Artificial Neural Networks and RL : Lecture 13

Wulfram Gerstner
EPFL, Lausanne, Switzerland

from brain-computing to neuromorphic computing

Objectives for today:

- three-factor learning rules can be implemented by the brain
- eligibility traces link correlations with delayed reward
- the dopamine signal has signature of the TD error
- local learning rules: 2-factor and three-factor

(- local learning rules in hardware/shifted to next week)

Reading for this week:

Sutton and Barto, Reinforcement Learning (MIT Press, 2nd edition 2018, also online)

Chapter: 15

Background reading:

(1) Fremaux, Sprekeler, Gerstner (2013) Reinforcement learning using a continuous-time actor-critic framework with spiking neurons *PLOS Computational Biol. doi:10.1371/journal.pcbi.1003024*

(2) *Gerstner et al. (2018)* Eligibility traces and plasticiy on behavioral time scales: experimental support for neoHebbian three-factor learning rules, *Frontiers in neural circuits* https://doi.org/10.3389/fncir.2018.00053

(3) Wolfram Schultz et al., (1997) A neural substrate of prediction and reward, SCIENCE, https://www.science.org/doi/full/10.1126/science.275.5306.1593

(4) Bert Offrein et al., 2020, Prospects for photonic implementations of neuromorphic devices and systems, *IEEE Xplore, https://ieeexplore.ieee.org/abstract/document/9371915*

Review: Biological Motivation of RL

Reinforcement Learning (RL)
→ **Learning by reward**

Field has two roots:
→ Optimziation/Markov
Decision Model
→ Biology

Questions for today?
- What elements of RL are ‘bio-plausible’?
- What elements can go to
neuromorphic hardware?

Previous slide.
one of the major drives of RL has been insights from biology and cognitive science: animals and humans are able to learn from rewards.

The question then is:

1. can we make the relation to biology more precise?

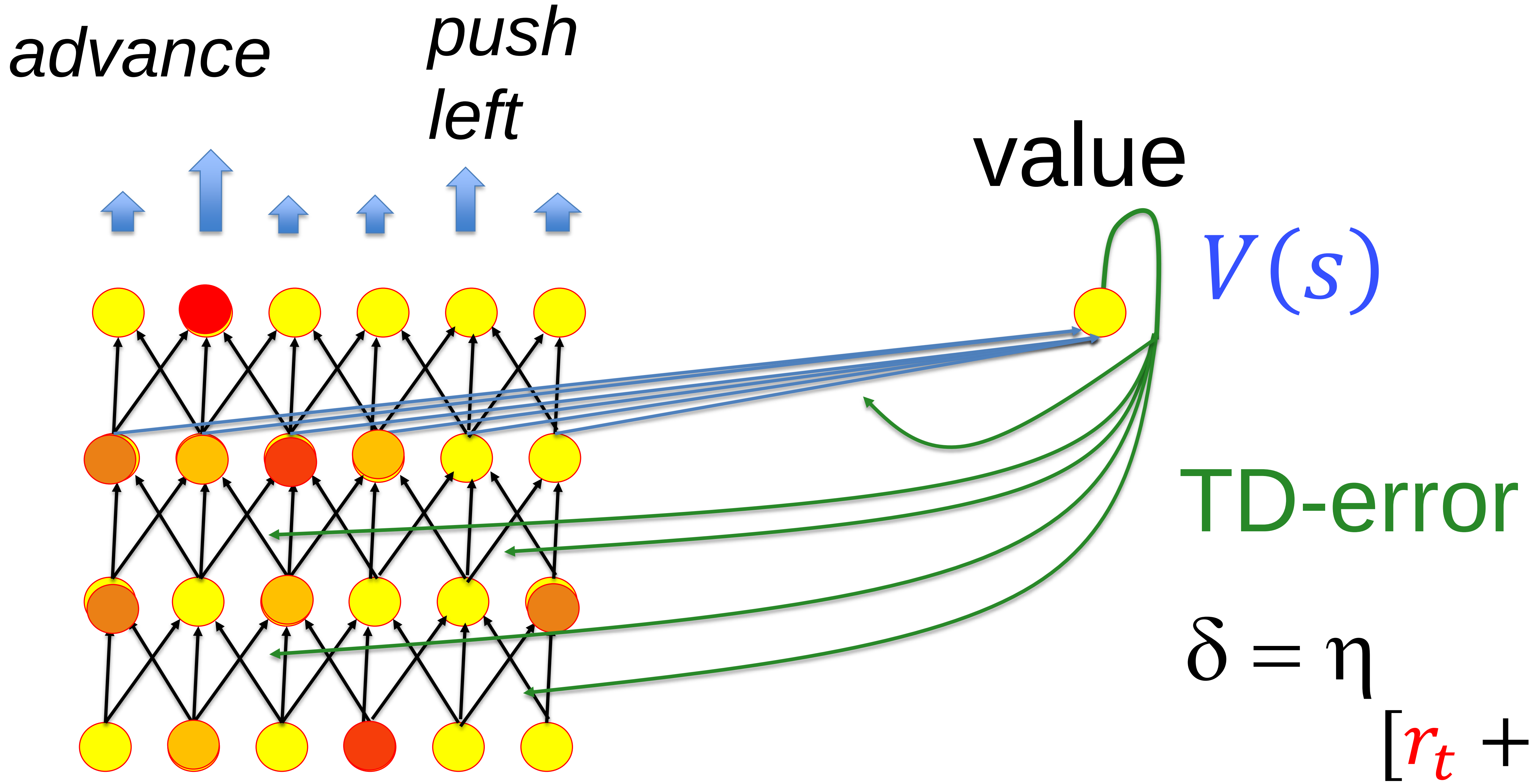
2. Can we exploit biological insights for unconvential computer hardware?

To answer these questions let us focus on the ‘Learning Rule’.

Review: Advantage Actor-Critic = ‘REINFORCE’ with TD signal

- Estimate $V(s)$
- learn via TD error

actions



advance

push left

value

$V(s)$

TD-error

$$\delta = \eta \, [r_t + \gamma V(s') - V(s)]$$

The update of parameters depends on the TD error!

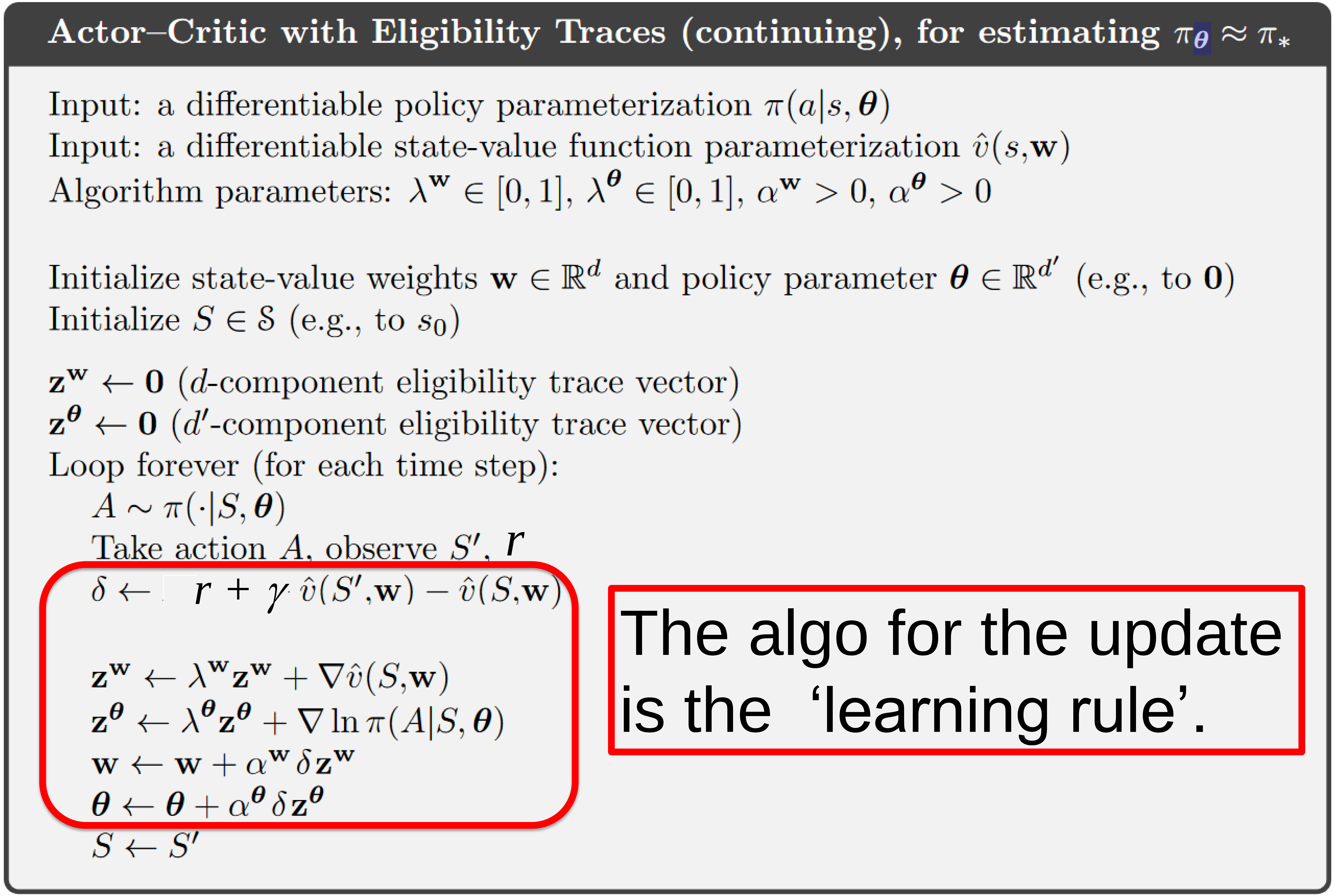
The algo for the update is called a ‘learning rule’.

Previous slide.
Let us focus on the ‘Learning Rule’ in the actor-critic setup.

There are weights w leading to the actor and other parameters θ leading to the critic.

Learning rule means that we analyze how these parameters change.

Review: Advantage Actor-Critic with Eligibility traces



Actor–Critic with Eligibility Traces (continuing), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

```
Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Algorithm parameters: $\lambda^{\mathbf{w}} \in [0,1]$, $\lambda^{\boldsymbol{\theta}} \in [0,1]$, $\alpha^{\mathbf{w}} > 0$, $\alpha^{\boldsymbol{\theta}} > 0$

Initialize state-value weights $\mathbf{w} \in \mathbb{R}^d$ and policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)
Initialize $S \in \mathcal{S}$ (e.g., to $s_0$)

$\mathbf{z}^{\mathbf{w}} \leftarrow \mathbf{0}$ ($d$-component eligibility trace vector)
$\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \mathbf{0}$ ($d'$-component eligibility trace vector)
Loop forever (for each time step):
    $A \sim \pi(\cdot|S,\boldsymbol{\theta})$
    Take action $A$, observe $S'$, $r$
    $\delta \leftarrow r + \gamma\,\hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$

    $\mathbf{z}^{\mathbf{w}} \leftarrow \lambda^{\mathbf{w}}\mathbf{z}^{\mathbf{w}} + \nabla\hat{v}(S,\mathbf{w})$
    $\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \lambda^{\boldsymbol{\theta}}\mathbf{z}^{\boldsymbol{\theta}} + \nabla\ln\pi(A|S,\boldsymbol{\theta})$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}}\,\delta\,\mathbf{z}^{\mathbf{w}}$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}}\,\delta\,\mathbf{z}^{\boldsymbol{\theta}}$
    $S \leftarrow S'$
```

The algo for the update is the ‘learning rule’.

Adapted from Sutton and Barto

Previous slide. Review from DeepRL1

Parameters in the advantage actor critic change proportional to

- The TD error delta
- The derivative of the log policy for the actor
- The derivative of the value function for the critic

In this version of the algo we also have eligibility traces. Set λ=0 to get a version without eligibility traces.

In the example on the next page eligibility traces are important.

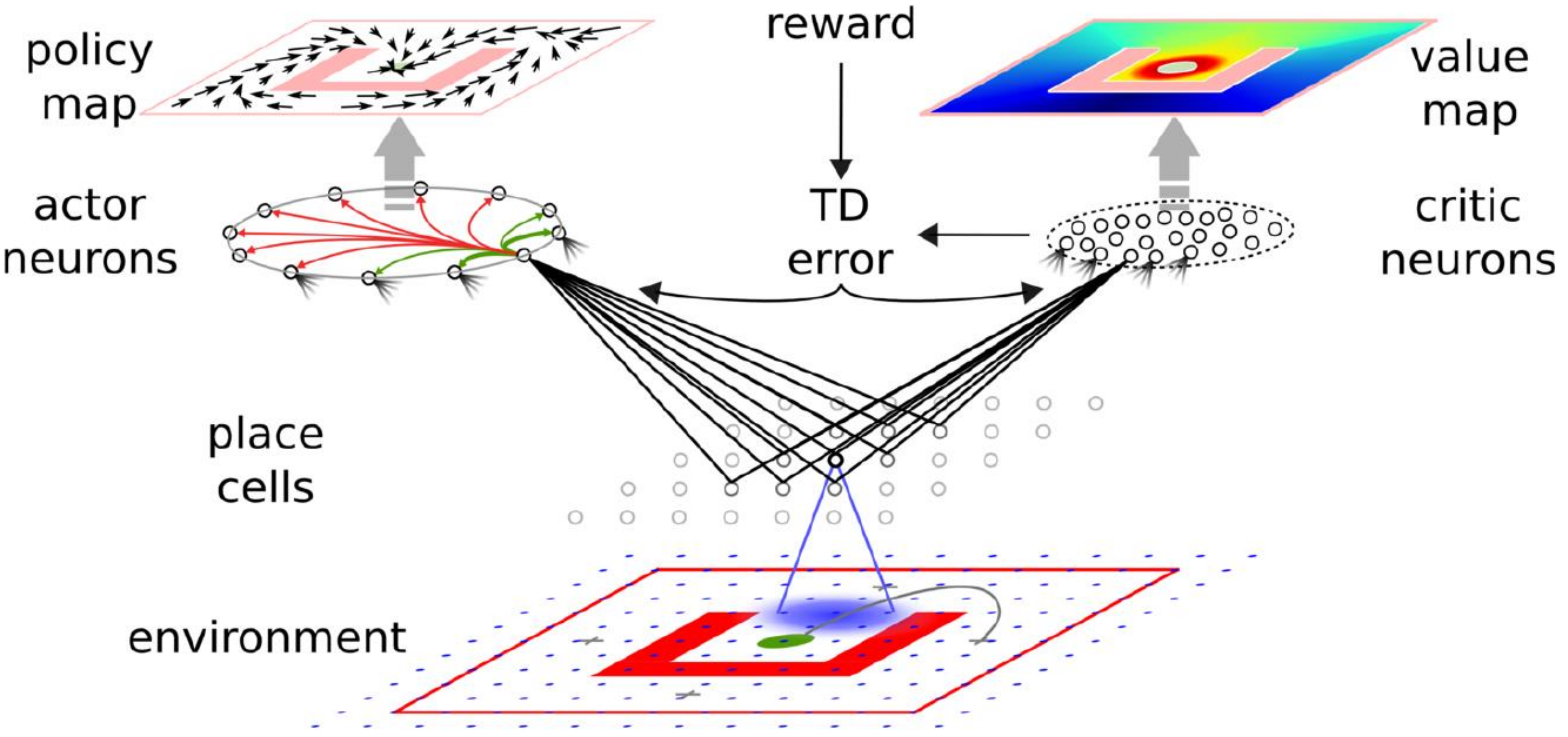
Review: Maze Navigation with Advantage Actor-Critic

Continuous action space:

ring of 360 action neurons

Value map:

Several identical neurons



Continuous state space:

Represented by Gaussian basis functions

Fremaux et al. (2013)

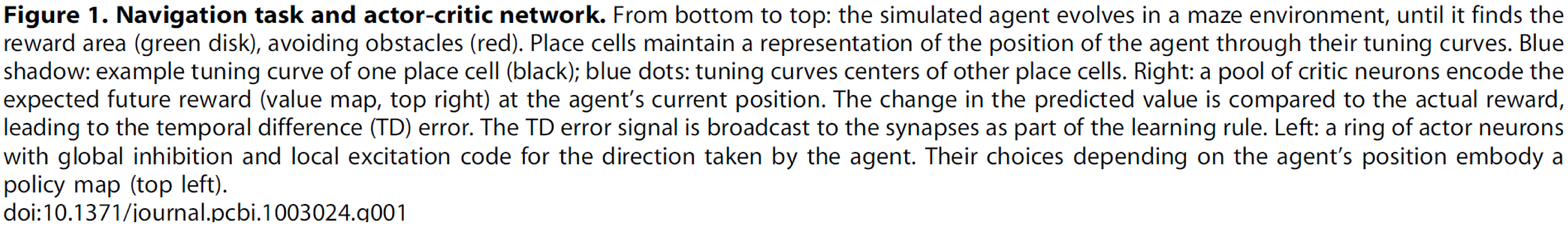


Figure 1. Navigation task and actor-critic network. From bottom to top: the simulated agent evolves in a maze environment, until it finds the reward area (green disk), avoiding obstacles (red). Place cells maintain a representation of the position of the agent through their tuning curves. Blue shadow: example tuning curve of one place cell (black); blue dots: tuning curves centers of other place cells. Right: a pool of critic neurons encode the expected future reward (value map, top right) at the agent's current position. The change in the predicted value is compared to the actual reward, leading to the temporal difference (TD) error. The TD error signal is broadcast to the synapses as part of the learning rule. Left: a ring of actor neurons with global inhibition and local excitation code for the direction taken by the agent. Their choices depending on the agent's position embody a policy map (top left).
doi:10.1371/journal.pcbi.1003024.g001

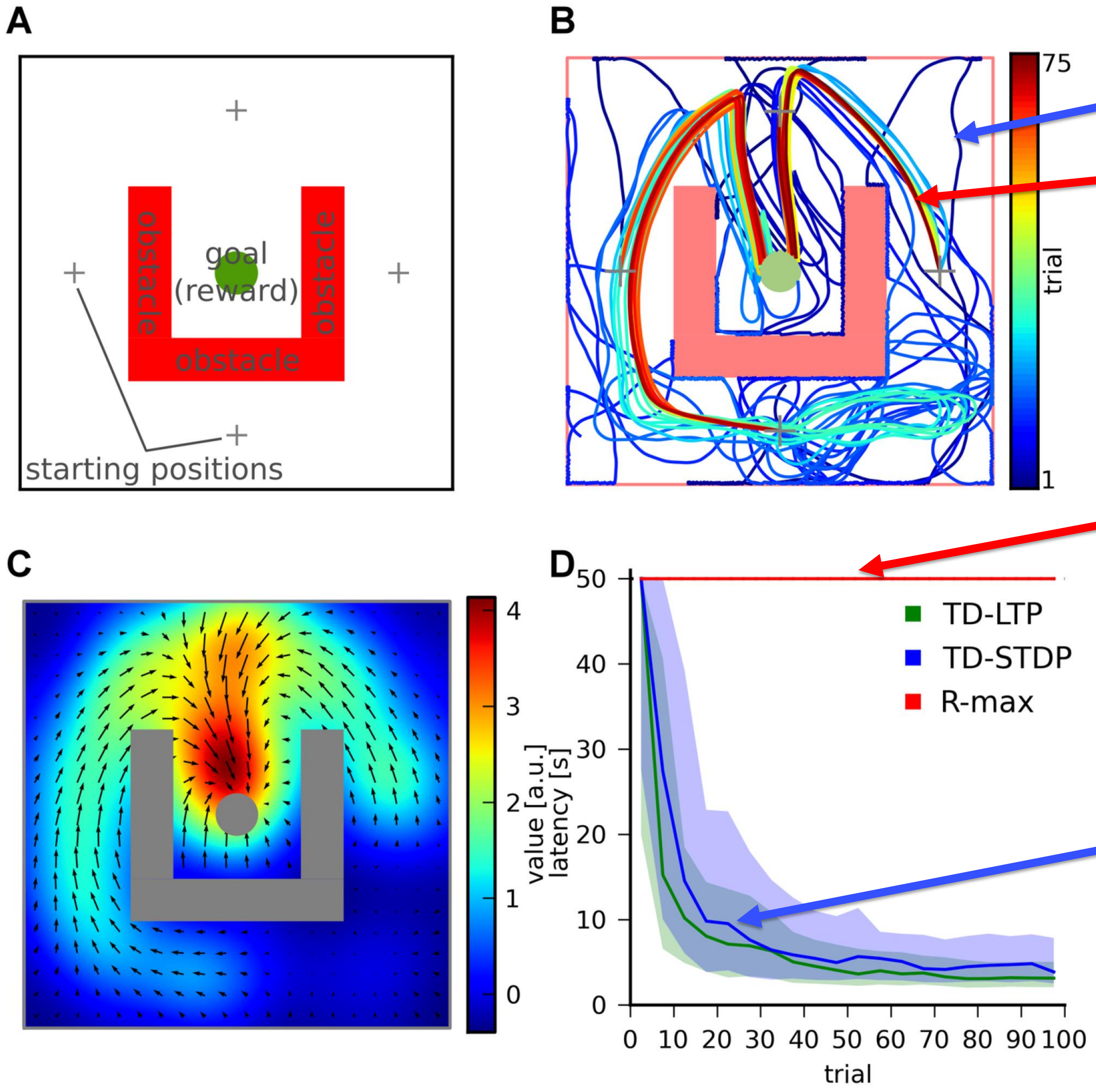
This network is not very deep, but it is powerful since states are represented by Gaussian basis functions. The parameters that need to be learnt are the weights to the actor and the connections to the critic.

Bottom: Gaussian basis functions are also called place cells.

Right: Critic could be a single neuron, but it is implemented in this application by pool of several independent identical neurons (that essentially learn the same value).

Left: action choices are represented by a ring of 360 neurons. In order to generalize well in action space neighboring neurons activate each other while neurons encoding opposite directions inhibit each other. This is a way to implement an inductive bias into the architecture: is direction 88 is good, then direction 89 is typically nearly as good.

Review: Advantage Actor-Critic with eligibility traces



early trial

Late trial

value map, critic,

R-max:
Policy gradient without Subtraction of bias (no critic). The goal was never found within 50s.

Advantage Actor Critic
After 25 trials, the goal was found within 20s.

Maze Navigation: Advantage Actor-Critic

Maze navigation learning task.
A: The maze consists of a square enclosure, with a circular goal area (green) in the center. A U-shaped obstacle (red) makes the task harder by forcing turns on trajectories from three out of the four possible starting locations (crosses).
B: Color-coded trajectories of an example TD-LTP agent during the first 75 simulated trials. Early trials (blue) are spent exploring the maze and the obstacles, while later trials (green to red) exploit stereotypical behavior.
C: Value map (color map) and policy (vector field) represented by the synaptic weights of the agent of panel B after 2000s simulated seconds.
D: Goal reaching latency of agents using different learning rules. Latencies of N~100 simulated agents per learning rule. The solid lines shows the median shaded area represents the 25th to 75th percentiles. The R-max agent were simulated without a critic and enters times-out after 50 seconds. All agents use eligibility traces of time scale 1 second.

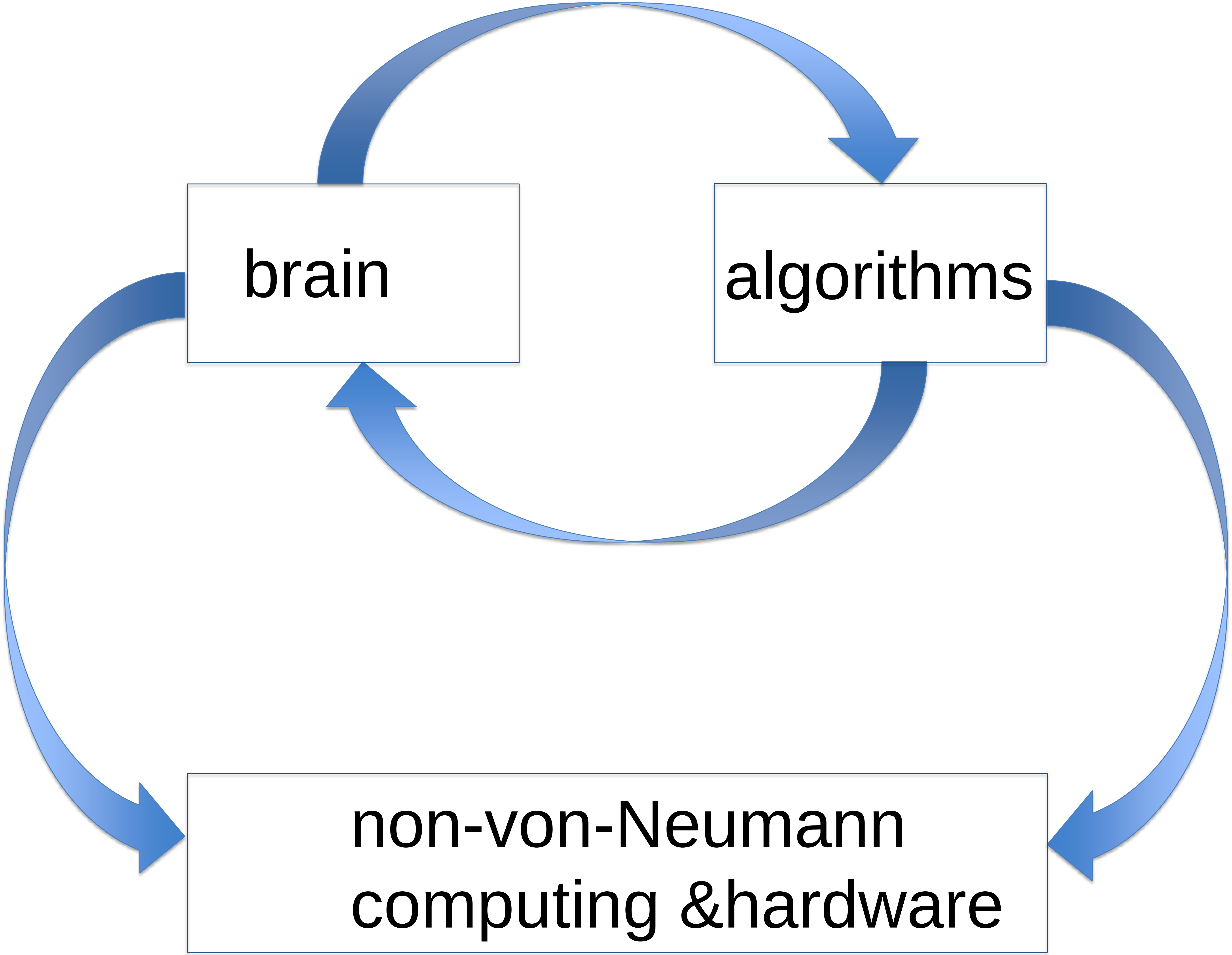
Fremaux et al. (2013)

Questions for today:

- does the brain implement reinforcement learning algorithms?
- Can the brain implement an actor-critic structure?
- What are ‘local learning rules’?
- Why are ‘local learning rules’ potentially important?
- What is a ‘Hebbian’ learning rule?
- What is a ‘three-factor’ learning rule?
- Can we use these ideas for Neuromorphic learning?

Previous slide. Your comments

Learning Rules



brain

algorithms

non-von-Neumann computing &hardware

Previous slide.

Program for this week.

In this introduction, we have reviewed some aspects of RL in an actor-critic structure, in particular the online ‘learning rule’, i.e., the algorithm for the parameter update after each step of the agent. In the following we focus on the learning rule and go back and forth between algorithms and the brain.

Having identified the basic aspects of the learning rule in RL, we now turn to the biology.

Artificial Neural Networks and RL : Lecture 13

from brain-computing to neuromorphic computing

Wulfram Gerstner
EPFL, Lausanne, Switzerland

1. Coarse brain anatomy

Previous slide.

Before we can make a link to Reinforcement Learning we need to know a bit more about the brain.

1. Coarse Brain Anatomy and Reinforcement Learning

Reinforcement learning needs:

- states / sensory representation

  → where are states encoded in the brain?
- action selection

  → where is action selection encoded in the brain?
- reward signals

  → how is reward encoded in the brain?

Previous slide.

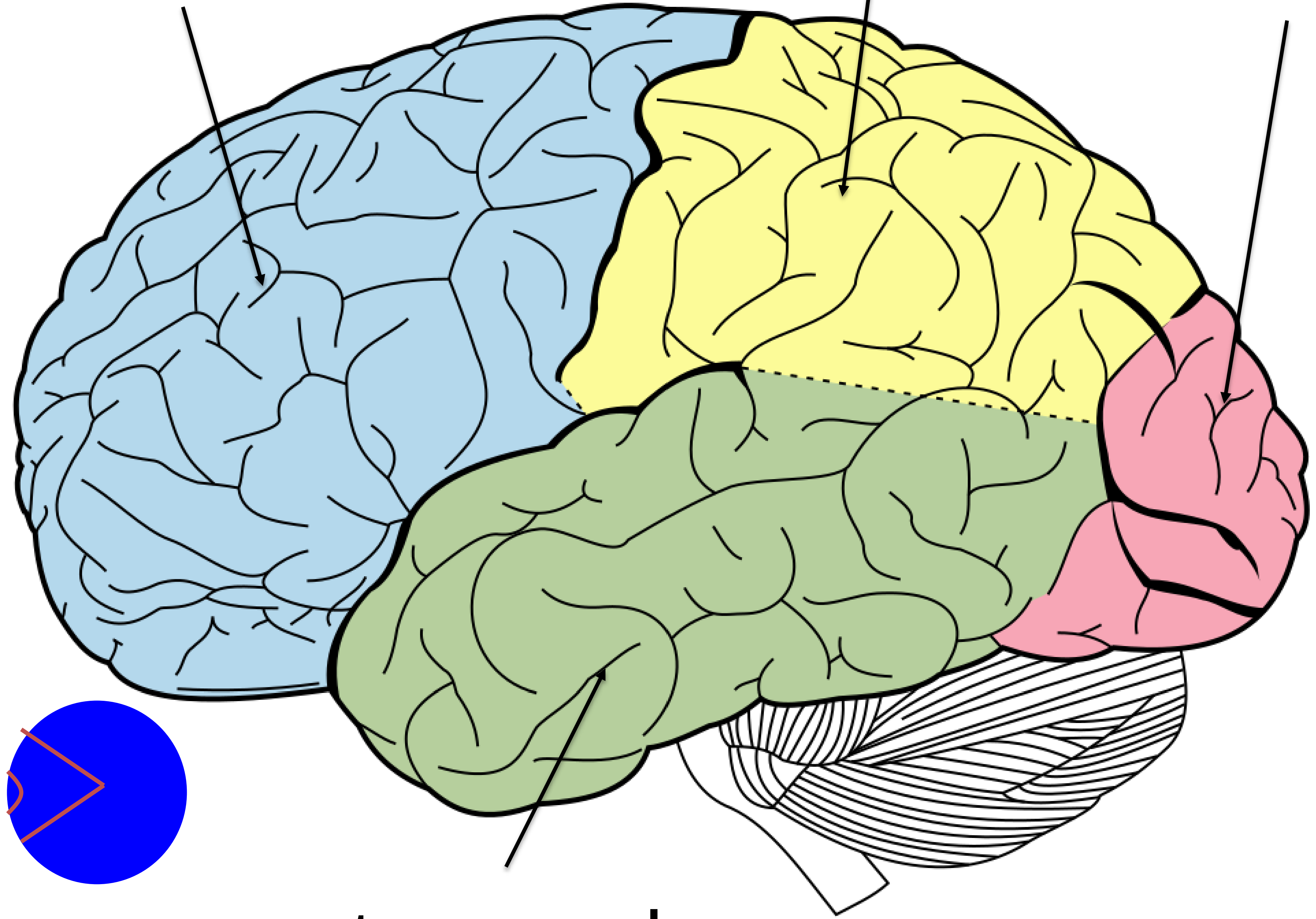
In reinforcement learning, the essential variables that define the update step of the learning rule are the states (defined by sensory representation), a policy for action selection, the actions themselves, and the rewards given by the environment.

If we want to link reinforcement learning to the brain, we will have to search for corresponding substrates and functions in the brain.
Therefore we now take a rather coarse and simplified look at the anatomy of the brain.

The Wikipedia articles give more information for those who are interested.

1. Coarse Brain Anatomy: Cortex

Sensory representation in visual/somatosensory/auditory cortex

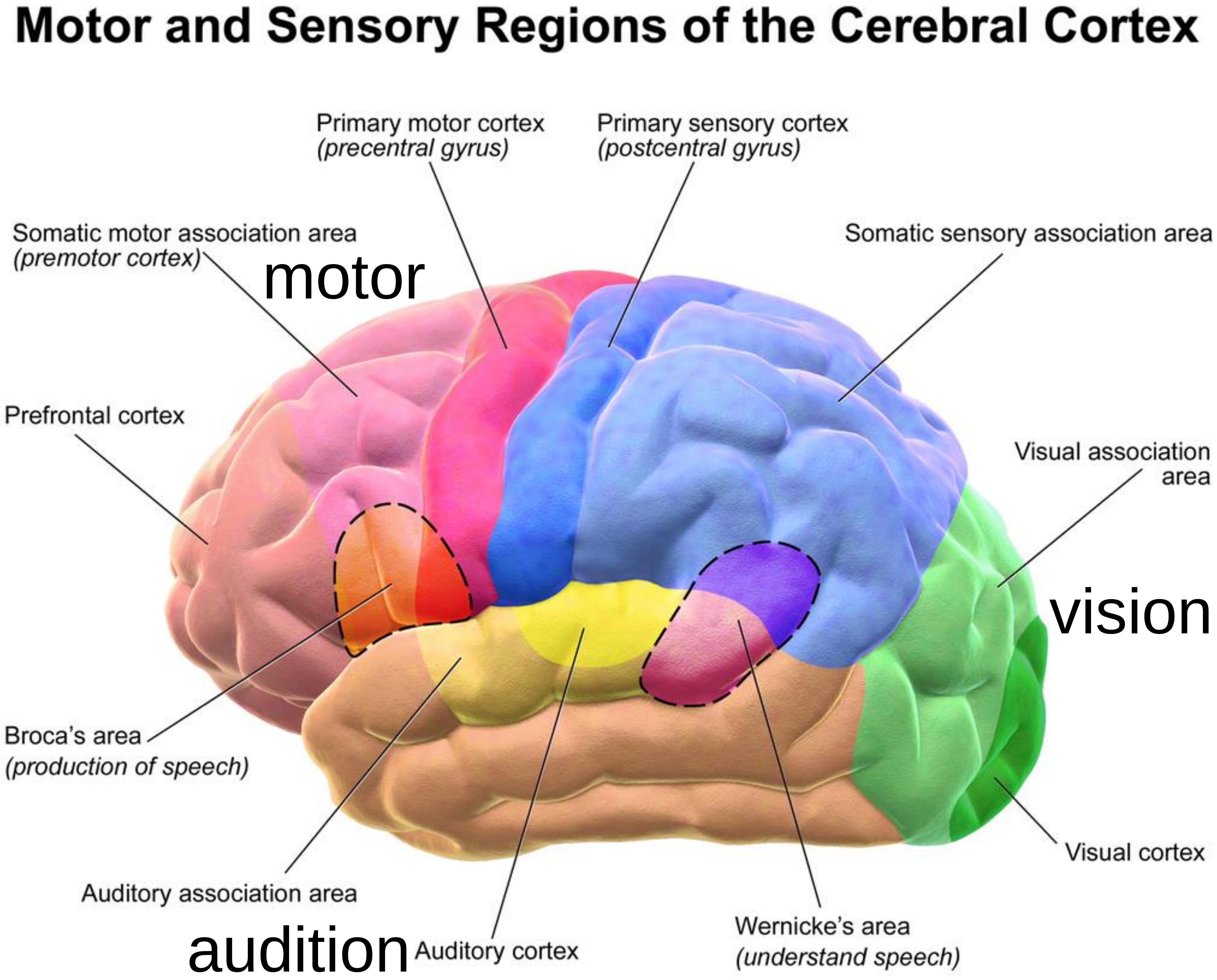


frontal cortex

parietal cortex

occipital cortex

temporal cortex



motor

vision

audition

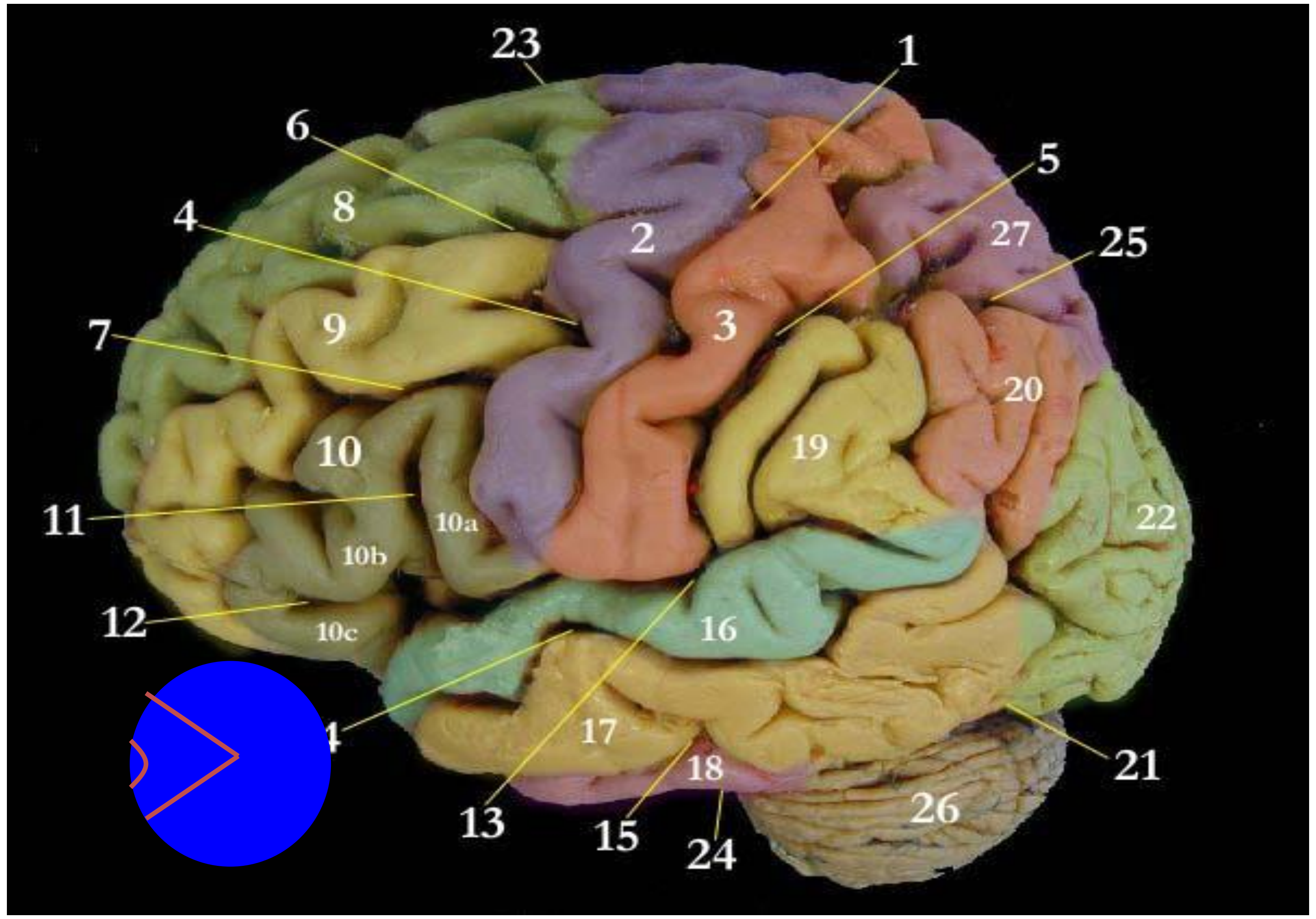
fig: Wikipedia

Previous slide.
Left: Anatomy. The Cortex is the part of the brain directly below the skull. It is a folded sheet of densely packed neurons. The biggest folds separate the four main part of cortex (frontal, Parietal, occipital, and temporal cortex)

Right: Functional assignments. Different parts of the brain are involved in different tasks. For example there several areas involved in processing visual stimuli (called primary and secondary cortex). Other areas are involved in audition (auditory cortex) or the presentation of the body surface (somatosensory cortex). Yet other areas are prepared in the preparation of motor commands for e.g., arm movement.

1. Coarse Brain Anatomy

- many different cortical areas
- but also several brain nuclei sitting below the cortex
- Some of these nuclei send dopamine signals
- Dopamine is related to **reward**, surprise, and pleasure
- Dopamine sent from: VTA and substantia nigra



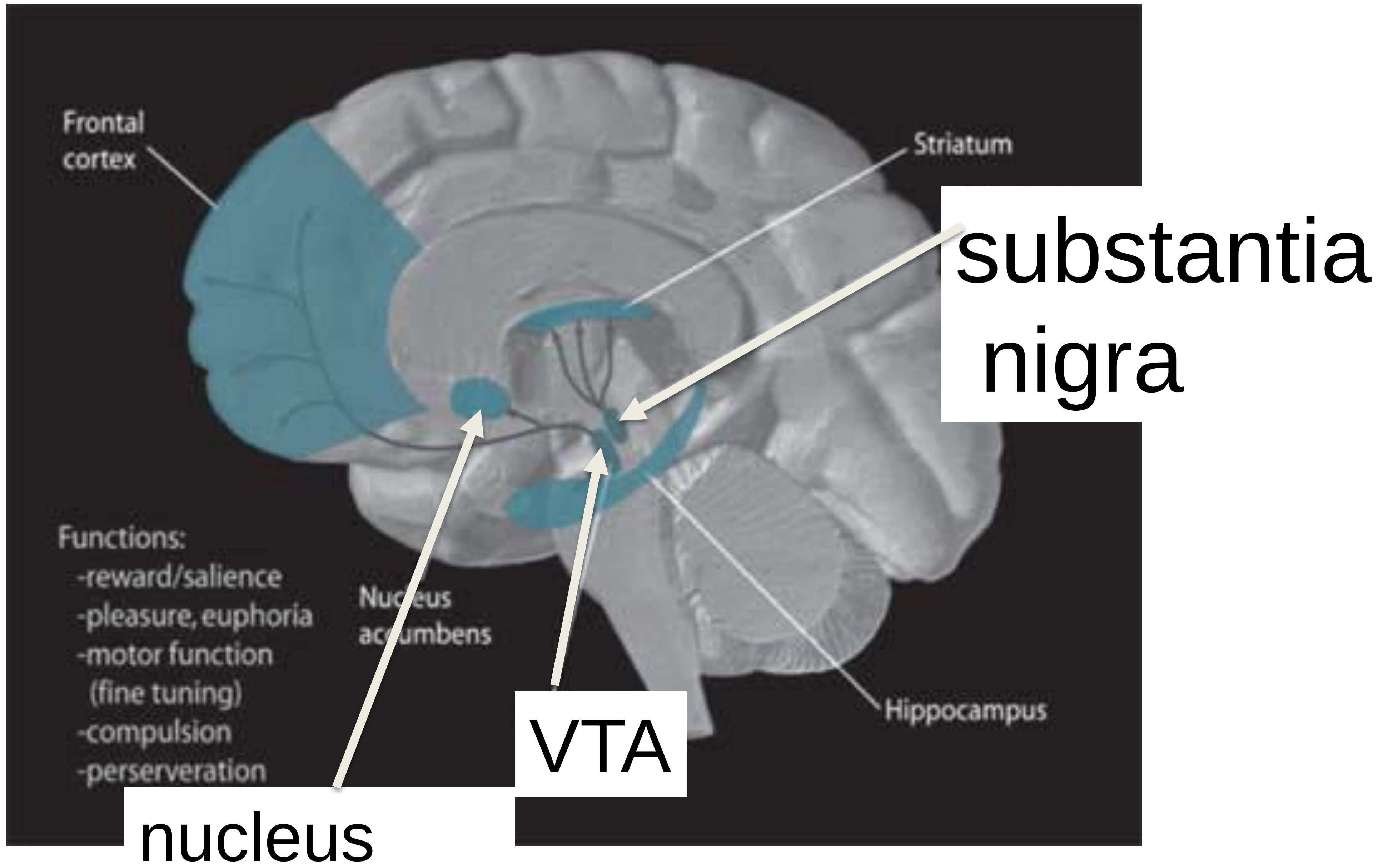


fig: Wikipedia commons

Previous slide.
Left: Anatomy. View on the folds of the cortex, and main cortical areas in different color.

