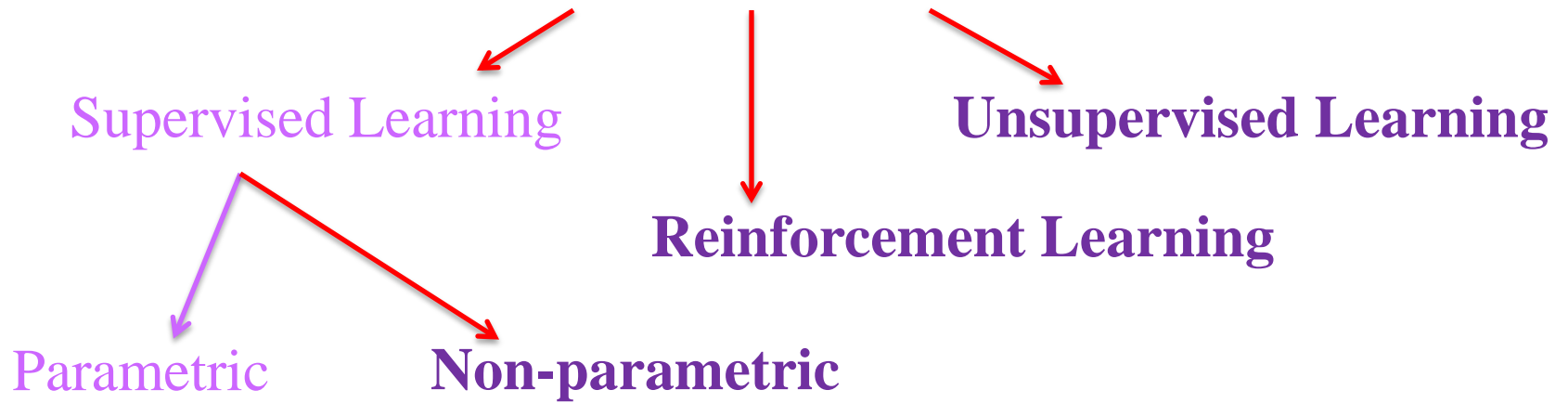


# Markov Decision Processes

Mausam

# Machine Learning



# Outline

Fri  
Clustering

K-means & Agglomerative

Mon

Expectation Maximization (EM)

Wed

Principle Component Analysis (PCA)

Fri

(MDPs)

