# ELL788 - Computational Perception and Cognition
# Term Paper

17 November 2015

# The Many Folds of Cognition: A Topological Perspective on Form Learning

Theme : Bayesian Cognition

Authors : Anshul Bawa, Sahil Loomba

## Abstract

*Given the resource constraints that human cognitive systems operate under, how do adults and children inductively learn from the percepts we encounter? Certainly, there are some heuristics involved in not just our inductive learning, but also in the way we do inference and reasoning. Can we explain away the use of such "shortcuts" by improving current computational cognitive models? This paper is an attempt in answering such questions, by taking Kemp et al.'s work on form learning as a neat illustration of some key ideas. We extend their work by stressing on the need to include topology theory (manifold learning) and hierarchical Bayesian modelling, by providing rigorous experimental results to support our stance.*

***Keywords***: Cognitive models, Manifold learning, Form learning, Bayesian cognition, Representativeness Heuristic, Availability Heuristic

**Caveat:**

- ➔ All scripts for Form Learning sourced from Kemp et. al's work at http://www.psy.cmu.edu/~ckemp/code/formdiscovery.html
- ➔ All scripts for Stochastic Neighbourhood Embedding sourced from van Der Maaten's dimensionality reduction toolbox written in MATLAB https://lvdmaaten.github.io/drtoolbox/
- ➔ Graphviz Library used for visualising networks
- ➔ All other scripts for experiments and analysis were written by the authors in MATLAB

# Introduction - Heuristics in Inductive Learning

In our initial proposal, we had proposed that a computational model of bounded rationality could possibly explain how humans learn abstract knowledge from very sparse data, something which machine intelligence has not been able to replicate [1]. We shift our focus from a decision-making formulation, where the discussion on bounded rationality would have been directly applicable, to the cognitive task of learning structure and form of data, and updation of one's belief with new evidence. Instead of establishing a cost optimisation problem on the number of samples taken for effective learning as in [1], we talk about constrained resources in the sense that humans can register only a compact number of features into a robust representation. The question that this paper tries to address is how do we perform inductive reasoning and arrive at generalisations. Topological Data Analysis (TDA) has been used before to explain parts of human perception, especially in vision, wherein a three-dimensional perception can be developed of a two-dimensional image falling on the retina [2]. Drawing a parallel with learning abstract representations from sparse data, we felt motivated to use TDA to better explain learning of categories and their relationships with other categories.

We show that our approach, which attempts to augment the form-learning framework of Kemp et al.[3] with manifold learning (TDA) in the feature space, provides insight into how our brains categorise, and begins to explain the simplifying information-processing shortcuts (heuristics) that humans employ while performing judgment tasks involving inductive generalisations in general, and while learning categories in particular.

The **representativeness heuristic** [4] is addressed, which is seen as a manifestation of our tendency to rely on a simpler representation of an experience while learning by induction. The representativeness heuristic says that subjects assess the similarity of objects and organise them based around the category prototype. This heuristic is used in cognitive systems because it is an easy computation [5]. This is also known as the typicality effect, and is an artifact of the nature of semantic memory. We show that manifolds, besides being better at learning the underlying form, also capture this typicality effect observed in our cognitive systems.

We also attempt to address the **availability heuristic**. This heuristic operates on the notion that if something can be easily recalled, it must be more important than alternatives that are not as readily recalled [10]. We demonstrate that our model captures the role of this heuristic in gist-extraction, under the reasonable assumption that visual memory has a better ease of retrieval than memory of concepts.

We initially set out with the puzzle of inductive leaps of reasoning in our cognition, the question of how complex and abstract underlying knowledge is learned from relatively sparse data. This points to the existence of some innate knowledge, which can be encoded in the form of priors on our reasoning system, like the hyperparameters of a hierarchical Bayesian network, or an underlying topological space like a manifold over which observations/data are smoothly embedded. Observation of similar or repeated experiences reinforce and validate the learnt manifolds.

# Motivation - Compact Knowledge Representation

The learning of categories occurs in semantic memory and not in episodic memory, and therefore its contents are not tied to any particular instance or experience [6]. Real-life observations involve thousands of observable features. However, not all of them are remembered, like a computer does. The mental representation does not enlist all these observed features, but all the observed features would nevertheless contribute implicitly to the encoding of a much simpler and lighter mental representation that has distilled out a small number of concepts, or knowledge points. These learnt concepts become the basis of how we relate different observations, and how we classify them.

Here is a demonstration of what we mean. On seeing an elephant for the first time, a child might notice numerous things about the elephant, like its long serpentine trunk, large and flappy ears, dull grey color, trunk-like feet, scaly and rough skin, curved and sharp pair of tusks, or diminutive tail. But the child will not memorise the elephant as a sum of its individual features. What is stored in semantic memory is the “gist” of experience, an abstract structure that applies to a wide variety of experiential objects and delineates categorical and functional relationships between such objects. This gist incorporates as many instances of objects experienced as possible without making the gist unnecessarily complex. Through the “manifolds” that we learn in our experiments extending the experiments of [3], we attempt to model this very gist that the cognitive system stores in its semantic memory. A complete theory of semantic memory should account not only for the representational structure of such gists, but also how they can be extracted from experience. We provide some pointers to work in the latter direction in our section on future work.

# Formulation - Extracting the Gist-of-Things

An operationalisation of what we mean by extracting the gist of numerous features is a dimensionality reduction into a handful of features. We foresaw and have demonstrated that a linear dimensionality reduction would be too simplistic, and that too much information is lost with even small reductions in dimensionality. Therefore, more sophisticated non-linear techniques of doing so are considered. We zeroed in on an algorithm for manifold learning, called stochastic neighbour embedding (SNE). The appendix contains more on SNE. The linear dimensionality reduction technique that has been pitted against SNE is principal component analysis (PCA).

Kemp et al. have proposed a graph grammar over cluster graphs to unify many different cognitively natural forms of categorisation into one framework [3]. One of the main datasets that they attempt to extract the underlying structure and form of, is a matrix of 102 features (anatomical, ecological and behavioural) for 33 different animals. Using this, they learn that a tree is the best form to categorise this data, and they also learn the structure of the tree [Figure 1], i.e, a classification of animals.

