Reinforcement Learning Models of the Basal Ganglia

Computational Models of Neural Systems

Lecture 6.2

David S. Touretzky

November, 2017

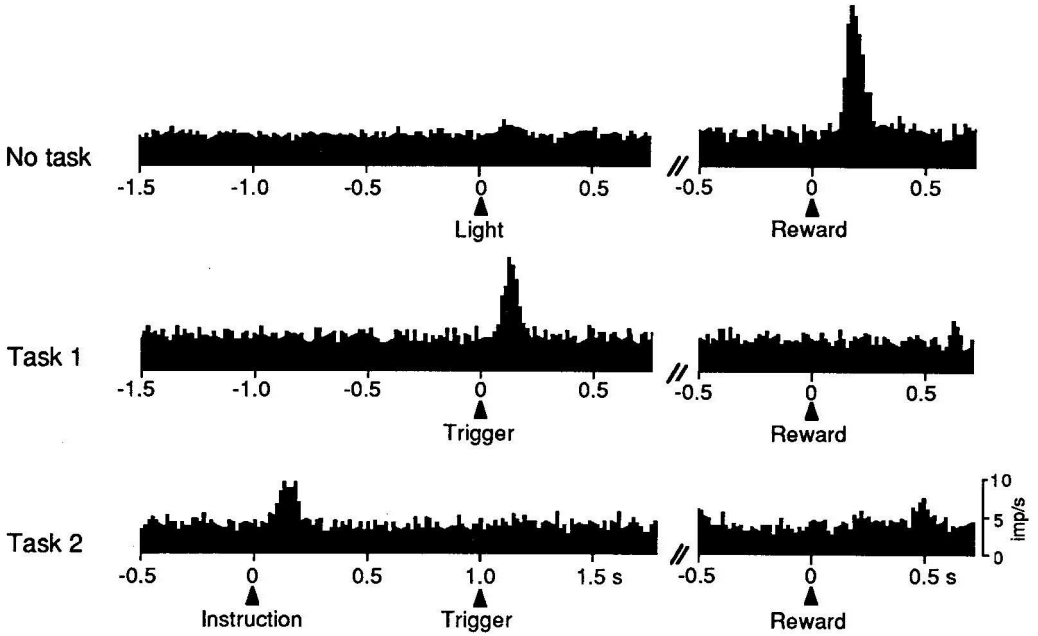
Dopamine Cells

- Located in SNc (substantia nigra pars compacta) and VTA (ventral tegmental area)
- Project to dorsal and ventral striatum, and also to various parts of cortex, especially frontal cortex.
- Respond (50-120 msec latency) with a short (< 200 msec) burst of spikes to:
  - Unpredicted primary reinforcer (food, juice)
  - Unpredicted CS (tone, light) that has become a secondary reinforcer
    - Reduced by overtraining; perhaps because environment now predicts
  - High intensity or novel stimuli
    - Response diminishes with repetition (loss of novelty)
  - For a few cells (less than 20%): aversive stimuli

What Do DA Cells Encode?

- Current theory says: reward prediction error.
  - Nicely explains response to unpredicted reinforcers
  - Novelty is somewhat rewarding to animals
  - Aversive stimuli? (prediction error)

- Teaching signal for striatum to learn to predict better.

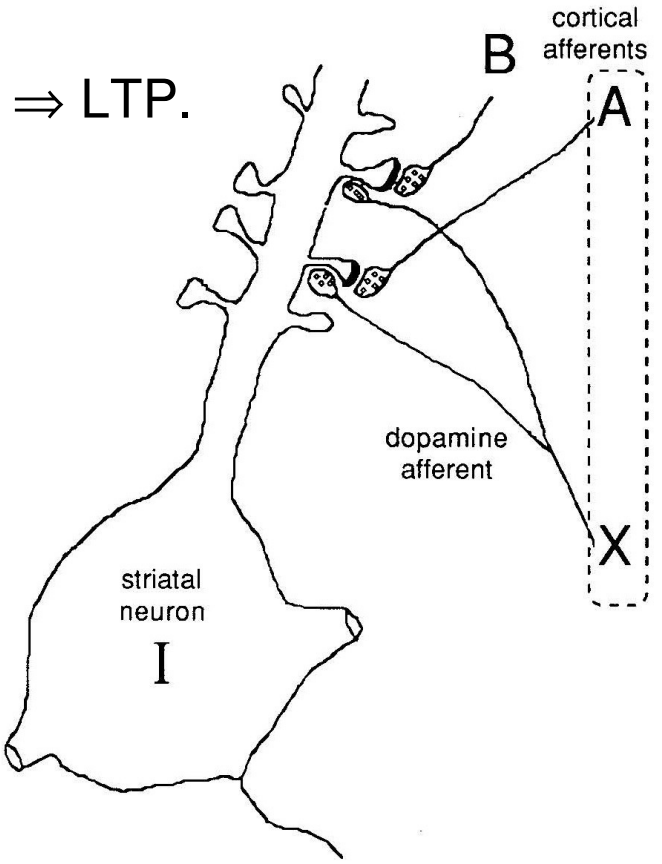


Specificity of Reward

- Schultz found all DA cells showed similar responses.
- But anatomy tells us that DA cells receive projections from different areas (cf. 5 or 21 parallel circuits in basal ganglia), so they should have different responses.
  - Maybe the problem is that his animals were only tested on a single task.
  - More recent experiments have shown that DA neurons can distinguish between more and less preferred rewards.

Dopamine Synapses

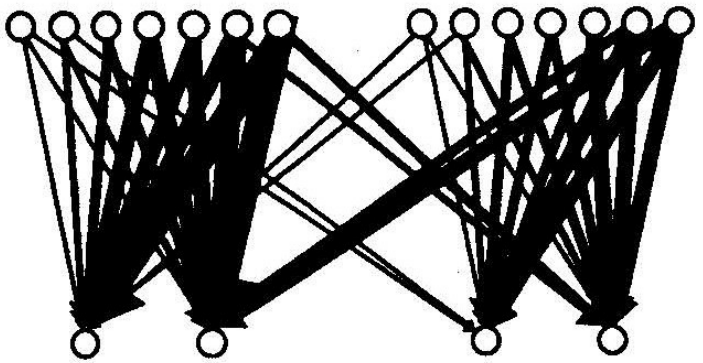
- Dopamine cells project to striatal spiny cells.
- Dopamine cells contact the spine neck; cortical afferents contact the spine head.
- Heterosynaptic learning rule?
  - Afferent input + subsequent dopamine input ⇒ LTP.
- Medium spiny cell:
  - 500-5,000 DA synapses
  - 5,000-10,000 cortical synapses



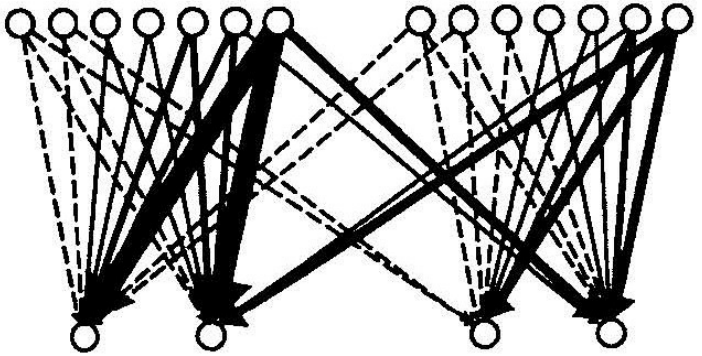
Effects of Dopamine

- Focusing: dopamine reduces postsynaptic excitability, which focuses attention on the striatal cells with strongest inputs.
- Dopamine probably causes LTP of the corticostriatal path, but only for connections that were recently active.
- Since dopamine release does *not* occur in response to predicted rewards, it cannot be involved in maintenance of learning.
  - What prevents extinction?
  - Perhaps a separate reinforcer signal in striatum.

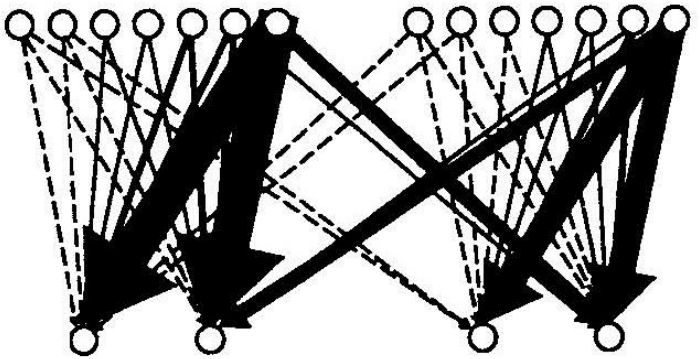
No dopamine activity



Dopamine-induced focussing



Dopamine-induced long term facilitation



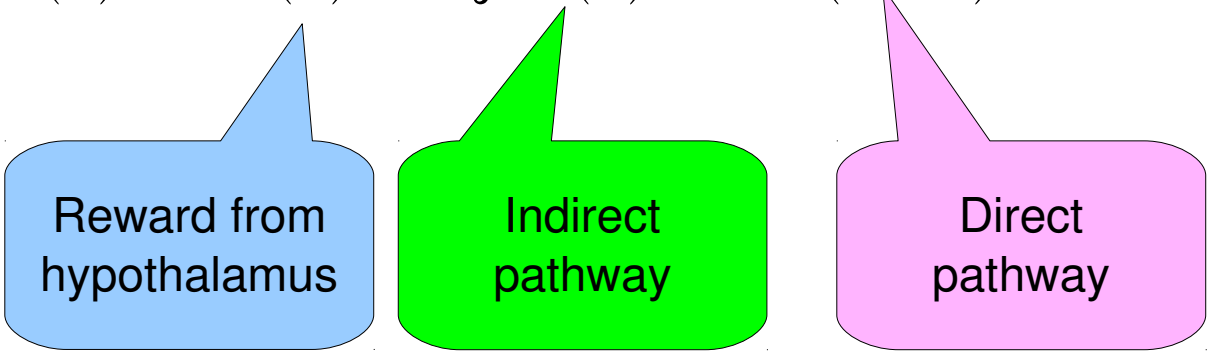
TD Learning Rule

- Goal: predict future reward as a function of current input $x_i(t)$.

$$V(t) = \sum_i w_i x_i(t)$$

- Reward prediction error $\delta(t)$:

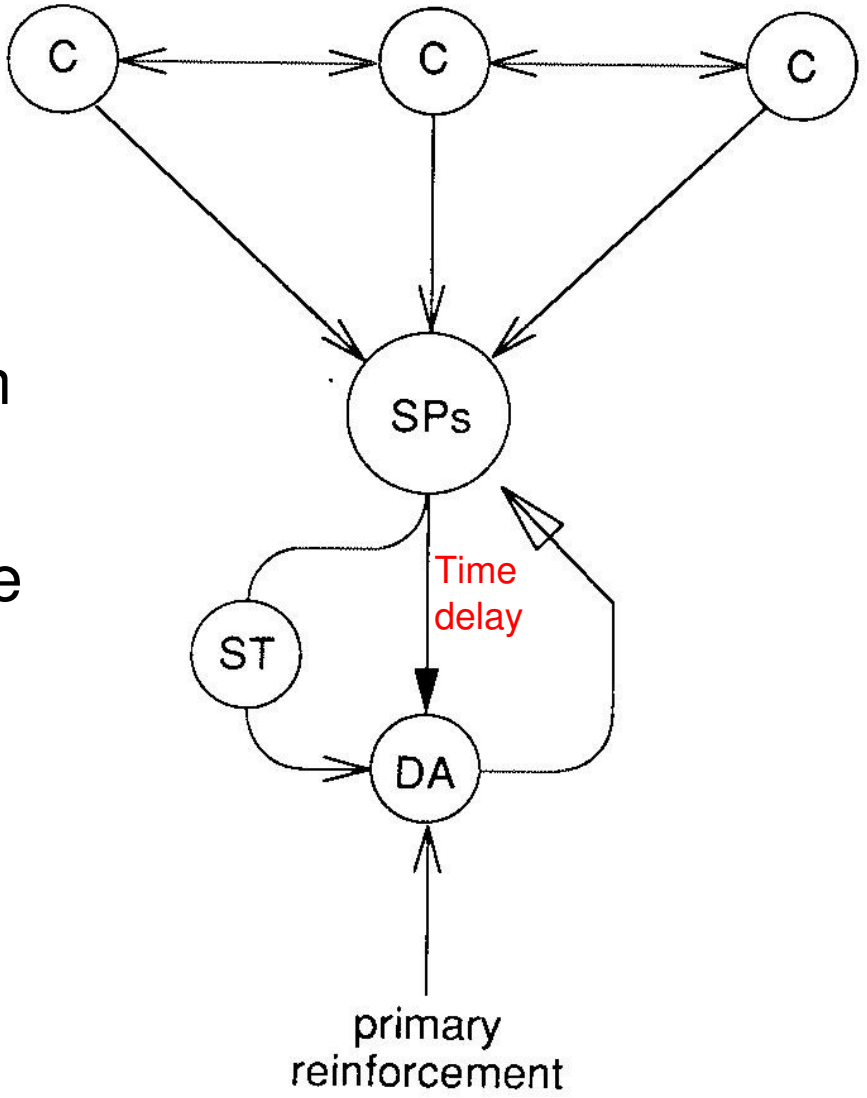
$$\delta(t) = r(t) + \gamma V(t) - V(t-1)$$

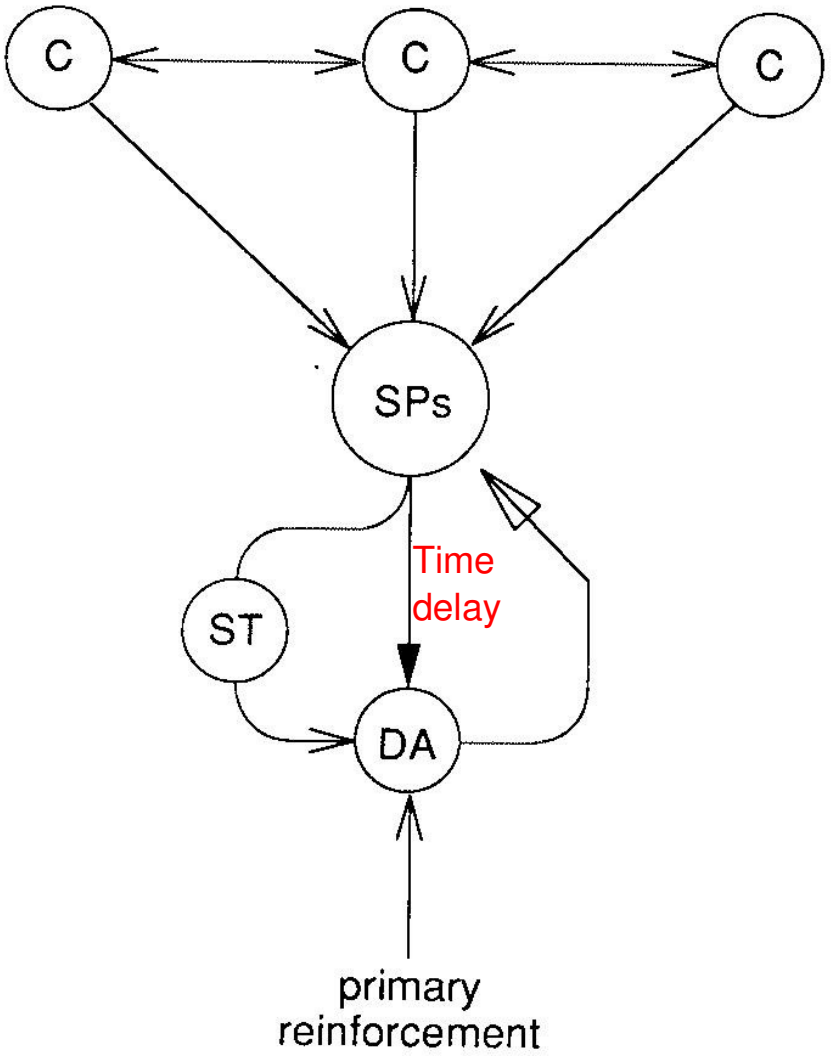


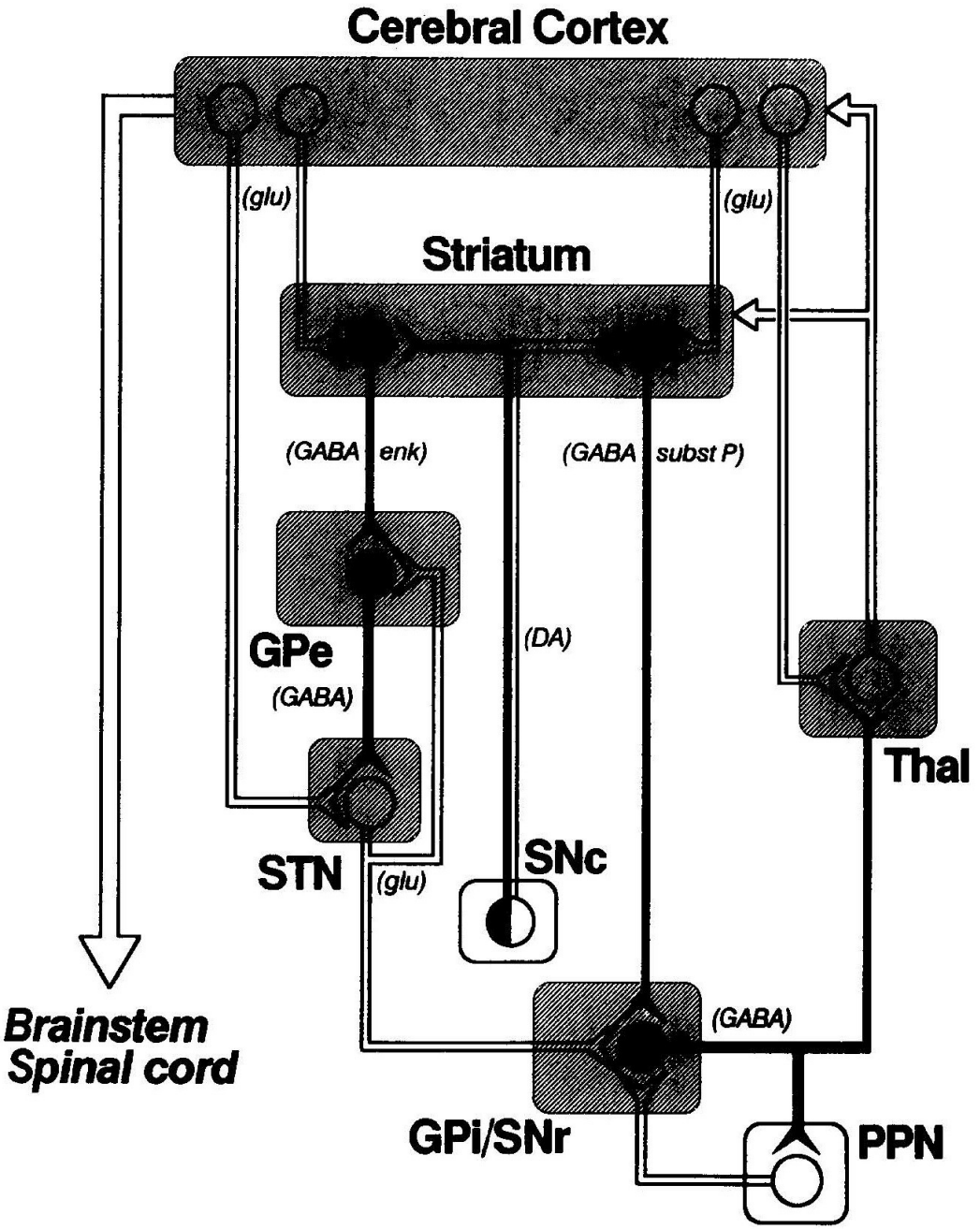
- Simplifying assumption: no discounting (γ equals 1).

