Reinforcement Learning

Odelia Schwartz
2020

Forms of learning?

Forms of learning

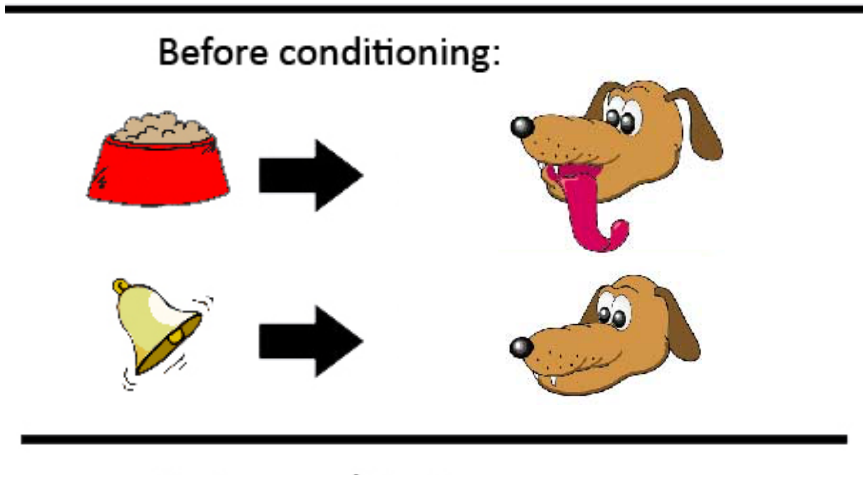
- Unsupervised learning
- Supervised learning
- Reinforcement learning

Forms of learning

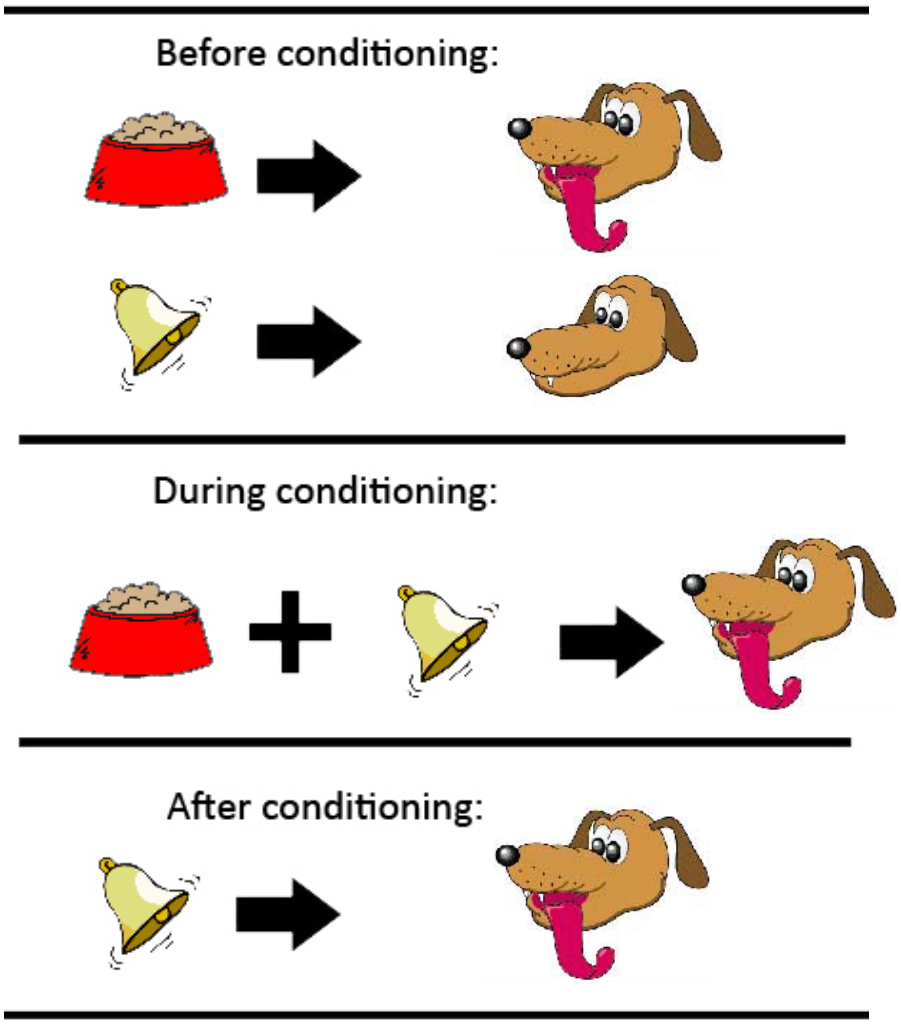
- Unsupervised learning
- Supervised learning
- Reinforcement learning

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
- Actual reward r (assigning a value to a reward)
- Predicted reward v

based on Dayan and Abbott book

Rescorla-Wagner rule (1972)

- Minimize difference between received reward and predicted reward
- Actual reward r (assigning a value to a reward)
- Predicted reward v

How can we minimize difference between actual and predicted?

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:

$$(r - v)^2$$

DILBERT: © Scott Adams/Dist. by United Feature Syndicate, Inc

In Niv and Schoenbaum 2009

Rescorla-Wagner rule (1972)

- Binary variable u (1 if stimulus is present; 0 if absent)

based on Dayan and Abbott book

Rescorla-Wagner rule (1972)

- Binary variable u (1 if stimulus is present; 0 if absent)
- Remember v is predicted reward
- Linear weight w

$$v = wu$$

based on Dayan and Abbott book

Rescorla-Wagner rule (1972)

- Binary variable u (1 if stimulus is present; 0 if absent)
- Remember v is predicted reward
- 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:

$$(r - v)^2$$

(average over presentations of stimulus and reward)

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:

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

Stimulus present

ε learning rate; associability of stimulus with reward

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:

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

Stimulus present

ε learning rate; associability of stimulus with reward

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

Rescorla-Wagner rule (1972)

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

Rescorla-Wagner rule (1972)

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

Later: dopaminergic neurons in Ventral Tegmental Area
Interpreted as encoding form of prediction error

- Update weight:

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

based on Dayan and Abbott book

- Update weight:

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

If stimulus always present,
can just omit u

based on Dayan and Abbott book

- Update weight:

- If a stimulus u is always presented, we can replace the weights w with the predicted reward v that we are updating…

- We've also now written the notation such that n is trial number

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

actual predicted

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?

- 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?

The predicted reward equals the actual current reward

- So if a stimulus is presented at trial n:

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

- What happens when it is smaller than 1?

- So if a stimulus is presented at trial n:

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

- What happens when it is smaller than 1?

Weights less heavily current reward

Acquisition and extinction

Assume the following experiment:

- Each trial stimulus is paired with reward or not paired with reward

From Dayan and Abbott book

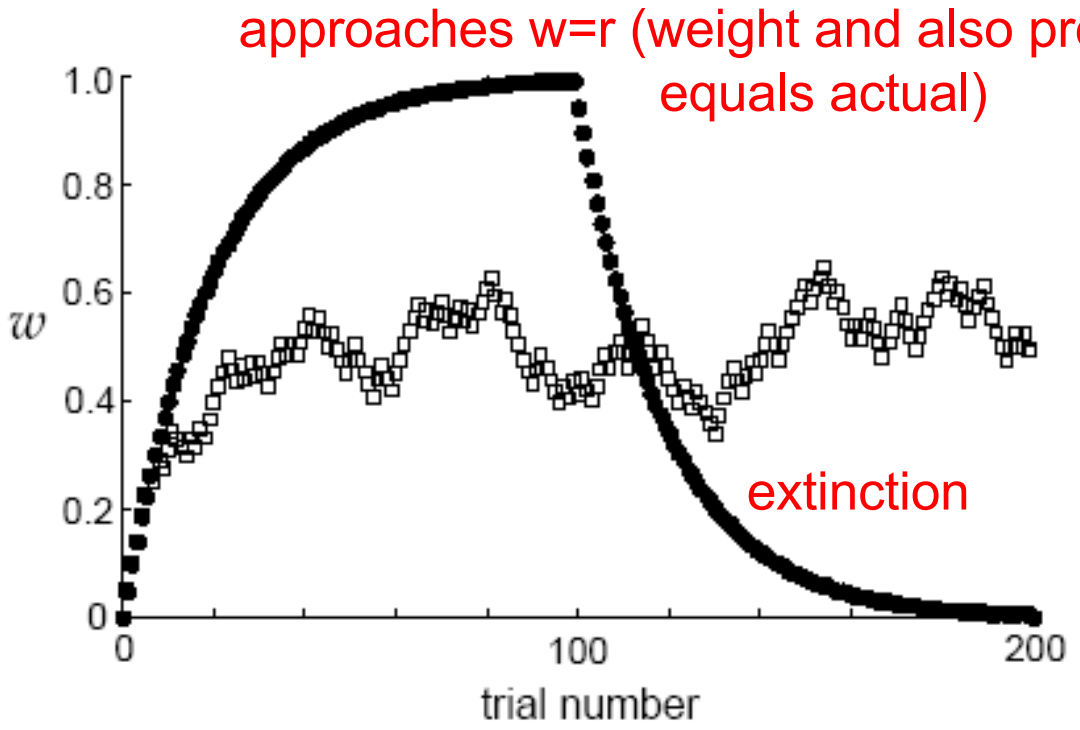
Acquisition and extinction

Assume the following experiment:

- Each trial stimulus is paired with reward or not paired with reward
- First 100 trials: reward (r=1) paired with stimulus
- next 100 trials: no reward (r=0) paired with stimulus

