6.036: Introduction to Machine Learning

MIT EECS
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

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

Lecture start: Tuesdays 9:35am

Who's talking? Prof. Tamara Broderick

Questions? Ask on Piazza: "lecture (week) 11" folder

Materials: slides, video will all be available on Canvas

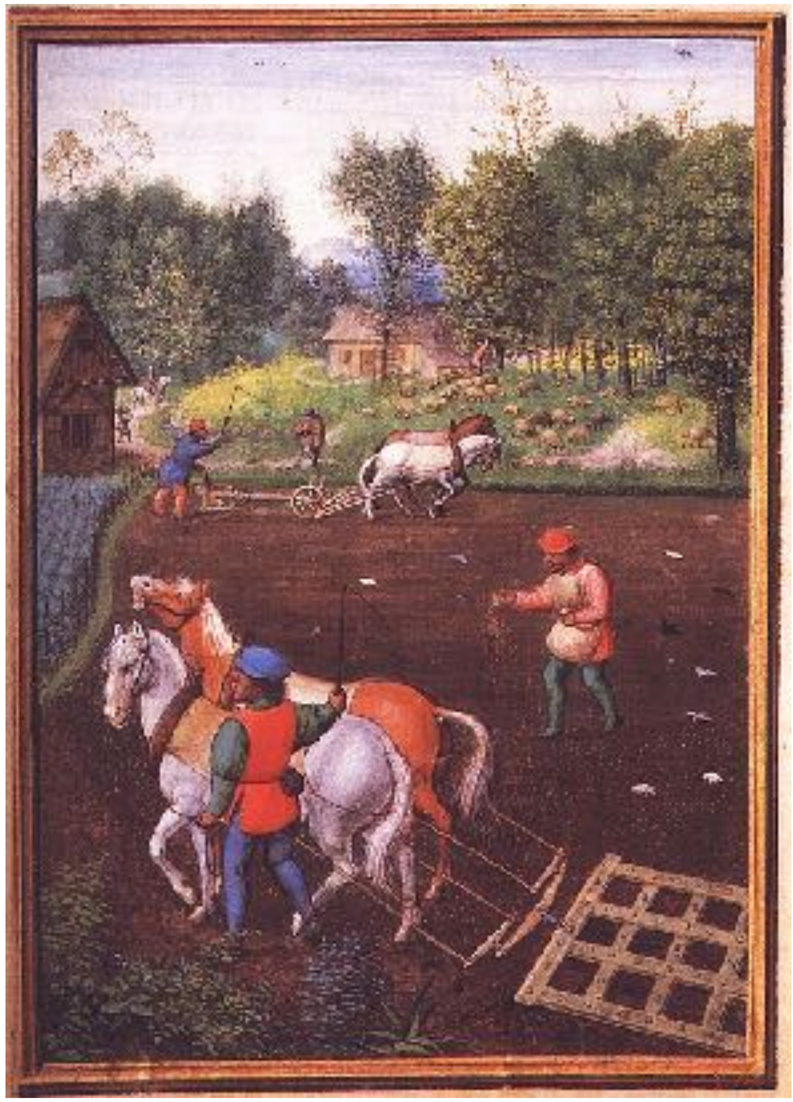
Live Zoom feed: https://mit.zoom.us/j/94238622313

Last Time(s)

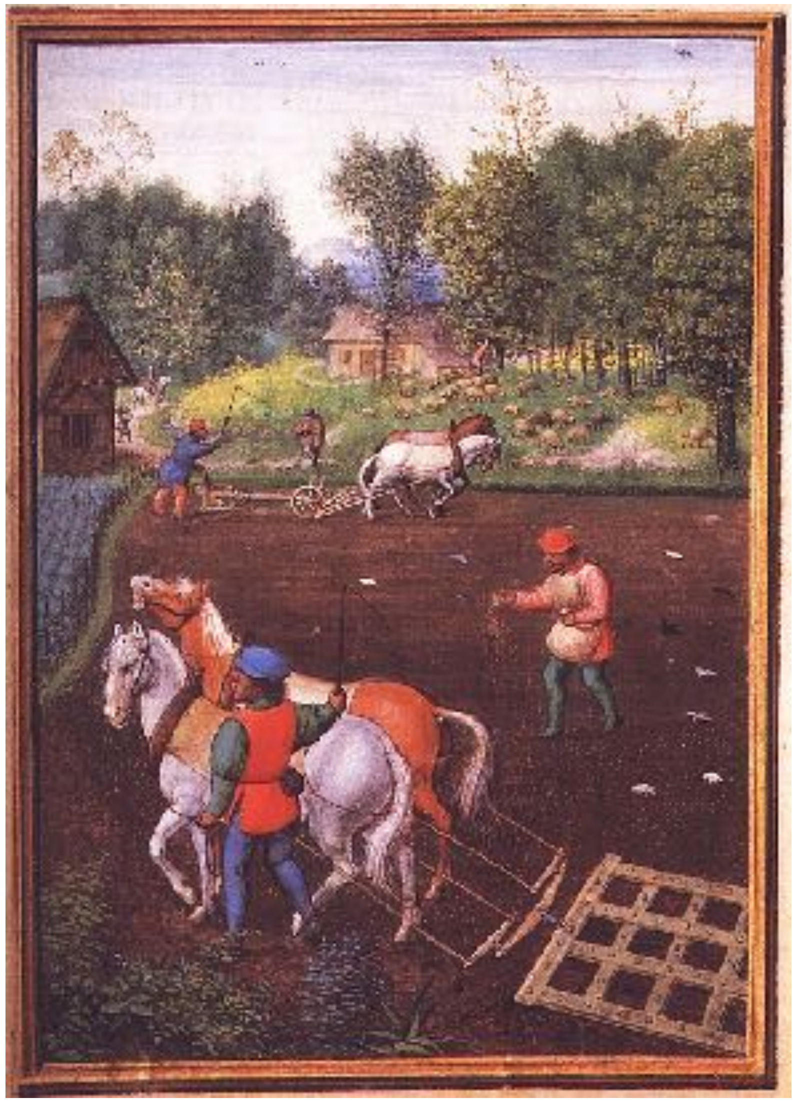
I. Supervised learning
II. Unsupervised learning
III. Decisions incur loss but don't have broader effect

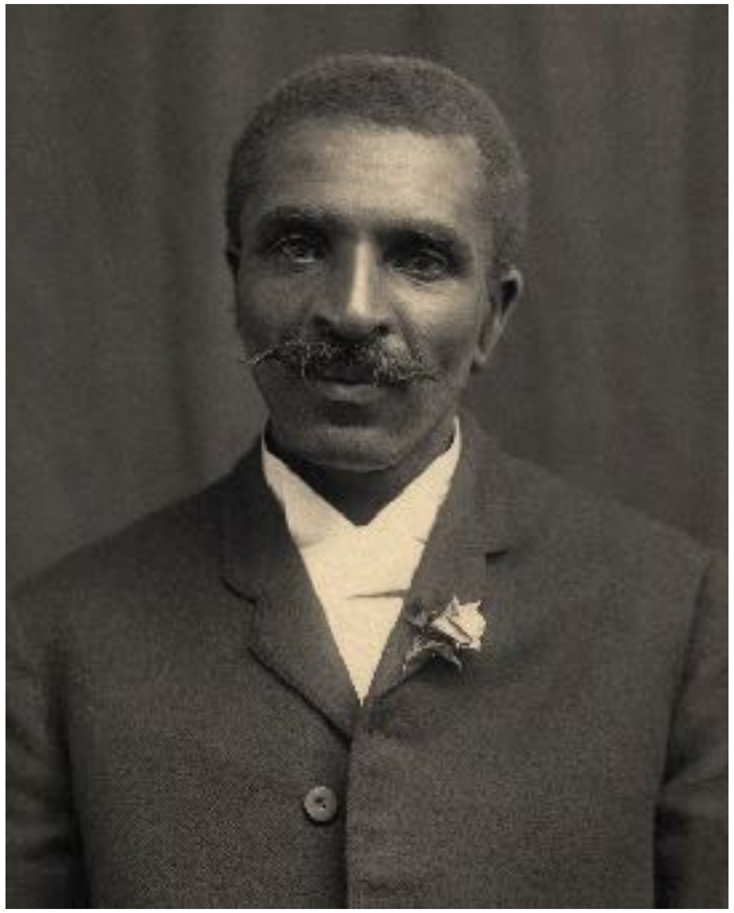
Today's Plan

I. Decisions change the state of the world
II. State machines
III. Markov decision processes (MDPs)

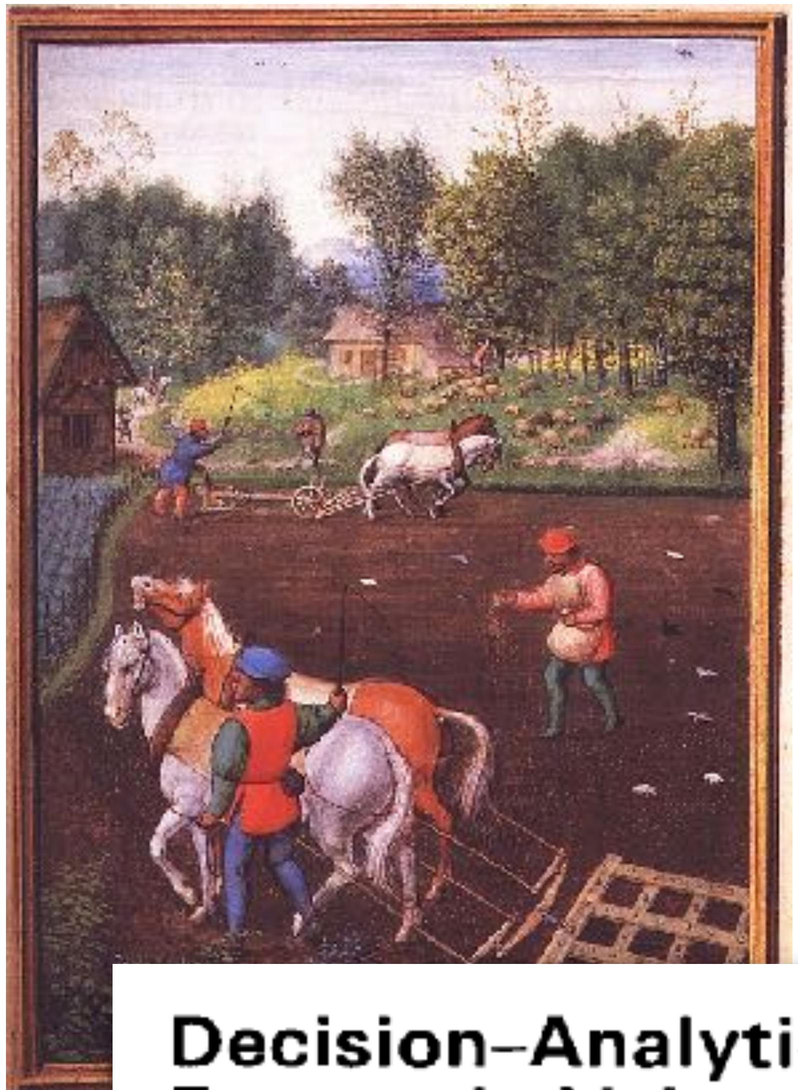


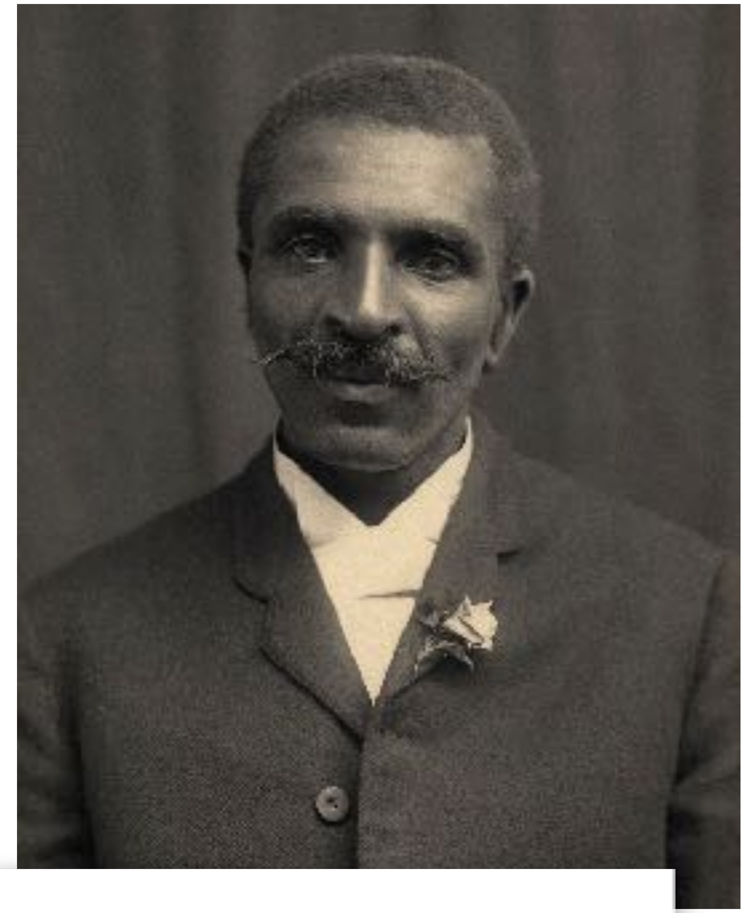
[https://en.wikipedia.org/wiki/Sowing#/media/File:Simon_Bening_-_September.jpg]





[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]

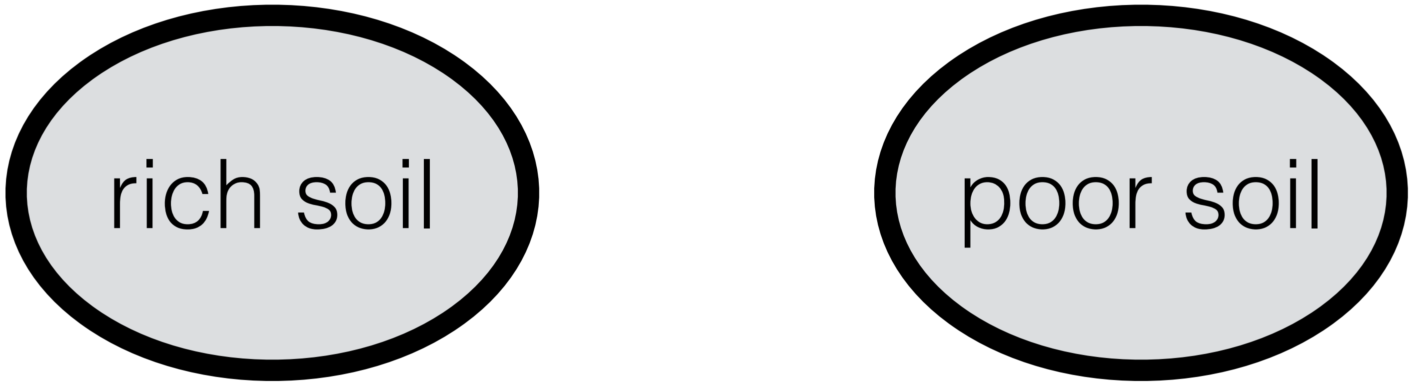
State Machine

State Machine

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

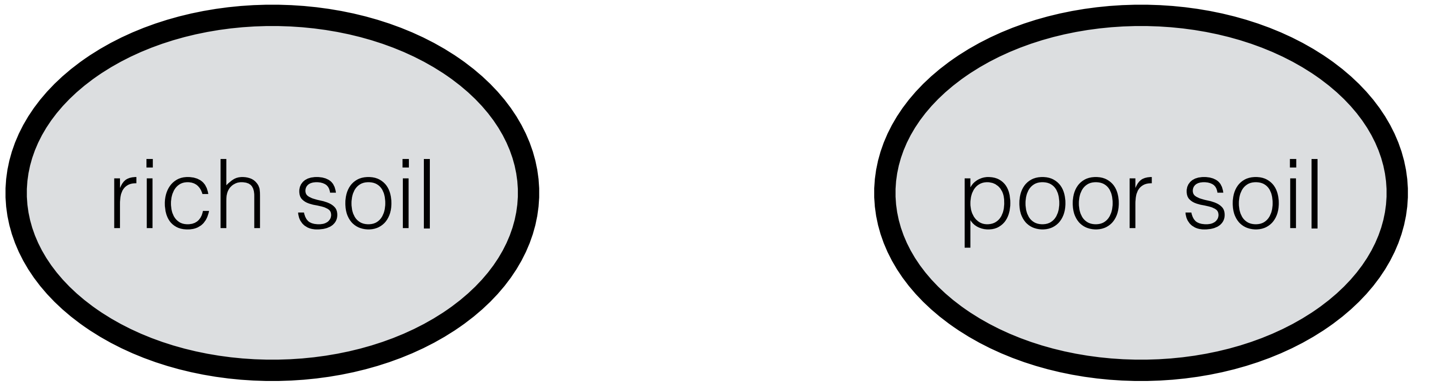
State Machine

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



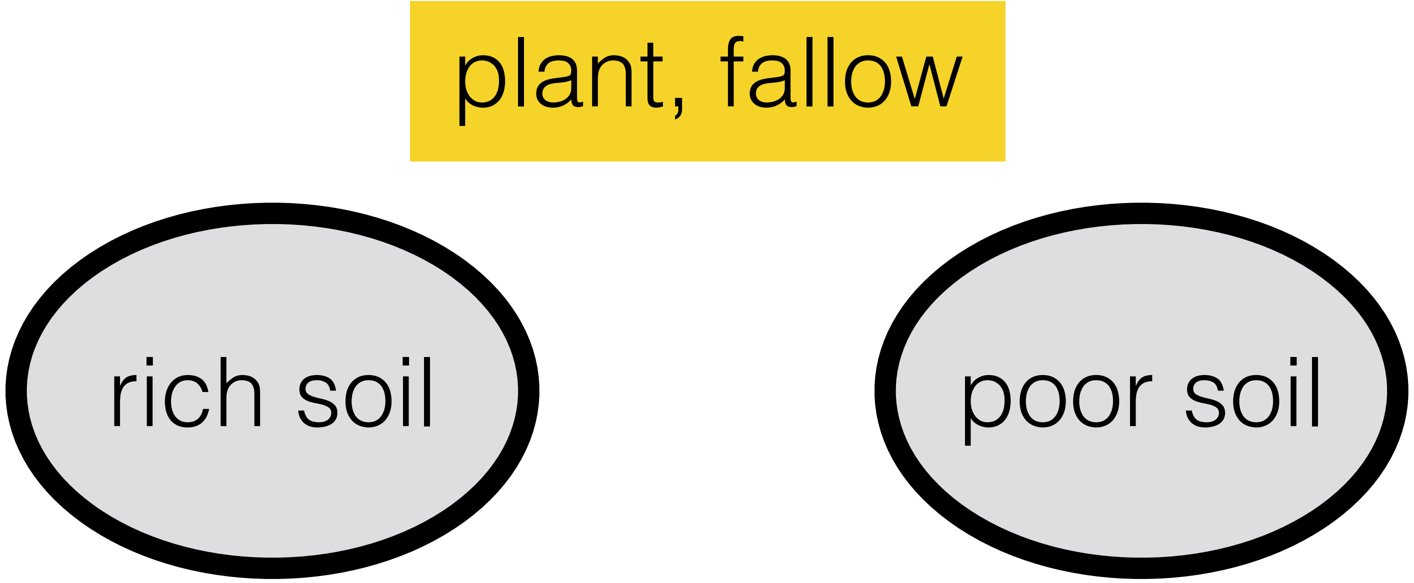
State Machine

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



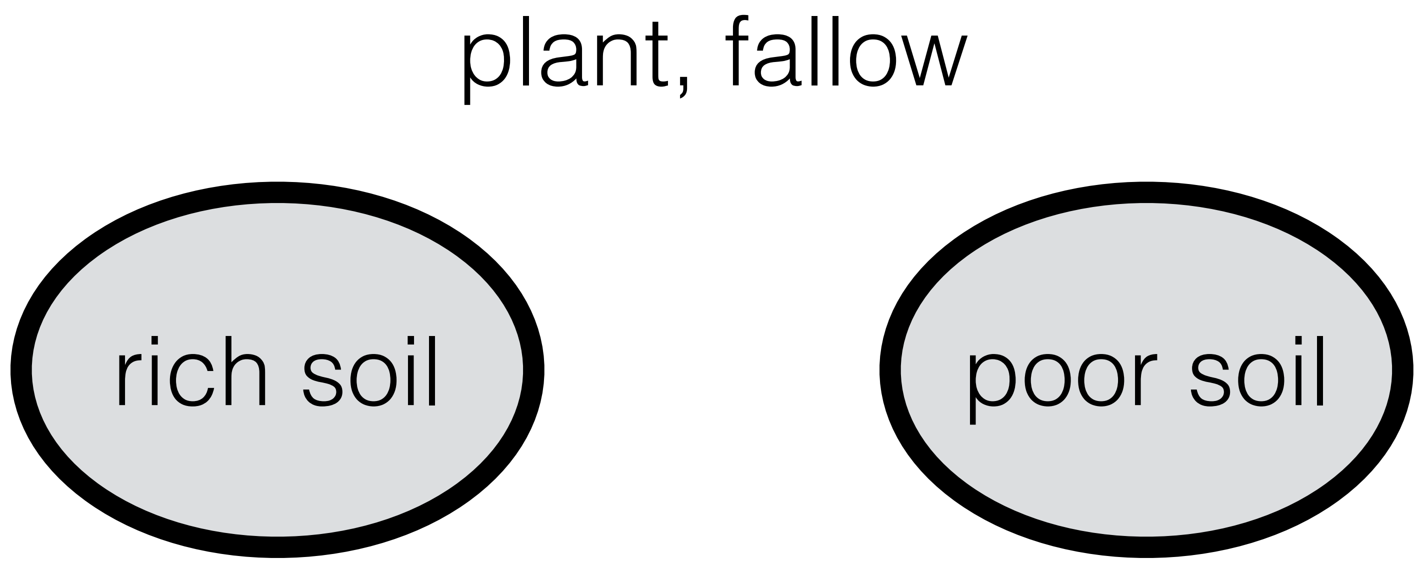
State Machine

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



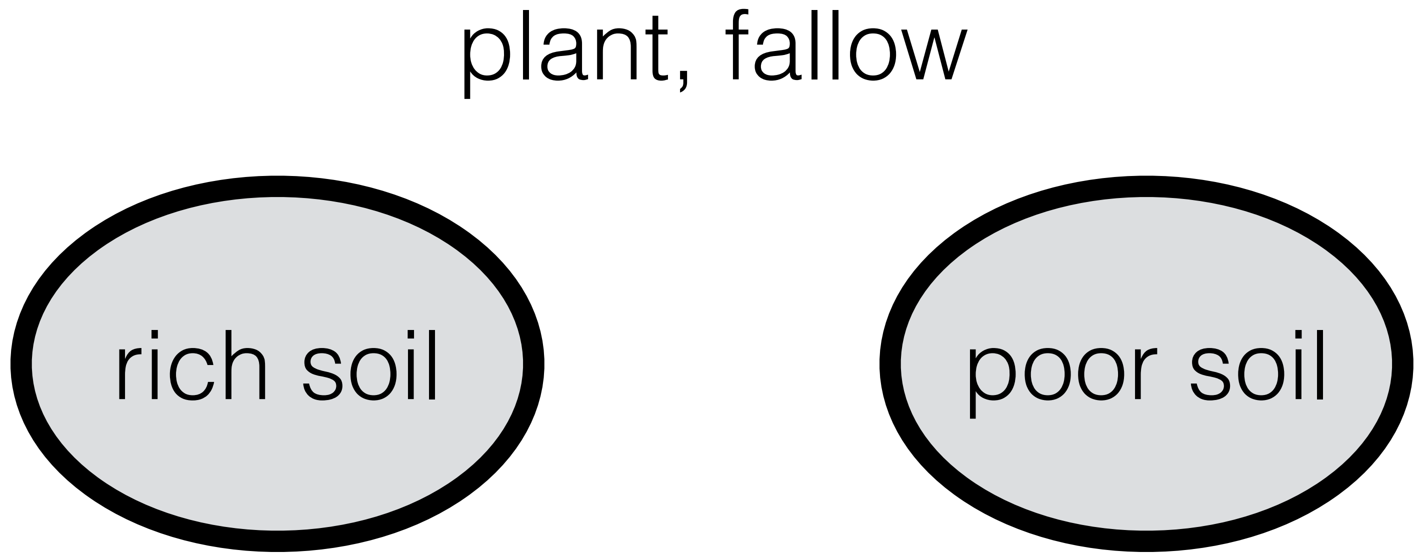
State Machine

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



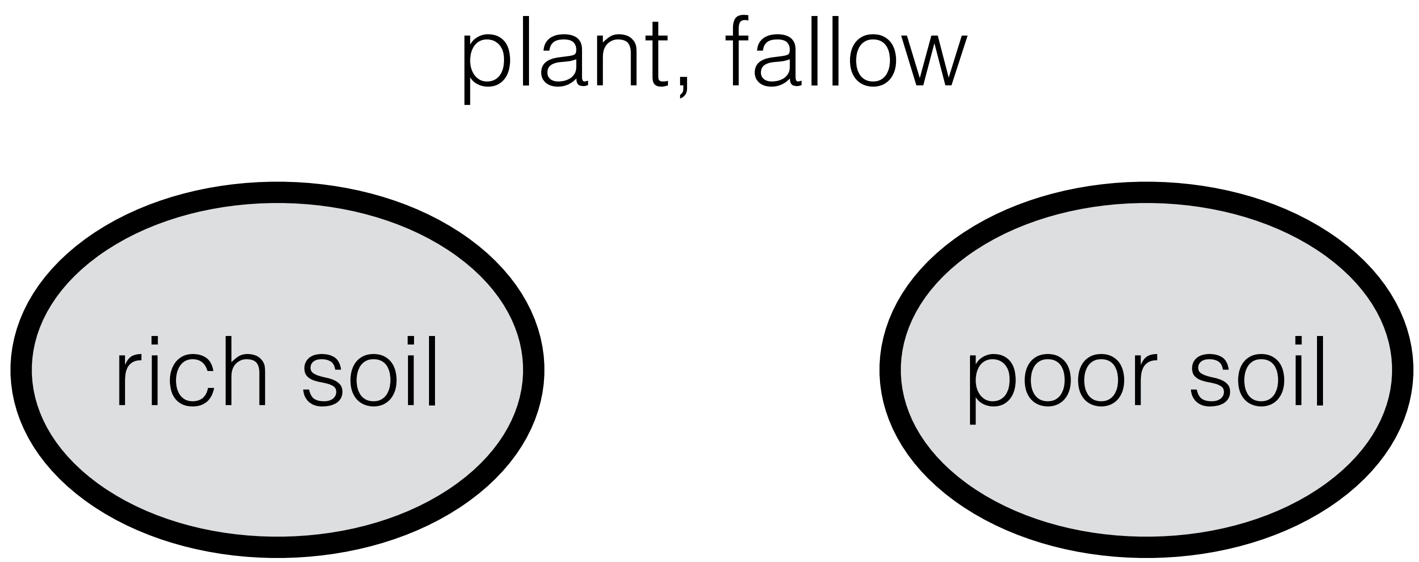
State Machine

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



State Machine

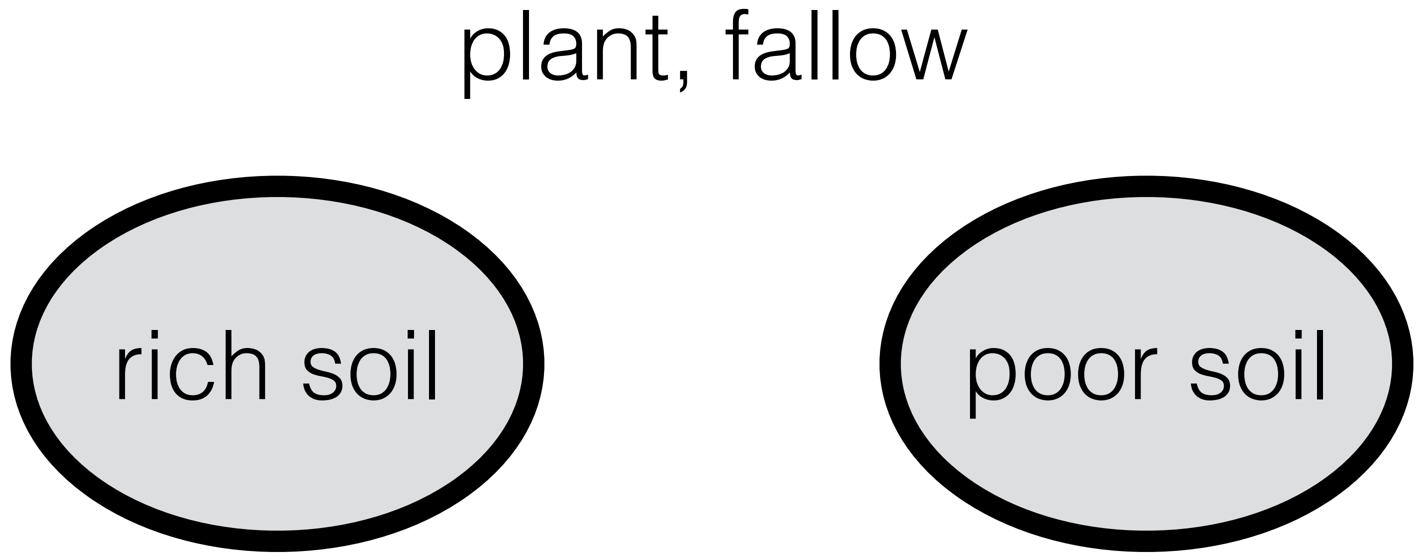
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state



