Memory and the statistical structure of the world

Samuel Gershman

Department of Brain and Cognitive Sciences, MIT

July 2014

Marco Polo

Marco Polo

Is this a funny-looking unicorn?

Big picture

- When do we modify old memories, and when do we create new ones?

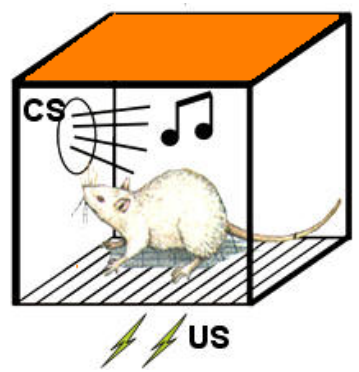
Big picture

- When do we modify old memories, and when do we create new ones?
- This question can be answered within a probabilistic computational framework: *we create new memories when we infer new latent causes in our environment*

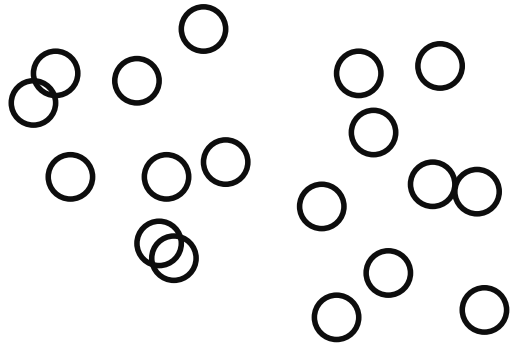
Big picture

- When do we modify old memories, and when do we create new ones?
- This question can be answered within a probabilistic computational framework: *we create new memories when we infer new latent causes in our environment*
- This principle has deep explanatory power across multiple domains

- Classical conditioning



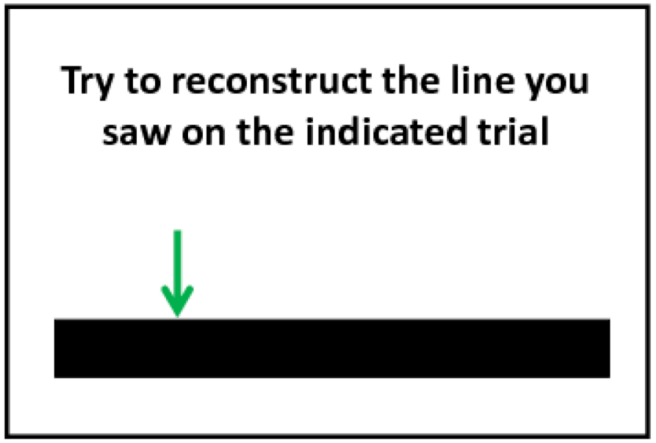
- Perceptual estimation



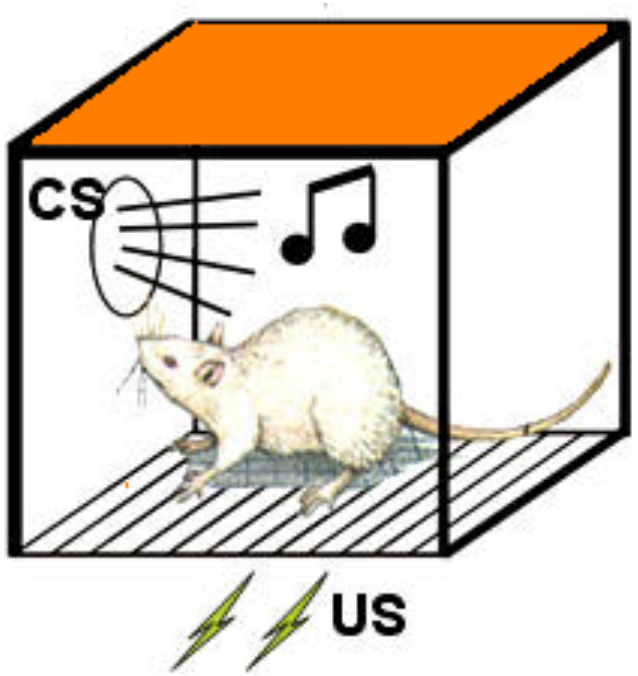
How many circles?

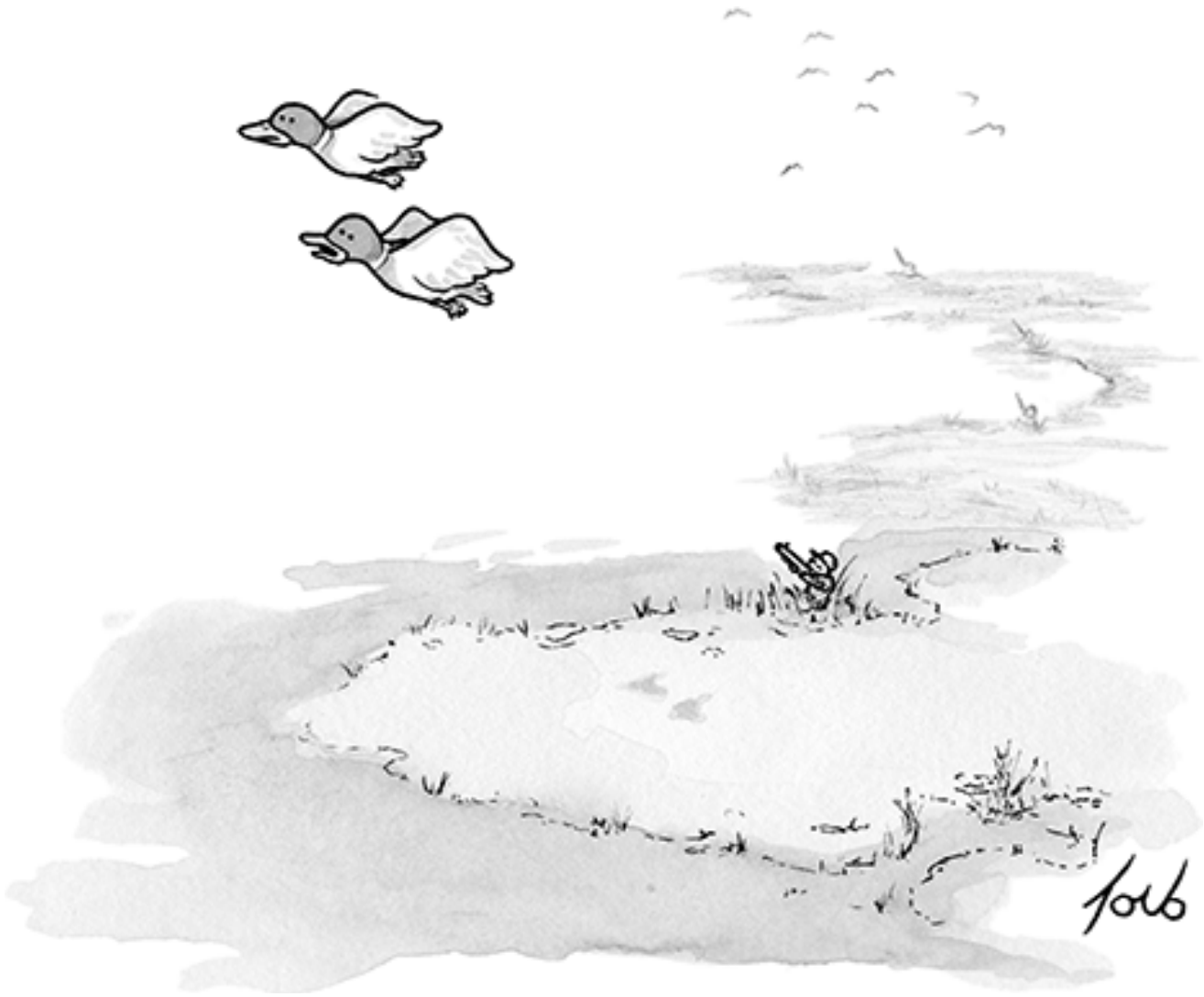
- Reconstructive memory

Reconstruction trial



What do animals learn during classical conditioning?

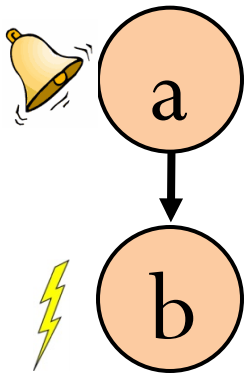




"It's that time of year when guys randomly explode."

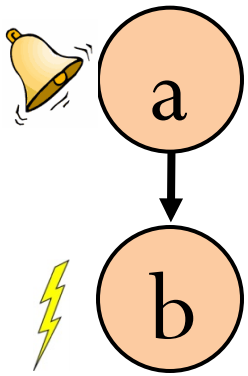
Some possibilities

Tone (a) causes shock (b)

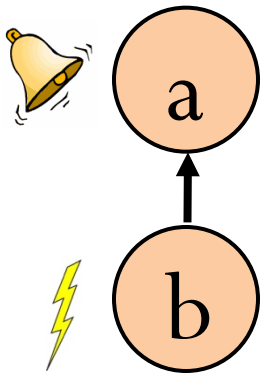


Some possibilities

Tone (a) causes shock (b)

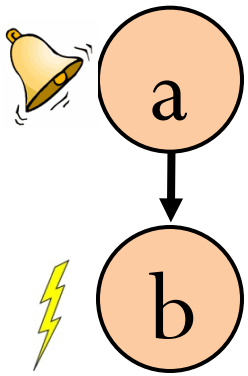


Shock causes tone

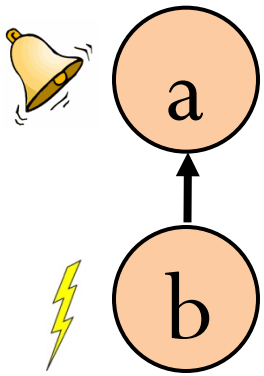


Some possibilities

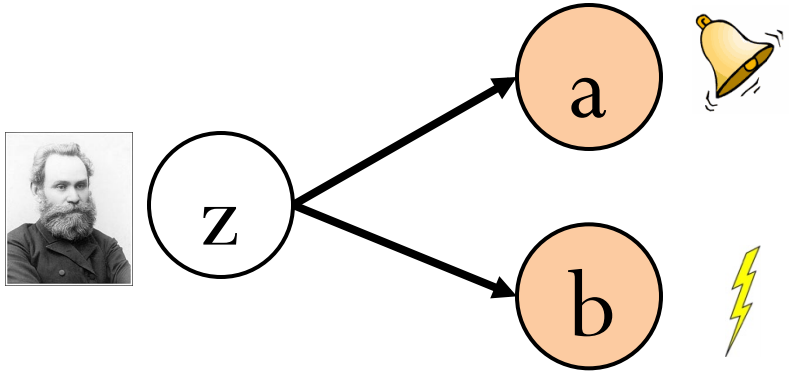
Tone (a) causes shock (b)



Shock causes tone

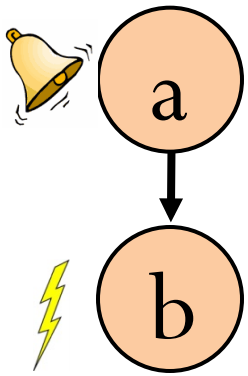


Something else (z) causes both tone and shock

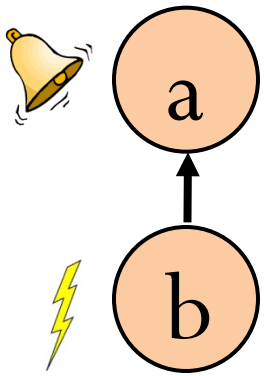


Some possibilities

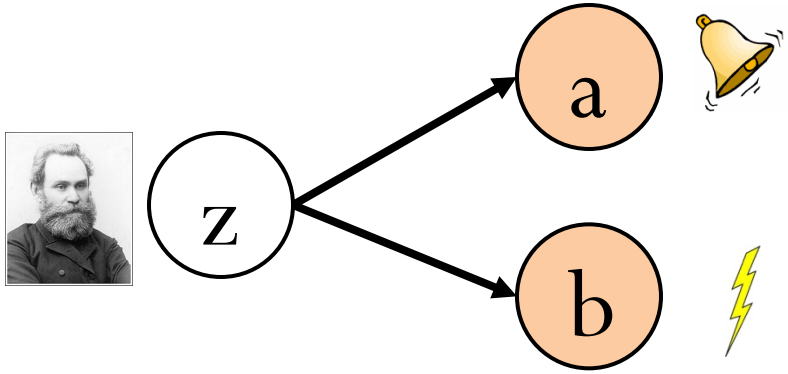
Tone (a) causes shock (b)



