E}^{(t)} = \sum_{i=1}^{t}(1 - \alpha)^{t-i}\alpha * \text{value}(s'^{(i)})$

$$\begin{aligned}
&(1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)}) \\
&= (1 - \alpha)\sum_{i=1}^{t-1}(1 - \alpha)^{t-1-i}\alpha\text{value}(s'^{(i)}) + \alpha\text{value}(s'^{(t)}) \\
&= \sum_{i=1}^{t}(1 - \alpha)^{t-i}\alpha\text{value}(s'^{(i)}) = \hat{E}^{(t)}
\end{aligned}$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

  - Can check: $\hat{E}^{(t)} = \sum_{i=1}^{t}(1 - \alpha)^{t-i}\alpha * \text{value}(s'^{(i)})$

$$\begin{aligned}
&(1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)}) \\
&= (1 - \alpha)\sum_{i=1}^{t-1}(1 - \alpha)^{t-1-i}\alpha\text{value}(s'^{(i)}) + \alpha\text{value}(s'^{(t)}) \\
&= \sum_{i=1}^{t}(1 - \alpha)^{t-i}\alpha\text{value}(s'^{(i)}) = \hat{E}^{(t)}
\end{aligned}$$

Aside: Estimating an expected value

- Expected value of a discrete random variable:

$$E_{\text{exact}} = \sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$$

  - Example: expected value of an insurance payout:

$$E_{\text{exact}} = \frac{1}{10{,}000}\$100{,}000 + \left(1 - \frac{1}{10{,}000}\right)\$0 = \$10$$

- An estimate of the expected value:

$$\tilde{E}^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \text{value}(s'^{(i)})$$

- Can write as a running average with $\alpha^{(t)} = 1/t$

$$\tilde{E}^{(0)} = 0; \tilde{E}^{(t)} = (1 - \alpha^{(t)})\tilde{E}^{(t-1)} + \alpha^{(t)} * \text{value}(s'^{(t)})$$

- Different running average, with constant $\alpha^{(t)} = \alpha$

$$\hat{E}^{(0)} = 0; \hat{E}^{(t)} = (1 - \alpha)\hat{E}^{(t-1)} + \alpha * \text{value}(s'^{(t)})$$

  - Can check: $\hat{E}^{(t)} = \sum_{i=1}^{t}(1-\alpha)^{t-i}\alpha * \text{value}(s'^{(i)})$

