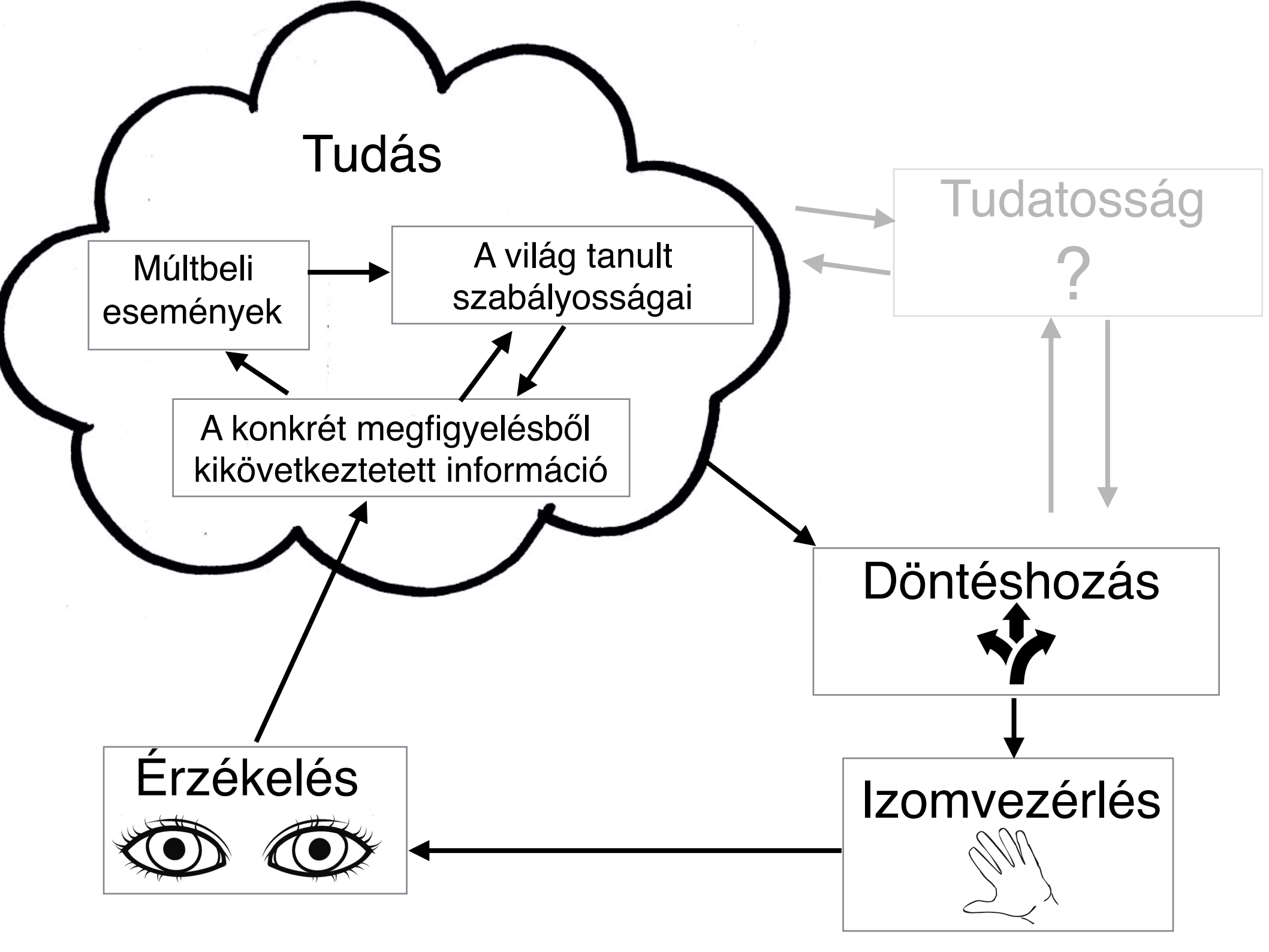
Megerősítéses tanulás


ALPHAGO
01:58:30
LEE SEDOL
01:59:41
Mind
Match
AlphaGo
Lee Sedol

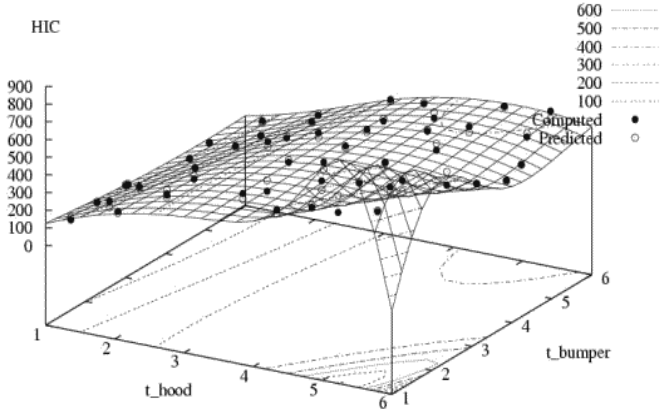

Tudás
Múltbeli események
A világ tanult szabályosságai
A konkrét megfigyelésből kikövetkeztetett információ
Tudatosság
?
Döntéshozás
Izomvezérlés
Érzékelés

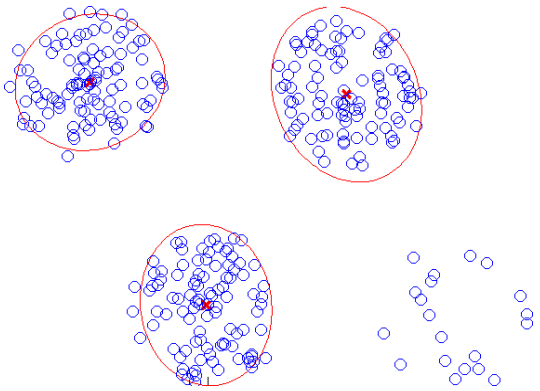
How to build a decision making model on top of the perceptual one?

