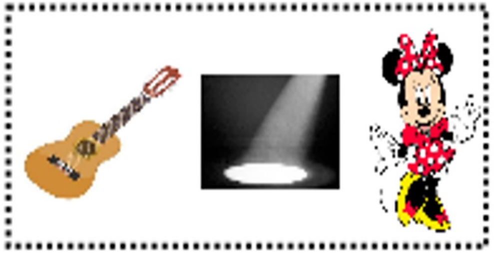
Reinforcement

Nisheeth

5th Jan 2018

*Interpret latent variables as **situations**. not **causes***

Index **situations** by stimuli co-occurrence patterns

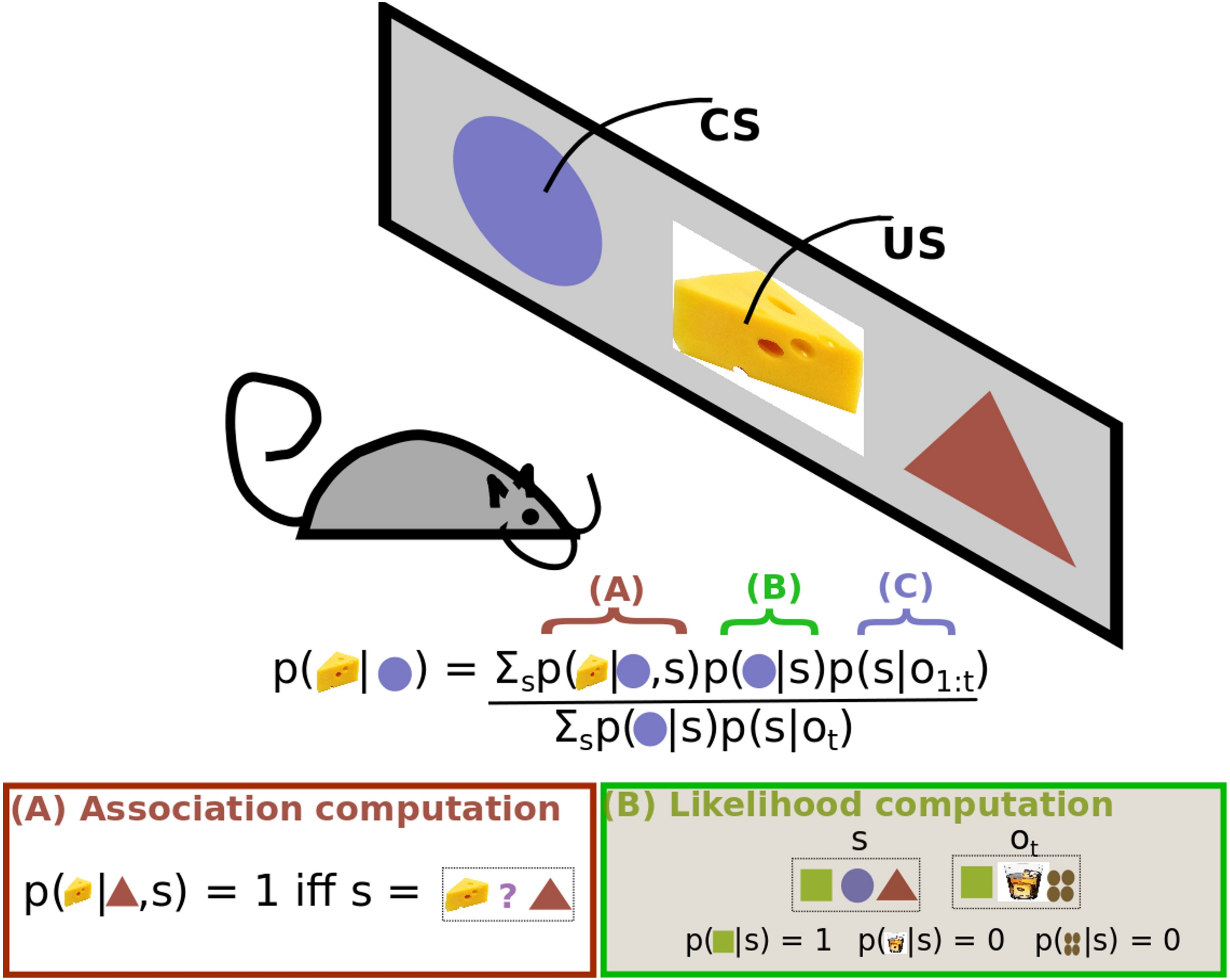


s = **Yeah!**

s = *Smoooth …*

s = **Yikes!!!**

