REINFORCEMENT LEARNING

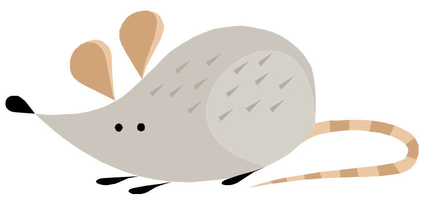
Eleni Vasilaki, University of Sheffield, UK

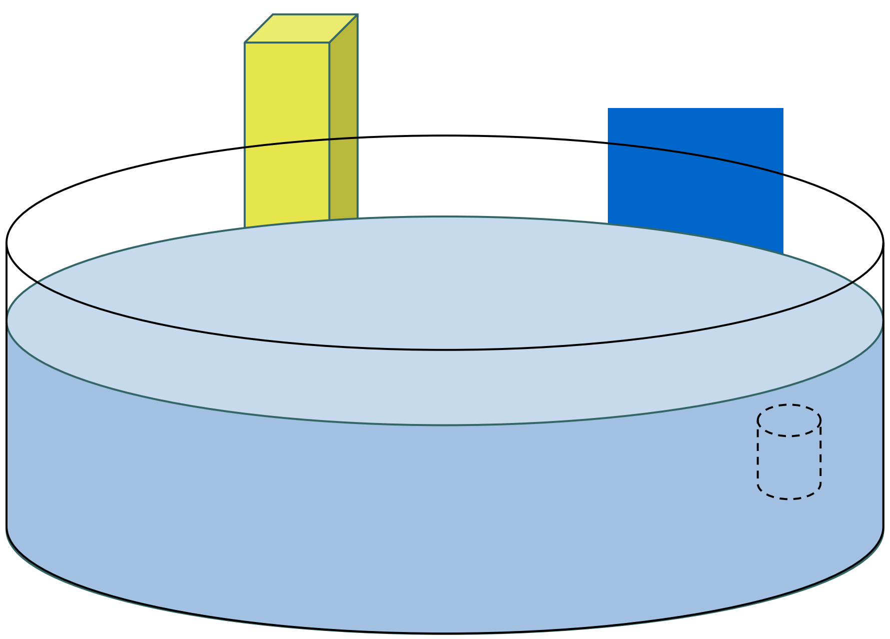
EPSRC
Engineering and Physical Sciences Research Council

