MLSS 2012 in Kyoto

Brain and Reinforcement Learning

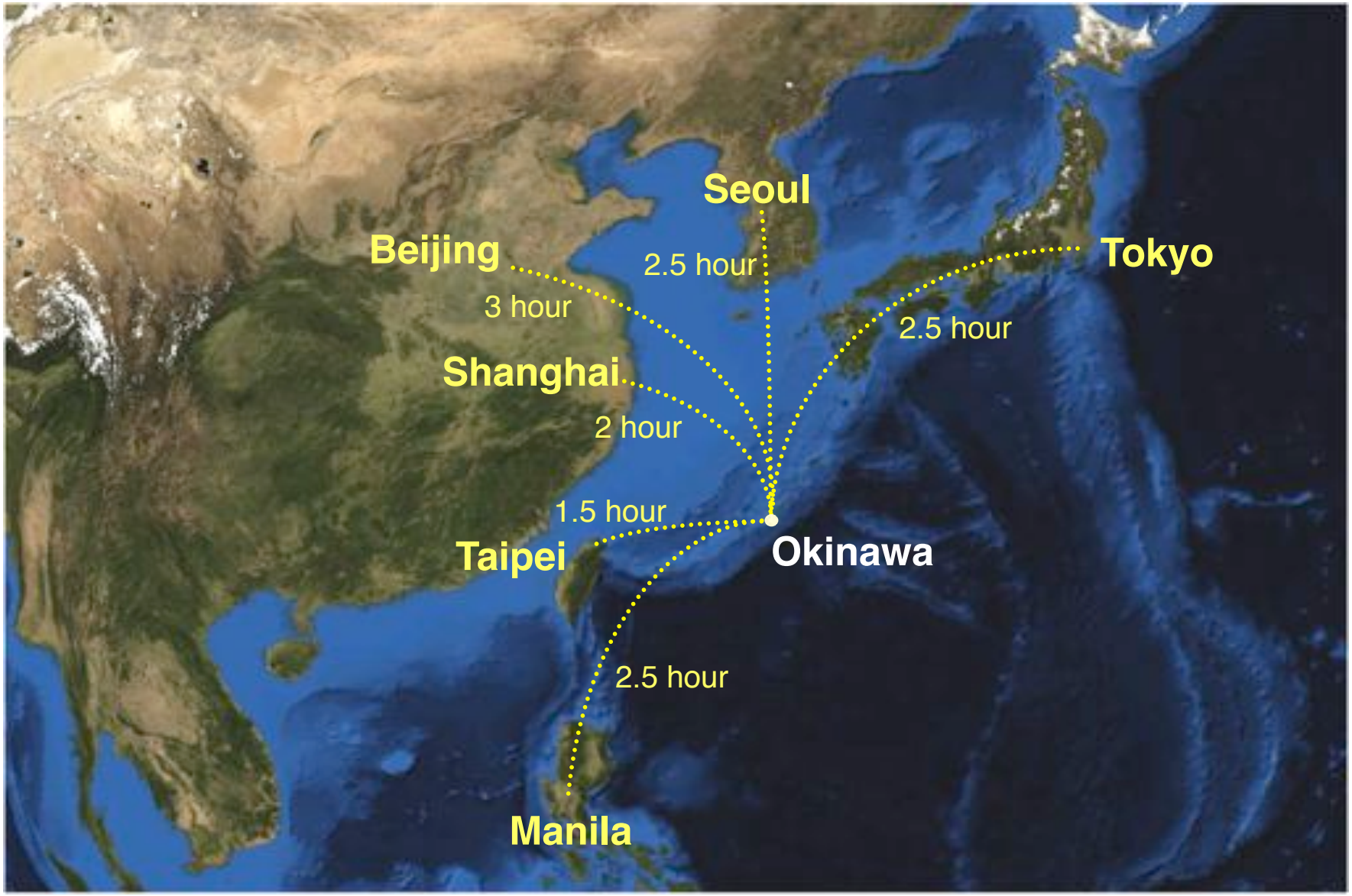
Kenji Doya

doya@oist.jp

Neural Computation Unit

Okinawa Institute of Science and Technology

Location of Okinawa



Okinawa Institute of Science & Technology

- Apr. 2004: Initial research
  - President: Sydney Brenner

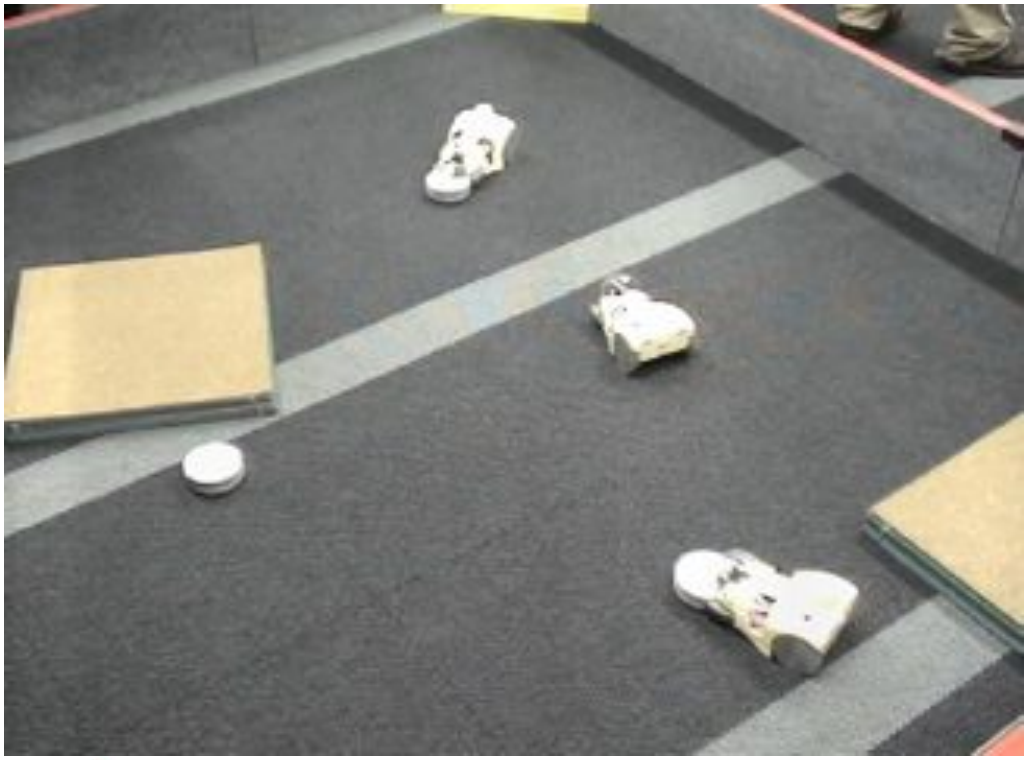
- Nov. 2011: Graduate university
  - President: Jonathan Dorfan

- Sept. 2012: Ph.D. course
  - 20 students/year

Our Research Interests

How to build adaptive, autonomous systems

- robot experiments

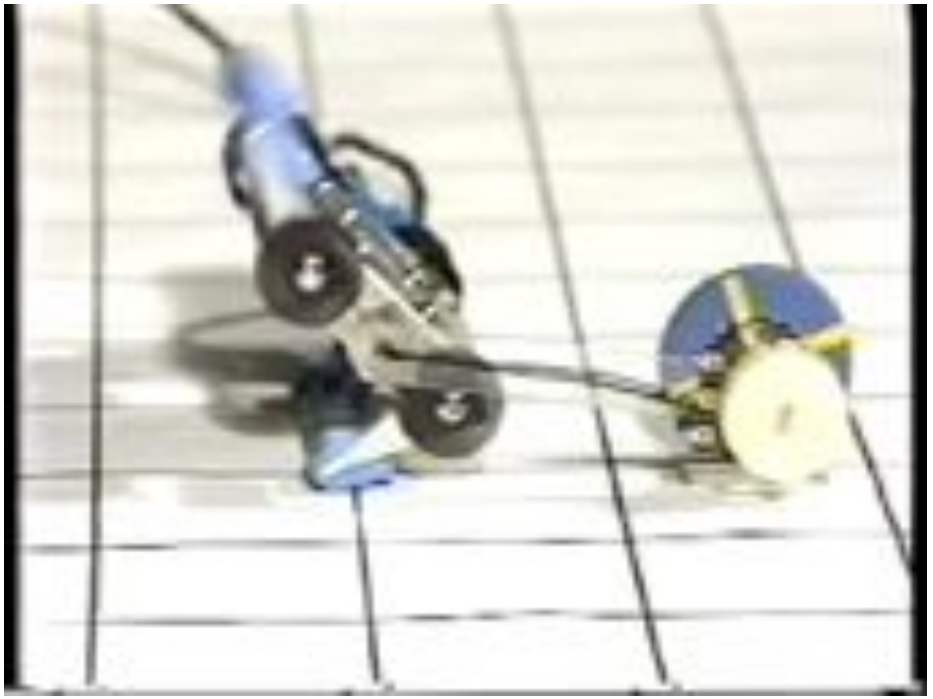


How the brain realizes robust, flexible adaptation

- neurobiology

Learning to Walk (Doya & Nakano, 1985)

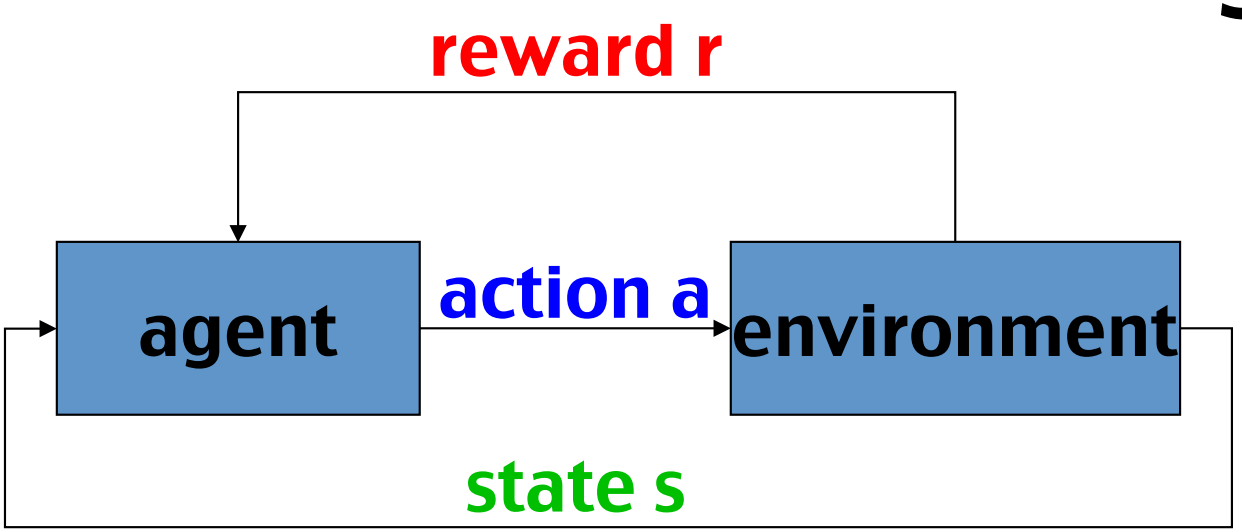
- Action: cycle of 4 postures
- Reward: speed sensor output



- Problem: a long jump followed by a fall

Need for long-term evaluation of action

Reinforcement Learning

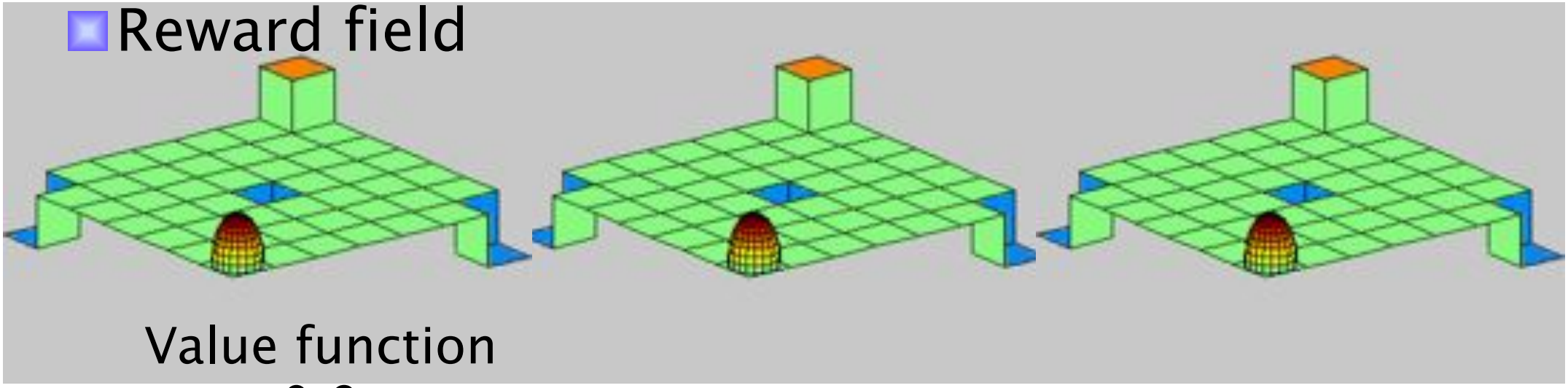


- Learn action policy: s → a to maximize rewards
- Value function: expected future rewards
  - $V(s(t)) = E[\, r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \gamma^3 r(t+3) + \ldots]$

    $0 \le \gamma \le 1$: discount factor $\gamma V(s(t+1))$
- Temporal difference (TD) error:
  - $\delta(t) = r(t) + \gamma V(s(t+1)) - V(s(t))$

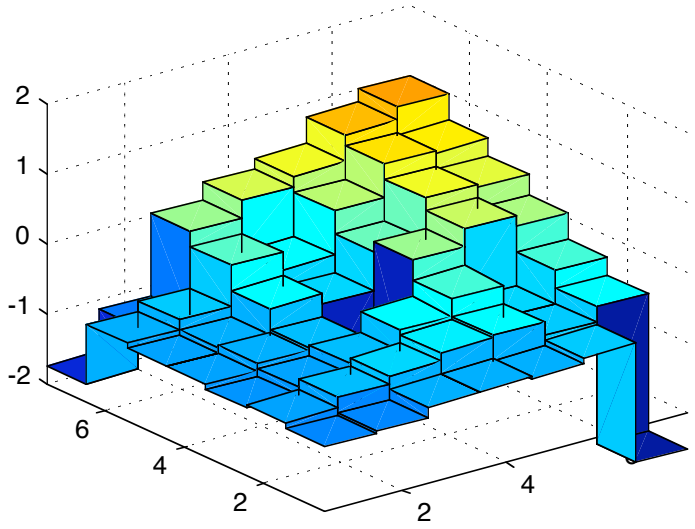
Example: Grid World

- Reward field

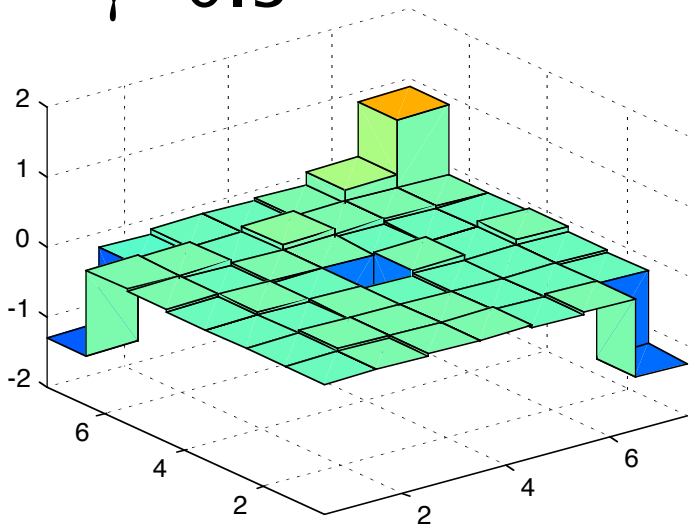


Value function

γ=0.9

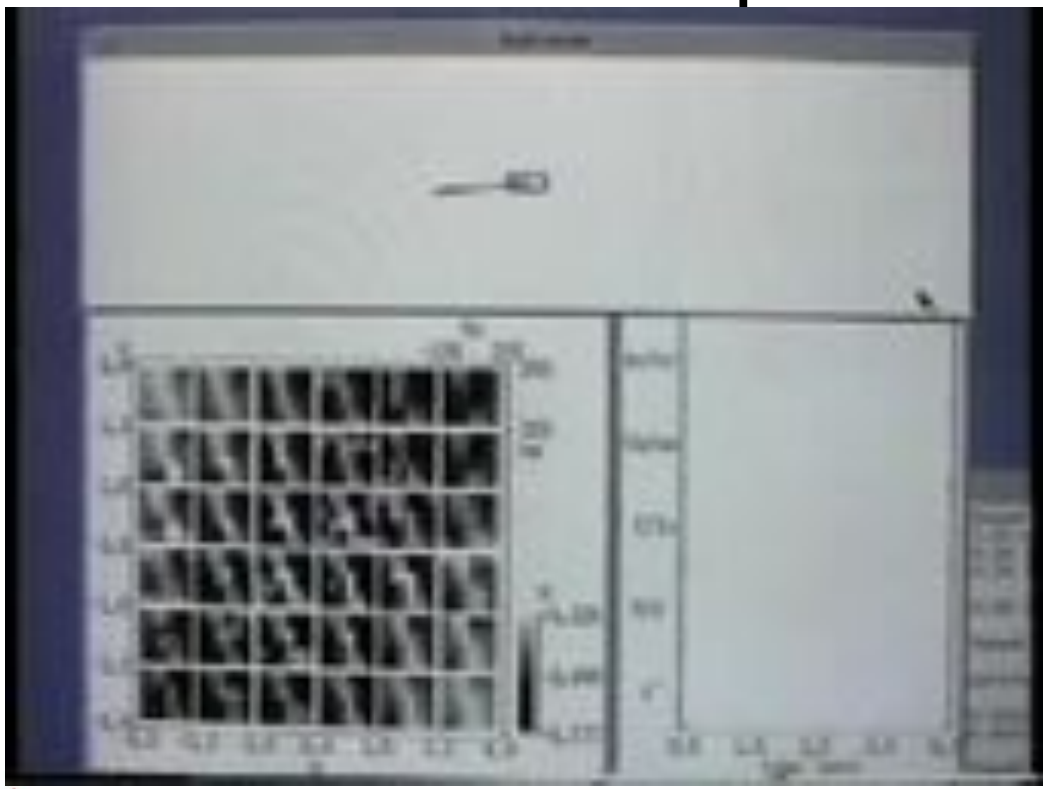


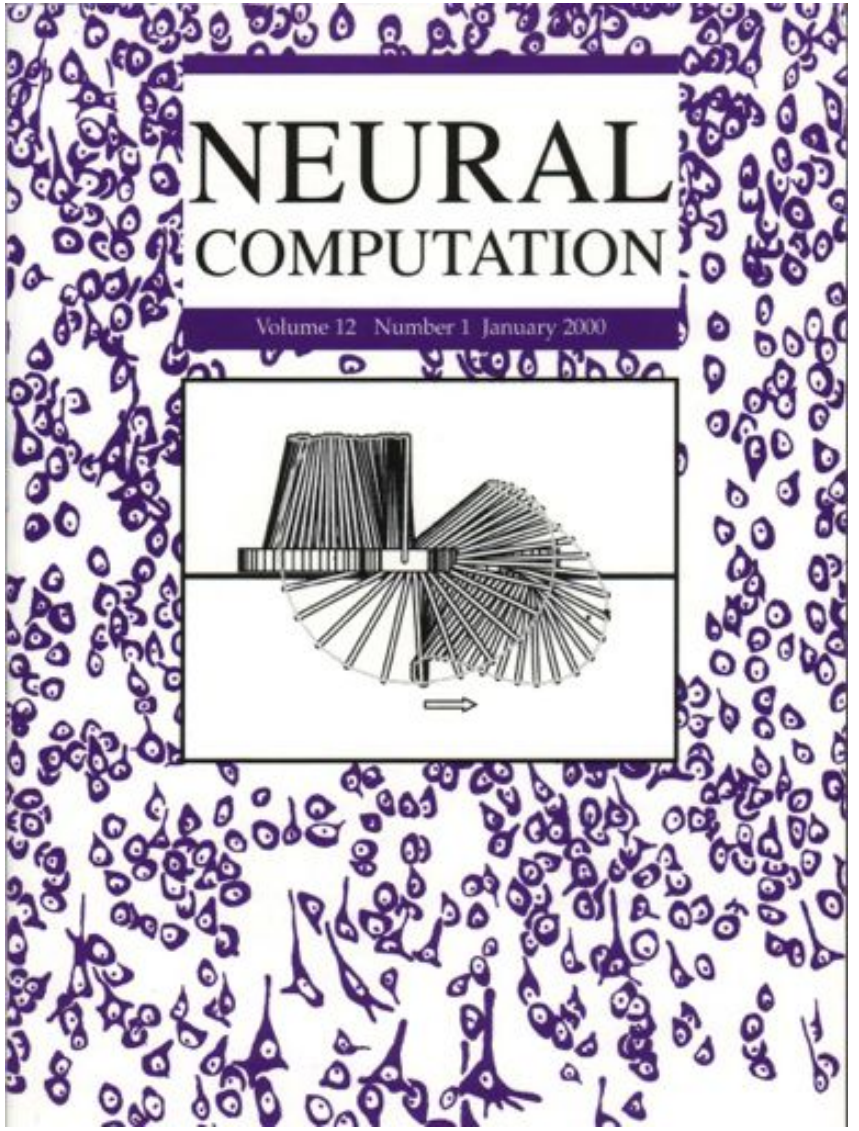
γ=0.3



Cart-Pole Swing-Up

- Reward: height of pole
- Punishment: collision
- Value in 4D state space



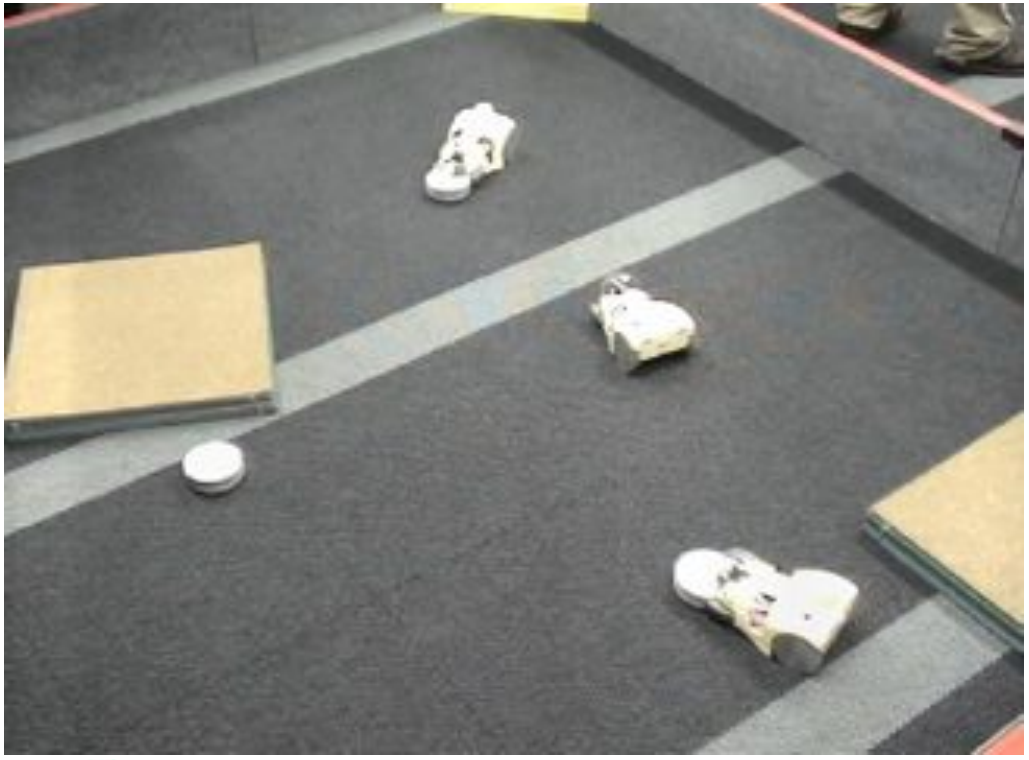


Learning to Stand Up (Morimoto & Doya, 2000)

- State: joint/head angles, angular velocity
- Action: torques to motors
- Reward: head height – tumble

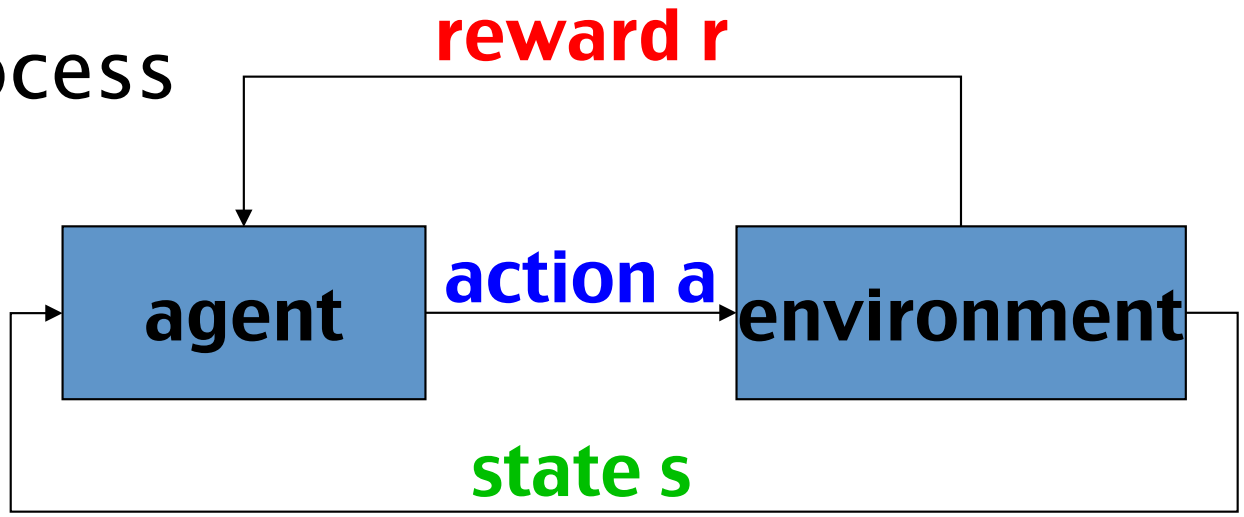
Learning to Survive and Reproduce

- Catch battery packs
  - survival
- Copy ‘genes’ by IR ports
  - reproduction, evolution



