A Critique of Pure Hierarchy: Uncovering Cross-Cutting Structure in a Natural Dataset*

J. L. McClelland
Department of Psychology, Stanford University
Stanford, CA, 94305, USA
E-mail: mcclelland@stanford.edu
www.psych.stanford.edu/~jlm

Z. Sadeghi
School of Electrical and Computer Engineering, University of Tehran,
Tehran, Iran
E-mail: zahra.sadeghi@ut.ac.ir

A. M. Saxe
Center for Brain Science, Harvard University
Cambridge, MA 02138, USA
E-mail: asaxe@stanford.edu
www.stanford.edu/~asaxe

How best can we understand – and visualize – the structure in multi-dimensional data? One common approach is to rely on hierarchical cluster analysis, either for theoretical or for more descriptive reasons. Here, we point out that an apparently revealing hierarchical clustering solution may well be compatible with structure that is not well characterized as a hierarchy. In particular, a hierarchical description can be equally consistent with cross-cutting rather than strictly hierarchical, or nested, structure. We offer an alternative approach, based on inspection of the feature vectors provided by a singular value decomposition (SVD) which allows a flexible mixture of hierarchical and cross-cutting dimensions and which can reveal whether dimensions are cross-cutting or nested. The SVD offers a more flexible representation than a hierarchy in that it can capture either hierarchical or cross-cutting structure or blends of these two structure types, or, indeed, many other structure types. We then introduce a refinement of the SVD approach based on sparse principal component analysis that leads to more easily interpretable dimensions. In our dataset, these dimensions correspond to aquatic vs. land animals, large vs. small animals, predators vs prey animals, and primates vs. other mammals.

Keywords: Semantic knowledge representation; Singular value decomposition; Sparse principal components analysis; Deep neural networks.

* This work is supported in part by a research fellowship from IROST to Z.S. and by fellowship support to A.S. from NSF IGERT Grant 0801700.

1. Background

How do humans represent knowledge of different types of items and their properties? One frequent approach is to assume this knowledge is represented in terms of a hierarchy. In some cases, a hierarchy may be used as a summary presentation of the structure in data, without a strong theoretical commitment. Other times, a hierarchy may be asserted to capture the knowledge that people use in making, for example, similarity judgments or property attributions[1]. In either case, we will suggest, a hierarchical presentation can be misleading if interpreted too literally, and it is at best ambiguous if the limitations of the hierarchical representation are understood. We discuss these issue and propose alternative ways to visualizing structure in data, focusing on a data set that has been used in several studies of the organization of semantic knowledge in the domain of mammals.

Our work draws on a body of prior work culminating with Kemp and Tenenbaum[1], and Glick[2,3]. Kemp and Tenenbaum found that a hierarchical tree structure provided a better fit to human property attribution data from several experiments in the domain of mammals compared to other structures they considered, but the Glick studies found that a hierarchy both failed to fully reflect the similarity structure in the data set and also missed aspects of human similarity and property attribution judgments. The evidence suggests that participants are sensitive to structure that cuts across the branches of the best-fitting taxonomic hierarchy, and is not adequately reflected within it. The focus here is not on an effort to contrast overall theoretical frameworks (Kemp and Tenenbaum's overall framework can be used to capture cross-cutting structure; see Shafto, Kemp, Mansinghka & Tenenbaum[4]), but on the effort to bring broader awareness to the limitations of a hierarchical representation of data. Our approach to overcoming these limitations reflects our interest in exploring ways to characterize structure that may be *quasi-regular,* and thus not fully consistent with *any* specific structure type[5], but the approach is ultimately completely empirical, and can be used to explore the structure in a data set regardless of any prior theoretical commitments.

Neural networks of the kinds we have often used in models are capable of capturing quasi-regular structure, thereby reproducing patterns of human behavior in several quasi-regular domains, such as single word reading and knowledge of objects and their properties. A limitation of the approach, however, is that knowledge in this form is stored in connection weights, and is often hard to intepret[6]. Recently, we have developed a mathematically explicit characterization of the knowledge learned in deep neural networks in terms of a weighted combination of aspects of the structure in a data set as revealed by singular value decomposition[7] (SVD). Our focus here is not on the neural network analysis, but on the SVD itself as a method for representing structure. We show how an SVD can be used to make apparent the underlying cross-cutting or nested structure in a data set that would be lost under a hierarchical clustering analysis, and we discuss an extension of the approach that can further increase the interpretability of the structure it uncovers in natural data.

In what follows, we first consider the data set that will be the focus of our analysis, and point out why no single hierarchy can adequately characterize the knowledge contained within it. Then, we describe how *both* the structure that the hierarchical model captures and the structure that it misses can be captured by a singular value decomposition of the data set. We then note that the singular value decomposition still lacks a degree of interpretability, leading us to explore a variant of the approach based on sparse principal components analysis (SPCA) that produces a more easily interpretable decomposition. The *discussion* section considers the broader implications of our findings for the interpretation of experimental observations and for alternative theoretical frameworks on the nature of human knowledge representations.

