ELSEVIER
Volume 120, Issue 1, July 2011
ISSN 0010-0277
pp. 1–148
COGNITION

Cognition 120 (2011) 1–25

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

A probabilistic model of cross-categorization

Patrick Shafto [a,*], Charles Kemp [b], Vikash Mansinghka [c], Joshua B. Tenenbaum [d]

[a] *Department of Psychological and Brain Sciences, 317 Life Sciences Building, University of Louisville, Louisville, KY 40292, United States*
[b] *Carnegie Mellon University, United States*
[c] *Navia Systems, United States*
[d] *Massachusetts Institute of Technology, United States*

ARTICLE INFO

Article history:
Received 21 April 2010
Revised 7 February 2011
Accepted 11 February 2011
Available online 4 March 2011

Keywords:
Cross-categorization
Categorization
Bayesian model

ABSTRACT

Most natural domains can be represented in multiple ways: we can categorize foods in terms of their nutritional content or social role, animals in terms of their taxonomic groupings or their ecological niches, and musical instruments in terms of their taxonomic categories or social uses. Previous approaches to modeling human categorization have largely ignored the problem of cross-categorization, focusing on learning just a single system of categories that explains all of the features. Cross-categorization presents a difficult problem: how can we infer categories without first knowing which features the categories are meant to explain? We present a novel model that suggests that human cross-categorization is a result of joint inference about multiple systems of categories and the features that they explain. We also formalize two commonly proposed alternative explanations for cross-categorization behavior: a features-first and an objects-first approach. The features-first approach suggests that cross-categorization is a consequence of attentional processes, where features are selected by an attentional mechanism first and categories are derived second. The objects-first approach suggests that cross-categorization is a consequence of repeated, sequential attempts to explain features, where categories are derived first, then features that are poorly explained are recategorized. We present two sets of simulations and experiments testing the models' predictions about human categorization. We find that an approach based on joint inference provides the best fit to human categorization behavior, and we suggest that a full account of human category learning will need to incorporate something akin to these capabilities.

1. Introduction

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 including meats and vegetables, or as a system of situational categories including 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 ensure that meals were nutritionally balanced and consulting the situational system to ensure that there were foods that were appropriate for the different times of the day (Ross & Murphy, 1999). In nearly every 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.

Several lines of behavioral research have shown how the ability to think about objects in multiple cross-cutting ways is critical to flexibility in inductive reasoning, planning, and problem solving. Heit and Rubinstein (1994) showed that when reasoning about anatomical properties

[*] Corresponding author. Tel.: +1 502-852-6197; fax: +1 502-852-8904.
E-mail address: p.shafto@louisville.edu (P. Shafto).

0010-0277/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2011.02.010

(e.g. "has a liver with two chambers that act as one") people draw upon taxonomic knowledge about animals to guide inferences, but when reasoning about behavioral properties (e.g. "usually travels in a back-and-forth, or zig-zag, trajectory") people draw upon ecological knowledge (see also Medin, Coley, Storms, & Hayes, 2003; Shafto & Coley, 2003). Barsalou (1983) demonstrated that people derive cross-cutting categories, such as 'things to take out of the house in case of a fire', in the service of goals, and that these goal-based categories are used in planning (Barsalou, 1991). Chi, Feltovich, and Glaser (1981) showed that expert physics students augment similarity-based event categories with categories based on abstract physical principles, and these abstract categories play an important role in expert problem solving.

Despite the empirical evidence that people categorize objects in multiple ways for different purposes, formal models of categorization and category learning have not typically addressed this dimension of cognitive flexibility. Most models represent only a single system of mutually exclusive categories for a given domain, and focus on simplified behavioral tasks where a single way of categorizing is sufficient to perform well (Anderson, 1991; Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, Palemeri, & McKinley, 1994; but see Martin & Billman, 1994). Consider, for example, the animals in Fig. 1. Standard approaches to categorization infer a single system of mutually exclusive categories such as that shown in Fig. 2. This captures intuitively compelling taxonomic categories such as mammals, birds, amphibians and reptiles, and invertebrates. However, while this is the only way that standard approaches can view the data, it is not the only way that people can view the data. People are able to see the domain in multiple ways: as taxonomic categories for sure, but also as ecological categories capturing animals that live primarily on land, sea, and air (see Fig. 3), and these kinds of categories are necessary to support basic cognitive functions such as inference (Heit & Rubinstein, 1994). In this paper, we investigate the cognitive basis of people's abilities to cross-categorize.

Cross-categorization presents a fundamental challenge: how a learner can infer categories without knowing in advance which features they are meant to explain? We propose that people address this challenge by jointly inferring one or more systems of categories and the features that the systems explain. We contrast this proposal with two intuitively compelling alternatives. The first proposes that cross-categorization arises as a consequence of selective attention. On this account, people attend to different features of the stimuli in different contexts, and categorize objects differently depending on the features currently under consideration. This general approach to cross-categorization builds on the many previous models that have emphasized the role of selective attention in categorization (e.g. Love, Medin, & Gureckis, 2004; Medin & Schaffer, 1978; Nosofsky, 1984; Shepard, Hovland, & Jenkins, 1961). The second alternative formalizes the intuition that people's cross-categorization behavior arises as a consequence of successive attempts to account for poorly

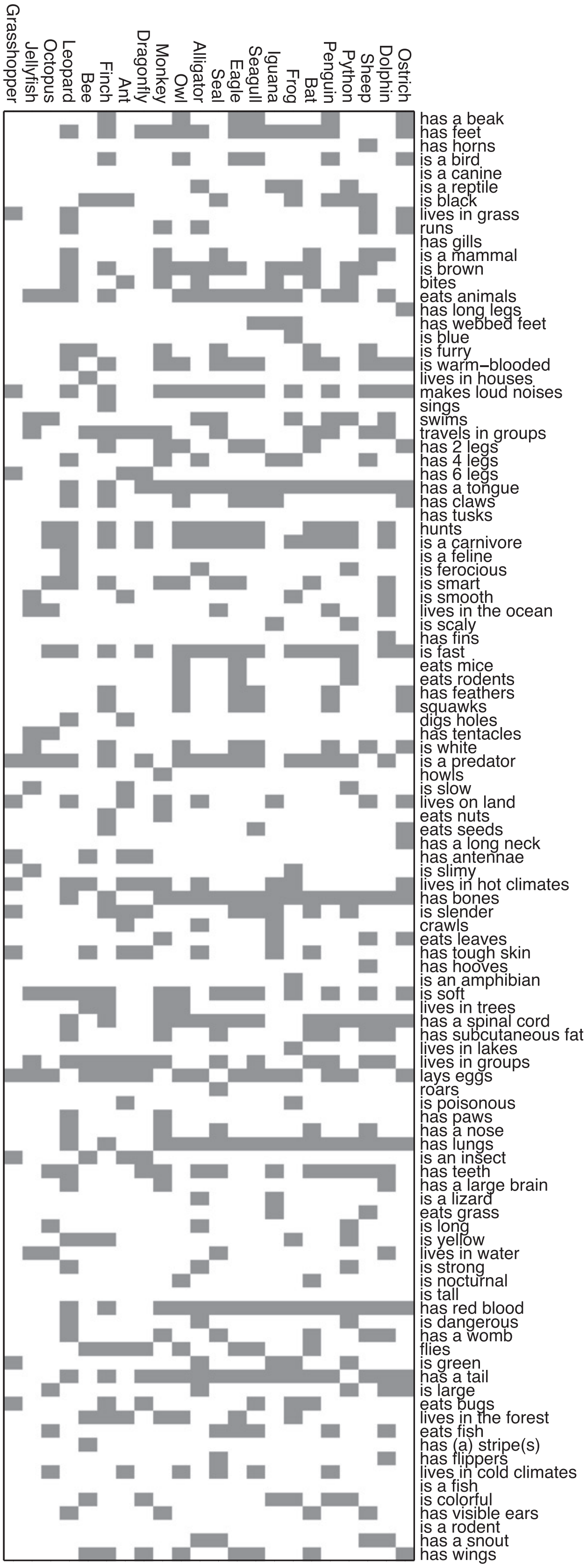


Fig. 1. Raw data provided to the models for the animals condition. For each object-feature pair, a gray square indicates that pair was judged to be true and a white square indicates that feature was judged to be false. For example, ostriches, penguins, and seagulls were all judged to have beaks.

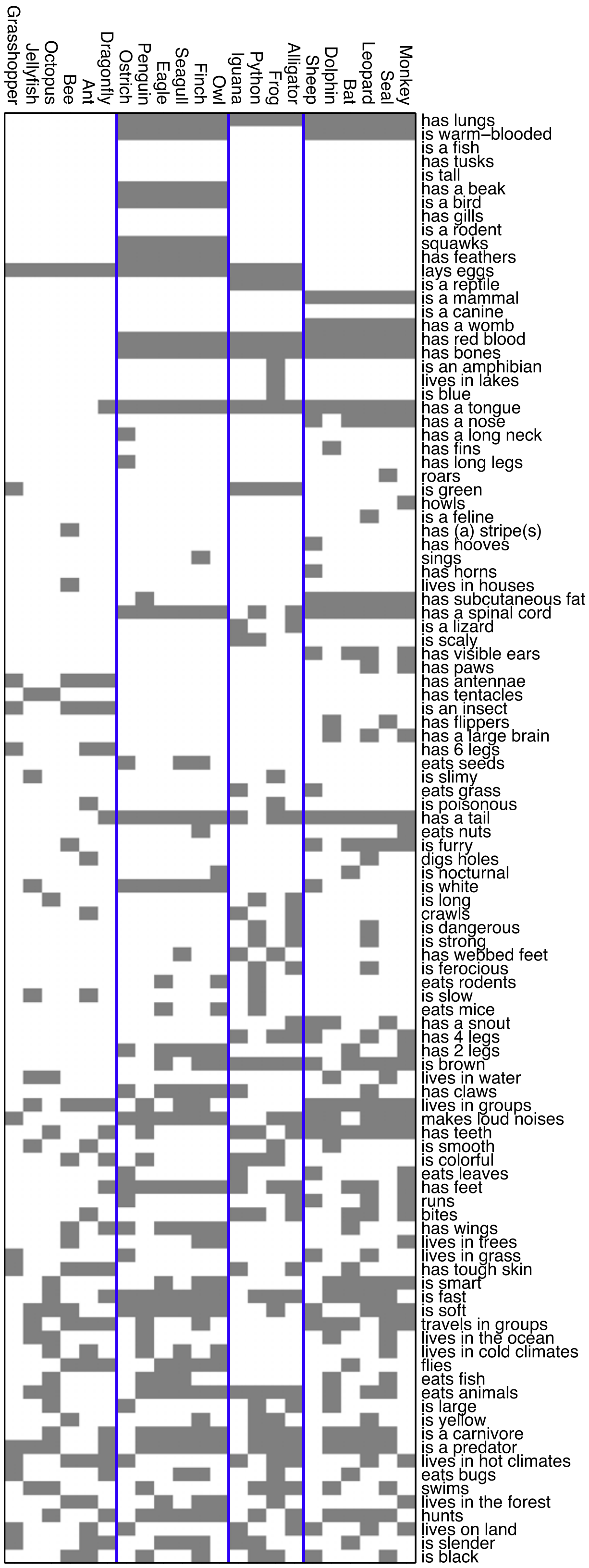


Fig. 2. The best solution to the animals data according to the single system model. The model finds four intuitively compelling categories: mammals, invertebrates, birds, and reptiles/amphibians. Features are sorted from those that are best explained by these categories to those that are most poorly explained. Note that while many features such as "has lungs" are well-explained, many that seem structured, such as "is a carnivore" are not.

explained features. On this account, people categorize objects, then identify features are not well-explained by these categories. They recategorize based on those poorly explained features, repeating if necessary, and the result is multiple different systems of categories which apply to different contexts. The critical difference between our approach and these two alternatives is that our approach alone relies on joint inference about categories and feature kinds. The selective attention approach first fixes a set of features then identifies categories that are supported by these features. The repeated recategorization approach first fixes a system of categories, then learns feature kinds and additional systems of categories to account for data that are not well-explained by the initial system. Our approach fixes neither the categories nor the feature kinds, and allows both to mutually constrain each other.

To illustrate the idea we develop a new model of categorization, CrossCat, that jointly infers cross-cutting systems of categories that best explain the data, together with the subset of features that each explains. We contrast this approach with a representative traditional approach to category learning, and two additional models which attempt to explain cross-categorization as a consequence of attentional mechanisms or sequential recategorization. We present two sets of experiments in which we investigate cross-categorization, and each model's ability to explain the human categorization behavior. To preview our results, we find that the CrossCat model accounts best for the data, suggesting that cross-categorization relies on joint inference about categories and features.

2. Empirical evidence for cross-cutting systems of categories

Research on categorization has tended to focus on either identifying category structure in real-world domains (e.g. Medin et al., 2005; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Ross & Murphy, 1999), or understanding category learning using artificially constructed stimulus sets (e.g. Shepard et al., 1961). Studies of real-world domains have shown that people learn richly structured categories, and have identified cross-cutting systems of categories in several domains. Studies of artificial category learning, on the other hand, rarely investigate how cross-cutting categories are learned. We therefore focus on the evidence suggesting that cross-cutting systems of categories underlie people's understanding of many real-world domains.