Shock causes tone



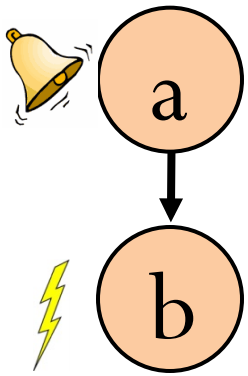
Something else (z) causes both tone and shock



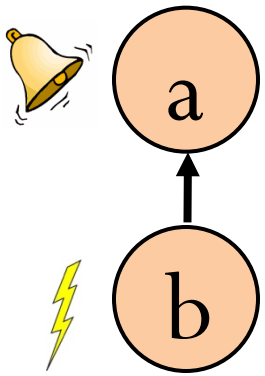
Too constrained

Some possibilities

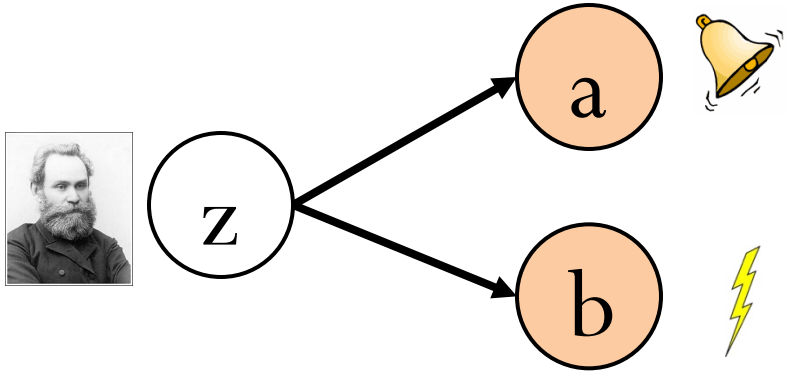
Tone (a) causes shock (b)



Shock causes tone



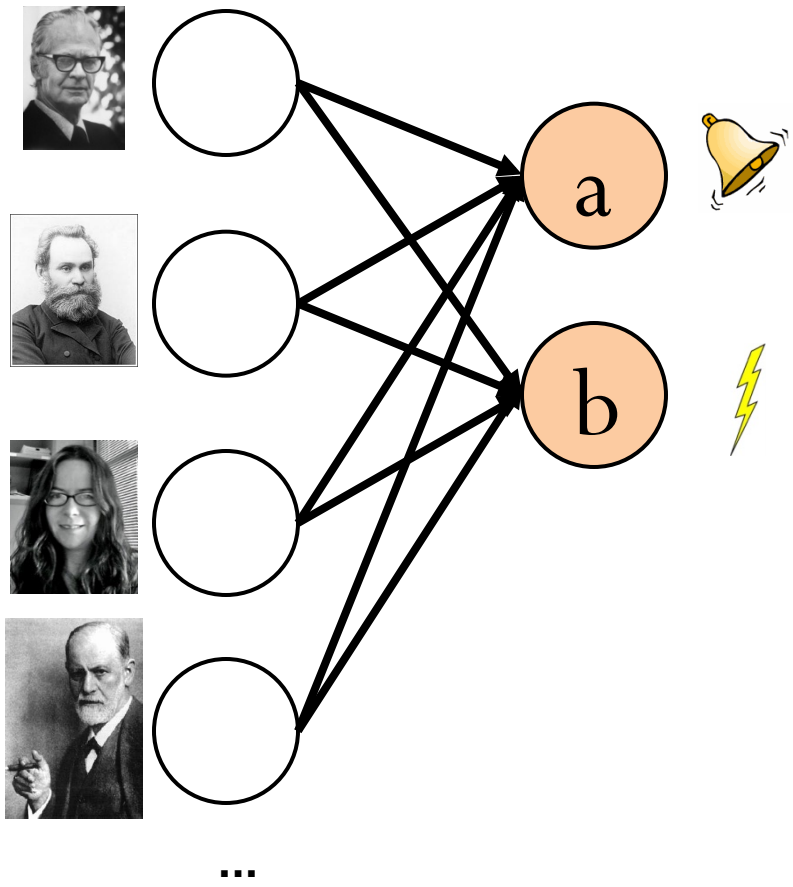
Something else (z) causes both tone and shock



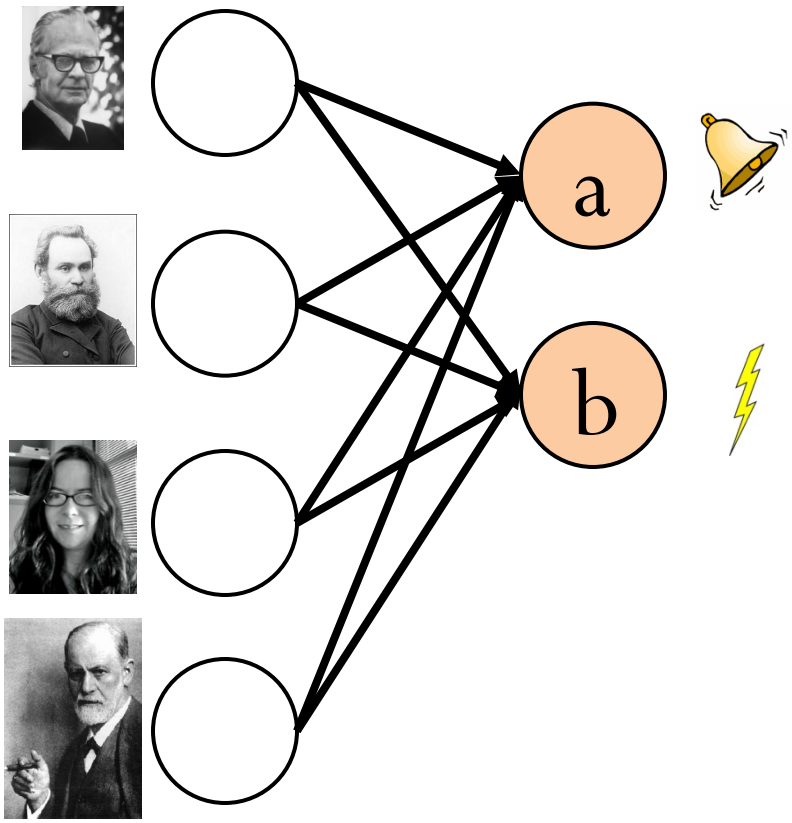
Too constrained

Too flexible?

Too flexible?



Too flexible?

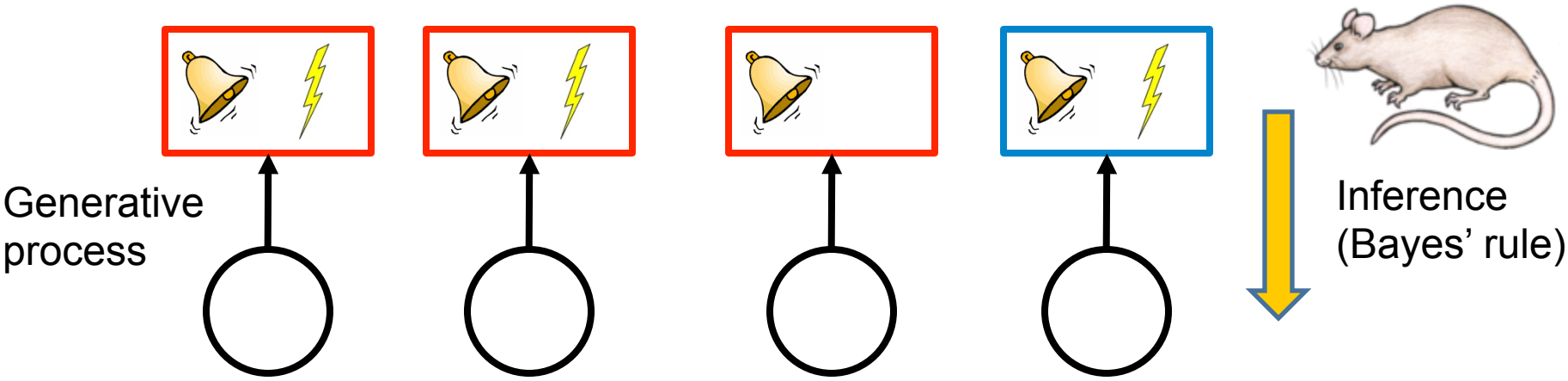


Hypothesis: Animals assume a generative model in which (1) the number of latent causes is unbounded, and (2) a small number of latent causes is more likely *a priori*.

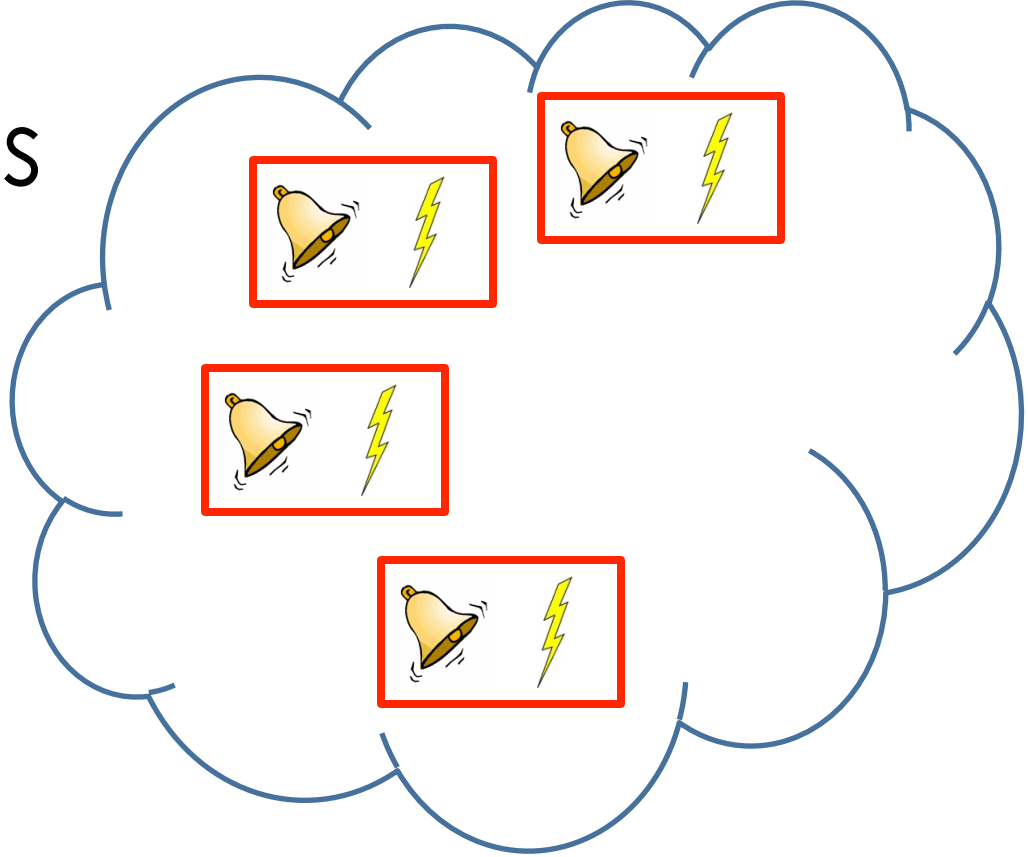
Inverting the generative model

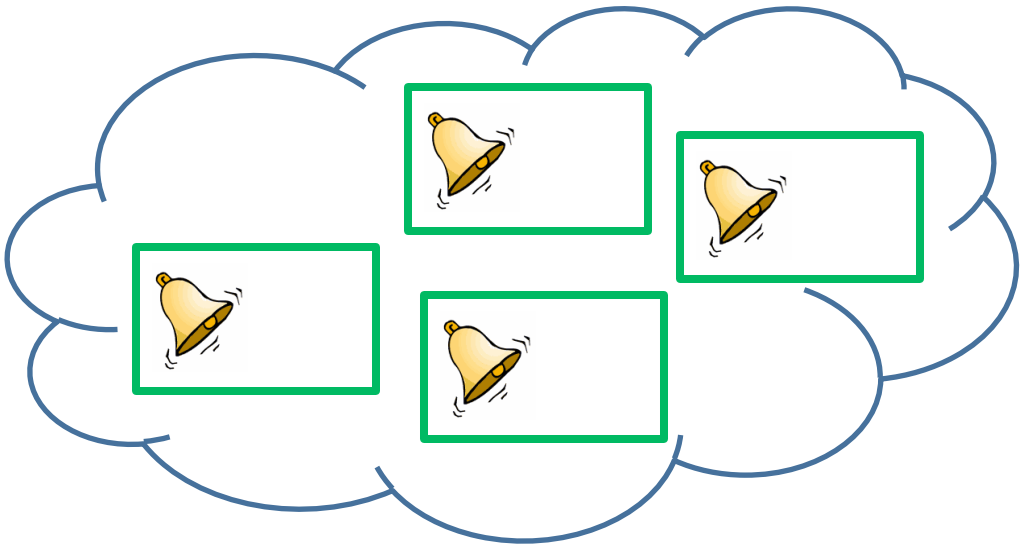
Bayes' rule inverts generative model to infer latent causes:

$$P(\text{cause} \mid \text{data}) \propto P(\text{data} \mid \text{cause})P(\text{cause})$$