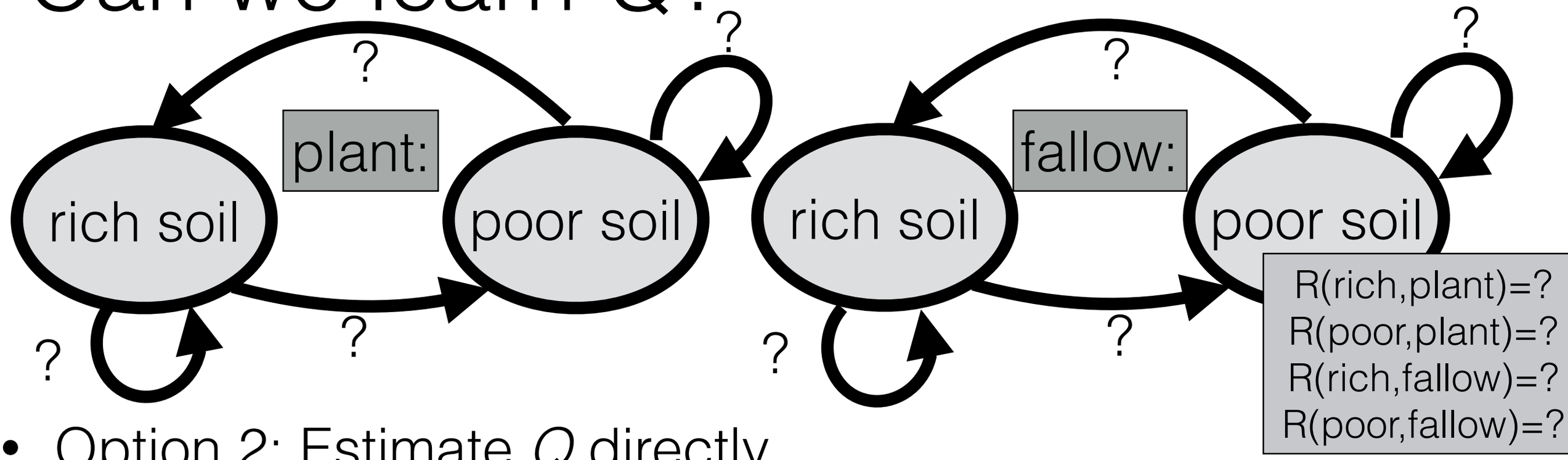
$$(1-\alpha)\hat{E}^{($$

$$= (1-\alpha$$ $$+ \alpha\text{value}(s'^{(t)})$$

$$= \sum_{i=1}^{t}(1-\alpha)^{t-i}\alpha\text{value}(s'^{(i)}) = \hat{E}^{(t)}$$

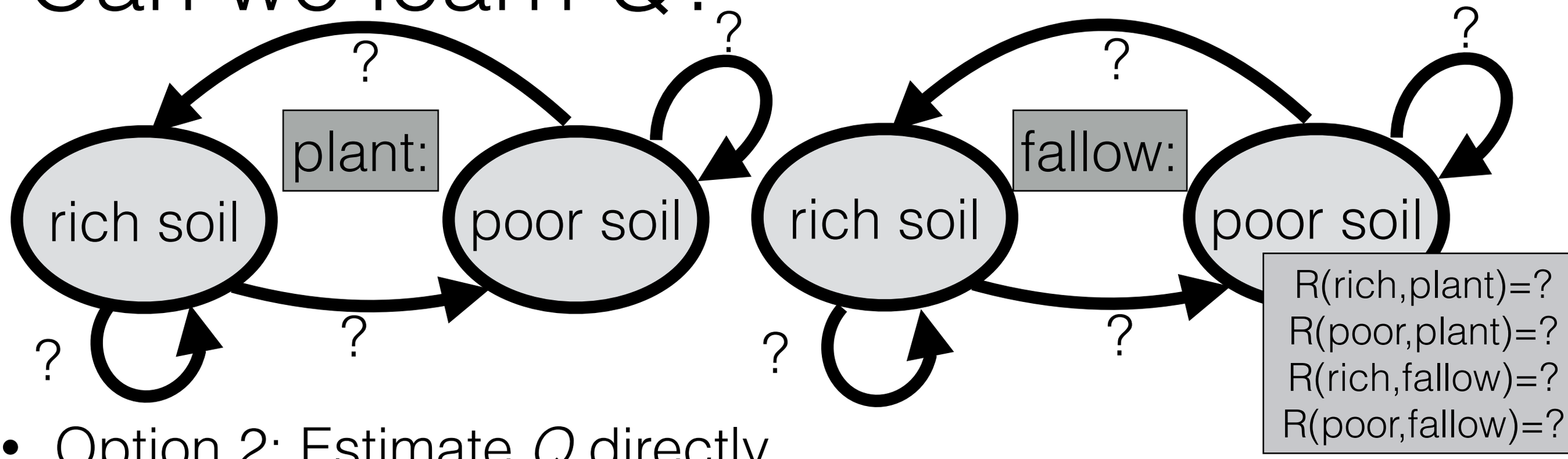
gives bigger weights to more recently observed values

Can we learn *Q*?



- Option 2: Estimate *Q* directly

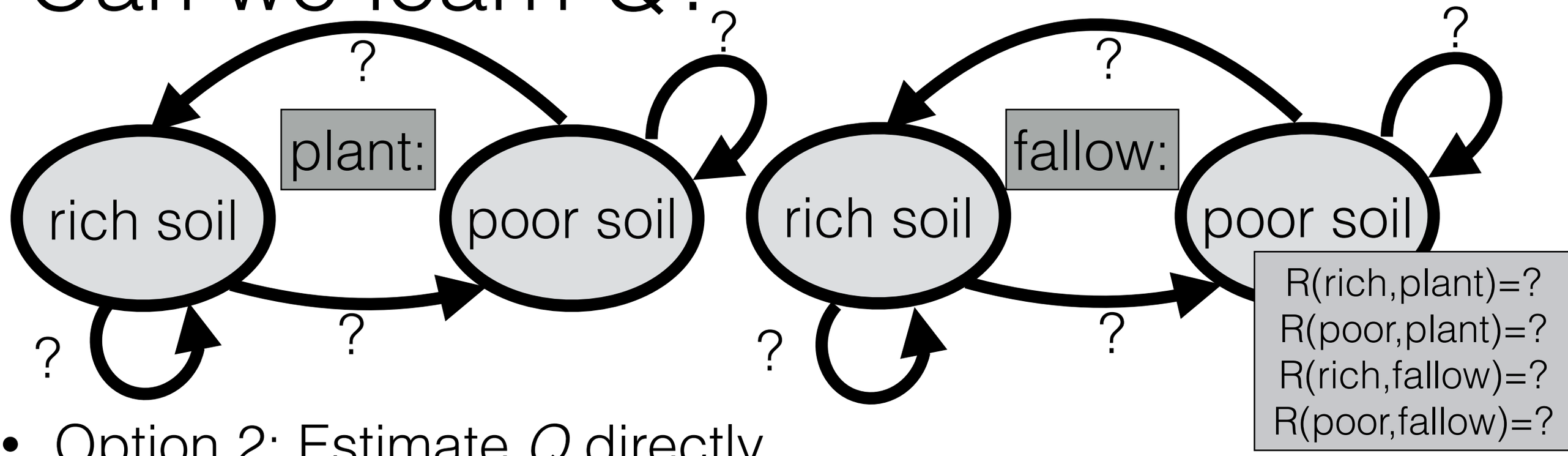
Can we learn *Q*?