Google DeepMind

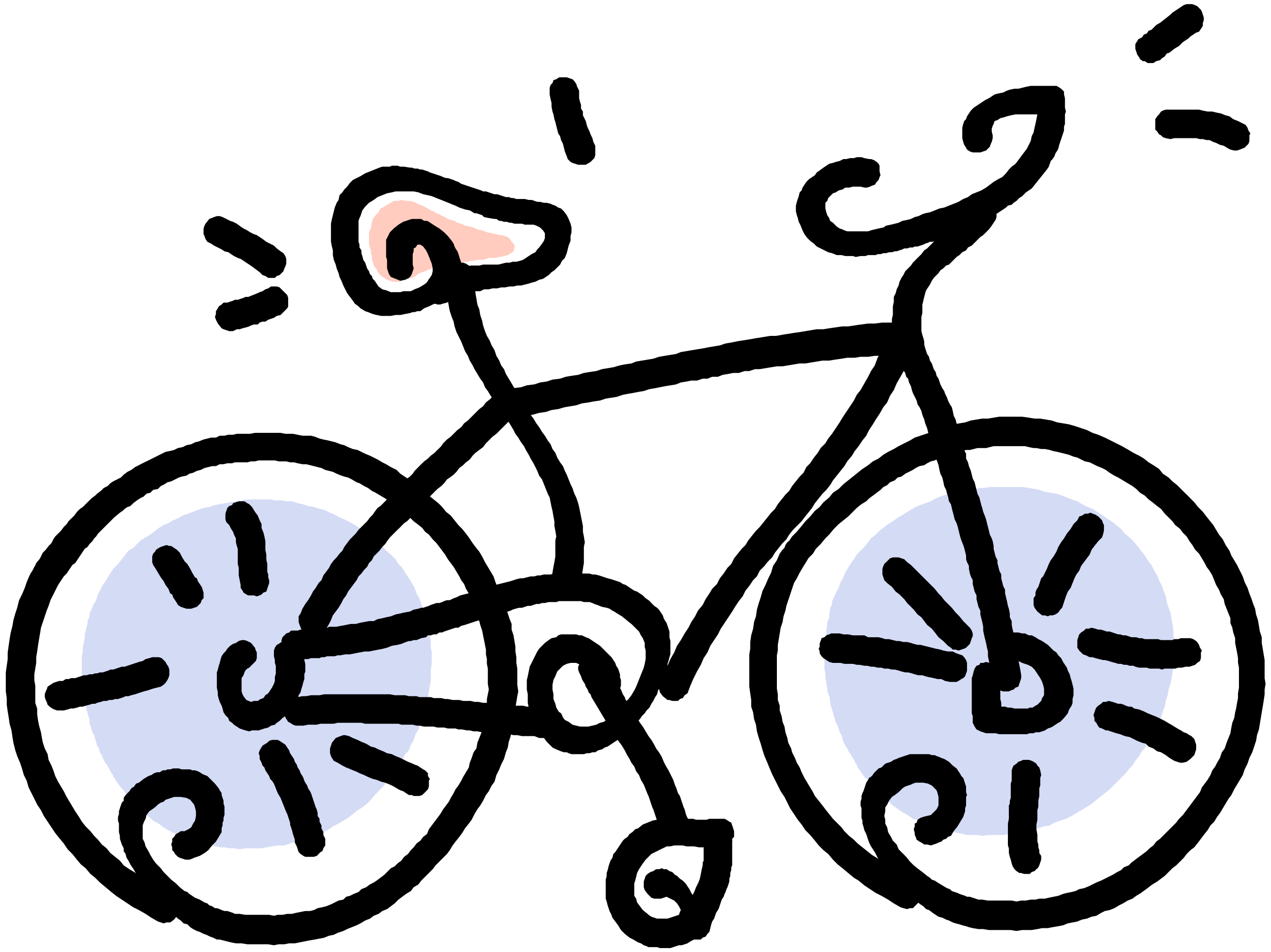
REINFORCEMENT LEARNING EXAMPLES

- Reward/Punishment
- Exploration/Exploitation



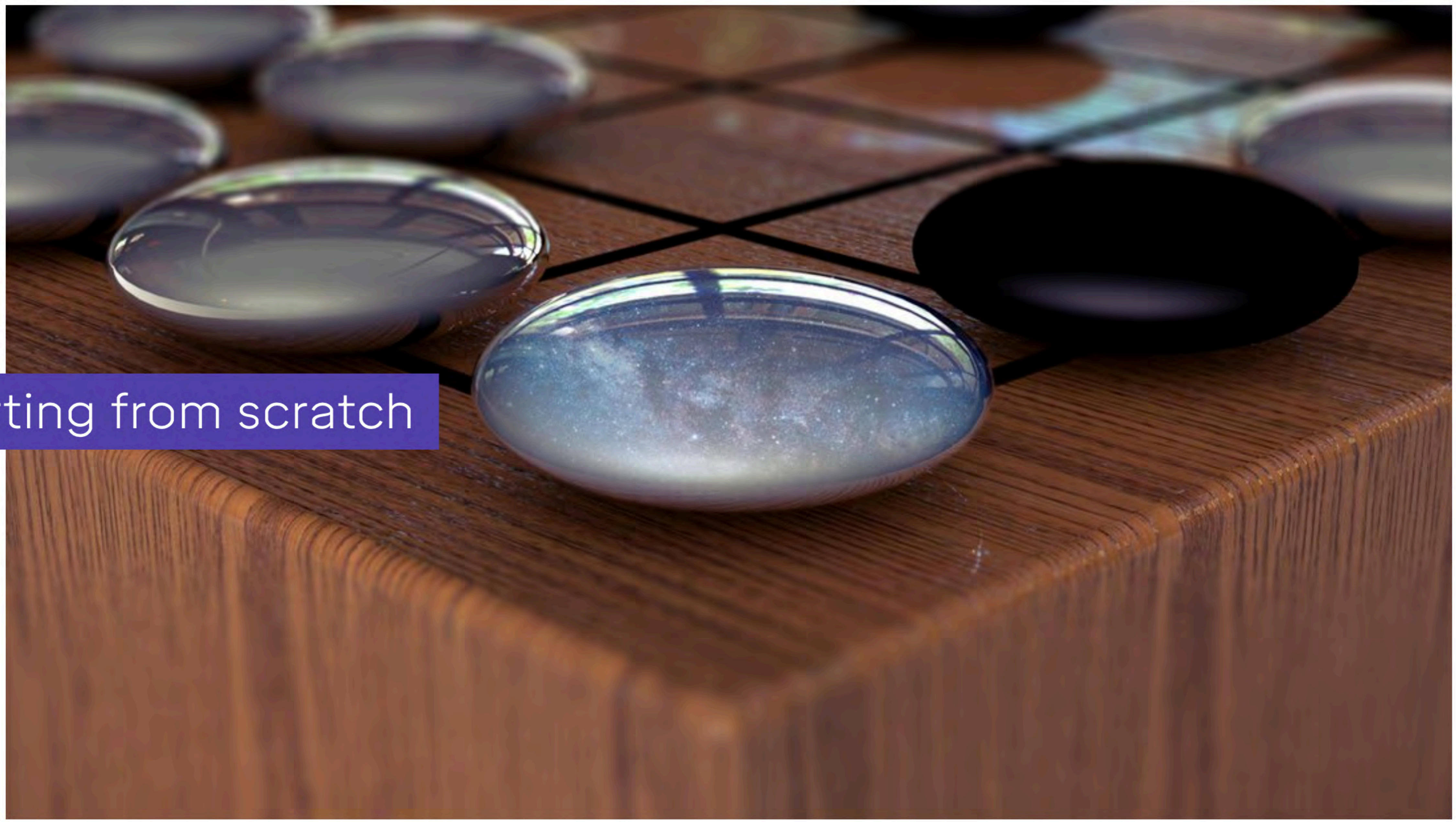


Morris, 1981



Sutton & Barto, 2018 http://incompleteideas.net/book/the-book.html

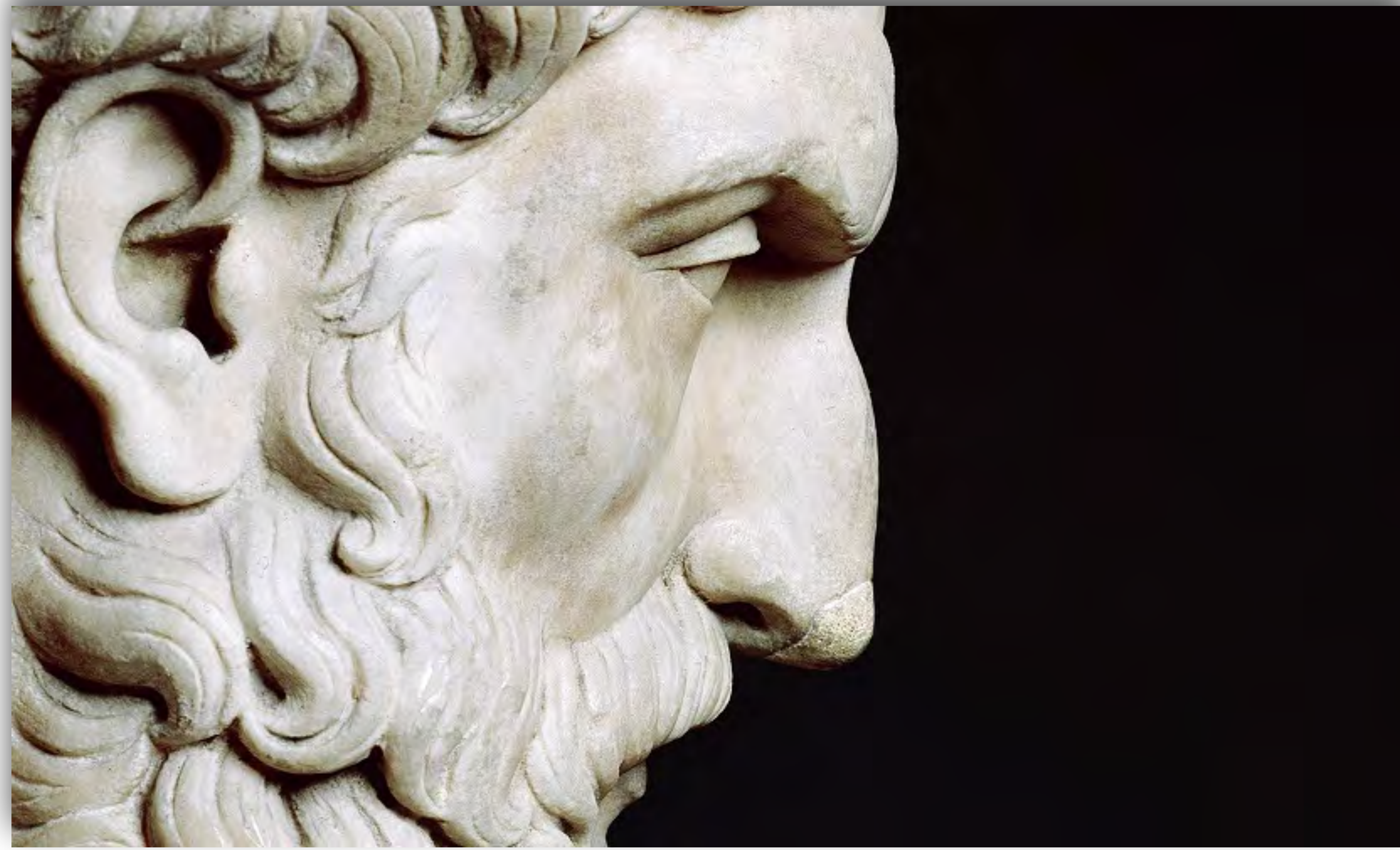
GOOGLE DEEPMIND ALPHAGO ZERO



https://deepmind.com/blog/alphago-zero-learning-scratch/

Silver et al, 2017 https://www.nature.com/articles/nature24270

EPICURUS & REINFORCEMENT LEARNING



http://www.defenseofreason.com

Vasilaki, 2017 https://arxiv.org/abs/1710.04582

EPICURUS & REINFORCEMENT LEARNING

“We say that pleasure is the beginning and the end of a happy life”

“For continual drinking and partying [..] do not produce a pleasant life, but sober reasoning which both examines the basis for every choice and avoidance […]”

— Epicurus’ Letter to Menoeceus
Diogenes Laertius, Lives of Eminent Philosophers

REWARD FUNCTION

$$R = r_t + r_{t+1} + r_{t+2} + \ldots + r_{t+N}$$

- Reward can be positive and negative.
- An action can bring a small immediate reward and a large future punishment.

Vasilaki, 2018 http://bit.ly/RL-happiness

Sutton & Barto, 2018 http://incompleteideas.net/book/the-book.html

DISCOUNT FACTOR

$$R_t = r_t + r_{t+1} + r_{t+2} + \ldots + r_{t+N}$$

$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \ldots$$

$$0 \leq \gamma < 1$$

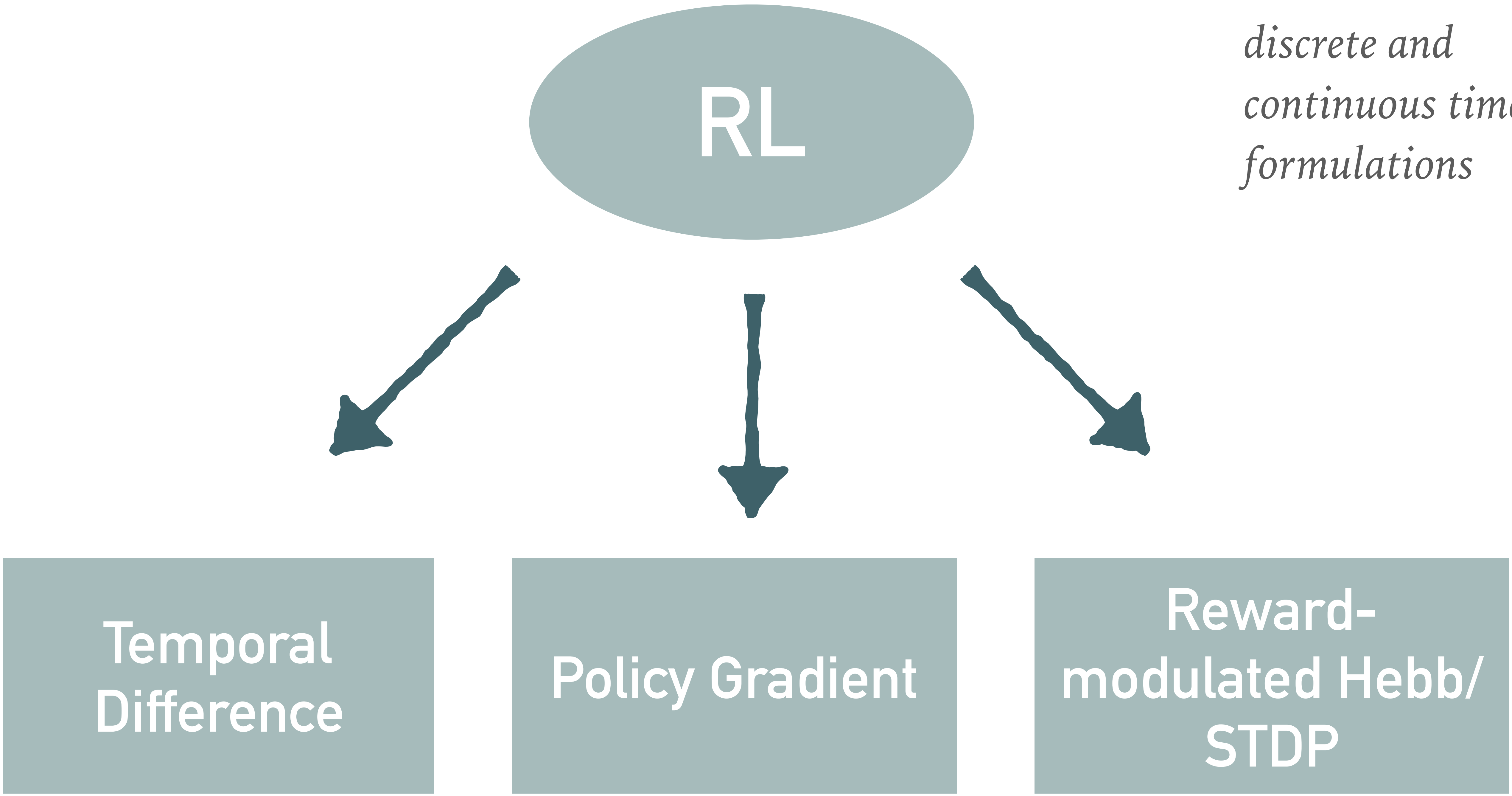
This is why I am impatient.

Vasilaki, 2018 http://bit.ly/RL-happiness

Sutton & Barto, 2018 http://incompleteideas.net/book/the-book.html

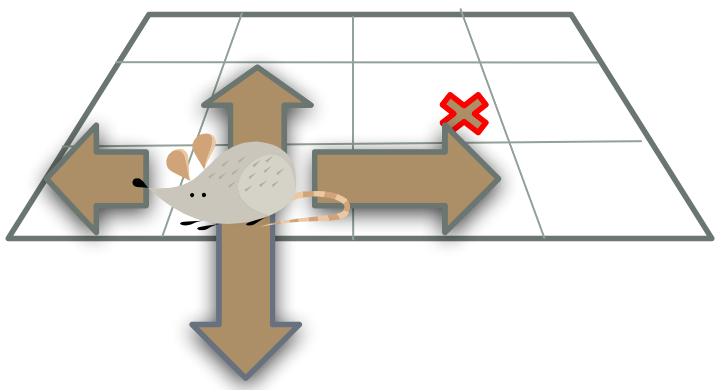
REINFORCEMENT LEARNING CATEGORIES

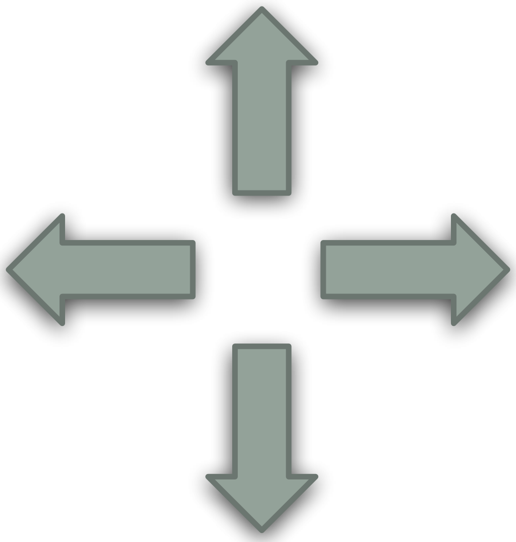
discrete and continuous time formulations



TEMPORAL DIFFERENCE LEARNING

States x Actions: 3x4x4=48





Maximize expected return

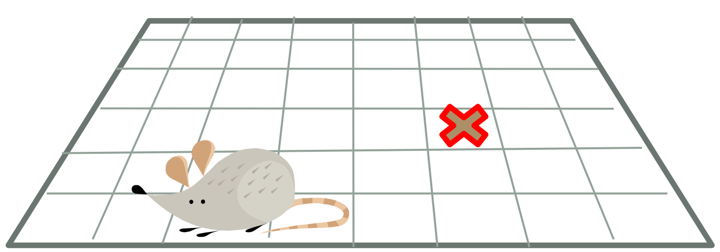
Q (state, action)

We do not know the Q values

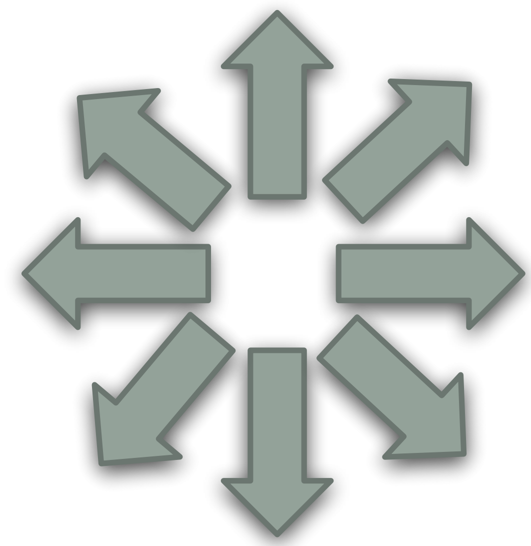
Sutton & Barto, 2018

TEMPORAL DIFFERENCE LEARNING

States x Actions: 6x8x8=384



Q (state, action)



Learning process slows down

Sutton & Barto, 2018

POLICY (π)

- How do I chose an action?
  - I know all the Q values
    - Exploit (max Q value, a.k.a Greedy policy)
  - I only have an estimate, and this may be wrong!
    - Explore (e.g. ε-Greedy, soft-max)