- We need to choose a target variable from the model that we will try to optimise - this will be called the reward
- Motor output will be modelled a fixed set of possible actions that make modifications in the environment
- A combination of inferred values for the latent variables is called a (perceived) state of the environment
- We have to figure out which action to choose in every state

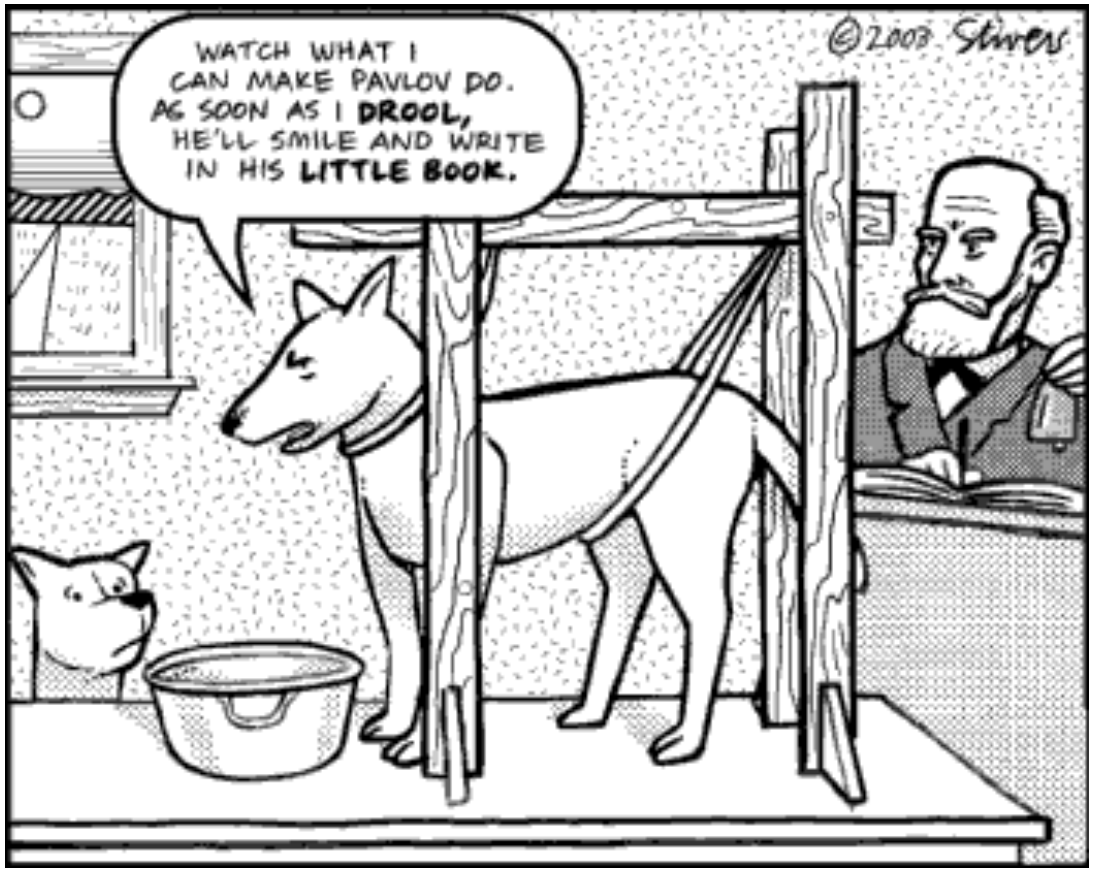
A tanulás alapvető típusai

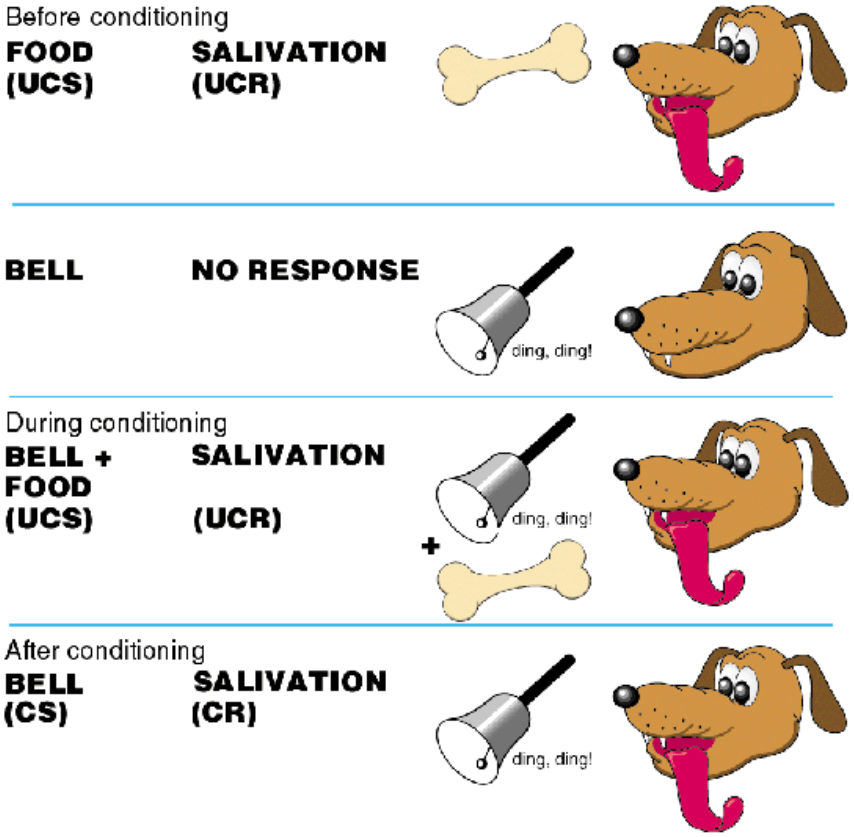
- Felügyelt
  - Az adat: bemenet-kimenet párok halmaza
  - A cél: függvényapproximáció, klasszifikáció
- Megerősítéses
  - Az adat: állapotmegfigyelések és jutalmak
  - A cél: optimális stratégia a jutalom maximalizálására
- Nem felügyelt, reprezentációs
  - Az adat: bemenetek halmaza