Simple TD Learning Model

- Barto, Adams, and Houk proposed a TD learning theory based on a simplified anatomical model.
- Striosomal spiny cells (SPs) learn to predict reinforcement.
- Dopamine cells (DA) generate the error signal.
- ST = subthalamic nucleus

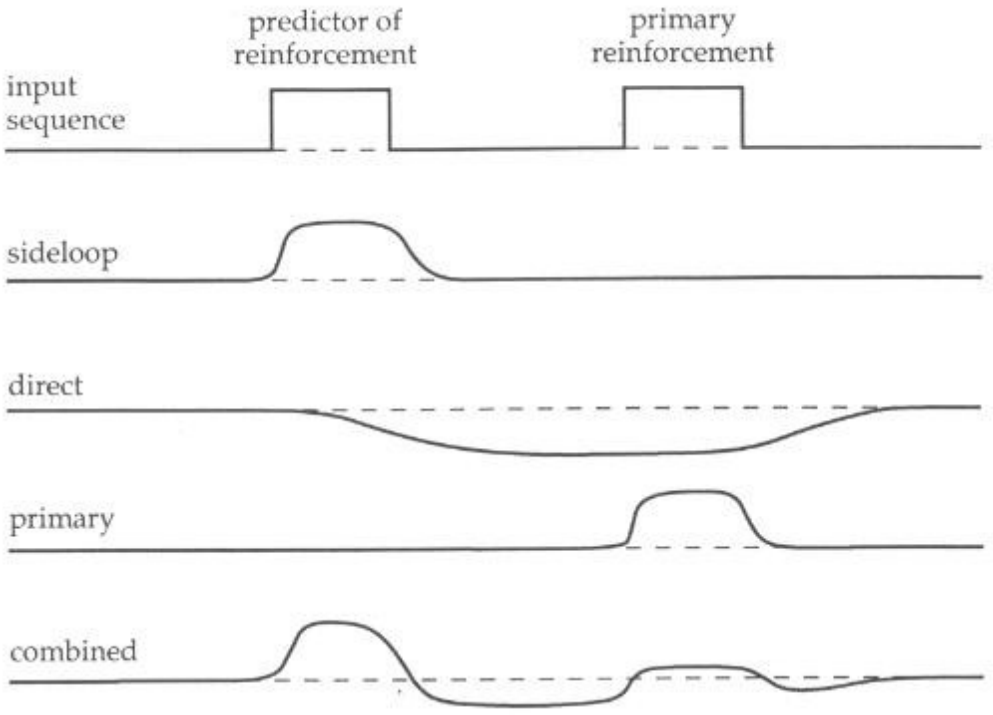



C
C
C
SPs
Time
delay
ST
DA
primary
reinforcement


Cerebral Cortex
(glu)
(glu)
Striatum
(GABA enk)
(GABA subst P)
GPe
(GABA)
(DA)
Thal
STN
(glu)
SNc
Brainstem
Spinal cord
(GABA)
GPi/SNr
PPN

Response to Reinforcers

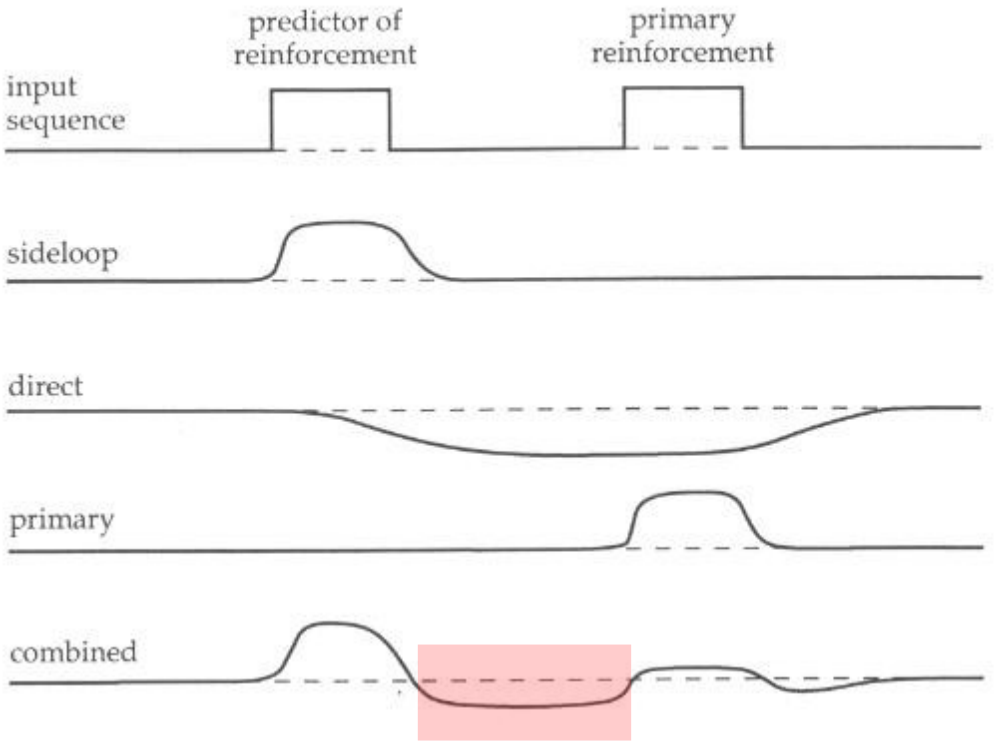
- Indirect path is fast: striatum to GPe to STN excites dopamine cells in SNc/VTA.
- Direct path must be slow and long lasting. $GABA_A$ inhibition only lasts 25 msec. Perhaps $GABA_B$ inhibition is used, but not conclusively demonstrated.



What's Wrong With This Model?

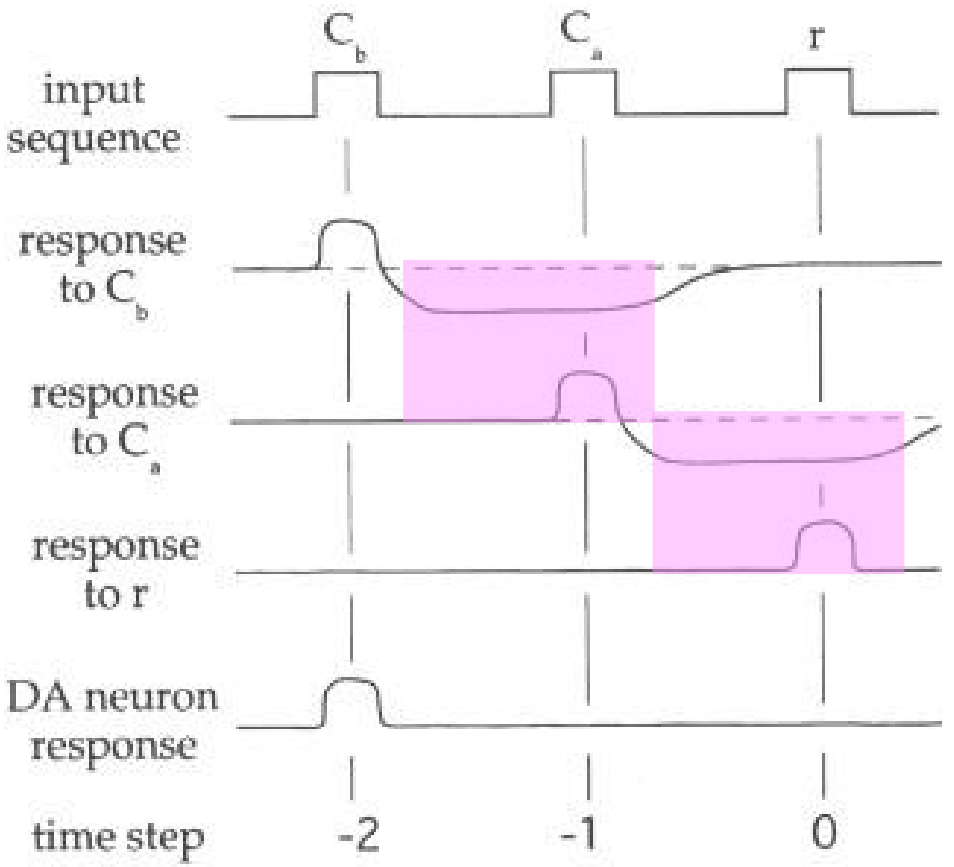
- Even $GABA_B$ inhibition may be too short lasting.

- The model predicts a decrease of dopamine activity preceding primary reward.



Responses to Earlier Predictors

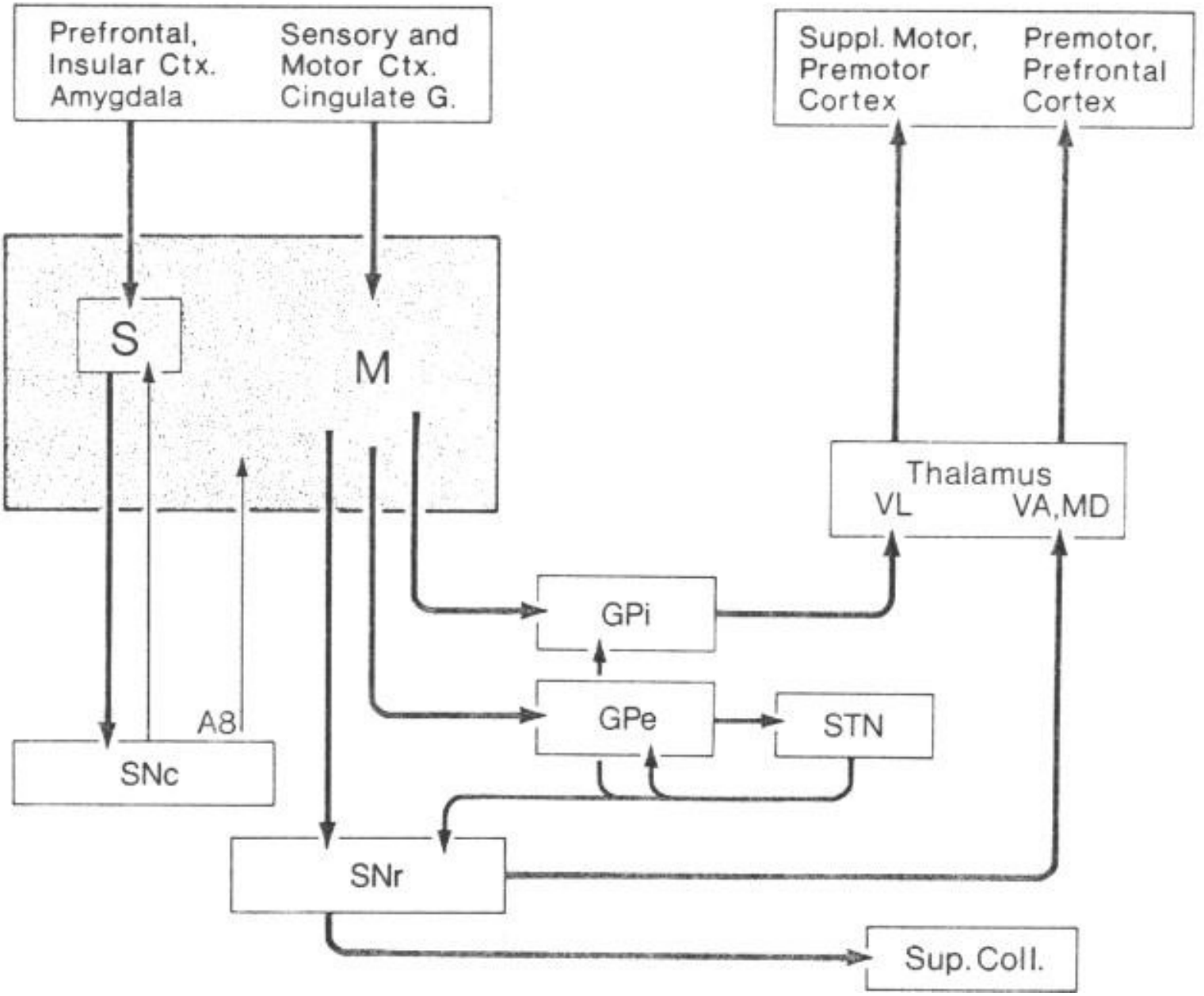
- Highly simplified model using fixed time steps.
- Timing is assumed to be just right for slow inhibition to cancel fast excitation: unrealistic.