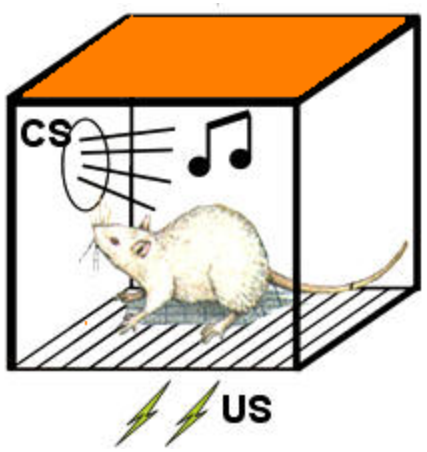
Conditioning as clustering





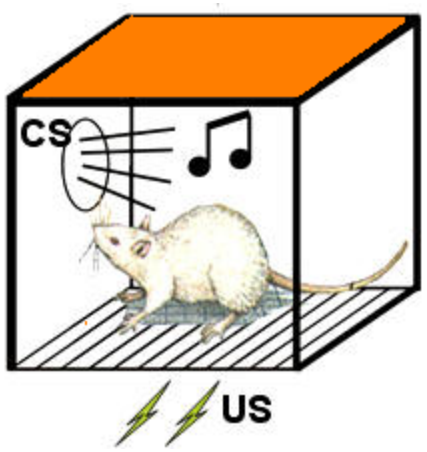
Case study: renewal

Acquisition (box A)

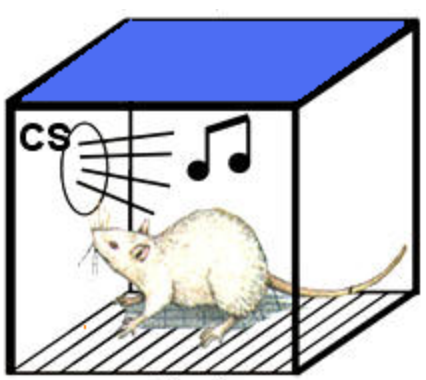


Case study: renewal

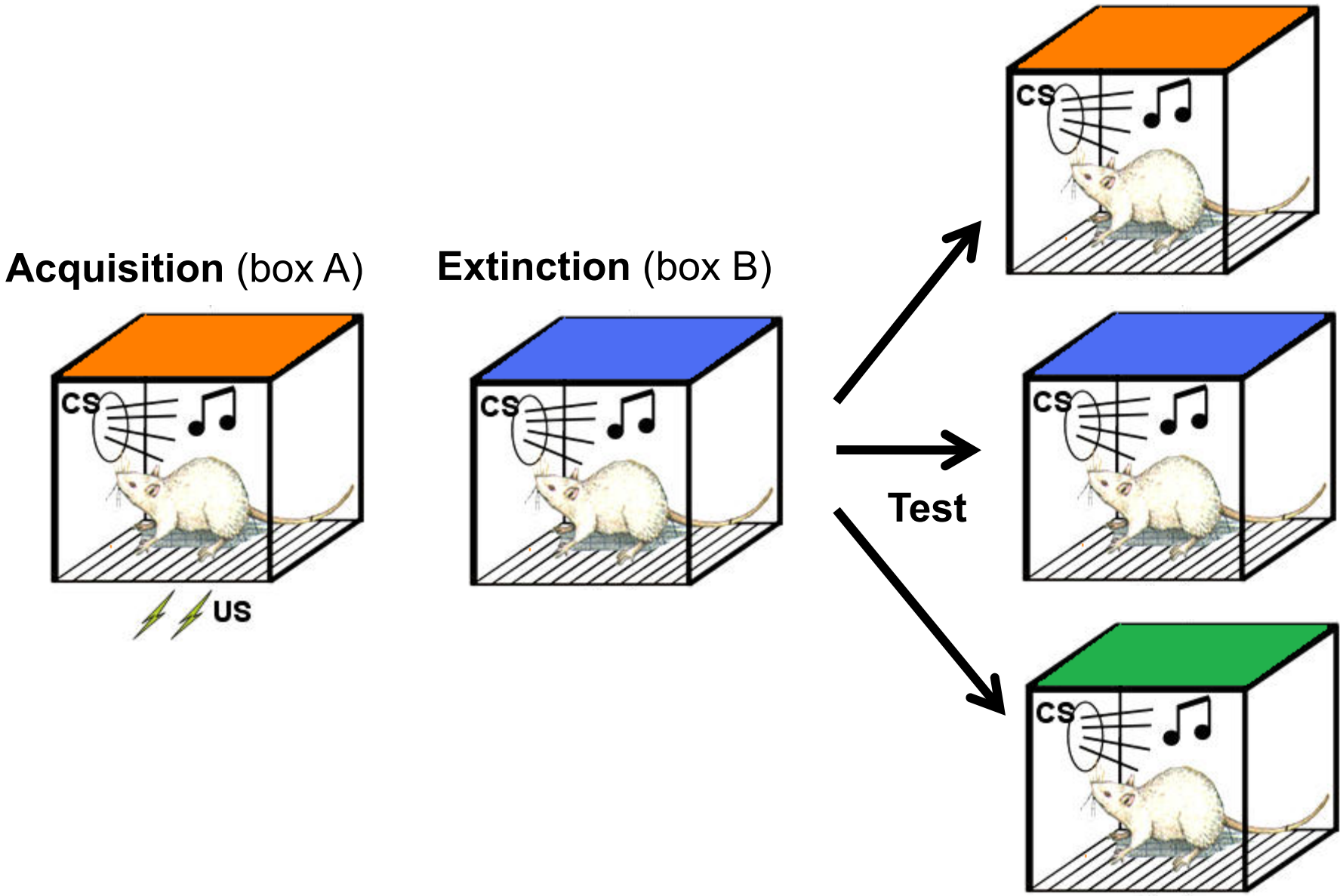
Acquisition (box A)



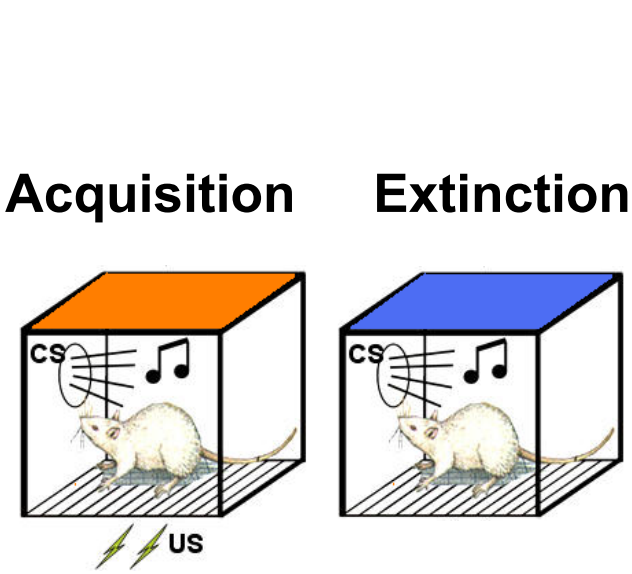
Extinction (box B)

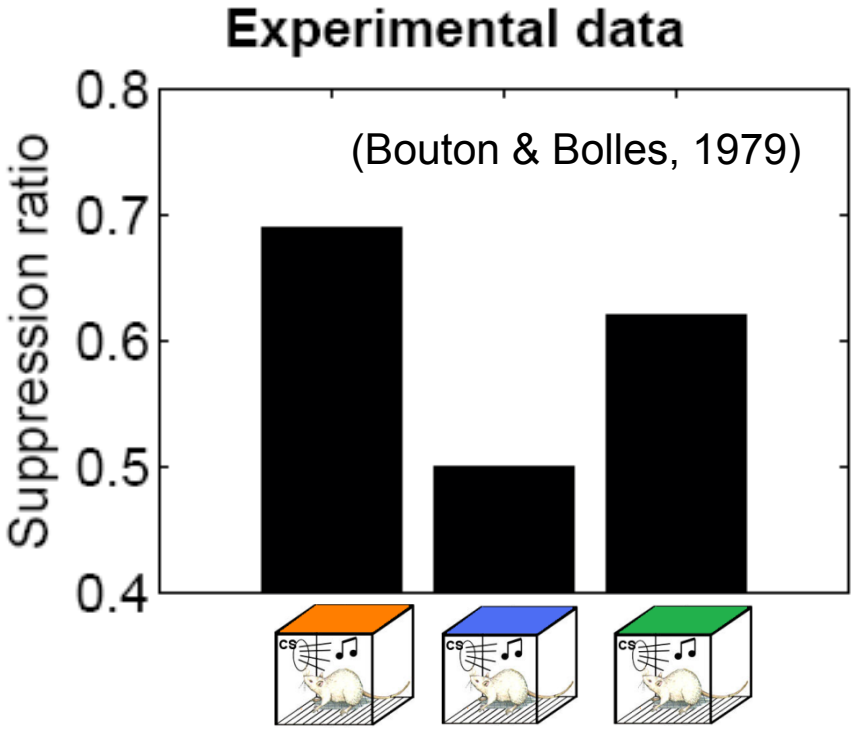


Case study: renewal

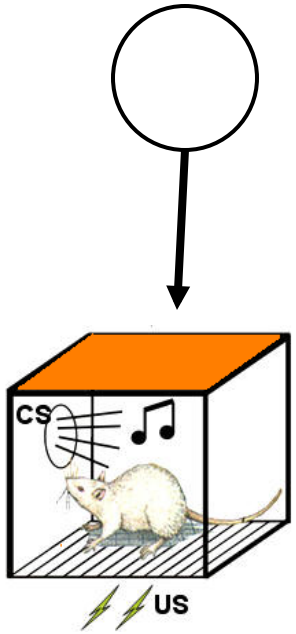


Conditioned responding is renewed!

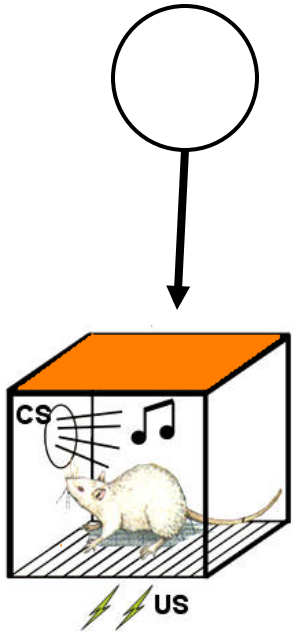




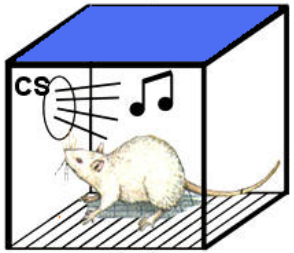
The rat hasn't unlearned its conditioned response; it has ***learned something new.***