Sutton & Barto, 2018

TEMPORAL DIFFERENCE LEARNING

$$R = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \dots$$

$$R = r_t + \gamma(r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots)$$

$$Q^\pi(s, a) = E_\pi\{R_t \,|\, s_t = s, a_t = a\}$$

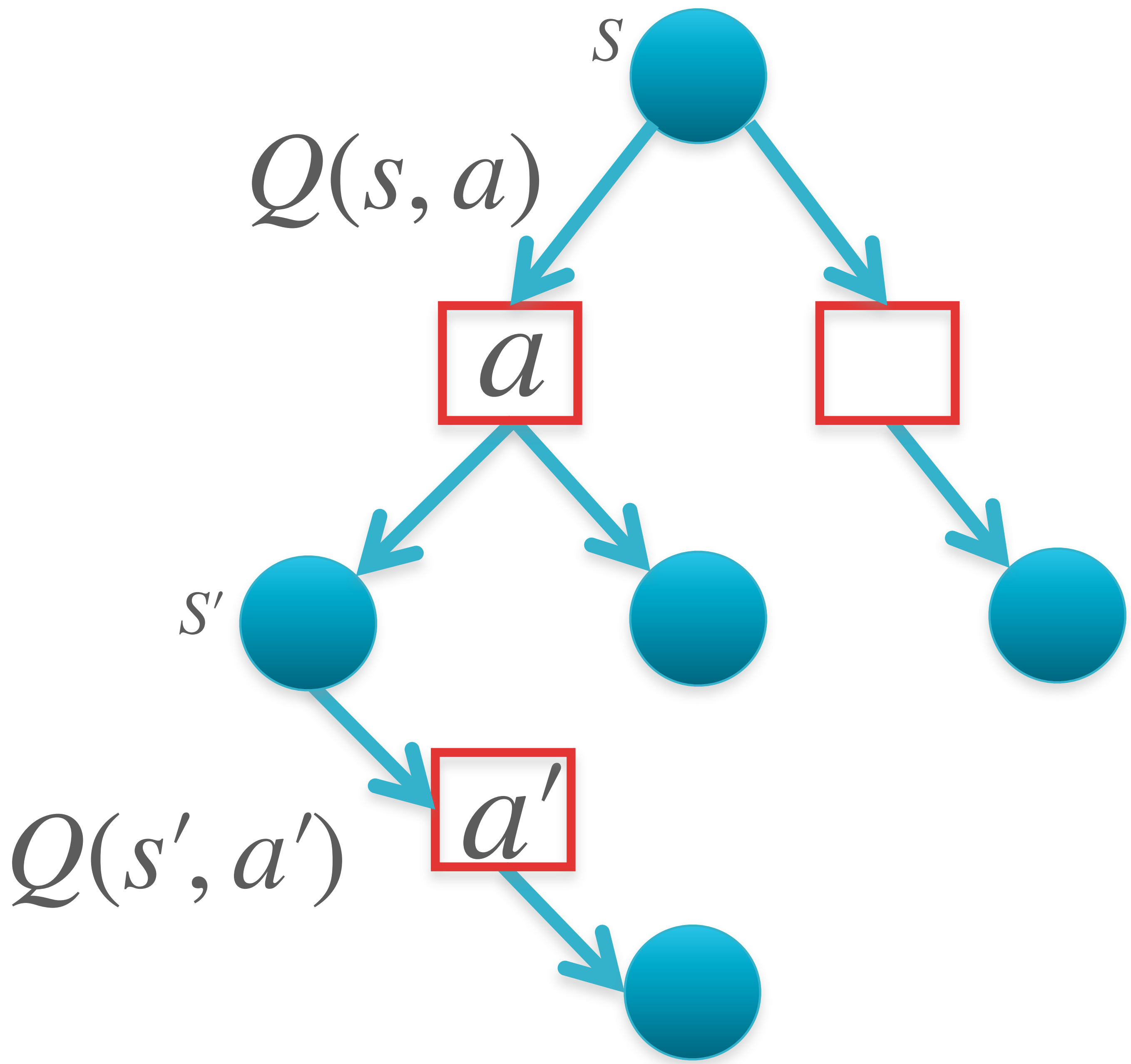
Bellman Equations,

Markovian property

Sutton & Barto, 2018

SARSA

$$\Delta Q(s,a) = \eta\Big((r + \gamma Q(s',a')) - Q(s,a)\Big)$$



Sutton & Barto, 2018

SARSA AND HAPPINESS

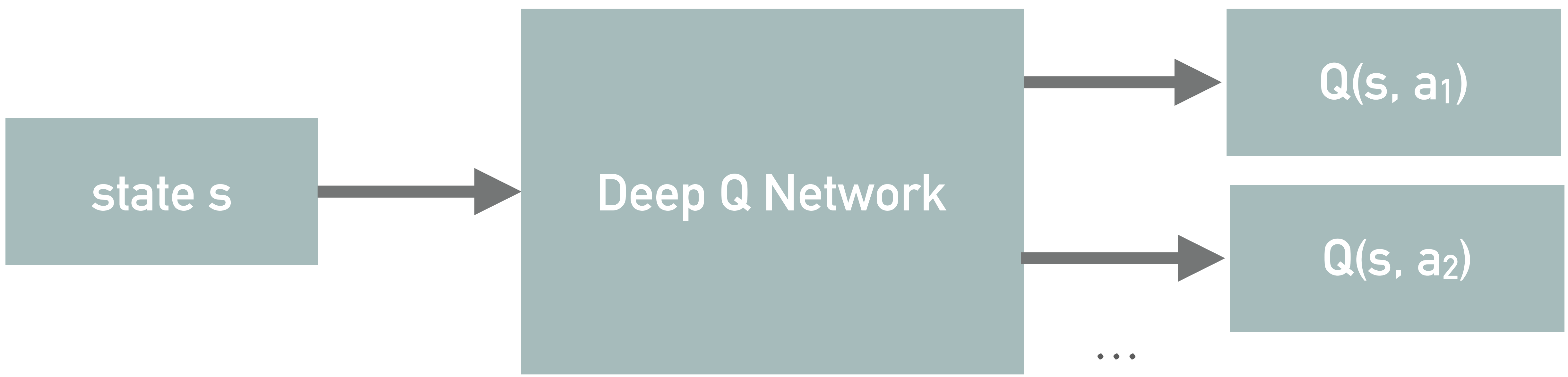
$$\Delta Q(s,a) = \eta \left(\underbrace{(r + \gamma Q(s',a'))}_{\text{What I "actually" get}} - \underbrace{Q(s,a)}_{\text{Anticipated reward}} \right)$$

A positive reward may feel like punishment

A negative reward may feel like reward

Vasilaki, 2018 http://bit.ly/RL-happiness

DEEP REINFORCEMENT LEARNING



$$L(\theta_i) = \Big(\underbrace{(r + \gamma Q(s', a', \theta_{i-1}))}_{\textit{Target}} - Q(s, a, \theta_i)\Big)^2$$

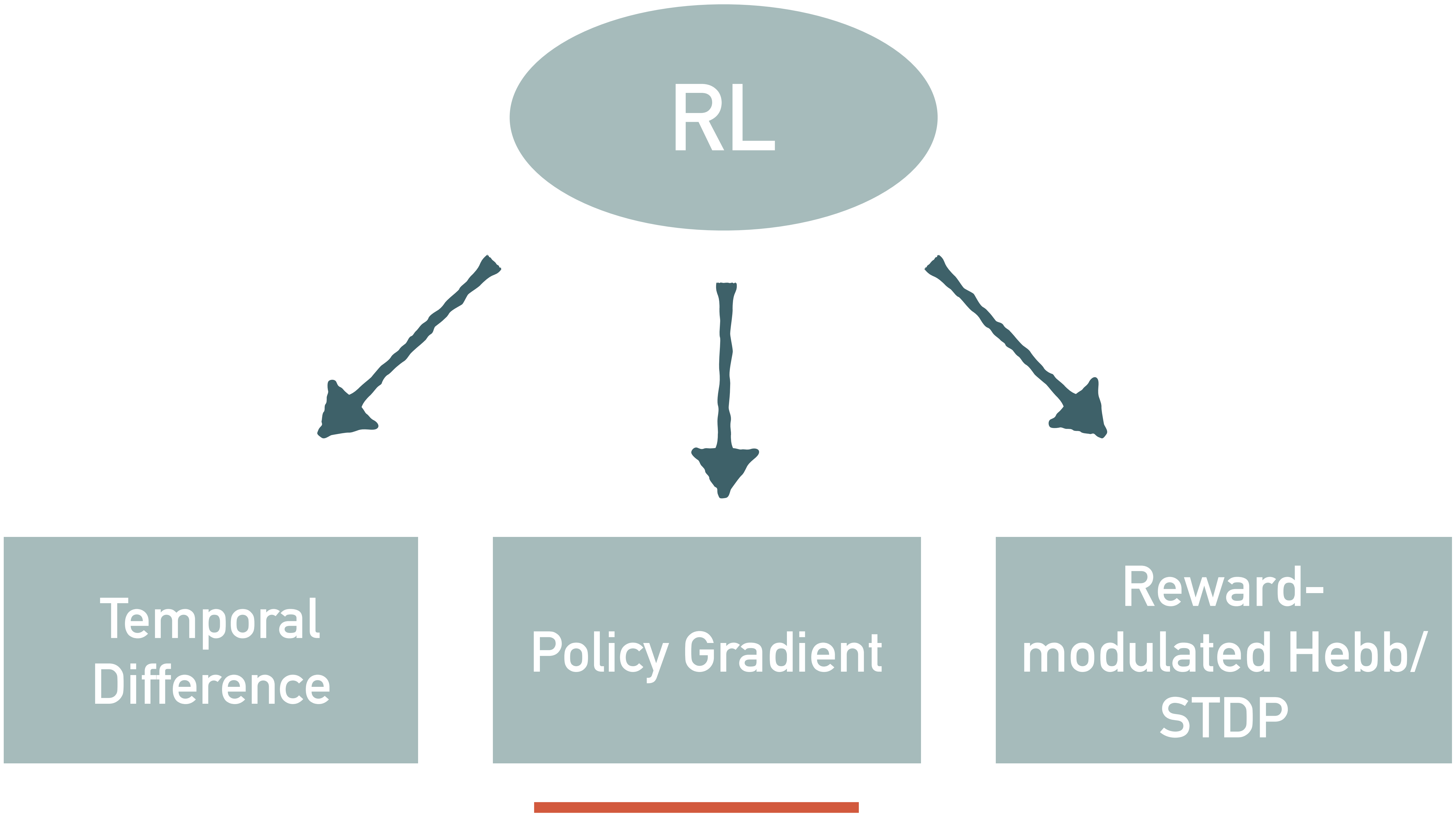
Zhao et al, IEEE SSCI 2016
Mnih et al, Nature 2015
Mnih et al, NIPS 2013

DEEP REINFORCEMENT LEARNING

https://deepmind.com/research/dqn/

Mnih et al, Nature 2015
Mnih et al, NIPS 2013

REINFORCEMENT LEARNING CATEGORIES



POLICY GRADIENT METHODS

$$\langle R \rangle_{x,y} = \sum_{x,y} R(x, y)P(y \,|\, x)P(x)$$

$$\langle \Delta w \rangle_{x,y} = \alpha \frac{\partial \langle R \rangle_{x,y}}{\partial w}$$

