

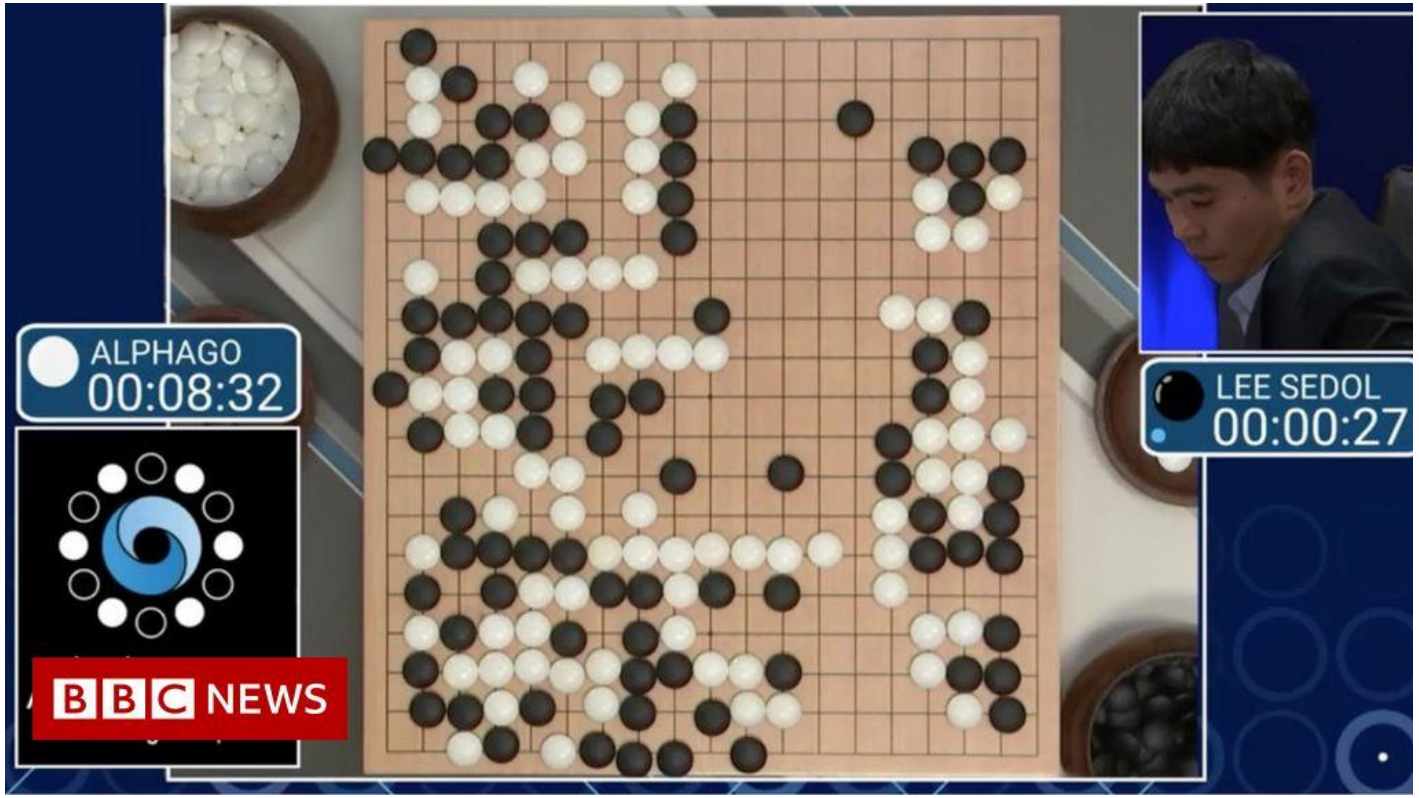
Strojové učenie II

prednáška 1 – Úvod do učenia posilňovaním

Ing. Ján Magyar, PhD.
Katedra kybernetiky a umelej inteligencie
Technická univerzita v Košiciach
2021/2022 letný semester

Prečo reinforcement?

- hranie hier
  - častý benchmark
  - zdrojom inovácií v RL
  - šachy, Go, StarCraft, ...



