Learning cross-cutting systems of categories

Patrick Shafto, Charles Kemp, Vikash Mansinghka, Matthew Gordon, & Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

Most natural domains can be represented in multiple ways: animals may be thought of in terms of their taxonomic groupings or their ecological niches and foods may be thought of in terms of their nutritional content or social role. We present a computational framework that discovers multiple systems of categories given information about a domain of objects and their properties. Each system of object categories accounts for a distinct and coherent subset of the features. A first experiment shows that our *CrossCat* model predicts human learning in an artificial category learning task. A second experiment shows that the model discovers important structure in two real-world domains. Traditional models of categorization usually search for a single system of categories: we suggest that these models do not predict human performance in our task, and miss important structure in our real world examples.

People explain different aspects of everyday objects in different ways. For example, steak is high in iron because it is a meat; however, it is often served with wine because it is a dinner food. The different ways of thinking about steak underscore different ways of thinking about the domain of foods: as a system of taxonomic categories like meats and vegetables, or as a system of situational categories like breakfast foods and dinner foods. If you were to plan meals for a family trip you would draw upon both of these systems of categories, consulting the taxonomy to insure that meals were nutritionally balanced and consulting the situational system to insure that there were foods that were appropriate for the different times of the day. In any domain, objects have different kinds of properties, and more than one system of categories is needed to explain the different relationships among objects in the domain.

Psychologists have experimentally confirmed that multiple systems of categories are needed to account for human behavior. Ross and Murphy (1999) showed that subjects draw on at least two different kinds of knowledge to categorize and reason about foods: knowledge about taxonomic categories and knowledge about foods that tend to be eaten together. Similarly, studies have shown that animals may be thought about in terms of taxonomic categories such as mammals and reptiles, or ecological categories such as predators and prey. For example, reasoning about anatomical properties appears to draw on taxonomic categories, but reasoning about disease transmission may rely on ecological categories (see Heit and Rubinstein, 1994; Shafto and Coley, 2003; Shafto et al., 2005).

Most previous models of categorization have attempted to discover a single system of categories within a given domain (but see Martin and Billman, 1994). This paper introduces *CrossCat*, a Bayesian framework for discovering multiple systems of categories — for example, discovering that foods can be organized into a system of taxonomic categories and a system of situational categories. A key feature of our approach is that we need not specify the number of systems of categories in advance, or the number of categories within each system: our model automatically discovers a representation of appropriate complexity.

To test our model, we studied human performance in an unsupervised learning task, and analyzed the structure of two real-world datasets: foods and animals. Our model provides a good account of human performance, and captures intuitively compelling structures in both of our datasets. Of the previous models that search for a single system of categories, our approach is related most closely to Anderson's rational analysis of categorization (Anderson, 1991). We compare our approach to this model throughout, and argue that models that rely on a single system of categories cannot provide an adequate account of human learning and reasoning.

A generative model for learning systems of categories

Assume we are provided with an array of objects and features — for example, the matrix of foods shown in Figure 1. Our goal is to organize the objects into one or more systems of categories, and to discover the features best explained by each system. A good solution for the food matrix is shown in Figure 2. There are two systems of categories: the first is a situational system partitioned into breakfast foods and dinner foods, and the second is a taxonomic system that includes starches, meats, and diary foods. Intuitively, the solution is a good one because each feature respects the structure of its associated category system; for example, "served with wine" discriminates perfectly between the situational categories, but is not as clean with respect to the taxonomic categories (half of the starches are served with wine and half are not).

More formally, CrossCat takes as input a list of O objects, a list of F features, and an O by F data matrix

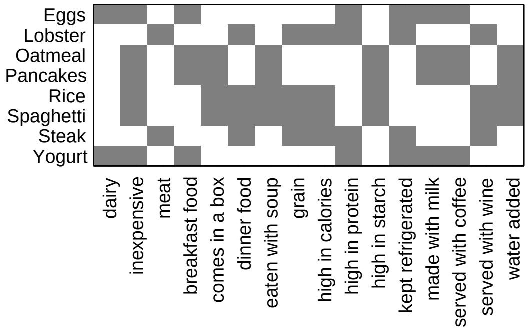

Figure 1: The food matrix used in simulations. Grey and white areas indicate true and false features, respectively.

