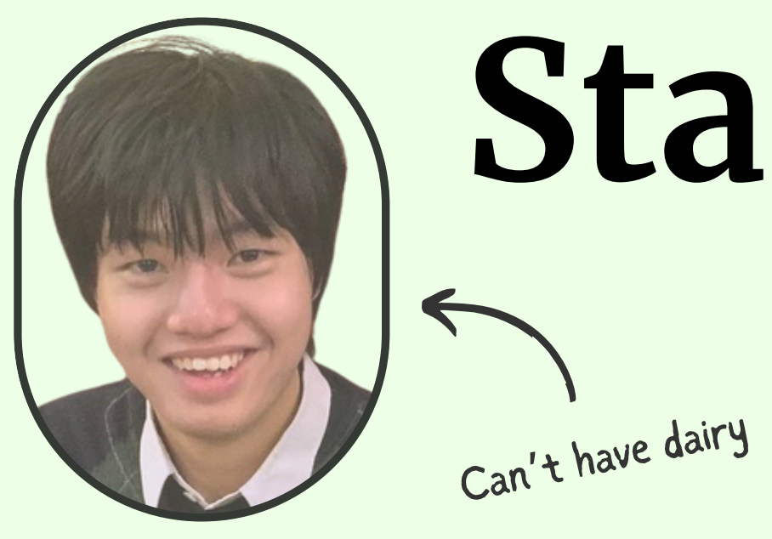
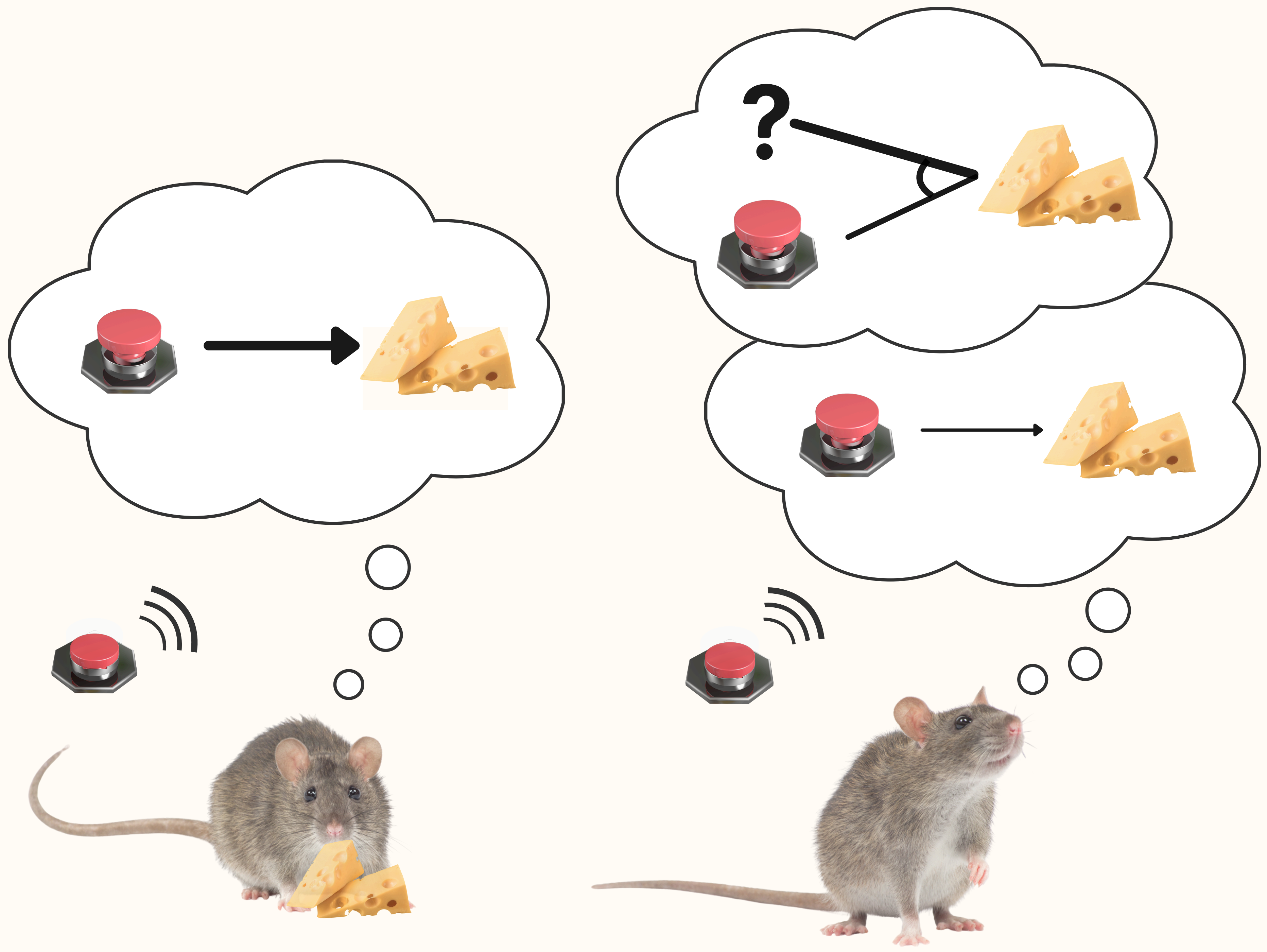