Ross and Murphy (1999) describe a series of experiments that provide evidence for people's use of both taxonomic and situation-based categories when thinking about foods. They showed that, when presented with individual foods (e.g. bagel) and asked to name categories to which they belong, people list about 50% taxonomic categories (e.g. grain) and 42% situation-based categories (e.g. eaten for breakfast). These results were supported by a variety of other measures indicating that people consider foods to be good examples of both taxonomic and situation-based categories, and that people reliably sort foods into systems of taxonomic and situation-based

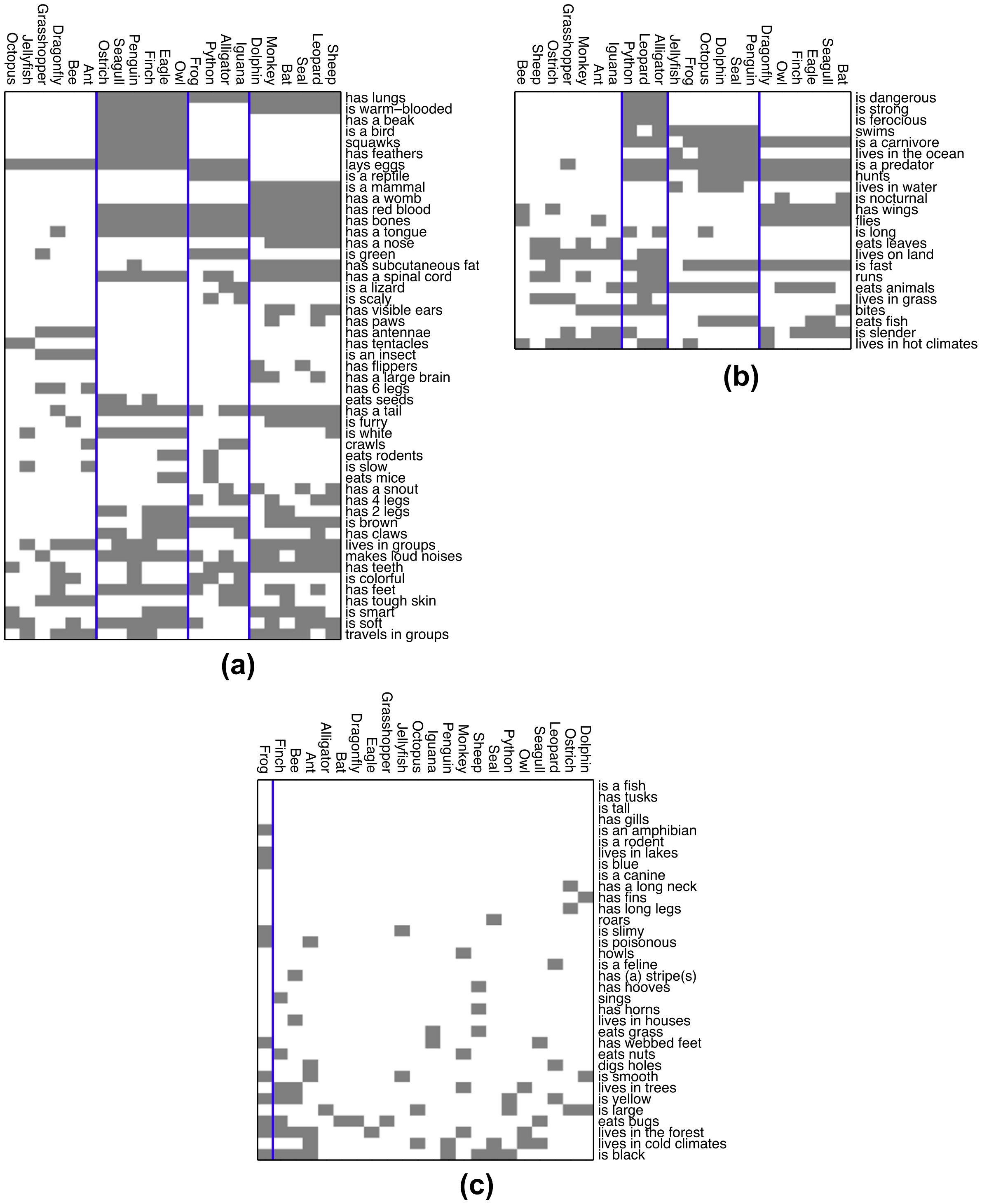


Fig. 3. The best solution for the animals data according to CrossCat. The solution includes three systems of categories. Panel (a) shows that CrossCat discovers a taxonomic system that is identical to that found by the single system model. Panel (b) shows a system of ecological categories including land predators, aquatic predators, aerial predators, and prey. This system accounts for features such as "is dangerous", "lives in the ocean", and "is a carnivore". Panel (c) shows a system mainly consisting of noisy and irrelevant features, in which nearly all of the objects are grouped together.

categories. Finally, Ross and Murphy showed that both taxonomic and situation-based category labels prime retrieval of category members, and that both kinds of categories can guide inferences in inductive reasoning. Similar results have been obtained by researchers investigating the domains of biology (Boster & Johnson, 1989; Medin et al., 2005; Proffitt, Coley, & Medin, 2000; Shafto & Coley, 2003) and person categorization (Nelson & Miller, 1995; Smith, Fazio, & Cejka, 1996; Zarate & Smith, 1990), and in research with children (Nguyen, 2007; Nguyen & Murphy, 2003). Together, these results provide a compelling demonstration that people spontaneously produce, consistently agree about, and reason based on multiple systems of categories.

There is compelling evidence that people cross-categorize. But what gives rise to these abilities? In the next section, we will discuss previous models of category learning in the context of cross-cutting categorization.

3. Previous approaches to category learning

Previous models of categorization have taken a wide variety of approaches to the problem of category learning, including those based on learning rules (e.g. Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Nosofsky et al., 1994), inferring prototypes (e.g. Posner & Keele, 1968), storing exemplars (e.g. Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1984, 1986), and simplicity-based approaches (Pothos & Chater, 2002; Pothos & Close, 2008), as well as many which interpolate between approaches (e.g. Anderson, 1991; Love et al., 2004). These approaches differ in many ways, including whether they focus on supervised or unsupervised learning (see Griffiths, Canini, Sanborn, & Navarro, 2007; Love, 2002). However, these models are similar in that they focus on learning a single system of categories (but see Pothos & Close, 2008), limiting their applicability to modeling learning in real-world domains.

Though they seem unlikely to account for cross-categorization, these approaches provide important insights into the basic problems of supervised and unsupervised category learning. In this paper, we focus on the problem of unsupervised learning. In unsupervised categorization, no labels are provided for the data; rather, based on the stimulus structure, participants are asked to discover both how many categories there are and which objects belong in which categories (cf. Rosch et al., 1976). This contrasts with supervised categorization, where examples are paired with category labels, and participants are tested on how long it takes to learn to correctly categorize examples within some error bounds. Both are important, though unsupervised categorization is arguably the more challenging because of the ambiguity about the number of categories. Additionally, it seems closer to the problems people face in the real world. We view supervised learning as a special case of unsupervised learning, though we will not discuss the extension of our model to the supervised (or semi-supervised) setting.

We discuss in detail one standard approach to unsupervised categorization, a variant of Anderson's (1991) Rational Model of Categorization that we call the single system model. We will use this as a basis from which we will explore some different explanations of people's abilities to cross-categorize. Particularly, we contrast three intuitive approaches that attempt to explain human cross-categorization as joint inference about multiple systems of categories and the features that they explain, as a consequence of attentional mechanisms operating over features, or as a result of sequential attempts to form categories that account for features that were not explained by previous systems.

3.1. The single system model

Anderson proposed a solution to the problem of unsupervised categorization based on a probabilistic model known in machine learning as the infinite mixture model (Rasmussen, 2000) and in statistics as the Dirichlet process mixture model (Neal, 1998). Here we call it the single system model. It embodies two intuitions about category structure in the world: the world tends to be clumpy, with objects clustered into a relatively small number of categories, and objects in the same category tend to have similar features.[1] Formalizing categorization as Bayesian inference over the hypothesis space of systems of categories described by this model automatically implements an Occam's razor-like tradeoff between two goals: minimizing the number of clusters posited and maximizing the relative similarity of objects within a cluster.

This approach has two additional aspects useful for modeling human categorization. First, as when people discover categories in the real world, the number of categories does not have to be known in advance. The model is able to infer the correct number of categories for a given set of objects, and that number grows naturally as new objects are introduced. Second, the model has a minimal number of free parameters (2) allowing for straightforward implementation, and parsimonious explanations for predicted phenomena.[2]

Here we present a sketch of the model's critical aspects needed to intuitively understand it and the extensions we develop for cross-categorization. Full mathematical details are provided in Appendix A.[3] The model assumes as input a matrix of objects and features, D, where entry $D_{o,f}$ contains the value of feature f for object o. These values could take on a variety of forms, however, we focus on the special case of binary features. Thus, for our purposes $D_{o,f}$ indicates whether object o has feature f. For example, Fig. 1 shows a data set representing the features of different kinds of animals.

The single system model assumes that there are an unknown number of categories that underlie the objects, and that objects within a category tend to have the same value for a given feature. The goal of the model is then to infer likely system of categories, w, or sometimes we may infer the single most probable system of categories (the maximum a posteriori solution or MAP).

The probability of an assignment of objects to categories given the data, $p(w|D)$ depends on two factors: the prior probability of the assignment of objects to categories, and the probability of observed data given the categories. Formally, the probability of a system of categories w given the data D is,

$$p(w|D,\alpha,\delta) \propto p(w|\alpha)p(D|w,\delta). \tag{1}$$

The probability of a particular system of categories $p(w|\alpha)$ captures a preference for a small number of categories relative to the total number of objects, and the

[1] Anderson's approach also involved implementing plausible constraints on cognition, but we will not address his proposed constraints in this paper.

[2] Nosofsky (1991) describes Anderson's model as having one parameter for each feature plus one coupling parameter. Because we see no *a priori* reason to have different parameter values for different features, we set all features to the same value. The coupling parameter is the second free parameter in our characterization.

[3] See also Appendix B for formal details of other implementations not discussed in the text.

strength of this preference is governed by the parameter α. The term $p(D|w,\delta)$ assesses the probability of the observed feature values, given the system of categories. The model assumes that, for objects in different categories, the values they take on a particular feature are independent (i.e. learning that reptiles tend to have legs does not tell you whether or not birds or mammals will have legs). Therefore, different categories can be treated separately. For simplicity, it is typically assumed that features are conditionally independent given the underlying system of categories, which means that within a category different features can also be treated separately. Thus, assessing the probability of the observed features given the categories reduces to assessing the probability of the observed values for a feature within a single category multiple times. The probability of the observed data given a system of categories $p(D|w,\delta)$ prefers that objects in a category have the same values for each feature, and the strength of this preference depends on the parameter δ.

The model captures a tradeoff between two competing factors. The term $P(w|\alpha)$ specifies a preference for simple solutions that use a small number of object categories. The term $P(D|w,\delta)$ favors solutions that explain the data well, and tends to prefer more complex solutions. By combining these terms, we arrive at a model that attempts to find the simplest solution that adequately accounts for the data. This approach has strong relations to simplicity-based approaches (Pothos & Chater, 2002; Pothos & Close, 2008), as well as prototype and exemplar models (see Nosofsky, 1991 for a detailed discussion). Fig. 2 shows how the tradeoff between these factors leads the model to group animals in Fig. 1 into a small number of familiar large-scale taxonomic categories: mammals, birds, reptiles and amphibians, and insects and invertebrates.

Once the prior $p(w|\alpha)$ and the likelihood $p(D|w,\delta)$ have been formalized, categorization can be treated as a problem of finding a w that has high probability in the posterior distribution $p(w|D, \alpha,\delta)$ (see Eq. (1)). In the this paper, we address this search problem using Markov Chain Monte Carlo (MCMC; Gelman, Carlin, Stern, & Rubin, 1995), a stochastic hill-climbing approach to inference, operating over the complete data set. Intuitively, the algorithm searches through the space of possible category assignments for w, and tends to spend more time searching around high probability assignments. Specifically, the algorithm imagines proposing small changes to the current solution; for example, moving an object from one category to another or splitting an existing category into two, and tends to choose the better of the imagined and current states. We are not committed to the psychological reality of this particular search process as a model of the learning process or development, and there are a number of psychologically plausible approaches for approximate inference (see also Anderson, 1991; Sanborn, Griffiths, & Navarro, 2006).

4. Formal models of cross-categorization

Although it seems clear that traditional models of categorization cannot account for people's ability to cross-categorize, there is little agreement on what does account for these abilities. Here we contrast three potential explanations: that cross-categorization is a consequence of joint inference about categories and the features they explain, that cross-categorization is a side effect of attentional mechanisms, or that cross-categorization is a consequence of repeated attempts to explain observed features. We present a new model, CrossCat, which is based on joint inference about cross-cutting systems of categories that explain different features. For example, in Fig. 3a the feature "lays eggs" is associated with the taxonomic system of categories, and the feature "is dangerous" is associated with the ecological system of categories. Although some features are plausibly associated with multiple systems—for example, the feature "has blubber" seems to depend on being a mammal and an aquatic creature—assuming that each feature is assigned to just one system provides a simple starting point for investigating cross-categorization. As discussed in the General Discussion, future approaches could relax this assumption in several ways.

To test whether full cross-categorization is necessary to account for people's behavior, we also introduce a family of models that attempt to explain people's behavior as a consequence of attentional mechanisms. In addition, we introduce a third model, based on sequential attempts to infer categories to account for poorly explained features. By contrasting the ability of these models to predict human behavior, we will be able to address the cognitive basis of cross-categorization. Particularly, is people's behavior more consistent with joint inference about cross-cutting systems of categories, or alternatives that are not fully committed to cross-categorization but may lead to apparent cross-categorization behavior?

4.1. Cross-categorization through joint inference over multiple systems of categories

The first approach is *CrossCat*, a model that infers multiple systems of categories. CrossCat extends the approach of the single system model by allowing there to be potentially many systems of categories, where each system categorizes the objects based on a subset of the features. Under this approach, systems of categories are considered contemporaneously, and exist to explain different subsets of features.

Given the animals data in Fig. 1, our goal is to infer how many systems best describe the data, to decide which features belong to which system, and to discover a categorization of the objects for each system. Our approach extends the single system model by adding a method for assigning features to different systems. The probability of a particular assignment of features to systems $p(s|\alpha)$ uses the same probabilistic process as the assignment of objects to categories. As for the single system model, the parameter α governs how strong a preference we have for a small number of systems. This is the key difference between CrossCat and the single system model. CrossCat allows there to be many different systems, while the single system model assumes that all features belong to the same system.

Within each system, CrossCat is effectively equivalent to the single system model. That is, the probability of a single system depends on the probability of the assignment of the objects to categories, $p(w|\alpha)$, and the probability of the observed values of the features in that system, given the categories. Importantly, however, there are as many sets of categories as there are systems in s. Thus, the probability of all of the categories $p(\{w\}|\alpha)$ is the product of the probabilities of each individual category w where the object categories for different systems may differ.

The final component is the probability of the data given the systems of categories, $P(D|\{w\},s,\delta)$. Each system of categories has an associated subset of features, and for that system there is a particular assignment of objects to categories. As for the single system model, we assume that values of features across categories are independent, and for convenience assume that different features are independent within a category.

Combining these three factors, we specify the probability of a particular assignment of features to systems and categories for each systems,

$$p(s,\{w\}|D,\delta,\alpha) \propto p(s|\alpha)p(\{w\}|\alpha)p(D|s,\{w\},\delta). \qquad (2)$$

Like the single system model, CrossCat trades off the preference for complexity, favored by the $p(D|\{w\},s,\alpha)$, against the preference for simplicity, favored by the priors $p(s|\alpha)$ and $p(\{w\}|\alpha)$, where α captures the strength of this preference. In cases where the data are best explained by a single system of categories, CrossCat arrives at the same solution as the single system model. However, when the data are best explained by multiple systems of categories, CrossCat is able to discover multiple systems, and which features are explained by each.

Because it is a generalization, CrossCat inherits many of the strengths of the single system model. It embodies simple assumptions about why categorization works: people have a preference for smaller numbers of categories that tend to explain the data well. CrossCat also has a minimal number of free parameters. Because we have no a priori grounds for expecting greater complexity in the number of systems or the numbers of categories per system, we assume that a single α parameter governs both the prior assigning objects to categories and the prior assigning features to systems. CrossCat thus has the same number of free parameters as the single system model.[4] It is also related to prototype and exemplar approaches, though the relationship is somewhat more complicated, and we return to this issue in the general discussion.

