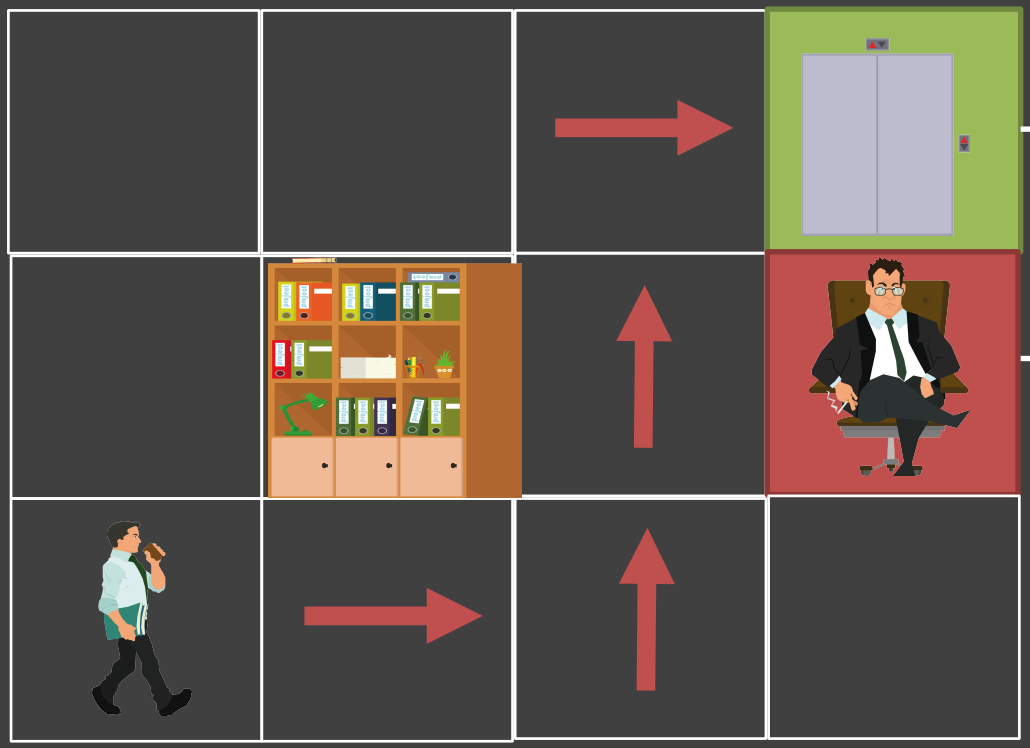
(PARTIALLY OBSERVABLE) MARKOV DECISION PROCESSES

Frederike Petzschner & Lionel Rigoux

What's the fastest way out?