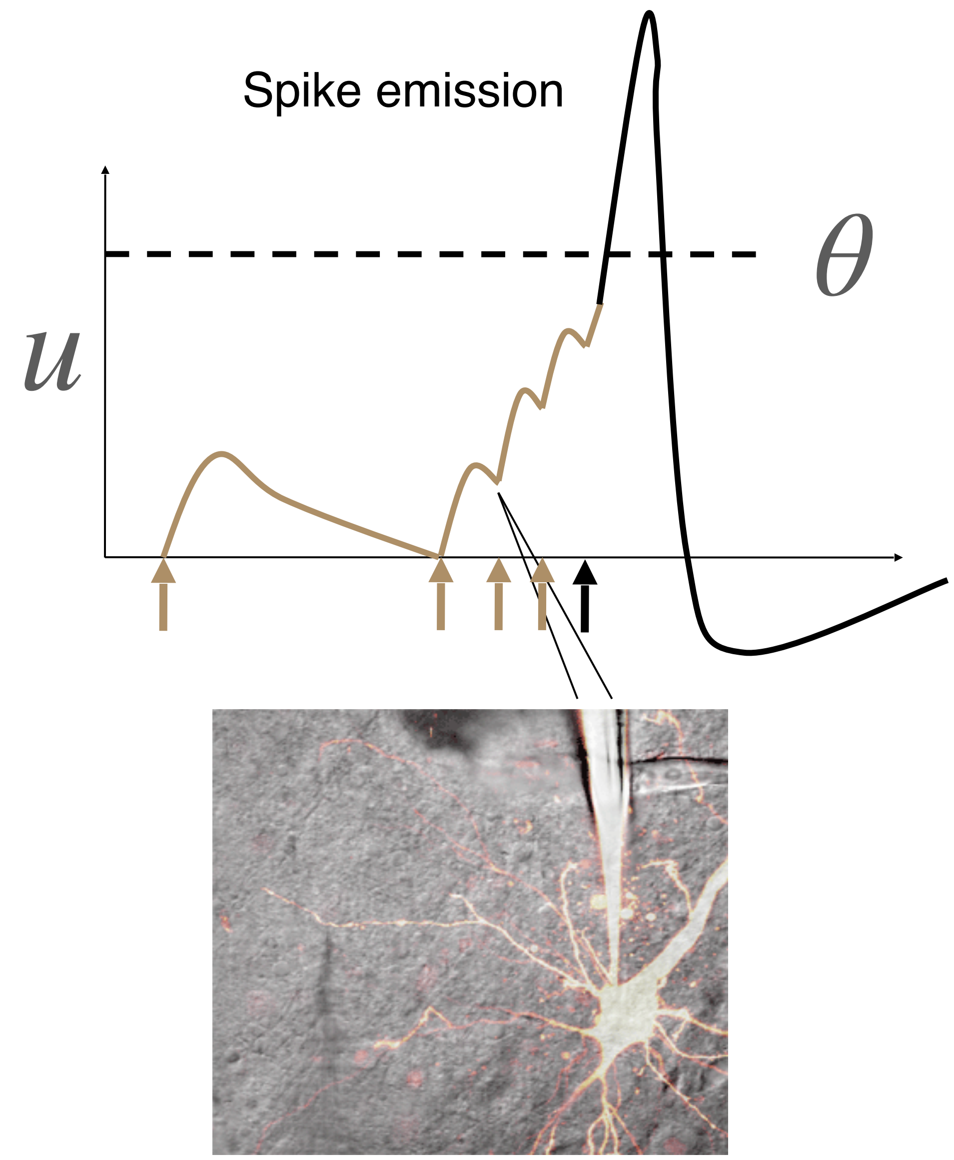
$$\frac{dw_{ij}}{dt} = \underbrace{\alpha}_{\text{Learning rate}} \overbrace{R(t)}^{\text{Reward}} \underbrace{e_{ij}(t)}_{\text{Eligibility trace}}$$

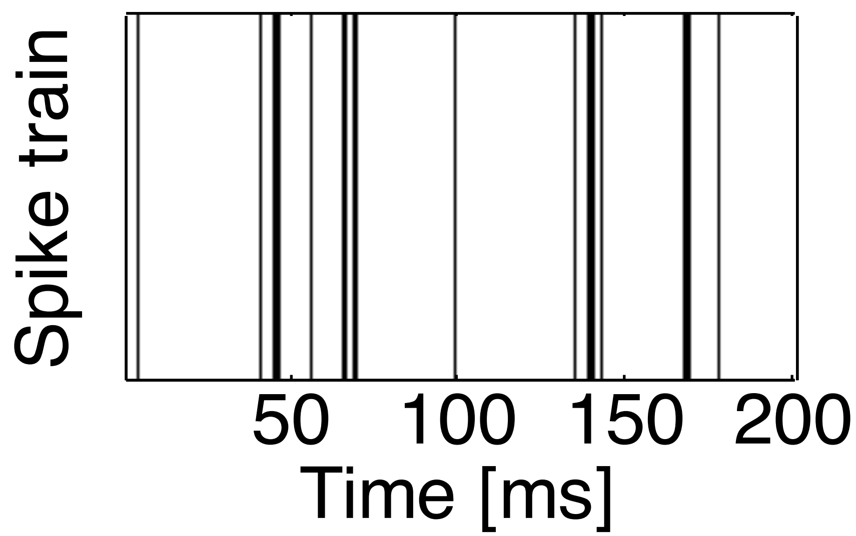
$$\frac{de_{ij}}{dt} = -\frac{e_{ij}}{\underbrace{\tau_e}_{\text{Time constant}}} + \left(\underbrace{Y_i(t)}_{\text{Spike}} - \overbrace{\rho(t)}^{\text{Instantaneous probability of firing}} \right) \overbrace{\sum_{t_j^f} \epsilon(t - t_j^f)}^{\text{Input at synapse ij}}$$

Williams 1992
Xie and Seung, 2004
Pfister et al. 2006
Florian 2007
Vasilaki et al, 2009

SPIKING NEURONS

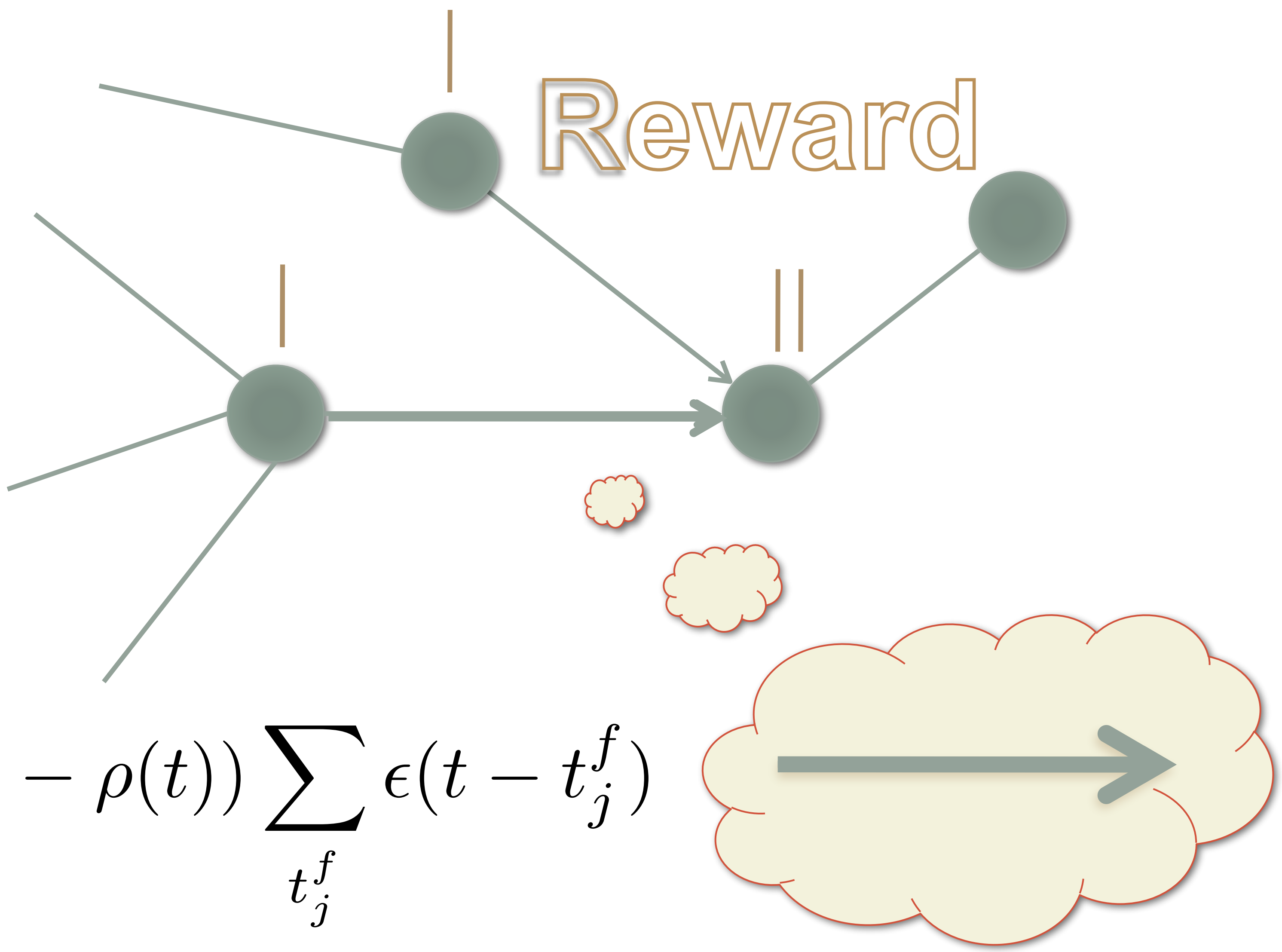
- Spikes
- Threshold





Stein, 1965
Gerstner et al, 2002

ELIGIBILITY TRACE



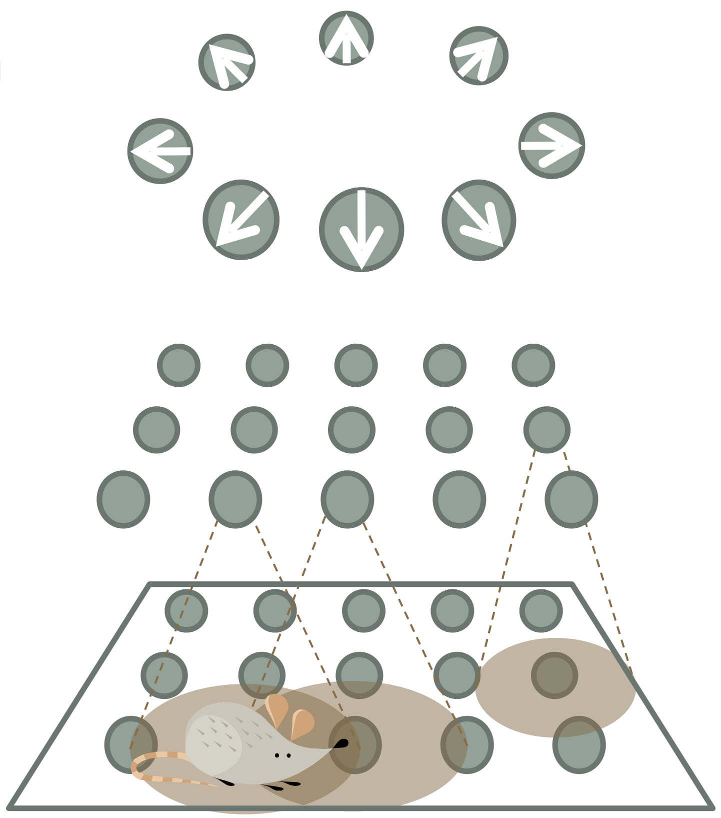
$$(Y_i(t) - \rho(t)) \sum_{t_j^f} \epsilon(t - t_j^f)$$

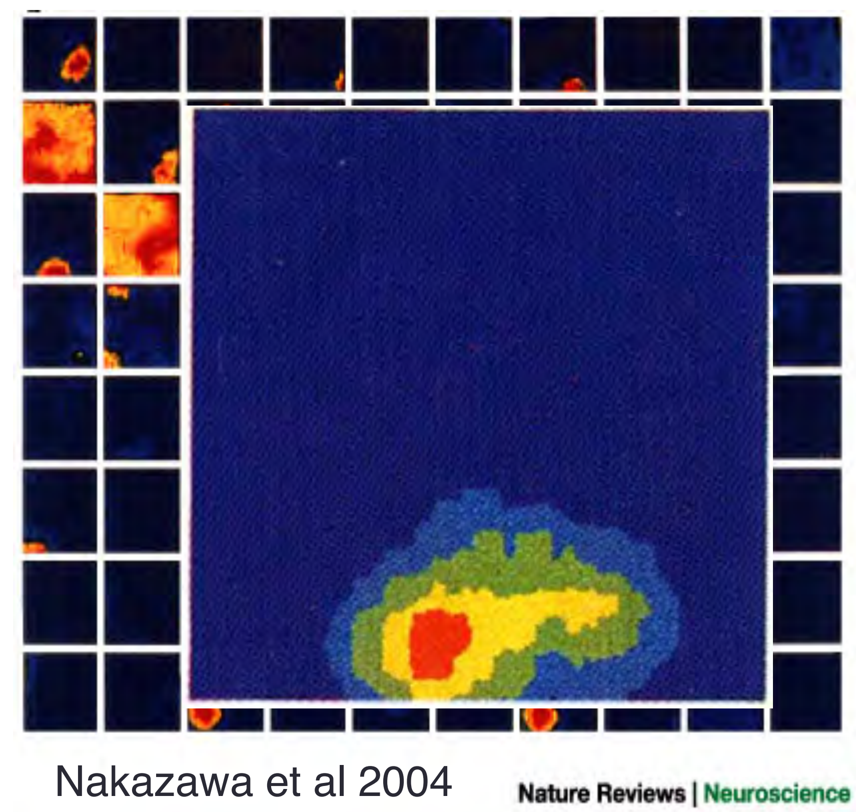
SPIKE BASED REINFORCEMENT LEARNING

Action encoded by the centre of mass of the activity.