By using multiple systems of categories, CrossCat can explain patterns in how features covary across objects that the single system model cannot explain—patterns that reflect different kinds of category structures underlying the domain. Fig. 3 shows many examples of this behavior. For example, bats and dolphins share some features because they are both mammals, while bats and eagles share other features because they are flying creatures. CrossCat explains the former pattern by grouping bats and dolphins together in the taxonomic system of categories (Fig. 3a), and the latter pattern by grouping bats and eagles together in the ecological system of categories (Fig. 3b).

As for the single system model, we perform inference using MCMC over the complete data set. The approach is broadly similar to inference in the single system model. The algorithm searches for better solutions by imagining moving either some objects to new categories or features to new systems, preferring moves that improve on the current state. Although we think that roughly similar heuristics may sometimes be used by human learners, we make no claims about the psychological reality of this particular inference algorithm, only the importance of joint inference over multiple systems of categories for explaining people's behavior. Importantly, the algorithm described tends to converge to solutions that jointly optimize the systems of categories and the features that they explain.

4.2. Features-first: attending to subsets of features

The second account we consider is based on the idea that cross-cutting systems may arise from a features-first approach. In this approach, attention to different subsets of information leads to identification of subsets of features, and then categories are derived. We formalize this intuition as an extension of the single system model. The attention model is essentially the same as the single system model, but instead of applying the model to the full data set, we consider only a subset of the features at any given time.[5] Thus, under the attention model, cross-categorization is explained by constraining features first, then inferring the best system of categories to explain those features.

We implemented two versions of this proposal, which correspond to different ways of choosing subsets of features. The first version follows the simplest possibility, with subsets chosen by random selection. The second version(s) implement feature choice based on a simple notion of search. The first feature is chosen at random, and subsequent features are chosen to align well with the previously selected feature. We implemented feature selection in a variety of different ways, which are summarized in detail in Appendix A (we tried a total of 6 versions which varied on how feature similarity was computed and whether choices were based on maximal or average similarity).[6] In the results, we present only the subset search model that showed the best performance.

In each of these models, a subset of the features is chosen. We formalize this choice using the same prior as in CrossCat, where the probability of choosing n features is the probability of choosing n features in one group, and the remaining $F - n$ features in a second group (which is ignored). The two models differ only in how the subset is chosen: in the subset attention model we consider the possibility of randomly choosing subsets of different sizes, and in the subset search model we consider choosing subsets of

[4] In all of the work we present in this paper, the parameters were set to $\alpha = 0.5$ and $\delta = 0.25$.

[5] We also implemented a features-first approach in which the single system model was used to categorize features, then applied to these subsets of features to infer categories. This model did not perform well, and we only report the subsets of features results.

[6] In addition to the six versions that are described in Appendix A, we tried an additional 12 versions of a related but different implementation, which did not perform better and are described in Appendix B.

similar features. For both versions, search is formalized as for the single system model, and full formal details are presented in Appendix A.

4.3. Objects-first: repeated recategorization

The final approach we consider suggests that cross-categorization arises from multiple applications of simple categorization, where each successive application focuses on the features that were not well-explained by previous systems of categories. On this account, feature kinds are derived from object categories; instead of being jointly optimized, categories are given priority and feature kinds are secondary.

We formalize these intuitions as an extension of the single system model, with an additional parameter, corresponding to the criteria for determining when a feature is well enough accounted for. For each iteration, the probability of categories is defined as for the single system model (see Eq. (1)). The number of iterations is determined by the parameter, together with the data. We formalize goodness of a feature, g, by considering the relative probability of each feature compared to the feature that is best explained by the system, $g = \frac{p(D_f|w,\delta)}{max_f[p(D_f|w,\delta)]}$. Thus, goodness ranges from 1, for the best explained feature, down in theory to 0. The value of the free parameter determines which features are good enough. All of the features with g less than the cutoff are recategorized, resulting in additional systems of categories. Within each run, search is formalized as described for the single system model, and full formal details are presented in Appendix A.

5. Artificial data simulations

To highlight differences between the models, we constructed three data sets. Each data set included eight objects and six features, and was designed to have either a single system or multiple systems of categories. The first data set was constructed to have two systems with three orthogonal categories each. Fig. 4a shows the raw data on the top. Note that for the first three features, clean blocks of features appear, while the last three features have a diagonal pattern. The best solution according to CrossCat is shown on the bottom, and shows that both sets of features resolve into perfectly clean systems of categories. A second data set (see Fig. 4b) was also constructed to have two systems of categories, but with a somewhat less clean solution. The first system was based on the first three features and had the same structure as in the previous example. The second system was based on the last three features, and was designed to have two categories. The two categories had one defining feature (in the figure, antenna), and two noisy variations. A third data set was constructed to have one system of two prototype-like categories. Fig. 4c shows the raw data on the top, and the MAP solution according to CrossCat on the bottom.

To illustrate why CrossCat prefers these solutions, take the case of Fig. 4b. The best solution has two systems, the first with three categories and the second with two categories. To discuss the components of the score, we use log probabilities, which range from negative infinity to zero, and are equivalent to probabilities of 0 and 1, respectively. Higher log probabilities thus indicate more probable outcomes. The total score for the best solutions is composed of five parts, the prior probability for the feature kinds $p(s|\alpha)$, prior probabilities for each of the systems of categories $p(\{w\}|\alpha)$, and the probability of the data given the system of categories for each of the two systems $p(d|s,\{w\},\delta)$. The best solution is broken down into these components in Fig. 5a. The prior for feature kinds $p(s|\alpha)$ contributes a log probability of -5.09. The prior on systems of categories $p(\{w\}|\alpha)$ prefers simplicity, so the probability of the three category system is less than the two category system, -9.67 versus -6.78, respectively. The data are more likely when, for a single feature, all of the objects in a category have the same value. Under the likelihood $p(d|s,\{w\},\delta)$, the three category system is more probable than the two category system, -8.51 and -11.46. This embodies the tradeoff between the prior and the likelihood – the three

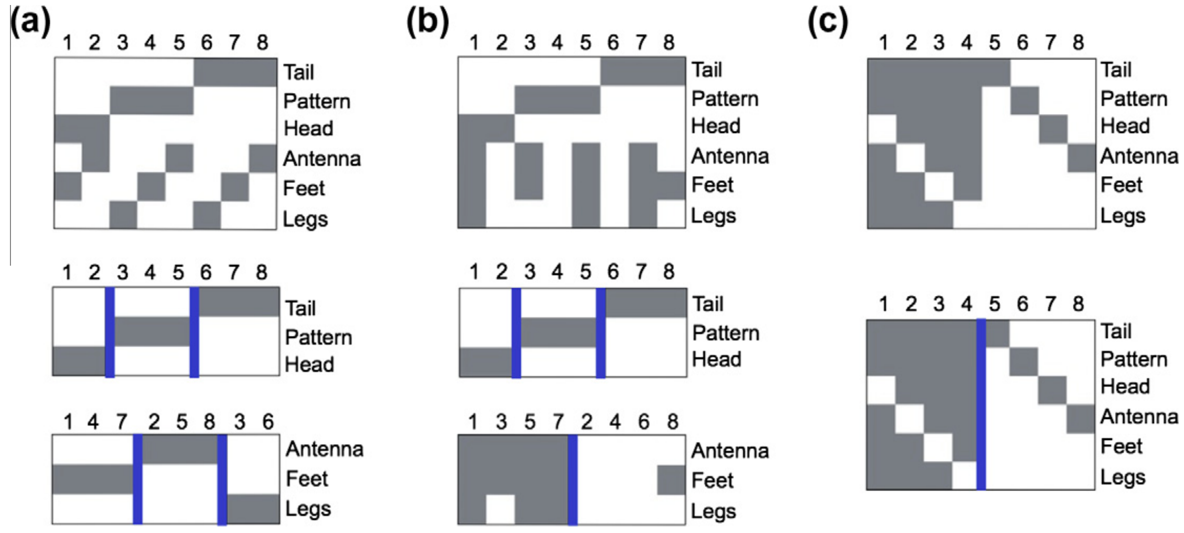


Fig. 4. Matrices representing the structure of the different stimulus sets. The matrices correspond to the stimuli for: (a) 3/3, (b) 3/2, and (c) 2 category conditions. Unsorted matrices are presented on the top, and on the bottom are the best solutions according to CrossCat.

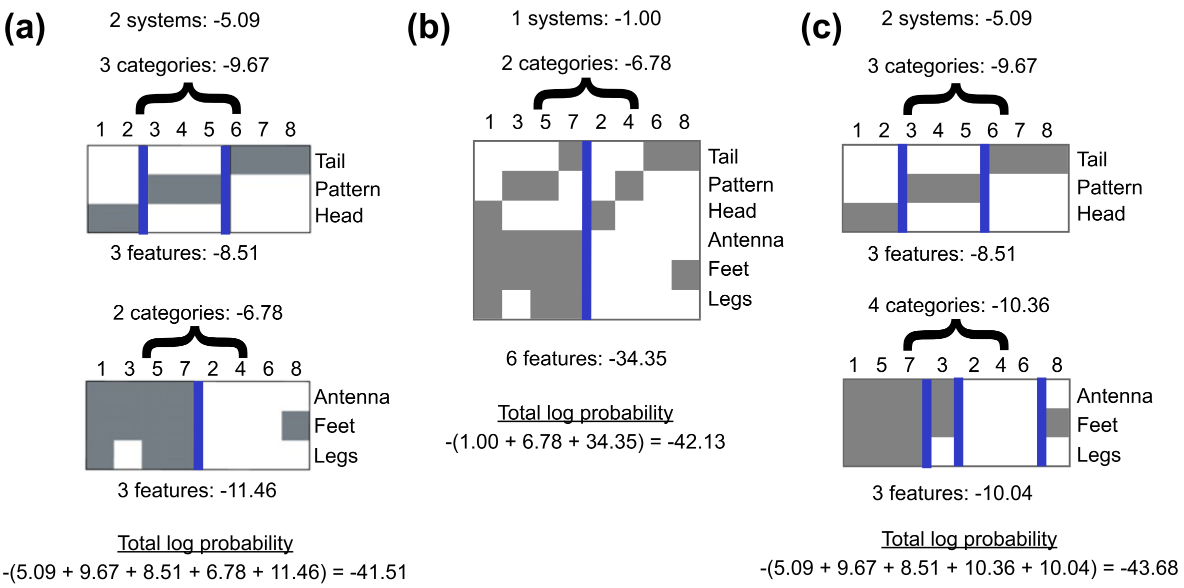


Fig. 5. Three possible solutions to the 3/2 data. The solutions correspond to: (a) the best solution, (b) a simpler solution that accounts for the data less well, and (c) a solution that accounts for the data better, but at the cost of greater complexity.

category system is more complex (and hence has a lower prior) but fits the data better, while the two category system is simpler but fits the data less well. The log probability of the total solution can be calculated by adding up these contributions, −41.51.

We can contrast this solution with possible alternatives where the structure is simpler or more complex to show how the prior trades off with the likelihood. For the first case, we consider a single system with the two categories discussed above, shown in Fig. 5b. There are three components: the prior on the feature kinds, the prior on the system of categories, and the likelihood of the data. Because the structure on features is simpler and there is one less system, the prior prefers this case over the solution described above. The log probability for the feature kinds and system of categories are $log(p(s|\alpha)) = -1.00$ and $log(p(w_1|\alpha)) = -6.78$. However, the simpler solution comes at the cost of explaining the data well, and as a result the log probability of the data under the likelihood is relatively low, $log(p(d|s,\{w\},\delta)) = -34.35$. The result is a total solution that is less probable, −42.13, than best solution.

Another set of alternative solutions could be formed by choosing a solution that is more complex, but fits the data better. Consider for example the case where we add two categories to the two category system in Fig. 5a. The result would be a four-category system that could explain the data without exceptions, as in Fig. 5c. In this case, the probability of the feature kinds, and the three category system are going to be identical to the original, $log(p(s|\alpha)) = -5.09$ and $log(p(w_1|\alpha)) = -9.67$. However, because the second system now has four categories, it has relatively low prior probability, $log(p(w_2|\alpha)) = -10.36$. The question is whether the ability to explain the data better overcomes this penalty. The probability of the data for the first system is unchanged, −8.51, while the probability of the data for the second system increases, −10.04. In this case, the ability to explain the data better does not outweigh the added complexity and the probability of the total solution, −43.68, is also lower than the log probability of the best solution.

5.1. Contrasting the predictions of CrossCat and the single system model

CrossCat's approach, based on multiple systems of contrasting categories, leads to distinctly different predictions than each of the other models. Contrasting CrossCat with the single system model provides insight into the effects of multiple systems of categories. For the data in Fig. 4c, the single system model and CrossCat predict the same solutions because the single system model is a special case of CrossCat, and the data only warrant a single system of categories. However, for the data in Fig. 4a and b, the models make sharply divergent predictions. Unlike CrossCat, the single system model cannot recognize multiple systems. Because the systems of categories are orthogonal to each other, the single system model cannot see any structure. The single system model predicts that the best solution is a single category with all of the objects.

For the data in Fig. 4b, CrossCat and the single system model again predict sharply different systems of categories. In this case, the two systems of categories are not orthogonal to each other. The single system model predicts that the simpler two category system is the best solution. However, the next four best systems of categories are variants of the best. For example, the second and third best solutions were 1357–24–68 and 17–2468–35. These results reflect the single system model's inability to recognize substructures that are less prominent in the data.

These predictions highlight important differences between CrossCat and the single system approach. As demonstrated by the third data set, CrossCat can learn that a single system of categories is the appropriate structure, when that is the best explanation for the data. However, when the data are best described by multiple systems of categories, as for the first two data sets, CrossCat learns which systems of categories best describe the data. The first data set shows that CrossCat learns rich category structure in a case when the single system model learns that there is essentially no structure. The second data set shows that even when there are not perfectly orthogonal structures, CrossCat's ability to learn different systems for different subsets of the features predicts markedly different solutions than a single system model. Critically, the predictions provided by CrossCat are not equivalent to taking the top *N* predictions of the single system model. CrossCat predicts a qualitative change in the ordering of hypotheses.

5.2. Contrasting the predictions of CrossCat and the subset attention and subset search models

In contrast with the single system model, the subset attention model can attend to subsets of features. However, the ability to attend to subsets of features does not lead to the same predictions as CrossCat. Because most of the possible subsets of the the data in Fig. 4a do not form good clusters, the random subset model predicts that a single cluster with all of the data is highly likely, like the single system model. For the data in Fig. 4b, the subset attention model predicts that the system with two categories is highly likely, but not the system with three categories that is predicted by CrossCat. Like the single system model, the feature subset model instead predicts that two systems that split the two category solution are also highly likely, 1357–24–68 and 17–2468–35. The preference for the two-category solution shows the influence of the prior's preference for simpler solutions. The subset attention model shows a stronger preference for the additional three category solutions because of the high probability of subsets of features that span the two systems, requiring explanation of the structure in both. For the data in Fig. 4c, the subset attention model, like CrossCat and the single system model, predicts that a two category solution 1234–5678 is highly likely. This preference is, however, not very strong because only subsets with three or more features will uniquely identify the two category structure.