Example

State Machine

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

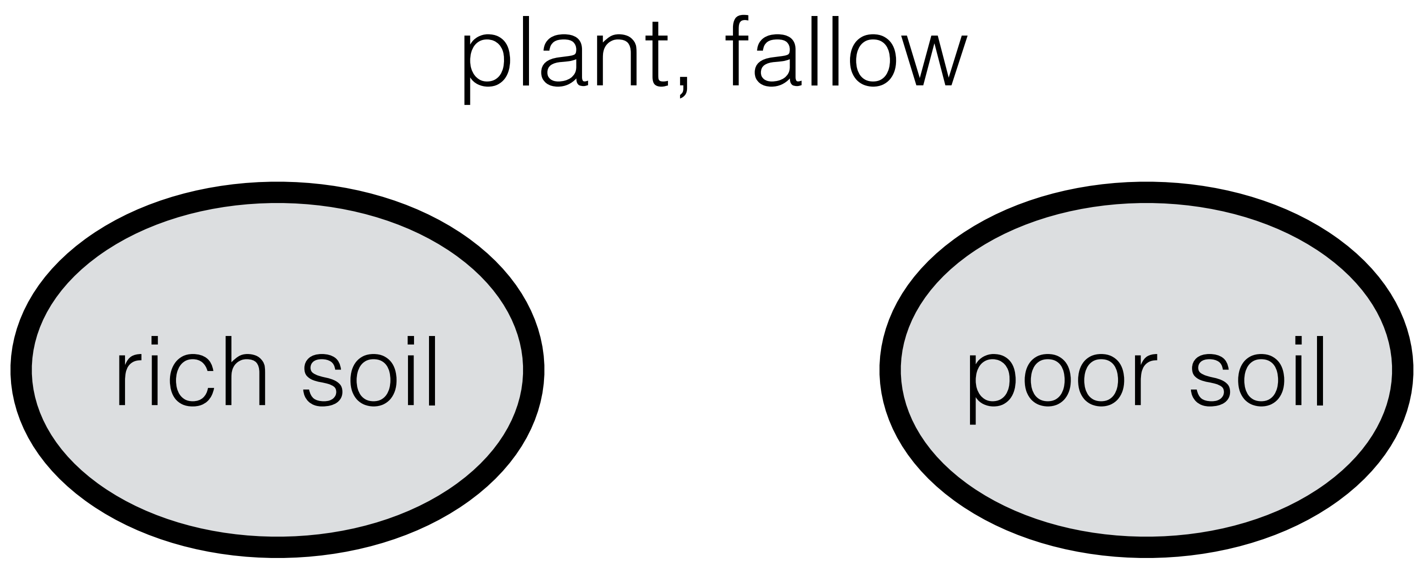


Example

$s_0 = \text{rich}$

State Machine

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

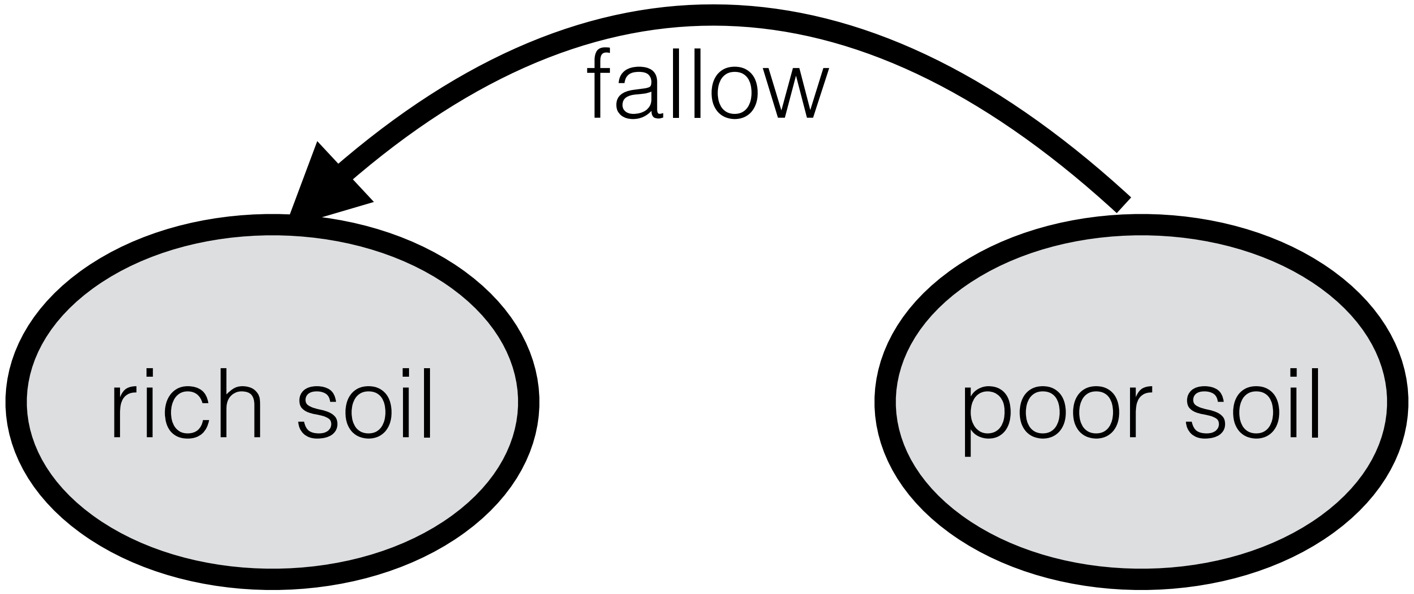


Example

$s_0 = \text{rich}$

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

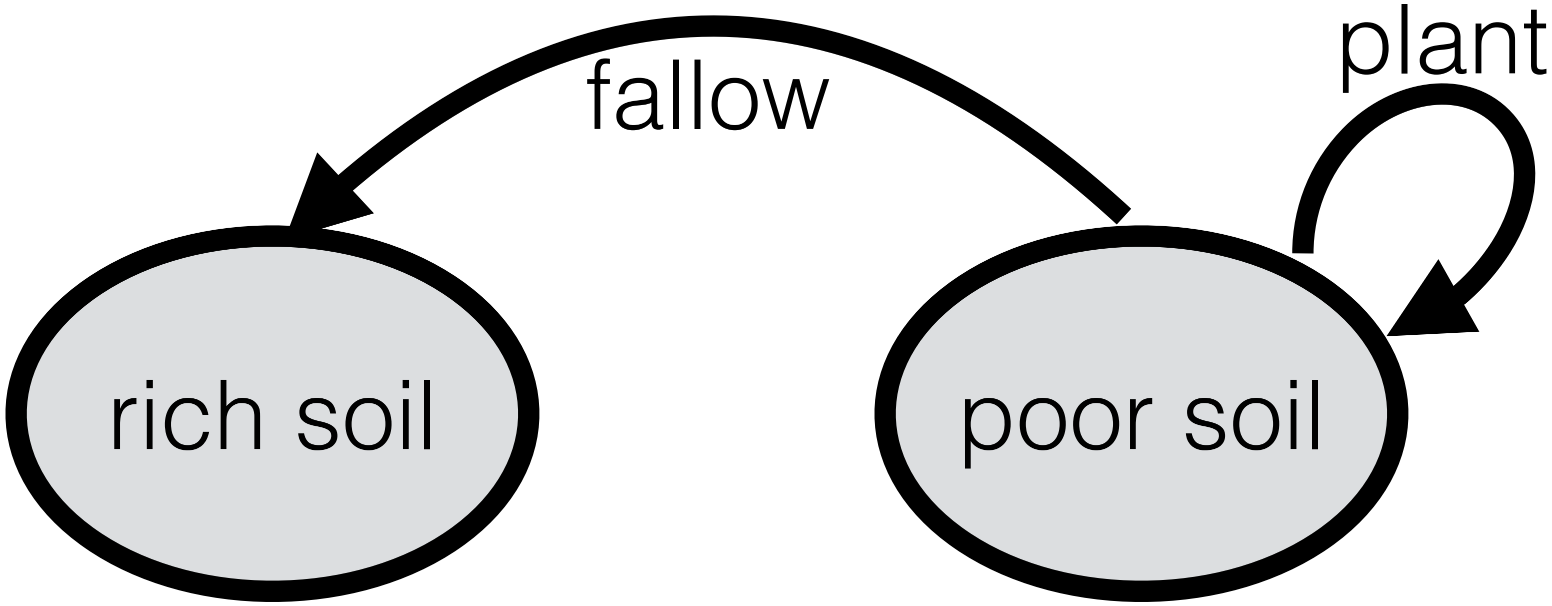


Example

$s_0 = \text{rich}$

State Machine

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

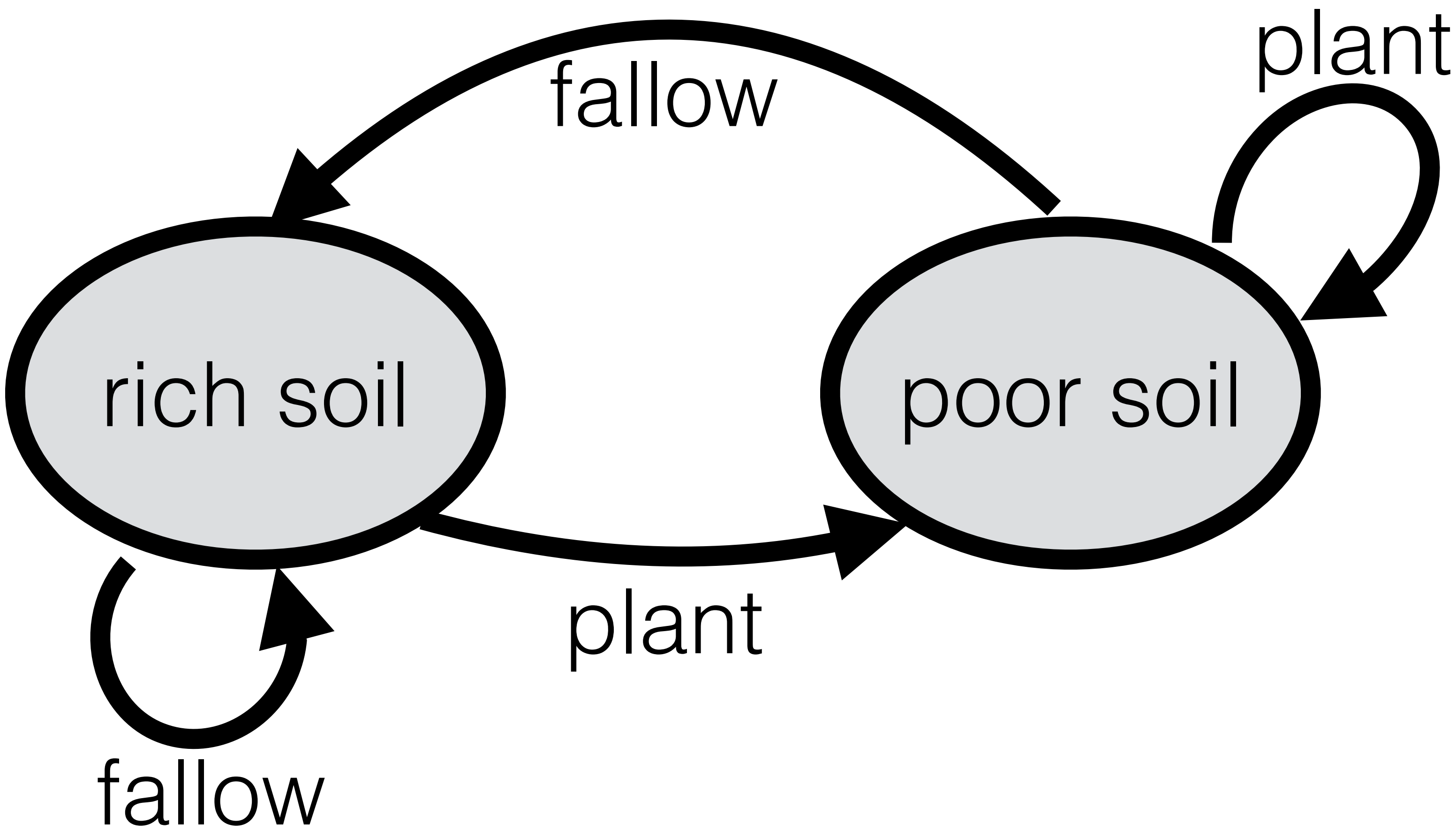


Example

$s_0 = \text{rich}$

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

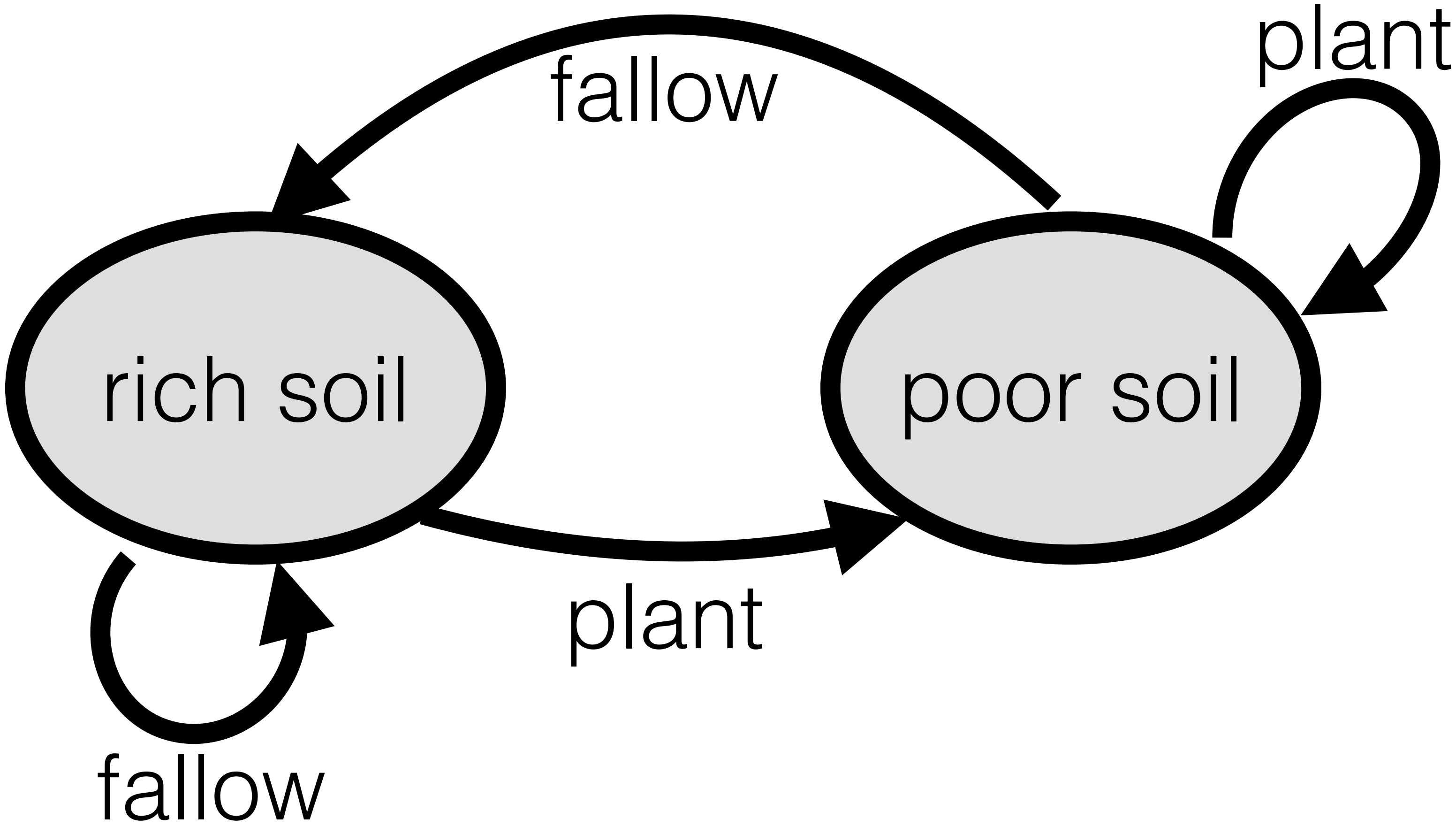


Example

$s_0 = \text{rich}$

State Machine

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



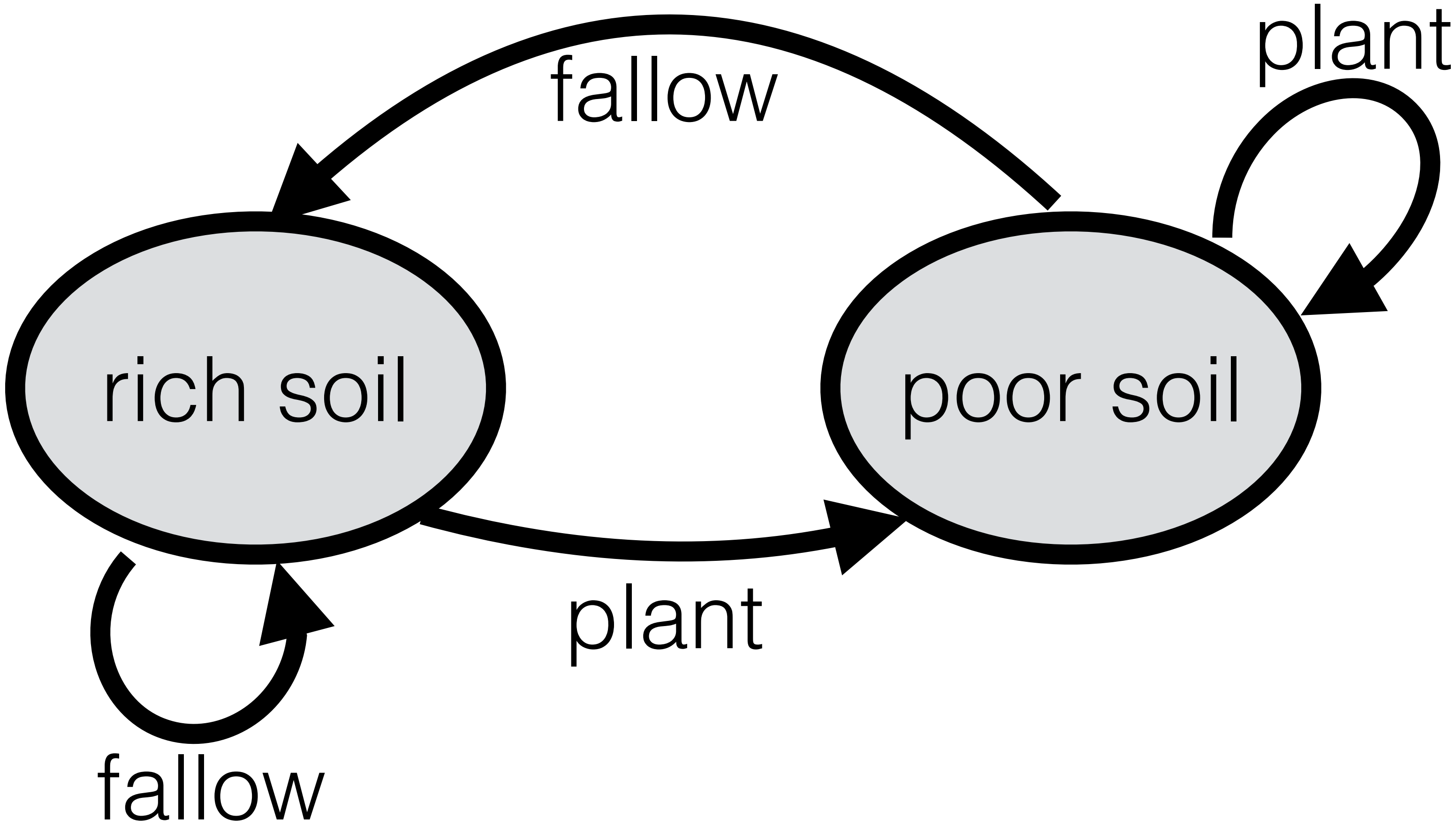
Example

$s_0 = \text{rich}$

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

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



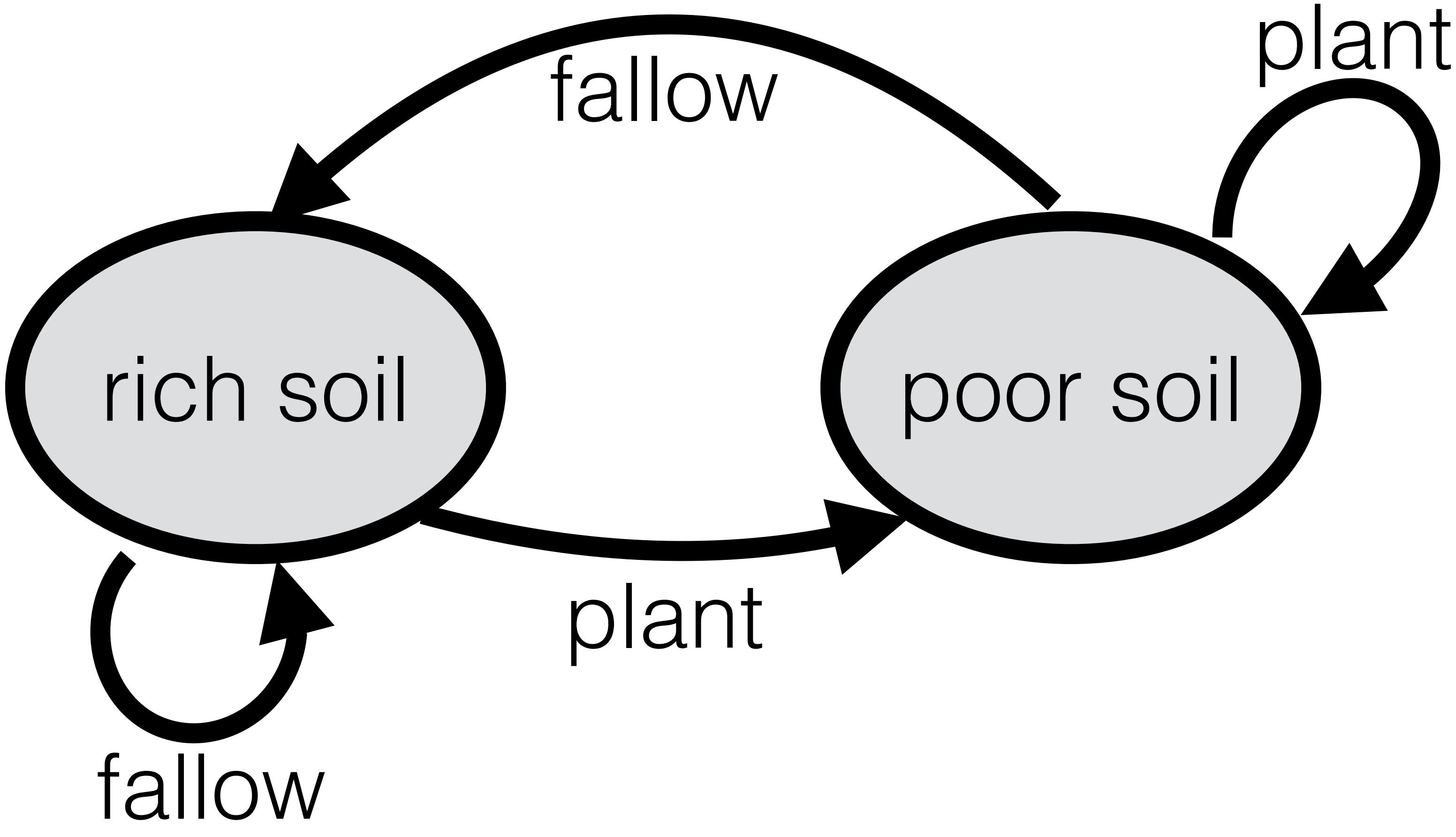
Example

$s_0 = \text{rich}$

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

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



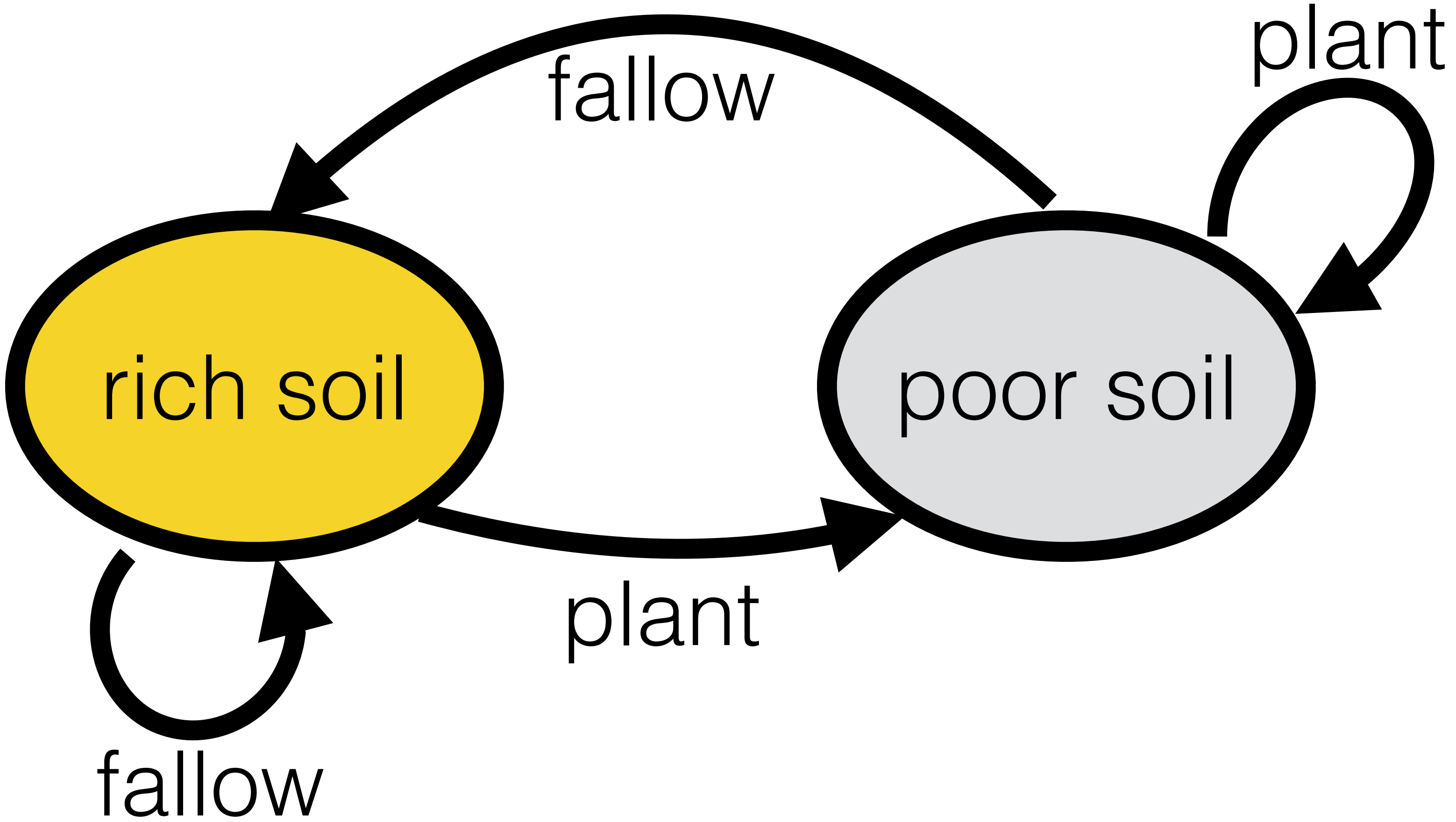
Example

$s_0 = \text{rich}$

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

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



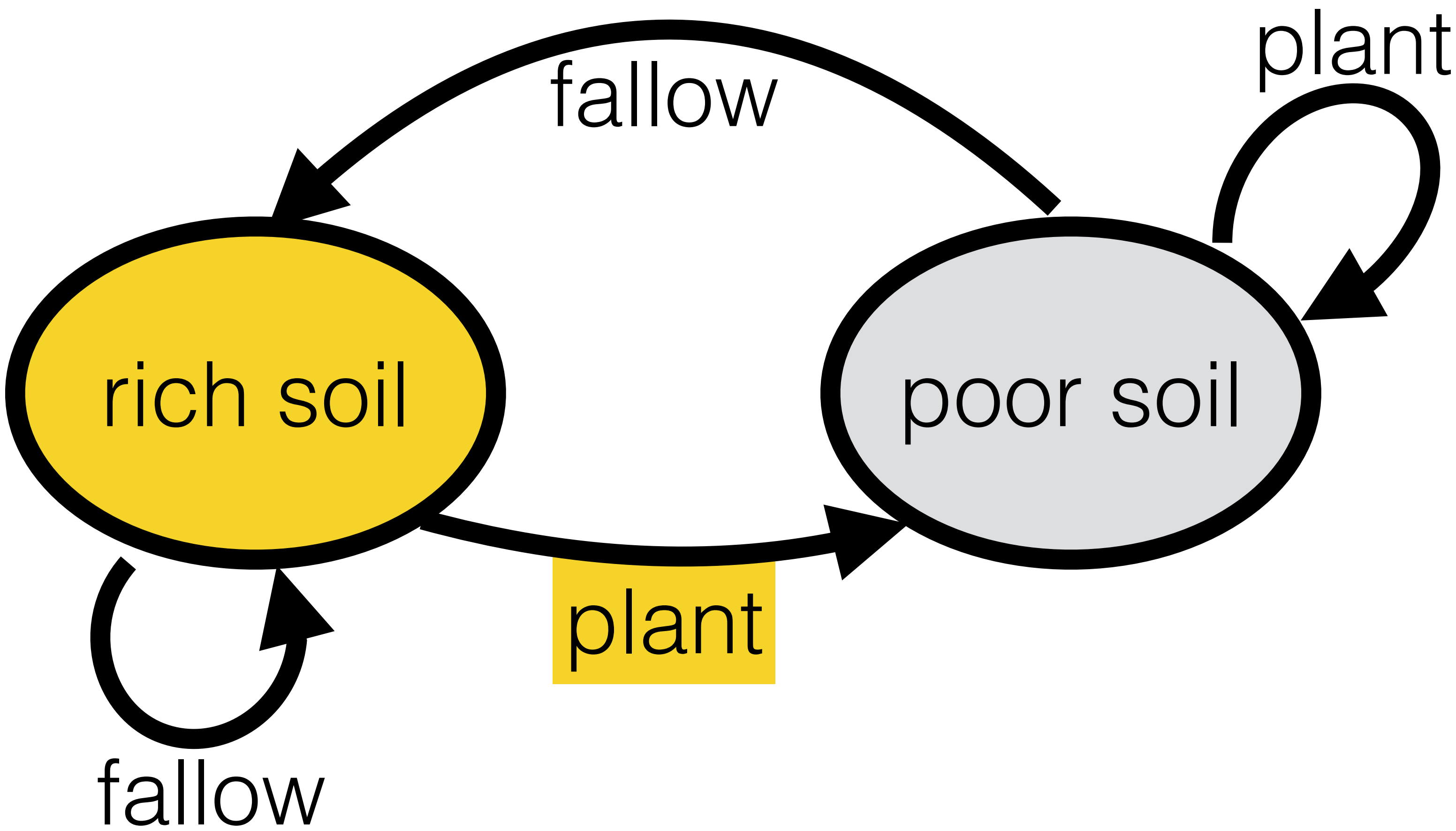
Example

$s_0 = \text{rich}$

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

State Machine

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



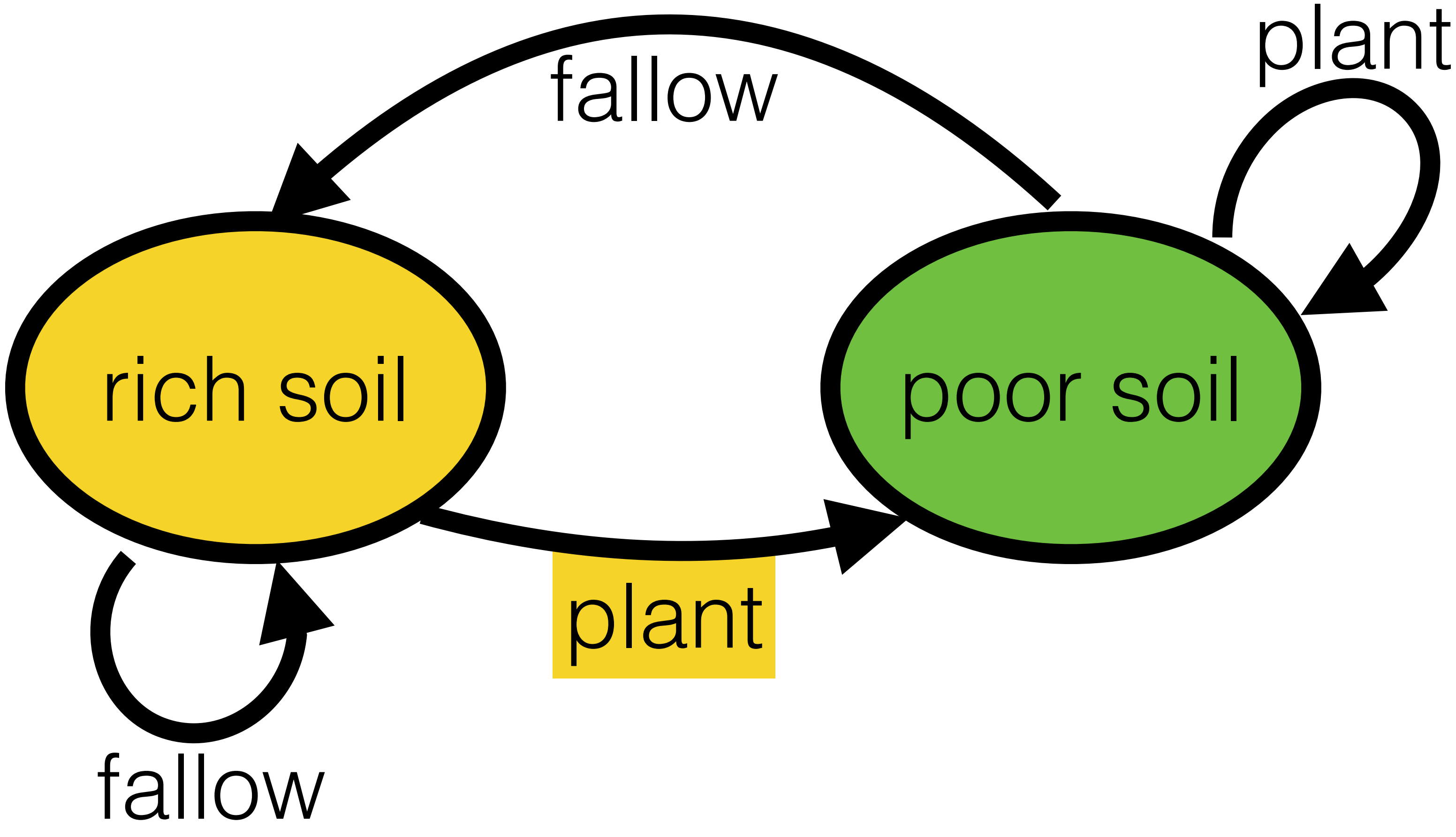
Example

$s_0 = \text{rich}$

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

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



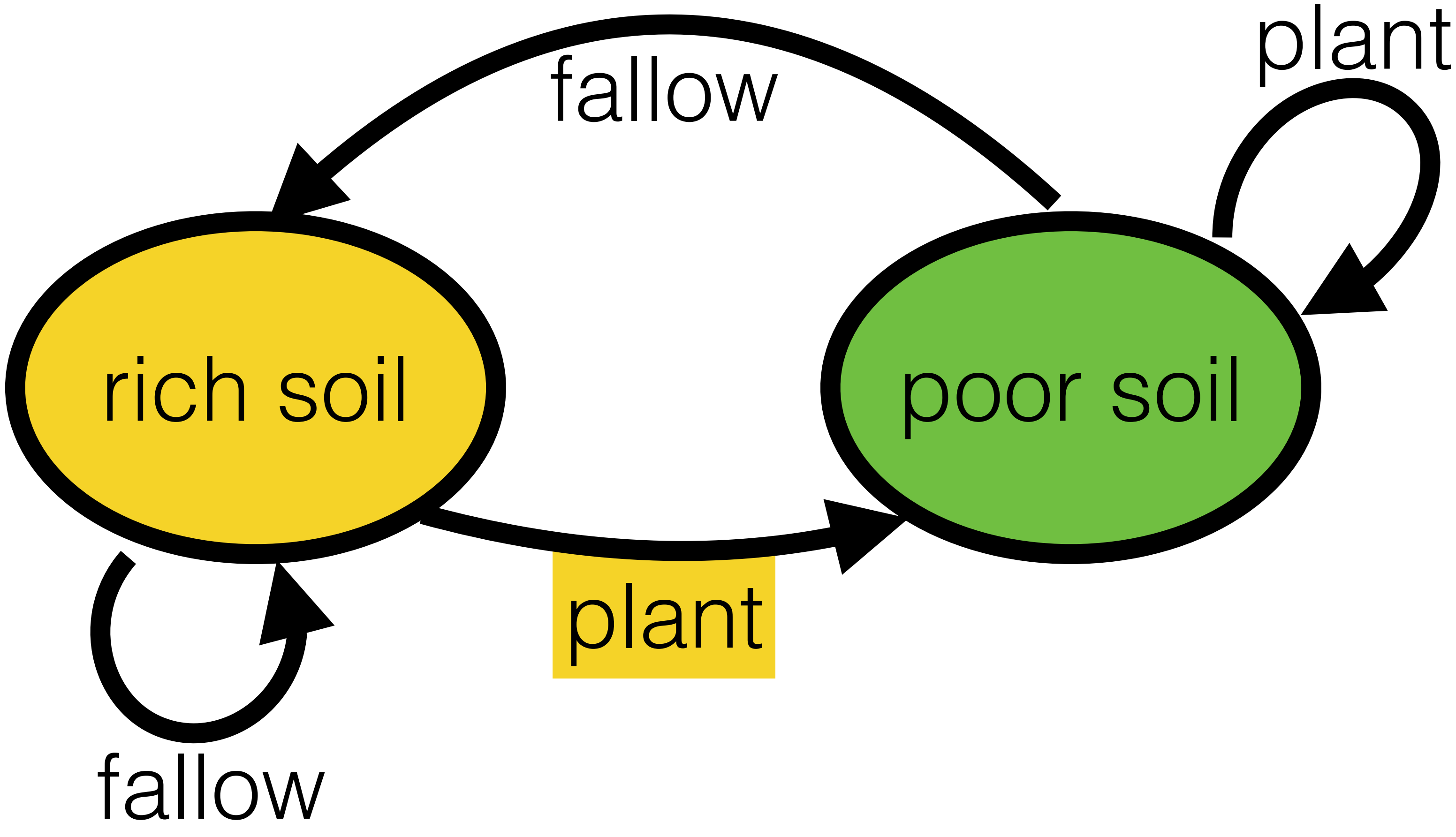
Example

$s_0 = \text{rich}$

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

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

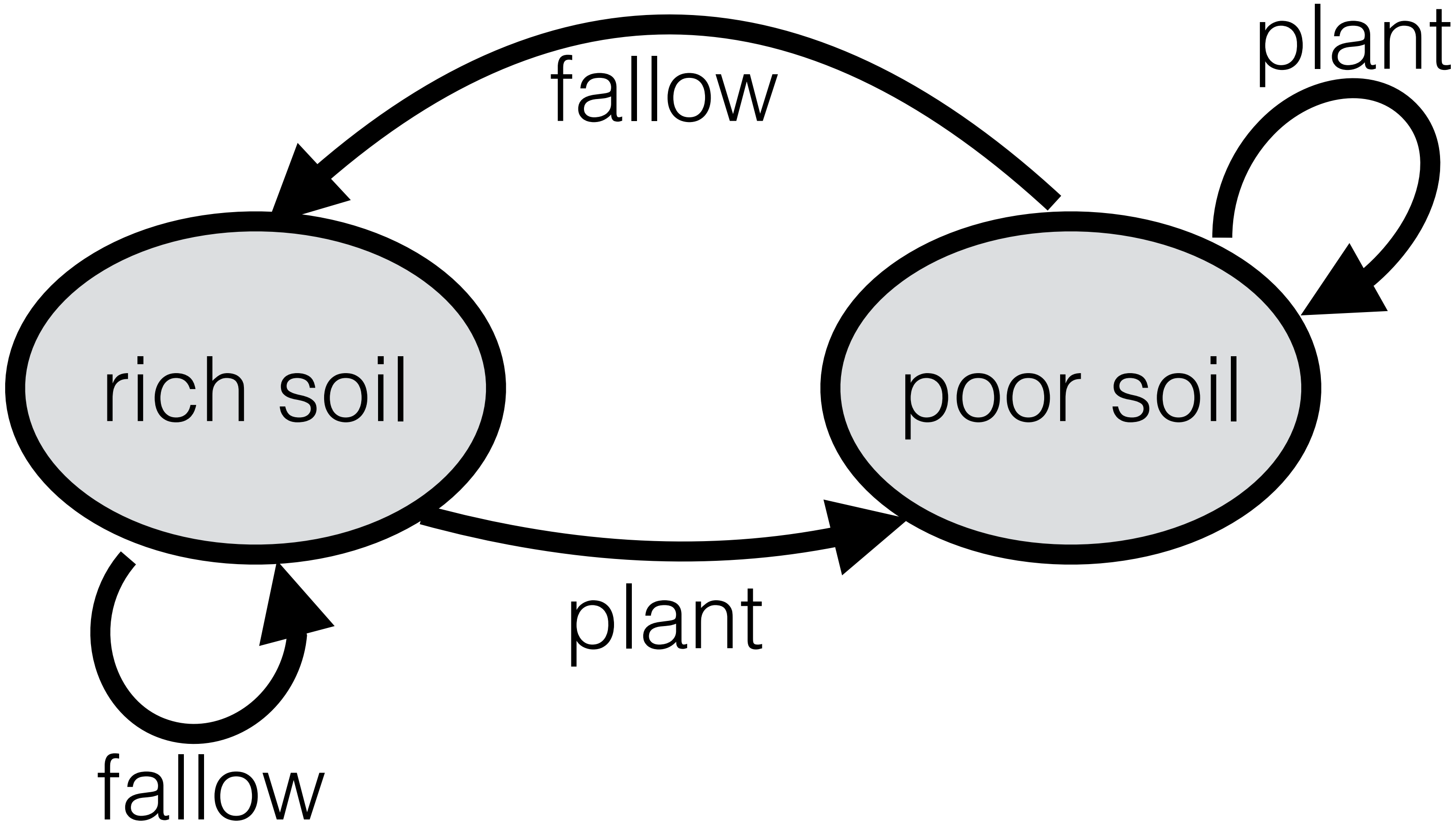


Example

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

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



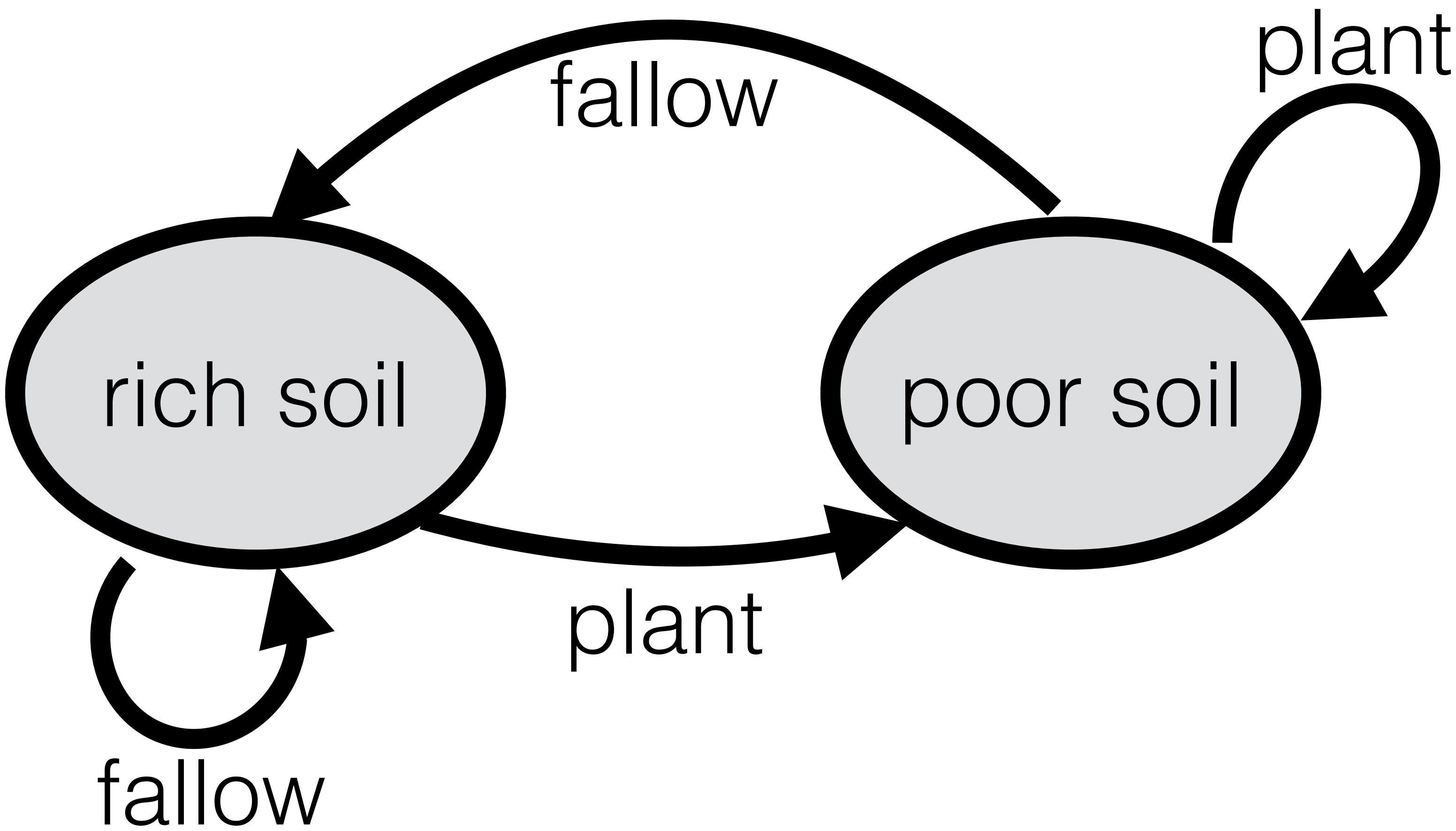
Example

$s_0 = \text{rich}$

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

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

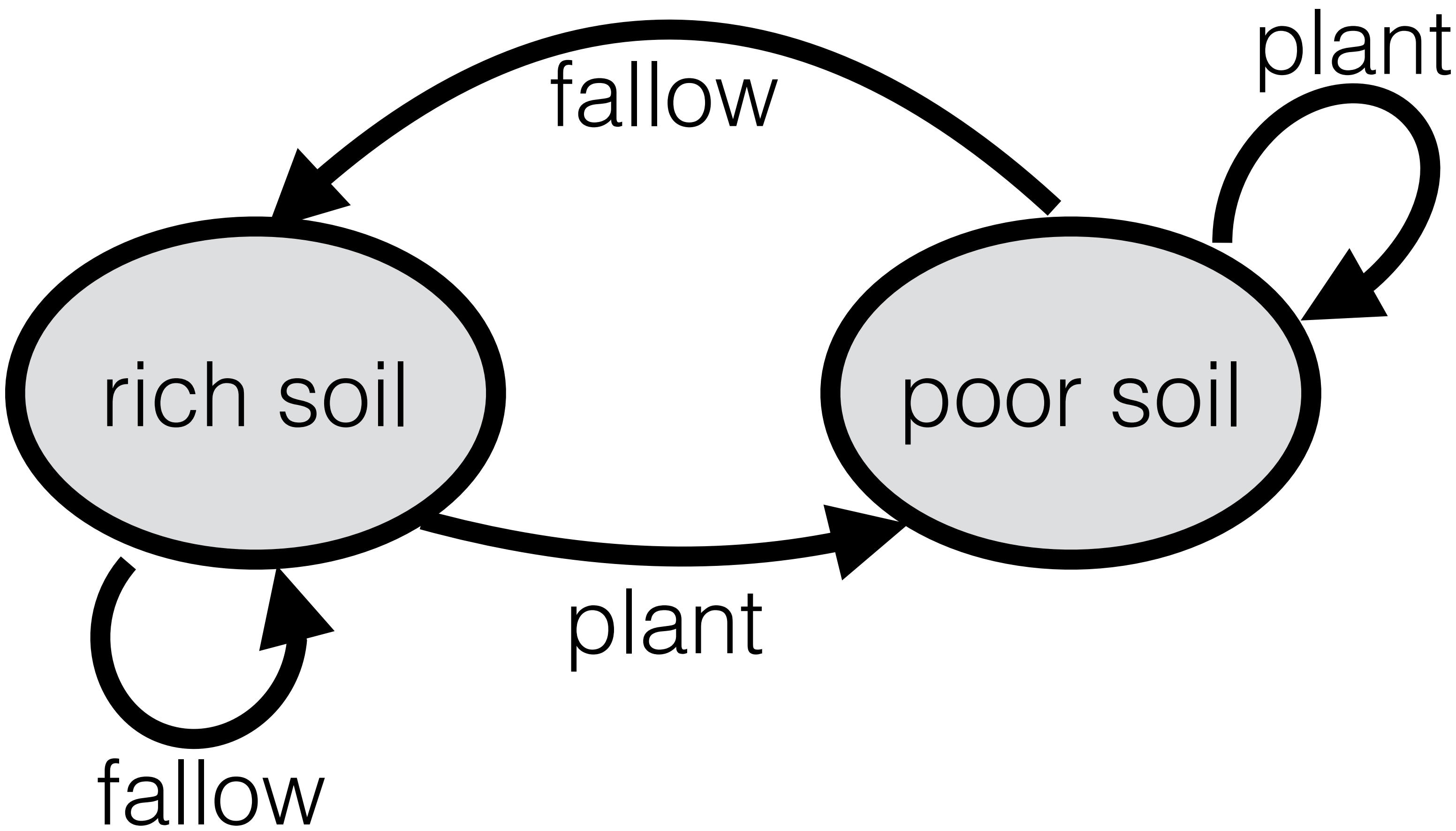


Example

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

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

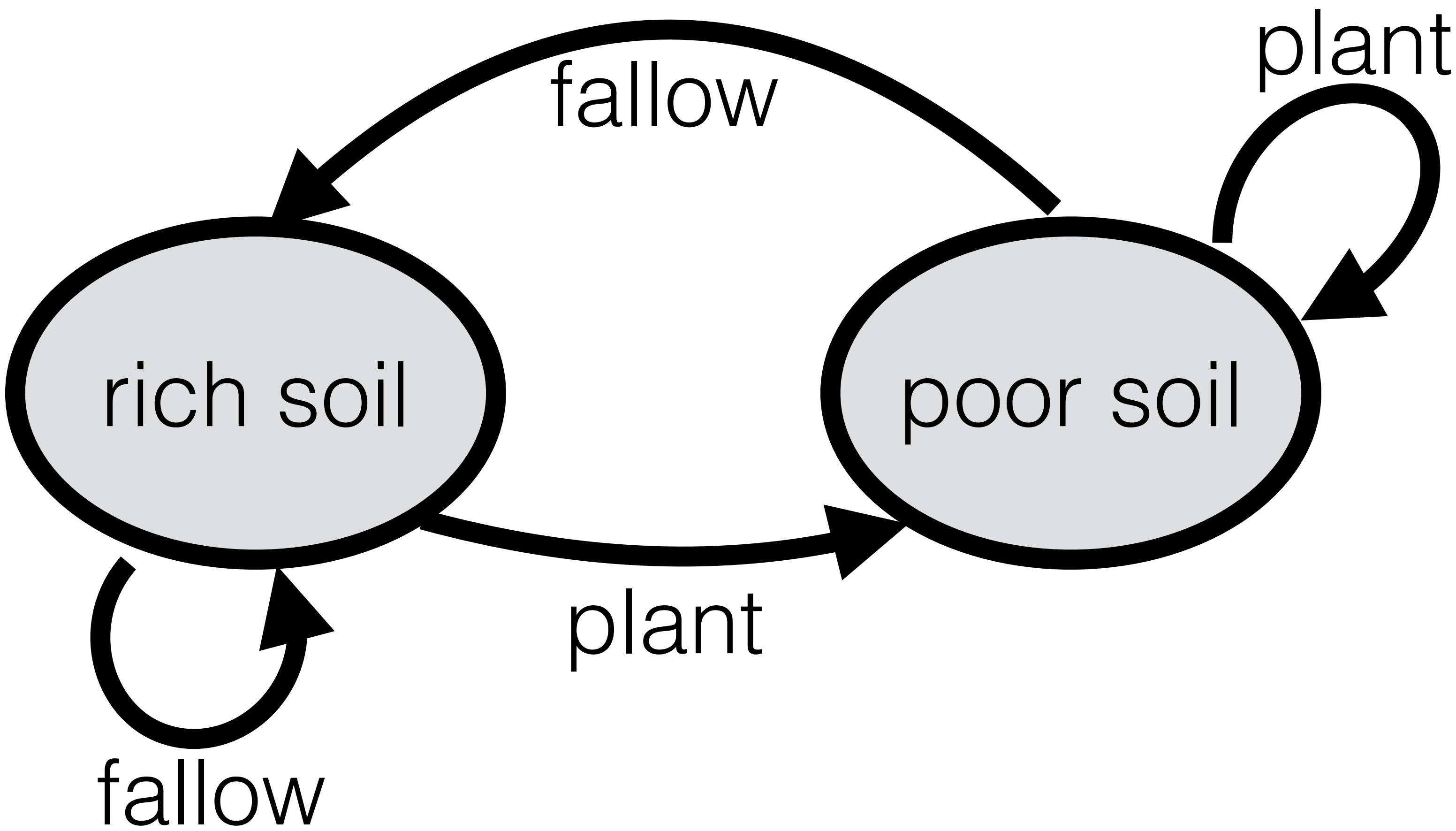


Example

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

State Machine

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

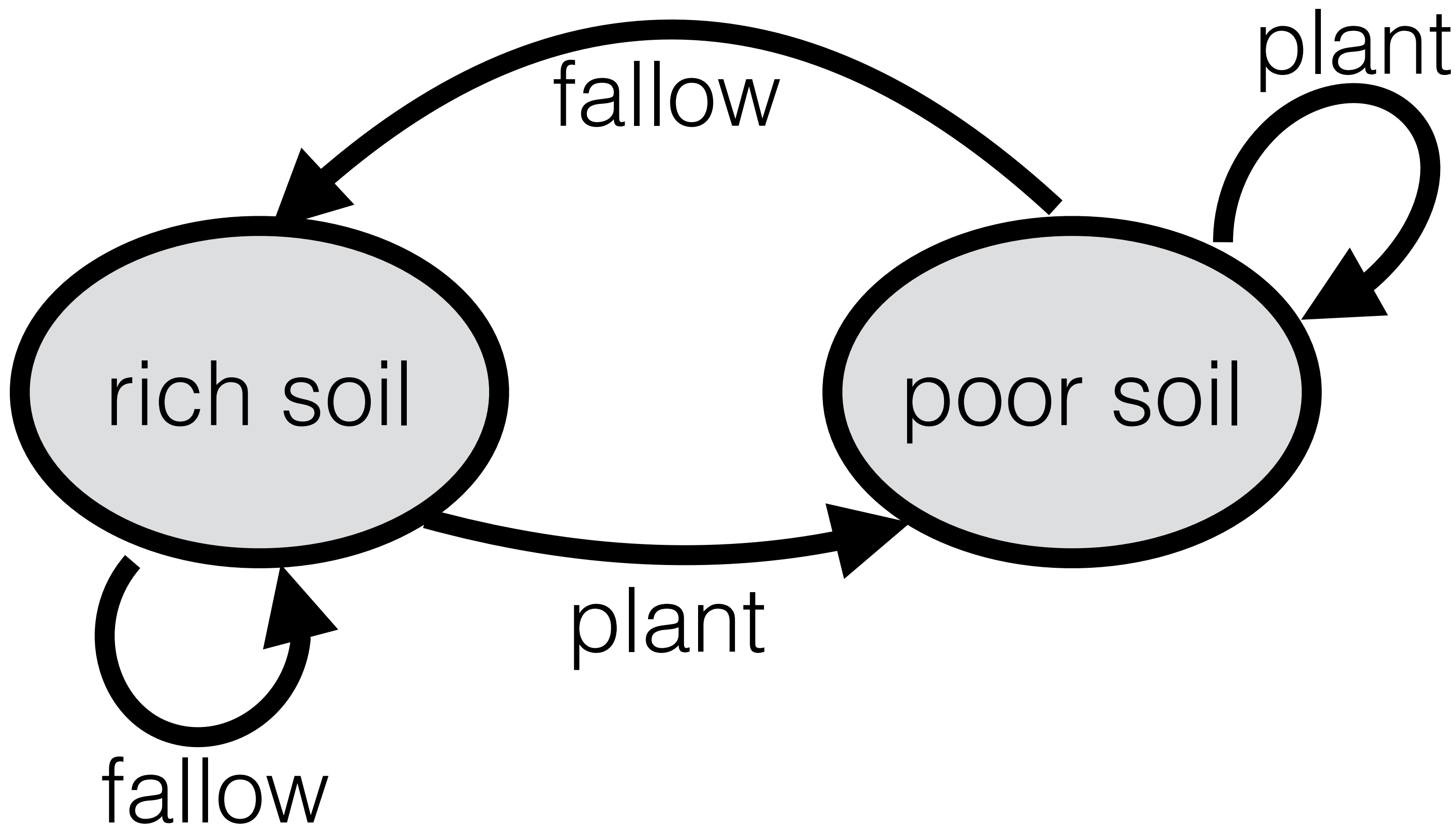


Example

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

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)

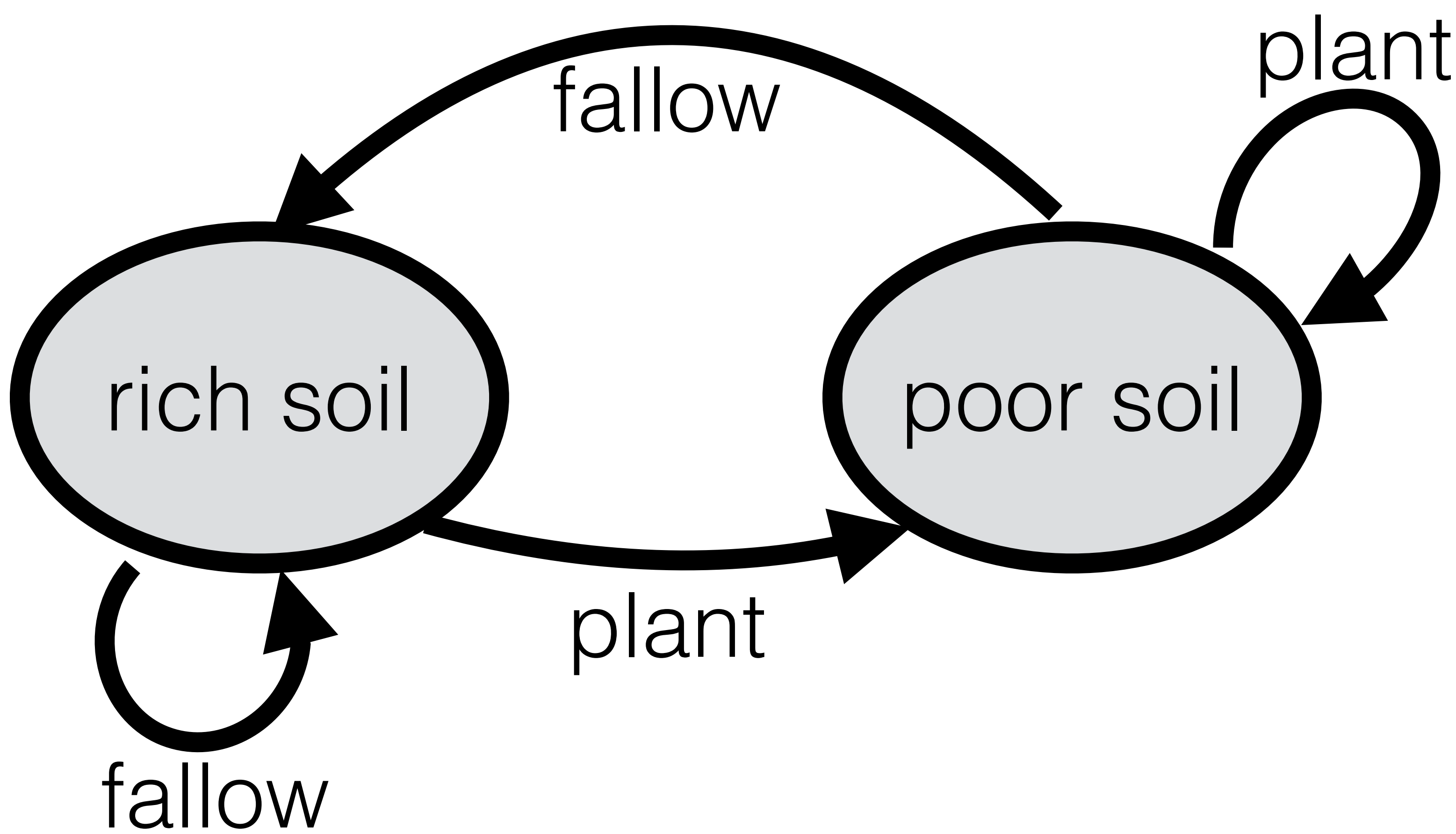


Example

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

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

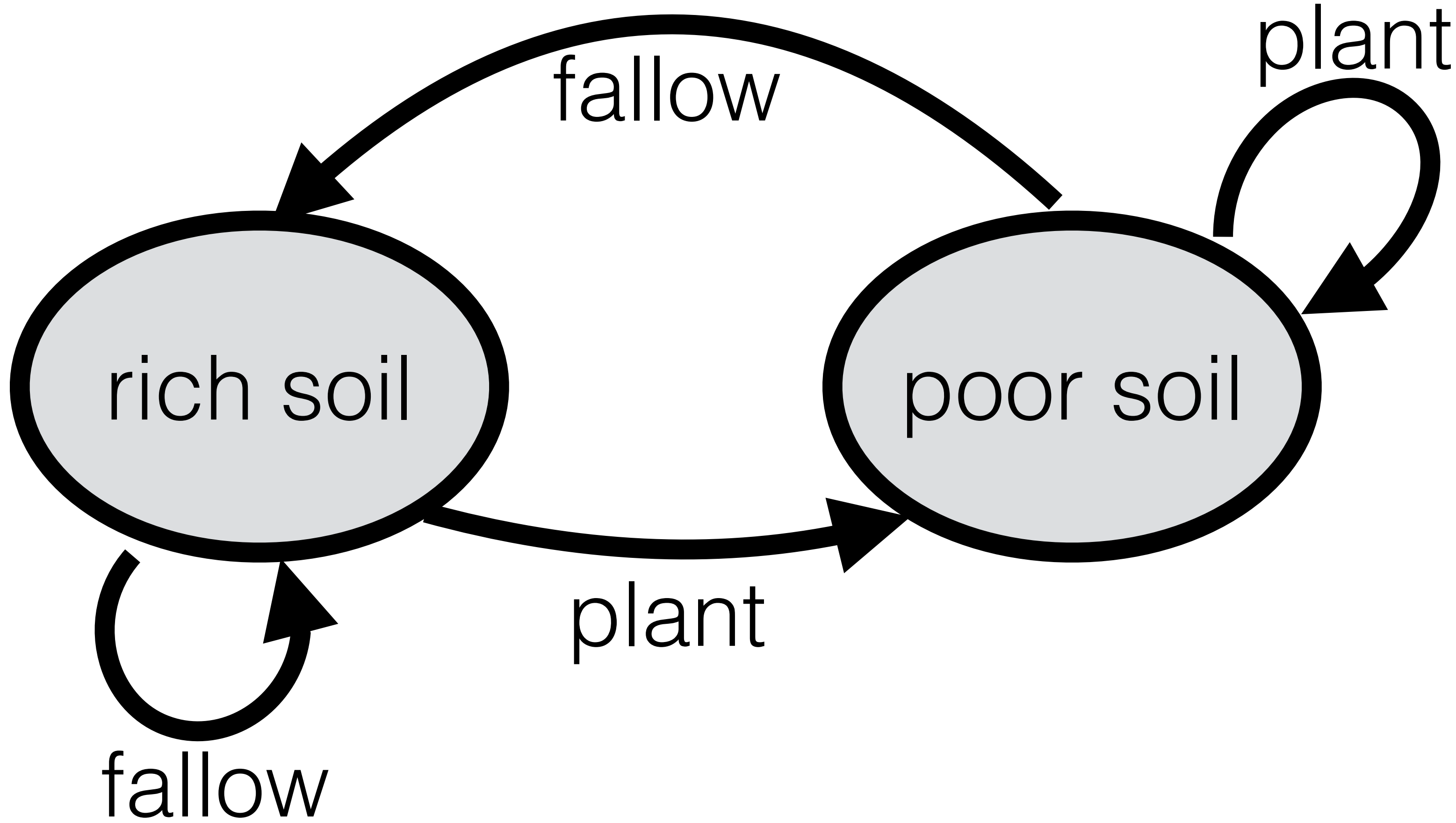
$s_0 = \text{rich}$

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

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

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

$s_0 = \text{rich}$

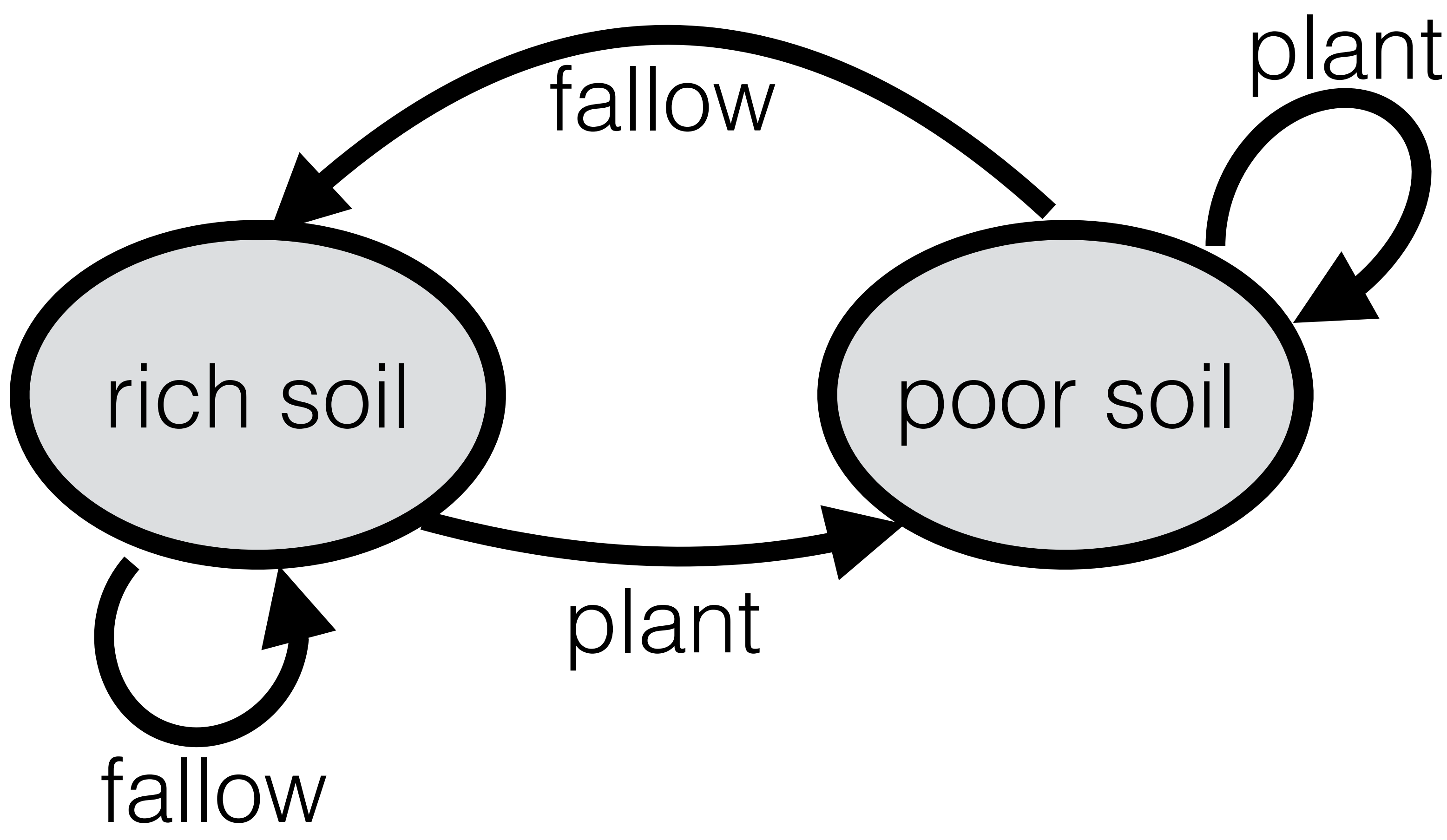
$s_1 = f(s_0, \text{plant}) = \text{poor};$

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

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

State Machine

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

$s_0 = \text{rich}$

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

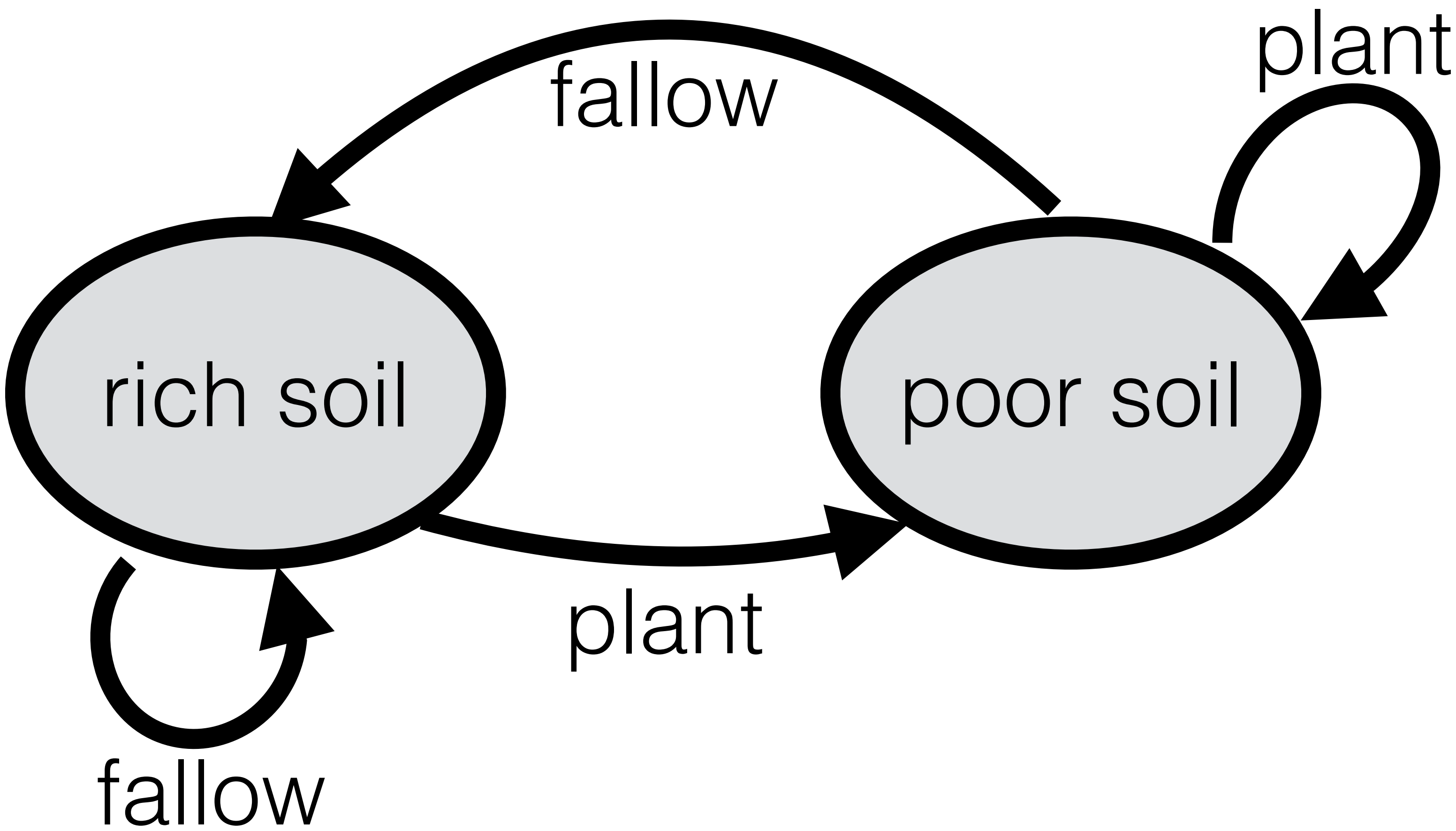
$y_1 = g(s_1) = \text{poor}$

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

$y_2 = g(s_2) = \text{rich}$

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)

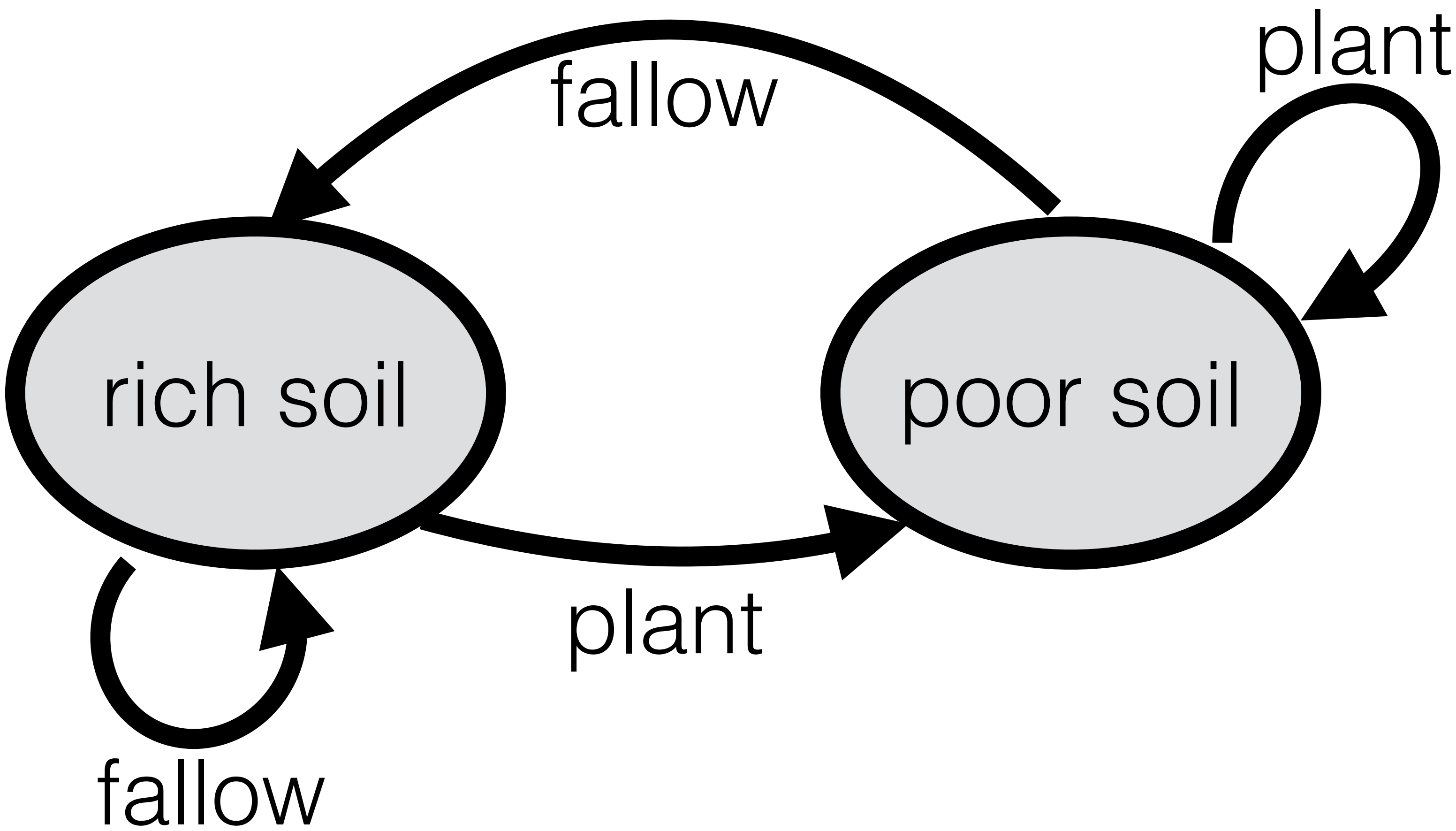


Example

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

State Machine

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

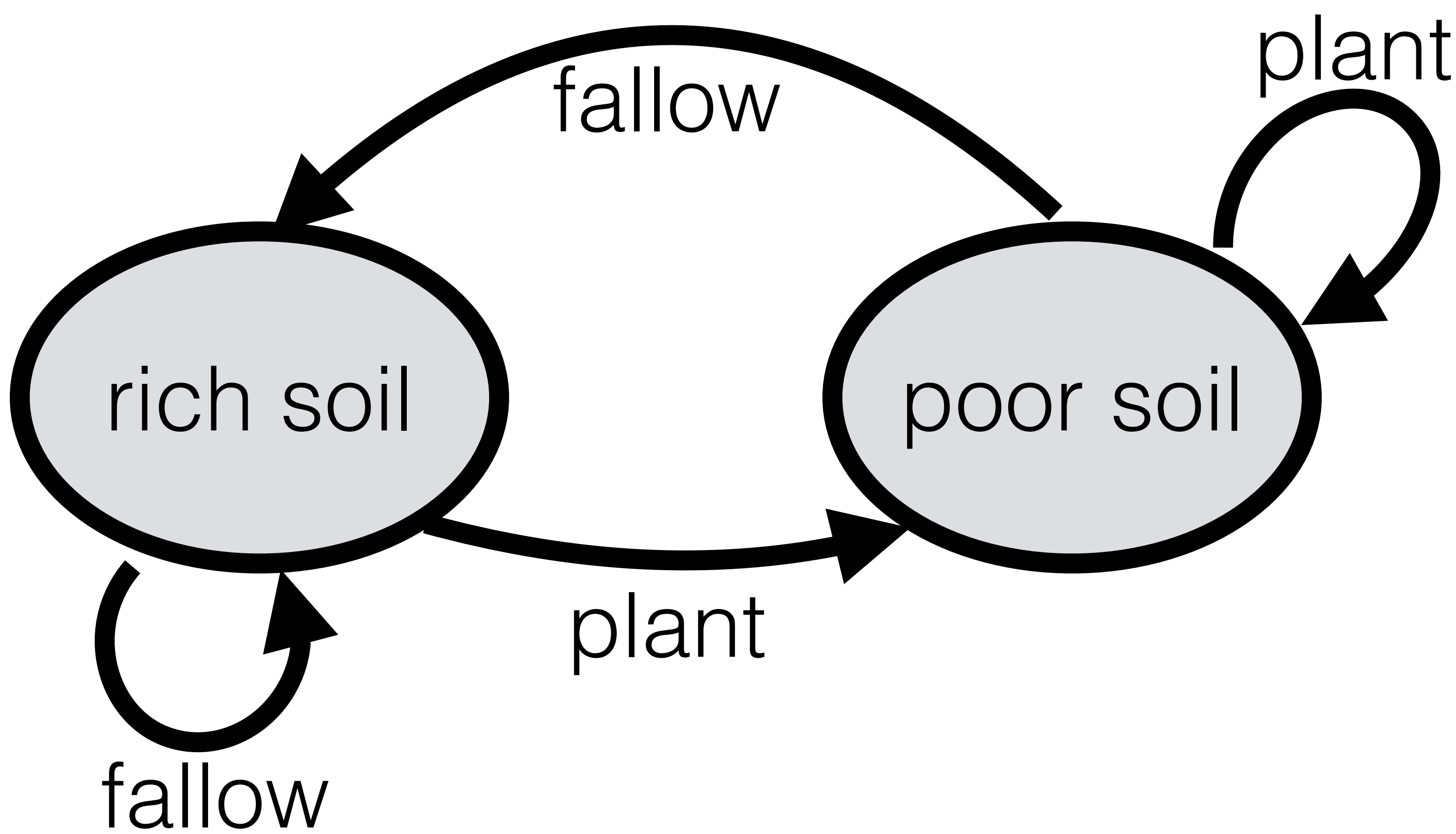


Example

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

State Machine

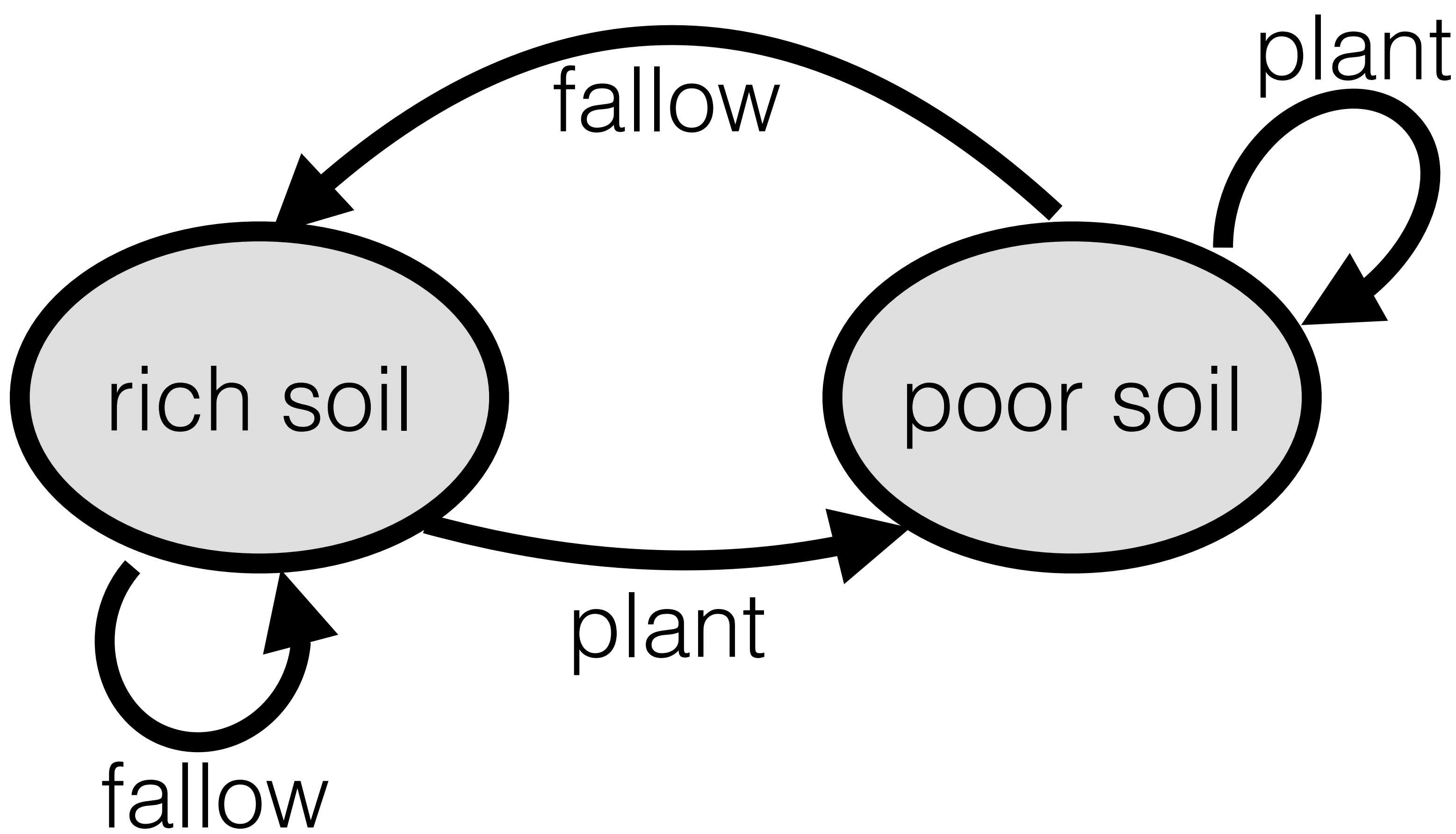
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$: transition function
- $\mathcal{Y}$: set of possible outputs
- $g : \mathcal{S} \rightarrow \mathcal{Y}$: output function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



Example

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

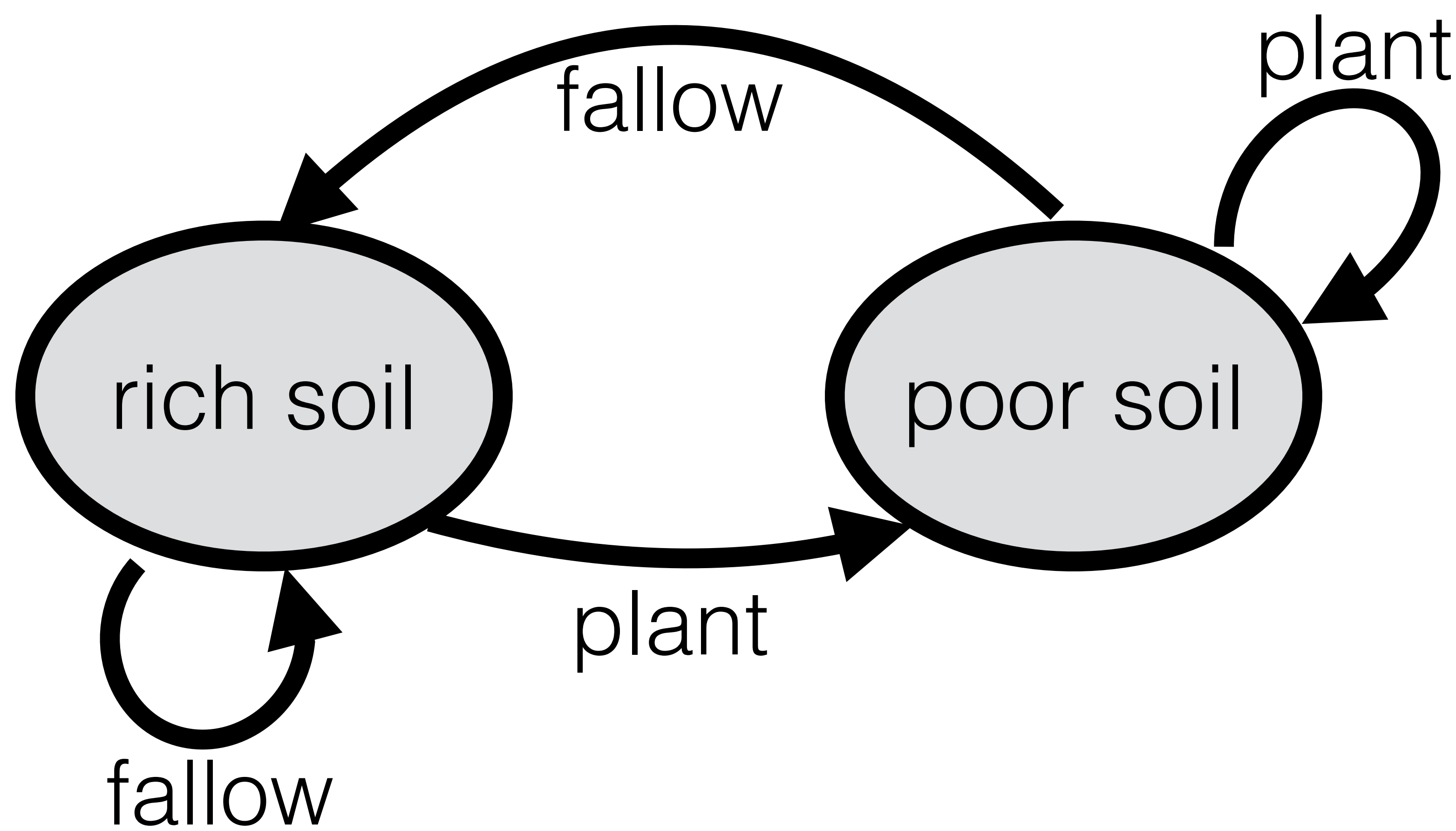
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function
- $\mathcal{Y}$: set of possible outputs
- $g : \mathcal{S} \to \mathcal{Y}$: output function
    - e.g.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



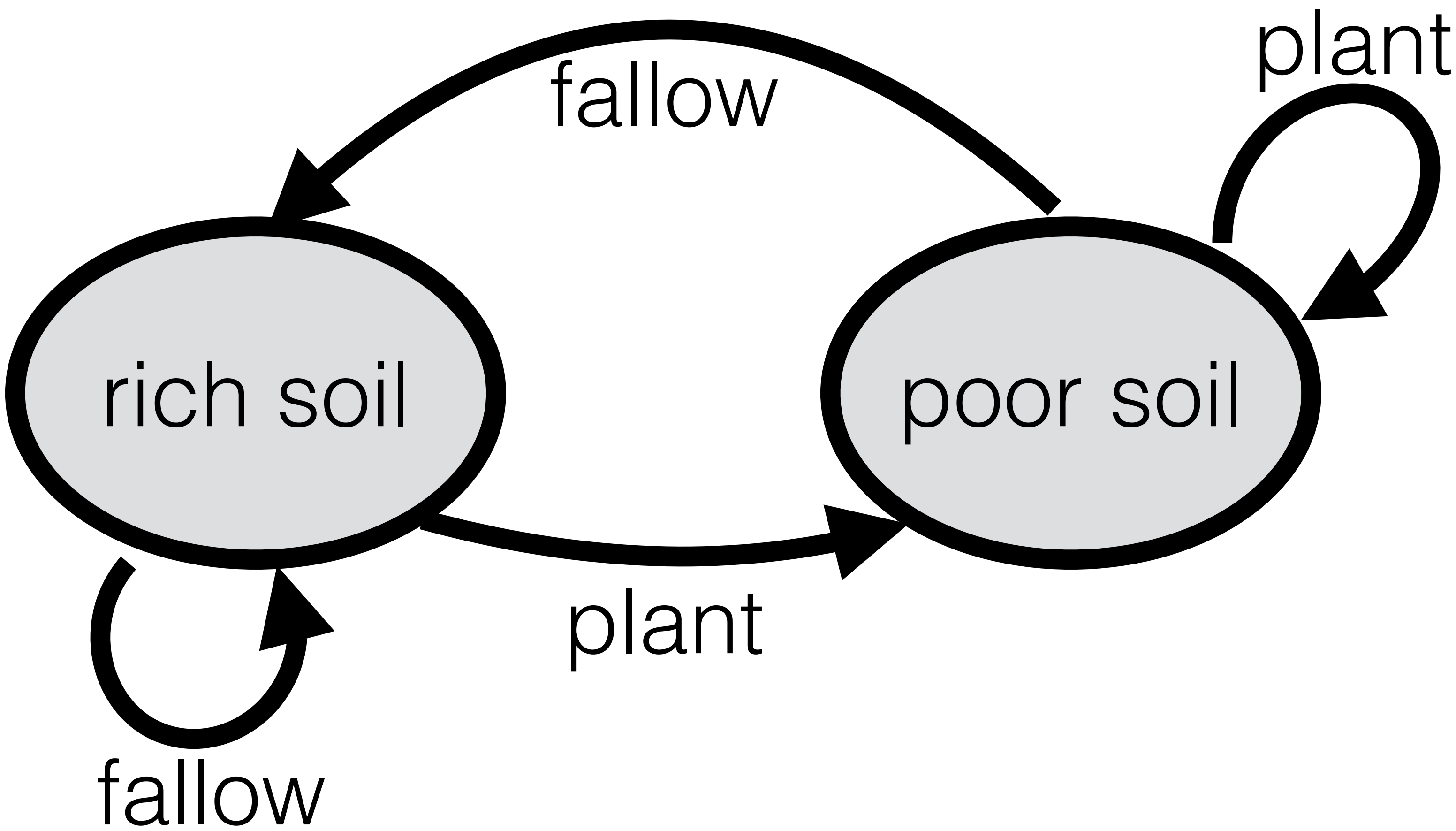
Example

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

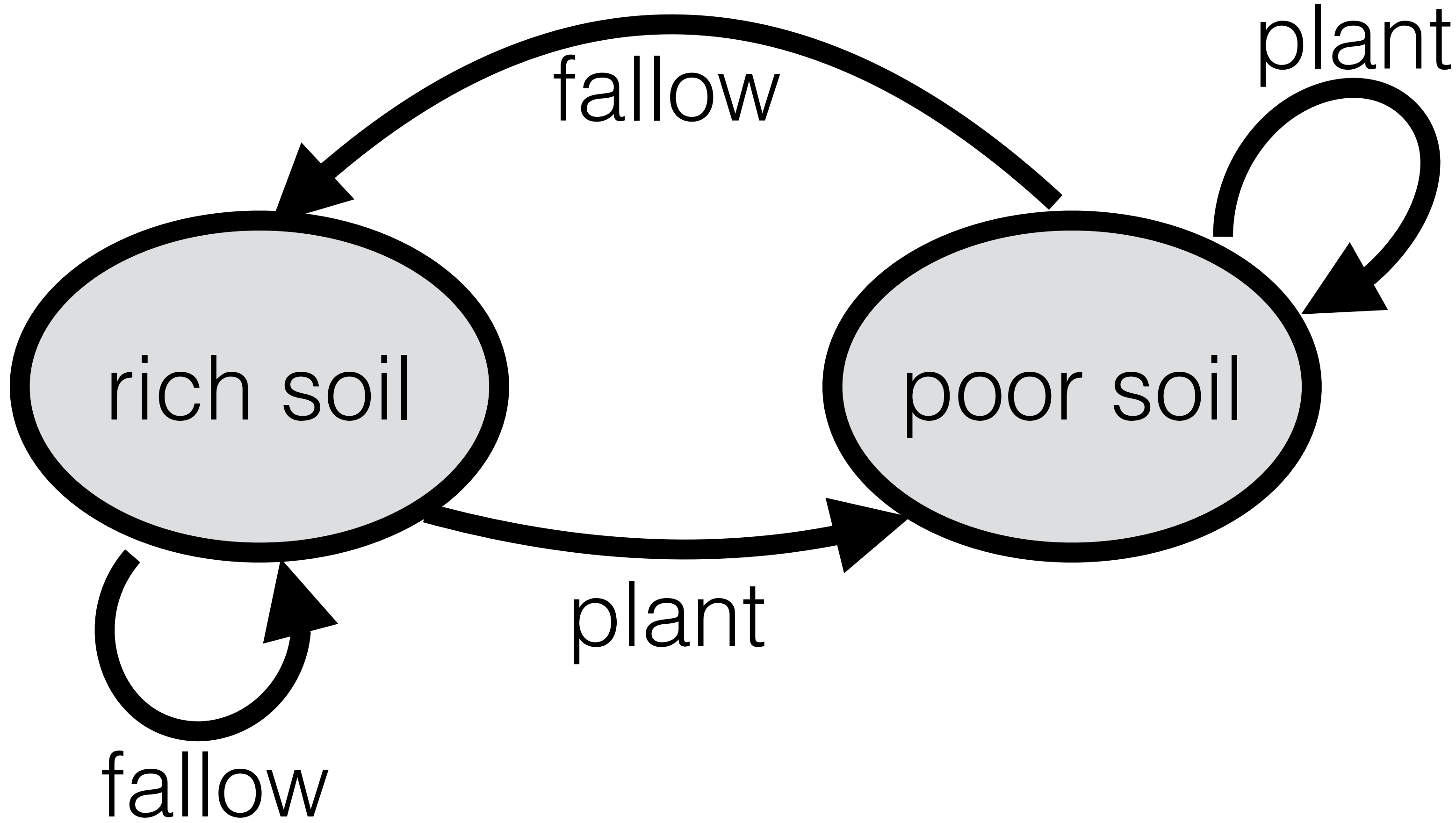
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function
- $\mathcal{Y}$: set of possible outputs
- $g : \mathcal{S} \to \mathcal{Y}$: output function
  - e.g. g(s)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



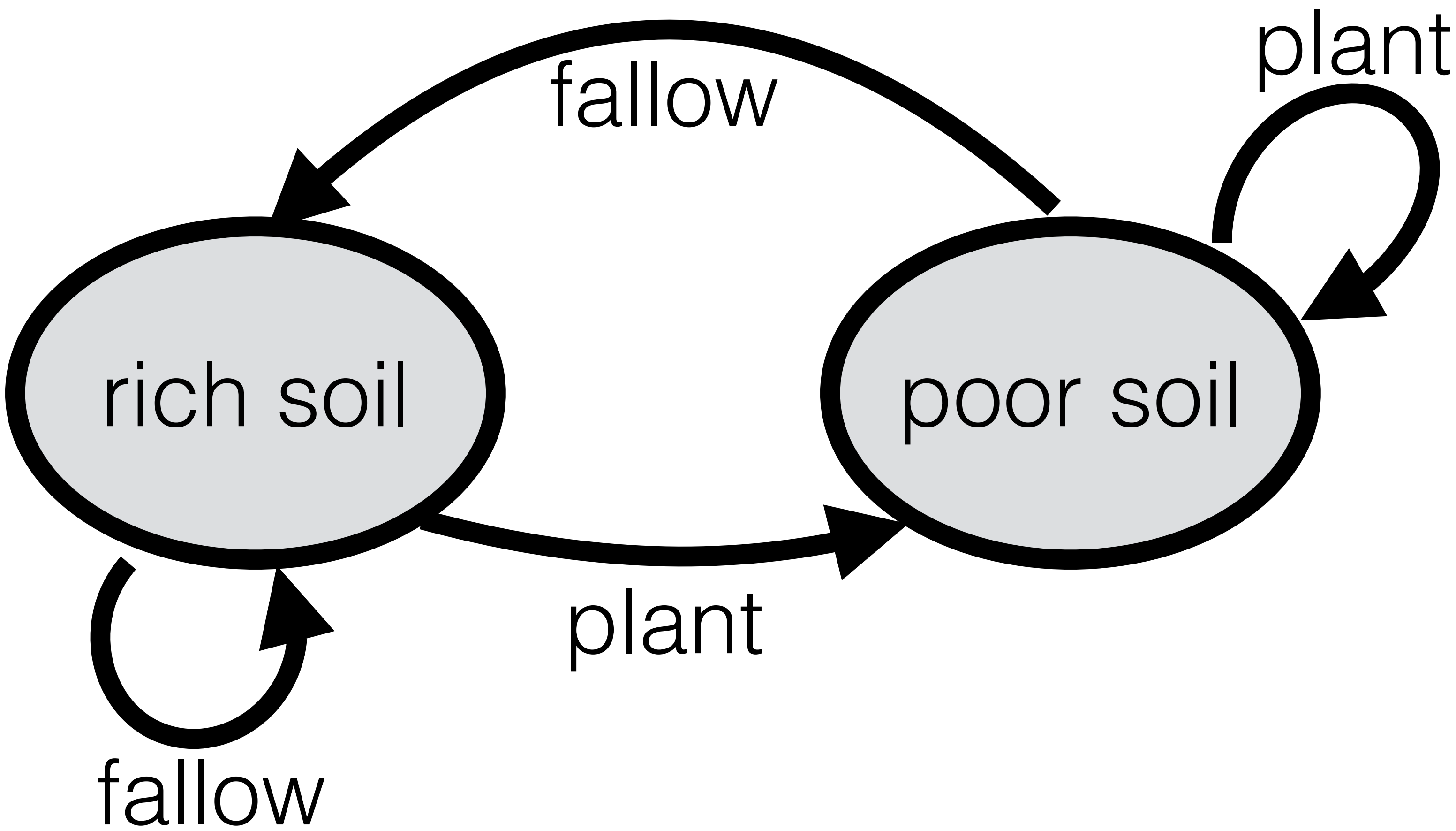
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function
- $\mathcal{Y}$: set of possible outputs
- $g : \mathcal{S} \to \mathcal{Y}$: output function
    - e.g.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



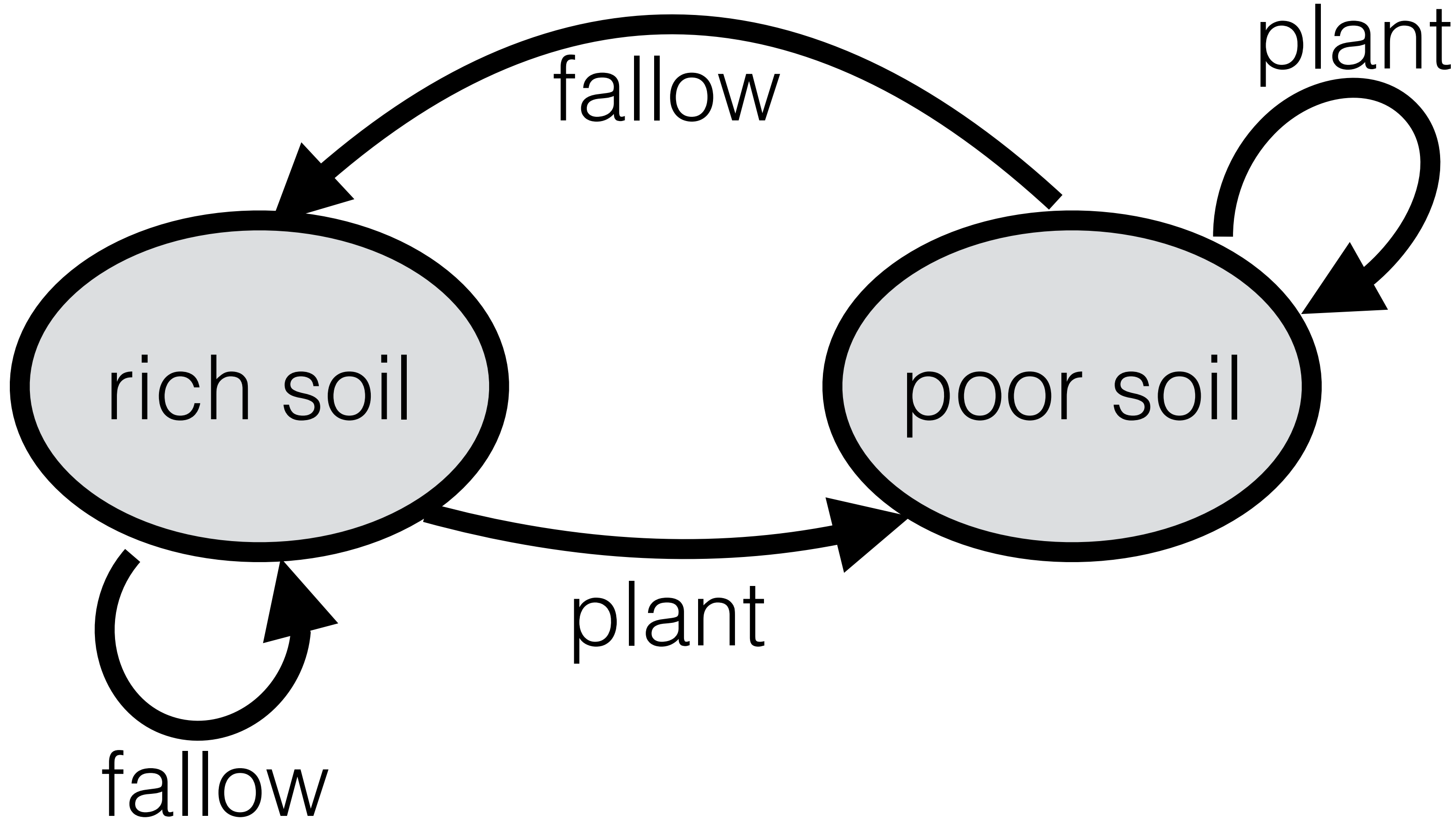
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function



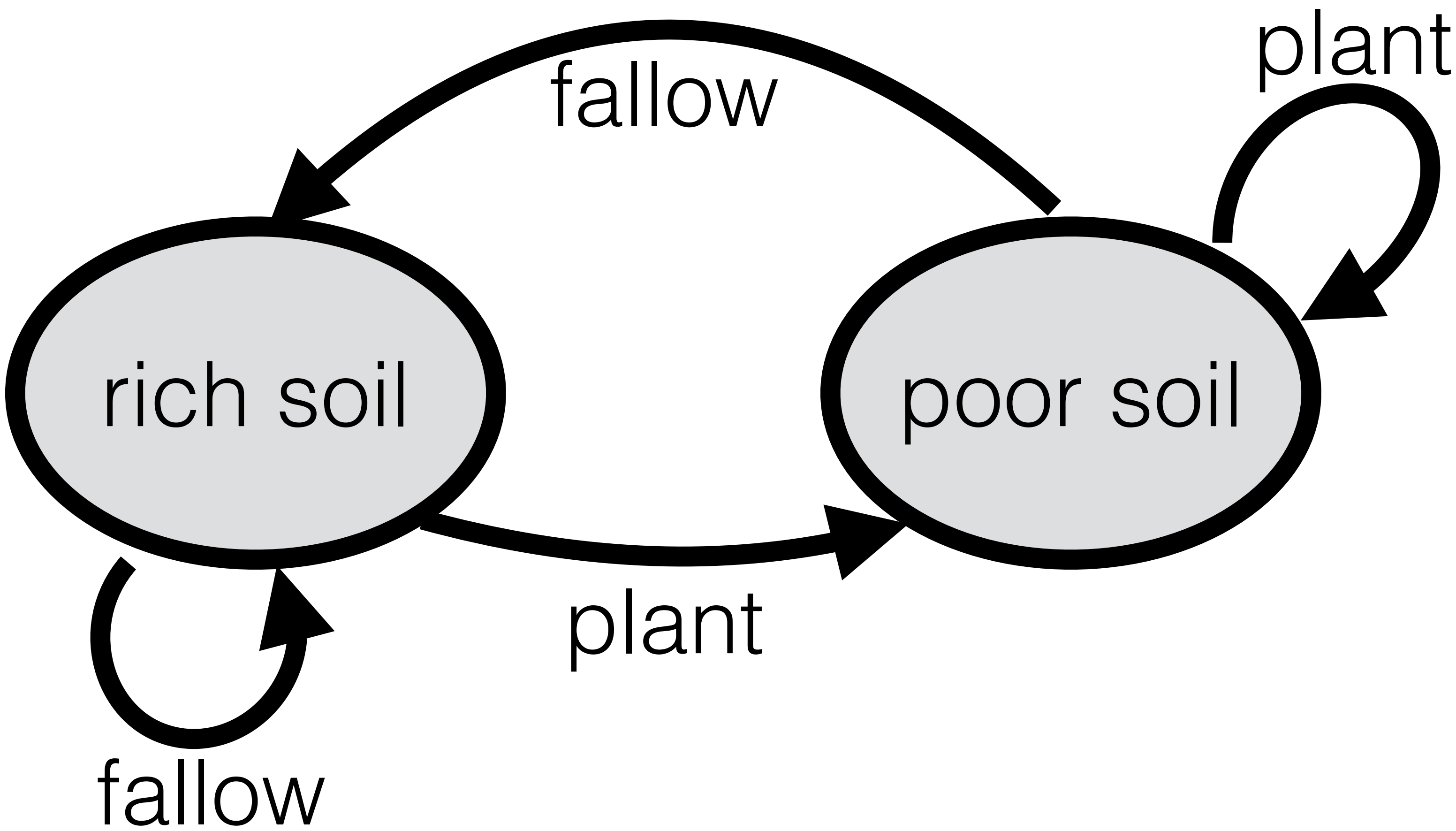
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function
- R reward function



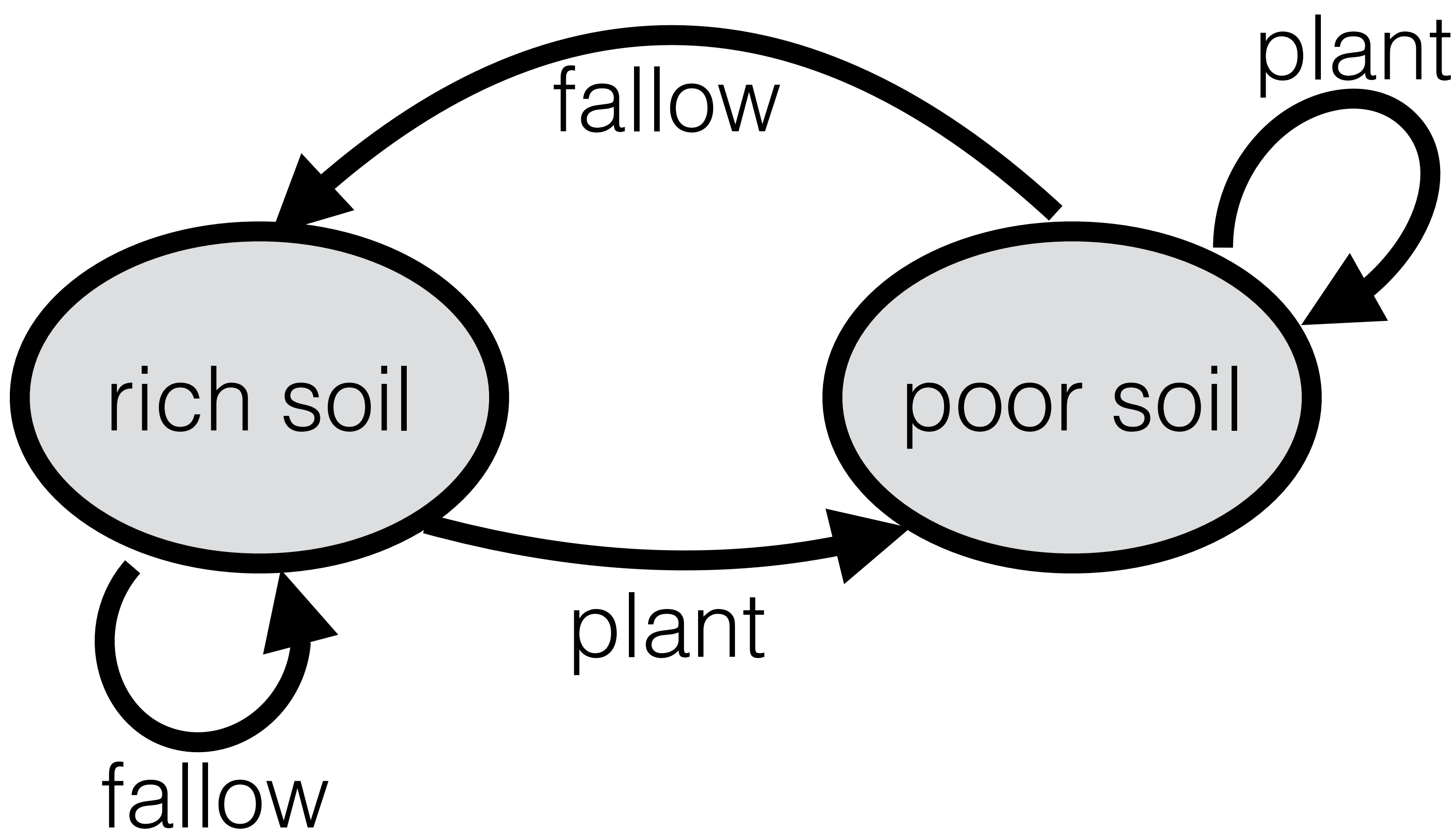
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$: transition function
- R reward function
  - e.g. # bushels in harvest



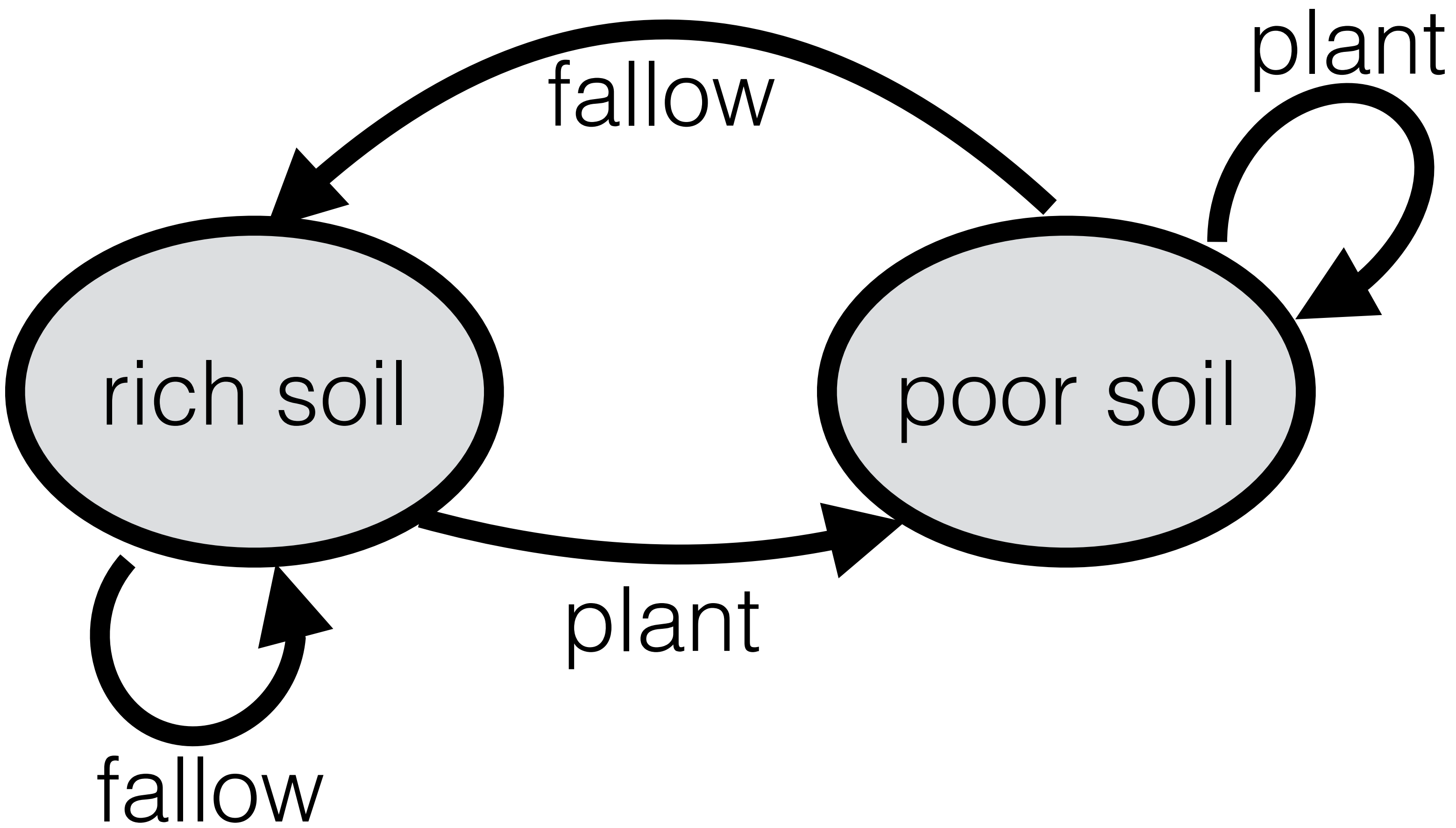
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function
- R reward function
    - e.g. # bushels in harvest



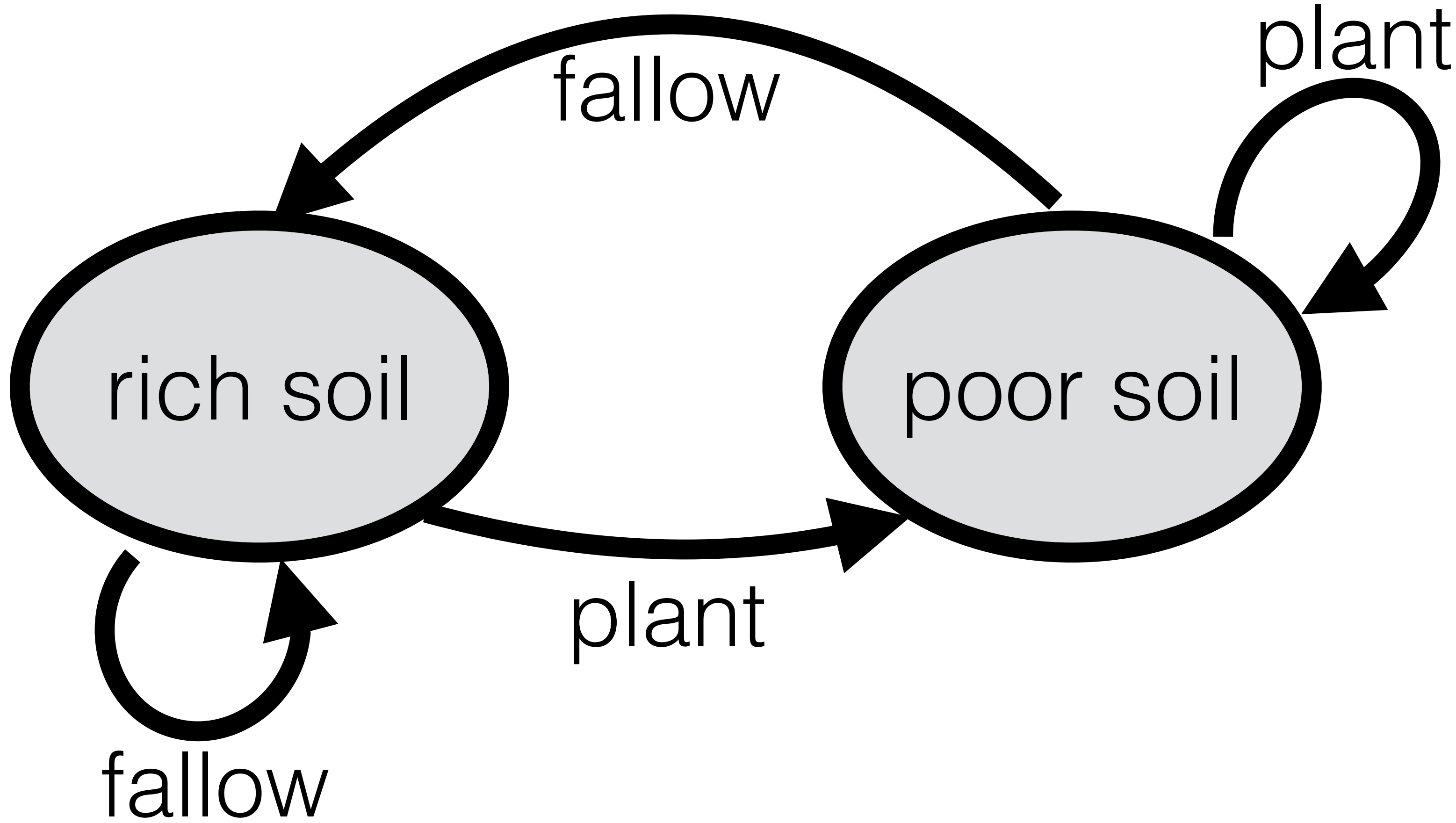
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$: transition function
- $R :$ $\rightarrow \mathbb{R}$ reward function
  - e.g. # bushels in harvest