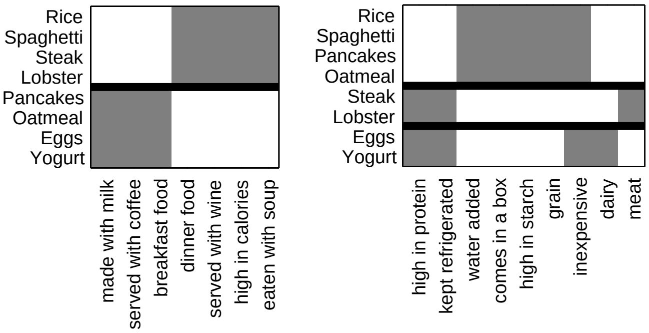

Figure 2: The food matrix sorted by the best solution under CrossCat: two systems corresponding to situational and taxonomic categorizations of the objects.

D, where the (o, f) entry is the value of feature f for object o. The output is a partition of the features into kinds, and for each kind, a categorization of the objects. In Figure 2, the model has found two kinds, with two and three categories of objects. The output is represented using a vector $\mathbf{z}$, where z_f is the kind for feature f (here, kind 1 corresponds to situational features and kind 2 corresponds to taxonomic features). Our model also returns the categories associated with each feature kind k in a vector $\mathbf{w}^k$. In the food example, eggs are a breakfast food (category 2 for kind 1) and a dairy product (category 3 for kind 2), so $w_1^1 = 2$ and $w_1^2 = 3$.

To specify our Bayesian model, we define a probabilistic process that generates the feature kinds, the systems of categories and the data matrix D. We then discover systems of categories by searching for the feature kinds $\mathbf{z}$ and categories $\mathbf{w}^k$ that are most probable given the data: in other words, that maximize

$$P(\mathbf{z}, \{\mathbf{w}^k\}|D) \propto P(\mathbf{z}, \{\mathbf{w}^k\}, D) \qquad (1)$$

$$= P(\mathbf{z}) \prod_{k=1}^{K} P(D^k|\mathbf{w}^k) P(\mathbf{w}^k) \qquad (2)$$

where K is the number of feature kinds in $\mathbf{z}$, and D^k is the portion of the data matrix which system k must explain. To complete the model, we must define three components: $P(\mathbf{z})$, a prior distribution on feature partitions, $P(\mathbf{w}^k)$, a prior on partitions of the objects into categories, and $P(D^k|\mathbf{w}^k)$, a process by which the data for each feature kind are generated. We consider each component in turn.

Intuitively, the prior on feature partitions $P(\mathbf{z})$ should assign some probability to all possible partitions, including the partition where all features belong to the same kind, and the partition where each feature belongs to its own kind. The prior, however, should favor the simpler partitions — those that use only a small number of kinds. We capture both intuitions by using a prior induced by the *Chinese restaurant process* or CRP, a standard tool from nonparametric Bayesian statistics and the basis for category discovery in Anderson's rational model. The CRP provides a mathematically principled way of discovering the right number of classes as well as their membership, and scales to arbitrarily large numbers of features. It can be thought of in terms of a seating scheme for a restaurant with an infinite number of tables. Each table corresponds to a group in a partition (here, a feature kind), and each person to enter the restaurant corresponds to an element to be partitioned (here, a feature). People are seated sequentially according to the following probabilities, where n_k is the number of people previously seated at table k and α is the distribution's single parameter:

$$P(z_i = k|z_1, \cdots, z_{i-1}) = \begin{cases} \frac{n_k}{i-1+\alpha} & \text{if } n_k > 0 \\ \frac{\alpha}{i-1+\alpha} & k \text{ is a new class} \end{cases}$$

Since the CRP is *exchangeable*, it induces a distribution $P(\mathbf{z})$ on complete class assignments that is invariant to the ordering of the features.

If we knew $\mathbf{z}$, the assignment of features to kinds, our problem would reduce to a series of K traditional categorization problems. For each feature kind k we could search for the partition $\mathbf{w}^k$ of the objects into categories that maximizes $P(\mathbf{w}^k|D^k) \propto P(D^k|\mathbf{w}^k)P(\mathbf{w}^k)$. Even though $\mathbf{z}$ must be inferred by our model, the last two terms in Equation 2 are identical to distributions needed by one traditional model of categorization, Anderson's rational model (Anderson, 1991), also known as the infinite mixture model (Rasmussen, 2002). Note that CrossCat reduces to this model when all features are assigned to a single kind. Our technical innovation, then, is to suggest that multiple mixture models, each on a different set of features, are needed to capture everyday knowledge about the structure of real-world domains.