Stacked DAGs for Sequential and Hierarchical Learning

Dominic Le | dominic.le@mail.utoronto.ca | dominicle.net
University of Toronto

Main Question

In extinction, how do previously learned associations get *rewritten*?



Background

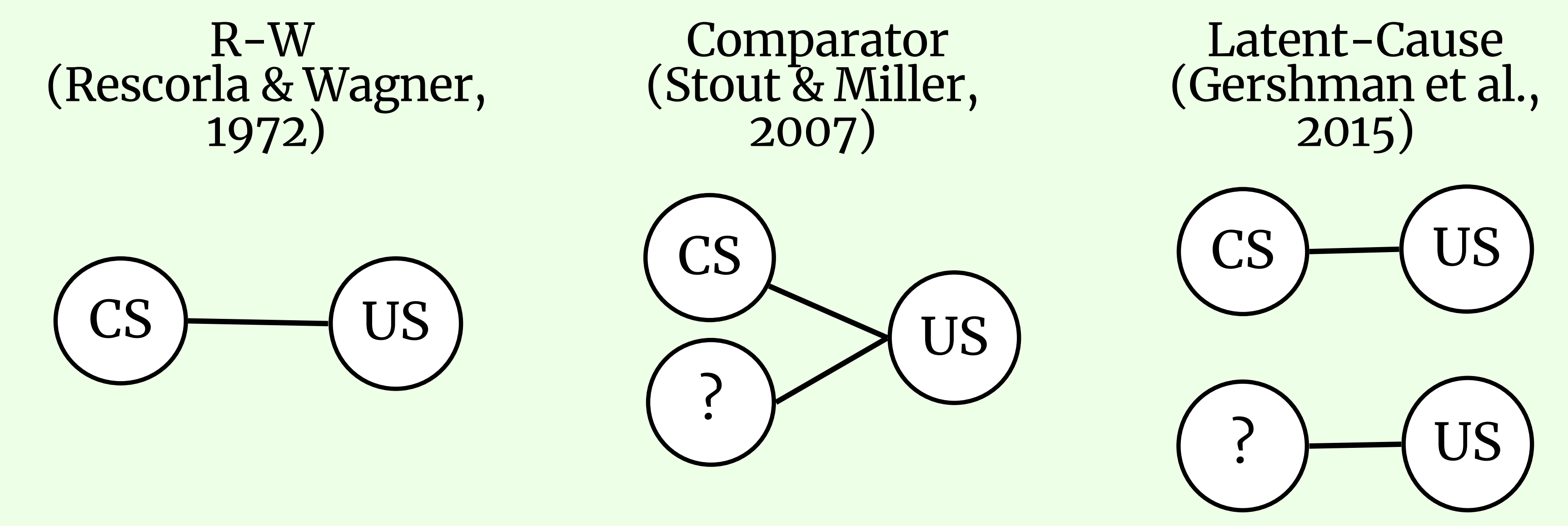
'Unlearning' [1]

Extinction is unlearning the association from the unconditioned stimuli (US) to the conditioned stimuli (CS) (as in classical conditioning and exposure therapy studies).

Spontaneous recovery is when the extinct US-CS association re-emerges (i.e. relapse).