Right: Below the cortex sit different nuclei. Some of these nuclei use dopamine as their signaling molecule. Important nuclei for dopamine are the Ventral Tegmental Area (VTA) and the Substantia Nigra pars compacte (SNc). These dopamine neurons send their signals to large areas of the cortex as well as to the striatum (and nucleus accumbens).
Since dopamine is involved in reward, these dopamine neurons will play a role in this lecture that links reinforcement learning and the brain.

In the next slides we will focus on striatum and hippocampus.

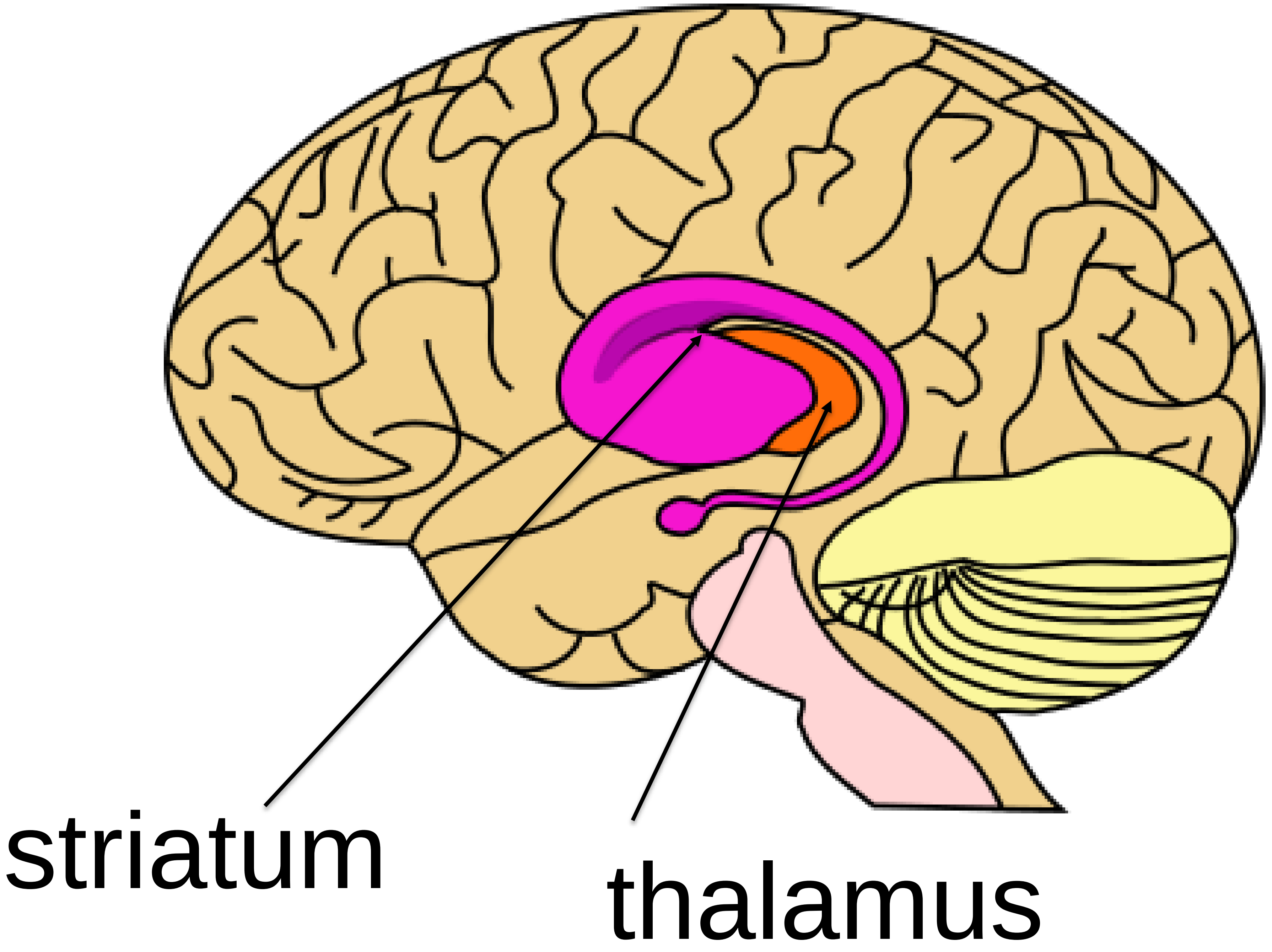
1. Coarse Brain Anatomy: Striatum

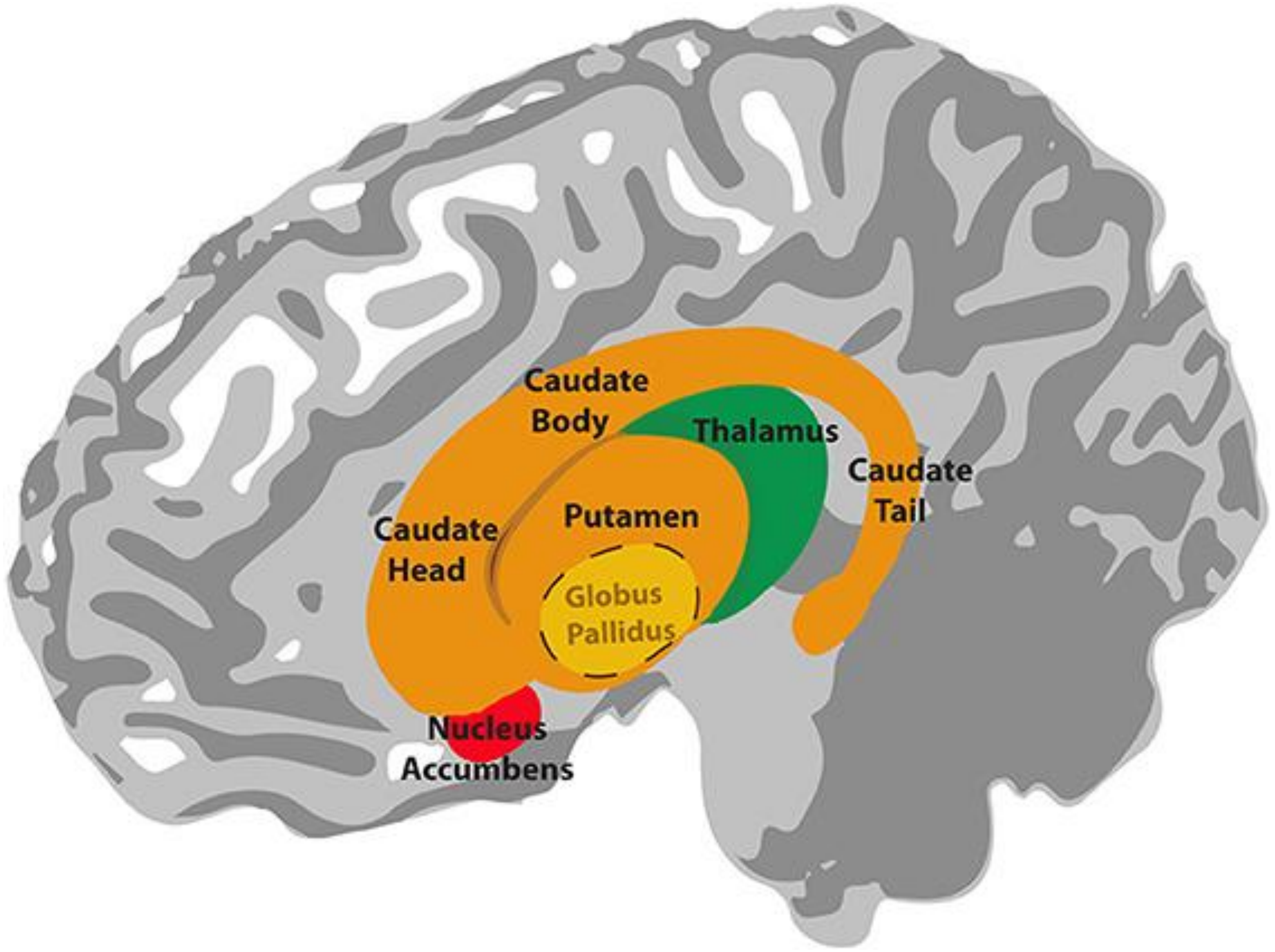
- Striatum sits below cortex
- Part of the ‘basal ganglia’
- **Dorsal striatum** involved in **action selection**, decisions

Striatum consists of

- Caudate (dorsal striatum)
- Putamen (dorsal striatum)

https://en.wikipedia.org/wiki/Striatum





Nucleus Accumbens is part of **ventral striatum**

fig: Wikipedia

Previous slide.
Left: Sketch of the Anatomical location of striatum and thalamus.
Right: the striatum lies also below the cortex. Since the striatum is involved in action selection it will play an important role in this lecture.
From Wikipedia:

The **striatum** is a nucleus (a cluster of neurons) in the subcortical basal ganglia of the forebrain. The striatum is a critical component of the motor and reward systems; receives glutamatergic and dopaminergic inputs from different sources; and serves as the primary input to the rest of the basal ganglia.

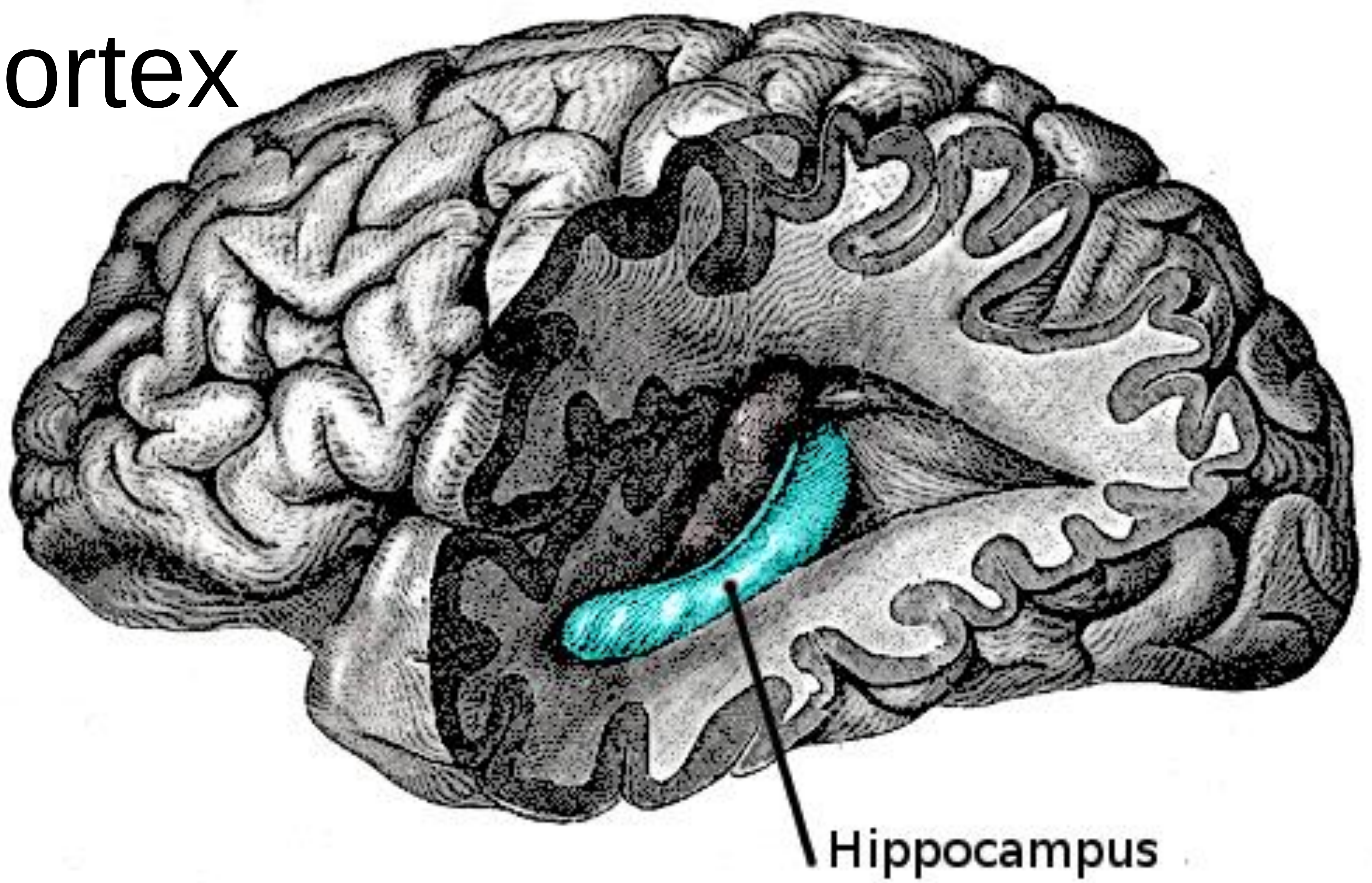
Functionally, the striatum coordinates multiple aspects of cognition, including both motor and action planning, decision-making, motivation, reinforcement, and reward perception.The striatum is made up of the caudate nucleus and the lentiform nucleus. The lentiform nucleus is made up of the larger putamen, and the smaller globus pallidus.

In primates, the striatum is divided into a **ventral striatum**, and a **dorsal striatum**, subdivisions that are based upon function and connections. The ventral striatum consists of the nucleus accumbens and the olfactory tubercle. The dorsal striatum consists of the caudate nucleus and the putamen. A white matter, nerve tract (the internal capsule) in the dorsal striatum separates the caudate nucleus and the putamen.[4] Anatomically, the term *striatum* describes its striped (striated) appearance of grey-and-white matter

1. Coarse Brain Anatomy: hippocampus

Hippocampus

- Sits below/part of temporal cortex
- Involved in memory
- Involved in spatial memory



Spatial memory:

knowing where you are,

knowing how to navigate in an environment

fig: Wikipedia

Henry Gray (1918) *Anatomy of the Human Body*

Previous slide.
From Wikipedia:

The **hippocampus** (named after its resemblance to the seahorse, from the Greek ἱππόκαμπος, "seahorse" from ἵππος *hippos*, "horse" and κάμπος *kampos*, "sea monster") is a major component of the brains of humans and other vertebrates. Humans and other mammals have two hippocampuses, one in each side of the brain. The hippocampus belongs to the limbic system and plays important roles in the consolidation of information from short-term memory to long-term memory, and in spatial memory that enables navigation. The hippocampus is located under the cerebral cortex (allocortical)[1][2][3] and in primates in the medial temporal lobe.

1. Coarse Brain Anatomy and Reinforcement Learning

Reinforcement learning needs:

- states / sensory representation → cortex?, hippocampus?

- action selection → striatum?, motor cortex?

- reward signals → dopamine?

Previous slide.

In reinforcement learning, the essential variables are the states (defined by sensory representation), a policy for action selection, the actions themselves, and the rewards given by the environment.

If we want to link reinforcement learning to the brain, we will have to search for corresponding substrates and functions in the brain.

The above rough ideas need to be defined during the rest of this lecture.

1. Quiz: Coarse Functional Brain anatomy

[ ] the brain = the cortex (synonyms)
[ ] the cortex consists of several areas
[ ] some areas are more involved in vision, others more in the representation of the body surface
[ ] below the cortex there are groups (clusters) of neurons
[ ] Hippocampus sends out dopamine signals
[ ] VTA and nucleus accumbens send out dopamine signals
[ ] dopamine is linked to reward, pleasure, surprise
[ ] striatum is involved in action selection

Previous slide. Your comments

Artificial Neural Networks and RL : Lecture 13

Wulfram Gerstner
EPFL, Lausanne, Switzerland

from brain-computing to neuromorphic computing

1. Coarse Brain Anatomy
2. **Synaptic Plasticity**

Previous slide.

Reinforcement Learning is, obviously, a form of ‘learning’. Learning is related to synaptic plasticity. Therefore this is our second topic.

2. Behavioral Learning

Learning actions:

→ riding a bicycle
→ play tennis
→ play the violin

‘models of action choice’

Remembering episodes

→ first day at EPFL
→ plan how to get home
→ reward-free

‘models of the world’

Previous slide.

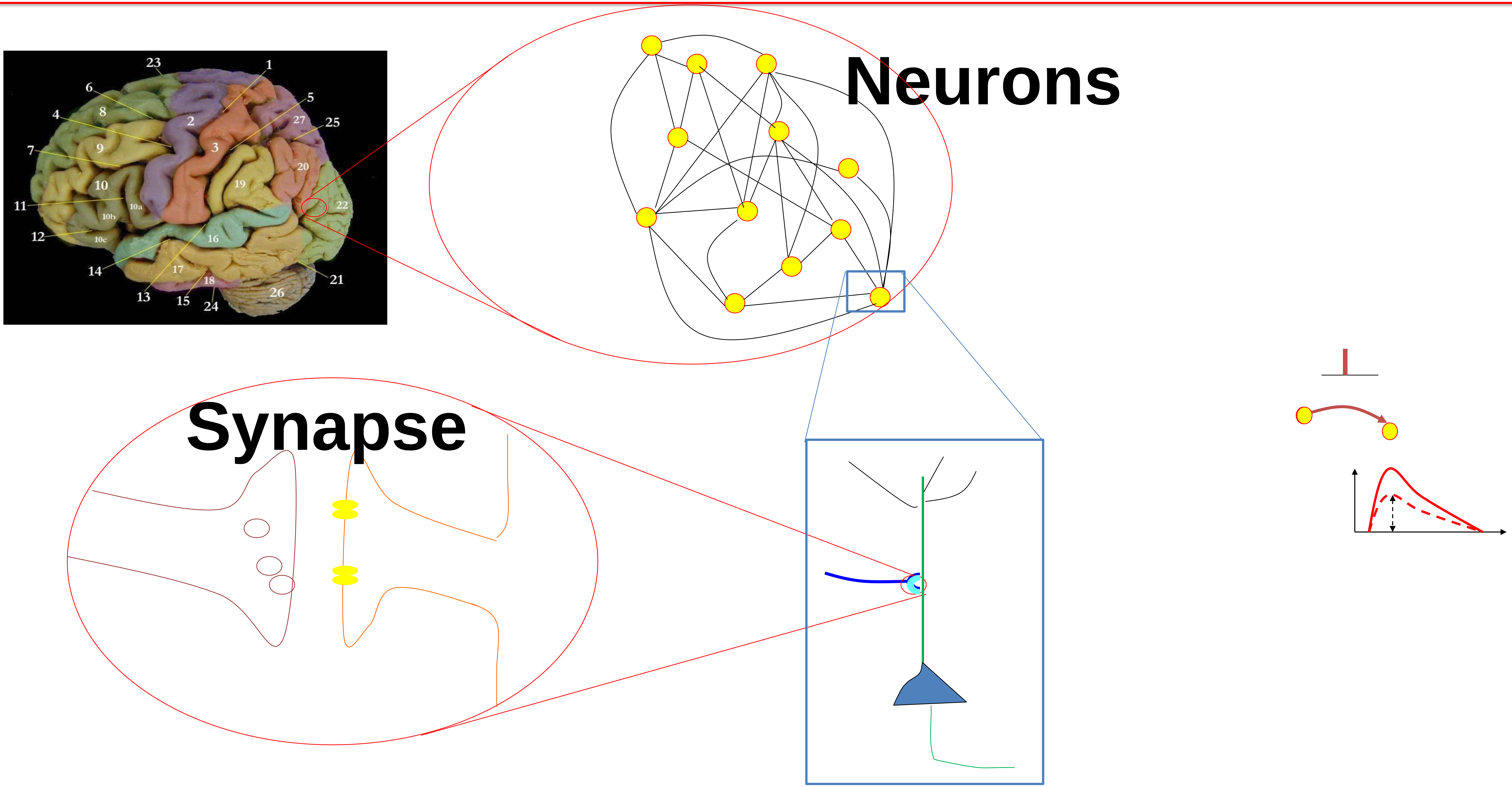
When we learn to ride a bike we learn with Reinforcement-like feedback, e.g., we don’t want to fall because falling hurts.

When we learn play the tennis or the violin we also get feedback via the observed outcome – which can be good or bad.

When we walk around a city for the first time we develop a model of the environment – even in the absence of any specific rewards (except, may be, that it is good to know how to find the way home).

All these are examples of learning. The last one might be unsupervised learning, but the others are clearly reinforcement learning.

2. Behavioral Learning – and synaptic plasticity



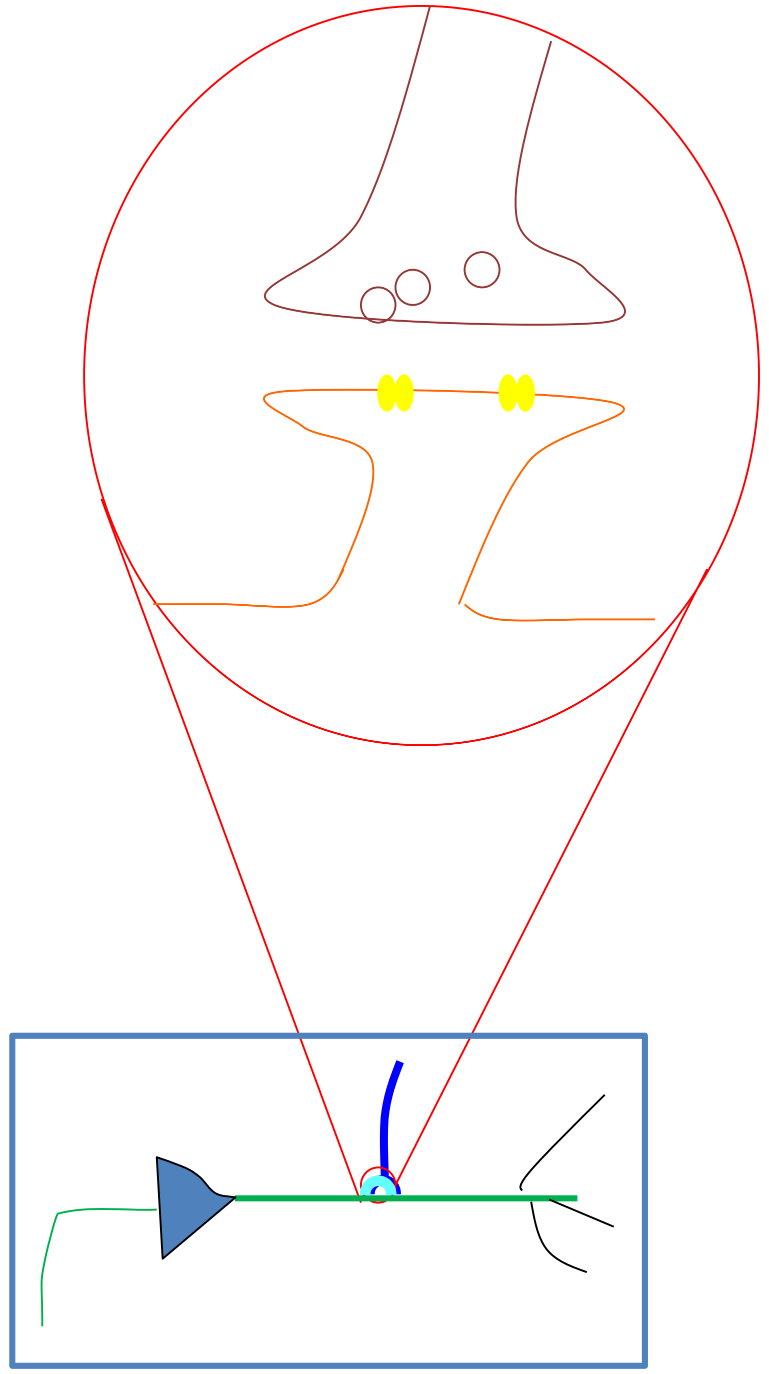
Synaptic Plasticity =Change in Connection Strength

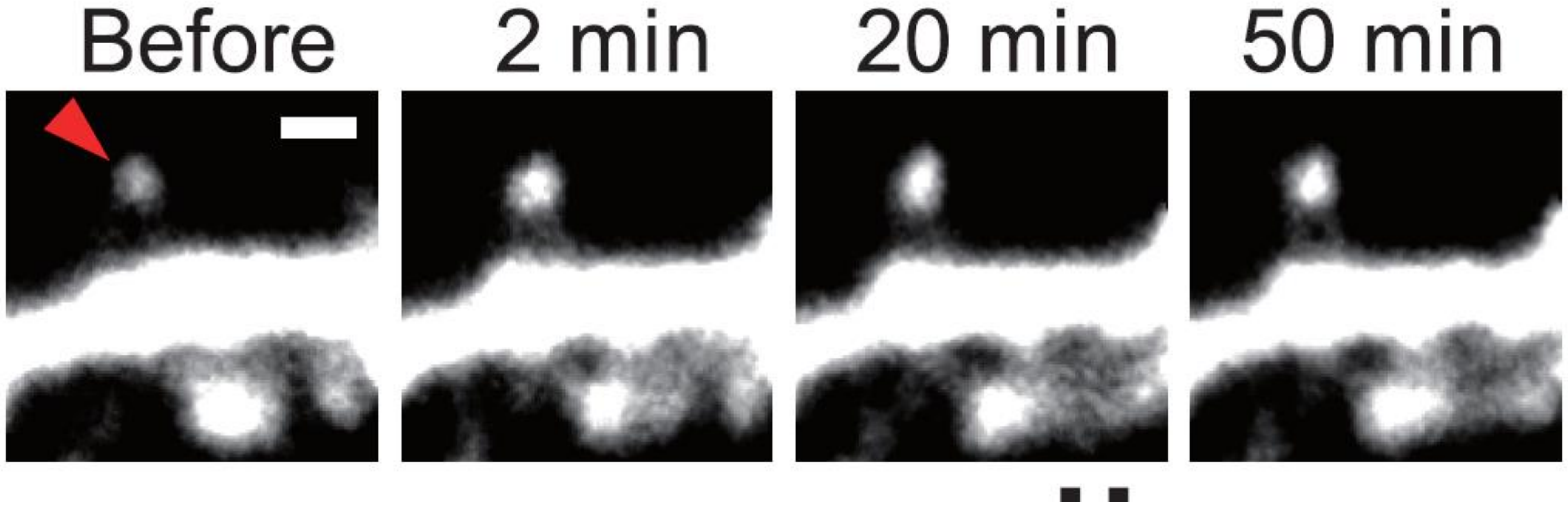
Previous slide.

When we observe learning on the level of behavior (we get better at tennis), then this implies that something has changed in our brain:
The contact points between neurons (called synapses) have changed. Synaptic changes manifest themselves as a change in connections strength.

Synaptic plasticity describes the phenomena and rules of synaptic changes.

2. Synaptic plasticity – structural changes





Yagishita et al.
Science, 2014

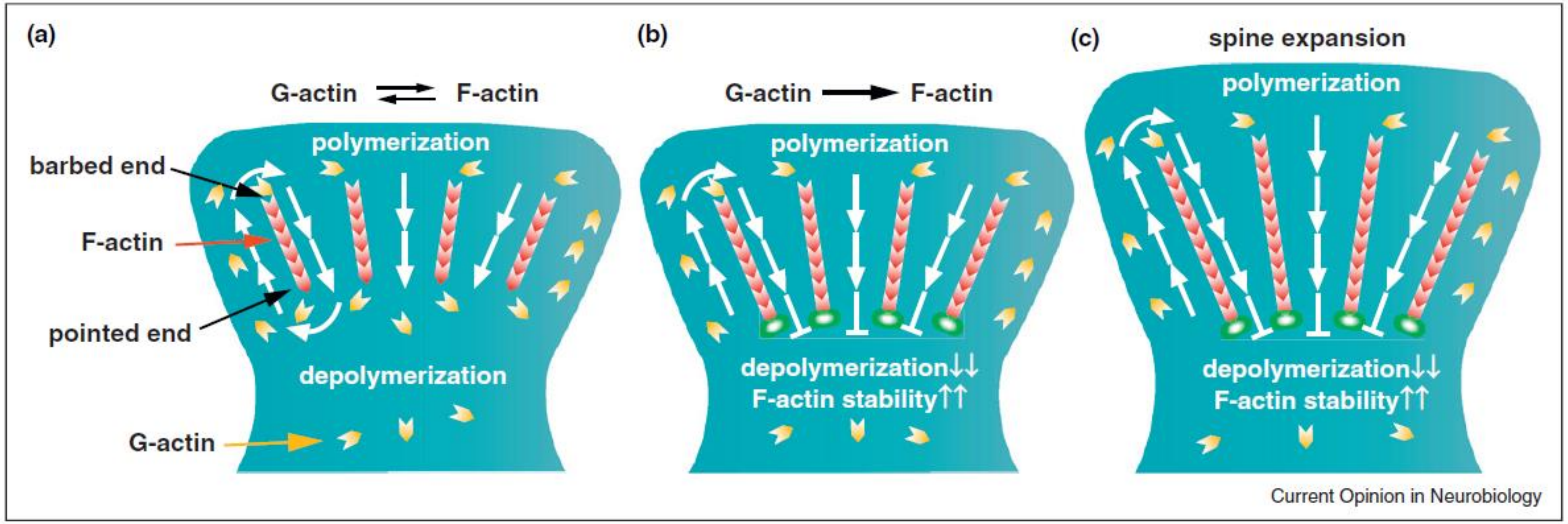
Previous slide.

The synaptic connection consists of two parts. The end of an axonal branch coming from the sending neuron; and the counterpart, a protrusion on the dendrite of the receiving neuron, called spine.

We refer to the sending neuron as presynaptic and to the receiving one as postsynaptic.

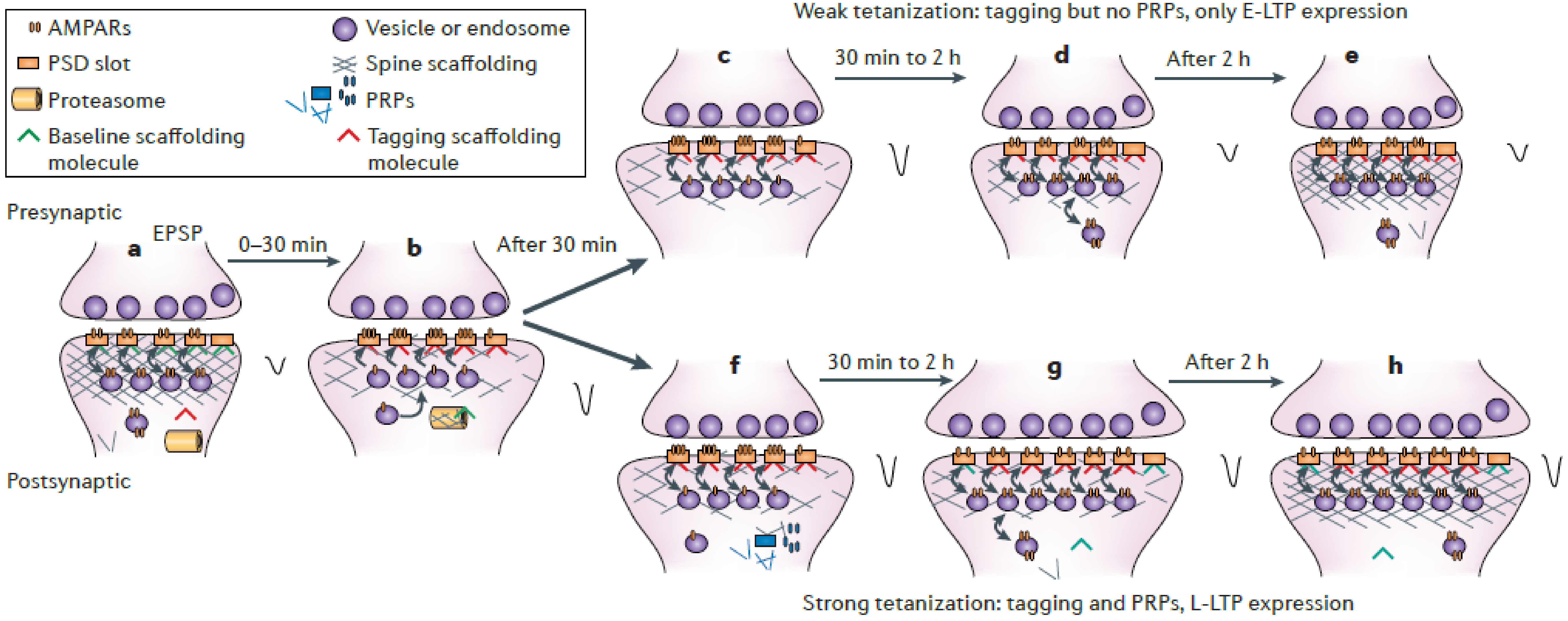
A change in the connection strength is observable with imaging methods as an increase in the size of the spine. The bigger spine remains big for a long time (here observed for nearly one hour).

2. synaptic plasticity – molecular changes



Current Opinion in Neurobiology

Bosch et al. 2012, Curr. Opinion Neurobiol.

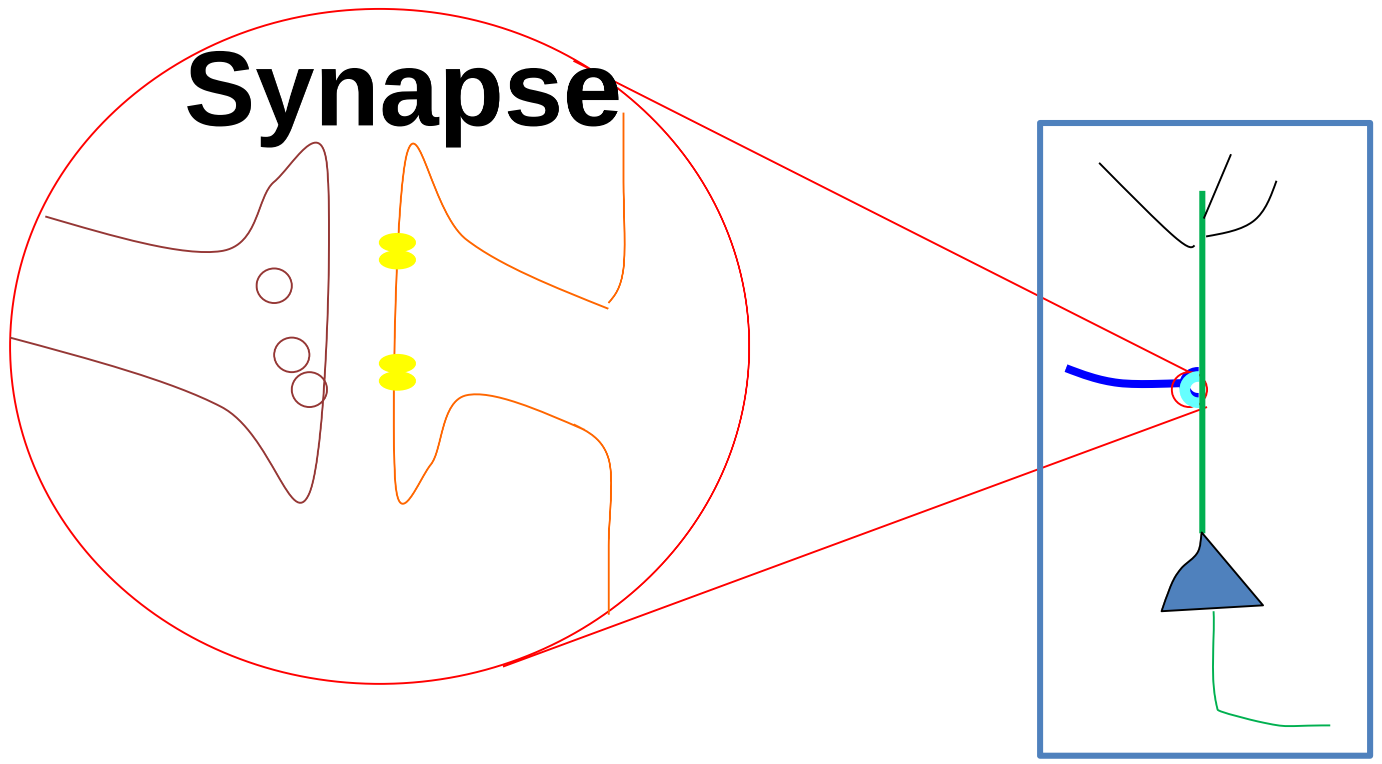

AMPARs
PSD slot
Proteasome
Baseline scaffolding molecule
Vesicle or endosome
Spine scaffolding
PRPs
Tagging scaffolding molecule
Weak tetanization: tagging but no PRPs, only E-LTP expression
Presynaptic
EPSP
Postsynaptic
a
0–30 min
b
After 30 min
c
30 min to 2 h
d
After 2 h
e
f
30 min to 2 h
g
After 2 h
h
Strong tetanization: tagging and PRPs, L-LTP expression

Redondo and Morris 2011, Nature Rev. Neurosci.

Previous slide. (not shown in class)

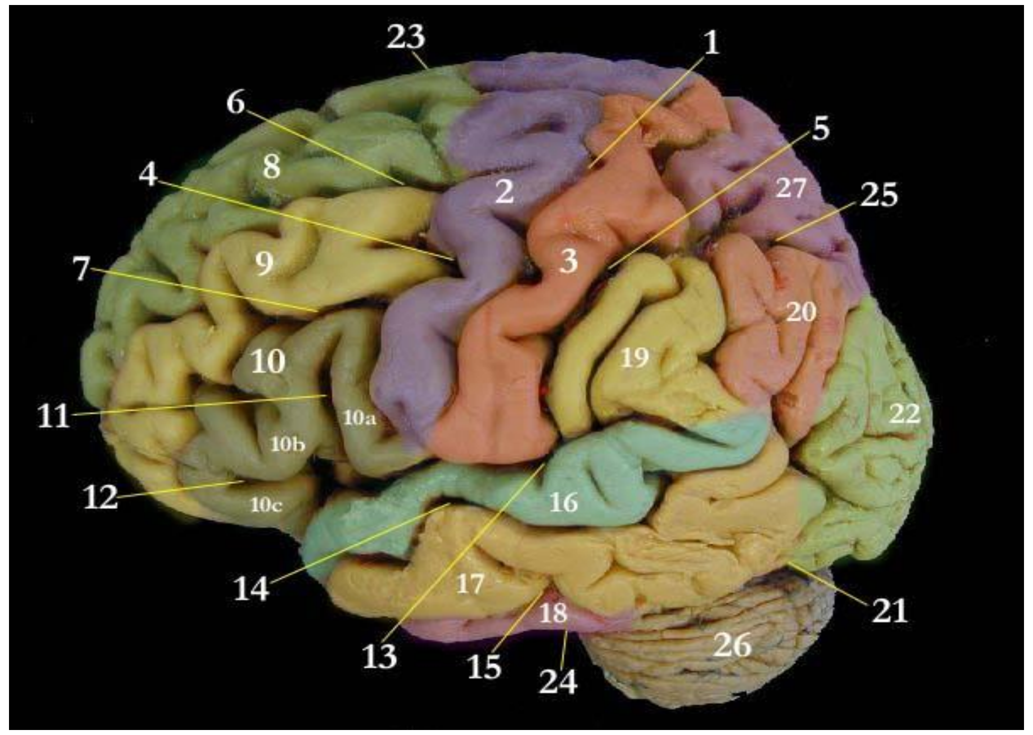
The actual molecular machinery inside the spine is very complicated – and of no further interest of us in the following.

2. synaptic plasticity – connections change



More space for fingers allocated in cortex
- musicians vs. non-musicians

Amunts et al. Human Brain Map. 1997
Gaser and Schlaug, J. Neuosci. 2003



More space allocated in hippocampus
- London taxi driver vs bus driver

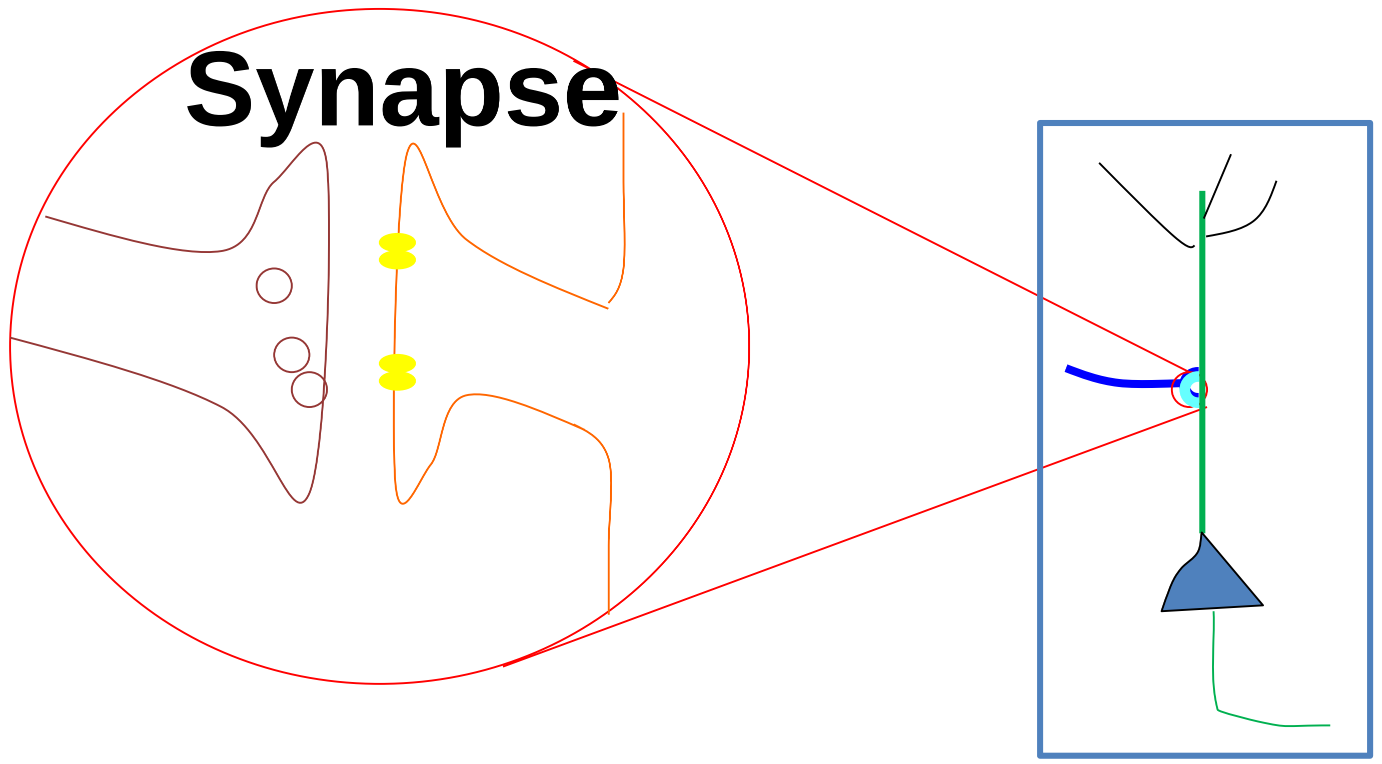
Macquire et al. Hippocampus 2006

Previous slide.

We said at the beginning of the lecture that different areas of the brain are involved in different tasks. For example, the somatosensory cortex represents the body surface. Nowadays one can measure that the size of the cortical area devoted to fingers is larger for musicians than for non-musicians. Since musicians are not born with a larger area, this result implies that synaptic plasticity can influence the function of the neurons in the brain.

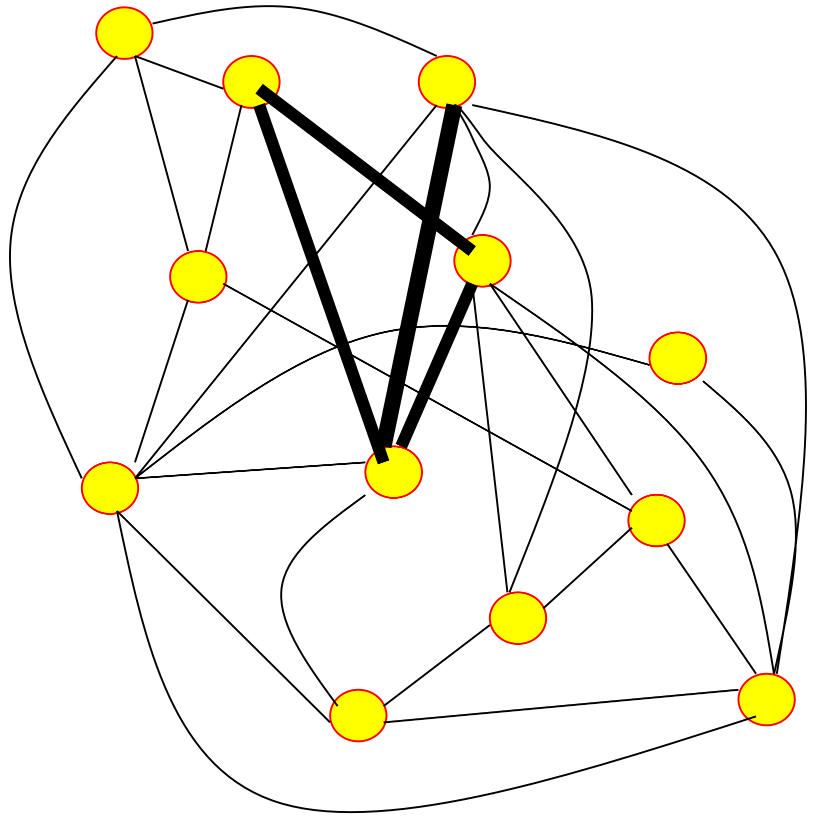
Similarly, hippocampus is involved in spatial navigation. Not surprisingly, London taxi drivers have a bigger hippocampus than London bus drivers.

