Generalization of within-category feature correlations

Nolan Conaway & Kenneth J. Kurtz
Department of Psychology, Binghamton University
Binghamton, NY 13905 USA

Abstract

Theoretical and empirical work in the field of classification learning is centered on a 'reference point' view, where learners are thought to represent categories in terms of stored points in psychological space (e.g., prototypes, exemplars, clusters). Reference point representations fully specify how regions of psychological space are associated with class labels, but they do not contain information about how features relate to one another (within- class or otherwise). We present a novel experiment suggesting human learners acquire knowledge of within-class feature correlations and use this knowledge during generalization. Our methods conform strictly to the traditional artificial classification learning paradigm, and our results cannot be explained by any prominent reference point model (i.e., GCM, ALCOVE). An alternative to the reference point framework (DIVA) provides a strong account of the observed performance. We additionally describe preliminary work on a novel discriminative clustering model that also explains our results.

Keywords: categorization; generalization; formal modeling

Introduction

Research on human classification learning is fundamentally interested in questions of representation: How do people represent categories? How does a category's internal structure influence its subjective difficulty? How do people generalize their knowledge about categories? Current research addressing these questions is centered around a 'reference-point' framework, whereby people are thought to acquire category knowledge associating stored perceptual referents (e.g., prototypes, exemplars) with individual categories. The success of reference point models (e.g., Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky, 1984; J. D. Smith & Minda, 2000) is unparalleled within the field, and as a result these models are widely considered to be definitive accounts of how categories are learned and represented (for reviews, see Murphy, 2002; Pothos & Wills, 2011; see also Kurtz, 2015).

Although reference point models differ from one another in a variety of ways, these models are comparable in that they assume that categories are represented by one or more points in a psychological space. On the extremes, prototype models represent categories with a central tendency (i.e., the average across known members), whereas exemplar models use specific observations. Many successful reference point models employ a selective attentional mechanism, enabling them to weight the importance of each stimulus dimension (Kruschke, 1992; Medin & Schaffer, 1978).

Importantly, however, reference point representations do not incorporate all aspects of class structure. While points of reference can be used to encode information about how regions of space are associated with known categories, they do not contain information about how features relate to one another (either globally or within a class). By consequence, reference point models are only sensitive to correlations between features insofar as those correlations are reflected in the distances between stored reference points.

Although there is evidence that people make use of feature correlations in natural concepts (Malt & Smith, 1984), research on correlation learning in a traditional artificial classification learning (TACL) setting has been mixed. Whereas Medin, Altom, Edelson, and Freko (1982) reported evidence that feature correlations influence classification of artificial categories, Murphy and Wisniewski (1989) later expanded upon that study and found little evidence of correlation learning, unless features are expected to be correlated. Finally, Anderson and Fincham (1996) reported that participants used correlations to infer values of missing features; though note that traditional reference point models are unable to simulate feature inference (Lee & Navarro, 2002).

In this paper, we report a classification learning experiment demonstrating that people represent correlations between features, beyond what can be explained in terms of stored reference points. In addition to providing evidence that learners do acquire knowledge about correlations between features, the classification performance we report demonstrates a systematic failure in the reference point framework. We bolster our empirical results with simulations using the Generalized Context Model (GCM; Nosofsky, 1984), an exemplar model that embodies the central tenets of the reference point view.

We also report simulations using the *DIV*ergent *A*utoencoder model (DIVA; Kurtz, 2007, 2015), a autoassociative network model that stands as a similarity-based alternative to the reference point framework. The DIVA model is fully instantiated as a connectionist network: as in traditional multilayer perceptron (MLP) architectures, DIVA is initialized with a input units encoding feature values, as well as a collection of hidden units enabling the learning of an internal representation (Rumelhart, Hinton, & Williams, 1986). DIVA's primary point of departure from these models lies in its learning objective: instead of learning representations to predict class responses, the DIVA model learns auto-associatively to predict feature values along divergent, category-specific output channels. Thus, DIVA's category representations are acquired for the purposes of making *feature* predictions rather than *class* predictions. Classification decisions are made using a secondary response rule: the probability of any given classification depends on the relative amount of feature prediction error (reconstruction error) across all categories, with better reconstructions leading to increased probability.

