The Role of Cross-cutting Systems of Categories in Category-based Induction

Neil A. Smith (neil.smith@louisville.edu)
Department of Computer Engineering & Computer Science
University of Louisville

Patrick Shafto (p.shafto@louisville.edu)
Department of Psychological & Brain Sciences
University of Louisville

Abstract

Prediction is arguably the most fundamental problem that people face. Having discovered that some object possesses a particular feature, how is it that people are able to accurately infer that another object exhibits the property? Psychologists have actively studied this reasoning process; yet, current models of induction cannot provide an explanation for the entirety of the related phenomena. One reason may be that current models fail to account for people's ability to assess multiple categories when making an inference. Building on previous research (Shafto et al., 2006), we present a model of inductive reasoning based on cross-cutting knowledge representation. We present an experiment that investigates the ability of this model to account for known inductive phenomena. We show that a model which assesses multiple kinds of knowledge explains the flexibility of human inference, better than models relying on a single kind of knowledge.

Keywords: Category-based induction; Bayesian model; Cross-categorization.

One of the most remarkable features of human reasoning is the ability to predict unobserved aspects of the world. For instance, consider learning that *mammals* have sesamoid bones. Based on this knowledge, people reliably predict that *wolves* are more likely to have this property than are *oxen* (Sloman, 1993). What knowledge supports these systematic predictions?

To understand the underlying knowledge, psychologists have studied people as they reason about novel properties (like "has sesamoid bones"). These efforts have uncovered a variety of systematic patterns, which constrain models describing how knowledge supports inference. For instance, Osherson et al. (1990) describe a "monotonicity" effect, that occurs when people are asked to reason about a novel property possessed by some object; for instance, consider learning that *penguins* have a novel property. When people are asked to predict whether another specimen or other specimens, such as *all birds*, will have the property, the strength of their prediction depends on the number of examples that have the property. So, if told that both *penguins* and *finches* have the property, people tend to predict that *all birds* are more likely to have it, than if only told about *penguins*. Interestingly, while this inference may appear to be quite tenable, observe that if the additional exemplar was changed from *finches* to *dolphins*, it would seem to temper a willingness to generalize the property to *all birds*, a phenomenon called "non-monotonicity".

In an attempt to explain this *category-based* induction, psychologists have developed computational models; however, extant models have yet to account for the entirety of the characteristic phenomena. One reason may be that previous models rely on a single kind of knowledge, but there is evidence that people use multiple kinds of knowledge to guide prediction (e.g. Ross and Murphy, 1999). For example, an attribute possessed by *penguins* and *dolphins* seems unlikely to be true of *all birds* because there is an alternative categorization that would explain the shared property; namely, *aquatic creatures*.

Building off of work by Shafto et al. (2006), we propose a novel model of category-based induction based on cross-cutting categories. We contrast this model with a well-known account of category-based induction, the feature-based induction model (FBIM; Sloman, 1993). We present an experiment comparing the performance of the model in predicting people's inferences on a number of known phenomena. Finally, we conclude by discussing implications of the findings for categorization, inductive reasoning, and learning.

Category-based Induction

Studies in category-based induction typically follow a paradigm introduced by Rips (1975), in which participants are asked to rate the strength of arguments of the form:

Zebras have sesamoid bones
Hippos have sesamoid bones
___
All mammals have sesamoid bones

The statements above the line indicate the premises of the argument, which are assumed to be true. The task is to assess the likelihood that the category below the line, the conclusion, has the property. The most often-cited literature on category-based induction makes use of so-called *blank properties*, such as "secretes uric acid crystals" or "travels in groups". The assumption in their use is that blank properties carry minimal a priori knowledge to guide people's inferences. In this paper, we choose an even more generic attribute for our stimuli, namely, "Property X". We use this not simply for convenience, but because evidence has shown that the *blank properties* may be influential in people's judgement making. For example, when Heit and Rubenstein (1994) asked people to reason about an anatomical or physiological property of animals (e.g. "secretes uric acid crystals"), people made stronger inferences when the animals were taxonomically related; however, when asked to reason about a behavioral property (e.g. "travels in groups"), people were more confident when the animals shared an ecological niche. Furthermore, we make use of a notation inspired by Tenenbaum, Kemp, and Shafto (2007), that denotes syllogisms, concisely as $P_1, \dots P_n \xrightarrow{prop} C$. Here, P_n represents the nth premise, *prop* is the property used, and C denotes the conclusion. Using this notation, the above argument would be represented as: $\textit{zebras,hippos} \xrightarrow{\textit{sesamoid bones}} \textit{mammals}$.

