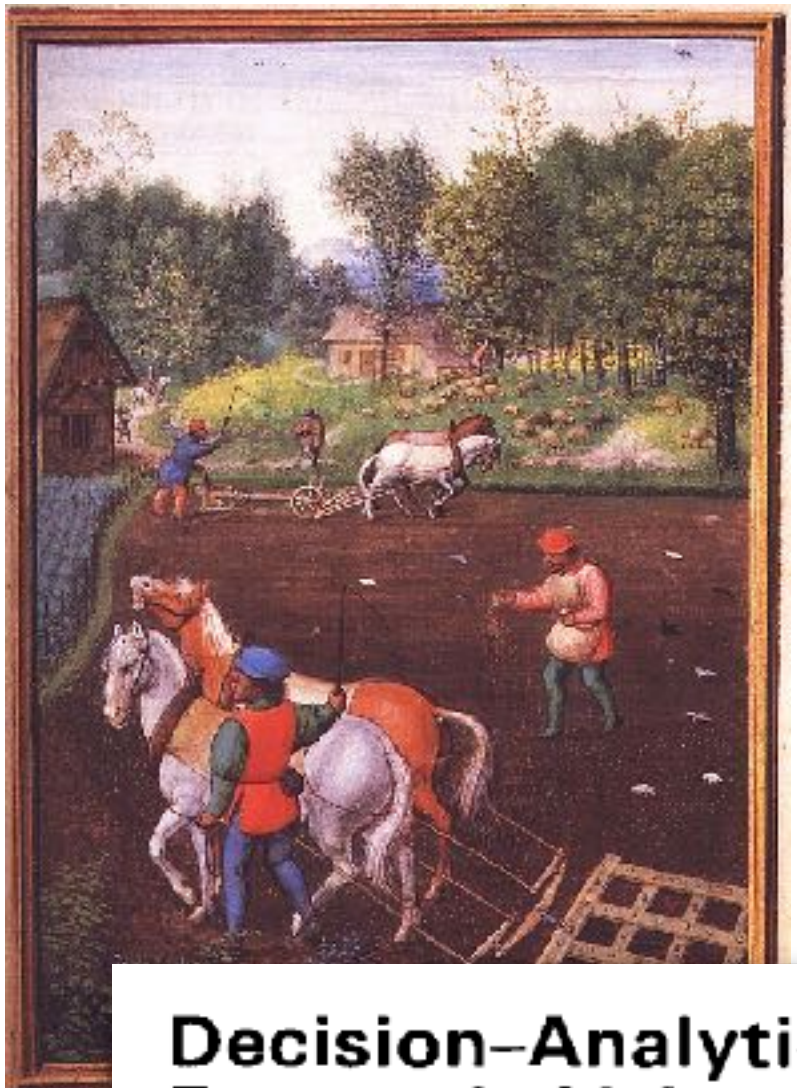
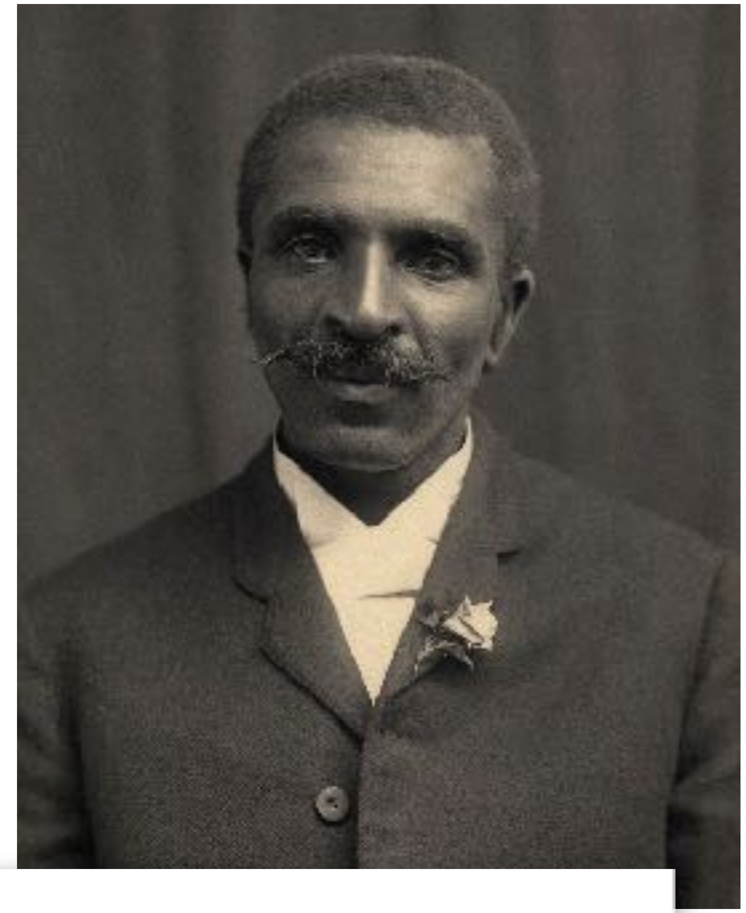
Markov Decision Process

Prof. Tamara Broderick

Edited From 6.036 Fall21 Offering





Decision–Analytic Assessment of the Economic Value of Weather Forecasts: The Fallowing/Planting Problem

RICHARD W. KATZ
National Center for Atmospheric Research, U.S.A.

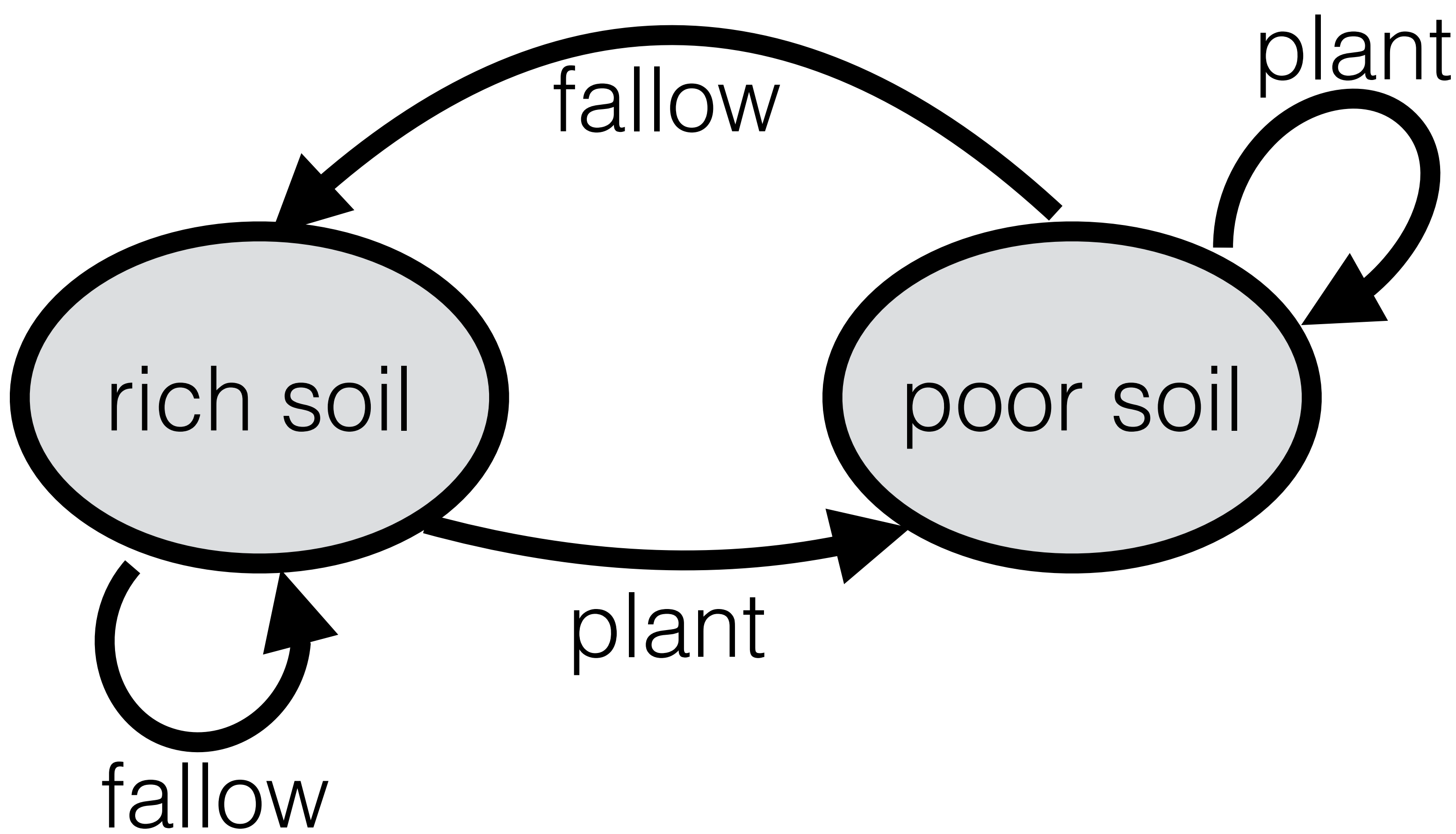
and

BARBARA G. BROWN* and ALLAN H. MURPHY
Oregon State University, U.S.A.

[https://en.wikipedia.org/wiki/Sowing#/media/File:Simon_Bening_-_September.jpg]
[https://en.wikipedia.org/wiki/George_Washington_Carver#/media/File:George_Washington_Carver_c1910_-_Restoration.jpg]
[Katz, Brown, Murphy 1987]

