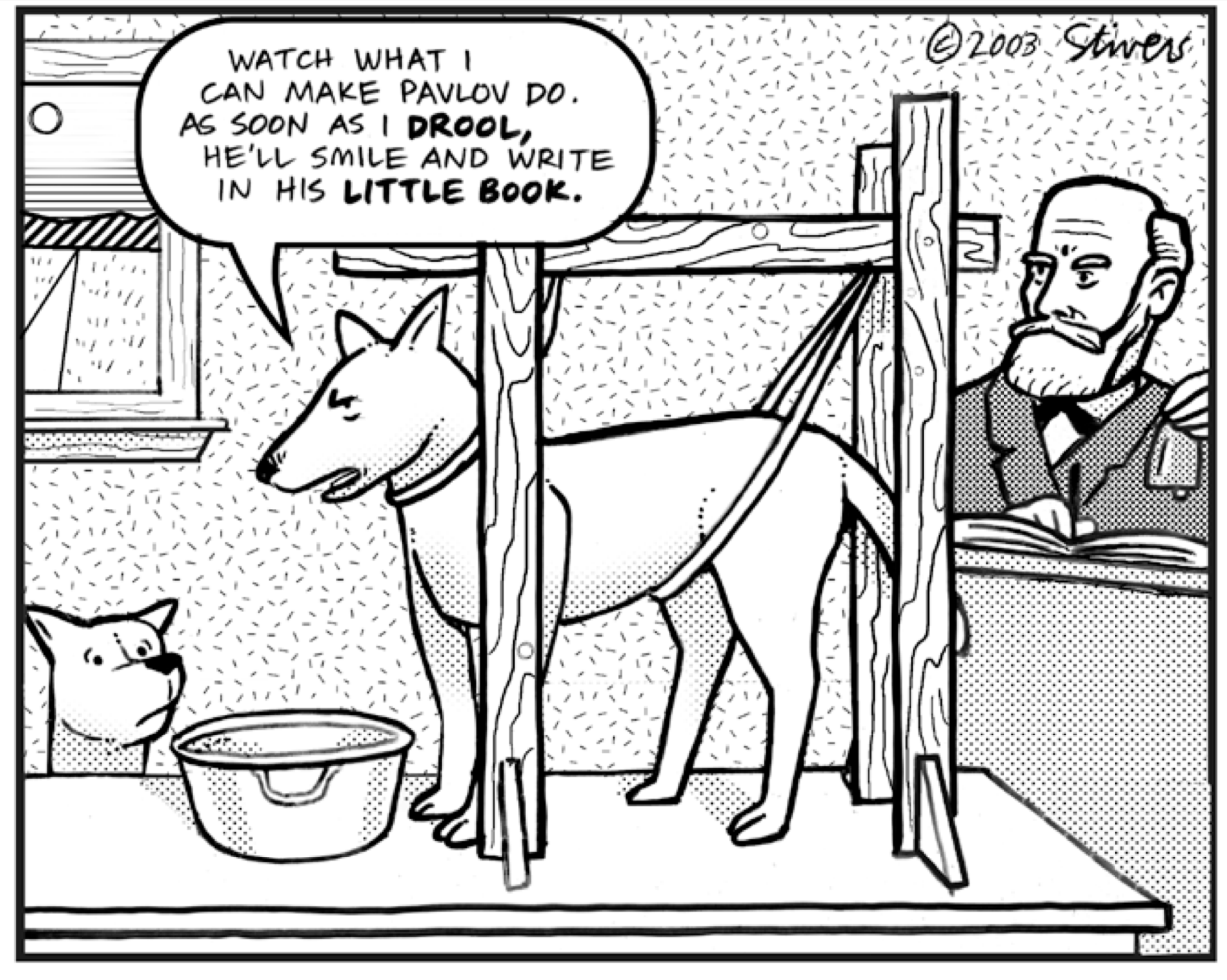
Temporal reinforcement learning



Slides adapted from:
Andra Geana
Jeff Cockburn
Michael Frank

Assume you're a brain

You need to eat things, you don't want to be eaten by other things, and you'd "like" to produce more brains.

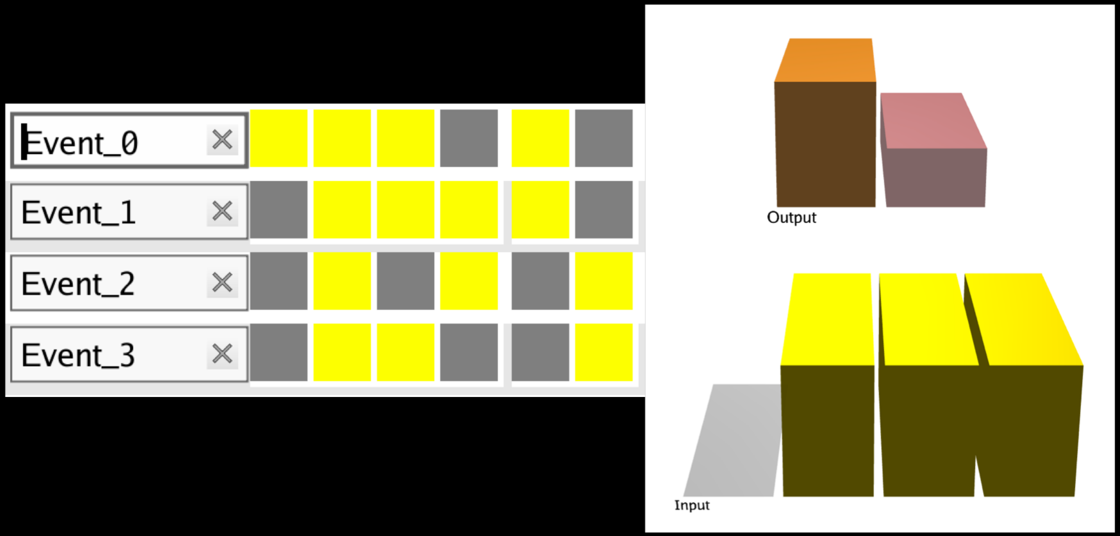
1. *What is learning?*
2. *Why is learning important?*
3. *What should you learn?*
4. *When should you learn?*

There are different types of "learning"

Learning: Remembering relationships

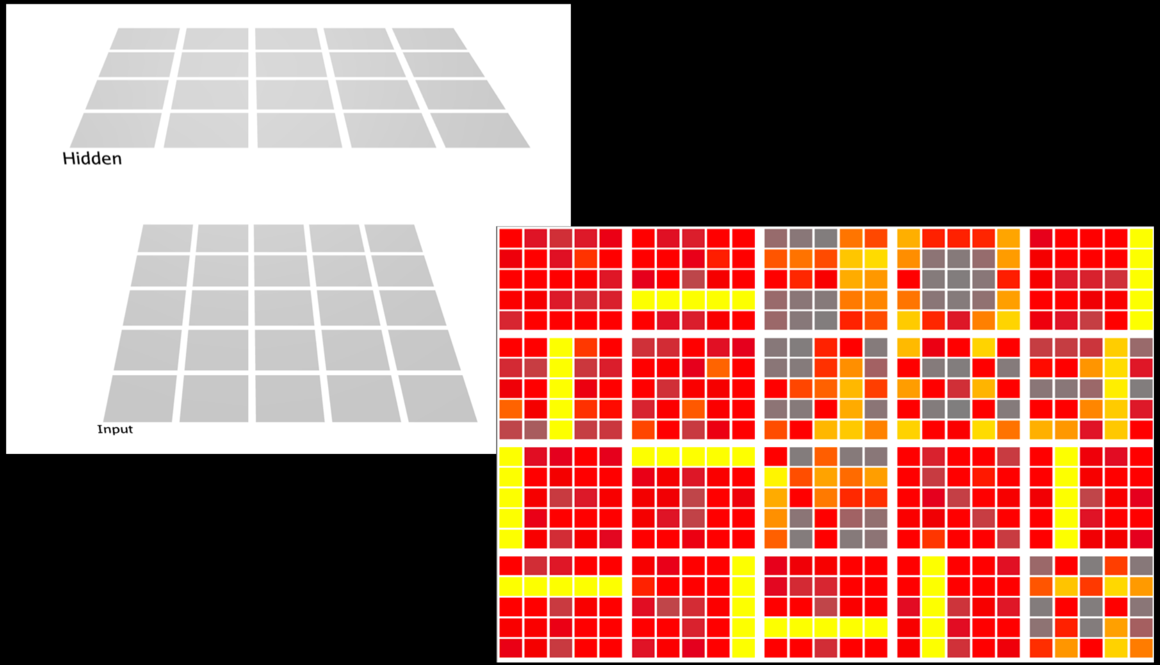
SUPERVISED

Someone tells you them (e.g. parents)

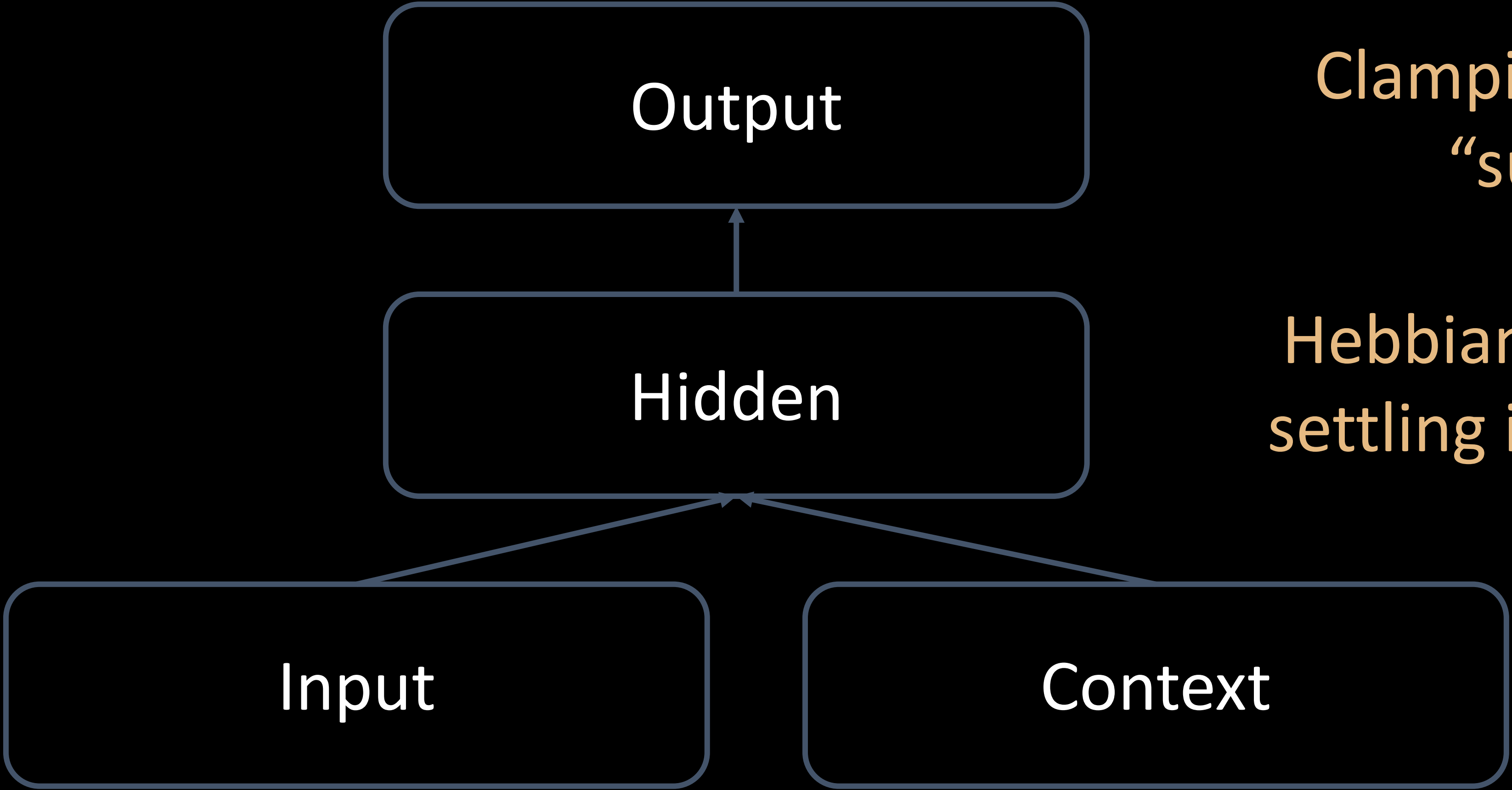


UNSUPERVISED

You see stuff out there



Learning example: SRN



Clamping the output is "supervising".

Hebbian plasticity during settling is "unsupervised".

Reinforcement learning: In-between

Features:

- No ‘teacher’
- No ‘correct’ answer given
- Only better/worse outcomes

What determines value?

Grounded in motivation

Some ‘wants’ are innate

Link events/actions to innate ‘wants’

Supervised? Un?

Link spans time!

