Psychological Review
2009, Vol. 116, No. 1, 20–58

© 2009 American Psychological Association
0033-295X/09/$12.00 DOI: 10.1037/a0014282

# Structured Statistical Models of Inductive Reasoning

Charles Kemp
Carnegie Mellon University

Joshua B. Tenenbaum
Massachusetts Institute of Technology

Everyday inductive inferences are often guided by rich background knowledge. Formal models of induction should aim to incorporate this knowledge and should explain how different kinds of knowledge lead to the distinctive patterns of reasoning found in different inductive contexts. This article presents a Bayesian framework that attempts to meet both goals and describe 4 applications of the framework: a taxonomic model, a spatial model, a threshold model, and a causal model. Each model makes probabilistic inferences about the extensions of novel properties, but the priors for the 4 models are defined over different kinds of structures that capture different relationships between the categories in a domain. The framework therefore shows how statistical inference can operate over structured background knowledge, and the authors argue that this interaction between structure and statistics is critical for explaining the power and flexibility of human reasoning.

*Keywords:* inductive reasoning, property induction, knowledge representation, Bayesian inference

Humans are adept at making inferences that take them beyond the limits of their direct experience. Even young children can learn the meaning of a novel word from a single labeled example (Heibeck & Markman, 1987), predict the trajectory of a moving object when it passes behind an occluder (Spelke, 1990), and choose a gait that allows them to walk over terrain they have never before encountered. Inferences like these may differ in many respects, but common to them all is the need to go beyond the information given (Bruner, 1973).

Two different ways of going beyond the available information can be distinguished. Deductive inferences draw out conclusions that may have been previously unstated but were implicit in the data provided. Inductive inferences go beyond the available data in a more fundamental way and arrive at conclusions that are likely but not certain given the available evidence. Both kinds of inferences are of psychological interest, but inductive inferences appear to play a more central role in everyday cognition. We have already seen examples related to language, vision, and motor control, and many other inductive problems have been described in the literature (Anderson, 1990; Holland, Holyoak, Nisbett, & Thagard, 1986).

This article describes a formal approach to inductive inference that should apply to many different problems, but we focus on the problem of property induction (Sloman & Lagnado, 2005). In particular, we consider cases where one or more categories in a domain are observed to have a novel property and the inductive task is to predict how the property is distributed over the remaining categories in the domain. For instance, given that bears have sesamoid bones, which species is more likely to share this property: moose or salmon (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Rips, 1975)? Moose may seem like the better choice because they are more similar biologically to bears, but different properties can lead to different patterns of inference. For example, given that a certain disease is found in bears, it may seem more likely that the disease is found in salmon than in moose—perhaps the bears picked up the disease from something they ate.

As these examples suggest, inferences about different properties can draw on very different kinds of knowledge. A psychological account of induction should answer at least two questions—what is the nature of the background knowledge that supports induction, and how is that knowledge combined with evidence to yield a conclusion? The first challenge is to handle the diverse forms of knowledge that are relevant to different problems. For instance, inferences about an anatomical property like sesamoid bones may be guided by knowledge about the taxonomic relationships between biological species, but inferences about a novel disease may be guided by ecological relationships between species, such as predator–prey relations. The second challenge is to explain how this knowledge guides induction. For instance, we need to explain how knowledge about ecological relationships ("bears eat salmon") is combined with evidence ("salmon have a disease") to arrive at a conclusion ("bears are likely to carry the disease").

Existing accounts of property induction usually emphasize just one of the questions we have identified. Theory-based approaches (Carey, 1985; Murphy & Medin, 1985) focus on the first question and attempt to characterize the knowledge that supports induction. Studies in this tradition have established that induction often draws on intuitive theories, or systems of rich conceptual knowledge, and

Charles Kemp, Department of Psychology, Carnegie Mellon University; Joshua B. Tenenbaum, Department of Brain and Cognitive Science, Massachusetts Institute of Technology.

This work was supported in part by the Air Force Office of Scientific Research under contracts FA9550-05-1-0321 and FA9550-07-2-0351. Charles Kemp was supported by the William Asbjornsen Albert Memorial Fellowship, and Joshua B. Tenenbaum was supported by the Paul E. Newton Chair. Many of the ideas in this article were developed in collaborations with Neville Sanjana, Patrick Shafto, Elizabeth Baraff-Bonawitz, and John Coley. We thank Erik Sudderth and Bob Rehder for valuable discussions, Sergey Blok and collaborators for sharing data described in Blok et al. (2007), and Steven Sloman for sharing his copy of the feature data described in Osherson et al. (1991).

Correspondence concerning this article should be sent to Charles Kemp, 5000 Forbes Avenue, Baker Hall 340T, Pittsburgh, PA 15217. E-mail: ckemp@cmu.edu

that different properties can invoke different intuitive theories. Theory-based approaches, however, rarely attempt to formalize the content of intuitive theories and are therefore unable to explain precisely how these theories are used for inductive inference. Statistical or similarity-based approaches (Heit, 1998; Osherson et al., 1990; Rips, 1975) offer complementary strengths—they often provide formal models of inductive inference, but they usually work with very limited kinds of knowledge. Similarity-based approaches, for instance, typically assume that the knowledge required for property induction can be captured by a single pairwise relation between the categories in a domain. We argue that each of these traditions needs the other. Theory-based approaches need statistical inference to explain how theories are acquired and used, and statistical approaches will remain limited in scope unless they can incorporate the content of intuitive theories.

This article develops a modeling framework that attempts to combine the strengths of the theory-based and statistical traditions. The problem of property induction can be modeled as a statistical inference about the probability of the conclusion given the observed premises. A Bayesian approach to this problem incorporates a prior distribution, and we suggest that this prior distribution is often generated by intuitive theories. Bayesian models are criticized in some contexts for relying on prior distributions (Edwards, 1972), but sensitivity to prior knowledge is a distinct advantage when modeling inductive reasoning. The prior distribution used by a Bayesian model can capture background knowledge of arbitrary sophistication, and different prior distributions can account for different patterns of reasoning in different inductive contexts.

To turn these ideas into a computational framework, we develop a general method for capturing some of the rich background knowledge embedded in intuitive theories. Our approach is summarized by Figure 1. For any given problem, the starting point is a structure representing the key relationships between categories in a domain. For example, the tree structure in Figure 1a captures knowledge about the taxonomic relationships among a group of biological species, the one-dimensional spaces in Figures 1b and 1c capture knowledge about the body weights of these species, and the directed graph in Figure 1d captures knowledge about predator–prey relationships.

In addition to knowledge about relationships between categories, a reasoner must also know how the distribution of a given property is likely to depend on these relationships. For example, a biological property like “has enzyme X132” is likely to respect taxonomic boundaries and will probably be shared by the species belonging to some subtree of the taxonomy in Figure 1a. Species which “ideally consume around 15 g of sodium per week in the wild” will probably have roughly similar weights and will fall within some subinterval of the one-dimensional space in Figure 1b. Two properties may depend on the same underlying structure in different ways—for example, a species will be “heavy enough to trigger an average pit trap” if its weight exceeds some threshold along the dimension in Figure 1c. Finally, a reasoner might know that properties like “carries leptospirosis” are likely to be transmitted along the links in the food web in Figure 1d but could also arise from other sources outside the web.