https://link.springer.com/article/10.3758/s13420-012-0080-8


CS
US
(A)
(B)
(C)
(A) Association computation
(B) Likelihood computation
s
o_t

Bayes 101

- Bayes theorem is a simple consequence of conditional probability factoring

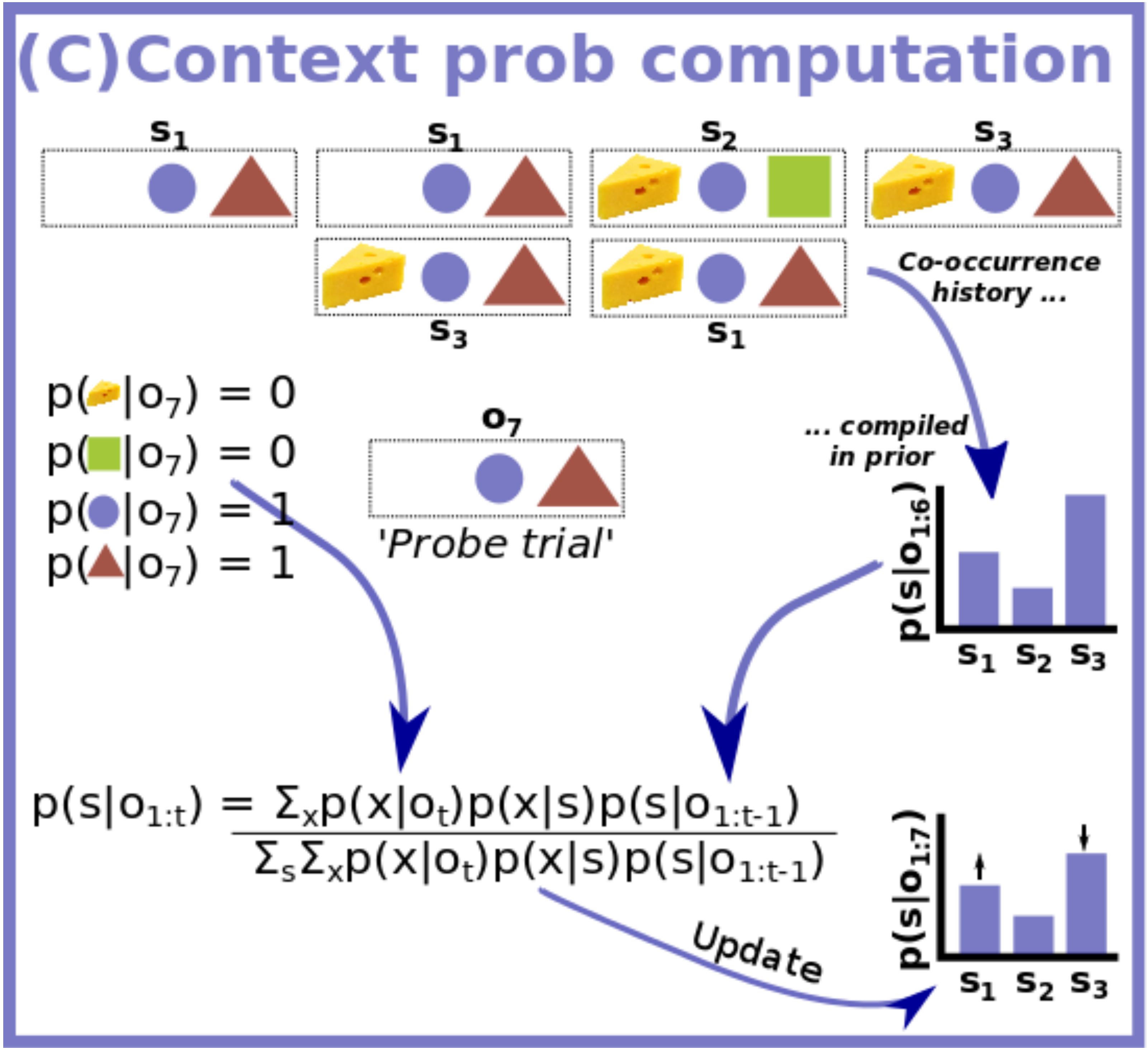
$$p(A \mid B)\,p(B) = p(B \mid A)\,p(A)$$