2. Synaptic plasticity: summary



- Connections can be strong or weak
- Strong connections have thick spines
- Synaptic plasticity

  = change of connection



Syn. Plasticity should enable Learning

- adapt to the statistics of task and environments (useful filters, allocate space etc)
- memorize facts and episodes
- learn models of the world
- learn motor tasks

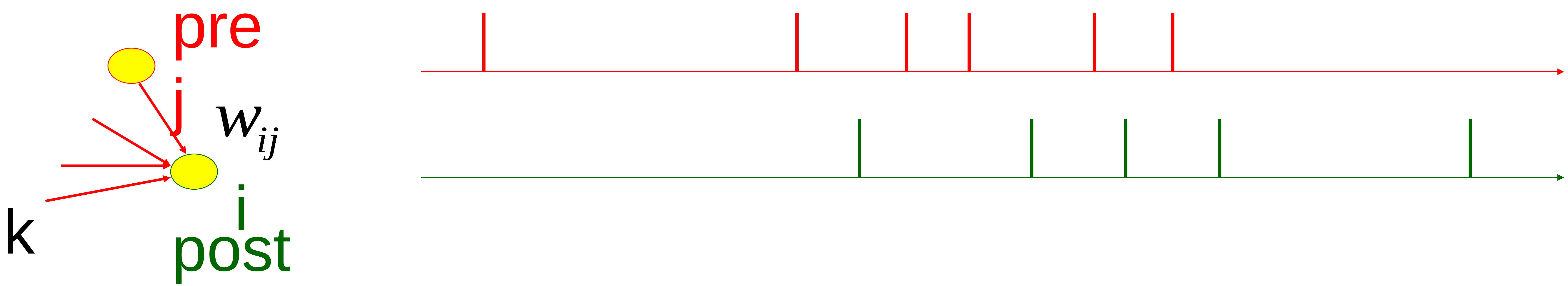
Previous slide.

Thus connections can be strong or weak – and synaptic plasticity describes the changes of one synapse from weak to strong or back.

The synaptic changes are thought to be the basis of learning – whatever the learning task at hand.

The question now is: Are the any rules that would predict whether and when a synapse gets stronger?

2. Hebb rule / Hebbian Learning



When an axon of cell j repeatedly or persistently takes part in firing cell i, then j's efficiency as one of the cells firing i is increased

Hebb, 1949

- local rule
- simultaneously active (correlations)

Previous slide.

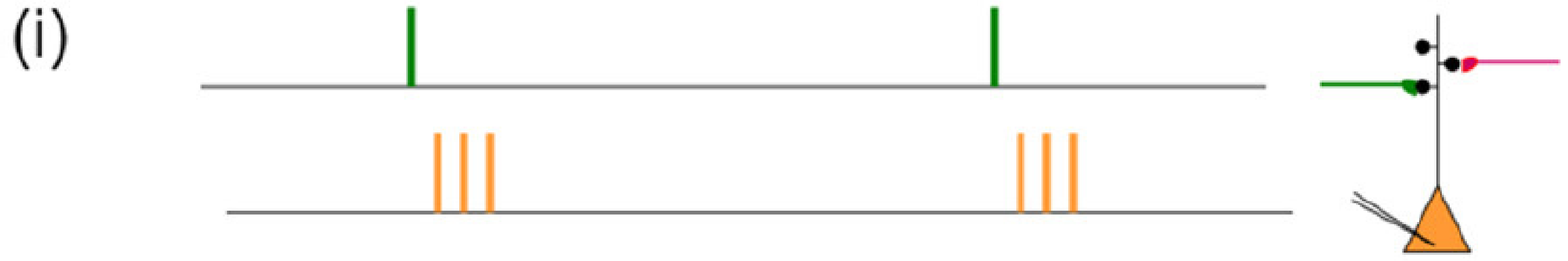
The Hebb rule is the classic rule of synaptic plasticity.

It is often summarized by saying: if two neurons are active together, the connection between those two neurons gets stronger.
Note that the original formulation of Hebb also has a ‘causal’ notion: ‘takes part in firing’ – which is more than just firing together.

Local rule means: changes only depend on information that is available at the synapse. The changes for the weight from j to i can depend on the activity of neuron j and the state (or activity) of neuron i, and the value of the weight itself, but for example not explicitly on the activity of another neuron k. Note that if k connects to i, the activity of i summarizes the influence of k. In other words, i may depend IMPLICITLY on k, but the weight changes do not depend EXPLICITLY on k.

2. Hebbian Learning (LTP)

Hebbian coactivation:
pre-post-post-post



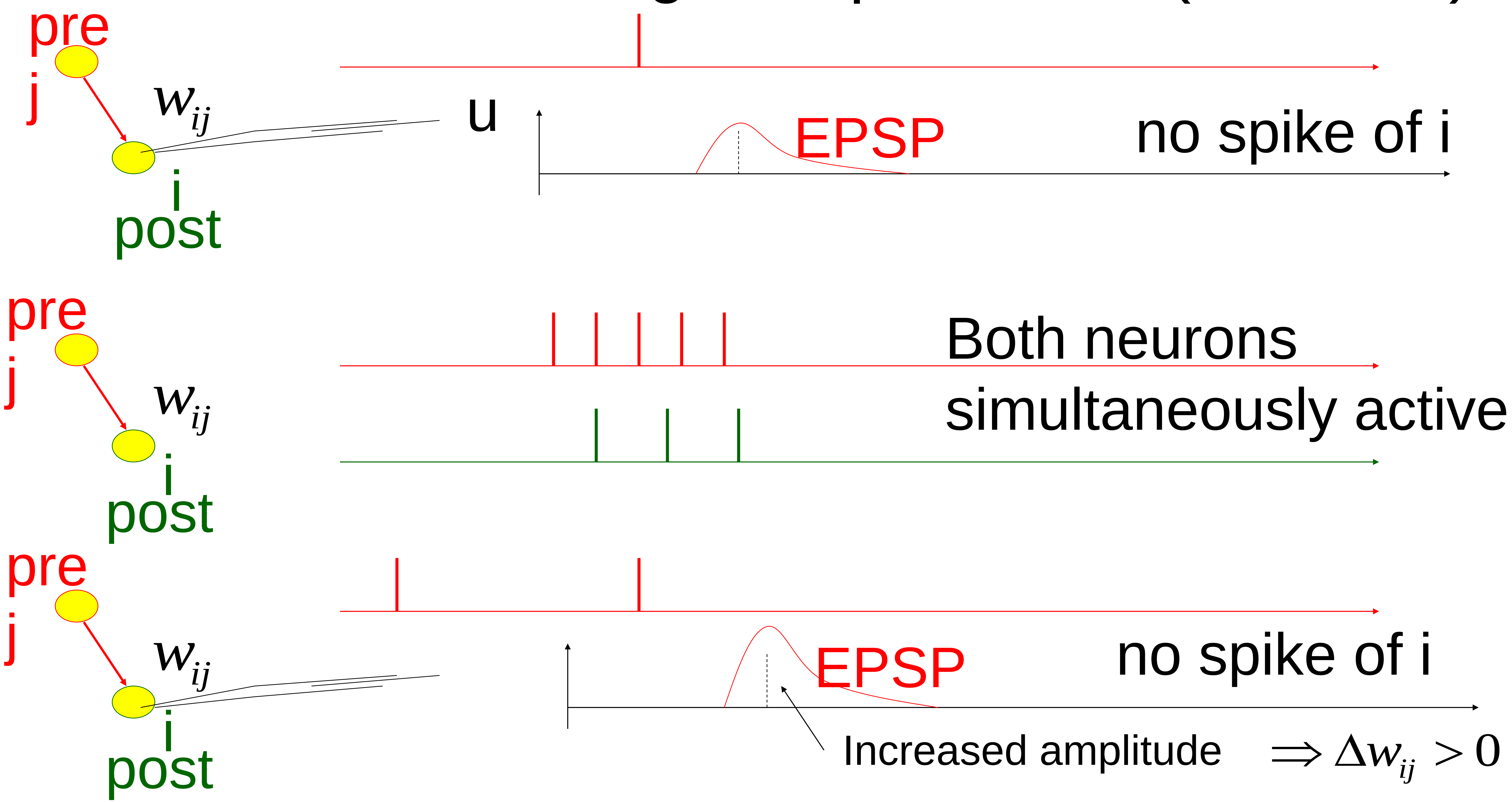
Previous slide.

The joint activation of pre- and postsynaptic neuron induces a strengthening of the synapses. A strong stimulus is several repetitions of a pulse of the presynaptic neuron, followed by three or four spikes of the postsynaptic neuron.

Hundreds of experiments are consistent with Hebbian learning.

2. Synaptic plasticity: Long-Term Potentiation (LTP)

Hebbian Learning in experiments (schematic)



Previous slide.

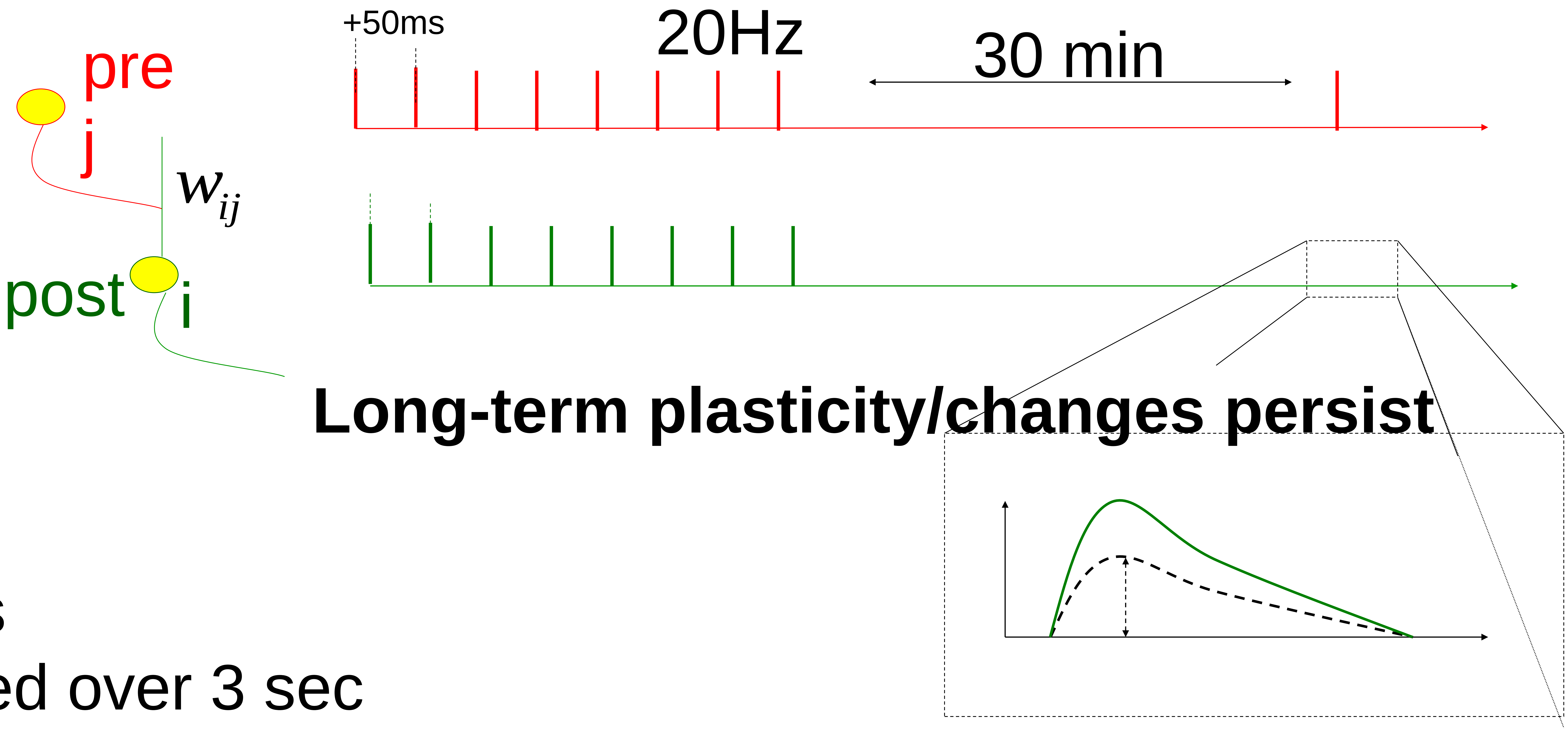
In a schematic experiment,

1) You first test the size of the synapse by sending a pulse from the presynaptic neurons across the synapses. The amplitude of the excitatory postsynaptic potential (EPSP) is a convenient measure of the synaptic strength. It has been shown that it is correlated with the size of the spine.

2) Then you do the Hebbian protocol: you make both neurons fire together

3) Finally you test again the size of the synapse. If the amplitude is bigger you conclude that the synaptic weight has increased.

2. Why the name ‘ Long-term plasticity ‘ (LTP)?



Long-term plasticity/changes persist

Changes

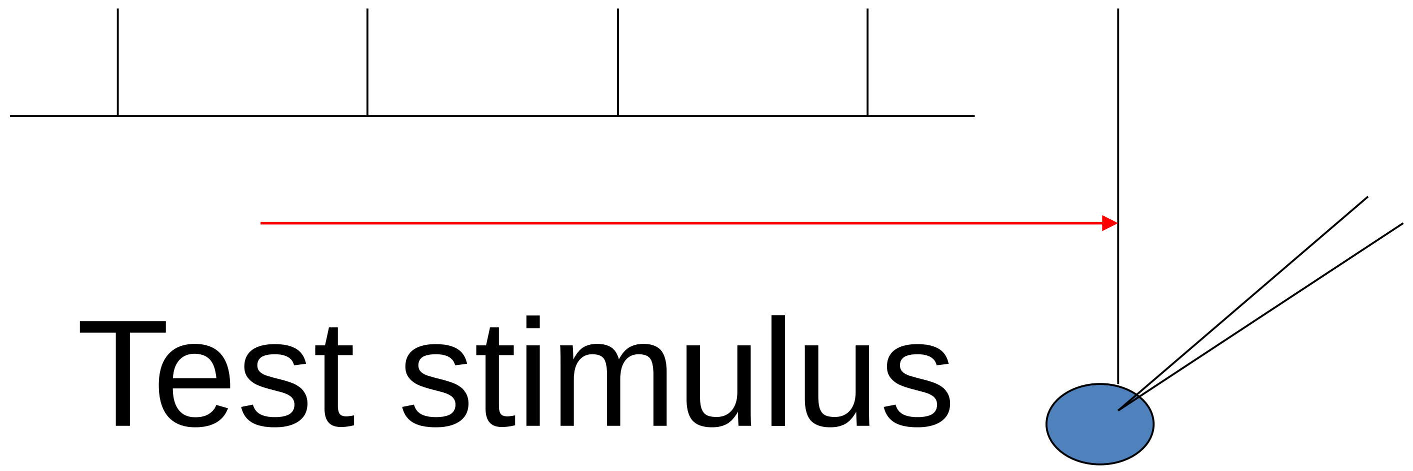
- induced over 3 sec
- persist over 1 – 10 hours (or longer?)

Previous slide.

Experimentalists talk about Long-Term Potentiation (LTP), because once the change is induced it persists for a long time. Interestingly, it is sufficient to make the two neurons fire together for just a few seconds.

Thus inducation of plasticity is rapid, but the changes persist for an hour or more.

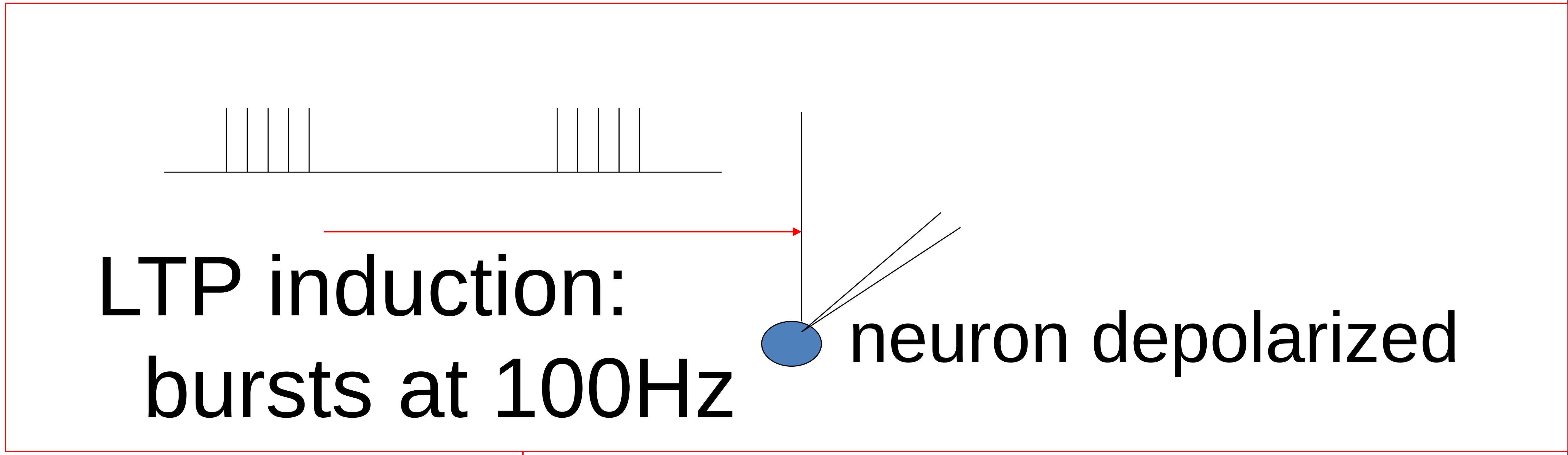
2. Classical paradigm of LTP induction – pairing



Test stimulus

At 0.1 Hz

neuron at -70mV



LTP induction:

bursts at 100Hz

neuron depolarized

Standard LTP

PAIRING experiment

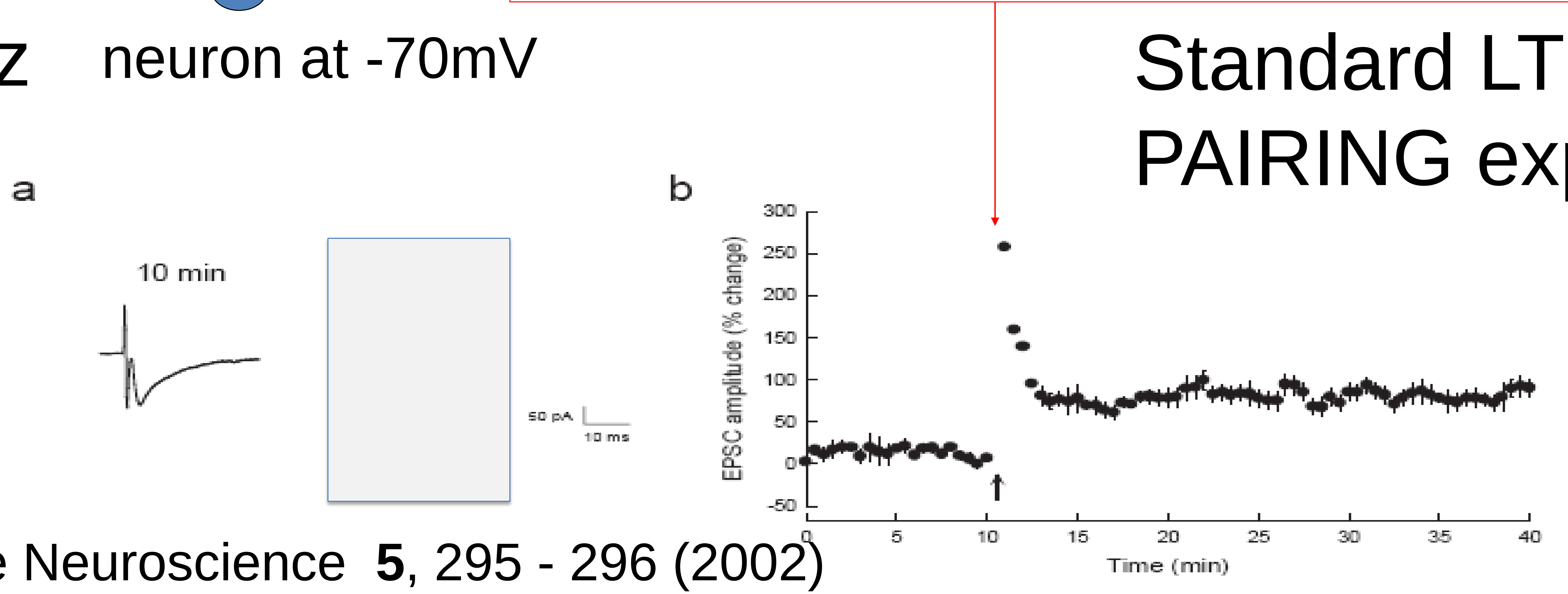


Fig. from Nature Neuroscience **5**, 295 - 296 (2002)

D. S.F. Ling, … & Todd C. Sacktor

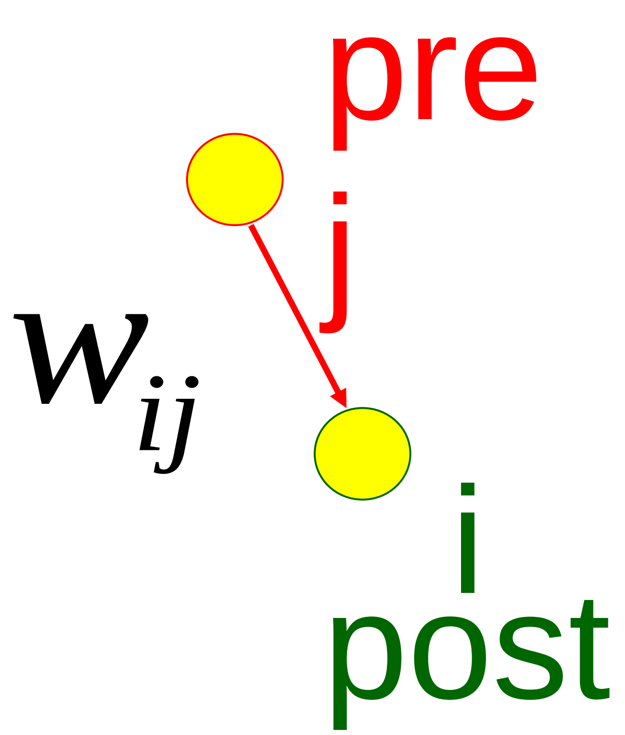
See also: Bliss and Lomo (1973), Artola, Brocher, Singer (1990), Bliss and Collingridge (1993)

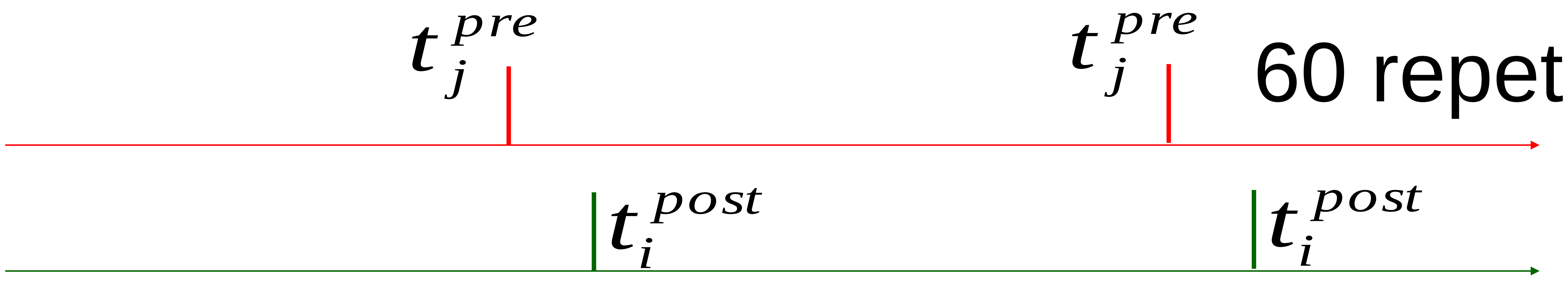
Previous slide. Not shown in class.

In one classic paradigm of LTP induction, the presynaptic fibers are strongly stimulated (with bursts of 100 pulses per second, repeated several times) while the postsynaptic neuron is stimulated with an electrode to put above its normal ‘resting potential’.

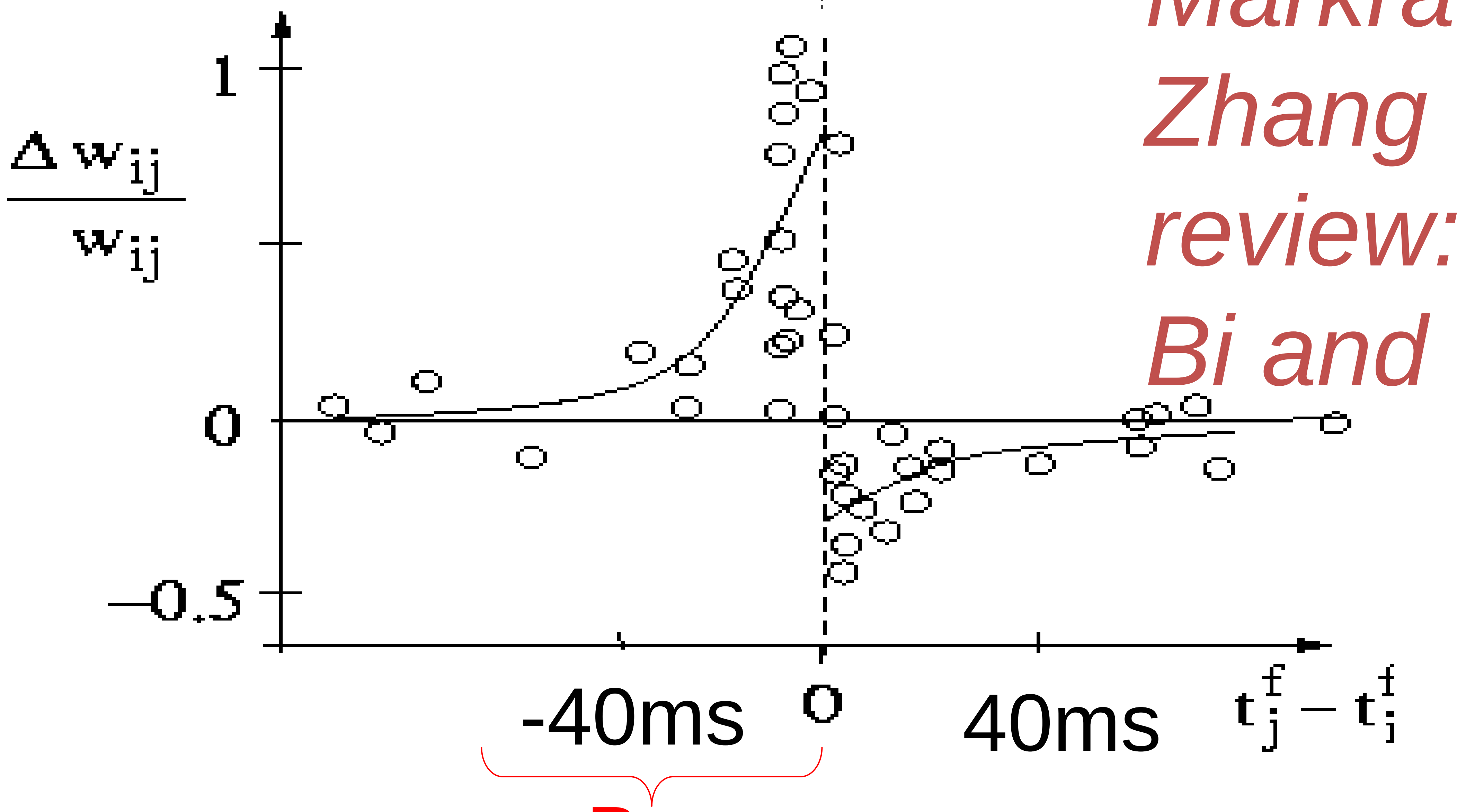
The size of the synapses is measured by the excitatory postsynaptic current (EPSC) which is itself proportional to the EPSP. After the stimulation (which lasts less than a minute) the synapse remains strong for a long time.
The initial transient is of no importance for our discussion.

2. Spike-timing dependent plasticity (STDP)



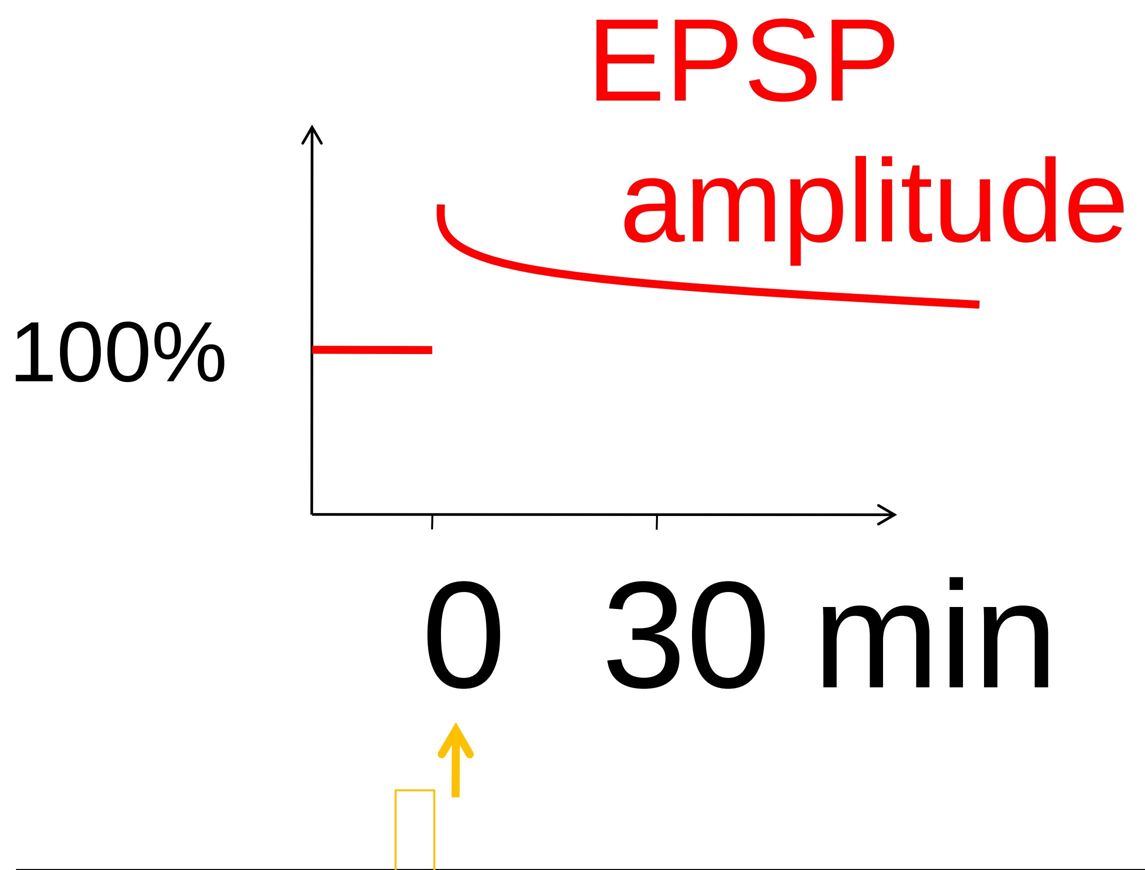


60 repetitions



Pre
before post

Markram et al, 1995,1997
Zhang et al, 1998
review:
Bi and Poo, 2001



Previous slide.
In the STDP paradigm of LTP induction, the presynaptic neuron is stimulated so that it emits a single spike, and the postsynaptic neuron is also stimulated so that emits a single spike – either a few milliseconds before or after the presynaptic spike. This stimulation protocol (for example pre-before-post) is then repeated several times.

The increase of the synaptic weight (induced by repeated pre-before-post) persists for a long time.
How much it increases (or decreases) depends on the exact timing of conicidences of pre- and post-spikes on the time scale of 10ms

Since the size of the increase depends on the relative timing of the two spikes, this induction protocol is called Spike-Timing-Dependent Plasticity (STDP).

2. Summary: Synaptic plasticity

Synaptic plasticity

- makes connections stronger or weaker
- can be experimentally induced
- needs ‘joint activation’ of the two connected neurons
- is induced rapidly, but can last for a long time
- Spike-timing dependent plasticity is one of many protocols

Hebb rule:

- ‘neurons that fire together, wire together’

S. Loewl and W. Singer, Science 1992

‘Local rule’:

- only the activity of sending and receiving neurons matters

Previous slide.

There are several experimental paradigms to induce synaptic changes.
Most of these paradigms are consistent with the Hebb rule:
Neurons that fire together, wire together, a slogan that was introduced by Loewl and Singer in 1992.

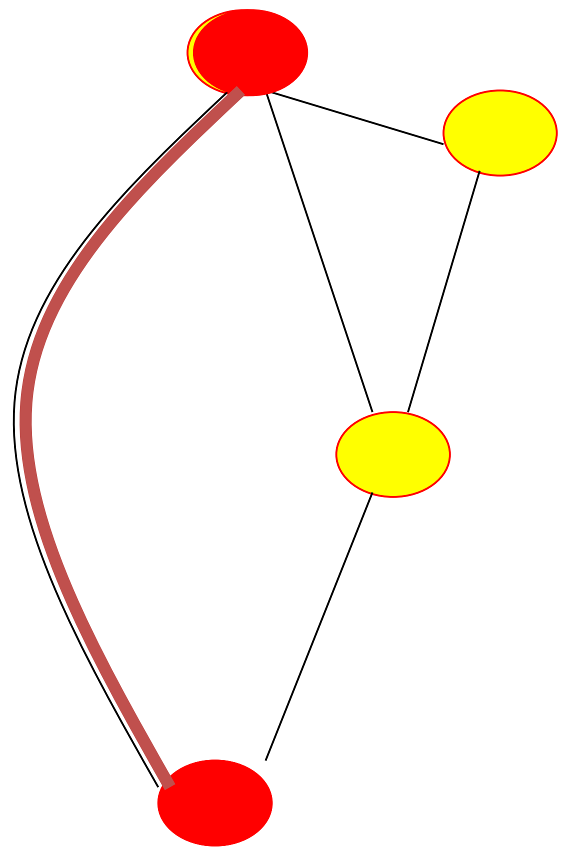
However, in all these Hebbian learning rules and their corresponding experimental paradigms, the role of reward is unclear and not considered.

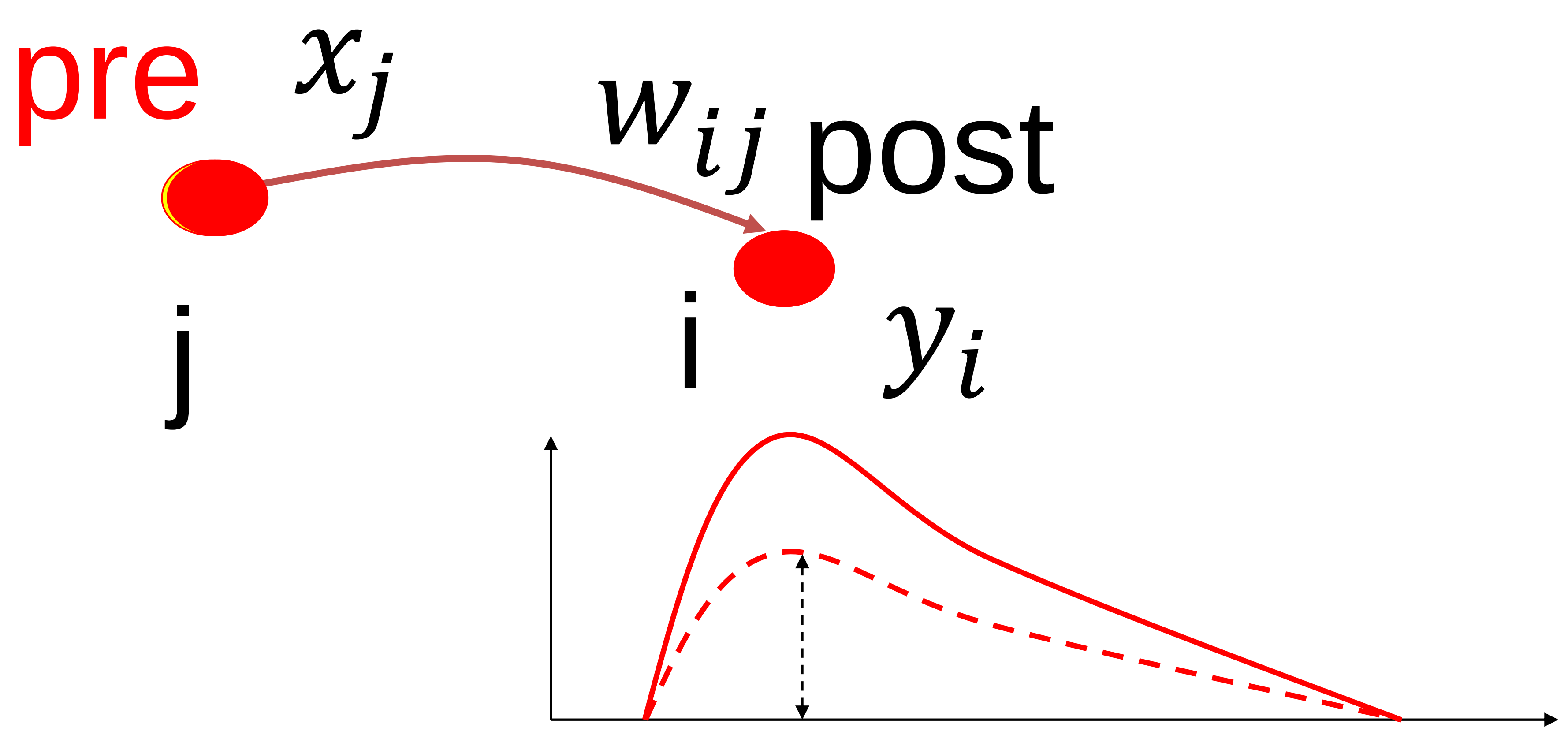
Hebbian rules are examples of ‘LOCAL’ learning rules.

- For the change of a connection from neuron j to neuron i, only the activity of these two neurons i and j matters, but not the activity of some other neuron k further away.
- Local means that only information that is locally available at the site of the synapse can be used to drive a weight change. What is available is the value of the weight itself, as well as the state of the postsynaptic neuron and the incoming spikes sent by the presynaptic neuron.

3. Synaptic Plasticity: depends on activity of pre and post

Hebbian Learning





$$w_{ij}\varepsilon\left(t-t_j^f\right)$$

- ‘local’ learning rule

- Changes depend on two factors: presynaptic and postsynaptic

$$\Delta w_{ij} = F(pre, post, w_{ij})$$

- Sensitive to conincidences ‘pre’ and ‘post’

$$\Delta w_{ij} = c\, x_j\, [y_i - b]$$

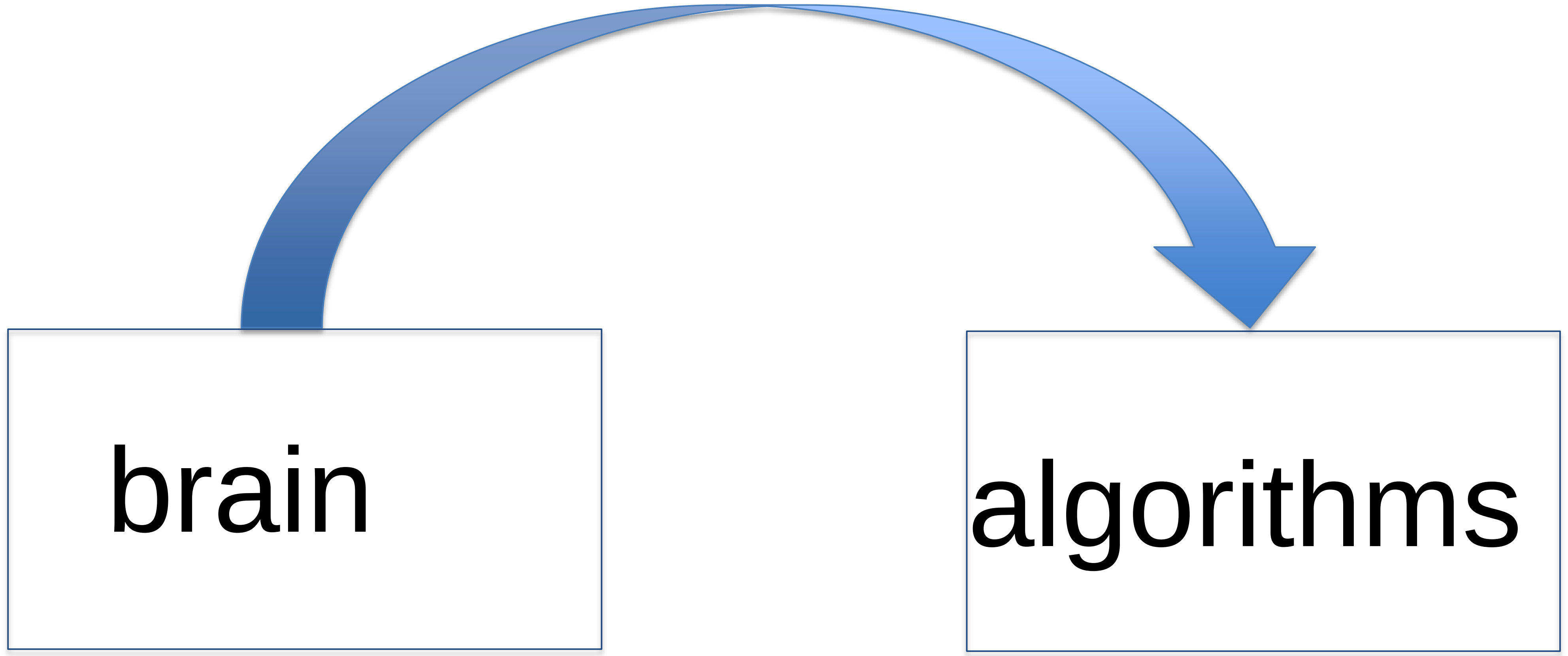
Previous slide.

In standard Hebbian learning, the change of the synaptic weight depends on presynaptic activity x_j (the presynaptic factor, pre) and the state of the postsynaptic neuron (for example $x_i - b$, an example of a postsynaptic factor, b is an arbitrary constant).

1. The rule is local: it depends only on information that is available at the synapse.
2. It is built from two factors: the multiplication of a presynaptic and a postsynaptic factor.
3. Note that it does not contain the notion of reward or success.

Now we want to see whether such rules can be mapped to the math we did in this class!

