(PARTIALLY OBSERVABLE) MARKOV DECISION PROCESSES

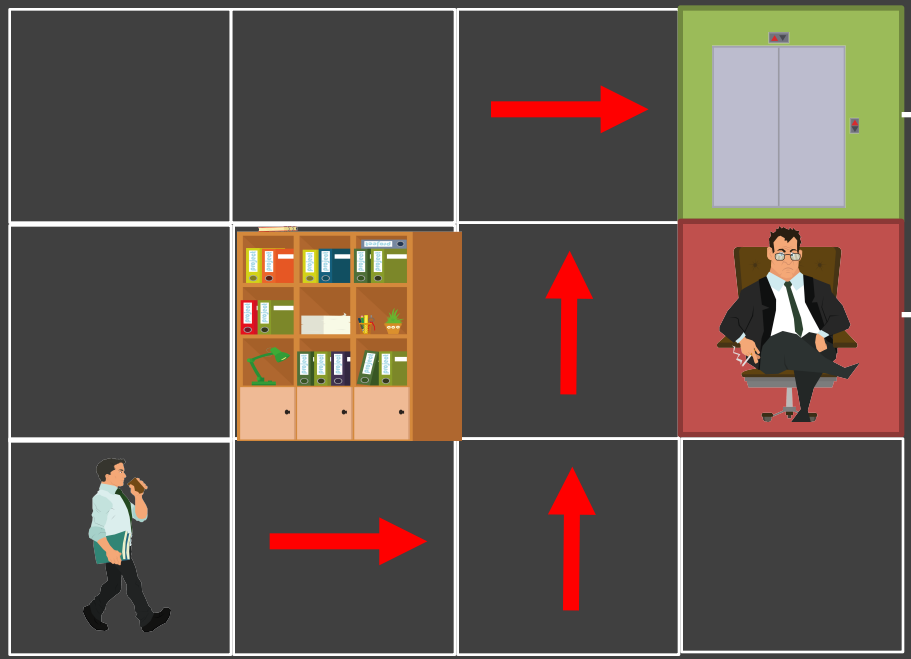
DR FREDERIKE PETZSCHNER

DR LIONEL RIGOUX

29 08 2017


elcome

What's the fastest way out?