Zdroj: https://www.bbc.com/news/technology-35785875

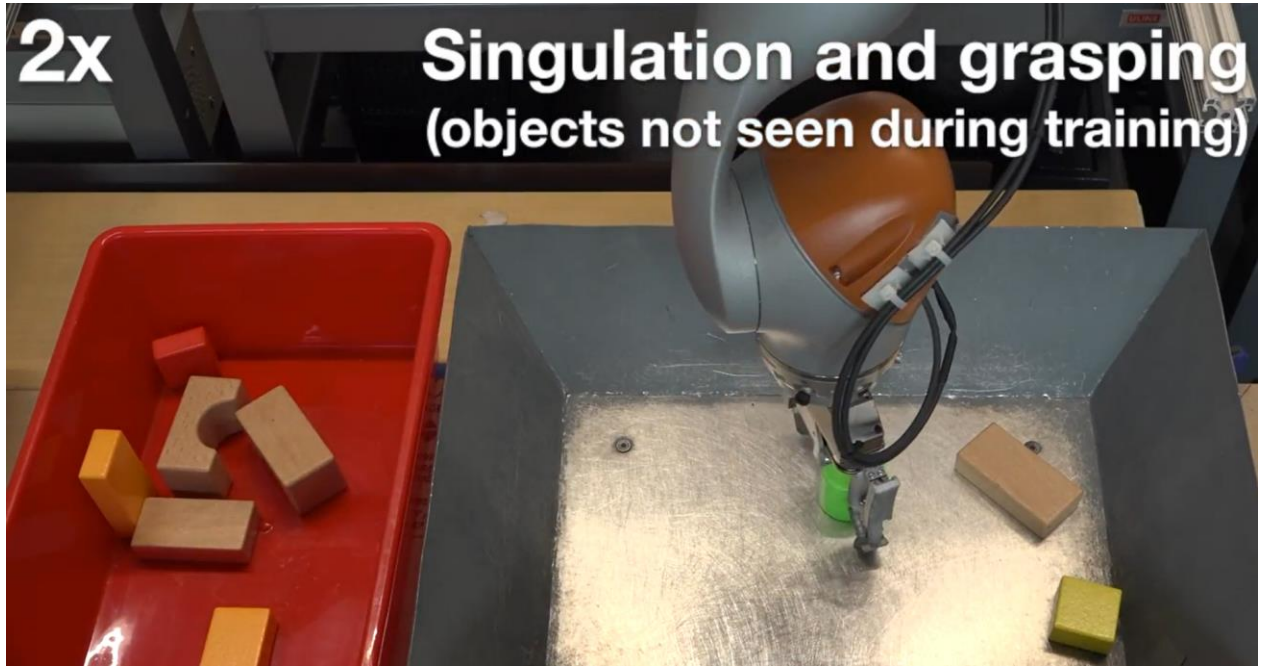
Prečo reinforcement?

- autonómne vozidlá
  - spracovanie kamerového obrazu
  - trénovanie na reálnych dátach
  - spojitý priestor akcií

Zdroj: AWS DeepRacer

Prečo reinforcement?

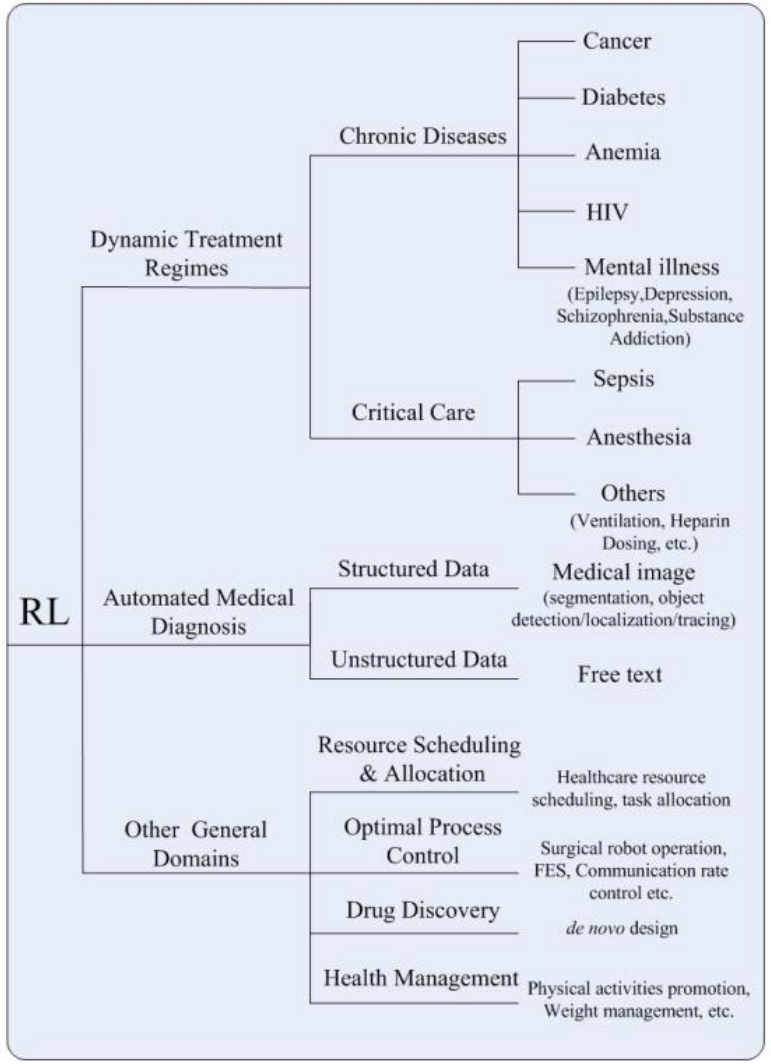
- riadenie robotov
  - veľký počet trénovacích chodov
  - pomocou simulácií
  - najmä pre priemysel