State Machine

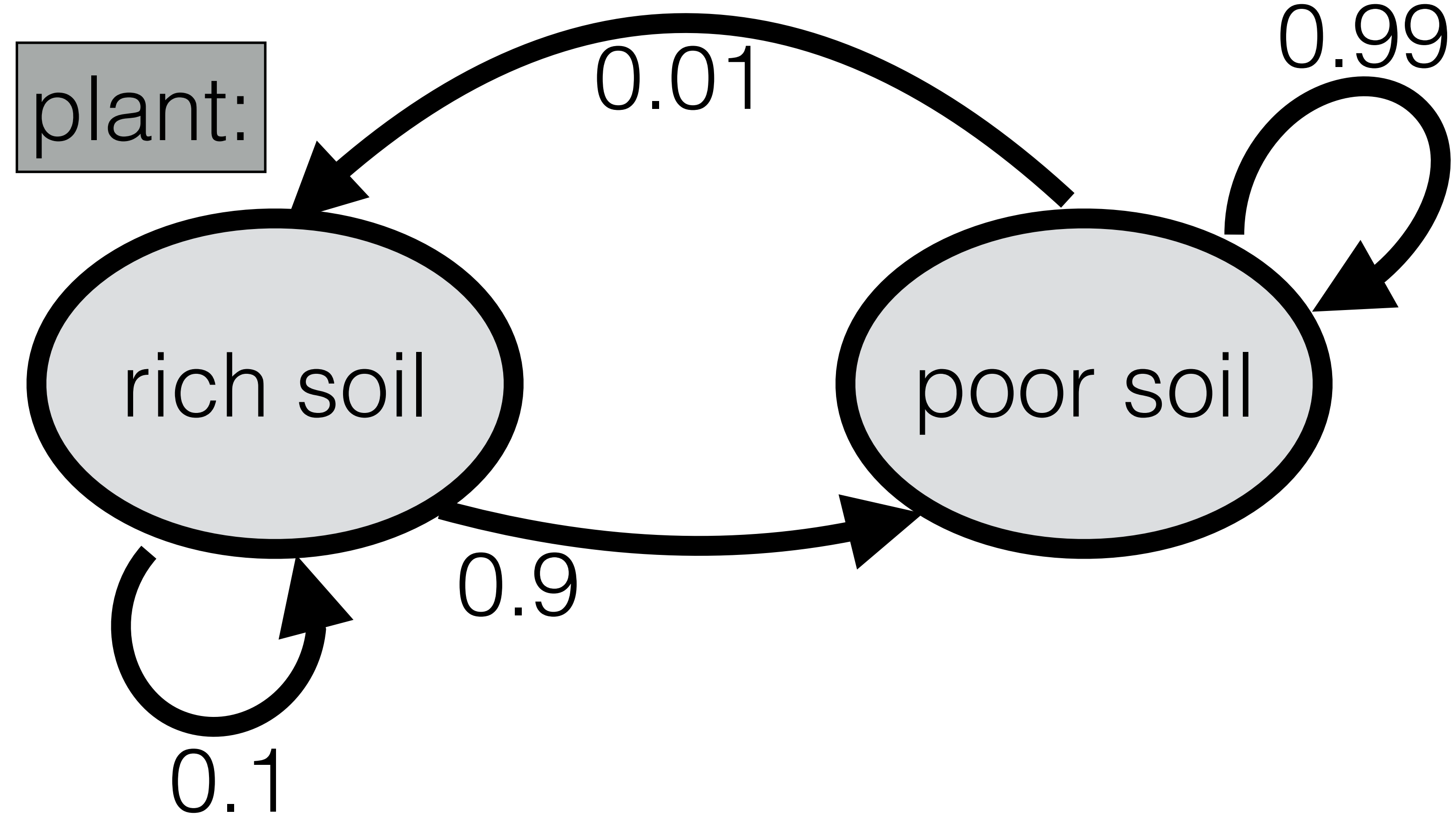
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function
- $\mathcal{Y}$: set of possible outputs
- $g : \mathcal{S} \to \mathcal{Y}$: output function
  - e.g. g(s) = s
  - e.g. g(s) = soil-moisture-sensor(s)



Example

$$s_0 = \text{rich}$$
$$s_1 = f(s_0, \text{plant}) = \text{poor};$$
$$y_1 = g(s_1) = \text{poor}$$
$$s_2 = f(s_1, \text{fallow}) = \text{rich};$$
$$y_2 = g(s_2) = \text{rich}$$

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T
  transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) = R(poor, fallow) = 0 bushels



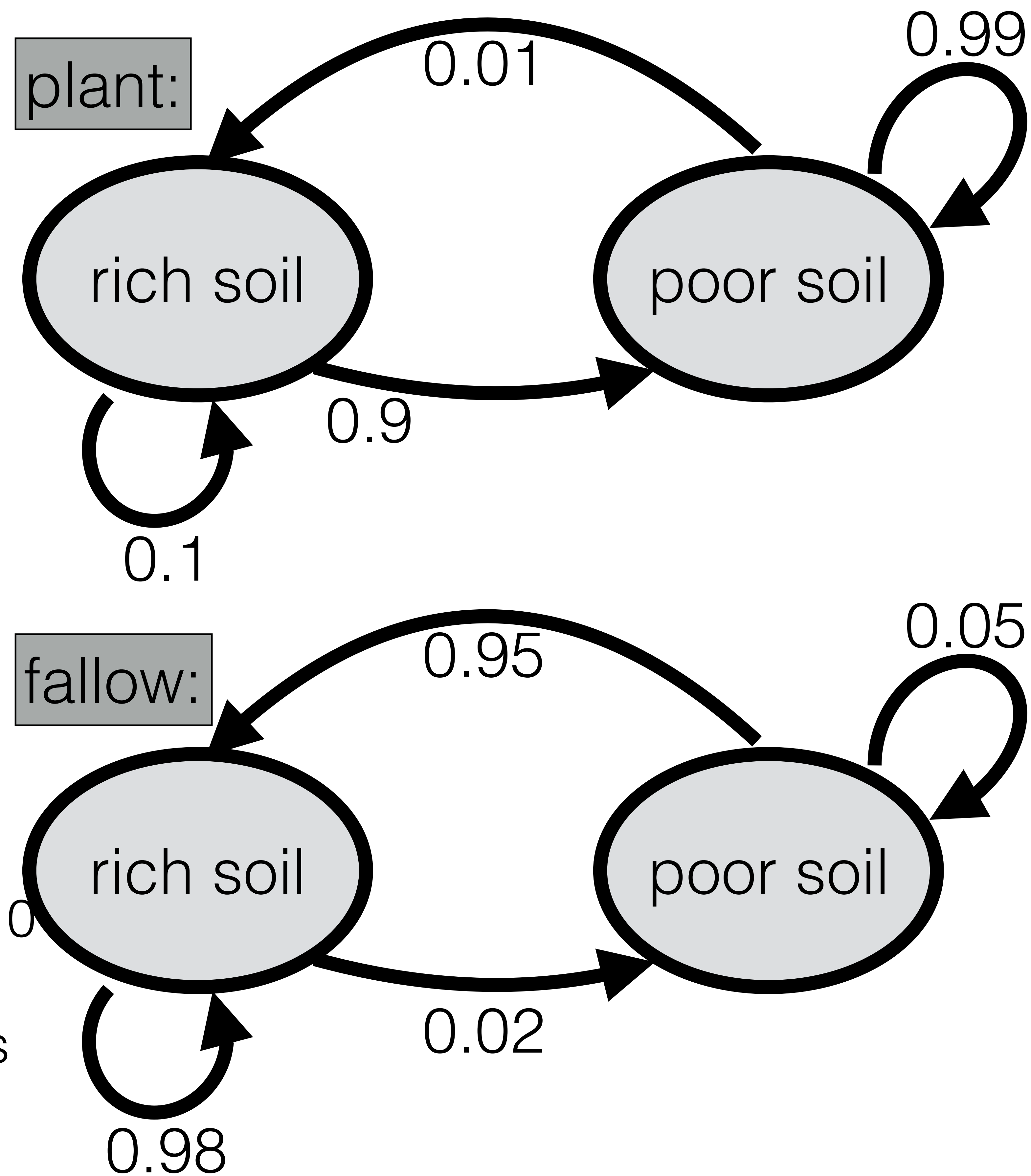
- Transition matrix for “plant” action:

end state

start state

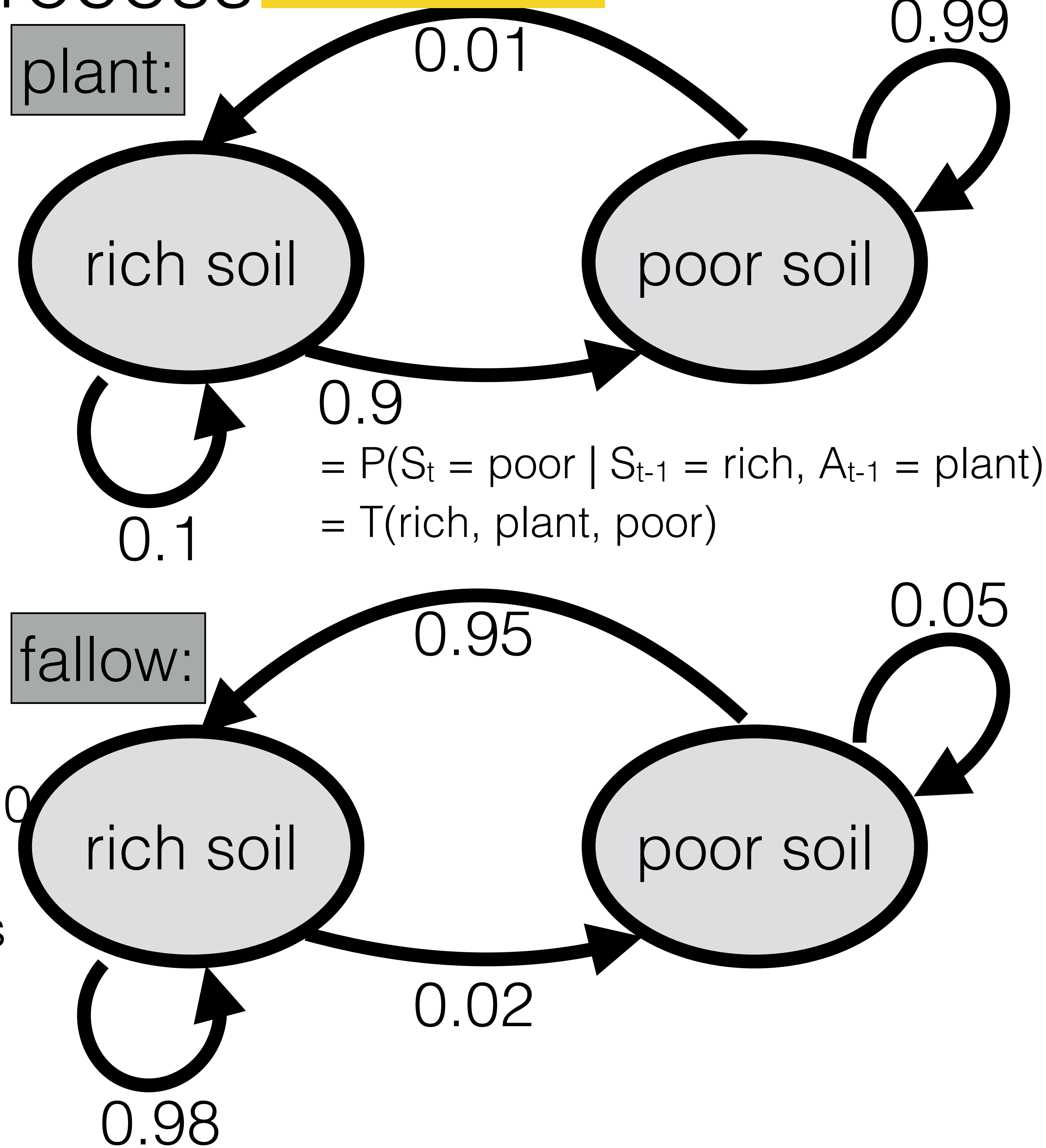
$$\begin{array}{r c} & \begin{array}{cc} \text{rich} & \text{poor} \end{array} \\ \begin{array}{r} \text{rich} \\ \text{poor} \end{array} & \left[\begin{array}{cc} 0.1 & 0.9 \\ 0.01 & 0.99 \end{array} \right] \end{array}$$

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
    - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) = R(poor, fallow) = 0 bushels



