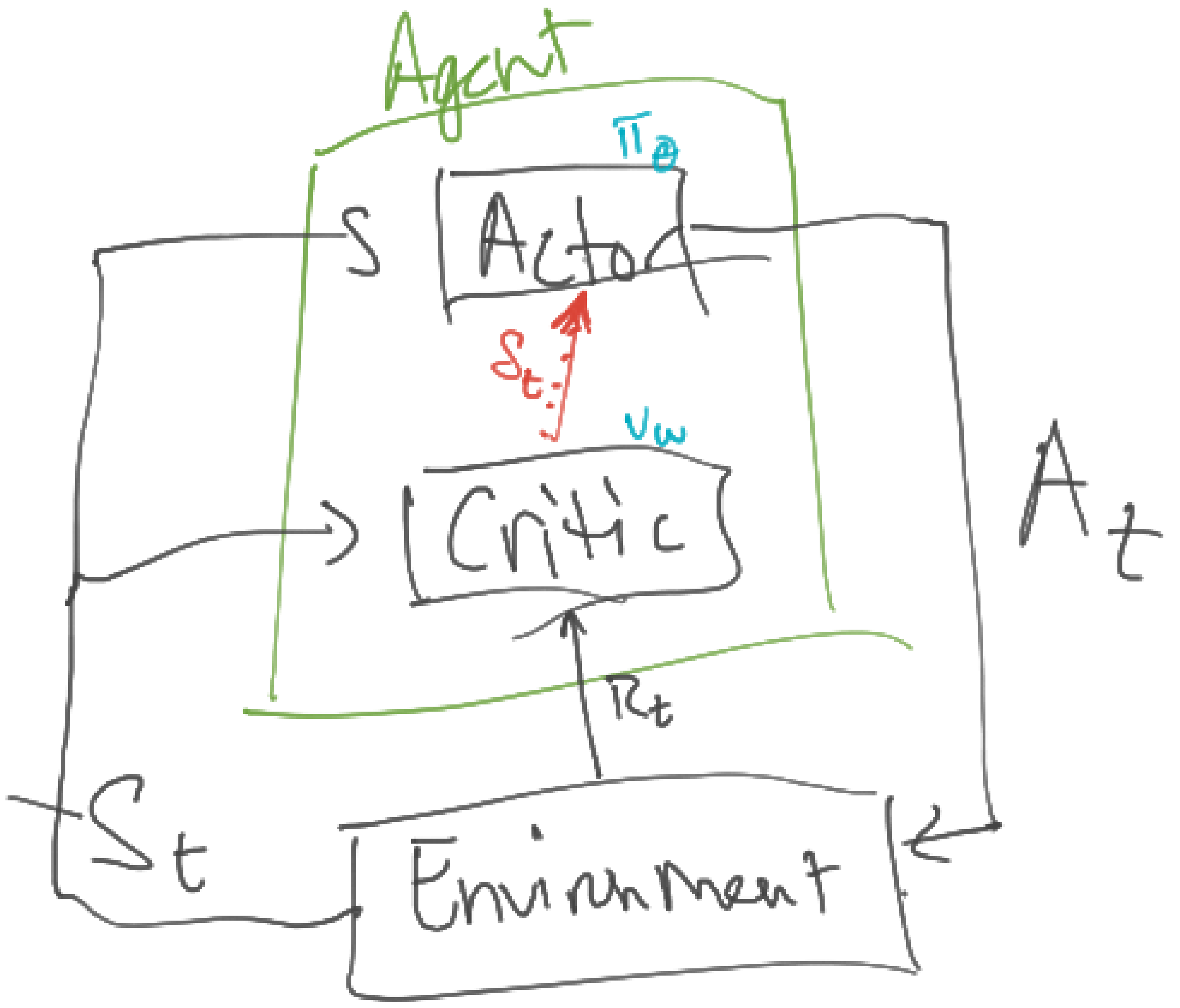
Actor - Critic

$$v^\pi(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| S_t = s ; \pi\right]$$

```
Hyperparams:  Initial policy params $\theta$
              Policy representation (ANN? Linear?)
              Actor step size $\alpha$
              Critic step size $\beta$.
              Value function representation, $v_w$.
              Initial value function weights $w$.

For each episode:
    For each time $t$
        Agent observes $S_t$
        Agent selects action $A_t$ using $\pi_\theta$
        Env responds with $S_{t+1}$ and $R_t$
        $\delta_t = R_t + \gamma v_w(S_{t+1}) - v_w(S_t)$   // TD-error
        $\forall i,\ \theta_i \leftarrow \theta_i + \alpha \boxed{\gamma^t} \delta_t \frac{\partial \ln(\pi_\theta(S_t, A_t))}{\partial \theta_i}$   // Actor update
