Reinforcement Learning

Reinforcement Learning

Reinforcement Learning:

Learning behaviors from reward signals

Types of Learning

less ← **Amount of Feedback** → more

Unsupervised

Extracting statistical structure from unlabeled data

Supervised

Learning with ground truth feedback

Types of Learning

less ← **Amount of Feedback** → more

Unsupervised

Extracting statistical structure
from unlabeled data

ICA
Sparse Coding
Boltzmann Machines
Hopfield Networks
PCA

Supervised

Learning with ground truth feedback

Perceptrons
CNNs
Deep Learning

Types of Learning

less ← **Amount of Feedback** → more

Unsupervised

Extracting statistical structure
from unlabeled data

ICA
Sparse Coding
Boltzmann Machines
Hopfield Networks
PCA

Reinforcement

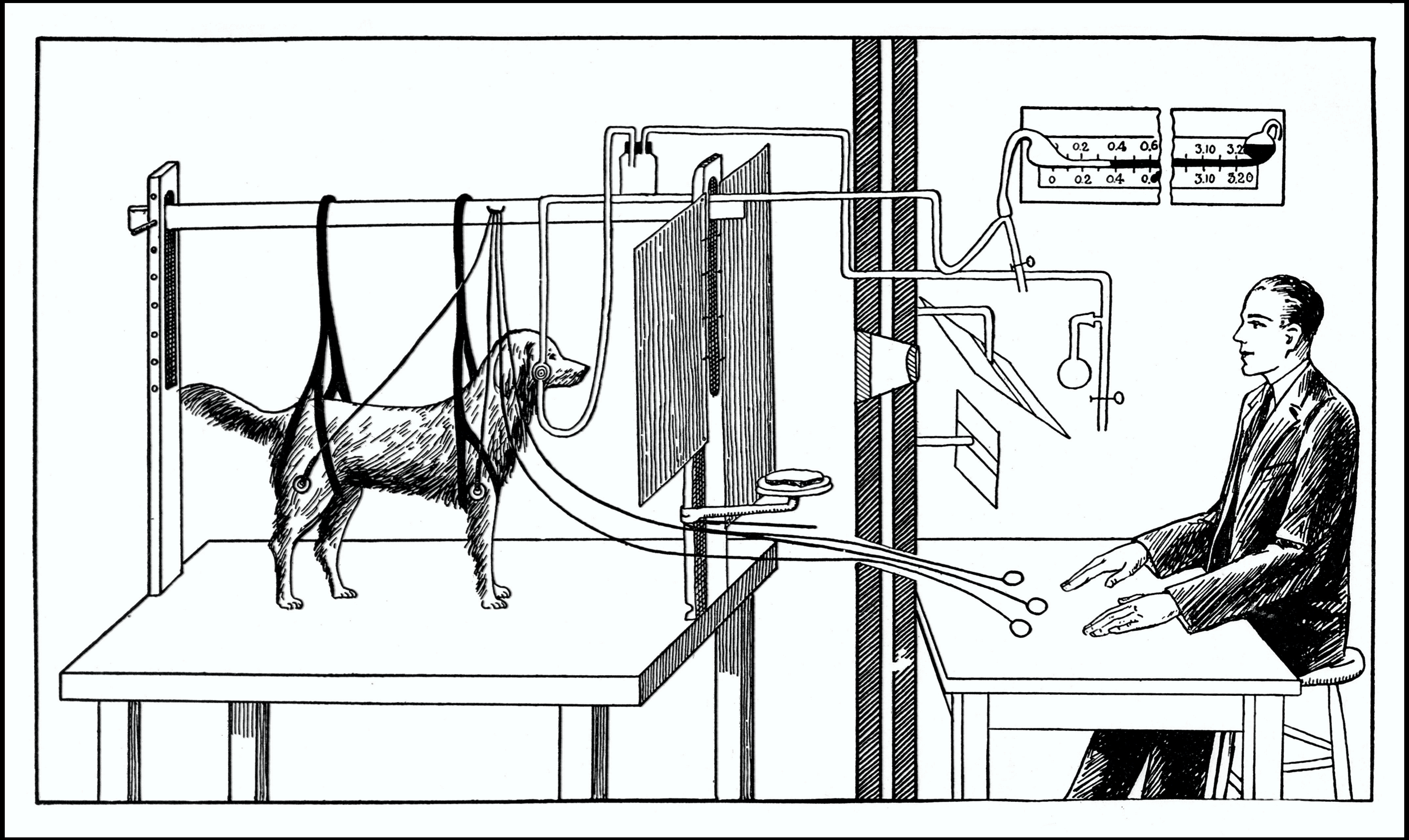
Learning behaviors
from reward signals

Supervised

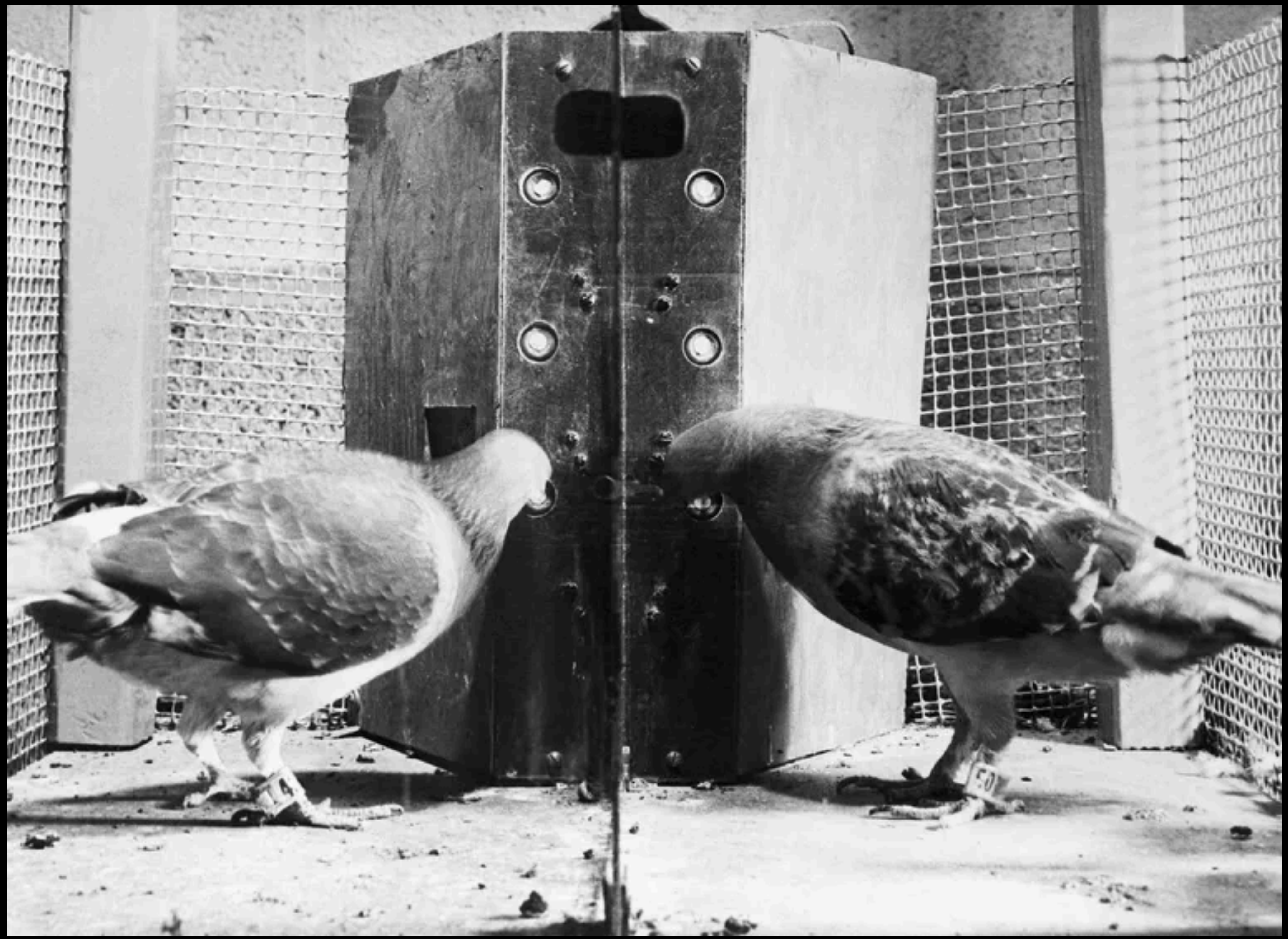
Learning with ground truth feedback

Perceptrons
CNNs
Deep Learning

Conditioning and Reinforcement



Conditioning and Reinforcement



Classical Conditioning Paradigms

	Pavlovian	Extinction	Partial	Blocking	Overshadow
Training	$A \rightarrow +$				
Test	$A \rightarrow +$				

