A mathematical theory of semantic development in deep neural networks

Andrew M. Saxe, James L. McClelland, and Surya Ganguli

Presented by:

Udith Haputhanthri(ge),
Electrical and Computer Engineering
Princeton University

Empirical Literature on semantic cognition

1. Acquisition of semantic knowledge
   a. Stage-like learning
      - Relative stasis followed by abrupt conceptual reorganization [6, 7]
   b. Hierarchical differentiation
      - Broader categorical distinctions are generally learned before finer-grained distinctions [1, 5]
   c. Illusory/ incorrect facts during the developmental stasis [2]
2. Organization of semantic knowledge
   a. Category membership is a graded quantity
   b. Item typicality is reproducible across individuals [8, 9]
      - And correlates with performance on diverse semantic tasks [10, 14]
   c. Coherent vs less-coherent categories
      - Coherent categories can be learned/ represented relatively easily [8, 15, 16]
3. Deployment of semantic knowledge
   a. Inductive generation (i.e. make decisions about novel items/ properties) [2,3]
   b. Inductive generalization systematically changes over time: becomes more specific with age [2,3,17-19]
4. Neural representations of semantic knowledge
   a. Similarity structure of neural population vectors in response to different stimuli
      - Example works: Inanimate objects are differentiated from animate objects [22, 23]
   b. Such neural similarity structure is preserved across humans and monkeys [24, 25]
   c. Correspondence between neural similarity patterns and behavioral similarity patterns [21]

Goal of the work: Reproduce those behaviors “in a mathematically rigorous way”

Existing attempts:

No “neural implementation” work out there,

a. That does a **mechanistic/ concrete analysis** to see if artificial neural networks also show these
(Though we know that neural networks can gradually extract semantic structures)

b. If they do, how/ why

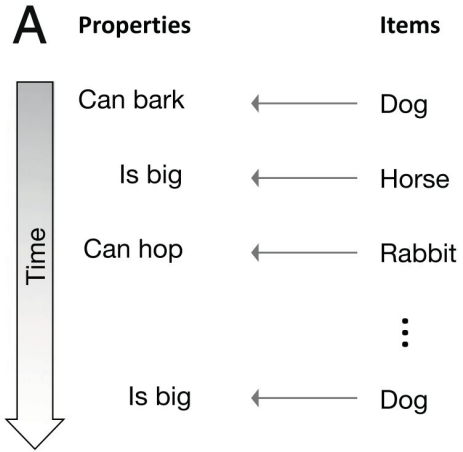
Goal of the work: Reproduce those behaviors “in a mathematically rigorous way”

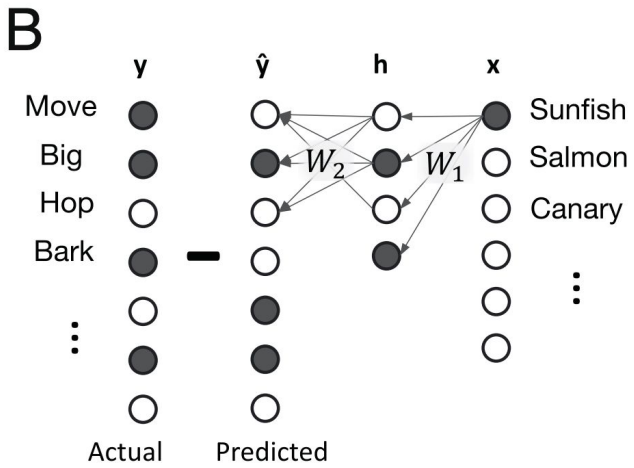
- This paper uses simple (i.e. mathematically tractable) artificial neural network model (i.e. deep linear neural network) and,

Gives exact analytical solutions describing the semantic development trajectory of,

1. Knowledge acquisition
2. Organization
3. Deployment
4. Neural representations

Method: Deep Linear Neural Networks





Developmental time (i.e. training)

- Network experience sequential episode (i.e. input and ground truth)

- Supervised learning

Linear network with 3 layers

- Inputs: x (think of it as a one-hot)
- Outputs: $\hat{y}$ (features of the item)
- Hidden features: h
- Weights: W_1, W_2

Method: Deep Linear Neural Networks

Network: $\hat{\mathbf{y}} = \mathbf{W}^2\mathbf{W}^1\mathbf{x}$

Loss : $SSE(\mathbf{W}^1, \mathbf{W}^2) = \frac{1}{2}\left\|\mathbf{y}^i - \hat{\mathbf{y}}^i\right\|^2$

Solving gradient descent for deep linear 3 layer network, for single sample :

$$\Delta\mathbf{W}^1 = \lambda{\mathbf{W}^2}^T\left(\mathbf{y}^i - \hat{\mathbf{y}}^i\right){\mathbf{x}^i}^T,\ \Delta\mathbf{W}^2 = \lambda\left(\mathbf{y}^i - \hat{\mathbf{y}}^i\right){\mathbf{h}^i}^T$$

Method: Deep Linear Neural Networks

$$\hat{\mathbf{y}} = \mathbf{W}^2\mathbf{W}^1\mathbf{x}$$

$$SSE(\mathbf{W}^1, \mathbf{W}^2) = \frac{1}{2}\left\|\mathbf{y}^i - \hat{\mathbf{y}}^i\right\|^2$$

Weights updates :

$$\Delta\mathbf{W}^1 = -\lambda\frac{\partial}{\partial\mathbf{W}^1}SSE(\mathbf{W}^1, \mathbf{W}^2)$$

$$\Delta\mathbf{W}^2 = -\lambda\frac{\partial}{\partial\mathbf{W}^2}SSE(\mathbf{W}^1, \mathbf{W}^2)$$

Solve the derivatives:

$$\Delta\mathbf{W}^1 = \lambda\mathbf{W}^{2\,T}\left(\mathbf{y}^i - \hat{\mathbf{y}}^i\right)\mathbf{x}^{iT}$$

$$\Delta\mathbf{W}^2 = \lambda\left(\mathbf{y}^i - \hat{\mathbf{y}}^i\right)\mathbf{h}^{iT}$$

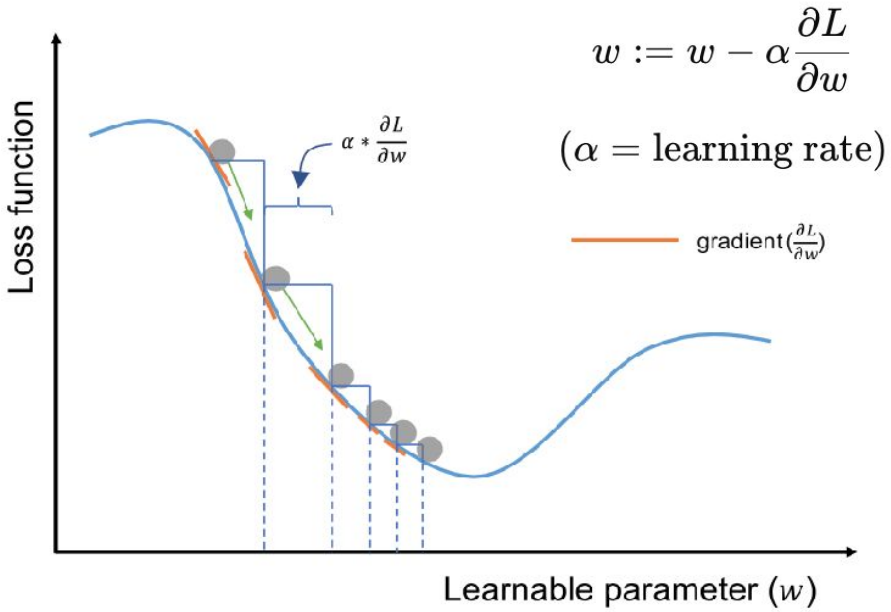


Image credit: Keshtkar, Kamyab. (2021).

Method: Deep Linear Neural Networks

Network: $\hat{\mathbf{y}} = \mathbf{W}^2\mathbf{W}^1\mathbf{x}$

Loss : $SSE(\mathbf{W}^1, \mathbf{W}^2) = \frac{1}{2}\left\|\mathbf{y}^i - \hat{\mathbf{y}}^i\right\|^2$

Solving gradient descent for deep linear 3 layer network, for single sample :

$$\Delta\mathbf{W}^1 = \lambda{\mathbf{W}^2}^T\left(\mathbf{y}^i - \hat{\mathbf{y}}^i\right){\mathbf{x}^i}^T,\ \Delta\mathbf{W}^2 = \lambda\left(\mathbf{y}^i - \hat{\mathbf{y}}^i\right){\mathbf{h}^i}^T$$

Method: Deep Linear Neural Networks

Network: $\hat{\mathbf{y}} = \mathbf{W}^2\mathbf{W}^1\mathbf{x}$

Loss : $SSE(\mathbf{W}^1, \mathbf{W}^2) = \frac{1}{2}\left\|\mathbf{y}^i - \hat{\mathbf{y}}^i\right\|^2$

Solving gradient descent for deep linear 3 layer network, for single sample :

$$\Delta\mathbf{W}^1 = \lambda {\mathbf{W}^2}^T\left(\mathbf{y}^i - \hat{\mathbf{y}}^i\right){\mathbf{x}^i}^T,\ \Delta\mathbf{W}^2 = \lambda\left(\mathbf{y}^i - \hat{\mathbf{y}}^i\right){\mathbf{h}^i}^T$$

When training for samples i ∈ {1,...,P}, average weight updates for the entire dataset (assuming small learning rate (λ)):

$$\Delta\mathbf{W}^1(t) = \lambda P\mathbf{W}^{2T}(\mathbf{\Sigma}^{yx} - \mathbf{W}^2\mathbf{W}^1\mathbf{\Sigma}^x)$$

$$\Delta\mathbf{W}^2(t) = \lambda P(\mathbf{\Sigma}^{yx} - \mathbf{W}^2\mathbf{W}^1\mathbf{\Sigma}^x)\mathbf{W}^{1T}$$