Researchers studying category-based induction have identified a number of phenomena, which have been used as benchmarks by which to compare the performance of models. Osherson et al. (1990) identified arguments as one of three types: general, specific, or mixed. The *general* class is said to be formed by those arguments, whose conclusion categories properly include all of their premise categories. For example, when reasoning about a *blank property* and given a conclusion category of *all dogs*, having the premise categories of *german shepards* and *chiuauas* would make the argument a general one. An argument is *specific*, if any natural category that properly includes one of the premise or conclusion categories, properly includes the others. To demonstrate, the previous example could be changed into a specific argument by replacing the category *all dogs* with *dalmations*. Finally, a *mixed* argument is any argument that is neither general nor specific. For instance, an argument with premise categories of *dolphins* and *octopuses*, and having a conclusion category of *all mammals* would be considered mixed.

Premise typicality. The more representative the premise categories are of the conclusion category, the stronger the argument. Since *eagles* are a typical bird, relative to *ostriches*, the argument $\textit{eagles} \xrightarrow{X} \textit{birds}$, is considered stronger than the argument would be with *ostriches* as its premise.

Premise diversity. Argument strength is thought to increase as the diversity between the premises increases. For example, compare: $\textit{sheep,dolphin} \xrightarrow{X} \textit{mammals}$ to the argument: $\textit{sheep,leopards} \xrightarrow{X} \textit{mammals}$. *sheep* and *leopards* do not represent the variety of *mammals* well—as their taxonomic relation is pronounced—thus, the former argument appears to be weaker than the latter. Historically, diversity has been a principle in the philosophy of science, which essentially states that a more diverse range of evidence better confirms a hypothesis than does the same amount of similar evidence.

Conclusion specificity. In cases where the conclusion category properly includes the premise categories, the argument with the more specific conclusion category will be considered stronger. For example, $\textit{finches} \xrightarrow{X} \textit{birds}$ is considered stronger than the argument with *animals* as the conclusion category.

Premise monotonicity. Argument strength will tend to increase with the addition of a premise, such that this premise is chosen from the lowest-level category that includes both the categories of the other premises, and the conclusion. To demonstrate: $\textit{bats} \xrightarrow{X} \textit{mammals}$ and $\textit{bats,leopards} \xrightarrow{X} \textit{mammals}$.

Non-monotonicity. Some arguments can be made weaker by adding a premise that converts them into mixed arguments. For example, given the argument $\textit{penguins} \xrightarrow{X} \textit{dolphins}$, if one were to add *finches* as a premise it may reduce the perceived strength of the argument.

Premise-conclusion asymmetry. Originally discovered by Rips (1975), this phenomenon occurs when an argument's premise and conclusion categories are inverted. So, $\textit{bats} \xrightarrow{X} \textit{leopards}$ would become $\textit{leopards} \xrightarrow{X} \textit{bats}$. It was found that the strength of each argument was not evenly rated.

Inclusion fallacy. A person commits an *inclusion fallacy* when they reason that an argument with a general conclusion category, is more cogent than one whose conclusion category is more specific. This phenomenon is termed a *fallacy* because it does not appear normatively rational. For example, observe the arguments: $\textit{crows} \xrightarrow{X} \textit{birds}$ and $\textit{crows} \xrightarrow{X} \textit{ostriches}$. Notice, *birds* is superordinate to *ostriches*, yet, people tend to rate the former argument stronger.

Premise-conclusion identity. Argument strength is absolute when the premise and conlcusion are identical. That is, when the argument is of the form $q \xrightarrow{X} q$.

Premise-conclusion inclusion. An argument whose premise categories are superordinate to the conclusion category is absolute. For example, the argument $\textit{animals} \xrightarrow{X} \textit{birds}$ demonstrates this effect.

