MIT EECS
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

# 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 9" category)

**Materials:** Will all be available at course website

## Last Time(s)

I. Regression, classification

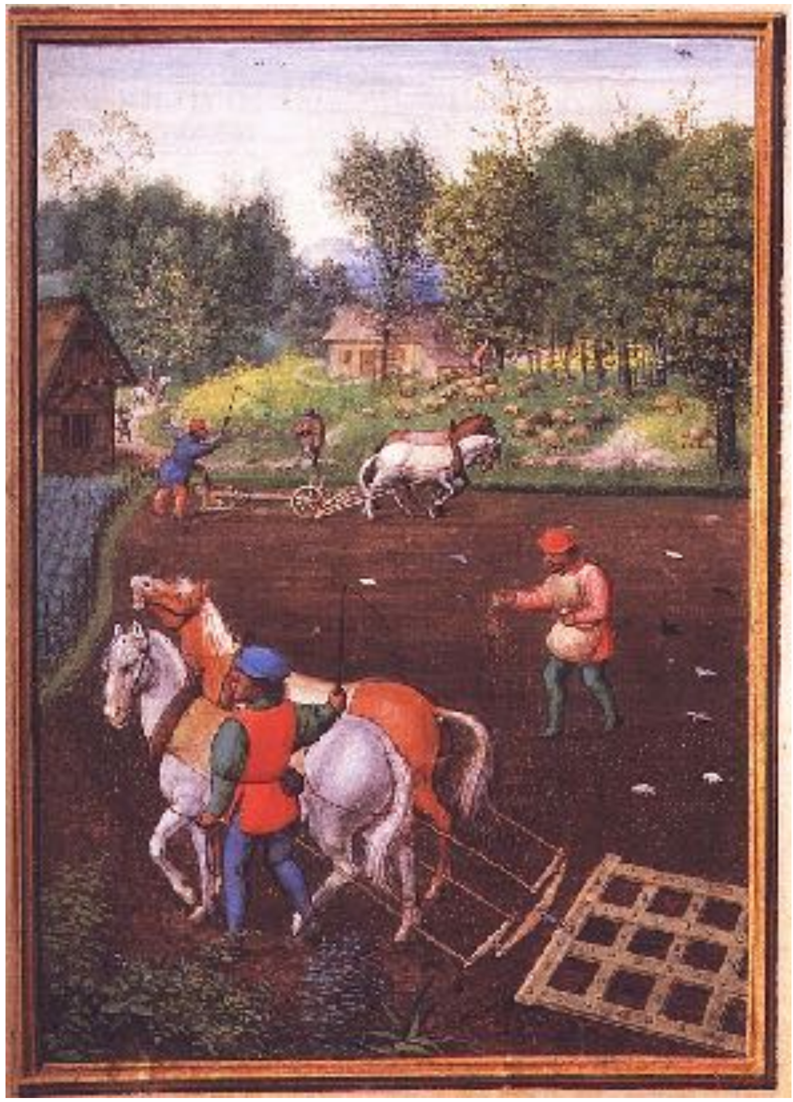
II. Decisions incur loss but don't have broader effect

## Today's Plan

I. Decisions change the state of the world

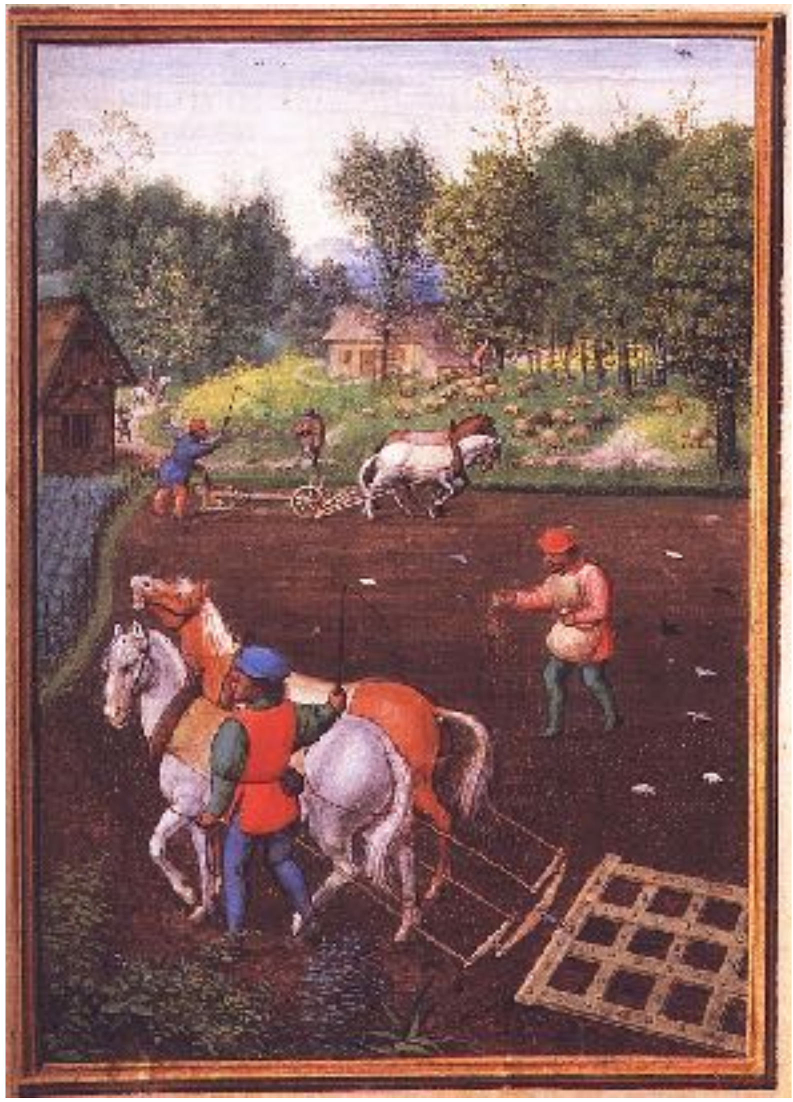
II. State machines

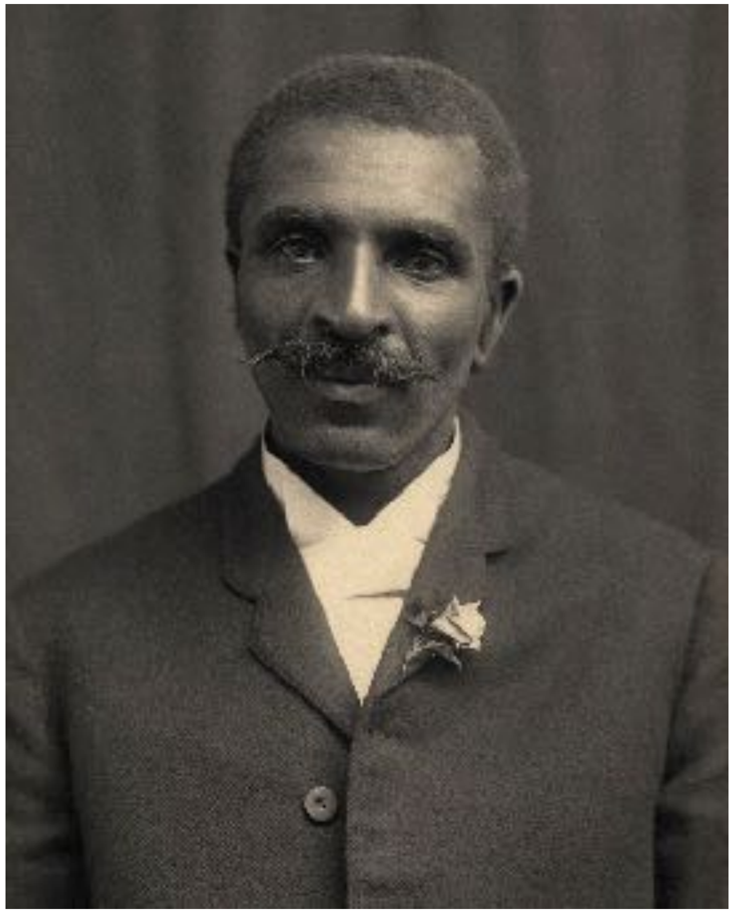
III. Markov decision processes (MDPs)



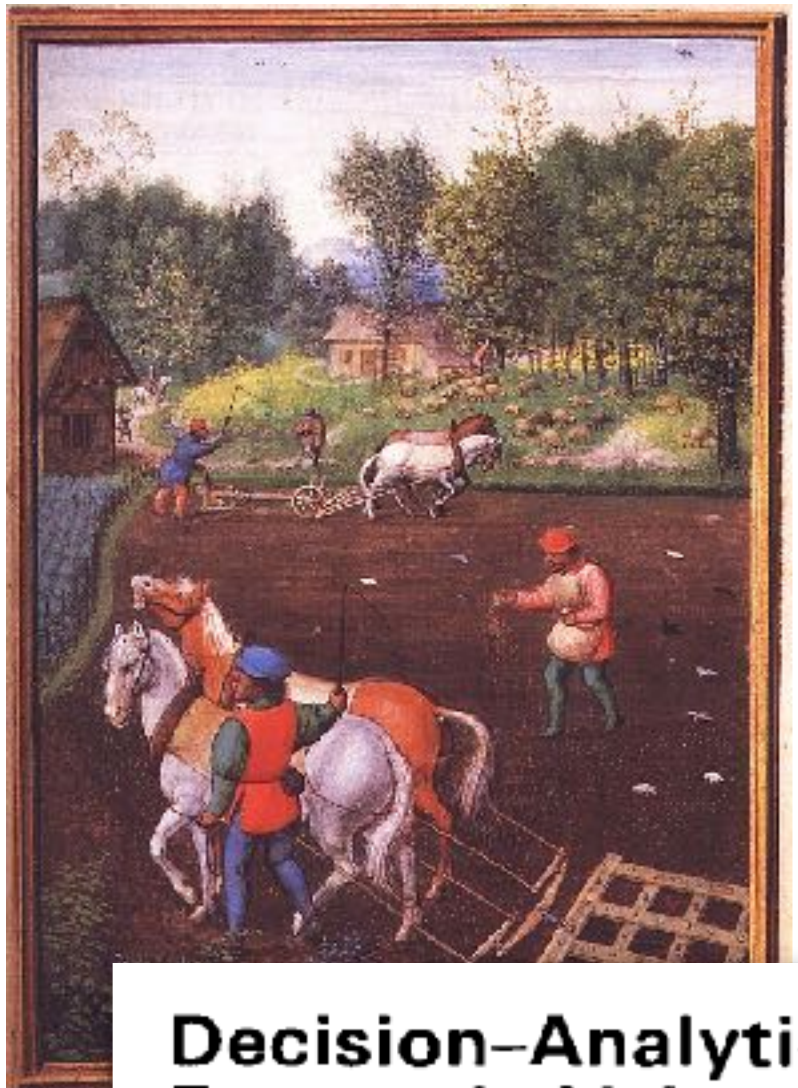


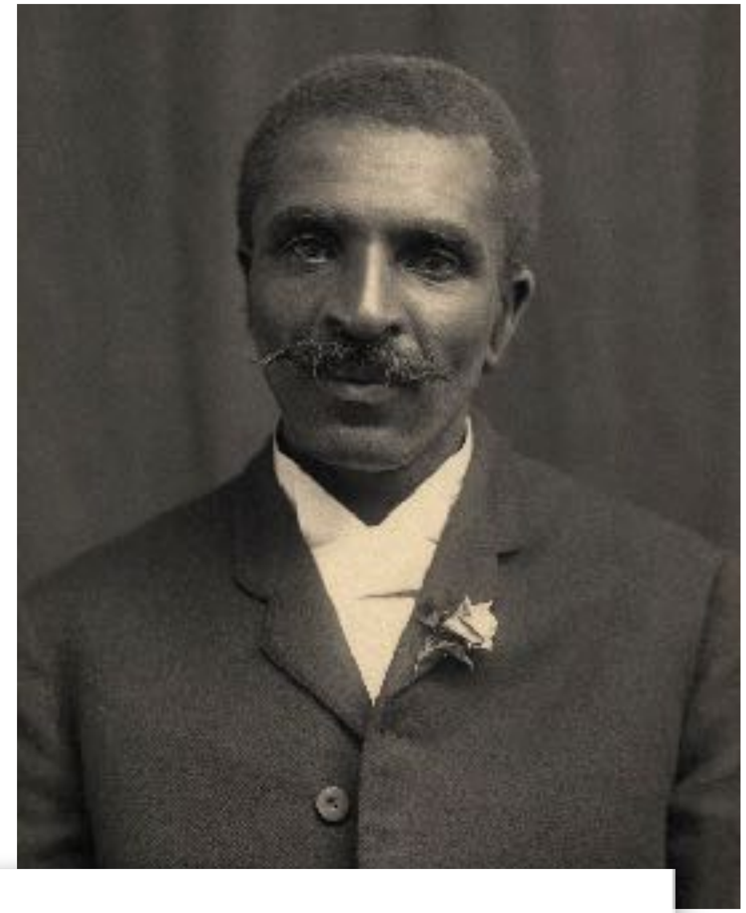
[ https://en.wikipedia.org/wiki/Sowing#/media/File:Simon_Bening_-_September.jpg ]





[ 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 ]





**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.*

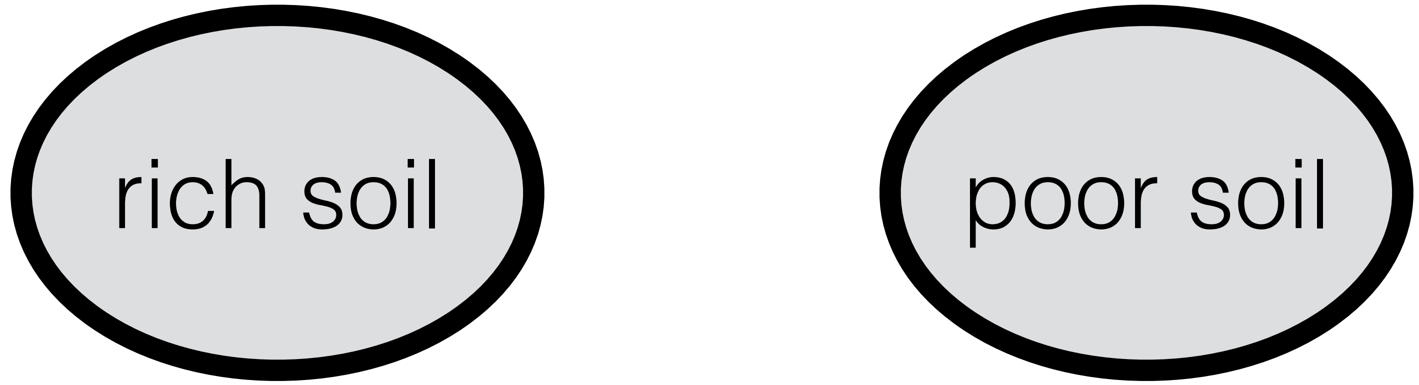
[ https://en.wikipedia.org/wiki/Sowing#/media/File:Simon_Bening_-_September.jpg ]
[ https://en.wikipedia.org/wiki/George_Washington_Carver#/media/File:George_Washington_Carver_c1910_-_Restoration.jpg ]
[ Katz, Brown, Murphy 1987 ]

# State Machine

- $\mathcal{S}$ = set of possible states

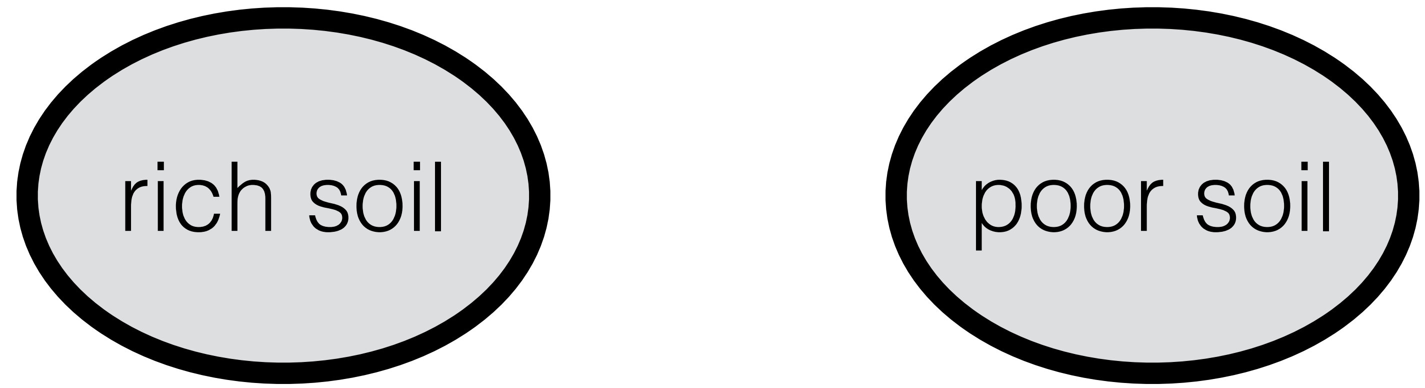
# State Machine

- $\mathcal{S}$ = set of possible states



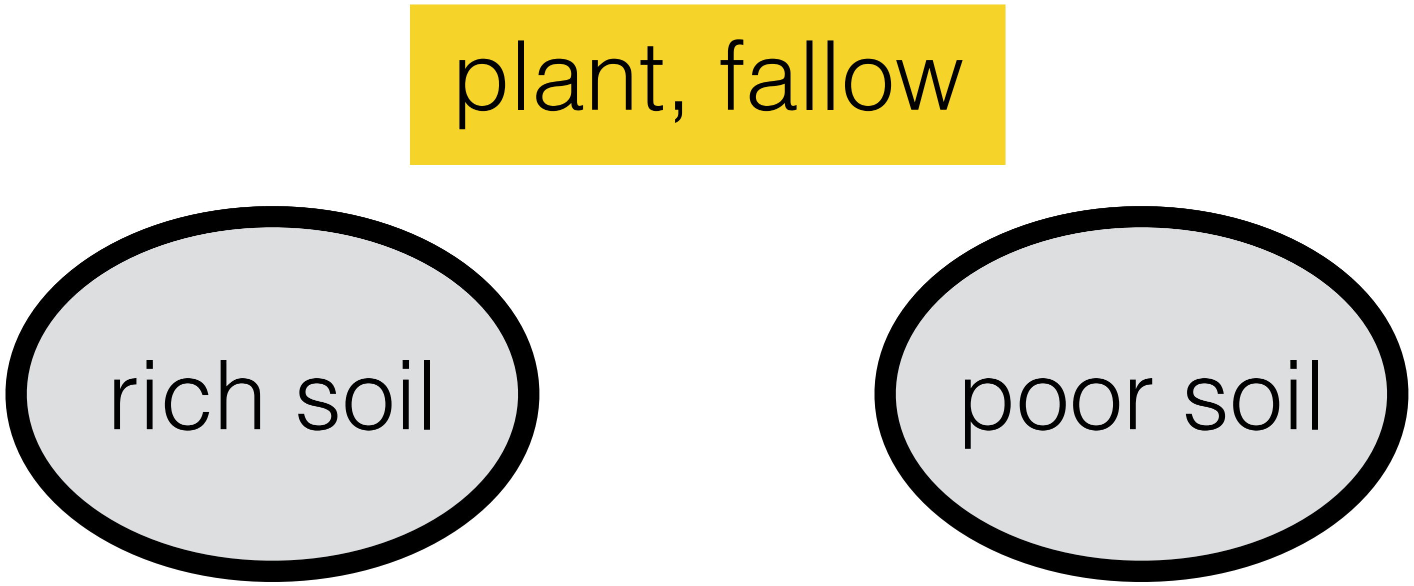
# State Machine

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs



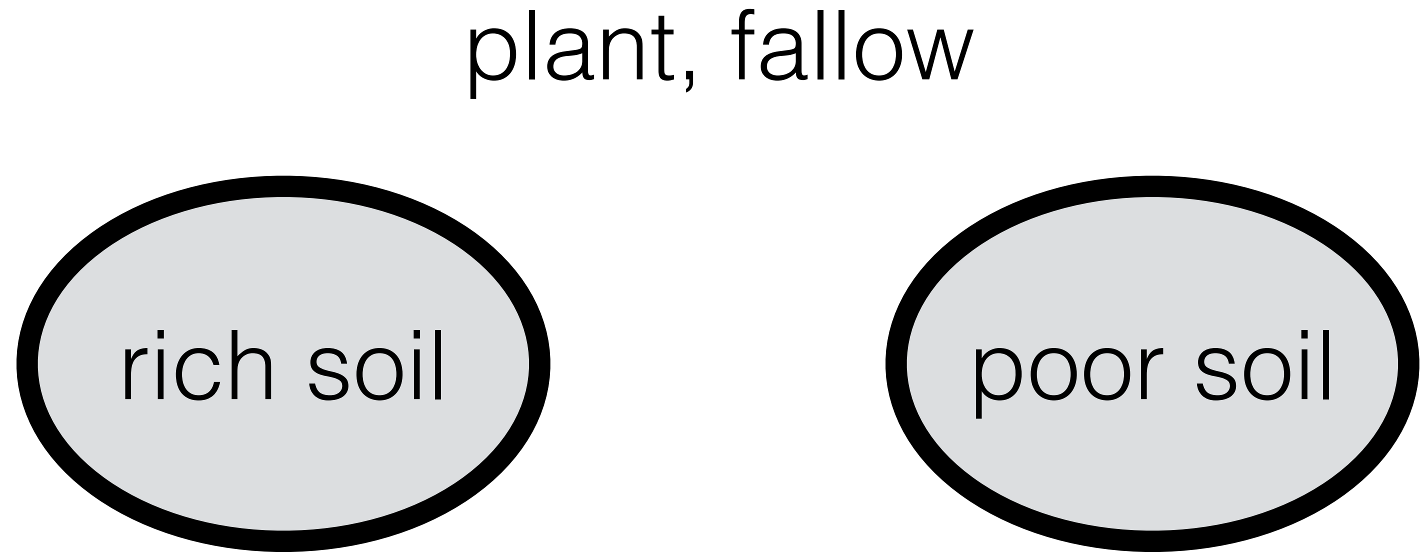
# State Machine

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs



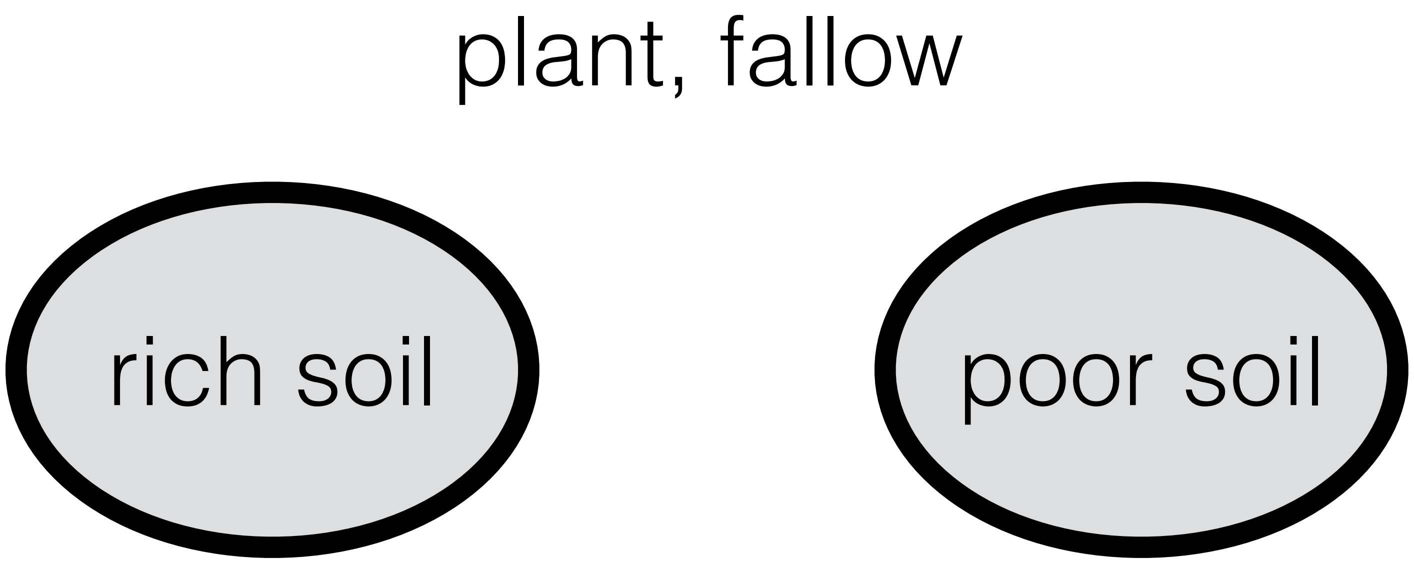
# State Machine

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs



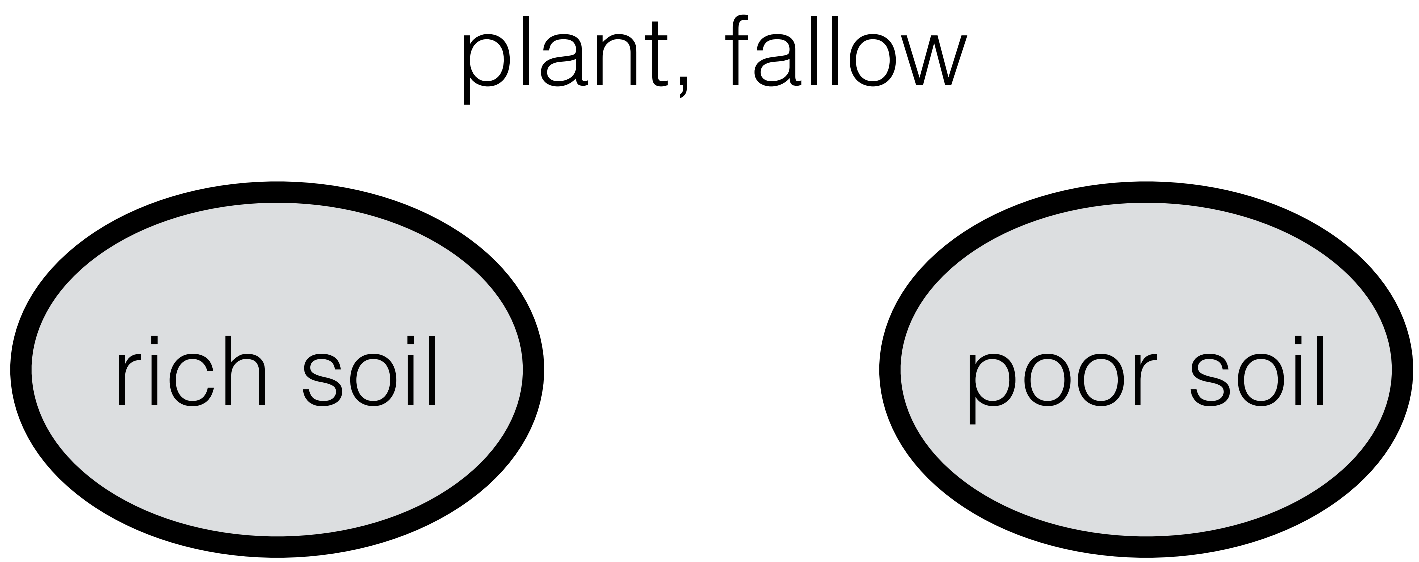
# State Machine

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state



# State Machine

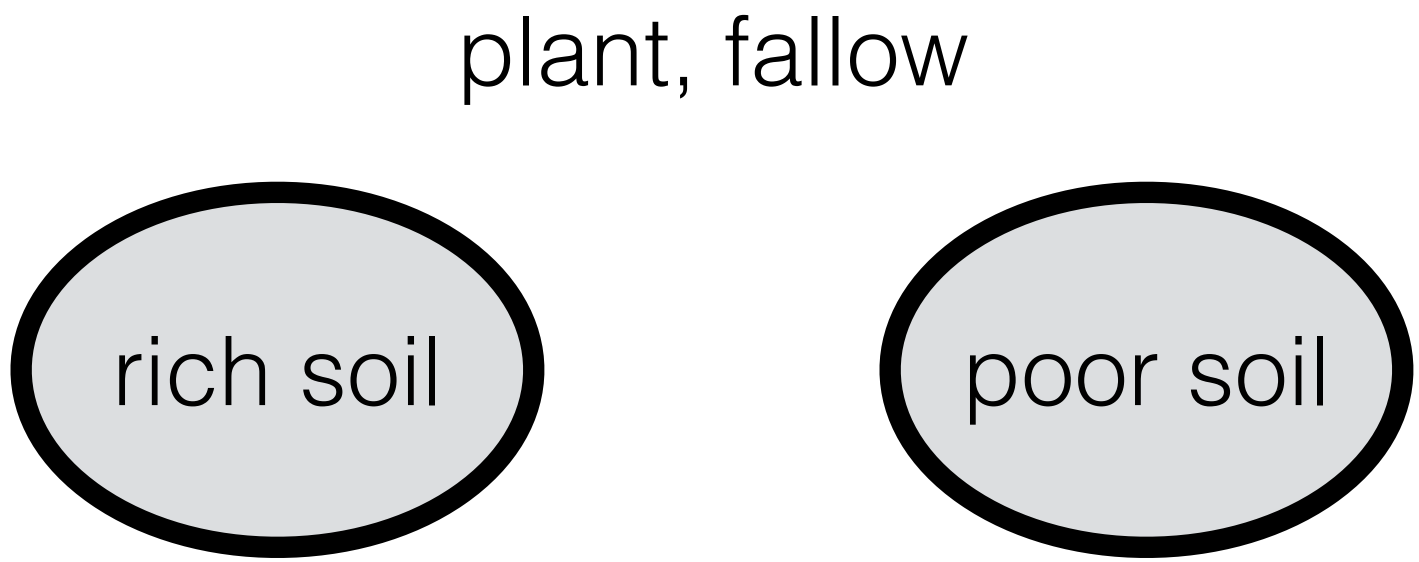
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state




Example

# State Machine

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state

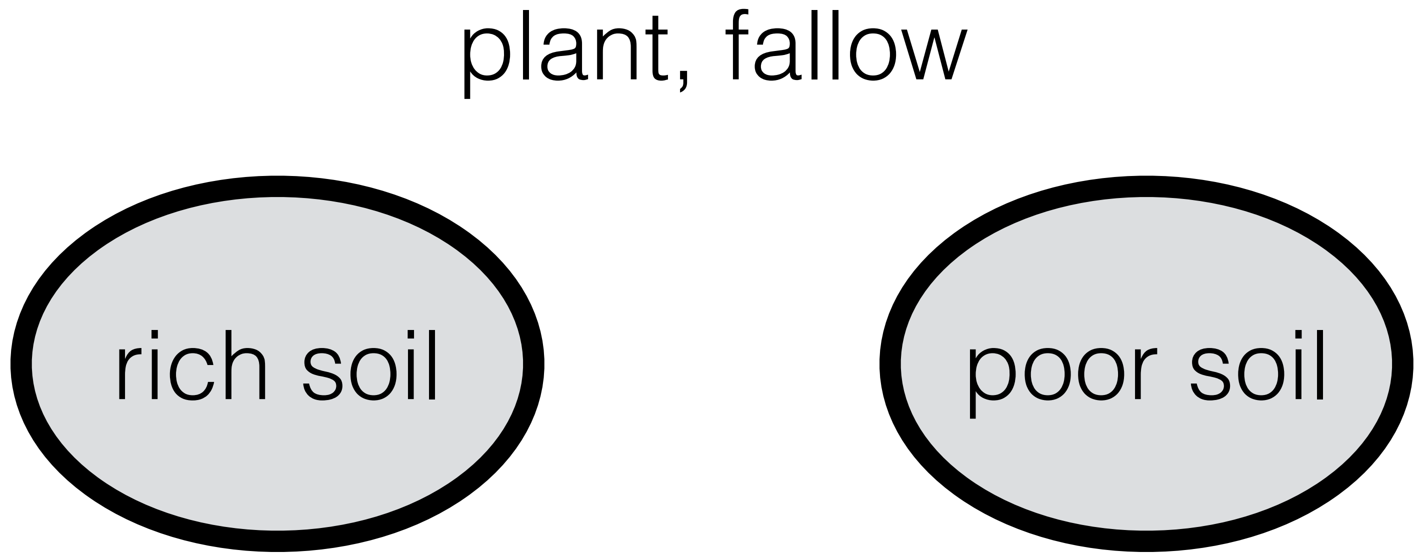



**Example**

$s_0 = \text{rich}$

# 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

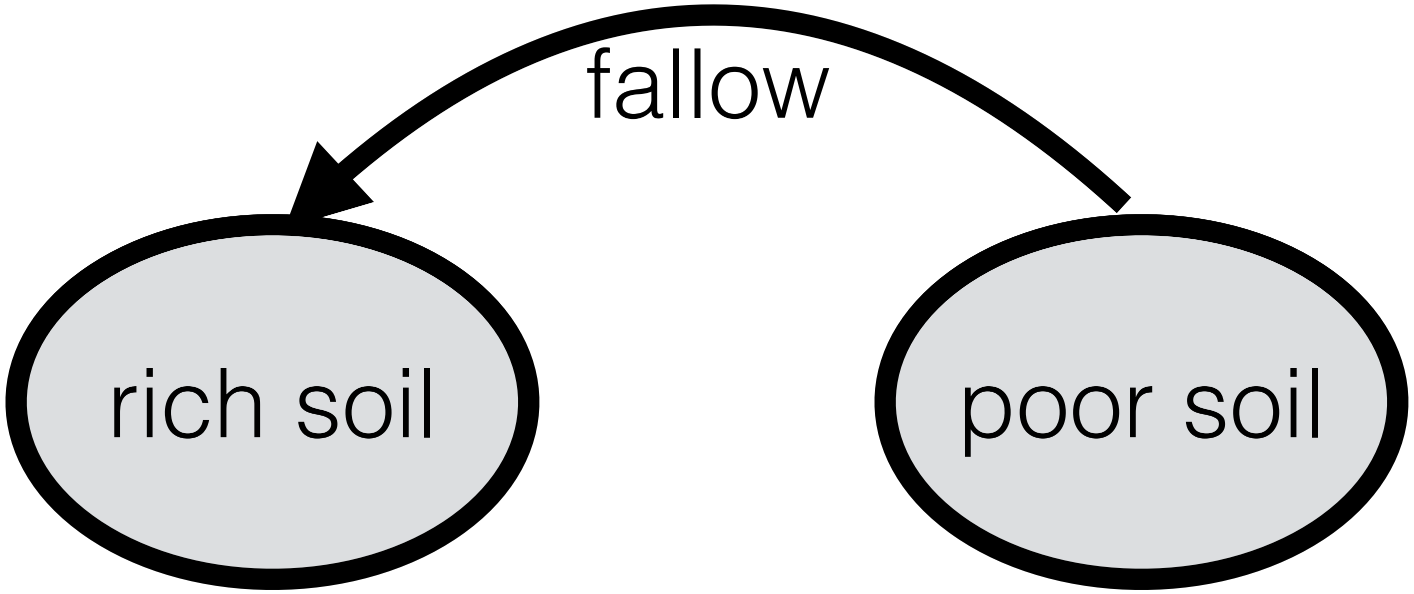



Example

$s_0 = \text{rich}$

# State Machine

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$ : transition function

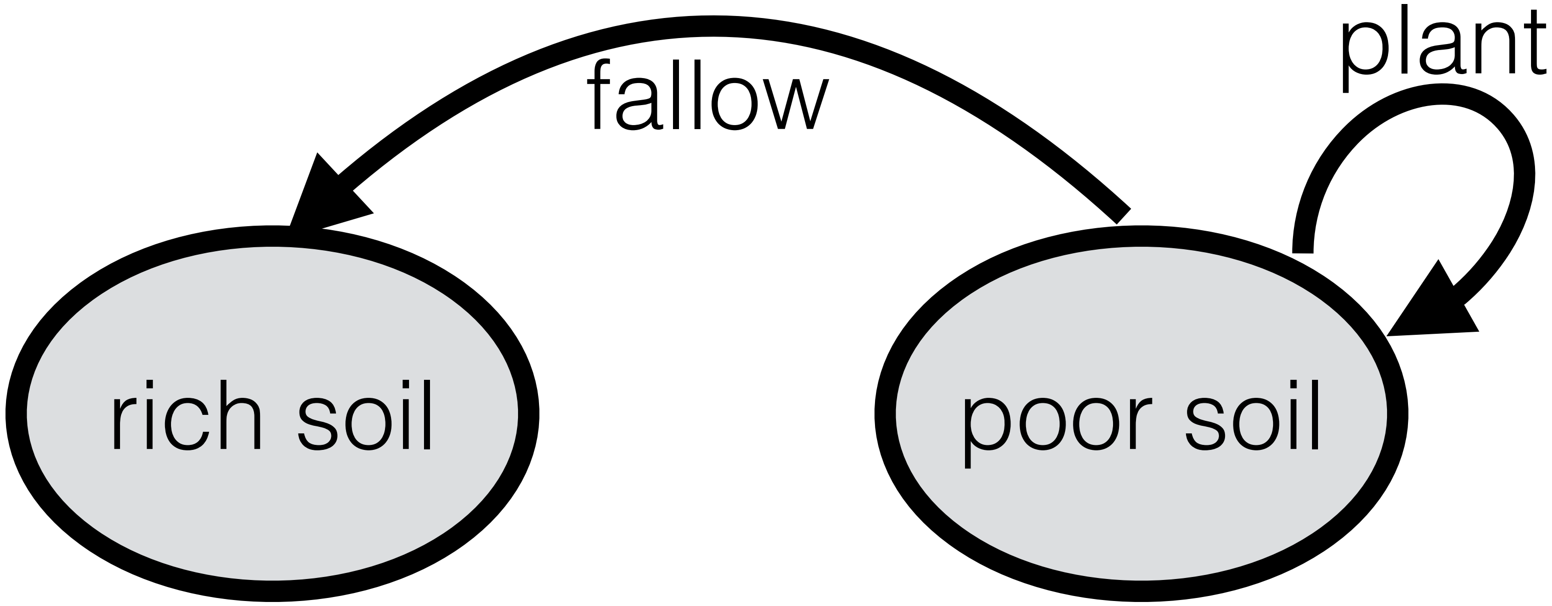



Example

$s_0 = \text{rich}$

# State Machine

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$ : transition function

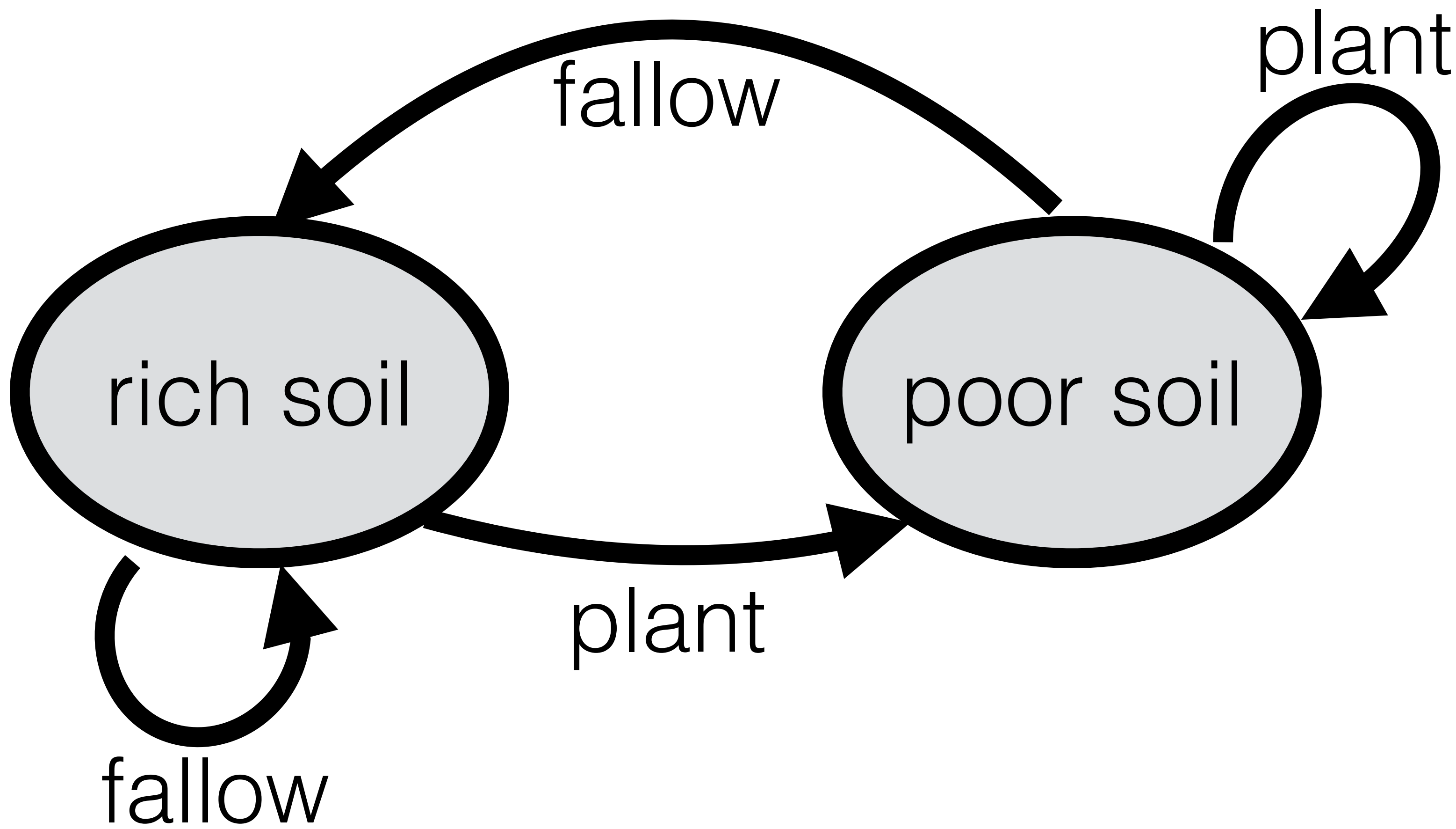



Example

$s_0 = \text{rich}$

# 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

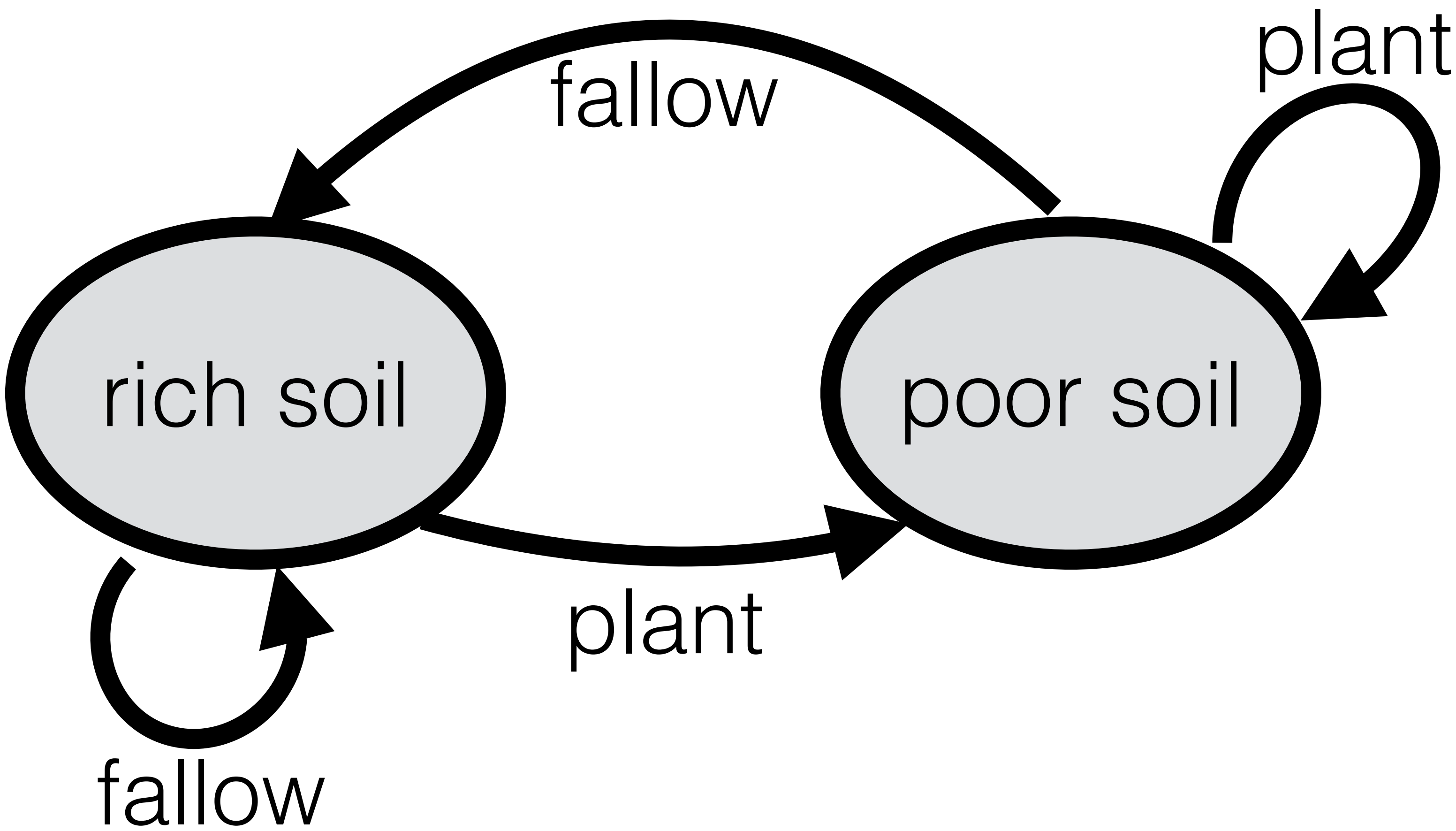



**Example**

$s_0 = \text{rich}$

# 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



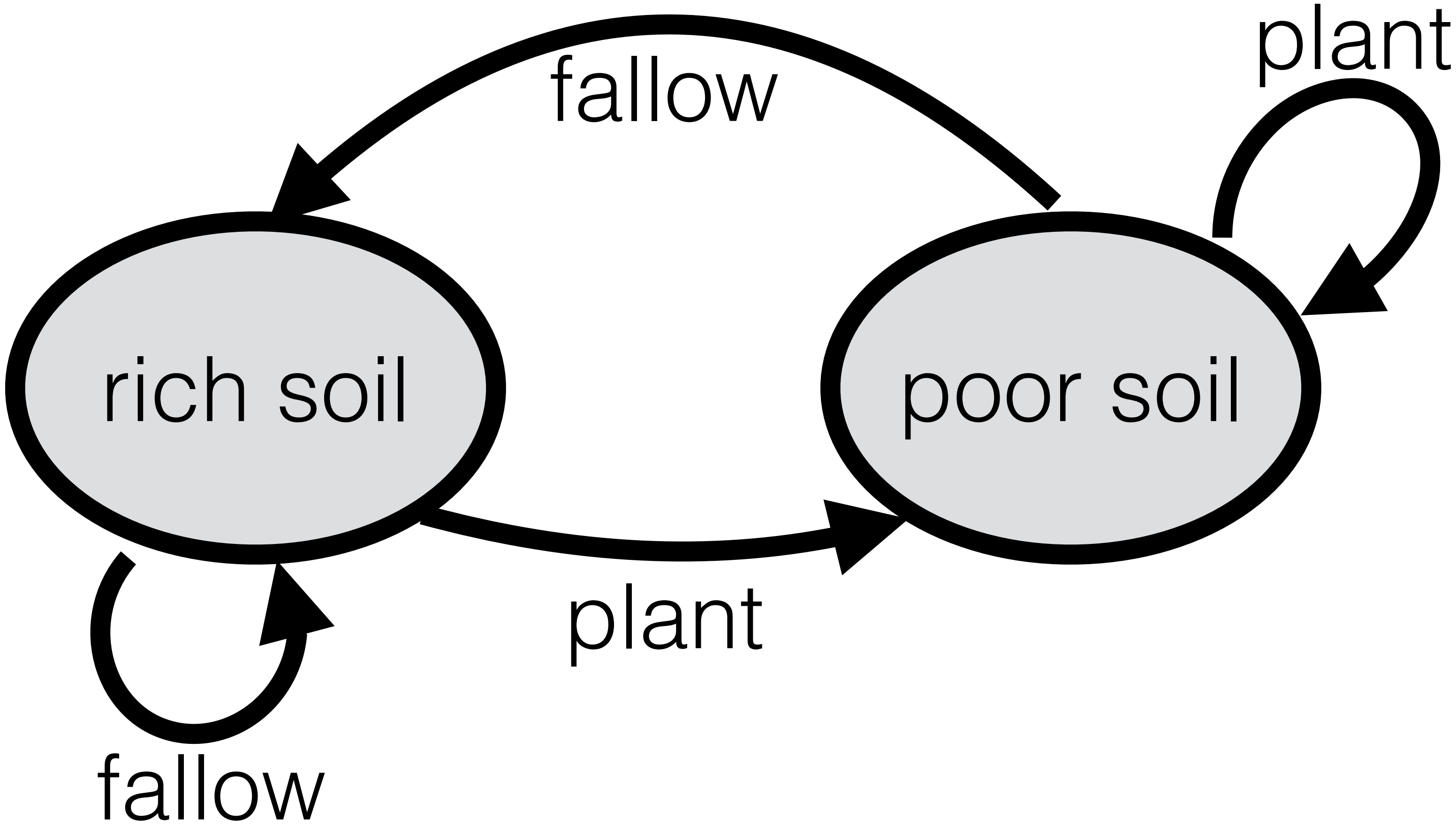

Example

$s_0 = \text{rich}$

$s_1 = f(s_0, \text{plant}) =$

# 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



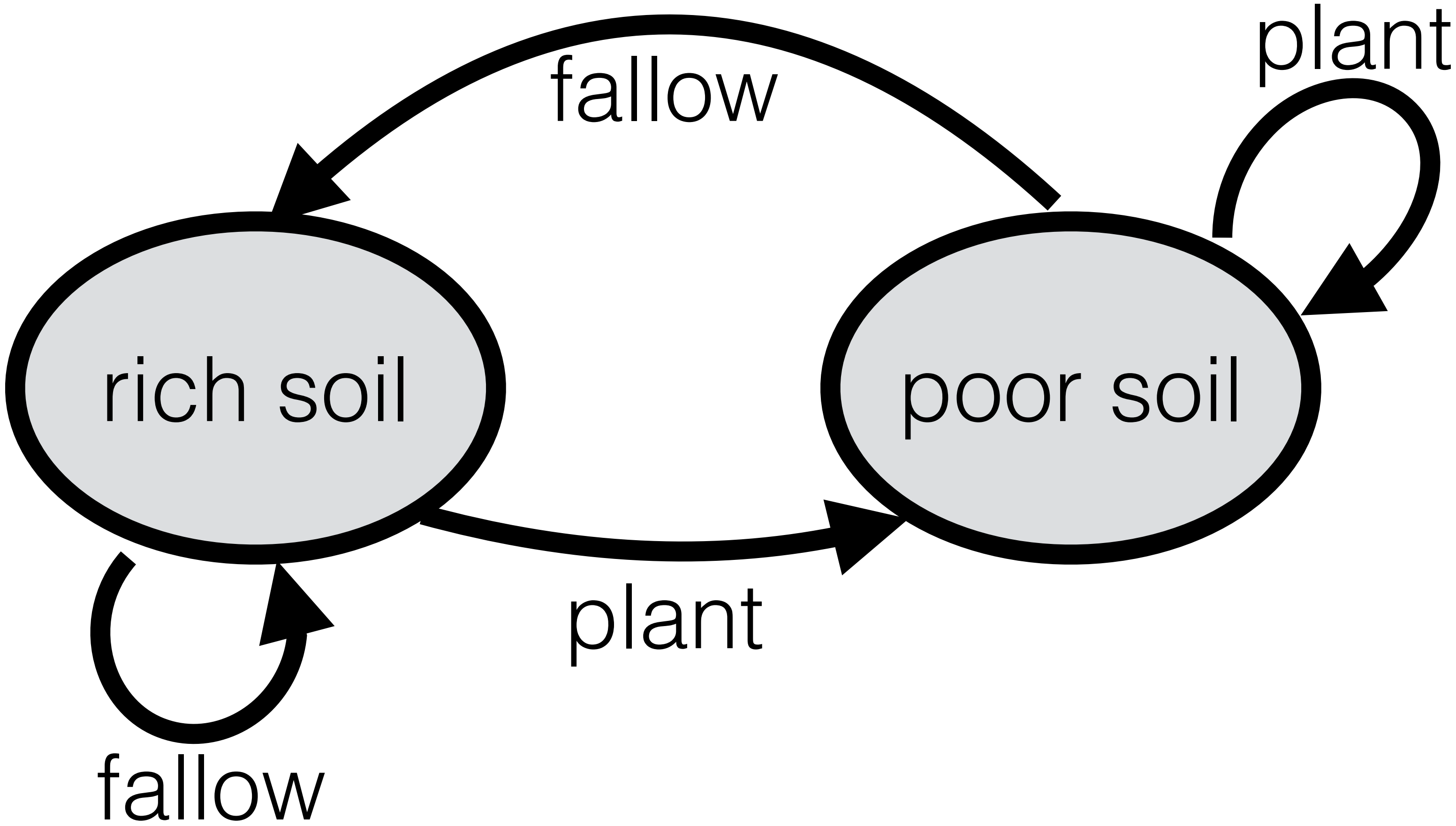

Example

$s_0 = \text{rich}$

$s_1 = f(s_0, \text{plant}) =$

# 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



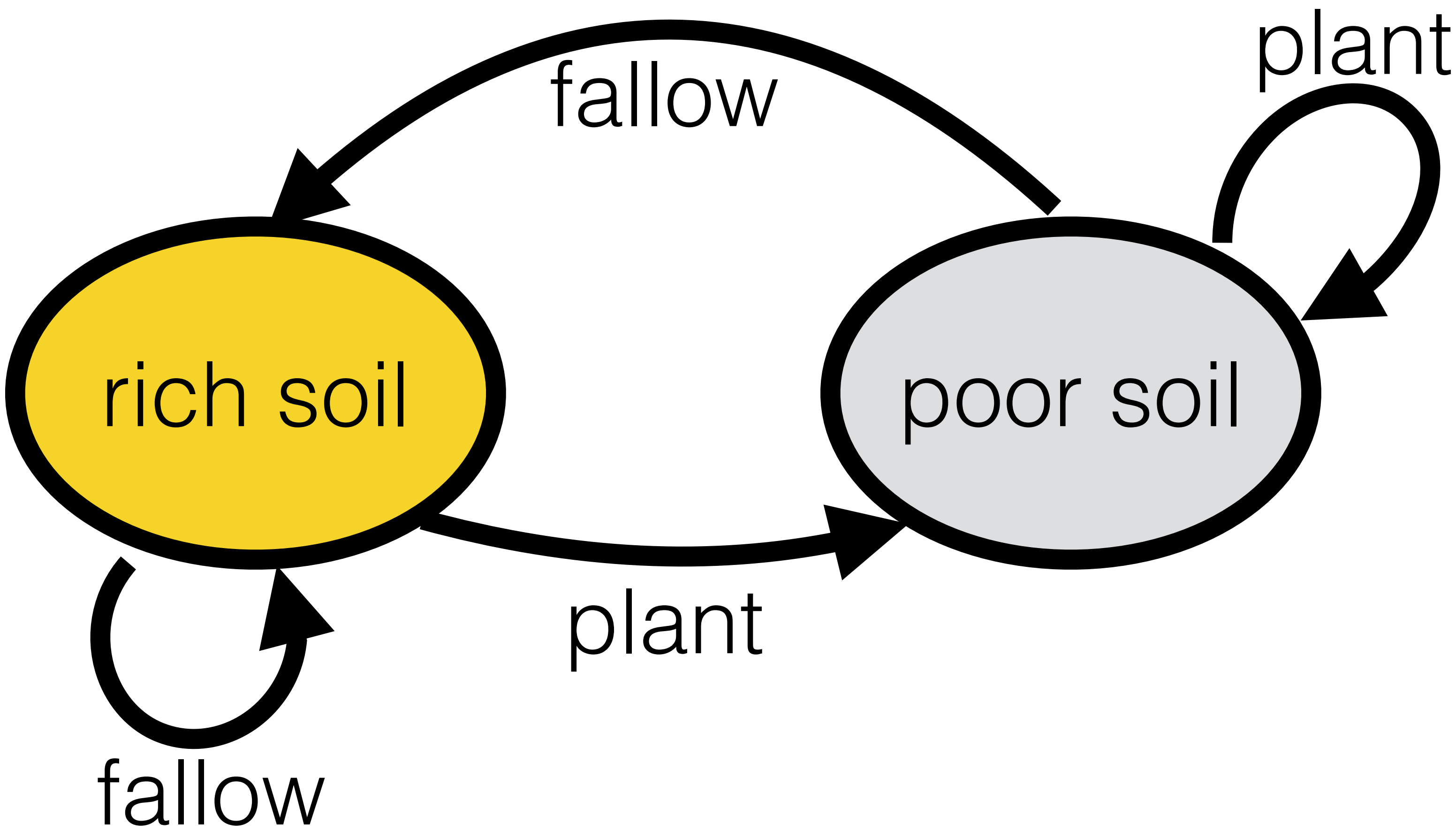

Example

$s_0 = \text{rich}$

$s_1 = f(s_0, \text{plant}) =$

# 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



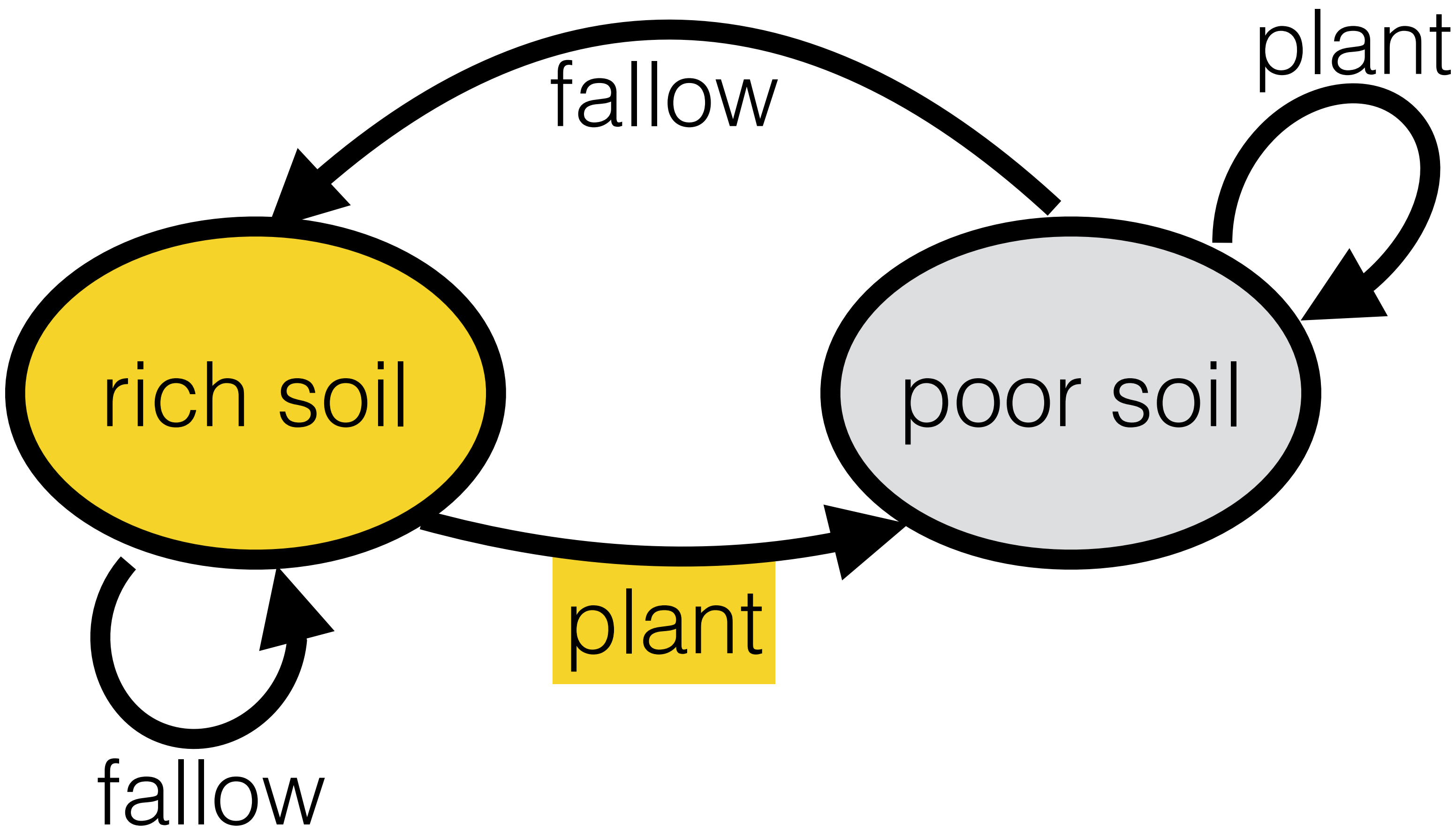

Example

$s_0 = \text{rich}$

$s_1 = f(s_0, \text{plant}) =$

# State Machine

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$ : transition function



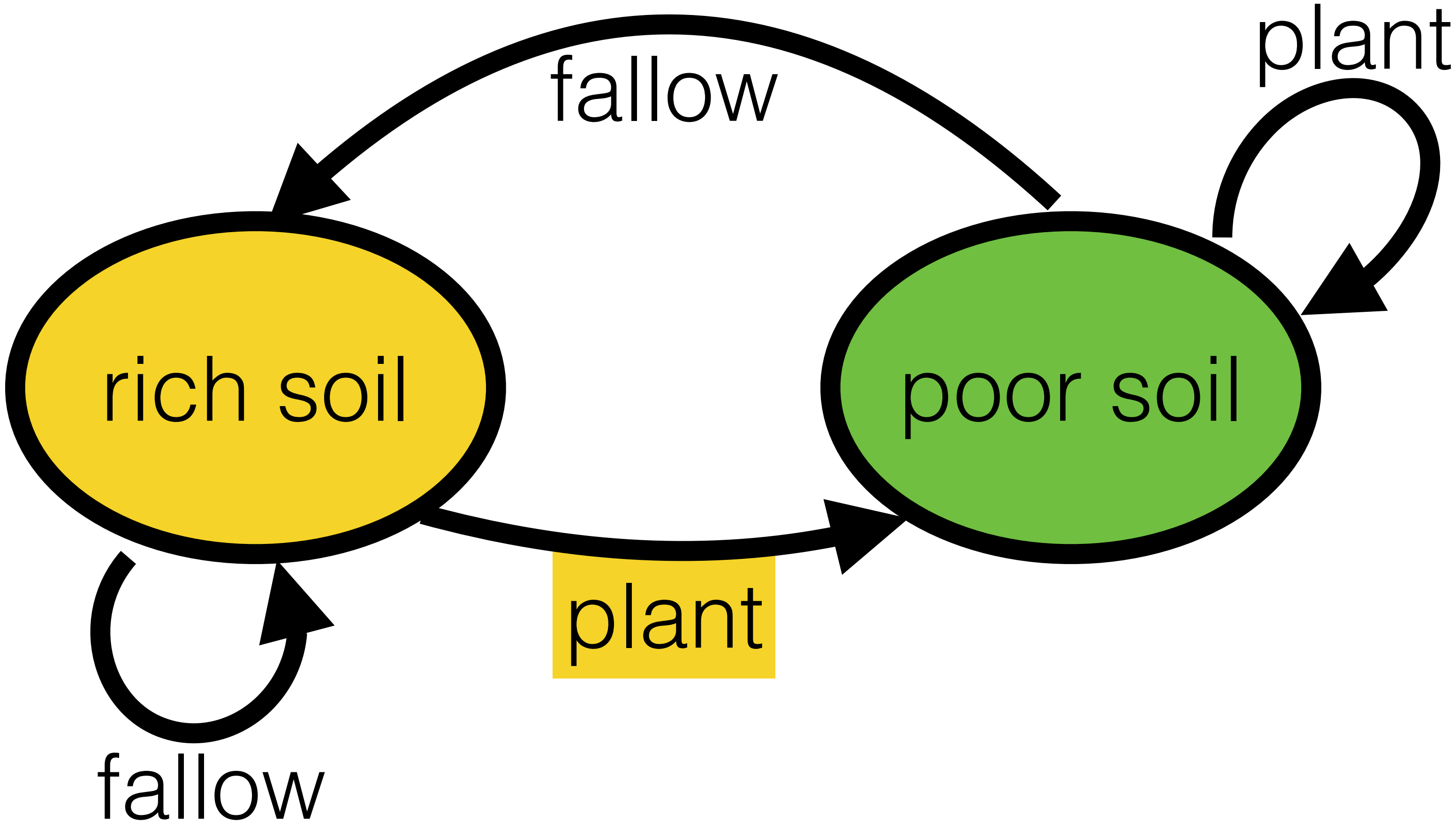

Example

$s_0 = \text{rich}$

$s_1 = f(s_0, \text{plant}) =$

# State Machine

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$ : transition function



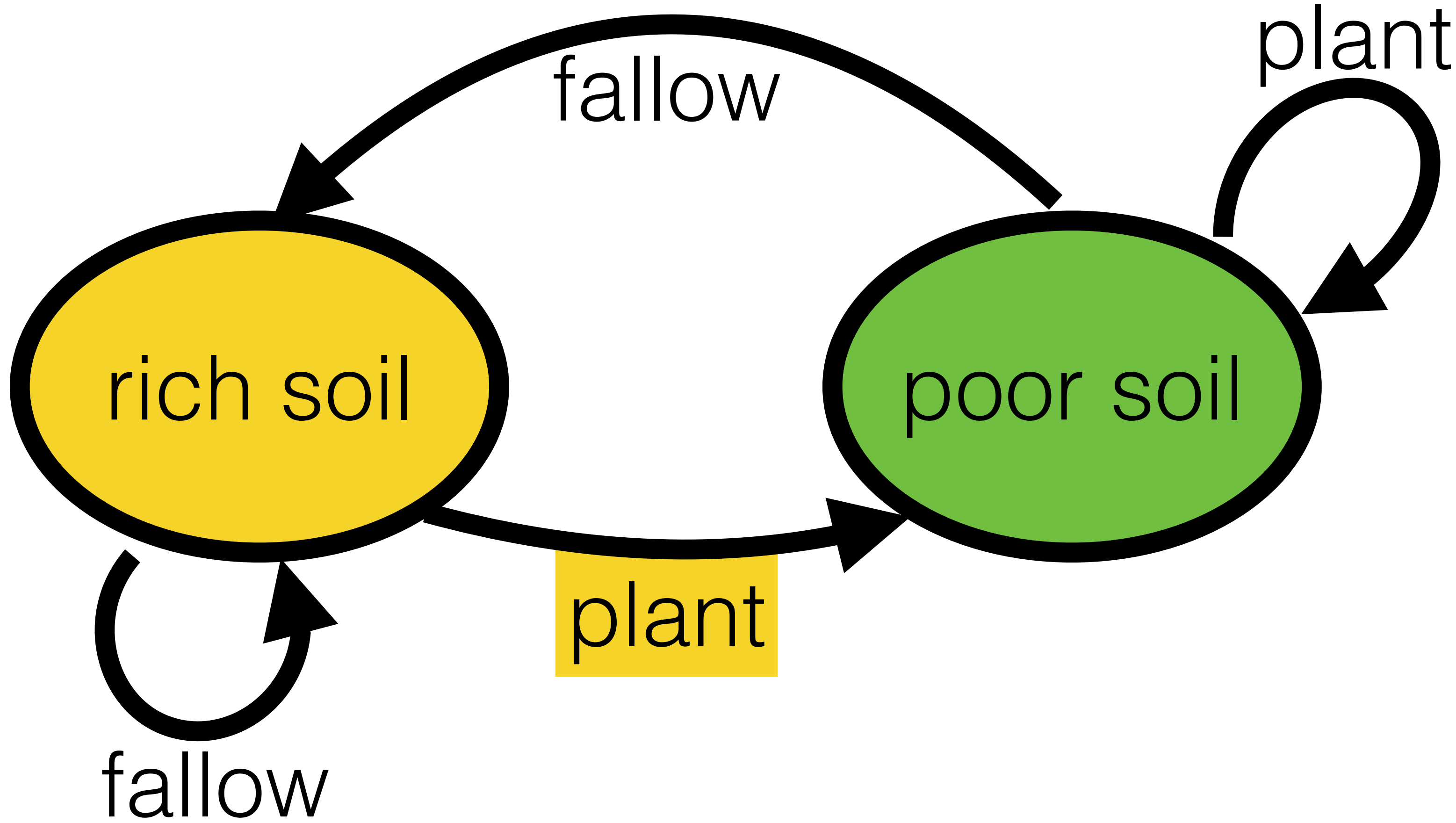

Example

$s_0 = \text{rich}$

$s_1 = f(s_0, \text{plant}) =$

# 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

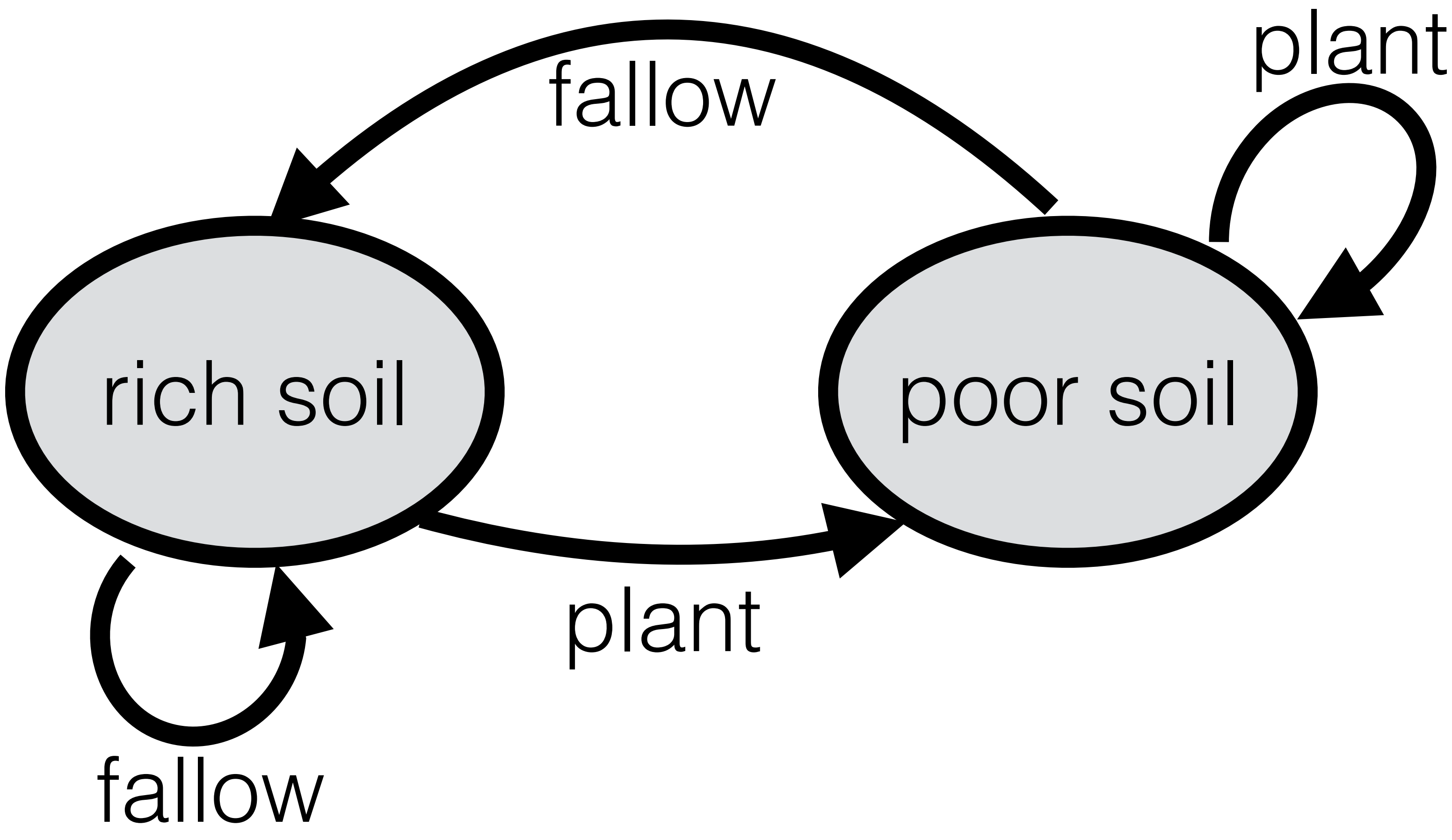



Example

$$s_0 = \text{rich}$$
$$s_1 = f(s_0, \text{plant}) = \text{poor}$$

# State Machine

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$ : transition function

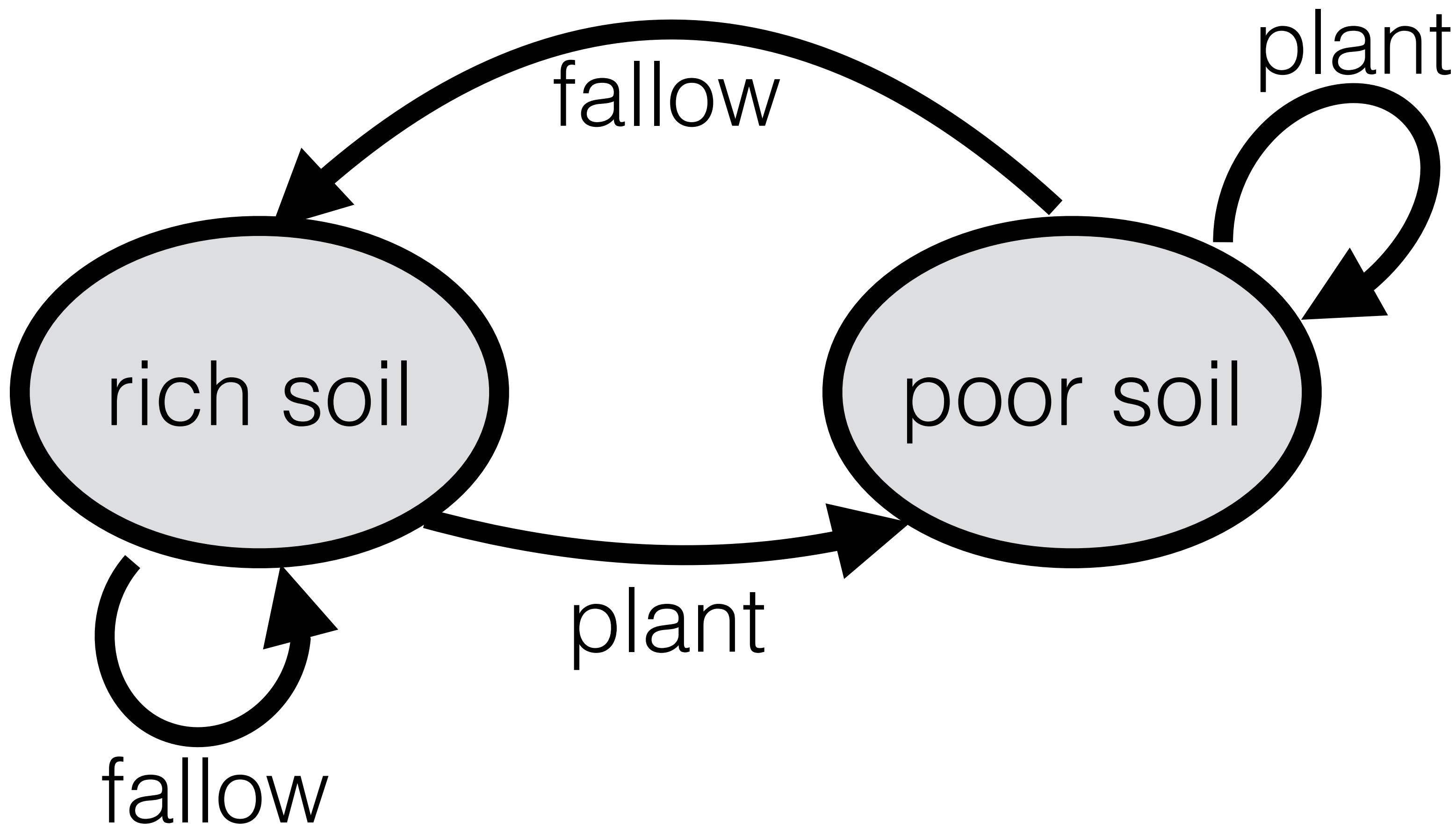



**Example**

$$s_0 = \text{rich}$$
$$s_1 = f(s_0, \text{plant}) = \text{poor}$$

# State Machine

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$ : transition function
- $\mathcal{Y}$ : set of possible outputs

