Reinforcement Learning

Odelia Schwartz
2016

Forms of learning?

Forms of learning

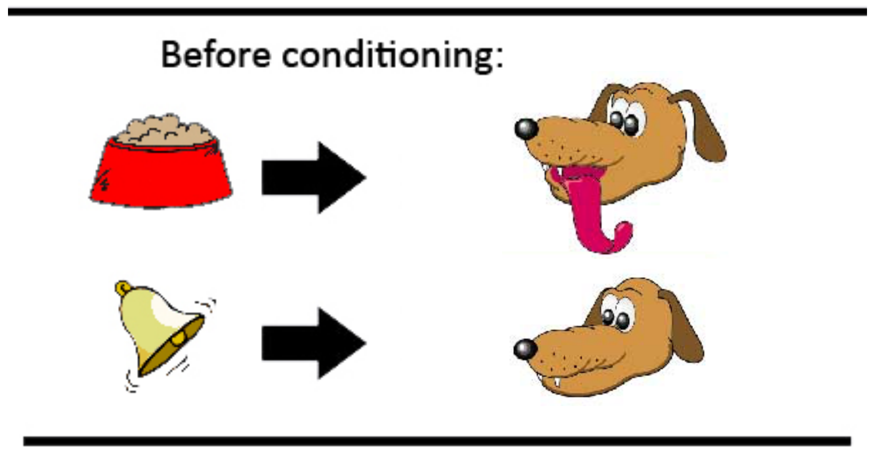
- Unsupervised learning
- Supervised learning
- Reinforcement learning

Forms of learning

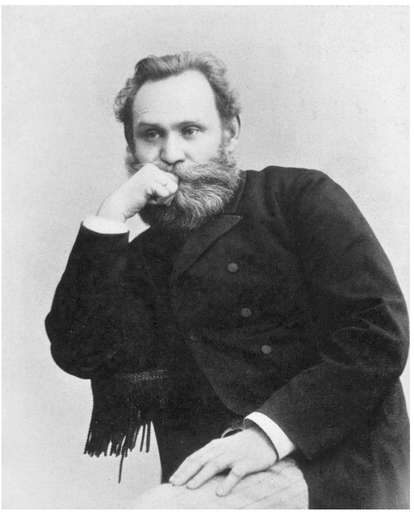
- Unsupervised learning
- Supervised learning
- Reinforcement learning

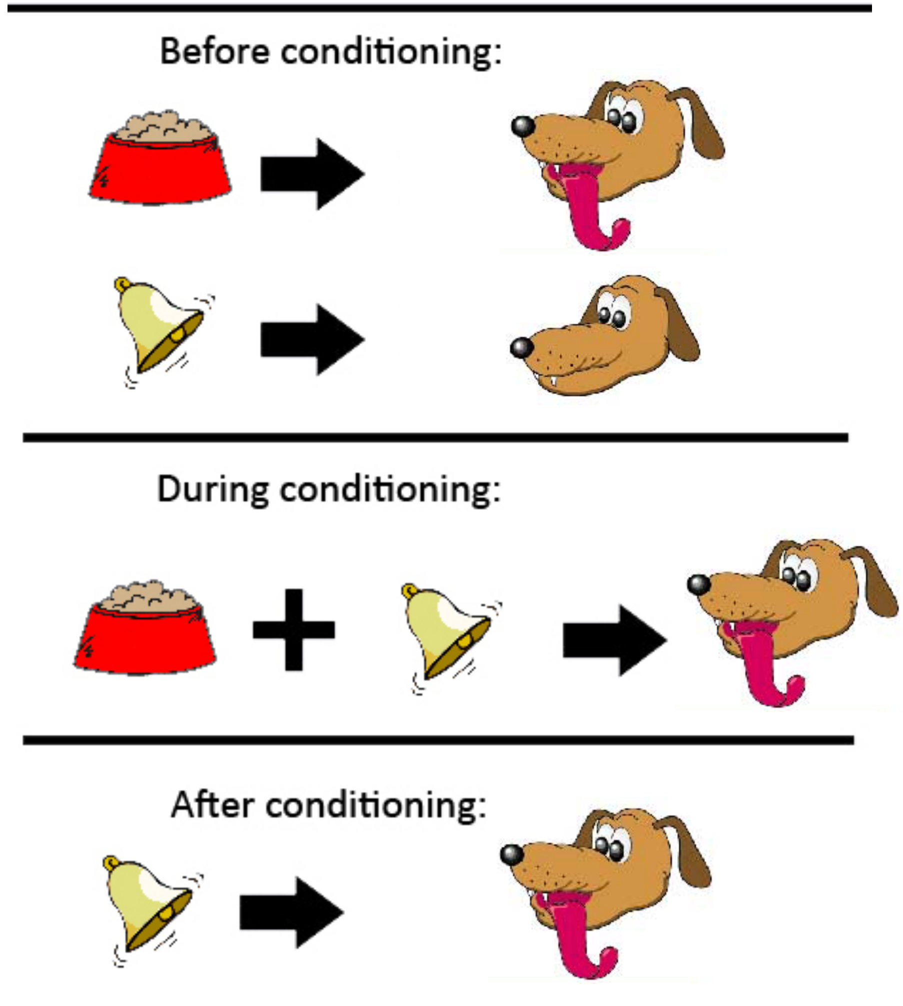
Another active field that combines computation, machine learning, neurophysiology, fMRI

Pavlov and classical conditioning



Pavlov and classical conditioning





Modern terminology

- Stimuli
- Rewards
- Expectations of reward: behavior is learned based on expectations of reward
- Can learn based on consequences of actions (instrumental conditioning); can learn whole sequence of actions (example: maze)

Rescorla-Wagner rule (1972)

- Can describe classical conditioning and range of related effects
- Based on simple linear prediction of reward associated with a stimulus (error based learning)
- Includes weight updating as in the perceptron rule we did in lab, but we learn from error in predicting reward

Rescorla-Wagner rule (1972)

- Minimize difference between received reward and predicted reward
- Binary variable u (1 if stimulus is present; 0 if absent)
- Predicted reward v
- Linear weight w

$$v = wu$$

- If stimulus u is present:

$$v = w$$

based on Dayan and Abbott book

Rescorla-Wagner rule (1972)

- Minimize squared error between received reward r and predicted reward v:

$$(r - v)^2$$

based on Dayan and Abbott book

Rescorla-Wagner rule (1972)

- Minimize squared error between received reward r and predicted reward v:

$$(r - v)^2$$

In Niv and Schoenbaum 2009

Rescorla-Wagner rule (1972)

- Minimize squared error between received reward r and predicted reward v:

$$(r-v)^2$$

(average over presentations of stimulus and reward)

- Update weight:

$$w \rightarrow w + \varepsilon(r-v)u$$

ε learning rate

Also known as delta learning rule: $\delta = r - v$

- Update weight:

$$w \rightarrow w + \varepsilon(r - v)u$$

- Simpler notation: if a stimulus is presented at trial n (we'll just take u as 1 and set v to w):

$$v_{n+1} = v_n + \epsilon(r_n - v_n)$$

based on Dayan and Abbott book

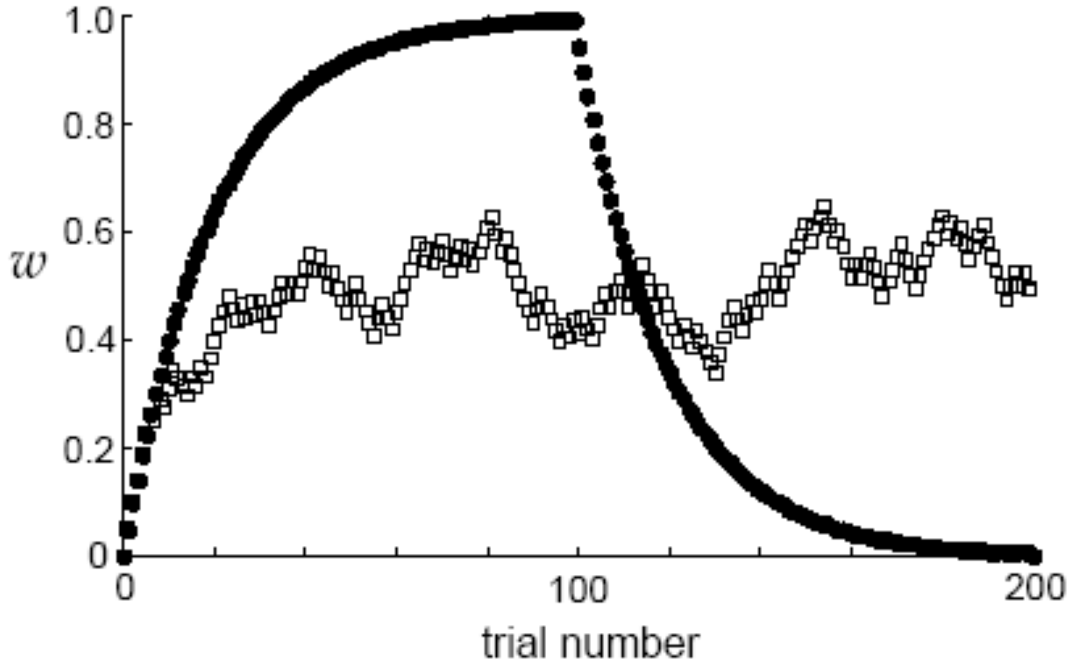
- So if a stimulus is presented at trial n:

$$v_{n+1} = v_n + \epsilon(r_n - v_n)$$

- What happens when learning rate = 1?

- What happens when it is smaller than 1?

