(PARTIALLY OBSERVABLE) MARKOV DECISION PROCESSES

DR FREDERIKE PETZSCHNER

DR LIONEL RIGOUX

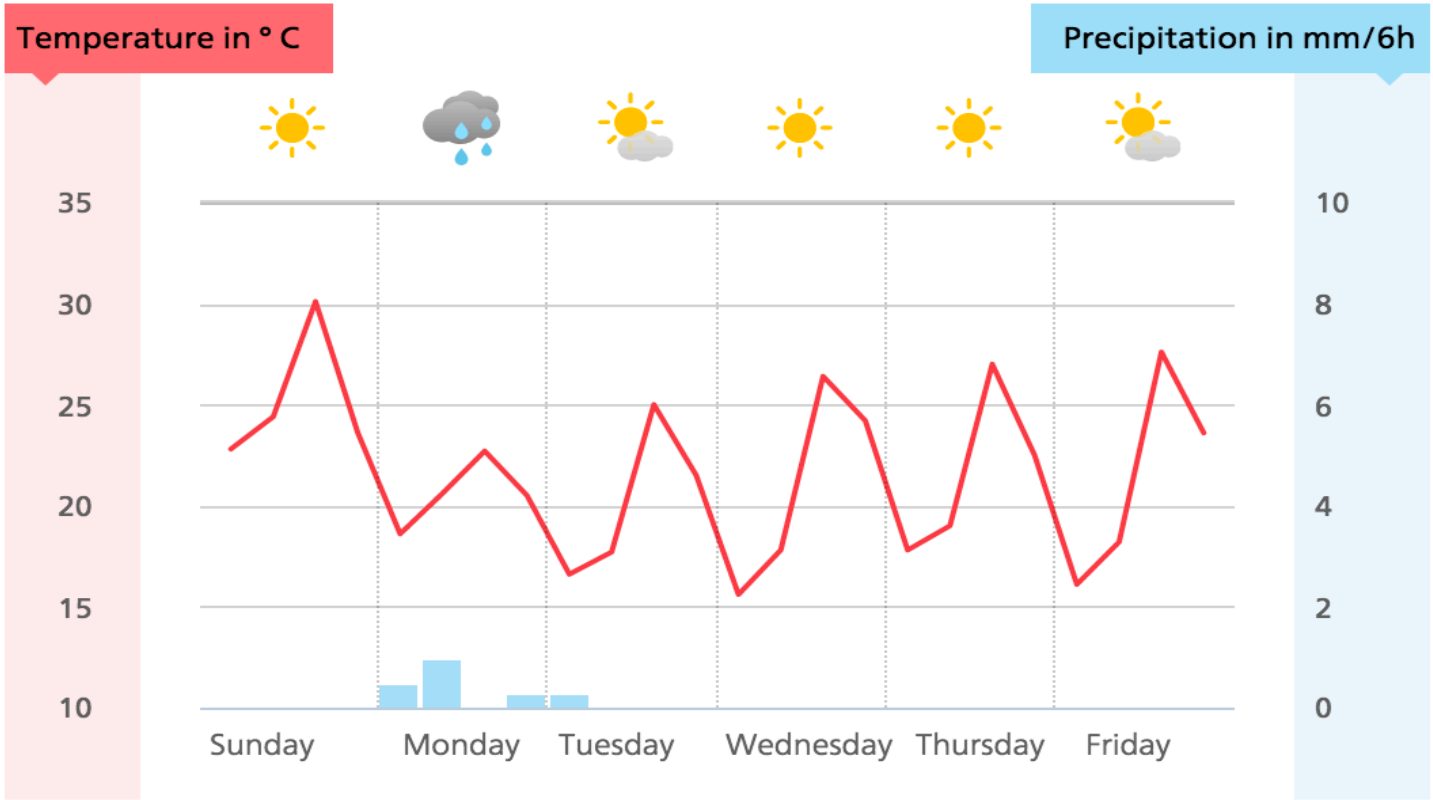
30082016

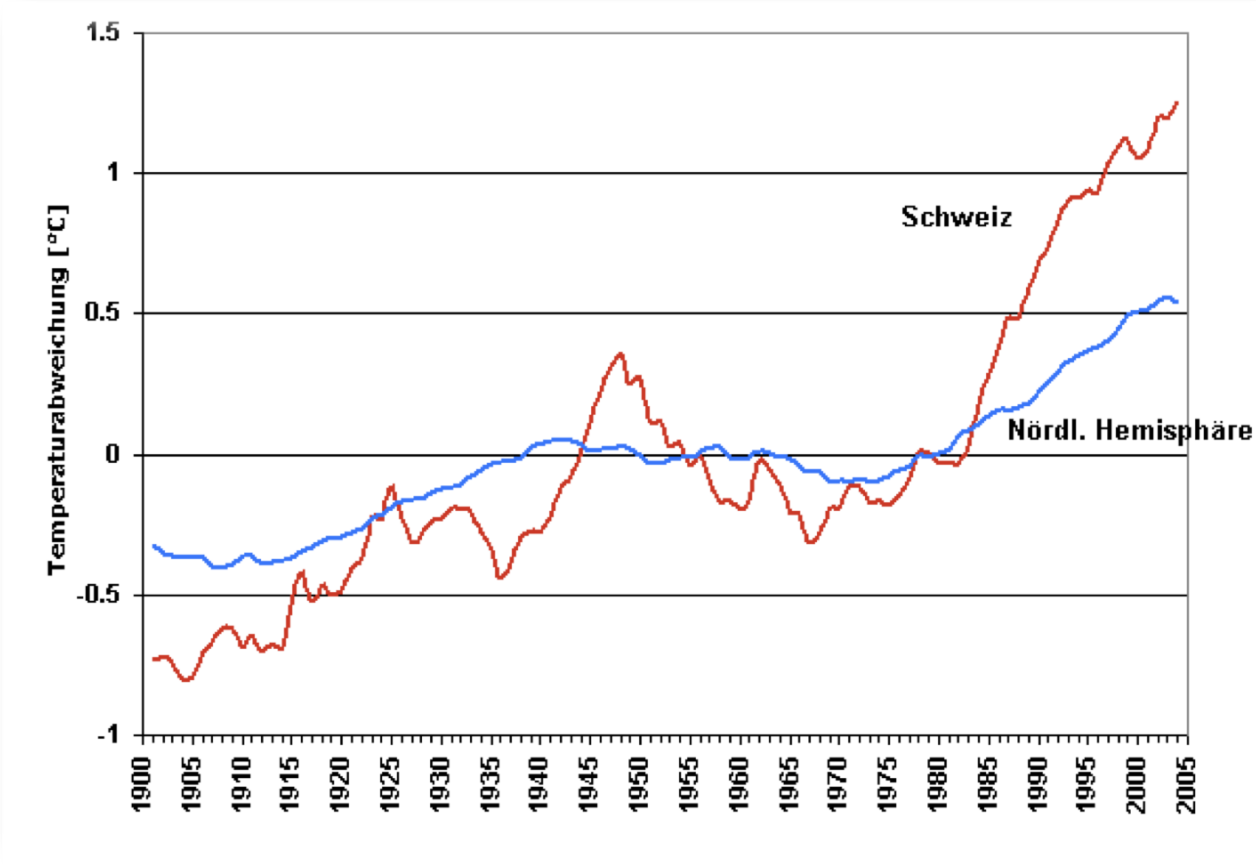
Local forecasts **Zürich**

28 August to 2 September 2016

Zip code or city

Choose location




Temperaturabweichung [°C]
1.5
1
0.5
0
-0.5
-1