- Option 2: Estimate *Q* directly
- Recall infinite horizon value iteration:

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

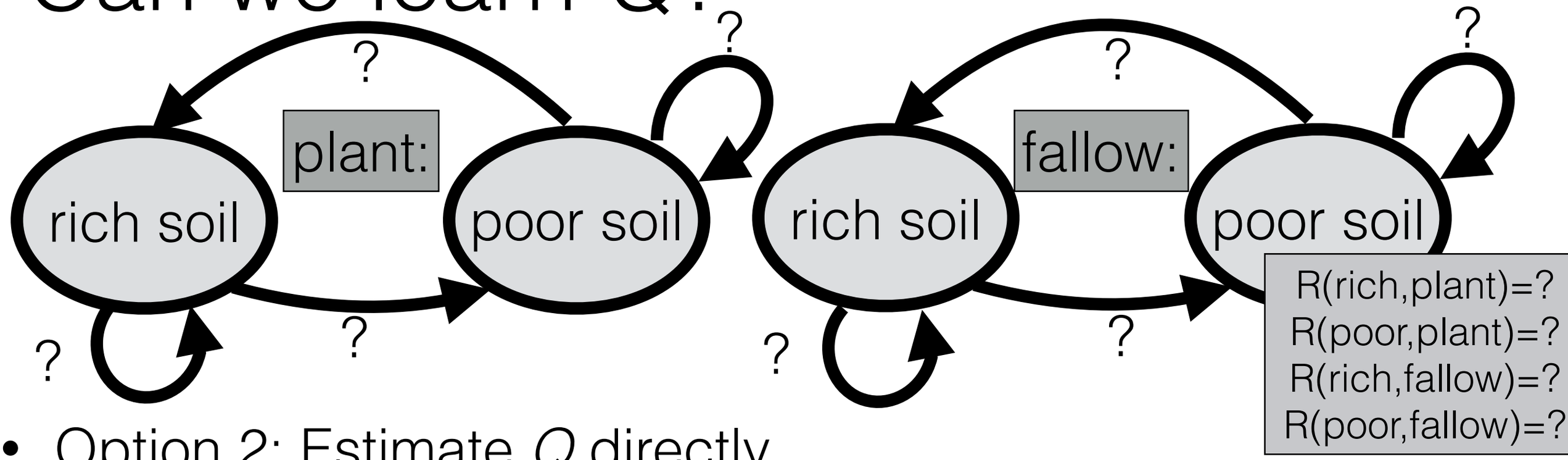
Can we learn *Q*?



- Option 2: Estimate *Q* directly
- Recall infinite horizon value iteration:

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

Can we learn *Q*?

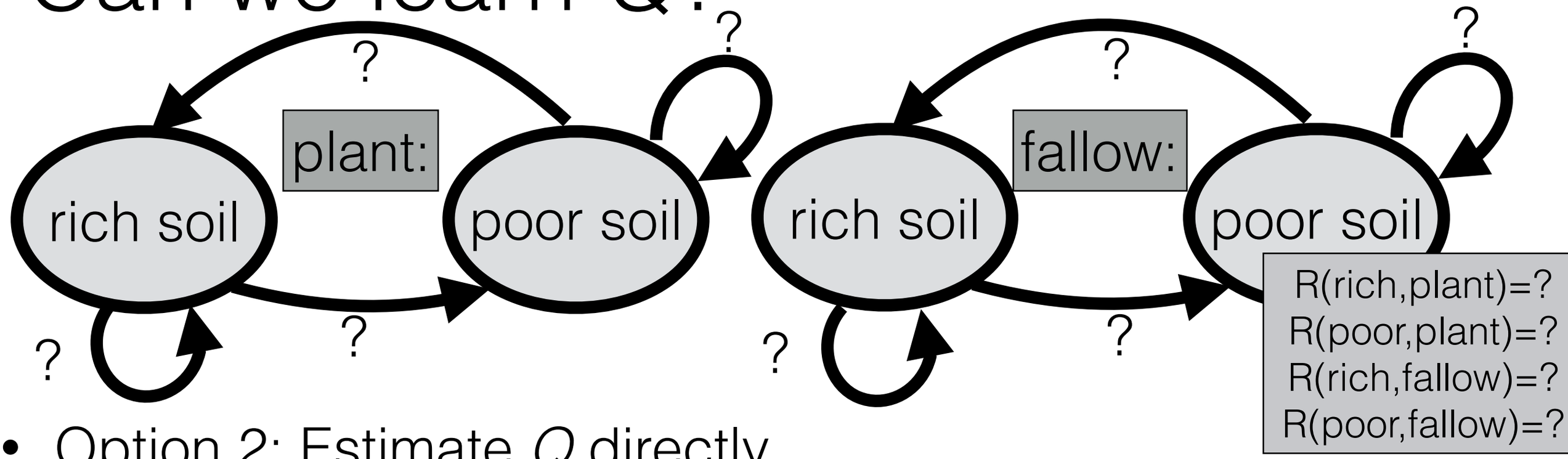


- Option 2: Estimate *Q* directly
- Recall infinite horizon value iteration:

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

  - An expected value: $\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$

Can we learn *Q*?