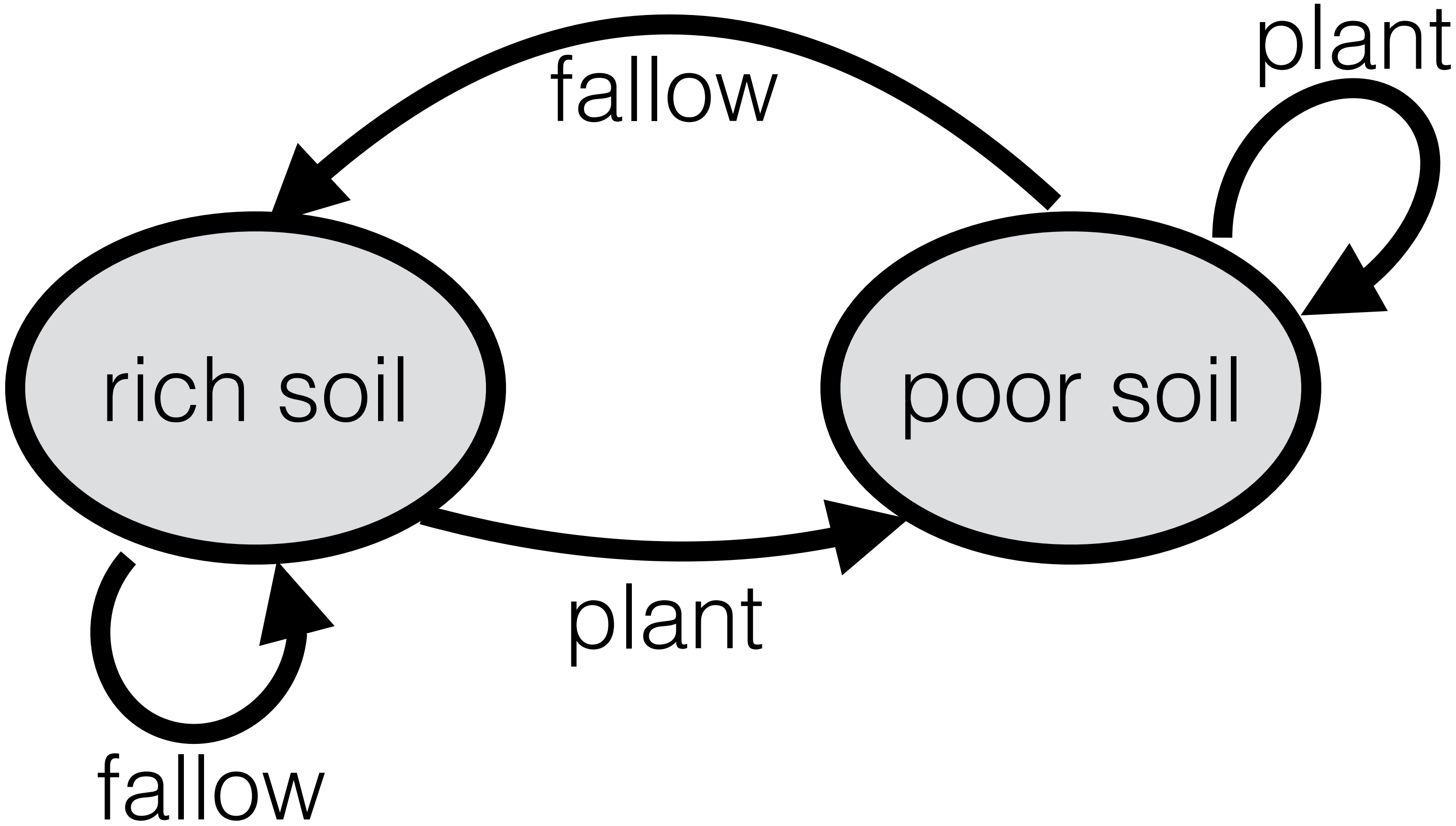



Example

$$s_0 = \text{rich}$$
$$s_1 = f(s_0, \text{plant}) = \text{poor}$$

# 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

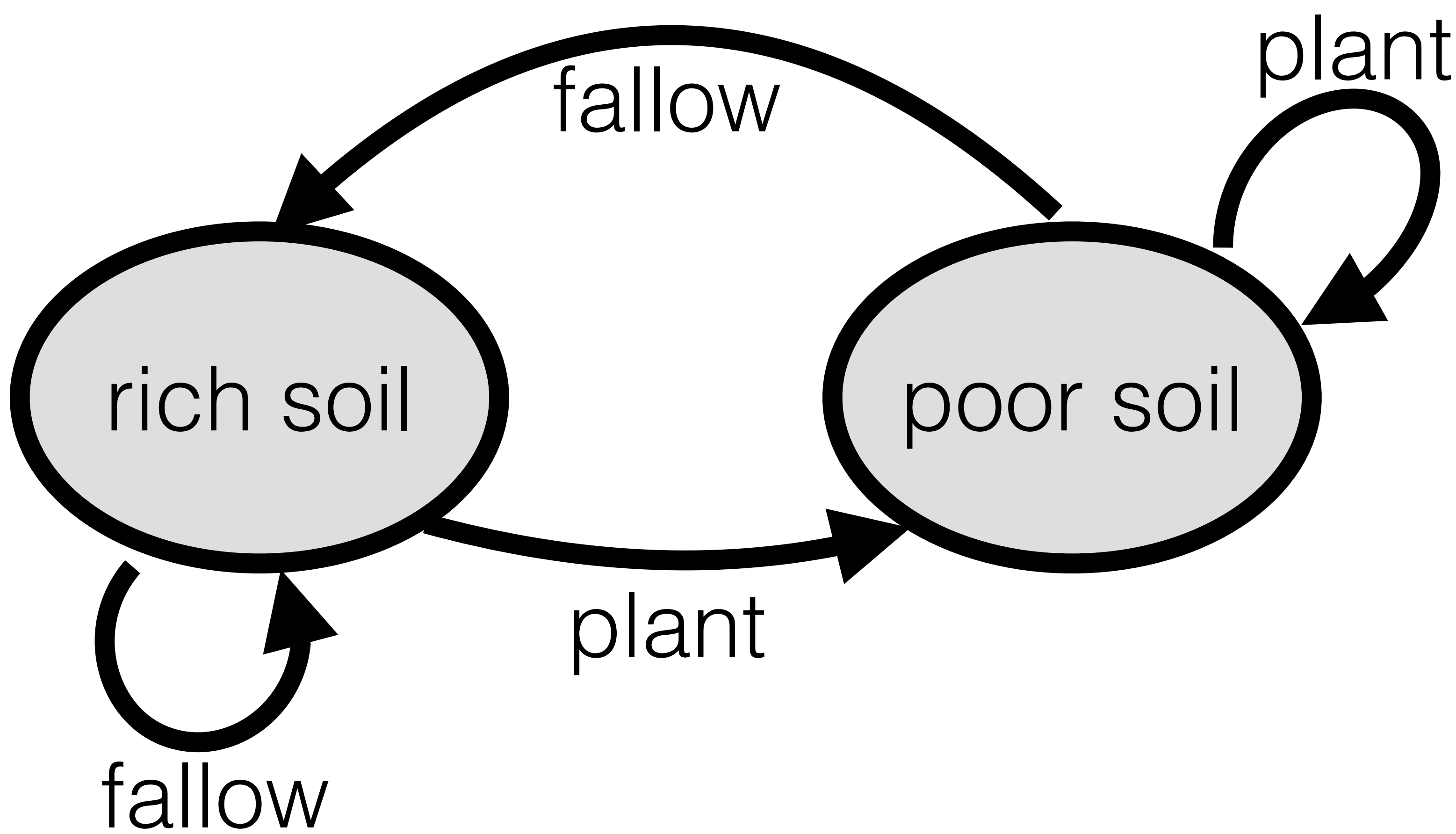



Example

$s_0 = \text{rich}$
$s_1 = f(s_0, \text{plant}) = \text{poor}$

# 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

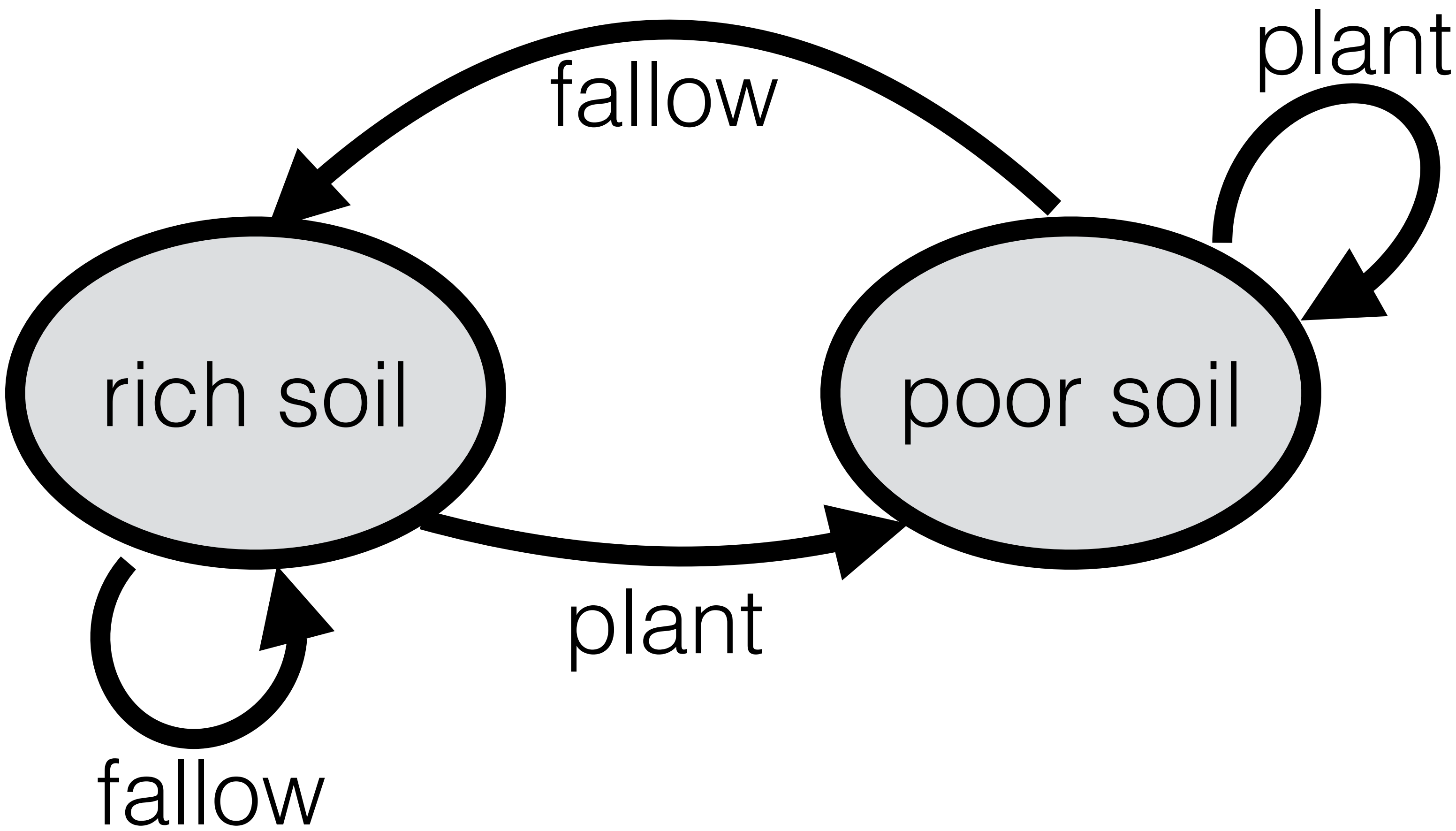



**Example**

$$s_0 = \text{rich}$$
$$s_1 = f(s_0, \text{plant}) = \text{poor}$$

# 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)

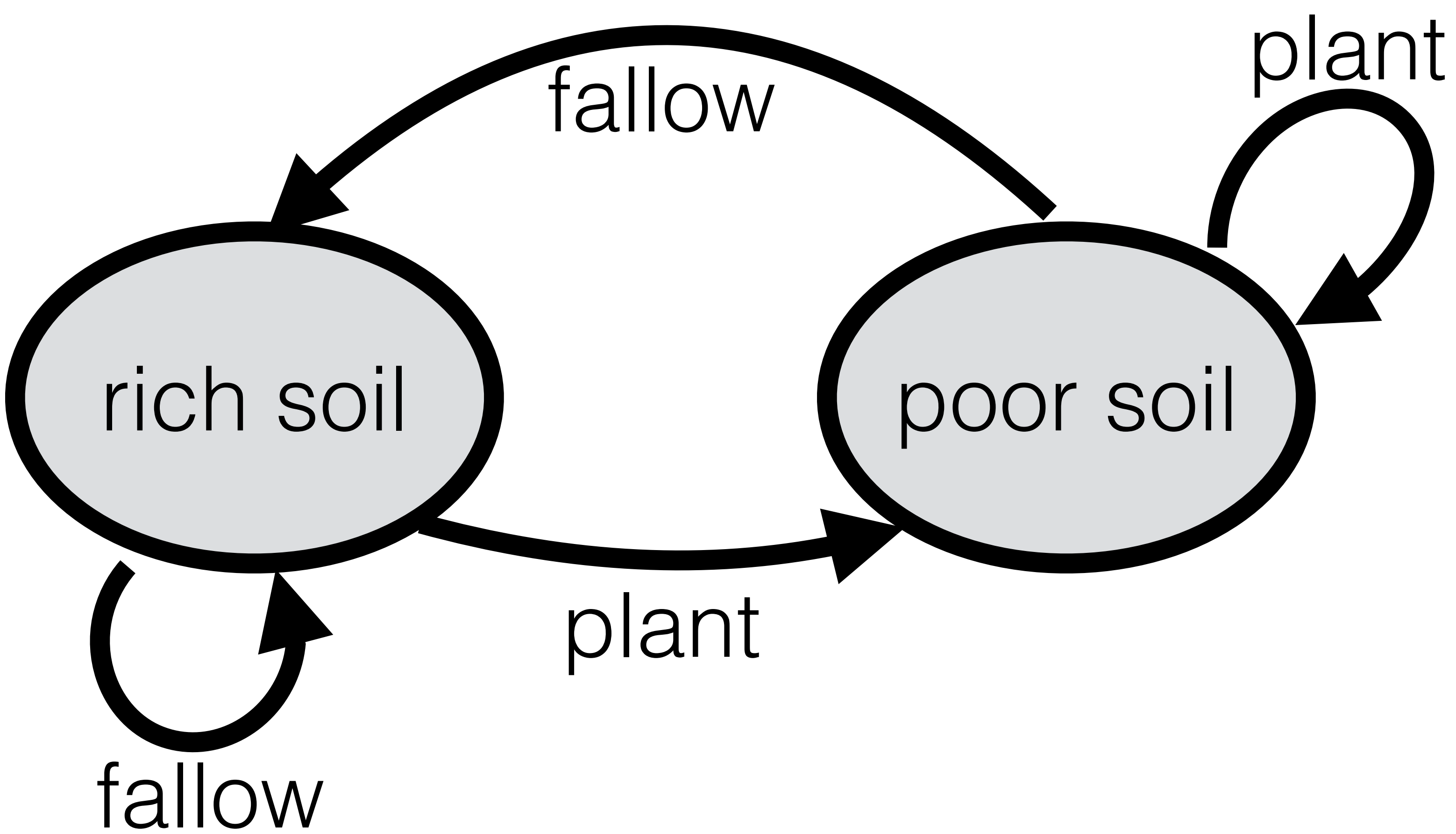



**Example**

$$s_0 = \text{rich}$$
$$s_1 = f(s_0, \text{plant}) = \text{poor}$$

# 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**

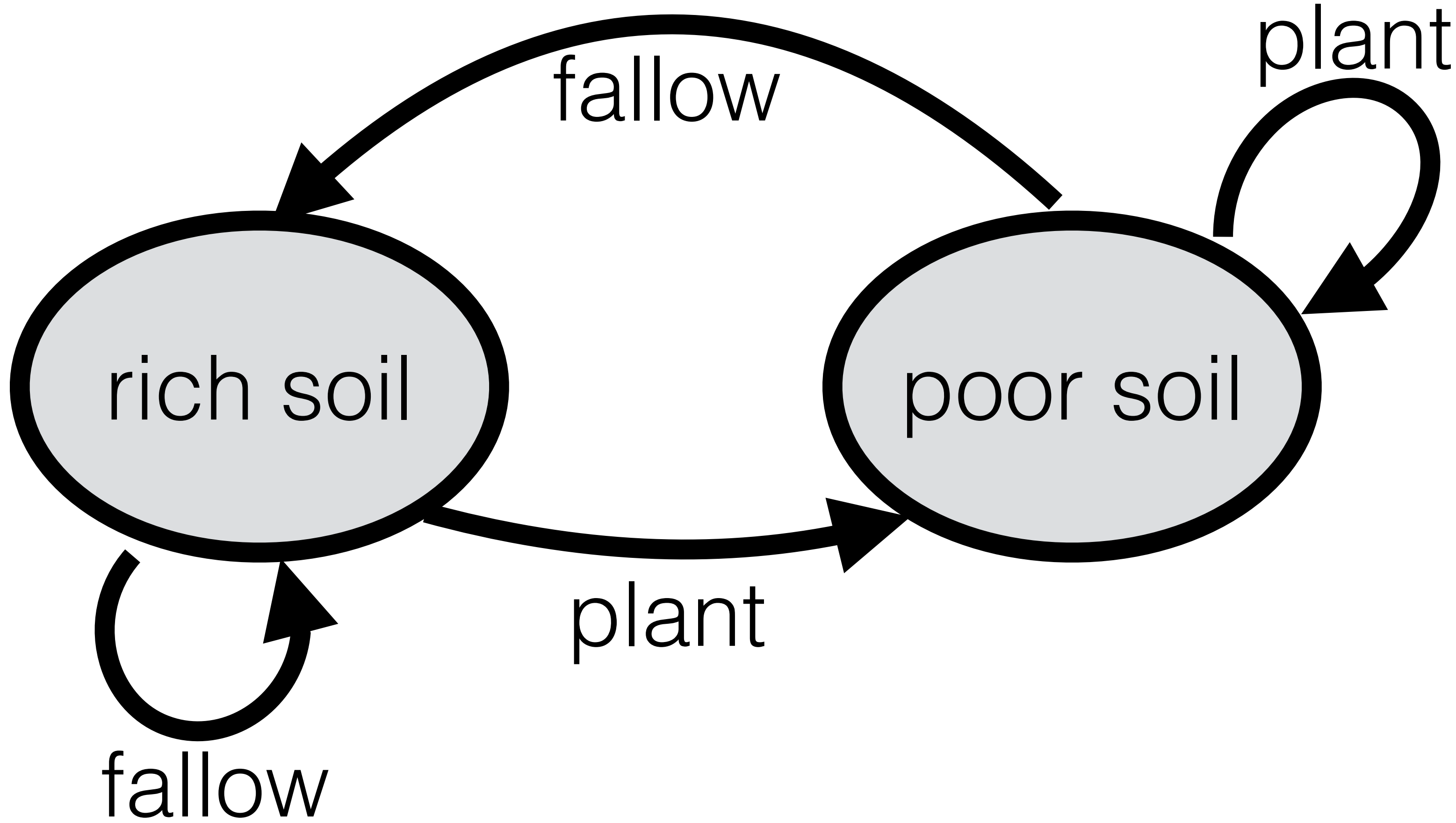
$s_0 = \text{rich}$

$s_1 = f(s_0, \text{plant}) = \text{poor};$

$y_1 = g(s_1) = \text{poor}$

# 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**

$s_0 = \text{rich}$

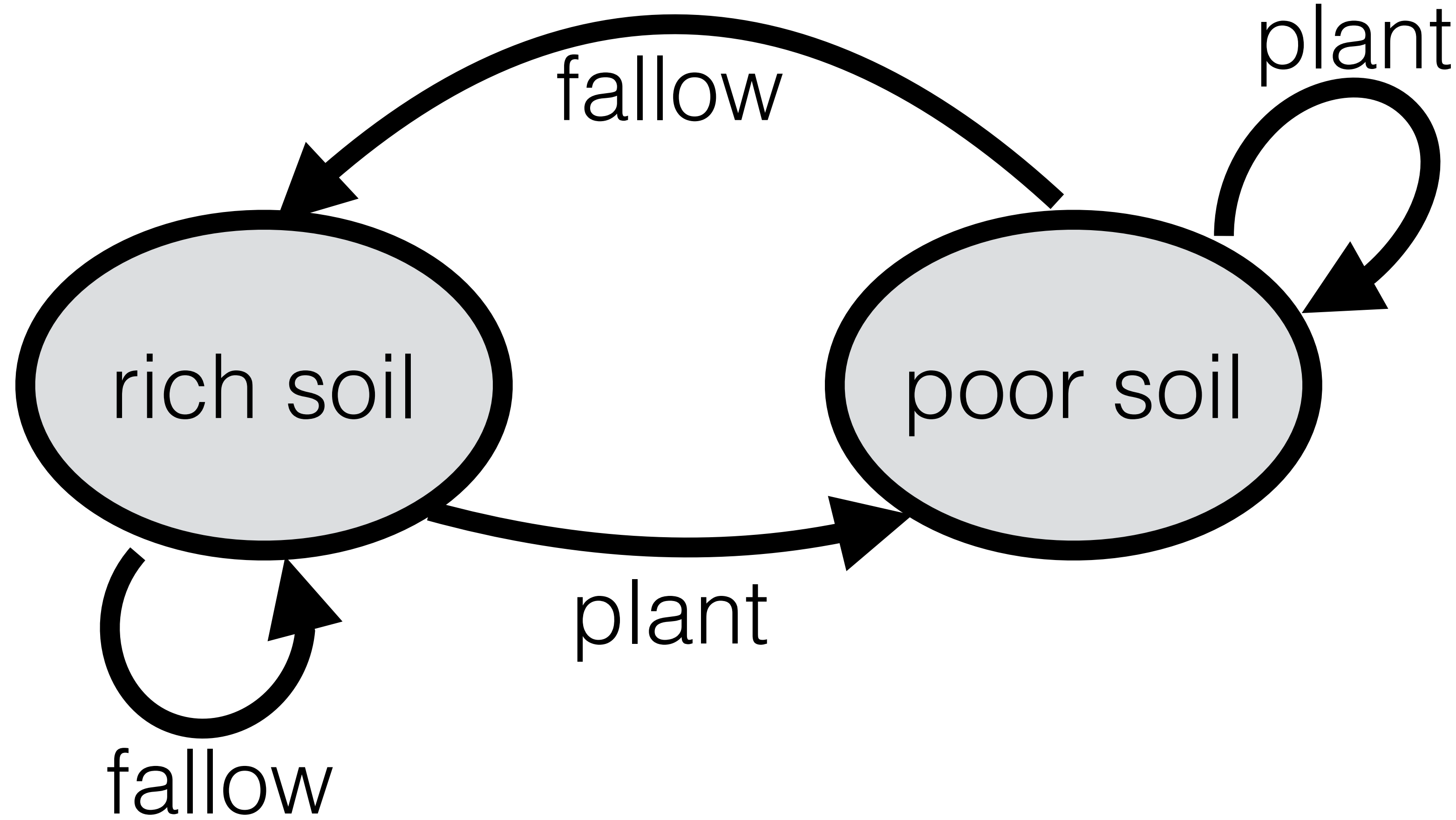
$s_1 = f(s_0, \text{plant}) = \text{poor};$

$y_1 = g(s_1) = \text{poor}$

$s_2 = f(s_1, \text{fallow}) = \text{rich}$

# State Machine

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

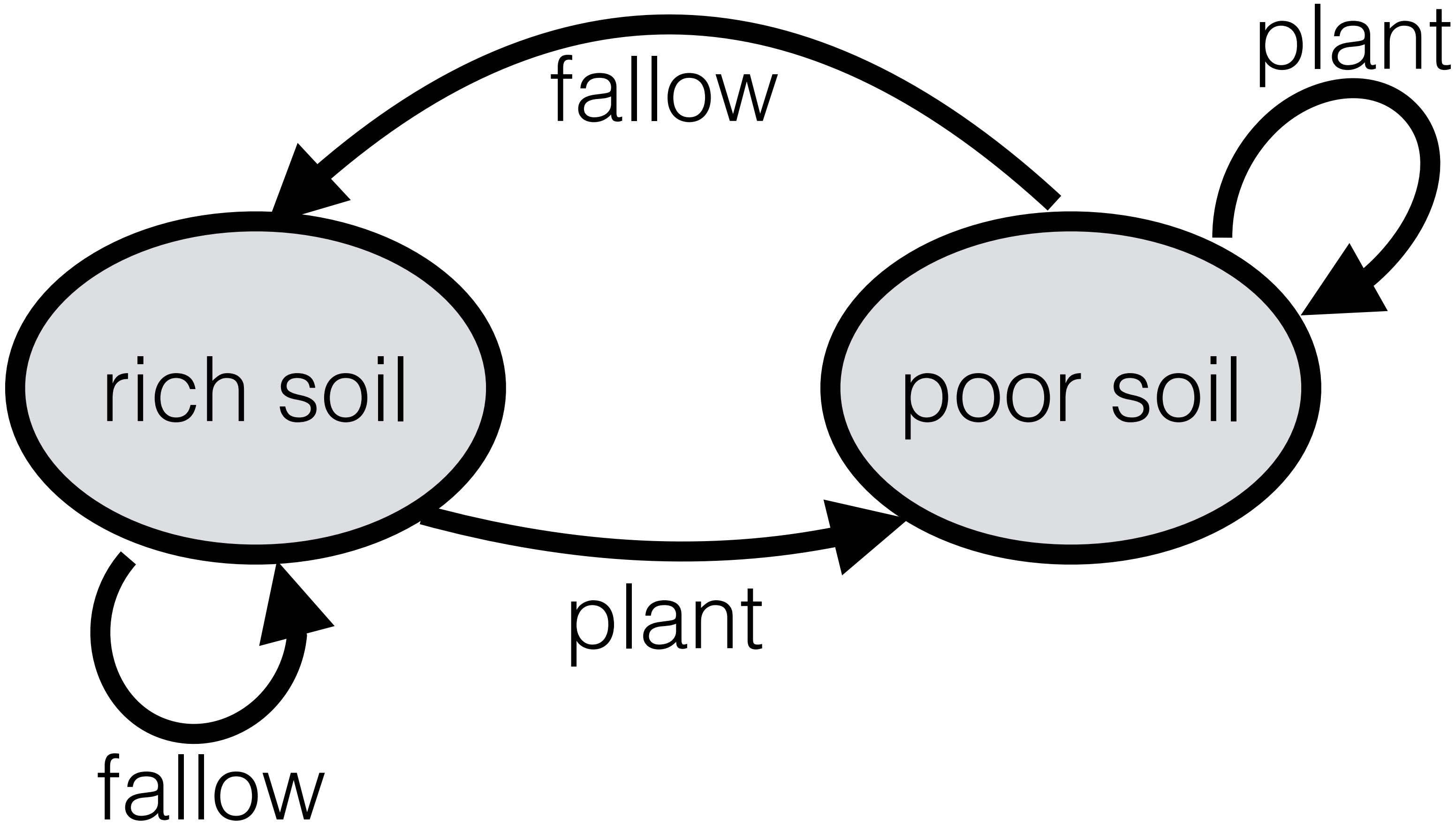



**Example**

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

# 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)

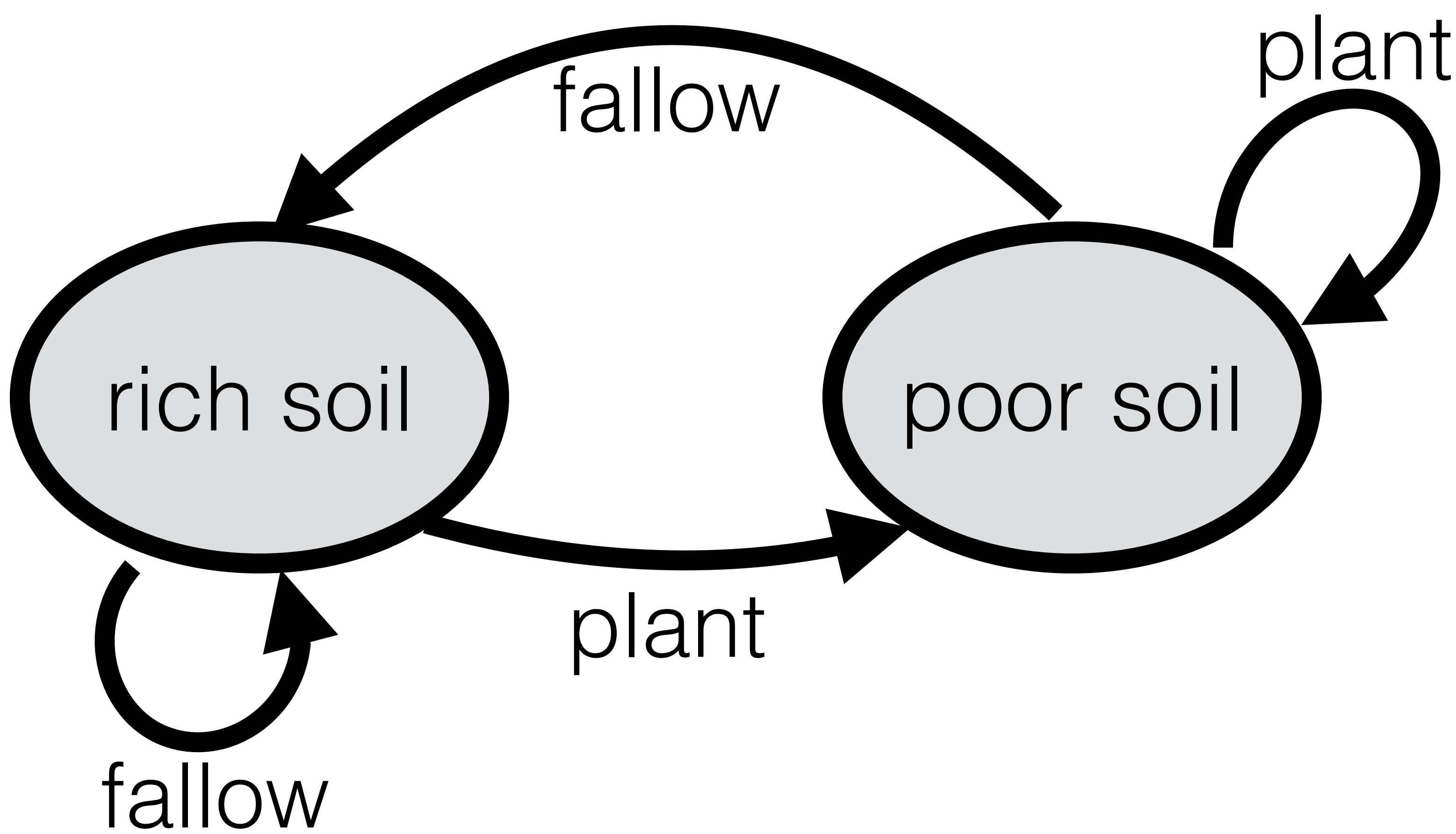



Example

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

# 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)

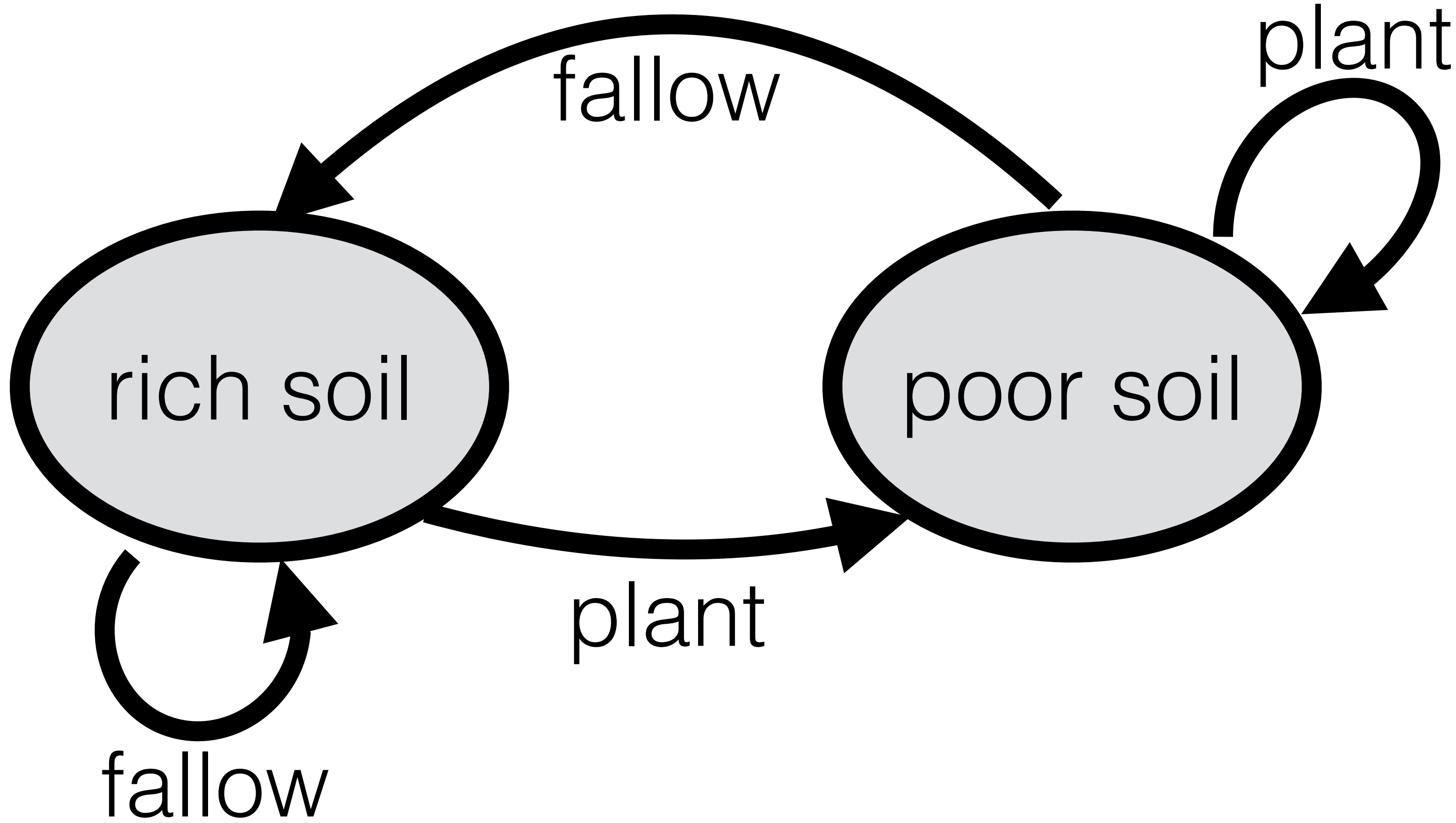



**Example**

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

# State Machine

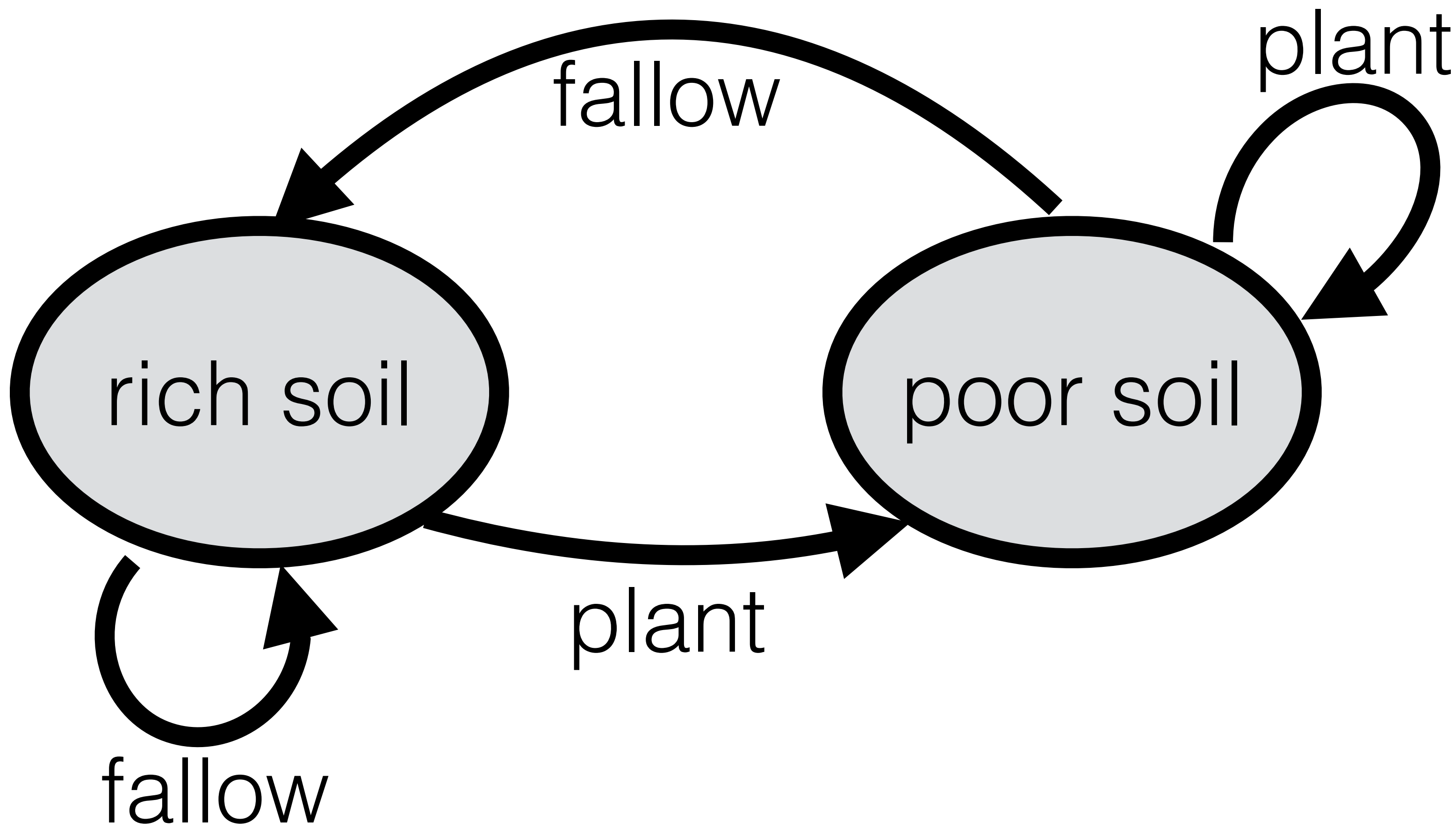
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$ : transition function
- $\mathcal{Y}$ : set of possible outputs
- $g : \mathcal{S} \rightarrow \mathcal{Y}$ : output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




**Example**

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

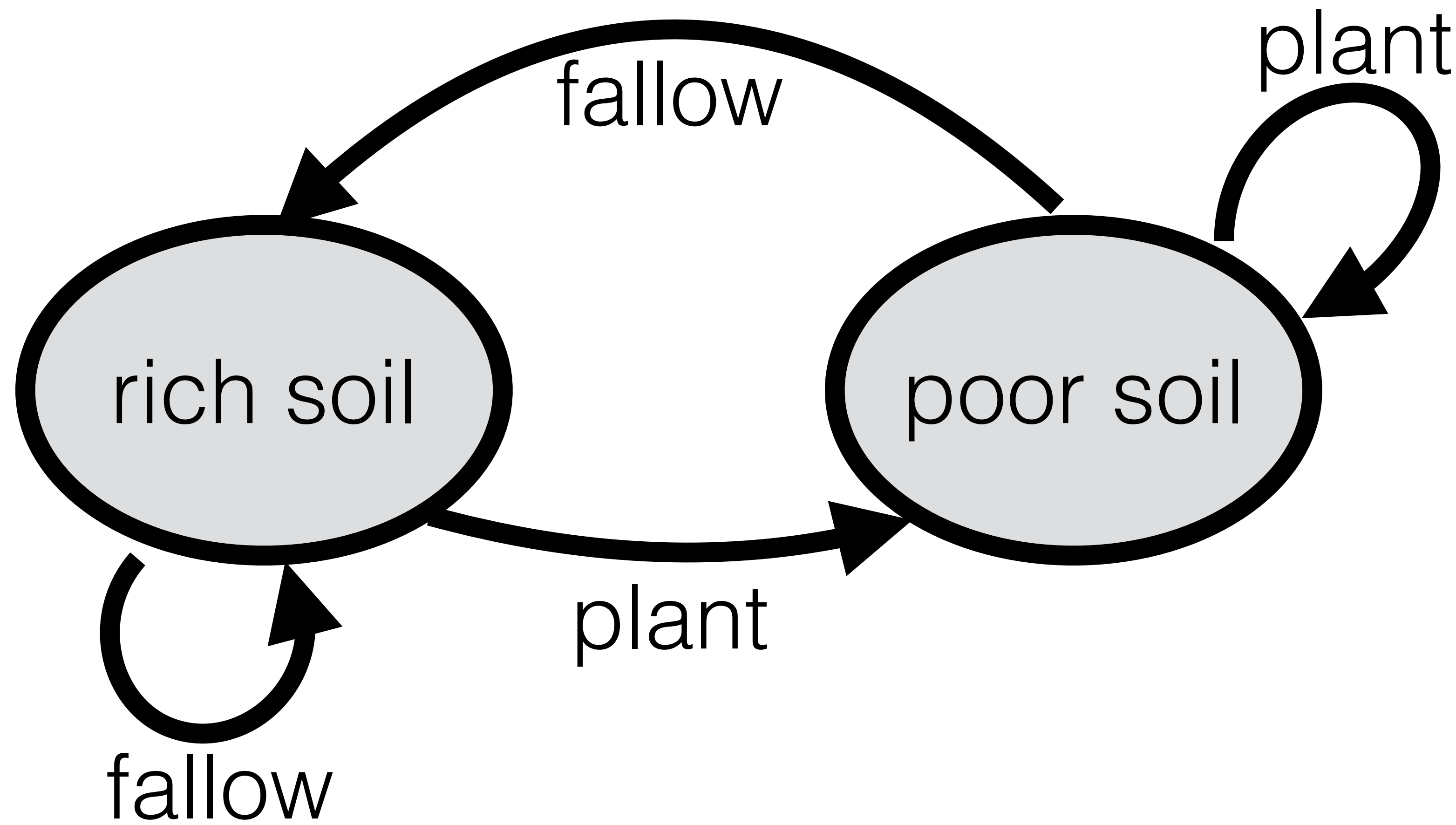
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ : transition function
- $\mathcal{Y}$ : set of possible outputs
- $g : \mathcal{S} \to \mathcal{Y}$ : output function
  - e.g.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



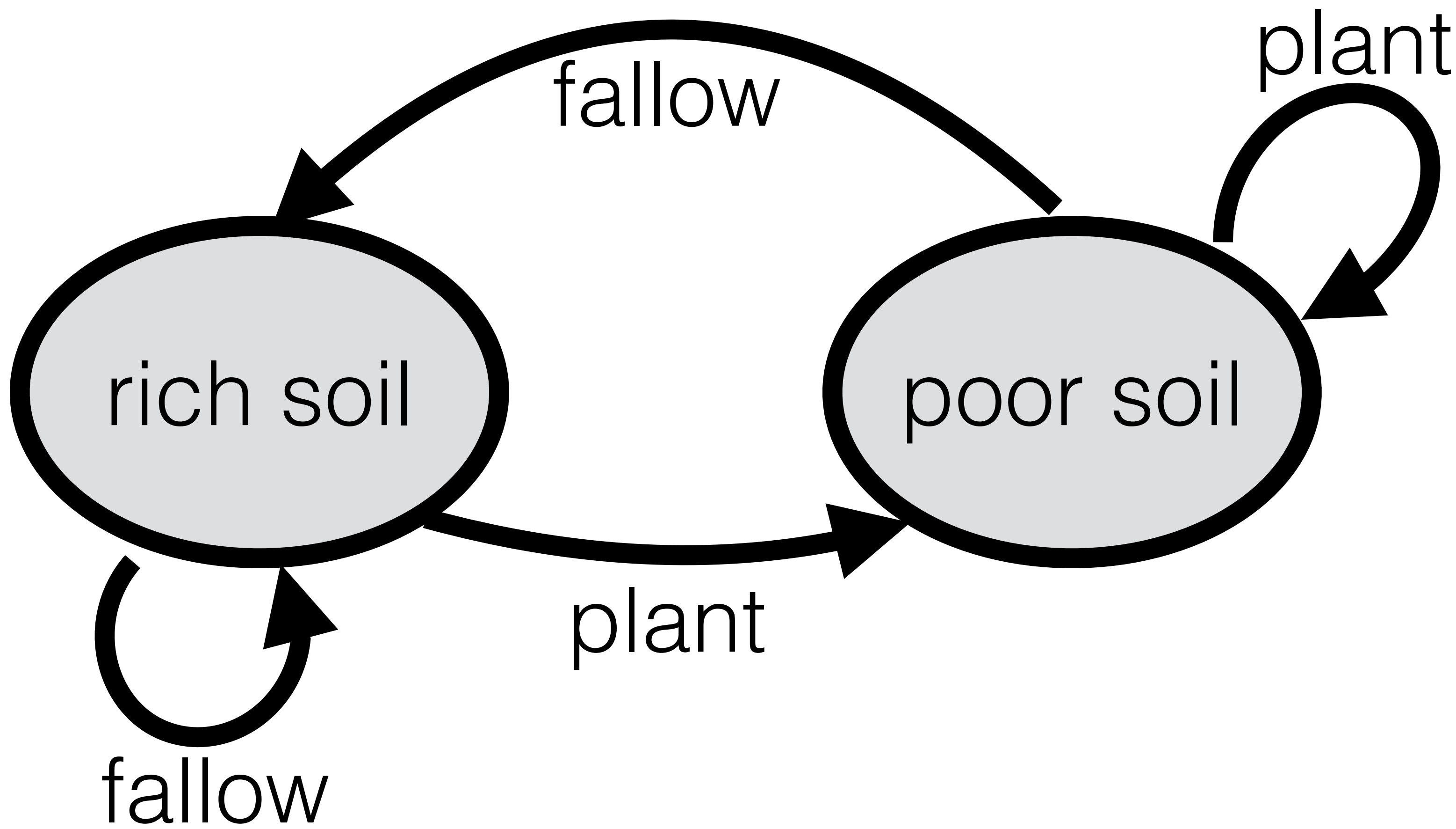

**Example**

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

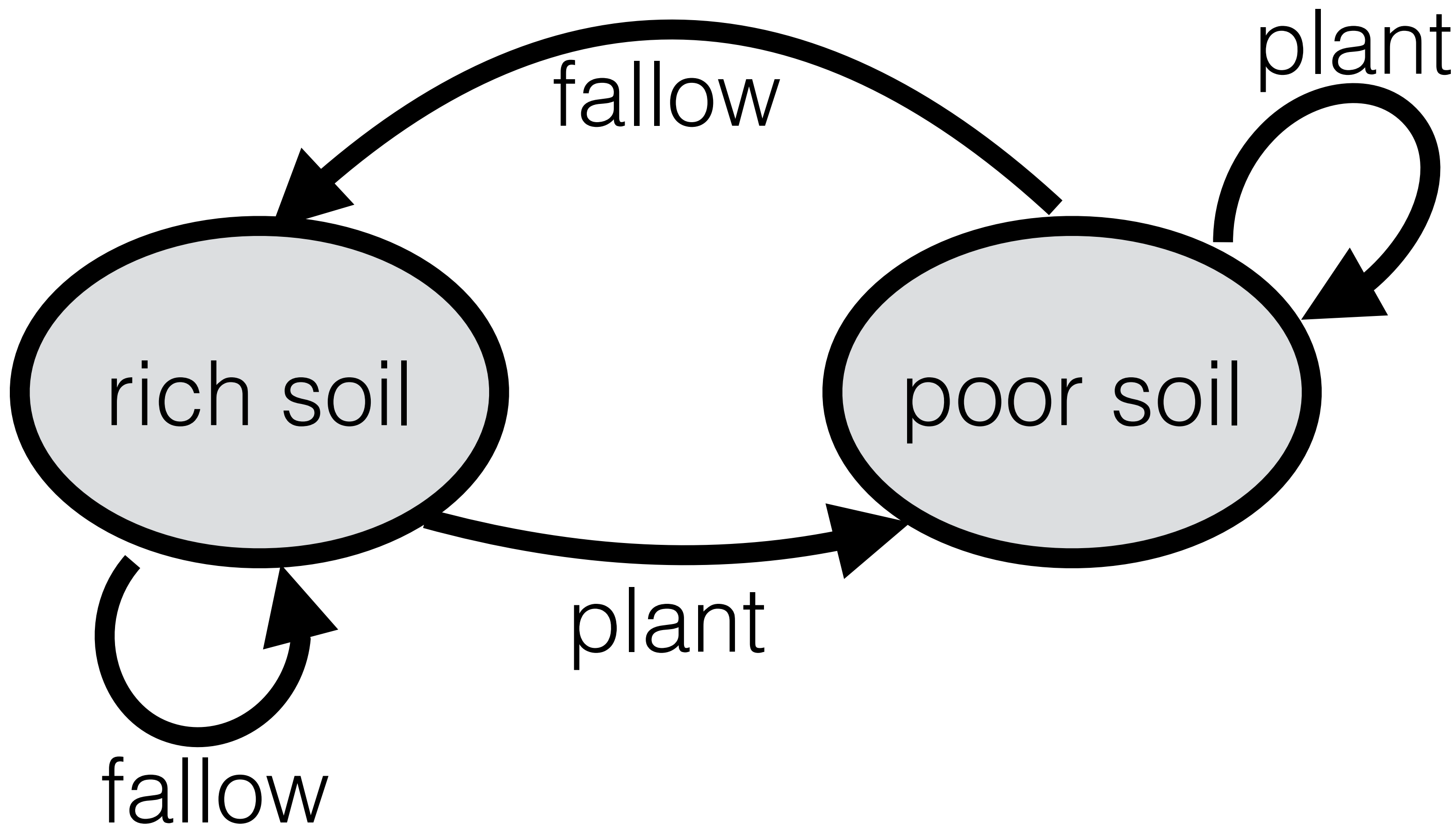
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$ : transition function
- $\mathcal{Y}$ : set of possible outputs
- $g : \mathcal{S} \rightarrow \mathcal{Y}$ : output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



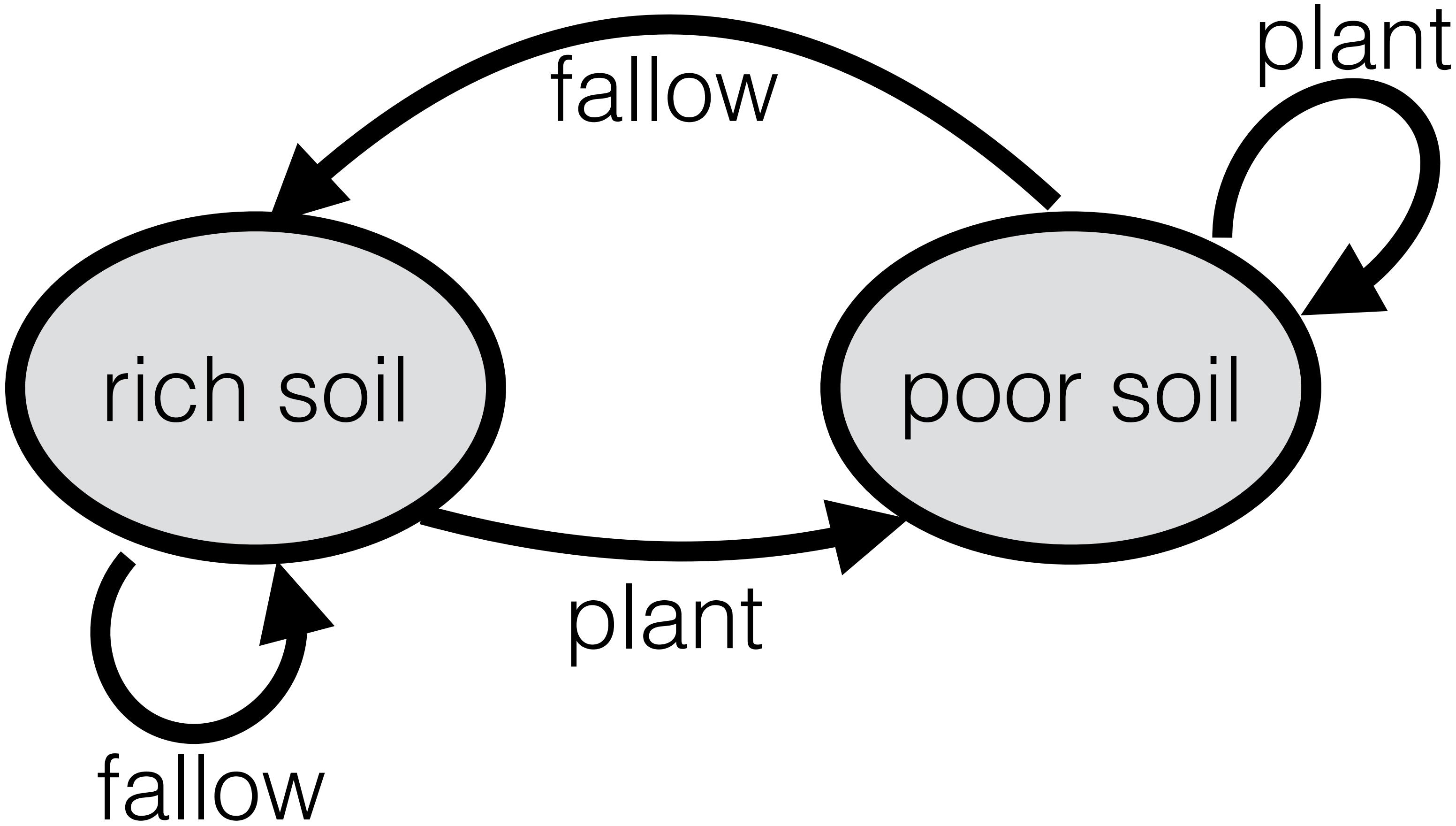
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ : transition function
- $\mathcal{Y}$ : set of possible outputs
- $g : \mathcal{S} \to \mathcal{Y}$ : output function
  - e.g.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



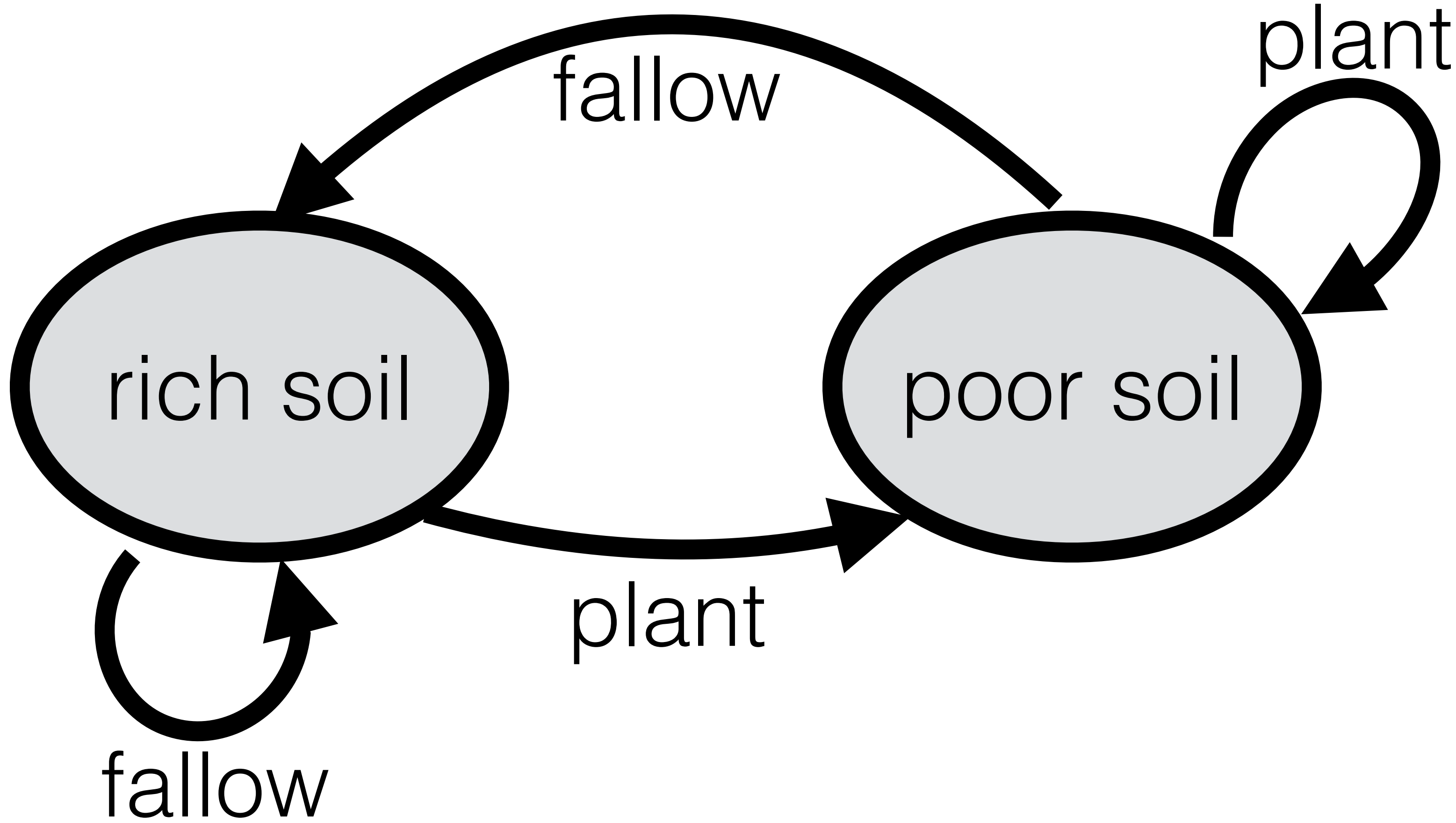
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ : transition function



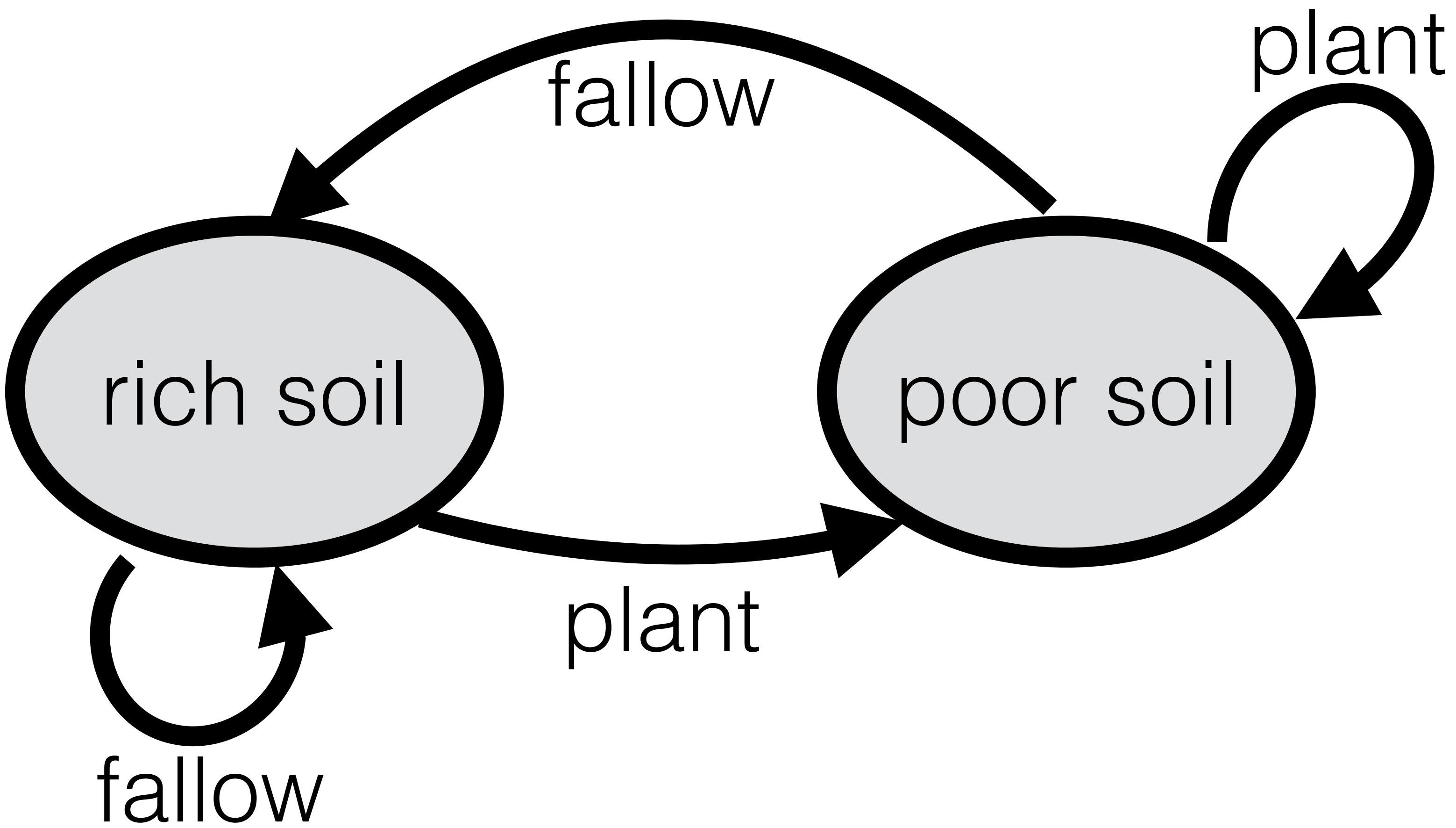
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ : transition function
- $R$ reward function



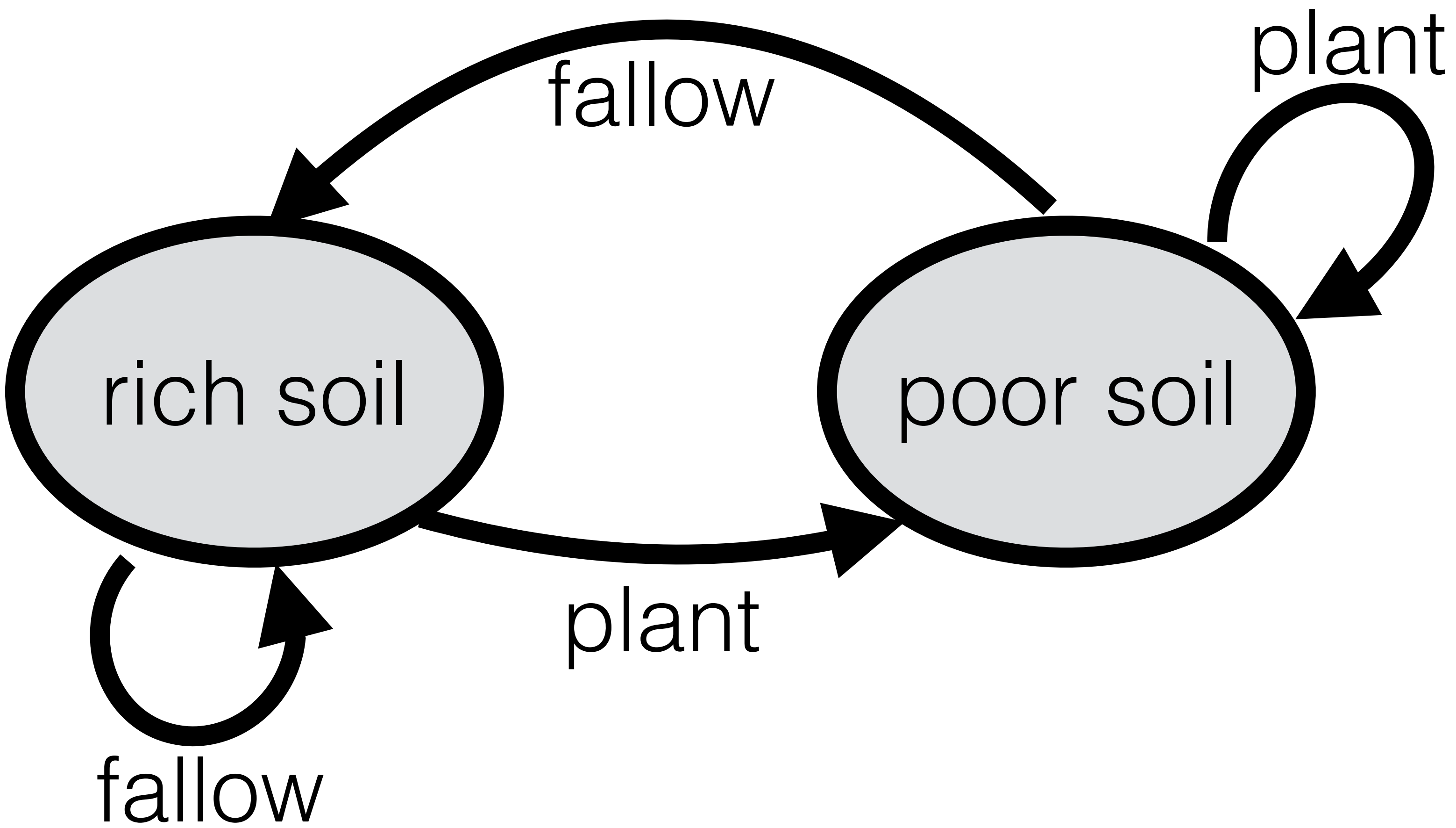
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$ : transition function
- $R$
  reward function
  - e.g. # bushels in harvest



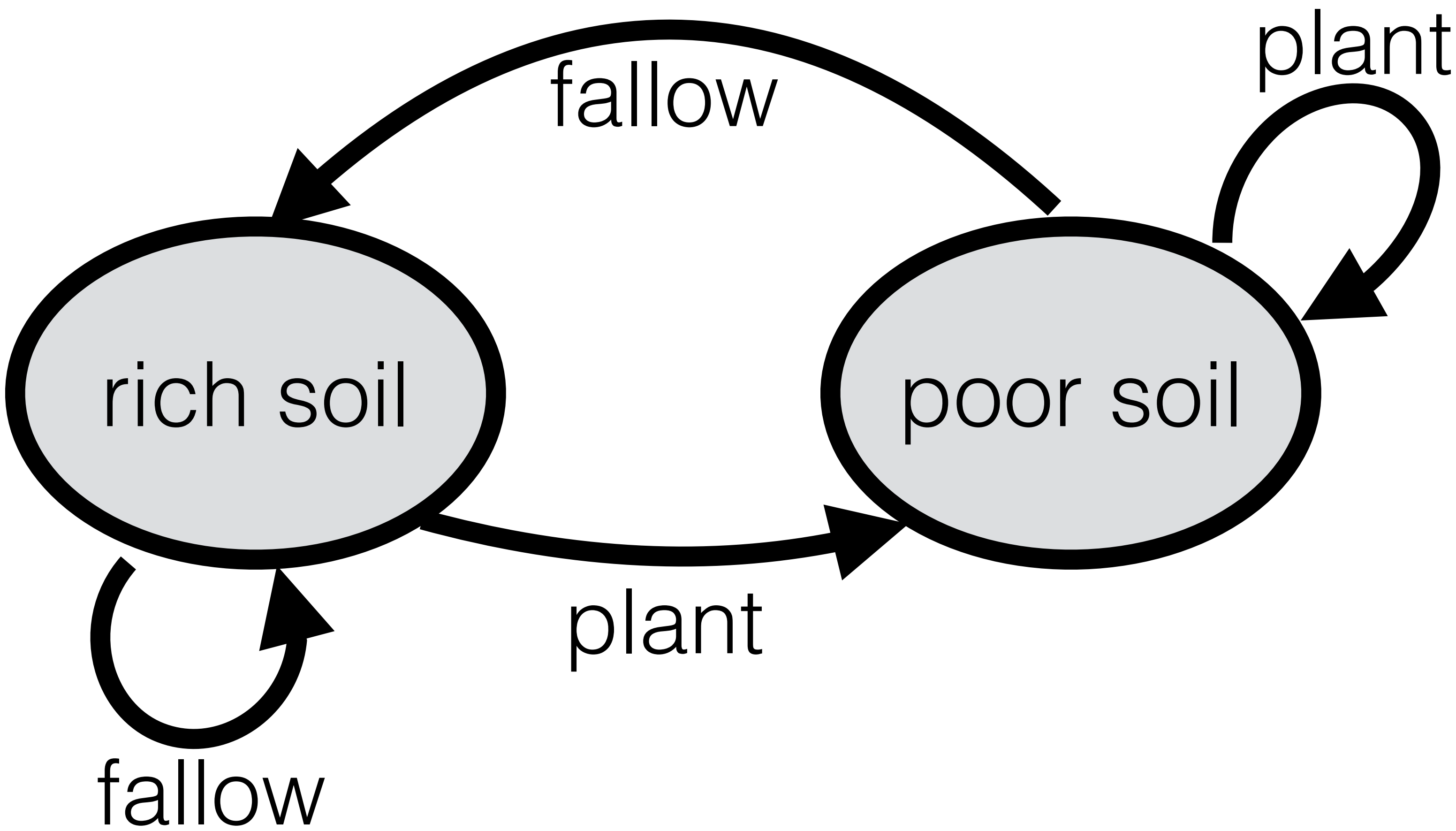
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ : transition function
- $R$ reward function
  - e.g. # bushels in harvest



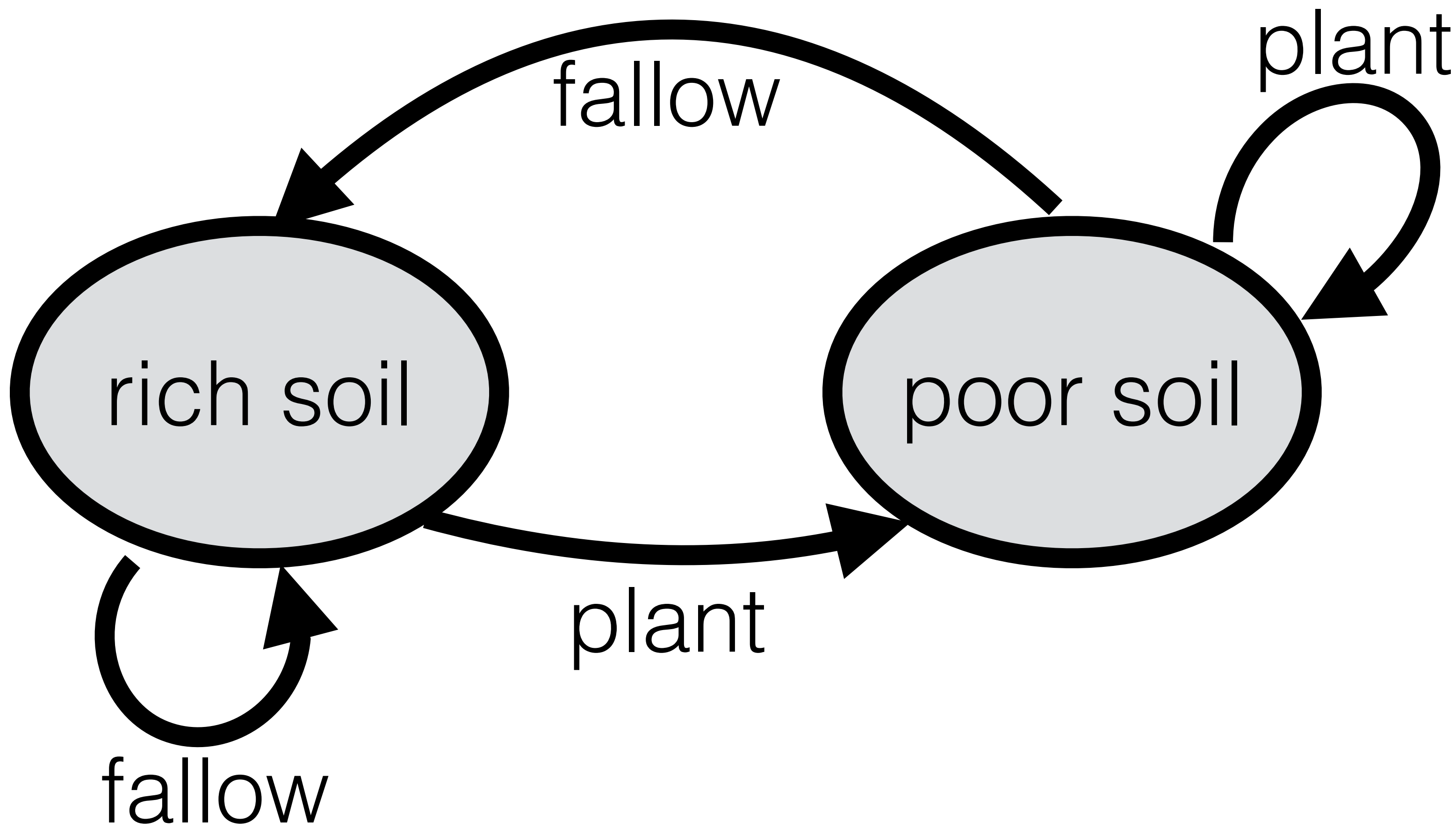
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ : transition function
- $R :$ $\to \mathbb{R}$
  reward function
  - e.g. # bushels in harvest



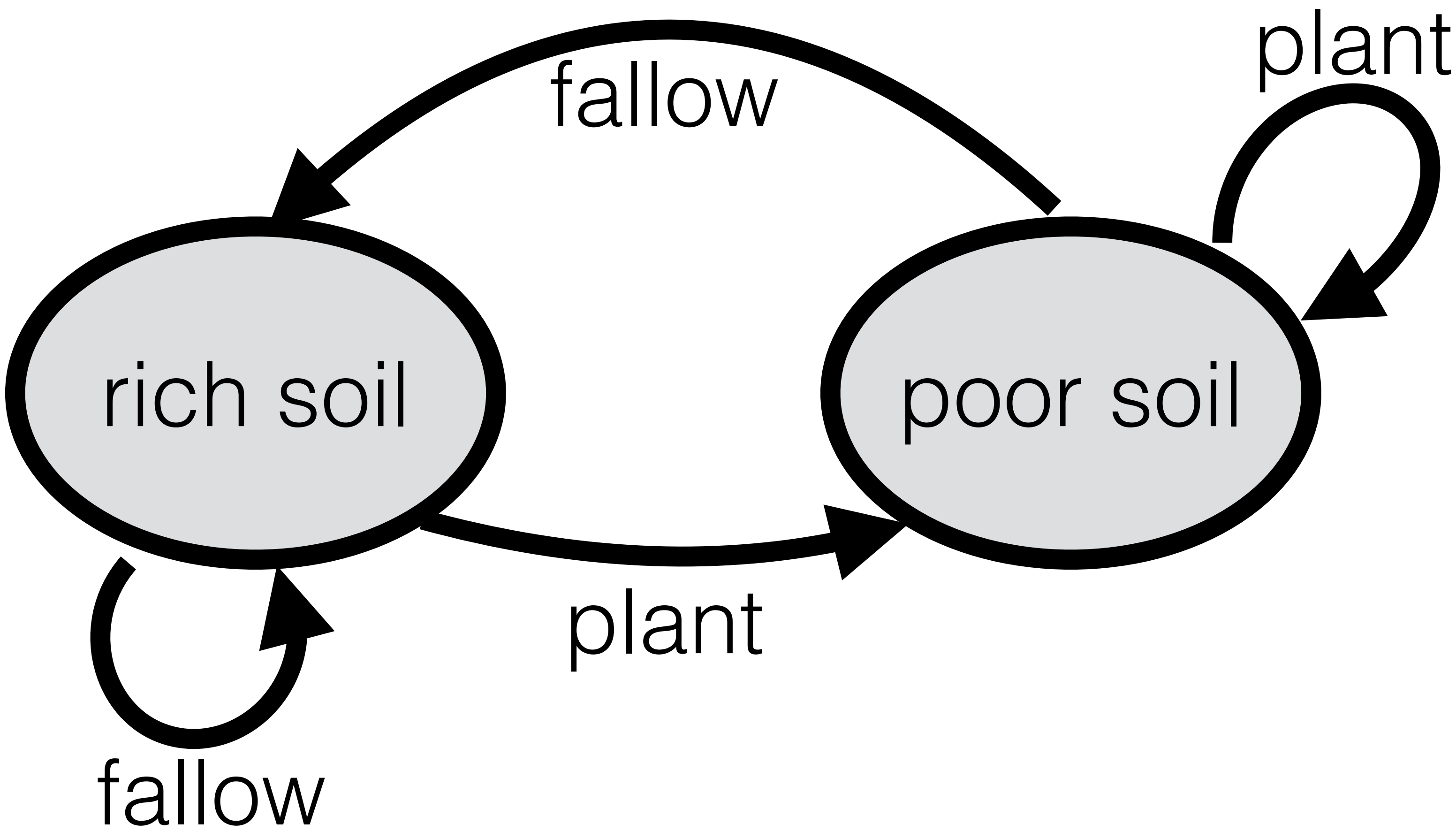
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ : transition function
- $R : \quad \mathcal{X} \to \mathbb{R}$ reward function
    - e.g. # bushels in harvest



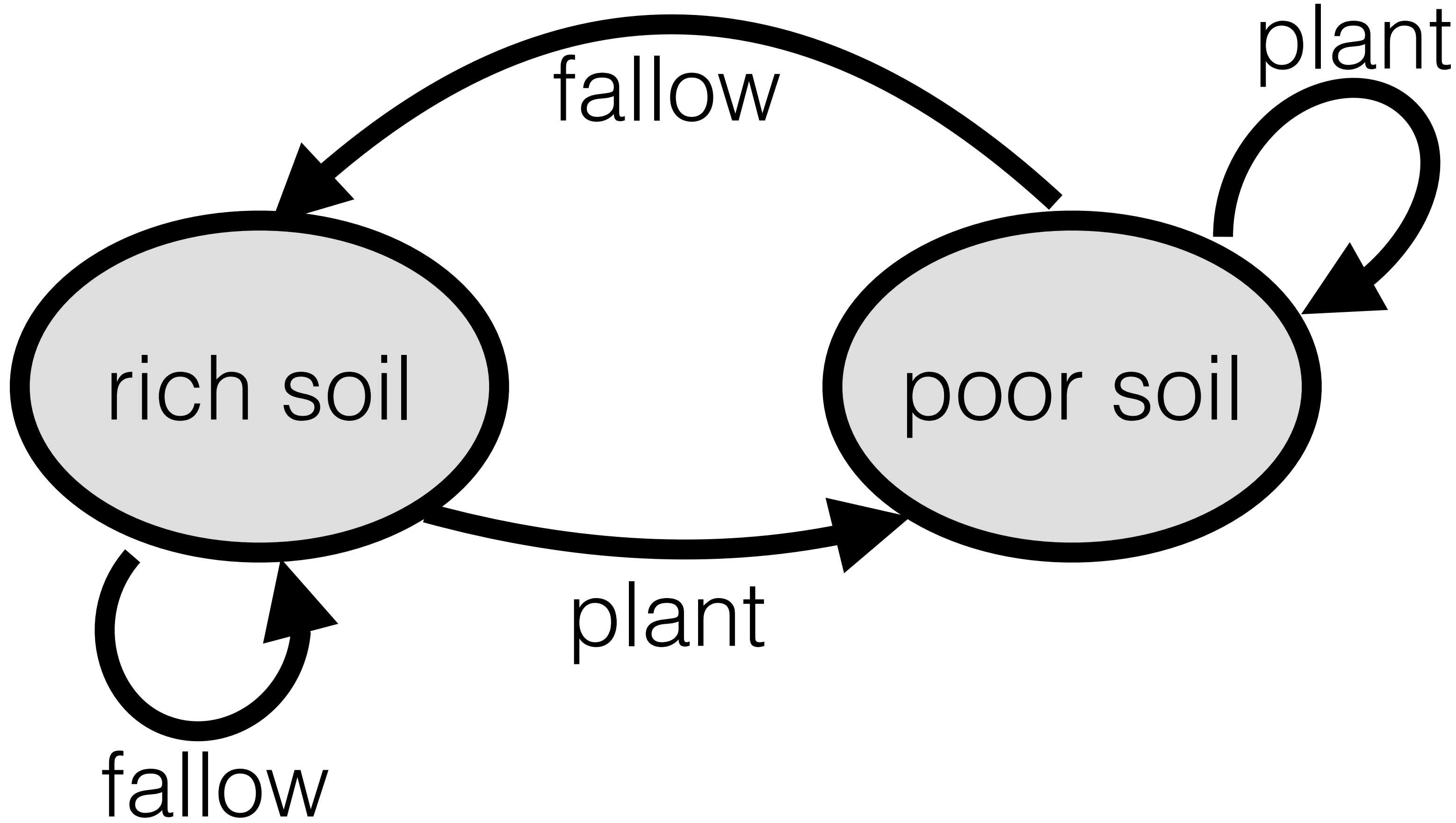
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ : transition function
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. # bushels in harvest



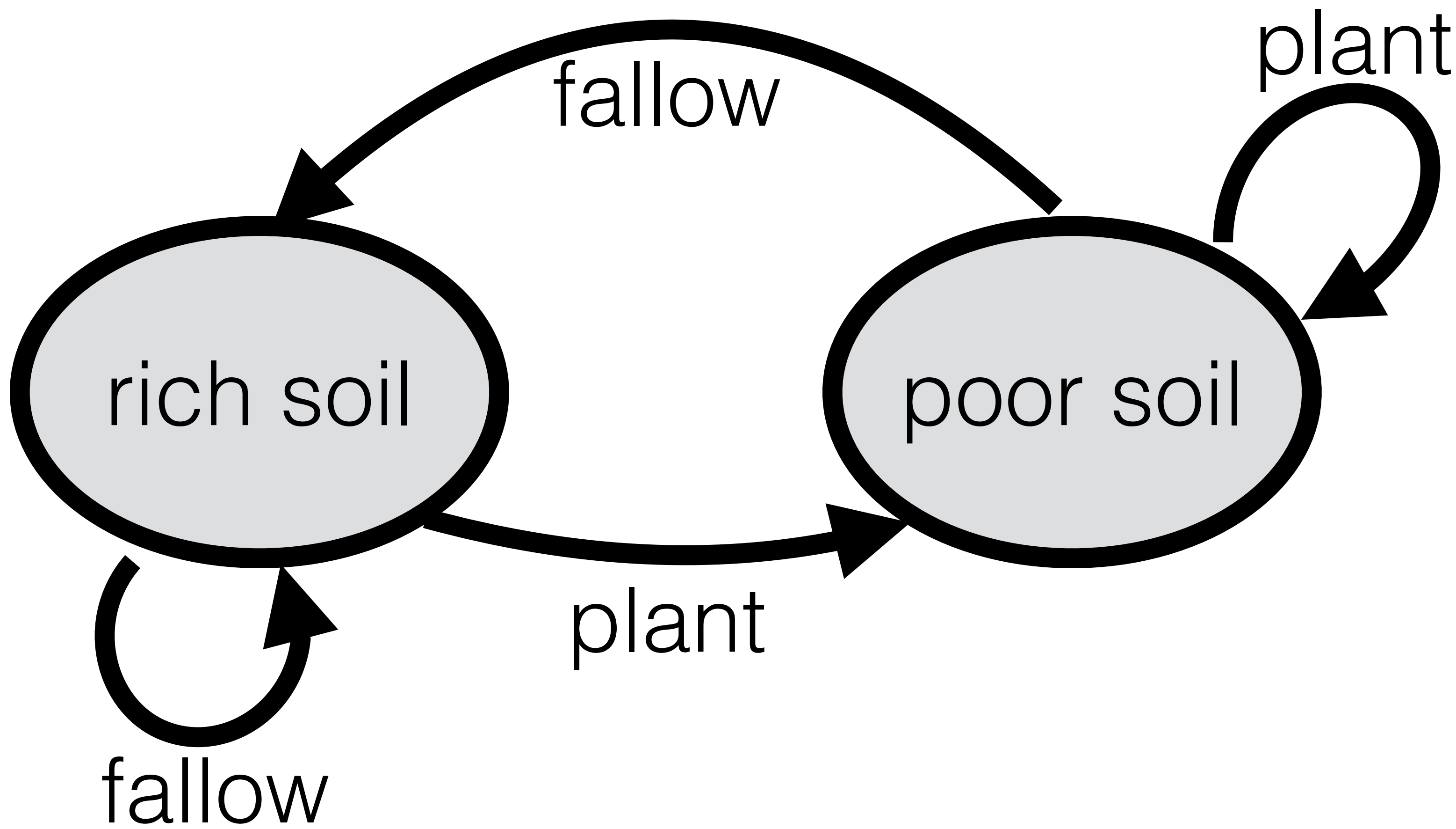
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ : transition function
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. # bushels in harvest



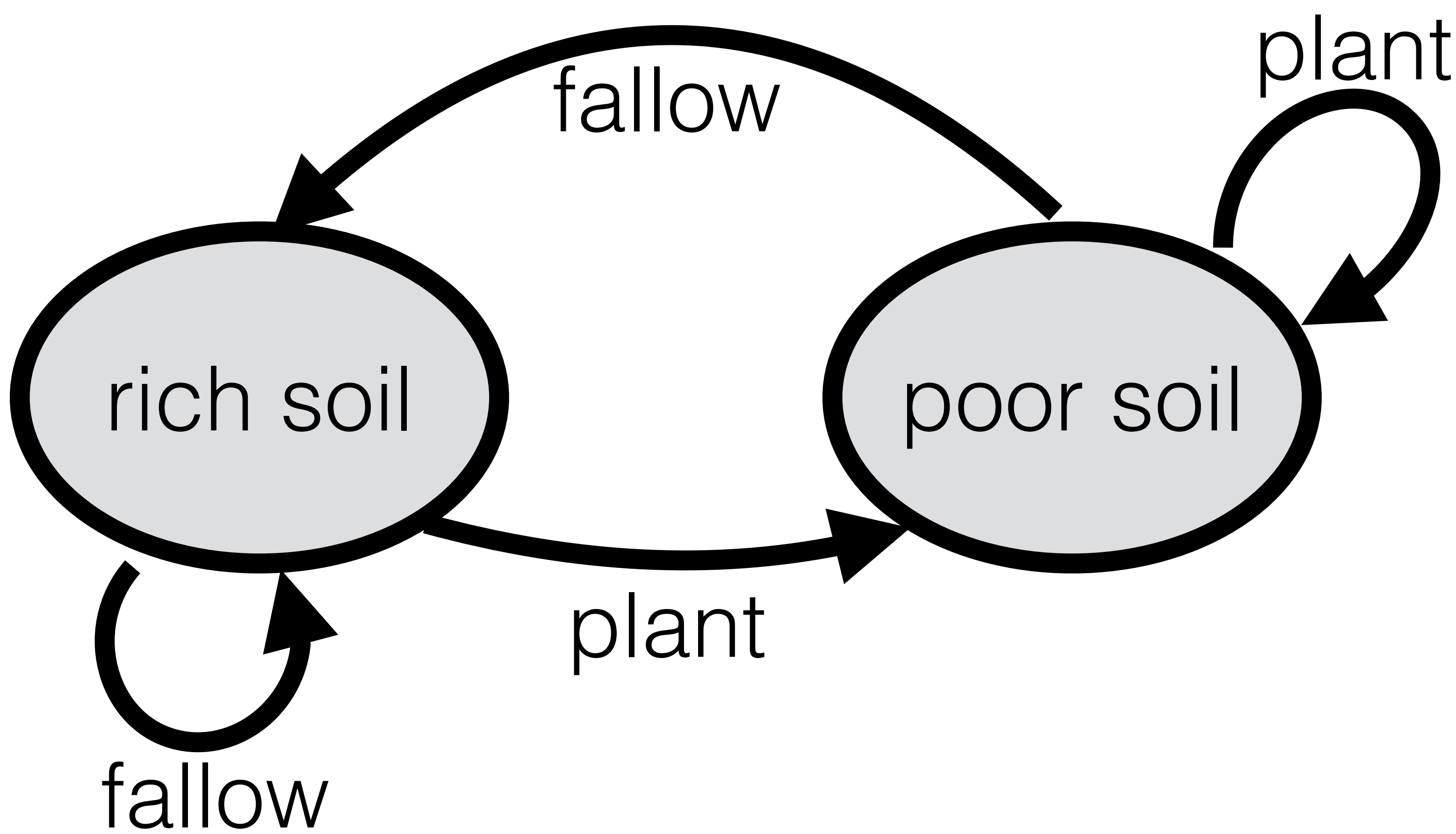
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ : transition function
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels



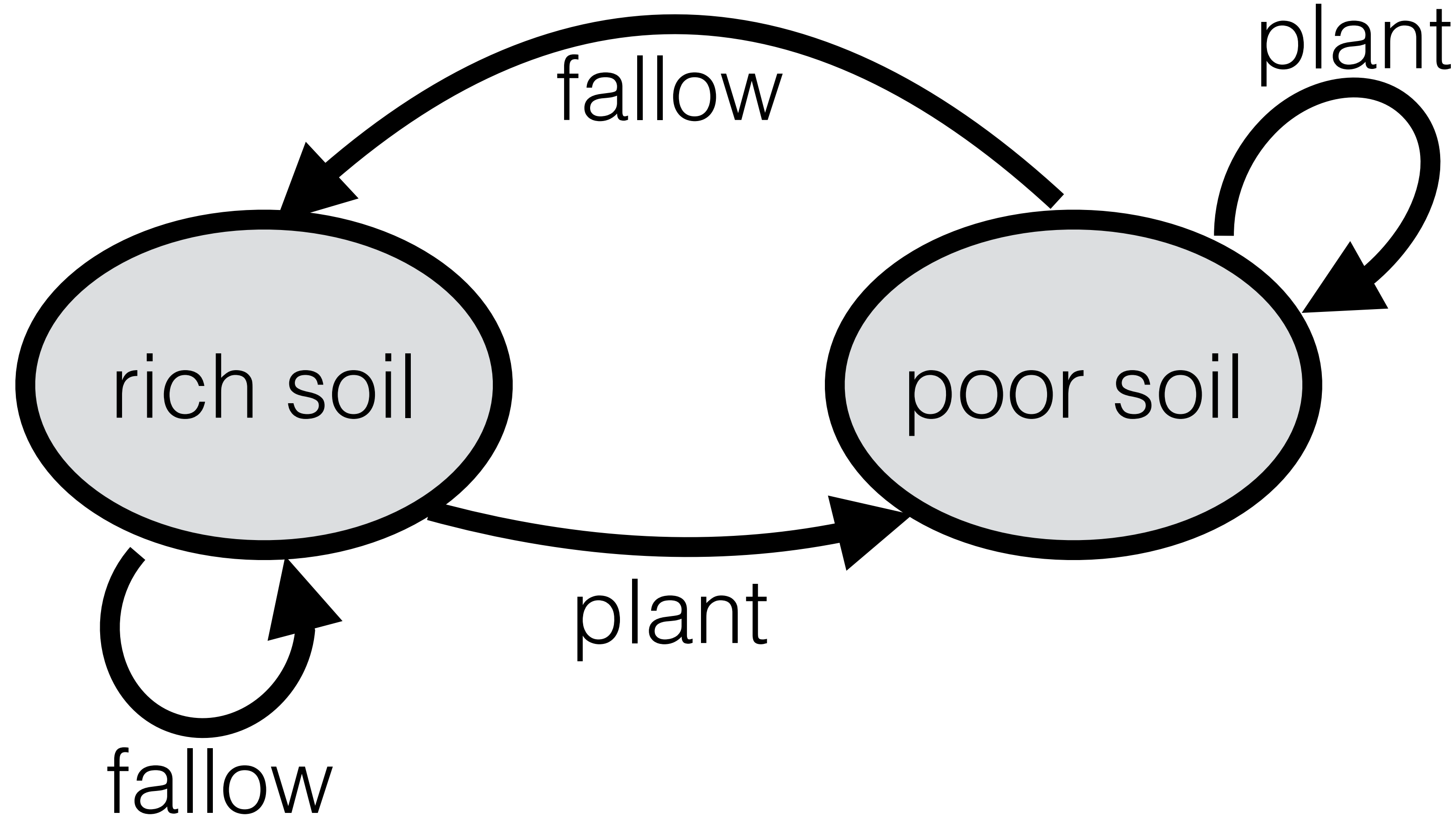
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ : transition function
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels;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