1900
1905
1910
1915
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
Schweiz
Nördl. Hemisphäre

Past

Present

Future

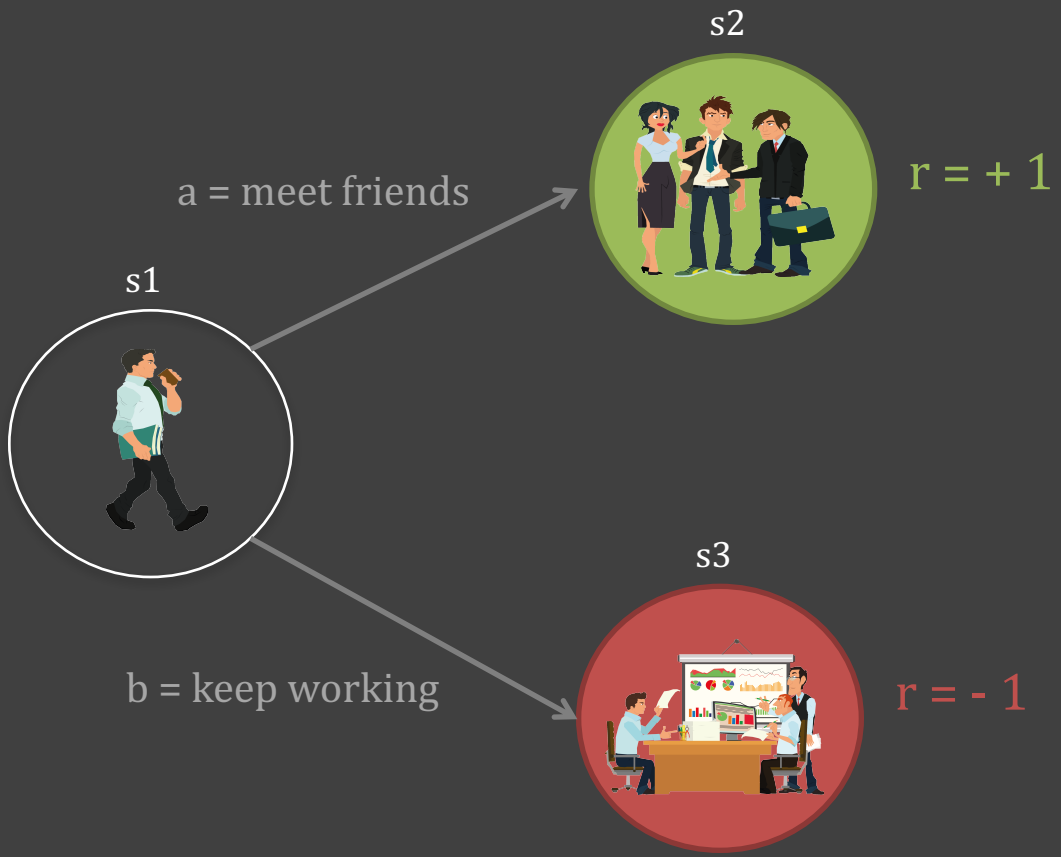
Countries will aim to keep global temperatures from rising more than 2° C (3.6° F) by 2100 with an ideal target of keeping temperature rise below 1.5° C (2.7° F).