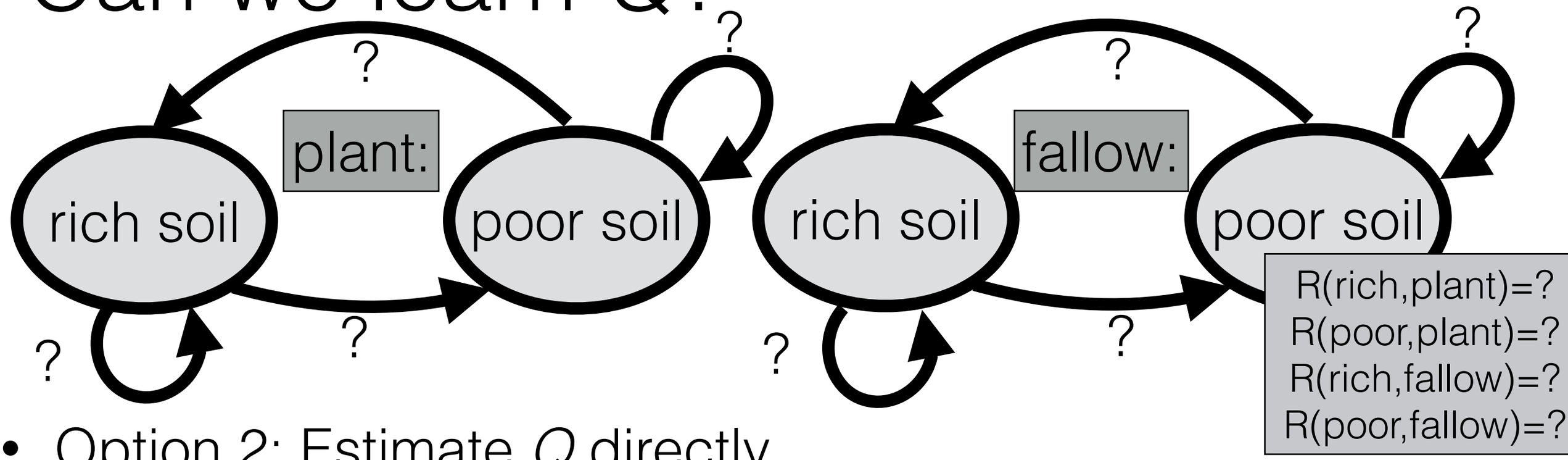
- Option 2: Estimate *Q* directly
- Recall infinite horizon value iteration:

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

  - An expected value: $\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$
- Update proposal: observe (*s*,*a*,*s*') & update estimate *Q*(*s*,*a*)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha \left(R(s,a) + \gamma \max_{a'} Q_{\text{old}}(s',a')\right)$$

Can we learn *Q*?



- Option 2: Estimate *Q* directly
- Recall infinite horizon value iteration:

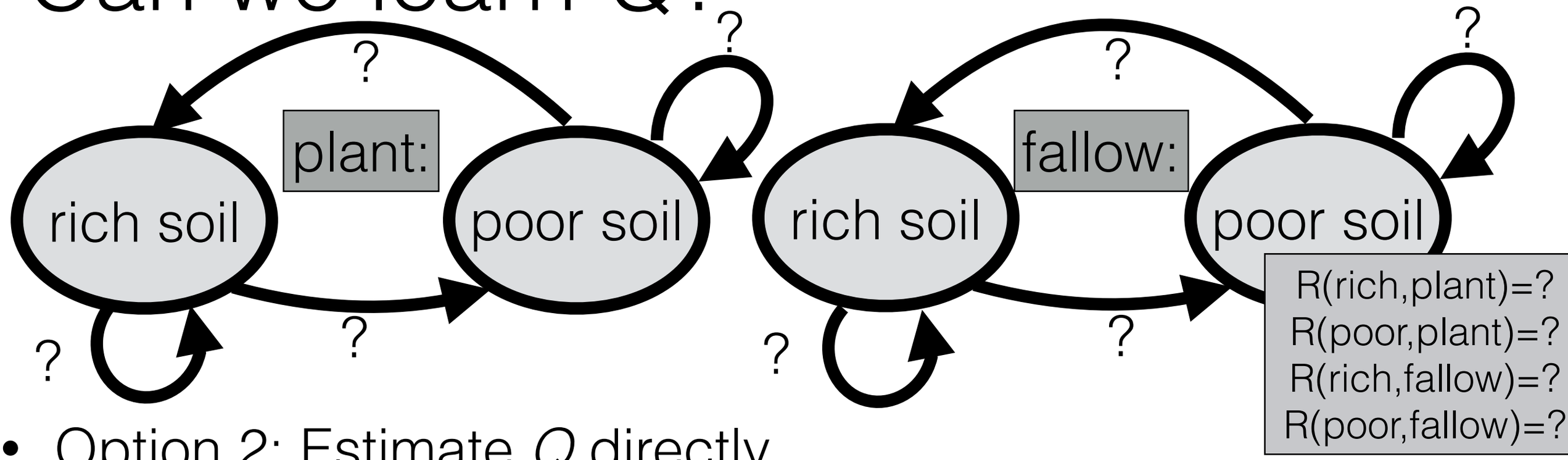
$$Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$$
$$= \sum_{s'} T(s,a,s') \left[R(s,a) + \gamma \max_{a'} Q_{\text{old}}(s',a')\right]$$