From Dayan and Abbott book

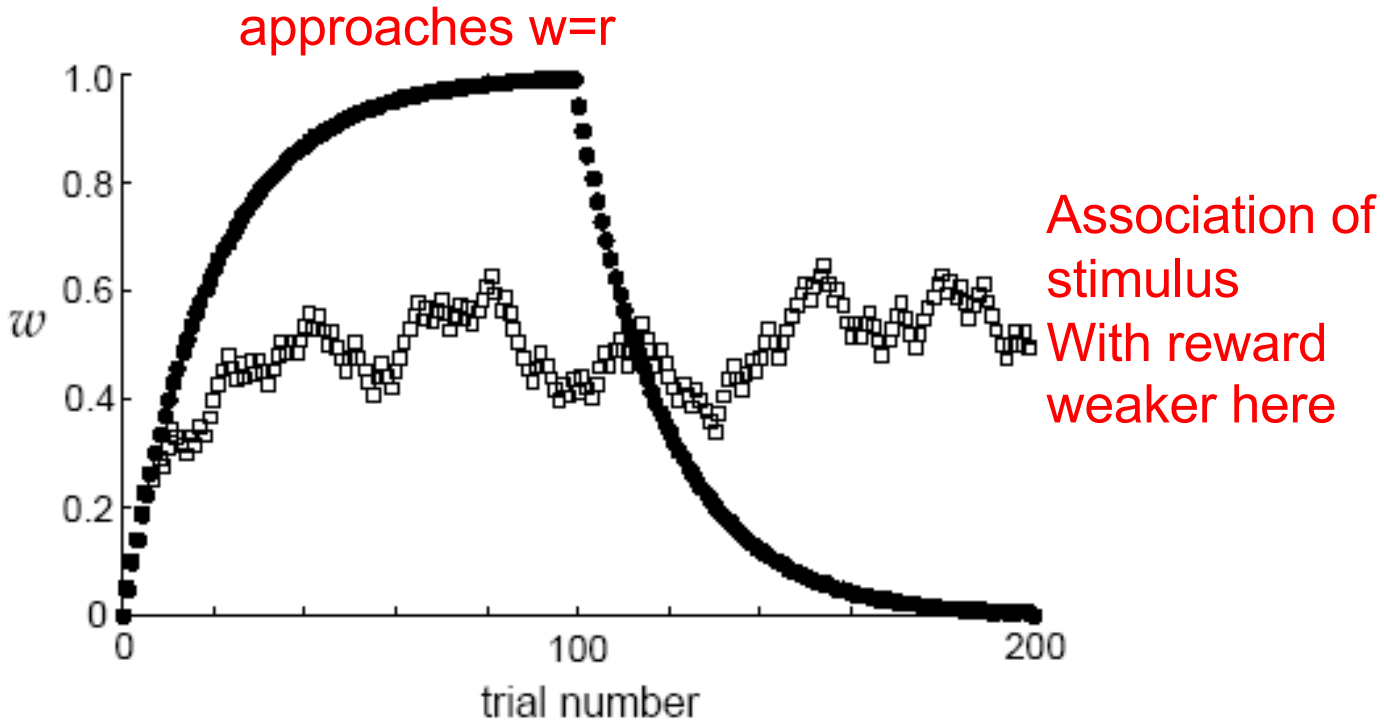
Acquisition and extinction



- Solid: First 100 trials: reward (r=1) paired with stimulus; next 100 trials no reward (r=0) paired with stimulus (learning rate .05)
- Dashed: Ignore for now