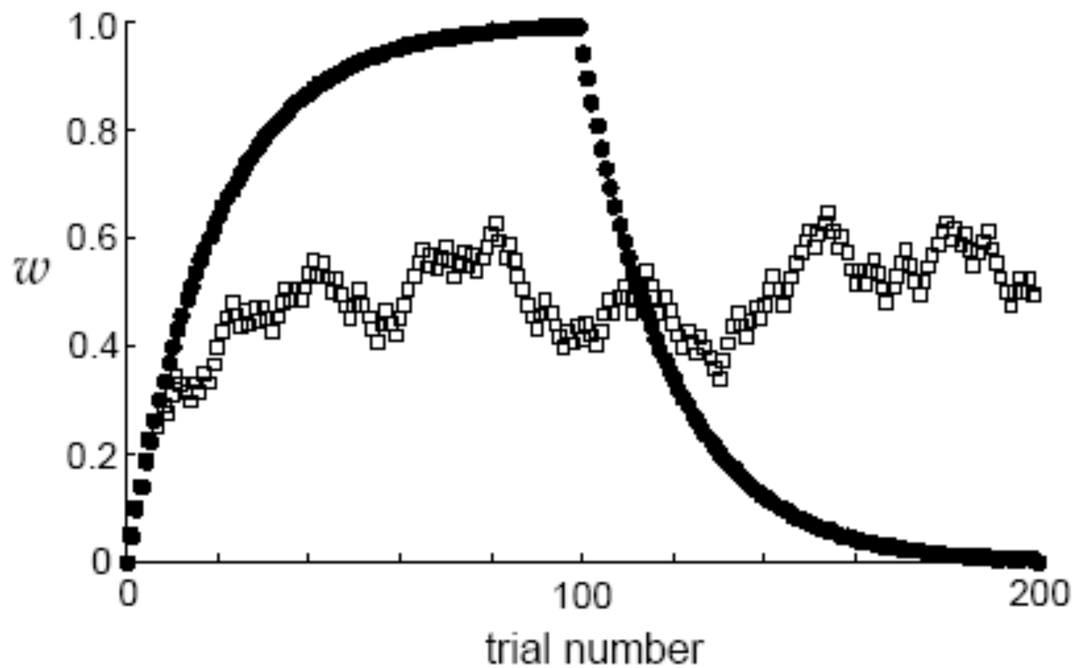
Acquisition and extinction



- Solid: First 100 trials: reward (r=1) paired with stimulus; next 100 trials no reward (r=0) paired with stimulus (learning rate .05)
- Dashed: Reward paired with stimulus randomly 50 percent of time

From Dayan and Abbott book

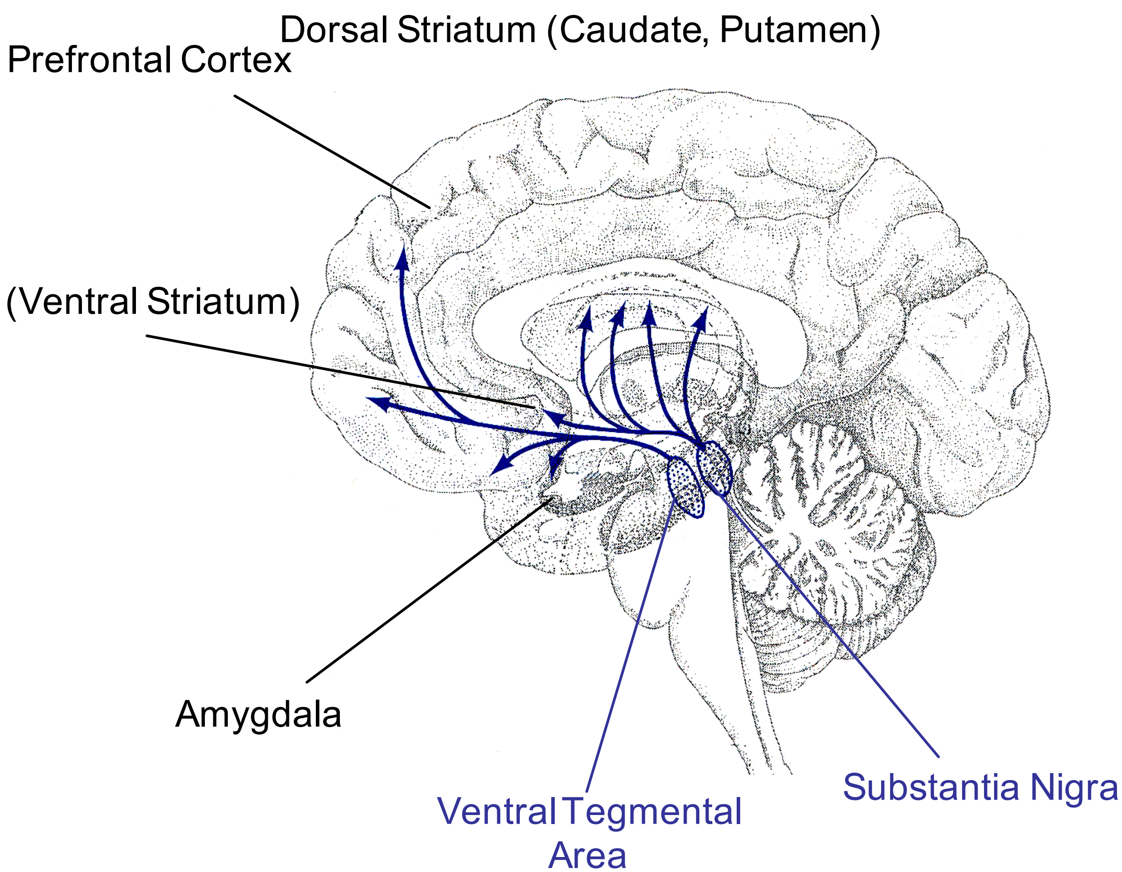
Acquisition and extinction



- Curves show w over time
- What is the predicted reward v and the error (r-v)?

From Dayan and Abbott book

Dopamine areas



From Dayan slides

Dopamine roles?

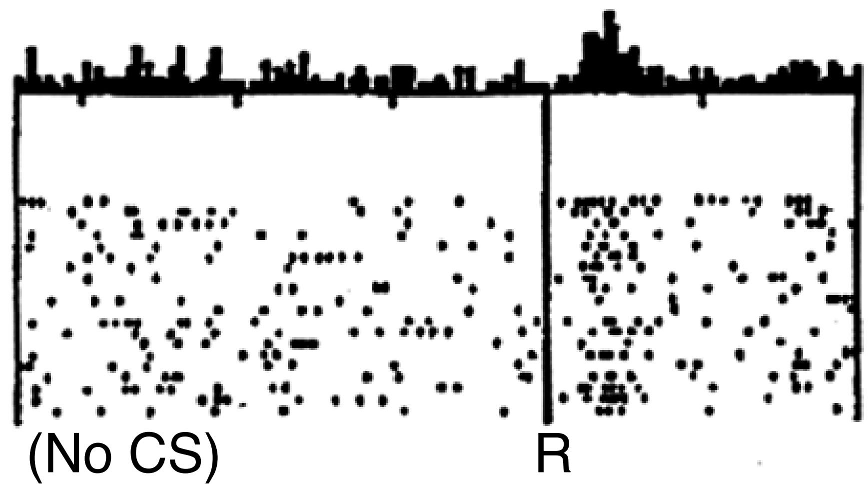
Dopamine roles?

Associated with…

- reward (we’ll see prediction error)
- self-stimulation
- motor control (initiation)
- addiction

VTA Activity of dopaminergic neurons

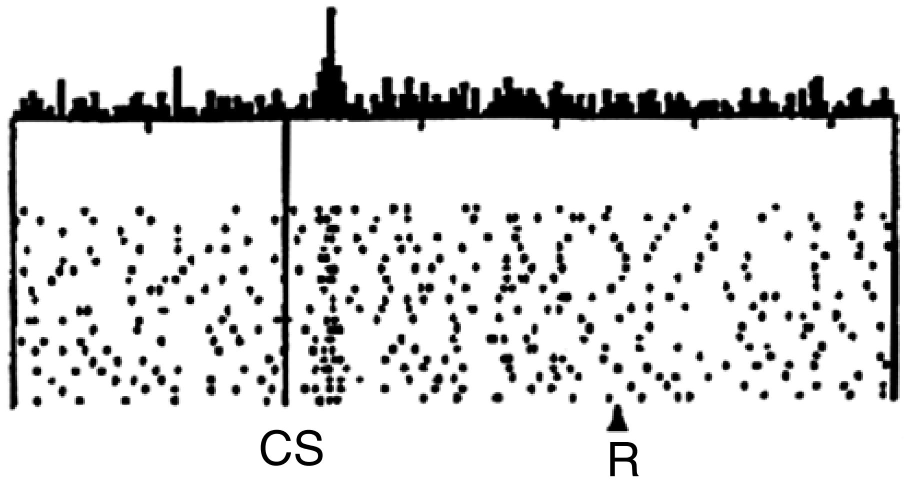
No prediction
Reward occurs



Before learning, reward is given in experiment, but animal does not predict (expect) reward (why is there increased activity after reward?)

Schultz, Dayan, Montague, 1997

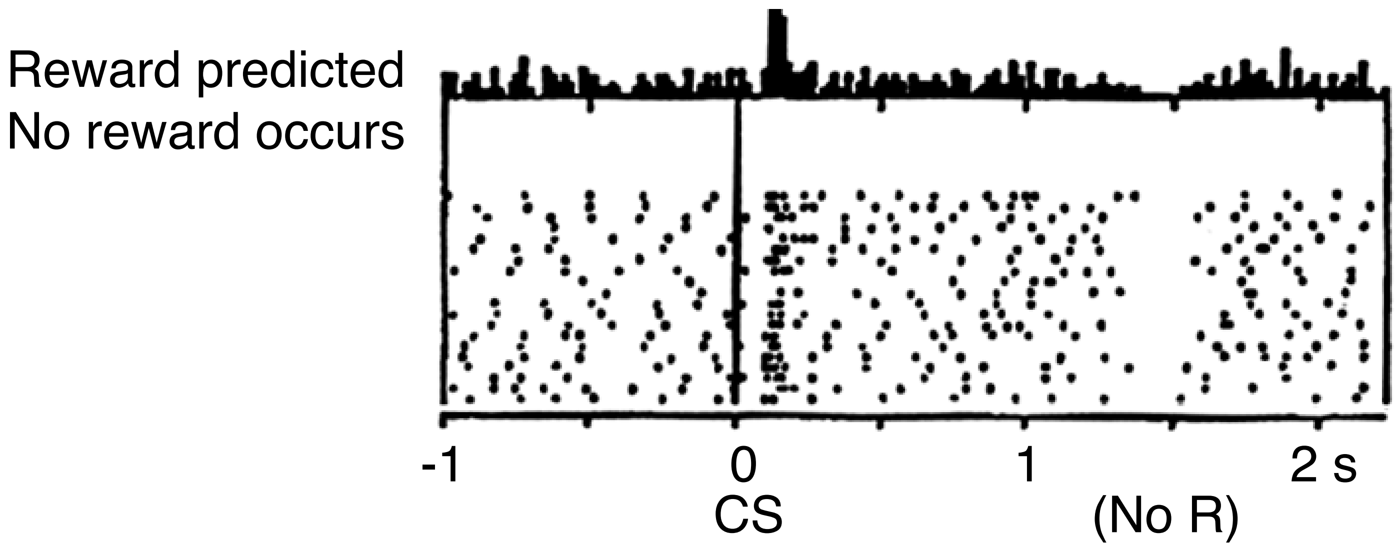
VTA Activity of dopaminergic neurons



After learning, conditioned stimulus predicts reward, and reward is given in experiment (why is activity fairly uniform after reward?)