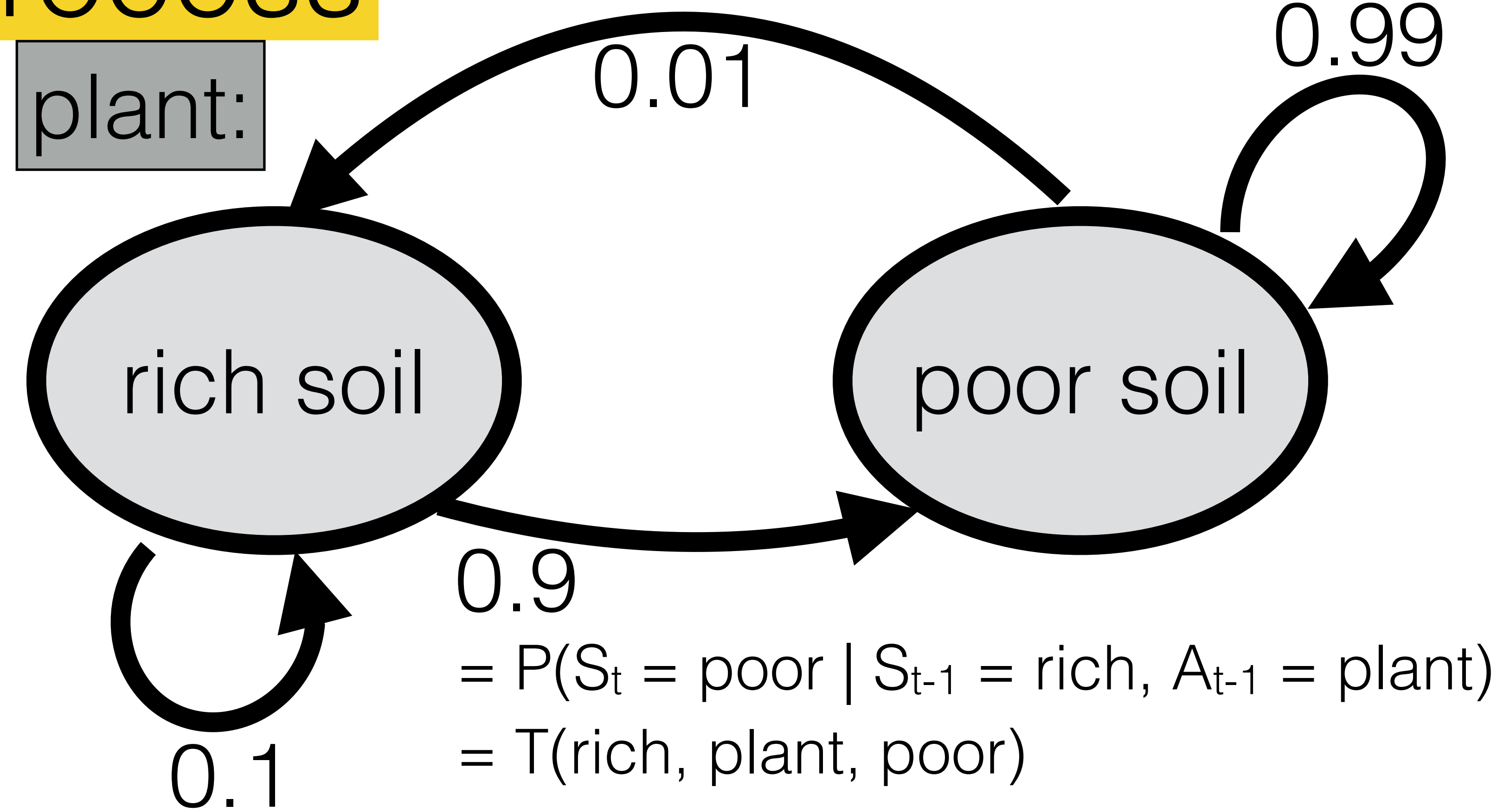
Markov Decision Process (basically)

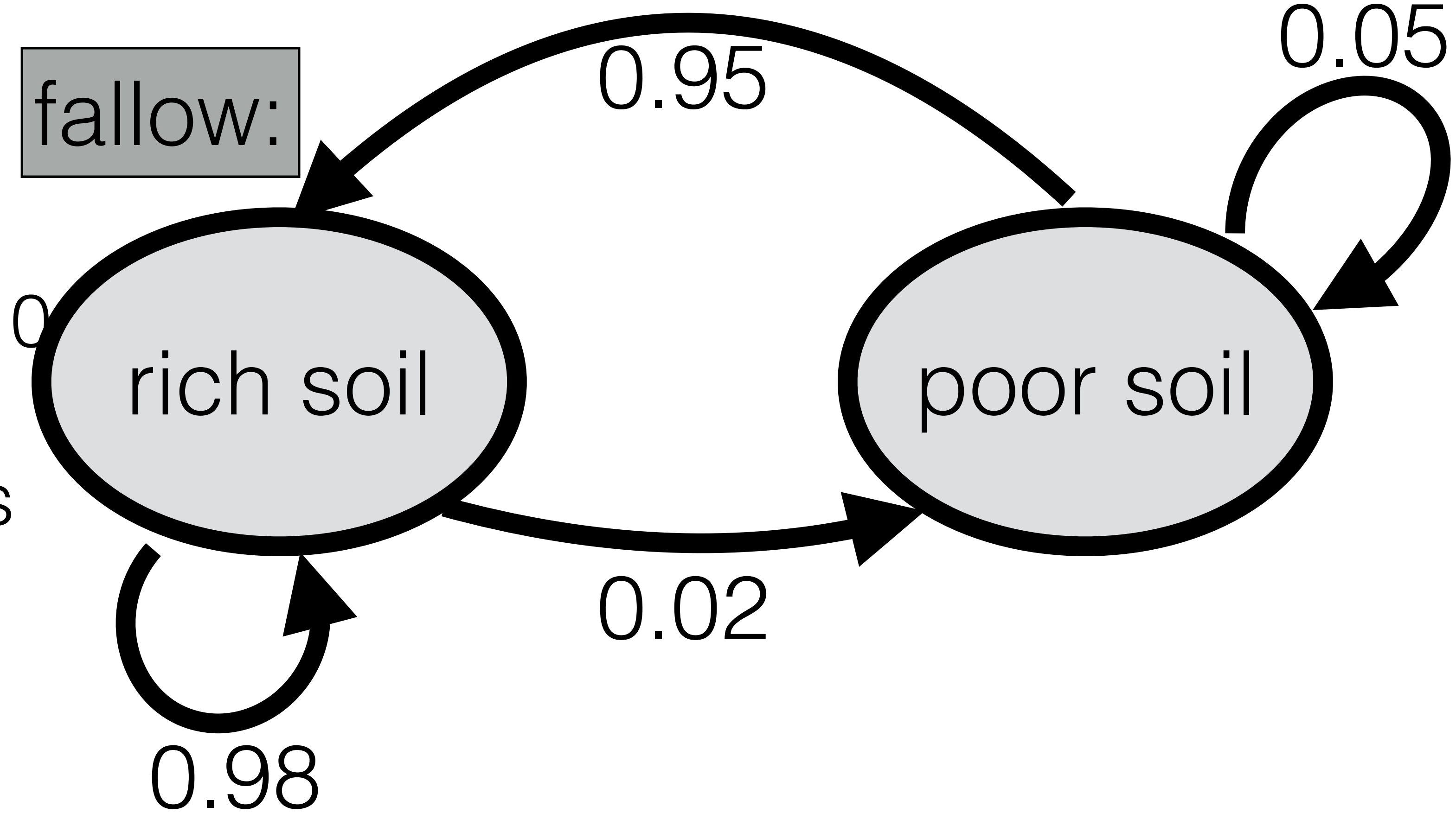
- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) = R(poor, fallow) = 0 bushels
- A discount factor



Markov Decision Process

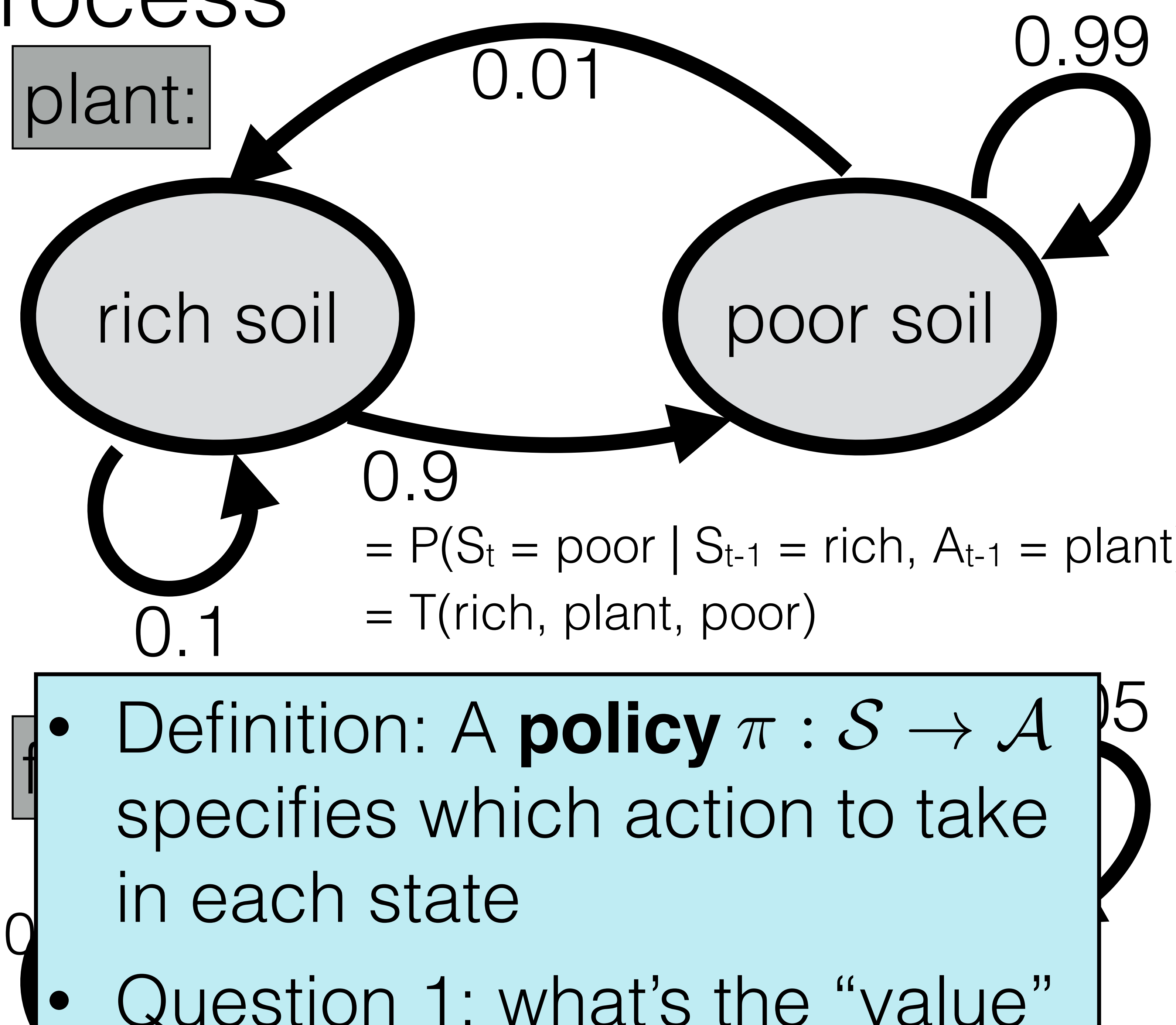
- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: reward function
    - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) = R(poor, fallow) = 0 bushels
- A discount factor





Markov Decision Process

- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) = R(poor, fallow) = 0 bushels
- A discount factor