PLANNINGPROBLEM

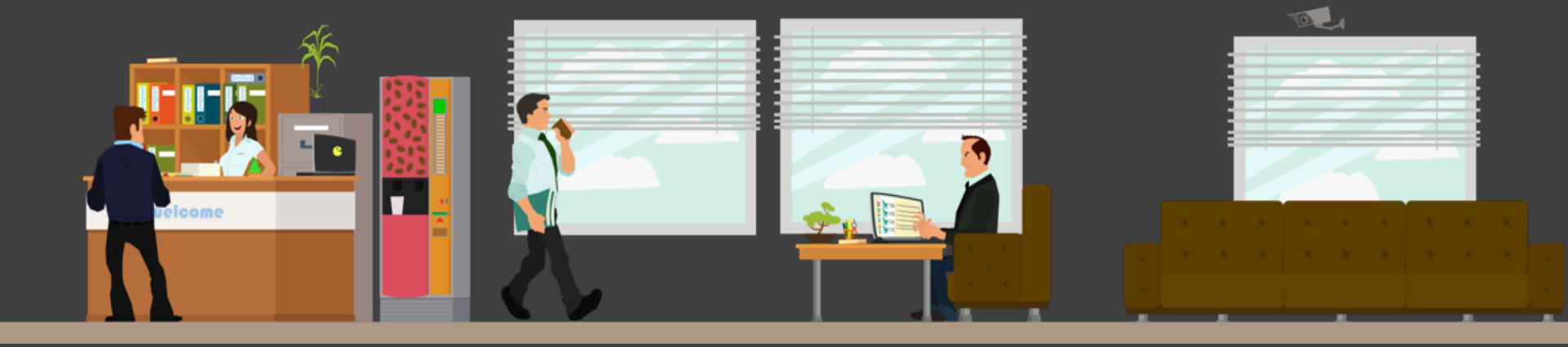
Given a complete and correct model of the world dynamics and a reward structure, what is the optimal way to behave?