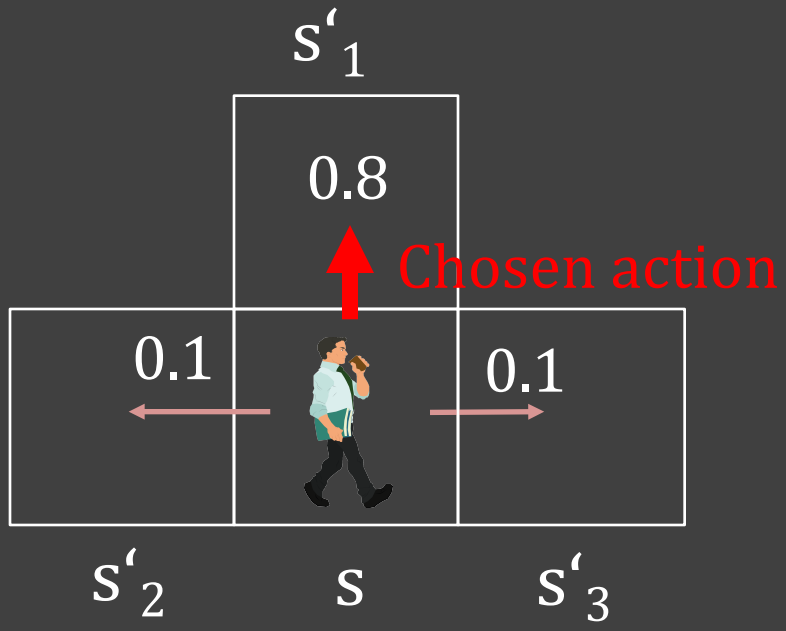
r = + 1

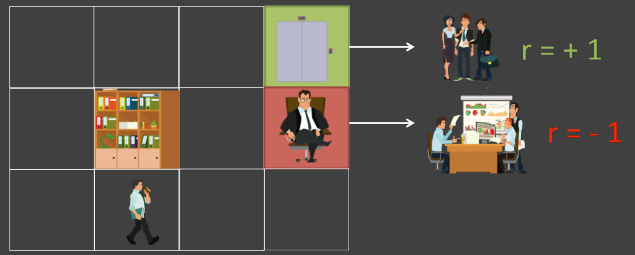
r = - 1

Plan: sequence of actions

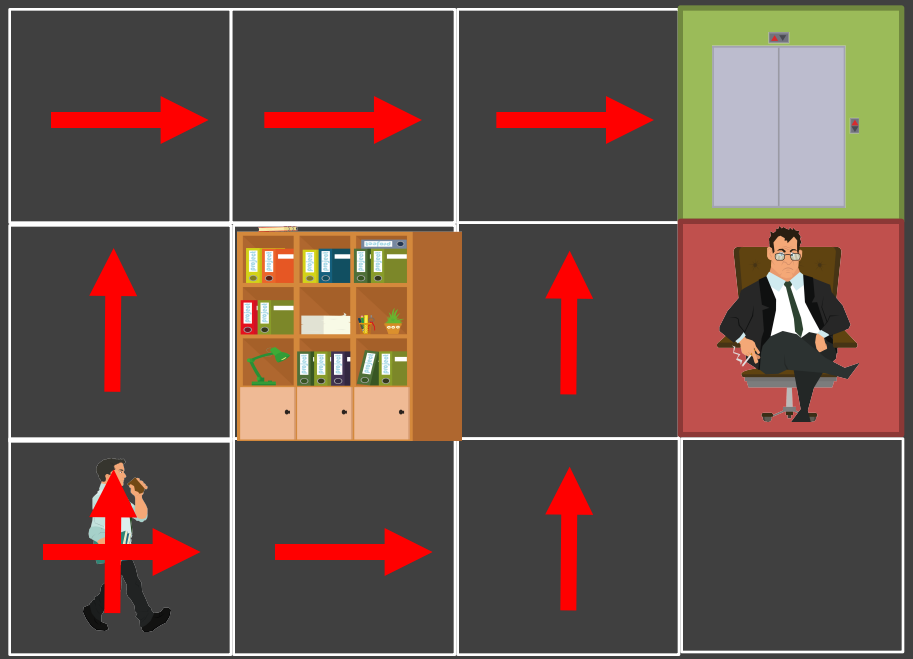
r_{step} = -0.04

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.

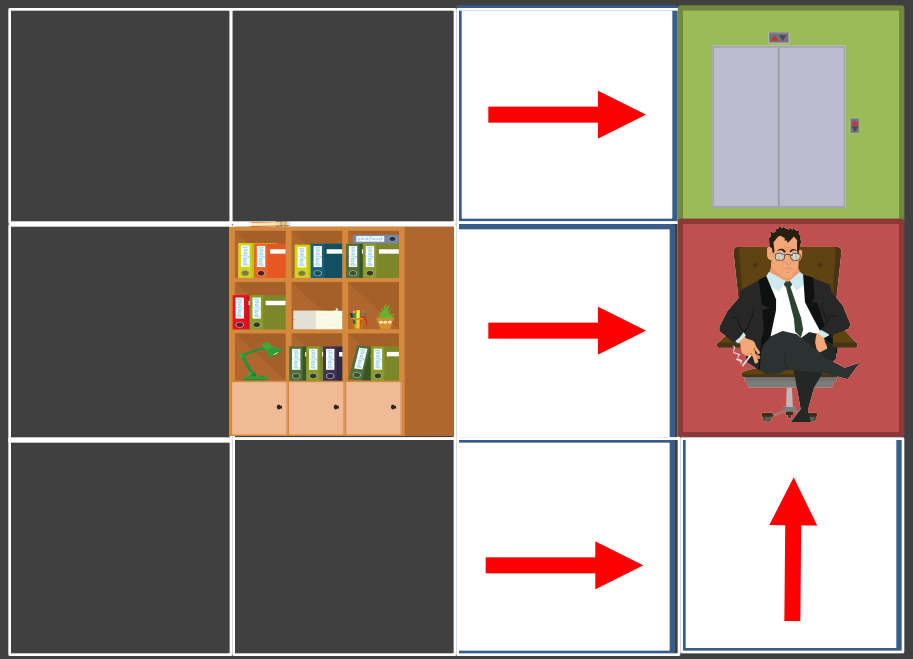
Whats Markovian about an MDP?