Learning Rules



brain

algorithms

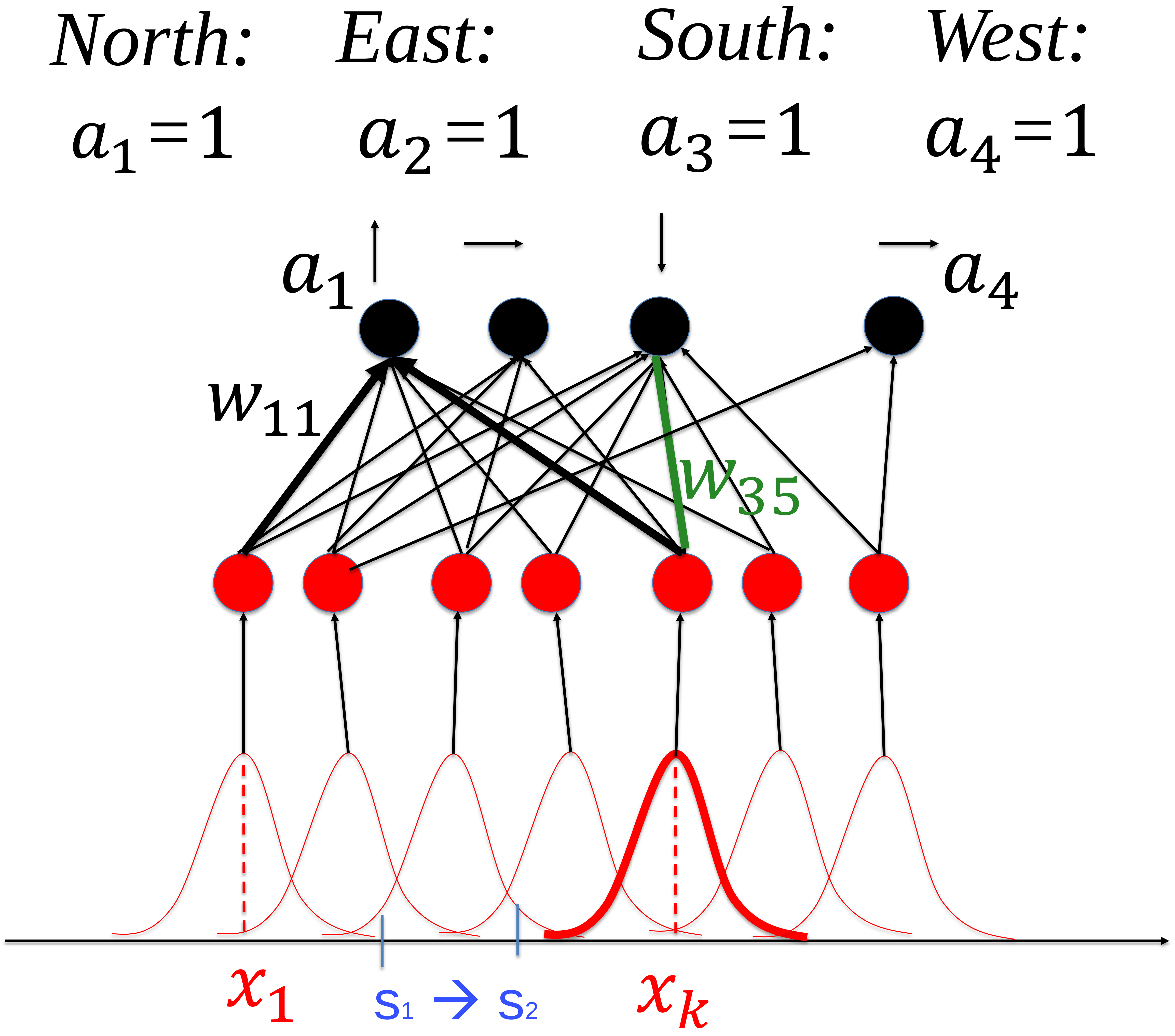
“Can the brain implement policy gradient?”

Previous slide.

After this introduction into the learning rules of the brain, let us now ask the following question:

Is the learning rule of policy gradient with softmax output consistent with what we know about learning rules in the brain

Policy gradient rule – can we interpret this as local rule?



Discrete actions with
1 hot coding

If at time *t*, the action
$a_i^t = 1$ is chosen then
$a_j^t = 0$ for all other
output neurons $j \neq i$

Action choice:
Softmax

(previous slide)

1. The policy is softmax:

this implies that output neurons interact interact such that the policy $\pi(a_i^t = 1|\vec{x})$ is normalized

$$\sum_i \pi(a_i^t = 1|\vec{x}) = 1$$

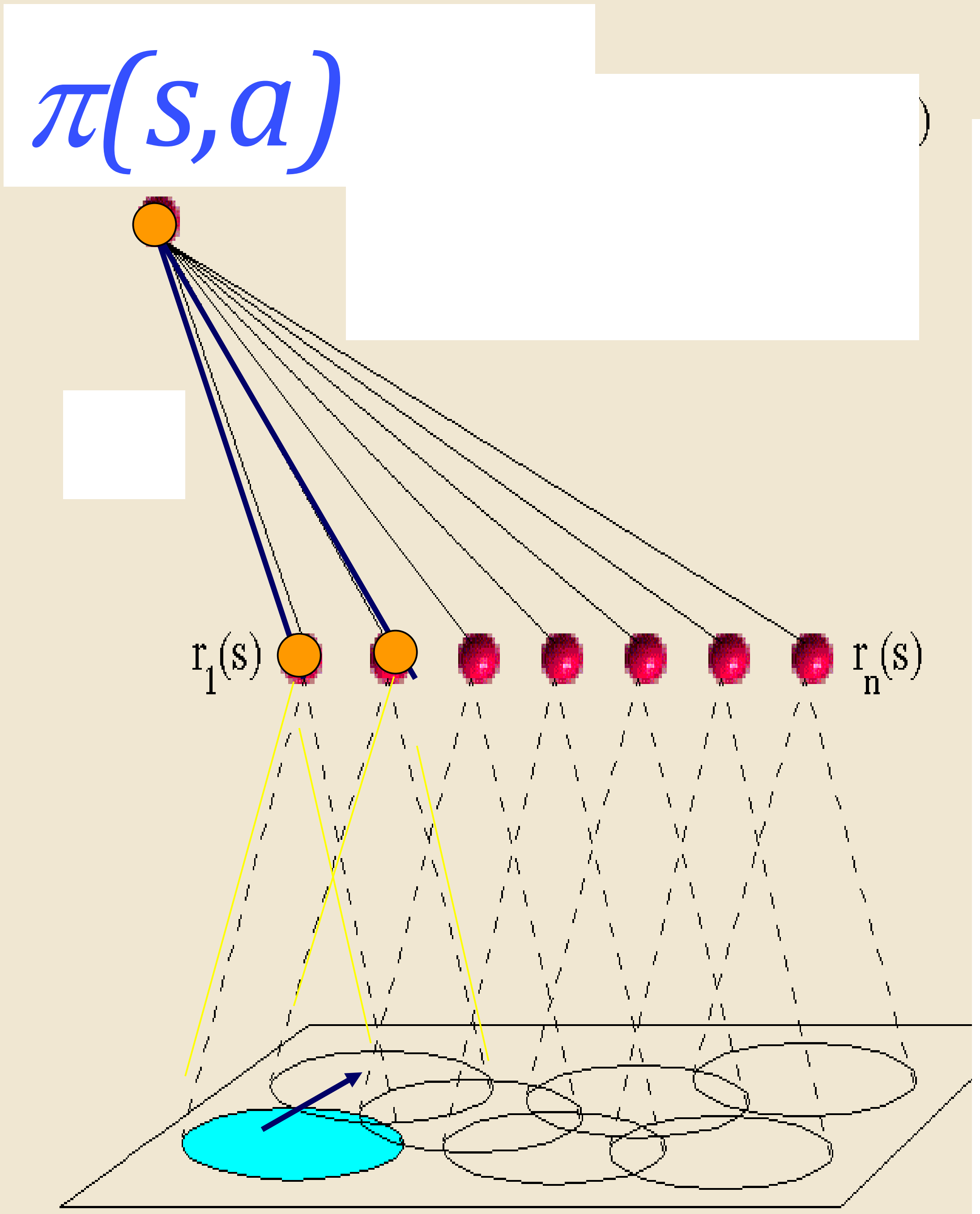
2. The coding is 1-hot:

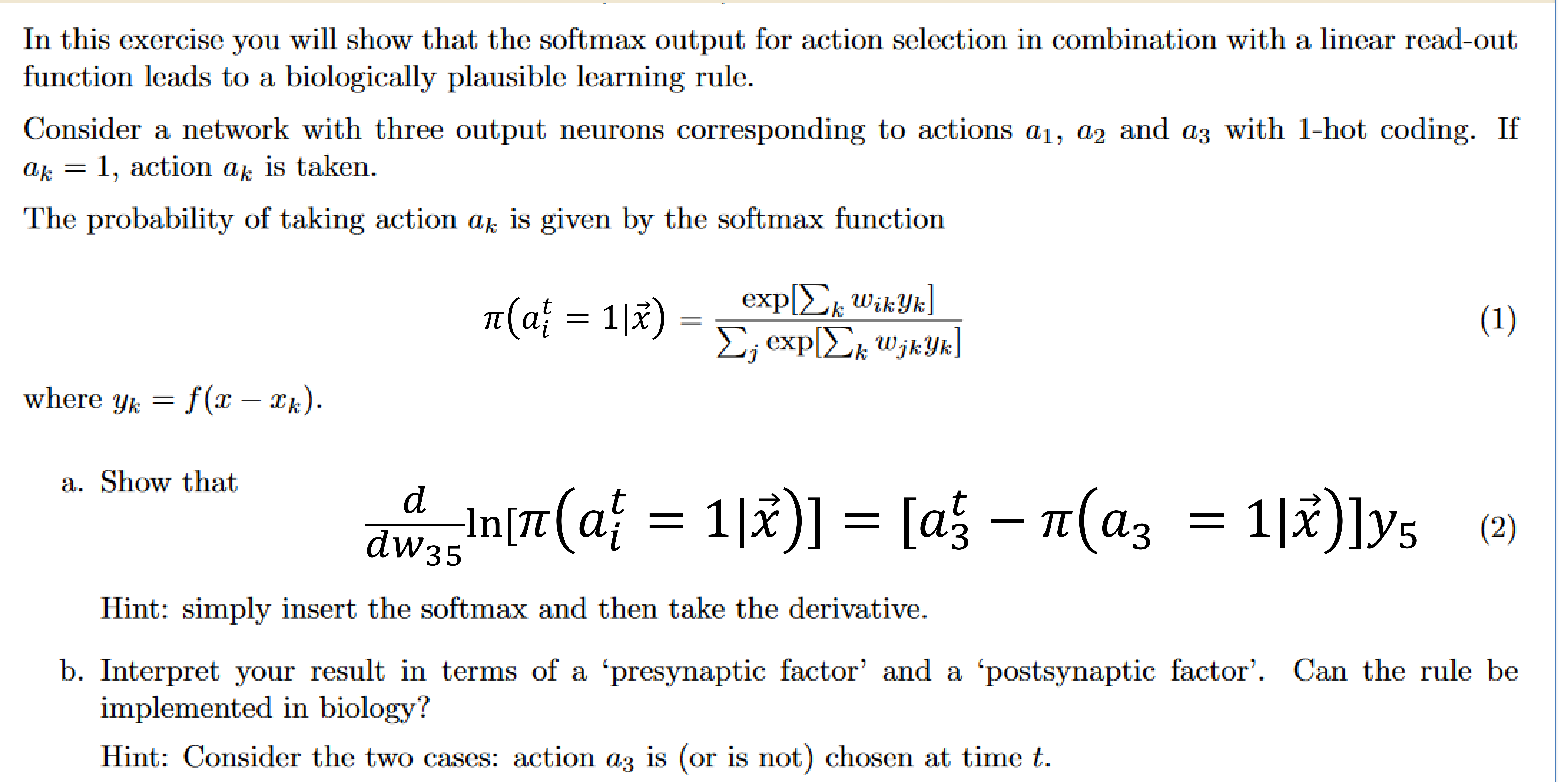
This implies that if at time t, the action $a_i^t = 1$ is chosen then neuron i sends immediately an output signal to all other neurons to inhibit their activity so that $a_j^t = 0$ for all other output neurons $j \neq i$.

Exercise: Continuous input representation

Take 15 minutes

Lecture continues at … 12h15





In this exercise you will show that the softmax output for action selection in combination with a linear read-out function leads to a biologically plausible learning rule.

Consider a network with three output neurons corresponding to actions a_1, a_2 and a_3 with 1-hot coding. If $a_k = 1$, action a_k is taken.

The probability of taking action a_k is given by the softmax function

$$\pi(a_i^t = 1|\vec{x}) = \frac{\exp[\sum_k w_{ik}y_k]}{\sum_j \exp[\sum_k w_{jk}y_k]} \qquad (1)$$

where $y_k = f(x - x_k)$.

a. Show that

$$\frac{d}{dw_{35}}\ln[\pi(a_i^t = 1|\vec{x})] = [a_3^t - \pi(a_3 = 1|\vec{x})]y_5 \qquad (2)$$

Hint: simply insert the softmax and then take the derivative.

b. Interpret your result in terms of a ‘presynaptic factor’ and a ‘postsynaptic factor’. Can the rule be implemented in biology?

Hint: Consider the two cases: action a_3 is (or is not) chosen at time t.

c. Go back to the advantage actor critic (slide 7), without eligibility trace and add ‘reward’. What is the resulting learning rule for parameter update?

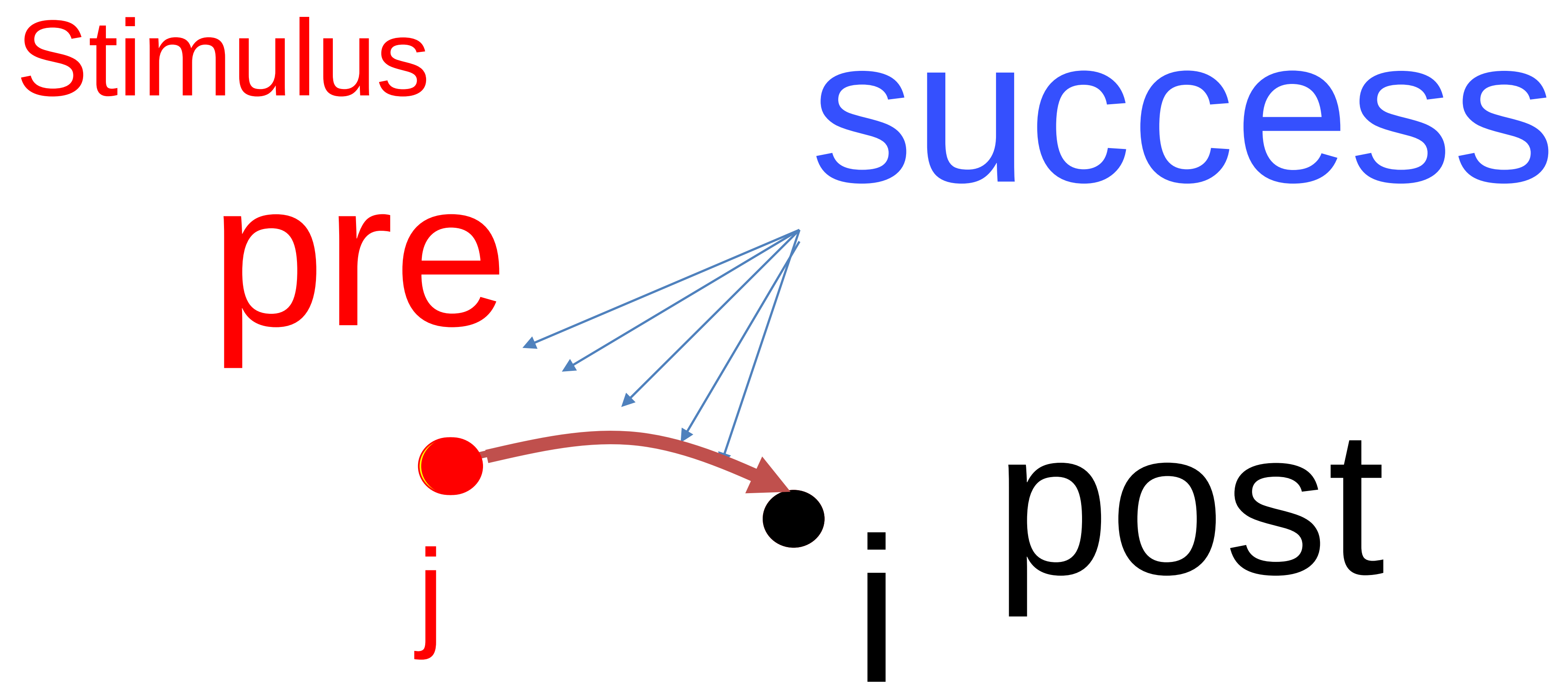
d. Add an eligibility trace to your result in c). How would you implement this rule in hardware?

Note:
one hidden layer;
only output weights
are learned

(previous slide)
Your notes.

Discussion of Exercise: Comparison with Biology

parameter = weight w_{ij}



Change depends on pre and post

Three factors: **success**

$$\Delta w_{ij} = \eta \quad S(a_i^t, \vec{x}) \overbrace{[\, a_i^t - \langle a_i(\vec{x}) \rangle]}^{\textbf{post}} \overset{\textbf{pre}}{x_j}$$

postsynaptic factor is

'activity – expected activity'

Previous slide.

Reinforcement Learning includes a set of very powerful algorithm – as we have seen in previous lectures. Here S denotes the success, which is reward (in REINFORCE) or reward minus baseline (in REINFORCE with baseline), or TD error (in the advantage actor-critic)

For today the big question is:
Is the structure of the brain suited to implement reinforcement learning algorithms?
If so which one? Q-learning or SARSA? How about Policy gradient?
Is the brain architecture compatible with an actor-critic structure?

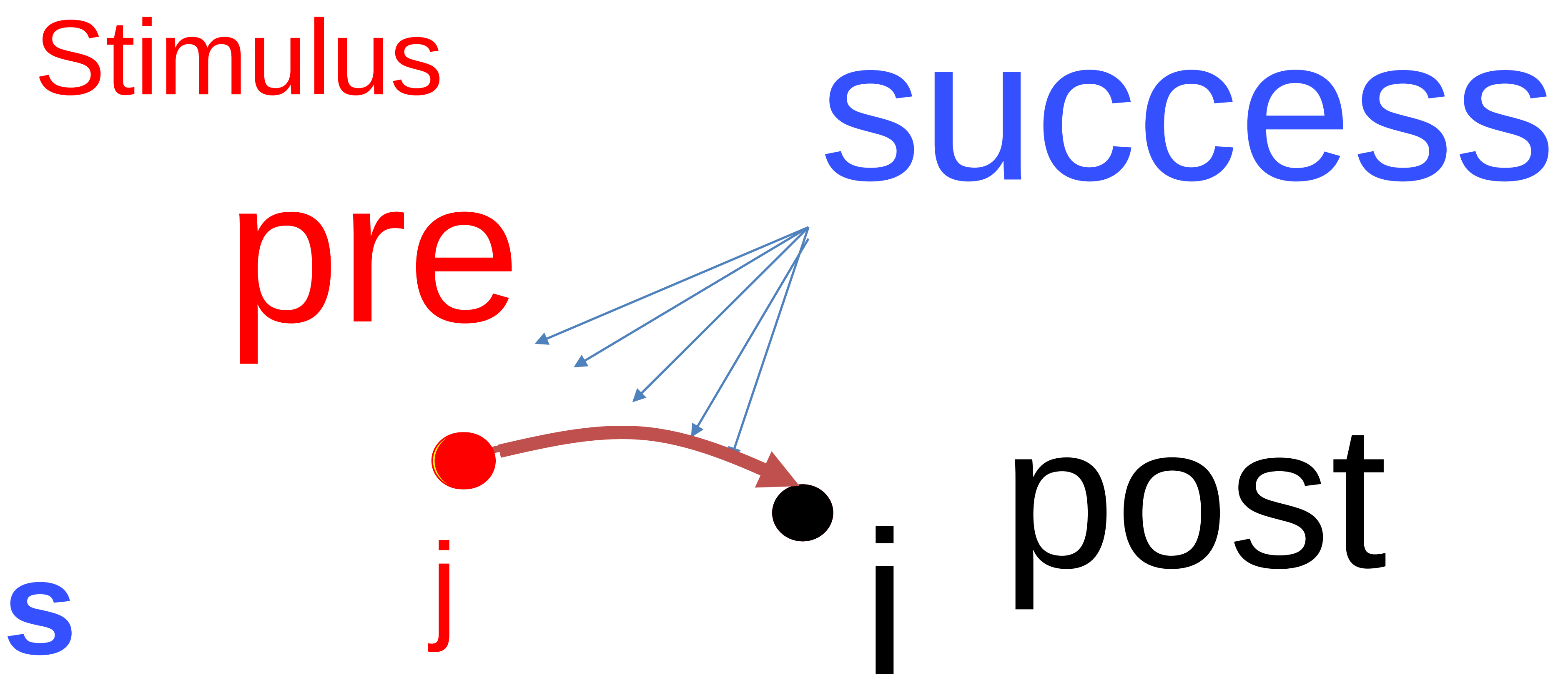
Could the brain implement backprop?

These are the questions we will address in the following.
And to do so, we have to first get a big of background information on brain anatomy.

Three-factor rule

Change depends

- Local factor **pre**
- Local factor **post**
- Global broadcast factor **success**
- **Success could be reward or TD error**



Stimulus

pre

success

j

i

post

Three factors: **success** **post** **pre**

$$\Delta w_{ij} = \eta \quad S(a_i^t,\vec{x}) \overbrace{[\, a_i^t - \langle a_i(\vec{x}) \rangle]} x_j$$

postsynaptic factor is

'activity – expected activity'

Previous slide.

The result of Reinforcement Learning with an actor-critic leads to a three-factor rule:

- A presynaptic factor, activity of the sending neuron, such as spike arrival at the synapse.
- A postsynaptic factor: its activity (output spikes, a=1 or inactive a=0) minus the ‘mean drive’ for this state $\langle y_i(\vec{x}) \rangle = \pi(a_i|\vec{x}\ )$
- In addition to the above two local factor (similar to a Hebb rule) there is one global broadcasting factor. The success.
- The success could be the reward itself (REINFORCE algorithm), or reward minus baseline (REINFORCE with state-dependent baseline), or the TD signal (advantage actor critic).

Artificial Neural Networks and RL : Lecture 13

Wulfram Gerstner
EPFL, Lausanne, Switzerland

from brain-computing to neuromorphic computing

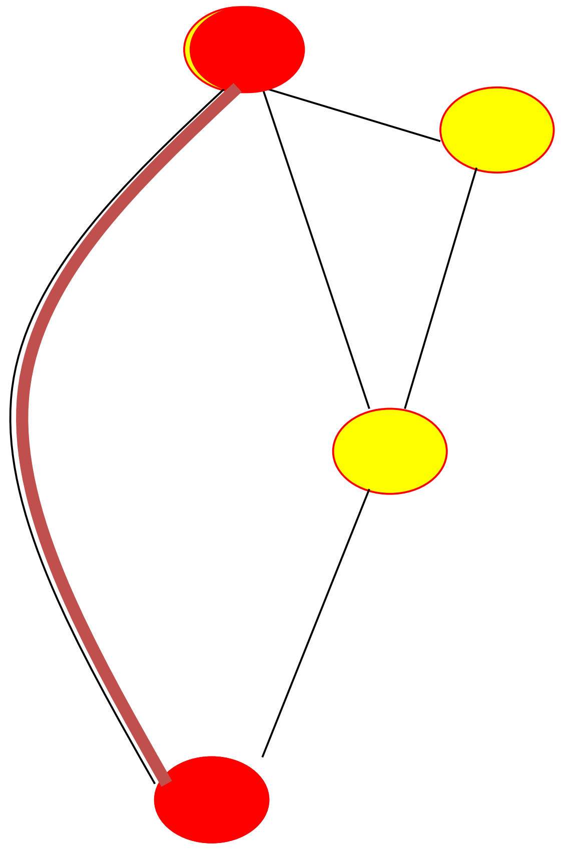
1. Coarse Brain Anatomy
2. Synaptic Plasticity
3. **Three-factor Learning Rules vs. 2-factor rules**

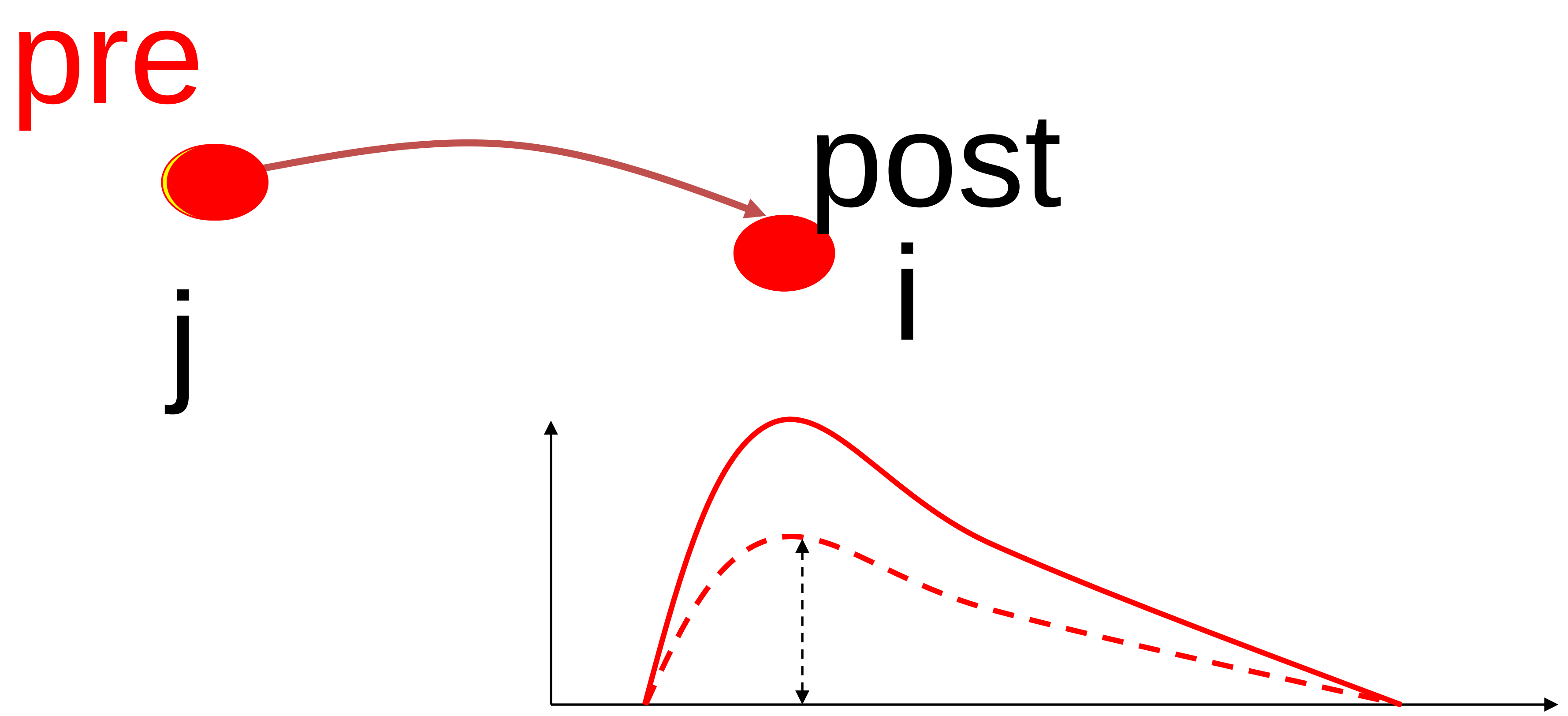
Previous slide.

Since Hebbian learning rules are limited, we have to extend the framework and include a ‘third factor’ that could represent reward.

3. Classification of synaptic changes: unsupervised learning

Hebbian Learning
= unsupervised learning





$$w_{ij}\varepsilon\left(t-t_j^f\right)$$

$$\Delta w_{ij} = F\left(pre, post, w_{ij}\right)$$

Previous slide.

In standard Hebbian learning, the change of the synaptic weight depends only on presynaptic activity (pre) and the state of the postsynaptic neuron (post). The rule is local, and does not contain the notion of reward or success.

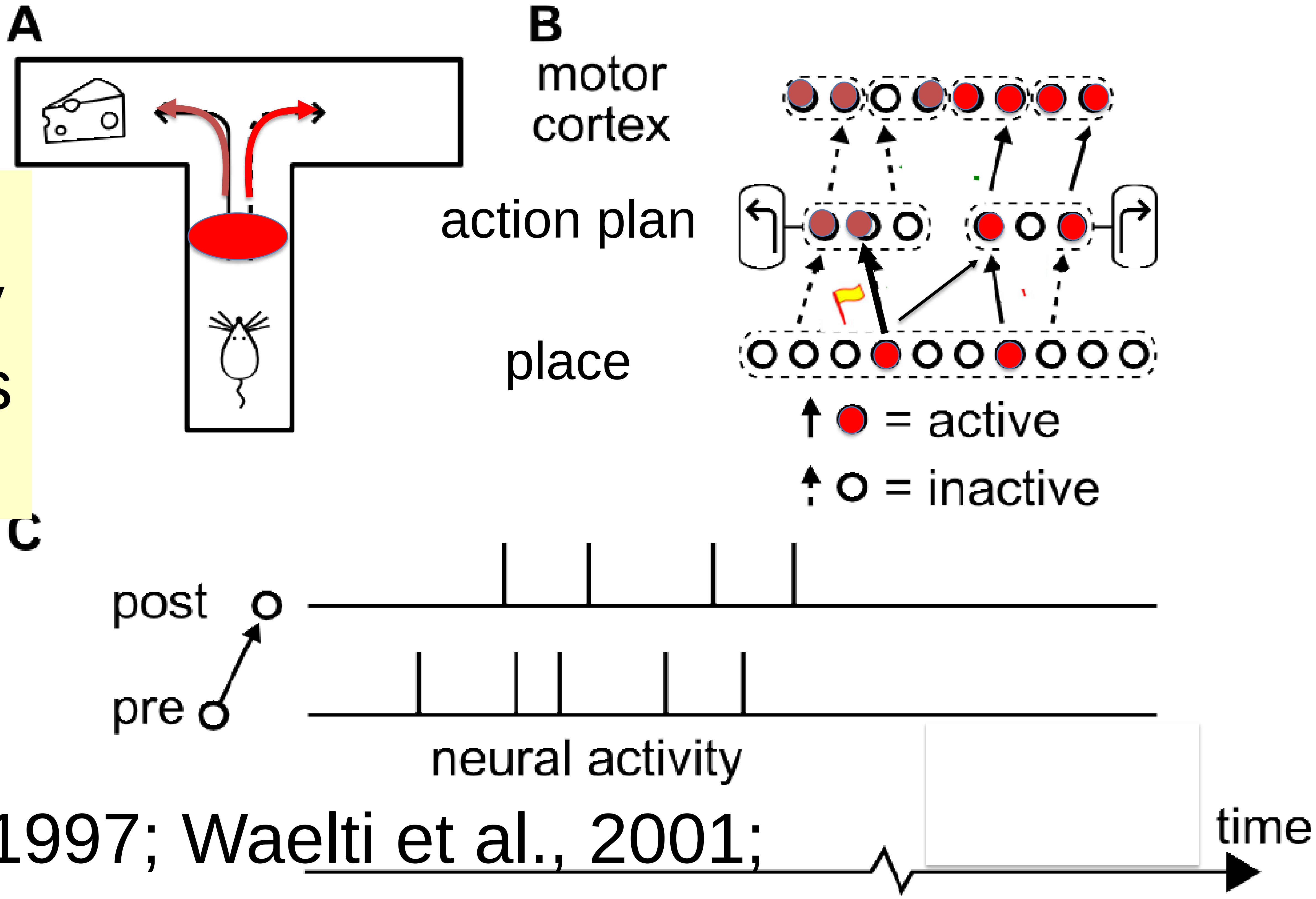
The value of the weight w_{ij} is measured by sending a test-pulse across the synapse.

Is Hebbian Learning sufficient? No!

Image: Fremaux and Gerstner, Front. Neur. Circ., 2015

Eligibility trace:
Synapse keeps memory of pre-post coincidences over a few seconds

Dopamine:
Reward/success



Schultz et al. 1997; Waelti et al., 2001;

→ **Reinforcement learning: success = reward – (expected reward)**

TD-learning, SARSA, Policy gradient (book: Sutton and Barto, 2018)

Previous slide.

Hebbian learning as it stands is not sufficient to describe learning in a setting were rewards play a role. If joint activity of pre- and post causes stronger synapses, the rat is likely to repeat the same unrewarded action a second time.

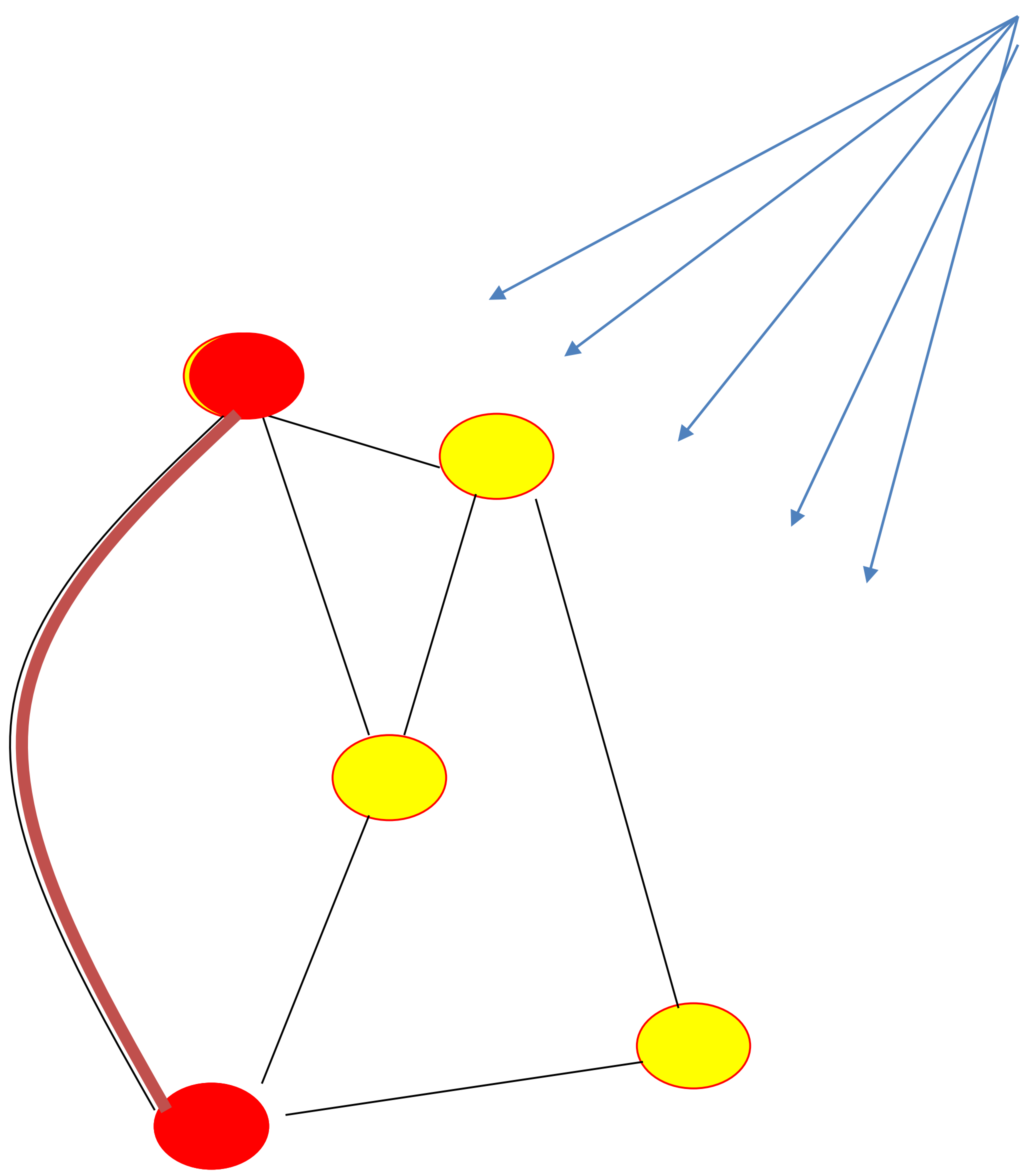
Hypothetical functional role of neuromodulated synaptic plasticity. Reward-modulated learning

(A) Schematic reward-based learning experiment.Ananimal learns to perform a desired sequence of actions(e.g.,move straight,then turn left) in a T-maze through trial-and-error with rewards (cheese)

(B) The current position (“place”) of the animal in the environment is represented by an assembly of active cells in the hippocampus.These cells feed neurons (e.g.,in the dorsal striatum) which code for high-level actions at the choice point,e.g., “turn left” or “turn right.” These neuronsin turn project to motorcortex neurons,responsible for the detailed implementation of actions. A success signal modulates (green arrows) the induction of plasticity

(C) Neuromodulator timing. While spikes occur on the time scale of milliseconds, the success signal may come a few seconds laters.

3. Classification of synaptic changes: Reinforcement Learning



SUCCESS

Reinforcement Learning
= reward + Hebb

$$\Delta w_{ij} \propto F(pre, post, SUCCESS)$$

local global

broadly diffused signal:
neuromodulator

Previous slide.

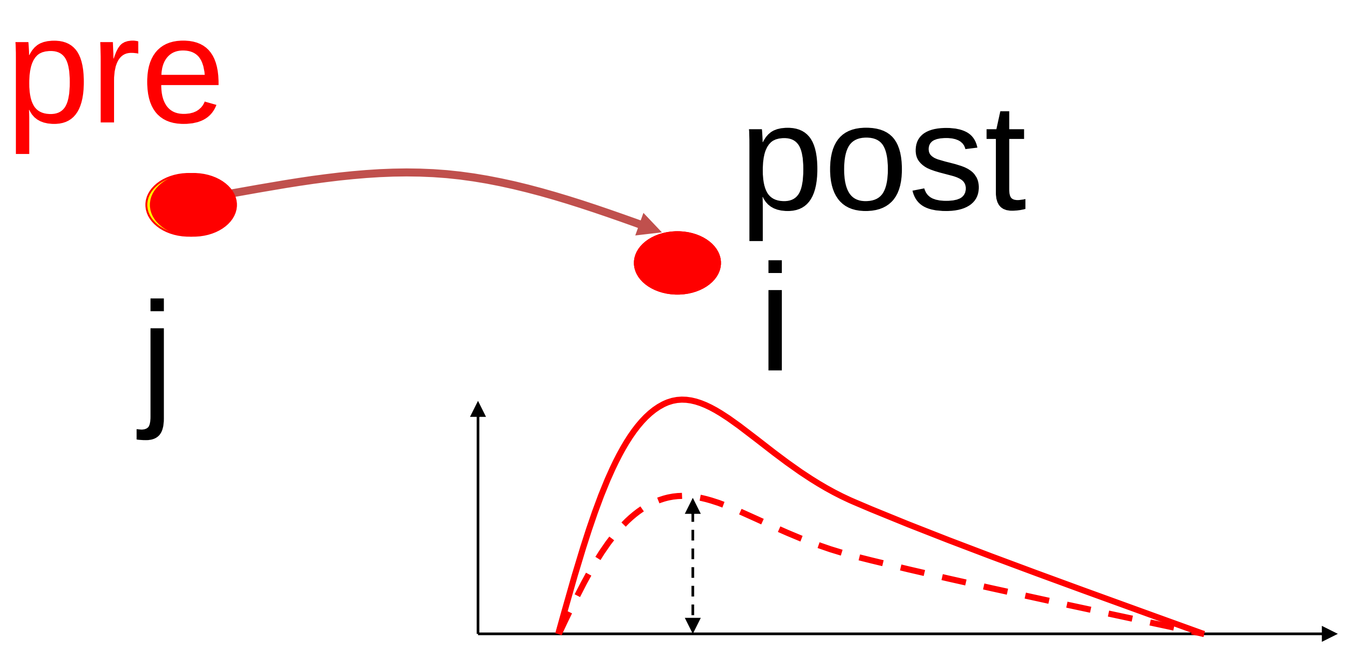
For the moment we say the reinforcement learning depends on three factors: the Hebbian pre- and postsynaptic factor plus a success signal related to reward.
We will get more precise later.

3. Classification of synaptic changes

unsupervised vs reinforcement

LTP/LTD/Hebb
Theoretical concept

- passive changes
- exploit statistical correlations

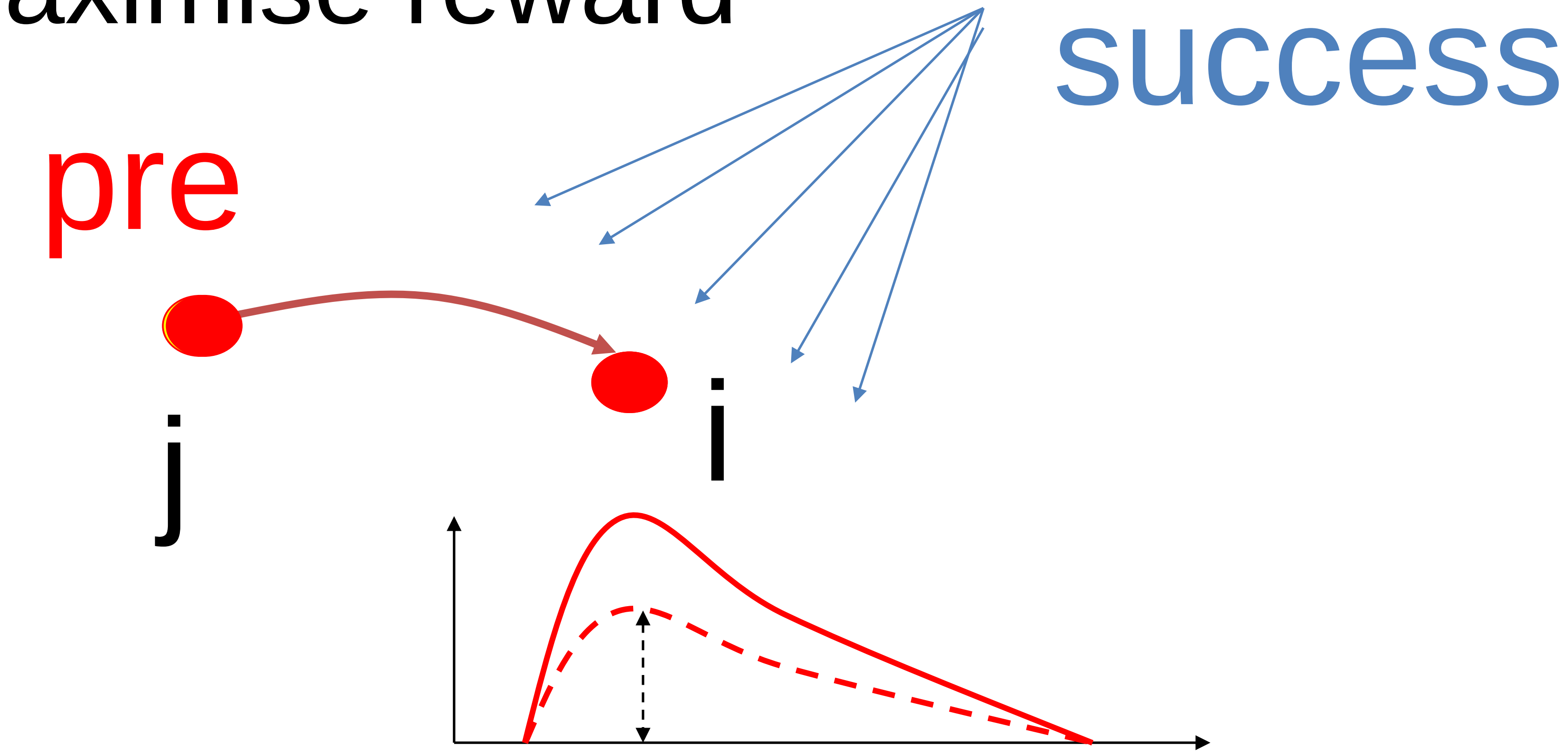


Functionality

-useful for development
(develop good filters)

Reinforcement Learning
Theoretical concept

- conditioned changes
- maximise reward



Functionality

- useful for learning
a new behavior

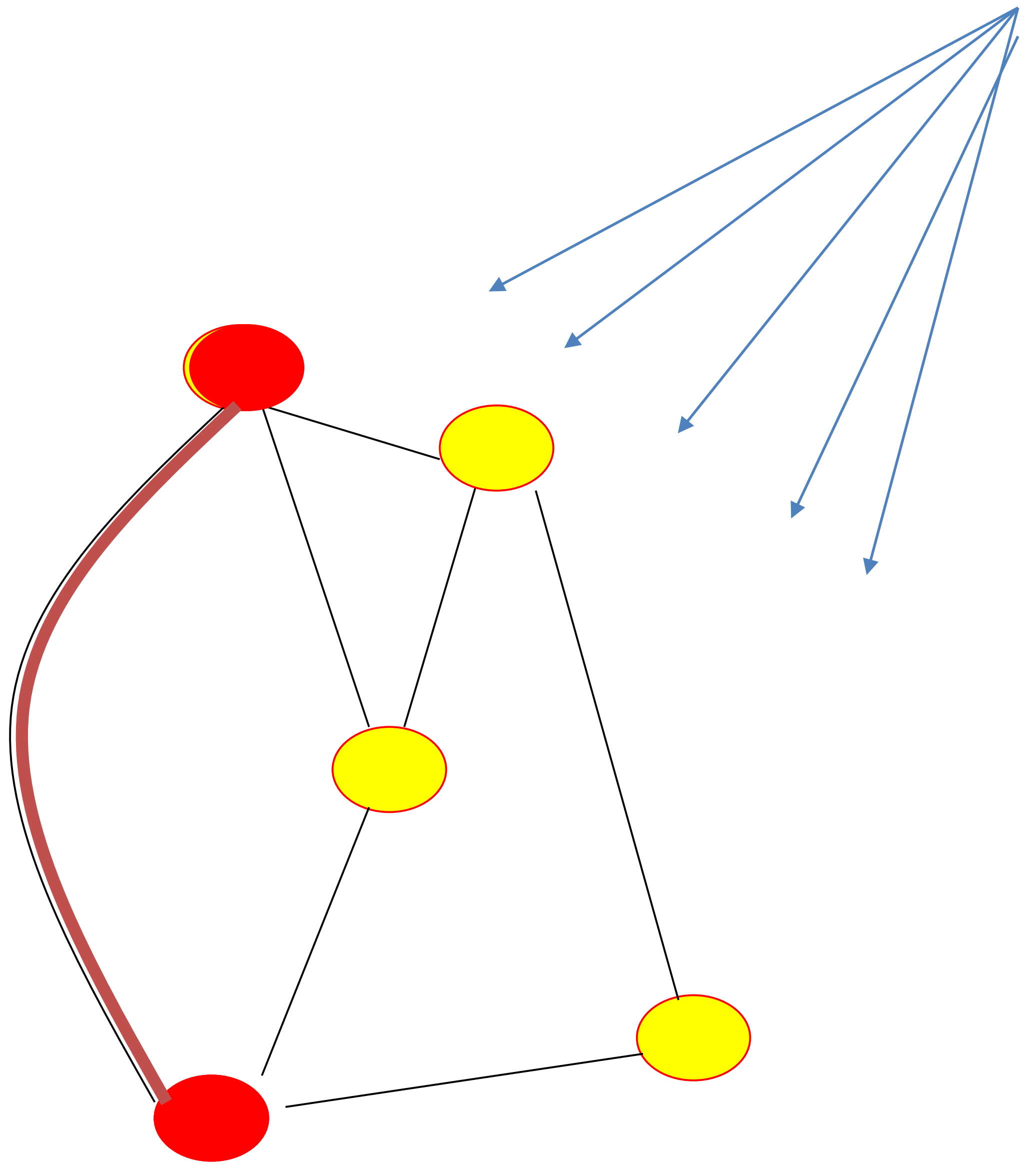
Previous slide.