- Lends itself easily to sequential updates

$$p(m \mid obs_{1:t}) = \frac{p(obs_t \mid m)\,p(m \mid obs_{1:t-1})}{\sum_m p(obs_t \mid m)\,p(m \mid obs_{1:t-1})}$$

- Great fit for cognitive modeling
  - Models interaction of already known with new data

Understand the Bayes Theorem first https://www.youtube.com/watch?v=HZGCoVF3YvM



The occurrence of the new evidence updates the poterior (as the prior for the next step)

Some RW failures explained by latent cause model

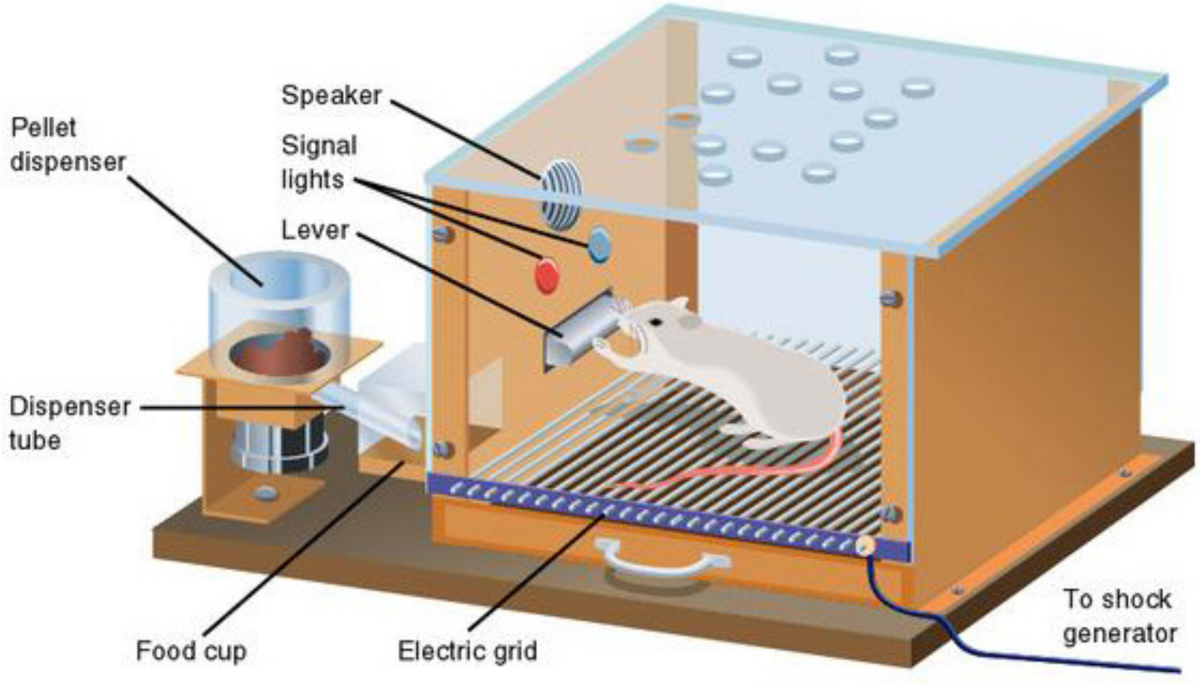
- Spontaneous recovery from extinction
- Facilitated reacquisition
- Conditioned inhibitor pairing
- Pre-exposure effect
- Higher order conditioning