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

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

Not every process is an MDP!

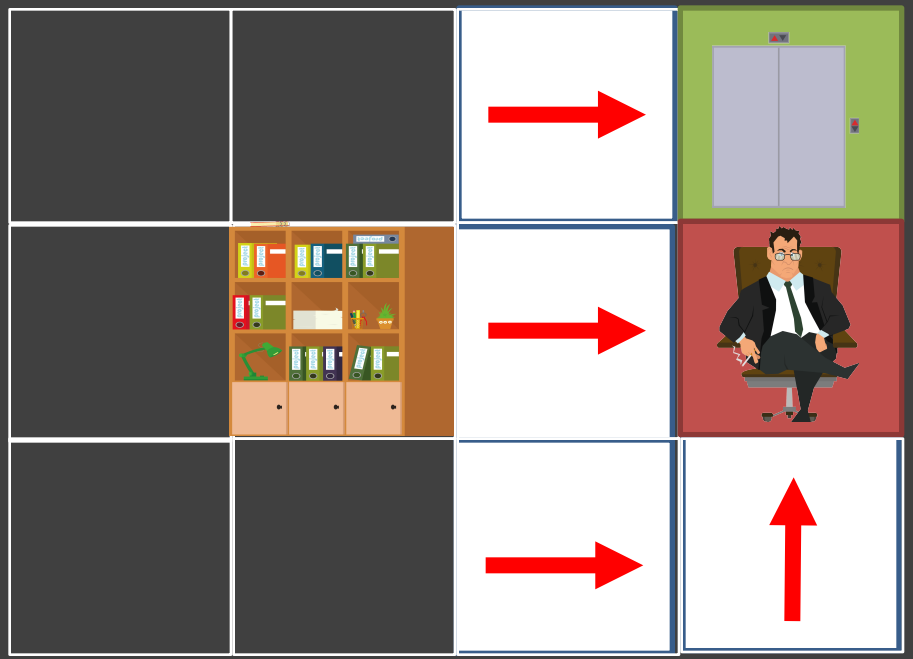
What if the reward structure of the world changes.



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

$r_{step} = -2$

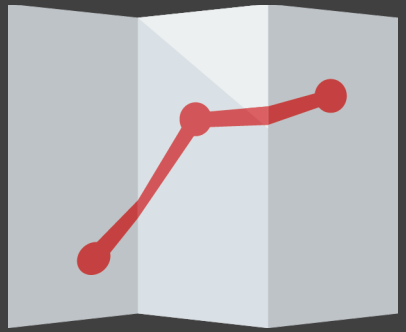
What if the reward structure of the world changes.



