

# **An introduction to Reinforcement Learning**

**Philipp Schartenbeck**

**AI Center, University of Tübingen**

# What is reinforcement learning (RL)?

Also, why are slides in English?

- RL is a **computational approach** to learning from **interactions** with the **environment**
  - Trial-and-error
  - Delayed reward
- Considers whole problem of **goal-directed** agent interacting with an **uncertain** environment
- RL agents
  - Have explicit goals
  - Sense aspects of their environments
  - Choose actions to influence their environments
- Very general

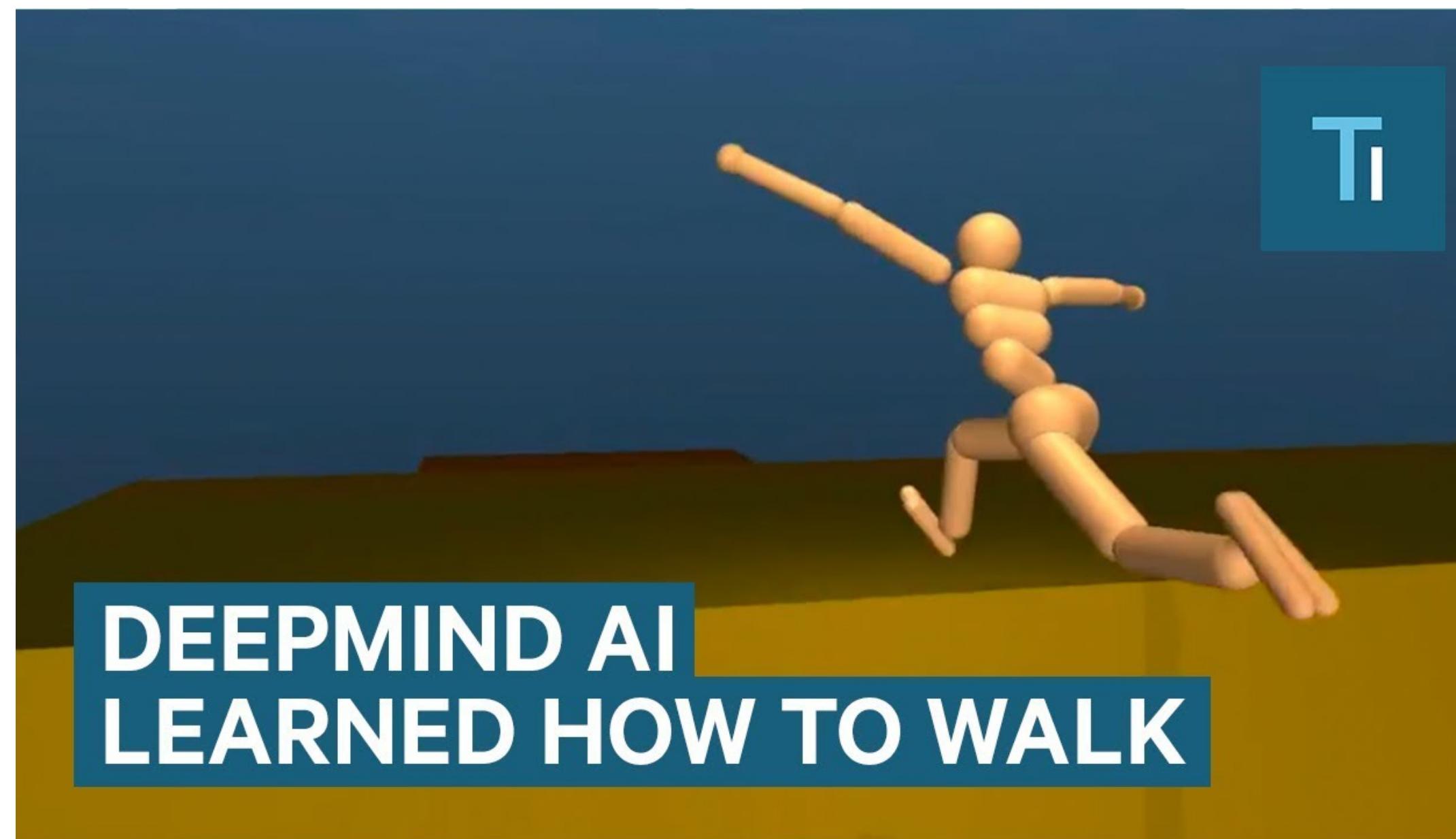
# What is reinforcement learning (RL)?

A few hours (+a bit of evolution) after birth:



# What is reinforcement learning (RL)?

This process is perhaps not too different from AI learning to walk:



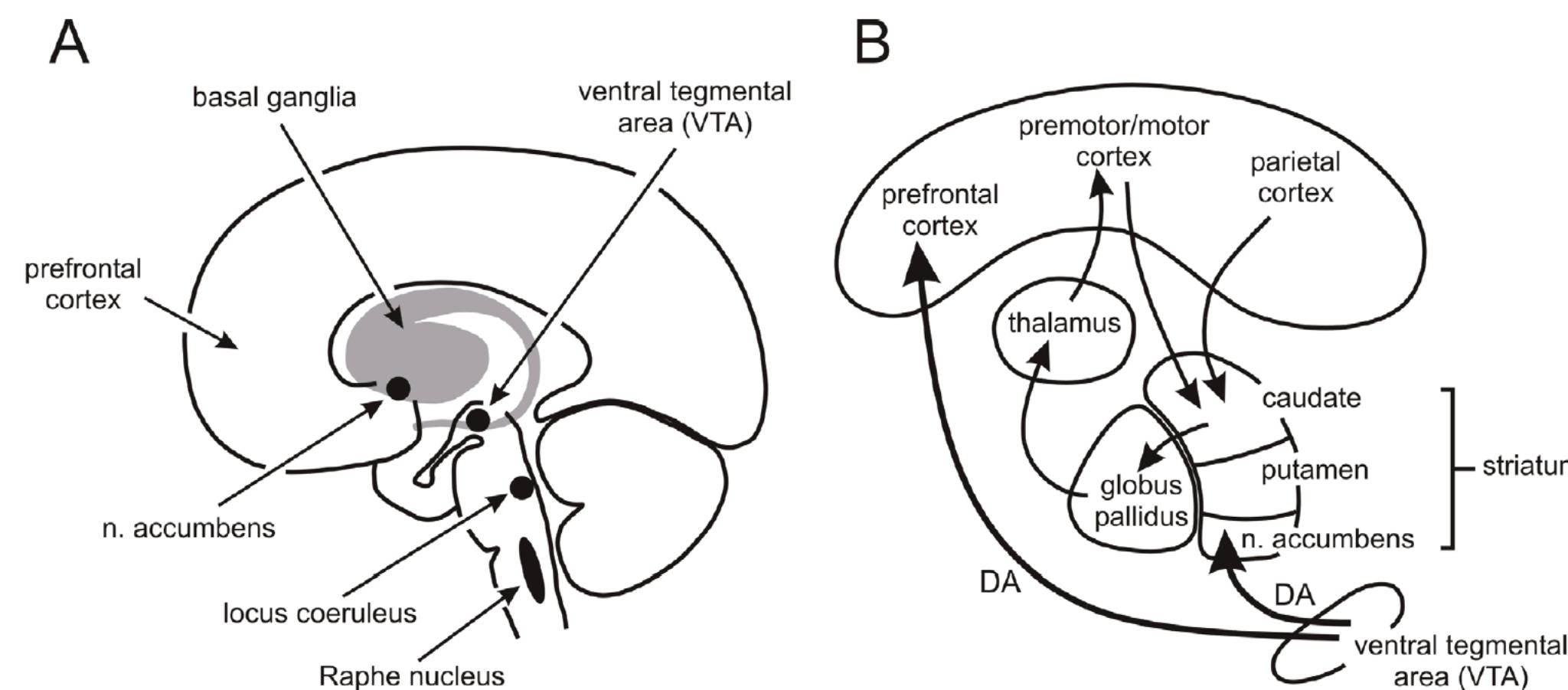
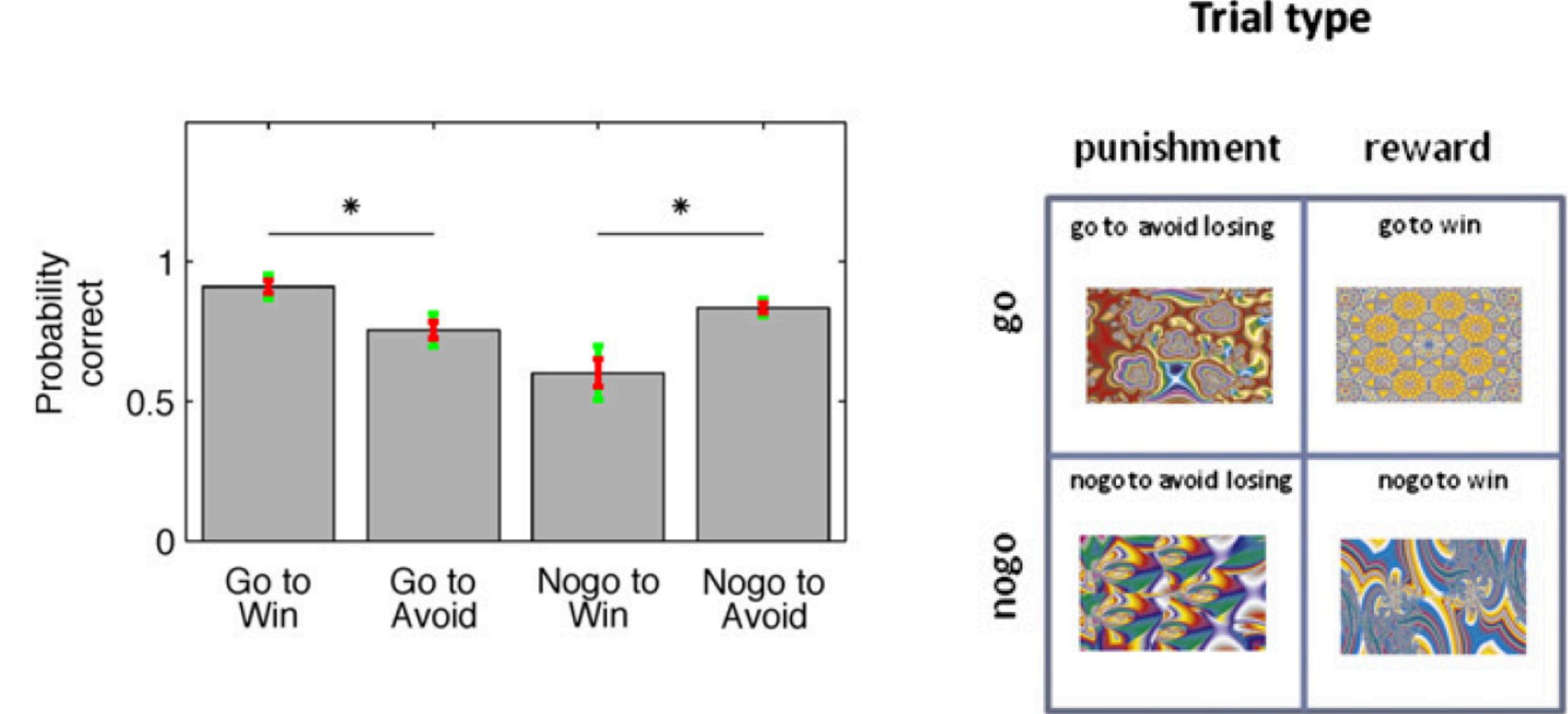
# What is reinforcement learning (RL)?

Learn useful actions:



# What is reinforcement learning (RL)?

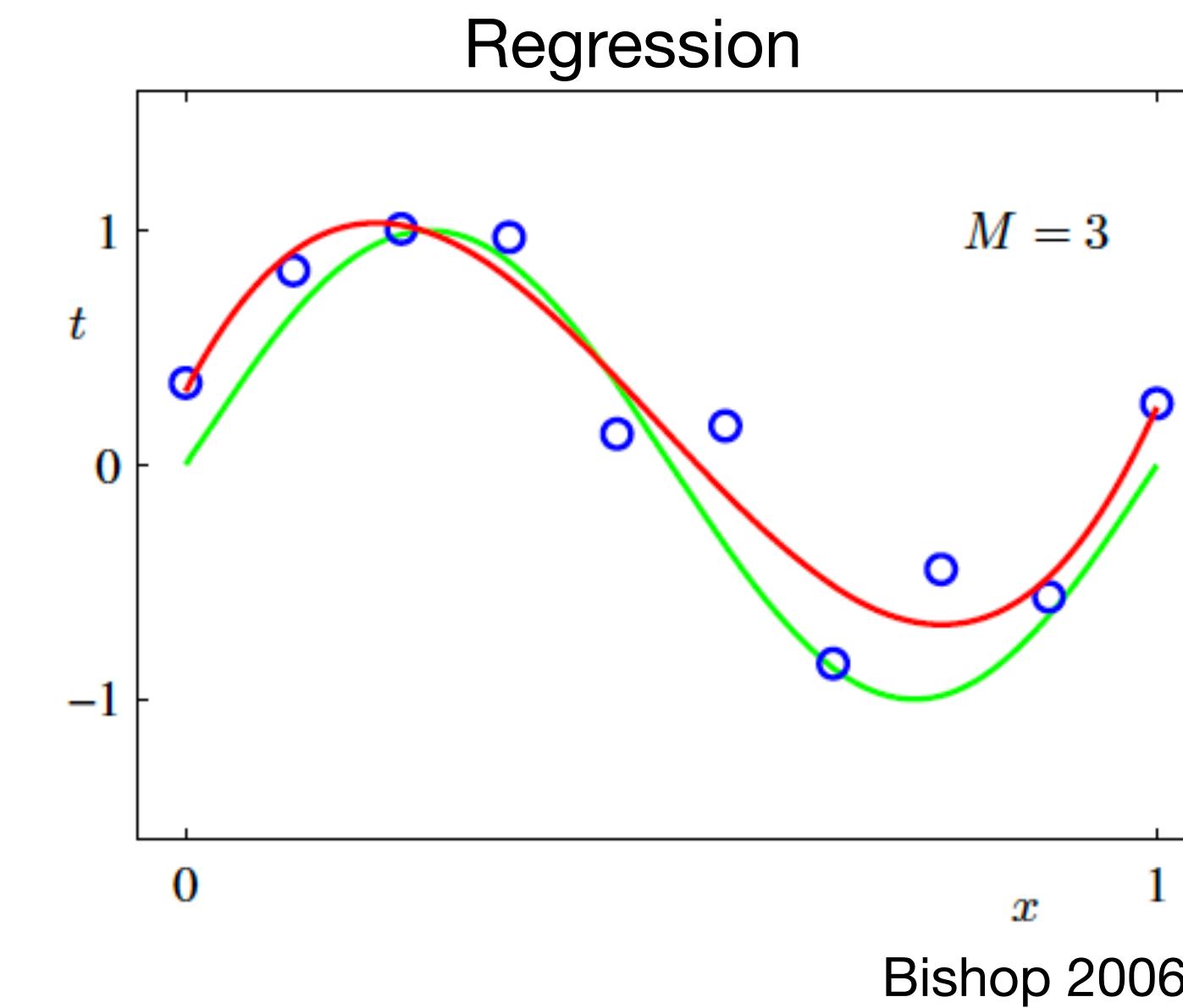
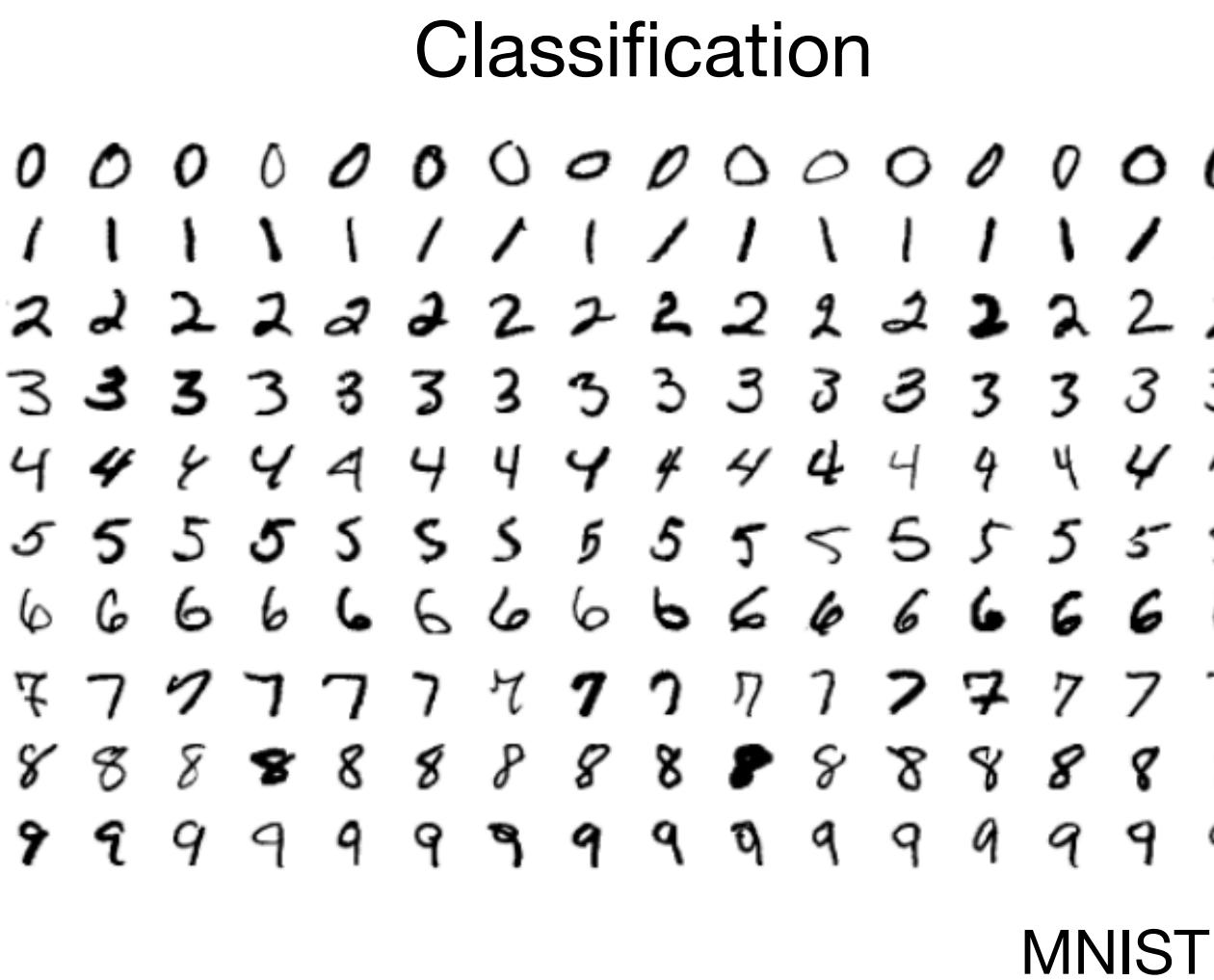
RL has been tremendously successful at explaining behaviour and psychological variables  
(More next session)



RL has been tremendously successful at explaining neuroscience  
(More in session after next)

# Types of (machine) learning: supervised learning

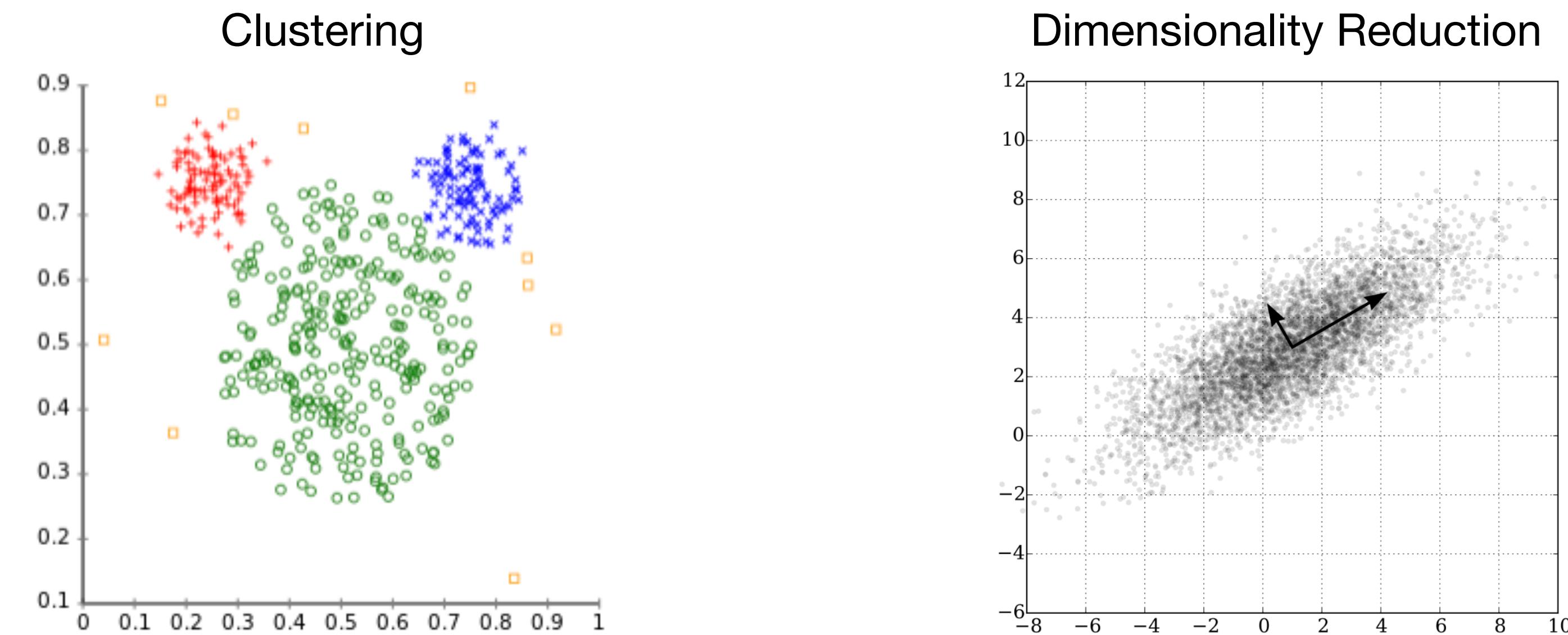
- Find correct labelling/prediction of data:



- That's not what we want though:
  - Want to learn from own *experience* by *interacting* with the world

# Types of (machine) learning: unsupervised learning

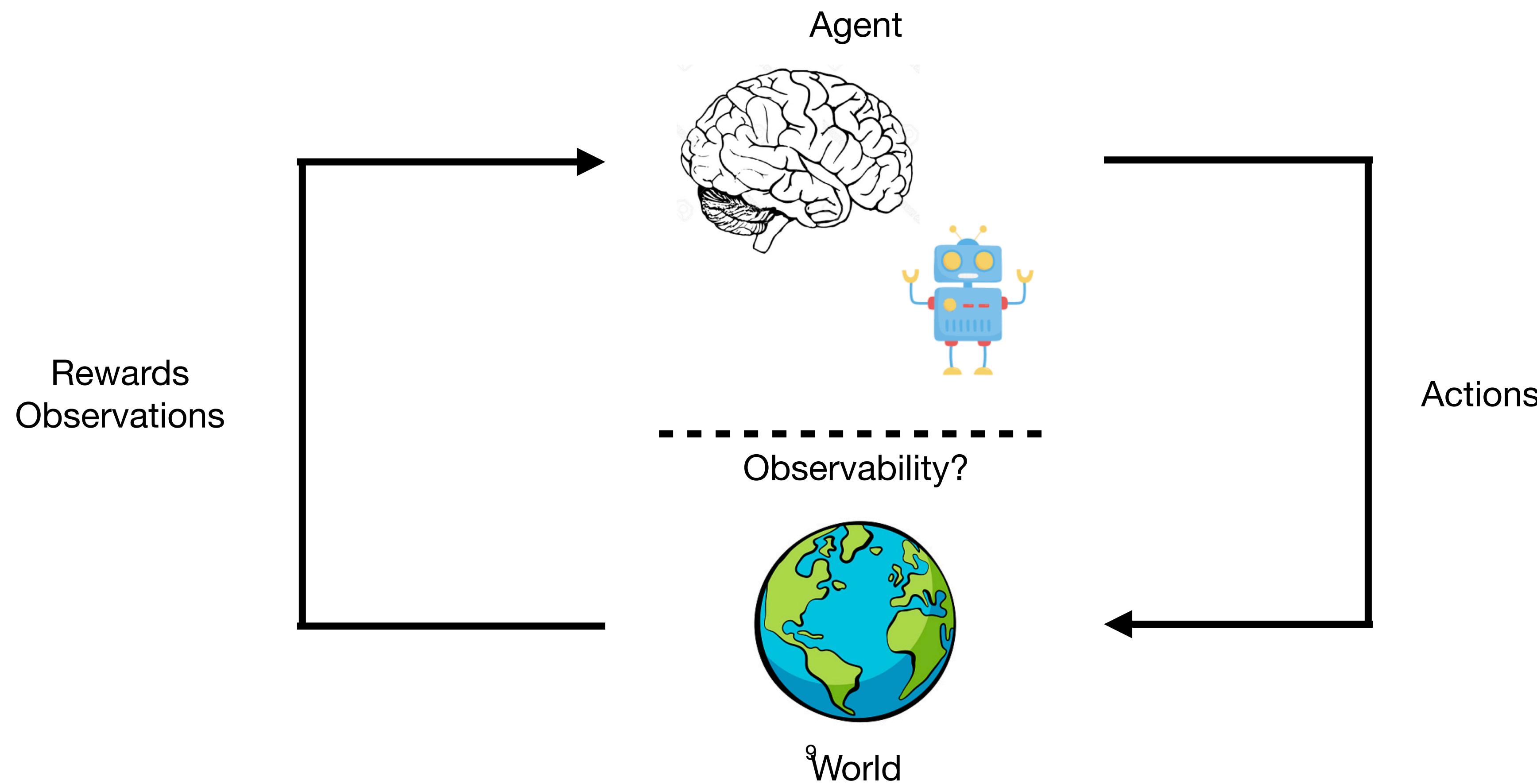
- Find structure in data:



- That's also not what we want:
  - Don't (necessarily) want to learn hidden structure, rather: maximise reward

# Types of (machine) learning: Reinforcement learning

- In RL, we want to learn good actions from interactions with the world



# RL as a Marrian system?

- RL has the ambition to provide a complete account of agency
  - There is much debate about that (e.g. Sliver et al. 2021: “Reward is enough”)
- Can RL account for Marr’s levels?
  - The **problem**: optimal prediction of future reward
  - The **algorithm**: temporal difference learning, Q-learning, model-based RL, ...
  - Neural **implementation**: Basal ganglia, dopaminergic system, replay, ...

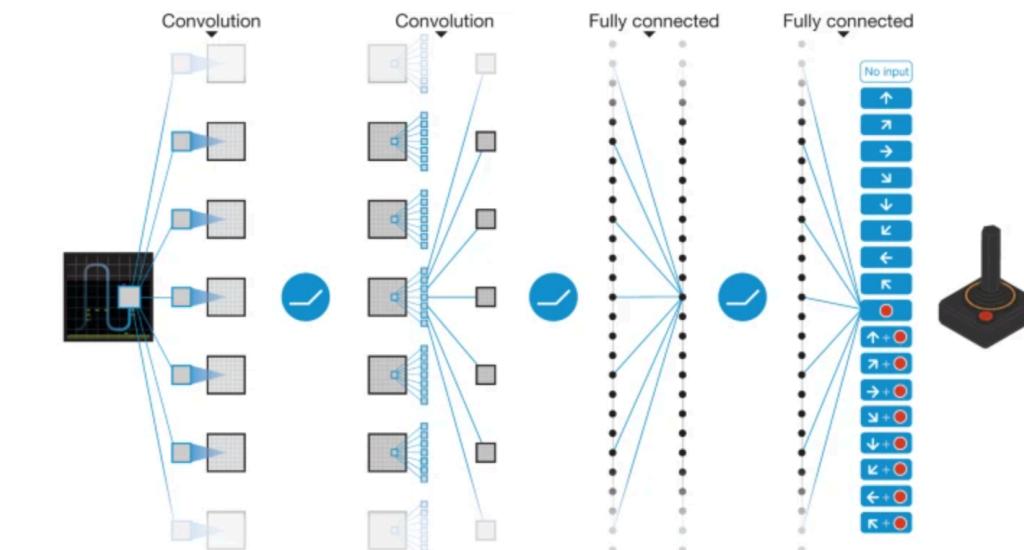
# Why should we know about RL?