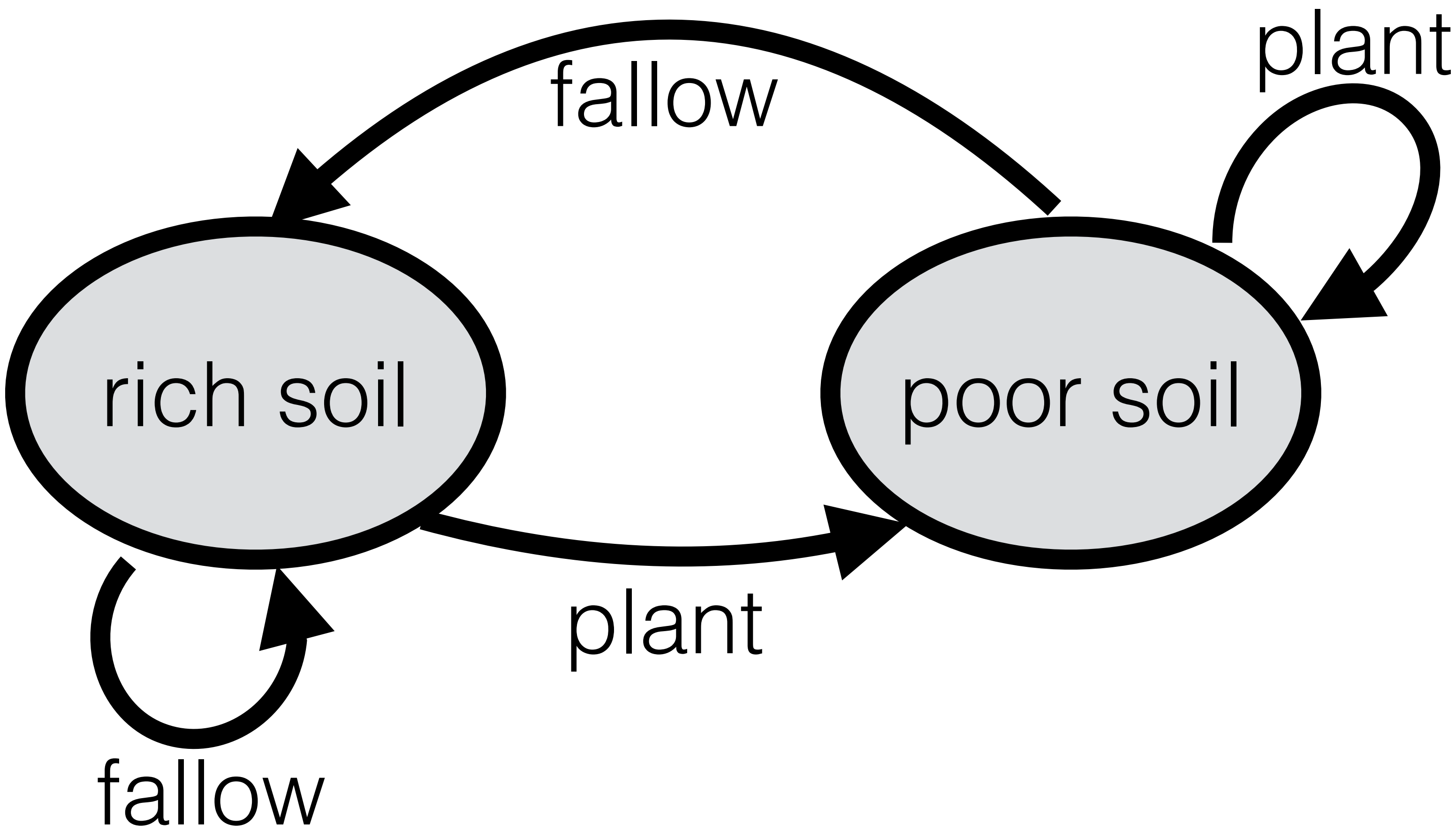
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ : transition function
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$ : reward function
  - e.g. R(rich, plant) = 100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



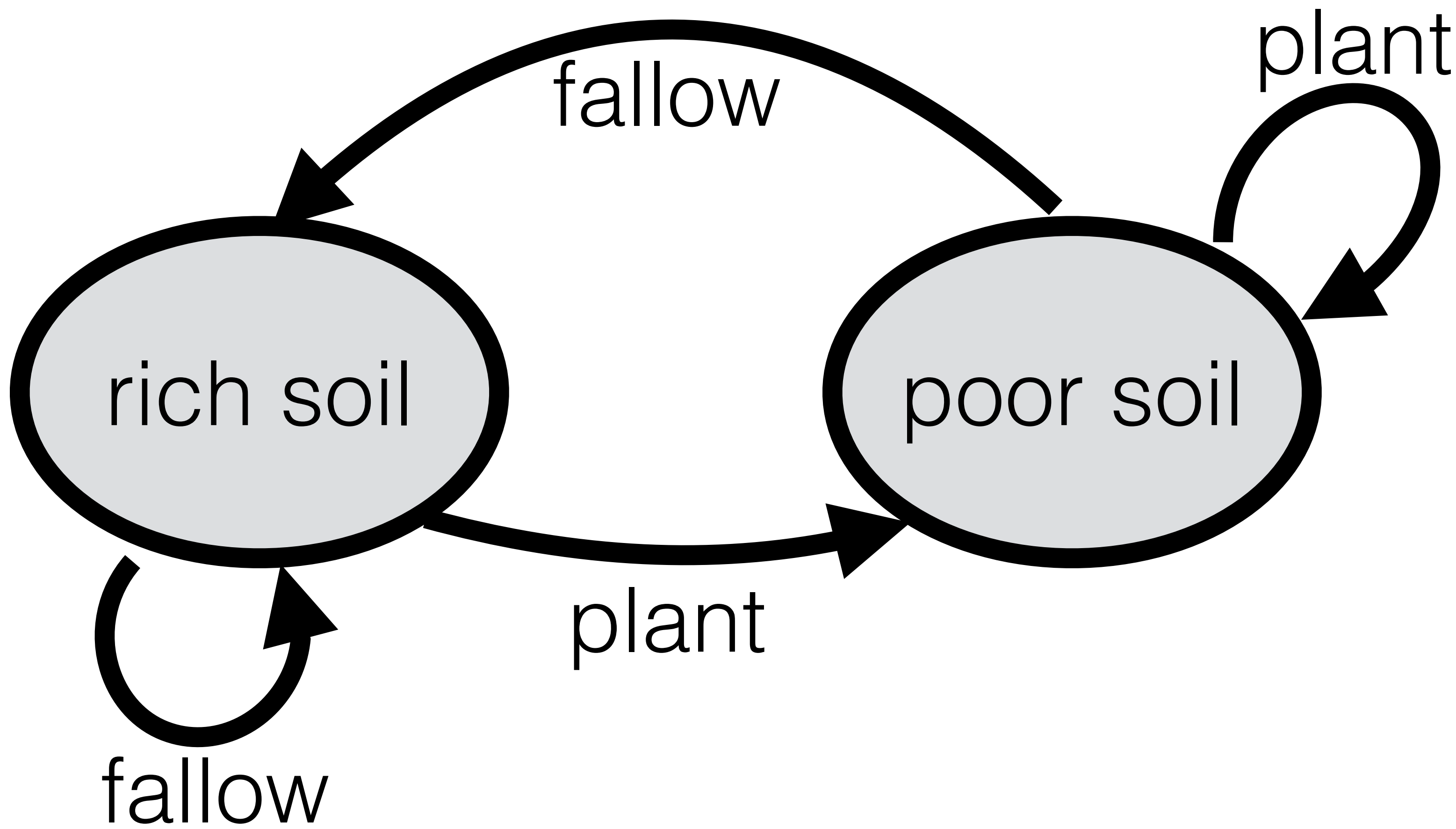
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ : transition function
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels;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



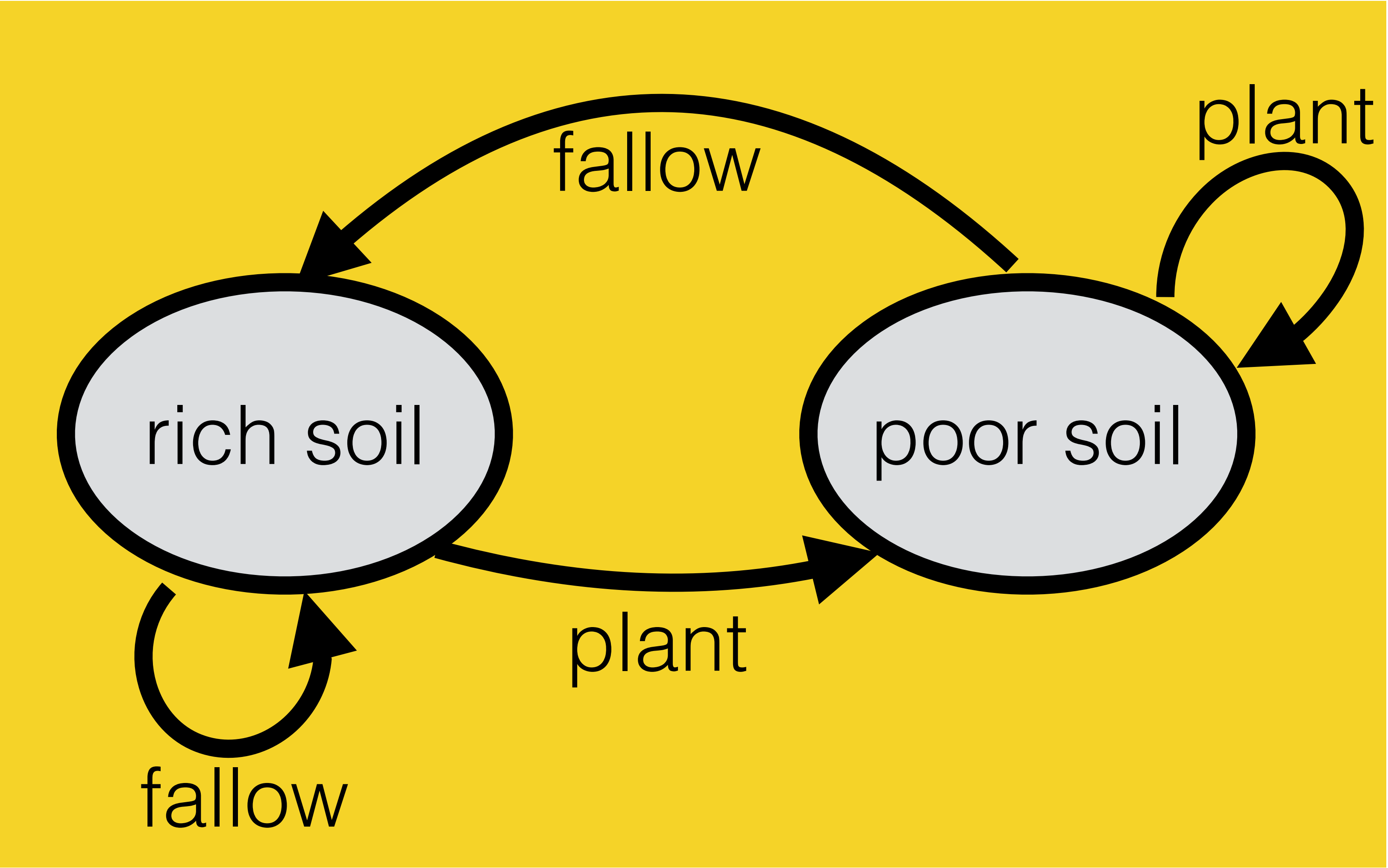
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$ : transition function
- $R : \mathcal{S} \times \mathcal{X} \rightarrow \mathbb{R}$: reward function
    - e.g. R(rich,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



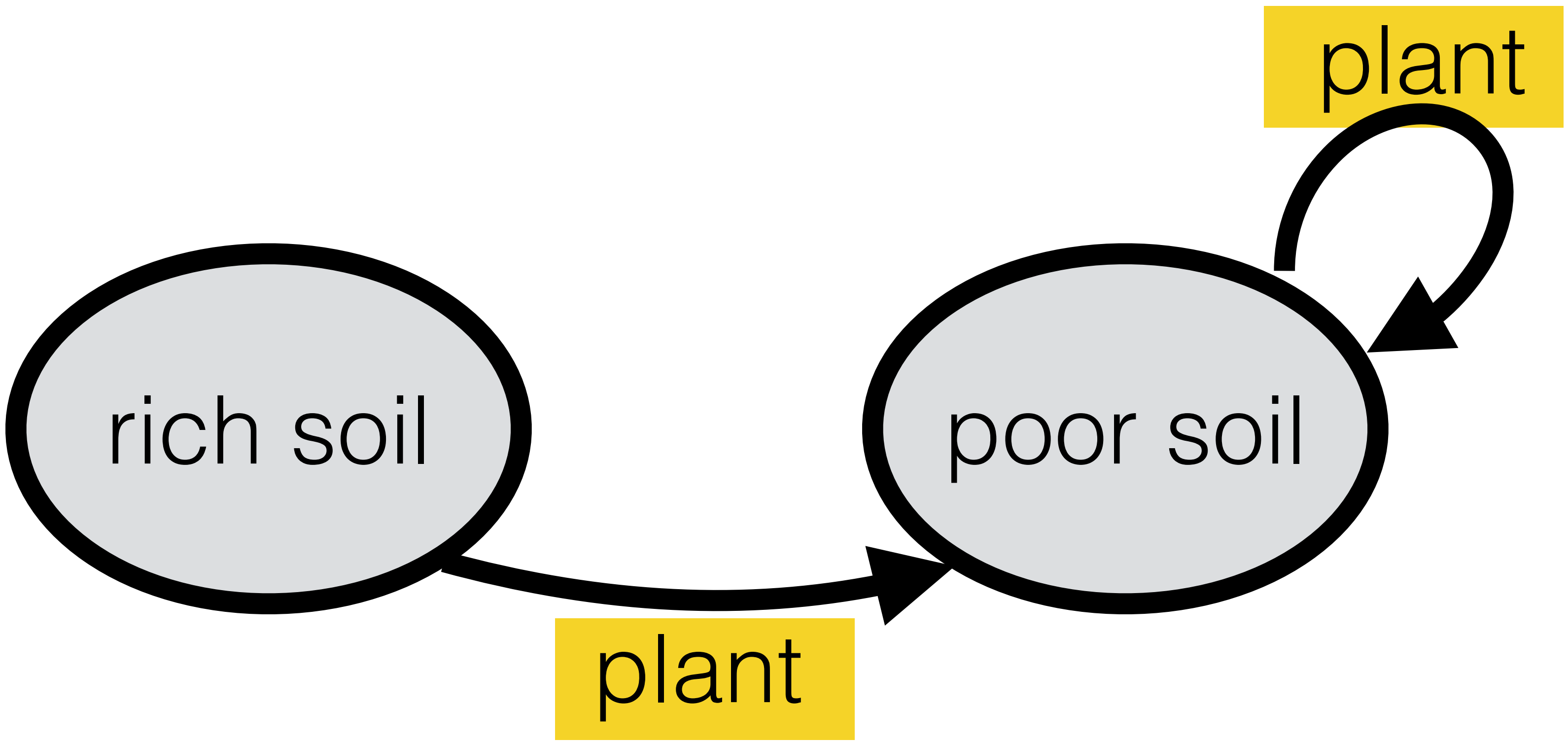
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$ transition model
- $R : \mathcal{S} \times \mathcal{X} \rightarrow \mathbb{R}$: reward function
    - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich,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



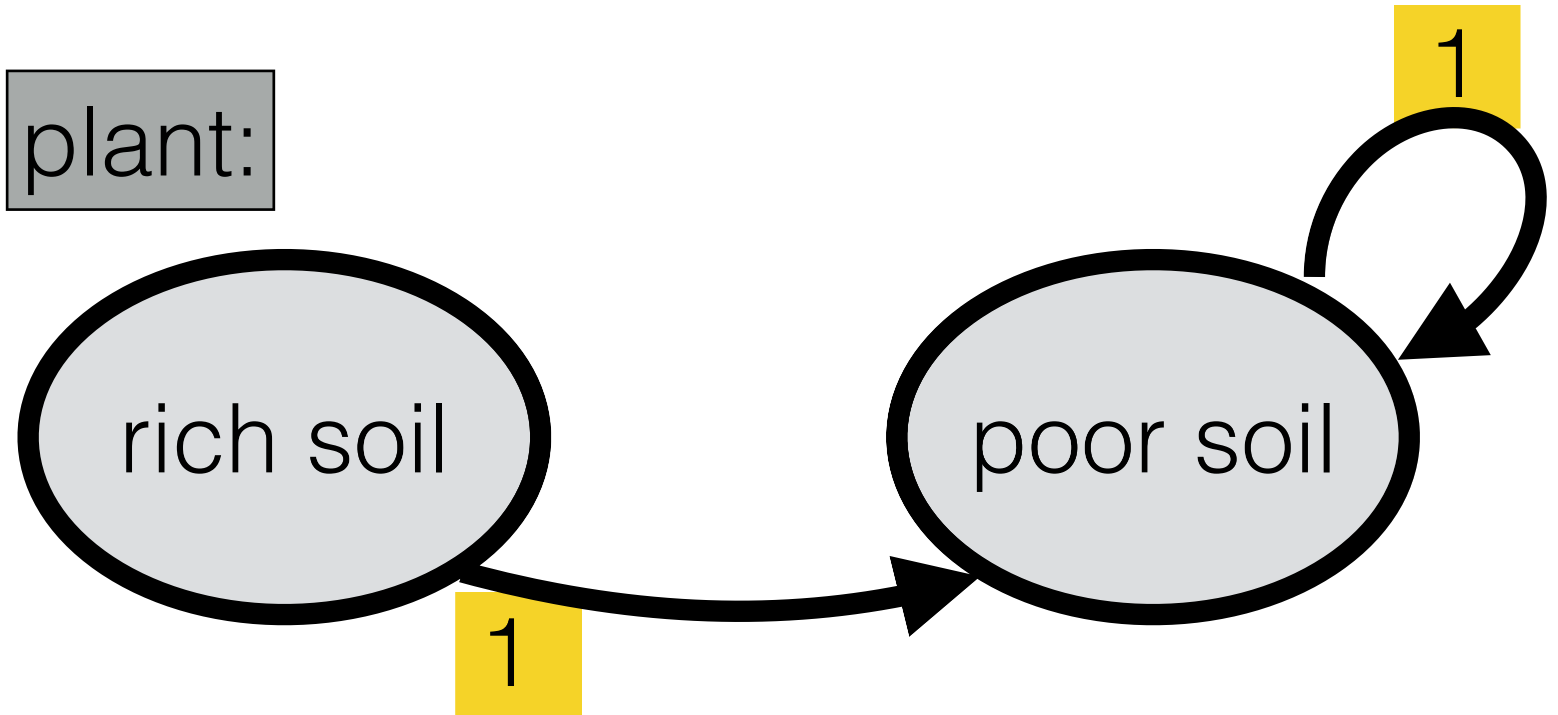
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$ transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) =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



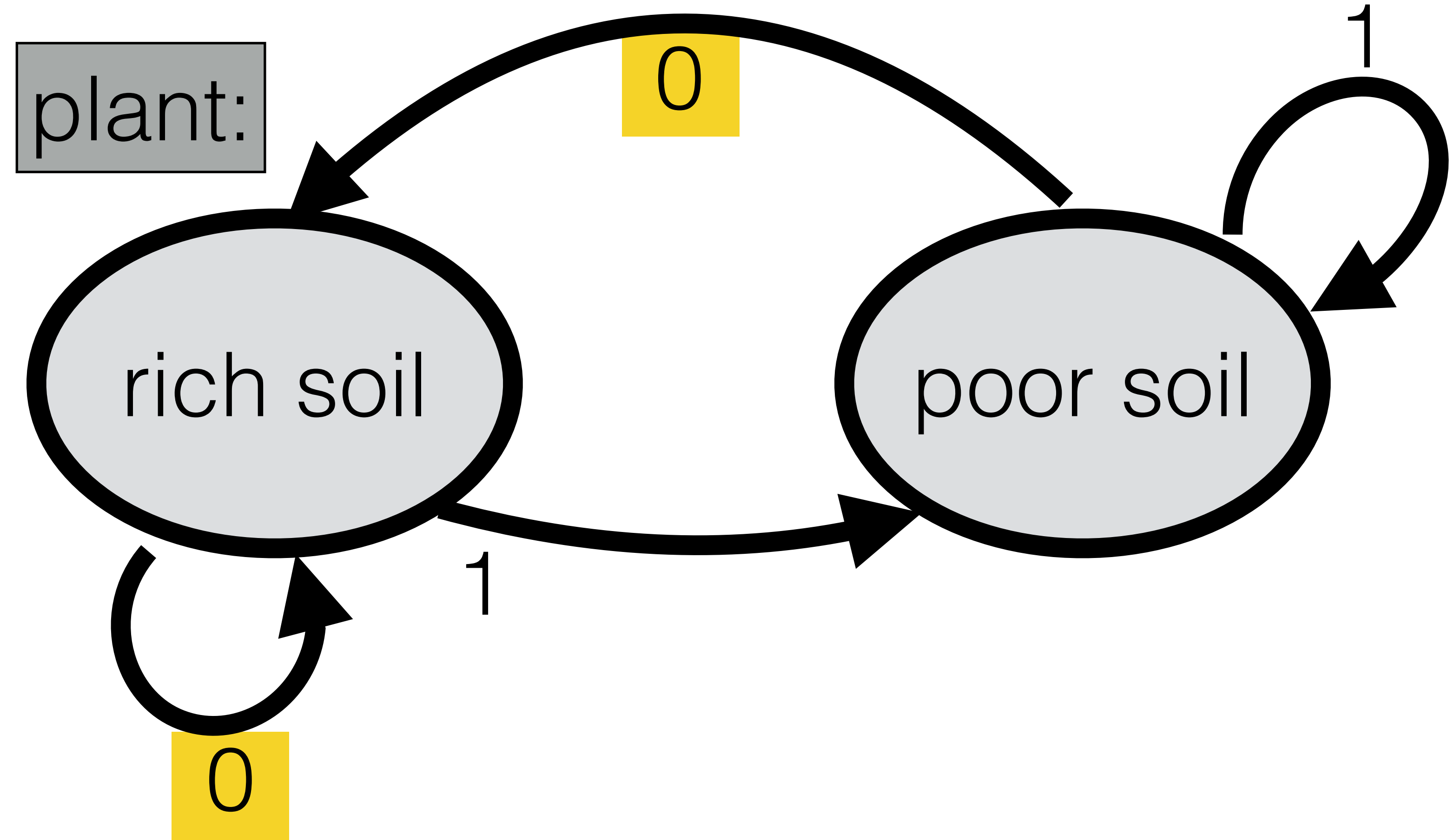
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$
  transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$ :
  reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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



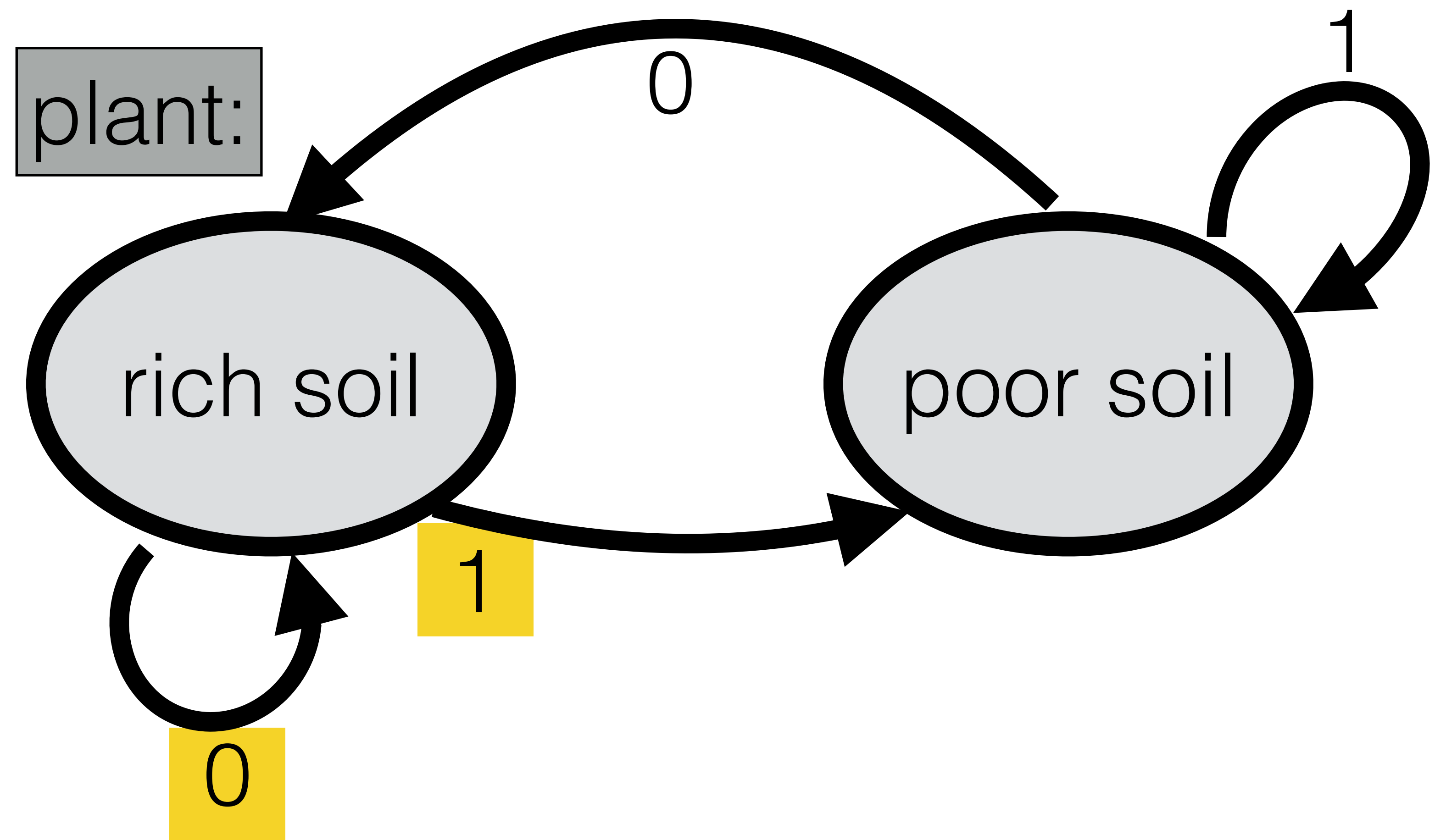
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$
  transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$:
  reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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



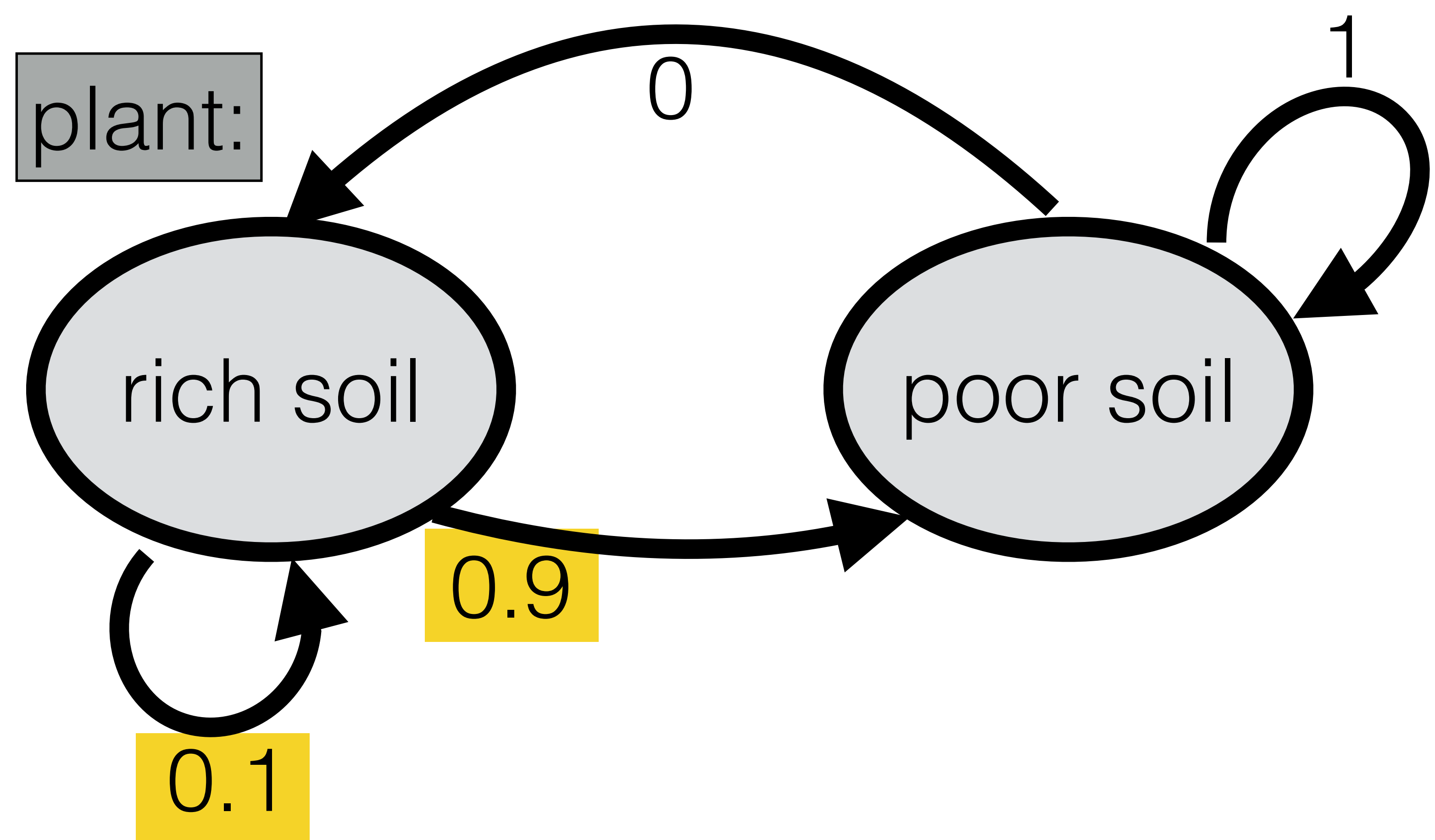
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$ transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
    - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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



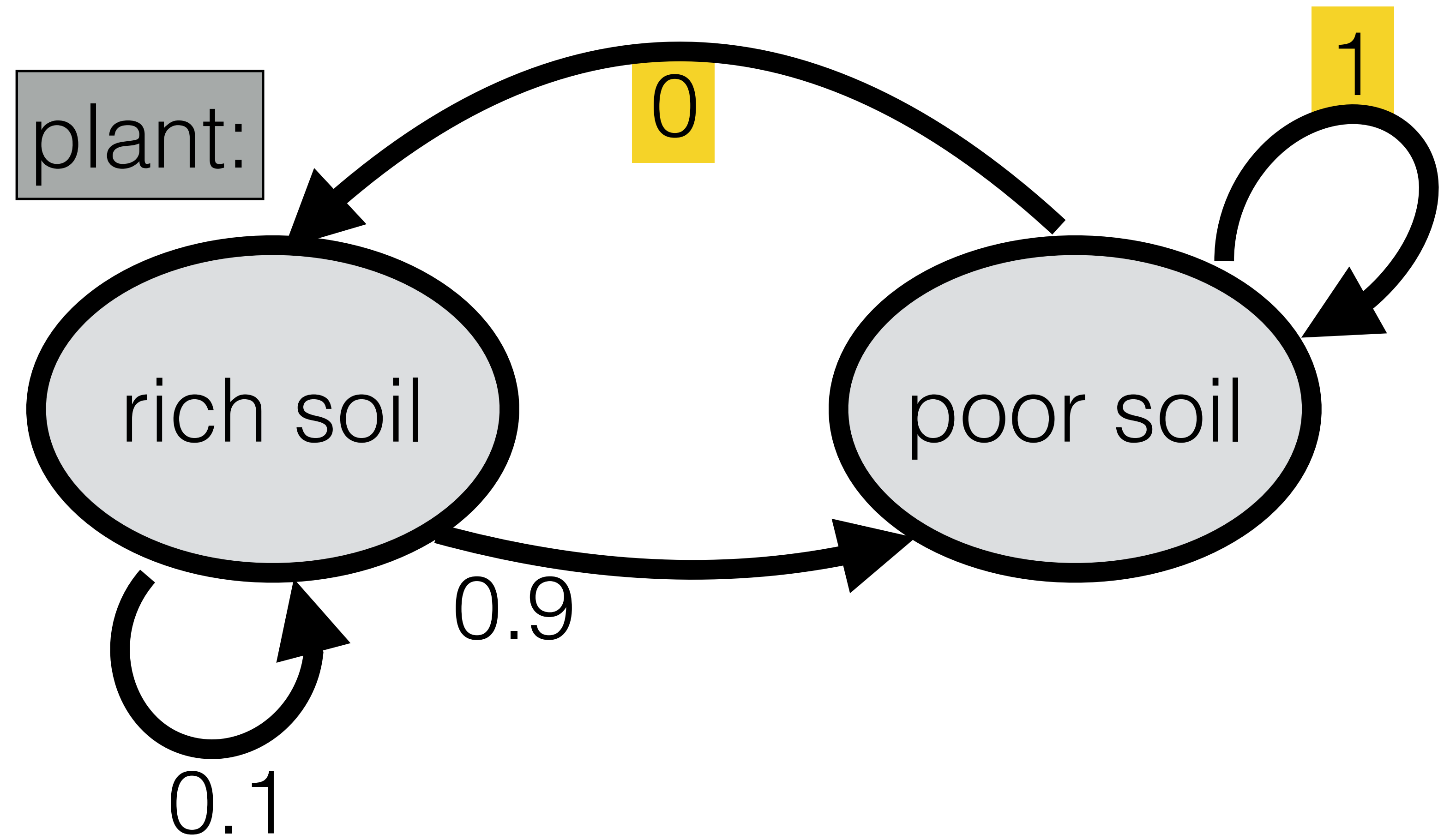
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$ transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) =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



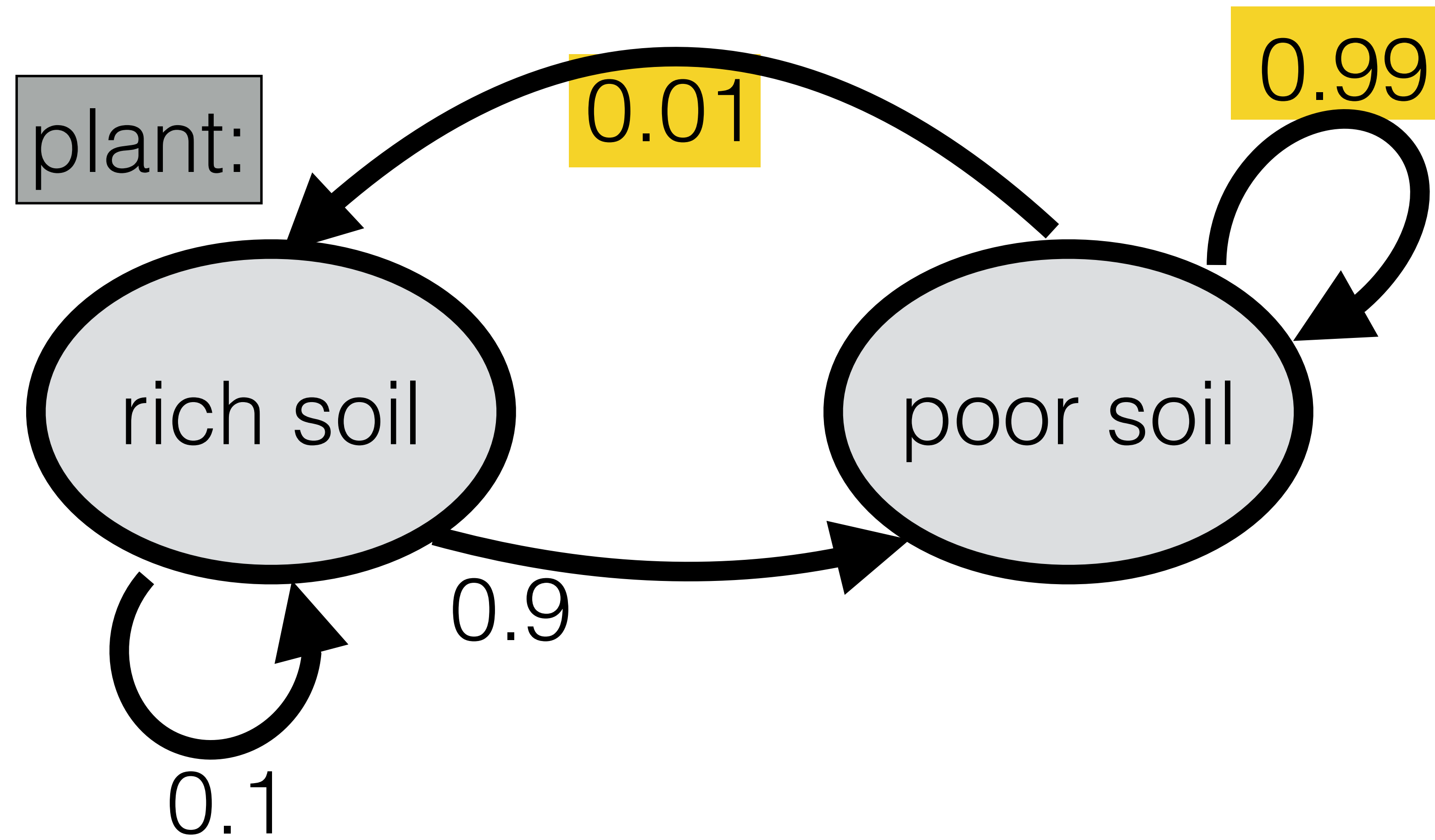
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$ transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
    - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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



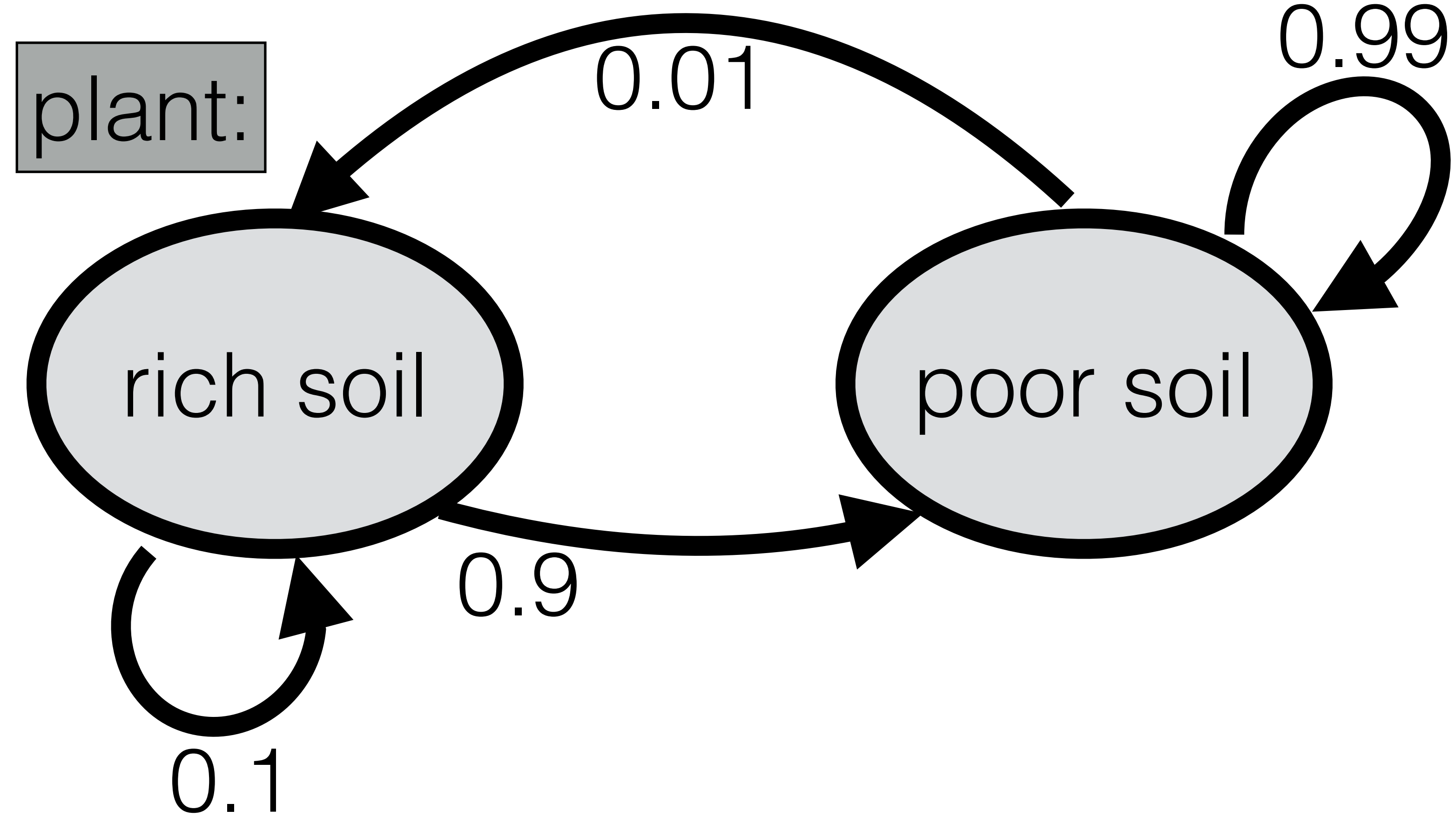
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$ transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
    - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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



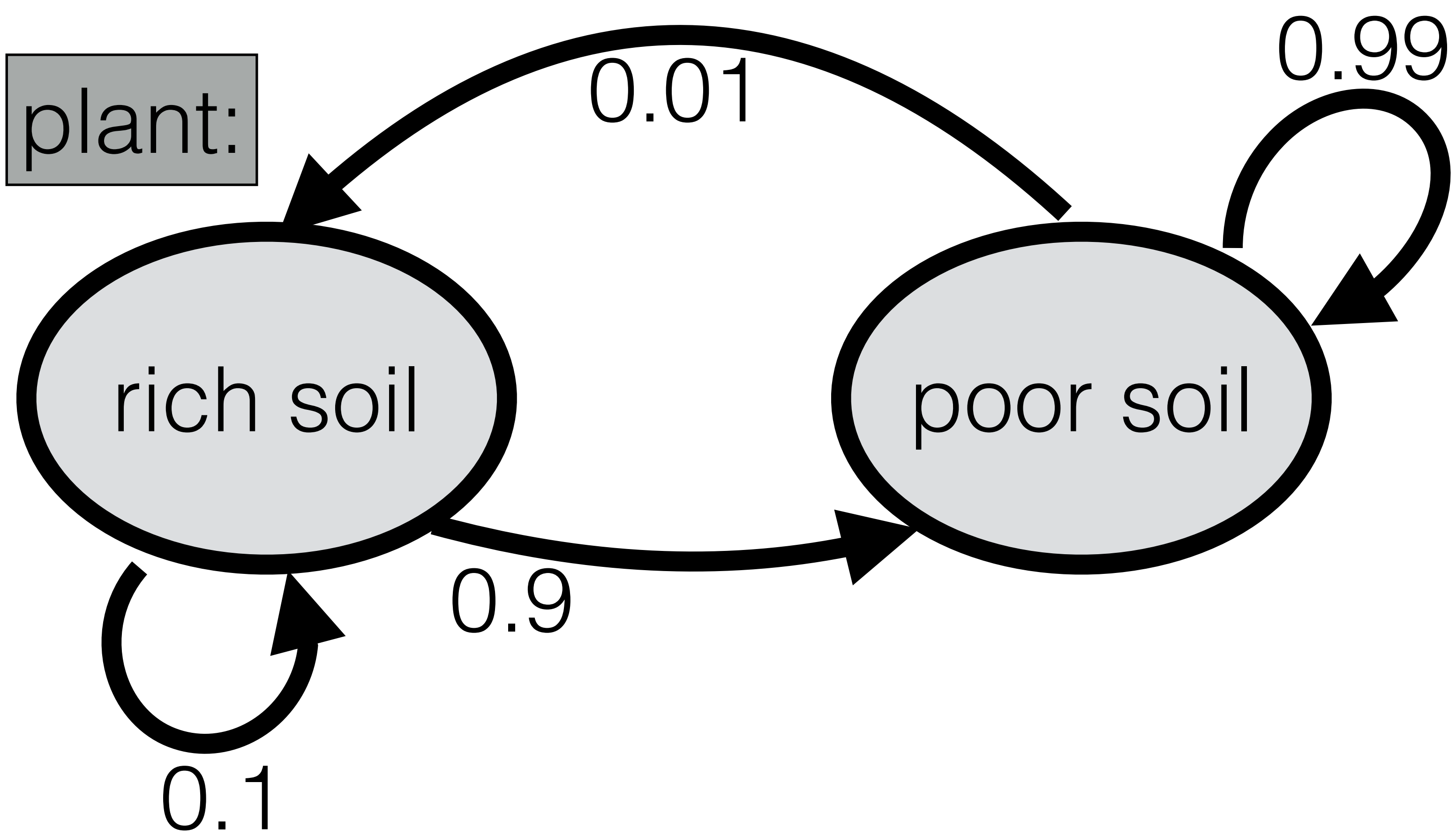
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$
  transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$:
  reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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



- $\mathcal{S}$ = 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

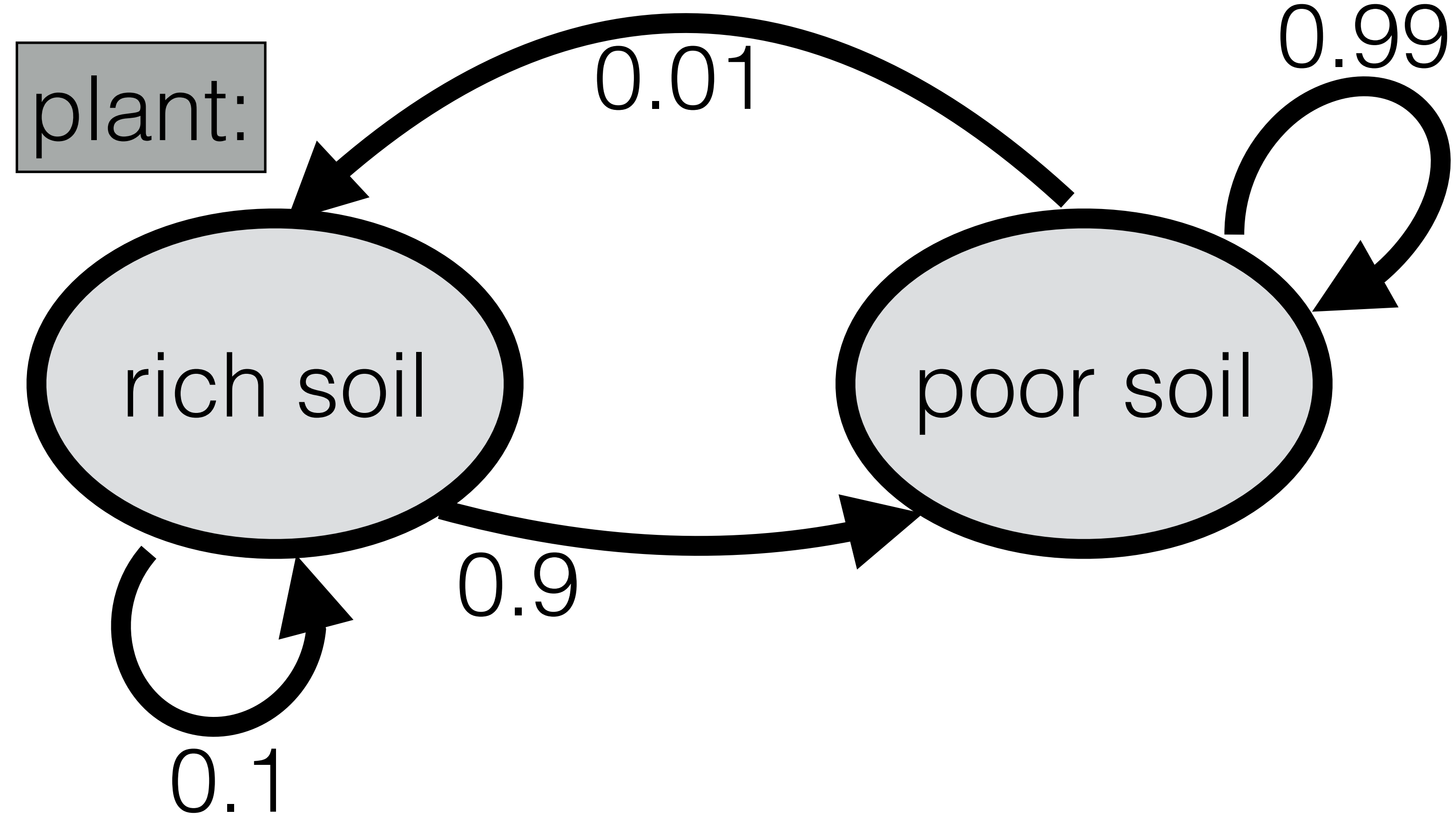


- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$ transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) =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




- Transition matrix for "plant" action:

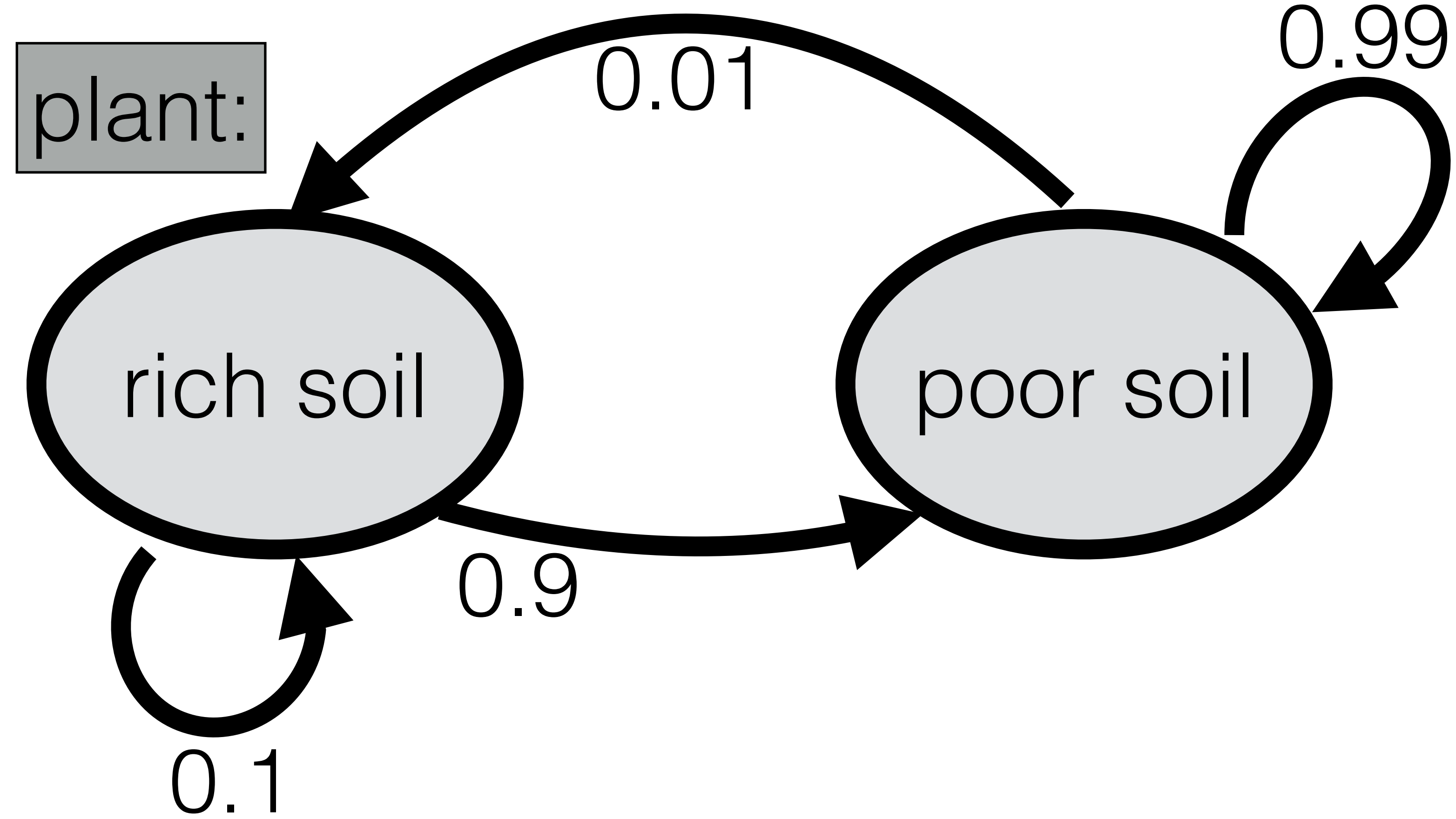
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$
  transition model
- $R : \mathcal{S} \times \mathcal{X} \rightarrow \mathbb{R}$:
  reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich,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




- Transition matrix for "plant" action:

$$\begin{array}{r c} & \begin{array}{cc} \text{rich} & \text{poor} \end{array} \\ \begin{array}{r} \text{rich} \\ \text{poor} \end{array} & \left[ \begin{array}{cc} \phantom{0.00} & \phantom{0.00} \\ \phantom{0.00} & \phantom{0.00} \end{array} \right] \end{array}$$

- $\mathcal{S}$ = 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

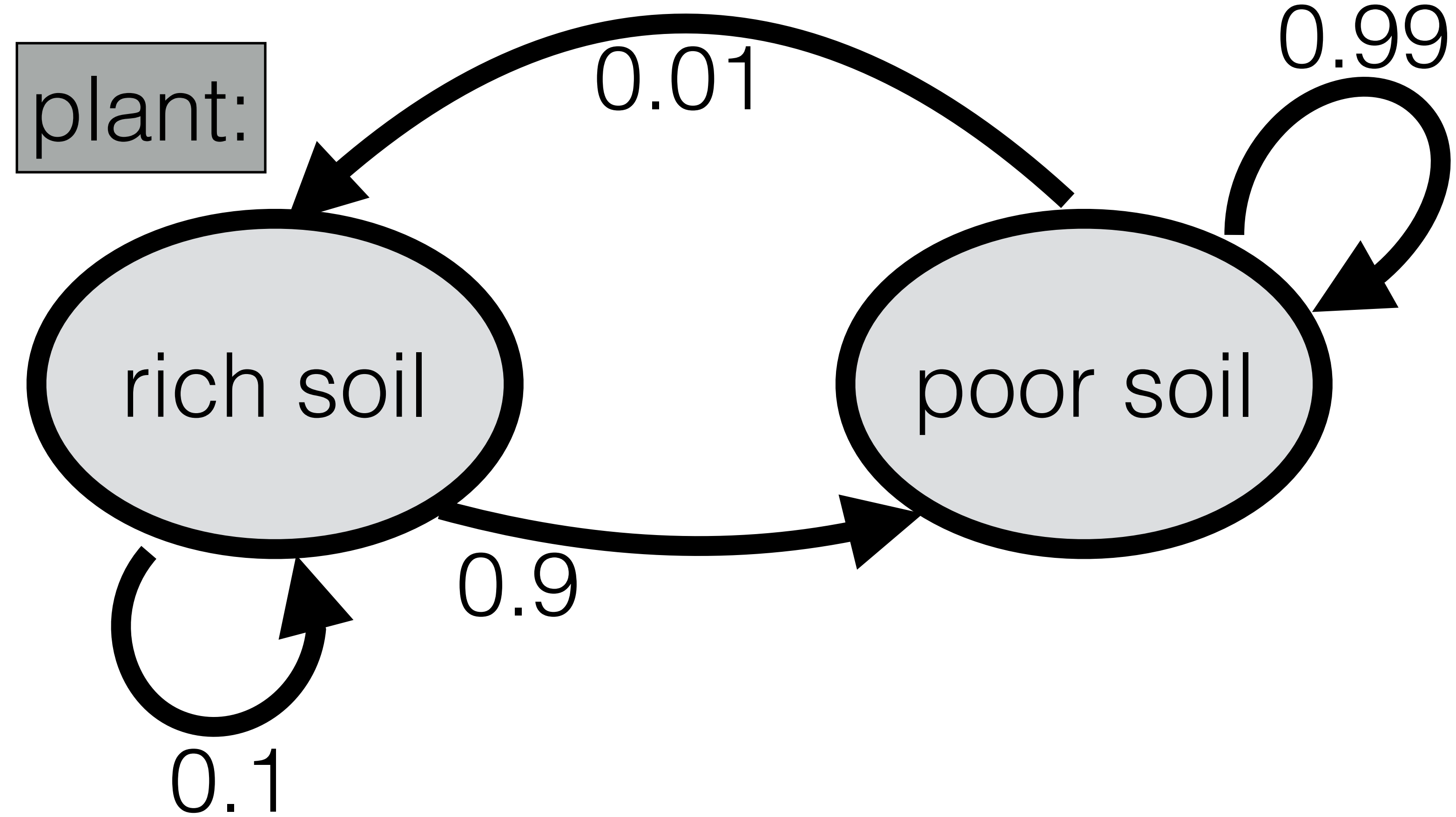



- Transition matrix for “plant” action:

start state

$$\begin{array}{c|c} & \begin{array}{cc} \text{rich} & \text{poor} \end{array} \\ \begin{array}{c} \text{rich} \\ \text{poor} \end{array} & \begin{bmatrix} \quad & \quad \\ \quad & \quad \end{bmatrix} \end{array}$$

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



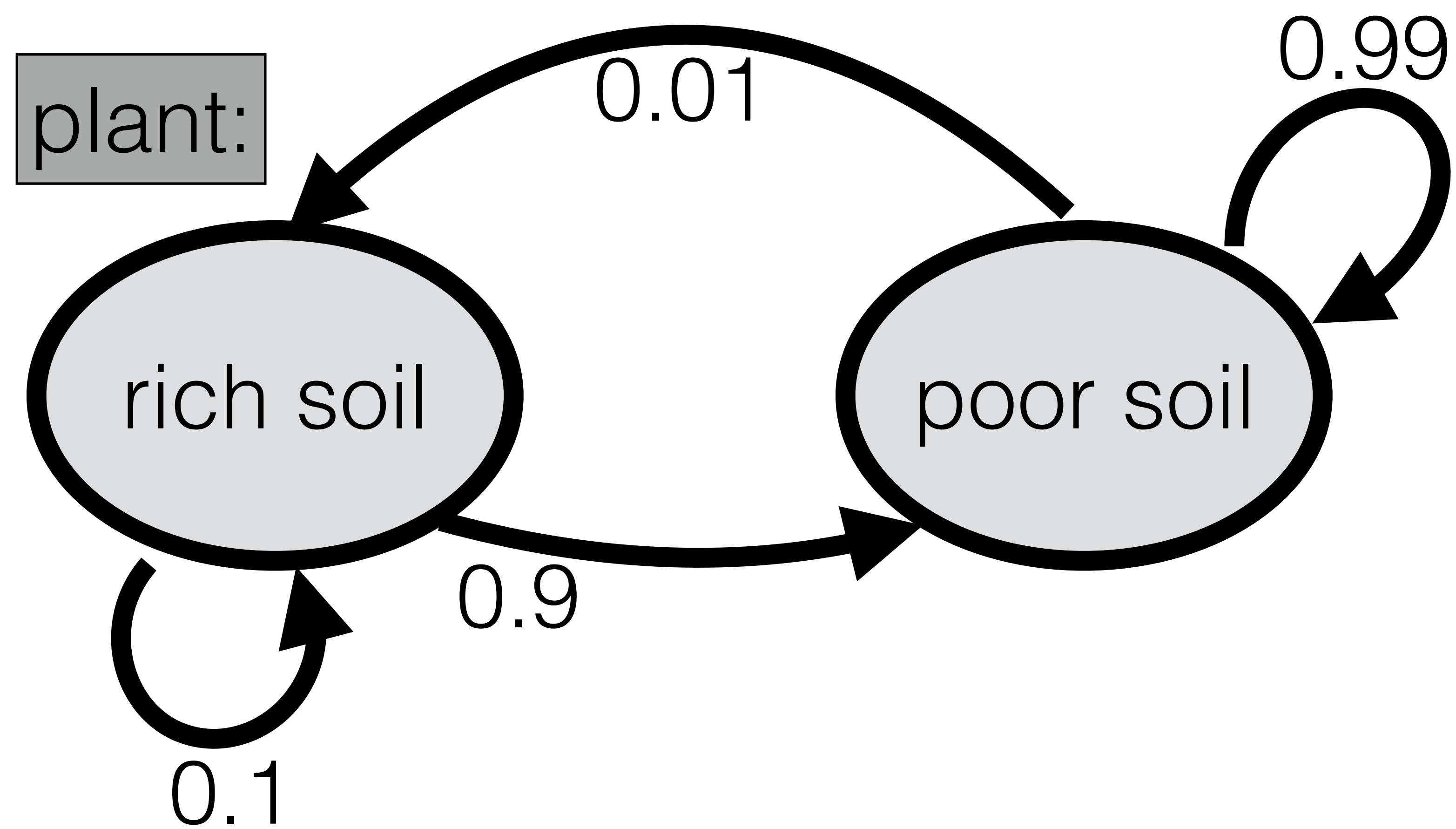

- Transition matrix for "plant" action:

*end state*

*start state*

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

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



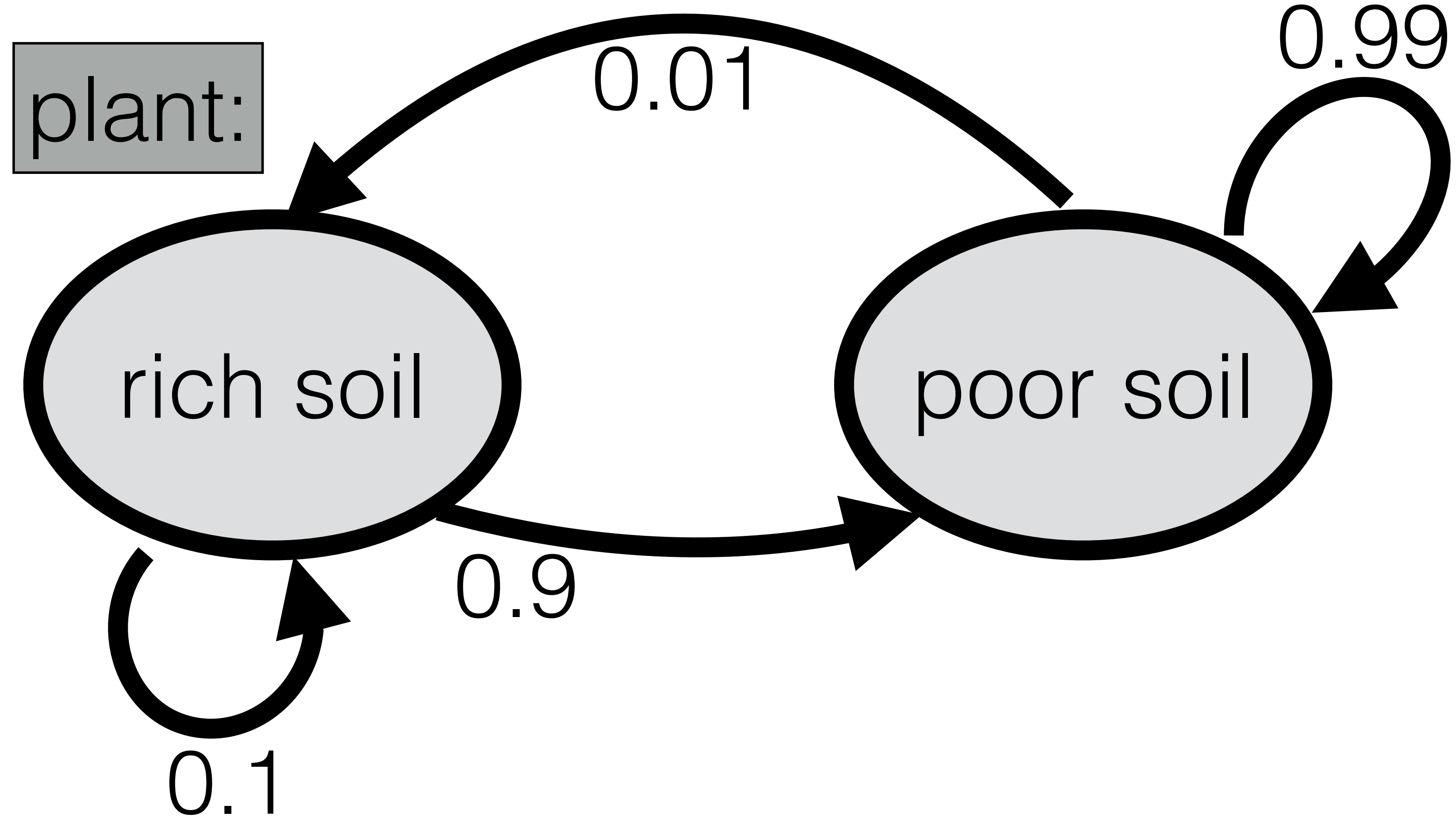

- Transition matrix for “plant” action:

*end state*

*start state*

$$\begin{array}{cc} & \begin{array}{cc} \text{rich} & \text{poor} \end{array} \\ \begin{array}{c} \text{rich} \\ \text{poor} \end{array} & \begin{bmatrix} \phantom{0.0} & \phantom{0.0} \\ \phantom{0.0} & \phantom{0.0} \end{bmatrix} \end{array}$$

- $\mathcal{S}$ = 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



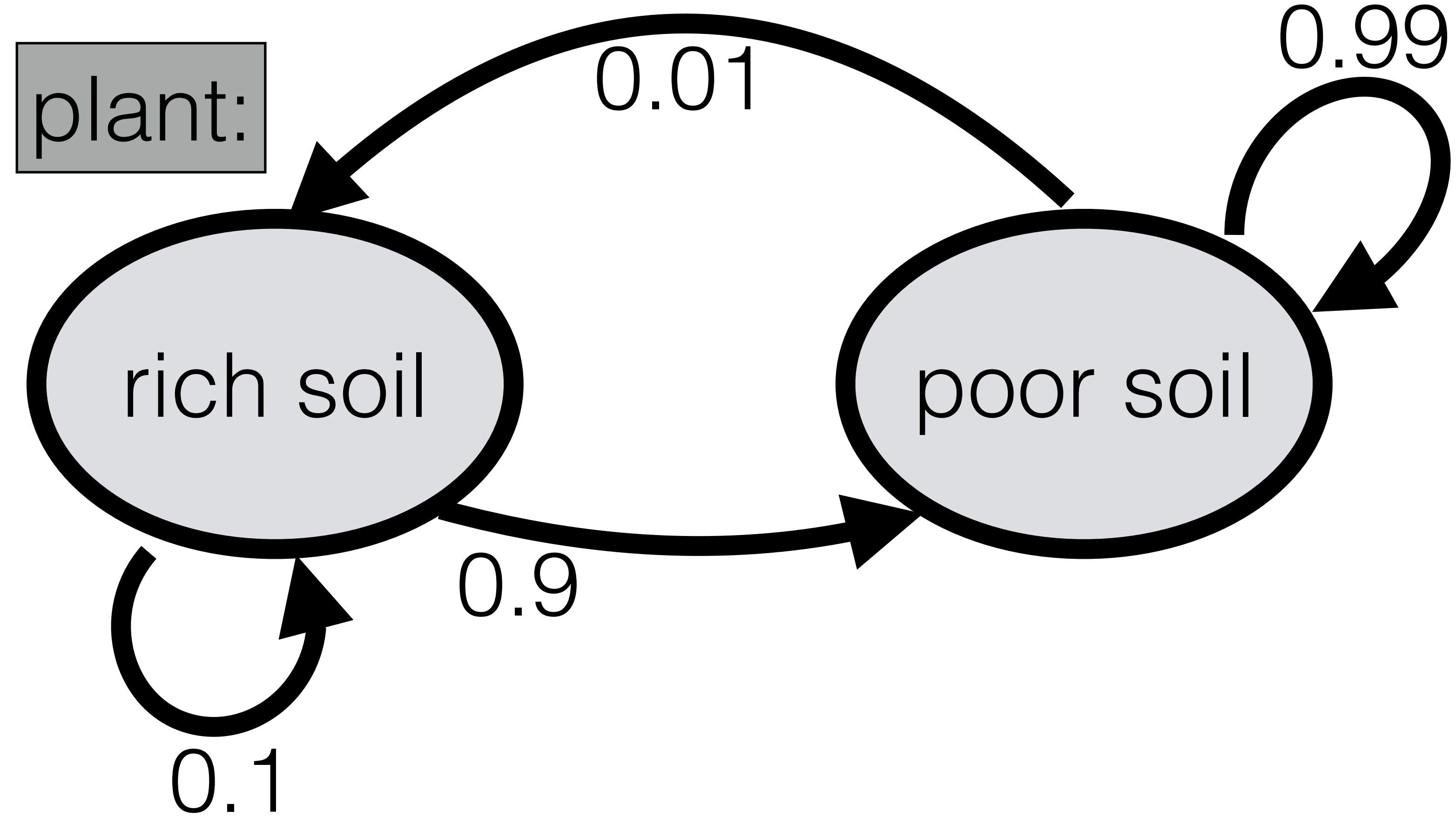

- Transition matrix for “plant” action:

*end state*

*start state*

$$\begin{array}{c c} & \begin{array}{cc} \text{rich} & \text{poor} \end{array} \\ \begin{array}{c} \text{rich} \\ \text{poor} \end{array} & \left[\begin{array}{cc} \phantom{0.9} & \boxed{0.9} \\ & \end{array}\right] \end{array}$$

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



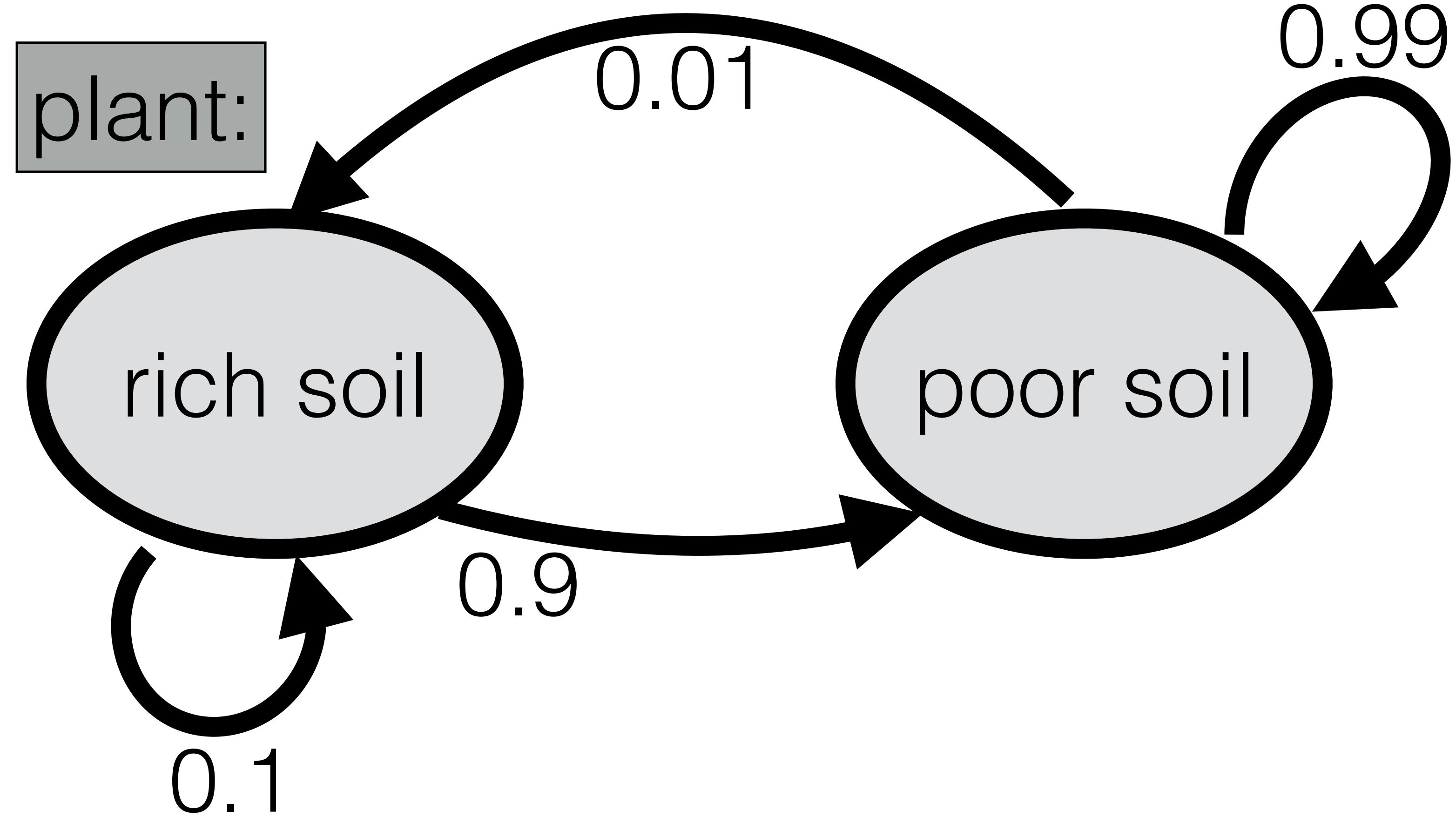

- Transition matrix for “plant” action:

end state

start state

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

- $\mathcal{S}$ = 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



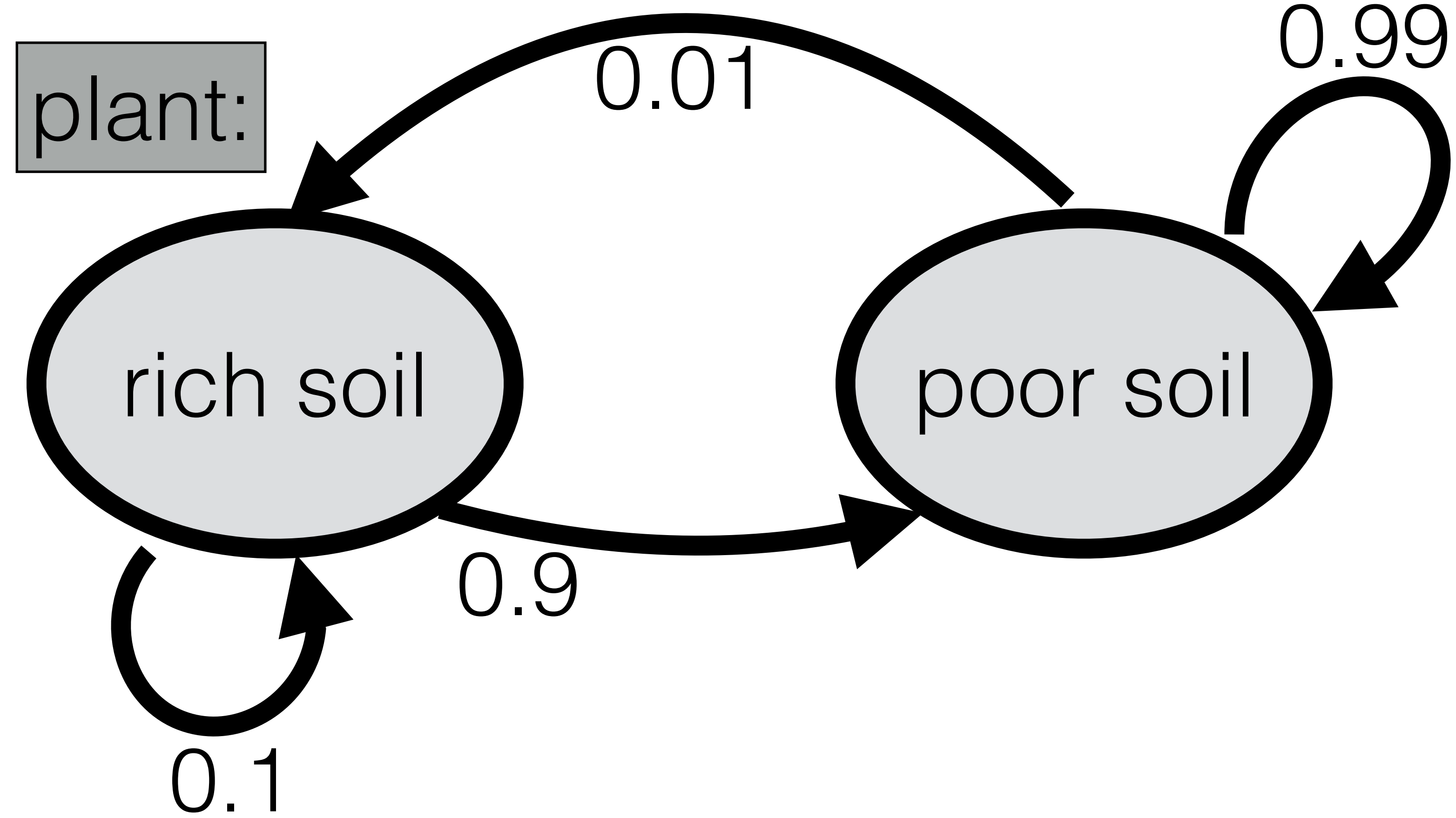

- Transition matrix for "plant" action:

*end state*

*start state*

$$\begin{array}{r c} & \begin{array}{cc} \text{rich} & \text{poor} \end{array} \\ \begin{array}{r} \text{rich} \\ \text{poor} \end{array} & \begin{bmatrix} 0.1 & 0.9 \\ 0.01 & 0.99 \end{bmatrix} \end{array}$$

- $\mathcal{S}$ = 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



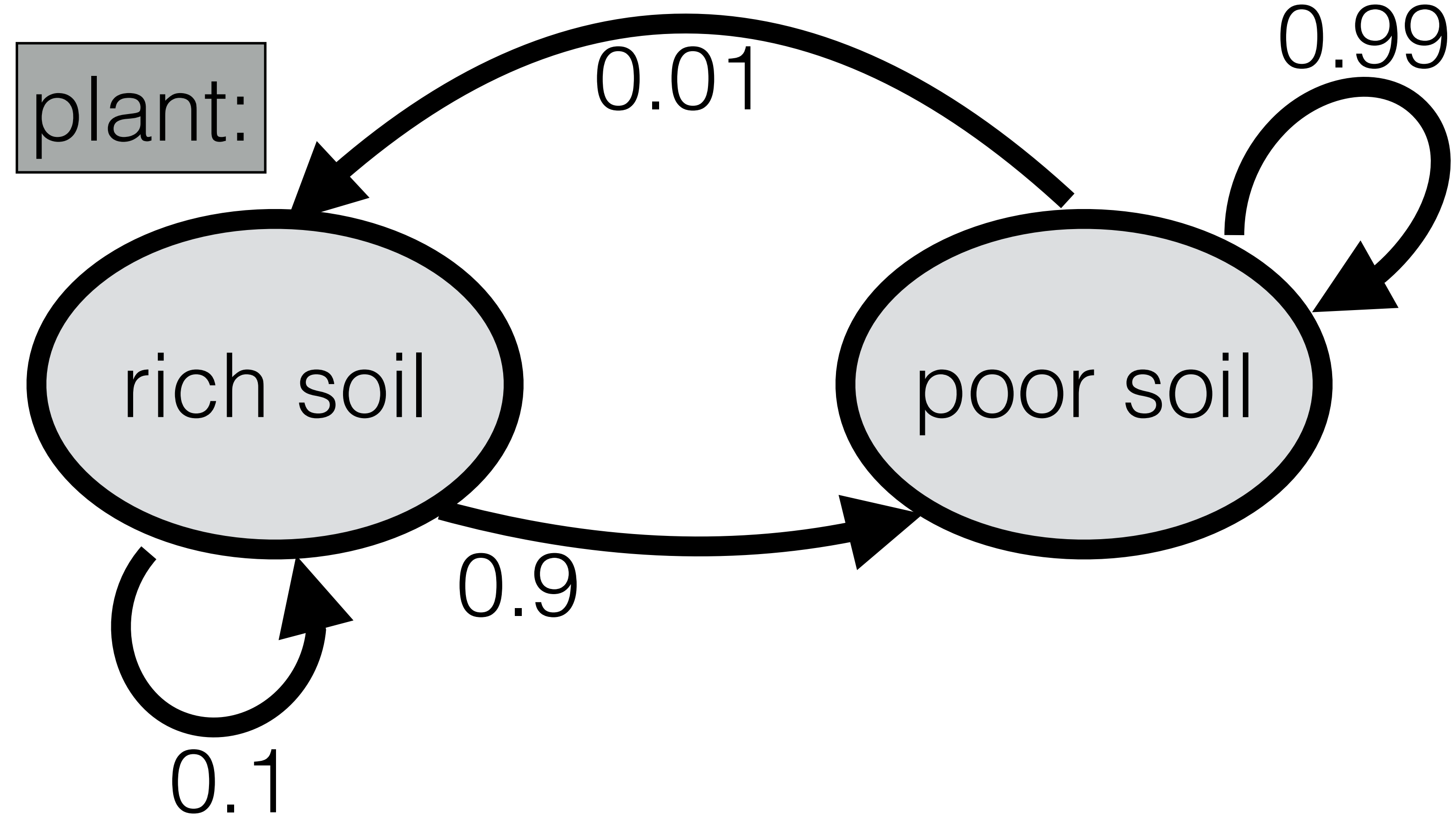

- Transition matrix for “plant” action:

*end state*

*start state*

$$\begin{array}{cc} & \begin{array}{cc} \text{rich} & \text{poor} \end{array} \\ \begin{array}{c} \text{rich} \\ \text{poor} \end{array} & \begin{bmatrix} 0.1 & 0.9 \\ 0.01 & 0.99 \end{bmatrix} \end{array}$$

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



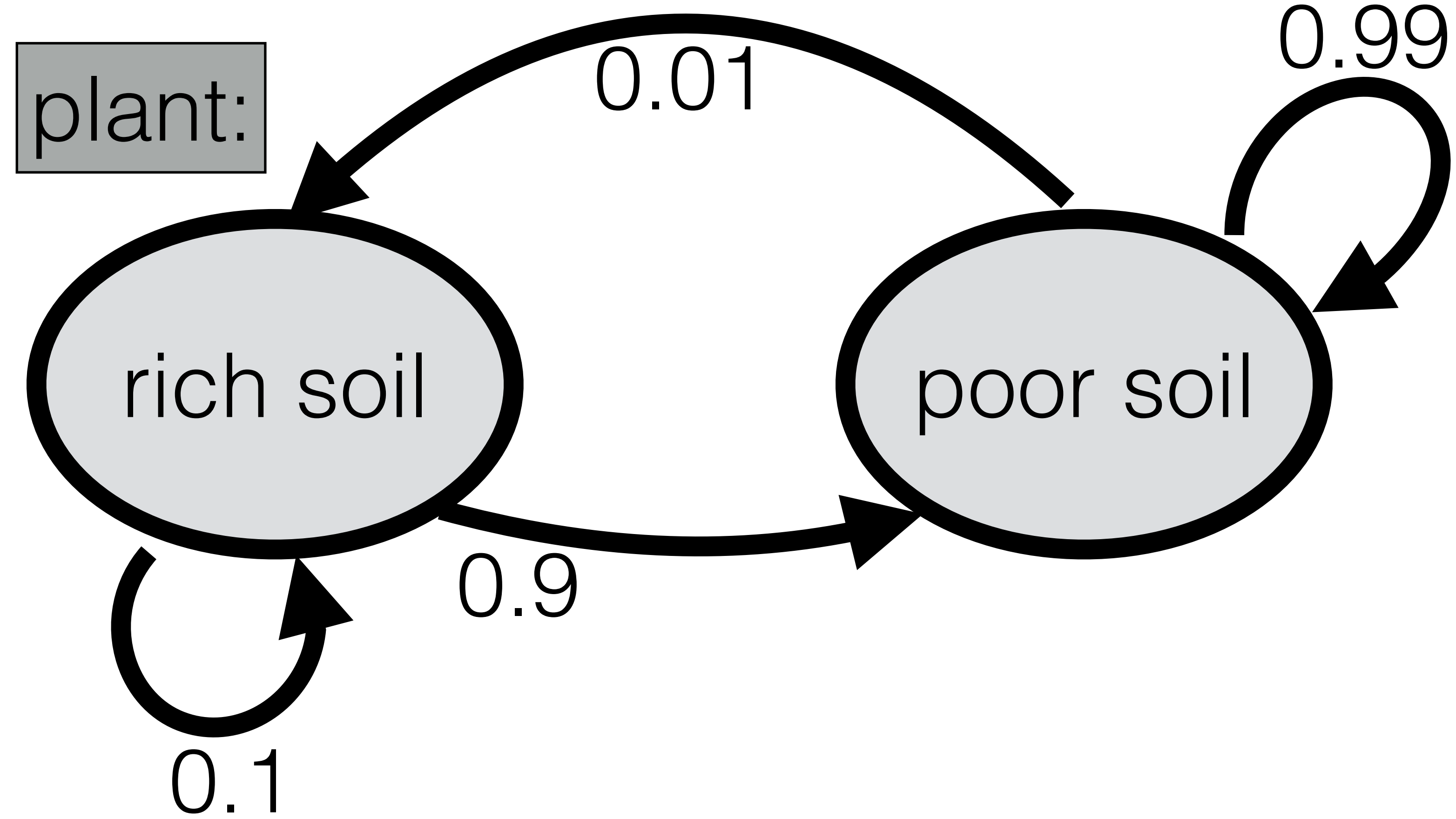

- Transition matrix for "plant" action:

*end state*

*start state*

$$\begin{array}{r} \\ \text{rich} \\ \text{poor} \end{array} \begin{array}{c} \begin{array}{cc} \text{rich} & \text{poor} \end{array} \\ \begin{bmatrix} \boxed{0.1} & 0.9 \\ \boxed{0.01} & 0.99 \end{bmatrix} \end{array}$$