Classical Conditioning Paradigms

	Pavlovian	Extinction	Partial	Blocking	Overshadow
Training	$A \rightarrow +$	$A \rightarrow +$ $A \rightarrow \cdot$			
Test	$A \rightarrow +$	$A \rightarrow \cdot$			

Classical Conditioning Paradigms

	Pavlovian	Extinction	Partial	Blocking	Overshadow
Training	$A \rightarrow +$	$A \rightarrow +$ $A \rightarrow \cdot$	$A \rightarrow +$ $AB \rightarrow +$		
Test	$A \rightarrow +$	$A \rightarrow \cdot$	$A \rightarrow \alpha +$		

Classical Conditioning Paradigms

	Pavlovian	Extinction	Partial	Blocking	Overshadow
Training	$A \to +$	$A \to +$ $A \to \cdot$	$A \to +$ $AB \to +$	$A \to +$ $AB \to +$	
Test	$A \to +$	$A \to \cdot$	$A \to \alpha+$	$A \to +$ $B \to \cdot$	

Classical Conditioning Paradigms

	Pavlovian	Extinction	Partial	Blocking	Overshadow
Training	$A \rightarrow +$	$A \rightarrow +$ $A \rightarrow \cdot$	$A \rightarrow +$ $AB \rightarrow +$	$A \rightarrow +$ $AB \rightarrow +$	$AB \rightarrow +$
Test	$A \rightarrow +$	$A \rightarrow \cdot$	$A \rightarrow \alpha+$	$A \rightarrow +$ $B \rightarrow \cdot$	$A \rightarrow \alpha+$ $B \rightarrow \beta+$

Rescorla-Wagner Model

s : Stimulus Variable

r : Actual Reward

v : Expected Reward

w : Weights

Rescorla & Wagner, 1972

Rescorla-Wagner Model

s : Stimulus Variable

r : Actual Reward

v : Expected Reward

w : Weights

Model

$$v = w^\top s$$

Rescorla & Wagner, 1972

Rescorla-Wagner Model

s : Stimulus Variable

r : Actual Reward

v : Expected Reward

w : Weights

Model

$$v = w^\top s$$

Objective

$$\min \sum \frac{1}{2}(r - v)^2$$

Rescorla & Wagner, 1972

Rescorla-Wagner Model

s : Stimulus Variable

r : Actual Reward

v : Expected Reward

w : Weights

Model

$$v = w^\top s$$

Objective

$$\min \sum \frac{1}{2}(r - v)^2$$

Learning Rule

$$\delta = r - v$$
$$w \rightarrow w + \epsilon \delta s$$

Rescorla & Wagner, 1972

Rescorla-Wagner Model

s : Stimulus Variable

r : Actual Reward

v : Expected Reward

w : Weights

Model	Objective	Learning Rule
$v = w^\top s$	$\min \sum \frac{1}{2}(r - v)^2$	$\boxed{\delta = r - v}$ $w \rightarrow w + \epsilon \delta s$

Prediction Error

Rescorla & Wagner, 1972

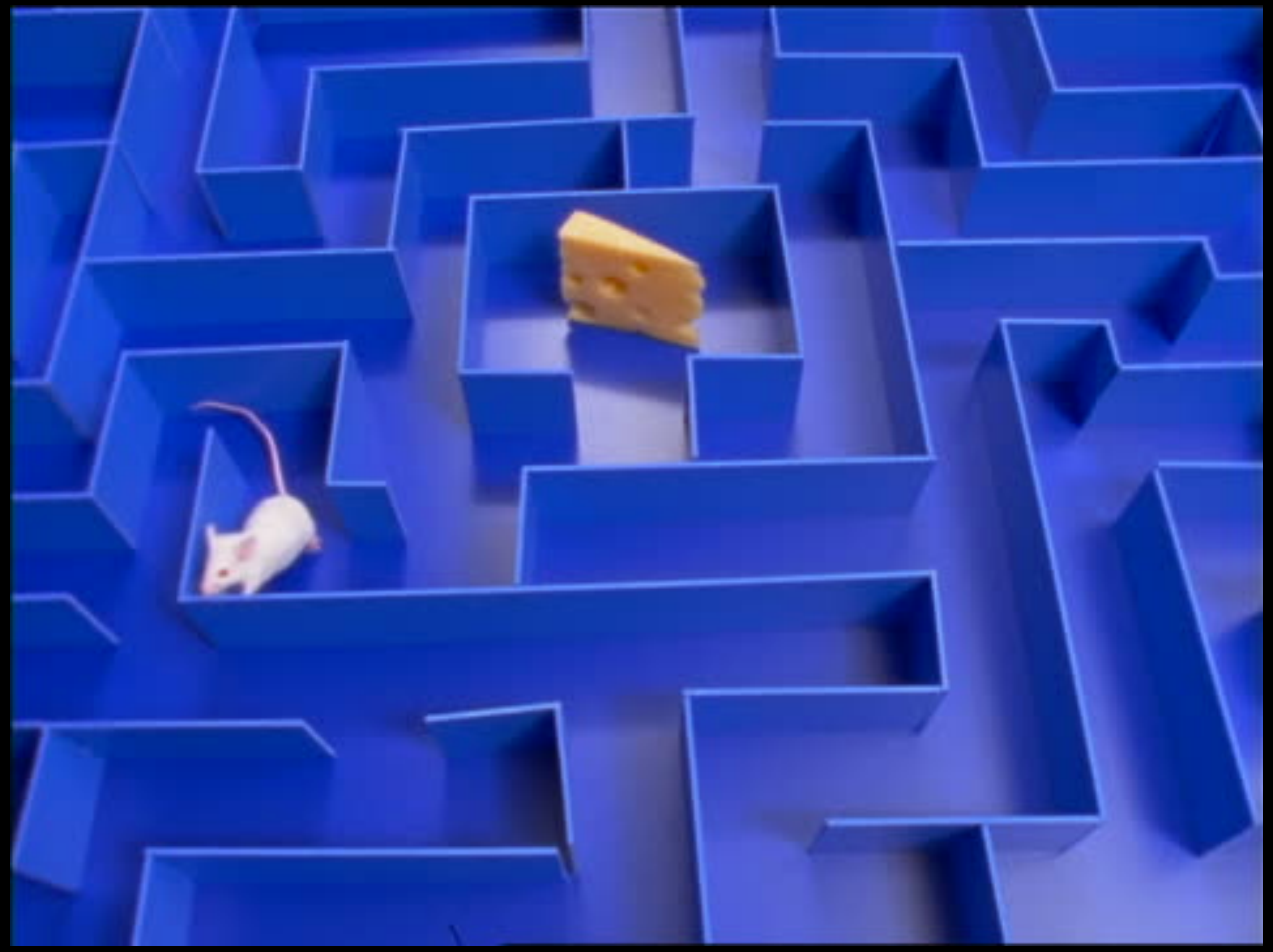
Classical Conditioning Paradigms

	Pavlovian	Extinction	Partial	Blocking	Overshadow
Training	$A \to +$	$A \to +$ $A \to \cdot$	$A \to +$ $AB \to +$	$A \to +$ $AB \to +$	$AB \to +$
Test	$A \to +$	$A \to \cdot$	$A \to \alpha +$	$A \to +$ $B \to \cdot$	$A \to \alpha +$ $B \to \beta +$

Model	Objective	Learning Rule
$v = w^\top s$	$\min \sum \frac{1}{2}(r - v)^2$	$\delta = r - v$ $w \to w + \epsilon \delta s$