Markov Decision Process (MDP)

- Markov decision process
  - state $s \in S$
  - action $a \in A$
  - policy $p(a|s)$
  - reward $r(s,a)$
  - dynamics $p(s'|s,a)$



- Optimal policy: maximize cumulative reward
  - finite horizon: $E[r(1) + r(2) + r(3) + \ldots + r(T)]$
  - infinite horizon: $E[r(1) + \gamma r(2) + \gamma^2 r(3) + \ldots]$

    $0 \leq \gamma \leq 1$: temporal discount factor
  - average reward: $E[r(1) + r(2) + \ldots + r(T)]/T$, $T \to \infty$

Solving MDPs

Dynamic Programming

- p(s'|s,a) and r(s,a) are known
- Solve Bellman equation

  $V(s) = \max_a E[r(s,a) + \gamma V(s')]$
  - V(s): value function

    expected reward from state s
- Apply optimal policy

  $a = \text{argmax}_a E[r(s,a) + \gamma V^*(s')]$
- Value iteration
- Policy iteration

Reinforcement Learning

- p(s'|s,a) and r(s,a) are unknown
- Learn from actual experience

  {s,a,r,s,a,r,…}
- Monte Carlo
- SARSA
- Q-learning
- Actor-Critic
- Policy gradient
- Model-based
  - learn p(s'|s,a), r(s,a) and do DP

Actor-Critic and TD learning

- Actor: parameterized policy: $P(a|s; w)$
- Critic: learn value function

$$V(s(t)) = E[\, r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \ldots]$$

  - in a table or a neural network
- Temporal Difference (TD) error:
  - $\delta(t) = r(t) + \gamma\, V(s(t+1)) - V(s(t))$
- Update
  - Critic: $\Delta V(s(t)) = \alpha\, \delta(t)$
  - Actor: $\Delta w = \alpha\, \delta(t)\, \partial P(a(t)|s(t);w)/\partial w$

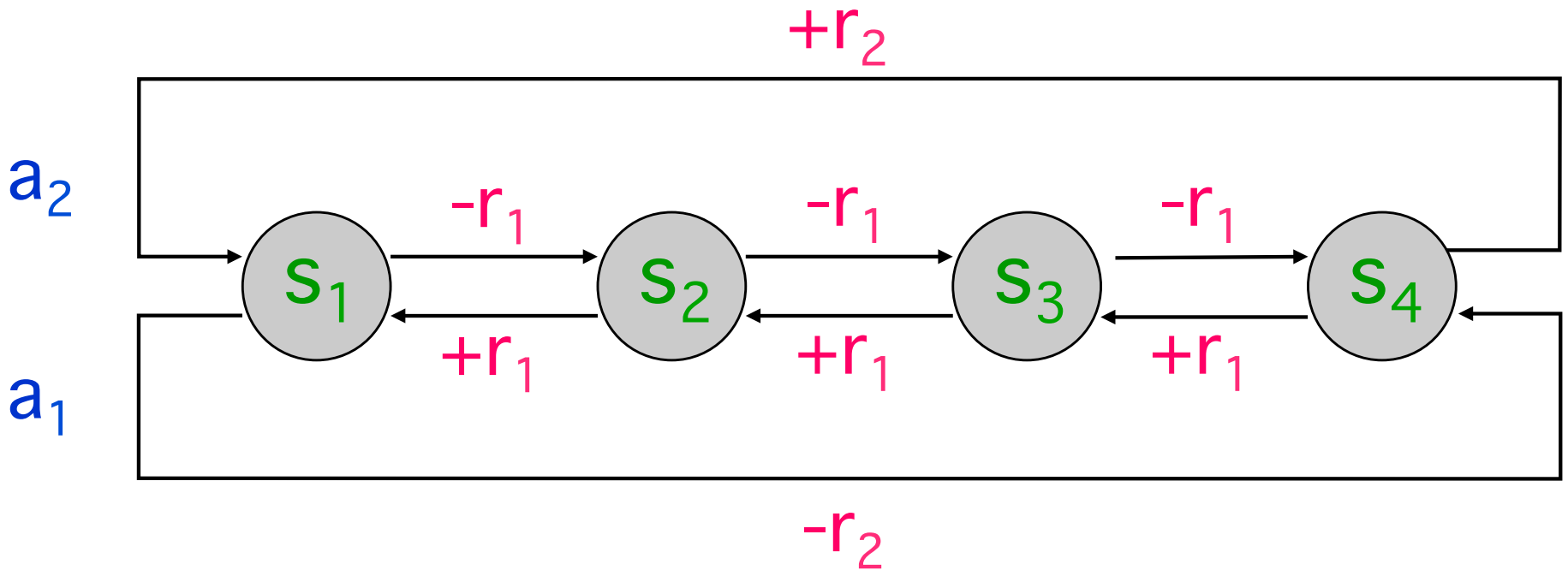
    … reinforce a(t) by $\delta(t)$

SARSA and Q Learning

- Action value function
  - $Q(s,a) = E[\, r(t) + \gamma r(t+1) + \gamma^2 r(t+2) \ldots | \, s(t)=s, a(t)=a]$
- Action selection
  - ε-greedy: $a = \text{argmax}_a \, Q(s,a)$ with prob $1-\varepsilon$
  - Boltzman: $P(a_i|s) = \exp[\beta Q(s,a_i)] \, / \, \Sigma_j \exp[\beta Q(s,a_j)]$
- SARSA: on-policy update
  - $\Delta Q(s(t),a(t)) = \alpha\{r(t)+\gamma Q(s(t+1),a(t+1))-Q(s(t),a(t))\}$
- Q learning: off-policy update
  - $\Delta Q(s(t),a(t)) = \alpha\{r(t)+\gamma \text{max}_{a'} Q(s(t+1),a')-Q(s(t),a(t))\}$

“Lose to Gain” Task

- N states, 2 actions



- if $r_2 >> r_1$, then better take a_2

Reinforcement Learning

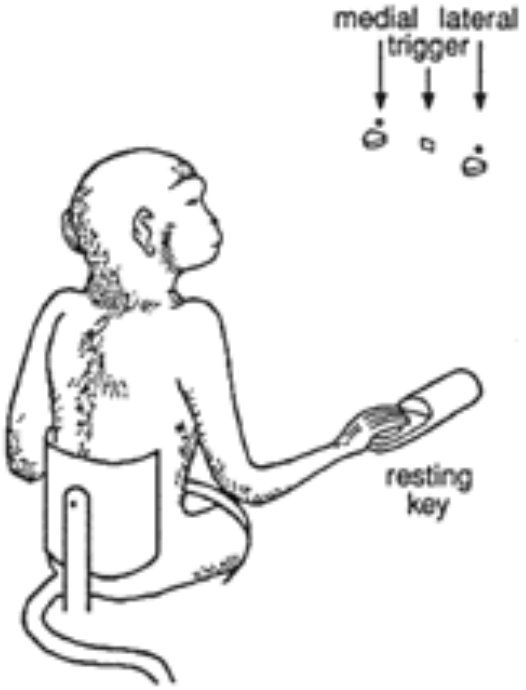
- Predict reward: *value function*
  - $V(s) = E[\, r(t) + \gamma r(t+1) + \gamma^2 r(t+2) \ldots | \, s(t)=s]$
  - $Q(s,a) = E[\, r(t) + \gamma r(t+1) + \gamma^2 r(t+2) \ldots | \, s(t)=s, \, a(t)=a]$
- Select action
  - *greedy*: $a = \text{argmax}\, Q(s,a)$
  - *Boltzmann*: $P(a|s) \propto \exp[\, \beta \, Q(s,a)]$
- Update prediction: *TD error*
  - $\delta(t) = r(t) + \gamma V(s(t+1)) - V(s(t))$
  - $\Delta V(s(t)) = \alpha \, \delta(t)$
  - $\Delta Q(s(t),a(t)) = \alpha \, \delta(t)$

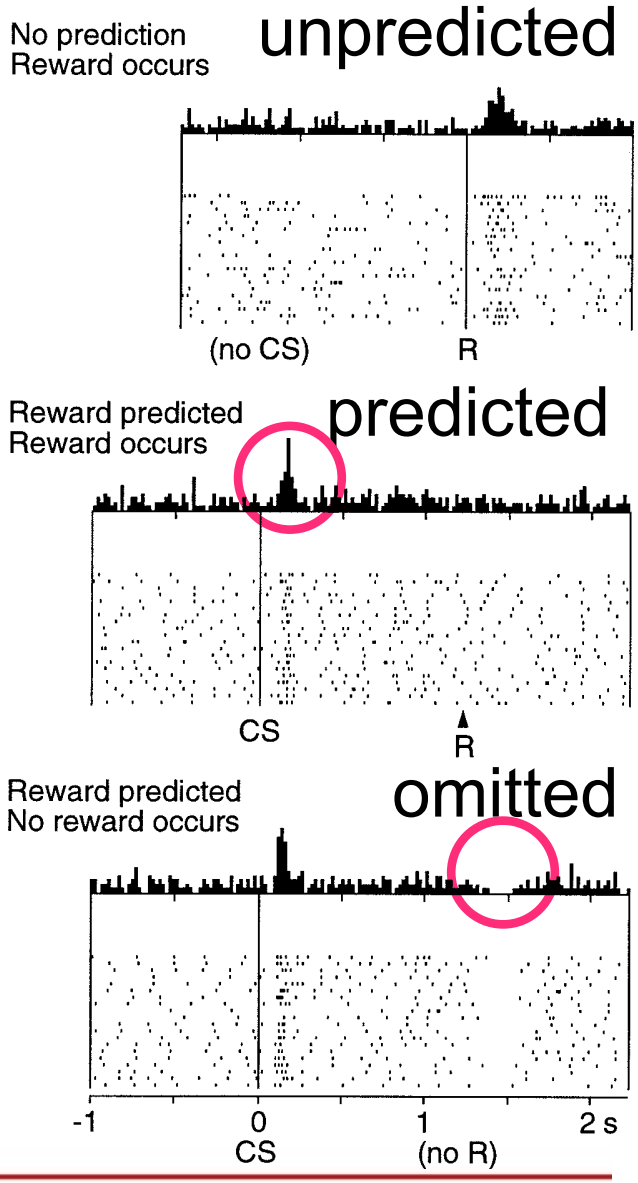
How to implement these steps?

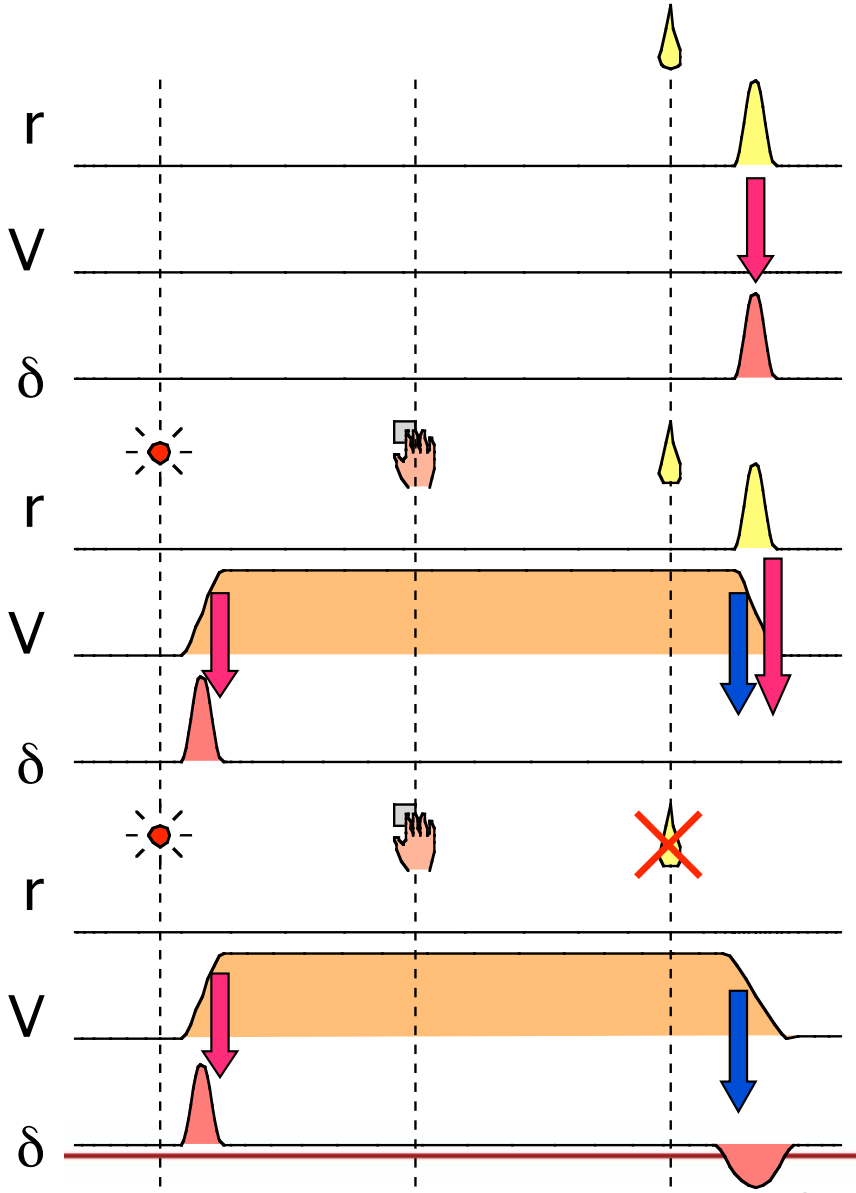
How to tune these parameters?

Dopamine Neurons Code TD Error

$$\delta(t) = r(t) + \gamma V(s(t+1)) - V(s(t))$$



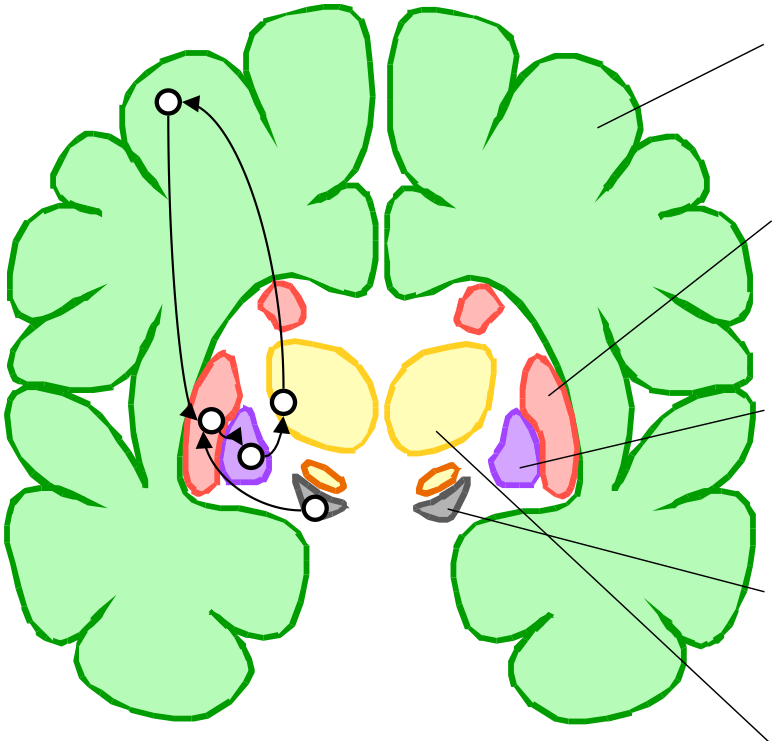




(Schultz et al. 1997)

Basal Ganglia for Reinforcement Learning?

(Doya 2000, 2007)



Cerebral cortex

state/action coding

Striatum

reward prediction

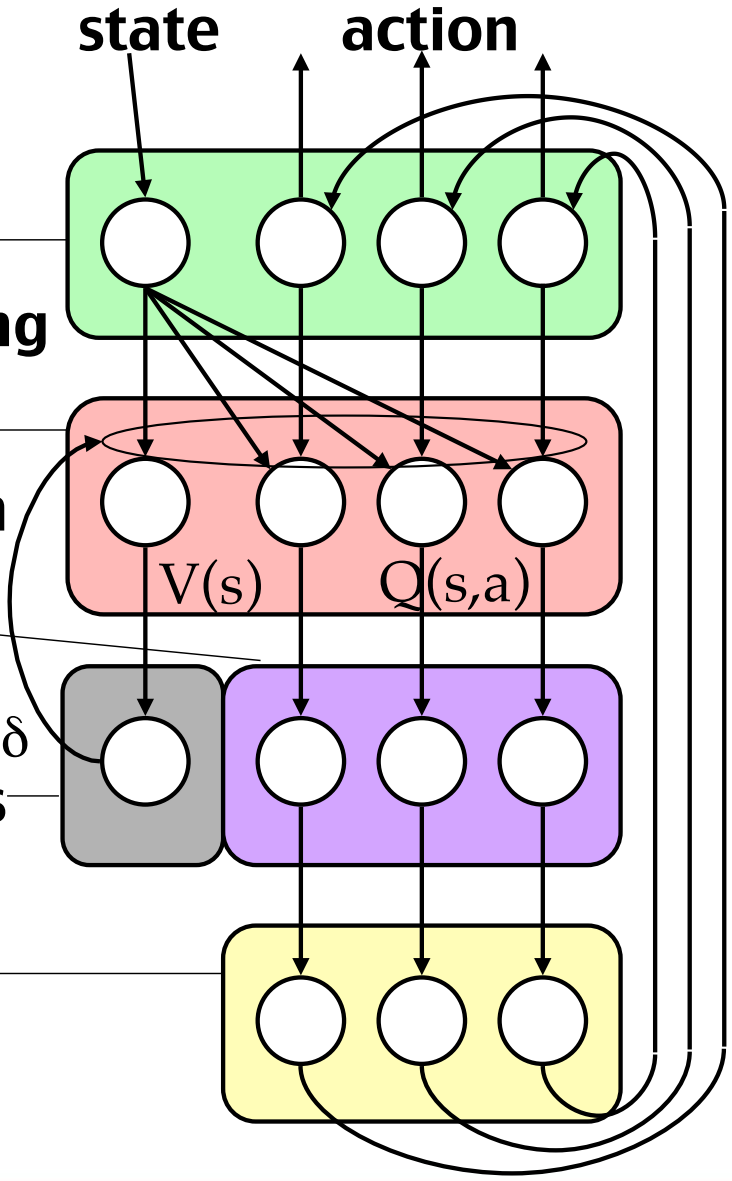
Pallidum

action selection

Dopamine neurons

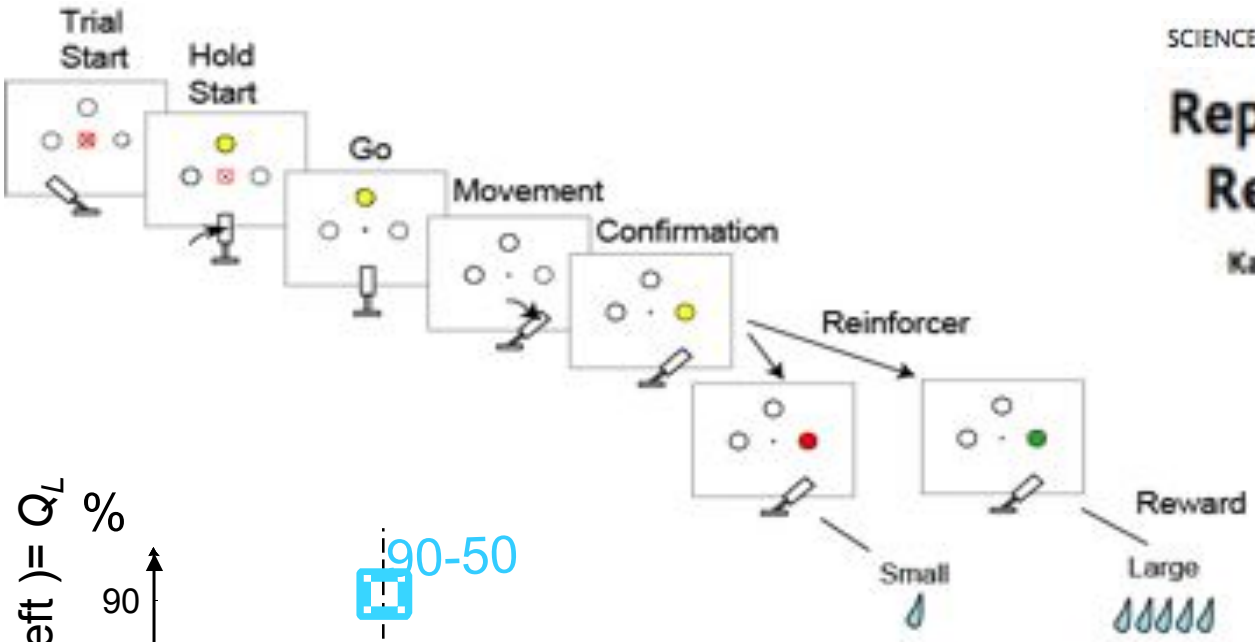
TD signal

Thalamus



Monkey Free Choice Task

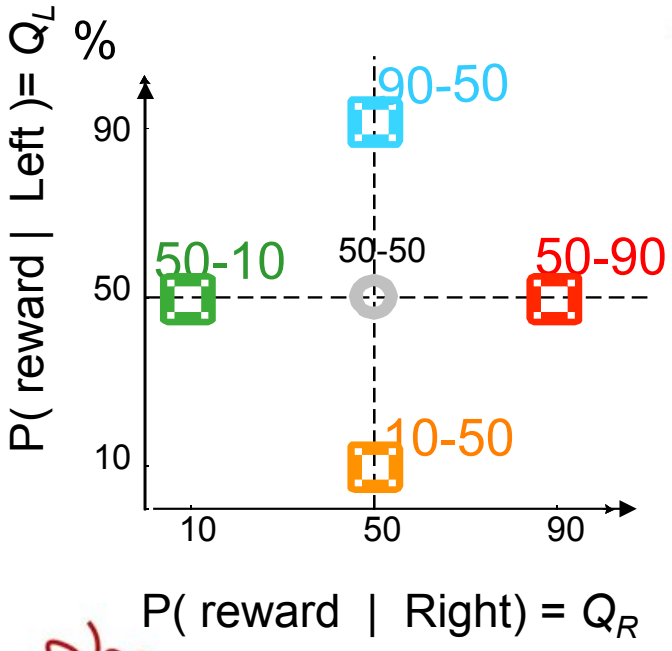
(Samejima et al., 2005)

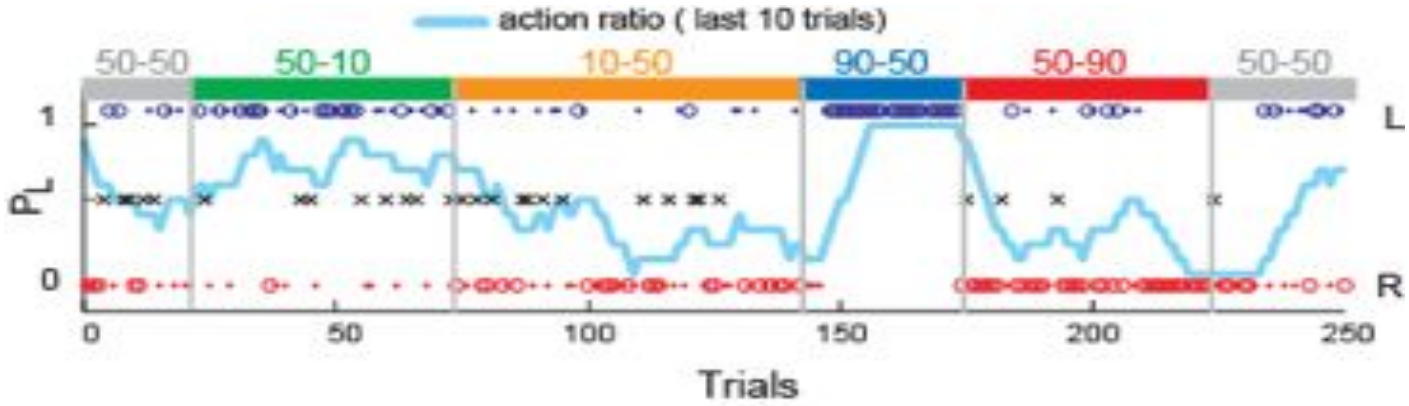


SCIENCE VOL 310 25 NOVEMBER 2005 1337

Representation of Action-Specific Reward Values in the Striatum

Kazuyuki Samejima,[1*†] Yasumasa Ueda,[2] Kenji Doya,[1,3] Minoru Kimura[2*]

