An introduction to Reinforcement Learning

3rd of May 2022

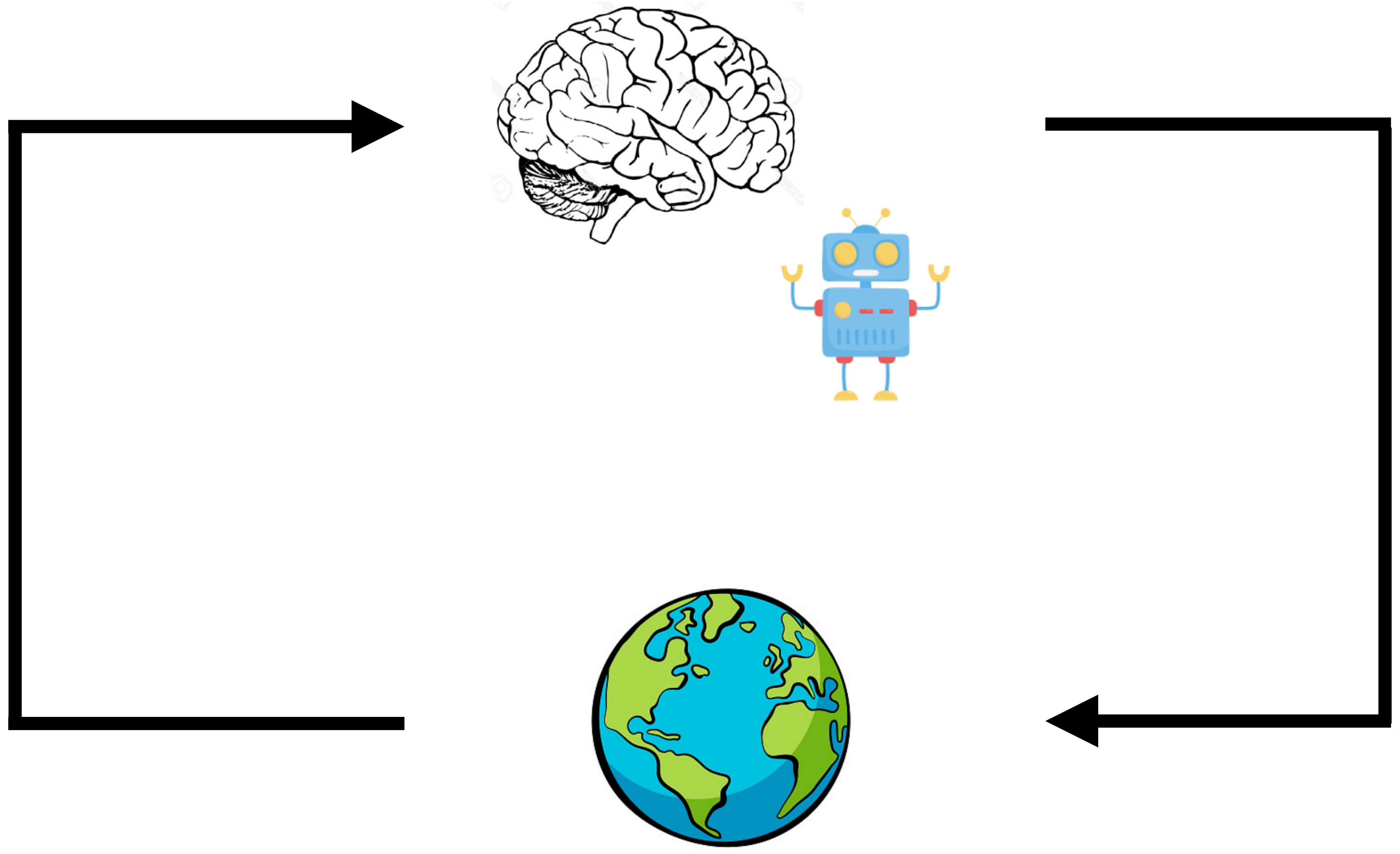
Recap: Basic setup: how to agents learn to act?