From Dayan and Abbott book

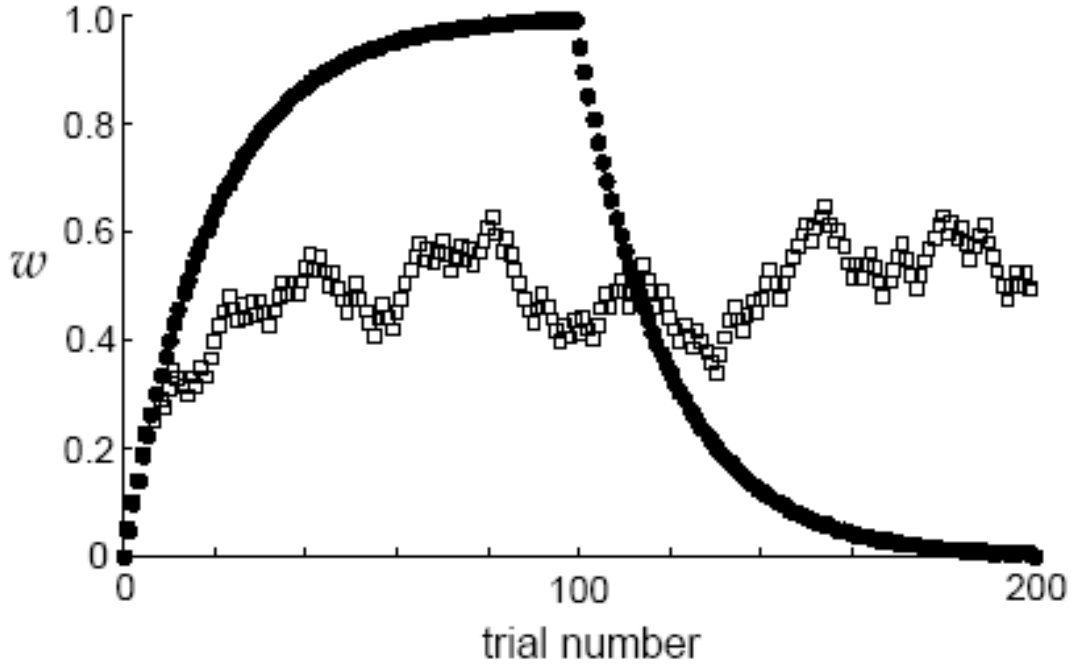
Acquisition and extinction



- 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

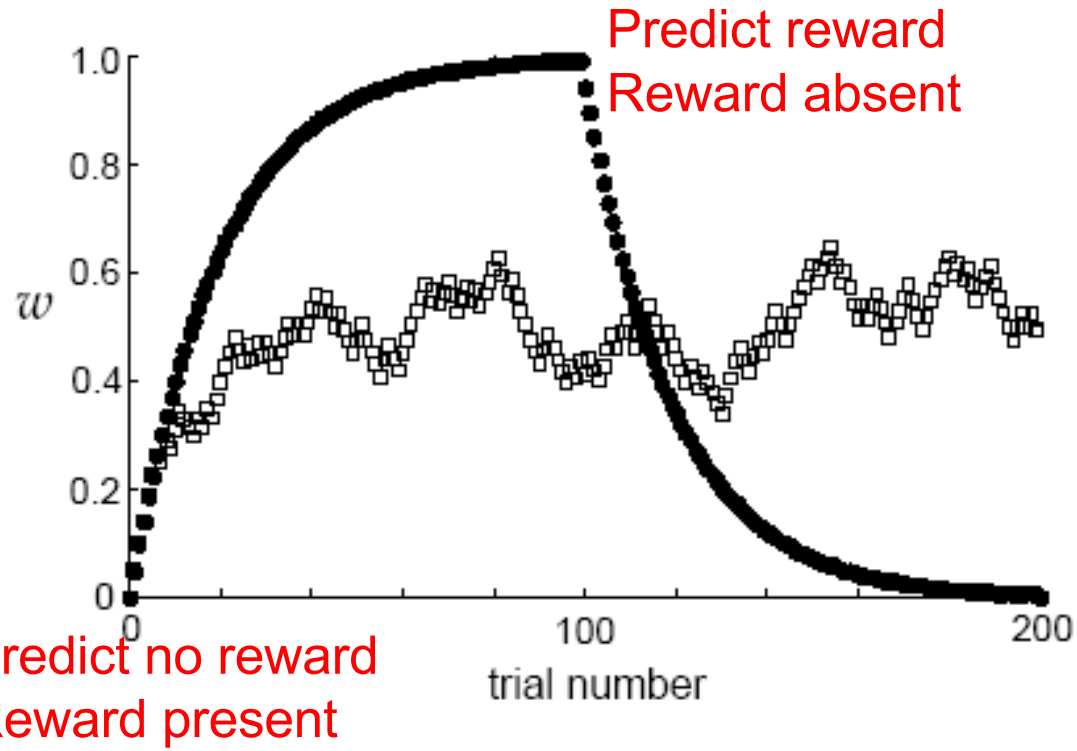
Acquisition and extinction



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

From Dayan and Abbott book

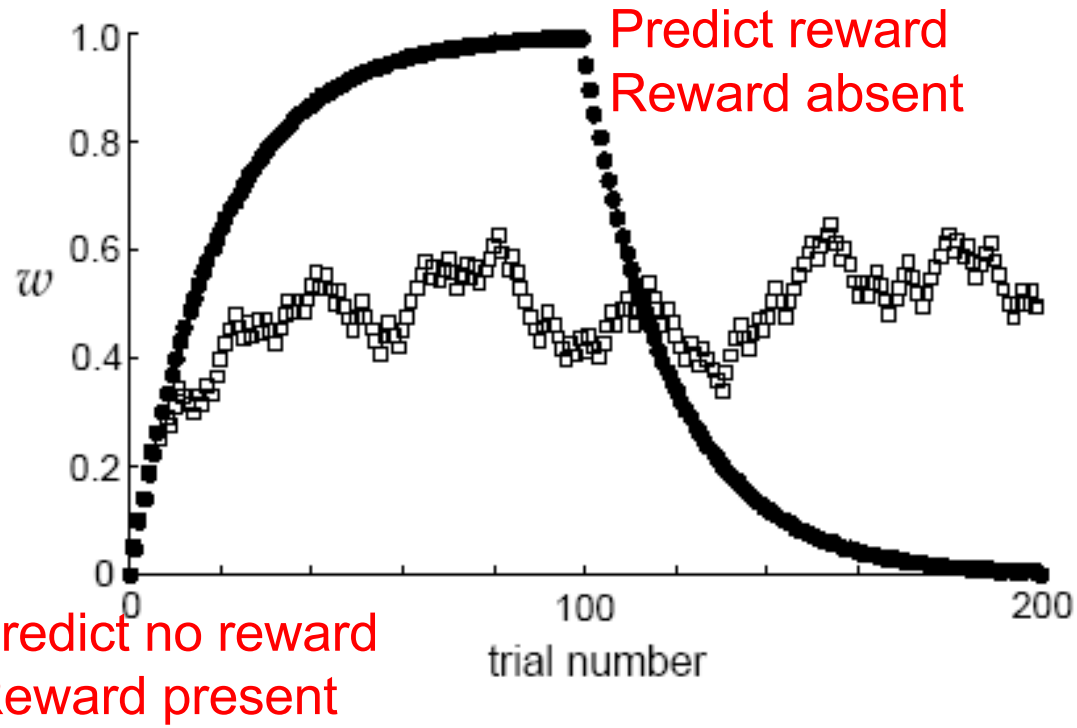
Acquisition and extinction



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

From Dayan and Abbott book

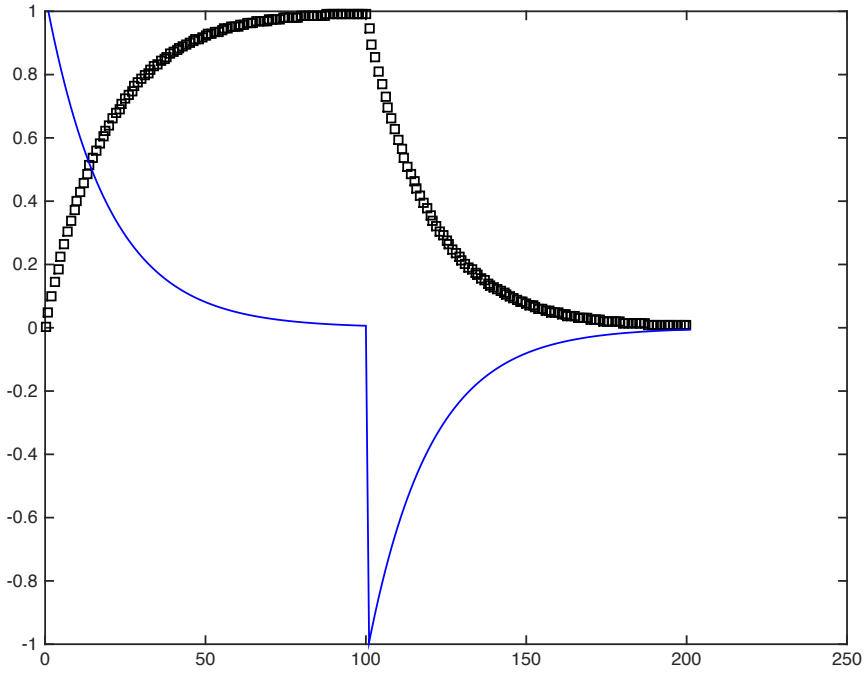
Acquisition and extinction



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

From Dayan and Abbott book

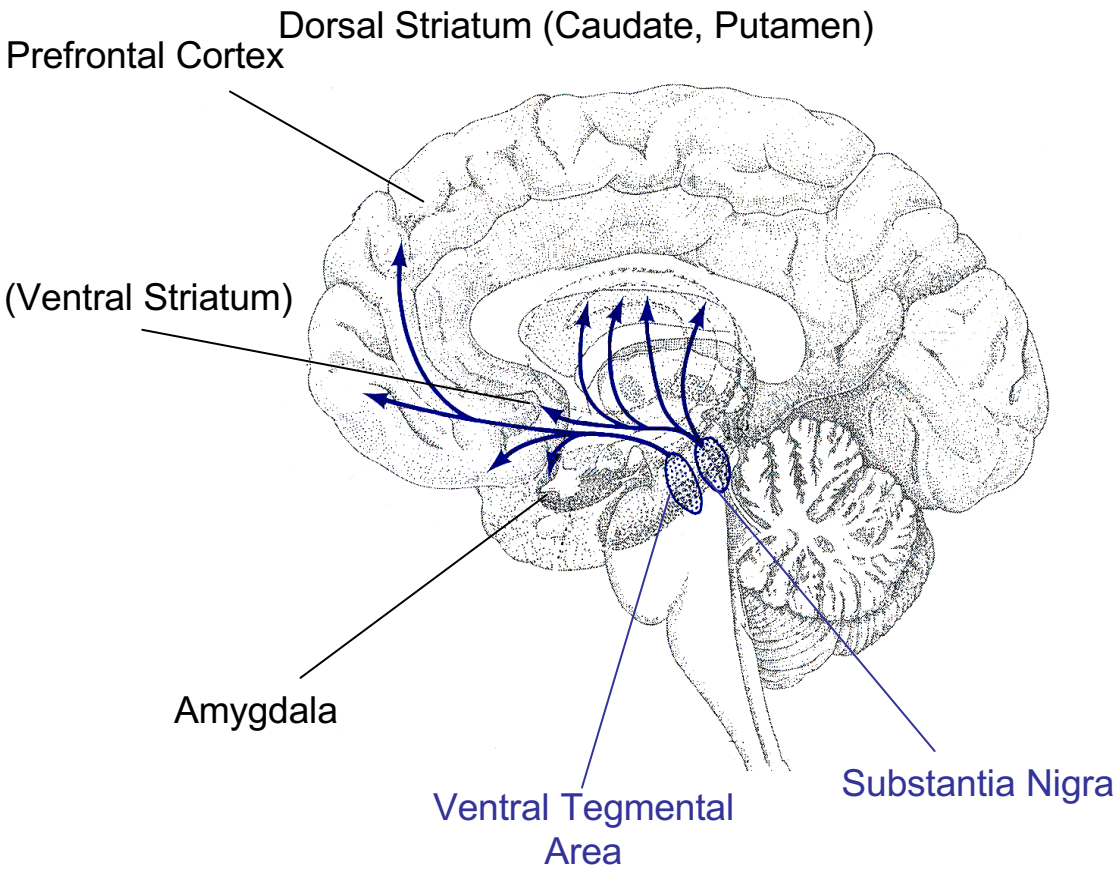
Acquisition and extinction



- Black curve: v
- Blue curve: (r-v)

From Dayan and Abbott book

Dopamine areas



From Dayan slides

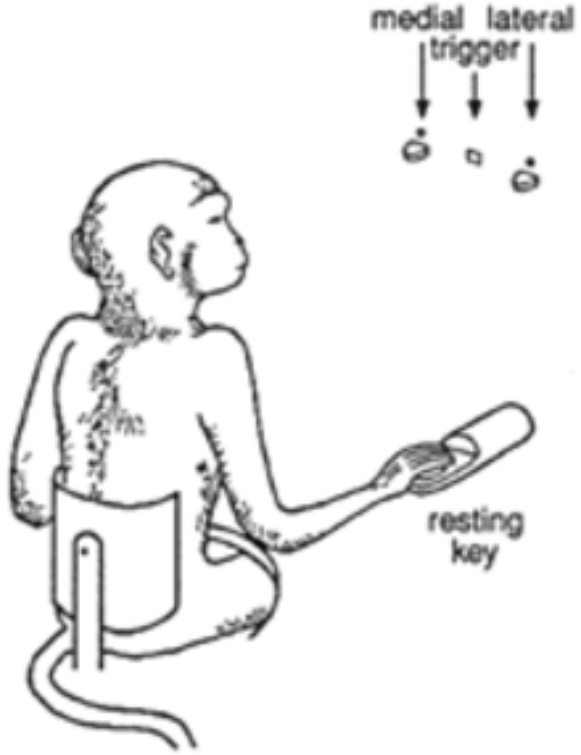
Dopamine roles?

Dopamine roles?

Associated with…

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

VTA Activity of dopaminergic neurons

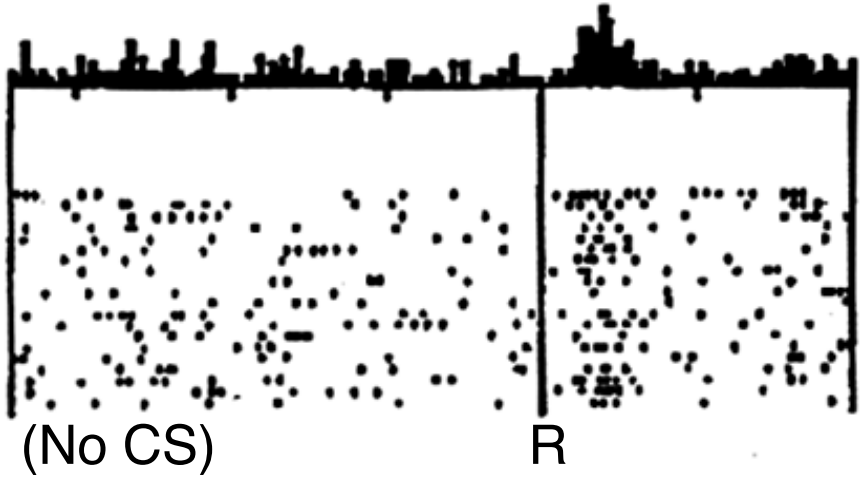


- Monkey trained to respond to light or sound for food and drink rewards (instrumental conditioning)
- Finger on resting key until sound is presented
- Then release key to get reward

Schultz, Dayan, Montague, 1997

VTA Activity of dopaminergic neurons

No prediction
Reward occurs

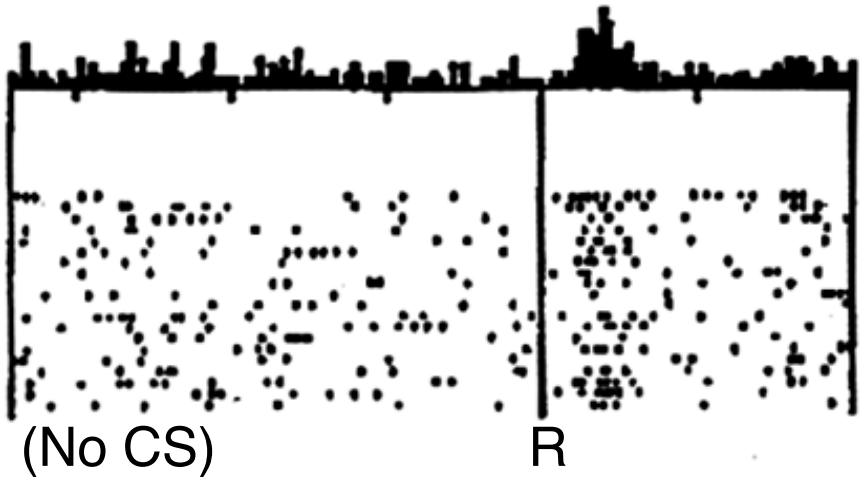


Before learning, reward is given in experiment, but animal does not predict (expect) reward (why is there increased activity after reward?)