Problem: Lack of Timing Information

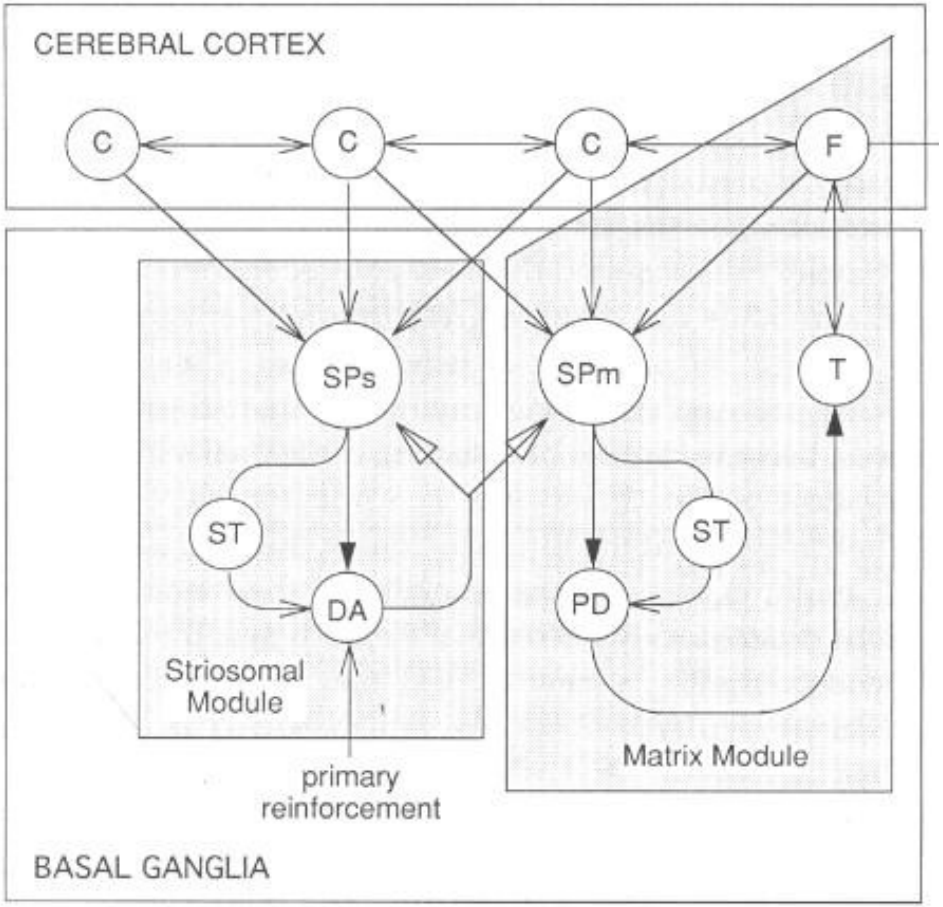
- The problem with this model is that a single striosomal cell is being asked to:
  - respond to a secondary reinforcer stimulus (indirect path), *and also*
  - predict the timing of the primary reward to follow (direct path)
- Need a more sophisticated TD model.
- If we use a serial compound stimulus representation, then the predicted timing of future rewards can be decoupled from response to the current stimulus.
- But this requires a major assumption about the striatum: it would have to function as a working memory in order to predict rewards based on stimulus history.

Review of Anatomy: Striosome vs. Matrix



Striatum As Actor/Critic System (Speculative)

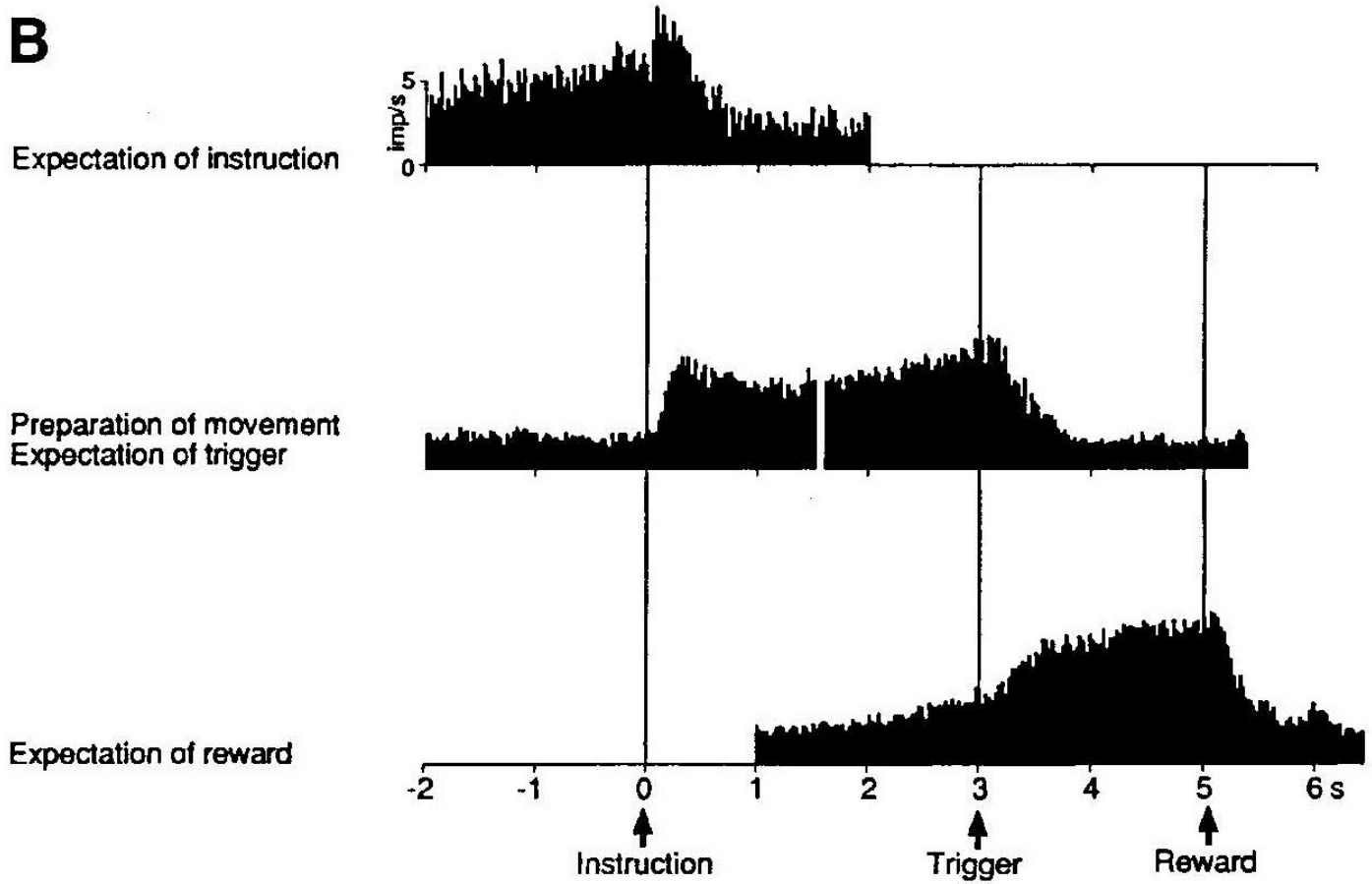
- Striosomal modules (critic) predict reward of selected action.
- Matrix modules (actor) select actions.
- Dopamine error signal trains critic to predict reward and matrix to select best action.



PD = pallidum

Striatal Representations

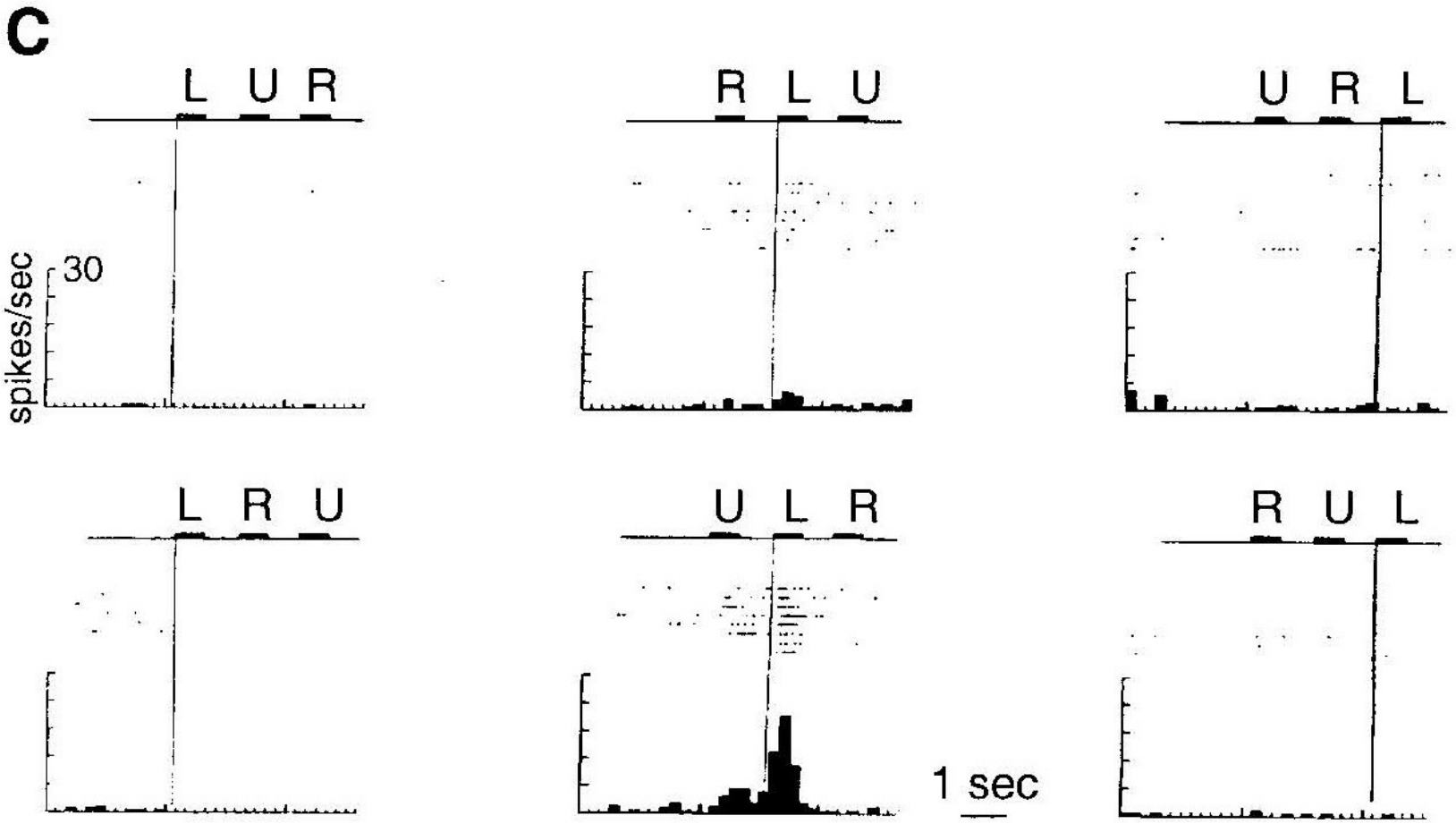
Expectation- and preparation-related striatal neurons:



Striatal Representations

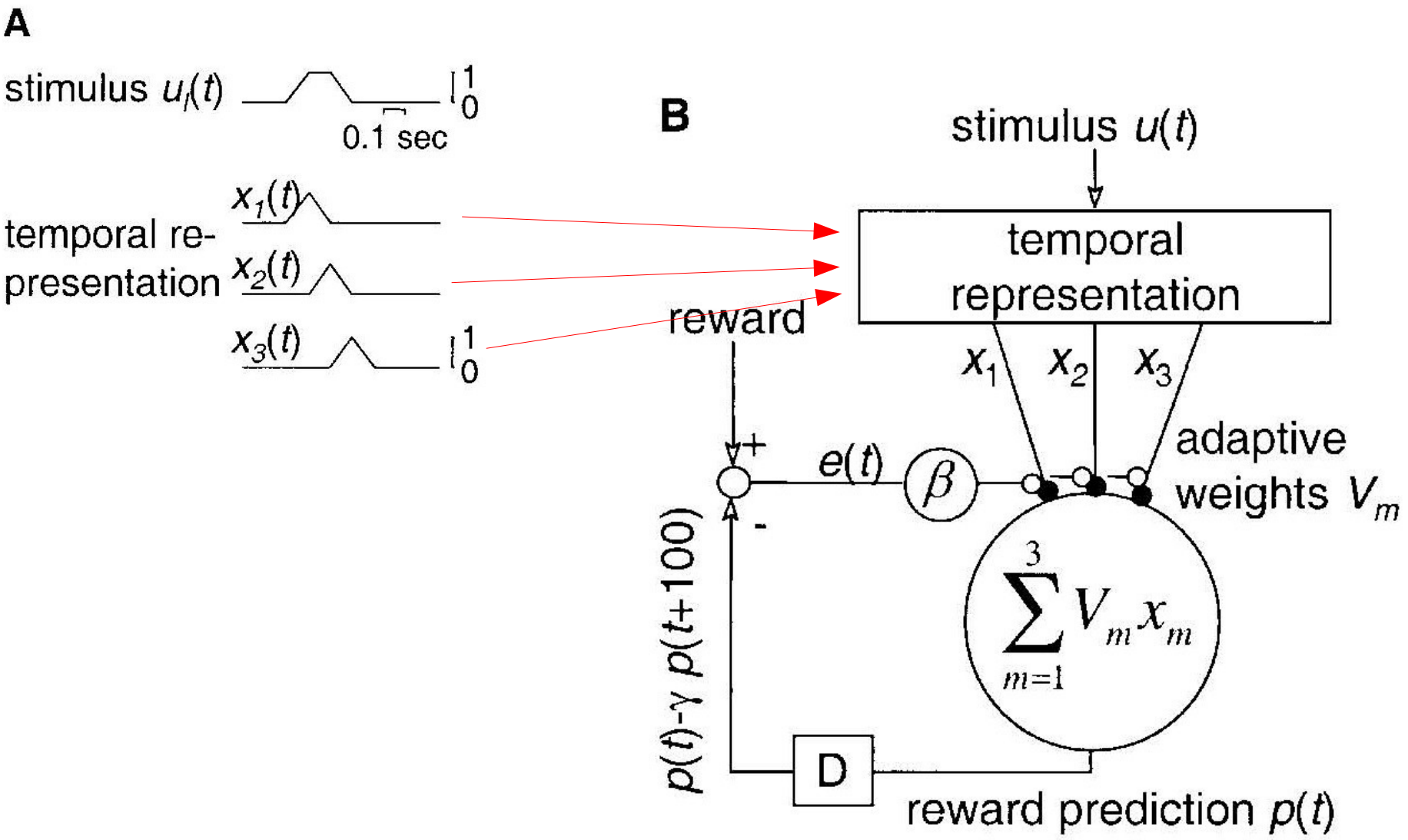
- Caudate neuron that responds to stimulus L only within the sequence U-L-R. Apicella found 35 of 125 caudate neurons responded to a specific target modulated by rank in sequence or co-occurrence with other targets.

Visual targets / levers: L=left, R=right, U=upper.