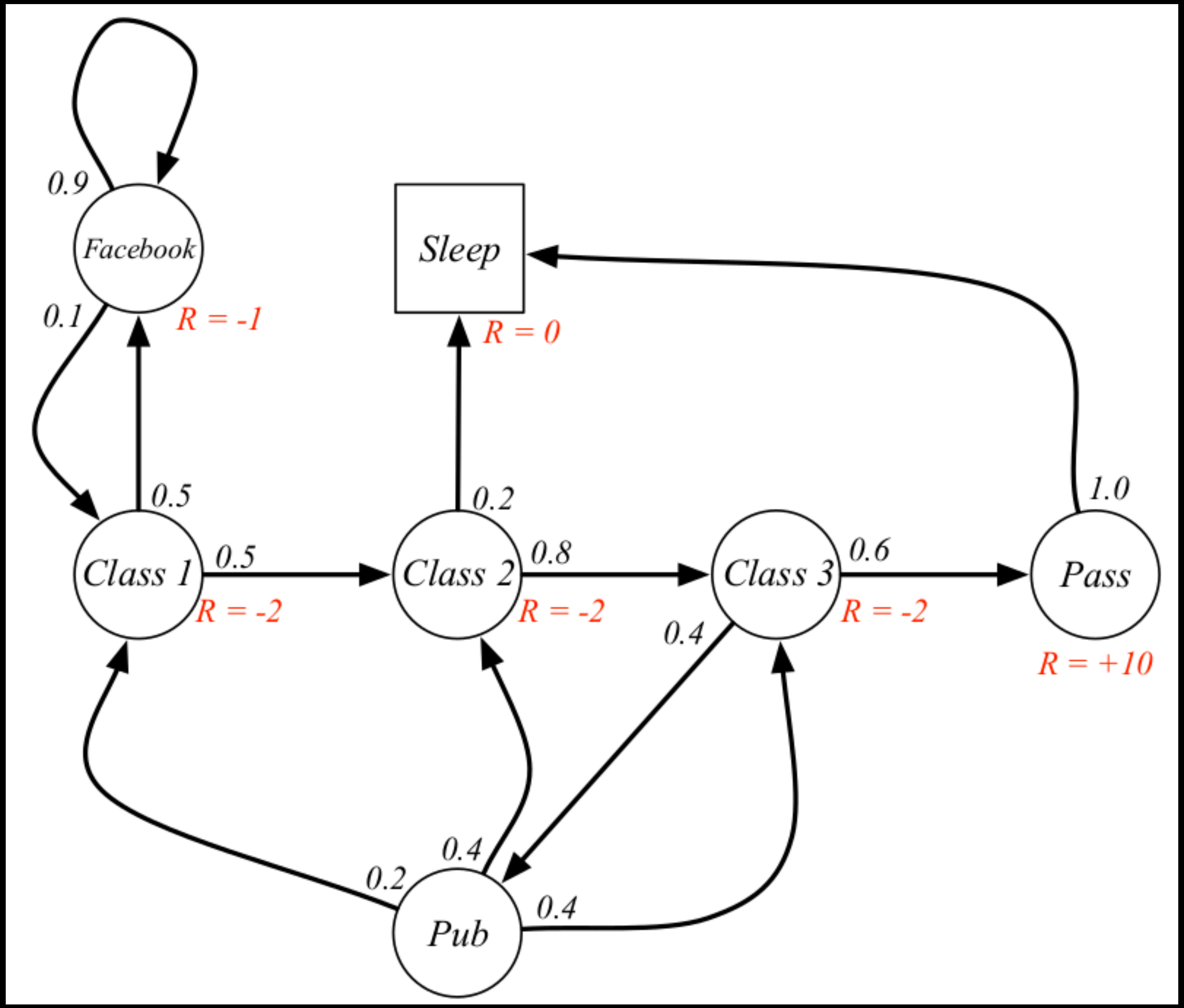
Predicting Future Reward:

Temporal Difference Learning



Bringing an Agent in the Loop

Bringing an Agent in the Loop



Predicting Future Reward:

Temporal Difference Learning

s : Stimulus Variable

r : Actual Reward

v : Expected Reward

w : Weights

Sutton & Barto, 1990

Predicting Future Reward:

Temporal Difference Learning

s_t : Stimulus Variable

r_t : Actual Reward

V_t : Expected Reward

w_t : Weights

Sutton & Barto, 1990

Predicting Future Reward:

Temporal Difference Learning

s_t : Stimulus Variable

r_t : Actual Reward

V_t : Expected Reward

w_t : Weights

Total Expected Reward

$$V_t = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i r_{t+i}\right]$$

Sutton & Barto, 1990

Predicting Future Reward:

Temporal Difference Learning

s_t : Stimulus Variable

r_t : Actual Reward

V_t : Expected Reward

w_t : Weights

Total Expected Reward

$$V_t = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i r_{t+i}\right]$$

Bellman Equation

$$V_t = r_t + \gamma \mathbb{E}\left[V_{t+1}\right]$$

Sutton & Barto, 1990

Predicting Future Reward:

Temporal Difference Learning

s_t : Stimulus Variable

r_t : Actual Reward

V_t : Expected Reward

w_t : Weights

Model

$$\hat{V}_t = w_t^\top s_t$$

Total Expected Reward

$$V_t = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i r_{t+i}\right]$$

Bellman Equation

$$V_t = r_t + \gamma \mathbb{E}\left[V_{t+1}\right]$$

Sutton & Barto, 1990

Predicting Future Reward:

Temporal Difference Learning

s_t : Stimulus Variable

r_t : Actual Reward

V_t : Expected Reward

w_t : Weights

Model

$$\hat{V}_t = w_t^\top s_t$$

Objective

$$\min \sum \frac{1}{2}(V_t - \hat{V}_t)^2$$

Total Expected Reward

$$V_t = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i r_{t+i}\right]$$

Bellman Equation

$$V_t = r_t + \gamma \mathbb{E}\left[V_{t+1}\right]$$

Sutton & Barto, 1990

Predicting Future Reward:

Temporal Difference Learning

s_t : Stimulus Variable

r_t : Actual Reward

V_t : Expected Reward

w_t : Weights

Model

$$\hat{V}_t = w_t^\top s_t$$

Objective

$$\min \sum \frac{1}{2}(V_t - \hat{V}_t)^2$$

Total Expected Reward

$$V_t = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i r_{t+i}\right]$$

Bellman Equation

$$V_t = r_t + \gamma \mathbb{E}\left[V_{t+1}\right]$$

Prediction Error

$$\delta_t = r_t + \gamma \mathbb{E}\left[V_{t+1}\right] - \hat{V}_t$$

Sutton & Barto, 1990

Predicting Future Reward:

Temporal Difference Learning

s_t : Stimulus Variable

r_t : Actual Reward

V_t : Expected Reward

w_t : Weights

Model	Objective	Update Rule
$\hat{V}_t = w_t^\top s_t$	$\min \sum \frac{1}{2}(V_t - \hat{V}_t)^2$	$w_{t+1} = w_t + \epsilon s_t \delta_t$