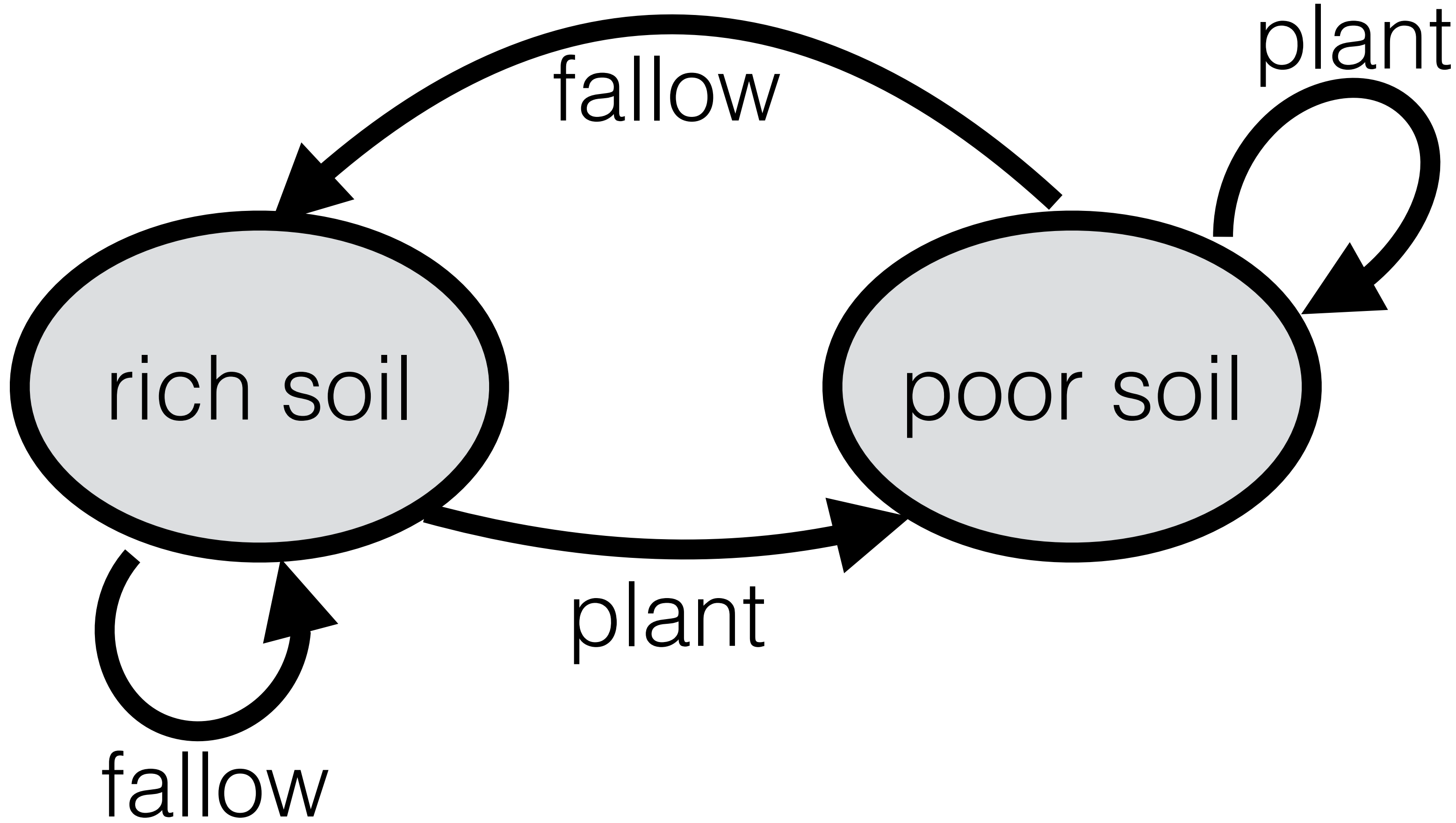
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function
- $R : \quad \mathcal{X} \to \mathbb{R}$ reward function
    - e.g. # bushels in harvest



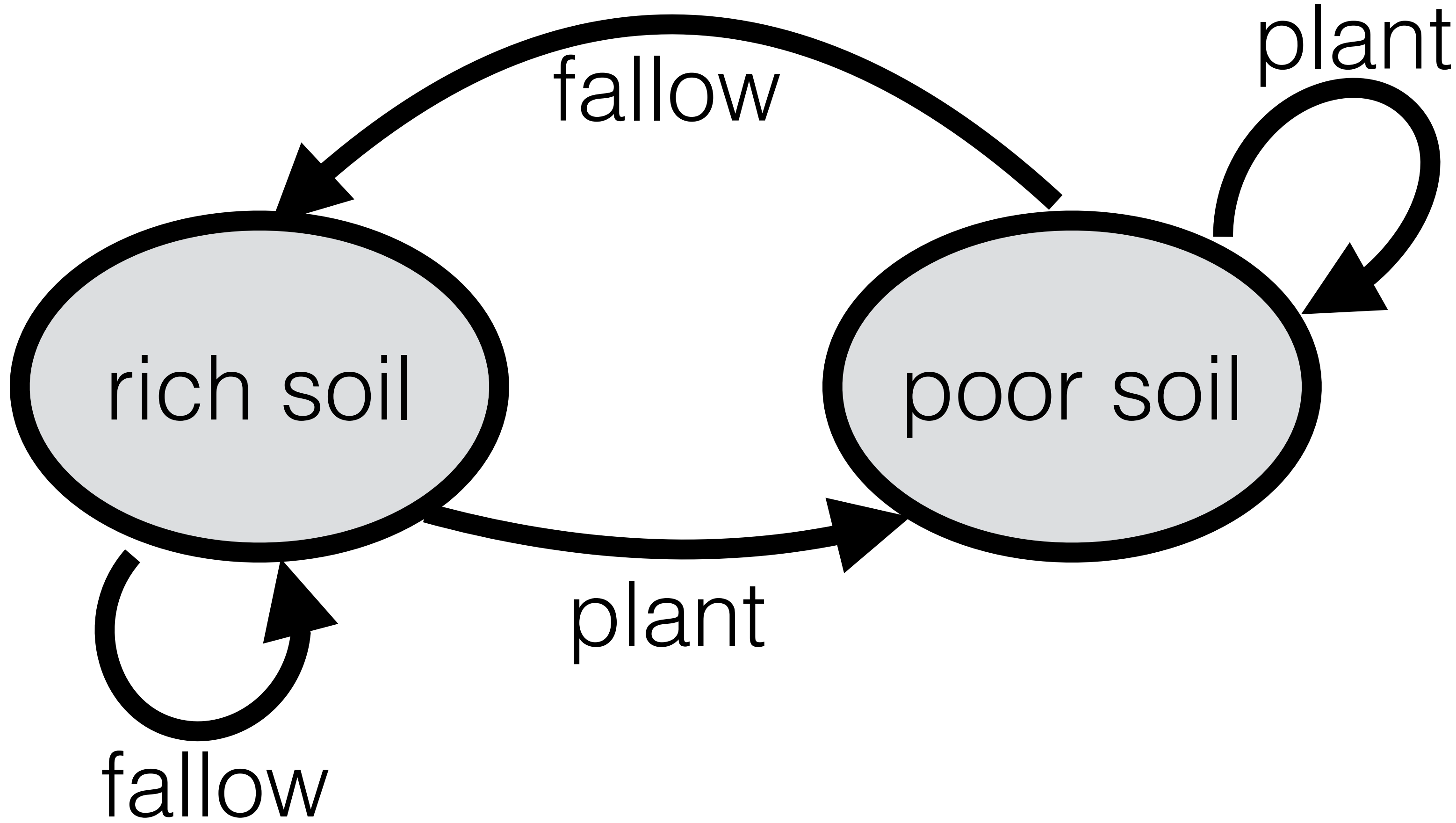
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. # bushels in harvest



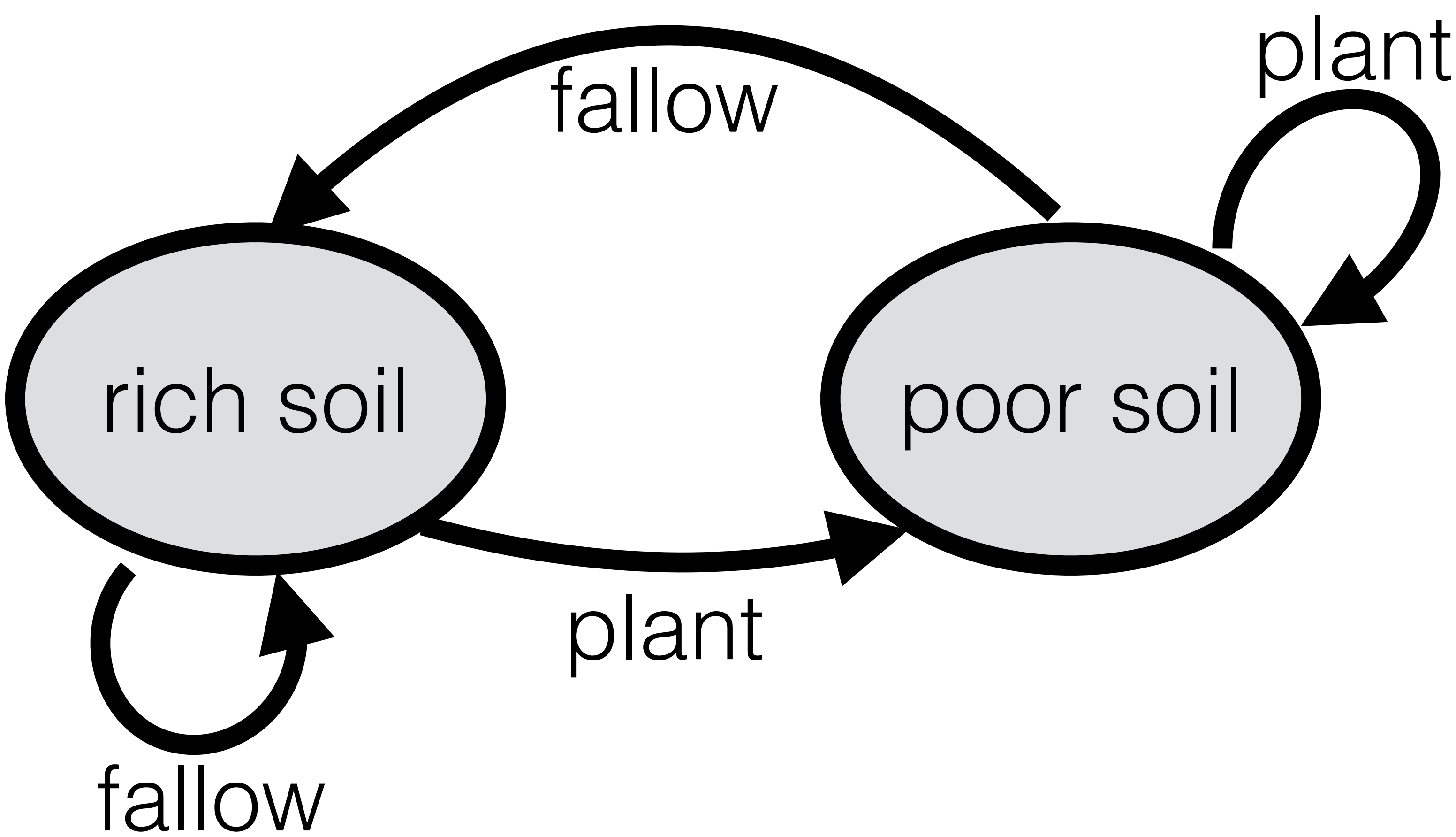
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. # bushels in harvest



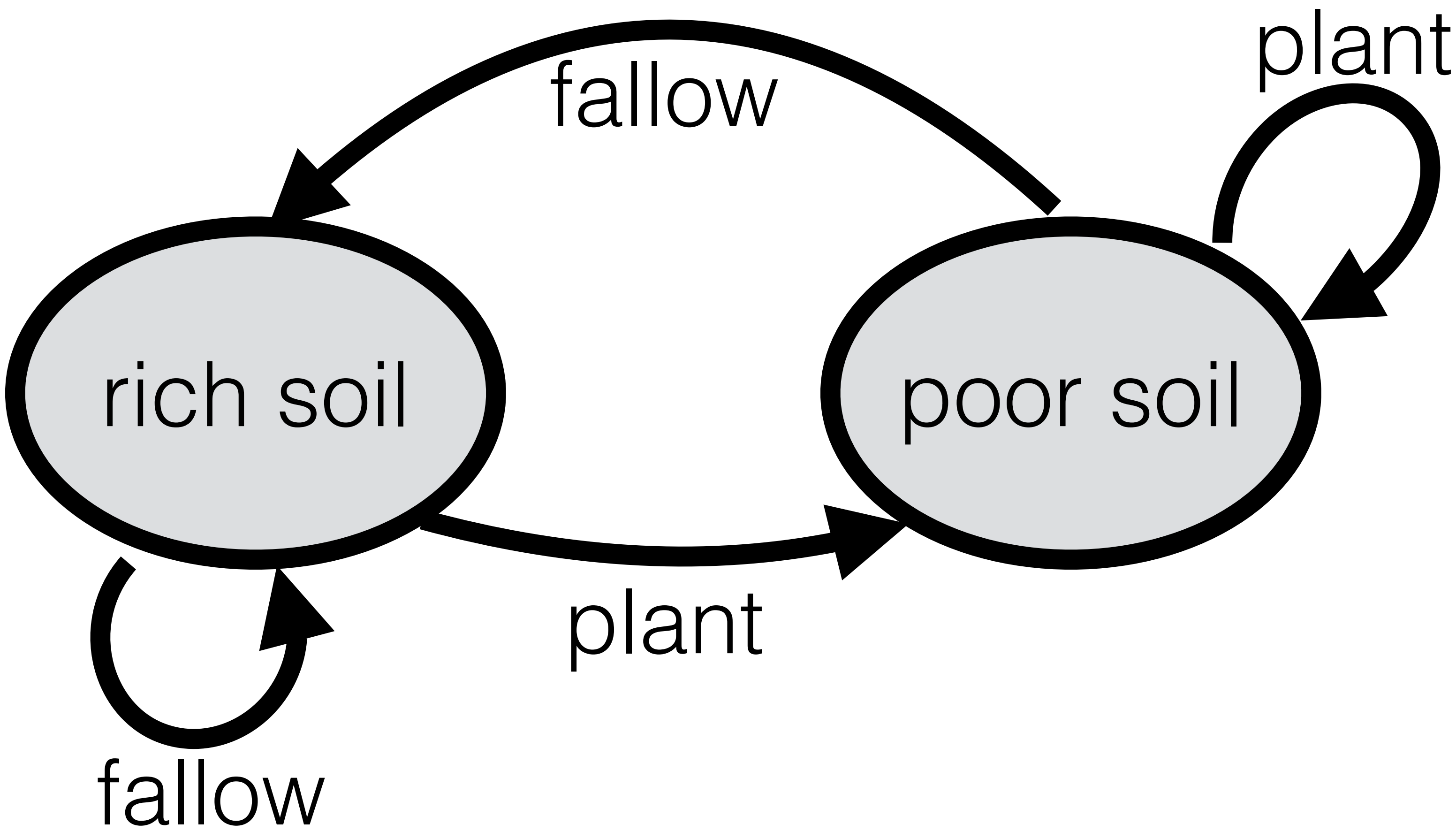
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels



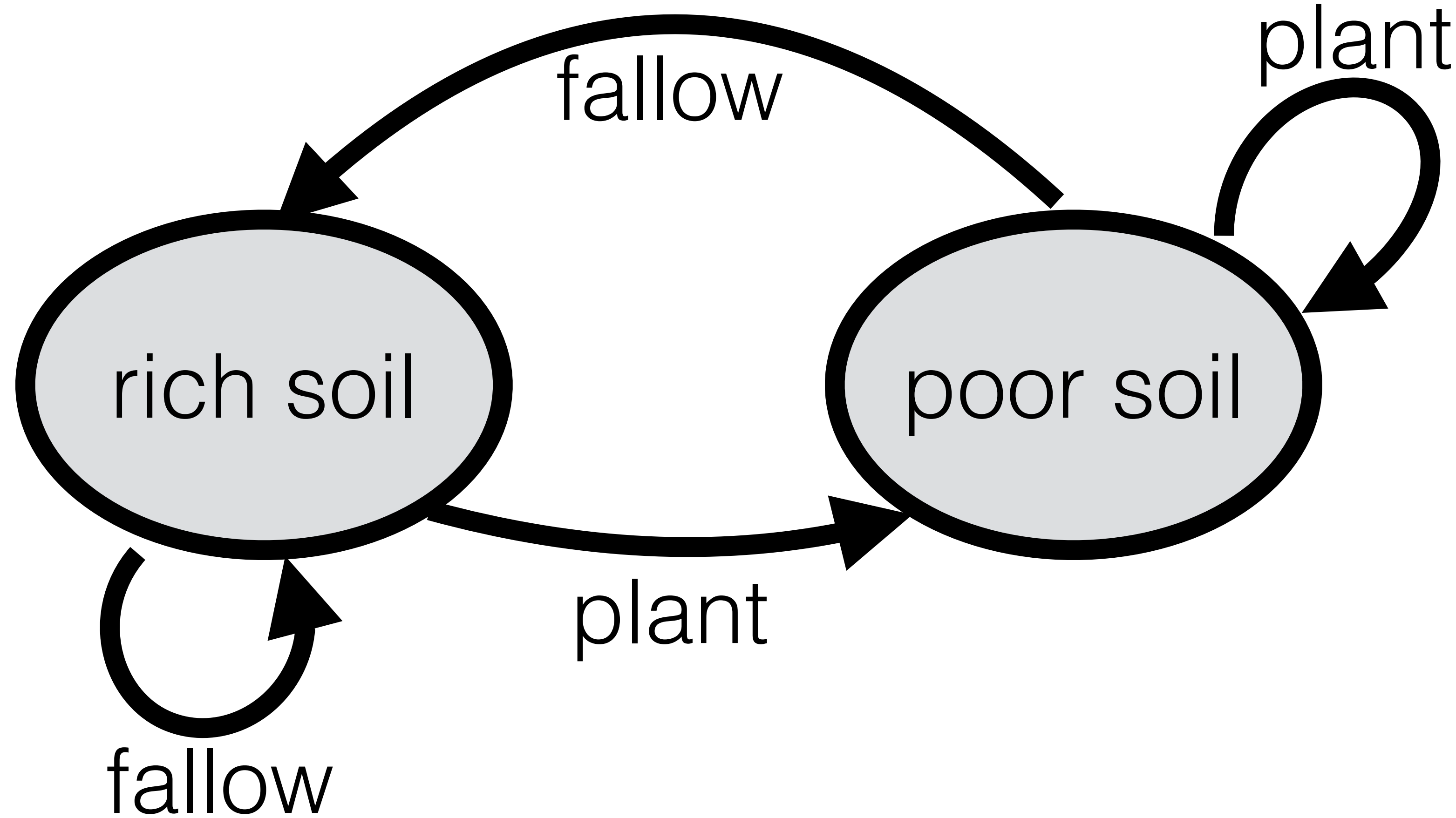
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels;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



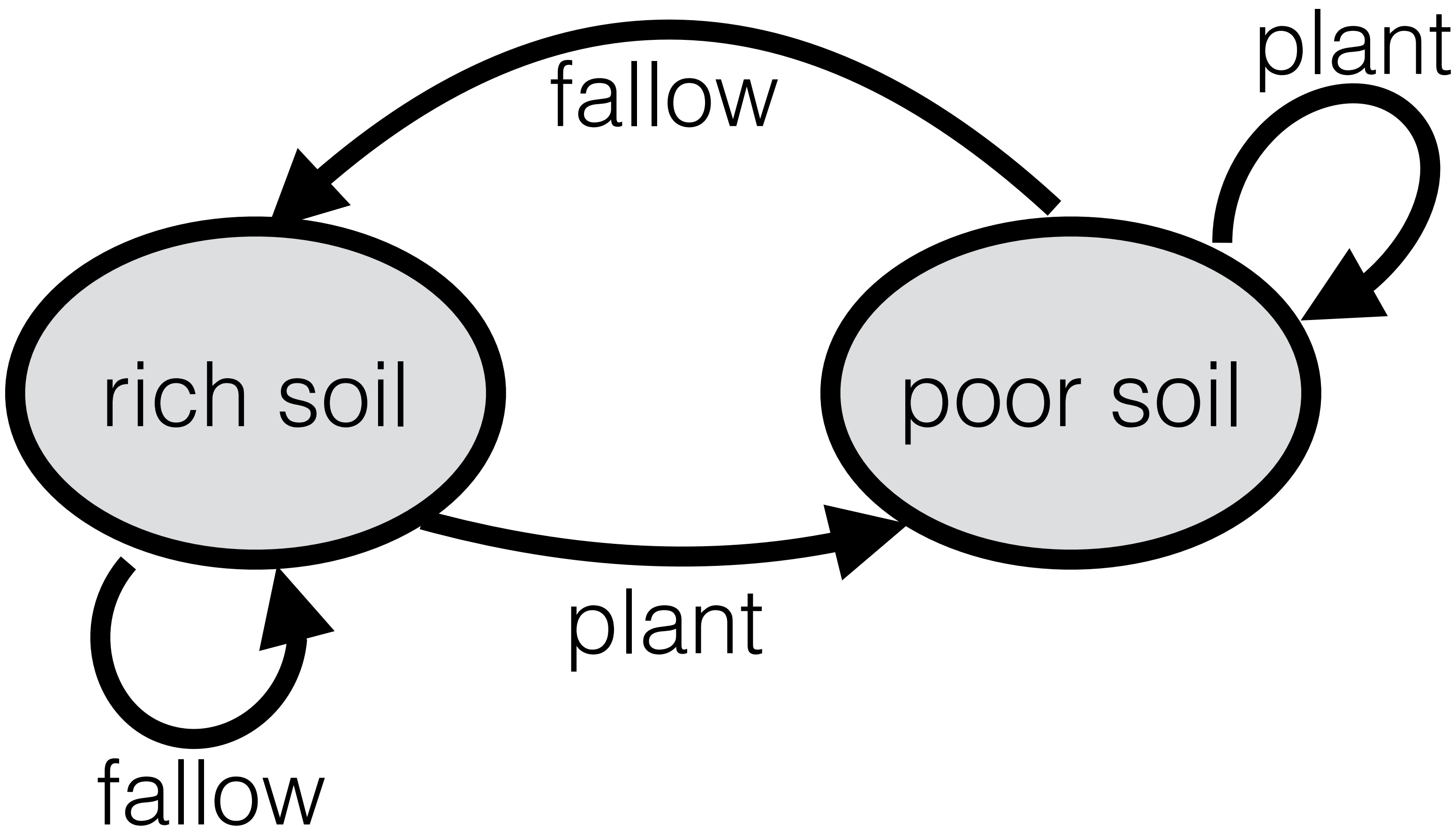
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels;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



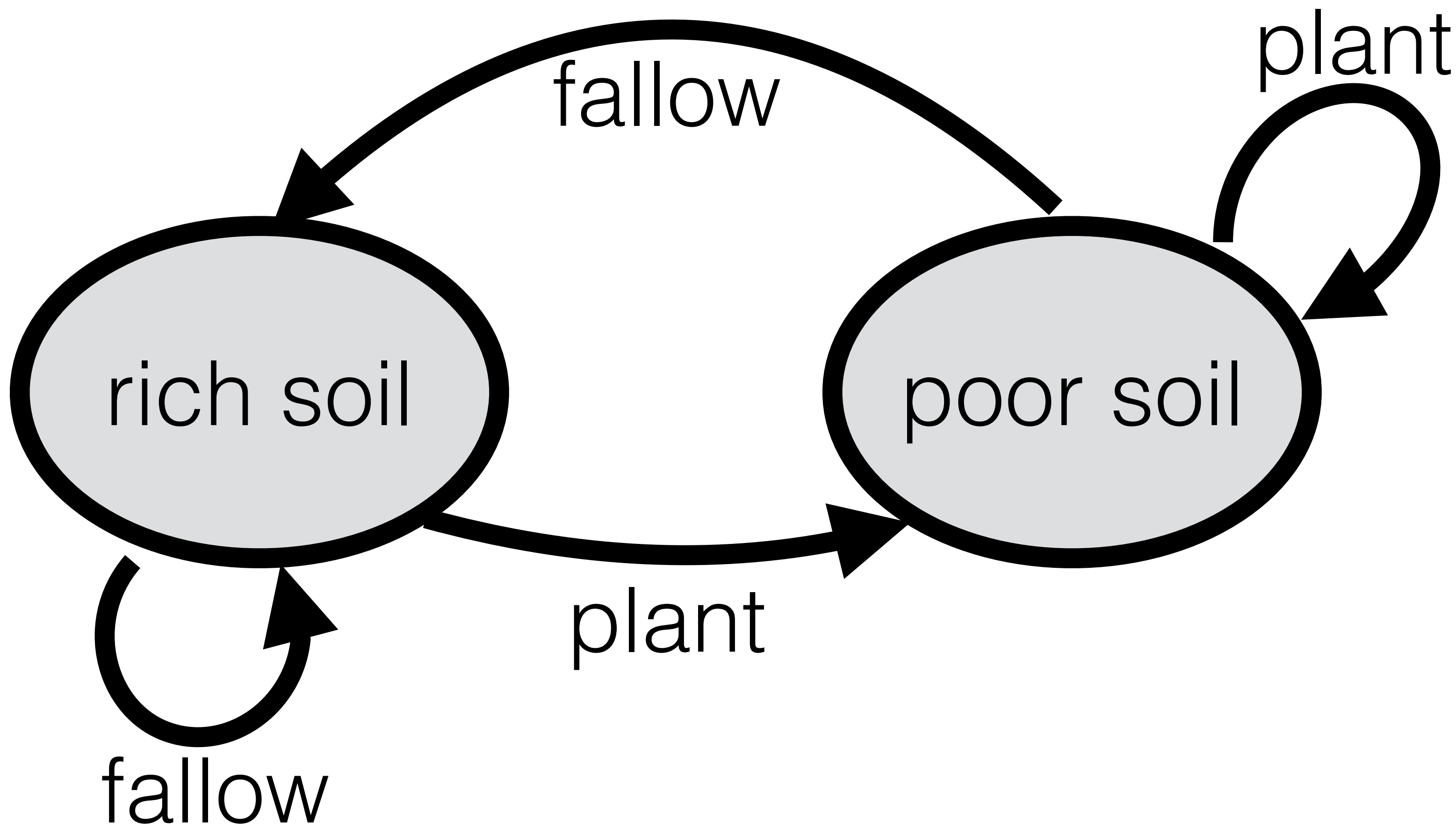
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels;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



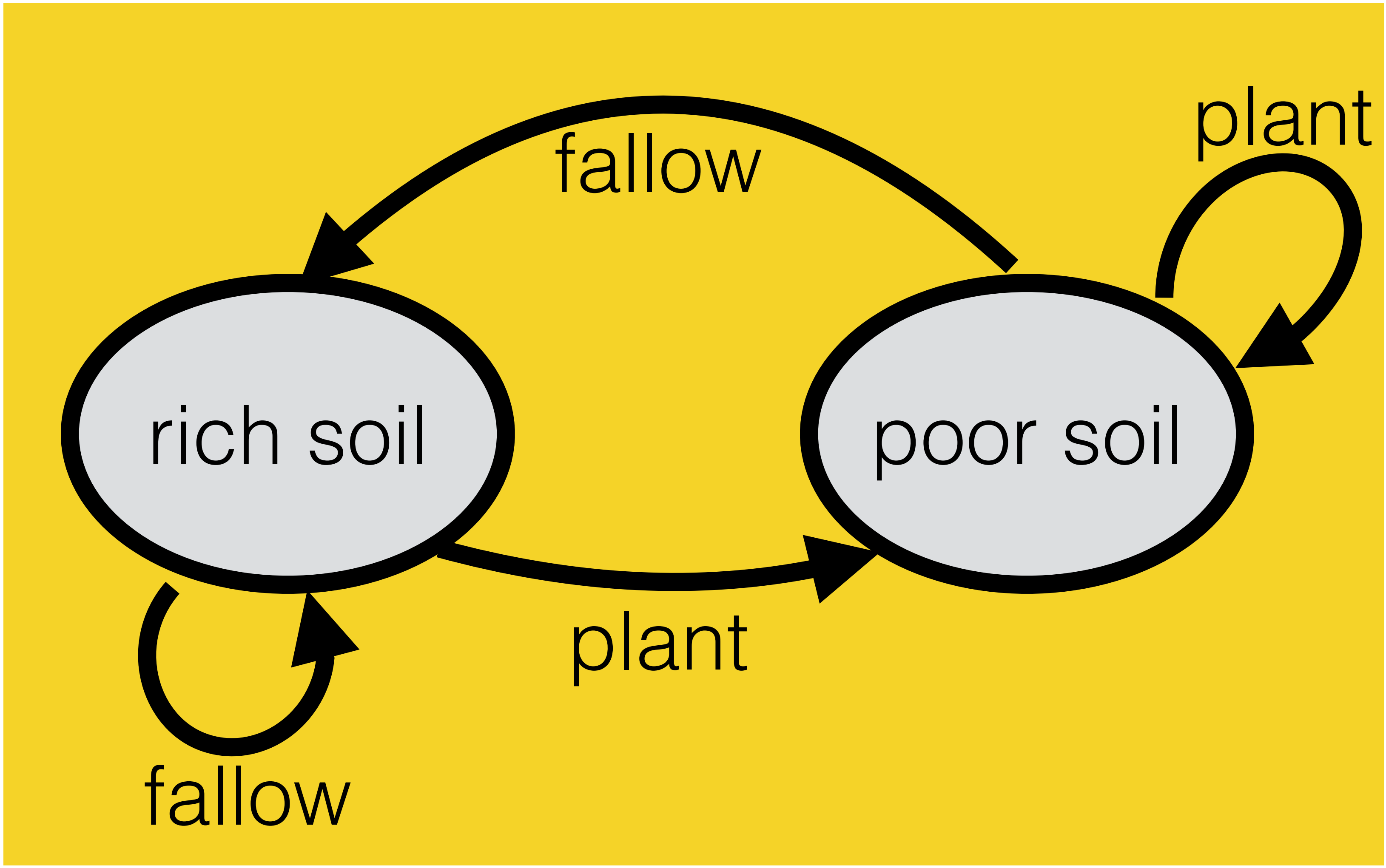
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$: transition function
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels;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



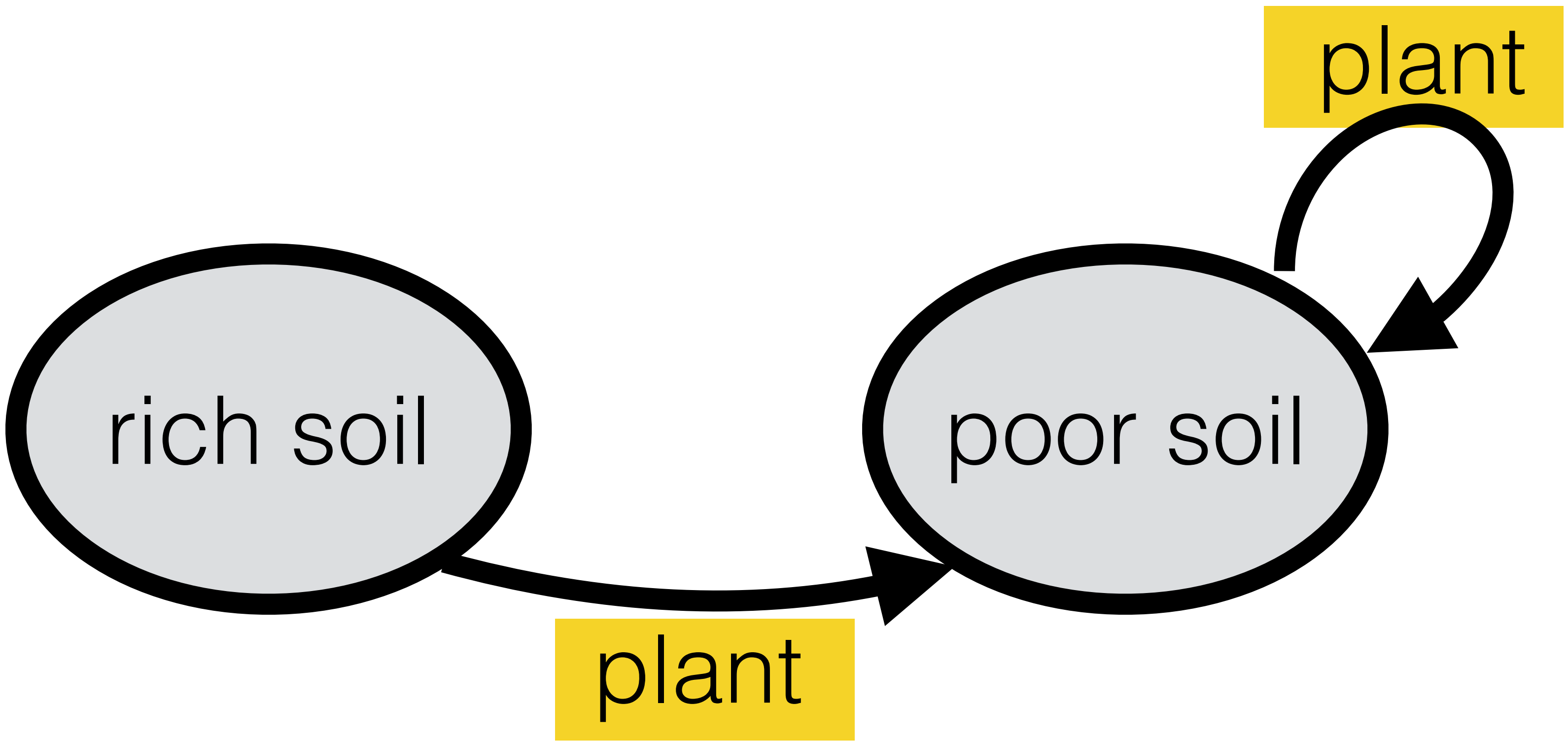
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T transition model
- $R : \mathcal{S} \times \mathcal{X} \rightarrow \mathbb{R}$: reward function
    - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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



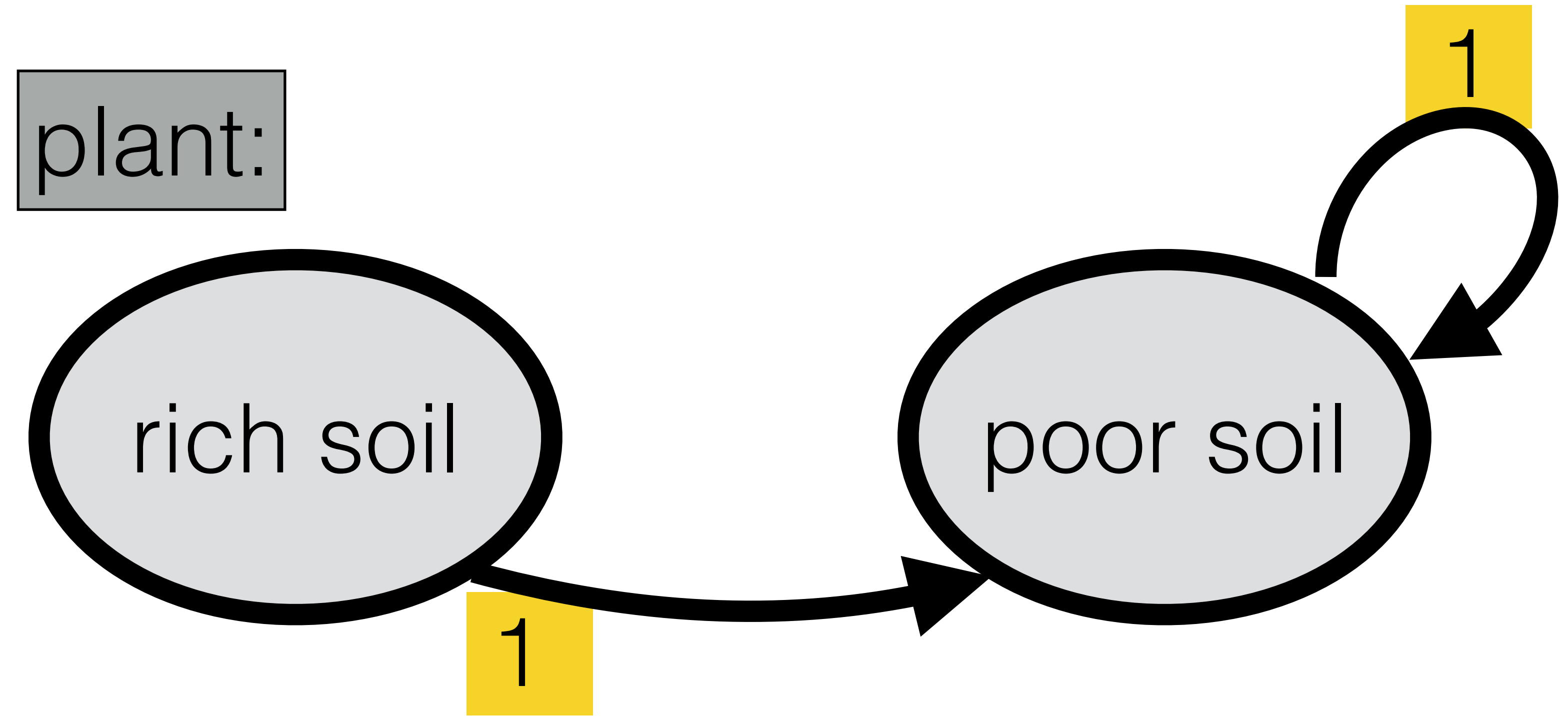
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) = R(poor,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



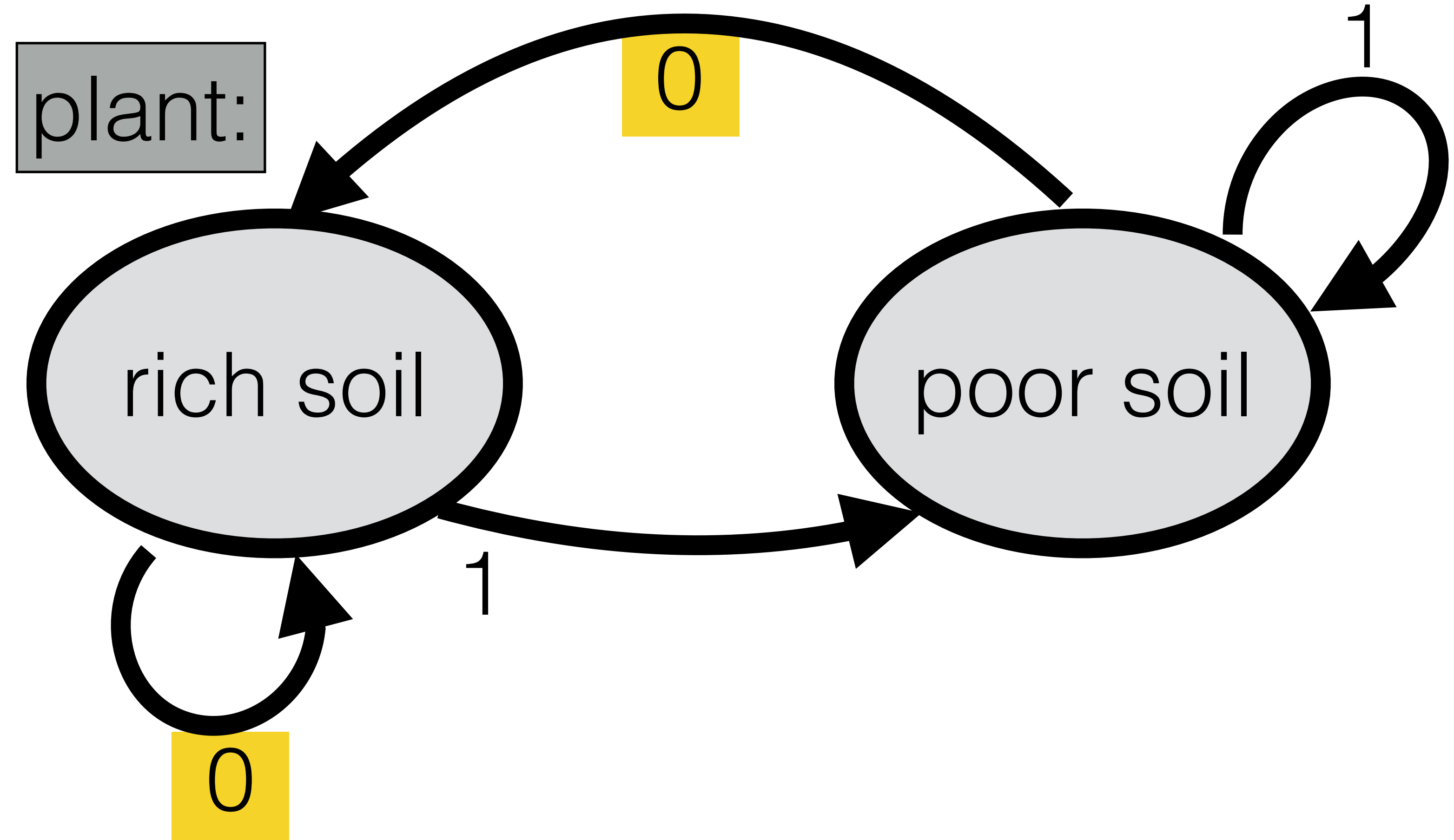
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T
  transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$:
  reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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



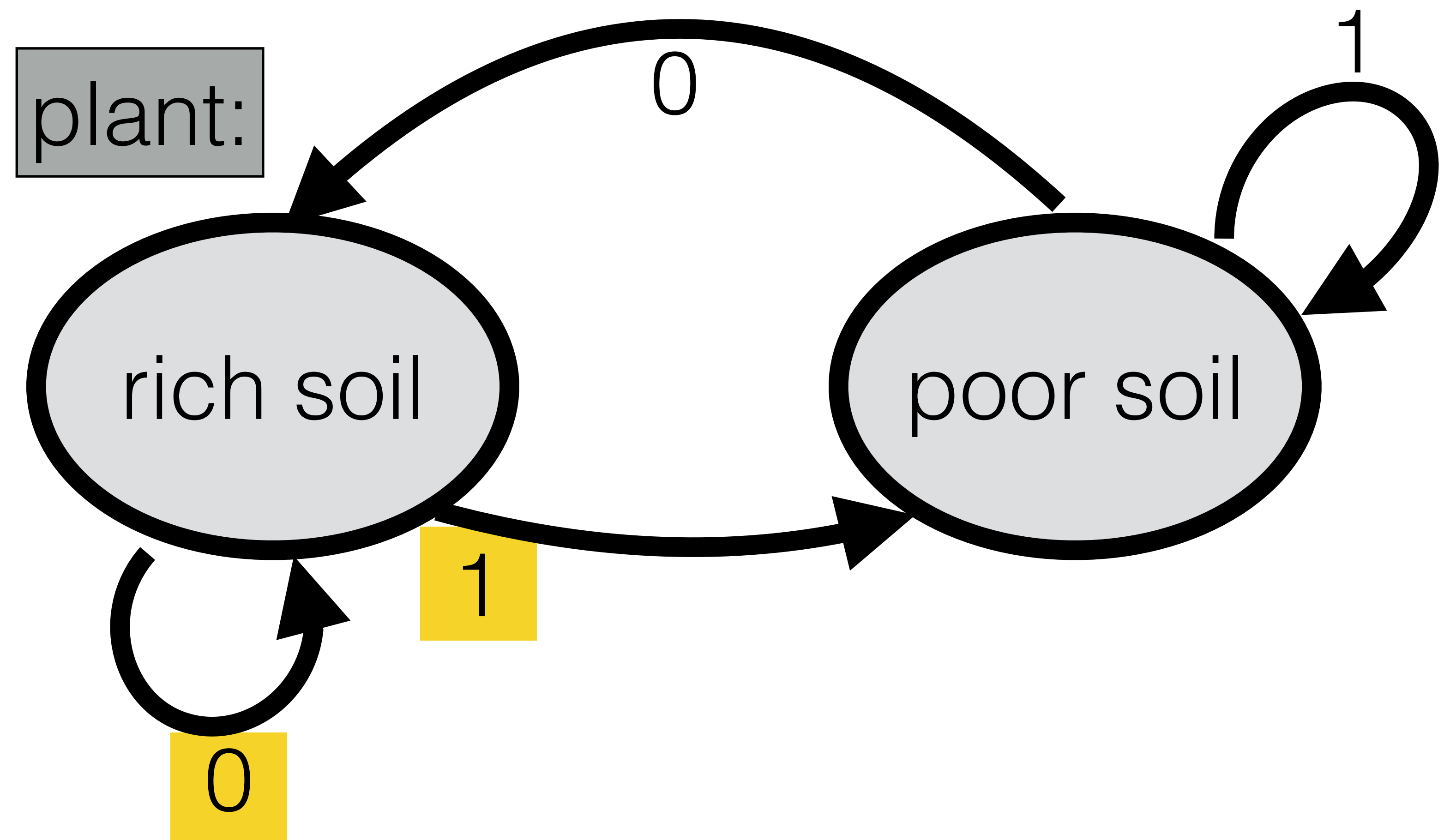
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T transition model
- $R : \mathcal{S} \times \mathcal{X} \rightarrow \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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



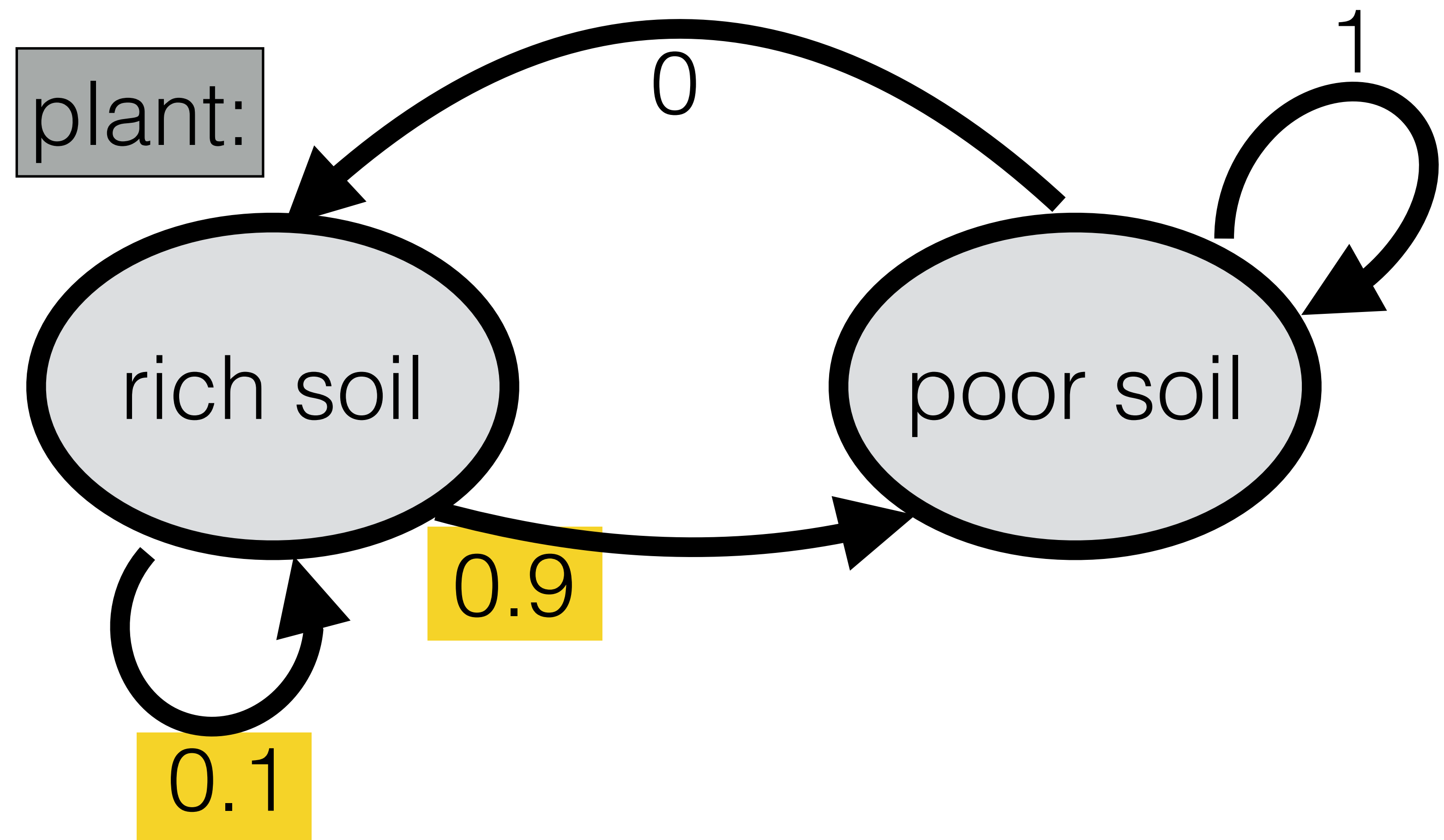
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T
  transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$:
  reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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



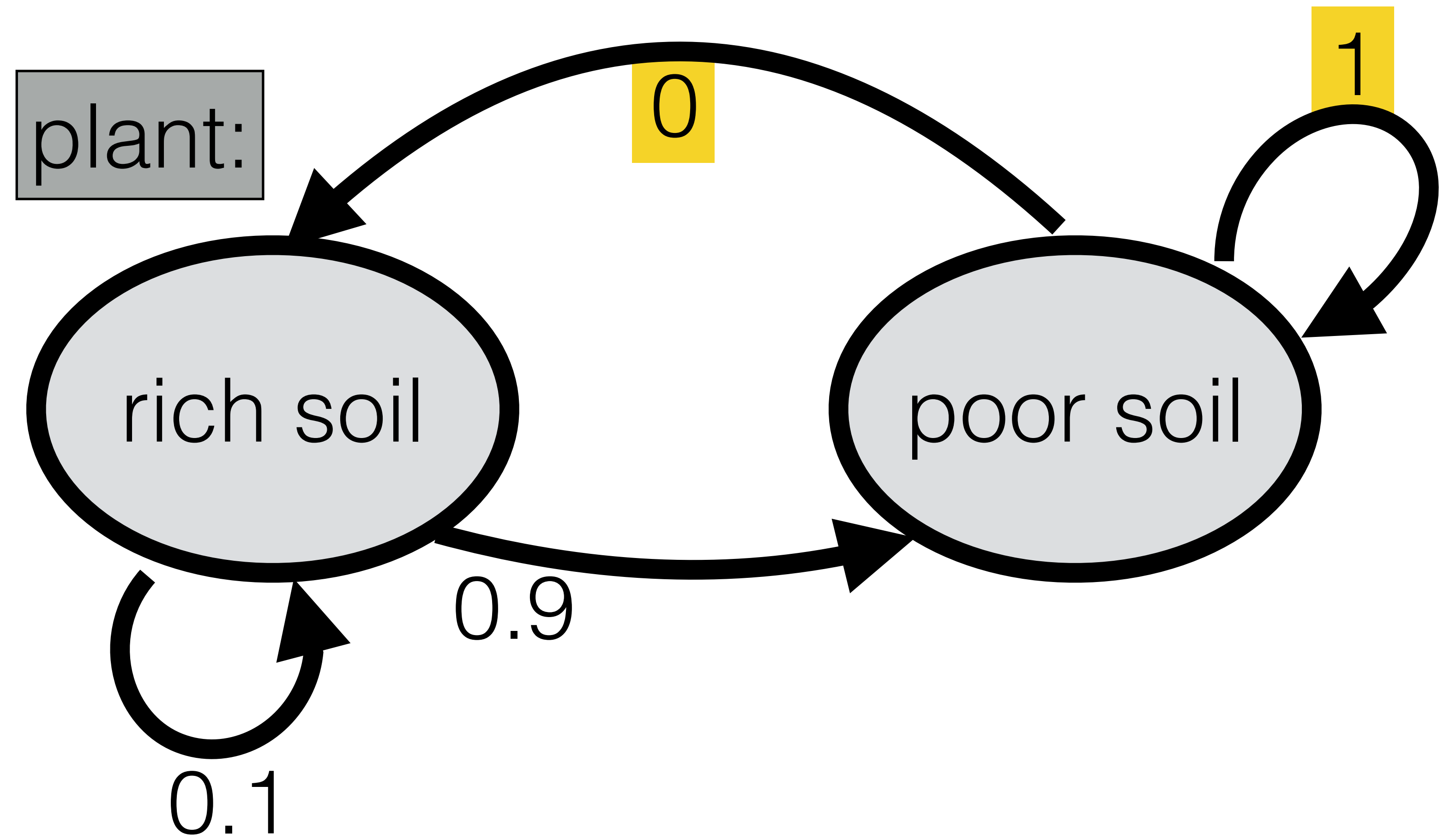
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) = R(poor,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



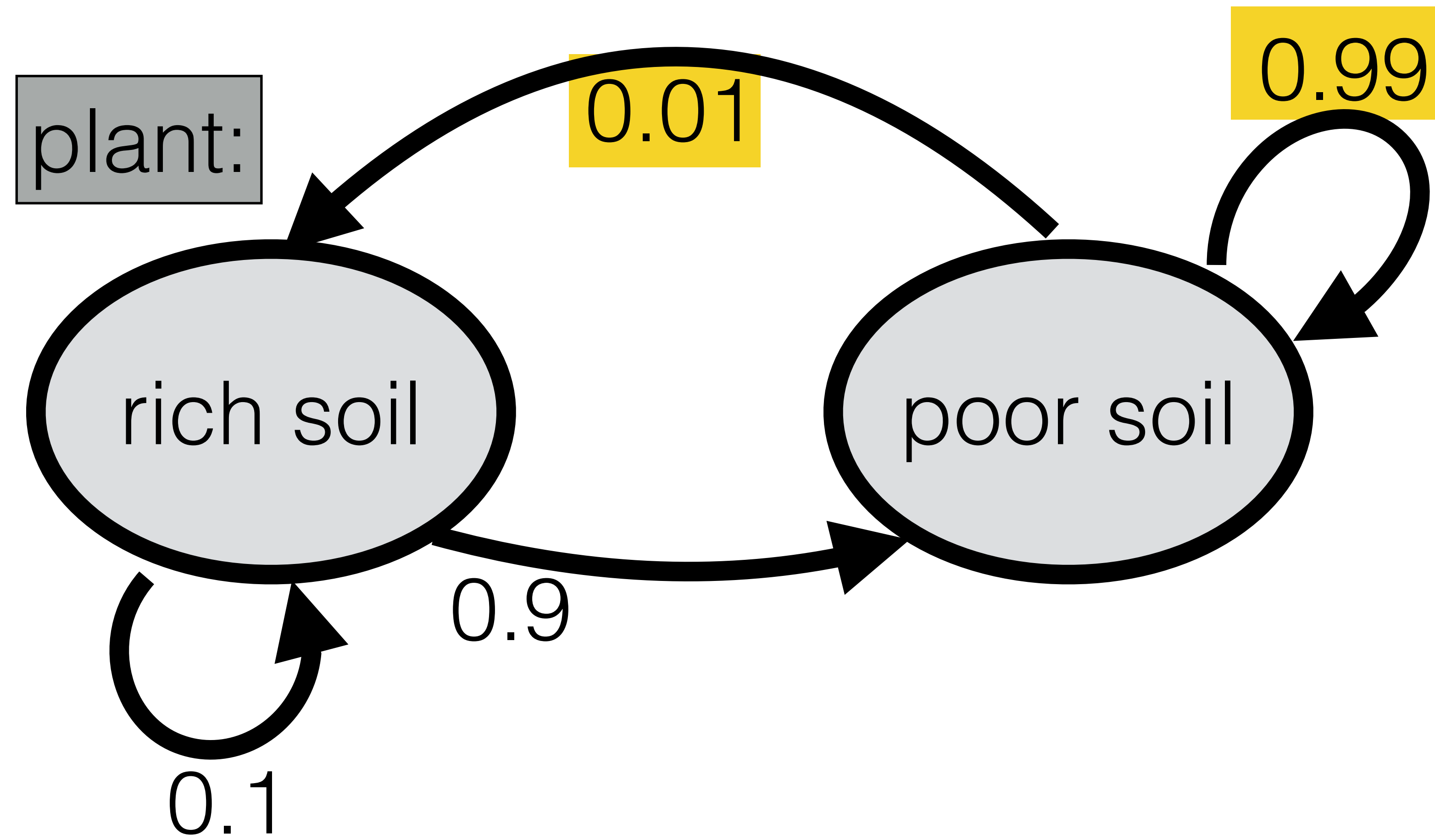
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
    - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) =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



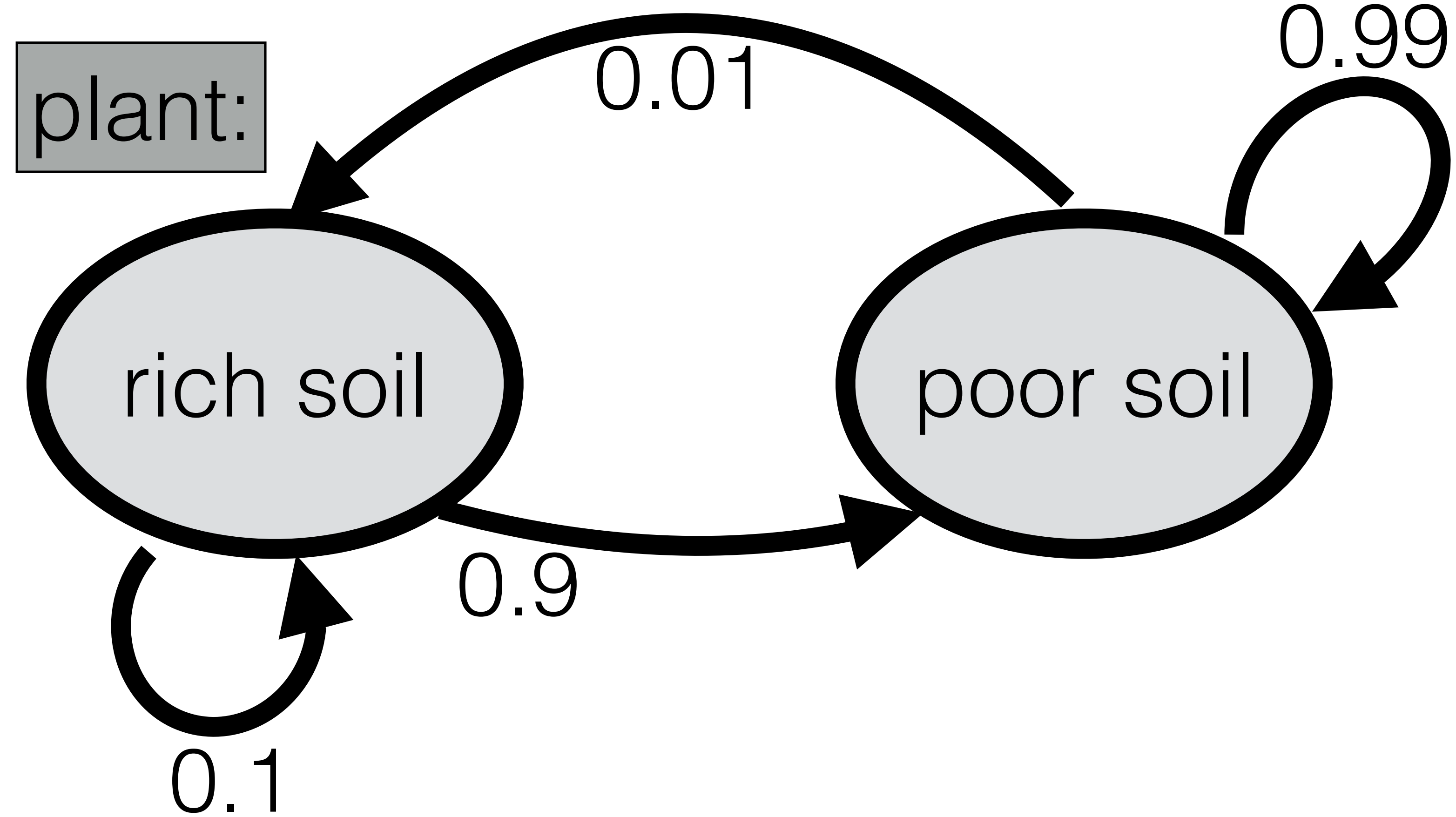
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) = R(poor,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



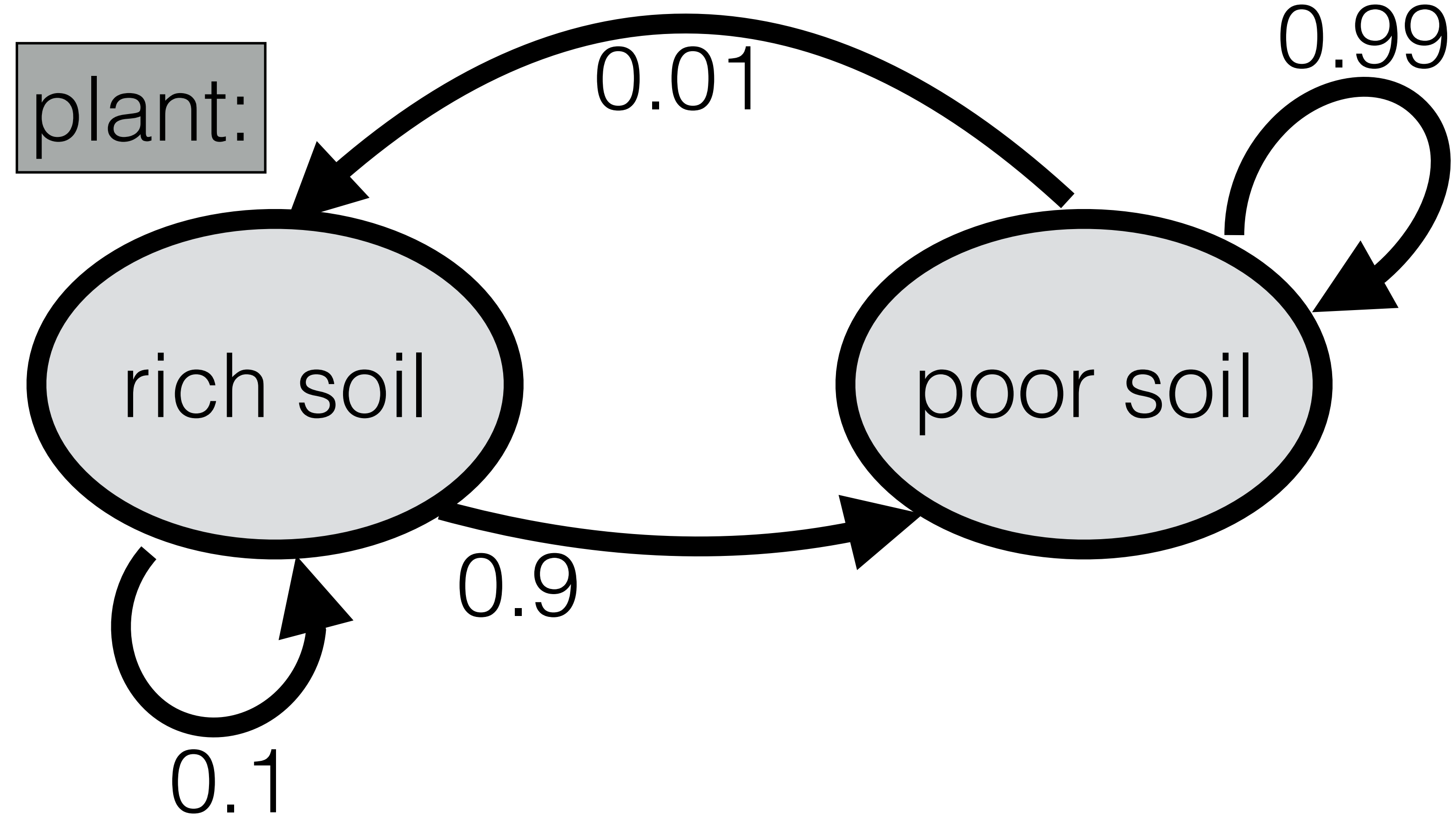
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) = R(poor,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

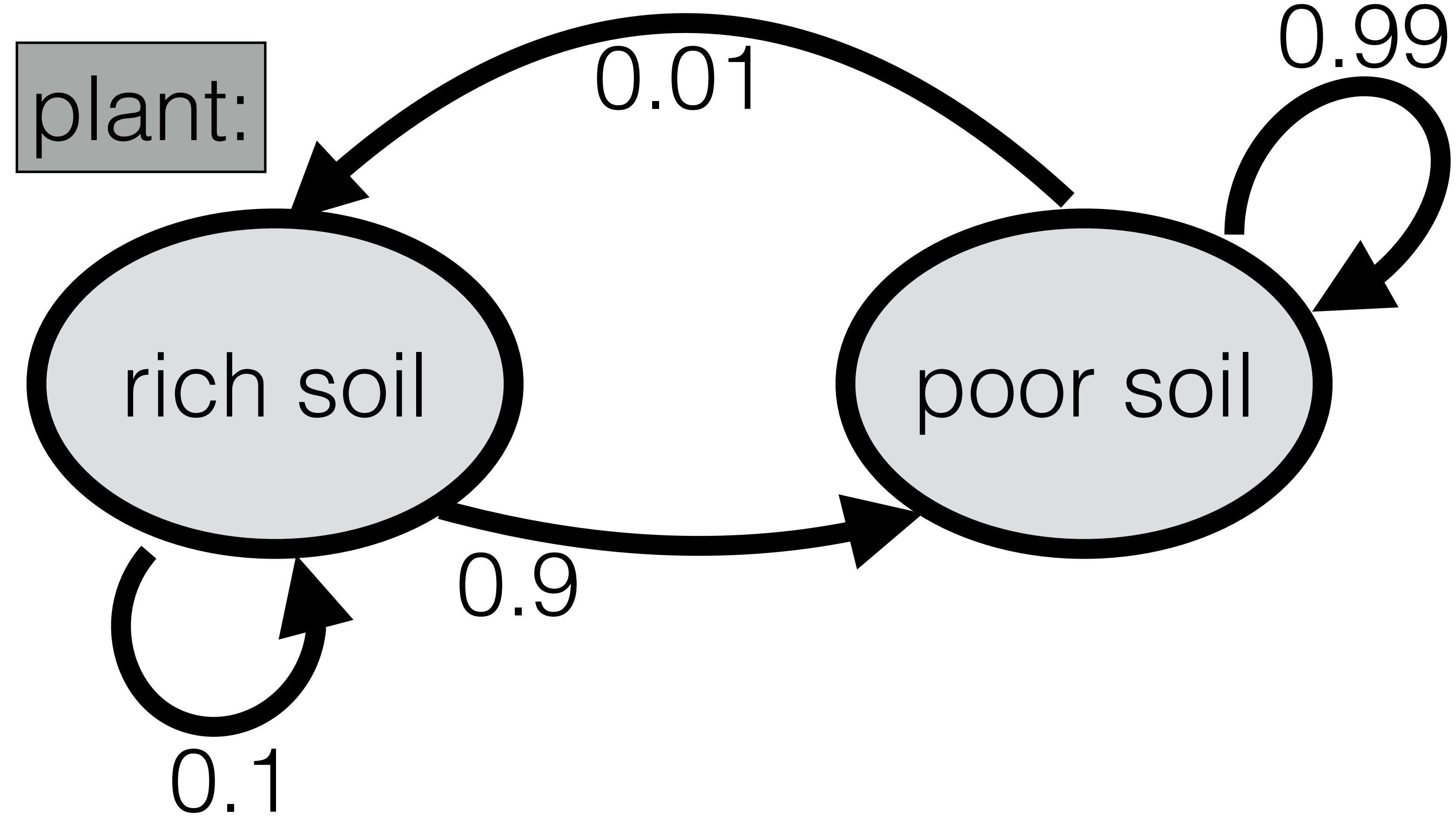


- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) = R(poor,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



- Transition matrix for "plant" action:

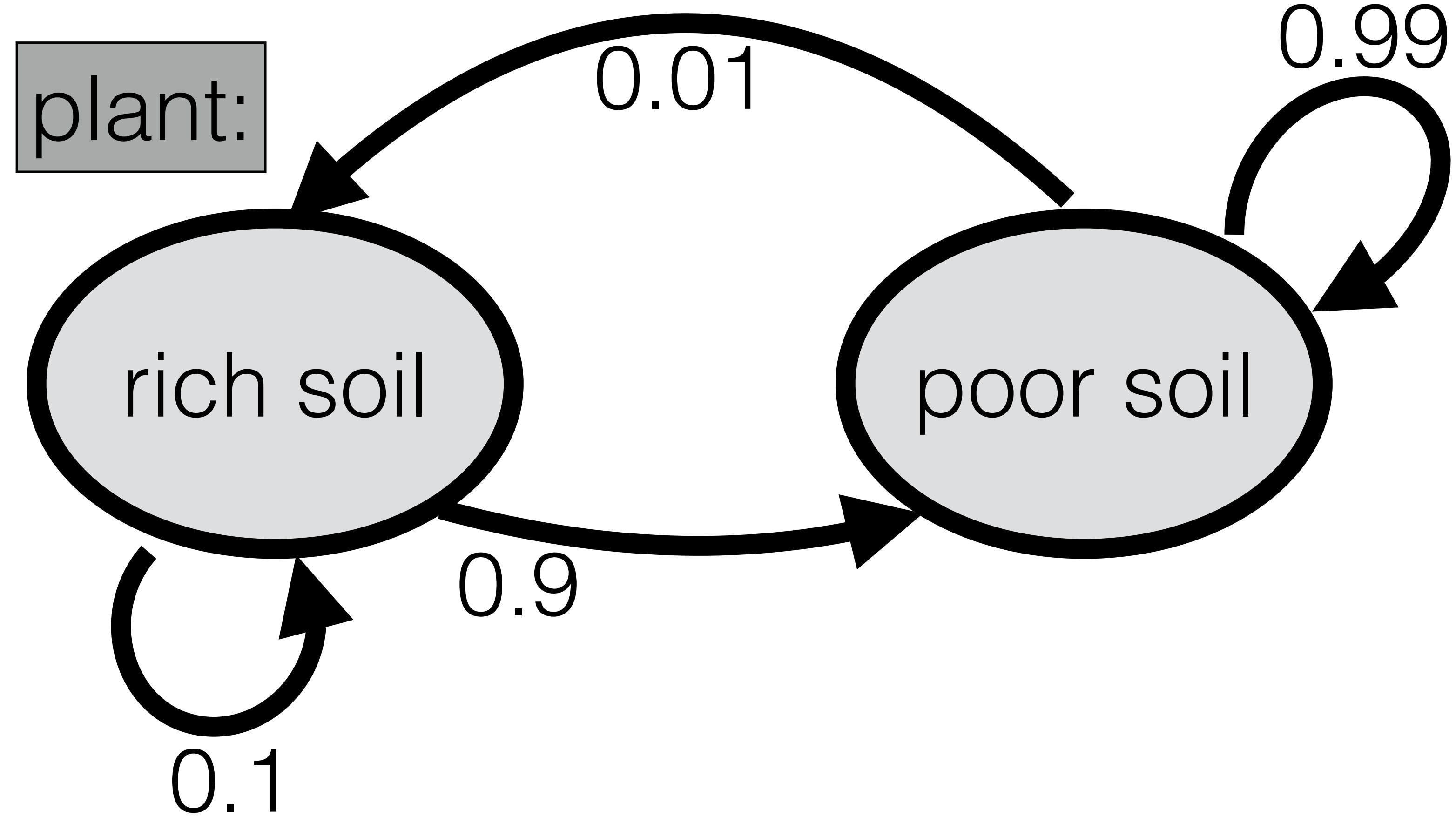
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T transition model
- $R : \mathcal{S} \times \mathcal{X} \rightarrow \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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



- Transition matrix for “plant” action:

	rich	poor
rich		
poor		

- $\mathcal{S}$ = 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

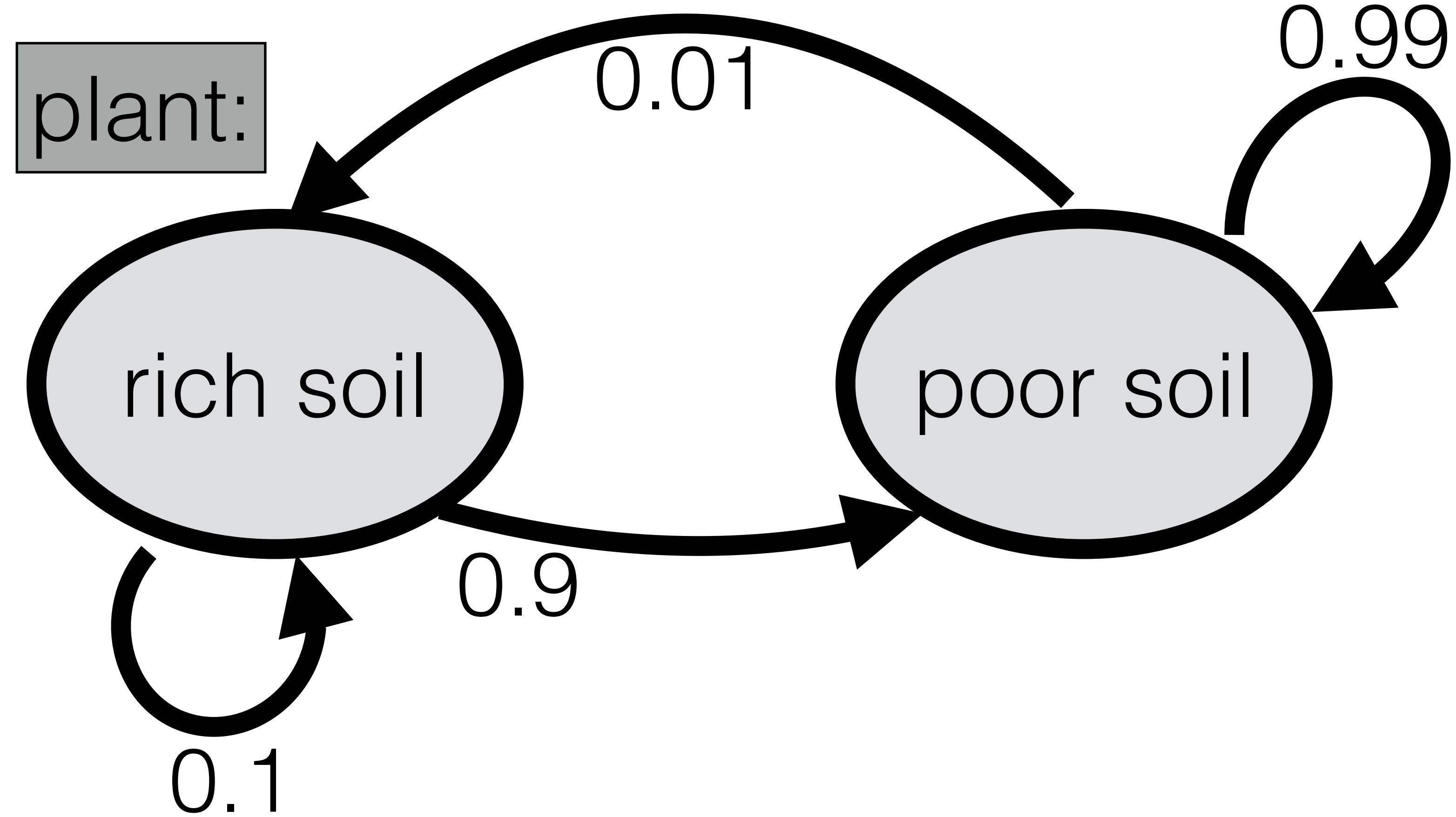


- Transition matrix for “plant” action:

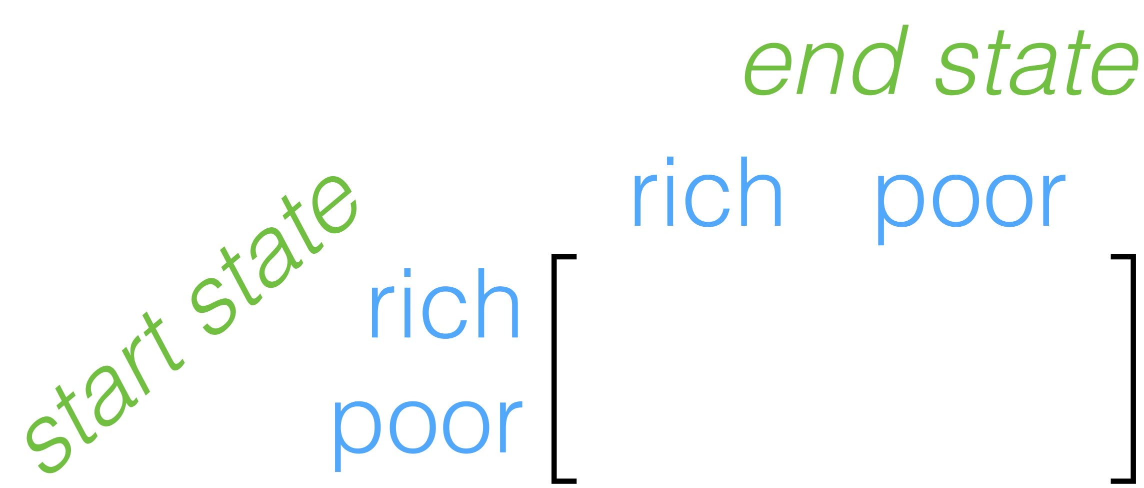
start state

	rich	poor
rich		
poor		

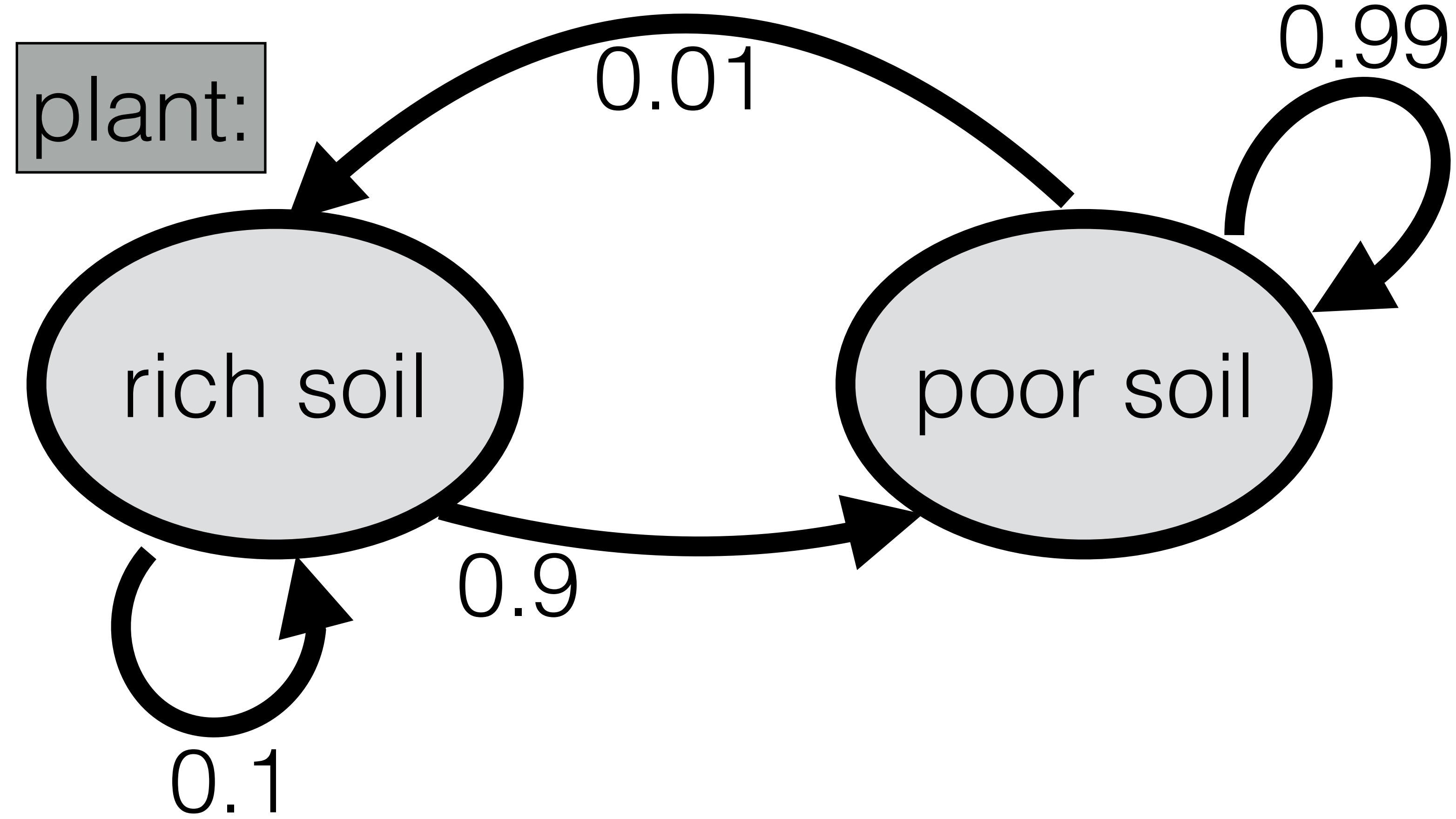
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) = R(poor,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



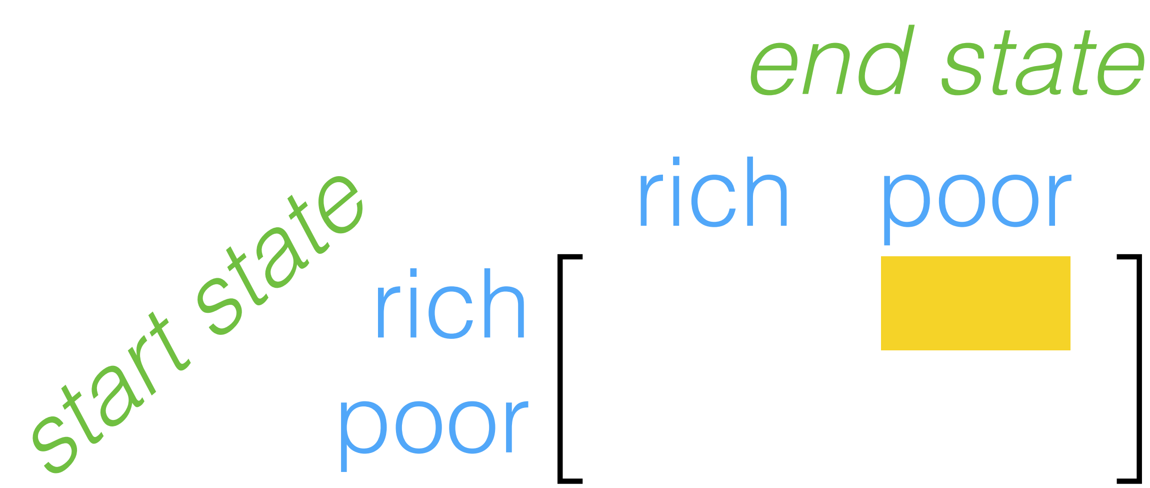
- Transition matrix for "plant" action:



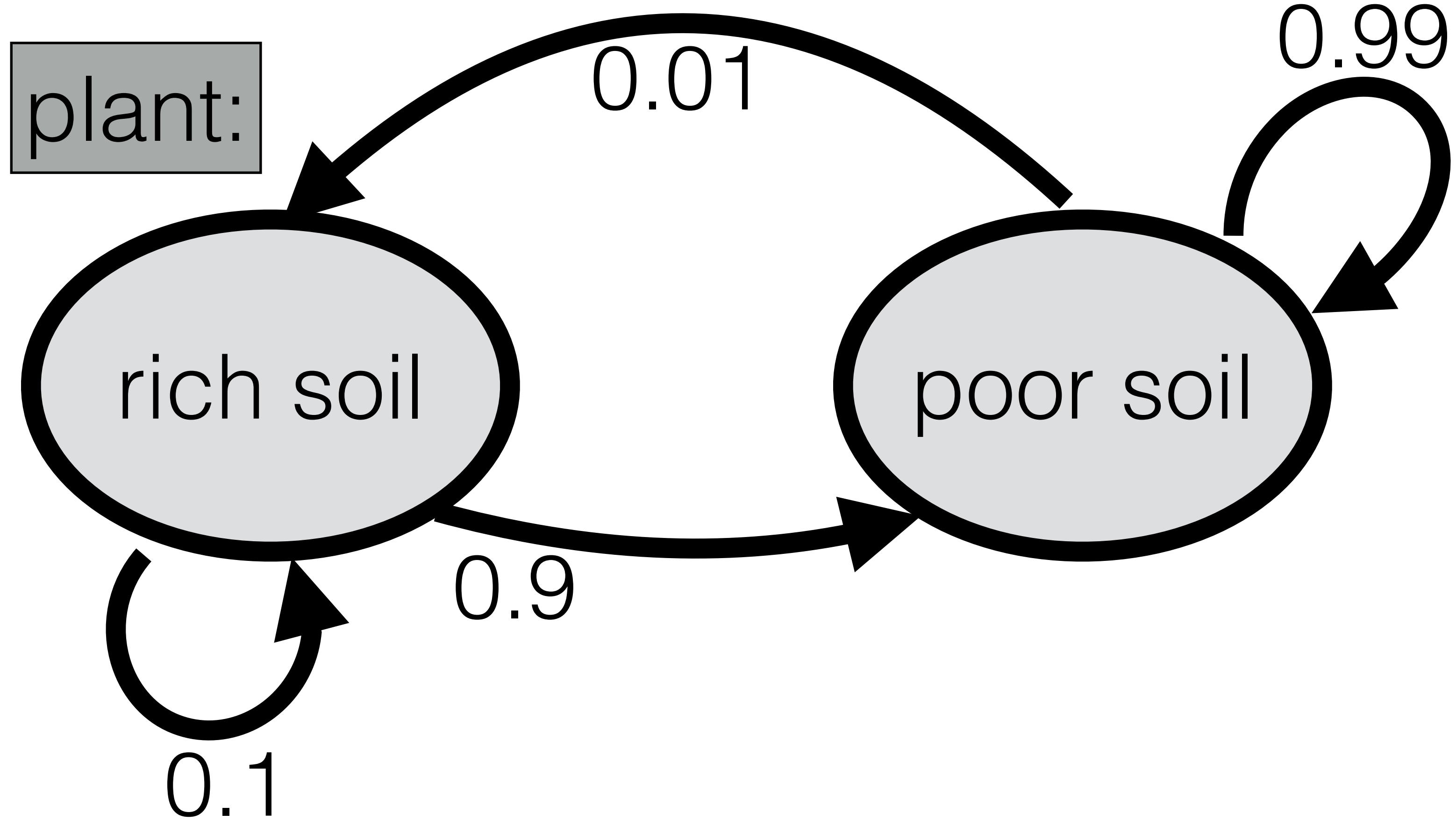
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T transition model
- $R : \mathcal{S} \times \mathcal{X} \rightarrow \mathbb{R}$: reward function
    - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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



- Transition matrix for "plant" action:



- $\mathcal{S}$ = 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

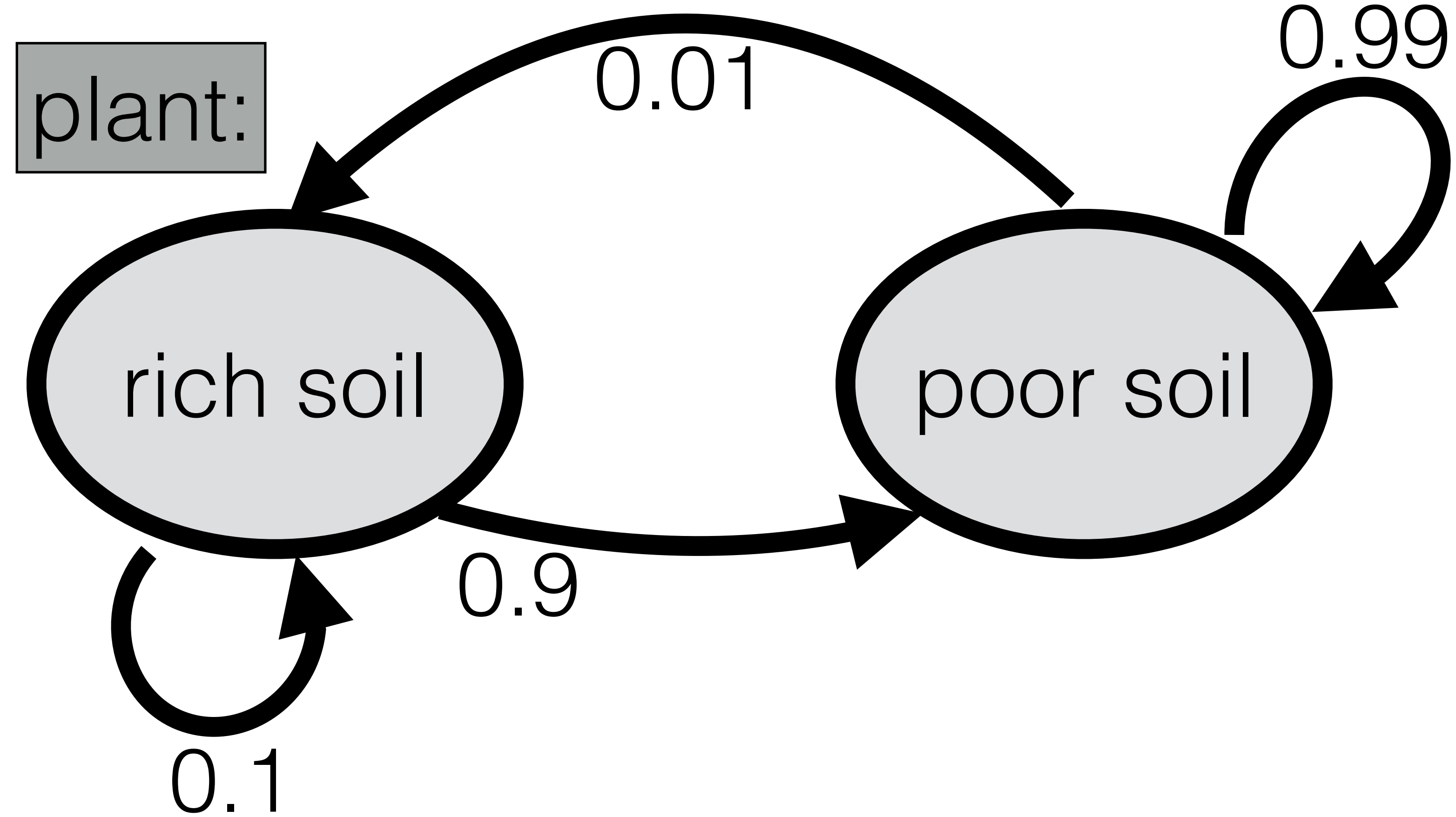


- Transition matrix for "plant" action:

end state / *start state*

	rich	poor
rich		0.9
poor		

- $\mathcal{S}$ = 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



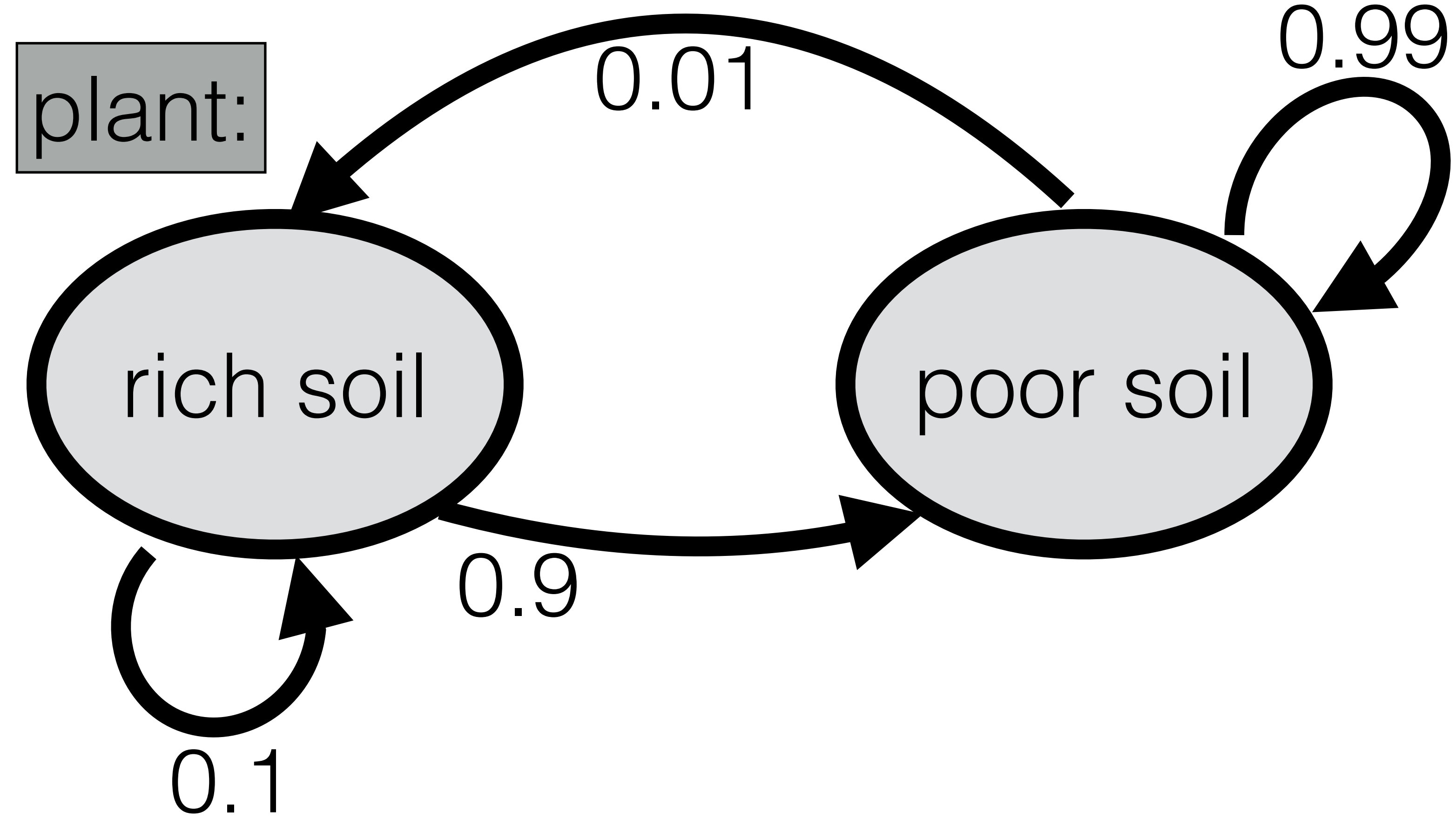
- Transition matrix for “plant” action:

end state

start state

	rich	poor
rich	0.1	0.9
poor		

- $\mathcal{S}$ = 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



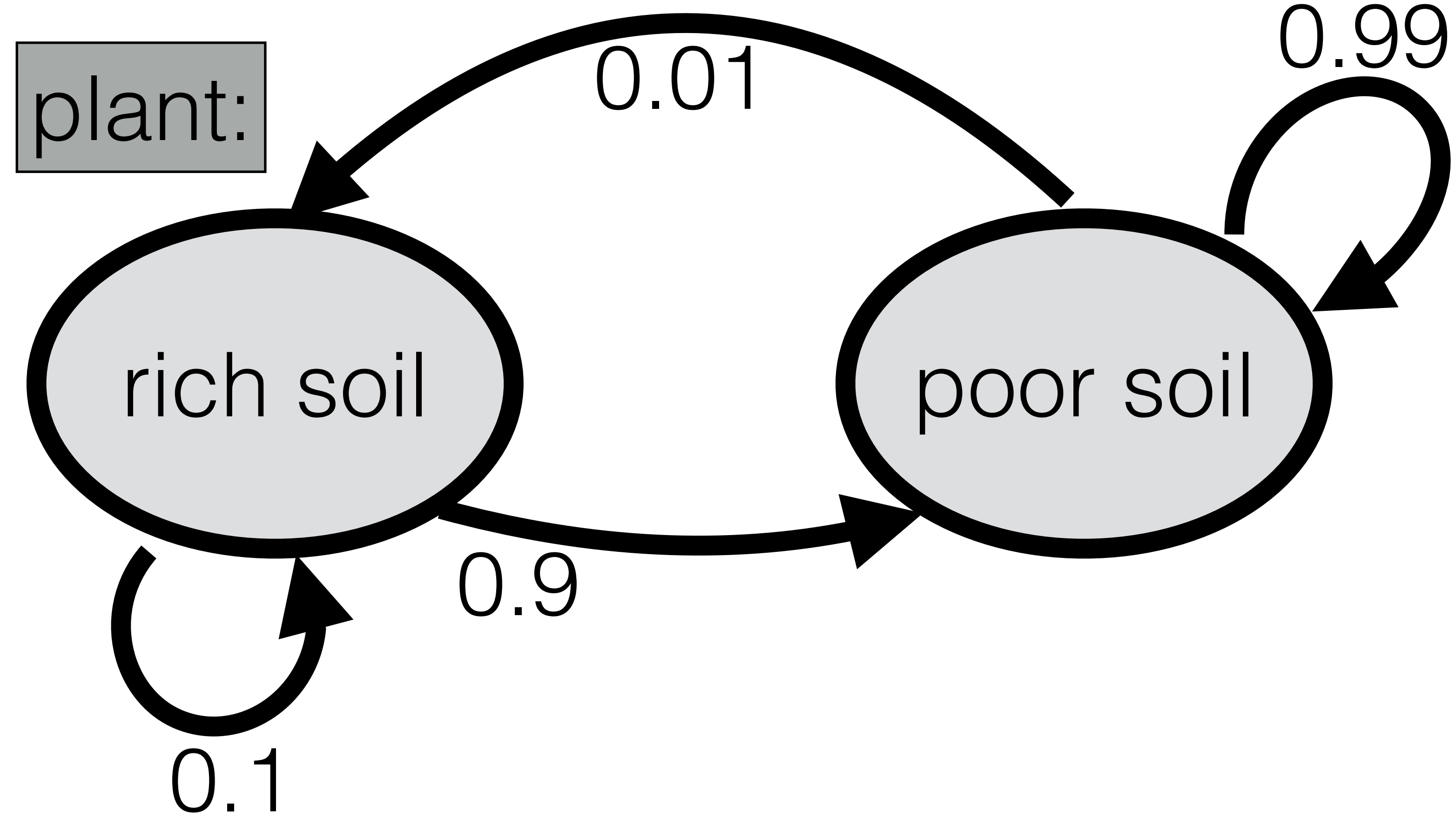
- Transition matrix for "plant" action:

end state

start state

	rich	poor
rich	0.1	0.9
poor	0.01	0.99

- $\mathcal{S}$ = 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

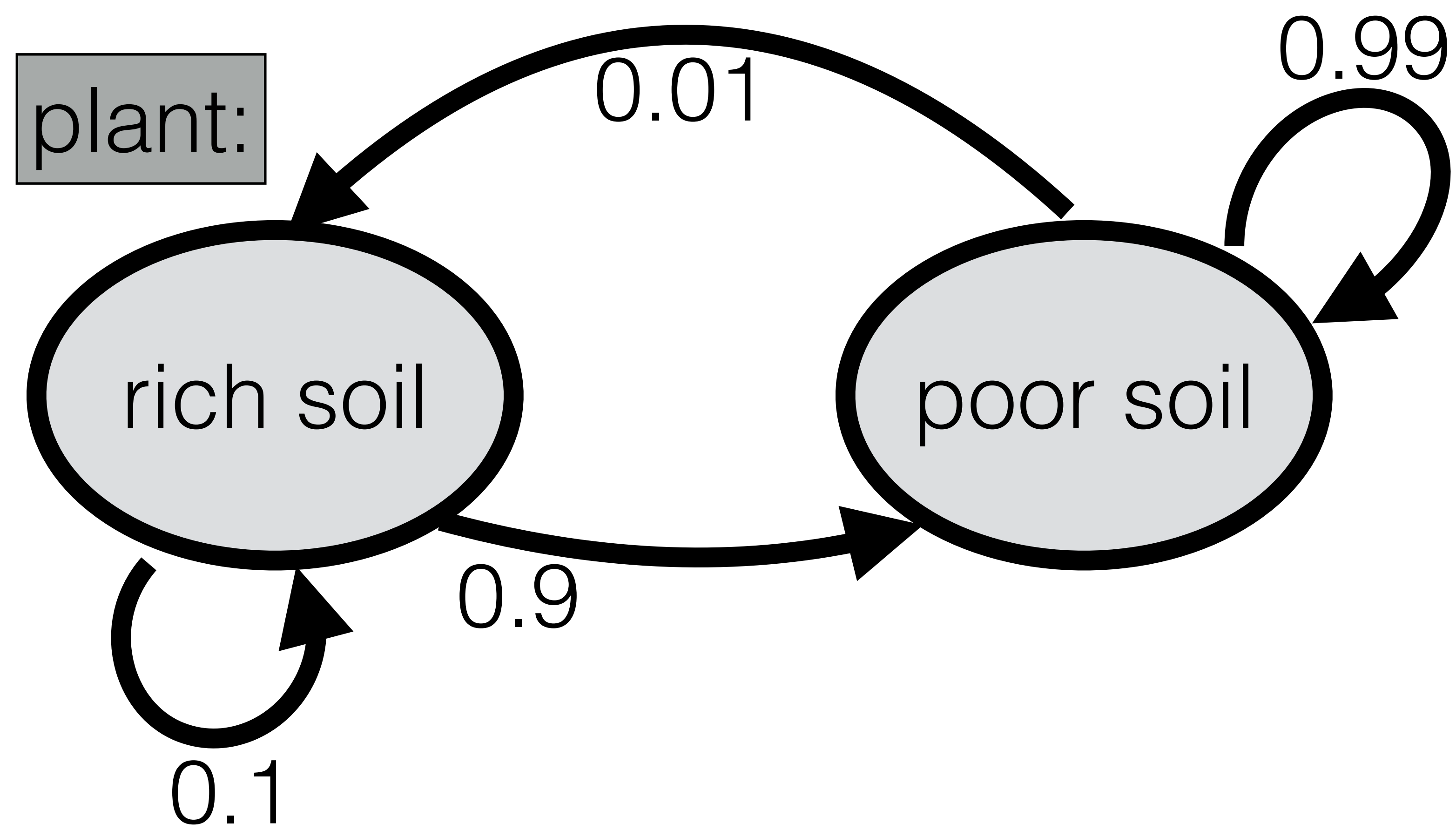


- Transition matrix for "plant" action:

end state

start state	rich	poor
rich	0.1	0.9
poor	0.01	0.99

- $\mathcal{S}$ = 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



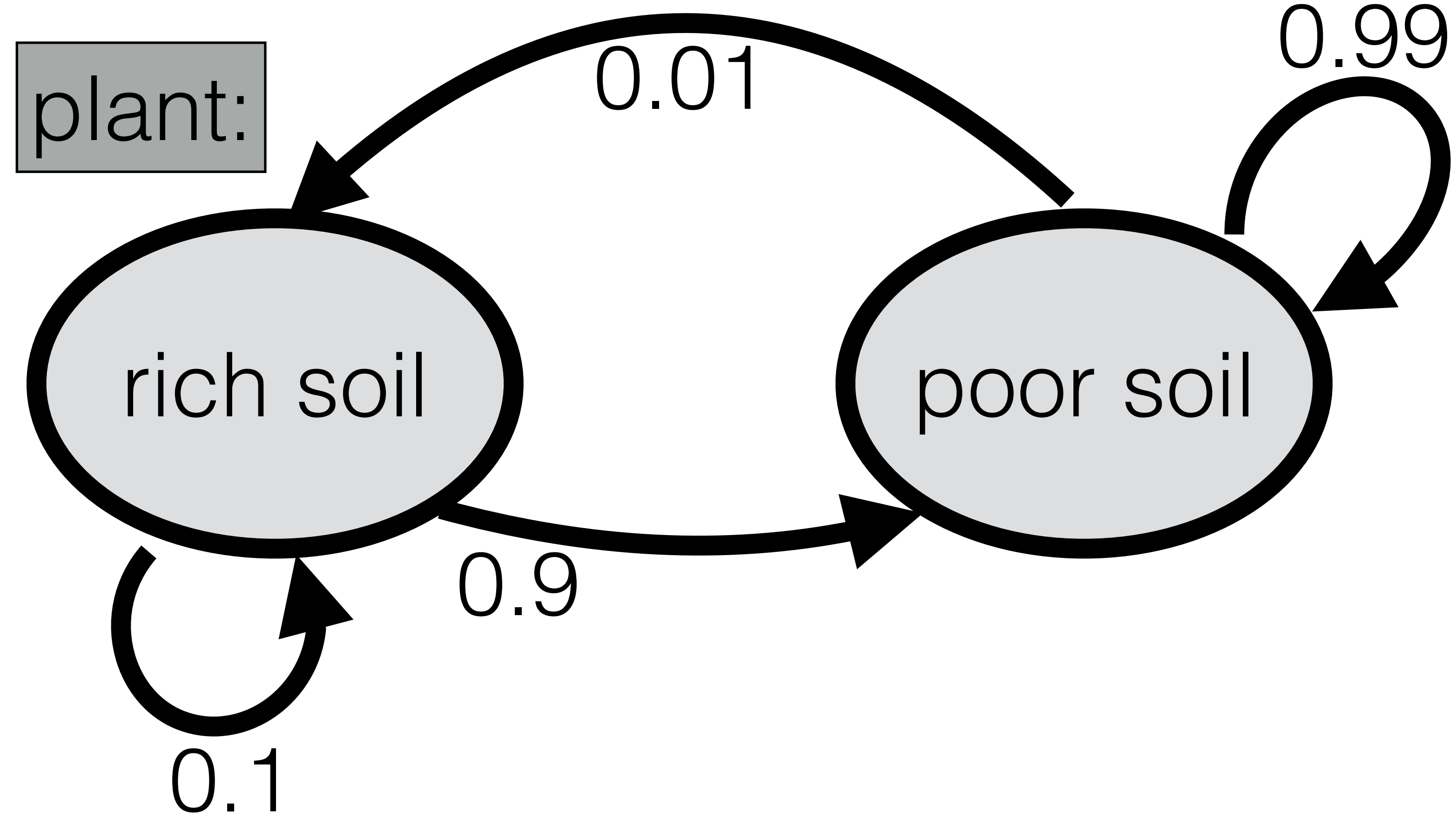
- Transition matrix for "plant" action:

end state

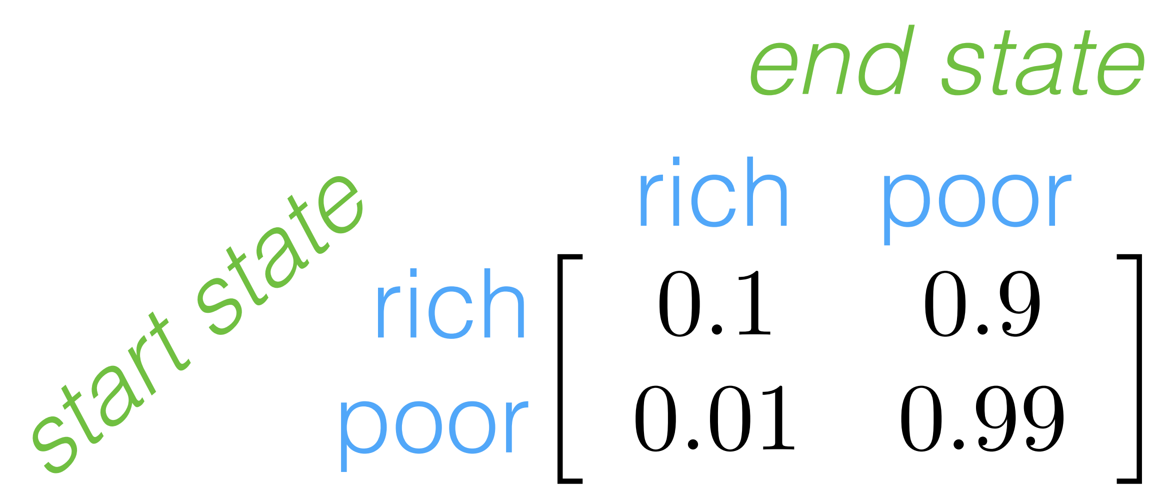
start state

	rich	poor
rich	0.1	0.9
poor	0.01	0.99

- $\mathcal{S}$ = 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

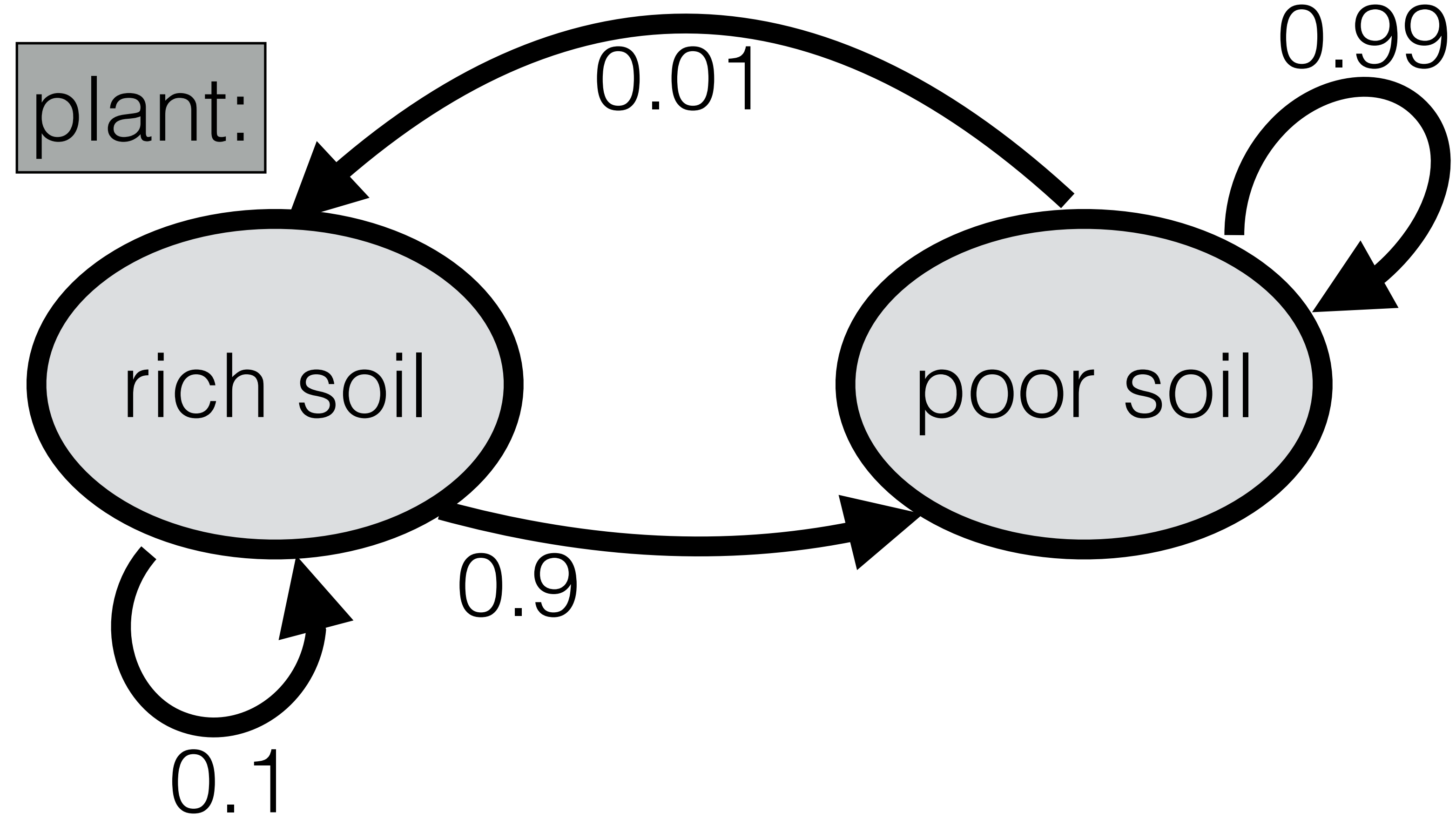


- Transition matrix for "plant" action:

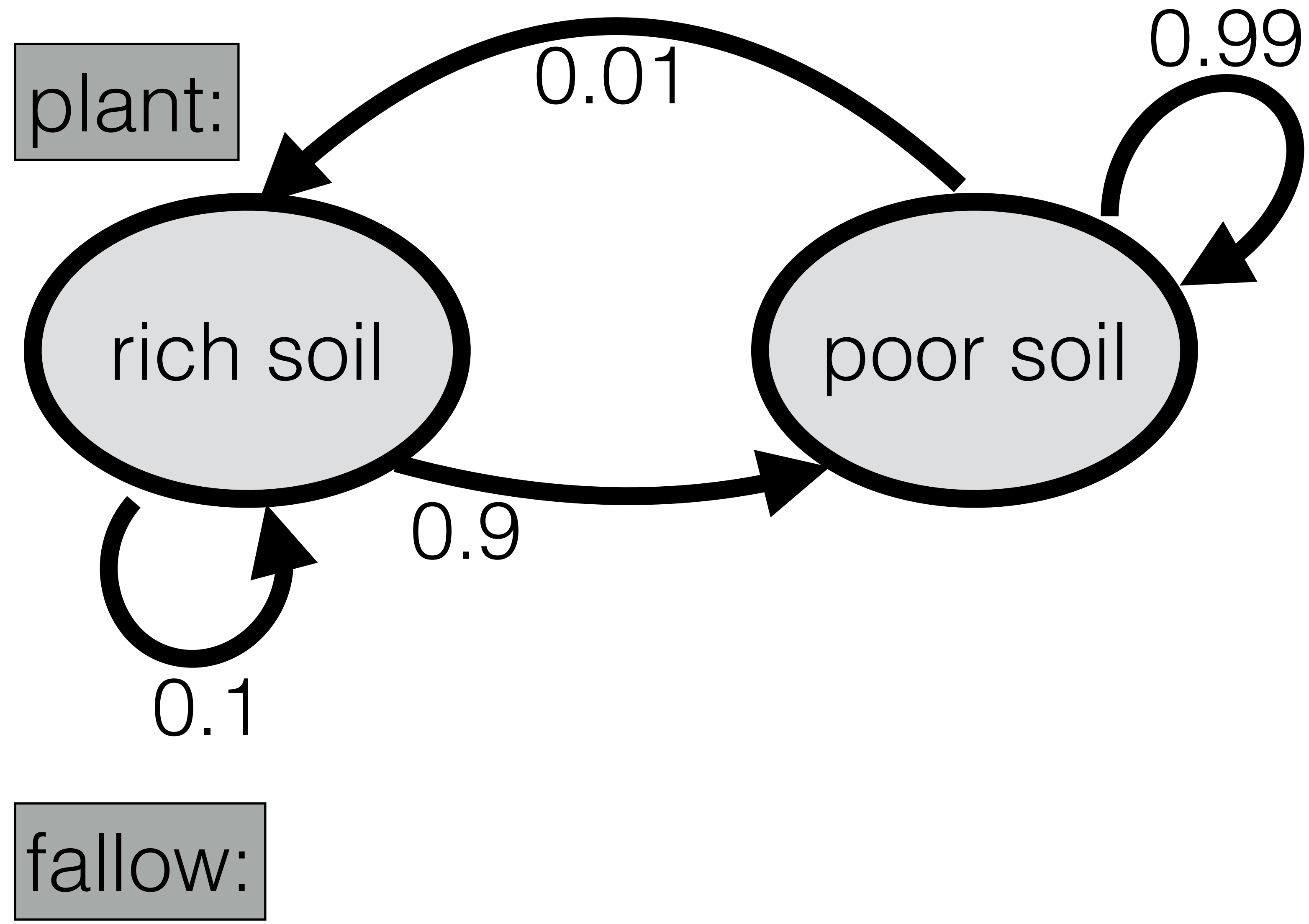


	rich	poor
rich	0.1	0.9
poor	0.01	0.99

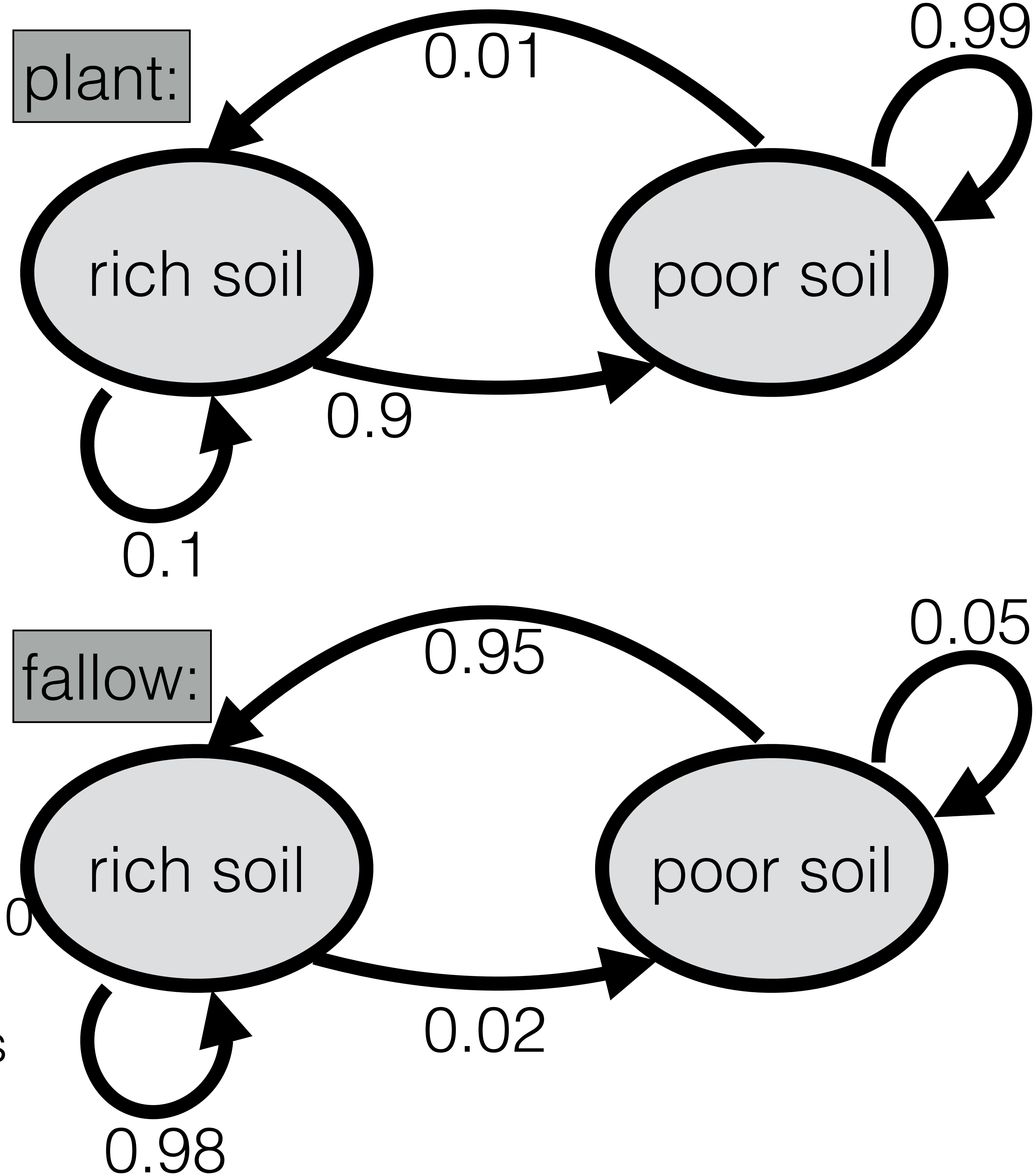
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T
  transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) =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



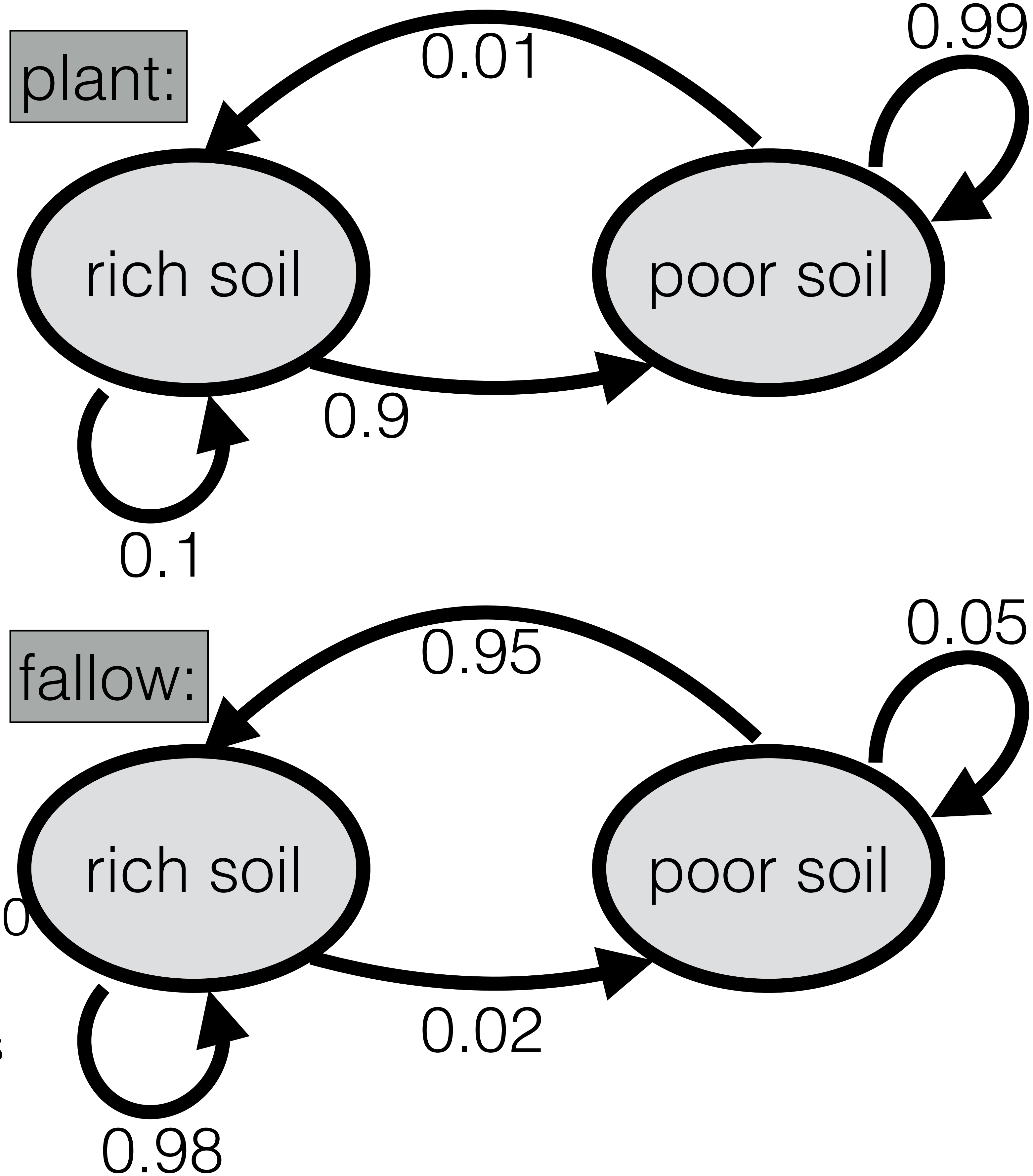
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T
  transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$:
  reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow)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



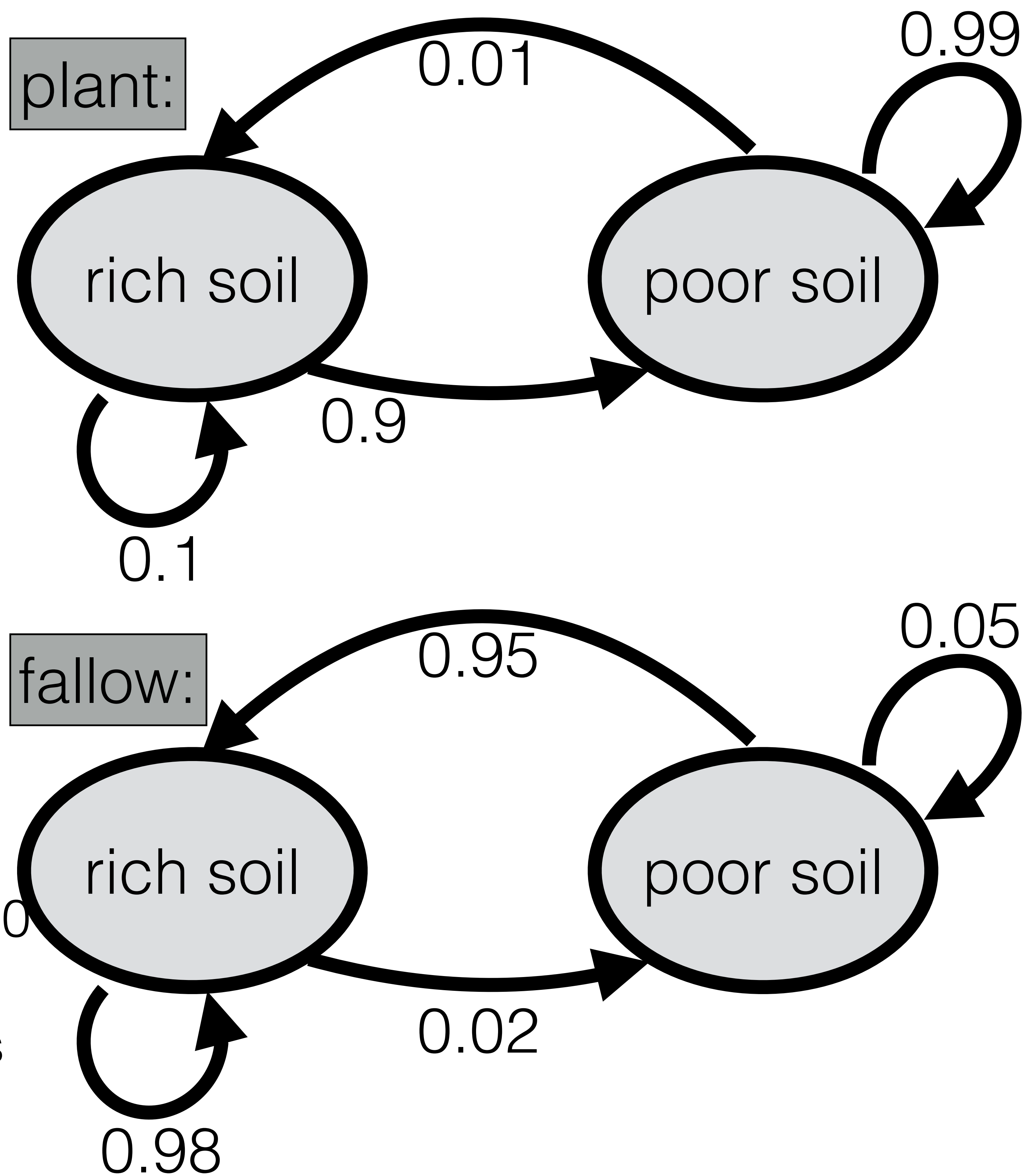
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T
  transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) =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



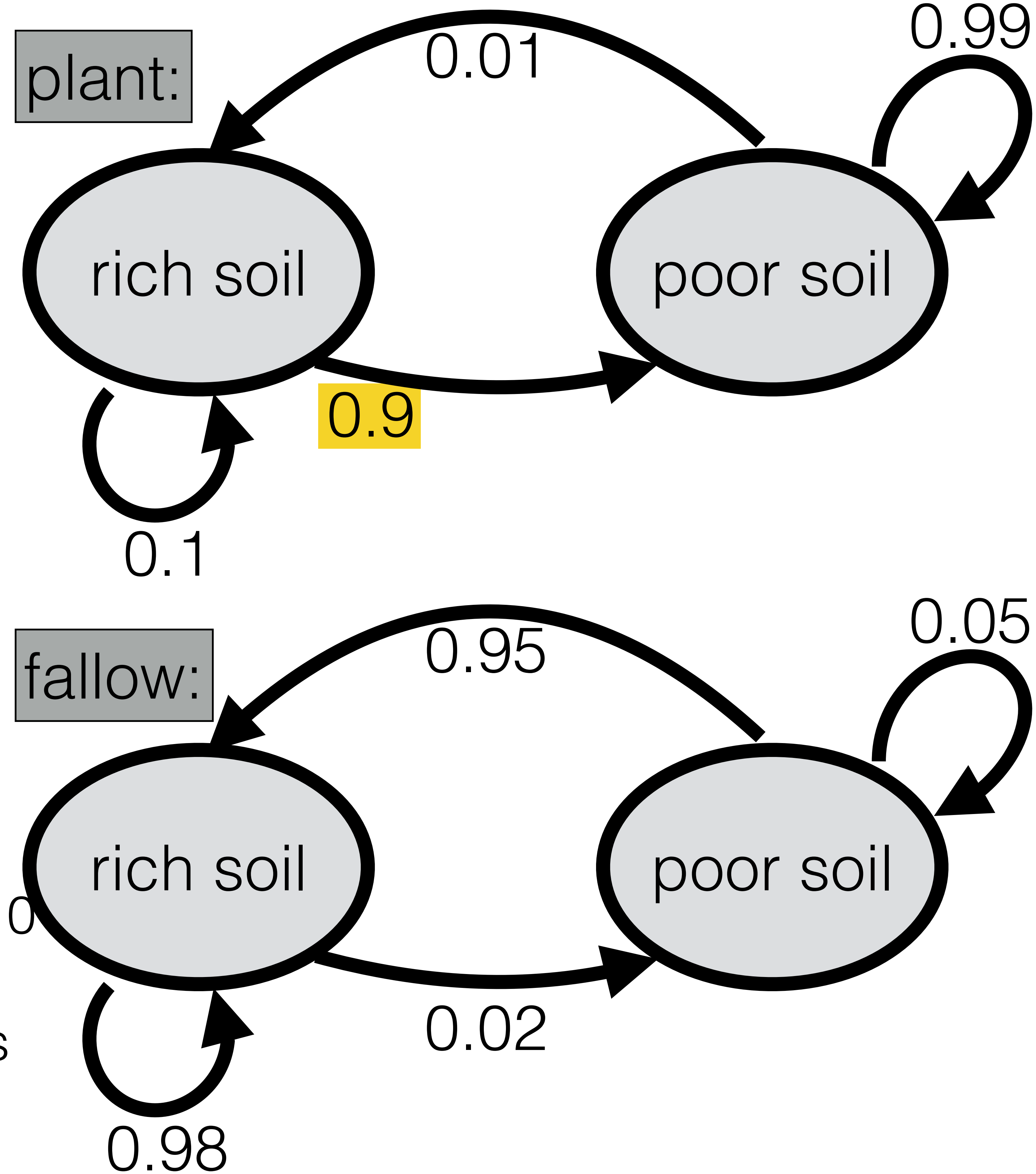
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- T transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) = 10 bushels; R(rich, fallow) = R(poor,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



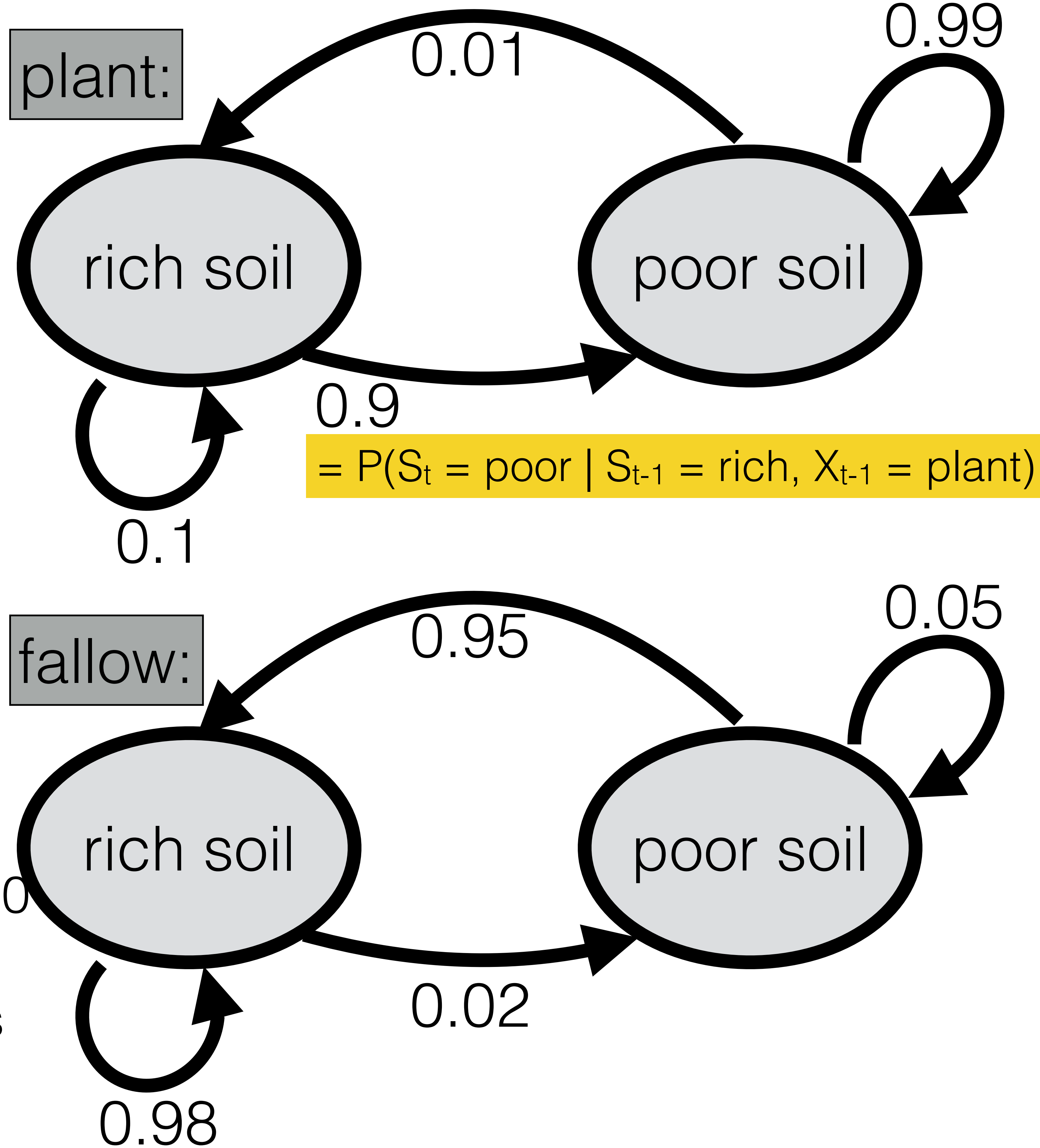
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
    - e.g. R(rich,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



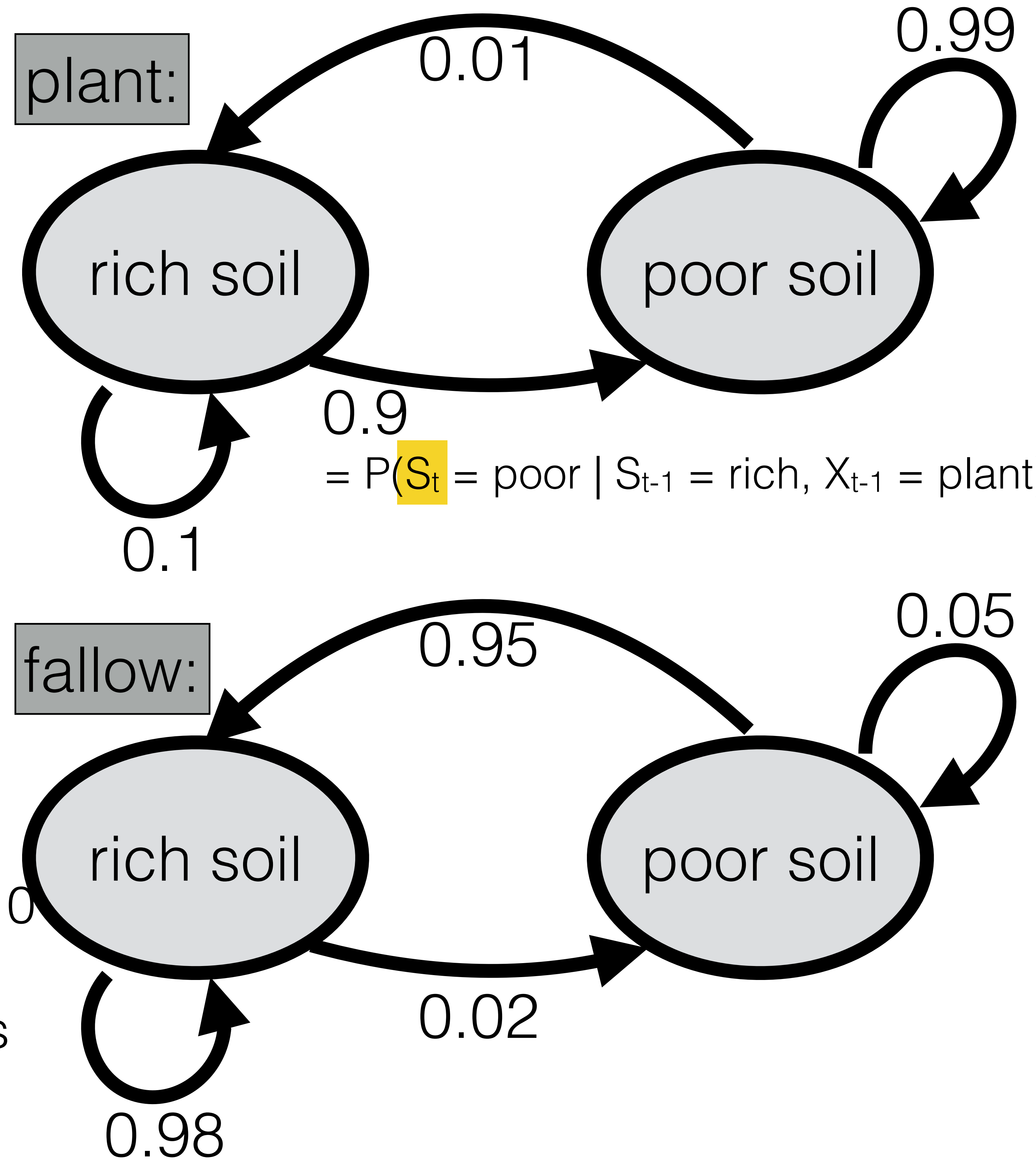
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
    - e.g. R(rich,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



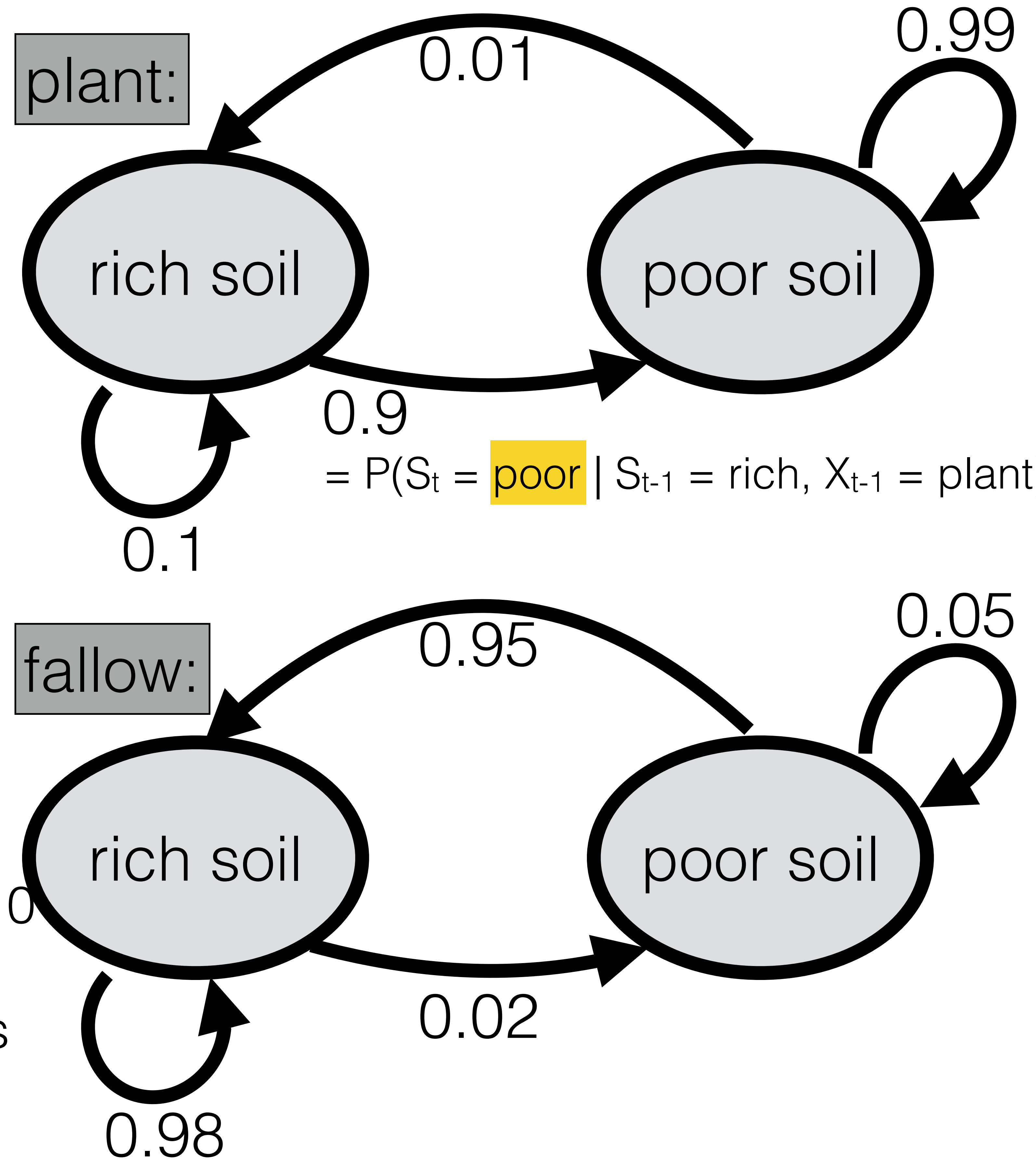
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
    - e.g. R(rich,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



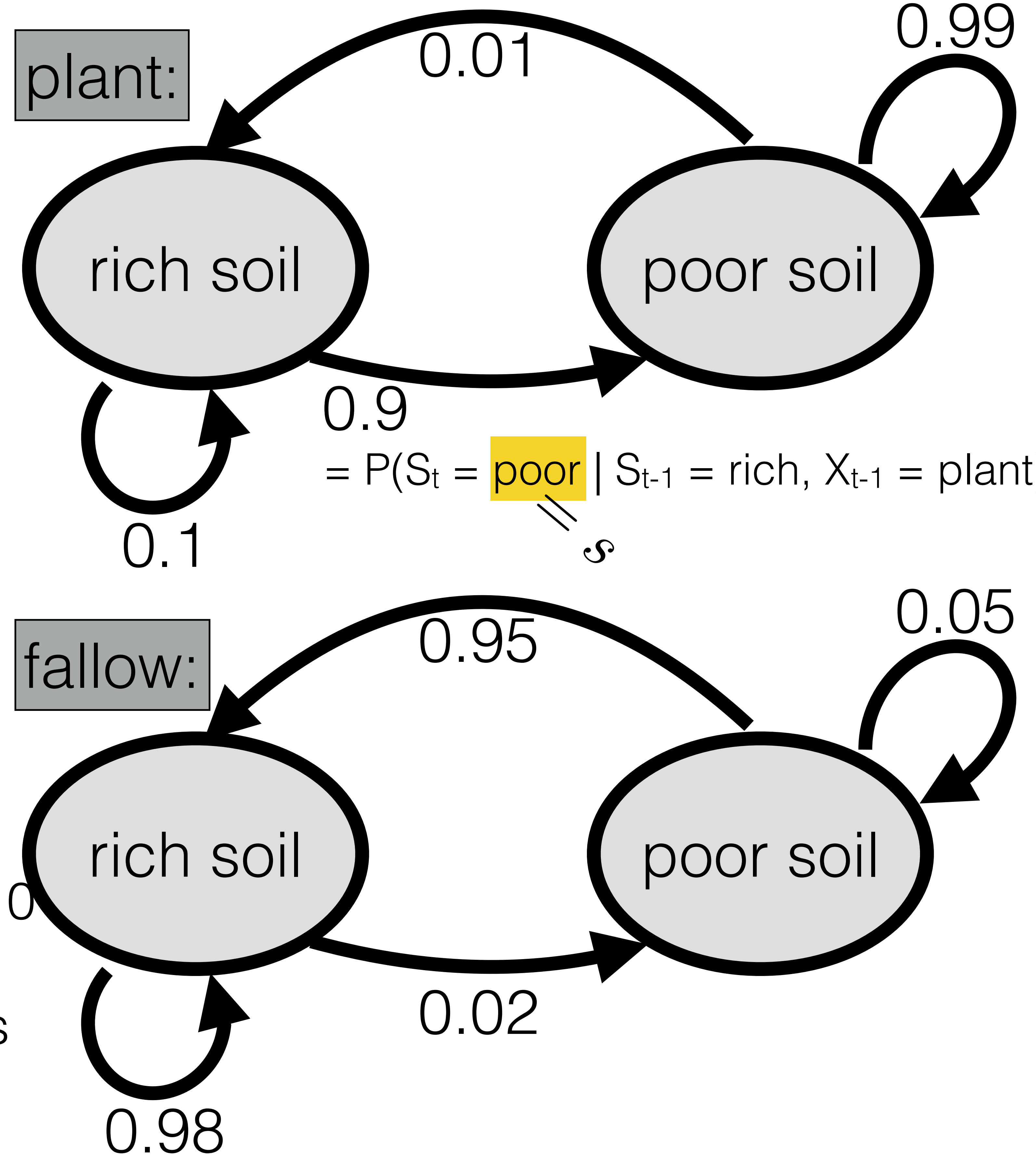
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
    - e.g. R(rich,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



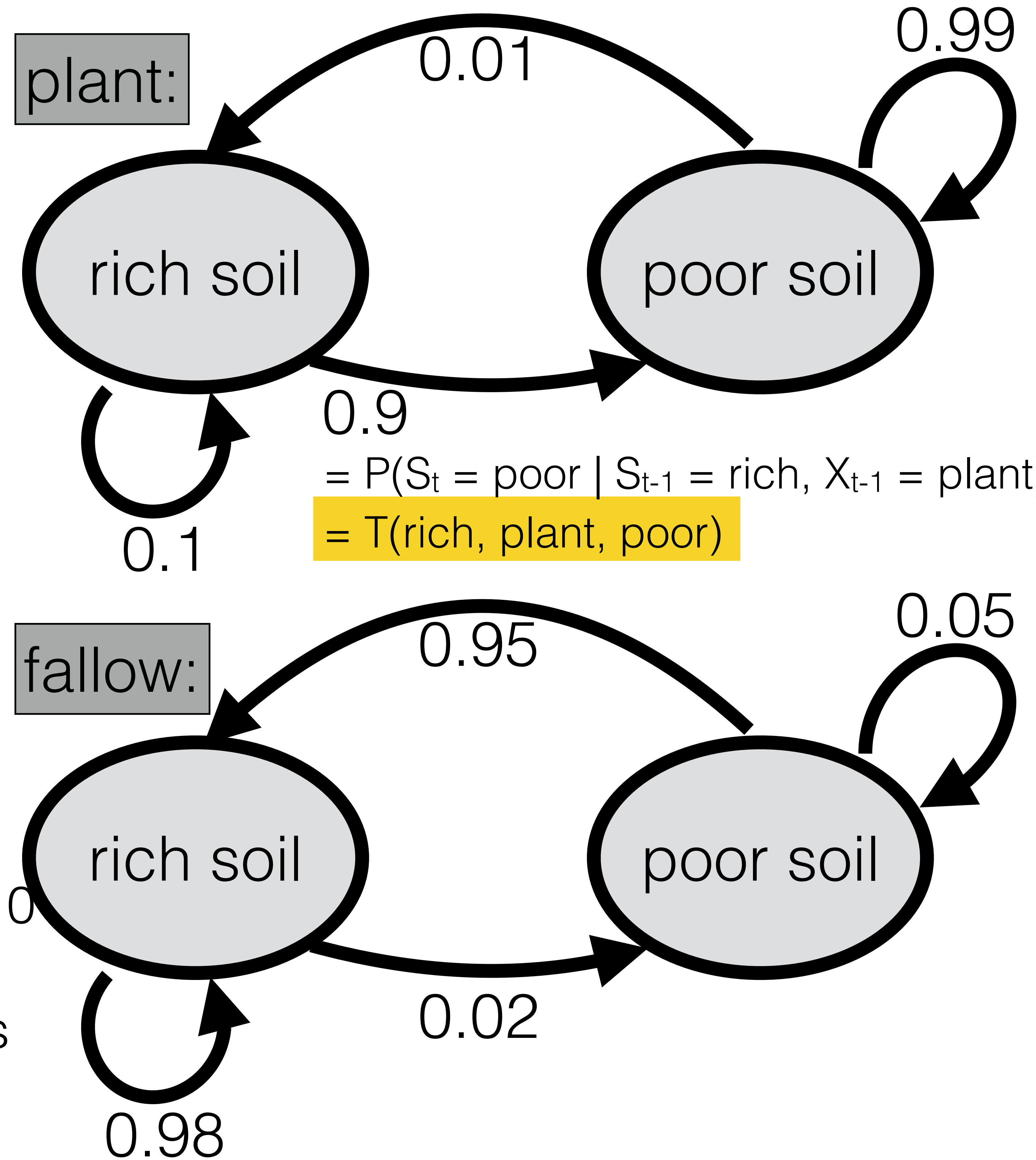
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant)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