Schultz, Dayan, Montague, 1997

VTA Activity of dopaminergic neurons

No prediction
Reward occurs

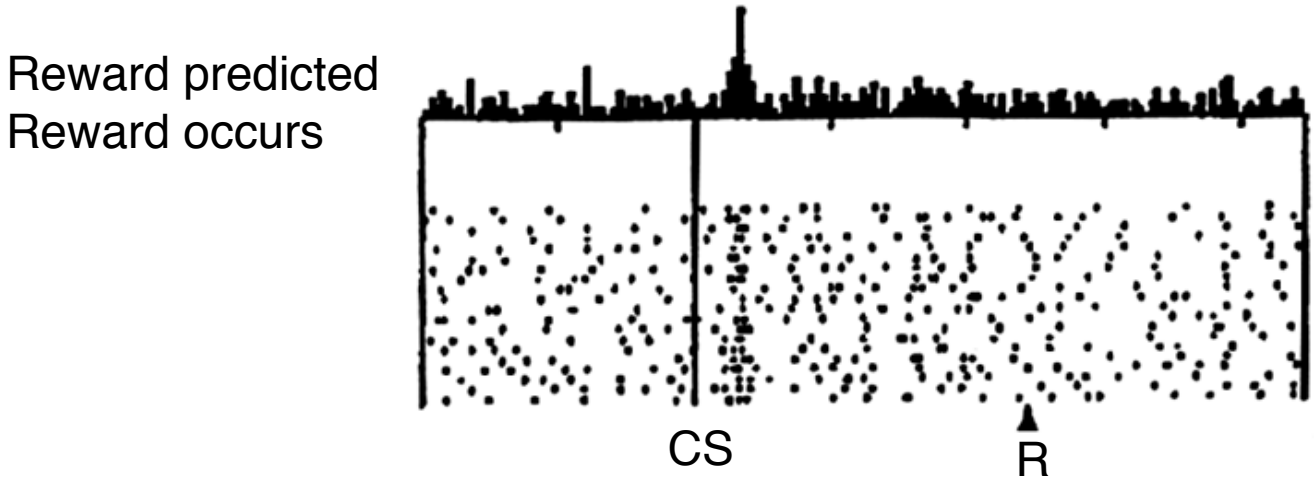


Before learning, reward is given in experiment, but animal does not predict (expect) reward (why is there increased activity after reward?)

Think r-v (actual minus predicted reward)

Schultz, Dayan, Montague, 1997

VTA Activity of dopaminergic neurons

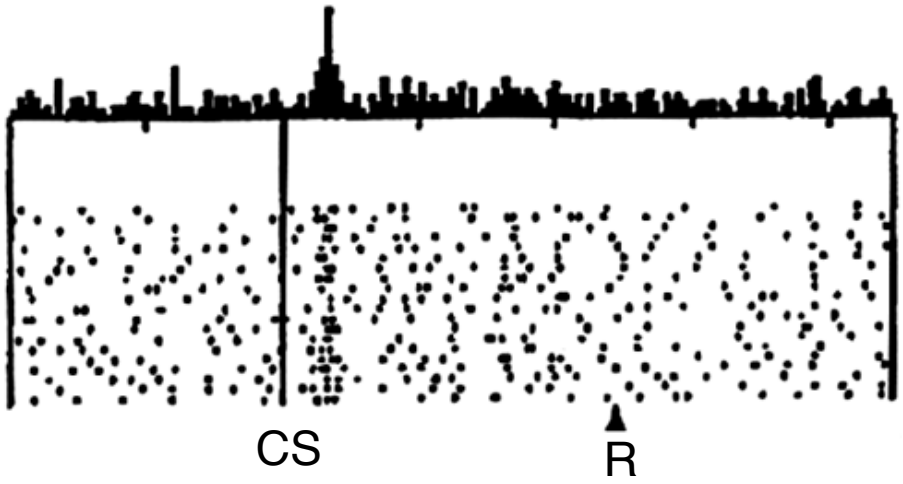


After learning, conditioned stimulus predicts reward, and reward is given in experiment (why is activity fairly uniform after reward?)

Schultz, Dayan, Montague, 1997

VTA Activity of dopaminergic neurons

Reward predicted
Reward occurs



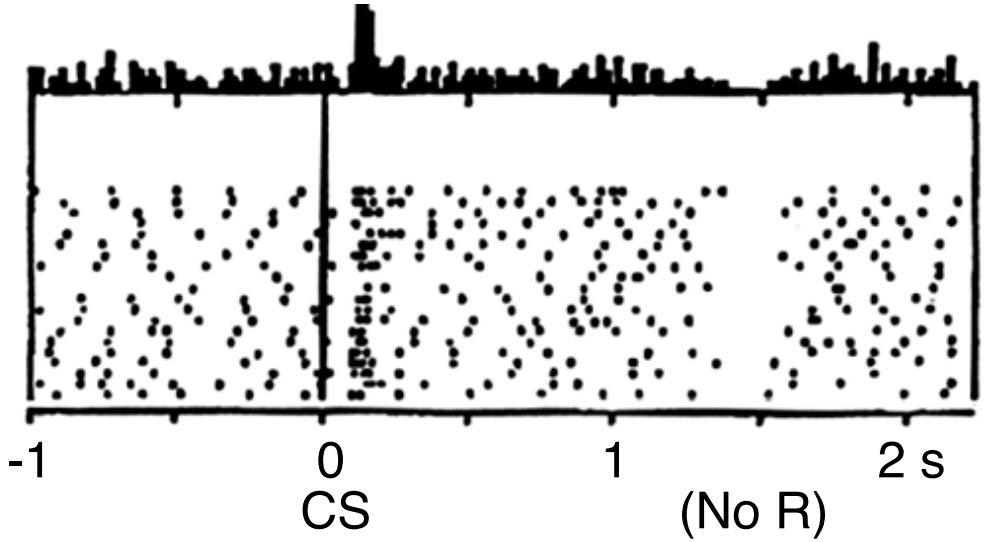
After learning, conditioned stimulus predicts reward, and reward is given in experiment (why is activity fairly uniform after reward?)

Think r-v (actual minus predicted reward)

Schultz, Dayan, Montague, 1997

VTA Activity of dopaminergic neurons

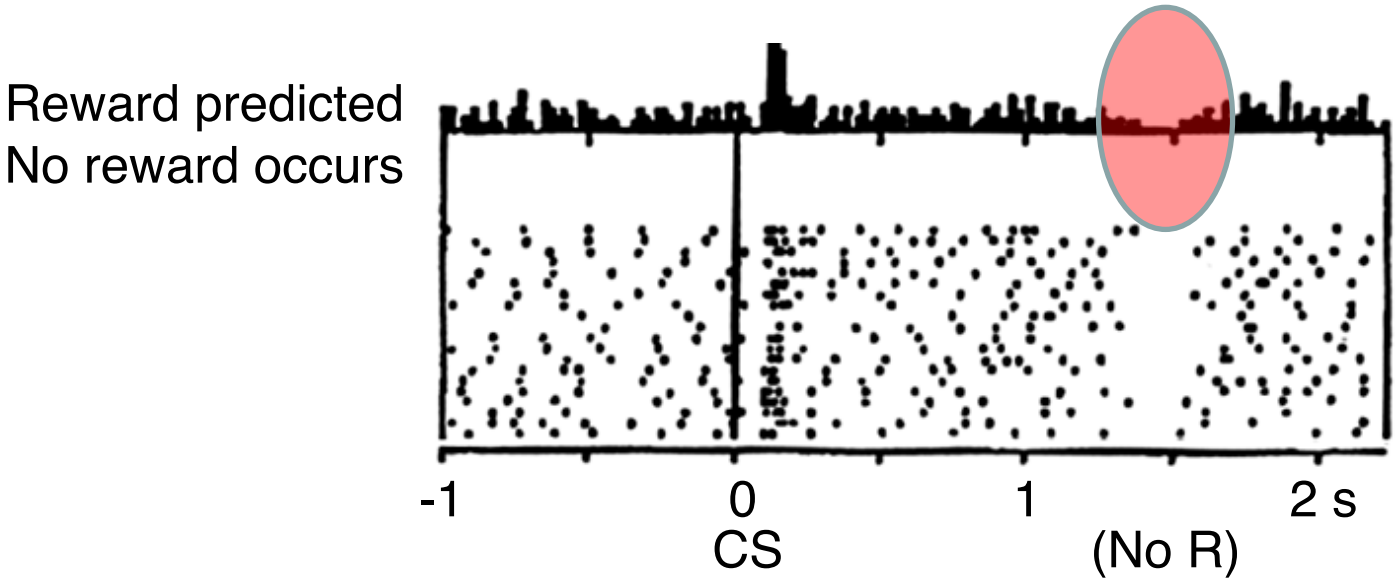
Reward predicted
No reward occurs



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

Schultz, Dayan, Montague, 1997

VTA Activity of dopaminergic neurons

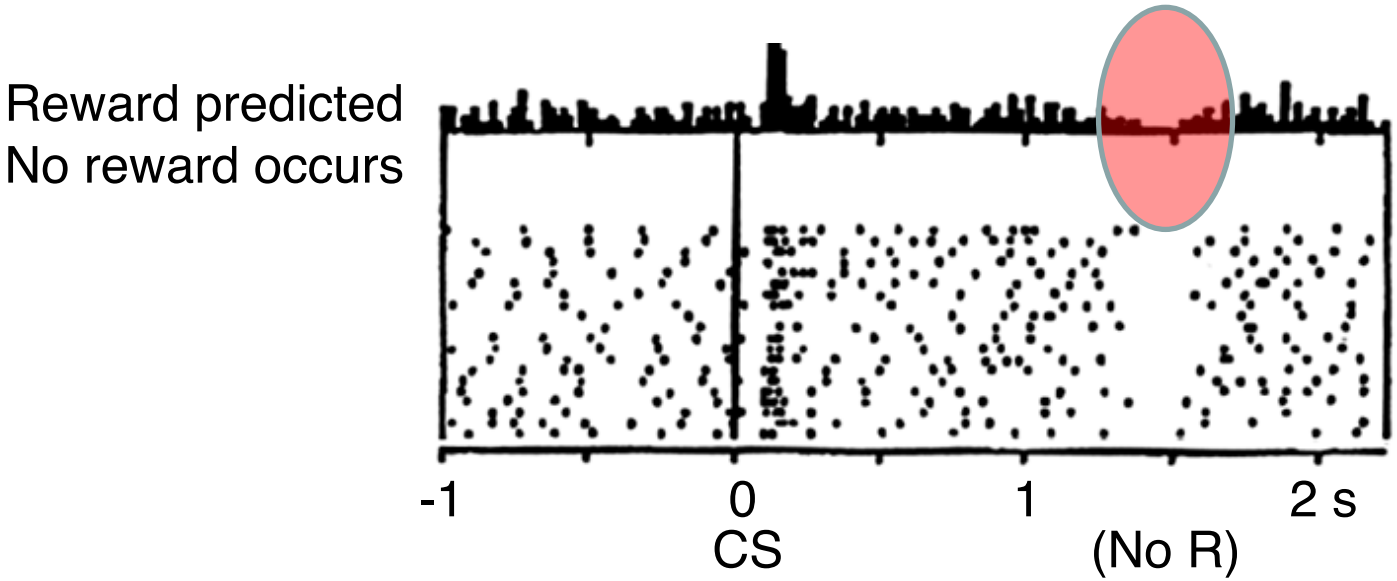


After learning, conditioned stimulus predicts reward so there is an expectation of reward, but no reward is given in the experiment