Agents can learn a **model of the environment** to make smarter decisions, e.g.:

$$P(s_{t+1} = s \mid s_t = s, a_t = a)$$

Based on a reward signal, agents learn **values of actions/states:**

$$V_\pi(s) = \mathbb{E}_\pi[R \mid s_0 = s]$$



Reward r_t

Action a_t

State s_t

Values can be **learnt** (simplified!!):

$$V(s) \leftarrow V(s) + \alpha \cdot (r - V(s))$$

Learning rate

Prediction error

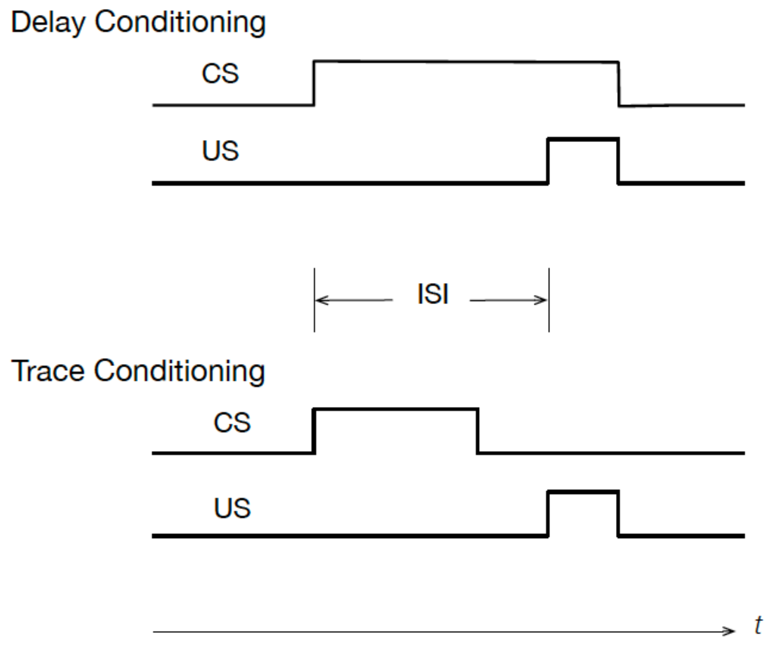
Action is governed by a **policy**:

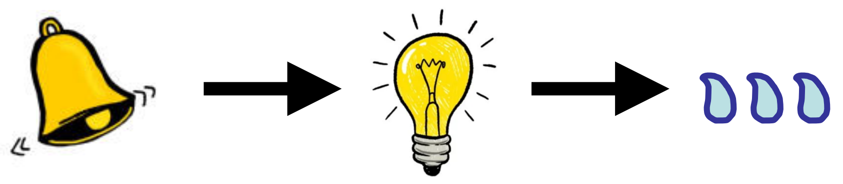
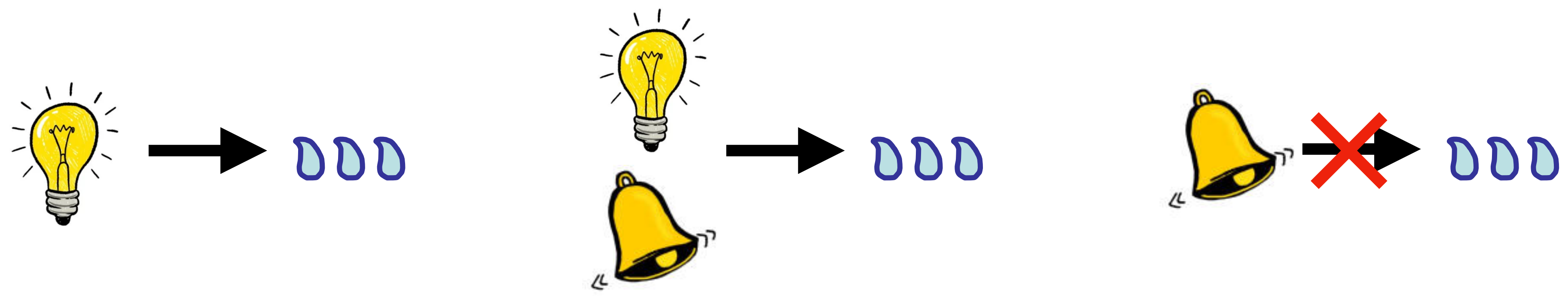
$$\pi(a, s) = P(a_t = a \mid s_t = s)$$

Recap: “Three” historical branches of RL

- Association learning, prediction (early 1900s)
- Optimal control (1950 onward)
- Learning *and* control (1980 onward)