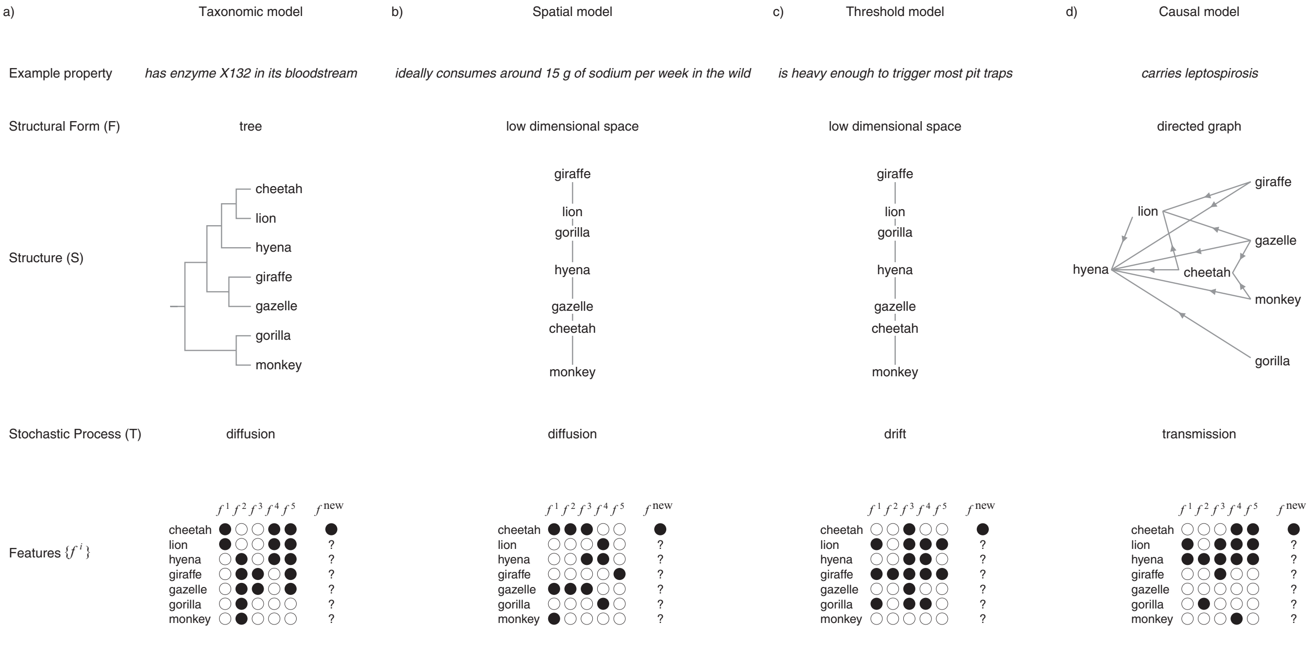

*Figure 1.* Structured statistical models for reasoning about four kinds of properties. The models rely on structures of three different forms, and each model combines a structure $S$ and a stochastic process $T$ to generate a prior distribution, $p(f|S,T)$, on properties. The bottom row shows properties with high prior probability according to each model. Each column represents the extension of a property—for example, property $f^1$ in the taxonomic data set is shared only by cheetahs and lions. The prior distribution $p(f|S,T)$ specified by each model can be used to make inferences about a partially observed property, $f^{\text{new}}$. For instance, a diffusion process over a tree captures the intuition that animals nearby in the tree are likely to have similar properties and predicts that lions are likely to have enzyme X132 if cheetahs are known to have this property.

We will formalize the knowledge just described by defining stochastic processes over the structures in Figure 1. Each stochastic process induces a prior distribution over possible extensions of the property of interest. The "diffusion process" captures the expectation that the property will be smoothly distributed over the underlying structure—in other words, for any pair of nearby categories, probably both or neither will have the property. The "drift process" captures the idea that categories towards one end of the underlying structure are most likely to have the property of interest. Finally, the "transmission process" captures the idea that diseases can be transmitted over a food web. The bottom row of Figure 1 shows some of the extensions with highest prior probability ($f^1$ through $f^5$) for four inductive contexts. Note, for example, that the highest probability extensions in the food web example (Figure 1d) all specify that the species at the top of the food web (here, the hyena) is likely to carry the disease.

A prior distribution on property extensions allows our framework to make inferences about sparsely observed properties. All of our models rely on domain-general Bayesian inference, but the different prior distributions summarized by Figure 1 lead to qualitatively different patterns of reasoning. Given that cheetahs have a certain property, for instance, the models predict that the animals most likely to share the property are lions (taxonomic model), gazelles (spatial model), giraffes (threshold model), and hyenas (causal model). The taxonomic and spatial models (Figures 1a and 1b) can be viewed as accounts of similarity-based reasoning, where similarity is defined in terms of proximity in the underlying structure. The threshold and causal models, however, capture inferences that cannot be explained by similarity alone and that often run counter to similarity.

It is important to note that we offer a modeling framework rather than a single model of induction. Our framework can be used to construct many specific models, and each of these models may make some contribution in its own right. Here, however, we put four of these models side by side (Figure 1) in order to argue for a unified approach to inductive reasoning. Some researchers suggest that the mind makes use of a large number of "content-specific, domain-specific inference engines" (Fiddick, Cosmides & Tooby, 2000, p. 12), and that domain-general principles can play little role in explaining human thought. We propose, however, that general-purpose principles can provide a foundation for explaining many aspects of human inductive reasoning. Our framework highlights two kinds of general principles: principles related to inference, and principles related to knowledge representation. Like many previous researchers (Oaksford & Chater, 2007), we suggest that inductive judgments can often be characterized as statistical inferences, and that domain-general statistical inference can help to explain responses to inductive problems that appear quite different on the surface. General proposals about knowledge representation are less common in the reasoning literature, but we suggest that the background knowledge that supports many instances of property induction can be captured by a structure that represents relationships between categories and a stochastic process that specifies how features are distributed over that structure (Figure 1). Note, however, that very different structures and stochastic approaches may be needed for different inductive contexts. An appropriate motto for our approach is "unity in diversity"—the models in Figure 1 share several important characteristics, but the differences between these models help to explain why patterns of inductive inference vary dramatically when the inductive context is changed.

Because we aim for a unifying approach to induction, many of the themes we emphasize have been previously discussed, including the role of background knowledge (S.A. Gelman & Markman, 1986; Heit & Rubinstein, 1994) and the idea that qualitatively different representations are needed to capture background knowledge about different inductive contexts (Pruzansky, Tversky, & Carroll, 1982; Shepard, 1980). Despite the acknowledged importance of these ideas, few if any formal approaches to inductive reasoning incorporate these insights. Previous formal models assume that background knowledge can be captured by a similarity measure and a taxonomy (Osherson, Stern, Wilkie, Stob, & Smith, 1991), a list of features (Sloman, 1993), a set of neural network weights (Rogers & McClelland, 2004), a list of prior probabilities (Heit, 1998), or a similarity measure together with a list of prior probabilities (Blok, Medin, & Osherson, 2007). Unlike our approach, none of these approaches provides a general framework for explaining how inductive inferences draw on the knowledge embedded in different kinds of structures. In addition, none of these previous approaches provides quantitative models of inductive reasoning over the diverse range of contexts that we consider.

The two central themes in our work, structure and statistics, are perennial topics of debate in the reasoning literature and in cognitive science more generally. Several authors have argued that structured approaches and statistical approaches can both help to explain inductive reasoning and that their strengths and weaknesses are complementary. Structured approaches are able to account for the sophistication and complexity of semantic knowledge (Keil, 1979), but these approaches are often seen as brittle and unable to cope with noise and exceptions (Rogers & McClelland, 2004). Statistical approaches, on the other hand, appear to cope well with noise and uncertainty, but often seem unable to capture core aspects of abstract semantic knowledge (Kemp & Tenenbaum, in press; Marcus, 1991). It seems natural to attempt to combine the strengths of both approaches, but how?