DIVA's design principles offer a unique account of human

category learning: rather than assuming people learn to predict class responses through association to stored points of reference, DIVA proposes that people learn representations of the observed regularities within each class. Accordingly, with regard to learning feature correlations, DIVA's predictions sharply contrast from those made by reference point models. Specifically, because DIVA is trained on feature prediction (rather than class prediction), it strongly relies on within-class feature correlations to aid learning. Thus, whereas reference point models do not encode any information about feature correlations, DIVA relies on those correlations in the service of minimizing feature prediction error.

Finally, we conclude our report with preliminary work on a novel discriminative clustering model that explains our results without acquiring knowledge of how features are correlated. Development of this model is ongoing, though our simulations results indicate that it may succeed as an account of human classification more generally.

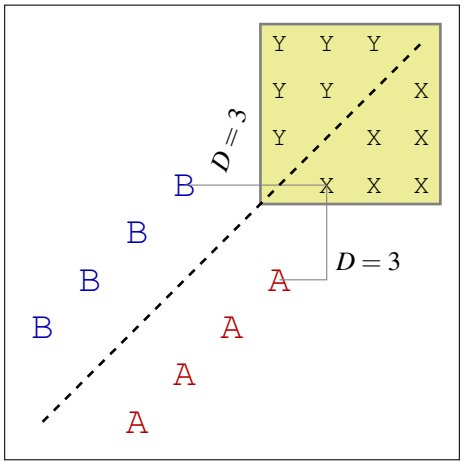


Figure 1: The Diagonal classification. X and Y indicate critical generalization items. Annotations illustrate equal city-block distance between the critical items and known category members.

The Current Study

In a *Diagonal* classification, the two categories ('A' and 'B') are organized along a diagonal boundary (see Figure 1). This classification is notable for its likeness to Information-Integration (Ashby & Maddox, 1990) and Condensation categories (Gottwald & Garner, 1972; Kruschke, 1993). The features are perfectly correlated within each category, but the training exemplars are isolated in one region of the stimulus space, allowing for generalization items that follow the diagonal boundary (labeled X & Y in Figure 1).

Our DIVA simulations revealed that the model typically extends the diagonal boundary to these items – exemplars on the Category A side of the extended boundary are more likely to be classified as members of Category A than exemplars on the other side, and vice-versa (i.e., $A \rightarrow X$, $B \rightarrow Y$). DIVA's performance is concisely explained by its design principles: because DIVA is trained on feature prediction (rather than class prediction), the model learns that the features can be used to predict one another. After training, DIVA reconstructs novel items following its knowledge of each category's feature correlations, affording an extension of the diagonal boundary.

We found that the GCM could not mimic DIVA's performance. Instead of extending the diagonal boundary outward, the GCM classifies each critical item with equal probability. A close examination of the Diagonal classification explains the GCM's performance: each of the critical items is equally distant (under a city-block metric) to known members of both categories. While a Euclidean metric would enable the GCM to extend each category to its critical items, the use of a city-block metric reflects the separable stimulus dimensions used in the behavioral experiment reported below (see Garner, 1974). City-block distance was also supported by the results of an independent pairwise similarity-rating study – stimuli X & Y were not rated as more similar to items on their own side of the diagonal boundary.

With evenly distributed selective attention, the GCM classifies each item with a probability of 0.5. Unequal allocation of attention results in uniform changes to the classification gradient, but not generalization of each category along the diagonal. For example, greater allocation to the vertical dimension results in increased Category B probability for the entire collection of critical items. Finally, other types of reference points (i.e., prototypes, clusters) also lead to the same performance. The critical items area are equally close to the category prototypes, as well as a variety of cluster configurations. The GCM's results therefore characterize a set of predictions made by reference point models more generally.