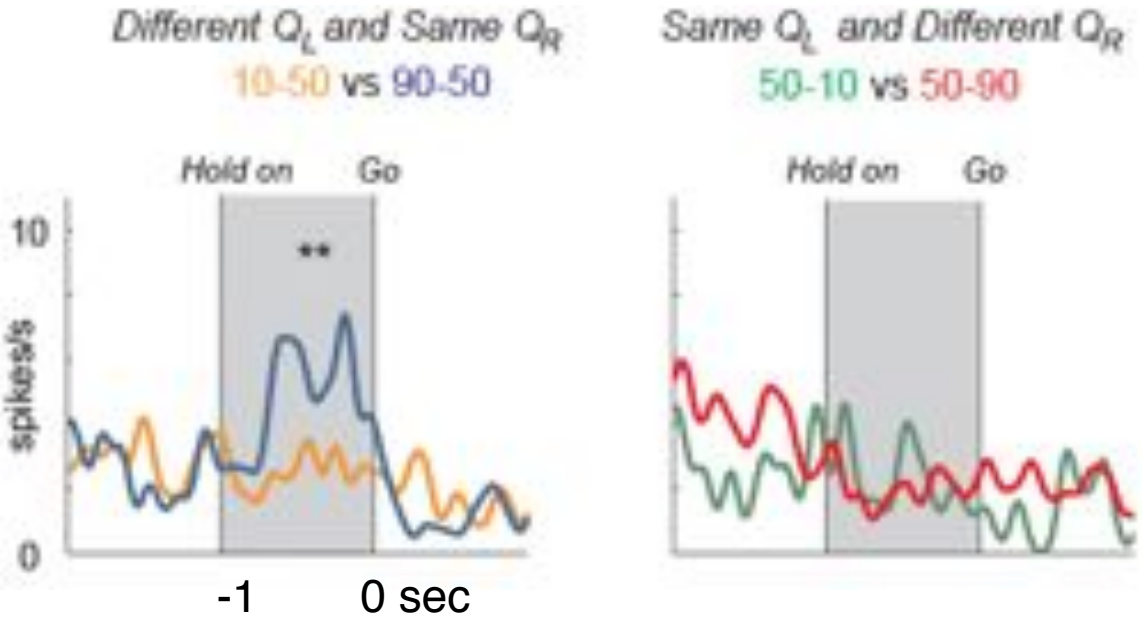




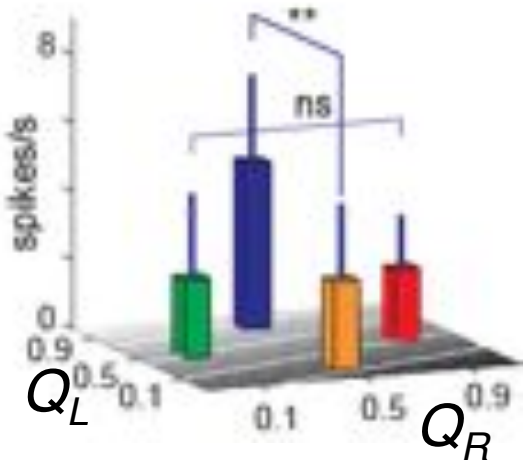
Dissociation of action and reward

Action Value Coding in Striatum

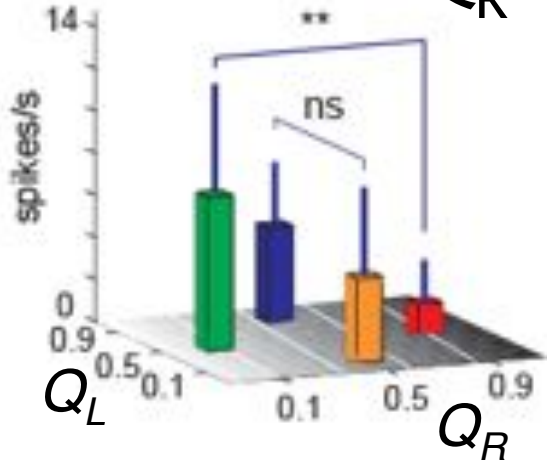
(Samejima et al., 2005)



Q_L neuron



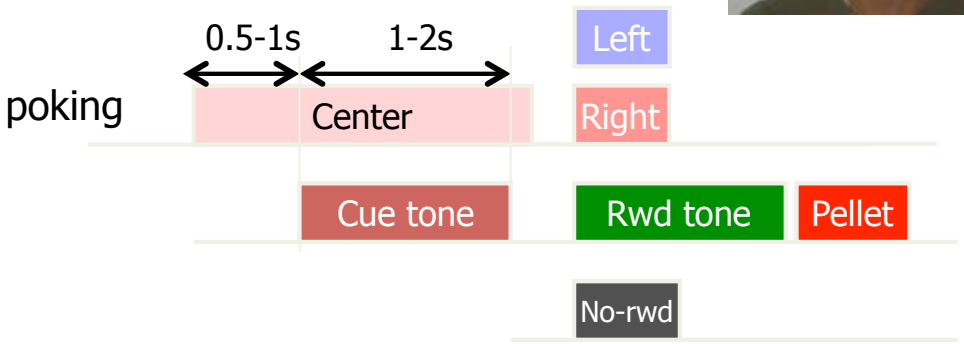
$-Q_R$ neuron



Forced and Free Choice Task

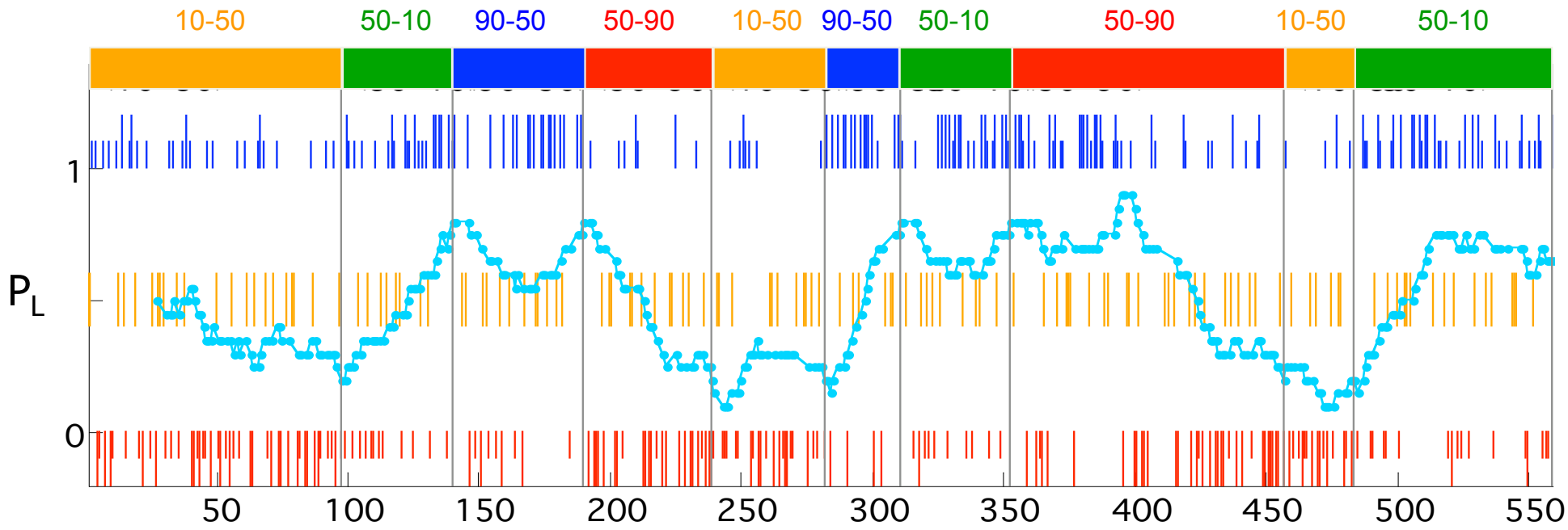
Makoto Ito

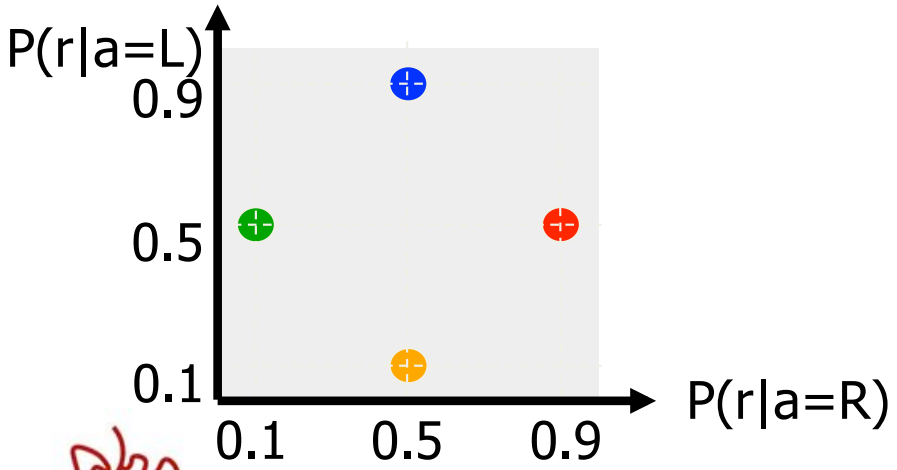
Left Center Right



Cue tone	Reward prob. (L, R)
Left tone (900Hz)	Fixed (50%,0%)
Right tone (6500Hz)	Fixed (0%, 50%)
Free-choice tone (White noise)	Varied (90%, 50%) (50%, 90%) (50%, 10%) (10%, 50%)

Time Course of Choice





Left for tone A

Right for tone A

Tone B

Generalized Q-learning Model

(Ito & Doya, 2009)

- Action selection

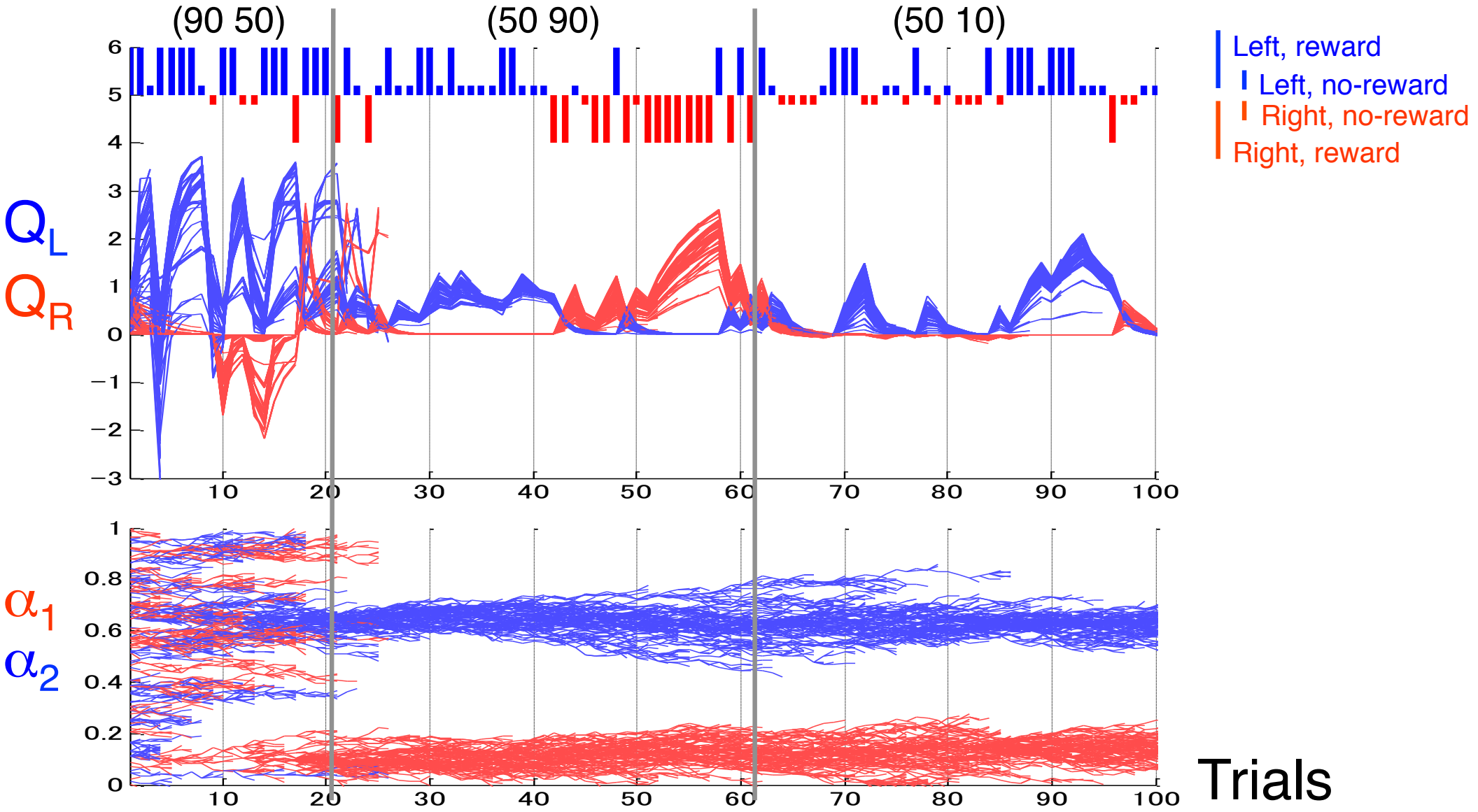
$$P(a(t)=L) = \exp Q_L(t)/(\exp Q_L(t)+\exp Q_R(t))$$

- Action value update: $i \in \{L,R\}$

$$Q_i(t+1) = \begin{cases} (1-\alpha_1)Q_i(t) + \alpha_1\kappa_1 & \text{if } a(t)=i,\ r(t)=1 \\ (1-\alpha_1)Q_i(t) - \alpha_1\kappa_2 & \text{if } a(t)=i,\ r(t)=0 \\ (1-\alpha_2)Q_i(t) & \text{if } a(t)\neq i,\ r(t)=1 \\ (1-\alpha_2)Q_i(t) & \text{if } a(t)\neq i,\ r(t)=0 \end{cases}$$

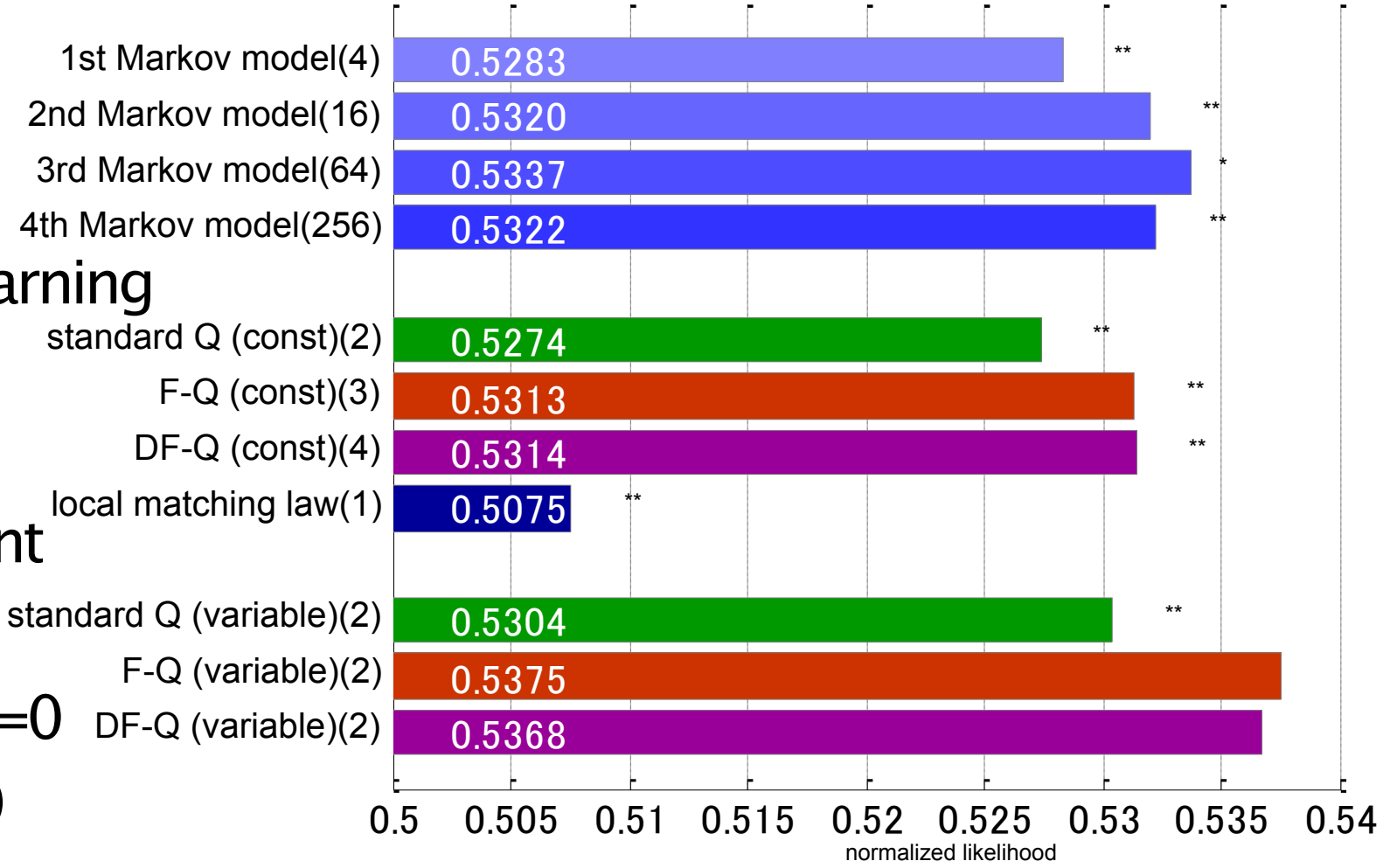
- Parameters
  - α_1: learning rate
  - α_2: forgetting rate
  - κ_1: reward reinforcement
  - κ_2: no-reward aversion

Model Fitting by Particle Filter

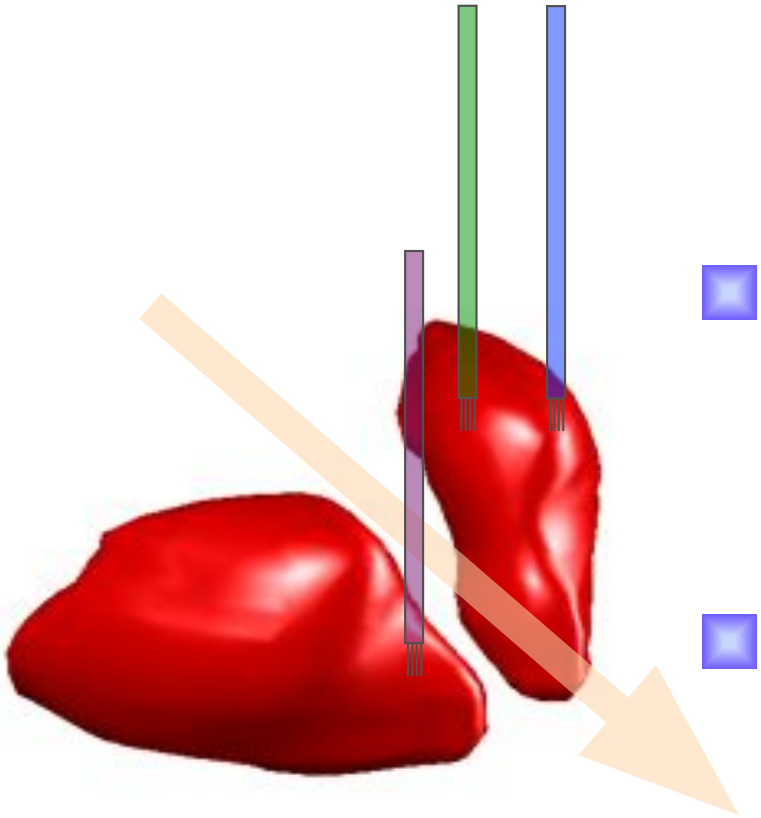


Model Fitting

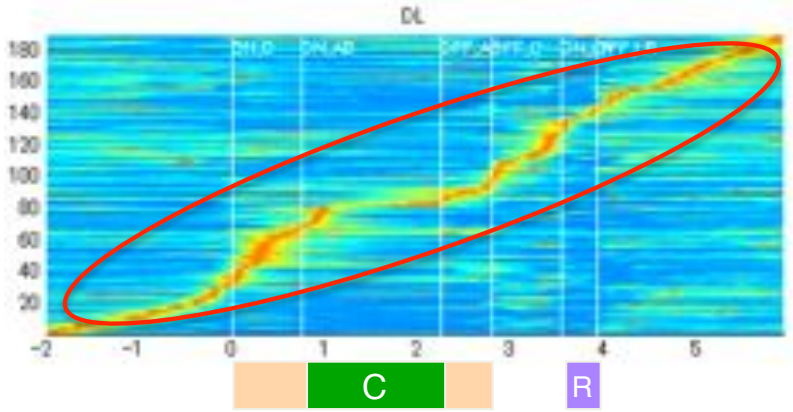
- Generalized Q learning
  - α_1: learning
  - α_2: forgetting
  - κ_1: reinforcement
  - κ_2: aversion
  - standard: $\alpha_2=\kappa_2=0$
  - forgetting: $\kappa_2=0$

