Introducere in Reinforcement Learning

Luciana Morogan

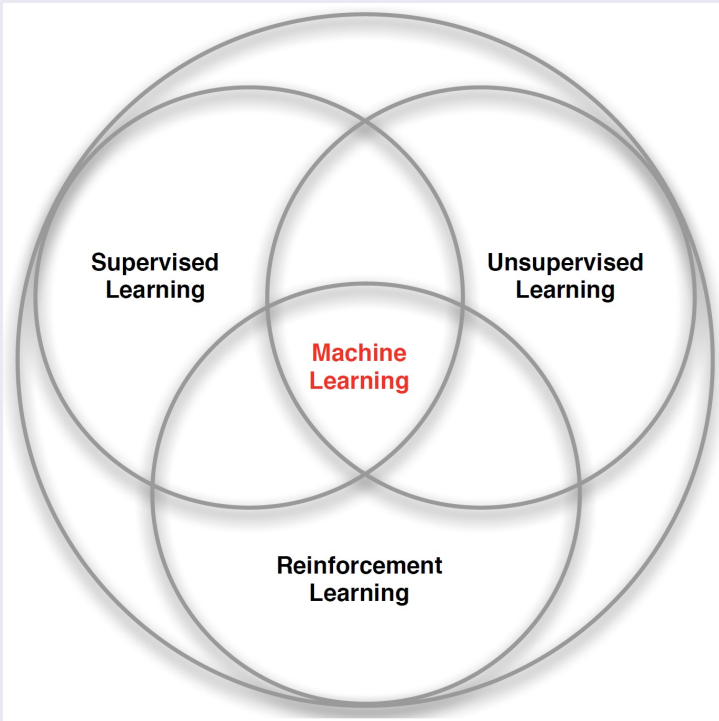
Academia Tehnica Militara

May 22, 2025

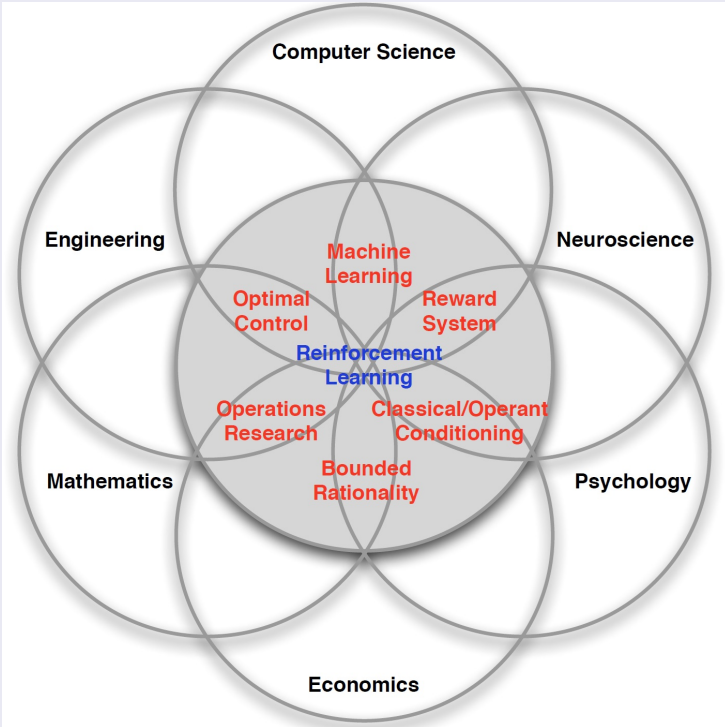
Cuprins

ML vs. RL

Ramurile ML



Fete ale RL



Sursa: curs RL - David Silver, DeepMind

RL $\neq$ Alte paradigme ML

- Nu exista un supervisor, doar un semnal de recompensa
- Feedback-ul este intarziat, nu instantaneu
- Timpul conteaza (datele sunt secventiale, nu independente si identic distribuite)
- Actiunile agentului influenteaza datele pe care le va primi ulterior

- *Exemple:* control centrala electrica, drone, elicoptere, roboti autonomi, invatare autonoma de jocuri (exp. tip Atari) etc.