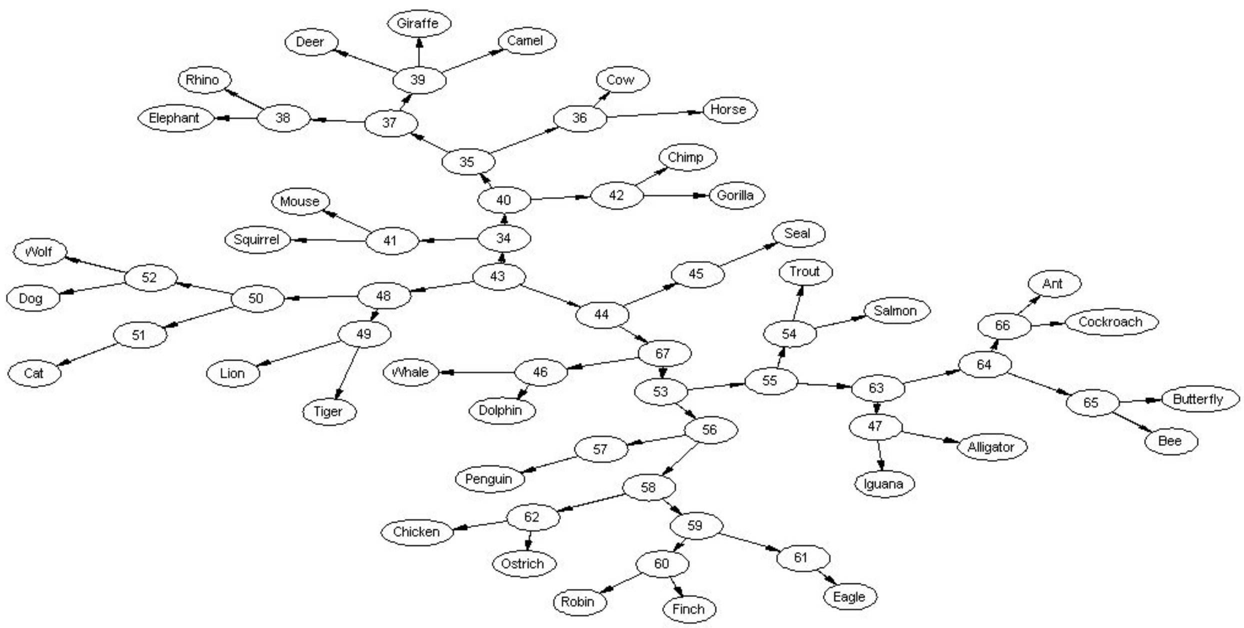


**Figure 1(a) - Ground Tree**

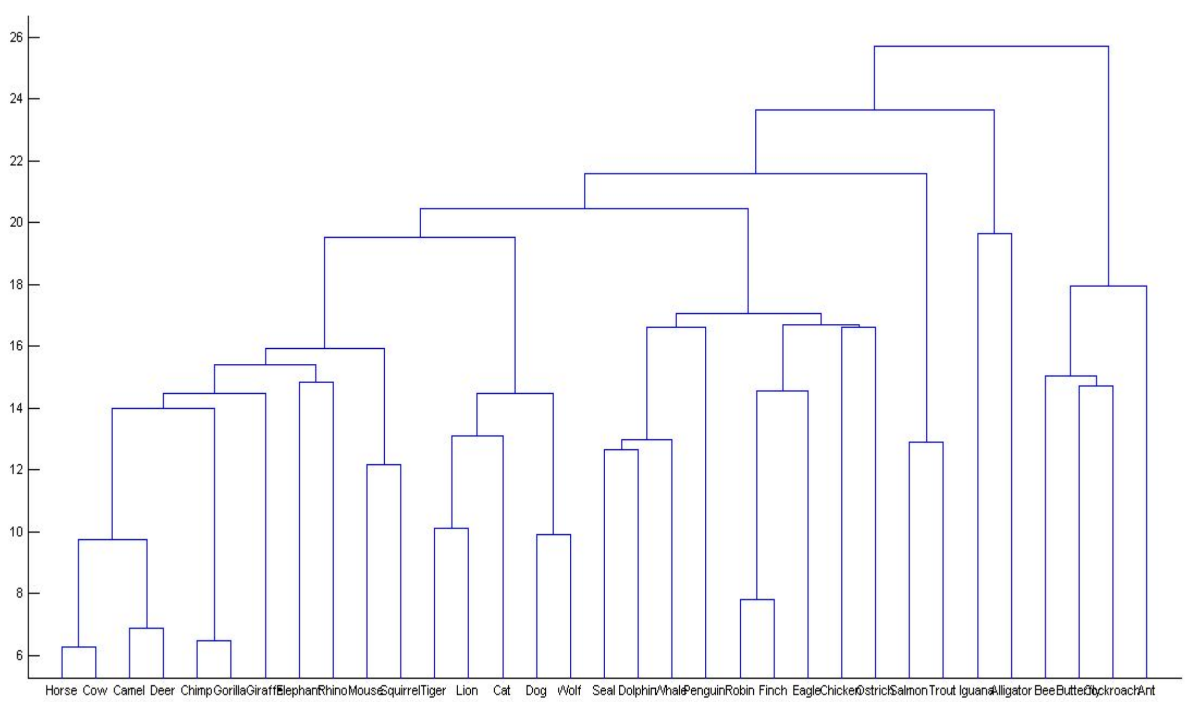


**Figure 1(b) - Ground HAC dendrogram**

They have attempted to model the development of a conceptualisation of categorisation by varying the number of features by choosing subsets of different sizes from the set of features. This is not entirely representative of how aggregation of information takes place cognitively. A smarter way of being selective with features is needed. As categorising over hundreds of features is cognitively very expensive, we hypothesise that in reality, our cognitive systems extract the gist of all these features before using this information for conceptual categorisation. Also, direct categorisation, as done by Kemp et al., leads to some erroneous inference on this animal dataset. We believe that equal priors on all the 8 forms, as used in their work, is not a good idea.

We have augmented their model with an initial step of dimensionality reduction. The kind of dimensionality reduction technique used is varied, to compare a topological space analysis method to a simplistic, linear technique. The number of dimensions of the reduced space is also varied in both techniques, and an analysis is attempted of the likelihoods and intuitiveness of various forms learnt over these reduced spaces.

As a metric of comparing outputs across these tweaks, their log-likelihood scores are not exclusively relied upon as a metric of the accuracy of form-learning, because some problematic results were seen that way. Two tree structures are compared to each other and to the natural tree form using a simple heuristic that we have formulated based on matching pairwise shortest distances between leaves of the tree (refer to the appendix for more). Dendrograms obtained on running hierarchical agglomerative clustering (HAC) on this data, are also provided as an aid to visualising similarity.

# Results - Why Manifolds Work

When the form learning procedure of [3] was performed over data of a reduced dimensionality, obtained after SNE, we observe that increasing dimensionality (from as low as 2 to higher) improves the similarity score of a tree form (to the “true” form, which is the tree structure obtained by ground features). But, such an improvement is not observed in the case of PCA [Table 1]. In fact, even the lowest dimensional SNE (SNE-2) performs better than all PCAs (2,4,8) which we tried for.

| **Dimensionality Reduction Method** | **Similarity of tree to true tree (lower is better)** |
|---|---|
| PCA-2 | 0.9303 |
| PCA-4 | 0.9327 |
| PCA-8 | 0.9396 |
| SNE-2 | 0.9280 |

| | |
|---|---|
| SNE-4 | 0.9249 |
| SNE-8 | 0.9063 |

**Table 1 - Similarity of tree forms**

It may be observed that the tree structure predicted over SNE-reduced data is fairly similar to the structure of the actual tree learnt over the full data, even for very low number of dimensions, by visualising the predicted trees below [Figure 2].