Action Cells

Place Cells



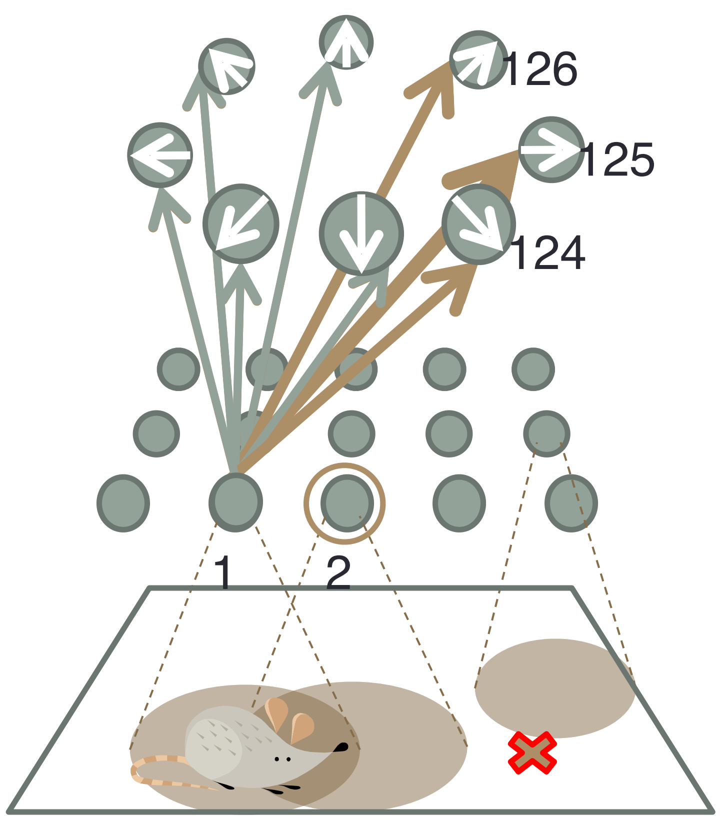


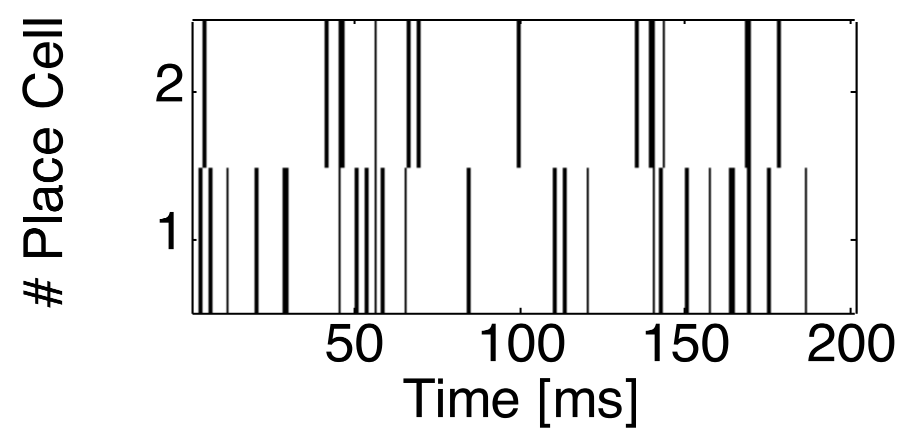
Nakazawa et al 2004

O'Keefe & Dostrovsky , 1971

Vasilaki et al, PLOS Comb Biol 2009

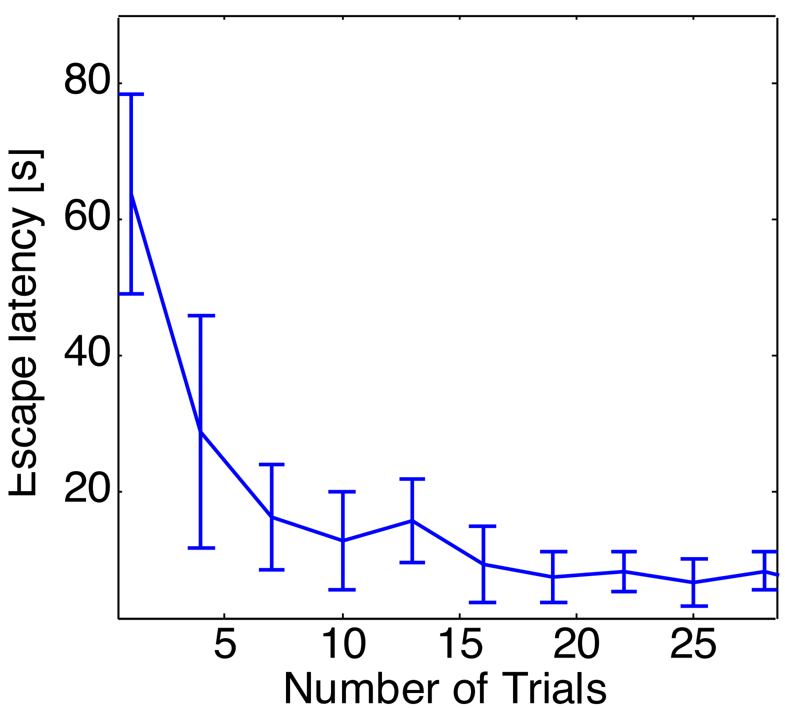
SPIKE BASED REINFORCEMENT LEARNING

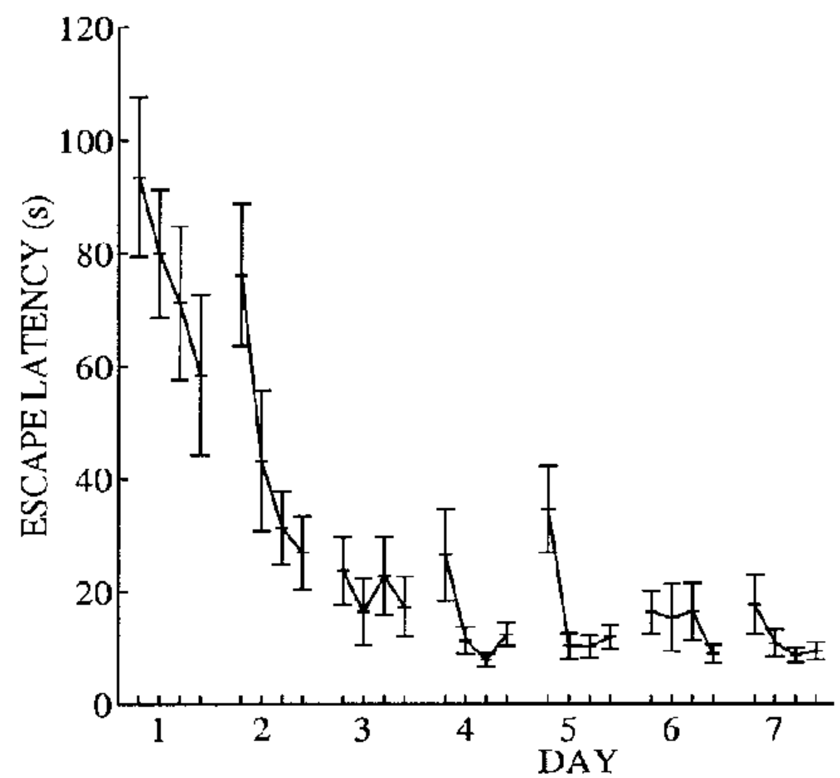




Vasilaki et al, PLOS Comb Biol 2009

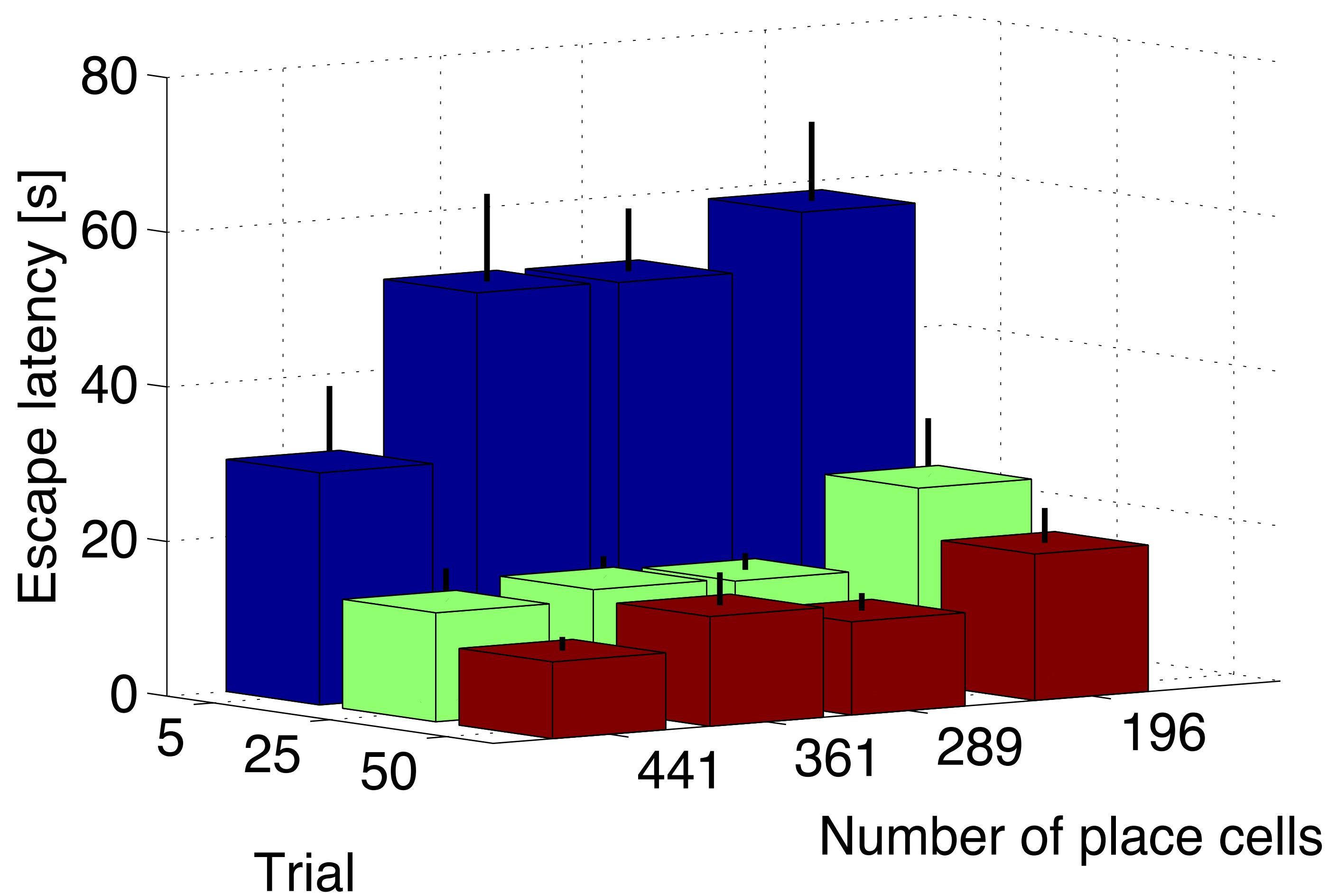
SIMULATION VS EXPERIMENT





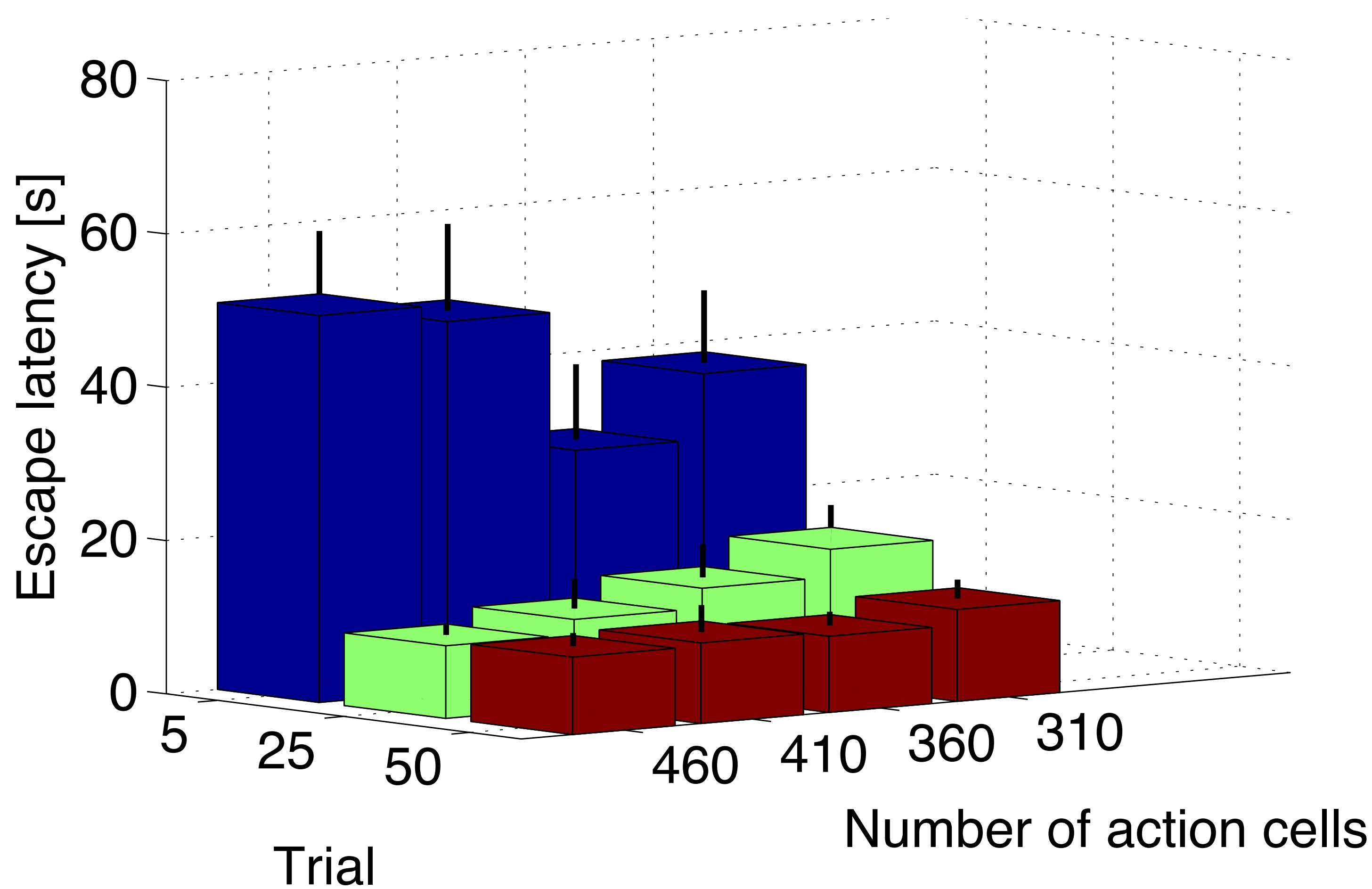