2. Dataset and limitations of representing it hierarchically

The data set that will be the focus of our analysis, here called the *50 mammal set*, came originally from Osherson *et al.*[8] and was augmented by Kemp and Tenenbaum[1], who provided it to us. It is worth noting at the outset that this data set is not objective data, but is instead a characterization of human knowledge, so that the effort to discover which sort of representation best characterizes this data set is an exercise in modeling human knowledge, not simply an exercise in modeling facts about objects in the world.

The data set was obtained by asking participants to rate the applicability of each of 85 different predicate terms to each of 50 different mammals, then taking the mean over participants of the ratings in each cell of the 50 animal by 85 property matrix. Figure 1 displays the matrix of correlations of the property vectors of the 50 animals. The items are arranged according to a generic hierarchical clustering solution displayed in Figure 2. That is, items within the same cluster are near each other. If the hierarchy adequately captures all of the structure in the training data, then, similar items should be near each other, and dissimilar items should be farther away. It should be evident from Figure 1 that the hierarchical clustering captures many strong similarity relations (reflected by dark red color near the main diagonal) but also misses many others (dark red colors *not* near the main diagonal). For example, the weasel (21st row or column of the array) is similar to, and grouped with other small mammals (raccoon, rat, weasel, squirrel, etc.), and it shares much in common with them. However, the weasel is almost as similar to the fox and some of the other predators as it is to the small mammals it is grouped with, a fact that is not captured in the hierarchical solution in Figure 2.

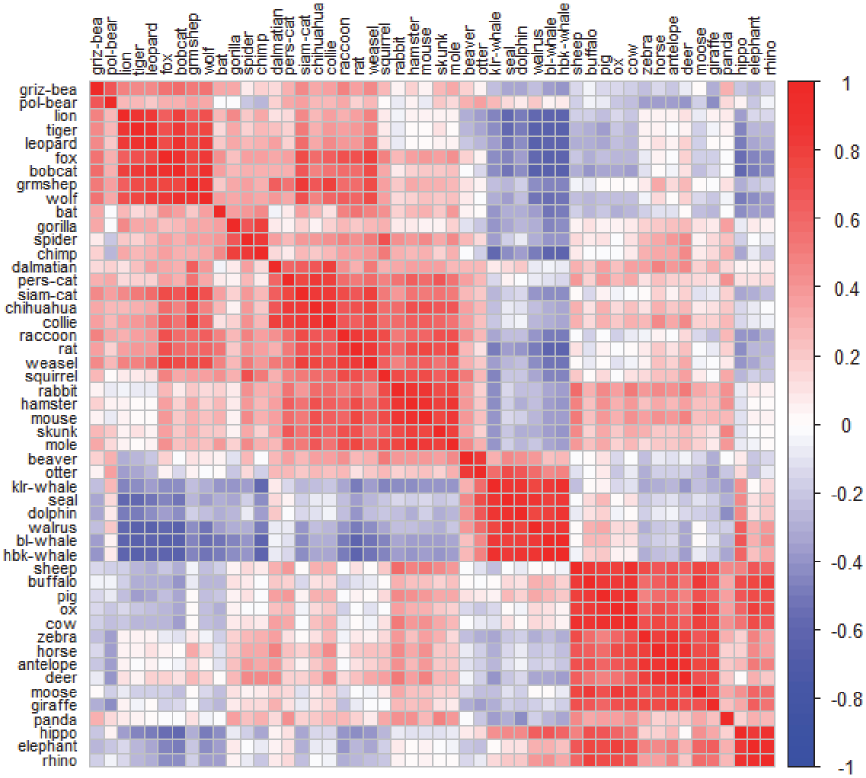

Fig. 1. Correlation matrix for the 50-mammal data set used in Kemp & Tenenbaum[1]. Ordering is based on the hierarchical clustering solution shown in Fig 2.

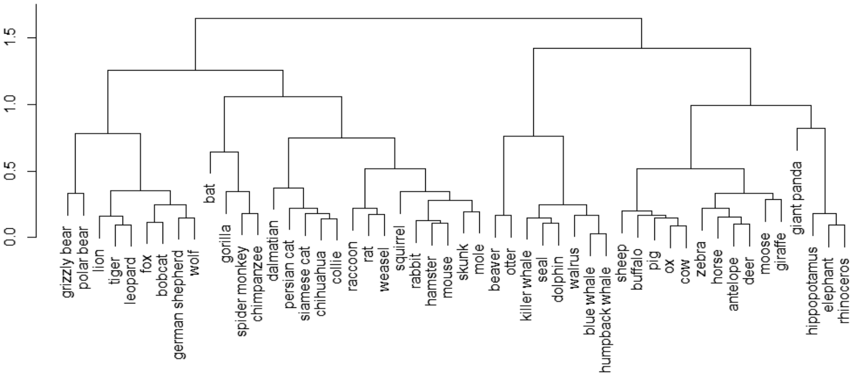

Fig. 2. Hierarchical clustering of the 50 mammal set, obtained with R's **hclust** function. Distances are 1-C, where C is the correlation of the two items' vectors of Manthattan distances to other items.