The subset search model performs differently from CrossCat, despite focusing on selecting groups of features that go together. To see why, consider the data in Fig. 4a. The features Tail and Head, which are in the same system according to CrossCat, are (anti)correlated, having 5 out of 8 features in contradiction. The features Tail and Legs, which are in opposite categories according to CrossCat, also have 5 out of 8 features in contradiction.[7] Any model of feature similarity will necessarily be based on the number of matching and mismatching values, and will therefore have difficulty with cases like this. This suggests that raw feature similarity does not adequately predict category structure. As a consequence, the subset search model(s) predict that categories like 12678–345 and 13467–258 are higher probability than categories like 12–345–678 and 147–258–36. Feature similarity does not provide as strong an inductive bias as joint inference over systems of categories.

5.3. Contrasting the predictions of CrossCat and the repeated recategorization model

The repeated recategorization model also does not perform like CrossCat. To see why, consider the data in Fig. 4. For the data in Fig. 4a the repeated recategorization model predicts that the best solution is placing all objects in the same category. This is because the repeated recategorization model is based on the single system model, and inferring categories based on all features leads to the same result. However, the repeated recategorization model can identify which features are not well captured by this solution, and consequently the repeated recategorization model also tends to favor solutions that focus on single features, which are not well-explained by a single category. Considering the data in Fig. 4c, the repeated recategorization model, unlike the other models, fails to converge robustly on the two category solution. Although it predicts that 1234–5678 is highly likely, it also predicts that a number of other solutions are nearly as likely. This is because the repeated recategorization model sets a cutoff value without contrasting possible solutions for bad features, and therefore splits off subsets of features that, while not perfect, would be best accounted for by a single system of categories. Because repeated recategorization derives feature kinds from object categories, it obtains quite different predictions than CrossCat.

6. Experiment 1: Categorization of artificial stimuli

We developed an unsupervised category learning experiment using a novel multiple sorting method, to explore which systems of categories people tend to form given the different data sets in Fig. 4. For the first data set, CrossCat predicts that people should generate the two orthogonal systems with three categories. The alternative models all predict that a system with a single category is the best solution. For the second data set, all models predict that the two category solution should be highly likely; however, only CrossCat predicts the three category solution 12–345–678. For the third data set, all of the models predict that a two category solution, 1234–5678, should be the most likely.

6.1. Method

6.1.1. Participants

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

6.1.2. Materials

Three sets of artificial stimuli, which we refer to as the 3/3, 3/2, and 2 category learning conditions, were created

[7] Note that features are two-valued—there are, for example, two different kinds of legs—and values are symmetric.

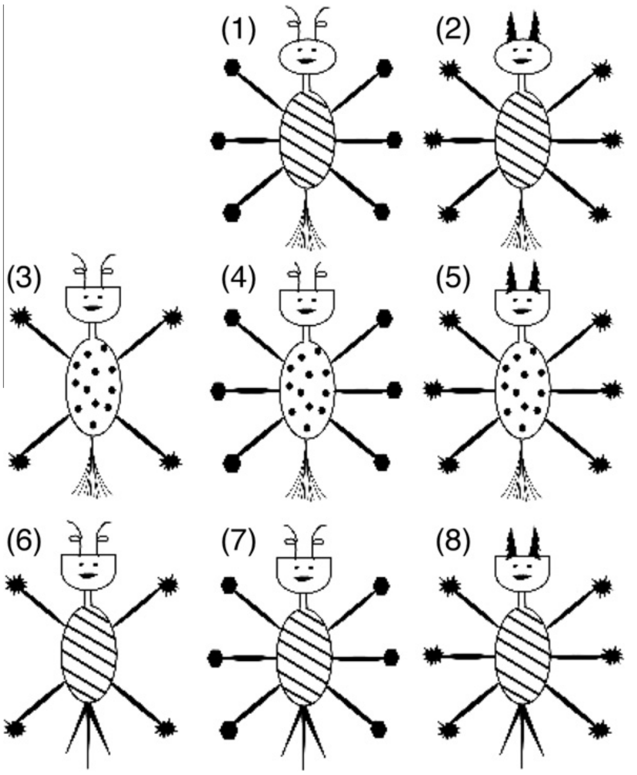


Fig. 6. Stimuli from the 3/3 condition. Note that three categories can be formed by either grouping the objects by row or by column. Stimuli numbers correspond to the numbers used in Figs. 4 and 5.

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 Fig. 6 for an example set of stimuli).

6.1.3. 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. As an example, the experimenter explained that foods can be organized into taxonomic categories (meats, dairy foods, or grains) or situational categories (breakfast foods, lunch foods or dinner foods). The experimenter then explained the 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 (based on size and texture). 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. This was included because people have a well-known tendency to sort based on single features in artificial categorization tasks, and there is good reason to believe that this is not representative of their real-world behavior (e.g. Rosch & Mervis, 1975). This manipulation was intended to help motivate participants to produce more real-world type behavior. Participants were told that they would be given a set of cards and that 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 decided that no additional sorts remained. Each participant sorted each set of stimuli, and the sets were presented in random order.

6.2. Results and discussion

We generated predictions for models by computing the marginal probability for each system of categories. Predictions were derived by enumerating and ranking all possible systems of categories. For CrossCat, this was limited to all solutions with up to two systems. For one system solutions, the marginal probability was simply the probability of that system given the data. For two system solutions, the probability of each system was computed by computing the probability of the whole solution and the probability of choosing that system from the solution (uniformly at random). The marginal probability of a system under CrossCat is the sum of the probabilities of each possible way of obtaining that system. Similarly, for the feature subset models, we summed the probabilities of different systems for each possible subset of features, weighted by the probability of choosing that subset of features. For full details, see Appendix A.

Though people produced a variety of sorts, responses converged on a relatively small subset. For the following analyses we focus only on the top five sorts produced by people in each learning condition. These capture the majority of the sorts produced by people and allow us to focus on the most reliable subset of the data.

Model predictions and human results for the 3/3 condition are plotted in Fig. 7a. The best solution for the 3/3 condition according to CrossCat contains two systems with three categories each, and these are the most likely categories according to the model. The single system model, subset attention model, subset search, and repeated recategorization all predict that a single category with all of the objects is highly likely. Because the systems are orthogonal, a model that is not able to focus on explaining subsets of the data separately, like the single system model, does not recognize the structure of the data. However, simply focusing on subsets is not enough to sharply identify this as the correct structure. The sorting data clearly show that people have no trouble discovering orthogonal systems of categories. To quantify the correspondence between the models and observed data, we correlated the predicted log probability of a system of categories with the observed frequency of the six sets, for each model. CrossCat accurately predicted the frequencies of different categories, $r = 0.91$. All of the other models fail to predict people's behavior (single system, $r = 0.16$; subset attention, $r = 0.16$; subset search, $r = -0.76$; repeated recategorization, $r = -0.88$).

The model predictions and human data from the 3/2 learning condition are shown in Fig. 7b. The MAP solution according to CrossCat contained two categories: one system of three clean categories, and one system of two categories with noise. CrossCat predicts these to be the most

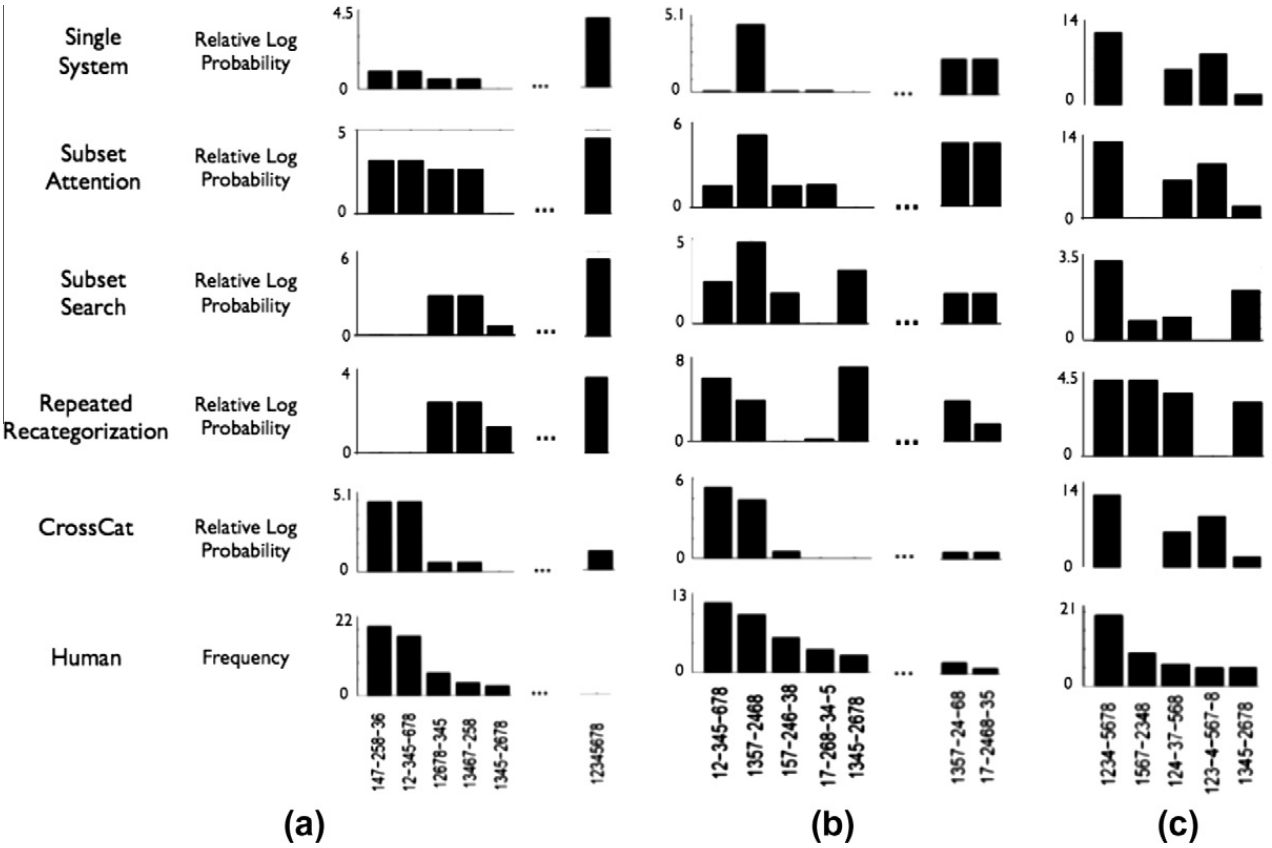


Fig. 7. Predictions and results for: (a) 3/3, (b) 3/2, and (c) 2 category conditions. The *x*-axes show the object clusterings, where different categories are separated by dashes. The *y*-axes represent the relative probability of the categories according to the models, or the frequency with which the solution was observed in people's clusterings. The object clusterings are sorted by frequency in human data. Highly probable solutions predicted by the models but infrequently observed in the human data are shown after the breaks in the *x*-axes.

likely sorts. The single system model again has difficulty with the fact that these categories are largely orthogonal. Because they are not perfectly opposed, the single system and subset attention models identify the simpler of the two systems as the most likely category. The predictions of the repeated recategorization model are similar to those of CrossCat, with the exception that it predicts an additional two category system is more likely than the two systems predicted by CrossCat. The human data shows that people were most likely to produce the two systems predicted by CrossCat, and overall fit was good, $r = 0.70$. Notably, the single system, subset attention, subset search, and repeated recategorization models all incorrectly predict that at least one other system should be more likely than the second best system under CrossCat, and the overall fits to the data suffer as a consequence (single system, $r = -0.06$; subset attention, $r = -0.06$; subset search, $r = -0.14$; repeated recategorization, $r = -0.28$).

In the 2 condition, all models agreed on the best solution, a single system with two categories. This was also the most common sort by people, chosen by two-thirds of participants and appearing more than twice as often as the second most frequent sort. The repeated recategorization model predicts a second solution to be as likely as likely as the first, which is not consistent with the human data. Overall, all models provided reasonably good fits to the human data, though the subset search model performed best (CrossCat, $r = 0.57$; single system, $r = 0.57$; subset attention, $r = 0.57$; subset search, $r = 0.76$; repeated recategorization, $r = 0.55$)

To further investigate the data, we consider, for each pair of objects, what is the probability that they will appear in the same category? Specifically, we compared the probabilities predicted by each model with the observed frequency in the human data. For the 3/3 condition, all models perform well, though the repeated recategorization performs worst (CrossCat, $r = 0.91$; single system, $r = 0.91$; repeated recategorization, $r = 0.78$; subset attention, $r = 0.92$; subset search, $r = 0.90$). Because the only possible structure in the data corresponds to the two orthogonal systems of categories, all models capture the pairwise relationships between objects. In contrast, in the 3/2 condition the attention-based models are unable to predict people's performance (CrossCat, $r = 0.88$; single system, $r = 0.87$; repeated recategorization, $r = 0.87$; subset attention, $r = -0.06$; subset search, $r = -0.11$). Because the two systems are not orthogonal, attending to subsets of features that span the systems can lead to systematically incorrect categorization predictions, while attending to the full set of features tends to lead to categories that are qualitatively similar to the categories in one system or the other. For the 2 condition, only CrossCat and the single system model provide accurate predictions of people's behavior (CrossCat, $r = 0.95$; single system, $r = 0.95$; repeated recategorization, $r = 0.26$; subset model, $r = 0.09$; subset search, $r = 0.07$). Notably, the repeated recategorization model fits poorly because it splits off imperfect features, and forms categories capturing individual variability rather than the overall prototype structure.

Note that the single system model does predict the probability with which pairs of objects are categorized together despite not learning the right categories. To see why, consider the last two categories in Fig. 7b. The category 1357–24–68 is almost the same as the category 1357–2468 in terms of the implied pairwise object relations. The single system model tends to assign high probability to these kinds of nearby wrong answers. Therefore, while its predictions about the probability of categorizing pairs of objects are quite good, its predictions about the categories that people will form are quite poor.

The two sets of analyses show that the attention-based and recategorization models do not capture the patterns observed in the human data. The analyses also show that while the single system model predicts the probability with which pairs of objects are categorized, it does not capture the actual categories that people form. Only CrossCat captures both the the probability of categorization pairs of objects and the actual categories that people form.

7. Real-world data simulations

Although controlled studies of artificial category learning provide some insight into cognition, our ultimate goal is to model real-world behavior. We therefore explored the models' predictions in two real-world domains: animals and musical instruments. This section describes modeling results for these domains, and the next section describes behavioral experiments testing the model predictions.

The starting point for each analysis is again a matrix of objects by features. We used feature-listing data that were collected prior to the work described in this paper. The animals matrix was based on feature-verification responses of a single participant, and the instruments data set was collected by DeDeyne et al. (in press).[8] For both data sets, a subset of objects was chosen to make the number more reasonable for participants to categorize multiple times, resulting in 22 objects for the animals data set and 13 objects for the musical instruments data set.

The raw data for the animals data set are shown in Fig. 1. The data represent the knowledge that, for example, ostriches and dolphins have lungs, while jellyfish and grasshoppers do not. The MAP solution according to the single system model is shown in Fig. 2. The solution contains a system of taxonomic categories, including mammals, invertebrates, birds, and reptiles/amphibians. These categories capture a subset of the features quite well, including 'has lungs' and 'squawks'. However, as can be seen on the right side of the figure, there are many features which are poorly explained by these categories. Some of these features seem likely to contain interesting structure based on ecological relations, including 'hunts', 'is a carnivore', and 'eats leaves'.