Schultz, Dayan, Montague, 1997

VTA Activity of dopaminergic neurons

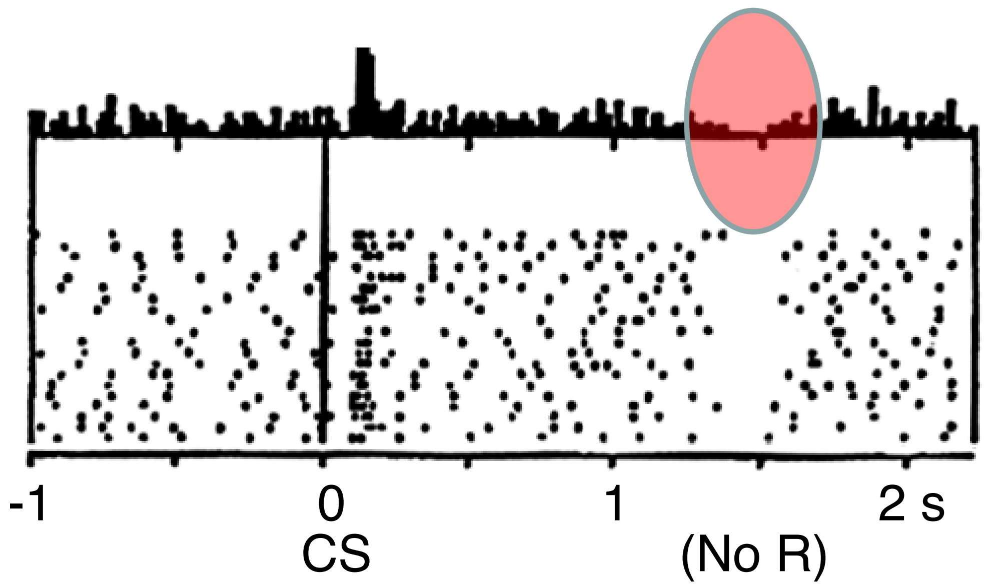


After learning, conditioned stimulus predicts reward so there is an expectation of reward, but no reward is given in the experiment

Schultz, Dayan, Montague, 1997

VTA Activity of dopaminergic neurons

Reward predicted
No reward occurs

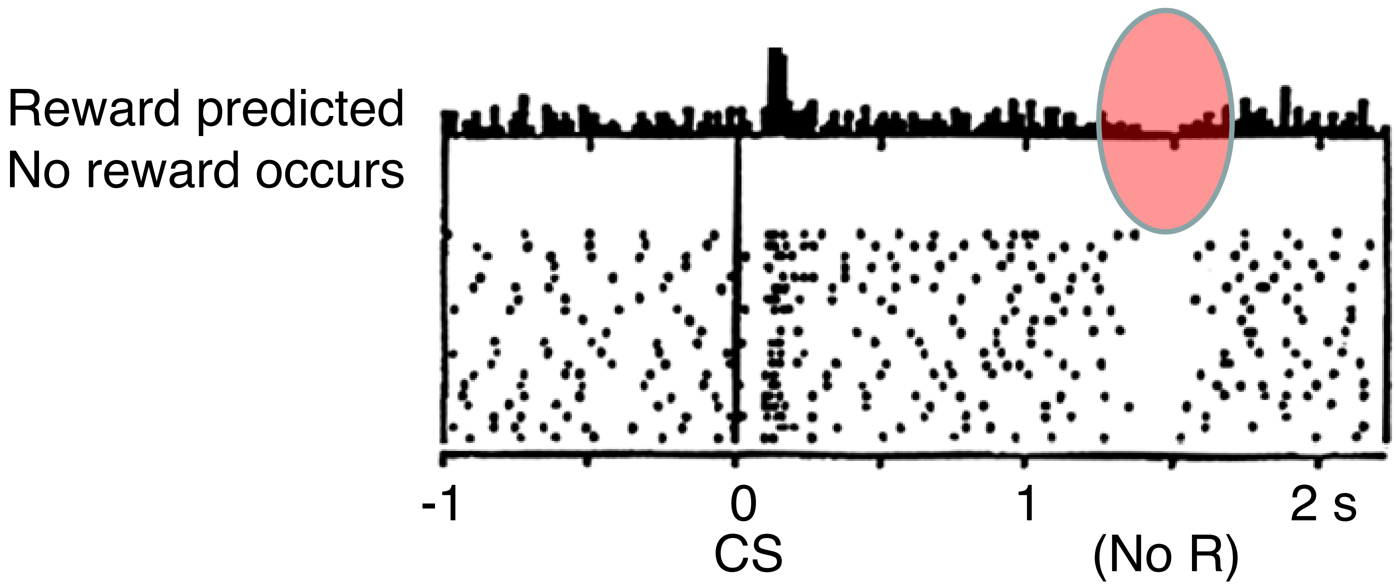


After learning, conditioned stimulus predicts reward so there is an expectation of reward, but no reward is given in the experiment

Why is there a dip? What are these neurons doing?

Schultz, Dayan, Montague, 1997

VTA Activity of dopaminergic neurons

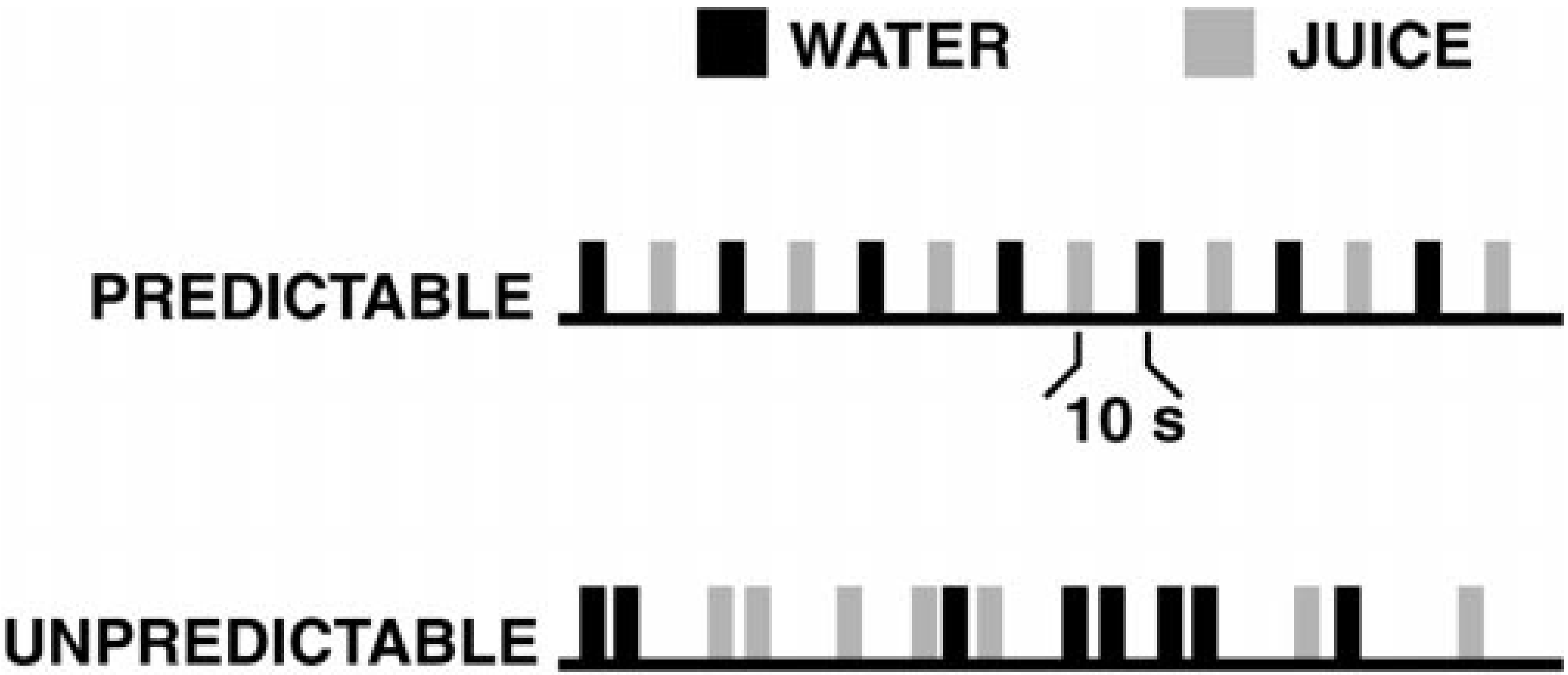


After learning, conditioned stimulus predicts reward so there is an expectation of reward, but no reward is given in the experiment

What are these neurons doing? Prediction error between actual and predicted reward (like r-v)

