Learning in the Brain: Eligibility Traces and 3-factor Rules

Wulfram Gerstner
EPFL, Lausanne

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Many years of work together with:
Claudia Clopath , Nicolas Fremaux, Michael Herzog, Richard Kempter,
Marco Lehmann, Jean-Pascal Pfister, Kerstin Preuschof,
Henning Sprekeler, Tim Vogels, Eleni Vasilaki, Friedemann Zenke

Funding acknowledged: ERC, Brain-i-Nets, HBP, Swiss Natl. Sci. Foundation, EPFL

Memory Formation

- stream of inputs
- lasts (sometimes)

How do we remember?

Examples:

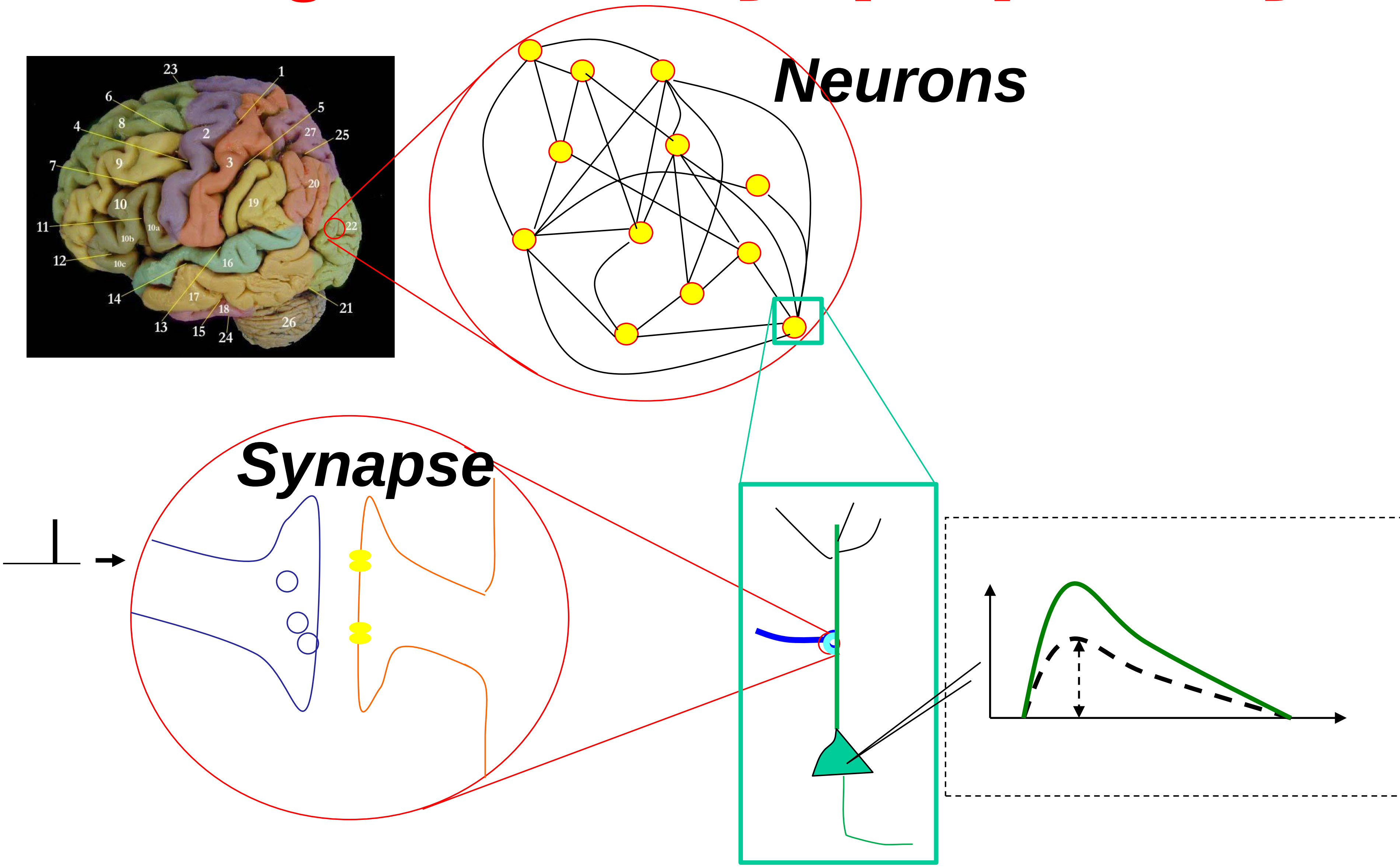
-Highway
-School
-Traumatic memories

Learning skills:

-table tennis, skiing, biking, piano

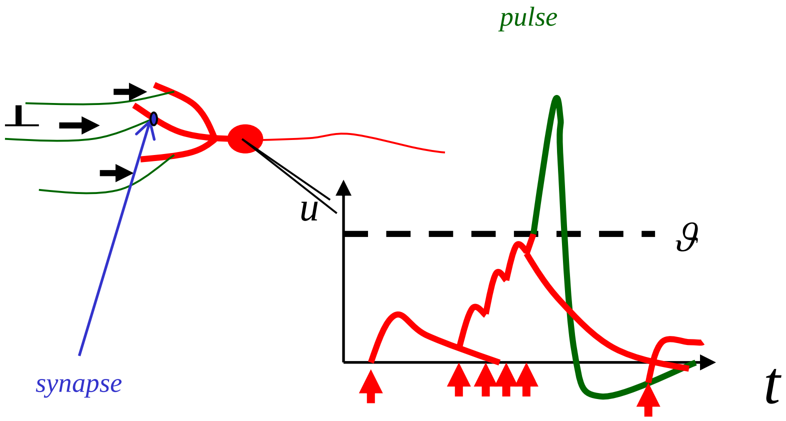
How do we learn?

Learning – based on synaptic plasticity



Synaptic Plasticity =Change in Connection Strength

The brain: neurons sum their inputs



- Spikes arrive
- Summation of Responses
- Threshold for spike emission

Eligibility Traces and 3-factor Rules of Synaptic Plasticity

√ - Introduction

- Hebbian Learning and STDP: a Framework

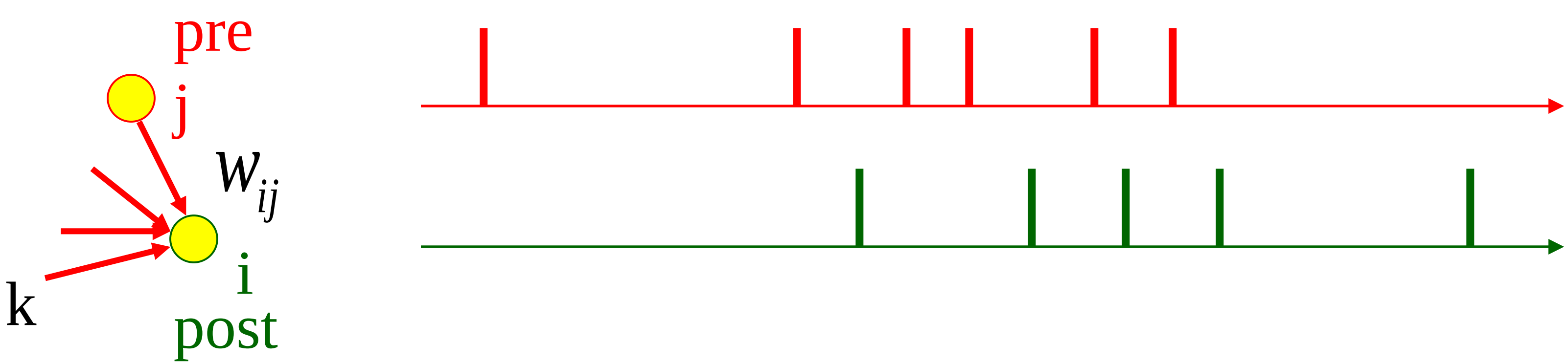
- 3-factor rules: a Framework

- Example: Learning in Mazes

- Example: Behavioral Eligibility Trace

- Summary

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*

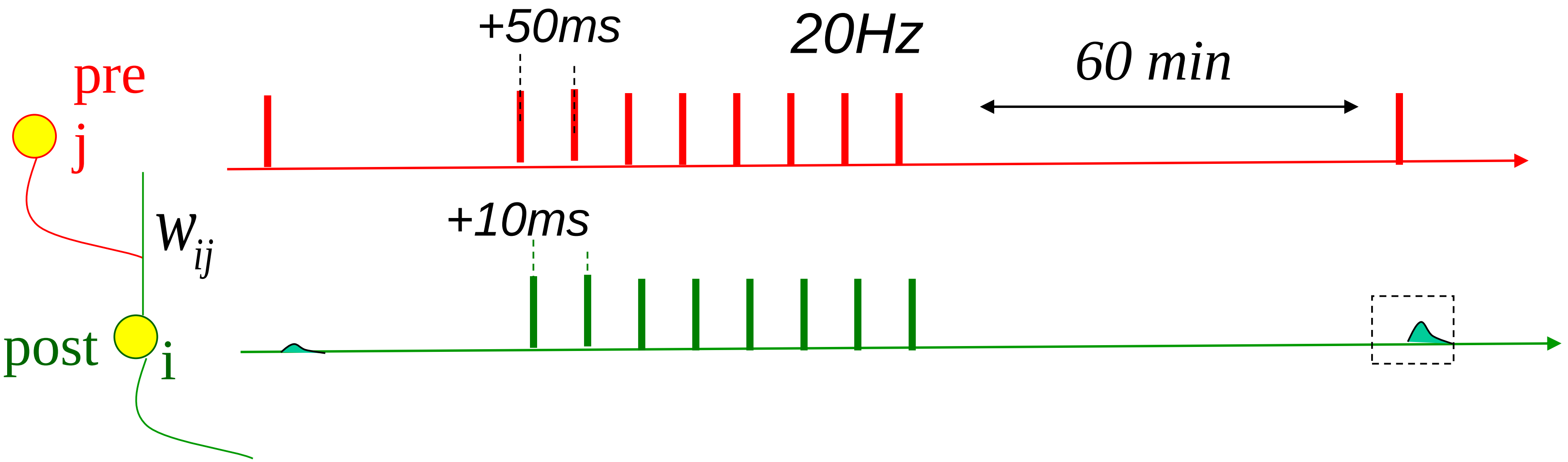
Hebb, 1949

***'active together'* → synapse strengthened**

Experiments: *Bliss and Lomo 1973, Levy and Stewart, 1983, ...*
Markram et al. 1997, Bi and Poo, 1998, ...

Reviews: *Bliss and Collingridge, 1993, Sjostrom et al. 2008...*
Markram et al. 2011, ...

Hebbian Learning with spikes

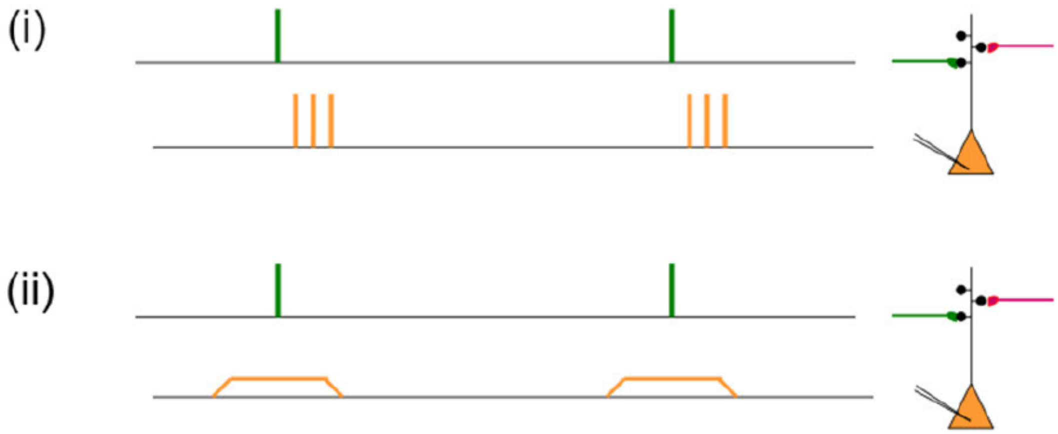


Long-Term Potentiation:

The effect lasts for a long time

(hours, days, weeks …)

Hebbian rules (2-factor rules)



Hebbian

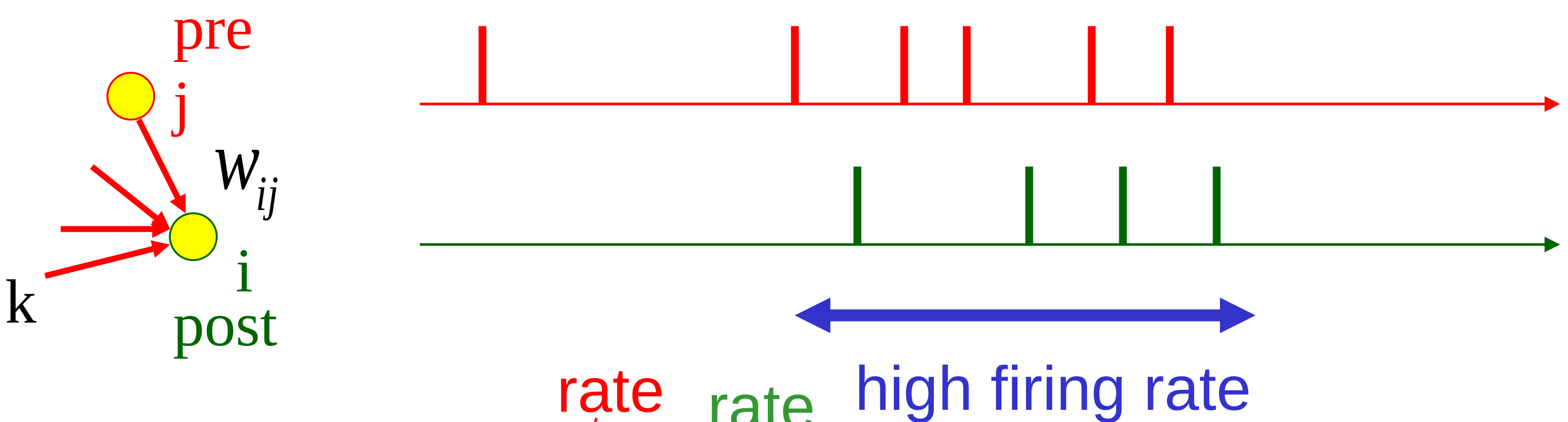
Hebbian

'active together' **→**

green synapse strengthened

(but not the red one)