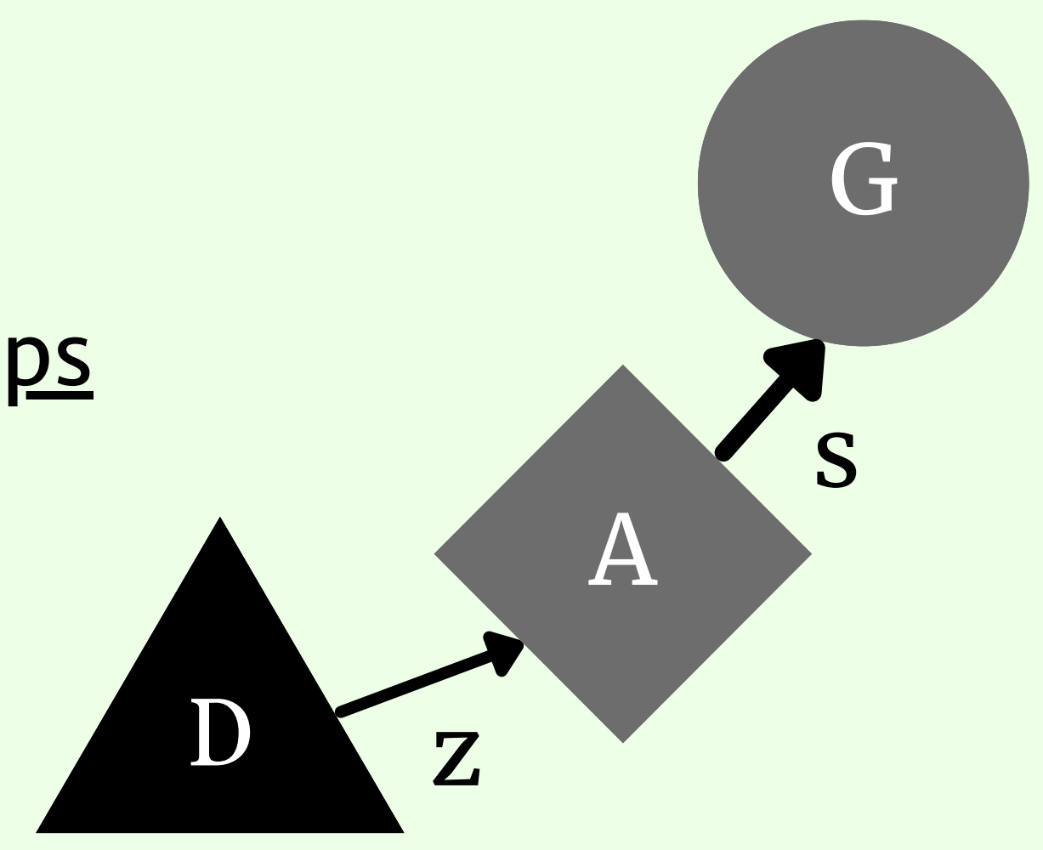
However, there are many ways to induce extinction (e.g., CS-alone, partial, context manipulations, deepened extinction).

Past Modeling Approaches



Directed Acyclic Graphs (DAGs) [2]