Schultz, Dayan, Montague, 1997

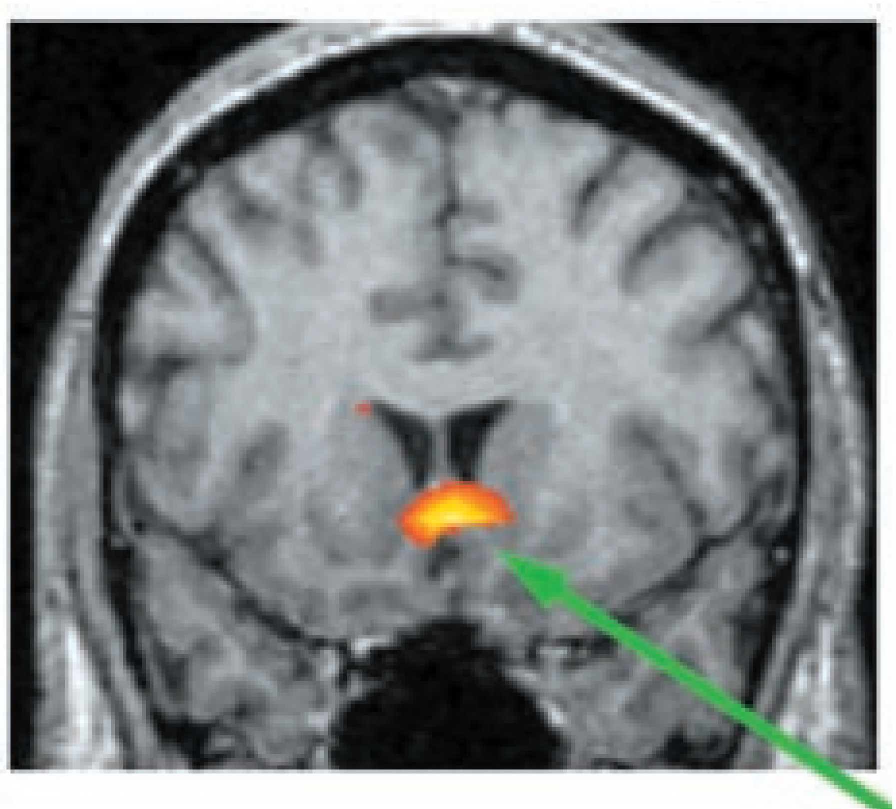
fMRI



Berns et al., 2001

BOLD fMRI signal

Juice Unpredictable - Predictable

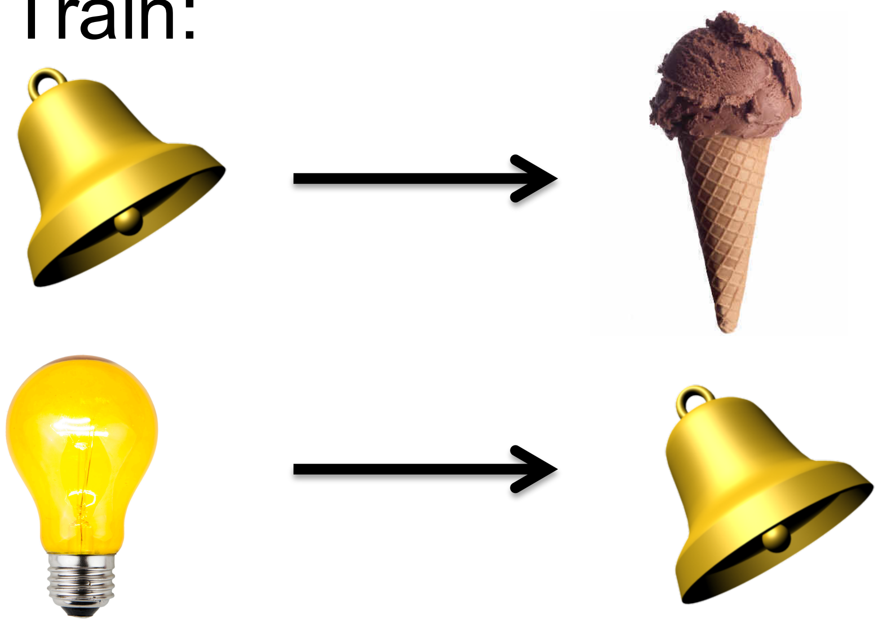


Greater BOLD response to unpredictable stimuli (in striatum)

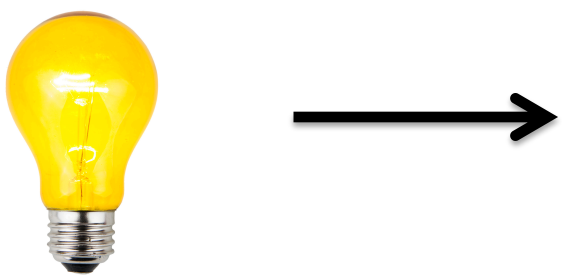
Berns et al., 2001; see also O'Doherty et al. 2003

Shortcomings of Rescorla-Wagner: Example: secondary conditioning

Train:



Test:

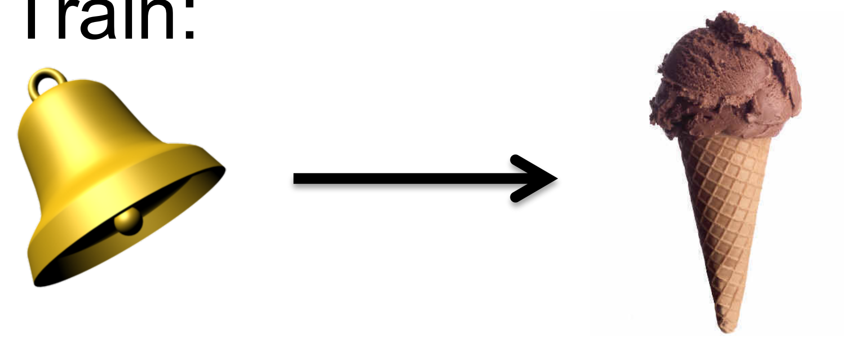


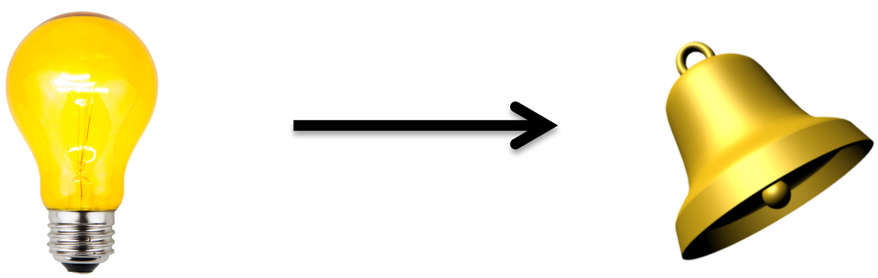
??

Based on Peter Dayan slides

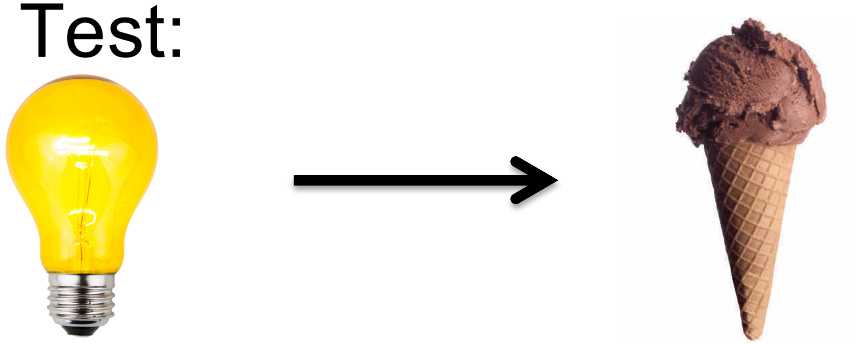
Shortcomings of Rescorla-Wagner: Example: secondary conditioning

Train:





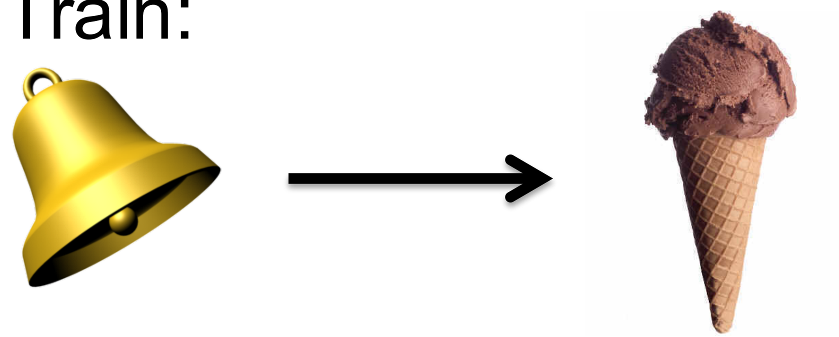
Test:

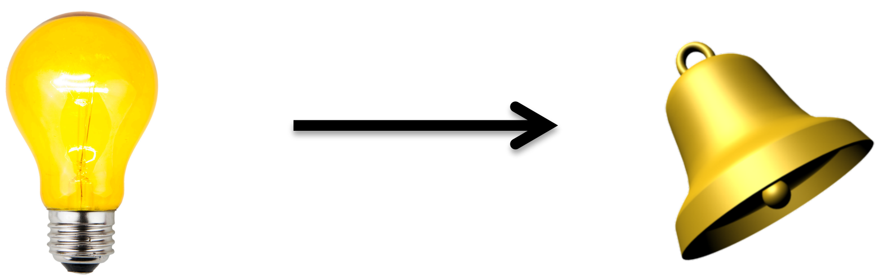


Animals learn (more generally, actions that lead to longer term rewards)

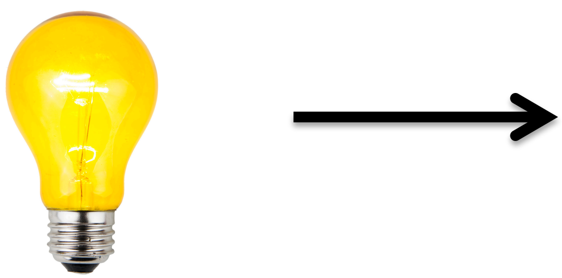
Shortcomings of Rescorla-Wagner: Example: secondary conditioning

Train:



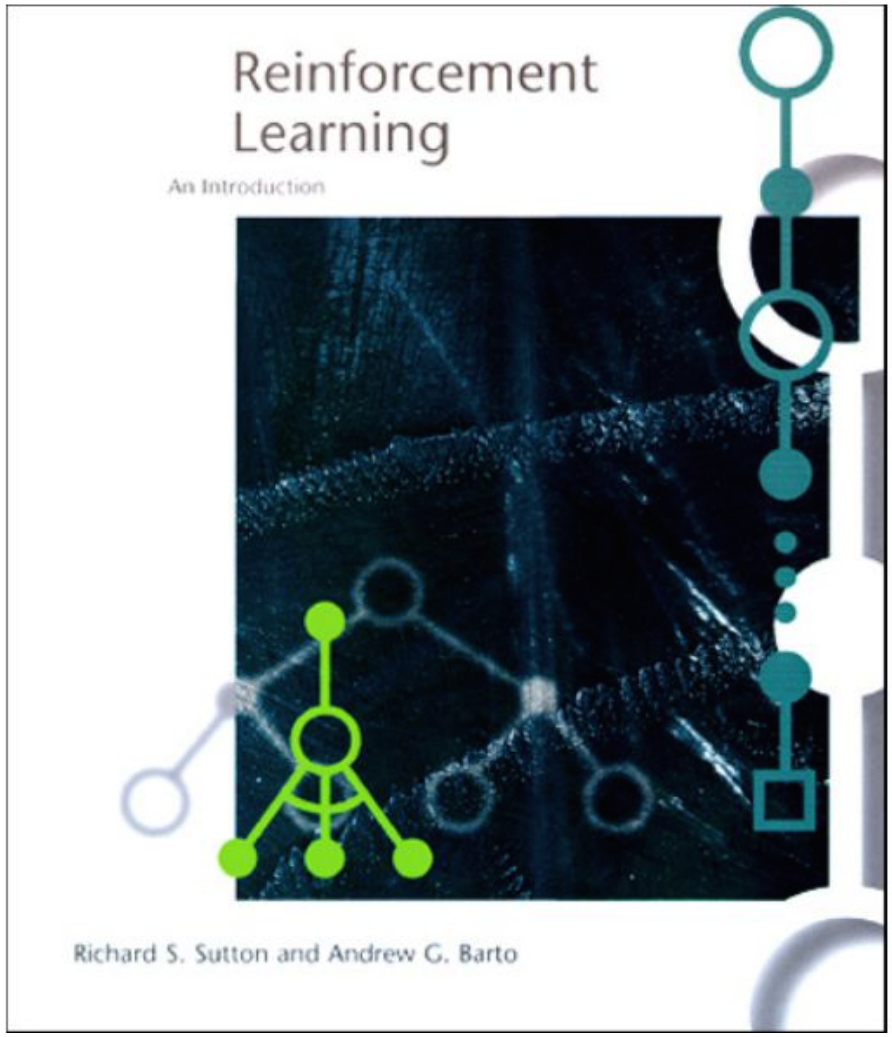


Test:



Rescorla-Wagner would predict no reward; only predicts immediate reward

1990s: Sutton and Barto (Computer Scientists)



1990s: Sutton and Barto (Computer Scientists)

