

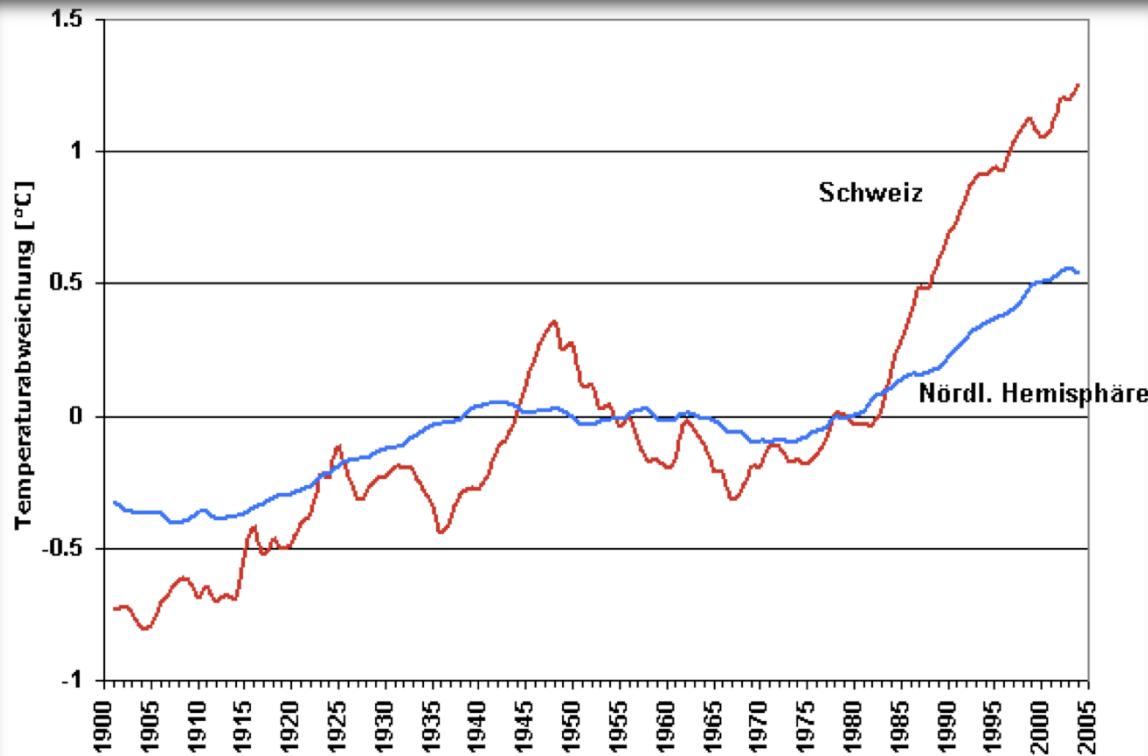
(PARTIALLY OBSERVABLE)
MARKOV DECISION PROCESSES

DR FREDERIKE PETZSCHNER

DRLIONEL RIGOUX

13122015

What happened from 30. November to 12. December 2015?



Past



Present



Future



Countries will aim to keep global temperatures from rising more than 2°C (3.6°F) by 2100 with an ideal target of keeping temperature rise below 1.5°C (2.7°F).