Following the infinite mixture model, our prior on a partition of the objects into categories, $P(\mathbf{w}^k)$ is induced by a CRP with hyperparameter β. The remaining term, $P(D^k|\mathbf{w}^k)$, is the standard likelihood for independent binary features in a mixture model:

$$P(D_k|\mathbf{w}^k) = \prod_{f}^{F_k} \prod_{c}^{C} \frac{Beta(n_{f,c} + \delta, \bar{n}_{f,c} + \delta)}{Beta(\delta, \delta)} \qquad (3)$$

where F_k is the number of features in kind k, C is the number of categories in $\mathbf{w}^k$, and $n_{f,c}$ and $\bar{n}_{f,c}$ are the number of true and false instances of feature f in category c. Intuitively, the term $P(D^k|\mathbf{w}^k)$ is largest when the feature values in kind k are well predicted by the object categories chosen for kind k: that is, when all members of the same object category tend to have the same values for all features in kind k.

Now that the three terms in Equation 2 have been specified, we can see that CrossCat captures a tradeoff between two competing factors. The terms $P(\mathbf{z})$ and $P(\mathbf{w}^k)$ specify a preference for simple solutions that use a small number of feature kinds and a small number of object categories within each kind. The term $P(D^k|\mathbf{w}^k)$ favors solutions that explain the data well, and tends to prefer more complex solutions. By combining all three terms, we arrive at a model that attempts to find the simplest solution that adequately accounts for the data.

All results in this paper were generated using a stochastic search algorithm with local proposals similar to those typical of a Gibbs sampler. Some proposals move features from one kind to another (possibly creating new kinds and systems of categories) and others move objects between categories (possibly creating new categories within existing kinds). Hyperparameters (α, β and δ) were set to 0.5 in all cases. Many algorithms could potentially implement the the computational theory we have described, and we make no claims about the psychological plausibility of the particular implementation that we chose.

Experiments

We present two experiments contrasting the performance of CrossCat and a conventional infinite mixture model. Our first experiment tests model predictions against human results in an unsupervised learning task. Use of artificial categories allows us to ask whether people can discover multiple systems of categories, and to test model predictions in a controlled setting. Our second experiment contrasts CrossCat and the infinite mixture model in two real-world domains. We compare the representations found by the two models to our intuitions about the structure of these domains, paying particular attention to the additional structure discovered by CrossCat. To the degree that CrossCat predicts human performance and our intuitions, we suggest that it provides a good characterization of how people learn and represent systems of categories.

Modeling human category learning

Three artificial bug stimulus sets were created for this experiment. An example of one of these sets is presented in Figure 3. The stimulus sets were designed to support different systems of categories over the objects, where each system accounts for a subset of the features.

Method

Participants: Ten individuals from the MIT community participated in this experiment. Participants were recruited via a mailing list, and included both students and non-students.

Materials: Three sets of artificial stimuli, which we refer to as 3-3, 3-2, and 2, were created for the experiment. Each set of stimuli included eight bugs that varied on six binary features: number of legs, kinds of feet, body patterns, kinds of tails, kinds of antennae, and head shapes (see Figure 3 for an example set of stimuli and Figure 4 for the matrices corresponding to the three conditions). The 3-3 condition was designed to have two orthogonal systems, each with three categories. For example, the rows in Figure 3 represent categories defined by the shape of their heads, kinds of tails, and body patterns, while the columns represent an orthogonal set of categories based on the number of legs, kind of antennae, and kind of feet. The 3-2 condition was designed to have two systems of categories, one system with three categories and the second system with two categories. The 2 condition was designed to have a single system of two categories.