Recompensa

- O recompensa R_t este un semnal scalar de feedback
- Defineste cat de bun este comportamentul agentului la momentul t
- Agentul incerca sa maximizeze cumulativ recompensa

Definitie (Ipoteza de Recompensa)

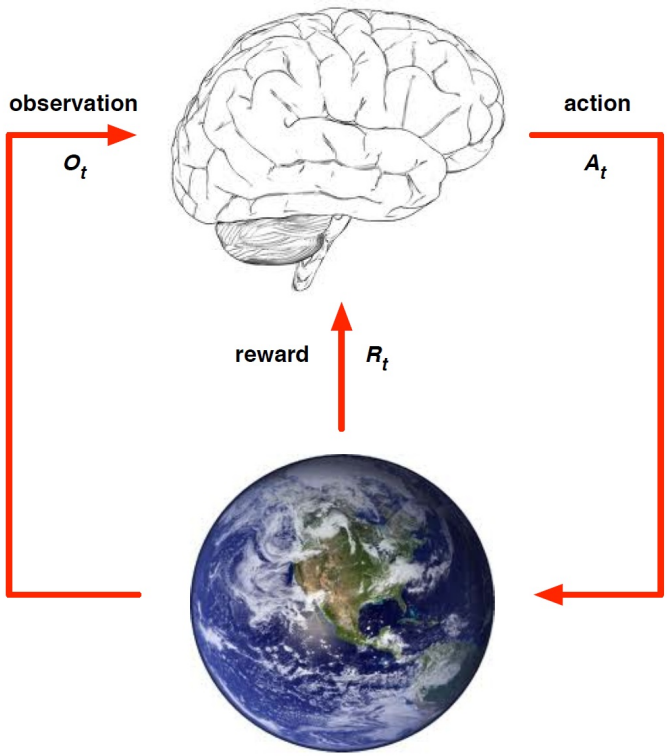
Toate obiectivele pot fi descrise prin maximizarea recompensei cumulative asteptate

- **Controlează o centrală electrică**
  - ✅ Recompensă pozitivă pentru producerea de energie
  - ❌ Recompensă negativă pentru depășirea pragurilor de siguranță
- **Fă un robot umanoid să meargă**
  - ✅ Recompensă pozitivă pentru mișcare înainte
  - ❌ Recompensă negativă pentru cădere
- **Joacă mai multe jocuri Atari mai bine decât oamenii**
  - ✅/❌ Recompensă pozitivă/negativă pentru creșterea/scăderea scorului

Secventialitatea in luarea deciziilor

- Obiectiv: alegerea acelor actiuni pentru maximizarea recompensei viitoare totale
- Actiunile pot avea consecinte pe termen lung
- Recompensa poate fi amanata
- Poate fi mai buna alegerea de a sacrifica recompensa imediata pentru maximizarea castigului pe termen lung

Agentul si mediul



- La fiecare pas t agentul
  - Executa actiunea A_t
  - Primeste observatia O_t
  - Primeste recompensa scalara R_t
- Mediul
  - Primeste actiunea A_t
  - Emite observatia O_{t+1}
  - Emite recompensa scalara R_{t+1}
- t este incrementat la fiecare pas al mediului

Istoric si stare

- Istoric:: secventa $H_t = O_1, R_1, A_1, \ldots, A_{t-1}, O_t, R_t$
- Tot ceea ce urmeaza depinde de istoric: agentul selecteaza actiunile, iar mediul observatiile/recompensele
- Stare:: totalitatea informatiei folosite pentru a determina ce tot ceea ce urmeaza

$$S_t = f(H_t)$$

Stari

Mediu

- *Starea mediului::* reprezentarea interna a mediului care, de obicei nu este vizibila agentului. Chiar daca ar fi vizibila, nu furnizeaza informatie relevanta