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

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)

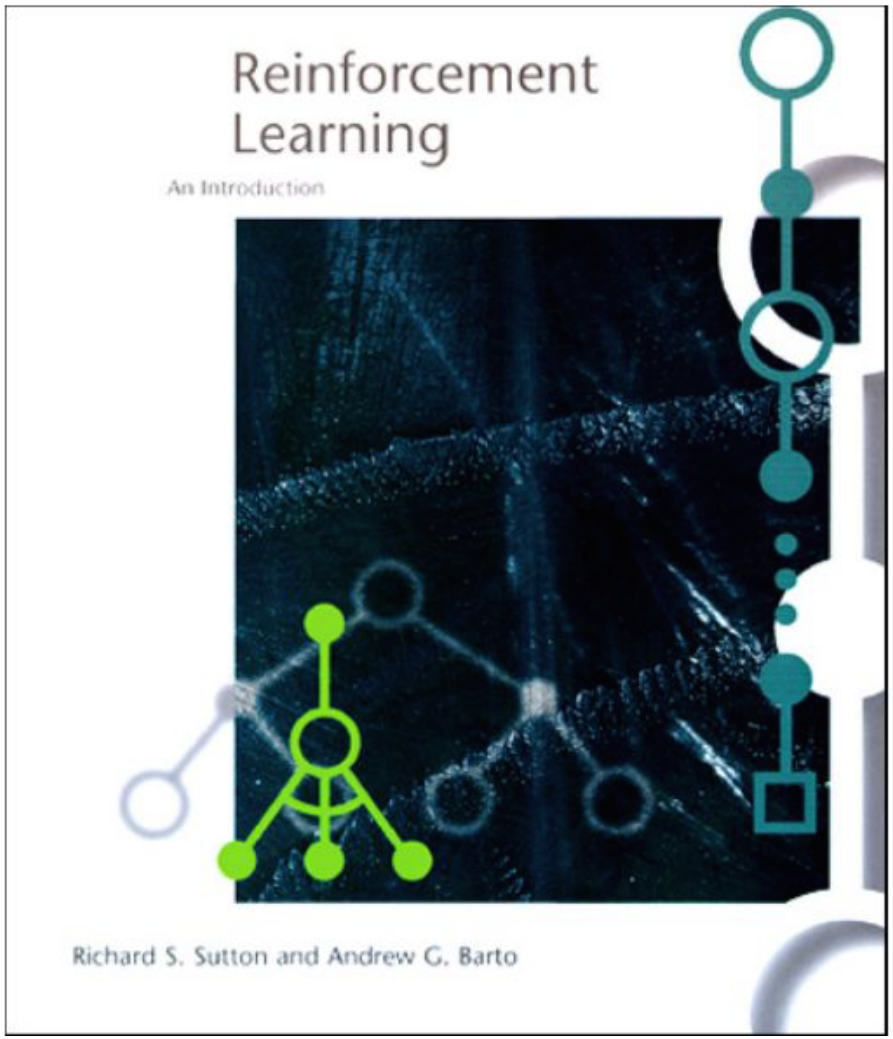
Shortcomings of Rescorla-Wagner: Example: secondary conditioning

Train:

Test:

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

1990s: Sutton and Barto (Computer Scientists)



Now also
New edition

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

From Dayan slides

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

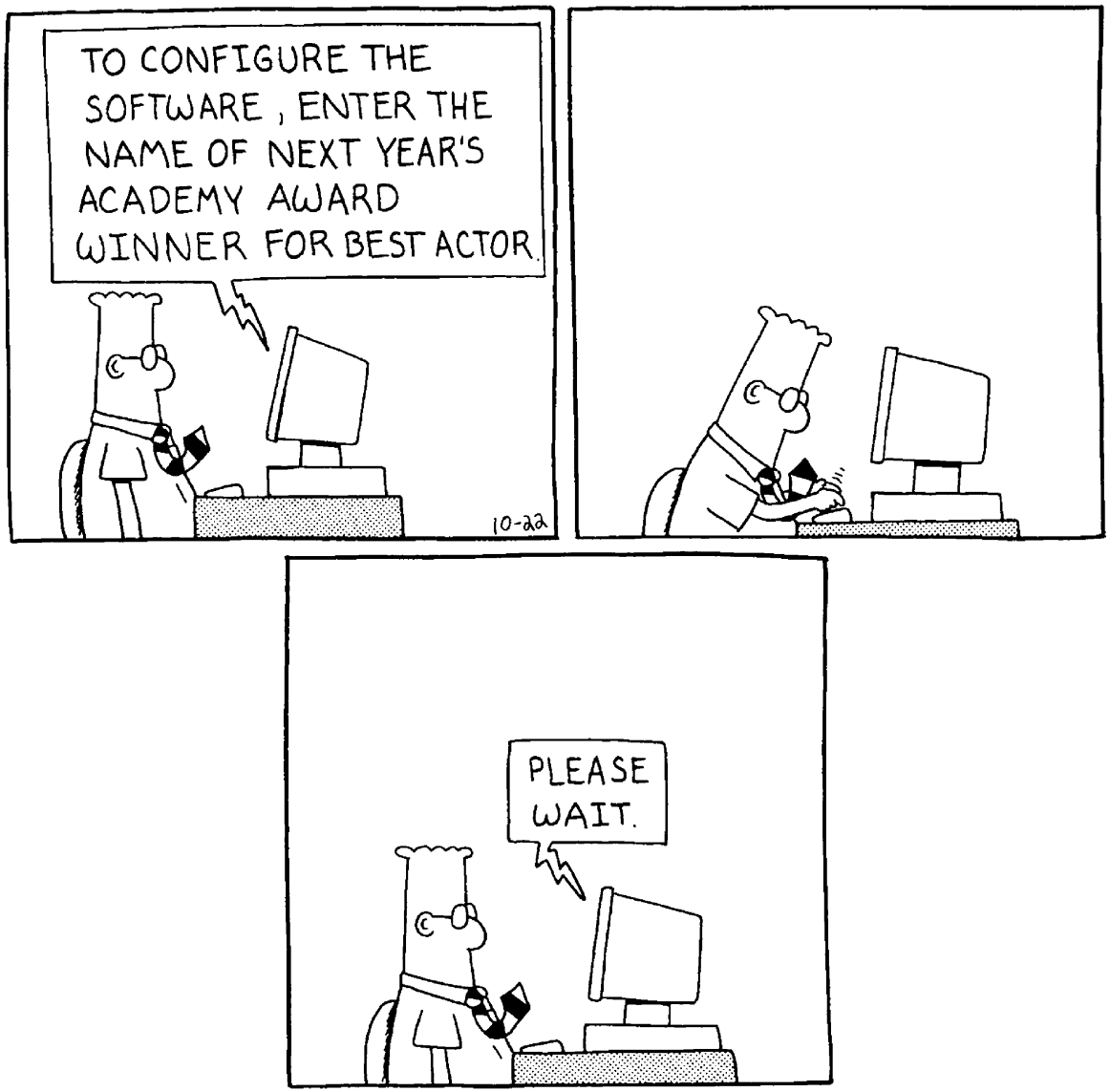
(here t represents time within a trial; reward can come at any time within a trial. Sutton and Barto interpret v_t as the prediction of total future reward expected from time t onward until the end of the trial)

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



DILBERT: © Scott Adams/Dist. by United Feature Syndicate, Inc

In Niv and Schoenbaum, Trends Cog Sci 2009

Temporal Difference Learning

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

(here t represents time within a trial)

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

$r_{t+1}; r_{t+2}; r_{t+3} \cdots$

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

Anticipated future reward at time t =
reward at time t + anticipated future rewards at time t

Based on Dayan slides; Daw slides

Temporal Difference Learning

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

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

Difference between what I anticipate at time t+1 and what I anticipate at time 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: $v_t \rightarrow v_t + \varepsilon(r_t + v_{t+1} - v_t)$

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

RV versus TD

- Rescorla-Wagner error: (n represents trial)

$$\delta_n = r_n - v_n$$

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

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

Name comes from!

Temporal Difference Learning

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

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

Name comes from!