A number of proposals follow a hybrid approach in which two systems of reasoning operate side-by-side in the human mind (Erickson & Kruschke, 1998; Kahneman & Tversky, 2002; Pinker, 1999; Sloman, 1996). One system might be specialized for rule-based, symbolic thought, whereas the other is specialized for implicit statistical judgments. Our work is fundamentally different and offers a single, integrated, *structured statistical* approach to modeling inductive reasoning. The models in Figure 1 demonstrate how representations with different structures can capture the background knowledge that is relevant to different inductive contexts and can define context-specific priors for statistical inferences about the distribution of unobserved properties. In each setting, a single structured statistical model helps to explain how induction is guided by sophisticated background knowledge and how people make flexible inferences in the face of noise and uncertainty.

We began with two fundamental questions about the knowledge that guides induction: What is this knowledge, and how is it used? We will address these questions by showing how background knowledge can be captured using structures and stochastic processes and how knowledge of this sort can be exploited by a Bayesian framework for property induction. From the outset, however, we should acknowledge a third question that is equally (if not more) fundamental: How is this background knowledge itself acquired by human learners? We will ultimately argue that the structured statistical approach to inductive reasoning also provides a principled account of learning and will show, for example, that our taxonomic model (Figure 1a) can be used to learn a tree-structured taxonomy given a set of observed biological properties. Our treatment of learning, however, is far from complete, and future work is needed to expand on the approach that we sketch.

The next section reviews previous empirical and modeling work on property induction, and we then introduce our structured statistical approach to this problem. To demonstrate the generality of our framework, we apply it to four inductive contexts and provide formal descriptions of the four models summarized by Figure 1. We compare each model to several alternatives and evaluate all of these models using behavioral data collected in a range of property induction experiments. In each case, our approach is either the only extant model or is competitive with the best published model for the task. We then turn to the question of how structured background knowledge might be acquired and show how some aspects of this question can be addressed within our Bayesian framework. Finally, we discuss the strengths and limitations of our framework and identify important areas for future work.

## Property Induction

This section introduces several phenomena that have emerged from empirical studies of property induction. We use the formula $P_1, \ldots P_n \rightarrow C$ (*property*) to represent an *n*-premise argument where $P_i$ is the *i*th premise, *C* is the conclusion, and *property* indicates the property used. For example,

> *Hippos have skin that is more resistant to penetration than most synthetic fibers.*
>
> *Therefore housecats have skin that is more resistant to penetration than most synthetic fibers.*

will be represented as *hippos* → *housecats (thick skin).* Similarly, the argument.

> *Hippos have skin that is more resistant to penetration than most synthetic fibers.*
>
> *Therefore all mammals have skin that is more resistant to penetration than most synthetic fibers.*

will be represented as *hippos* → *mammals (thick skin).*

### Inductive Phenomena

The most systematic studies of property induction have generally used *blank properties.* For arguments involving animal species, blank properties are properties that are recognized as biological but about which little else is known—for example, "has enzyme X132." For reasons described below, we usually classify arguments according to inductive context instead of the blankness of the property involved. Arguments like *horses* → *cows (enzyme)* belong to the *default biological context.* We call it the default context because it remains active even when a recognizably biological property ("has enzyme X132") is replaced with a completely blank property ("has property P"). Other researchers, including Rosch (1978), have suggested that people assume a default context when no context is specified.

Although many inductive contexts are possible, the default biological context has probably received more attention than any other. Studies that explore this context have identified many qualitative phenomena: Osherson et al. (1990) identify 13, and Sloman (1993) adds several others. Here we mention just three. *Premise-conclusion similarity* is the effect that argument strength increases as the premises become more similar to the conclusion: for example, *horses* → *dolphins (enzyme*) is weaker than *seals* → *dolphins (enzyme). Typicality* is the effect that argument strength increases as the premises become more typical of the conclusion category. For example, *seals* → *mammals (enzyme*) is weaker than *horses* → *mammals (enzyme)* because seals are less typical mammals than horses. Finally, *diversity* is the effect that argument strength increases as the diversity of the premises increases. For example, *horses, rhinos* → *mammals (enzyme*) is weaker than *seals, squirrels* → *mammals (enzyme).*

Explaining inductive behavior in the default biological context is a challenging problem. Even if we find some way of accounting for all the phenomena individually, it is necessary to find some way to compare their relative weights. Which is better, an argument that is strong according to the typicality criterion or an argument that is strong according to the diversity criterion? The problem is especially difficult because arguments that are strong according to one criterion may be weak according to another—for example, *seals, squirrels* → *mammals (enzyme*) has premises that are quite diverse but are not very typical of the conclusion. For reasons of this sort, our primary measure of model performance will consider quantitative predictions across collections of many arguments. Qualitative contrasts between pairs of arguments are also useful, and we will use them to demonstrate the implications of different modeling assumptions, but the best way for a formal model to prove its worth is to make accurate quantitative predictions (cf. Meehl, 1978).

Blank properties may be useful for designing controlled experiments and for exploring default knowledge about the categories in a domain, but humans can reason about many different kinds of properties, including properties that play a role in rich and specialized systems of prior knowledge. The strength of an argument often depends critically on the property involved. We distinguish between two ways in which a property can influence the strength of an argument. First, it is possible to change the property but preserve the inductive context. Blok et al. (2007) show that *rabbits* → *tigers (24 chromosomes*) is weaker than *rabbits* → *tigers* (*≥20 chromosomes*) because *24 chromosomes* ("has exactly 24 chromosomes") is less likely a priori than *≥20 chromosomes* ("has more than 20 chromosomes"). Second, changing the property will often alter the inductive context. Many researchers have described related effects (S. A. Gelman & Markman, 1986; Heit & Rubinstein, 1994; Shafto & Coley, 2003; Macario, 1991; Smith, Shafir, & Osherson, 1993), and here we mention just three. S. A. Gelman and Markman (1986) show that children reason differently about biological properties (e.g., "has cold blood") and physical properties (e.g., "weighs one ton")—for example, *brontosaurus* → *triceratops (cold blood*) is relatively strong, but *brontosaurus* → *triceratops (weighs one ton*) is relatively weak. Smith et al. (1993) show that prior knowledge about the plausibility of premises and conclusions alters patterns of inference—for instance, *house cats* → *hippos (P*) is strong if P = "has skin that is more resistant to penetration than most synthetic fibers" but weak if P = "has a visual system that fully adapts to darkness in less than five minutes." Finally, Shafto and Coley (2003) show that disease properties draw on causal knowledge about predator–prey interactions and are treated differently from arguments that invoke the default biological context. For example, *gazelles → lions (sesamoid bones*) is weaker than *gazelles → lions (babesiosis),* where babesiosis is a malaria-like disease.

Researchers have also suggested that the very same property can trigger different inductive contexts when the premise and conclusion categories change. Consider the arguments *flies → bees (P)* and *flowers → bees (P*) where P is a completely blank predicate ("has property P"). The first argument triggers the default biological context, but the second argument invokes knowledge about feeding relations (Medin, Coley, Storms, & Hayes, 2005). For this reason we will classify arguments according to the inductive context they trigger instead of the property they use.

This brief survey of the literature suggests that property induction depends on the inductive context in subtle and intricate ways. Despite the complexity of these phenomena, psychologists have made some progress towards explaining the extreme context-sensitivity of inductive reasoning. The theory-based approach (Carey, 1985; Keil, 1989; Murphy & Medin, 1985) claims that induction is supported by *intuitive theories,* or systems of "causal relations that collectively generate or explain the phenomena in a domain" (Murphy, 1993, p. 177). Induction appears so complex because people have many intuitive theories, each of which has different inductive consequences, and because small changes to inductive problems can mean that very different theories are triggered.