$$S_t^e$$

Agent

- *Starea agentului::* Reprezentarea interna a agentului, informatie folosita de algoritmii de RL si poate fi orice functie a istroricului

$$S_t^a = f(H_t)$$

Stare Markov

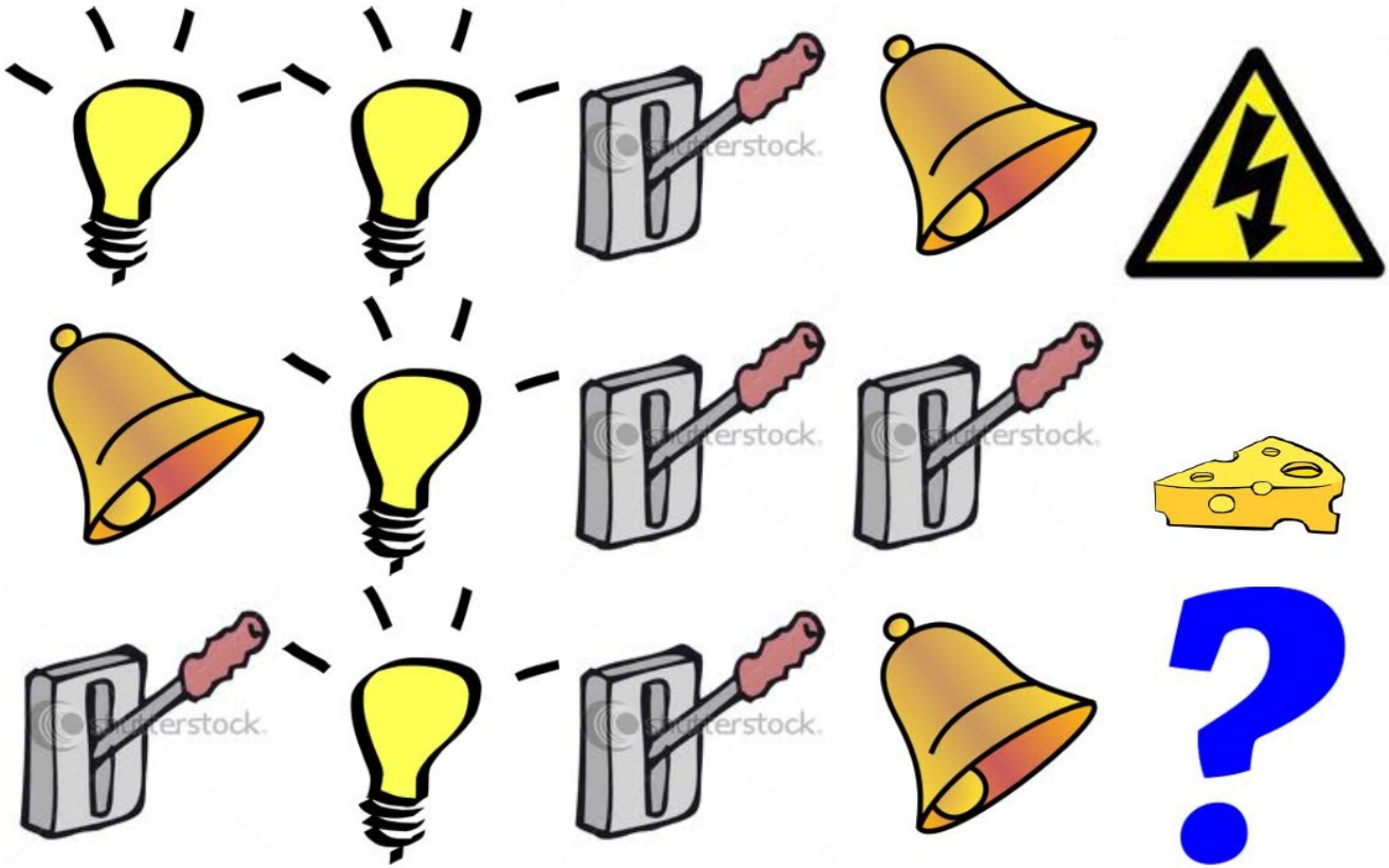
Stare Markov

O stare S_t este Markov dc si numai dc

$$\mathcal{P}[S_{t+1}|S_t] = \mathcal{P}[S_{t+1}|S_1, ..., S_t]$$

- "Fiind dat prezentul, viitorul este independent de trecut" :: o stare Markov contine toata informatia utila din histroy (istoric)
- Starea mediului S_t^e este Markov
- History-ul H_t este Markov