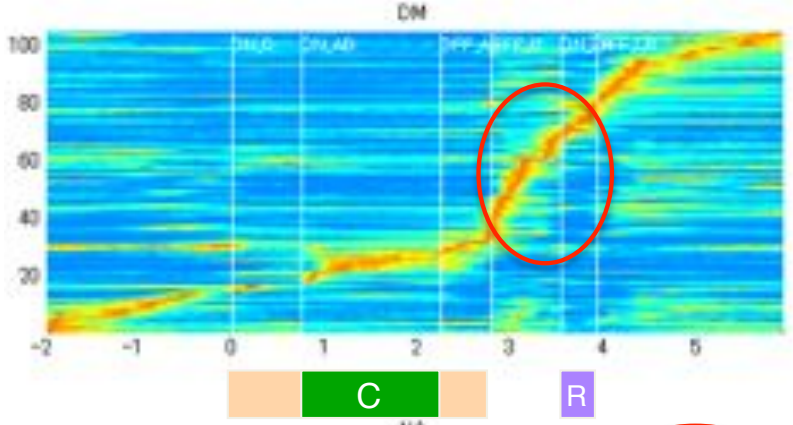


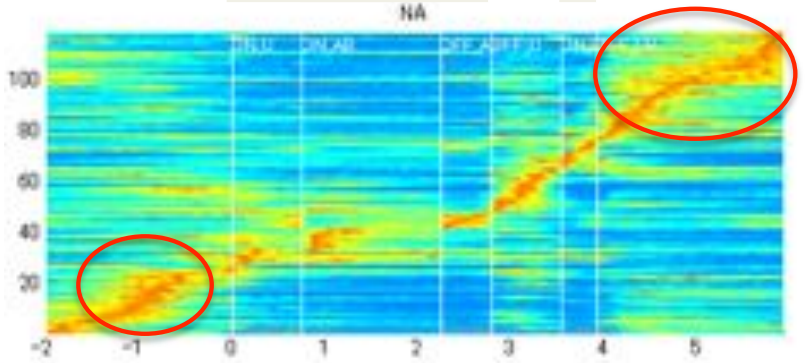
Neural Activity in the Striatum



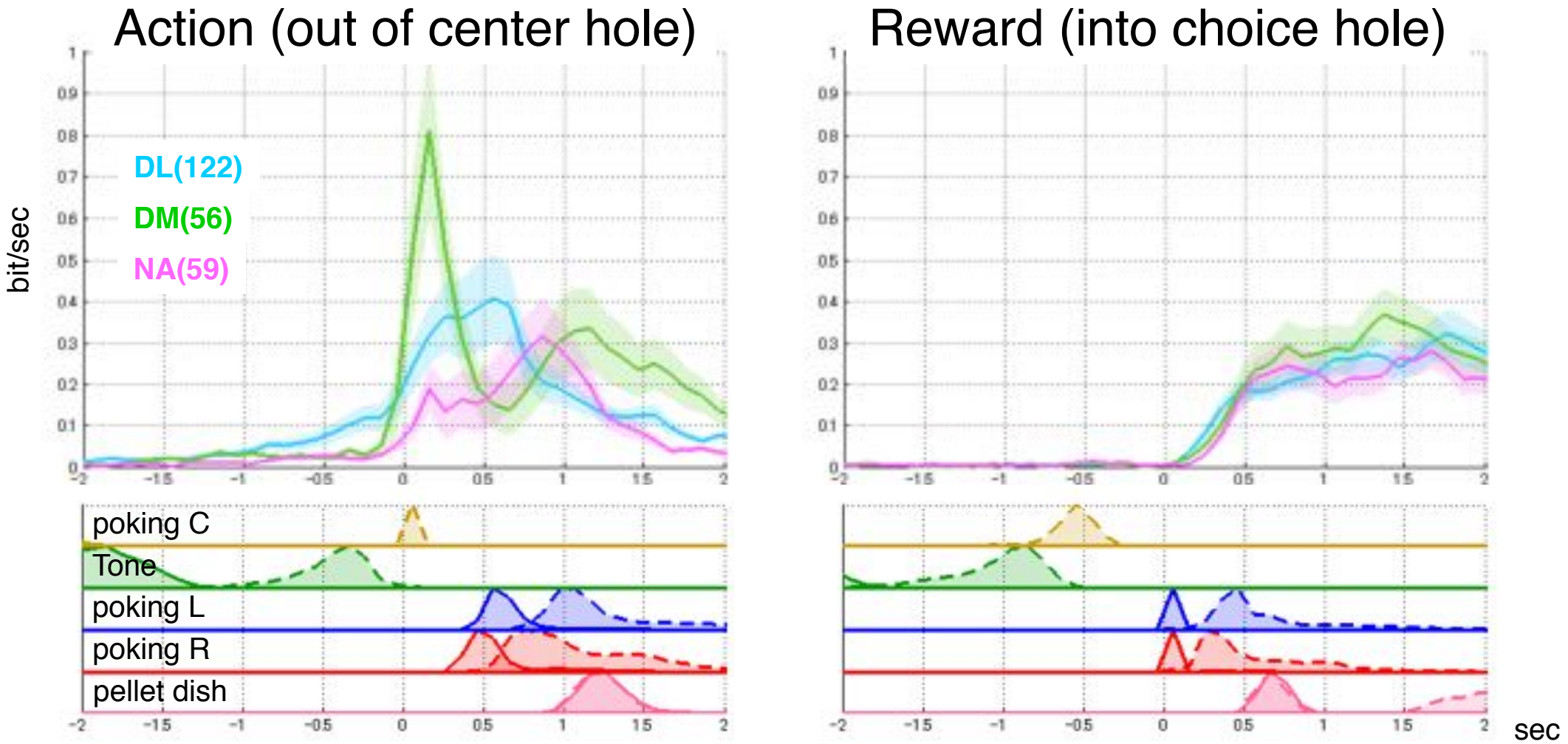
- Dorsolateral
- Dorsomedial
- Ventral



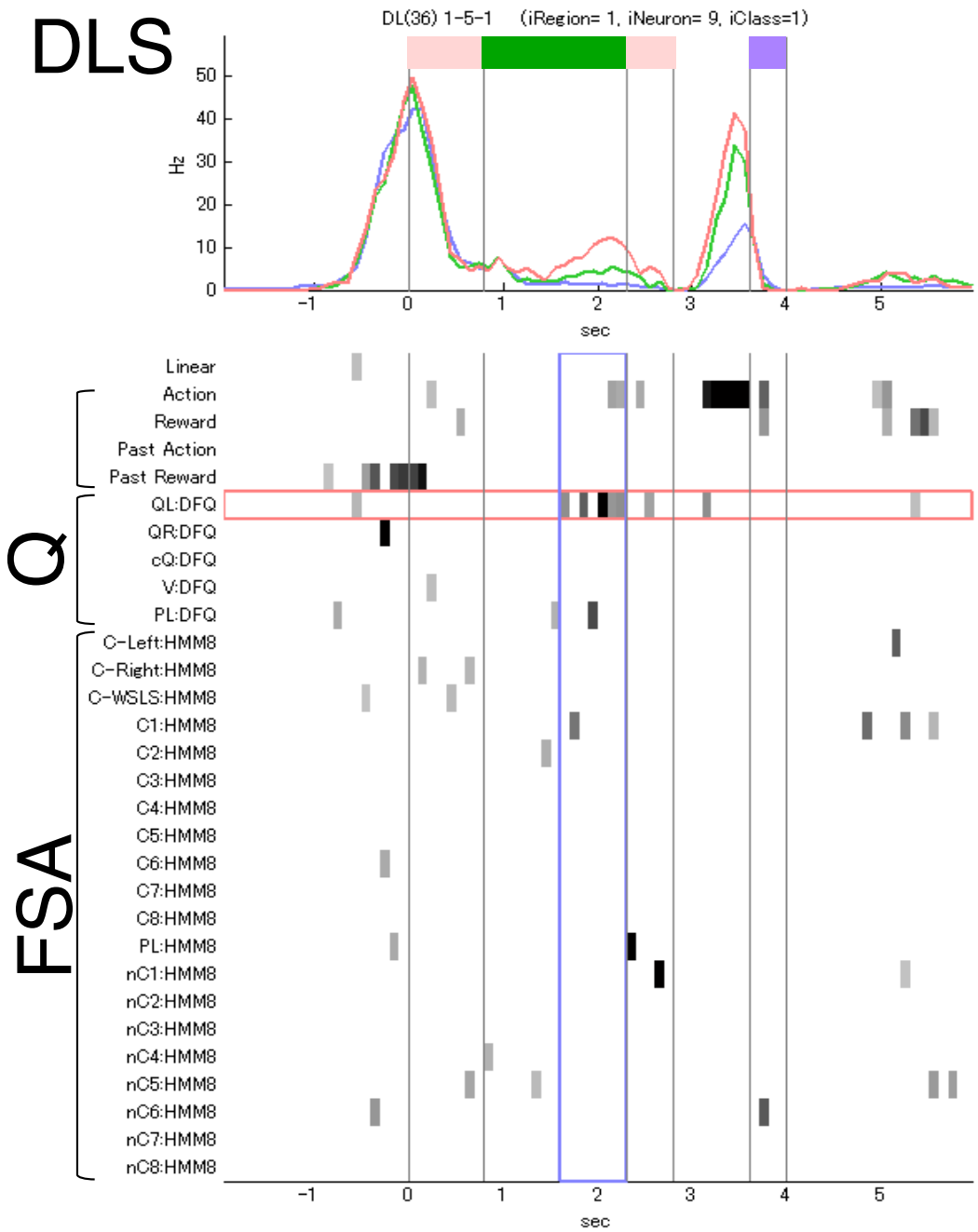


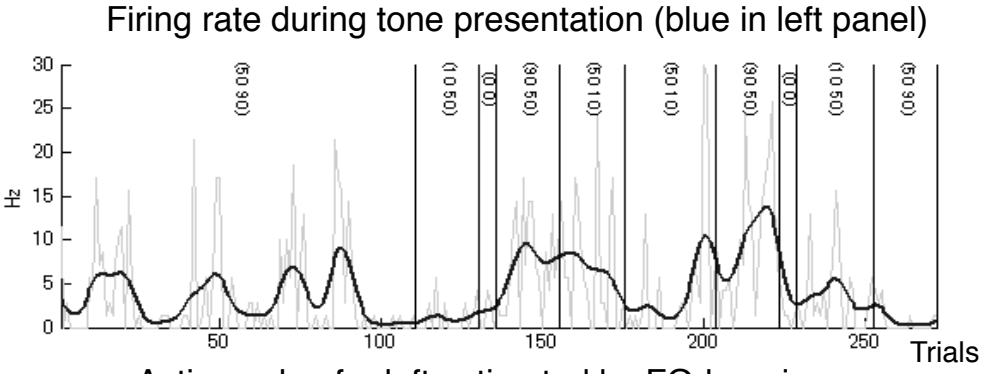


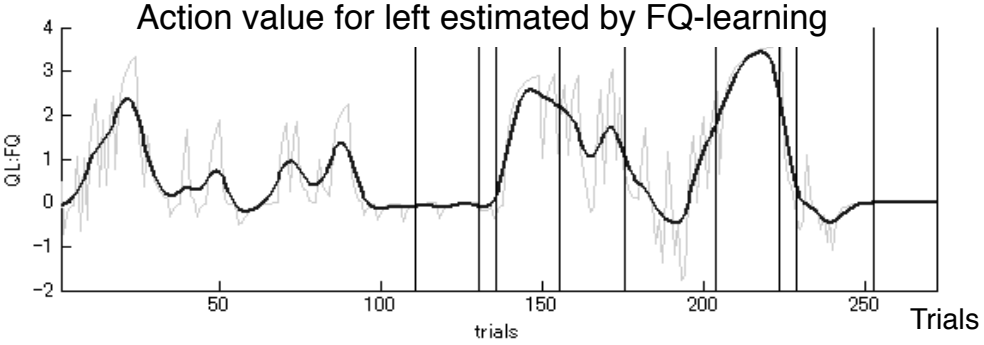
Information of Action and Reward

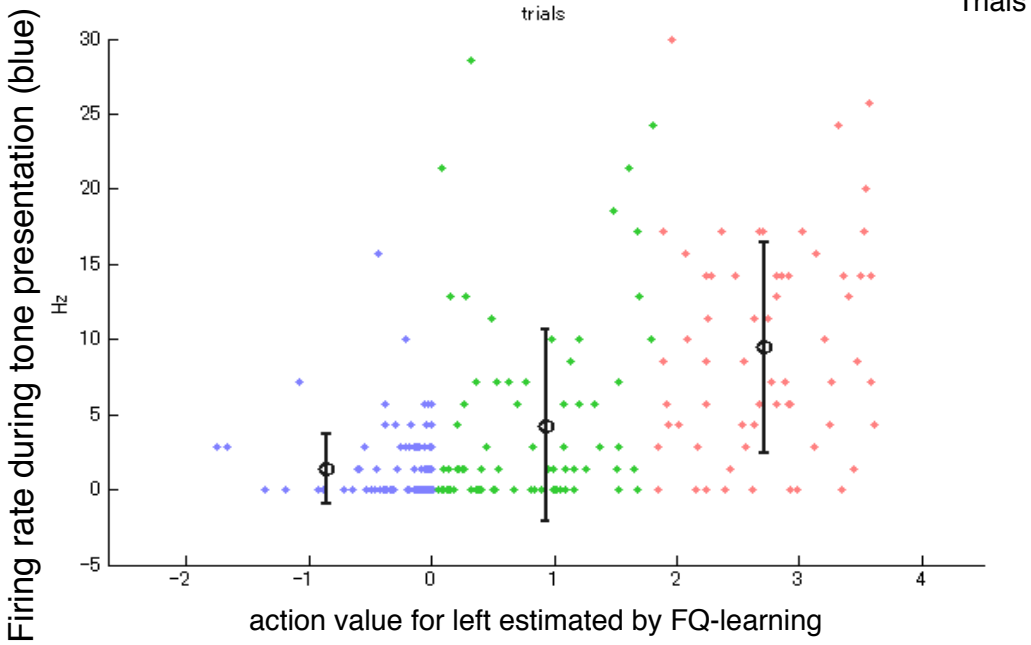


Action value coded by a DLS neuron

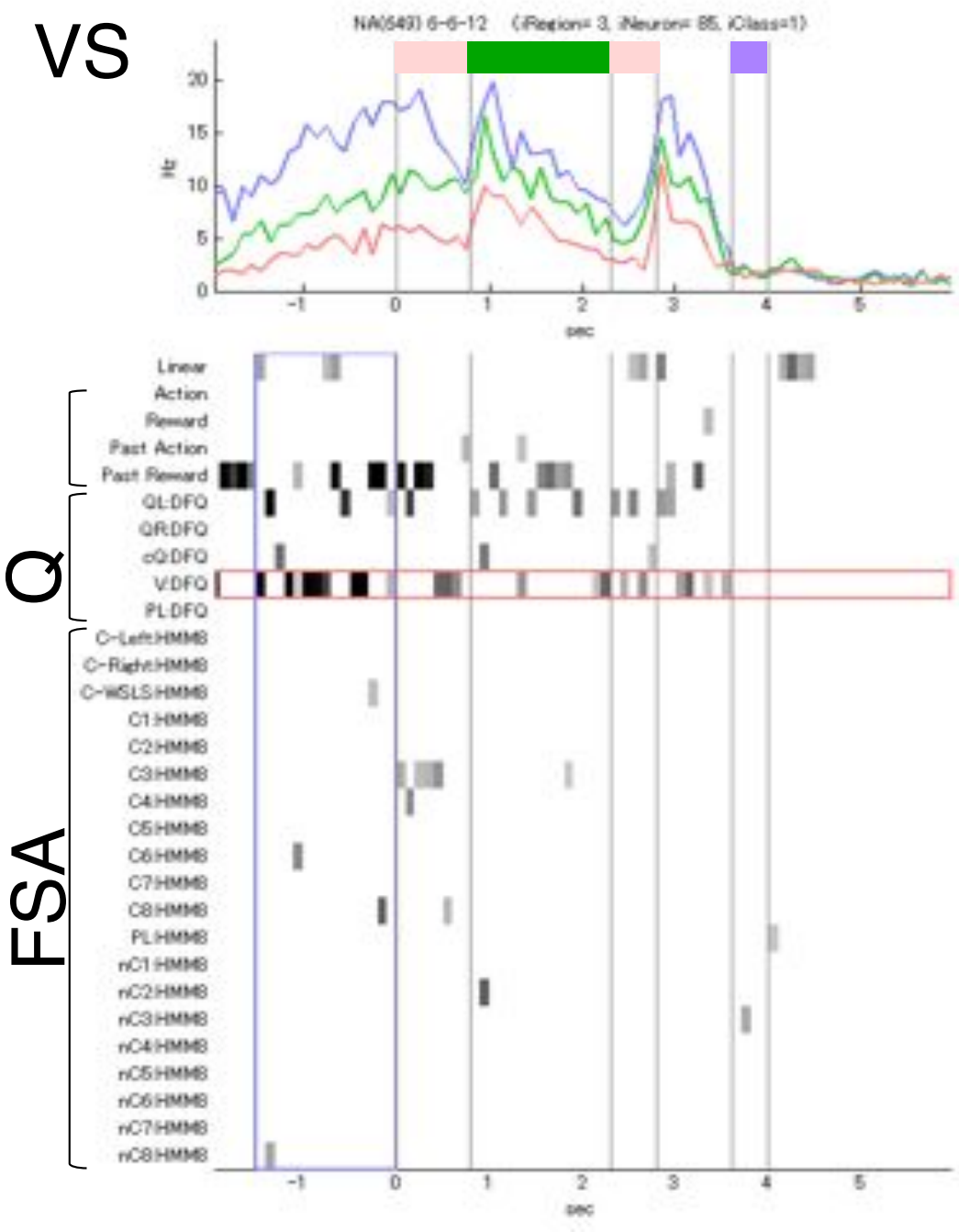


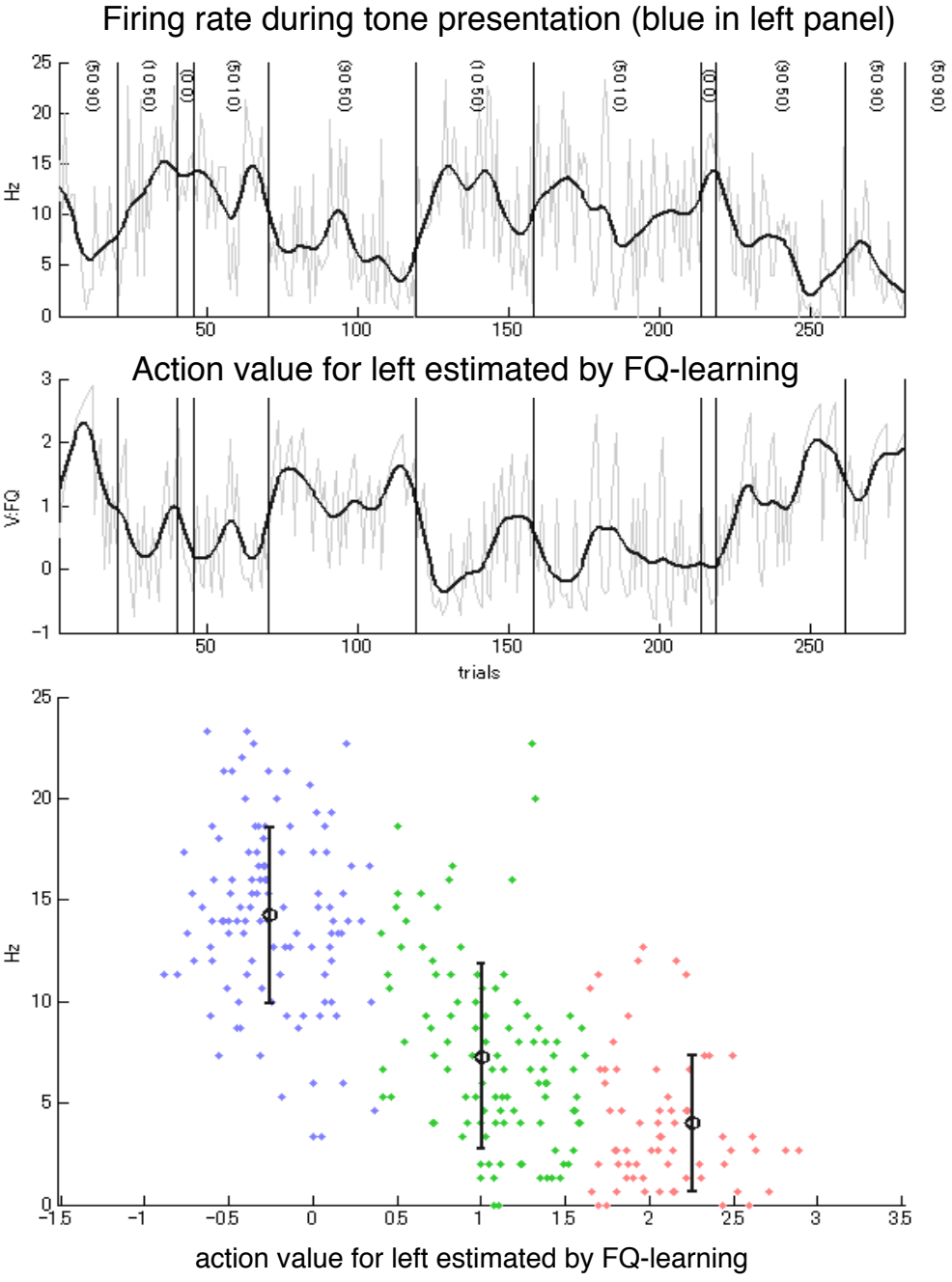






State value coded by a VS neuron





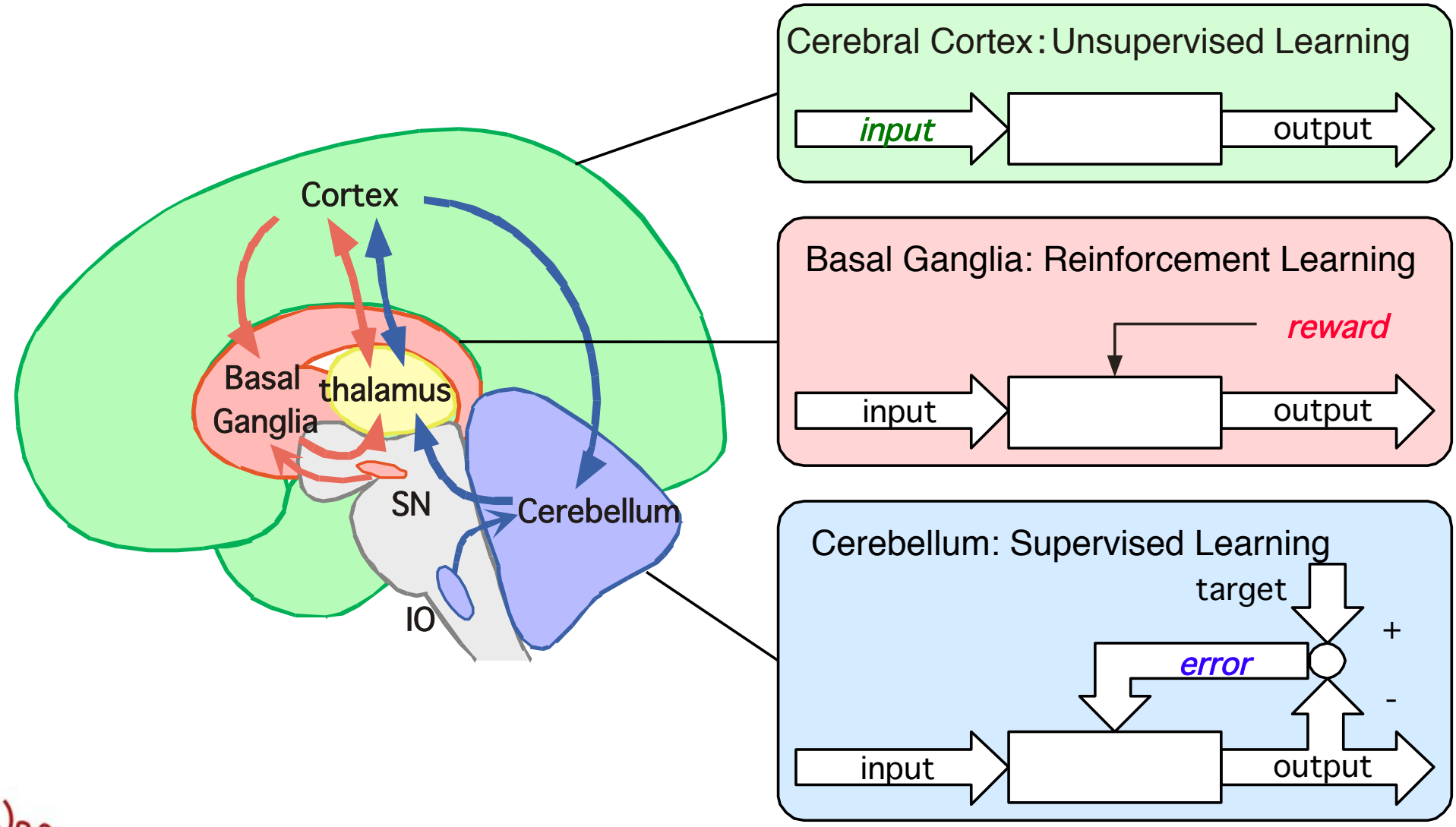
Hierarchy in Cortico-Striatal Network

(Voorn et al., 2004)

- Dorsolateral striatum – motor
  - early action coding
  - what action to take?
- Dorsomedial striatum - frontal
  - action value
  - in what context?
- Ventral striatum - limbic
  - state value
  - whether worth doing?

Specialization by Learning Algorithms

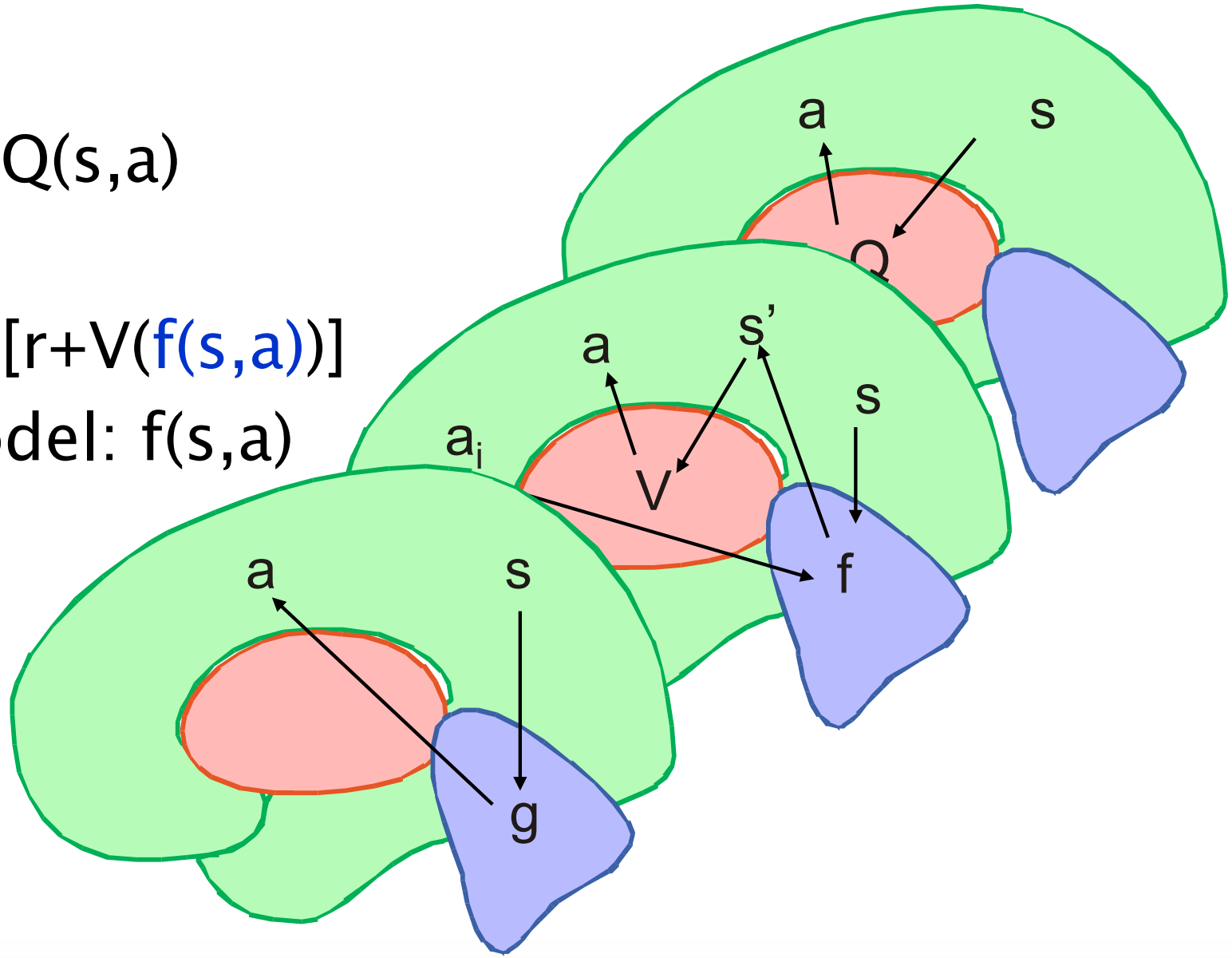
(Doya, 1999)



Multiple Action Selection Schemes

- Model-free
  - $a = \text{argmax}_a\ Q(s,a)$
- Model-based
  - $a = \text{argmax}_a\ [r+V(f(s,a))]$

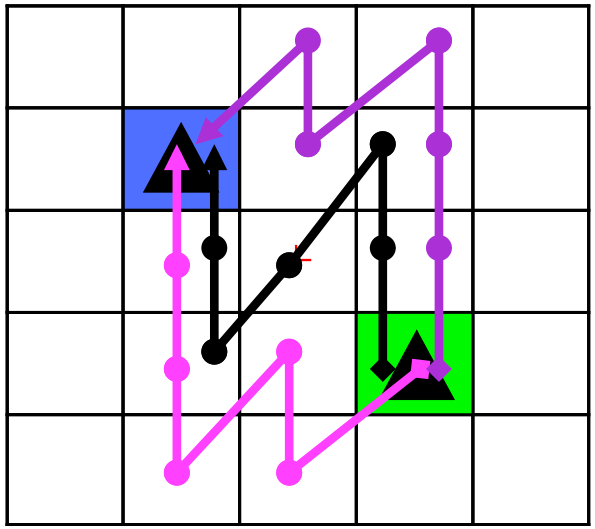
    forward model: f(s,a)
- Lookup table
  - $a = g(s)$