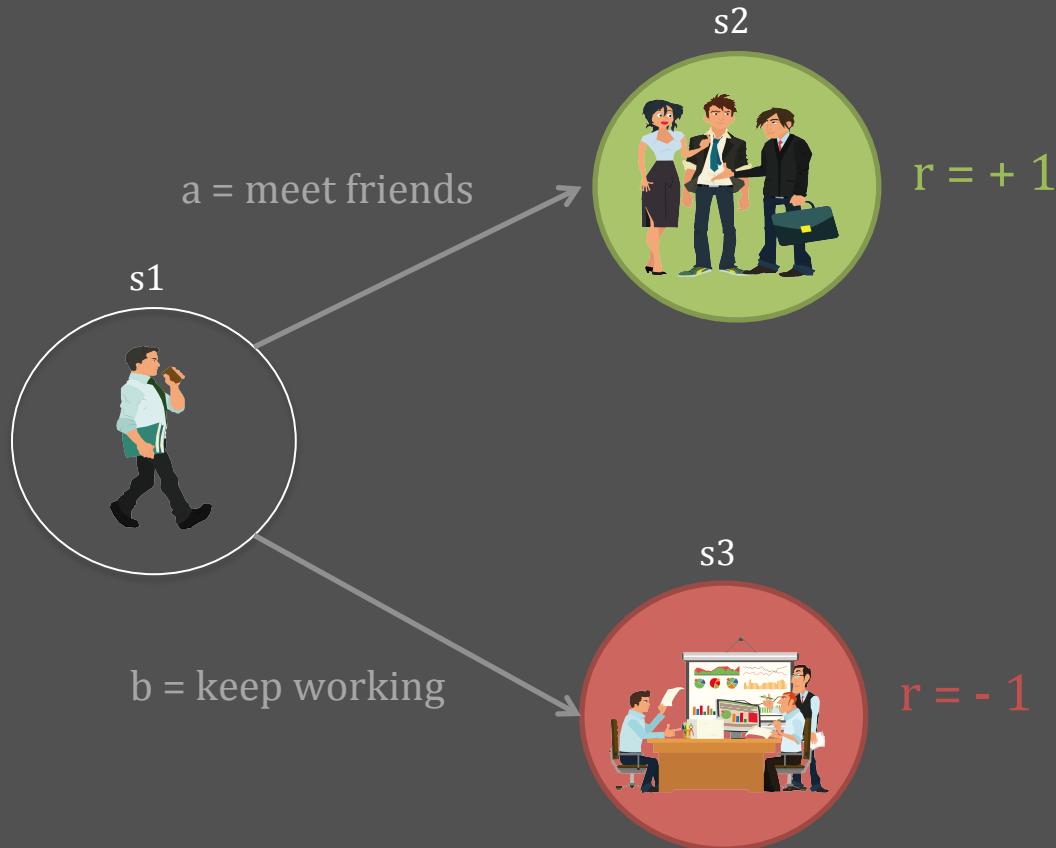


PLANNINGPROBLEM

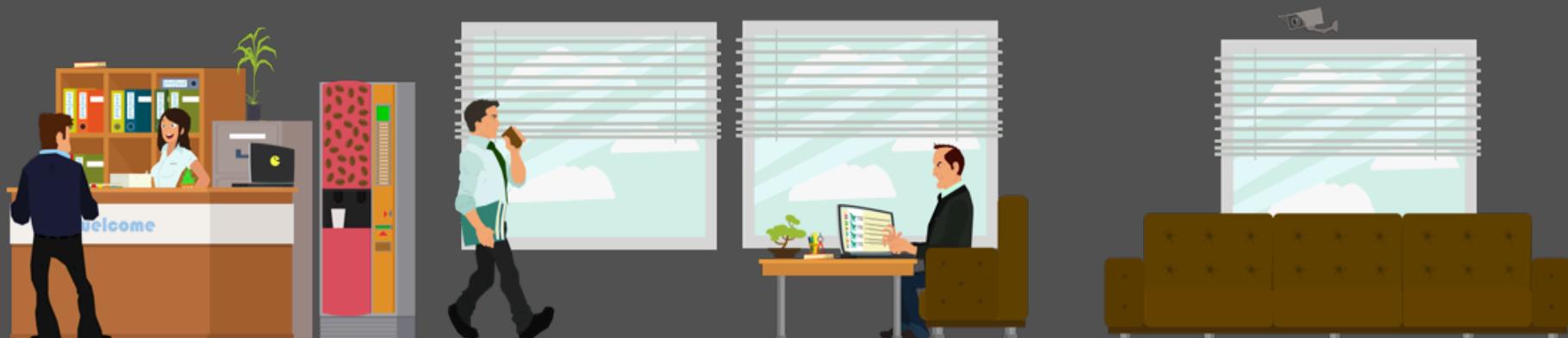
Given a complete and correct model
of the world dynamics and a reward
structure, what is the optimal way
to behave?

