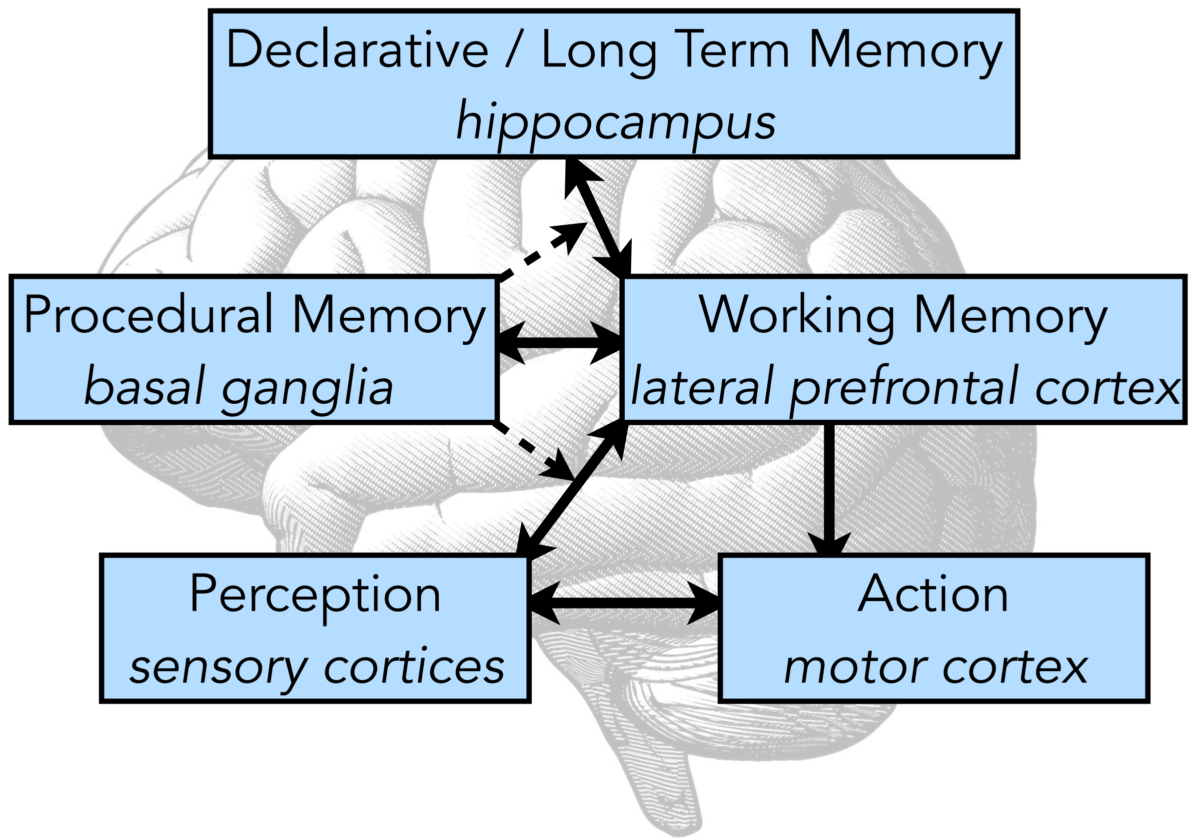
High Dimensional Vector Spaces as the Architecture of Cognition

Matthew A. Kelly - Penn State
Nipun Arora & Robert L. West - Carleton University
David Reitter - Penn State & Google Research

Common Model of Cognition

A ***blueprint*** for realizing a ***cognitive architecture*** and associated brain areas (Laird et al., 2017, Steine-Hanson, Koh, & Stocco, 2018; Stocco et al., 2018).



Thesis

The key parts of **cognitive architectures** - **declarative** and **procedural** memory - and their key abilities - **learning**, **memory retrieval**, **judgement**, and **decision-making** - can be realized as algebraic operations on vectors in a high-dimensional space.

Types of Models

- ***Deep learning*** has an impressive ability to process data to find *patterns*, but does not model *high-level cognition* and tends to be inscrutable.

- ***Symbolic architectures*** can capture the complexities of *high-level cognition* and provide *explainable* theories, but have limited ability to detect *patterns* or *learn*.

- **Vector-symbolic architectures**, where *symbols* are represented as *vectors*, bridge the gap between approaches.

Holographic Declarative Memory (HDM)

- based on the **BEAGLE** (Jones & Mewhort, 2007) and **DSHM** (Rutledge-Taylor et al., 2014)
- candidate for realizing **declarative memory** and aspects of **procedural memory**
- can be integrated with **deep-learning**
- can be integrated with **ACT-R**
- uses ***holographic reduced representations*** (Plate, 1995) to instantiate complex concepts in high-dimensional vectors.

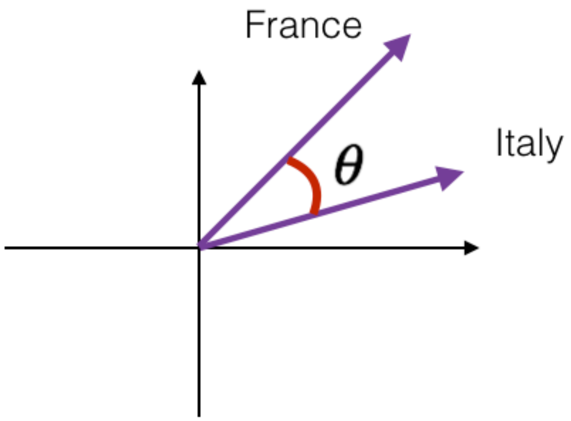
Overview

- How does **HDM** work?
- Recency, primacy, decay, and **free recall**
- **Fan effect** and interference
- Probability estimation and the **conjunction fallacy**
- Surprise and learning **iterated decision**

Vectors

- **e** ***environment vector***, represents an item, randomly generated.
- **m** ***memory vector***, constructed from environment vectors to encode associations between items in the environment.
- **q** ***cue vector***, constructed from environment vectors to encode a question asked of memory.