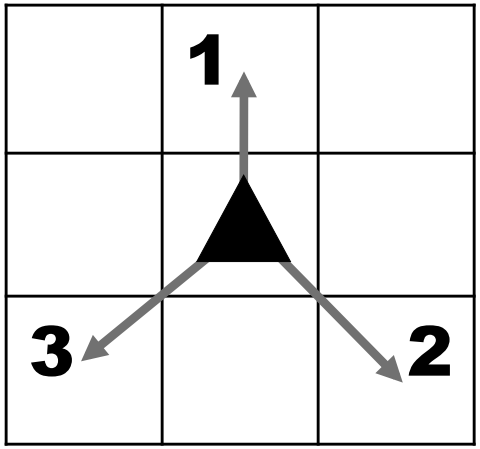


‘Grid Sailing’ Task

Alan Fermin

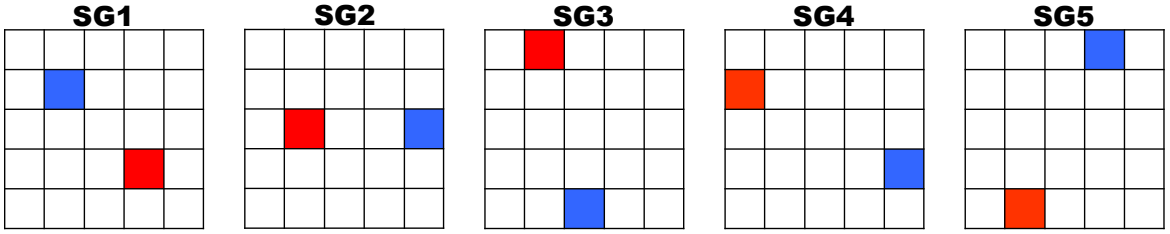
- Move a cursor to the goal
  - 100 points for shortest path
  - -5 points per excess steps
- Keymap
  - only 3 directions
  - non-trivial path planning
- Immediate or delayed start
  - 4 to 6 sec for planning
  - timeout in 6 sec



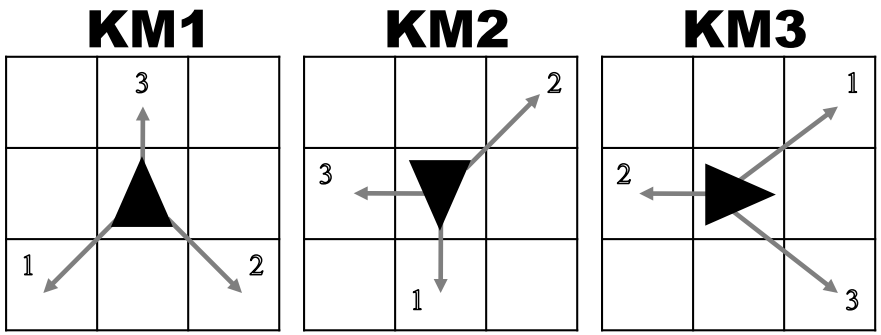


Task Conditions

Start-Goal Positions



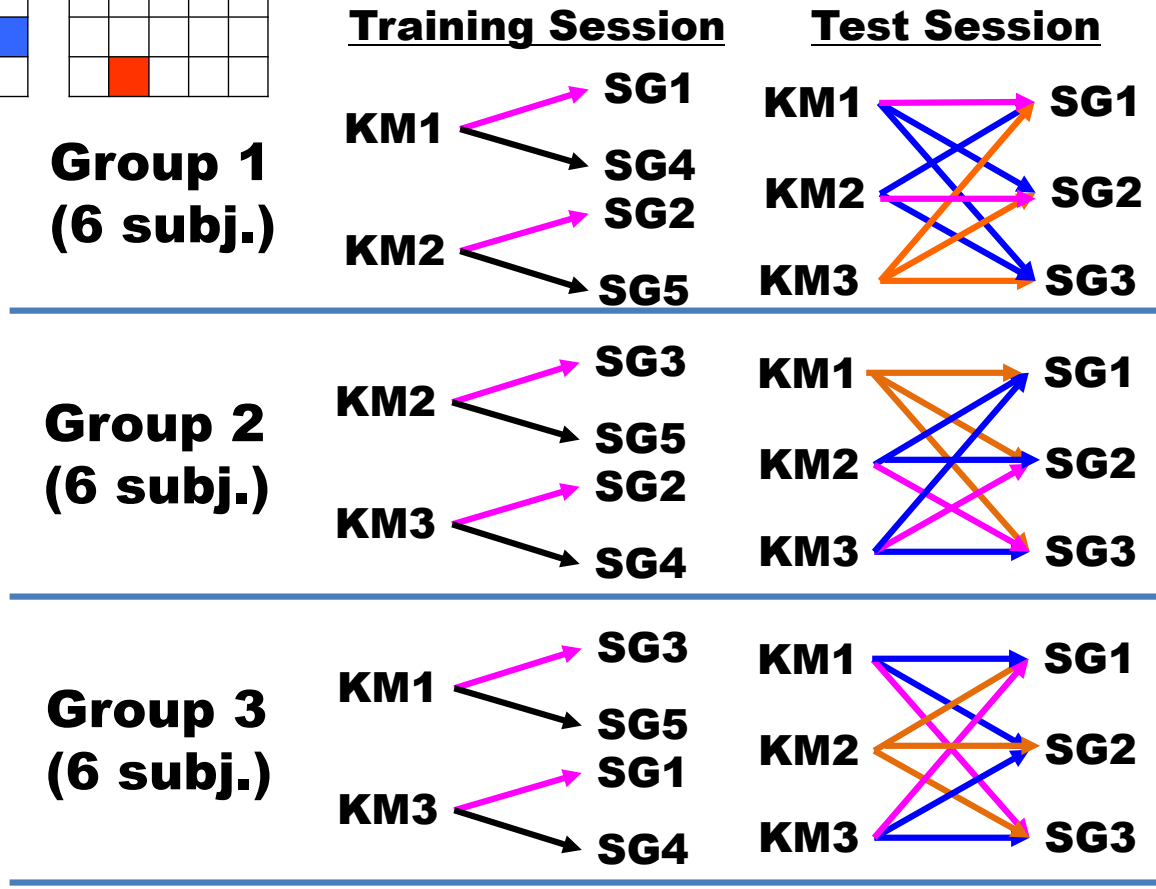
Key-Maps



Test Conditions

- Cond 1: New Key-Map
- Cond 2: Learned Key-Maps
- Cond 3: Learned KM-SG

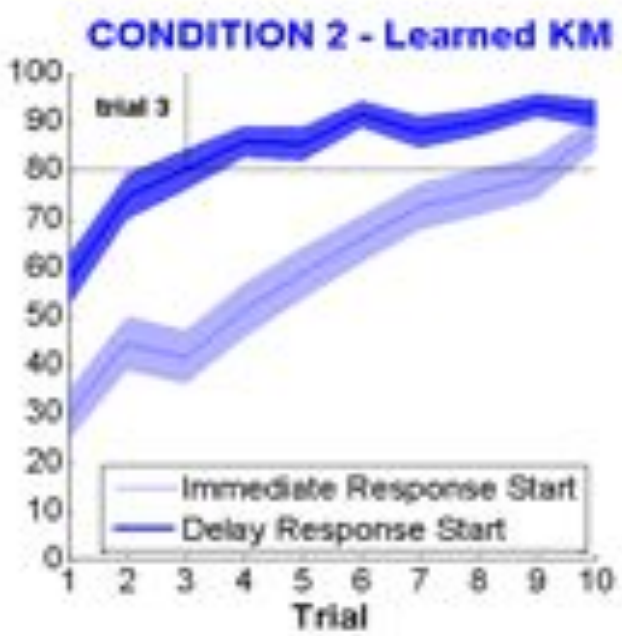
KM-SG combinations



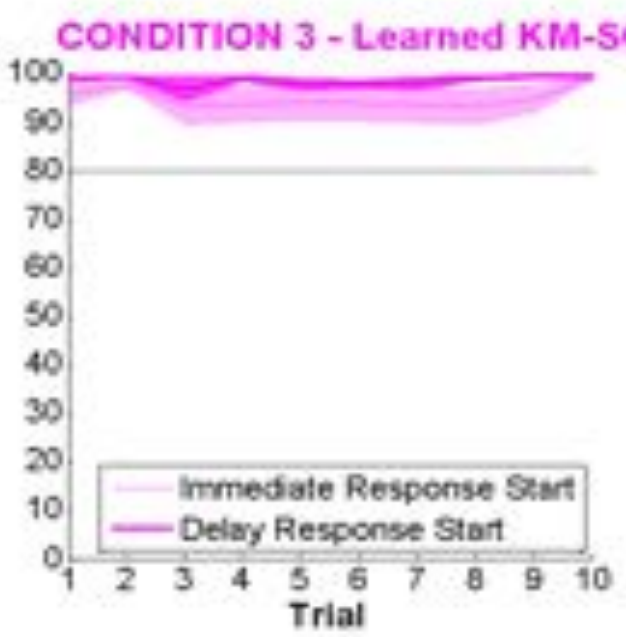
Effect of Pre-start Delay Time

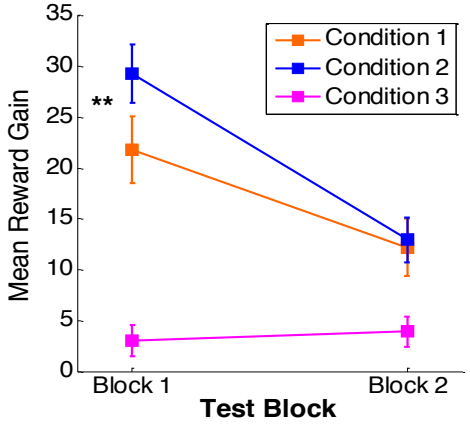
New

Learned key-map

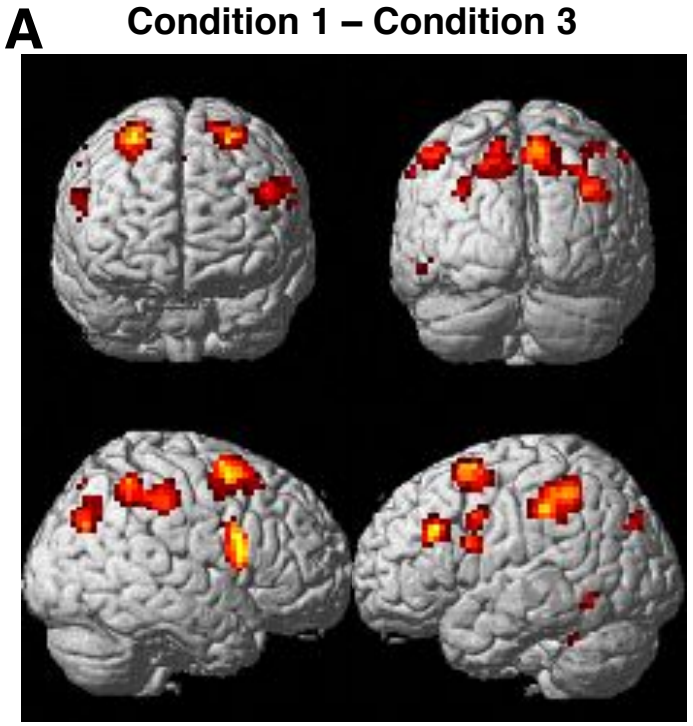


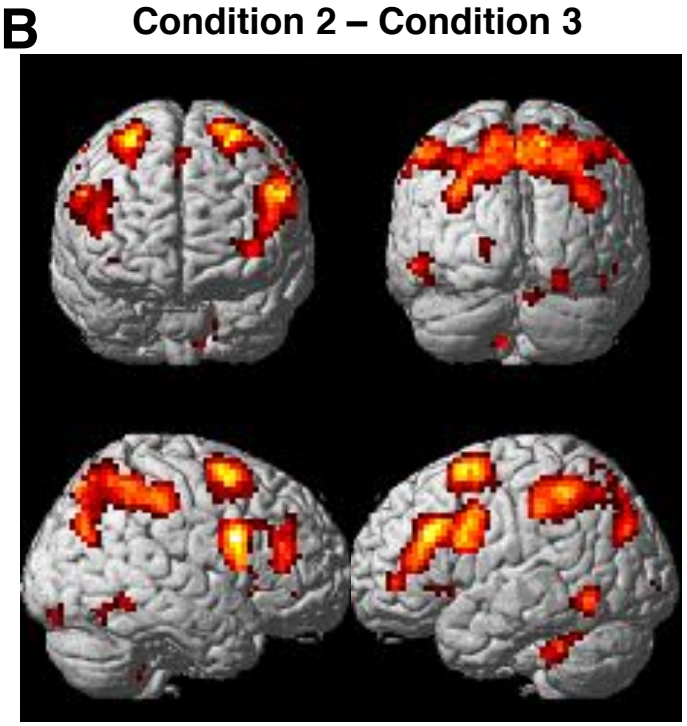
Learned

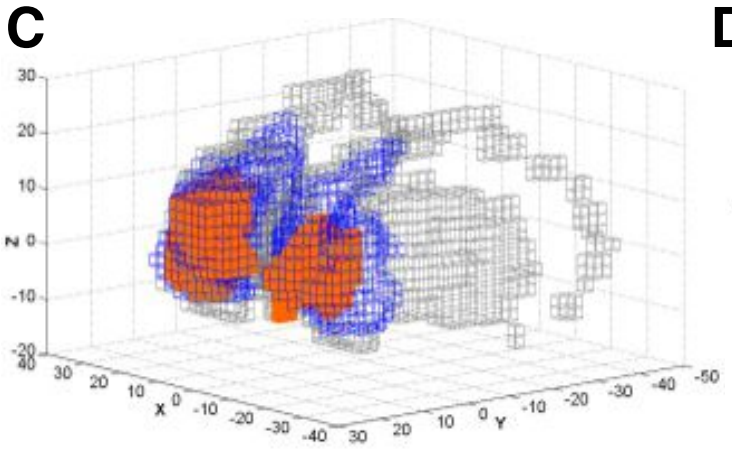


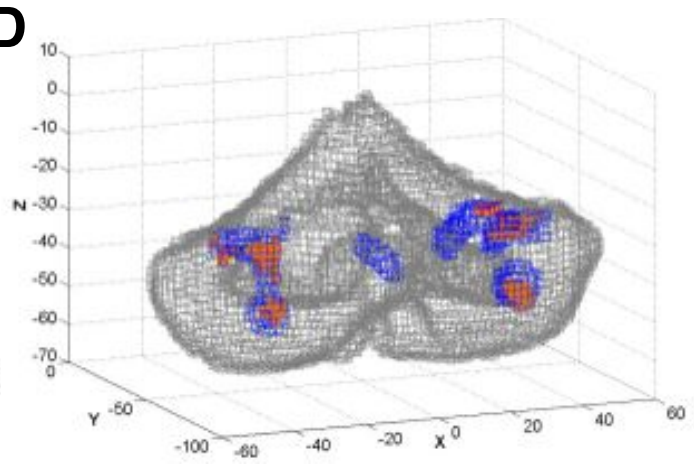


Delay Period Activity







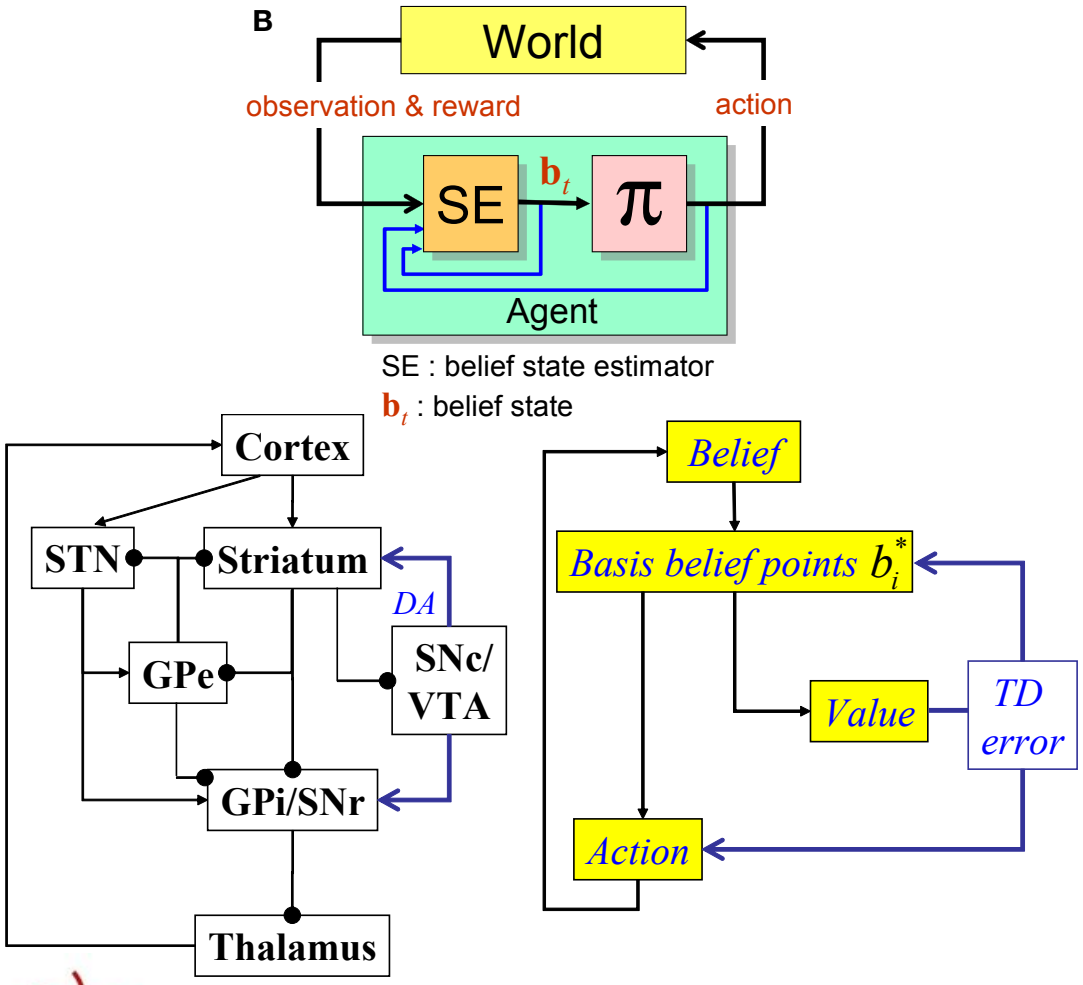


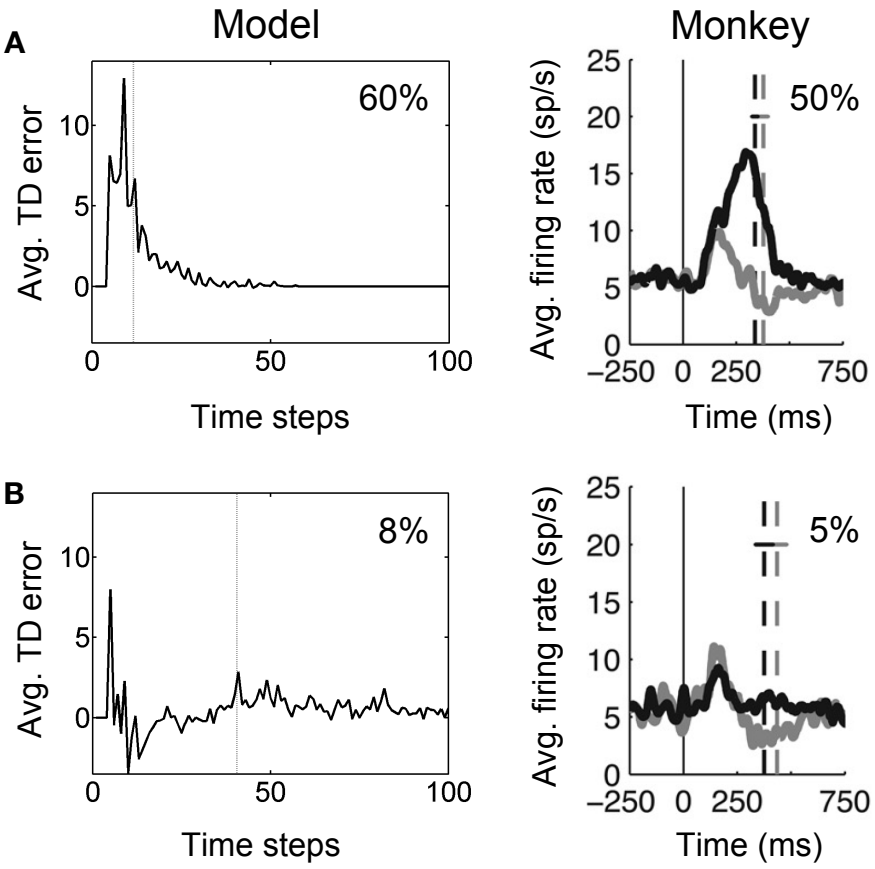
- Condition 2
  - DLPFC
  - PMC
  - parietal
  - anterior striatum
  - lateral cerebellum

POMDP by Cortex-Basal Ganglia

Rao (2011)

- Belief state update by the cortex

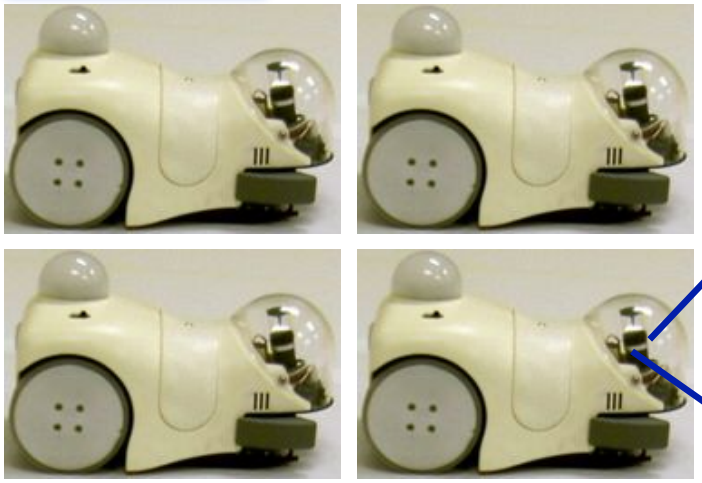




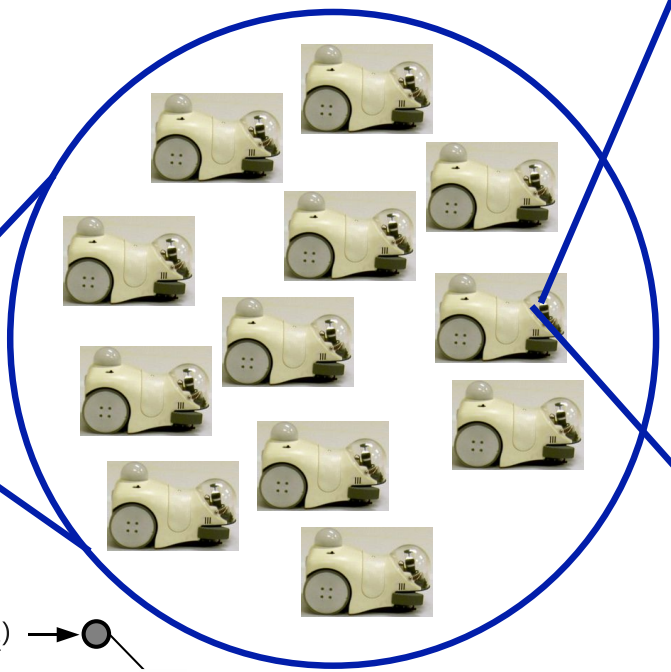
Embodied Evolution (Elfwing et al., 2011)