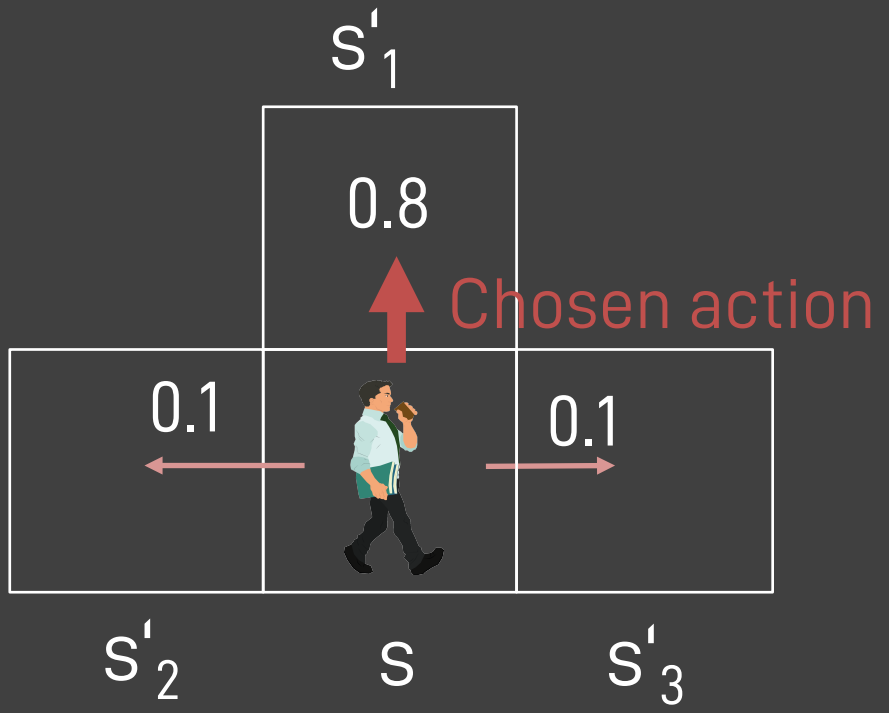
r = + 1

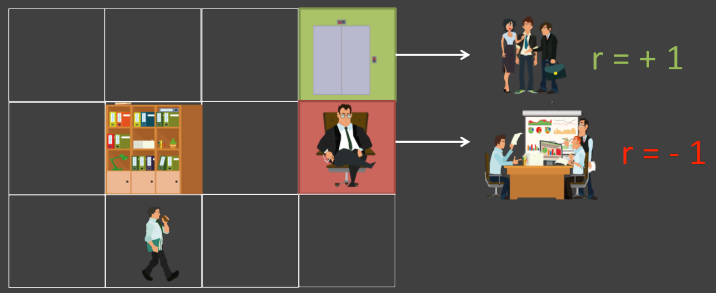
r = - 1

Plan: sequence of actions

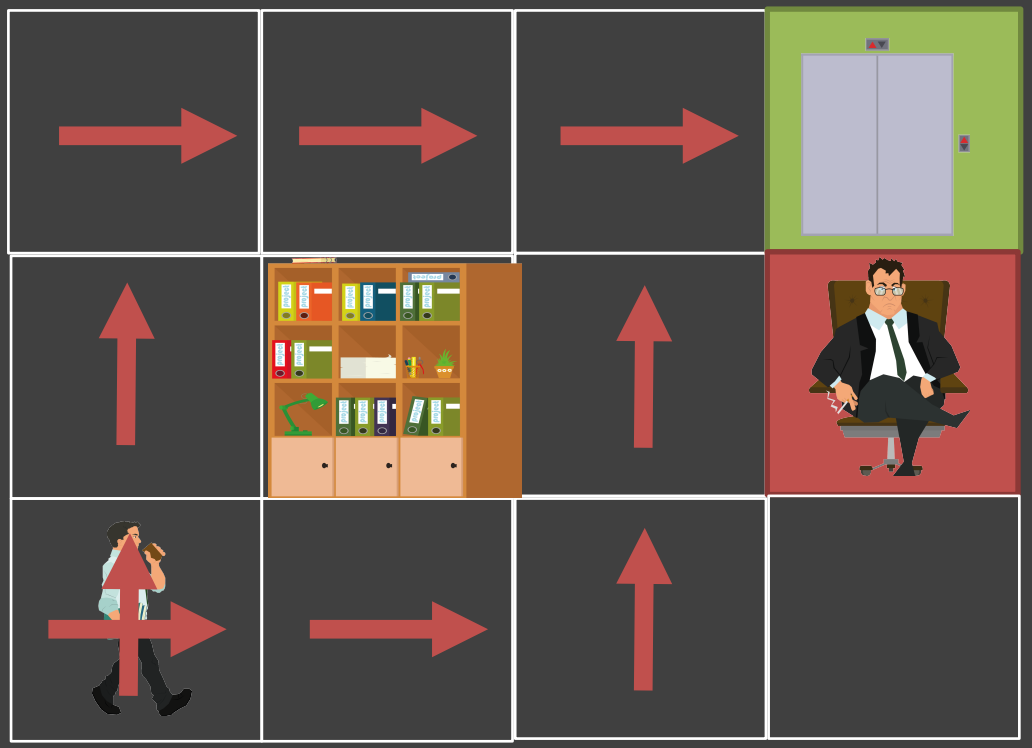
r_{step} = -0.04

Nondeterministic Action Rule





What's is the reliability of the action sequence: up, up, right, right, right?



r = + 1

r = - 1

Answer: 0.32776

$= 0.8^5 + 0.1^4\, 0.8$

$r_{step} = -0.04$

The MDP is defined by:

States s → S (state space) (Start state; Maybe: terminating state)

Actions a → A (action space)

Transition Function: T (s, a, s'): P(s'|s,a)

Reward Function: R (s, a, s'), R (s, a), R (s)

Policy $\pi(s)$ → a

π^* → optimal policy

MDP is a nondeterministic search problem.