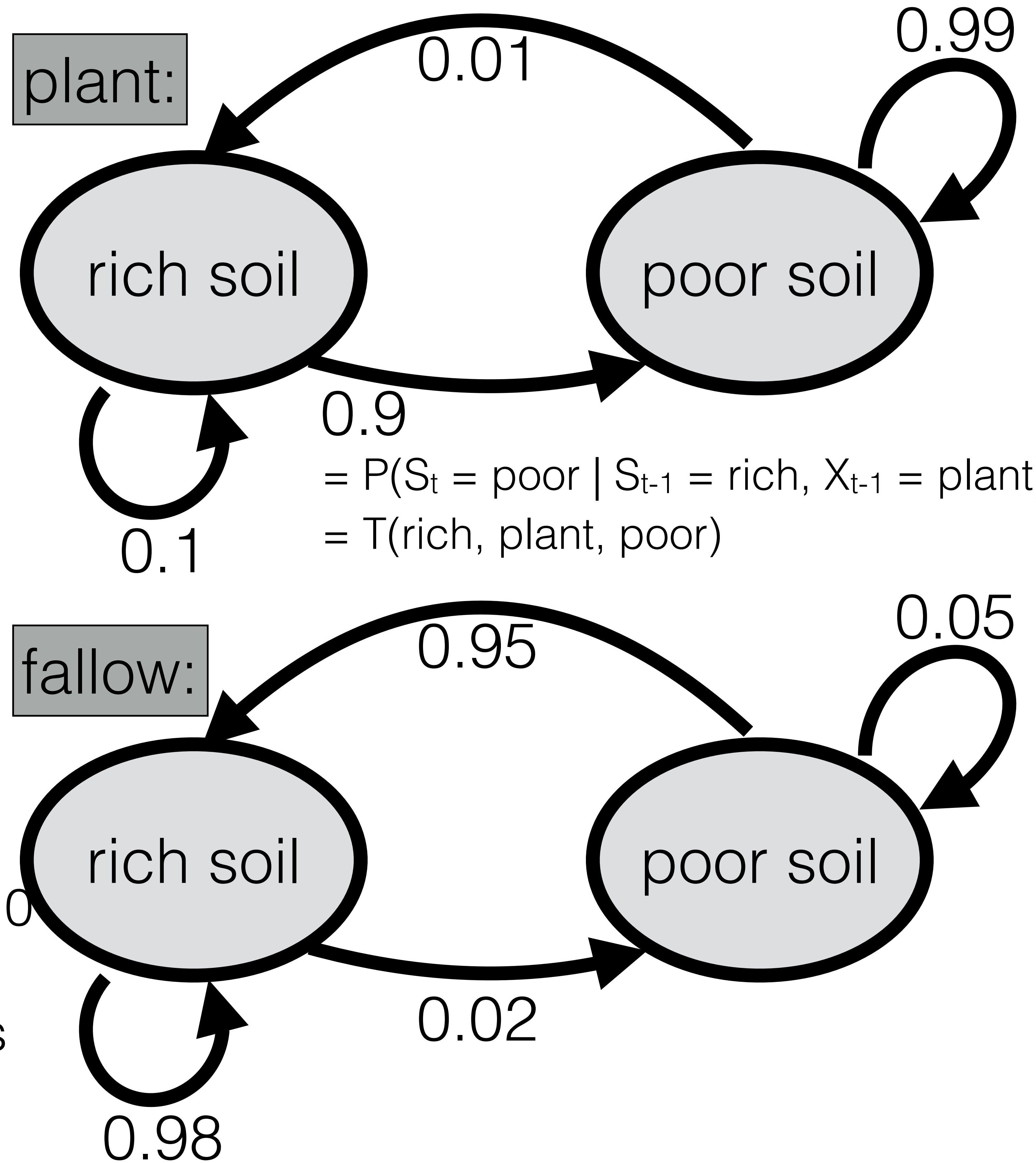
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \rightarrow \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \rightarrow \mathbb{R}$: reward function
 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

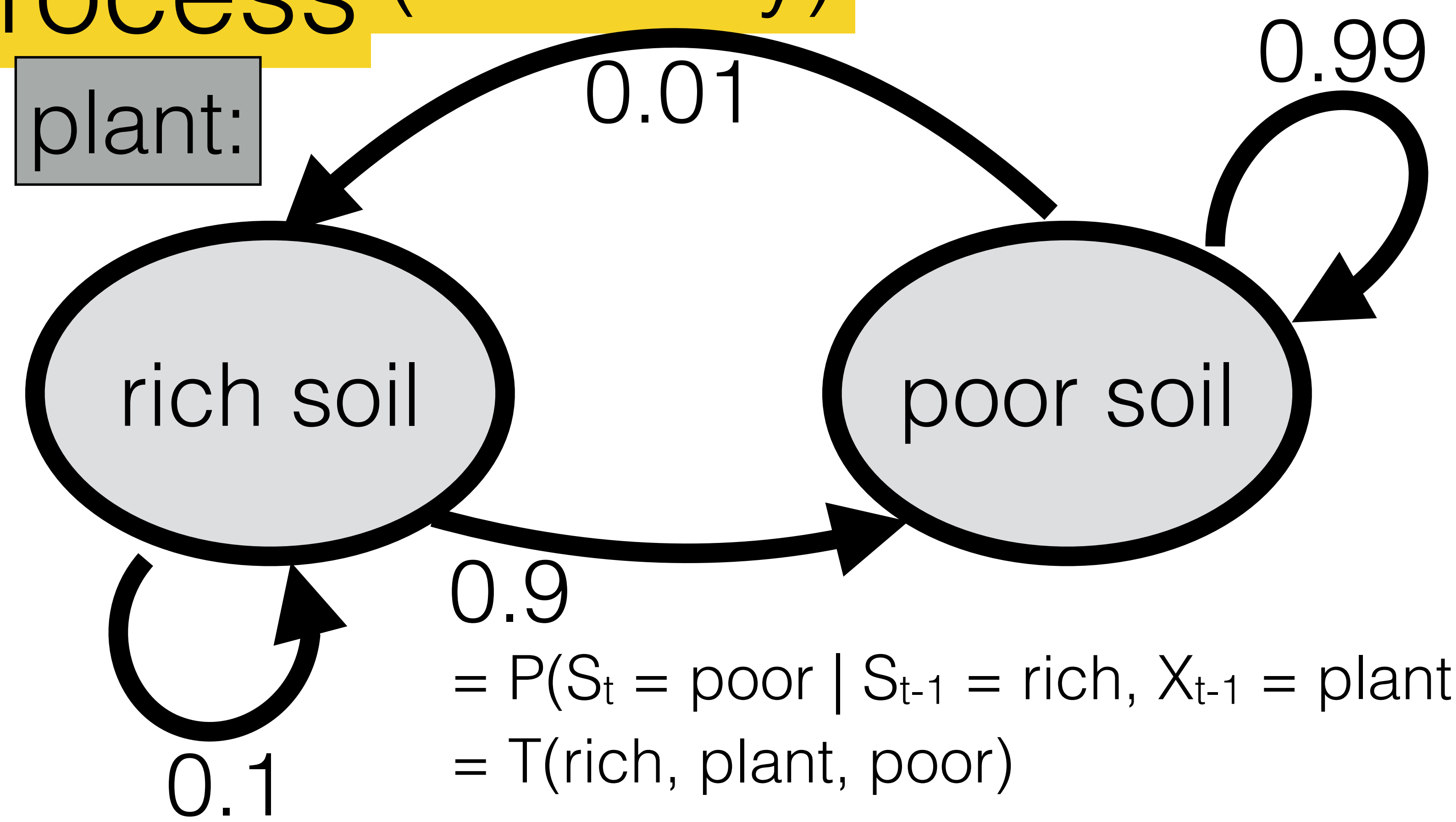


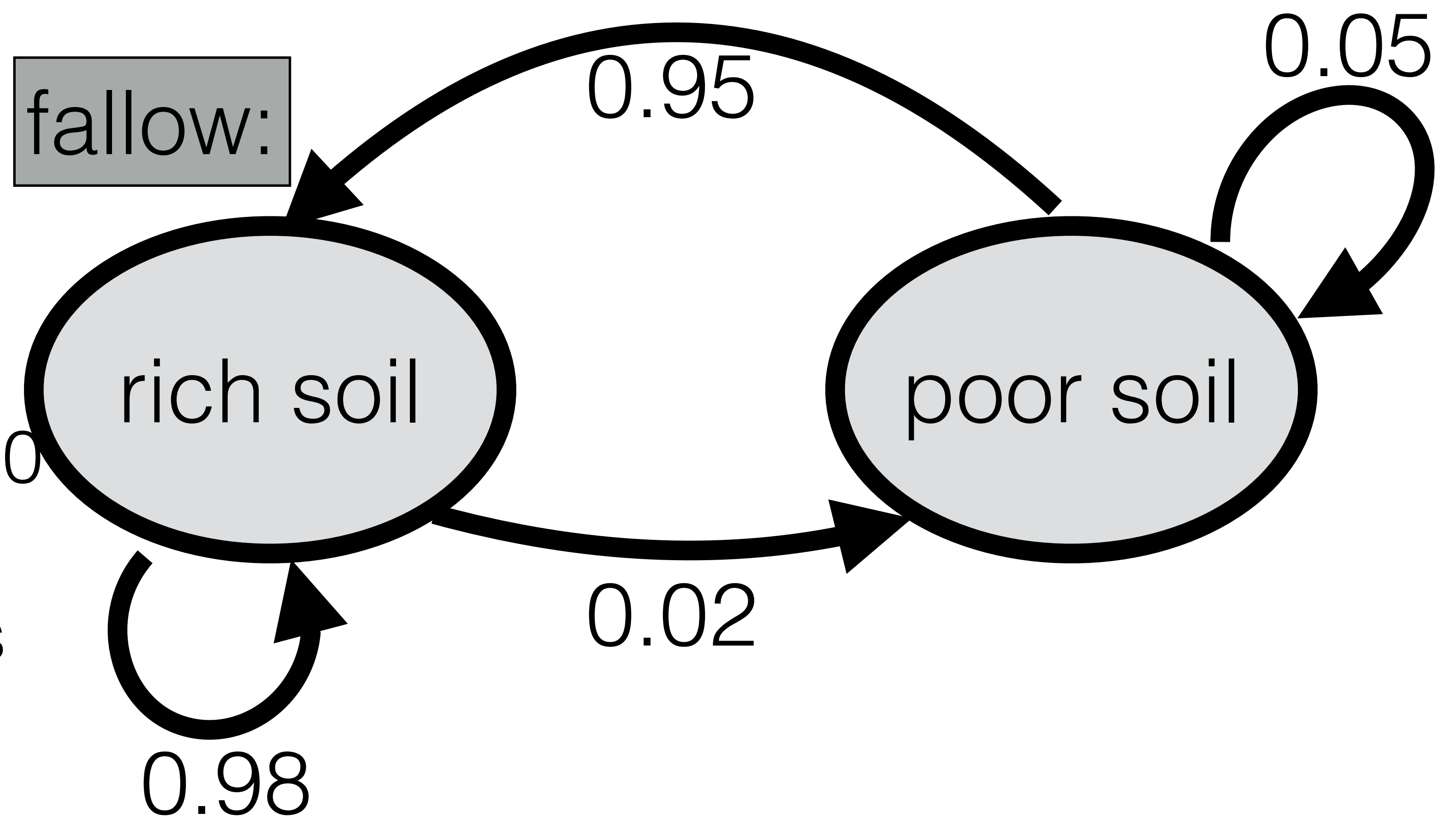
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant)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



Markov Decision Process (basically)

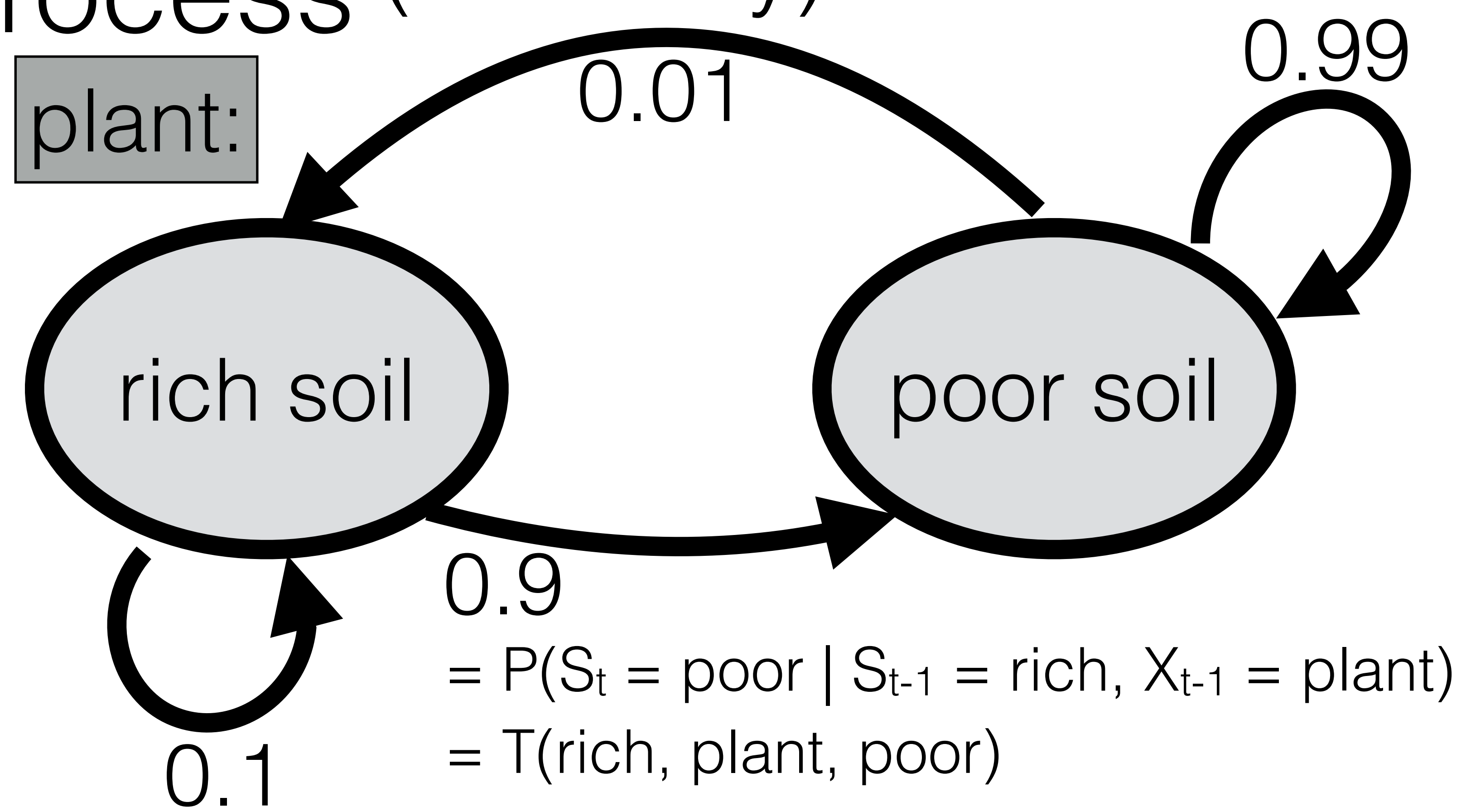
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant)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

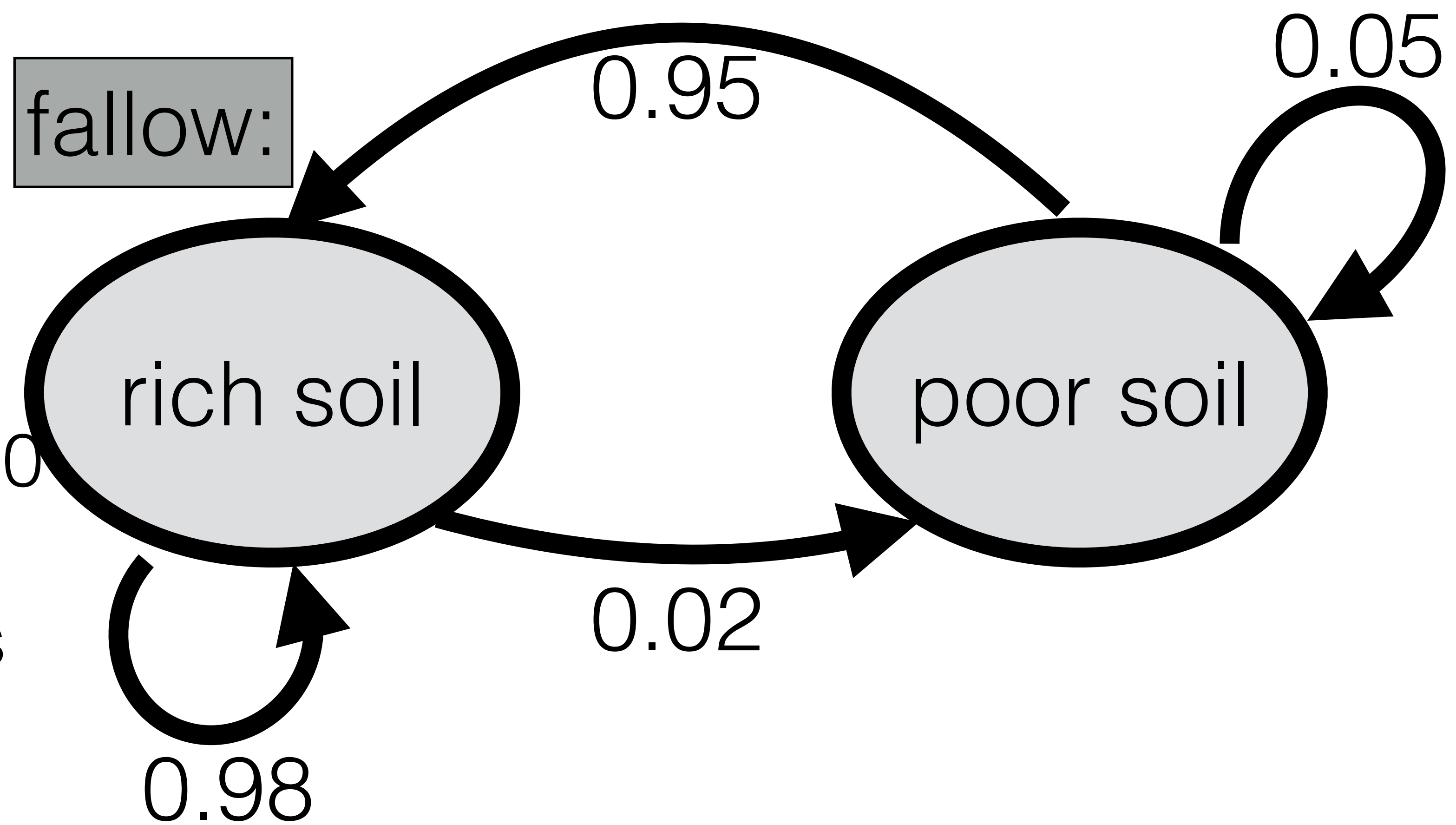




Markov Decision Process (basically)

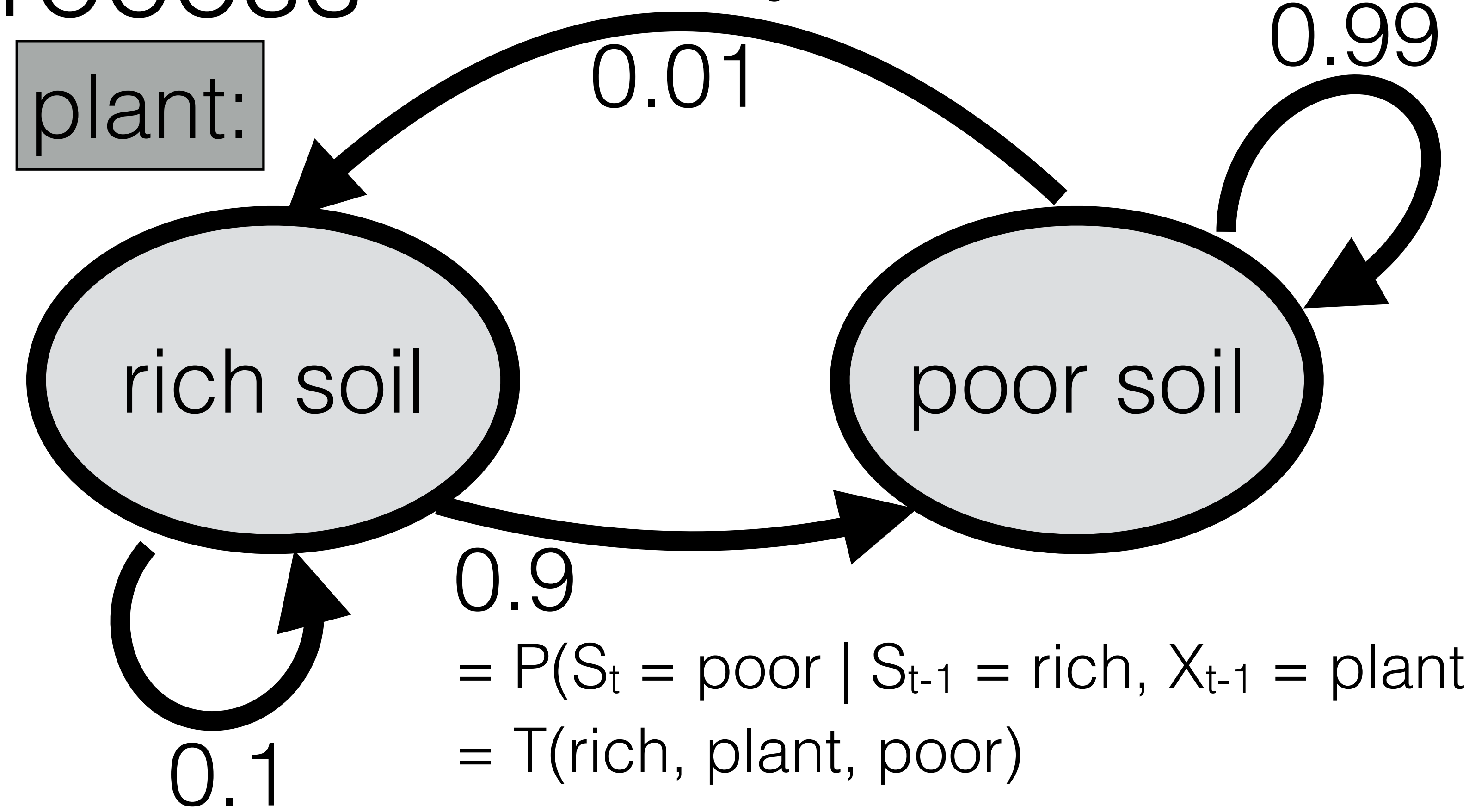
- $\mathcal{S}$ = set of possible states
- $\mathcal{X}$ = set of possible inputs
- $s_0 \in \mathcal{S}$: initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant)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

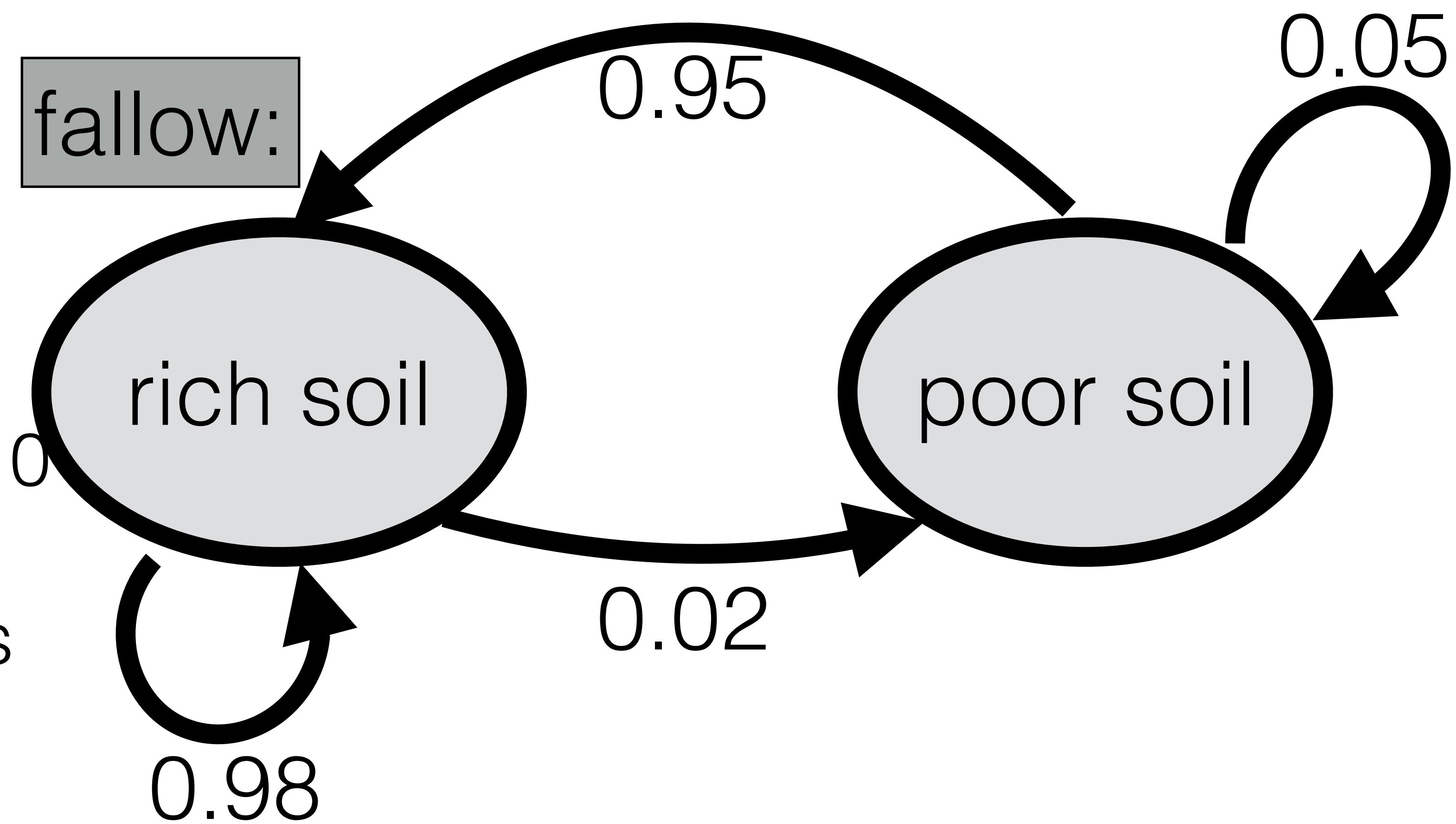




Markov Decision Process (basically)

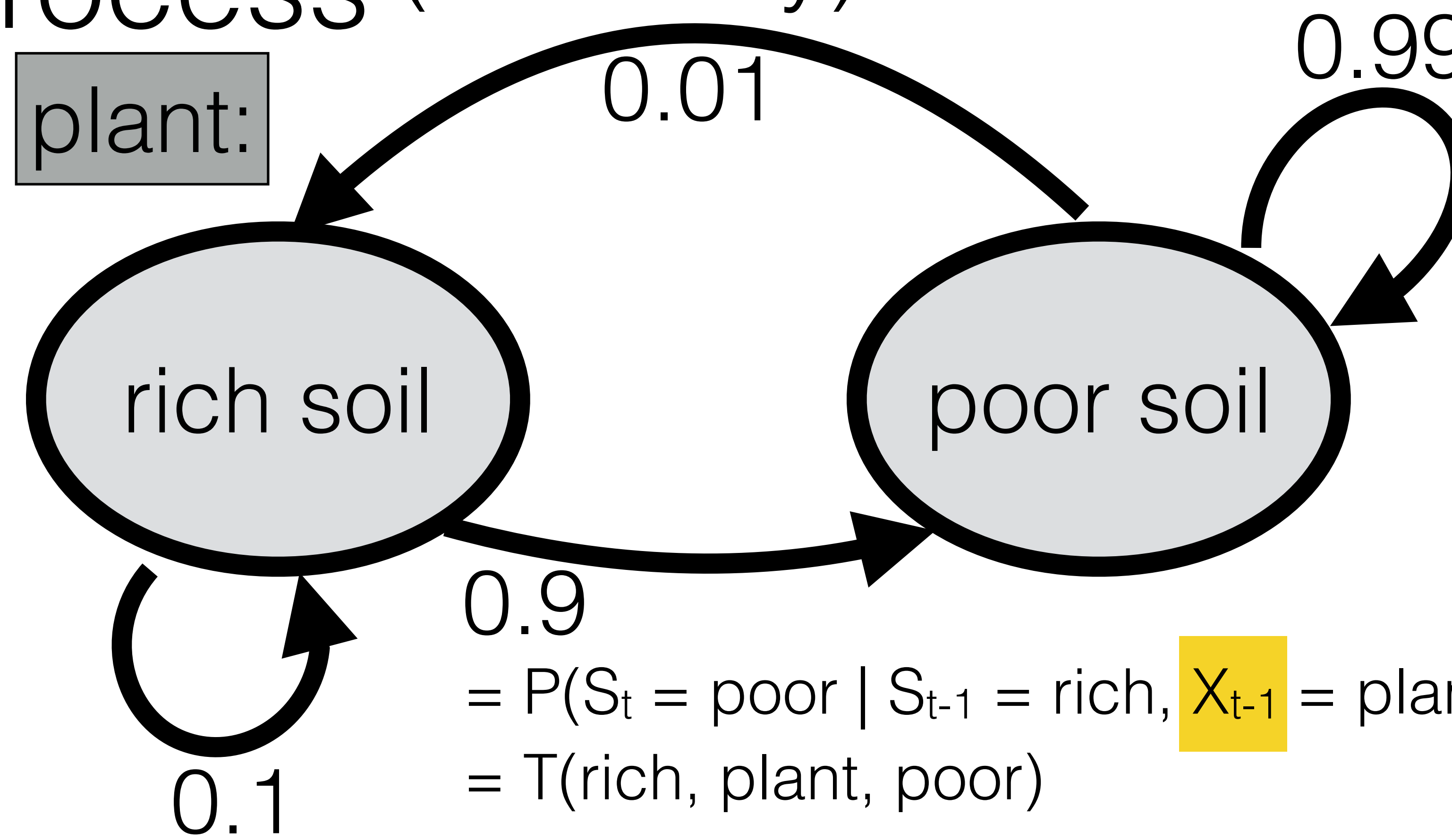
- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $s_0 \in \mathcal{S}$: initial state
- $T : \mathcal{S} \times \mathcal{X} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{X} \to \mathbb{R}$: reward function
    - e.g. R(rich,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





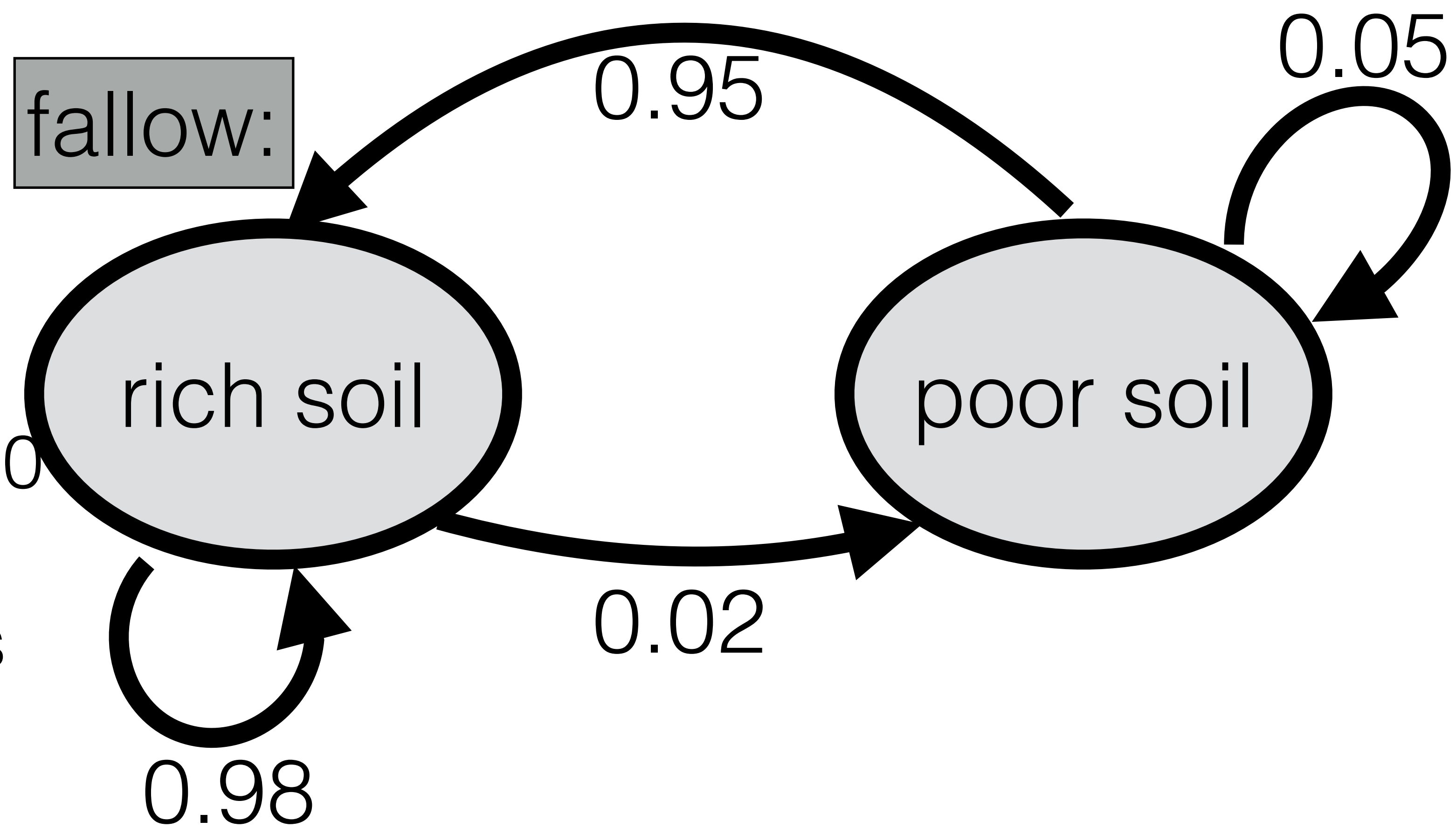
Markov Decision Process (basically)

- $\mathcal{S}$ = 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



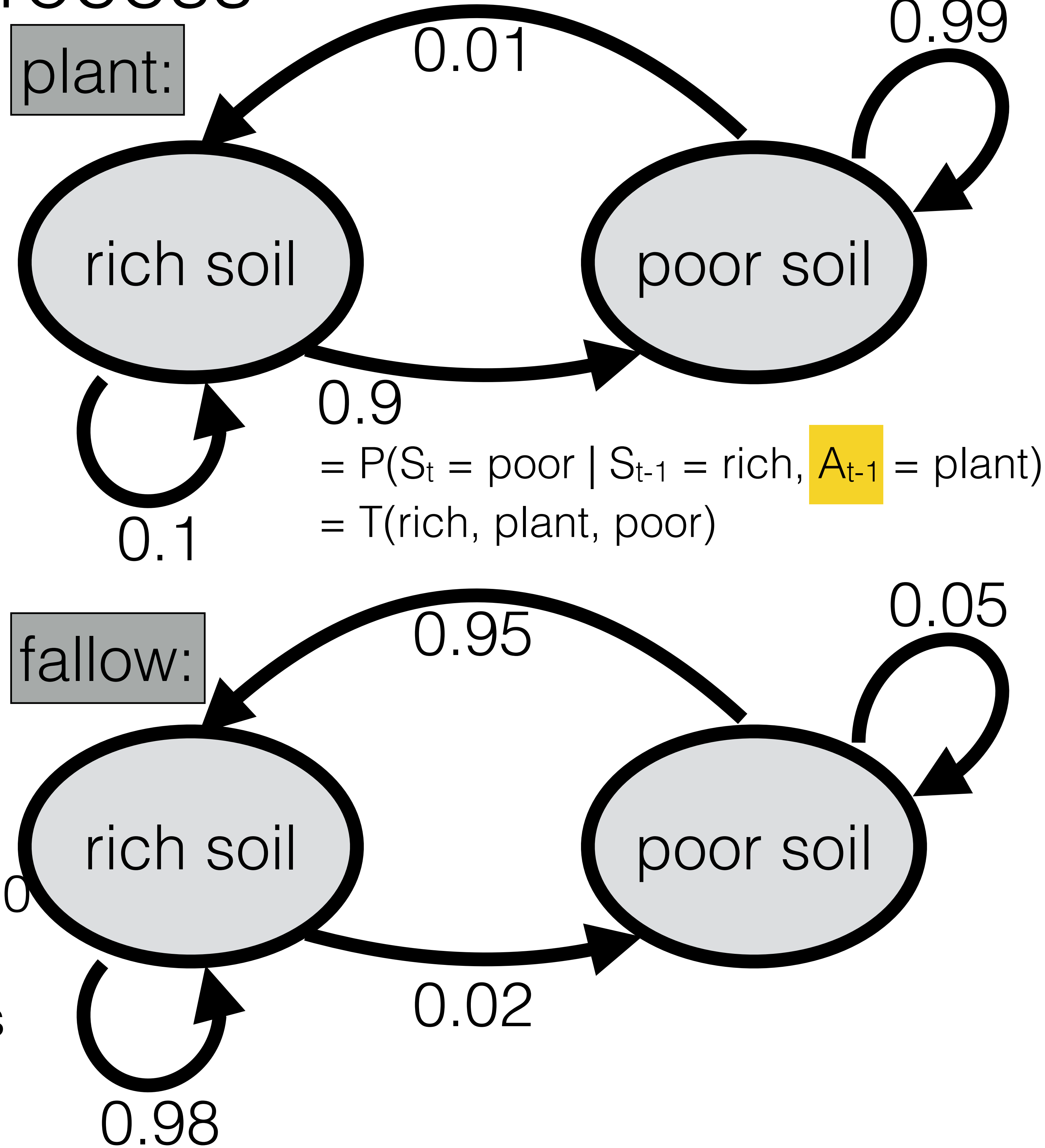
$= P(S_t = \text{poor} \mid S_{t-1} = \text{rich}, X_{t-1} = \text{plant})$

$= T(\text{rich, plant, poor})$



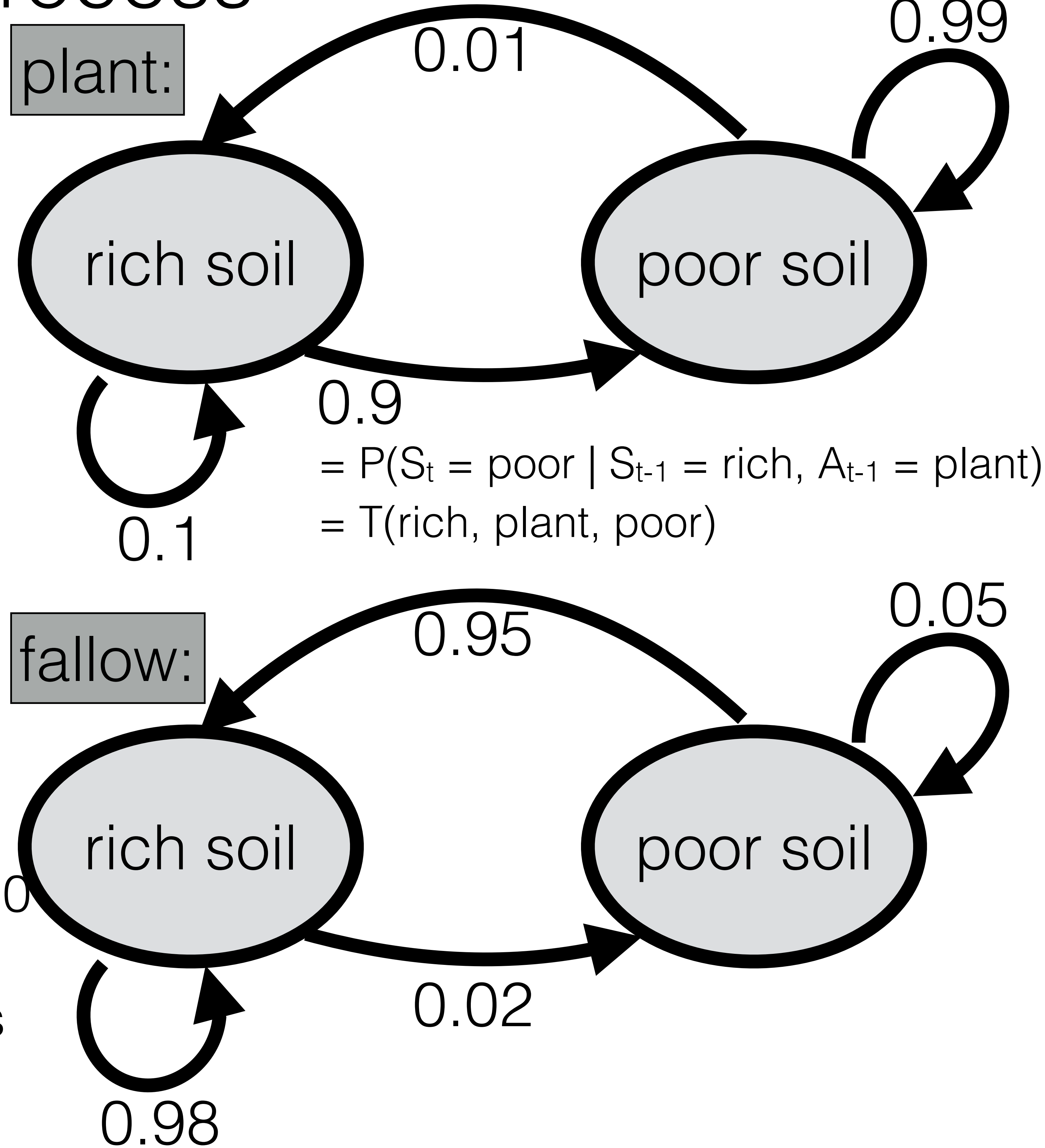
Markov Decision Process (basically)

- $\mathcal{S}$ = 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



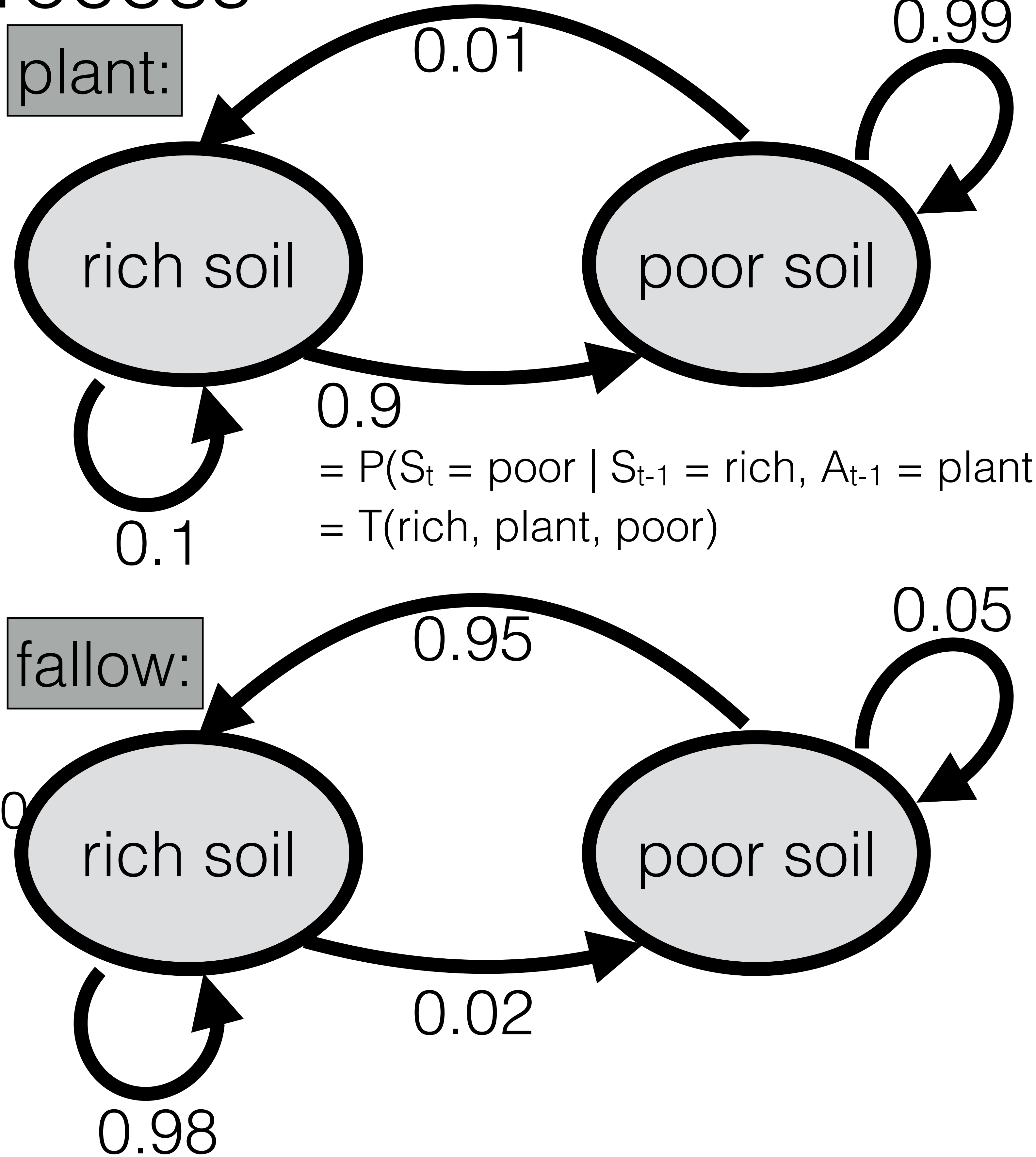
Markov Decision Process (basically)

- $\mathcal{S}$ = 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



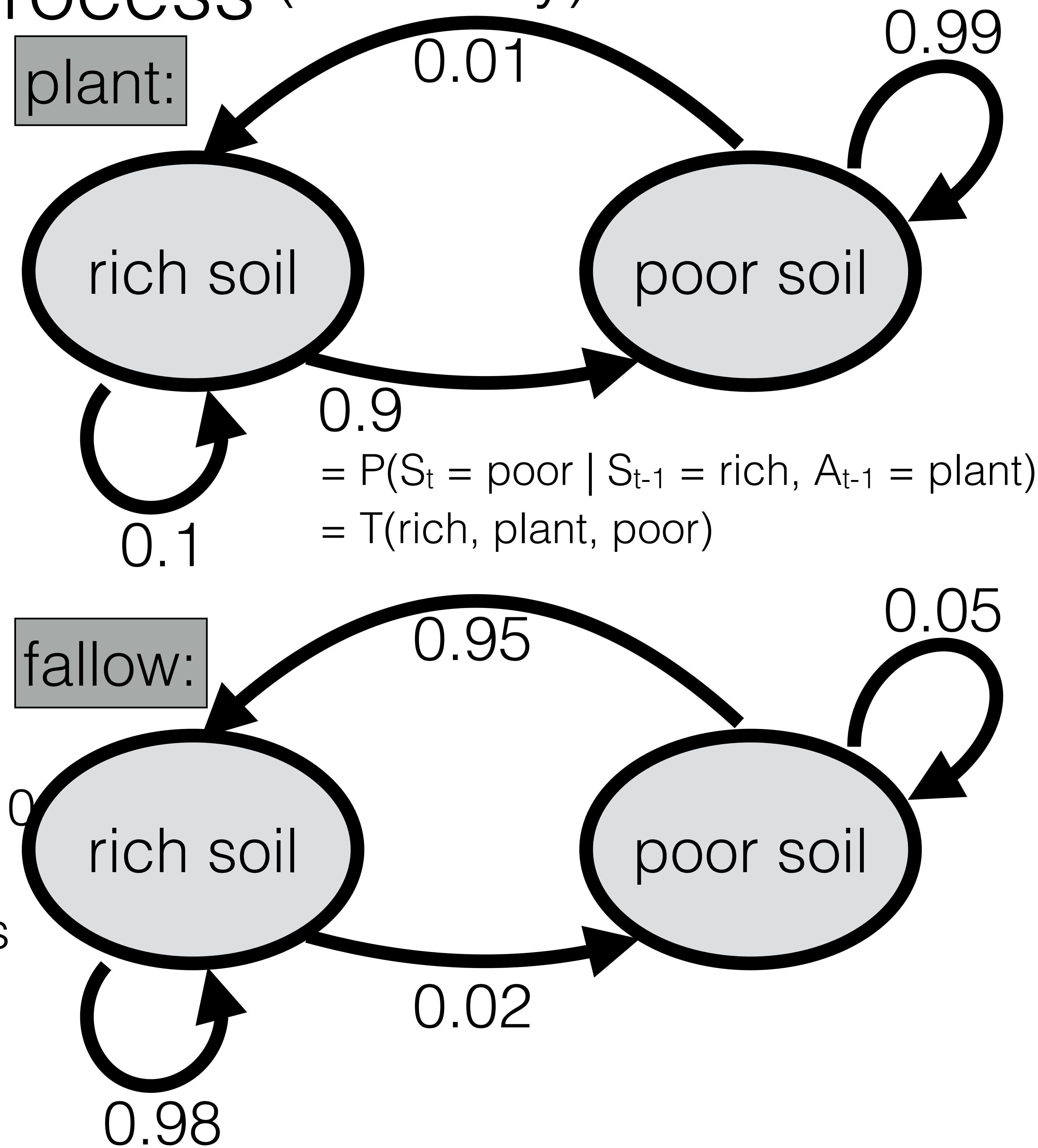
Markov Decision Process (basically)

- $\mathcal{S}$ = 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



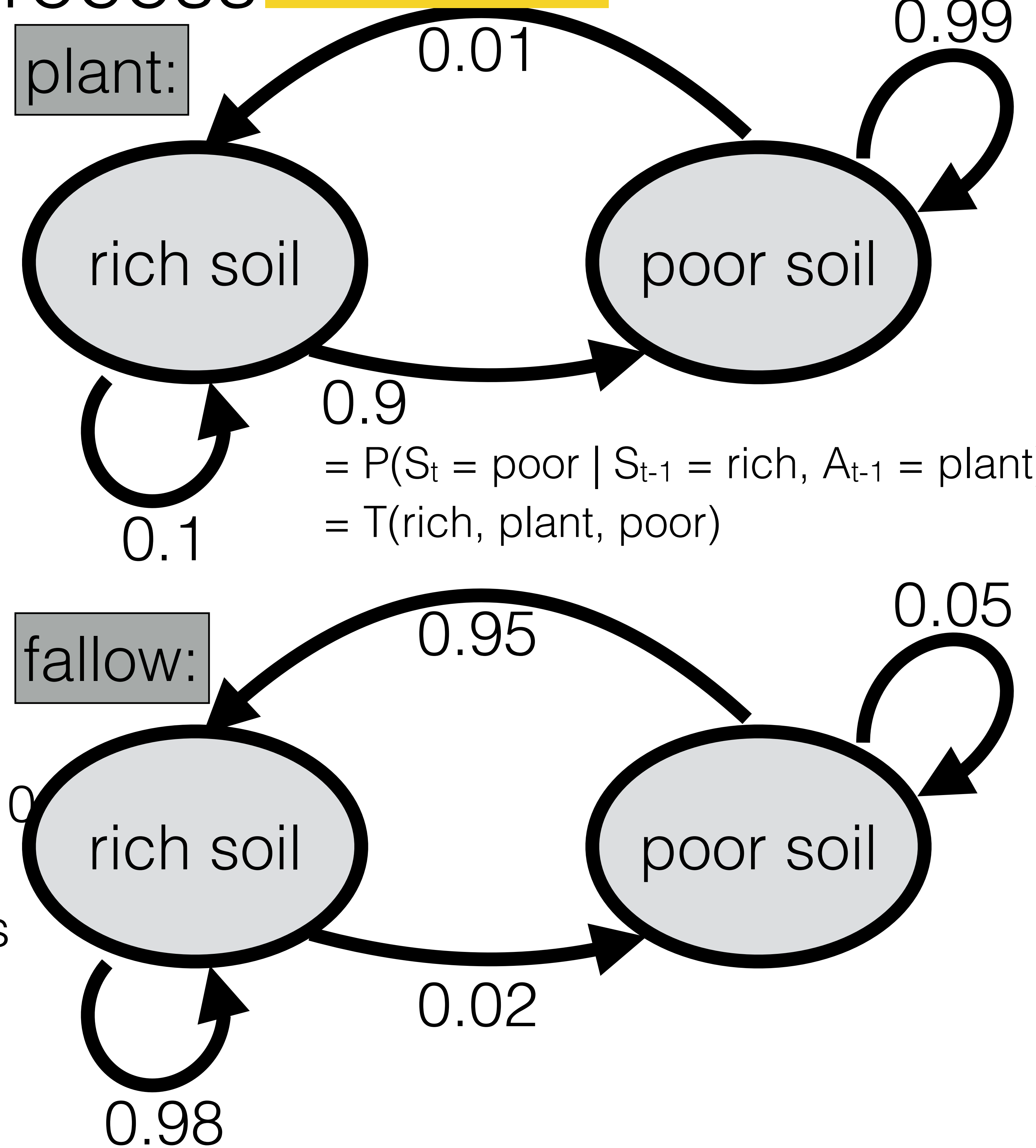
Markov Decision Process (basically)

- $\mathcal{S}$ = 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



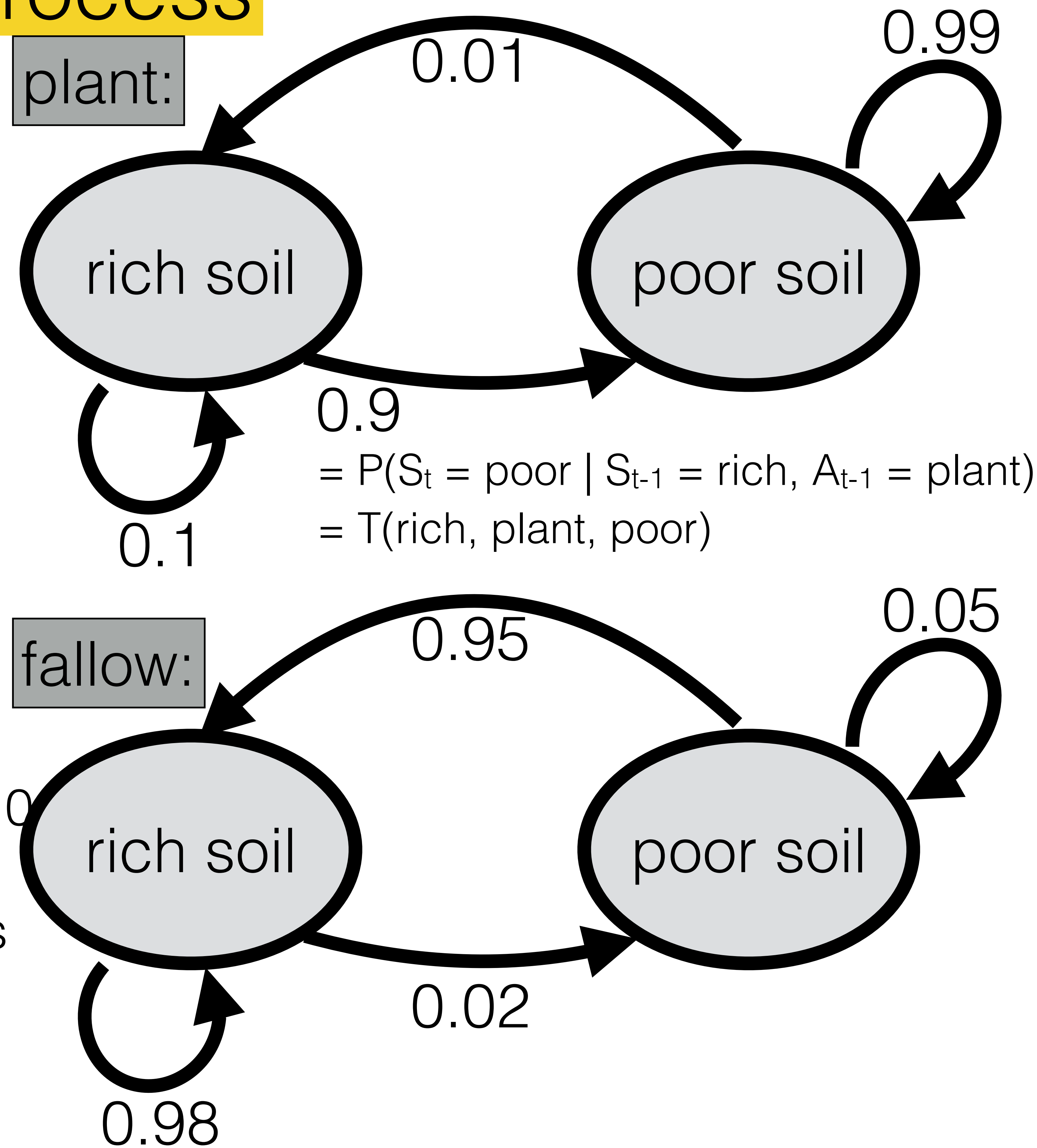
Markov Decision Process (basically)

- $\mathcal{S}$ = 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



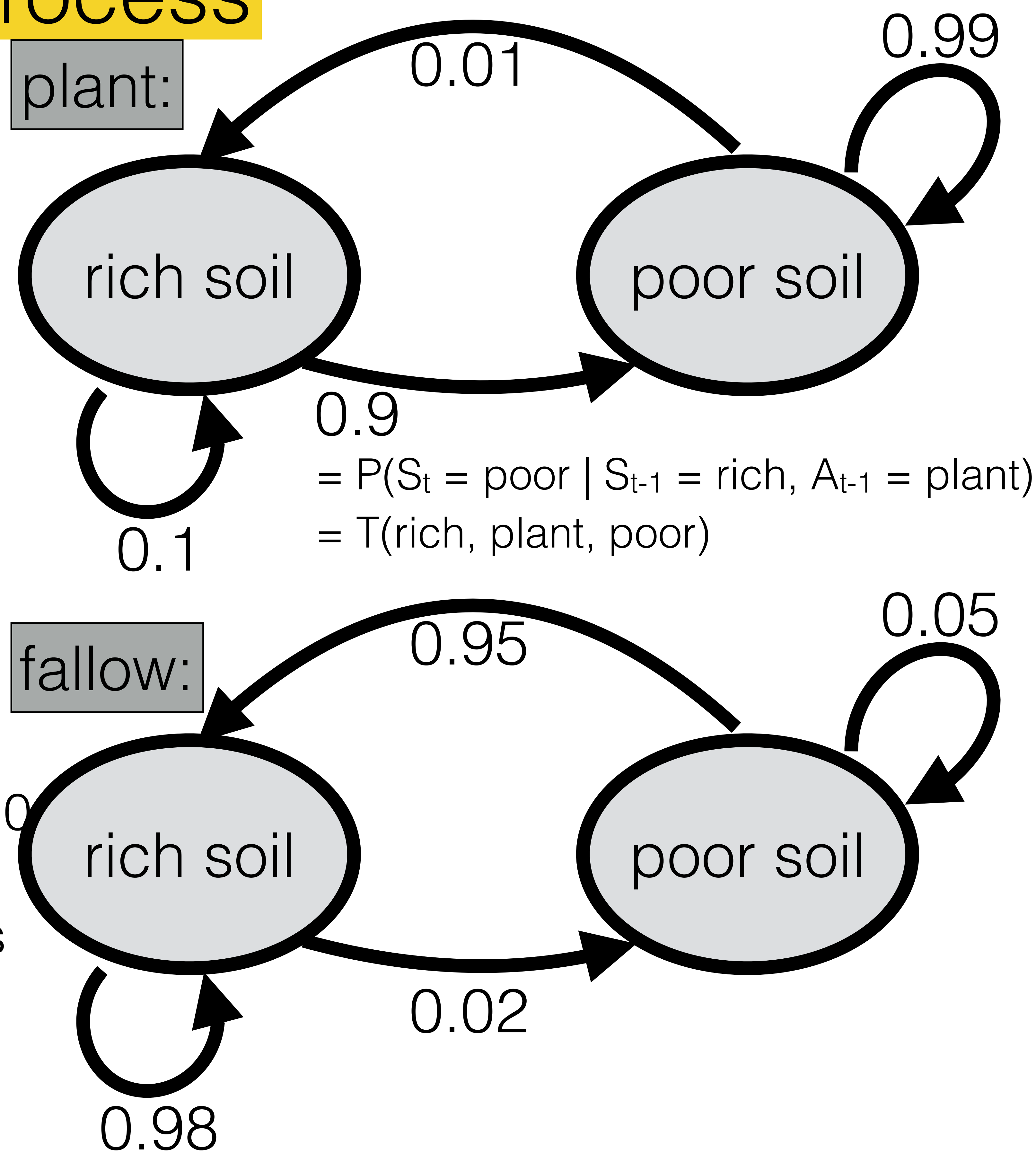
Markov Decision Process

- $\mathcal{S}$ = 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



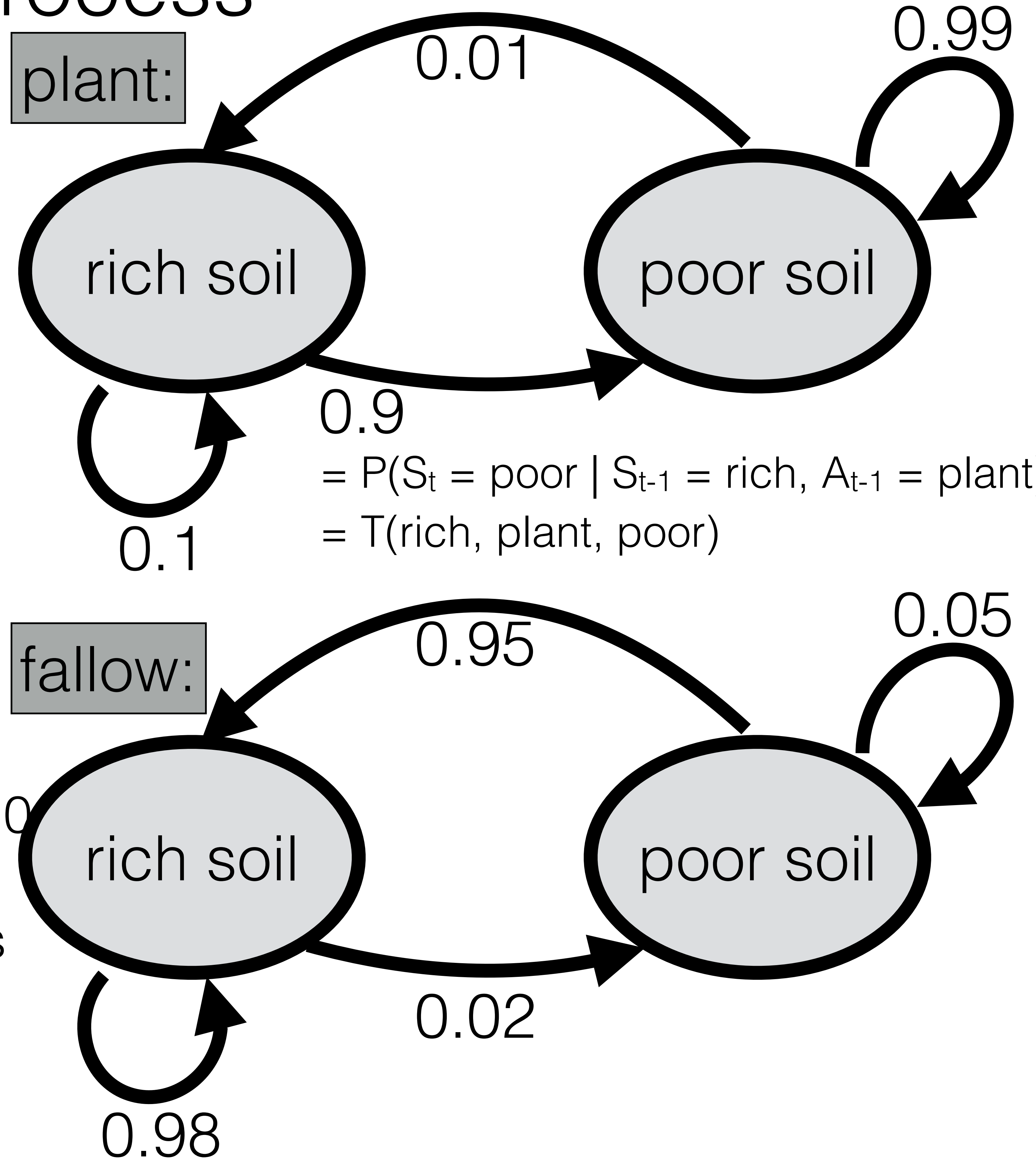
Markov Decision Process

- $\mathcal{S}$ = 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



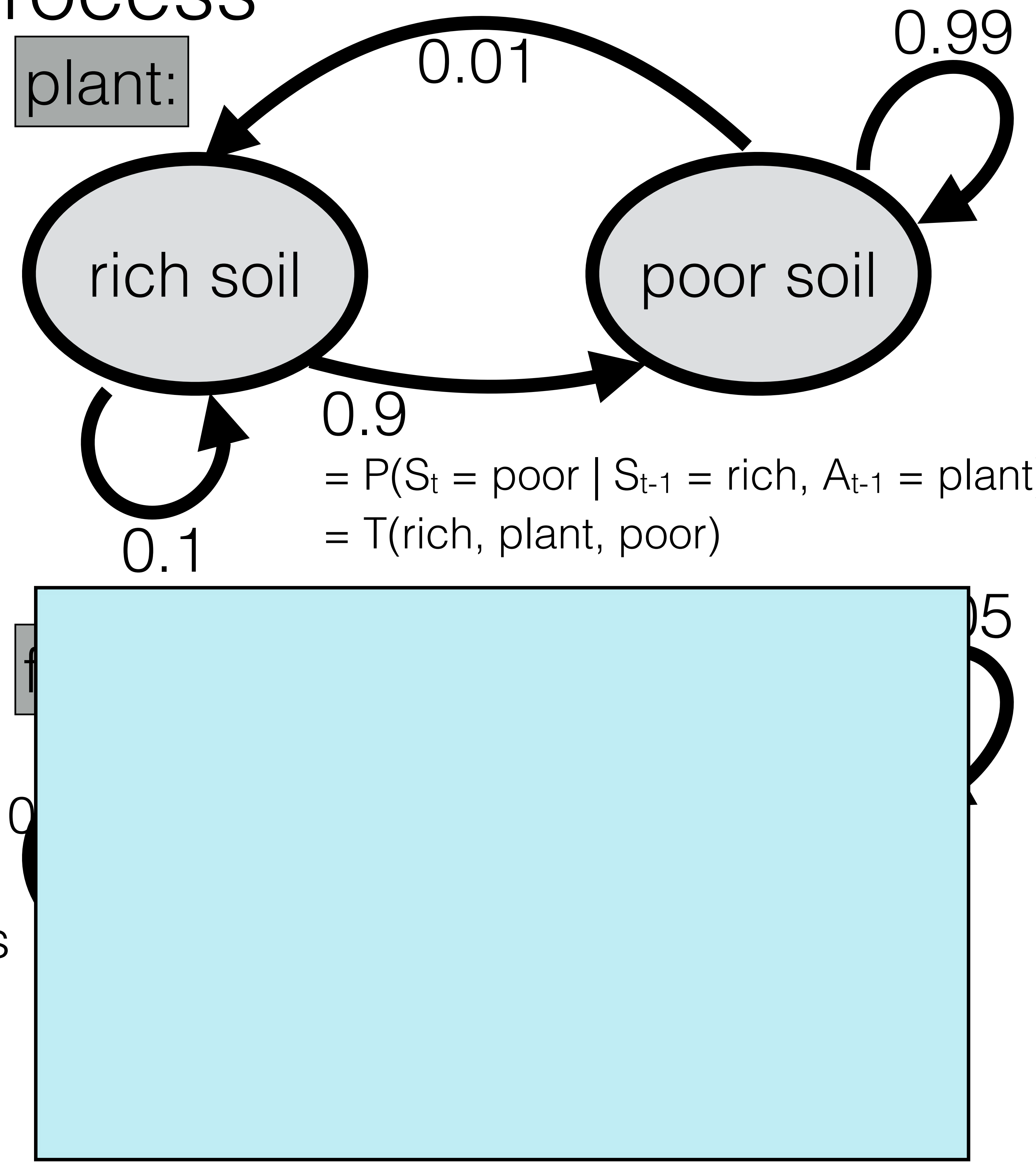
Markov Decision Process

- $\mathcal{S}$ = 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



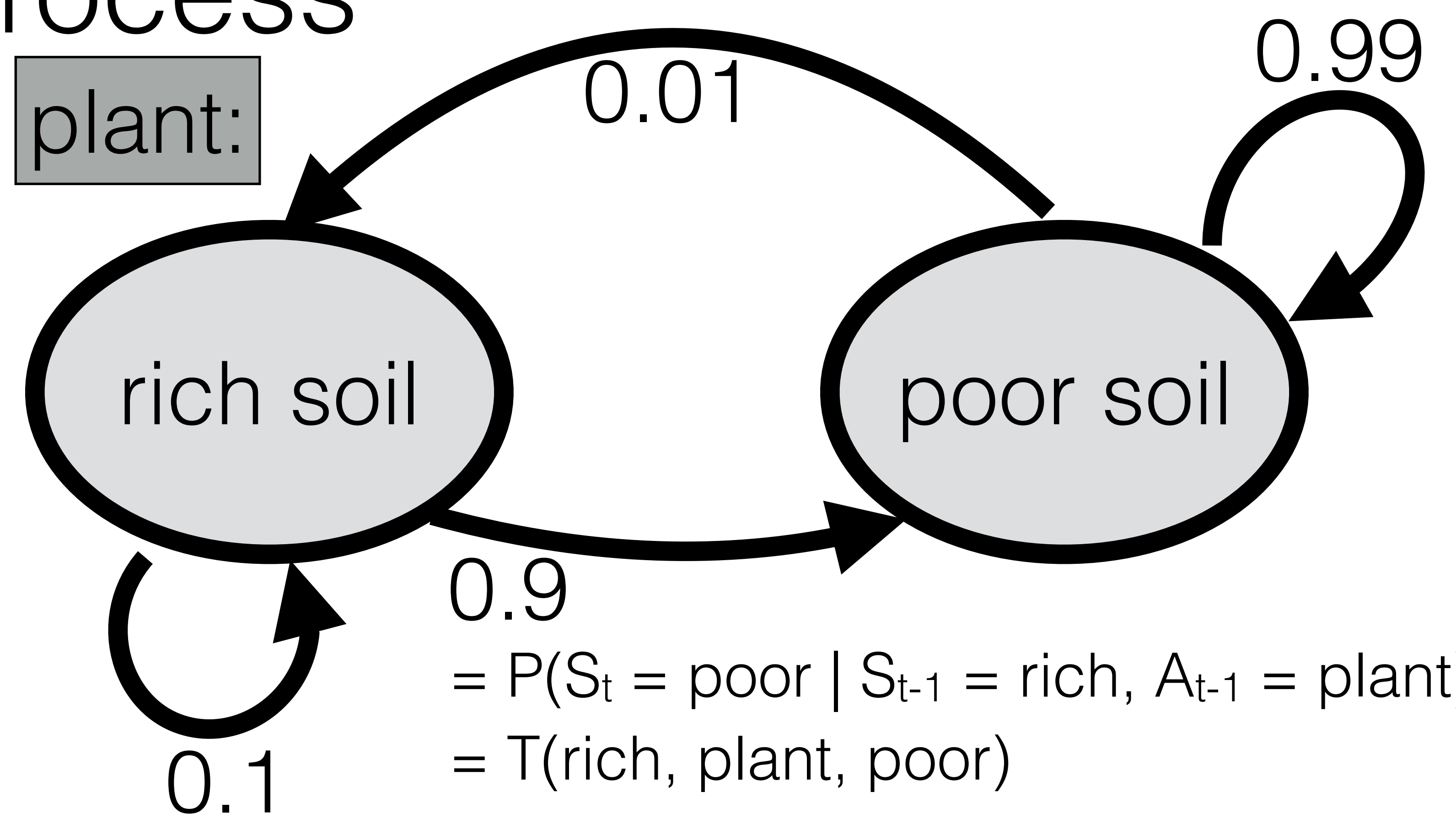
Markov Decision Process

- $\mathcal{S}$ = 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

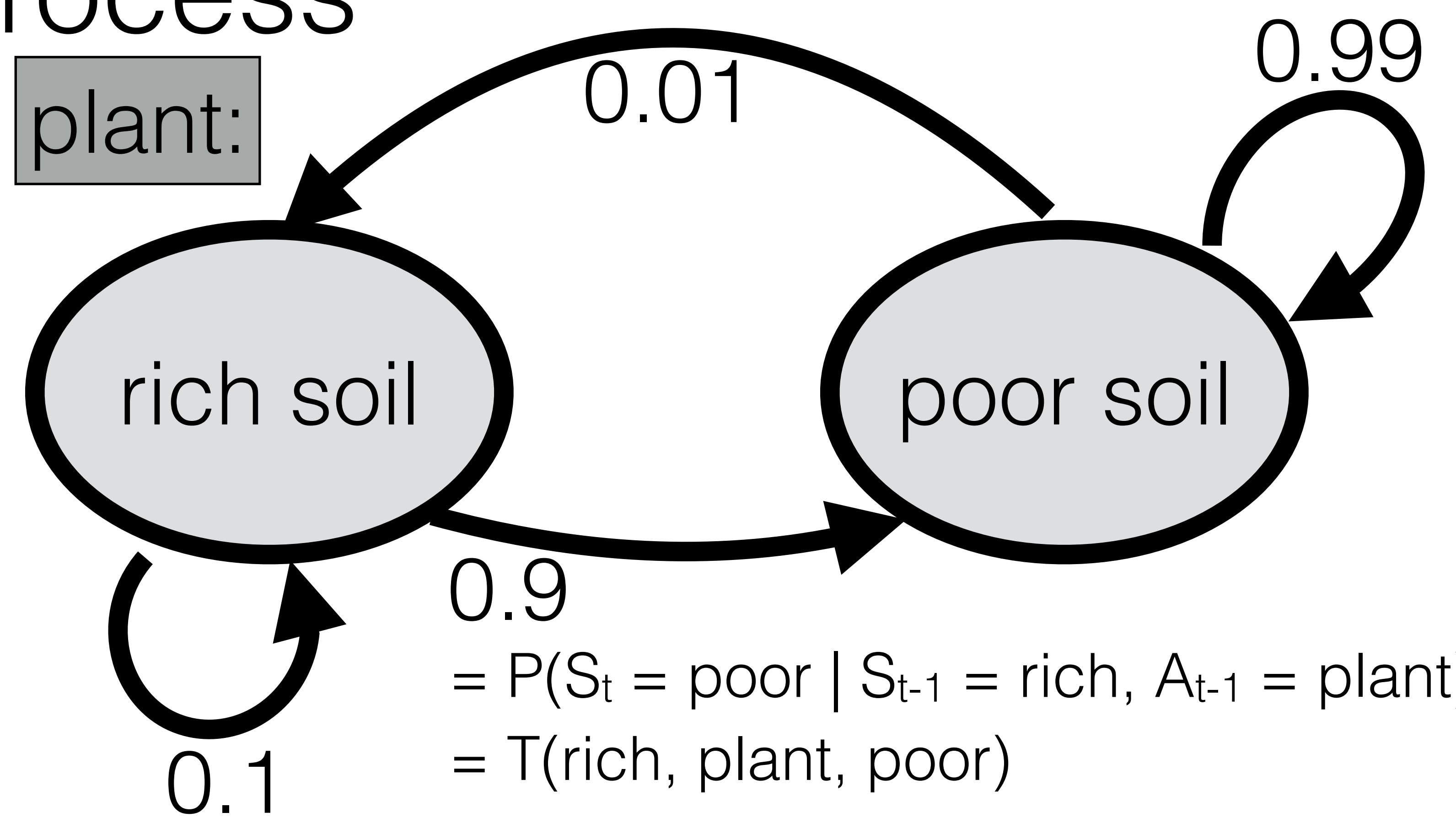
- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) =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

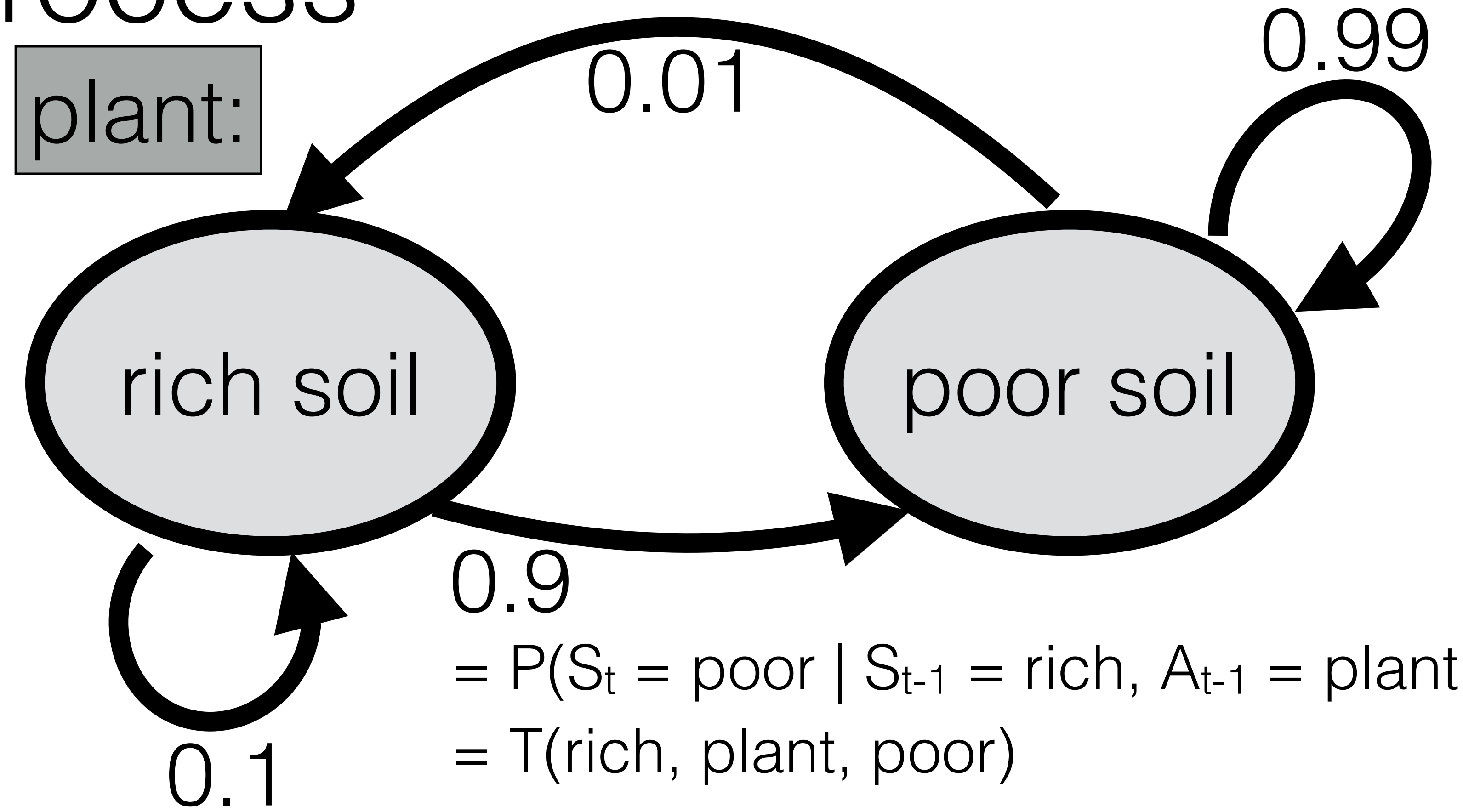
- $\mathcal{S}$ = set of possible states
- $\mathcal{A}$ = set of possible actions
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$: transition model
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: reward function
  - e.g. R(rich, plant) = 100 bushels; R(poor, plant) =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

Expectation

Expectation

- Suppose a random variable *R* has *m* possible values: $r_1, \ldots, r_m$

Expectation

- Suppose a random variable *R* has *m* possible values: $r_1, \ldots, r_m$
  - Example: a lottery pays $r_1 =$ 40*10^6 USD if you win and $r_2 =$ -2 USD if you lose.