Littman & Cassandra

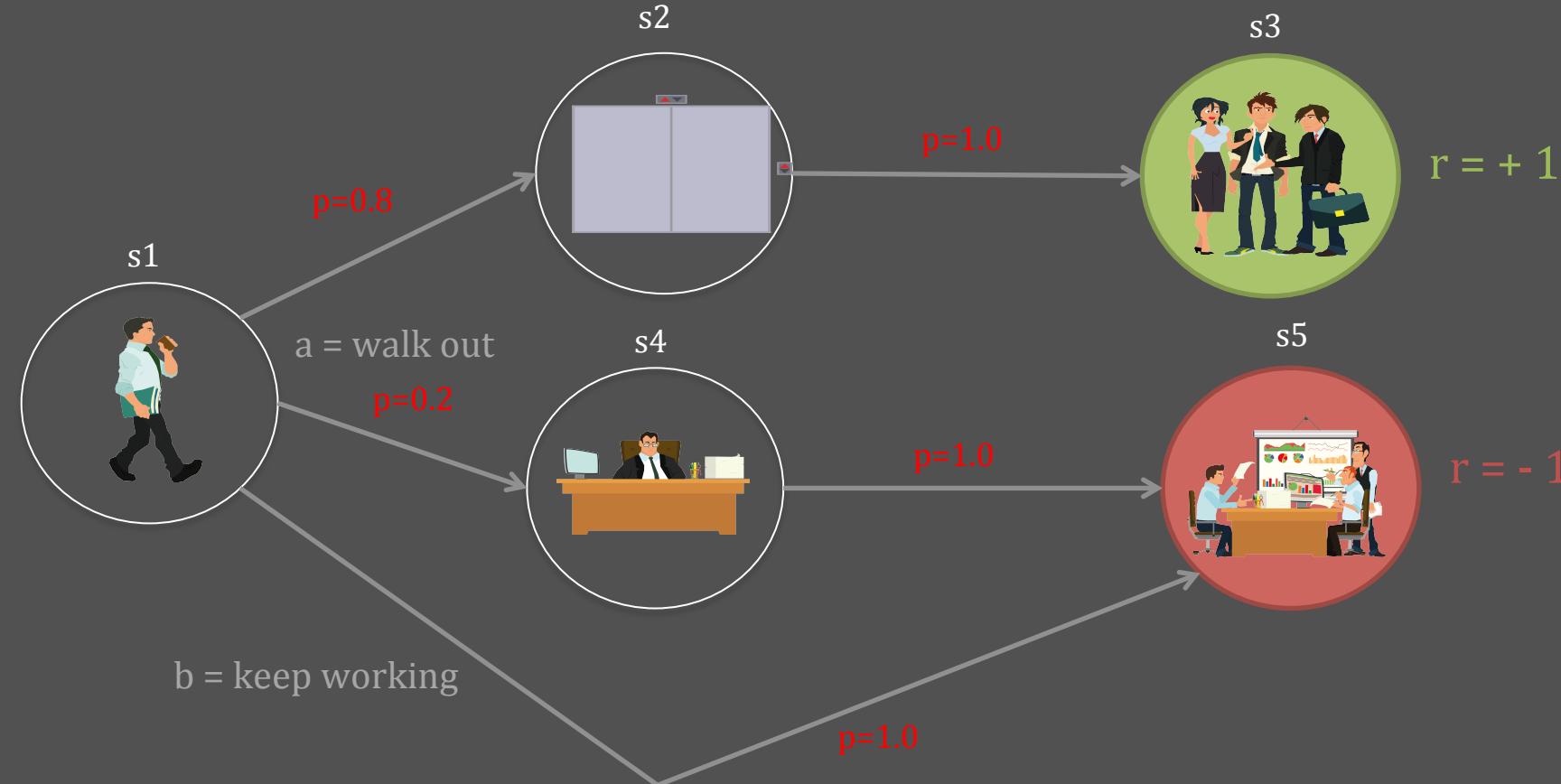
DETERMINISTIC WORLD



NONDETERMINISTIC ACTIONS



NONDETERMINISTIC WORLD



The MDP is defined by:

States $s \rightarrow S$ (state space)