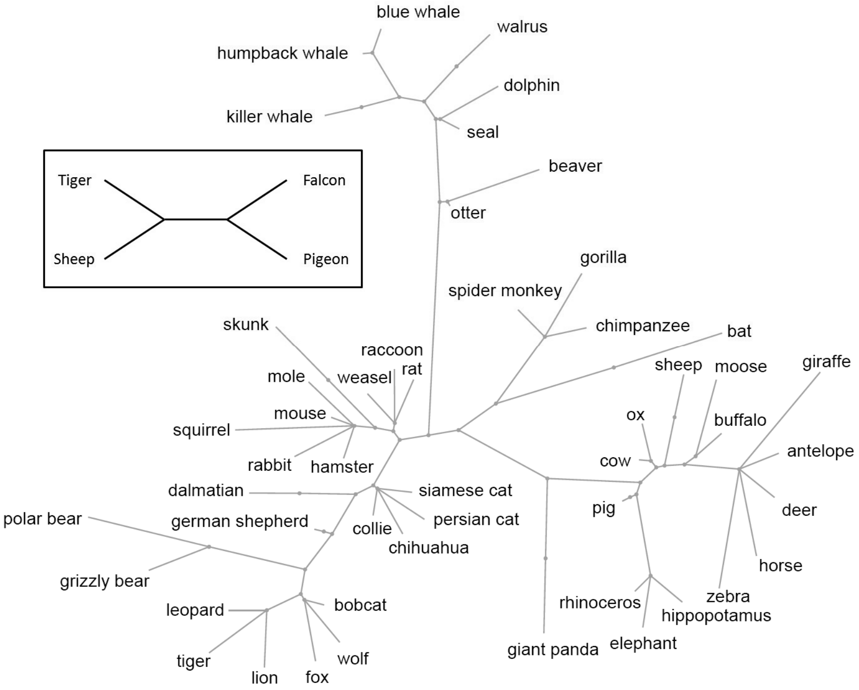

Fig. 3. The best fitting hierarchical diffusion model reported by Kemp and Tenenbaum[1], Fig 3, p. 28, Copyright 2009 American Psychological Association. Reprinted with permission. Inset shows the structure of the minimal hierarchical model that would be fit to both data sets shown in Fig 4. The structure is similar to that in Fig 2, but the bat and primates placed slightly differently.

No single hierarchy can capture the similarity of the weasel to both the small mammals and the predators at the same time, as the weasel can only be placed in one cluster, while (for example) the squirrel and fox belong in these two different clusters. The hierarchical tree is thus a Procrustean bed for this data set, forcing items to fit into a structure that does not suit them well. Figure 3 presents the similar hierarchical solution found by Kemp and Tenenbaum[1]. Their solution is based on a probabilistic property diffusion process, such that properties of items constrain each other by an amount that depends only on the length of the connecting path between items. This length can be directly measured from the graph in Figure 3. It will be seen that their solution has the same difficulty as the more generic hierarchical clustering solution shown in Figure 2. Their model is also a hierarchical model, in that items are connected by a single branching network, such that there is only one path between each pair of animals (there is no privileged root of their tree, but one might imagine one near the center of the figure, where the main vertical branch rises from the main left-right axis). The weasel is again placed close to the other small mammals, forcing the distance to the fox and other predators to be long.

Glick[2,3] pointed to these aspects of the data and went on to show that participants' judgments of the likelihood that a biological property attributed to one animal would also hold of another animal were not fully accounted for by Kemp and Tenenbaum's hierarchical tree. Such judgments were obtained in a property attribution task, which presented a putatively generic premise attributing a fictitious property to a familiar class of objects, such as 'Horses have enzyme X132 in their bloodstream'. Participants then rated the likelihood that other target animals would also have this property. Kemp & Tenenbaum[1] found that their hierarchy provided a better account of previously existing data on such attributions on several subsets of mammals compared to the raw similarities among these items in the 50-mammal dataset. However, Glick used a different subset of animals, including the rat, the fox, and the weasel, and found that participants' ratings did not adhere well to the predictions of the hierarchical model. Glick also obtained similarity judgments between all possible pairs of items in a data set consisting of 50 mammals, 30 birds, and 20 fish. While a hierarchical model with separate branches for mammals, birds, and fish captures the strongest dimensions in this data, it fails to pick up on other factors that are clearly present in the similarity ratings. For example, a mammal and a bird were judged more similar if they were similar in size or ferocity. This similarity cannot be captured in a hierarchical tree, given that all the birds are on one branch of the tree and all of the mammals are on the other. In short a tree has no way of revealing cross-cutting similarity.

To state the limitations of a tree-like structure more formally, only the average value of a property can propagate between sets of items at different ends of a single connecting pathway. In the hierarchy, birds and mammals are on different branches connected by a single pathway (as in the minimal hierarchical model shown in the inset in Figure 3), so that only average values of birds can propagate to mammals and *vice-versa.*

3. An alternative account of semantic structure

As an alternative to a hierarchical approach, we offer the view that semantic structure might best be captured by a more flexible system of representation that can be sensitive to multiple types of structure that may be present in a data set. Our scheme grows out of our past efforts to understand the knowledge learned in deep (i.e., multi-layer) neural networks. Importantly, in our current approach, the knowledge is characterized at a level more abstract than any particular neural network, bringing out the underlying structural relationships present in the data set. The approach relies on the concepts of singular value decomposition (SVD) and the closely related concepts of principal component analysis. Singular value decomposition is also used in the well-known Latent Semantic Analysis approach to semantic representation[9]. We apply it here, not to matrices of word co-occurrences in texts, but to the item-feature matrix, to directly capture the semantic relations among items based on their properties. We focus on an intuitive presentation; for technical details, see Saxe *et al.*[7].