Feature exclusion. Premises having no overlapping features with the conclusion will have no effect on the perceived cogency of the argument (Sloman, 1993). Observe the following: $\textit{leopards,monkeys} \xrightarrow{X} \textit{sheep}$. When people were asked to choose the stronger argument between one like that above, and an argument whose second premise included a less similar exemplar, such as *dolphins*, people tended to choose the former as the stronger argument. This phenomenon demonstrates a boundary condition on the diversity principle; though an added premise may lead to more diverse evidence, if the additional premise shares fewer salient features with the conclusion, it will fail to strengthen the argument.

Inclusion similarity. The strength of an argument whose premise category includes the conclusion category will vary depending upon the perceived similarity between the premise and conclusion categories (Sloman, 1993). Sloman (1993) demonstrates this effect with an argument similar to *mammals* $\xrightarrow{X}$ *bats*, and asking subjects to compare against an argument, like *animals* $\xrightarrow{X}$ *leopards*. Since, *leopards* are more typical of the category *mammals*, people tend judge the latter argument as stronger.

Non-diversity. Some arguments with more diverse premises might be judged weaker than those with more similar premises. Consider the arguments: *leopards,seals* $\xrightarrow{X}$ *dolphins* and *leopards,jellyfish* $\xrightarrow{X}$ *dolphins*. While *leopards* and *seals* are more similar than *leopards* and *jellyfish*, people tend to choose the first argument as the stronger one. Note that this particular case cannot be attributed to a *feature exclusion* effect, since salient features in the premise categories are shared with those of the conclusion category (e.g. *seals*, *jellyfish* and *dolphins* are all aquatic animals).

Models of category-based induction

A number of models have been proposed to account for the observed phenomena (Osherson et al., 1990; Sloman, 1993; Medin et al., 2003). These models differ in a number of ways, including in the proposed reasoning mechanisms, and the underlying knowledge representations. Here we focus on the model that provides the most complete account of the known phenomena within a fully specified computational model, the feature-based induction model (Sloman, 1993).

Feature-Based Induction Model (FBIM)

The feature-based induction model (Sloman, 1993), while theoretically expressed as a connectionist network, relies on concepts of *similarity* and *coverage* to explain argument strength. Because it is a connectionist model, the FBIM does not assume an overarching category structure. Rather, the notions of similarity and coverage are in relation to the features of the premise and conclusion categories—not the *whole* of each category. The features are formally represented by a vector, **F**, of numerical elements, f_i, such that each value encodes the absence or presence of a particular feature as 0 or 1. For example, the category, *Robins* would be encoded, as $\mathbf{F}(Robins) = [f_1(Robins)...f_n(Robins)]$.

In this model, *similarity* can be thought of as a function on feature matches and mismatches, and *coverage* as the extent to which the premise features overlap those of the conclusion features. For single premise arguments, the strength of the argument can be expressed as

$$S = \frac{\mathbf{F}(P_1) \cdot \mathbf{F}(C)}{|\mathbf{F}(C)|^{\mathbf{2}}} \qquad (1)$$

The numerator, $\mathbf{F}(P_1) \cdot \mathbf{F}(C)$, yields a scalar that is given by the dot product between the two vectors, and the term $|\mathbf{F}(C)|$, returns the length of the vector. This can be thought of geometrically, as the projection of the vector of premise features, $\mathbf{F}(P_1)$, onto the vector of the conclusion features, $\mathbf{F}(C)$.

This model can account for many of the documented phenomena. For example, the FBIM inspired both the *inclusion similarity* and *feature exclusion* effects. To demonstrate the model, consider the argument: *sheep* $\xrightarrow{X}$ *mammals*, the first step in obtaining the argument strength is to encode the premise and conclusion categories, for example:

$$\mathbf{F}(sheep) = [f_1(sheep) = \text{“}has\ hooves\text{”}, ...] = [1, ...]$$

$$\mathbf{F}(mammal) = [f_1(mammal) = \text{“}is\ furry\text{”}, ...] = [1, ...]$$

The next step, in words, is to calculate the argument strength that is expressed as the ratio in Equation 1. This is the proportion of features in **F**(*mammals*) that is also in **F**(*sheep*), so that the larger this proportion, the stronger the argument is perceived. Since, the premise category *sheep* has a larger number of shared features with the conclusion category *mammals* than, say, a premise category of *bats*, the former argument will yield a higher rating than the latter.