**Markov Decision Processes**

Mon

**Reinforcement Learning (RL)**

Wed

Instance-Based Learning

Fri

SVMs & Summary

# Planning Agent

Static vs. Dynamic



Fully  
vs.  
Partially  
Observable

Deterministic  
vs.  
Stochastic

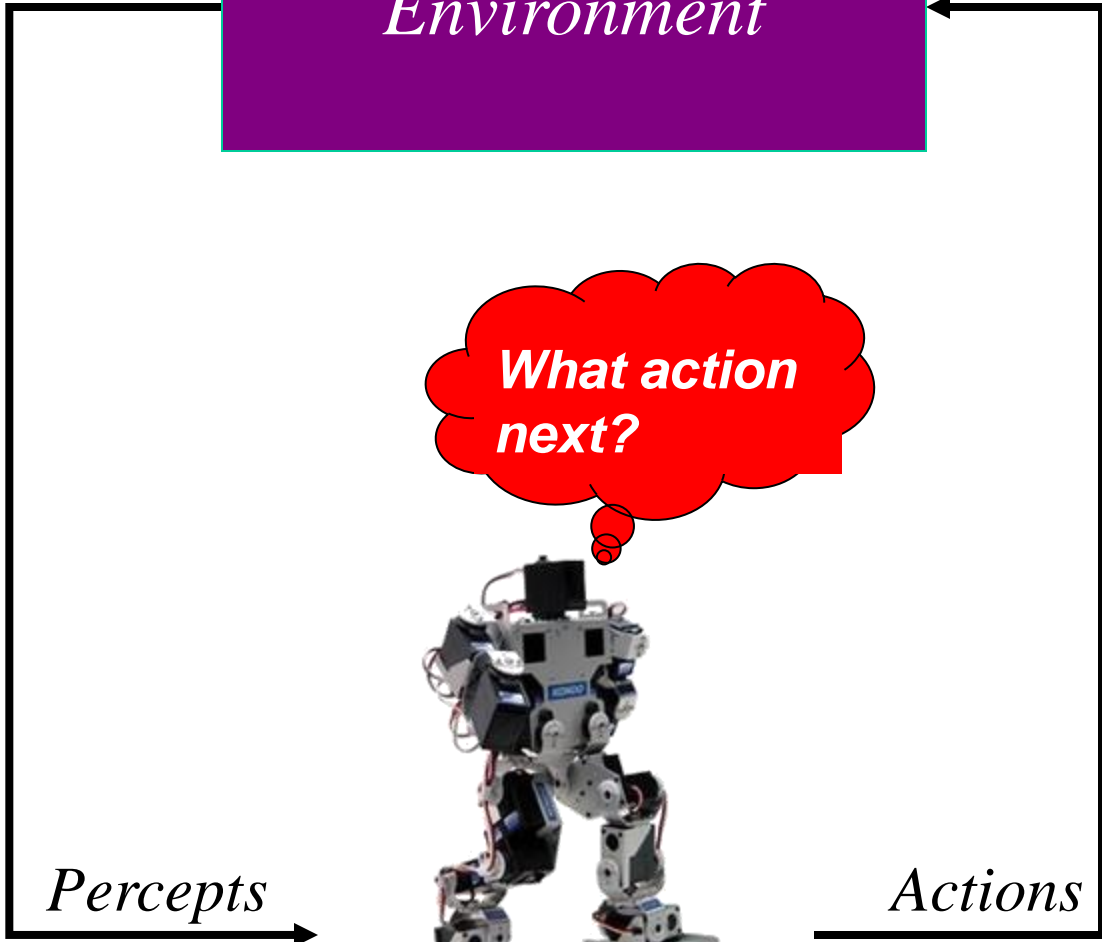
Perfect  
vs.  
Noisy

Instantaneous  
vs.  
Durative

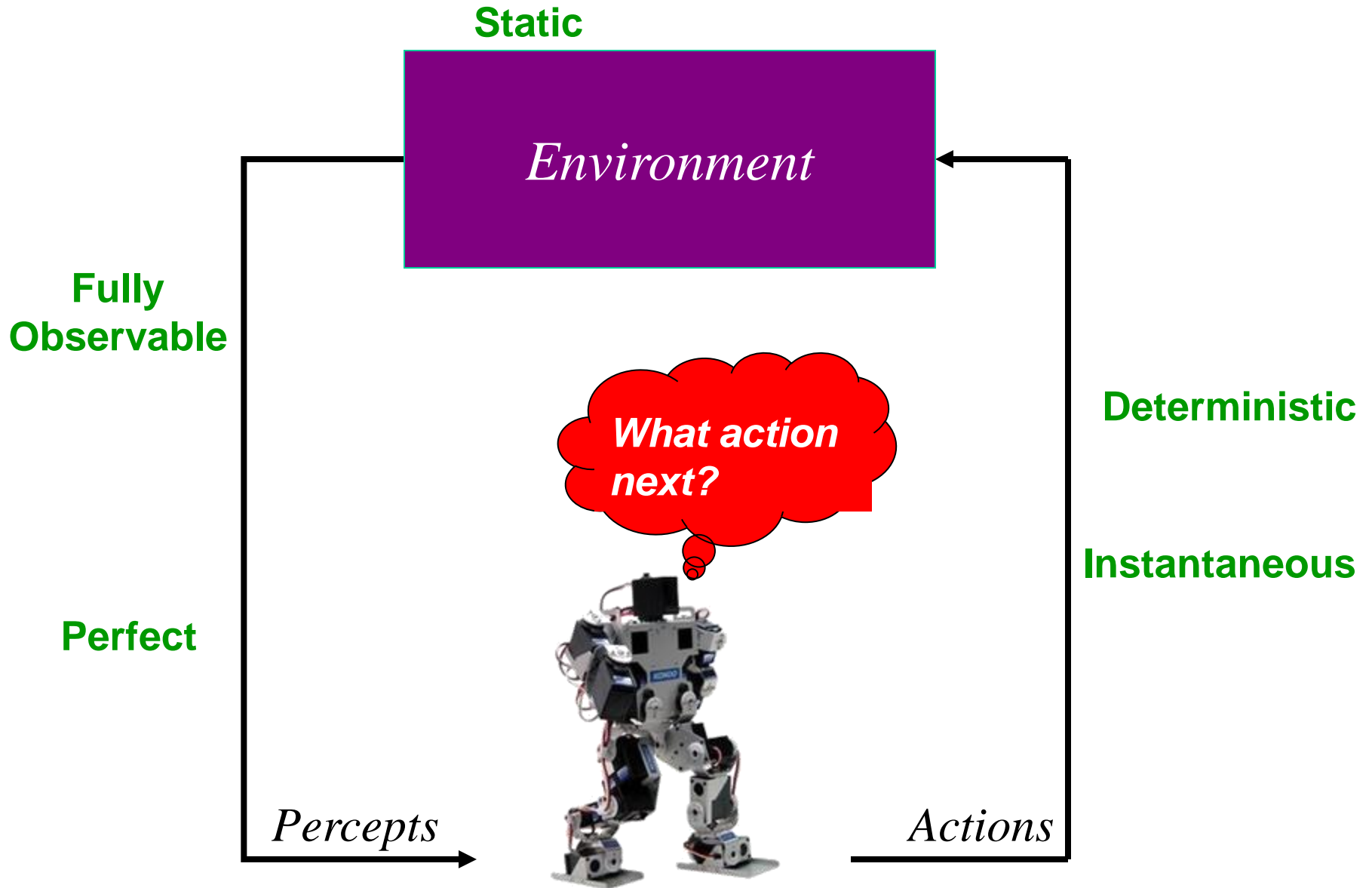


*Percepts*

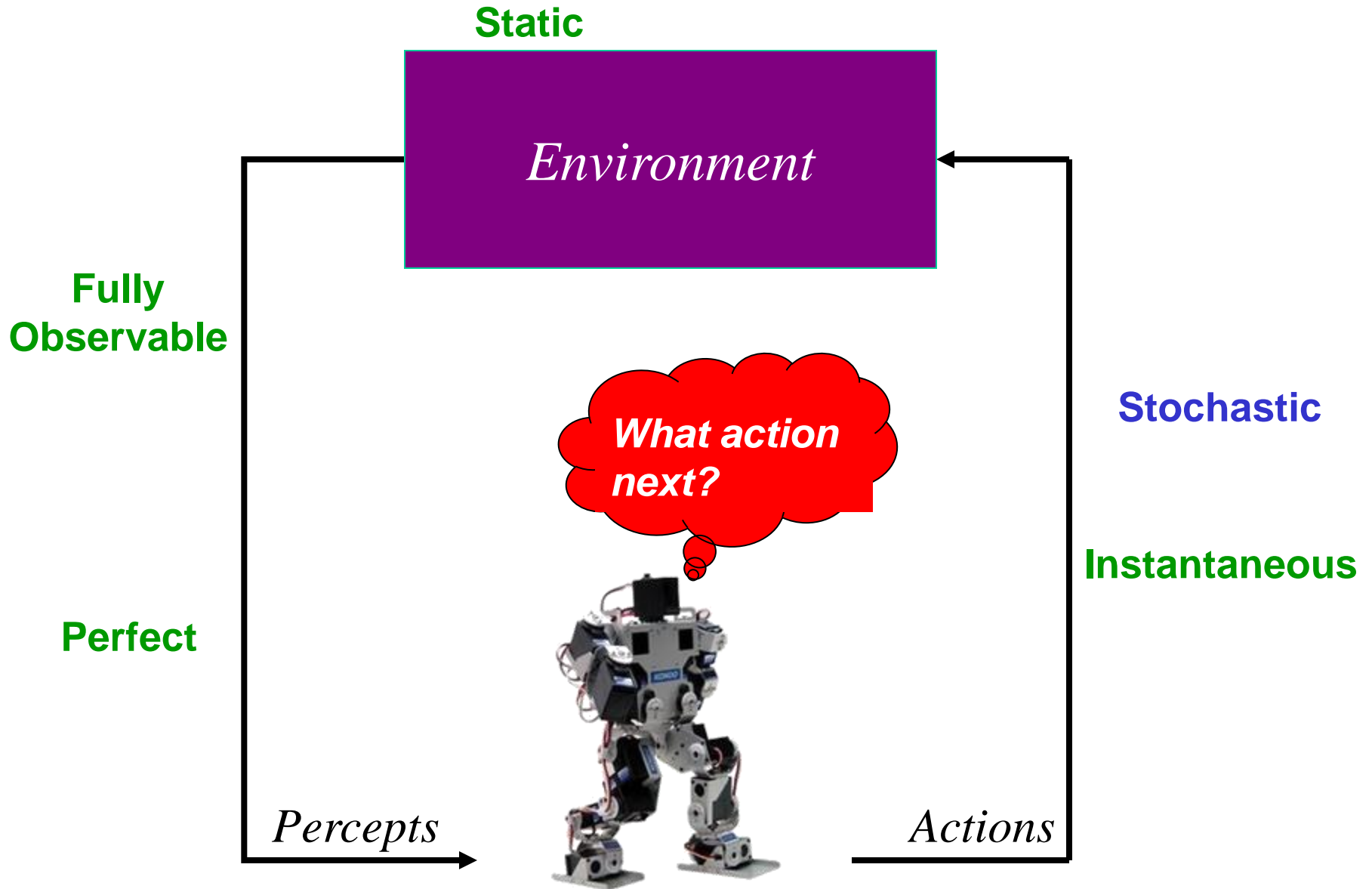
*Actions*



# Classical Planning



# Stochastic Planning: MDPs



# Markov Decision Process (MDP)

- $\mathcal{S}$ : A set of states
- $\mathcal{A}$ : A set of actions
- $\Pr(s'|s,a)$ : transition model
- $\mathcal{C}(s,a,s')$ : cost model
- $\mathcal{G}$ : set of goals
- $s_0$ : start state
- $\gamma$ : discount factor
- $\mathcal{R}(s,a,s')$ : reward model

# Objective of an MDP

- Find a policy  $\pi: \mathcal{S} \rightarrow \mathcal{A}$
- which optimizes
  - minimizes  $\left( \begin{array}{c} \text{discounted} \\ \text{or} \\ \text{undiscount.} \end{array} \right)$  expected cost to reach a goal
  - maximizes  $\left( \begin{array}{c} \text{discounted} \\ \text{or} \\ \text{undiscount.} \end{array} \right)$  expected reward
  - maximizes  $\left( \begin{array}{c} \text{discounted} \\ \text{or} \\ \text{undiscount.} \end{array} \right)$  expected (reward-cost)
- given a \_\_\_\_\_ horizon
  - finite
  - infinite
  - indefinite
- assuming full observability