Examining these predictions more methodically, we conducted a 'grid-search' to evaluate DIVA and GCM performance under a range of parameter settings. At each point in the search, the models were tested on classification of novel items X & Y. To quantify each parameterization's degree of diagonal extension, we calculated the difference score in the Category A classification probability for the critical items on either side of the boundary ($X - Y$). Positive difference scores indicate systematic generalization of each class, and neutral (≈ 0) scores indicate uniform generalization. Plotting these scores as a density curve across points in the search (Figure 2) reveals strongly systematic behavior: whereas DIVA nearly always generalized each class outward, the GCM produced identical responses to X & Y under every parameter setting.

In what follows, we report a behavioral study testing the predictions made by DIVA and the GCM on generalization of the Diagonal classification. If human learners represent categories solely in terms of reference points, then generalization should be uniform: participants should be no more likely to produce a Category A response to items on the A-side of the diagonal than on the opposite side (i.e., $X \approx Y$). However, if participants acquire knowledge about how the features relate to one another within each category (as predicted by DIVA),

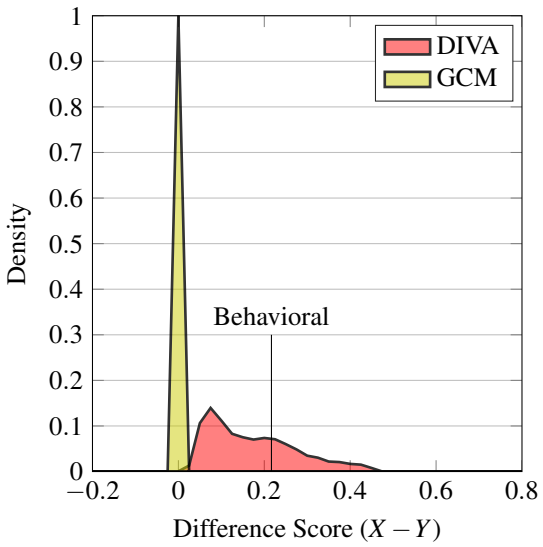


Figure 2: 'Grid search' simulations with DIVA and the GCM. Densities in this plot reflect predictions about the classification of critical items, across many different parameter settings. Positive scores indicate that each category was generalized to the items on its side of the boundary, $A \rightarrow X$, $B \rightarrow Y$.

then we should observe systematic generalization of the diagonal boundary outward (i.e., $A \rightarrow X$, $B \rightarrow Y$).

Participants and Materials. 30 undergraduates from Binghamton University participated in fulfillment of a course requirement. Stimuli were squares varying in shading and size (see Figure 3). The separability of these features justifies a city-block metric (Garner, 1974). An independent scaling study verified that the dimensions are nearly equal in perceptual salience. The assignment between perceptual and conceptual dimensions was counterbalanced across participants.

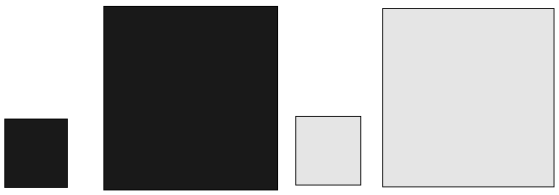


Figure 3: Sample stimuli.

Procedure. Participants completed 96 training trials (12 blocks consisting of the 8 training examples). On each trial, a stimulus was presented on a computer screen and learners were prompted to make a classification decision by clicking one of two buttons (labeled 'Alpha' and 'Beta'). After selecting a class label, learners were given feedback on their response. Following the training phase, participants completed 81 generalization trials consisting of items sampled at 9 positions on each dimension. All of the training examples were included (intermixed). Feedback was not provided during the generalization phase. Participants were informed that there would be test trials prior to beginning the experiment.