Population

Robots



Virtual agents
15-25



Genes

Weights for top layer NN

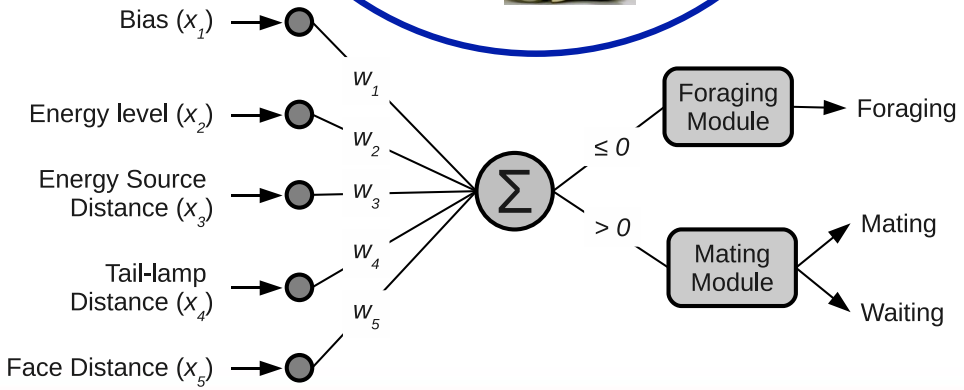
$w_1, w_2, \ldots, w_n$

Weights shaping rewards

$v_1, v_2, \ldots, v_n$

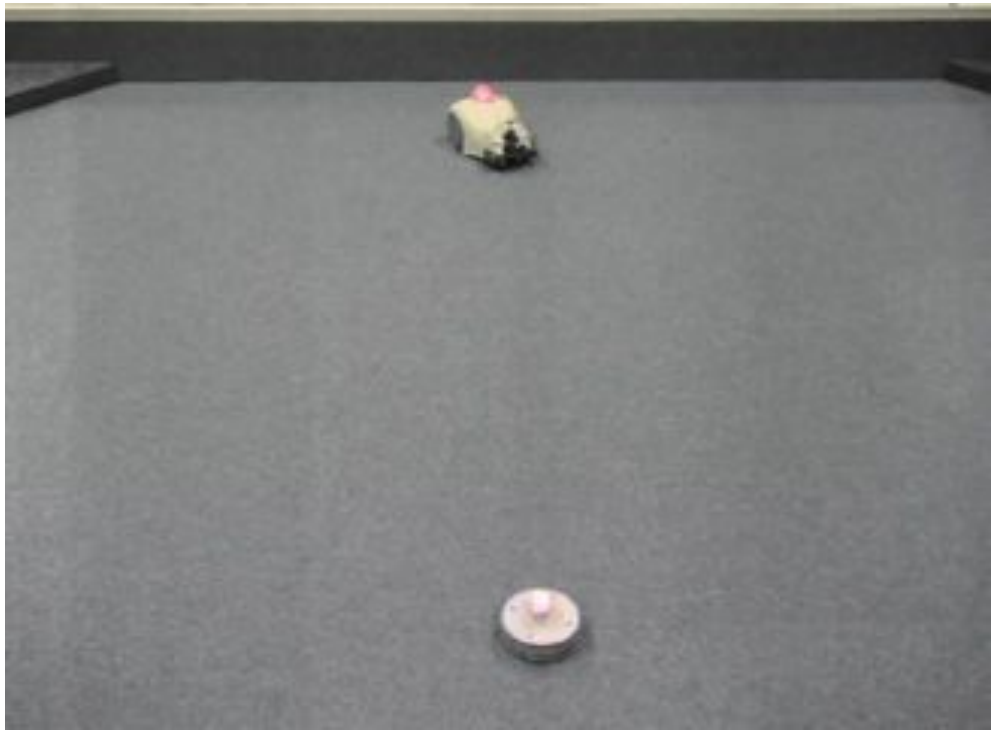
Meta-parameters

$\alpha\gamma\lambda\tau_k\tau_0$

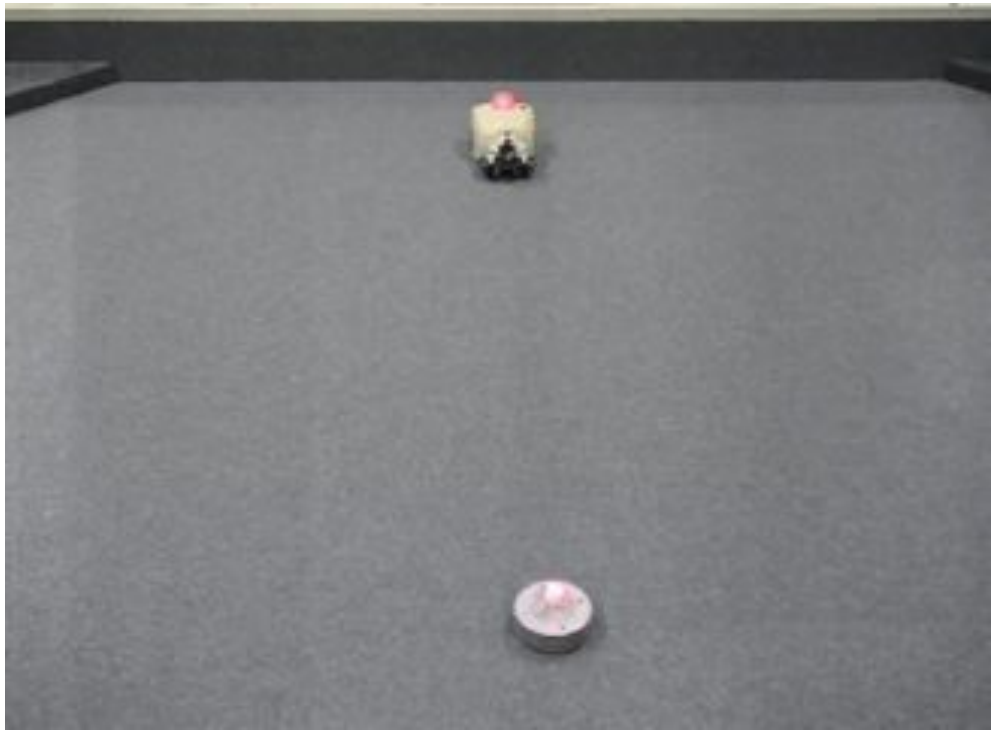


Temporal Discount Factor γ

- Large γ
  - reach for far reward



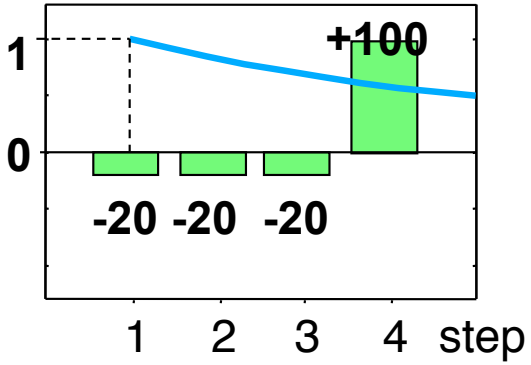
- Small γ
  - only to near reward



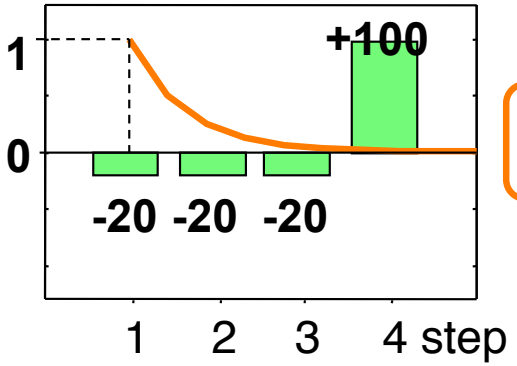
Temporal Discount Factor γ

- $V(t) = E[\, r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \gamma^3 r(t+3) + \ldots]$
  - controls the 'character' of an agent

γ large



γ small



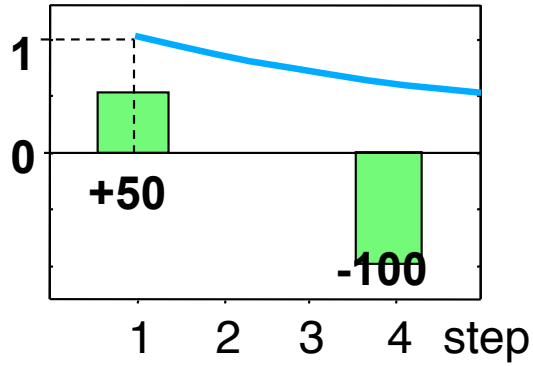
Depression?

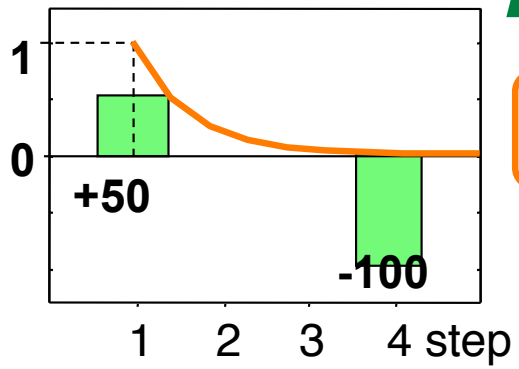
no pain, no gain!

V =18.7

better stay idle

V =-25.1





Impulsivity?

stay away from danger

V = -22.9

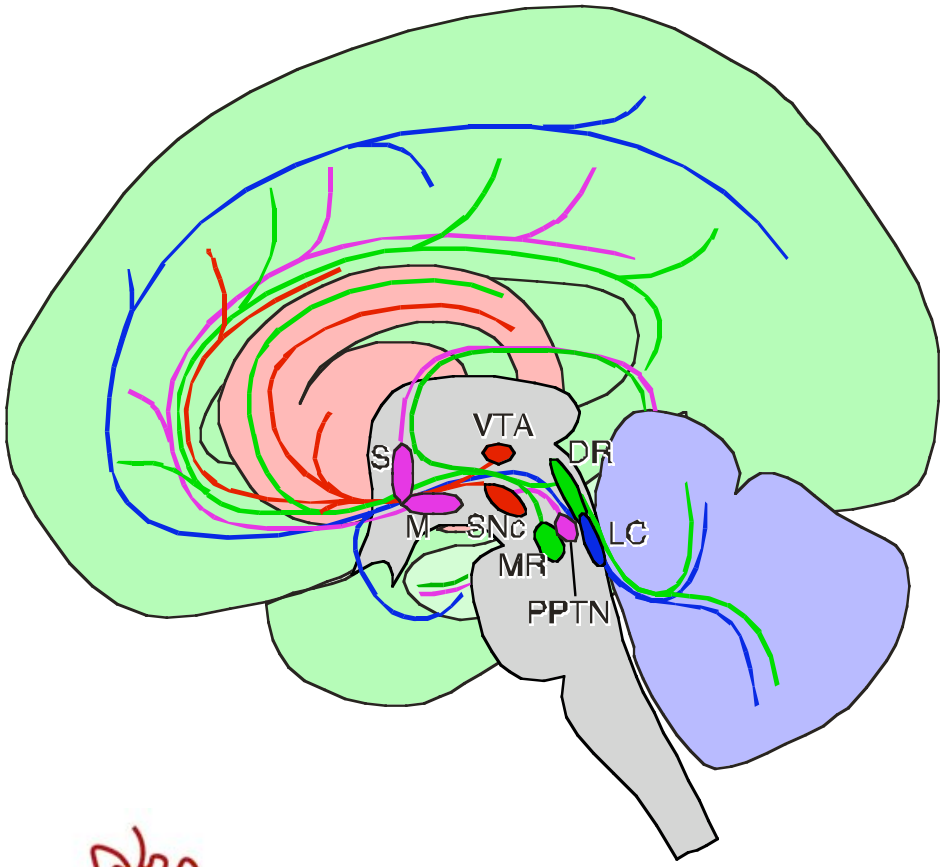
can't resist temptation

V = 47.3

Neuromodulators for Metalearning

(Doya, 2002)

- *Metaparameter* tuning is critical in RL
  - How does the brain tune them?



Dopamine: TD error δ

Acetylcholine: learning rate α

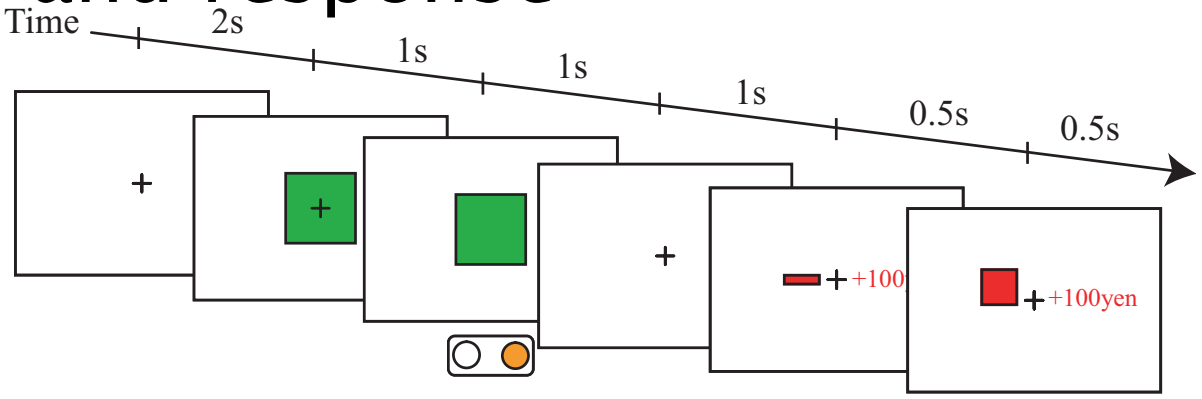
Noradrenaline: exploration β

Serotonin: temporal discount γ

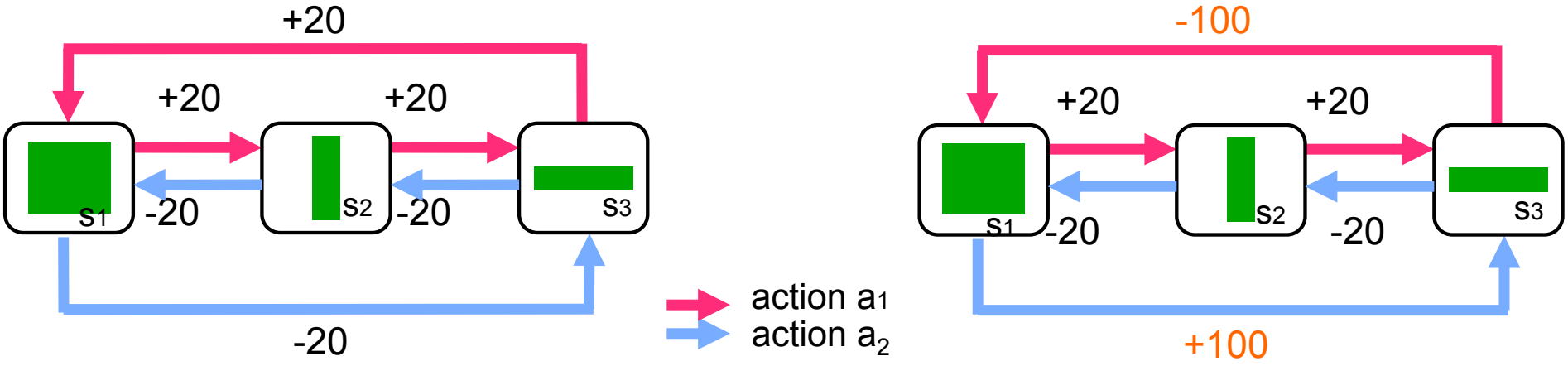
Markov Decision Task

(Tanaka et al., 2004)

- Stimulus and response

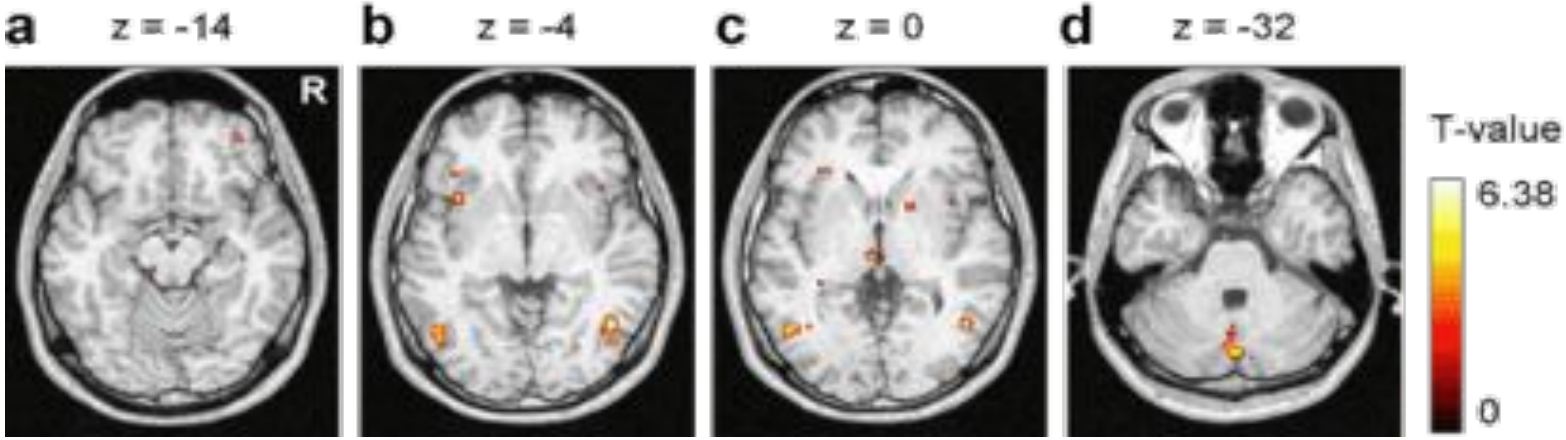


- State transition and reward functions



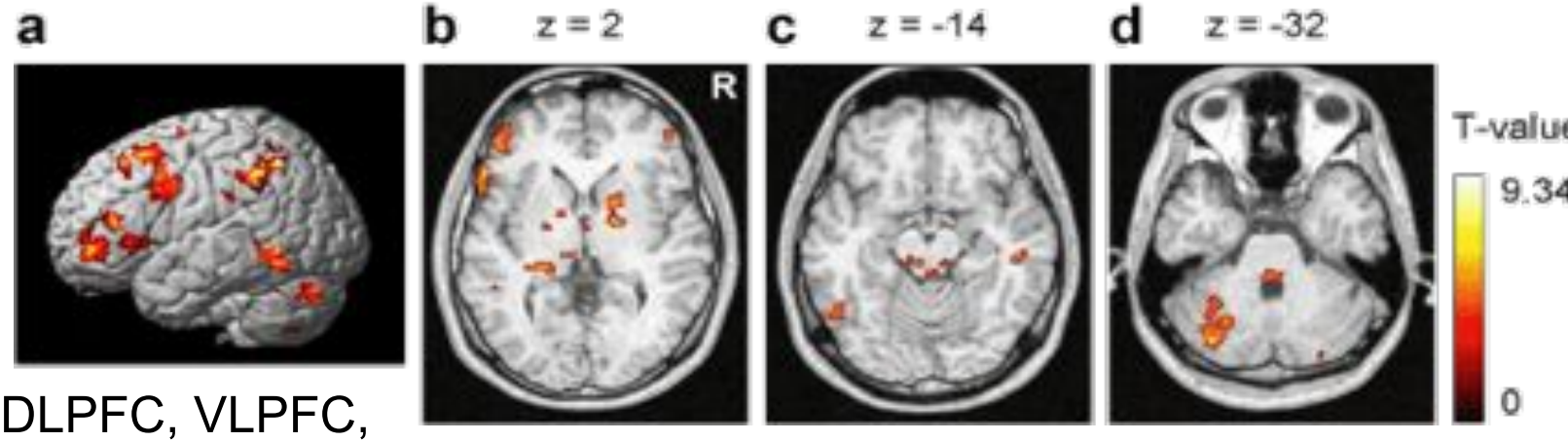
Block-Design Analysis

SHORT vs. NO Reward
(p < 0.001 uncorrected)



OFC Insula Striatum Cerebellum

LONG vs. SHORT
(p < 0.0001 uncorrected)



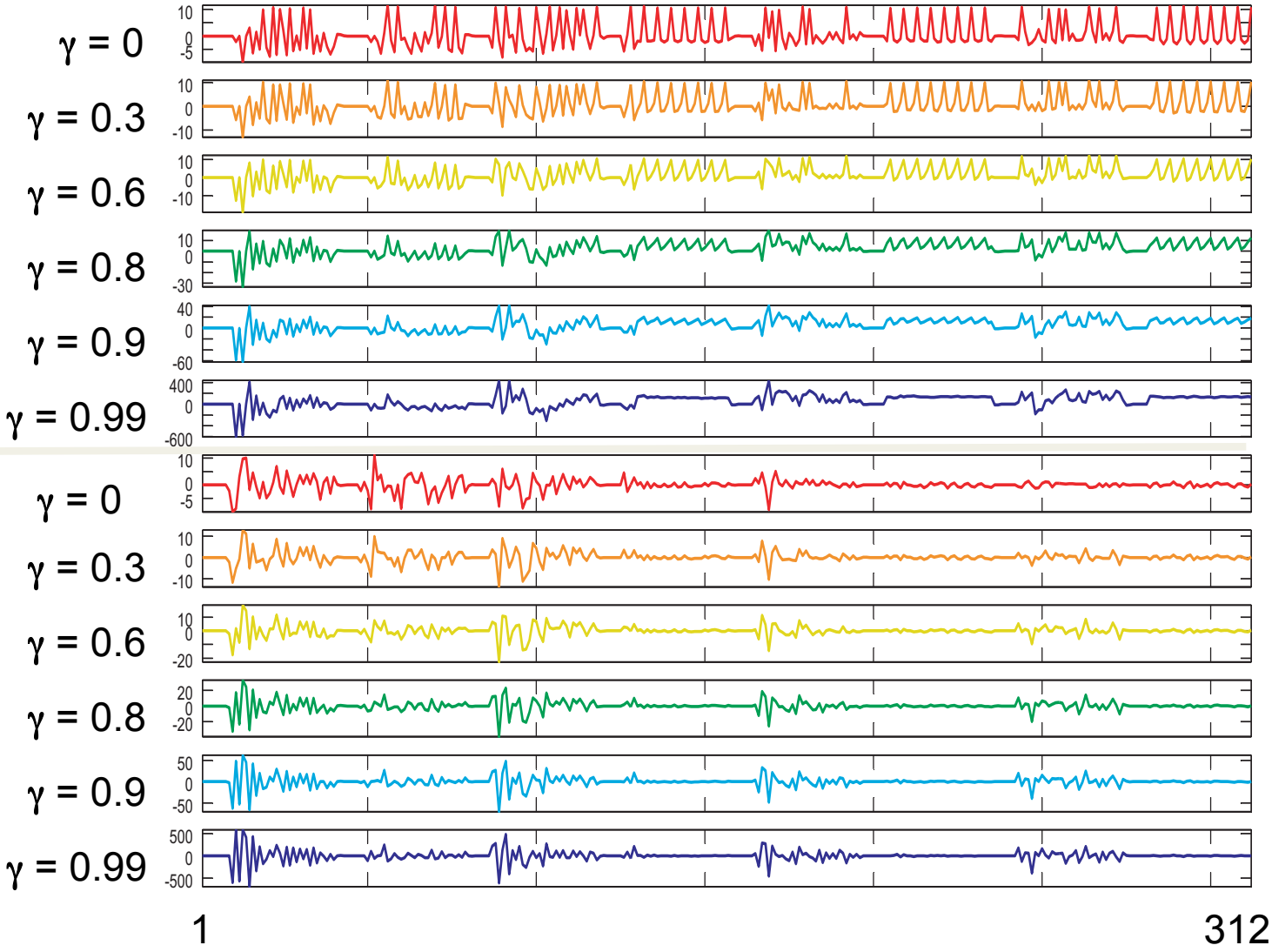
DLPFC, VLPFC, IPC, PMd Striatum **Dorsal raphe** Cerebellum

- Different areas for immediate/future reward prediction

Model-based Explanatory Variables

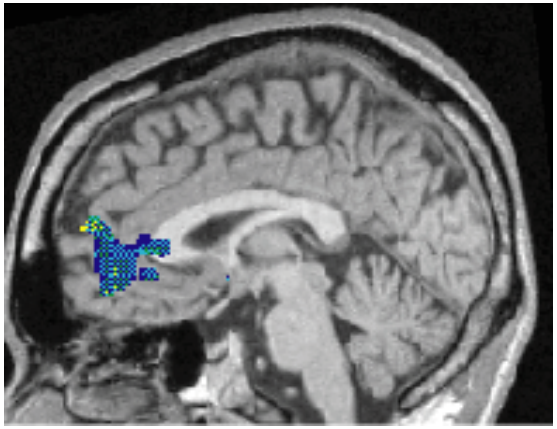
- Reward prediction V(t)

- Reward prediction error δ(t)

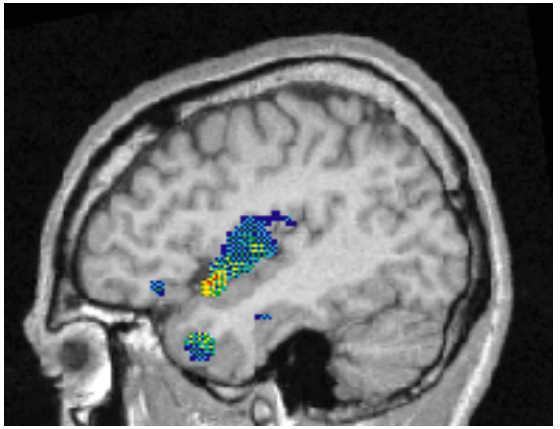


Regression Analysis

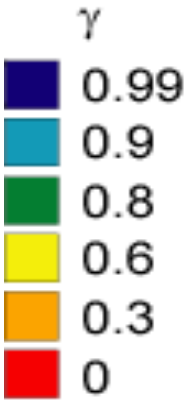
- Reward prediction V(t)



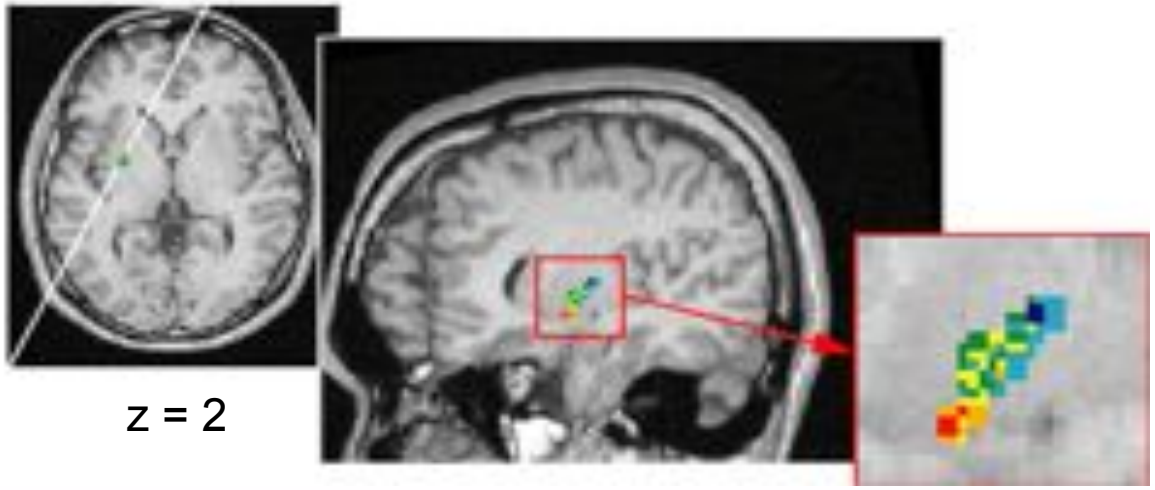
x = -2 mm



x = -42 mm



- Reward prediction error δ(t)