Littman & Cassandra

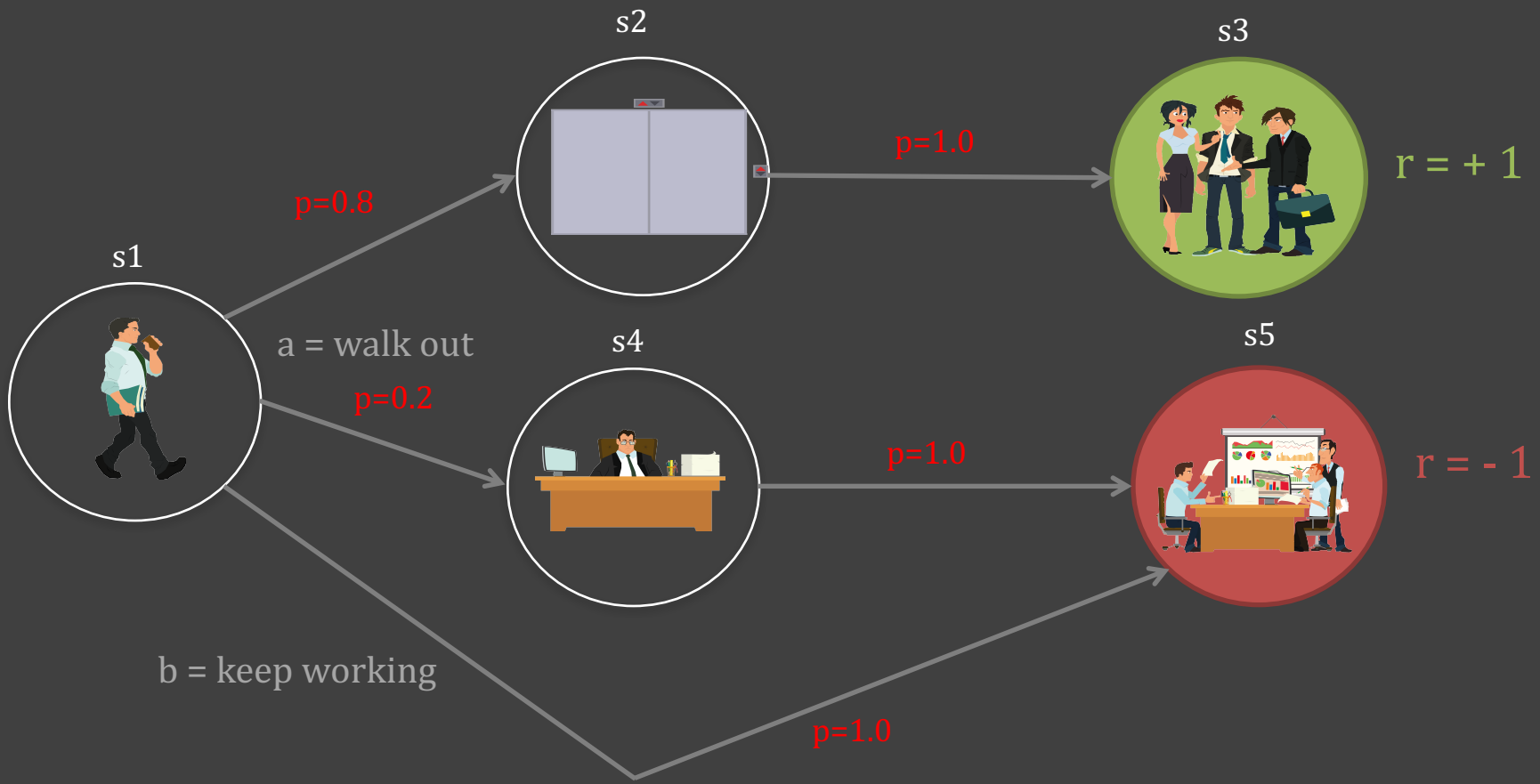
DETERMINISTICWORLD



NONDETERMINISTICACTIONS



NONDETERMINISTICWORLD



The MDP is defined by:

States s → S (state space)

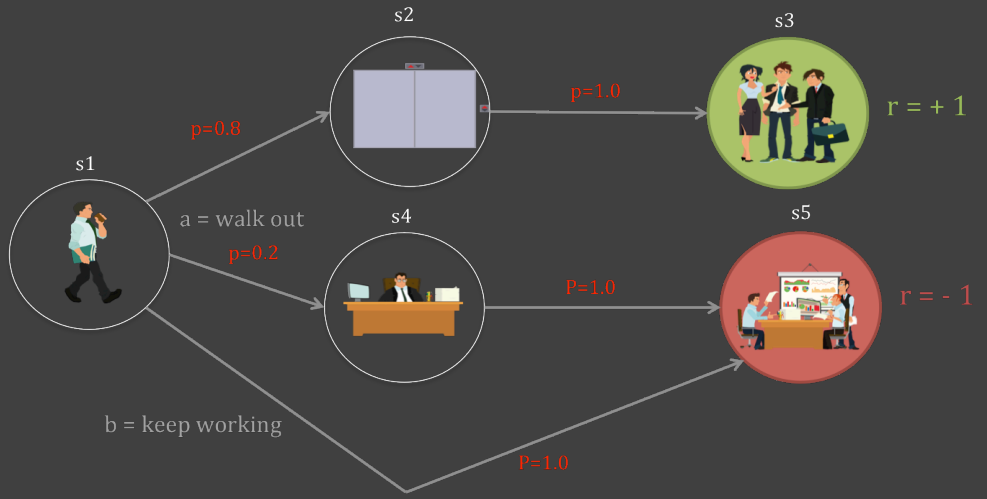
Actions a → A (action space)

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
s1	a	s4	0.2
s1	a	s2	0.8
s1	b	s5	1

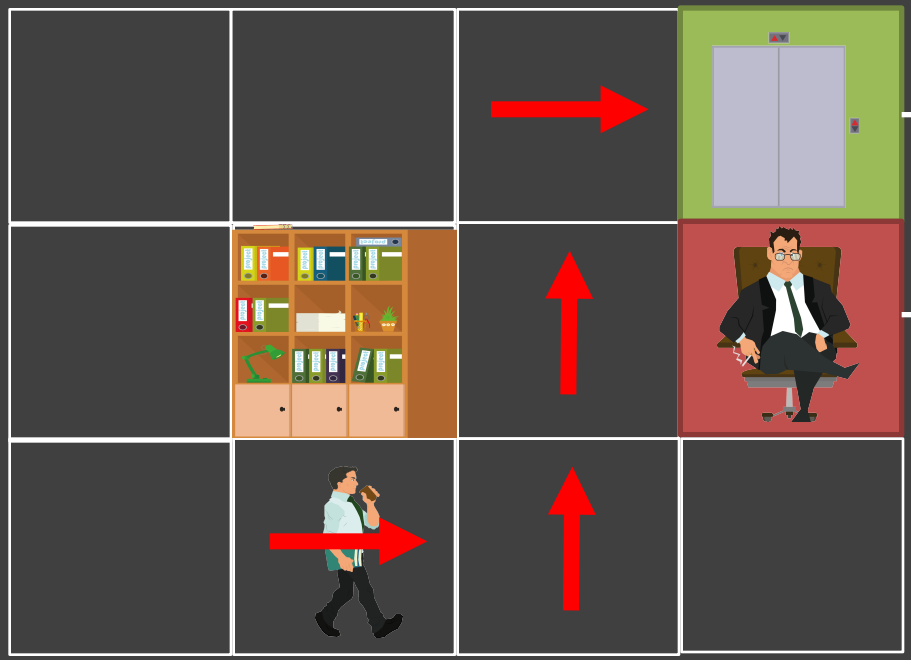
WHATSMARKOVABOUTANMDP?