Fig. 3 shows shows that the MAP solution according to CrossCat contains three systems. The first system is identical to the categories predicted by the single system model (Fig. 3a). In addition, CrossCat learns two more systems that apply to features that are poorly captured by the single system model. Fig. 3b shows a roughly ecological system of categories, including categories of land predators, aquatic predators, aerial predators, and prey. These categories indicate, for example, that bats and eagles and seals and penguins have many features in common. Fig. 3c shows a third system which contains features that are noisy (e.g. 'is large') or irrelevant (e.g. 'is a rodent').[9] CrossCat has learned that the features in this system have little discriminative ability, and therefore should not be attended to.

The raw data for the musical instruments data set are shown in Fig. 8, and and indicate, for example, that banjos have strings but that drums do not. The MAP solution according to the single system model is shown in Fig. 9. The solution includes five roughly taxonomic categories including wind instruments, two categories of string instruments, and two categories of percussion instruments. This solution attempts to accommodate two different kinds of information: information about taxonomic categories, and information about the social uses of the instruments.

The best solution for the musical instruments data according to CrossCat is shown in Fig. 10. The best solution includes two systems. The first system corresponds to a taxonomic categorization of musical instruments: wind, percussion, and string instruments. The second system approximates categories based on social concerns. Although not perfect, they reflect instruments used in more classical settings, such as cellos, harps, bassoons, and clarinets, as well as those that are found in popular (rock) music settings, drums and bass. The remaining splits reflect intermediate groups such as violins and flutes, which can be found together in folk music and classical music.

The attention and repeated recategorization models do not lend themselves to presentations of single solutions, because any possible solution depends on random processes. However, these models do lead to quantifiable predictions that will be contrasted with those of CrossCat and the single system model. We will test these predictions in Experiment 2.

8. Experiment 2: Categorization of real-world domains

8.1. Method

8.1.1. Participants

25 University of Louisville undergraduates participated in exchange for extra credit.

[8] Four people participated in the feature verification task in their study. We thresholded the data at two: if two people said that a feature was true, then we considered it to be true, otherwise it was considered false.

[9] The uninformative system does split off frog to capture the amphibian characteristics that do not apply to reptiles.

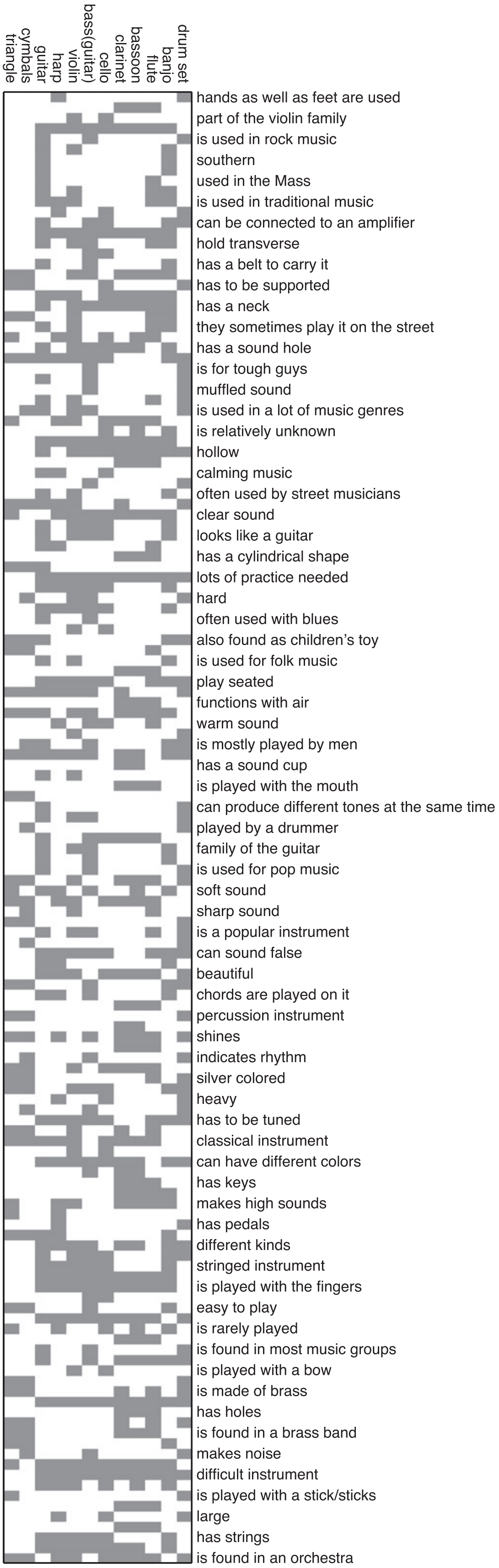


Fig. 8. Raw data provided to the models for the instruments condition. These data are derived from human judgments collected in DeDeyne et al. (in press). Due to space constraints, only every other feature is labeled.

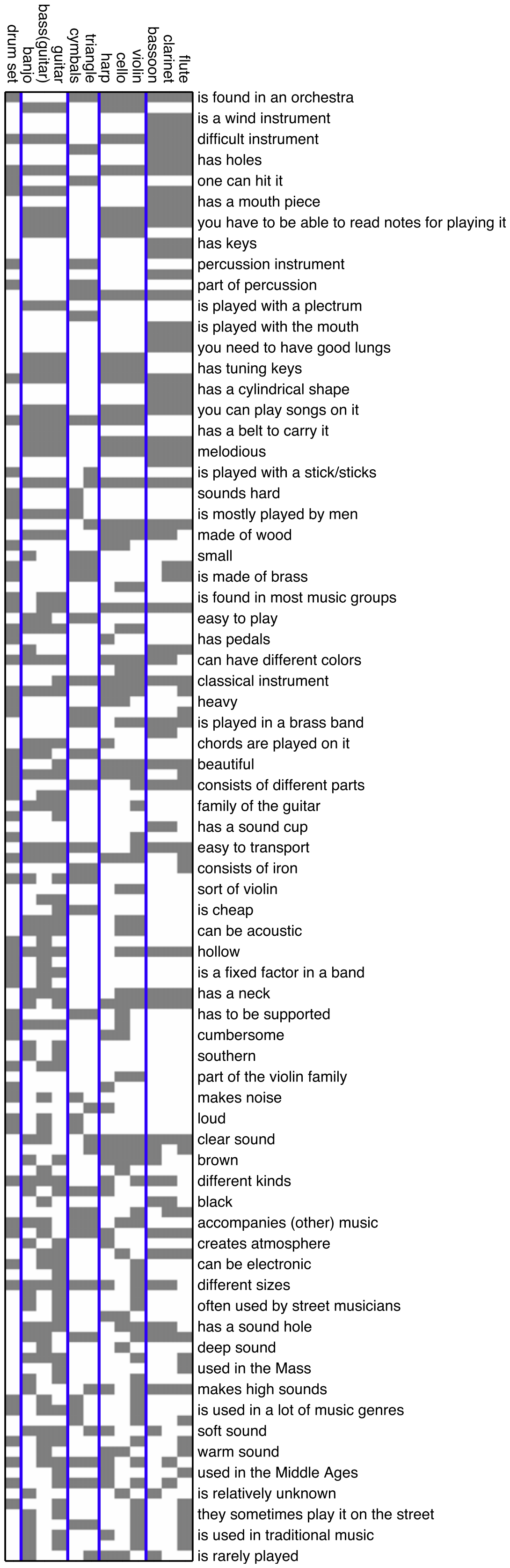


Fig. 9. The best solution to the instruments data according to the single system model. The model finds a roughly taxonomic system of the instruments into wind, string, and percussion categories. This solution divides the string and percussion instruments each into two groups. Features are sorted from best explained at the top, to most poorly explained at the bottom.

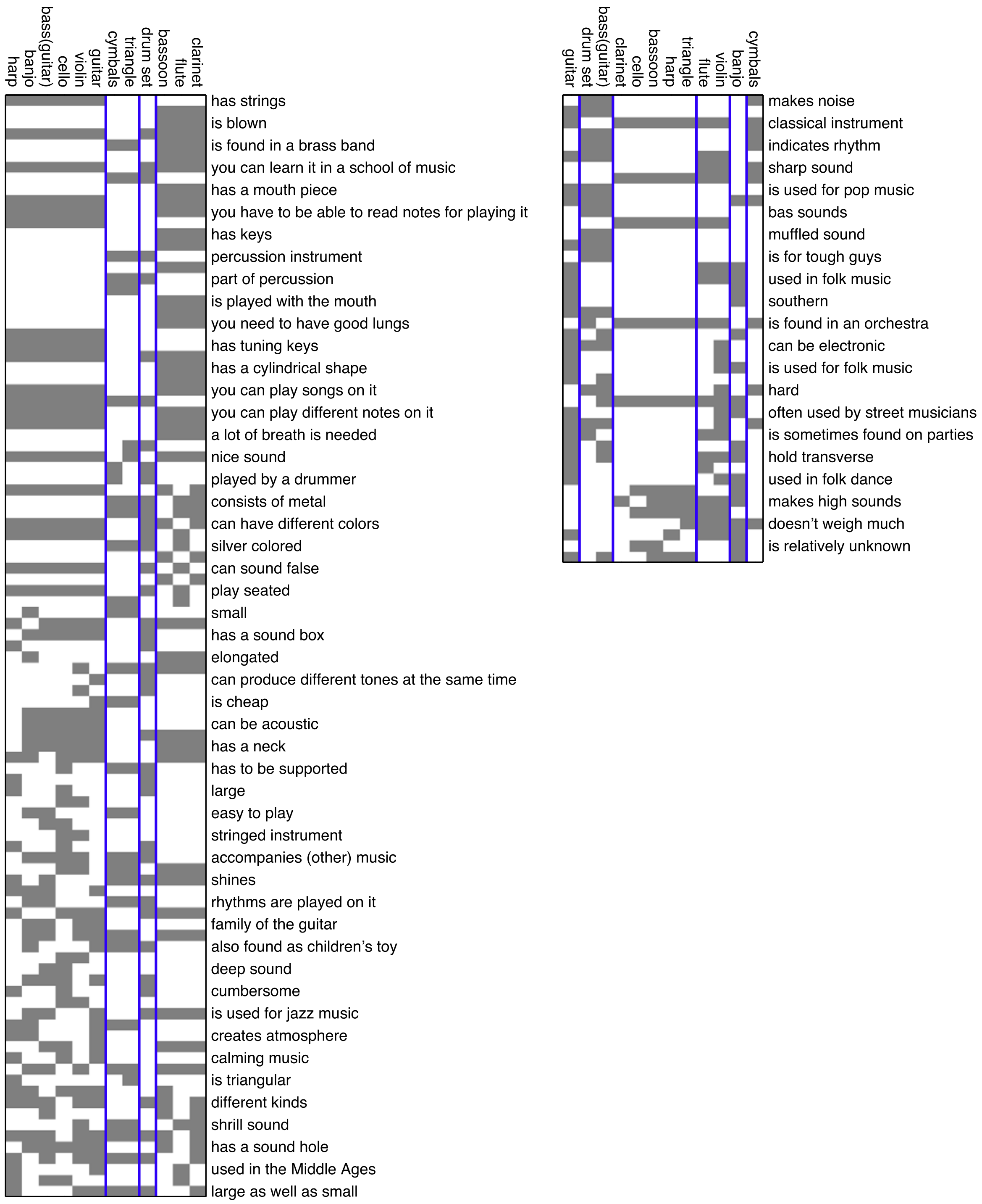


Fig. 10. The best solution found by CrossCat for the instruments data, containing two systems of: (a) taxonomic, and (b) social use categories. The system of taxonomic categories includes wind, percussion, and string instruments. The system based on social use includes categories of classical instruments such as cello, harp, bassoon, and clarinet, and the rhythm section (drums and bass) of rock bands are grouped.

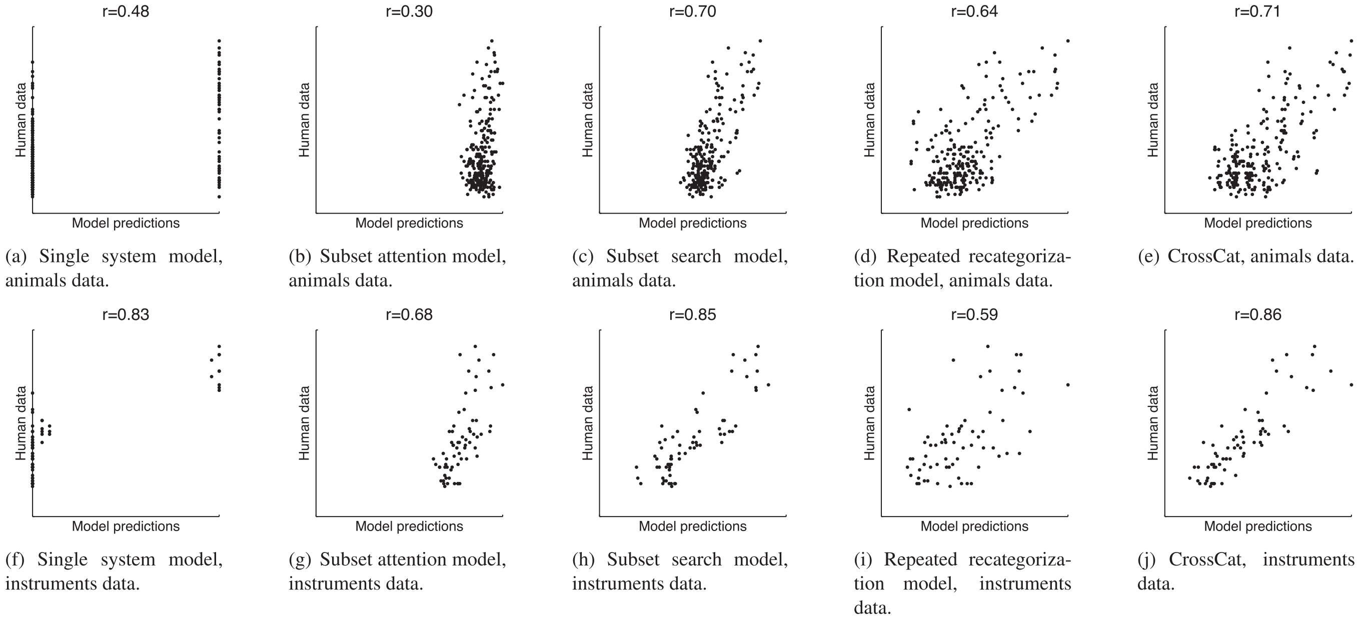
8.1.2. Materials

The experiment included two sets of items: an animals set and an instruments set. The entities used in both sets are listed in Figs. 1 and 8.

8.1.3. Procedure

The experiment was conducted on computers using MATLAB. Participants sat at a computer and were told that we were interested in how people categorize objects, specifically the different ways that people categorize the same set of objects. We then gave them the example of categorizing foods in groups based on taxonomic categories such as meats, dairy, and grains or in situational categories such as breakfast, lunch, and dinner foods.