This article attempts to develop a unifying computational framework that accounts for many of the phenomena mentioned above. We draw inspiration from the theory-based approach, and part of our task is to build formal models of the knowledge that guides induction. A unifying account, however, must also find a way to reconcile this knowledge-based approach with the many existing results that describe inductive inferences about blank predicates.

### Formal Models

The formal approach to property induction extends at least as far back as the work of Rips (1975). Perhaps the best-known approach is the similarity-coverage model (SCM) of Osherson et al. (1990). The SCM proposes that the strength of an inductive argument is a linear combination of two factors: the similarity of the conclusion to the premises and the extent to which the premises cover the smallest superordinate taxonomic category including both premises and conclusion. Much of the background knowledge required by the model is captured by a measure of pairwise similarity between the categories in a given domain.

Instead of founding a model on similarity, an appealing alternative is to start with a collection of features. In some settings it will be necessary to assume that the features are extracted from another kind of input (linguistic input, say), but in general the move from similarity to features is a move towards models that can be directly grounded in experience. The feature-based model of Sloman (1993) computes inductive strength as a normalized measure of feature overlap between conclusion and example categories. More recently, Rogers and McClelland (2004) have presented a feature-based approach to semantic cognition that uses a feedforward connectionist network with two hidden layers.

Even though different nonblank properties can often lead to very different patterns of inductive projection, most formal models of property induction can handle only the default inductive context evoked by blank predicates. An exception is the Gap model (Smith et al., 1993), and SimProb (Blok et al., 2007) is a related model that has been developed more recently. Both of these models make inferences about properties that we will call "threshold properties." Each of these properties corresponds to a threshold along a familiar dimension—for instance, "has skin that is more resistant to penetration than most synthetic fibers" is closely related to the dimension of skin toughness. For a given domain and threshold property, SimProb uses a measure of pairwise similarity between the categories in the domain and a vector indicating the a priori probability that the threshold property applies to each category. Supplying the model with appropriate prior probabilities allows it to make different inferences about different properties.

Heit (1998) described a Bayesian approach that can also incorporate prior knowledge about the property of interest. Heit's model includes two components: a prior distribution and an inference engine. The inference engine relies on statistical inference, as described in the next section. No formal method for generating the prior is given, but Heit suggests that the prior may be a simple function of a set of properties retrieved from memory, and we will evaluate the performance of a memory-based Bayesian model that is inspired by this suggestion.

The starting point for our approach is closely related to the work of Heit (1998). Our framework relies on Bayesian inference, but we emphasize the importance of choosing and formalizing an appropriate prior. Most of the knowledge that supports induction is captured by the prior, and a computational theory should be as explicit as possible about the knowledge it assumes. We argue that very different kinds of knowledge are needed in different inductive contexts, and show how representations with different structures can generate priors that capture many kinds of knowledge.

The four models described in this article (Figure 1) build on our previous attempts to develop structured statistical models of property induction. We previously developed a tree-based model that captures many of the same intuitions as our taxonomic model (Kemp & Tenenbaum, 2003; Tenenbaum, Kemp, & Shafto, 2007), and the causal model has been described elsewhere (Shafto, Kemp, Baraff, Coley, & Tenenbaum, 2005; Shafto, Kemp, Bonawitz, Coley, & Tenenbaum, 2008). Here we introduce several new models and model variants that emerge from our framework, and compare them both within and across a number of different inductive contexts. Our goal is to provide a comprehensive account of the structured statistical approach to property induction, and to evaluate how well this approach captures the very different patterns of reasoning observed in different inductive contexts.

## Bayesian Property Induction

We develop a computational framework that attempts to handle many of the phenomena described by previous studies of property induction. Suppose that we are working in a domain with a set of known categories, such as a set of biological species. The problem of property induction can be modeled as an inference about a partially observed feature vector—some categories may be known to have the property or feature, but the status of others is unknown. We take a Bayesian approach and assume that a learner begins with a prior distribution over the space of all logically possible feature vectors. The learner observes several labeled examples—

categories observed to have the feature are positive examples and categories observed not to have the feature are negative examples. Based on this evidence, the learner updates his or her distribution over the space of possible feature vectors, and the updated distribution (also known as the posterior distribution) can be used to predict the probability that any given category has the feature of interest.

We formalize this Bayesian approach by specifying a framework with two components: a recipe for specifying prior distributions and an engine for inductive inference. The inference engine implements domain-general statistical inference and remains the same regardless of the inductive context. Different priors, however, are needed for different inductive contexts. Even though different inductive problems may draw on very different kinds of knowledge, we suggest that this knowledge can often be formalized using stochastic processes (e.g., diffusion, drift, or transmission) defined over structures that capture relationships between the categories in a domain (Figure 1).

### *The Bayesian Inference Engine*

The Bayesian approach to induction is extremely general and can be applied to problems which appear quite different on the surface. We describe an engine for Bayesian inference that has previously been used to develop algorithms for machine learning (Haussler, Kearns, & Schapire, 1994) and to model concept learning (Shepard, 1987; Tenenbaum & Griffiths, 2001) and inductive reasoning (Heit, 1998) in humans.

Assume that we are working within a finite domain containing $n$ categories. We will use a running example from the biological domain where the categories are four species: cheetahs, lions, gorillas and monkeys. Suppose that we are interested in a novel property or feature (we use these terms interchangeably). Our framework can handle continuous-valued features, but we focus on the case where the novel feature can be represented as an $n$-place vector $f$ that assigns 1 to each category that has the feature and 0 to all remaining categories. Because there are $n$ categories, the number of distinct feature vectors $f$ is $2^n$, and the 16 possible feature vectors for our running example are shown in Figure 2a. Assume for now that the prior probability $p(f)$ of each feature vector is known. The prior in Figure 2a roughly captures the idea that cheetahs and lions are expected to have similar features and that the same holds for gorillas and monkeys.

Suppose that we observe $l_X$, a label vector for the categories in some set $X$. For instance, the case where $X = \{cheetah, monkey\}$ and $l_X = [1, 0]$ indicates that cheetahs have the novel feature but that monkeys do not (Figure 2b). The observations in $l_X$ can be treated as a partial specification of the full feature vector $f$ that we want to infer. Given these observations, Bayes' rule specifies how our prior distribution $p(f)$ can be updated to produce a posterior distribution $p(f|l_X)$ on the feature vector $f$:

$$p(f\,|l_X) = \frac{p(l_X|f)p(f)}{\sum_f p(l_X|f)p(f)} \tag{1}$$

where the sum in the denominator is over all possible feature vectors f.