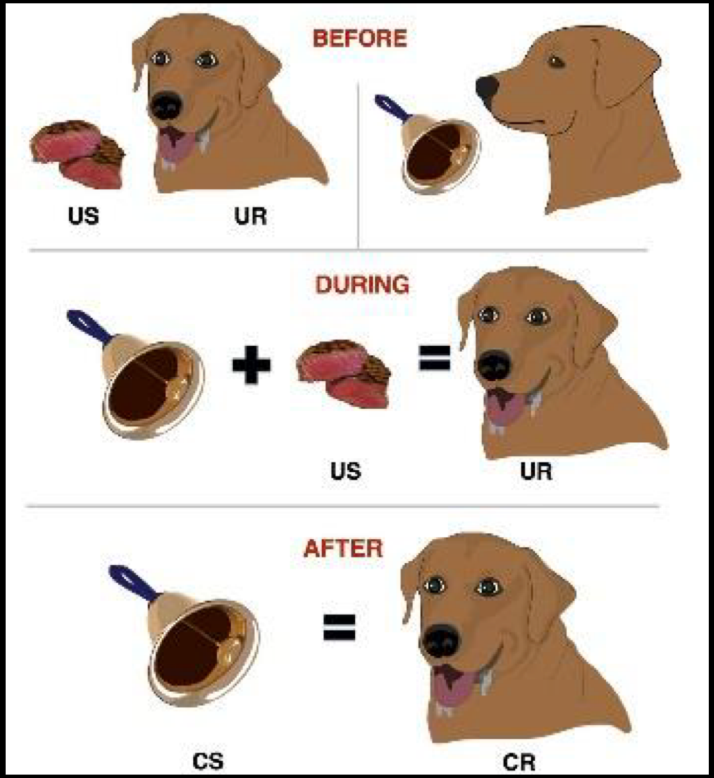
Reinforcement Learning

- Reinforcement (term from operant conditioning)
  - Something that makes it more likely that a certain response will re-occur

Reinforcement Learning

- Reinforcement (term from operant conditioning)
  - Something that makes it more likely that a certain response will re-occur

Conditioning: training an organism to respond in specific ways to certain stimuli (shaping behavior)



Classical (Pavlovian):
train involuntary responses (CR)

Instrumental (operant):
train voluntary responses (actions)

Reinforcement Learning

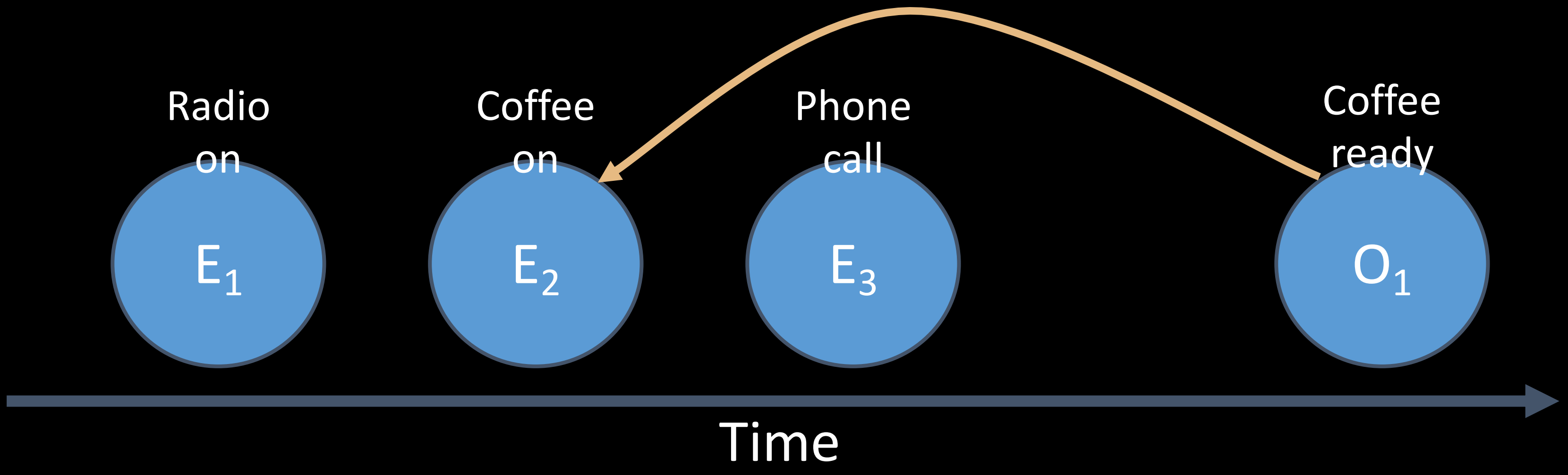
Using trial-and-error to learn how to map situations to actions so as to maximize a numerical reward signal

Sutton & Barto, 2nd Ed, 2018

Learning about delayed outcomes

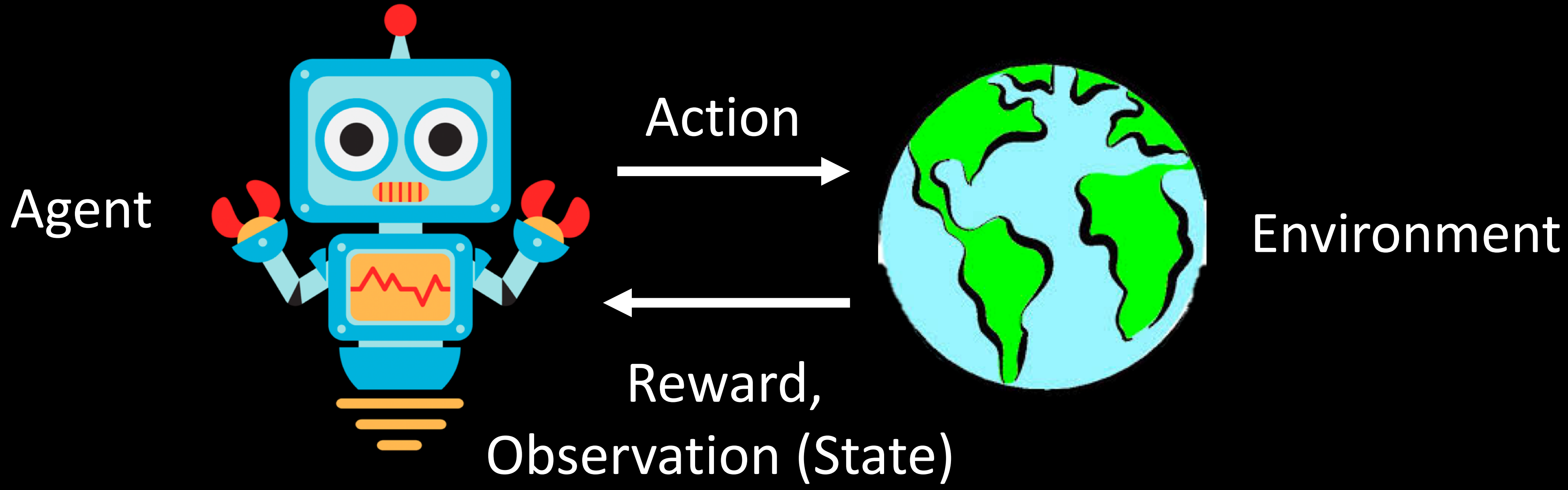
- The results of actions are often delayed
- Irrelevant information abounds
- Want “good” results

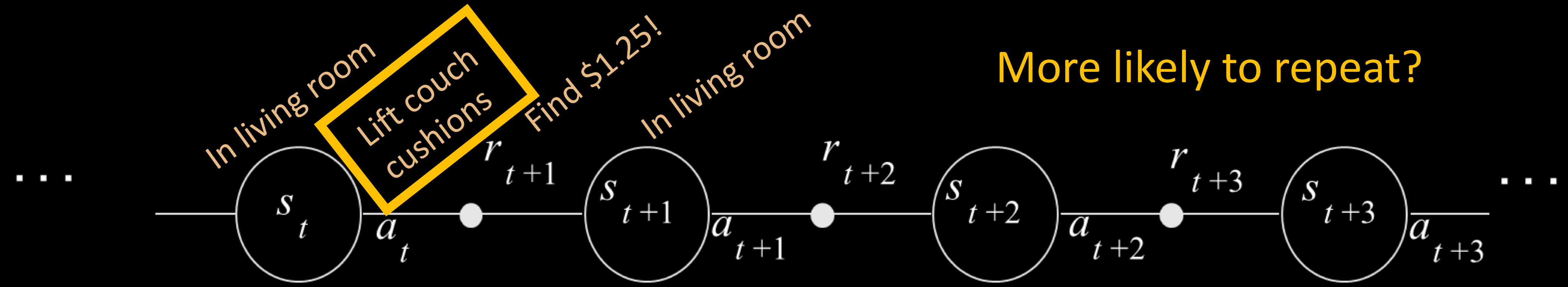
Goal: Learn to predict event outcomes



Referred to as: Temporal credit assignment

Temporal difference learning (informally)

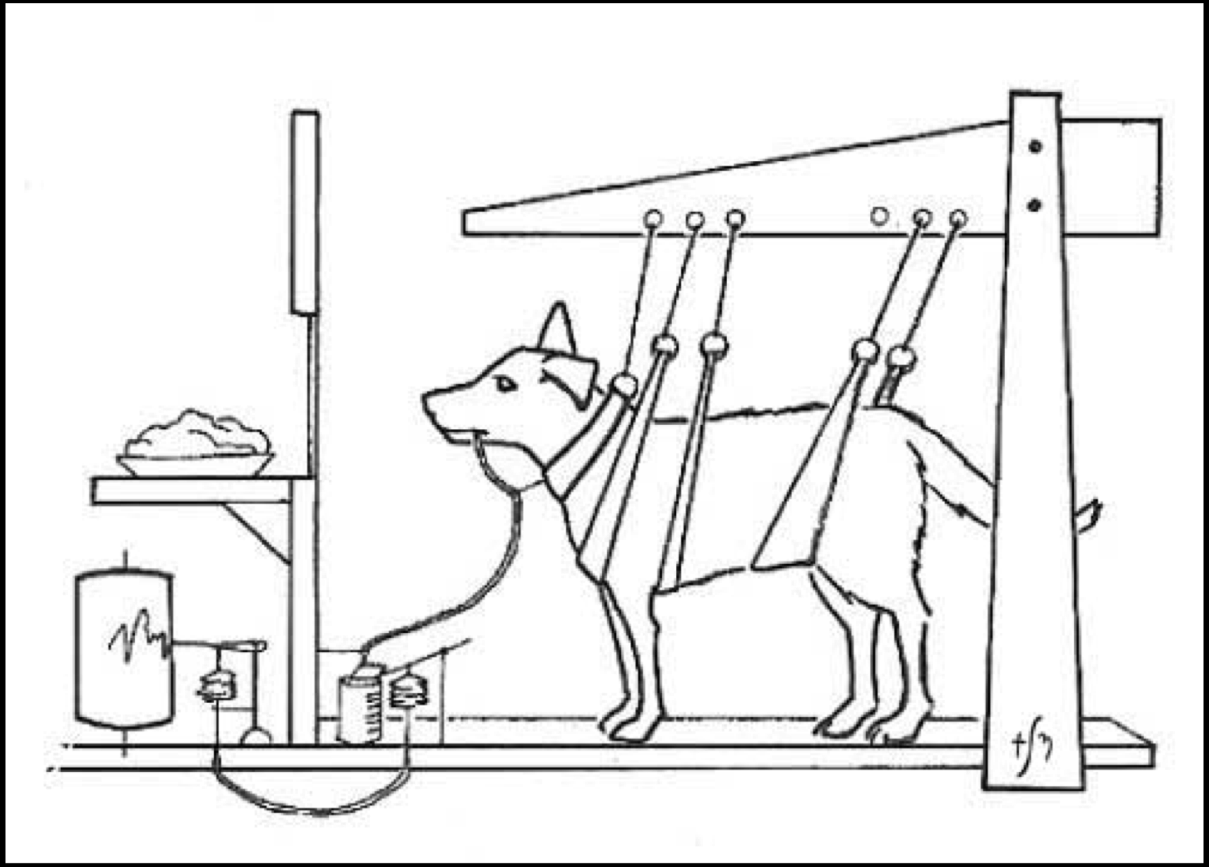




More likely to repeat?

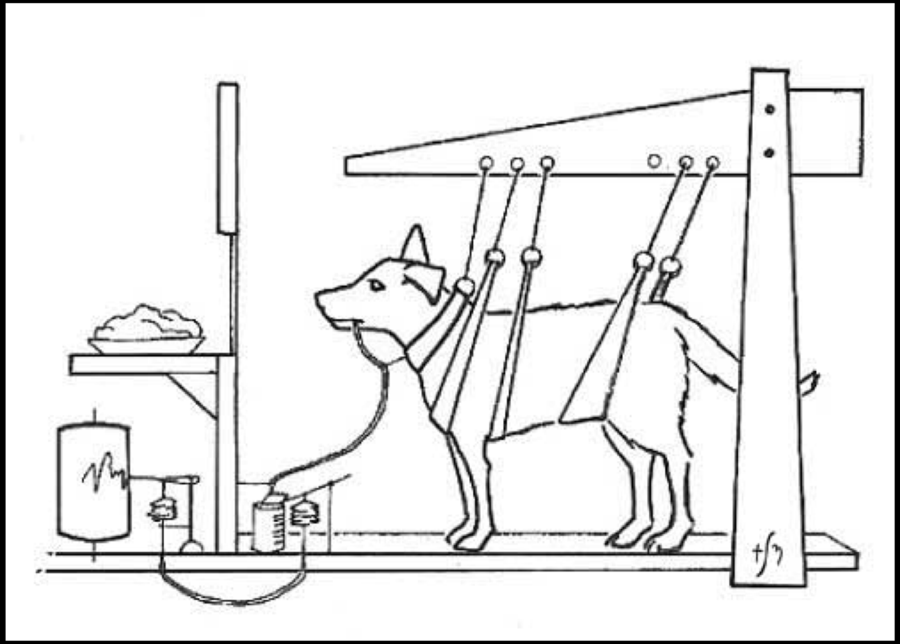
Classical Conditioning

What did Pavlov learn about what his dog learned?



Classical Conditioning

What did Pavlov learn about what his dog learned?