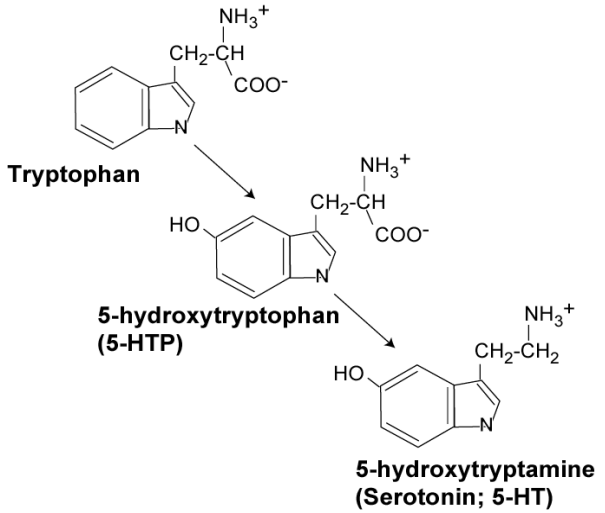


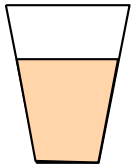
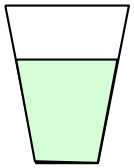
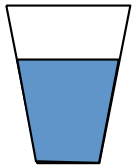
z = 2

Tryptophan Depletion/Loading

(Tanaka et al., 2007)

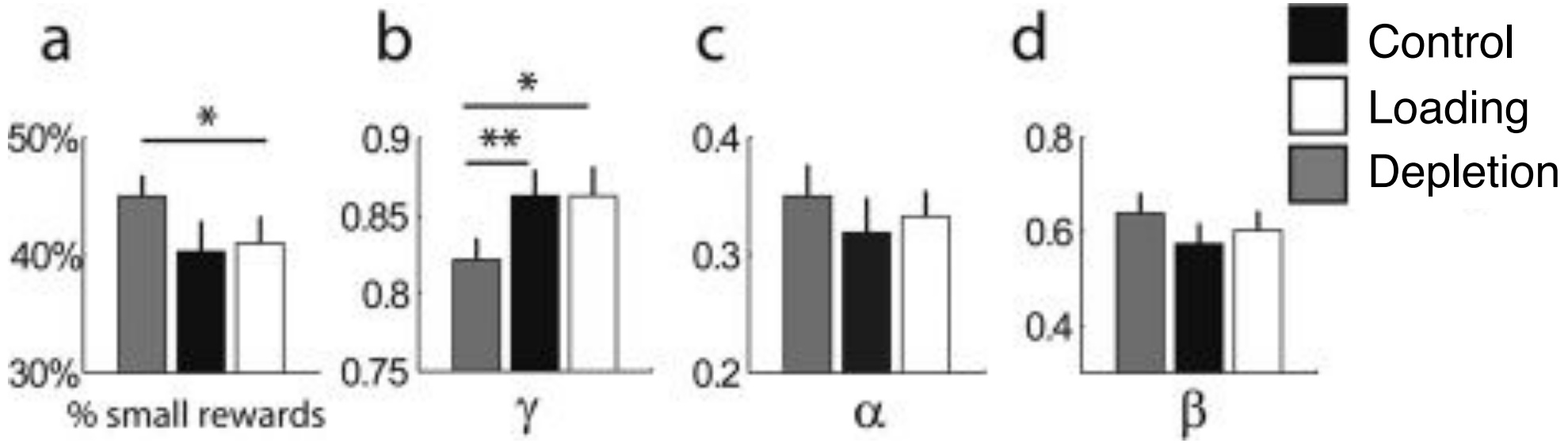
- Tryptophan: precursor of serotonin
  - depletion/loading affect central serotonin levels
    (e.g. Bjork et al. 2001, Luciana et al. 2001)
- 100 g of amino acid mixture
  - experiments after 6 hours



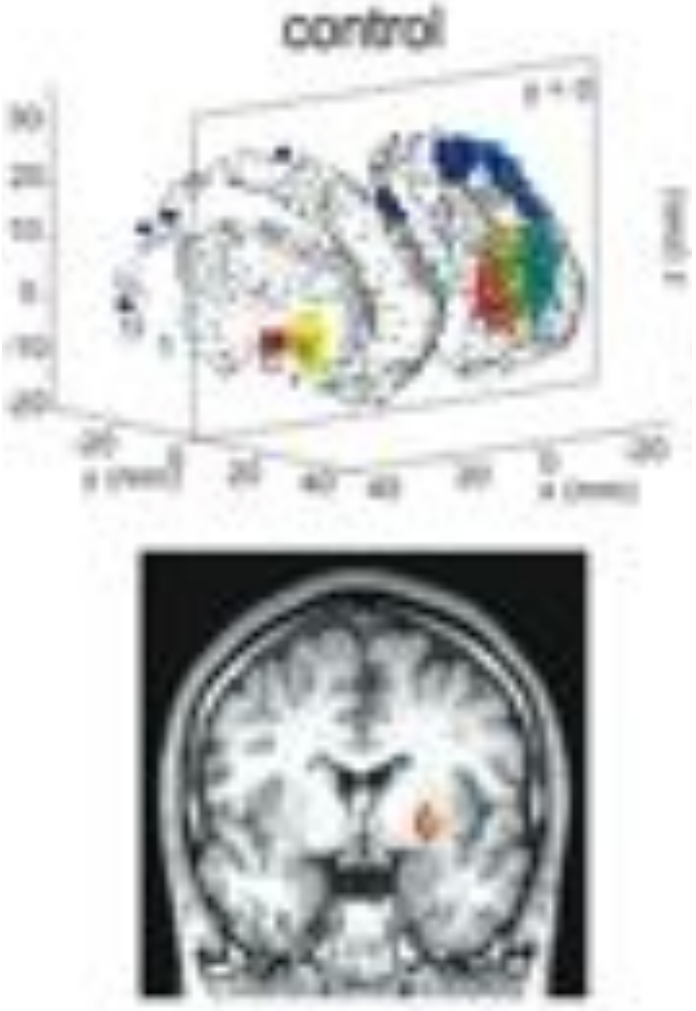
Day1: Tr-	Day2: Tr0	Day3: Tr+
		
No tryptophan (Depletion)	2.3g of tryptophan (Control)	10.3g of tryptophan (Loading)

Behavioral Result (Schweighfer et al., 2008)

- Extended sessions outside scanner

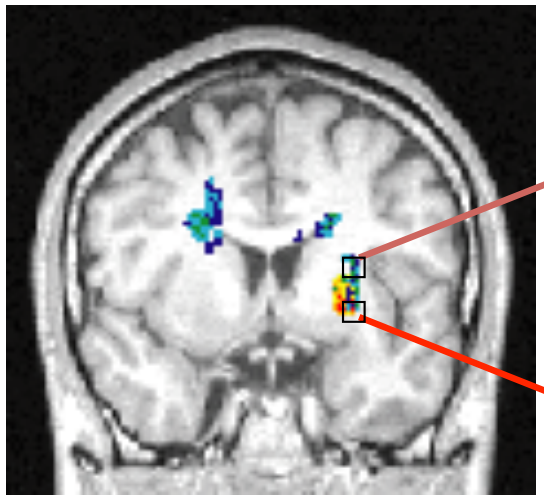


Modulation by Tryptophan Levels

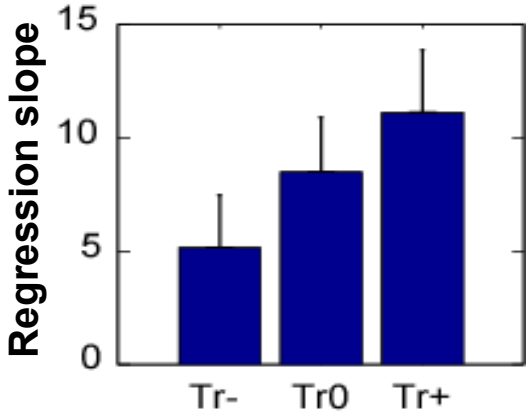


Changes in Correlation Coefficient

- ROI (region of interest) analysis

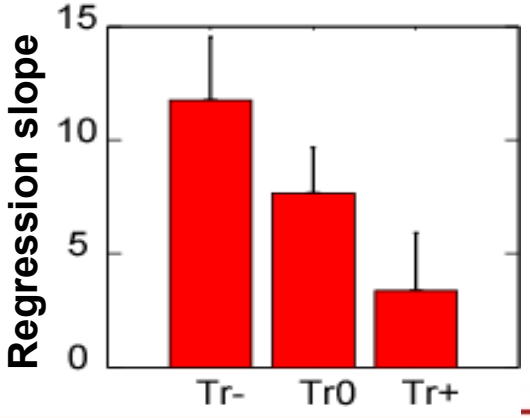


γ = 0.99
(16, 2, 28)



Tr- < Tr+
correlation with V at **large** γ
in **dorsal** Putamen

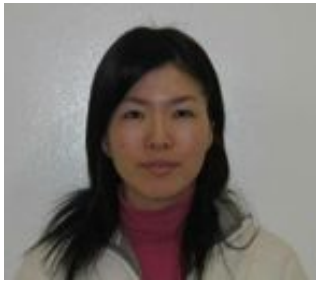
γ = 0.6
(28, 0, -4)

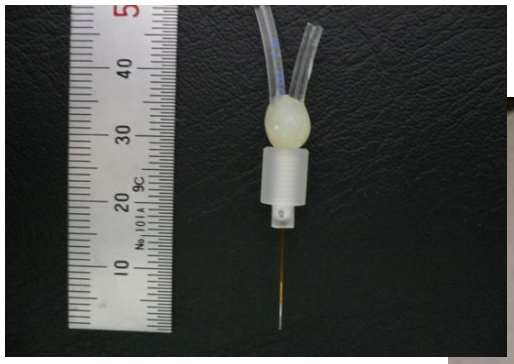


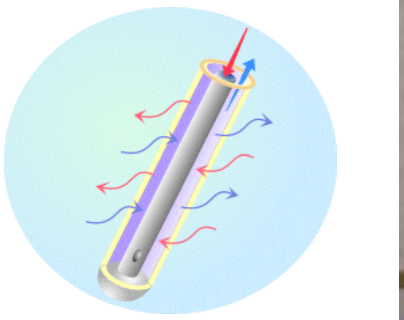
Tr- > Tr+
correlation with V at **small** γ
in **ventral** Putamen

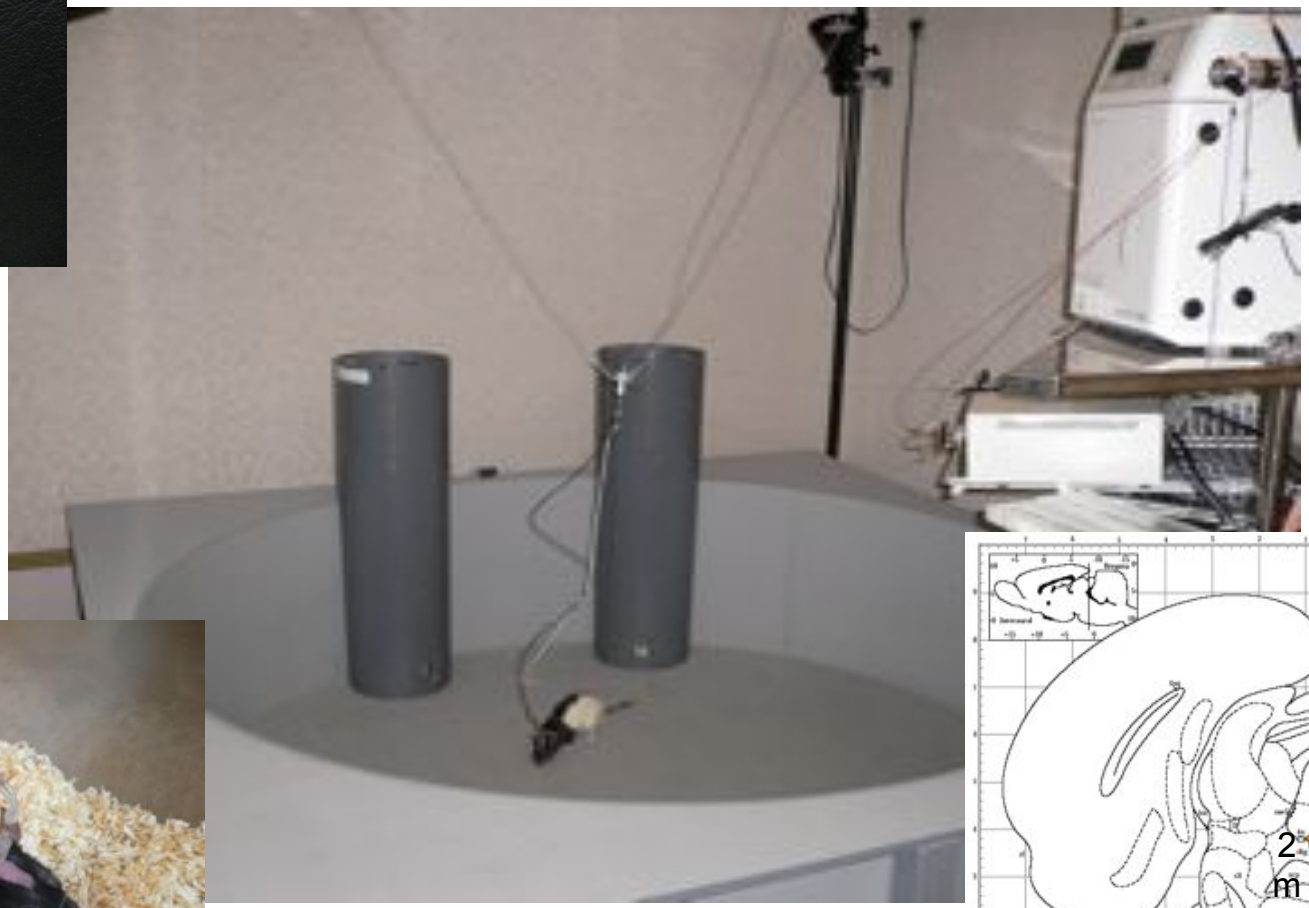
Microdialysis Experiment

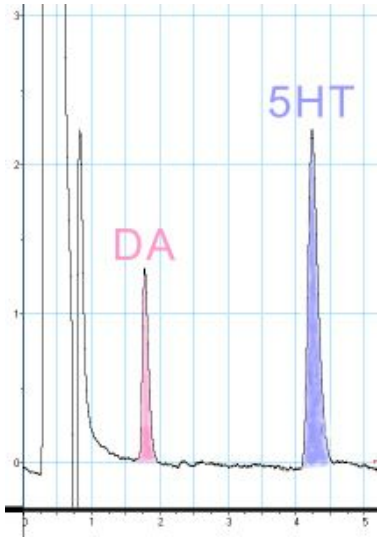
(Miyazaki et al 2011 EJNS)

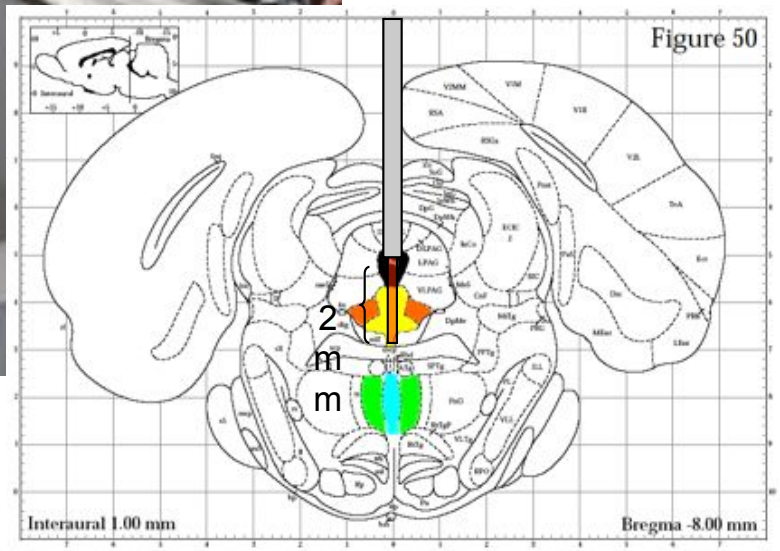


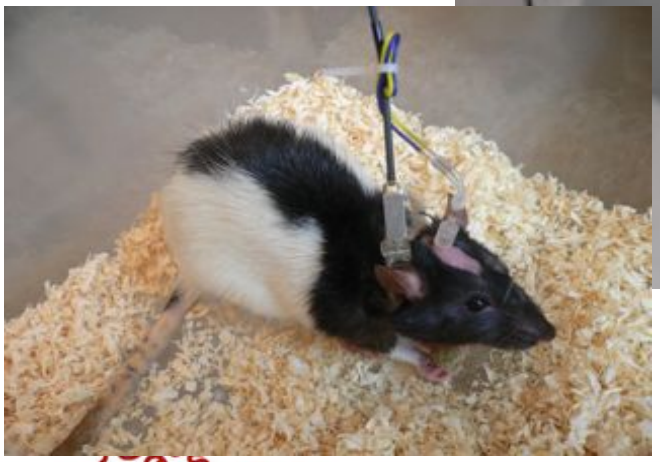






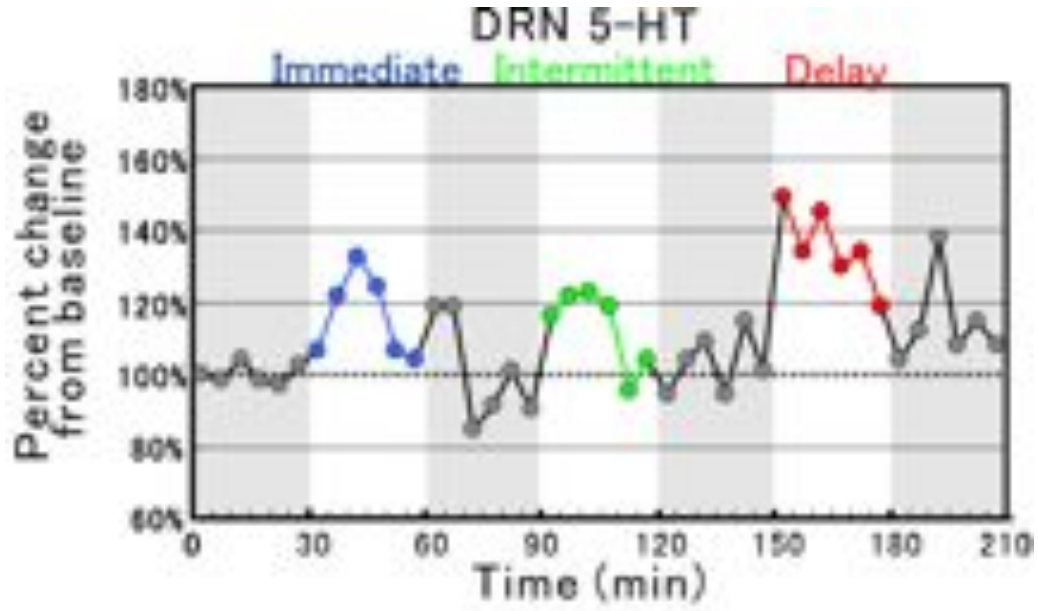




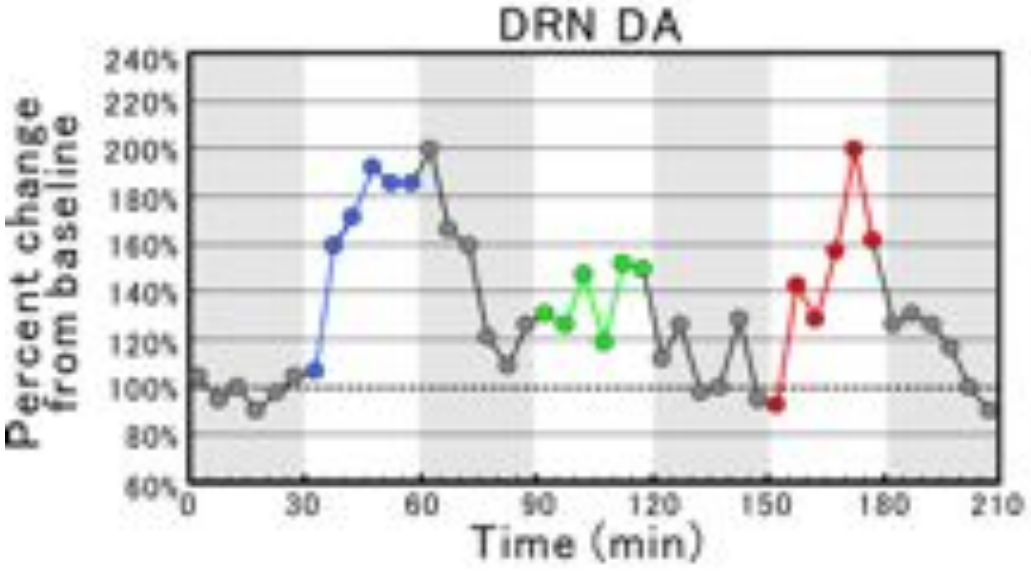


Dopamine and Serotonin Responses

Serotonin

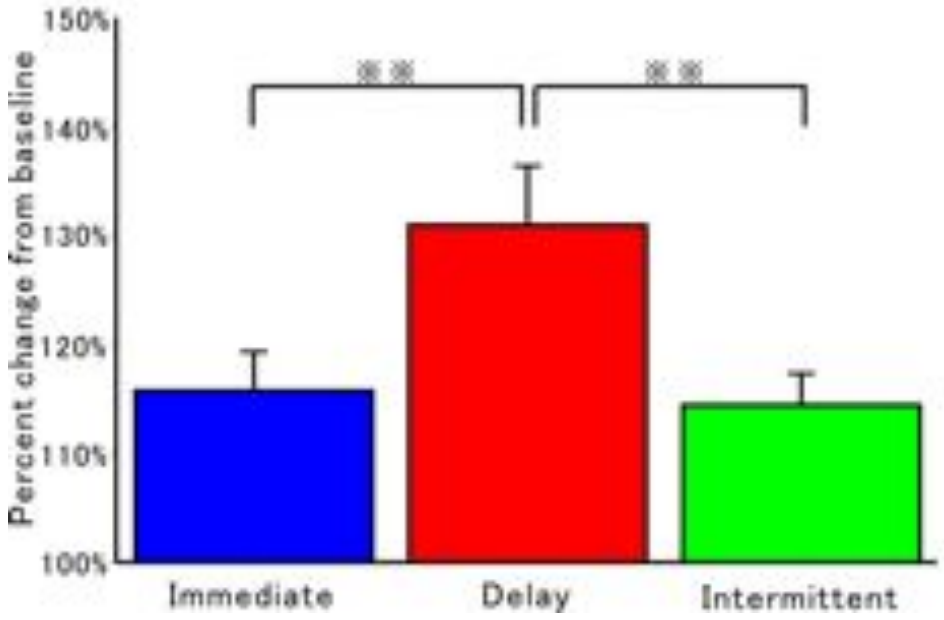


Dopamine

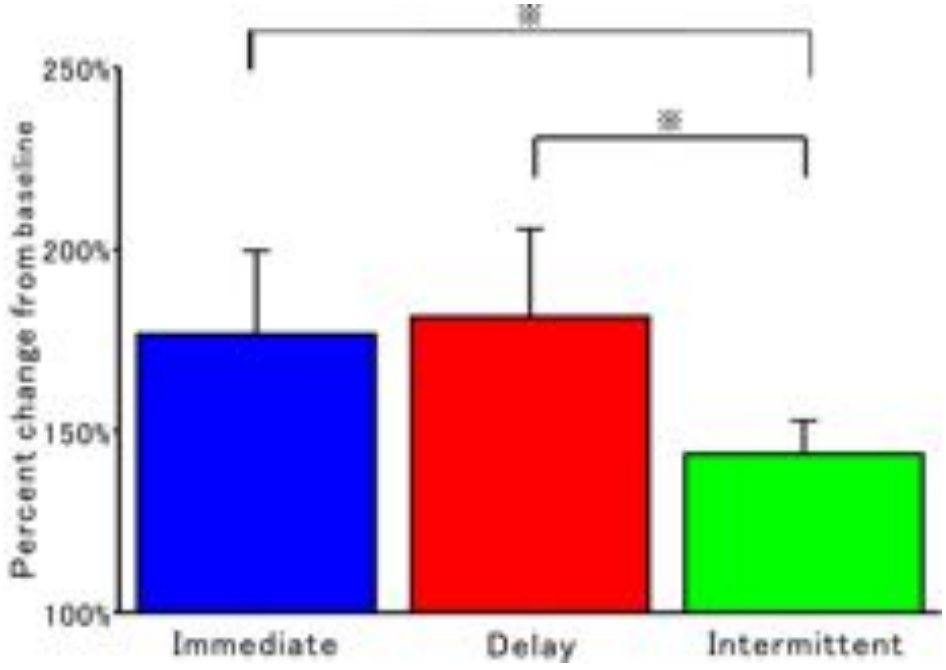


Average in 30 minutes

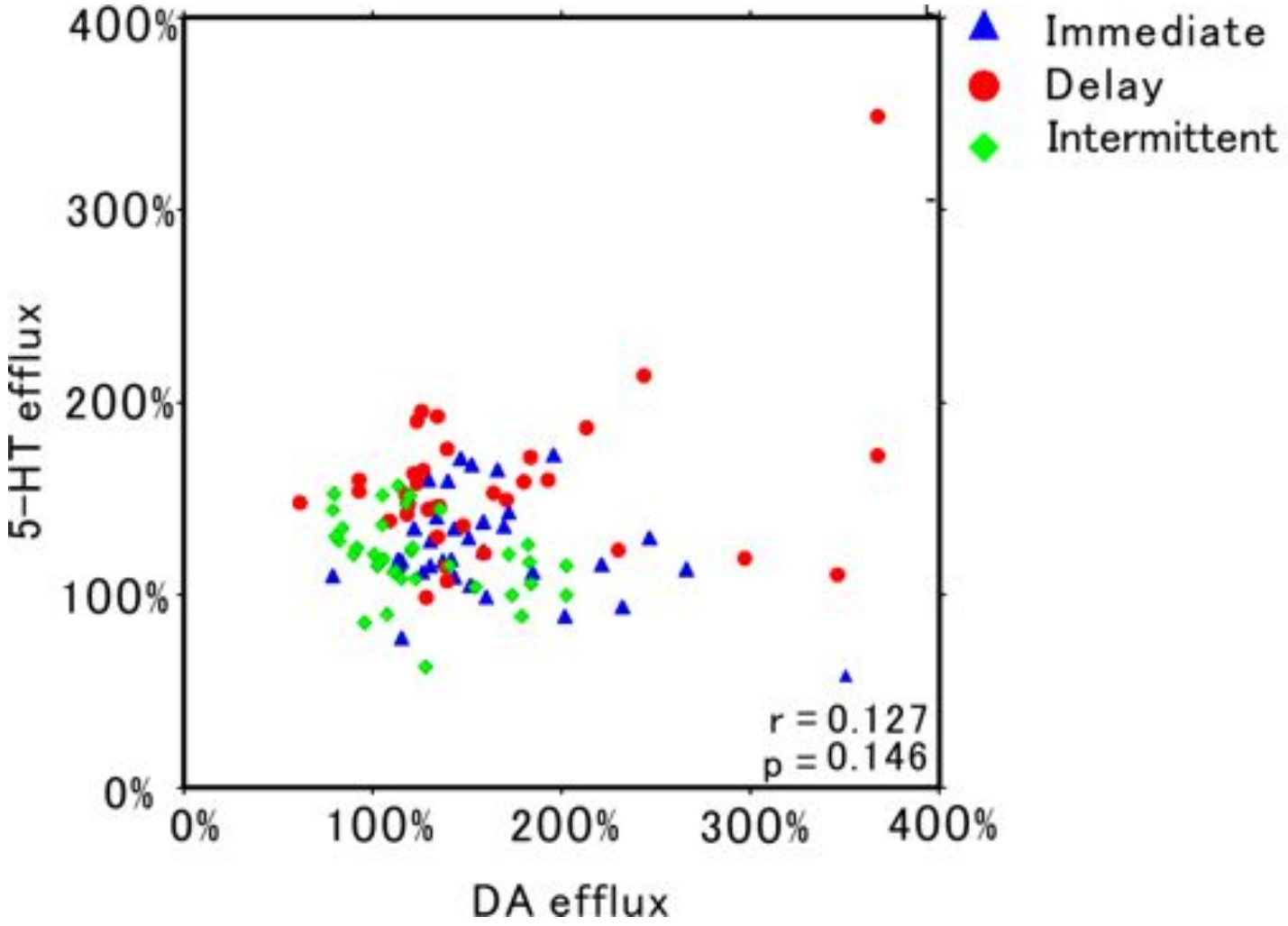
- Serotonin (n=10)