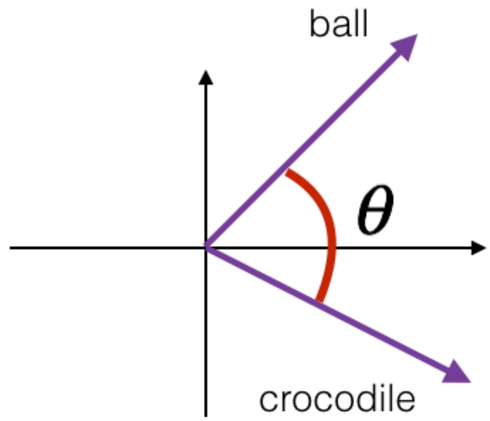
Similarity



France and Italy are quite similar

θ is close to 0°

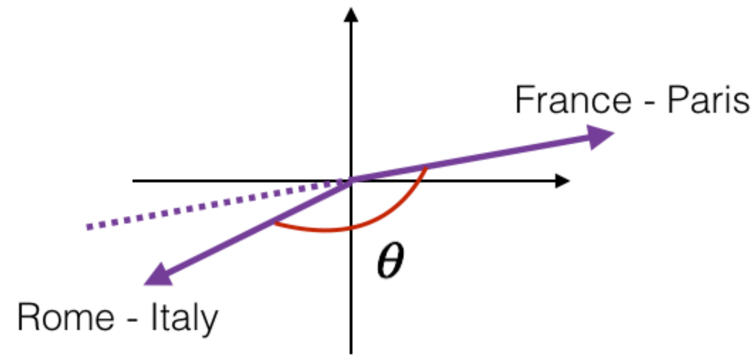
$\cos(\theta) \approx 1$



ball and crocodile are not similar

θ is close to 90°

$\cos(\theta) \approx 0$



the two vectors are similar but opposite
the first one encodes (city - country)
while the second one encodes (country - city)

θ is close to 180°

$\cos(\theta) \approx -1$

- Similarity is calculated as the cosine of a pair of vectors.
- Cosine = 1 if there's an angle of 0° and the vectors are **identical**
- Cosine = 0 if there's an angle of 90° and the vectors are **unrelated**

Activation (ACT-R DM)

- Activation in ACT-R is a sum of **base level activation** and **spreading activation**.

$$A_i = B_i + \sum_{j=1}^{n} W_j S_{ji}$$

- **Spreading activation** estimates the probability of the chunk conditional on the cue.

$$B_i = \ln(\sum_{j=1}^{n} t_j^{-d})$$

- **Base level activation** estimates the unconditional probability of the chunk, given the frequency and recency of its occurrence. Decays (*d*) over time (*t*).

Activation (HDM)

- Weight of old memories $\mathbf{m}_{i\text{-}1}$ is shrunk by $0 < \alpha < 1$ when a new memory $\mathbf{v}_i$ is added.

$$\mathbf{m}_i = \alpha \mathbf{m}_{i-1} + \mathbf{v}_i$$

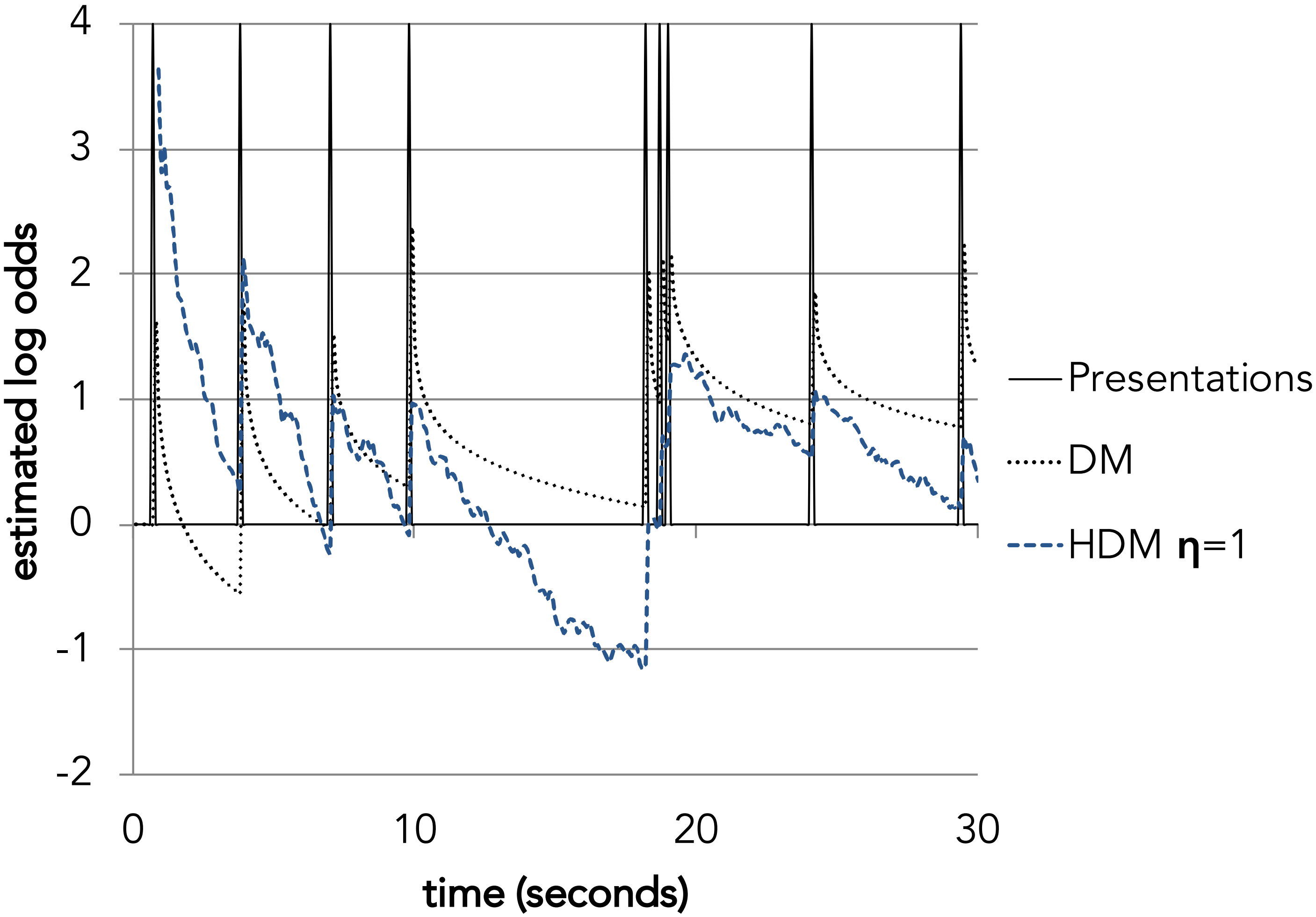
- Over time, an amount η of random noise $\mathbf{n}$ is added to memory $\mathbf{m}$.