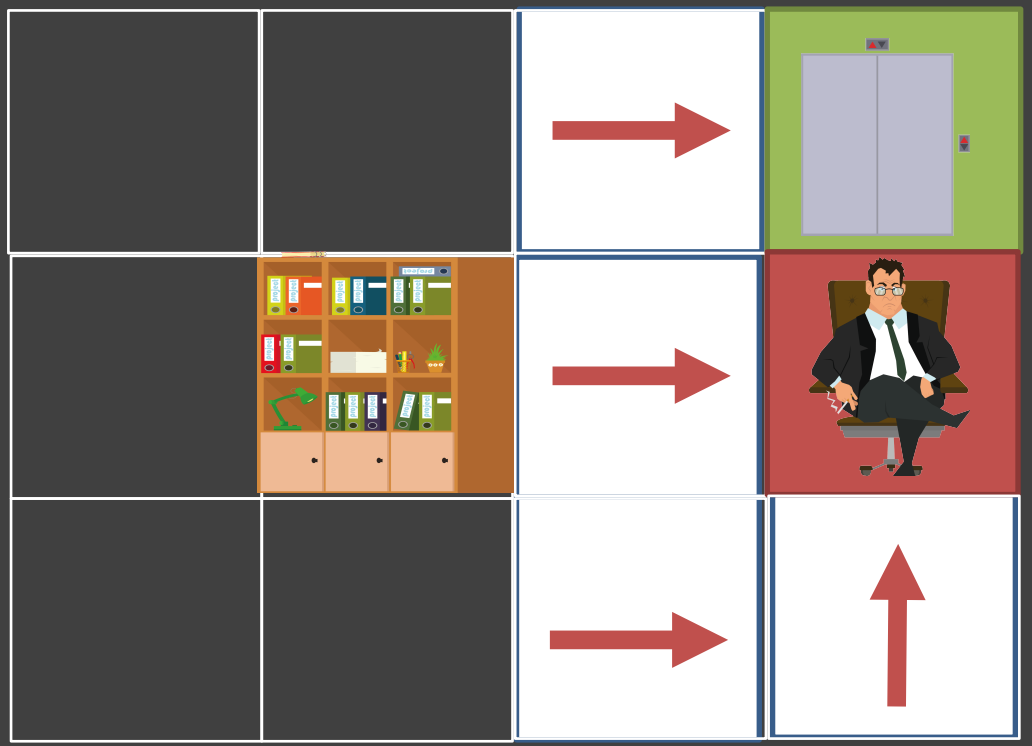
Whats Markovian about an MDP?

Future and Past are independent.
Action outcomes only depend on your current state.

$$P(S_{t+1} = s' \mid S_t = s_t, A_t = a_t, S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}, \ldots) = P(S_{t+1} = s' \mid S_t = s_t, A_t = a)$$

Not every process is an MDP!

What if the reward structure of the world changes.

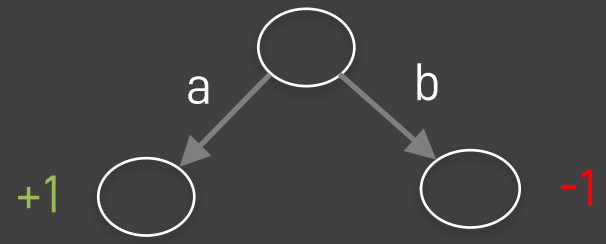


Quiz: what's the best strategy in the four white fields?

$r_{step} = -2$

We have been looking at sequences of reward

[-1] < [+1]