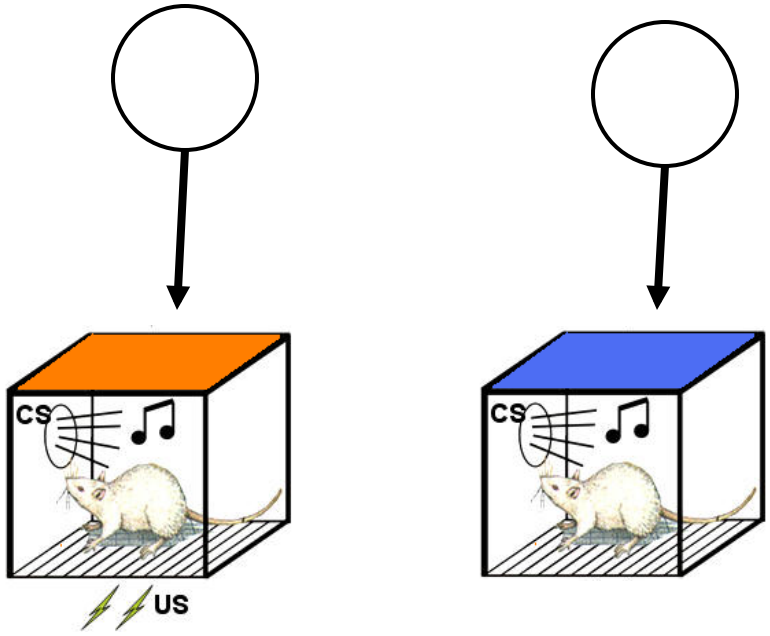
Acquisition



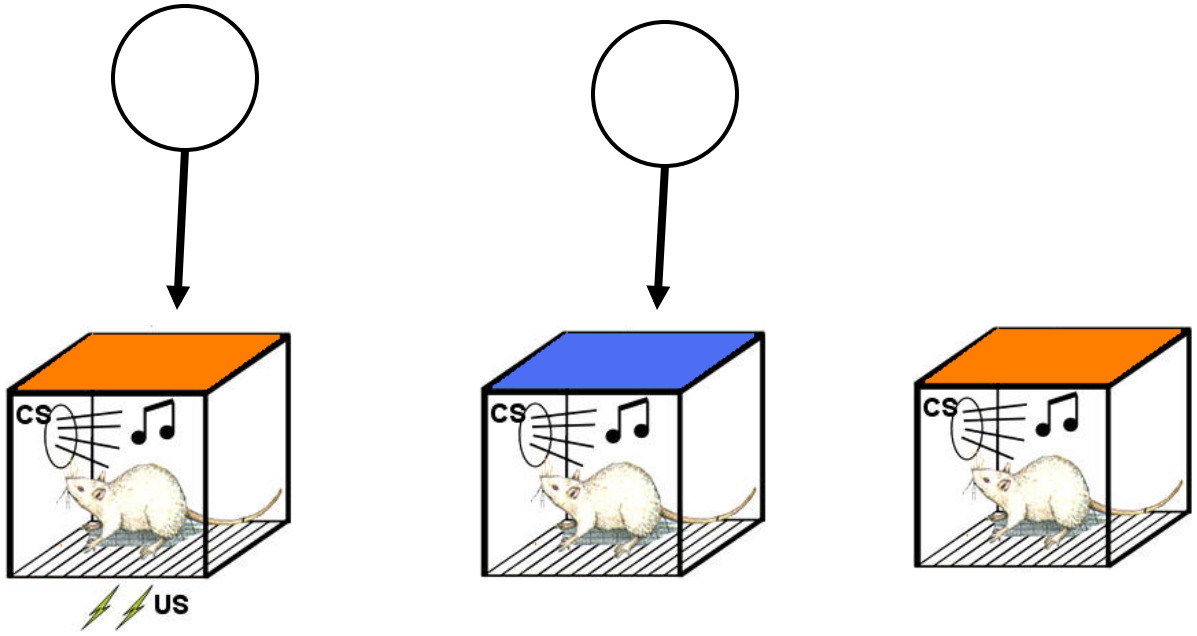
Acquisition



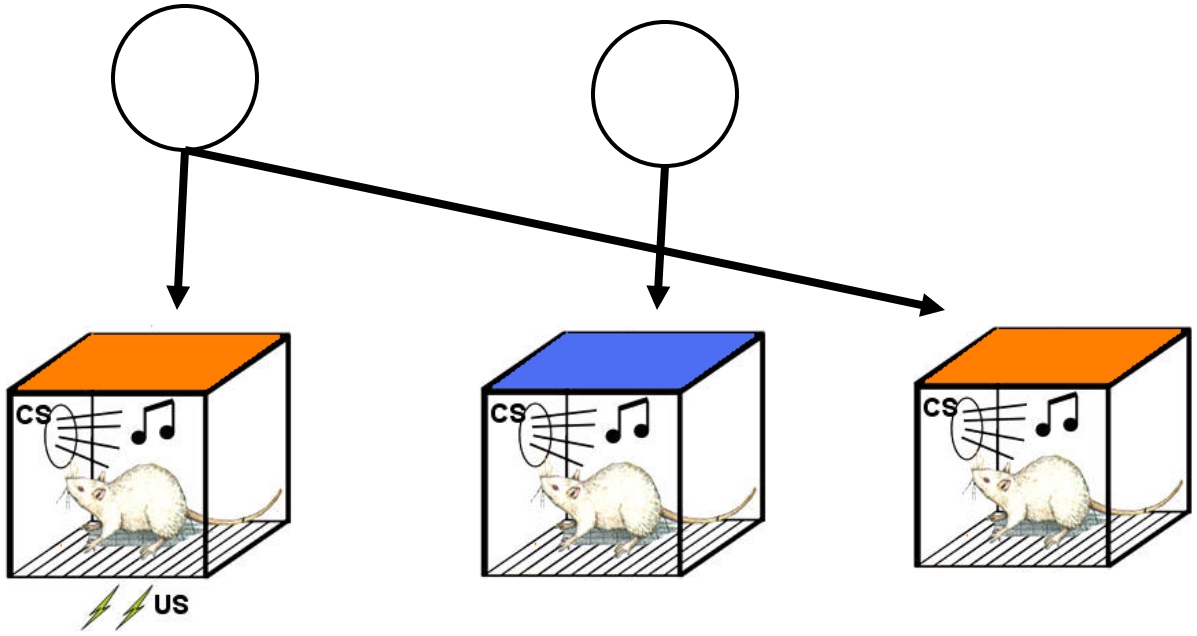
Extinction



Acquisition **Extinction**

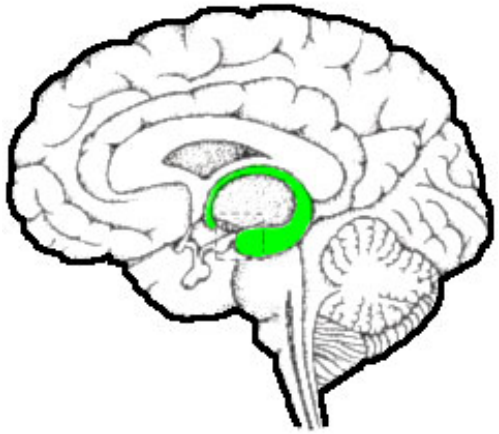


Acquisition **Extinction** **Test**

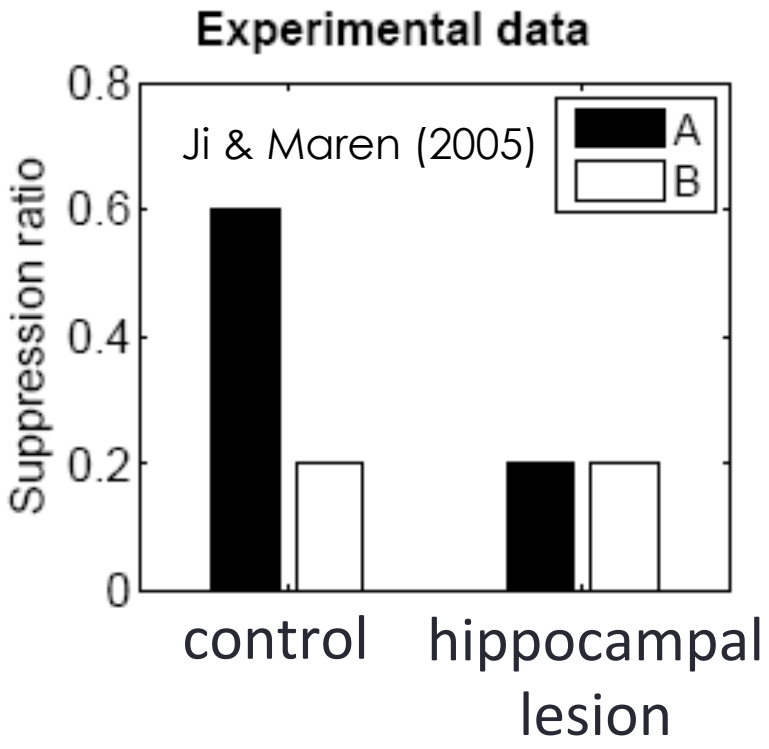

CS
US
CS
CS
Acquisition
Extinction
Test

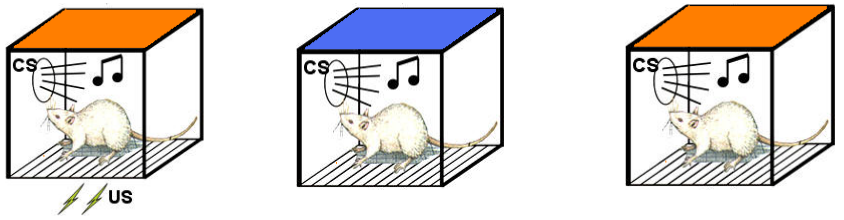
Clustering in the brain

Hippocampus supports the ability to flexibly infer new latent causes



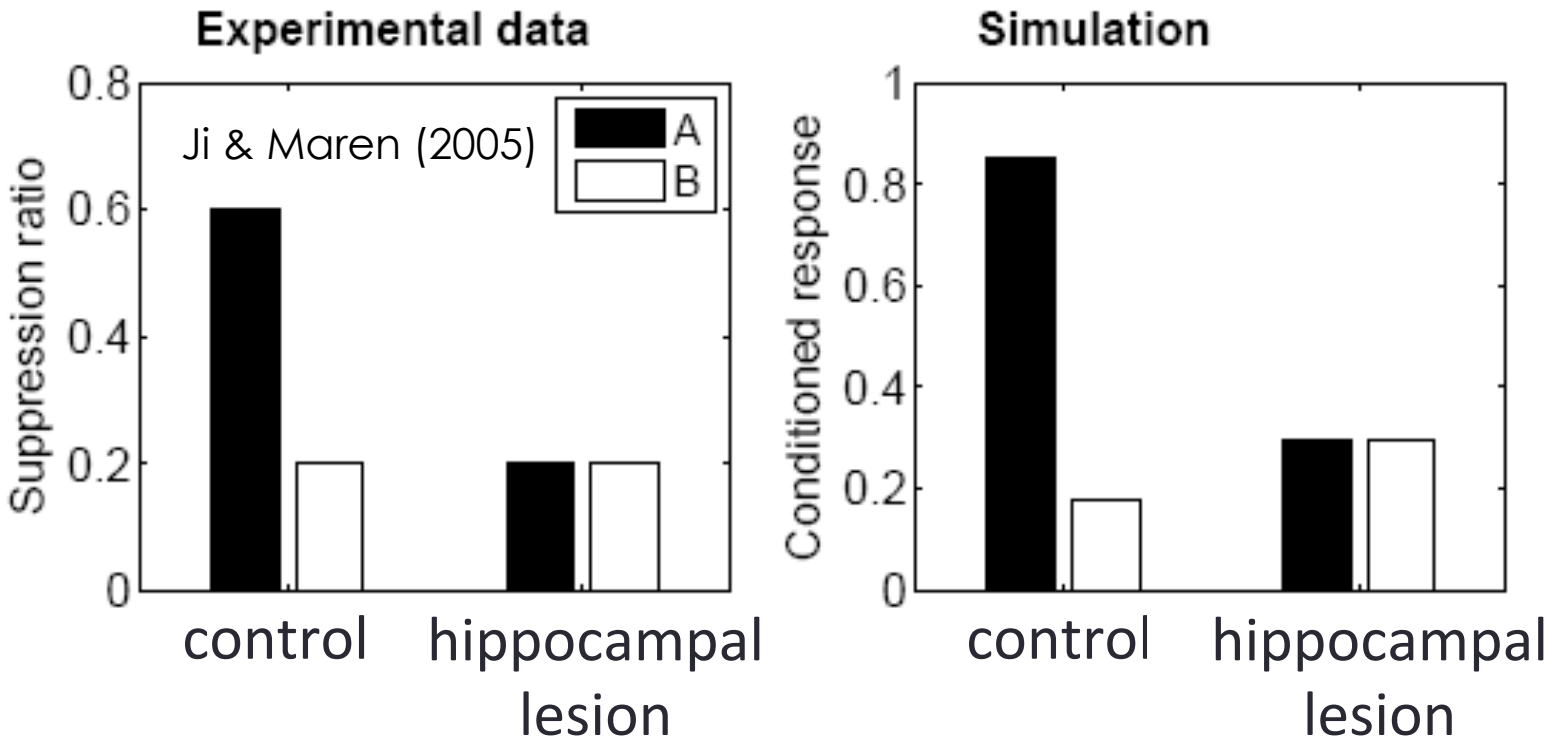
Pre-training lesions of hippocampus abolish renewal