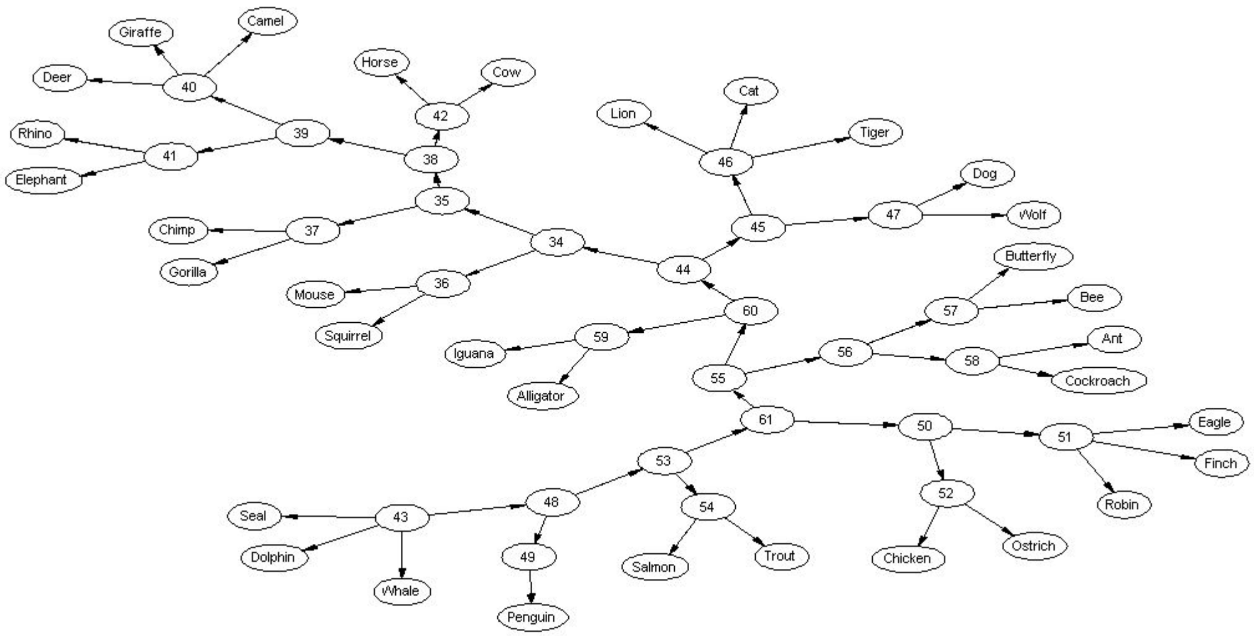


**Figure 2(a) - SNE-16 tree**

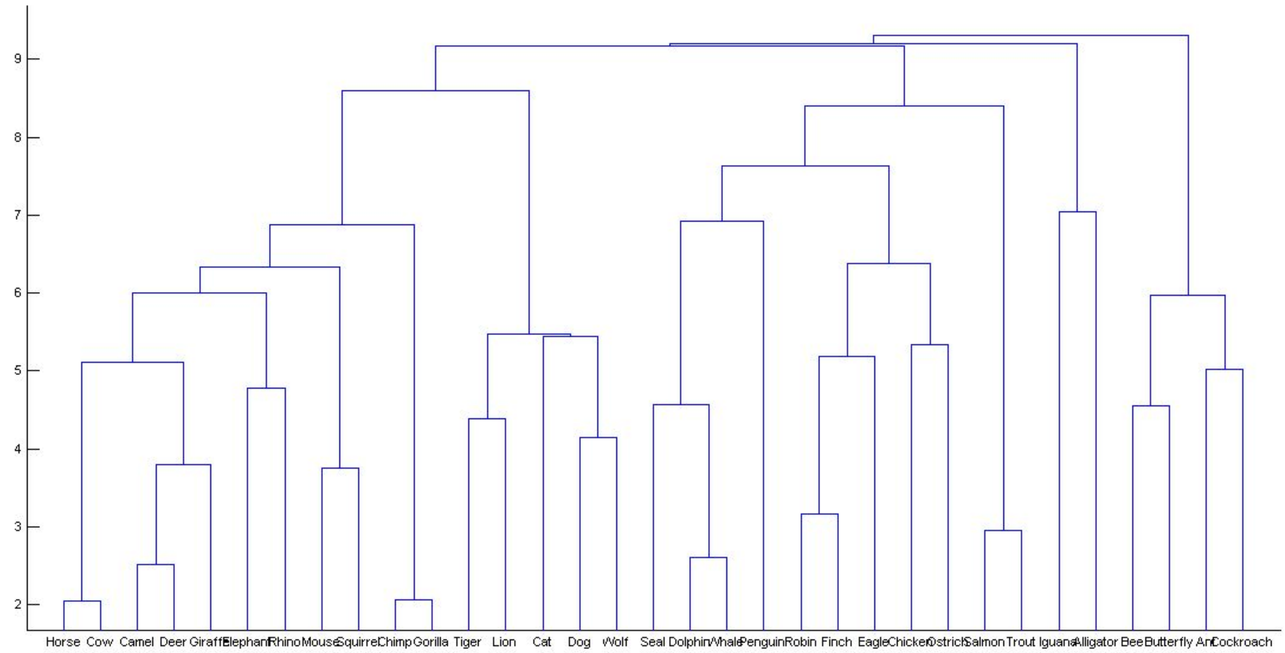


**Figure 2(b) - SNE-16 HAC dendrogram**

The corresponding tree structure predicted over PCA-reduced data has many more dissimilarities to the actual tree [Figure 3].