Day 1

A bell rings:

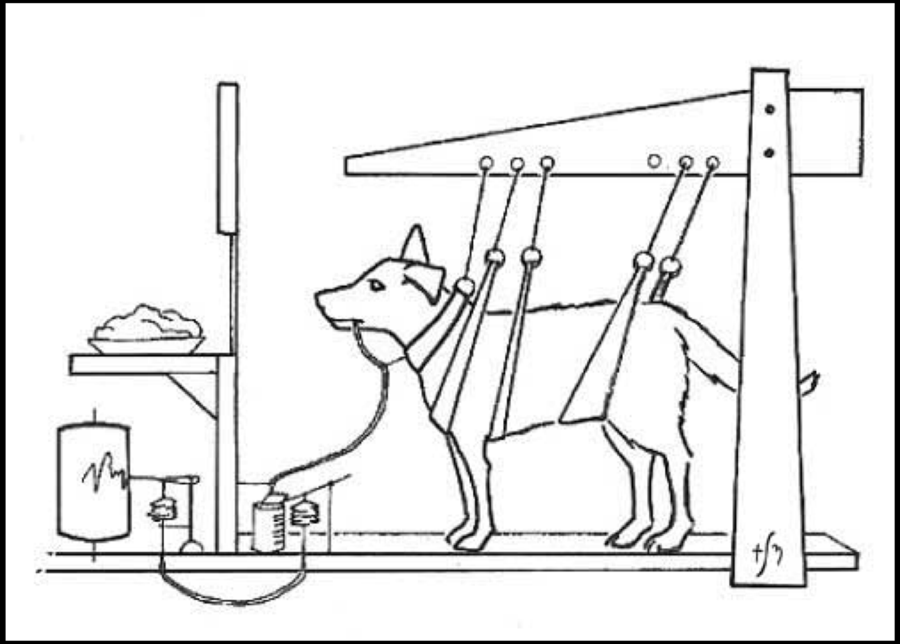
- Was that: expected / unexpected ?
- Is this: Better / Worse / Neutral to what you expected?
- Is there anything to learn?

Food arrives:

- Was that: expected / unexpected ?
- Is this: Better / Worse / Neutral to what you expected?
- Is there anything to learn?

Classical Conditioning

What did Pavlov learn about what his dog learned?



Day 2

A bell rings:

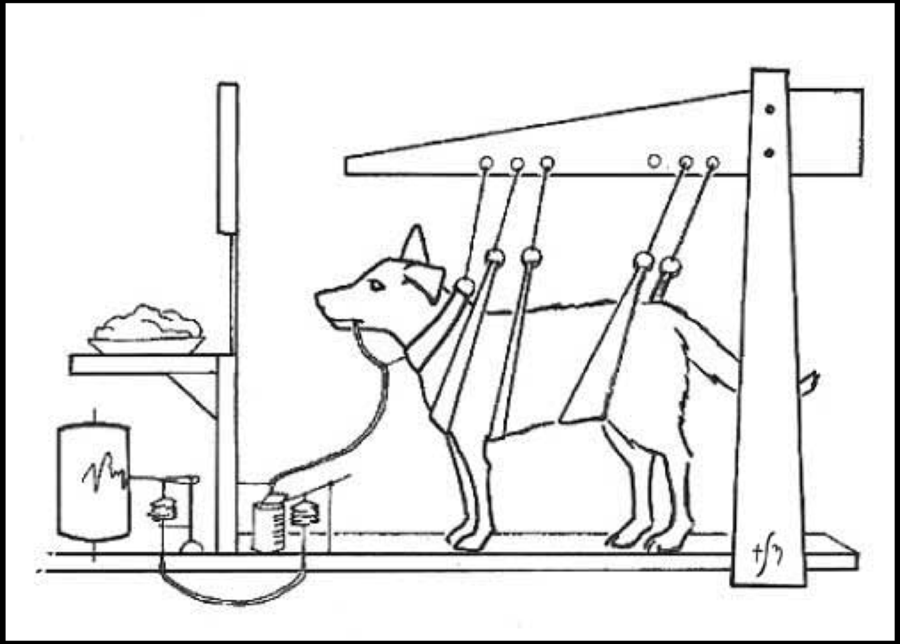
- Was that: expected / unexpected ?
- Is this: Better / Worse / Neutral to what you expected?
- Is there anything to learn?

Food arrives:

- Was that: expected / unexpected ?
- Is this: Better / Worse / Neutral to what you expected?
- Is there anything to learn?

Classical Conditioning

What did Pavlov learn about what his dog learned?



Day 3

A bell rings:

- Was that: expected / unexpected ?
- Is this: Better / Worse / Neutral to what you expected?
- Is there anything to learn?

Food arrives:

- Was that: expected / unexpected ?
- Is this: Better / Worse / Neutral to what you expected?
- Is there anything to learn?

Rescorla & Wagner (1972)

Why would you not always want a learning rate of 1?

- **Prediction error or "surprise" driven learning rule**

Prediction Error
difference b/n observed and predicted

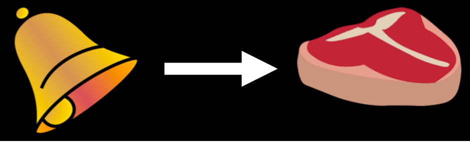
Predicted Value
of state at time t

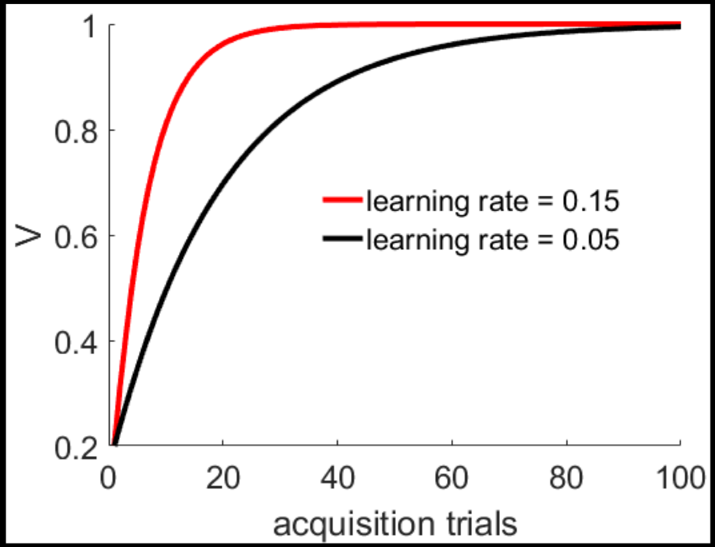
$$\delta_t = r_t - V_t(s)$$

$$V_{t+1}(s) = V_t(s) + \alpha\delta_t$$

Updated Value
at time t+1 of state

Learning rate: how much do we care about each new data point?

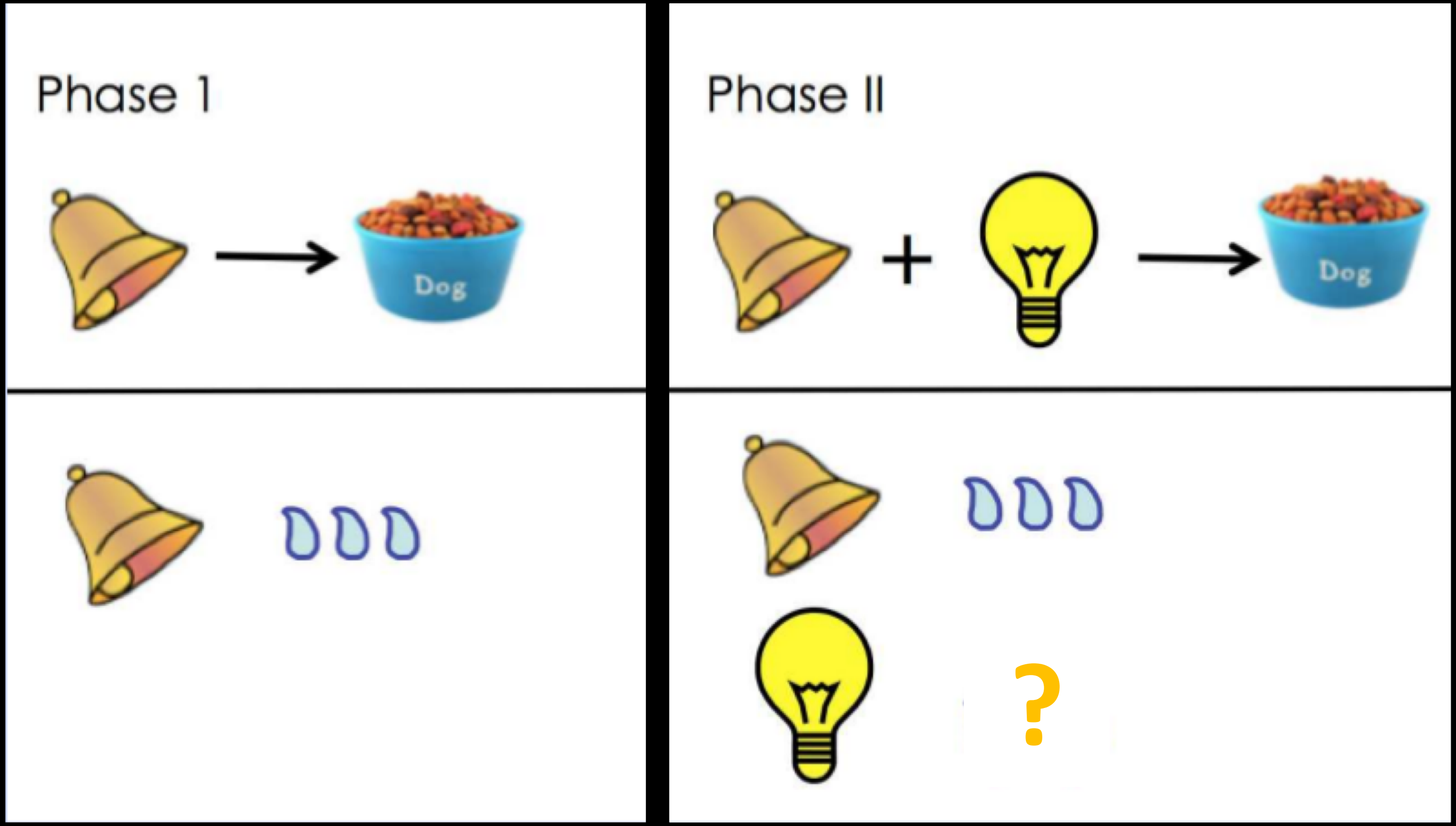




Higher learning rates means we learn (***and forget***) faster!

Blocking

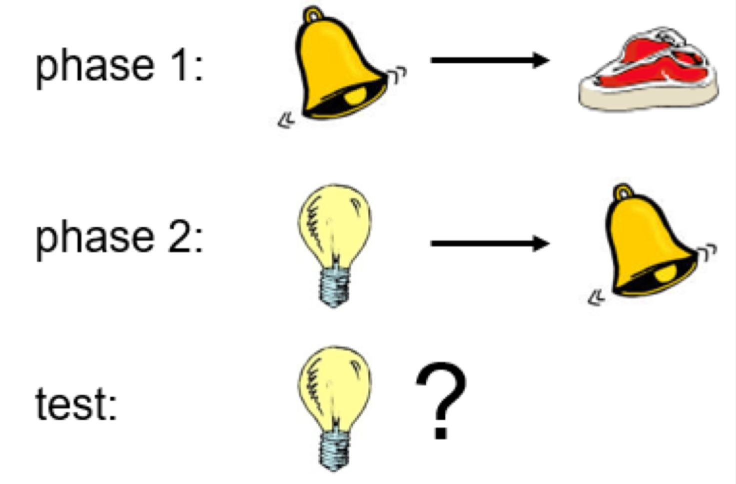
$$\delta_t = r_t - V_t(s)$$
$$V_{t+1}(s) = V_t(s) + \alpha\delta_t$$



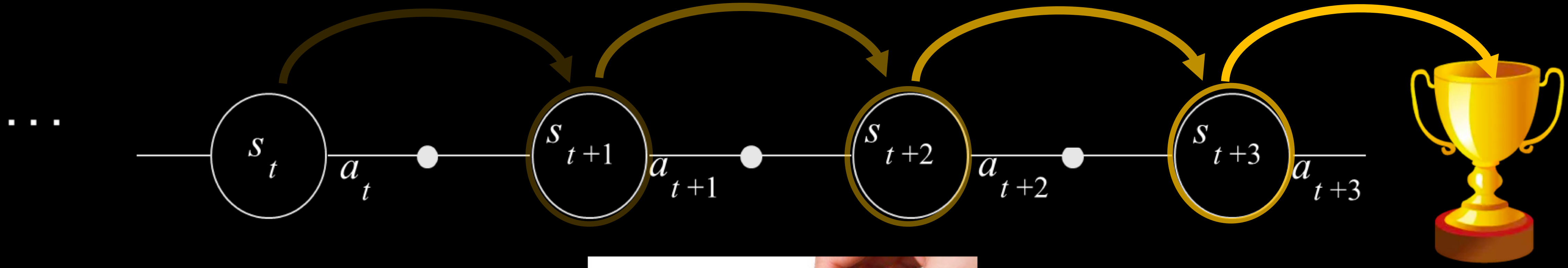
Problems with Rescorla Wagner

$$\delta_t = r_t - V_t(s)$$
$$V_{t+1}(s) = V_t(s) + \alpha\delta_t$$

- Trial-level learning and prediction
- Scaling problem
- Can't explain effects like associative bias
  - Rats more likely to associate light and shock or flavor and poisoning
- Can't explain second order conditioning
  - *Why?*



How can RL handle sequences of states?



Rather than trial-level predicting based on immediate rewards,
learn to predict (discounted) future rewards!