The likelihood term $p(l_X|f)$ may vary from setting to setting depending on the process by which the categories in the observation set $X$ are generated and the process by which the labels $l_X$ are generated for those examples. The general discussion

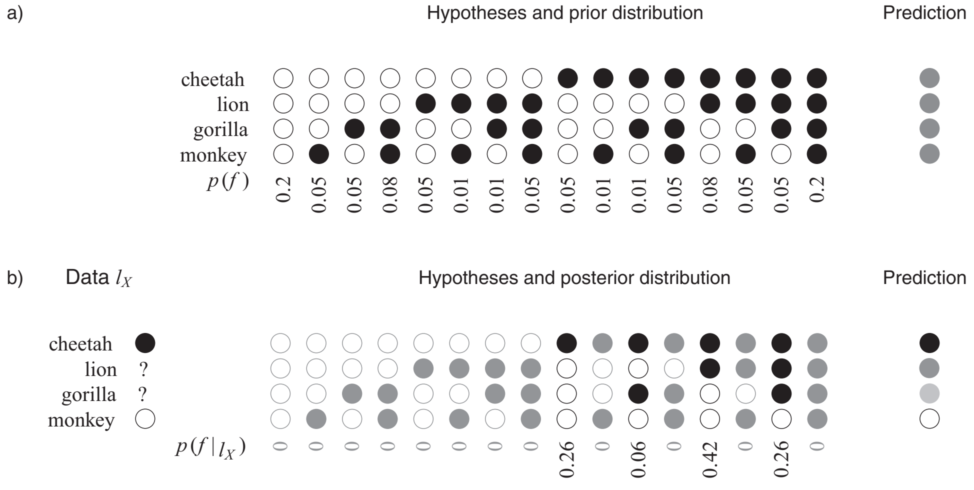

*Figure 2.* (a) Given a domain with four categories, there are 16 distinct hypotheses about the extension of a novel feature. The bottom row of the table shows a prior distribution $p(f)$ over the 16 possible feature vectors. The grayscale vector on the far right shows predictions about individual categories computed by summing over the space of hypotheses. Based on the prior alone, the probability that cheetahs have the novel feature is 0.5, and the 3 remaining entries in the prediction vector are also 0.5. (b) Bayesian property induction. After observing a label vector $l_X$ that indicates that cheetahs have the novel feature but monkeys do not, 12 feature vectors are no longer possible and have been grayed out. The posterior distribution $p(f|l_X)$ can be computed by renormalizing the prior distribution $p(f)$ on the 4 feature vectors that remain. The prediction vector now indicates that cheetahs definitely have the feature, that monkeys definitely do not have the feature, and that lions (0.68) are more likely to have the feature than gorillas (0.32).

returns to this issue, but for now we make two generic assumptions—we assume that $X$ is sampled at random from all subsets of the domain and that the labels $l_X$ are generated without noise. It follows that:

$$p(l_X|f) \propto \begin{cases} 1, & \text{if } f_X = l_X \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $f_X$ specifies the values assigned to set $X$ by feature vector $f$. Equation 2 indicates that label vectors $l_X$ that are inconsistent with the true feature vector $f$ are never observed and that all remaining label vectors are equally likely to be observed. The distribution in Equation 2 is specified up to a normalizing constant that depends on the number of observation sets $X$, which in turn depends on the size of the domain.

Given our assumptions about the likelihood $p(l_X|f)$, Equation 1 is equivalent to

$$p(f|l_X) = \begin{cases} \dfrac{p(f)}{\sum_{f:f_X=l_X} p(f)}, & \text{if } f_X = l_X \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where the sum in the denominator is over the set $\{f : f_X = l_X\}$ of all feature vectors $f$ that are consistent with the label vector $l_X$. Intuitively, Equation 3 is the result of renormalizing the prior distribution $p(f)$ to account for the fact that some feature vectors $f$ are inconsistent with the observations in $l_X$ and therefore now have zero posterior probability. The posterior distribution for our running example is shown in Figure 2b.

Given the posterior distribution $p(f|l_X)$, we can compute $p(f_i = 1|l_X)$, or the posterior probability that category $i$ has the novel feature:

$$p(f_i = 1|l_X) = \sum_{f:f_i=1} p(f|l_X) \tag{4}$$

$$= \frac{\sum_{f:f_i=1,\, f_X=l_X}}{\sum_{f:f_X=l_X} p(f)} \tag{5}$$

where Equation 5 follows from Equation 3. Intuitively, Equation 5 states that the posterior probability $p(f_i = 1|l_X)$ is equal to the proportion of feature vectors consistent with $l_X$ that also set $f_i = 1$, where each feature vector is weighted by its prior probability $p(f)$. The prediction vector in Figure 2b shows predictions $p(f_i = 1|l_X)$ for our running example. For example, the posterior probability that lions have the novel feature given that cheetahs have the feature but that monkeys do not is $\dfrac{0.08 + 0.05}{0.05 + 0.01 + 0.08 + 0.05} = 0.68$.

Other inferences can be formulated similarly. For example, after observing $l_X$, we can compute $p(f_Y = 1|l_X)$, or the posterior probability that all categories in set $Y$ have the novel feature:

$$p(f_Y = 1|l_X) = \frac{\sum_{f:f_Y=1,\, f_X=l_X} p(f)}{\sum_{f:f_X=l_X} p(f)} \tag{6}$$

Intuitively, Equation 5 states that the posterior probability $p(f_Y = 1|l_X)$ is equal to the proportion of feature vectors consistent with $l_X$ that also set $f_Y = 1$, where each feature vector is weighted by its prior probability $p(f)$. In Figure 2b, for instance, the posterior probability that lions and gorillas have the novel feature given that cheetahs have the feature but that monkeys do not is $\dfrac{0.05}{0.05 + 0.01 + 0.08 + 0.05} = 0.26$. To compute the probability that all members of a superordinate category (e.g., "mammals") have a certain feature, we use Equation 6 where set $Y$ includes all members of the superordinate category in the data set under consideration. In Figure 2, "all mammals" includes only 4 species, but each of the animal data sets we consider later includes between 6 and 10 mammal species.

Equations 5 and 6 form part of a computational theory of property induction. If a reasoning system starts with a prior distribution $p(f)$ and knows that the categories in $X$ and the label vector $l_X$ were generated according to the process described by Equation 2, then the probabilities produced by Equations 5 and 6 will be normatively correct. Our working assumption is that human inductive reasoning can be explained as an approximation to Bayesian inference, and we expect that the probabilities produced by these equations will approximate the generalizations made by human reasoners. We make no claim, however, that the equations capture the *mechanisms* that support human reasoning. For example, each equation includes a sum that can range over many different feature vectors $f$, and we do not suggest that people explicitly consider a large set of possible feature vectors when ranking the strength of an argument. Although we do not attempt to provide a process model, a complete account of inductive reasoning should provide explanations at each of the three levels described by Marr (1982), and future work can explore how the computations in Equations 5 and 6 can be approximated by psychologically plausible processes.

### *Generating a Prior*

The prior distribution $p(f)$ should capture expectations about the property or feature of interest. In Figure 2, for instance, the feature vectors with high prior probability should indicate either that cheetahs and lions both have the property or that neither species has the novel property. Formalizing the relevant prior knowledge may initially seem like a difficult problem—if there are $n$ species, somehow we need to generate $2^n$ numbers, one for each possible feature vector. Simply listing these numbers provides little insight—for instance, it does not seem helpful to propose that human knowledge about a set of 50 species is faithfully represented by a list of $2^{50}$ numbers. Instead, we develop an approach where the prior $p(f)$ is generated by two kinds of background knowledge: knowledge about relationships between the categories in a domain and knowledge about how the property of interest depends on these relationships. These two aspects of background knowledge can be formalized as a structure $S$ and a stochastic process $T$ defined over this structure (Figure 1). By combining different structures and stochastic processes, we can capture different kinds of knowledge and account for inductive inferences in many different contexts.

Relationships between categories can be captured by many kinds of structures. One prominent approach focuses on representations expressed as sentences in a compositional language, such as predicate logic (Fodor & Pylyshyn, 1988; Gentner, 1983). Here we take a more inclusive approach, and allow any representation that captures relationships between categories to qualify as a structure. A central theme of our work is that different kinds of structures can capture different kinds of relationships between categories. Tree structures can capture taxonomic relationships between categories, multidimensional spaces can capture proximity relationships between categories, graphs can capture directed relationships between categories, and formulae in predicate logic can capture all of these relationships and many others besides. Our framework has room for all of these representations, but we focus here on graph structures and continuous spaces.