Method: Deep Linear Neural Networks

Network: $\hat{\mathbf{y}} = \mathbf{W}^2\mathbf{W}^1\mathbf{x}$

Loss : $SSE(\mathbf{W}^1, \mathbf{W}^2) = \frac{1}{2}\left\|\mathbf{y}^i - \hat{\mathbf{y}}^i\right\|^2$

Solving gradient descent for deep linear 3 layer network, for single sample :

$$\Delta\mathbf{W}^1 = \lambda {\mathbf{W}^2}^T\left(\mathbf{y}^i - \hat{\mathbf{y}}^i\right){\mathbf{x}^i}^T,\ \Delta\mathbf{W}^2 = \lambda\left(\mathbf{y}^i - \hat{\mathbf{y}}^i\right){\mathbf{h}^i}^T$$

When training for

Note: Highly complex learning dynamics in the linear model

1. Coupled nonlinear differential equations
2. Up to cubic weight interactions

es for the entire dataset (assuming small learning rate (λ)):

$(\mathbf{\Sigma}^{yx} - \mathbf{W}^2\mathbf{W}^1\mathbf{\Sigma}^x)$

$- \mathbf{W}^2\mathbf{W}^1\mathbf{\Sigma}^x)\mathbf{W}^{1T}$

When learning rat

limit

(where τ = 1/(Pλ))

$$\tau\frac{d}{dt}\mathbf{W}^1 = {\mathbf{W}^2}^T\left(\mathbf{\Sigma}^{yx} - \mathbf{W}^2\mathbf{W}^1\mathbf{\Sigma}^x\right)$$

$$\tau\frac{d}{dt}\mathbf{W}^2 = \left(\mathbf{\Sigma}^{yx} - \mathbf{W}^2\mathbf{W}^1\mathbf{\Sigma}^x\right){\mathbf{W}^1}^T$$

Method: Deep Linear Neural Networks

Goal: Analyze how W^1, W^2 evolves with the time
Problem: They are high-dimensional
Solution: Do some reduction using Singular Value Decomposition

Why SVD specifically?

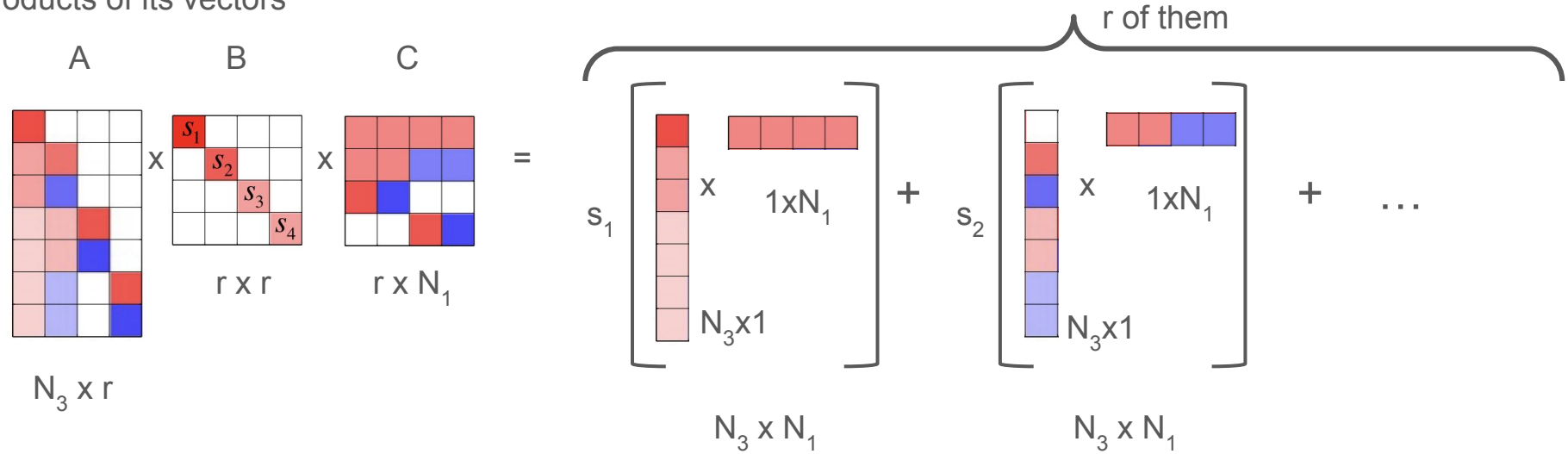
Method: Singular Value Decomposition

Let's consider at SVD of the dataset correlation matrix : $\mathbf{\Sigma}^{yx} = \mathbf{USV}^T = \sum_{\alpha=1}^{\min(N_1,N_3)} s_\alpha \mathbf{u}^\alpha \mathbf{v}^{\alpha\,T}$

Here,

U: a set of "left singular vectors" (shape: N3 x r)
V: a set of "right singular vectors" (shape: N1 x r)
S: diagonal matrix with singular values in the diagonal (shape: r x r)

Note: We can decompose any A x B x C matrix multiplication (with diagonal matrix B) into a sum of outer products of its vectors



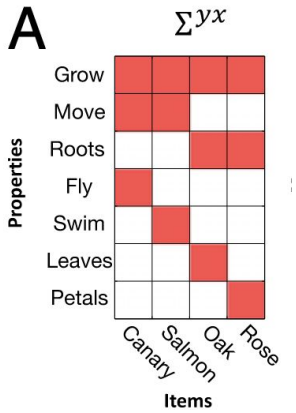
Method: Singular Value Decomposition

Let's consider at SVD of the dataset correlation matrix : $\mathbf{\Sigma}^{yx} = \mathbf{USV}^T = \sum_{\alpha=1}^{\min(N_1,N_3)} s_\alpha \mathbf{u}^\alpha \mathbf{v}^{\alpha T}$

Here,

U: a set of "left singular vectors" (shape: N3 x r)
V: a set of "right singular vectors" (shape: N1 x r)
S: diagonal matrix with singular values in the diagonal (shape: r x r)

How interpretable those U, V, S matrices **[super cool]**



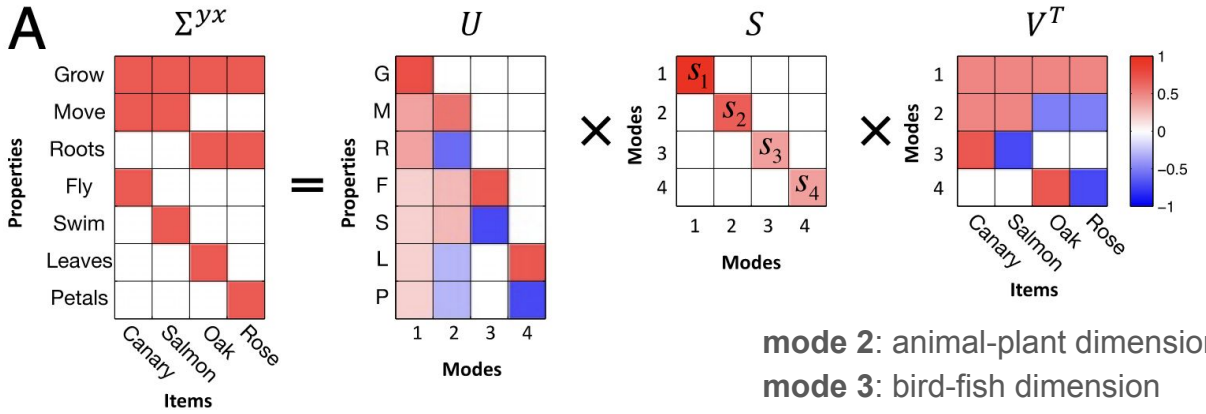
Method: Singular Value Decomposition

Let's consider at SVD of the dataset correlation matrix : $\mathbf{\Sigma}^{yx} = \mathbf{USV}^T = \sum_{\alpha=1}^{\min(N_1,N_3)} s_\alpha \mathbf{u}^\alpha \mathbf{v}^{\alpha T}$

Here,

U: a set of "left singular vectors" (shape: N3 x r)
V: a set of "right singular vectors" (shape: N1 x r)
S: diagonal matrix with singular values in the diagonal (shape: r x r)

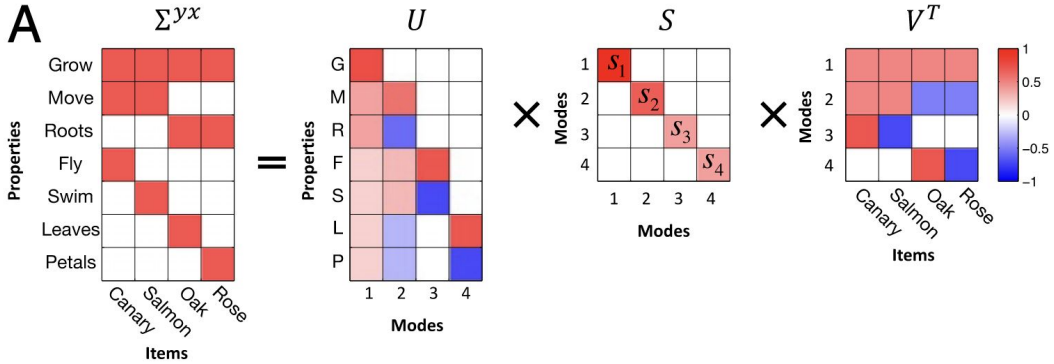
How interpretable those U, V, S matrices **[super cool]**



mode 2: animal-plant dimension
mode 3: bird-fish dimension
mode 4: flower-tree dimension