Zdroj: https://www.youtube.com/watch?v=W4joe3zzglU

Prečo reinforcement?

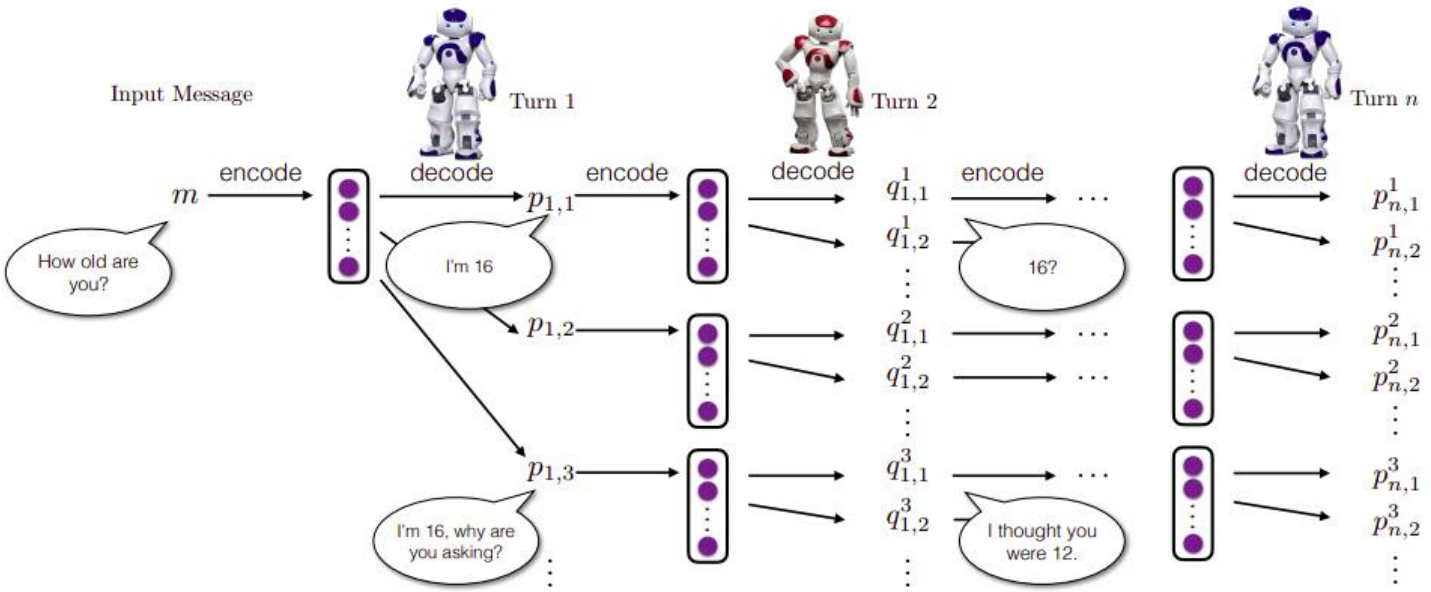
- medicína
  - prispôsobenie liečby
  - automatizovaná diagnostika
  - optimalizácia procesov
  - pridelenie zdrojov