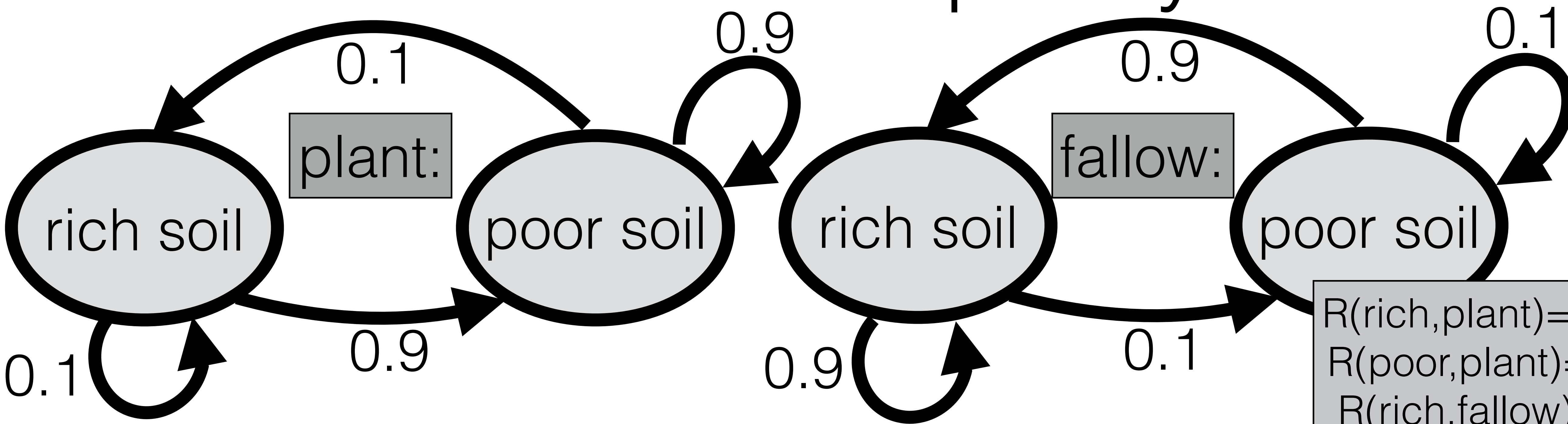
- Definition: A **policy** $\pi : \mathcal{S} \to \mathcal{A}$ specifies which action to take in each state
- Question 1: what's the "value" of a policy?
- Question 2: what's the best policy?

Expectation

- Suppose a random variable *R* has *m* possible values: $r_1, \ldots, r_m$
  - Example: a lottery pays $r_1 =$ $40*10^6$ USD if you win and $r_2 =$ -2 USD if you lose.
  - Question: if I could play this lottery a limitless number of times, how much could I expect to make each time I play, on average?
- Suppose $R = r_i$ with probability p_i
  - So we always have $\sum_{i=1}^{m} p_i = 1$
  - Example continued: $p_1 = 3.4*10^{-9}$
- Then the *expectation* of *R* is $\mathbb{E}[R] = \sum_{i=1}^{m} p_i r_i$
  - Example: $\mathbb{E}[R] =$ $3.4*10^{-9}$ x $40*10^6$ + (1 - $3.4*10^{-9}$) x -2

  = -1.86 USD

What's the value of a policy?

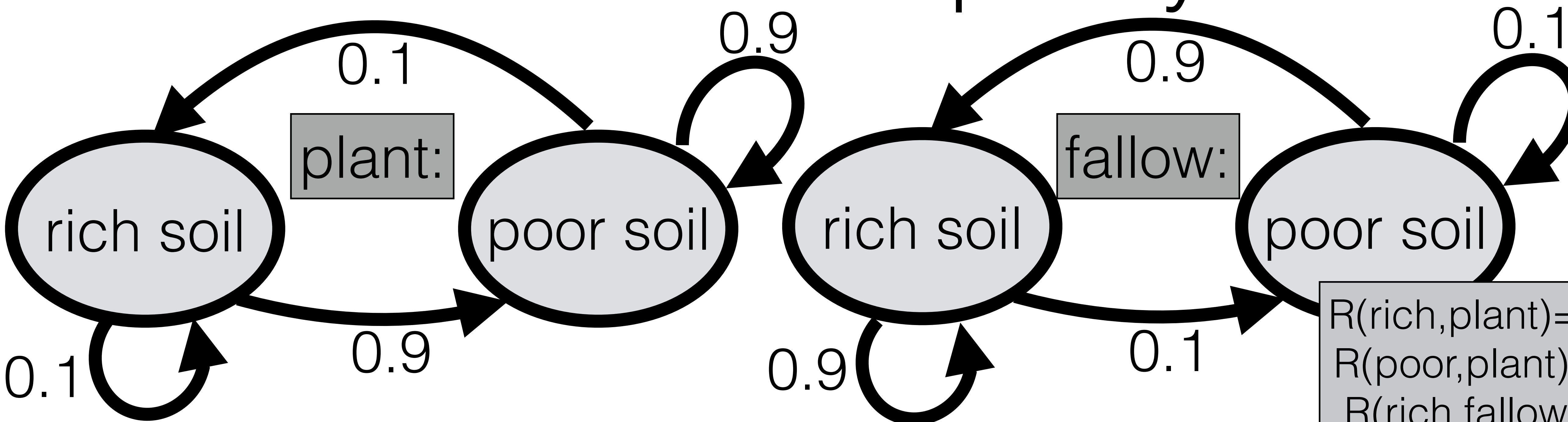


R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

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

I'm renting a field for *h* growing seasons. Then it will be destroyed to make a strip mall.

What's the value of a policy?



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

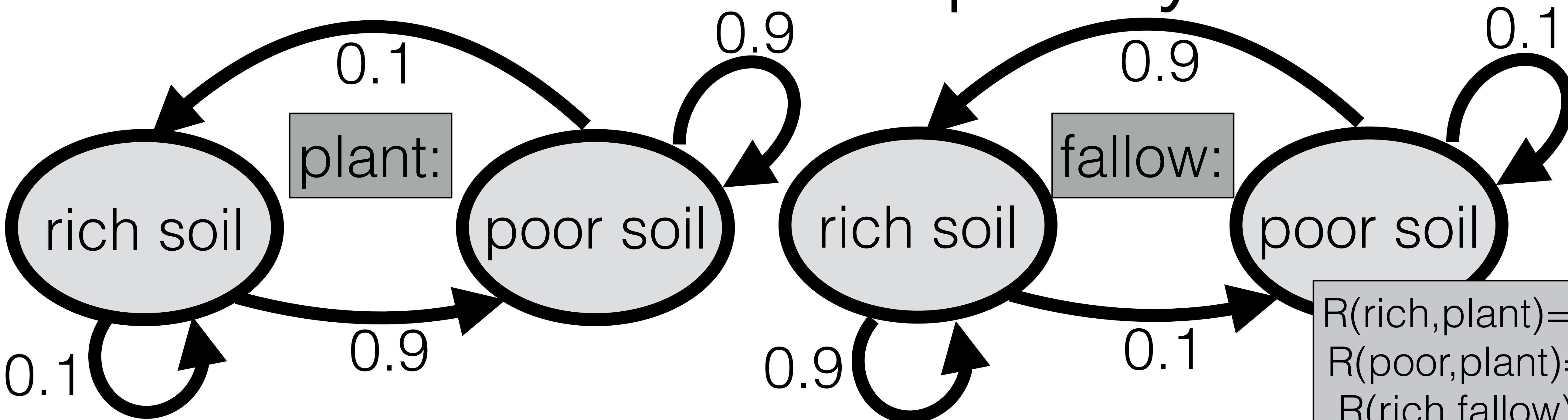
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy π starting at *s*

Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

$$V_\pi^0(s) = 0;\ \underbrace{V_\pi^h(s)}_{\text{value of the policy with } h \text{ steps left}} = \underbrace{R(s, \pi(s))}_{\text{value of the policy on this time step}} + \underbrace{\sum_{s'} T(s, \pi(s), s') \cdot V_\pi^{h-1}(s')}_{\text{(expected) value of the policy across all future time steps}}$$

$$V_{\pi_A}^1(\text{rich}) = 100;\ V_{\pi_A}^1(\text{poor}) = 10;\ V_{\pi_B}^1(\text{rich}) = 100;\ V_{\pi_B}^1(\text{poor}) = 0$$

What's the value of a policy?



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

- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy π starting at *s*

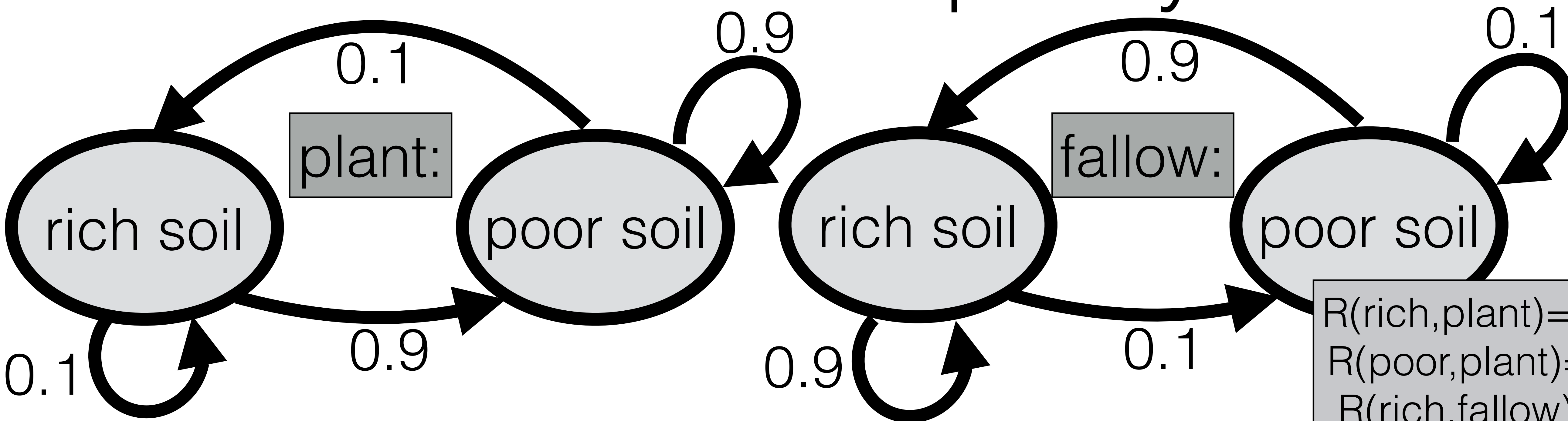
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

$$V_{\pi_A}^1(\text{rich}) = 100; V_{\pi_A}^1(\text{poor}) = 10; V_{\pi_B}^1(\text{rich}) = 100; V_{\pi_B}^1(\text{poor}) = 0$$

$$\begin{aligned} V_{\pi_A}^2(\text{rich}) &= R(\text{rich}, \pi_A(\text{rich})) + T(\text{rich}, \pi_A(\text{rich}), \text{rich}) V_{\pi_A}^1(\text{rich}) \\ &\quad + T(\text{rich}, \pi_A(\text{rich}), \text{poor}) V_{\pi_A}^1(\text{poor}) \\ &= 100 + (0.1)(100) + (0.9)(10) \\ &= 119 \end{aligned}$$

What's the value of a policy?