Synaptic Plasticity (rate models)



rate rate

$$\frac{dw_{ij}}{dt} = F(w_{ij}; \nu_j^{pre}, \nu_i^{post})$$

- local rule

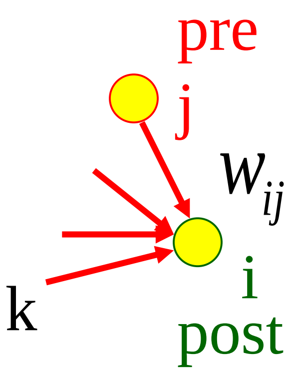
- simultaneously active

Hebbian

$$\frac{dw_{ij}}{dt} = a_0 + a_1^{pre}\nu_j^{pre} + a_1^{post}\nu_i^{post} + a_2^{corr}\nu_j^{pre}\nu_i^{post} +$$

depend on w_{ij}

Induction of plasticity



- **homosynaptic/Hebb** *('pre'* ***and*** *'post')*
- **heterosynaptic plasticity** *(pure 'post'-term)*
- **transmitter-induced** *(pure 'pre'-term)*

Hebbian

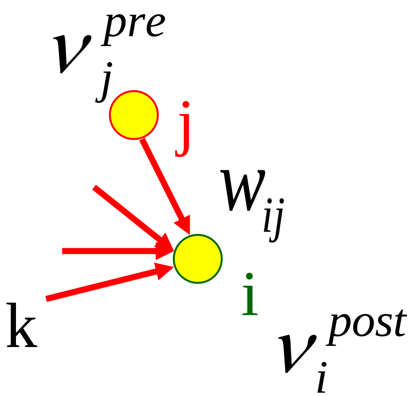
$$\frac{dw_{ij}}{dt} = a_0 + a_1^{pre}\nu_j^{pre} + a_1^{post}\nu_i^{post} + a_2^{corr}\nu_j^{pre}\nu_i^{post} +$$

Hebbian

$$+a_3^{BCM}\nu_j^{pre}(\nu_i^{post})^2 \quad +a_4^{post}(w_{ij})[\nu_i^{post}]^4$$

$$+ \ldots$$

Hebbian Learning: rate models (1980-1990)



all 4 are Hebbian models
all 4 are local models
many other combinations possible

$$\frac{dw_{ij}}{dt} = a_2^{corr} \nu_j^{pre} \nu_i^{post}$$

$$\frac{dw_{ij}}{dt} = a_2^{corr} \nu_j^{pre} \nu_i^{post} - c$$

$$\frac{dw_{ij}}{dt} = a_2^{corr} \nu_j^{pre} (\nu_i^{post} - \vartheta)$$

$$\frac{dw_{ij}}{dt} = a_3^{BCM} \; \nu_j^{pre} (\nu_i^{post})^2$$

pre	on	off	on	off
post	on	on	off	off
	+	0	0	0
	+	-	-	-
	+	0	-	0
	+	0	0	0

Induction of plasticity

1. **Hebbian Learning is a ‘framework’**
   → not a single ‘rule’

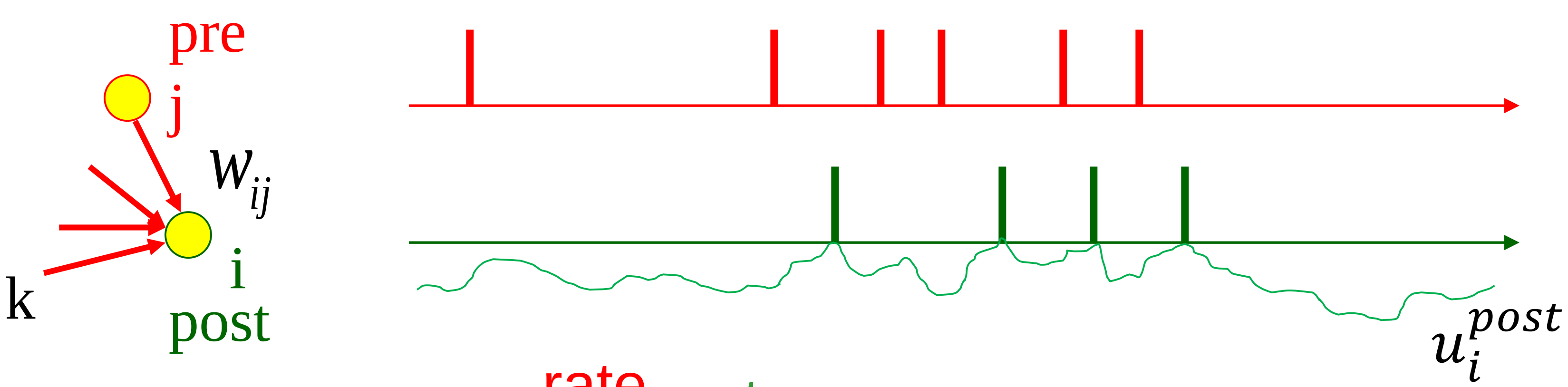
2. **Hebbian = all ‘local’ rules with at least one homosynaptic term** $(\nu_j^{pre})^n(\nu_i^{post})^m$
   → *‘pre’* **and** *‘post’* **together**

3. **Hebbian Learning may also contain other terms**
   → heterosynaptic plasticity *(pure ‘post’-term)*
   → transmitter-induced *(pure ‘pre’-term)*

4. **Suitable combination of these terms enables formation of memories (assemblies) in networks, as well as modeling of spine dynamics**

Zenke et al., Nat. Comm., 2015, Deger et al. Cerebral Cortex, 2018;

Synaptic Plasticity (models)



rate rate

$$\frac{dw_{ij}}{dt} = F(w_{ij;} \nu_j^{pre}, \nu_i^{post})$$

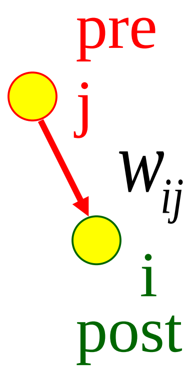
- **local rule**
- **simultaneously active**

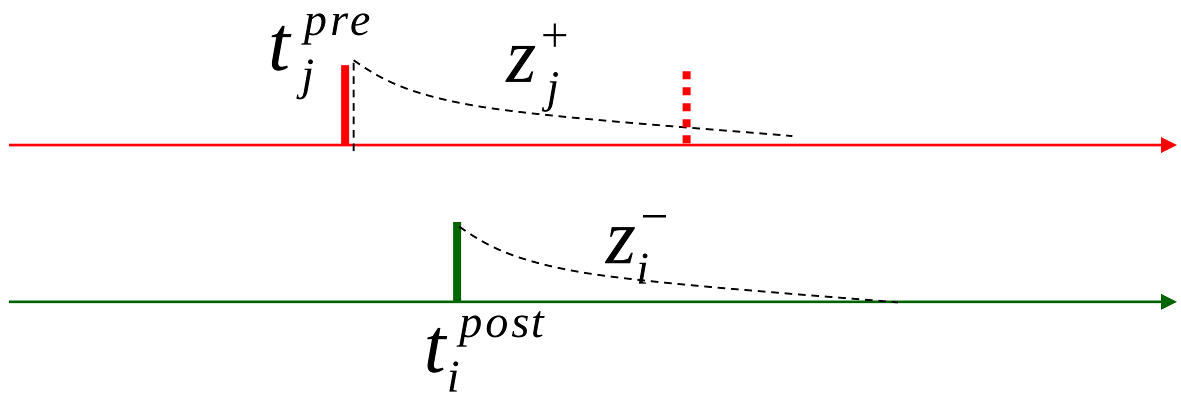
BUT: Firing Rate is not all

$$\frac{dw_{ij}}{dt} = F(w_{ij}; \mathrm{spikes}_j^{pre}, spikes_i^{post}, u_i^{post})$$

voltage

Spike-timing dependent plasticity: ‘traces’ for STDP





$$\tau_+ \frac{d}{dt} z_j^+ = -z_j^+ + \delta(t - t_j^{pre})$$

jump at presyn. spike

$$\tau_- \frac{d}{dt} z_i^- = -z_i^- + \delta(t - t_i^{post})$$

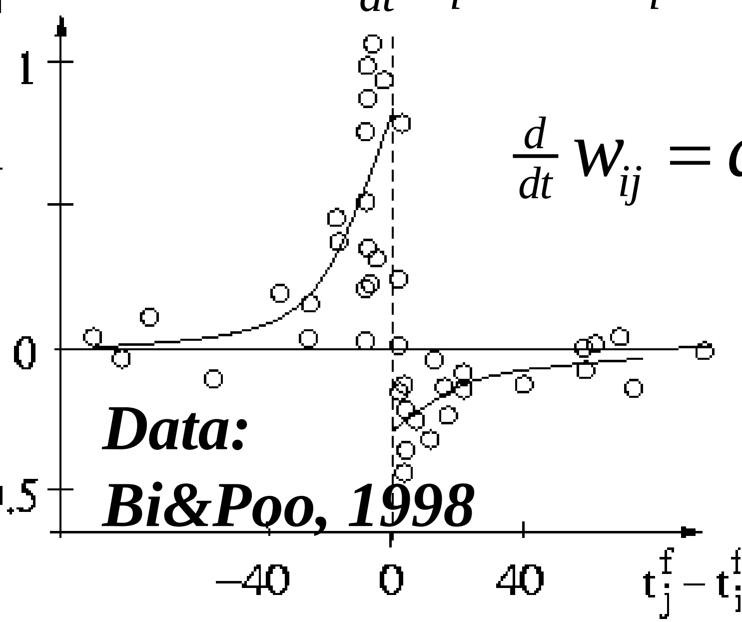
jump at postsyn. spike

Hebbian

$$\frac{d}{dt} w_{ij} = a(w_{ij}) z_j^+ \delta(t - t_i^{post}) - b(w_{ij}) z_i^- \delta(t - t_j^{pre})$$

pre-before-post

post-before-pre

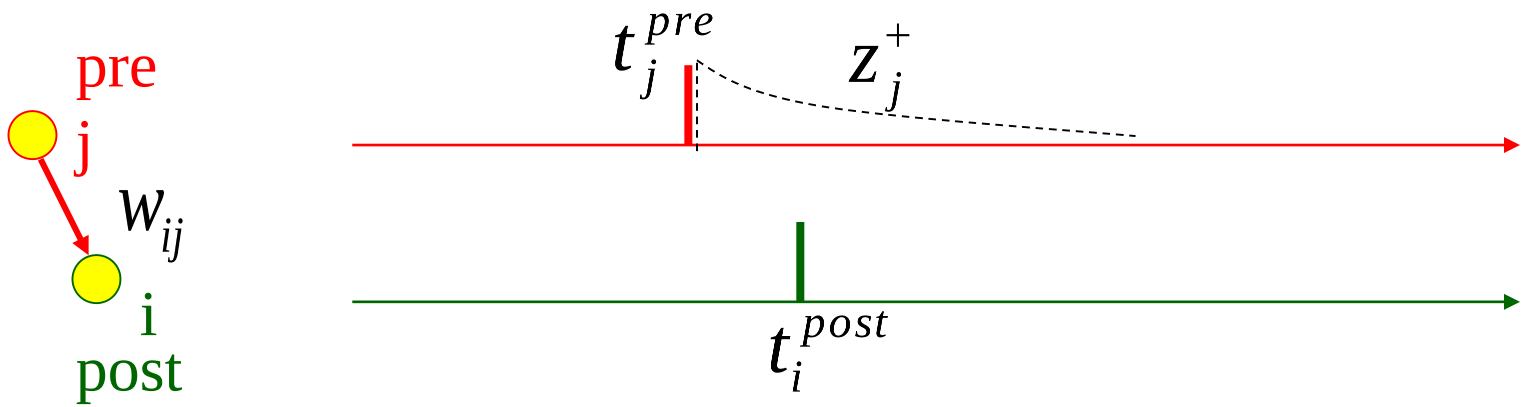


Data: Bi&Poo, 1998

Simple STDP models

(Gerstner et al. 1996, Kempter et al. 1999, Kistler-van Hemmen 2000, … Song-Miller-Abbott 2000, Senn et al. 2001,)

Local rules with spikes



- Each spike arrival leaves a trace at the synapse
- Trace 'read out' at moment of post-spike
- Implements PRE and POST 'together'

STDP = spike-based 'Hebbian' learning
→ not a single rule, but a framework
→ many terms combined in common
 STDP rules
 (homosynaptic, heterosynaptic, …)

Summary this part: HEBBian Learning

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

- Framework for local learning rules

- PRE and POST: homosynaptic (‘Hebbian’)
- POST-only: heterosynaptic
- PRE-only: transmitter-induced

- PRE **stands for**:
    - spike arrival at synapse
    - trace left by neurotransmitted at synapse
- POST **stands for**
    - BPAP
    - voltage at location of synapse
    - trace left by voltage (e.g., Ca, 2^{nd} messenger)

Memory

1) How do we remember?

1') How is memory generated in synapses?

1'') Spine dynamics, LTP/LTD experiments?