Total Expected Reward

$$V_t = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i r_{t+i}\right]$$

Bellman Equation

$$V_t = r_t + \gamma \mathbb{E}\left[V_{t+1}\right]$$

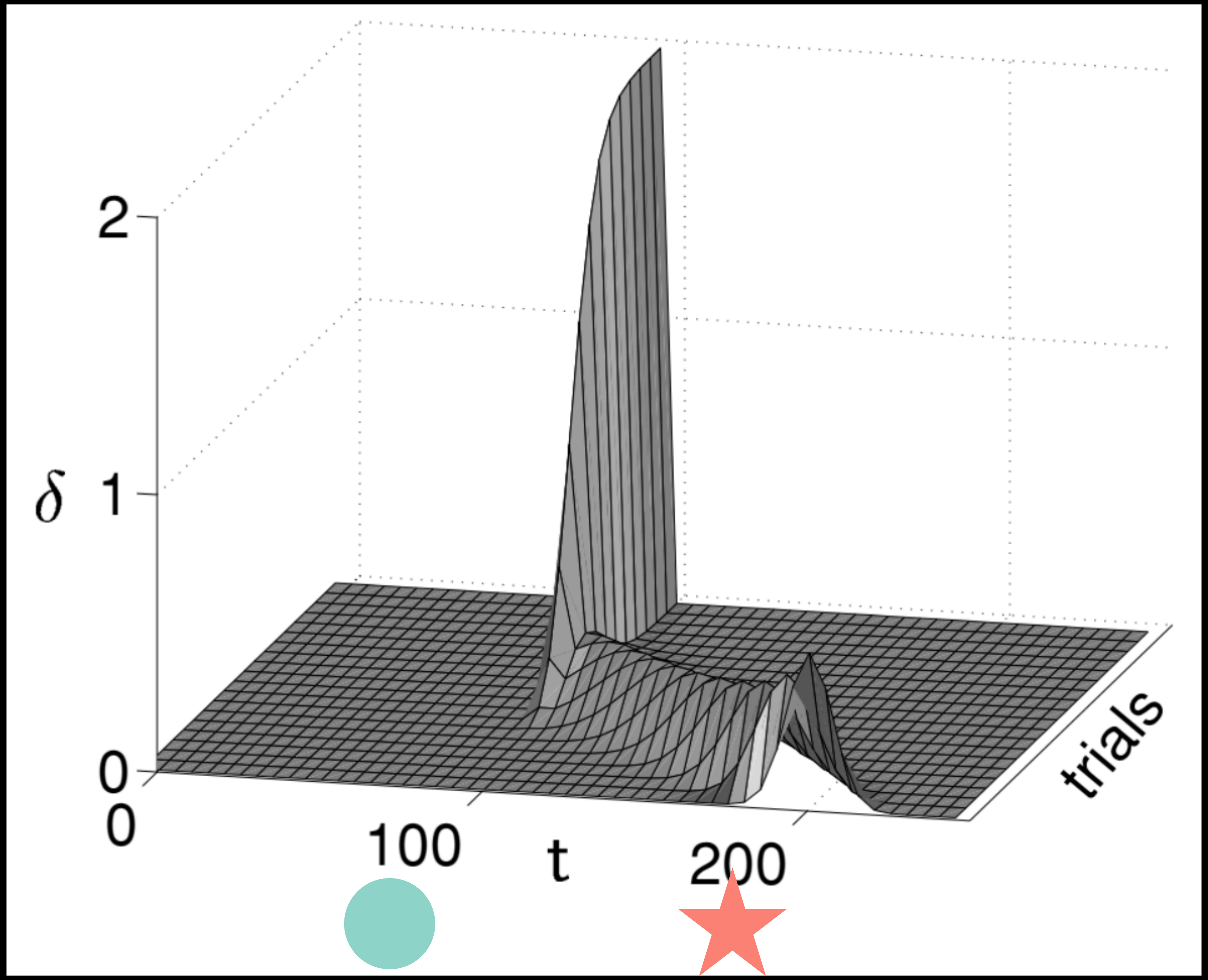
Prediction Error

$$\delta_t = r_t + \gamma \mathbb{E}\left[V_{t+1}\right] - \hat{V}_t$$

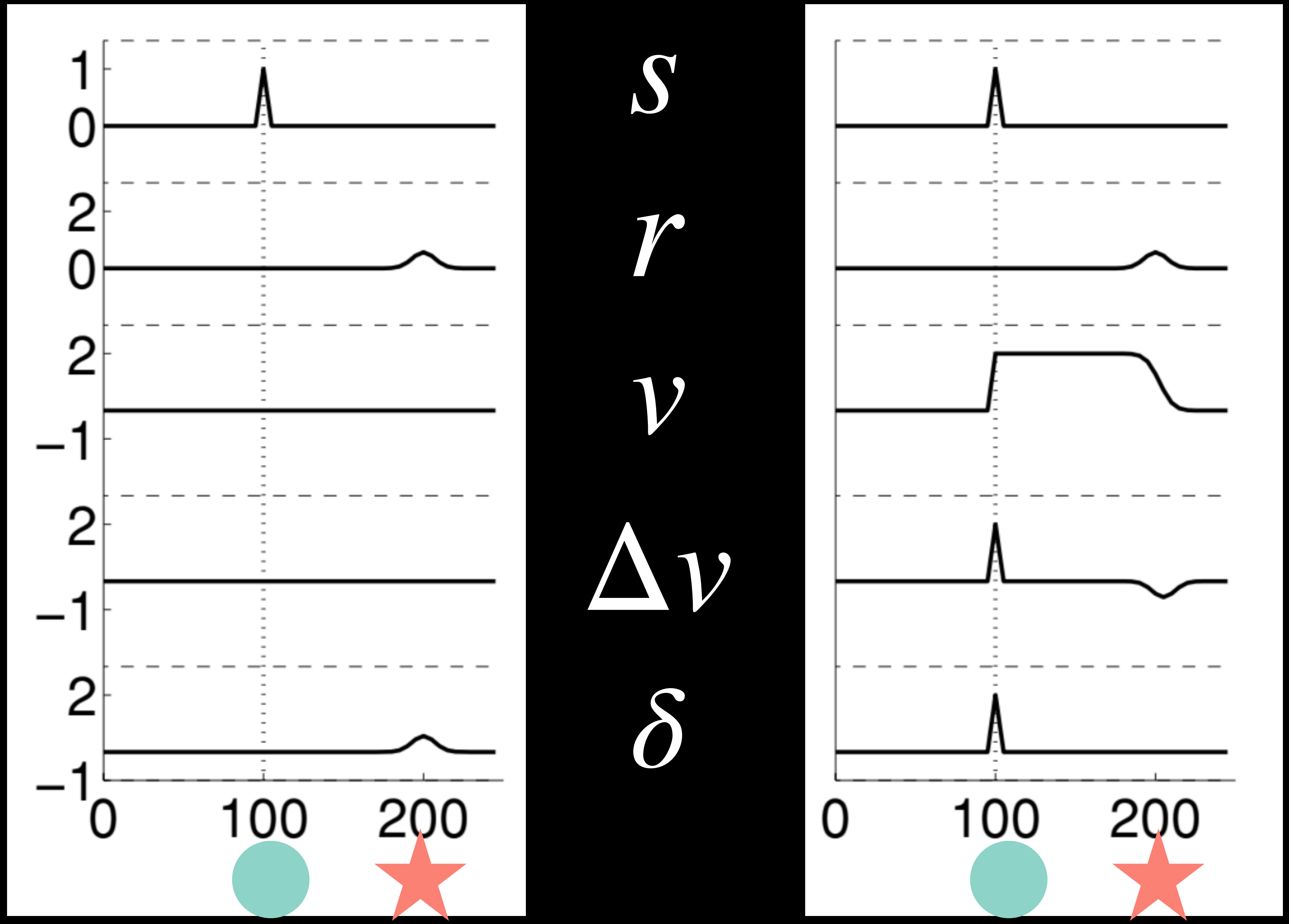
Sutton & Barto, 1990

Predicting Future Reward:

Temporal Difference Learning



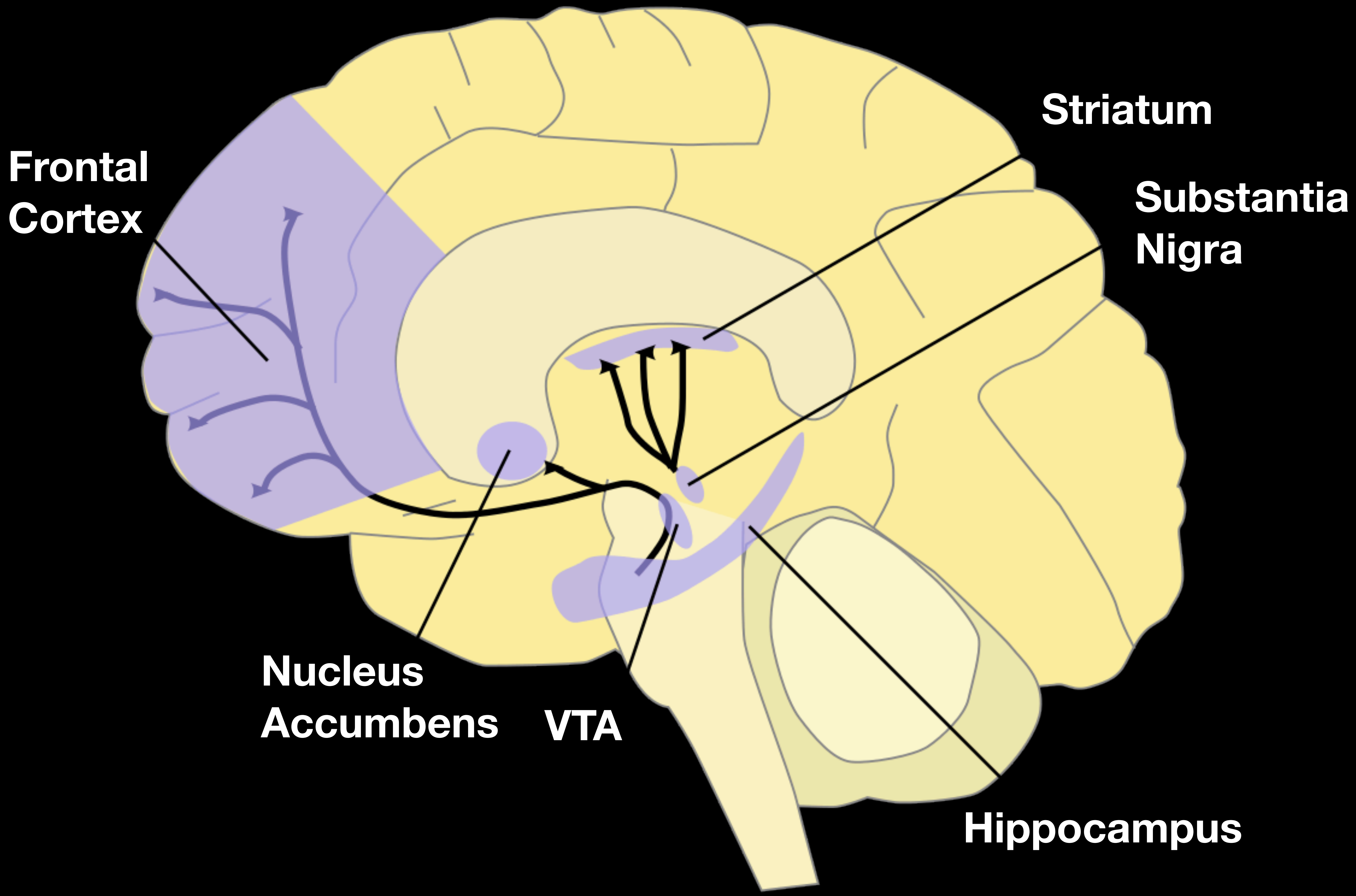
Prediction Error



Before Training

After Training

TD Learning in the Brain



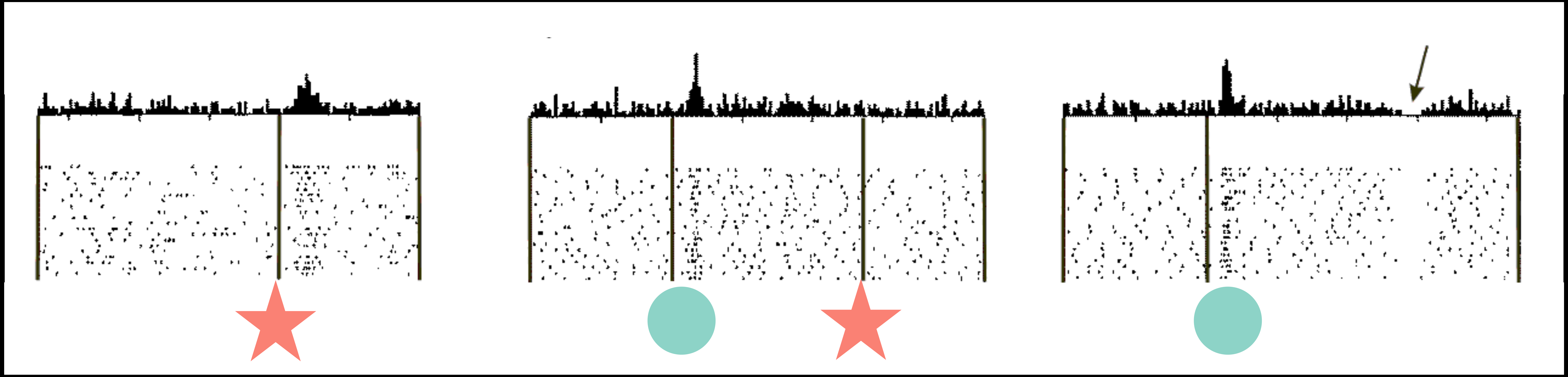
TD Learning in the Brain

TD Learning in the Brain

Novelty Response
Reward, No Stimulus

After Learning
Stimulus + Reward

After Learning
Stimulus, No Reward

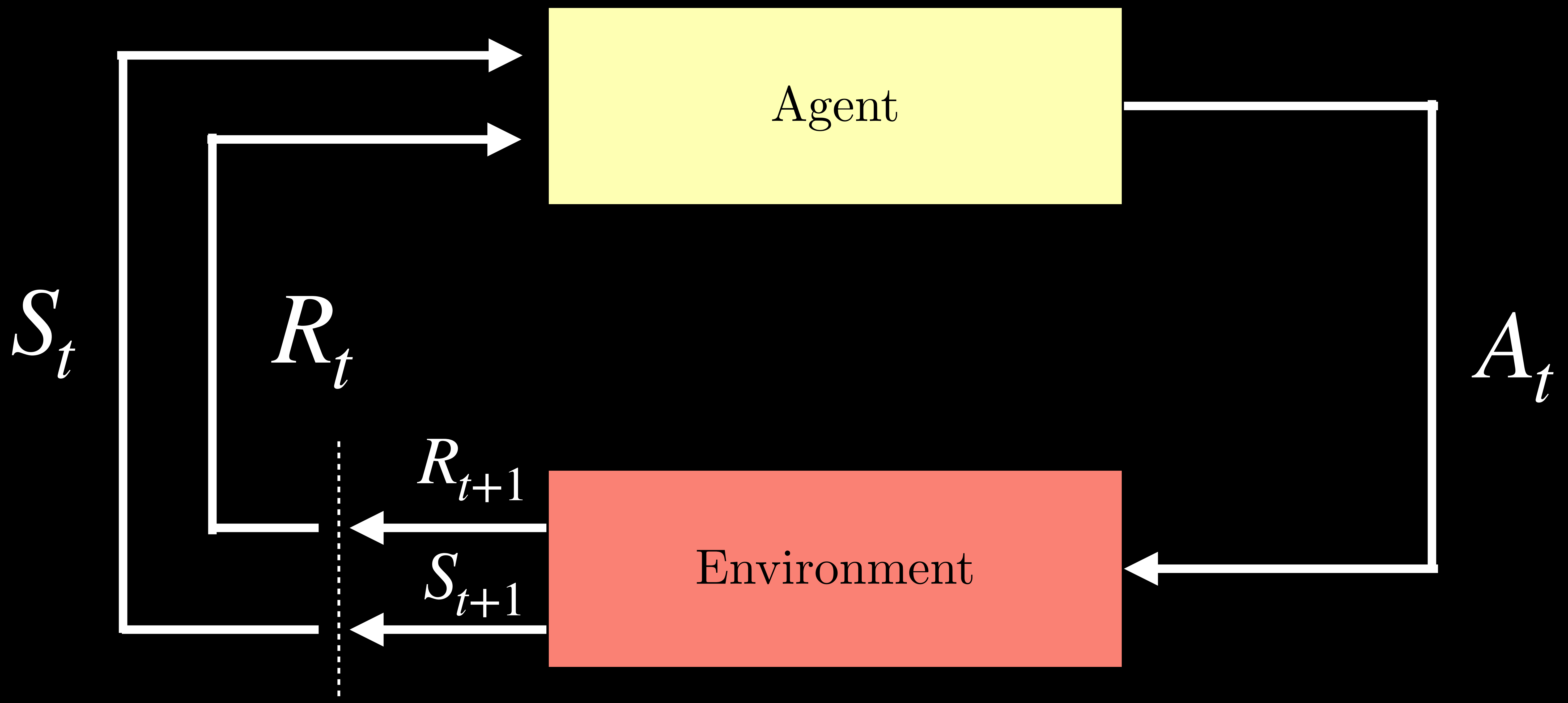


Bringing an Agent in the Loop

S_t : State

R_t : Reward

A_t : Actions