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

- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy π starting at *s*

Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

$$V_{\pi_A}^1(\text{rich}) = 100; V_{\pi_A}^1(\text{poor}) = 10; V_{\pi_B}^1(\text{rich}) = 100; V_{\pi_B}^1(\text{poor}) = 0$$

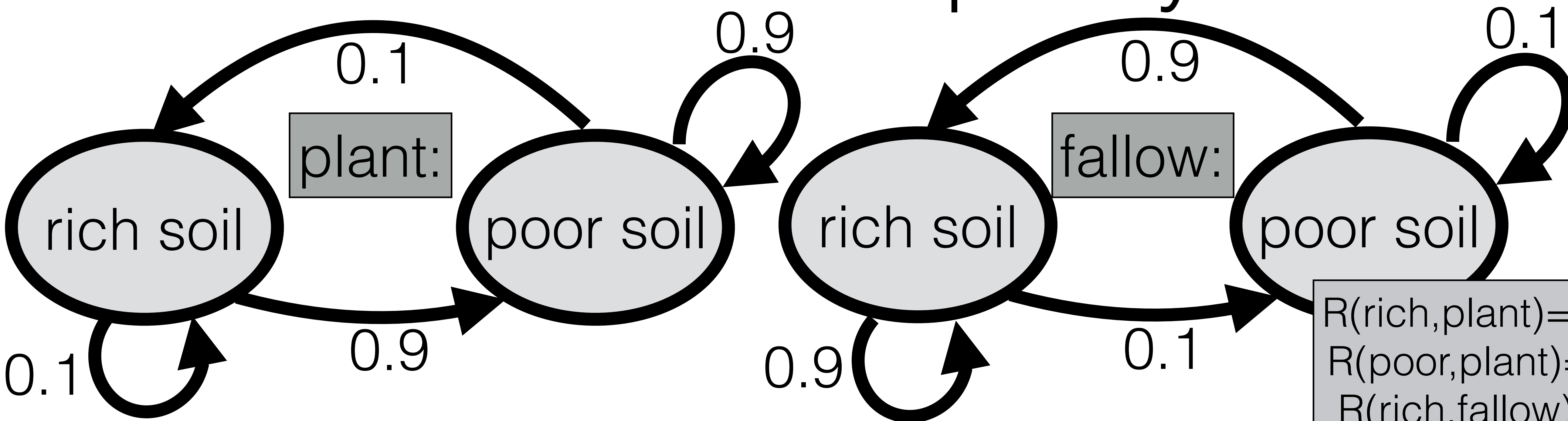
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

$$V_{\pi_A}^3(\text{rich}) = 138; V_{\pi_A}^3(\text{poor}) = 48; V_{\pi_B}^3(\text{rich}) = 192; V_{\pi_B}^3(\text{poor}) = 108$$

Who wins?

I.e. at least as good at all states and strictly better for at least one state

What's the value of a policy?



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

- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy π starting at *s*

Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

$$V_{\pi_A}^1(\text{rich}) = 100; V_{\pi_A}^1(\text{poor}) = 10; V_{\pi_B}^1(\text{rich}) = 100; V_{\pi_B}^1(\text{poor}) = 0$$

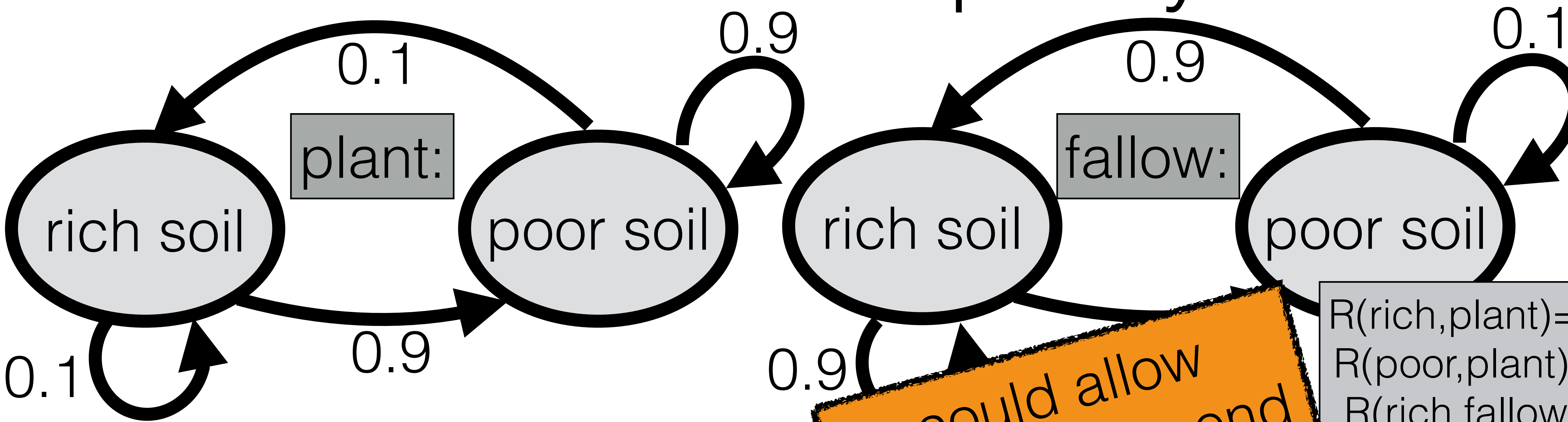
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

$$V_{\pi_A}^3(\text{rich}) = 138; V_{\pi_A}^3(\text{poor}) = 48; V_{\pi_B}^3(\text{rich}) = 192; V_{\pi_B}^3(\text{poor}) = 108$$

Who wins? $\pi_A >_{h=1} \pi_B; \pi_A <_{h=3} \pi_B;$ Neither policy wins for *h* = 2

- I.e. at least as good at all states and strictly better for at least one state

What's the value of a policy?



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

could allow policy to depend on horizon ("non-stationary")

- *h*: horizon (e.g. how many gro
- $V_\pi^h(s)$: value (expected reward ing at *s*

Dueling farmers! π_A: always plan nt if rich, else fallow

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

$$V_{\pi_A}^1(\text{rich}) = 100; V_{\pi_A}^1(\text{poor}) = 10; V_{\pi_B}^1(\text{rich}) = 100; V_{\pi_B}^1(\text{poor}) = 0$$

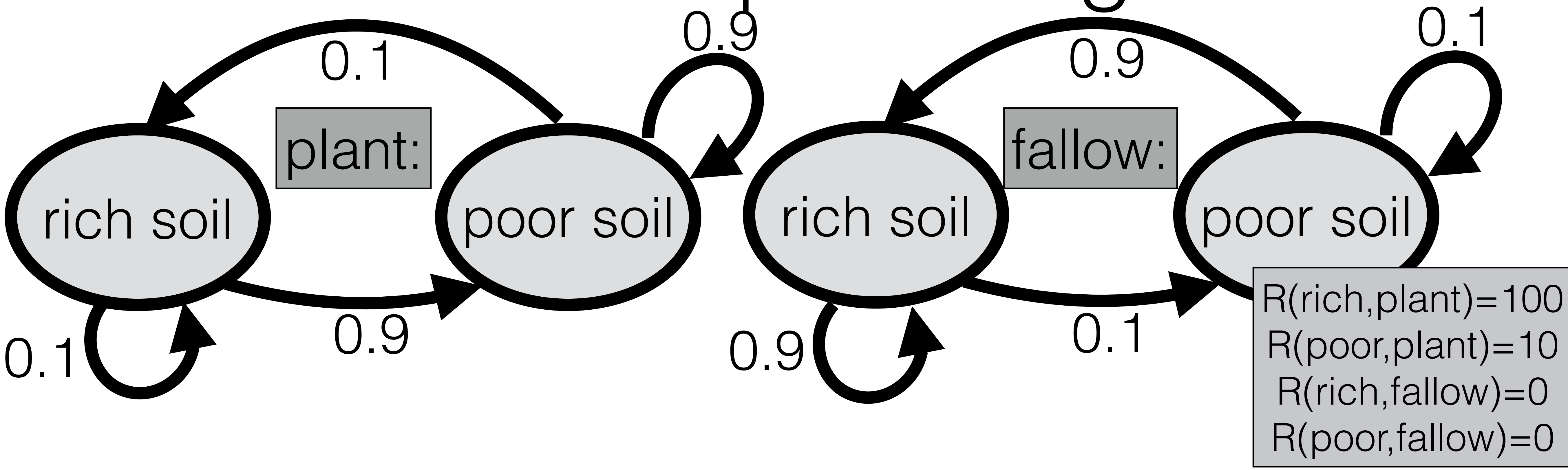
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

$$V_{\pi_A}^3(\text{rich}) = 138; V_{\pi_A}^3(\text{poor}) = 48; V_{\pi_B}^3(\text{rich}) = 192; V_{\pi_B}^3(\text{poor}) = 108$$

Who wins? $\pi_A >_{h=1} \pi_B ; \pi_A <_{h=3} \pi_B$ value of delayed gratification

9 I.e. at least as good at all states and strictly better for at least one state

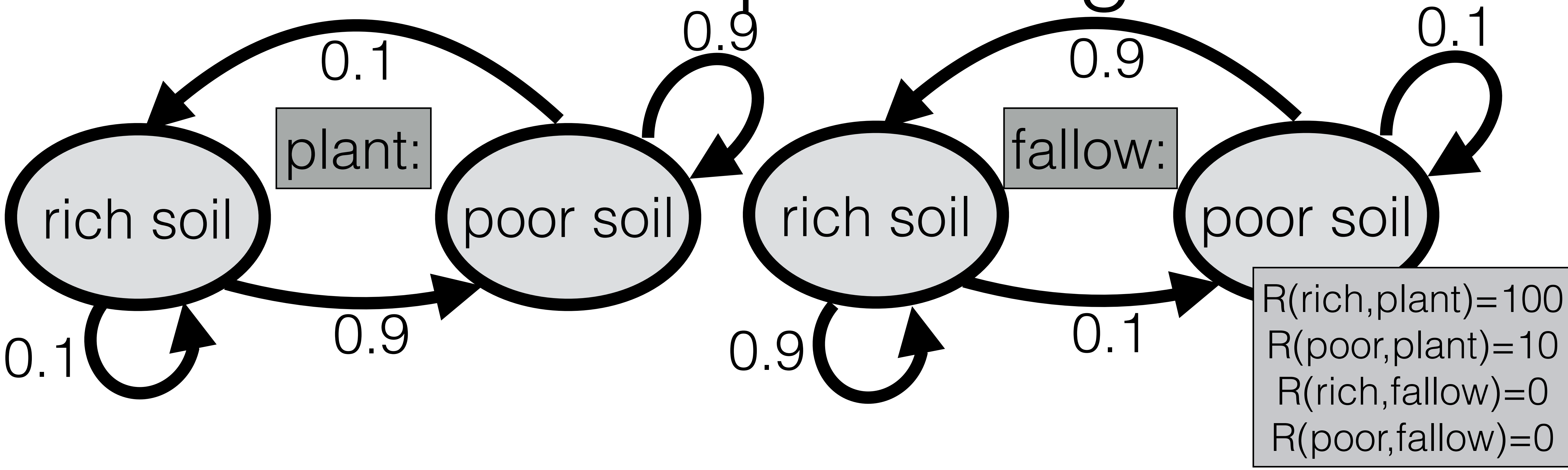
What if I don't stop farming?



- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels
  - Example: What's the value of 1 bushel per year forever?

$$\underbrace{V}_{\text{value for all future}} = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma + \gamma^2 + \cdots) = \underbrace{1}_{\text{value \textit{on} first time step}} + \underbrace{\gamma V}_{\text{value \textit{after} first time step}}$$

What if I don't stop farming?



- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels
  - Example: What's the value of 1 bushel per year forever?

$$V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma + \gamma^2 + \cdots) = 1 + \gamma V$$

$V = 1/(1-\gamma)$ E.g. $\gamma = 0.99 \Rightarrow V = 1/0.01 = 100$ bushels

- $V_\pi(s)$: expected reward with policy π starting at state *s*

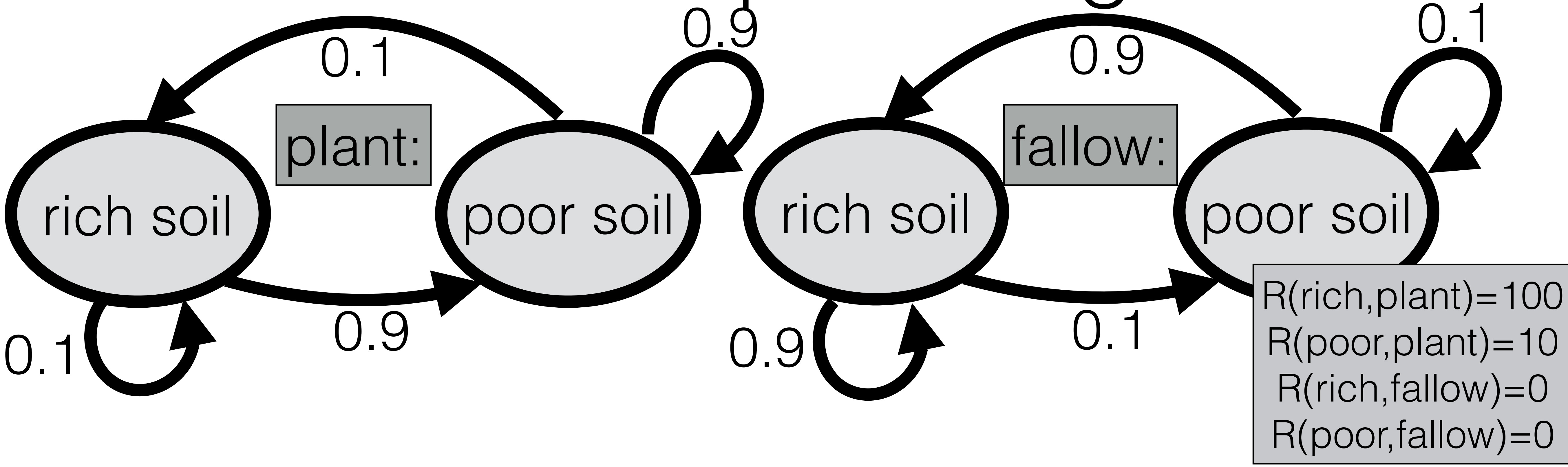
$$V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$$

policy value for all future

policy value on first time step

(expected) policy value *after* first time step

What if I don't stop farming?

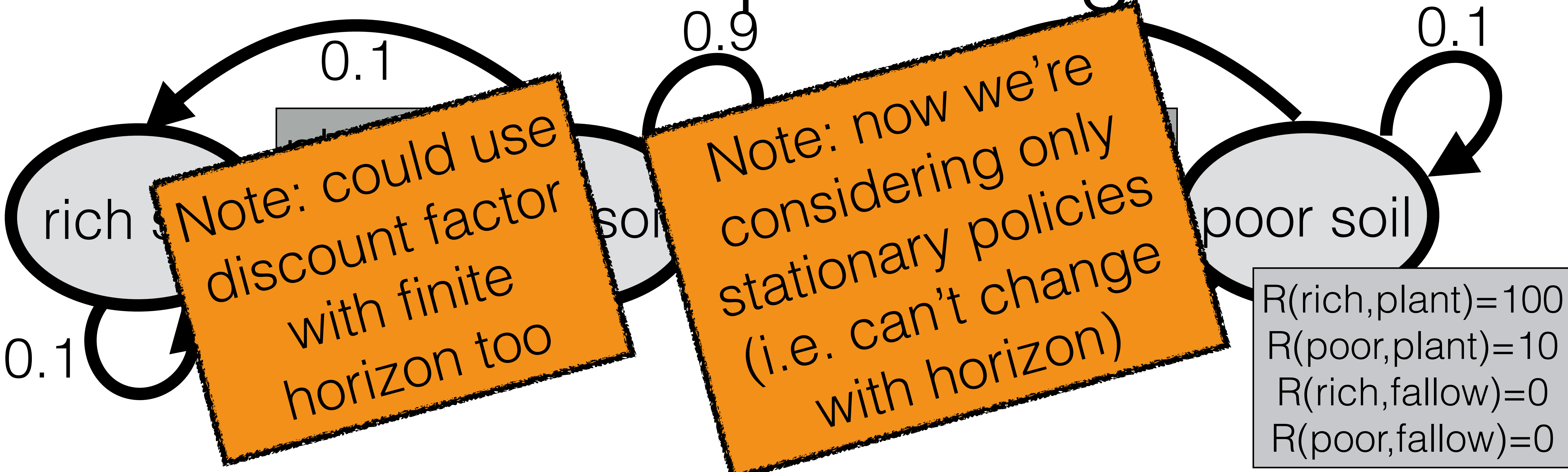


- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels
  - Example: What's the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma + \gamma^2 + \cdots) = 1 + \gamma V$
    $V = 1/(1-\gamma)$ E.g. $\gamma = 0.99 \Rightarrow V = 1/0.01 = 100$ bushels
- $V_\pi(s)$: expected reward with policy π starting at state *s*

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

  - $|\mathcal{S}|$ linear equations in $|\mathcal{S}|$ unknowns

What if I don't stop farming?

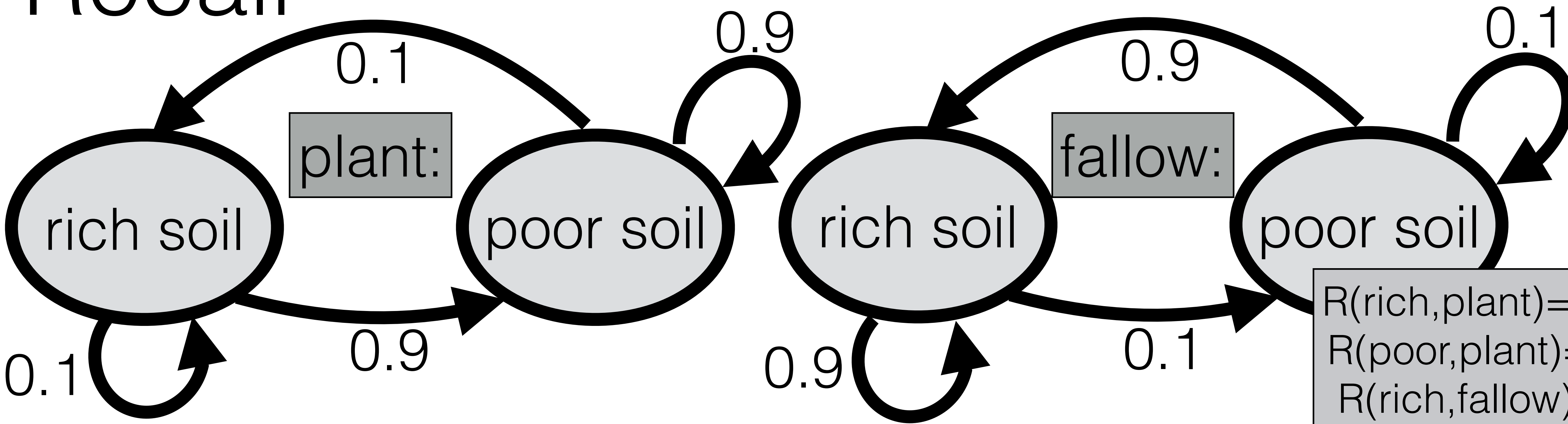


- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels
  - Example: What's the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma + \gamma^2 + \cdots) = 1 + \gamma V$
    $V = 1/(1-\gamma)$ E.g. $\gamma = 0.99 \Rightarrow V = 1/0.01 = 100$ bushels
- $V_\pi(s)$: expected reward with policy π starting at state *s*

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

  - $|\mathcal{S}|$ linear equations in $|\mathcal{S}|$ unknowns

Recall



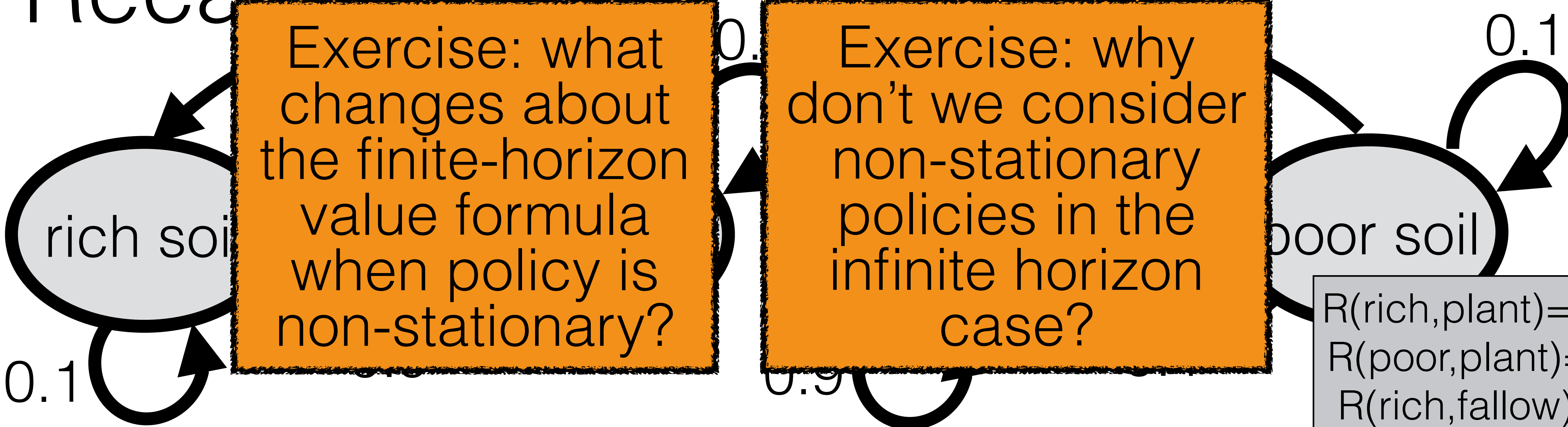
R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

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

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

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

Recall



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

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

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

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

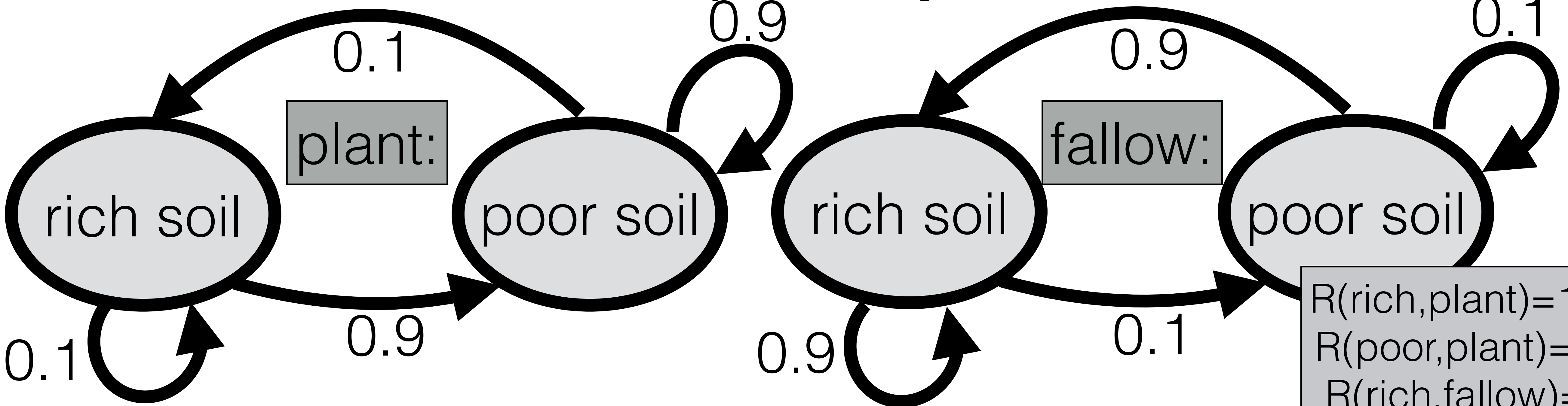
$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi^{h-1}(s')$$

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

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

- Next question: What's the best policy?