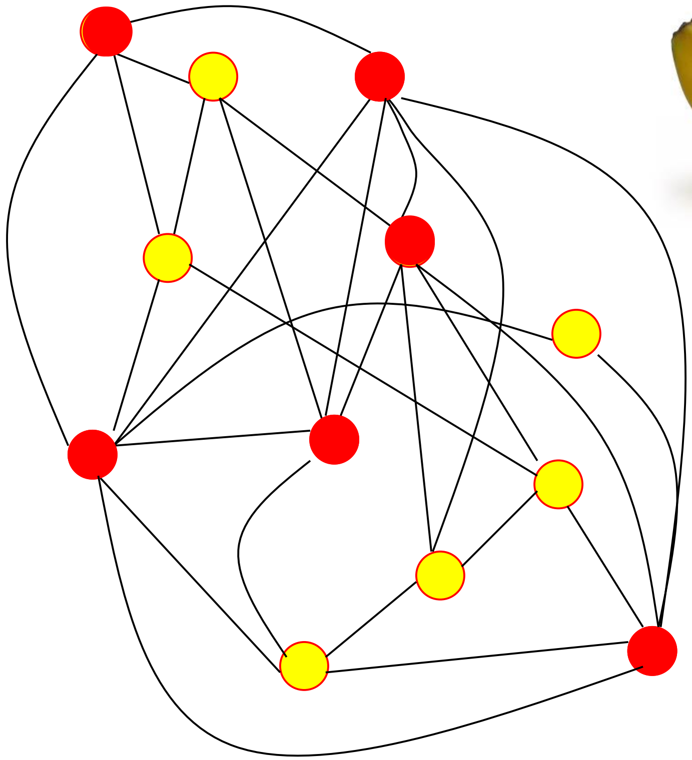
→ Hebbian learning
is a good candidate

→ Build synaptic plasticity models of the form

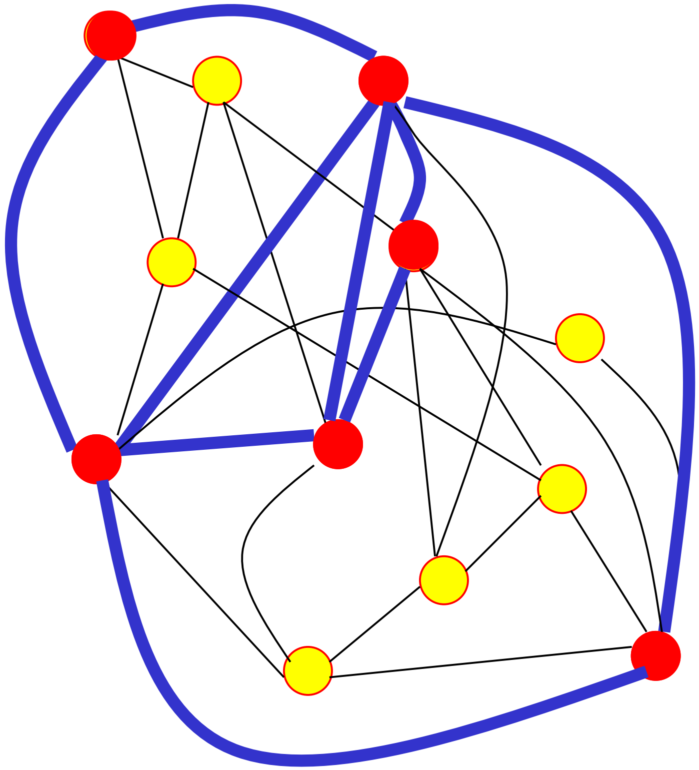
$$\frac{dw_{ij}}{dt} = F(w_{ij}; \mathrm{PRE}_j \;, \mathrm{POST}_i \;)$$

<u>Does this really work?</u>

Hebbian Learning



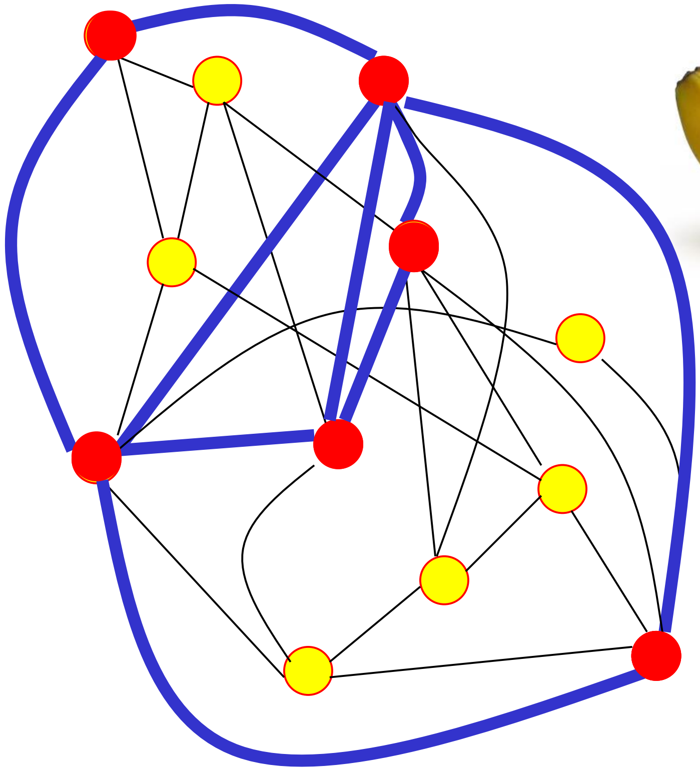
Hebbian Learning

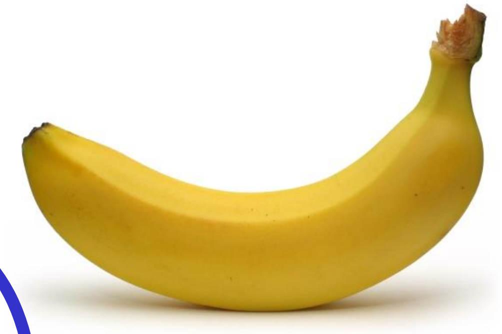


item memorized

Hebbian Learning

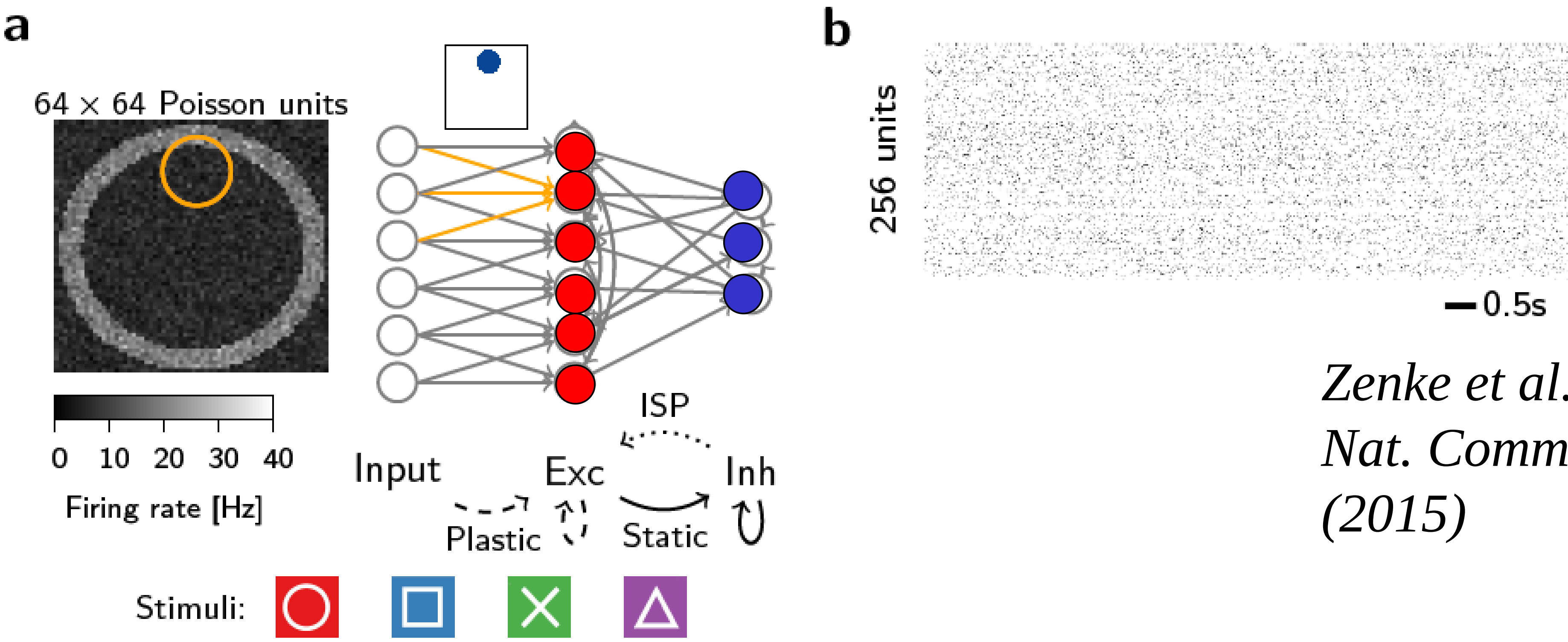
Recall:
Partial info





item recalled

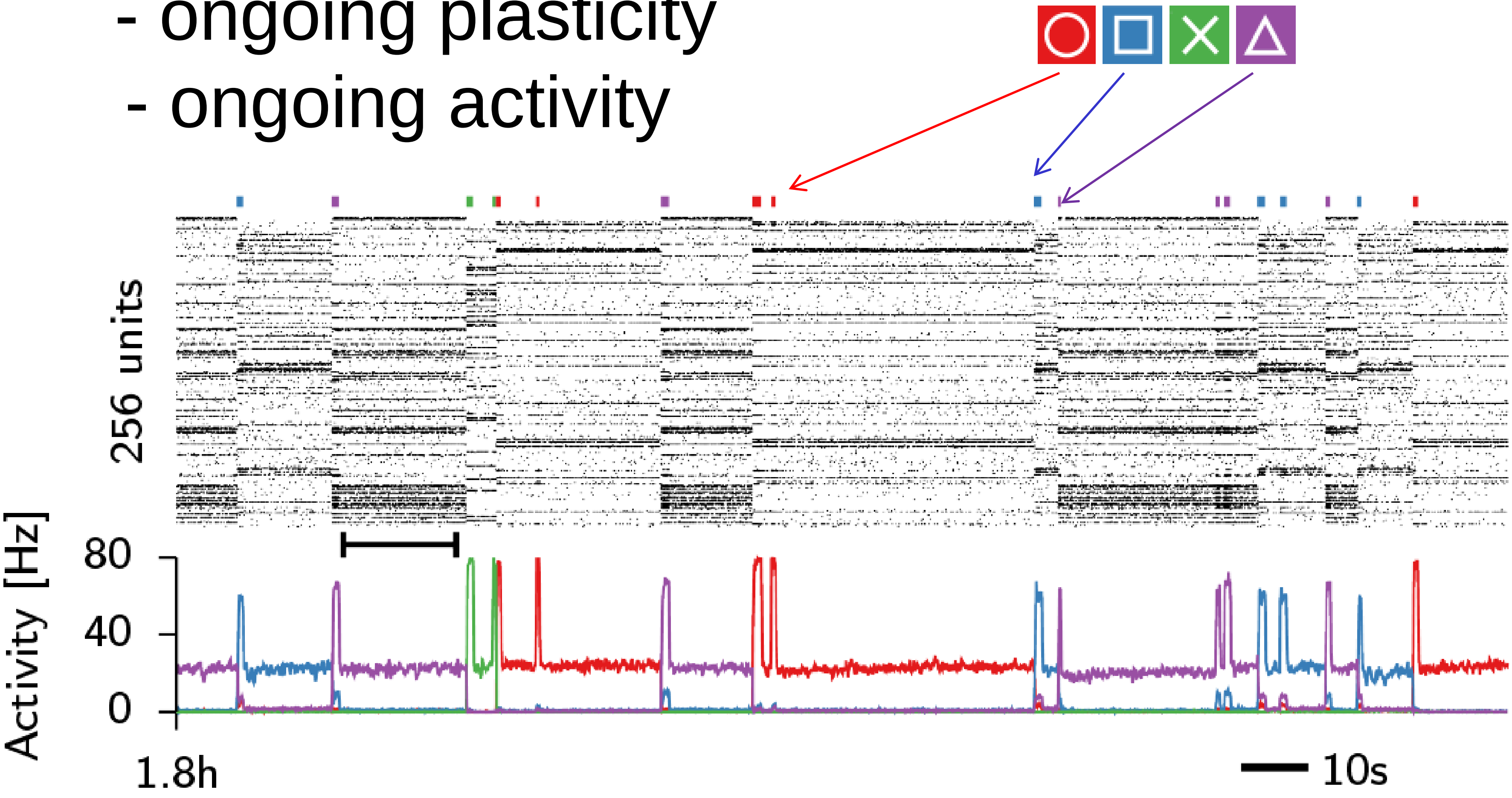
Plasticity in feedforward /recurrent connections



Zenke et al., Nat. Comm. (2015)

Stable memory recall despite

- ongoing plasticity
- ongoing activity



Zenke et al., Nat. Comm. (2015)

Summary this part:

- **Local learning rules (generalized Hebb)**
  - STDP + other terms: ‘orchestrated’
  - terms with two factors (pre and post) induce weight change
  - needs heterosynaptic plasticity (post-only)
    → controls weight growth/ network activity
- **stable memory formation**
- **stable recall despite**
  - ongoing plasticity
  - ongoing activity

Eligibility Traces and 3-factor Rules of Synaptic Plasticity

√ - Introduction

√ - Hebbian Learning: a Framework

→ - 3-factor rules: a Framework

- Example: Learning in Mazes

- Example: Behavioral Eligibility Trace

- Summary

Memory

1)   How do we remember?

1') How is memory generated in synapses?

1'') Spine dynamics, LTP/LTD experiments?

→ Hebbian learning
is a good candidate

Learning skills:

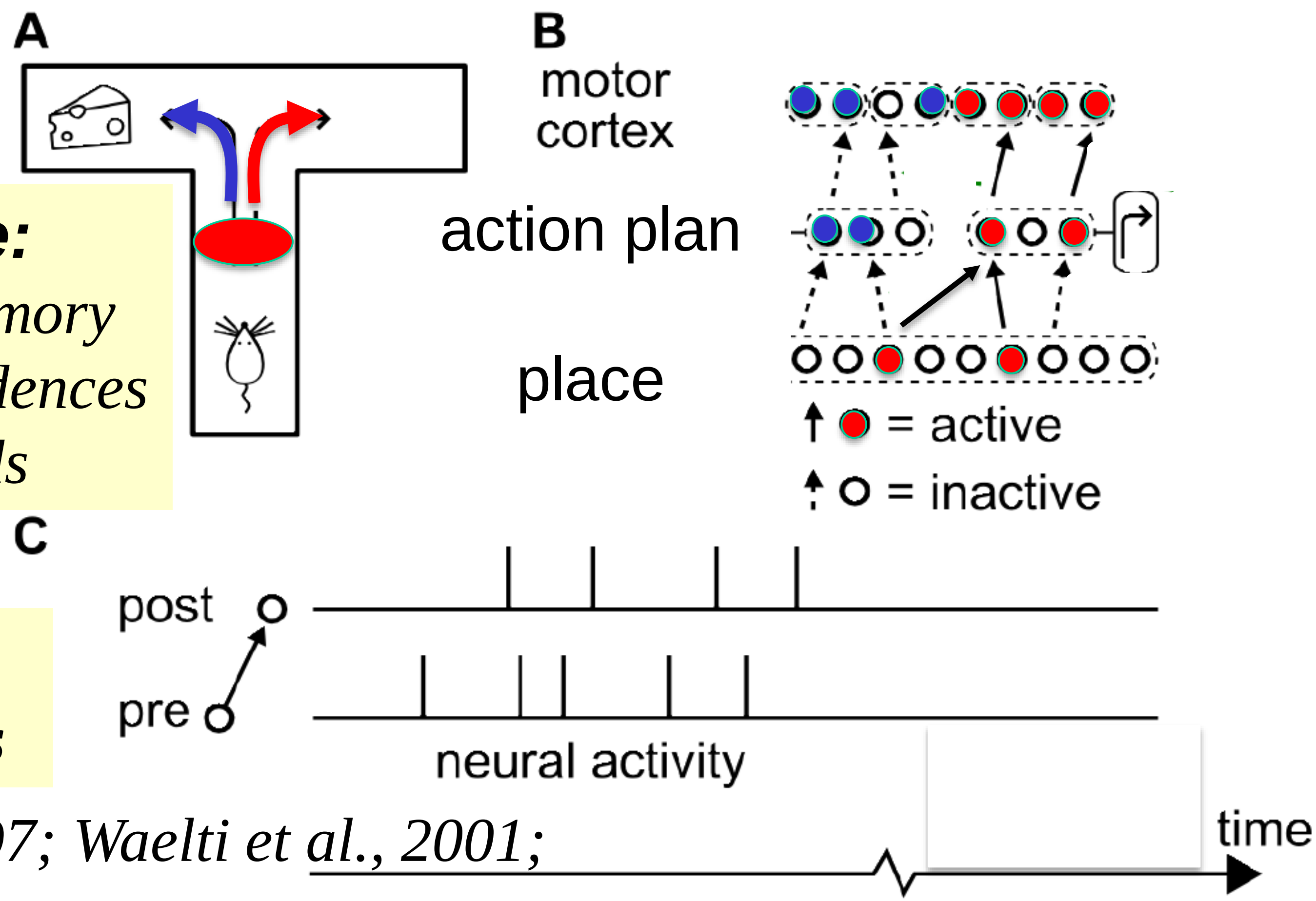
-table tennis, skiing, biking

2) How do we learn?