        $\forall j,\ w_j \leftarrow w_j + \beta \delta_t \frac{\partial v_w(S_t)}{\partial w_j}$   // critic update.
```

Theory says to include
In practice it is bad. Almost nobody includes this term.



Psychology

Sutton & Barto
2nd edition.

RLDM

Operant conditioning: learning process through which the strength of a behavior is modified by reward or punishment.

(control)

(searching for a better policy)

- learning a policy.

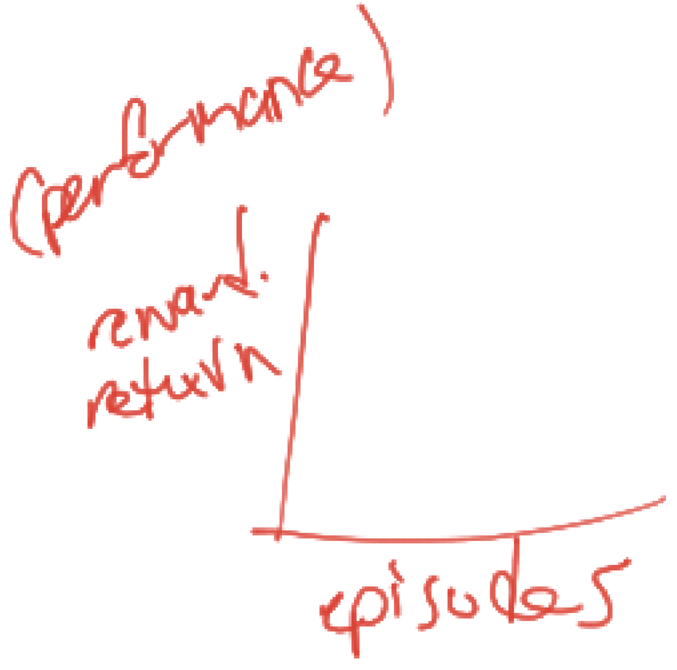
Classical conditioning: learning procedure in which a biologically potent stimulus (e.g., food) is paired with a previously neutral stimulus (e.g., bell).

(prediction)

- learning a value function.

- Thorndike's Puzzle Boxes.