History: Learning to predict reward

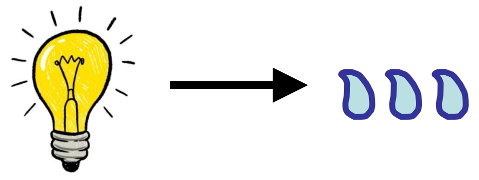


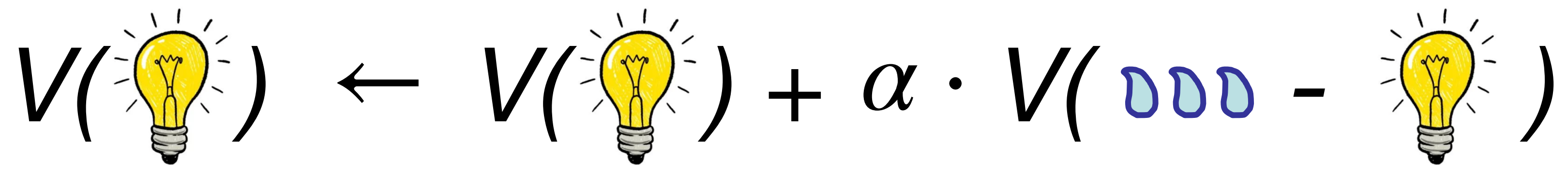
- **Classical** (Pavlovian) **conditioning** (roughly) in domain of algorithms for **prediction**
    - Algorithms for **control: instrumental** (operant) **conditioning**
- At least two interesting phenomena in classical conditioning from algorithmic perspective:
    - **Higher-order conditioning**  Temporal Difference (TD) Learning
    - **Blocking**  Rescorla-Wagner Learning

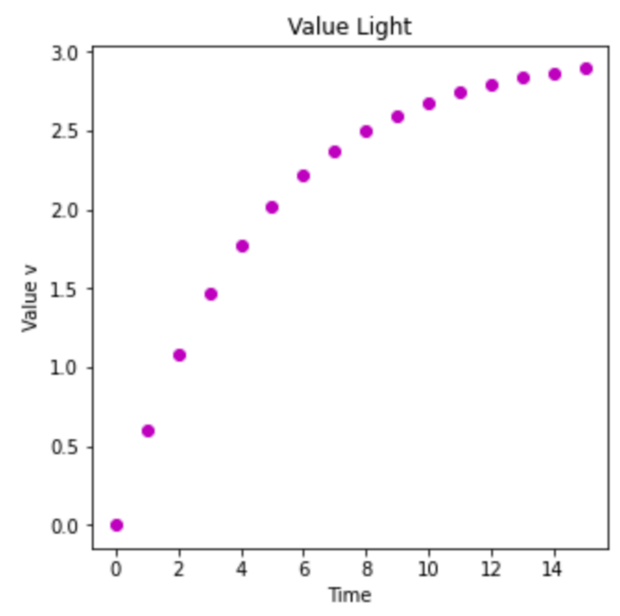
Basics of Learning: Blocking and Rescorla-Wagner Learning

Learn associative strength between a CS and US

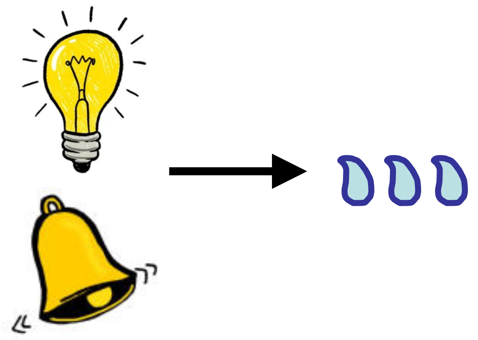
$$V(s) \leftarrow V(s) + \alpha \cdot (r - V(s))$$

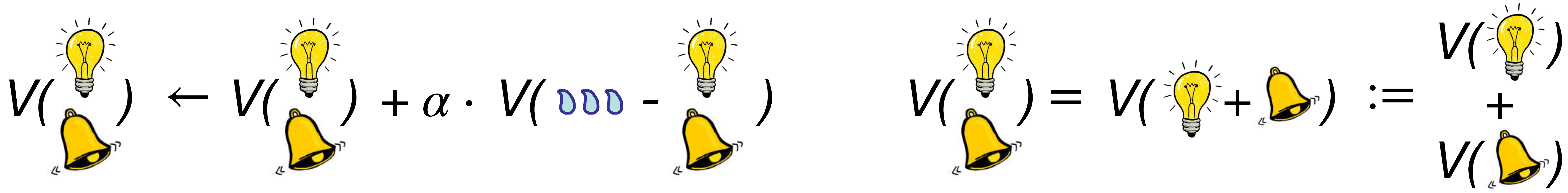


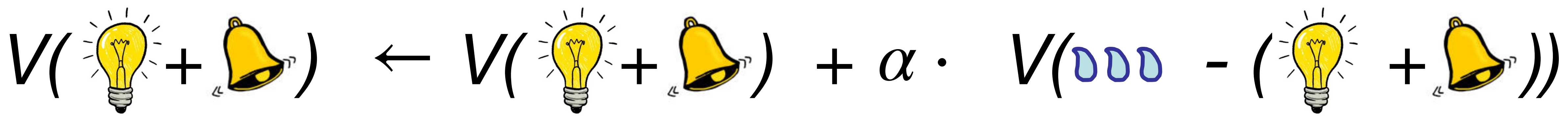




Introduce a second CS:







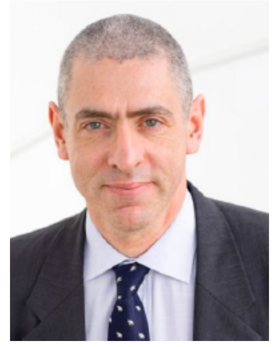
What does the value of the sound CS look like at different stages of learning?

Coding: Python, Google Collab

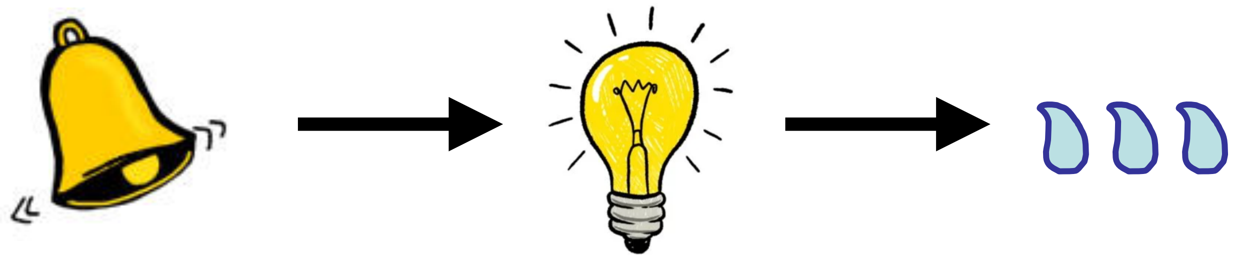
https://github.com/schwartenbeckph/RL-Course/tree/main/2022_05_03

History: Learning *and* Control


- **TD learning**, Actor-critic architecture (Sutton & Barto, 1981, 1982)
- Chris Watkins 1989 (+ Peter Dayan 1992): integrate dynamic programming with online learning

  - **Q learning**
- Key idea: use *experience* and *own value estimates*!
  - One example: secondary reinforcement



Basics of Learning: Higher order conditioning and TD learning (next time)

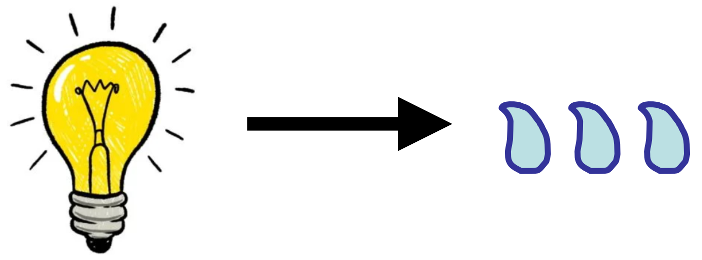
- Extends Rescorla–Wagner model
  - Address how within-trial *and* between-trial timing relationships among stimuli influence learning
  - How can higher-order conditioning arise?
- Real-time
  - t labels time steps within *or* between trials

$$V(s_t) \leftarrow V(s_t) + \alpha \cdot (r + \gamma \cdot V(s_{t+1}) - V(s_t))$$