Note: Is this true for any dataset: **No**. Generally true for data generated via binary trees

Method: Singular Value Decomposition



mode 2: animal-plant dimension
mode 3: bird-fish dimension
mode 4: flower-tree dimension

We can intuitively interpret those U, S, V as follows:

1. α : a categorical distinction/ mode “embedded” in the statistical structure of the dataset
2. v^{α}_{i} (α-th column of V is v^{α}. i-th item is v^{α}_{i}) : how aligned the i-th item with the category-α
   - Let’s call v^{α} “object-analyzer” vectors
3. u^{α}_{m} (α-th column of U is u^{α}. m-th feature is u^{α}_{m}) : how important m-th feature for the category-α
   - Let’s call u^{α} “feature synthesizer” vectors
4. s^{α} : α-th singular value -> how much category-α explains the dataset

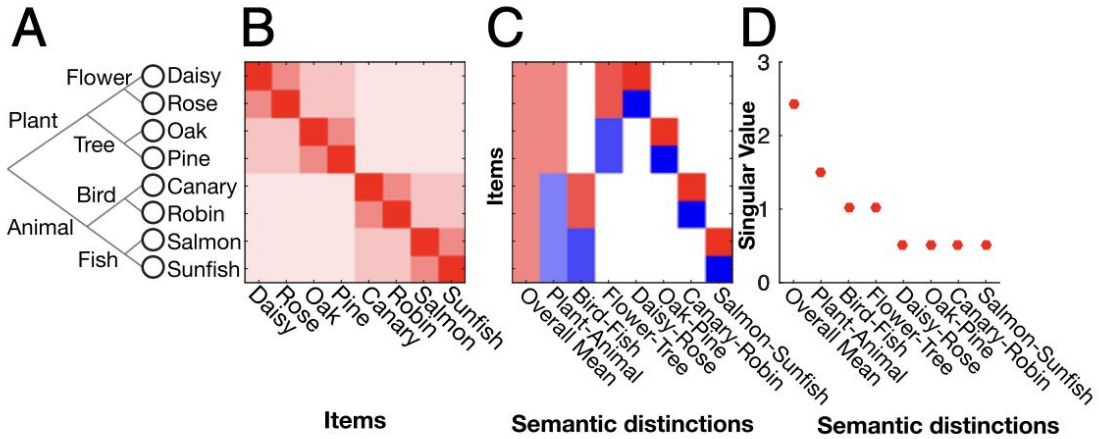
Summary:

1. SVD gives V, U matrices that act as bases for objects and features respectively
2. Let’s use them to understand how weight matrices evolve

Method: Singular Value Decomposition

More visualizations to show that: **SVD V matrix reveals semantic distinctions that mirrors the hierarchical taxonomy**

- This depends on how we generated the dataset:
  This case: dataset is created with a **branching diffusion process with an evolutionary dynamics**



Method: Deep Linear Neural Networks - decomposition of weight matrices

Lets fix U, V and decompose W^1, W^2: $\mathbf{W}^2(t)\mathbf{W}^1(t) = \mathbf{U}\mathbf{A}(t)\mathbf{V}^T$

Note:

1. This is “not” SVD of $W^2(t)W^2(t)$.
   - why - U, V are obtained from $\Sigma^{xy}=USV^T$. And they do not change with t.
2. So, A(t) doesn’t not have to be diagonal

However,

- under certain assumptions [small learning rate (λ <<1), near zero random weight initializations]
  - We can show that, **non-diagonal elements of A(t) decay to 0**
- So, we can approximate A(t) with a diagonal matrix (α-th diagonal component : $a_\alpha(t)$).

$$\mathbf{W}^2(t)\mathbf{W}^1(t) = \mathbf{U}\mathbf{A}(t)\mathbf{V}^T = \sum_{\alpha=1}^{N_2} a_\alpha(t)\,\mathbf{u}^\alpha \mathbf{v}^{\alpha\,T}$$

Note: this is super similar to the form of SVD we had: $\mathbf{\Sigma}^{yx} = \mathbf{U}\mathbf{S}\mathbf{V}^T = \sum_{\alpha=1}^{\min(N_1,N_3)} s_\alpha \mathbf{u}^\alpha \mathbf{v}^{\alpha\,T}$

Method: Deep Linear Neural Networks - decomposition of weight matrices

$$\mathbf{W}^2(t)\mathbf{W}^1(t) = \mathbf{U}\mathbf{A}(t)\mathbf{V}^T = \sum_{\alpha=1}^{N_2} a_\alpha(t)\,\mathbf{u}^\alpha \mathbf{v}^{\alpha\,T}$$

SVD of input-output correlations:

$$\mathbf{\Sigma}^{yx} = \mathbf{U}\mathbf{S}\mathbf{V}^T = \sum_{\alpha=1}^{\min(N_1,N_3)} s_\alpha \mathbf{u}^\alpha \mathbf{v}^{\alpha\,T}$$

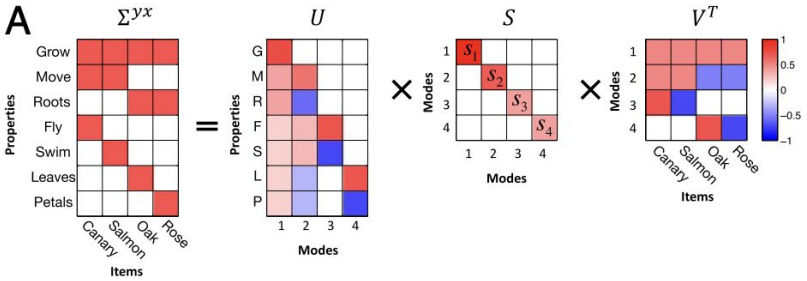
Let's call $a_\alpha(t)$ "**effective singular values**"

Allows us to interpret $a_\alpha(t)$ similar to s_α where,

- $a_\alpha(t)$: At t-th epoch, "according to the model", how much category-α to explains the dataset

  [**recall**: s^α : how much category-α actually explains the dataset]

singular values of the dataset



Method: Deep Linear Neural Networks - decomposition of weight matrices

- Now we can convert the forward-transformation ($W^1(t)*W^2(t)$) into few time-dependent scalars (i.e. effective singular values, $a_\alpha(t)$ for all α),

* - We can convert dynamics of weights evolution into dynamics of effective singular values

$$\tau \frac{d}{dt} a_\alpha = 2a_\alpha (s_\alpha - a_\alpha)$$

Solving this, we can get,

$$a_\alpha(t) = \frac{s_\alpha e^{2s_\alpha t/\tau}}{e^{2s_\alpha t/\tau} - 1 + s_\alpha / a_\alpha^0}$$

Side note: a sigmoidal growth (i.e. $1/ (1+e^{-x})$)

Meaning: We know “exactly” how and when model learns all the categorical distinctions α in the dataset

Lets visualize this !!!

Method: Deep Linear Neural Networks - decomposition of weight matrices

Before visualizing $a_\alpha(t)$, let's get effective singular value evolution equations for single-layer/ shallow network.

Network : $\hat{y} = \mathbf{W}^s \mathbf{x}$

Loss : $SSE(\mathbf{W}^s) = \frac{1}{2} \left\| \mathbf{y}^i - \hat{\mathbf{y}}^i \right\|^2$

Weight update : $\Delta \mathbf{W}^s = \lambda (\mathbf{y}^i - \hat{\mathbf{y}}^i) \mathbf{x}^{iT}$

Continuous weight evolution : $\tau \frac{d}{dt} \mathbf{W}^s = \mathbf{\Sigma}^{yx} - \mathbf{W}^s \mathbf{\Sigma}^x$
(Note: linear over W)

Evolution of effective singular values : $\tau \frac{d}{dt} b_\alpha = s_\alpha - b_\alpha$

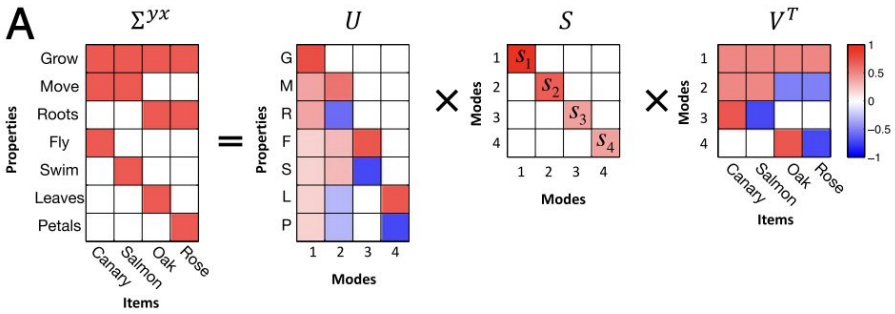
Solutions for the effective singular values : $b_\alpha(t) = s_\alpha \left(1 - e^{-t/\tau}\right) + b_\alpha^0 e^{-t/\tau}$

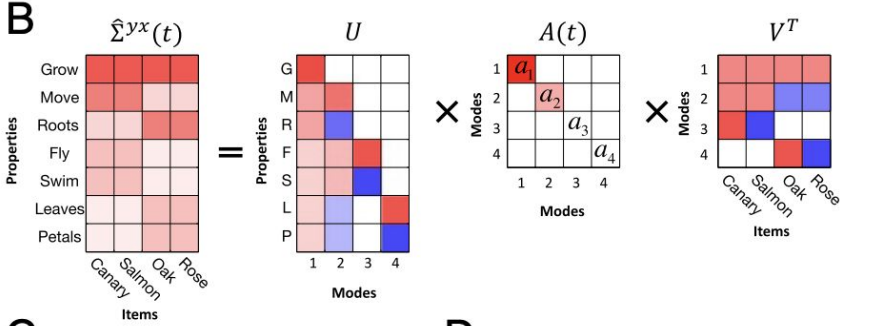
- **Side note**: $(1\text{-}e^x)$ growth