Acquisition (A) **Extinction (B)** **Test (A)**

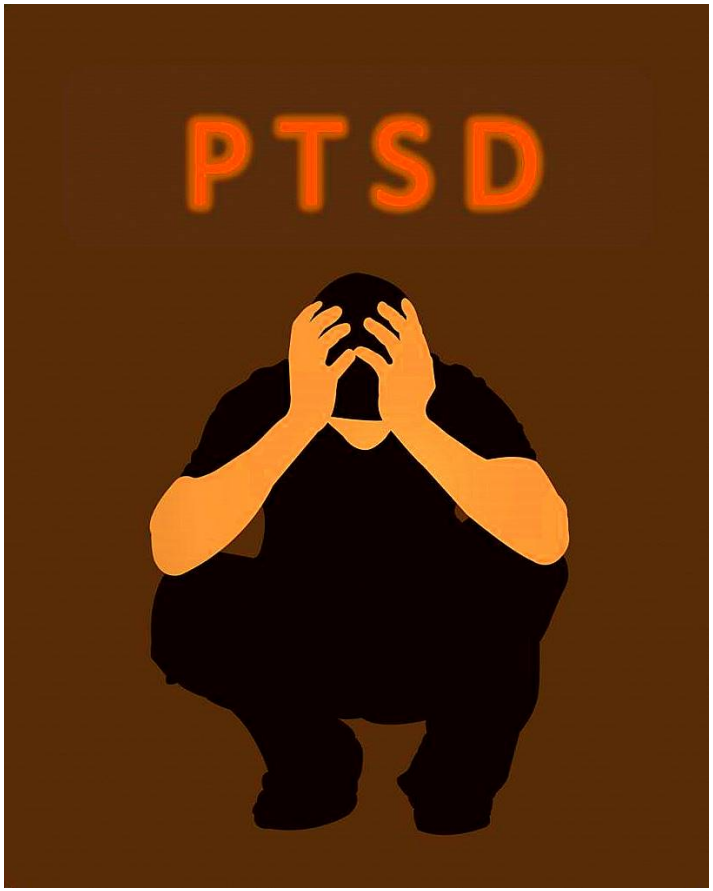
Pre-training lesions of hippocampus abolish renewal



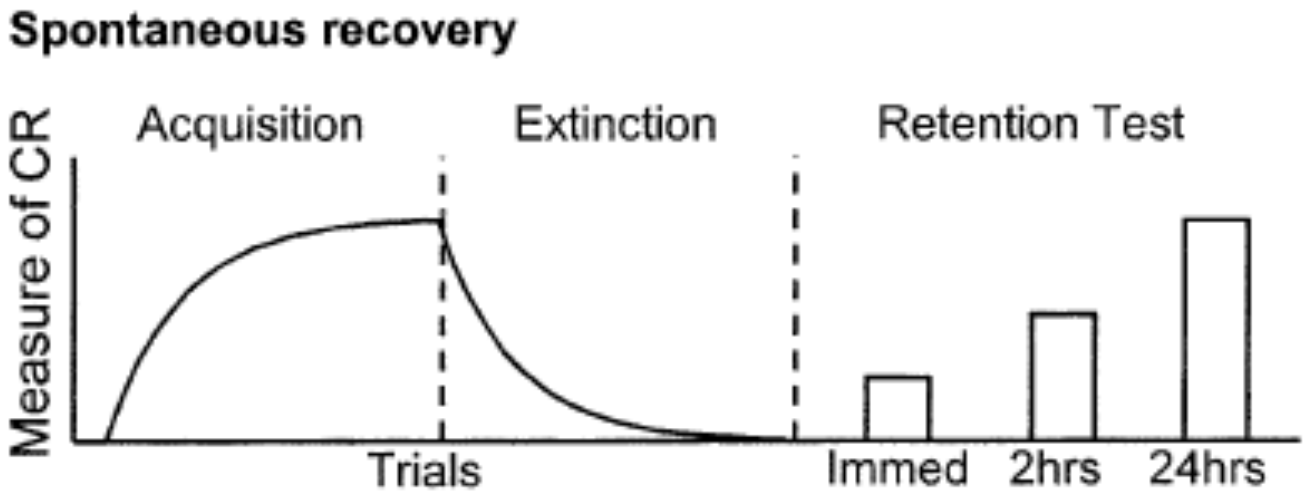
Hippocampal lesions handicap the model's ability to infer new clusters

Gershman, Blei & Niv (2010), *Psych Review*

Why are memories hard to modify?

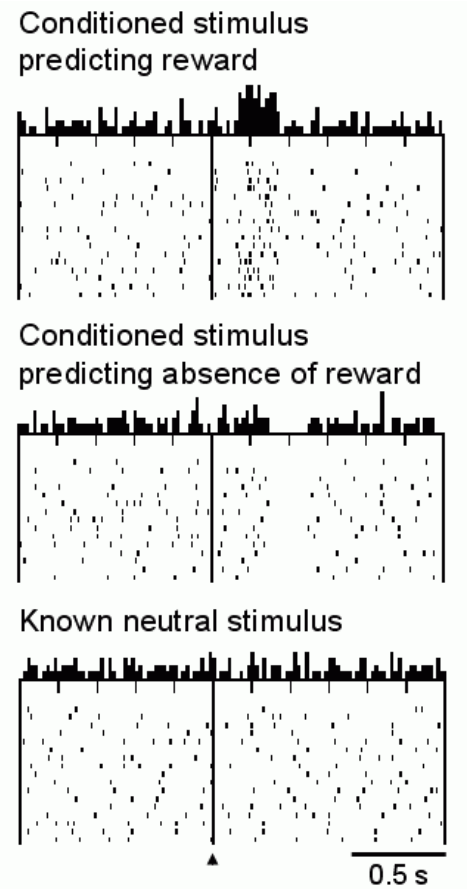


Relapse in classical conditioning



Myers & Davis (2002)

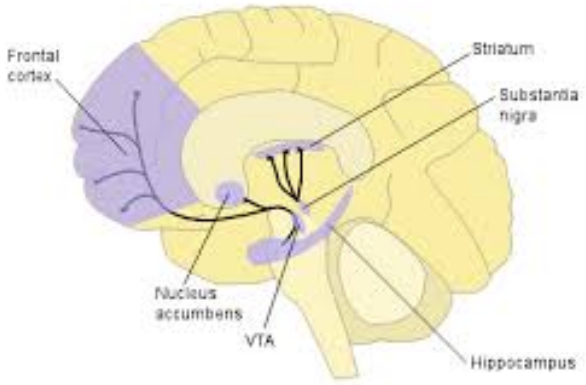
Prediction errors and learning



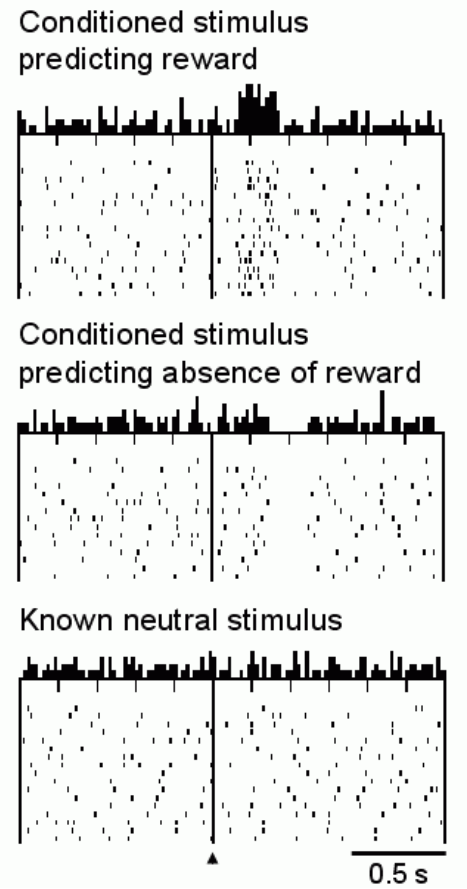
Schultz (1998)

Rescorla-Wagner model

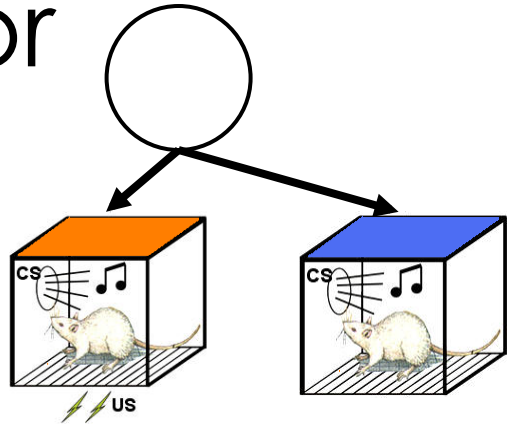
$$V \leftarrow V + \eta\,[\underbrace{\text{outcome} - V}_{\text{prediction error}}]$$



An alternative view: two roles for prediction error

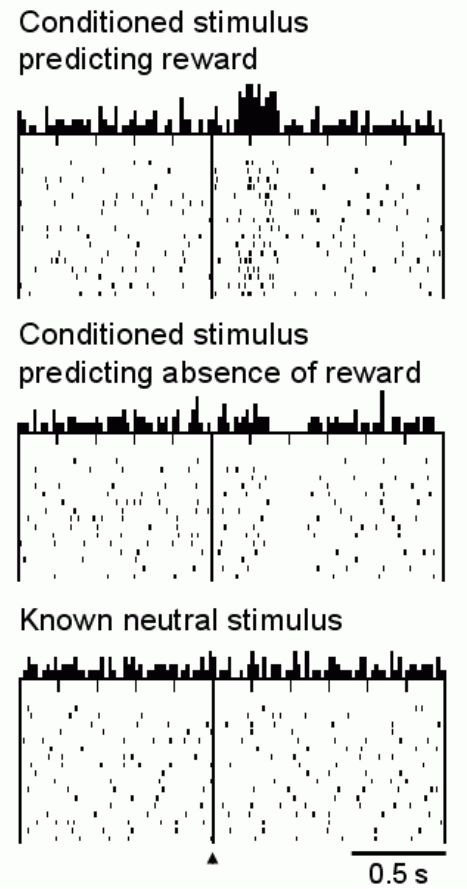


When errors are small:
memory modification



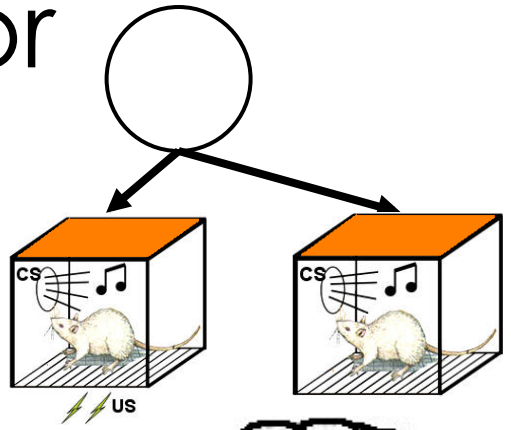
Schultz (1998)

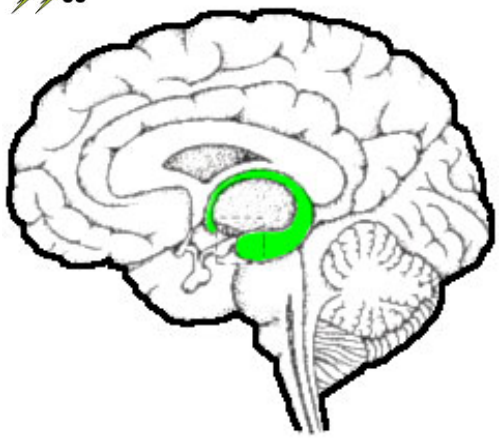
An alternative view: two roles for prediction error

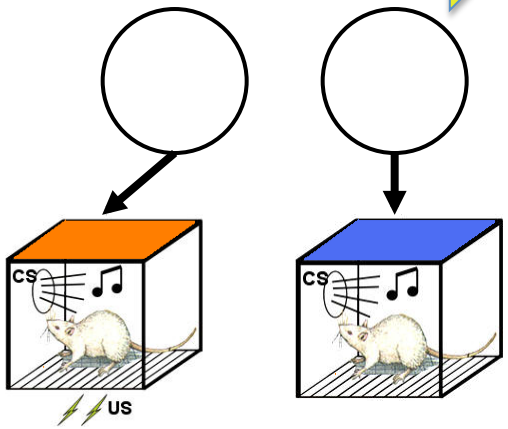


Schultz (1998)