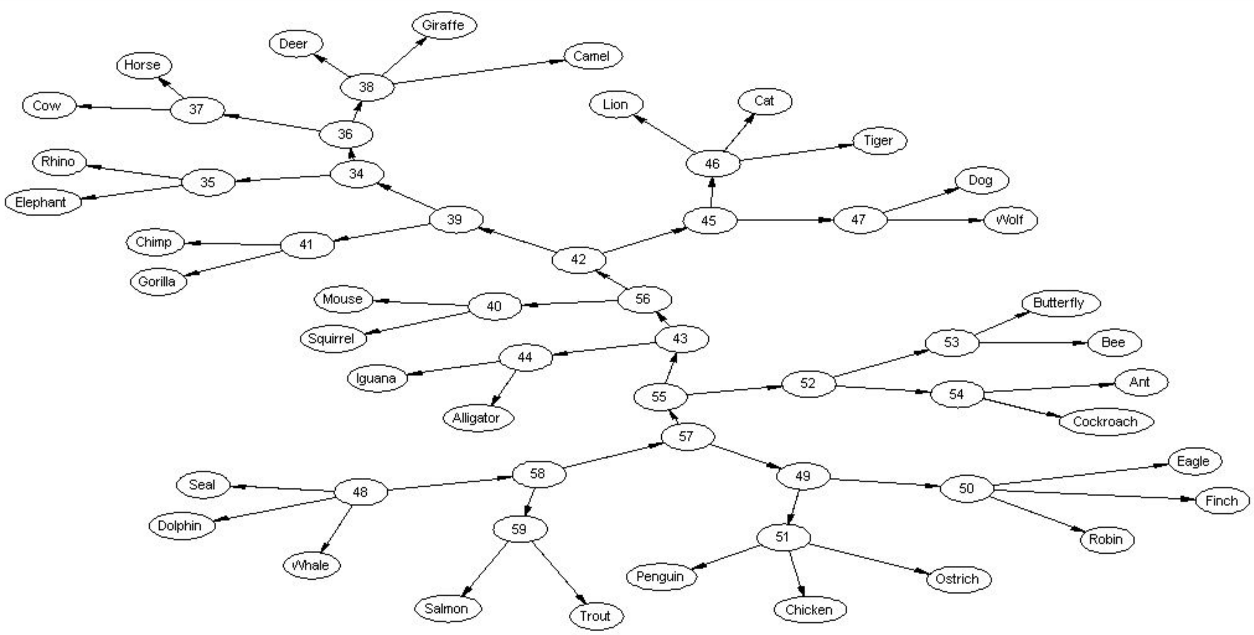


**Figure 3(a) - PCA-16 tree**

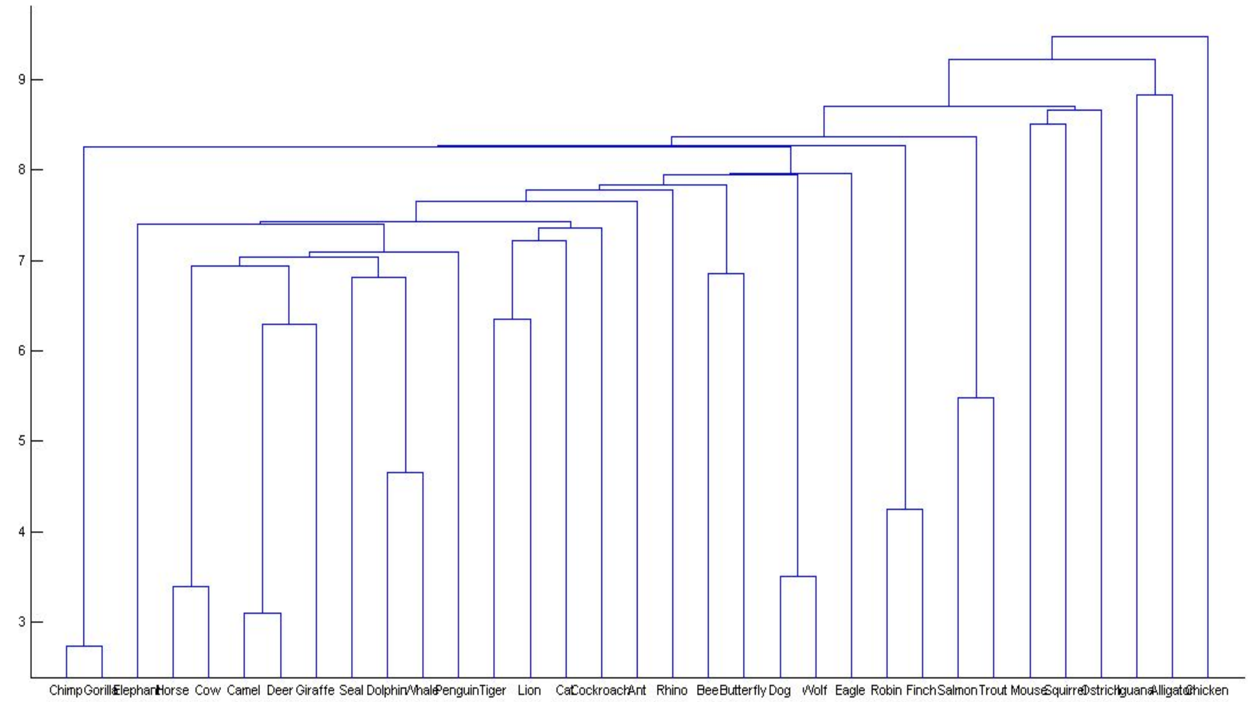


**Figure 3(b) - PCA-16 HAC dendrogram**

It is observed that out of the eight forms considered, their algorithm originally assigns a higher likelihood to a ring form than to a tree form, albeit by a small margin [Table 2] [Figure 4]. However, after dimensionality reduction by SNE on the same dataset, the tree turns out to be a likelier form than a ring, which means our preprocessing of gist-extraction corrects the erroneous deduction mentioned above. So SNE-reduced data corrects the false form-learning (ring over tree) that occurs over the raw dataset in [3].

| Dataset | Log of relative likelihood of tree w.r.t ring (higher is better) |
|---|---|
| Ground | -2.7 |
| SNE-4 | 1.7 |