2')What are good learning rules?

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;

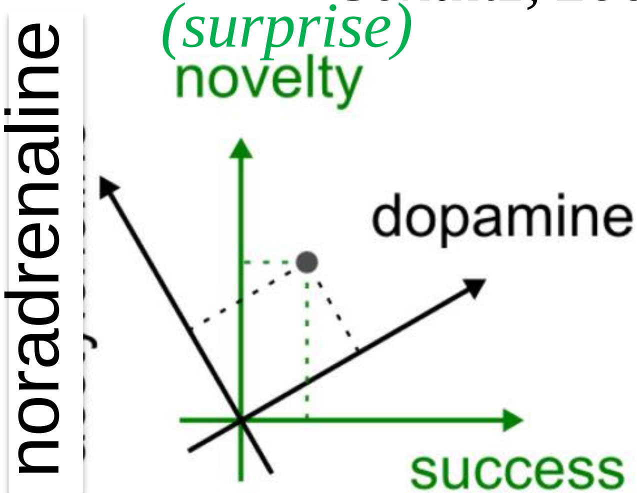
→ *Reinforcement learning: success = reward – (expected reward)*

TD-learning, SARSA, Policy gradient *(book: Sutton and Barto, 1997/2018)*

- **4 or 5 neuromodulators**
- **near-global action**

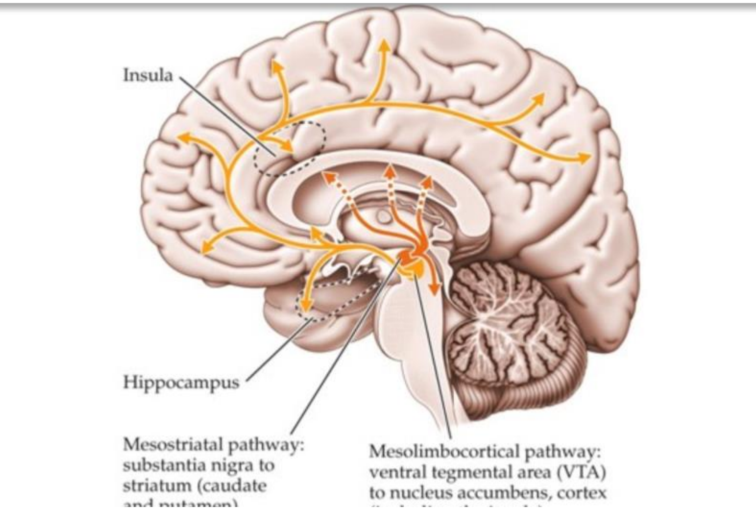
Dopamine/reward/TD:
Schultz et al., 1997,
Schultz, 2002

(surprise)

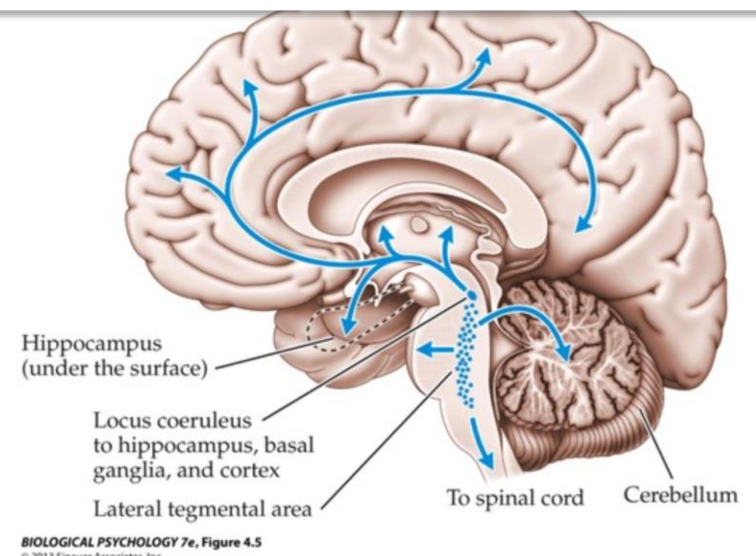


(reward – exp. reward)

Dopamine



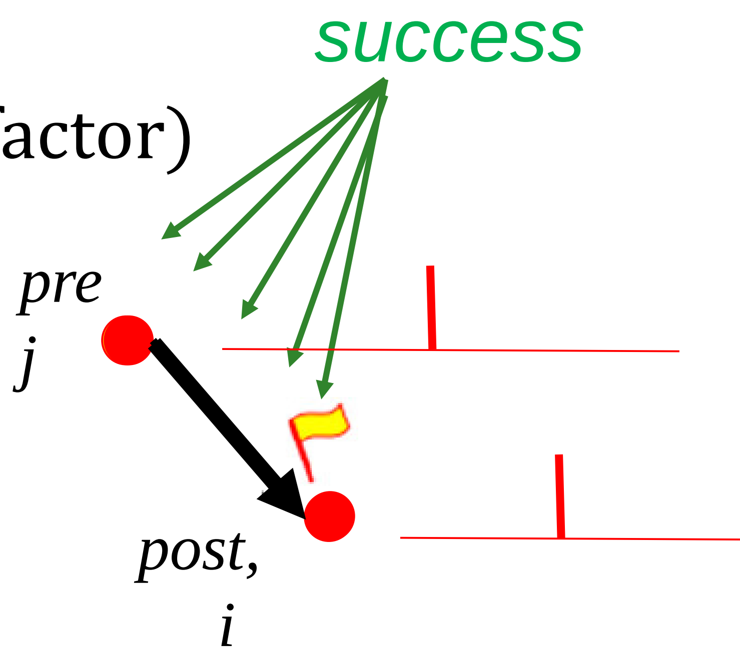
Noradrenaline



Three-factor rules ('neo-Hebbian')

Crow 1968; Barto 1985
Schultz et al. 1997; Waelti et al., 2001;
Reynolds and Wickens 2002;
Lisman et al. 2011

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



3rd factor: neuromodulators

- ***Dopamine***
- *Acetylcholine*
- *Noradrenaline*

Three-factor STDP

(for reinforcement learning)

Success signal:
reward – expected reward

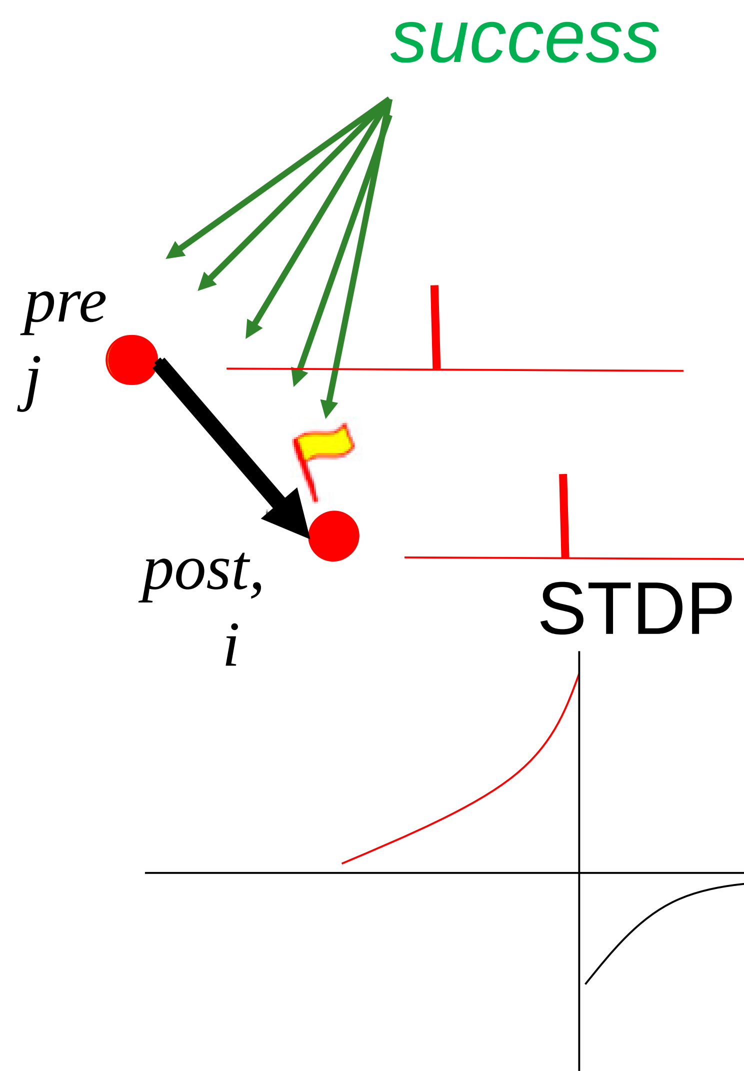
$$\Delta w_{ij} \propto F(pre, post, 3rd\ factor)$$

$$\tau \frac{d}{dt} e_{ij} = STDP_{ij} - e_{ij}$$

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

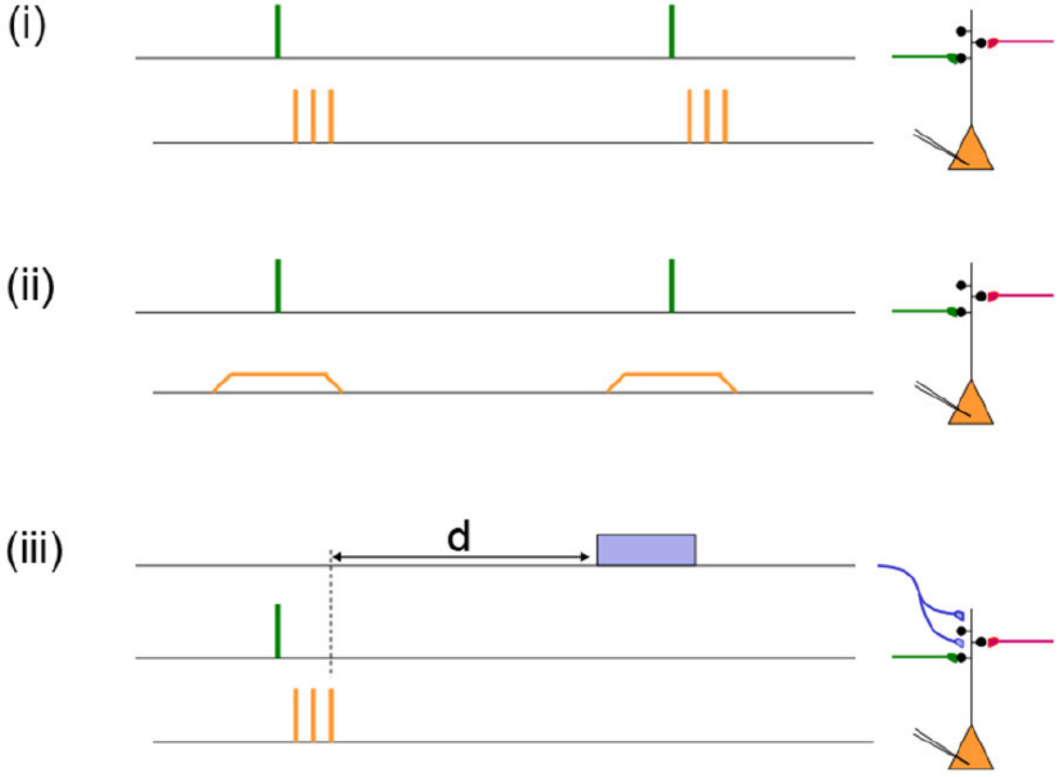
Success signal

Hebb rule/eligibility trace



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

Eligibility traces and 3-factor rules



Hebbian

Hebbian

3-factor

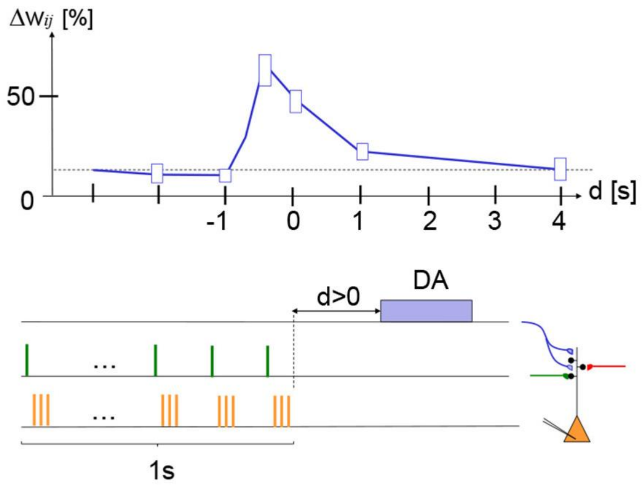
Neuromodulator can come with a delay of 1s - 5s