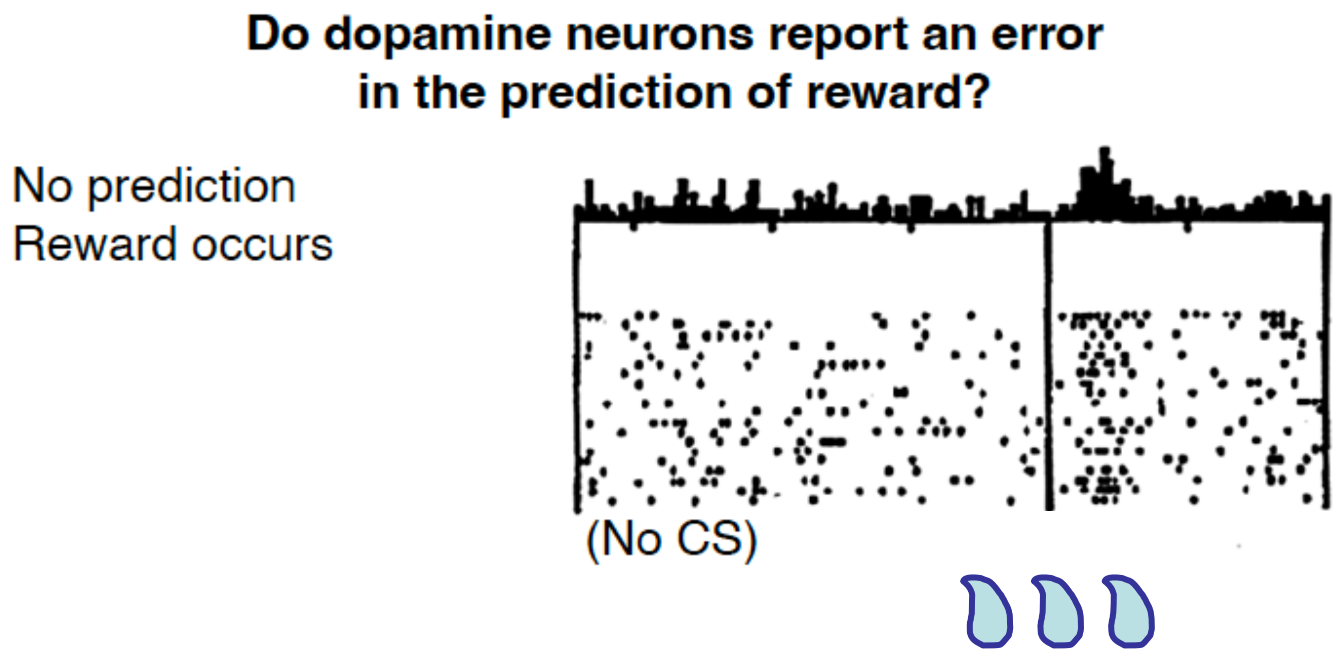
  - Think of time between t and t+1 as a small time interval, say .01 second
- Solves:
  - Higher order conditioning
  - NO blocking if CS_2 is moved before previously learnt CS_1
  - A lot of other things..

RL success story: Dopamine (a primer)

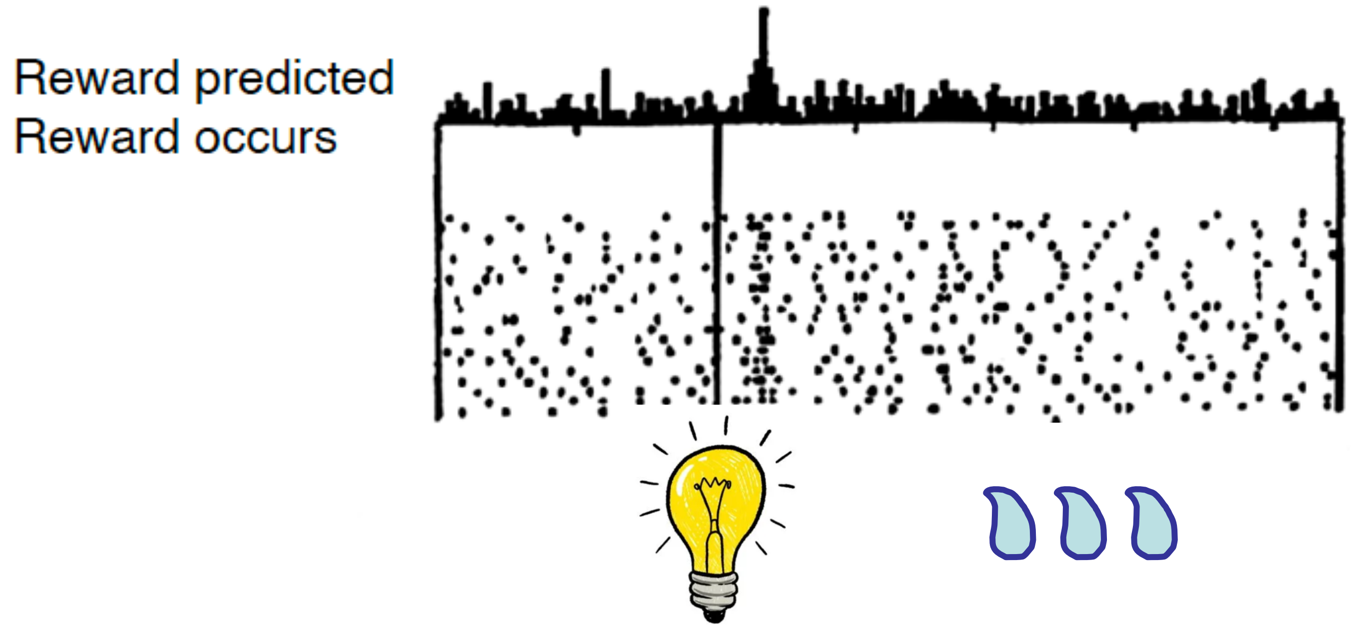
Can RL tell us anything about the brain?