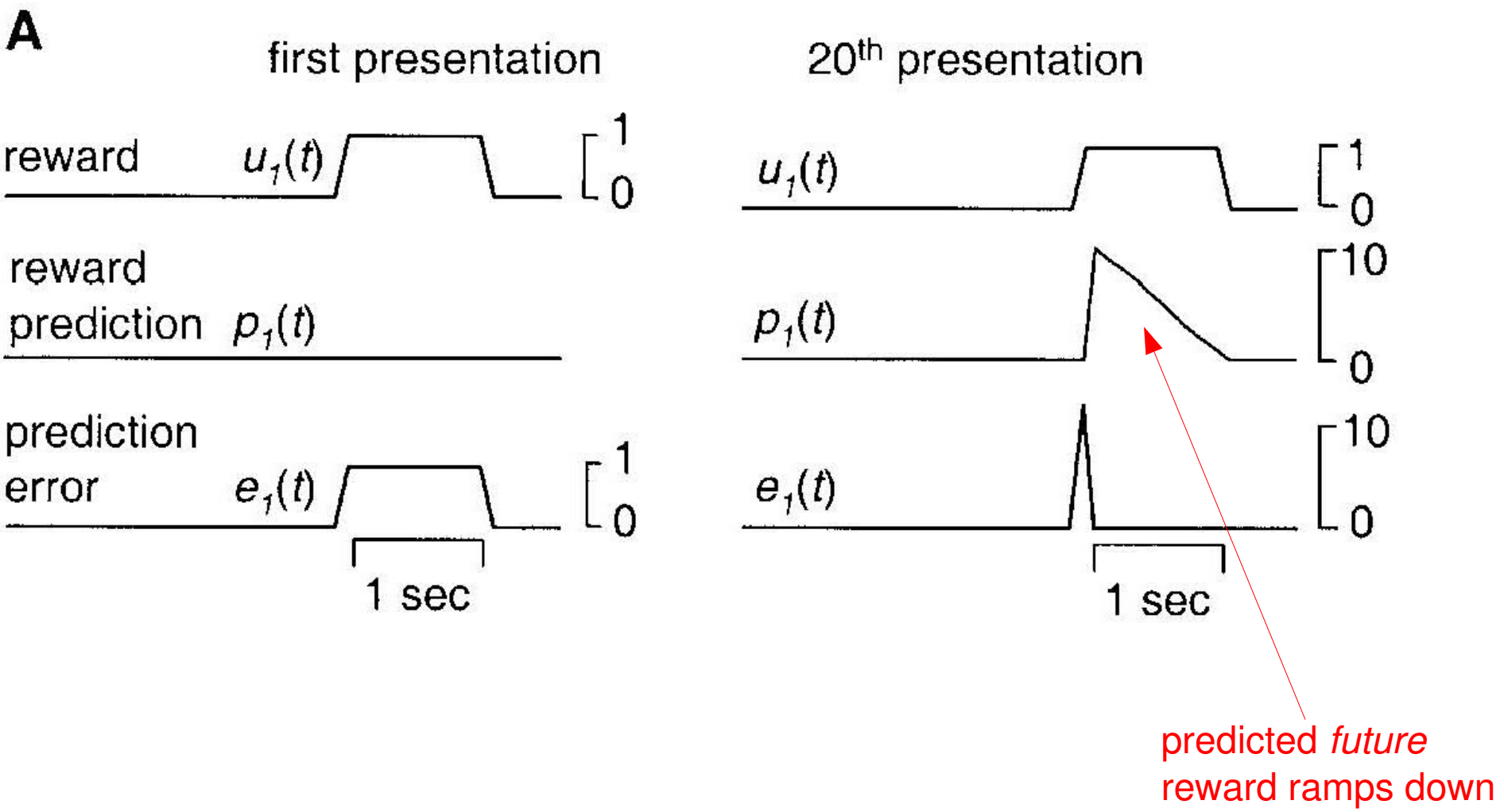


Suri & Schultz TD Model

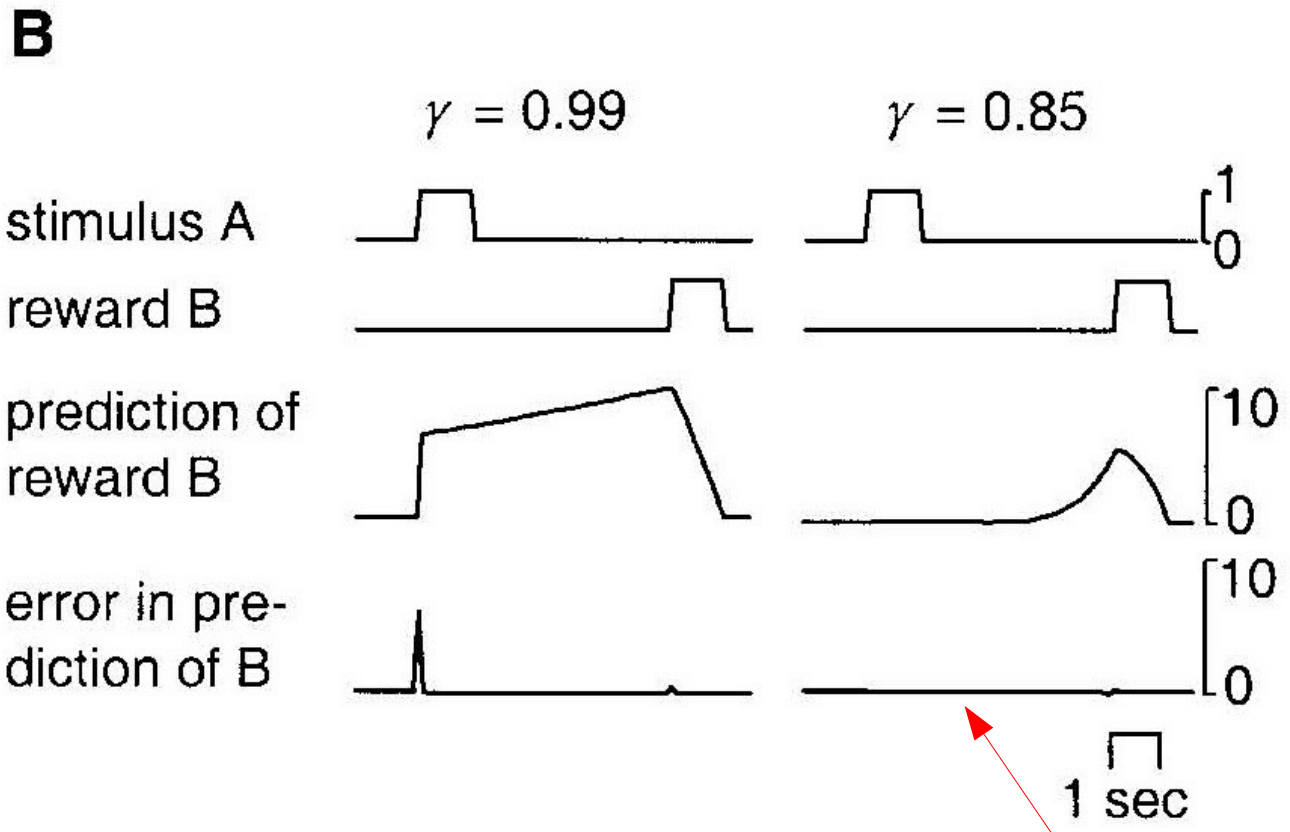
Complete serial compound representation can learn timing.



TD Reward Prediction

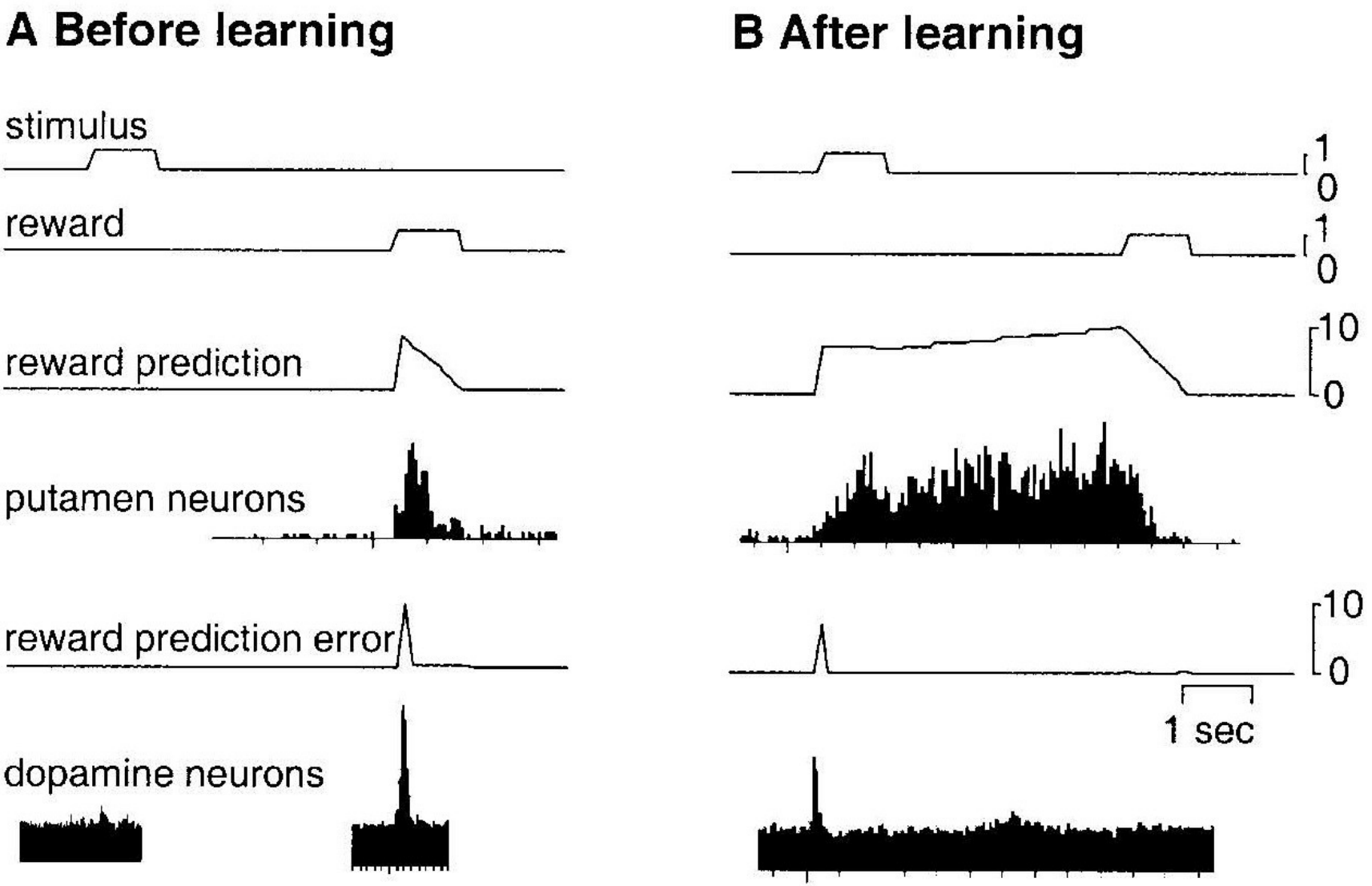


Discounting Rate Shapes the Reward Prediction

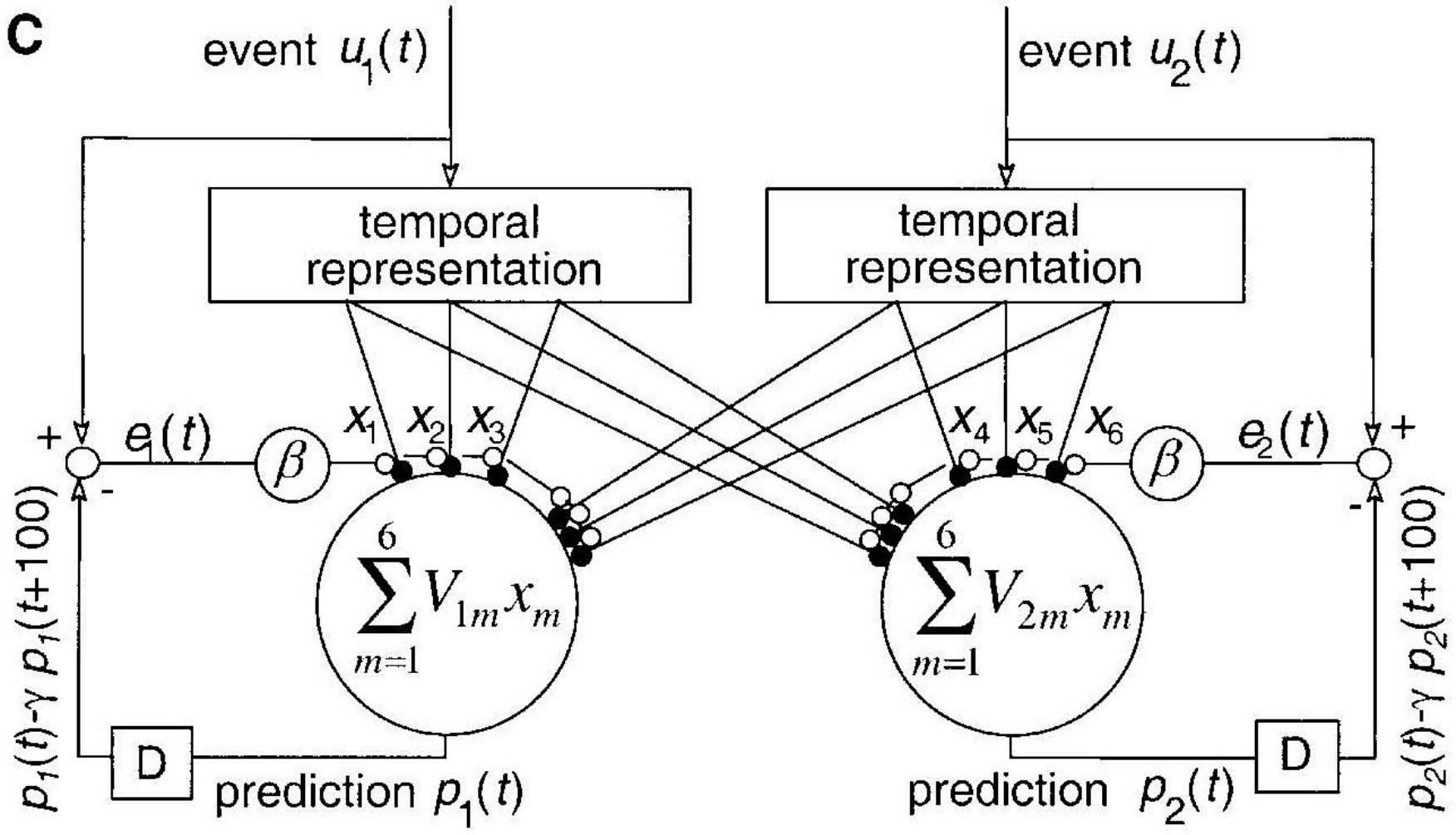


Error near zero everywhere because reward fully discounted and prediction ramps up slowly.