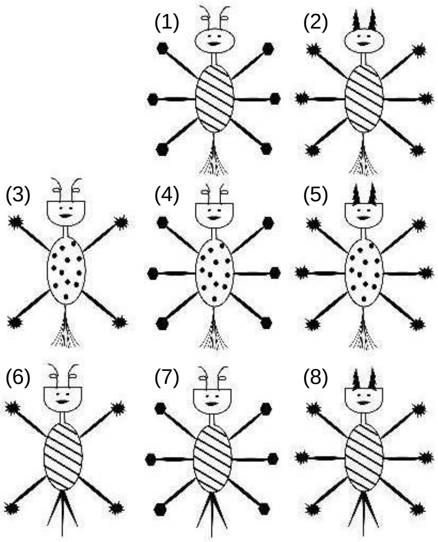

Figure 3: Stimuli from the 3-3 condition. Note that three categories can be formed by either grouping the objects by row or by column. Stimuli numbers correspond to the numbers used in Figures 4 and 5.

Procedure: There were two phases to the experiment: training and testing. In the training phase, participants were told that we were interested in different ways of categorizing a single set of objects, and different ways of categorizing foods was given as an example. The experimenter then explained the sequential sorting task with two examples. In the first example, the experimenter showed two ways of categorizing a set of cards with two orthogonal feature dimensions. In the second example, the experimenter showed the participant two prototype categories using stimuli from Yamauchi and Markman (2000). The experimenter explained that this was a good set of categories because it captured most of the information about the objects. Participants were told that they would be given a set of cards and they would be asked to sort them into categories. They would then be asked if there was another way of sorting the objects that "captured new and different information" about the objects. After each sort, they would be asked if there was another way of sorting the objects until they refused. When they

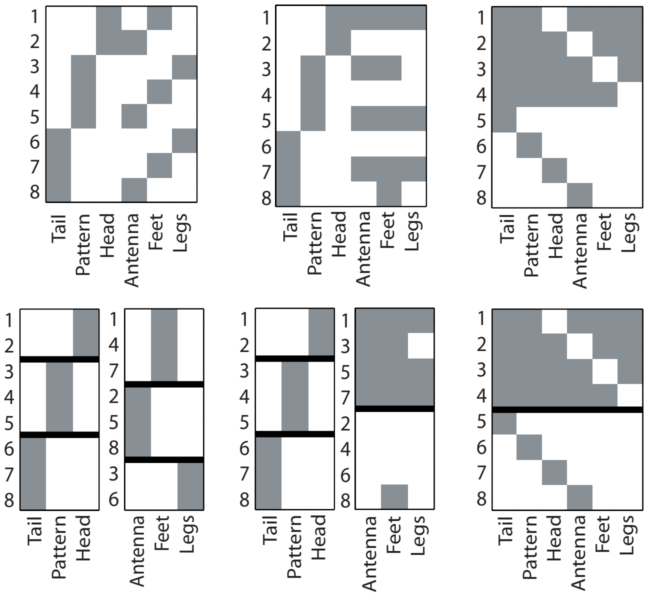

Figure 4: Matrices representing the structure of the different stimulus sets. The matrices correspond to the stimuli for 3-3, 3-2, and 2 conditions, from the left to the right. Unsorted matrices are presented on the top, and on the bottom are the best solutions according to the CrossCat model.

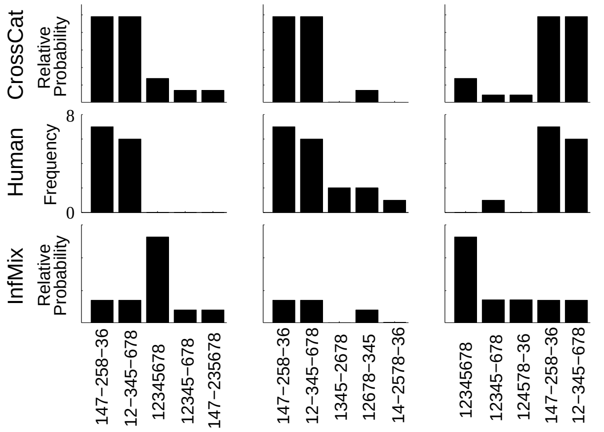

Figure 5: Results from the 3-3 condition. In each column, the results are sorted by a different standard: from left to right, the results are sorted by CrossCat predictions, human results, and the infinite mixture model's predictions. Note that the y-axis scale is the same across each row and the model solutions are plotted according to their relative probability, while the human data are plotted by frequency. On the x-axis is the partition of the objects, with different categories separated by dashes. For example, in the case at the far left, stimuli 1, 4, and 7 were in the same category, while stimuli 2, 5, and 8 were in another.

refused to sort further, participants were asked to rate the goodness of each system of categories (though we do not present those results here). Each participant sorted each set of stimuli, and the sets were presented in random order.