Expectation

- Suppose a random variable *R* has *m* possible values: $r_1, \ldots, r_m$
  - Example: a lottery pays $r_1 =$ 40*10^6 USD if you win and $r_2 =$ -2 USD if you lose.
  - Question: if I could play this lottery a limitless number of times, how much could I expect to make each time I play, on average?

Expectation

- Suppose a random variable *R* has *m* possible values: $r_1, \ldots, r_m$
  - Example: a lottery pays $r_1 =$ 40*10^6 USD if you win and $r_2 =$ -2 USD if you lose.
  - Question: if I could play this lottery a limitless number of times, how much could I expect to make each time I play, on average?
- Suppose $R = r_i$ with probability p_i

Expectation

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

Expectation

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

Expectation

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

Expectation

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

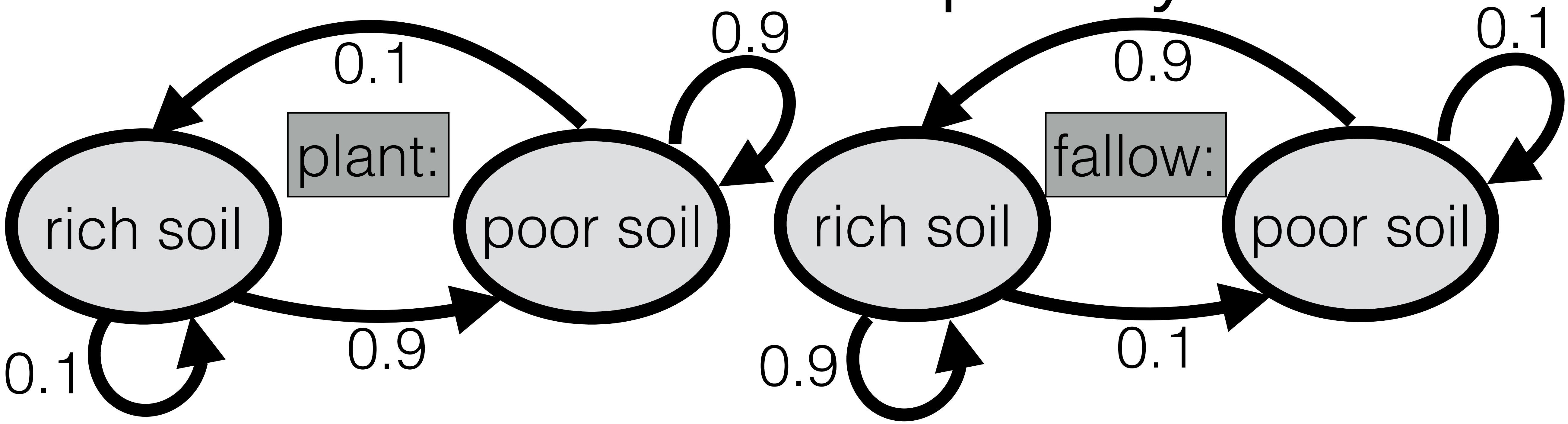
Expectation

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

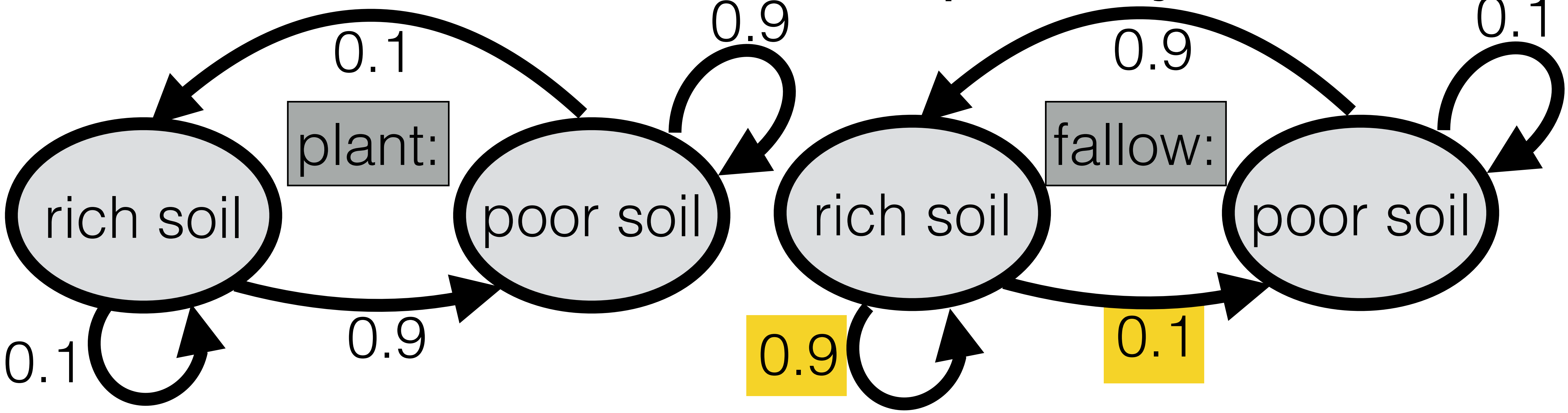
    = -1.86 USD

What's the value of a policy?

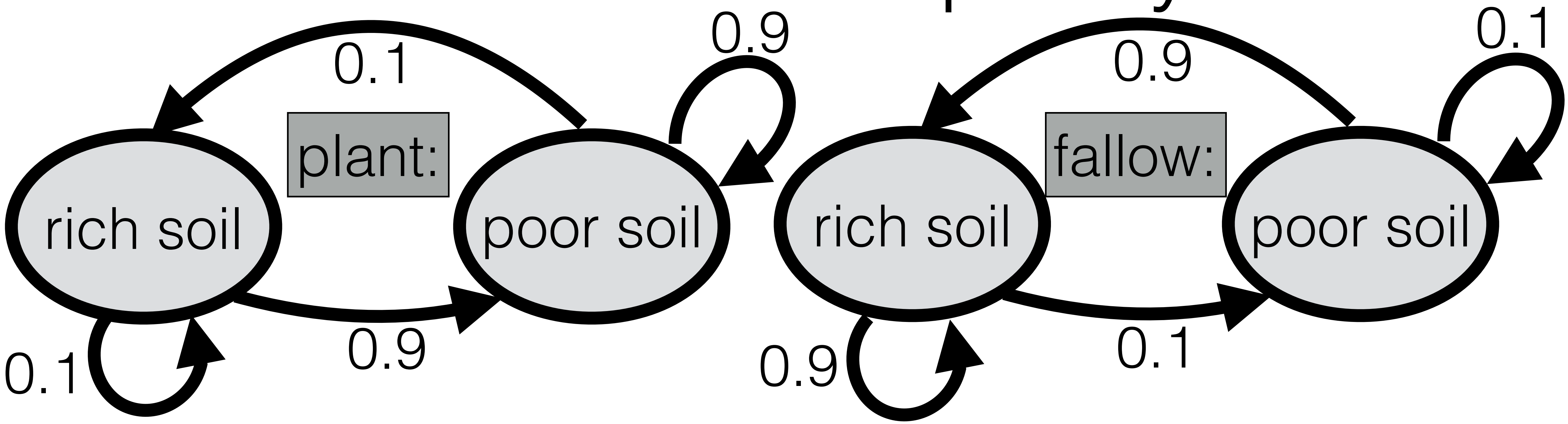
What's the value of a policy?



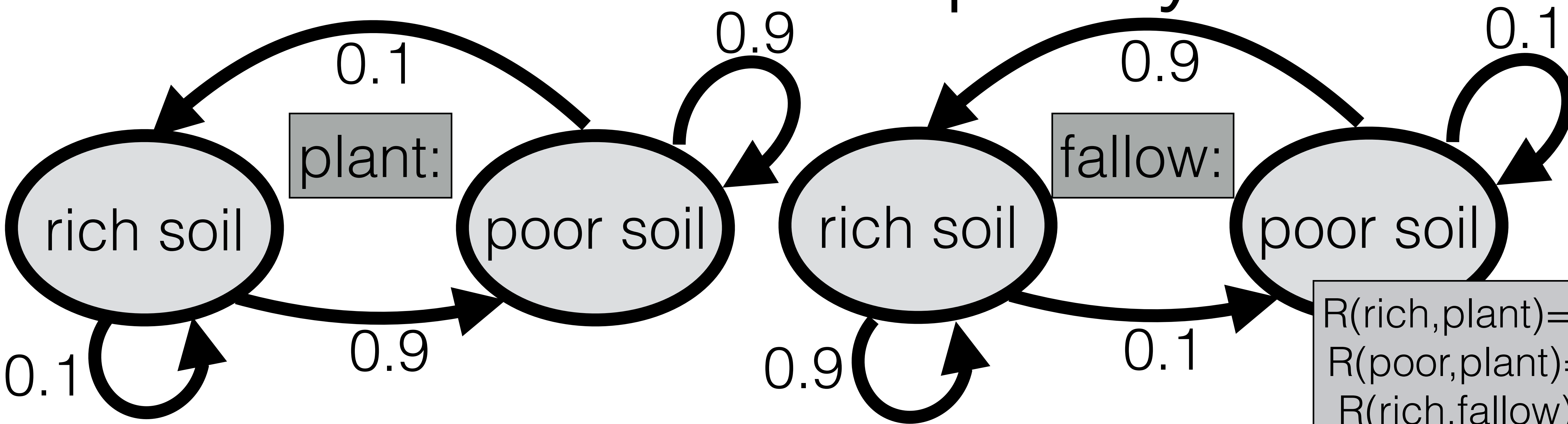
What's the value of a policy?



What's the value of a policy?

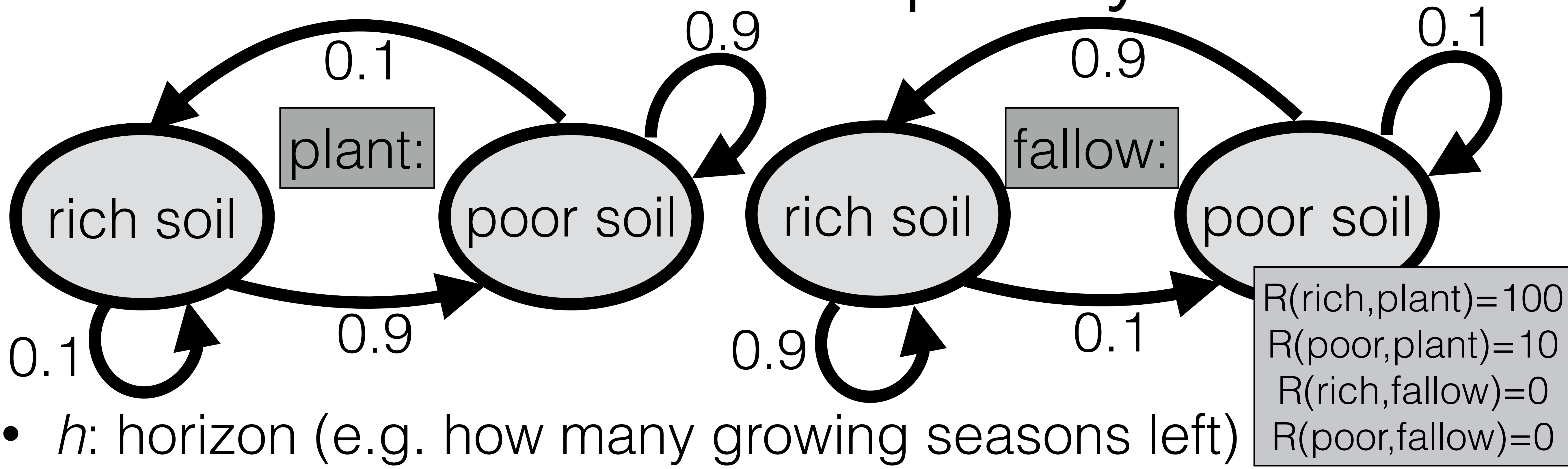


What's the value of a policy?



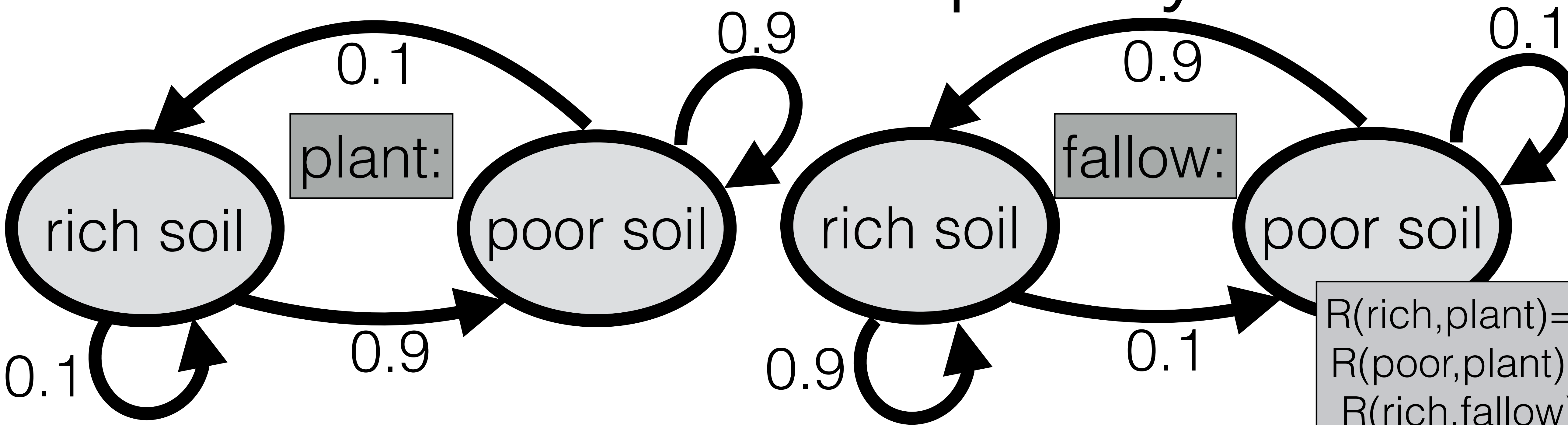
R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

What's the value of a policy?



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

What's the value of a policy?

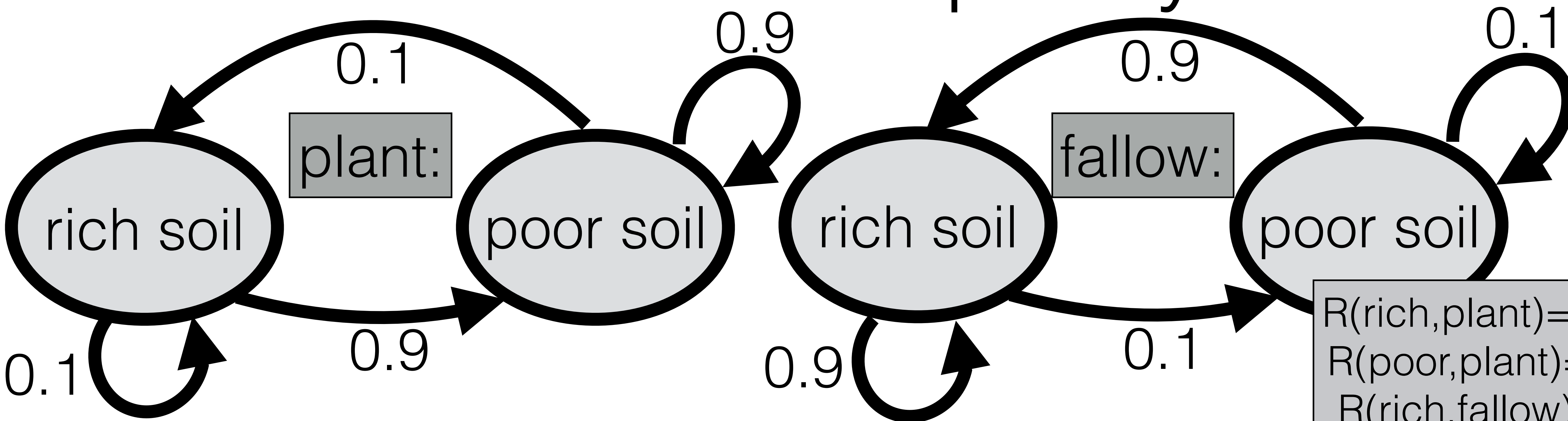


R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

- *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.

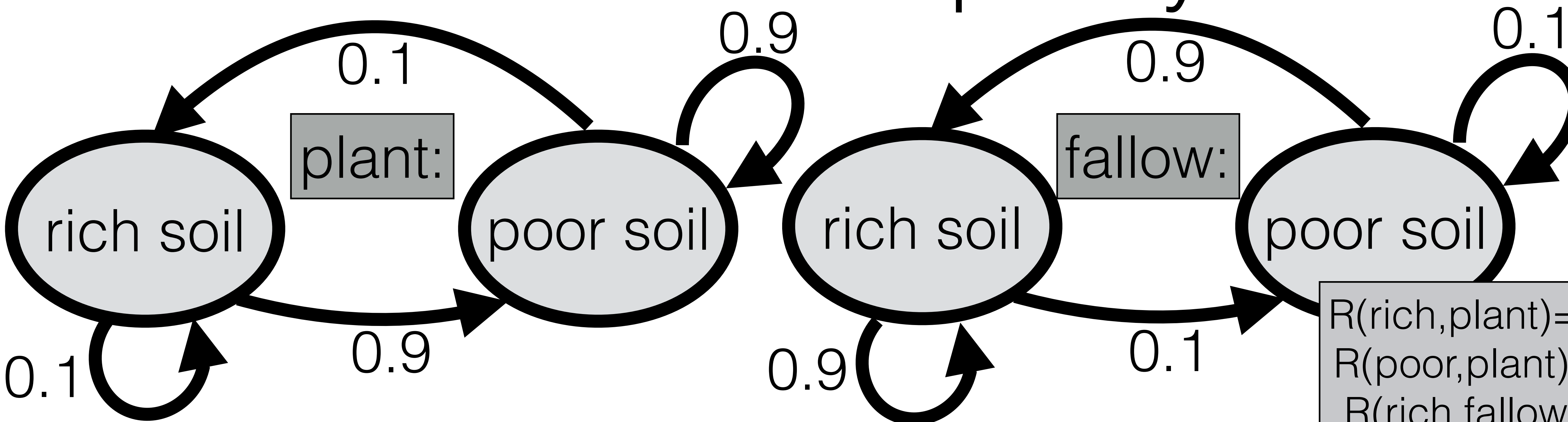
What's the value of a policy?



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

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

What's the value of a policy?

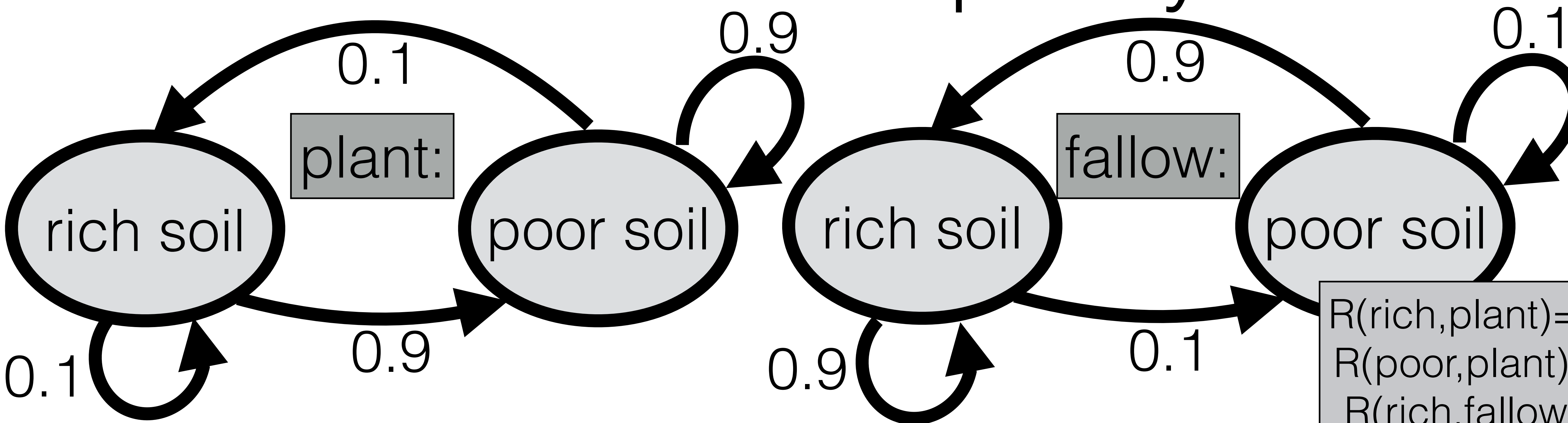


R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

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

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

What's the value of a policy?



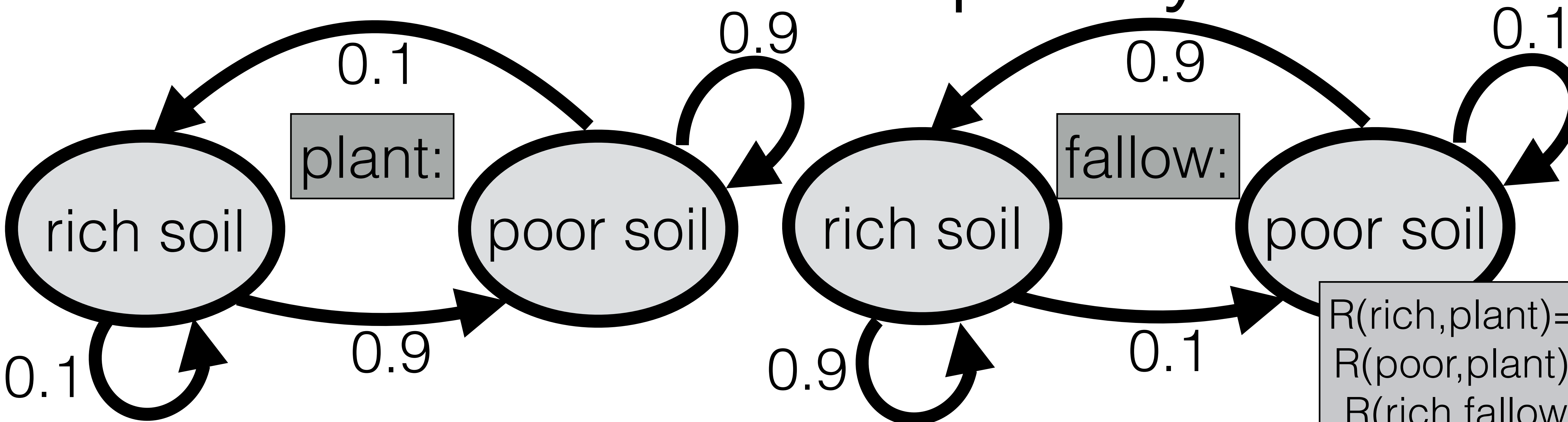
R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

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

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

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

What's the value of a policy?



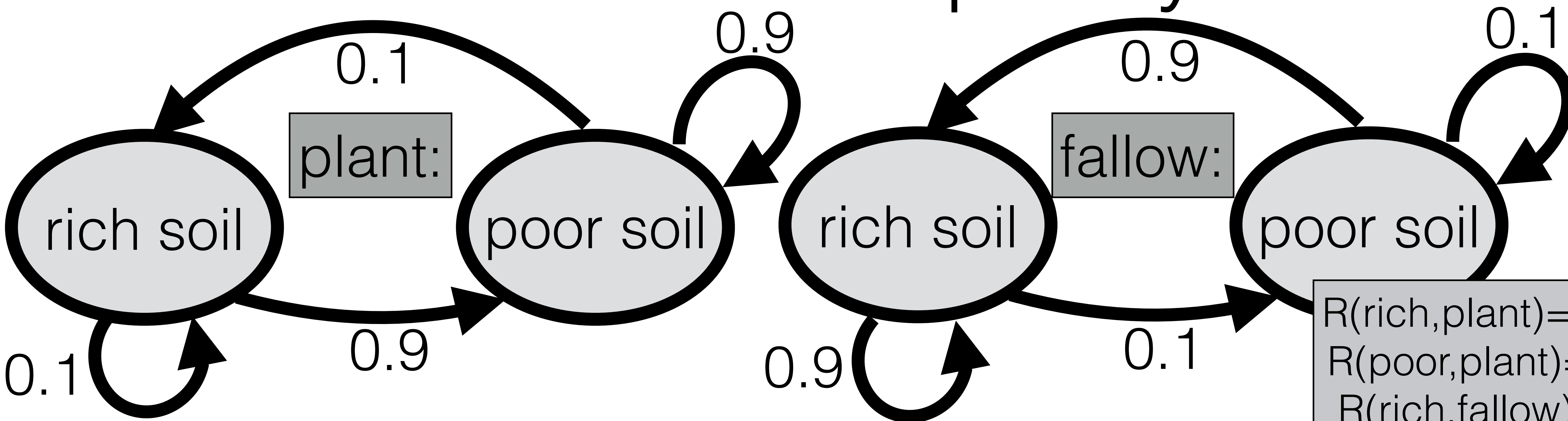
R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

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

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

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

What's the value of a policy?



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

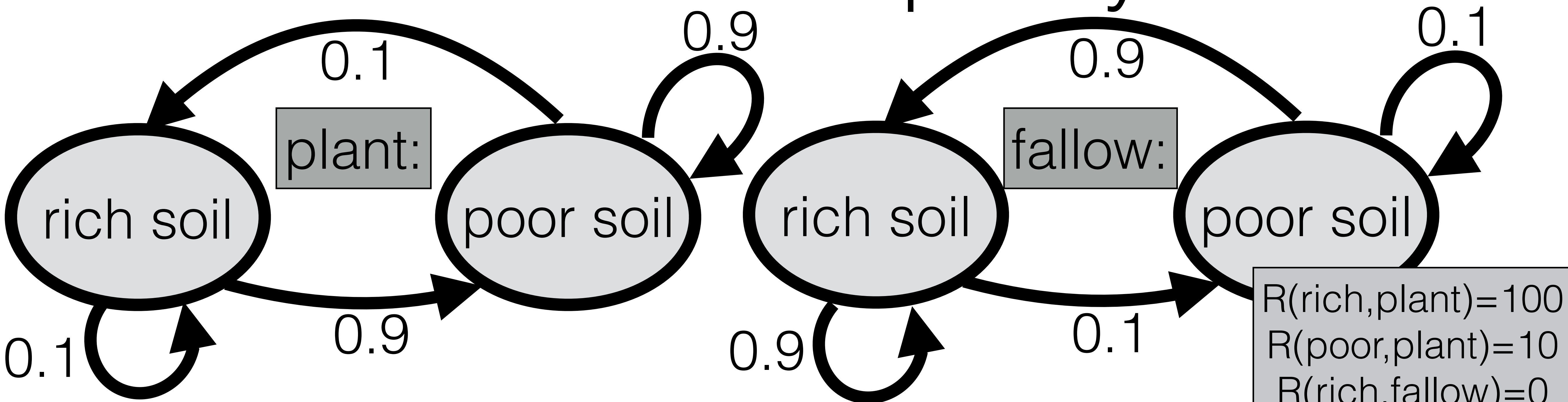
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy π starting at *s*

Dueling farmers! π_A : always plant; π_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?



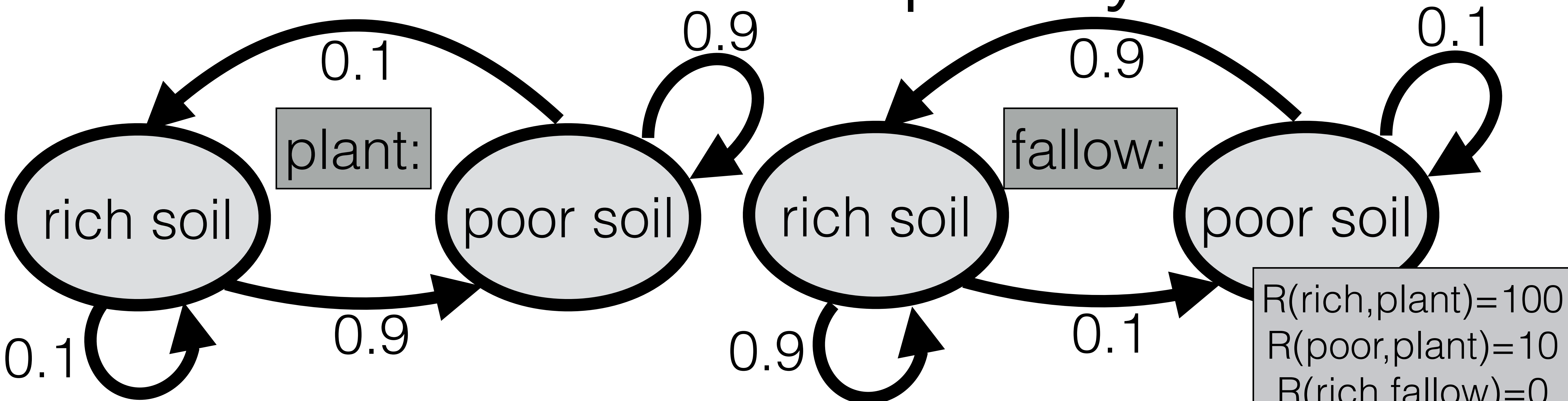
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy π starting at *s*

Dueling farmers! π_A : always plant; π_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?



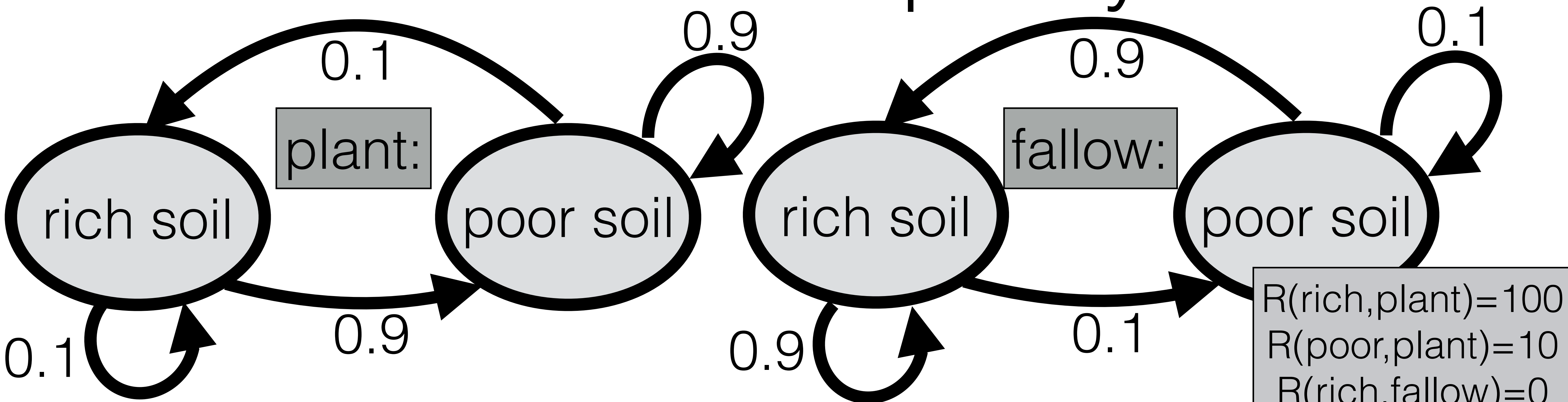
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy π starting at *s*

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

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

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

What's the value of a policy?



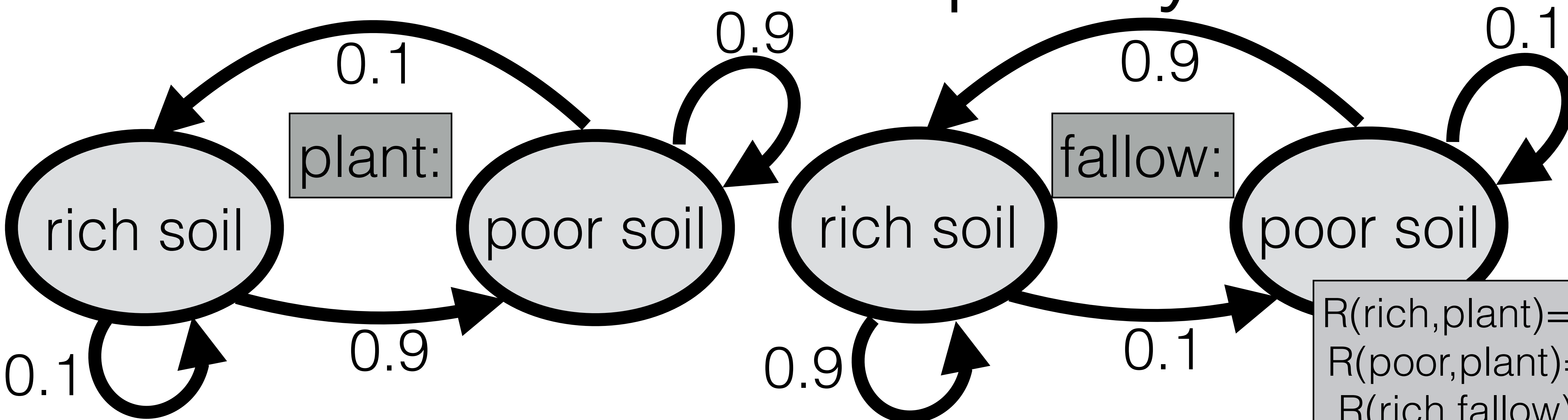
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy π starting at *s*

Dueling farmers! π_A : always plant; π_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

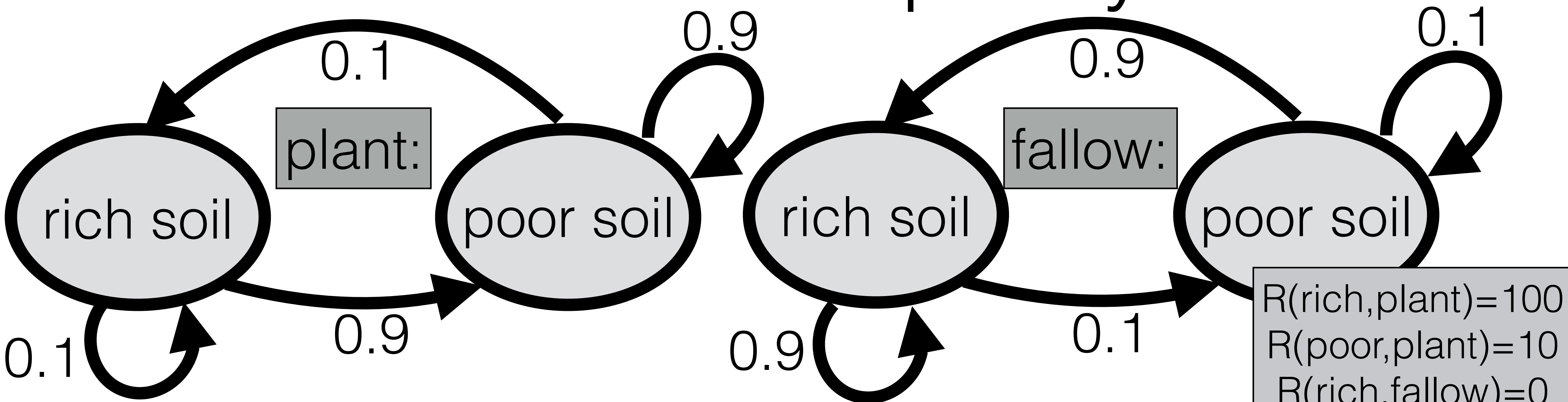
- *h*: horizon (e.g. how many growing seasons left)
- $V_{\pi}^{h}(s)$: value (expected reward) with policy π starting at *s*

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

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

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

What's the value of a policy?



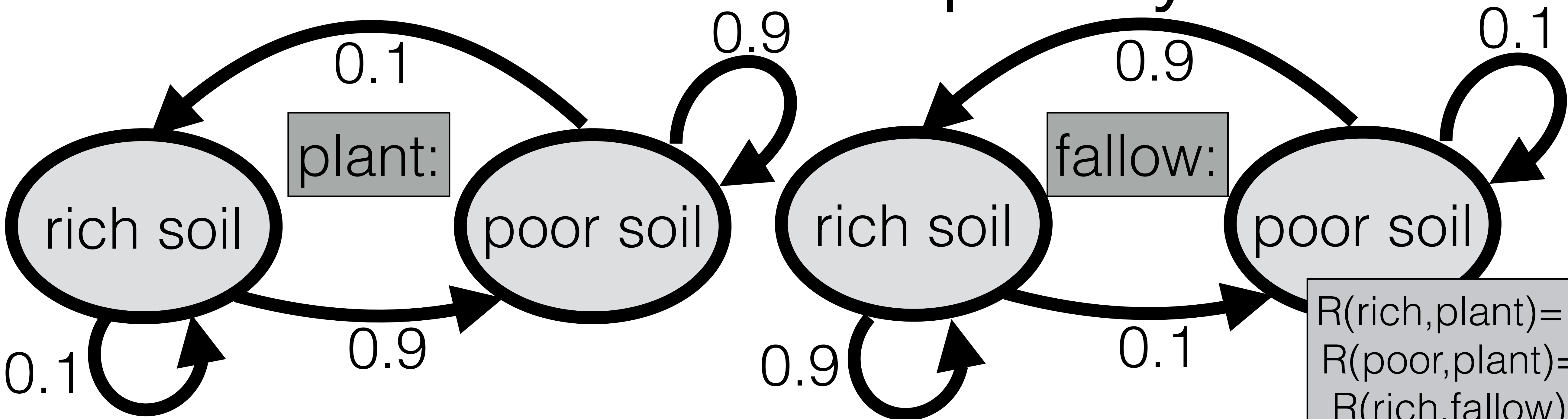
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy π starting at *s*

Dueling farmers! π_A : always plant; π_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

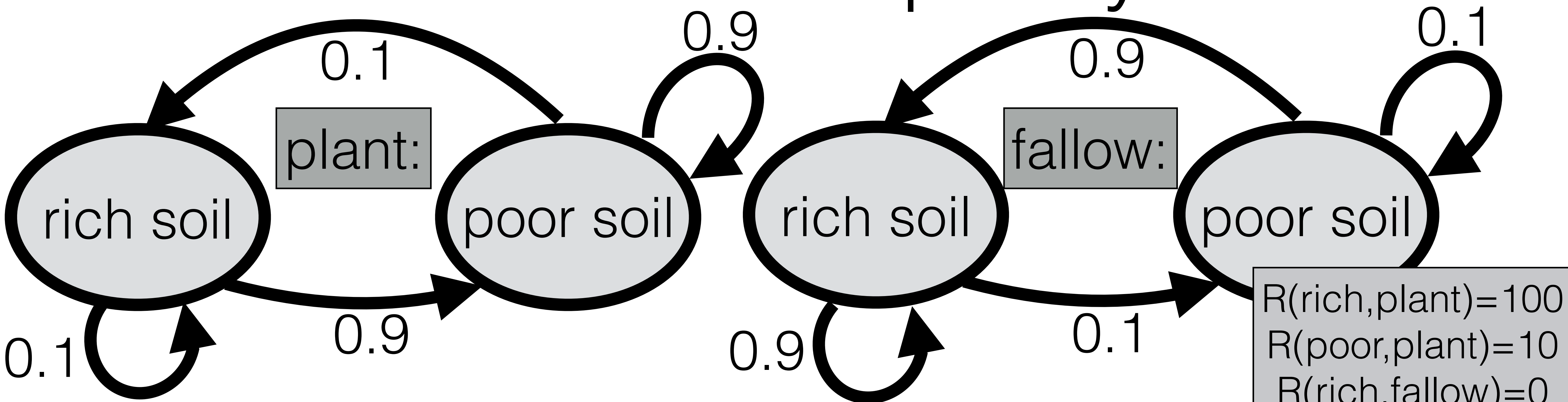
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy π starting at *s*

Dueling farmers! π_A : always plant; π_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?



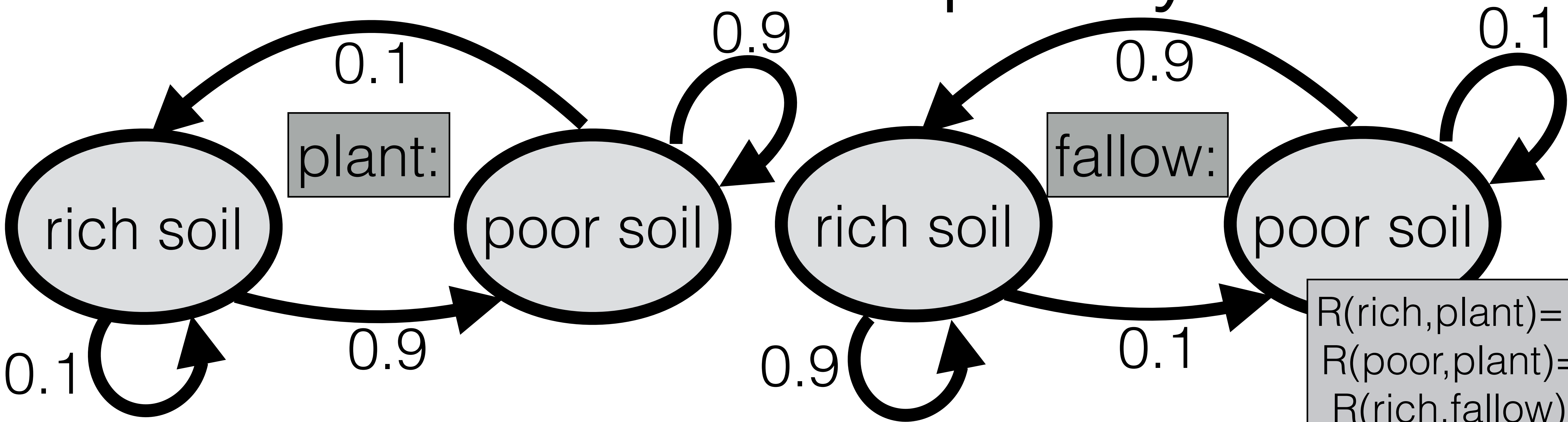
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy π starting at *s*

Dueling farmers! π_A : always plant; π_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

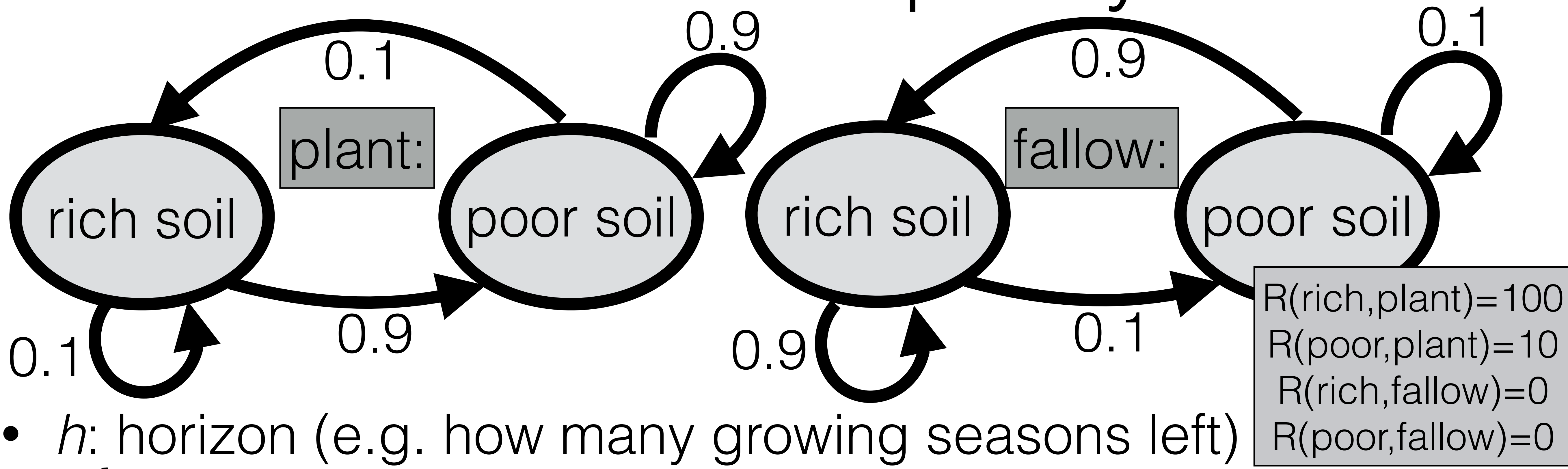
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy π starting at *s*

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

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

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

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 π starting at *s*

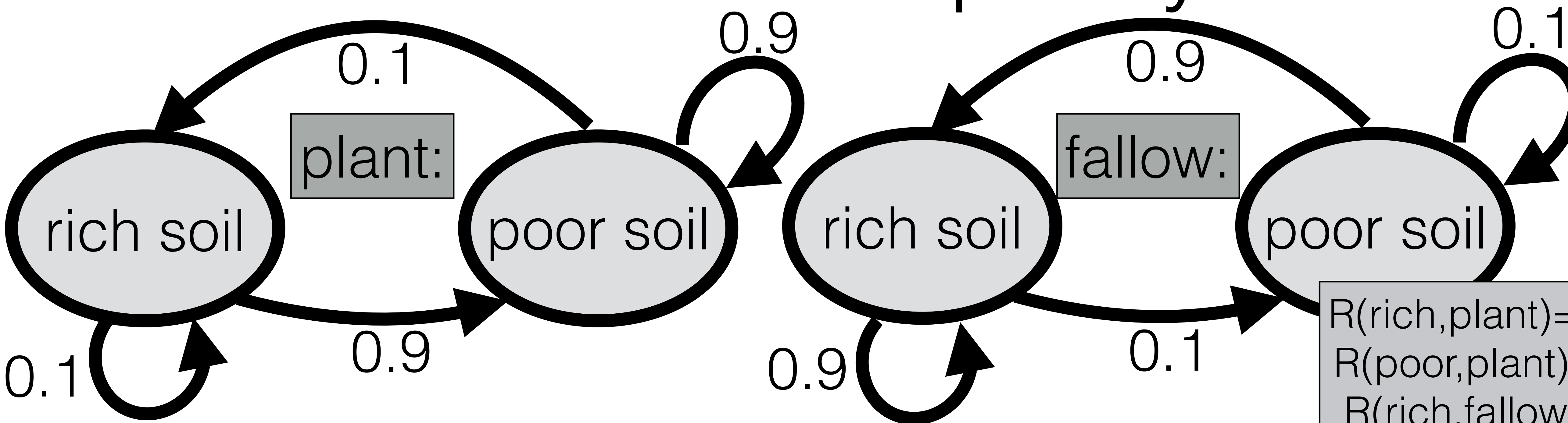
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

value of the policy with *h* steps 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)
- $V_\pi^h(s)$: value (expected reward) with policy π starting at *s*

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

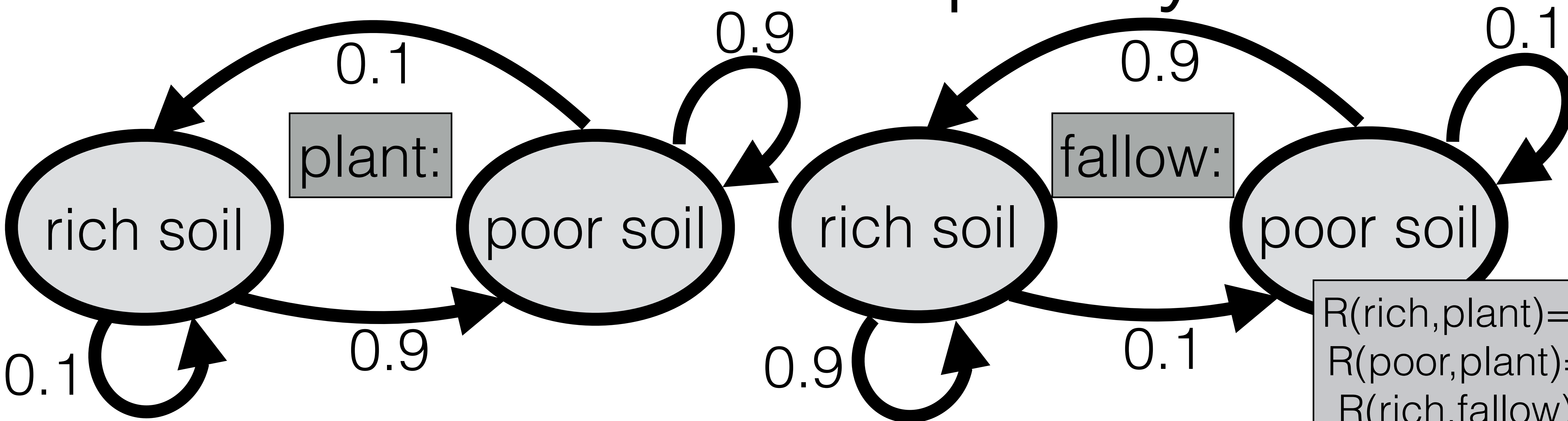
$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_\pi^{h-1}(s')$$

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

value of the policy with *h* steps left

value of the policy on this time step

What's the value of a policy?



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

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

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

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

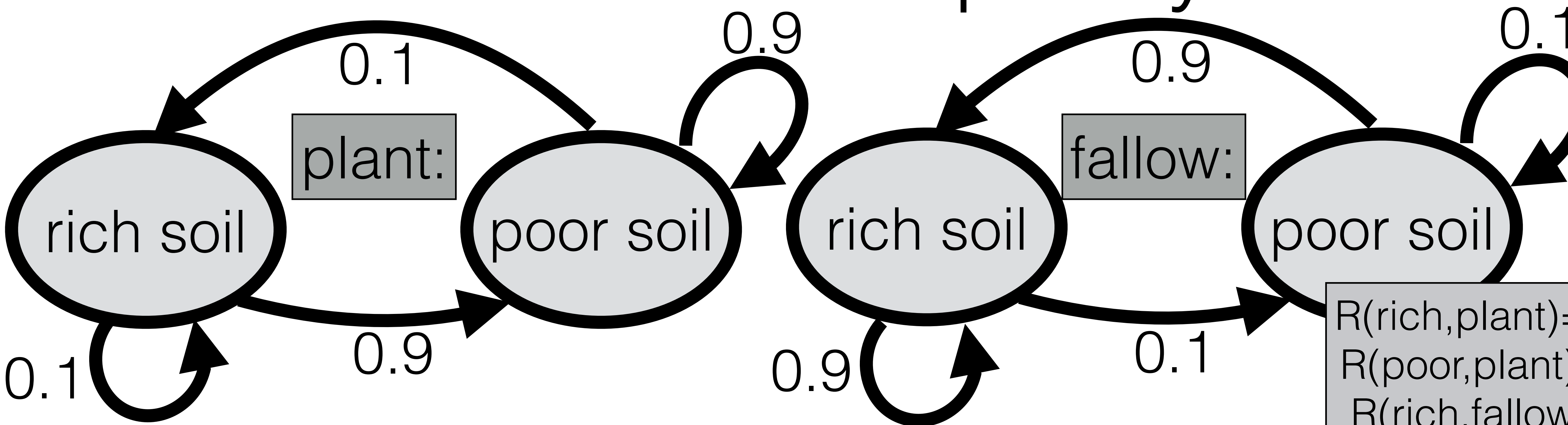
$$V_{\pi_A}^1(\text{rich}) = 100; V_{\pi_A}^1(\text{poor}) = 10; V_{\pi_B}^1(\text{rich}) = 100; V_{\pi_B}^1(\text{poor}) = 0$$

value of the policy with *h* steps left

value of the policy on this time step

(expected) value of the policy across all future time steps

What's the value of a policy?



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

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

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

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

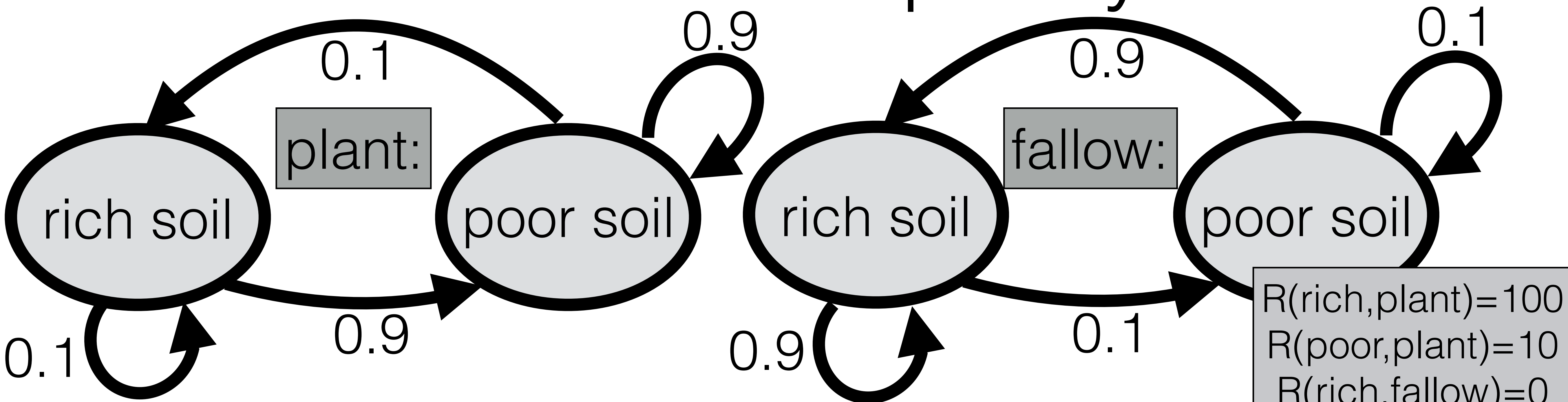
$$V_{\pi_A}^1(\text{rich}) = 100; V_{\pi_A}^1(\text{poor}) = 10; V_{\pi_B}^1(\text{rich}) = 100; V_{\pi_B}^1(\text{poor}) = 0$$

value of the policy with *h* steps left

value of the policy on this time step

(expected) value of the policy across all future time steps

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 π starting at *s*

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

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

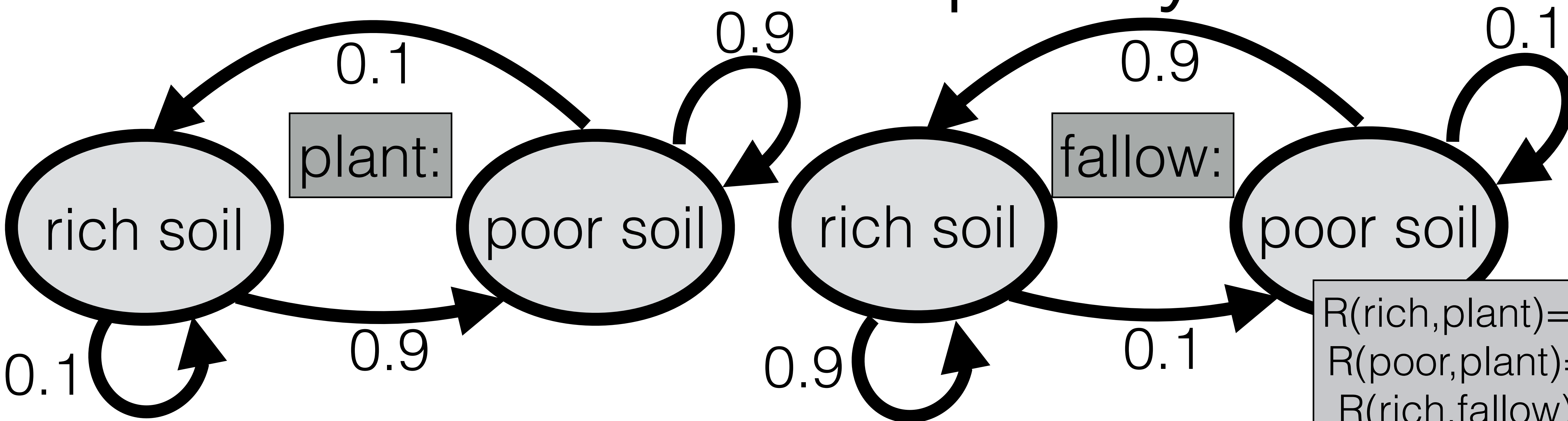
$$V_{\pi_A}^1(\text{rich}) = 100; V_{\pi_A}^1(\text{poor}) = 10; V_{\pi_B}^1(\text{rich}) = 100; V_{\pi_B}^1(\text{poor}) = 0$$

value of the policy with *h* steps left

value of the policy on this time step

(expected) value of the policy across all future time steps

What's the value of a policy?



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

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

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

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

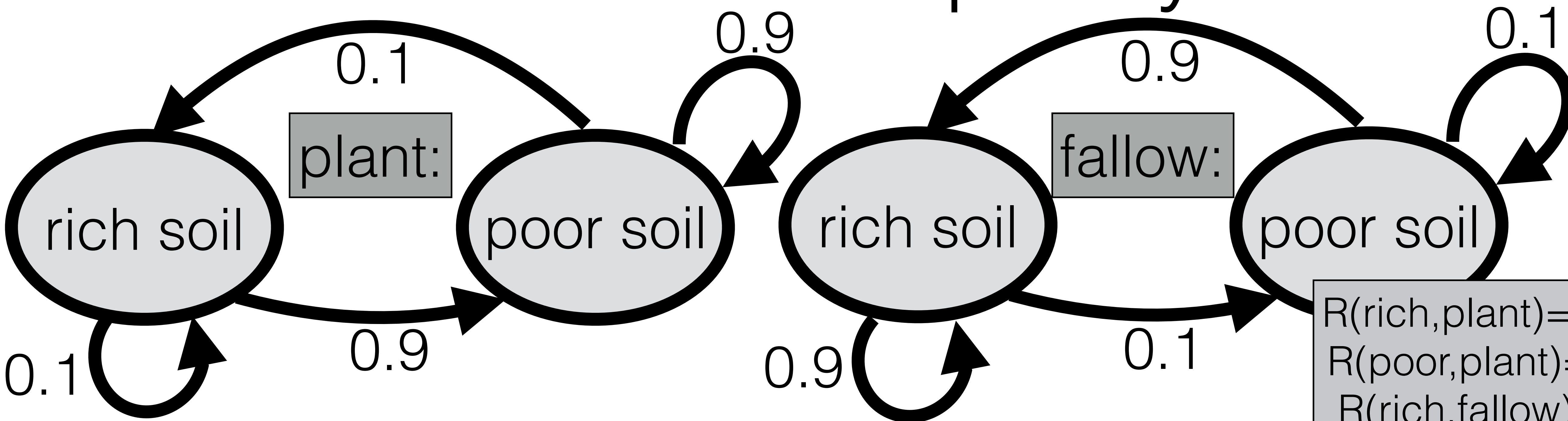
$$V_{\pi_A}^1(\text{rich}) = 100; V_{\pi_A}^1(\text{poor}) = 10; V_{\pi_B}^1(\text{rich}) = 100; V_{\pi_B}^1(\text{poor}) = 0$$

value of the policy with h steps left

value of the policy on this time step

(expected) value of the policy across all future time steps

What's the value of a policy?



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

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

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

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

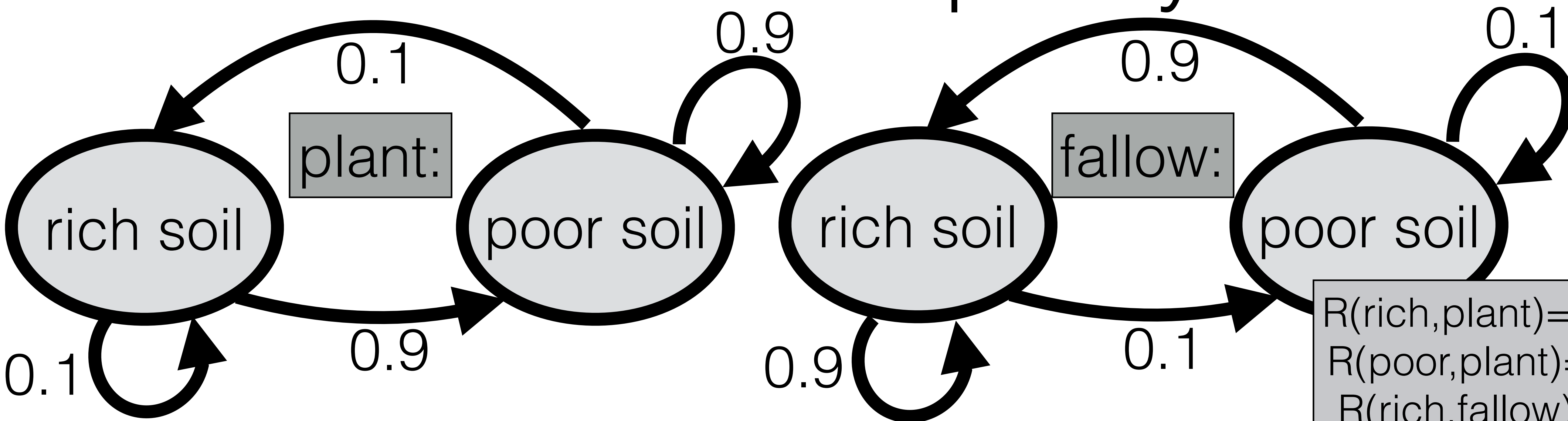
$$V_{\pi_A}^1(\text{rich}) = 100; V_{\pi_A}^1(\text{poor}) = 10; V_{\pi_B}^1(\text{rich}) = 100; V_{\pi_B}^1(\text{poor}) = 0$$

value of the policy with *h* steps left

value of the policy on this time step

(expected) value of the policy across all future time steps

What's the value of a policy?



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

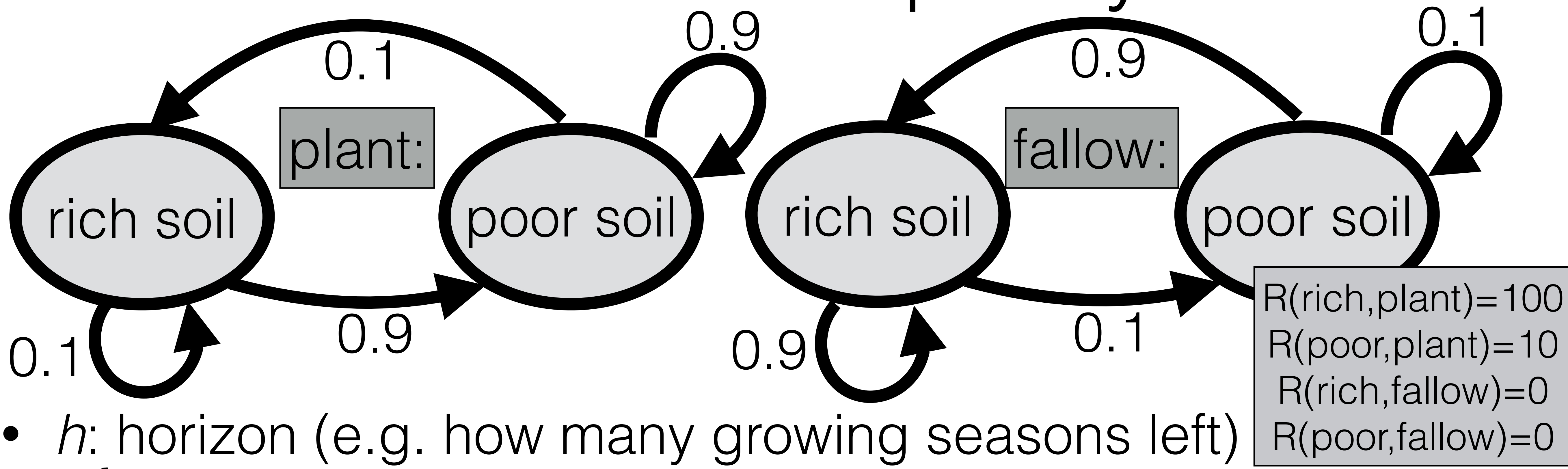
- *h*: horizon (e.g. how many growing seasons left)
- $V_\pi^h(s)$: value (expected reward) with policy π starting at *s*

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

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

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

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 π starting at *s*

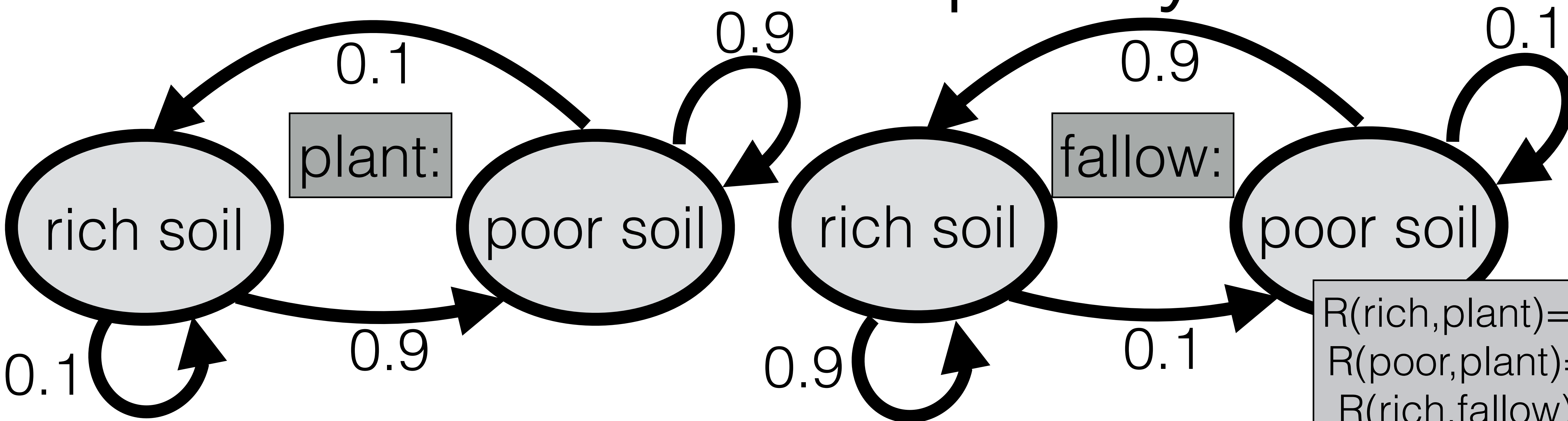
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

What's the value of a policy?