When errors are small:
memory modification







When errors are large:
memory formation

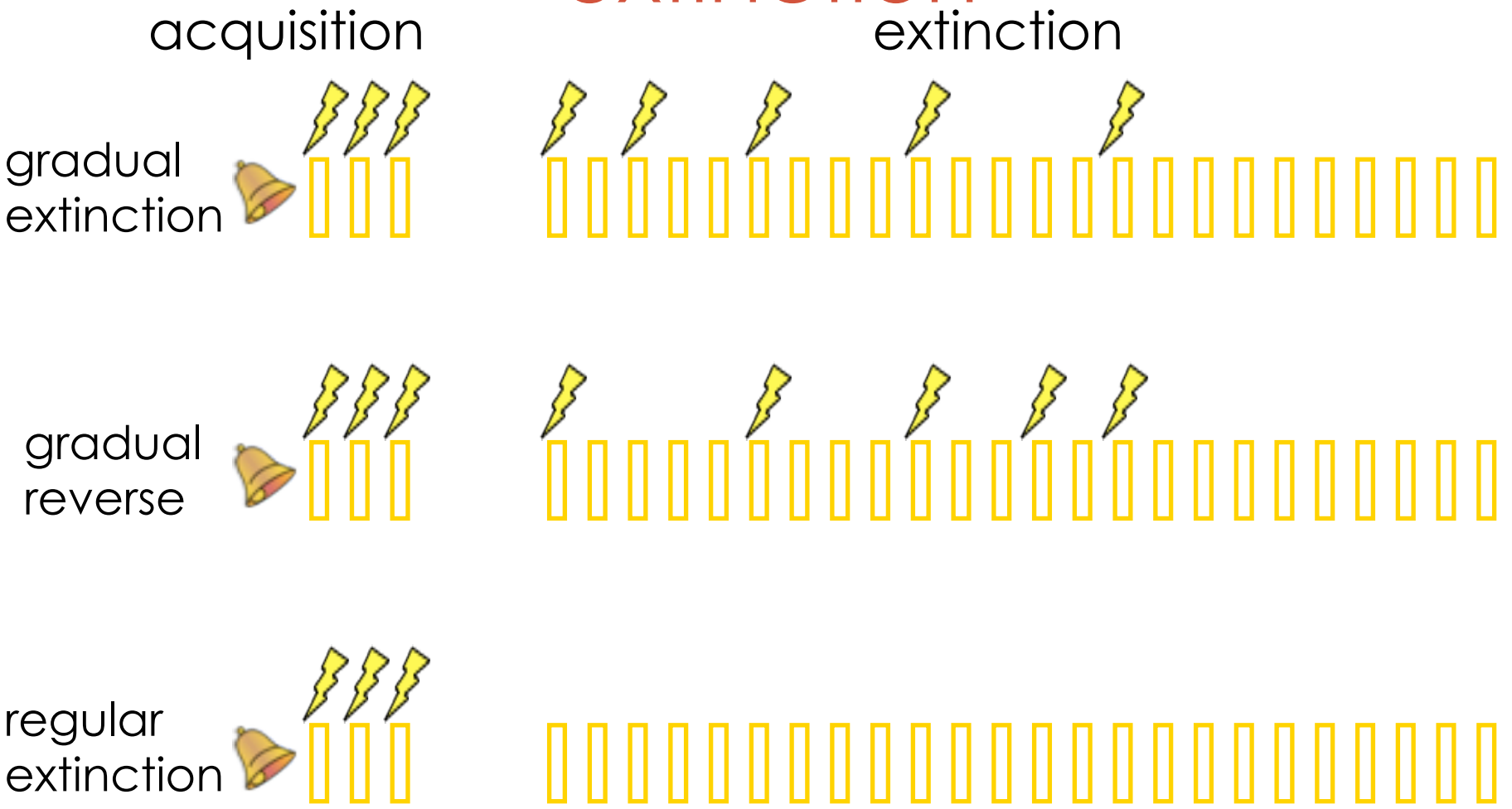
How to erase a fear memory

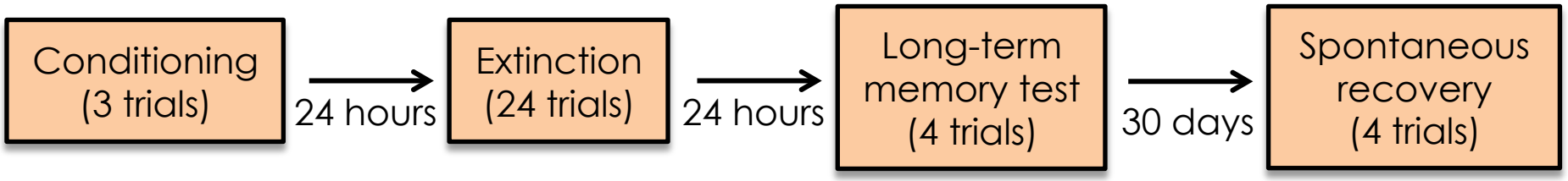
- Prediction errors should be large enough to drive learning, but not so large that a new latent cause is inferred.

How to erase a fear memory

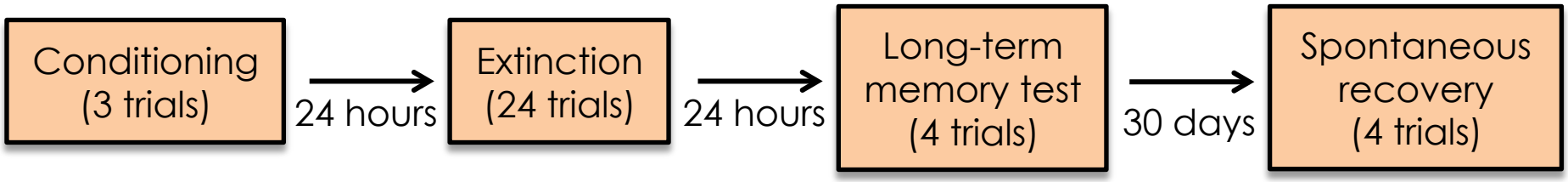
- Prediction errors should be large enough to drive learning, but not so large that a new latent cause is inferred.
- Titrate prediction errors by extinguishing gradually.

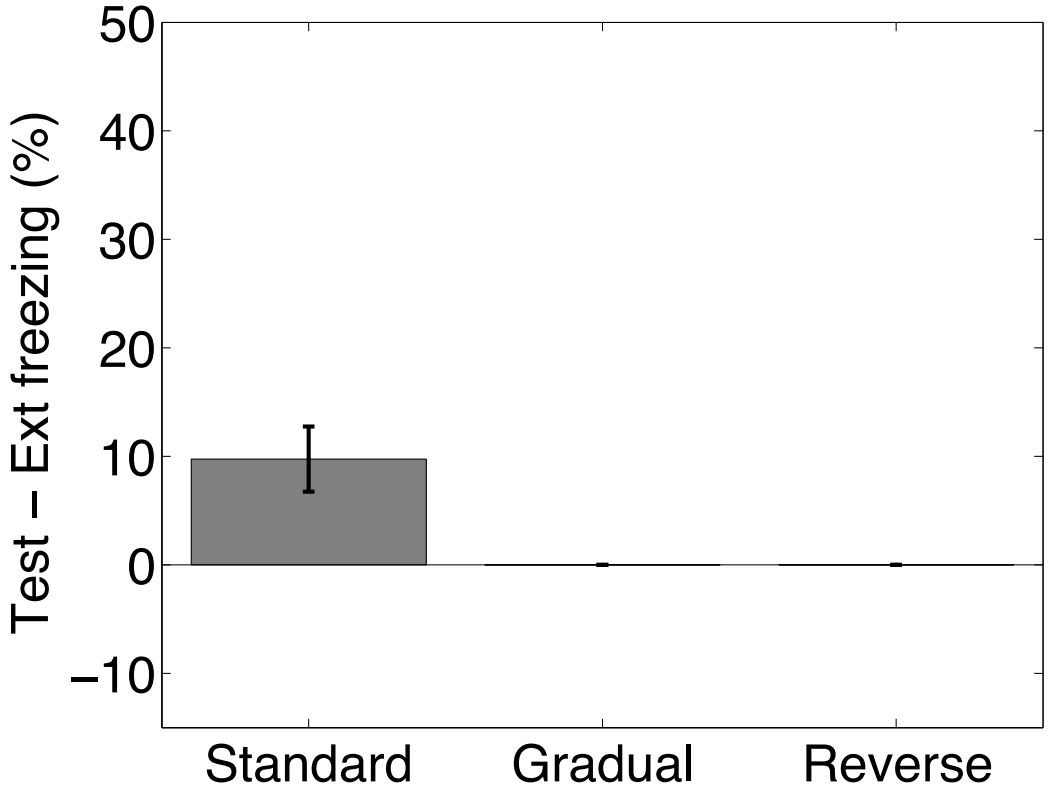
Testing the model: gradual extinction



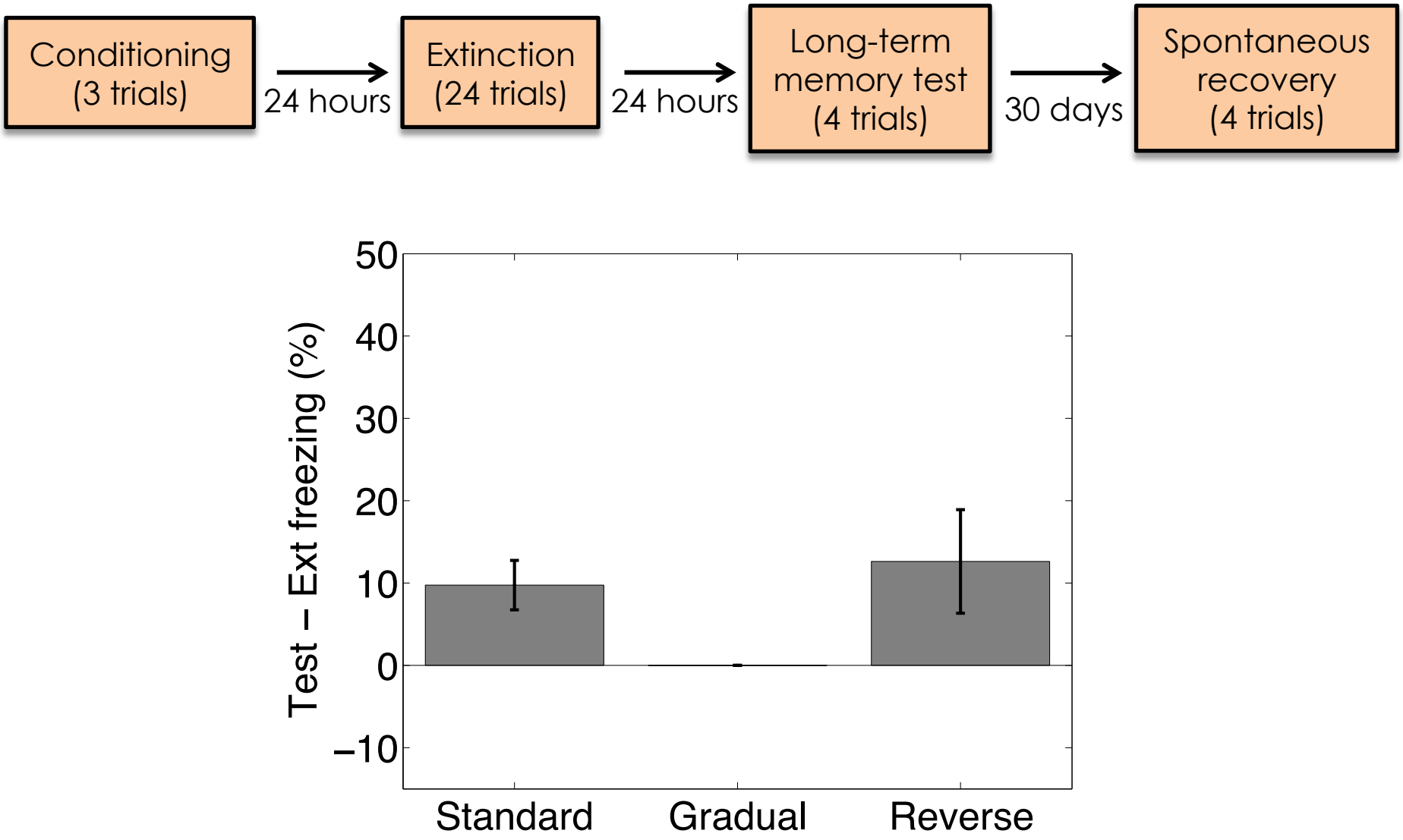


Gershman, Jones, Norman, Monfils & Niv (2013)