**Table 2 - Relative likelihood of tree and ring forms**

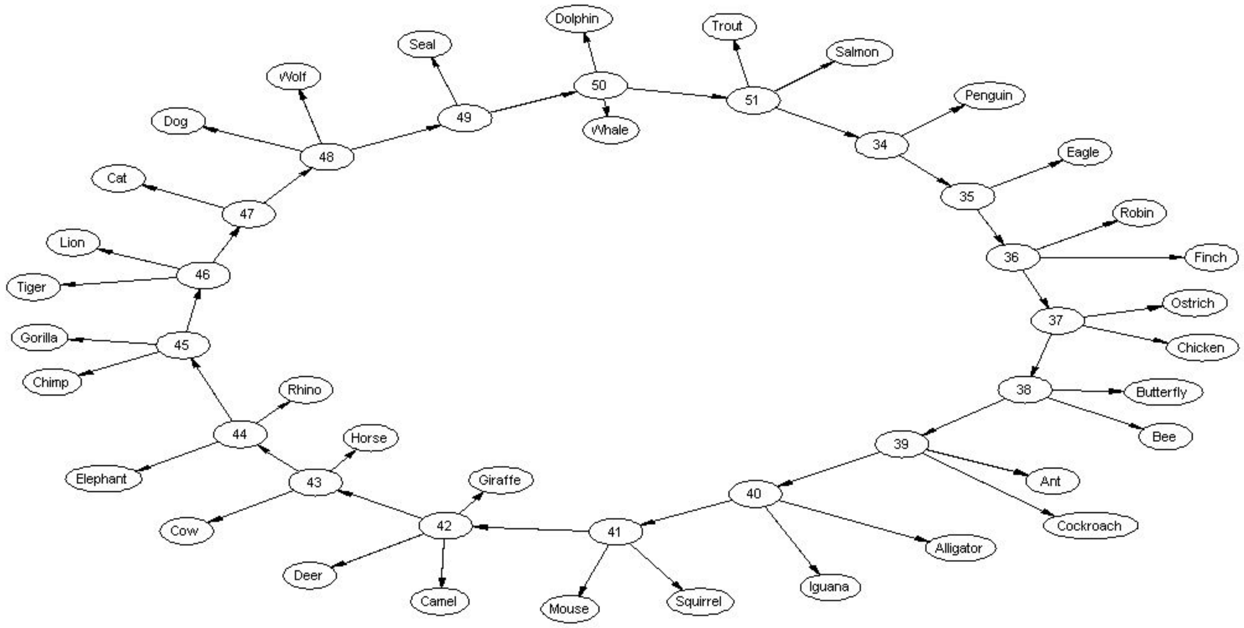


**Figure 4(a) - Ground Ring**

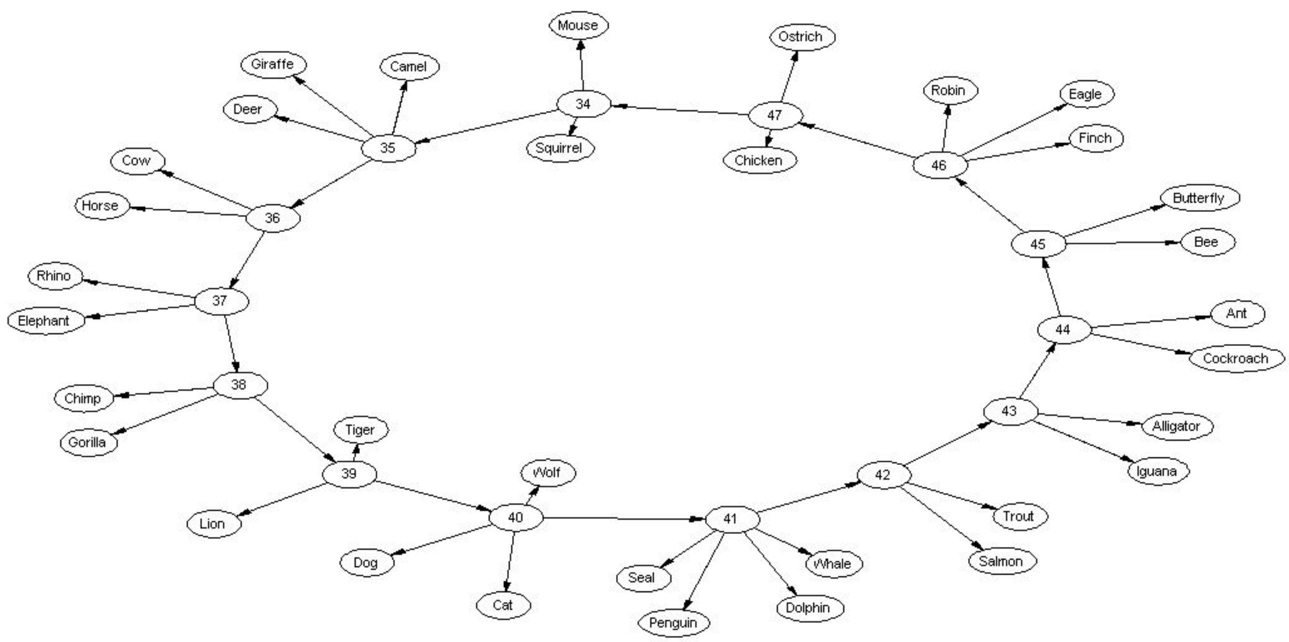


**Figure 4(b) - SNE-16 Ring**

It is seen that when we reduce dimensions of the feature space from 102 to 4 using SNE, the reduced dimensions indeed do capture the essence of a lot of the features. One of the dimensions (Dimension 3 in Figure 5) corresponds to the habitat of animals. Animals of aquatic habitat have high negative values in this dimension, animals that fly are in the low negative to low positive range, and terrestrial animals have high positive values. Another dimension

(Dimension 4 in Figure 5) corresponds to their eating habits, with carnivores getting a high negative value and herbivores getting a low negative to high positive value. Multiple features in the set of 102 have weak to strong correlations to these feature categories, and performing SNE extracts the gist of all these features as seen above with the more abstract features of habitat and food habits. Dimension 1 and Dimension 2, that were extracted by SNE are the dot colour and dot size, respectively.

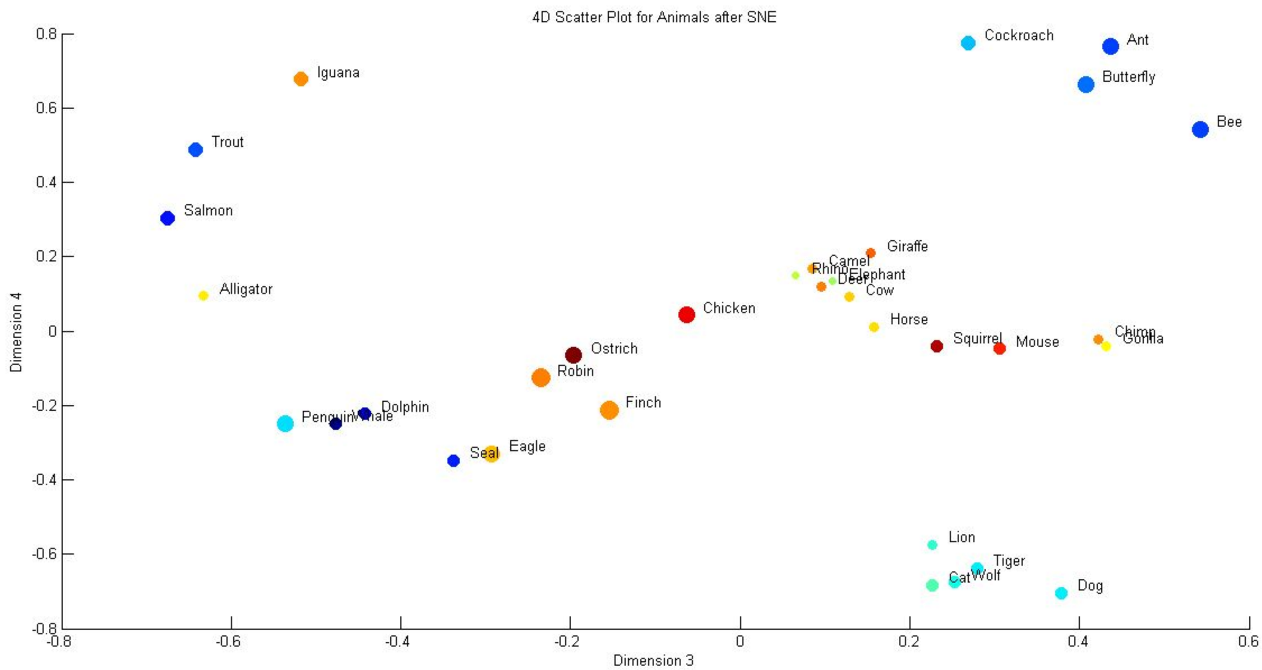


**Figure 5 - Scatter Plot for SNE-4**

Interestingly, no such intuitive correlates could be arrived at on the 4 dimensions discovered using PCA. (Refer to additional plots attached along with this report.)