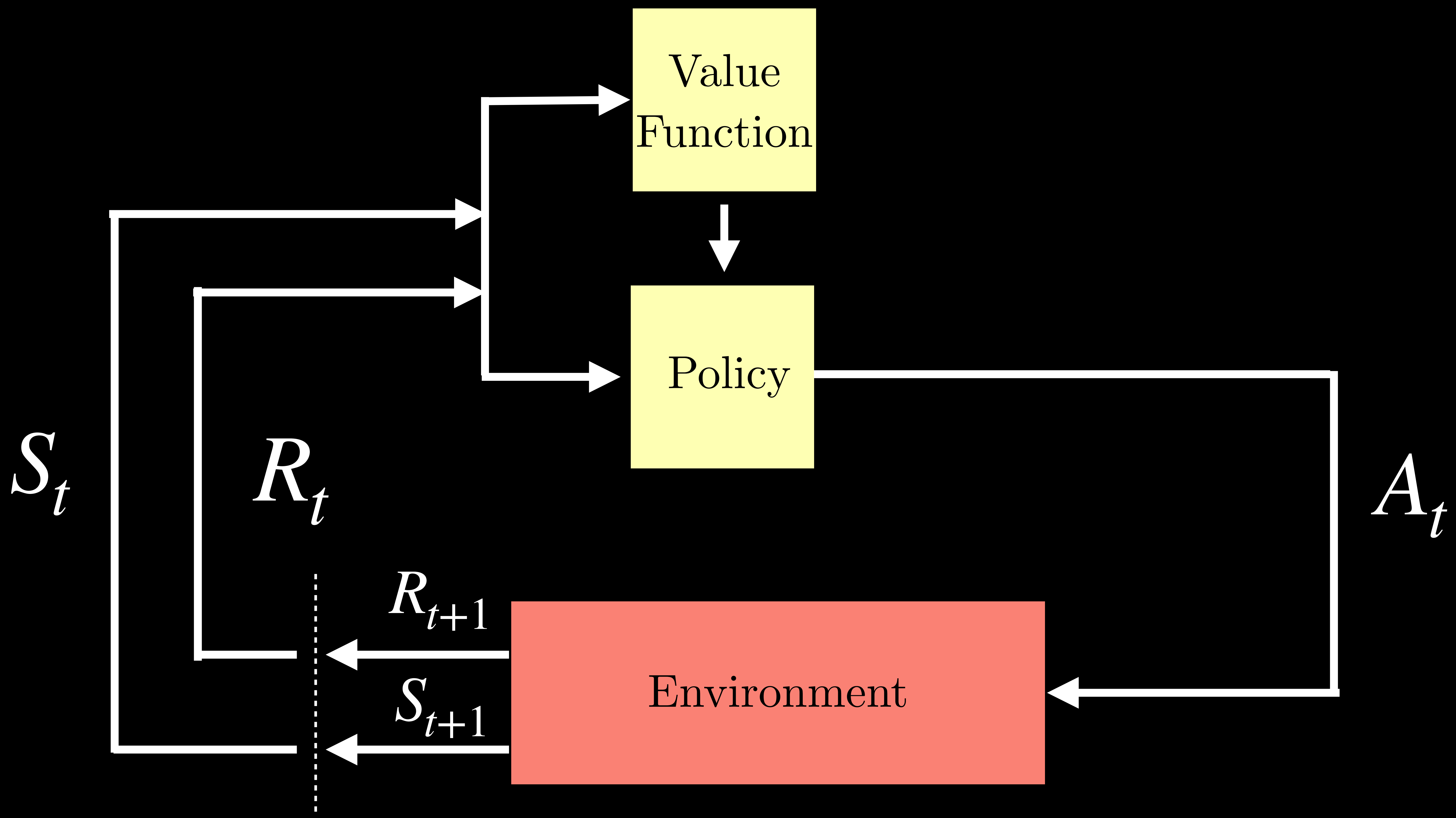
Bringing an Agent in the Loop

The Actor-Critic Algorithm

S_t : State

R_t : Reward

A_t : Actions



Bringing an Agent in the Loop

The Actor-Critic Algorithm

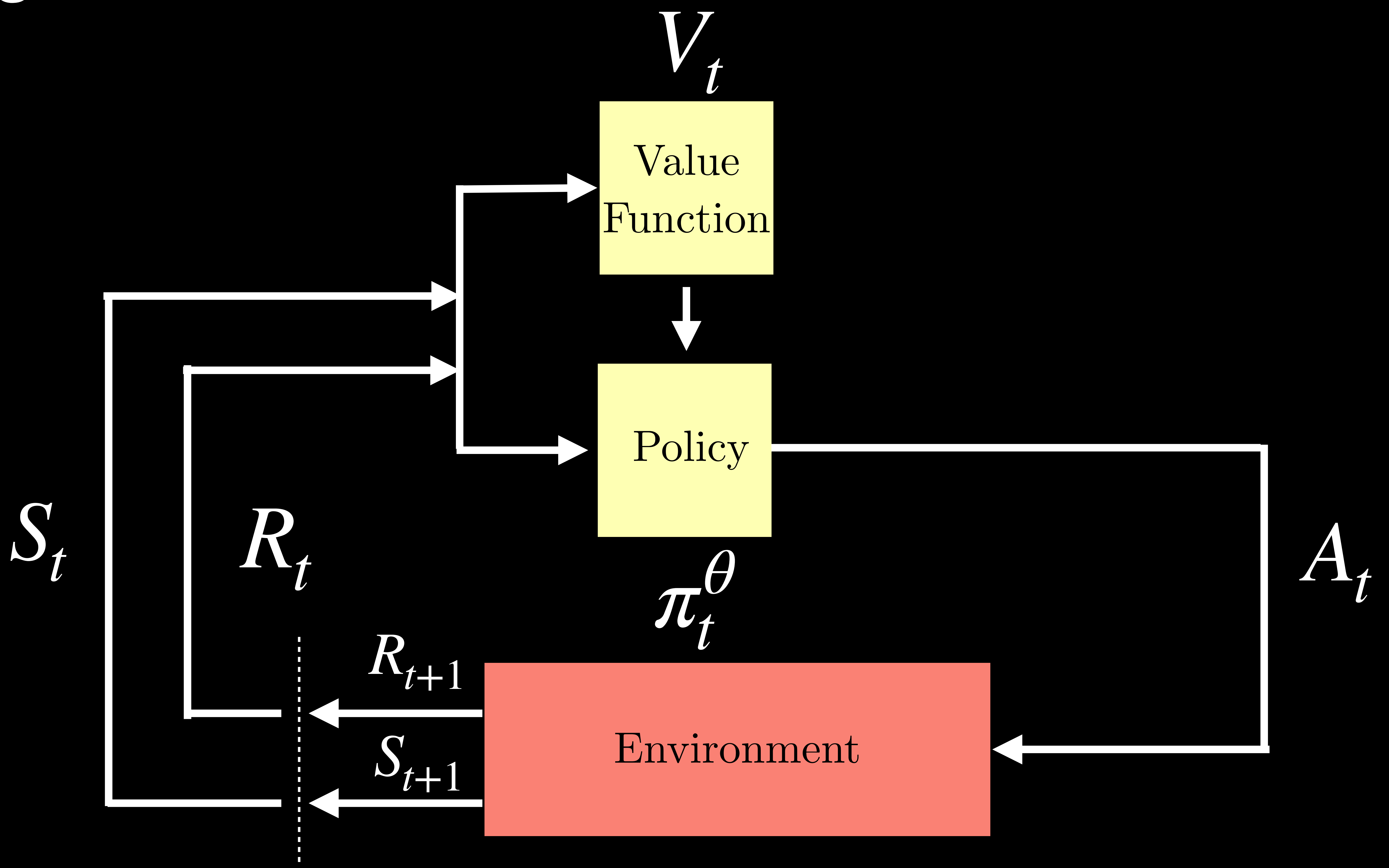
S_t : State

R_t : Reward

A_t : Actions

π_t^θ : Policy

V_t : Value Function



Bringing an Agent in the Loop

The Actor-Critic Algorithm

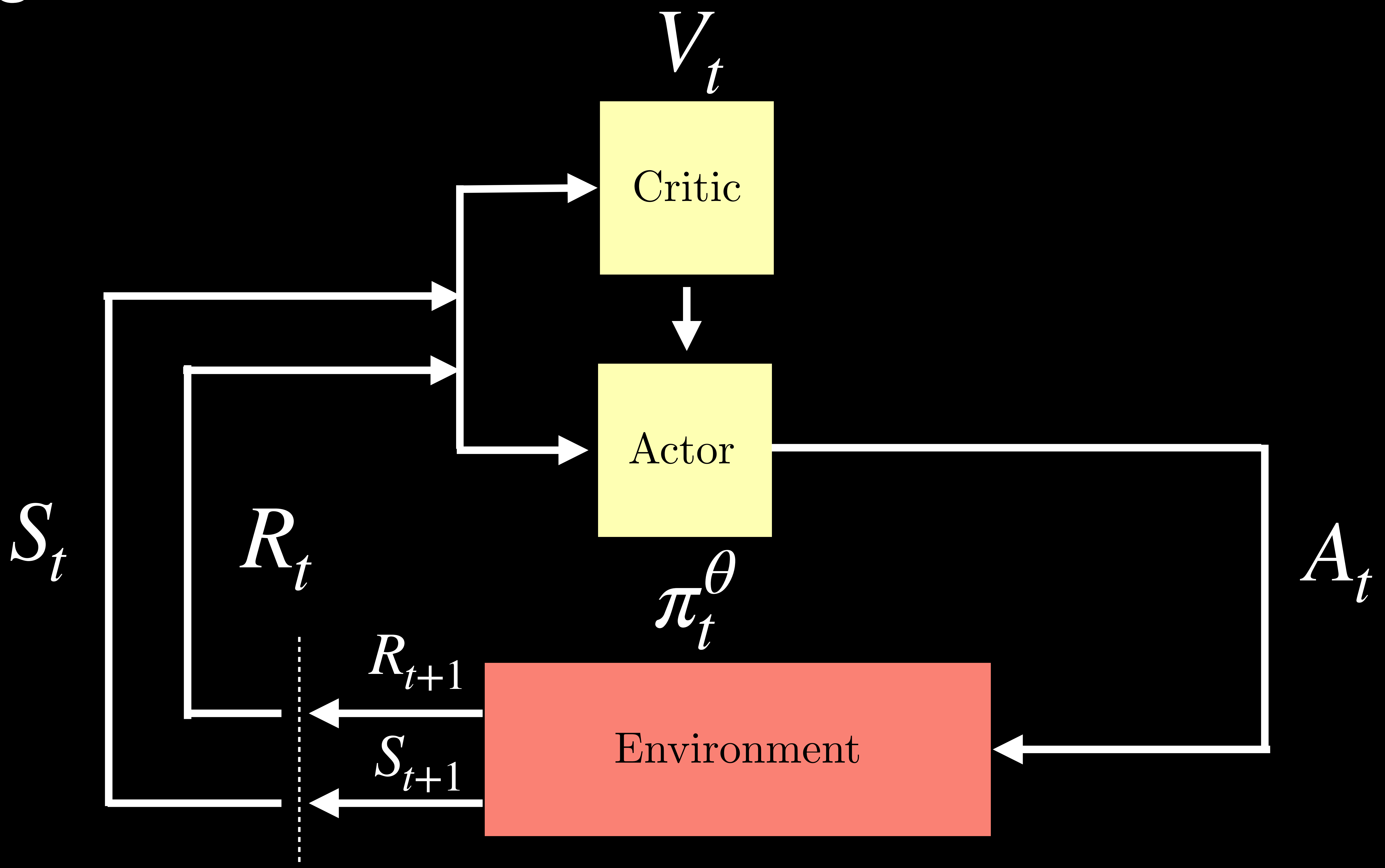
S_t : State

R_t : Reward

A_t : Actions

π_t^θ : Policy