The structures used by our four models are shown in Figure 1. The taxonomic model uses a tree structure to capture taxonomic relationships between biological species, and the spatial model uses a low-dimensional structure to capture proximity relationships between categories. The example in Figure 1 shows a one-dimensional structure that captures the body weights of a set of species, but later we consider a problem where the underlying structure is a mental map, or a two-dimensional representation of the geographic relationships between American cities. The threshold model uses a one-dimensional structure that corresponds to an underlying dimension, such as weight or strength. Finally, the causal model uses a directed graph that captures predator–prey relationships among a group of biological species.

The second component of each model in Figure 1 is a stochastic process $T$ that indicates how the property of interest depends on structure $S$. Our first two models rely on a stochastic process—the diffusion process—that leads to a version of similarity-based reasoning. Structures such as graphs and low-dimensional spaces induce a notion of distance—for any two categories, the representation indicates whether or not the categories are relatively close. The diffusion process formalizes the idea that nearby categories in a structure will tend to have similar properties. When defined over the structures in Figures 1a and 1b, the diffusion process generates a prior favoring features that are shared by all and only the species that belong to some connected region of the representation.

Our remaining two models use stochastic processes that depart from the idea of similarity-based reasoning. The threshold model uses a process—the drift process—which assumes that properties are more likely to be found in certain regions of a given structure. For example, if the representation is a one-dimensional structure, the drift process can specify that the property is more likely to be found towards one end of the structure than the other. The prior generated by this process will favor features that are shared by all categories beyond some point along the structure (Figure 1c). Our final model uses a stochastic process which assumes that properties are noisily transmitted along the arrows in a causal graph. We use this process to capture the common sense notion that diseases often spread from prey to predator through a food web.

We will occasionally use the term *theory* to refer to a structured statistical model that generates a prior $p(f)$ for Bayesian inference. Each of the theories we consider is a combination of a structure and a stochastic process defined over that structure. Some of the most typical theories are systems of causal knowledge, but we use the term more broadly for any set of principles that guides inductive inference (Murphy & Medin, 1985). The term acknowledges our intellectual debt to the previous theory-based approaches, but our usage is more inclusive than that of some authors (Carey & Spelke, 1996; Keil, 1989) in at least two respects. First, we do not require theories to be consciously accessible. Second, we use the term for systems of knowledge (e.g., mathematical knowledge or knowledge of kinship relations) that are not explicitly causal.

It is worth noting that some of the theories summarized in Figure 1 may not meet all the criteria required by previous descriptions of theories, but we view the similarities between our approach and previous theory-based approaches as more important than the differences. Previous theory-based approaches have presented experimental evidence that inductive inferences often draw on relatively sophisticated systems of knowledge—systems that exceed the expressive power of many formal models of inductive reasoning. Bayesian models, however, emphasize the role of prior knowledge, and we will show how these models can incorporate knowledge-rich priors that capture some aspects of intuitive theories.

## Taxonomic Reasoning

We apply our framework first to the default biological context. Generating a prior for any inductive context involves a two-step procedure. First we must identify the structure that best captures the relevant relationships between the categories in the domain. Next we must identify a stochastic process that captures knowledge about how properties tend to be distributed over this representation.

A natural representation for the default biological context is a tree structure where the animals are located at the leaves of the tree. We will work with the tree in Figure 3a, and later we discuss how this structure might be acquired. There are at least two reasons to believe that tree structures play a role in biological reasoning. Anthropologists have shown that cultures all over the world organize living kinds into folk taxonomies, and some have argued further that the tendency to organize living kinds into tree structures is innate (Atran, 1998). If we assume that people develop representations that are well matched to the world, there are also scientific considerations that support the choice of a tree structure. Naturalists since Linnaeus have known that tree representations account well for the distribution of biological features, and the ultimate reason why trees are suitable may be that living kinds are the outcome of a branching process—the process of evolution.

A process for generating properties over the tree should satisfy the intuition that properties will usually be smooth over the tree—in other words, that species nearby in the tree will tend to have similar properties. The process, however, should allow species to share a property even if they are very distant in the tree—as a biologist might say, we need to allow for the possibility of convergent evolution. We also need to allow for exceptions—for example, even though penguins may be located near all the other birds in the tree, we know that they lack some important avian features. All of these requirements can be captured by a process that we will call the *diffusion process.* Similar stochastic processes have previously been used to model response times in decision-making tasks (Ratcliff, 1978), but we use the notion of diffusion to capture background knowledge about the distribution of properties over a structure. We describe the diffusion process as a recipe for generating a single feature vector. If we create a large sample of feature vectors by following the recipe many times, the prior probability of any feature vector is proportional to the number of times it appears in the sample.

To introduce the diffusion process, consider first the case where the underlying structure is a linear graph instead of a tree. The top panel of Figure 4a shows one such graph, where each node in the

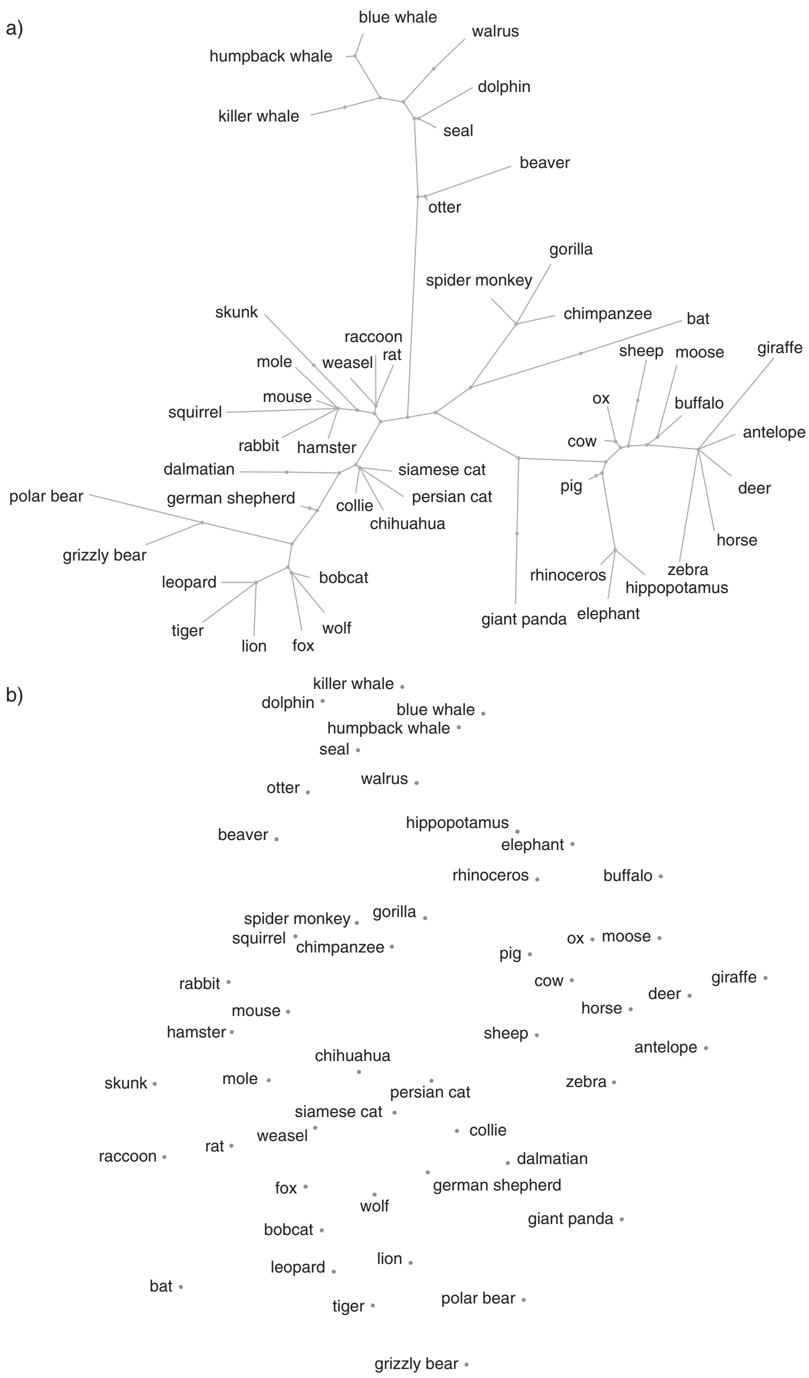

*Figure 3.* (a) A tree and (b) a two-dimensional representation learned from a matrix *D* of human feature ratings that includes 50 species and 85 features.