Association vs reinforcement

- Association: things that occur together in the world, occur together in the mind
  - Tested using classical conditioning
  - Environment acts on the observer
- Reinforcement: actions that are rewarded become desirable in future
  - Tested using operant/instrumental conditioning
  - Observer acts on the environment

Operant conditioning

- Observers act upon the world, and face consequences
  - Consequences can be interpreted as rewards



Modeling classical conditioning

- Most popular approach for years was the Rescorla-Wagner model

$$\Delta V_X^{n+1} = \alpha_X \beta(\lambda - V_{tot})$$

$$V_X^{n+1} = V_X^n + \Delta V_X^{n+1}$$

Some versions replace V_{tot} with V_x; what is the difference?

- Could reproduce a number of empirical observations in classical conditioning experiments

Can modify to accommodate reward prediction

- Original equation
  - Update size based on *associative strength* available

$$V_X^{n+1} = V_X^n + \alpha(\lambda - V_{tot})$$

- Bush-Mosteller model of reinforcement, for action *a*

$$V_a^{n+1} = V_a^n + \alpha(R^n - V_a^n)$$

Generalized reinforcement learning

- Bush Mosteller style models simply update value based on a discounted average of received rewards
  - Useless in trying to predict the value of sequential events, e.g. A → B → reward
- A more generalized notion of reward learning was needed
  - Temporal difference learning
  - Other flavors of reinforcement learning (out of scope)

Reinterpreting the learning gradient

- In Bush Mosteller, the reward prediction error is driven by the difference between
  - A discounted average of received rewards
  - The current reward
- In TD learning, RPE is the difference between
  - Expected value of discounted future rewards