- $\mathcal{S}$ = 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



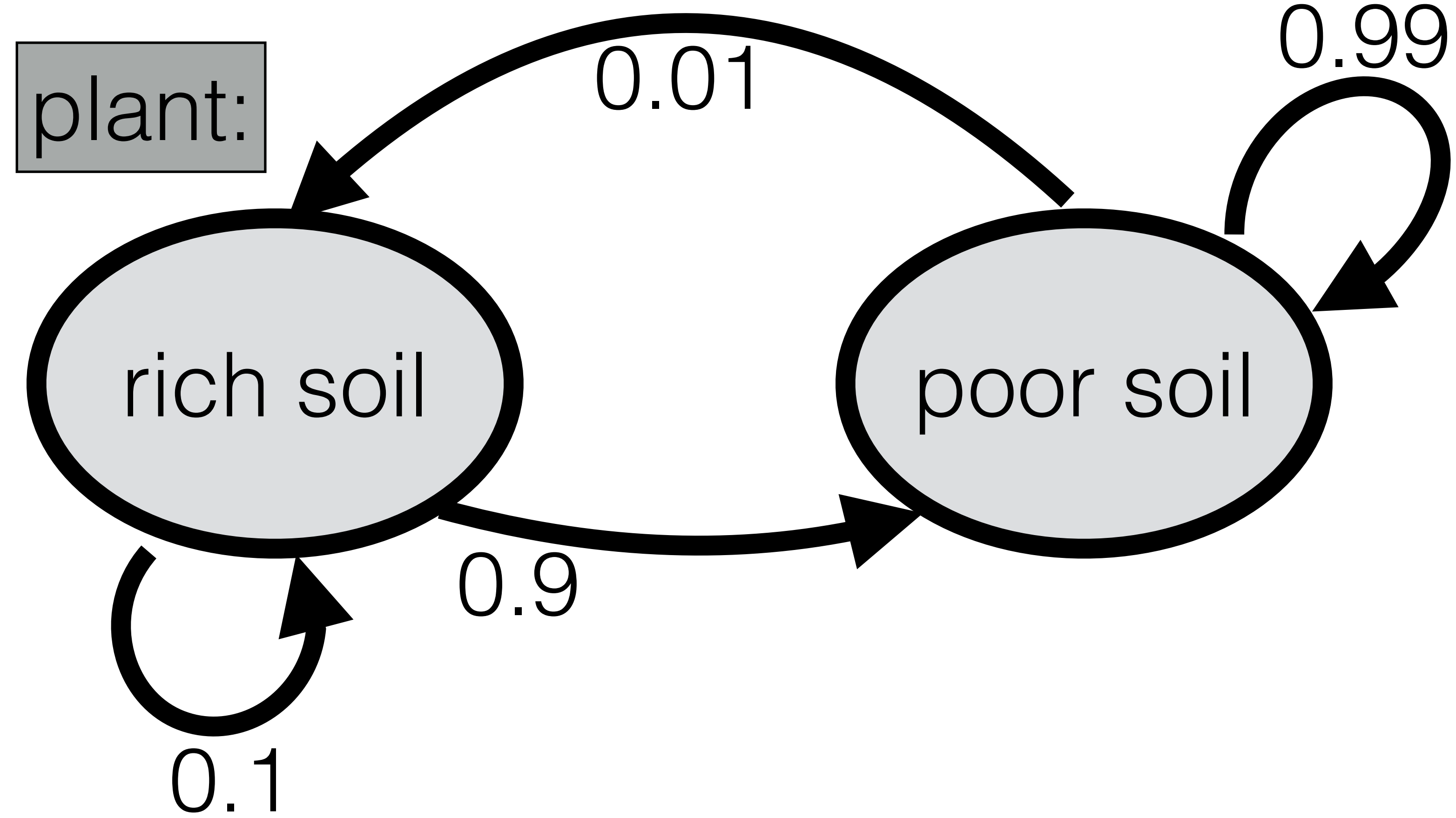

- Transition matrix for “plant” action:

*end state*

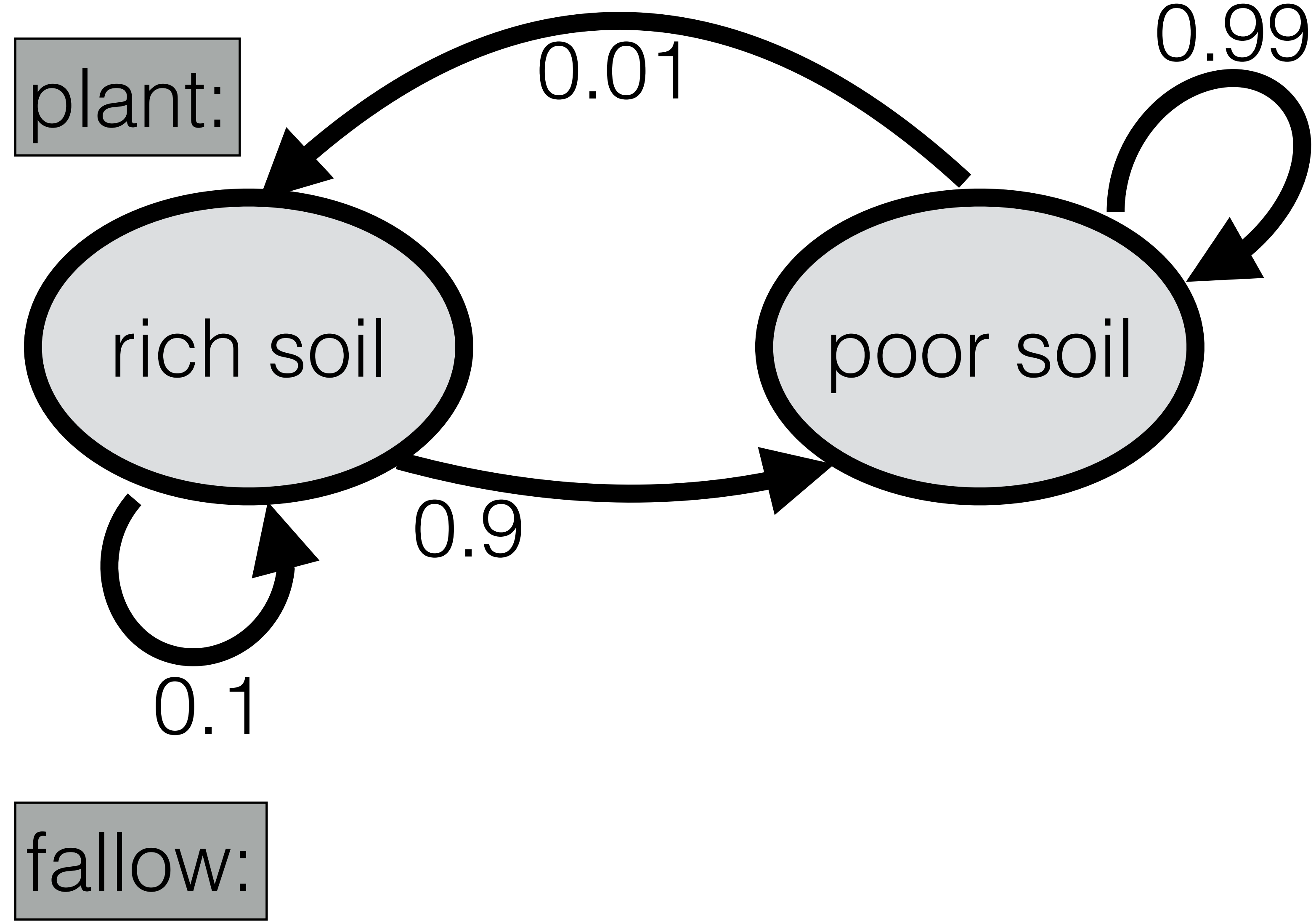
*start state*

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

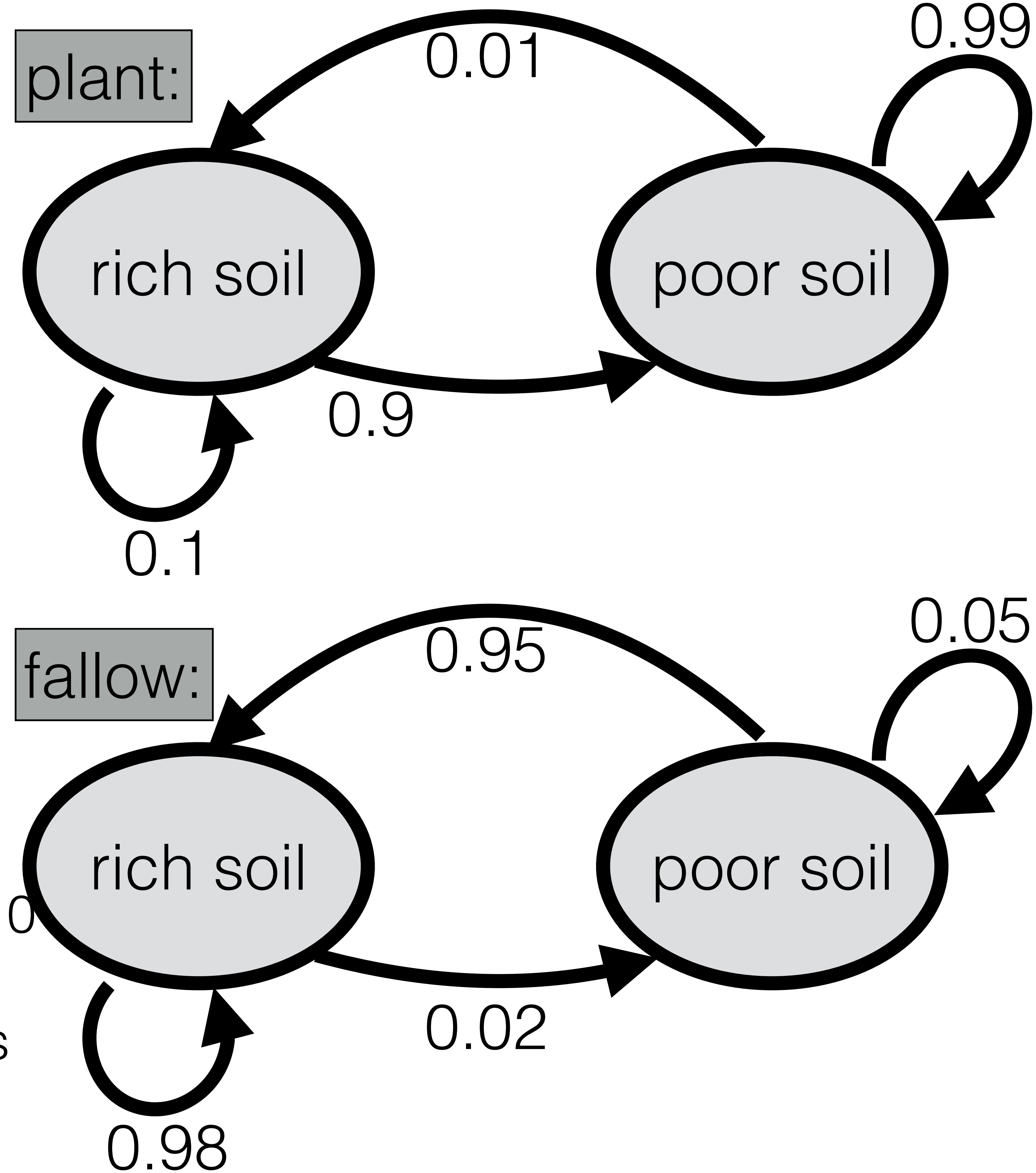
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$ transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) =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



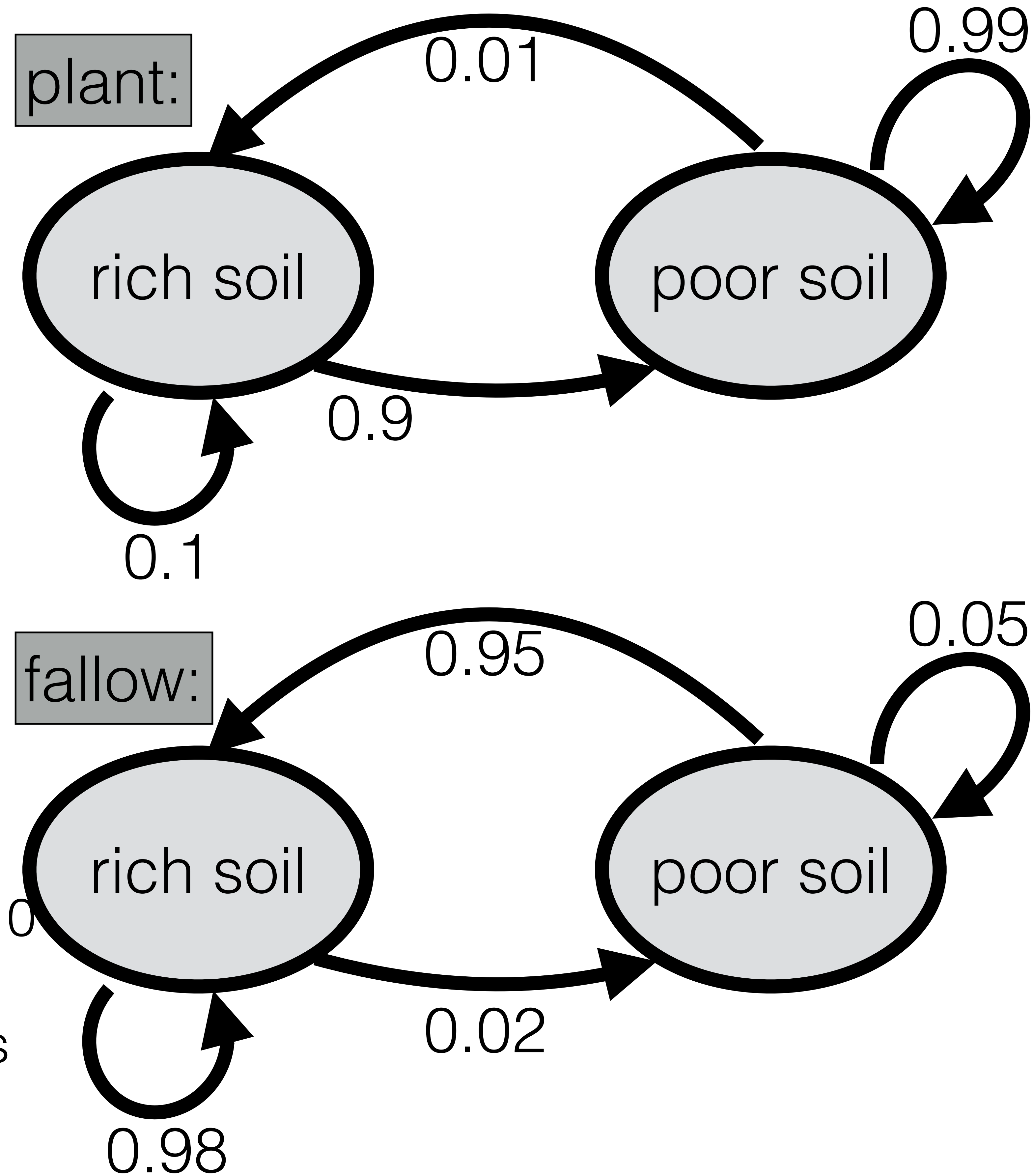
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$
  transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$:
  reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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



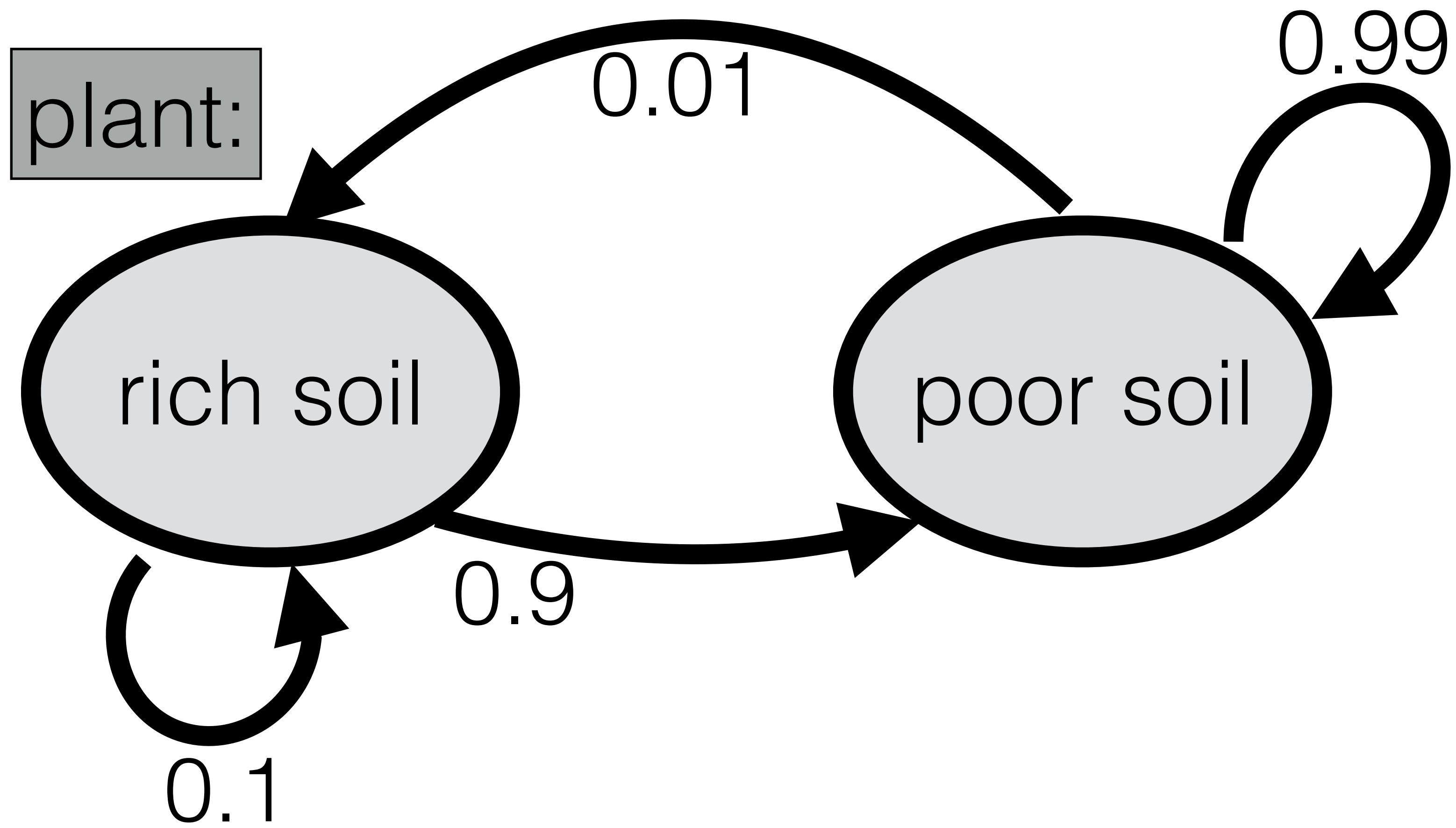
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$
  transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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

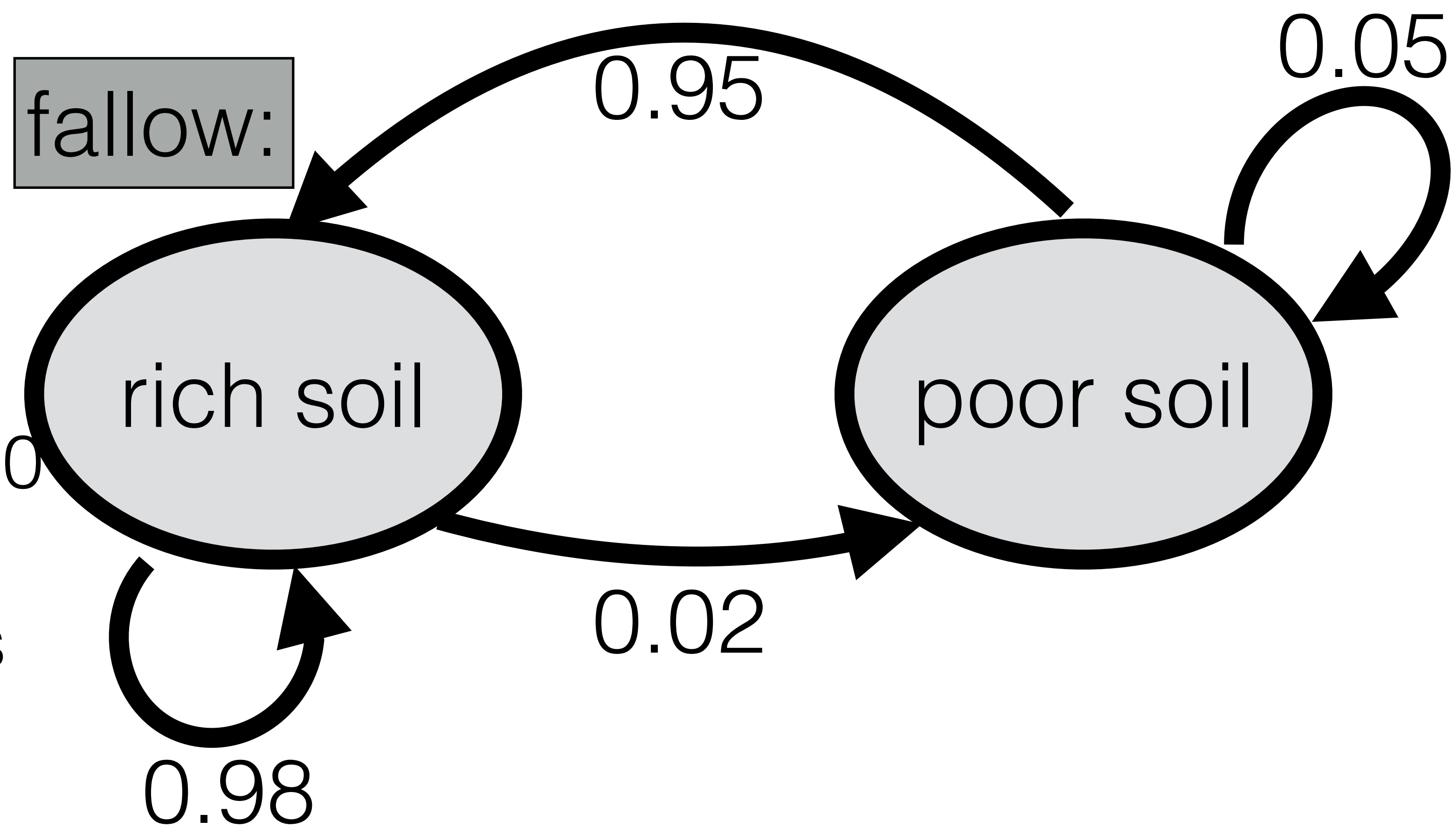


- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T$ transition model
- $R : \mathcal{S} \times \mathcal{X} \rightarrow \mathbb{R}$: reward function
    - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich,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

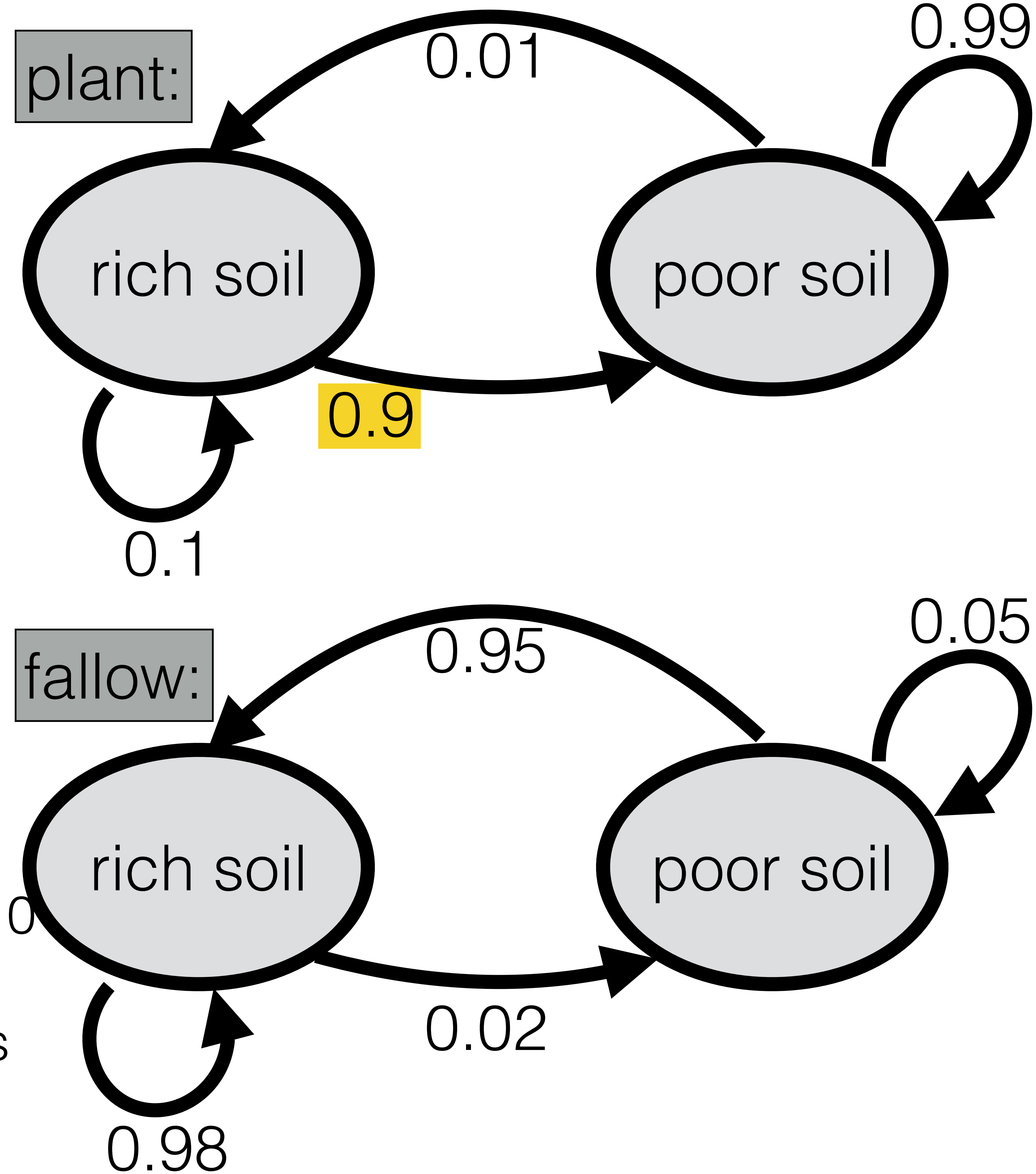


- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich,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

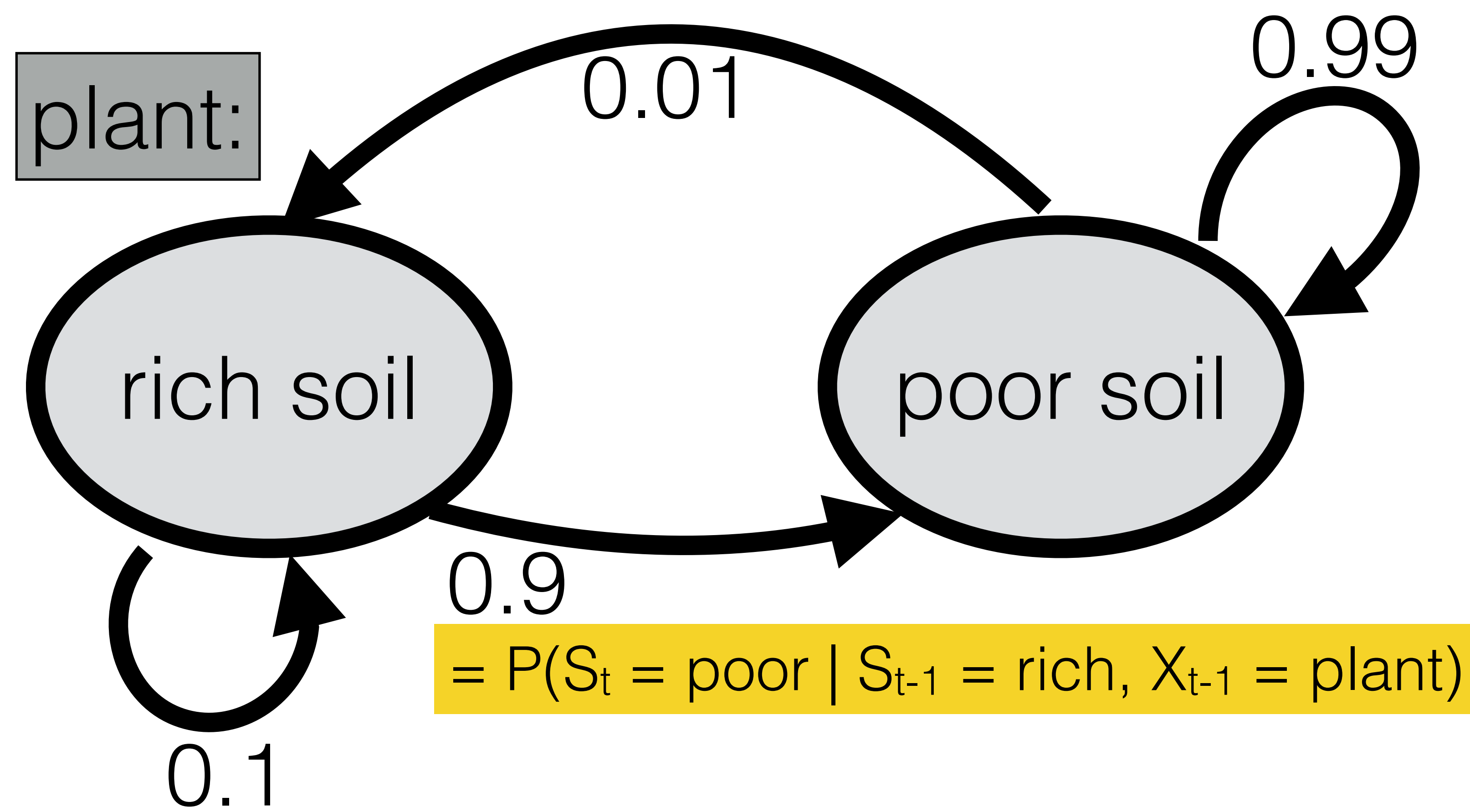


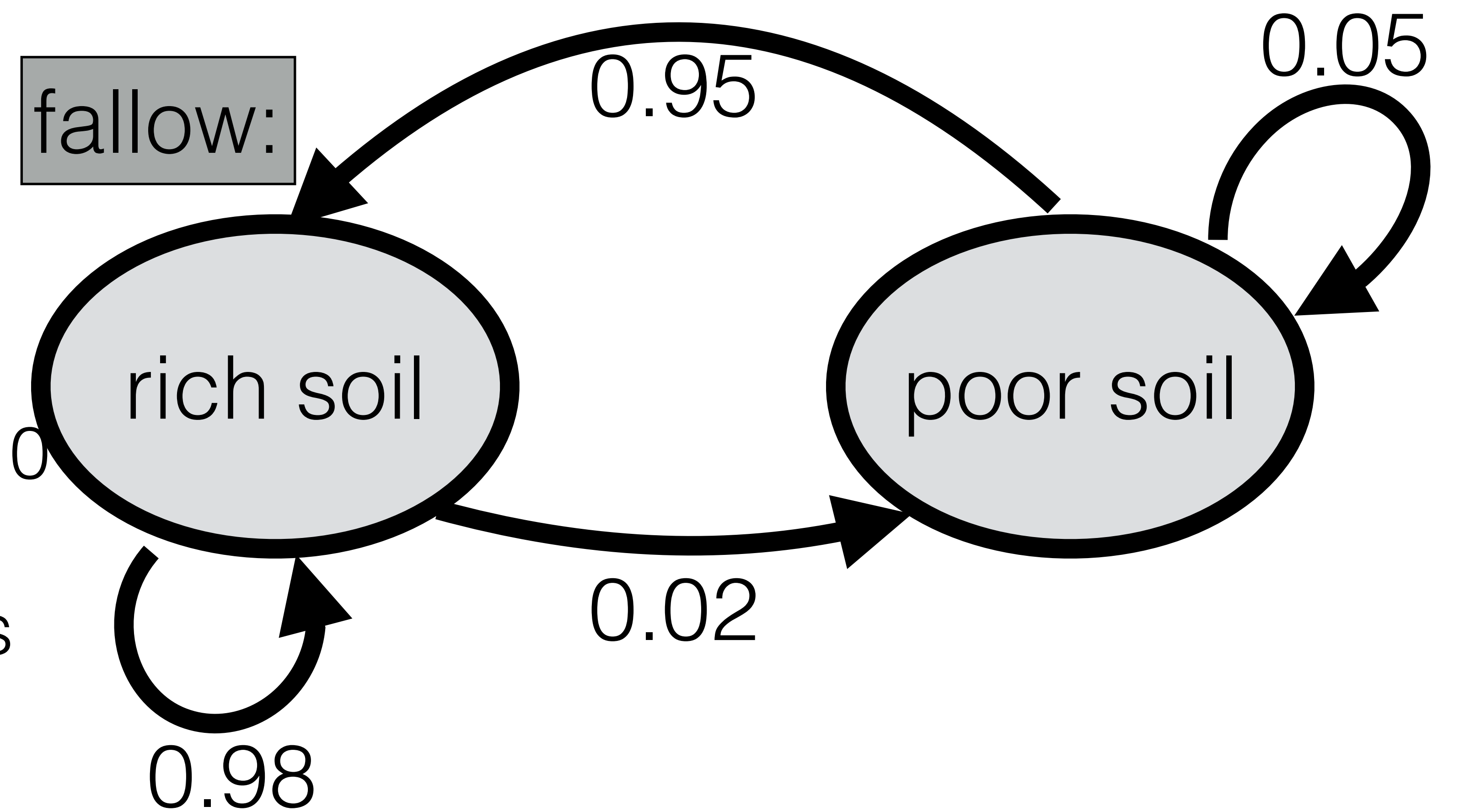



- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich,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

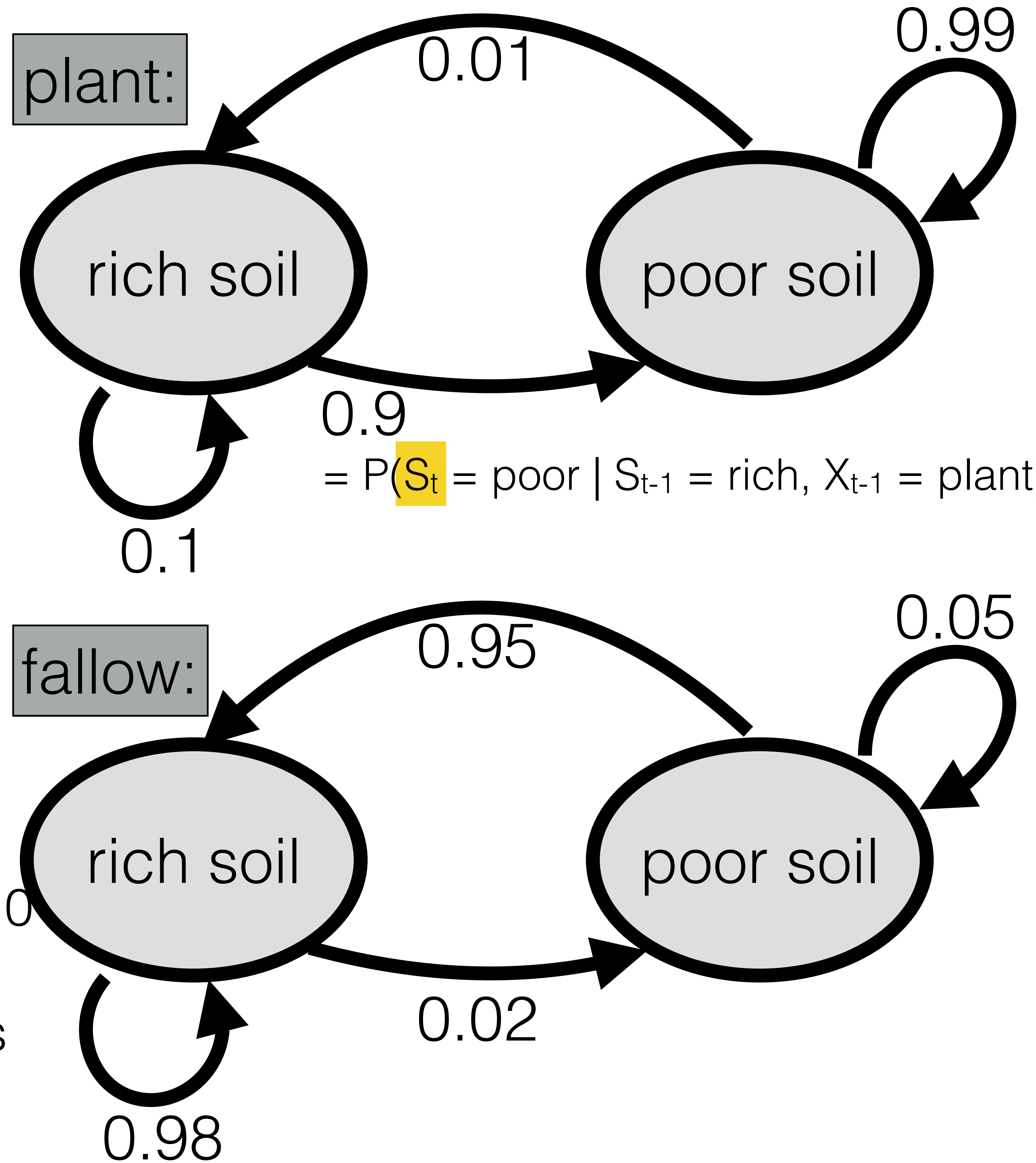


- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich,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

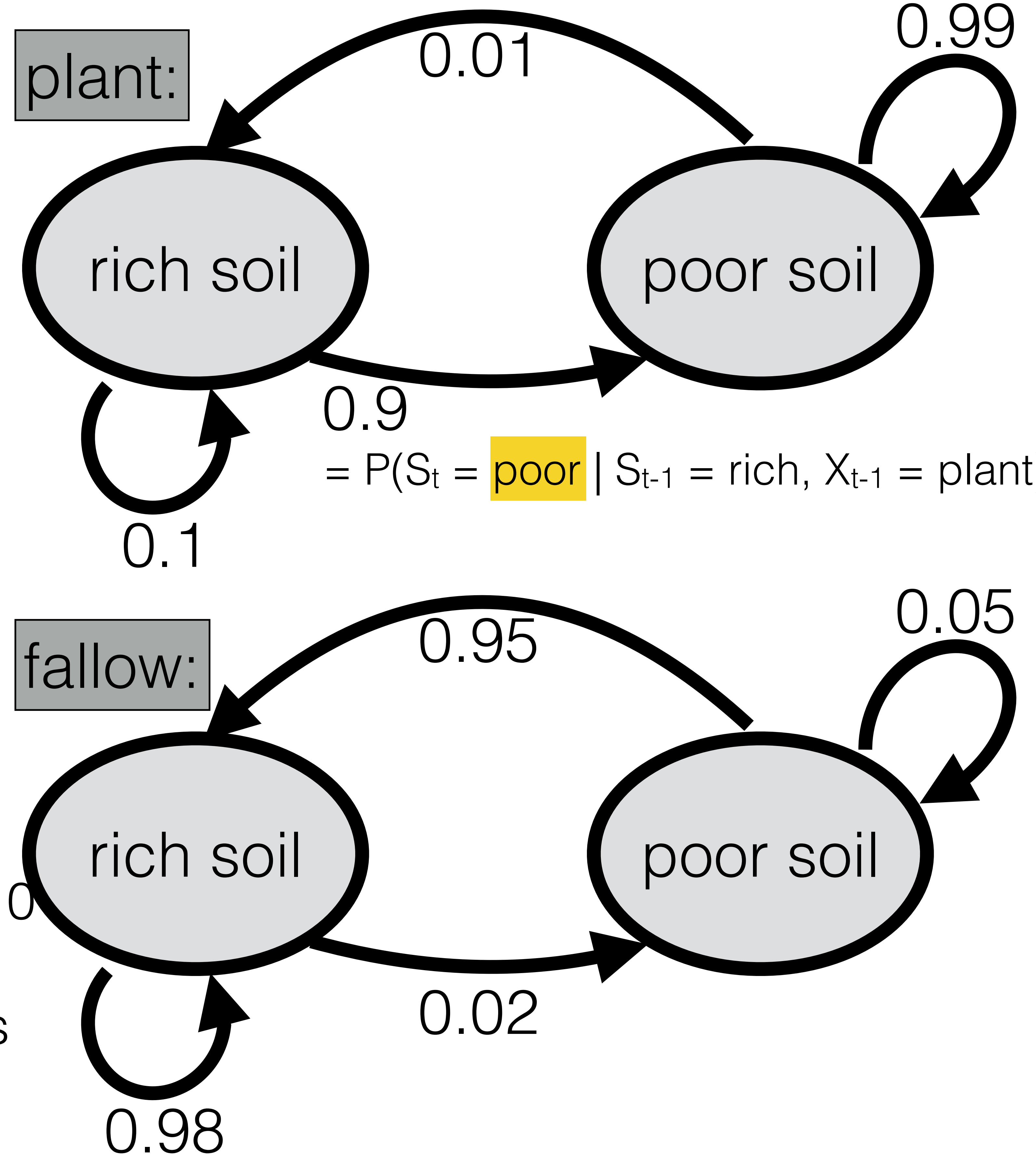


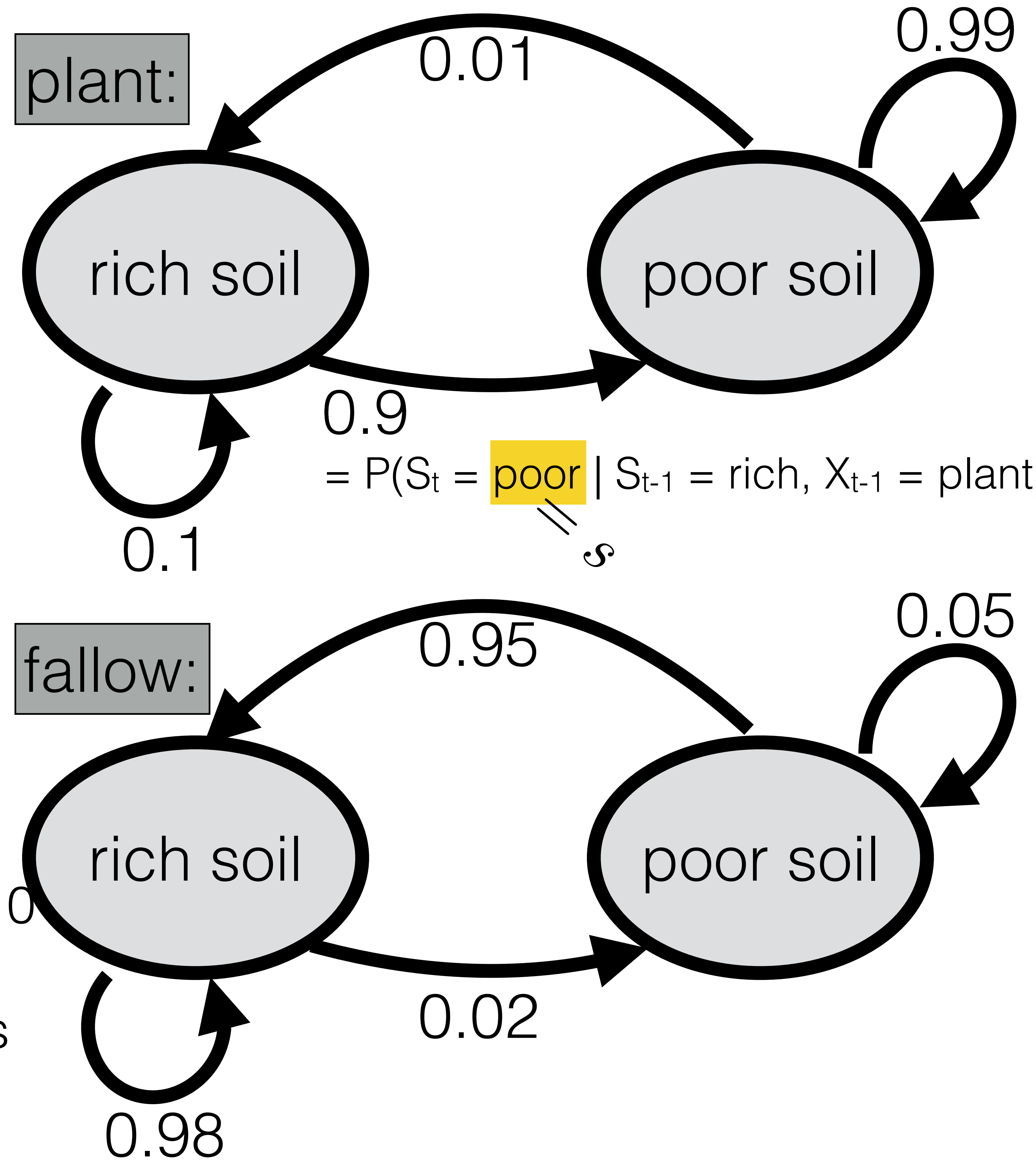
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
    - e.g. R(rich,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



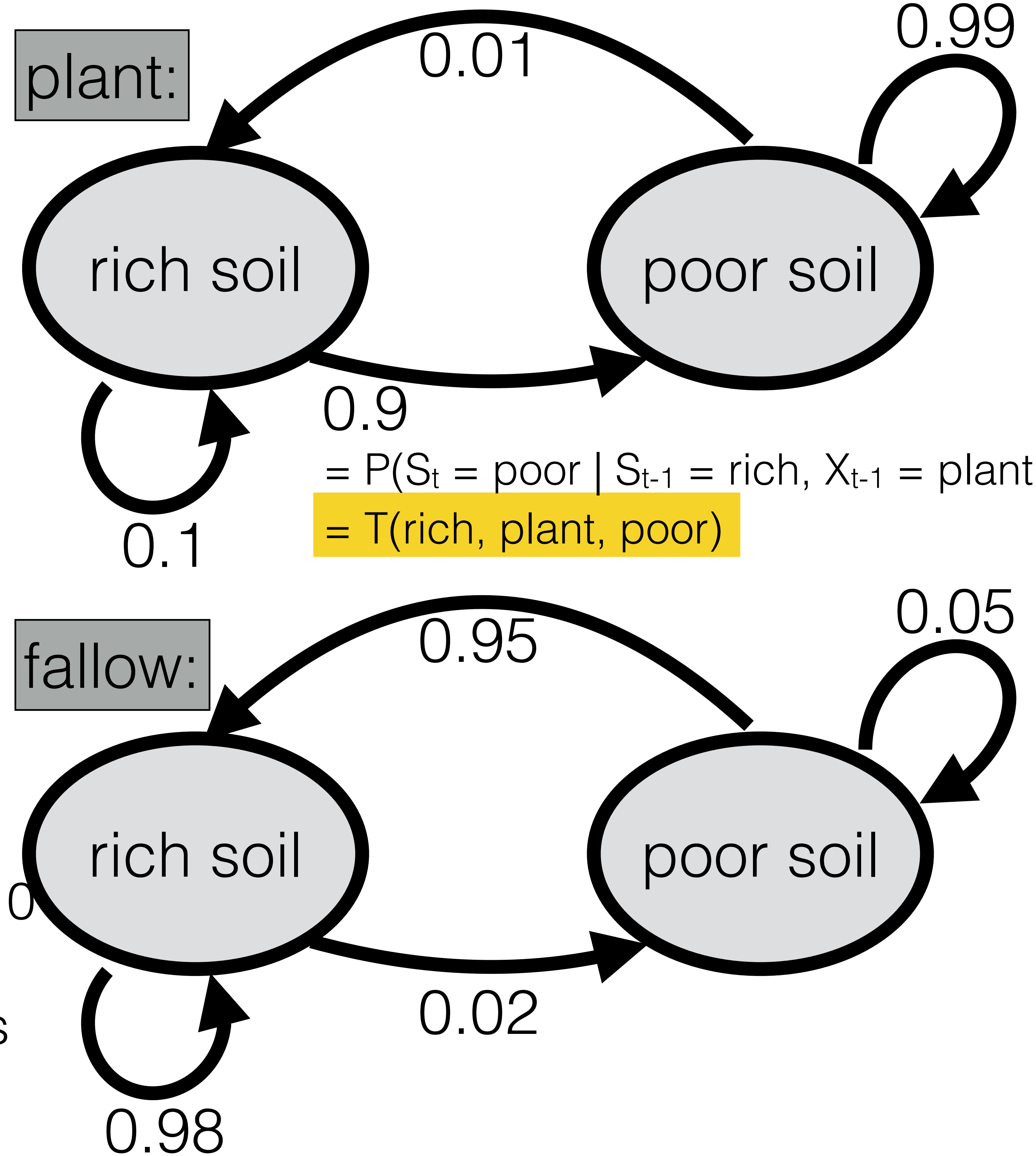
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich,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



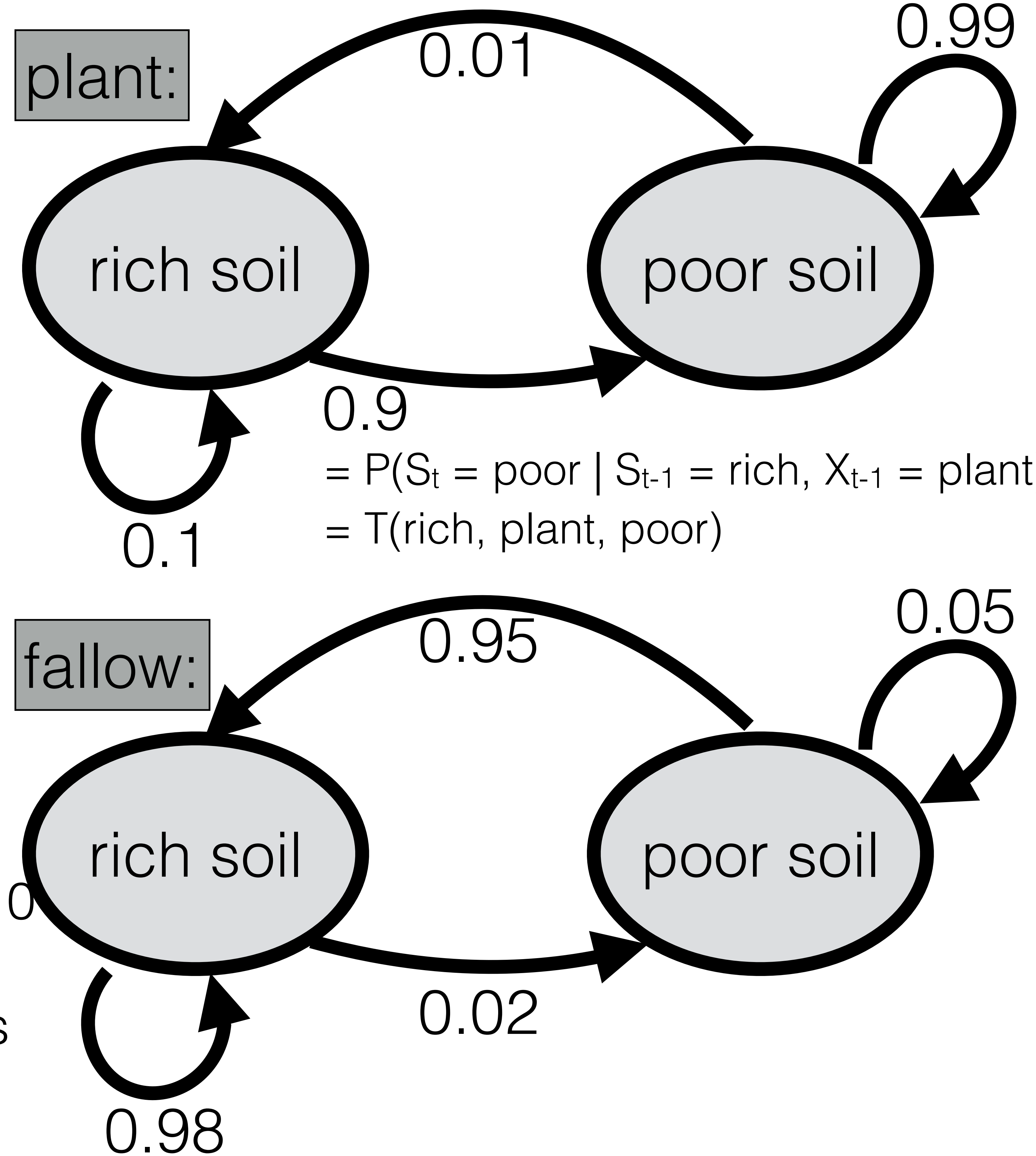
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich,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



- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: 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

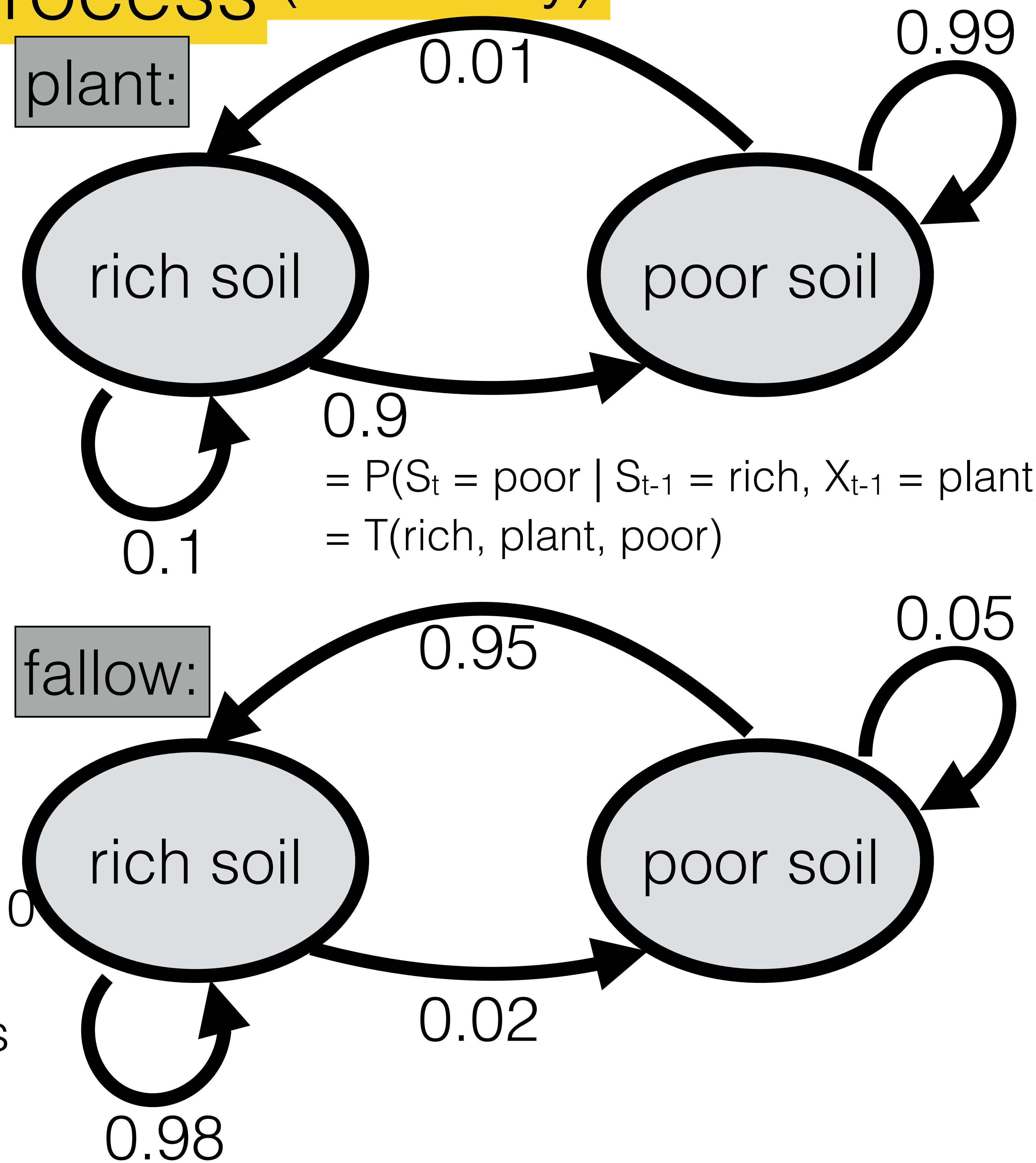


- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
    - e.g. R(rich,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



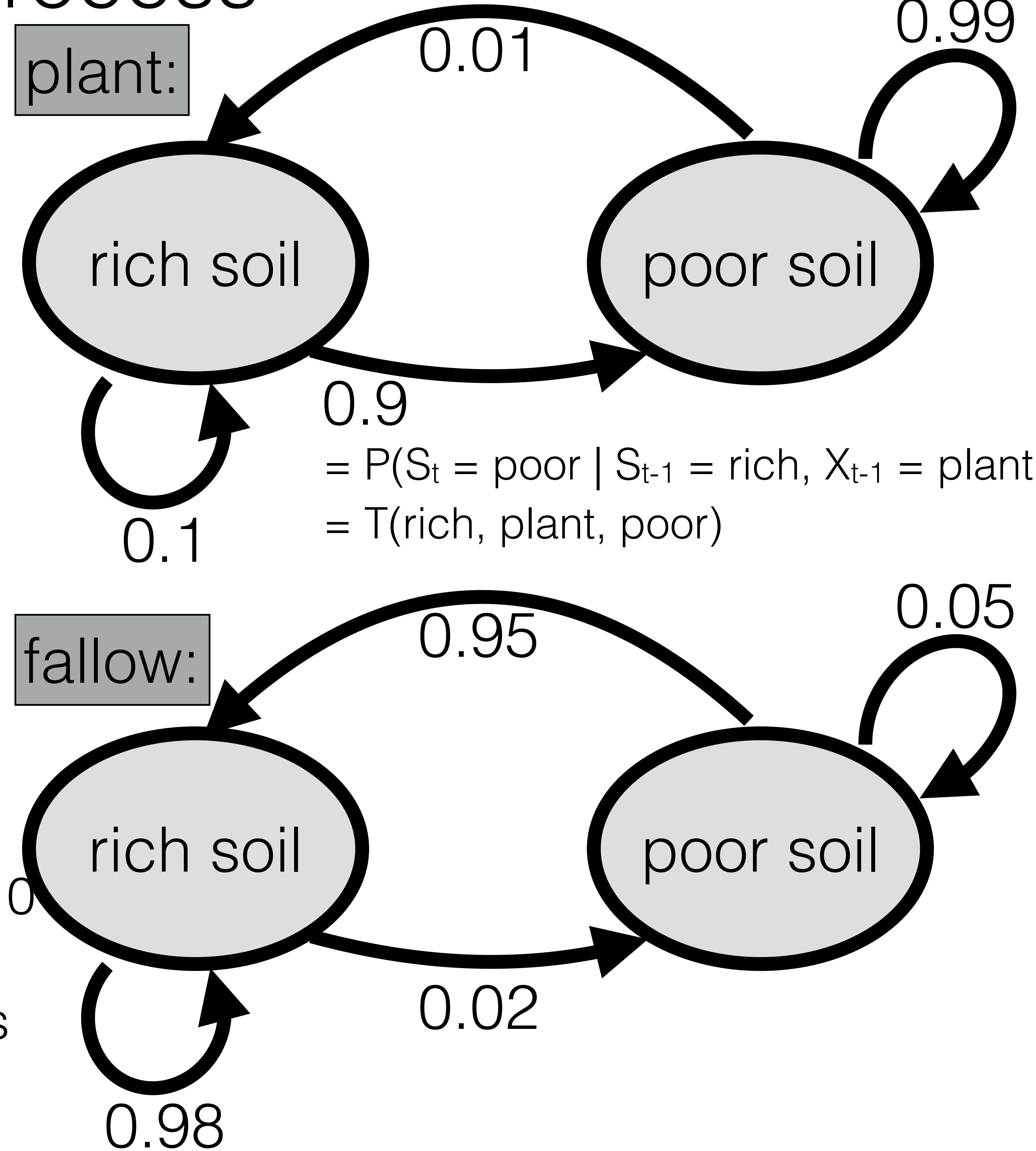
# Markov Decision Process (basically)

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$ : 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



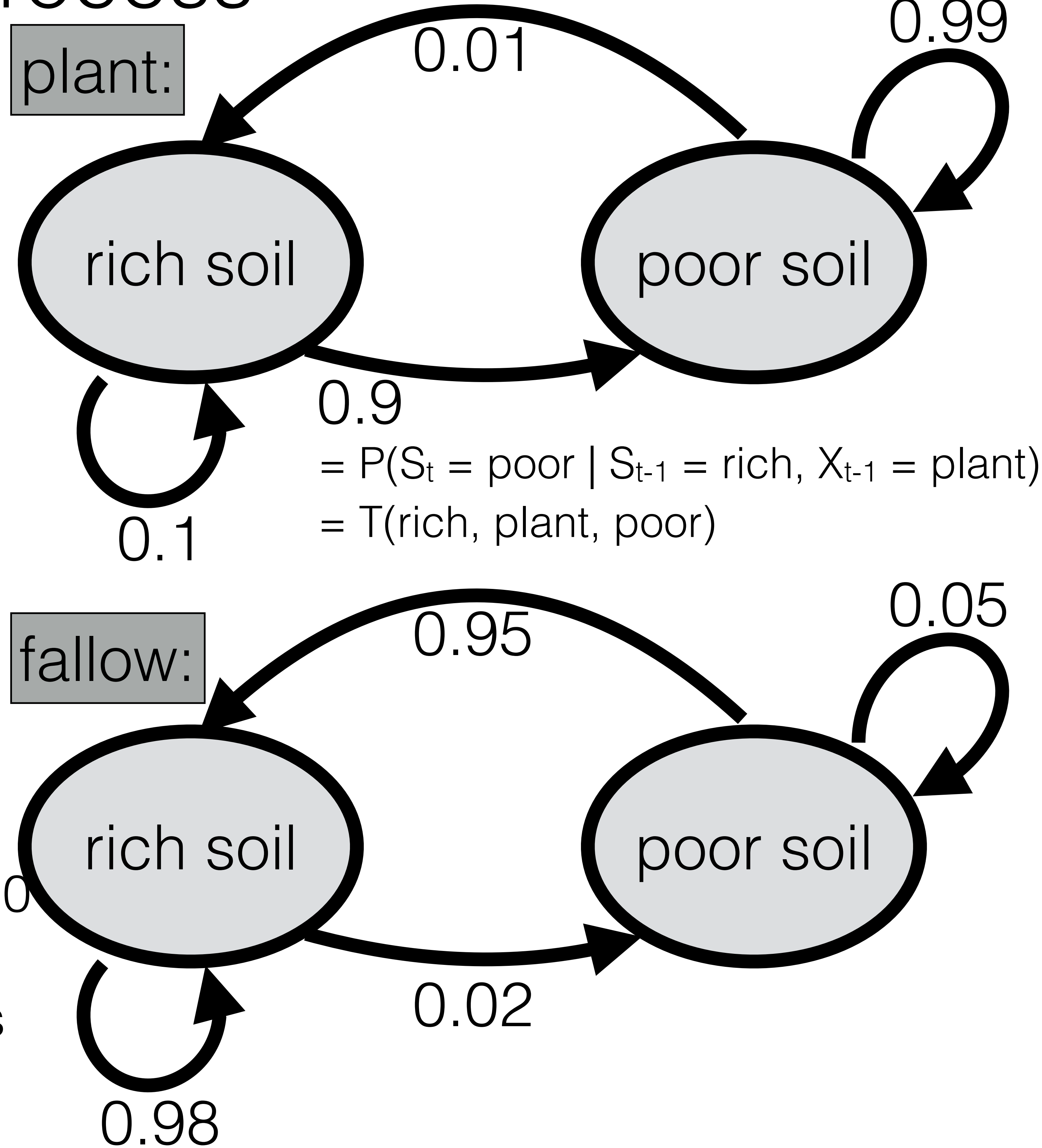
# Markov Decision Process (basically)

- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$ : initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: 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



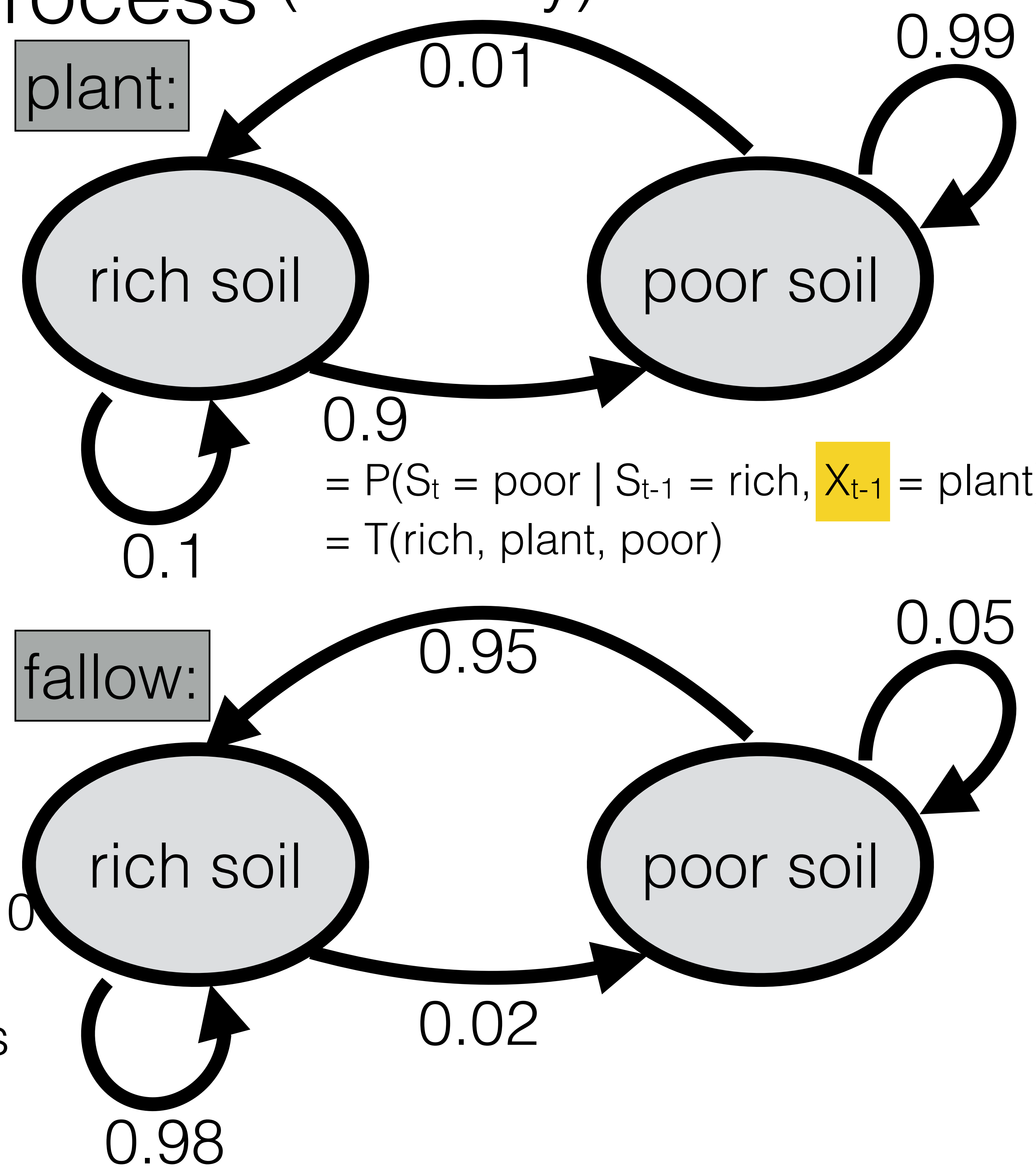
# Markov Decision Process (basically)

- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $s_0 \in \mathcal{S}$ : initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: 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



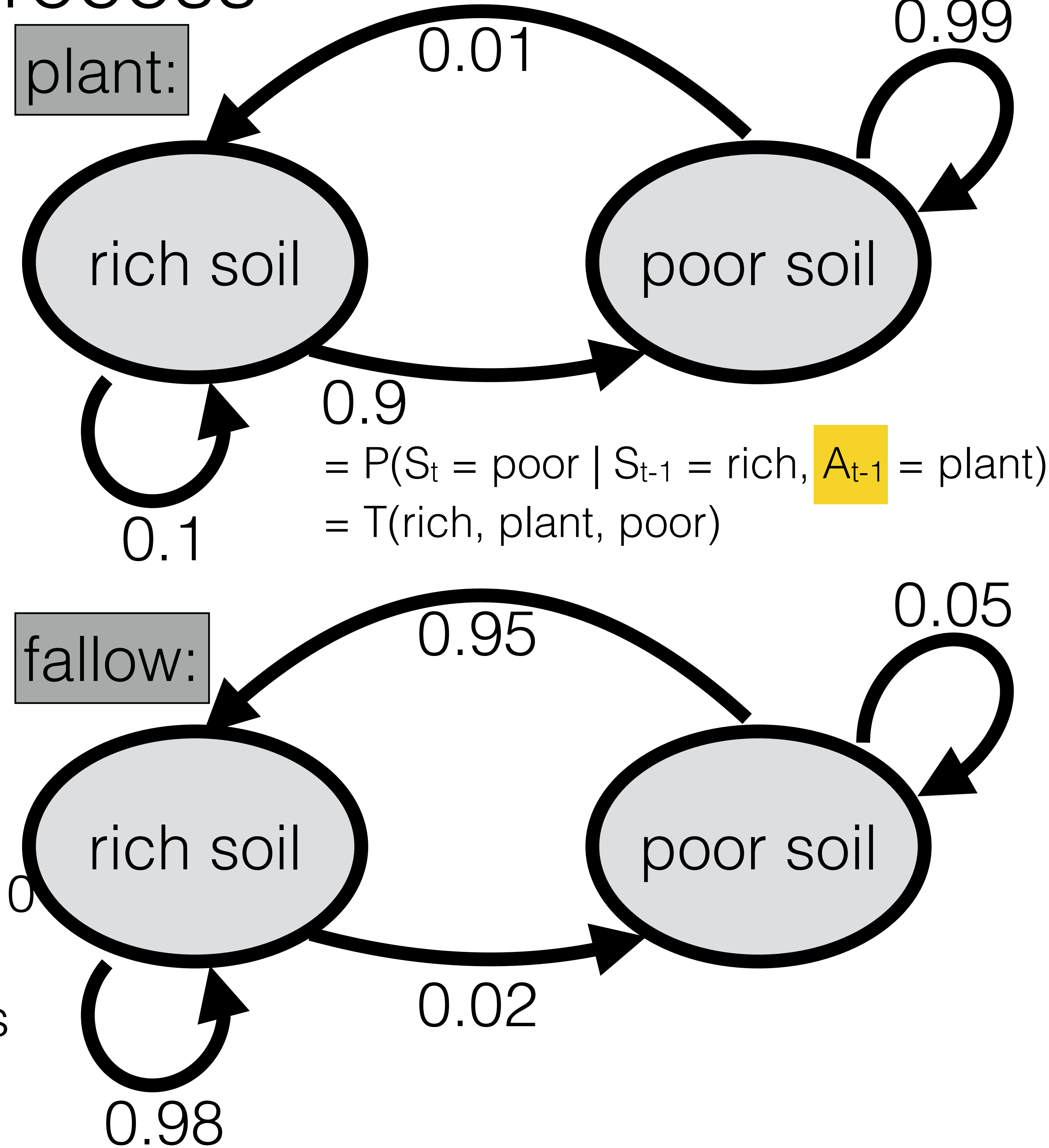
# Markov Decision Process (basically)

- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $s_0 \in \mathcal{S}$ : initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: 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



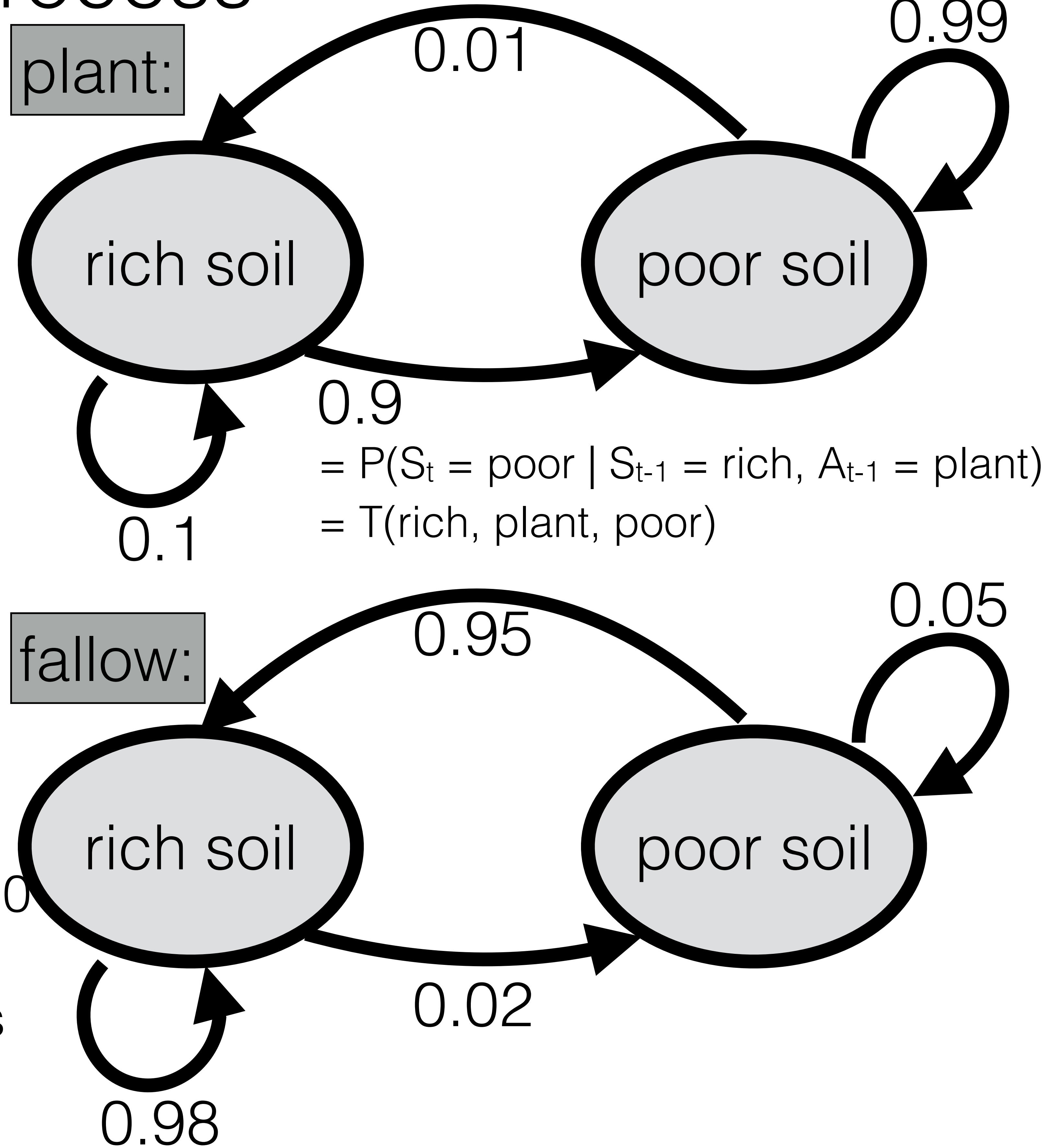
# Markov Decision Process (basically)

- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $s_0 \in \mathcal{S}$ : initial state
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ : 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



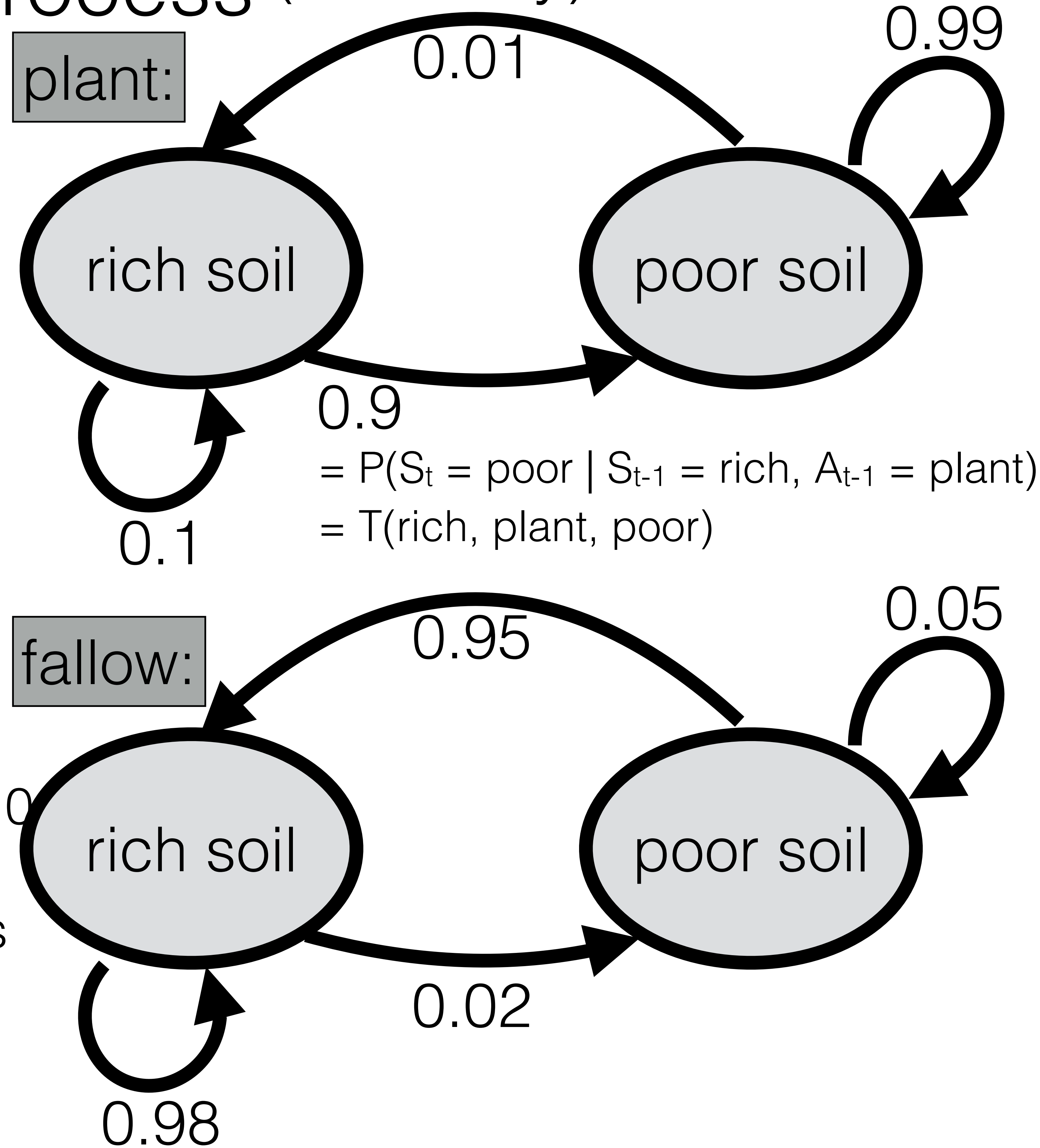
# Markov Decision Process (basically)

- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $s_0 \in \mathcal{S}$ : initial state
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ : 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



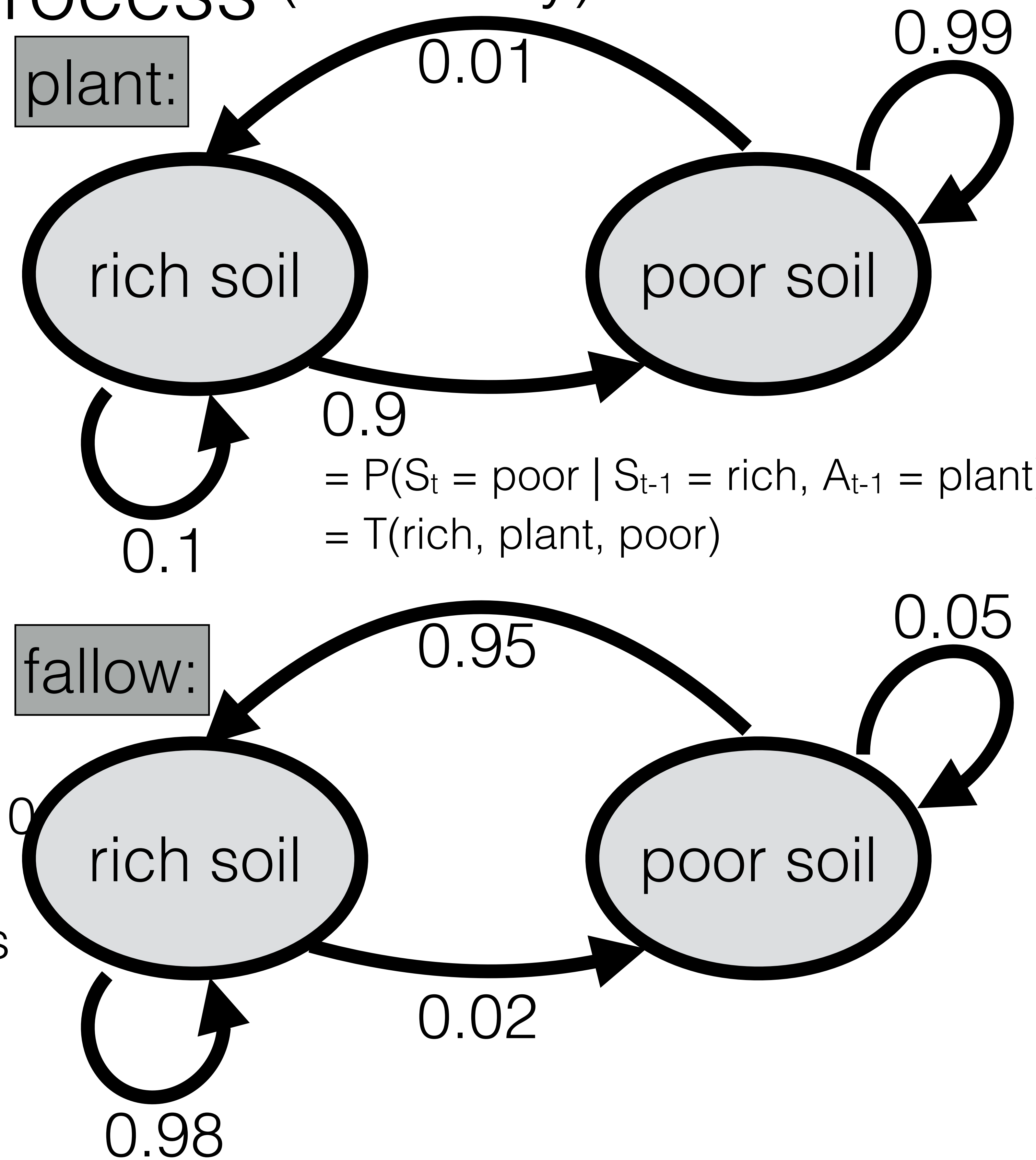
# Markov Decision Process (basically)

- $\mathcal{S}$ = 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



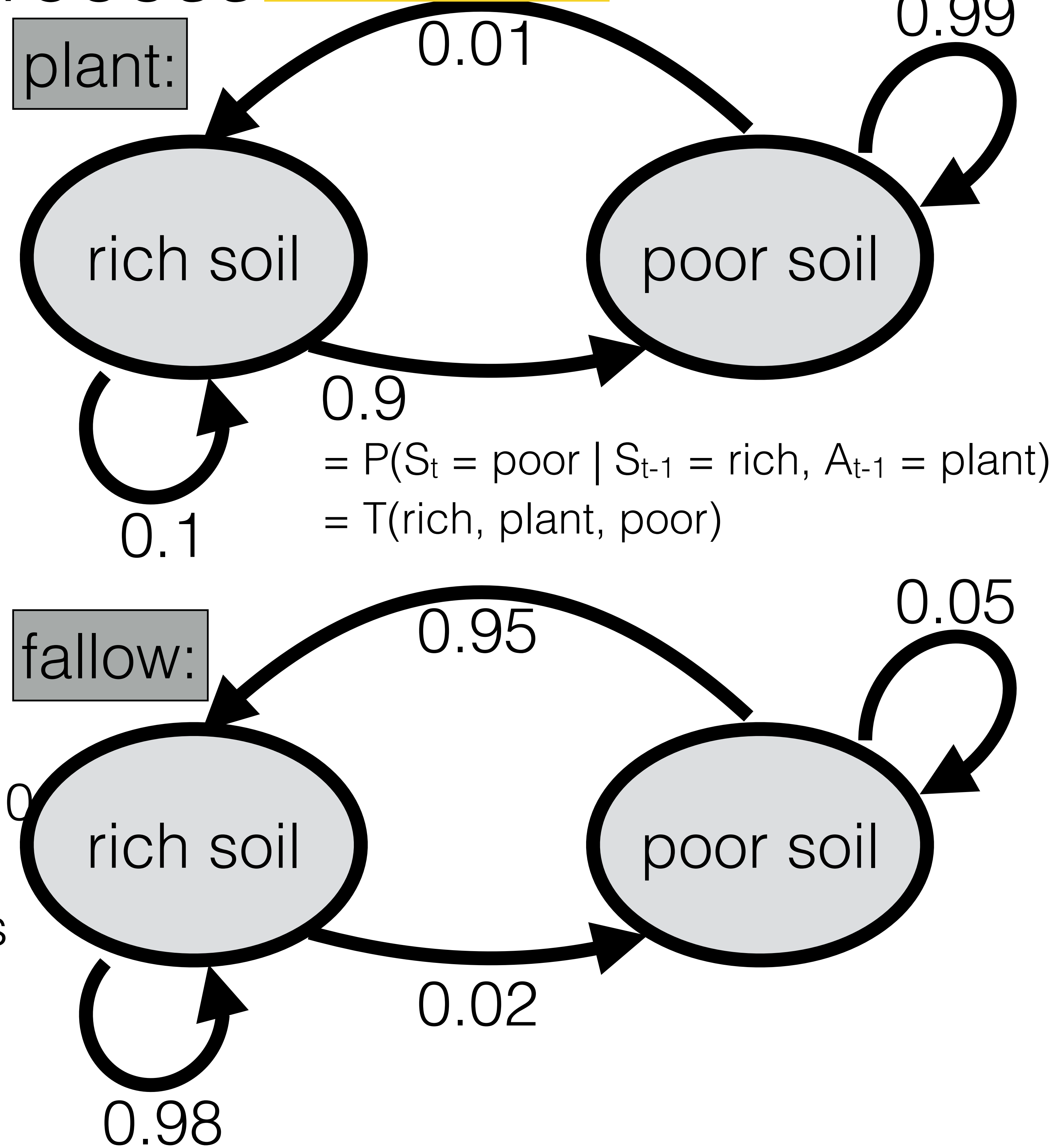
# Markov Decision Process (basically)