This does not mean the standard Hebbian learning is wrong: in fact it is very useful for the development of generic synaptic connections, e.g., to make neurons develop good filtering properties that pick up relevant statistical signals in the stream of input.

The three-factor rules are relevant for learning novel behaviors via feedback through reward.

3. Three-factor rule of Hebbian Learning

= Hebb-rule gated by a neuromodulator



Neuromodulators: Interestingness, surprise; attention; novelty

$$\Delta w_{ij} \propto F(\overset{\uparrow}{pre}, \overset{\uparrow}{post}, \overset{\uparrow}{MOD})$$

local global

Previous slide.

The three-factor rules have a Hebbian component: pre- and postsynaptic activity together, but in addition the third factor which is related to neuromodulators.

There are several neuromodulators in the brain

Neuromodulator projections

- 4 or 5 neuromodulators
- near-global action

Dopamine/reward/TD:
Schultz et al., 1997,
Schultz, 2002

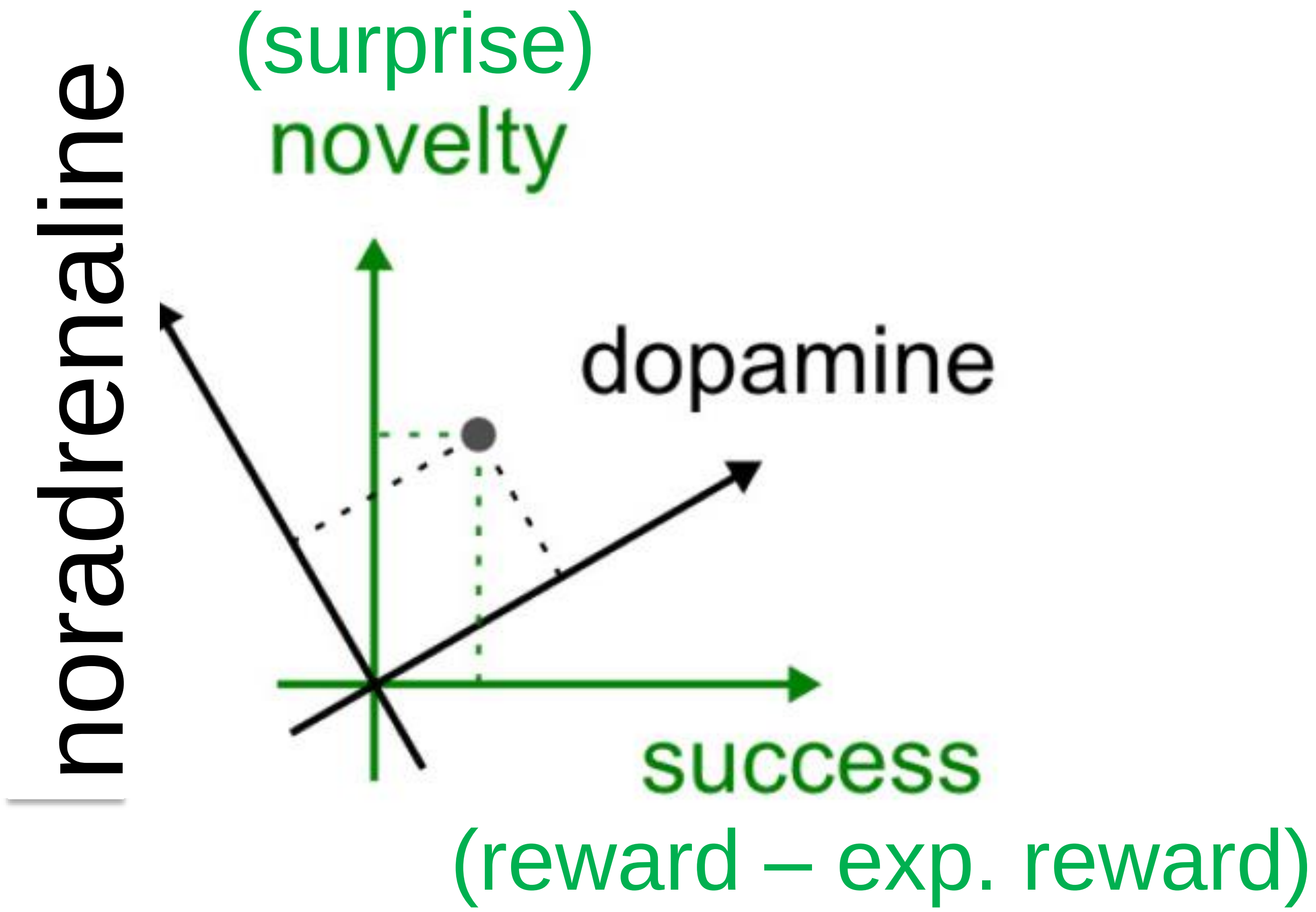
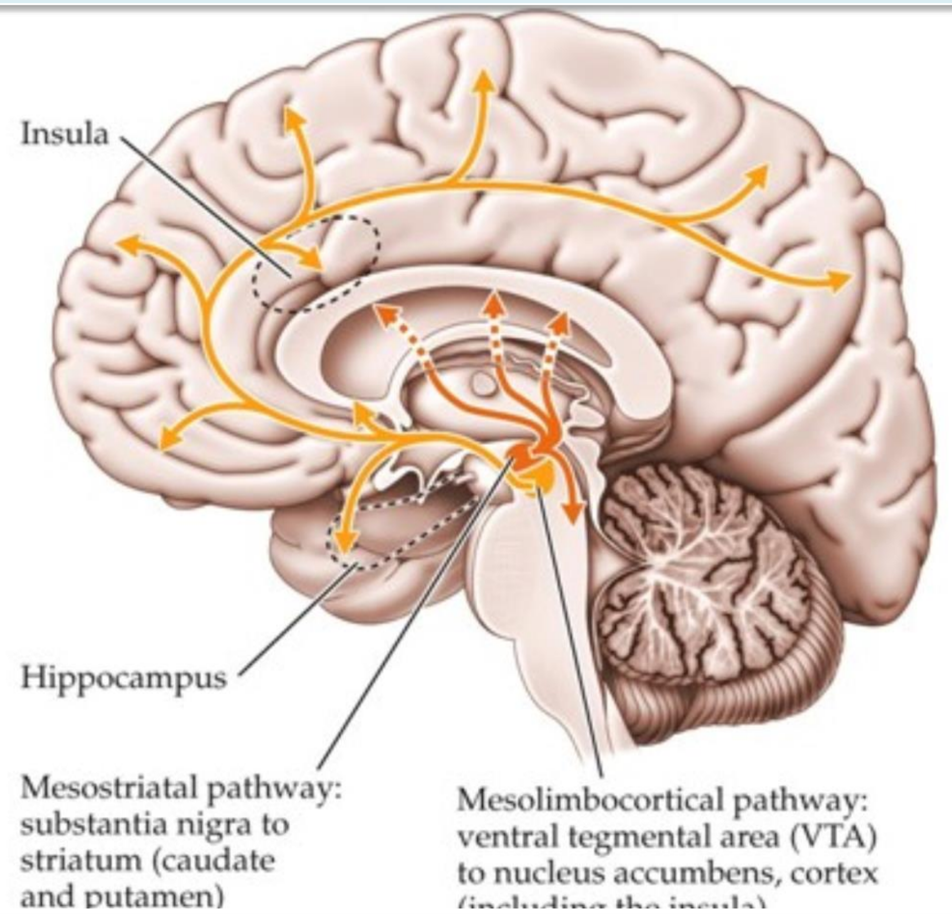


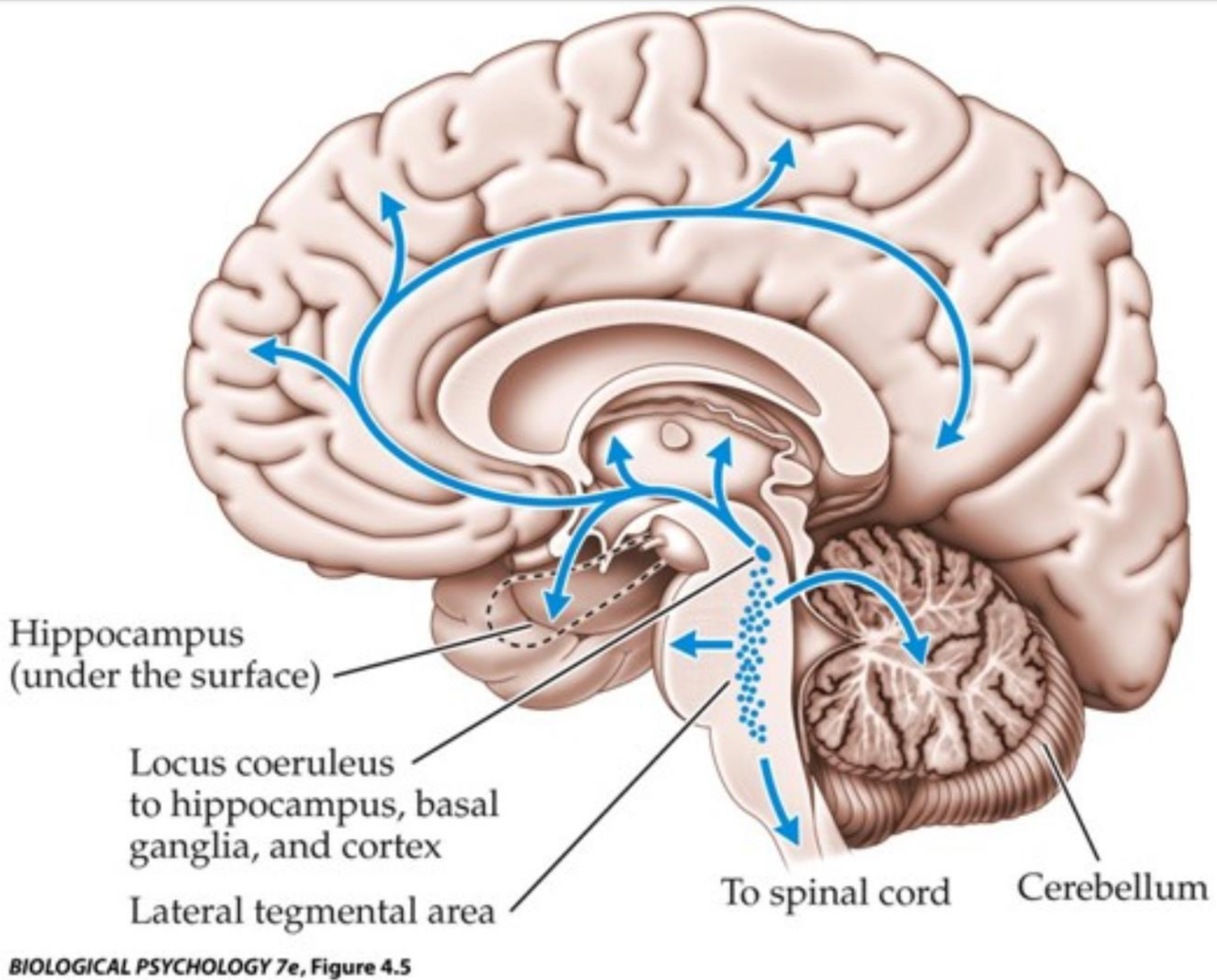
Image:
Fremaux and Gerstner, Frontiers (2016)

Image: Biological Psychology, Sinauer

Dopamine (DA)


Insula
Hippocampus
Mesostriatal pathway: substantia nigra to striatum (caudate and putamen)
Mesolimbocortical pathway: ventral tegmental area (VTA) to nucleus accumbens, cortex (including the insula)

Noradrenaline (NE)


Hippocampus (under the surface)
Locus coeruleus to hippocampus, basal ganglia, and cortex
Lateral tegmental area
To spinal cord
Cerebellum
BIOLOGICAL PSYCHOLOGY 7e, Figure 4.5

Previous slide.

The most famous neuromodulator is dopamine (DA) which is related to reward, as we will see.

But there are other neuromodulators such as noradrenaline (also called norepinephrine, NE) which is related to surprise.

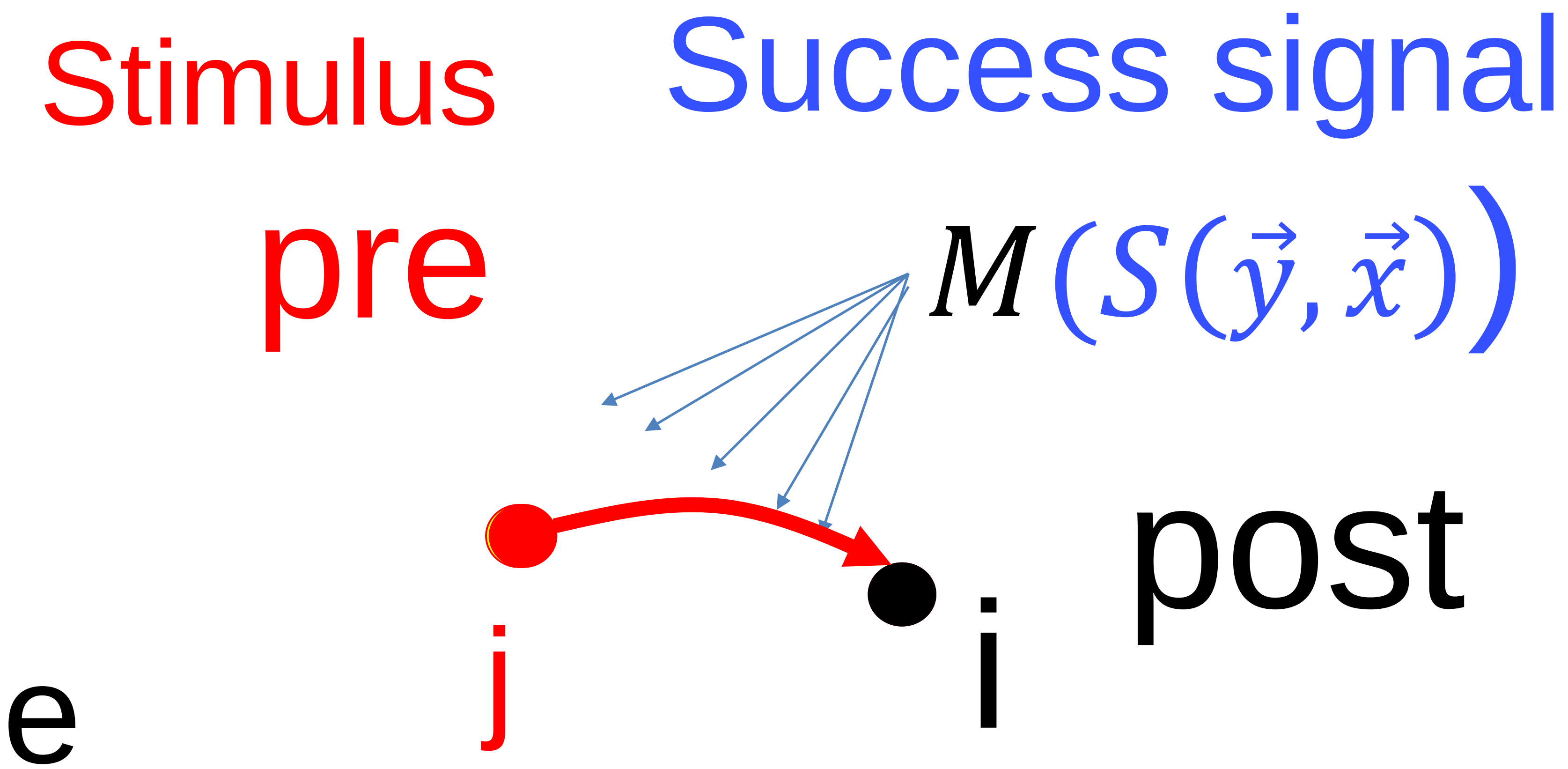
Left: the mapping between neuromodulators and functions is not one-to-one. Indeed, dopamine also has a ‘surprise’ component.

Right: most neuromodulators send axons to large areas of the brain, in particular to several cortical areas. The axons branch out in thousands of branches.
Thus the information transmitted by a neuromodulator arrives nearly everywhere.
In this sense, it is a ‘global’ signal, available in nearly all brain areas.

3. Formalism of Three-factor rules with eligibility trace

x_j = activity of presynaptic neuron

y_i = activity of presynaptic neuron



Step 1: co-activation sets eligibility trace

$$\Delta z_{ij} = \eta \; f(y_i)\, g(x_j)$$

Step 2: eligibility trace decays over time

$$z_{ij} \leftarrow \lambda \, z_{ij}$$

Step 3: eligibility trace translated into weight change

$$\Delta w_{ij} = \eta \, M\big(S(\vec{y},\vec{x})\big) z_{ij}$$

Previous slide.

Three-factor rules are implementable with eligibility traces.

The joint activation of pre- and postsynaptic neuron sets a ‘flag’. This step is similar to the Hebb-rule, but the change of the synapse is not yet implemented.

The eligibility trace decays over time

However, if a neuromodulatory signal M arrives before the eligibility trace has decayed to zero, an actual change of the weight is implemented.

The change is proportional to

- the momentary value of the eligibility trace
- the value of the success signal

The success signal can be broadcasted by a neuromodulator the signals

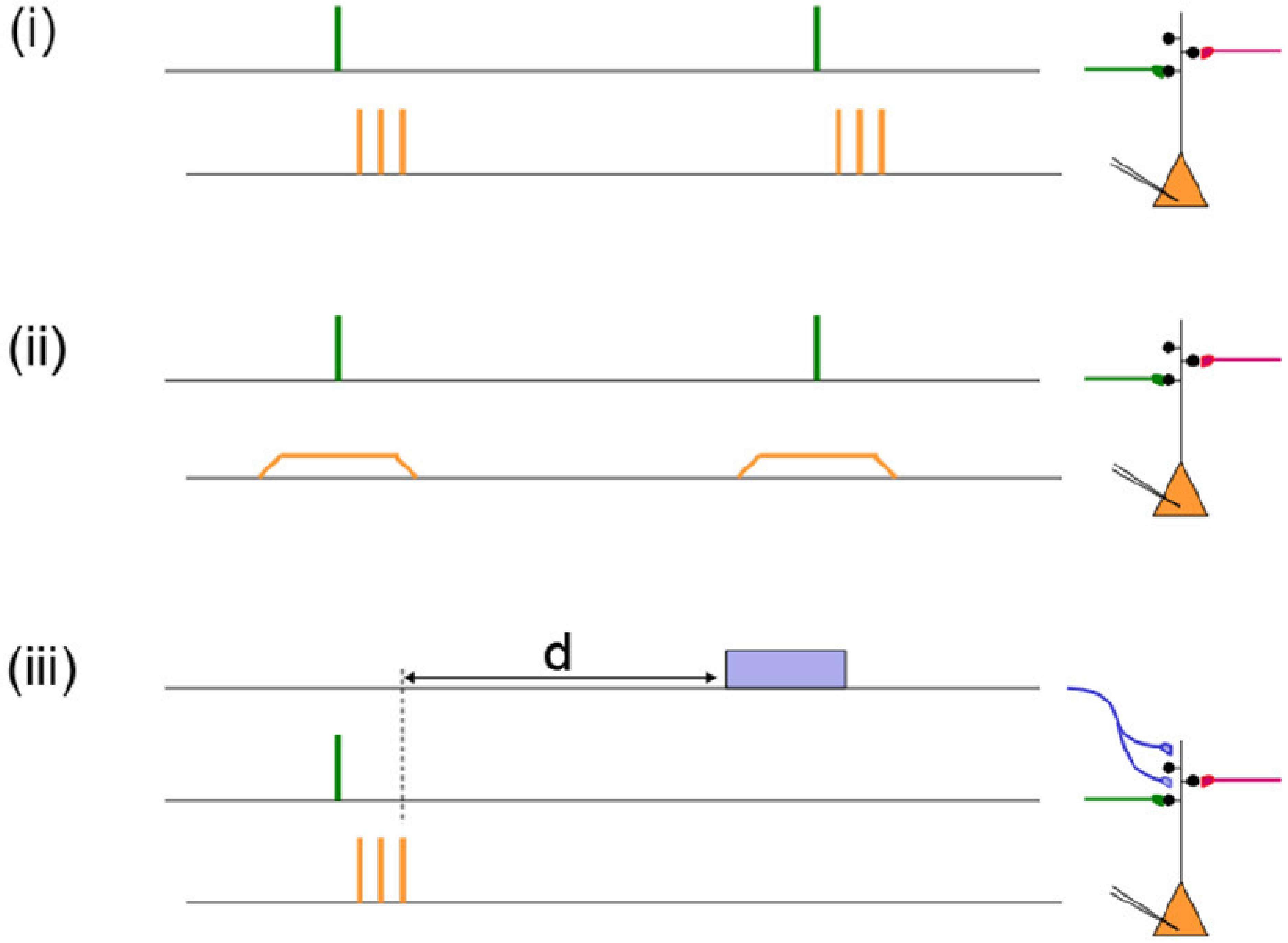
- Reward minus reward-baseline OR
- TD-error

3. Hebbian LTP versus Three-factor rules

Hebbian coactivation:
pre-post-post-post

Hebbian coactivation:
but no post-spikes

Scenario of three-factor rule: Hebb+modulator



Neuromodulator can come with a delay of 1s -

Image: Gerstner et al. (2018, review paper in Frontiers)

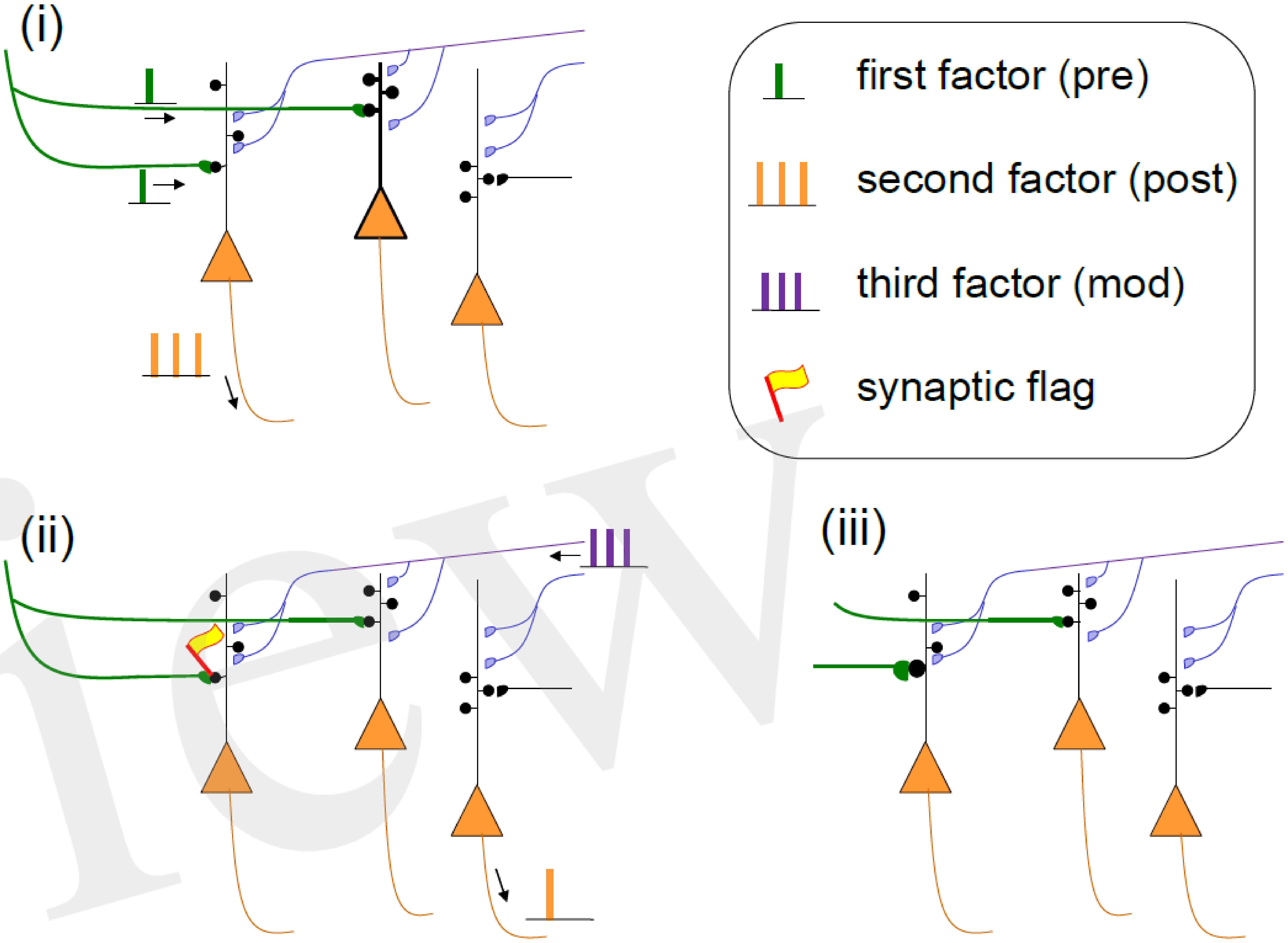
Previous slide.

The joint activation of pre- and postsynaptic neuron sets a ‘flag’. This step is similar to the Hebb-rule, but the change of the synapse is not yet implemented.
Note that joint activation can imply spikes of pre- (green) and postsynaptic (orange) neuron (top);
Or spikes of a presynaptic neuron combined with a weak voltage increase in the postsynaptic neuron (middle).

Bottom: three-factor rule only if a neuromodulatory signal M arrives before the eligibility trace has decayed to zero, an actual change of the weight is implemented. The neuromodulater arrives through the branches

The ideas of three-factor rules can be traced back over several decades.
Early papers were Crow 1968, Barto, 1983/1985, Schultz 1997,
First experimental papers Schultz 1997

3. Three-factor rules: synaptic flags and delayed reward (mod)



synaptic flag plays role of eligibility trace

Fig: Gerstner et al. 2018

Previous slide.

Specificity of three-factor learning rules.

(i) Presynaptic input spikes (green) arrive at two different neurons, but only one of these also shows postsynaptic activity (orange spikes).

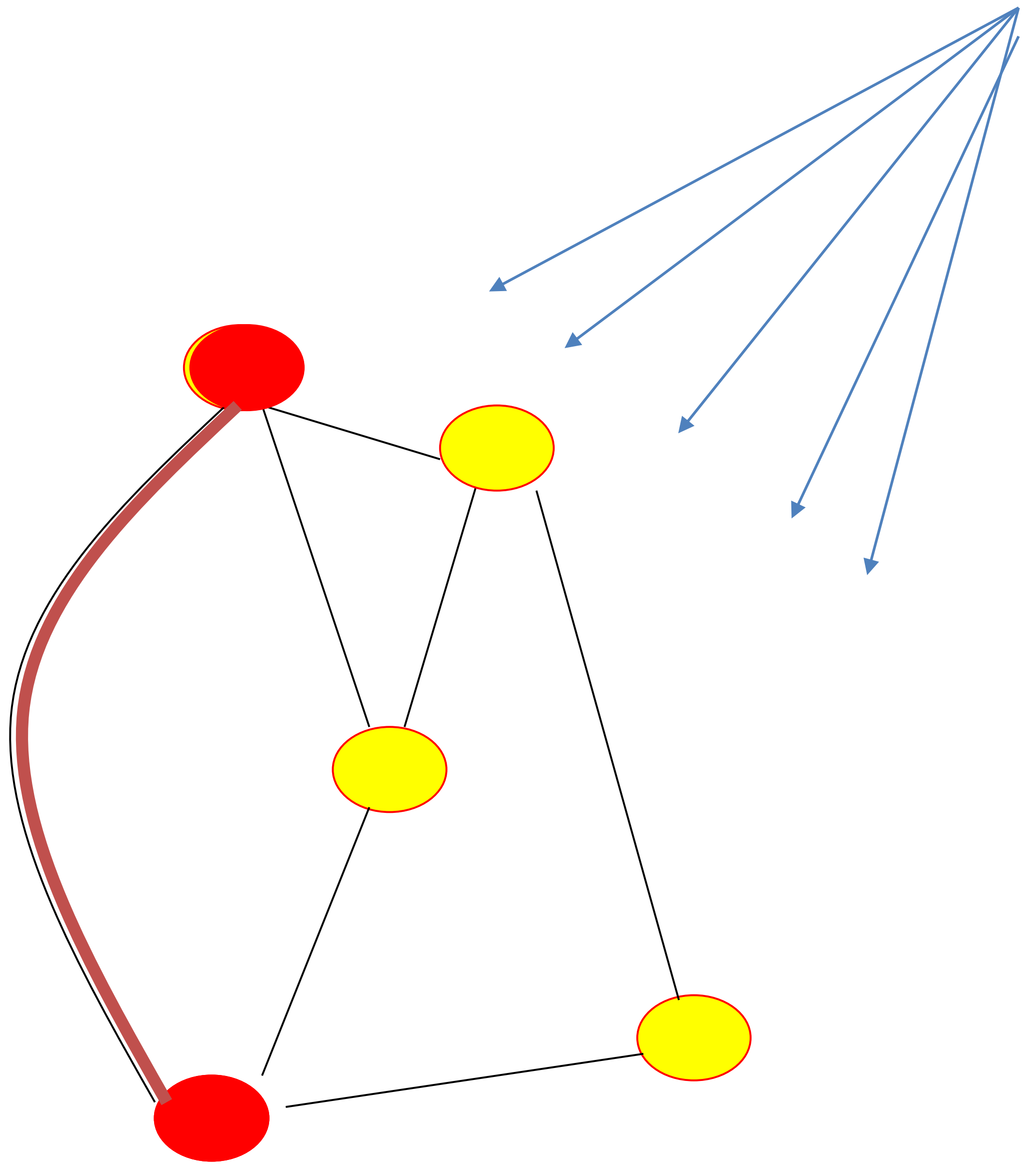
(ii) A synaptic flag is set only at the synapse with a Hebbian co-activation of pre- and postsynaptic factors; the synapse become then eligible to interact with the third factor (blue). Spontaneous spikes of other neurons do not interfere.

(iii) The interaction of the synaptic flag (eligibility trace) with the third factor leads to a strengthening of the synapse (green).

Fig caption: Gerstner et al. 2018

3. Recent experiments for Three-factor rules

Neuromodulators for reward, interestingness, surprise; attention; novelty



Step 1: co-activation sets eligibility trace

Step 2: eligibility trace decays over time

Step 3: delayed neuro-Modulator:
eligibility trace translated into weight change

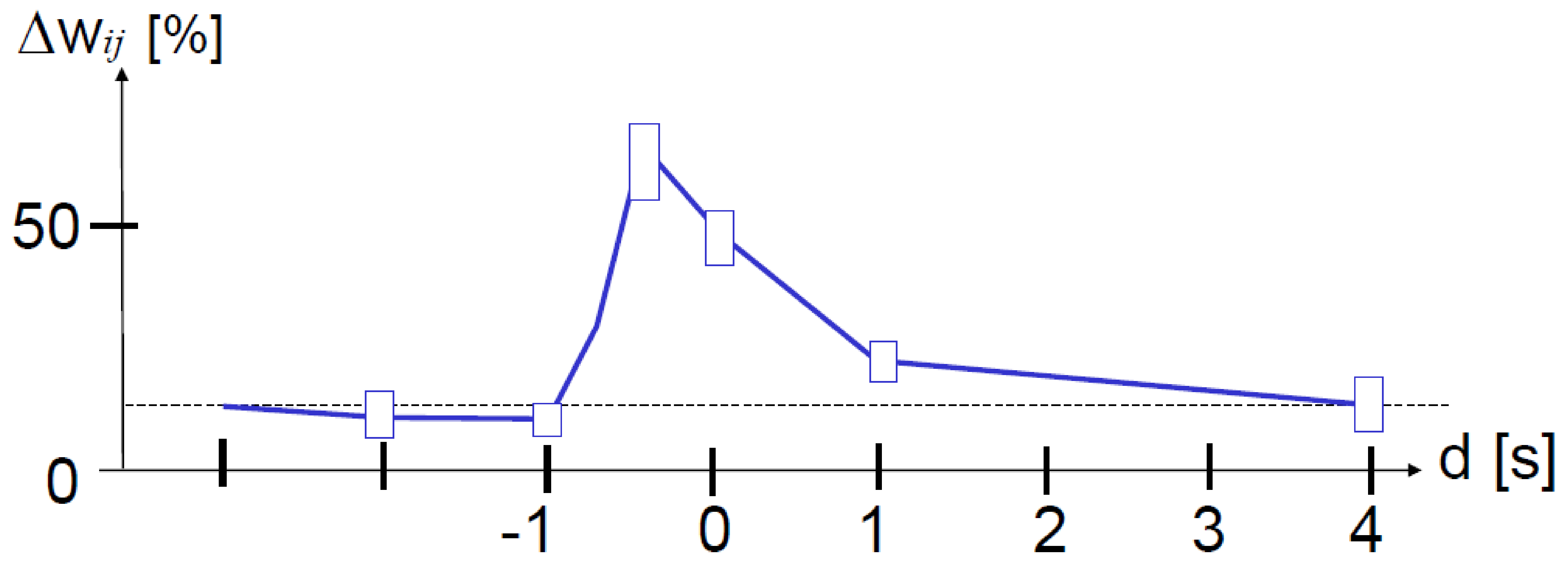
Previous slide.

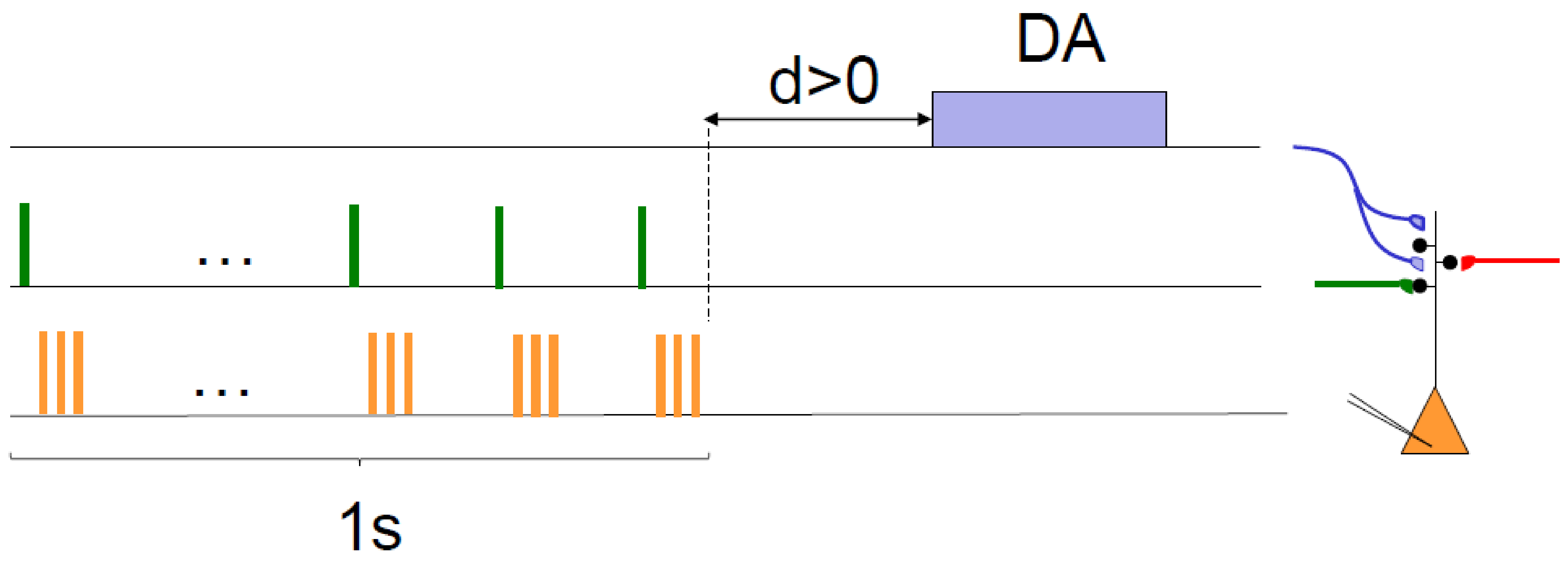
three-factor learning rules are a theoretical concept.

But are there any experiments? Only quite recently, a few experimental results were published that directly address this question.

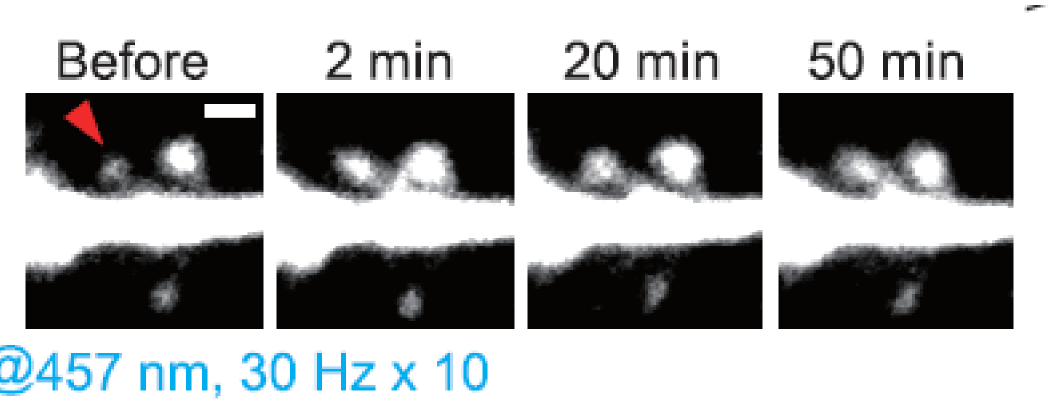
3. Three-factor rules in striatum: eligibility trace and delayed Da

Yagishita et al. 2014



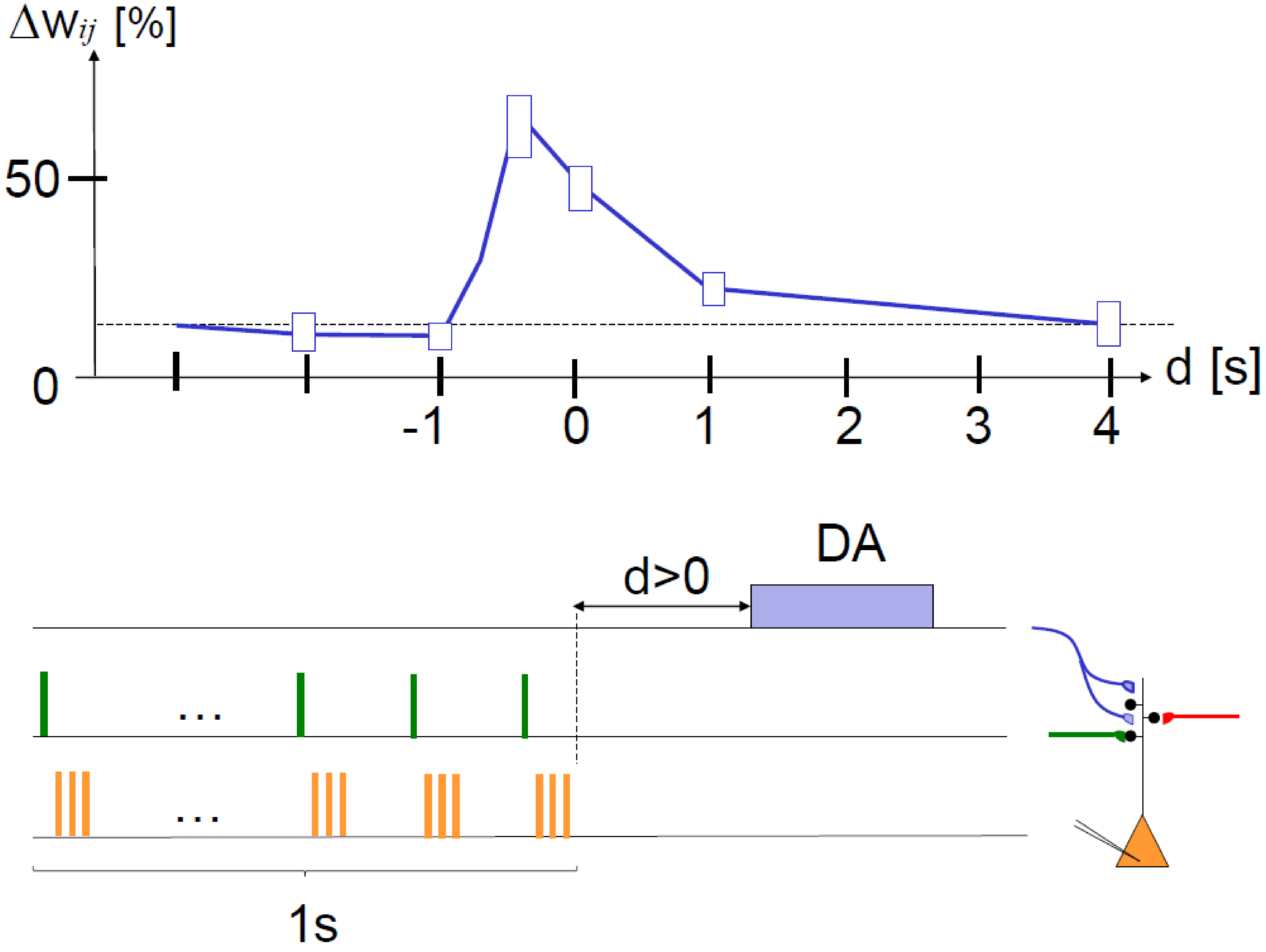


Reminder:
Striatum involved
in action selection



-Dopamine can come with a delay of 1s
-Long-Term stability over at least 50 min.