- Lots of reasons
- Powerful framework to understand individual differences in behaviour - individual parameters?
  - Differences in action selection (randomness, heuristics)
  - Mechanisms to find the value of states
  - Mechanisms of learning

- Exploit vs. Explore



- Interesting modern applications based on ('deep') RL



Mnih et al. 2015

# **Course Structure**

# This seminar: components

- Most of this is first time material - tell me if something doesn't work
  - Should be fun and open for suggestions
- Theory (key reference: [Sutton & Barto, 1998](#))
- Research (key papers)
- Coding (Python)

# Dates and topics

Options

- Just cancel
- Coding exercise
- Two sessions (coding) on 17th or 31st

19.04.2022	
26.04.2022	
03.05.2022	
10.05.2022	
17.05.2022	
Probably not around	24.05.2022
	31.05.2022
	Holiday
14.06.2022	
21.06.2022	
28.06.2022	
05.07.2022	
12.07.2022	
19.07.2022	
26.07.2022	

Basics,  
theory

Applications,  
other aspects

# Dates and topics

## Topics (some flexibility)

• Intro RL, Python	19.04.2022	Basics, theory
• Basics of learning theories, psychology	26.04.2022	
• Learning about different options, neuroscience • Primer on Temporal Difference (TD) learning	03.05.2022	
• Markov Decision Processes	10.05.2022	
• Basics of control • Dynamic Programming, TD learning • Action selection	17.05.2022	
• Other important aspects • Model-based vs. Model-free • Exploration vs. Exploitation	24.05.2022	
	31.05.2022	

# Dates and topics

## Topics (some flexibility)

• Some coding	14.06.2022	
• Role of different parameters	21.06.2022	
• Model-fitting	28.06.2022	
• If possible: parameter recovery, model comparison	05.07.2022	
• ‘Advanced’ topics and current applications	12.07.2022	
• Planning, Dyna, replay	19.07.2022	
• Clever ways of planning, tree-search etc	26.07.2022	
• Deep RL		
• Future directions, limitations, current research		

Applications,  
important aspects

# Key resources

- Sutton and Barto 1998 Reinforcement Learning: An Introduction
- My GitHub
- Other resources
  - David Silver's course at UCL
  - Other great courses on RL in Tuebingen with slightly different focus, e.g. by Georg Martius

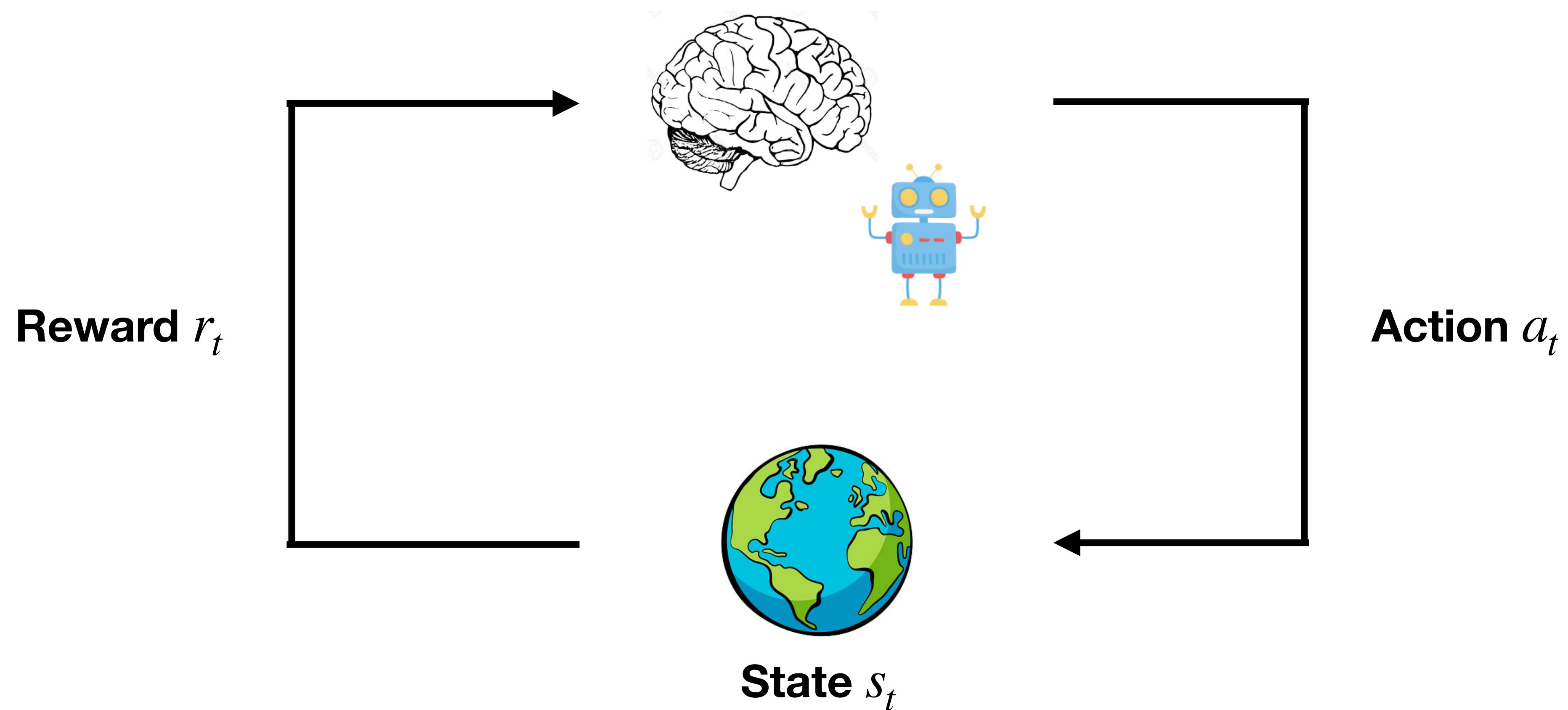
# Evaluation

- Some flexibility
- Essay at end of the course
  - e.g. modelling of a simple task
  - Review on specific application, topic
- Additional possibilities
  - Smaller coding exercises
  - Presentation