$$U(s_0, s_1, s_2, \dots) = R(s_0) + R(s_1) + R(s_2) + \cdots$$

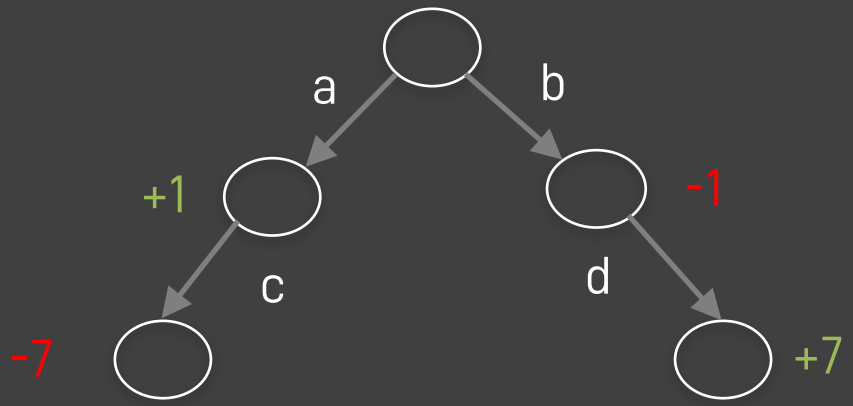
$$U(s_0, s_1, s_2, \dots) = \sum_{t=0}^{\infty} R(s_t)$$

Sequences of rewards

Utility depends on all successor states !!! (longterm)

[+1-7] < [-1+7]

[-6] < [+6]

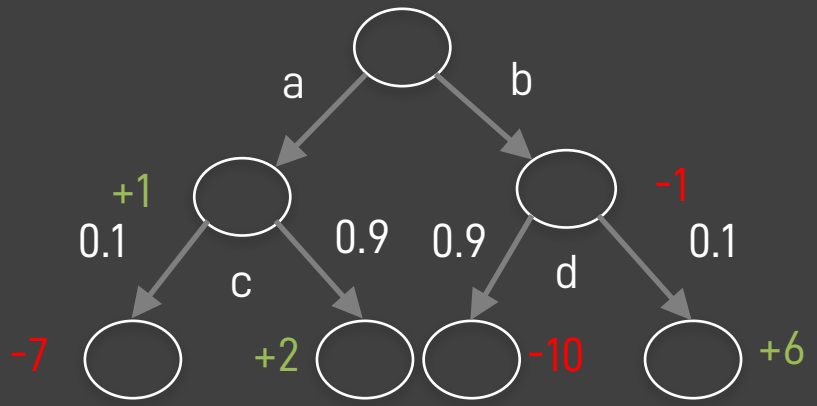


$$U(s_0, s_1, s_2, \dots) = \sum_{t=0}^{\infty} R(s_t)$$

Utility of a sequence of states in non-deterministic worlds

[+1,-7*0.1 + 0.9*2] < [-1,+6*0.1 -10*0.9]

[+2.1] < [-9.8]



$$U(s_0, s_1, s_2, \dots) = E\left[\sum_{t=0}^{\infty} R(s_t)\right]$$

Sequences of rewards

[+1,+20,+1,+20,+1+1,.....]

[+1,+1,+1,+1,+1+1,.....]