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

$r_{step} = -2$

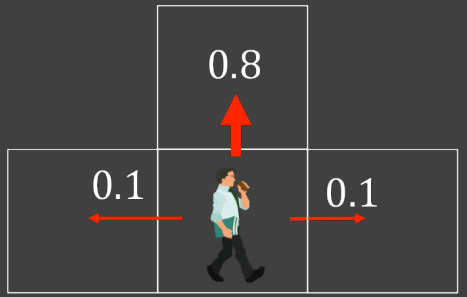
Policies



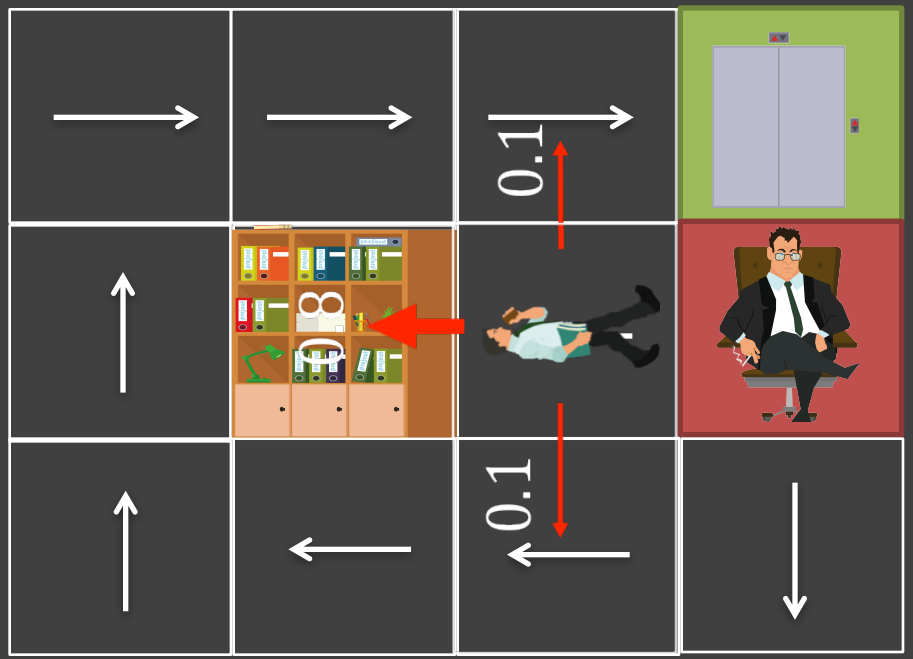
We want a plan! But this is not a deterministic world!
Plan: mapping from states to actions

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

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



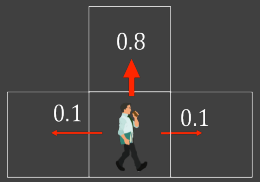
POLICIES



r = + 1

r = - 1

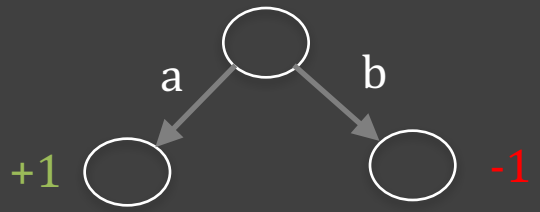
step cost: r = - 0.01



VALUES

Optimal policy: maximizes the expected value:

[-1] < [+1]

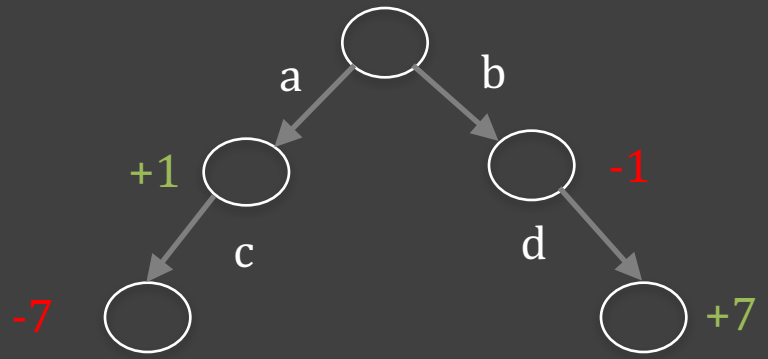


EXPECTEDVALUE

Optimal policy: maximizes the expected value

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

[-6] < [+6]

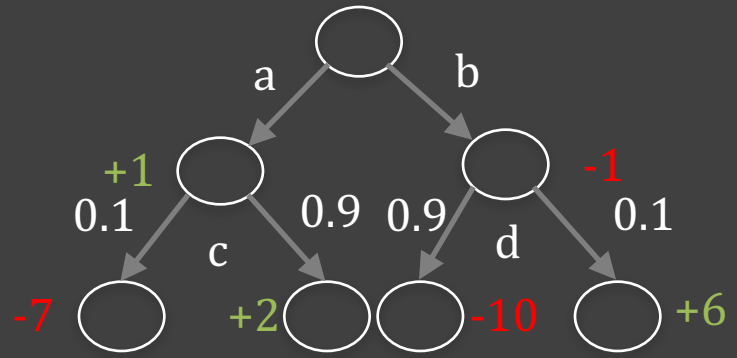


Value depends on all successor states !!!