- $\mathcal{S}$ = 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 (basically)

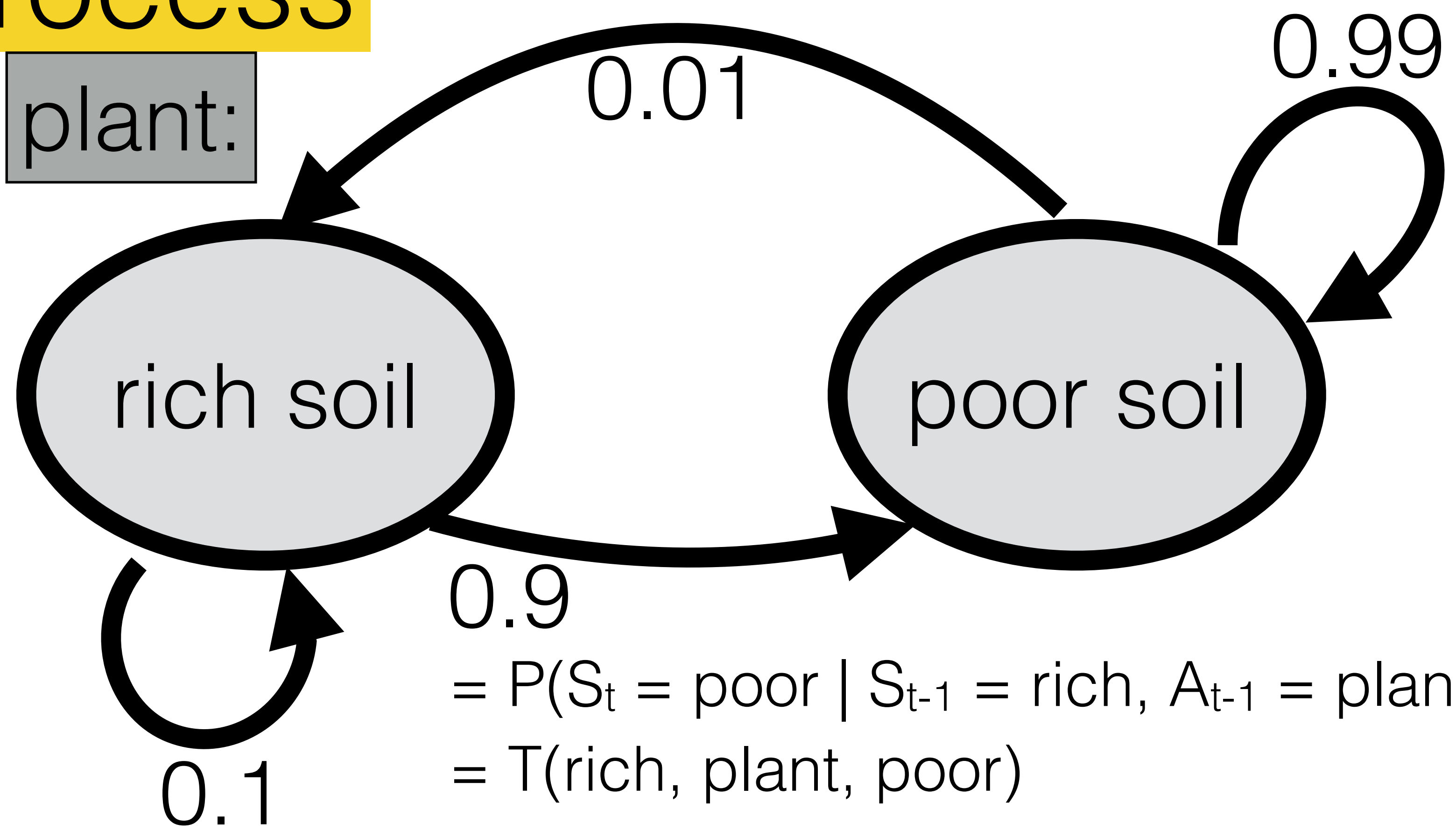
- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) =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



# Markov Decision Process

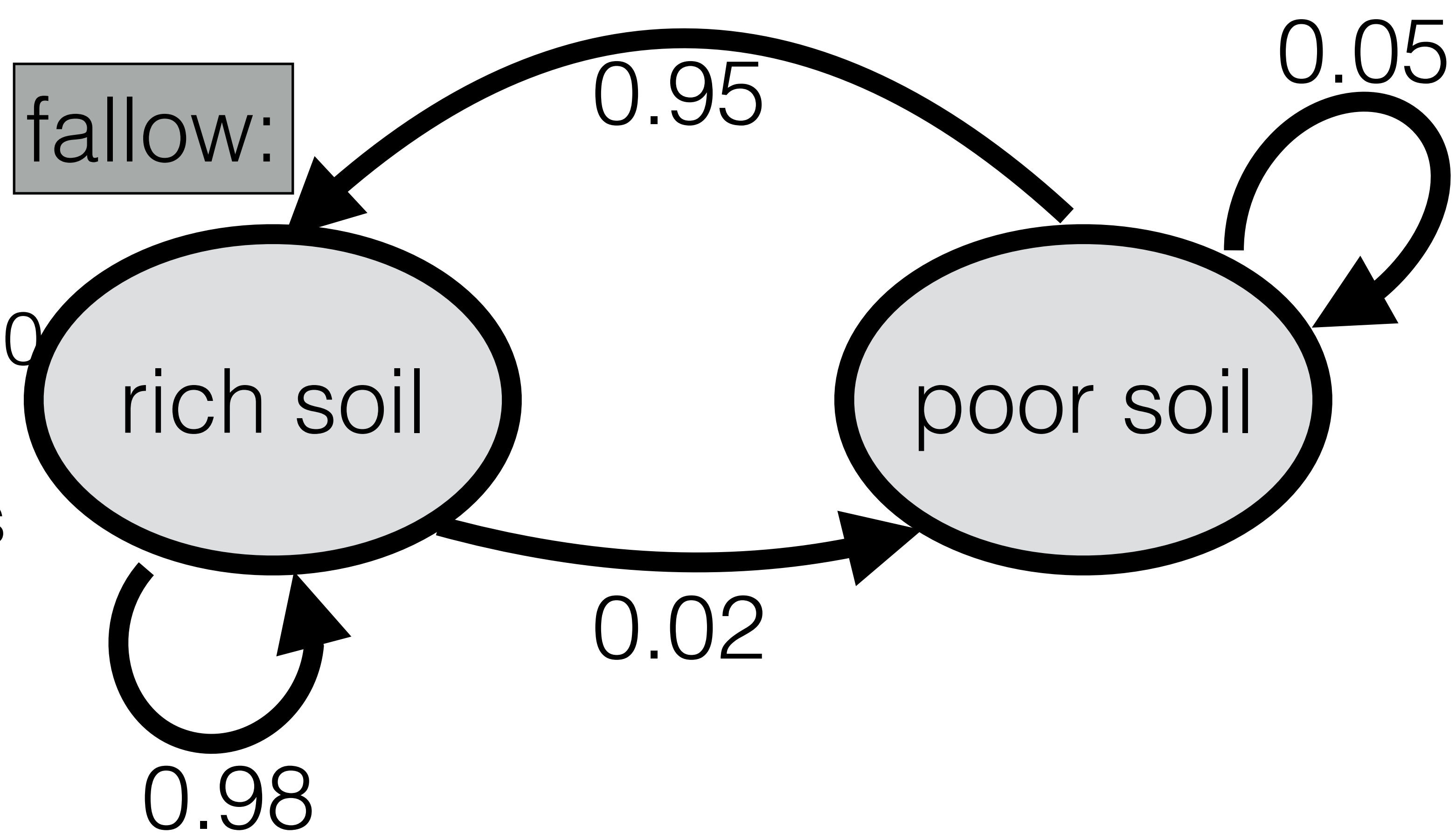
- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ : transition model
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ : reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant)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

plant:



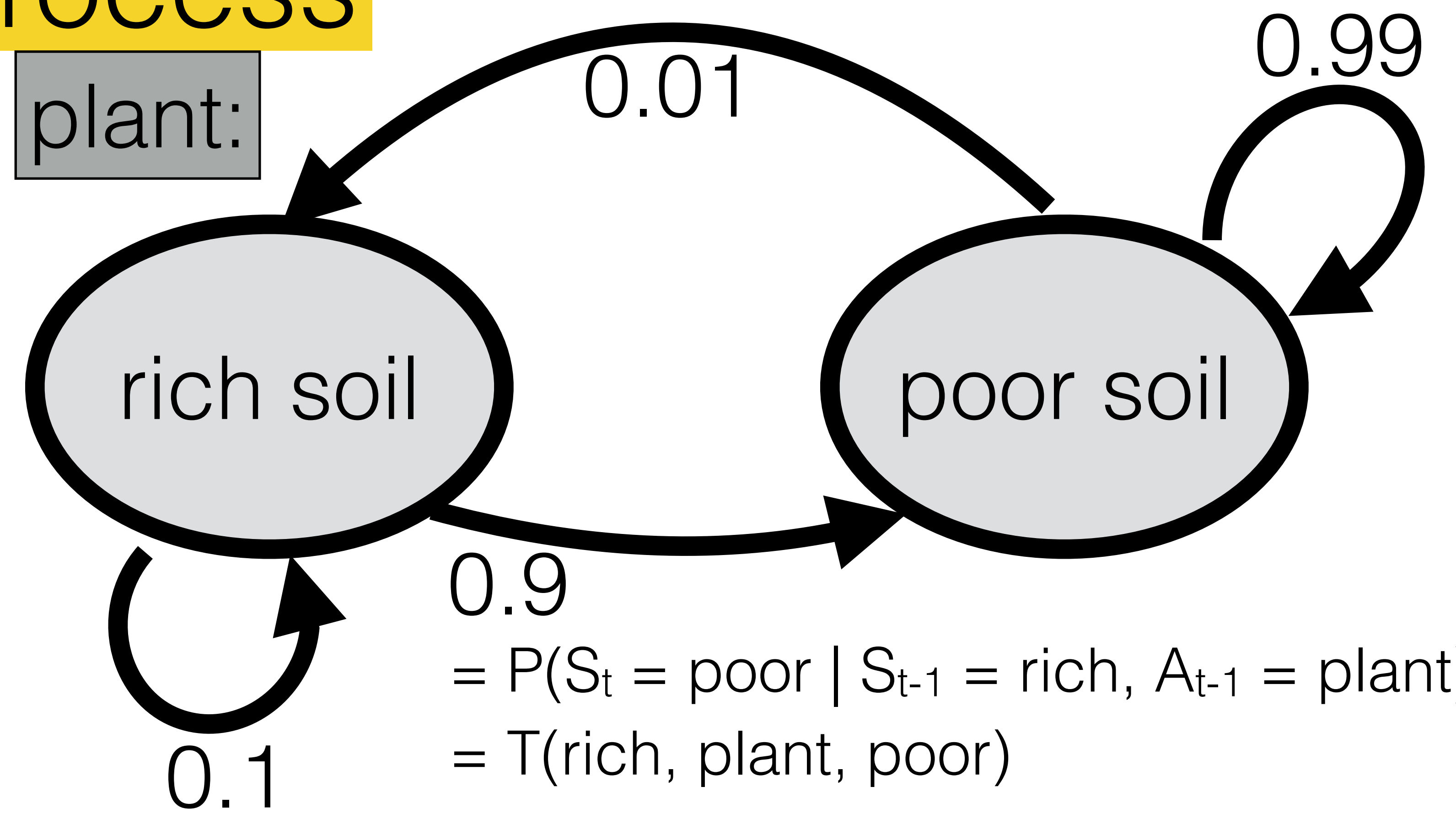

= $P(S_t = \text{poor} \mid S_{t-1} = \text{rich}, A_{t-1} = \text{plant})$
= T(rich, plant, poor)

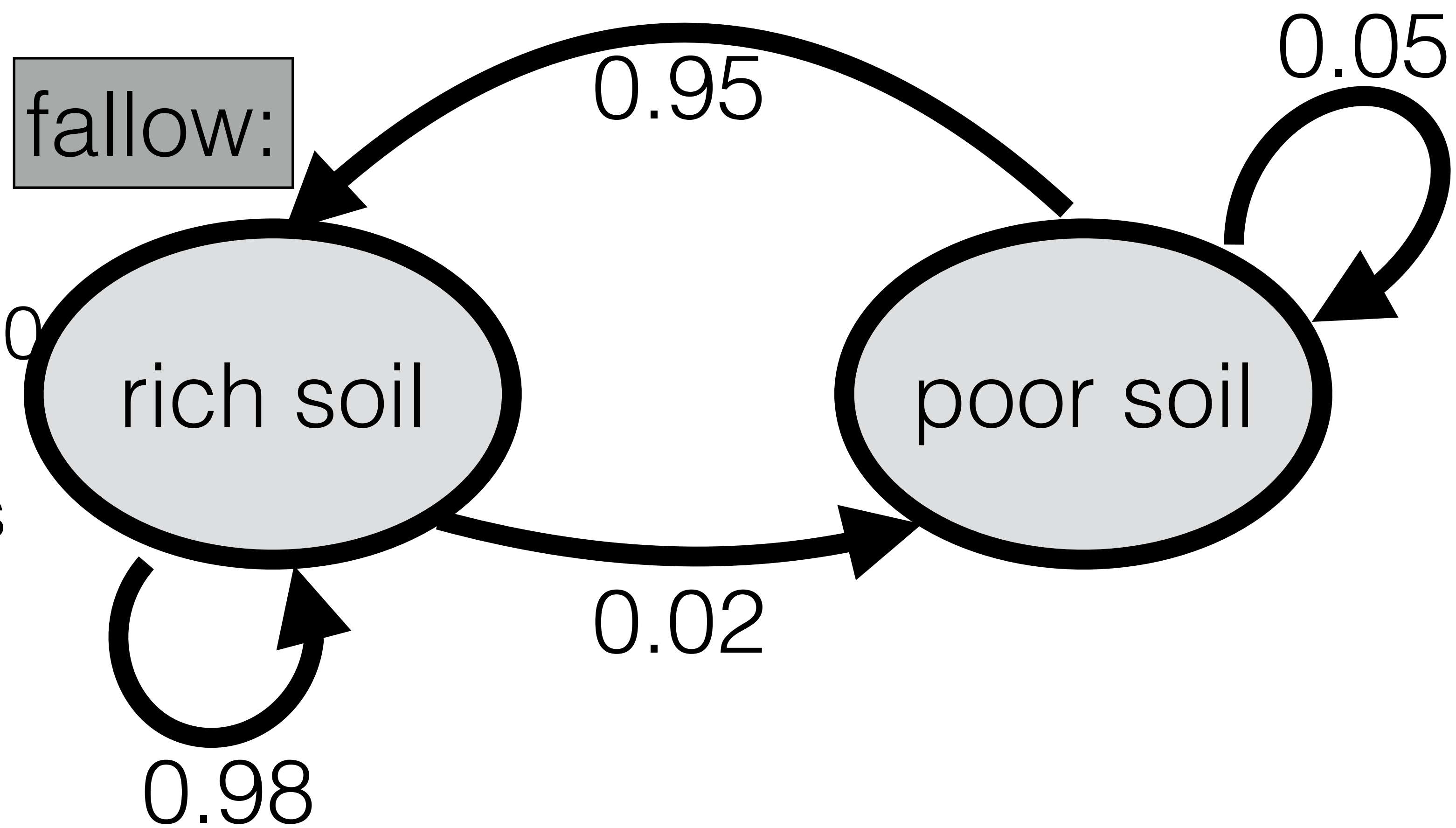
fallow:



# Markov Decision Process

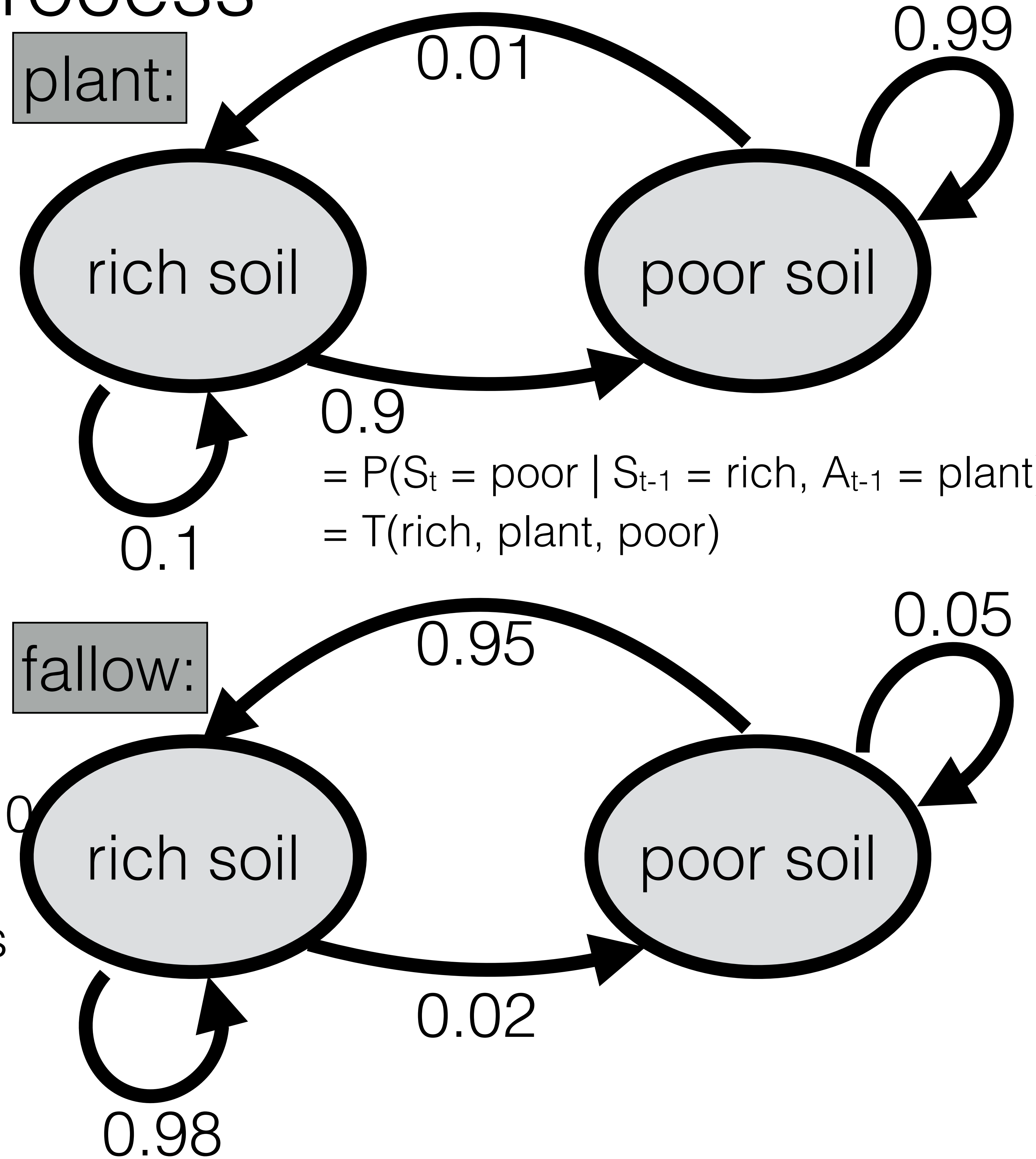
- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ : transition model
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ : reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant)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



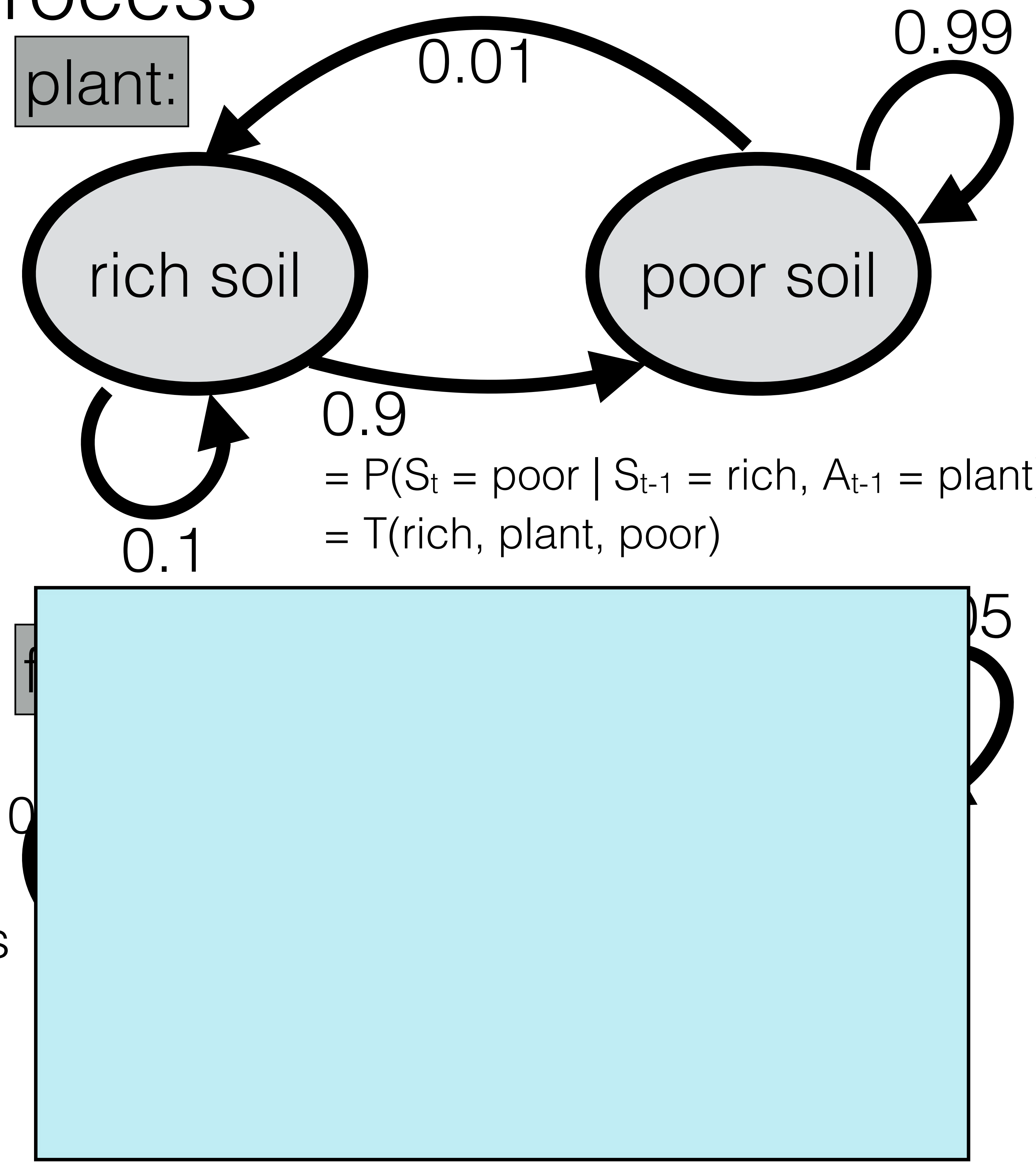
# Markov Decision Process

- $\mathcal{S}$ = 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



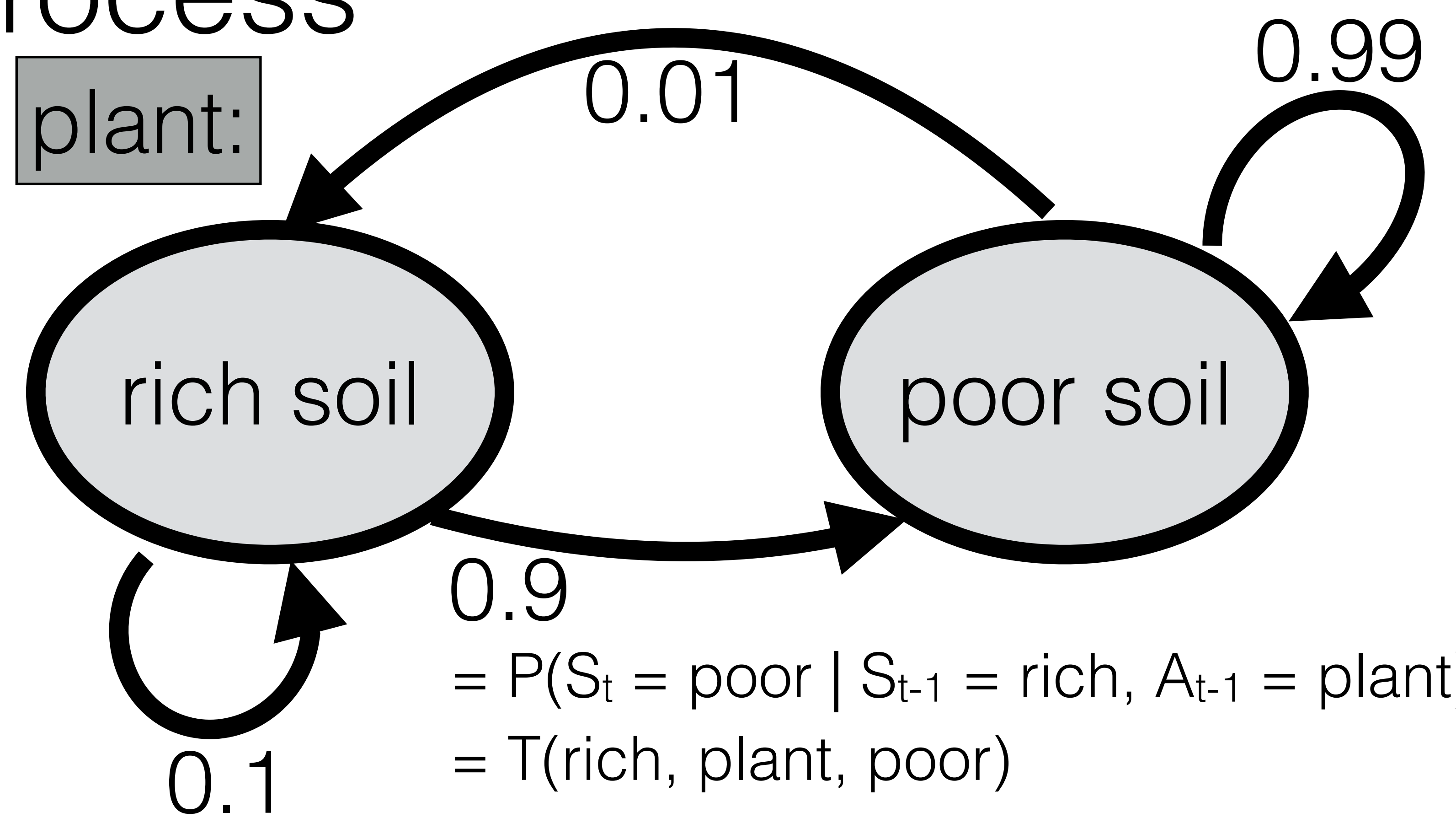
# Markov Decision Process

- $\mathcal{S}$ = 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

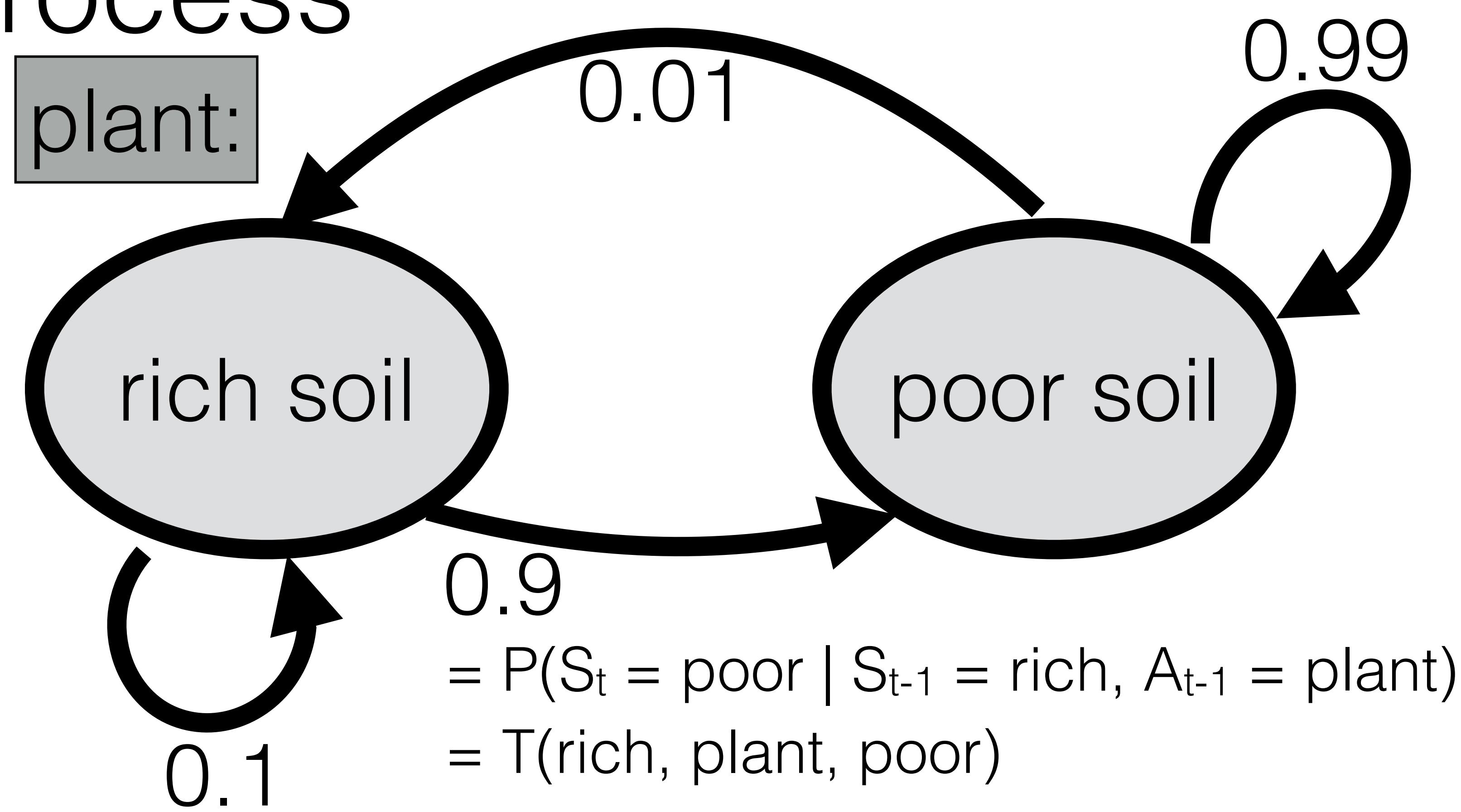
- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) =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

# Markov Decision Process

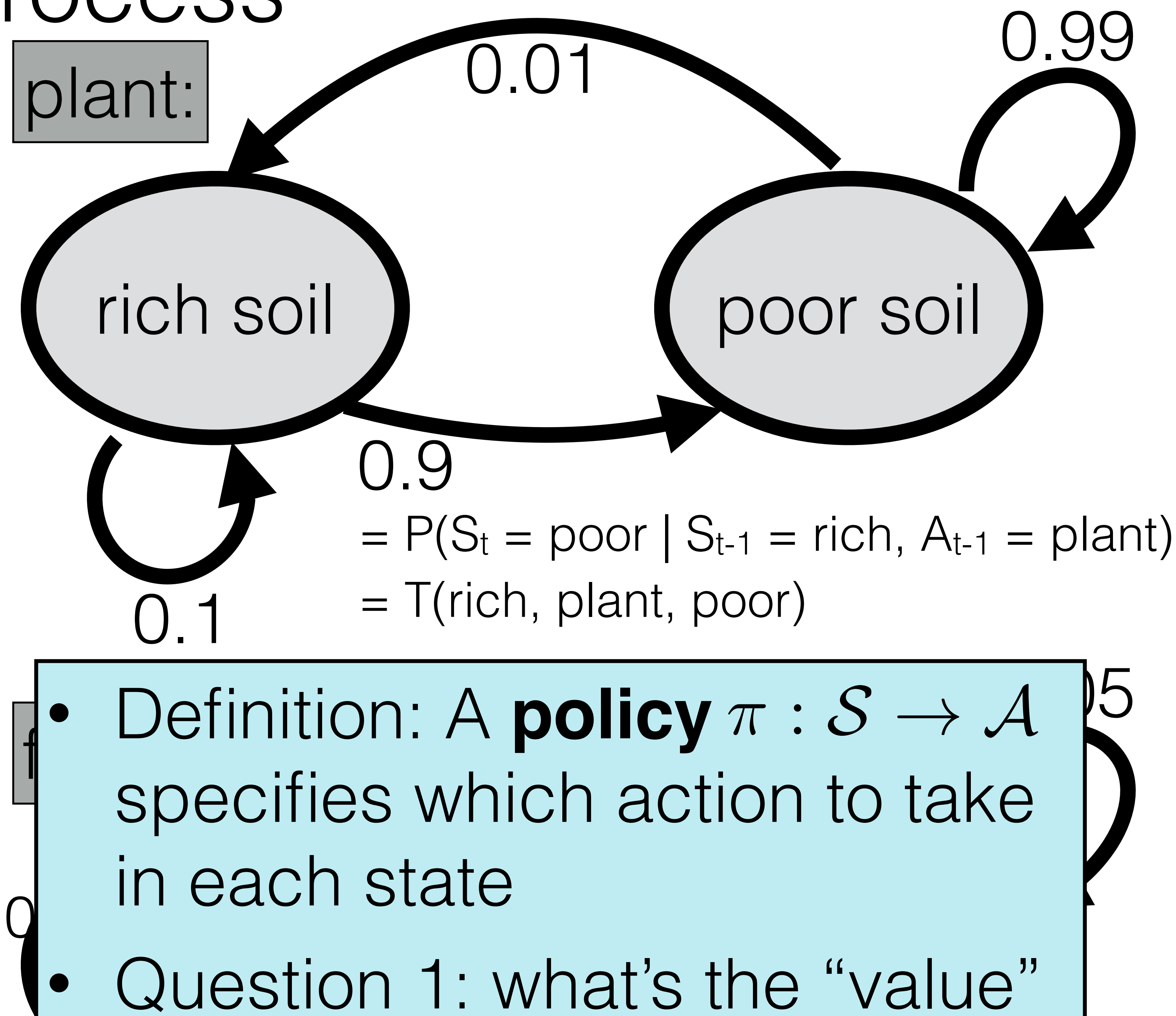
- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ : transition model
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ : reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant)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

# Markov Decision Process

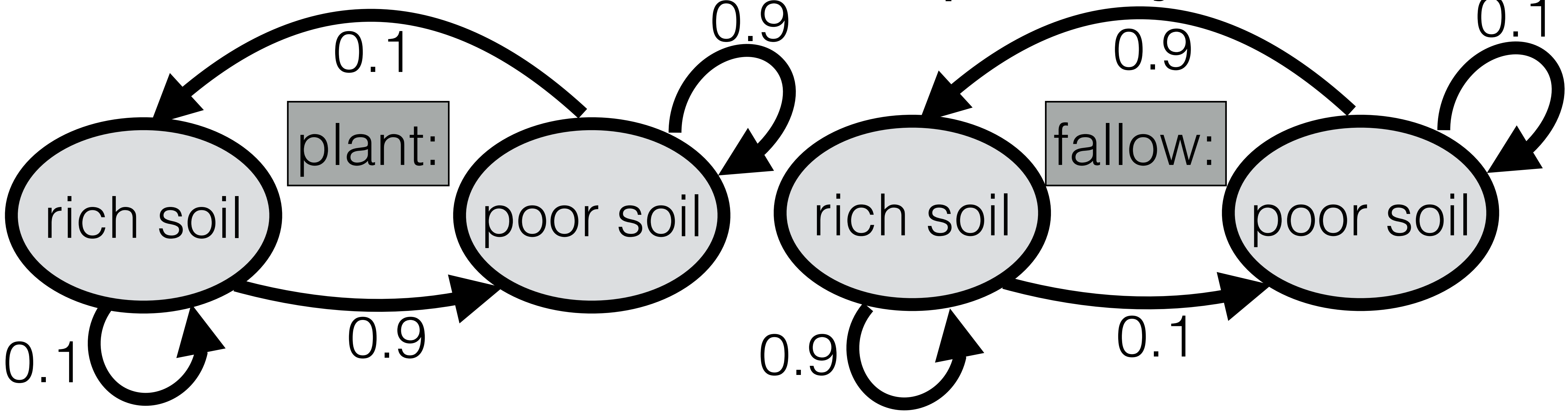
- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ : transition model
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ : reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant)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



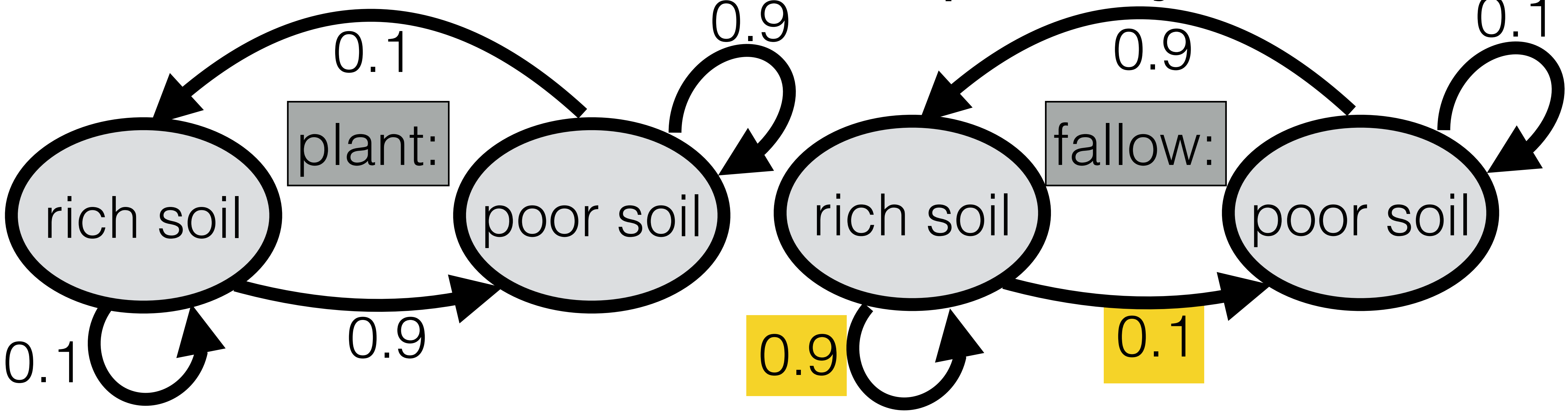

- Definition: A **policy** $\pi : \mathcal{S} \to \mathcal{A}$ specifies which action to take in each state
- Question 1: what's the "value" of a policy?
- Question 2: what's the best policy?

# What’s the value of a policy?

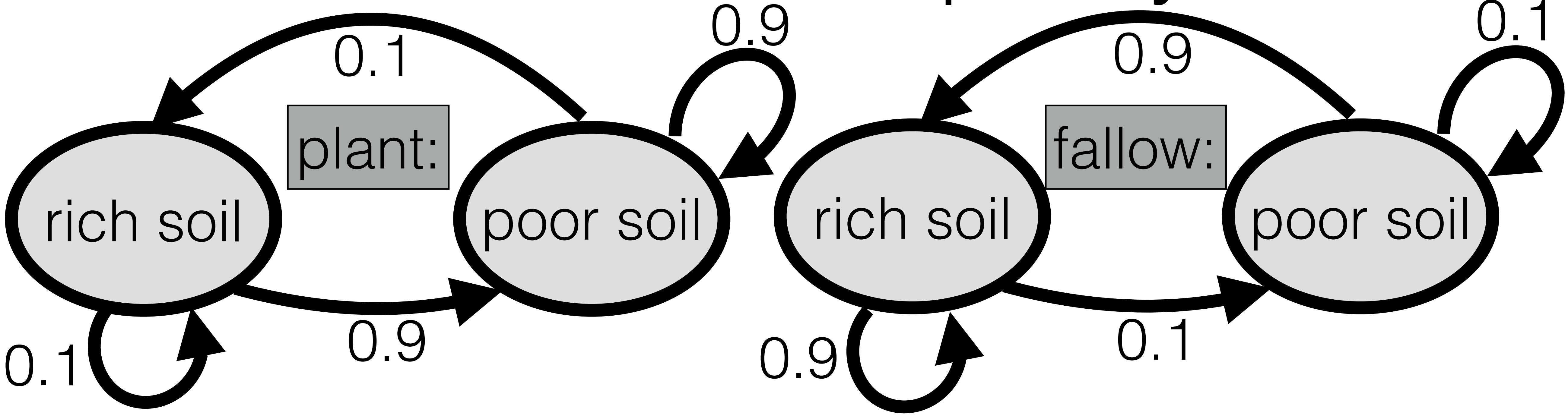
# What's the value of a policy?



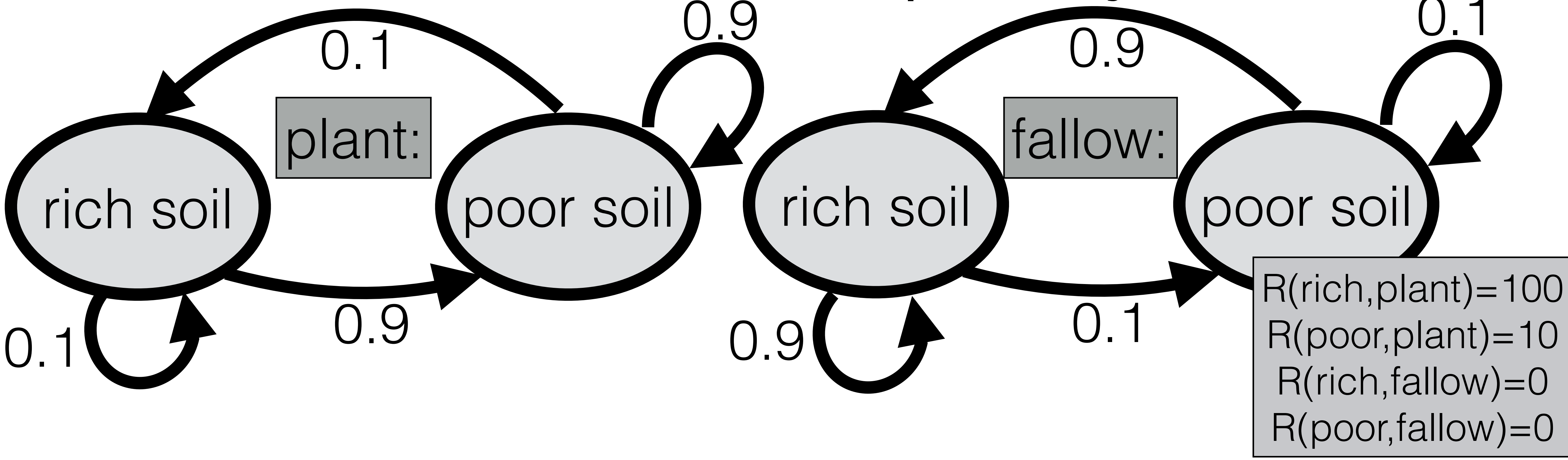
# What's the value of a policy?



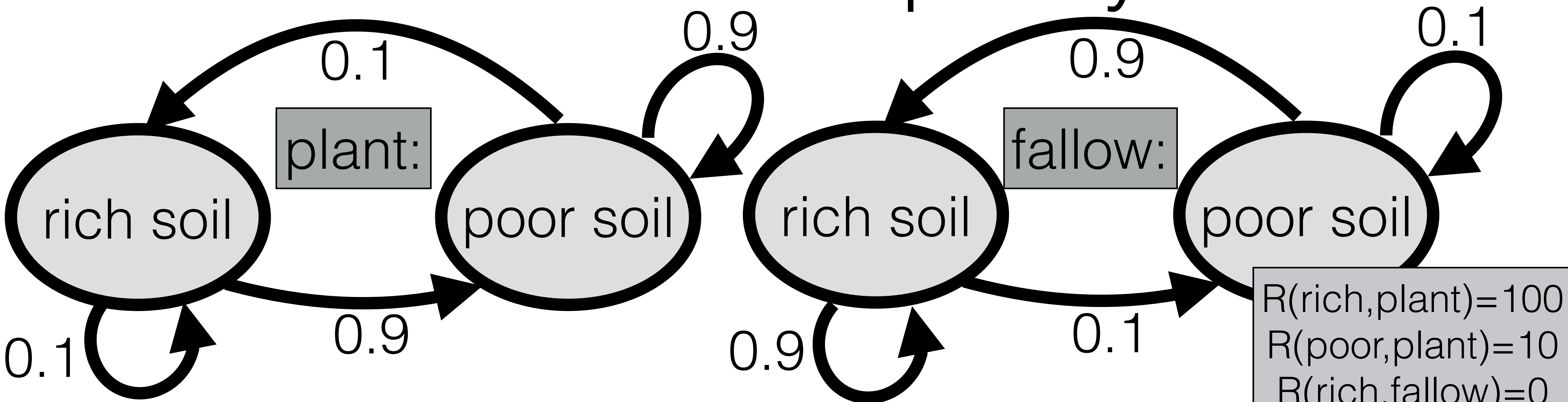
# What's the value of a policy?



# What’s the value of a policy?

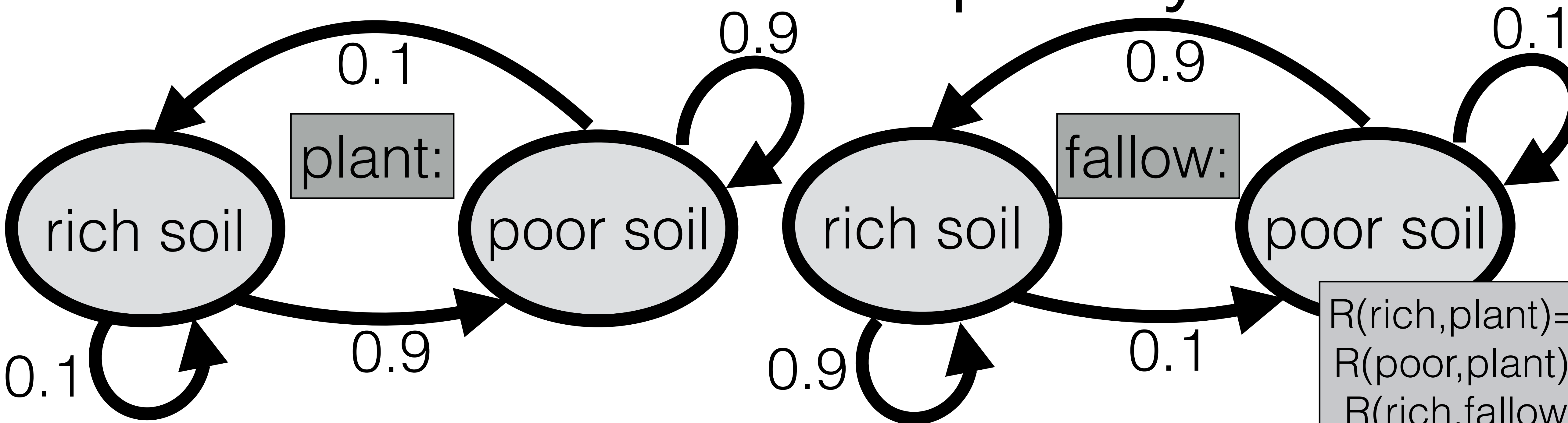


# What's the value of a policy?




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

# What’s the value of a policy?

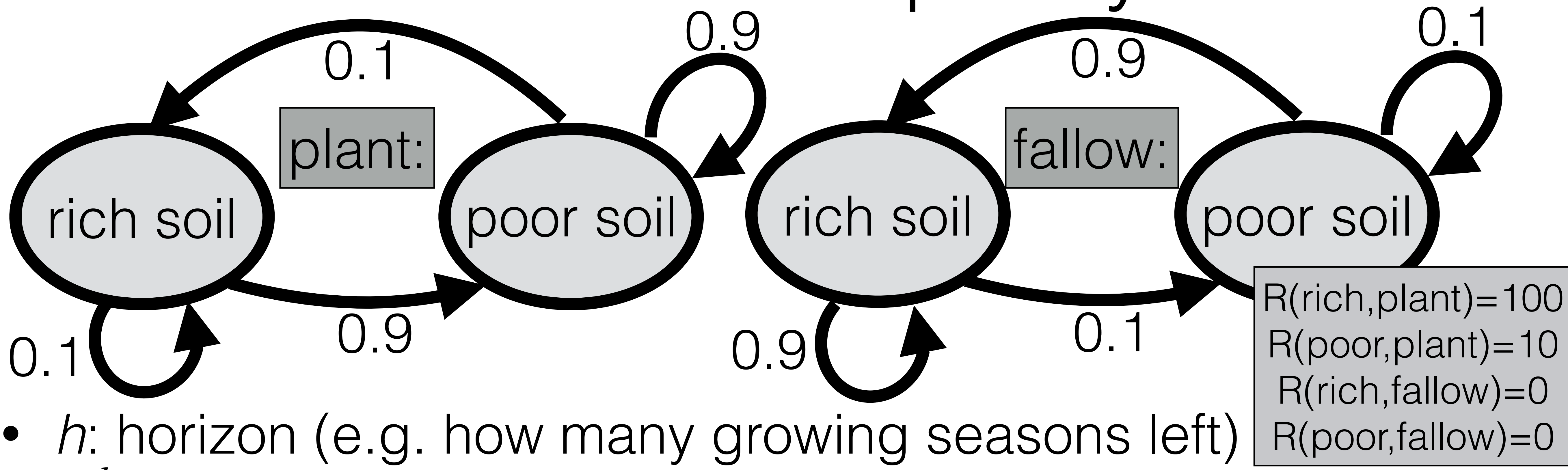



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

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

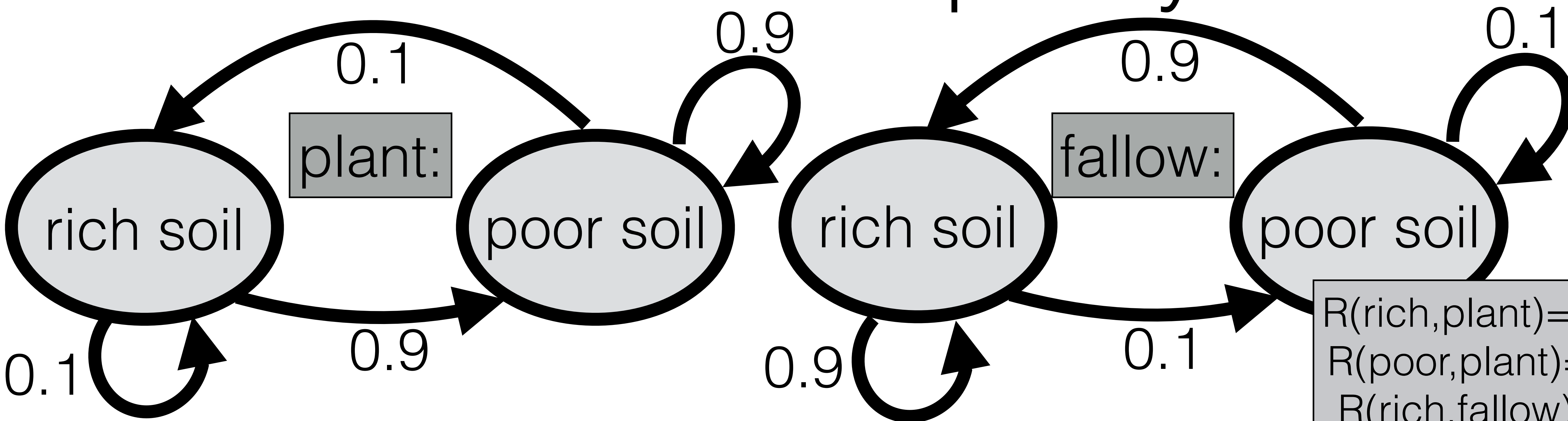
I’m renting a field for *h* growing seasons. Then it will be destroyed to make a strip mall.

# What’s the value of a policy?




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

# What’s the value of a policy?

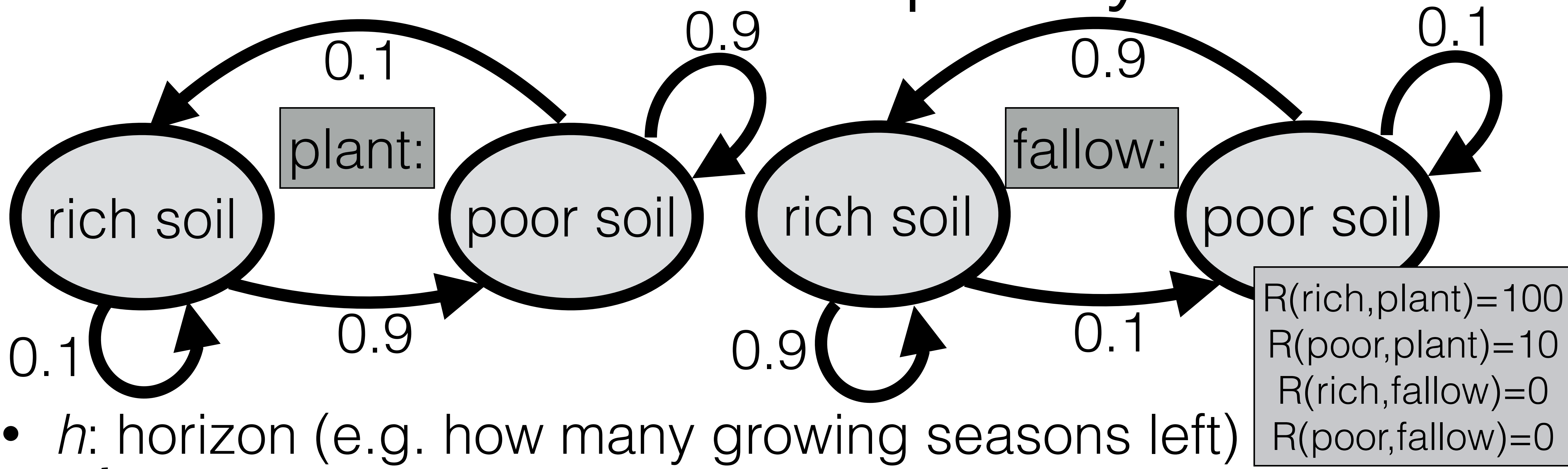



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

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

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

# What's the value of a policy?

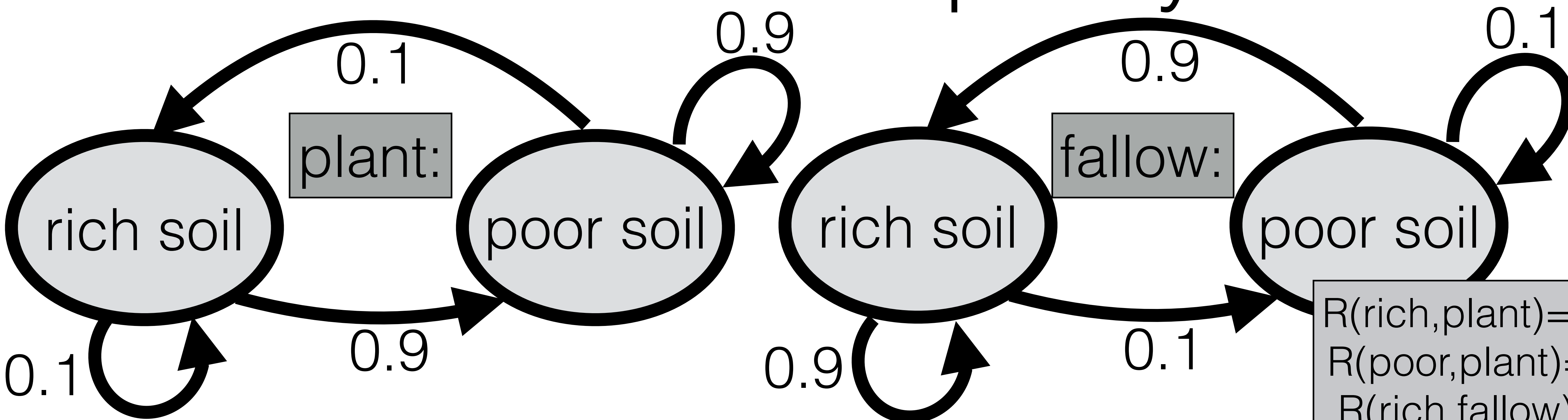



- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

# What's the value of a policy?



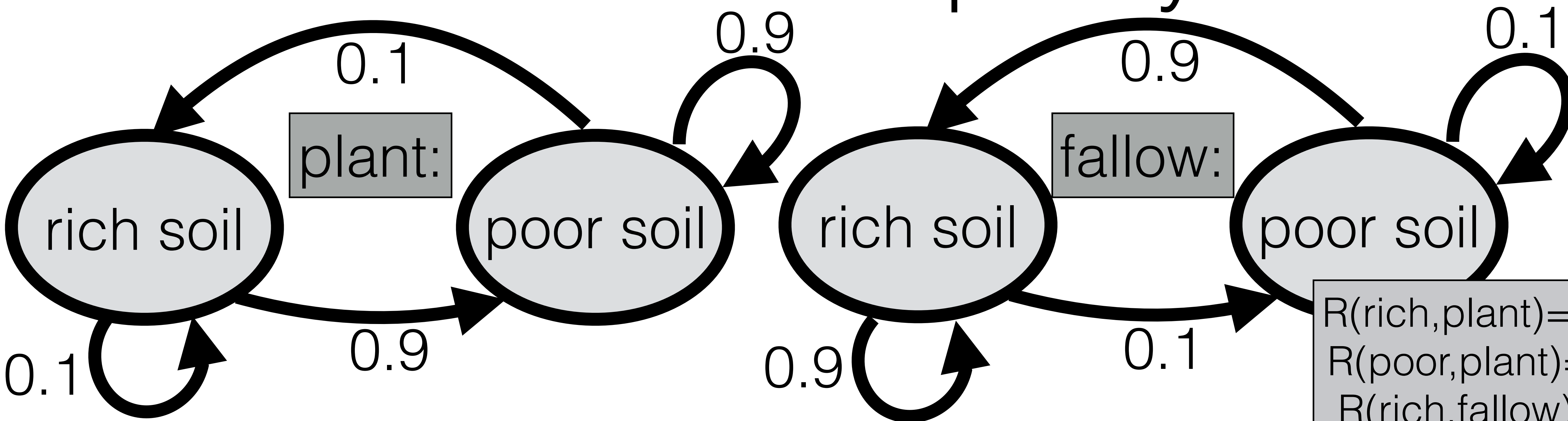

R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

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

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

$$V_\pi^0(s) = 0; V_\pi^1(s) = R(s, \pi(s))$$

# What's the value of a policy?




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

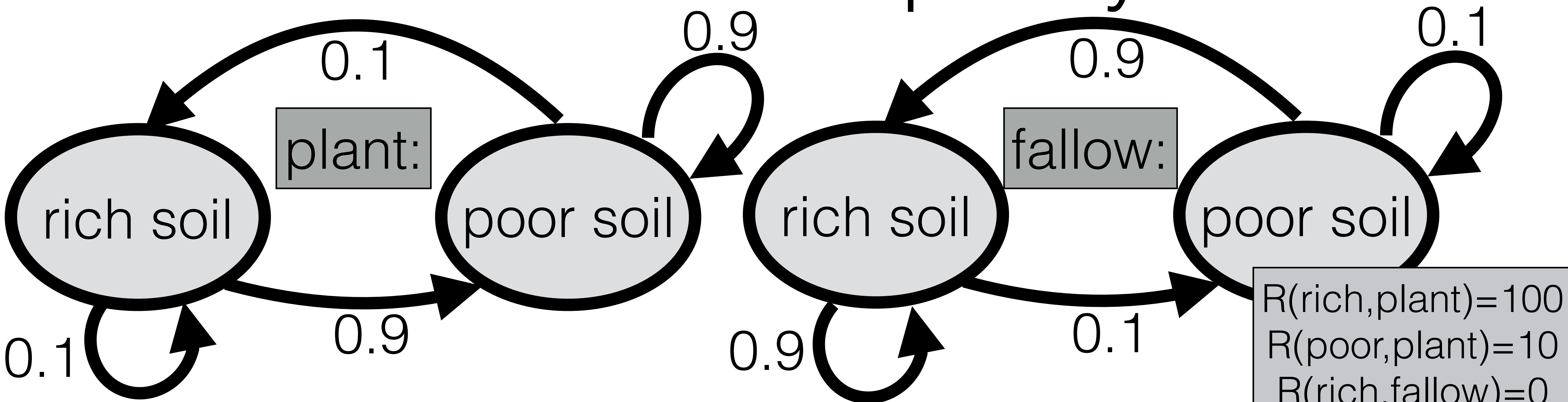
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

$$V_\pi^0(s) = 0; V_\pi^1(s) = R(s, \pi(s))$$

$$V_{\pi_A}^1(\text{rich}) =$$

# What's the value of a policy?



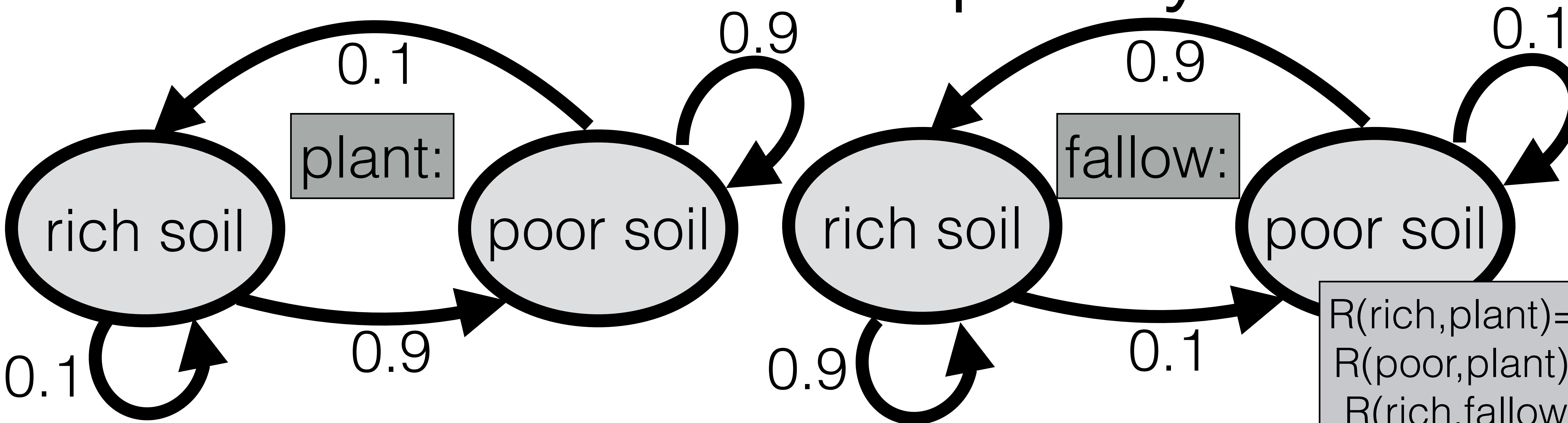

- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

$$V_\pi^0(s) = 0; V_\pi^1(s) = R(s, \pi(s))$$

$$V_{\pi_A}^1(\text{rich}) = R(\text{rich}, \pi_A(\text{rich})) =$$

# What’s the value of a policy?




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

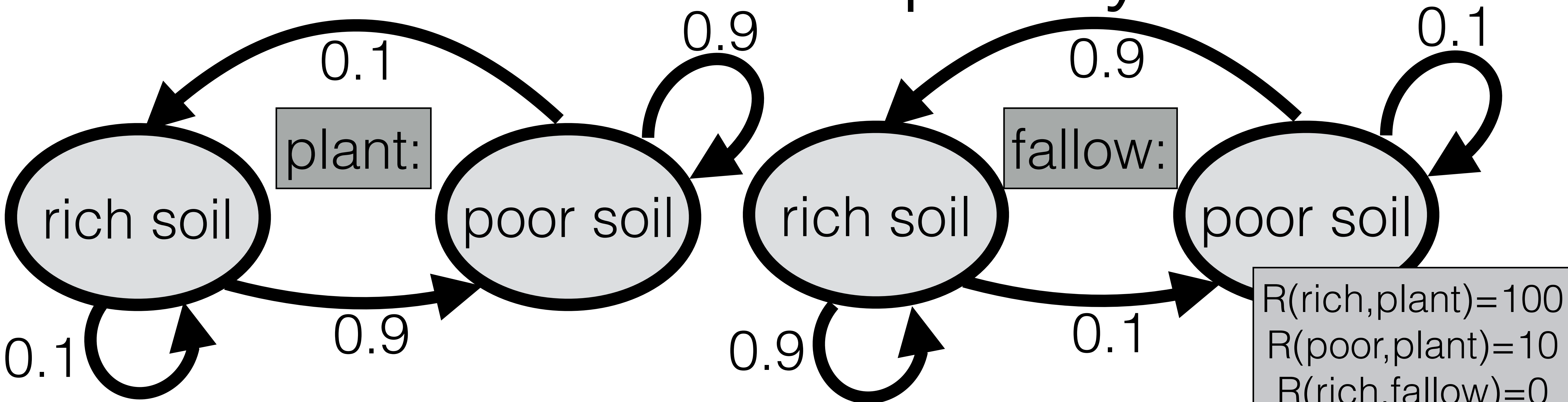
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

$$V_\pi^0(s) = 0; V_\pi^1(s) = R(s, \pi(s))$$

$$V_{\pi_A}^1(\text{rich}) = R(\text{rich}, \pi_A(\text{rich})) = 100$$

# What’s the value of a policy?



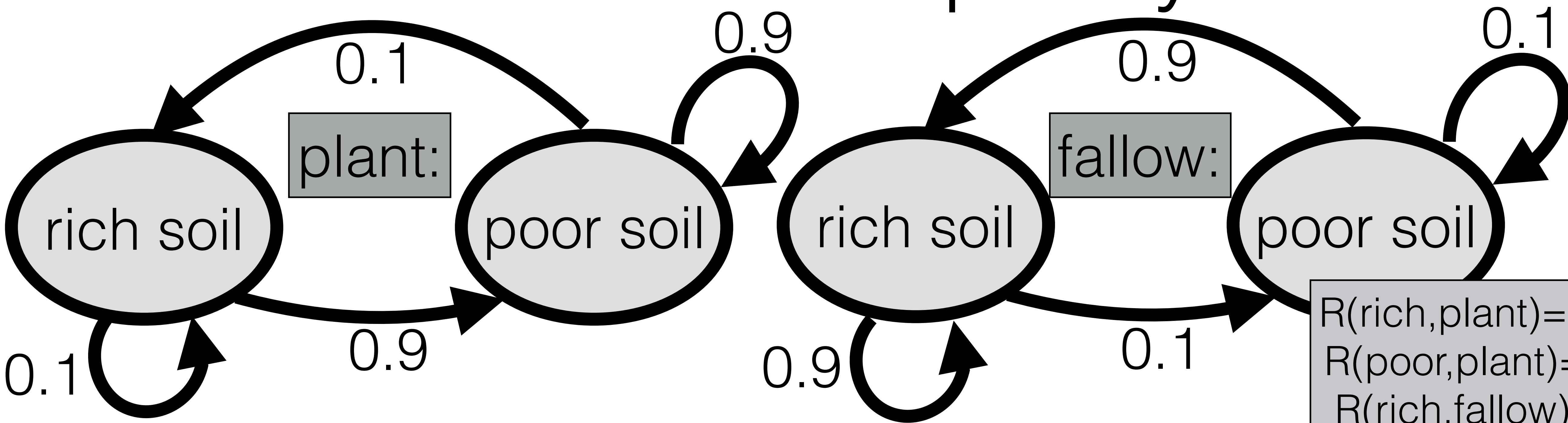

- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

$$V_\pi^0(s) = 0; V_\pi^1(s) = R(s, \pi(s))$$

$$V_{\pi_A}^1(\text{rich}) = 100$$

# What's the value of a policy?




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

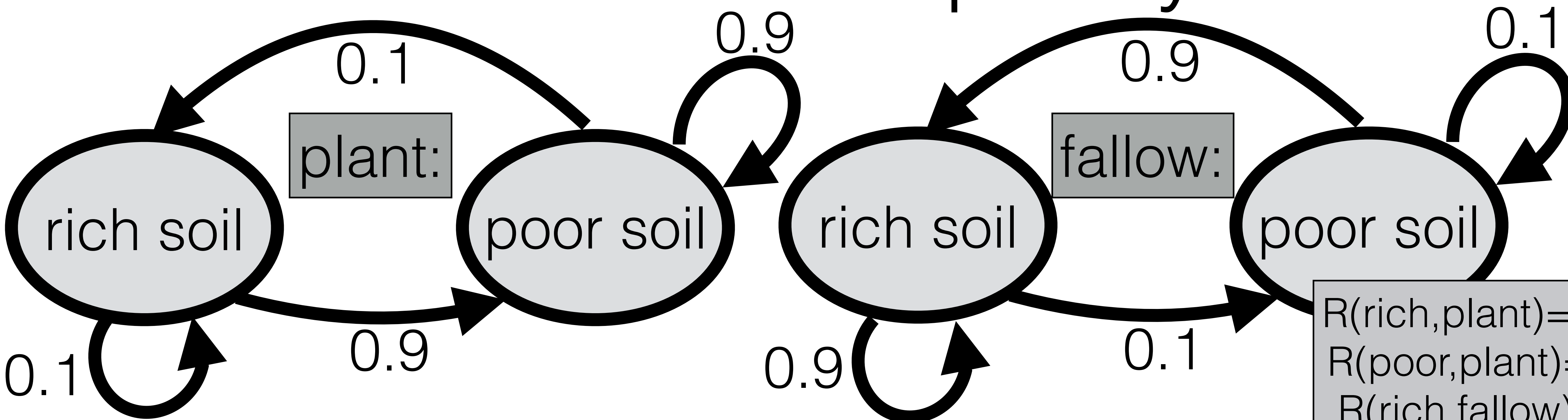
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

$$V_\pi^0(s) = 0; V_\pi^1(s) = R(s, \pi(s))$$

$$V_{\pi_A}^1(\mathrm{rich}) = 100; V_{\pi_A}^1(\mathrm{poor}) =$$

# What's the value of a policy?




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

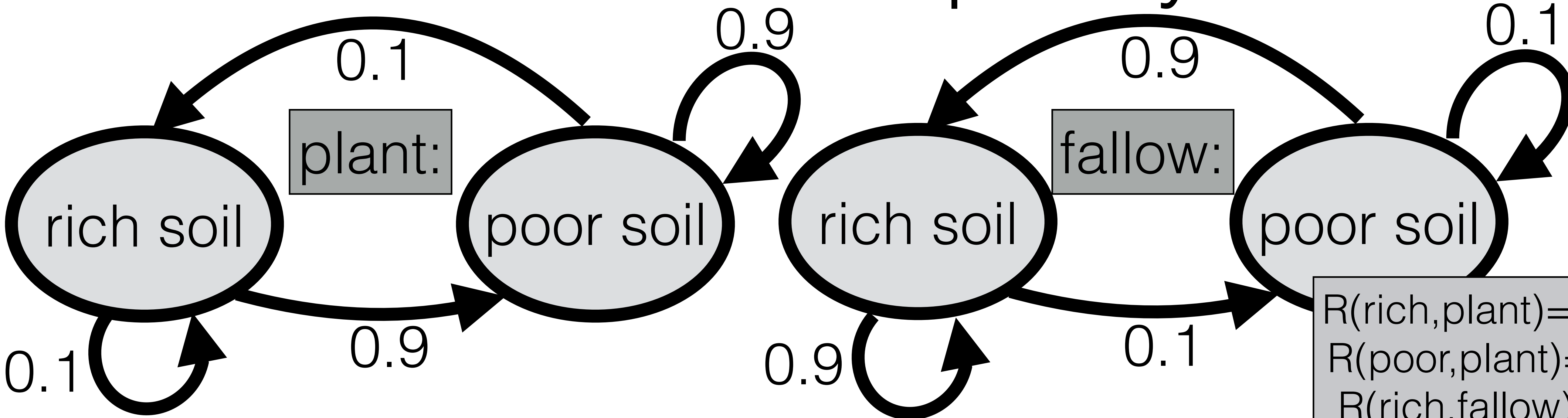
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

$$V_\pi^0(s) = 0; V_\pi^1(s) = R(s, \pi(s))$$

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

# What's the value of a policy?




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

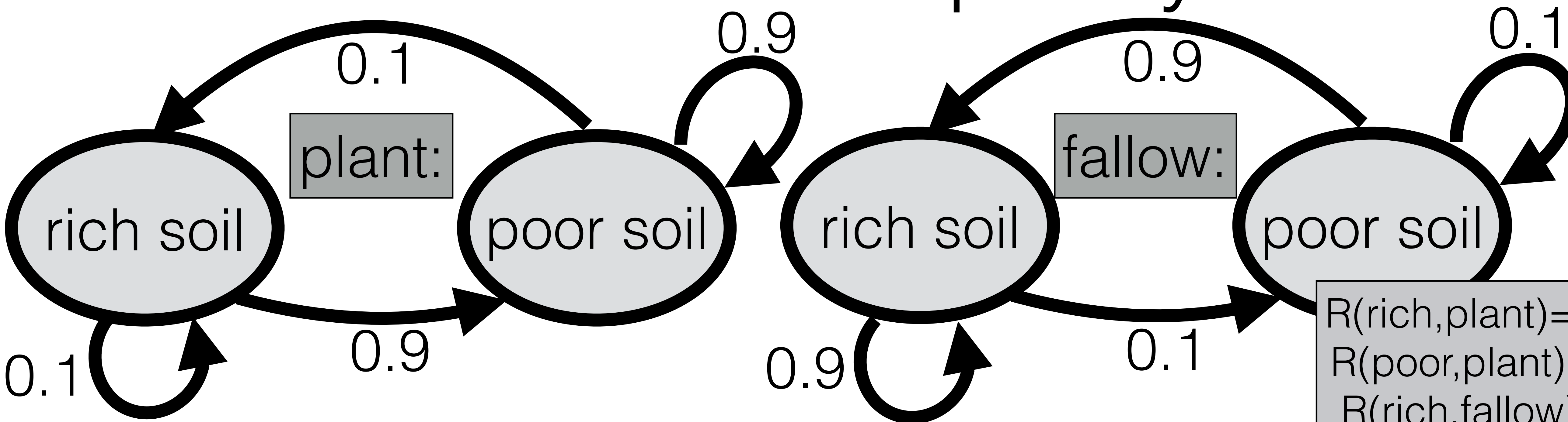
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

$$V_\pi^0(s) = 0; V_\pi^1(s) = R(s, \pi(s))$$

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

# What's the value of a policy?




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

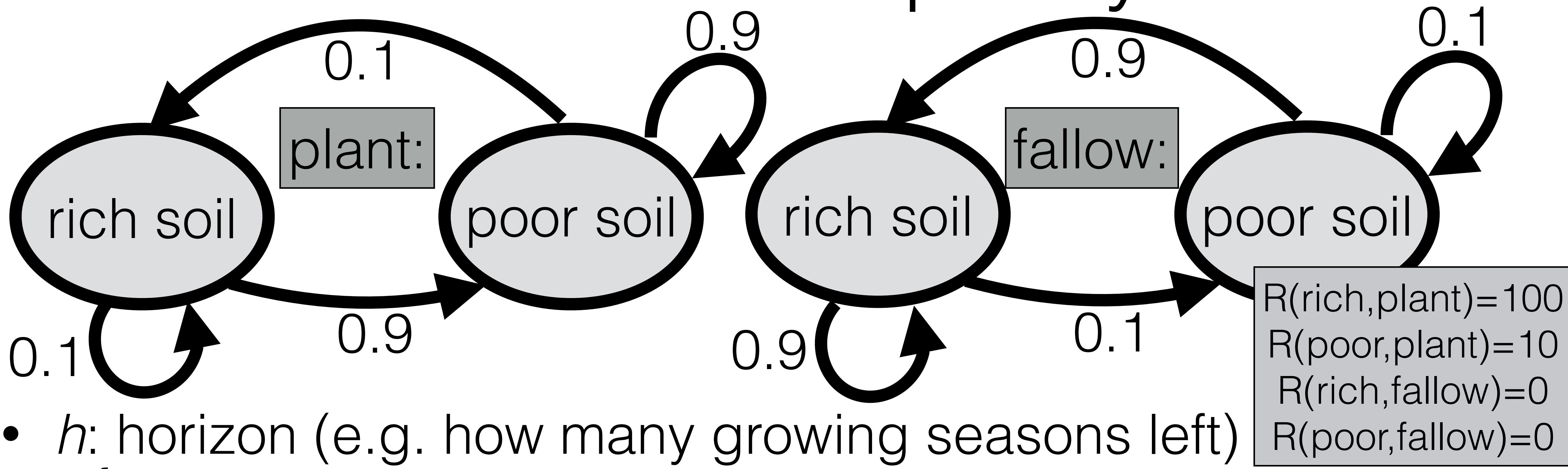
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

$V_\pi^0(s) = 0; V_\pi^1(s) = R(s, \pi(s))$

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

# What's the value of a policy?



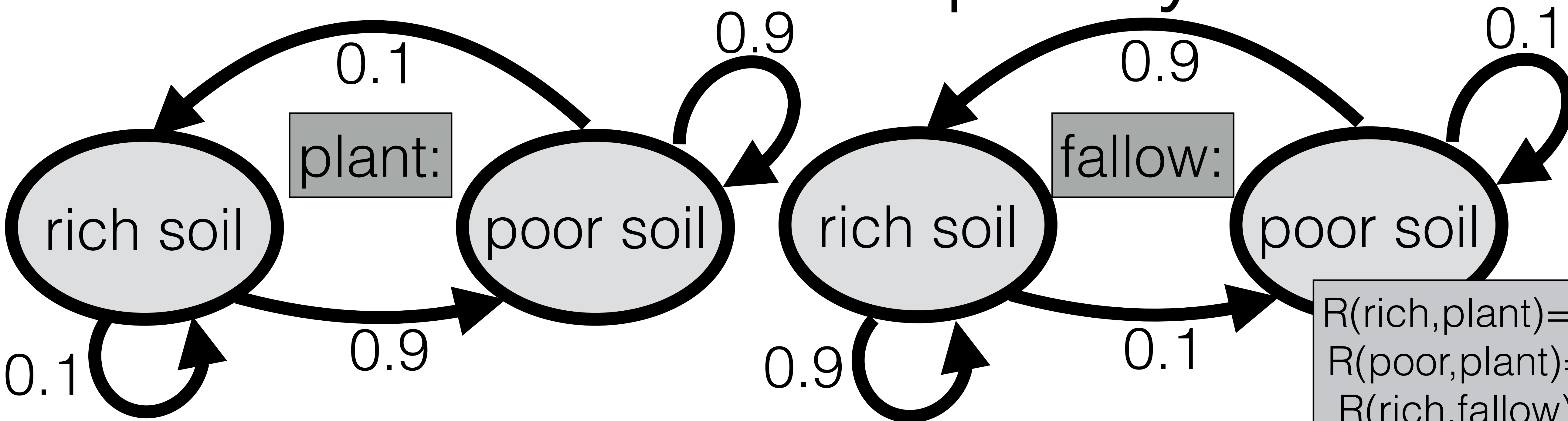

- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

# What’s the value of a policy?




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

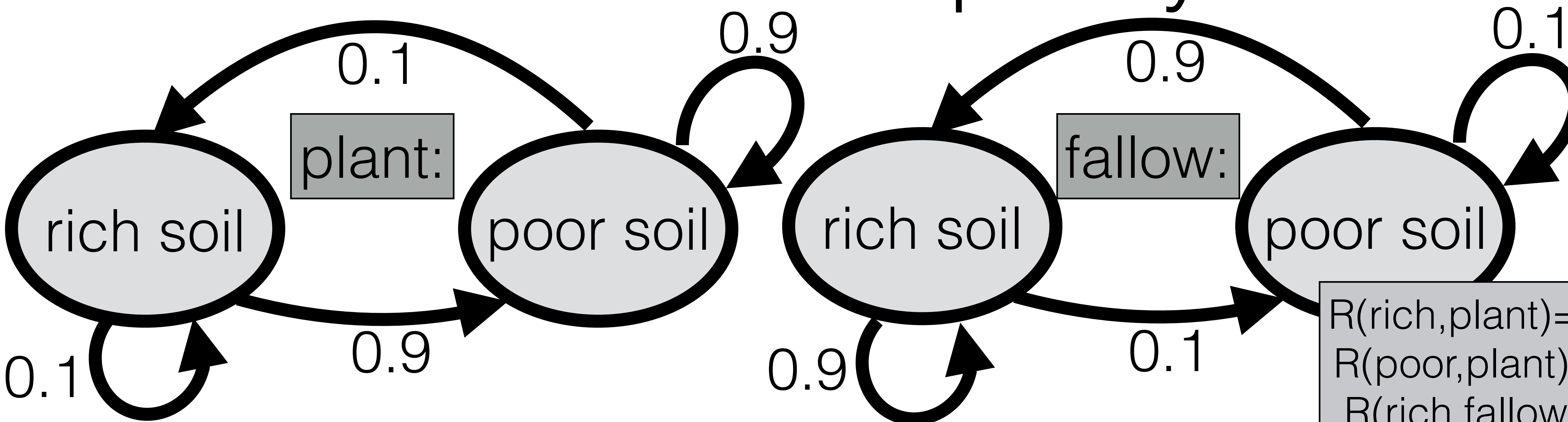
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

# What's the value of a policy?




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

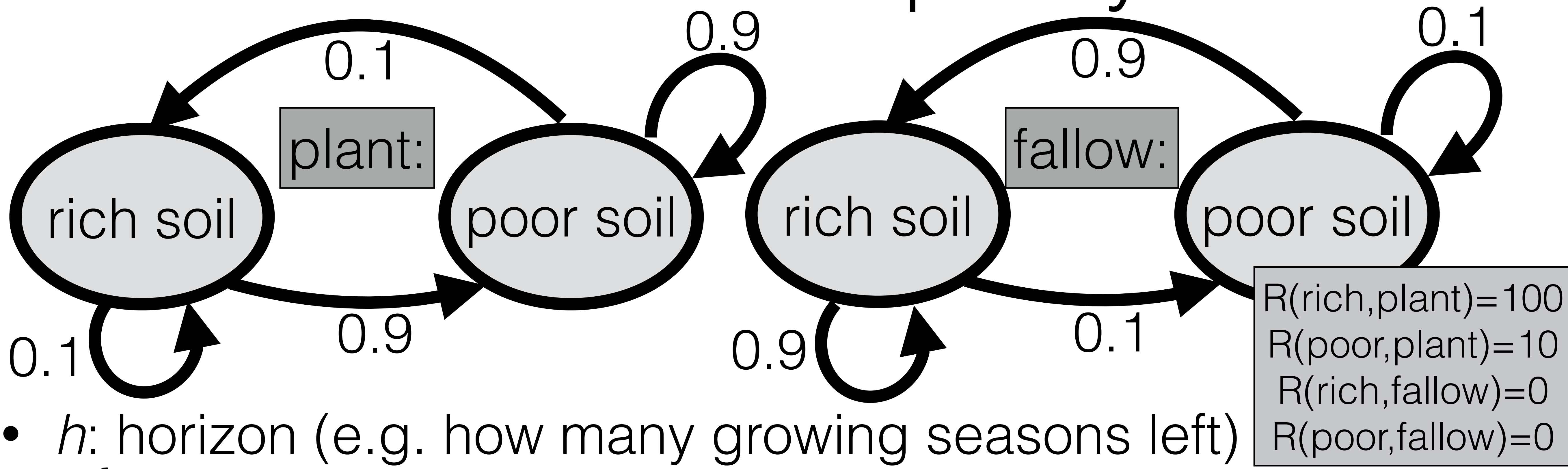
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