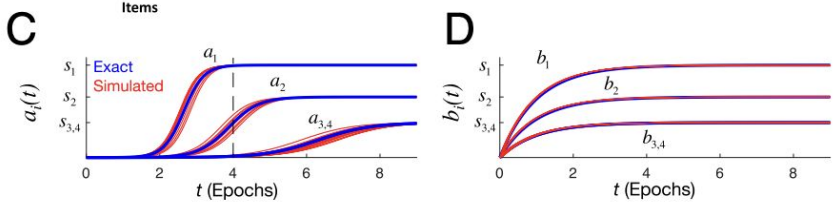
Empirical Literature on semantic cognition

1. Acquisition of semantic knowledge
    a. Stage-like learning
        - Relative stasis followed by abrupt conceptual reorganization [6, 7]
    b. Hierarchical differentiation
        - Broader categorical distinctions are generally learned before finer-grained distinctions [1, 5]
    c. Illusory/ incorrect facts during the developmental stasis [2]
2. Organization of semantic knowledge
    a. Category membership is a graded quantity
    b. Item typicality is reproducible across individuals [8, 9]
        - And correlates with performance on diverse semantic tasks [10, 14]
    c. Coherent vs less-coherent categories
        - Coherent categories can be learned/ represented relatively easily [8, 15, 16]
3. Deployment of semantic knowledge
    a. Inductive generation (i.e. make decisions about novel items/ properties) [2,3]
    b. Inductive generalization systematically changes over time: becomes more specific with age [2,3,17-19]
4. Neural representations of semantic knowledge
    a. Similarity structure of neural population vectors in response to different stimuli
        - Example works: Inanimate objects are differentiated from animate objects [22, 23]
    b. Such neural similarity structure is preserved across humans and monkeys [24, 25]
    c. Correspondence between neural similarity patterns and behavioral similarity patterns [21]

Acquisition: Rapid Stage-like Learning







Deep linear network

Shallow linear network

$$a_\alpha(t) = \frac{s_\alpha e^{2s_\alpha t/\tau}}{e^{2s_\alpha t/\tau} - 1 + s_\alpha / a_\alpha^0}$$

$$b_\alpha(t) = s_\alpha \left(1 - e^{-t/\tau}\right) + b_\alpha^0 e^{-t/\tau}$$

Observations: Sigmoidal vs exponential trajectory

1. For deep linear model:
   - Modes with strong explanatory power (higher s_α) learns faster
     - I.e. rapid stage-like transitions
2. For shallow linear model
   - All modes learn simultaneously

Time taken to achieve almost-best performance for given category/ mode-α

Deep linear model

$$t(s_\alpha, \epsilon) = \frac{\tau}{s_\alpha} \ln \frac{s_\alpha}{\epsilon}$$

higher s_α : learns faster

Shallow linear model

$$t(s_\alpha, \epsilon) = \tau \ln \frac{s_\alpha}{\epsilon}$$

Weakly associated with s_α

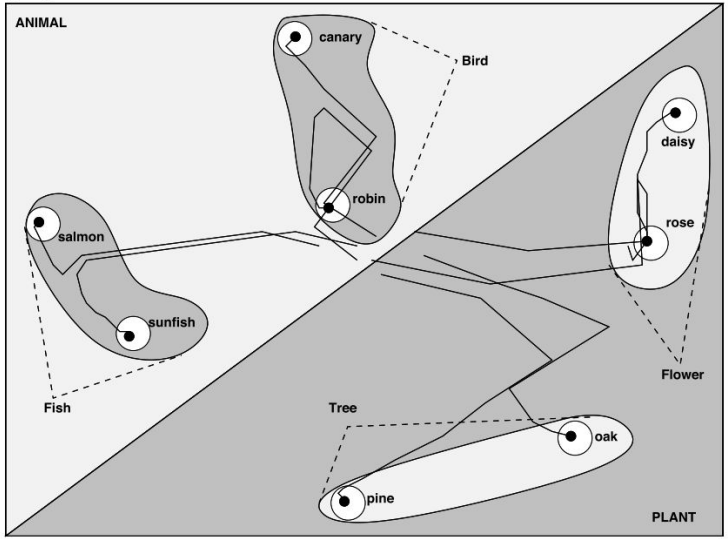
Acquisition: Progressive Differentiation in Hierarchical Structure

Previous work: Empirical results with deep nonlinear networks [4]

Questions:

1. How/ why?
2. Do simple linear models show this?

Goal: Find out how “exactly” hidden representations evolve during the training - with deep linear networks



From:
[4] Rogers, Timothy & Mcclelland, James. (2004). Semantic Cognition: A Parallel Distributed Processing Approach.

Acquisition: Progressive Differentiation in Hierarchical Structure

How:

$$\mathbf{W}^2(t)\mathbf{W}^1(t) = \mathbf{U}\mathbf{A}(t)\mathbf{V}^T = \sum_{\alpha=1}^{N_2} a_\alpha(t)\,\mathbf{u}^\alpha \mathbf{v}^{\alpha\,T}$$

1. We can solve $W^1(t)$, $W^2(t)$ using the effective singular value matrix, and V

$$\begin{aligned} \mathbf{W}^1(t) &= \mathbf{Q}\sqrt{\mathbf{A}(t)}\mathbf{V}^T, \\ \mathbf{W}^2(t) &= \mathbf{U}\sqrt{\mathbf{A}(t)}\mathbf{Q}^{-1} \end{aligned}$$

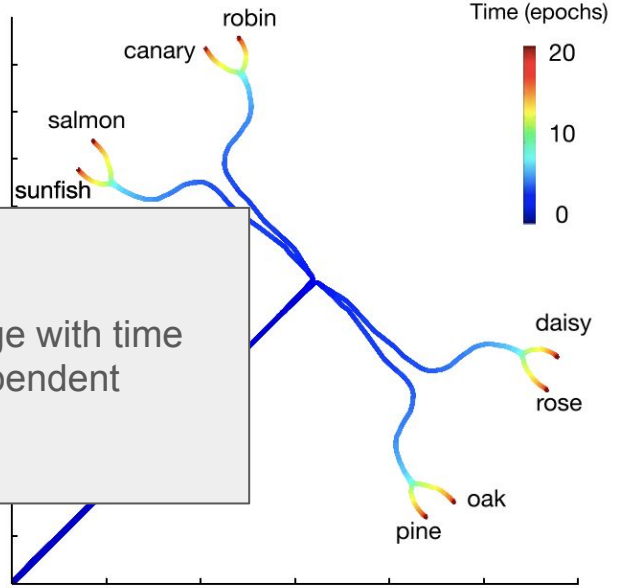
2. Then can compute hidden representations analytically : $h_i(t) = \mathbf{W}^1(t)\mathbf{x}_i$

$$h_i^\alpha(t) = \sqrt{a^\alpha(t)}\mathbf{v}_i^\alpha$$

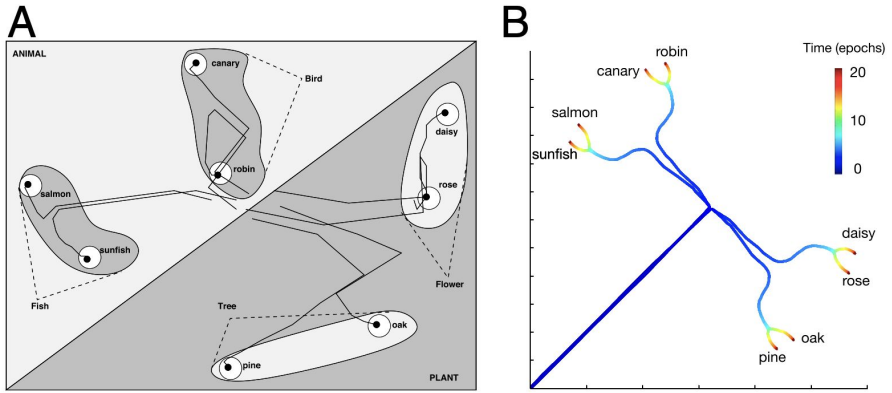
$$\mathbf{h}_i(t) = \begin{pmatrix} h_i^1(t) \\ h_i^2(t) \\ h_i^3(t) \\ \vdots \\ h_i^\alpha(t) \\ \vdots \\ h_i^{N_2}(t) \end{pmatrix}$$

Note:

1. v^α_i: does not change with time
2. $a^\alpha(t)$: only time-dependent quantity



Acquisition: Progressive Differentiation in Hierarchical Structure

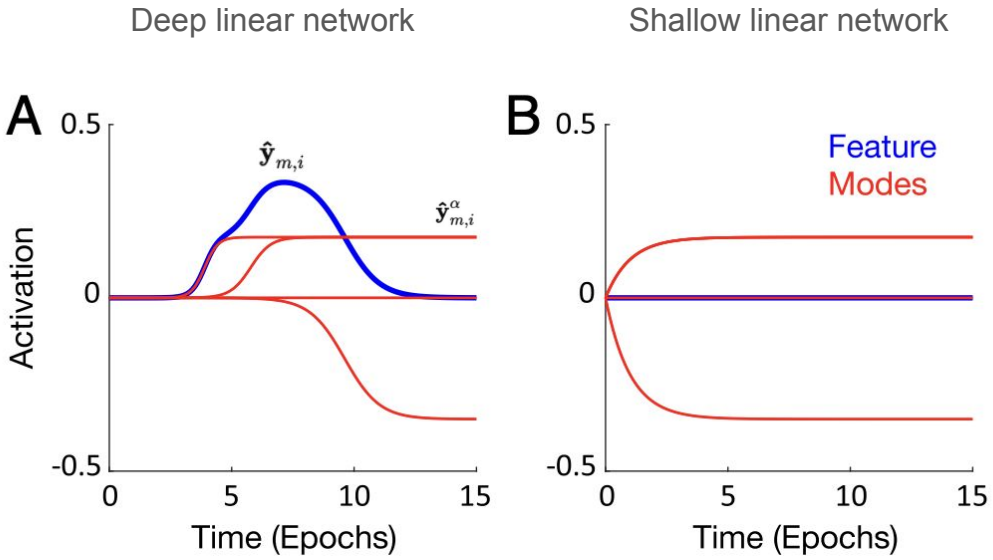


Summary:

1. Network does not have to be super complex to have hierarchical differentiation over training
2. This work theoretically showed that, deep “linear” networks can explain learning dynamics of nonlinear counter-parts

Super cool fact: Those categories are **not explicitly given to the network** (i.e. networks are learned to predict features of an item). **Learning was able to extract those “statistical structure encoded in the dataset”**

Acquisition: Illusory Correlations



Blue: predicted value of feature m = “can_fly” for item “salmon” during learning ($\hat{\mathbf{y}}_{m,i}$)

Red: Contribution from each mode (summation is the actual output) ($\hat{\mathbf{y}}^{\alpha}_{m,i}$)

Here, $\hat{\mathbf{y}}_{m,i} = \sum_{\alpha} \hat{\mathbf{y}}^{\alpha}_{m,i}$

Question: Why this is happening even when the network has a incremental, error-correcting learning process

Intuition:

- We minimize global error (i.e. across all properties/ items)
- Predicting “on-average” good results : could sometime result in transient increase in errors for specific properties/ items

Conclusion:

1. Even simple deep-linear networks could have illusory correlations
2. Shallow-linear networks could not have those.

Empirical Literature on semantic cognition

1. Acquisition of semantic knowledge ✅
   a. Stage-like learning
      - Relative stasis followed by abrupt conceptual reorganization [6, 7]
   b. Hierarchical differentiation
      - Broader categorical distinctions are generally learned before finer-grained distinctions [1, 5]
   c. Illusory/ incorrect facts during the developmental stasis [2]
2. Organization of semantic knowledge
   a. Category membership is a graded quantity
   b. Item typicality is reproducible across individuals [8, 9]
      - And correlates with performance on diverse semantic tasks [10, 14]
   c. Coherent vs less-coherent categories
      - Coherent categories can be learned/ represented relatively easily [8, 15, 16]
3. Deployment of semantic knowledge
   a. Inductive generation (i.e. make decisions about novel items/ properties) [2,3]
   b. Inductive generalization systematically changes over time: becomes more specific with age [2,3,17-19]
4. Neural representations of semantic knowledge
   a. Similarity structure of neural population vectors in response to different stimuli
      - Example works: Inanimate objects are differentiated from animate objects [22, 23]
   b. Such neural similarity structure is preserved across humans and monkeys [24, 25]
   c. Correspondence between neural similarity patterns and behavioral similarity patterns [21]

Organization: Duality between “item typicality” and “category prototype”

Goal:

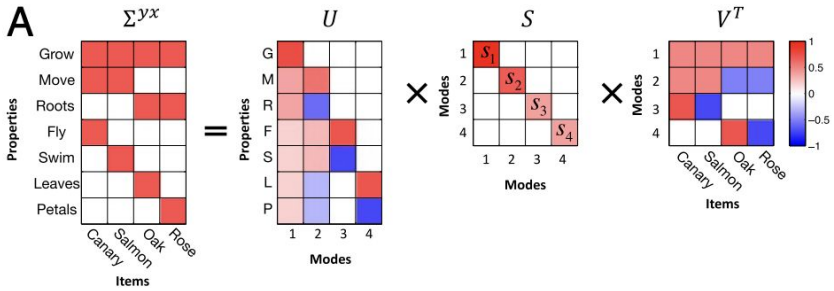
1. A “natural” mathematical definition for **item typicality**
2. Which **improves task performance**
    - i.e. typical item should give a highest performance for the tasks that are favourable for the category

Organization: Duality between “item typicality” and “category prototype”

Intuitive meaning for **item typicality** for category-α (say, category/ dim-α : animal(+)-plant(-) axis)

a. Canary, salmon, etc -> higher (+)
b. Oak, rose, etc -> higher (-)

This also is given by $\mathbf{v}^{\alpha}_{i}$



We can mathematically show that,

$$\hat{\mathbf{y}}^{\alpha}_{m,i} \leftarrow \mathbf{u}^{\alpha}_{m} \boldsymbol{s}_{\alpha} \mathbf{v}^{\alpha}_{i} \qquad \text{where,} \qquad \hat{\mathbf{y}}_{m,i} = \sum_{\alpha} \hat{\mathbf{y}}^{\alpha}_{m,i}$$

So, for a typical item i (i.e. a parrot over penguin for the category-α birds),

$$|\mathbf{v}^{\alpha}_{i}| > |\mathbf{v}^{\alpha}_{j}| \Rightarrow |\hat{\mathbf{y}}^{\alpha}_{i,m}| > |\hat{\mathbf{y}}^{\alpha}_{j,m}| \Rightarrow |\hat{\mathbf{y}}_{i,m}| > |\hat{\mathbf{y}}_{j,m}|$$

Meaning: Typical item gives higher performance, if the task is monotonic with response

e.g. In a desert, you need to find something to eat. If you have to decide between a plant and sand, which one are you more likely to eat to survive?

Organization: Duality between “item typicality” and “category prototype”

Previous definitions of item typicality : weighted sum of category-specific features present/ absent in the item

E.g. If an item can fly, eat worms, cannot bark, .. -> item could be a typical bird

Problem: which features are relevant (i.e. weighting scheme): rely on prior knowledge

But here, we can prove that,

$$\mathbf{v}_i^{\alpha} = \frac{1}{Ps_{\alpha}} \sum_{m=1}^{N_3} \mathbf{u}_m^{\alpha} \mathbf{o}_m^i$$

o^i_m: i-th item, m-th feature

How: For specific case of X= I,

Consider Σ^{xy}= (1/P)O, where O = [$y^1 \ldots y^{N1}$]

Meaning: weights for the feature-m is given by our feature synthesizer vector

[recall: u^{α}_m : how important m-th feature for the category-α]

Note: Definitions only valid for data generated by binary trees - so that singular dimensions explain the categories

Organization: Duality between “item typicality” and “category prototype”

Category Prototype: ideal set of feature vector that represents category-α

Previous theories:

- Is the features of the best exemplar of the category/ prototypical objects
- Weighted average of features of all the items : more typical items have higher weights

Problem: Weighting is rely on prior knowledge

We also can prove that,

$$\frac{1}{Ps_\alpha}\sum_{i=1}^{N_1}\mathbf{v}_i^\alpha \mathbf{o}_m^i$$

Organization: Duality between “item typicality” and “category prototype”

Category Prototype: ideal set of feature vector that represents category-α

Previous theories:

- Is the features of the best exemplar of the category/ prototypical objects
- Weighted average of features of all the items : more typical items have higher weights

Problem: Weighting is rely on prior knowledge

We also can prove that,

$$\mathbf{u}_m^\alpha = \frac{1}{Ps_\alpha} \sum_{i=1}^{N_1} \mathbf{v}_i^\alpha \mathbf{o}_m^i$$

Meaning: Take features from all the items, weighted by how typical the item is (v^α_i), and then average -> this gives the “feature synthesizer” vector

Conclusion: Feature synthesizer vector (u^α) is actually the category prototype of the category-α

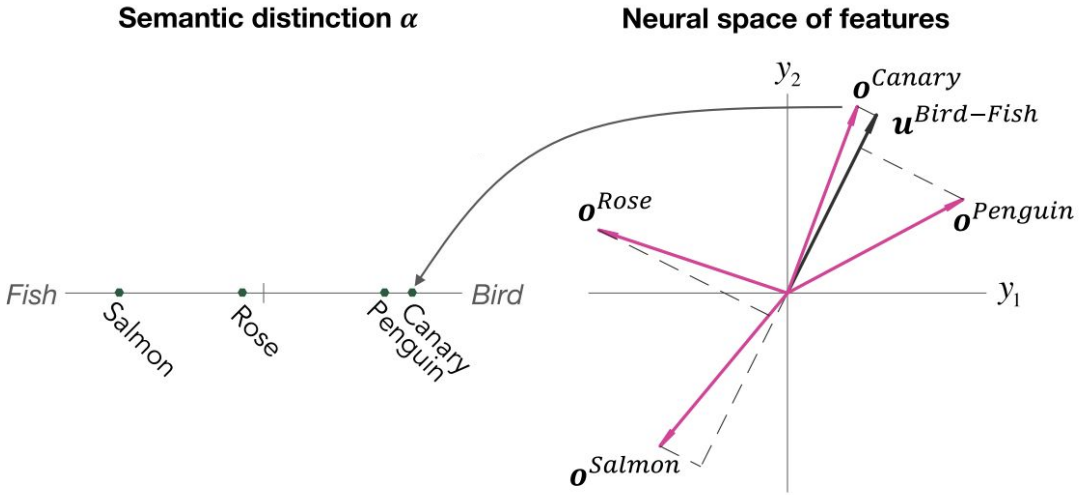
Organization: Duality between “item typicality” and “category prototype”

1. Item typicality is the cosine-similarity between category prototype and object feature vector

$$\mathbf{v}_i^{\alpha} = \frac{1}{Ps_{\alpha}} \sum_{m=1}^{N_3} \mathbf{u}_m^{\alpha} \mathbf{o}_m^i$$

2. Category prototype is the weighted average of object feature vectors weighted by that objects typicality

$$\mathbf{u}_m^{\alpha} = \frac{1}{Ps_{\alpha}} \sum_{i=1}^{N_1} \mathbf{v}_i^{\alpha} \mathbf{o}_m^i$$



Organization: Category Coherence

- Categories naturally learned are not arbitrary
  - They are coherent
  - Can efficiently represent the structure of the world
- **What is coherence (intuitively):**
  - Set of the things that are red have less coherence that the category of dogs
- **Questions:**
  - When is a category learned
  - What determines its coherence
- **Previous definitions**: Coherent categories consist of tight clusters of items that share many features AND highly distinct from other categories with different features [8, 15]
- **Problem**: definition could be circular [3, 16, 17]
  - To know which items are category members → need to know what features are important for that category
  - To know which features are important → need to know which items are members
- **Goal:** Give a definition for simple model of disjoint categories → demonstrate how NNs figure out above circular conflict

Organization: Category Coherence

Method:

Consider a dataset with N_0 objects and N_f features.