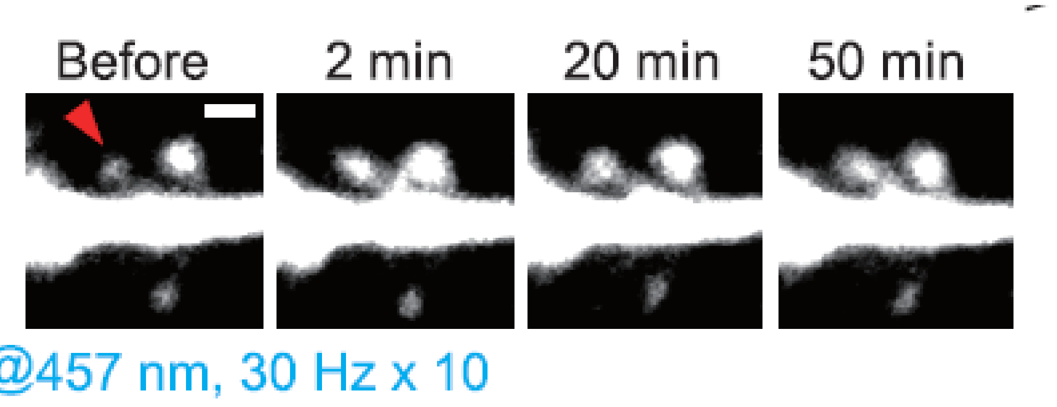
3. Three-factor rules in striatum: eligibility trace and delayed Da



Yagishita et al. 2014

In striatum medial spiny cells, stimulation of presynaptic glutamatergic fibers (green) followed by three postsynaptic action potentials (STDP with pre-post-post-post at +10ms) repeated 10 times at 10Hz yields LTP if dopamine (DA) fibers are stimulated during the presentation (d < 0) or shortly afterward (d = 0s or d = 1s) but not if dopamine is given with a delay d = 4s; redrawn after Fig. 1 of (Yagishita et al., 2014), with delay d defined as time since end of STDP protocol.

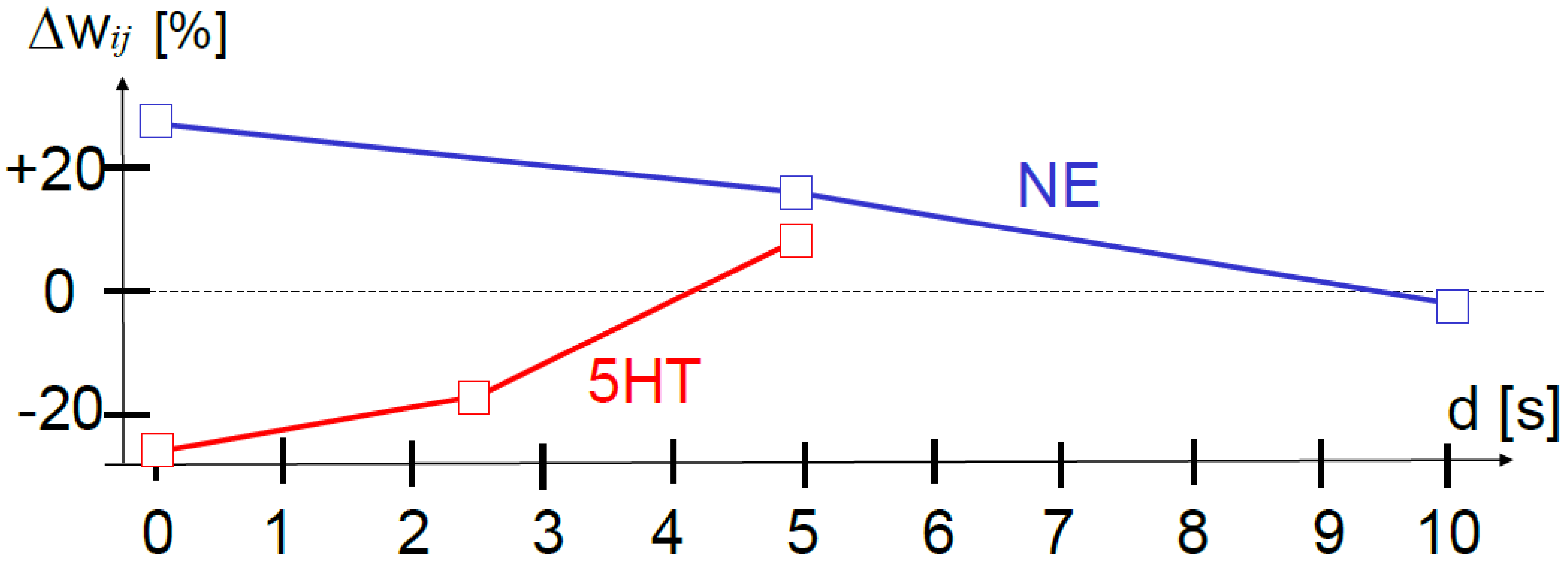
Lower left: the image from the beginning of this lecture comes from this experiment of Yagishita.



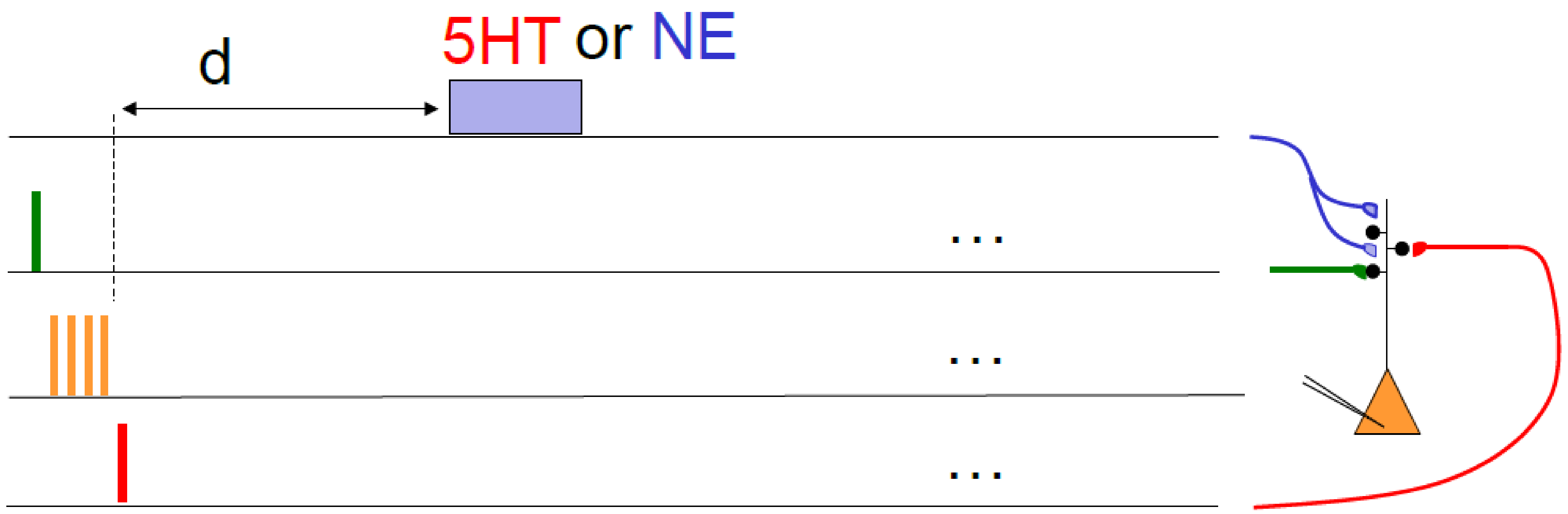
-Dopamine can come with a delay of 1s

-Long-Term stability over at least 50 min.

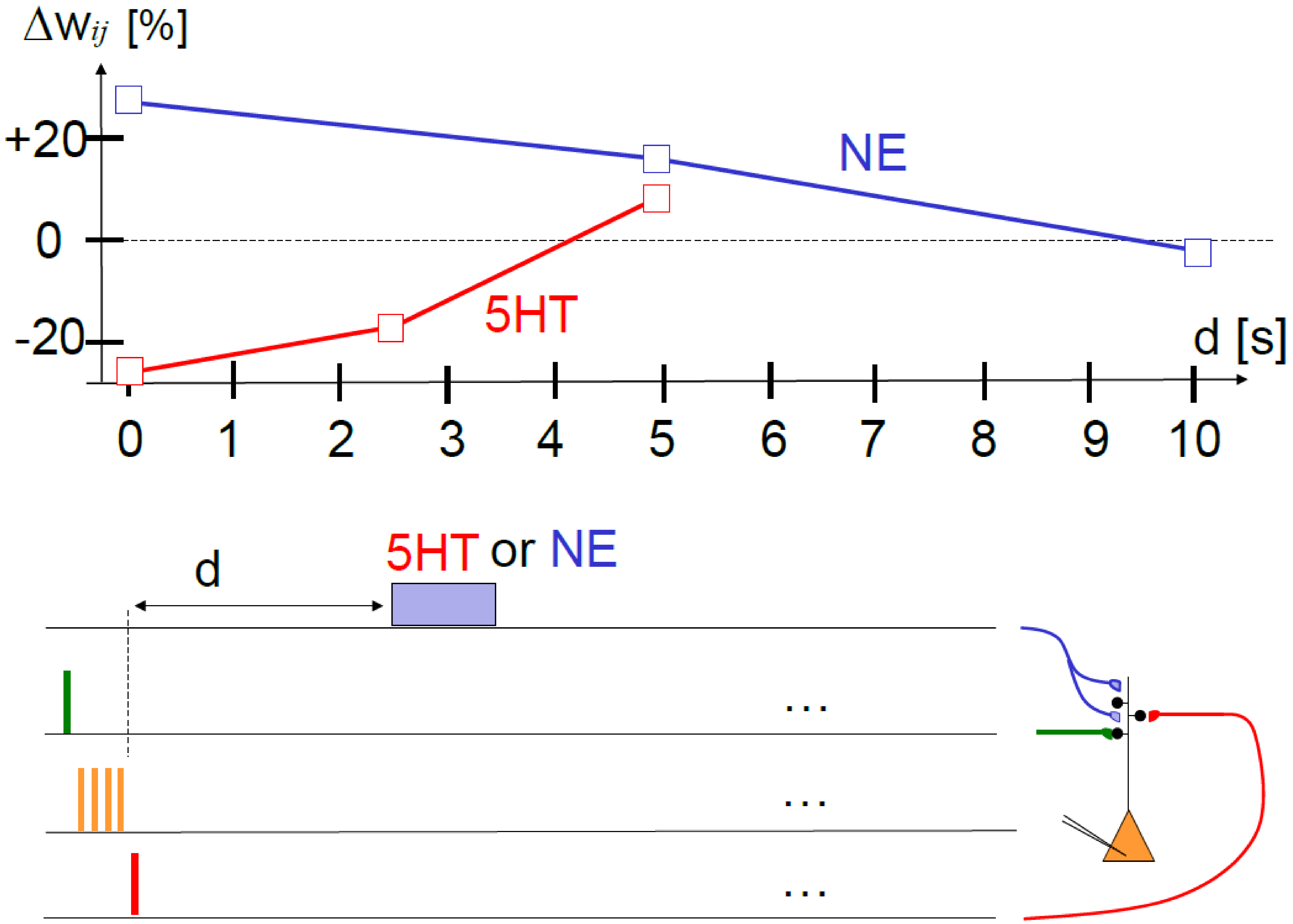
3. Three-factor rules in cortex: eligibility trace and delayed NE



(He et al., 2015).



3. Not shown in class: second example



In cortical pyramidal cells, stimulation of two independent presynaptic pathways (green and red) from layer 4 to layer 2/3 by a single pulse is paired with a burst of four postsynaptic spikes (orange).

If the pre-before-post stimulation was combined with a pulse of norepinephrine (NE) receptor agonist isoproterenol with a delay of 0 or 5s, the protocol gave LTP (blue trace).

If the post-before-pre stimulation was combined with a pulse of serotonin (5-HT) of a delay of 0 or 2.5s, the protocol gave LTD (red trace).

(He et al., 2015).

3. Three-factor rules: summary

Lecture continues at 14h15

Three factors are needed for synaptic changes:

- Presynaptic factor = spikes of presynaptic neuron
  or the effect of spike arrival at the synapse
- Postsynaptic factor = spikes of postsynaptic neuron
  or increased voltage or a function of both
- Third factor = Neuromodulator such as dopamine

Previous slide.

three-factor learning rules are a theoretical concept.

But recent experiments show that the brain really can implement three-factor rules. Importantly, the third factor (neuromodulator) can come with a delay of one or two seconds after the Hebbian induction protocol that sets the eligibility trace.

Quiz. Synaptic Plasticity and Learning Rules

Standard Long-term potentiation

[ ] has an acronym LTP

[ ] takes more than 10 minutes to induce

[ ] lasts more than 30 minutes

[ ] depends on presynaptic activity AND on state of postsynaptic neuron

Learning rules in the brain

[ ] Hebbian learning depends on presynaptic activity AND on state of postsynaptic neuron

[ ] Reinforcement learning depends on neuromodulators such as dopamine indicating reward

[ ] Three-factor rule: presynaptic signal, postsynaptic signal, and neuromodulator signal (e.g., DA) MUST arrive at the same time.

Previous slide.

Your comments.

Artificial Neural Networks and RL : Lecture 13

Wulfram Gerstner
EPFL, Lausanne, Switzerland

from brain-computing to neuromorphic computing

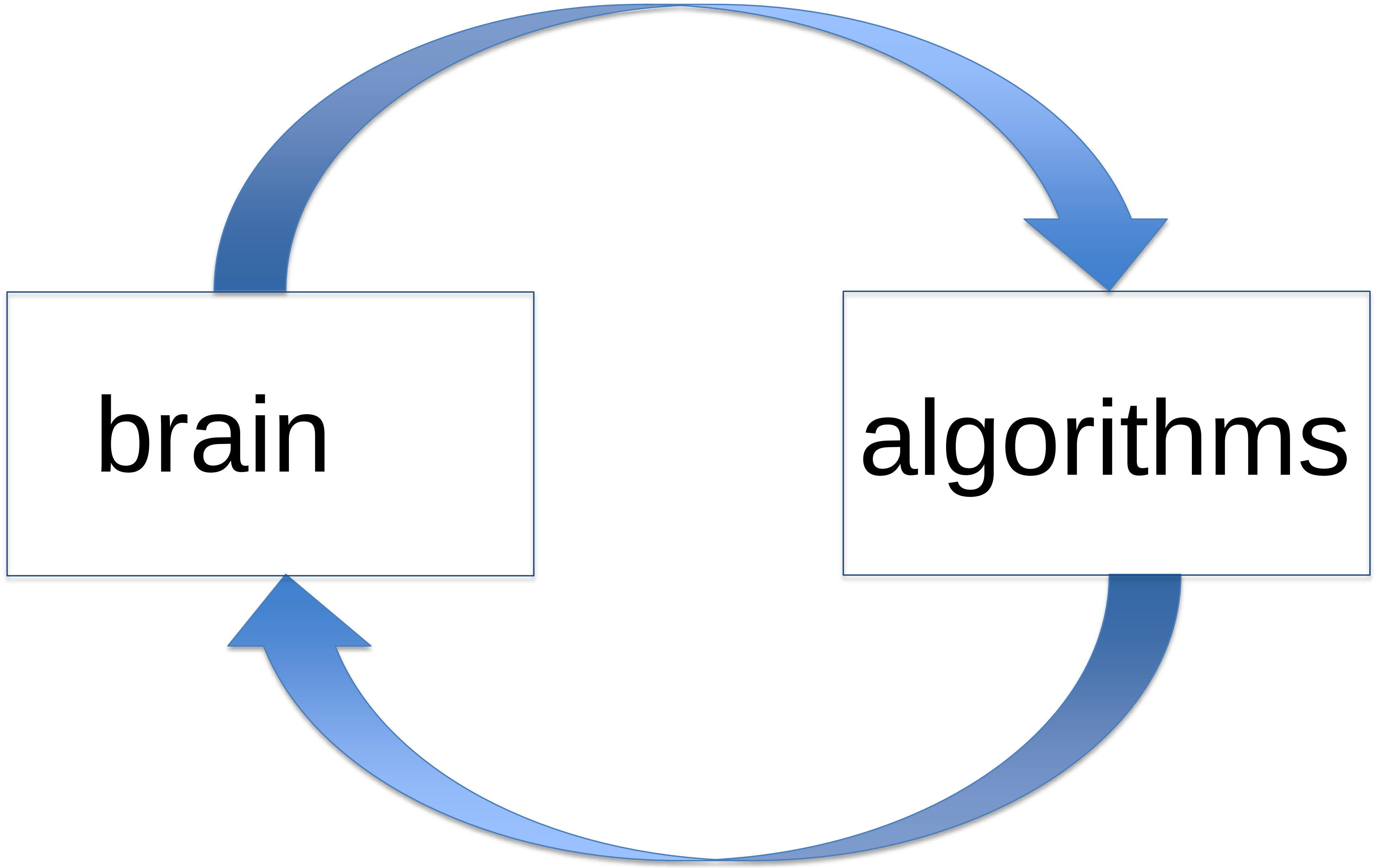
1. Coarse Brain Anatomy
2. Synaptic Plasticity
3. Three-factor Learning Rules
4. **Policy Gradient with Eligibility Traces Revisited**

Previous slide.

I now want to show that reinforcement learning with policy gradient gives rise to three-factor learning rules.

Learning Rules

3-factor learning rules



brain

algorithms

Advantage Actor-Critic with eligibility traces

Previous slide.
We will now compare the learning rule of the advantage actor critic with eligibility traces to the three-factor rules of the brain.

We bring together the actor-critic with eligibility traces and the results of exercise 1 today.

4. Eligibility traces from Policy Gradient (Exercise today)

Run episode.

At each time step, observe state s_t, action a_t, reward r_t

1) Update eligibility trace

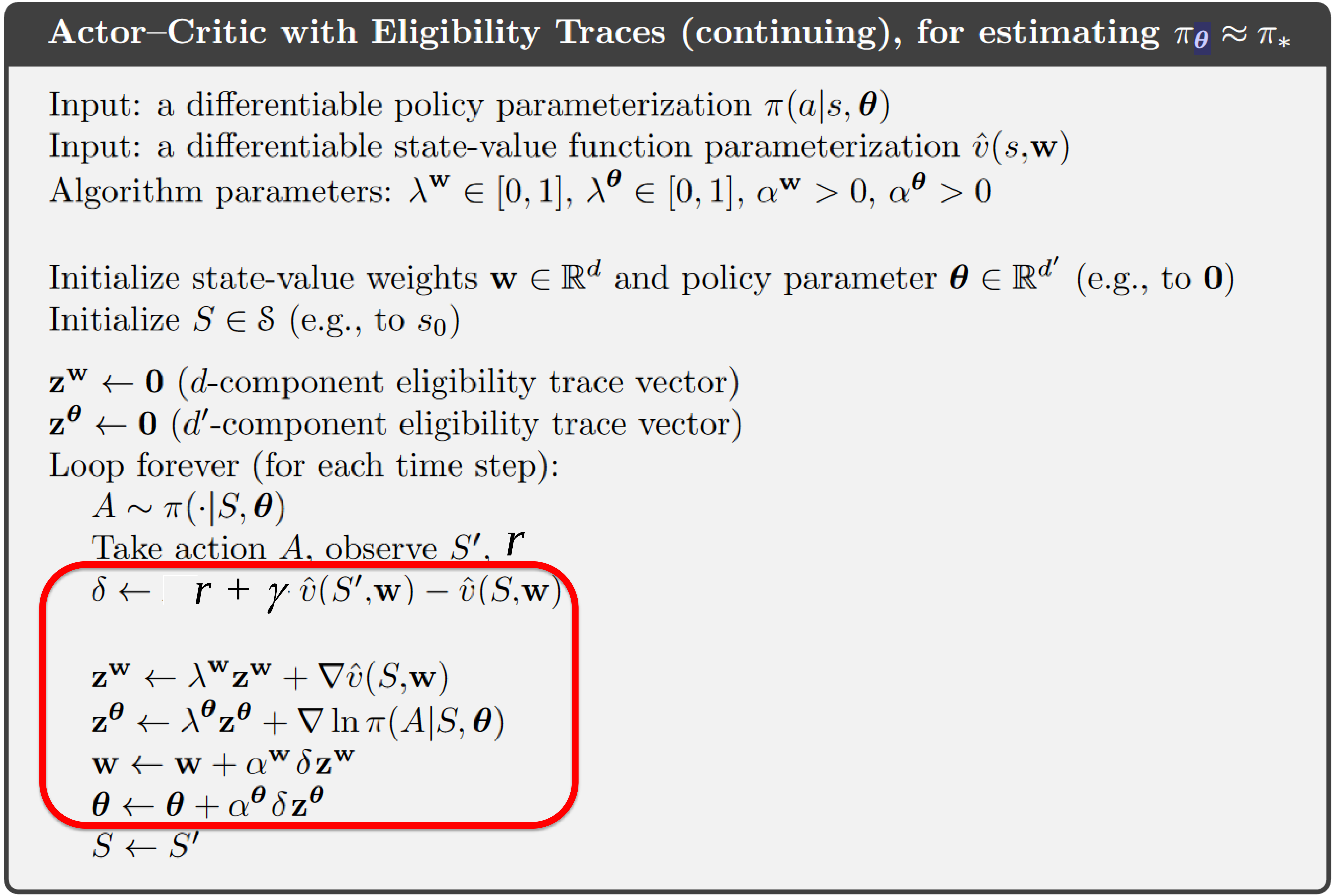
$z_k \leftarrow z_k \lambda$ decay of **all** traces

$z_k \leftarrow z_k + \frac{d}{dw_k} \ln[\pi(a_t|s_t, w_k)]$ increase of **all** traces

2) update parameters

$\Delta w_k = \eta \; r_t \; z_k$

Previous slide. repetition of the exercises from week 10 and Exercise of Today
Leads to the algo on slide 7

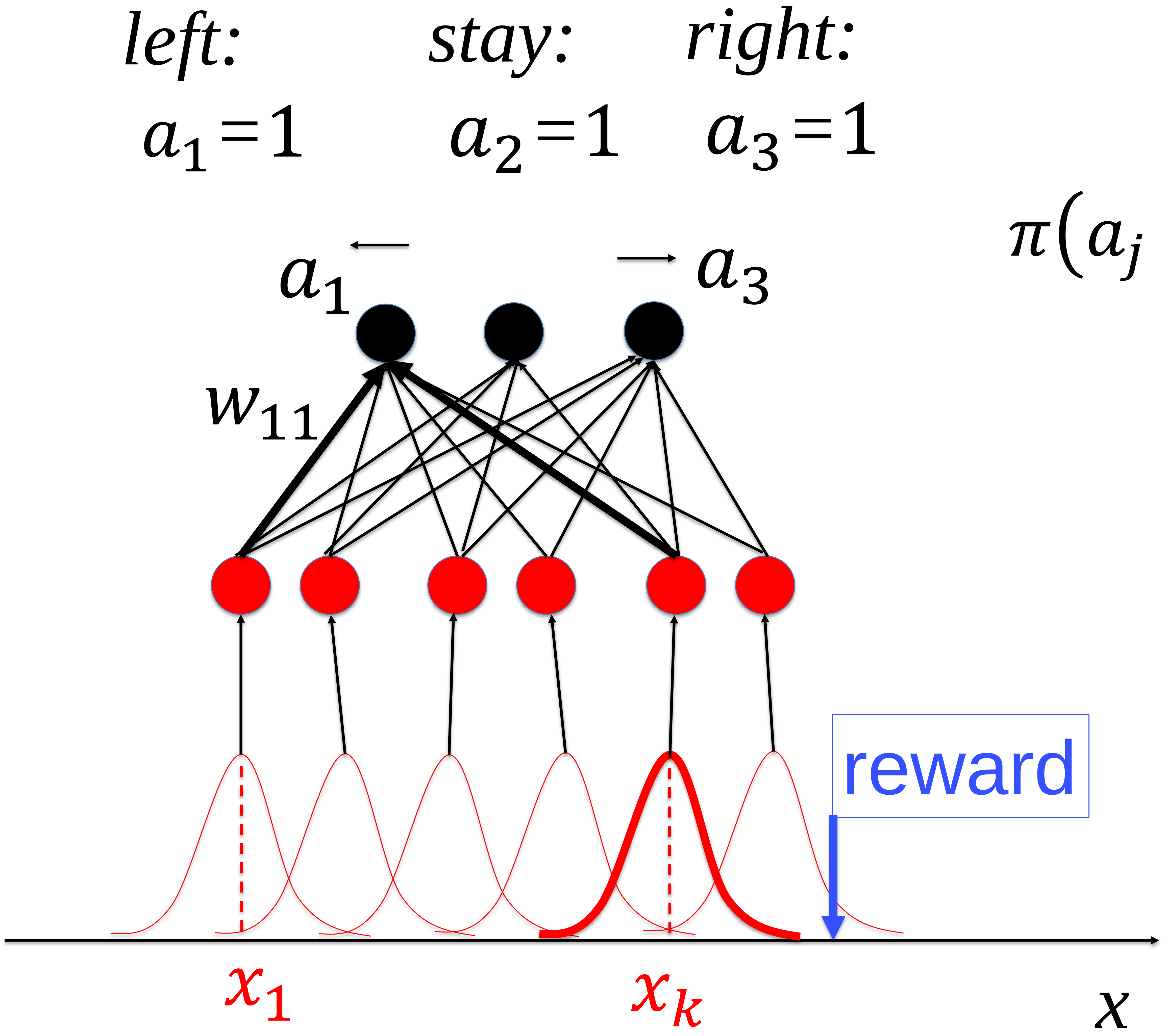


Actor–Critic with Eligibility Traces (continuing), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

```
Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Algorithm parameters: $\lambda^{\mathbf{w}} \in [0,1]$, $\lambda^{\boldsymbol{\theta}} \in [0,1]$, $\alpha^{\mathbf{w}} > 0$, $\alpha^{\boldsymbol{\theta}} > 0$

Initialize state-value weights $\mathbf{w} \in \mathbb{R}^d$ and policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)
Initialize $S \in \mathcal{S}$ (e.g., to $s_0$)

$\mathbf{z}^{\mathbf{w}} \leftarrow \mathbf{0}$ ($d$-component eligibility trace vector)
$\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \mathbf{0}$ ($d'$-component eligibility trace vector)
Loop forever (for each time step):
    $A \sim \pi(\cdot|S,\boldsymbol{\theta})$
    Take action $A$, observe $S'$, $r$
    $\delta \leftarrow r + \gamma\, \hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$

    $\mathbf{z}^{\mathbf{w}} \leftarrow \lambda^{\mathbf{w}} \mathbf{z}^{\mathbf{w}} + \nabla \hat{v}(S,\mathbf{w})$
    $\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \lambda^{\boldsymbol{\theta}} \mathbf{z}^{\boldsymbol{\theta}} + \nabla \ln \pi(A|S,\boldsymbol{\theta})$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}}\, \delta\, \mathbf{z}^{\mathbf{w}}$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}}\, \delta\, \mathbf{z}^{\boldsymbol{\theta}}$
    $S \leftarrow S'$
```

Adapted from Sutton and Barto

4. Example: Linear activation model with softmax policy



parameters

$$\pi(a_j = 1 | \vec{x}, \theta) = softmax[\sum_k w_{jk}\, y_k]$$

$$y_k = f(\vec{x} - x_k)$$

f =basis function

Previous slide.
Suppose the agent moves on a linear track.
There are three possible actions: left, right, or stay.

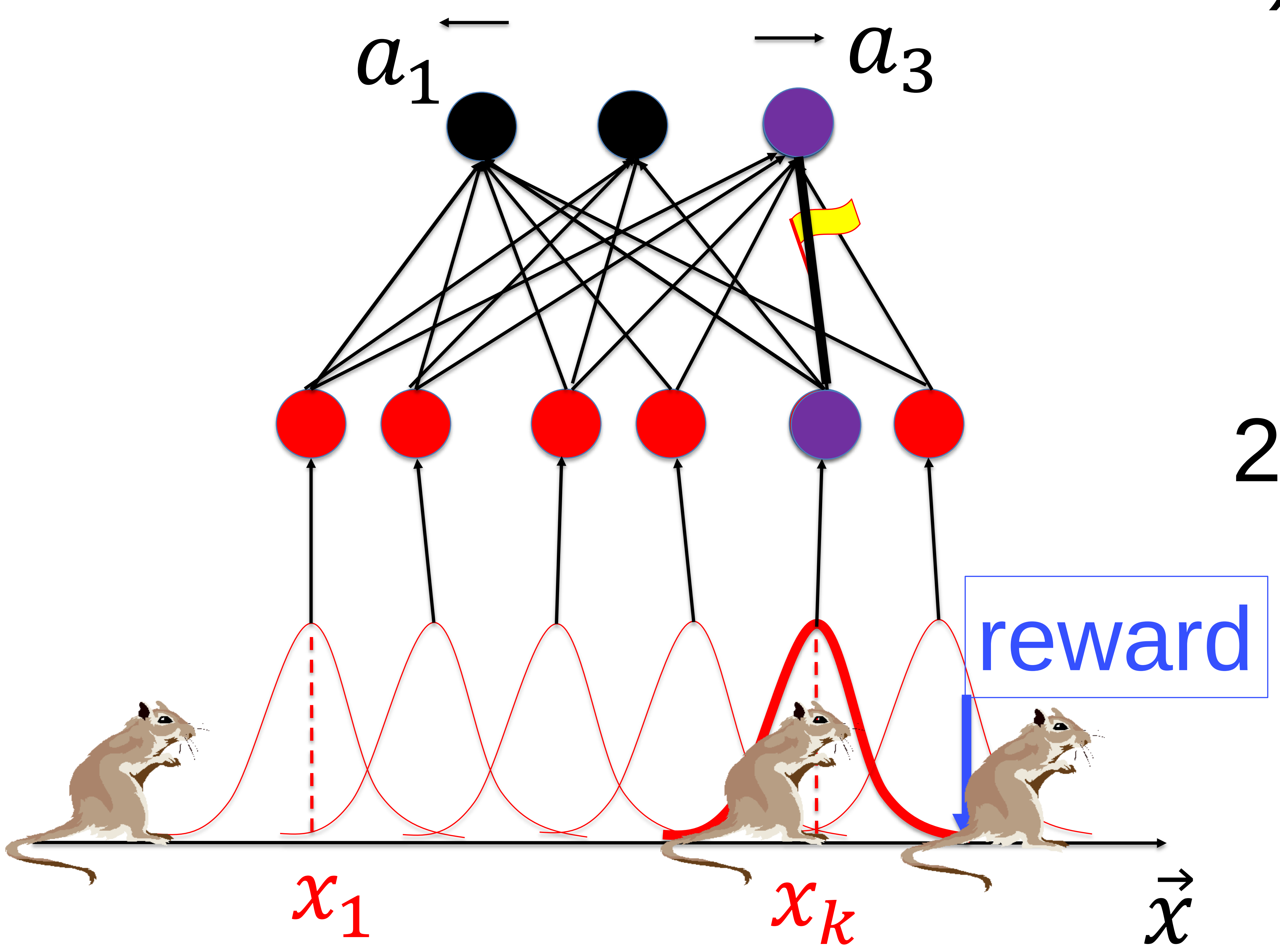
The policy is given by the softmax function. The total drive of the action neurons is a linear function of the activity y of the hidden neurons which in turn depends on the input x. The activity of hidden neuron k is f(x-x_k). The basis function f could for example be a Gaussian function with center at x_k.

4. Example: Linear activation model with softmax policy

left: $a_1=1$

stay: $a_2=1$

right: $a_3=1$



0) Choose action $a_i \in \{0,1\}$

1) Update eligibility trace

$$z_{ik} \leftarrow z_{ik}\,\lambda$$

$$z_{ik} \leftarrow z_{ik} + \frac{d}{dw_k}\ln[\pi(a_i^t = 1|\vec{x})]$$

2) update weights

$$\Delta w_{lk} = \eta\ r_t\ z_{lk}$$

Already done in Exercise 1

→ Three-factor rule with eligibility traces

Previous slide.
This is the result of the in-class exercise (Exercise 1 of this week).
Importantly, the update of the eligibility trace is a local learning rule that depends on a presynaptic factor and a postsynaptic factor.
The reward is the third factor and has no indices (since it acts as a global factor, broadcasted to all neurons and synapses).

4. Summary: 3-factor rules derived from Policy Gradient

- Policy gradient with one hidden layer and linear softmax readout yields a 3-factor rule
- Eligibility trace is set by joint activity of presynaptic and postsynaptic neuron
- Update happens proportional to the eligibility trace and to either reward r (REINFORCE) or TD error (Adv. Actor-Critic)
- The presynaptic neuron represents the state
- The postsynaptic neuron the action
- True online rule
  → could be implemented in biology
  → can also be implemented in parallel asynchr. hardware

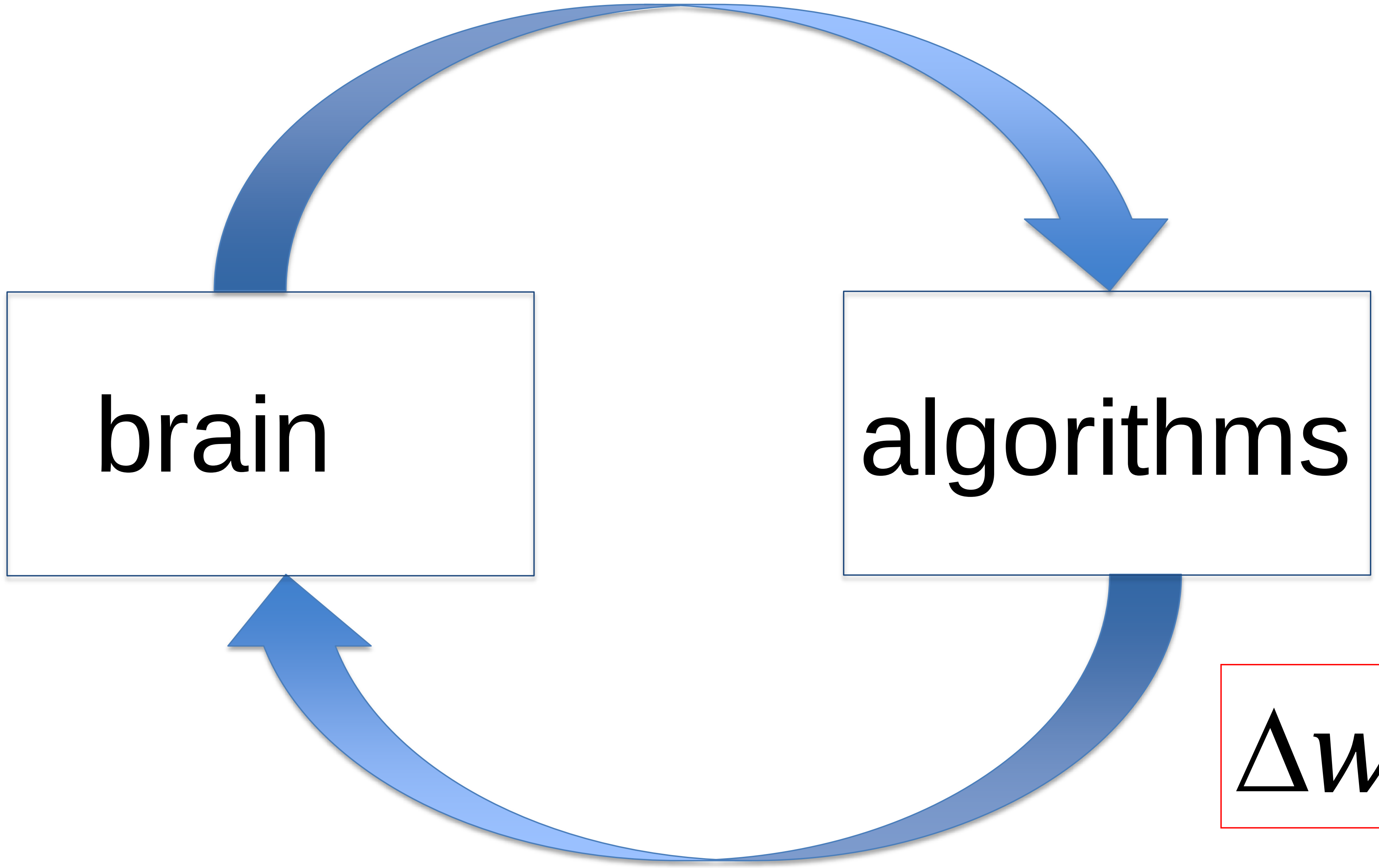
Previous slide.
Summary: A policy gradient algorithm in a network where the output layer has a linear drive with softmax output leads to a three-factor learning rule for the connections between neurons in the hidden layer and the output.

These three factor learning rules are important because they are completely asynchronous, local, and online and could therefore be implemented in biology or parallel hardware.

The global modulator could present either the reward r directly (in the style of the REINFORCE algorithm); or it could present the TD error (which yields an interpretation as advantage actor-critic.

Which one of the two possibilities would fit the dopamine signal?
This is the next question

Learning Rules

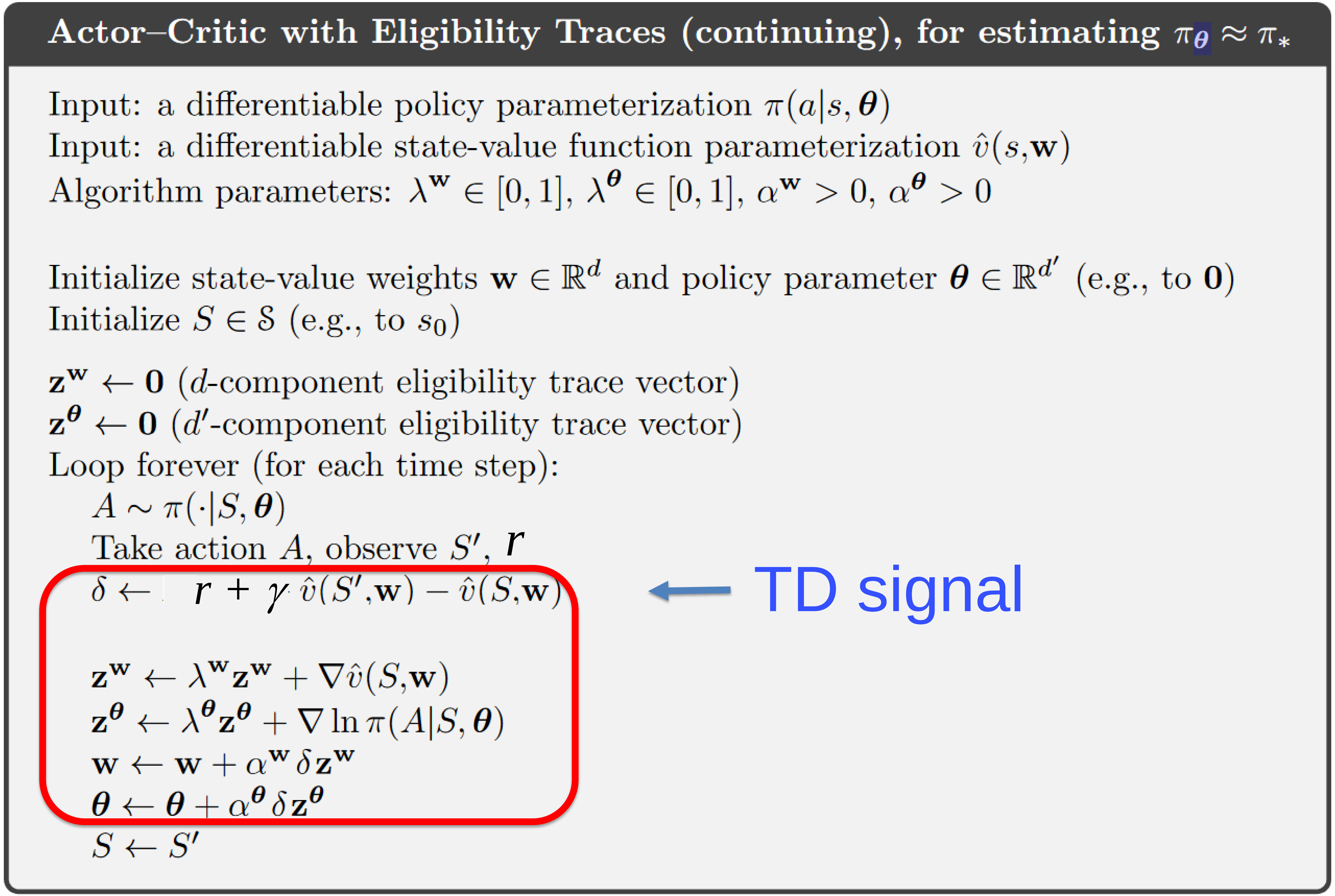


$$\Delta w_{lk} = \eta \;\; r_t \; z_{lk}$$

The learning rule of the (normal) actor-critic with eligibility traces (REINFORCE WITH BASELINE) is consistent with a brain-like three-factor rule.

Updates proportional to the reward r (minus baseline).

Review: Advantage Actor-Critic with Eligibility traces

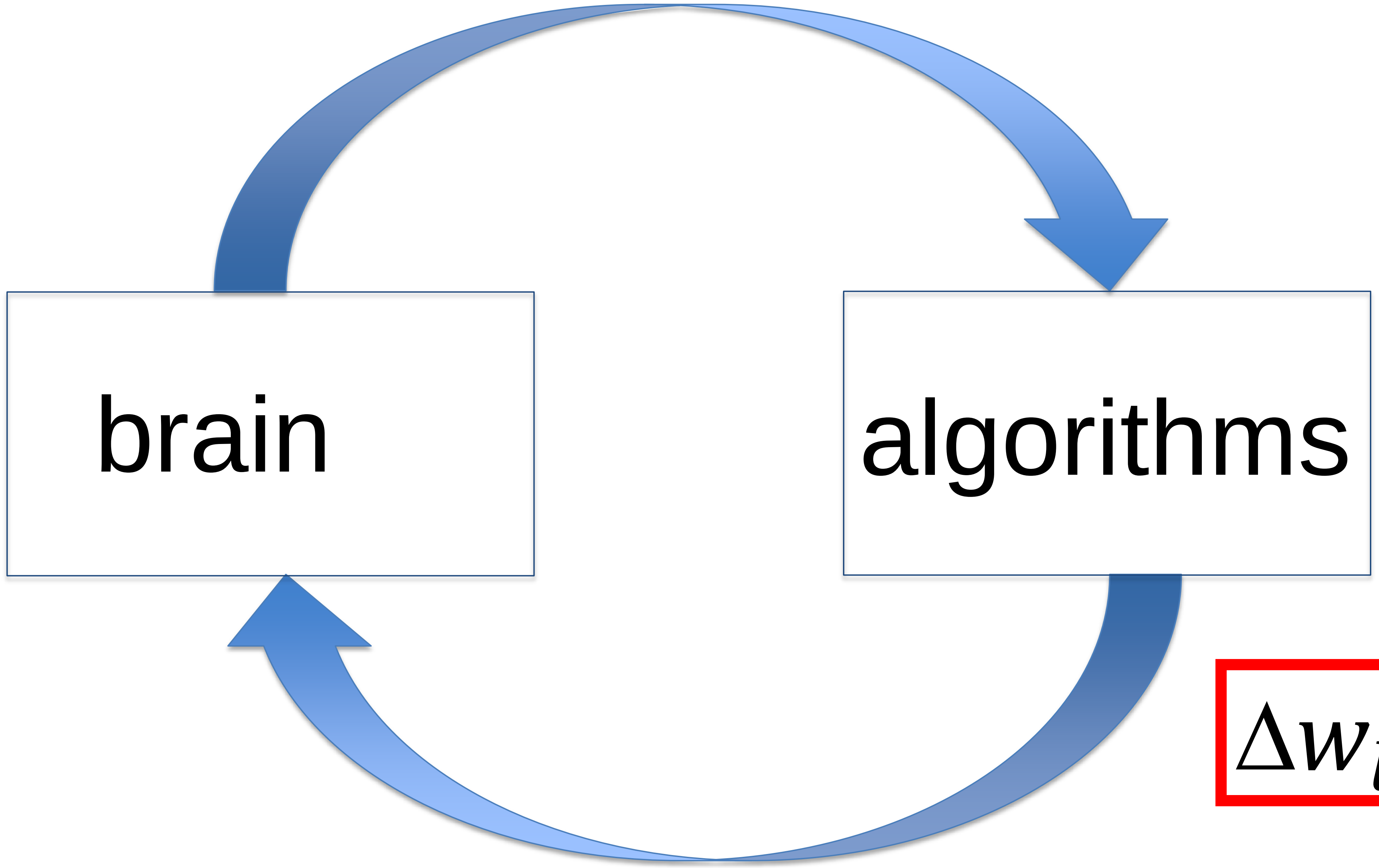


```
Actor–Critic with Eligibility Traces (continuing), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Algorithm parameters: $\lambda^{\mathbf{w}} \in [0,1]$, $\lambda^{\boldsymbol{\theta}} \in [0,1]$, $\alpha^{\mathbf{w}} > 0$, $\alpha^{\boldsymbol{\theta}} > 0$

Initialize state-value weights $\mathbf{w} \in \mathbb{R}^d$ and policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)
Initialize $S \in \mathcal{S}$ (e.g., to $s_0$)

$\mathbf{z}^{\mathbf{w}} \leftarrow \mathbf{0}$ ($d$-component eligibility trace vector)
$\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \mathbf{0}$ ($d'$-component eligibility trace vector)
Loop forever (for each time step):
    $A \sim \pi(\cdot|S,\boldsymbol{\theta})$
    Take action $A$, observe $S'$, $r$
    $\delta \leftarrow r + \gamma\, \hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$

    $\mathbf{z}^{\mathbf{w}} \leftarrow \lambda^{\mathbf{w}} \mathbf{z}^{\mathbf{w}} + \nabla \hat{v}(S,\mathbf{w})$
    $\mathbf{z}^{\boldsymbol{\theta}} \leftarrow \lambda^{\boldsymbol{\theta}} \mathbf{z}^{\boldsymbol{\theta}} + \nabla \ln \pi(A|S,\boldsymbol{\theta})$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}}\, \delta\, \mathbf{z}^{\mathbf{w}}$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}}\, \delta\, \mathbf{z}^{\boldsymbol{\theta}}$
    $S \leftarrow S'$
```

← TD signal

Adapted from Sutton and Barto

Learning Rules



TD signal

$$\boxed{\Delta w_{lk} = \eta \;\; \delta_t \; z_{lk}}$$

The learning rule of the **advantage** actor-critic with eligibility traces is consistent with a brain-like three-factor rule

Condition: the brain can broad-cast a TD signal!