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

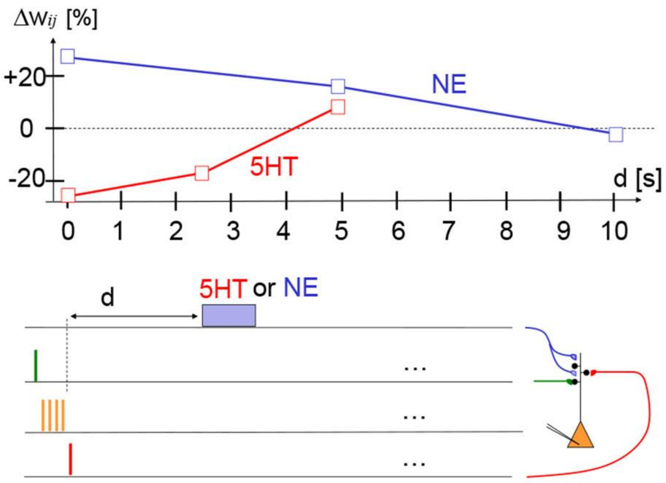
Eligibility traces and 3-factor rules

Neuromodulator can come with a delay of 1s – 5s

Striatum



Cortex



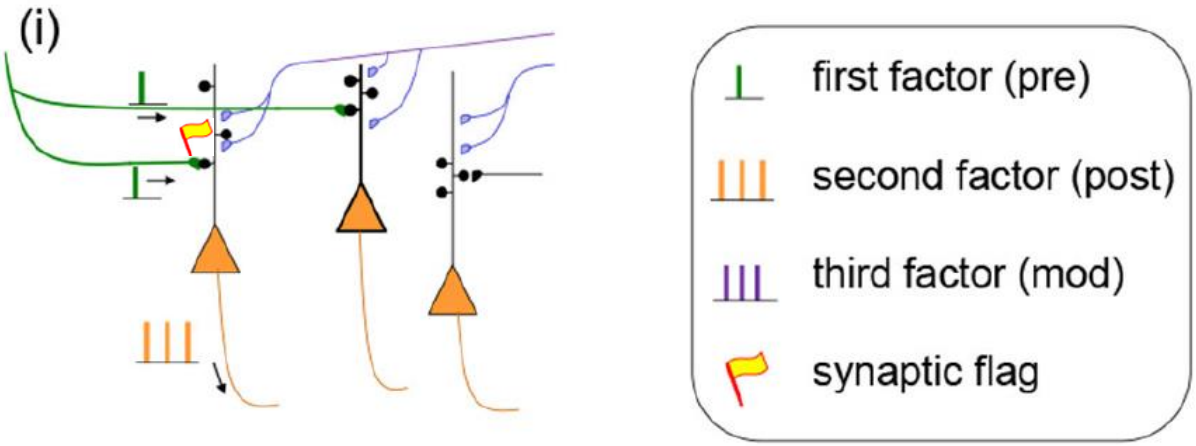
Yagishita et al. 2014 (Kasai lab)

He et al. 2015 (Kirkwood lab)

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

Eligibility traces and 3-factor rules

Selectivity of 3-factor rules

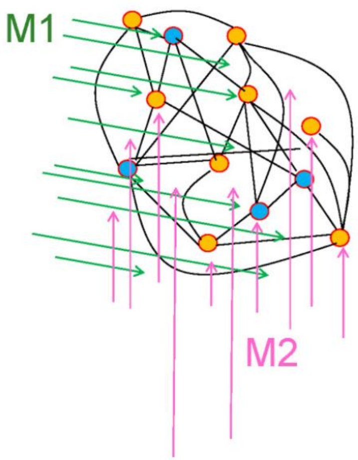


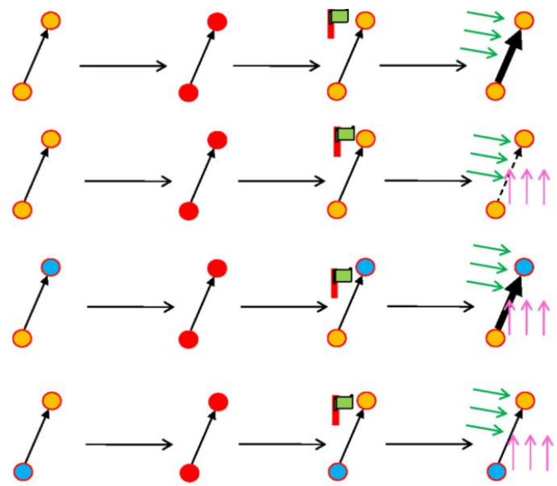
synaptic changes are selective

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

Specificity of 3-factor rules with two neuromodulators

• = *active*





LTP

LTD

LTP

no change

LTP

LTD (heterosyn.)

Two types of neurons (blue, orange)
Two types of neuromodulators (M1, M2)
→ Many combinations!

Summary this part:

- **3-factor learning rules: a framework**
  - two local factors (pre and post)
  - one ‘global’ factor (same for many neurons)

- **Global factor**
  - reward minus expected reward
  - could also be surprise (ongoing work)

- **Generalization to several global factors**
  - neuromodulators dopamine, Ach, Noradrenaline
  - ‘emotional’ brain states modulate learning
    (surprise, reward, exciting, good, progess …)

Eligibility Traces and 3-factor Rules of Synaptic Plasticity

√ - Introduction

√ - Hebbian Learning: a Framework

√ - 3-factor rules: a Framework

- Example: 2 Simulation studies

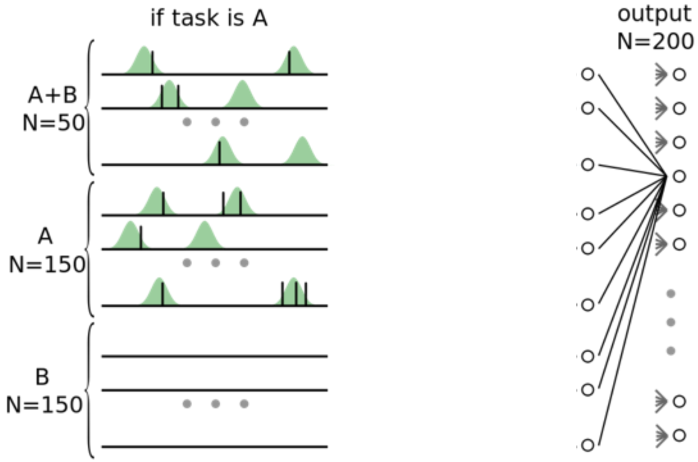
- Example: Behavioral Eligibility Trace

- Summary and Conclusions

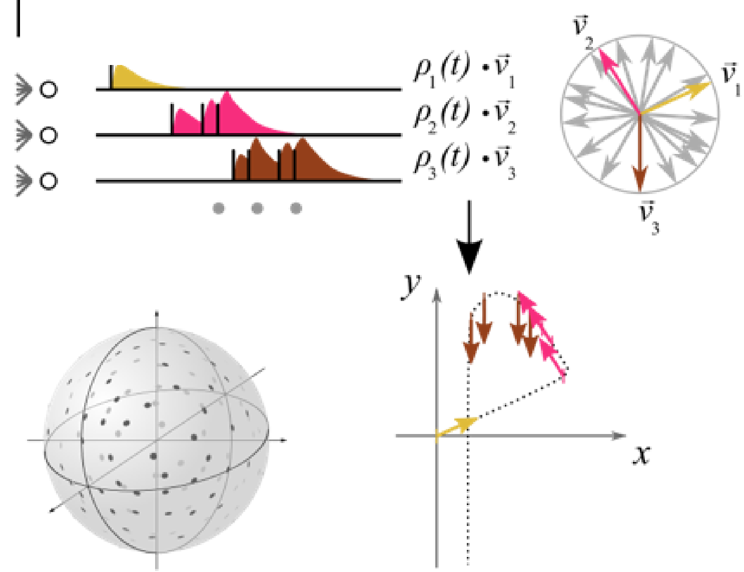
What are the learning rules of the brain?

Example 1:
Table tennis serve

Learning spatial trajectories

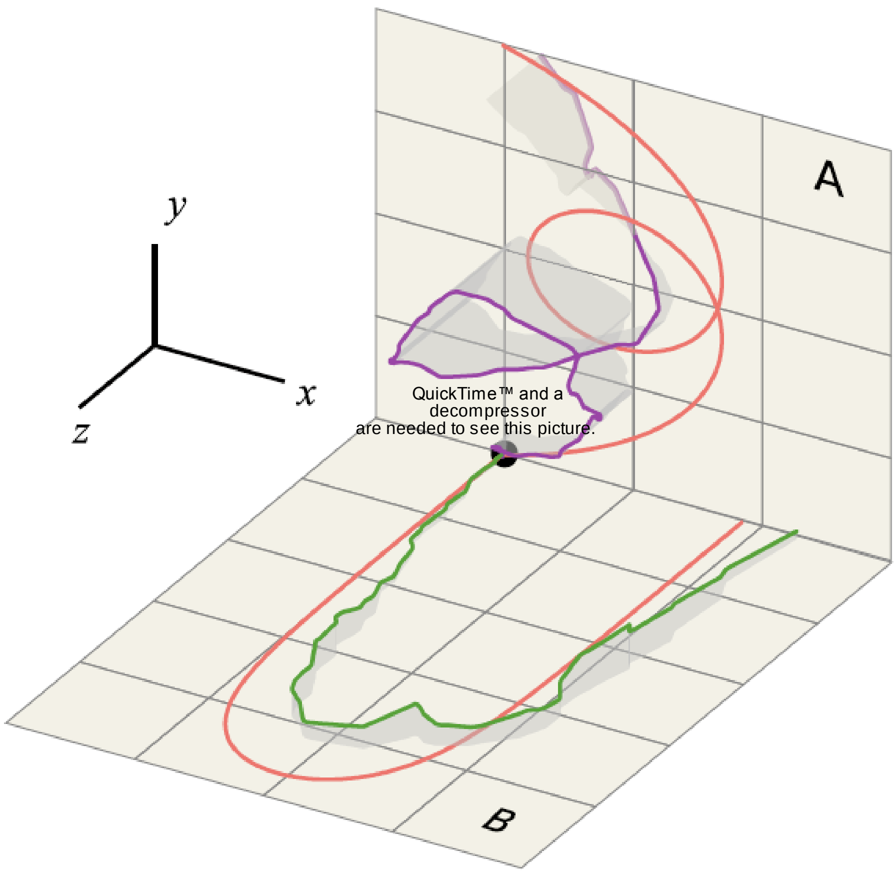


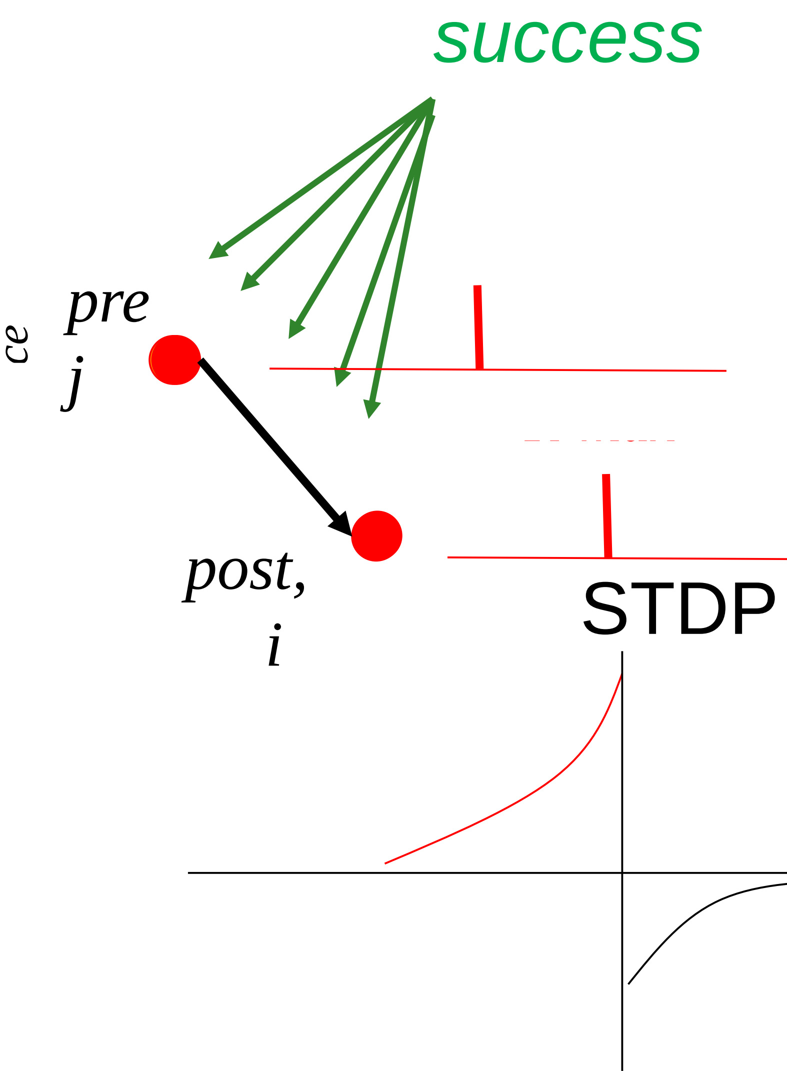
Population vector coding of movements



- *70’000 synapses*
- *1 trial =1 second*
- *Output to trajectories via population vector coding*
- *Single **reward at the END of each trial** based on similarity with a target trajectory*

Learning spatial trajectories





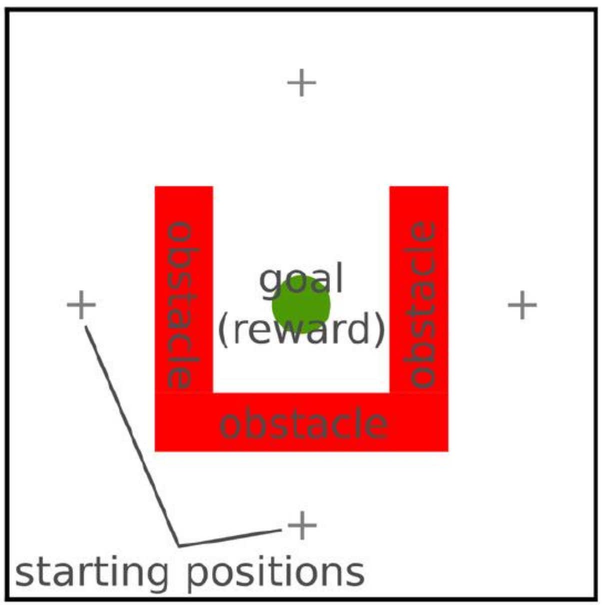
What are the learning rules of the brain?