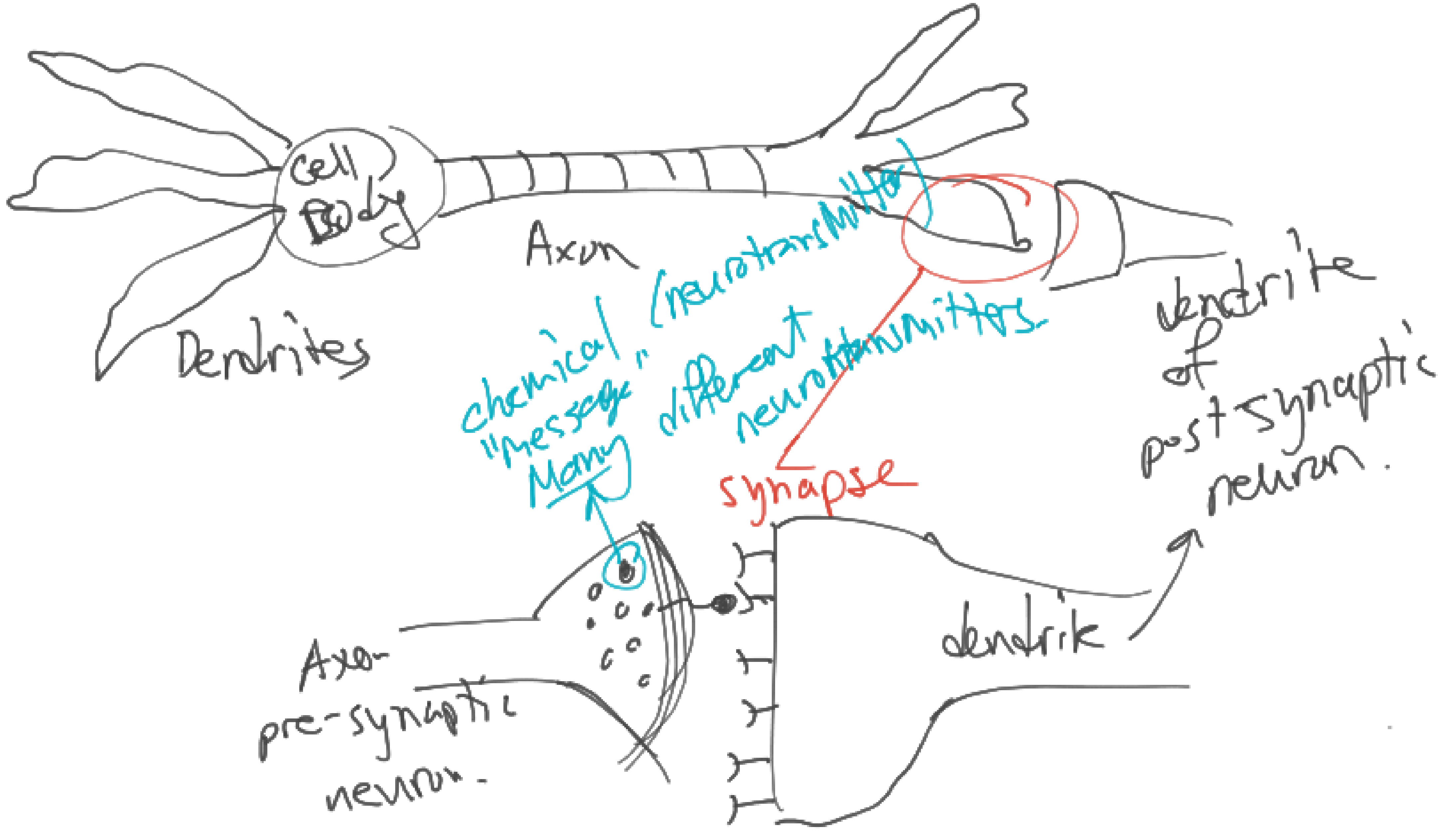
1898



Neuroscience

Dopamine

- Contraction of 3,4 - dihydroxyphenethylamine
- Chemical
- Neurotransmitters.


Cell Body
Axon
Dendrites
chemical "message" (neurotransmitter)
Many different neurotransmitters.
synapse
dendrite of post synaptic neuron.
Axon
pre-synaptic neuron.
dendrite

Dopamine neuron
(Dopaminergic neuron)
is a neuron that emits the neurotransmitter dopamine.

Two clusters in mammals:

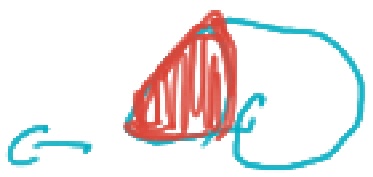
SNpc and VTA

SNpc ↓ substantia nigra pars compacta

VTA ↓ ventral tegmental area.

$\text{SNpc} \xrightarrow{\text{dopamine}} \text{Striatum}$

↳ coordinates motor & action planning, decision making, motivation.