- Rescorla-Wagner

VERSUS

- Temporal Difference Learning:

  Predict value of future rewards (not just current)

Temporal Difference Learning

- Predict value of future rewards
- Predictions are useful for behavior
- Generalization of Rescorla-Wagner to real time
- Explains data that Rescorla-Wagner does not

Based on Dayan slides

Rescorla-Wagner

Want $v_n = r_n$ (here n represents a trial)

Error $\delta_n = r_n - v_n$

$$v_{n+1} = v_n + \varepsilon\delta_n$$

Temporal Difference Learning

Want $v_t = r_t + r_{t+1} + r_{t+2} + r_{t+3} \ldots.$

(here t represents time within a trial; reward can come at any time within a trial. Sutton and Barto interpret v_t as the prediction of total future reward expected from time t onward until the end of the trial)

Based on Dayan slides; Daw slides

Temporal Difference Learning

Want $$v_t = r_t + r_{t+1} + r_{t+2} + r_{t+3} \ldots$$

(here t represents time within a trial; reward can come at any time within a trial. Sutton and Barto interpret v_t as the prediction of total future reward expected from time t onward until the end of the trial)

Prediction error:

$$\delta_t = (r_t + r_{t+1} + r_{t+2} + r_{t+3} \ldots) - V_t$$

Temporal Difference Learning

Want $v_t = r_t + r_{t+1} + r_{t+2} + r_{t+3} \ldots.$

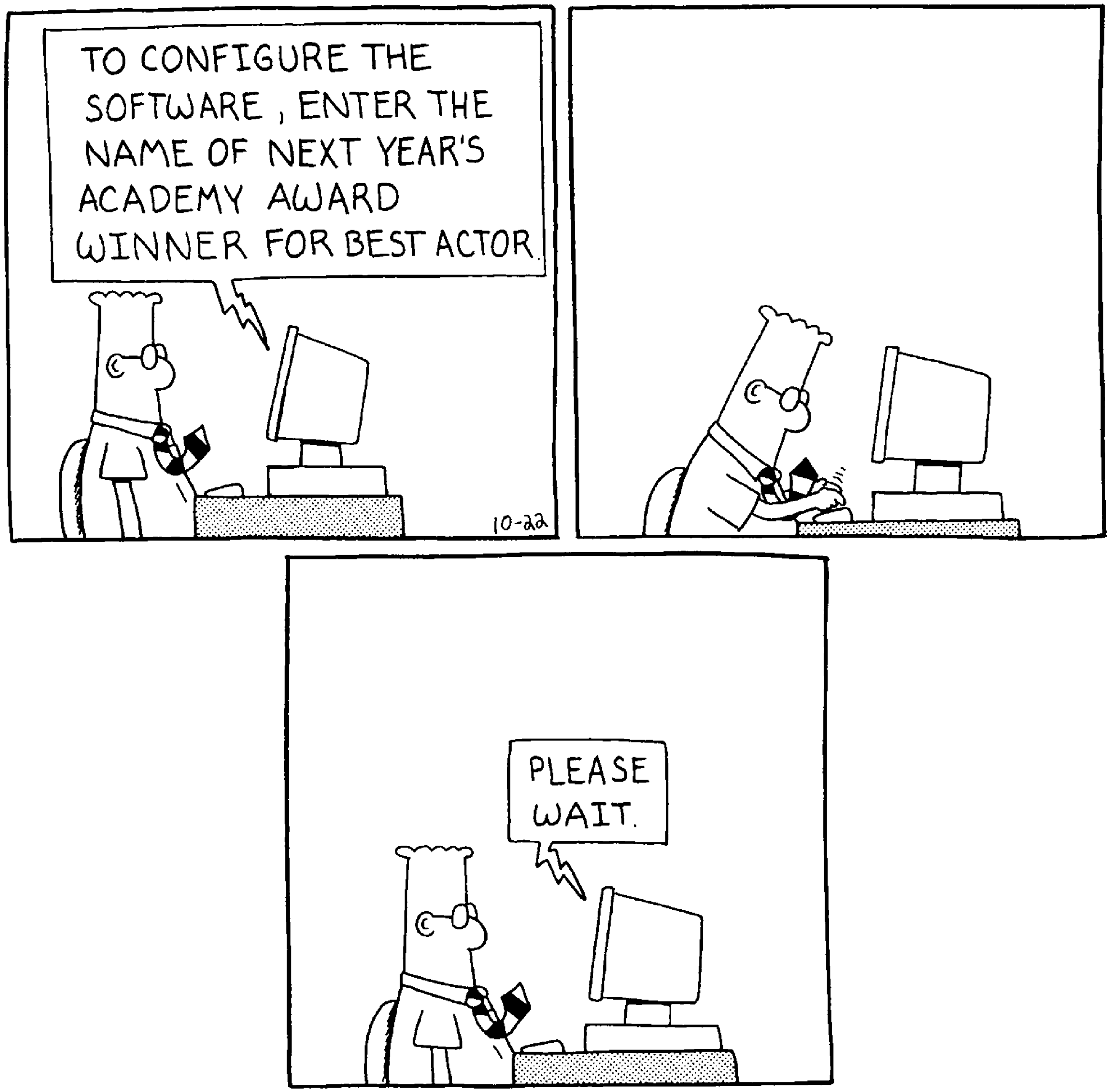
(here t represents time within a trial; reward can come at any time within a trial. Sutton and Barto interpret v_t as the prediction of total future reward expected from time t onward until the end of the trial)

Prediction error:

$$\delta_t = (r_t + r_{t+1} + r_{t+2} + r_{t+3} \ldots.) - V_t$$

Problem??

Based on Dayan slides; Daw slides



In Niv and Schoenbaum, Trends Cog Sci 2009

Temporal Difference Learning

Want $$v_t = r_t + r_{t+1} + r_{t+2} + r_{t+3} \ldots.$$

(here t represents time within a trial)

But we don't want to wait forever for all future rewards…

$$r_{t+1}; r_{t+2}; r_{t+3} \ldots.$$

Temporal Difference Learning

Want $$v_t = r_t + r_{t+1} + r_{t+2} + r_{t+3} \ldots$$

(here t represents time within a trial)

Recursion "trick": $$v_t = r_t + v_{t+1}$$

Based on Dayan slides; Daw slides

Temporal Difference Learning

From recursion want:

$$v_t = r_t + v_{t+1}$$

Error:

$$\delta_t = r_t + v_{t+1} - v_t$$

Temporal Difference Learning

From recursion want: $$v_t = r_t + v_{t+1}$$

Error: $$\delta_t = r_t + v_{t+1} - v_t$$

Update: $$v_t \to v_t + \varepsilon(r_t + v_{t+1} - v_t)$$

$$= (1- \ \varepsilon)v_t + \varepsilon(r_t + v_{t+1})$$

RV versus TD

- Rescorla-Wagner error: (n represents trial)

$$\delta_n = r_n - v_n$$

- Temporal Difference Error: (t is time within a trial)

$$\delta_t = r_t + \boxed{v_{t+1} - v_t}$$

Name comes from!

Temporal Difference Learning

- Temporal Difference Error: (t is time within a trial)

$$\delta_t = r_t + \boxed{v_{t+1} - v_t}$$

Name comes from!