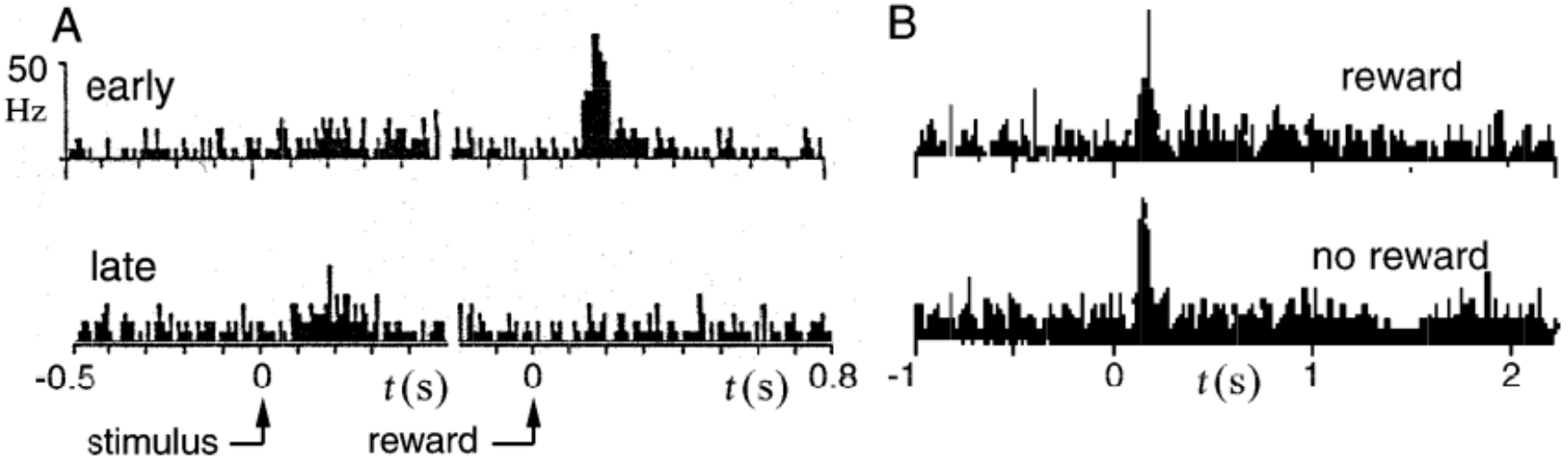
Classical conditioning





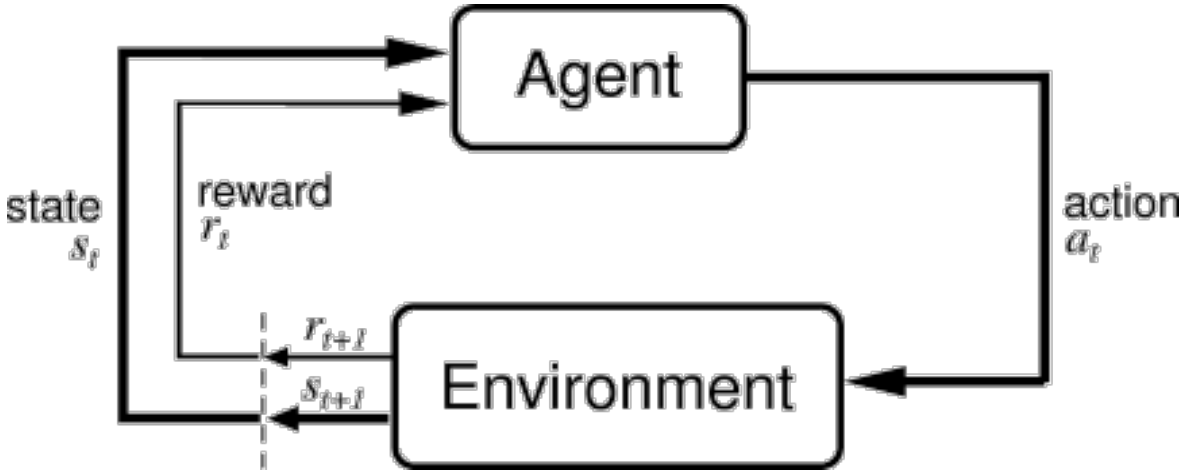
- Simplest case of learning from reward

Reward encoding in the cortex



- Dopaminerg neurons in the monkey's cortex respond according to the learned association between indicator variables and reward
- Activity is proportional to surprise

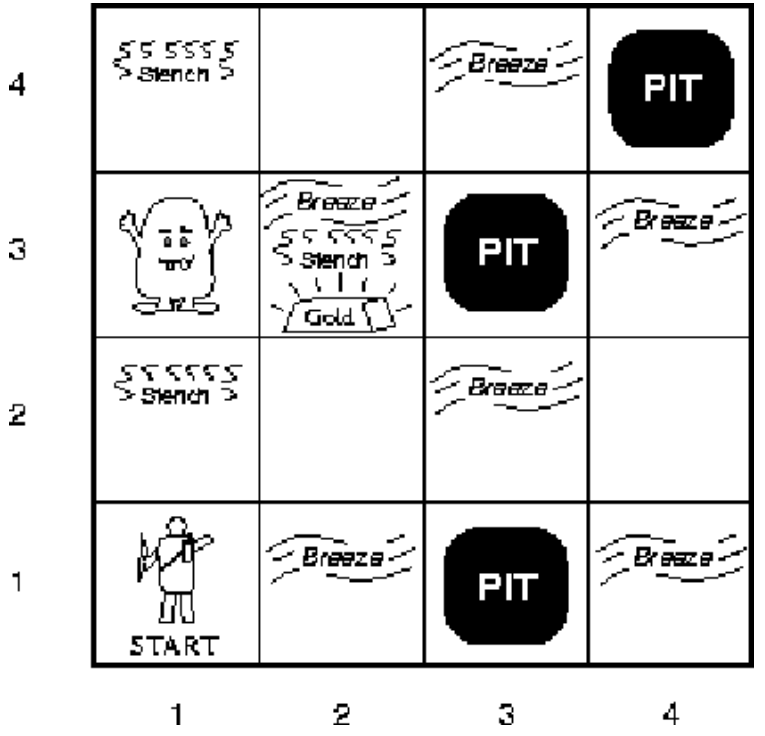
Az ágens-környezet modell



- Ágens: a tanulórendszer
- Környezet: minden, amit nem belsőleg állít elő
- Bemenet: aktuális állapot, jutalom
- Kimenet: cselekvés

Learning the value of states

- Simplest case: there is a finite number of states of the environment
- each state *s* is a combination of inferred values for the latent variables of the mental model (e.g. we take the *a posteriori* most probable values)
- for each state we assign a value, *V(s)*, that encodes the desirability of that state
- after each decision, at time *t*, we update the estimation of state values using previous estimations and the reward
- intuition: a value of a state is determined by the reward, and the value of other states that are accessible from it through a few actions



Russell & Norvig, 1995

$$V_{t+1}(s_t) = V_t(s_t) + \alpha[r_{t+1} + \gamma V_t(s_{t+1}) - V_t(s_t)]$$

A jutalom hosszútávú maximalizációja

- Összesített jutalom: valamilyen becslést kell alkalmazni a jövőre nézve
- Különbség a felügyelt tanulástól: egyiknél azt tudjuk, hogy a képtérben milyen messze vagyunk az optimálistól, a másiknál arra kapunk egy becslést, hogy a paramétertérben mennyire térünk el az optimálistól

Temporal Difference learning

- We learn from prediction error
- The value of the state we stepped on makes the previously visited state more similar to itself
- We can propagate to earlier states too

$$V_{t+1}(s_t) = V_t(s_t) + \epsilon \left[r_{t+1} + \gamma V_t(s_t + 1) - V_t(s_t) \right]$$

Model-based and model-free RL

- We have to use a strategy knowing the values of states
  - We need to know which action leads to which state
  - We can learn that from trying too
  - Or alternatively, we can learn the value of state-action pairs instead
- Model-free or Q-learning
  - has more parameters, needs less evaluations

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \epsilon \left[r_{t+1} + \gamma \max_a Q_t(s_t + 1, a) - Q_t(s_t, a_t) \right]$$

Tanulási szabályok

- Off-policy Q-learning

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha [r_{t+1} + \gamma \max_a Q_t(x_{t+1}, a) - Q_t(s_t, a_t)]$$

- On-policy szabály (SARSA): nem a lehető legjobb, hanem mindig az aktuálisan választott cselekvés alapján tanulunk