Temporal difference learning (more formal)

Prediction Error
difference b/n outcome and predicted

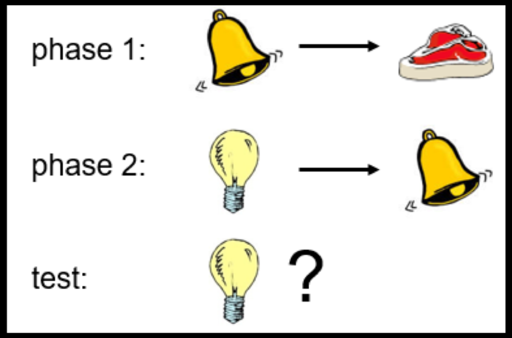
Outcome
Predicted value

Expectations
Predicted value

$$\delta_t = \overbrace{r_t + \boldsymbol{V(s_{t+1})}} - V(s_t)$$

$$V(s_t) = V(s_t) + \alpha\delta_t$$

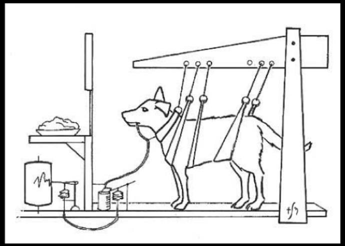
$$V(s_{t+1}) = sum[R_{future}]$$



$$\delta_t = r_t + \boldsymbol{V(s_{t+1})} - V(s_t)$$
$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$$

$\alpha = 1$

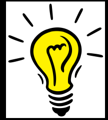
Pavlov's dog



V	V	V
0	0	1

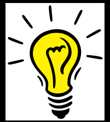
t = 0

r=0

$V(S_t) = V($$) = 0$

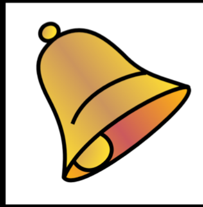
$V(S_{t+1}) = V($$) = 0$

$\delta = 0 + 0 - 0 = 0$

$V($$) = 0$

t = 1

r=0



$V(S_t) = V($$) = 0$

$V(S_{t+1}) = V($$) = 0$

$\delta = 0 + 0 - 0 = 0$

$V($$) = 0$

t = 2

r=1

$V(S_t) = V($$) = 0$

$V(S_{t+1}) =$ None

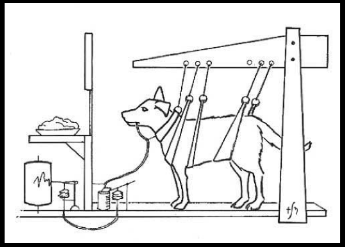
$\delta = 1 - 0 = 1$

$V($$) = 1$

Pavlov's dog

$$\delta_t = r_t + V(s_{t+1}) - V(s_t)$$
$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$$

$\alpha = 1$



V 	V 	V 
0	0	1
0	1	1

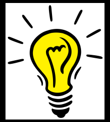
t = 0

r=0

$V(S_t) = V($  $) = 0$

$V(S_{t+1}) = V($  $) = 0$

$\delta = 0 + 0 - 0 = 0$

$V($  $) = 0$

t = 1

r=0

$V(S_t) = V($  $) = 0$

$V(S_{t+1}) = V($  $) = 1$

$\delta = 0 + 1 - 0 = 1$

$V($  $) = 1$

t = 2

r=1

$V(S_t) = V($  $) = 1$

$V(S_{t+1}) =$ None

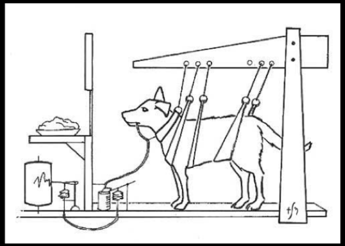
$\delta = 1 - 1 = 0$

$V($  $) = 1$

$$\delta_t = r_t + V(s_{t+1}) - V(s_t)$$
$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$$

$\alpha = 1$

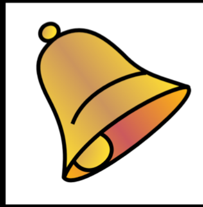
Pavlov's dog



r=0

t = 0

r=0



t = 1

r=1

t = 2

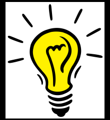
V 	V 	V 
0	0	1
0	1	1
1	1	1

t = 0

$V(S_t) = V($  $) = 0$

$V(S_{t+1}) = V($  $) = 1$

$\boxed{\delta = 0 + 1 - 0 = 1}$

$V($  $) = 1$

"bootstrapping"

t = 1

$V(S_t) = V($  $) = 1$

$V(S_{t+1}) = V($  $) = 1$

$\delta = 0 + 1 - 1 = 0$

$V($  $) = 1$

t = 2

$V(S_t) = V($  $) = 1$

$V(S_{t+1})$ = None

$\delta = 1 - 1 = 0$

$V($  $) = 1$

Uses own estimation of future reward to learn *even* in the absence of immediate reward

Consider a long sequence

$$\delta_t = r_t + \boldsymbol{V(s_{t+1})} - V(s_t)$$
$$V(s_t) = V(s_t) + \alpha\delta_t$$

If $V(s_{t+1}) = sum[R_{future}]$

You are guaranteed all the kibbles…100 time points from now

Consider a long sequence

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$

$$V(s_t) = V(s_t) + \alpha\delta_t$$

Solution: discount the value of future reward!

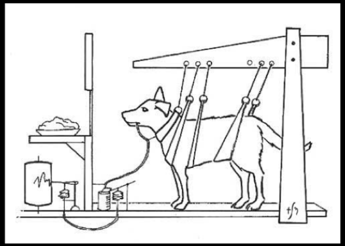
$V(s_t)$ = sum of discounted rewards

$= r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \ldots$

Pavlov's dog

$$\delta_t = r_t + \gamma \boldsymbol{V}(\boldsymbol{s_{t+1}}) - V(s_t)$$
$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$$

$\alpha = 1$
$\gamma = 0.5$



V 	V 	V 
0	0	1

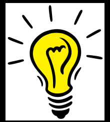
t = 0

r=0

$V(S_t) = V($  $) = 0$

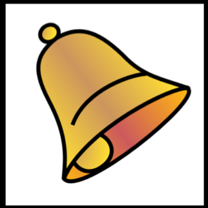
$V(S_{t+1}) = V($  $) = 0$

$\delta = 0 + (0.5)0 - 0 = 0$

$V($  $) = 0$

t = 1

r=0



$V(S_t) = V($  $) = 0$

$V(S_{t+1}) = V($  $) = 0$

$\delta = 0 + (0.5)0 - 0 = 0$

$V($  $) = 0$

t = 2

r=1

$V(S_t) = V($  $) = 0$

$V(S_{t+1})$ = None

$\delta = 1 - 0 = 1$

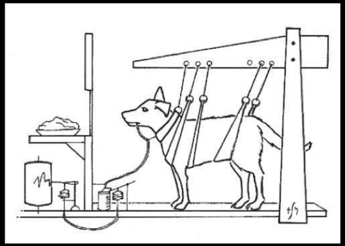
$V($  $) = 1$

$$\delta_t = r_t + \gamma \boldsymbol{V(s_{t+1})} - V(s_t)$$
$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$$

Pavlov's dog

$\alpha = 1$

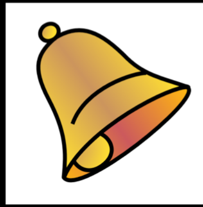
$\gamma = 0.5$



r=0

t = 0

r=0



t = 1

r=1

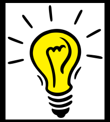
t = 2

V 	V 	V 
0	0	1
0	0.5	1

$V(S_t) = V($  $) = 0$

$V(S_{t+1}) = V($  $) = 0$

$\delta = 0 + (.5)0 - 0 = 0$

$V($  $)$ = 0

$V(S_t) = V($  $) = 0$

$V(S_{t+1}) = V($  $) = 1$

$\delta = 0 + .5(1) - 0 = .5$

$V($  $)$ = 0.5

$V(S_t) = V($  $) = 1$

$V(S_{t+1})$ = None

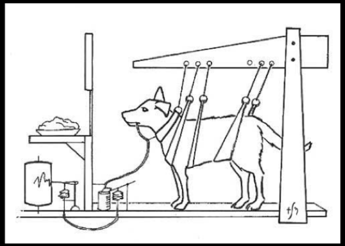
$\delta = 1 - 1 = 0$

$V($  $)$ = 1

Pavlov's dog

$$\delta_t = r_t + \gamma \boldsymbol{V(s_{t+1})} - V(s_t)$$
$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$$

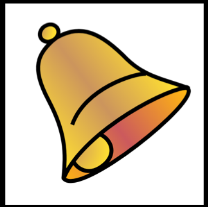
$\alpha = 1$
$\gamma = 0.5$



r=0

t = 0

r=0



t = 1

r=1

t = 2

V	V	V
0	0	1
0	0.5	1
.25	.5	1

$V(S_t) = V($$) = 0$

$V(S_{t+1}) = V($$) = 0.5$

$\delta = 0 + (.5).5 - 0 = .25$

$V($$)$ = .25

$V(S_t) = V($$) = 0.5$

$V(S_{t+1}) = V($$) = 1$

$\delta = 0 + (.5)1 - .5 = 0$

$V($$)$ = 0.5

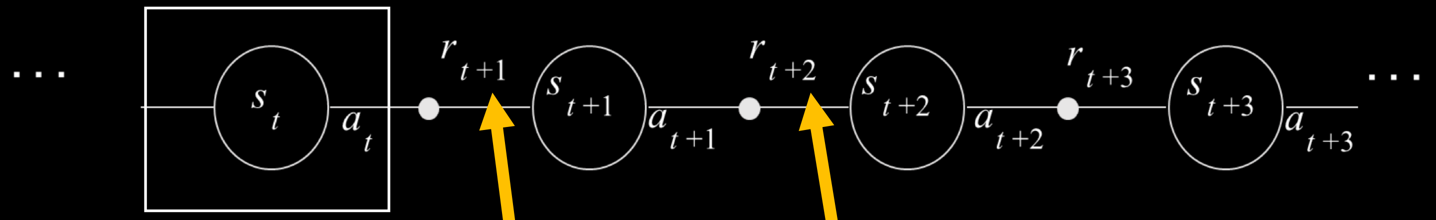
$V(S_t) = V($$) = 1$

$V(S_{t+1})$ = None

$\delta = 1 - 1 = 0$

$V($$)$ = 1

TD-Learning



- Value of state expressed in future expected rewards

$$V_t = E\left[\sum_{i=t}^{T} \gamma^{i-t} r_i\right] = E[\gamma^0\, r_t + \gamma^1 r_{t+1} + \gamma^2 r_{t+2}\ \ldots]$$

$$= E[\gamma^0\, r_t] + \gamma E[r_{t+1} + \gamma^1 r_{t+2}\ \ldots]$$

$$= E[r_t] + \gamma V_{t+1}$$

"Bellman Equation"

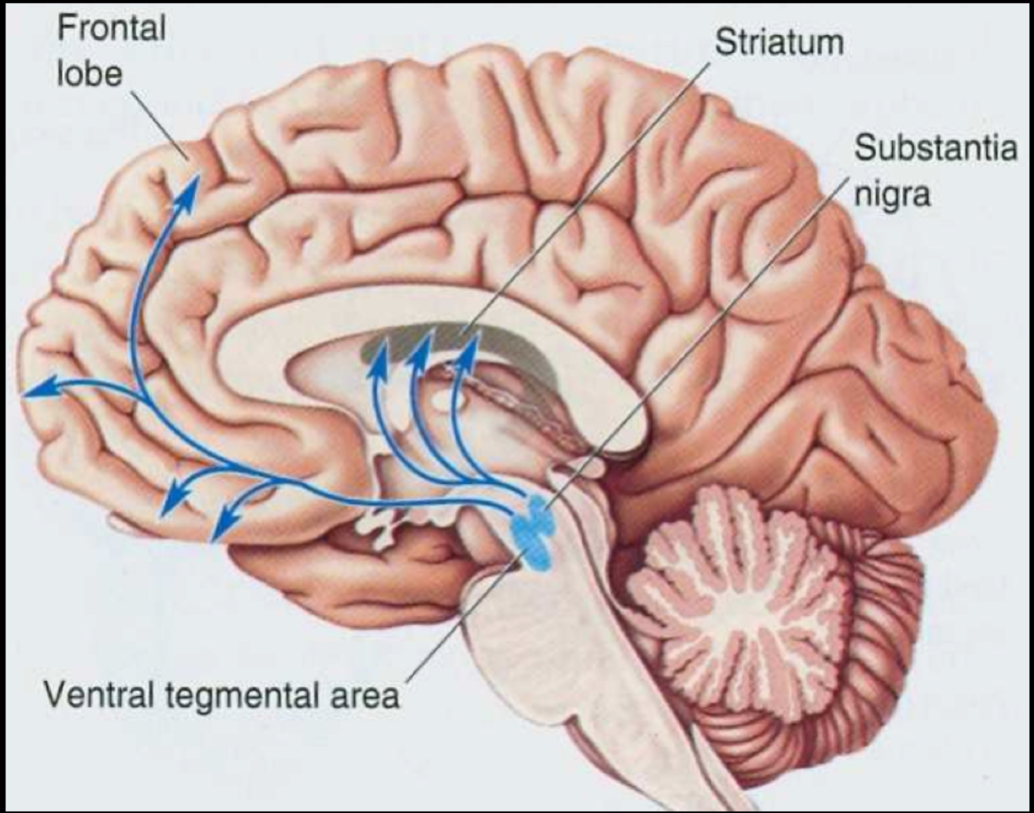
$$\delta_t = r_t + \gamma \hat{V}_{t+1}(s) - \hat{V}_t(s)$$

$$\hat{V}_{t+1}(s) = \hat{V}_t(s) + \alpha \delta_t$$

"temporal difference"

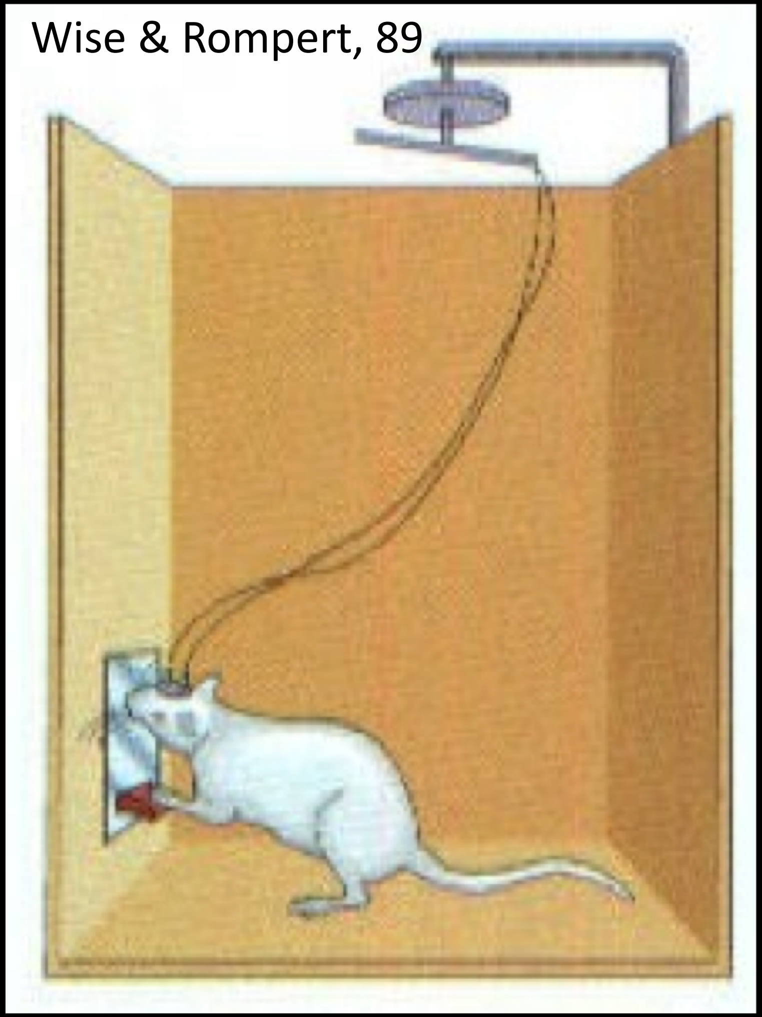
That’s cool, Alana, but what about the brain?