Previous slide.
The main difference between standard REINFORCE (potentially with baseline subtraction) and the Advantage Actor Critic is that
in the advantage actor-critic the global modulator re-presents the TD error.

We now show that the TD signal is consistent with the dopamine signal!

Artificial Neural Networks and RL : Lecture 13

Wulfram Gerstner

EPFL, Lausanne, Switzerland

from brain-computing to neuromorphic computing

1. Coarse Brain Anatomy
2. Synaptic Plasticity
3. Three-factor Learning Rules
4. Policy Gradient with Eligibility Traces Revisited
5. **Dopamine as a Third Factor**

Previous slide.

So far the third factor remained rather abstract. We mentioned that a neuromodulator such as dopamine could be involved. Let us make this idea more precise and show experimental data.

5. Neuromodulators as Third factor

Three factors are needed for synaptic changes:

- Presynaptic factor = spikes of presynaptic neuron
- Postsynaptic factor = spikes of postsynaptic neuron or increased voltage
- Third factor = Neuromodulator such as dopamine

Presynaptic and postsynaptic factor ‘select’ the synapse.

→ a small subset of synapses becomes ‘eligible’ for change.

The ‘Third factor’ is a nearly global signal

→ broadcast signal, potentially received by all synapses.

Synapses need all three factors for change

Previous slide.
Before we start let us review the basics of a three-factor learning rule. We said that the third factor could be a neuromodulator such as dopamine.

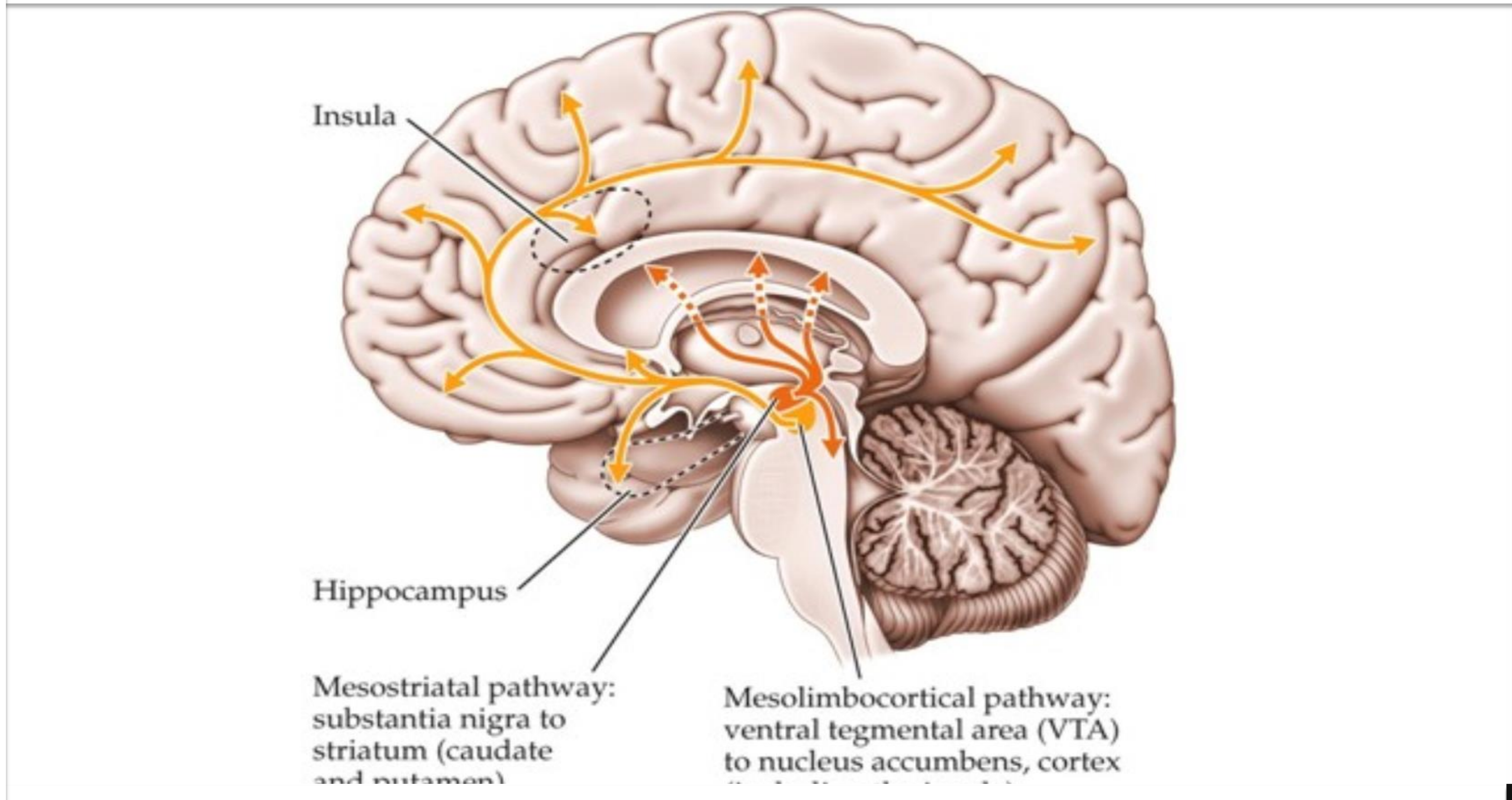
Review: Reward information

Neuromodulator **dopamine**: - is nearly globally broadcasted
- signals reward minus expected reward

‘success signal’

Schultz et al., 1997,
Waelti et al., 2001
Schultz, 2002

Dopamine


Insula
Hippocampus
Mesostriatal pathway: substantia nigra to striatum (caudate and putamen)
Mesolimbocortical pathway: ventral tegmental area (VTA) to nucleus accumbens, cortex

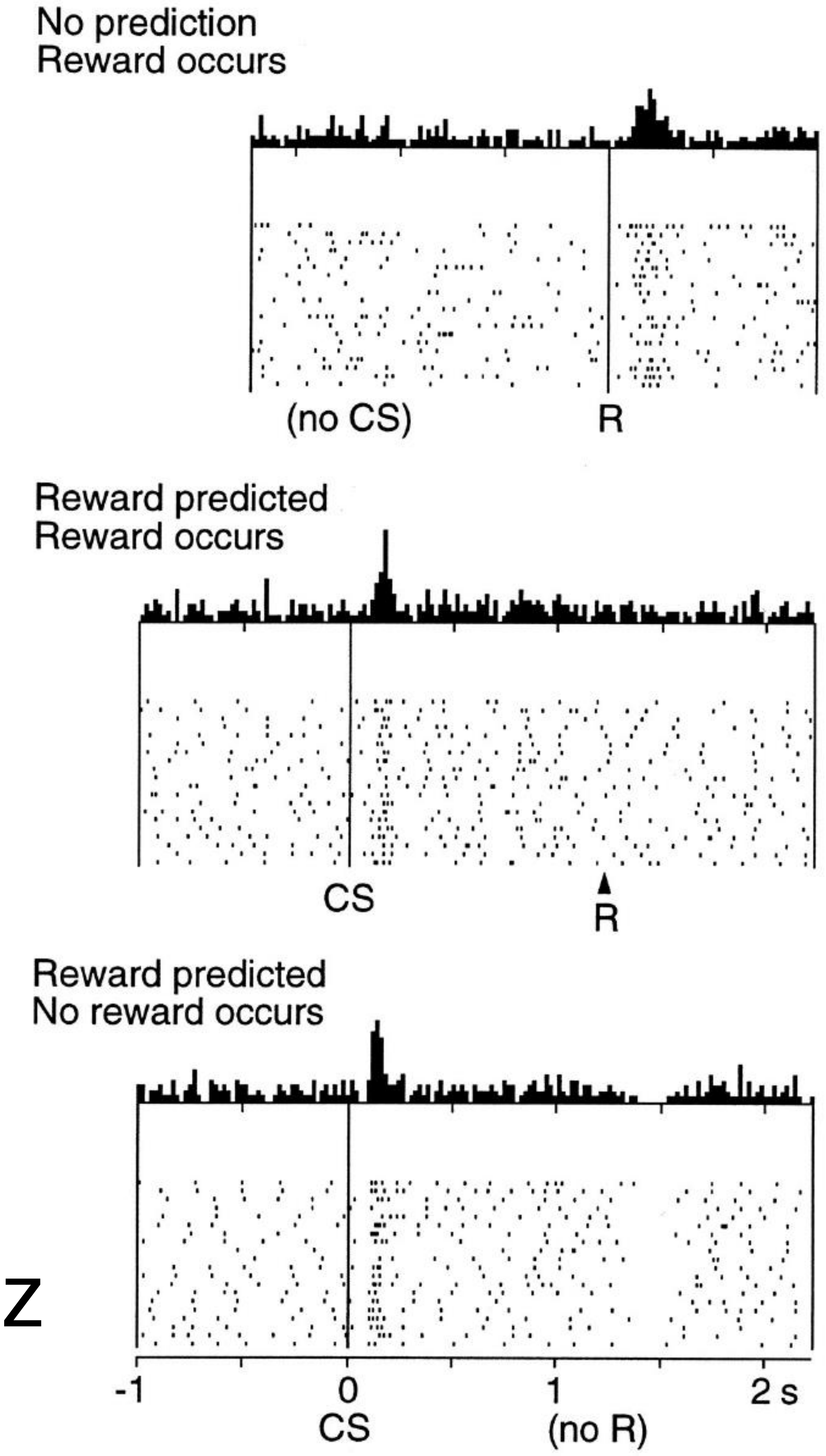
Previous slide. Dopamine neurons send dopamine signals to many neurons and synapses in parallel in a broadcast like fashion.

5. Dopamine as Third factor

Conditioning:

red light →1s→reward

CS:
Conditioning
Stimulus



No prediction
Reward occurs

Reward predicted
Reward occurs

Reward predicted
No reward occurs

Sutton book, reprinted from W. Schultz

5. Dopamine as Third factor

This is now the famous experiment of W. Schultz.
In reality the CS was not a red light, but that does not matter

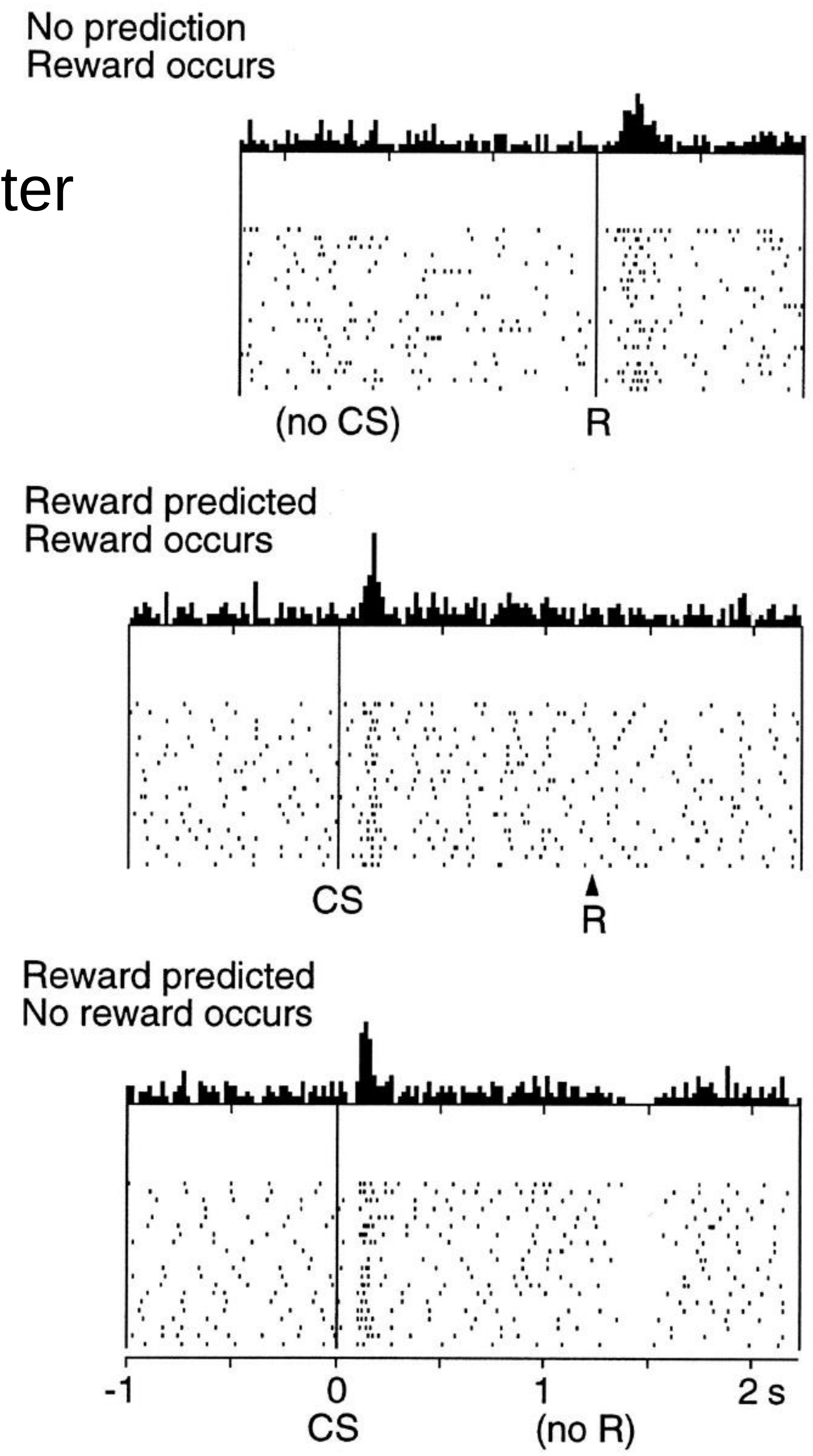


Figure 15.3: The response of dopamine neurons drops below baseline shortly after the time when an expected reward fails to occur. Top: dopamine neurons are activated by the unpredicted delivery of a drop of apple juice. Middle: dopamine neurons respond to a conditioned stimulus (CS) that predicts reward and do not respond to the reward itself. Bottom: when the reward predicted by the CS fails to occur, the activity of dopamine neurons drops below baseline shortly after the time the reward is expected to occur. At the top of each of these panels is shown the average number of action potentials produced by monitored dopamine neurons within small time intervals around the indicated times. The raster plots below show the activity patterns of the individual dopamine neurons that were monitored; each dot represents an action potential. From Schultz, Dayan, and Montague, A Neural Substrate of Prediction and Reward, *Science*, vol. 275, issue 5306, pages 1593-1598, March 14, 1997. Reprinted with permission from AAAS.

5. Summary: Dopamine as Third factor

- Dopamine signals ‘reward minus expected reward’
- Dopamine signals an ‘event that predicts a reward’
- Dopamine signals approximately the TD-error

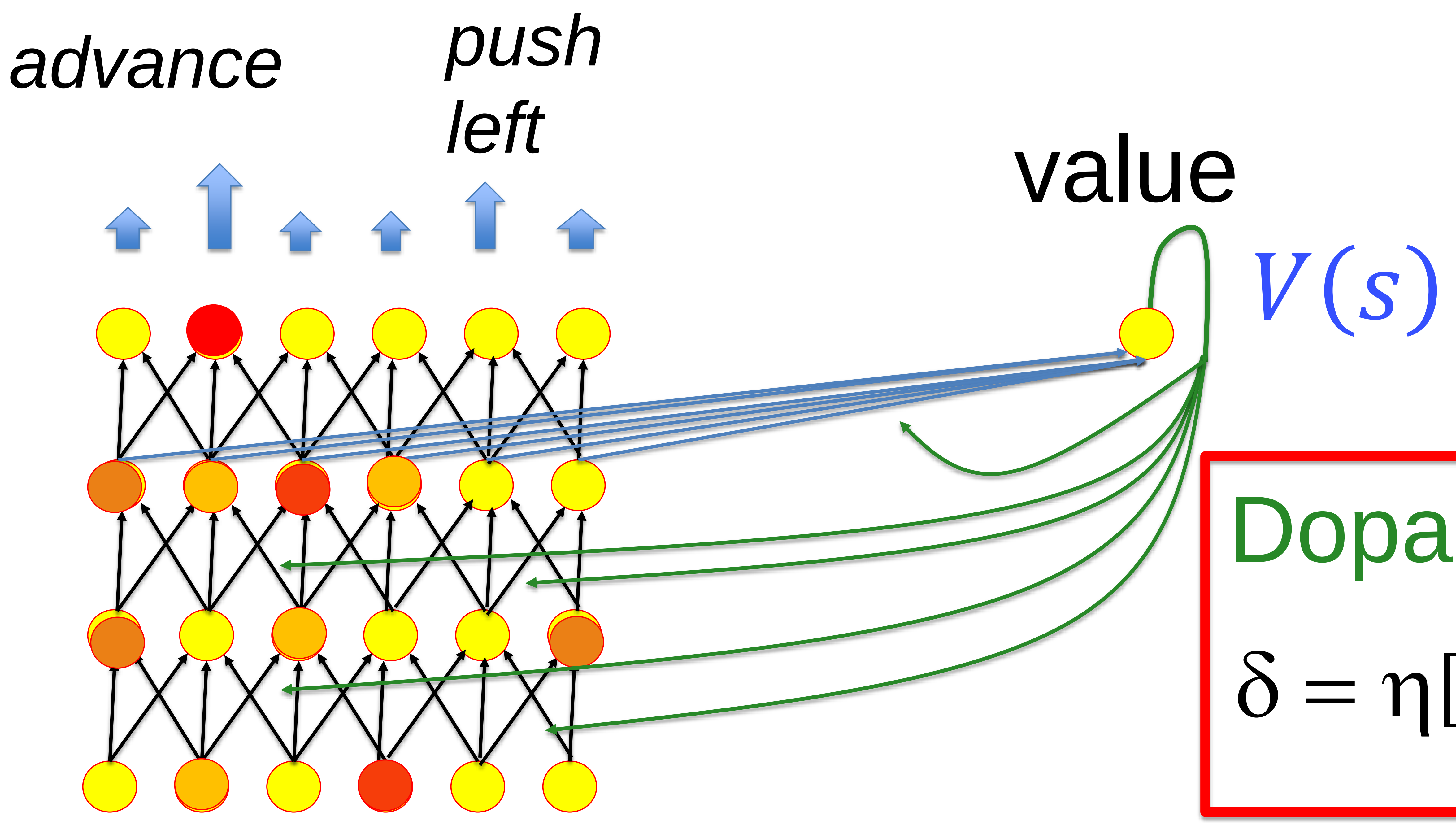
$$DA(t) = [\underbrace{r(t)-(\,V(s)-\gamma\, V(s'))}_{\text{TD-delta}}]$$

Previous slide.

The paper of W. Schultz has related the dopamine signal to some basic aspects of Temporal difference Learning. The Dopamine signal is similar to the TD error.

5. Application: Advantage Actor-Critic = update with TD signal

- Estimate $V(s)$
- learn via TD error



Dopamine = TD-error

$$\delta = \eta[r_t + \gamma V(s') - V(s)]$$

Previous slide.

Review of actor-critic architecture

5. Combine Eligibility Traces with TD in Advantage Actor-Critic

Idea:

- keep memory of previous ‘candidate updates’
- memory decays over time
- Update an **eligibility trace for each parameter**

$z_k \leftarrow z_k \, \lambda$ decay of **all** traces

$z_k \leftarrow z_k + \frac{d}{dw_k} \ln[\pi(a|s, w_k)]$ increase of **all** traces

- update **all** parameters:

$$\boxed{\Delta w_k = \eta \; \underbrace{[r - (V(s) - \gamma V(s'))]}_{\text{TD-delta}} \; z_k}$$

→ policy gradient with eligibility trace and TD error

Previous slide.

Review of algorithm with actor-critic architecture and policy gradient with eligibility traces and TD.

5. Summary: Eligibility Traces with TD in Actor-Critic

Three-factor rules:

Presynaptic and postsynaptic factor ‘select’ the synapse.

→ a small subset of synapses becomes ‘eligible’ for change.

The ‘Third factor’ is a nearly global broadcast signal

→ potentially received by all synapses.

Synapses need all three factors for change

The ‘Third factor’ can be the Dopamine-like TD signal

→ Need actor-critic architecture to calculate $\gamma V(s') - V(s)$

→ Dopamine signals $[r_t + \gamma V(s') - V(s)]$

Previous slide.

The three factor rule, dopamine, TD signals, value functions now all fit together.

Artificial Neural Networks and RL : Lecture 13

Wulfram Gerstner

EPFL, Lausanne, Switzerland

from brain-computing to neuromorphic computing

1. Coarse Brain Anatomy
2. Synaptic Plasticity
3. Three-factor Learning Rules
4. Policy Gradient with Eligibility Traces Revisited
5. Dopamine as a Third Factor
6. **Example: Navigation in a Maze (Model Study)**

- What is the task?
- How are ‘states’ represented?
- How are ‘actions’ represented?
- How is the ‘learning rule’ represented

Previous slide.

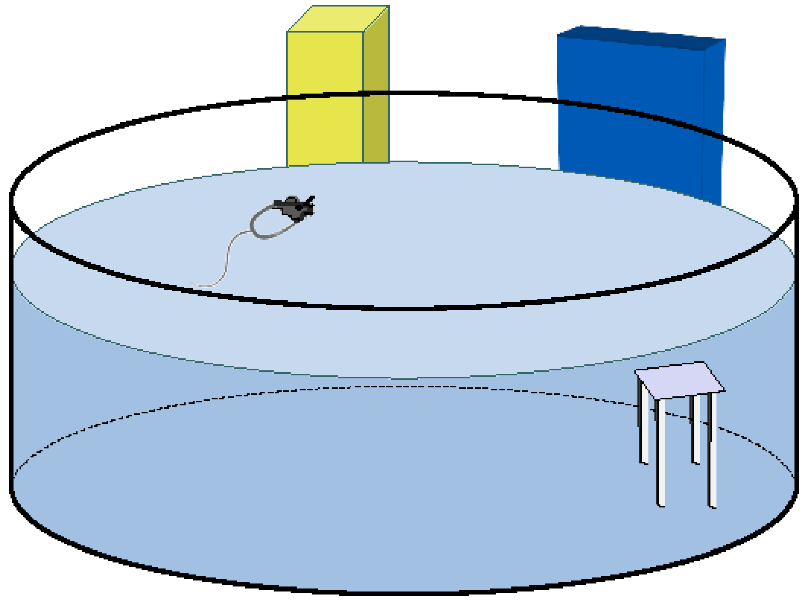
We said that the three factor rule, dopamine, TD signals, value functions now all fit together. Let’s apply this to the problem of navigation in a maze.

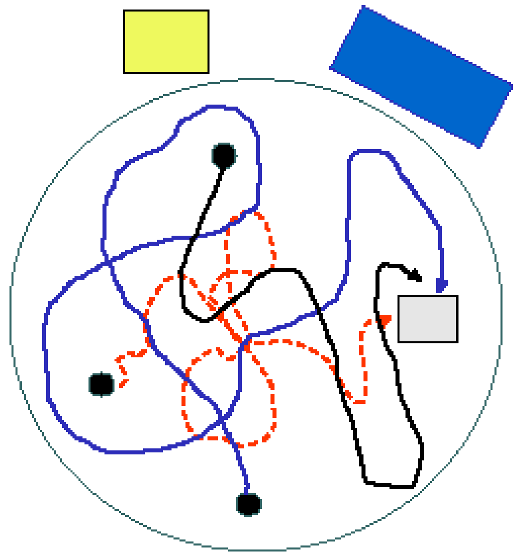
For biological plausibility we have to consider:

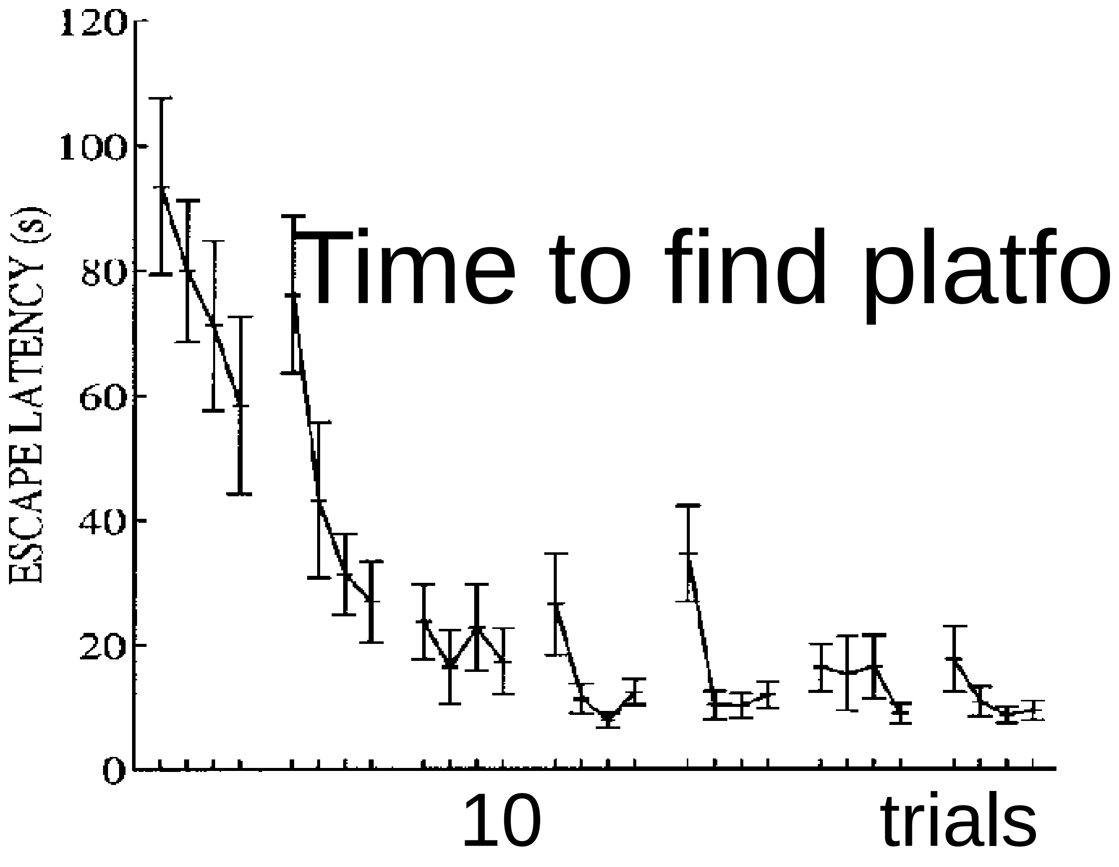
- Representation of states
- Representation of actions
- Representation of TD signal and learning rule

Review: TASK = conditioning in the Morris Water Maze

Morris Water Maze







Rats learn to find the hidden platform

(Because they like to get out of the cold water)

Foster, Morris, Dayan 2000

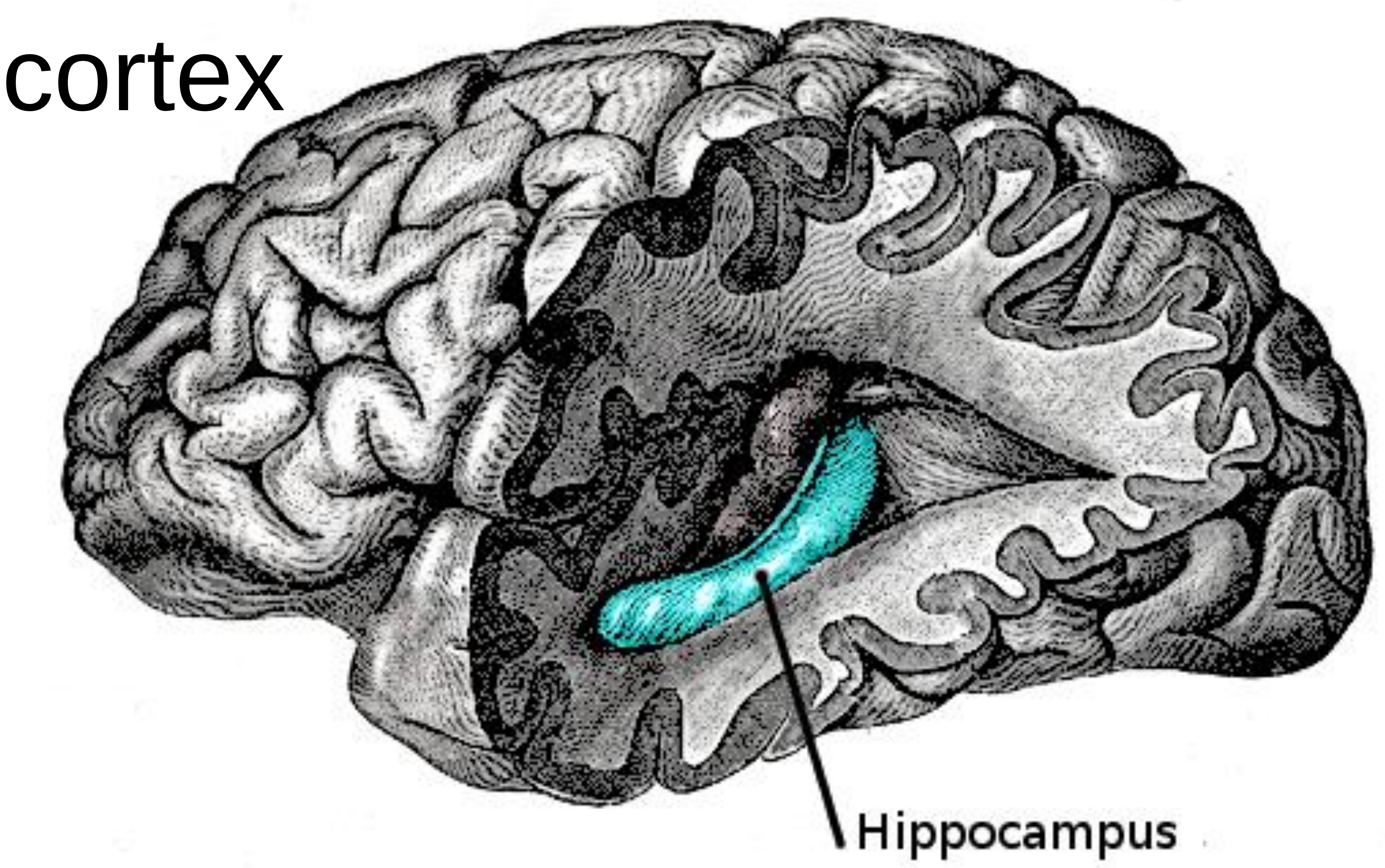
Previous slide.
Behvioral experiment in the Morris Water Maze.
The water is milky so that the platform is visible.

After a few trials the rat swims directly to the platform

6. Representation of momentary state: hippocampus

Hippocampus

- Sits below/part of temporal cortex
- Involved in memory
- Involved in spatial memory



Spatial memory:

knowing where you are,

knowing how to navigate in an environment

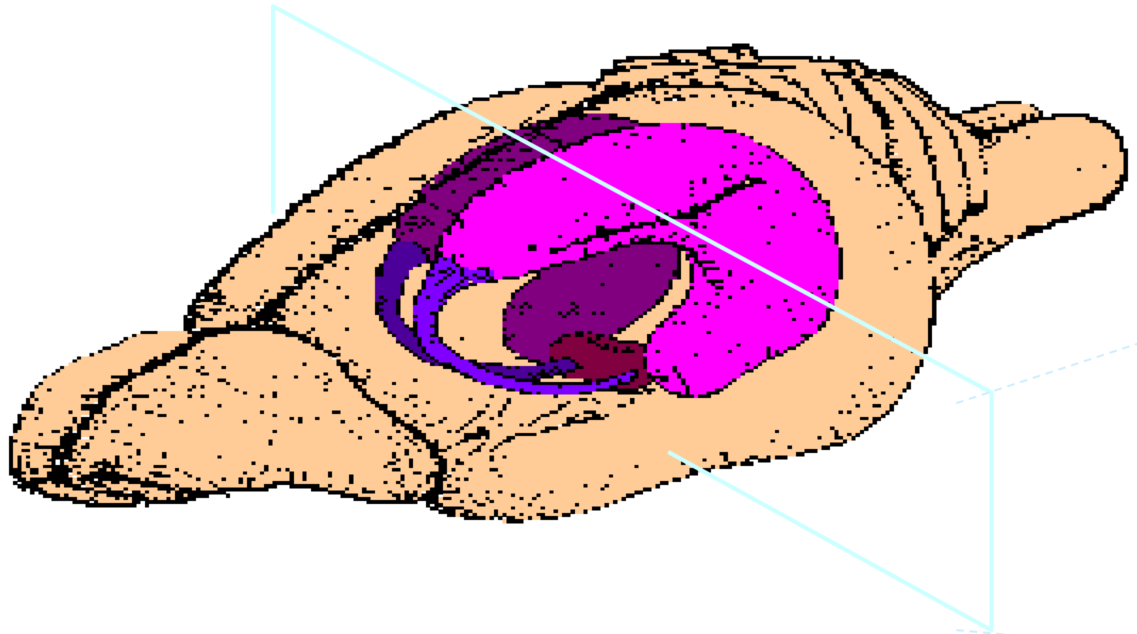
fig: Wikipedia

Henry Gray (1918) *Anatomy of the Human Body*

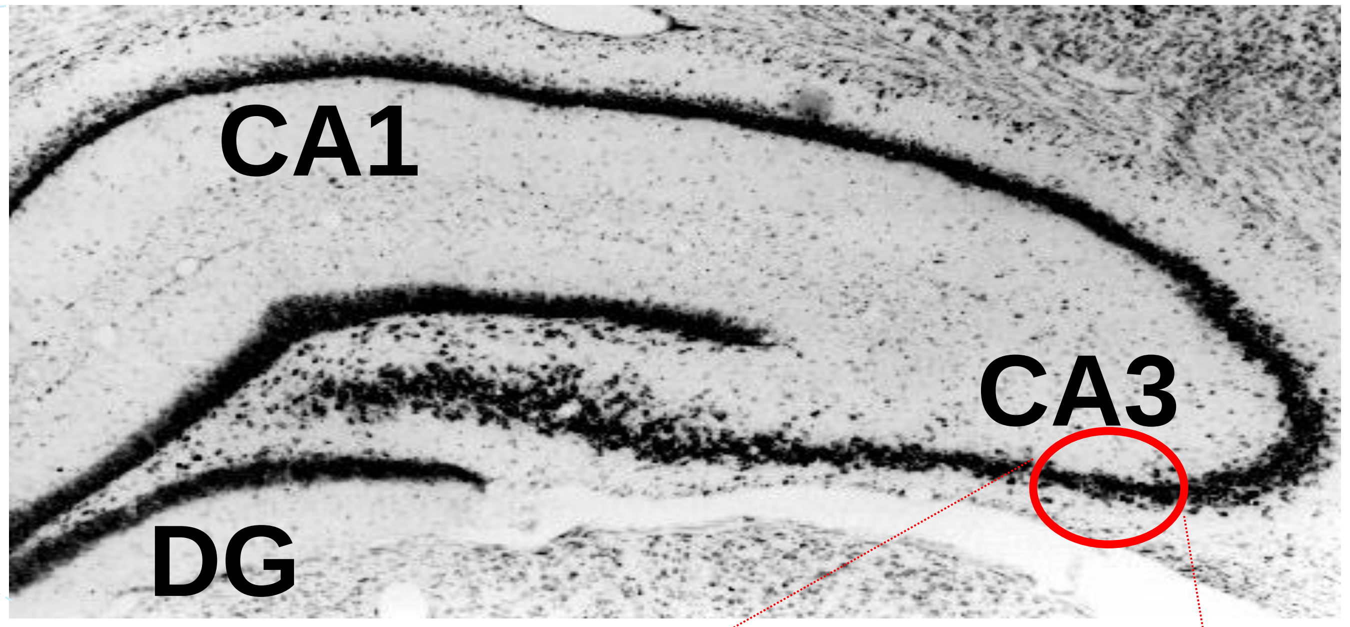
Previous slide.

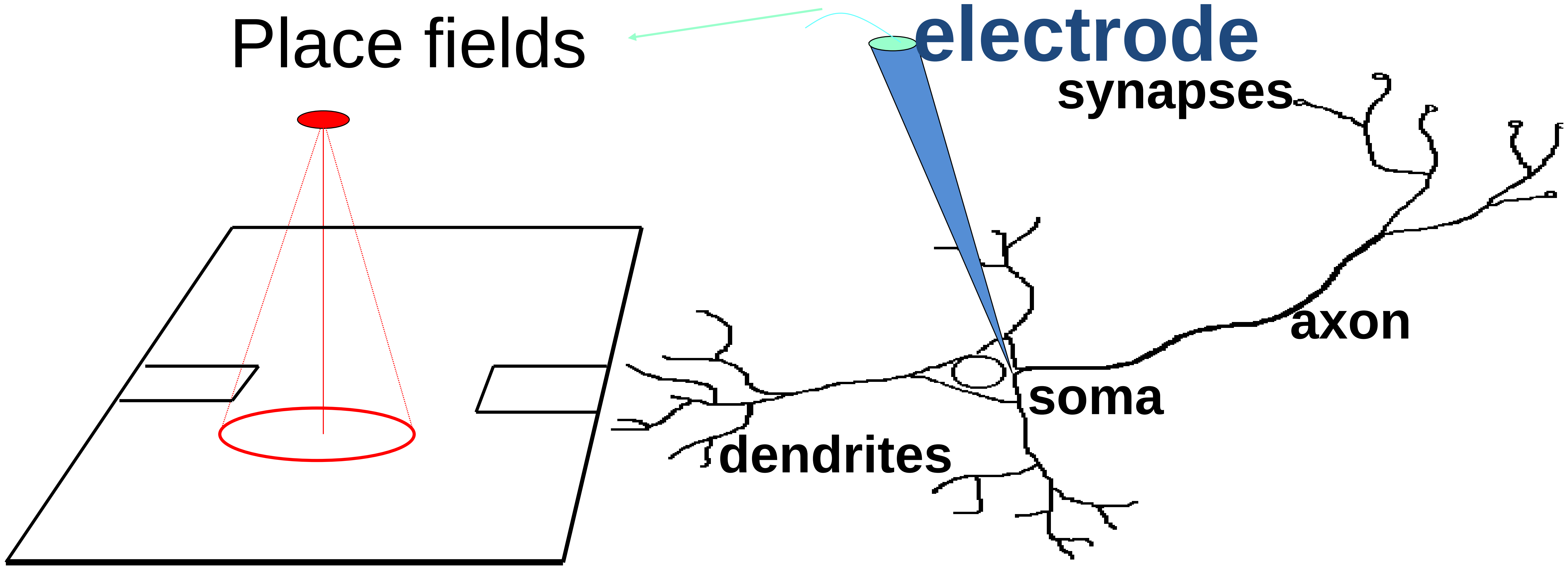
the problem of navigation needs the spatical representation of the hippocampus.