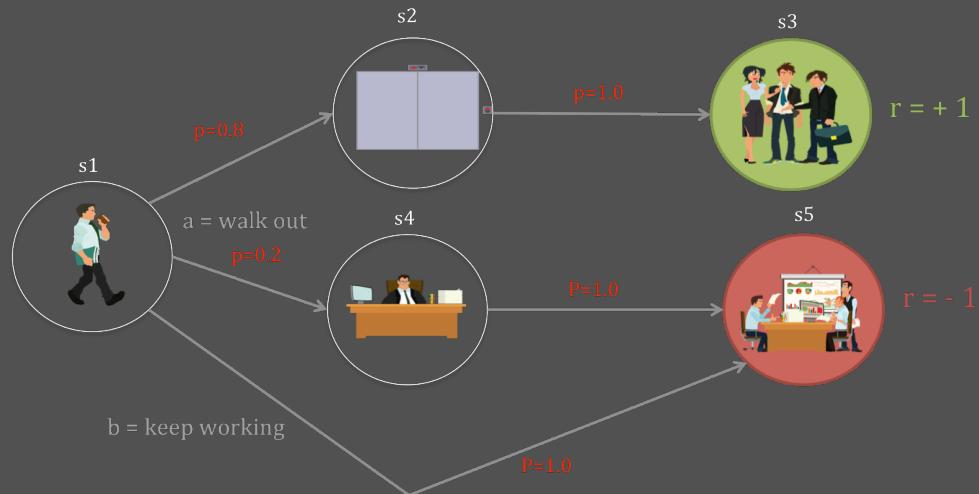
Actions $a \rightarrow A$ (action space)

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

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

Start state

Maybe: terminating state



MDP is a nondeterministic search problem.

s	a	s'	p
s_1	a	s_4	0.2
s_1	a	s_2	0.8
s_1	b	s_5	1

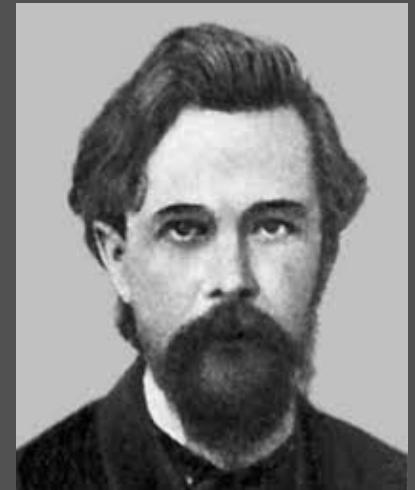
WHATS MARKOV ABOUT AN MDP?

Future and Past are independent.

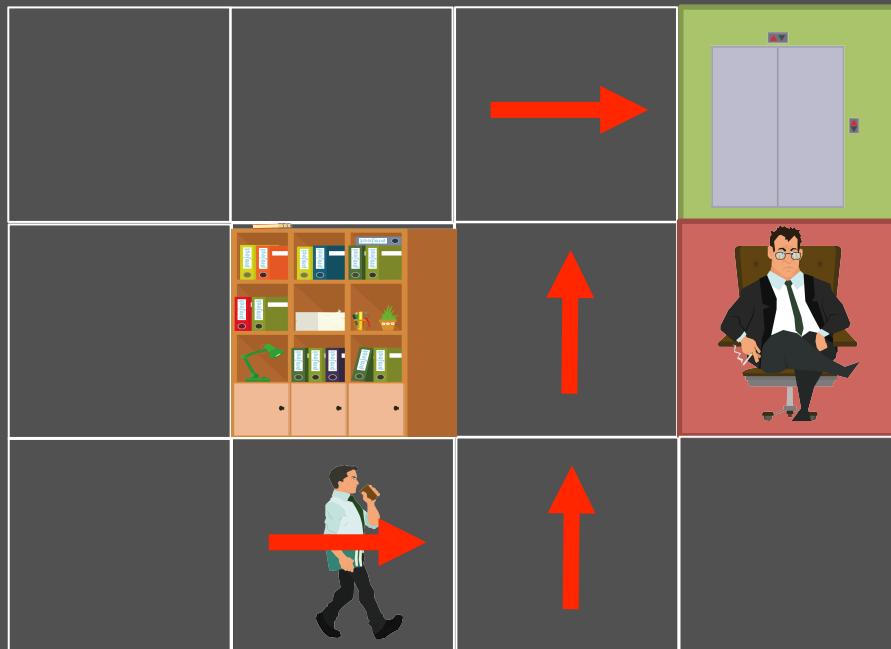
Action outcomes only depend on your current state.