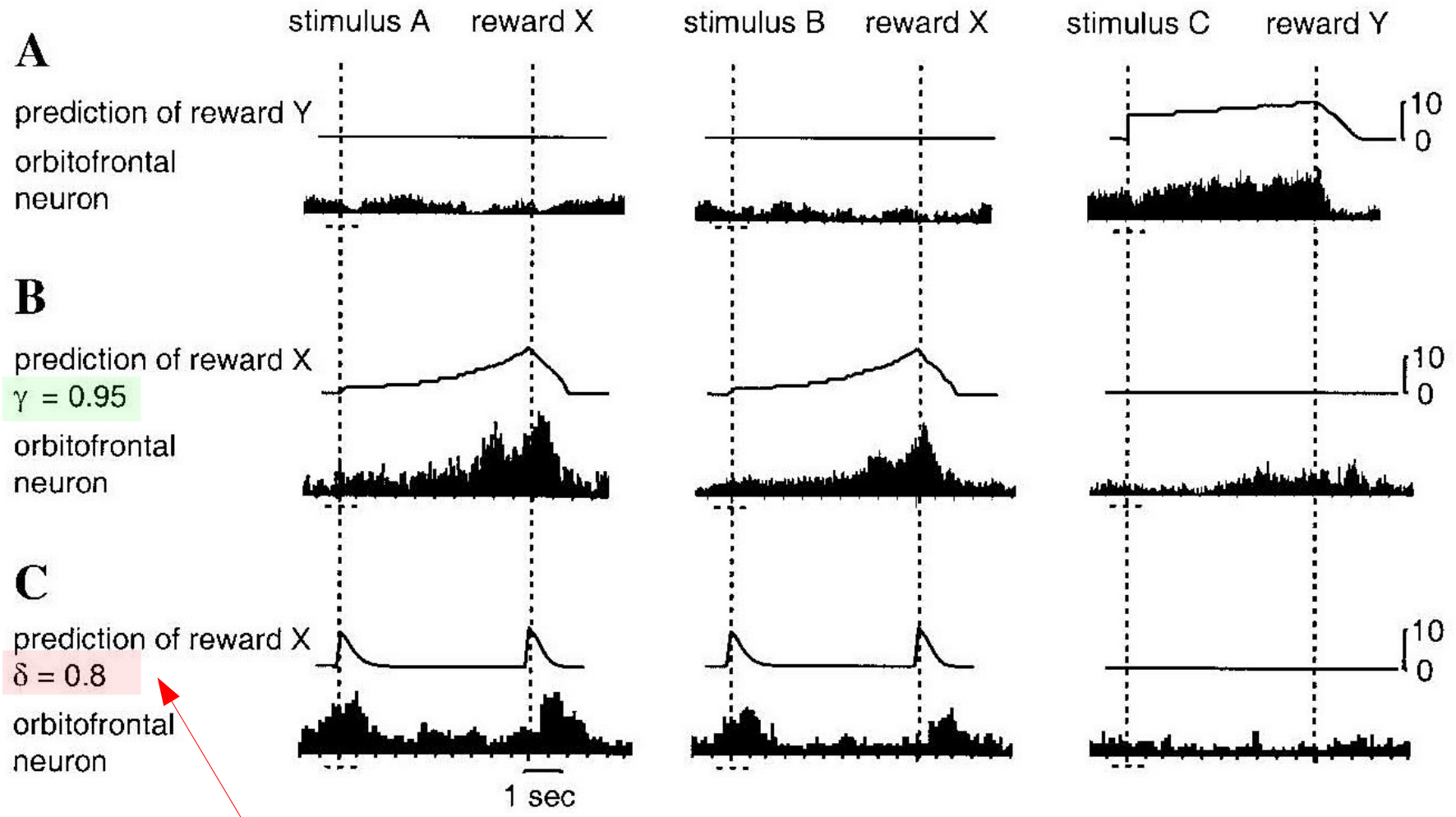
Effects of Learning



Separate Model For Each Reward Type



Varying Model Parameters Allows Reward Prediction to fit Orbitofrontal Cortex Data

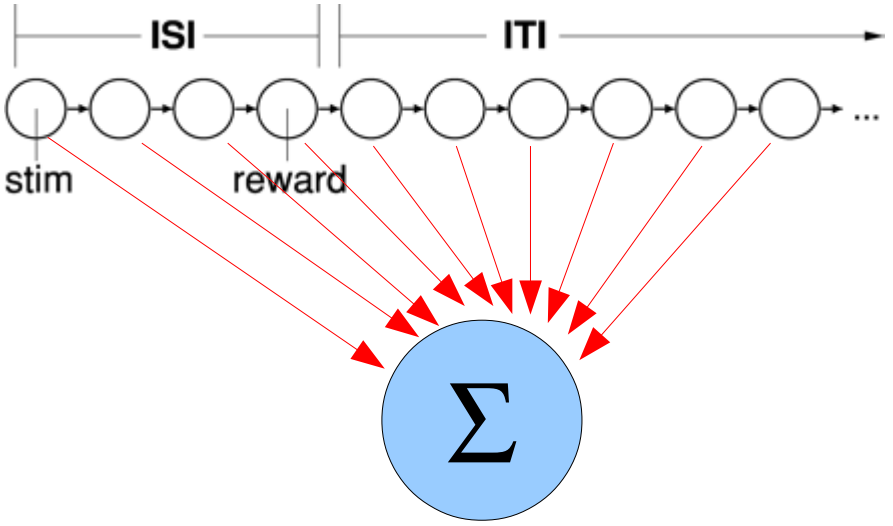


representation decay, but long eligibility trace

Reward X and reward Y are two different liquids.

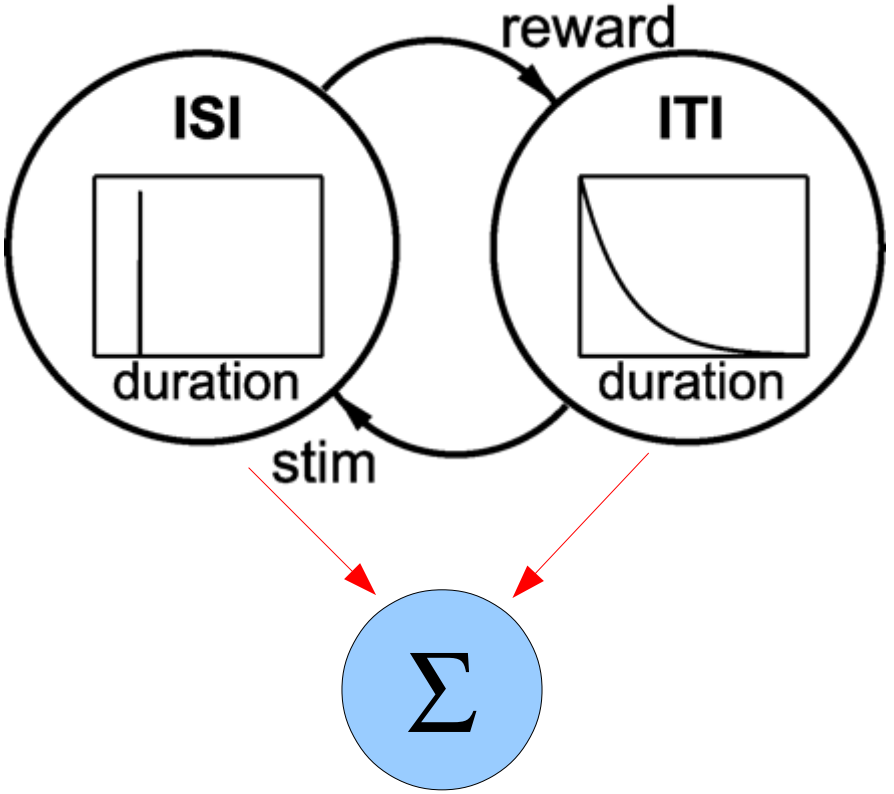
Problems With the Suri & Schultz TD Model

- Correctly predicts pause after omitted reward, but incorrectly predicts pause after early reward.
- Can't handle experiments with variable inter-stimulus intervals: predicts same small negative error at each time step where reward could occur and same large positive response where it does occur.
- The source of these problems is that the complete-serial-compound (delay line) representation is too simplistic.