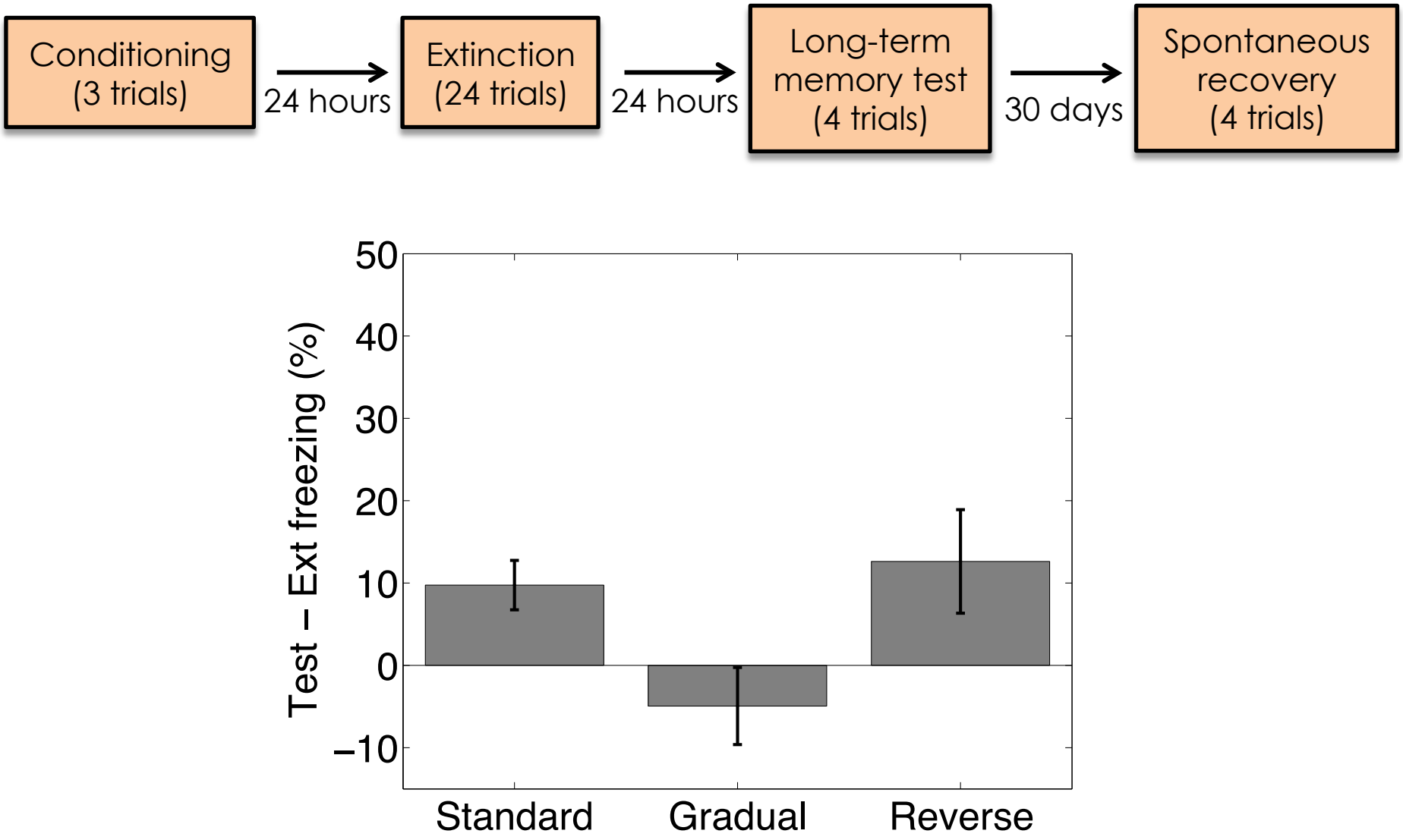




Gershman, Jones, Norman, Monfils & Niv (2013)

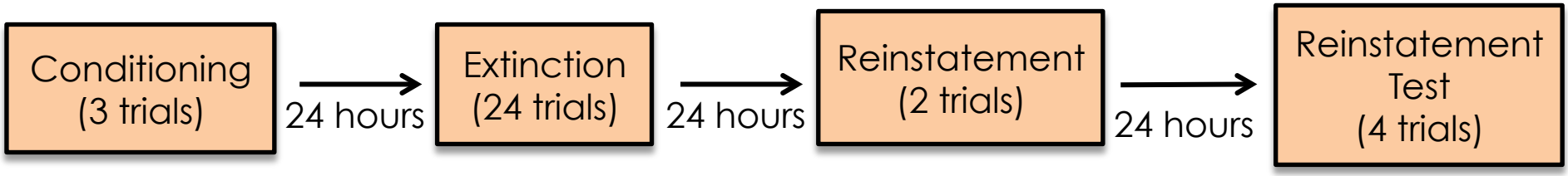


Gershman, Jones, Norman, Monfils & Niv (2013)



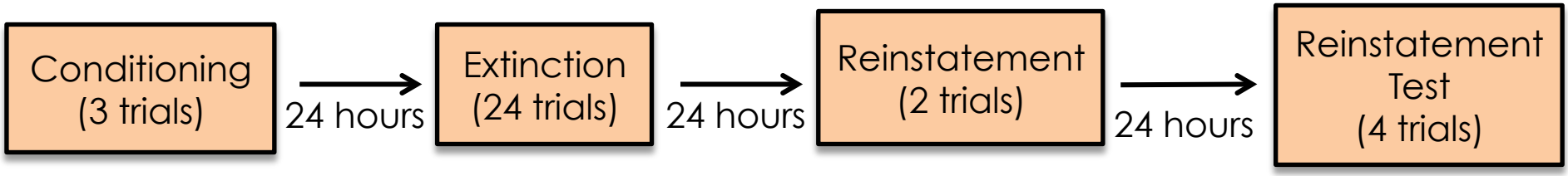
Gershman, Jones, Norman, Monfils & Niv (2013)

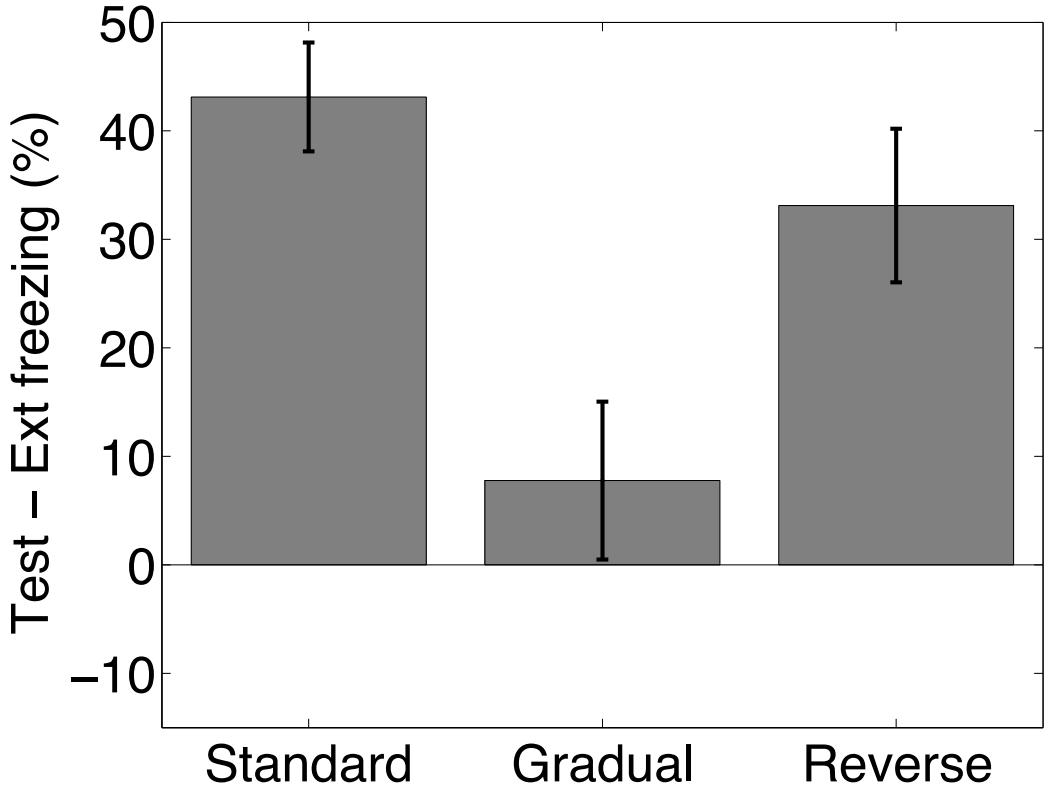
Reinstatement design



Gershman, Jones, Norman, Monfils & Niv (2013)

Reinstatement design

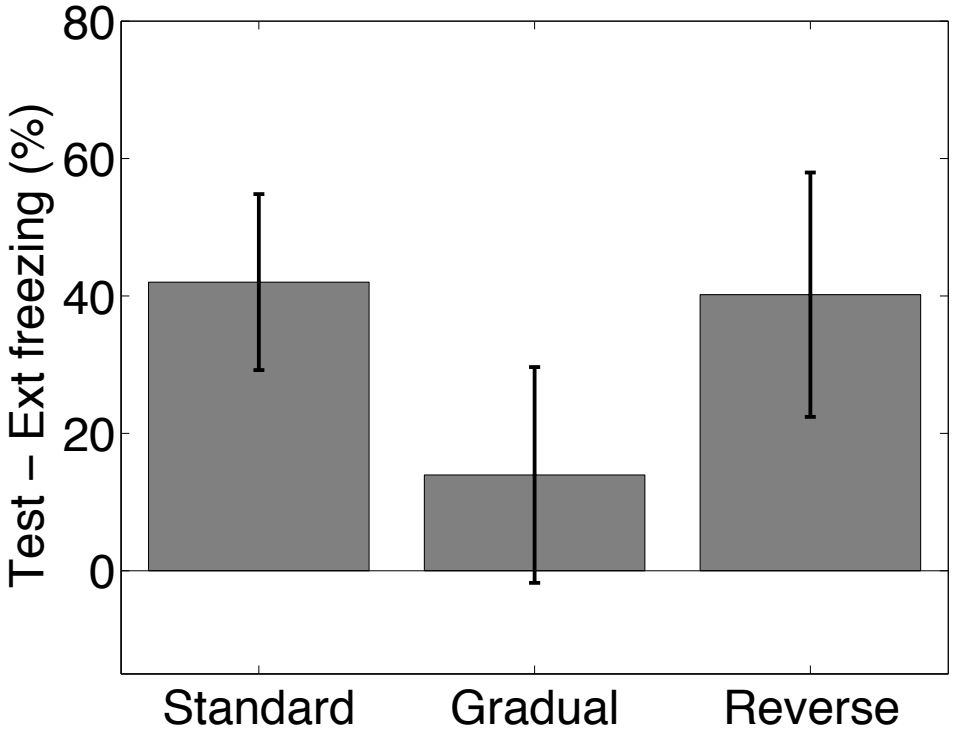




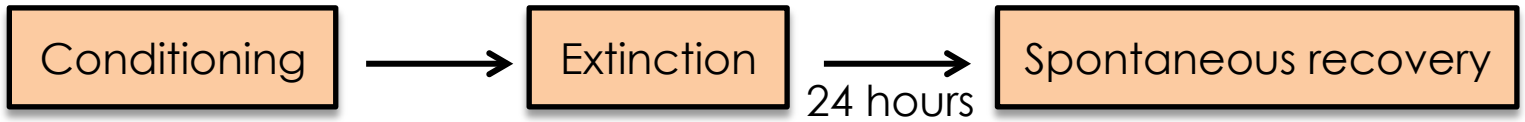
Gershman, Jones, Norman, Monfils & Niv (2013)

Gradual extinction in humans

Gradual extinction in humans



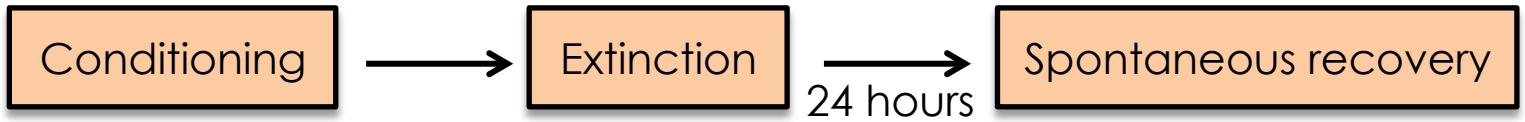
Predicting spontaneous recovery in humans



(partial reinforcement)

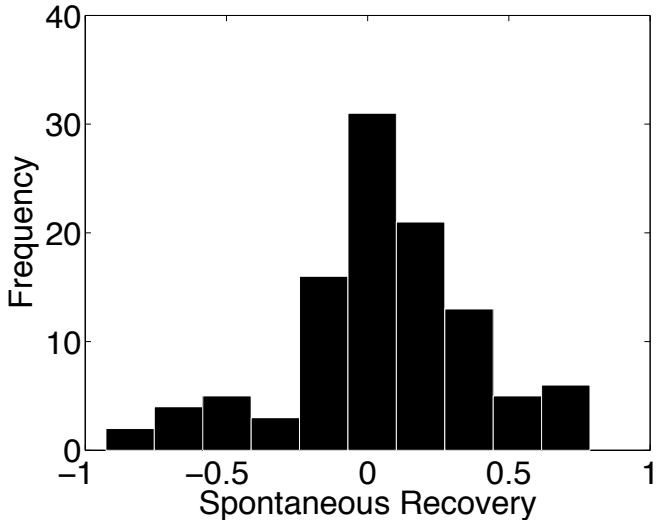
Gershman & Hartley (submitted)

Predicting spontaneous recovery in humans



(partial reinforcement)

Why do some people show a return of fear, and some don't?