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

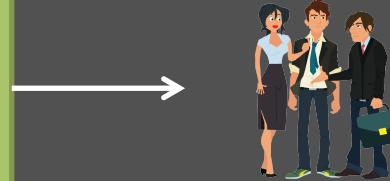
Not every process is an MDP!



What's the best way out?



step cost: $r = -0.01$



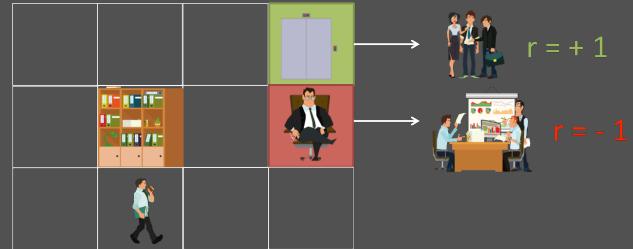
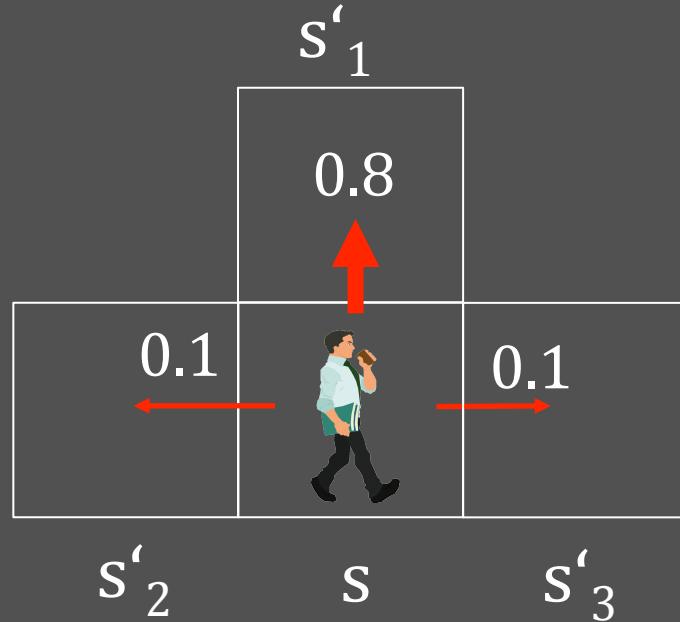
$$r = +1$$