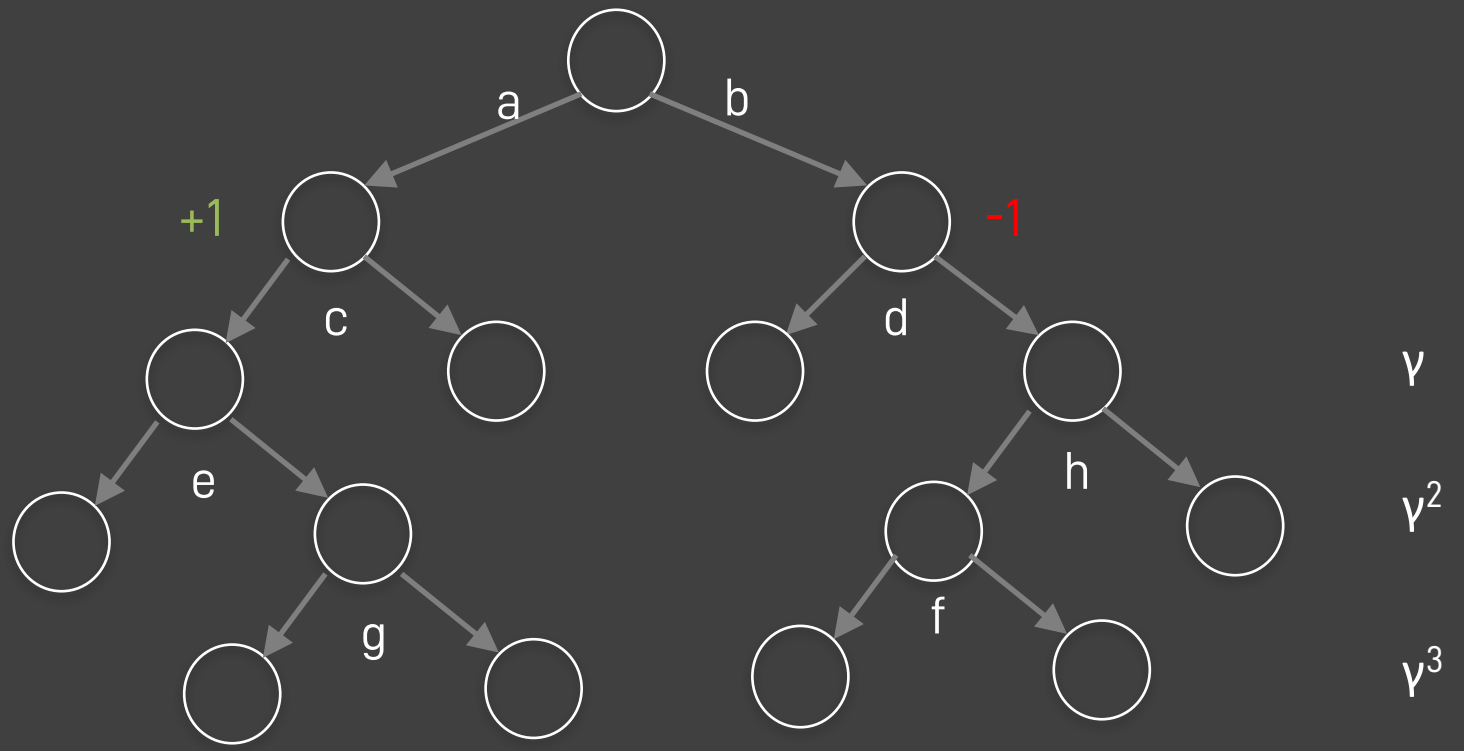
Discounting

1 γ γ^2

$$U(s_0, s_1, s_2, \dots) = E\left[\sum_{t=0}^{\infty} y^t R(s_t)\right]$$

$0 < \gamma < 1$

Discounting



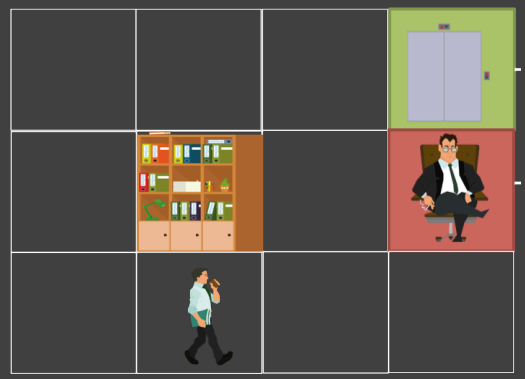
Policies

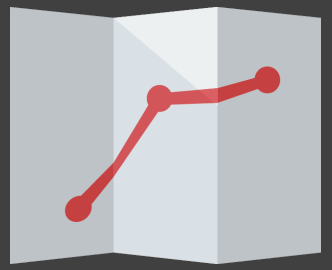
We want a plan! But this is not a deterministic world!
Plan: mapping from states to actions

Policy π: states → actions

- It's like an if-then-plan
- look-up table

$$U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} y^t R(s_t) \,\middle|\, \pi\right]$$





Optimal Policy π^*: states → actions

- maximized expected utility

$$\pi^* = \max_{\pi}\left(E\left[\sum_{t=0}^{\infty} y^t R(s_t) \,\middle|\, \pi\right]\right)$$

$$U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} y^t R(s_t) \,\middle|\, \pi\right]$$

$$U^{\pi}(s) = R(s) + y \max_a \sum_{s'} T(s, a, s') \cdot U(s')$$

BELLMAN EQUATION

Let's find the optimal policy….