According to our approach, the content of a data set is characterized in terms of a set of underlying dimensions. Each item is placed at a point along each of the dimensions, and each dimension is associated with a vector of features. The placement of an item on a dimension can be considered a classification scheme. For example, for the toy data set in Figure 4 (top) one classification vector would classify items on a bird-mammal dimension (positive end = mammal, negative end = bird). Associated with each classification vector there is a corresponding feature vector. The bird-mammal feature vector assigns high likelihood to *four legs* and *fur* and low likelihood to *wings* and *feathers* for mammals, and the reverse pattern for birds. The total strength of this relationship, expressing how much of the overall variation in the item-property matrix it accounts for, is captured by the strength of the singular value (SV) associated with this dimension (Figure 4, top center).

A virtue of this approach is that it allows for other dimensions either to be hierarchically embedded within the classification captured by a stronger dimension or to cut across such dimensions. In the data set on the top left of Figure 4, we see that there are two predators and two pray animals, and in each

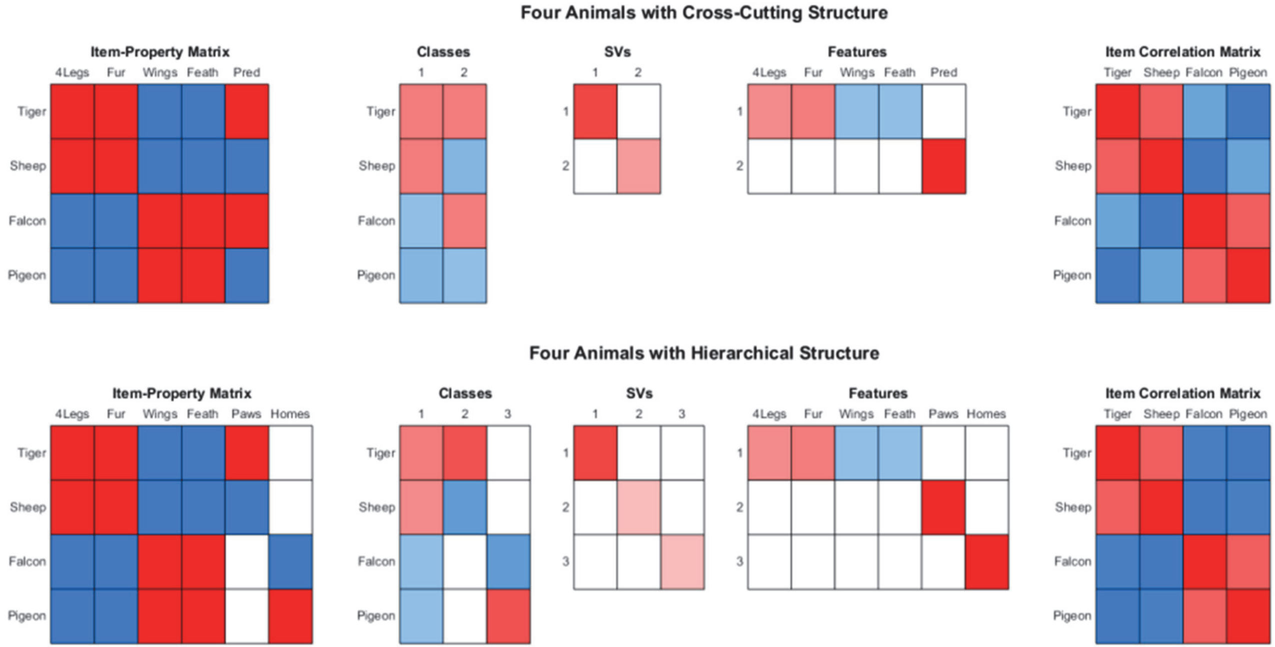

Fig. 4. A data set with cross-cutting structure (top row) and another with hierarchical structure (bottom row). Left panels are the data sets themselves. Singular value decomposition (middle panels) brings out the differences in the structure, revealing a second cross-cutting dimension in the first case and two embedded dimensions in the second case. These different structures are associated with distinct correlation matrices (right panels) revealing cross-cutting structure in the top row but not in the bottom row. A hierarchical model like that of Kemp and Tenenbaum cannot capture these cross-cutting similarities in the data set shown in the top row, and would produce a correlation matrix like the one associated with the truly hierarchical data set in both cases.

case one is a mammal and the other is a bird. Thus predator status is a cross-cutting dimension. It can be represented as a single additional classification dimension (positive for predators, negative for prey) and a corresponding feature vector. In this simplified example, there is a single *predator* feature, but the scheme will accommodate sets of features such as *has sharp claws, eats animals*, etc., that tend to occur in predators but not in non-predators.