RL in the brain: Dopamine system



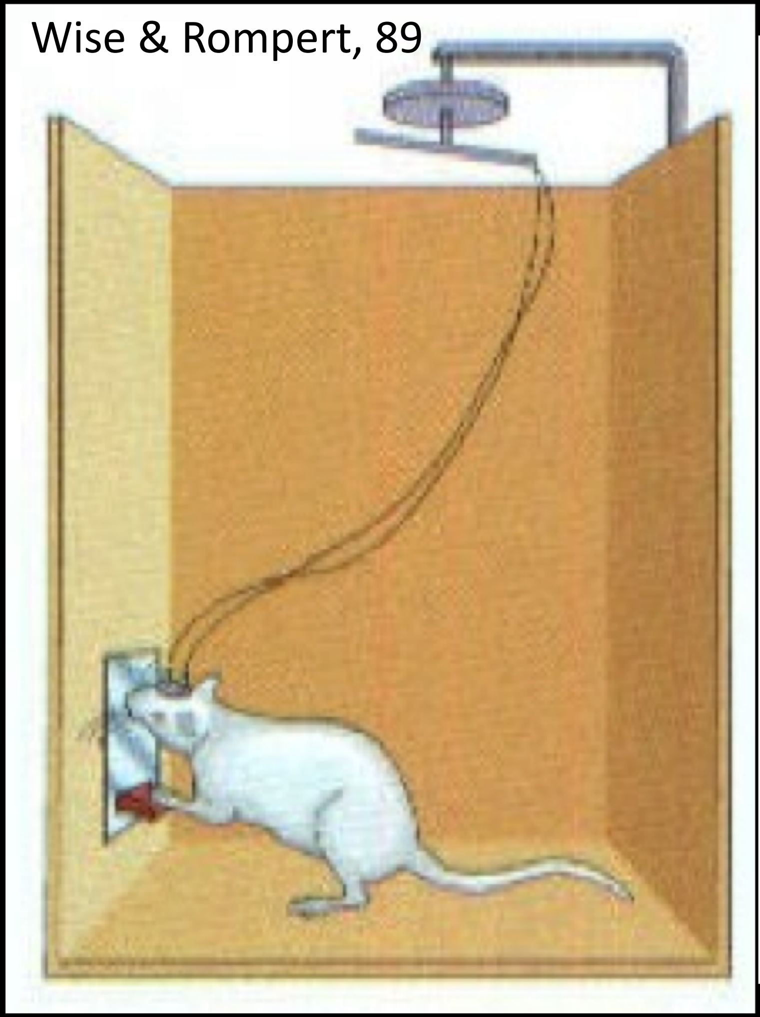
What is dopamine doing?

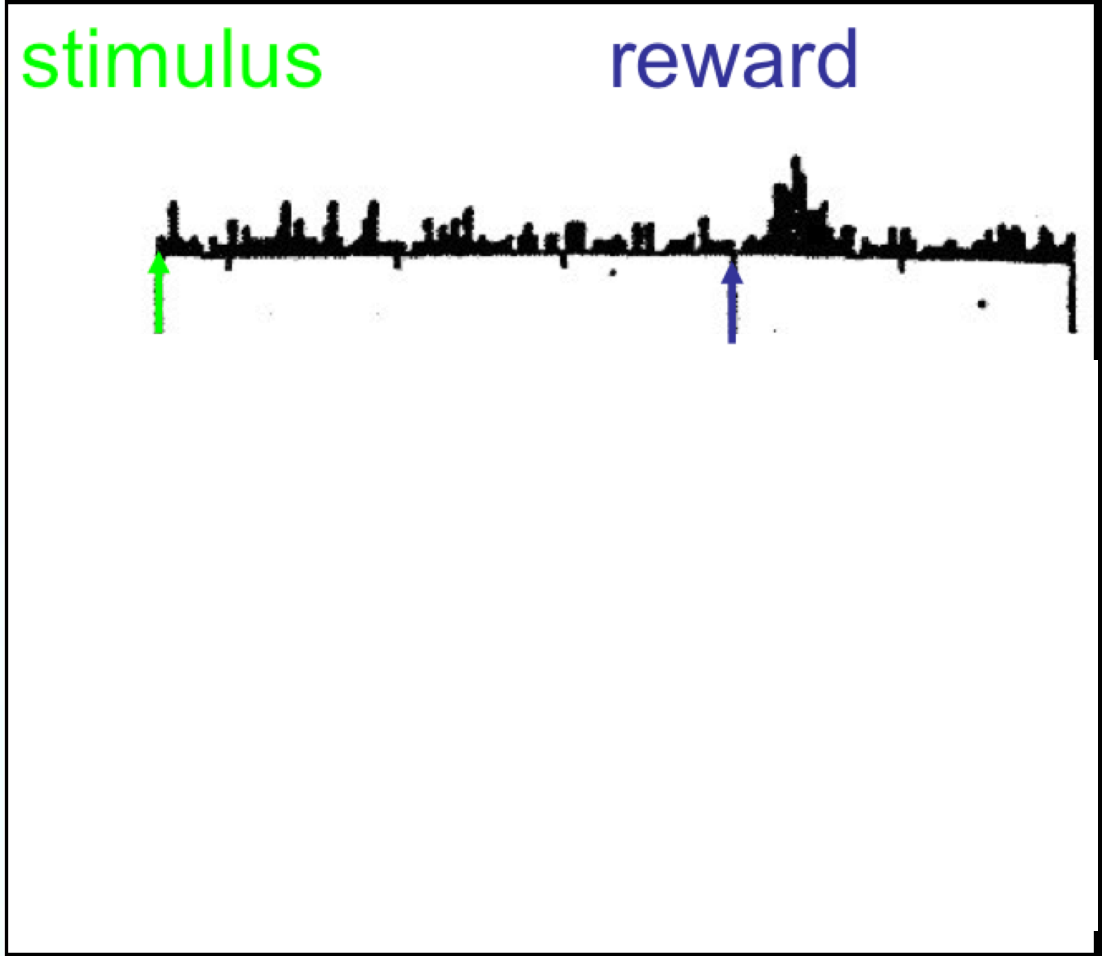
Dopamine carries the brain's reward signal



What is dopamine doing?

Dopamine carries the brain's reward signal

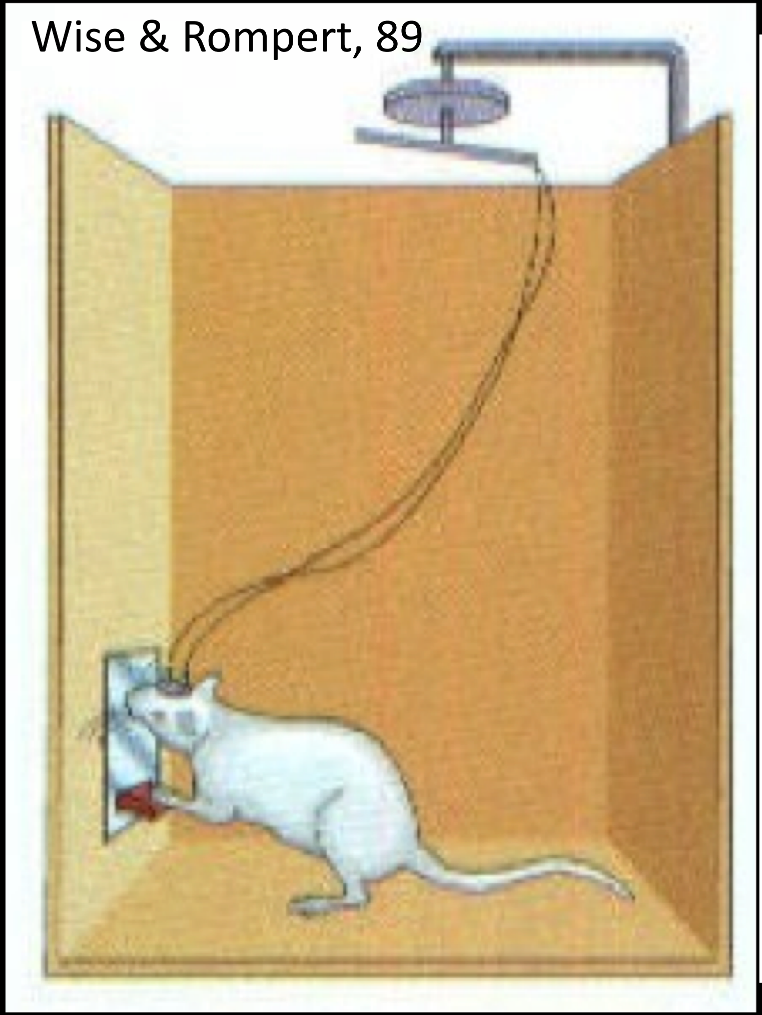


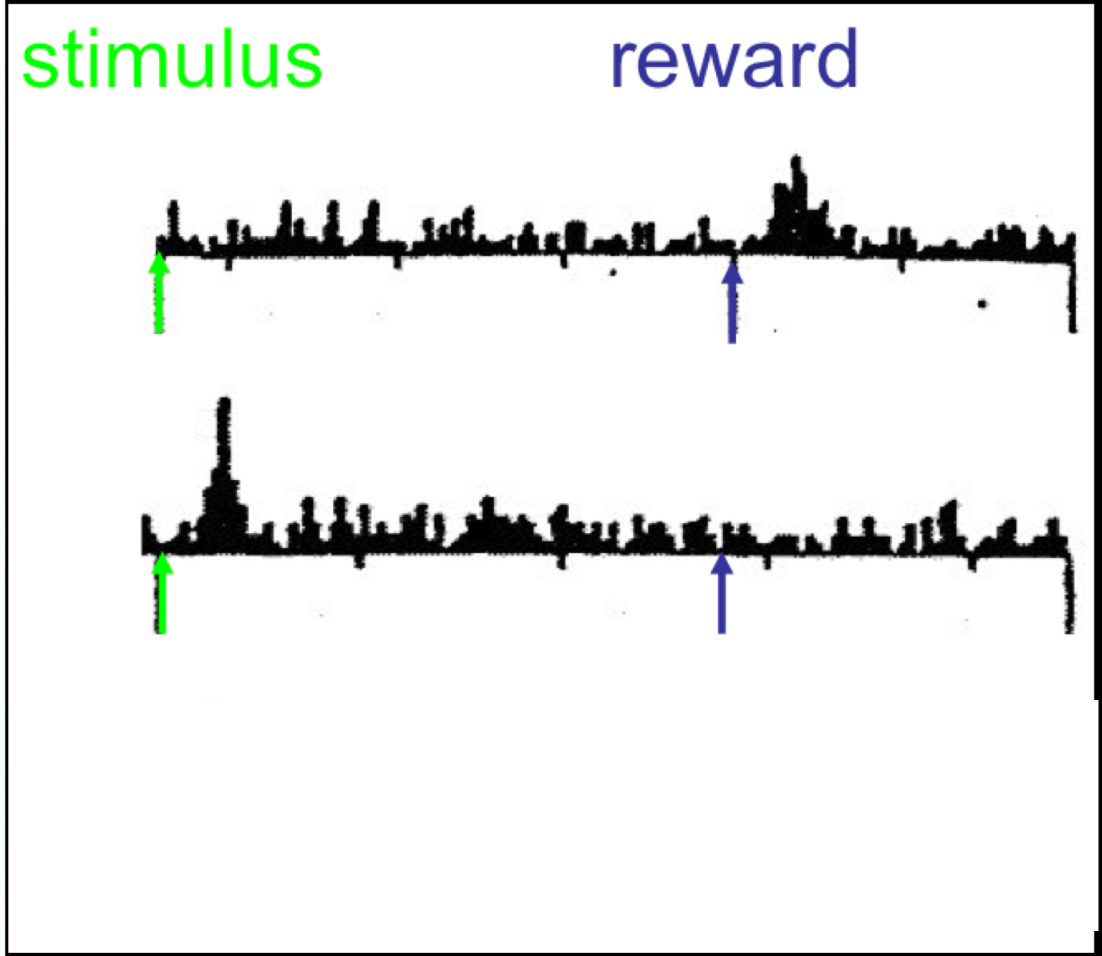


Schultz et. al, 98

What is dopamine doing?