# What's the value of a policy?



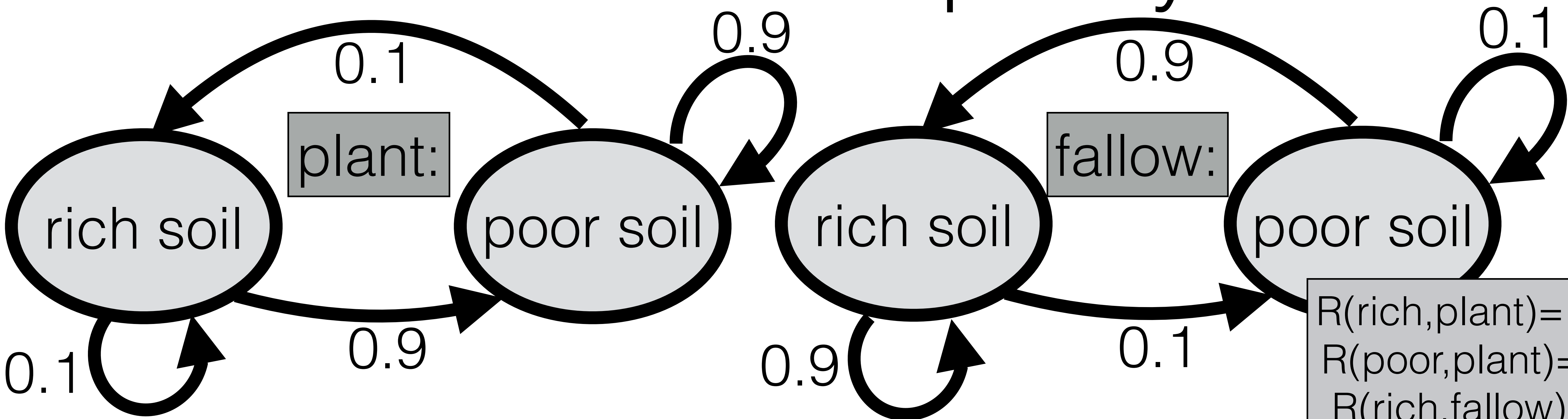

- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

# What's the value of a policy?




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

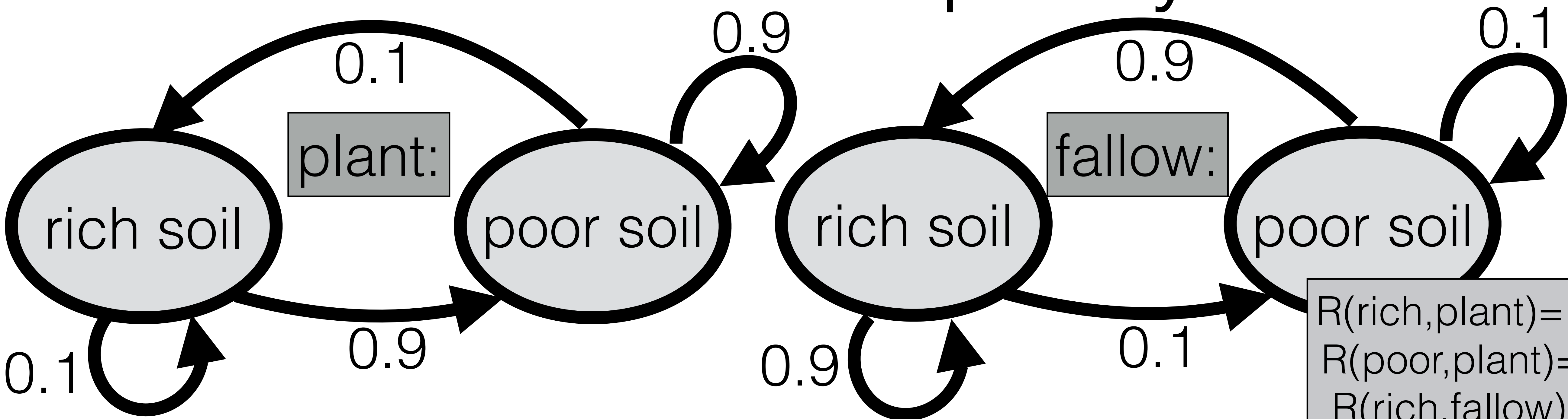
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

# What's the value of a policy?




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

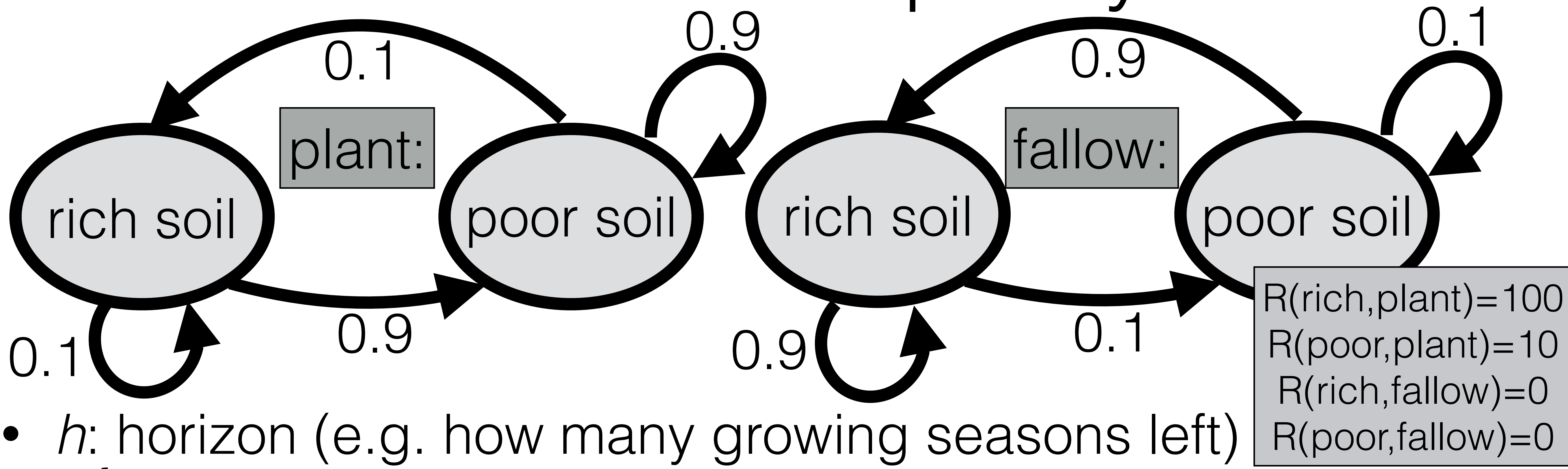
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

# What's the value of a policy?



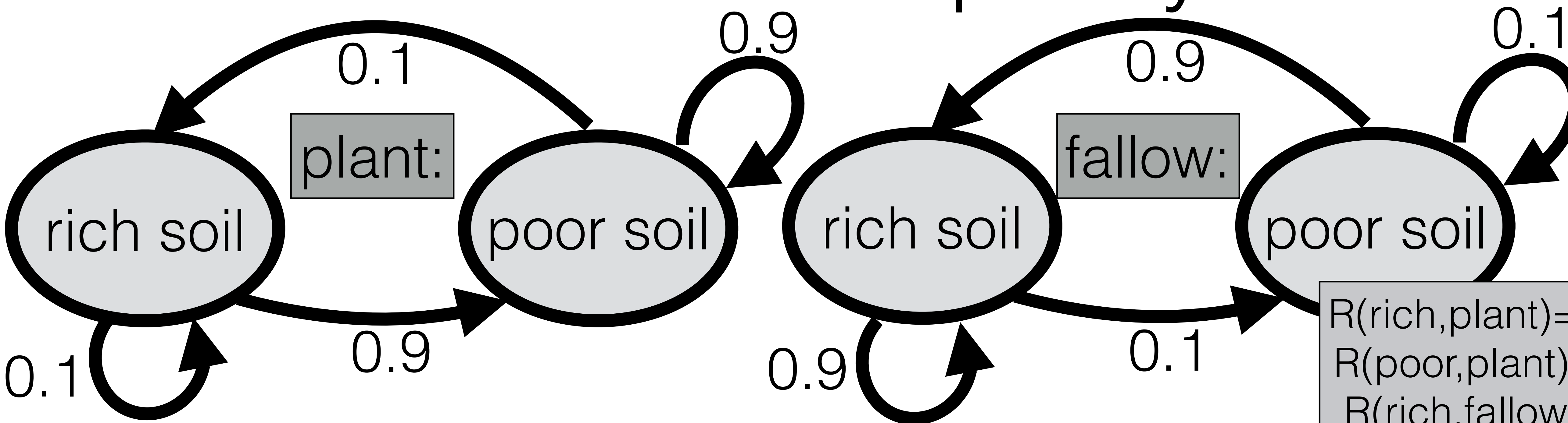

- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

# 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 $\pi$ starting at *s*

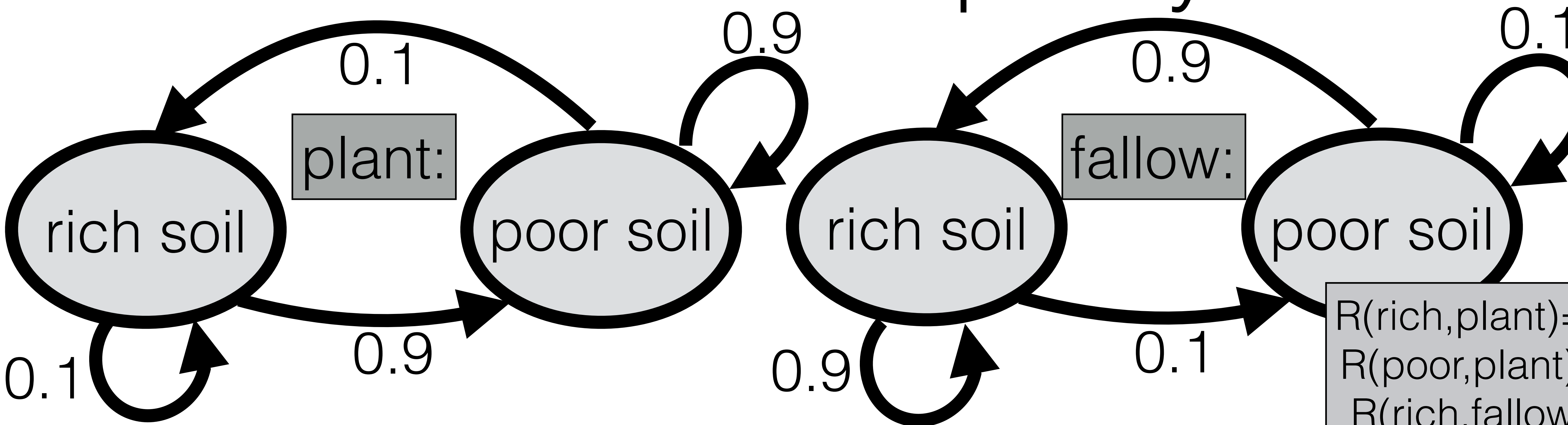
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

$$V_{\pi_A}^2(\text{rich}) =$$

# 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 $\pi$ starting at $s$

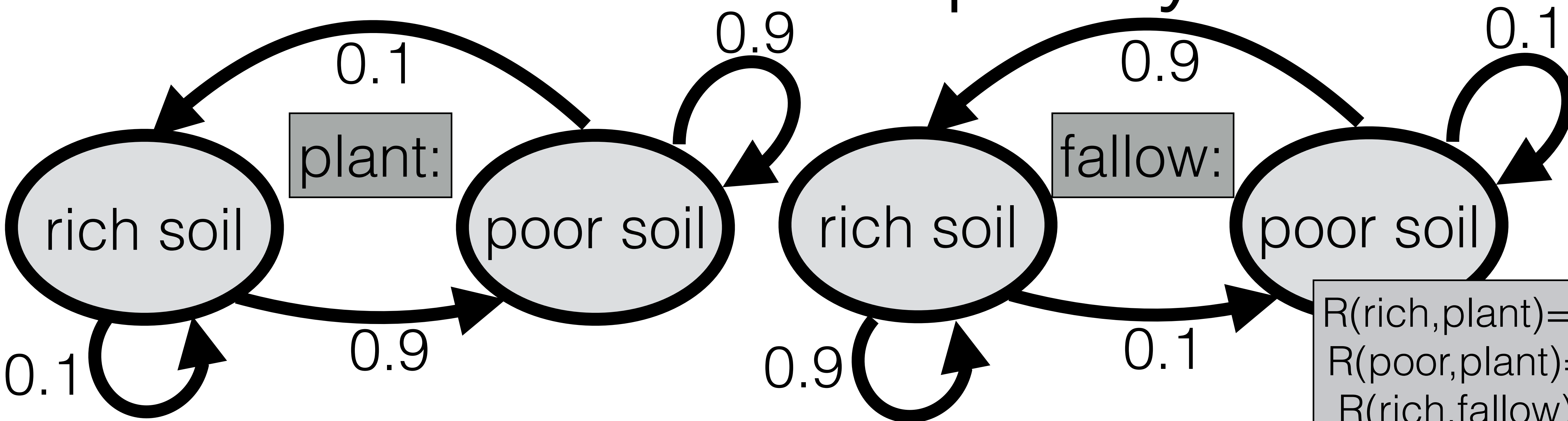
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

$$V_{\pi_A}^2(\text{rich}) = R(\text{rich}, \pi_A(\text{rich})) +$$

# 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 $\pi$ starting at *s*

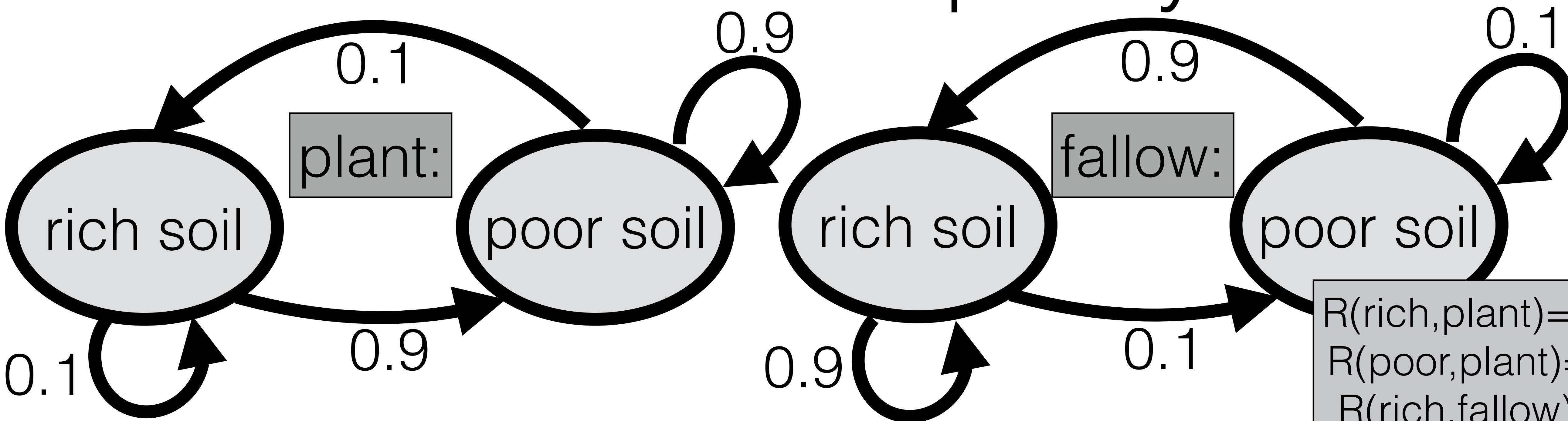
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

$$V_{\pi_A}^2(\text{rich}) = R(\text{rich}, \pi_A(\text{rich})) + T(\text{rich}, \pi_A(\text{rich}), \text{rich}) V_{\pi_A}^1(\text{rich}) \\ + T(\text{rich}, \pi_A(\text{rich}), \text{poor}) V_{\pi_A}^1(\text{poor})$$

# 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 $\pi$ starting at $s$

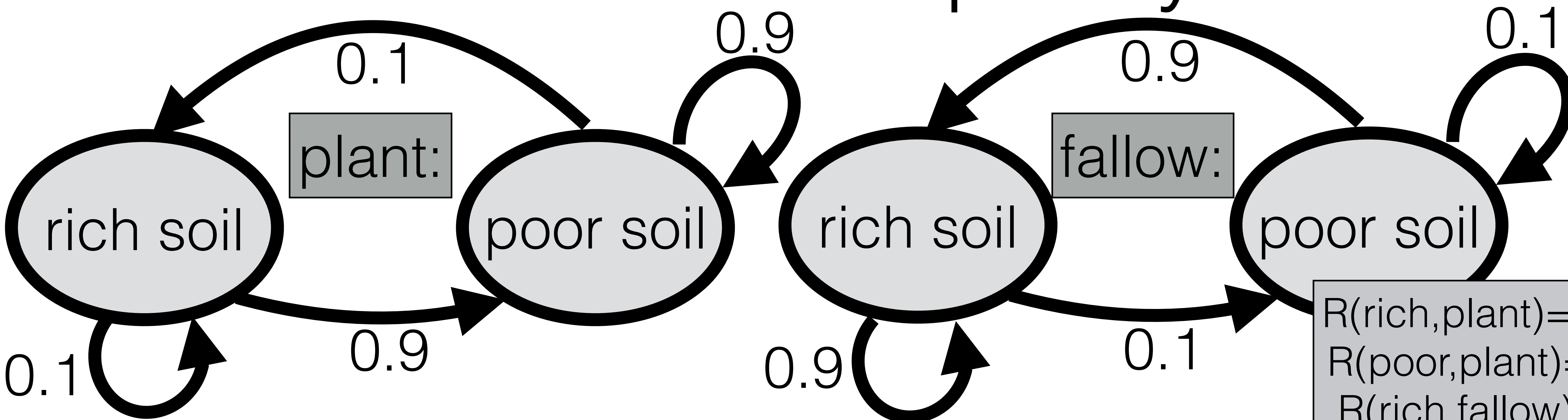
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

$$V_{\pi_A}^2(\text{rich}) = R(\text{rich}, \pi_A(\text{rich})) + T(\text{rich}, \pi_A(\text{rich}), \text{rich}) V_{\pi_A}^1(\text{rich})$$
$$+ T(\text{rich}, \pi_A(\text{rich}), \text{poor}) V_{\pi_A}^1(\text{poor})$$

# 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 $\pi$ starting at $s$

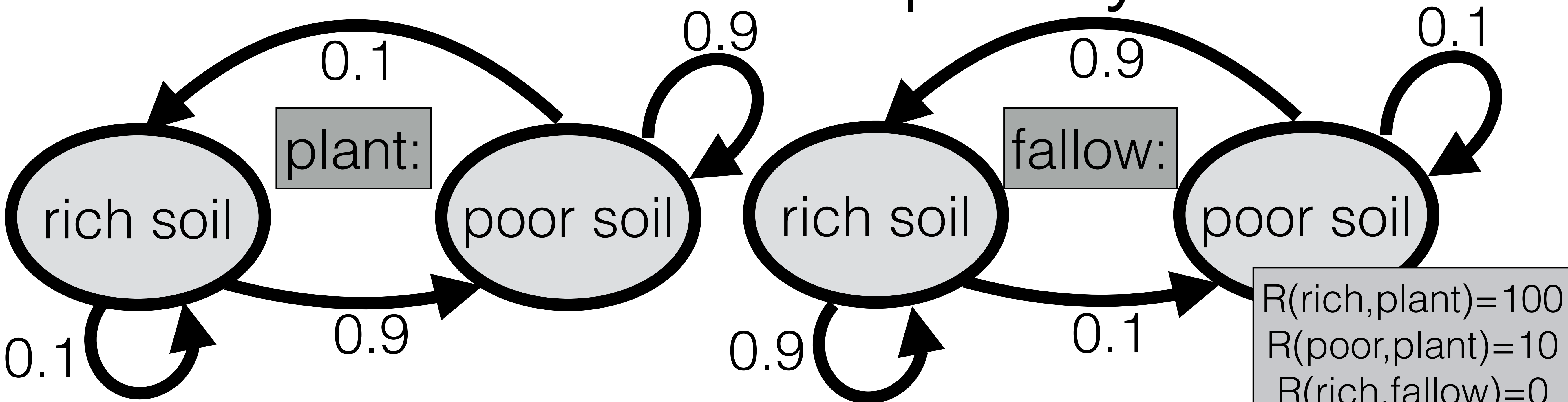
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

$$V_{\pi_A}^2(\text{rich}) = R(\text{rich}, \pi_A(\text{rich})) + T(\text{rich}, \pi_A(\text{rich}), \text{rich}) V_{\pi_A}^1(\text{rich})$$
$$+ T(\text{rich}, \pi_A(\text{rich}), \text{poor}) V_{\pi_A}^1(\text{poor})$$

# What's the value of a policy?




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

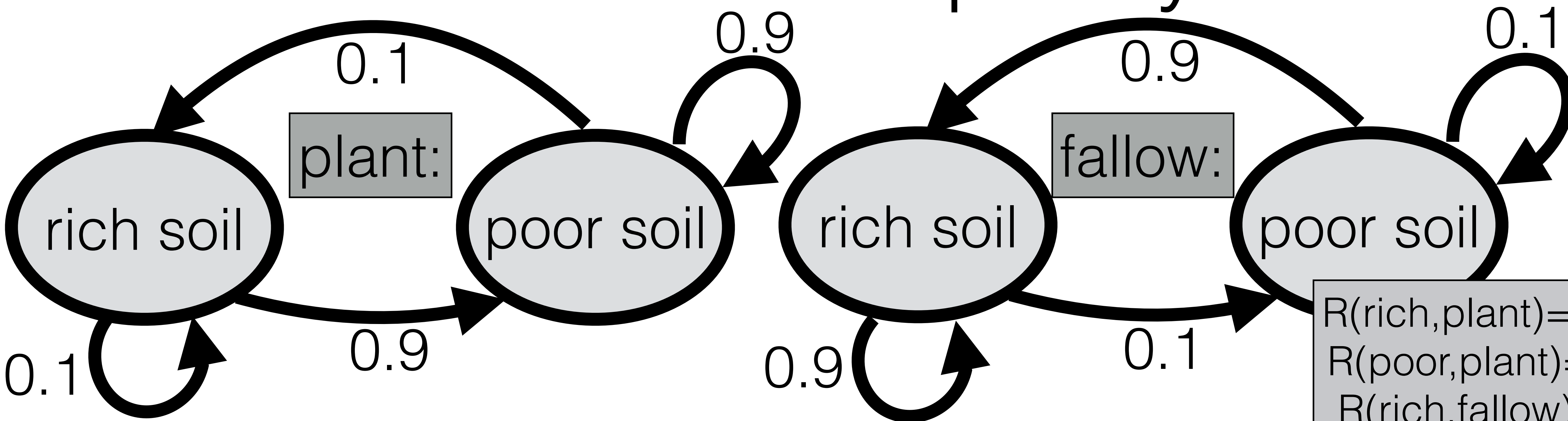
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

$$V_{\pi_A}^2(\text{rich}) = R(\text{rich}, \pi_A(\text{rich})) + T(\text{rich}, \pi_A(\text{rich}), \text{rich}) V_{\pi_A}^1(\text{rich})$$
$$+ T(\text{rich}, \pi_A(\text{rich}), \text{poor}) V_{\pi_A}^1(\text{poor})$$

# 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 $\pi$ starting at *s*

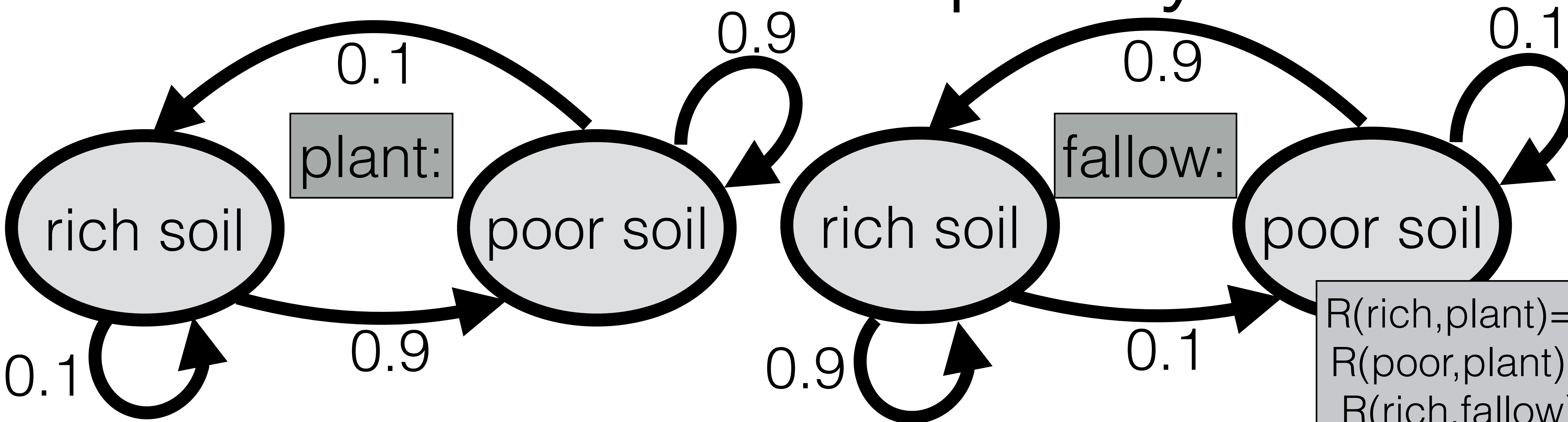
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

$$V_{\pi_A}^2(\text{rich}) = R(\text{rich}, \pi_A(\text{rich})) + T(\text{rich}, \pi_A(\text{rich}), \text{rich}) V_{\pi_A}^1(\text{rich})$$
$$+ T(\text{rich}, \pi_A(\text{rich}), \text{poor}) V_{\pi_A}^1(\text{poor})$$

# 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 $\pi$ starting at *s*

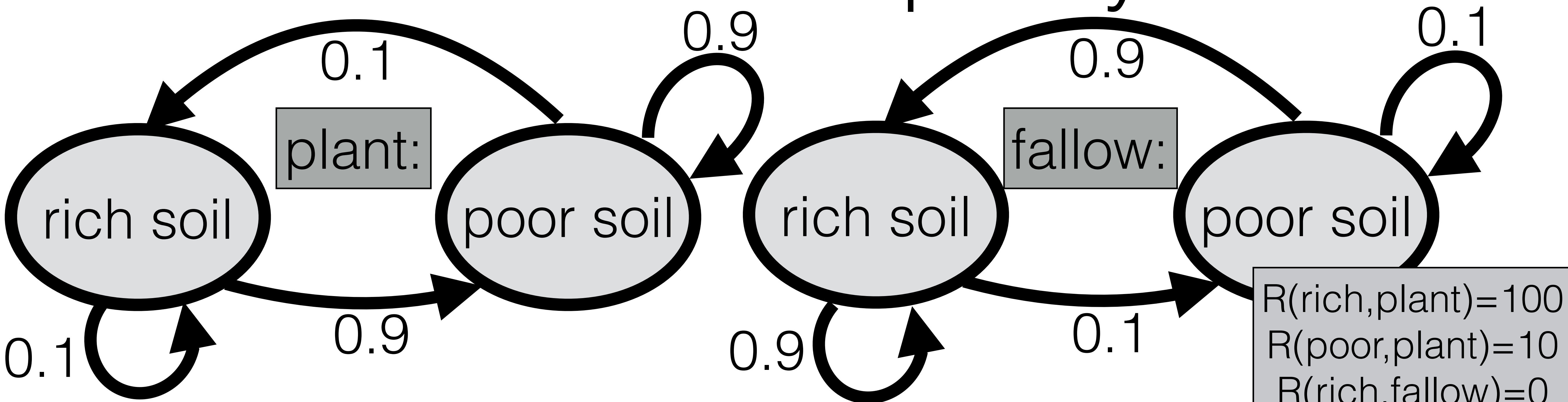
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

$$\begin{aligned} V_{\pi_A}^2(\text{rich}) = R(\text{rich}, \pi_A(\text{rich})) &+ T(\text{rich}, \pi_A(\text{rich}), \text{rich}) V_{\pi_A}^1(\text{rich}) \\ &+ T(\text{rich}, \pi_A(\text{rich}), \text{poor}) V_{\pi_A}^1(\text{poor}) \end{aligned}$$

# What's the value of a policy?




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

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

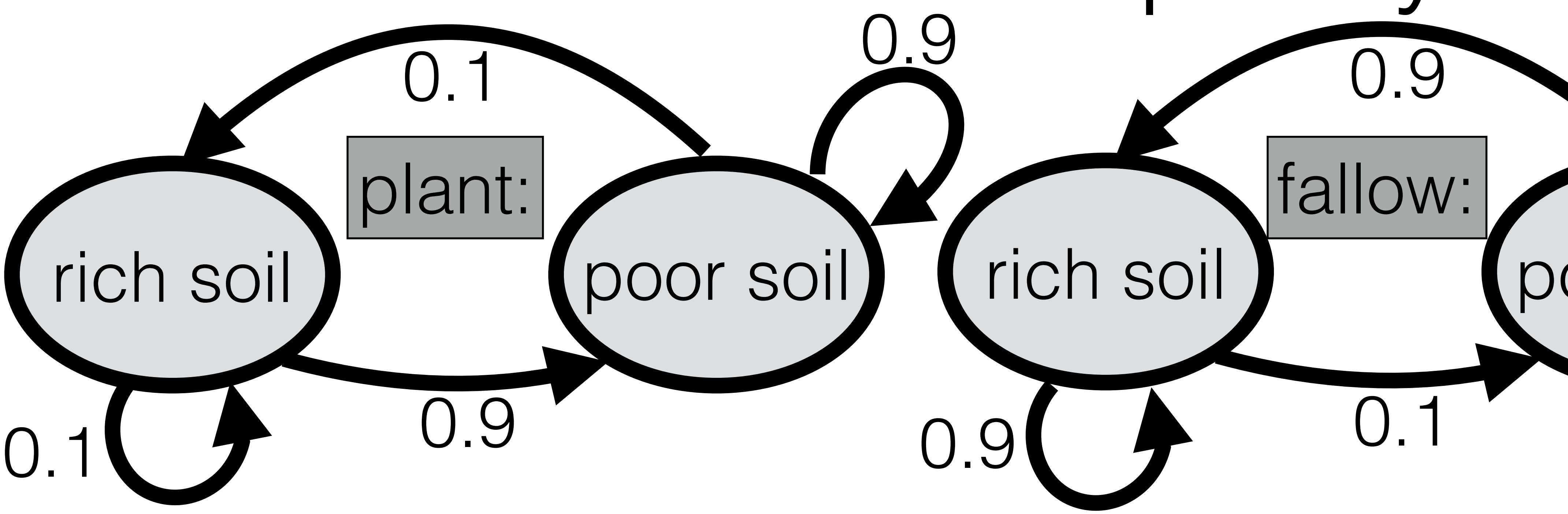
$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_\pi^{h-1}(s')$$

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

$$\begin{aligned} V_{\pi_A}^2(\text{rich}) = R(\text{rich}, \pi_A(\text{rich})) &+ T(\text{rich}, \pi_A(\text{rich}), \text{rich}) V_{\pi_A}^1(\text{rich}) \\ &+ T(\text{rich}, \pi_A(\text{rich}), \text{poor}) V_{\pi_A}^1(\text{poor}) \end{aligned}$$

$$= 100 + (0.1)(100) + (0.9)(10)$$

# 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 $\pi$ starting at *s*

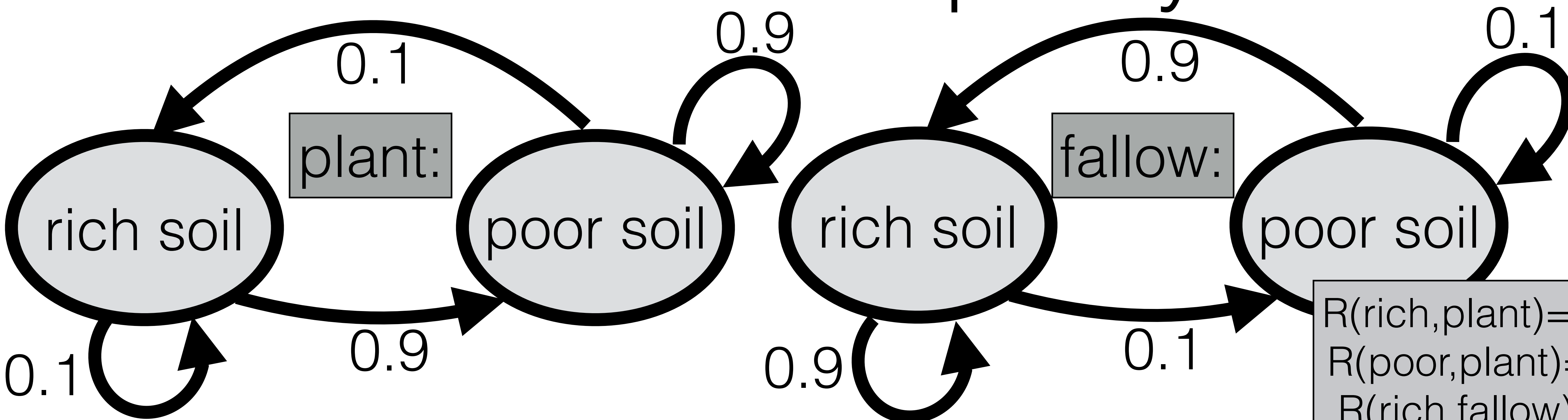
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

$$V_{\pi_A}^2(\text{rich}) = R(\text{rich}, \pi_A(\text{rich})) + T(\text{rich}, \pi_A(\text{rich}), \text{rich}) V_{\pi_A}^1(\text{rich})$$
$$+ T(\text{rich}, \pi_A(\text{rich}), \text{poor}) V_{\pi_A}^1(\text{poor})$$
$$= 100 + (0.1)(100) + (0.9)(10)$$

# 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 $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

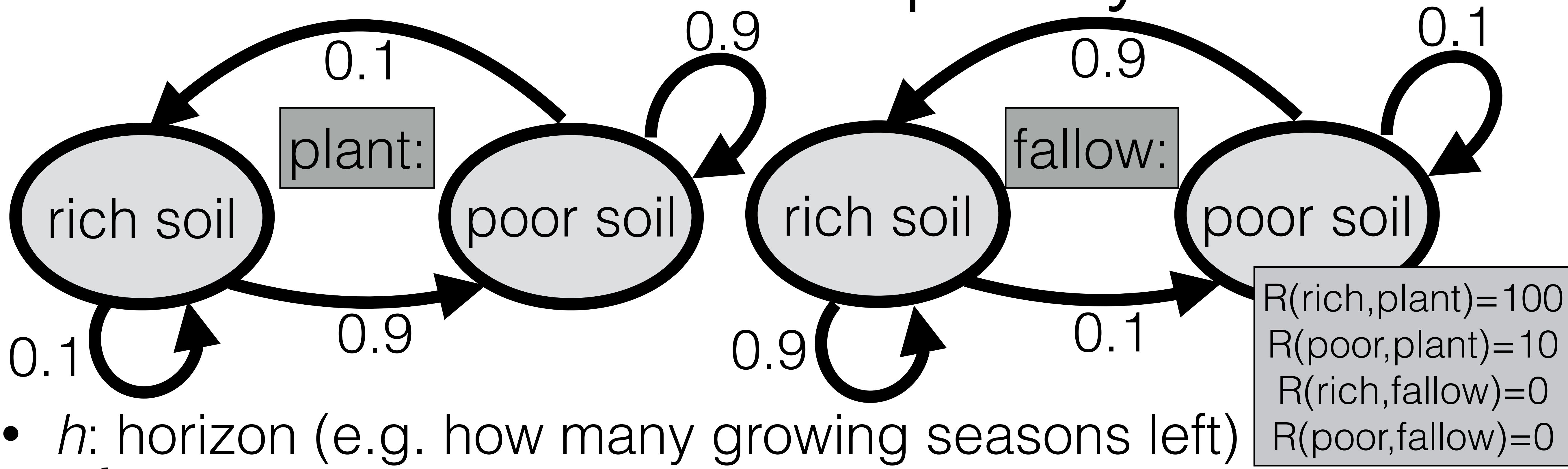
$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_\pi^{h-1}(s')$$

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

$$V_{\pi_A}^2(\text{rich}) = R(\text{rich}, \pi_A(\text{rich})) + T(\text{rich}, \pi_A(\text{rich}), \text{rich}) V_{\pi_A}^1(\text{rich})$$
$$+ T(\text{rich}, \pi_A(\text{rich}), \text{poor}) V_{\pi_A}^1(\text{poor})$$

= 100 + (0.1)(100) + (0.9)(10)

# What's the value of a policy?




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

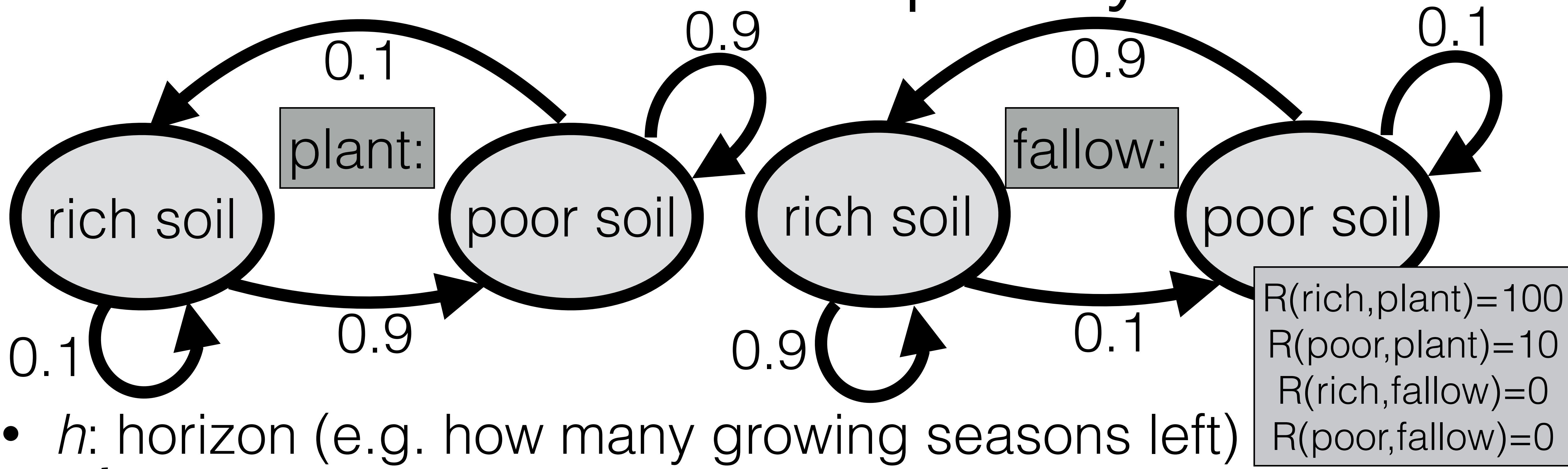
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

$$V_{\pi_A}^2(\text{rich}) = R(\text{rich}, \pi_A(\text{rich})) + T(\text{rich}, \pi_A(\text{rich}), \text{rich}) V_{\pi_A}^1(\text{rich})$$
$$+ T(\text{rich}, \pi_A(\text{rich}), \text{poor}) V_{\pi_A}^1(\text{poor})$$
$$= 100 + (0.1)(100) + (0.9)(10)$$

# What's the value of a policy?




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

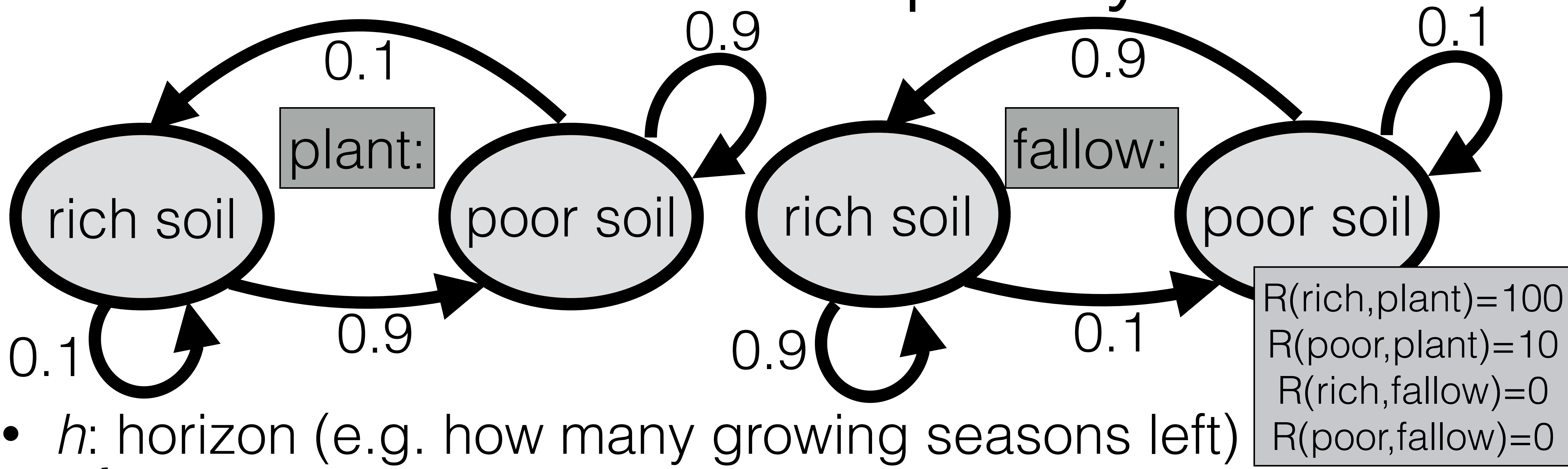
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

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

# What's the value of a policy?




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

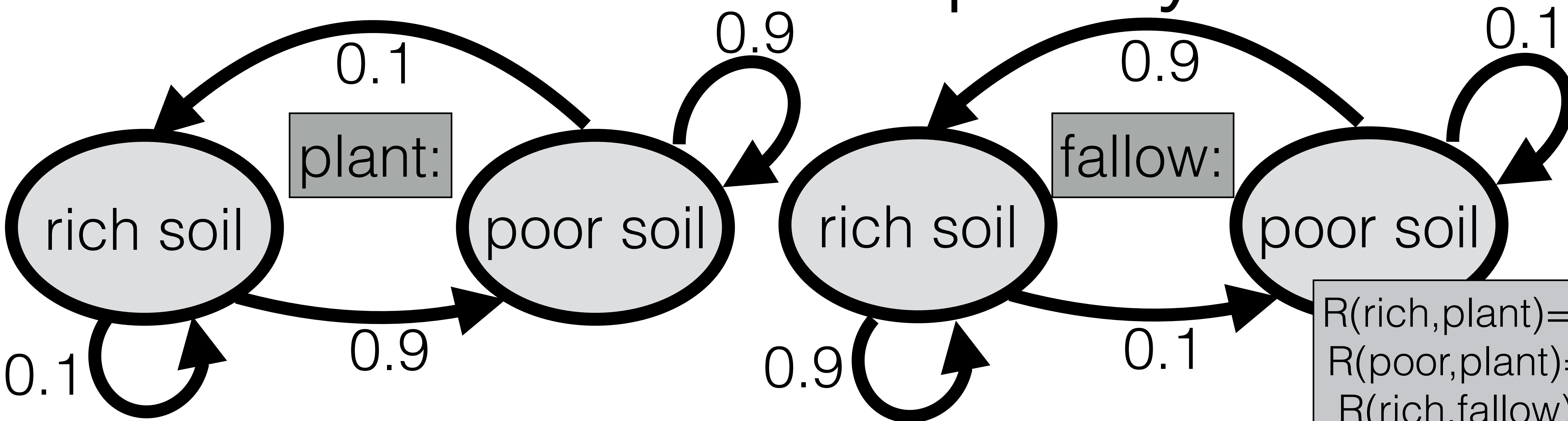
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

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

# 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 $\pi$ starting at *s*

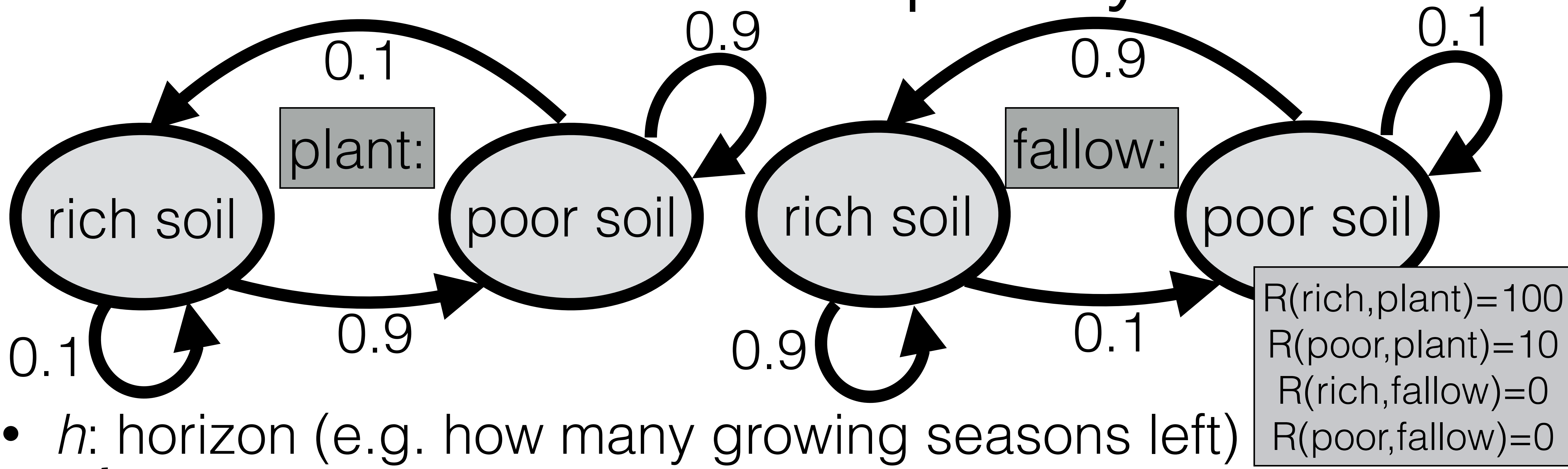
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

$$V_{\pi_A}^2(\text{rich}) = R(\text{rich}, \pi_A(\text{rich})) + T(\text{rich}, \pi_A(\text{rich}), \text{rich}) V_{\pi_A}^1(\text{rich})$$
$$+ T(\text{rich}, \pi_A(\text{rich}), \text{poor}) V_{\pi_A}^1(\text{poor})$$
$$= 100 + (0.1)(100) + (0.9)(10)$$

# What's the value of a policy?




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

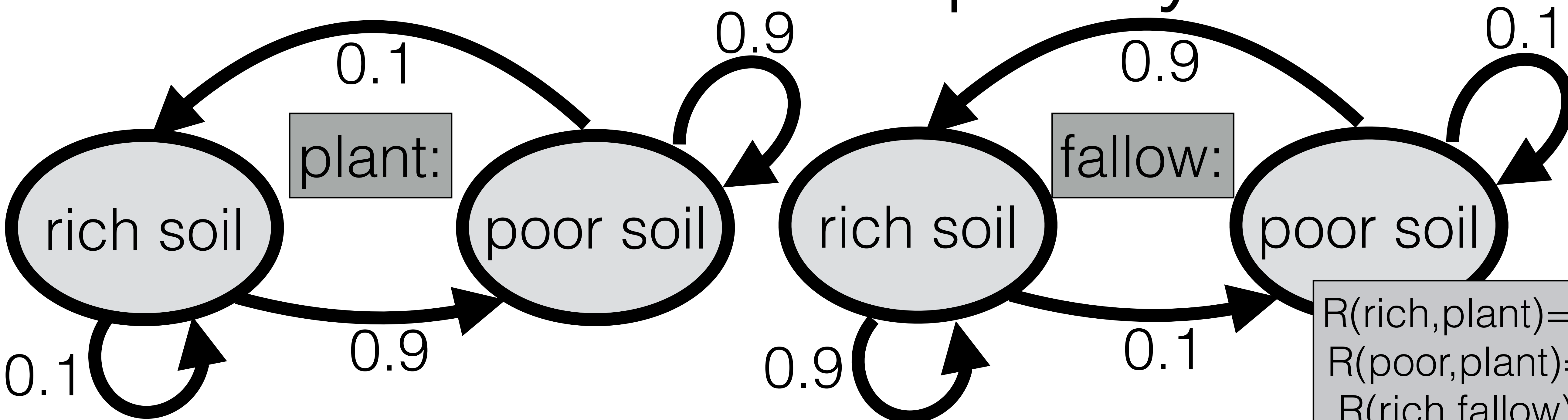
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

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

# 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 $\pi$ starting at *s*

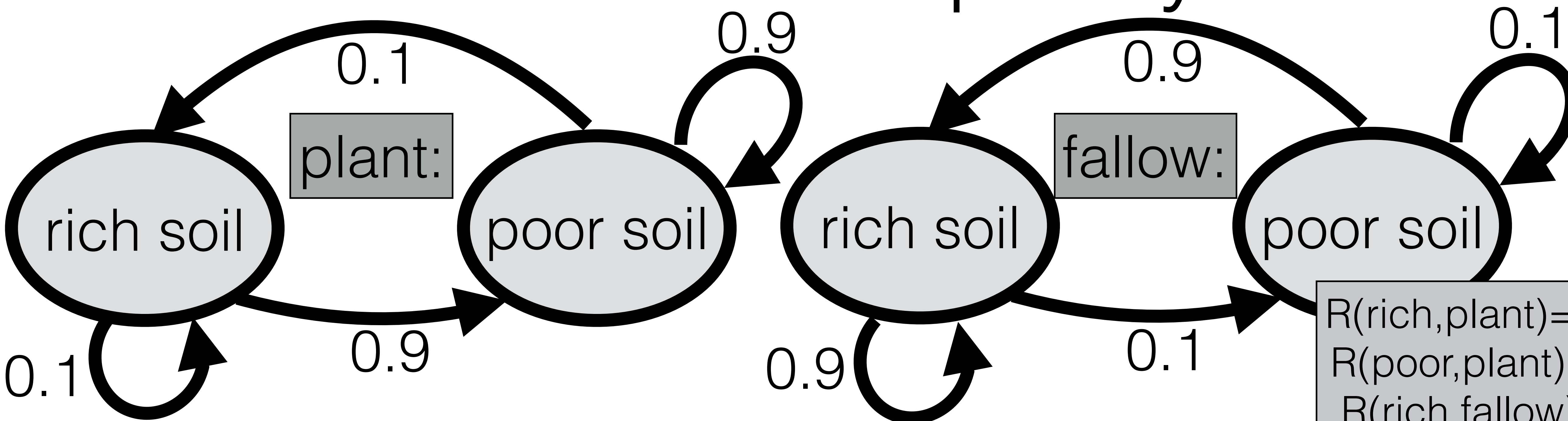
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

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

# 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 $\pi$ starting at *s*

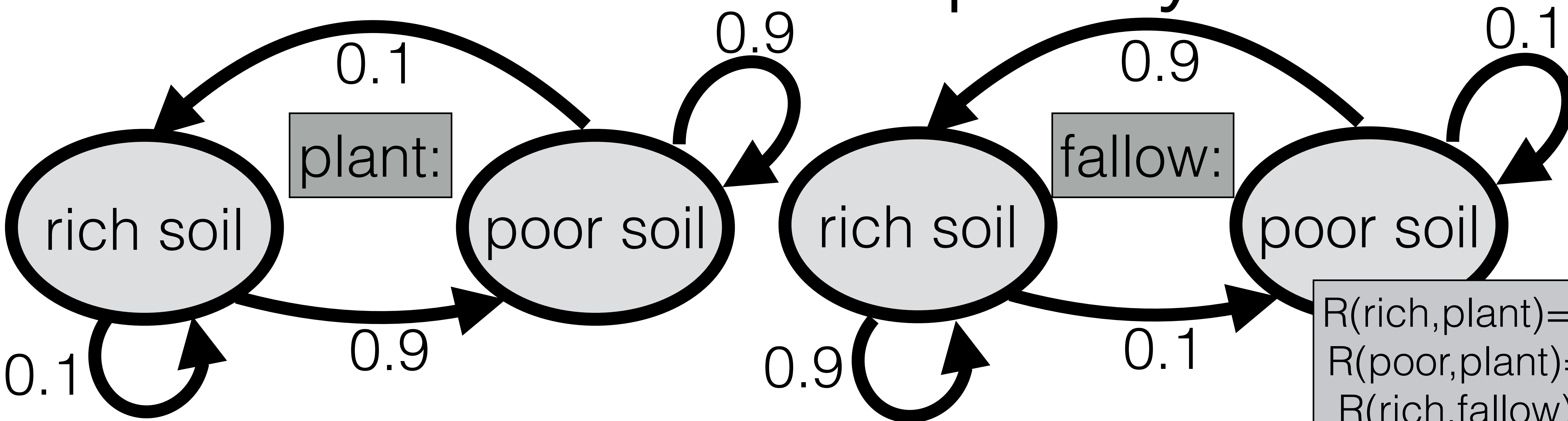
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

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

# 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 $\pi$ starting at *s*

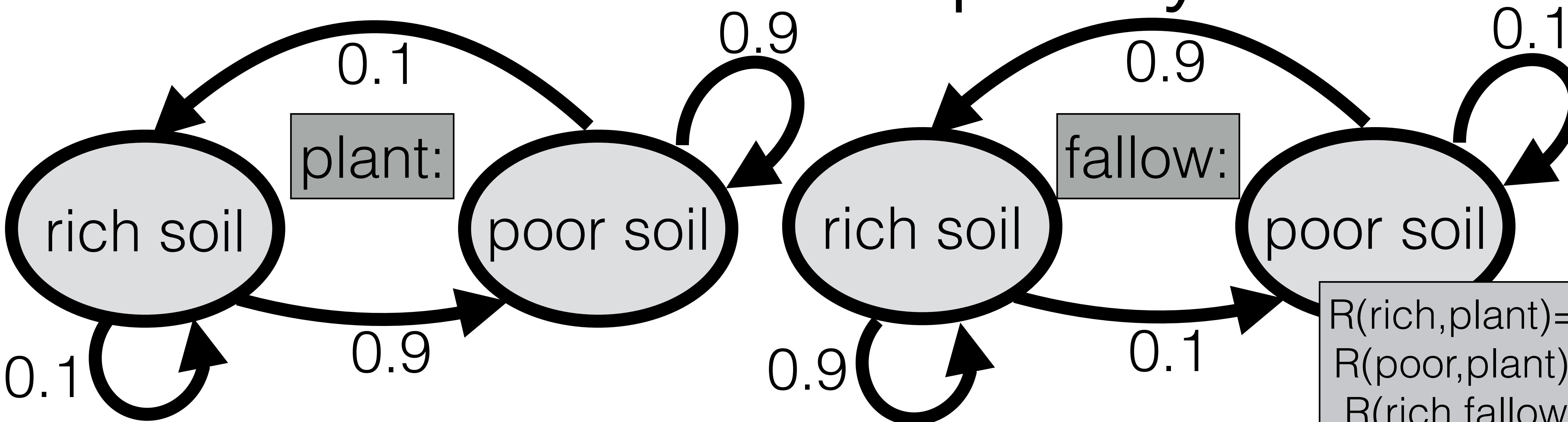
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

$$V_{\pi_A}^2(\text{rich}) = 119$$

# 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 $\pi$ starting at *s*

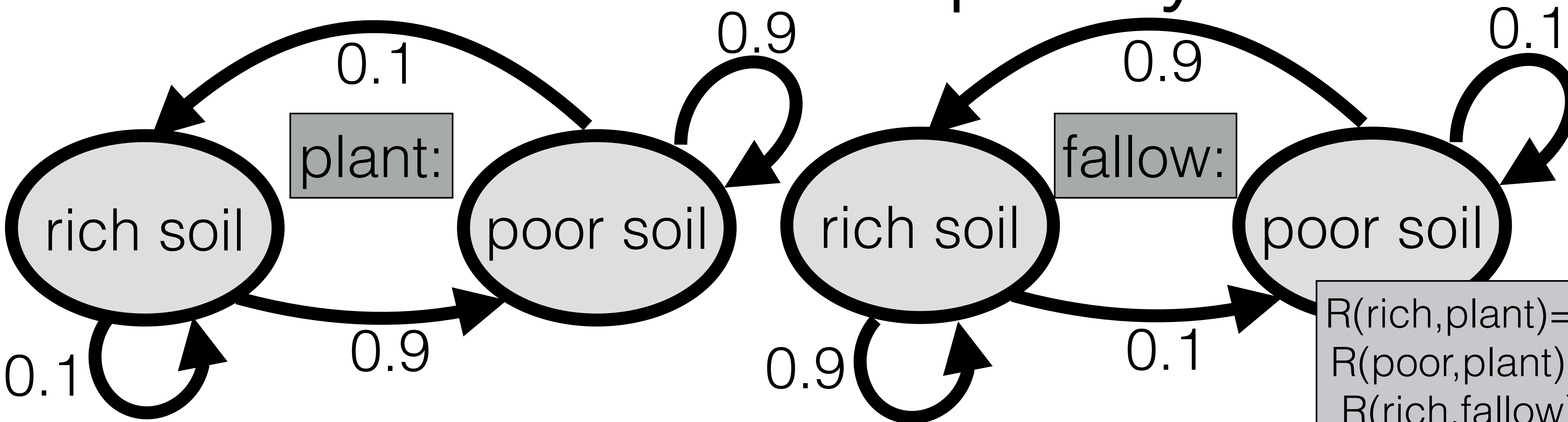
Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

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

# 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 $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

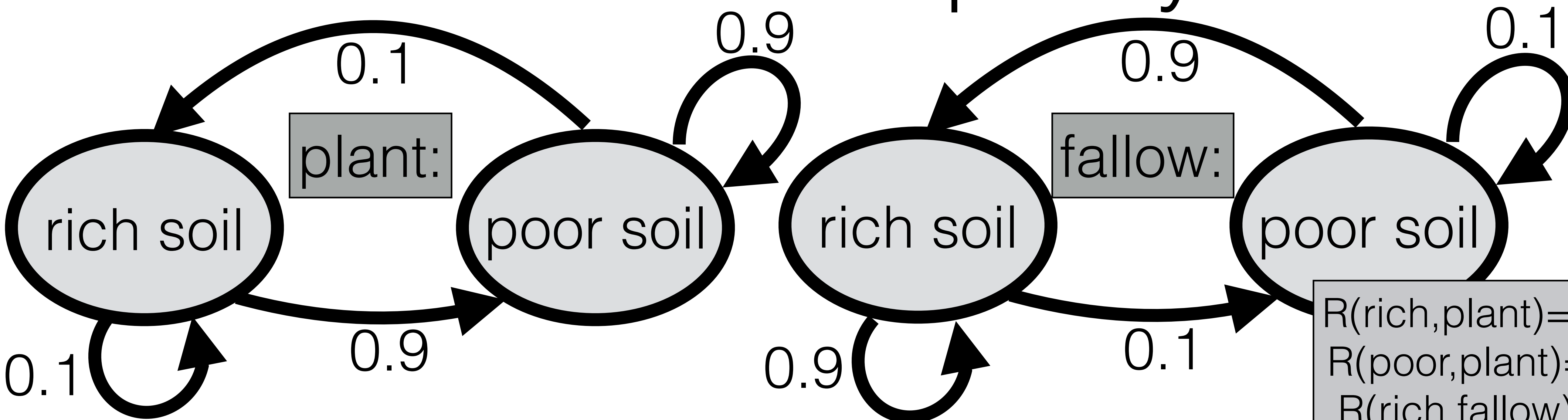
$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_\pi^{h-1}(s')$$

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

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

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

# 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 $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

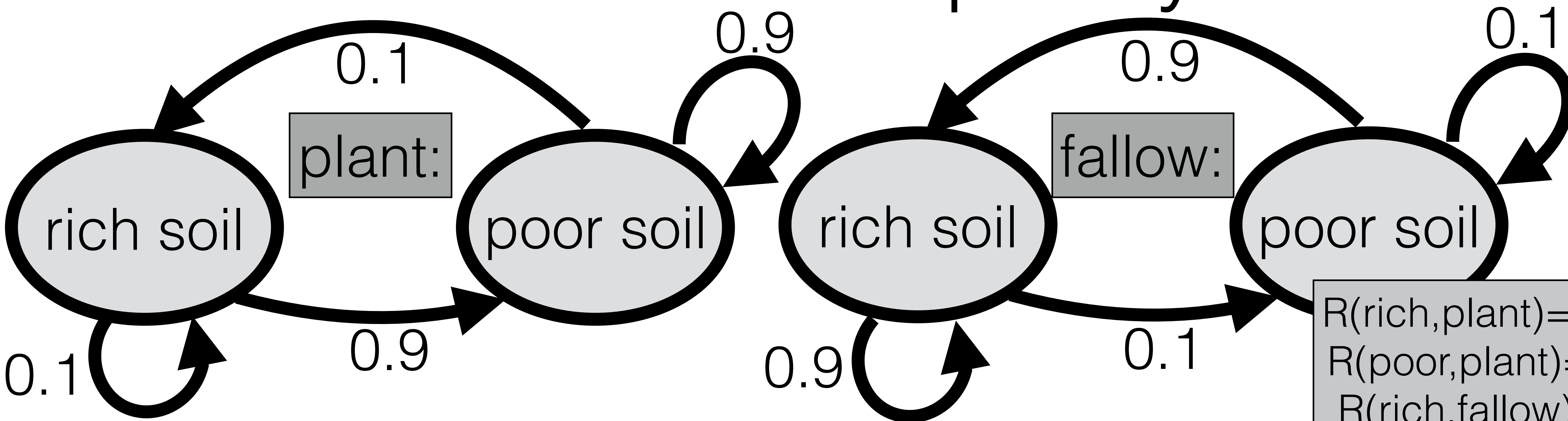
$$V_{\pi_A}^1(\text{rich}) = 100; V_{\pi_A}^1(\text{poor}) = 10; V_{\pi_B}^1(\text{rich}) = 100; V_{\pi_B}^1(\text{poor}) = 0$$

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

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

Who wins?

# 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 $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

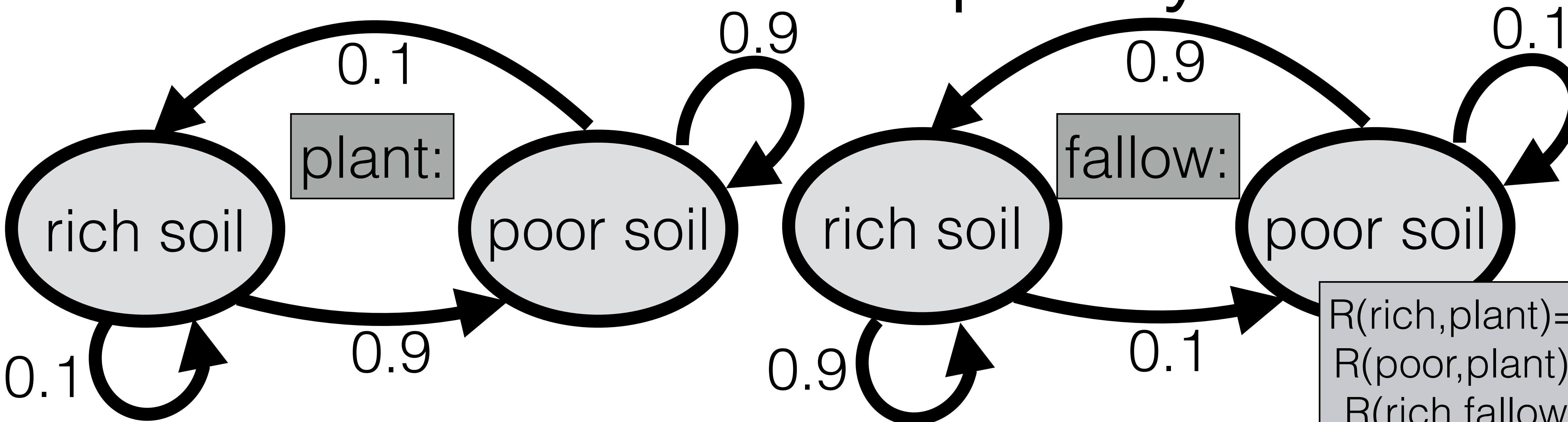
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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 $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

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

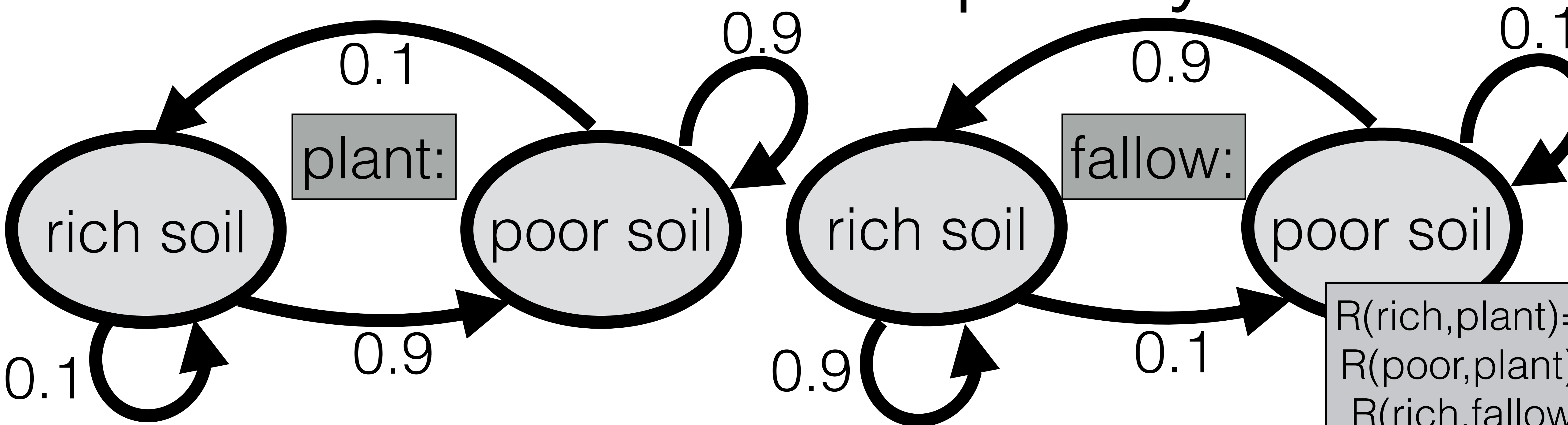
$$V_{\pi_A}^3(\text{rich}) = 138; V_{\pi_A}^3(\text{poor}) = 48; V_{\pi_B}^3(\text{rich}) = 192; V_{\pi_B}^3(\text{poor}) = 108$$

Who wins?

$h=1$

- 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 $\pi$ starting at *s*

**Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow**

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

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

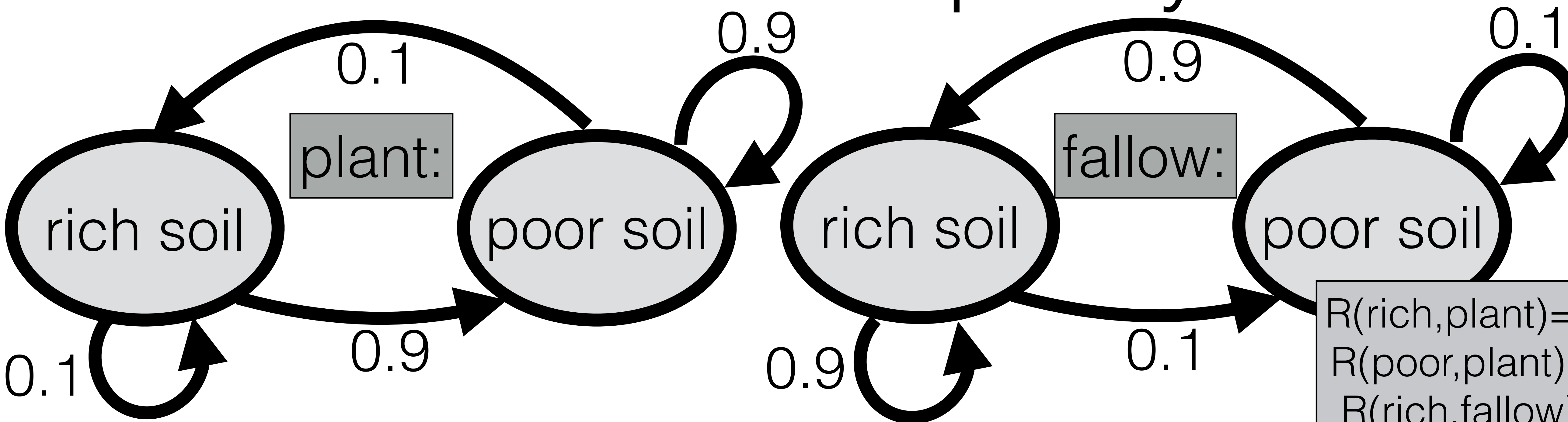
$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$

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

Who wins? $\pi_A >_{h=1} \pi_B$

- 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 $\pi$ starting at *s*

**Dueling farmers!** $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

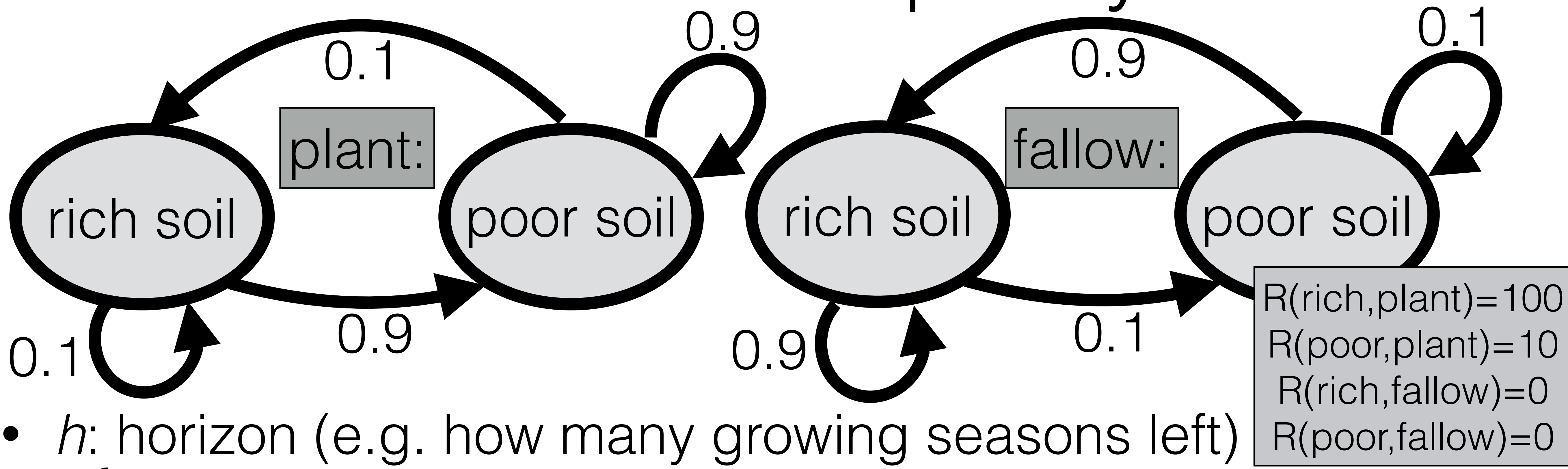
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

**Who wins?** $\pi_A >_{h=1} \pi_B$ $\quad$ $h=3$

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

# What's the value of a policy?




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

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

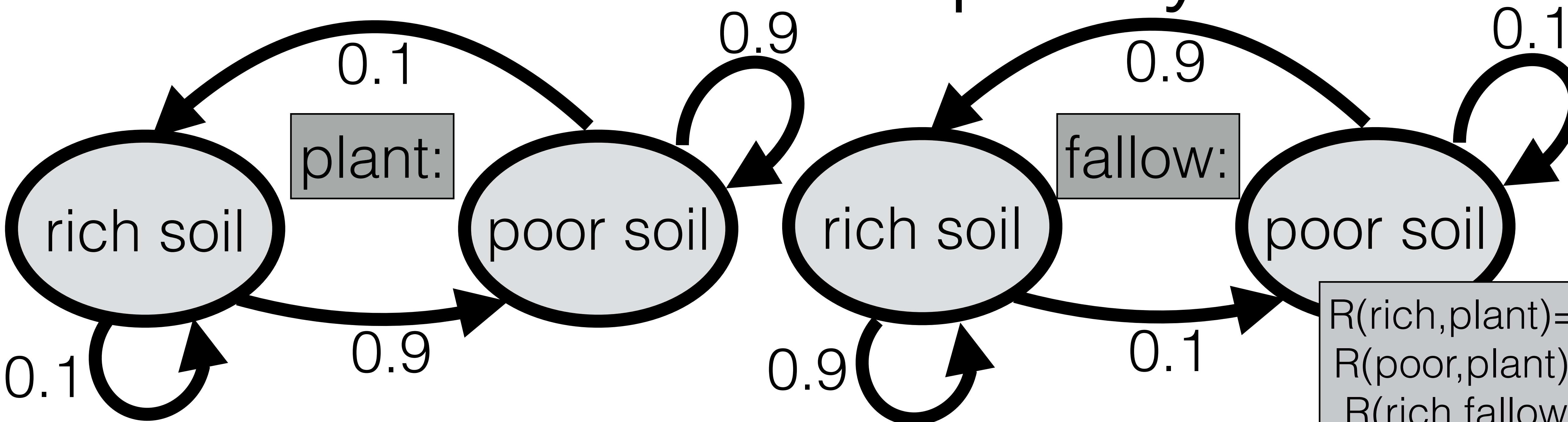
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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

- 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 $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

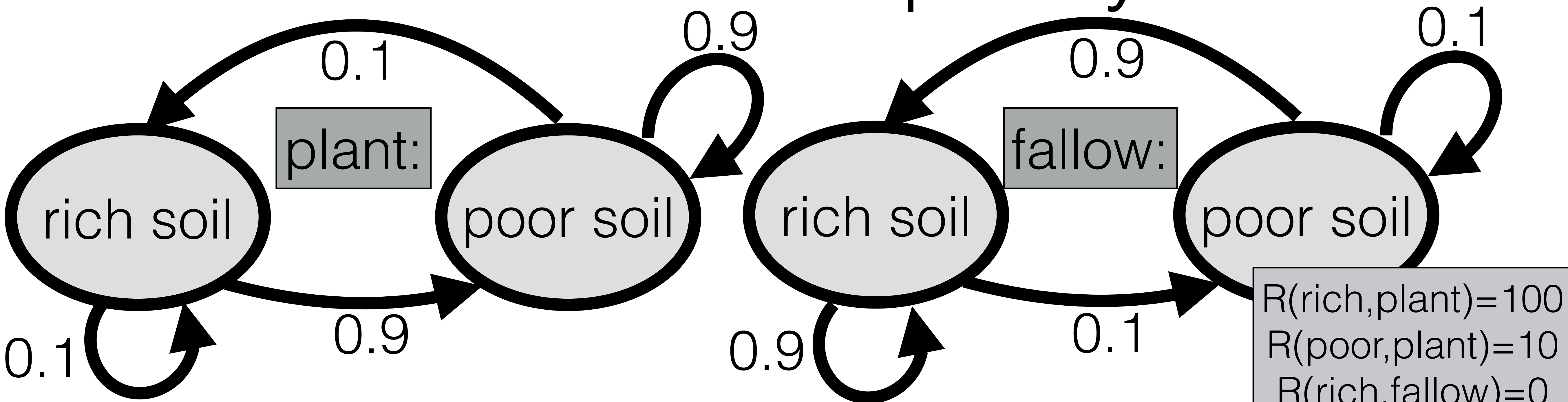
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

Who wins? $\pi_A >_{h=1} \pi_B; \pi_A <_{h=3} \pi_B;$ $h{=}2$

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

# What's the value of a policy?




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

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

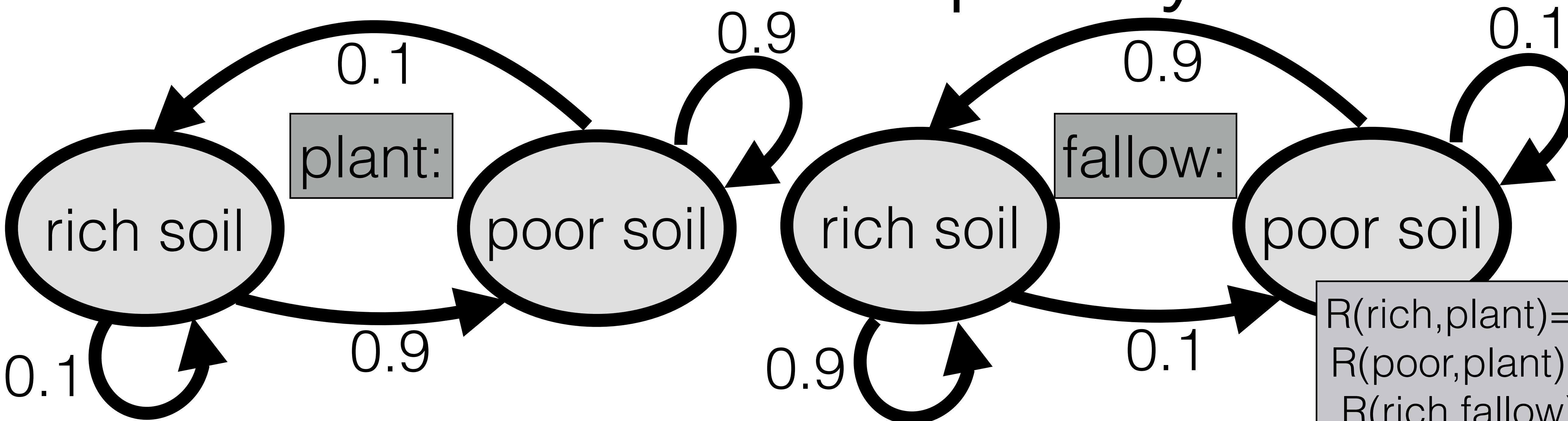
$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$

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

Who wins? $\pi_A >_{h=1} \pi_B; \pi_A <_{h=3} \pi_B;$ No 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

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

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

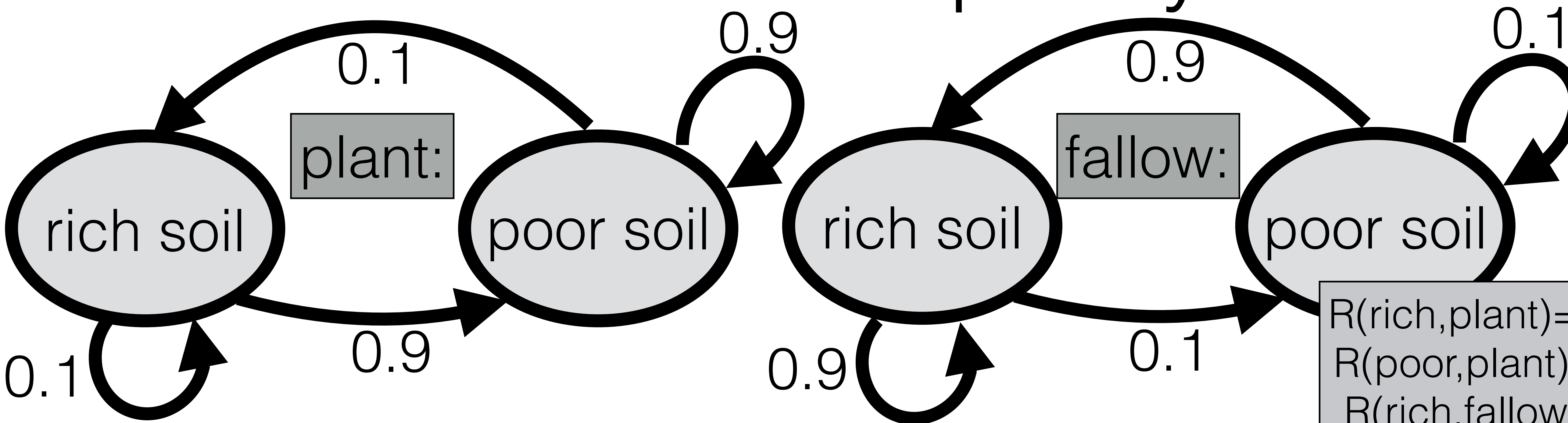
$$V_{\pi_A}^{2}(\text{rich}) = 119; V_{\pi_A}^{2}(\text{poor}) = 29; V_{\pi_B}^{2}(\text{rich}) = 110; V_{\pi_B}^{2}(\text{poor}) = 90$$

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

Who wins? $\pi_A >_{h=1} \pi_B; \pi_A <_{h=3} \pi_B;$ No 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

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

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

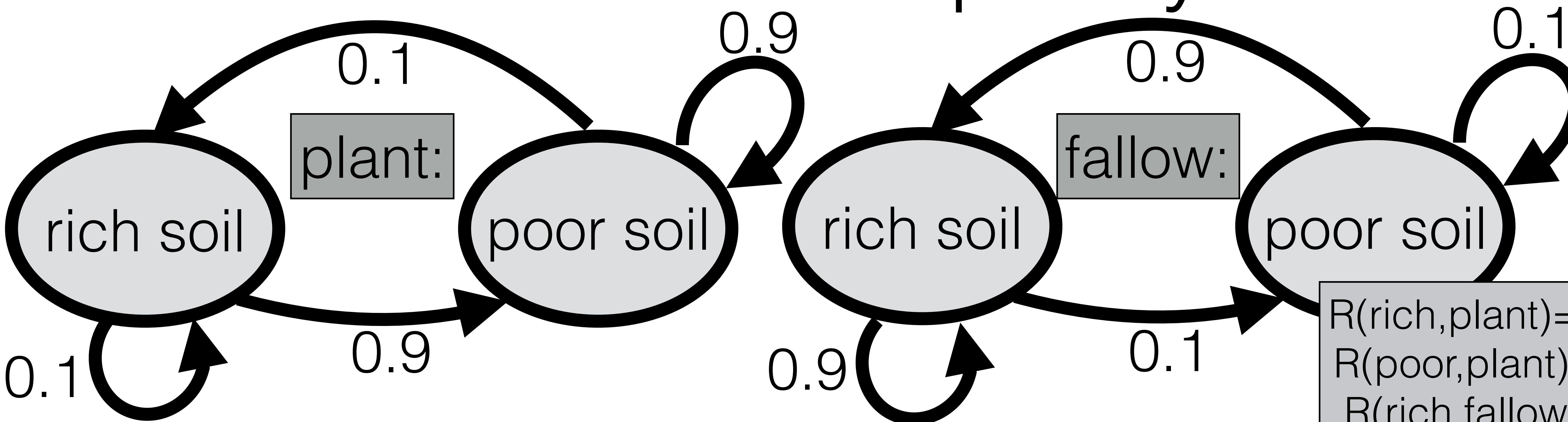
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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

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 $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

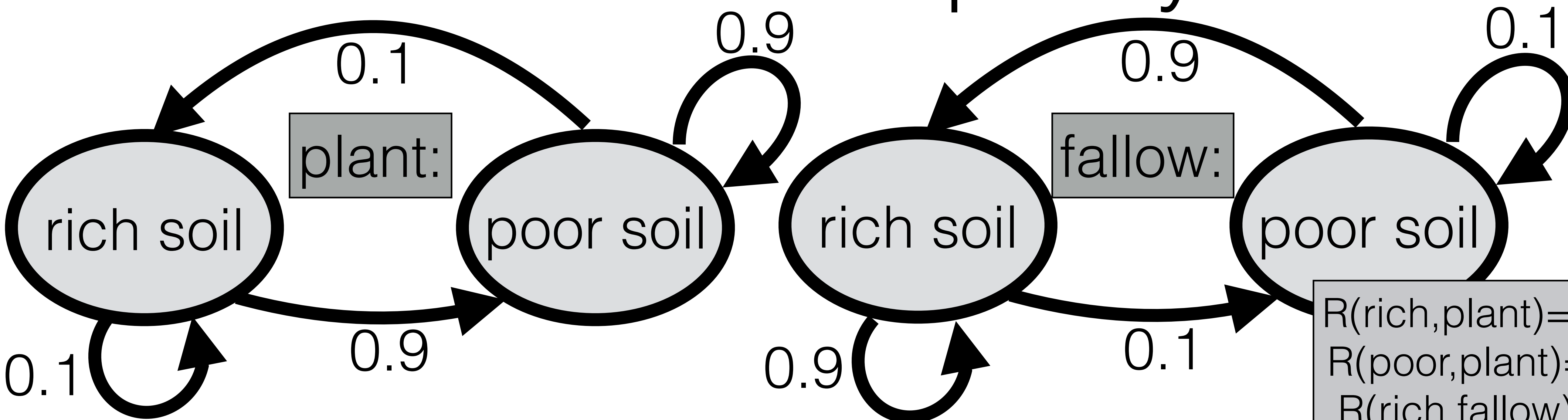
$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$