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha [r_{t+1} + \gamma Q_t(x_{t+1}, a_{t+1}) - Q_t(s_t, a_t)]$$

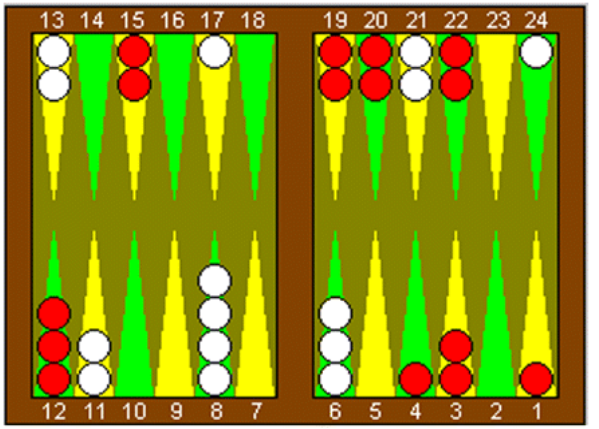
Exploration vs. exploitation

- When we don’t know anything about the environment yet, it doesn’t make sense to repeat the first series of actions that led to some reward
- When we know more, we can just use the best strategy we found
- Usual way: act randomly in the beginning, gradually increase of probability of choosing the action we think is best

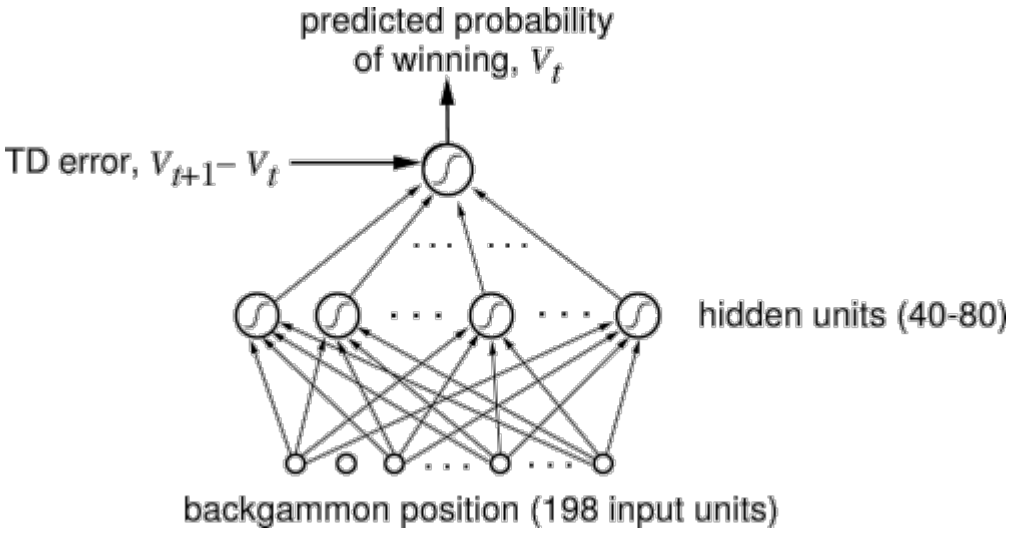
Representation of the value function

- If there are not so many, we can use a table
- With large and continuous spaces
  - we can only represent state variables
  - we need to generalise to states never visited
  - feedforward neural networks are a good choice
  - we have to construct a desired output for backpropagation at each step from the prediction error

TD tanulás neurális hálózattal



- Gerald Tesauro: TD-Gammon
  - Előrecsatolt hálózat
  - Bemenet: a lehetséges lépések nyomán elért állapotok
  - Kimenet: állapotérték (nyerési valség)
- Minden lépésben meg kell határozni a hálózat kimeneti hibáját
  - Reward signal alapján
- Eredmény: a legjobb emberi játékosokkal összemérhető



TD neurális reprezentációval

- Prediction error felhasználása a tanuláshoz
- Az állapotérték frissítése neurális reprezentációban:

$$w(\tau) \leftarrow w(\tau) + \varepsilon\delta(t)u(t-\tau) \qquad \delta(t) = \sum_t r(t+\tau) - v(t)$$

- A prediction error kiszámítása
  - A teljes jövőbeli jutalom kellene hozzá
  - Egylépéses lokális közelítést alkalmazunk

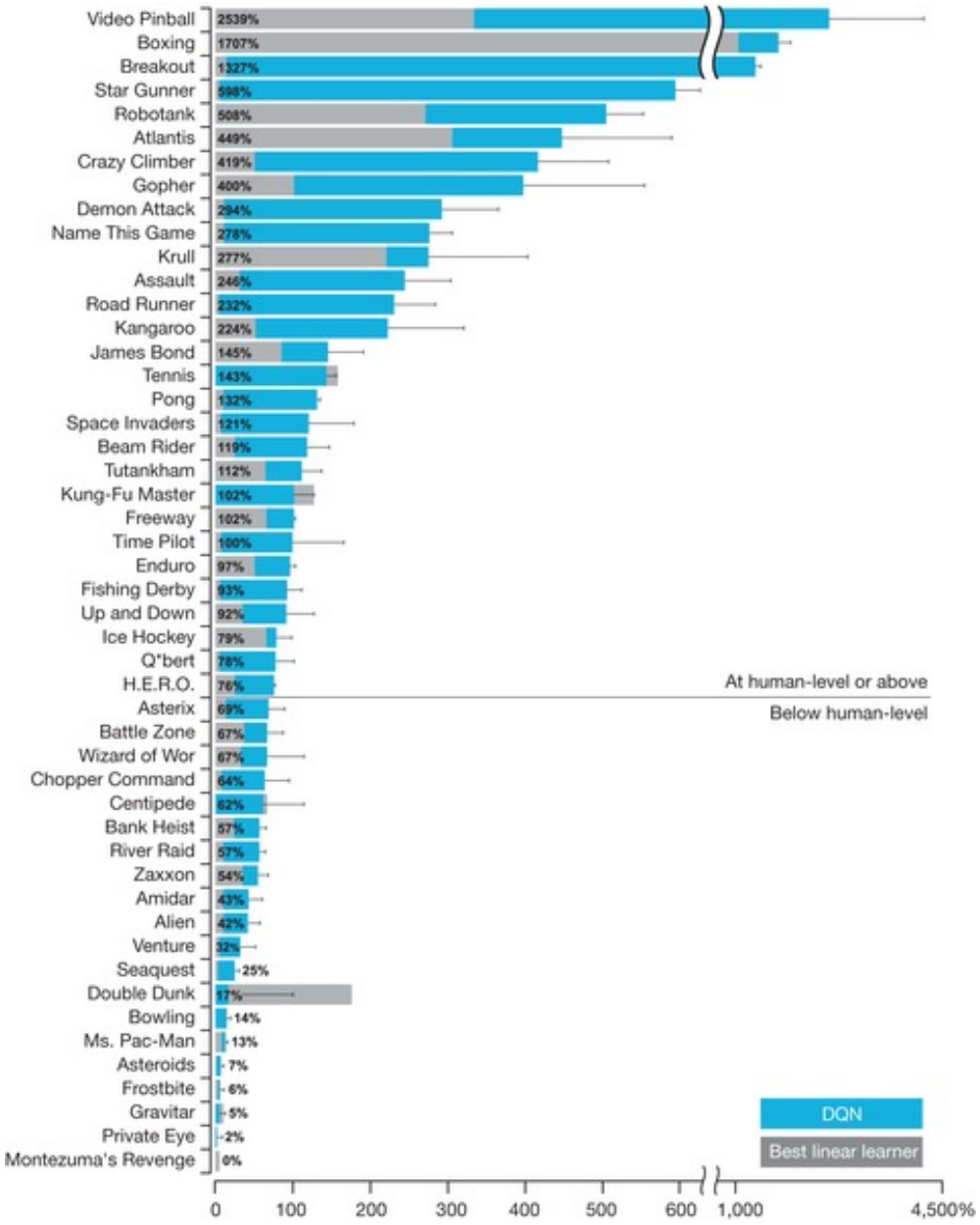
$$\sum_t r(t+\tau) - v(t) \approx r(t) + v(t+1)$$