Example 2:
Maze task

Spiking 3-factor rules

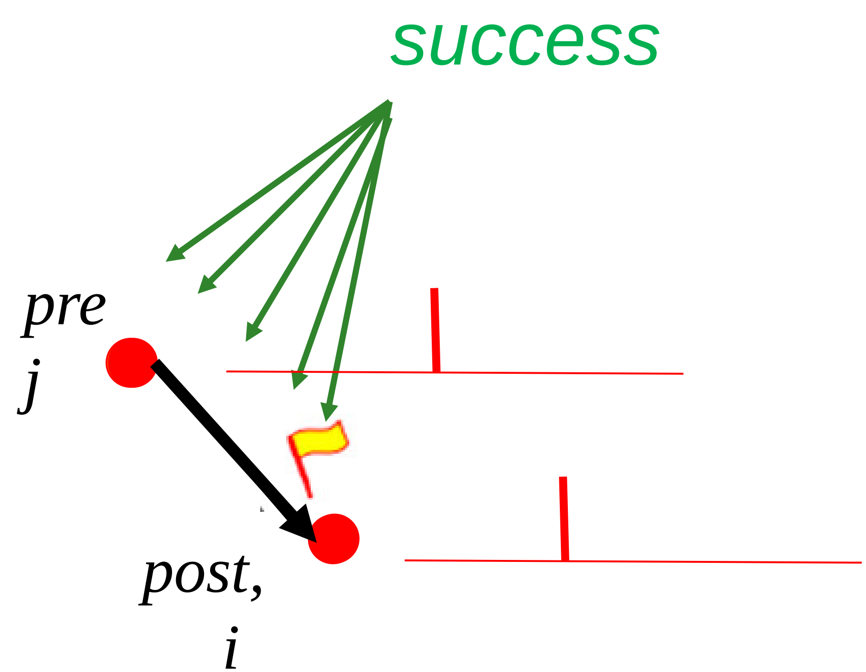
Fremaux et al. 2013

Maze Task



3-factor rule

- STDP sets eligibility trace
- success induces LTP



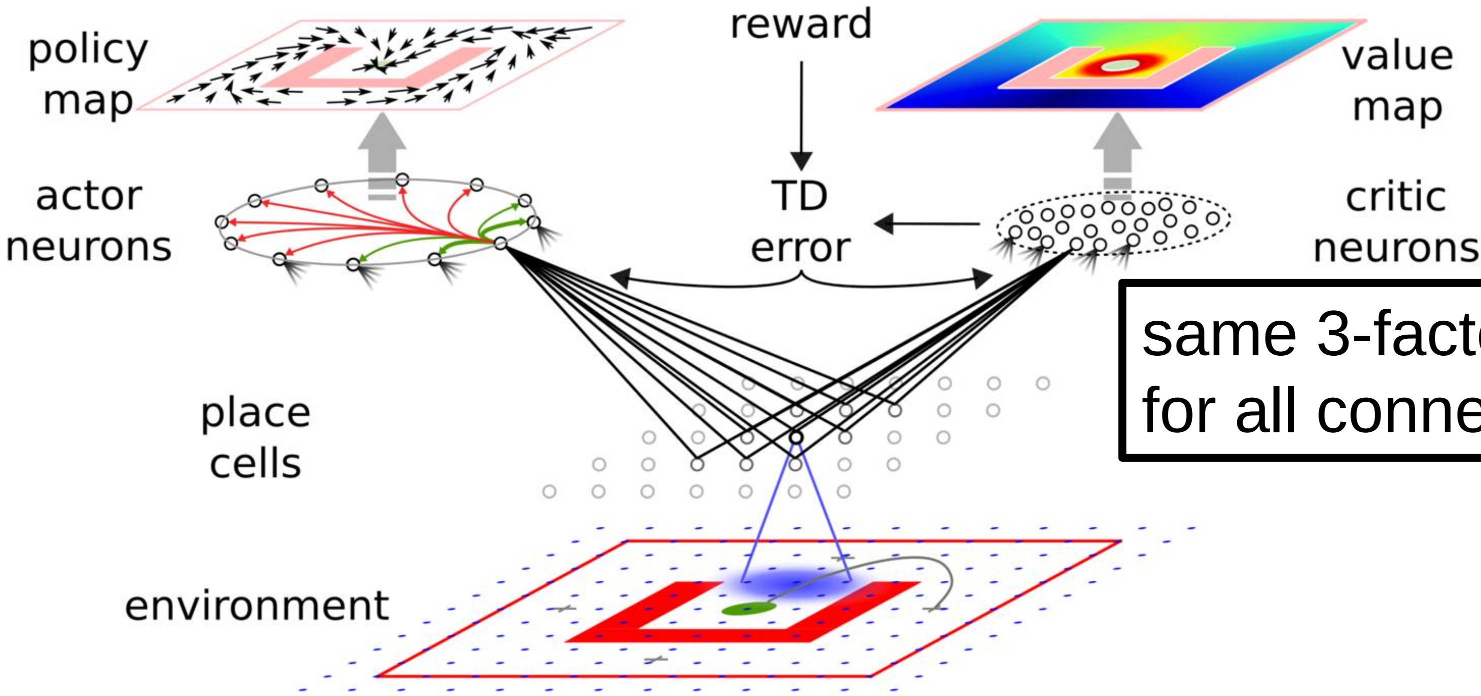
Spiking 3-factor rules

Fremaux et al. 2013

actor-critic-architecture

Continuous action space:
ring of stochastically spiking neurons

Value map:
Independent stoch. spiking n

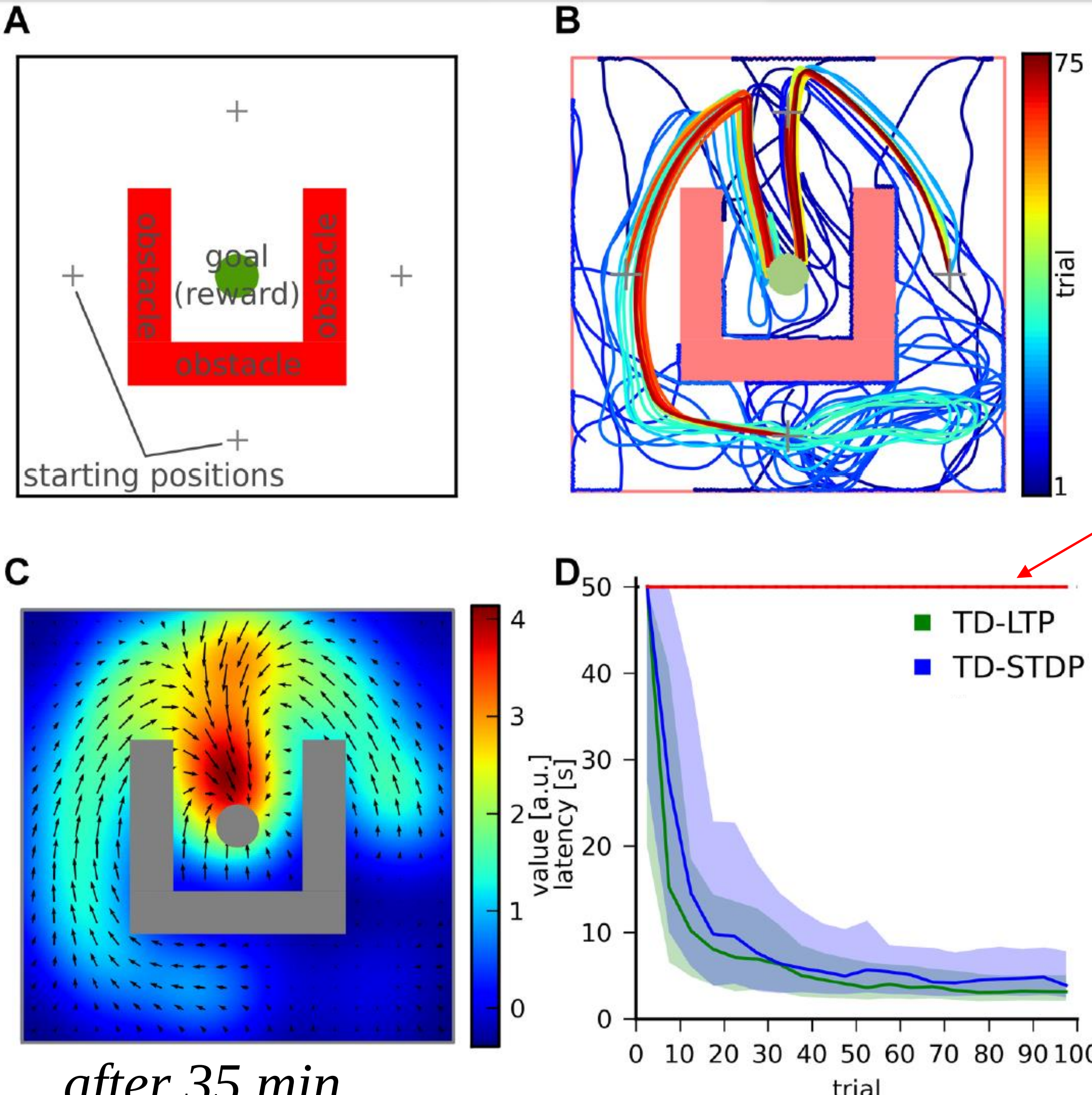


same 3-factor rule
for all connections

Continuous state space:
Represented by spiking place cells

Performance in Maze

Fremaux et al. 2013



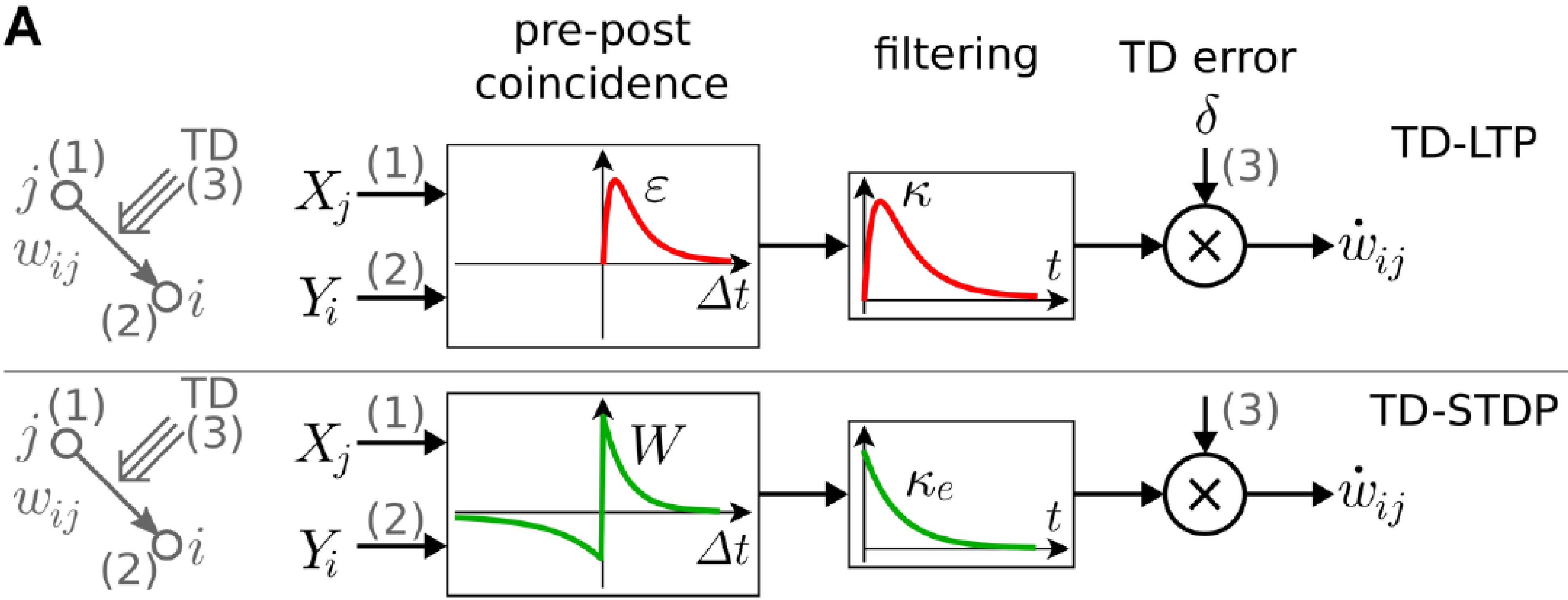
no learning for standard policy gradient

3-factor rule in actor-critic architecture

after 35 min

Spiking 3-factor rules

Fremaux et al. 2013



coactivation
= sets
eligibility
trace

filtering
= decay
of eligibility
trace

Spiking 3-factor rules

Implementation in Biology?

cf. Watabe-Uchida et al. 2017

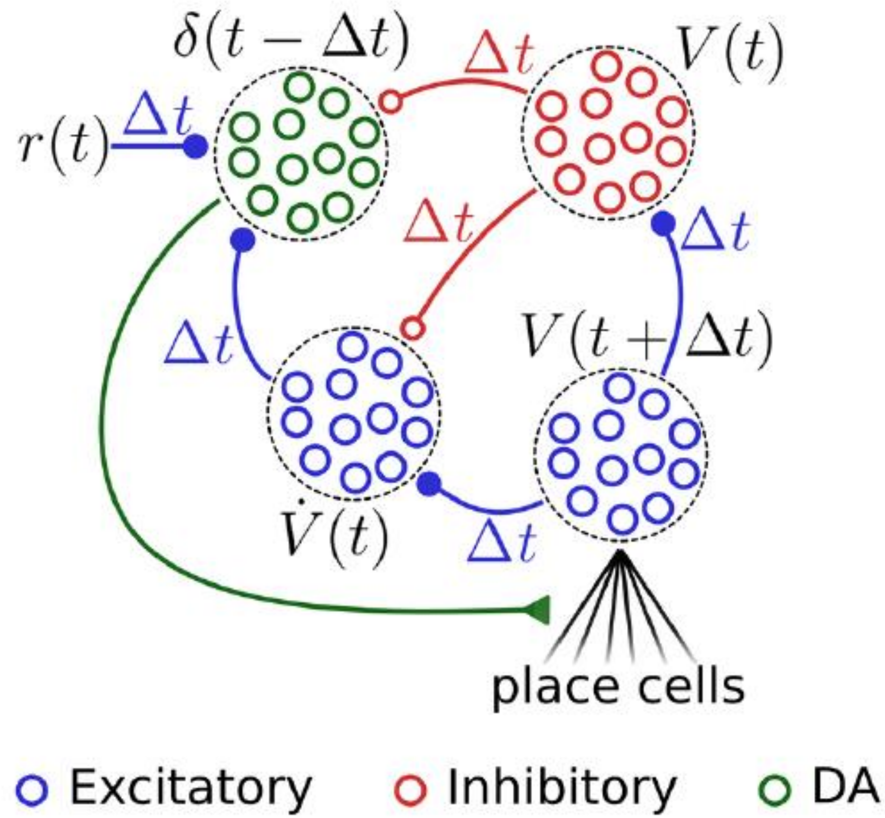
TD error is calculated using

reward

$$\delta = r + \gamma\, V(t+\Delta t) - V(t)$$

value of next position - value of present position

C



Fremaux et al. 2013

Summary this part:

- **3-factor learning rules: a framework**
  - two local factors (pre and post, e.g, STDP)
  - one 'global' factor (same for all neurons)
  - details of STDP rule do NOT matter

- **Global factor is TD error, based on value *V***
  - reward minus expected reward (TD-error)
  - value calculated by critic
  - value estimation builds up in tens of trials
  - same learning rule for critic and actor

- **Time scale of eligibility trace: a few seconds**
  - consistent with experiments (Yagishita)

Eligibility Traces and 3-factor Rules of Synaptic Plasticity

√ - Introduction

√ - Hebbian Learning: a Framework

√ - 3-factor rules: a Framework

√ - Example: Learning in Mazes

→ - Example: Human Eligibility Traces

- Summary and Conclusions

TD-learning

Sutton and Barto 2018

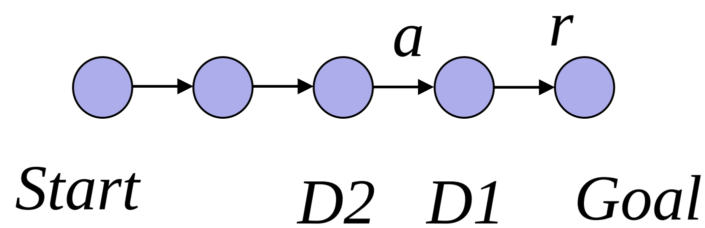
TD error is calculated using reward

$$\delta = r + \gamma\, V(t + \Delta t) - V(t)$$

value of next position - value of present position

$$\delta = r + \gamma\, V(s_{n+1}) - V(s_n)$$

Linear track



Initialize before 1st episode

assume that all

- *V* values are zero *(V=0)*
- *V* values updated with δ

$$V(s_n) \leftarrow V(s_n) + \mu\,\delta$$

→ *after episode 1 only value of D1 is updated, but not value of D2*

Standard TD learning (TD-0) is slow!