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

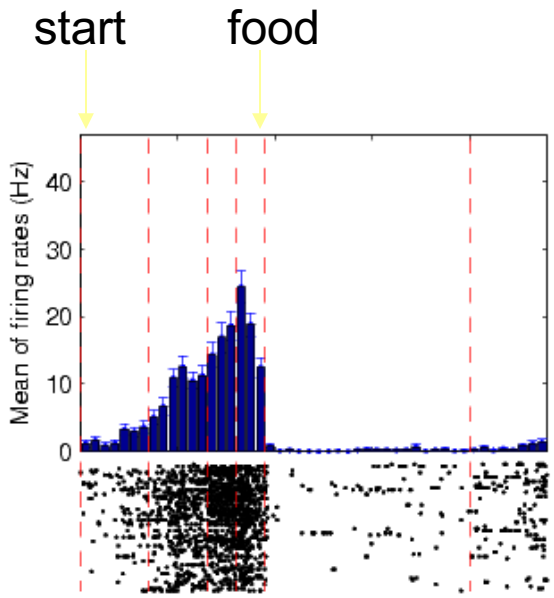
$v_{t+1} > v_t$ Got better

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

Based on Daw slides

Temporal Difference Learning

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:

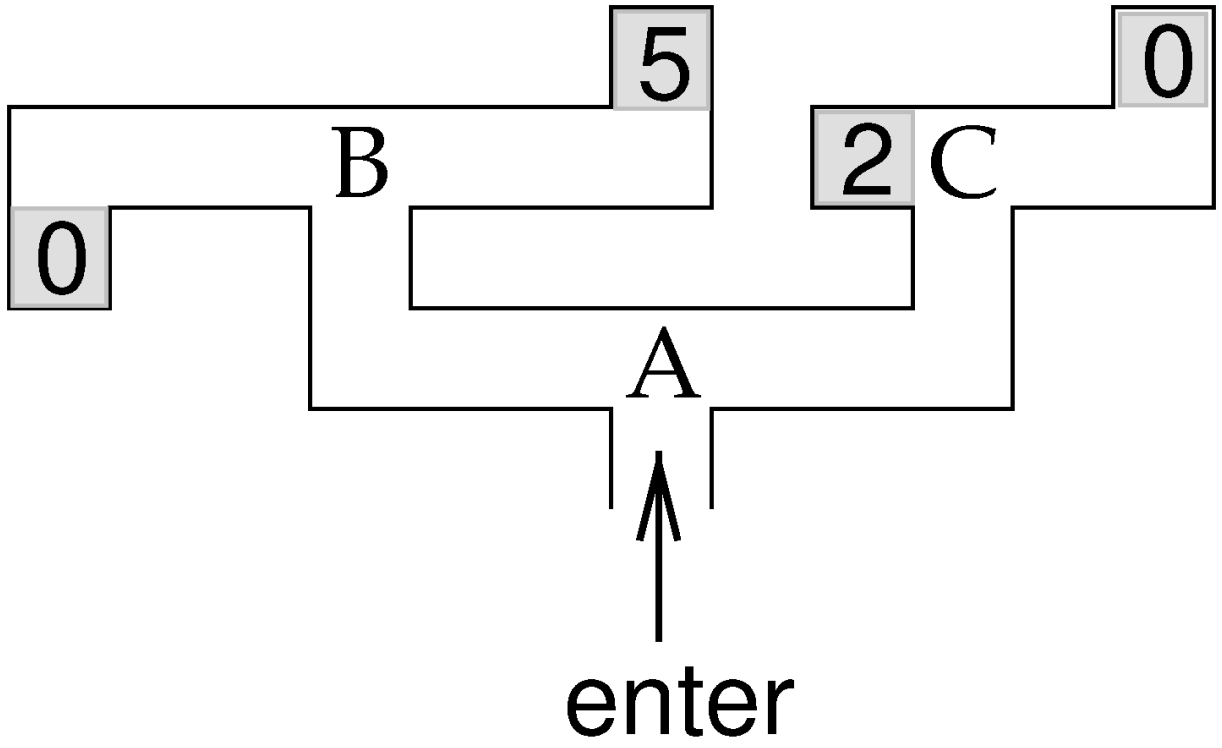
- Problem: Predict future reward
- Algorithm: Temporal Difference Learning (generalization of Rescorla-Wagner)
- Implementation: Dopamine neurons signaling error in reward prediction

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

Reward based on sequence of actions

Recent example in machine learning

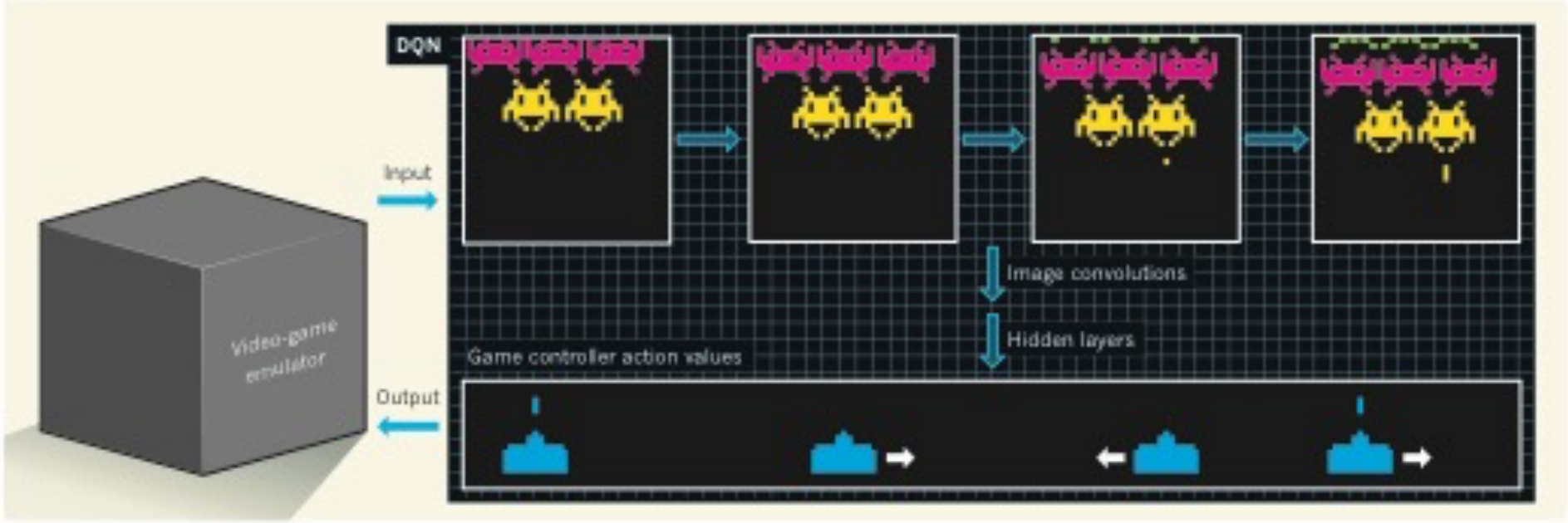
LETTER

doi:10.1038/nature14236

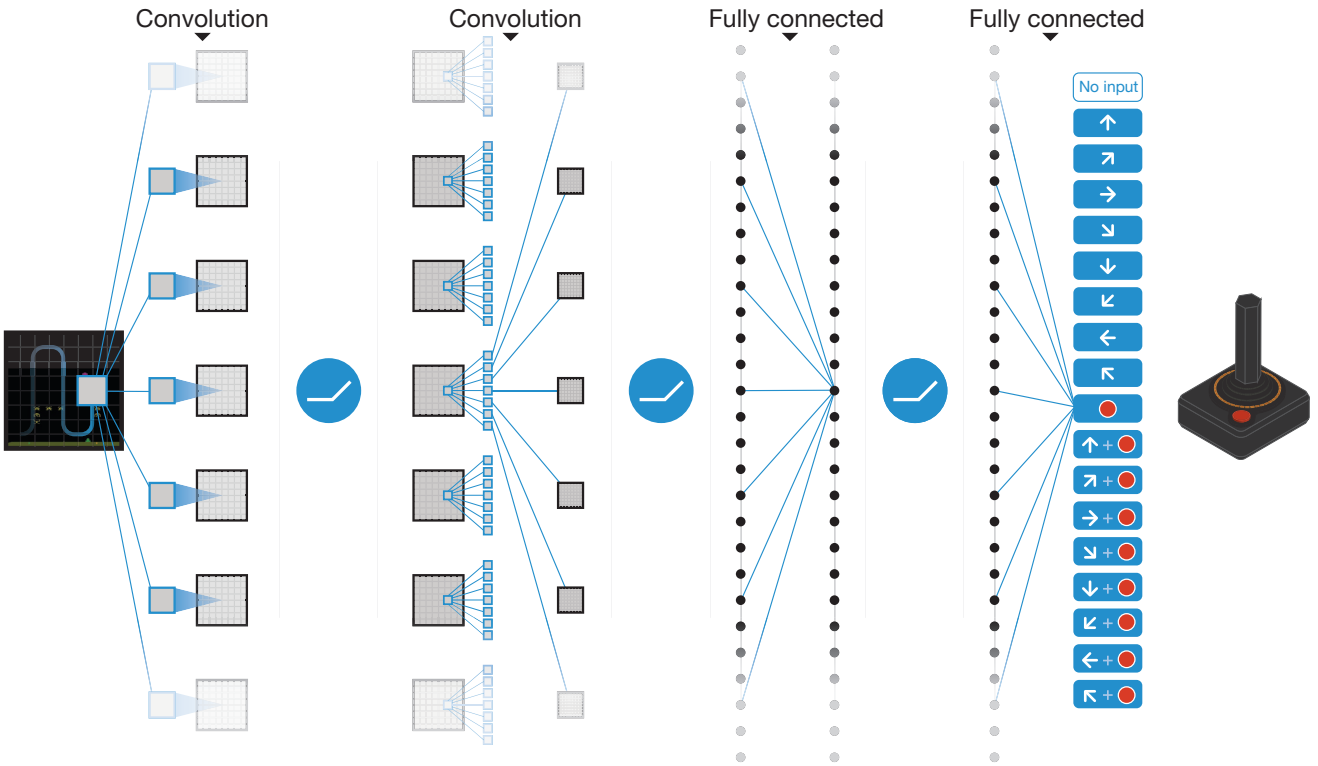
Human-level control through deep reinforcement learning

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]