  - An expected value: $\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$

- Update proposal: observe (*s*,*a*,*s*') & update estimate *Q*(*s*,*a*)

$$Q(s,a) = (1-\alpha) Q(s,a) + \alpha \left(R(s,a) + \gamma \max_{a'} Q_{\text{old}}(s',a')\right)$$

Can we learn *Q*?



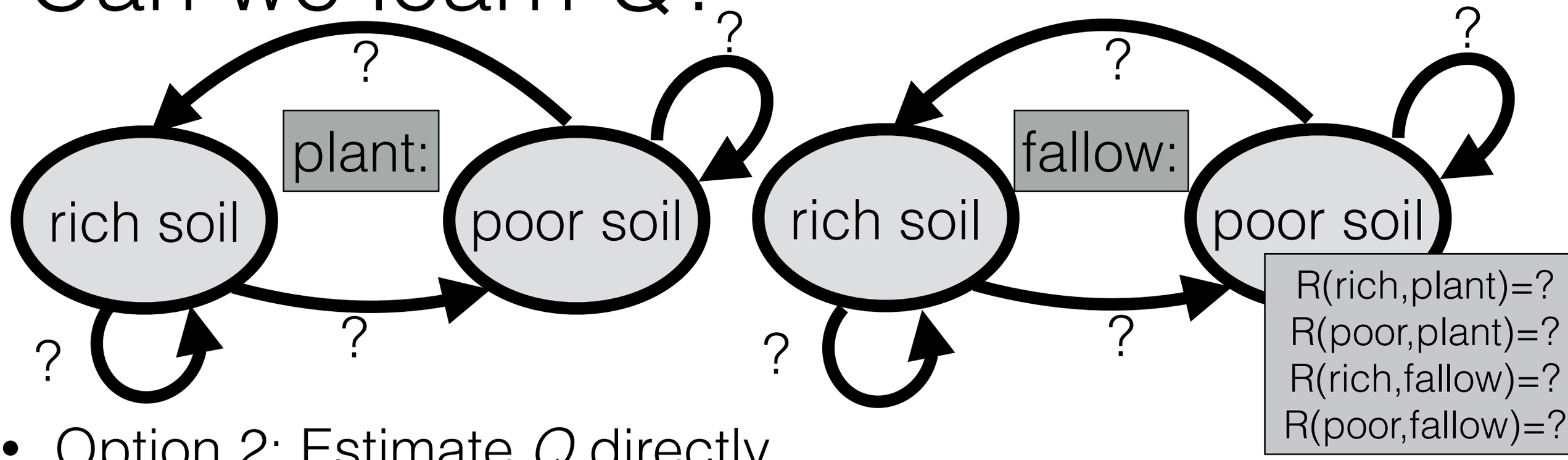
- Option 2: Estimate *Q* directly
- Recall infinite horizon value iteration:

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

  - An expected value: $\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$
- Update proposal: observe (*s*,*a*,*s*') & update estimate *Q*(*s*,*a*)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha \left(R(s,a) + \gamma \max_{a'} Q_{\text{old}}(s',a')\right)$$

Can we learn *Q*?