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

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

- 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 $\pi$ starting at *s*

Dueling farmers! $\pi_A$ : always plant; $\pi_B$: plant if rich, else fallow

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

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

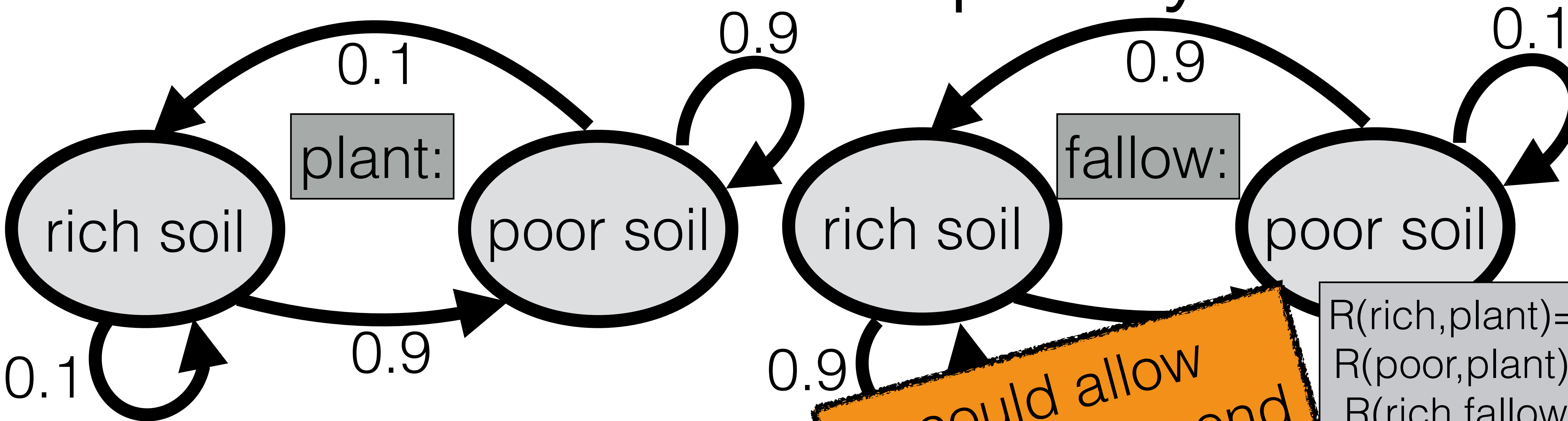
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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

- 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 gro
- $V_\pi^h(s)$: value (expected reward ing at *s*

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

Dueling farmers! $\pi_A$ : always plar ant if rich, else fallow

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

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

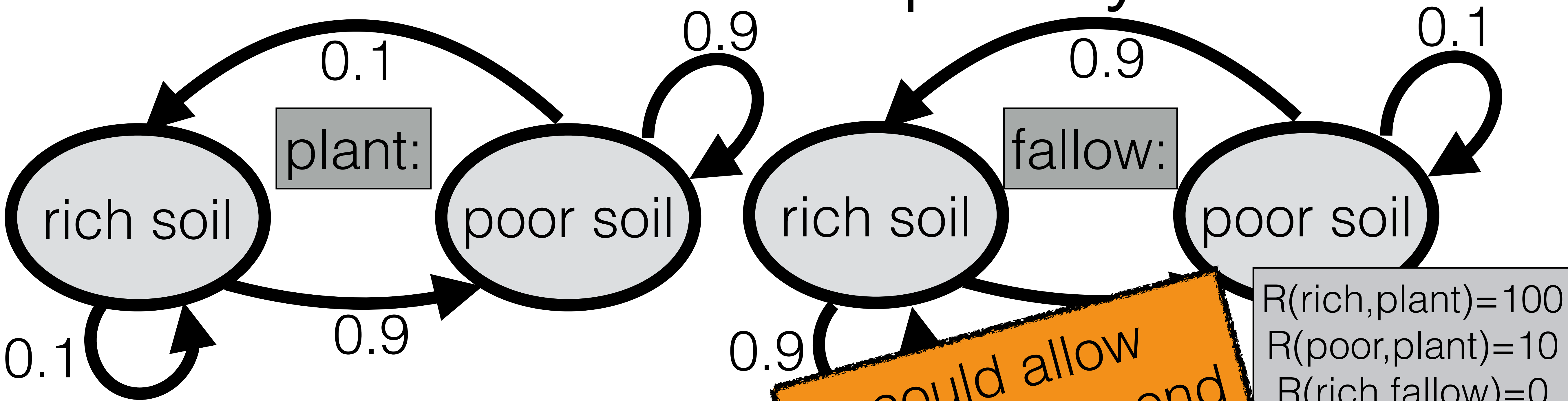
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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

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

# What's the value of a policy?




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

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

Dueling farmers! $\pi_A$: always plan ... nt if rich, else fallow

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

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

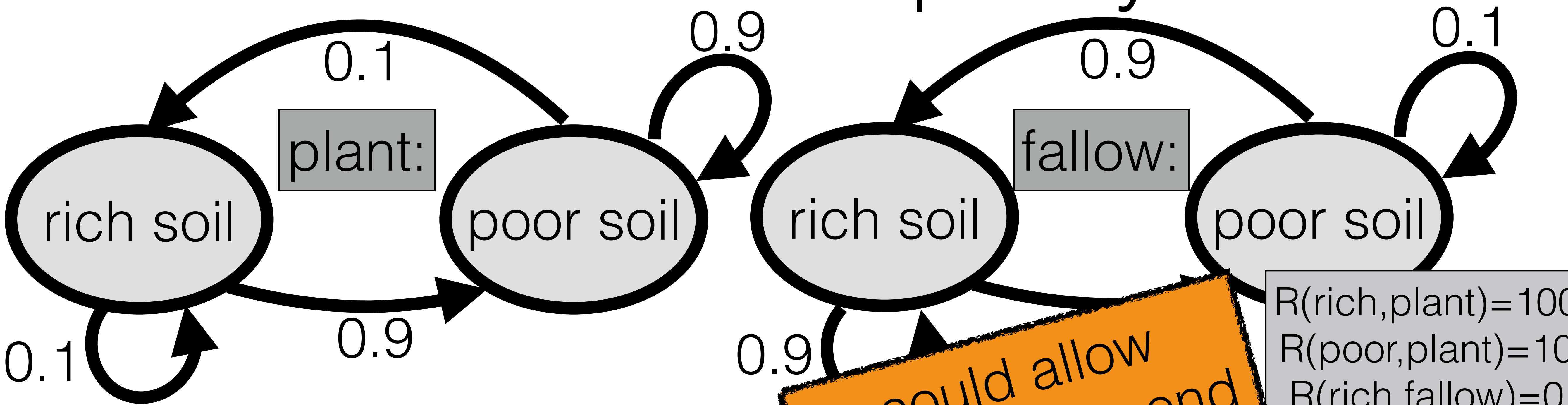
$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$

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

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

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

# What's the value of a policy?




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

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

Dueling farmers! $\pi_A$ : always plan nt if rich, else fallow

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

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

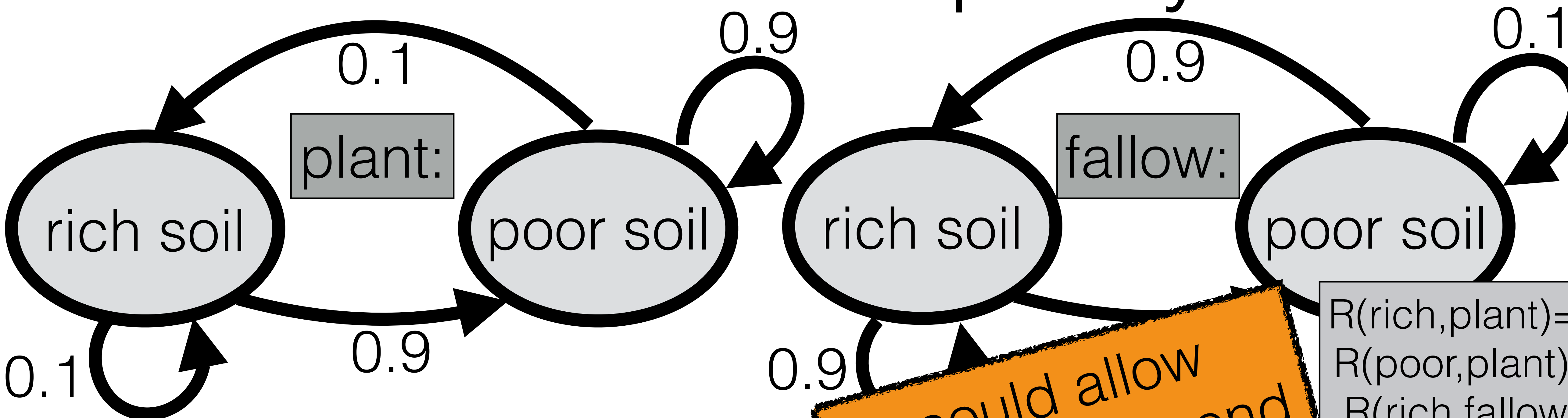
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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

 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! $\pi_A$ : always plan nt if rich, else fallow

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

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

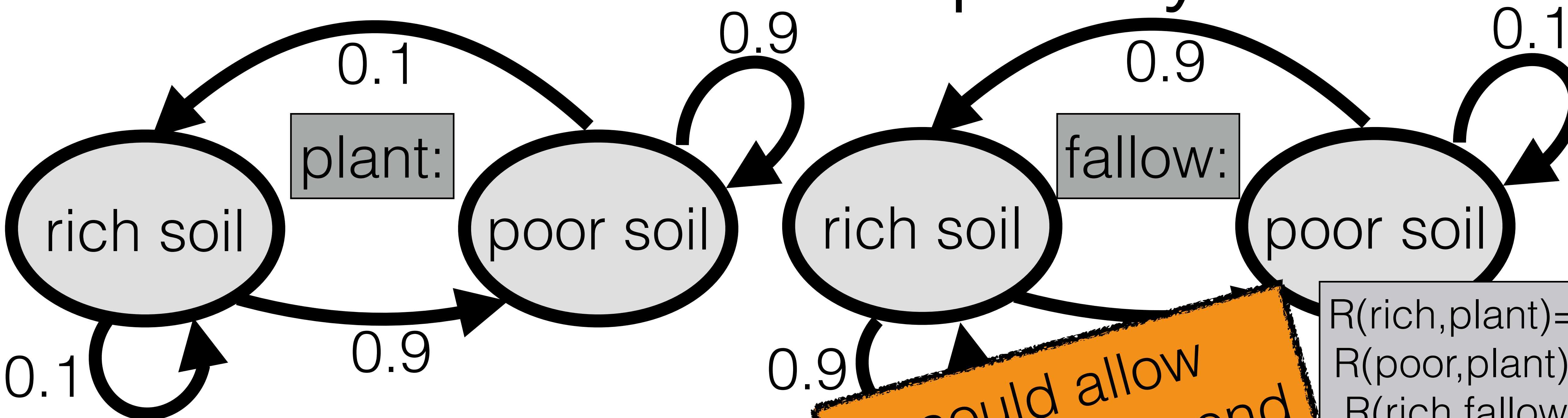
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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

- 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 gro
- $V_\pi^h(s)$: value (expected reward ing at *s*

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

Dueling farmers! $\pi_A$ : always plan nt if rich, else fallow

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

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

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

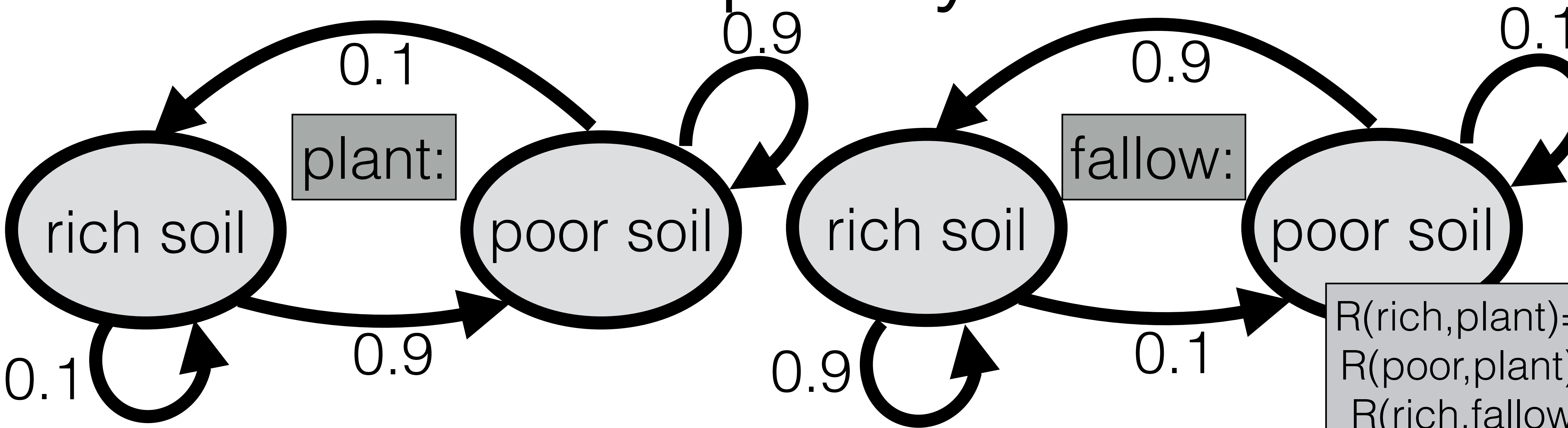
$$V_{\pi_A}^3(\text{rich}) = 138; V_{\pi_A}^3(\text{poor}) = 48; V_{\pi_B}^3(\text{rich}) = 192; V_{\pi_B}^3(\text{poor}) = 108$$

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

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

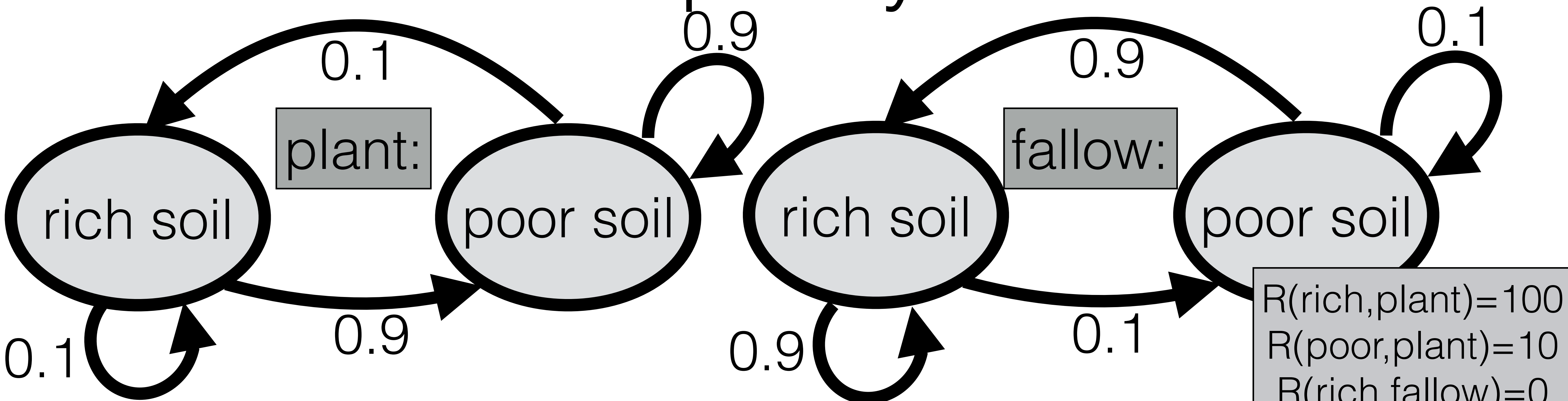
# What's the best policy?

# What's the best policy?



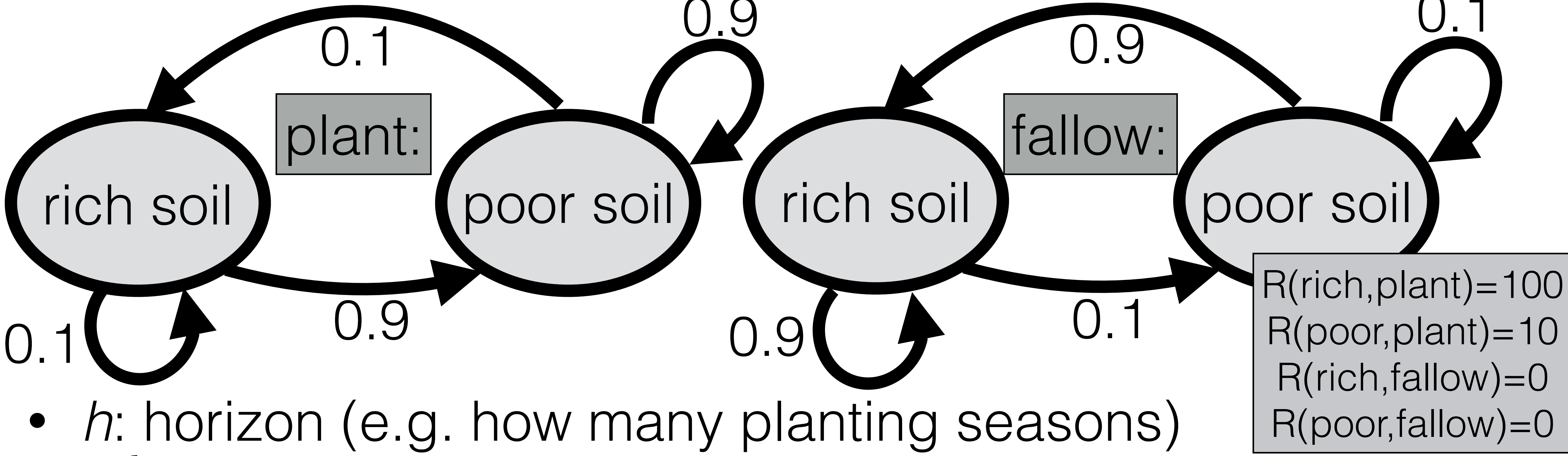

R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

# What's the best policy?



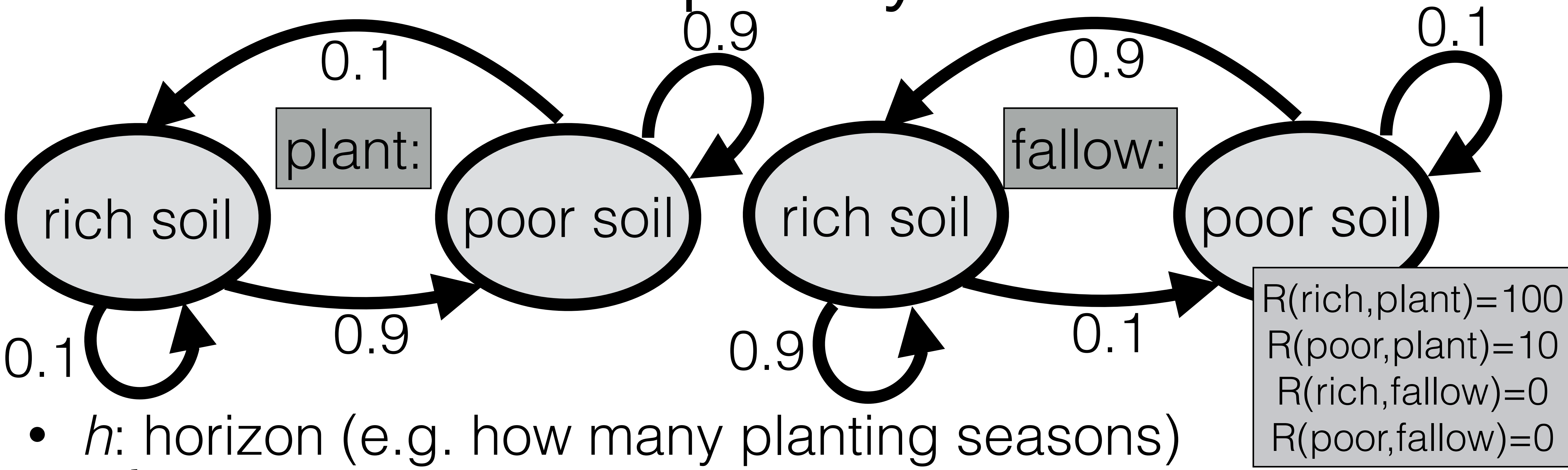

- *h*: horizon (e.g. how many planting seasons)

# What's the best policy?



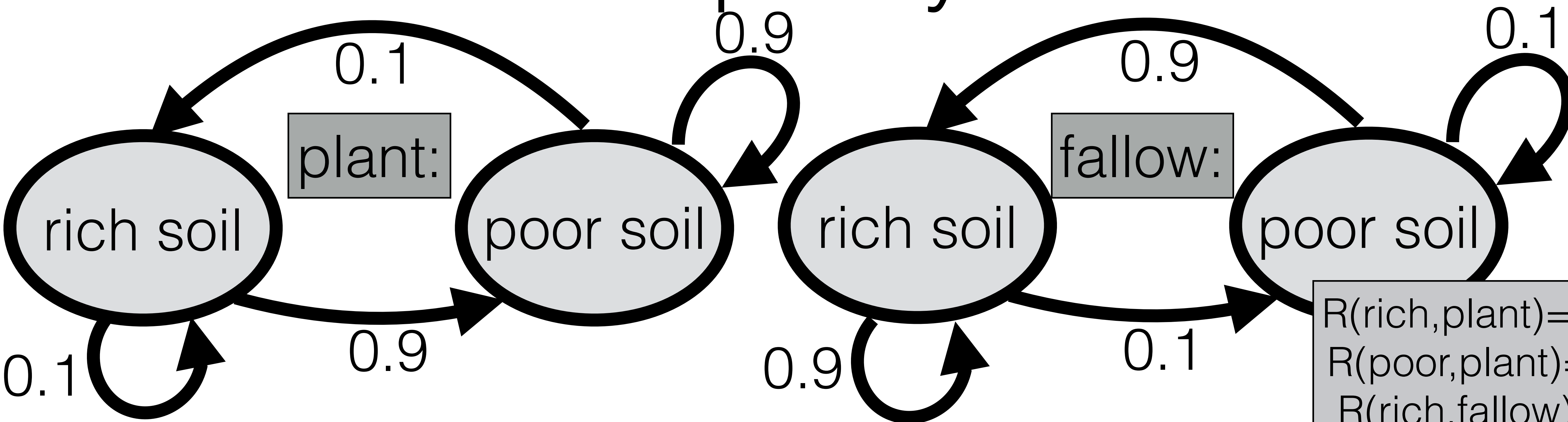

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left

# What's the best policy?




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

# What’s the best policy?

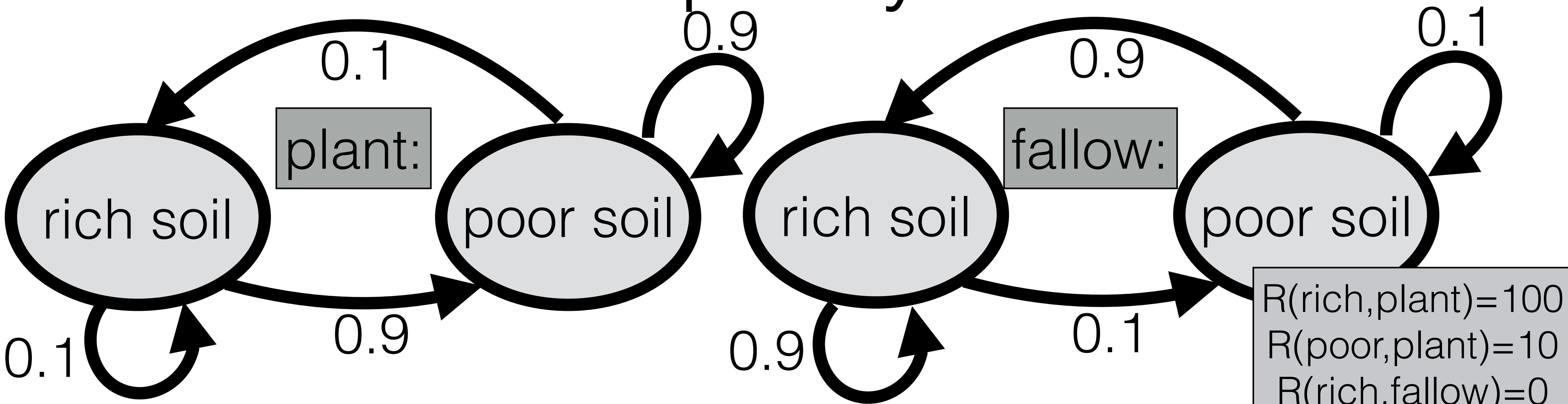



R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *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**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

Compare to: $V_\pi^h(s)$

# What's the best policy?

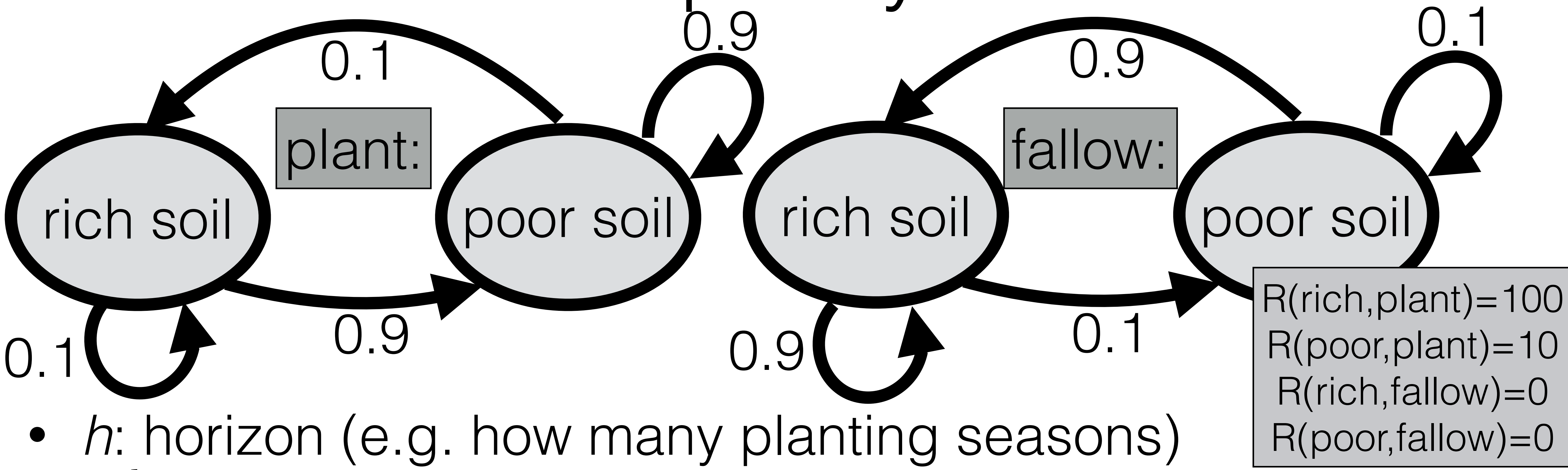



- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s, a)$

Compare to: $V_\pi^h(s)$

Note: there can be more than one optimal policy

# What's the best policy?



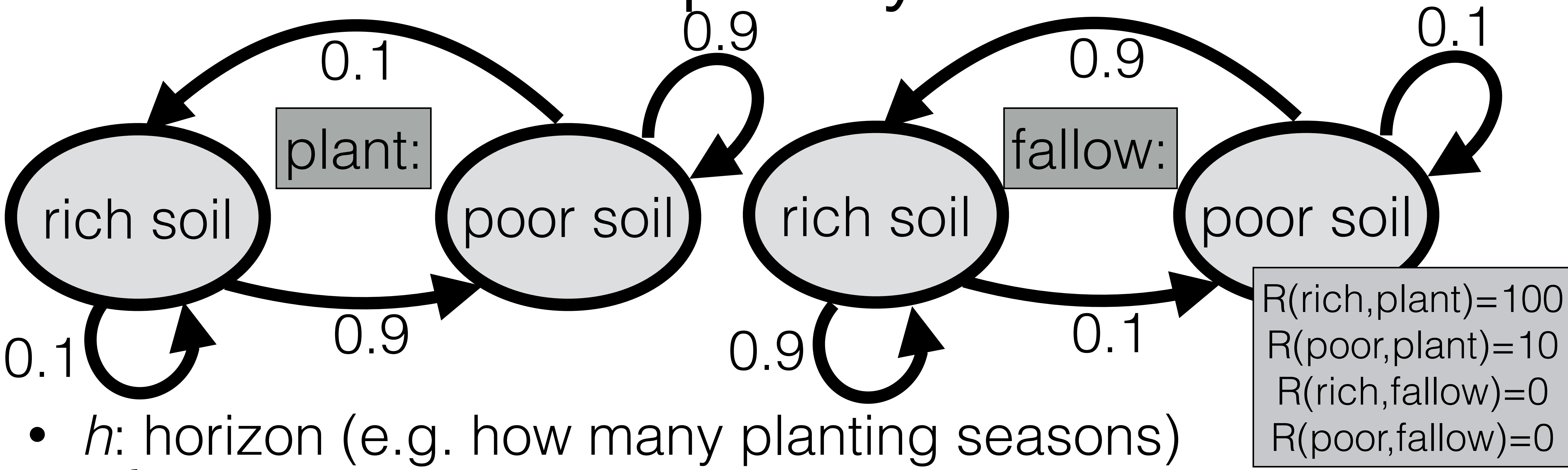

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s, a)$

Compare to: $V_\pi^h(s)$

Note: there can be more than one optimal policy

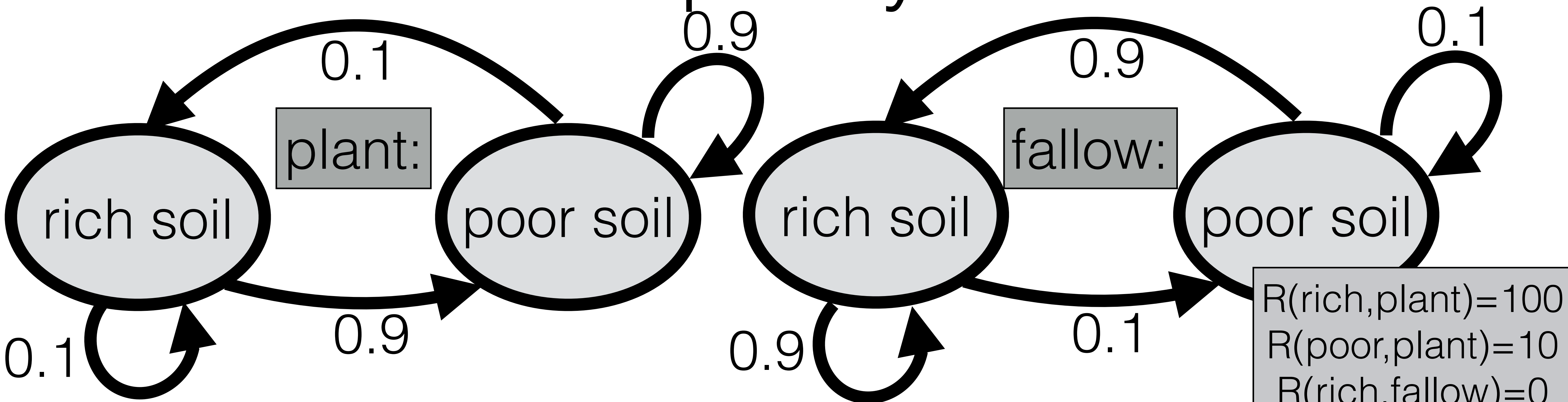
Note: the optimal policy may be non-stationary

# What's the best policy?




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

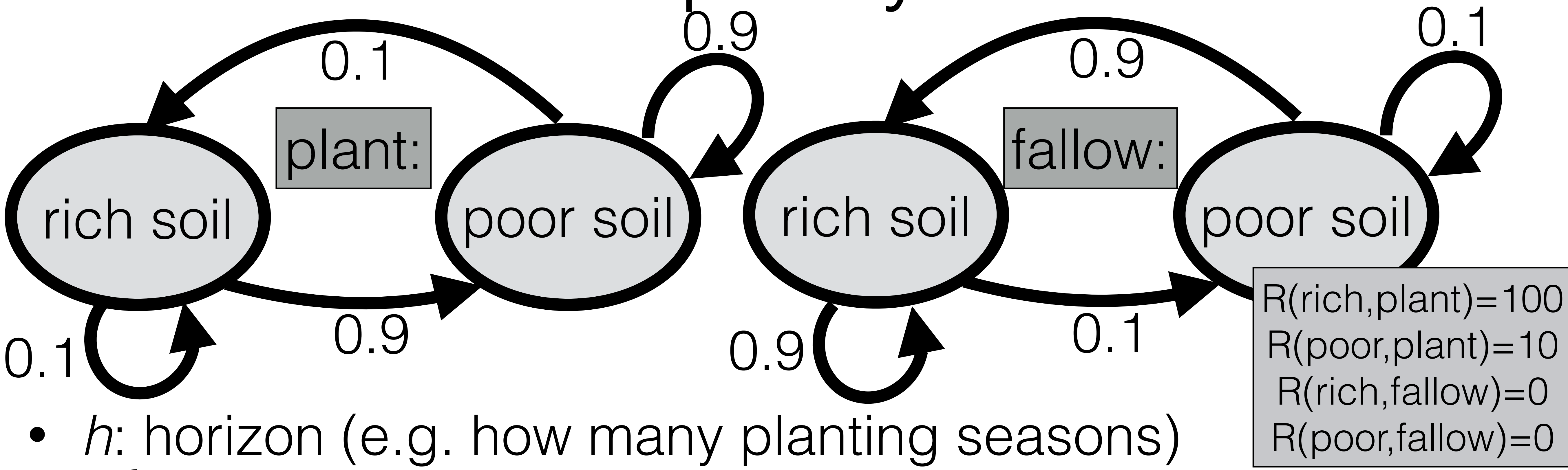
# What's the best policy?




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

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

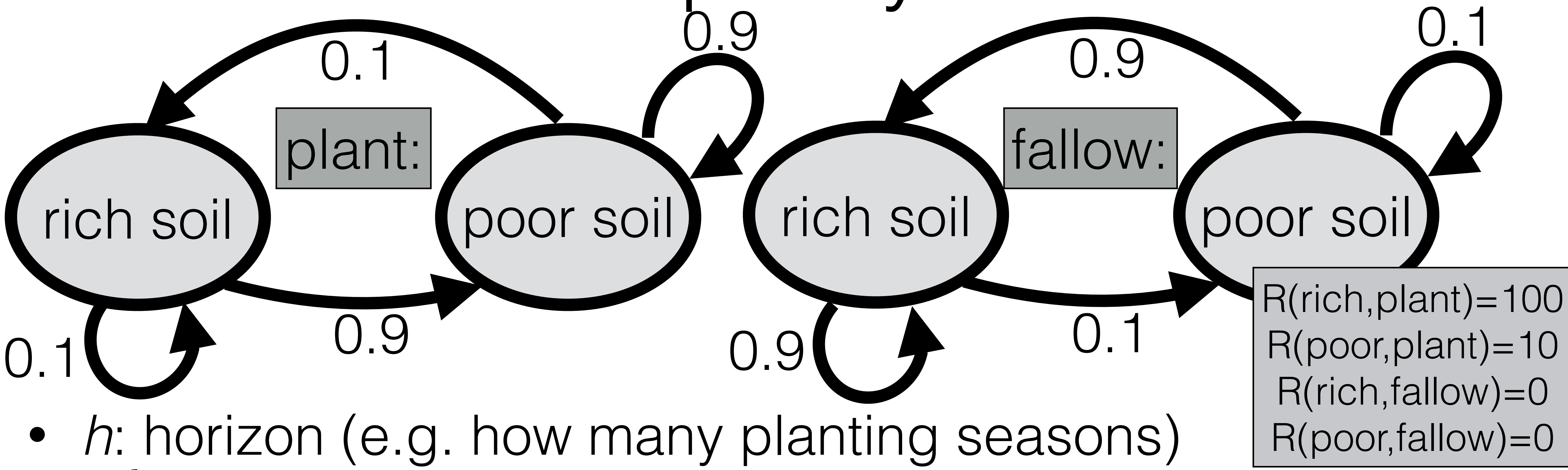
# What’s the best policy?




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

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

# What's the best policy?

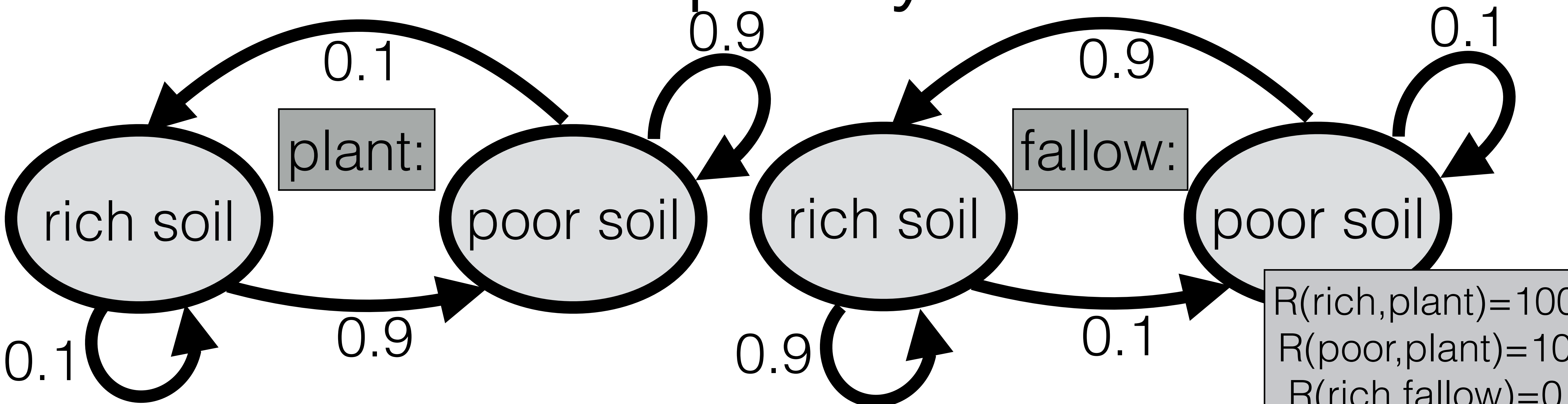



- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

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

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

# What's the best policy?

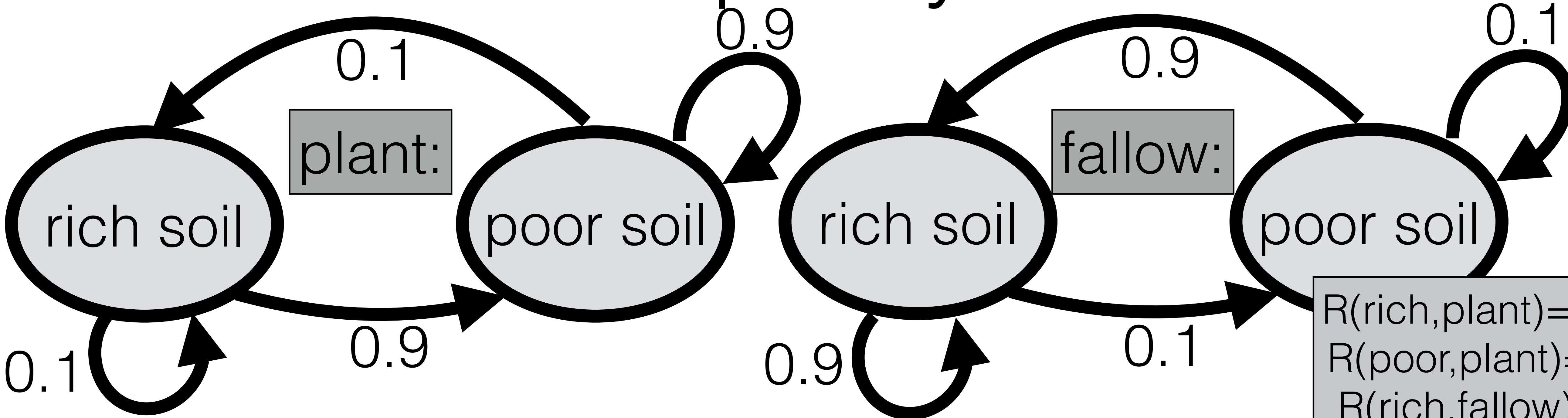



- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With $Q$, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

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

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

# What's the best policy?



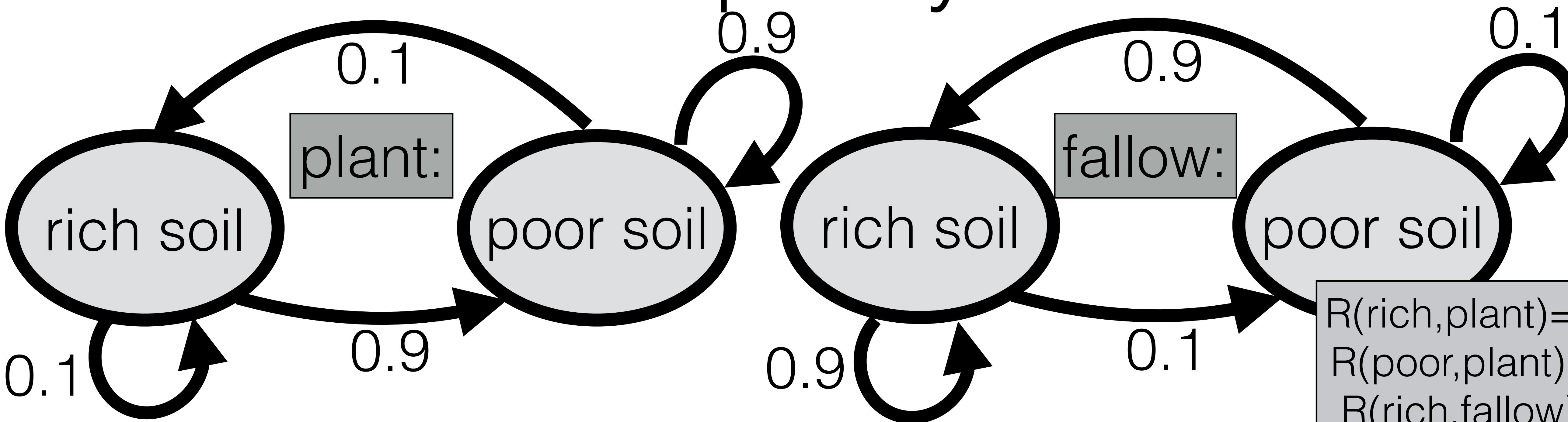

R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

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

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

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

# What's the best policy?




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

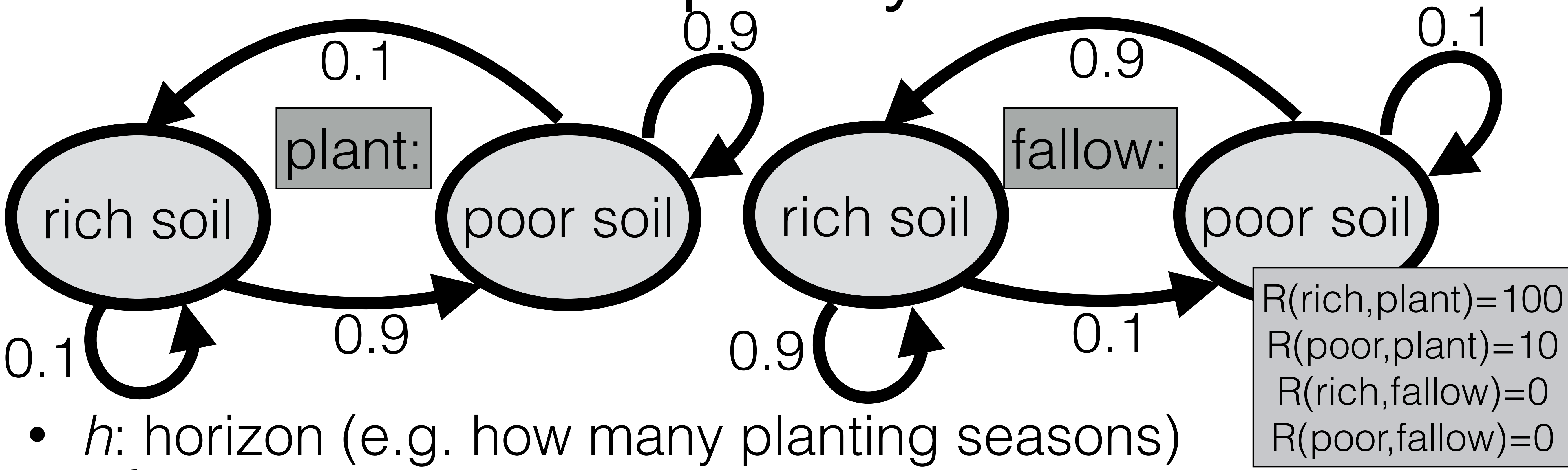
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

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

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

What's best?

# What's the best policy?




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

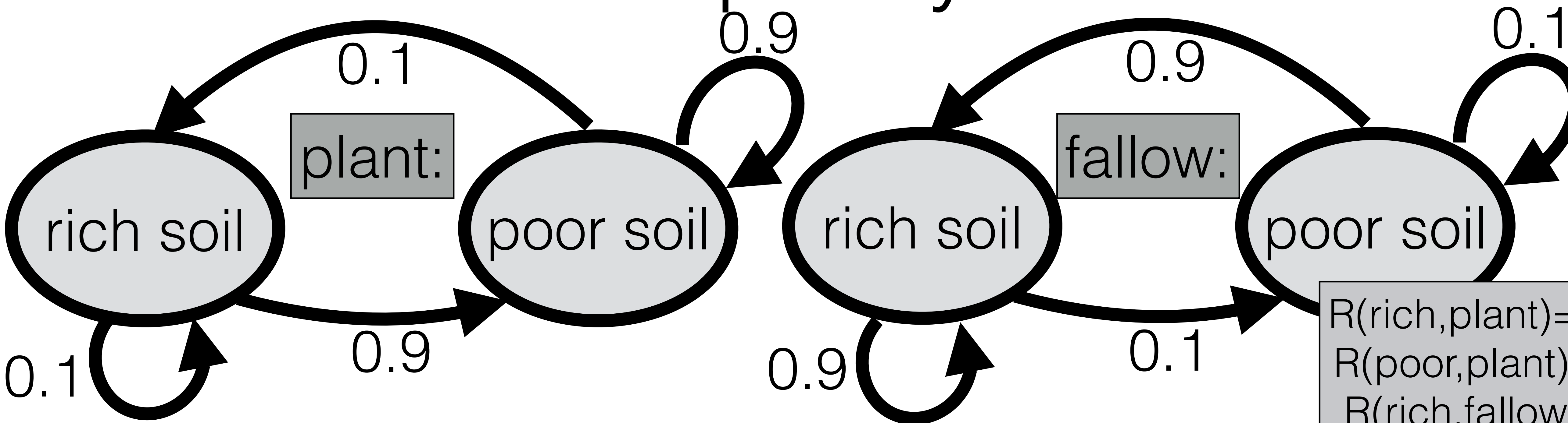
$$Q^0(s,a) = 0; Q^1(s,a) = R(s,a)$$

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

What's best?

$\pi_1^*$

# What's the best policy?




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

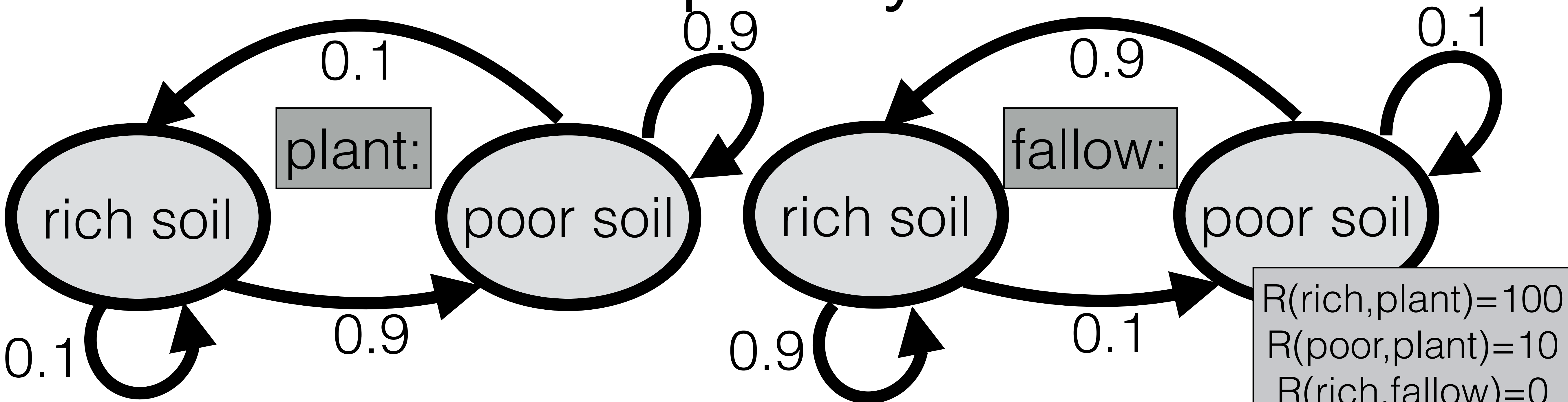
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s, a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s, a)$

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

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

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

# What's the best policy?




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

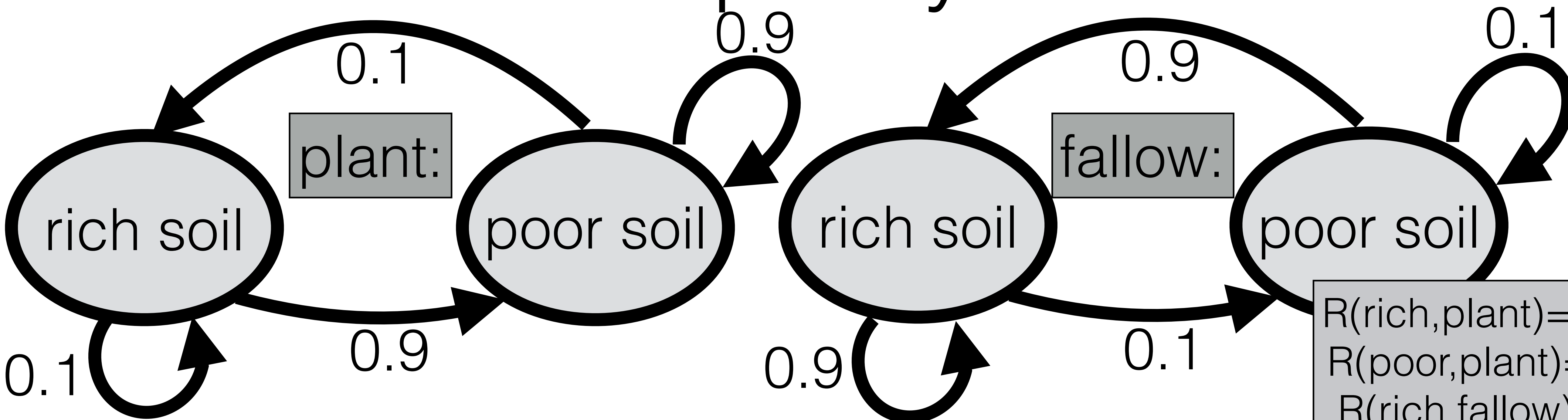
$Q^0(s, a) = 0$; $Q^1(s, a) = R(s, a)$

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

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

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

# What's the best policy?




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**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

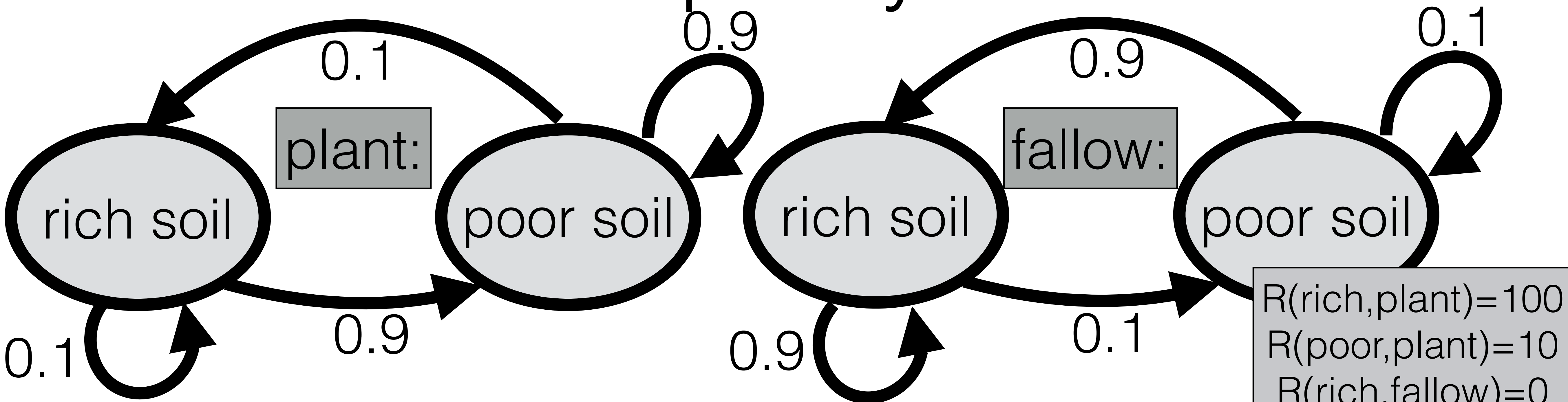
$Q^0(s,a) = 0;$ $Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$

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

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

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

# What's the best policy?




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

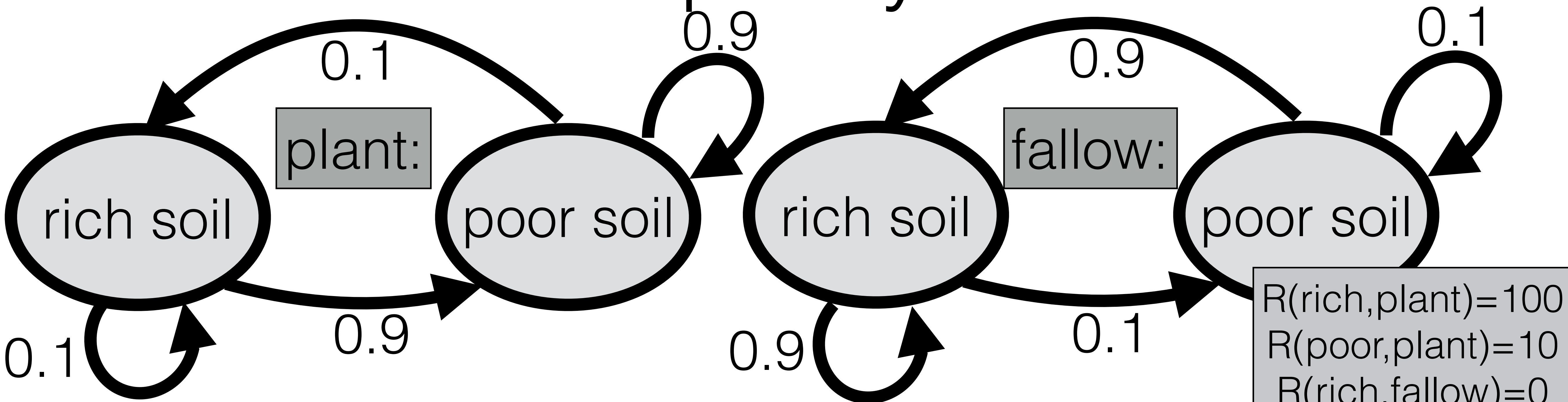
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

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

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

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

# What's the best policy?



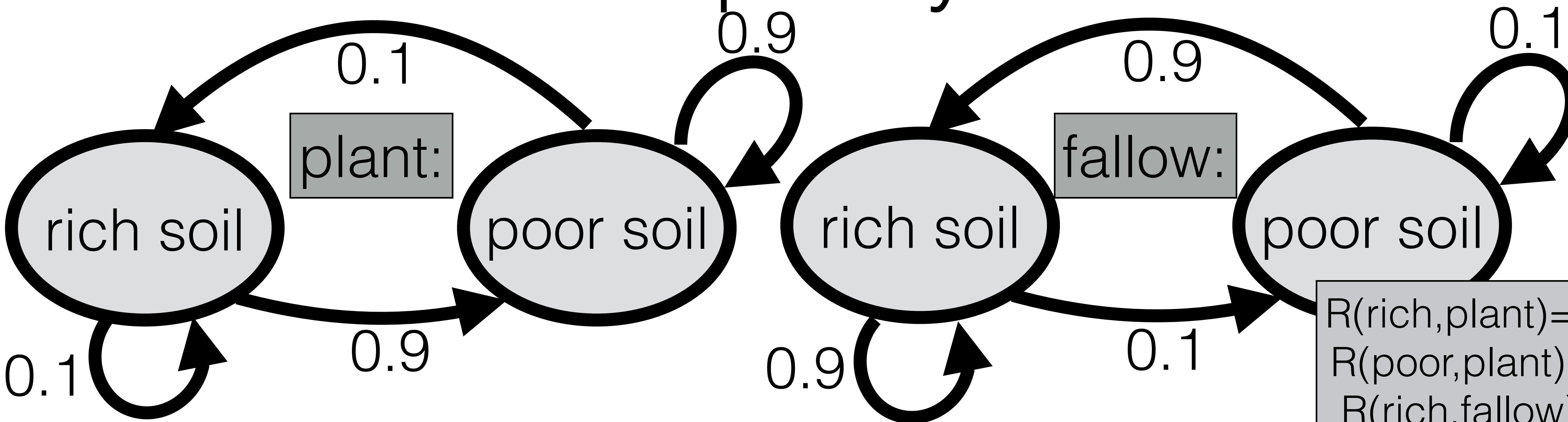

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi^*_h(s) = \arg\max_a Q^h(s,a)$

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

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

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

# What's the best policy?




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

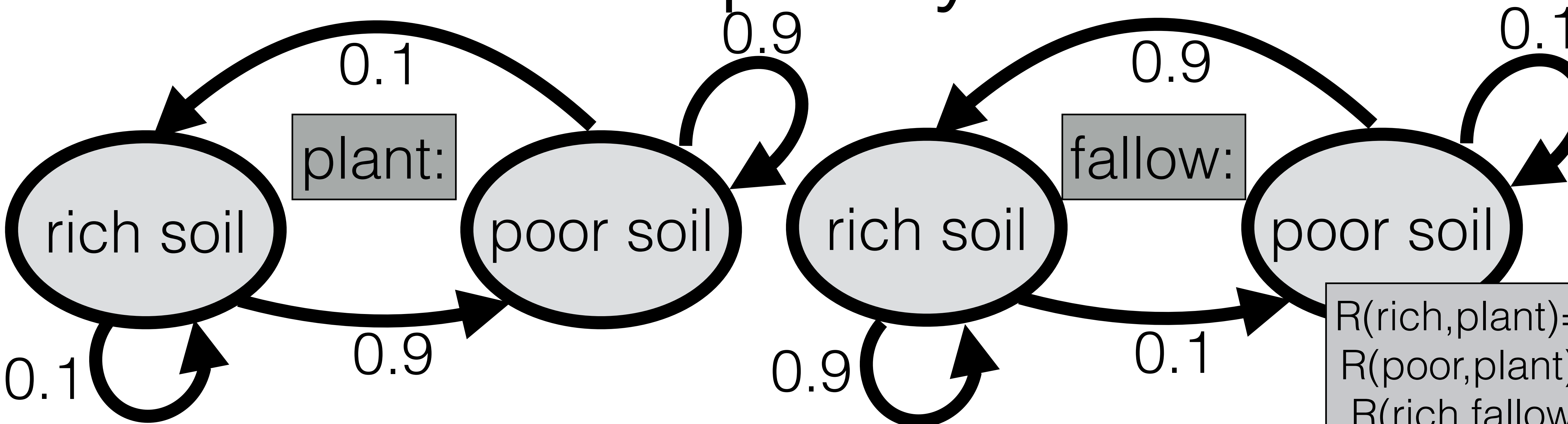
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

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

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

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

# What's the best policy?




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

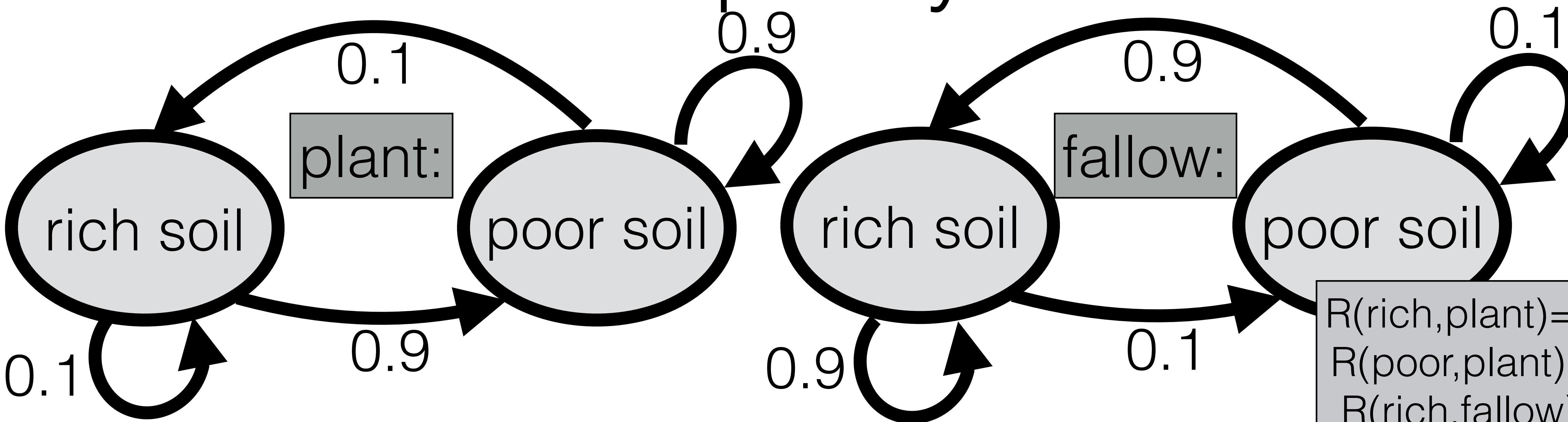
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

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

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

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

# What's the best policy?




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

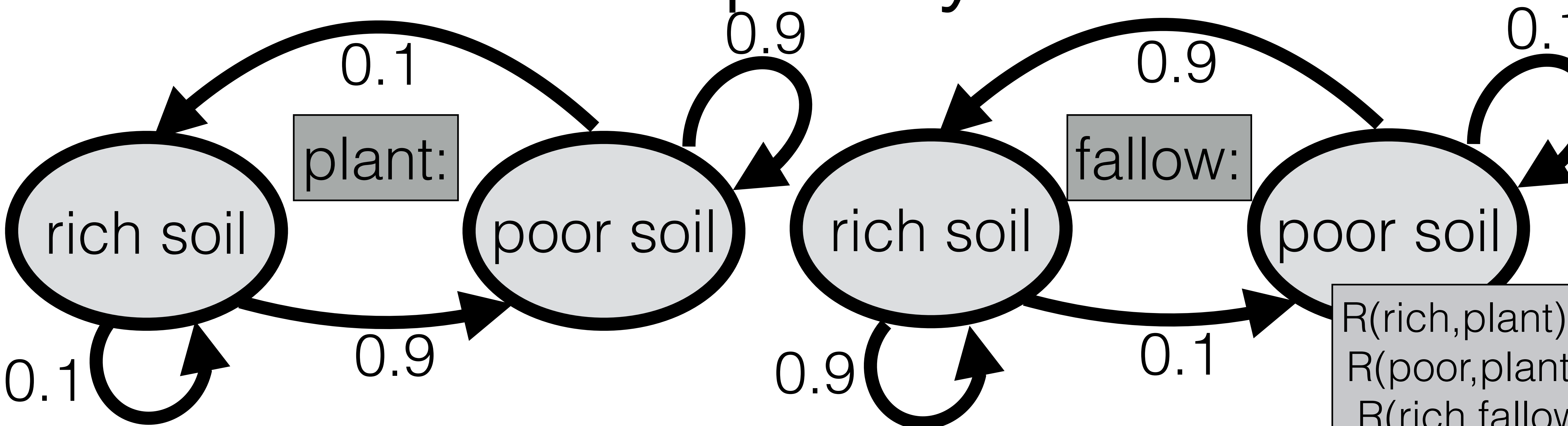
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi^*_h(s) = \arg\max_a Q^h(s,a)$

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

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

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

# What's the best policy?




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

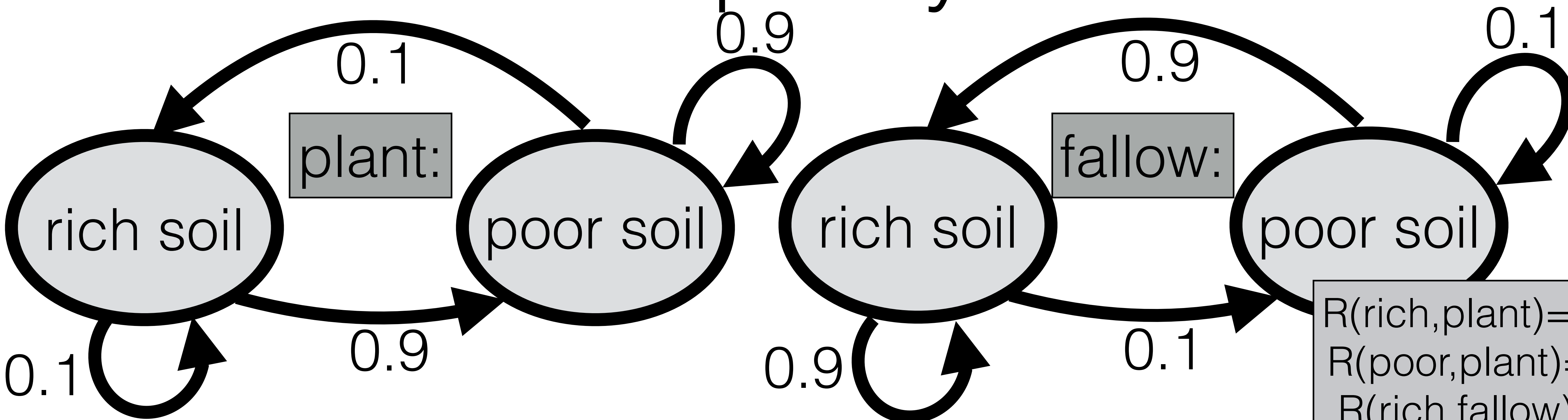
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

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

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

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

# What’s the best policy?




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

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

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

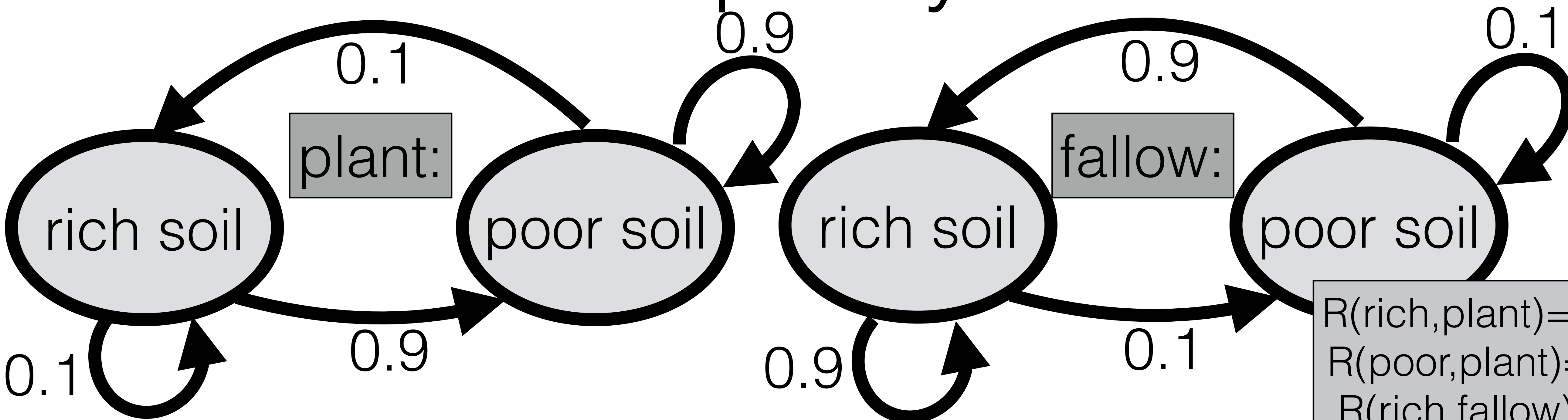
$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$

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

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

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

# What's the best policy?




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

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

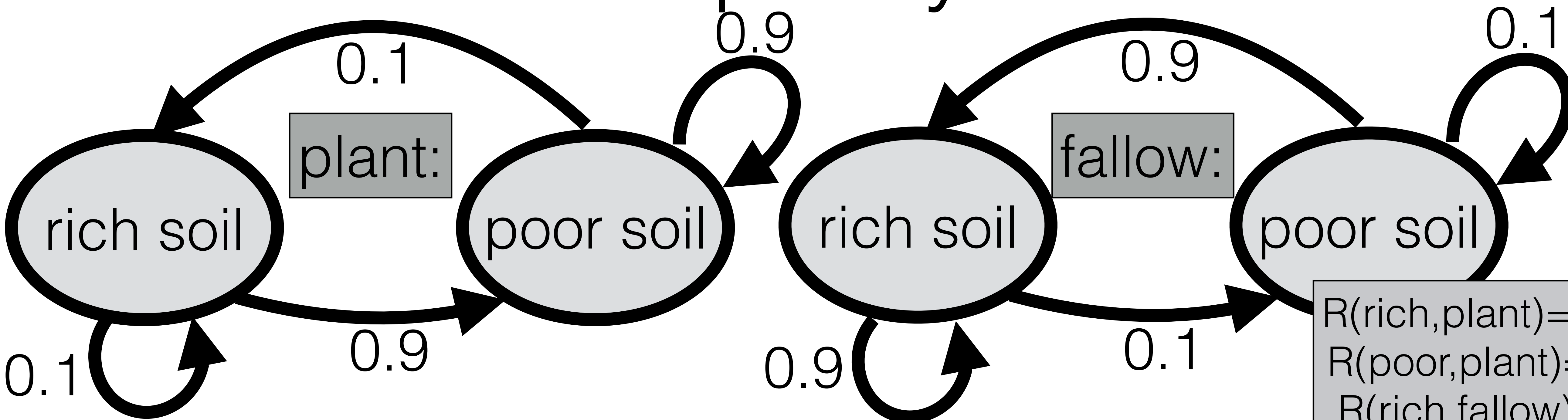
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

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

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

# What's the best policy?




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**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

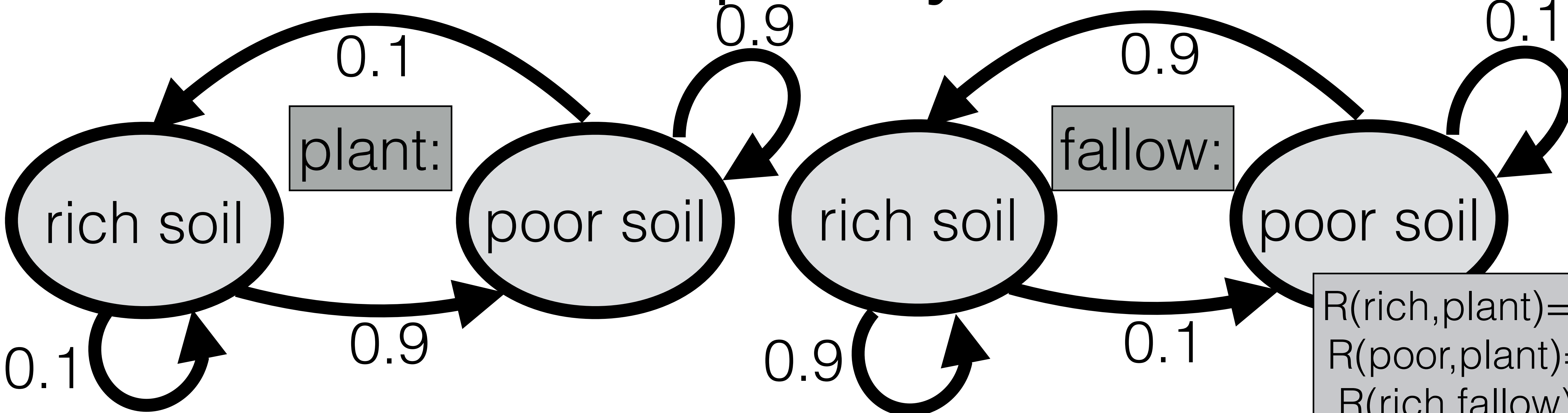
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

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

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

# What's the best policy?




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**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

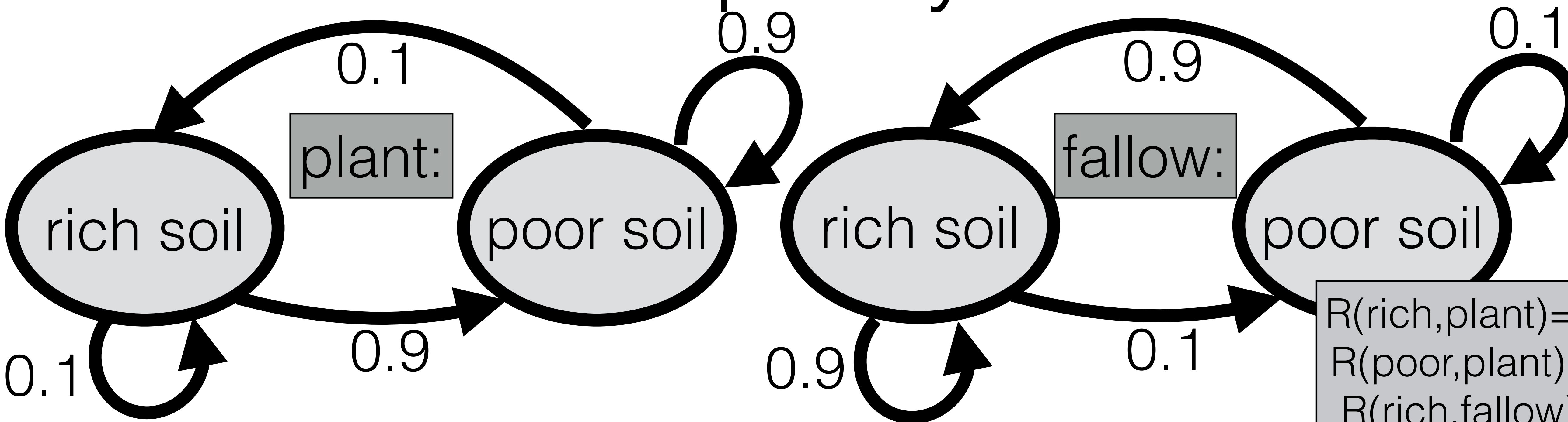
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

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

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

# What's the best policy?




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**: $\pi^*_h(s) = \arg\max_a Q^h(s,a)$

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

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

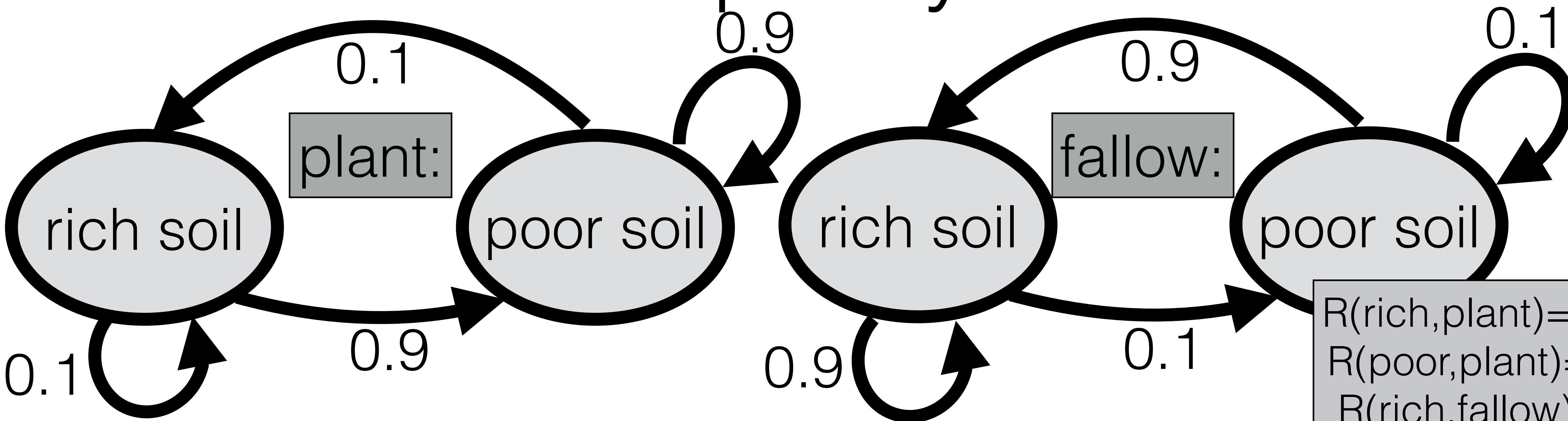
$$Q^1(\text{poor},\text{plant}) = 10; Q^1(\text{poor},\text{fallow}) = 0$$

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

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

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

# What's the best policy?




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**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

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

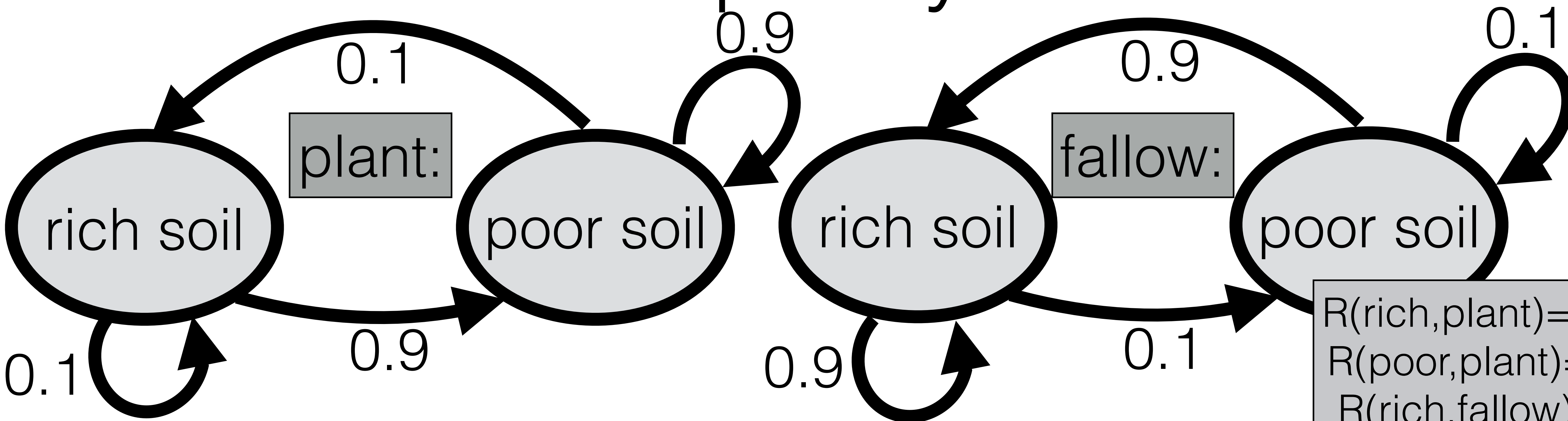
$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$

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

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

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

# What's the best policy?




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**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

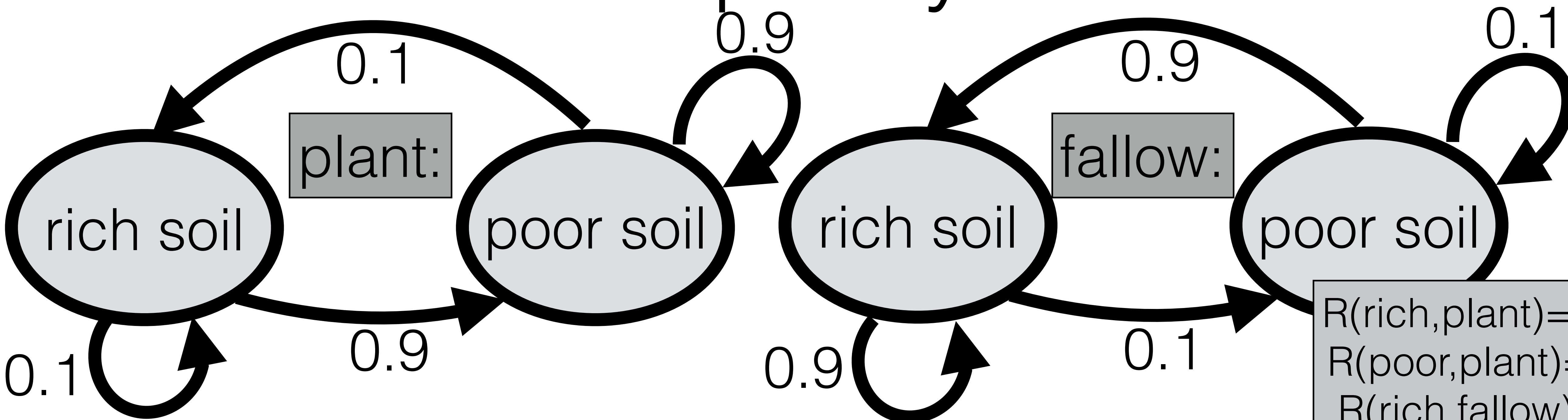
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

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

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

# What's the best policy?




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**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

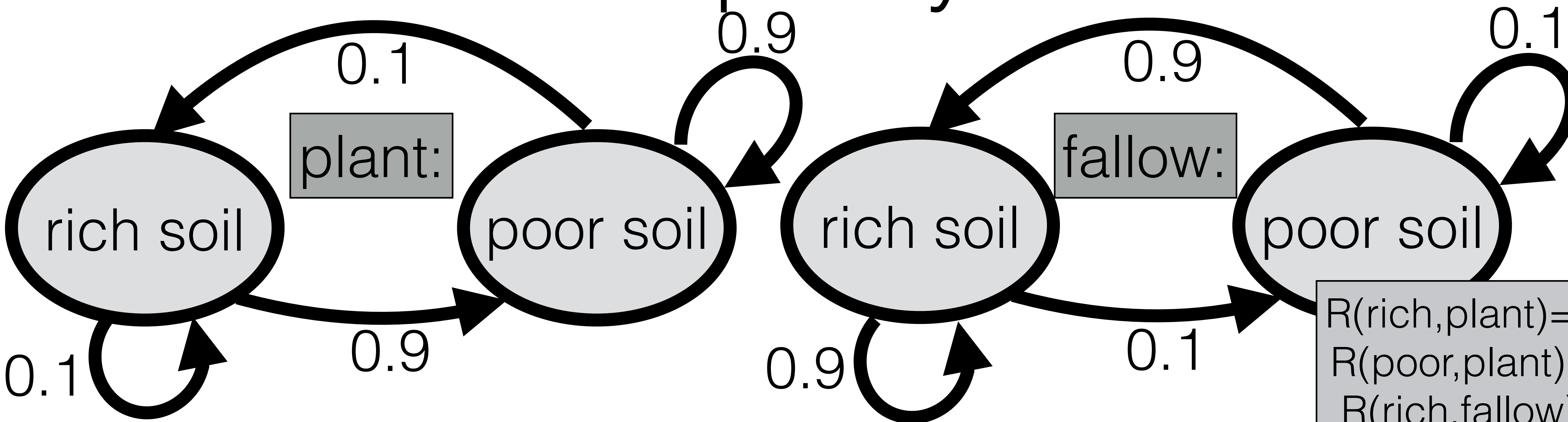
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

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

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

# What's the best policy?




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

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

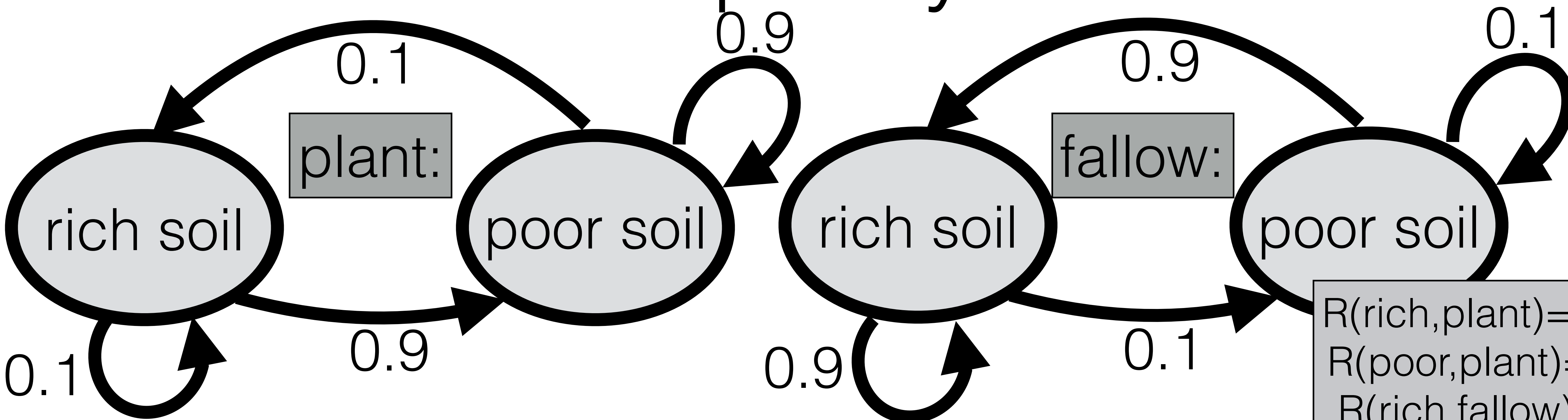
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