Alternatively, additional dimensions can be embedded within each of the higher-level dimensions, as represented in the alternative data set shown in the second row of Figure 4. In this case, one feature distinguishes among the two mammals (e.g., *has paws*), while the other distinguishes between the two birds (*flies home* or *homes*). What is critical here is that additional dimensions can either be cross-cutting or they can be further sub-categorizing dimensions. In the latter case, they show up in the SVD as two separate dimensions, each with its own classification and feature-specification vector. Thus, the SVD is capable of distinguishing these two types of structure. However, a hierarchical clustering analysis will necessarily treat both kinds of dimensions the same, failing to distinguish the two types of structure. Specifically, the two different SVDs each fully capture the structure present in the corresponding data set, as illustrated by the correlation matrices reconstructed from each SVD on the right of each row of the figure. In contrast, the hierarchical tree for both data sets has the structure shown in the inset to Figure 3 with a corresponding correlation matrix like that appropriate for a data set where the structure is truly hierarchical, as in the second row of Figure 4.

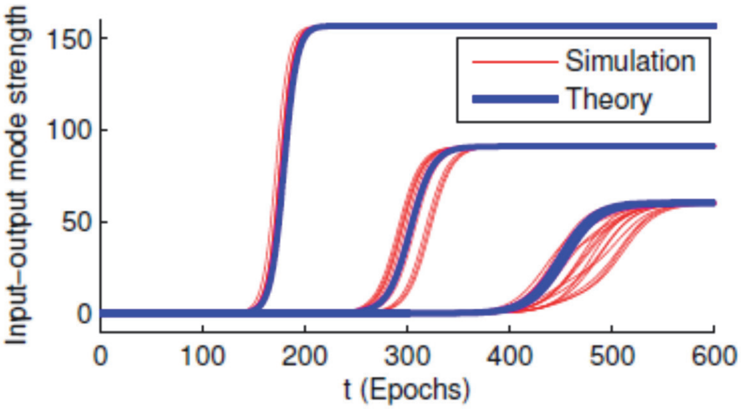

Fig. 5. Time course of learning in a linear network trained with a set of item-feature pairs characterized by three singular dimensions, each with different strengths. From Figure 3, p. 1274 of Saxe, McClelland and Ganguli[7]. Copyright 2013, the authors. Reprinted with permission.

In addition to its use in flexibly capturing these and other types of structure which can be mixed within the same data set, the SVD approach can also be linked to the gradual learning process that occurs in a deep neural network. Rogers and McClelland[10] studied this process as it arises in the semantic network model introduced by Rumelhart and Todd[11]. A subsequent mathematical analysis by Saxe, McClelland and Ganguli[7] demonstrated that the learning dynamics exhibited by such networks can be well-captured with a simplified (i.e., linearized) version of the same deep network. Space prevents a detailed review here; suffice it to say that the time course of learning in both the original, non-linear network and the simplified linear version of the network, when trained on a given set of training examples, is well captured by the statement that the learned connection weights learn the singular dimensions that characterize the data set in a stage-like fashion, as shown in Figure 5. Quantitatively, the strength of the representation of each singular dimension increasing with time according to a highly *non-linear* function:

$$s(t) = (Se^{2St/\tau})/(e^{2St/\tau} - 1 + S/s_0\ )$$

where $s(t)$ represents the strength of the dimension after time t, S represents the strength of the underlying singular value, s_0 represents the initial strength of the dimension (which will have a small random value in a randomly initialized neural network), and τ is a time constant depending on the network's learning rate. According to this function, shown for a training set characterized by three singular values in Figure 5, each singular dimension is acquired after a time that depends on the size of its singular value. This kind of pattern is similar to the stage-like progression seen in many developmental transitions[12] and in non-linear neural networks like those applied to development by Rogers and McClelland[10] or McClelland[13].

4. Uncovering the structure in the 50-mammal data set

A difficulty with singular value decomposition when applied to a natural data set (such as the 50-mammal data set) is that the resulting dimensions are not necessarily easily interpretable. While these dimensions may be the true dimensions that characterize the structure in the data and are the ones the deep network successively extracts, they do not necessarily correspond to easily interpretable aspects of the underlying data. We do not find this problematic, since we expect natural structure to be *quasi-regular,* rather than truly systematic[14,15]. We can still ask, though, how cognitively interpretable dimensions might arise from experience with a set of items. One possibility is that such dimensions are constrained to account for strong patterns of co-variation present in a body of experiences while relying only on a small number of features. The motivation for this might either be cognitive economy or simplicity in describing these dimensions to others. To explore this possibility, we explored the dimensions discovered by sparse principal component analysis[16] (SPCA). Our explorations employed a variant of SPCA that extracts components sequentially (**spca** function from the SpaSM toolbox[17]).

First we note how standard principal component analysis (PCA) relates to the SVD approach we have described thus far. A standard PCA when applied to a data set such as ours returns a set of *coefficient vectors* that each correspond to the *feature vectors* returned by the SVD of the same data set, as well as a set of *score* that play the role of the *classification vectors* in the SVD, but with each score vector equal to the corresponding classification vector of the SVD scaled by the corresponding singular value. Each successive coefficient vector is chosen to be orthogonal to all previous co-efficient vectors and to maximize the amount of residual or remaining variance it explains.