The FBIM can demonstrate many of the documented phenomena; however, it is not without limitations. For instance, Sloman ran correlations between his model and human judgements, and found that in 3 out of 5 cases, Osherson et al.'s Similarity Coverage Model (SCM; 1990) had stronger fit—although, he did provide a defense for this finding (Sloman, 1993). Further, the basic FBIM cannot account for non-monotonicity. [1]

Induction by cross-categorization

Shafto et al. (2006) introduced *CrossCat*, a model of cross-categorization. Given data, the model infers a partitioning of features into different kinds, and, for each *feature-kind*, the model infers a categorization of the objects. This model differs from previous models in that it considers multiple systems of categories to guide an inference, but maintains key similarities to previous approaches. The FBIM considers the entire set of features of the premise and conclusion categories when assessing argument strength. The SCM (Osherson et al., 1990) relies on a taxonomy of categories that applies in all contexts. CrossCat strikes a balance between these approaches allowing flexible use of knowledge to guide inferences like the FBIM, and allowing structured representations to guide inference like the SCM.

CrossCat is formally defined as a model that takes as input a list of features, F, a list of objects, O, and an $O \times F$ object-feature matrix, D. Each entry, $(o, f) \in D$, encodes the value of feature f for object o. For example, given that $o_1 =$“crow” and $f_1 =$“has a beak”, then $D(o_1, f_1) = 1$. The goal is to make inferences about two kinds of situations that correspond to specific and general arguments. We deal with specific arguments first. For specific arguments, the goal is, for a novel feature y, with some observed entries y_{obs} and some unobserved entries y_{unobs}, predict the unobserved entries based on the observed data D and observed entries y_{obs}. Under the model, this prediction is mediated by inferences

[1] An extended model to address this issue was proposed but not tested.

about the likely cross-categorized representations, r,

$$P(y_{unobs}|D,y_{obs}) \propto \sum_r P(y_{unobs}|r)P(r|D,y_{obs}). \qquad (2)$$

That is, representations r, that are probable given the data provide the most weight when predicting y_{unobs}.

These predictions rely on inferring likely cross-categorized representations, r. Under the model, r is composed of two parts: a vector z, of length F, where z_f designates the kind k of feature f, and a set of vectors, $\{w\}$, where w^k contains the categories for kind k. Given a data set, D, the model infers likely combinations of z and $\{w\}$, by approximating the posterior probability of $P(z,\{w\}|D,y)$. The model is specified generatively, by first choosing a partition of the features z, then for each kind $k \in z$, choosing a categorization of the objects w^k, and prior probabilities for each feature in that kind, and then generating the data D^k that correspond to that kind. In a departure from the Shafto et al. (2006) model, we include hyperpriors on the strength s and balance b of feature values. Following Kemp, Perfors, and Tenenbaum (2007), we assume an exponential distribution on the strength parameter, providing a strong expectation that each feature's value will tend to be the same within a category, and a uniform balance.

Formally, given a data set, D, the model infers z, a partition of features into kinds, and $\{w\}$, where w^k contains the categories for kind k, subject to

$$P(z,\{w\}|D,s,b) \propto P(z,\{w\},D,s,b) \qquad (3)$$

$$= P(z)\prod_{k=1}^{K} P(D^k|w^k,s,b)P(w^k)P(s^k)P(b^k) \qquad (4)$$

where K is the number of feature-kinds in z, D^k is the portion of D that must be explained by system k, and $P(D^k|w^k)$ is the process that generates the data for each feature-kind. The prior distribution on feature partitions, and the prior on objects into categories are denoted by $P(z)$ and $P(w^k)$, respectively. Finally, s^k and b^k represent hyperpriors on the feature values. To evaluate the posterior probability, we must specify each component of Equation 4.