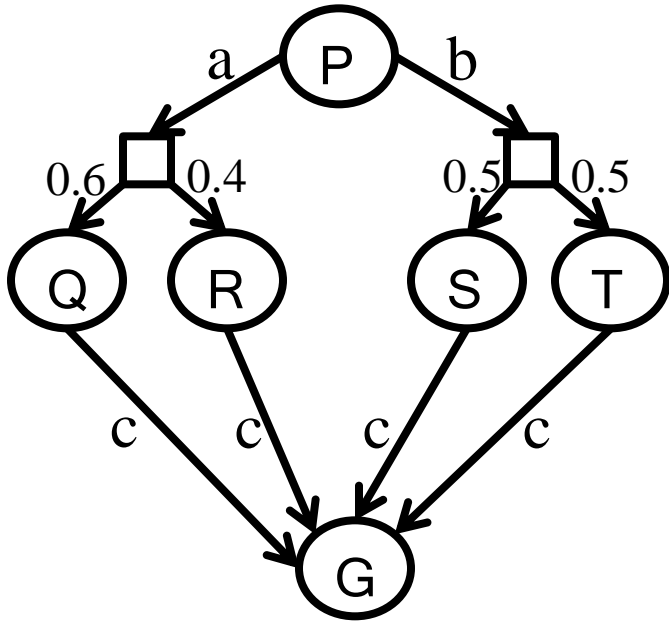
## Role of Discount Factor ( $\gamma$ )

- Keep the total reward/total cost finite
  - useful for infinite horizon problems
- Intuition (economics):
  - Money today is worth more than money tomorrow.
- Total reward:  $r_1 + \gamma r_2 + \gamma^2 r_3 + \dots$
- Total cost:  $c_1 + \gamma c_2 + \gamma^2 c_3 + \dots$

# Examples of MDPs

- Goal-directed, Indefinite Horizon, Cost Minimization MDP
  - $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{C}, \mathcal{G}, s_0 \rangle$
  - Most often studied in planning, graph theory communities
- Infinite Horizon, Discounted Reward Maximization MDP **most popular**
  - $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$
  - Most often studied in machine learning, economics, operations research communities
- Oversubscription Planning: Non absorbing goals, Reward Max. MDP
  - $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{G}, \mathcal{R}, s_0 \rangle$
  - Relatively recent model

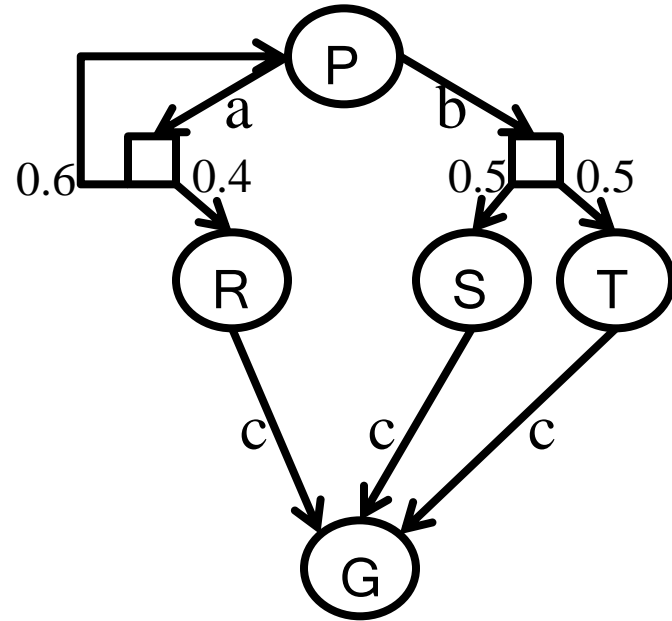
# AND/OR Acyclic Graphs vs. MDPs



$$C(a) = 5, C(b) = 10, C(c) = 1$$

$$V(Q/R/S/T) = 1$$

- $V(P) = 6$  – action a



- infinite loop
- $V(R/S/T) = 1$
- $Q(P,b) = 11$
- $Q(P,a) = \text{????}$
- suppose I decide to take a in P
- $Q(P,a) = 5 + 0.4 * 1 + 0.6 Q(P,a)$
- $\rightarrow = 13.5$