- Dopamine (n=8)



Serotonin v.s. Dopamine



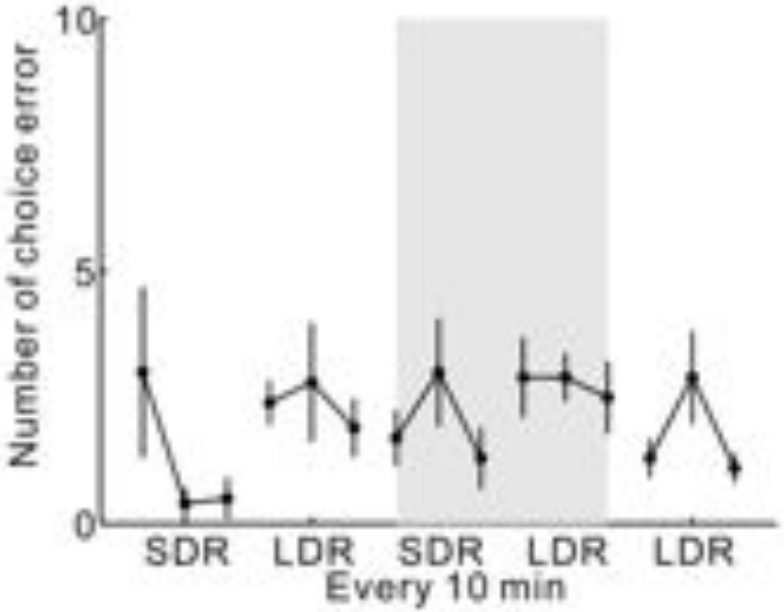
- No significant positive or negative correlation

Effect of Serotonin Suppression

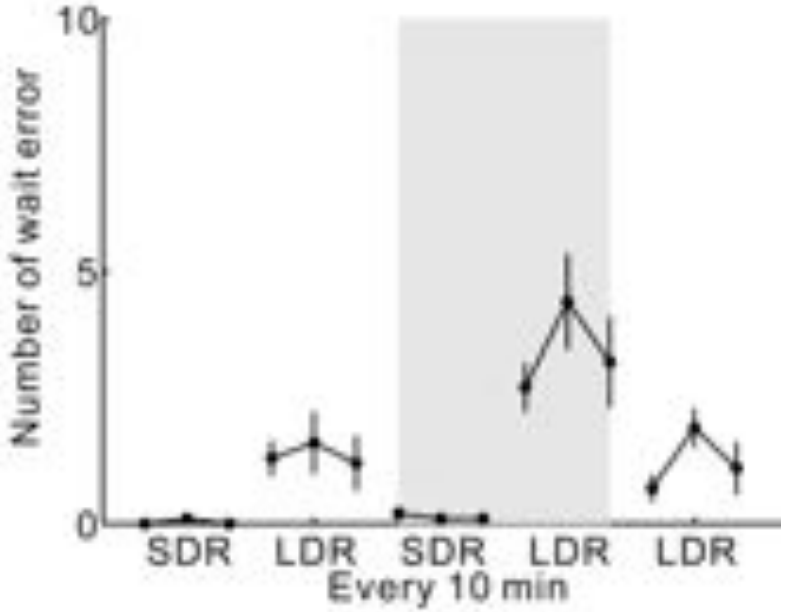
(Miyazaki et al. JNS 2012)

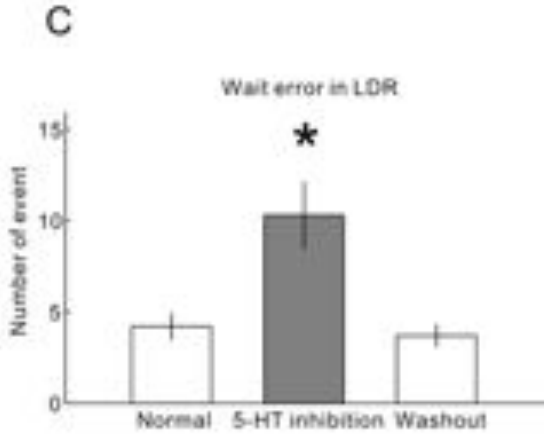
- 5-HT_{1A} agonist in DRN
- Wait errors in long-delayed reward condition

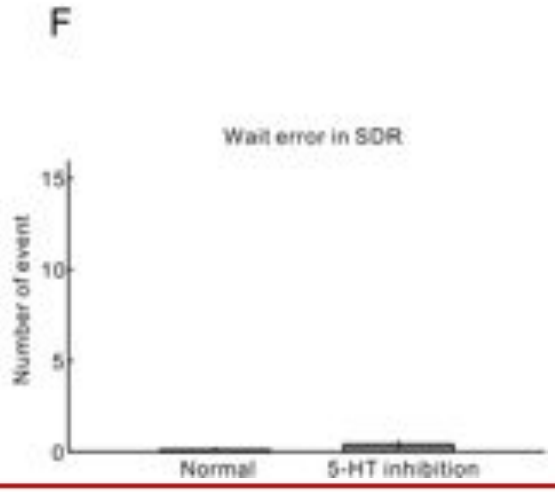
choice error



wait error

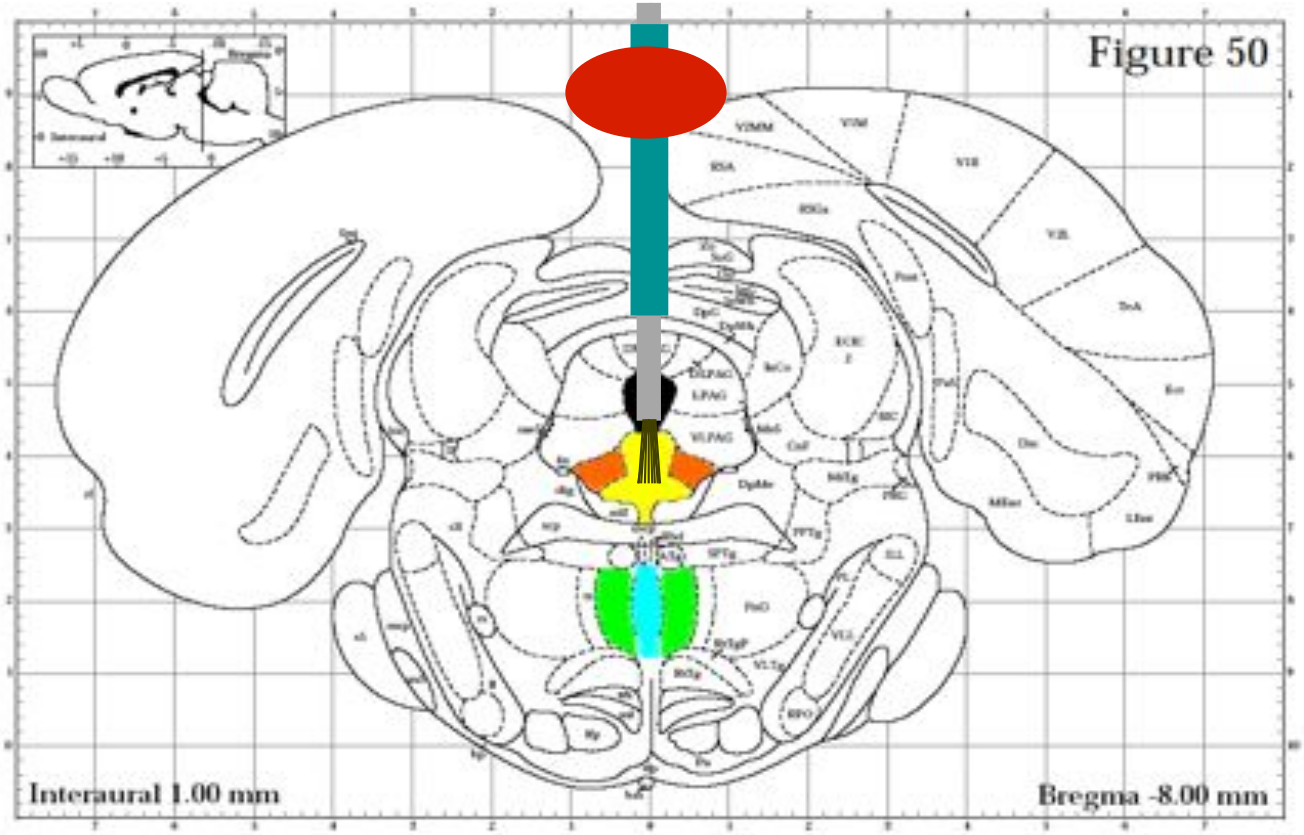




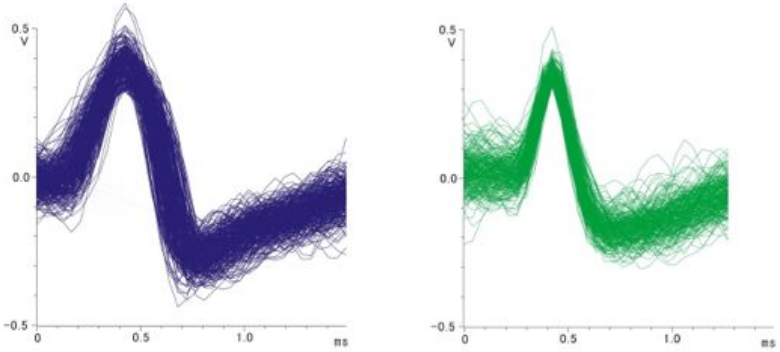


Dorsal Raphe Neuron Recording

(Miyazaki et al. 2011 JNS)



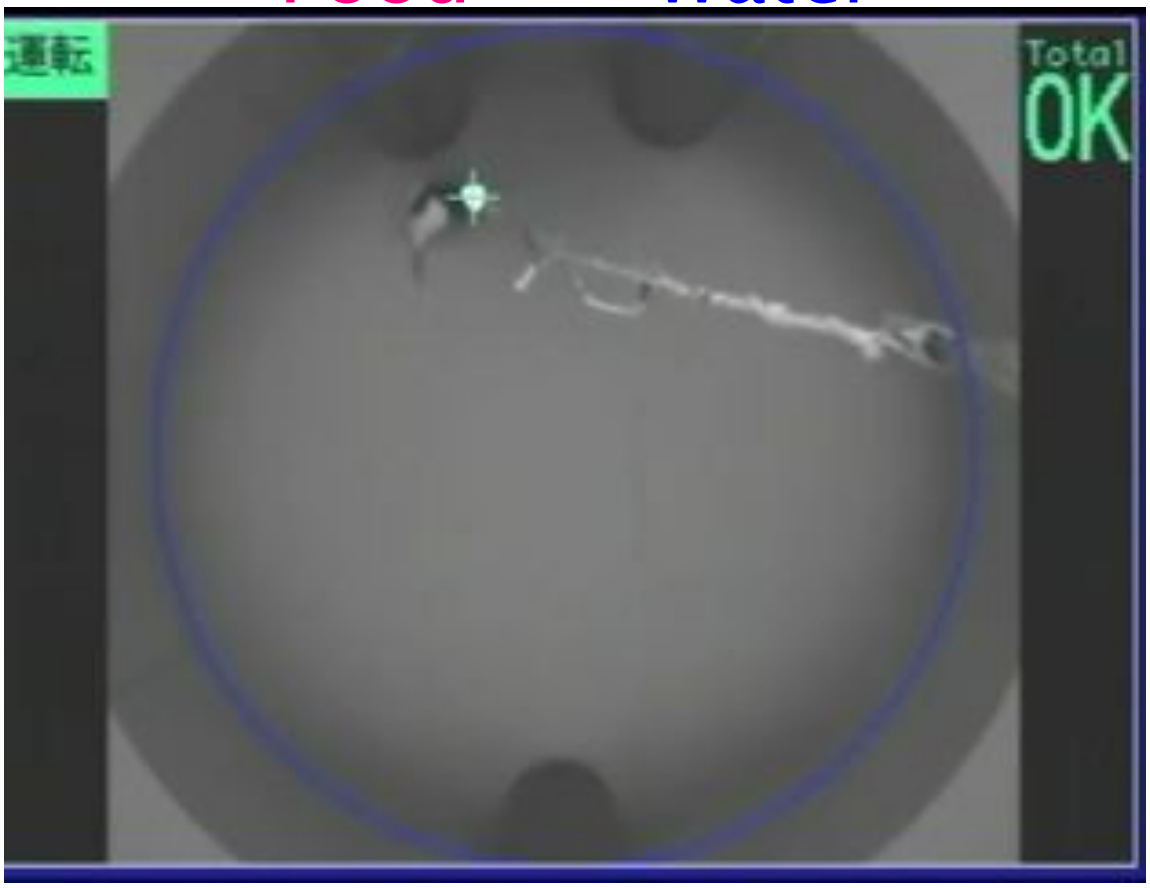
- Putative 5-HT neurons
  - wider spikes



  - low firing rate
  - suppression by 5-HT1a agonist

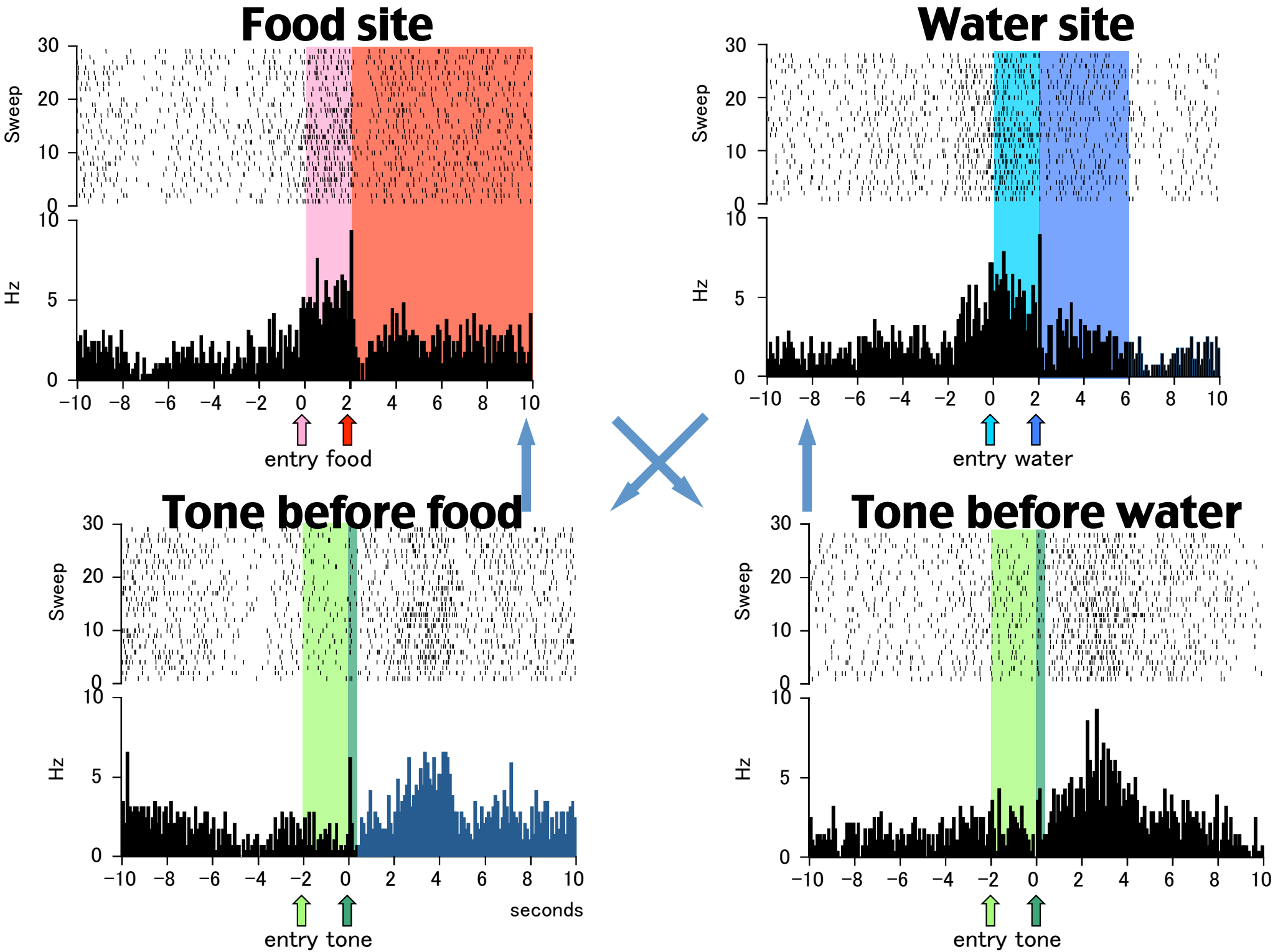
Delayed Tone-Food-Tone-Water Task

Food Water



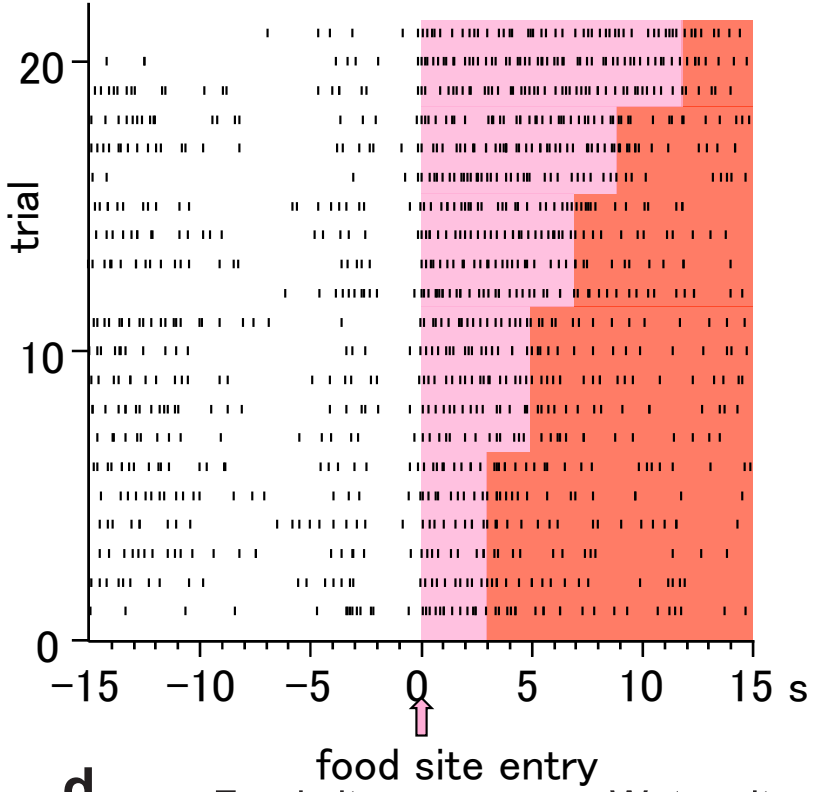
Tone

- 2 ~ 20 sec delays

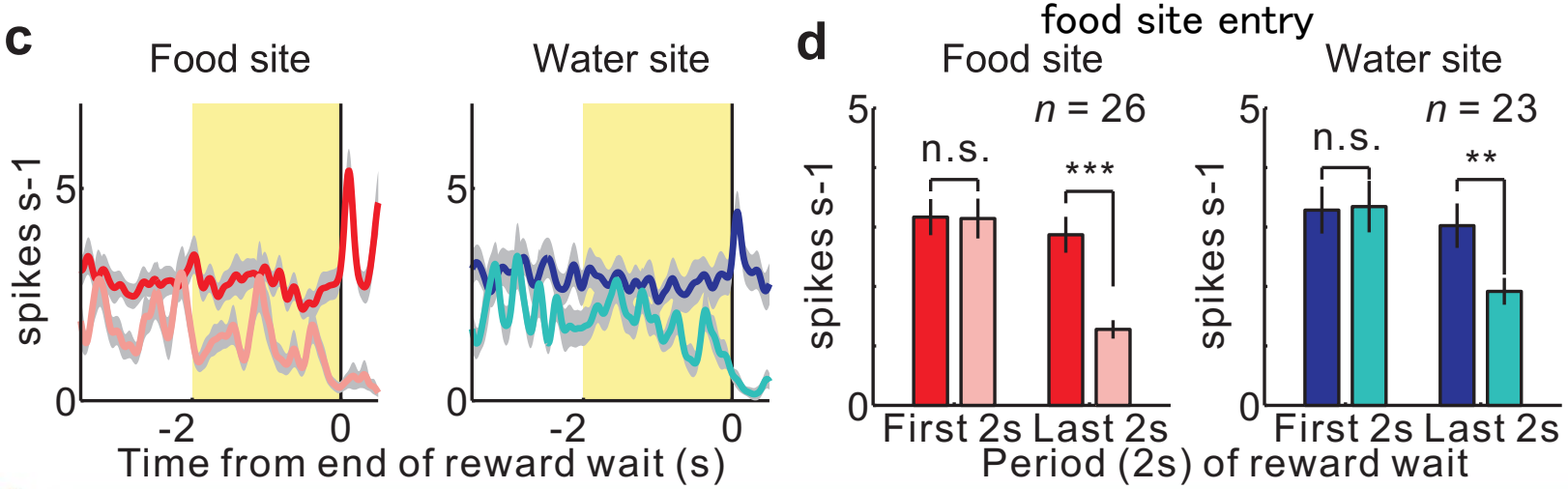

Food site
Water site
Tone before food
Tone before water
Sweep
Hz
entry food
entry water
entry tone
entry tone
seconds

Error Trials in Extended Delay

- Sustained firing



- Drop before wait error



Serotonin Experiments: Summary

- Microdialysis
  - higher release for delayed reward
  - no apparent opponency with dopamine
  - lower release cause waiting error

Consistent with discounting hypothesis

- Serotonin neuron recording
  - higher firing during waiting
  - firing stops before giving up

More dynamic variable than a 'parameter'

- Question: regulation of serotonin neurons
  - algorithm for regulation of patience

Collaborators

- ATR → Tamagawa U
  - Kazuyuki Samejima
- NAIST→CalTech→ATR→Osaka U
  - Saori Tanaka
- CREST → USC
  - Nicolas Schweighofer
- OIST
  - Makoto Ito
  - Kayoko Miyazaki
  - Katsuhiko Miyazaki
  - Takashi Nakano
  - Jun Yoshimoto
  - Eiji Uchibe
  - Stefan Elfwing
- Kyoto PUM
  - Minoru Kimura
  - Yasumasa Ueda
- Hiroshima U
  - Shigeto Yamawaki
  - Yasumasa Okamoto
  - Go Okada
  - Kazutaka Ueda
  - Shuji Asahi
  - Kazuhiro Shishida
- U Otago → OIST
  - Jeff Wickens