Zdroj: Reinforcement Learning in Healthcare: A Survey

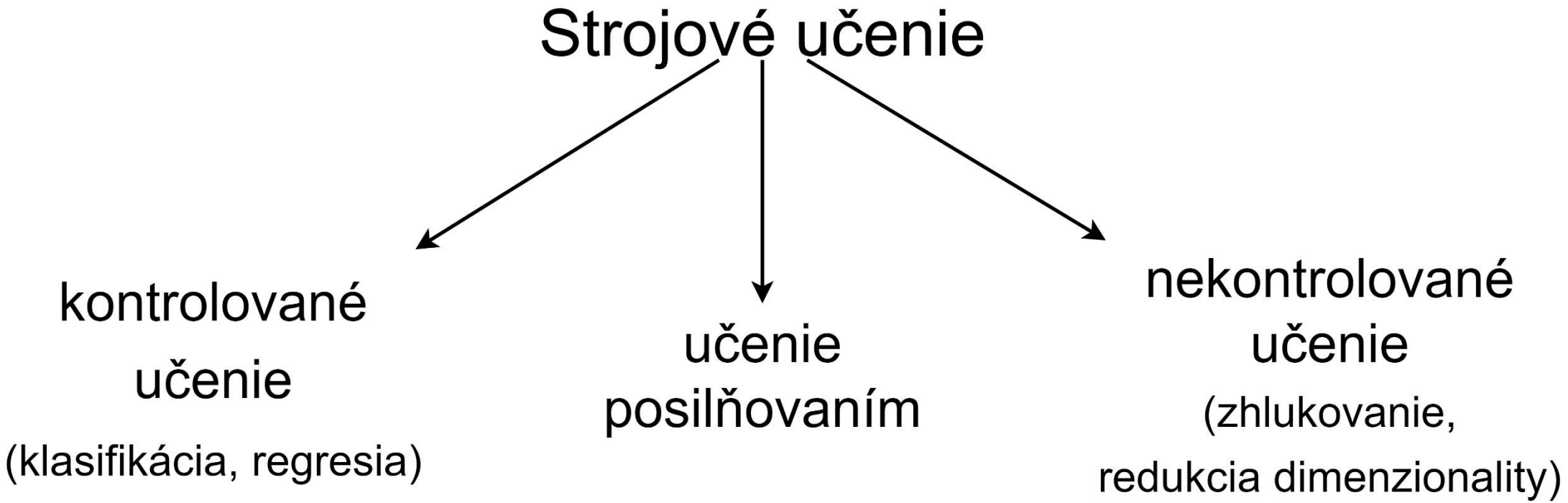
Prečo reinforcement?

- sociálne správanie
  - spracovanie prirodzeného jazyka
  - afektívne systémy
  - personalizácia



Zdroj: Deep Reinforcement Learning for Dialogue Generation

Strojové učenie a RL

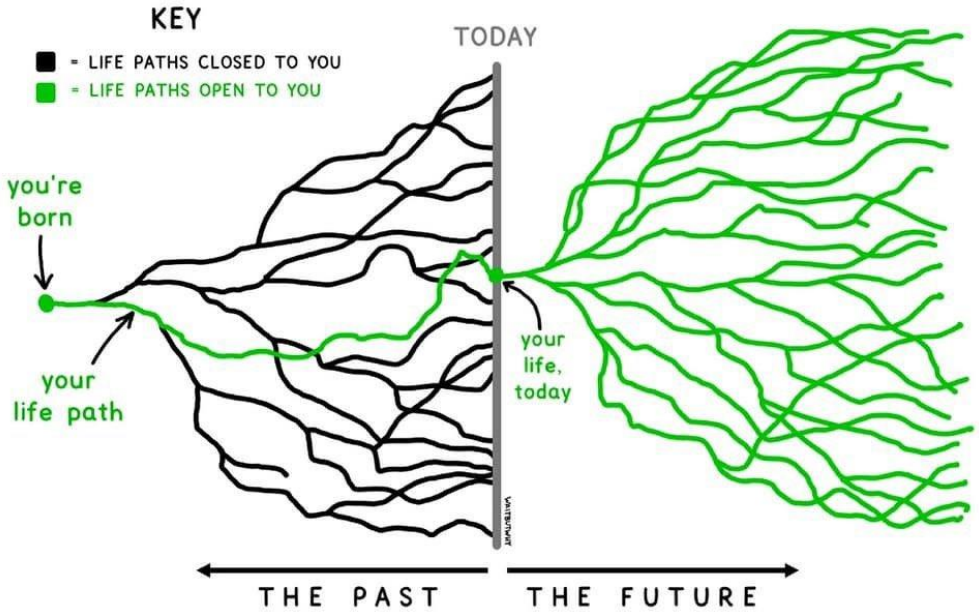


Základné charakteristiky RL

- cieľom je natrénovať agenta, ktorý reaguje na prostredie tak, aby dosiahol určený cieľ
- učenie je umožnené cez interakciu s prostredím
- agent dostáva spätnú väzbu

Interakcia s prostredím

- agent získa skúsenosti súčasne s učením
- agent riadi interakciu štýlom pokus-omyl – potrebuje úspechy aj neúspechy
- akcie môžu ovplyvniť budúce možnosti agenta



Spätná väzba

- zriedkavá / po každej akcii
- oneskorená / okamžitá – často určená pre postupnosť akcií
- ťažko odhadnúť (ne)správnosť akcií agenta
- pre agenta je často relevantná kumulatívna odmena

Čo je potrebné k RL?