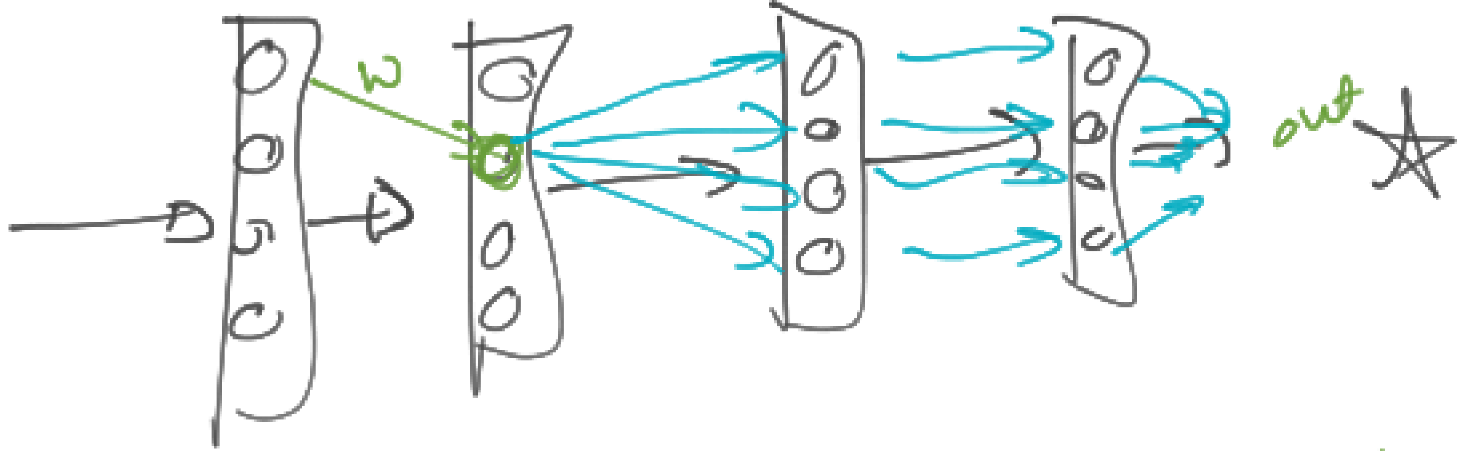
$\text{VTA} \xrightarrow{\text{dopamine}}$ numerous areas including prefrontal cortex → planning, personality, decision making

Reward Prediction Error, Hypothesis for dopamine.

TD error.

Olds & Milner 1954 : Dopamine $\propto$ reward.

Brains probabably do not implement backpropagation.



$\frac{\partial out}{\partial w}$

Δ_j terms propagated backwards through the network.

→ Information does not seem to pass backwards down the axon.

Each "neuron" can updatue using only its input, output, and δ_t.

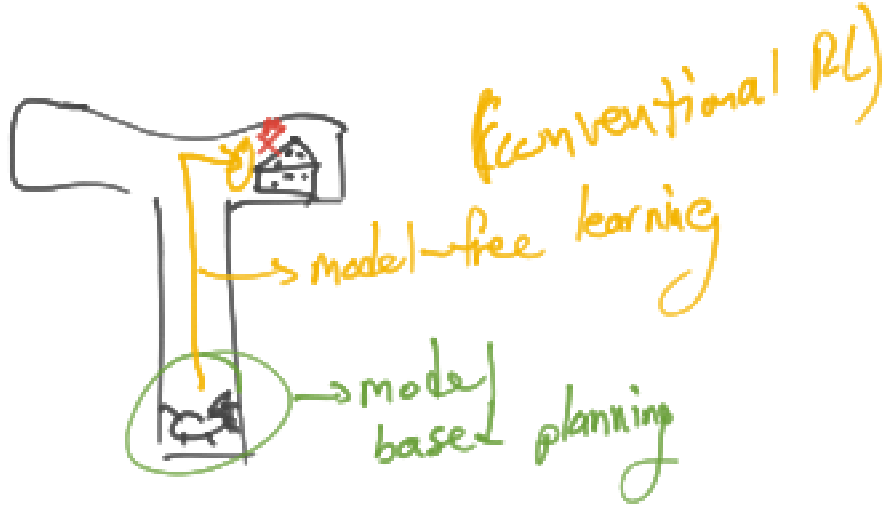
Duplicate of network in reverse.