- Option 2: Estimate *Q* directly
- Recall infinite horizon value iteration:

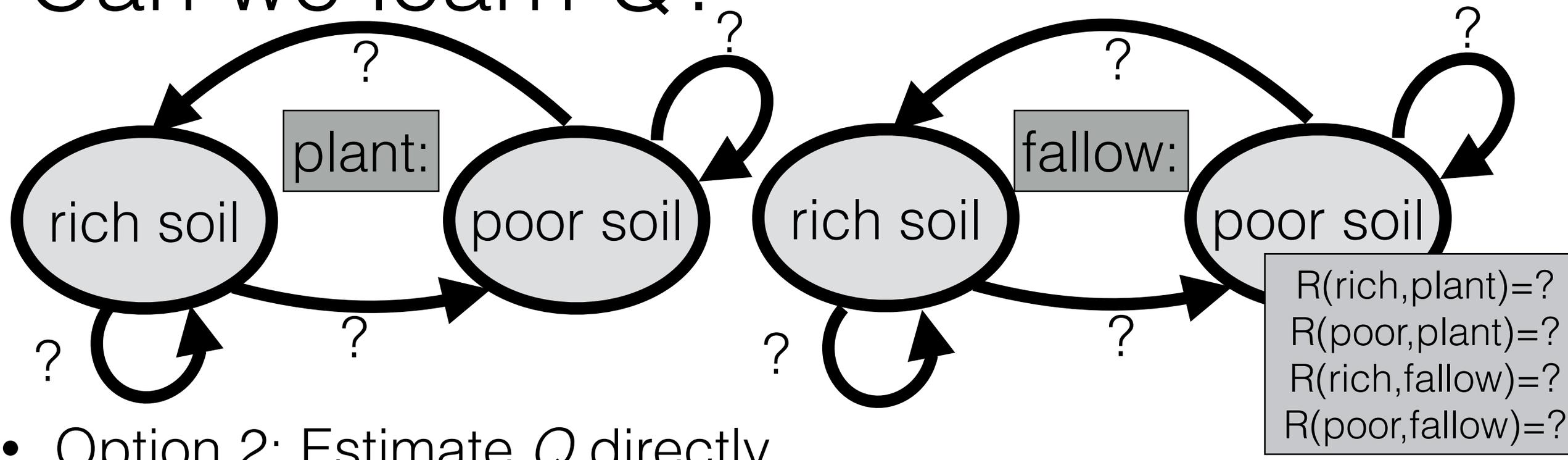
$$Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$$
$$= \sum_{s'} T(s,a,s') \left[R(s,a) + \gamma \max_{a'} Q_{\text{old}}(s',a')\right]$$

  - An expected value: $\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$

- Update proposal: observe (*s*,*a*,*s*') & update estimate *Q*(*s*,*a*)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha \left(R(s,a) + \gamma \max_{a'} Q_{\text{old}}(s',a')\right)$$

Can we learn *Q*?



- Option 2: Estimate *Q* directly
- Recall infinite horizon value iteration:

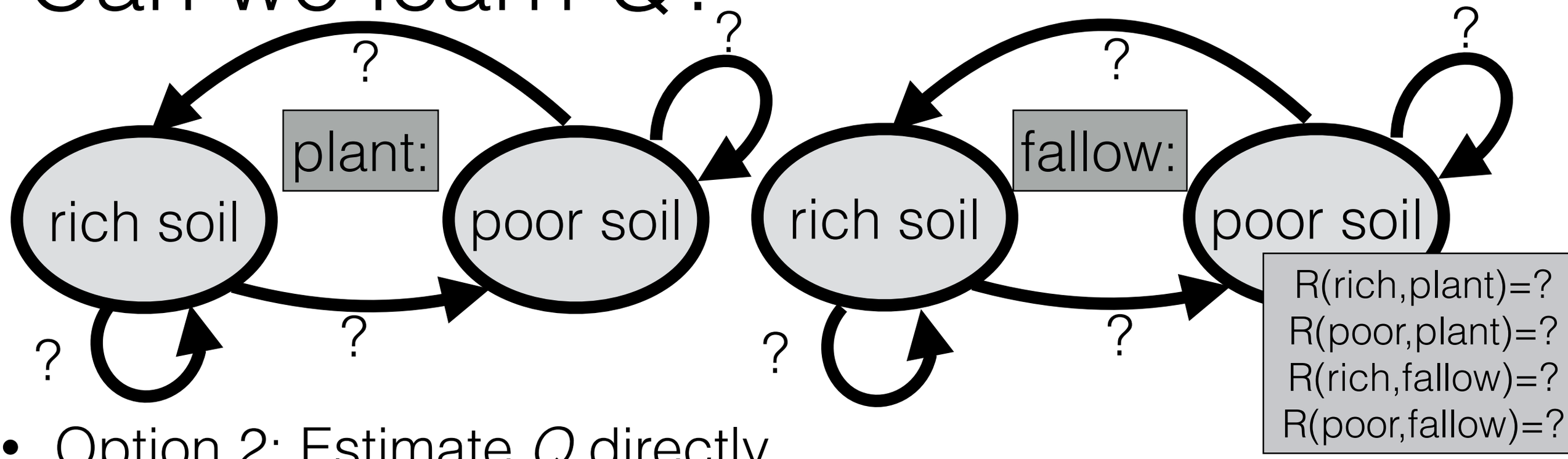
$$Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$$
$$= \sum_{s'} T(s,a,s') \left[R(s,a) + \gamma \max_{a'} Q_{\text{old}}(s',a')\right]$$

  - An expected value: $\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$

- Update proposal: observe (*s*,*a*,*s*') & update estimate *Q*(*s*,*a*)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha \left(R(s,a) + \gamma \max_{a'} Q_{\text{old}}(s',a')\right)$$

Can we learn *Q*?