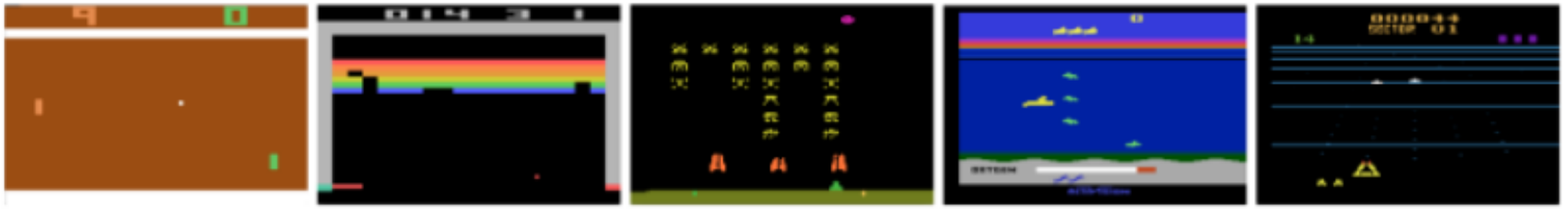
  - Ha a környezet megfigyelhető, akkor az optimális stratégiához konvergál
- A hibát visszaterjeszthetjük a korábbi állapotokra is

Learning to play computer games

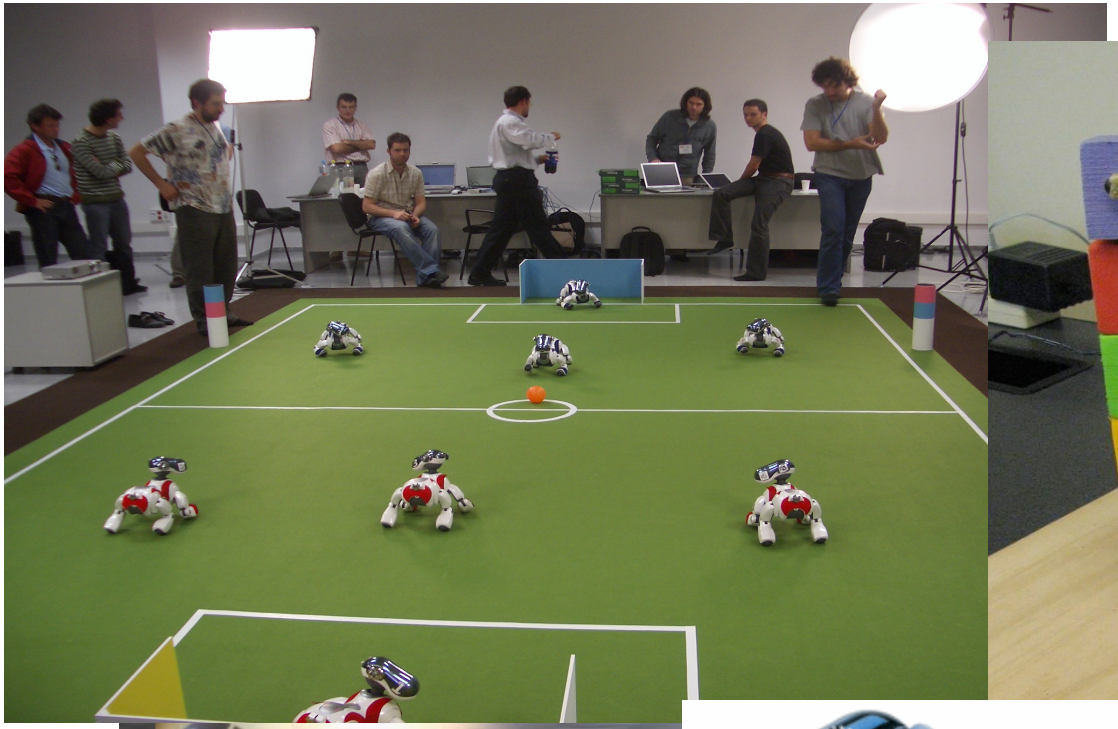
- reinforcement learning combined with deep networks
- observation: computer screen & score