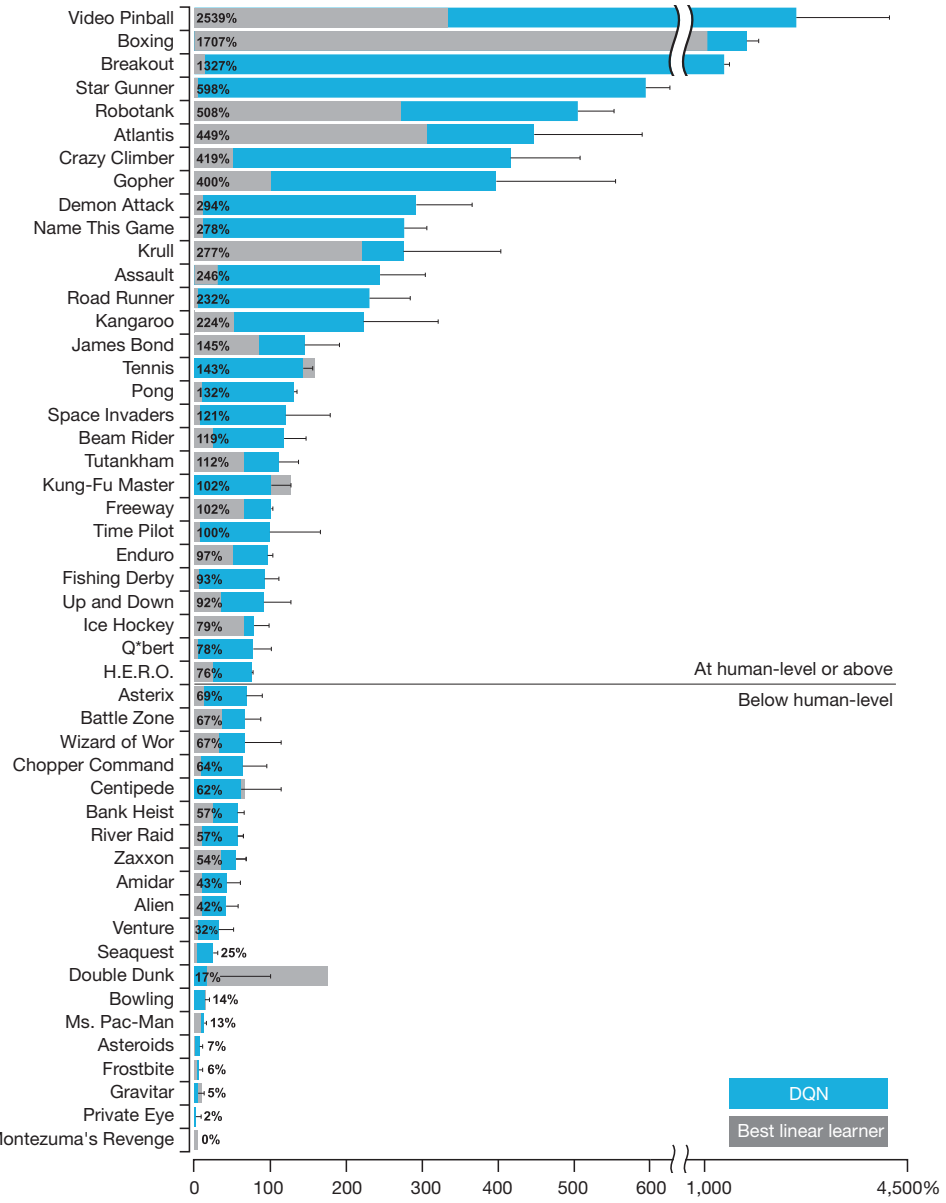
Mnih et al. Nature 518, 529–533; **2015**



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



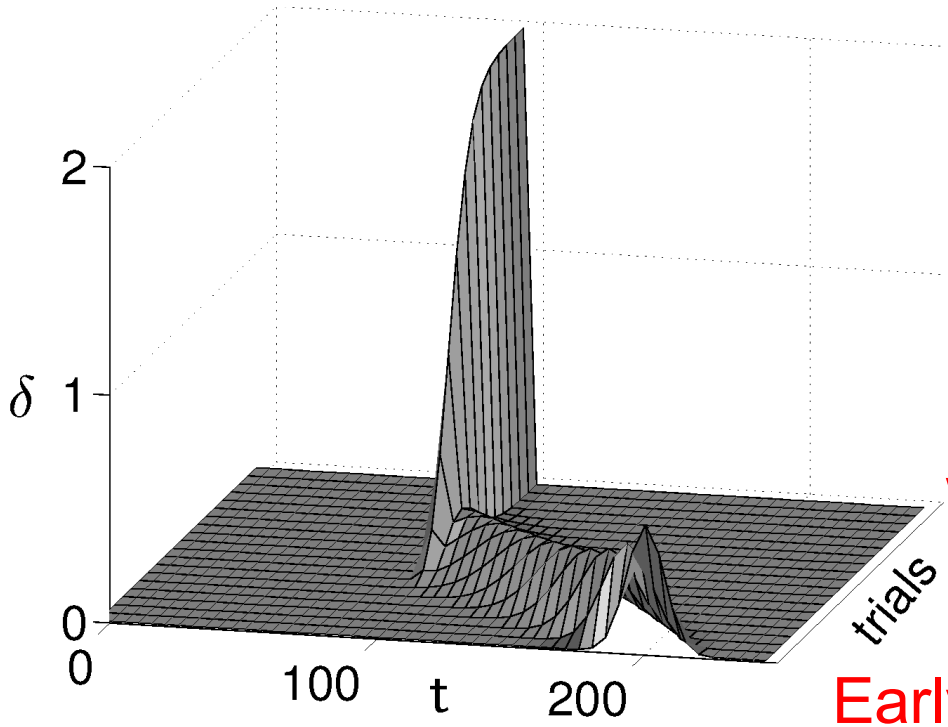
Mnih et al. Nature 518, 529–533; **2015**