$$F^{n} = R^{n+1} + \gamma R^{n+2} + \gamma^{2} R^{n+3} + ...$$

  - Information suggesting the expectation is mistaken

http://www.scholarpedia.org/article/Temporal_difference_learning

The TD learning algorithm

- Bush Mosteller algorithm

$$V_a^{n+1} = V_a^n + \alpha(R^n - V_a^n)$$

- TD algorithm

$$V_a^{n+1} = V_a^n + \alpha(F^n - V_a^n)$$

- Discounted future rewards not available instantaneously
  - Use math trick

$$F^n = R^{n+1} + \gamma F^{n+1}$$

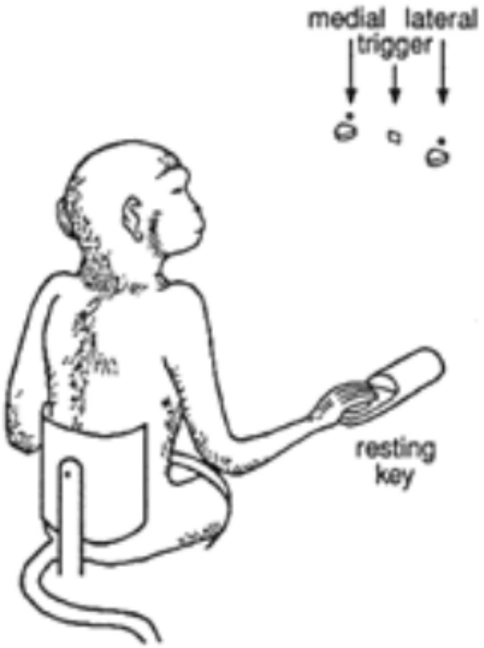
The TD reward prediction error

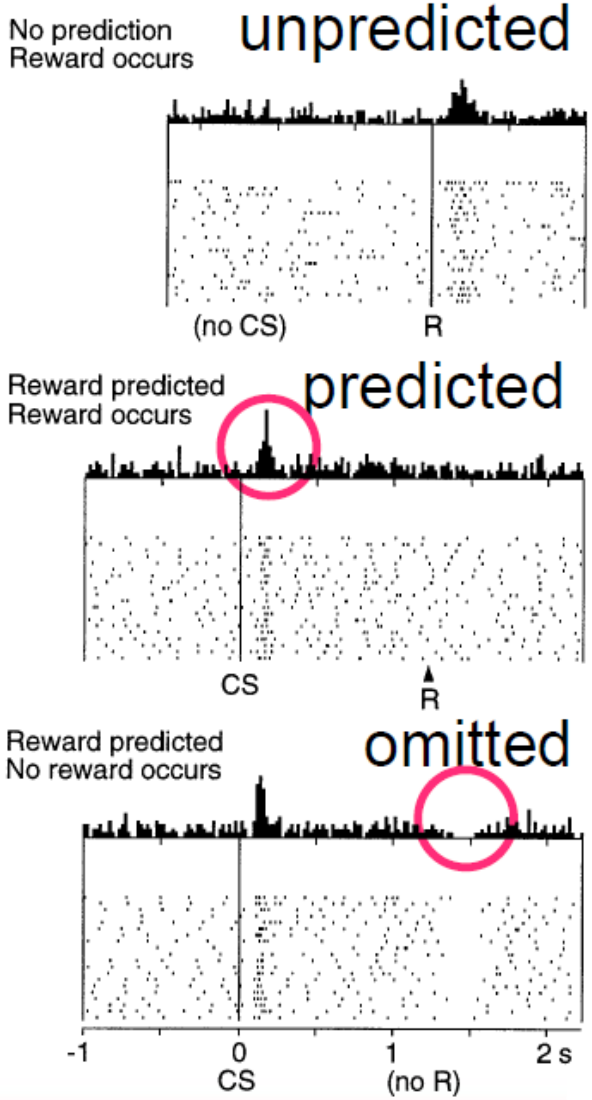
$$\delta^{n\square 1} \square R^{n\square 1} \square \gamma V^{n\square 1} - V^{n}$$

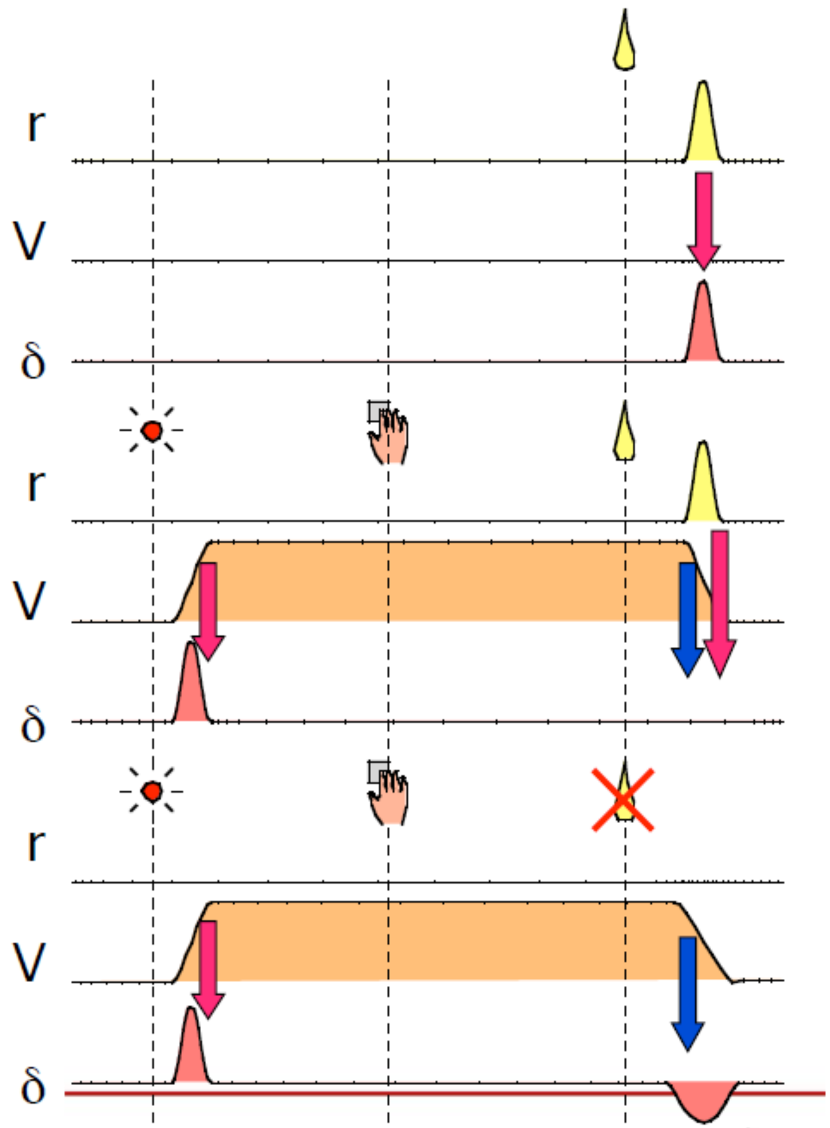
Learning continues until reward expectations are perfectly aligned with received reward