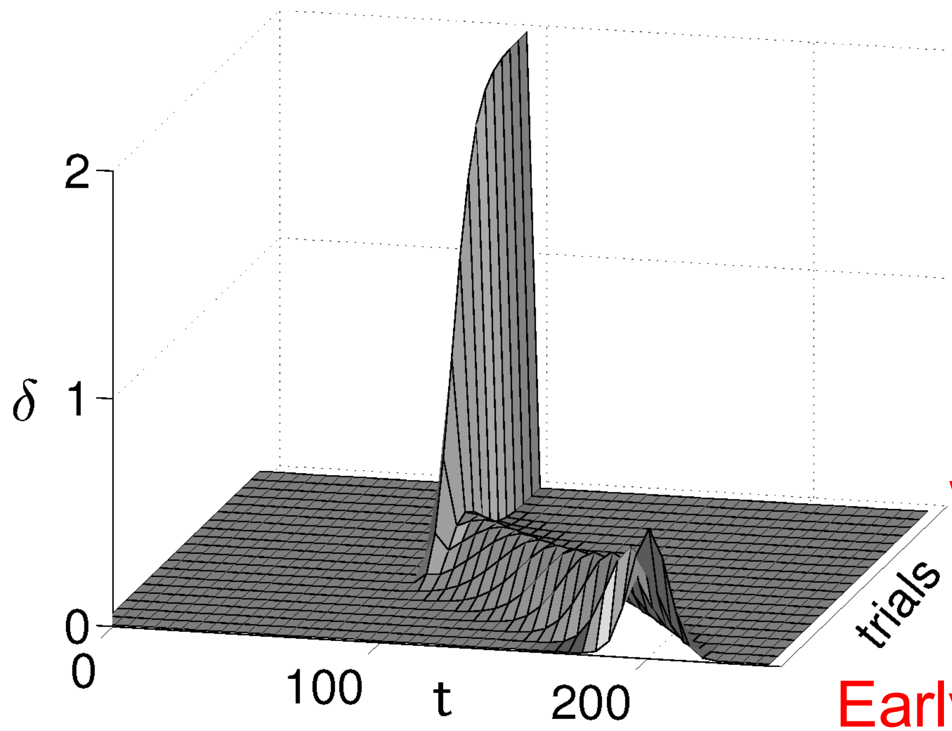
$v_{t+1} = v_t$ Predictions steady

$v_{t+1} > v_t$ Got better

$v_{t+1} < v_t$ Got worse

Based on Daw slides

Temporal Difference Learning



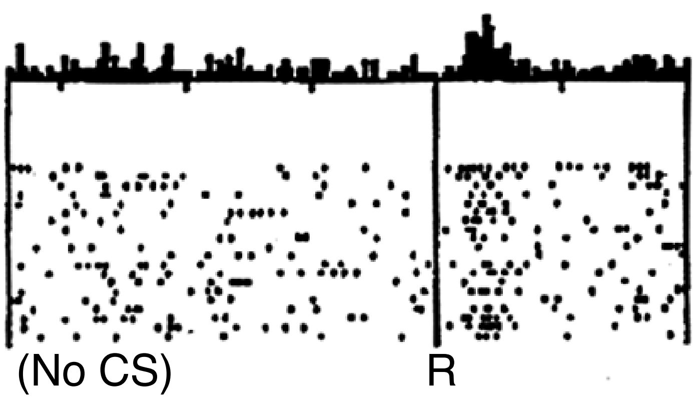
Late trials: error when stimulus

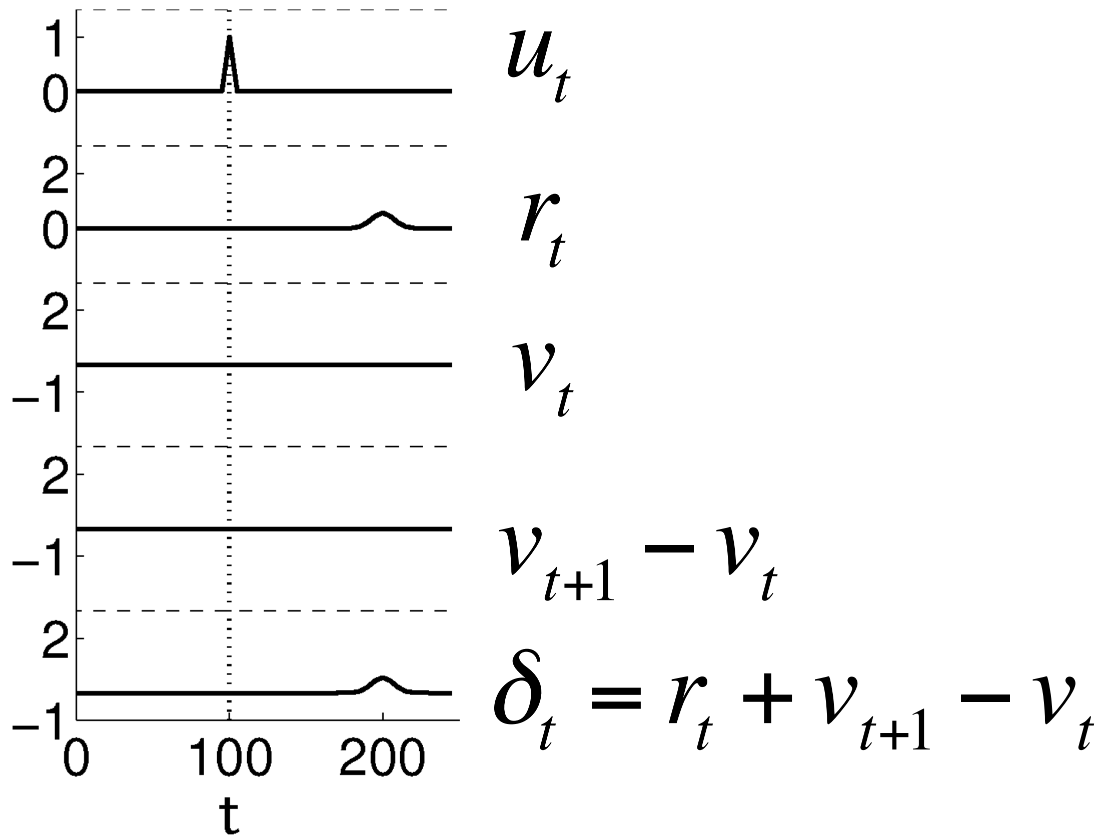
Early trials: error when reward

Dayan and Abbott Book: Surface plot of prediction error (stimulus at 100; reward at 200)