The tree form has a higher score than any other form except partitions in SNE-4 (i.e. SNE with a reduction to 4 dimensions), but that difference is expected, because reducing the dimensions of data would naturally favour a simplistic form more. In fact, the nonlinear dimensionality reduction method that our approach uses to operationalise gist-extraction induces a sort of clustering over the reduced feature space [7], making the algorithm of Kemp et al. favour a partition form disproportionately. Because of the kind of complications that arise above with the relative scores of various forms, it was decided not to rely on their log-likelihood scores as a metric of the accuracy of form-learning after dimensionality reduction.

Another validation of SNE as a tool for gist extraction is the fact that it preserves (across a reduction of dimensions) the clusters that form learning predicts on the data. To this end, hierarchical agglomerative clustering is performed separately on ground data and the reduced-dimensionality data (SNE-2) to analytically separate data points into clusters for both of them. Upon reducing dimensionality, not only are the clusters preserved, the relative centres of clusters are also largely preserved. The latter has been measured by assigning the most central sample in every cluster as a category representative. As shown below, out of the 8 category representatives, the original dataset and SNE-4 data match on 6 of them, whereas the original and PCA-4 data match on only 2 of them. The rationale for using the most central samples of a cluster stems from the representativeness heuristic mentioned above. Brains identify various clusters by their most typical members, and the fact that SNE largely preserves clusters and categories points to the fact that it preserves the semantics of our conceptualisation of categories.

This categorisation was taken a step further and isolated the discriminating features of each cluster. A feature is called a discriminating feature for a cluster if it is switched on (or off) only for samples belonging to that cluster and is switched off (or on) for all other cluster samples. We have identified 36 such discriminating features around the 8 clusters that were segregated using HAC on data after reducing it by SNE-2 [Table 4]. These features thus characterise on what basis does SNE categorise different samples, and thus unravels the underlying clustering structure. It is observed that a majority (23 out of 36) of these features are, in fact, anatomical and specifically features of visible anatomy (22 out of 23). In the original set, only 59 out of 102 features were anatomical, but more importantly, within them, 49 were features of visible anatomy [Table 5]. Equally significant is the reduction in the fraction of behavioural features from ground data (24 out of 102) to the reduced-dimension data (5 out of 36).

| S.No | List of discriminating features | Type of feature |
|---|---|---|
| 1 | has a large brain | Anatomical (visible) |
| 2 | has 6 legs | Anatomical (visible) |
| 3 | has a nose | Anatomical (visible) |
| 4 | has paws | Anatomical (visible) |
| 5 | has antennae | Anatomical (visible) |
| 6 | is long | Anatomical (visible) |
| 7 | is large | Anatomical (visible) |
| 8 | has tusks | Anatomical (visible) |
| 9 | is slender | Anatomical (visible) |
| 10 | has horns | Anatomical (visible) |
| 11 | has hooves | Anatomical (visible) |
| 12 | is poisonous | Anatomical |
| 13 | is soft | Anatomical (visible) |
| 14 | is black | Anatomical (visible) |
| 15 | is a rodent | Anatomical (visible) |
| 16 | has webbed feet | Anatomical (visible) |
| 17 | is a feline | Anatomical (visible) |

| | | |
|---|---|---|
| 18 | is an insect | Anatomical (visible) |
| 19 | is scaly | Anatomical (visible) |
| 20 | is furry | Anatomical (visible) |
| 21 | has flippers | Anatomical (visible) |
| 22 | is colorful | Anatomical (visible) |
| 23 | is a canine | Anatomical (visible) |
| 23 | is strong | Behavioural |
| 25 | howls | Behavioural |
| 26 | travels in groups | Behavioural |
| 27 | is dangerous | Behavioural |
| 28 | digs holes | Behavioural |
| 29 | eats grass | Eating habits |
| 30 | eats leaves | Eating habits |
| 31 | eats bugs | Eating habits |
| 32 | eats fish | Eating habits |
| 33 | lives in lakes | Habitat |
| 34 | lives in ocean | Habitat |
| 35 | lives in water | Habitat |
| 36 | lives in houses | Habitat |

**Table 4 : List of discriminating features and their types**

| Fraction of total | All features | Discriminating features |
|---|---|---|
| Anatomical | 0.57 | 0.64 |
| Visible anatomical (as fraction of anatomical) | 0.83 | 0.96 |
| Behavioural | 0.24 | 0.13 |

**Table 5 : Fractions of different kinds of features in all and discriminating features**

This demonstrates something very interesting about what kind of features we give more importance to in categorising objects. Anatomical features, which are visible and hence are easier to store and retrieve from memory, play a disproportionately larger role in how humans reason about categorisation. Conversely, behavioural features, which are not as easy to retrieve and store, get downplayed in object categorisation. This is a pertinent example of the availability heuristic at play. What feature are identified as discriminating one category from another is coloured by the ease of retrieval of that feature in the cognitive system. The correlation of ease of retrieval with the degree to which a word arouses an image has been demonstrated by Paivio in [11]. This ability to explain availability heuristic is a case for SNE-induced topology as a better model of human cognition than was earlier believed. A more concrete evidence of the parallels would require a more theoretical and topological understanding of the workings of SNE. These manifolds seem to not only answer how concepts are represented, but also what sort of irrationalities creep in because of the approximations that made and relied upon, in the interest of conserving cognitive resources.

# Tying it all together - Manifolds, Typicality, Bayesian Cognition and Form Learning

Having established the power of manifolds in better understanding human cognition in a compact representation setting, we hoped to merge our analysis with [3]'s framework, but it was found to be impossible to reconcile the generality of intuitive "gist" features with the generality of a grammar over various forms. For instance, there is a steep tradeoff when reducing dimensionality, between preserving a better likelihood of the true form vis-à-vis other forms like partitions and keeping the features simple/generic enough to be considered as corresponding to what is stored in semantic memory. Some possible future work could look at trying to incorporate semantic memory considerations into the probability distribution of various forms. This probability distribution could be conditioned on something that achieves the gist extraction that has been implemented in our work through a topological analysis on the dataset of [3]. But the problem in using this as a hyperparameter is that of introducing circularity, of using gist features derived from the entire data to predict its structure. How this gist knowledge is updated when humans gather more data and experience, is something that can be modelled, and a framework that bootstraps this gist knowledge to probability distributions over structural forms, could reconcile the two.

It is seen that, for this particular dataset, SNE-16 actually improves the accuracy of learning the correct tree structure than if the same is learnt directly from ground data. To substantiate this claim, a **cognitive study** was conducted**,** wherein 100 respondents were presented with two images: Image1-SNE-16-Tree and Image2-Ground-Tree. The survey results (See Figure 6 below) convey that a significant majority of all respondents (40/100) were indifferent to the two animal-tree representations, another sizeable chunk (38/100) preferred the SNE-16 version, while a minority (22/100) preferred the ground version[1].