Dopamine Neurons Code TD Error

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



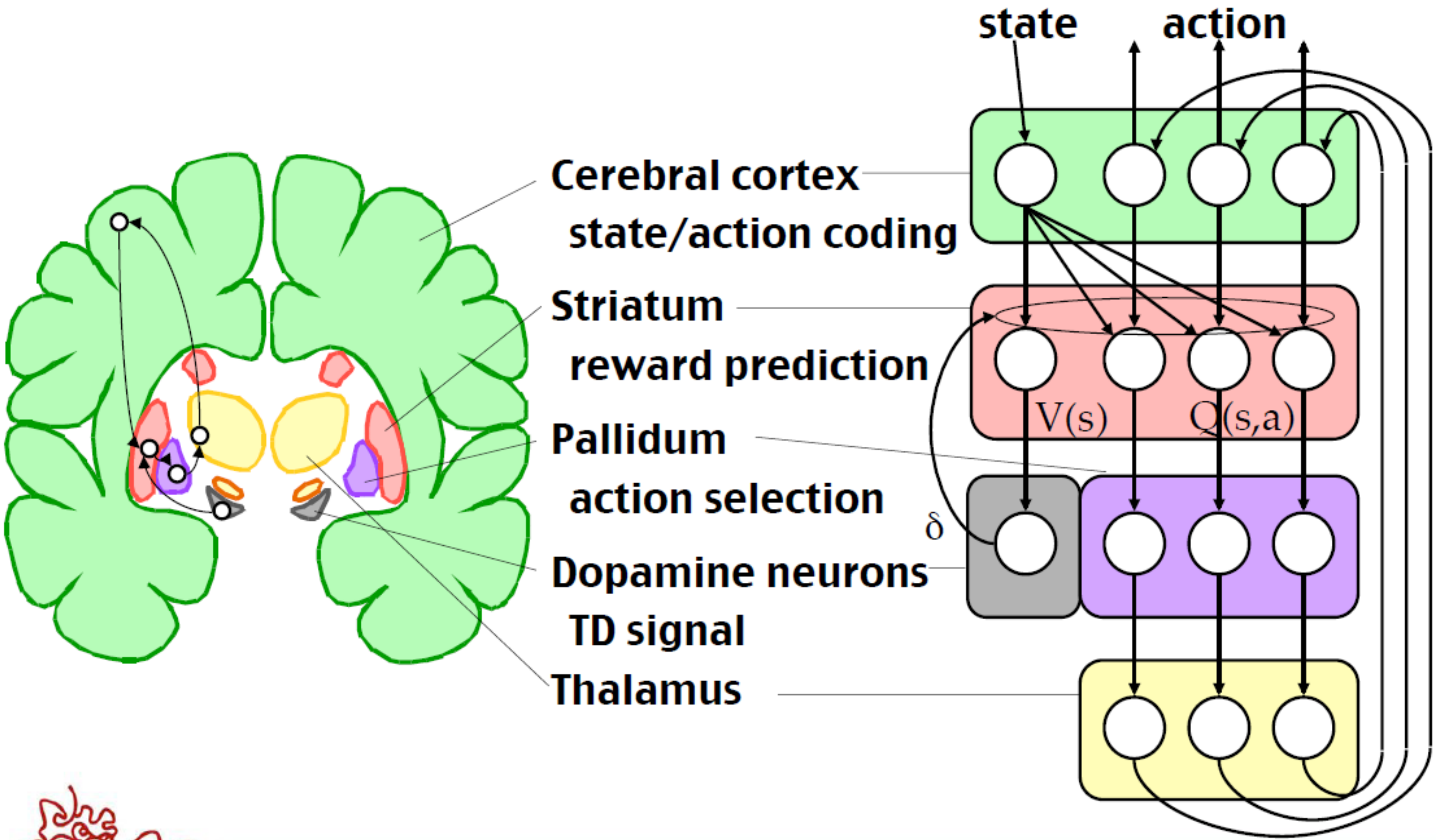




(Schultz et al. 1997)

Basal Ganglia for Reinforcement Learning?

(Doya 2000, 2007)



http://www.pnas.org/content/108/Supplement_3/15647.full

Addiction as a computational process gone awry

David A. Redish, Science 2004

Under natural circumstances, the temporal difference signal is the following:

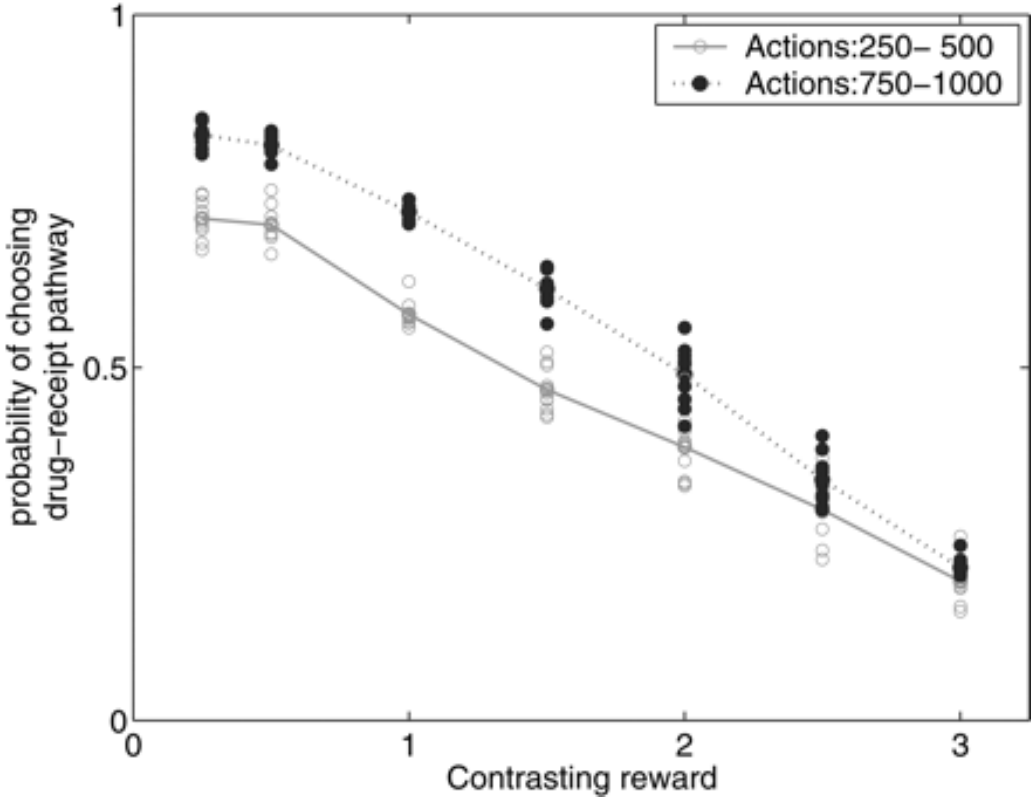
$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