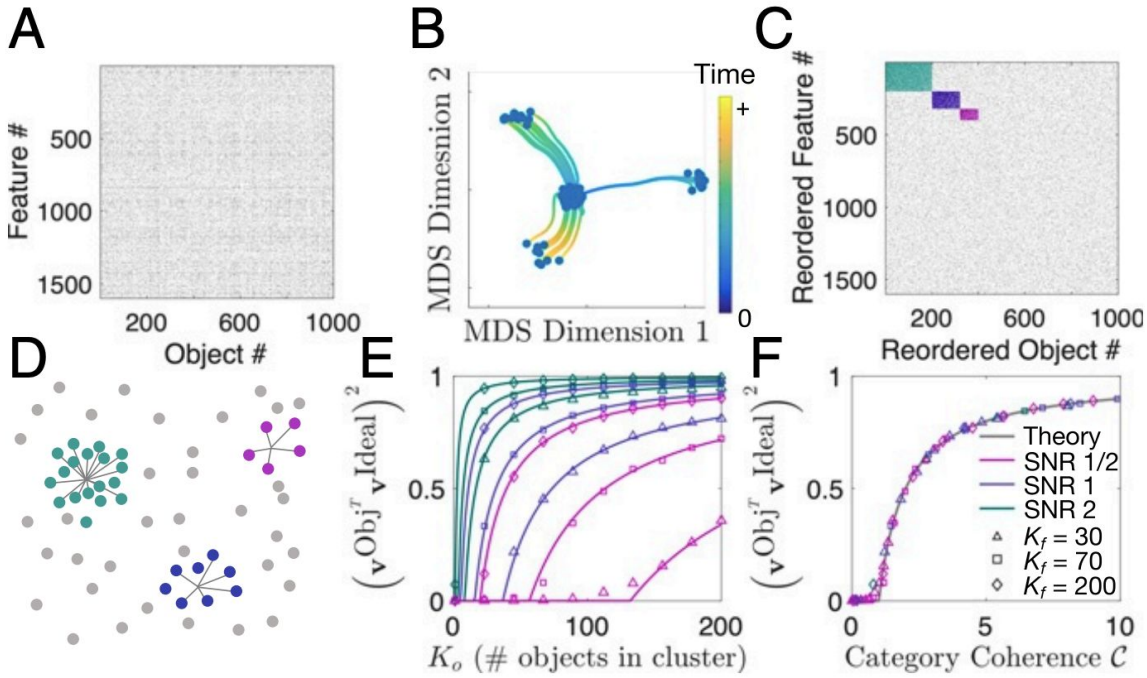
- Now consider a category in the dataset: with subset K_f features tend to occur higher prob. p in a subset of K_0 items.
- Background features occurs with low probability q in background items when they are not part of the category

Question: what values of K_f, K_0, p, q, N_f, N_0 → such a category can be learned, and how accurately?

Proposed theoretical solution:

$$\mathcal{C} = \mathrm{SNR}\frac{K_o K_f}{\sqrt{N_o N_f}} \qquad \text{where,} \qquad \mathrm{SNR} \equiv \frac{(p-q)^2}{q(1-q)}$$

Organization: Category Coherence



$$\mathcal{C} = \mathrm{SNR}\frac{K_o K_f}{\sqrt{N_o N_f}} \quad \text{where,} \quad \mathrm{SNR} \equiv \frac{(p-q)^2}{q(1-q)}$$

A: noisy dataset with 3 disjoint categories

B: evolution of internal representations: reveals 3 clusters

C: We can use V to find out which items are most typical for each category-α : and then reorder A. We should be able to observe 3 distinct categories

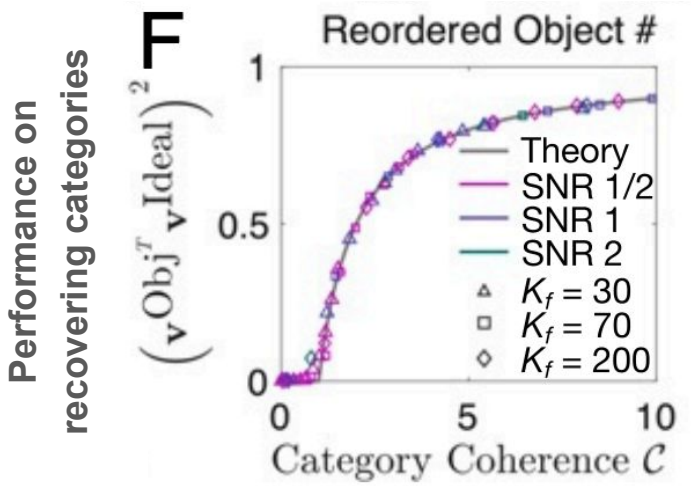
D: Based on item typicality, we can highlight the items

E: Performance in recovering such categories. Solid: theoretical, Symbols: Empirical

F: Change x-axis to C: all experiments collapse to one curve

- **Meaning**: Recovery depend on all the independent variables through coherence variable
- **Threshold behavior:** C <1: category cannot be learned

Organization: Category Coherence



$$\mathcal{C} = \mathrm{SNR}\frac{K_o K_f}{\sqrt{N_o N_f}}$$ where,

$$\mathrm{SNR} \equiv \frac{(p-q)^2}{q(1-q)}$$

Threshold behavior: C <1: category cannot be learned

If SNR= 1: threshold occur when

$$K_0 K_f = \sqrt{N_0 N_f}$$

Meaning:

- It's pretty relaxed

  i.e. if #features = 1600, #data= 1000, an **small category with #features= 40, #data= 40**

  - √(1600 * 1000) < 40*40

  is easily learnable even by a deep linear network

Organization: Category Coherence

how NNs solve the circularity problem:

[**Recall**: what circularity again: have important features ↔ being category member]

- simultaneously build object analyzers and feature synthesizers → gradually build accurate representations without needing prior knowledge

Summary:

- **Quantitative definition of category coherence** → give bounds to category-learning performance in neural networks
- Consistent with previous notions:
  - coherent categories have large subsets of items, with high probability large subsets of features Those features do not co-occur with other categories

Organization: Category Coherence

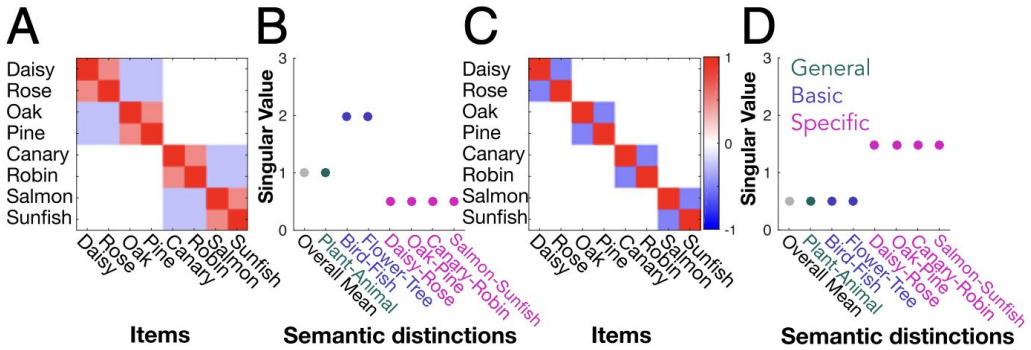


Fig. 8. From similarity structure to category coherence. (*A*) A hierarchical similarity structure over objects in which categories at the basic level are very different from each other due to a negative similarity. (*B*) For this structure, basic-level categorical distinctions acquire larger singular values, or category coherence, and therefore gain an advantage in both learning and task performance. (*C*) Now, subordinate categories are very different from each other through negative similarity. (*D*) Consequently, subordinate categories gain a coherence advantage. See *SI Appendix* for formulas relating similarity structure to category coherence.