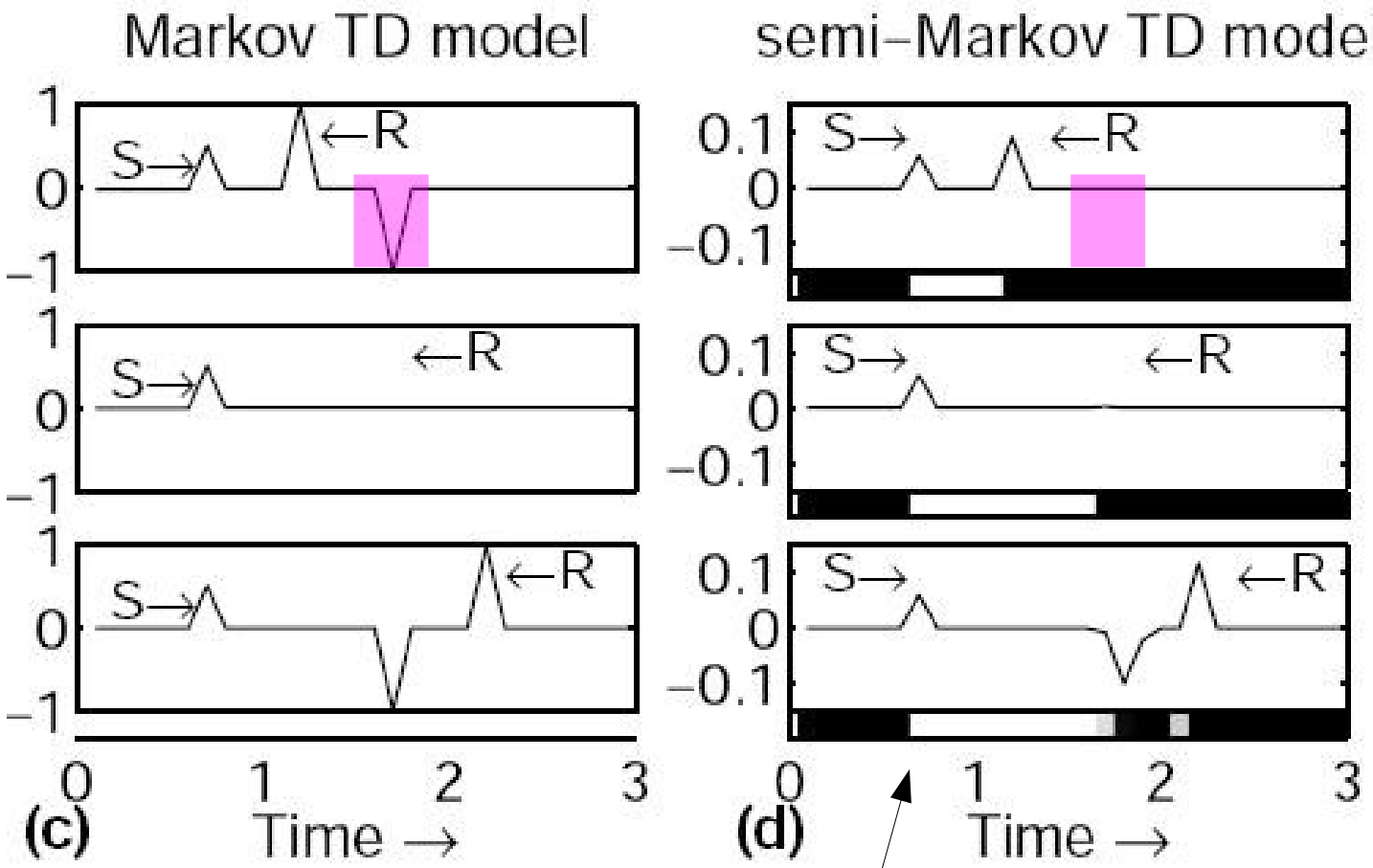


Daw, Courville, and Touretzky (2003, 2006)

- Replace CSC with a Hidden Semi-Markov Model (HSMM) to handle early rewards correctly.
- Each state has a distribution of dwell times.
- Early reward forces an early state transition.



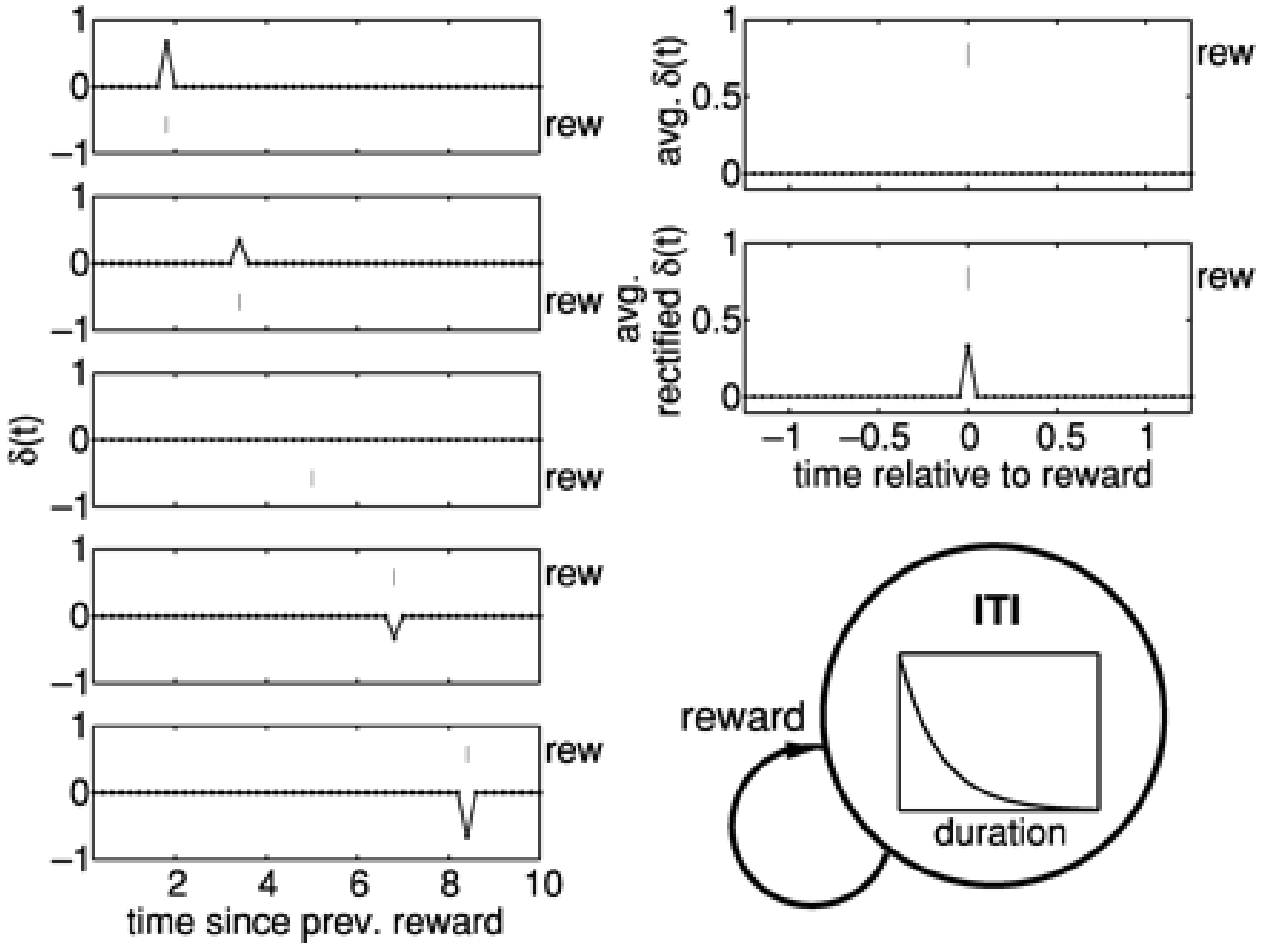
Early, Timely, and Late Rewards



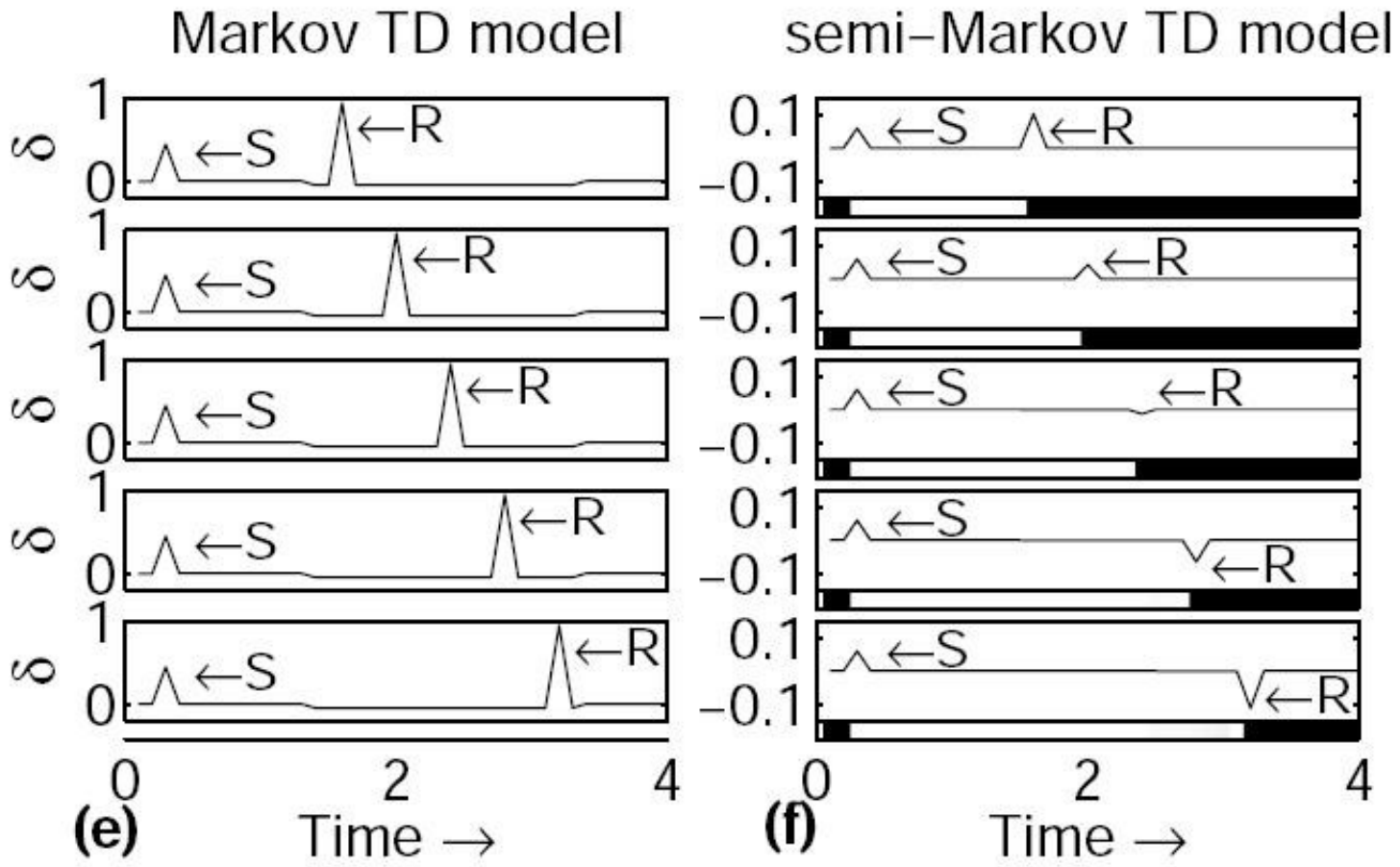
Black = ITI state, white = ISI state;
gray indicates uncertainty.

Unisgnalled Rewards at Poisson Intervals

- Mean reward prediction error is zero, but mean partially rectified error (simulated dopamine signal) is positive, matching the data.



Variable ISI



The hidden semi-Markov model shows reduced dopamine response when the reward appears later vs. earlier, in qualitative agreement with the animal data.

Summary

- Dopamine seems to encode several things: reward prediction error, novelty, and even aversive stimuli.
- The TD learning model does a good job of explaining dopamine responses to primary and secondary reinforcers.
- To properly account for timing effects the simple CSC representation must be replaced with something better.
- Example: Hidden Semi-Markov Models
  - Markov model = states plus transitions
  - “Hidden” means the current state must be inferred
  - “Semi-” means dwell times are drawn from a distribution; transitions do not occur deterministically
- But learning HSMMs is a hard problem: what are the states?
- How is an HSMM learned? Cortex!