- stavy – aktuálny status prostredia
- akcie – kroky, ktorými agent môže ovplyvniť prostredie
- prechody – určujú aktualizáciu stavu prostredia
- politika – určuje voľbu akcie
- odmena – definuje cieľ interakcie
- (model) – umožňuje predikovať chovanie prostredia

Stav

- typy stavu
  - stav prostredia s_t^e
  - agentov stav prostredia s_t^a
  - pozorovanie prostredia o_t
- plná pozorovateľnosť
  - $s_t^e = s_t^a = o_t$
- čiastočná pozorovateľnosť
  - $s_t^e \neq s_t^a$
  - agent aktualizuje s_t^a na základe predošlých pozorovaní
  - $s_t^a = (P[s_t = s_1], P[s_t = s_2], \dots, P[s_t = s_n])$


HIGH SCORE
1690
27250

Akcie

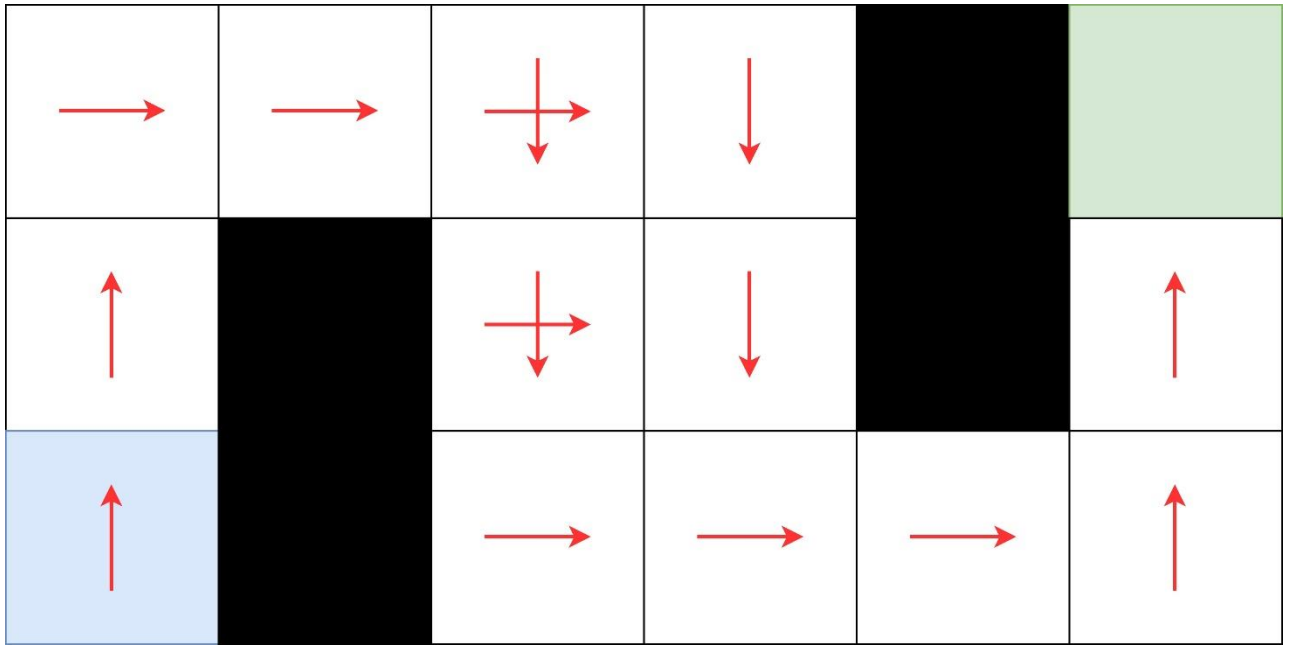
- množina akcií, ktoré sú agentovi k dispozícii: $a \in A$
- často diskrétne akcie, ale priestor akcií môže byť aj spojitý
- v každom stave môžu byť dostupné všetky akcie, alebo môžu byť aj limitované
- množina akcií je vždy daná
- nízkoúrovňové / vysokoúrovňové
- fyzické / mentálne

Prechody

- ak agent vyberie niektorú akciu, prostredie na ňu zareaguje a aktualizuje svoj stav
- aktualizácia stavu je popísaná prechodom $T: S \times A \rightarrow S$
- deterministické / nedeterministické
- v niektorých problémoch môžu byť časovo závislé