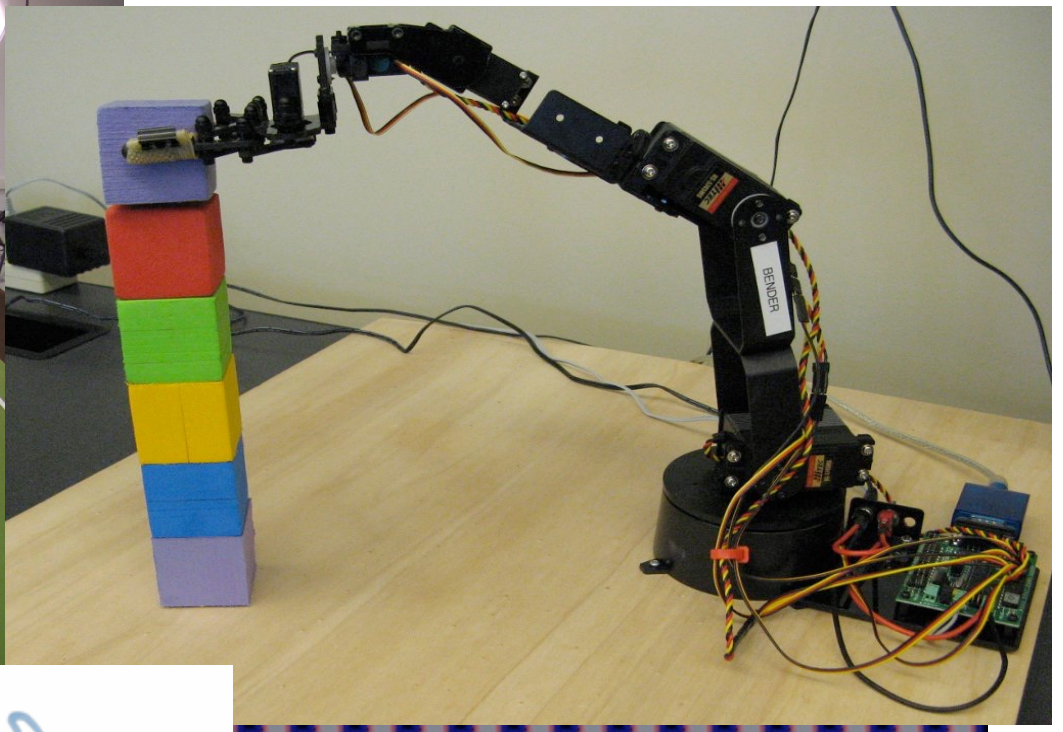
Mnih et al, 2015

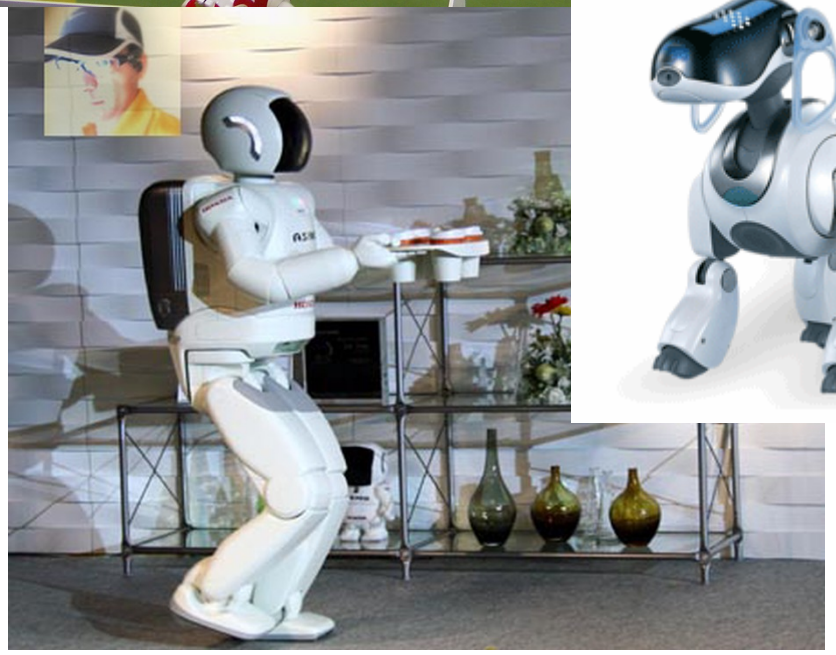


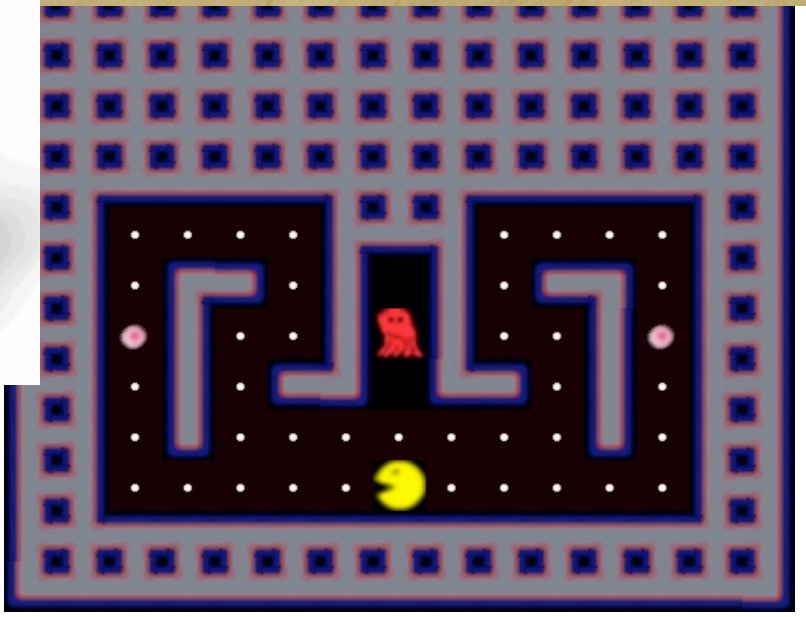


Gépi példák





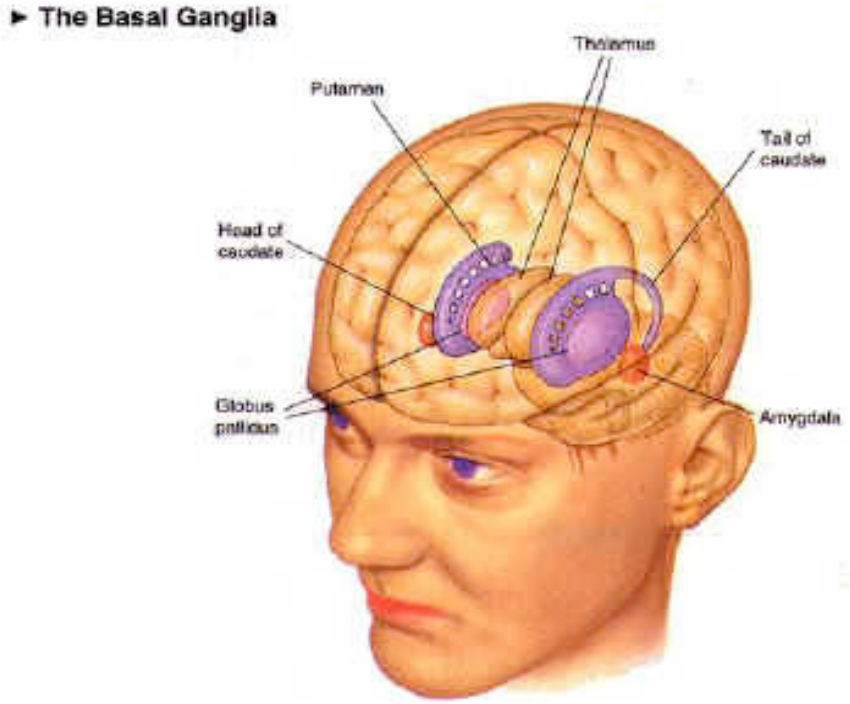


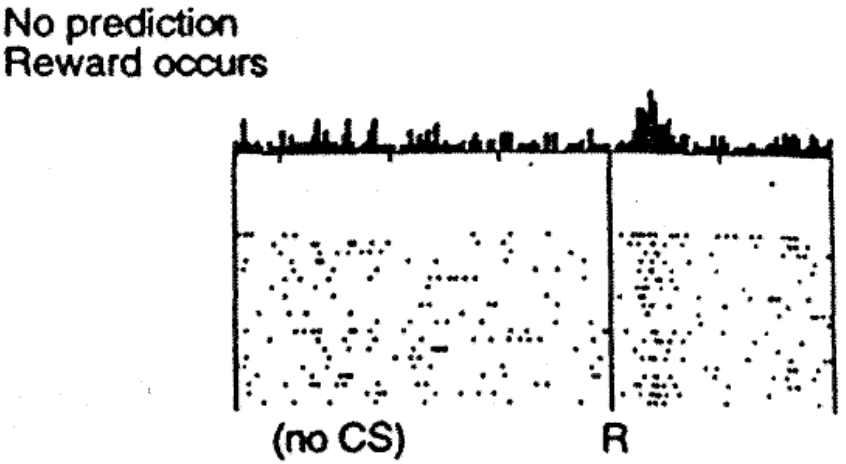