Temporal Difference Learning

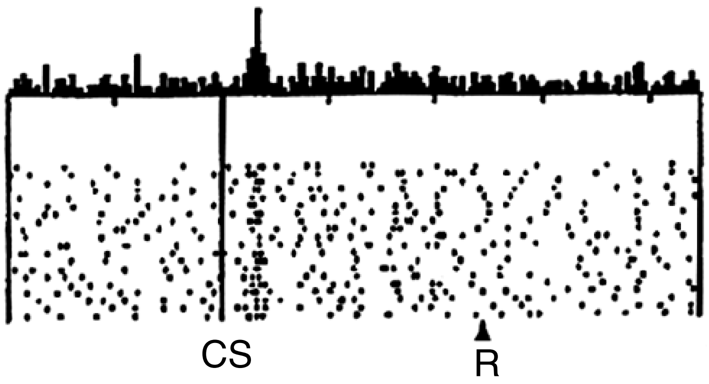
Before learning

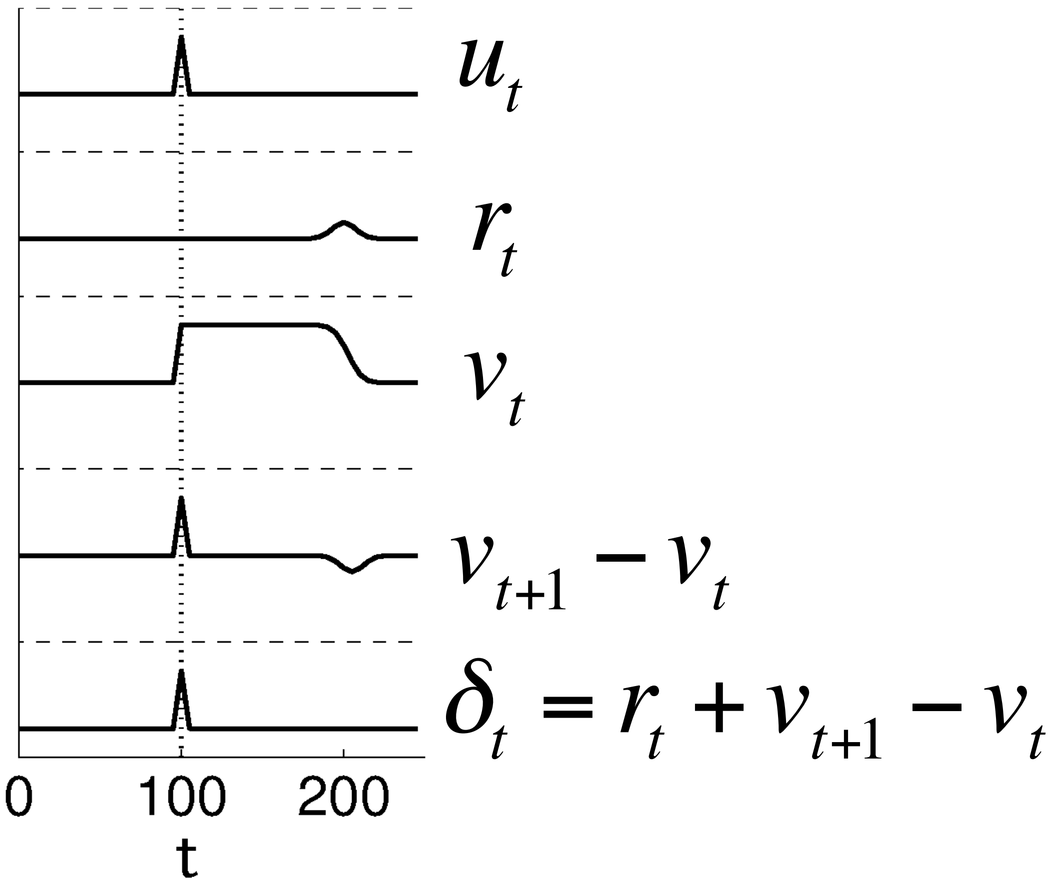




Temporal Difference Learning

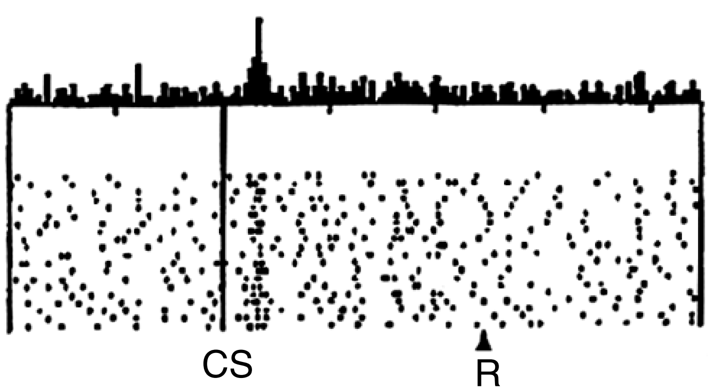
After learning

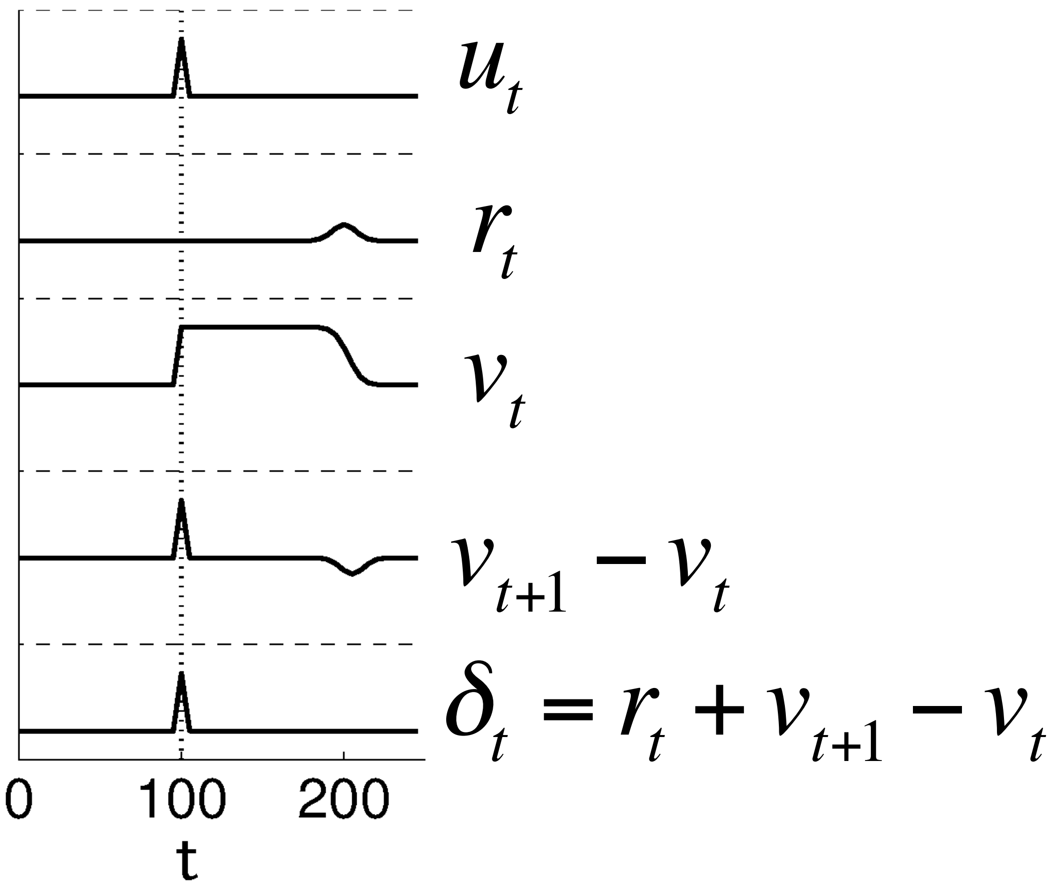




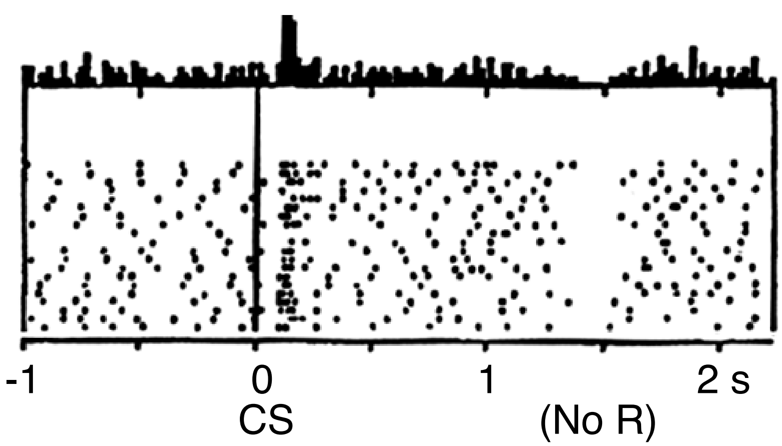
Temporal Difference Learning

After learning

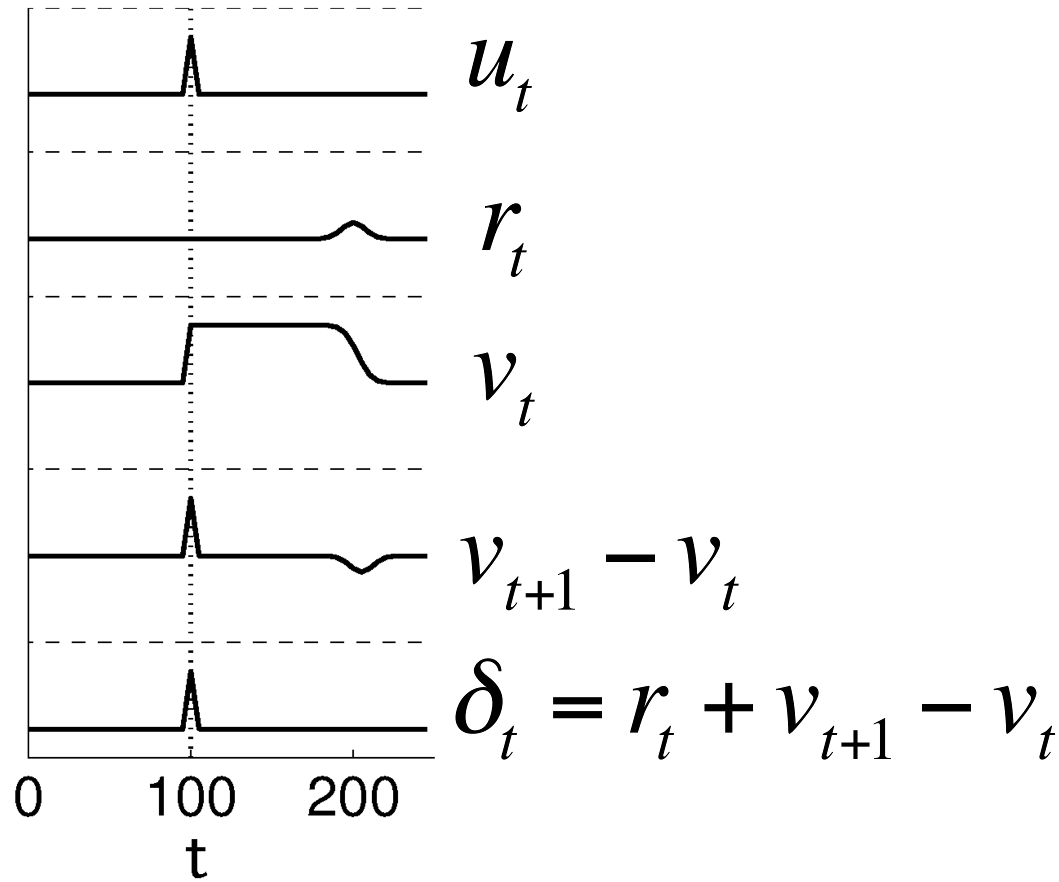




Temporal Difference Learning



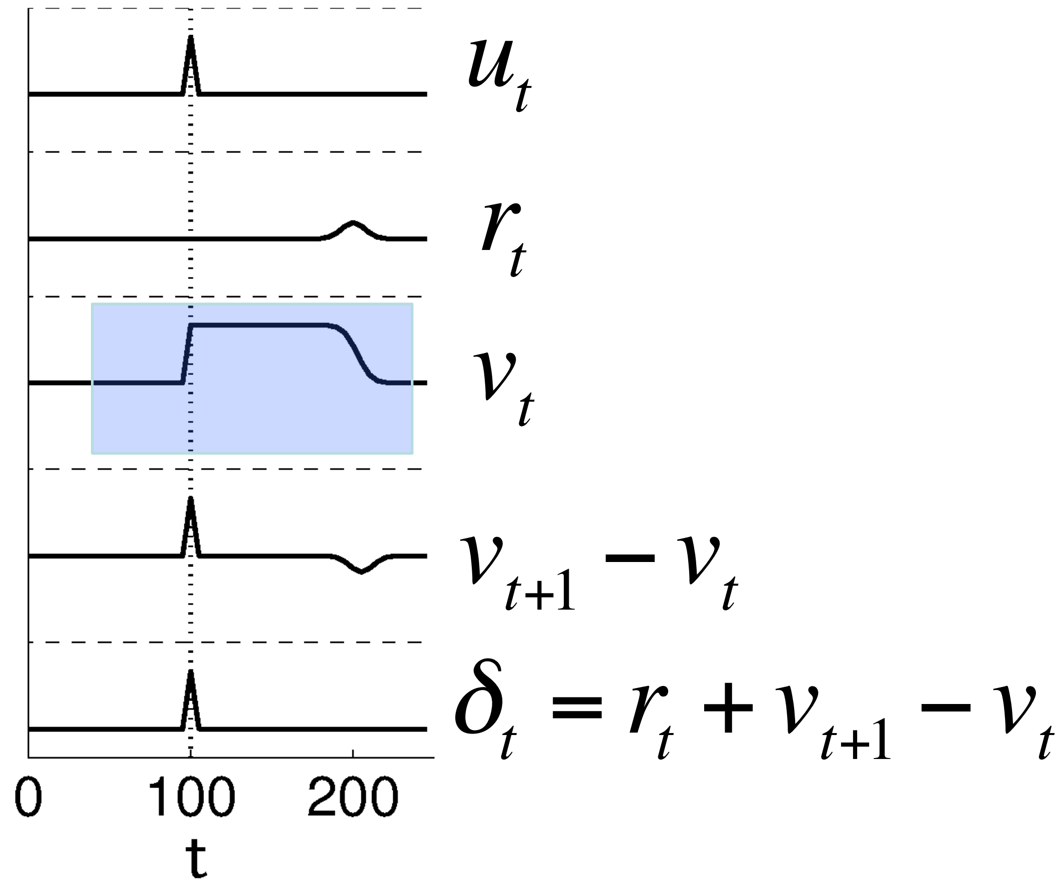
After learning



What should change here?

Temporal Difference Learning

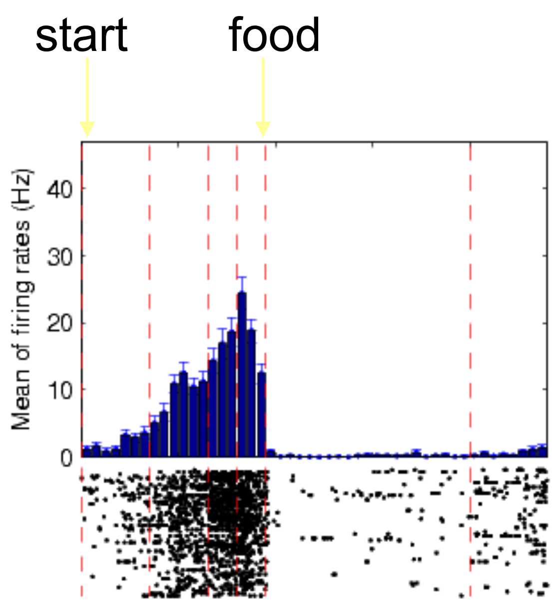
After learning



What about anticipation of future rewards?

Temporal Difference Learning

Striatal neurons (activity that precedes rewards and changes with learning)



(Daw)

What about anticipation of future rewards?

From Dayan slides

Summary

Marr's 3 levels:

- Problem: Predict future reward
- Algorithm: Temporal Difference Learning (generalization of Rescorla-Wagner)
- Implementation: Dopamine neurons signaling error in reward prediction

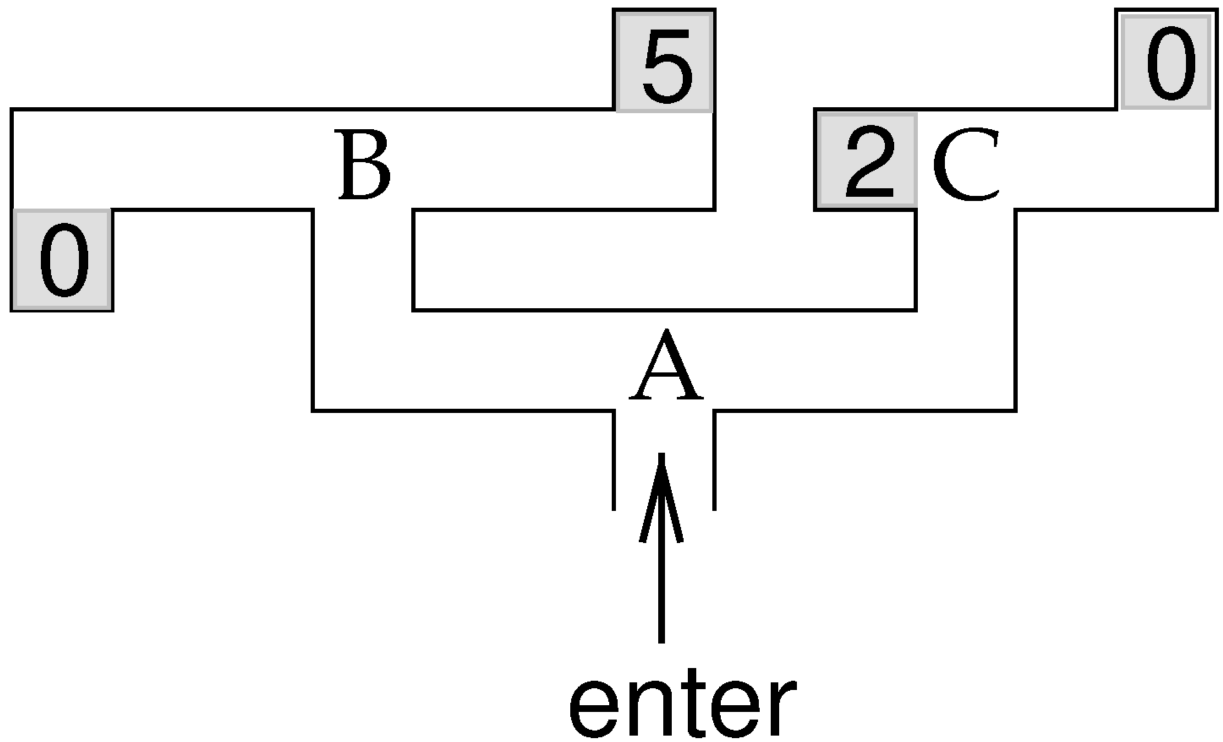
Based on Dayan slides

What else

- Applied in more sophisticated sequential decision making tasks with future rewards