[1] Study results can be found here:
https://docs.google.com/forms/d/1BDpFLn-eVkF35FHtZq02N6xAwjMMjWYGbP96PFqL6ZI/viewanalytics

Pick the tree that, according to you, is the best representation of animal relationships.


**Figure 6 - Results of the Cognitive Study; Image1 refers to SNE-16 tree version, Image2 refers to Ground tree version**

In the following section, we elaborate on how this approach can be tied back to the Bayesian framework of cognition. Belief updation in a Bayesian model takes the form of updation of priors to posteriors, and these posteriors acting as priors for further observations. Humans start off by learning simple categorisation over a simple manifold over few dimensions. As more and more categories are observed, the mental representation of a cluster is slowly substituted by a singular sample, which is the most typical member of that cluster. A new manifold is then learnt over these category representatives along with the new data points. This new manifold, though more complex than the initial one, is still robust due to the smoothing out of category outliers.

In essence, what is being suggested is a Full Bayesian Learning approach to form learning, rather than the Maximum a Posteriori (MAP) approach with equal-priors used by Kemp et al. In Bayesian Learning theory, learning can be done in three ways, as tabulated below [Table 6]. In the limit of infinite data, ML estimation suffices for a good estimation of parameters $\vartheta$. However, as amply elaborated above, human cognition must make sense from few data points. Thus, we must move to MAP, and eventually to Full Bayesian estimation, wherein the priors on $\vartheta$ are adjusted over time.

| **Learning Method** | **Key Idea (X is data and $\vartheta$ are parameters)** |
| --- | --- |
| Maximum Likelihood (ML) | max $P(X/\vartheta)$ |
| Maximum a posteriori (MAP) | max $P(X/\vartheta).P(\vartheta)$ |
| Full Bayesian Learning | mean $P(X/\vartheta).P(\vartheta)$ |

**Table 6 : Various Bayesian Learning Methodologies**

The representativeness heuristic preserves the robustness of learning manifolds (and updating our beliefs around them on observing novel data) by hinging our conceptualisation of categories around category representatives. A more holistic model can perhaps fit into this analysis the transformation of what is an outlier at first to a new category in itself with observation of other similar objects. This holistic model is illustrated below with an example.

Consider a child who has never learnt anything about "animals". Thus initially, one can assume as Kemp et al. do, that the child has a uniform prior distribution on all 8 forms in which this animal knowledge base is compactly represented. Now once a **manifold** (and the eventual form and structure) of these animals is learnt, the form learning model returns likelihoods of the forms, which should be set as the prior distribution for the next batch of learning for the child. Notice how doing this essentially attaches a hyperparameter to our **Bayesian** model. Assume a number of iterations have passed, and now a manifold embeds animals in the child's cognitive system. Now when a new animal is observed by the child, how will the child embed this into the manifold? Intuition suggests that the child compares compact features of the new animal to the discriminating features of a cluster on the manifold, that is, relying on representativeness and availability heuristics, hence showing the **typicality** effect and a bias towards ease of retrieval. Embedding of this new animal on the manifold could have essentially three effects: (a) no effect to the category representatives, and thus to the manifold, (b) a slight shift in the category representations (c) a big dent on the manifold, which can possibly lead to a paradigm shift in the **form and structure learnt** on the entire animal knowledge base and thus on the new priors of these forms. Note how the idea of a "learnt manifold" translates to having learnt the form-priors and the category representatives in our model. Examples of each of these three outcomes could be: (a) introduction of "tortoise" to a system which already has "turtle" embedded, (b) introduction of egg-laying "platypus" to a system which has "mammals" embedded, (c) introduction of "bridge" animals such as "pterodactyls", which could bridge "reptiles" to "birds" and induce a relationship among different categories, moving the prior away from a partition-form and towards a tree-form.

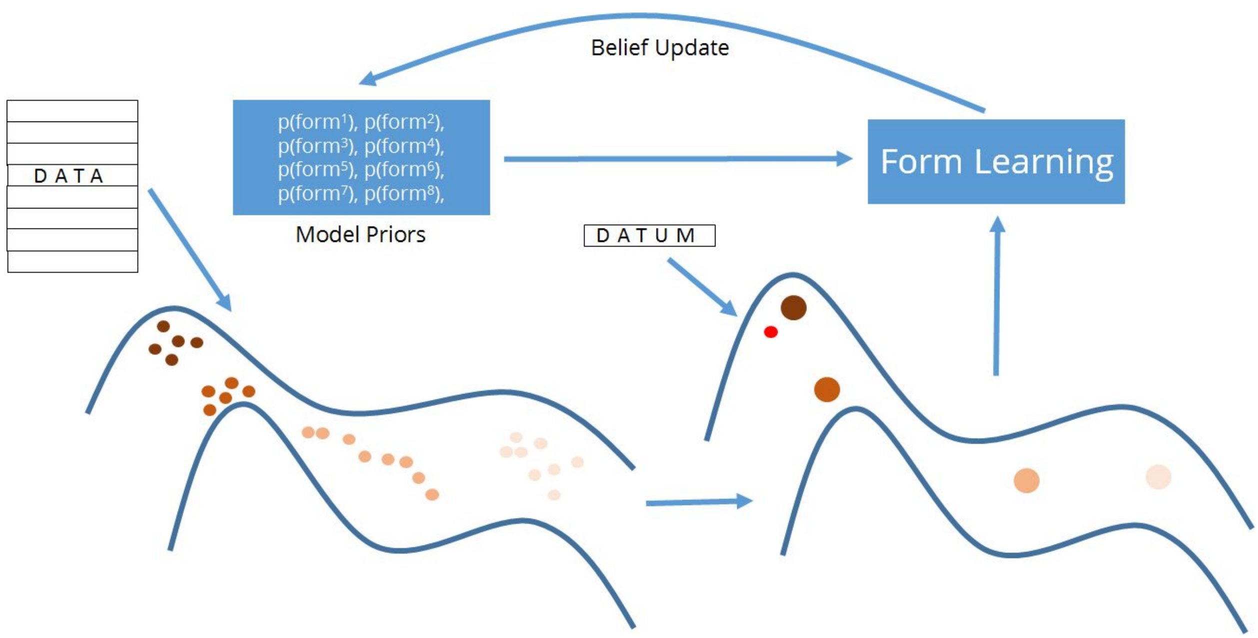


**Figure 7 -A schematic of the suggested extension to Kemp et al.'s cognitive model for Form Learning, as explained in this section**

So how does this inductive learning take place? Illustratively, one can imagine that if a child knows of only dogs and cats, he/she might only care for the feature "barks/meows", thus forming partitions, structured as two clusters. The manifold too here is very simple, since dimensions are very low. But as more animals are observed, the features increase, the manifold becomes more complicated, and priors of more complicated forms become larger. And for inference, the suggested category-representative comparison approach is used to fit where the newer animals lie on the manifold. The ease of retrieval is another way by which our cognitive systems avoid complex manifolds in favour of simpler ones.