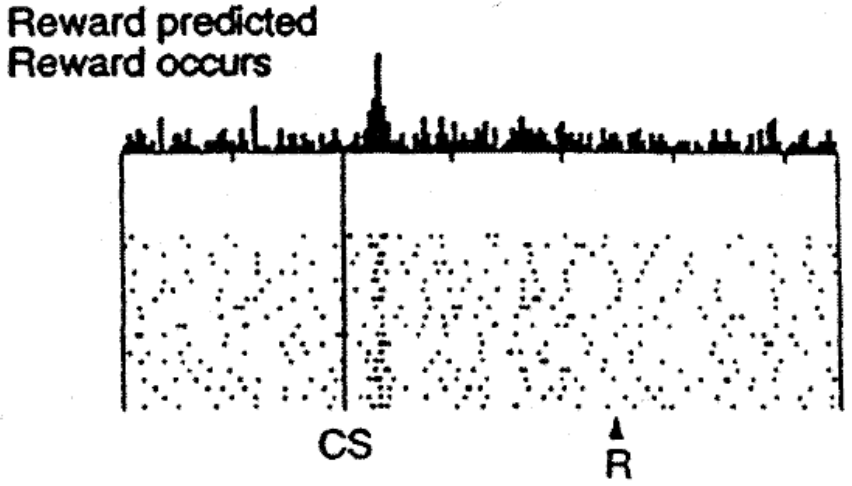
The effect of reward in dopaminerg cell of basal ganglia

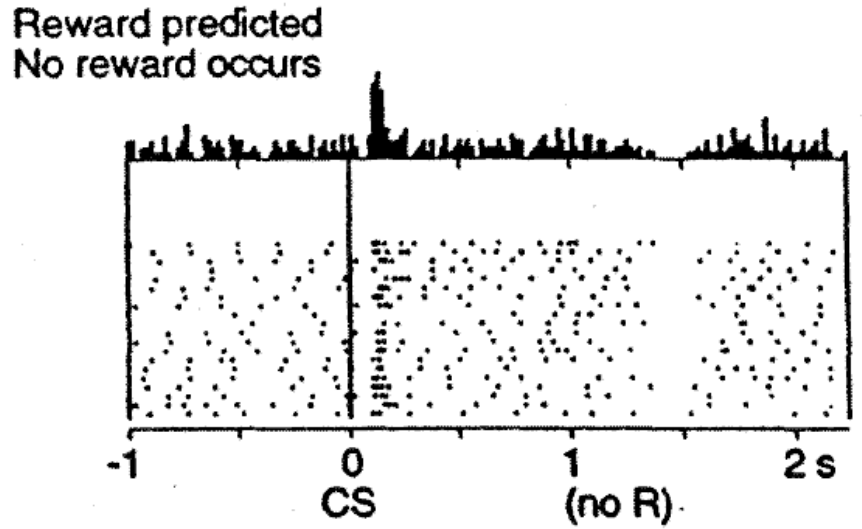
An interpretation:

Dopamine cells signals the difference between the expected and received reward.