$$\mathbf{m}_t = \mathbf{m}_{t-1} + \eta \mathbf{n}$$

- Cosine similarity C between vectors approximates root probability.
- Activation A estimates the log odds.

$$A = \ln\left(\frac{C^2}{1 - C^2}\right)$$

Activation



Activation of an item in memory over time in **ACT-R DM** and **HDM** as the item is repeatedly presented to the model.

Free Recall

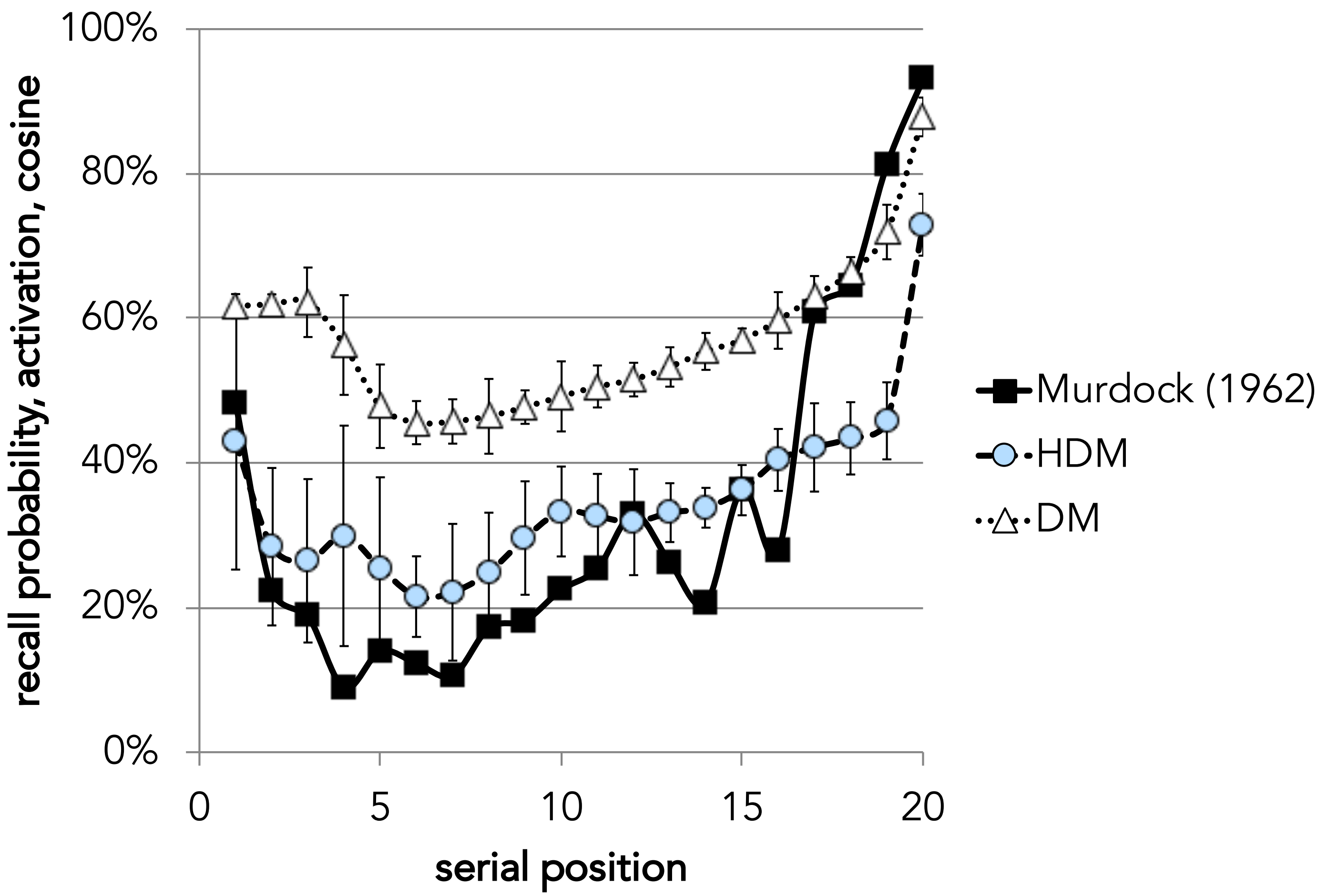
Participants & models presented 20 words at a rate of 1 word / 2 s.

Participants report back the list in any order.

Noise is added to HDM vectors over time.