Dopamine carries the brain's reward signal

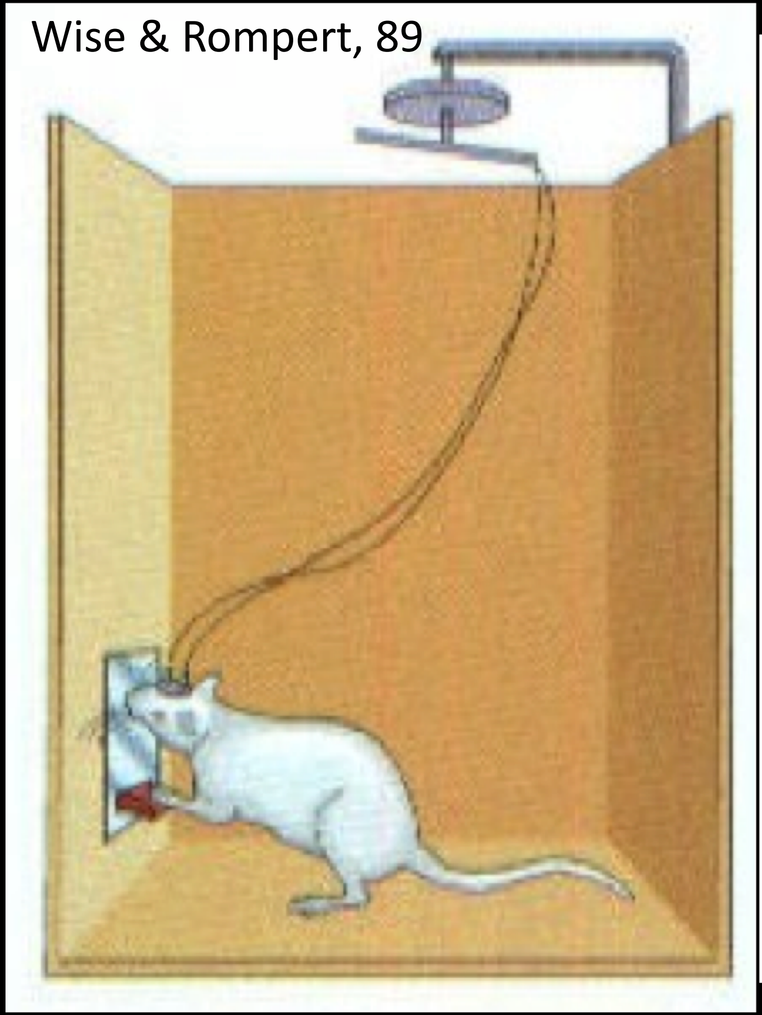


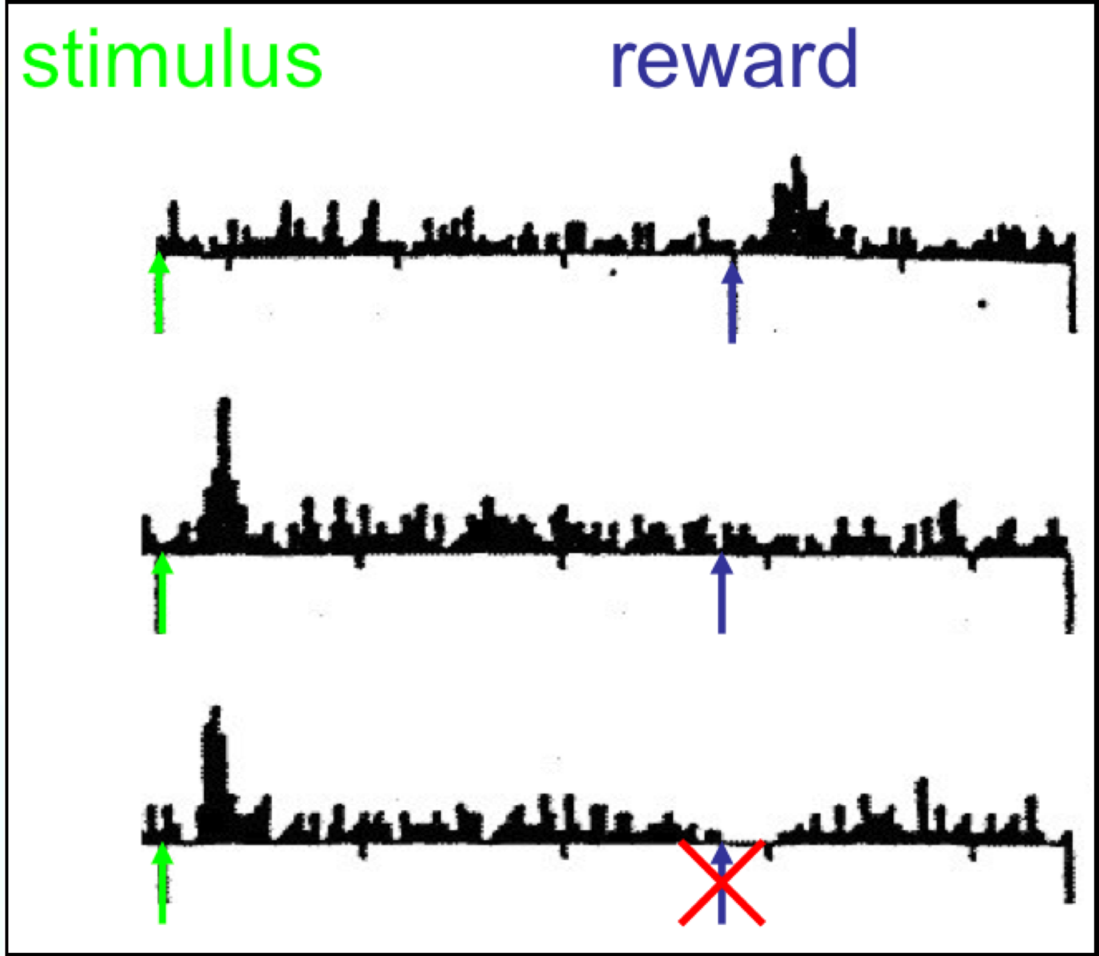


Schultz et. al, 98

What is dopamine doing?

Dopamine carries the brain's reward signal



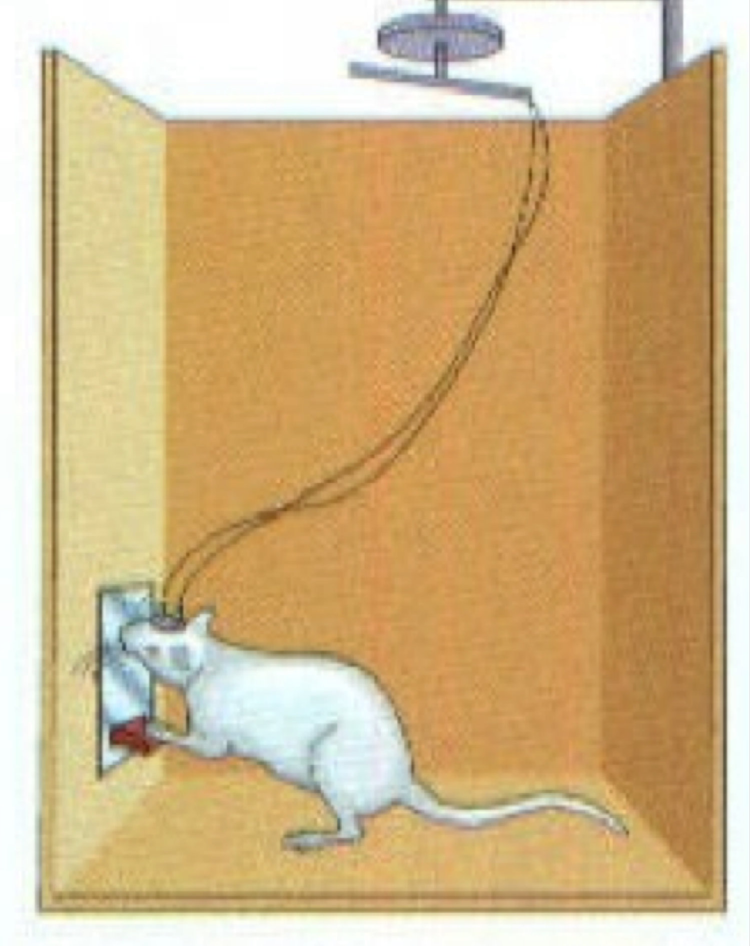


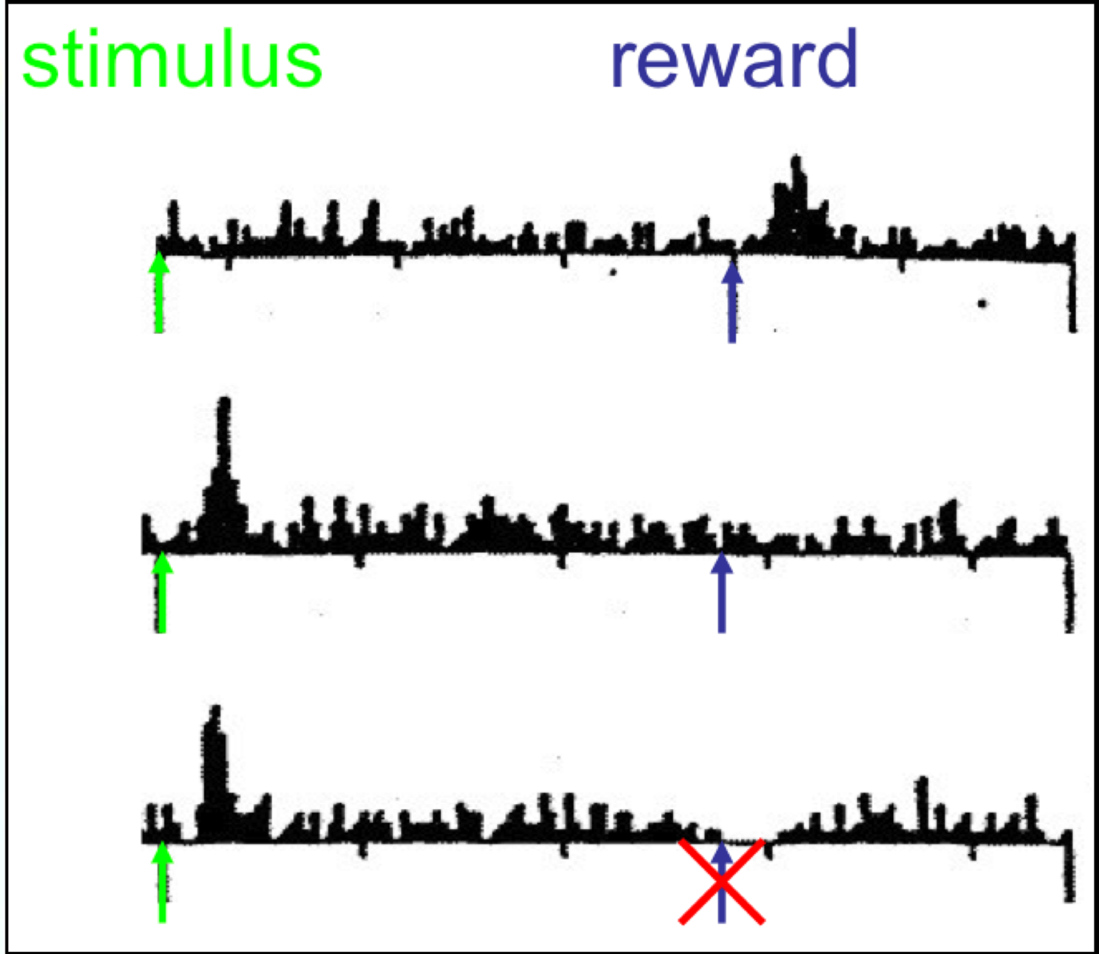
Schultz et. al, 98

What is dopamine doing?