People participated in the animals and the instruments categorization tasks in counterbalanced order. A set of words (either animals or instruments) appeared on the screen in random locations. Participants could group the



(a) Single system model, animals data.

(b) Subset attention model, animals data.

(c) Subset search model, animals data.

(d) Repeated recategorization model, animals data.

(e) CrossCat, animals data.

(f) Single system model, instruments data.

(g) Subset attention model, instruments data.

(h) Subset search model, instruments data.

(i) Repeated recategorization model, instruments data.

(j) CrossCat, instruments data.

Fig. 11. Scatterplots showing the predicted (*x*-axis) and observed (*y*-axis) probabilities of objects being in the same cluster for the single system, subset attention, subset search, repeated recategorization, and CrossCat models on the animals and instruments data. The plots show that while the models vary in how accurately they predict the human data, in both conditions CrossCat performed as well as, or better than the alternative models.

words by clicking and dragging. They could then indicate categories by drawing rectangles around groups of words. When they completed categorizing the words, they were asked whether there were other ways of categorizing that captured new and different information about the objects. If participants indicated that there were, a new screen appeared with the same words in random positions. Their previous categories remained visible for reference, but could no longer be modified. This process repeated as long as participants indicated there were other important sets of categories. When participants indicated that there were no more important ways of categorizing the objects, the experiment was completed. Upon completion, we debriefed and thanked participants.

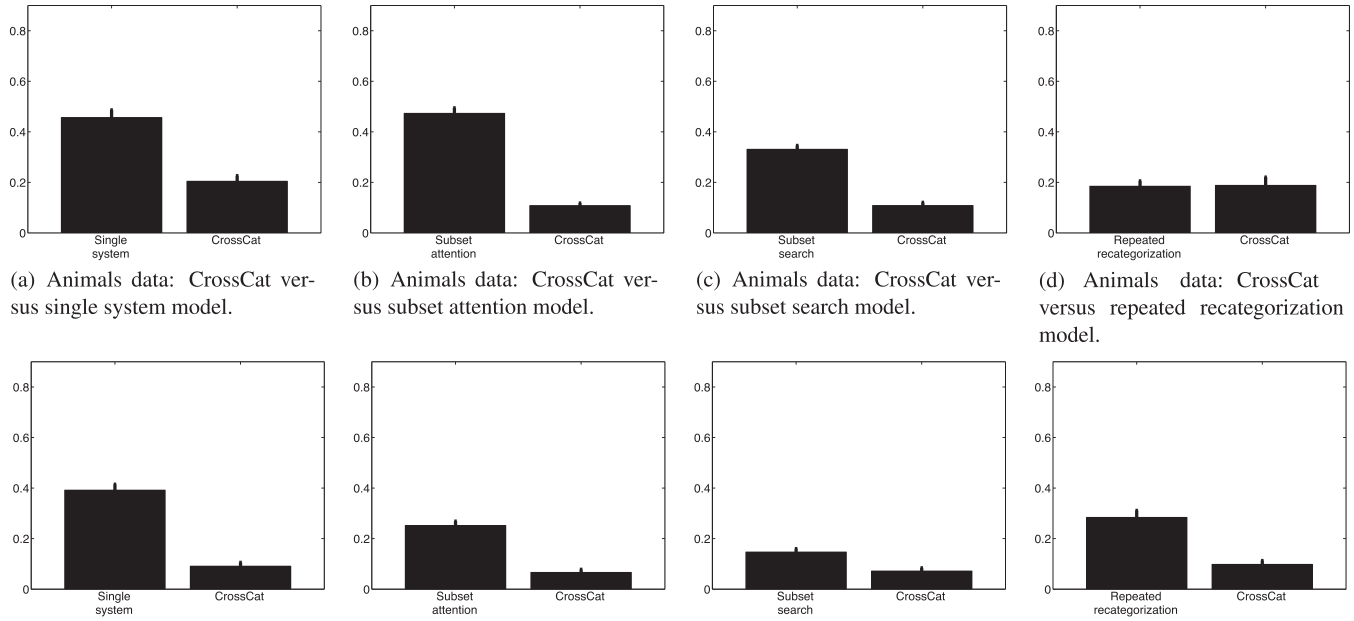
8.2. Results and discussion

We compare the categories people produced to the model predictions, addressing a series of questions. First, we will investigate the predicted and observed probabilities of categorizing pairs of objects together, allowing us to characterize which model best fits people's behavior and which pairs of items show the greatest differences between models. Second, we will compare the accuracy of CrossCat to the alternative models on the pairs of items for which their predictions differ the most. Third, we will compare the observed distribution of numbers of categories per system to the distribution predicted by the models.

To test how well the models predict the human data, we investigated which pairs of objects are more or less likely to be categorized together, collapsing across all of the categories people formed. This will allow us to quantitatively compare the models' predictions and human data for each pair of objects, to identify and to test specific items where the model predictions diverge. All model predictions were generated using simulations. Samples from the posterior distributions were obtained via MCMC (see Appendix A for details). From these samples, we estimated the marginal probability that two objects belong together by computing the proportion of times the pair was categorized together. For the single system model, this marginal probability is the proportion that cases where each pair of objects was assigned to the same category. For CrossCat, if a sample had multiple systems of categories, each system of categories was treated separately. Then the proportion of co-occurrences were computed as for the single system model. For the subset attention and the subset search models, subsets were drawn and then samples were drawn given that subset of features. This was repeated multiple times for different samples of features, and the predictions for each subset were weighted by the probability of sampling that feature kind. Full details are provided in Appendix A.

The scatterplots in Fig. 11 compare the model predictions and human data for each data set. For the animals data, the single system model makes sharp predictions, predicting for all pairs that they will almost deterministically be categorized either together or separately.[10] The model shows a reasonable correlation, $r = 0.48$, but it is clearly not accounting in detail for the variability in human behavior, as indicated by the points clustering on the left and right sides of the plot. The subset attention model performs considerably worse, $r = 0.30$. Attention to subsets of features leads to considerable blurring of the overall

[10] To ensure that this was not due to a failure of the MCMC to mix properly, we implemented an Importance sampler that sampled from a relaxed version of the posterior distribution, and weighted samples appropriately (see Appendix A for details). The predictions of both approaches converged, and the predictions from the importance sampler are shown for the single system model for both data sets.



(a) Animals data: CrossCat versus single system model.

(b) Animals data: CrossCat versus subset attention model.

(c) Animals data: CrossCat versus subset search model.

(d) Animals data: CrossCat versus repeated recategorization model.

(e) Musical instruments data: CrossCat versus single system model.

(f) Musical instruments data: CrossCat versus subset attention model.

(g) Musical instruments data: CrossCat versus subset search model.

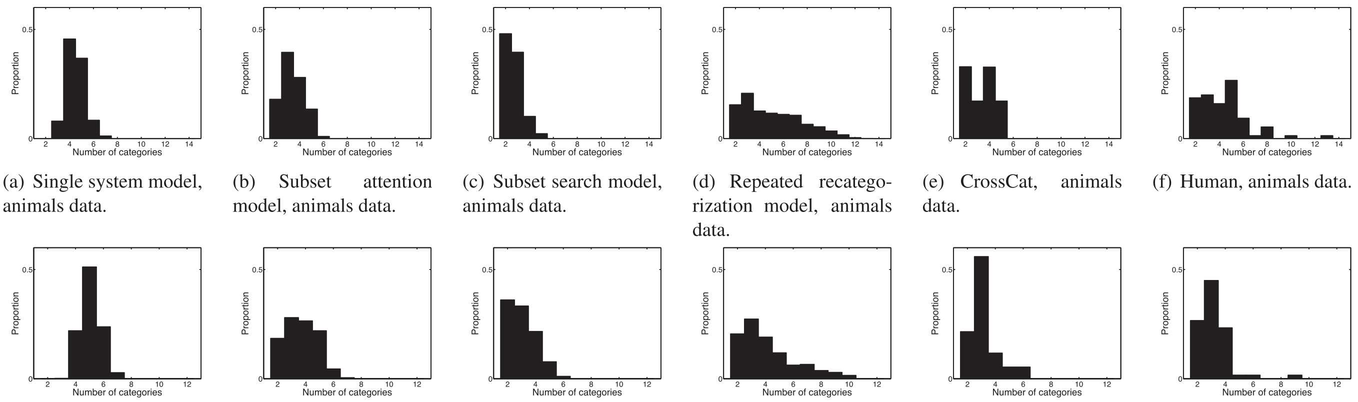
(h) Musical instruments data: CrossCat versus repeated recategorization model.

Fig. 12. Bar charts showing the average absolute deviation between the model predictions and human data for items where CrossCat differed from each other model. Panels (a)–(d) contrast CrossCat with each of the other models on the animals data. CrossCat's predictions are significantly more accurate than the single system model, the subset attention model, and the subset search model and comparable to the repeated recategorization model for the animals data. Panels (e)–(h) contrast CrossCat with each of the other models on the musical instruments data. CrossCat's predictions are again significantly more accurate than all four models.

structure, which results in much poorer fits. The subset search model fits human data well, $r = 0.70$, as does the repeated recategorization model, $r = 0.64$. CrossCat fits the human data most closely, 0.71. These results suggest that the subset search, repeated recategorization, and CrossCat capture the gross ordering of the data.

For the musical instruments data, the models show a different pattern of fits. The single system model, $r = 0.83$, and CrossCat, $r = 0.86$, both provide very close fits to the human data. Also, the subset search model provides a strong fit to the data, $r = 0.85$. In contrast, the subset attention model, $r = 0.68$ and repeated recategorization model, $r = 0.59$, provide a less accurate fit to the human data. The subset attention model provides the weakest fit because it is unable to discern good features from bad features, and incorporation of subsets of features that are not representative of the whole lead to incorrect predictions. Repeated recategorization also performs poorly on the musical instruments. Whereas for the animals data, repeated categorization captured a successively more detailed picture of the data, for the instruments data, successive categories focus on idiosyncratic details. The repeated recategorization model can only identify when features are not well-explained; however, this does not necessarily imply that there is important structure that has yet to be identified. These results suggest that both CrossCat and the subset search model predict the ordering of the pairs of objects; however, inspection of the subset search scatterplots suggests that, at least for the animals data, subset search does not accurately predict the probabilities with which people group pairs of objects together. We address this issue in the next set of analyses.

To provide more insight into differences between CrossCat and the other models, we identified the 25 pairs of objects for which CrossCat's predictions deviated most sharply from each other model, for each data set. Then, for each set of items, we computed the average absolute deviation between the model predictions and human data, and the results are presented in Fig. 12. The sharpest differences between CrossCat and the single system model focused on animals that share ecological similarities, such as penguins and seals, and bats and seagulls, and instruments that are used in similar contexts, such as clarinet and harp, and cello and bassoon. Fig. 12a and d show that CrossCat provides significantly closer fits to the human data for these cases, $t(24) = 4.92$, $p < 0.001$ and $t(24) = 8.95$, $p < 0.001$. The sharpest differences between CrossCat and the subset attention model focus on items that are very different from each other. For the animals data, these include pairs such as octopus and sheep, and dolphin and grasshopper. For the musical instruments, these cases include flute and drums, and triangle and guitar. Fig. 12b and e show that CrossCat was significantly more accurate for the animals items $t(24) = 23.09$, $p < 0.001$ and for the instruments items, $t(24) = 7.07$, $p < 0.001$. CrossCat also fared well when contrasted with the subset search model. The main area of disagreement on the animals was items that do not belong together, such as ostrich and frogs, and penguins and sheep, and CrossCat performed significantly better than the subset search



(a) Single system model, animals data.

(b) Subset attention model, animals data.

(c) Subset search model, animals data.

(d) Repeated recategorization model, animals data.

(e) CrossCat, animals data.

(f) Human, animals data.

(g) Single system model, instruments data.

(h) Subset attention model, instruments data.

(i) Subset search model, instruments data.

(j) Repeated recategorization model, instruments data.

(k) CrossCat, instruments data.

(l) Human, instruments data.

Fig. 13. Histograms showing the proportion of systems containing different numbers of categories for the single system, subset attention, subset search, repeated recategorization, and CrossCat and the human results for the animals and instruments domains. For the animals domain, the human data show a relatively broad distribution spread between two and six categories, a pattern that is best approximated by CrossCat. For the instruments data, the human data shows a peak at three categories, a pattern best predicted by CrossCat.

model on these items, $t(24) = 25.64$, $p < 0.001$. For the instruments, the main differences were among various string instruments such as banjo and cello and guitar and harp. On these items, CrossCat was significantly more accurate than the subset search model, $t(24) = 3.40$, $p < 0.005$. The differences between CrossCat and the repeated recategorization model on the animals data were negligible, $t(24) = -0.08$, *ns*. However, for the instruments, the largest differences were on both taxonomically related pairs such as violin and harp and drums and cymbals and socially related items such as harp and basson. CrossCat performed significantly better than the repeated recategorization model on these, $t(24) = 7.27$, $p < 0.001$. These results suggest that although the subset attention model provided good correlations, CrossCat more accurately predicted which pairs of objects are more likely to belong in the same category.

To get a better sense of how well the models predict the structure of people's systems of categories, we compared the observed number of categories per system to the predictions of all of the models. Fig. 13 shows the distributions predicted by the models, and the frequencies observed for both domains. To quantify the strength of the relationships between the models and the human data, we ran correlations between observed and predicted proportions. The single system model's predictions tend to overestimate the number of categories and show relatively poor fits to the human data, $r = 0.68$ and $r = -0.10$ for the animals and instruments data. The subset attention model shows moderate fits to the human data, $r = 0.79$ and 0.82 but fails to capture the flatness of the distribution in the animals data and the sharp peak in the instruments data. The subset search model performs similarly on average, $r = 0.62$ and $r = 0.94$, trading off a better fit to the instruments against a worse fit to the animals data. The repeated recategorization model shows good quantitive fits, 0.81 and 0.94, but fails to capture the qualitative differences in people's performance on the two data sets. CrossCat provides the closest fit to the human data, $r = 0.81$ and $r = 0.95$, capturing both the relative flatness of the animals distribution and the peak in the instruments distribution.

Taken together, these results suggest that people's cross-categorization abilities are based on joint inference about multiple systems of categories and the features that they explain. The single system and subset attention model provide relatively poor fits to human data. The subset search model provides reasonable correlations to the human data, but has a strong preference for small numbers of categories, which leads the model to overestimate the probabilities of pairs of objects being categorized together. The repeated recategorization model failed to account for human data on the musical instruments, and failed to capture the qualitative aspects of people's sorting behavior. Only CrossCat provides an accurate fit to people's judgments across all three measures, a consequence of joint inference over multiple systems of categories.