The idea is that the drug (especially dopaminergic drugs like cocaine) may induce a small temporal difference signal directly (D), such that:

$$\delta_t = \max\left(r_{t+1} + \gamma V(s_{t+1}) - V(s_t) + D_t, D_t\right)$$

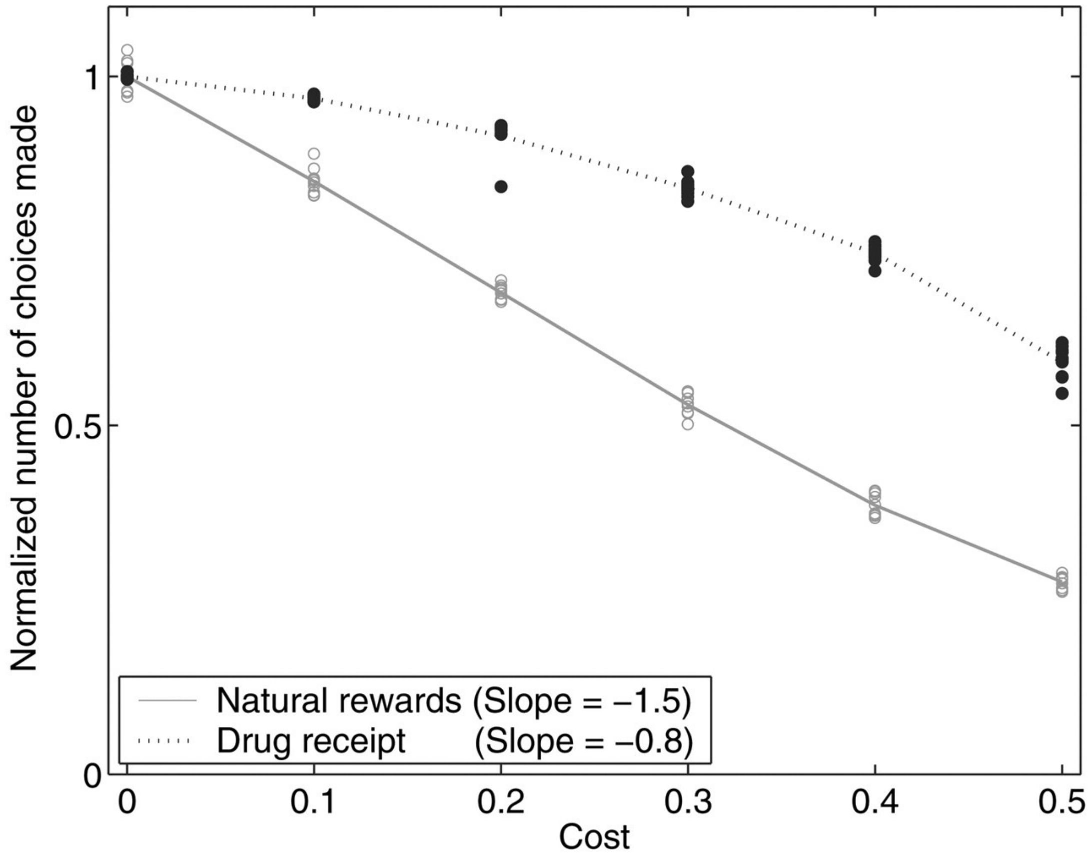
In the beginning the temporal difference signal is high, because of the high reward value of the drug (rational addiction theory). But with longer use, the reward value might sink, and negative consequences would normally reduce the non-adaptive behavior. But because d is always at least D, the behavior can not be unlearned.

Increased wanting (not more liking)



The model predicts that with continued use, the drug-seeking behavior becomes more insensitive to contrasting reward.

Decreased Elasticity



Elasticity is a term from economics. It measures how much the tendency to buy products decreases, as the price increases.

Because drug-seeking can not easily be unlearned, the behavior become less and less elastic with prolonged drug use.

Open questions

- Exploration, curiosity
- Locus of control and its effects
- Sub-goal construction and state hierarchy construction