Dopamine carries the brain's ~~reward~~ signal **reward prediction error**

Wise & Rompert, 89

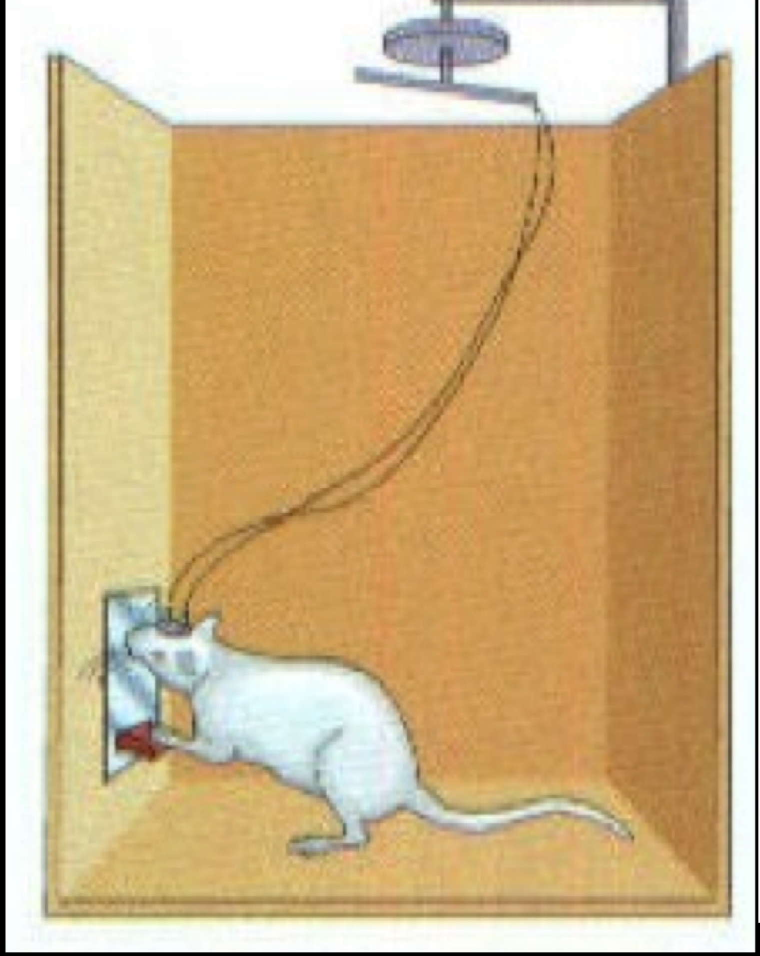


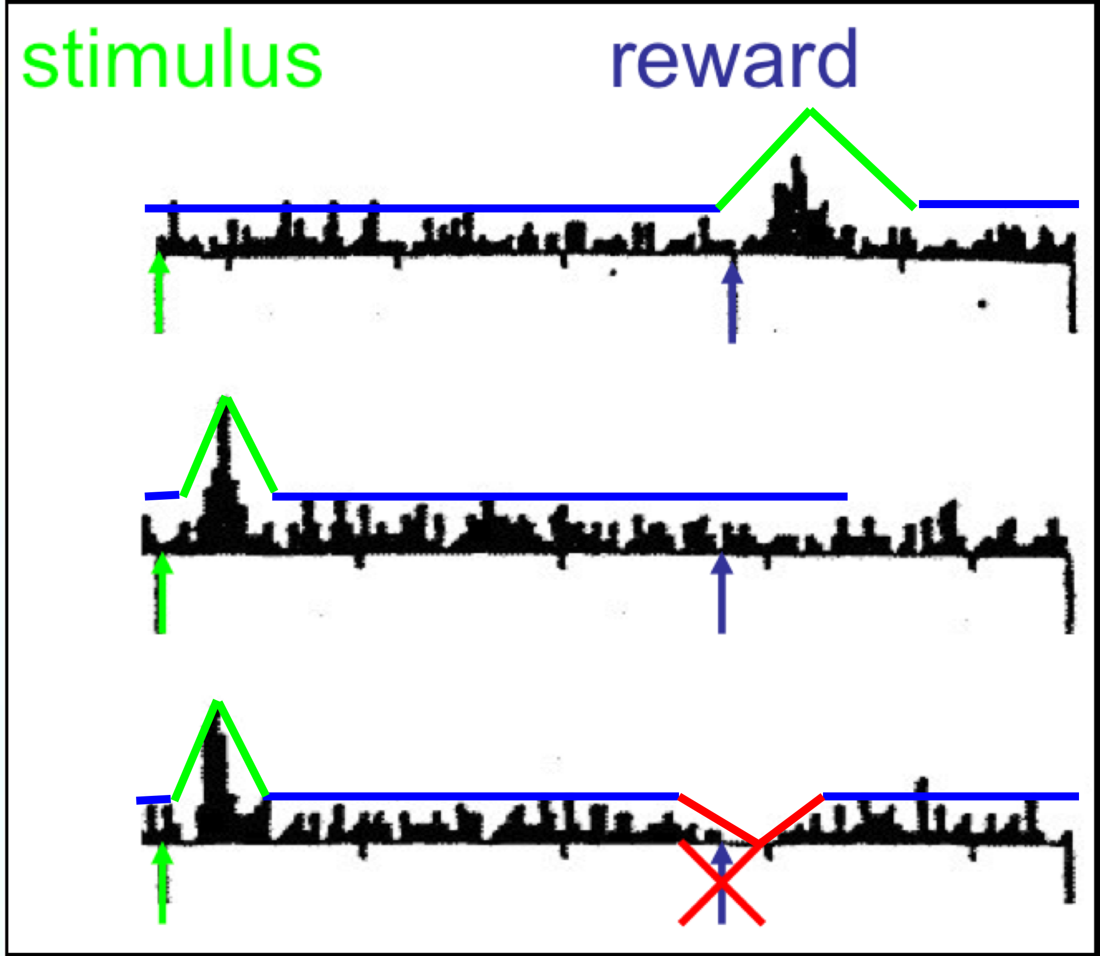


Schultz et. al, 98

What is dopamine doing?

Dopamine carries the brain's ~~reward~~ signal **reward prediction error**


Wise & Rompert, 89



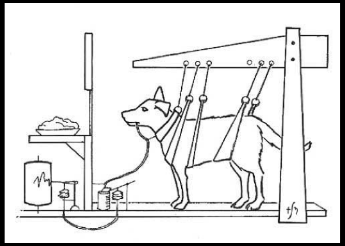
Schultz et. al, 98

$$\delta_t = r_t + \gamma \boldsymbol{V(s_{t+1})} - V(s_t)$$
$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$$