- Option 2: Estimate *Q* directly
- Recall infinite horizon value iteration:

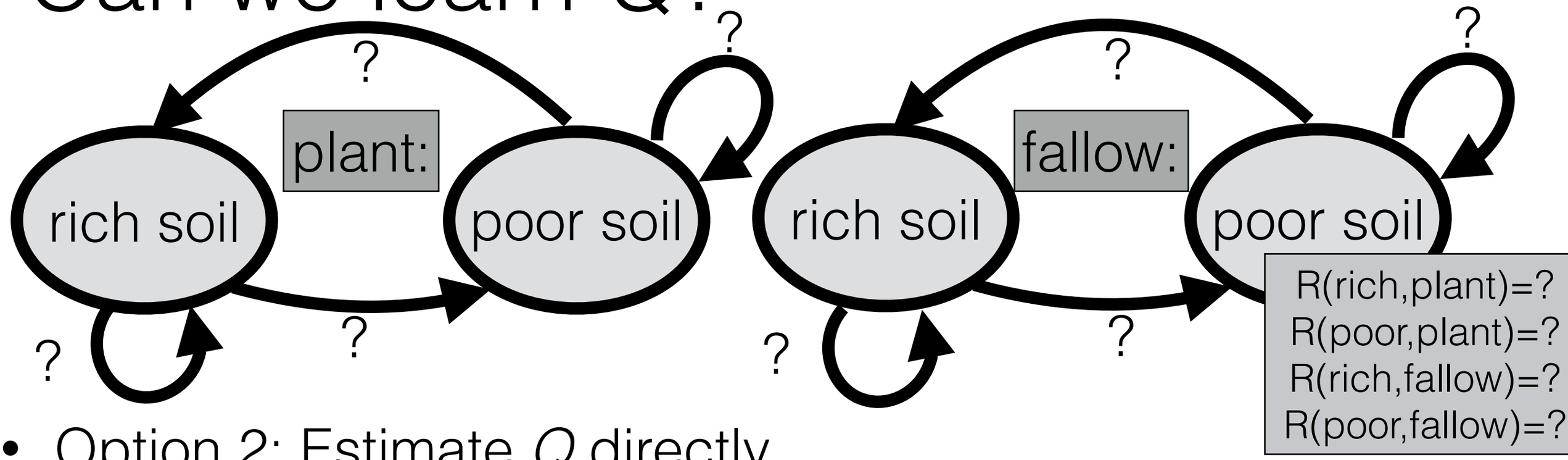
$$Q_{\text{new}}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{a'} Q_{\text{old}}(s',a')$$
$$= \sum_{s'} T(s,a,s') \left[R(s,a) + \gamma \max_{a'} Q_{\text{old}}(s',a')\right]$$

  - An expected value: $\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$

- Update proposal: observe (*s*,*a*,*s*') & update estimate *Q*(*s*,*a*)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha \left(R(s,a) + \gamma \max_{a'} Q_{\text{old}}(s',a')\right)$$

Can we learn *Q*?



- Option 2: Estimate *Q* directly
- Recall infinite horizon value iteration:

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

  - An expected value: $\sum_{s' \in \mathcal{S}} P(S' = s') * \text{value}(s')$

- Update proposal: observe (*s*,*a*,*s*') & update estimate *Q*(*s*,*a*)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha \left(R(s,a) + \gamma \max_{a'} Q_{\text{old}}(s',a')\right)$$

- Update proposal: observe (*s*,*a*,*s*') & update estimate *Q*(*s*,*a*)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha \left(R(s,a) + \gamma \max_{a'} Q(s',a')\right)$$

Yes, we can learn Q

Yes, we can learn Q

`Q-learning`

Yes, we can learn Q

`Q-learning(`$\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$`)`

Yes, we can learn Q

`Q-learning(`$\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$`)`

Yes, we can learn Q

`Q-learning(`$\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$`)`

Yes, we can learn Q

`Q-learning(`$\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$`)`

Yes, we can learn Q

`Q-learning(`$\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$`)`

Yes, we can learn Q

`Q-learning(`$\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$`)`

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s, a) = 0$

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s, a) = 0$

  Initialize $s = s_0$

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s, a) = 0$

  Initialize $s = s_0$

  while True

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s, a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