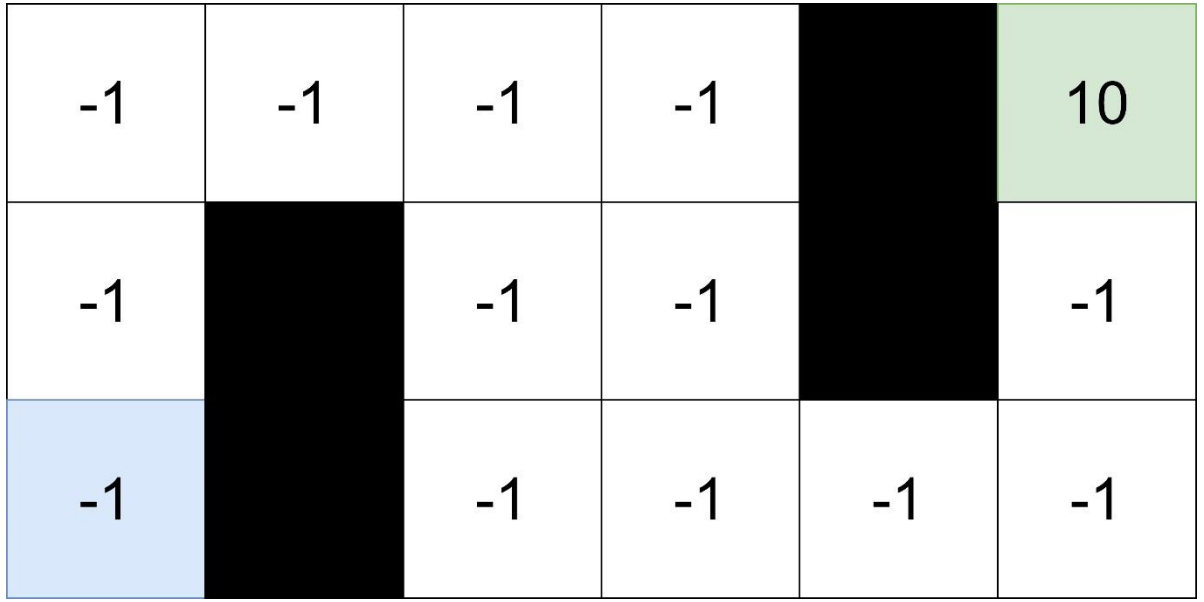
Politika

- mapuje stav na akciu agenta $\pi: S \rightarrow A(s)$
- deterministická $a = \pi(s)$
- stochastická $\pi(a|s) = P[A_t = a|S_t = s]$



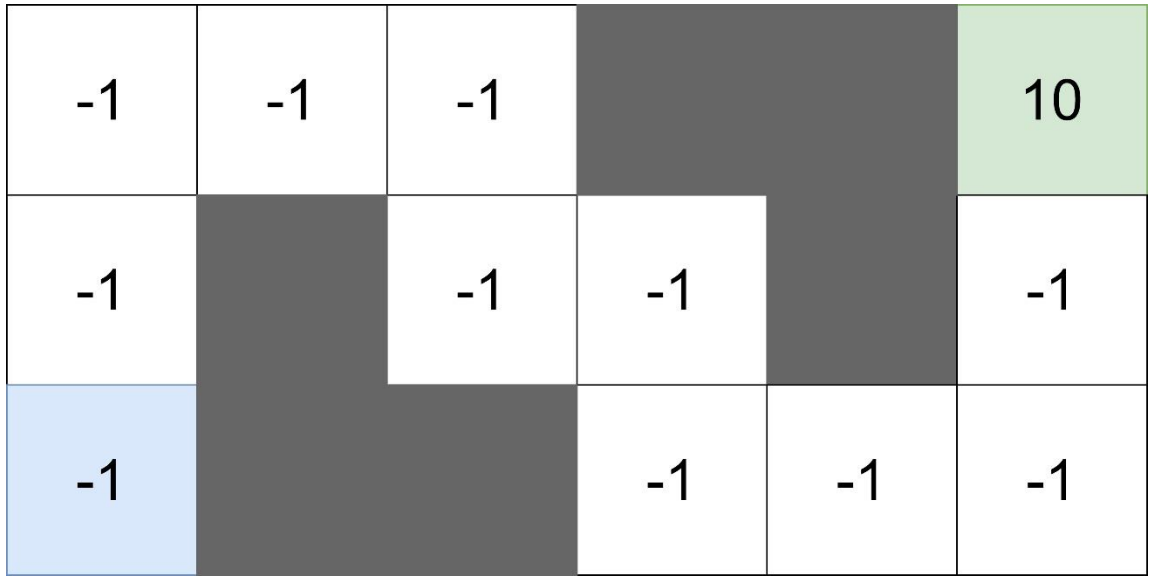
Odmena

- agent ju obdrží po zásahu do prostredia
- číselná hodnota
  - kladná – odmena, nulová – neutrálna, záporná – trest
- deterministická / stochastická
- agent by mal maximalizovať kumulatívnu odmenu