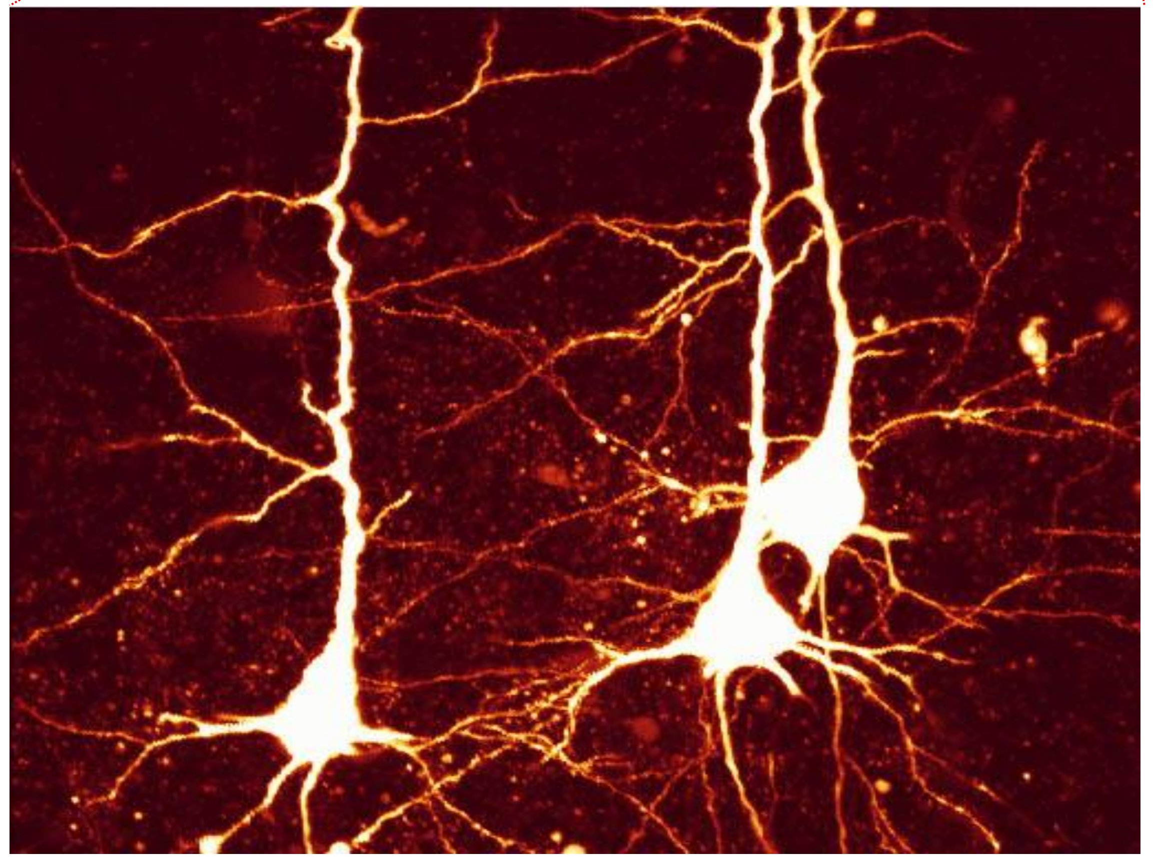
6. Place cells in rat hippocampus



rat brain







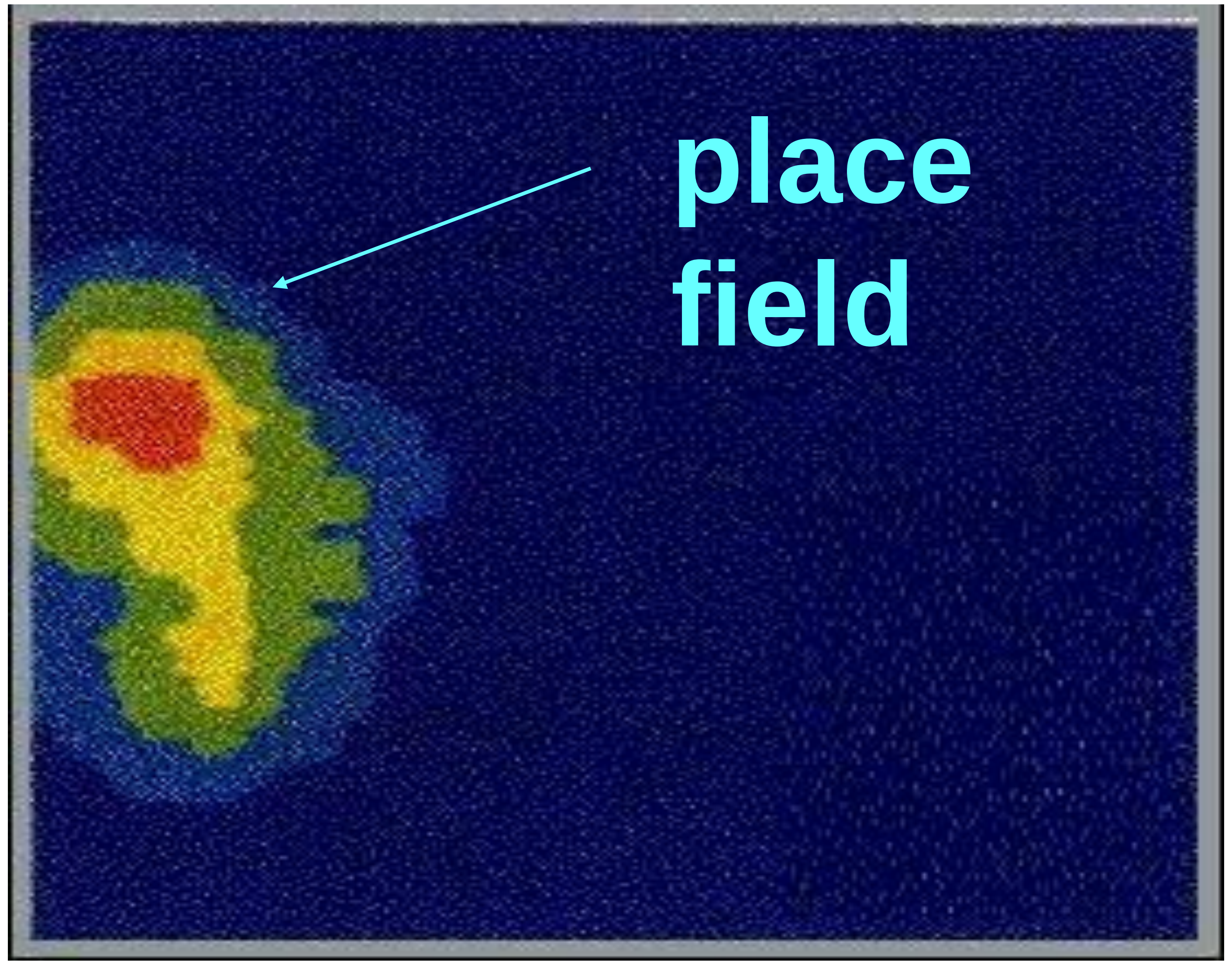
pyramidal cells

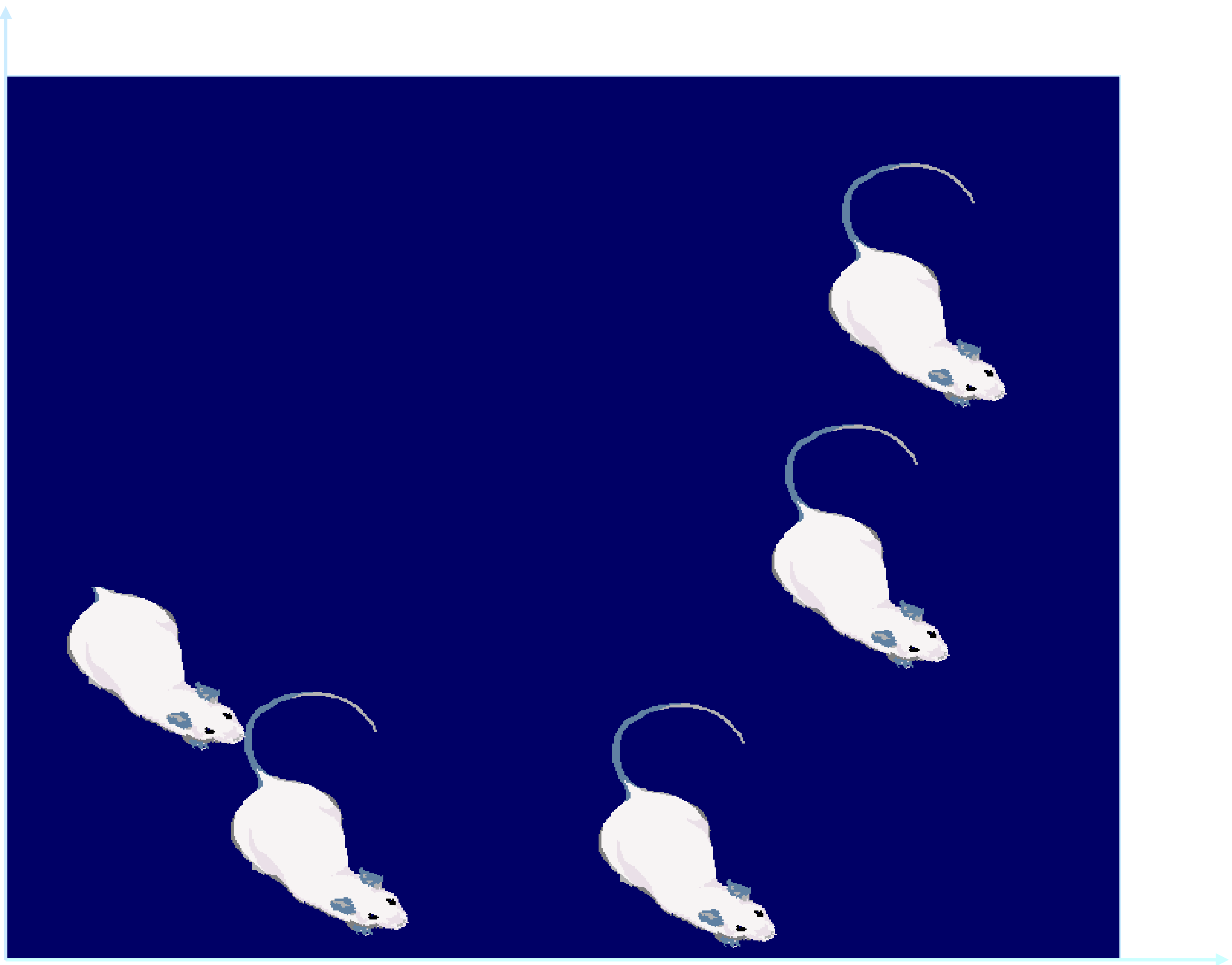
Previous slide.

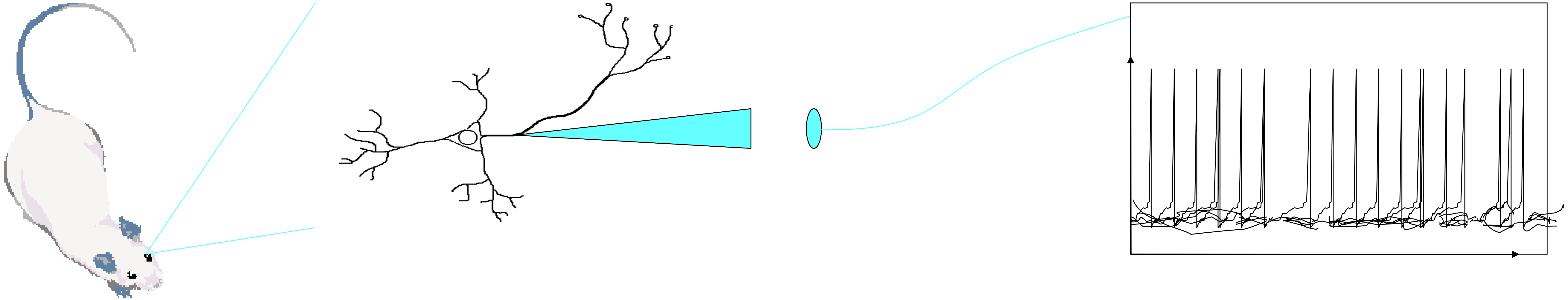
the hippocampus of rodents (rats or mice) looks somewhat different to that of humans. Importantly, cells in hippocampus of rodents respond only in a small region of the environment. For this reason they are called place cells. The small region is called the place field of the cell.

6. Hippocampal place cells

Main property: encoding the animal’s location







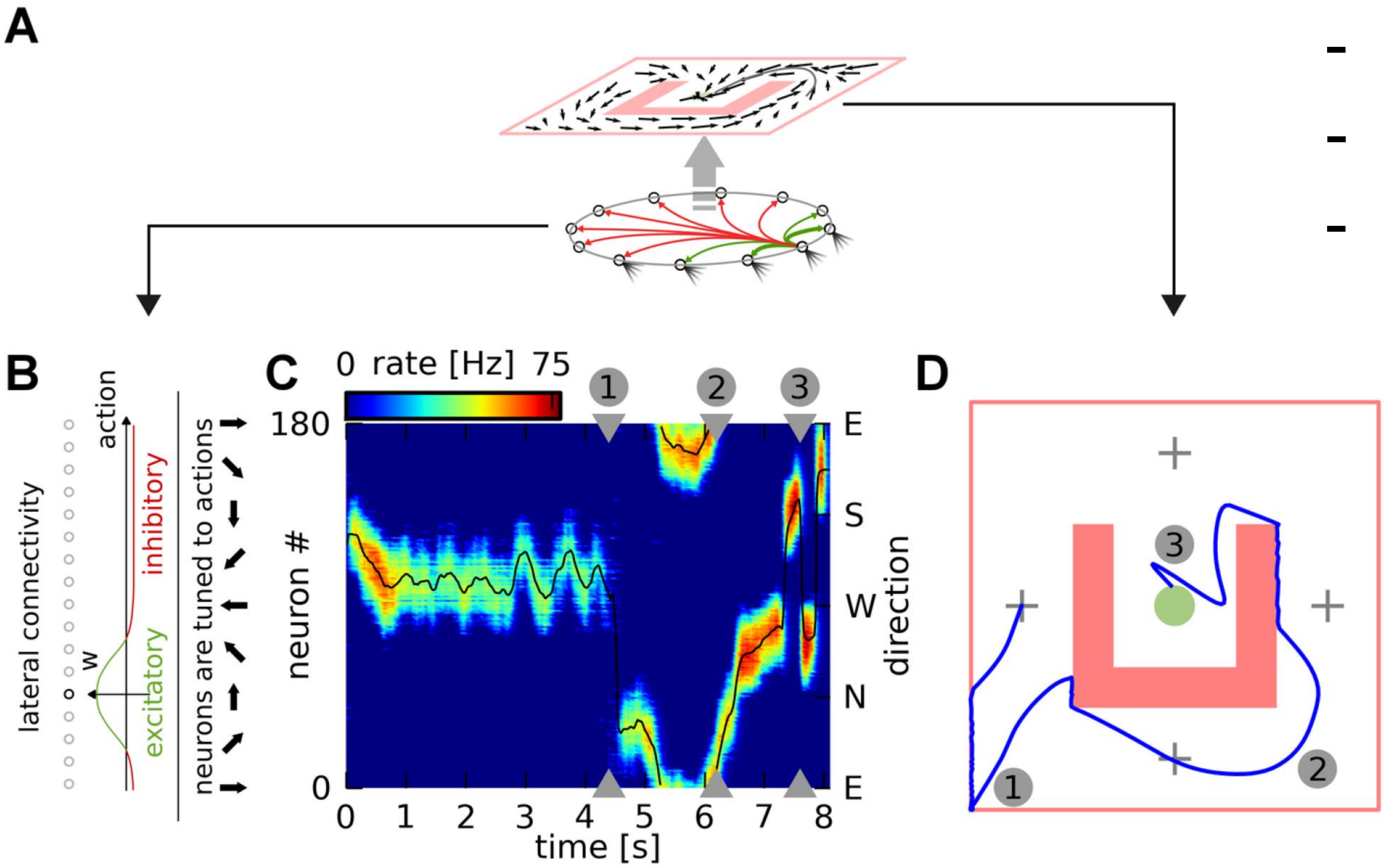
Previous slide.

Left: experimentally measured place field of a single cell in hippocampus.

Right: computer animation of place field

6. Representation of actions: Ring of spiking actor neurons

Note: no need to formally define a softmax function



- Local excitation
- Long-range inhibition
- Not a formal softmax
- Could be a model of action selection in striatum

Fremaux et al. (2013)

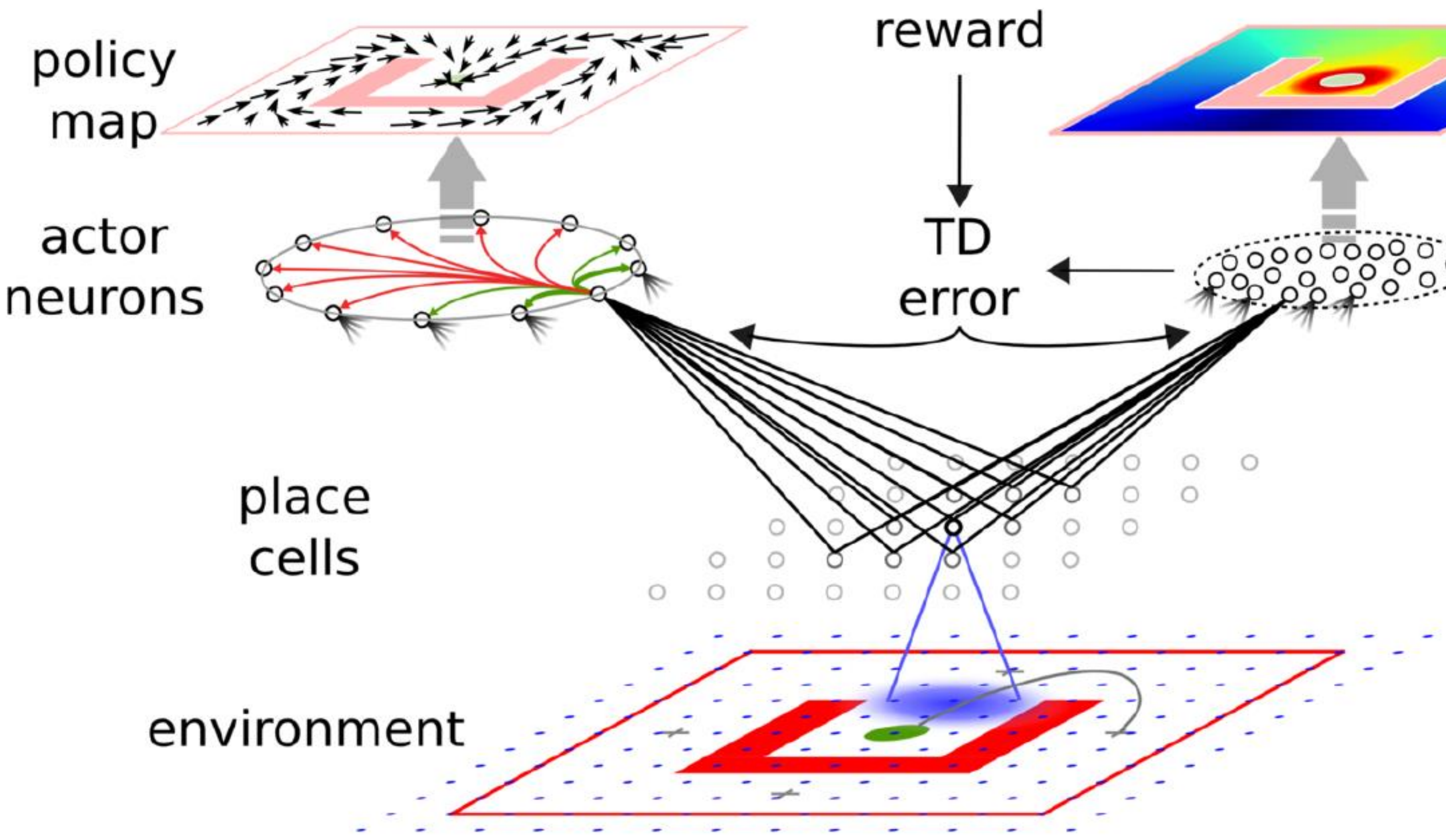
6. Ring of actor neurons

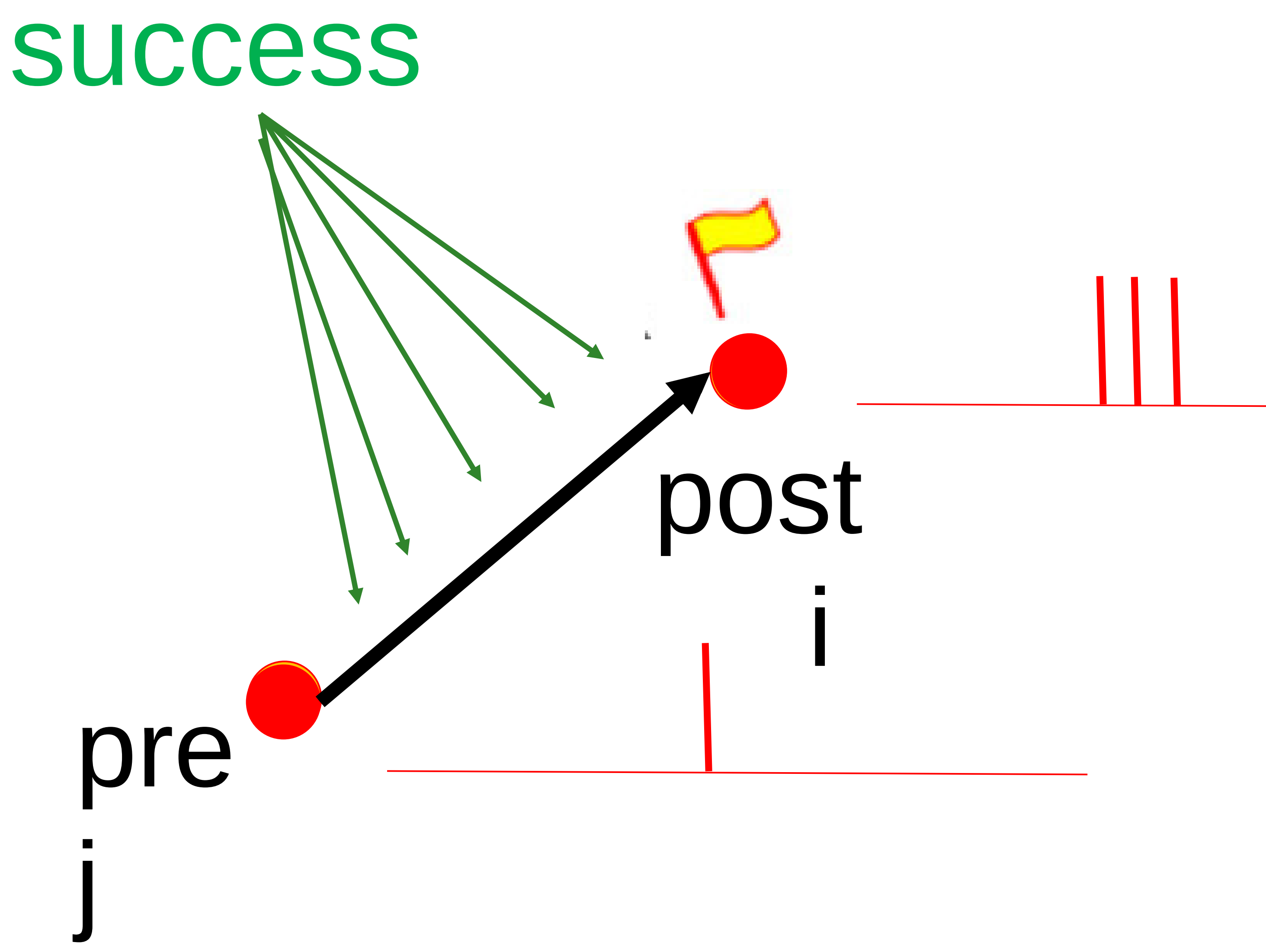
Actor neurons (previous slide).
A: A ring of actor neurons with lateral connectivity (bottom, green: excitatory, red: inhibitory) embodies the agent’s policy (top).
B: Lateral connectivity. Each neuron codes for a distinct motion direction. Neurons form excitatory synapses to similarly tuned neurons and inhibitory synapses to other neurons.
C: Activity of actor neurons during an example trial. The activity of the neurons (vertical axis) is shown as a color map against time (horizontal axis). The lateral connectivity ensures that there is a single bump of activity at every moment in time. The black line shows the direction of motion (right axis; arrows in panel B) chosen as a result of the neural activity.
D: Maze trajectory corresponding to the trial
shown in C. The numbered position markers match the times marked in C.

.

Fremaux et al. (2013)

6. Representation of Learning rule: Spikes + Eligibility trace





Fremaux et al. (2013)

Figure 1. Navigation task and actor-critic network. From bottom to top: the simulated agent evolves in a maze environment, until it finds the reward area (green disk), avoiding obstacles (red). Place cells maintain a representation of the position of the agent through their tuning curves. Blue shadow: example tuning curve of one place cell (black); blue dots: tuning curves centers of other place cells. Right: a pool of critic neurons encode the expected future reward (value map, top right) at the agent's current position. The change in the predicted value is compared to the actual reward, leading to the temporal difference (TD) error. The TD error signal is broadcast to the synapses as part of the learning rule. Left: a ring of actor neurons with global inhibition and local excitation code for the direction taken by the agent. Their choices depending on the agent's position embody a policy map (top left).

doi:10.1371/journal.pcbi.1003024.g001

6. Learning rule: Three-factor STDP for reward-based learning

Success signal:
TD error

$$\frac{dw_{ij}}{dt} = F(w_{ij}; \mathrm{PRE}_j\ , \mathrm{POST}_i\ \ , 3rd)$$

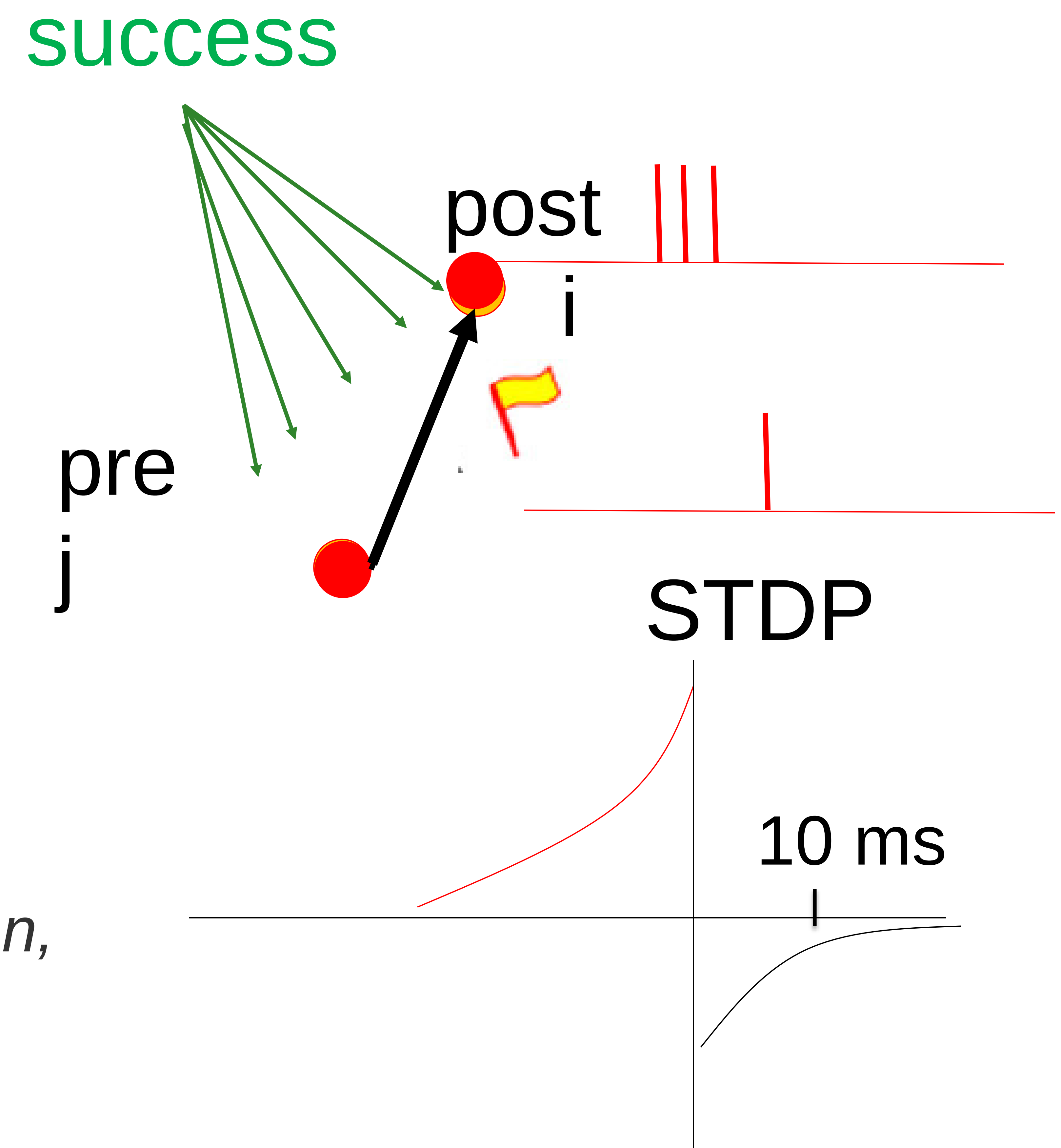
$$\tau \frac{d}{dt} e_{ij} = \mathrm{HEBB}_{ij} - e_{ij}$$

1s

$$\frac{d}{dt} w_{ij} = e_{ij} \cdot S(t)$$

Success signal

Hebb rule/eligibility trace



Xie and Seung 2003 Izhikevich, 2007; Florian, 2007; Legenstein et al., 2008,
Fremaux et al. 2010, 2013

6. Learning rule with TD in Actor-Critic for spiking neurons

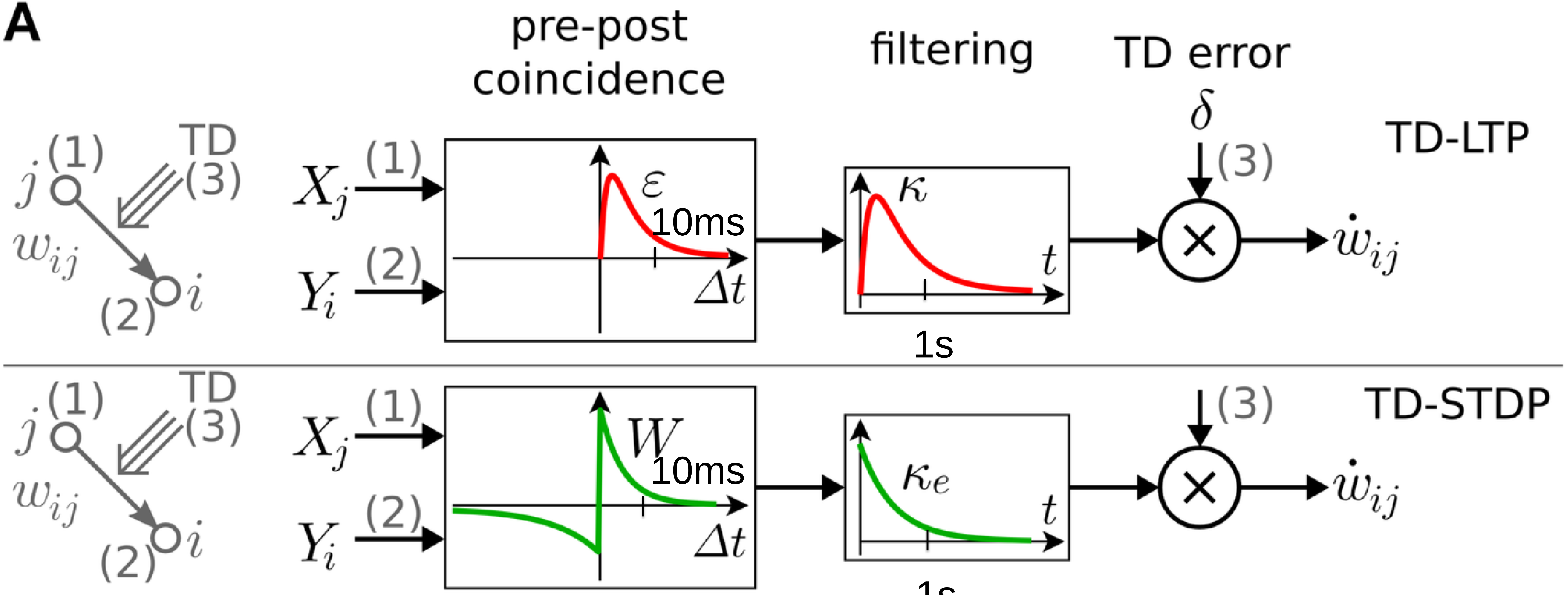
Learning rule with three factors (previous slide) based on spikes

1. In biology, neurons communicate by spikes (short electrical pulses).
2. Synaptic changes depend on the relative timing of the spikes of the sending (pre) and the receiving (post) neuron: Spike-Timing-Dependent Plasticity (STDP). Strong changes occur only if pre- and postsynaptic spikes coincide within +/- 20 ms.
3. STDP is used to set the eligibility trace. The eligibility trace decays on a much slower time scale of 1s.
4. Un success signal is necessary to transform the eligibility trace into an actual weight change.

Therefore weights increase if a success signal occurs within roughly one second after a coincident activity of pre- and postsynaptic neuron.

Fremaux et al. (2013)

6. Two variants of spike-based three-factor Learning rules



Condition for setting eligibility trace: 10 ms

Decay of eligibility trace : 1s

Fremaux et al. (2013)

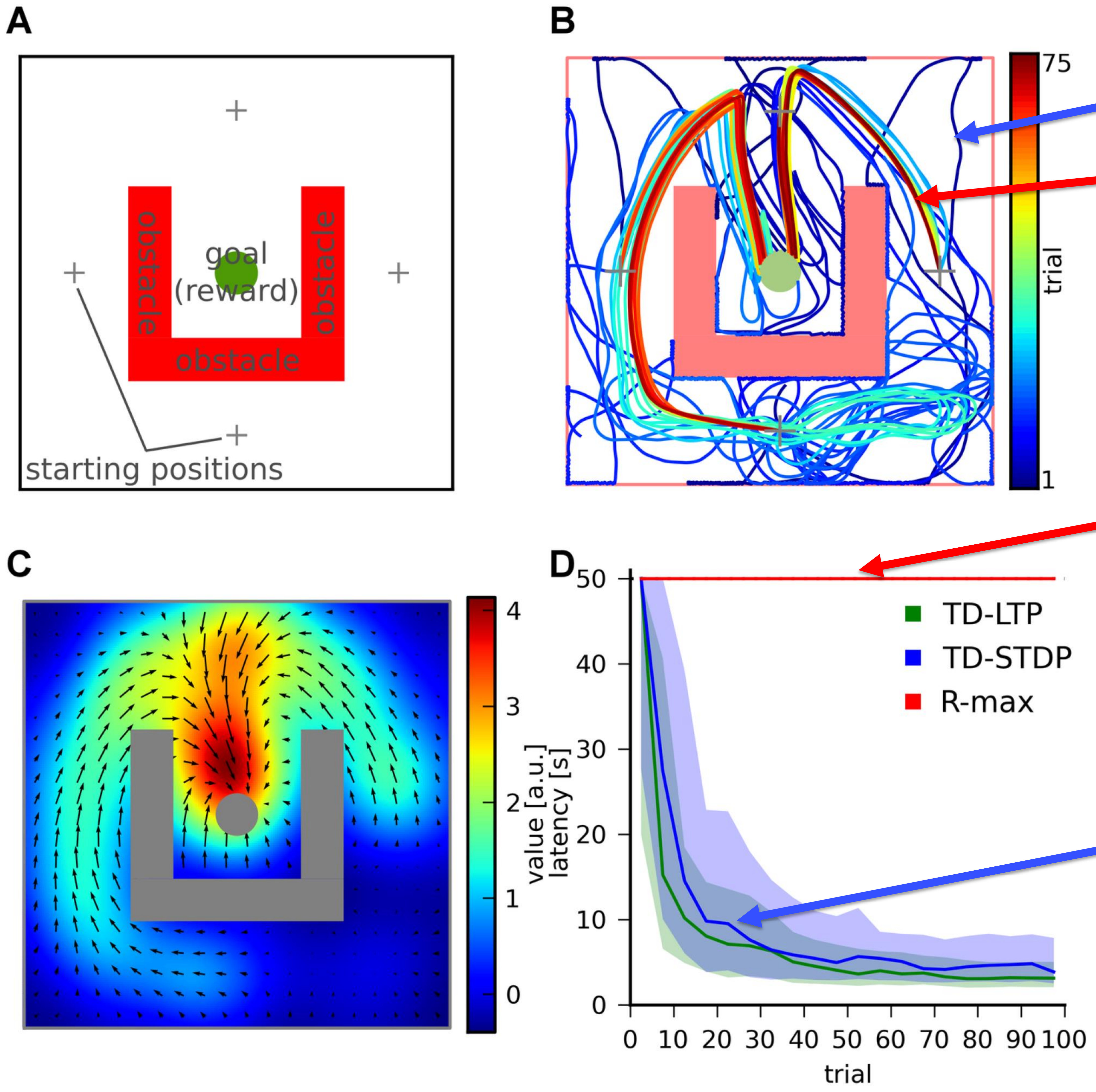
6. Learning rule with TD in Actor-Critic for spiking neurons

A: Learning rule with three factors (previous slide). We consider two different variants
Top: TD-LTP is the learning rule resulting from policy gradient. It works by passing the presynaptic spike train X_j (factor 1) and the postsynaptic spike train Y_i (factor 2) through a coincidence window ε. Spikes are counted as coincident if the postsynaptic spike occurs within after a few ms of a presynaptic spike. The result of the pre-post coincidence measure is low-pass-filtered by passing it through a kernel (which yields the eligibility trace, decaying of 1s), and then multiplied by the TD error $\delta(t)$ (factor 3) to yield the learning rule which controls the change of the synaptic weight w_ ij .

Bottom: TD-STDP is closer to biology and consists of a TD-modulated variant of STDP. The main difference with TD-LTP is the presence of a post-before-pre component in the coincidence window. As before, coincidences with 10ms set the eligibility trace

Fremaux et al. (2013)

6. Maze Navigation with TD in Actor-Critic with spiking neurons



early trial

Late trial

value map

R-max:
Policy gradient without the critic. The goal was never found within 50s.

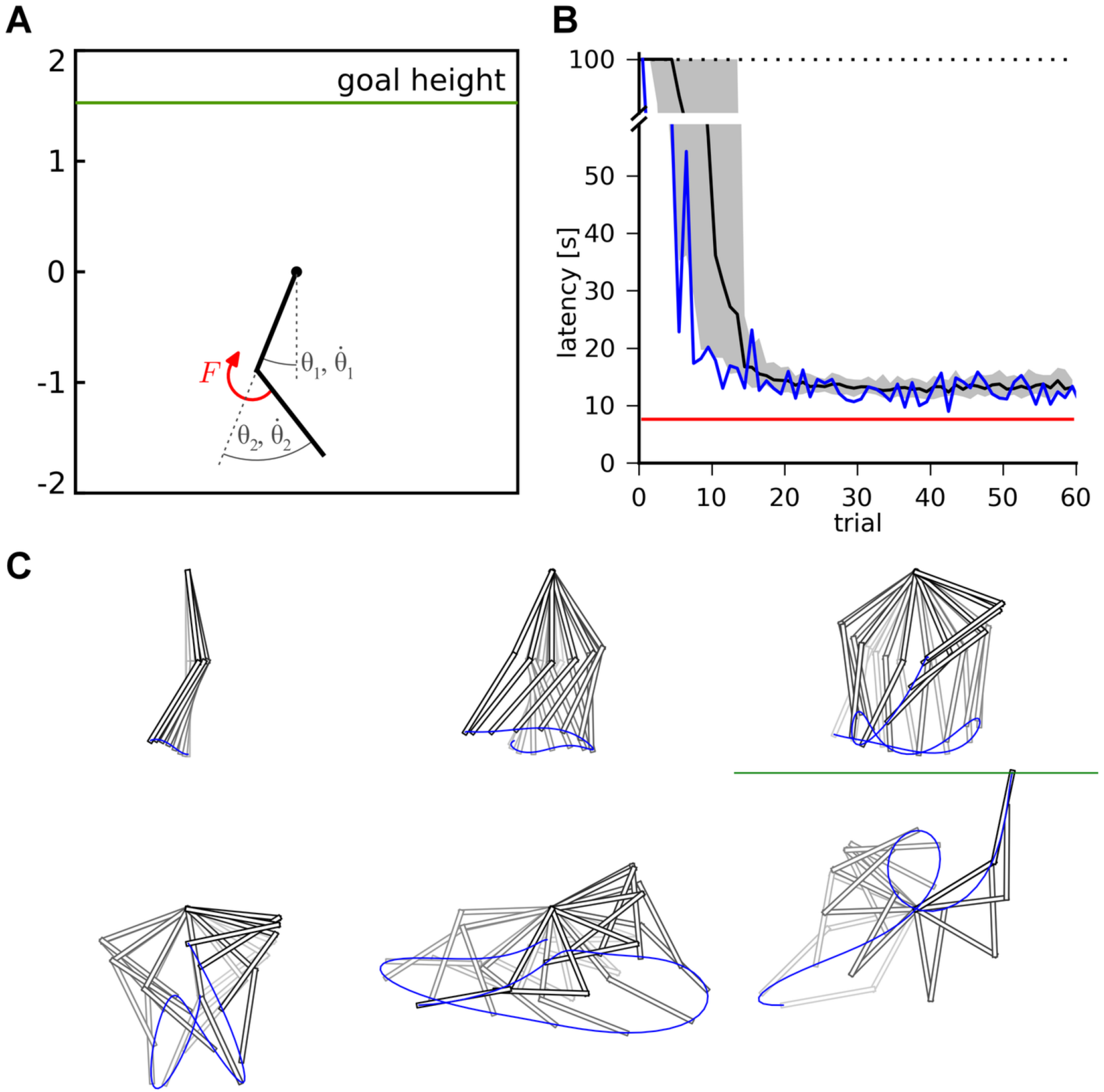
TD-STDP:
After 25 trials, the goal was found within 20s.

6. Maze Navigation with TD in Actor-Critic with spiking neurons

Maze navigation learning task. Both TD rules (TD-LTP and TD-STDP) work equally well. Hence, details of how the eligibility trace is set do not matter.
A: The maze consists of a square enclosure, with a circular goal area (green) in the center. A U-shaped obstacle (red) makes the task harder by forcing turns on trajectories from three out of the four possible starting locations (crosses).
B: Color-coded trajectories of an example TD-LTP agent during the first 75 simulated trials. Early trials (blue) are spent exploring the maze and the obstacles, while later trials (green to red) exploit stereotypical behavior.
C: Value map (color map) and policy (vector field) represented by the synaptic weights of the agent of panel B after 2000s simulated seconds.
D: Goal reaching latency of agents using different learning rules. Latencies of N=100 simulated agents per learning rule. The solid lines shows the median shaded area represents the 25th to 75th percentiles. The R-max learning rule is standard policy gradient agent without a critic and enters times-out after 50 seconds. Hence it is important that the 3rd factor is TD and not just ‘raw’ reward.

Fremaux et al. (2013)

6. Acrobot task with TD in Actor-Critic with spiking neurons



Fremaux et al. (2013)

Previous slide.
Application of the same model (spiking three-factor rule) to the Acrobot task.

6. TD in Actor-Critic with spiking neurons

- Learns in a few trials (assuming good representation)
- Works in continuous time.
- No artificial ‘time steps’
- Works with spiking neurons
- Works in continuous space and for continuous actions
- Uses a biologically plausible 3-factor learning rule
- Critic implements value function
- TD signal calculated by critic and broadcasted to network
- Actor neurons interact via synaptic connections
- No need for algorithmic ‘softmax’
- 3-factor rules with TD as global signal work much better than standard policy gradient (REINFORCE)

Fremaux et al. (2013)

Previous slide.

Summary of findings

6. Summary Learning in the Brain vs RL algo

Advantage Actor-Critic Reinforcement learning needs:

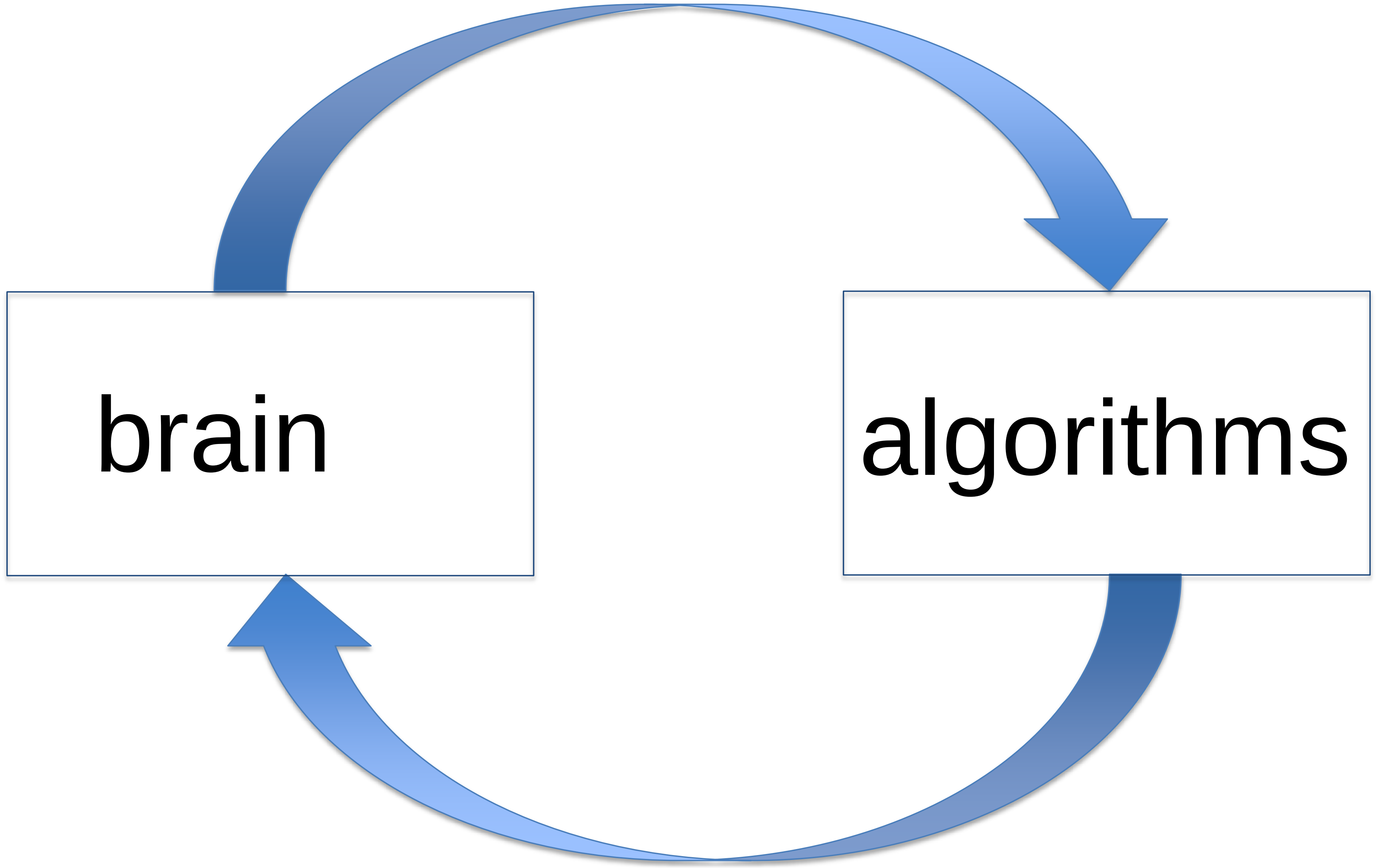
- states / sensory representation

- action selection

- value function/critic

- broad cast of TD error

- TD error calculation



6. Summary

Several aspects of TD learning in an actor-critic framework can be mapped to the brain:

Sensory representation: Cortex and Hippocampus
Actor : Dorsal Striatum
Critic : Ventral Striatum (nucleus accumbens)

TD-signal: Dopamine

Learning in a few trials (not millions!) possible, if the sensory presentation is well adapted to the task

6. Summary

Learning outcome: RL learning rules and the brain

- **three-factor learning rules can be implemented by the brain**
  - → synaptic changes need presynaptic factor, postsynaptic factor and a neuromodulator (3rd factor)
  - → actor-critic and other policy gradient methods give rise to very similar three-factor rules
- **eligibility traces as 'candidate parameter updates'**
  - → set by joint activation of pre- and postsynaptic factor
  - → decay over time
  - → transformed in weight update if dopamine signal comes
- **the dopamine signal has signature of the TD error**
  - → responds to reward minus expected reward
  - → responds to unexpected events that predict reward