What's the best policy? Finite horizon



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

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

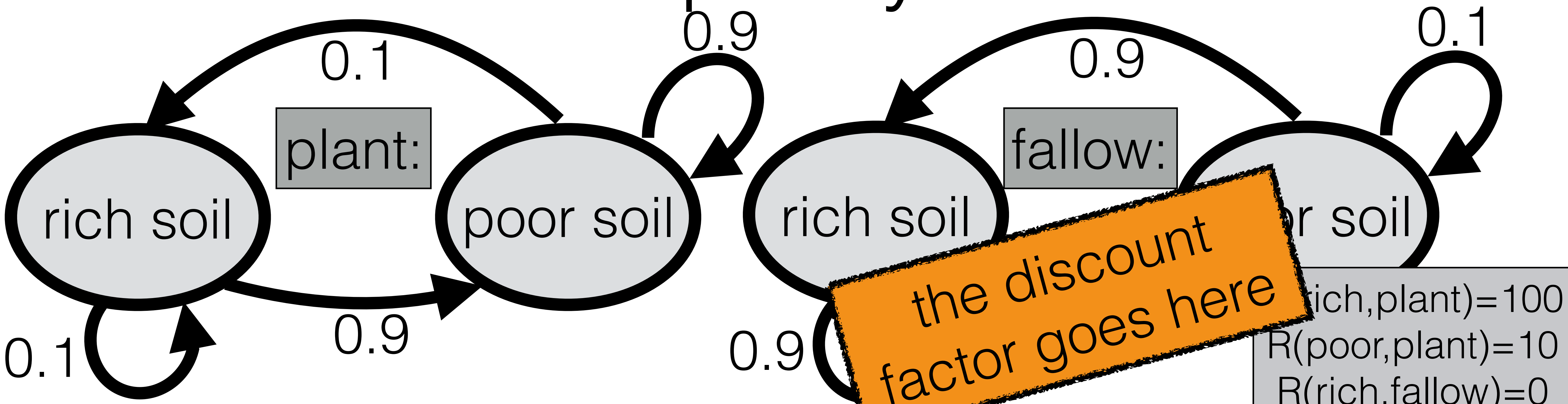
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

What's best?

Any *s*, $\pi_1^*(s) = \text{plant}$

What's the best policy? Finite horizon



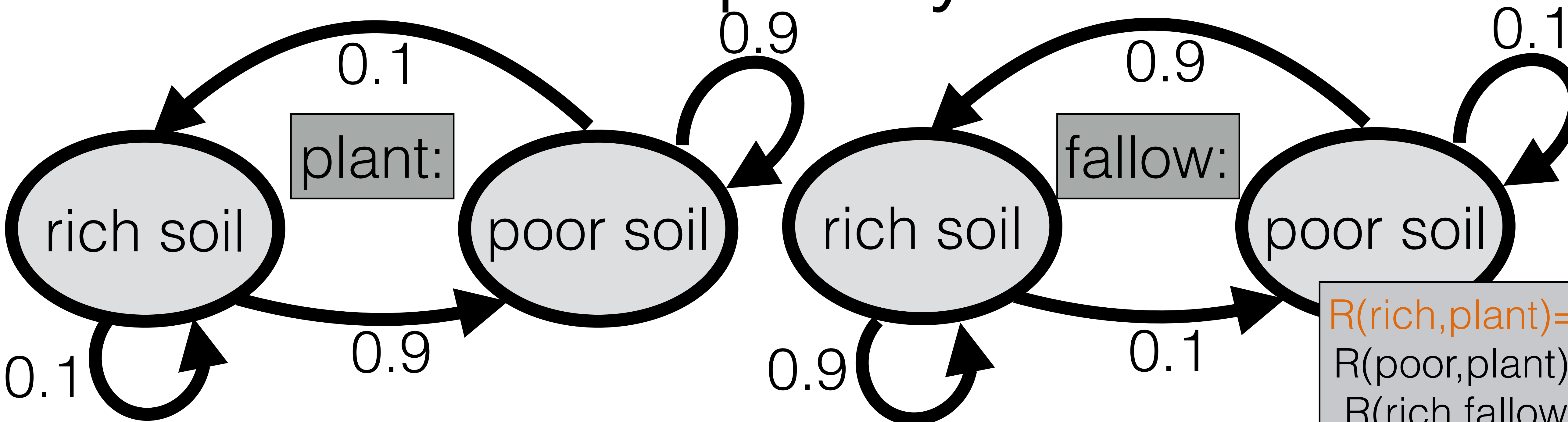
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

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

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

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

What's the best policy? Finite horizon



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

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

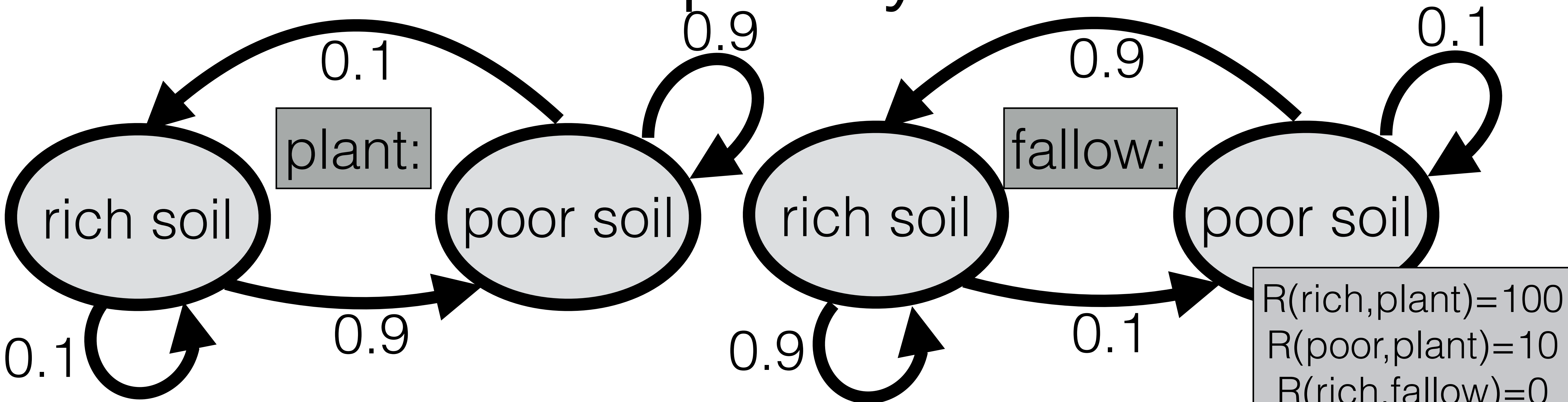
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

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

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

What's the best policy? Finite horizon



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

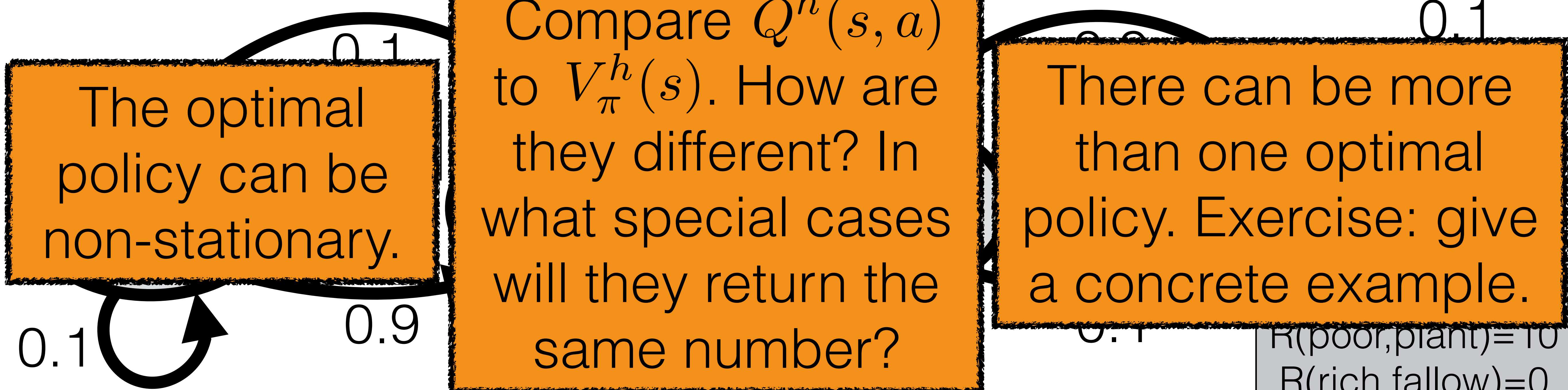
$$Q^0(s, a) = 0; Q^h(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \max_{a'} Q^{h-1}(s', a')$$

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

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

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

What's the best policy? Finite horizon



The optimal policy can be non-stationary.