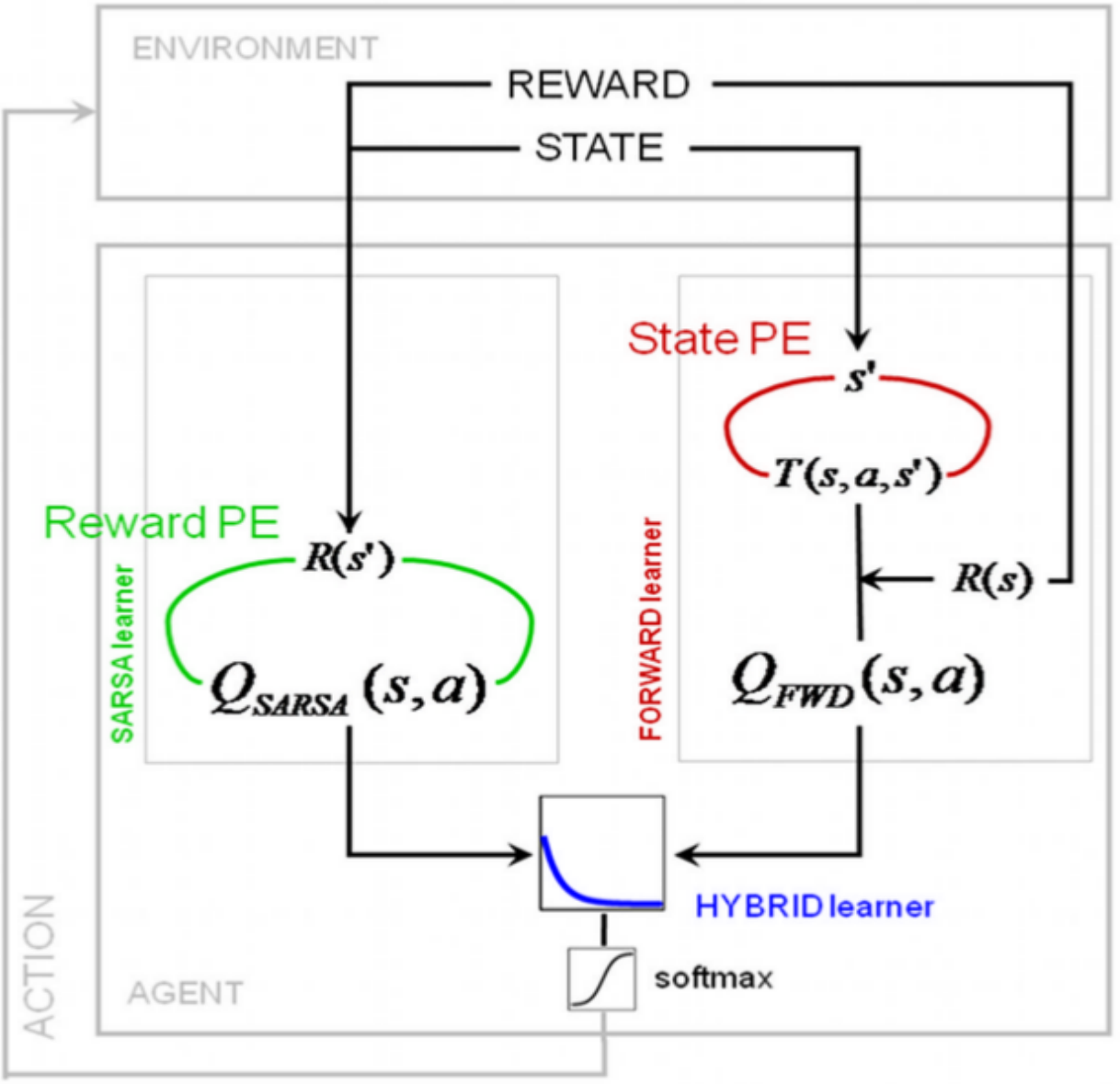




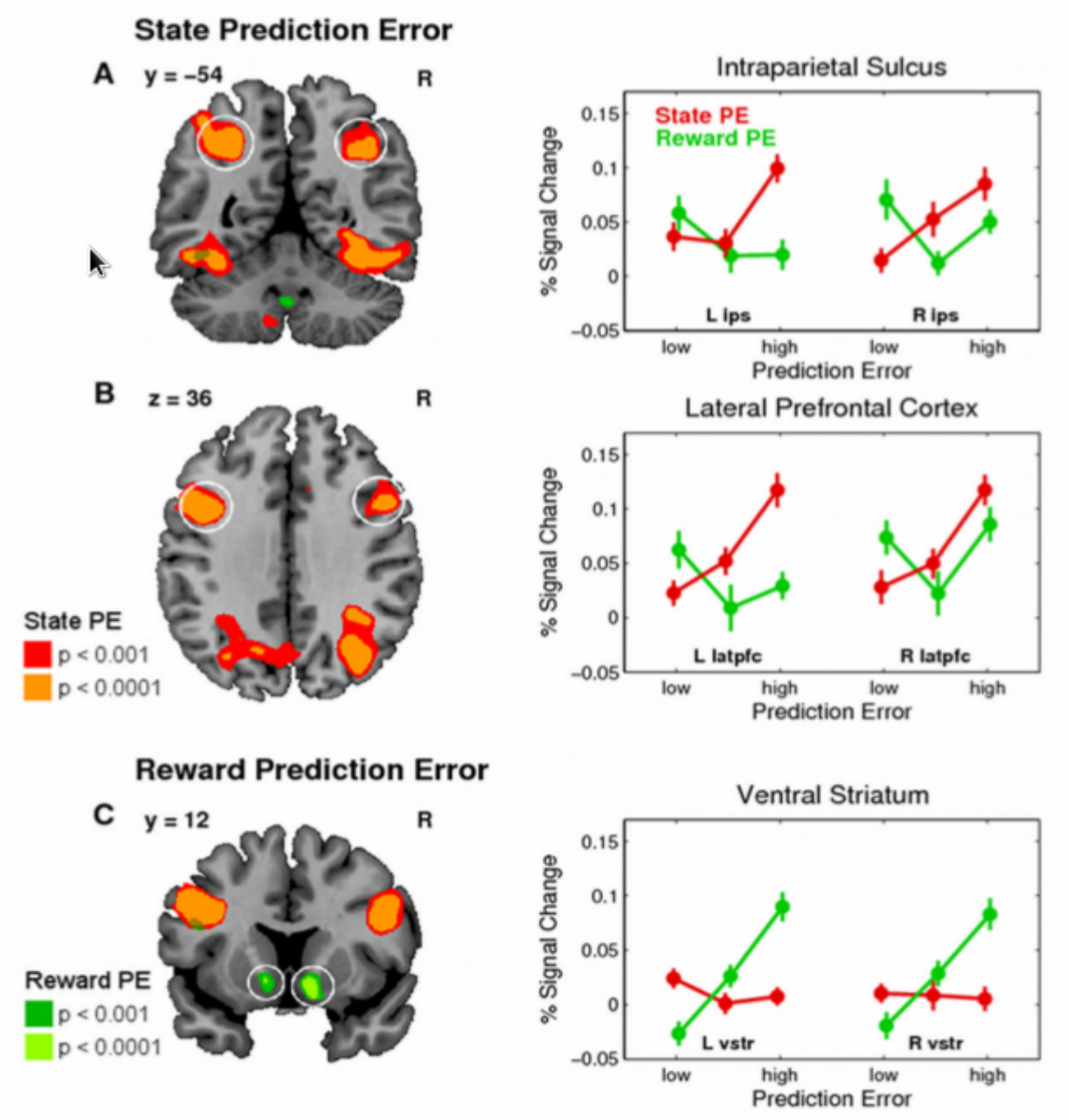




- Glascher, Daw Dayan, O'Doherty, *Neuron,* 2010.
- Modellalapú és modell nélküli algoritmusok predikciós hibájájával való korreláció



- fMRI eredmények



Mi az, amit tényleg tudunk az agy működéséről?

- Nagyon keveset
- Anatómiáról sokat tudunk
    - de nem ismerjük az idegsejtek pontos összeköttetési hálózatát (connectome)
    - a lokális összeköttetési mintázatok is csak részben ismertek
- Dinamikáról is sokat tudunk
    - le tudjuk írni egyes neuronok és hálózatok elektromos és kémiai viselkedését
    - de nehezen tudjuk ezt idegrendszeri funkciókhoz kötni
- Lokalizációról is sokat tudunk
    - alacsony szintű érzékelés, motorvezérlés, epizodikus memória, stb
    - sok funkcióról csak sejtések vannak
- Sok receptív mezőt ismerünk
    - feltéve, hogy a megfelelő stimulustulajdonságokat rögzítettük, amikor leírtuk őket
    - objektumokról és magasabbrendű koncepciókról csak tippjeink vannak
- Sokat tudunk arról, hogy hogy kell hasonló problémákat megoldani, mint az agy
    - specializált megoldásokkal, általános problémamegoldó nincs
    - fogalmunk sincs, hogyan fordíthatnánk le ezeket biofizikai implementációra
- Egy kicsit arról is tudunk, hogy hogyan nézhet ki a mentális modell, és hogyan lehet következtetni benne

HF

- Add meg a kincskeresős játékhoz tartozó állapot- és cselekvésteret
- Valósítsd meg valamilyen programnyelven az ágens és a környezet szimulált interakcióját
- Q-learning segítségével tanuld meg, melyik állapotban melyik cselekvés a leghasznosabb
- Eredményedet ábrákkal illusztráld