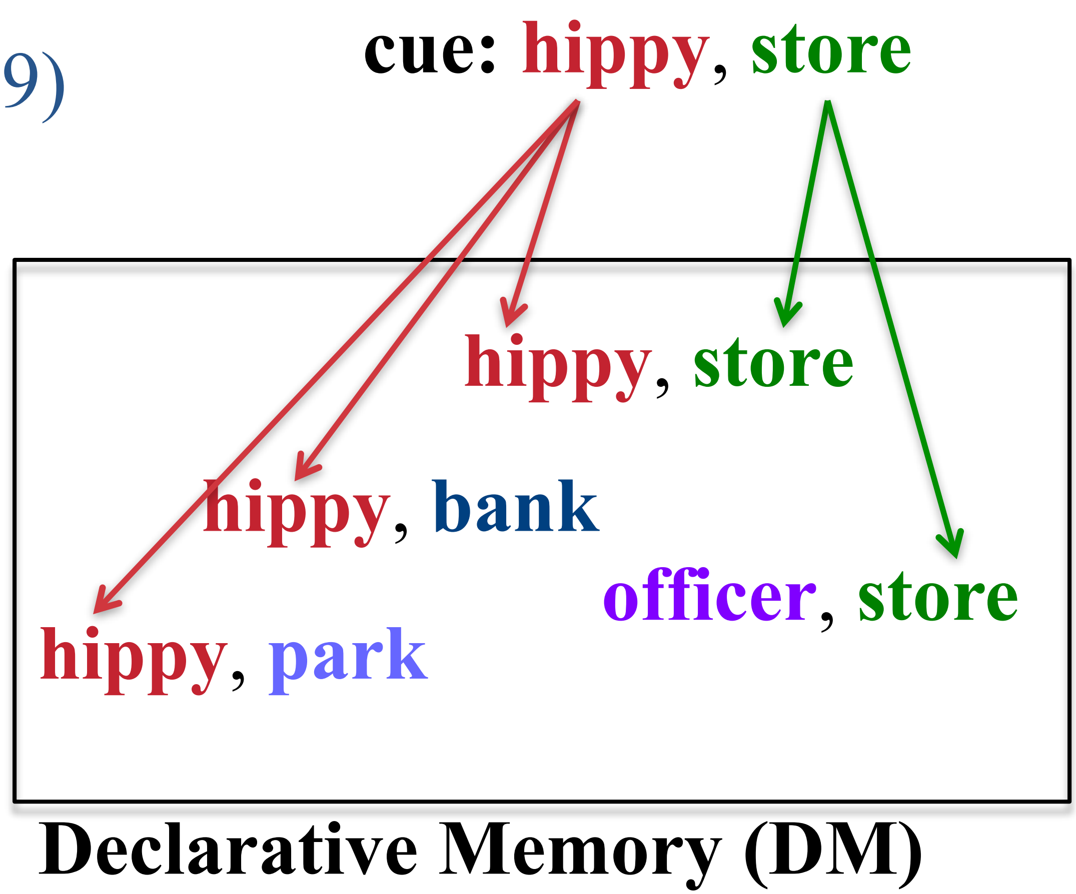
10 runs per model.

Human data from Murdock (1962).



Fan Effect

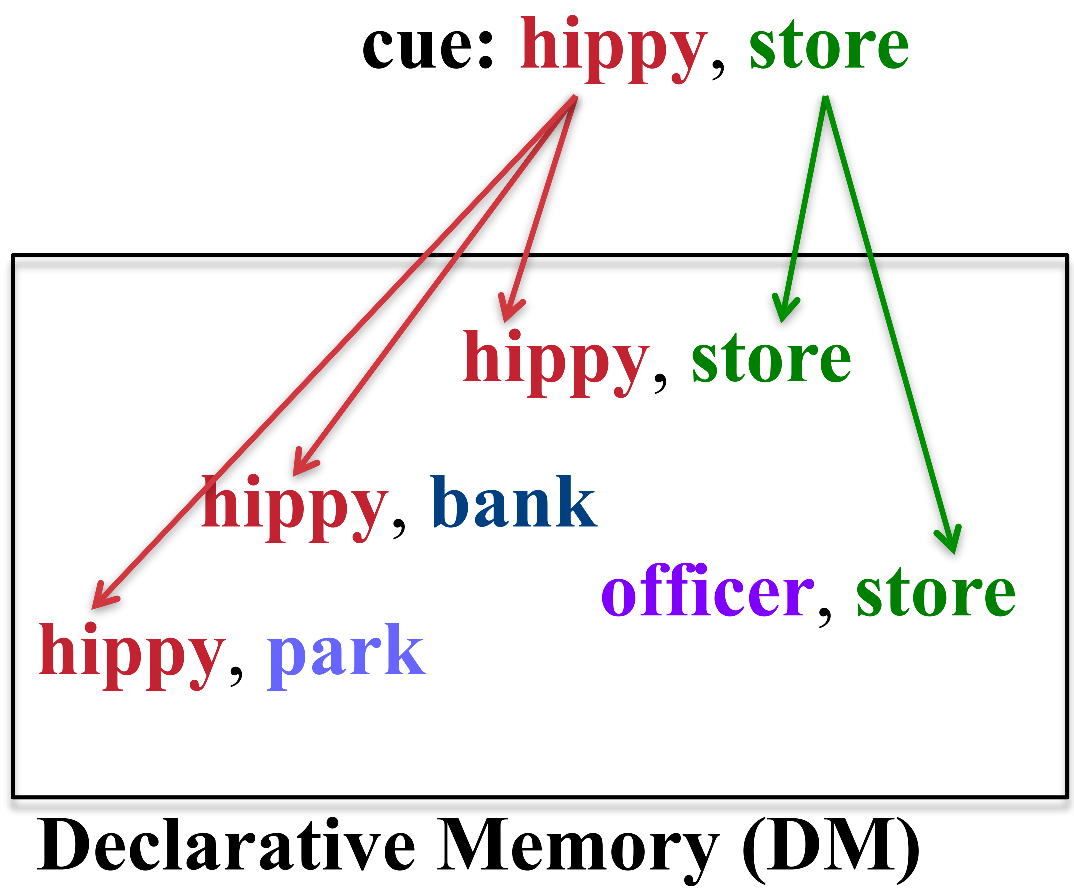
- Anderson & Reder (1999)
- **Study Set**: a list of object - location pairs (e.g., ***hippy in park***, ***officer in store***)
- **Test Set:** participants must determine which pairs were studied.