Pavlov's dog

$\alpha = 1$

$\gamma = 0.5$



V 	V 	V 
0	0	1
0	0.5	1
.25	.5	1

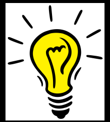
r=0

t = 0

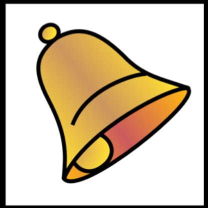
$V(S_t) = V($  $) = 0$

$V(S_{t+1}) = V($  $) = 0.5$

$\delta = 0 + (.5).5 - 0 = .25$

$V($  $) = .25$

r=0



t = 1

$V(S_t) = V($  $) = 0.5$

$V(S_{t+1}) = V($  $) = 1$

$\delta = 0 + (.5)1 - .5 = 0$

$V($  $) = 0.5$

r=1

t = 2

$V(S_t) = V($  $) = 1$

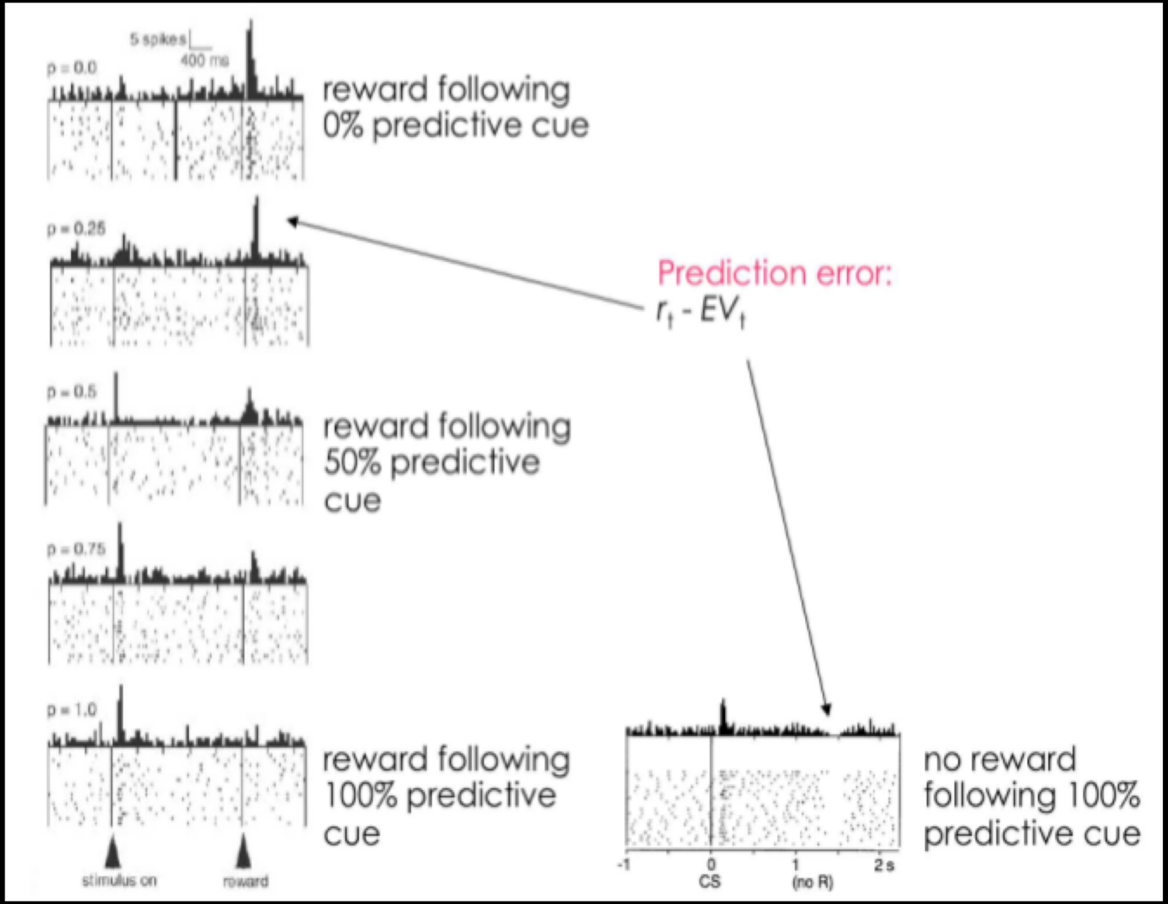
$V(S_{t+1}) = \text{None}$

$\delta = 1 - 0 = 0$

$V($  $) = 1$

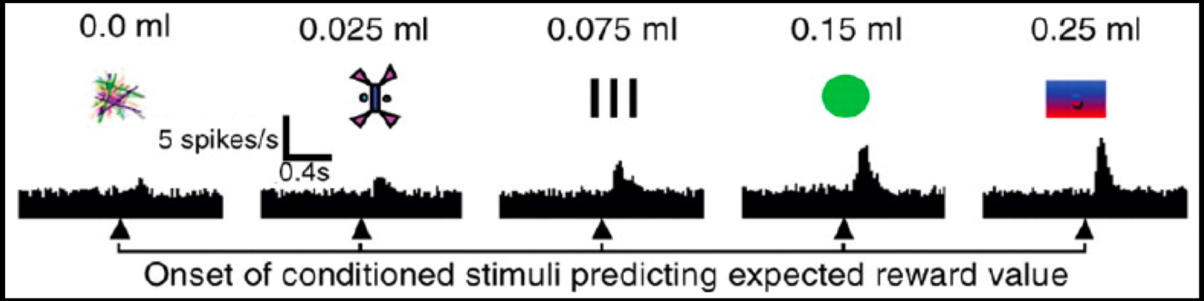
PE-hypothesis of DA

- Should be sensitive to probability of reward and magnitude of reward

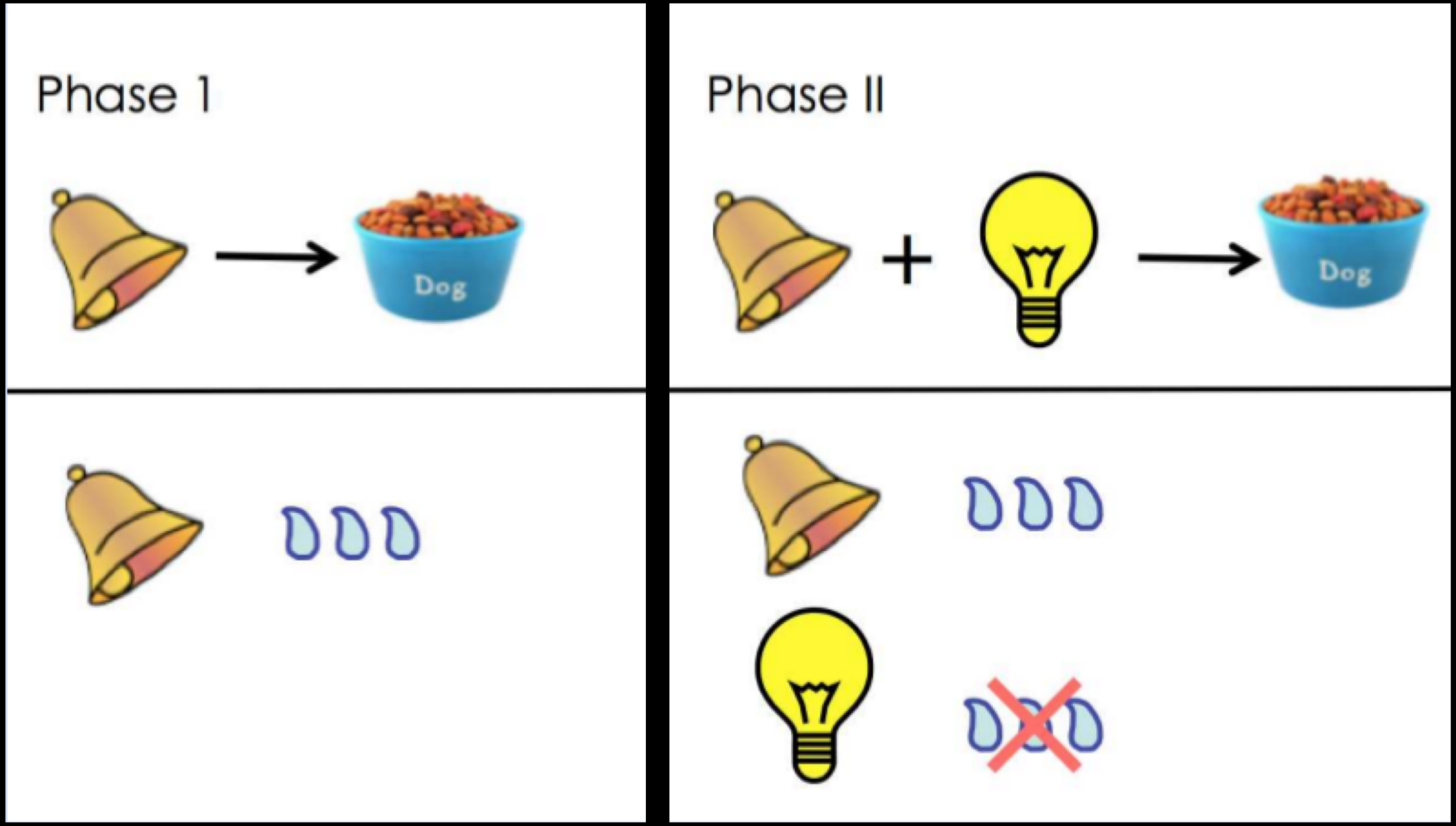


Fiorillo et al, 2003

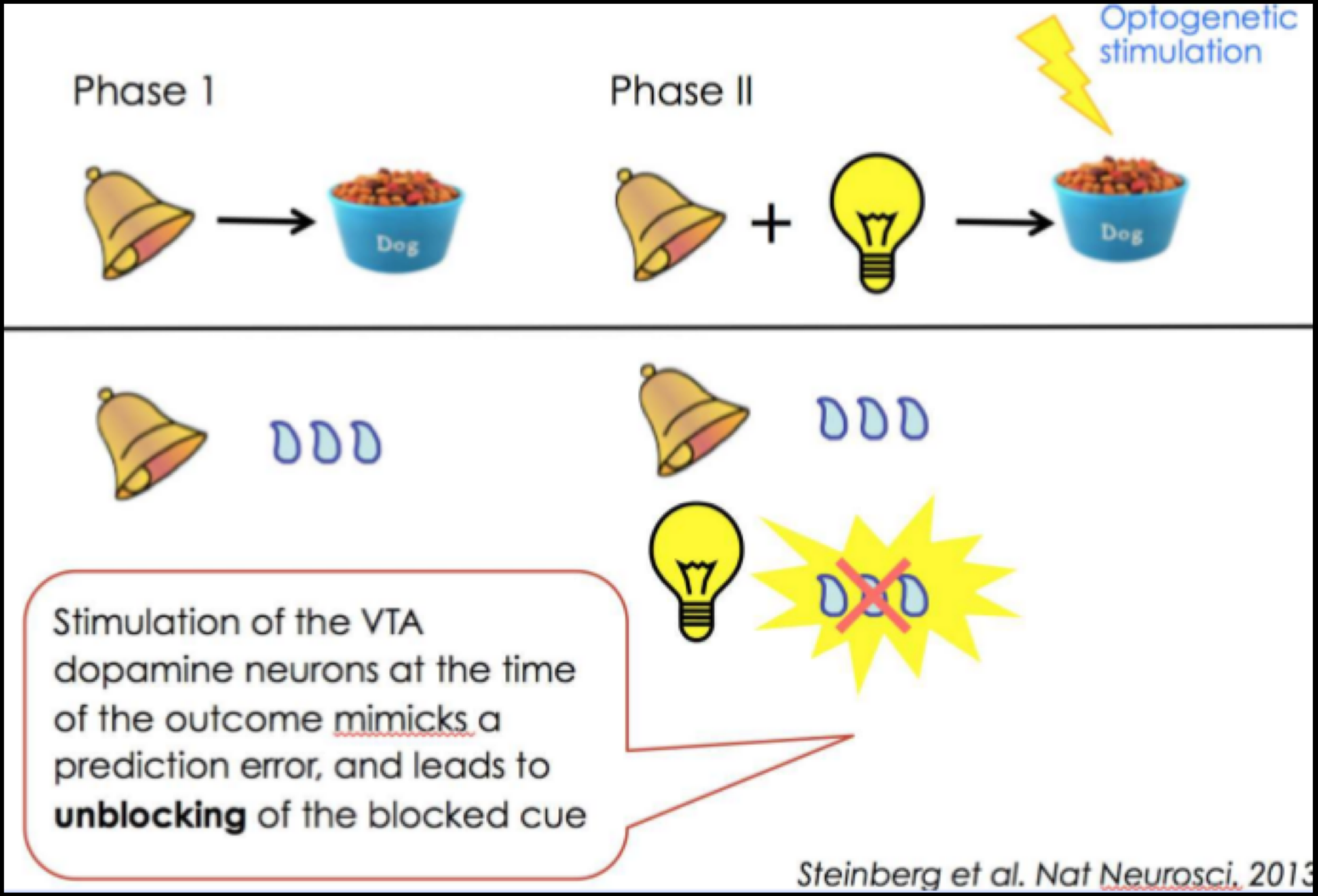
Tobler et al, 2005



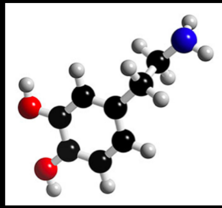
Blocking returns

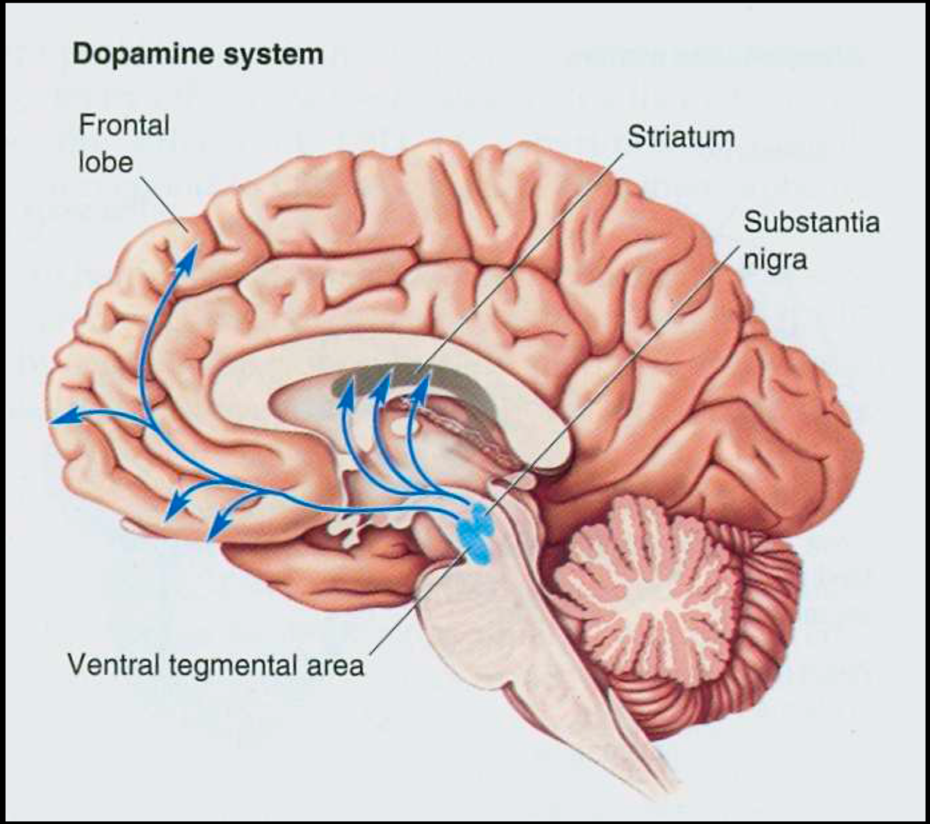


Blocking: DA induces learning



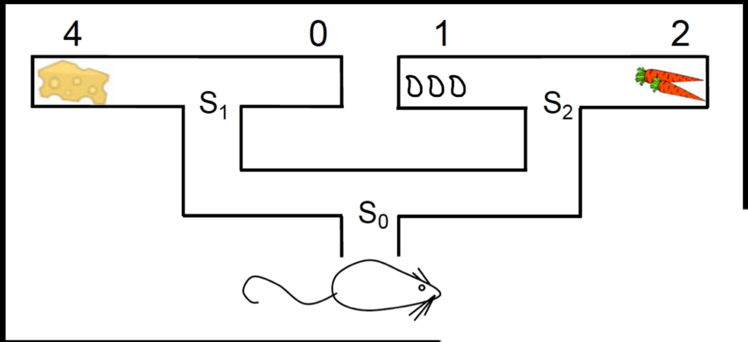
How are dopamine-based RPE signals used to select actions?





Will consider biological implementation in basal ganglia later

Q-Learning



The learning rate, i.e. that extent to which new information overrides old information. This is a number between 0 and 1.

The Q function we are updating, based on state s and action a at time t.

The reward earned when transitioning from time t to the next next turn, time t+1.

The value of the action that is estimated to return the largest (i.e. maximum) total future reward, based on all possible actions that can be made in the next state.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} + \lambda \max_{a} Q(s_{t+1}, a) - Q(s_t, a_t)\right]$$

The arrow operator means update the Q function to the left. This is saying, add the stuff to the right (i.e. the difference between the old and the new estimated future reward) to the existing Q value. This is equivalent in programming to A = A+B.

The discount rate. Determines how much future rewards are worth, compared to the value of immediate rewards. This is a number between 0 and 1.

The existing estimate of the Q function, (a.k.a. current the action-value).

Alternative - SARSA: takes into account actual choice on next time step, “on policy”

Learning which actions to take

Critic

V(s_t)

expected value of being in state s_t

Actor

Q(s_t,a_t)

preference (weight) for taking action **a_t** in state **s_t**

Analogy: Player/coach

Learning which actions to take

Critic

$V(s_t)$

expected value of being in state s_t

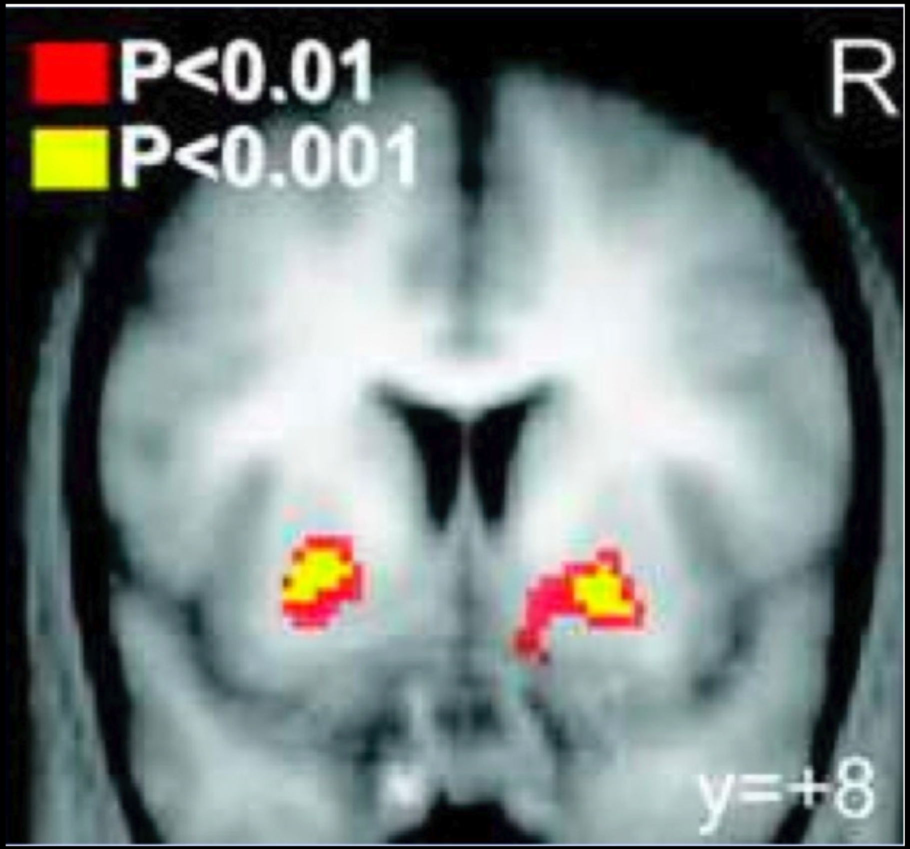
Actor

$Q(s_t,a_t)$

preference (weight) for taking action a_t in state s_t

1) $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$
2) $V(s_t) \leftarrow V(s_t) + \delta_t$
3) $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \delta_t$

fMRI & reward prediction errors



Ventral striatum correlates with reward prediction error: Critic!
Dorsal striatum correlates when actor involved

O'Doherty et al, 2004

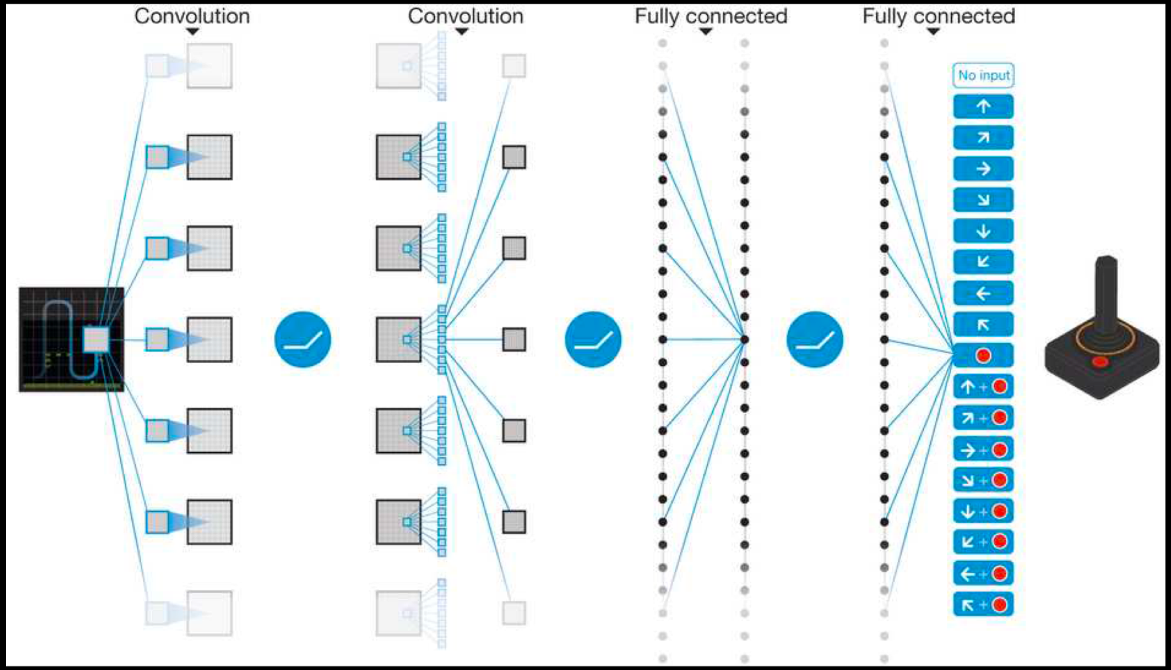
Google Deep Mind RL Network ("DQN") Plays Atari

LETTER

doi:10.1038/nature14236

Human-level control through deep reinforcement learning

Volodymyr Mnih[1]*, Koray Kavukcuoglu[1]*, David Silver[1]*, Andrei A. Rusu[1], Joel Veness[1], Marc G. Bellemare[1], Alex Graves[1], Martin Riedmiller[1], Andreas K. Fidjeland[1], Georg Ostrovski[1], Stig Petersen[1], Charles Beattie[1], Amir Sadik[1], Ioannis Antonoglou[1], Helen King[1], Dharshan Kumaran[1], Daan Wierstra[1], Shane Legg[1] & Demis Hassabis[1]



Assume you're a brain - revisited

You still need to eat things, you still don't want to be eaten by other things, and you'd still "like" to produce more brains.

1. *What is learning?*
2. *Why is learning important?*
3. *What should you learn?*
4. *When should you learn?*