Results. By the end of the training phase, most participants had successfully mastered the categories. On average, participants were 89.2% ($SE = 2.5$) accurate during the last training block. Only two of 30 participants failed to reach greater than 6/8 correct during the final training block. Aggregate training data is depicted in the left panel of Figure 4.

As a test of whether learners extended the diagonal boundary, we compared the average number of 'A' classification responses made to critical items X & Y (Figure 4). Learners were more likely to produce an 'A' response to X than to Y, $t(29) = 5.02$, $p < 0.001$, $d = 0.56$. We then compared the difference score we obtained behaviorally ($X - Y$) to our earlier results with DIVA and the GCM: as shown in Figure 2, the difference score we observed cannot be produced by the GCM, but is fully explained by DIVA. Aggregate generalization data is depicted in Figure 5.

Summary

We provided classification training to human learners on a Diagonal classification (Figure 1). Learners systematically extended each category to novel items that are equidistant to known exemplars from both categories. The generalization we observed can be considered evidence that learners acquire knowledge about feature correlations, and they apply that knowledge during generalization. Our results are also inconsistent with the notion that category knowledge solely consists of exemplar, prototype, or cluster reference points.

Our results are concisely explained by the DIVA model. Because DIVA's is principally autoassociative, the model relies on within-class internal regularities (such as feature correlations) to support reconstruction learning. In our above simulations, we found that the model frequently generalized the diagonal boundary outward, just was we observed in the behavioral study. However, it is customary to assess models using a post-hoc parameter fitting process. In the next section, we describe a more formal examination of the performance by DIVA and the GCM.

Simulations

The overall goal of the following simulations is to formally evaluate the performance of DIVA and the GCM in terms of quantitative fit to our observed generalization behavior. Before proceeding, it is worth noting that DIVA and the GCM are not fully comparable: unlike the GCM, DIVA's category representations are constrained via back-propagation learning (Rumelhart et al., 1986), and its performance is stochastic. However, model performance on generalization testing can still be compared to assess whether our results can be explained under a reference point scheme.

We used parameter optimization techniques to find each model's best fit to the observed generalization data using of a mean-squared error (MSE) metric. We used a hill climb-