Importanta reprezentarii unei stari - exemplu



Mediu total/partial observabil

Total

- Agentul *observa direct* starea mediului $O_t = S_t^a = S_t^e$

Partial

- Agentul *observa indirect* starea mediului (exp. drona vede doar parte a mediului, un jucator de poker vede doar cartile de pe masa...)
- Agentul trebuie sa isi construiasca propria reprezentare care poate fi
  - Istoricul complet: $S_t^a = H_t$
  - Presupuneri asupra starii mediului

    $S_t^a = (\mathcal{P}[S_t^e = s_1], ..., \mathcal{P}[S_t^e = s_n])$

  - O retea neuronala (recurenta aici): $S_t^a = \sigma(S_{t-1}^a W_s + O_t W_o)$
  - ...

Agent RL - componente

Un agent RL poate cuprinde (una sau mai multe):

- *Politica* (en. *Policy*): functia ce defineste comportamentul agentului
- *Functia valoare* (en. *Value function*): functie ce defineste cat de buna este o stare si/sau actiune
- *Model*: reprezentarea mediului de catre agent

Politica

- O politica este o functie ce defineste comportamentul agentului
- mapare de la stare la actiune
- *Politica deterministă*: $a = \pi(s)$
- *Politica stocastica*: $\pi(a|s) = P[A_t = a|S_t = s]$

Functia valoare (evaluare)

- Functia valoare defineste cat de buna este o stare si/sau actiune
- Predictie a recompensei viitoare
- Ajuta in alegerea/decizia legata de actiunea viitoare
- $v_\pi(s) = \mathbb{E}_\pi \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \mid S_t = s \right]$

Modelul

- Modelul defineste reprezentarea mediului pentru agent, predictie a ceea ce mediul ar putea face prin
  - Predictie a starii urmatoare (Tranzitii $\rightarrow$ Dinamica)

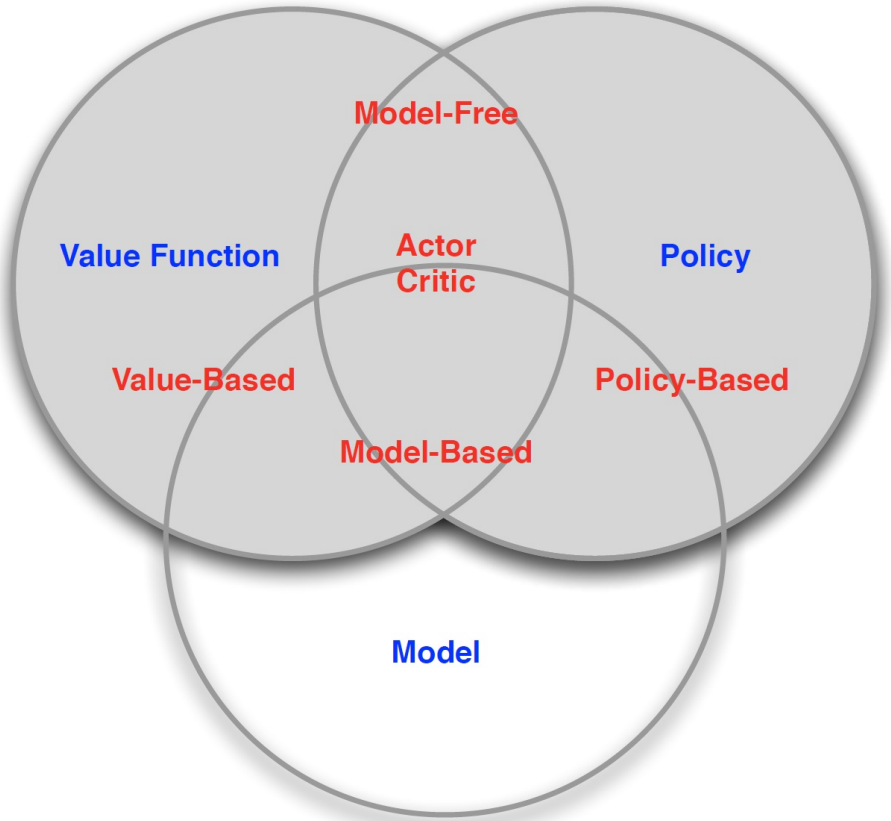
$$\mathcal{P}^a_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$$