In conclusion, the primary contributions of our paper hint at use of techniques which better explain the kinds of heuristics used in human cognition, learning and inference, under bounded constraints. Namely, techniques from manifold learning, which reduce the feature space into a compact form, without losing out on cluster and representativeness information. Then, the incorporation of the typicality effect into Kemp et al.'s model of form learning has also been suggested, by presenting an extension to their idea via inclusion of hyperparameters to the Bayesian model in the shape of prior probabilities of the forms, which are learnt along with the manifold as and when the agent encounters more data which gets embedded in that manifold. It has been shown that in addition to modelling how simplifying approximations are made in reasoning and remembering about the world, this model also seems to make some progress in capturing the inaccuracies (and deviations from perfect rationality) that arise due to these assumptions.

# Appendix

## Stochastic Neighbour Embedding

Stochastic Neighbour Embedding (SNE) is a probabilistic nonlinear method of placing objects described by high-dimensional feature vectors into a low-dimensional space while preserving neighborhood information. It works by centering a Gaussian distribution on each object in the high-dimensional space. The densities under this Gaussian define a probability distribution over neighbors of the object. The embedding aims to approximate this distribution on the low dimensional space, when the same operation is performed on the latter. A cost function is then employed for a simple gradient that adjusts positions in the reduced space. Unlike other dimensionality reduction methods, especially linear ones, this probabilistic framework simplifies representing each object by a mixture of widely separated low-dimensional 'images'. [7]

## Principal Component Analysis

Principal component analysis (PCA) is a linear dimensionality reduction procedure that uses an orthogonal transformation to convert/reduce a set of observations of variables (potentially correlated) into a set of values of linearly uncorrelated variables called principal components, which are orthogonal because they are the eigenvectors of a symmetric covariance matrix. The first principal component has the largest possible variance and accounts for as much of the variability in the data as possible. Each succeeding component has the highest variance possible constrained to orthogonality with preceding components. [8]

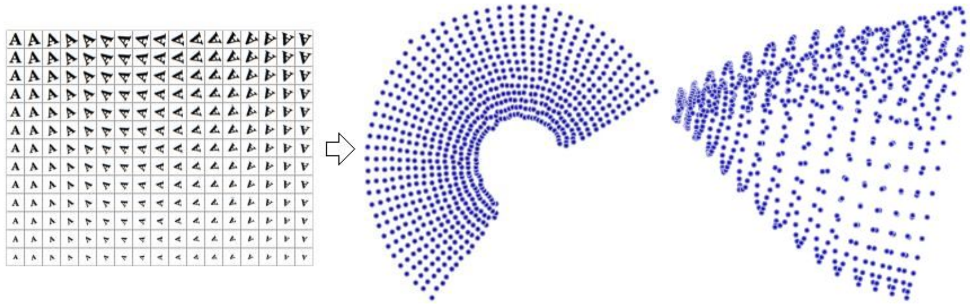

**Figure 8 - On a scaled and rotated dataset of "A"s, the (a) manifold dimensions (two) clearly expose the 2D nature of the data (radius and angle), while (b) PCA dimensions (two) are unable to discriminate effectively;**
**Source: Wikipedia/Nonlinear_dimensionality_reduction [9]**

## Learning Structural Forms

Using a graph grammar over cluster graphs, Kemp et al. have managed to represent many of the cognitively natural structures of categories in a simple framework. They assign uniform priors to 8 structural forms, and iterate over all of these forms. Within a form, they use the production rule for that form to randomly vary cluster partitions and greedily choose the best cluster graph within that form, the one that maximises the following probability. This is the joint probability of the structure and form that explain given data. Data is more likely to come from a structure if the features of that data vary smoothly over that structure. Since the production rules are exceedingly simple, this model is an intuitive framework for understanding how cognition performs categorisation. [3]

$$P(S, F|D) \propto P(D|S)P(S|F)P(F)$$

## Hierarchical Agglomerative Clustering

Hierarchical agglomerative clustering (HAC) is a method of unsupervised clustering which takes the bottom-up approach. It starts by initialising all data points as clusters. Then, depending on some distance metric (say Euclidean), it clusters pairs of clusters together iteratively, till the entire dataset becomes a cluster in itself. The distance between two clusters could also be done in a variety of ways, namely distance between averages, closest, farthest, etc. Inherently, this technique imposes a hierarchical structure on the data, and thus, the clusters can be neatly visualised as dendrograms, where the data points are the leaves of the dendrograms, and edge lengths represent distance between the clusters.

## Measuring Tree Similarities

Measuring graph similarity is an NP-hard problem, which involves establishing a polymorphism between two graph structures. However, neither do we wish to tackle an NP-hard problem, nor should we, since the two tree structures need only be similar in the way the leaves relate to each other, and not the entire trees themselves (note that the difference between a hierarchy form and a tree form is in whether there are clusters at non-leaf nodes). Therefore, to estimate a measure of similarity between two trees, we employ a simple heuristic. We find the shortest distance between pairs of all leaves in the two trees, and take an absolute difference, sum them over, and proportion it by the sum of shortest distances in the ground-truth tree. Higher is the value of this metric, the more dissimilar are the two trees.

# References


1. Edward Vul, Noah Goodman, Thomas L. Griffiths, Joshua B. Tenenbaum, *One and Done? Optimal Decisions From Very Few Samples*, Cognitive Science 38 (2014) 599–637, DOI: 10.1111/cogs.12101 https://stanford.edu/~ngoodman/papers/VulEtAl2009.pdf
2. Wikipedia: https://en.wikipedia.org/wiki/Topological_data_analysis
3. Charles Kemp, Joshua B. Tenenbaum, *The discovery of structural form*, PNAS, August 5, 2008, vol. 105, no. 31, 10687–10692, www.pnas.org/cgi/doi/10.1073/pnas.0802631105
4. Daniel Kahneman, Amos Tversky, *Subjective Probability: A Judgment of Representativeness*, Cognitive Psychology 3, 430454 ( 1972) http://datacolada.org/wp-content/uploads/2014/08/Kahneman-Tversky-1972.pdf
5. Thomas Gilovich, Kenneth Savitsky, *Like Goes With Like*, https://web.viu.ca/burnleyc/Psychology%20331/Like%20Goes%20with%20Like.pdf
6. Walter Kintsch, *The role of knowledge in discourse comprehension: A construction-integration model*, Psychological Review (1988), Vol 95(2), 163-182.
7. Geoffrey Hinton, Sam Roweis, *Stochastic Neighbour Embedding* https://www.cs.nyu.edu/~roweis/papers/sne_final.pdf
8. Wikipedia: https://en.wikipedia.org/wiki/Principal_component_analysis
9. Wikipedia: https://en.wikipedia.org/wiki/Nonlinear_dimensionality_reduction
10. Anthony Esgate, David Groome (2005). *An Introduction to Applied Cognitive Psychology*. Psychology Press. p. 201. ISBN 978-1-84169-318-7.
11. Allan Paivio, *Mental imagery in associative learning and memory*. Psychological Review, Vol 76(3), May 1969, 241-263. http://dx.doi.org/10.1037/h0027272