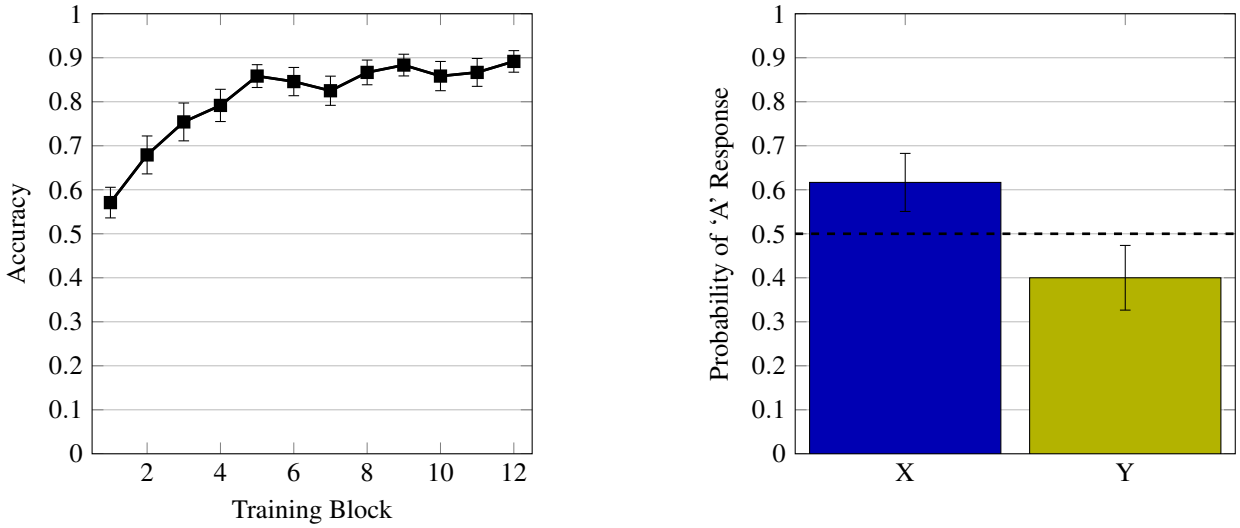


Figure 4: *Left*: Behavioral training accuracy. *Right*: Behavioral responses to critical items *X* & *Y*. Error bars reflect ±1 SE.

ing procedure to fit the GCM over four parameters: exemplar specificity (c), response determinism (γ; Nosofsky & Zaki, 2002), and attention strengths for features 1 and 2 (W_1 and W_2). DIVA's behavior is stochastic, however, which precludes the use of hill-climbing. As a result, we search for DIVA's best fit using a 'grid-search' technique to generate predictions along a range of settings for its four parameters: number of hidden units, learning rate, initial weight range, and a focusing parameter, β (Kurtz, 2015). At each point, DIVA was initialized 2000 times with random small-valued weights and a random presentation sequence. The model used logistic hidden units and linear outputs.

Overall, DIVA was able to provide a stronger fit to the full behavioral gradient than the GCM. DIVA's best fit ($MSE = 0.006$) was achieved with 3 hidden units, learning rate = 0.55, weight range ± 0.5, and $\beta = 3$. The GCM's best fit ($MSE = 0.0075$) was achieved with $c = 5.304$, $\gamma = 1.055$, $W_1 = 0.507$, $W_2 = 0.493$. Model performance was further differentiated the responses to the critical items (rather than the entire gradient): DIVA achieved $MSE = 0.0037$, and the GCM achieved $MSE = 0.02$. Beyond quantitative fitting, however, it is important to acknowledge that DIVA's predictions match the qualitative patterns of observed performance – the model extends each category to novel items along the diagonal boundary. Conversely, the GCM's generalization is completely neutral for these items. Each model's best performance is depicted in Figure 5.

A discriminative clustering account.

Although traditional reference point models are unable to explain our results, we have recently implemented a novel discriminative clustering account that successfully captures the observed generalization performance. The advance made by this account lies in the realization that the reference points associated with each category need not be localized as the exemplars, the category's central tendency, or the central tendency of select clusters of exemplars. Instead, the location of reference points may discriminatively reflect the proximity of opposite-category members. In doing so, this model may provide a account of the effects of contrast categories on conceptual representation (Davis & Love, 2010; Levering & Kurtz, 2006; Palmeri & Nosofsky, 2001).

Overall, the design of the discriminative clustering model is similar to SUSTAIN (Love et al., 2004): the model's reference point representation consists of a collection of clusters, and classification is based on each category's association the clusters. The model begins training with no internal representation, and recruits clusters in when it makes a poor classification decision. The primary departure from SUSTAIN concerns the localization of the clusters: on each trial, clusters belonging to the correct category are moved toward the presented exemplar, and clusters belonging to incorrect categories are move away from the presented exemplar. By allowing the clusters to move discriminatively from members of the opposite category, they become more similar to critical items on the category's side of the diagonal boundary, producing the observed pattern of generalization. Sample performance from this model is depicted in Figure 6.

Discussion

The reference point framework has dominated research and theory in category learning for the past 30 years. Reference point representations are useful in that they specify how regions of space are associated with class labels, but they do not encode information about global or within-class regularities, such as feature correlations.