  - Predictie a recompensei imediat urmatoare

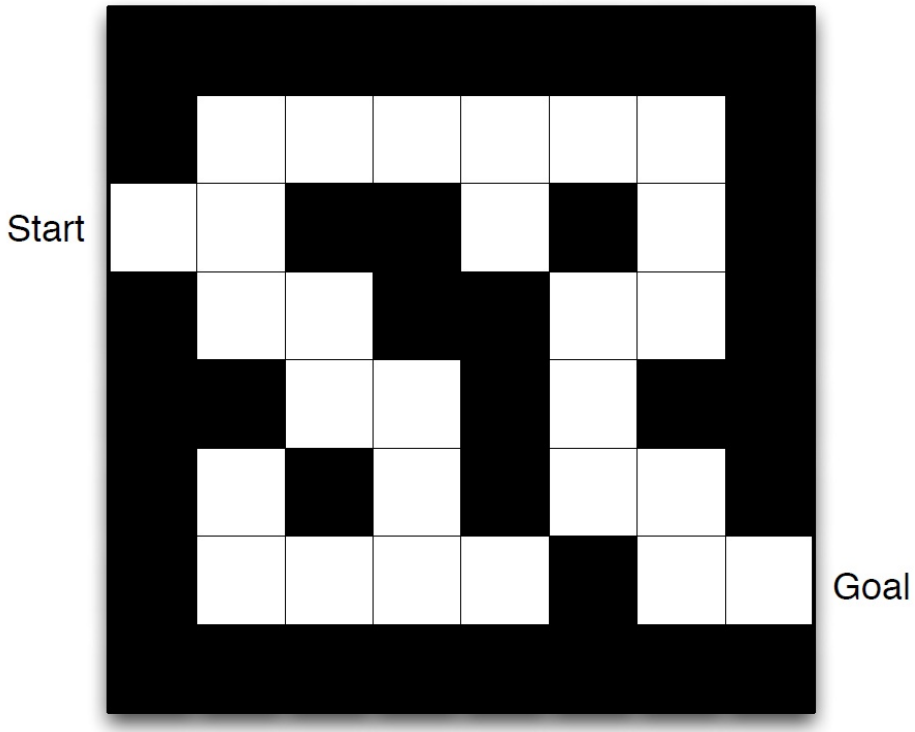
$$\mathcal{R}^a_s = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$$

Tipuri de agenti

- Bazati pe valoare (Value Function)
- Bazati pe politica (Policy)
- Actor Critic (Policy + Value Function)
- Model free (Policy +/- Value Function)
- Model based (Policy +/- Value Function + Model)