graph represents a category. The bottom panel shows a feature vector *f* that assigns a value (1 or 0) to each node in the graph. To generate this feature, we first generate a continuous feature vector *y* that is expected to vary smoothly over the graph. We then threshold the *y* vector at 0 to produce a binary feature *f*. The same basic approach can be used to generate features over any undirected graph, including a graph that takes the form of a tree. Intuitively, the diffusion process is most likely to generate features *f* that assign the same value to neighboring nodes in the underlying graph.

Consider now the case where the underlying structure is a tree rather than a linear graph. More formally, suppose that we are working with a set of *n* species and that we have a tree structure *S* where the species are located at the leaves of the tree. To generate

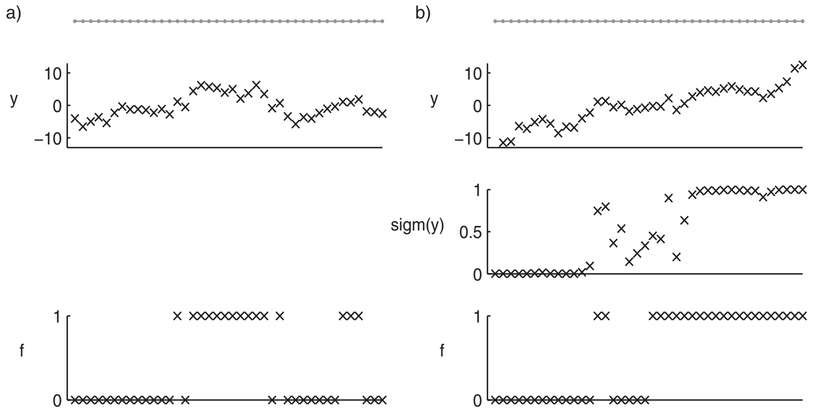

*Figure 4.* Generating binary features from two stochastic processes defined over a linear graph. Each node in the graph represents a category. (a) Sampling a binary feature vector *f* from a diffusion process (Equations 7-8). We first generate a feature vector *y* that assigns a continuous value to each category and that varies smoothly over the graph. We then threshold this continuous feature vector at 0 to create the binary feature vector in the bottom panel. (b) Sampling a binary feature from a drift process (Equation 10). We generate a continuous feature *y* as before, pass it through the sigmoid function *sigm*(·), then toss a coin with weight *sigm*($y_i$) to determine whether the category at position *i* has the feature.

a binary feature *f,* we first generate a continuous feature *y* that includes a real-valued label for every node in graph *S.* The feature *y* is drawn from a prior that ensures that it tends to vary smoothly over structure *S.* Formally, *y* is drawn from a multivariate Gaussian distribution with zero mean and a covariance matrix that encourages nearby nodes in the tree to have similar labels. Appendix A reviews the definition of a covariance matrix, and Figure 5c shows a visual representation of a covariance matrix defined over the

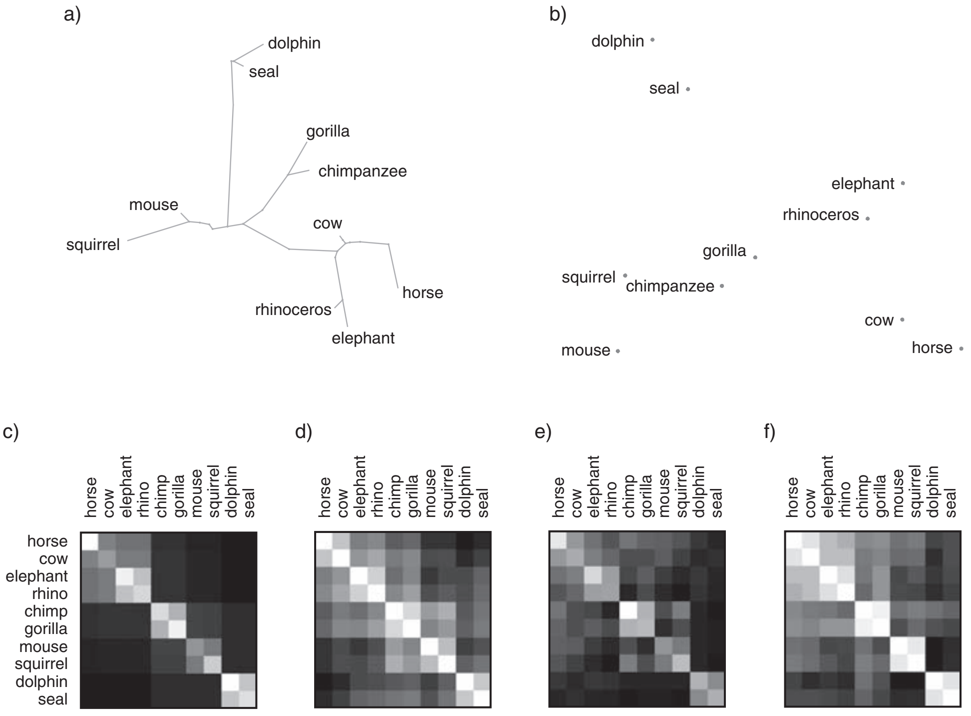

*Figure 5.* Structures and covariance matrices for a set of 10 species. (a)–(b): Substructures of the representations in Figure 3. (c)–(e): Covariance matrices for the taxonomic, spatial, and raw covariance models. White cells indicate pairs of species with high covariance. The matrices in (c) and (d) are generated by diffusion processes over the structures in (a) and (b). The matrix in (e) is a subset of the 50 by 50 covariance matrix 1/85$DD^{\mathrm{T}}$. (f): Human similarity ratings collected by Osherson et al. (1990).

10-species tree in Figure 5a. If we sample continuous features $y$ using this covariance matrix, pairs with high covariance (e.g., chimps and gorillas) will tend to have similar feature values. For instance, if chimps have a value on some dimension that exceeds the mean, the feature value for gorillas on the same dimension is also expected to be higher than normal. Appendix A describes in detail how a covariance matrix can be defined over any graph structure, but the basic intuition is that large entries in the covariance matrix will correspond to pairs of nodes that are nearby in the underlying structure.