Human Eligibility Traces

Psychophysics experiment
to check this

- 10 different states (clip art images)
- 2 buttons cause transitions
  (take left and right knee)

You see the first image in 1s! NOW!!!

Work together with Kerstin Preuschoff and Michael Herzog, and students Marco Lehmann and He Xu

Human Eligibility Traces

+

Human Eligibility Traces

+

Human Eligibility Traces

+

Human Eligibility Traces

+

Human Eligibility Traces

Human Eligibility Traces

+

Human Eligibility Traces

+

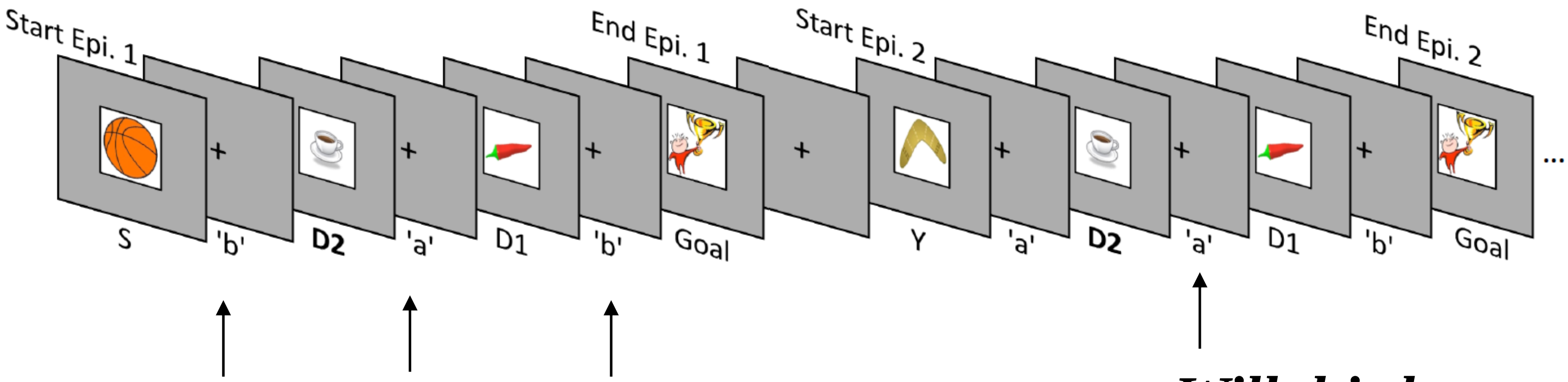
Human Eligibility Traces

Have you seen this image before?

Will you take the same action?

Press the button/knee!

Human Eligibility Traces *Lehmann et al. 2019*



You have chosen action ‘ ’

Will this be your next action?

Human Eligibility Traces *Lehmann et al. 2019*

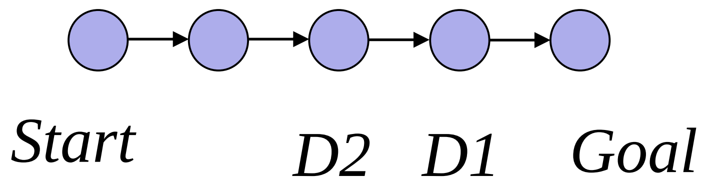
TD error is calculated using

reward

$$\delta = r + \gamma\, V(t+\Delta t) - V(t)$$

value of next position - value of present position

Linear track



Initialize before 1st episode

assume that all

- *V* values are zero *(V=0)*
- *V* values updated with δ

→ *after episode 1 only D1 is updated, but not D2*

Standard TD learning (TD-0) is slow!
Eligibility traces make TD learning fast!

Q-learning

Sutton and Barto, 2018

TD error is calculated using

reward

$$\delta = r + \gamma\, V(t+\Delta t) - V(t)$$

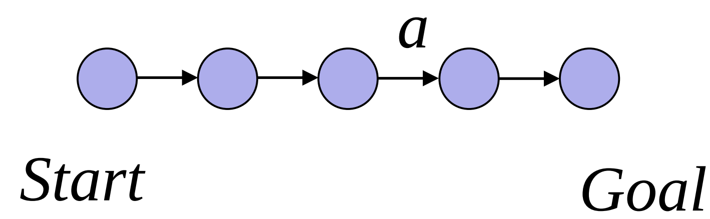
value of next position - value of present position

$$\delta = r + \gamma\, V(s_{n+1}) - Q(s_n, a_n)$$

action value

$$V(s_{n+1}) = \max_a Q(s_{n+1}, a\;\;)$$

Linear track



Standard update rule, TD-0

$$Q(s_n, a_n) \leftarrow Q(s_n, a_n) + \mu\, \delta$$

Update with eligibility trace

$$Q(s, a\;) \leftarrow Q(s, a\;) + \mu\, \delta\, e(s, a)$$

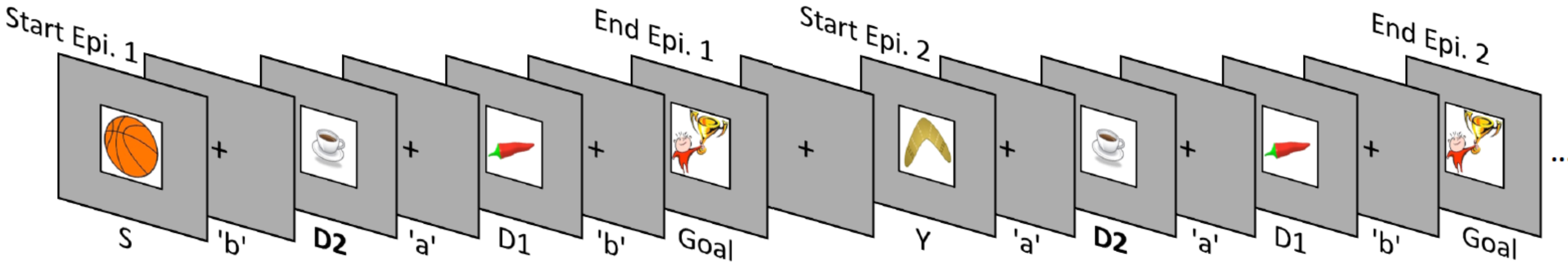
$$e(s, a) = 1 \;\boldsymbol{if}\; (s, a) = (s_n, a_n)$$

else

exponential decay

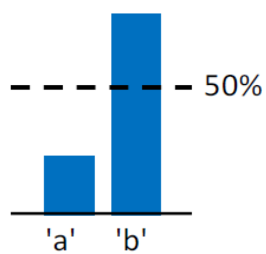
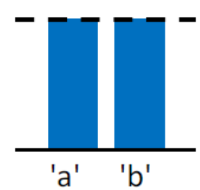
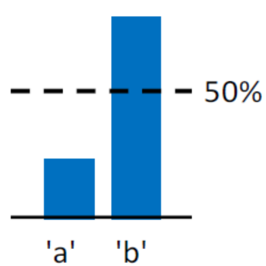
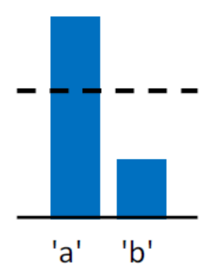
Human Eligibility Traces

Lehmann et al. 2019



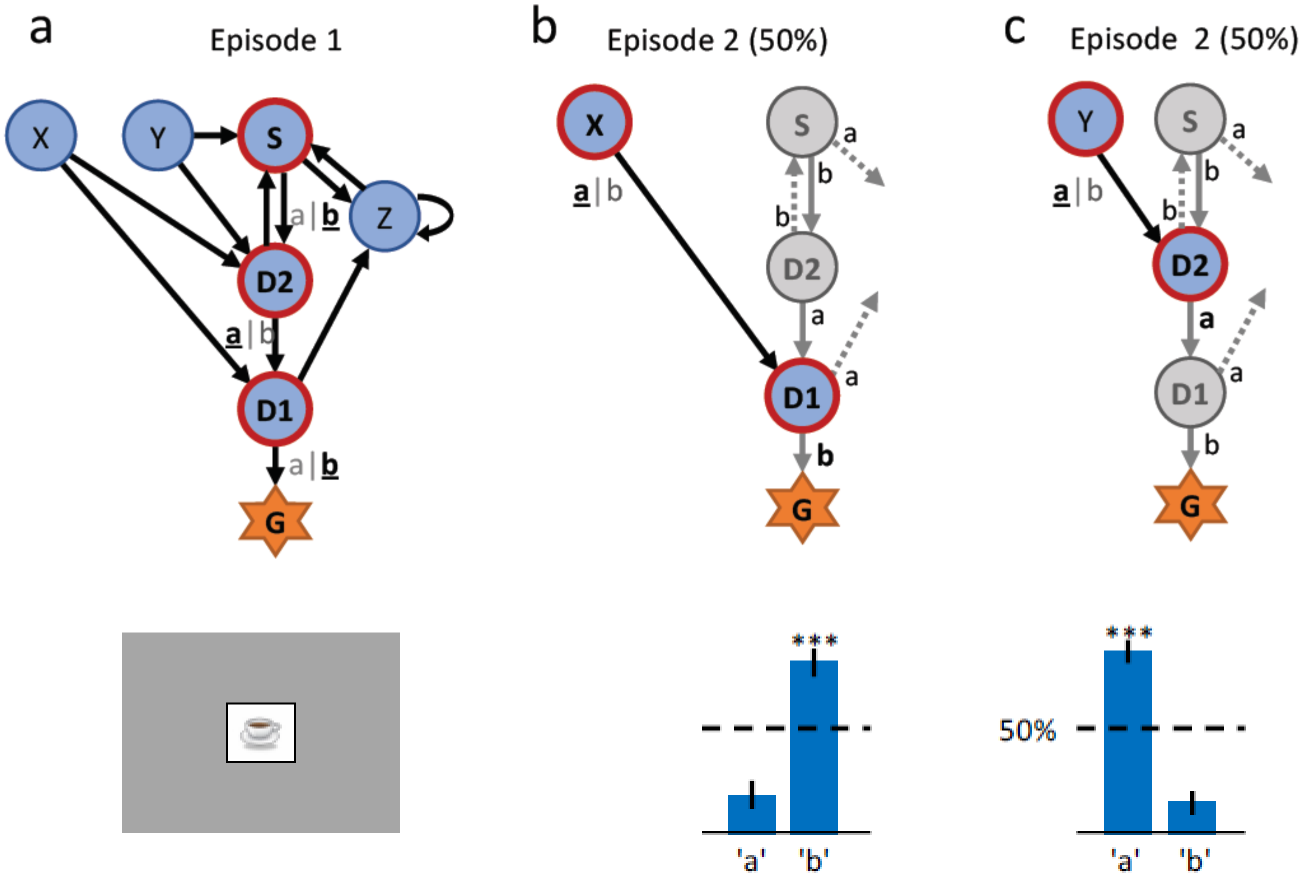
In state D2: action 'a '

Will a be your next action?