EXPECTEDVALUE

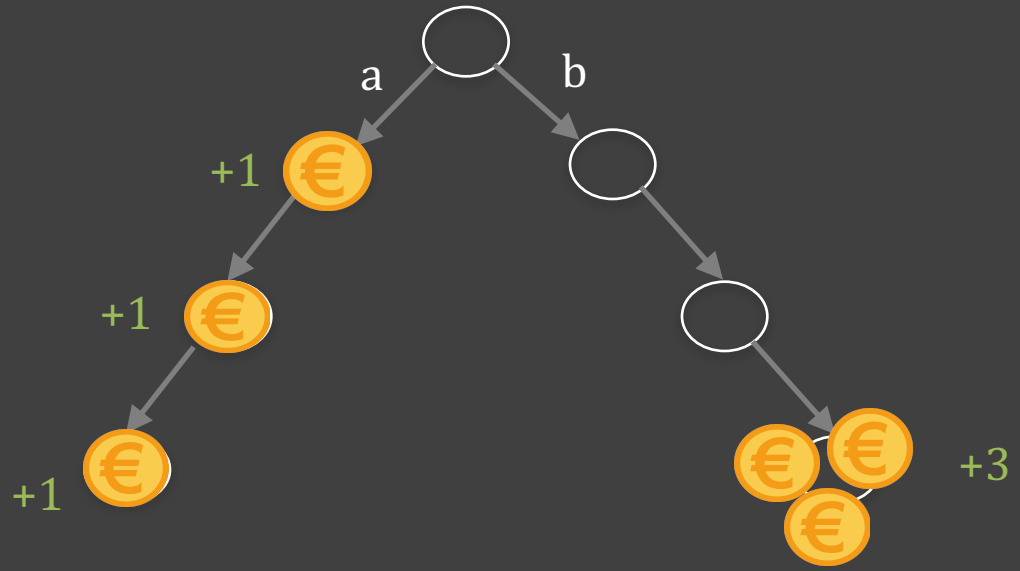
[+1,-7*0.1 + 0.9*2] < [-1,+6*0.1 -10*0.9]

[+2.1] < [-9.8]

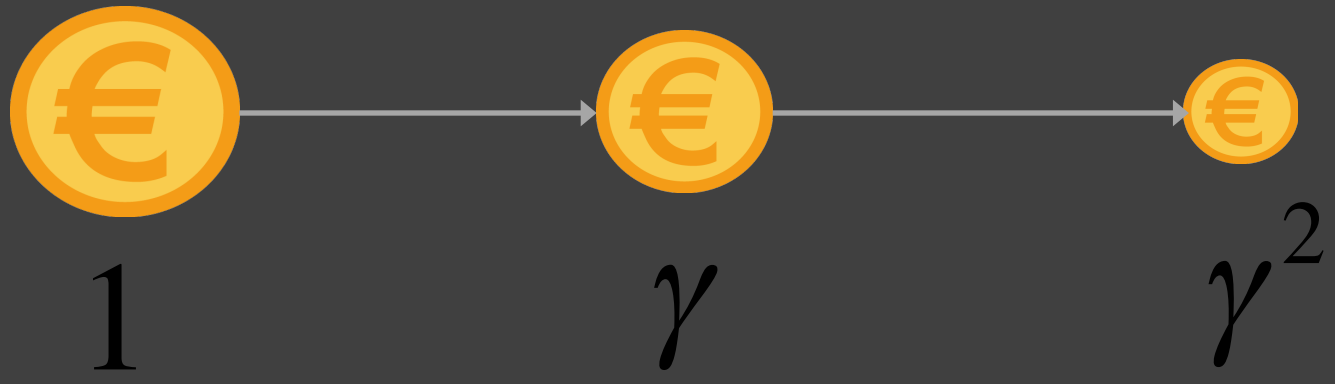


EXPECTEDVALUE

[+3] = [+3]