Results

We tallied the number of times each categorization appeared for each stimulus set across participants. On average, people sorted each stimulus set 2.3 times, with an average of 2.5 categories per sort. Numbers varied across the three conditions and trends were in the direction predicted by the model, but we did not formally analyze these results due to lack of statistical power.

We compared human performance to the predictions of our model and the infinite mixture model. For each model, the hyperparameters were set to 0.5. Predictions for the infinite mixture model were derived by enumerating and ranking possible solutions by the probability of the solution given the data and parameters. Predictions for CrossCat were derived by enumerating and ranking all solutions including one or two ways of categorizing the objects. For the two-partition solutions, the probability of the whole solution contributed to the scores of both partitions.

Model predictions and human results for the 3-3 condition and the 3-2 condition are plotted in Figures 5 and 6 respectively. The best solution for the 3-3 condition according to CrossCat contains two categories, which are also the modal sorts made by people, as can be seen in columns 1 and 2. The third column shows that the best solutions according to the infinite mixture model were not preferred by either people or CrossCat. Notably, the best solution according to the infinite mixture model is to put all objects in the same category, highlighting the model's inability to deal with orthogonal systems of categories.

In the 3-2 condition (see Figure 6), the best solution according to CrossCat contained two categories, both of which were the modal solutions found by people. The second column shows model predictions ordered by the most frequent sorts by people. The infinite mixture model only predicts one of the two modal solutions according to people, while CrossCat predicted both of the two most frequent sorts by people. The human plot also reveals that the third most frequent sorting made by people was not particularly probable under either model. The third column shows the best sorts according to the infinite mixture model. The second and third best solutions according to the infinite mixture model do not predict human data, and are not particularly probable according to CrossCat.

In the 2 condition, both models agreed on the best solution, a single system with two categories, 1234 − 5678. This was also the most common sort by people, appearing twice as often as the second most frequent sort.

The results suggest that people's sorts cannot be explained by the infinite mixture model or any model that relies on a single system of mutually exclusive categories. The infinite mixture model performs well when there is only a single system of categories, but it is unable to predict results when there are multiple categorizations of the objects, as shown in the 3-3 and 3-2 conditions. In the 3-3 case, the infinite mixture model is unable to predict either of the modal sorts made by people. In contrast, CrossCat predicts the modal sorts by people in all three conditions, suggesting that the model captures the logic of how people discover multiple systems

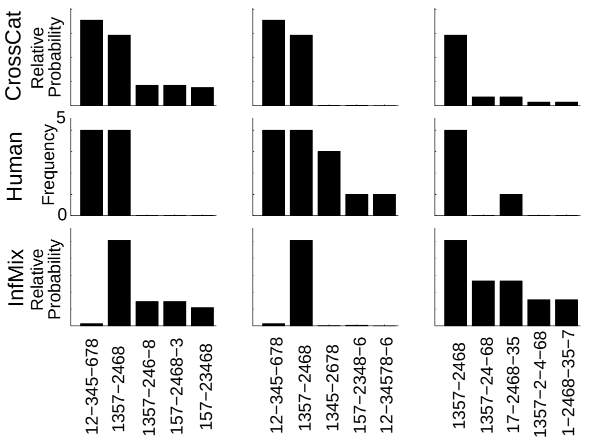

Figure 6: Results from the 3-2 condition. In each column, the results are sorted by a different standard: from left to right, the results are sorted by CrossCat predictions, human results, and the infinite mixture model. Note that the y-axis scale is the same across each row and the model solutions are plotted according to their relative probability, while the human data are plotted by frequency. On the x-axis is the partition of the objects, with different categories separated by dashes.

of categories when learning about a novel domain.

Discovering multiple representations for real-world domains

For this experiment, object-feature matrices were developed for two domains: foods and animals. These results include simulations for CrossCat and the infinite mixture model, and we contrast the systems of categories found by each model, paying particular attention to whether additional structure found by CrossCat reflects our intuitions about the structure of each domain.