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

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

Not every process is an MDP!

What's the best way out?



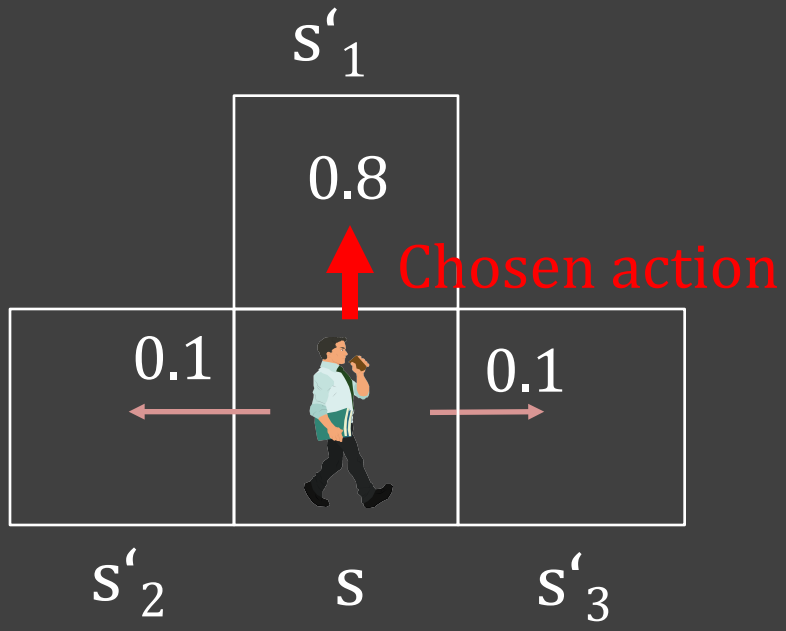
r = + 1

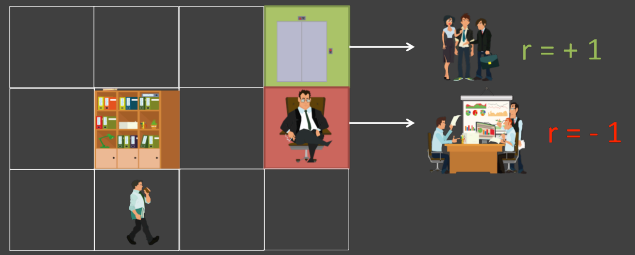
r = - 1

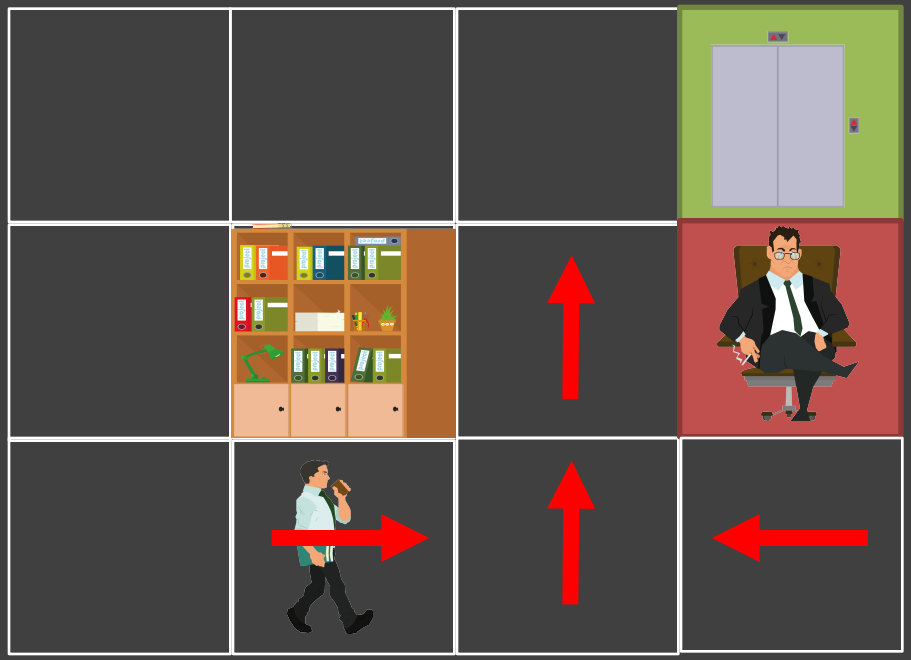
Plan: sequence of actions

step cost: r = - 0.01

NONDETERMINISTICACTIONRULE





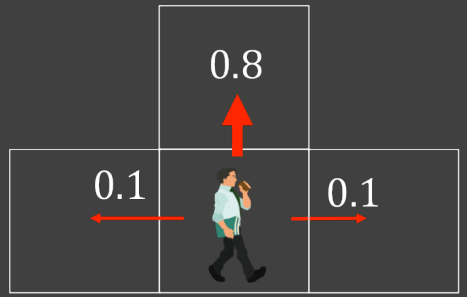


r = + 1

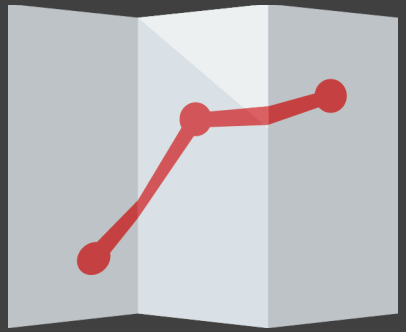
r = - 1

Action sequence:

step cost: → r = - 0.01