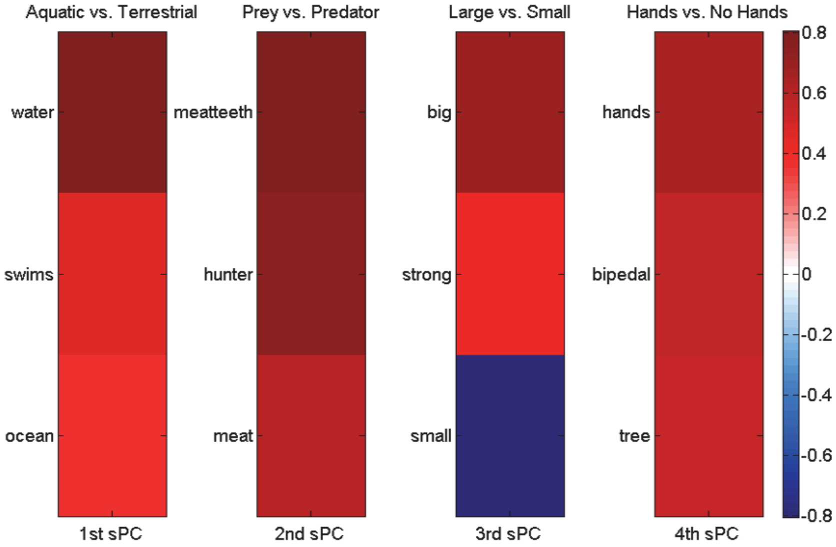

Fig. 6. Feature loadings on the first four sparse coefficient vectors found by the **spca** function (K >= 4, delta = 5, stop = -3) of Sjöstrand *et al.*[17].

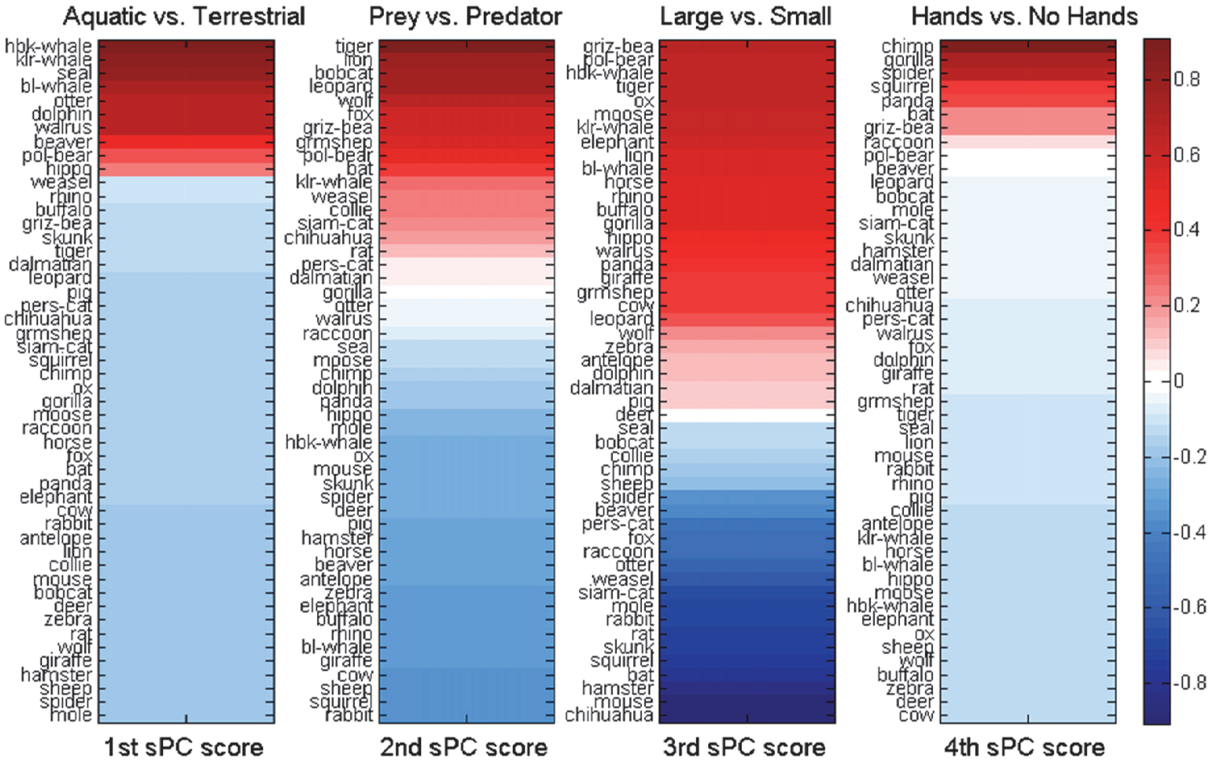

Fig. 7. Item scores on the first four sparse principal components.

The sequential SPCA algorithm we have used is like PCA in maximizing variance explained by each successive coefficient vector and in choosing orthogonal directions, but the selection of vectors is constrained so that the coefficient vectors are sparse.[†] We experimented with different sparsities (number of non-zero coefficients) and found that the clearest results were obtained using a sparsity of 3. The first four resulting feature and score vectors are shown in Figures 6 & 7. The first dimension identifies aquatic vs. non-aquatic mammals; the second identifies predators vs. prey; and the third picks out the size dimension. It is important to note that these are *cross-cutting* dimensions. The whales are all highly aquatic, but whereas the killer whale lies at one end of the predator/prey dimension, the blue whale lies near the other end. Similarly, though the seal is highly aquatic, it is well separated from the whales in size.