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

Silver et al. 2016

Temporal Difference Learning



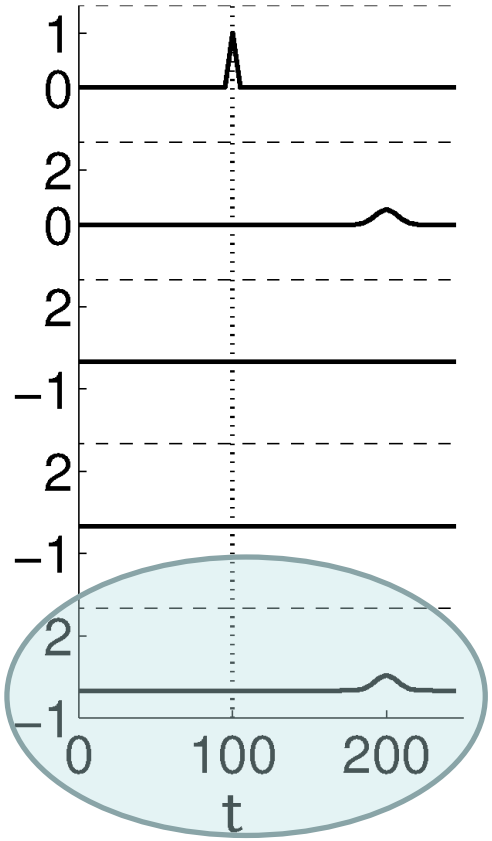
Late trials: error when stimulus

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



u_t Stimulus

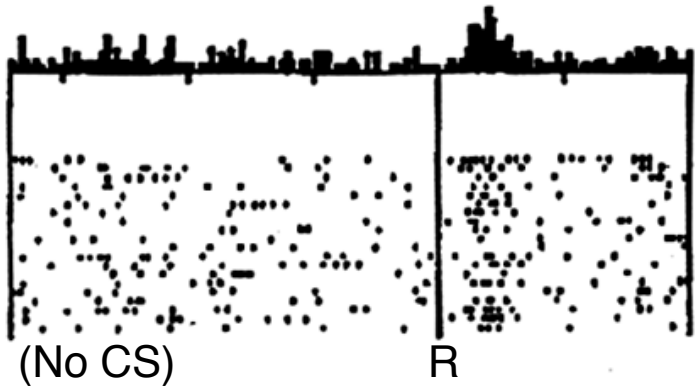
r_t Actual reward

v_t Prediction total reward from t to end of trial

$v_{t+1} - v_t$ Difference t+1 and t

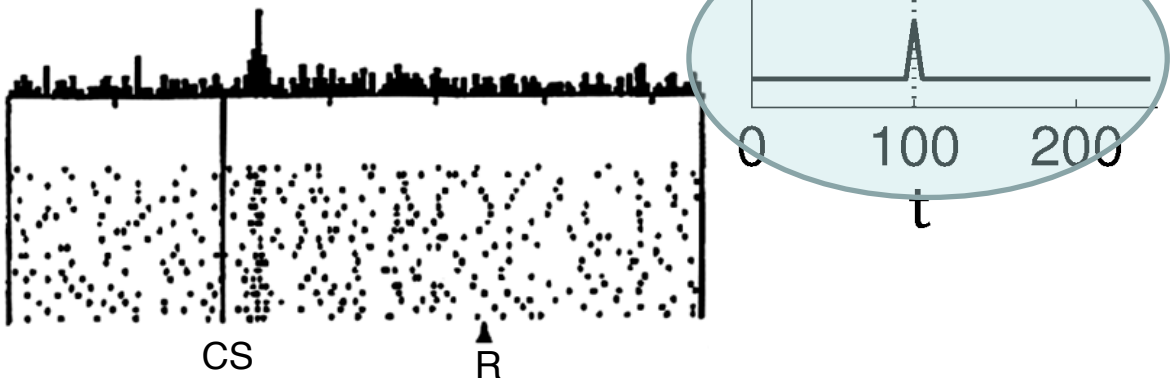
$\delta_t = r_t + v_{t+1} - v_t$

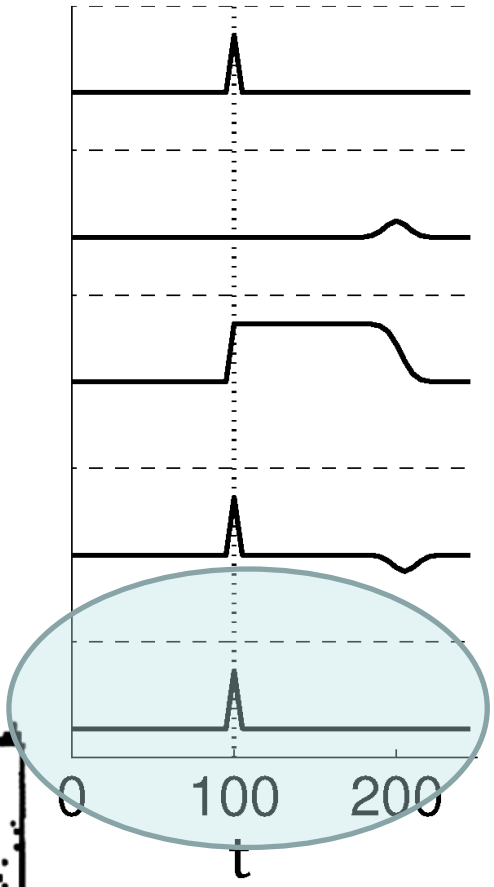
Prediction error



Temporal Difference Learning

After learning





u_t Stimulus

r_t Actual reward

v_t Prediction total reward from t to end of trial

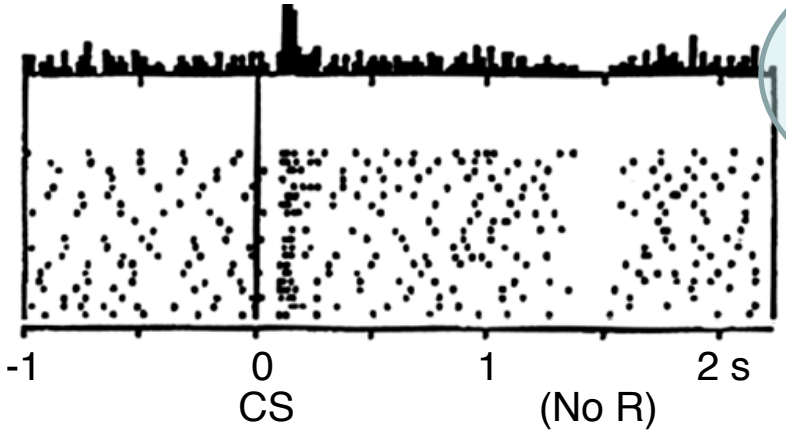
$v_{t+1} - v_t$ Difference t+1 and t

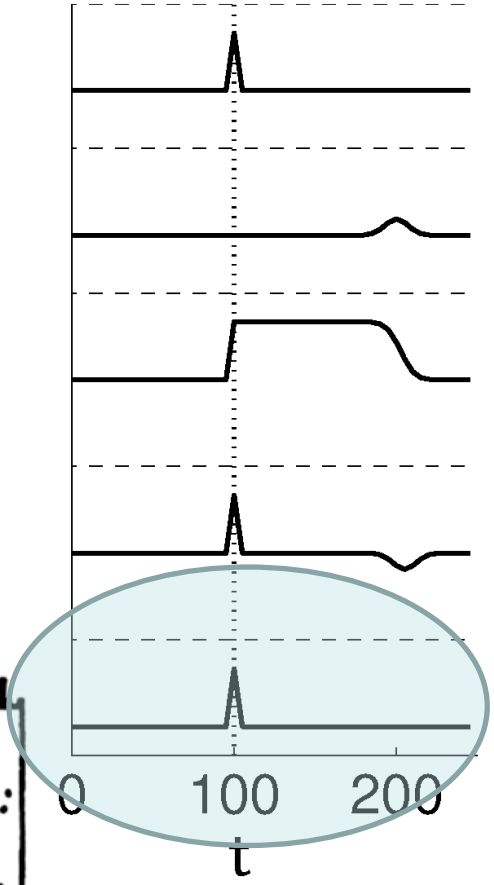
$\delta_t = r_t + v_{t+1} - v_t$

Prediction error

Temporal Difference Learning

After learning





u_t Stimulus

r_t Actual reward

v_t Prediction total reward from t to end of trial

$v_{t+1} - v_t$ Difference t+1 and t

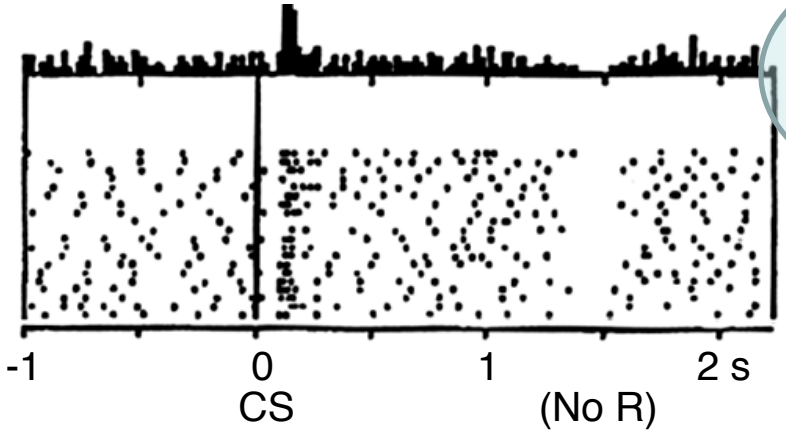
$\delta_t = r_t + v_{t+1} - v_t$

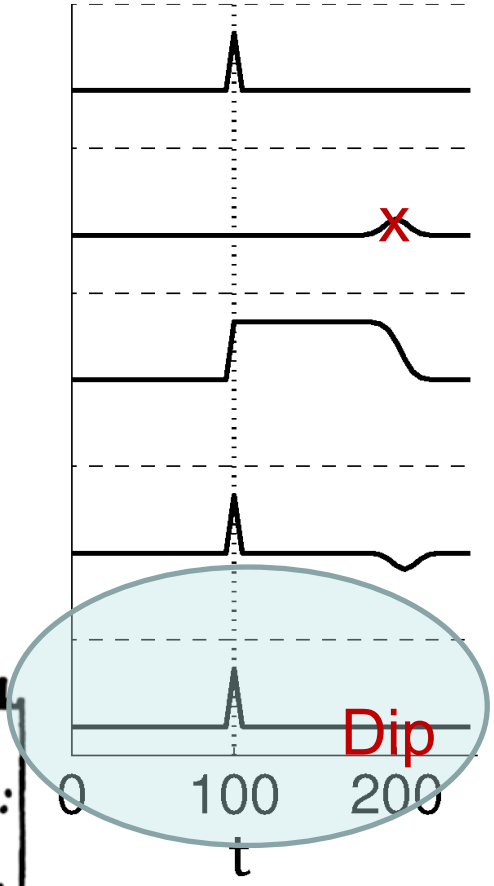
Prediction error

What should change in figure above to match data?

Temporal Difference Learning

After learning, and no reward





u_t Stimulus

r_t Actual reward

v_t Prediction total reward from t to end of trial

$v_{t+1} - v_t$ Difference t+1 and t

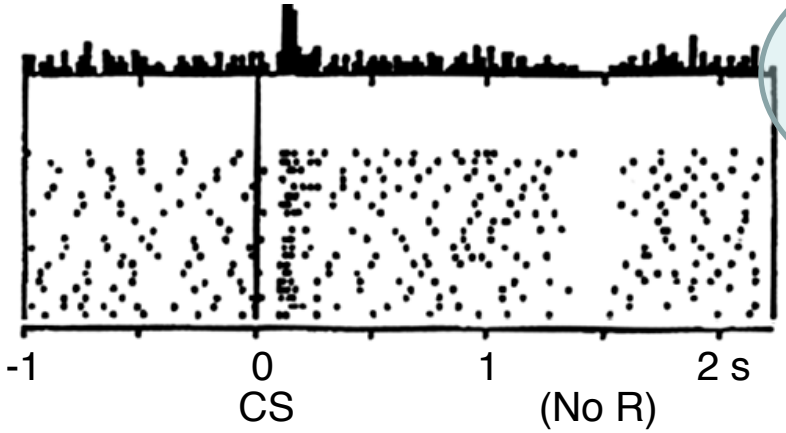
$\delta_t = r_t + v_{t+1} - v_t$

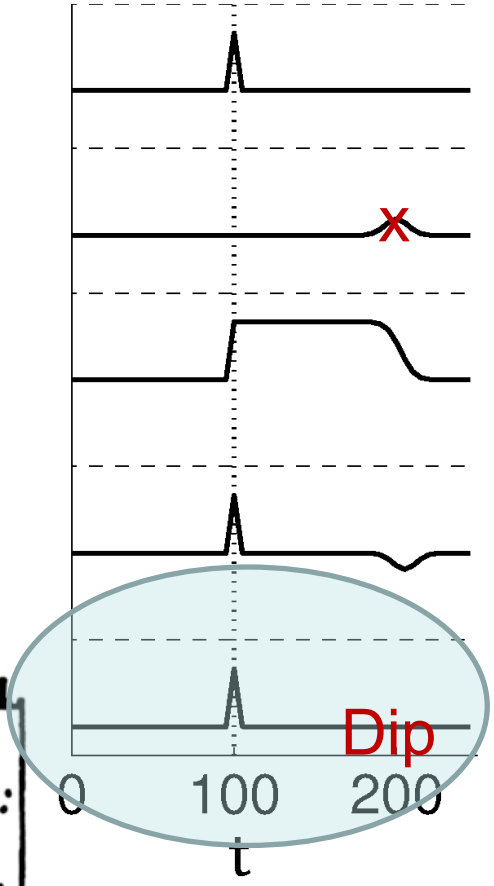
Prediction error

Here reward is 0 and
Prediction error should dip

Temporal Difference Learning

After learning, and no reward





u_t Stimulus

r_t Actual reward

v_t Prediction total reward from t to end of trial

$v_{t+1} - v_t$ Difference t+1 and t

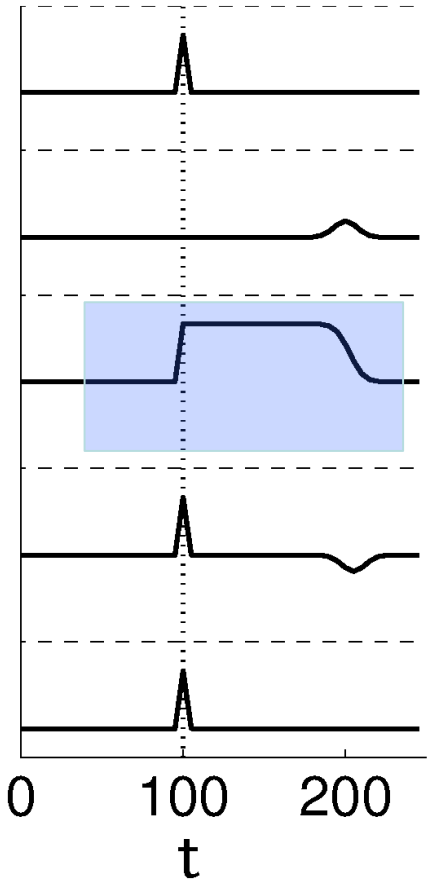
$\delta_t = r_t + v_{t+1} - v_t$

Prediction error

Neural responses related to prediction error

Temporal Difference Learning

After learning



What about anticipation of future rewards?