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

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

# What's the best policy?




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

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

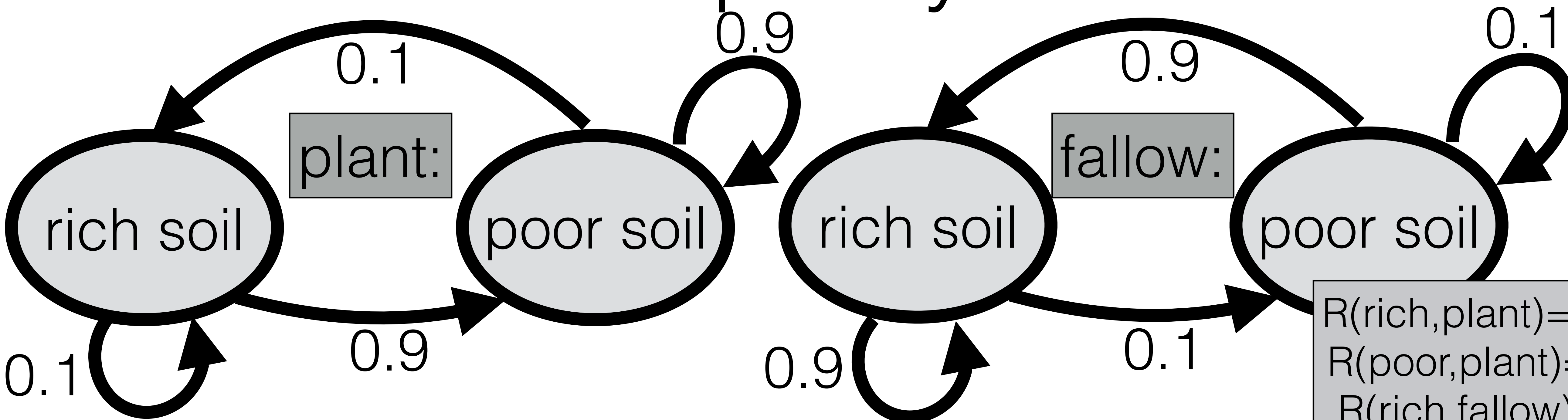
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

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

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

# What's the best policy?




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

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

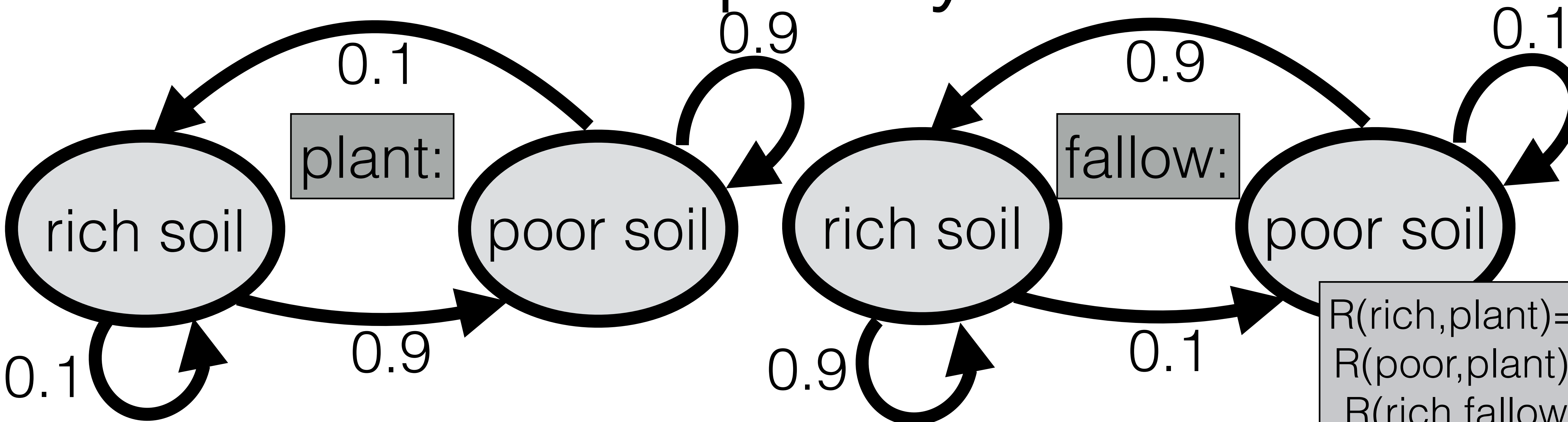
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

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

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

# What's the best policy?




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

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

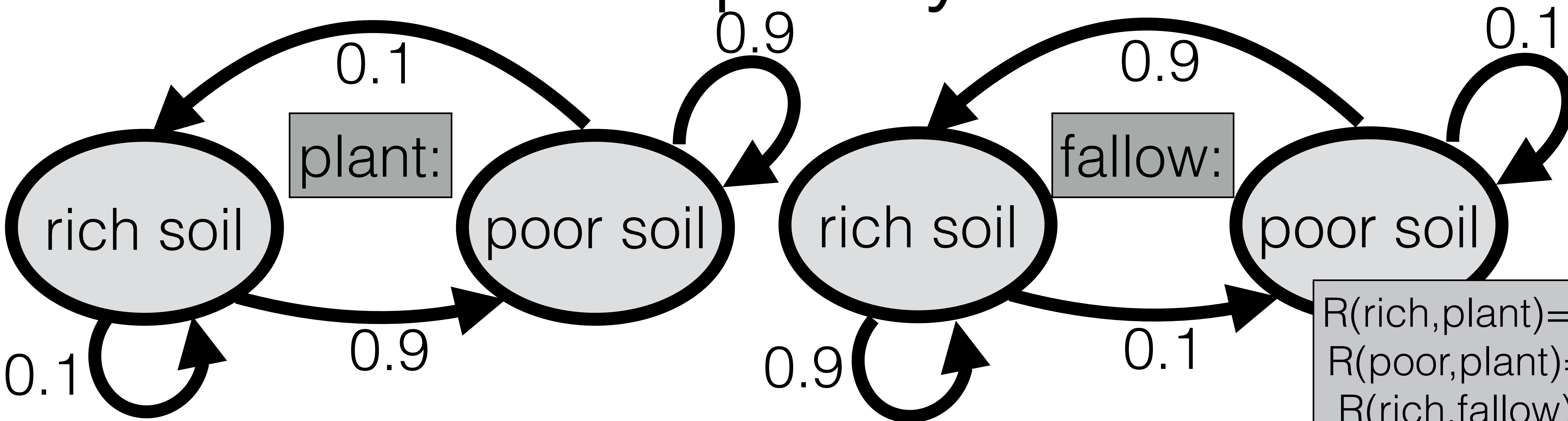
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

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

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

# What's the best policy?




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

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

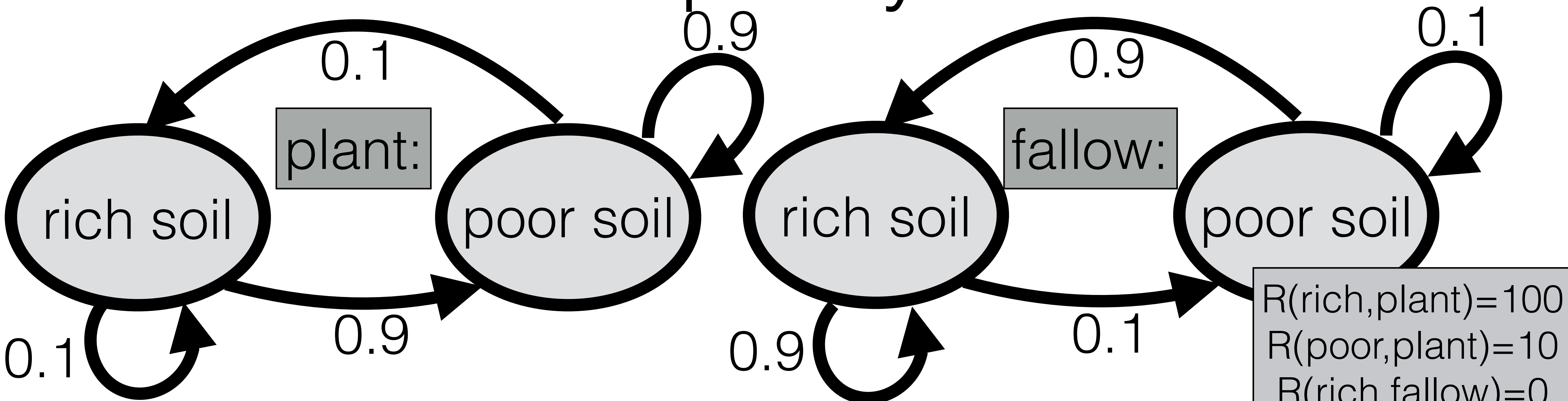
$$Q^0(s, a) = 0; Q^h(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \max_{a'} Q^{h-1}(s', a')$$

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

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

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

# What's the best policy?




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

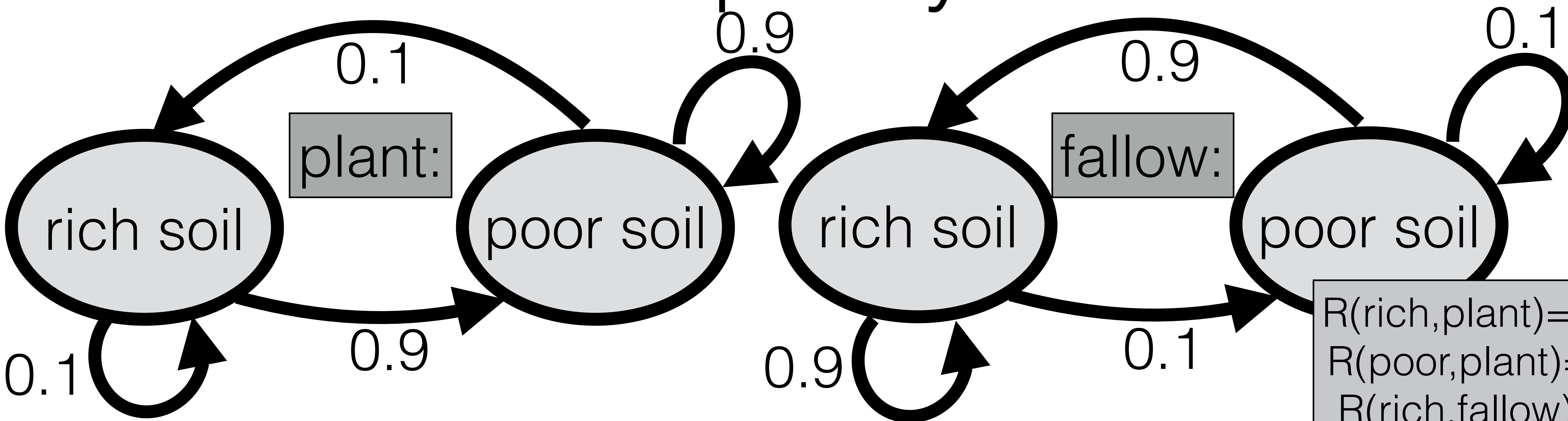
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

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

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

# What's the best policy?




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

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

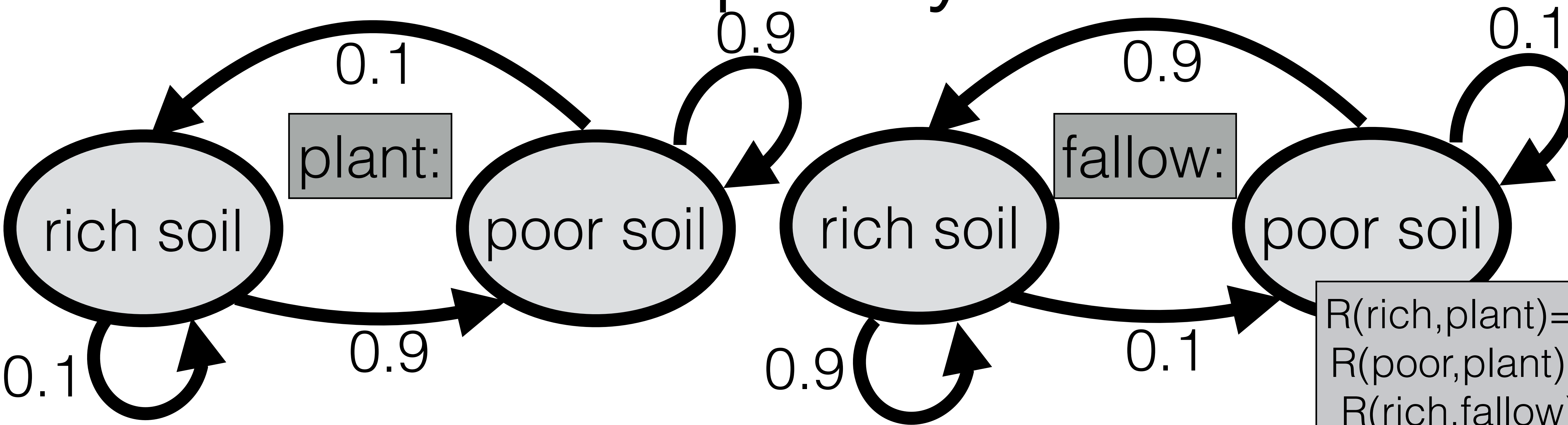
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

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

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

# What's the best policy?




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

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

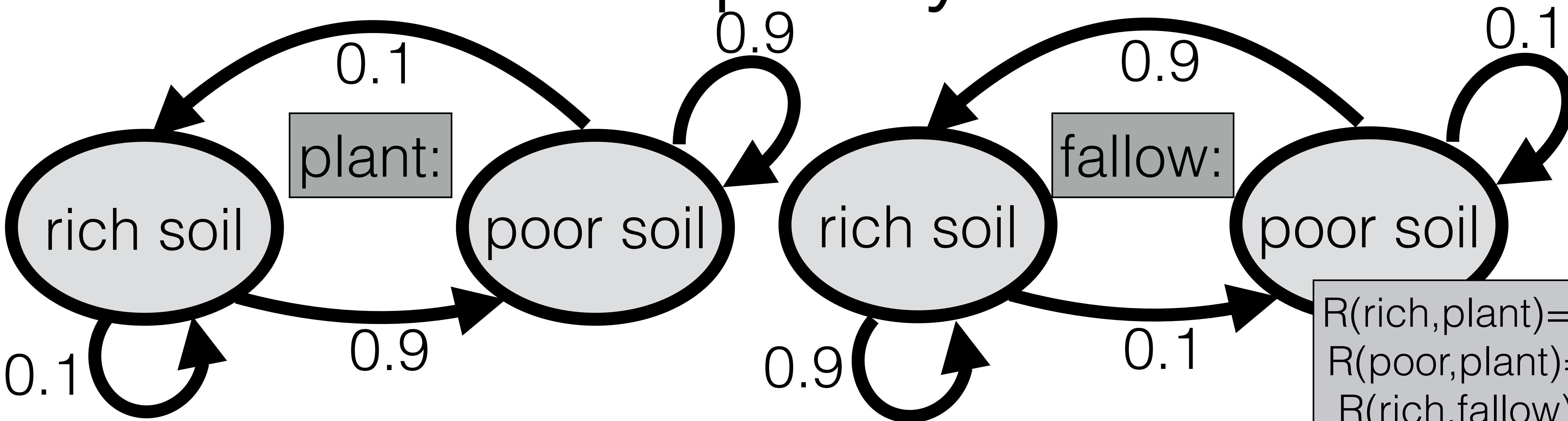
$$Q^0(s, a) = 0; Q^h(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \max_{a'} Q^{h-1}(s', a')$$

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

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

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

# What's the best policy?




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

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

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

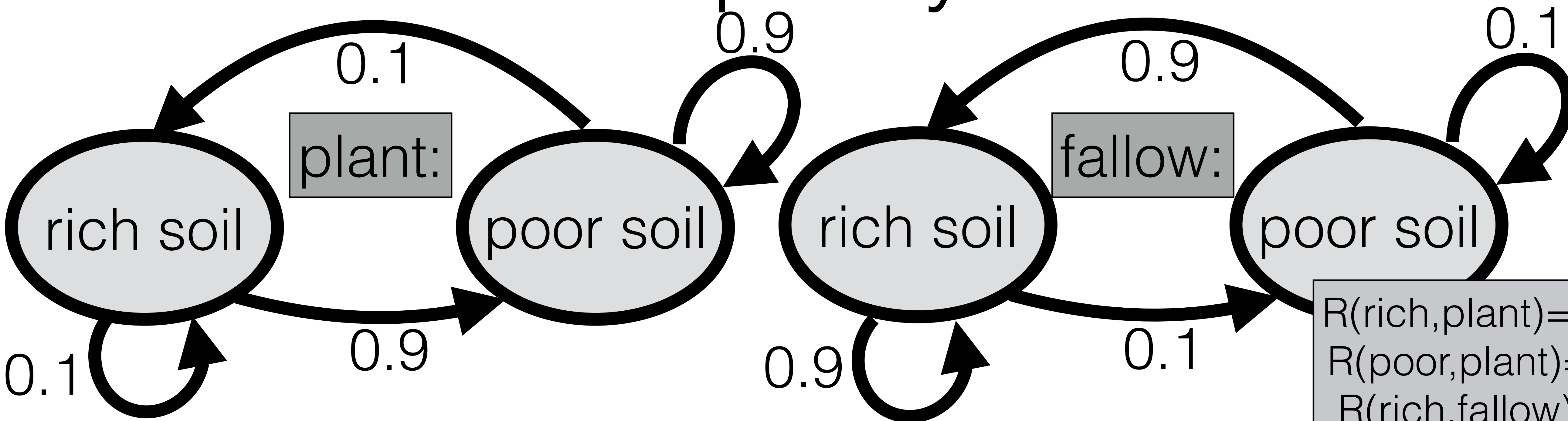
$$Q^1(\text{rich}, \text{plant}) = 100; Q^1(\text{rich}, \text{fallow}) = 0;$$

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

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

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

# What's the best policy?




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

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

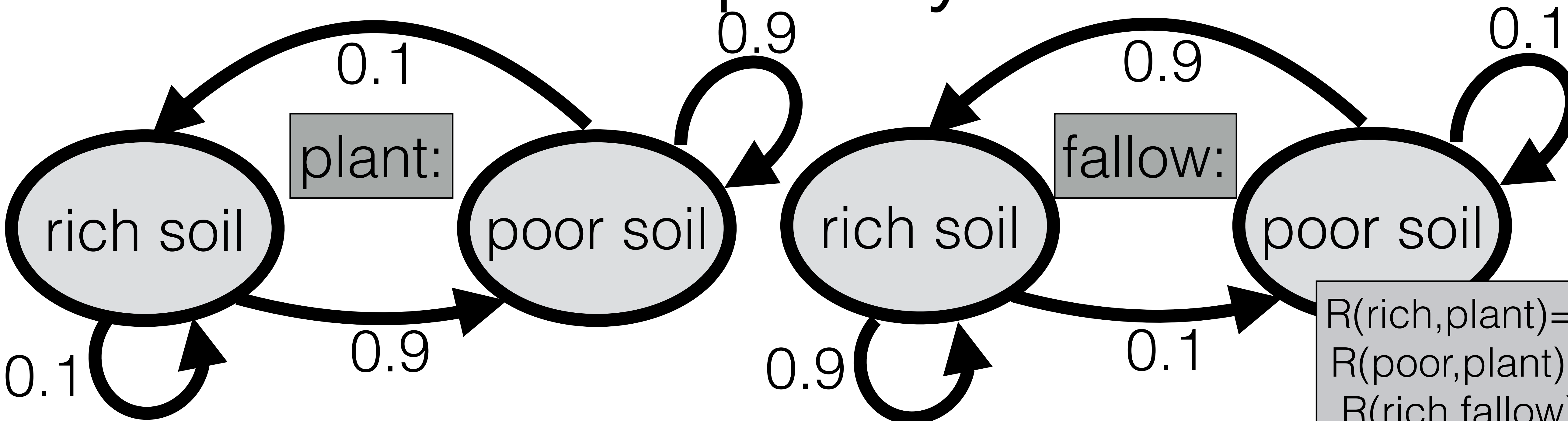
$$Q^0(s, a) = 0; Q^h(s, a) = R(s, a) + \sum_{s'} T(s, a, s') \max_{a'} Q^{h-1}(s', a')$$

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

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

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

# What’s the best policy?




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

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

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

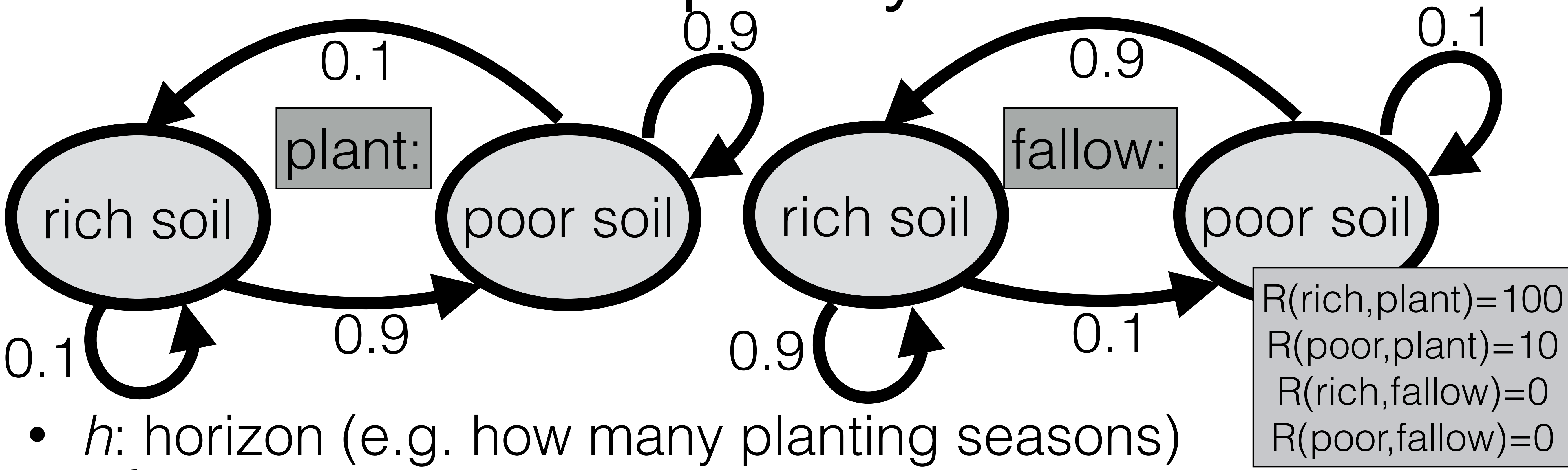
$$Q^1(\text{rich},\text{plant}) = 100; Q^1(\text{rich},\text{fallow}) = 0;$$

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

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

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

# What's the best policy?



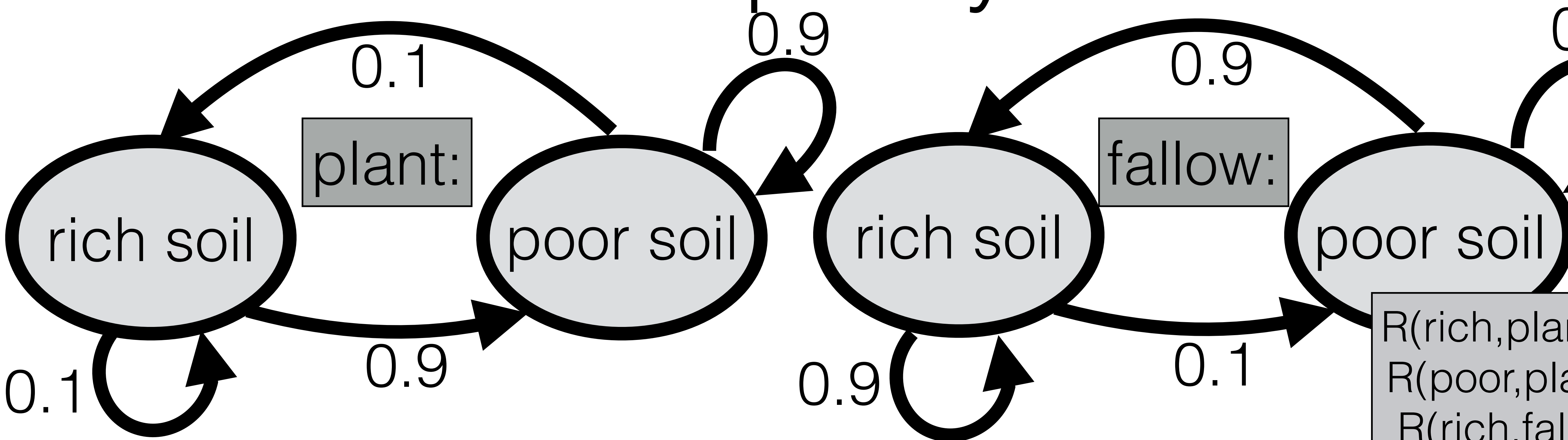

- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi^*_h(s) = \arg\max_a Q^h(s,a)$

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

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

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

# What's the best policy?




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

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

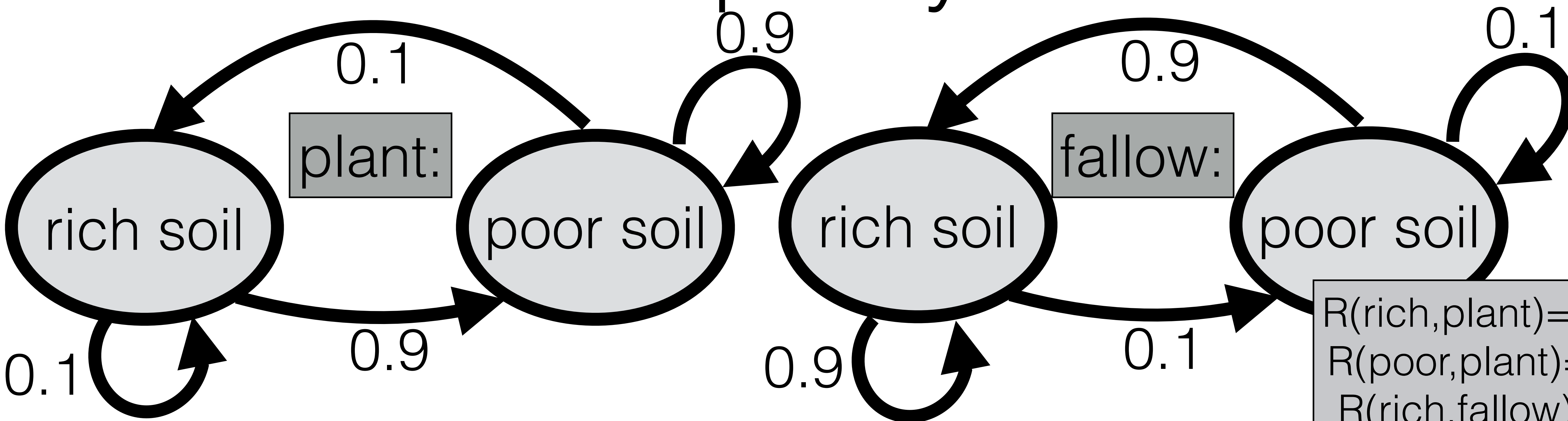
$$Q^0(s,a) = 0; Q^h(s,a) = R(s,a) + \sum_{s'} T(s,a,s') \max_{a'} Q^{h-1}(s',a')$$

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

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

$\pi_2^*$

# What's the best policy?




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

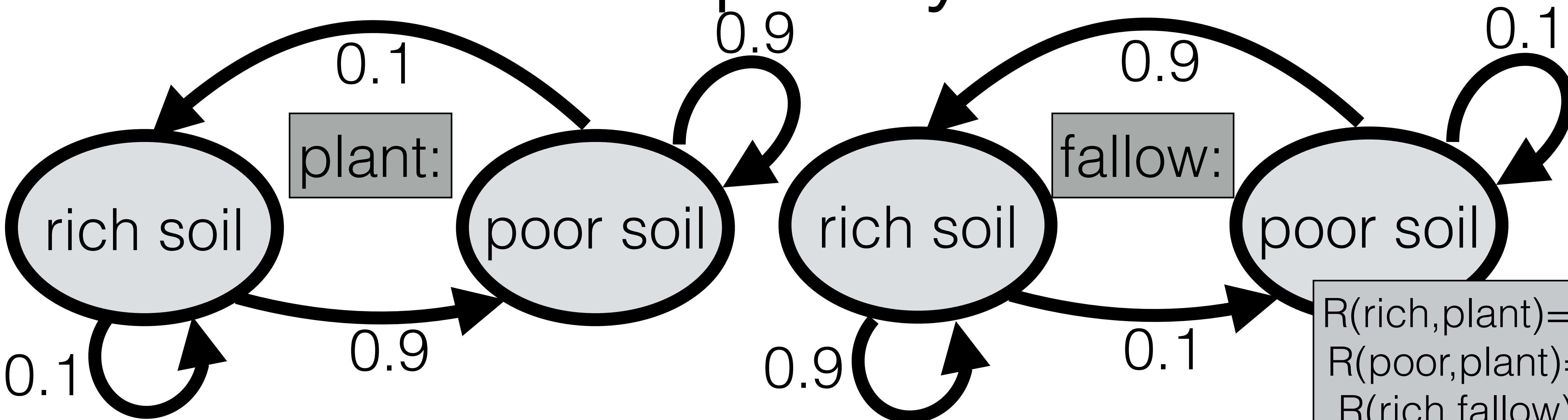
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

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

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

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

# What's the best policy?




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

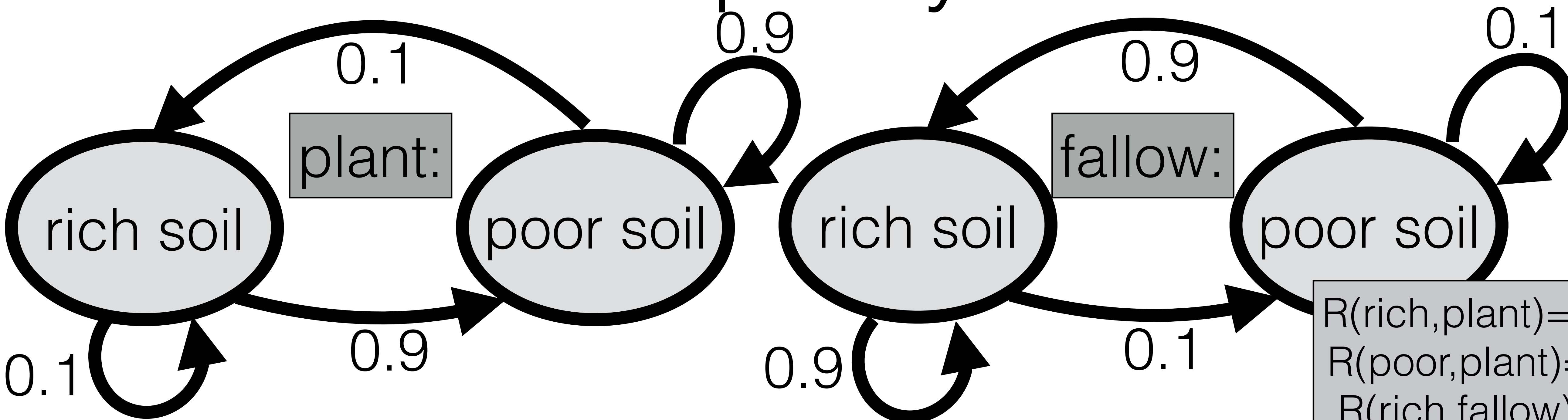
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

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

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

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

# What's the best policy?




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

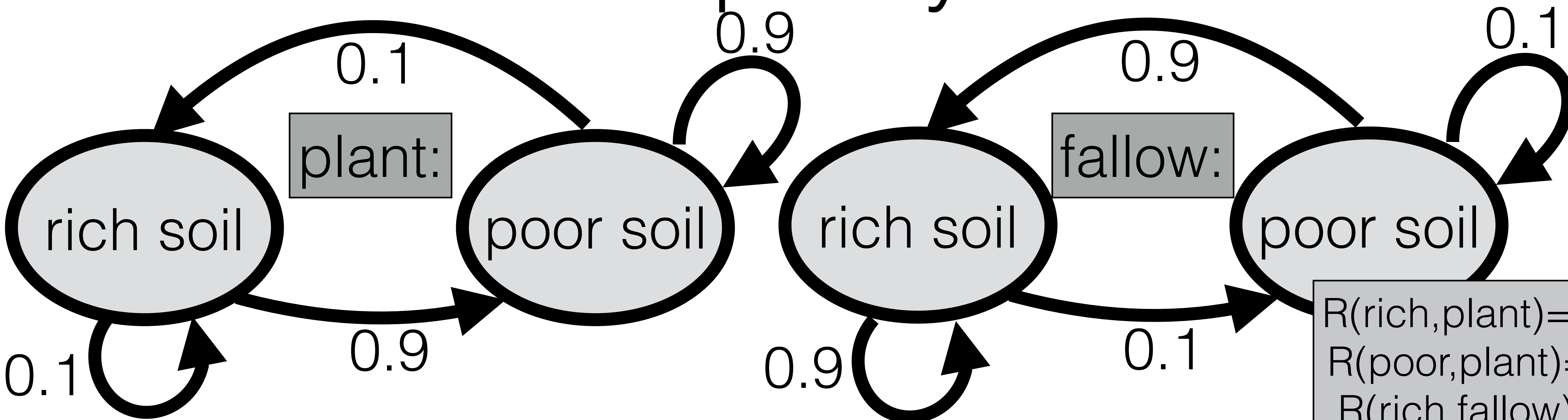
- *h*: horizon (e.g. how many planting seasons)
- $Q^h(s,a)$: expected reward of starting at *s*, making action *a*, and then making the "best" action for the *h*-1 steps left
- With *Q*, can find **an optimal policy**: $\pi_h^*(s) = \arg\max_a Q^h(s,a)$

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

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

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

# What's the best policy?




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

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

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

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

"finite-horizon value iteration"

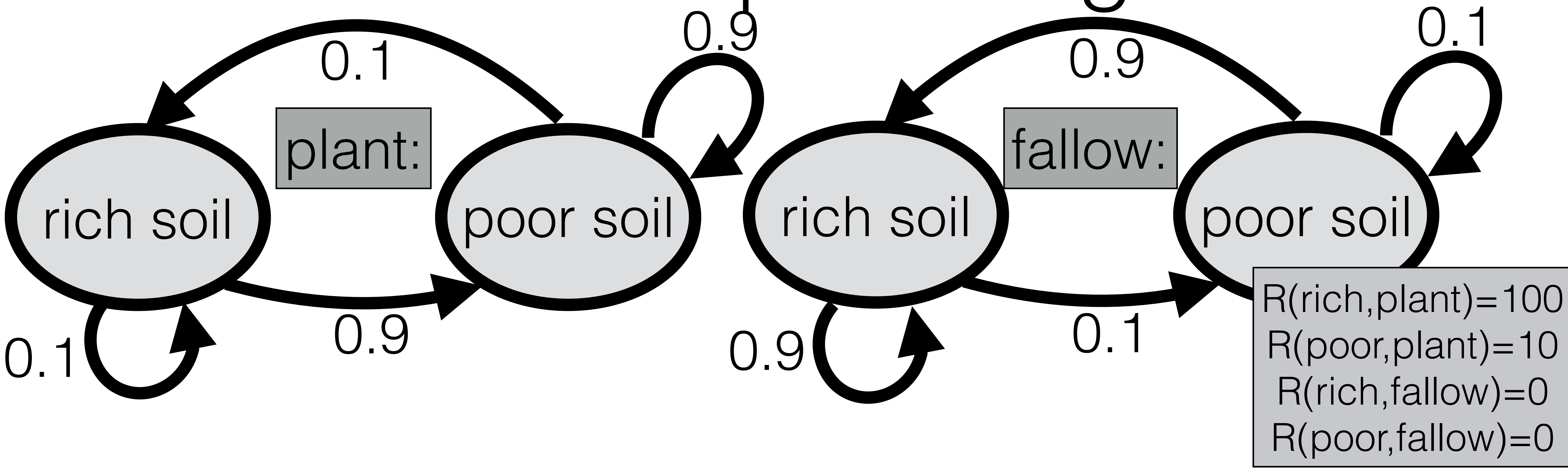
What's best? Any *s*, $\pi_1^*(s) = \text{plant}; \pi_2^*(\text{rich}) = \text{plant}, \pi_2^*(\text{poor}) = \text{fallow}$

# What if I don’t stop farming?

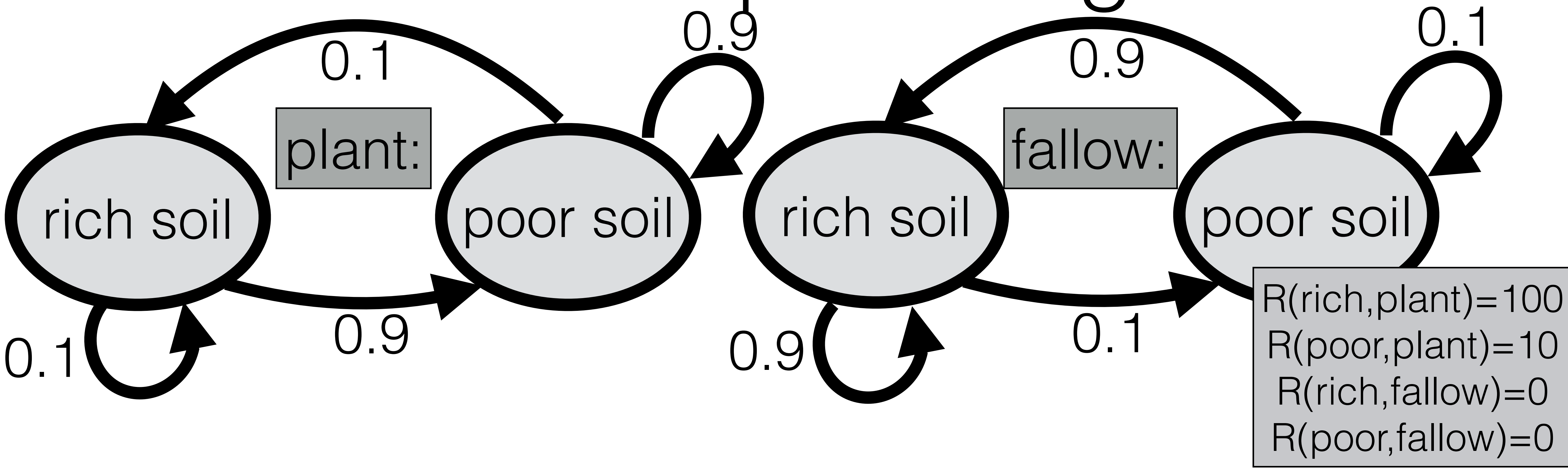
# What if I don’t stop farming?

Good news! No strip mall, and I get to keep the farm forever

# What if I don't stop farming?

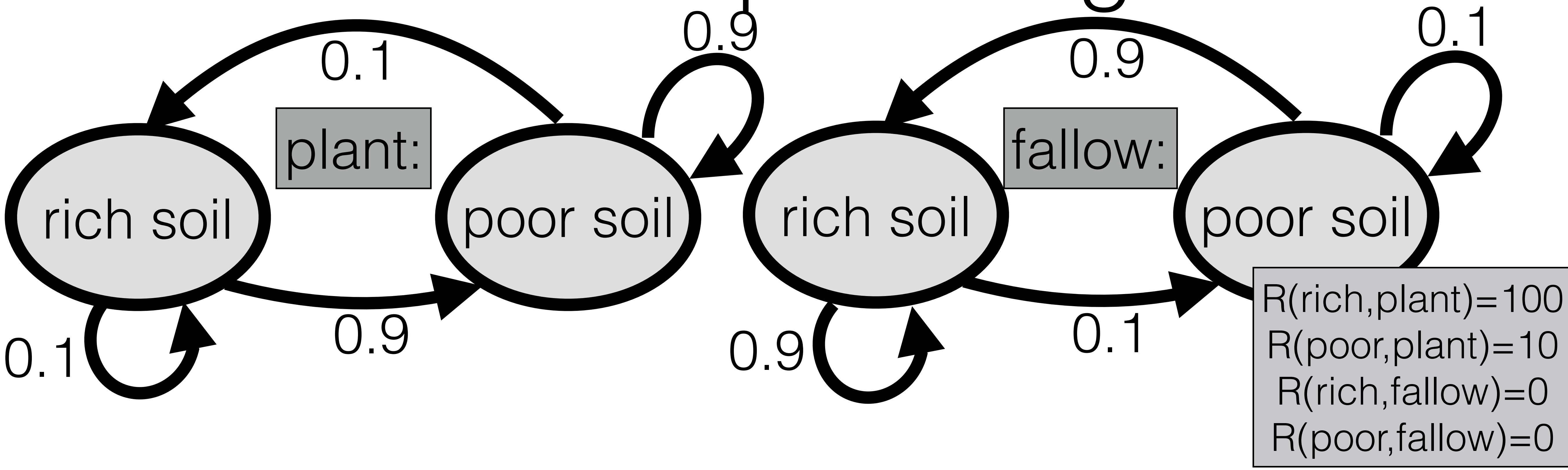


# What if I don’t stop farming?



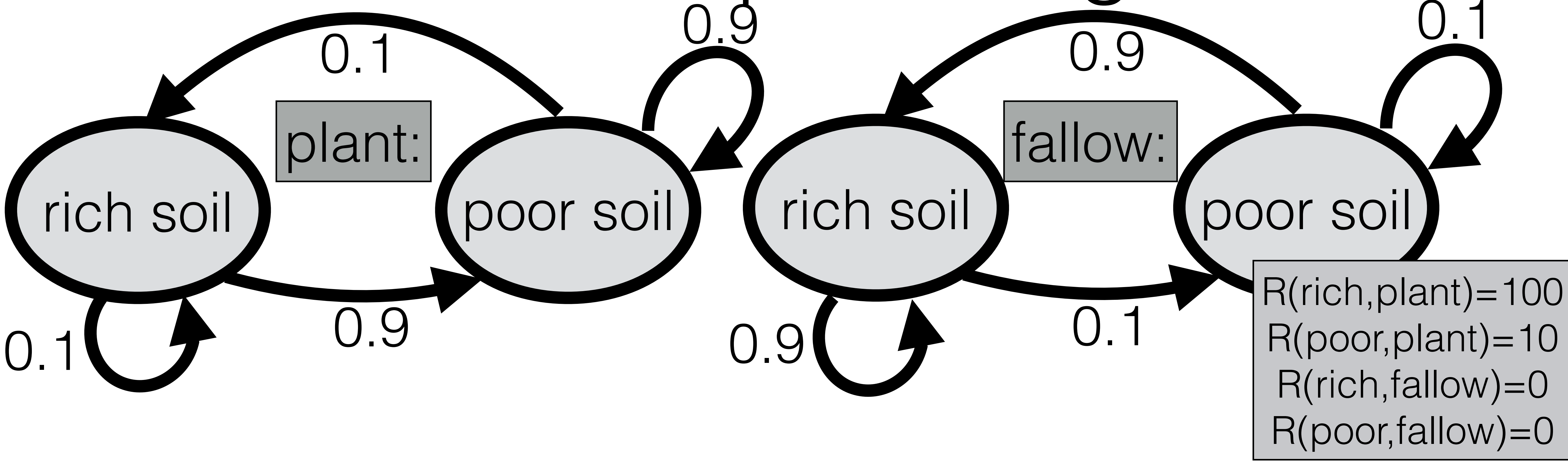

- Problem: 1,000 bushels today > 1,000 bushels in ten years

# What if I don't stop farming?



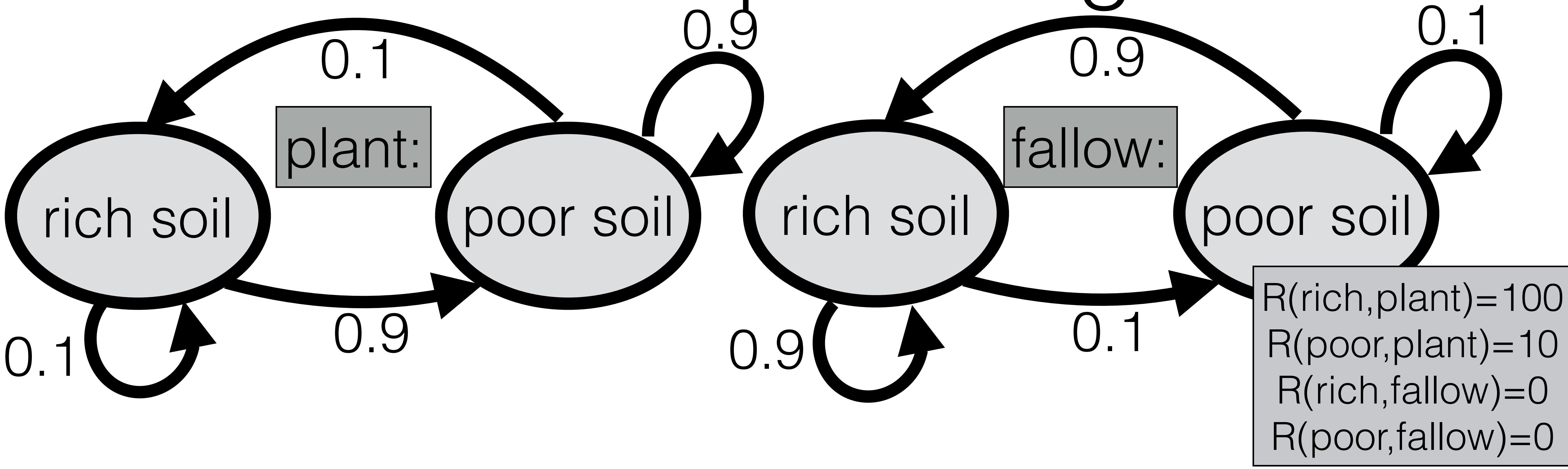

- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$

# What if I don't stop farming?



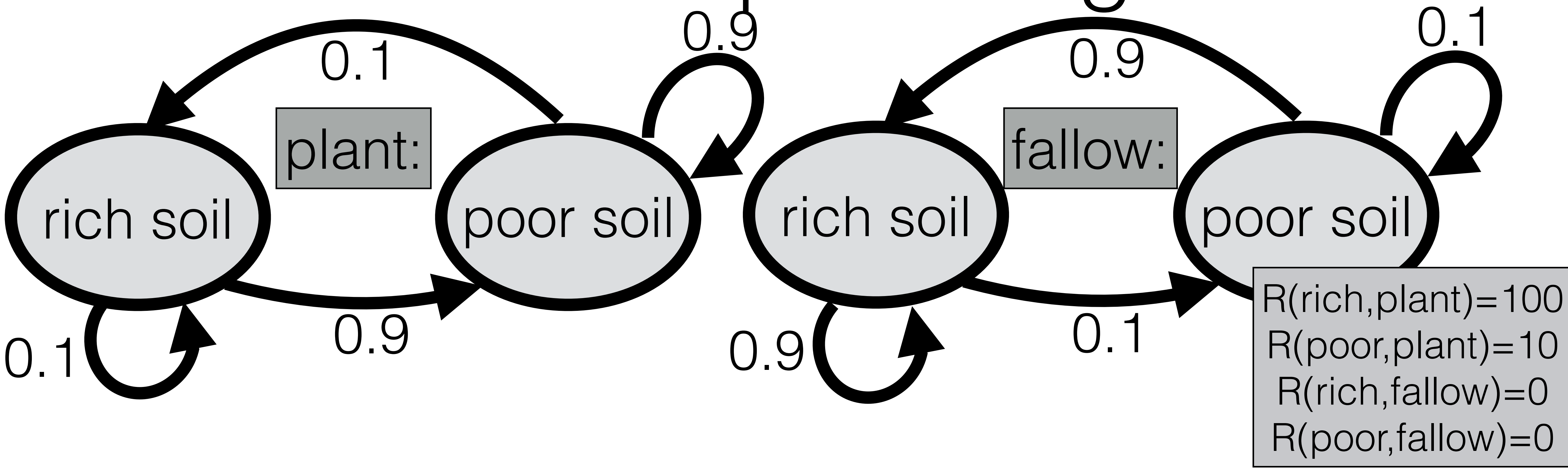

- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: $\gamma^t$ bushels

# What if I don’t stop farming?



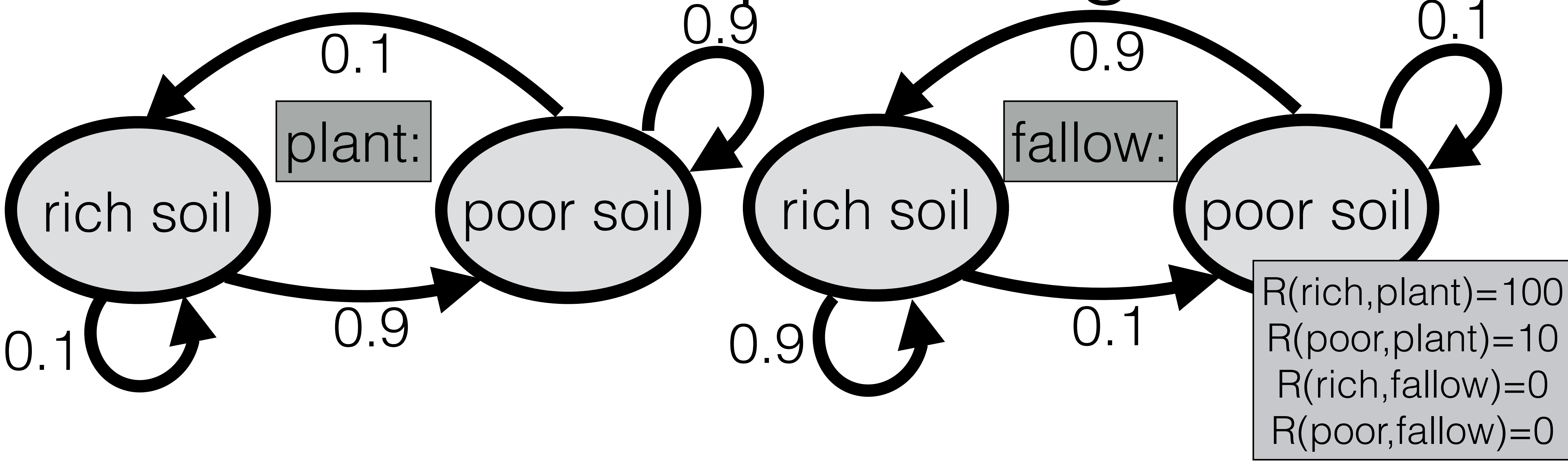

- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: $\gamma^t$ bushels
  - Example: What’s the value of 1 bushel per year forever?

# What if I don’t stop farming?



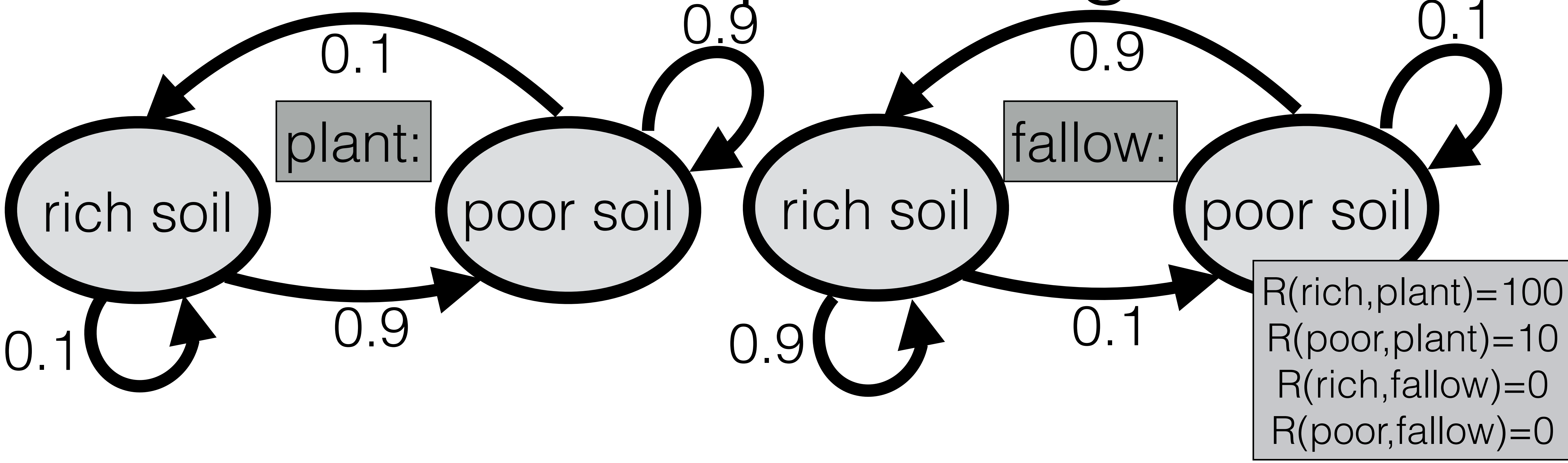

- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after $t$ time steps: $\gamma^t$ bushels
  - Example: What’s the value of 1 bushel per year forever?
    $V$

# What if I don't stop farming?



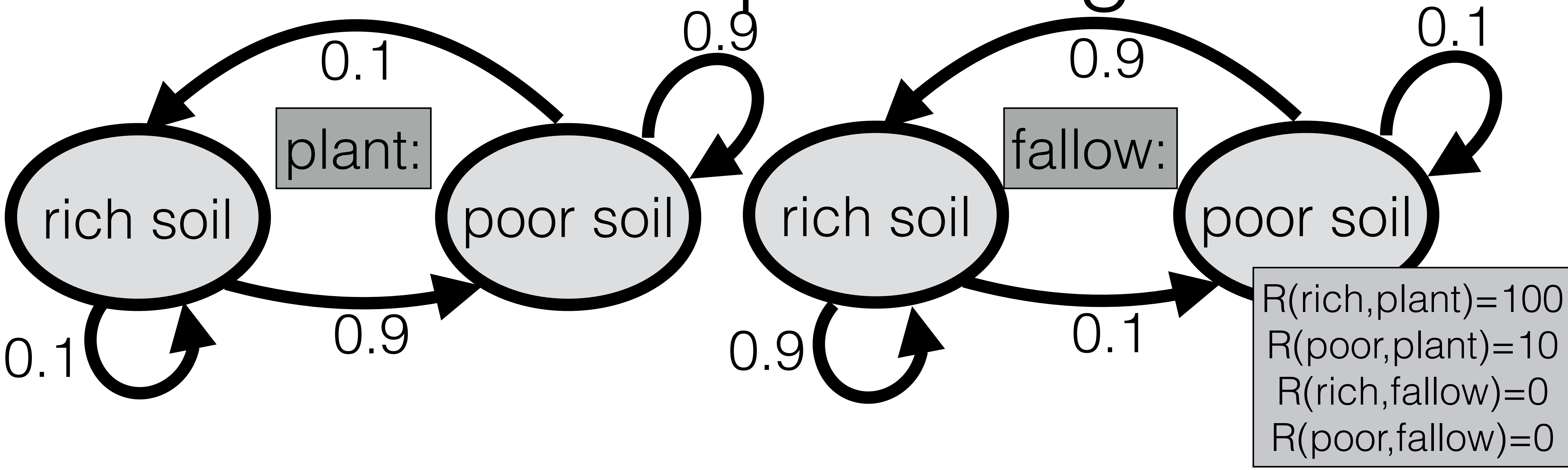

- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after $t$ time steps: $\gamma^t$ bushels
  - Example: What's the value of 1 bushel per year forever? $V = 1 + \gamma + \gamma^2 + \cdots$

# What if I don’t stop farming?



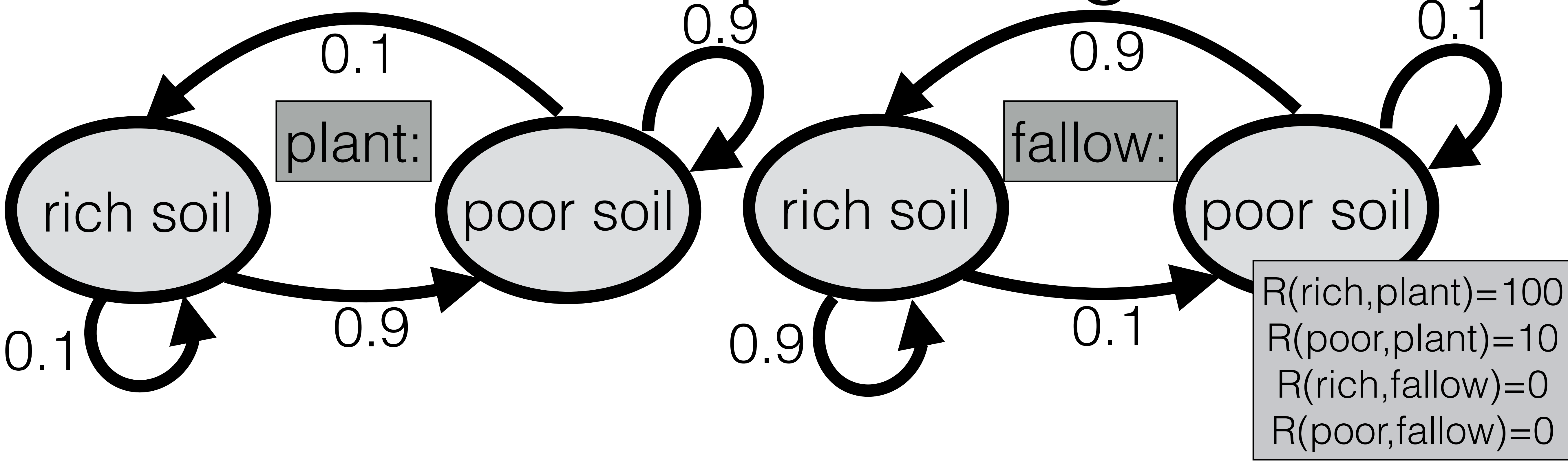

- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after $t$ time steps: $\gamma^t$ bushels
  - Example: What’s the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma + \gamma^2 + \cdots)$

# What if I don’t stop farming?



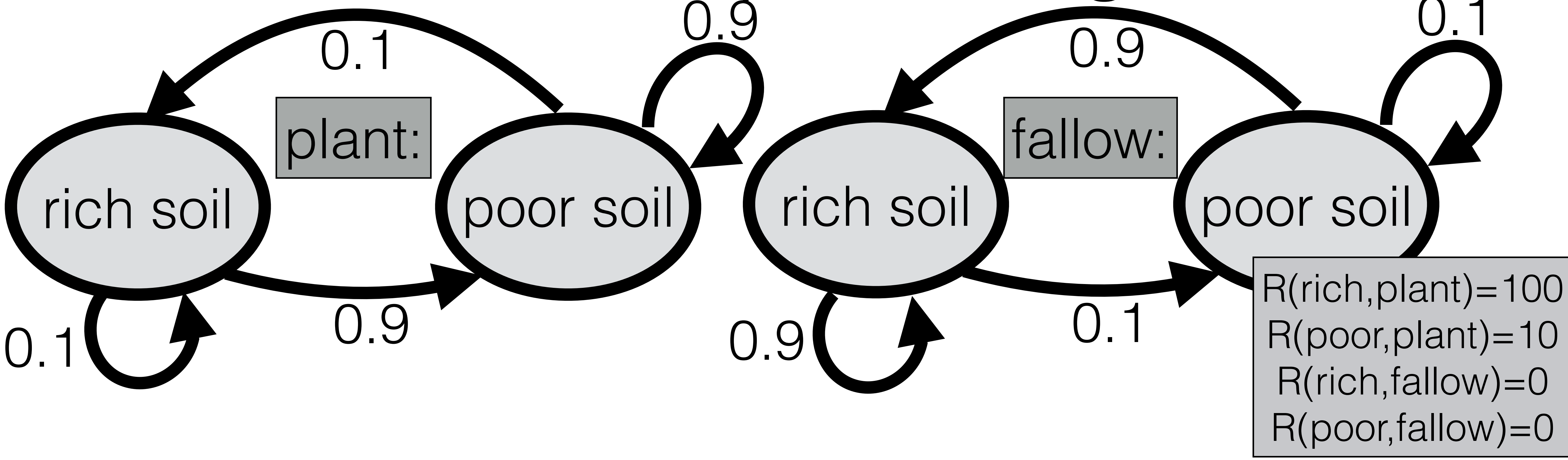

- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after $t$ time steps: $\gamma^t$ bushels
  - Example: What’s the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma + \gamma^2 + \cdots) = 1 + \gamma V$

# What if I don't stop farming?



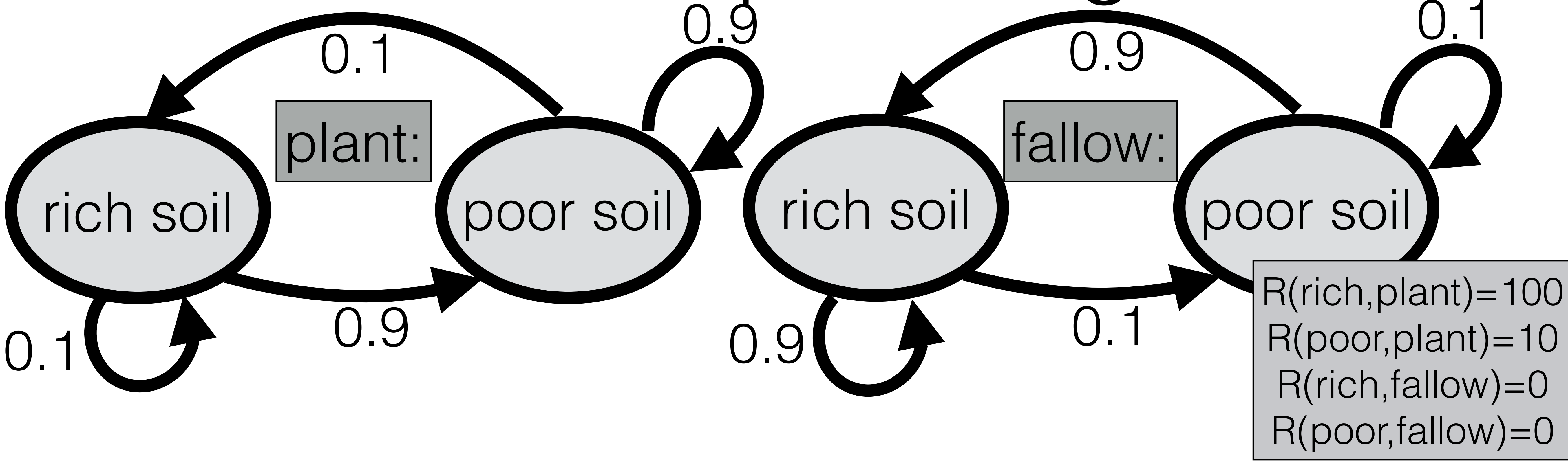

- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after $t$ time steps: $\gamma^t$ bushels
  - Example: What's the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma + \gamma^2 + \cdots) = 1 + \gamma V$
    $V = 1/(1-\gamma)$

# What if I don't stop farming?



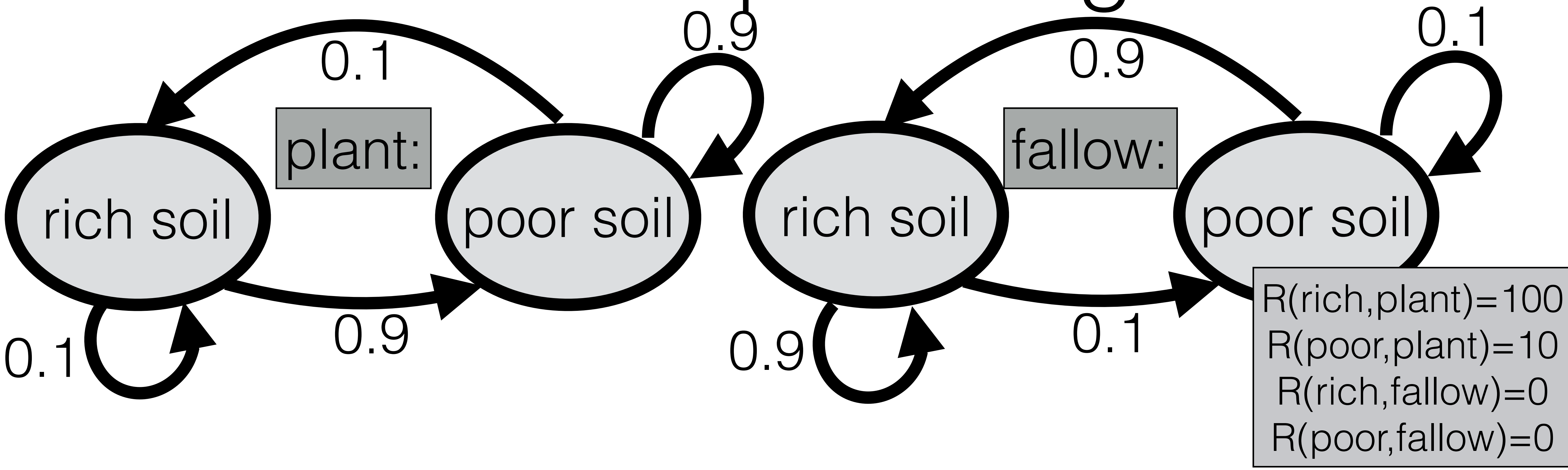

- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: $\gamma^t$ bushels
  - Example: What's the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma + \gamma^2 + \cdots) = 1 + \gamma V$
    $V = 1/(1 - \gamma)$ E.g. $\gamma = 0.99$

# What if I don’t stop farming?



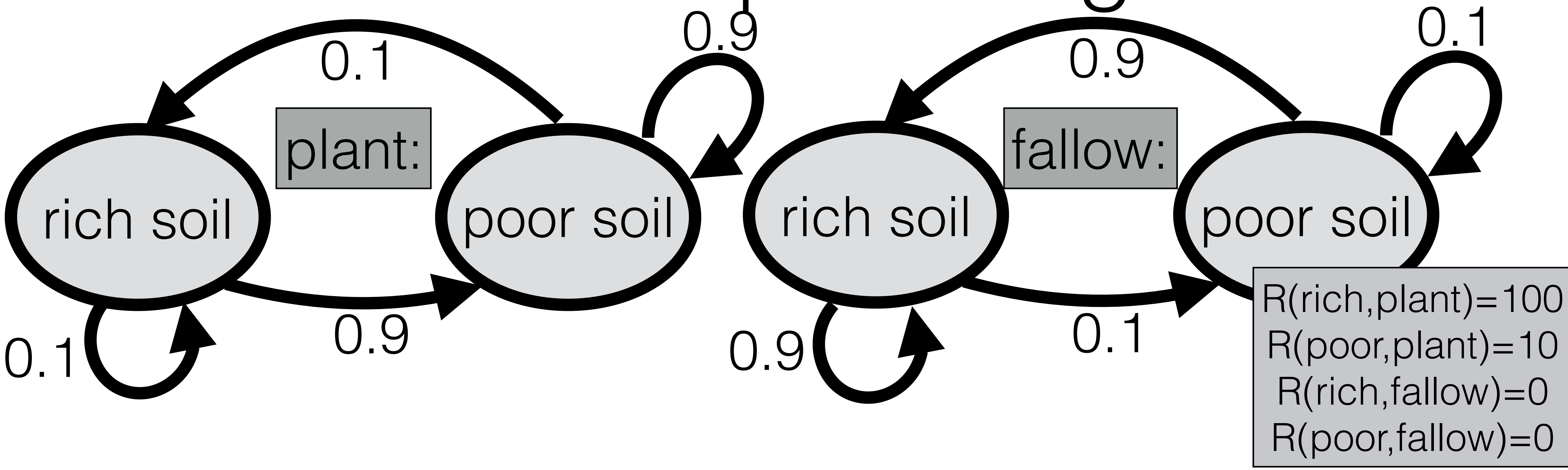

- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: $\gamma^t$ bushels
  - Example: What’s the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma + \gamma^2 + \cdots) = 1 + \gamma V$
    $V = 1/(1-\gamma)$ E.g. $\gamma = 0.99 \Rightarrow V = 1/0.01$

# What if I don’t stop farming?



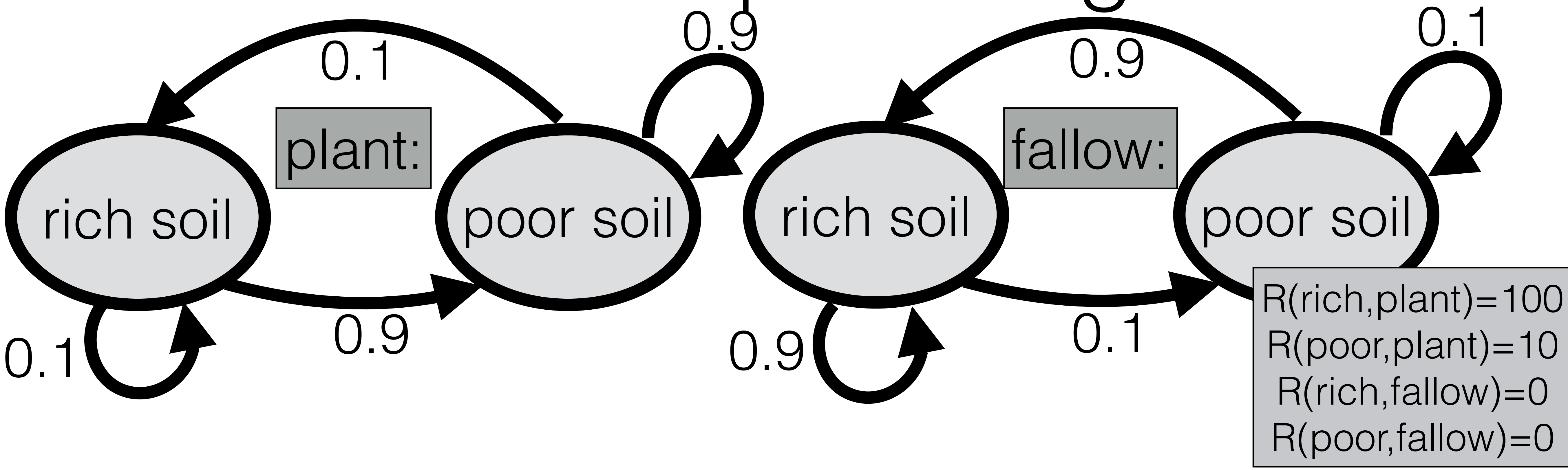

- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: $\gamma^t$ bushels
  - Example: What’s the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 +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

# 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: $\gamma^t$ 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 $\pi$ starting at state *s*

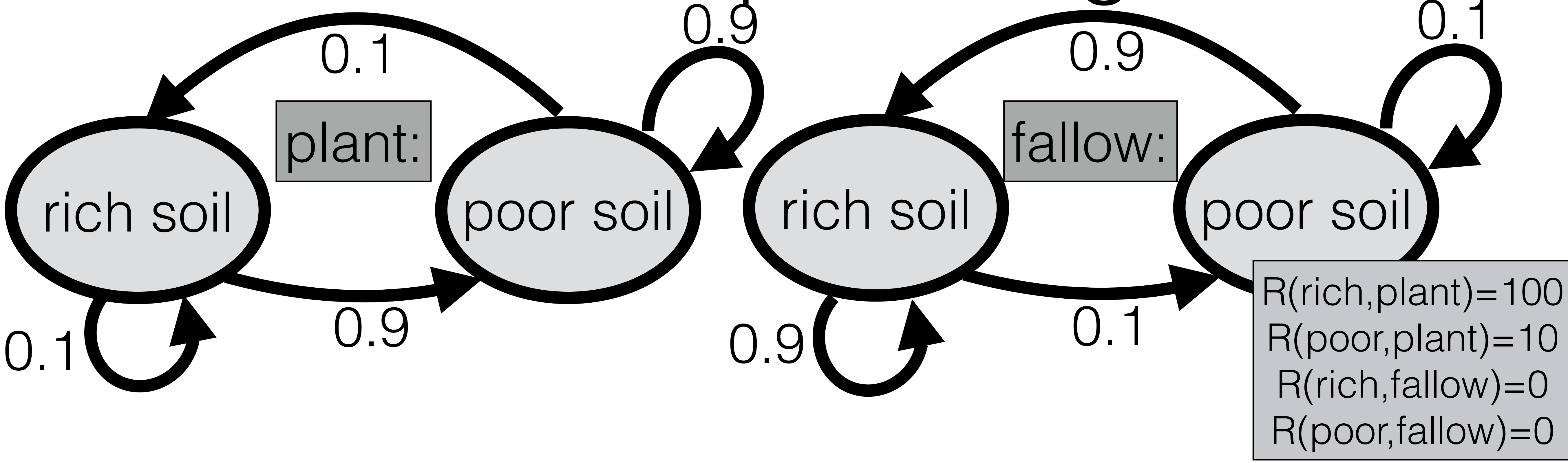
# What if I don't stop farming?




- 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: $\gamma^t$ 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 $\pi$ starting at state *s*

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

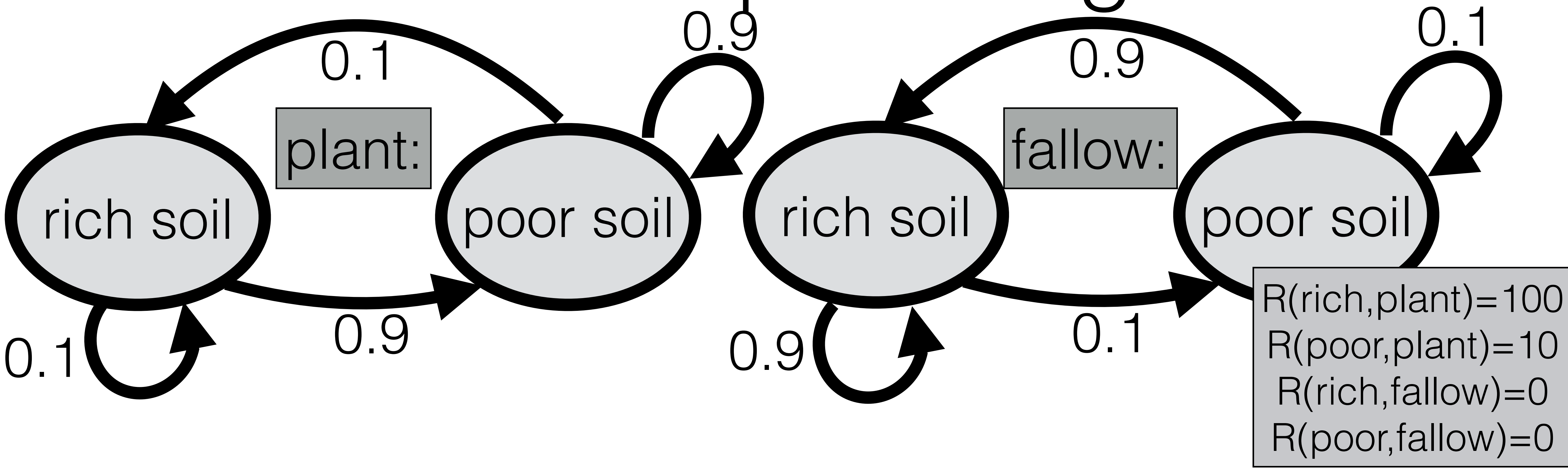
# What if I don't stop farming?




- 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: $\gamma^t$ 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 $\pi$ starting at state *s*

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

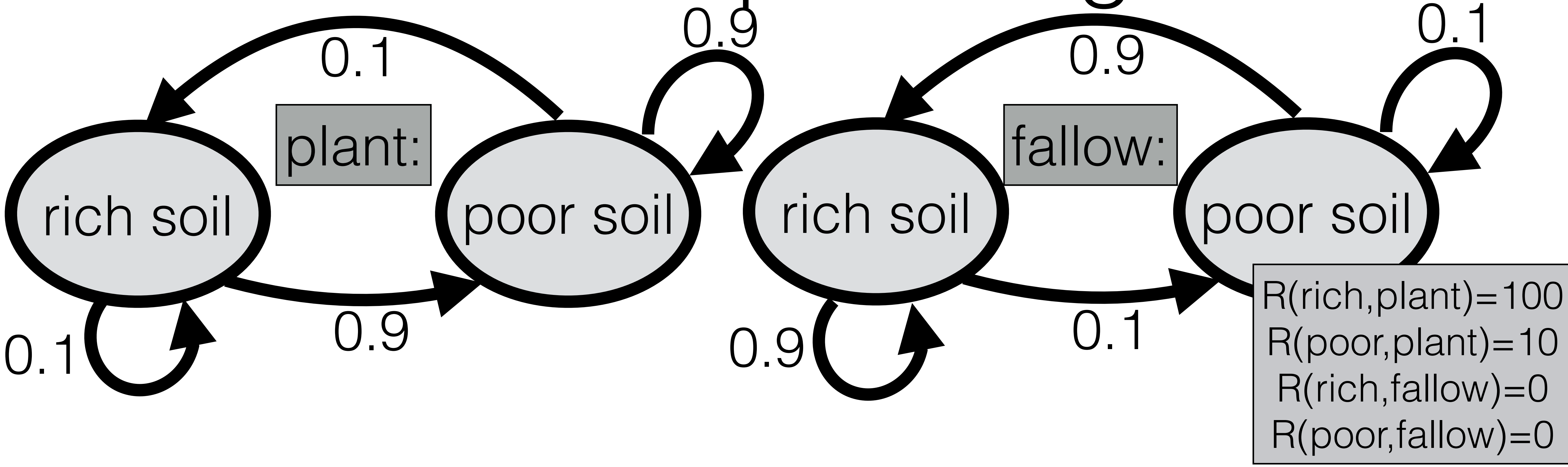
# What if I don't stop farming?




- 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: $\gamma^t$ 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 $\pi$ starting at state *s*

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

# What if I don't stop farming?



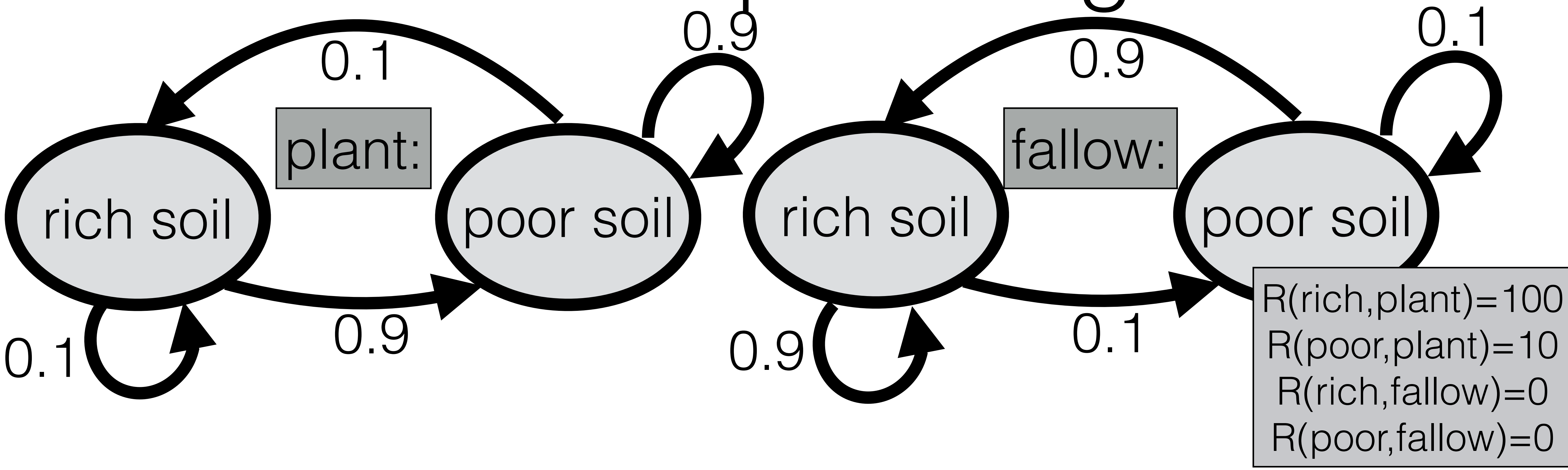

- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: $\gamma^t$ bushels
  - Example: What's the value of 1 bushel per year forever?

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

    $V = 1/(1-\gamma)$ E.g. $\gamma = 0.99 \Rightarrow V = 1/0.01 = 100$ bushels
- $V_\pi(s)$: expected reward with policy $\pi$ starting at state *s*

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

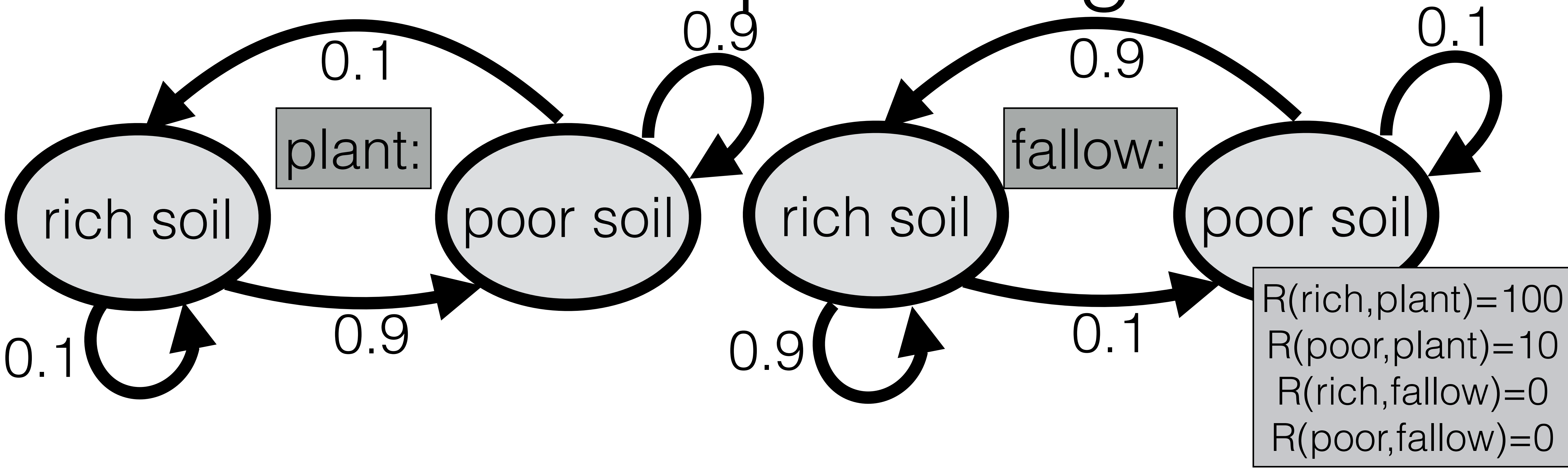
# What if I don't stop farming?




- 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: $\gamma^t$ 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 $\pi$ starting at state *s*

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

# What if I don't stop farming?

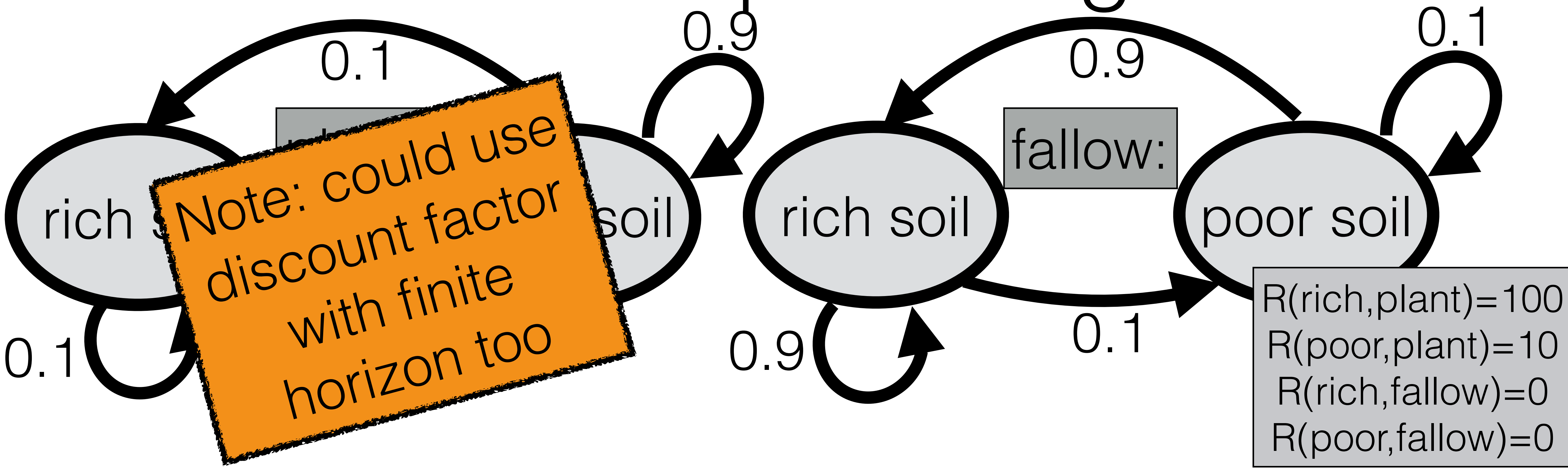



- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: $\gamma^t$ bushels
  - Example: What's the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 +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