The fourth dimension is also interpretable, and can be thought of in different ways. We have named this dimension after its most highly-weighted feature, hands, short for 'has hands'. The dimension also has large weights for 'bipedal' and 'tree', short for 'is bipedal', and 'lives in trees'. This cluster of features most strongly picks out the primates in the data set -- the chimp, the gorilla, and the spider monkey (note that size cuts across this dimension, since the three primates are of very different sizes). However, primarily because of the association (in the responses of the human participants who made the property attribution ratings) of all three types of primates with trees, the tree feature is included in the cluster, with the consequence that other, non-primate species that are also associated with trees score reasonably high on this dimension. This illustrates that the dimensions are ultimately defined by correlations and cannot be seen as necessarily separating characteristic from defining features or as selecting features of a homogeneous type. Dimensions of this kind might influence implicit or intuitive cognitive judgments, even if they do not necessarily capture the kinds of dimensions a scientist might use to organize information about the properties of objects[18].

The SPCs capture information also captured in Kemp and Tenenbaum's hierarchical analysis, as well as information not captured there. In particular, the aquatic dimension picked out by the SPCA is essentially the same dimension as the one captured along the vertical axis of Kemp and Tenenbaum's tree (see

[†] It should be noted that imposing a sparsity constraint of the feature vectors used in a PCA is not the same as imposing a sparsity constraint on the number of active units in a distributed representation of a given input[21]. The words *sparse* and *sparsity* are used in both contexts but the entities these terms describe are different in the two cases and it is useful to realize that there is a distinction between these usages.

Figure 3). The ‘hands’ dimension is associated with the small cluster of primates also found in Kemp and Tenenbaum’s tree. However, the prey-predator and size dimensions are not well captured in the tree. The tree picks out a set of large non-predators (lower right), but there are large and small prey and predator animals sprinkled throughout the top and bottom-left branches, as well as size variations within many branches.

In a further analysis, we found that each of the first few dimensions the SPCA discovers is more similar to the corresponding standard, non-sparse PC than to any of the other standard PCs, although the sparse PCs do have some cross-similarity with several non-sparse PCs. This finding is shown in Figure 8, which shows the projection of each of the first 5 sparse PCs onto each of the first five non-sparse PCs. We see the strongest positive values along the diagonal, but we also find non-zero off-diagonal projections. As examples, the first sparse PC projects most strongly onto the first standard PC but also has a fairly strong negative projection onto the second standard PC, while the second sparse PC has moderately strong positive projections onto the first and third standard PCs. Looking at these results in another way, we can say that the first standard PC has non-negligible projections onto all of the first 5 sparse PCs, indicating that the first standard PC reflects a mixture of the interpretable dimensions identified by the sparse principal component analysis.

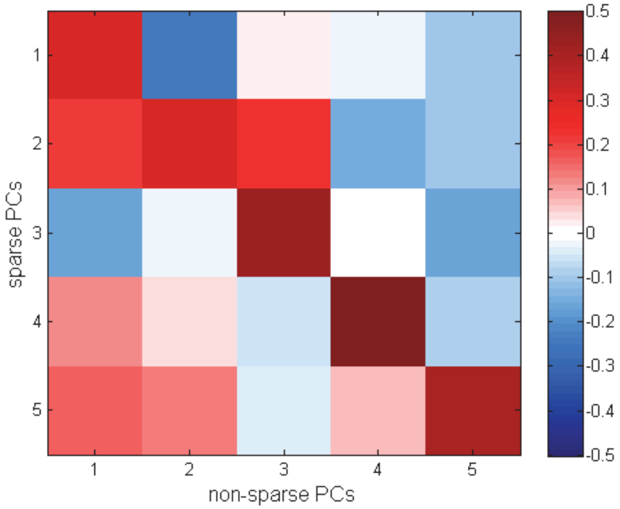

Fig. 8. Projections of the first five sparse principal components onto the first five non-sparse principal components of the 50 mammal data set.

5. Discussion

Our results highlight the fact that a tree structure may often provide an imperfect guide to the full structure present in a data set. In particular, a hierarchical tree is bound to hide semantic distinctions that cut across levels of the tree. While Kemp & Tenenbaum found some support for the view that human subjects, like their hierarchy, discard this cross-cutting information when making property attribution judgments, Glick[2] probed this matter further, and showed that such cross-cutting information does influence attribution judgments as well as item similarity judgments. These points led us to argue for models capable of representing blended, hybrid structural forms that can include both hierarchical as well as cross-cutting structure. In other work[19], we have also found that the more flexible approach to structure characterization offered by singular value decomposition and principal component analysis may help reveal cross-category associations that further contribute to semantic knowledge about items, beyond the properties of the items themselves.