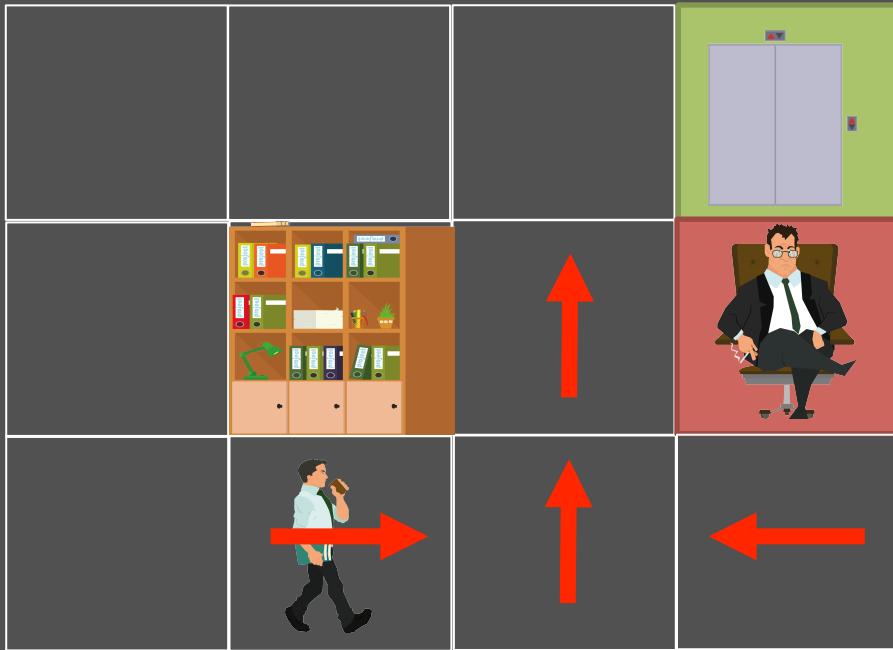


$$r = -1$$

Plan: sequence of actions

NONDETERMINISTIC ACTION RULE

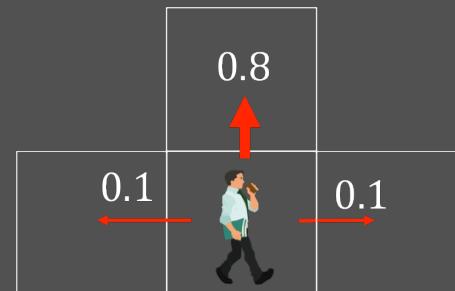




$r = + 1$

$r = - 1$

Action sequence:



POLICIES

We want a plan! But this is not a deterministic world!

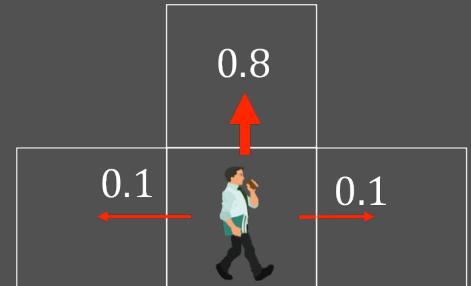
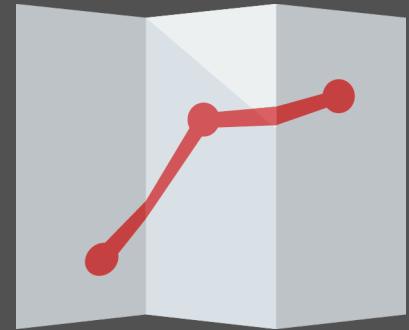
Plan: mapping from states to actions

Policy π : states \rightarrow actions

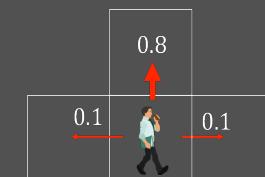
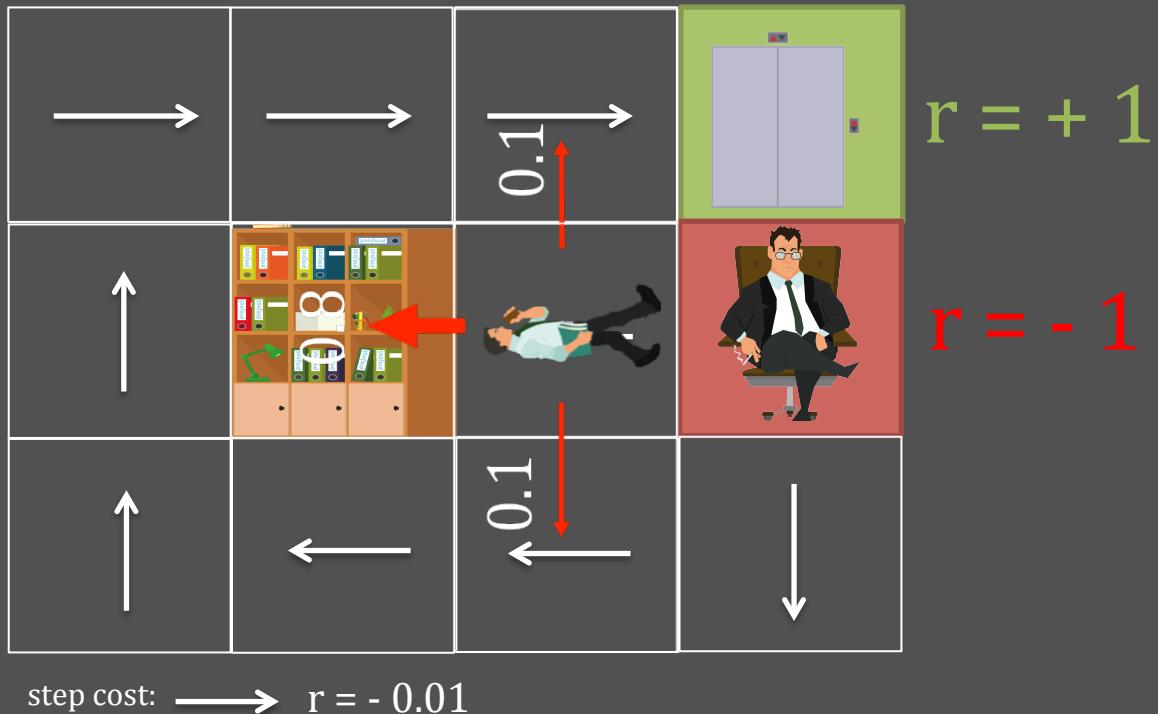
- it's like an if-then-plan
- look-up table