Model

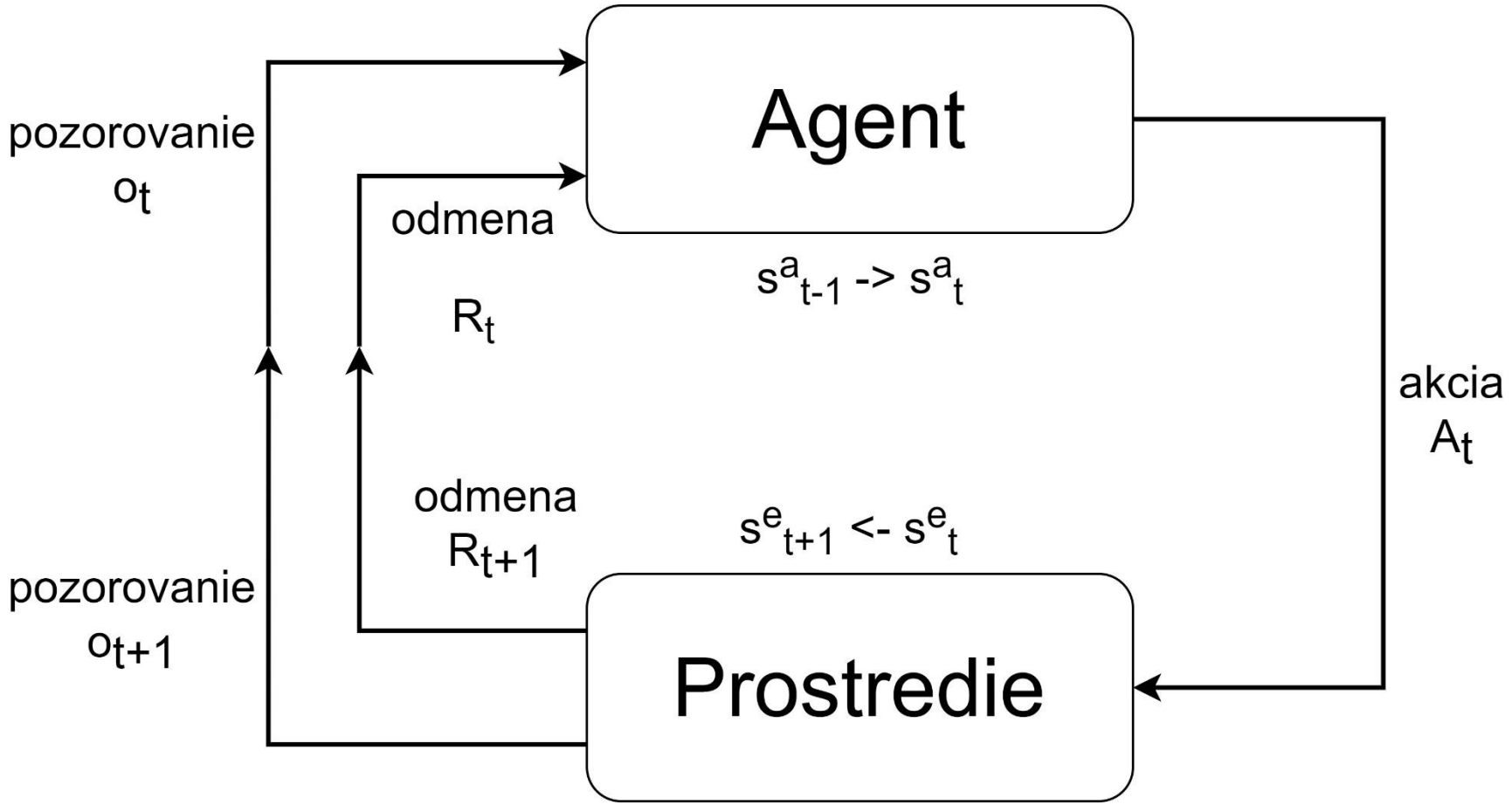
- agentova nepovinná reprezentácia prostredia
- predikuje dynamiku a odmenu
- nie je perfektný
- model môže byť poskytnutý, alebo ho agent sám zostrojí, alebo ho vôbec nepotrebuje



Sekvenčné rozhodovanie

- agent by mal akcie vyberať tak, aby maximalizoval budúcu kumulatívnu odmenu
- agent v kroku t
  - získava odmenu r_t
  - obdrží pozorovanie o_t
  - updatuje svoju reprezentáciu s_t^a
  - vyberie a vykoná akciu a_t
- prostredie v kroku t
  - na základe akcie a_t zmení svoj stav na s_{t+1}^e
  - pošle informáciu o novom stave o_{t+1}
  - pošle informáciu o odmene r_{t+1}