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

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

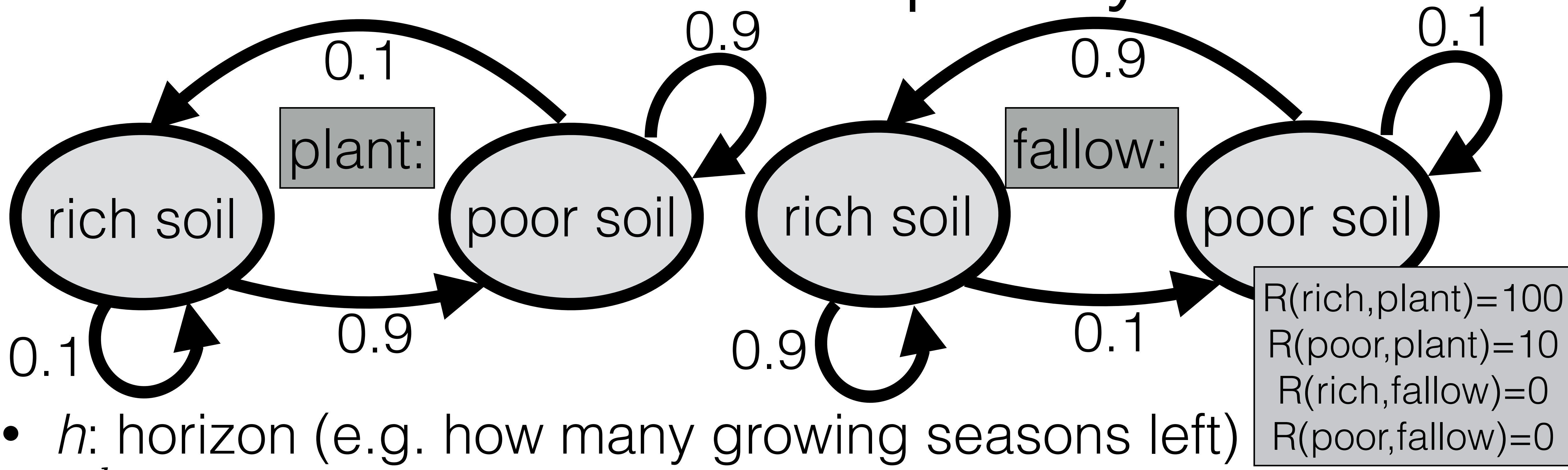
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

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 π starting at *s*

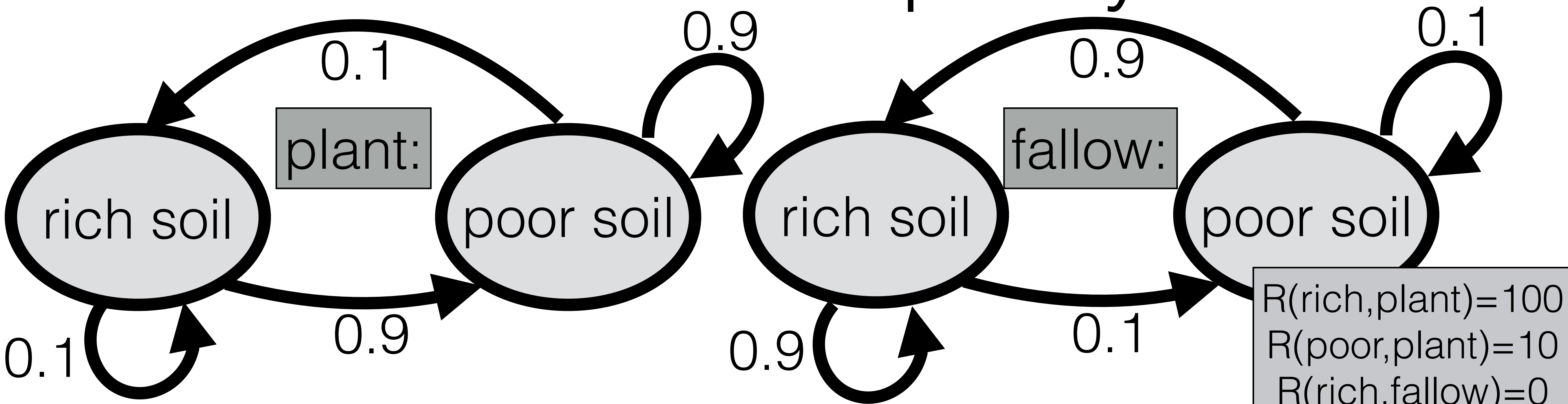
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

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 π starting at *s*

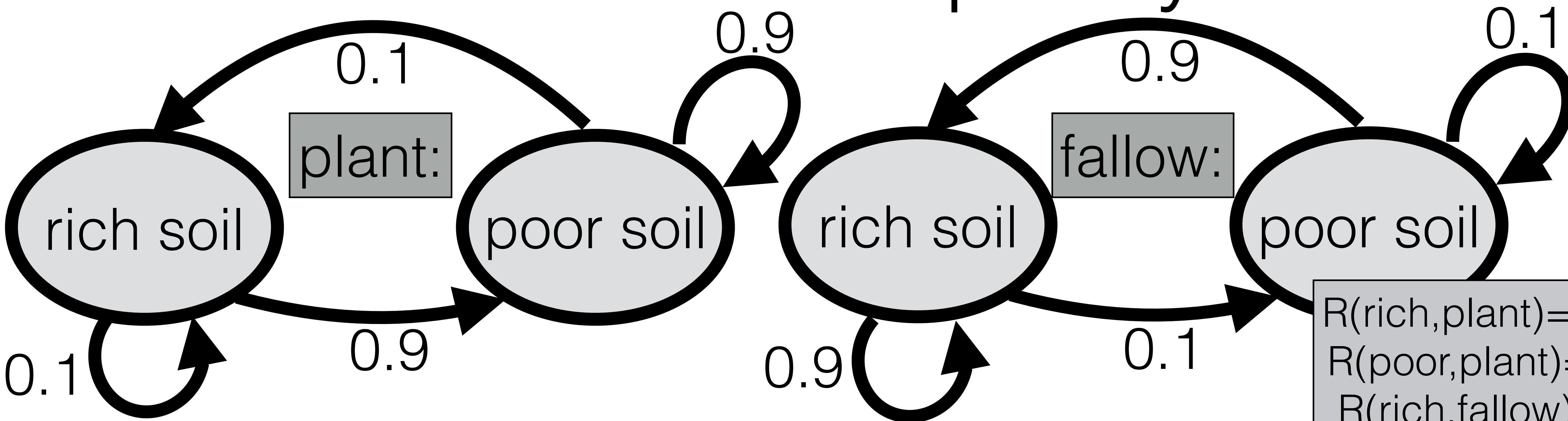
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

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 π starting at *s*

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

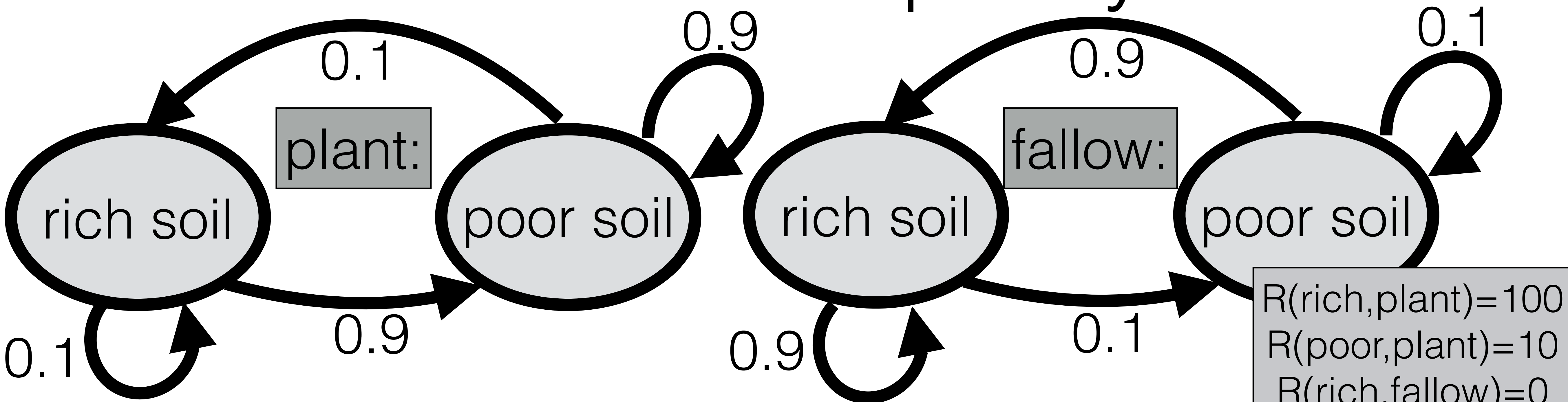
$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_\pi^{h-1}(s')$$

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

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

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

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 π starting at *s*

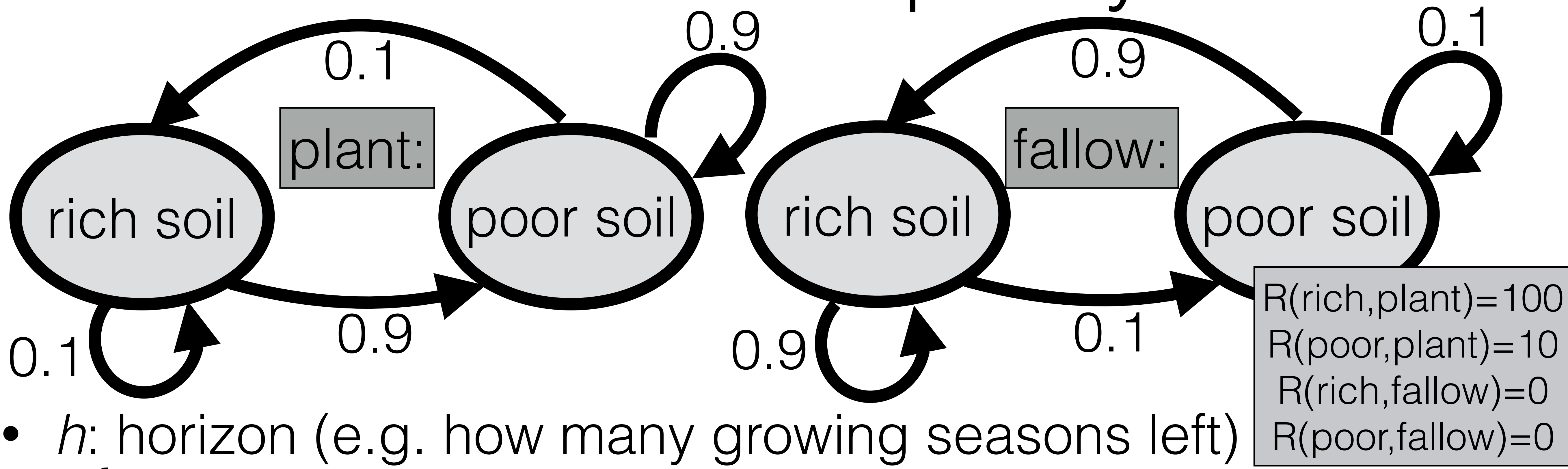
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

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 π starting at *s*

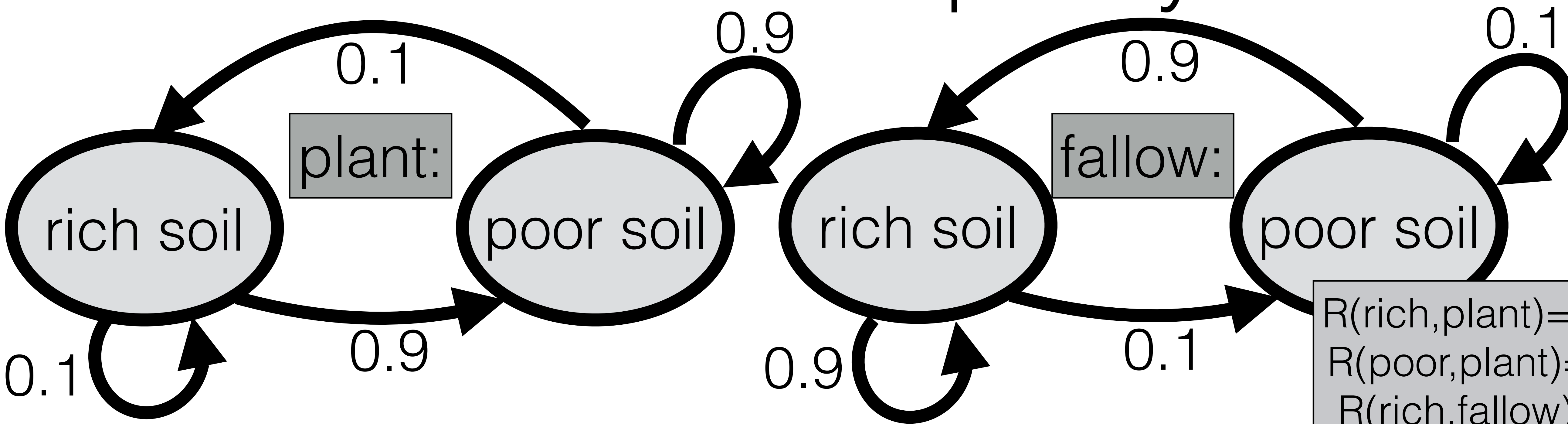
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

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 π starting at *s*

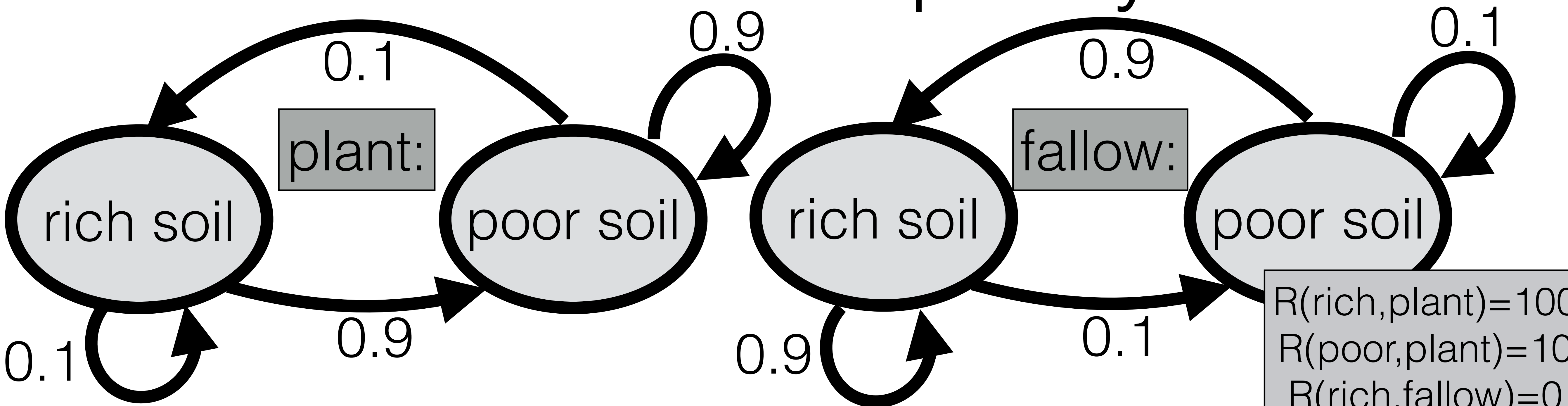
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

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 π starting at *s*

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

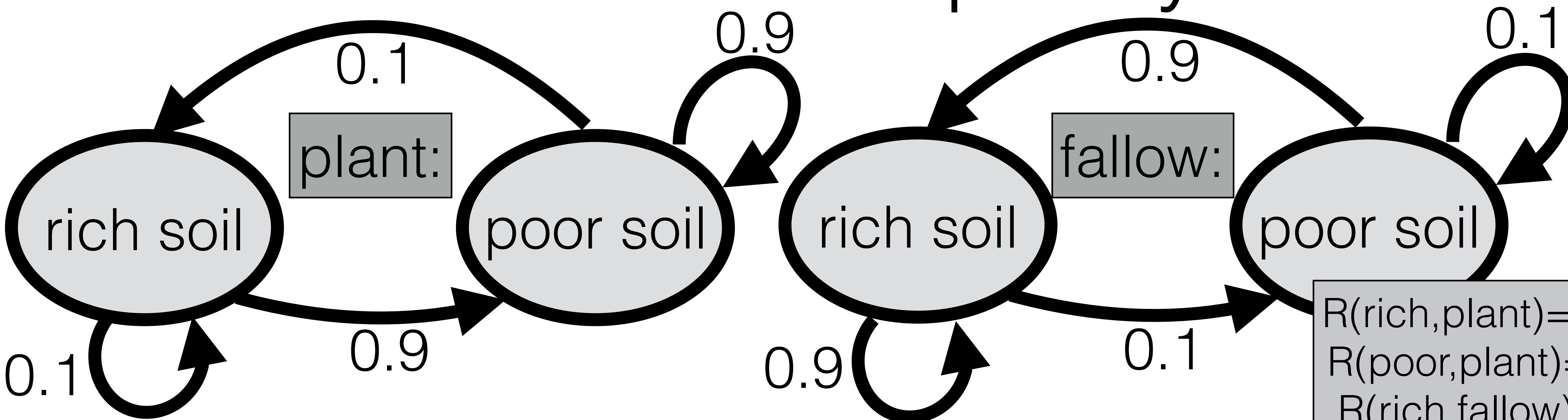
$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_\pi^{h-1}(s')$$

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

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

= 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 π starting at *s*

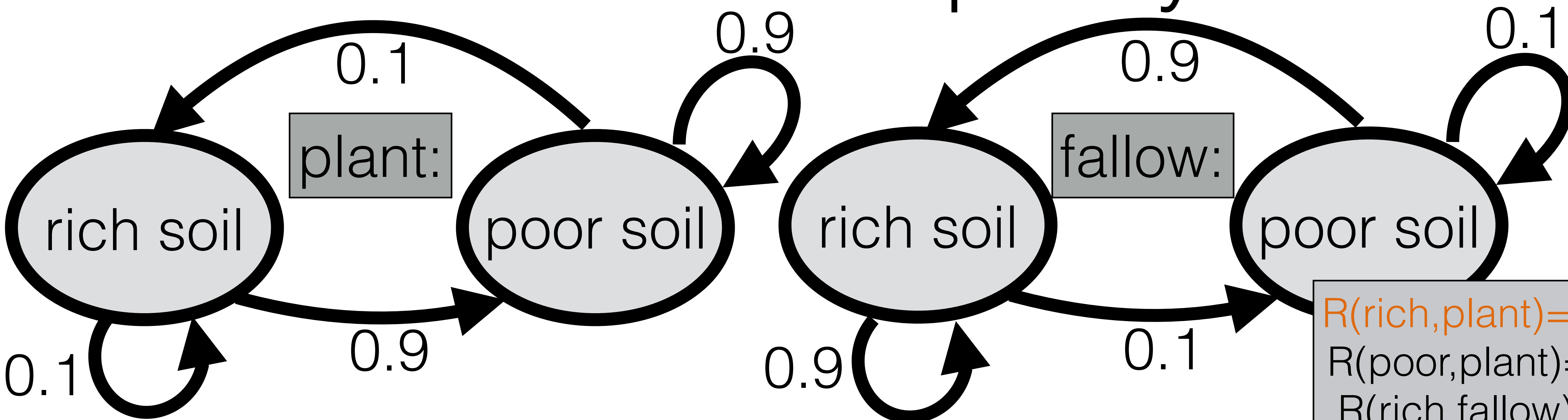
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

What's the value of a policy?



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

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

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

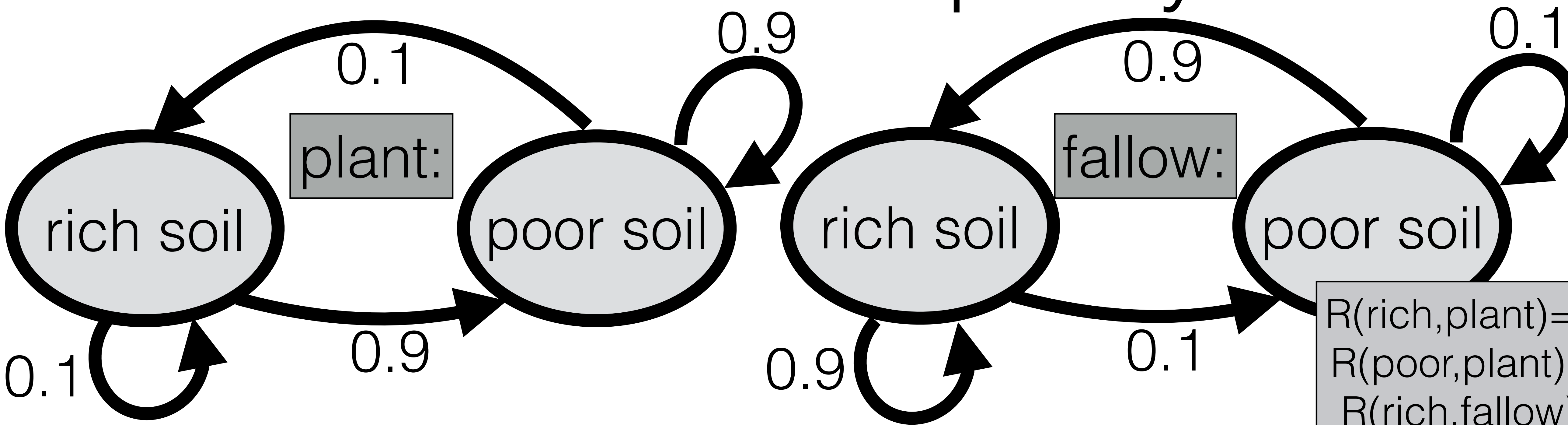
$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_\pi^{h-1}(s')$$

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

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

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

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 π starting at *s*

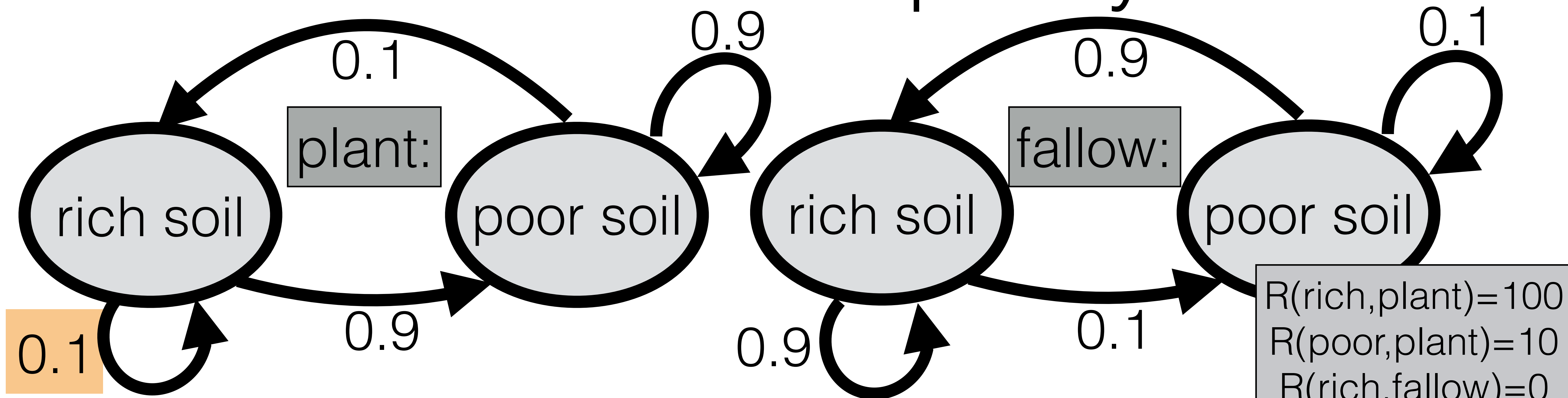
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

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 π starting at *s*

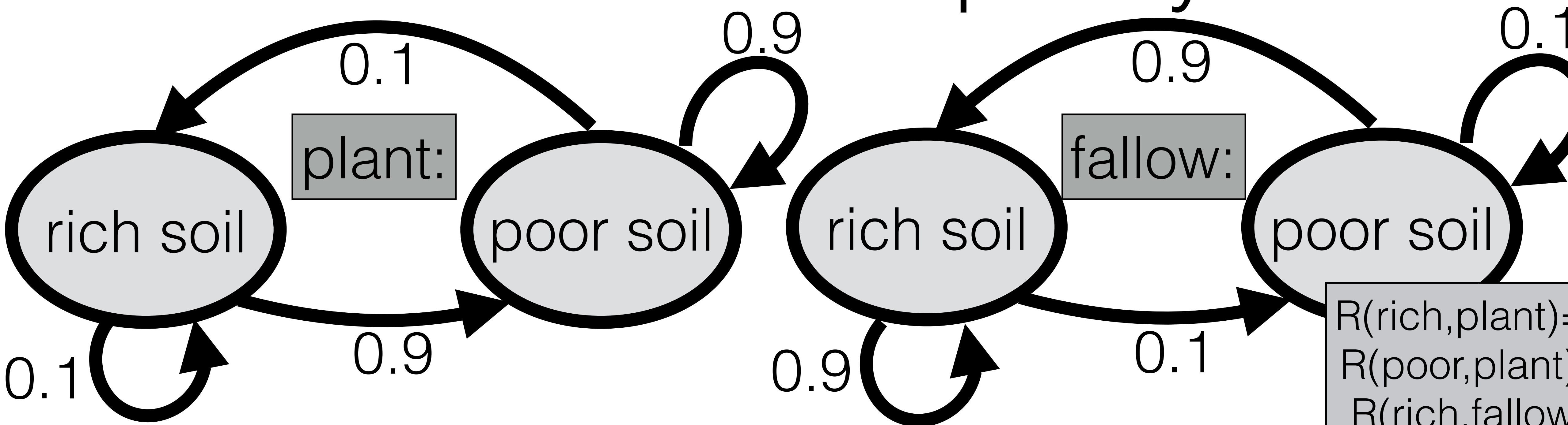
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

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 π starting at *s*

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

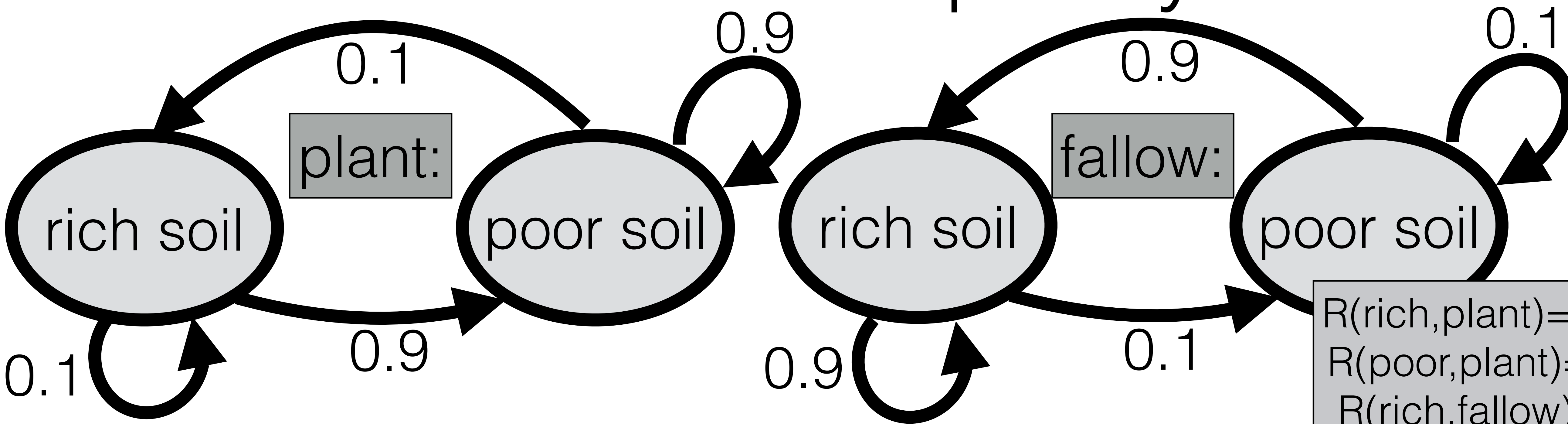
$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \textstyle\sum_{s'} T(s, \pi(s), s') \cdot V_\pi^{h-1}(s')$$

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

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

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

What's the value of a policy?



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

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

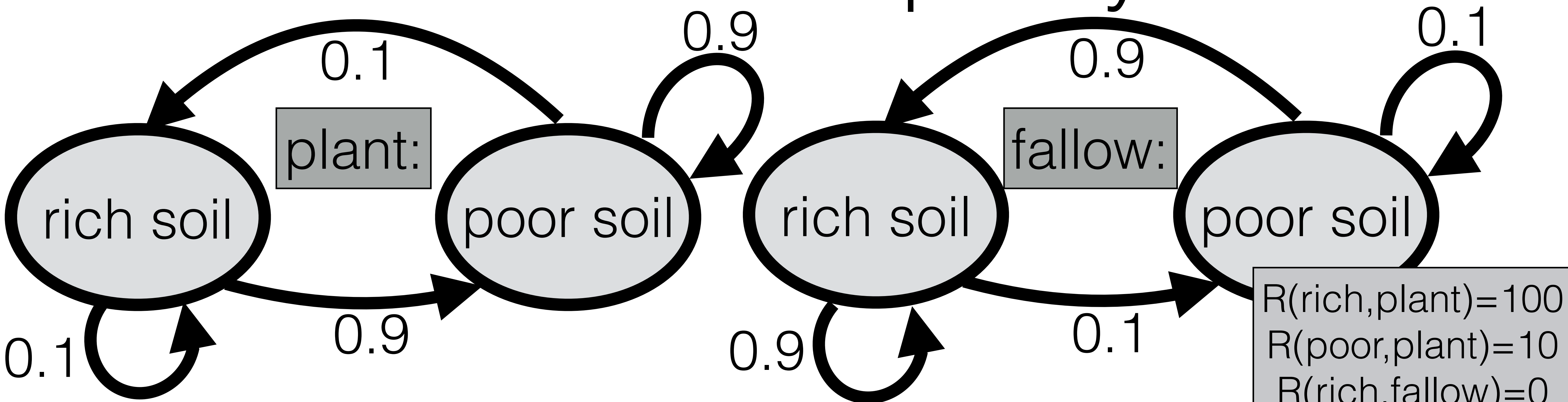
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

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 π starting at *s*

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

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

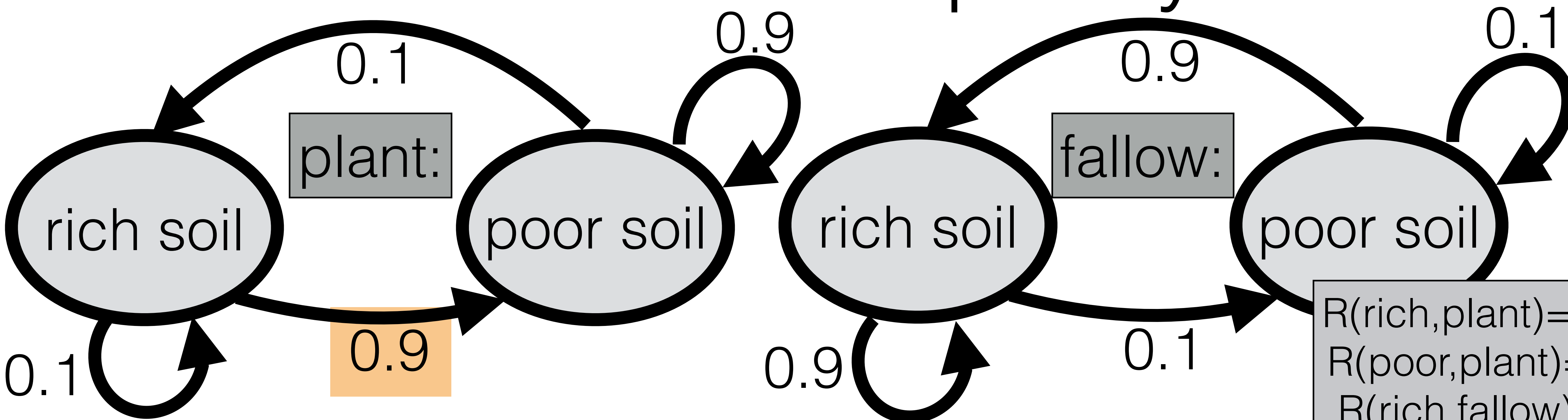
$$V_{\pi_A}^1(\text{rich}) = 100; V_{\pi_A}^1(\text{poor}) = 10; V_{\pi_B}^1(\text{rich}) = 100; V_{\pi_B}^1(\text{poor}) = 0$$

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

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

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

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 π starting at *s*

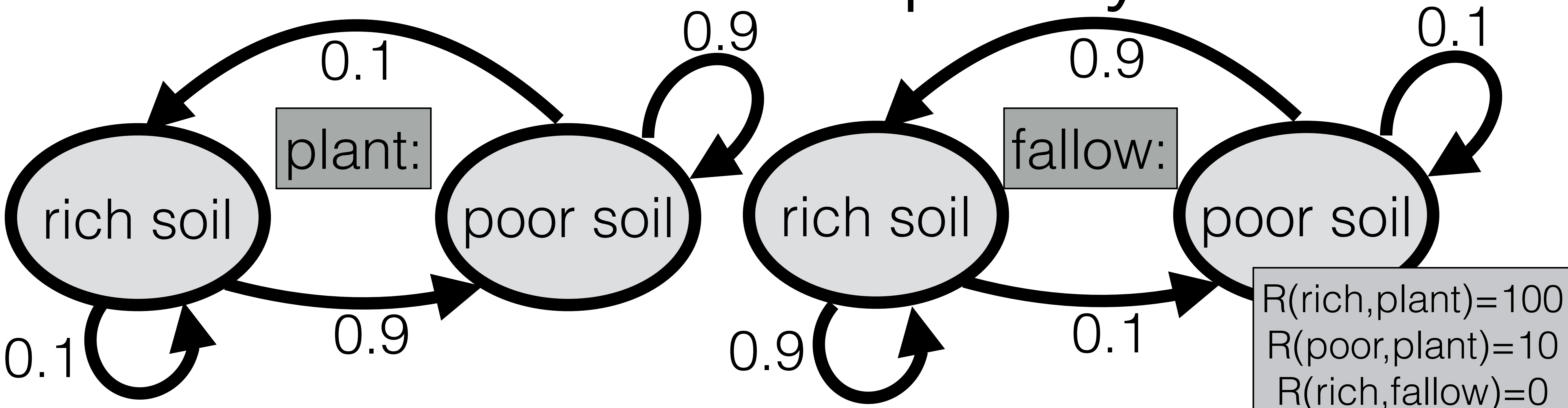
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

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 π starting at *s*

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

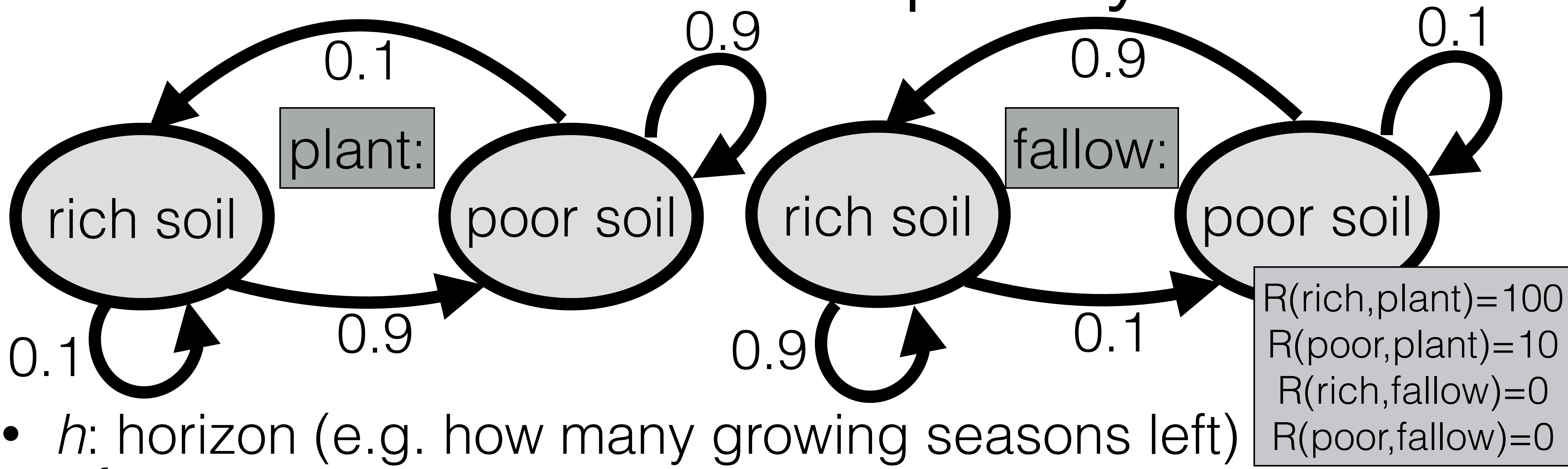
$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \textstyle\sum_{s'} T(s, \pi(s), s') \cdot V_\pi^{h-1}(s')$$

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

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

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

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 π starting at *s*

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

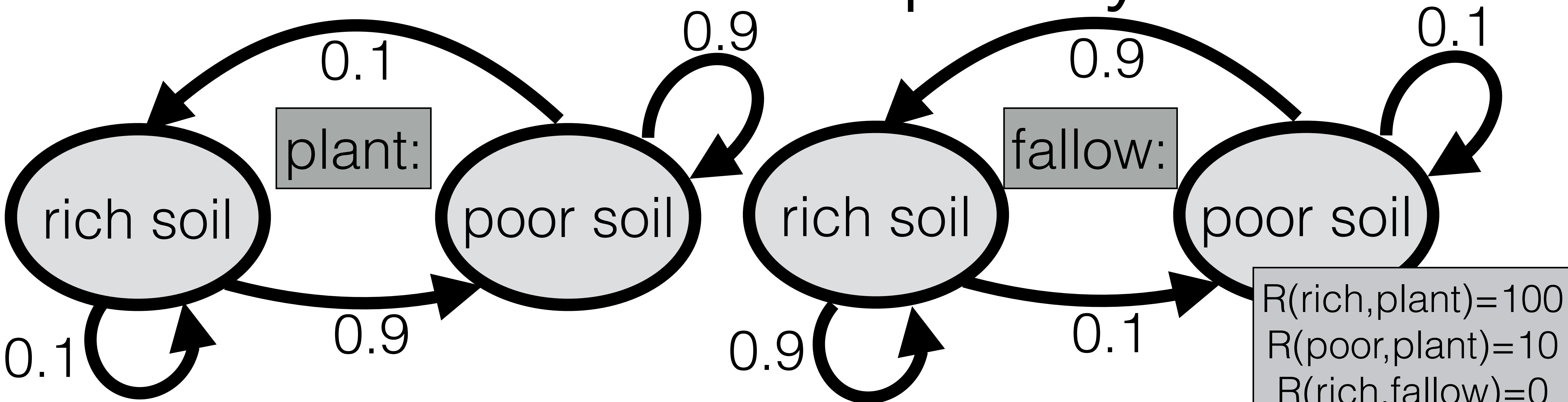
$$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_\pi^{h-1}(s')$$

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

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

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

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 π starting at *s*

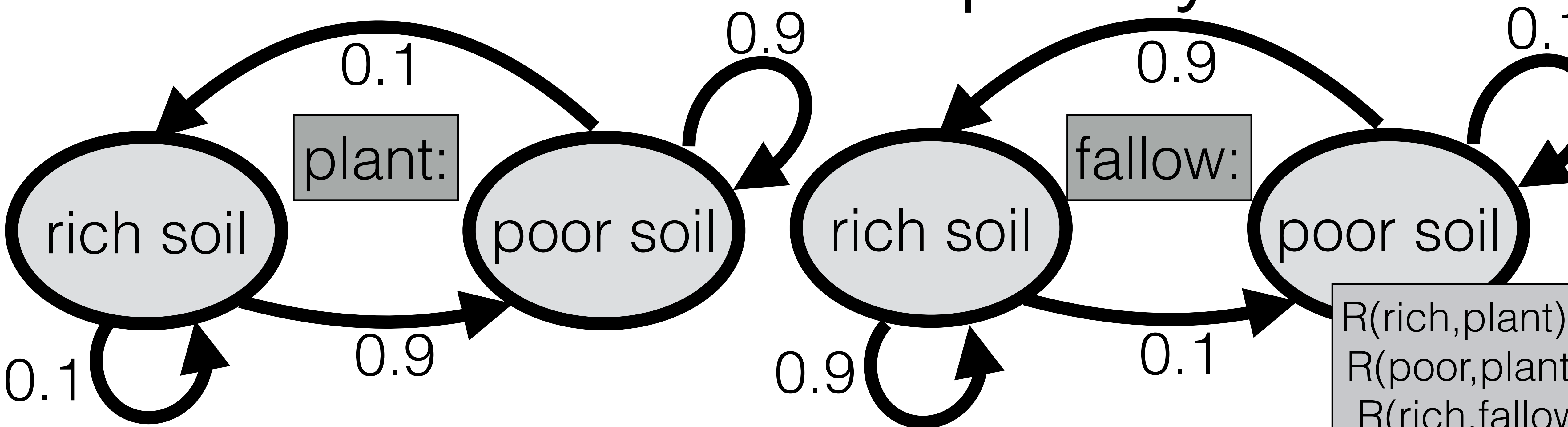
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

What's the value of a policy?



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

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

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

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

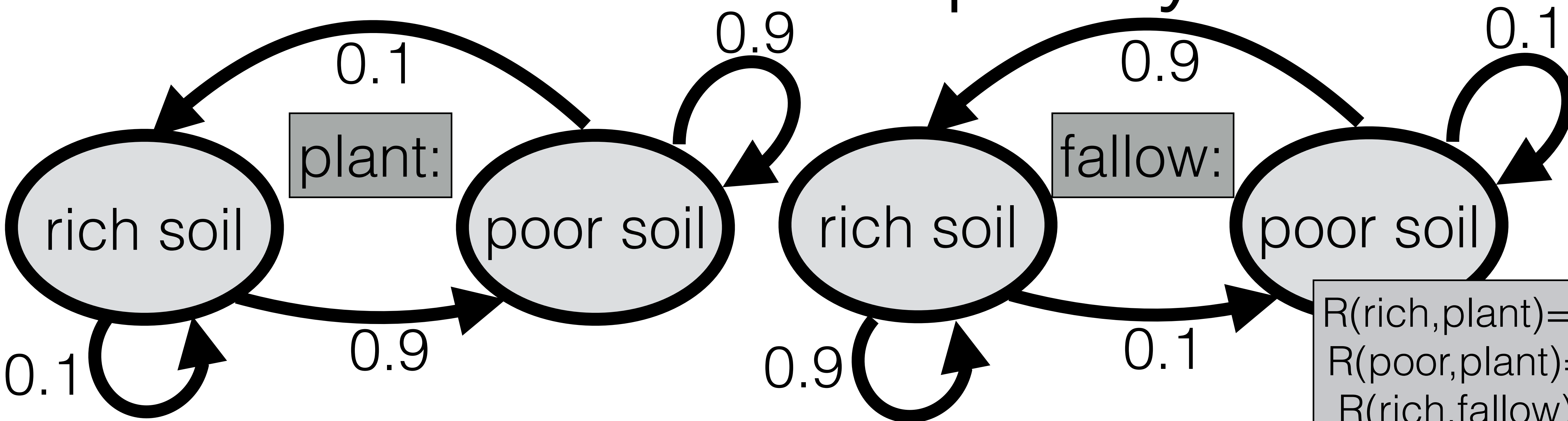
$$V_{\pi_A}^1(\text{rich}) = 100; V_{\pi_A}^1(\text{poor}) = 10; V_{\pi_B}^1(\text{rich}) = 100; V_{\pi_B}^1(\text{poor}) = 0$$

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

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

= 119

What's the value of a policy?



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

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

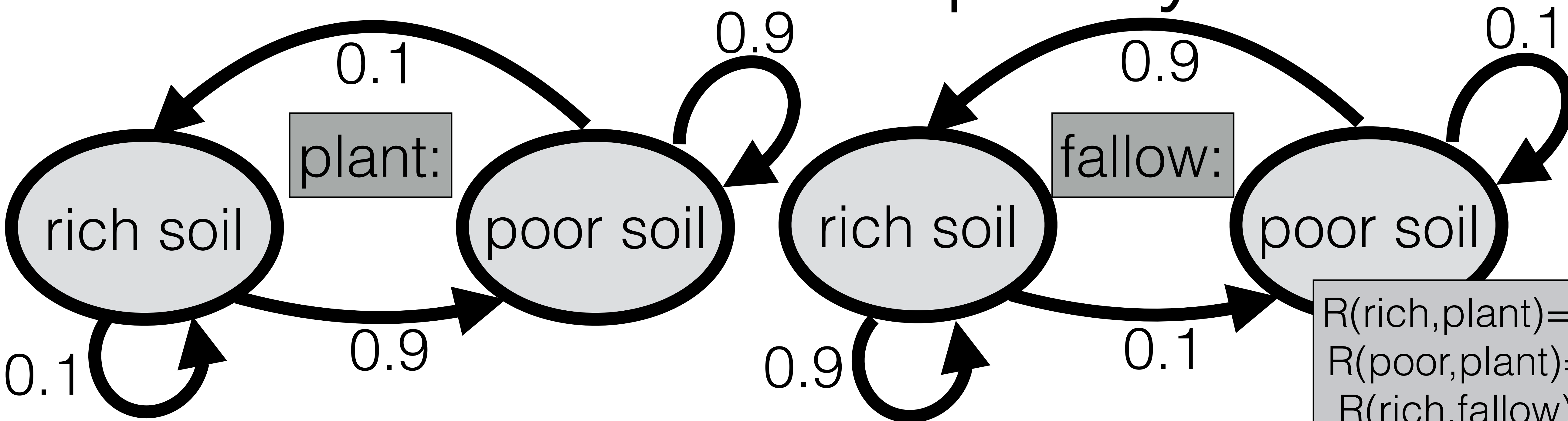
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

What's the value of a policy?



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

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

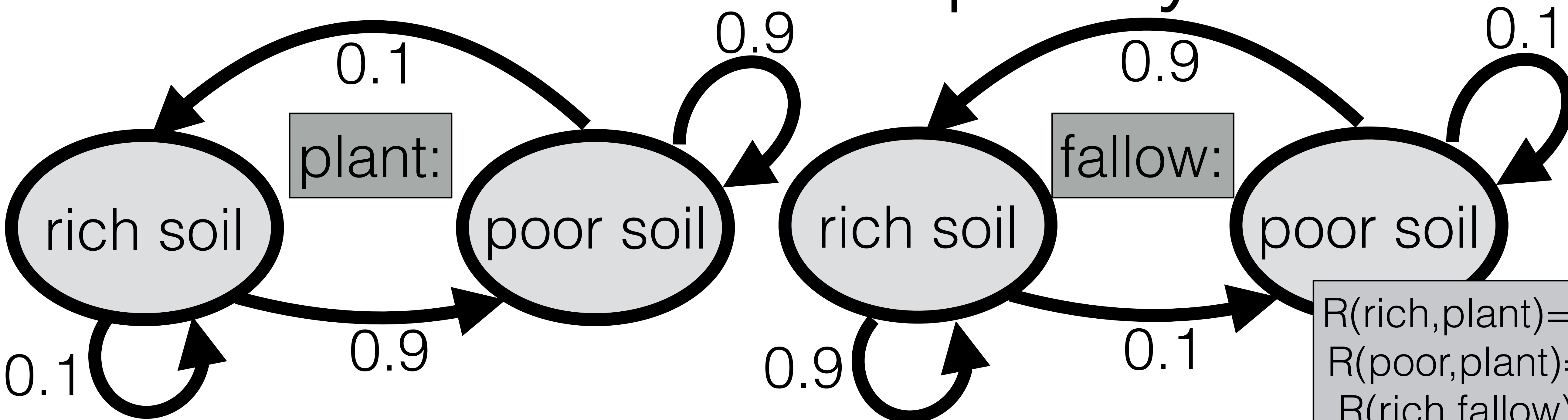
Dueling farmers! π_A : always plant; π_B: plant if rich, else fallow

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

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

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

What's the value of a policy?



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

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

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

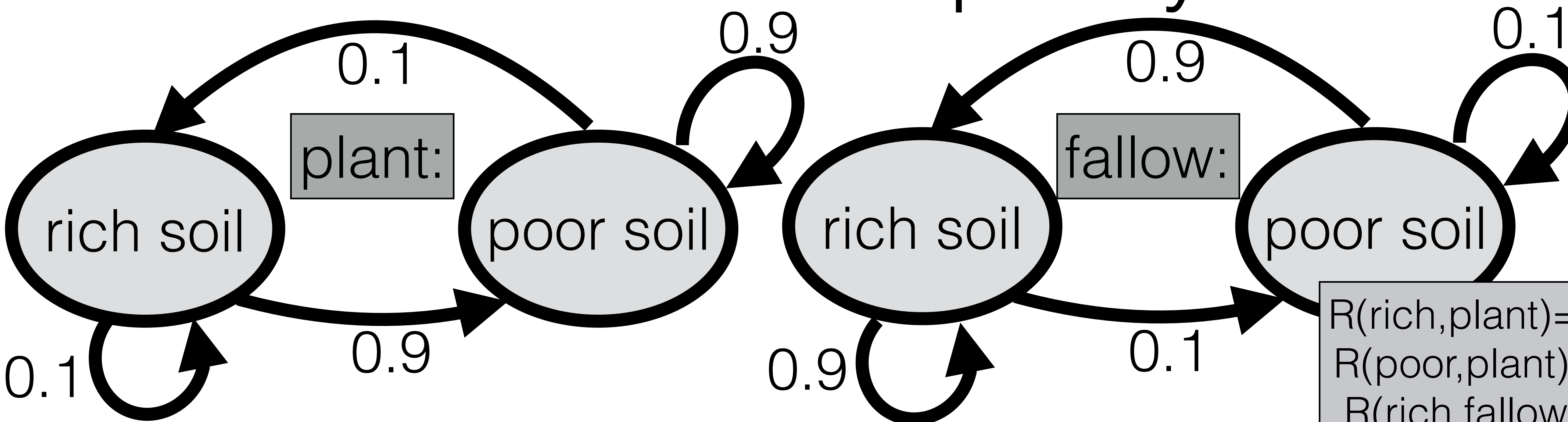
$V_\pi^0(s) = 0; V_\pi^h(s) = R(s, \pi(s)) + \sum_{s'} T(s, \pi(s), s') \cdot V_\pi^{h-1}(s')$

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

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

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

What's the value of a policy?



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

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

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

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

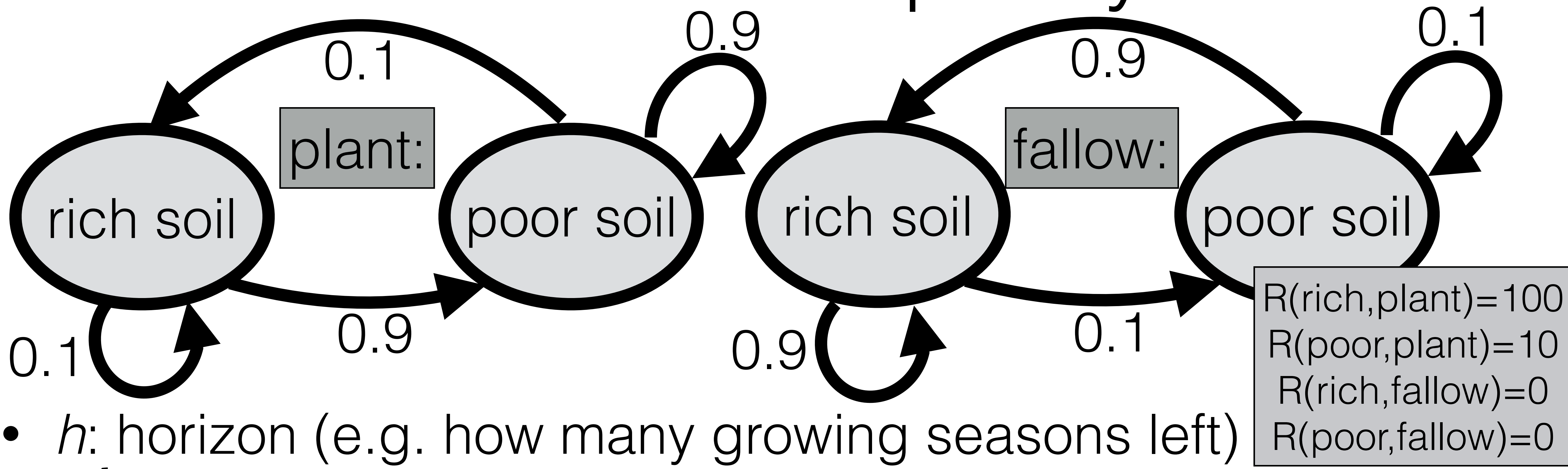
$$V_{\pi_A}^1(\text{rich}) = 100; V_{\pi_A}^1(\text{poor}) = 10; V_{\pi_B}^1(\text{rich}) = 100; V_{\pi_B}^1(\text{poor}) = 0$$

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

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

Who wins?

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 π starting at *s*

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

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

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

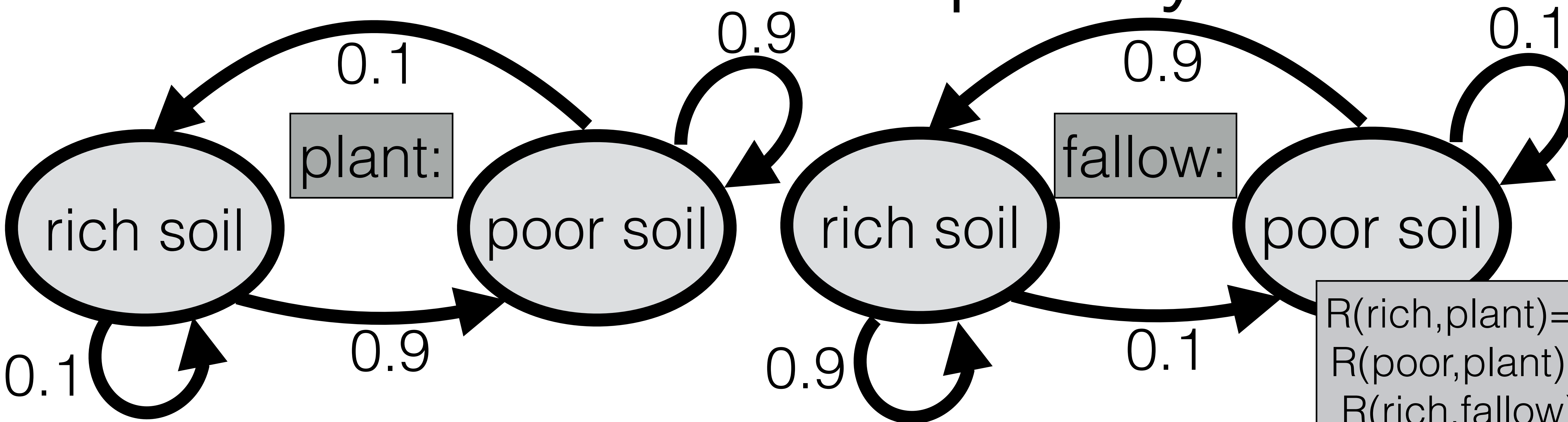
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

Who wins?

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

What's the value of a policy?



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

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

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

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

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

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

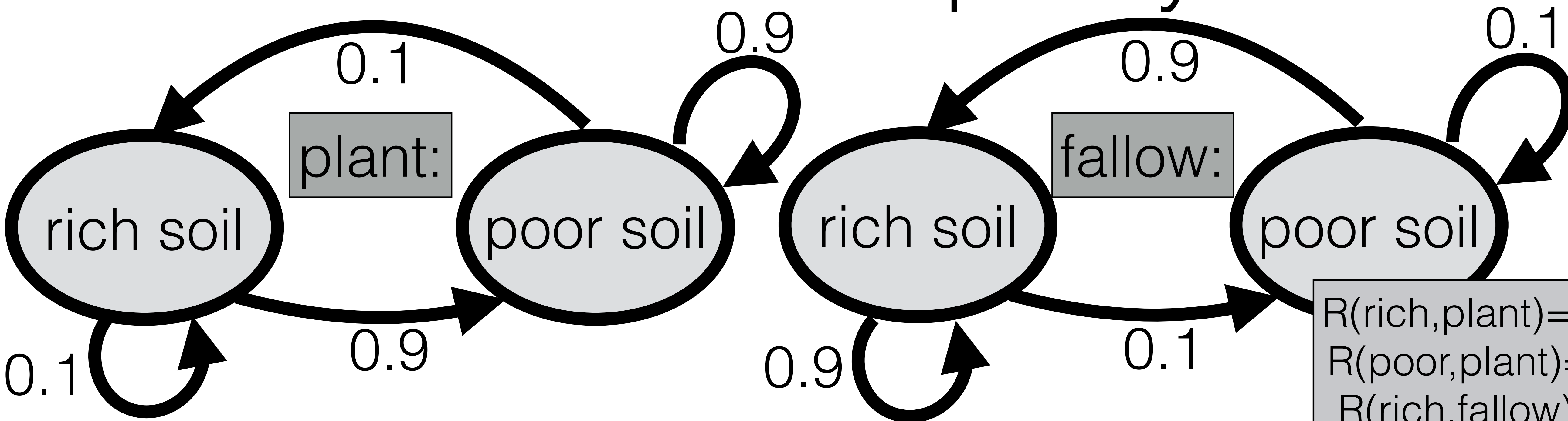
$$V_{\pi_A}^3(\text{rich}) = 138; V_{\pi_A}^3(\text{poor}) = 48; V_{\pi_B}^3(\text{rich}) = 192; V_{\pi_B}^3(\text{poor}) = 108$$

Who wins?

$h=1$

- 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 π starting at *s*

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

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

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

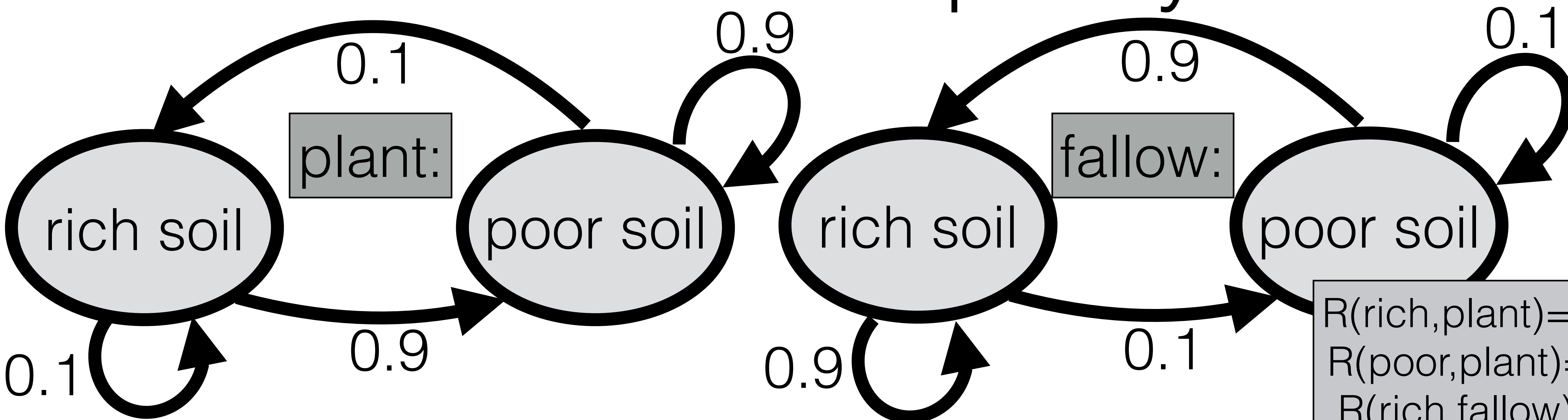
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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

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

What's the value of a policy?



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

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

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

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

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

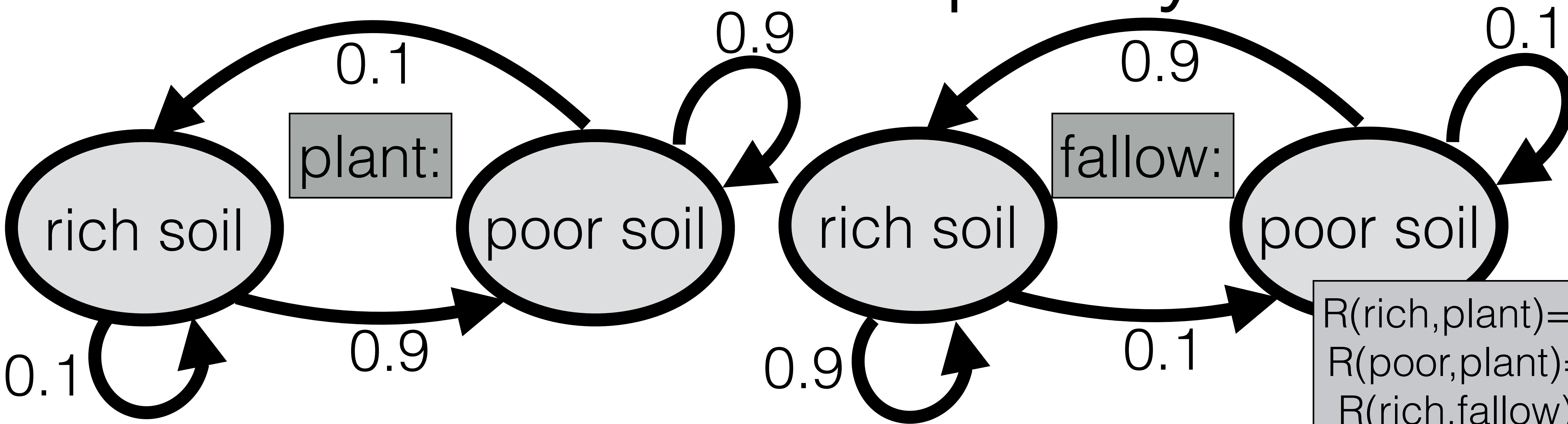
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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



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

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

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

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

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

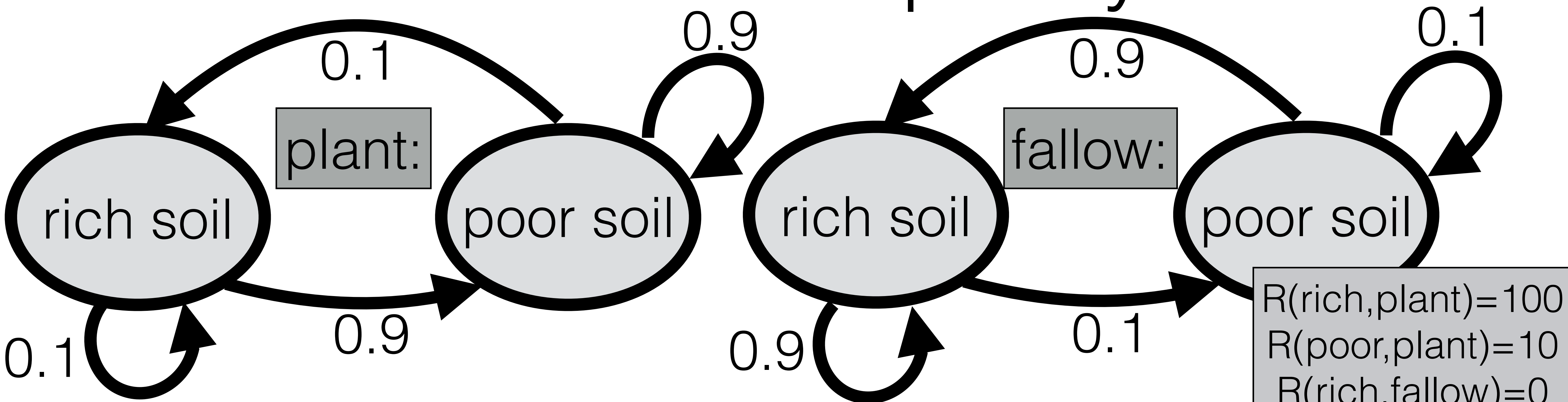
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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

- 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 π starting at s

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

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

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

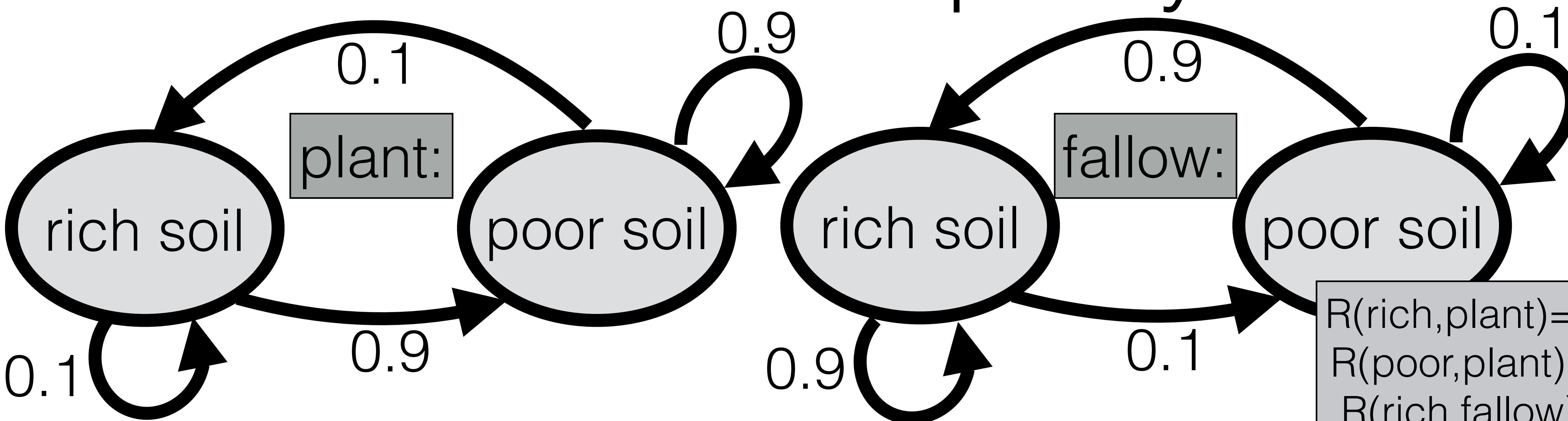
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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

- 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 π starting at *s*

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

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

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

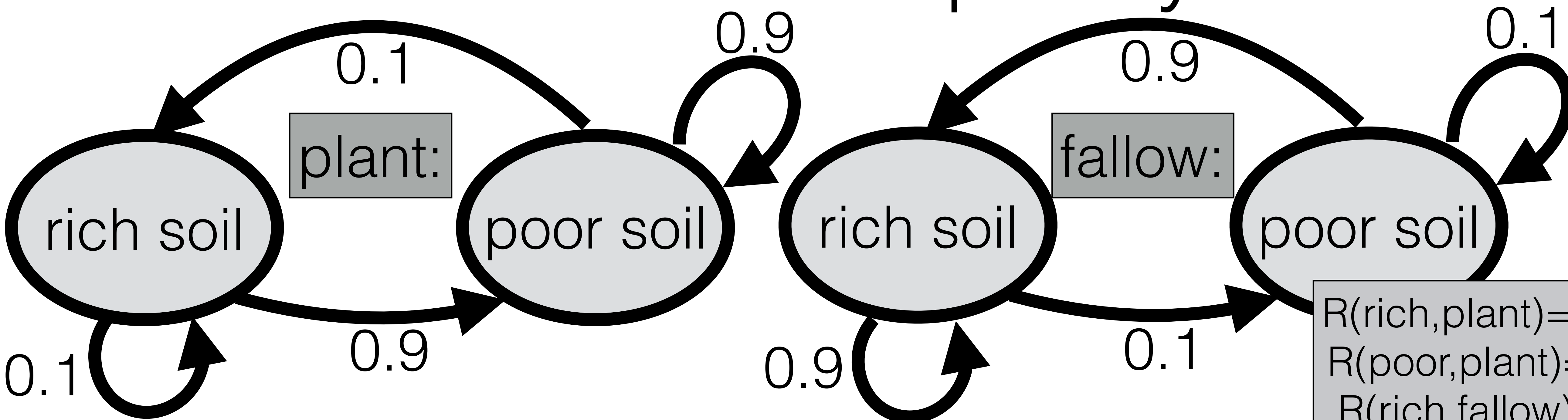
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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

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

What's the value of a policy?



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

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

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

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

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

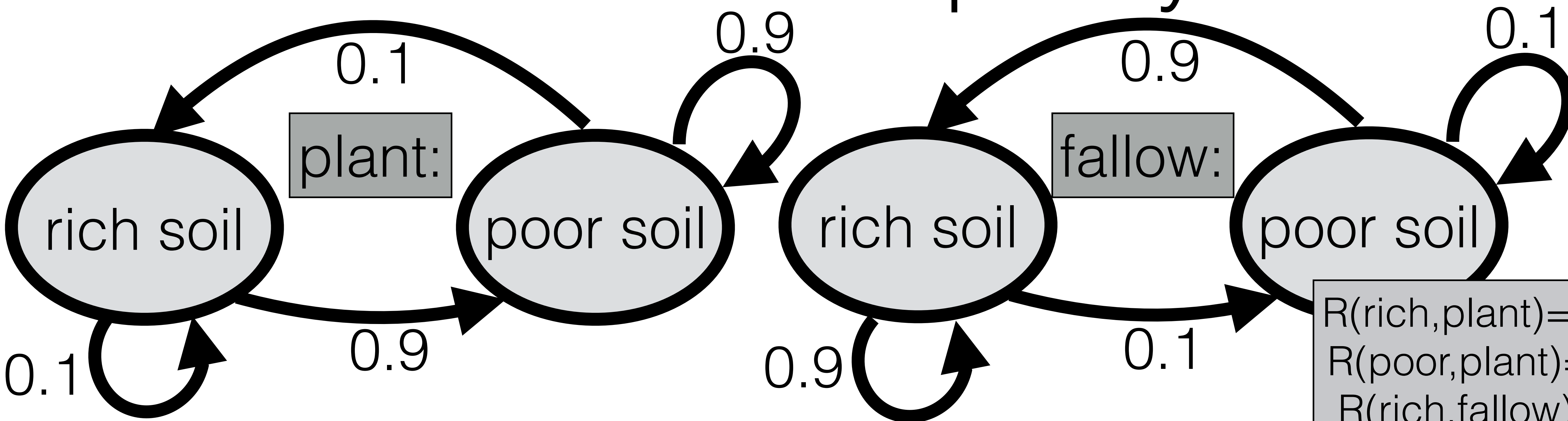
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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

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

What's the value of a policy?