Optimal Policy π^* : states \rightarrow actions

- maximized expected value



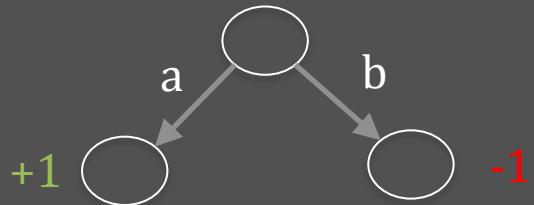
POLICIES



VALUES

Optimal policy: maximizes the expected value:

$[-1] < [+1]$

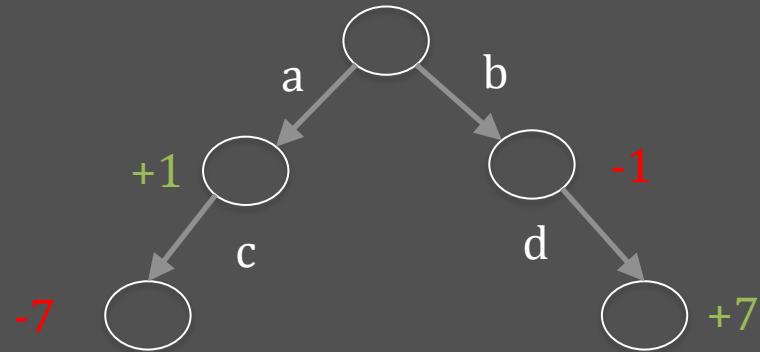


EXPECTEDVALUE

Optimal policy: maximizes the expected value

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

$$[-6] < [+6]$$

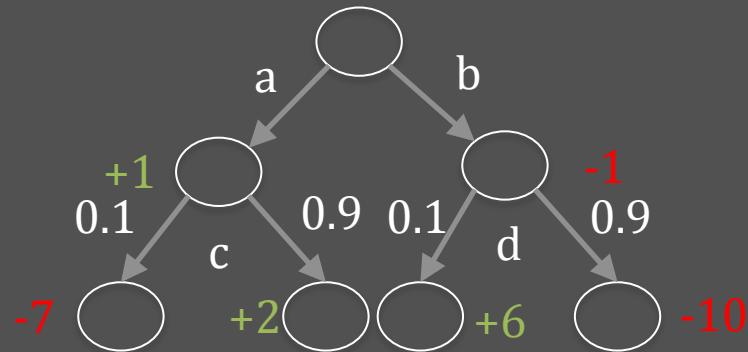


Value depends on all successor states !!!

EXPECTED VALUE

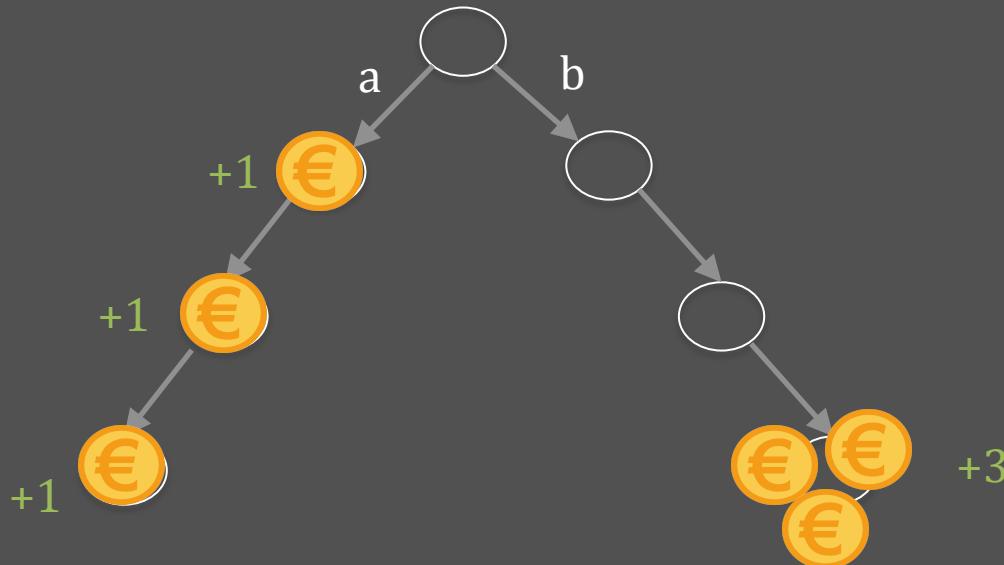
$$[+1, -7*0.1 + 0.9*2] < [-1, +6*0.1 - 10*0.9]$$