# Bellman Equations for MDP<sub>1</sub>

- $\langle \mathcal{S}, \mathcal{A}, \mathcal{Pr}, \mathcal{C}, \mathcal{G}, s_0 \rangle$
- Define  $J^*(s)$  {optimal cost} as the minimum expected cost to reach a goal from this state.
- $J^*$  should satisfy the following equation:

$$J^*(s) = 0 \text{ if } s \in \mathcal{G}$$

$$J^*(s) = \min_{a \in \mathcal{A}_p(s)} \sum_{s' \in \mathcal{S}} \mathcal{Pr}(s'|s, a) [\mathcal{C}(s, a, s') + J^*(s')]$$

## Bellman Equations for MDP<sub>2</sub>

- $\langle \mathcal{S}, \mathcal{A}, \mathcal{Pr}, \mathcal{R}, s_0, \gamma \rangle$
- Define  $V^*(s)$  {optimal **value**} as the **maximum** expected **discounted reward** from this state.
- $V^*$  should satisfy the following equation:

$$V^*(s) = \max_{a \in \mathcal{A}_p(s)} \sum_{s' \in \mathcal{S}} \mathcal{Pr}(s'|s, a) [\mathcal{R}(s, a, s') + \gamma V^*(s')]$$

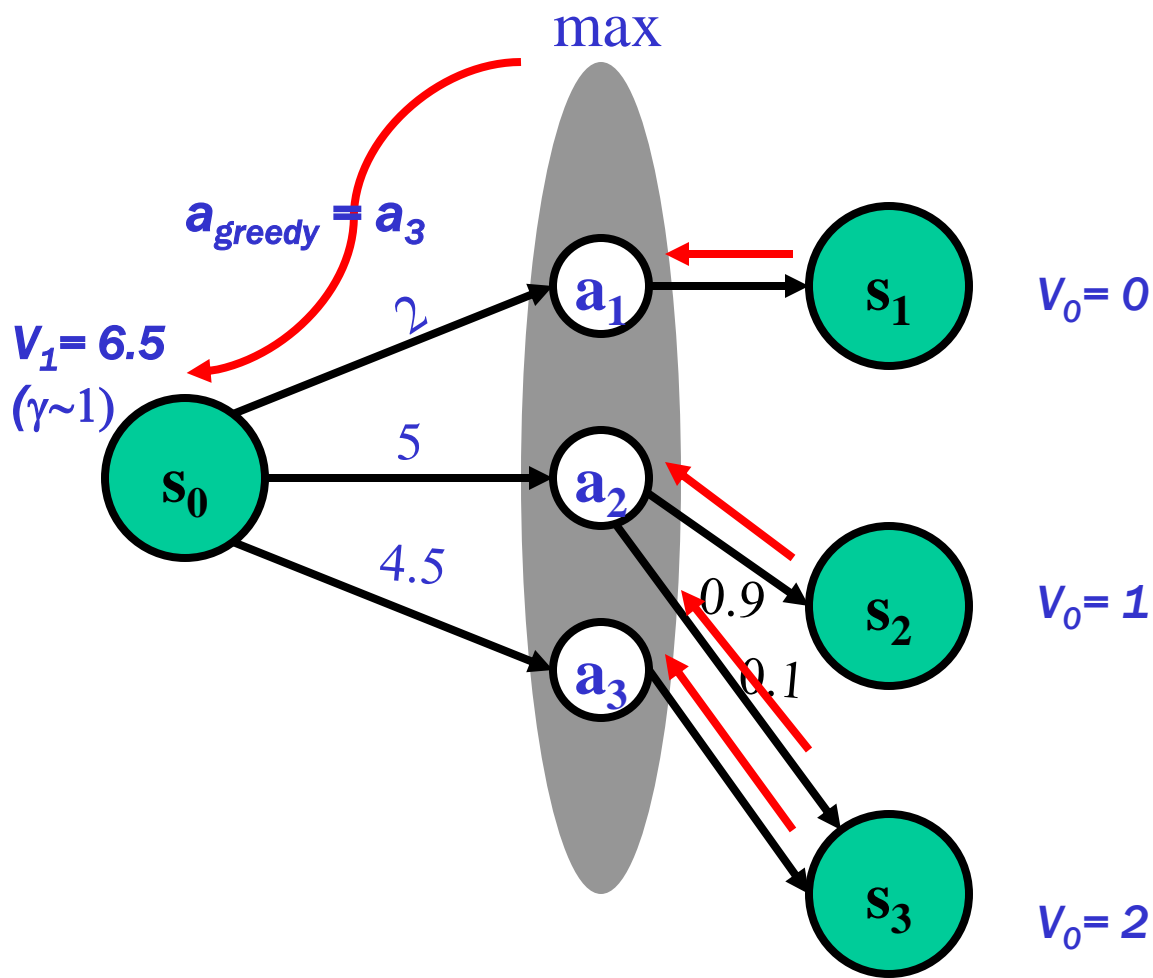
## Bellman Backup (MDP<sub>2</sub>)