Organization: Learning different statistical structures

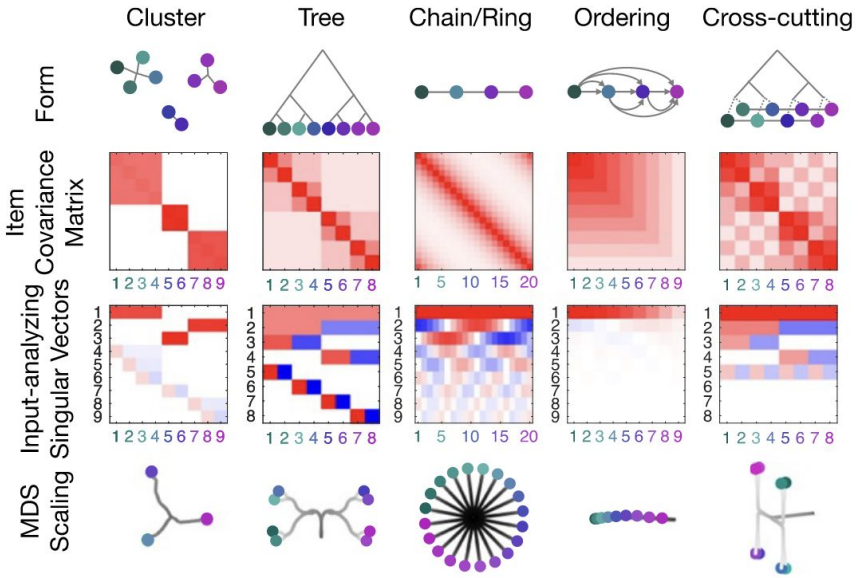
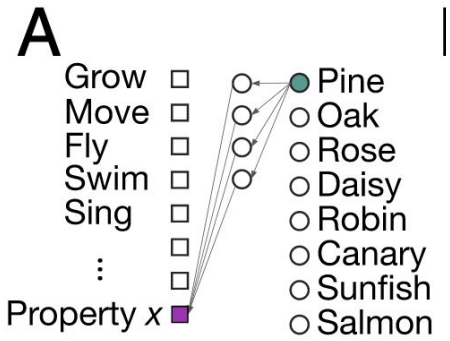


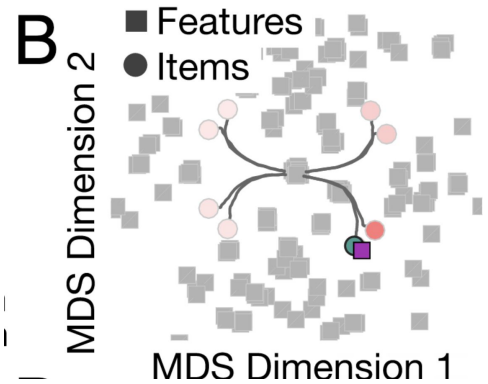
Fig. 9. Representation of explicit structural forms in a neural network. Each column shows a different structure. The first four columns correspond to pure structural forms, while the final column has cross-cutting structure. First row: The structure of the data generating PGM. Second row: The resulting item-covariance matrix arising from either data drawn from the PGM (first four columns) or designed by hand (final column). Third row: The input-analyzing singular vectors that will be learned by the linear neural network. Each vector is scaled by its singular value, showing its importance to representing the covariance matrix. Fourth row: MDS view of the development of internal representations.

Empirical Literature on semantic cognition

1. Acquisition of semantic knowledge ✅
   a. Stage-like learning
      - Relative stasis followed by abrupt conceptual reorganization [6, 7]
   b. Hierarchical differentiation
      - Broader categorical distinctions are generally learned before finer-grained distinctions [1, 5]
   c. Illusory/ incorrect facts during the developmental stasis [2]
2. Organization of semantic knowledge ✅
   a. Category membership is a graded quantity
   b. Item typicality is reproducible across individuals [8, 9]
      - And correlates with performance on diverse semantic tasks [10, 14]
   c. Coherent vs less-coherent categories
      - Coherent categories can be learned/ represented relatively easily [8, 15, 16]
3. Deployment of semantic knowledge
   a. Inductive generation (i.e. make decisions about novel items/ properties) [2,3]
   b. Inductive generalization systematically changes over time: becomes more specific with age [2,3,17-19]
4. Neural representations of semantic knowledge
   a. Similarity structure of neural population vectors in response to different stimuli
      - Example works: Inanimate objects are differentiated from animate objects [22, 23]
   b. Such neural similarity structure is preserved across humans and monkeys [24, 25]
   c. Correspondence between neural similarity patterns and behavioral similarity patterns [21]

Deployment: Inductive generalization





1. A novel feature (property x) is observed for a familiar item [A]
2. Learning assigns the novel feature (i.e. x) a neural representation [B]
    a. This places the feature in semantic similarity space near that object (i.e. Pine)
    b. Network then **inductively projects** that novel feature to other familiar items with closer hidden representations

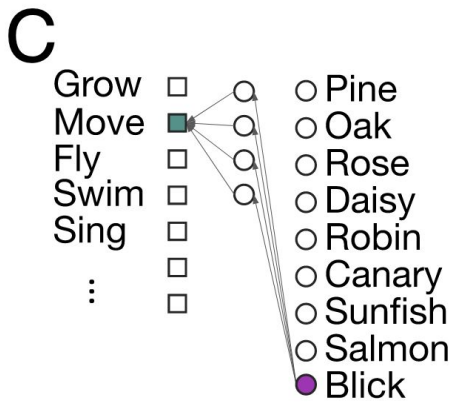
e.g. Tigers have 5 toes -> tigers are close to cats -> cats also have 5 toes

Analytical solution:

familiar item: i, novel feature m,
If item-j close to item-i : item-j also have the novel feature m

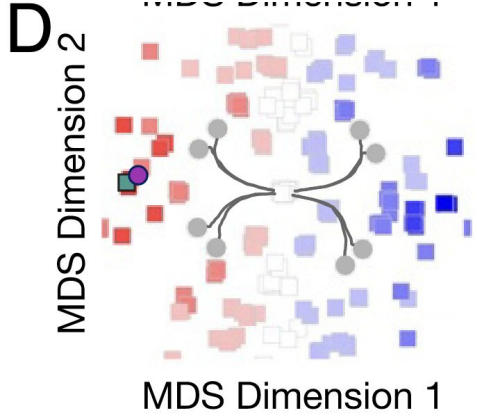
$$\hat{\mathbf{y}}_{m,j} = \frac{\mathbf{h}_j^T \mathbf{h}_i}{\|\mathbf{h}_i\|^2}$$

Deployment: Inductive generalization



1. A novel item (a blick) possesses a familiar feature
2. Learning assigns the novel item a neural representation
   a. This places the item in semantic similarity space near the that feature
   b. Other features are **inductively projected** to that item

e.g. a new animal which can bark -> bark is close to having 4 legs -> animal might have 4 legs

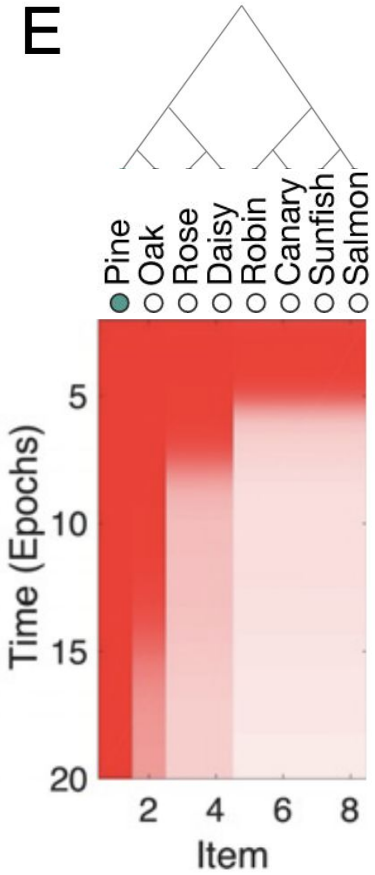


Analytical solution:

novel item: i, familiar feature m,
If feature-m is closer to feature-n, item i also has the feature-n

$$\hat{\mathbf{y}}_n = \mathbf{h}_n^T \mathbf{h}_m / \|\mathbf{h}_m\|^2$$

Deployment: Development shift in patterns of inductive generalization



Combining 2 previous observations:

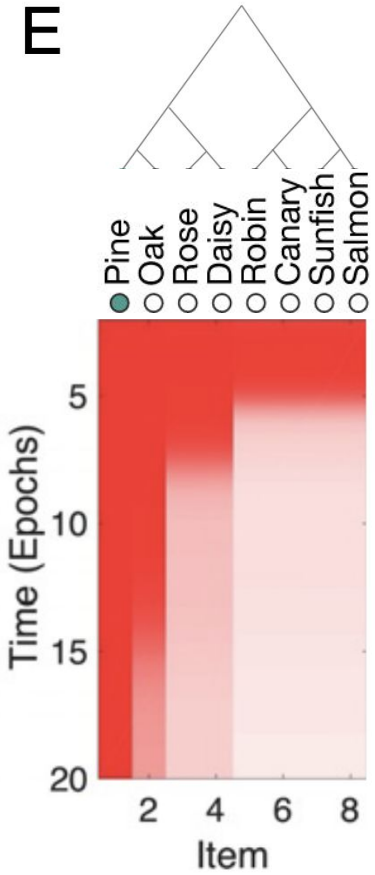
1. Networks learns hierarchical differentiation
2. During the learning, when network sees a new feature for a known item-i, it does inductive projection
    - items that are closer to the item-i -> predict the new feature

This naturally explains:

developmental shift in patterns of inductive projection from broad to specific,

Which is empirically observed in children [2, 3, 17, 18]

Deployment: Development shift in patterns of inductive generalization



1. **t = 7** : don't know the difference between trees and flowers
   - **Meaning**: Trees, flowers have closer representations
2. When it learns the new feature "pine" has, it project that feature to all "similar" items
   - And **that is, the entire plant group**
3. When it learns to differentiate flowers and plants (**t=10**), now the new feature of pine only projected into other trees (i.e. oak)

Empirical Literature on semantic cognition

1. Acquisition of semantic knowledge ✅
   a. Stage-like learning
      - Relative stasis followed by abrupt conceptual reorganization [6, 7]
   b. Hierarchical differentiation
      - Broader categorical distinctions are generally learned before finer-grained distinctions [1, 5]
   c. Illusory/ incorrect facts during the developmental stasis [2]
2. Organization of semantic knowledge ✅
   a. Category membership is a graded quantity
   b. Item typicality is reproducible across individuals [8, 9]
      - And correlates with performance on diverse semantic tasks [10, 14]
   c. Coherent vs less-coherent categories
      - Coherent categories can be learned/ represented relatively easily [8, 15, 16]
3. Deployment of semantic knowledge ✅
   a. Inductive generation (i.e. make decisions about novel items/ properties) [2,3]
   b. Inductive generalization systematically changes over time: becomes more specific with age [2,3,17-19]
4. Neural representations of semantic knowledge
   a. Similarity structure of neural population vectors in response to different stimuli
      - Example works: Inanimate objects are differentiated from animate objects [22, 23]
   b. Such neural similarity structure is preserved across humans and monkeys [24, 25]
   c. Correspondence between neural similarity patterns and behavioral similarity patterns [21]