Assignments of features to partitions z are evaluated via a chinese restaurant process (CRP) prior

$$P(z_i = k|z_1,...,z_{i-1}) = \begin{cases} \frac{n_k}{i-1+\alpha}, & \text{if } n_k > 0 \\ \frac{\alpha}{i-1+\alpha}, & k \text{ is a new class.} \end{cases} \qquad (5)$$

This process depends on a parameter α, which controls the strength of the preference for a small number of partitions. As $\alpha \rightarrow 0$, the process tends to produce small numbers of categories. Throughout, we set this parameter to .5, a moderate preference for simpler structures. Assignments of objects to categories for each kind w^k, are also evaluated via the CRP.

In this paper, we consider only binary features, and we therefore choose a Beta-Bernoulli model evaluating the probabilities of feature values. For simplicity, a feature is assumed to be a priori independent, and for a given feature, values in different categories are independent. Thus, the probability of $D(o \in c,f)$ depend only on the number of true and false values of the feature for those objects, and the parameters s and b,

$$P(D(o \in c,f)|w,z) = \frac{Beta(\#true + sb, \#false + s(1-b))}{Beta(sb, s(1-b))} \qquad (6)$$

where s and b represent the strength (measured in number of prior observations) and balance (the expected proportion of true values), and $Beta$ represents the Beta function.

We used Markov Chain Monte Carlo (MCMC) to generate predictions. The algorithm proposes moving features between kinds, objects between categories, and changing the values of s and b. In each iteration, the algorithm tends to prefer values that improve on the current state. Although we believe roughly similar heuristics may be used by people, we emphasize that our main claim is about the importance of cross-cutting category structure, not the behavior of this particular inference algorithm.

For specific arguments, the model predicts that, for a novel property, objects in the same category for that property's kind will be more likely to have the same feature.

For the general arguments, the goal is, for a novel feature y, with some observed entries y_{obs}, predict the probability that a general category—e.g. *birds*—has the property. This case is rather different because we need to predict, (a) whether the general category *birds* is sensible given the data, and (b) whether each and every bird is likely to have the property given the category exists. Formally,

$$P(y_{genExt}|D,y_{obs}) \propto \qquad (7)$$

$$\sum_r P(y_{genExt}|birds \in r)P(birds \in r|D,y_{obs}) \qquad (8)$$

where y_{genExt} represents the extension of the general category, and $birds \in r$ indicates that the category birds exists in the representation r. Otherwise, the details are as described above.

Experiment

In our experiment, we investigate how people perceive the cogency of one argument, relative to another. That is, when people are given two arguments and asked to judge which of the two is stronger, and by what magnitude. We compare the results of human judgements to CrossCat, the FBIM, and a conventional infinite mixture model (IMM; Rasmussen, 2000). The IMM differs from CrossCat in that it does not discover systems of categories, rather, it proposes a single clustering of the objects using the entire set of features as a basis.

Method

Participants: Fifteen subjects were recruited from the University of Louisville community, including both students and

Phenomenon	Stimuli	Human	IMM	CC	FBIM
Typicality	1. *eagles→birds* vs. *ostriches→birds*	**3.43**	**.071**	**.099**	**.040**
	2. *leopards→mammals* vs. *bats→mammals* †	**.567**	-.035	**.228**	**.108**
Diversity	3. *sheep,dolphins→mammals* vs. *sheep,leopards→mammals*	-1.7	**.029**	**.193**	**.033**
	4. *eagles,penguins→birds* vs. *eagles,owls→birds*	**.833**	**.051**	**.025**	**.037**
Conclusion Specificity	5. *finches→birds* vs. *finches→animals*	**4.37**	**.092**	**.649**	**.235**
	6. *dolphins→mammals* vs. *dolphins→animals*	**5.60**	**.118**	**.602**	**.152**
Premise Monotonicity	7. *penguins,eagles→birds* vs. *penguins→birds*	**4.50**	-.039	**.233**	**.038**
	8. *bat,leopards→mammals* vs. *bats→mammals*	**3.90**	**.009**	**.217**	**.053**
Non-monotonicity	9. *dolphins→mammals* vs. *dolphins,octopuses→mammals*	**4.43**	**.020**	**.041**	-.019
	10. *penguins→dolphins* vs. *penguins,finches→dolphins* †	**.033**	**.026**	**.033**	.000
Asymmetry	11. *leopards→bats* vs. *bats→leopards*	**.467**	**.011**	**.086**	**.113**
	12. *eagles→penguins* vs. *penguins→eagles*	**.667**	**.011**	**.093**	**.012**
Feature Exclusion	15. *leopards,sheep→monkeys* vs. *leopards,ants→monkeys*	**3.40**	**.002**	**.037**	-.016
	16. *leopards,monkeys→sheep* vs. *leopards,dolphins→sheep*	**2.97**	**.953**	**.061**	-.006
Non-diversity	17. *leopards,seals→dolphins* vs. *leopards,jellyfish→dolphins*	**4.13**	**.020**	**.004**	-.027
	18. *finches,ostriches→owls* vs. *finches,bats→owls* †	**.633**	-.007	**.049**	-.010