Coagent Networks

↳ extention of "learning automata"

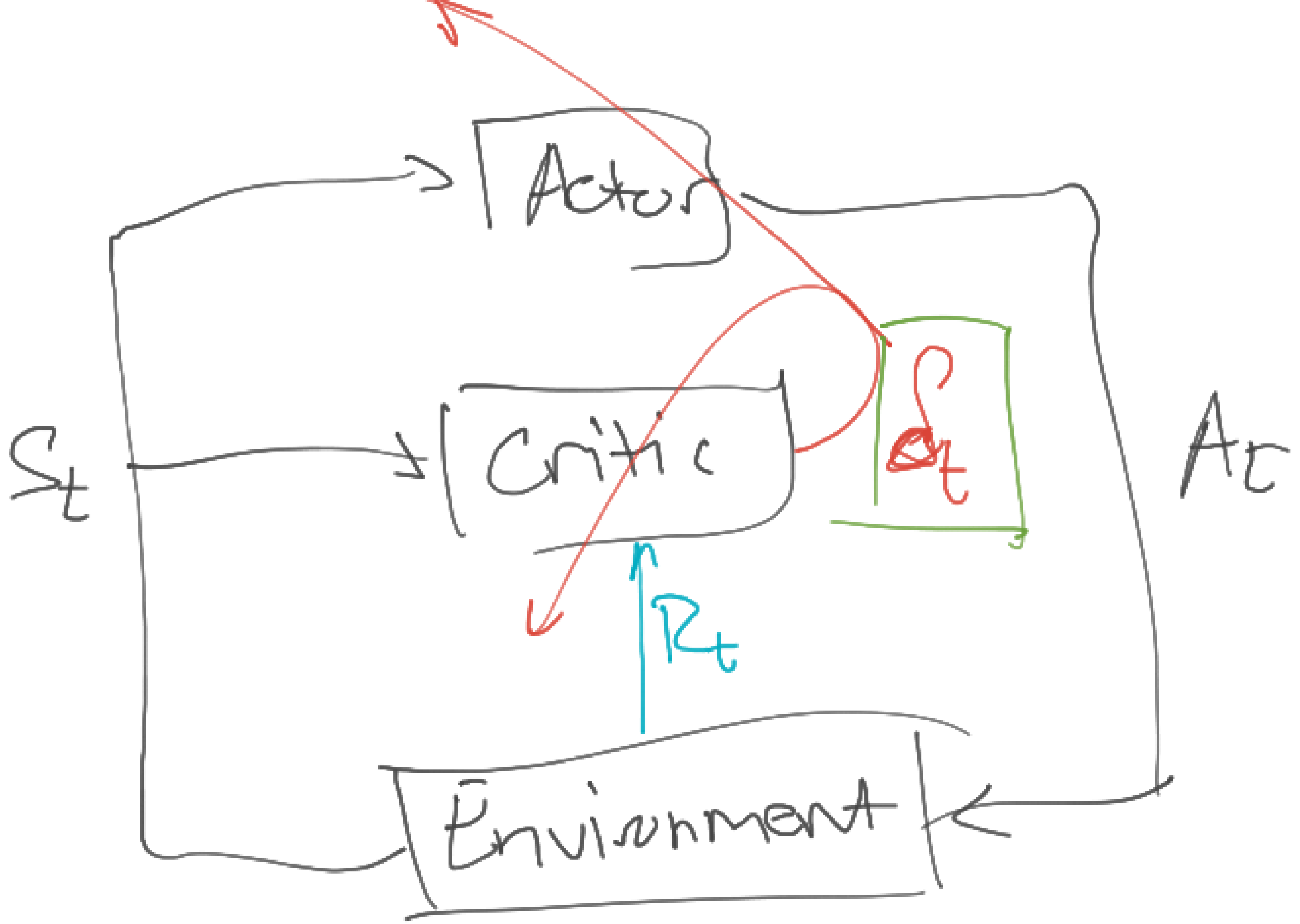
↳ training RL networks without backprop.

Reward devaluation Studies.



→ early learning is model-based planning

→ transitions to a model-free policy over time.



Treat sugar

V(cake in front)

v^{π} heroin (drugs) =

Addiction

Octopi