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

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

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

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

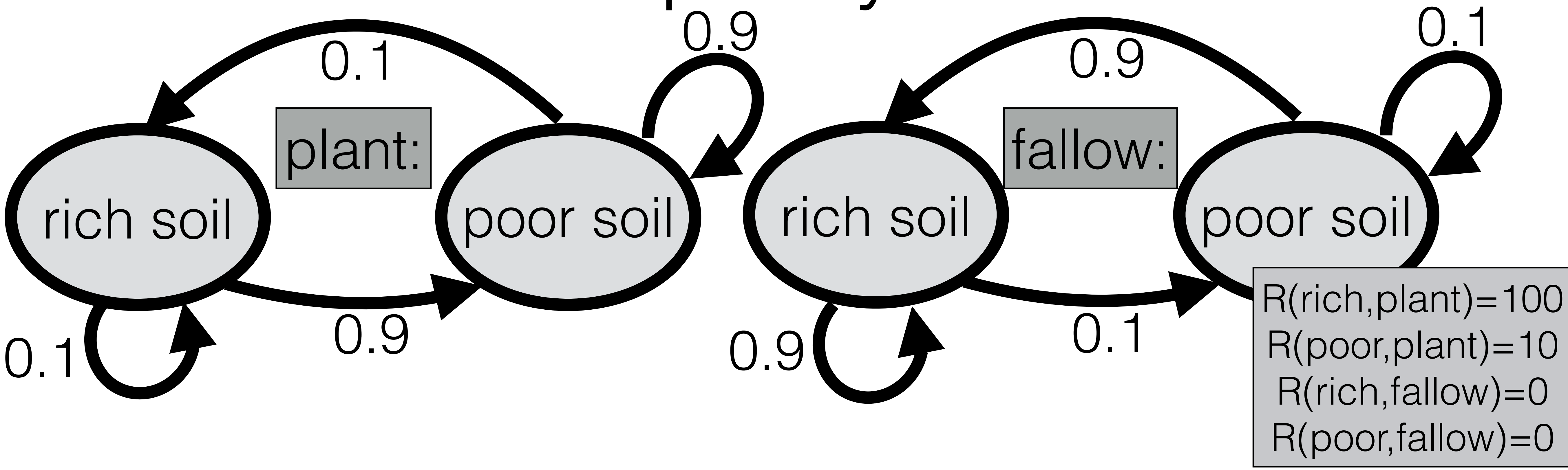
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

"finite-horizon value iteration"

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

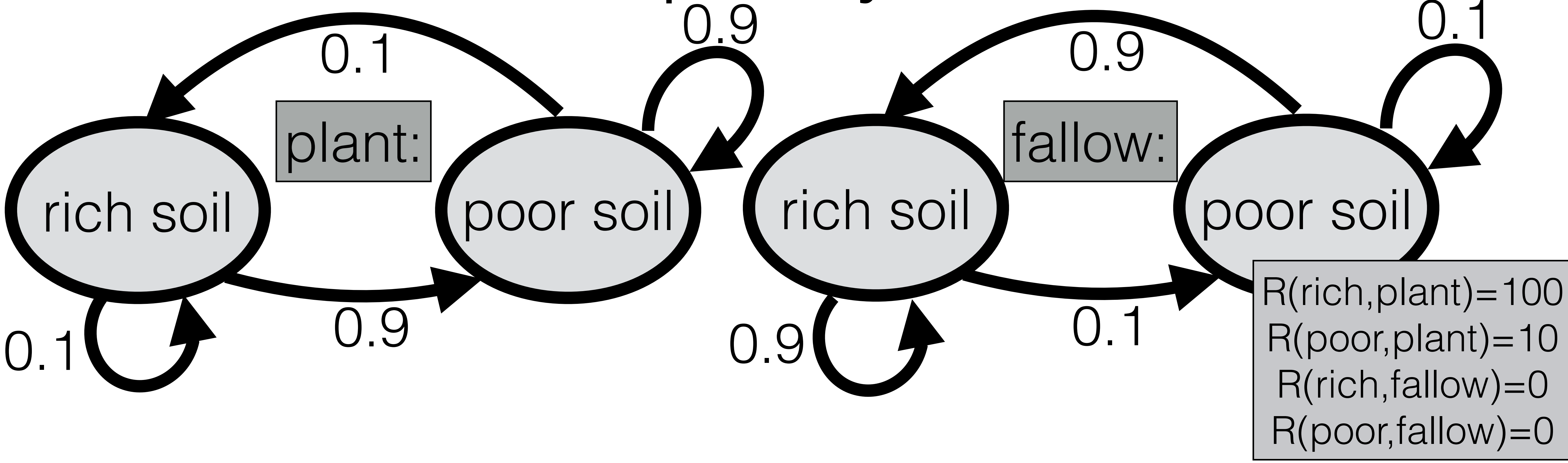
What's the best policy? Infinite horizon



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

Recall farmer A and farmer B from last time

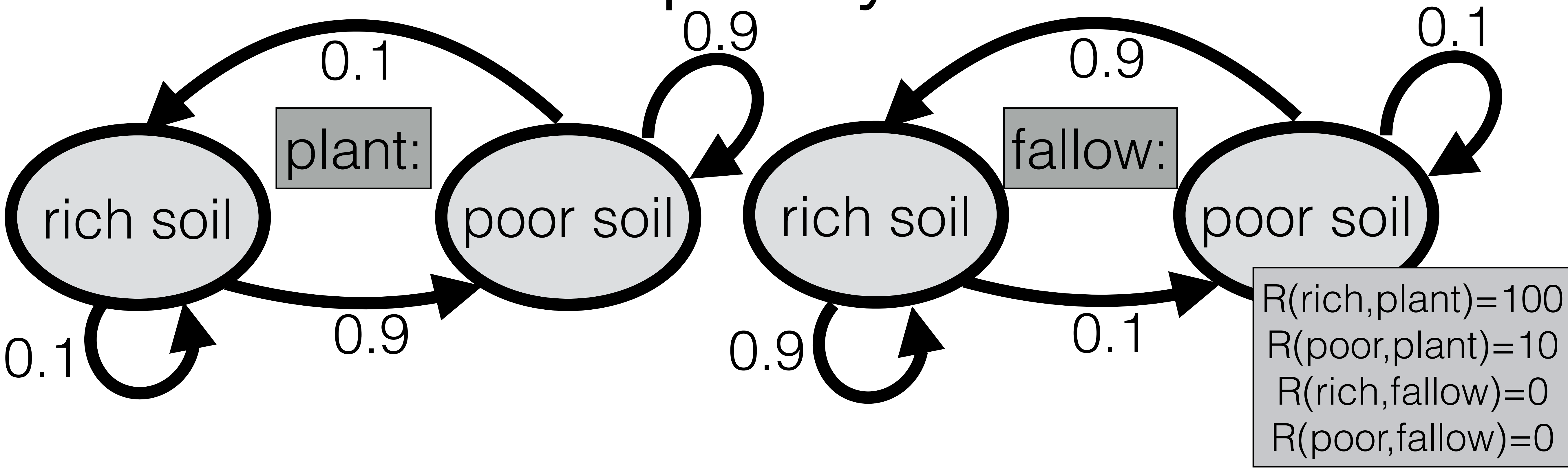
What's the best policy? Infinite horizon



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

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

What's the best policy? Infinite horizon



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

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

Infinite-Horizon Value Iteration

- Recall the finite-horizon case:

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

- A similar flavor for the infinite-horizon case:

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

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

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

  while True

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

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

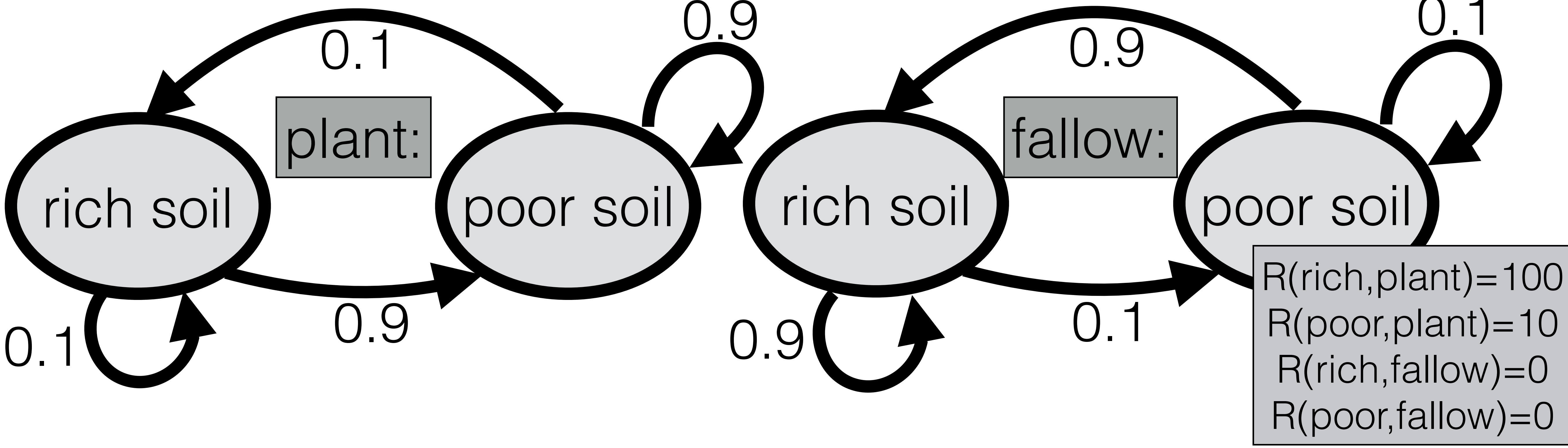
    if $\max_{s,a} |Q_{\text{old}}(s,a) - Q_{\text{new}}(s,a)| < \epsilon$

      return Q_{new}

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

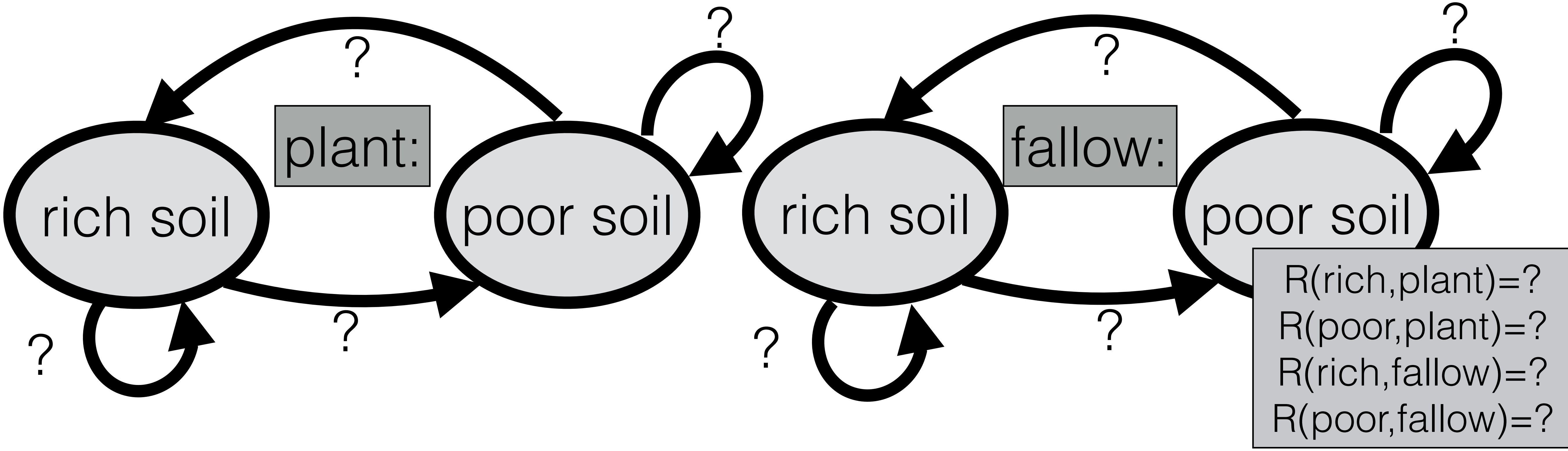
In real code, always cap the # of iterations

But what do we know at the start?



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

But what do we know at the start?



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