**What are you most interested in? Any other Ideas?**

# **Intro to intro to RL**

# Basic setup



# Basic setup: how to agents learn to act?

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

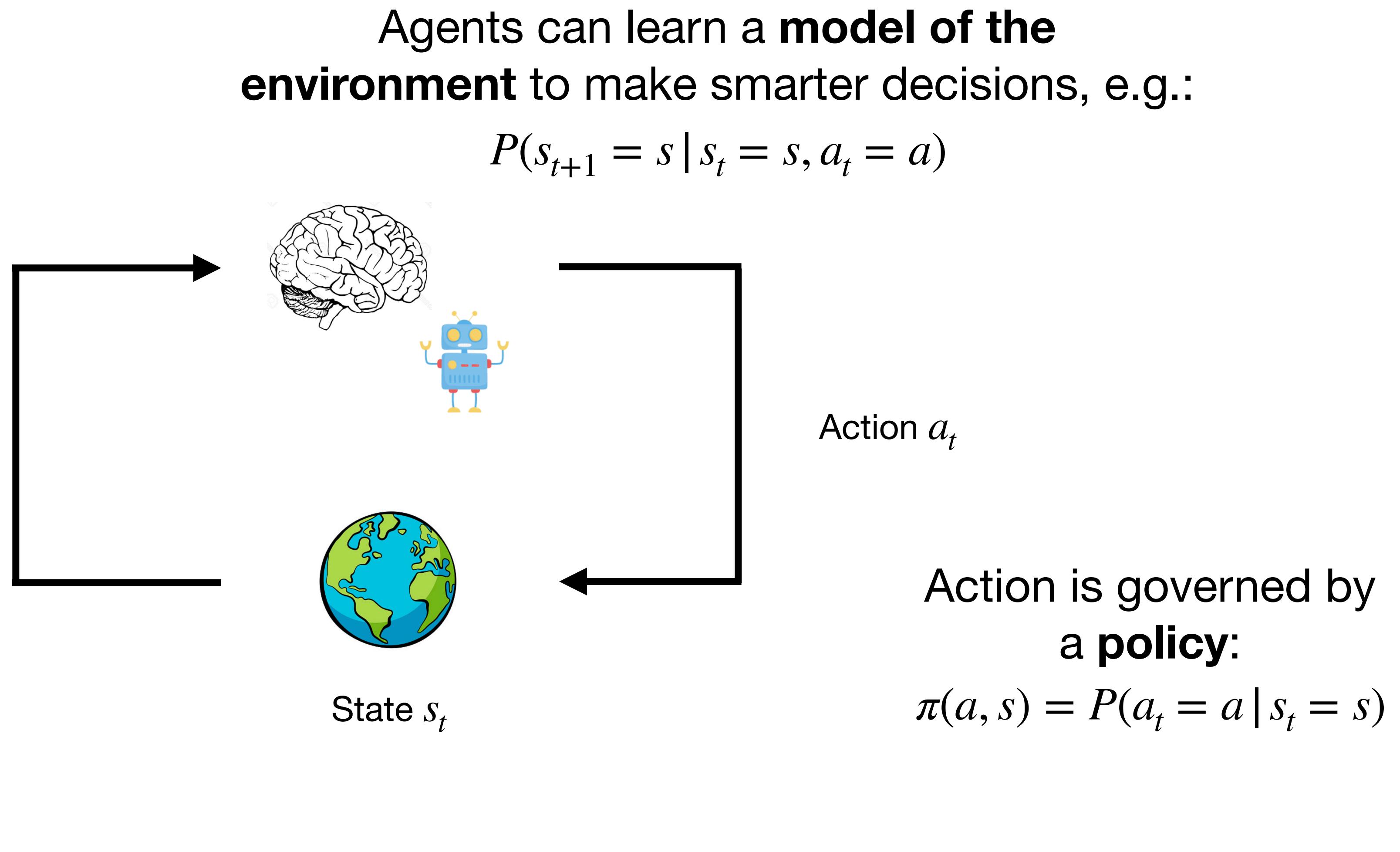
$$V_{\pi}(s) = \mathbb{E}_{\pi}[R | s_0 = s]$$

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

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

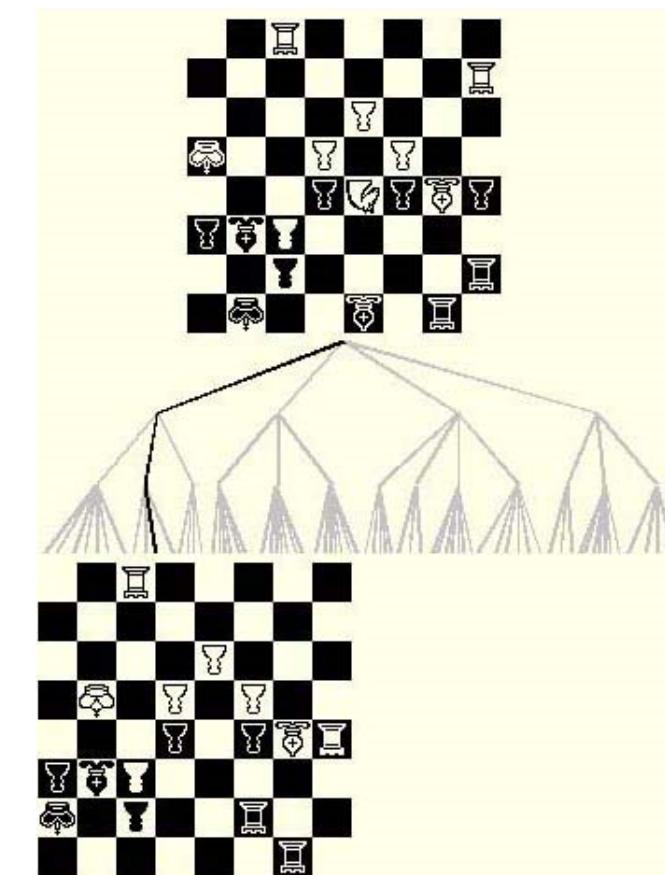
Learning rate

Prediction error



# More Examples

- **Chess:** what is...
  - The state?
  - An action?
  - A reward?
- How can values be learned over time?
- How could a model of the environment be useful?



Other relevant components:

- tree search
- position evaluation
- situation memory

*Taken from Peter Dayan*

# Examples extended..

- **Learn how to walk:** what is...
  - The state?
  - An action?
  - A reward?
- How can values be learned over time?
- How could a model of the environment be useful?



# Examples extended..

- Other examples (see Sutton & Barto, pp 4-5):
- Adaptive controller adjusts parameters of a petroleum refinery's operation in real time
  - Optimise yield/cost/quality trade-off
  - Objective: specified marginal costs
  - Without sticking strictly to pre-defined set points
- Mobile robot decides to search for trash to collect or find its way back to battery recharging station
  - Decision based on current charge level of battery and how quickly recharger has been found in the past.
- Prepare breakfast
  - Subgoals, hierarchies
  - Conditional behaviour
  - Sense/access bodily states

# Key features of all these examples

- (Danger of repeating myself): **Interaction** between active decision-making agent and its environment
  - Agent seeks to achieve a **goal**
  - **Uncertainty** about its environment
- Take into account **indirect, delayed consequences** of actions
  - Requires foresight or planning
- Need to **monitor** environment frequently
- **Judge progress** toward goal based on what can be sensed directly
- Use experience to **improve performance** over time (online vs. offline learning)
  - Basis for adjusting behaviour to exploit specific features of the task