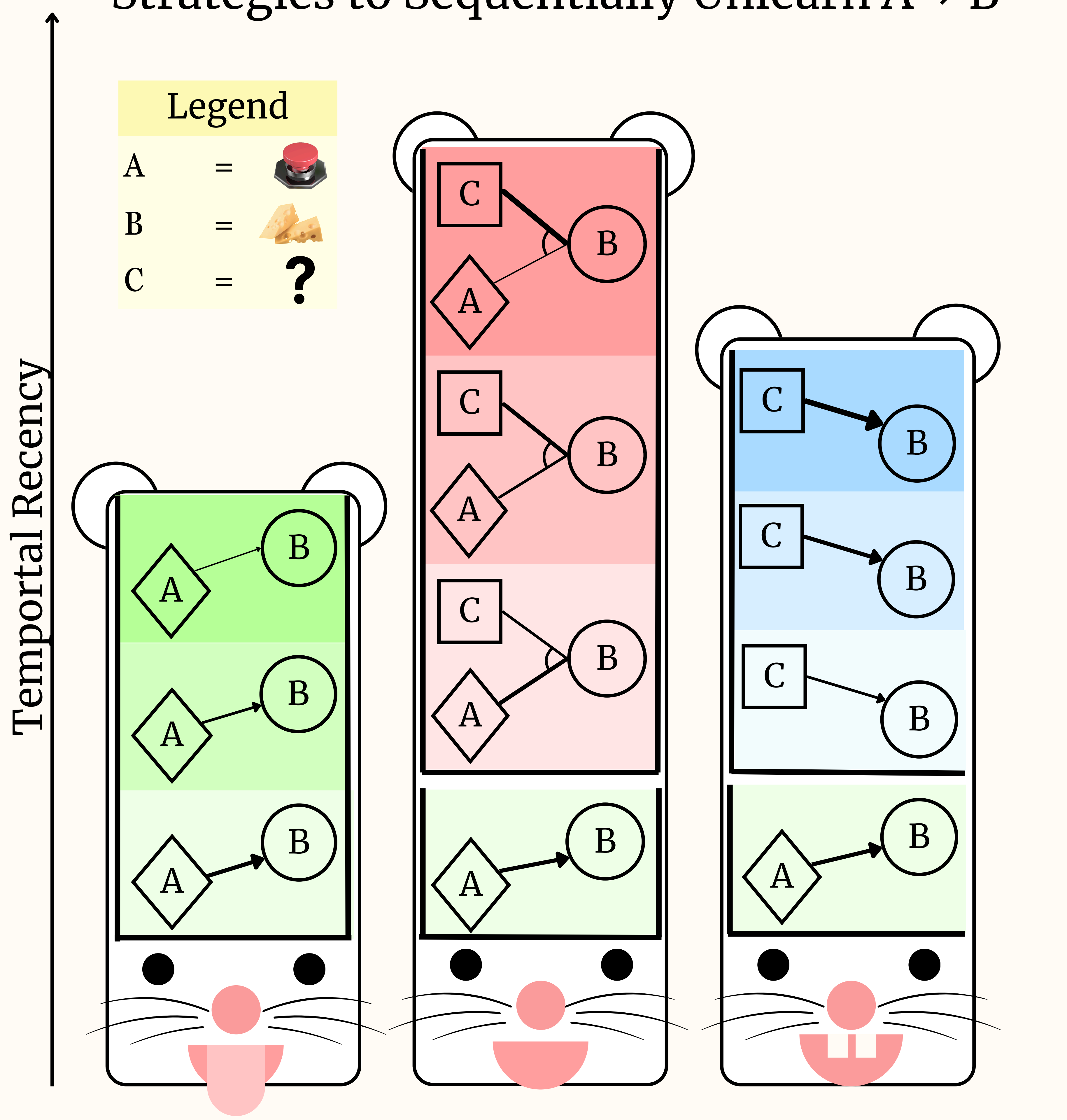
Effectively represents conditional relationships – essential for many cause-and-effect paradigms (e.g. models of causation, counterfactuals, neural activations)



Stacks

Learning forms an initial model A. If new evidence conflicts with A, the system either updates A's parameters or stacks a new structure B on top.

Strategies to Sequentially Unlearn A → B



Within-Stack: CS-alone

Between-Stack: Competition

Between-Stack: Latent-Cause

Implications

1. One structure at a time.
   - Learners switch models serially as contexts or expectations change.
2. Ranked stack.
   - The first-learned model anchors a queue of stored structures.
3. Preserved and accessible.
   - Older models reactivate based on recency and relevance.
4. No overwriting.
   - Conflicting input builds new structures.

Future Work

1. Merging structures (e.g. through similarity).
2. Backtracking conditions (e.g. when to quit branching).
3. Test/simulate model (e.g., context-switching, access delays).

Takeaway

Stacked DAGs may be a flexible framework to model dynamic and iterative structural learning unifying mechanisms from compatible theories.

1. Gottlieb, D. A. (2012). Pavlovian Conditioning. In Encyclopedia of the Sciences of Learning (pp. 2563–2567). Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-1428-6_1041
2. Danks, D. (2014). Unifying the mind: Cognitive representations as graphical models. the MIT press.

$ Thank you to the Social Sciences and Humanities Research Council (SSHRC), the University of Toronto Faculty of Arts & Science, the Department of Philosophy for funding this research under the University of Toronto Excellence Award
☺ Special thanks to Sara Aronowitz for her mentorship and supervision