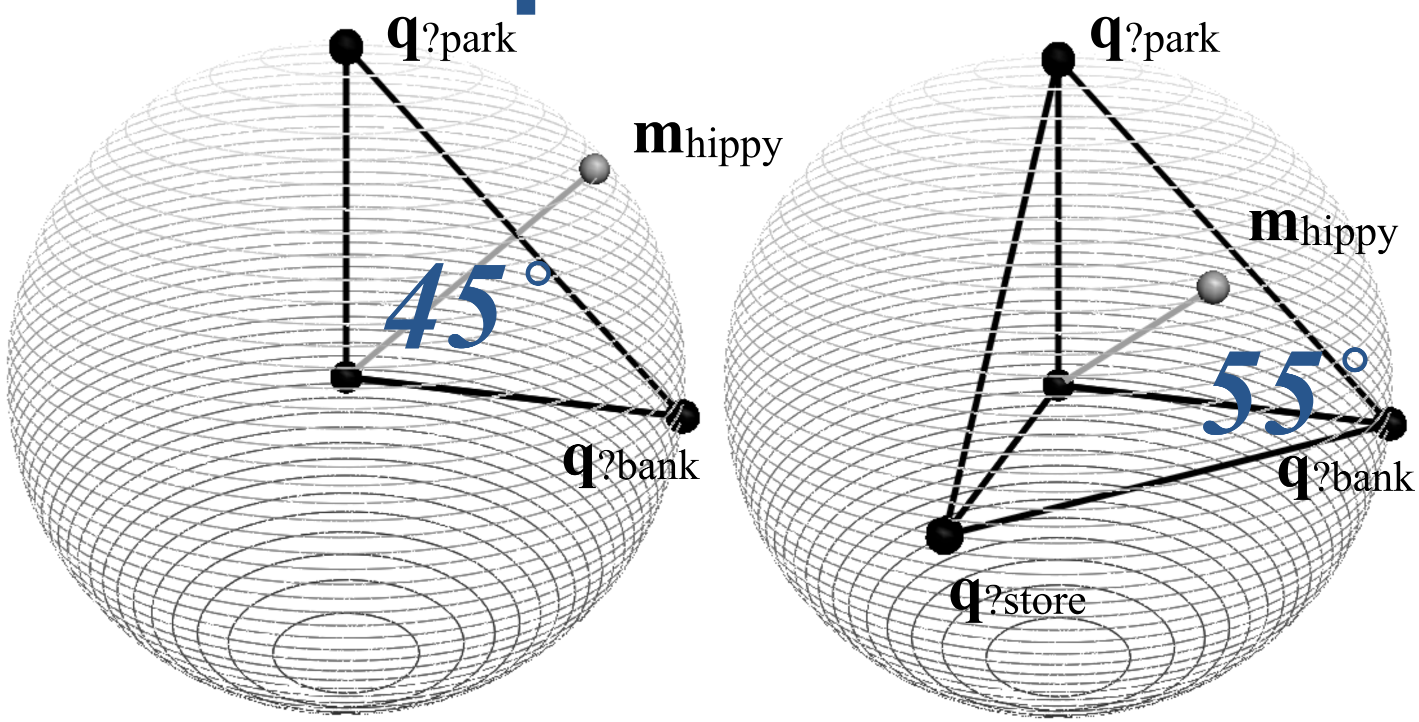
Fan Effect

- **Effect:** participants are ***slower*** to judge pairs that contain items that occur in more pairs in the study set
  (i.e., have a ***higher fan***).

- **Theory:** ***availability*** of an item in memory with respect to a cue is related to the item's ***probability conditional*** on the cue.



Vector Space Geometry



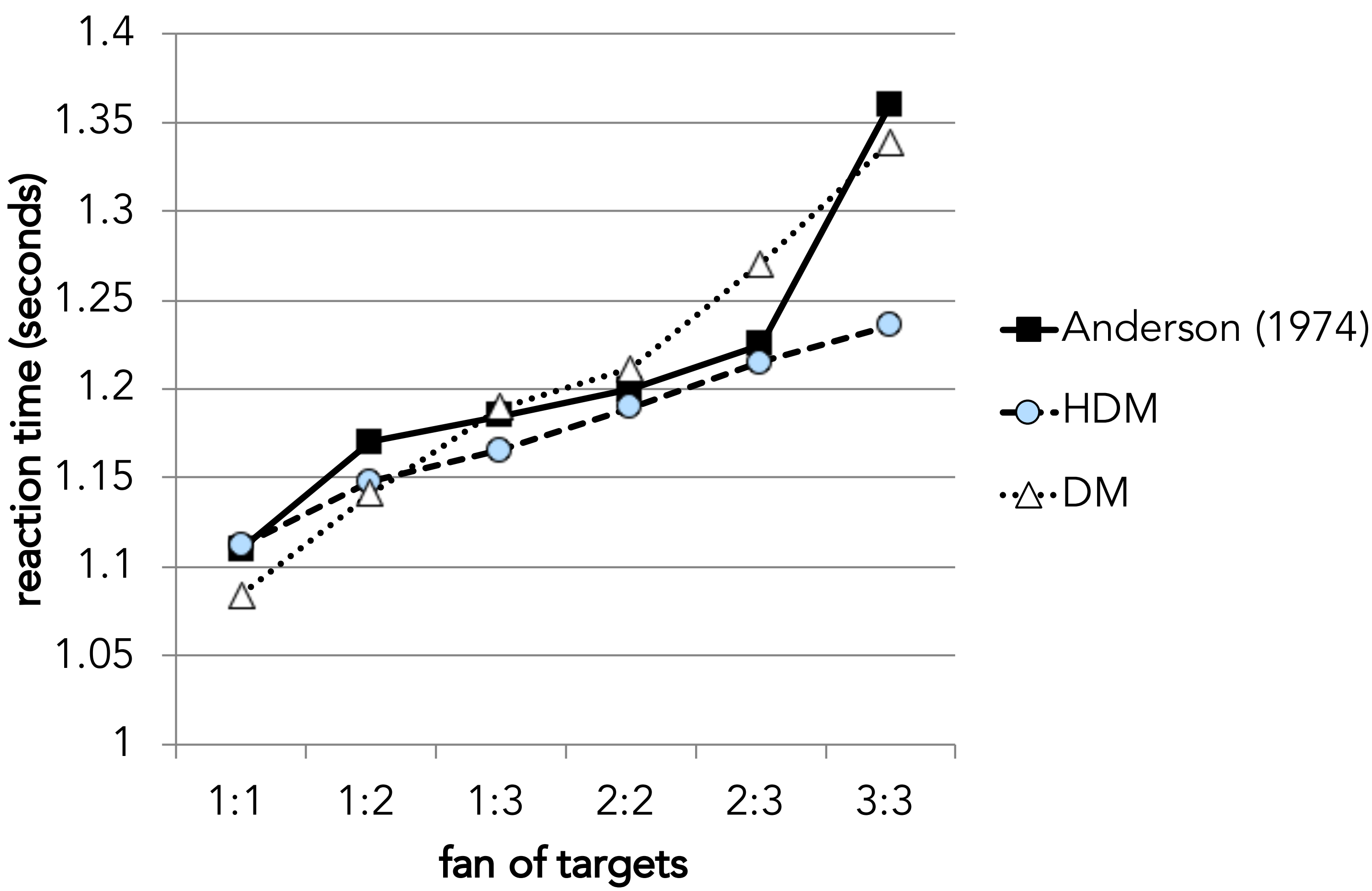
- **m**hippy with a fan of 2 (*left*) or 3 (*right*)
- *Cosine* is the **root** of the **fan**.
- *Cosine* is approximately the root **conditional probability** of an item in memory given the **cue**.

$$\text{cosine} = \frac{v_i}{\sqrt{v_1^2 + \ldots + v_i^2 + \ldots + v_n^2}}$$

Fan Effect

Response time for targets in the fan effect task (*left*).