The foods matrix contained 8 foods and 16 features (see Figure 1). The foods used were a subset of those in Ross and Murphy (1999), and the features were obtained by asking 7 participants to list features for each of the full set of objects in Ross and Murphy (1999). The objects and features used in this simulation were a subset chosen by the authors. The matrix was filled in by an undergraduate assistant, and a grey patch at matrix entry (i, j) indicates that feature j is true of object i.

The results (see Figure 2) are evocative of how the model could account for the results in Ross and Murphy (1999). In particular, these results show that CrossCat finds two systems of categories corresponding to taxonomic and situational groupings of foods, similar to those produced by people. Ross and Murphy (1999) also showed that people's inferences differed when reasoning about different properties: inferences about novel nutrients were guided by taxonomic knowledge and inferences about novel uses were based on situational knowledge. Our results suggest how these different kinds of inferences could be accounted for; by inferring the feature kind for a novel property and inferring unobserved values for the premises.

The second simulation addressed the domain of animals. The data matrix contained 22 animals and 106 features. This dataset is a subset of a larger matrix collected for an unrelated project. All of the features from the original data set were included in our simulation, and the animals were chosen to be representative of the original set.

The best solution found by CrossCat (Figure 7) identifies (a) a taxonomic system of categories, (b) a set of uninformative features and (c) an ecological system of categories. As a natural byproduct of Bayesian inference, our model computes the predictability of each feature given its associated system of categories. We arrange the features in order of decreasing predictability, so that features on the left in each system are generally the most diagnostic.

The taxonomic system is supported by appropriate features — 'has bones', 'is warm-blooded', 'lays eggs', etc. and divides the animals into birds, reptiles/amphibians, mammals, and invertebrates. The ecological system is best supported by features like 'is dangerous', 'is carnivorous', etc, and more weakly supported by features like 'lives in water' and 'flies'. These features are natural indicators of the animal categories it finds: prey, land predators, sea predators and air predators. Note that these categories nicely cross-cut the taxonomic ones. The third system consists of features of two kinds: those which were generally absent in the dataset (e.g. 'is canine', since no dogs were included) and those that were noisy with respect to the taxonomic and ecological systems. Interestingly, this system isolates one creature, 'frog', in its own category; this is because it is the lone animal with several features that are uninformative about all other animals in the set.

All three systems are intuitively appropriate and explain different important aspects of the domain. The infinite mixture model, when run on the same data, finds only the taxonomic categorization and would generalize the non-taxonomic features far more conservatively as it is forced to explain them as noisy taxonomic features.

Discussion

We presented a model that discovers multiple systems of categories in a single domain. Our model combines two insights that may seem incompatible at first. The vast majority of categorization models learn a single system of non-overlapping categories (e.g. Anderson, 1991), and one of the reasons for the popularity of this approach is that many real-world categories are mutually exclusive. There is no animal, for example, that is both a mammal and a reptile, and no food that is both a meat and a vegetable. The second insight is that categories can overlap. This approach also seems natural, since real-world categories do overlap: an animal can be a bird and a pet, and bacon is both a meat and a breakfast food.

Our model resolves the apparent contradiction between these perspectives. To a rough approximation, categories may often be organized into multiple systems of mutually exclusive categories. The first perspective recognizes the structure that is present within each of these systems, and the second perspective recognizes