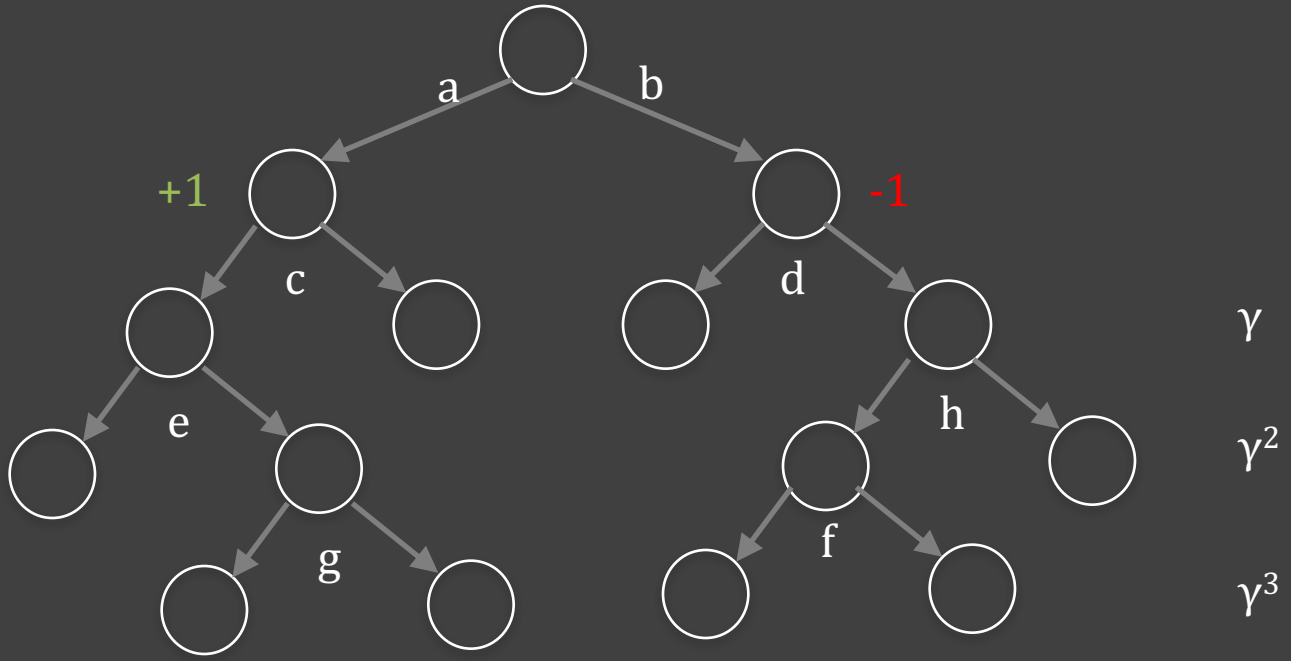
DISCOUNTING



$0 < \gamma < 1$

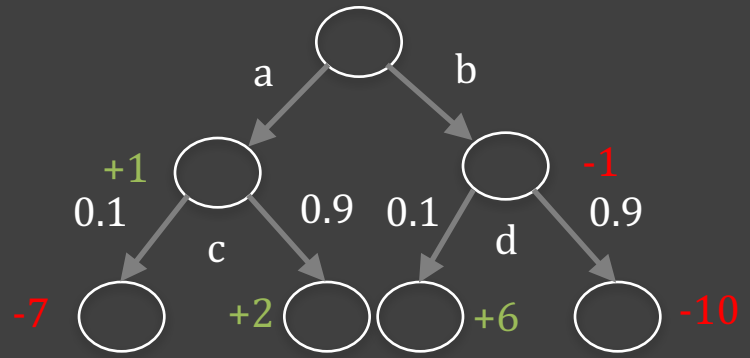
$$V([r_0, r_1, r_2, \ldots r_n]) = r_0 + \gamma r_1 + \gamma^2 r_2 + \ldots \gamma^n r_n$$

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