We reported an experiment suggesting that human learners acquire knowledge of within-class feature correlations. Specifically, after training on a Diagonal classification, participants systematically generalized in accord with each cat-

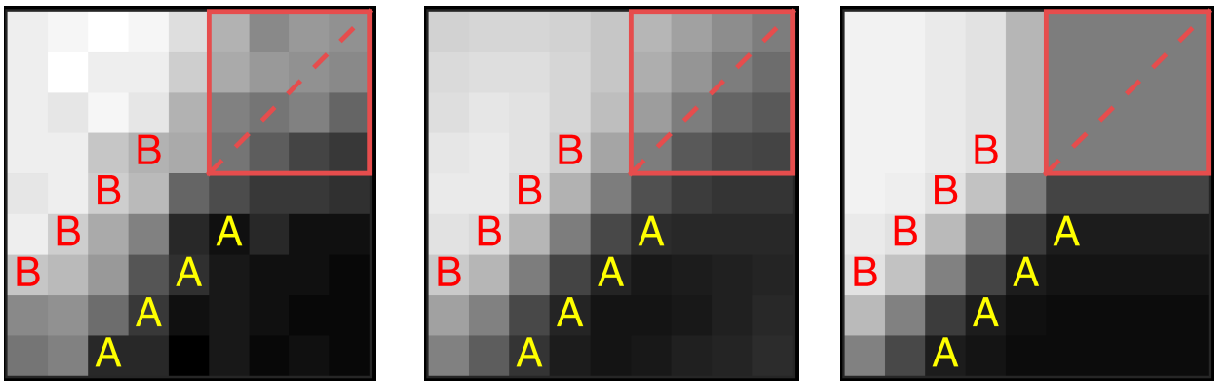


Figure 5: *Left*: Behavioral generalization. *Center & Right*: Best-fit predictions from DIVA and the GCM.

egory's internal structure to exemplars that are equally similar to members of both categories. The observed performance is inconsistent with prominent reference point models (e.g., Kruschke, 1992; Love et al., 2004; Nosofsky, 1984; J. D. Smith & Minda, 2000), and indicates that learners may acquire knowledge about categories that cannot be represented under a reference point scheme. The *Div*ergent *A*utoencoder model (DIVA; Kurtz, 2007, 2015) succinctly explains the observed performance: because DIVA is chiefly an autoassociator, the model depends on within-class regularities (such as feature correlations) to aid in feature prediction. Accordingly, DIVA's generalization shows strong sensitivity to the within-category feature correlations.

Although existing reference point models fail to match our results, we introduced a novel discriminative clustering model capable of producing the observed generalization. Development of this model is ongoing, though its results here show promise. Unlike existing cluster-based approaches (i.e., Love et al., 2004), the cluster locations in our model are optimized both for similarity to same-category exemplars, and for dissimilarity to opposite-category exemplars. Thus, the model's clusters are gradually moved outward in the stimulus space, taking on the value of a category ideal. From this location, clusters are more similar to novel items on their side of the diagonal boundary, affording generalization of each class. Future work will attempt to dissociate the predictions made by DIVA and the discriminative clustering model.

This report adds to an accumulating body of evidence against the idea that category learners solely acquire knowledge in the form of reference points. Research in function learning (DeLosh, Busemeyer, & McDaniel, 1997) has, for example, demonstrated key flaws in the account of extrapolation put forward by reference point similarity – studies on on rule-based generalization have revealed similar flaws in a category learning context (e.g., Erickson & Kruschke, 2002). Traditional reference point approaches are also unable to explain the effects of category variability on generalization (Cohen, Nosofsky, & Zaki, 2001; E. E. Smith & Sloman, 1994). Finally, our recent work (Conaway & Kurtz, 2015) has uncovered unique generalization behavior that cannot be explained via similarity to reference points. Taken as a whole, these findings raise a substantive challenge to theories of human category learning based on similarity to reference points.

Acknowledgments

We thank Sarah Laszlo and Gregory Murphy for helpful comments on this work. We additionally thank Sarah Laszlo for access to her server. We also thank the members of the Learning and Representation in Cognition (LaRC) lab at Binghamton University, as well as the BU Modeling meeting.