# Key Elements (more details later): Policy

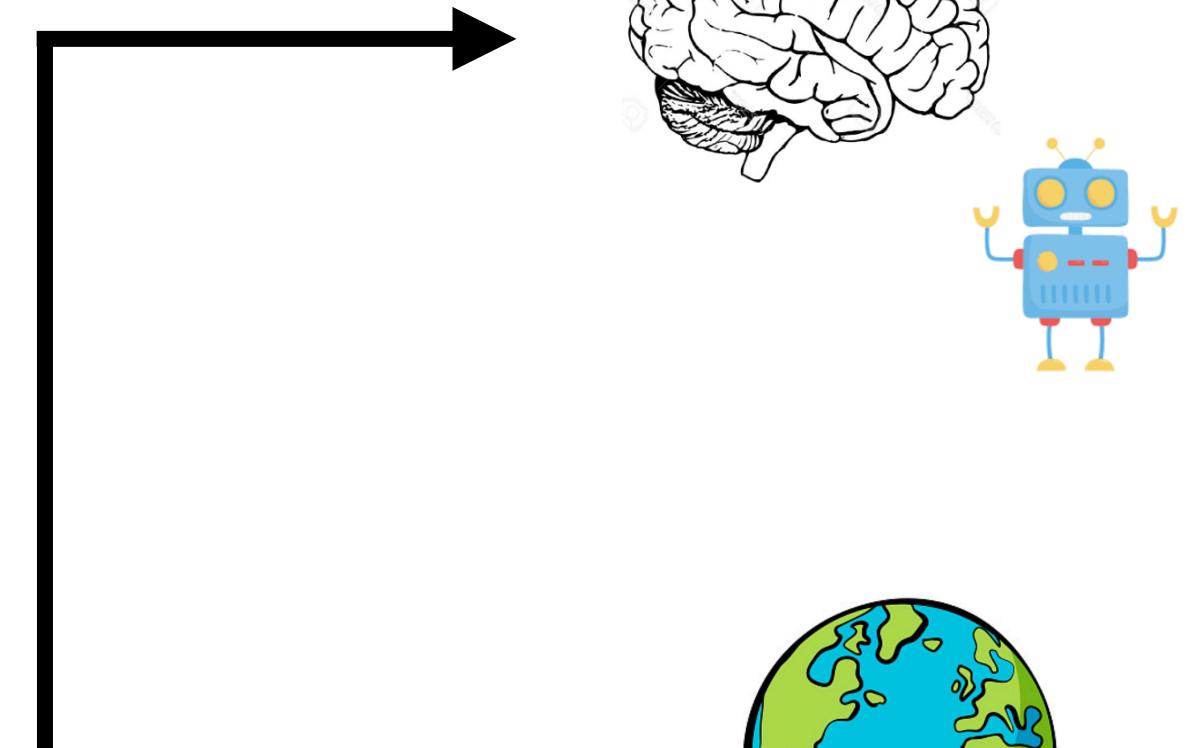
- Defines agent's **way of behaving** at a given time
  - Sufficiently determines behaviour
  - Often stochastic (determine probabilities for each action)
- Mapping from perceived states of the environment to actions to be taken when in those states
  - Cf., stimulus–response rules or associations
- Can be simple or difficult
  - Lookup table vs. extensive search

Action is governed by  
a **policy**:

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

# Key Elements (more details later): Reward signal

- = **goal** of a reinforcement learning problem
  - Single number from environment on each time-step
- SOLE objective is to maximise the total reward over the long run
- Primary basis for altering the policy
  - If action is followed by low reward, then the policy may be changed
- Often: stochastic function of state of environment and actions taken



# Key Elements (more details later): Value function

- = what is good in the **long run**

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

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

- Total amount of reward an agent can expect to accumulate over the future, starting from a given state
  - Long-term desirability of states
  - Taking into account states that are likely to follow and rewards available in those states
- A state might yield a low immediate reward but can still have a high value (because regularly followed by other states that yield high rewards)

# Reward vs. Value

- Rewards = primary, values = secondary
  - Rewards are the basis for estimating value
- BUT action selection is based on value, not immediate reward - why?
- It is more difficult to determine value than reward
  - Methods for value estimation are a central problem in RL

Reward  $r_t$

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

# Key Elements (more details later): Model of Environment

- Mimics **behaviour of environment**

- Allows to predict how the environment will behave

- E.g.: given a state and action, model can predict the resultant next state and next reward

- Models are used for planning

- Considering possible future situations before they are actually experienced

- Methods for solving reinforcement learning problems that use models and planning are called model-based methods

- Simpler model-free methods are explicitly trial-and-error learners (opposite of planning)

# ‘Multiple Systems’ in RL

- **Model-based RL**
  - Build a forward model of the task and outcomes
  - Search in the forward model
    - Optimal use of information
    - Computationally ruinous
- **Model-free RL**
  - learn values, which summarise future worth
    - computationally trivial
    - bootstrap-based; so statistically inefficient
- learn both – select according to uncertainty

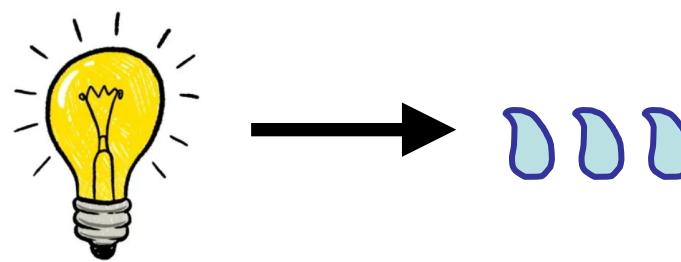
# What are the limits of RL?

- How do we define a state? Are all states perceivable?
- What about problems that cannot be solved via learning (e.g. inference)?
- Is reward to explain behaviour/cognition/brains?

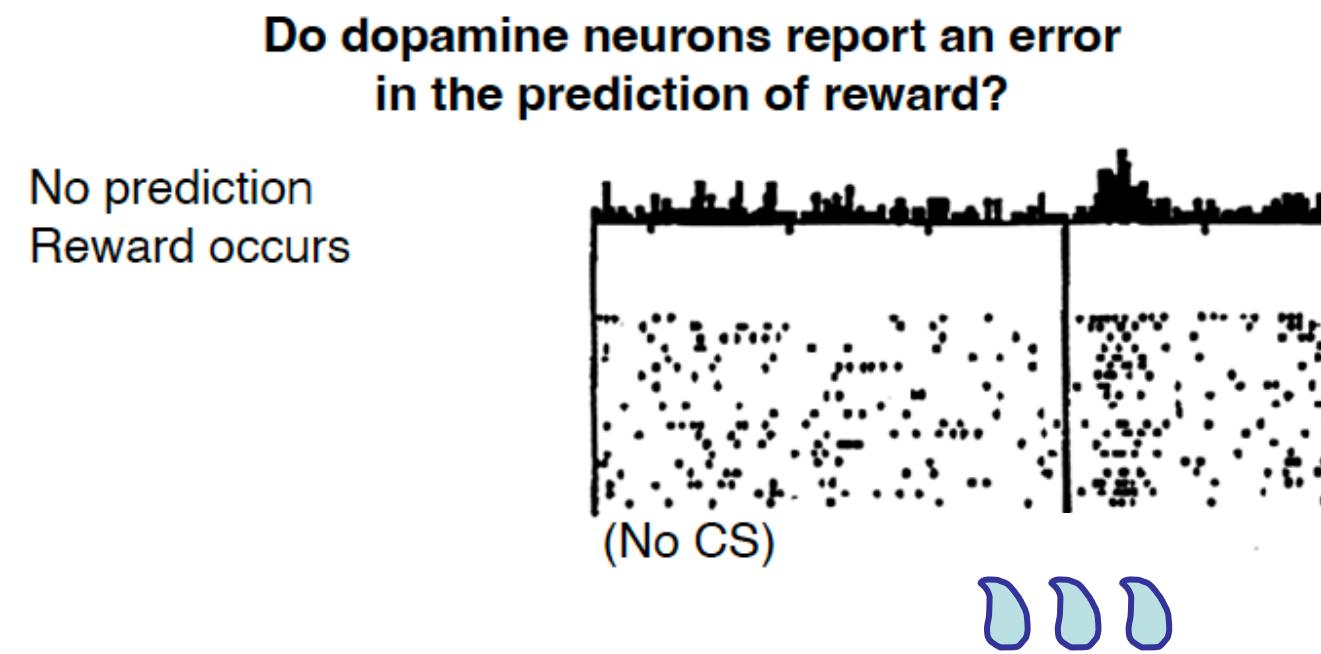
# **RL success story: Dopamine (a primer)**