Table 1: Mean Predictions of Human and Models.

non-students. The students were offered course credit for participating.

Design and Materials: The dataset used to make model comparisons was an object-feature matrix that was filled in by two coders for an unrelated project. The matrix consisted of 22 animals and 106 features, such that each entry *(i, j)* of the matrix contained a 1 or 0, indicating whether or not feature *j* was true of animal *i*.

Printed surveys were created for this experiment that consisted of eighteen questions in total. The questions were formulated to appear much like those, that elicited the documented phenomena in the subjects of Osherson et al. (1990) and Sloman (1993)—excluding premise-conclusion identity, premise-conclusion inclusion, inclusion similarity, and the inclusion fallacy. Each question consisted of two arguments, designated *A* and *B*, and a line with eleven hatch marks. Below the far-right hatch mark, a letter *B* was placed. Similarly, below the far-left hatch mark the letter *A* was placed.

Procedure: Subjects were handed a survey and a writing utensil. They were told that for each question they should read both of the arguments (i.e. *A* and *B*), and evaluate which one they believe to be stronger. Once they determined the stronger argument, they were asked to indicate on the line how much stronger they believed their choice to be, relative to the alternate. They were informed that argument strength began at the midpoint of the line and increased toward the argument that they deemed stronger.

Results

Participant ratings for all cases of the stimuli were tallied. An exact binomial sign test for each question, revealed that at least one example for all cases was significant ($p < .05$), excluding premise diversity and asymmetry. Both of the cases demonstrating asymmetry were marginal[2] (p = .0625 for both). The second examples of typicality, non-monotonicity, and non-diversity were not found to be significant. These stimuli are presented in Table 1, as number 2, number 10, and number 18, respectively, and are denoted by †. Subject judgements for our first example of premise diversity (i.e. number 3 in Table 1) were not in the expected direction.

We performed two runs of CrossCat (referenced in Table 1 as CC) and the IMM, each using 40 samples, for all of the arguments. The results of these simulations were averaged and are reported in Table 1. For each question, we expected a particular argument to be rated stronger than its alternate, consistent with the induction phenomena. For each of the questions, the difference between the ratings of argument A and B was calculated and averaged across participants. These ratings were compared to the model predictions, which are reported as $\log(\mathrm{arg}_{strong}/arg_{weak})$. Here, arg_{strong} is the argument that we expected would have the stronger rating, similarly arg_{weak} is the weaker rated argument. This measure provides an intuitive way of demonstrating whether a particular model was able to predict the outcome for each of our questions, showing positive values where the model was successful and negative values where the algorithm failed. In Table 1, we report the results that are in the expected direction in bold font. For all, but one of the arguments found to be statistically significant, our model is consistent with the induction phenomena; whereas, the IMM fails to predict the stronger argument in one of our cases—monotonicity—and the FBIM fails for four cases—non-monotonicity, feature exclusion, and non-diversity. For the remaining statistically significant argument, shown as number 3, in Table 1,

[2] Osherson et al. (1990) asked people to make judgements using the same paradigm that we employed, and also used a binomial sign test of significance. Their cases of asymmetry yielded a nonsignificant difference in the predicted direction.

we predicted that this argument would demonstrate the diversity phenomenon, as did CrossCat, the IMM, and the FBIM; however, subjects judged the argument with the less diverse premises to be stronger.