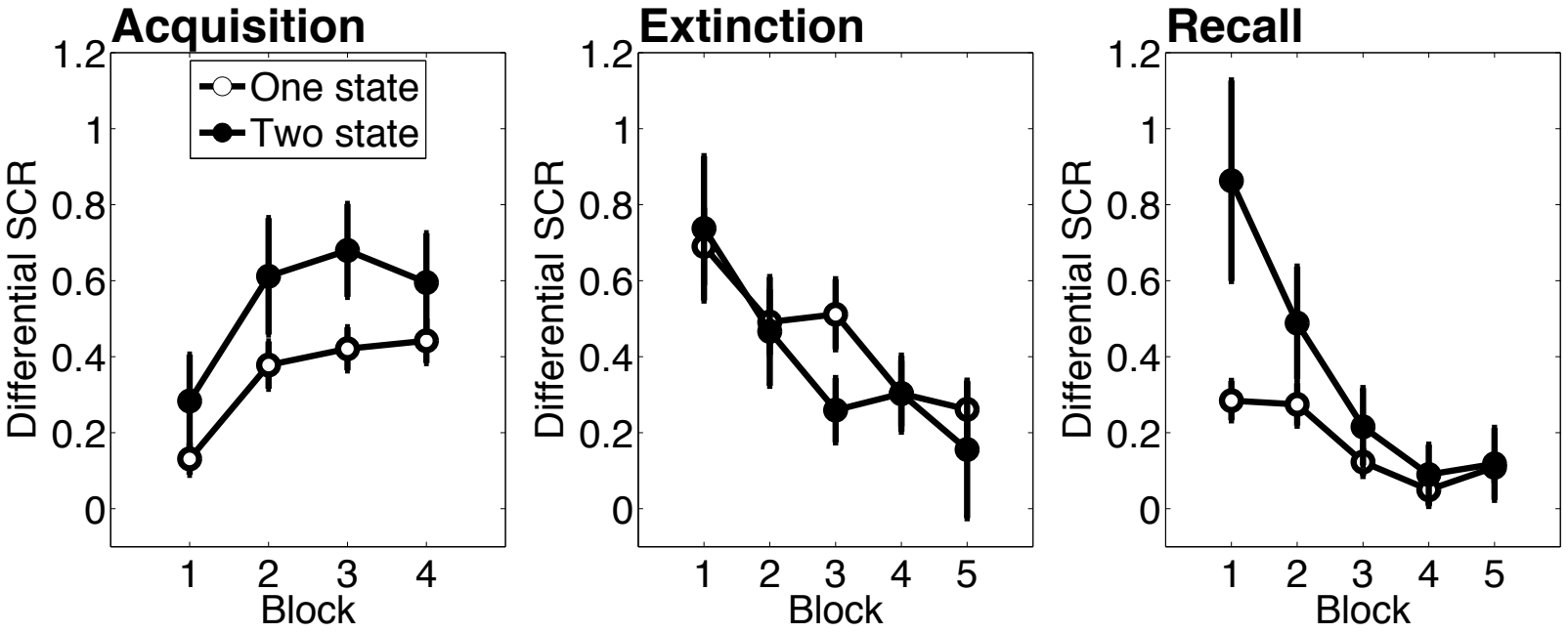


Gershman & Hartley (submitted)

Predicting spontaneous recovery in humans

Conditioning → Extinction → (24 hours) → Spontaneous recovery

(partial reinforcement)

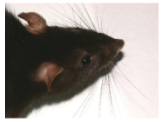


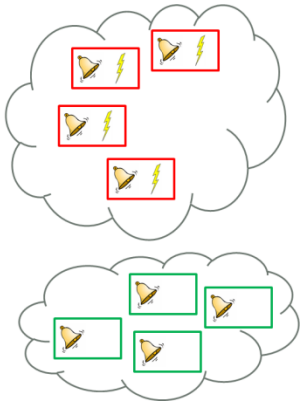
Model, fit only to conditioning & extinction data, divides subjects into two groups

Gershman & Hartley (submitted)

Summary

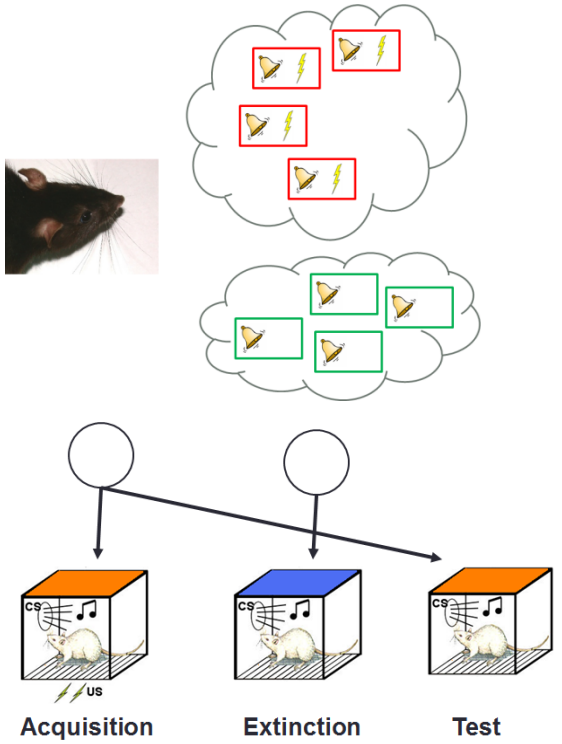
- Conditioning as clustering





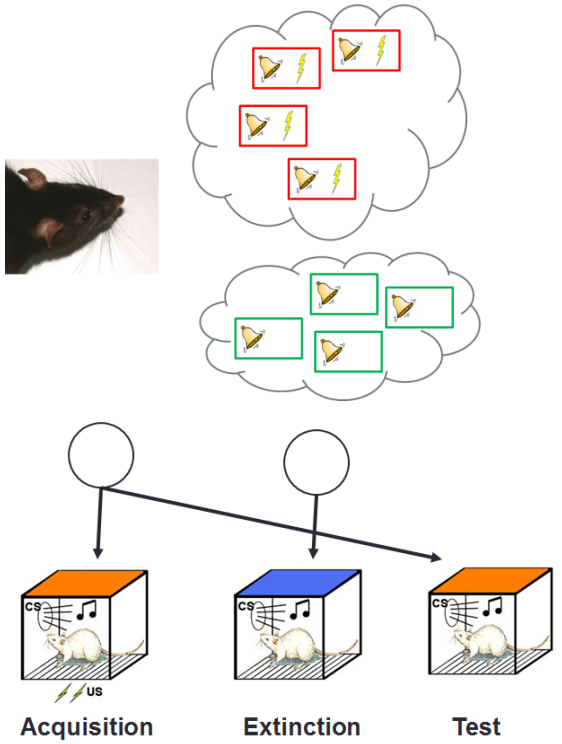
Summary

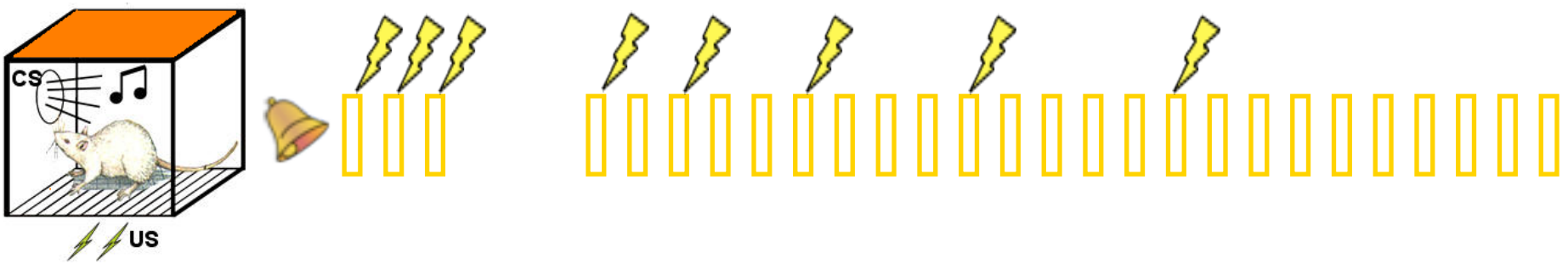
- Conditioning as clustering
- Memories reflect inferences about latent causes



Summary

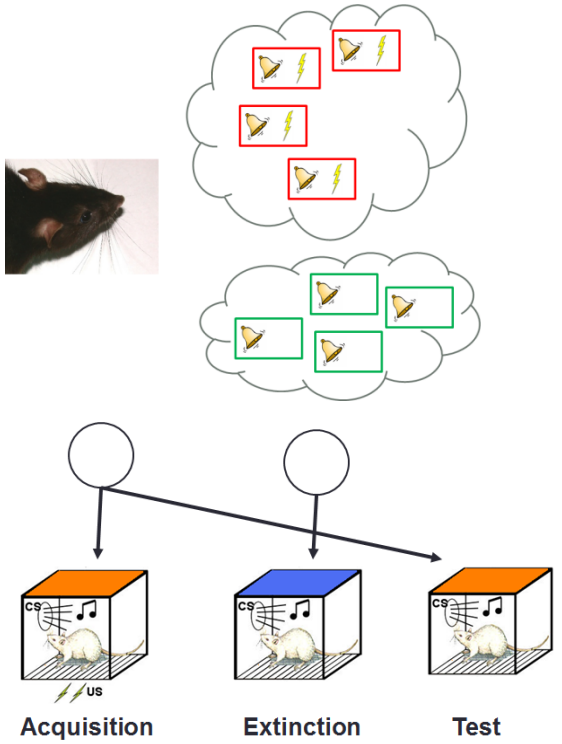
- Conditioning as clustering
- Memories reflect inferences about latent causes
- Gradual extinction prevents the return of fear

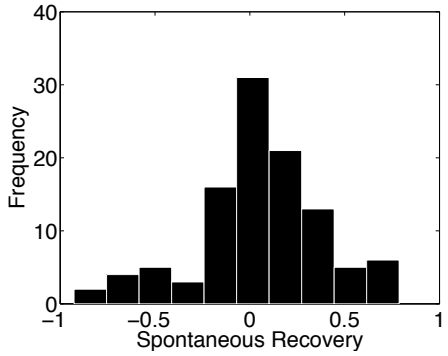




Summary

- Conditioning as clustering
- Memories reflect inferences about latent causes
- Gradual extinction prevents the return of fear
- Explaining individual differences in the return of fear





Big picture

- When do we modify old memories, and when do we create new ones?
- This question can be answered within a probabilistic computational framework: *we create new memories when we infer new latent causes in our environment*
- This principle has deep explanatory power across multiple domains

Acknowledgments

Ken Norman (Princeton)

Yael Niv (Princeton)

David Blei (Princeton)

Marie Monfils (UT Austin)

Carolyn Jones (UT Austin)
Cate Hartley (Sackler Institute)
Liz Phelps (NYU)