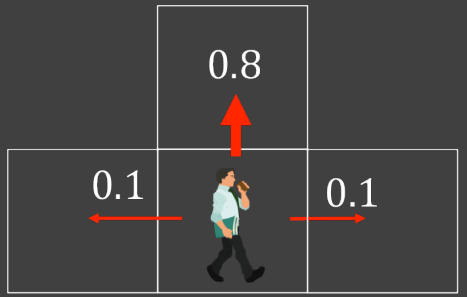
POLICIES



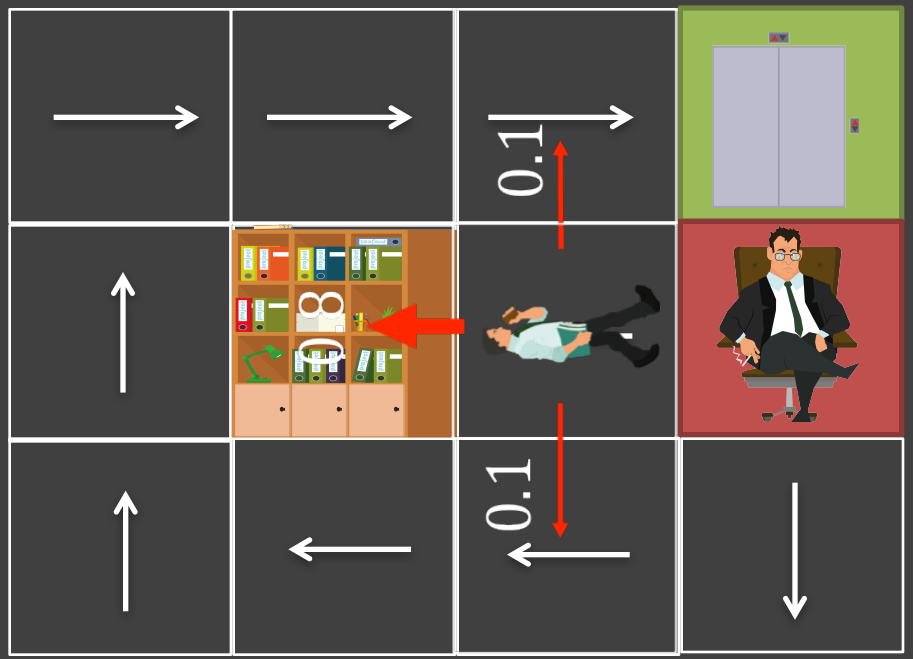
We want a plan! But this is not a deterministic world!
Plan: mapping from states to actions

Policy π: states → actions
- It's like an if-then-plan
- look-up table

Optimal Policy π*: states → actions
- maximized expected value



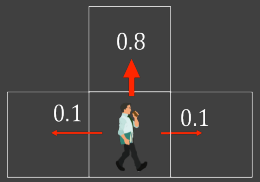
POLICIES



r = + 1

r = - 1

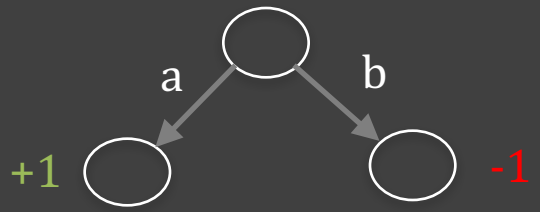
step cost: r = - 0.01



VALUES

Optimal policy: maximizes the expected value:

[-1] < [+1]

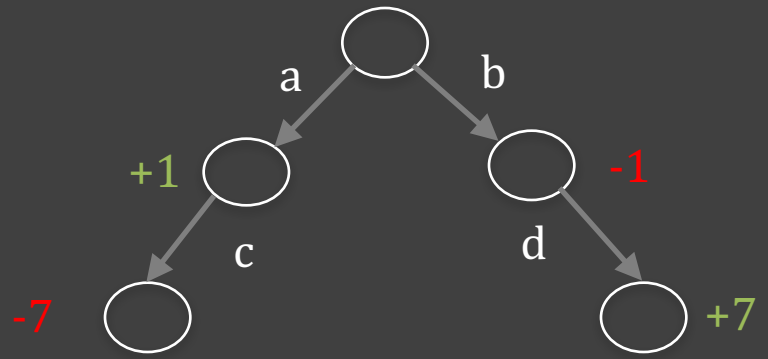


EXPECTEDVALUE

Optimal policy: maximizes the expected value

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

[-6] < [+6]

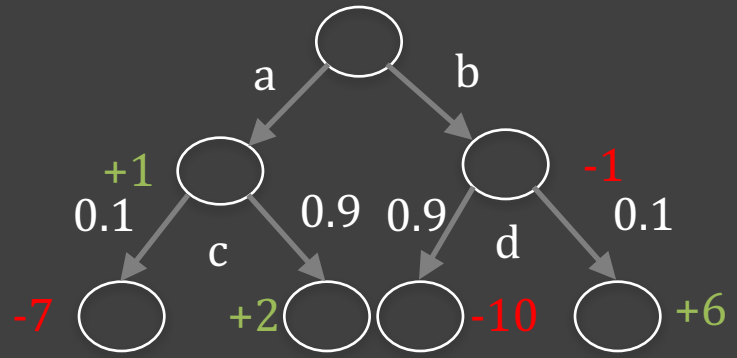


Value depends on all successor states !!!

EXPECTEDVALUE

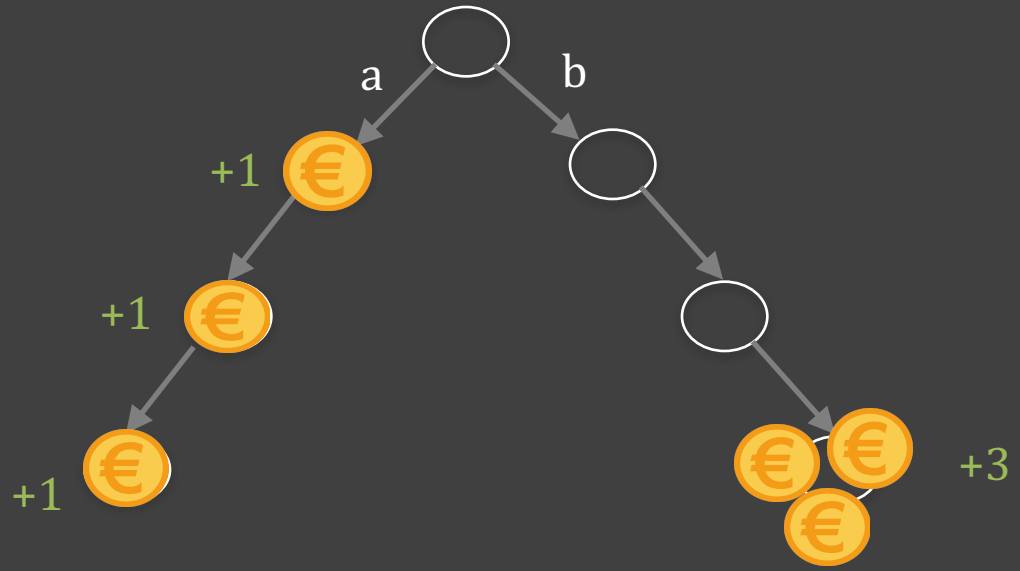
[+1,-7*0.1 + 0.9*2] < [-1,+6*0.1 -10*0.9]

[+2.1] < [-9.8]