Neural Representations: representation, behavior similarity

Observations:

a. Similarity structure of neural population vectors in response to different stimuli
   - Example works: Inanimate objects are differentiated from animate objects [22, 23]

b. Such neural similarity structure is preserved across humans and monkeys [24, 25]

c. Correspondence between neural similarity patterns and behavioral similarity patterns [21]

Questions:

1. Why is **representational similarity conserved**?
2. When the **brain mirrors the behavior**?

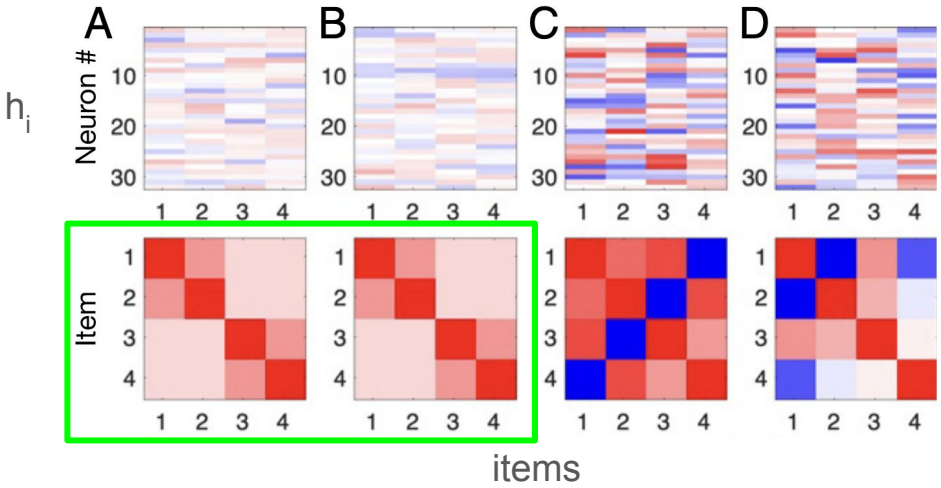
Neural Representations: conservation of representation similarity

Q1: Why is **representational similarity conserved**?

Under the assumption: **start learning from small random initial weights**

- We can prove that, **representational similarity matrices** should be consistent, even when hidden features are very different

(e.g. Biological analogue: all humans think cat faces are similar to tiger faces)



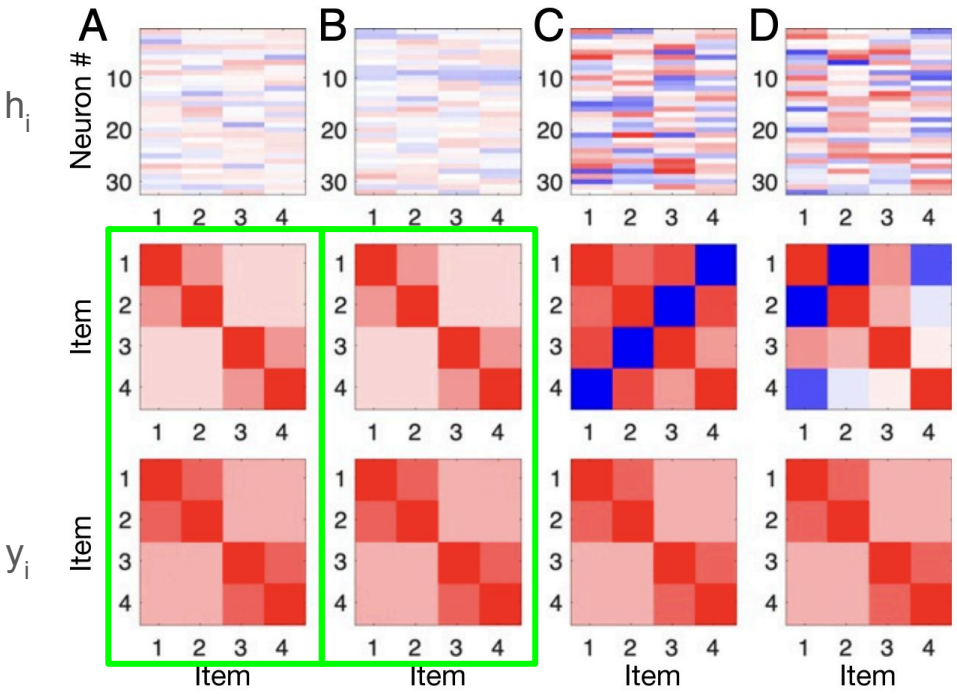
A, B: starts with small norm random weights

C, D: large-norm random weights

Why assumption: allows network to have minimum norm weights -> optimally implement the desired task

Neural Representations: brain mirror behavior

Q2: When the brain mirrors the behavior?



- We can also prove that, if the **network learns optimal smallest weights**,

$$\boldsymbol{\Sigma}^{\hat{y}} = \left(\boldsymbol{\Sigma}^{h}\right)^2$$

-> **Representation correlations** and **behavior correlations** have the same structure

Neural Representations: Optimal learning in brains

In the biological brain,

1. Conversation of representation similarity is observed in biological brain.
2. Correlations between behavior and representations also observed in biological brain

For the linear networks,

- Both can be observed if the learning is optimal with minimal weights

Does this mean: Learning in the brain is optimal with minimal synaptic strengths [?]

Summary: Semantic Cognition **Empirical Literature aligns with Deep Linear Networks**

1. Acquisition of semantic knowledge ✅
   a. Stage-like learning
      - Relative stasis followed by abrupt conceptual reorganization [6, 7]
   b. Hierarchical differentiation
      - Broader categorical distinctions are generally learned before finer-grained distinctions [1, 5]
   c. Illusory/ incorrect facts during the developmental stasis [2]
2. Organization of semantic knowledge ✅
   a. Category membership is a graded quantity
   b. Item typicality is reproducible across individuals [8, 9]
      - And correlates with performance on diverse semantic tasks [10, 14]
   c. Coherent vs less-coherent categories
      - Coherent categories can be learned/ represented relatively easily [8, 15, 16]
3. Deployment of semantic knowledge ✅
   a. Inductive generation (i.e. make decisions about novel items/ properties) [2,3]
   b. Inductive generalization systematically changes over time: becomes more specific with age [2,3,17-19]
4. Neural representations of semantic knowledge ✅
   a. Similarity structure of neural population vectors in response to different stimuli
      - Example works: Inanimate objects are differentiated from animate objects [22, 23]
   b. Such neural similarity structure is preserved across humans and monkeys [24, 25]
   c. Correspondence between neural similarity patterns and behavioral similarity patterns [21]

Thank you !!

References [All the citations are from the paper: Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. Proceedings of the National Academy of Sciences.]

[1] 1. Keil F (1979) Semantic and Conceptual Development: An Ontological Perspective (Harvard Univ Press, Cambridge, MA).
[2] Carey S (1985) Conceptual Change In Childhood (MIT Press, Cambridge, MA).
[3] Murphy G (2002) The Big Book of Concepts (MIT, Cambridge, MA).
[4] Rogers TT, McClelland JL (2004) Semantic Cognition: A Parallel Distributed Processing Approach (MIT Press, Cambridge, MA).
[5] Mandler J, McDonough L (1993) Concept formation in infancy. Cogn Dev 8:291–318.
[6] Inhelder B, Piaget J (1958) The Growth of Logical Thinking from Childhood to Adolescence (Basic Books, New York).
[7] Siegler R (1976) Three aspects of cognitive development. Cogn Psychol 8:481–520.
[8] Rosch E, Mervis C (1975) Family resemblances: Studies in the internal structure of categories. Cogn Psychol 7:573–605.
[9] Barsalou L (1985) Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. J Exp Psychol Learn Mem Cogn 11: 629–654.
[10] Rips L, Shoben E, Smith E (1973) Semantic distance and the verification of semantic relations. J Verbal Learn Verbal Behav 12:1–20.
[14] Osherson D, Smith E, Wilkie O, Lo´ pez A, Shafir E (1990) Category-based induction. Psychol Rev 97:185–200.
[15] Rosch E (1978) Principles of Categorization in Cognition and Categorization, eds Rosch E, Lloyd B (Lawrence Erlbaum, Hillsdale, NJ), pp 27–48.

References

[16] Murphy G, Medin D (1985) The role of theories in conceptual coherence. Psychol Rev 92:289–316.
[17] Keil F (1991) The emergence of theoretical beliefs as constraints on concepts. The Epigenesis of Mind: Essays on Biology and Cognition, eds Carey S, Gelman R (Psychology Press, New York).
[18] Carey S (2011) Pre´ cis of 'the origin of concepts'. Behav Brain Sci 34:113–124.
[19] Gelman S, Coley J (1990) The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. Dev Psychol 26:796–804.
[20] Edelman S (1998) Representation is representation of similarities. Behav Brain Sci 21:449–467.
[21] Kriegeskorte N, Kievit R (2013) Representational geometry: Integrating cognition, computation, and the brain. Trends Cogn Sci 17:401–412.
[22] Carlson T, Simmons R, Kriegeskorte N, Slevc L (2014) The emergence of semantic meaning in the ventral temporal pathway. J Cogn Neurosci 26:120–131.
[23] Connolly A, et al. (2012) The representation of biological classes in the human brain. J Neurosci 32:2608–2618.
[24] Mur M, et al. (2013) Human object-similarity judgments reflect and transcend the primate-IT object representation. Front Psychol 4:128.
[25] Kriegeskorte N, et al. (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126–1141.