- Yes, quite a lot.
- Particularly, it looks like dopamine (DA) is a key neurotransmitter for reward learning
    - Schultz, Dayan & Montague (1997): 

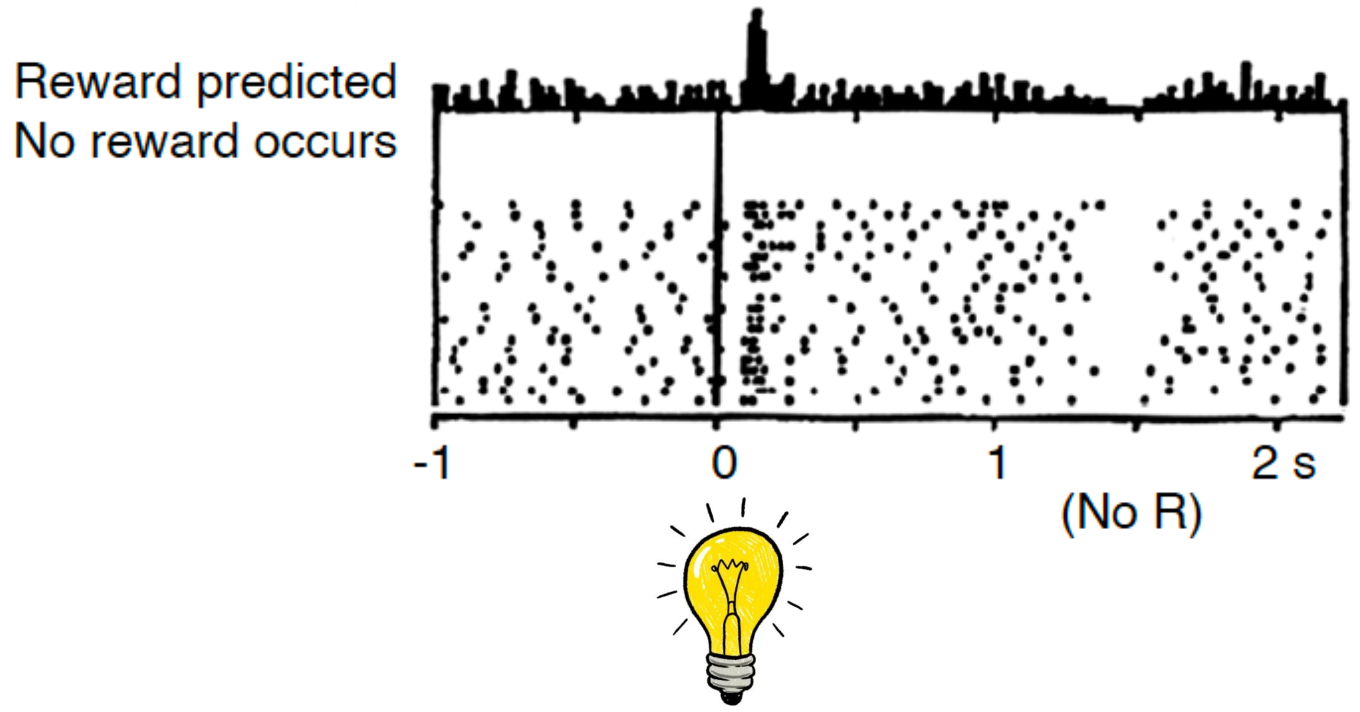
Dopamine neurons signal immediate reward



BUT: after training…
- DA signal reward prediction
- But not correctly predicted reward!



AND: it signals the unexpected omission of a reward!



This provides strong evidence that DA signals a **reward prediction error**

(Note: it is $r + V_{t+1} - V_t$ rather than $r - V_t$ though..)