The FBIM cannot account for the non-monotonicity cases, as its bounded by its mathematical definition. In words, any added premise features can only lead to a stronger argument. However, given the context of *penguins* and *dolphins*, features of *finches* do not seem applicable; yet the FBIM considers those features, and any overlap between those features and the features of *dolphins*, leads to a stronger prediction.

The FBIM can explain instances of the feature exclusion phenomenon when there is no overlap between the diversifying premise category, and the conclusion category. While Sloman (1993) was able to demonstrate this effect using the Osherson et al. (1990) dataset, our dataset possessed overlapping features between *ants* and *monkeys*, and between *dolphins* and *sheep*. Thus, the FBIM made stronger predictions for the arguments having more diverse premises; whereas, subjects rated the less diverse arguments stronger.

Finally, the FBIM fails to account for non-diversity when the more diverse premise category shares salient features with the conclusion category, such as *leopards,seals→dolphins* vs. *leopards,jellyfish→dolphins*. We predicted that despite the more diverse premises, the argument with premise categories of *leopards* and *seals* would be rated stronger by subjects, and their ratings agreed with our intuition. CrossCat can give an account for the arguments for which other models of induction fail to explain. CrossCat tends to apply those features that are salient, given the context. So, for instance, in the context of *penguins* and *finches*, the likelihood of adding features of *dolphins* is reduced, since features of birds are more salient.

Discussion

People are remarkable for their ability to accurately predict unobserved aspects of the world. Research in category-based induction seeks to explain people's success. Though many previous models of induction have been proposed, none explain the extant phenomena. In this paper, we have presented a novel model of category-based induction, based on cross-cutting categorization. We have shown that this model out performs two well-established models, the FBIM and IMM, accounting for the greatest number of documented phenomena.

Previous approaches have explored the extremes of knowledge representation. Models such as the SCM (Osherson et al., 1990) maintain a strict taxonomy that is applied across all situations, and cannot explain phenomena that require more flexibility. Models such as the FBIM have no structured representation, and therefore cannot discern predictive features from those that are idiosyncratic. Our approach balances these two extremes, maintaining a strong knowledge representation, but allowing for potentially many sets of categories that guide inference in different contexts.

We showed that our model is able to demonstrate many of the known phenomena that is associated with category-based induction. While we believe that our model is a step in the right direction, there are limitations. For instance, CrossCat only identifies category structures, and does not discover other more richly-structured knowledge representations found in real-world domains, such as tree structures. Further, the model does not take into account the fact that features can sometimes be the cause of other features, and futher use that knowledge to guide prediction. Both of these ideas represent important areas for future work.

References

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

Kemp, C., Perfors, A., & Tenebaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, *10*, 307–321.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Pyschonomic Bulletin and Review*, *100*, 254–278.

Osherson, D., Smith, E. E., Wilkie, O., & Shafir, E. (1990). Category-based induction. *Pyschological Review*, *75*, 185–200.

Rasmussen, C. E. (2000). The infinite gaussian mixture model. In *In advances in neural information processing systems 12* (pp. 554–560). MIT Press.

Rips, L. (1975). Inductive judgements about natural categories. *Journal of Verbal Learning and Verbal Behavior*, *14*, 665–681.

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

Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in the natural world. novices to experts, naïve similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 307–321.

Shafto, P., Kemp, C., Baraff, E., Coley, J. D., & Tenenbaum, J. B. (2005). Context-sensitive reasoning. In *Proceedings of the twenty-seventh annual conference of the cognitive science society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Shafto, P., Kemp, C., Mansinghka, V., Gordon, M., & Tenenbaum, J. B. (2006). Learning cross-cutting systems of categories. In *Proceedings of the twenty-eighth annual conference of the cognitive science society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 231–280.

Tenenbaum, J. B., Kemp, C., & Shafto, P. (2007). Theory-based bayesian models of inductive reasoning. In E. Heit & A. Feeney (Eds.), *Inductive reasoning.* New York: Cambridge University Press.