PREDICTION	Behavior, Episode 2	
	State D1	State D2
Hypothesis: No eligibility trace		
Hypothesis: With eligibility trace		

Behavioral Eligibility Traces

Lehmann et al. 2019



action in state *D1*

action in state *D2*

Human Eligibility Traces *Lehmann et al. 2019*

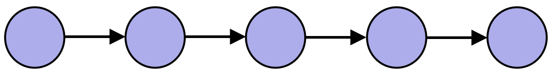
We find **1-shot learning**:
learned action bias after a single episode,
even in state D2 (two actions away from goal)

1-shot learning is compatible with eligibility traces
but not with TD-0, or Q-0, or SARSA-0

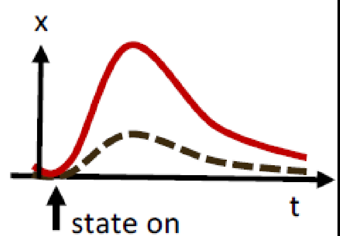
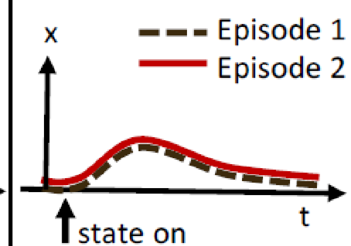
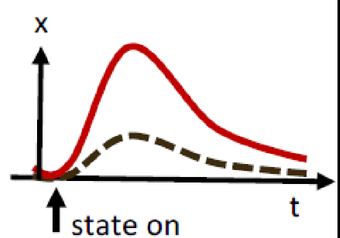
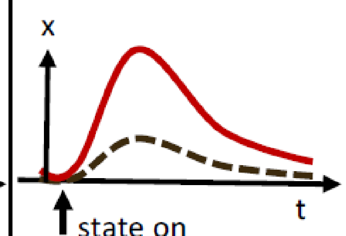
Fit model to behavioral data:
eligibility trace has a time scale of about 10s

Physiological Eligibility Traces

Linear track



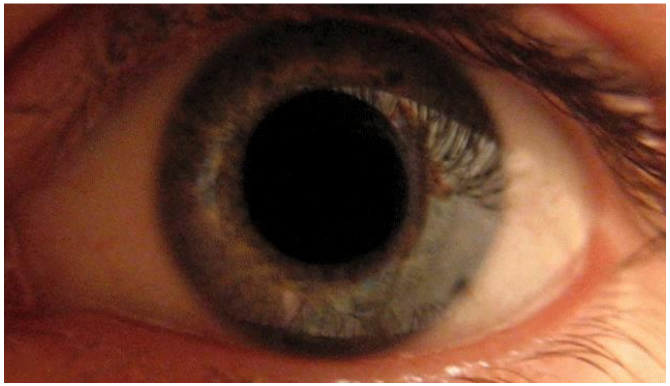
Start *D2* *D1* *Goal*

PREDICTION	Physiology, Episode 1 vs. Episode 2	
	State D1	State D2
Hypothesis: No eligibility trace		
Hypothesis: With eligibility trace		

Prediction valid for any signal that reflects
- value V
Or
- reward prediction error δ

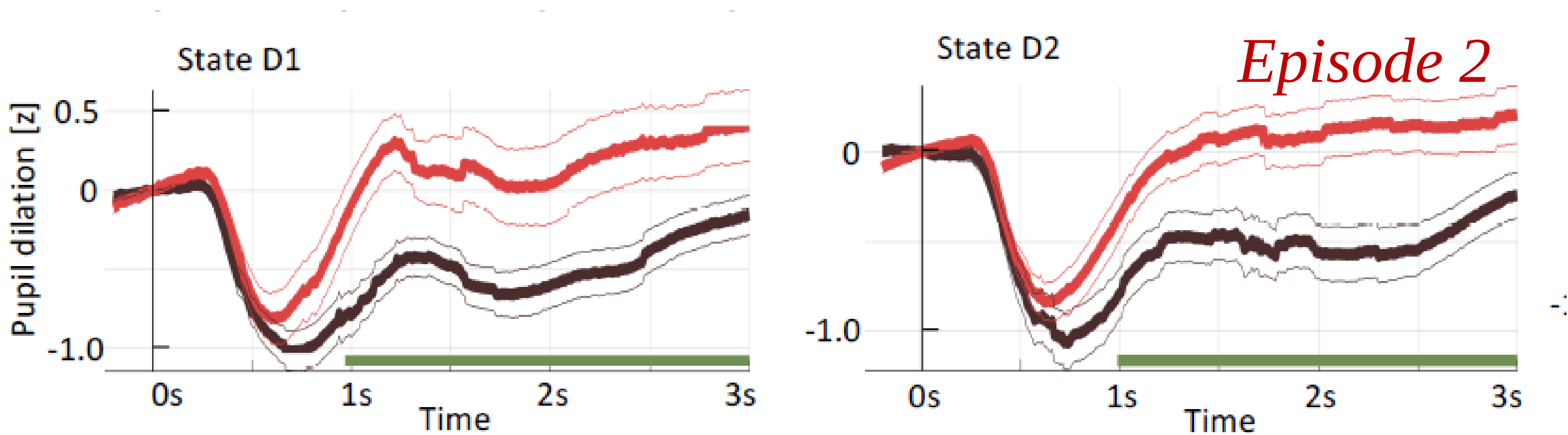
Lehmann et al. 2019

Human Eligibility Traces *Lehmann et al. 2019*



Pupil diameter is a measure for
- Light intensity
- Memory load
- Engagement
- Surprise

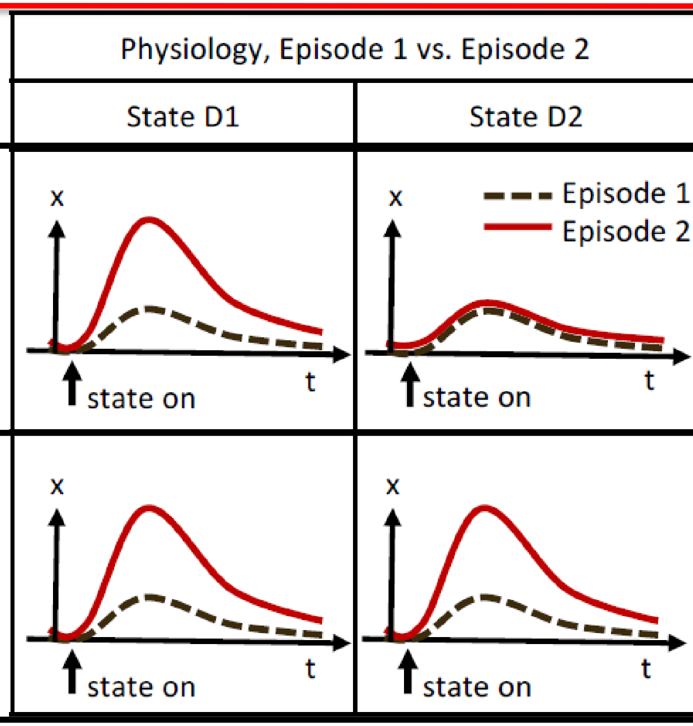
→ **Learning-related signal**



Physiological Eligibility Traces

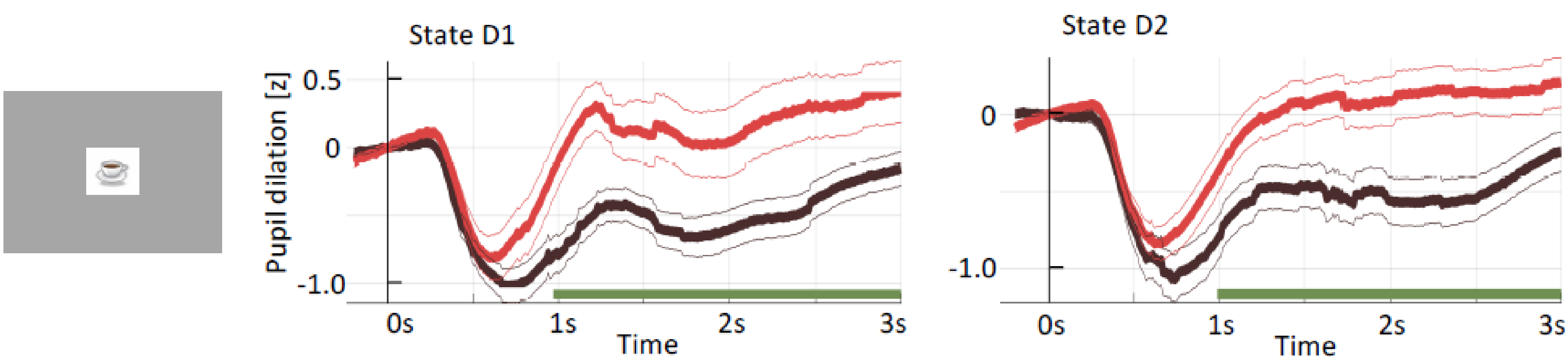
Hypothesis:
No eligibility trace

Hypothesis:
With eligibility trace



Prediction valid for any signal that reflects
- value V
Or
- reward prediction error δ

Lehmann et al. 2019



Summary this part

- **Decay Time of eligibility trace: 10 seconds**
  - derived from human behavioral experiment
  - a bit longer than slice experiments (Yagishita)

- **Reinforcement learning models with eligibility trace make qualitatively different predictions than those without**
  - experimental data in favor of eligibility traces

- **A reward a few seconds later influences** state-action associations
  - consistent with 3-factor rule framework

Eligibility Traces and 3-factor Rules of Synaptic Plasticity

√ - Introduction

√ - Hebbian Learning: a Framework

√ - 3-factor rules: a Framework

√ - Example: Learning in Mazes

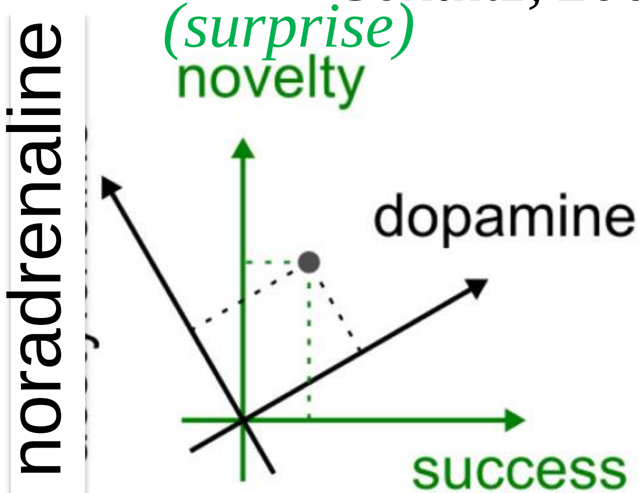
√ - Example: Behavioral Eligibility Trace

→ - Summary and Conclusions

- **4 or 5 neuromodulators**
- **near-global action**

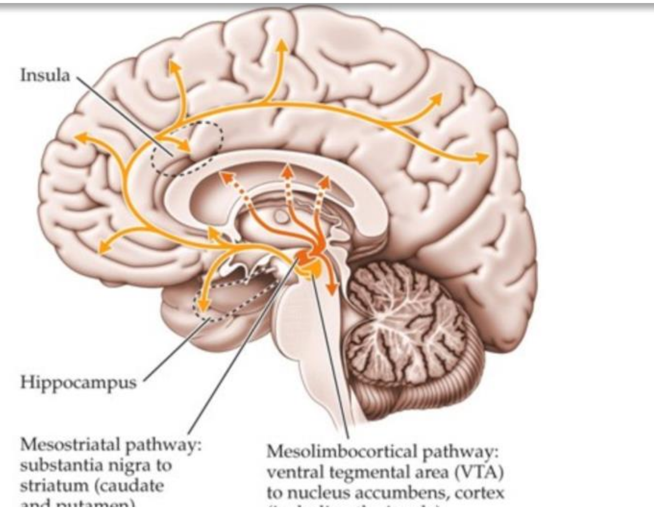
Dopamine/reward/TD:
Schultz et al., 1997,
Schultz, 2002

(surprise)

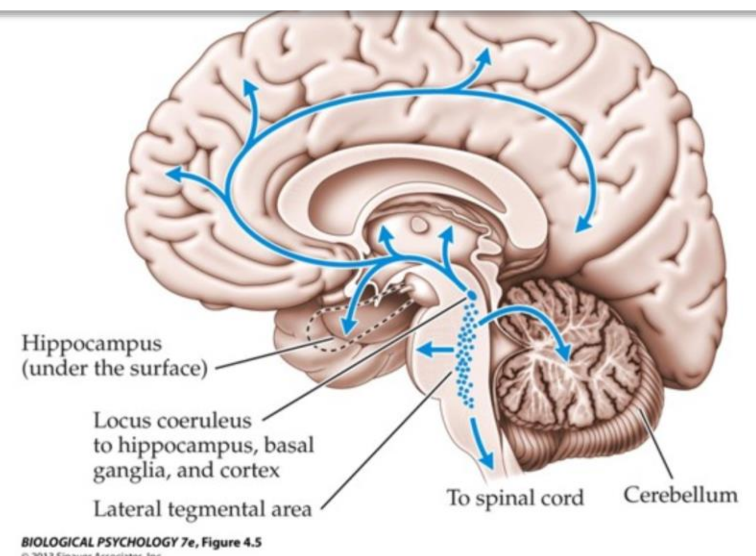


(reward – exp. reward)

Dopamine



Noradrenaline



Enjoy the images!

Why are we surprised?

- We always make models
- We know that the models are not perfect
- Surprise enables us to adapt the models

→ Hypothesis:

Surprise boosts learning (3rd factor)

NOTE: no reward!!!!

Conclusions:
Eligibility Traces and 3-factor Rules

1. Synaptic Plasticity is more than Hebb:
→ Hebb + 3rd factor

2. The 3rd factor can be reward or surprise or …
→ Neuromodulator(s), emotional states

3. Time scale of eligibility traces can be measured
→ 10s in human behavior
→ 1s – 5s in slices (striatum, cortex, hippocampus?)

4. Reinforcement learning can be fast
→ a few trials

Historical Remark: Interactions Theory Experiment

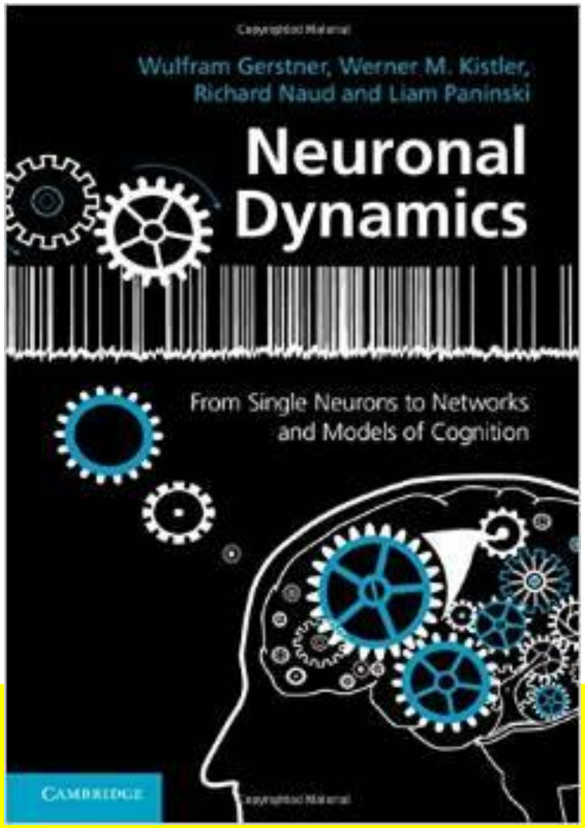
3-factor rule: a conceptual model

Crow 1968 - words
Klopf 1972; Barto et al. 1983, Barto 1985 – neuronal model
Watkins 1989, Dayan 1991 – abstract mathematical TD model
Wiliams 1992 – abstract mathematical model (policy gradient)
Forster and Dayan 2000, Arleo and Gerstner - hippocampal model
Xie and Seung 2004, Izhikevich 2007, Legenstein et al. 2009,
Vasilaki et al, 2009, ... : continuous time spiking three-factor rules

Prediction versus Assumption

- 3-factor rule
- Time scale of eligibility traces

Thanks to
Claudia Clopath (now Imperial College London)
Eleni Vasilaki (now Univ. of Sheffield)
Tim Vogels (now Univ. of Oxford)
J-P. Pfister (now Univ. of Bern)
Friedemann Zenke (now FMI, Basel)
Nicolas Fremaux (start-up)
Henning Sprekeler (now TU Berlin)



The End

Textbook: ***Neuronal Dynamics*** (Cambridge)
with W.M. Kistler, R. Naud, L. Paninski