9. General discussion

Although there is broad consensus that people do cross-categorize, there has not been consensus on how best to explain this ability. Our approach to addressing this question has been based on identifying common intuitive explanations of cross-categorization, and formalizing and contrasting models based on these intuitions. The accounts we have considered form a continuum from simply learning a single system of categorization, through two approaches which treat inferences about categories and features separately, to joint inference about systems of categories and the features they explain. In formalizing these explanations, we have focused on a single family of models, based on a version of Anderson's rational model (1991), systematically modifying aspects of the basic model to implement the different explanations. This approach has allowed us to minimize differences between the models, and has provided a means by which we may investigate which forms support human-like cross-categorization.

We have contrasted each of the models in two sets of experiments: the first focusing on artificial category learning, and the second on categorization in real-world domains. The artificial category learning experiment used stimuli that were designed to either support or not support multiple systems of categories. The real-world data were chosen to represent very different domains of knowledge: biological kinds and artifacts. In each of the two sets of experiments, the data induced sharp differences among all of the models. For the artificial categories, each of the alternative approaches showed characteristic divergences from the human data. Unlike people, the single system model was unable to recognize orthogonal systems of categories. The subset attention model, while able to identify orthogonal systems by only attending to a few features, incorrectly predicted a number of noisy and unnecessarily complex systems as a result of attending to subsets of features. Repeated recategorization failed to identify orthogonal structure, and was unable to account for situations when there was only a single system of categories. The subset search model failed to capture behavior on the data sets with multiple systems, because feature similarity does not necessarily map onto good category structure. Only CrossCat, the approach based on joint inference over multiple systems of categories, was able to account for people's behavior across the three data sets.

The real-world data also supported the importance of joint inference over multiple systems of categories. The single system approach provided a reasonable correlation with the human data for the musical instruments, but the fit was rather poor for the animals. The subset attention model showed poor fits to human behavior because random subsets of features do not necessarily capture representative structure. The subset search model provided strong correlations, but more detailed analyses showed that the model produced systematically fewer systems of categories and consequently overestimated the probability with which pairs of objects belong together. The repeated recategorization model provided good fits for the animals data, but it was systematically distracted by idiosyncratic details and therefore failed to capture people's behavior for the instruments data. CrossCat provided close fits to the human data in both cases, providing comparable or superior performance across all analyses. These results suggest that simple attentional mechanisms cannot explain human cross-categorization. Rather, a model that jointly infers multiple systems of categories provides the best explanation of people's behavior.

Beyond these specific findings, our work advances the understanding of category learning in several ways, which we will elaborate in the following. First, we discuss the intuitions for two assumptions, mutual exclusivity of categories and systems, that underlie our implementation of cross-cutting categorization. Second, we will discuss relations to, and implications for, previous approaches to category learning and the notion of selective attention. We will conclude by discussing future directions.

9.1. Mutual exclusive versus overlapping categories

Cross-categorization, and our model in particular, combine two ideas that may seem incompatible at first. The vast majority of categorization models learn a single system of non-overlapping categories, 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 ideas 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.

Cross-categorization resolves the apparent contradiction between these perspectives. 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 that categories from different systems may overlap. Our model therefore inherits much of the flexibility that overlapping categories provide, without losing the idea that the categories within any given system are often disjoint.

In their studies of food categories, Ross and Murphy (1999) noted a difference in the apparent structure of taxonomic and situation-based categories. Whereas foods tended to be rated as good members of one taxonomic category, some foods were considered reasonably good members of multiple situation-based categories. Based on these results, one might ask whether our assumption of mutually exclusive categories is too strong.

There are several reasons to believe that the mutual exclusivity assumption may not be too strong. Close inspection of Ross and Murphy's (1999) situation-based categories shows a rather heterogeneous group including 'breakfast foods', 'snacks', 'healthy foods', and 'junk foods'. It seems reasonable that membership in this assortment of categories would not be exclusive, as they do not appear to form a coherent system. It is, however, possible that even if one selected a more targeted subset of these categories such as breakfast, lunch, and dinner foods, one may find that membership is more mixed than for taxonomic categories. Foods may not be breakfast or lunch foods, but potentially both. CrossCat is able to handle this possibility; categories can be mutually exclusive (within a system of categories) or overlapping (across systems). Situation-based categories may not form a single coherent system, and if not, CrossCat should find that different systems account for different situation-based categories.

In formalizing our model, we assumed that features are not shared across systems. This assumption provides a simple starting point for exploring cross-categorization, but will ultimately need to be relaxed. Consider, for example, the feature "has blubber". This feature is possessed by a subset of animals defined by the intersection of mammals and aquatic creatures, and should probably be linked with both the taxonomic system of categories and the ecological system in Fig. 3. Another instance in the domain of animals is the feature "is large", which seems potentially related to taxonomic categories (mammals tend to be large relative to the others) and ecological categories (aerial creatures tend to be not large). These are two cases where our assumption of mutual exclusivity seems incorrect, and there are likely more. One way to develop an extension of CrossCat that handles cases like these is to replace the current prior on feature assignments (Eq. (3) in the Appendix) with a prior that allows features to be assigned to multiple systems of categories. As described in the appendix, the current prior is based on a procedure known as the Chinese Restaurant process, and an extension of this procedure called the Indian Buffet process (Griffiths & Ghahramani, 2006) allows features to be assigned to multiple systems of categories. Once features can be assigned to multiple systems of categories, some assumption must be made about how these multiple systems combine to predict whether the feature applies to a given object. One simple approach is a noisy-AND model—for example, an animal is expected to have blubber only if its taxonomic category AND its ecological category both indicate that blubber is a possibility. Other assumptions are possible, however, and future studies can explore which set of assumptions best captures human abilities.

Allowing features to be shared across multiple systems of categories is a natural direction for future work, but pursuing this direction is likely to leave the primary conclusions of the present paper unchanged. We focused on a comparison between CrossCat, an approach that jointly infers systems of categories and assignments of features to these categories, and two alternatives which suggest that cross-categorization emerges as a consequence of selective attention or repeated recategorization. Our data suggest that joint inference provides the closest account of human abilities, and refinements of the modeling approach can only increase the strength of this conclusion.

9.2. Implications for models of category learning

There have been many models of categorization and category learning proposed in the literature. These focus on the problem of learning a single system of categories, and therefore cannot account for the results showing that people learn and use cross-cutting systems of categories. In the following, we consider relations between exemplar and prototype approaches and simplicity and attention-based approaches and our own, addressing the possibility that extended versions of these approaches can handle cross-cutting systems of categories.

9.2.1. Exemplar and prototype approaches

One benefit of developing an approach that includes the single system model (a version of Anderson's rational model) as a special case is that there are strong ties between this model and both prototype and exemplar models (see Nosofsky, 1991). Nosofsky showed that, under certain conditions, the Rational model (Anderson, 1991) can be viewed as a generalization of an exemplar approach (Medin & Schaffer, 1978, specifically the Context model).[11] Depending on the value of a single parameter (α in our discussion of the single system model), the Rational model interpolates between storing each example individually, an exemplar-based approach, and grouping all of the examples into a single category, a prototype-based approach. Between these extremes, the model learns a single system with multiple categories, each represented by a single prototype.

CrossCat extends this approach by allowing multiple systems of categories. In our formalization, we have assumed that the α value is the same for categories and systems, since there seemed to be no *a priori* grounds for different settings. For the sake of comparison, we consider setting these parameters separately, and refer to them as α_c and α_s. In the case where α_s, the parameter for systems, ensures that all features are of the same kind (i.e. is set to 0), then CrossCat will only learn a single system of categories, and the relationships between traditional exemplar and prototype approaches are recovered as discussed above. As α_s goes to infinity and α_c goes to zero, we would learn multiple systems of identical categories. Each system would have a single category including all of the objects, and therefore would learn the average value for each feature. For independent features, the result is similar to the case where all features were together. As both α_c and α_s go to infinity, we are memorizing individual object-feature pairings.

This highlights the core difference between our approach and previous prototype and examplar-based models: our approach learns about the structure of both objects *and* features. Exemplar and prototype-based approaches represent two extremes of our model. Our suggestion is that this continuum is not enough (see also Murphy, 1993). It is also necessary to consider that different constellations of features be summarized differently, that different categories and representations are necessary to capture different ways of thinking about objects.

9.2.2. Simplicity-based approaches

Our approach is also related to simplicity-based approaches (Pothos & Chater, 2002; Pothos & Close, 2008). The simplicity-based approach attempts to learn the system of categories that (almost) allows within-category similarity to be greater than between category similarity for all pairs of objects. The model directly trades-off simpler category structure against violations of the constraint that within-category similarity be greater than between category similarity.

The spirit of trading off complexity of structure versus fit to the data is a unifying principle of our approaches. Indeed, CrossCat can account for many of the results presented by Pothos and Close (2008). In one experiment, the researchers presented participants with a category learning task in which participants observed artificial stimuli that varied on two continuous dimensions. Two data sets were created: one for which the best system of categories integrated information based on both feature dimensions, and the other for which the best categories treated the features independently (Figs. 7 and 8 in Pothos & Close, 2008). They found that in the case of dependent dimensions people's categories were closer to the two-dimensional categories than the unidimensional categories. In

[11] Note that this equivalence no longer holds for the unsupervised Generalized Context model (Pothos & Bailey, 2009).

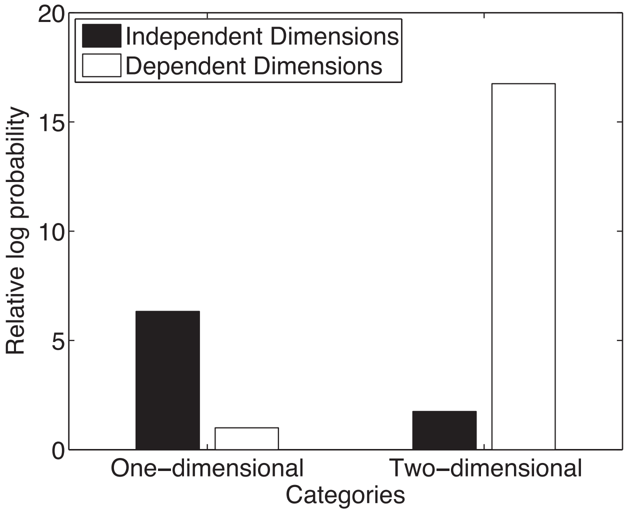


Fig. 14. CrossCat predictions on data from Pothos and Close (2008). For data with dependent (correlated) dimensions, CrossCat predicts two-dimensional categories. For data with independent dimensions, CrossCat predicts one-dimensional categories.

the independent dimensions case, people's categories were closer to the predicted unidimensional categories than the two-dimensional categories. Adapting our approach to handle continuous, potentially correlated data, CrossCat makes similar predictions.[12] Fig. 14 shows that CrossCat prefers categories based on a single dimension for the data that are independent, and a category based on both dimensions for the dependent data. This underscores the similarities of the approaches.

However, there are important differences between the simplicity-based approach as applied to categorization based on different feature dimensions in Pothos and Close (2008) and CrossCat. Whereas Pothos and Close's approach does admit multiple categories for a single data set, it is based on identifying categories that account for a subset of the features, and ignore the rest. In this sense, their model is similar to the subset attention model. Consequently, the simplicity approach described in (see Pothos & Chater, 2002) for a full description]pothosClose08 cannot account for our results.[13] In particular, on the artificial data, the simplicity account makes patterns of predictions (and errors) that are qualitatively similar to those of the subset attention model, providing a reasonable explanation of the 3/3 and and 2 category systems, but failing to predict the pattern of results observed in our 3/2 condition. These results suggest that Pothos and Close's approach does not capture the qualitative effects that we have observed here. We believe that this is because the model represents systems of categories independently, a representational commitment comparable to those in the subset attention models. It seems possible that an alternative formulation, based on joint inference about systems of categories, may be able to account for our data, and we believe that this would provide further insight into the representational basis of cross-categorization.

9.3. Cross-categorization and selective attention

Our results suggest that cross-categorization provides a more accurate account of human behavior than selective attention mechanisms; however, we do not consider this a point of discontinuity with previous research. Categorization with selective attention can be viewed as a simple kind of cross-categorization, where one system of categories is learned for one group of features (those that are attended to), and all objects are lumped into the same category for another set of features (those that are not attended to). Cross-categorization can be viewed as a generalization of this basic idea. Rather than adopting this distinction between features based on relevance or irrelevance, Cross-categorization proposes that all features are relevant, but in different contexts. In any one context, this is similar to selective attention. However, cross-categorization is also different from selective attention. Cross-categorization takes the basic goal of categorization – to discover the structure of the world – and generalizes it to allow for the possibility that there may be more than one explanation for structure in a domain.

Cross-categorization suggests that there may be more than one explanation to the question of why objects are the way they are. This represents a hypothesis about the world, that multiple explanations exist, and leads directly to the question of which categories best explain a given feature. This competition between systems of categories to explain features leads to a different view of selective attention. Rather than explaining features as relevant or irrelevant in a given context, cross-categorization attempts to explain *why* a feature is irrelevant; what is the alternative explanation? When there is a system of categories that better explains that feature and sufficiently many other features, we infer that it is irrelevant in this context because there is some other causal process that explains its structure. In this sense, where selective attention suggests that the problem of carving nature at its joints is a matter of attenuating noise, cross-categorization suggests that nature is composed of many different sets of joints, reflecting a complex of causal histories, and the problem is to infer the joints that separate these histories.

9.4. Conclusion

People think of objects as belonging to many different categories, and the ability to think about objects in different ways plays a critical role in how we categorize, reason, plan, and solve problems. This cross-categorization pre-

[12] To generate predictions, we implemented a generalized version of CrossCat that allowed multiple observations per cell (binomial data), and considered the Pothos and Close data to be on an interval-level scale. The modified version of CrossCat is similar in detail to the version discussed here, with the exception that the probability of the data (see Eq. (4)) must be modified to include binomial coefficients that depend on the number of observations per cell.

[13] We implemented the simplicity model described in Pothos and Chater (2002), and scored the systems of categories shown in Fig. 7 under their model. The simplicity model cannot be straightforwardly applied to data sets of the size of the real-world data we considered because the number of subsets of 106 features (as in the case of the animals data) is more than a nonillion (10^{30}) and running the simplicity model would require searching over all subsets of the feature space. This practical issue should not, however, be interpreted as an argument against the model.

sents a chicken-or-egg problem. How can people infer categories without knowing the relevant subset of features in advance? We have presented and tested four possible explanations for people's ability to cross-categorize: accounts based on simply learning a single system of categories, using feature-based attentional mechanisms to guide multiple category formation, using object-derived categories to identify feature kinds, and joint inference about multiple systems of categories and the features they explain. Our results suggest that people's behavior is best explained by approaches that jointly infer multiple systems of categories.

Though it is important to understand basic aspects of thought in simple settings, in order to understand the flexibility of human reasoning, we must attempt to model learning of the more robust and complex knowledge that underlies flexible learning and reasoning. We focused on developing an account of how cross-cutting categories could be inferred from bottom-up information. Top-down information, such as intuitive theories and verbal statements, also plays a key role in cross-categorization. It is important to continue to develop these approaches toward the goal of modeling human abilities in more real-world situations, and we believe that models that can learn multiple systems of categories to account for different aspects of knowledge are a step in the right direction.