- Given an estimate of  $V^*$  function (say  $V_n$ )
- Backup  $V_n$  function at state  $s$ 
  - calculate a new estimate ( $V_{n+1}$ ) :

$$Q_{n+1}(s, a) = \sum_{s' \in \mathcal{S}} Pr(s'|s, a) [\mathcal{R}(s, a, s') + \gamma V_n(s')]$$
$$V_{n+1}(s) = \max_{a \in Ap(s)} [Q_{n+1}(s, a)]$$

- $Q_{n+1}(s, a)$  : value/cost of the strategy:
  - execute action  $a$  in  $s$ , execute  $\pi_n$  subsequently
  - $\pi_n = \operatorname{argmax}_{a \in Ap(s)} Q_n(s, a)$

# Bellman Backup



$$Q_1(s, a_1) = 2 + 0 \gamma$$
$$Q_1(s, a_2) = 5 + \gamma 0.9 \times 1$$
$$+ \gamma 0.1 \times 2$$
$$Q_1(s, a_3) = 4.5 + 2 \gamma$$

## Value iteration [Bellman'57]

- assign an arbitrary assignment of  $V_0$  to each state.
  - repeat
    - for all states  $s$ 
      - compute  $V_{n+1}(s)$  by Bellman backup at  $s$ .
  - until  $\max_s |V_{n+1}(s) - V_n(s)| < \epsilon$
- Iteration n+1**
- Residual(s)**
-



# Comments

- Decision-theoretic Algorithm
- Dynamic Programming
- Fixed Point Computation
- Probabilistic version of Bellman-Ford Algorithm
  - for shortest path computation
  - $MDP_1$  : Stochastic Shortest Path Problem
- Time Complexity
  - one iteration:  $O(|S|^2|A|)$
  - number of iterations:  $\text{poly}(|S|, |A|, 1/(1-\gamma))$
- Space Complexity:  $O(|S|)$
- Factored MDPs = Planning under uncertainty
  - exponential space, exponential time

# Convergence Properties

- $V_n \rightarrow V^*$  in the limit as  $n \rightarrow \infty$
- $\varepsilon$ -convergence:  $V_n$  function is within  $\varepsilon$  of  $V^*$
- **Optimality**: current policy is within  $2\varepsilon\gamma/(1-\gamma)$  of optimal

# Policy Computation

$$\begin{aligned}\pi^*(s) &= \operatorname{argmax}_{a \in A_p(s)} Q^*(s, a) \\ &= \operatorname{argmax}_{a \in A_p(s)} \sum_{s' \in \mathcal{S}} \mathcal{P}r(s'|s, a) [\mathcal{R}(s, a, s') + \gamma V^*(s')]\end{aligned}$$

# Applications

- Stochastic Games
- Robotics: navigation, helicopter maneuvers...
- Finance: options, investments
- Communication Networks
- Medicine: Radiation planning for cancer
- Controlling workflows
- Optimize bidding decisions in auctions
- Traffic flow optimization
- Aircraft queueing for landing; airline meal provisioning
- Optimizing software on mobiles
- Forest firefighting
- ...

# Extensions

- Reinforcement Learning
  - learning the probability and rewards
  - acting while learning - connections to psychology