Vasilaki et al, PLOS Comb Biol 2009

SCALING PROPERTIES



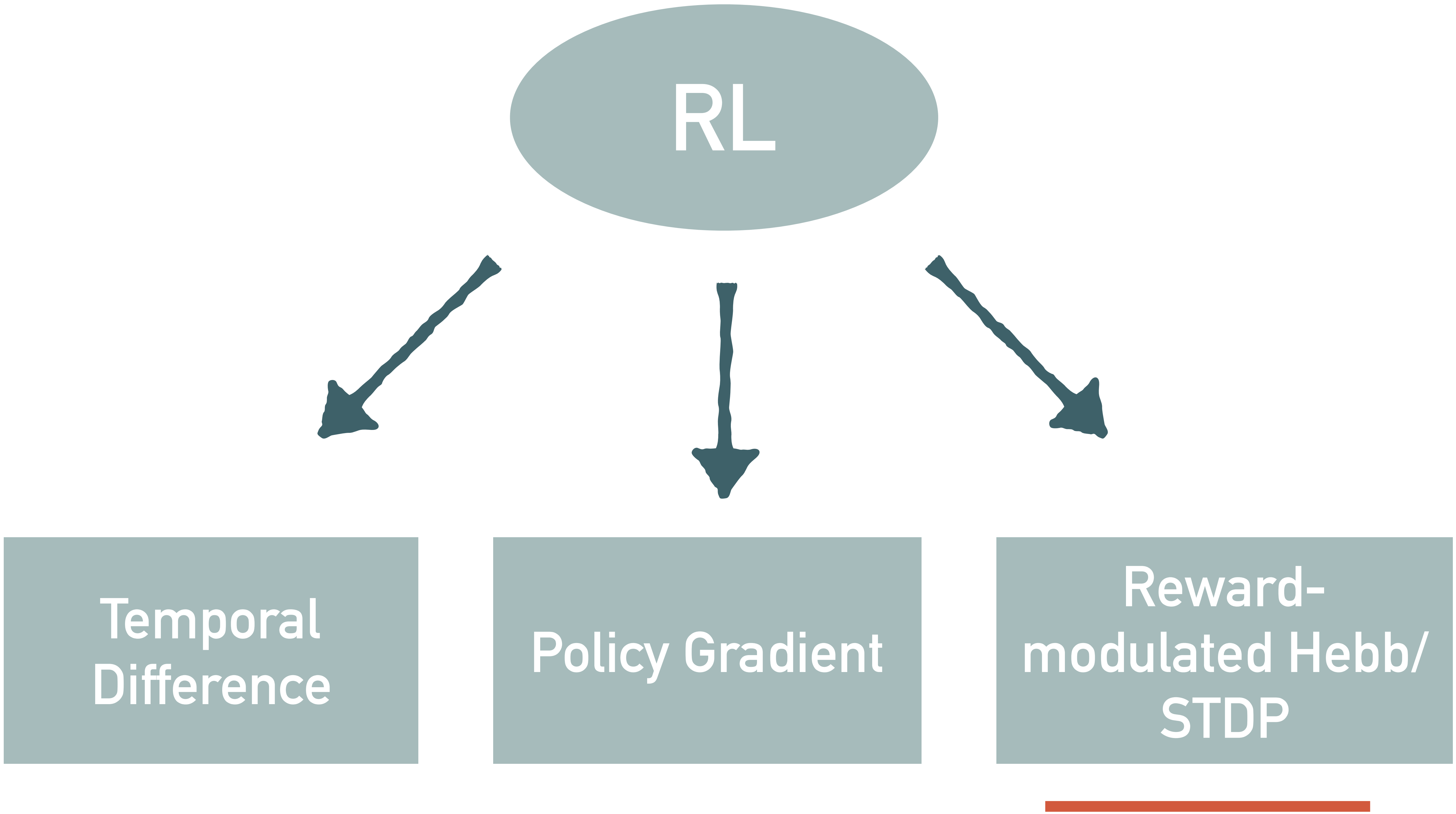
Vasilaki et al, PLOS Comb Biol 2009

SCALING PROPERTIES



Vasilaki et al, PLOS Comb Biol 2009

REINFORCEMENT LEARNING CATEGORIES



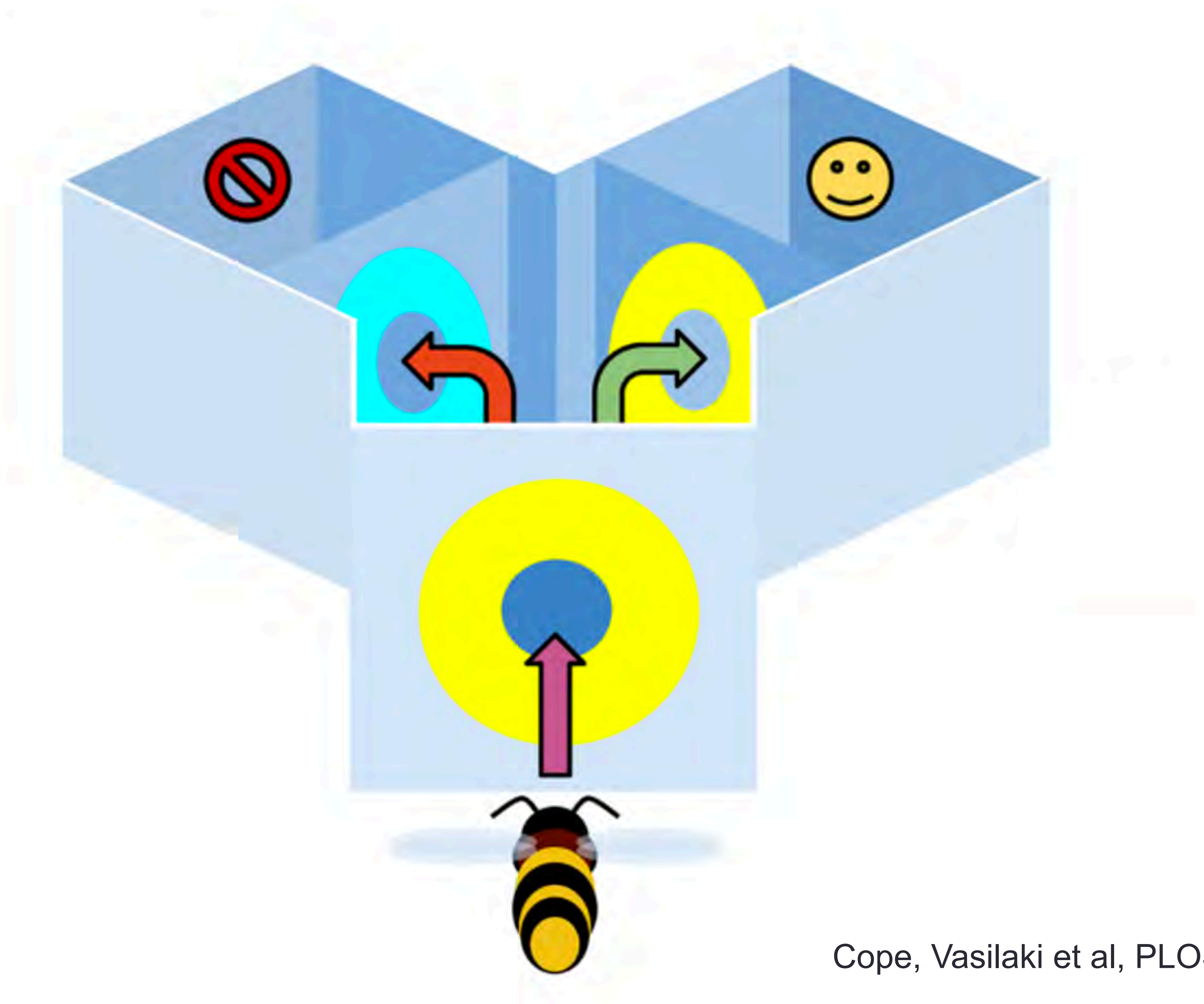
REWARD MODULATED HEBB

"Neurons that fire together wire together"

Correlations

$$\Delta w = g(reward) \times f(presynaptic\ activity,\ postsynaptic\ activity)$$