V_t : Value Function



Bringing an Agent in the Loop

The Actor-Critic Algorithm

S_t : State

R_t : Reward

A_t : Actions

π_t^θ : Policy

V_t : Value Function

Policy

$$a_t \sim \pi_\theta$$

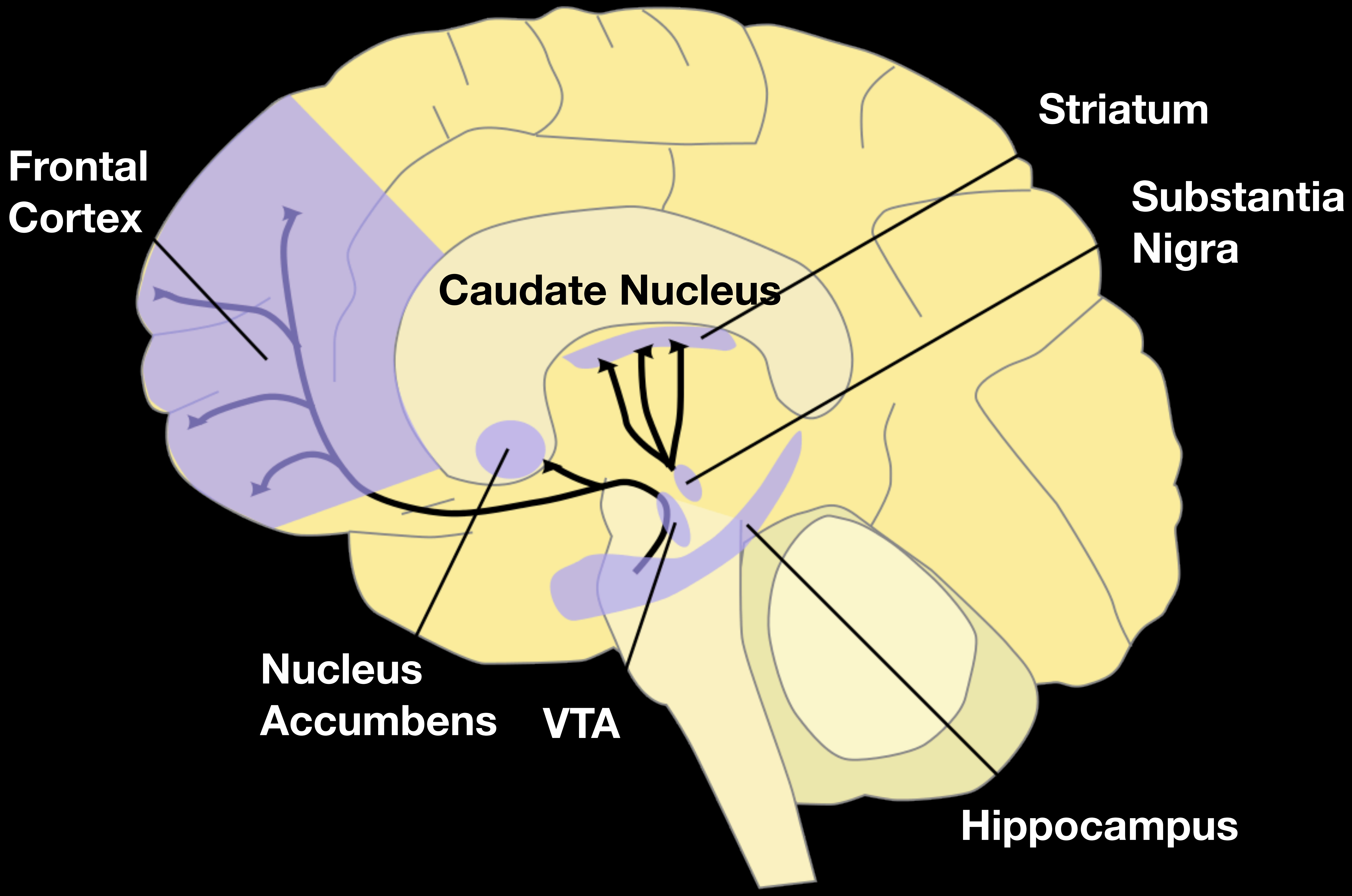
$$\pi_t^\theta = P(a_t = a \,|\, s_t = s)$$

$$P(a \,|\, s) = \frac{e^{\theta(s,a)}}{\sum_b e^{\theta(s,b)}}$$

Policy Update Rule

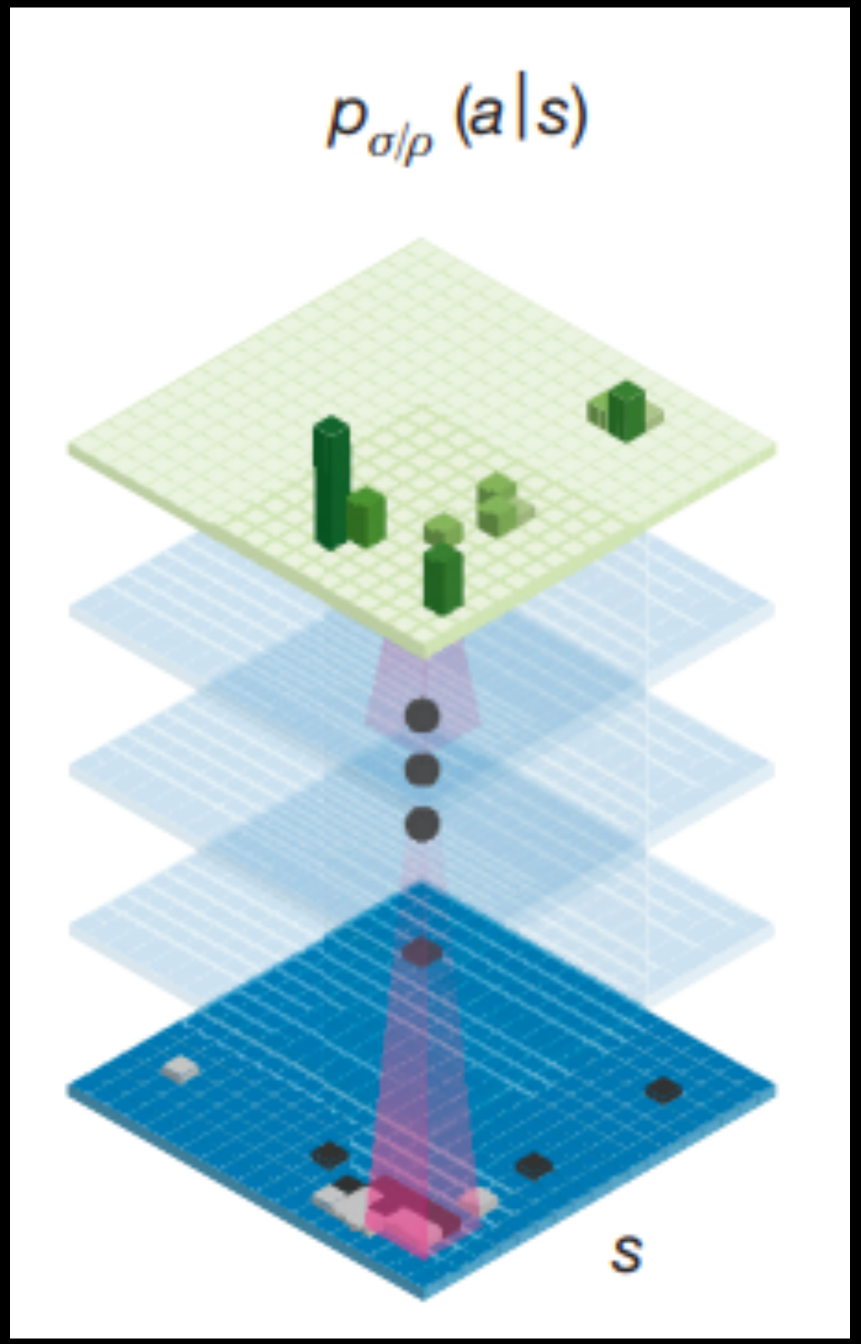
$$\theta(s_t, a_t) \leftarrow \theta(s_t, a_t) + \epsilon\delta_t$$