Interakcia medzi agentom a prostredím



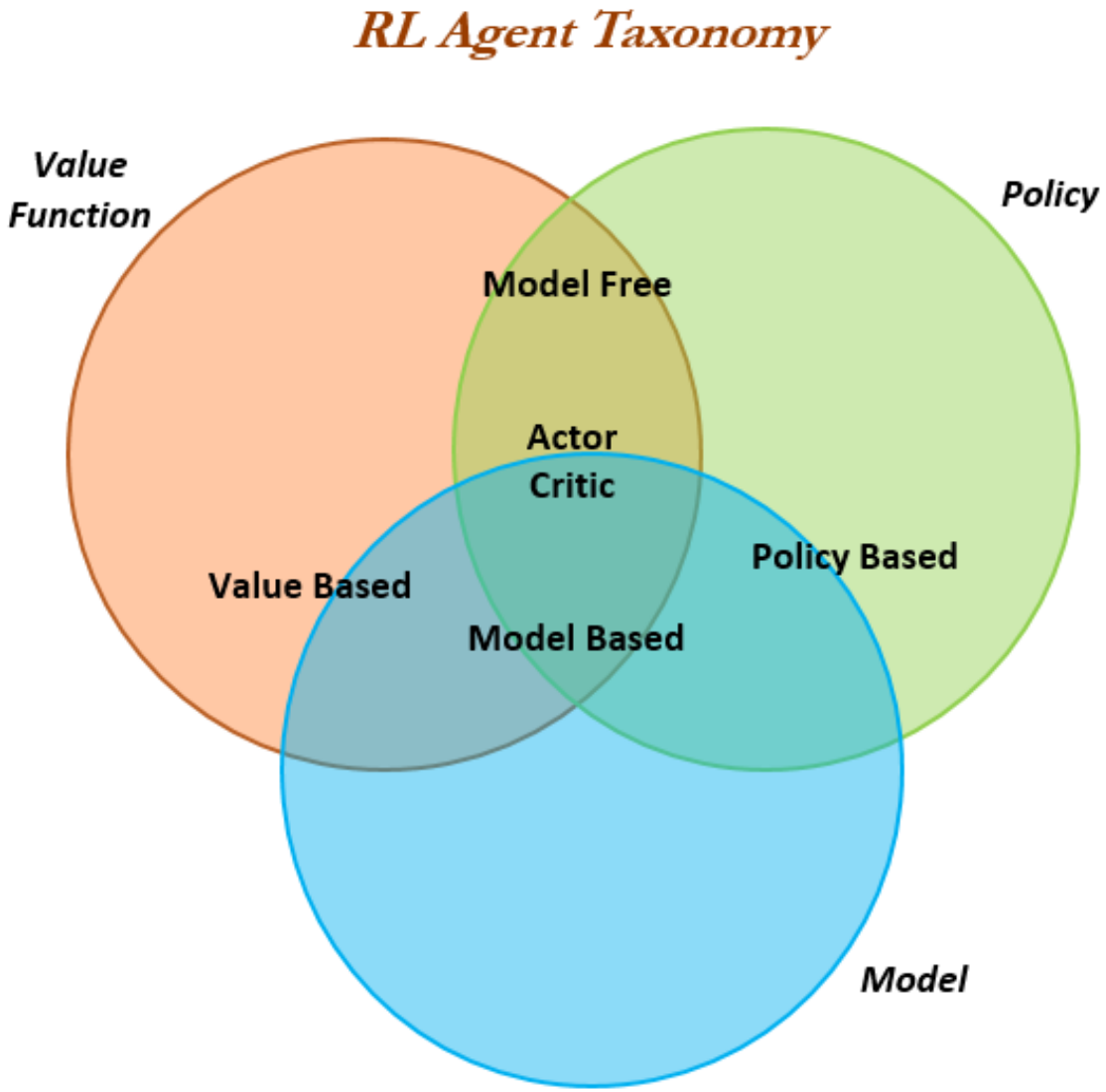
Rozhranie agent-prostredie

- fyzické rozhranie medzi prostredím a agentom (človekom, robotom, atď.)
- je možné, že agent iba vyberie akciu, ale je to už prostredie, ktoré ju vykoná
- poloha rozhrania je daná tým, čo agent ovláda a vie

Explorácia a exploatácia

- mal by sa agent spoliehať na existujúce vedomosti alebo skúmať nové možnosti riešenia?
- trénovanie prebieha počas získavania skúseností (pokus-omyl)
- exploatácia – agent využíva známe informácie
- explorácia – získavanie viac informácií o prostredí
- ideálne je kombinovať prístupy
- nechceme stratiť príliš z kumulatívnej odmeny

Taxonómia RL



Zdroj: Numerical Computing with Python (O'Reilly)