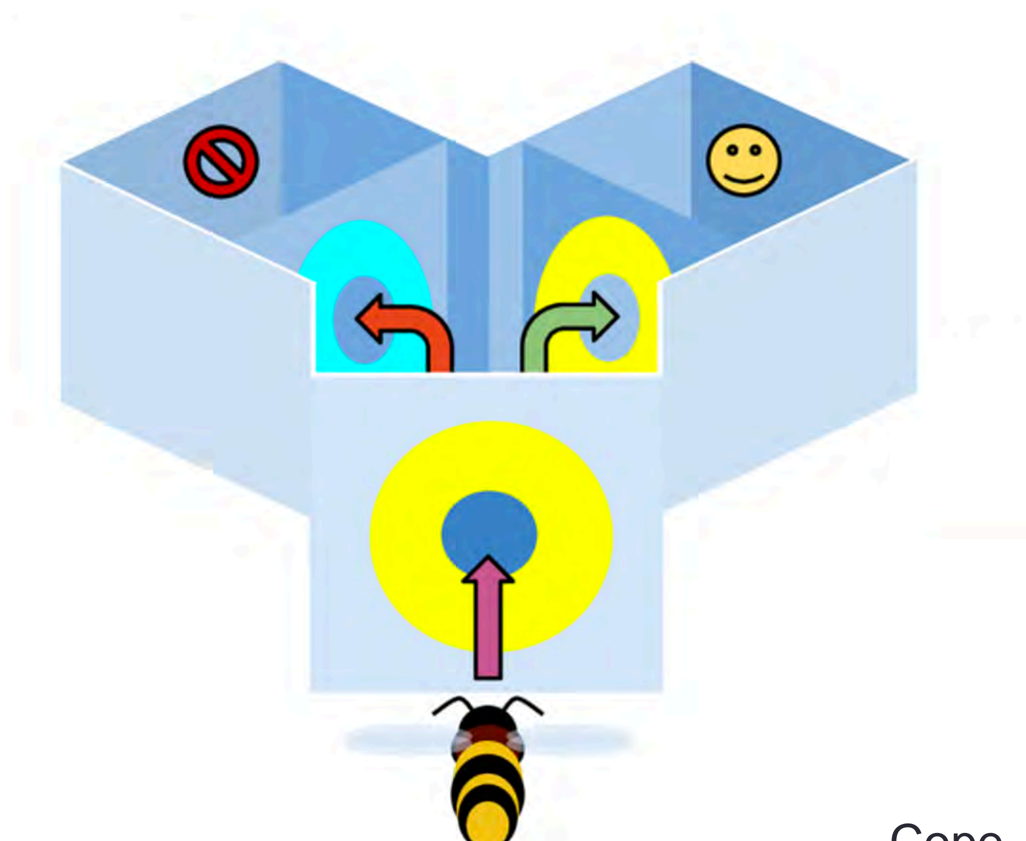
TO BEE OR NOT TO BEE



Cope, Vasilaki et al, PLOS Comb Biol 2018

TO BEE OR NOT TO BEE

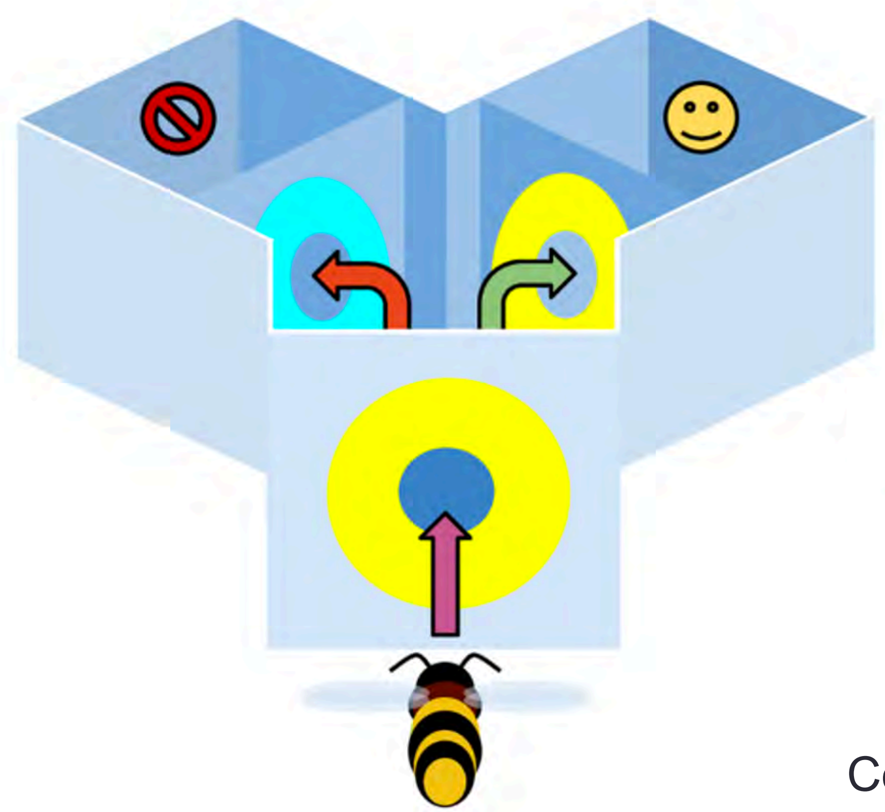
- Learns with visual stimuli
- Generalises to odours
  - Does leaning “sameness” require higher order cognition?



Cope, Vasilaki et al, PLOS Comb Biol 2018

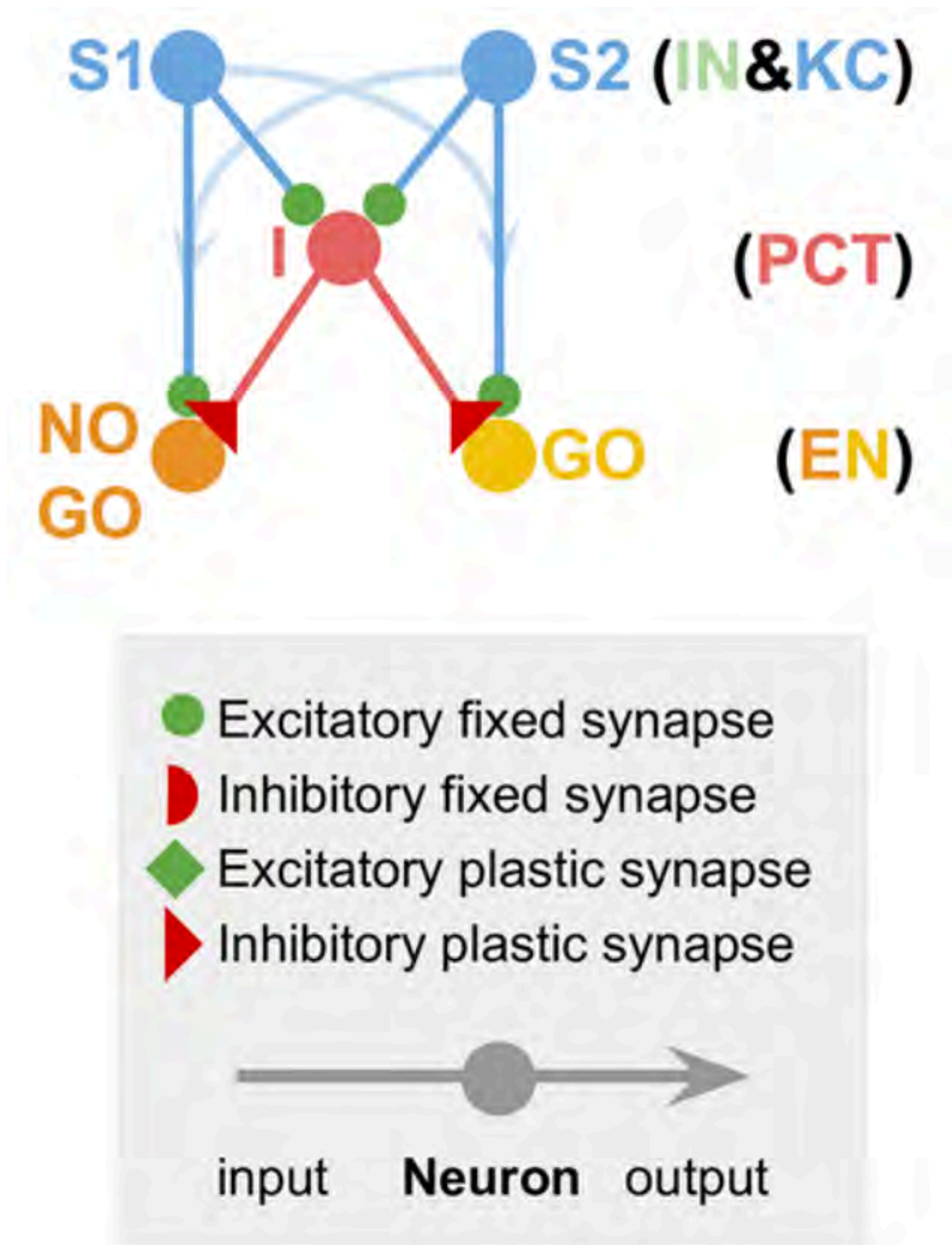
TO BEE OR NOT TO BEE

- Key idea: it learns to go to the pattern of the highest/lowest activity.
- "Stimulus-specific adaptation" - like observations. Repetitive stimulus results in lower activity.