- n equations
- n unknowns
- but…

How to act optimal? – value iteration

Step 1: Start with arbitrary utilities. Take the correct first action
Step 2: Update utilities based on their neighbours

Keep going….until convergence (Adding truth to wrong)

„An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.“

– Bellman, 1957

Let's do it - value iteration

$$U^{\pi}(s) = R(s) + y \max_{a} \sum_{s'} T(s,a,s') \cdot U(s')$$

$$U_{t+1}^{\pi}(s) = R(s) + y \max_{a} \sum_{s'} T(s,a,s') \cdot Ut(s')$$

$$U_{t=1}(white) = -.04 + .5 \cdot (0 + 0 + .8 \cdot 1) = .36$$

$$U_{t=2}(white) = -.04 + .5 \cdot (.1 \cdot .36 + .8 \cdot 1 - .1 \cdot .04) = .376$$

step cost: ⟶ r = - 0.04 y= 0.5 $U_0(s) = 0$

SUM

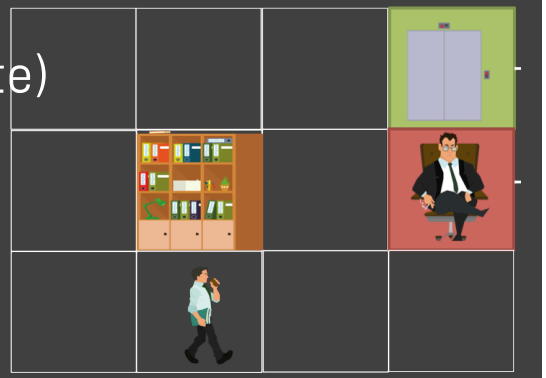
MDP: Non-deterministic search problem

States s → S (state space) (Start state; Maybe: terminating state)

Actions a → A (action space)

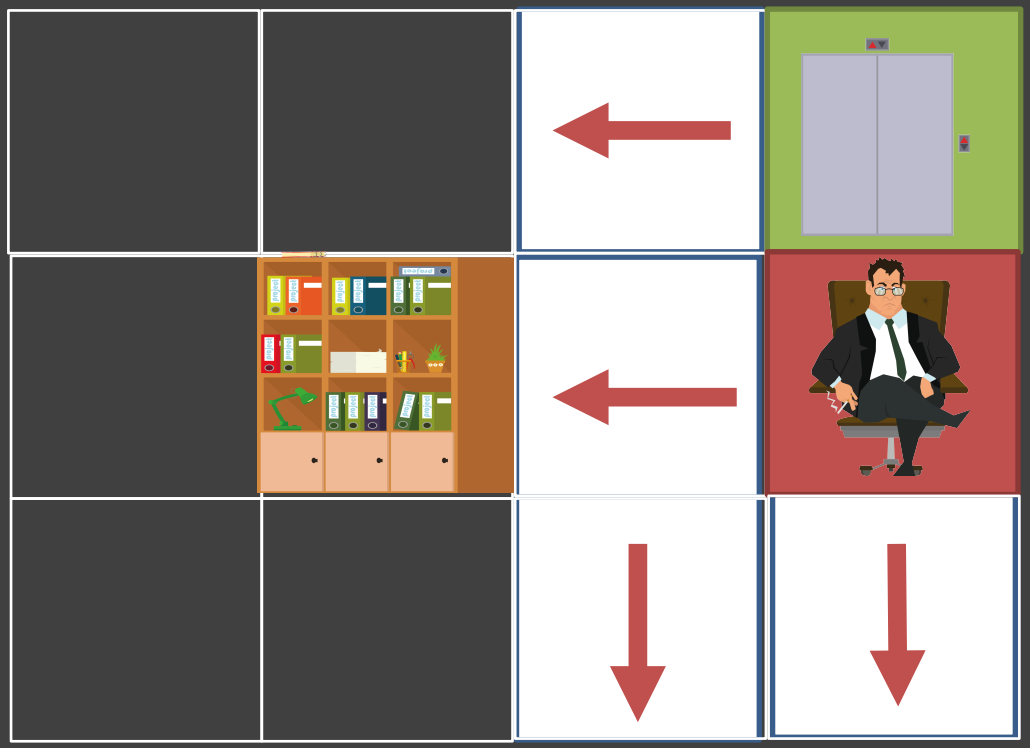
Transition Function: T (s, a, s'): P(s'|s,a)

Reward Function: R (s, a, s'), R (s, a), R (s)



→ POMDP: Uncertainty about states

What if the reward structure of the world changes.



Quiz what's the best strategy in the four white fields?

$r_{step} = +2$