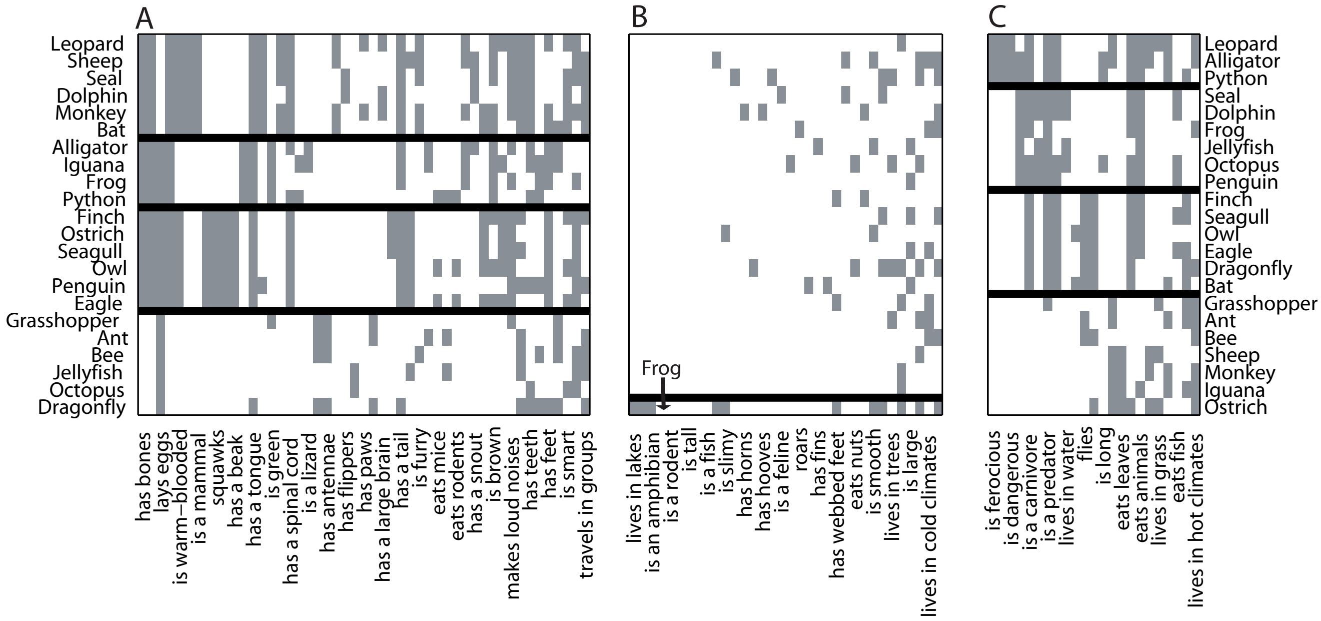

Figure 7: The best solution found by CrossCat for the animals data includes three systems of categories: (a) a taxonomic system, (b) a set of uninformative features and (c) an ecological system. Objects are labeled for the taxonomic and ecological systems. Features are presented in order of decreasing predictability. Due to space constraints, only every other feature is labeled.

that categories from different systems may overlap. Our model therefore inherits much of the flexibility that overlapping categories provide without losing the insight that the categories within any given system are often disjoint.

This knowledge is one structural constraint that can guide induction. Everyday intuitive leaps may also draw upon other structural constraints. The biological domain, for example, may be organized as a taxonomic tree when reasoning about the distribution of anatomical properties, but as a food web when reasoning about the distribution of novel diseases (Shafto and Coley, 2003; Shafto et al., 2005). Given a hypothesis space including several different structures, an extended version of our model should be able to group features into kinds and discover the structures that best explain each collection of features.

We have assumed that a given feature is related to only one of the many possible representations of a domain. In Figure 7, for example, some features are taxonomic, others are related to the system of ecological categories, but no feature is simultaneously taxonomic and ecological. Real-world features, however, may depend upon several systems of categories: whether an animal catches a disease may depend upon what it eats (i.e. its ecological category) and upon the genetic susceptibility shared by members of its taxonomic category. Extensions of our model can allow features to depend on one or more systems of categories, and we intend to pursue this in our future work.

Everyday human inference is remarkably flexible and accurate. The success of human reasoning is relies on our ability to acquire rich systems of categories, and to draw upon the system that is most relevant to any given task. Even state-of-the-art formal models fall far short of matching these abilities, but we believe that models with the ability to discover multiple systems of categories represent a step in the right direction.

Acknowledgments
We thank NTT communication sciences labs, the DARPA BICA program, the William Asbjornsen Albert memorial fellowship (C.K.), and the Paul E. Newton Chair (J.B.T.) for supporting this work.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98:409–429.

Heit, E. and Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:411–422.

Martin, J. D. and Billman, D. O. (1994). Acquiring and combining overlapping concepts. *Machine Learning*, 16:121–155.

Rasmussen, C. E. (2002). The infinite Gaussian mixture model. In *Advances in Neural Processing Systems 13*. MIT Press.

Ross, B. H. and Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, 38:495–553.

Shafto, P. and Coley, J. D. (2003). Development of categorization and reasoning in the natural world: Novices to experts, naive similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29:641–649.

Shafto, P., Kemp, C., Baraff, E., Coley, J. D., and Tenenbaum, J. B. (2005). Context-sensitive reasoning. In *Proceedings of the 27th annual conference of the Cognitive Science Society.*

Yamauchi, T. and Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26:776–795.