- Foundation of a lot of active research in machine learning, computational neuroscience, Biology, Psychology

More sophisticated tasks



Dayan and Abbott book

Recent example in machine learning

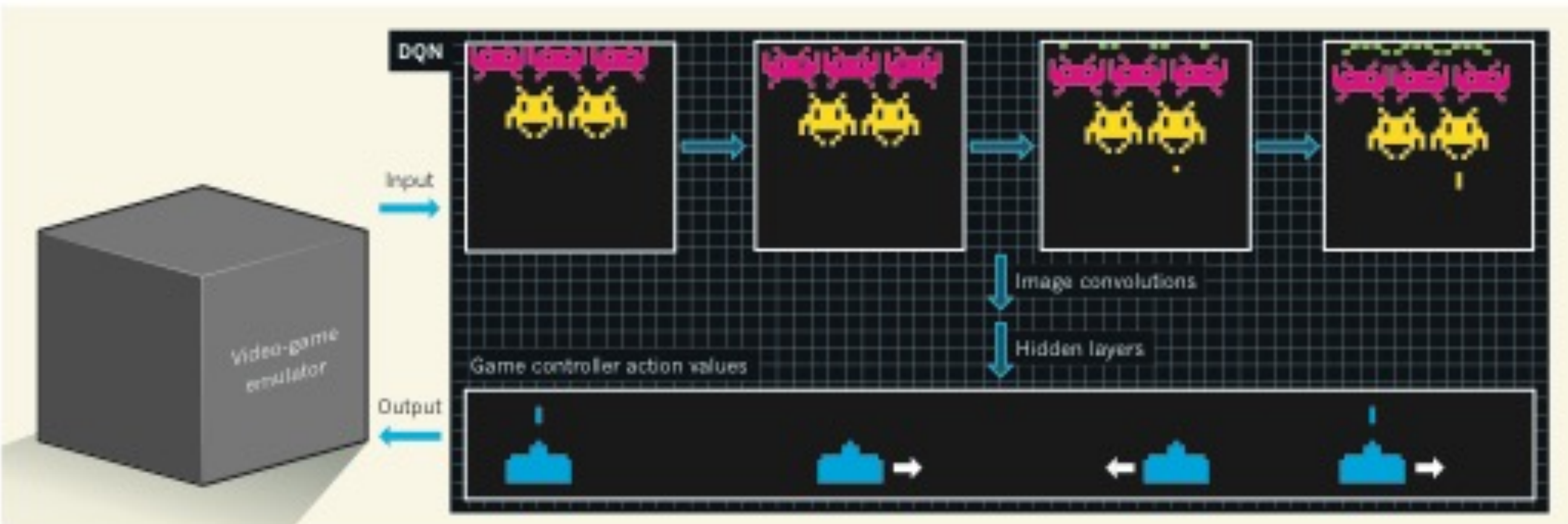
LETTER

doi:10.1038/nature14236

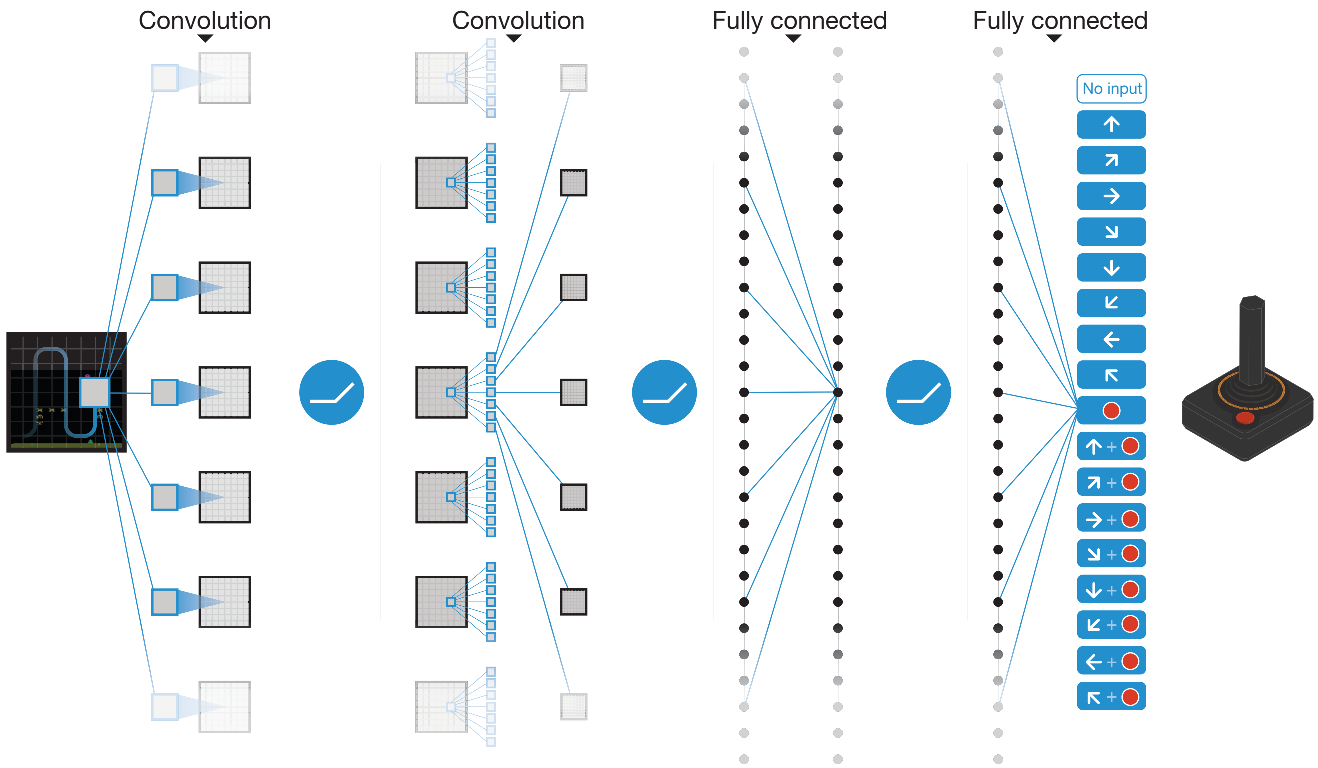
Human-level control through deep reinforcement learning

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]

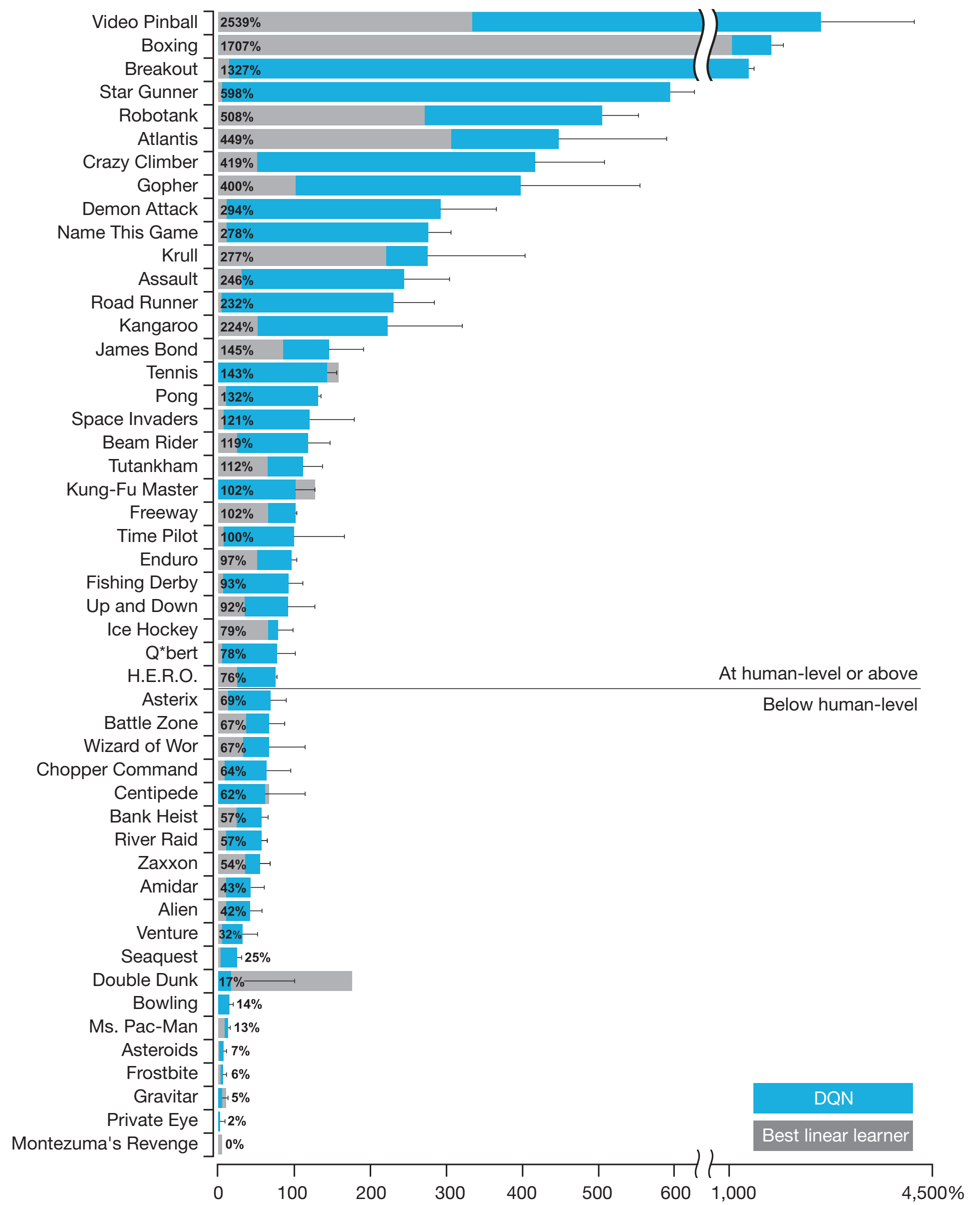
Mnih et al. Nature 518, 529–533; **2015**



Scholkopf. News and Views; Nature **2015**



Mnih et al. Nature 518, 529–533; **2015**



Mnih et al. Nature 518, 529–533; **2015**