As distance to **q** goes up, retrieval of **m** slows.

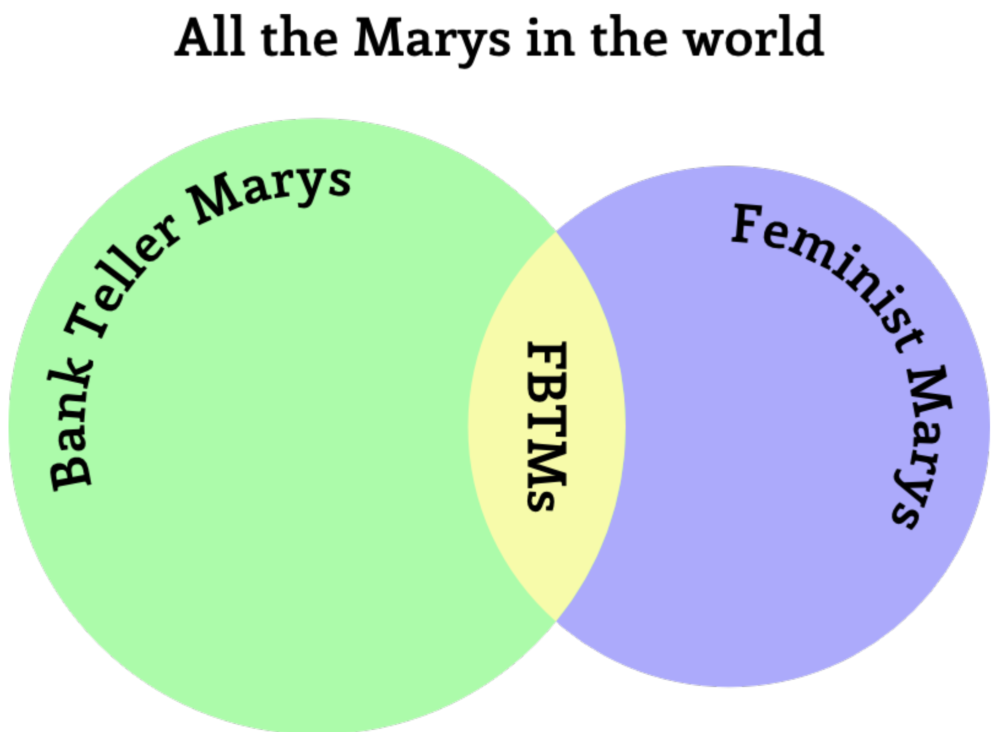


Conjunction Fallacy

People read a description of **Linda** and are asked to judge how likely she is to be:

- a **bank teller**,
- a **feminist**,
- **both** a bank teller ***and*** a feminist.

People say she's most likely **both** (Tversky & Kahneman, 1983), which is ***impossible***.



Conjunction Fallacy

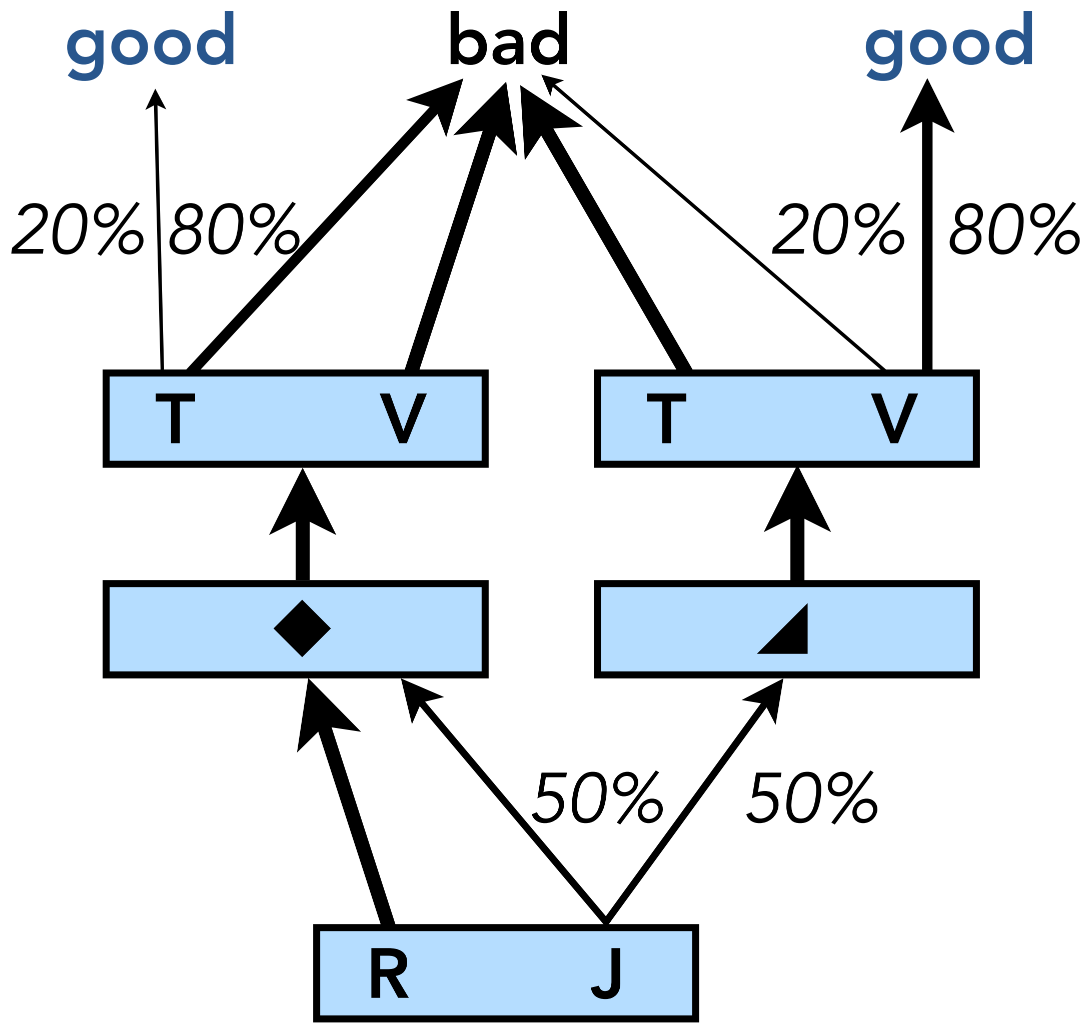
- We train HDM on a corpus of novels (Novels; 145 million words) or the British National Corpus (BNC; 100 million words).
- Is **Linda** (the sum of all words that describe Linda) more similar to **bank teller** (bank + teller) or **bank teller and feminist** (bank + teller + feminist)?
- For both models:

$$\text{cosine}(\mathbf{m}_{\text{Linda}}, \mathbf{m}_{\text{feminist}} + \mathbf{m}_{\text{bankteller}}) > \text{cosine}(\mathbf{m}_{\text{Linda}}, \mathbf{m}_{\text{bankteller}})$$

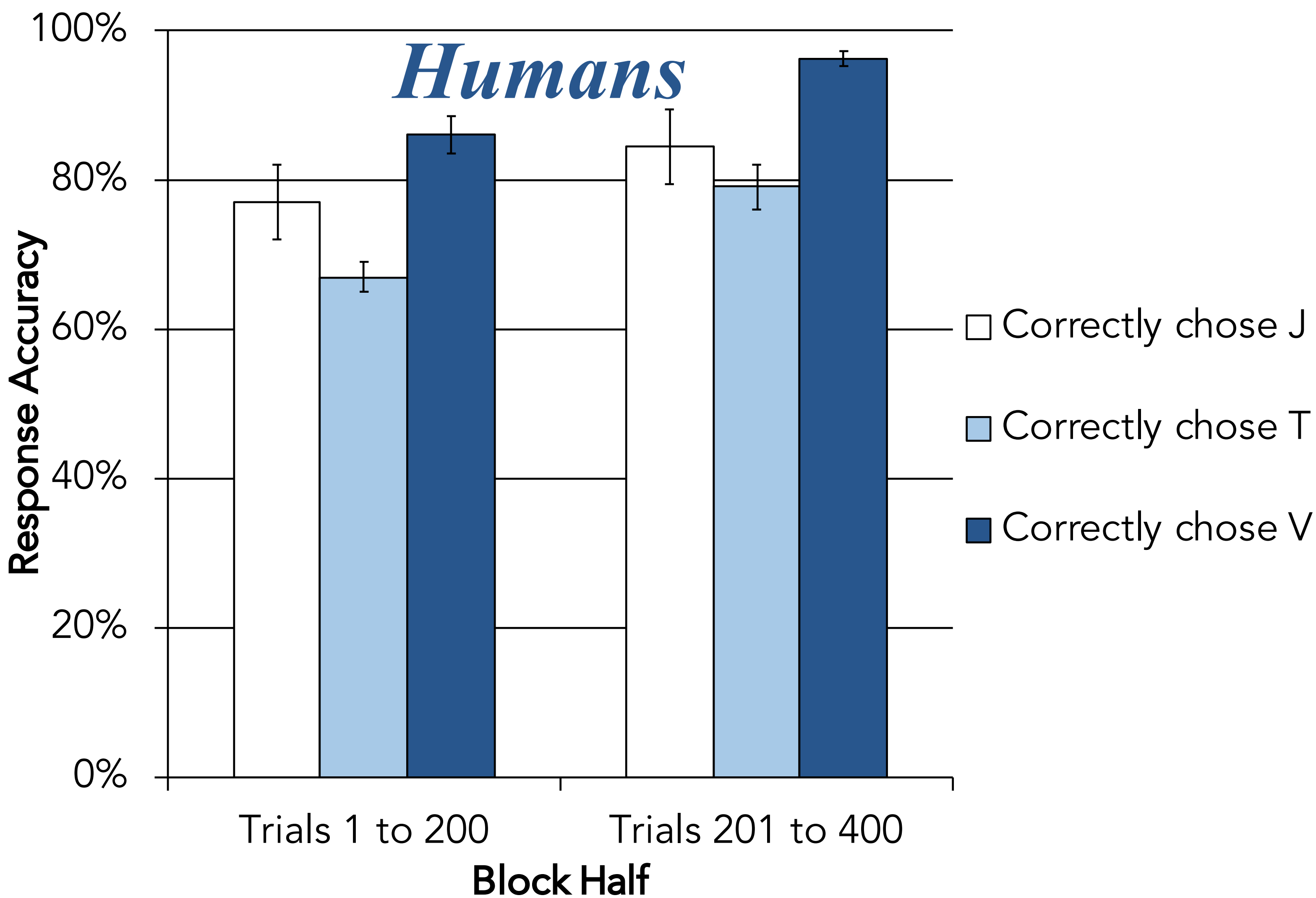
- **BNC**: 0.72 > 0.65, **Novels**: 0.68 > 0.62