E.g. ϵ-**greedy**

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

E.g. ϵ-**greedy**

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

E.g. ϵ-**greedy**

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a')\right)$$

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

    $Q(s,a) = (1-\alpha)Q(s,a) + \alpha\,(r + \gamma \max_{a'} Q(s',a'))$

    s = s′

E.g. ϵ-**greedy**

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a')\right)$$

    s = s′

E.g. ϵ-**greedy**

- Another perspective on the update:

$$Q(s,a) = Q(s,a) - \alpha\left[Q(s,a) - \left(r + \gamma \max_{a'} Q(s',a')\right)\right]$$

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a')\right)$$

    s = s′

E.g. ϵ-**greedy**

- Another perspective on the update:

$$Q(s,a) = Q(s,a) - \alpha\left[Q(s,a) - \left(r + \gamma \max_{a'} Q(s',a')\right)\right]$$

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

    $Q(s,a) = (1-\alpha)Q(s,a) + \alpha\,(r + \gamma \max_{a'} Q(s',a'))$

    s = s′

E.g. ϵ-**greedy**

- Another perspective on the update:

$$Q(s,a) = Q(s,a) - \alpha\,[Q(s,a) - (r + \gamma \max_{a'} Q(s',a'))]$$

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\,(r + \gamma \max_{a'} Q(s',a'))$$

s = s′

E.g. ϵ-greedy

- Another perspective on the update:

$$Q(s,a) = Q(s,a) - \alpha\,[Q(s,a) - (r + \gamma \max_{a'} Q(s',a'))]$$

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\,(r + \gamma \max_{a'} Q(s',a'))$$

s = s′

E.g. ϵ-greedy

- Another perspective on the update:

$$Q(s,a) = Q(s,a) - \alpha\,[Q(s,a) - (r + \gamma \max_{a'} Q(s',a'))]$$

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a')\right)$$

s = s′

E.g. ϵ-**greedy**

- Another perspective on the update:

$$Q(s,a) = Q(s,a) - \alpha\left[Q(s,a) - \left(r + \gamma \max_{a'} Q(s',a')\right)\right]$$

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\,(r + \gamma \max_{a'} Q(s',a'))$$

s = s′

E.g. ϵ-greedy

- Another perspective on the update:

$$Q(s,a) = Q(s,a) - \alpha\,[Q(s,a) - (r + \gamma \max_{a'} Q(s',a'))]$$

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a')\right)$$

s = s′

E.g. ϵ-**greedy**

- Another perspective on the update:

$$Q(s,a) = Q(s,a) - \alpha\left[Q(s,a) - \left(r + \gamma \max_{a'} Q(s',a')\right)\right]$$

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

    $Q(s,a) = (1-\alpha)Q(s,a) + \alpha\,(r + \gamma \max_{a'} Q(s',a'))$

    s = s′

E.g. ϵ-**greedy**

- Another perspective on the update:
$Q(s,a) = Q(s,a) - \alpha\,[Q(s,a) - (r + \gamma \max_{a'} Q(s',a'))]$

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a')\right)$$

    s = s′

E.g. ϵ-**greedy**

- Another perspective on the update:

$$Q(s,a) = Q(s,a) - \alpha\left[Q(s,a) - \left(r + \gamma \max_{a'} Q(s',a')\right)\right]$$

"temporal difference learning"

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a')\right)$$

    s = s′

E.g. ϵ-**greedy**

- Another perspective on the update:

$$Q(s,a) = Q(s,a) - \alpha\left[Q(s,a) - \left(r + \gamma \max_{a'} Q(s',a')\right)\right]$$

"temporal difference learning"

Note: Algorithm never learns/stores *T* or *R;* just *Q*

Yes, we can learn Q

Q-learning($\mathcal{S}, \mathcal{A}, s_0, \gamma, \alpha$)

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

    Initialize $Q(s,a) = 0$

  Initialize $s = s_0$

  while True

    a = select_action(s,Q)

    r,s′ = execute(a)

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\,(r + \gamma \max_{a'} Q(s',a'))$$

    s = s′

E.g. ϵ-**greedy**

Note: This algorithm doesn't quite take actions according the policy we're learning

- Another perspective on the update:

$$Q(s,a) = Q(s,a) - \alpha\,[Q(s,a) - (r + \gamma$$

"temporal difference learning"

Note: Algorithm never learns/stores *T* or *R;* just *Q*

Some terminology

- *Reinforcement learning* (RL): learning (to maximize rewards) by interacting with the world
  - Contrast with *supervised learning*
- *Model-based RL*: uses explicit conception of next state and reward given current state and action
  - "Model" used many different ways in machine learning
  - Contrast with *Model-free RL*
- *Q-learning*
  - Contrast with the *Q** function (expected reward of starting at *s*, making action *a*, and then making the "best" action ever after)
  - Contrast with (any horizon) *value iteration*