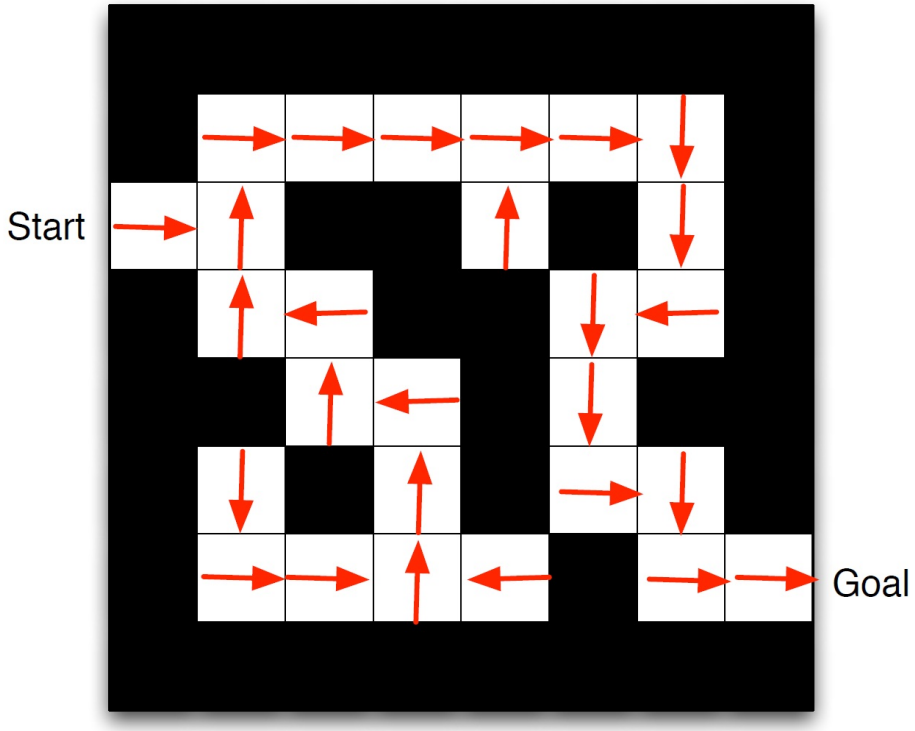


Exemplu (1)



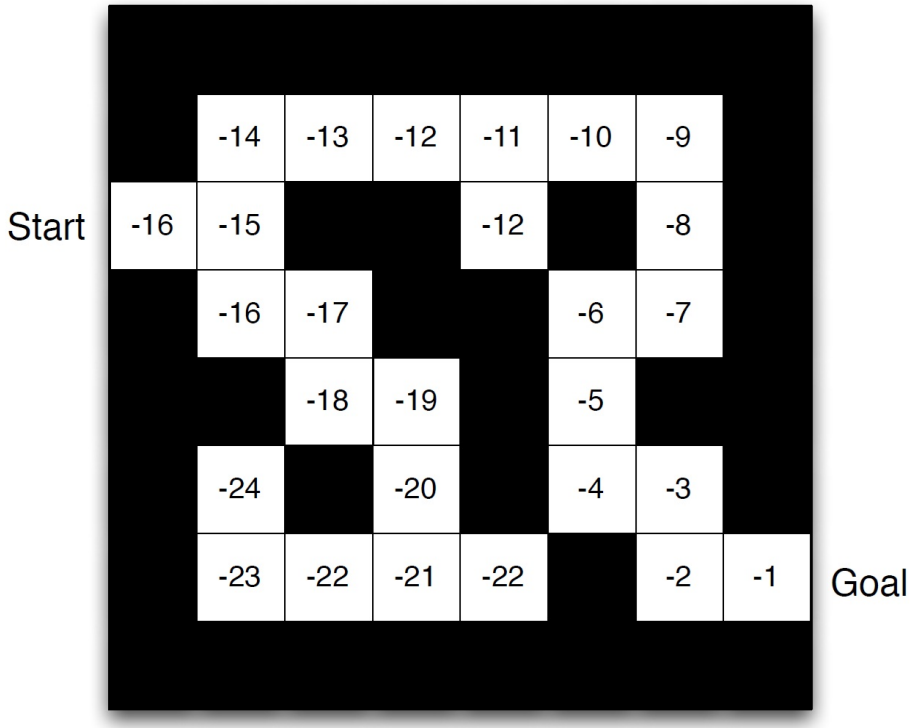
- Starea: locatia agentului
- Actiuni: N, S, E, V
- Recompensa: -1 per time-step