Iterated Decision Task

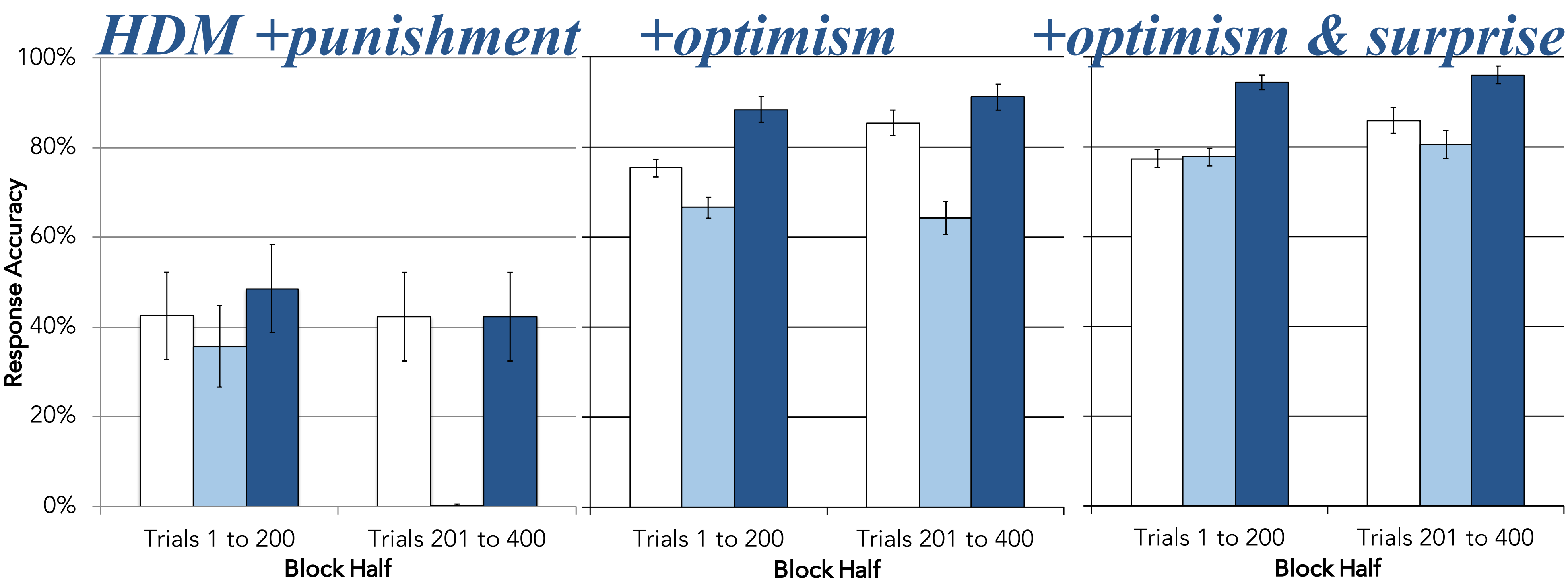
- Walsh and Anderson (2011)'s iterated binary decision task.
- Participants **learn** to make **choices** between arbitrary symbols that stochastically yield **positive feedback**.



Iterated Decision Task

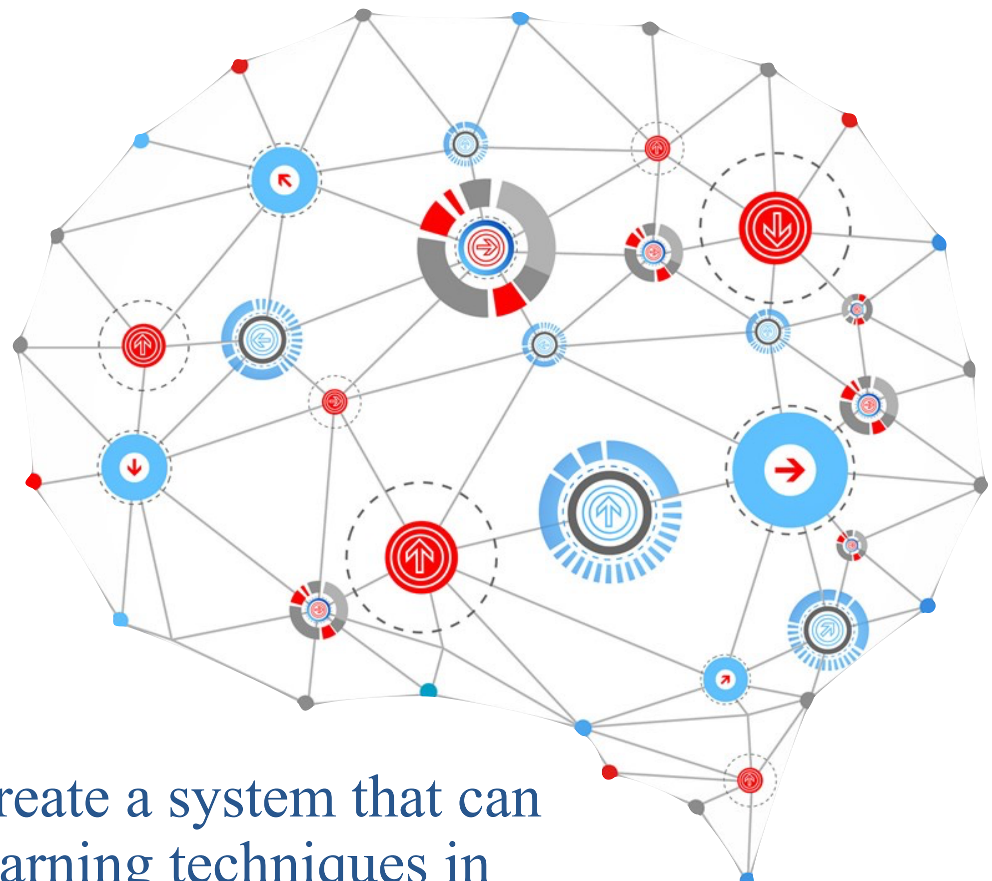


Iterated Decision Task



HDM learns to make the correct decisions when biased to explore the space using ***optimism*** and to self-correct using ***surprise***.

Conclusion



Our intent is to advance toward a **cognitive architecture** that is capable of modelling humans **at all scales of learning**, from the half-hour lab experiment to skills acquired over a **lifetime**.

By re-implementing ACT-R's **declarative memory** using **distributional semantics**, we create a system that can be integrated with modern machine learning techniques in **deep learning** while retaining ***long-term memory***, ***single-trial learning***, ***reasoning***, ***decision-making***, and other cognitive capacities associated with high-level cognition.

Thanks!