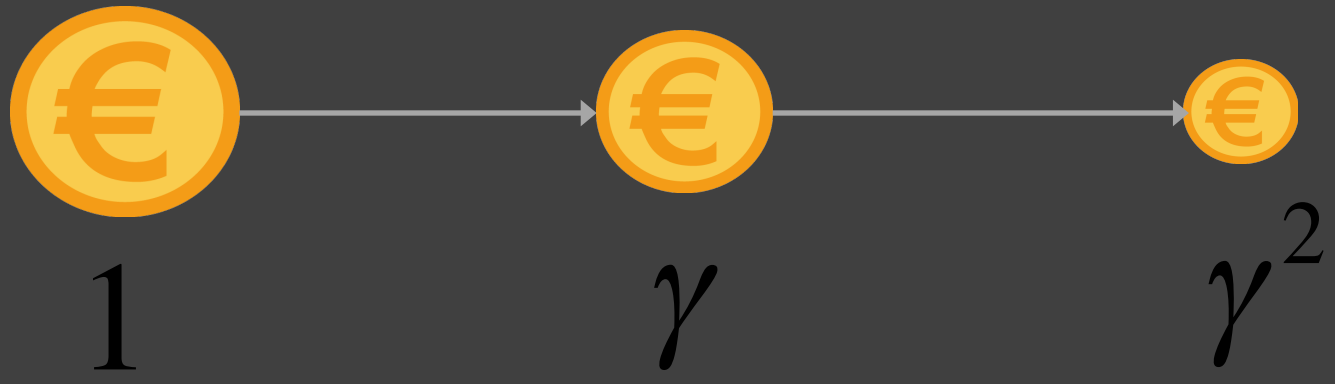


EXPECTEDVALUE

[+3] = [+3]



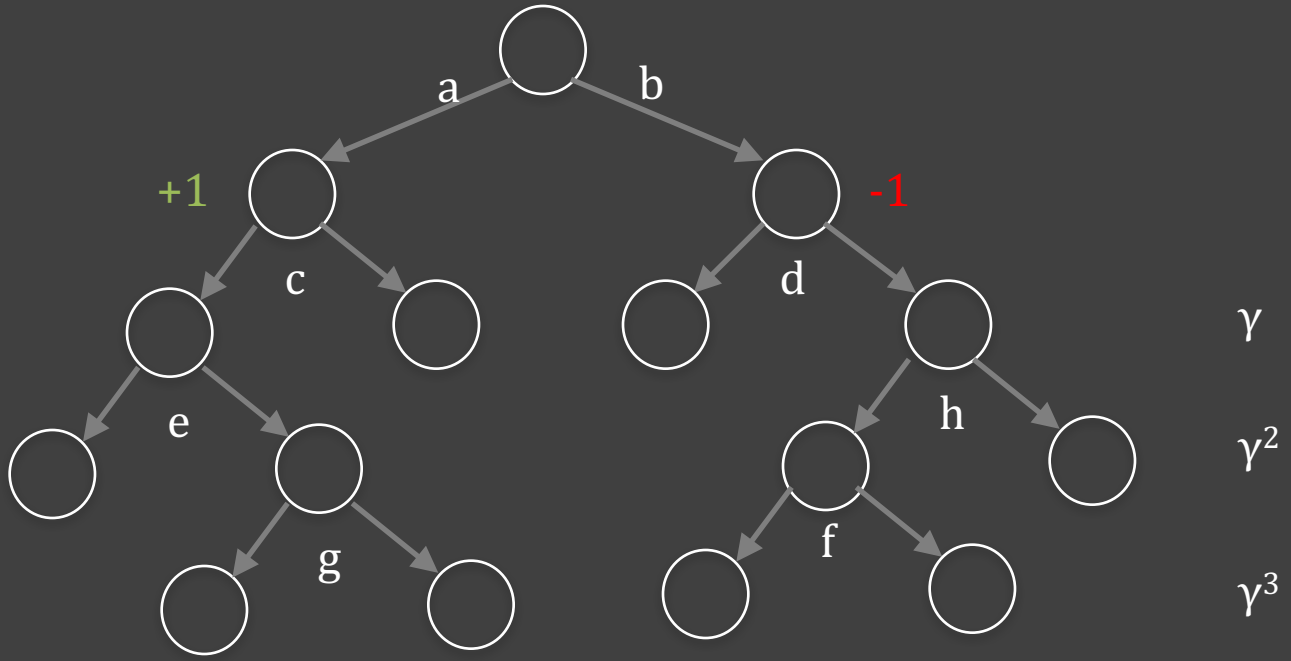
DISCOUNTING



$0 < \gamma < 1$

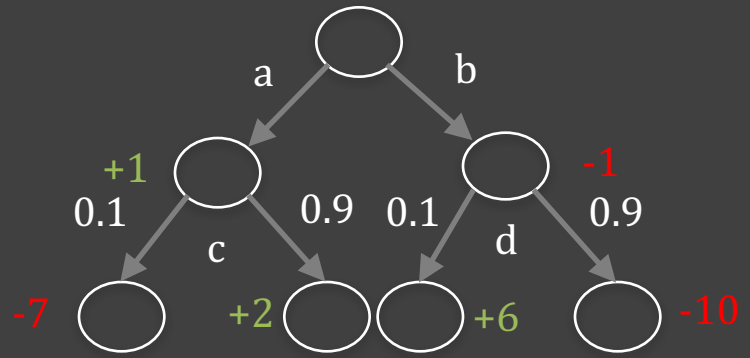
$$V([r_0, r_1, r_2, \ldots r_n]) = r_0 + \gamma r_1 + \gamma^2 r_2 + \ldots \gamma^n r_n$$

DISCOUNTING



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

BELLMANEQUATION



„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