Actor-Critic in the Brain

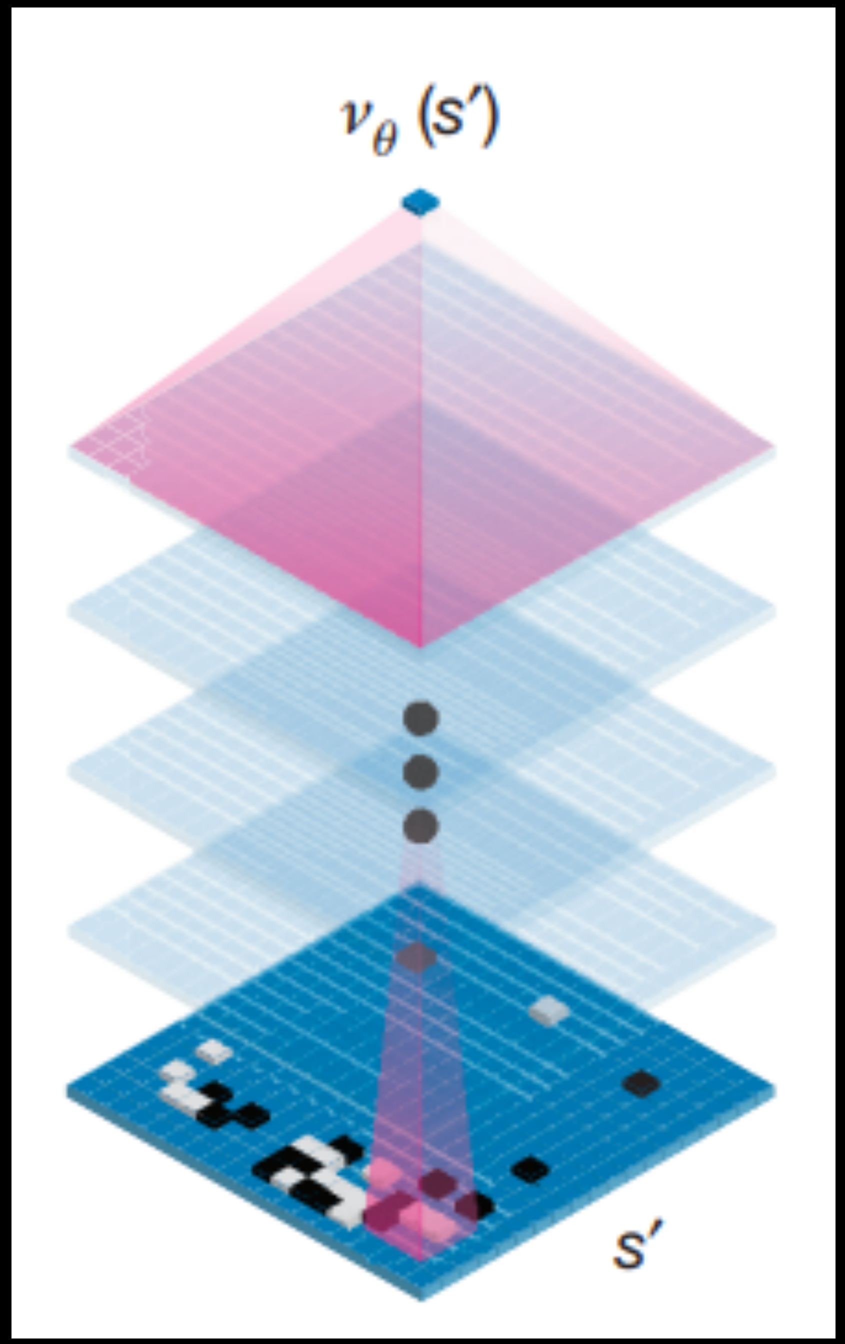


Deep Reinforcement Learning

Policy Network



Value Network

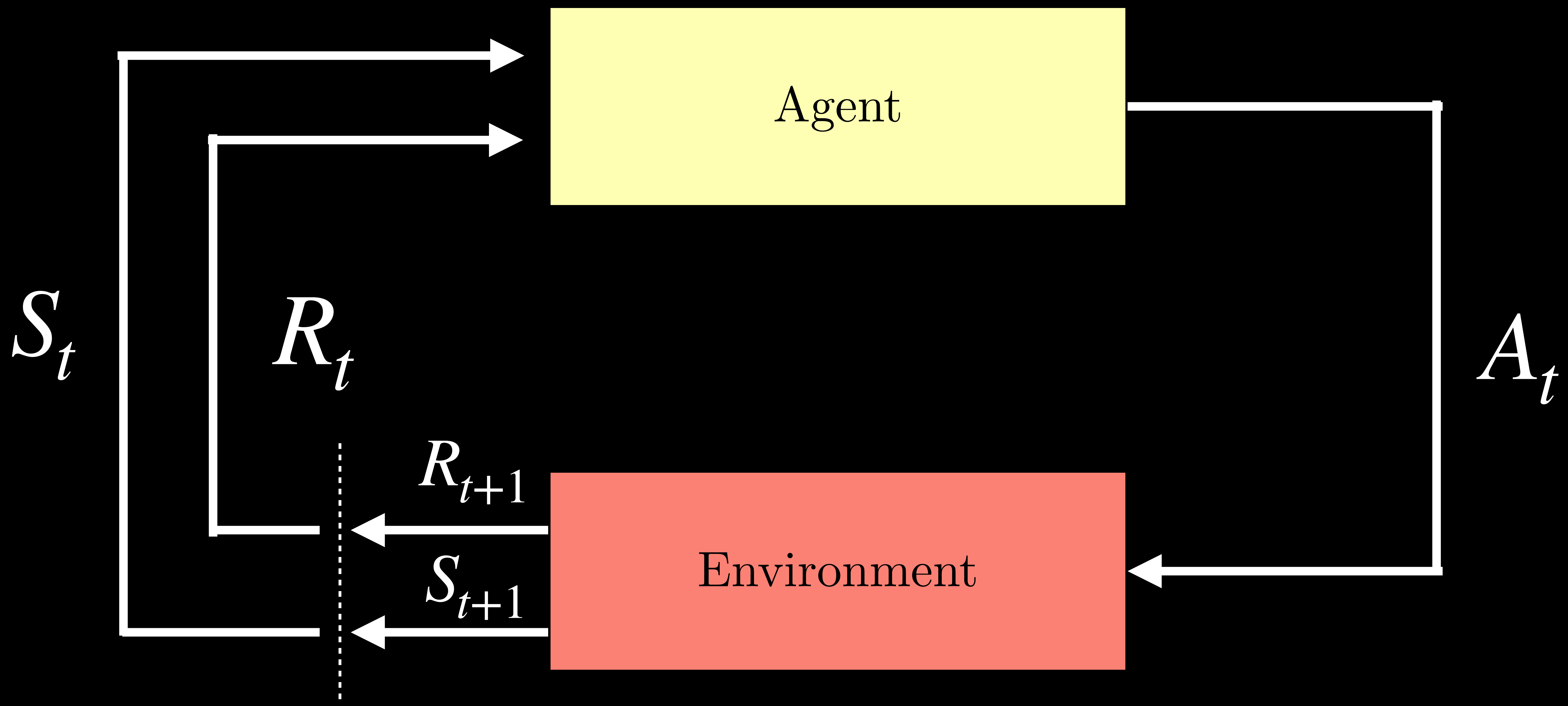


Model-Based Reinforcement Learning

S_t : State

R_t : Reward

A_t : Actions



Model-Based Reinforcement Learning

S_t : State

R_t : Reward

A_t : Actions