Cope, Vasilaki et al, PLOS Comb Biol 2018

SAMENESS, DIFFERENCE AND TRANSFER OF LEARNING

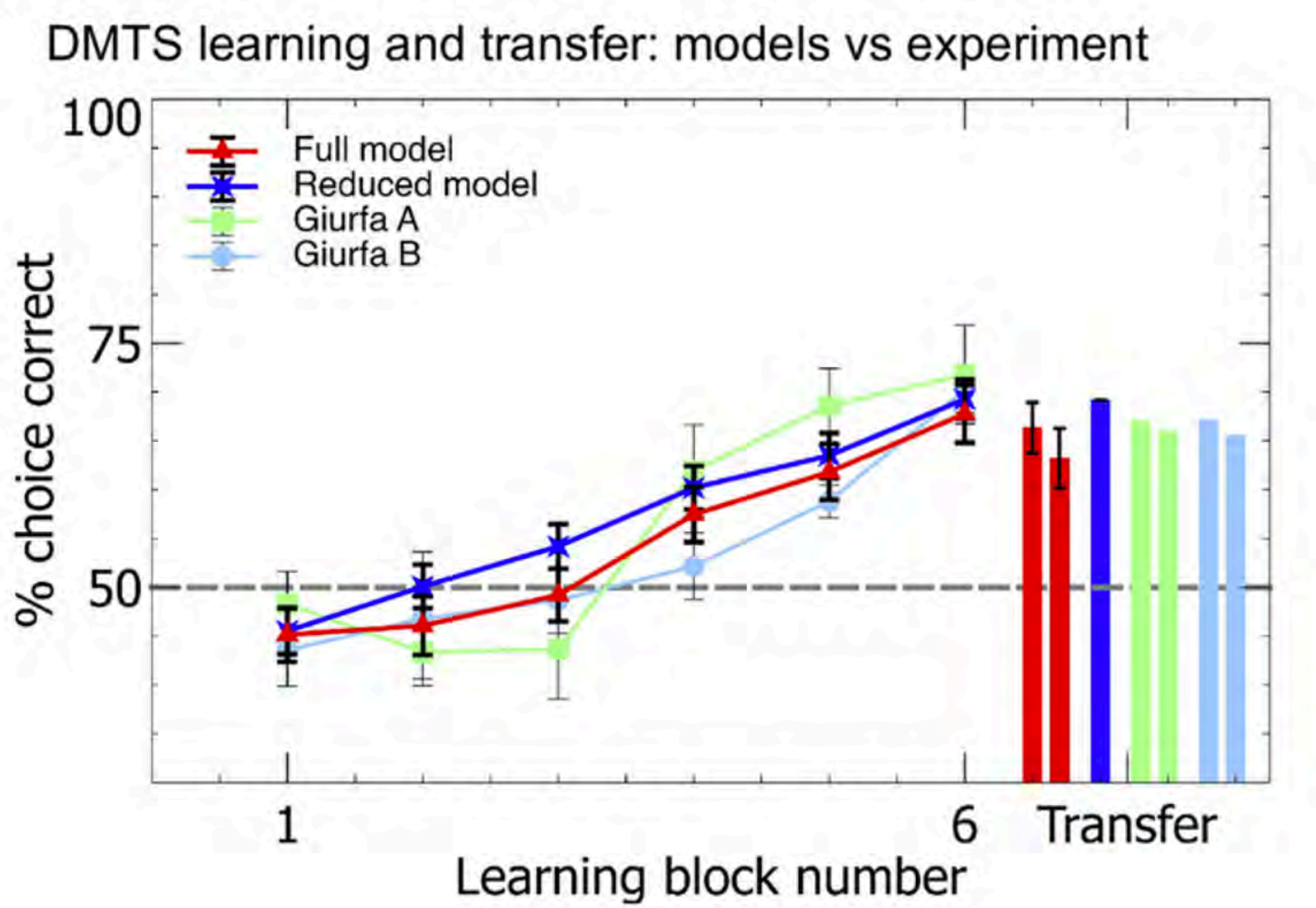


$$P(GO) = \frac{1}{1 + e^{-(c-d)(GO-NOGO)}}$$

$$P(NOGO) = 1 - P(GO)$$

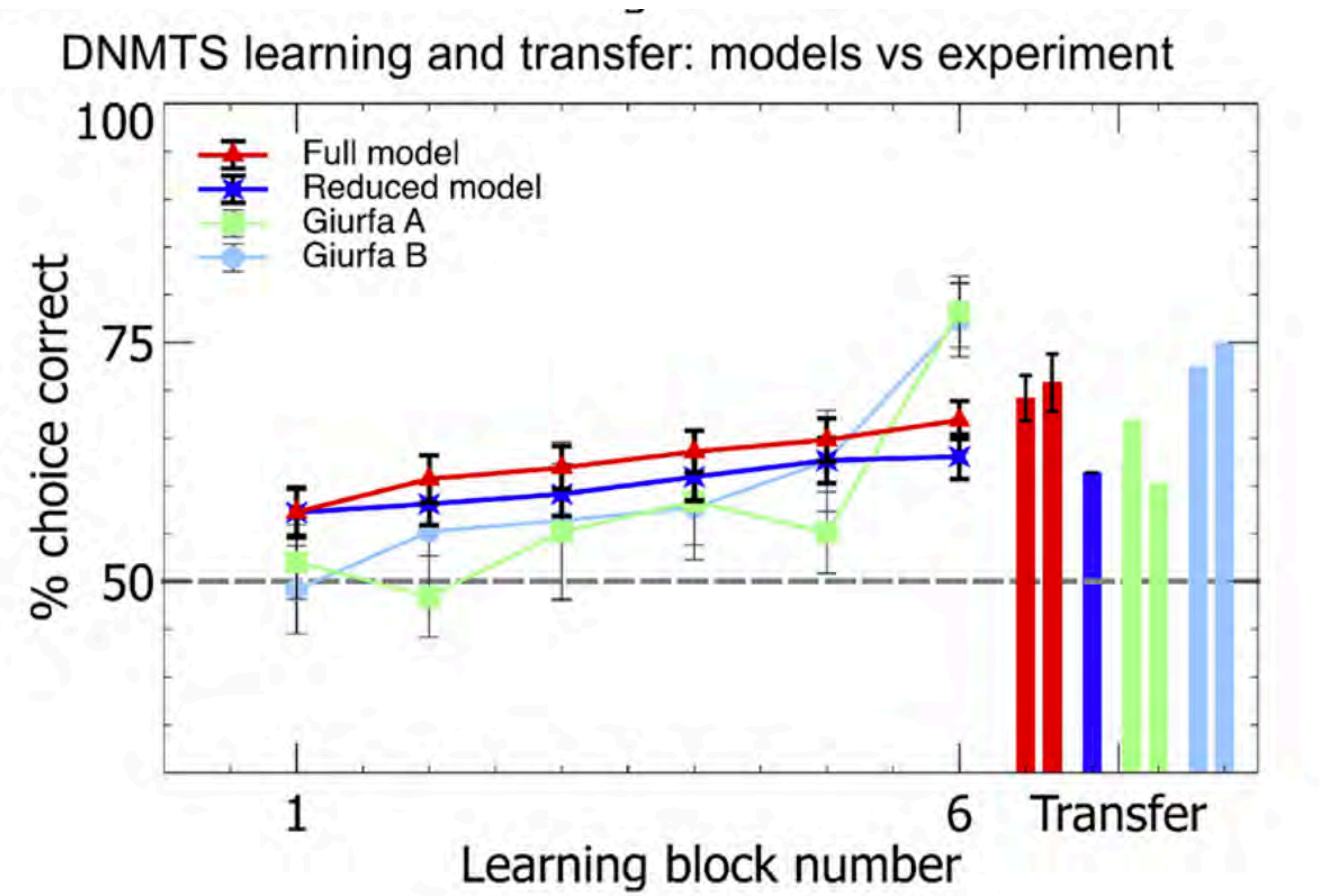
$$\Delta w^i = -\eta \ (R - R_b) \times presynaptic\ activity \times postsynaptic\ activity$$

SAMENESS, DIFFERENCE AND TRANSFER OF LEARNING



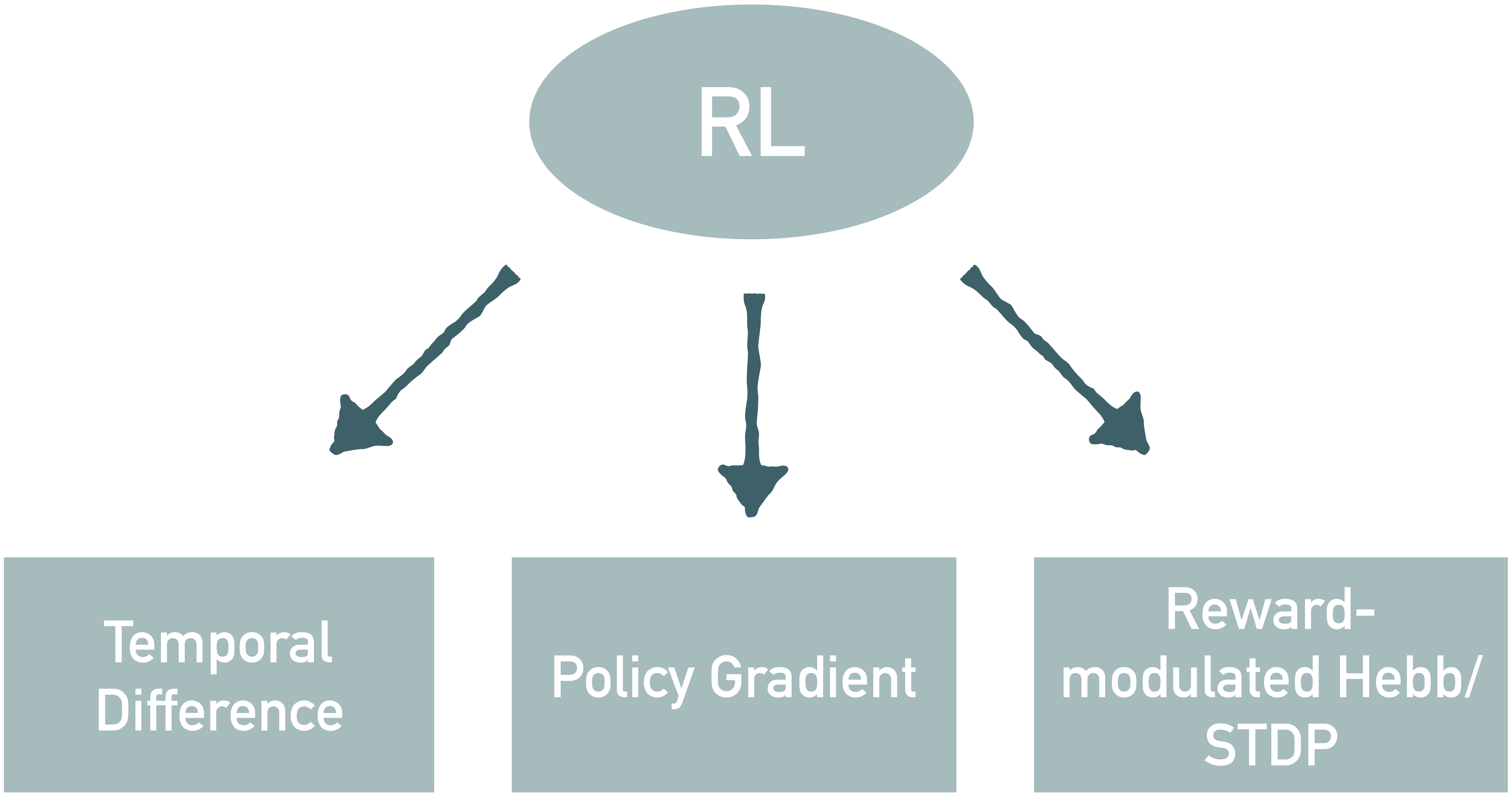
Cope, Vasilaki et al, PLOS Comb Biol 2018

SAMENESS, DIFFERENCE AND TRANSFER OF LEARNING



Cope, Vasilaki et al, PLOS Comb Biol 2018

SUMMARY


RL
Temporal Difference
Policy Gradient
Reward-modulated Hebb/STDP

SUMMARY

- Eligibility & Future Rewards
  - Not necessary in TD learning but can speed up (depending on the task)
  - May be necessary in Policy Gradient learning (depending on the formalism)
  - May be necessary in reward-modulated Hebb/STDP (depending on the formalism)

ACKNOWLEDGEMENTS

plos.org

PLOS | COMPUTATIONAL BIOLOGY

Browse | Publish | About

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Spike-Based Reinforcement Learning in Continuous State and Action Space: When Policy Gradient Methods Fail

Eleni Vasilaki, Nicolas Frémaux, Robert Urbanczik, Walter Senn, Wulfram Gerstner

Published: December 4, 2009 • https://doi.org/10.1371/journal.pcbi.1000586

ACKNOWLEDGEMENTS

plos.org

PLOS | COMPUTATIONAL BIOLOGY

Browse | Publish | About

OPEN ACCESS   PEER-REVIEWED

RESEARCH ARTICLE

Abstract concept learning in a simple neural network inspired by the insect brain

Alex J. Cope, Eleni Vasilaki, Dorian Minors, Chelsea Sabo, James A. R. Marshall, Andrew B. Barron

Version 2

Published: September 17, 2018 • https://doi.org/10.1371/journal.pcbi.1006435

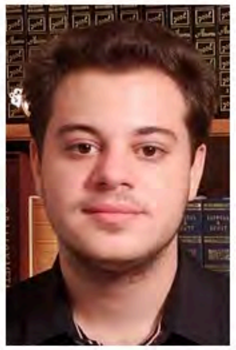
• >> See the preprint

Postdoctoral Fellows

Dr Natacha Vanattou-Saïfoudine — with INI, Zurich

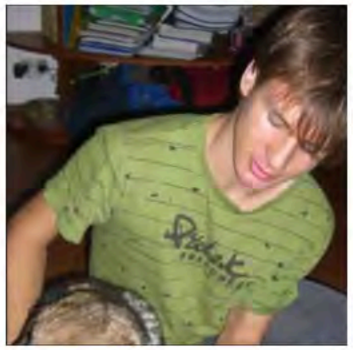
Dr Alex Cope — with Brains on Board project

PhD students

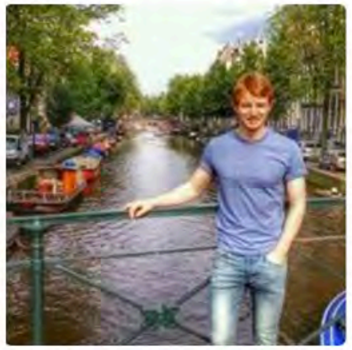


Avgoustinos Vouros

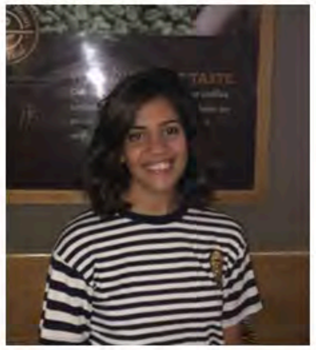
Chao Han



Luca Manneschi



Matthew Whelan — with Sheffield Robotics



Nada Abdelrahman — with Biomedical Sciences