Appendix A. Formal details for the single system, subset attention, subset search, and CrossCat models

A.1. Single system model

Both the single system model and CrossCat can be viewed as probabilistic generative models for possible categories and observed data. The single system model is composed of two parts.

The first is a prior on categories in systems, $p(w|\alpha)$, which has a single parameter α that represents the strength of our preference for small categories. This is implemented using a Chinese Restaurant Process (CRP) prior. The CRP is a sequential generative process which assumes there are a potentially unbounded number of categories, only some of which we have observed so far. Imagine a new object is encountered, and we need to decide which category it belongs in. The CRP states that the probability of assigning the object to an existing category is proportional to the number of objects in that category, and the probability of assigning that object to a new category depends on a single parameter, α.[14] More formally, the probability that object o_i is assigned to category c given the assignments of objects $w_{o_1 \ldots o_{i-1}}$ is

$$P(w_i = c|w_1, \cdots, w_{i-1}) = \begin{cases} \frac{n_c}{i-1+\alpha} & \text{if } w_c > 0 \\ \frac{\alpha}{i-1+\alpha} & c \text{ is a new category} \end{cases} \tag{3}$$

where n_c is the number of objects in category c. The key aspects of this prior are that it allows for anywhere between 1 and O categories, where O is the number of objects so far, and the number of categories can grow naturally as new data are observed.

The second part is a model for generating observations of a feature for all objects within a category, $p(D_{o\in c,f}|\delta, w)$, where δ specifies the strength of our preference that all objects in a category tend to have the same values for features. As δ goes from 1 down to 0, the preference gets stronger. Formally, the probability of the observed values of feature f for objects in a category c is

$$P(D_{o\in c,f}|w, \delta) = \frac{Beta(n_{c,f} + \delta, \bar{n}_{c,f} + \delta)}{Beta(\delta, \delta)} \tag{4}$$

where *Beta* indicates the beta function and $n_{c,f}$ is the number of cases for which the feature is observed.

The full generative model is then,

$$p(D, w|\alpha, \delta) = p(w|\alpha) \prod_{c=1}^{max(w)} p(D_{o\in c,f}|\delta, w). \tag{5}$$

In general, we observe D and want to make inferences about w. Using Bayesian inference, the probability $p(w|D, \alpha, \delta)$ is,

$$p(w|D, \alpha, \delta) \propto p(w|\alpha) \prod_{c=1}^{max(w)} p(D_{o\in c,f}|\delta, w) \tag{6}$$

where the normalization constant can be obtained by summing over possible sets of categories. This is not analytically tractable, and there are a variety of methods for approximate inference. One method which is gauranteed to converge to the correct answer is Gibbs sampling. In Gibbs sampling, we imagine probabilistically changing the category assignment of one object at a time, based on the fit between the object and each possible category. Formally,

$$p(w_i = c|w_{-i}, D, \alpha, \delta) \propto p(w'|\alpha) \prod_{c=1}^{max(w')} p(D_{o\in c,f}|\delta, w') \tag{7}$$

where w_{-i} are the assignments of all other objects, and w' is the proposed system of categories with $w_i = c$ and all other objects remaining the same. This is normalized based on all possible assignments of object i to categories c, where c could be a new category with only object i.

Cycling through the objects, the algorithm tends to move to sets of categories that have higher probability. In the long run, by storing sample categories, we are guaranteed that the proportion of samples of w will converge to $p(w|D, \alpha, \delta)$. We can then use the samples from the Gibbs sampler to answer questions about how likely two objects are to belong to the same category.

A.2. CrossCat

CrossCat extends the single system model by allowing multiple systems of categories. Within one particular system, CrossCat is the same as the single system model, aiming to discover the best system of categories for the objects based on the given features. However, in CrossCat, each

[14] This can be thought of as the number of imaginary seed objects that start new categories. Larger values indicate that we think there are more categories.

system contains a different categorization of the objects, and CrossCat learns how many systems are necessary and which features belong together in which systems.

Eq. 5 defines the generative process for the single system model. CrossCat extends this, by first choosing an assignment of features to some number of systems with probability $p(s|\alpha)$ using the CRP, then for each system choosing assignments of objects to some number of categories, each with probability $p(w|\alpha)$, and then generating data with probability $p(D|s,w,\delta)$. Combining all of the pieces, the probability of a particular set of systems and categories is given by

$$P(s,\{w\}|D,\alpha,\delta) \propto P(s|\alpha)\prod_{k=1}^{max(s)} P(w^k|\alpha) \times \prod_{c=1}^{max(w^k)} P(D_{o\in c_k, f\in k}|w^k,s,\delta) \qquad (8)$$

where w^k is the categorization associated with system k and $D_{o\in c_k, f\in k}$ is the subset of matrix that includes the features in system k and the objects in category c_k.

Inference can be performed using an extended version of the algorithm used for the single system model. Because within a system, CrossCat is the same as the single system model, the same Gibbs sampling algorithm is used to move objects within a system. It is alternated with MCMC (specifically, Metropolis-Hastings) proposals that allow features to change systems, including a potentially new system. The proposal works by selecting a feature at random, and assessing the probability of moving it to each existing system (excluding the one that it came from), as well as a new system (with randomly drawn categories). We choose a proposal system randomly in proportion to how well the feature fits the system. This defines the proposal probability for the Metropolis-Hastings moves. The probability that we accept the proposal depends on the probability of proposing the move, the probability of reversing the move (via the same proposal scheme), and the probability of the old and new states. Like the Gibbs sampling algorithm, this MCMC algorithm tends to move to better states, and is guaranteed in the long run to converge to the probability distribution $P(s,\{w\}|D,\alpha,\delta)$. Using samples drawn from this algorithm, we can approximate the probability of two objects being in the same category.

A.3. Subset attention model

The subset attention model extends the single system model by choosing a subset of n features, and running the single system model. The size of the subset was allowed to vary, but similar to the CRP prior for features in CrossCat, subsets were weighted based on their size. The weights allowed the size of the subset to vary, while maintaining the same number of free parameters as in CrossCat. To minimize differences among the models, the weights were based on the CRP prior probability of partitioning the features into two groups, one of size n and the other of size $F - n$. The full probability of a subset n and associated categories w is

$$p(s,w|D,\delta,\alpha) \propto p(s|\alpha)p(w|\alpha)p(D_s|s,w,\delta). \qquad (9)$$

where s specifies the subset of features under consideration, $p(s|\alpha)$ is the probability of choosing that subset (computed as described above), and D_s is the subset of the data that includes only the features currently under consideration.

To generate predictions for the artificial data sets, we scored systems of categories using Eq. 9, with subsets ranging from 1 to 6 features. We considered the five most common solutions provided by people, as well as any solutions that were higher probability than people's choices according to the single system model. As for CrossCat, the subset attention model's predictions reflects the probability of each solution, given all possible subsets, weighted by the prior probability of a subset of that size. The figures show the relative scores, where zero is defined to be the one system that had lowest probability out of these possibilities.

To generate predictions for the real-world data sets, we considered subsets, n, of between 1 and 10 features. For each size, we chose a subset and generated samples as described for the single system model. This resulted in a set of samples for that subset of features. We conducted 20 runs for each n, where each run resulted in 100 samples. The samples were weighted by the prior probability of that subset, and the sum of the weighted frequencies of object co-occurrences was computed, approximating the probability that each pair of objects belonged in the same category, under the model.

A.4. Subset search model

The subset search model extends the subset attention model by systematically (as opposed to randomly) choosing subsets of features. An initial feature was chosen randomly, and subsequent features were chosen to be similar. This model includes an additional parameter, which determines how similar is similar enough. Since there is no correct setting of this parameter *a priori*, we consider multiple possible values and weight and combine the results (see below). All other details of implementation were identical to those of the subset attention model, with the exception that the probability of a subset of features depended on the selection process in addition to the CRP contribution.

To implement feature search, we considered two variables: how similarity between features was determined and whether feature selection was based on average or maximal similarity between the currently selected features and new features. Feature similarity was formalized using the rand index (Rand, 1971), the adjusted rand index (Hubert & Arabie, 1985), and the wallace index (Denoeud, Garreta, & Guenoche, 2005). Either maximum or average similarity between new and already chosen features was used to determine how similar a new feature was to a group of already chosen features (cf. Osherson, Smith, Wilkie, López, & Shafir, 1990). Finally, given the feature of similarity between new features and the group, a single feature was deterministically chosen. This process was repeated until the the similarity measure fell below the critical value. The critical value was varied between 0.1 and 0.9 on different runs of the model. The samples based on runs at all values were combined and weighted based on the CRP. The combination of all of these concerns led to 6 unique models (3 × 2).

Simulations were conducted for each model for both the artificial and real-world data sets. Models performed comparably on the artificial data. For the real-world data, there was some variability in performance. Generally, most models performed comparably, and the only notable difference was that the model using the adjusted Rand index with average similarity performed better than the others on the animals data (as measured by correlation coefficient), and the results for that model are presented throughout the paper.

A.5. Repeated recategorization model

The repeated recategorization model predicts multiple categorizations based on initially categorizing the objects, identifying features that are poorly explained, recategorizing based on those features, and repeating. We formalized this as an extension of the single system model, with modications as follows. First, the initial categorization was obtained by running the single system model and recording 100 samples from the posterior and the MAP solution. For each feature we computed the log probability of the feature given the categories in the MAP solution. Features that are better explained by the categories have higher log probability. The log probabilities were converted into a measure of relative goodness, by computing the probability of each feature relative to the best feature. The result was goodness ratings that range from 1, which were the most likely features given the categorization, to (in theory) 0. The model was implemented with a free parameter that determined the cutoff for bad features. We varied this parameter from 0.9 down to 0.1 in increments of 0.1. Features with goodness scores below the cutoff were rerun using the single system model, and the process was repeated until there were no more bad features.

Predictions were derived for the artificial category learning and real-world data sets by considering the samples of categories obtained by applying the repeated recategorization model 3 times for each setting of the parameter and merging the samples that resulted from each run. Given the systems of categories, the predictions for the real-world data were derived as for CrossCat: we computed the proportion of solutions in which each pair of objects was in the same category. For the articial data sets, we computed the relative frequency of the different categorizations in the samples from the posterior. Because under Gibbs sampling a category is sampled in proportion to its probability, ranking the categorizations by frequency approximates the rank order of their probabilities.

Appendix B. Formal details for additional alternative models

To convince ourselves that cross-cutting categories were the best explanation of the data, we implemented an additional set of instantiations of the subset search idea. This included 12 different variants of the basic idea of subset search, which are detailed below.

B.1. Alternative subset search models

The main difference between these versions of the model and the ones discussed in the text is that these allow probabilistic selection of features by first choosing a number of features, then choosing features one at a time based on similarity. To allow for uncertainty about the right number of features, we weighted samples based on the probability of a group of features of this size under the CRP and generated predictions based on the weighted average.

To implement feature search, we considered three variables: how similarity between features was determined, whether feature selection was based on average or maximal similarity between the currently selected features and new features, and whether features were chosen probabilistically or deterministically. Feature similarity was formalized using the rand index (Rand, 1971), the adjusted rand index (Hubert & Arabie, 1985), and the wallace index (Denoeud et al., 2005). Either maximum or average similarity between new and already chosen features was used to determine how similar a new feature was to a group of already chosen features (cf. Osherson et al., 1990). Finally, given the similarity between new features and the group, a single feature was chosen. Either the maximally similar feature was chosen (deterministically), or features were chosen in proportion to their relative similarity (stochastically). The combination of all of these concerns lead to 12 unique models (3 × 2 × 2).

Simulations were conducted for each model for both the artificial and real-world data sets. All models performed comparably on the artificial data. For the real-world data, there was some variability in performance. Generally, several models performed comparably, and the only notable difference was a slight tendency for stochastic feature selection to outperform deterministic feature selection. The model using the adjusted rand index, average similarity, and stochastic feature selection performed best, as measured by average correlations on the real-world data sets, but the results were not better than those discussed in the text.

References

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

Barsalou, L. W. (1983). Ad hoc categories. *Memory and Cognition, 11*, 211–227.

Barsalou, L. W. (1991). Deriving categories to achieve goals. In G. Bower (Ed.). *The psychology of learning and motivation: Advances in research and theory* (Vol. 27, pp. 1–64). New York: Academic Press.

Boster, J. S., & Johnson, J. (1989). Form or function: A comparison of expert and novice judgements of similarities among fish. *American Anthropologist, 91*, 866–889.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121–152.

DeDeyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., Voorspoels, W., et al. (in press). Exemplars by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavioral Research Methods*.

Denoeud, L., Garreta, H., & Guenoche, A. (2005). Comparison of distance indices between partitions. In *Applied stochastic models and data analysis*.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman and Hall.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science, 32*, 108–154.

Griffiths, T., & Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In *Advances in neural information processing systems 11*.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proceedings of the 29th annual conference of the Cognitive Science Society*.

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

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.

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

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review, 9*, 829–835.

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

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

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychological Bulletin and Review, 10*, 517–532.

Medin, D. L., Ross, N., Atran, S., Coley, J. D., Proffitt, J., & Blok, S. (2005). Folkbiology of freshwater fish. *Cognition, 92*, 1–37.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 100*, 254–278.

Murphy, G. L. (1993). A rational theory of concepts. In I. V. Mechelen, J. Hampton, R. Michalski, & P. Theuns (Eds.), *The psychology of learning and motivation: Categorization by humans and machines*.

Neal, R. (1998). Markov chain sampling methods for Dirichlet process mixture models.

Nelson, L. J., & Miller, D. T. (1995). The distinctiveness effect in social categorization: You are what makes you unusual. *Psychological Science, 6*, 246–249.

Nguyen, S. (2007). Cross-classification and category representation in children+s concepts. *Developmental Psychology, 43*, 719–731.

Nguyen, S., & Murphy, G. (2003). An apple is more than a fruit: Cross-classification in children+s concepts. *Child Development, 74*, 1–24.

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Nosofsky, R. M. (1991). Relation between the rational model and the context model of categorization. *Psychological Science, 2*, 416–421.

Nosofsky, R. M., Palemeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53–79.

Osherson, D., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*(2), 185–200.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 241–248.

Pothos, E. M., & Bailey, T. M. (2009). Predicting category intuitiveness with the rational model, the simplicity model, and the generalized context model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1062–1080.

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26*, 303–343.

Pothos, E. M., & Close, J. (2008). One or two dimensions in spontaneous classification: A simplicity approach. *Cognition, 107*, 581–602.

Proffitt, J. B., Coley, J. D., & Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(4), 811–828.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*, 846–850.

Rasmussen, C. (2000). The infinite Gaussian mixture model. In *Advances in neural processing systems 12*.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382–439.

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

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In *Proceedings of the 28th annual conference of the Cognitive Science Society*.

Shafto, P., & 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.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, 75.

Smith, E. R., Fazio, R. H., & Cejka, M. (1996). Accessible attitudes influence categorization of multiply categorizable objects. *Journal of Personality and Social Psychology, 71*, 888–898.

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

Zarate, M. A., & Smith, E. R. (1990). Person categorization and stereotyping. *Social Cognition, 8*, 161–185.