References

Anderson, J. R., & Fincham, J. M. (1996). Categorization and sensitivity to correlation. *Journal of experimental psychology: Learning, Memory, and Cognition*, *22*(2), 259.

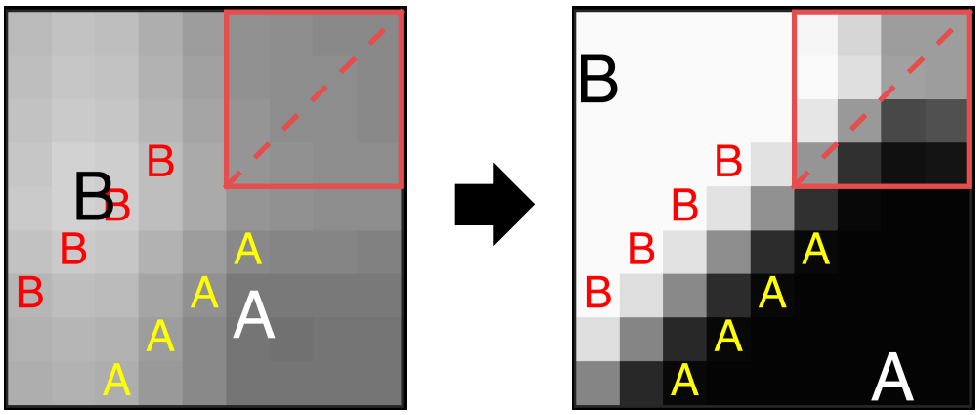
Ashby, F. G., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 598–612.

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory & Cognition*, *29*(8), 1165–1175.

Conaway, N. B., & Kurtz, K. J. (2015). A dissociation between categorization and similarity to exemplars. In D. C. Noelle & R. Dale (Eds.), *Proceedings of the 37th annual conference of the cognitive science society* (pp. 435–440). Austin, TX: Cognitive Science Society.

Davis, T., & Love, B. C. (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science*, *21*(2), 234–242.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: the sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *23*, 968–986.



Trial 8 Post-training

Figure 6: Sample performance from the discriminative clustering model. Learned clusters are labeled using large, black and white 'A' and 'B' characters. The model is capable of learning multiple clusters, but a single-cluster solution is typical for this classification. *Left*: generalization at the beginning of Block #2. *Right*: generalization at the end of the training phase.

Erickson, M. A., & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review*, *9*(1), 160–168.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Gottwald, R. L., & Garner, W. R. (1972). Effects of focusing strategy on speeded classification with grouping, filtering and condensation tasks. *Perception and Psychophysics*, *11*, 179-182.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*(1), 3–36.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin and Review*, *14*, 560–576.

Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. *Psychology of Learning and Motivation*, *63*, 77–114.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin and Review*, *9*, 43–58.

Levering, K., & Kurtz, K. J. (2006). The influence of learning to distinguish categories on graded structure. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (pp. 1681–1686). Mahwah, NJ: Erlbaum.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of verbal learning and verbal behavior*, *23*(2), 250–269.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(1), 37.

Medin, D. L., & Schaffer, M. M. (1978). A context theory of classification learning. *Psychological Review*, *85*, 207–238.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge: MIT press.

Murphy, G. L., & Wisniewski, E. J. (1989). Feature correlations in conceptual representations. *Advances in cognitive science*, *2*, 23–45.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *10*(1), 104–114.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, memory, and cognition*, *28*(5), 924.

Palmeri, T. J., & Nosofsky, R. M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *The Quarterly Journal of Experimental Psychology: Section A*, *54*(1), 197–235.

Pothos, E. M., & Wills, A. J. (2011). *Formal approaches in categorization*. Cambridge University Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Smith, E. E., & Sloman, S. A. (1994). Similarity-versus rule-based categorization. *Memory & Cognition*, *22*(4), 377–386.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 3.