Flexible structure is naturally captured by deep neural network models, but the dimensions these models find are not necessarily easily interpretable. We are not discomforted by the possibility that natural structure is not perfectly characterizable in terms that are easily interpretable, and we stress that these dimensions might still underlie human similarity and property attribution judgments, even if they are difficult to fully describe in words; further research on this topic will certainly be warranted.

In the meantime, the present work has focused on finding a way of projecting the knowledge that is captured in a deep neural network onto dimensions that may be more easily described. Our investigations indicate that such projections can be provided by including a sparsity constraint limiting the number of features used in each representational dimension. These dimensions bear some similarity to the corresponding non-sparse dimensions, but are more easily interpretable, and may be useful as a projection of the real knowledge underlying intuitive semantic judgments into dimensions that can be described in words. These dimensions are, however, still fundamentally derived from purely statistical considerations, and can combine information of very different types (as in the case of the fourth dimension, which combines ‘has hands’ and ‘lives in trees’, properties that happen to correlate in the ratings for the primates in the data set).

Implications for modeling frameworks. We would like to stress that our critique of a pure hierarchy as a representation of semantic knowledge is not intended as a critique of any particular approach to understanding human knowledge representation. The points made here and in the work of Glick[2] are criticisms of hierarchical models, whether they come from the tradition of structured probabilistic models, as the model of Kemp and Tenenbaum does, or from more generic hierarchical data analysis methods, as shown in Figure 2. Hierarchical models of whatever provenance obscure cross-cutting structure, and such structure plays a role in the similarity judgments and property attributions of human subjects. As noted earlier, Kemp, Tenenbaum, and colleagues have discussed ways of capturing combinations of structure types[20,4]. A structured probabilistic model of human semantic knowledge different from the one they presented in their 2009 paper could potentially address most of our concerns.

We would note, however, that it may be possible to represent and be sensitive to structure without relying on *any* explicit structural form[5]. A singular valued decomposition or deep neural network can, as we have seen, capture either hierarchical or cross-cutting structure or a blend of hierarchical and cross-cutting structure *without explicitly entertaining either structure type.* Indeed, the range of structure types that can be captured in this approach includes many other structure types, as well as arbitrary blends of such structure types, and even data sets that do not strictly conform to, but only approximate, such structure types. Thus, any explicit specification of a particular structure type can be misleading unless it is understood as a descriptive approximation rather than an actual characterization of the representation underlying human semantic cognition.

References

1. Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, **116,** 1, 20–58.
2. Glick, J. J. (2011). *Uncovering the organization of semantic structure with similarity and inductions*. Doctoral dissertation, Psychology department, Stanford University.
3. Glick, J. J., & McClelland, J. L. (2010). Semantics in the wild: Context-sensitive inferences about mammals. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (p.668). Austin, TX: Cognitive Science Society.
4. Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, **120**, 1, 1–25.
5. McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T.T., Seidenberg, M. S., & Smith, L. B. (2010). Letting Structure Emerge: Connectionist and Dynamical Systems Approaches to Understanding Cognition. *Trends in Cognitive Sciences*, **14**, 348–356.

6. McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, **2,** 6, 387-395.
7. Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013). Learning hierarchical category structure in deep neural networks. In M. Knauff, M. Paulen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the Cognitive Science Society*. (pp. 1271-1276). Austin, TX: Cognitive Science Society.
8. Osherson, D. N., Stern, J., Wilkie, O., Stob, M., & Smith, E. E. (1991). Default probability. *Cognitive Science*, **15**, 251–269.
9. Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, **104**, 2, 211.
10. Rogers, T. T. & McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach.* Cambridge, MA: MIT Press.
11. Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (Eds.), *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience* (pp. 3-30). Cambridge, MA: MIT Press.
12. McClelland, J. L. (1994). The interaction of nature and nurture in development: A parallel distributed processing perspective. In P. Bertelson, P. Eelen, & G. d'Ydewalle (Eds.), *International Perspectives on Psychological Science, Volume 1: Leading Themes*. United Kingdom: Erlbaum.
13. McClelland, J. L. (1989). Parallel distributed processing: Implications for cognition and development. In Morris, R. (Ed)., *Parallel distributed processing: Implications for psychology and neurobiology.* (pp. 8-45). New York: Oxford University Press.
14. Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, **96**, 4, 523-568.
15. McClelland, J. L. (2015). Capturing gradience, continuous change, and quasi-regularity in sound, word, phrase, and meaning. In B. MacWhinney & W. O'Grady (Eds.), *The Handbook of Language Emergence*, Chaper 2, pp. 54-80. Hoboken, NJ: John Wiley & Sons.
16. Zou, H., Hastie, T. & Tibshirani, R. (2006) Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, **15**, 2, 265-286.

17. Sjöstrand, K., Clemmensen, L. H., Larsen, R., & Ersbøll, B. (2012). Spasm: A matlab toolbox for sparse statistical modeling. *Journal of Statistical Software*, *VV* (Issue I).
18. Goldberg, R. F., & Thompson-Schill, S. L. (2009). Developmental "roots" in mature biological knowledge. *Psychological Science*, **20**, 4, 480-487.
19. Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015). You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia*, **76***,* 52-61.
20. Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *PNAS*, **105**, 31, 10687-10692.
21. Olshausen, B. A. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, **381**, 6583, 607-609.