# Can RL tell us anything about the brain?

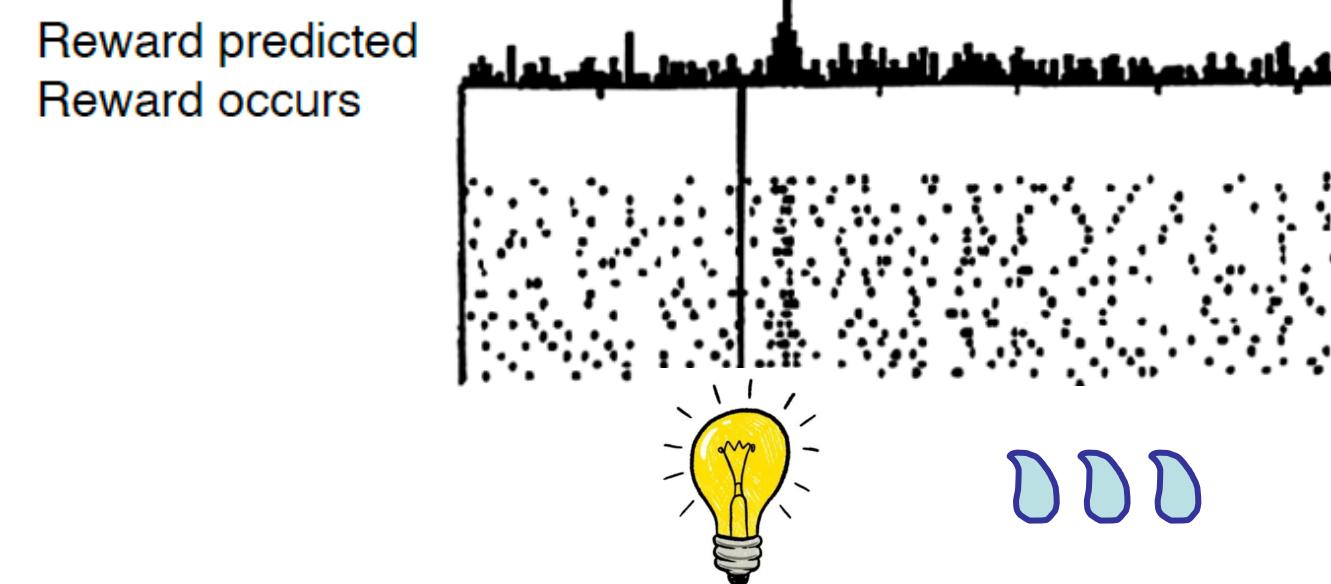
- Yes, quite a lot.
- Particularly, it looks like dopamine (DA) is a key neurotransmitter for reward learning
  - Schultz, Dayan & Montague (1997):



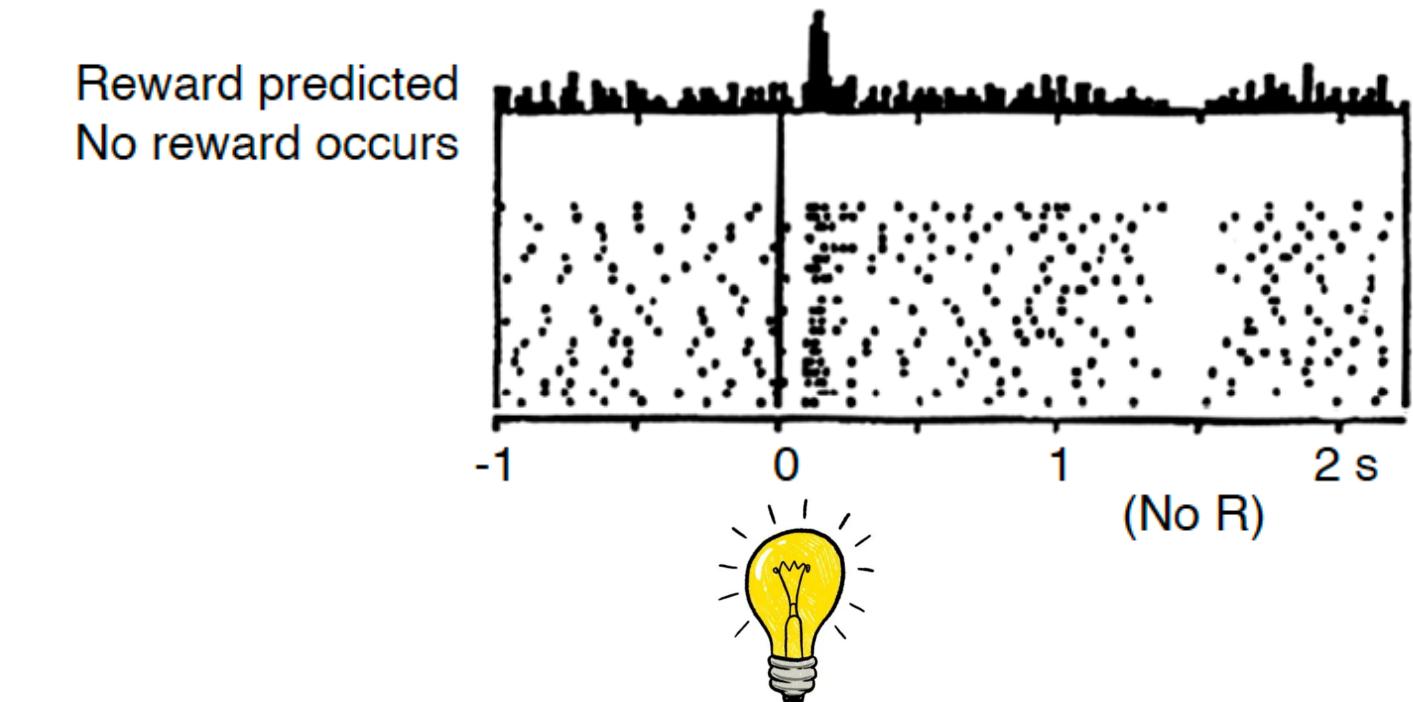
Dopamine neurons signal immediate reward



- 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..)

# Coding: Python, Google Collab

[https://github.com/schwartenbeckph/RL-Course/tree/main/2022\\_04\\_19](https://github.com/schwartenbeckph/RL-Course/tree/main/2022_04_19)