Exemplu (2)



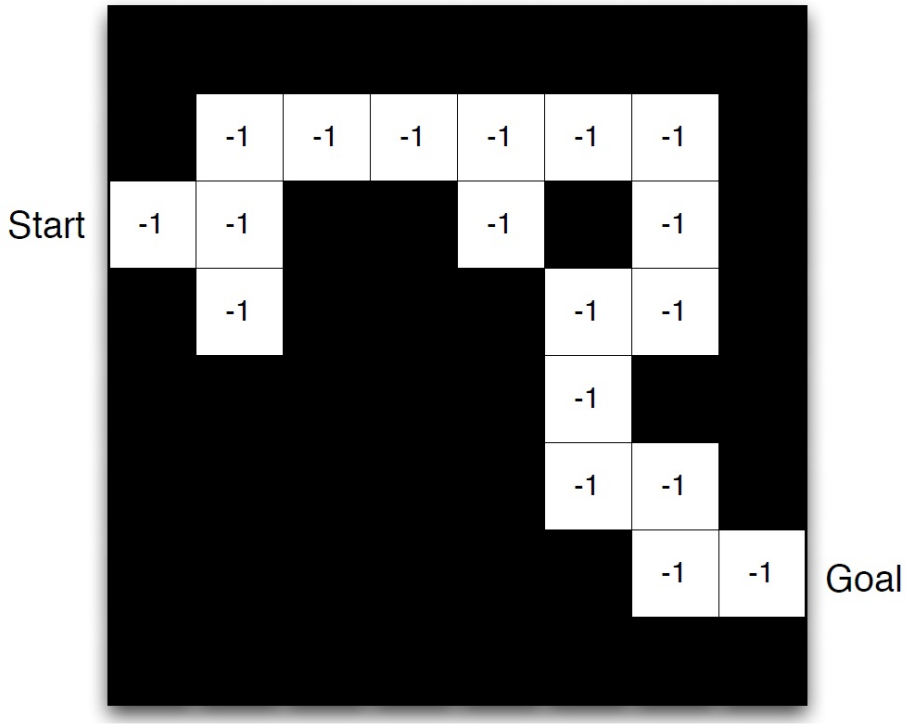
- Pentru fiecare stare s, sagetile arata politica $\pi(s)$

Exemplu (3)



- Pentru fiecare stare s, numarul din interiorul starii arata valoarea lui $v_\pi(s)$

Exemplu (4)



- Layout-ul labirintului reprezinta modelul tranzitiilor $\mathcal{P}^a_{ss'}$
- Valorile (aceleasi) -1 reprezinta recompensa imediata $\mathcal{R}^a_s$ pentru fiecare stare s

Probleme fundamentale in luarea secventiala de decizii

RL

- Mediul este initial necunoscut
- Agentul interactioneaza cu mediul
- Agentul isi imbunatateste politica

Planning

- Mediul este initial cunoscut
- Prin model agentul efectueaza computatii (nu sunt necesare interactiuni cu mediul)
- Agentul isi imbunatateste politica

Exemple??

Explorare si Exploatare

- RL:: invatare de tip trial-and-error:: Agentul incearca sa descopere o politica buna din experientele/interactiunile sale asupra mediului si fara a pierde prea multa recompensa pe parcursul invatarii
- Prin **explorare** se cauta informatie despre mediu
- Prin **exploatare** se foloseste informatia cunoscuta pentru maximizarea recompensei

Exemple

- Unde fac sapaturi (dupa petrol)
  Explorare:
  Exploatare:

- Cum joc un joc video?
  Explorare:
  Exploatare:

- Unde mancam?
  Explorare:
  Exploatare:

Predictie si Control

Predictie

- Evalueaza viitorul, fiind data o politica

Control

- Optimizeaza viitorul, fiind data cea mai buna politica