$$[+2.1] < [-9.8]$$

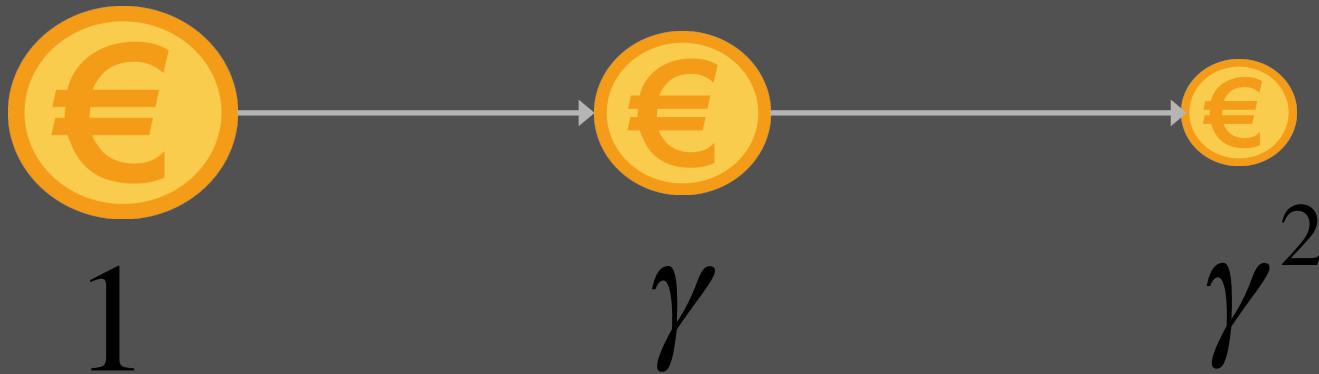


EXPECTED VALUE

[+3] = [+3]



DISCOUNTING

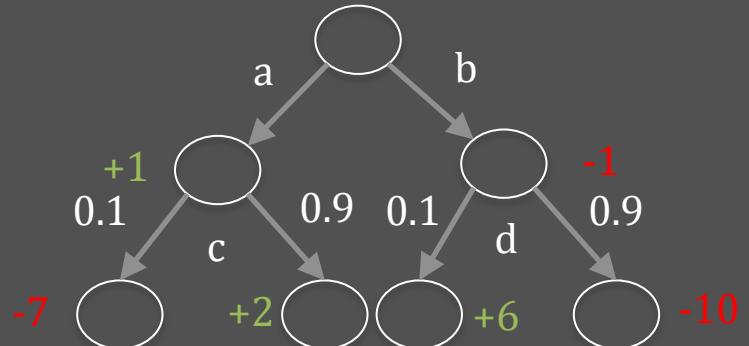


$$0 < \gamma < 1$$

$$V([r_0, r_1, r_2, \dots, r_n]) = r_0 + \gamma r_1 + \gamma^2 r_2 + \dots + \gamma^n r_n$$

$$V^*(s) = \max_a \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V^*(s')]$$

BELLMAN EQUATION



„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

How to act optimal?

Step 1: Take the correct first action

Step 2: Keep being optimal

SUM

MDP: Non-deterministic search problem
Uncertainty about performing actions
Discounting



→ POMDP: Uncertainty about states

THANKYOU