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

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

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

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

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

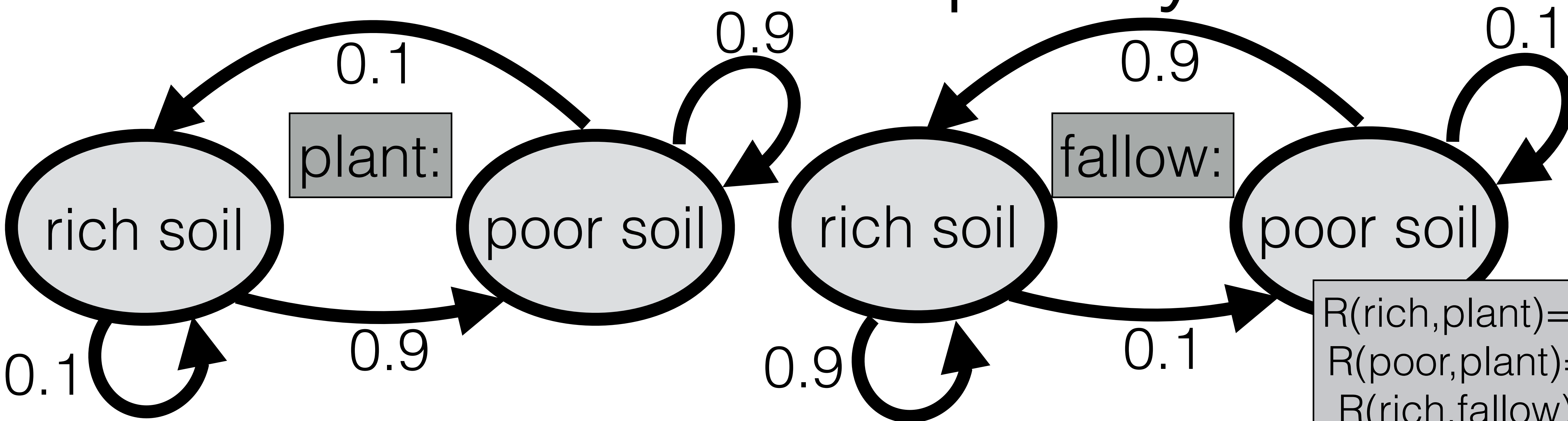
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

Who wins? $\pi_A >_{h=1} \pi_B ; \pi_A <_{h=3} \pi_B$ 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 π starting at *s*

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

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

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

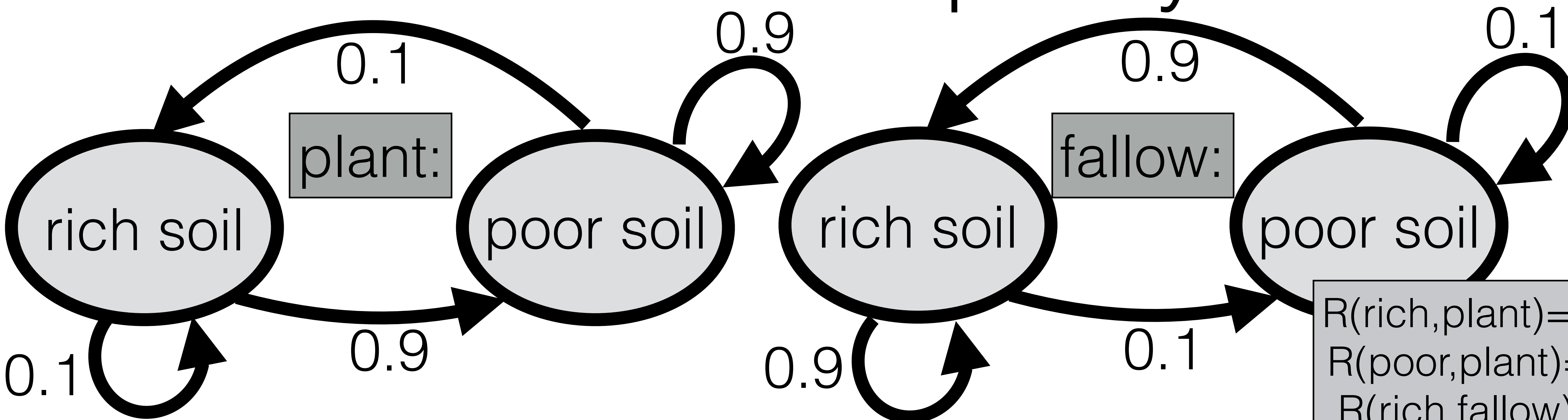
$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$

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

Who wins? $\pi_A >_{h=1} \pi_B; \pi_A <_{h=3} \pi_B$ 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 π starting at *s*

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

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

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

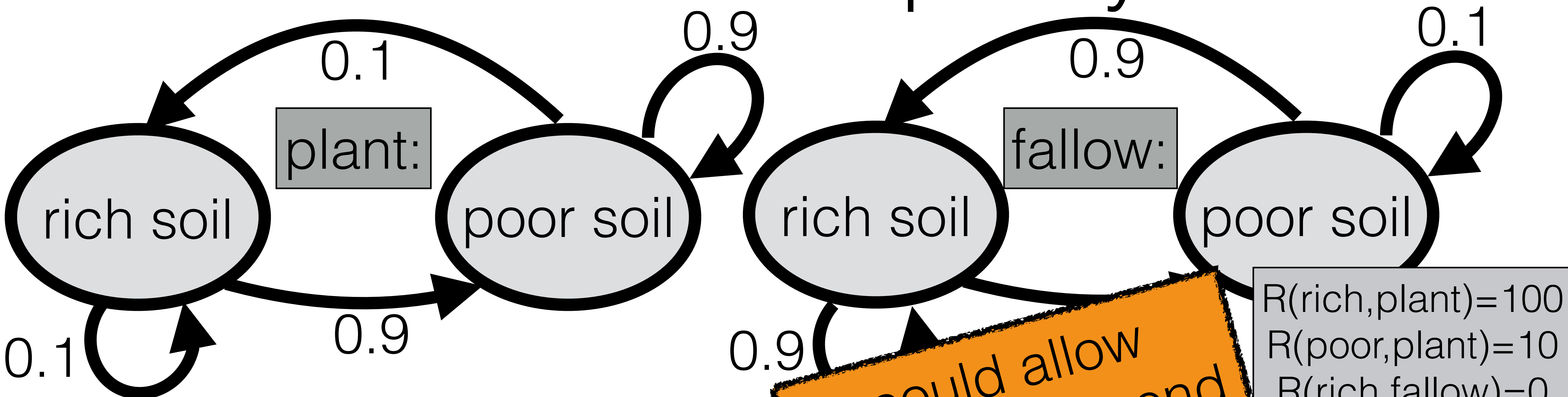
$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$

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

Who wins? $\pi_A >_{h=1} \pi_B; \pi_A <_{h=3} \pi_B$ 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! π_A: always plan nt if rich, else fallow

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

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

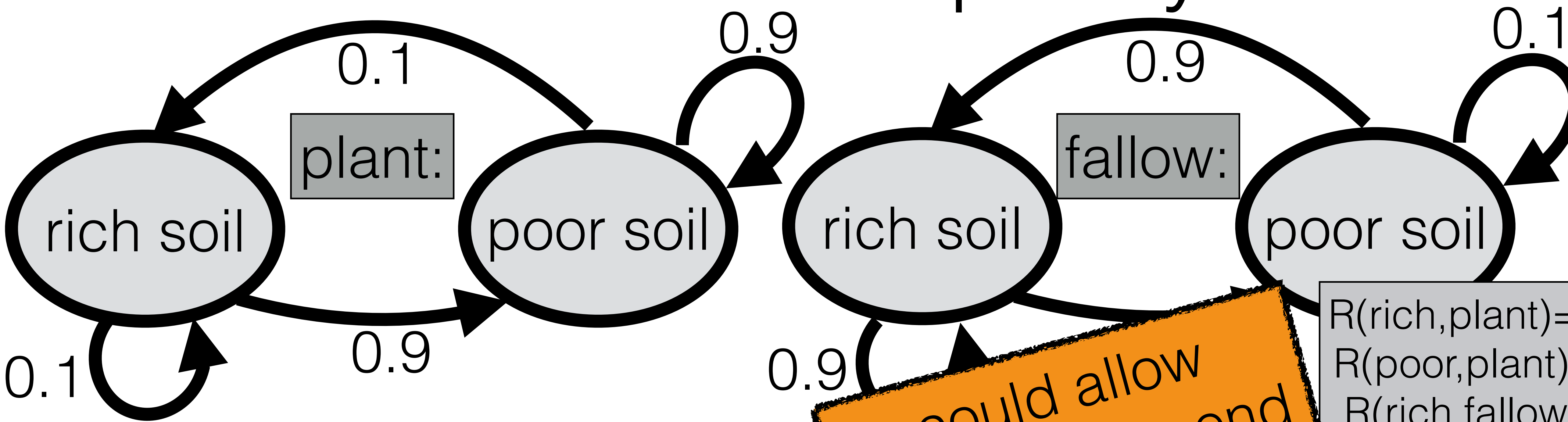
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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! π_A : always plan rich, else fallow

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

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

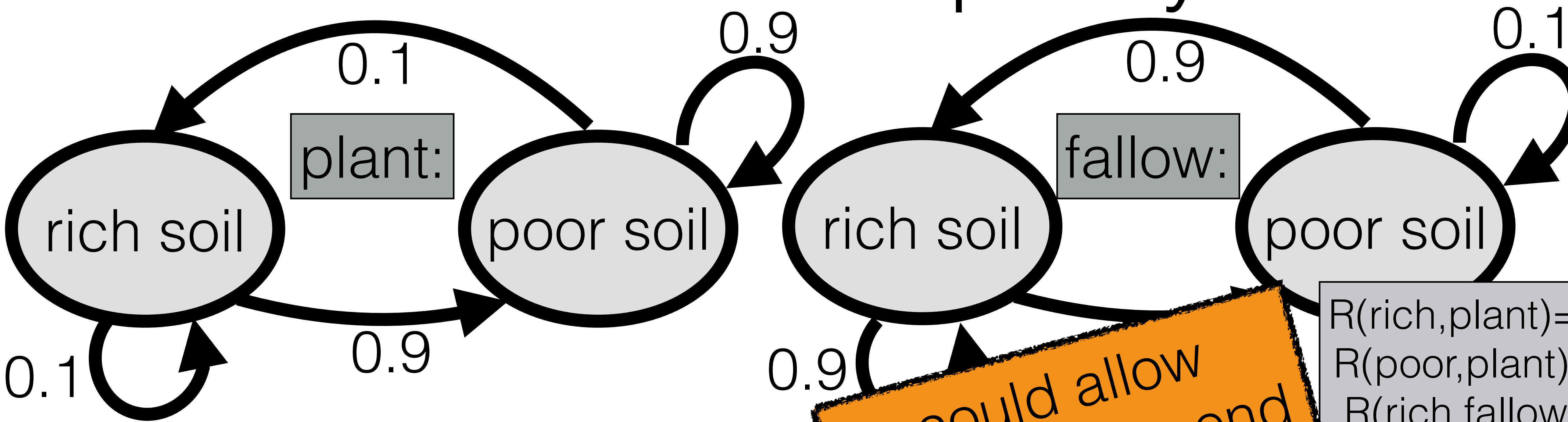
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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! π_A : always plan … nt if rich, else fallow

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

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

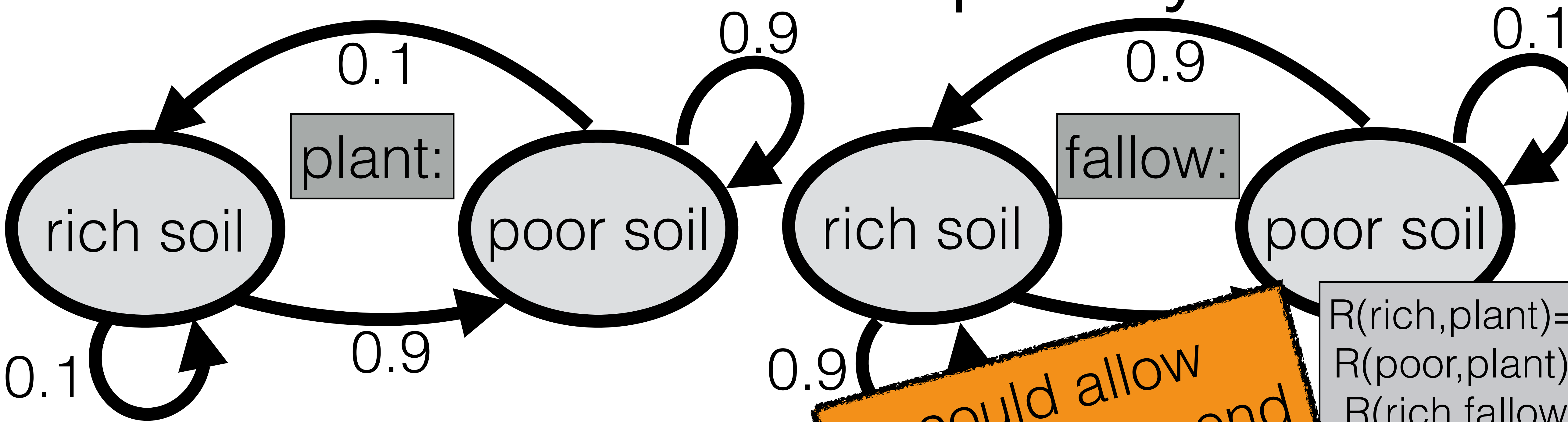
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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

 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! π_A : always plan rich, else fallow

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

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

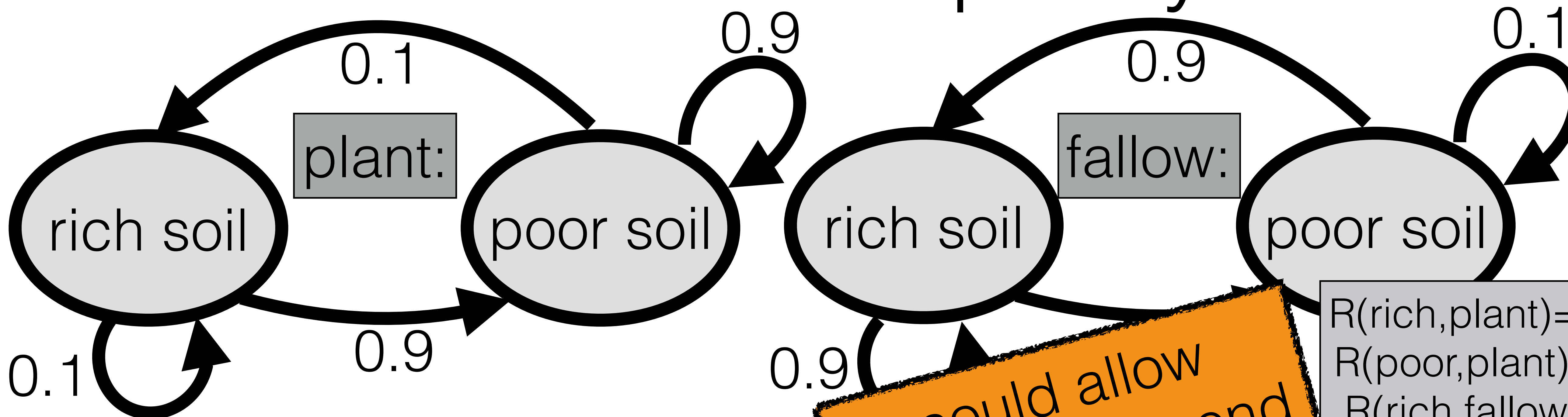
$$V_{\pi_A}^2(\text{rich}) = 119; V_{\pi_A}^2(\text{poor}) = 29; V_{\pi_B}^2(\text{rich}) = 110; V_{\pi_B}^2(\text{poor}) = 90$$

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

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! π_A: always plan… …nt if rich, else fallow

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

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

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

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

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

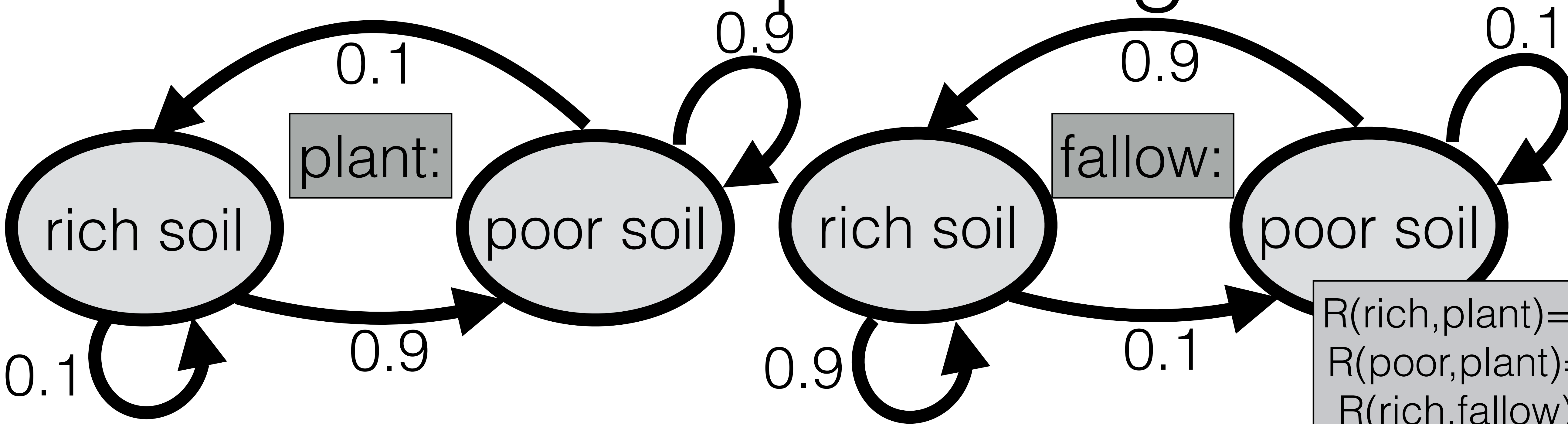
9 I.e. at least as good at all states and strictly better for at least one state

What if I don’t stop farming?

What if I don’t stop farming?

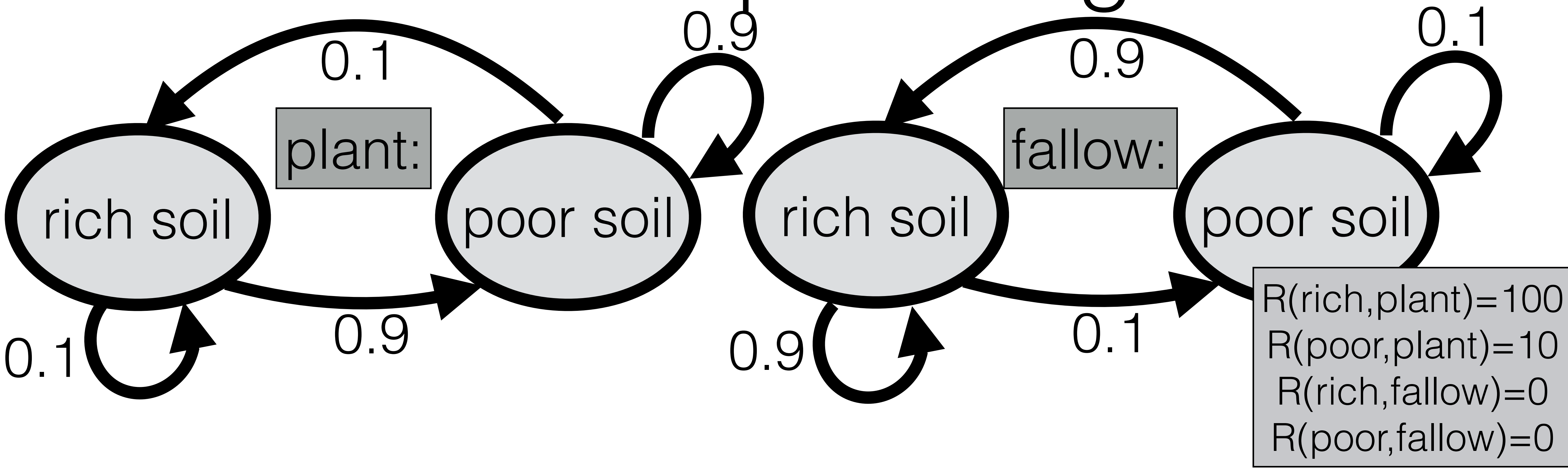
Good news! No strip mall, and I get to keep the farm forever

What if I don't stop farming?



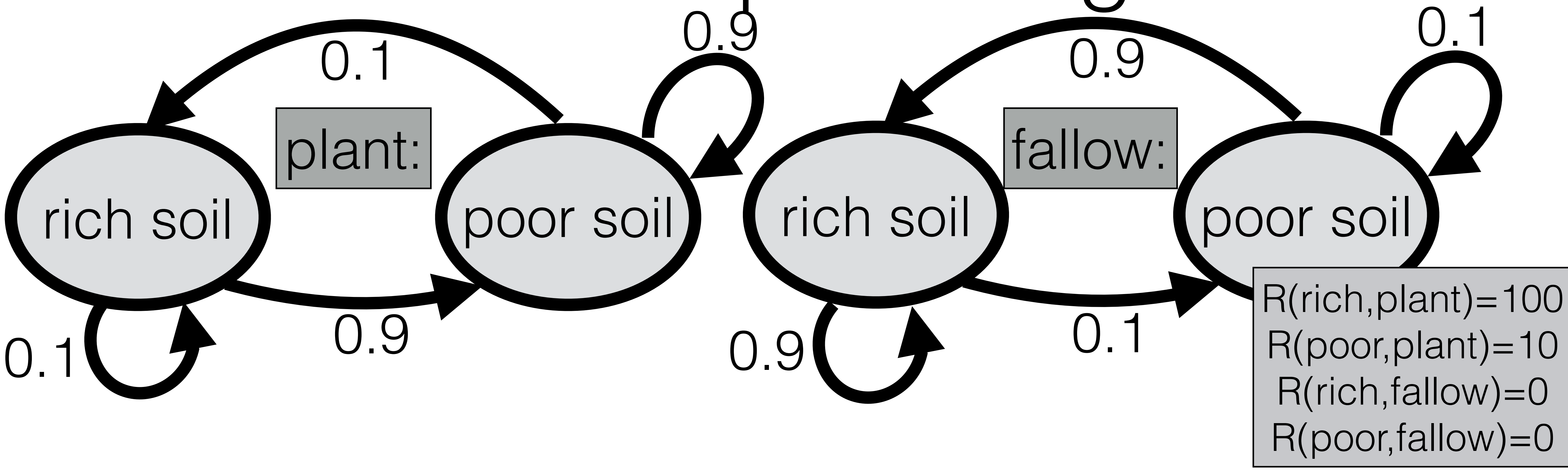
R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

What if I don't stop farming?



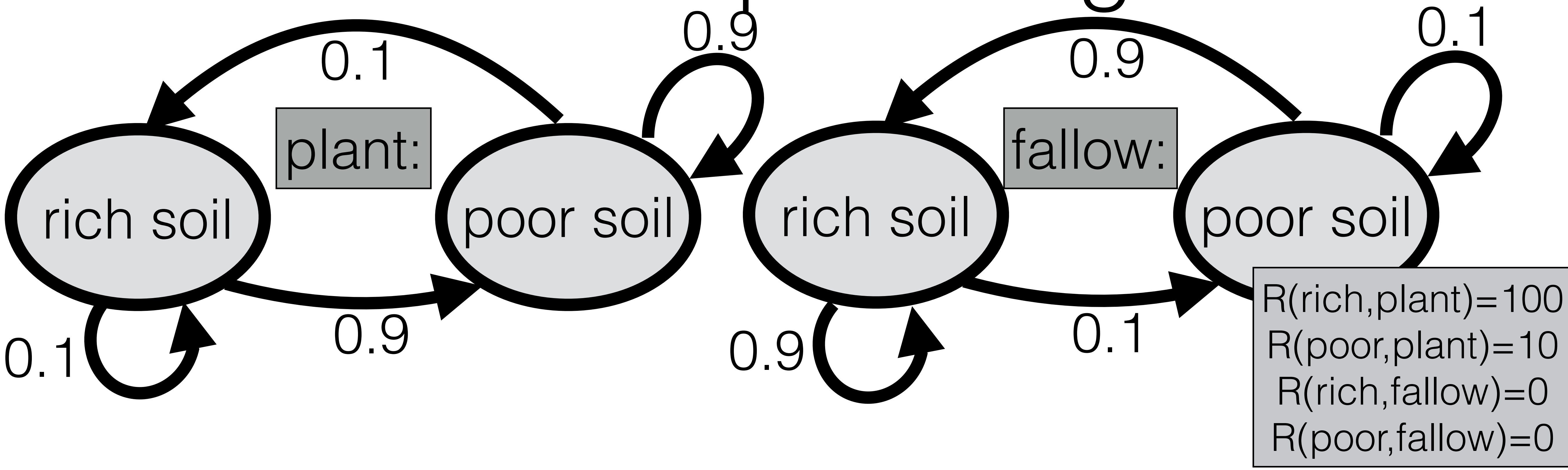
- Problem: 1,000 bushels today > 1,000 bushels in ten years

What if I don't stop farming?



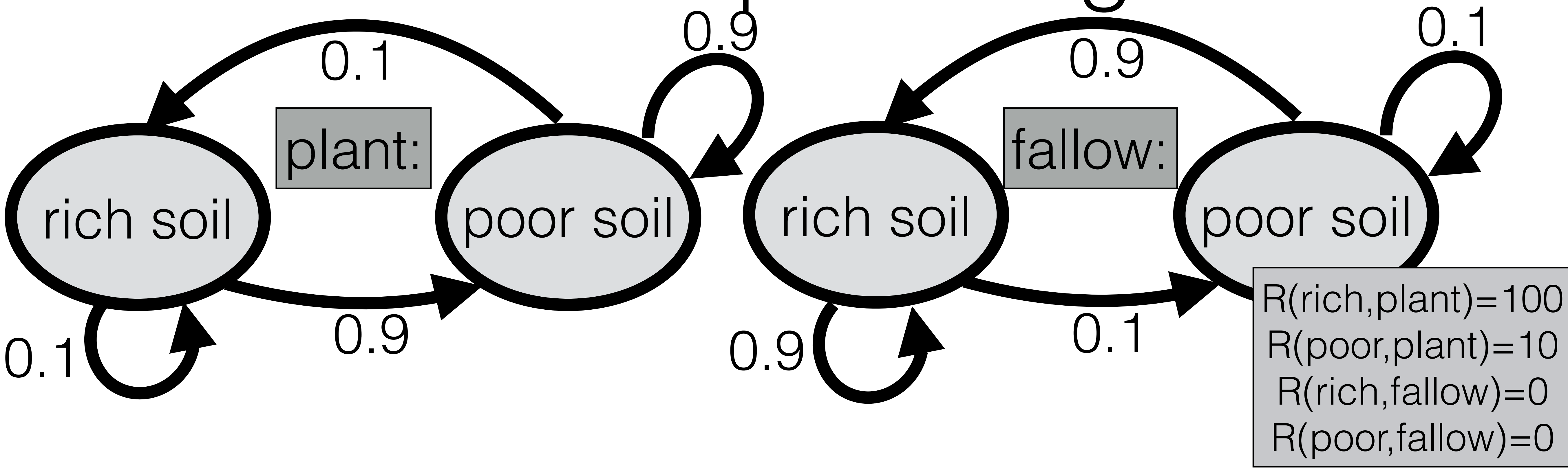
- Problem: 1,000 bushels today > 1,000 bushels in ten years
    - A solution: **discount factor** $\gamma : 0 < \gamma < 1$

What if I don't stop farming?



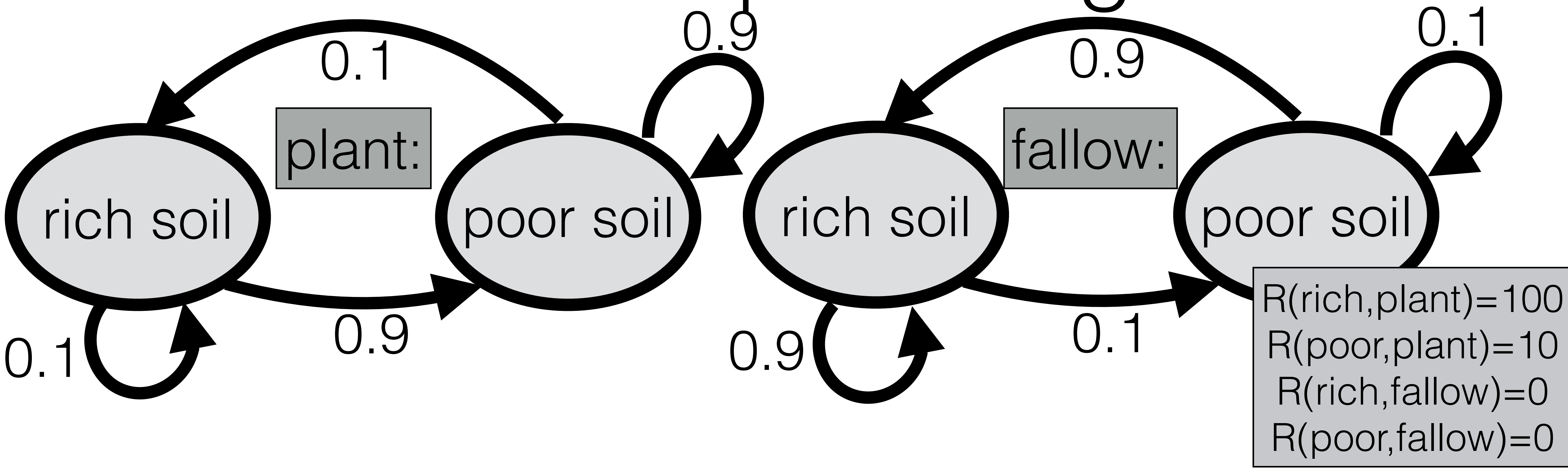
- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels

What if I don't stop farming?



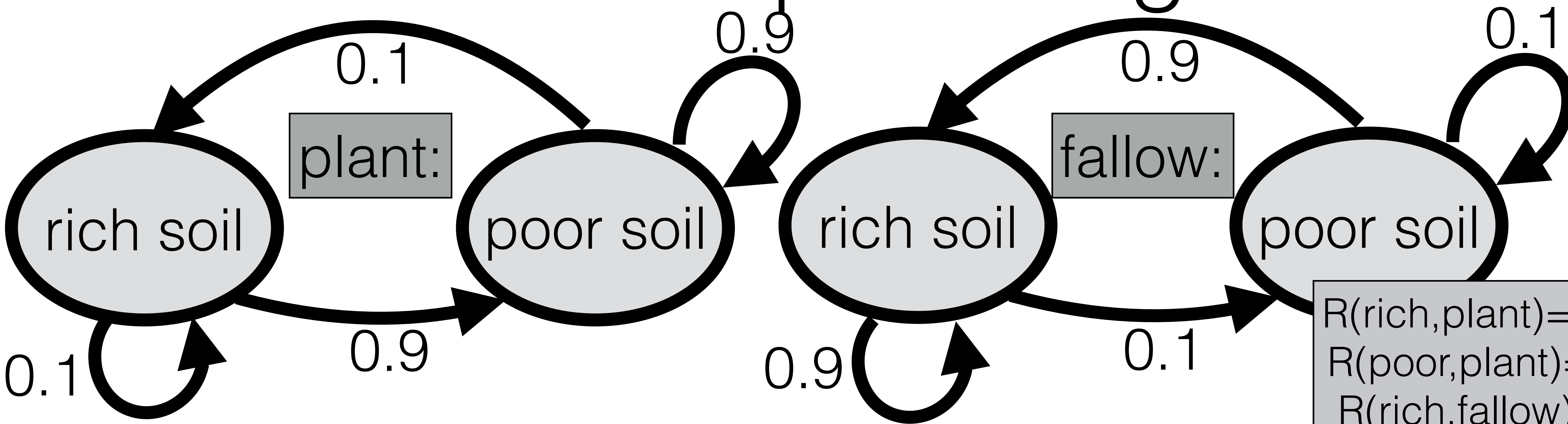
- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels
  - Example: What's the value of 1 bushel per year forever?

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: γ^t bushels
  - Example: What's the value of 1 bushel per year forever? V

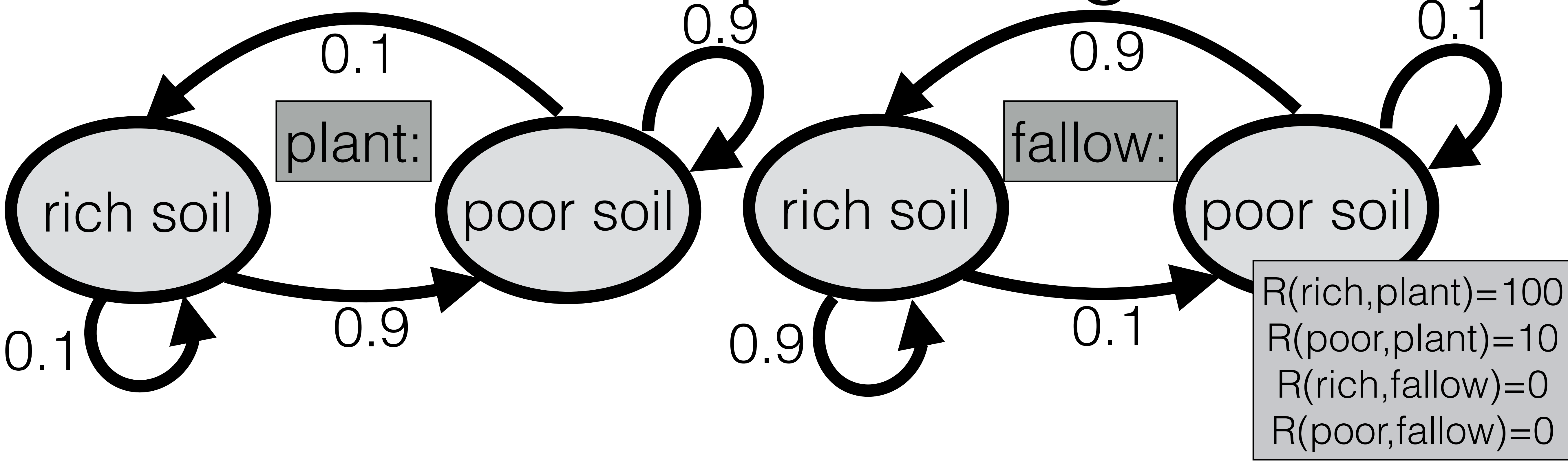
What if I don't stop farming?



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

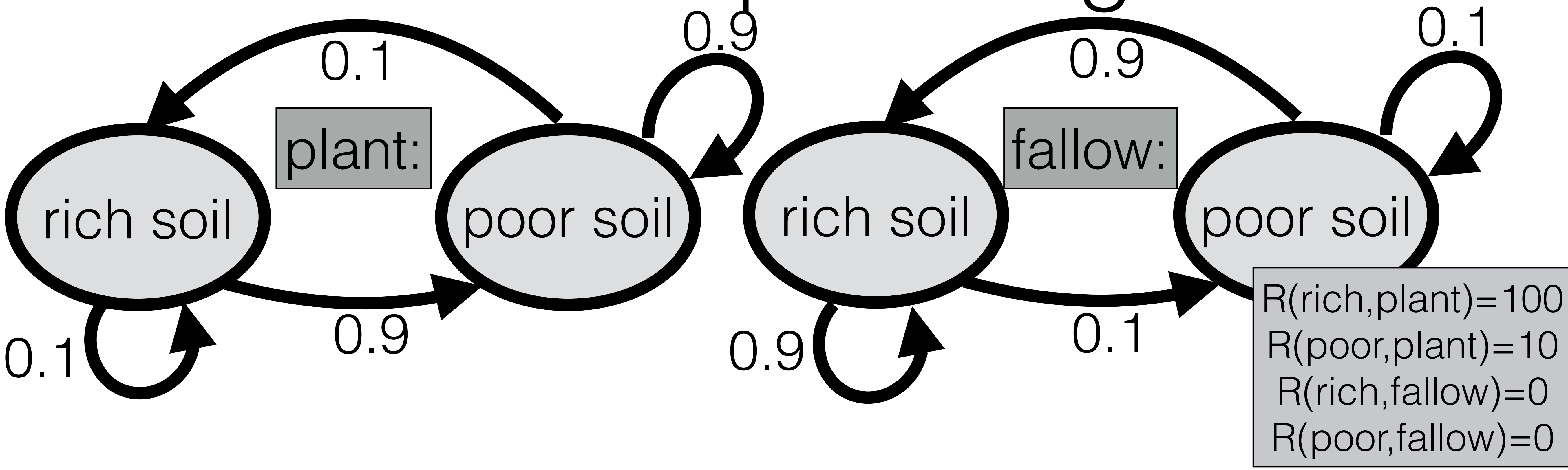
- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels
  - Example: What's the value of 1 bushel per year forever? $V = 1 + \gamma + \gamma^2 + \cdots$

What if I don't stop farming?



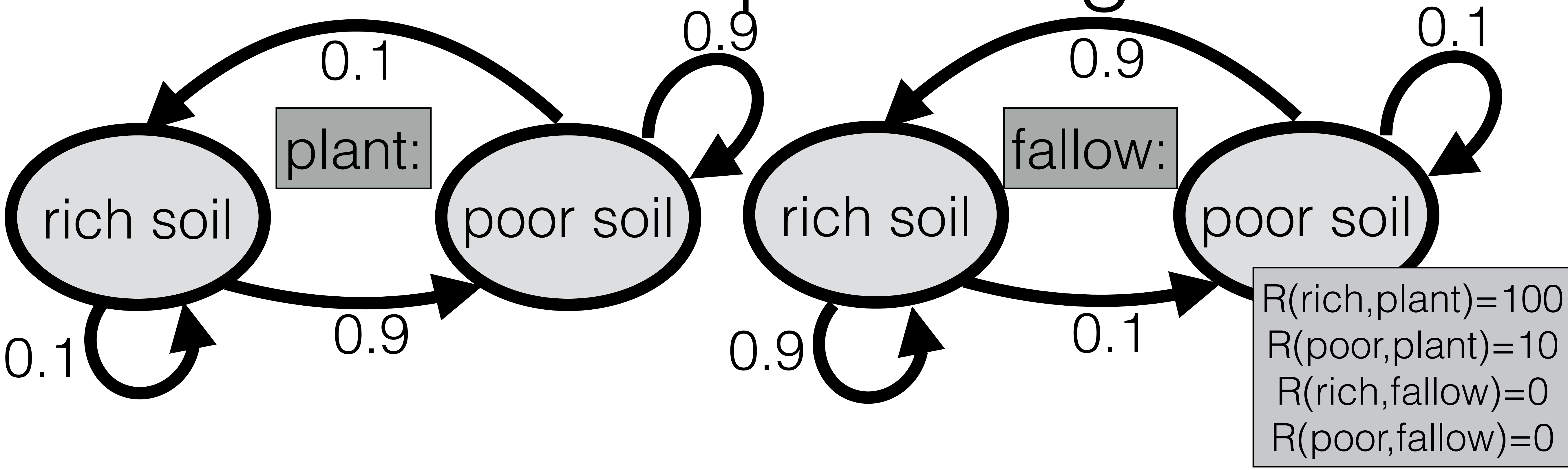
- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after t time steps: γ^t bushels
  - Example: What's the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma +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

What if I don't stop farming?



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

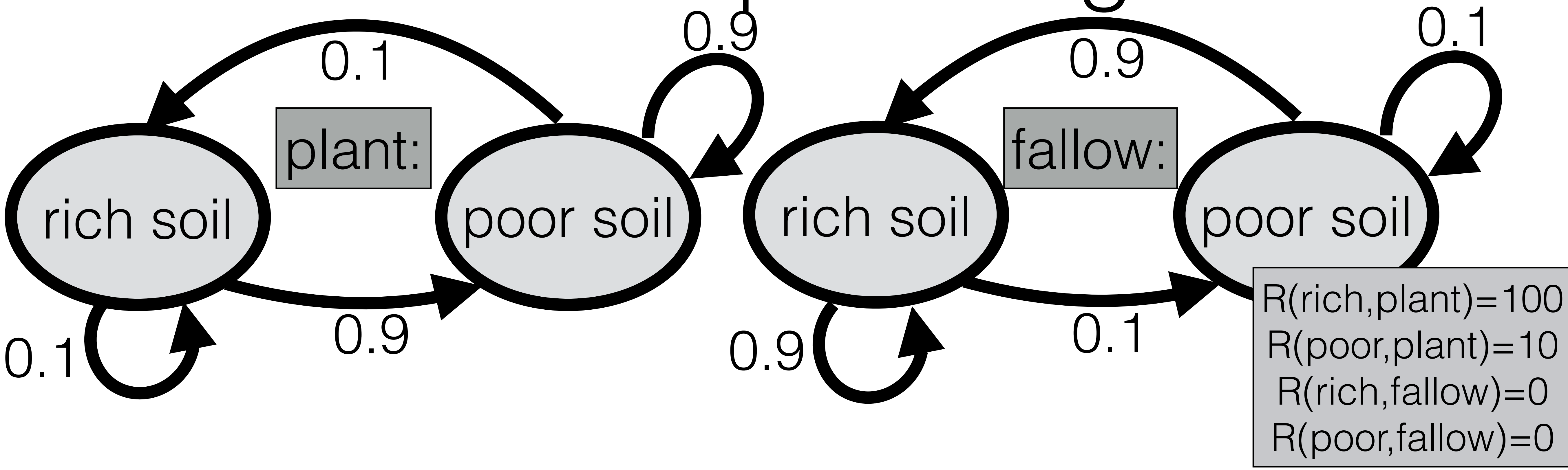
What if I don't stop farming?



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

$$\underbrace{V}_{\text{value for all future}} = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma + \gamma^2 + \cdots) = 1 + \gamma V$$

What if I don't stop farming?



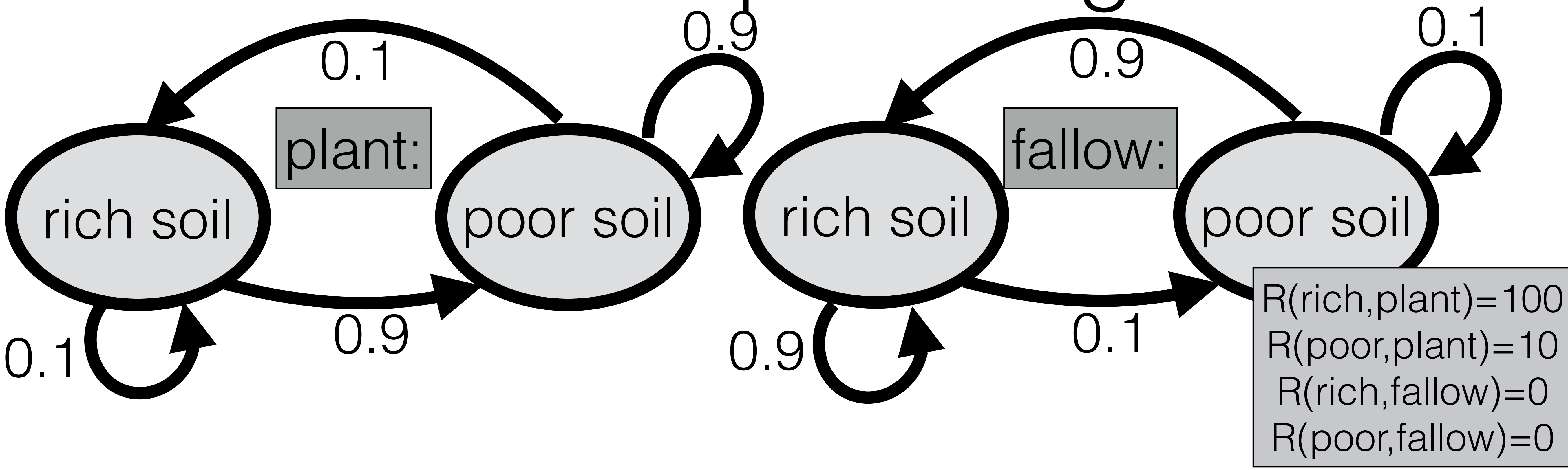
- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels
  - Example: What's the value of 1 bushel per year forever?

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

value for all future

value *on* first time step

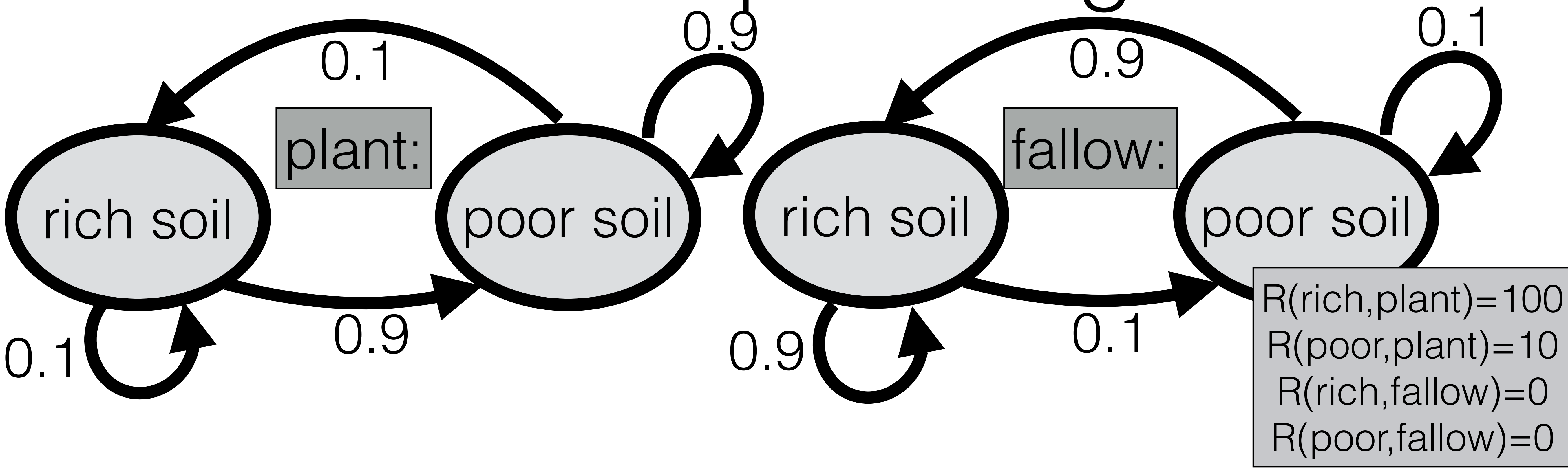
What if I don't stop farming?



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

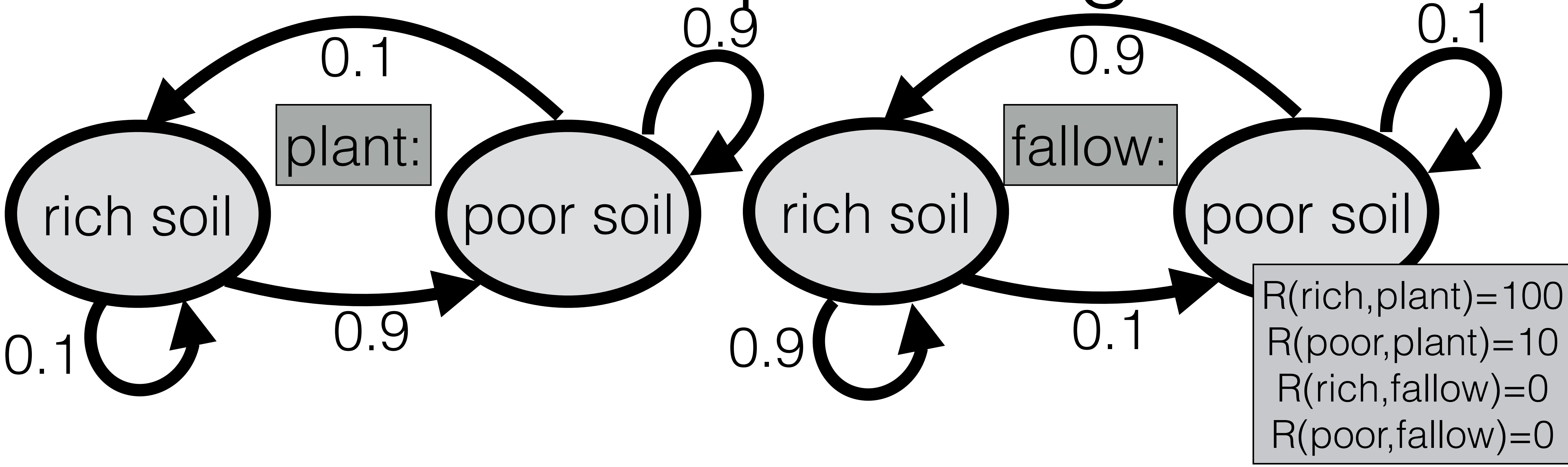
$$\underbrace{V}_{\text{value for all future}} = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma + \gamma^2 + \cdots) = \underbrace{1}_{\text{value } on \text{ first time step}} + \underbrace{\gamma V}_{\text{value } after \text{ first time step}}$$

What if I don't stop farming?



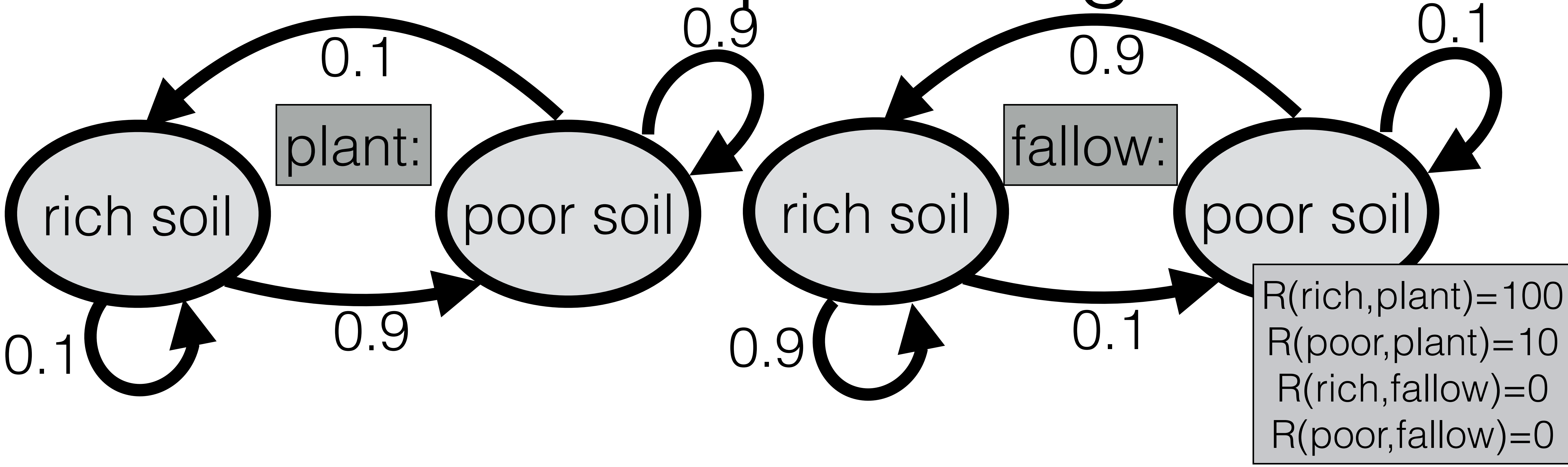
- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels
  - Example: What's the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma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

What if I don’t stop farming?



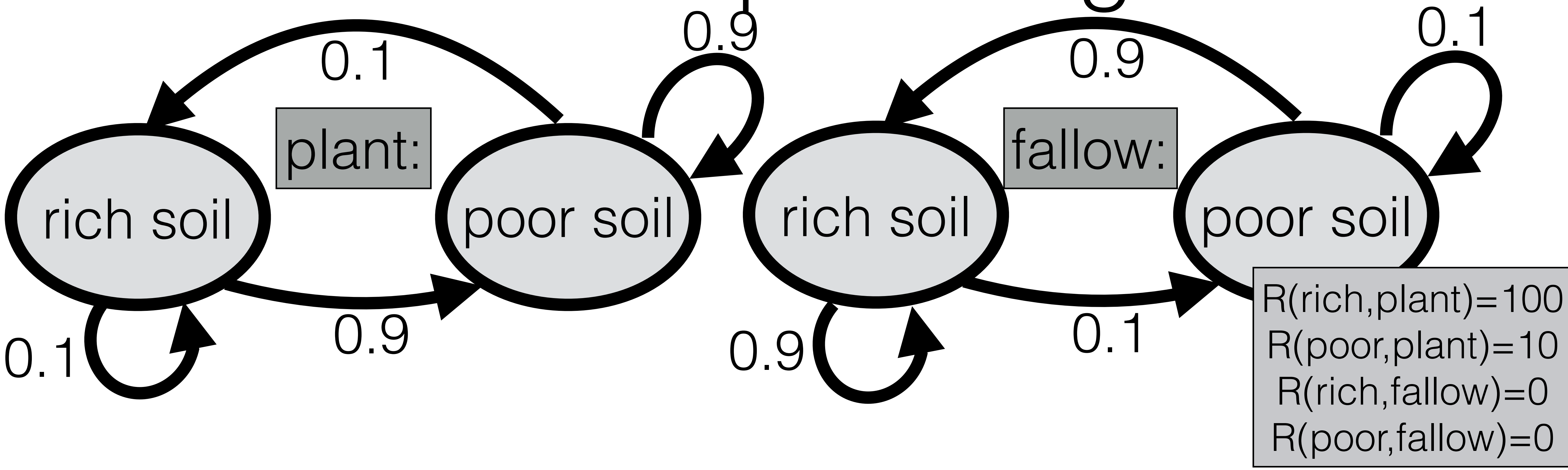
- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels
  - Example: What’s the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma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

What if I don't stop farming?



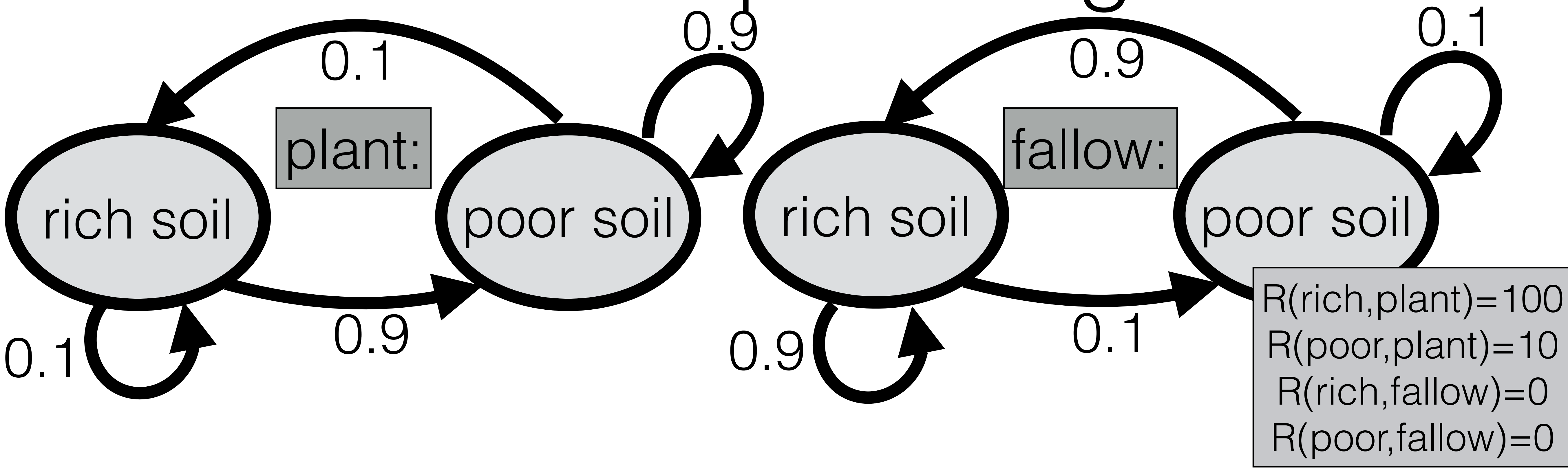
- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels
  - Example: What's the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma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

What if I don't stop farming?



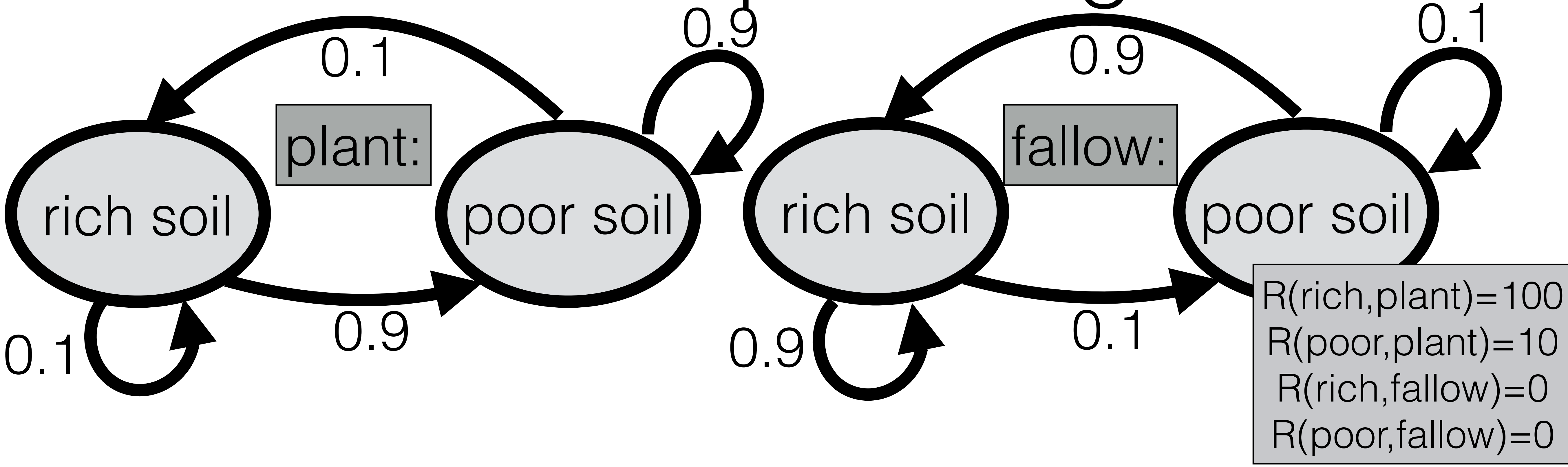
- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels
  - Example: What's the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma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

What if I don't stop farming?



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

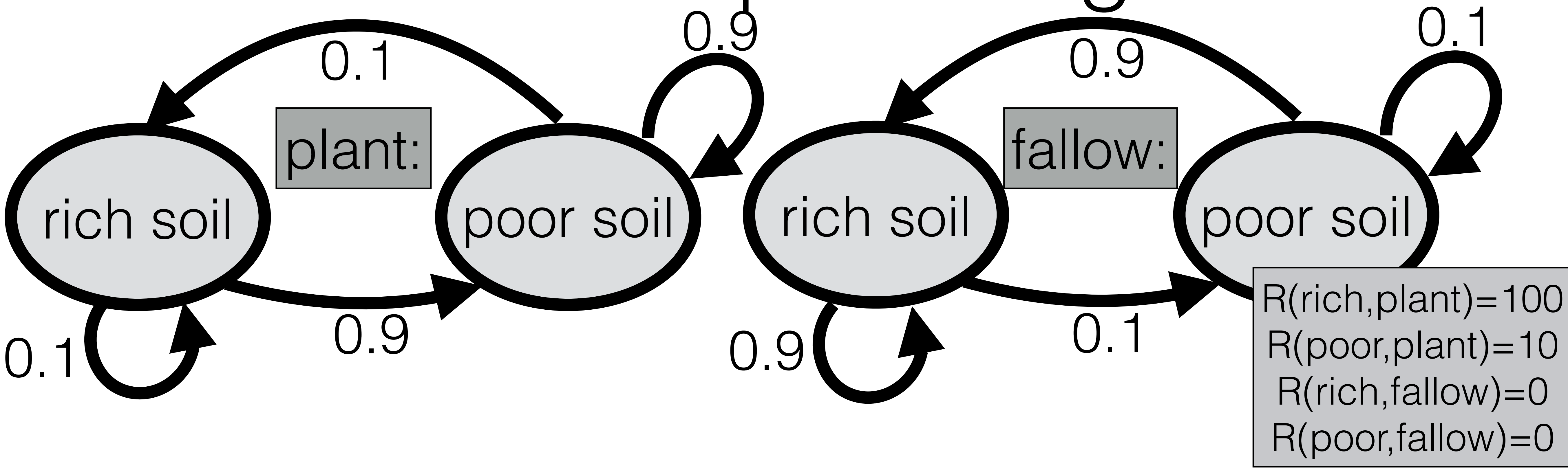
What if I don't stop farming?



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

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

What if I don't stop farming?



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

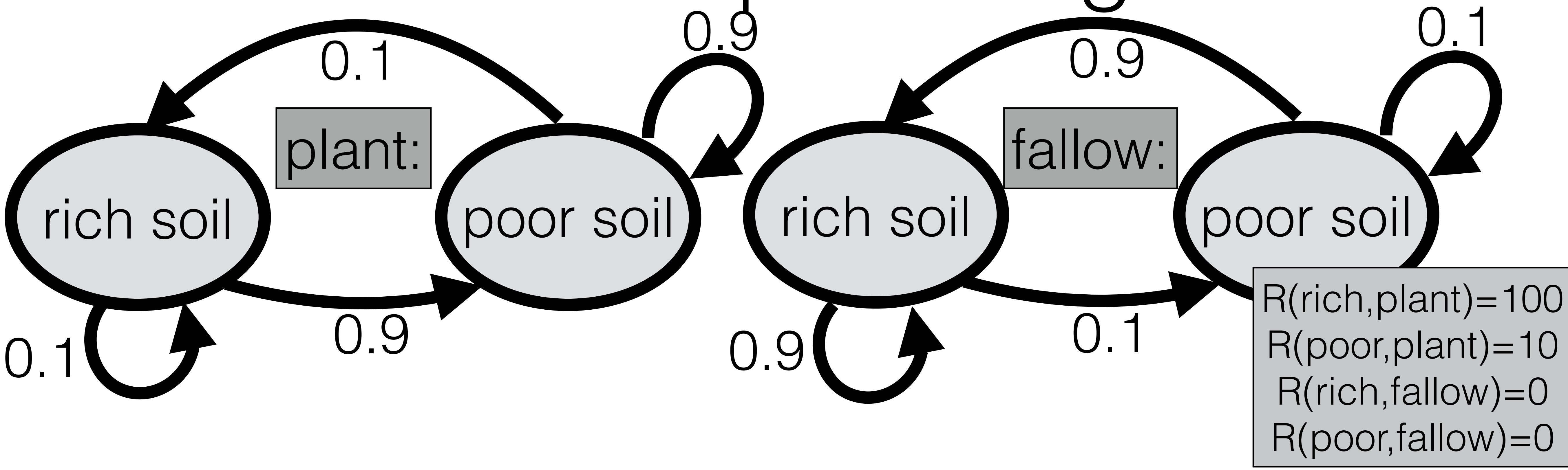
    $$V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma + \gamma^2 + \cdots) = 1 + \gamma V$$

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

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

policy value for all future

What if I don't stop farming?



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

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

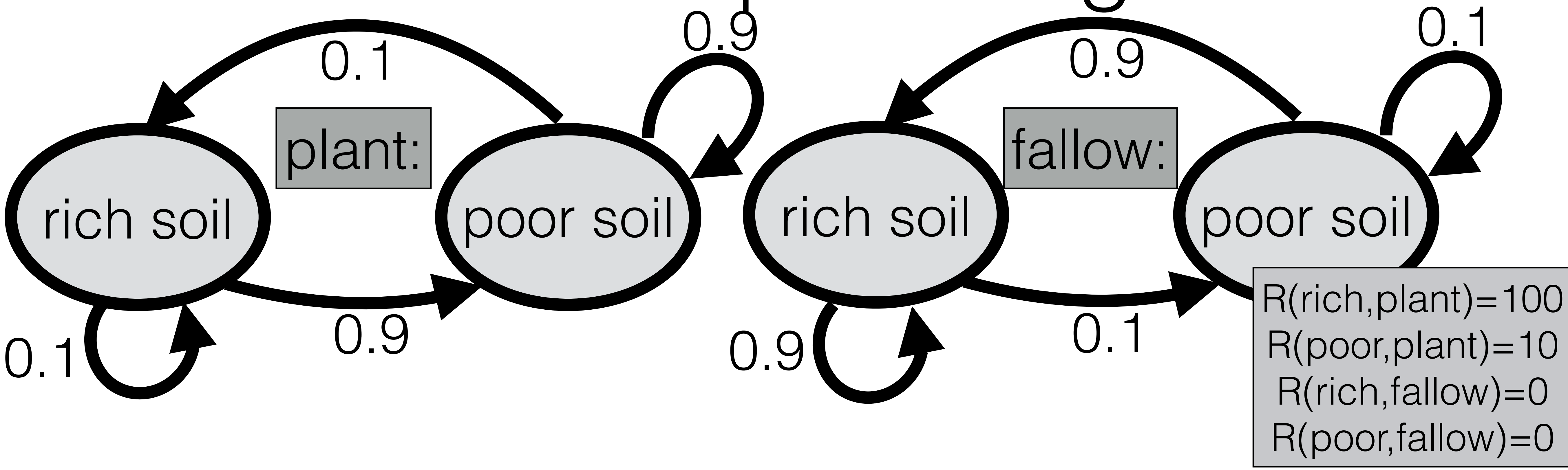
    $V = 1/(1 - \gamma)$ E.g. $\gamma = 0.99 \Rightarrow V = 1/0.01 = 100$ bushels
- $V_\pi(s)$: expected reward with policy π starting at state *s*

$$\underbrace{V_\pi(s)}_{\text{policy value for all future}} = \underbrace{R(s, \pi(s))}_{\text{policy value on first time step}} + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$$

policy value for all future

policy value on first time step

What if I don't stop farming?

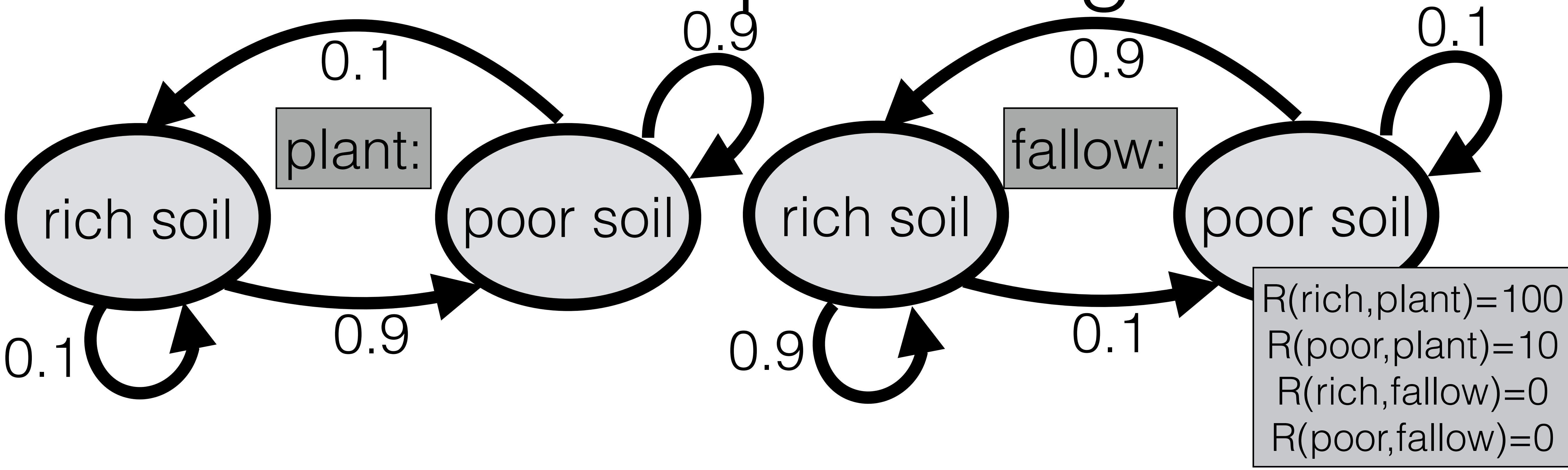


- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels
  - Example: What's the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma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

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

policy value for all future | policy value on first time step | (expected) policy value *after* first time step

What if I don't stop farming?

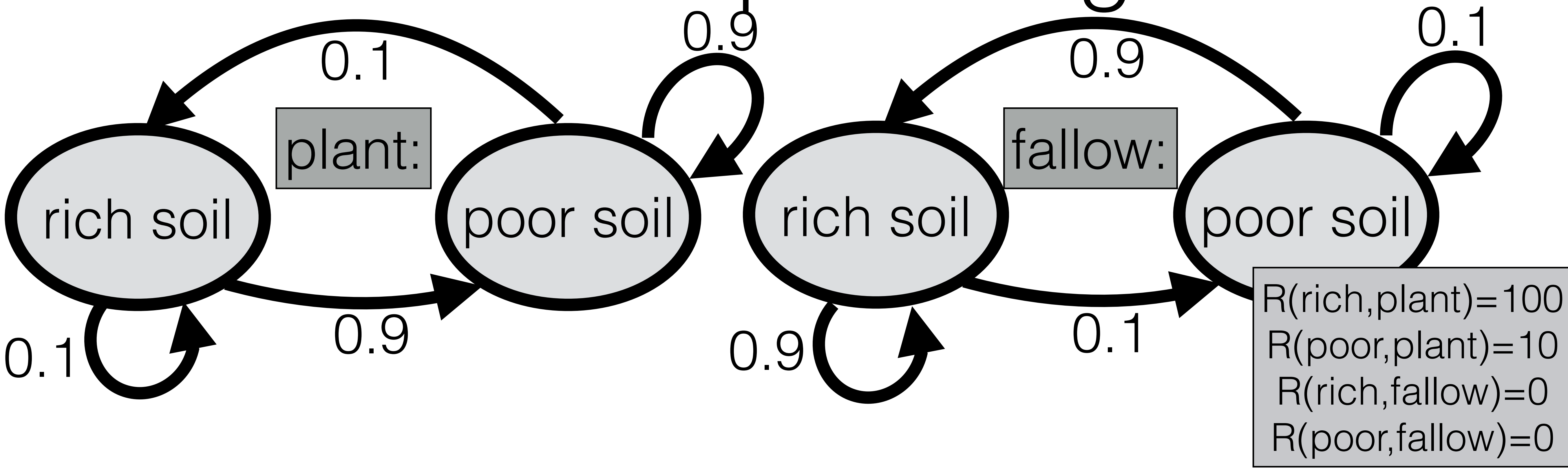


- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels
  - Example: What's the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma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

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

policy value for all future | policy value on first time step | (expected) policy value *after* first time step

What if I don't stop farming?



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

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

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

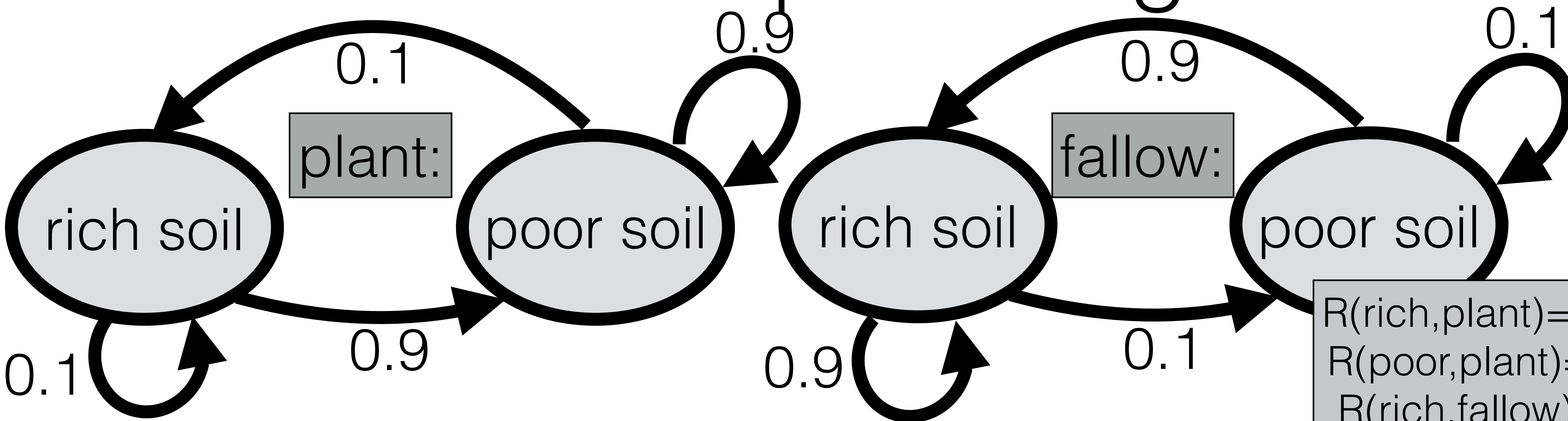
$$V_\pi(s) = R(s, \pi(s)) + \gamma \sum_{s'} T(s, \pi(s), s') V_\pi(s')$$

policy value for all future

policy value on first time step

(expected) policy value *after* first time step

What if I don't stop farming?



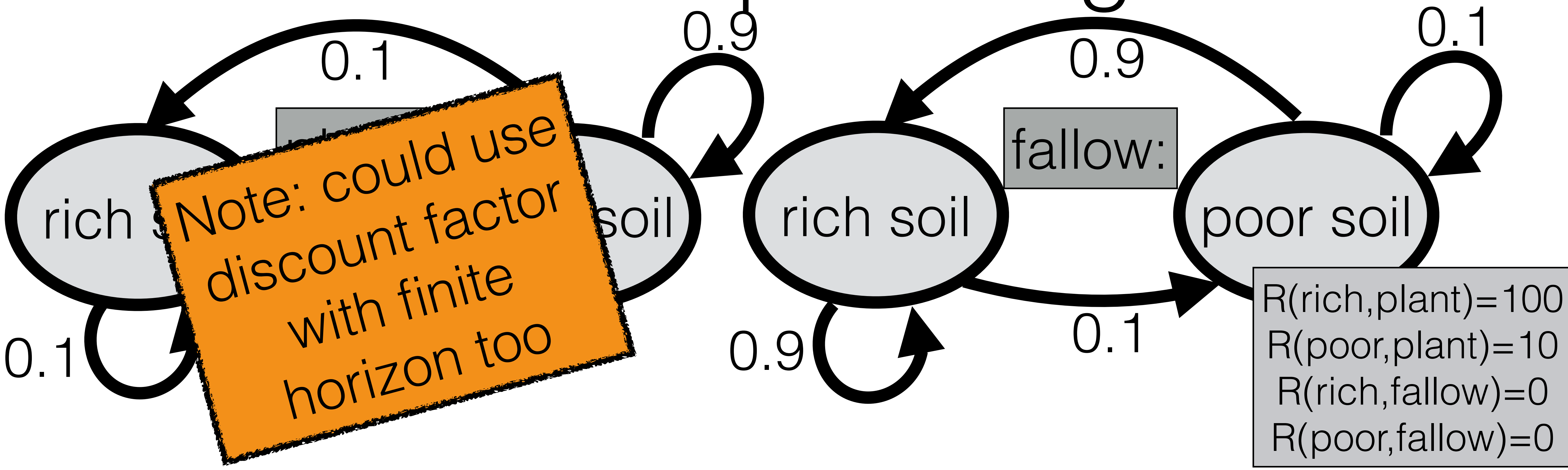
R(rich,plant)=100
R(poor,plant)=10
R(rich,fallow)=0
R(poor,fallow)=0

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

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

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

What if I don't stop farming?

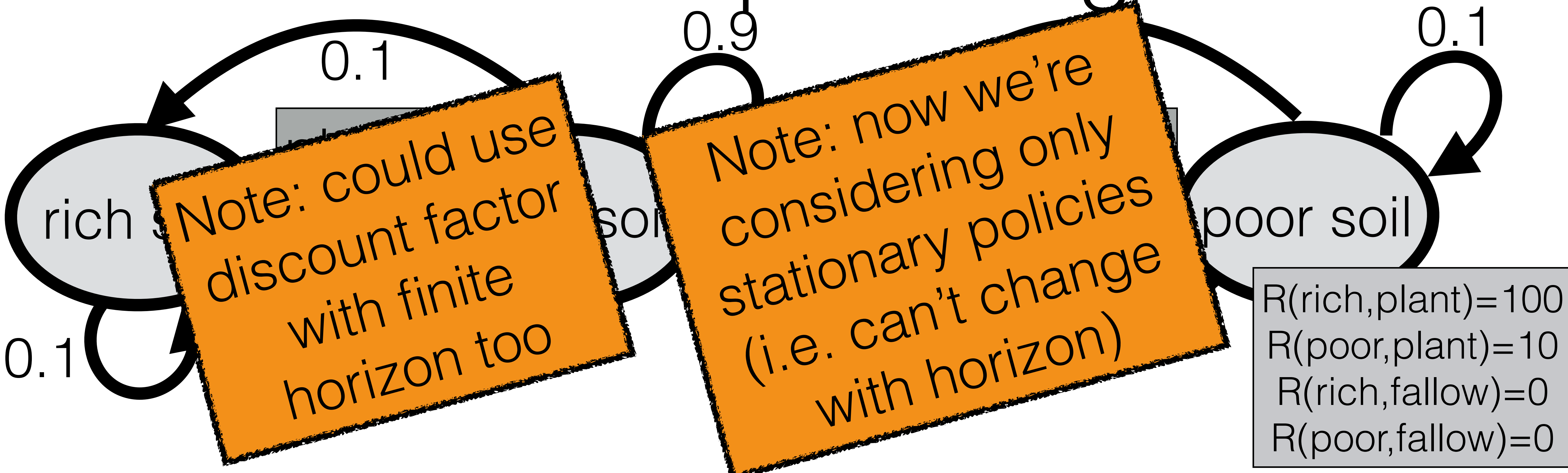


- Problem: 1,000 bushels today > 1,000 bushels in ten years
  - A solution: **discount factor** $\gamma : 0 < \gamma < 1$
  - Value of 1 bushel after *t* time steps: γ^t bushels
  - Example: What's the value of 1 bushel per year forever?
    $V = 1 + \gamma + \gamma^2 + \cdots = 1 + \gamma(1 + \gamma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

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

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

What if I don't stop farming?



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

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

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