After generating a continuous feature $y$, we convert it to a binary vector $f$ by thresholding at zero. The complete generative model can be written as:

$$y \sim N(0,\ K) \tag{7}$$

$$f_i = \Theta(y_i) \tag{8}$$

where $f_i$ is the feature value for category $i$, and $\Theta(y_i)$ is 1 if $y_i \geq 0$ and 0 otherwise. Equation 7 indicates that $y$ is drawn from a zero-mean Gaussian distribution with covariance matrix $K$, and Equation 8 indicates that $f$ is generated by thresholding the continuous vector $y$. This generative process will assign nonzero probability to all of the $2^n$ possible binary features, but the features with high prior probability will tend to be consistent with the covariance matrix $K$. As described in Appendix B the definition of $K$ uses a single free parameter, σ, which captures the extent to which feature values are expected to depart from the mean of the Gaussian distribution in Equation 7. We set $\sigma = 5$ throughout this article.

The Bayesian model with a prior $p(f)$ defined by a diffusion process over a tree can be called the *tree + diffusion* model, but we will also refer to it as the taxonomic model. The first column of Figure 6 shows generalization curves predicted by this model. The

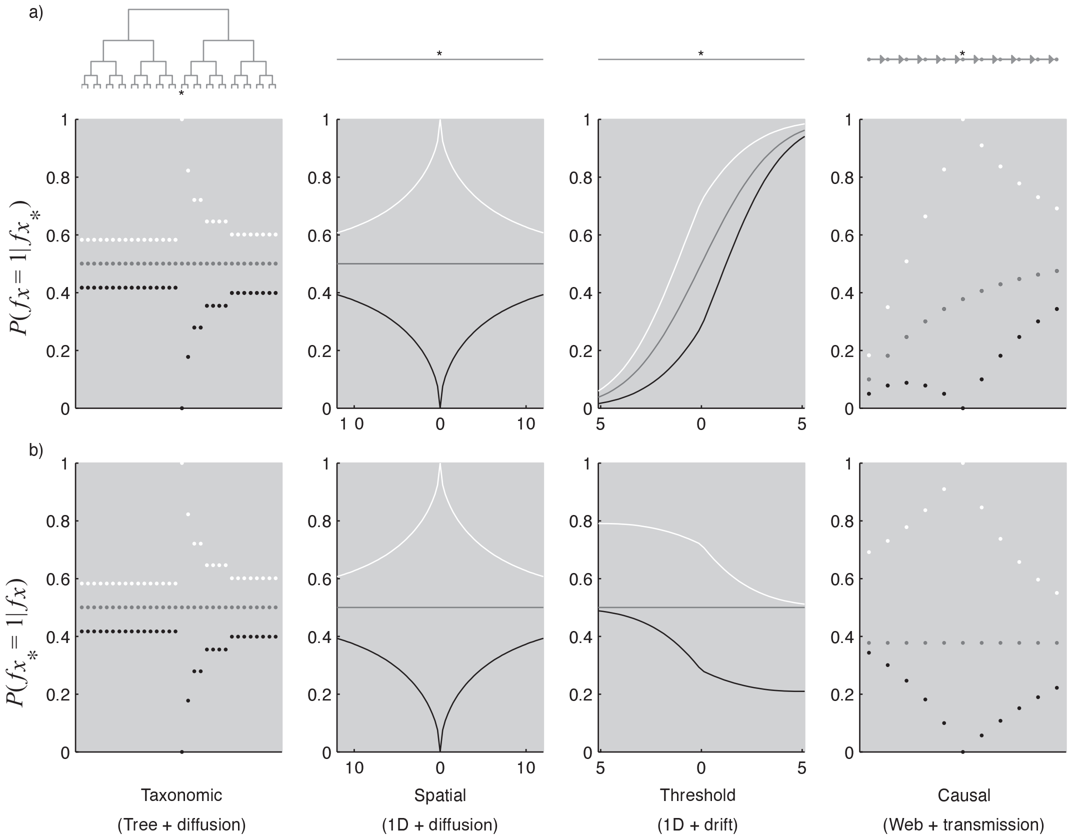

*Figure 6.* Generalization curves for four models: a diffusion process defined over a tree, a diffusion process defined over a one-dimensional space, a drift process defined over a one-dimensional space, and a transmission process defined over a food chain. (a): Projections *from* a category ($x_*$) located at the asterisk. The category is observed to have a novel property ($f_{x_*} = 1$, white curve) or observed not to have a novel property ($f_{x_*} = 0$, black curve), and predictions are made about categories located elsewhere in the structure. In the taxonomic case, observing that $x_*$ has a property supports the prediction that nearby categories in the tree will also have the property (white curve). The gray curve shows predictions before any positive or negative examples have been observed. (b): Projections *to* category $x_*$ after a single positive example (white curve) or a single negative example (black curve) is observed elsewhere in the structure. In the case of the threshold model, observing that a category at the far left of the dimension has a property (white curve) provides good evidence that $x_*$ will also have the property. The gray line shows the prior probability that category $x_*$ has the novel property.

first column shows inferences about 32 categories that lie at the leaves of a balanced binary tree. Before any observations have been made, the gray curve in Figure 6a shows that each category is predicted to have the novel property with probability 0.5. The white curve in Figure 6a shows predictions after the category at the asterisk (category $x_*$) is observed to have the property of interest. As expected, the model predictions fall off smoothly over the tree—the probability that the category closest to $x_*$ has the property is 0.8, and the probability falls below 0.6 for the categories most distant from $x_*$. The black curve shows predictions when the category at the asterisk is observed not to have the novel property, and again predictions vary smoothly over the tree. The curves in Figure 6b show inductive projections *to* the category marked with an asterisk. For instance, given that the leftmost category in the tree is known to have a property, the leftmost white point indicates the probability that category $x_*$ will also have this property. In the case of the taxonomic model, the curves in Figures 6a and Figure 6b are identical, indicating that one-premise arguments remain equally strong when the premise and conclusion are exchanged.

The diffusion process leads to several qualitative phenomena that are evident in Figure 6 and summarized in Table 1. Symmetry is the first of these phenomena—for all pairs of categories ($a$, $b$), the diffusion process predicts that $a \rightarrow b$ and $b \rightarrow a$ are equally strong. For example, a diffusion process over the tree in Figure 1 predicts that *cheetahs → monkeys (enzyme)* and *monkeys → cheetahs (enzyme)* will be equally strong. The diffusion process also leads to a distance effect—the strength of a one-premise argument decreases as the distance between premise and conclusion increases. For example, the taxonomic model in Figure 1 predicts that *cheetahs → lions (enzyme)* is stronger than *cheetahs → monkeys (enzyme).* Four versions of the distance effect are shown in Table 1, and the generalization curves for the taxonomic model (Figure 6) satisfy each of these conditions.

Table 1
*Qualitative Phenomena That Emerge From the Diffusion, Drift, and Transmission Processes*

| | a b c | Diffusion | Drift | Transmission |
|---|---|---|---|---|
| Symmetry | ($a \rightarrow b$, $b \rightarrow a$) | = | > | > |
| Distance | D1. ($a \rightarrow b$, $a \rightarrow c$) | > | < | > |
| | D2. ($b \rightarrow c$, $a \rightarrow c$) | > | < | > |
| | D3. ($c \rightarrow b$, $c \rightarrow a$) | > | > | > |
| | D4. ($b \rightarrow a$, $c \rightarrow a$) | > | > | > |

*Note.* Each process is defined over a linear structure, and $a$, $b$, and $c$ are categories located somewhere along the structure. Each row of the table represents a comparison between two arguments: Symmetry, for example, compares the strengths of $a \rightarrow b$ and $b \rightarrow a$. Of the three processes, only the diffusion process predicts that $a \rightarrow b$ and $b \rightarrow a$ are equal (=) in strength. The drift and transmission processes both predict that $a \rightarrow b$ is stronger (>) than $b \rightarrow a$. For the drift process, we have assumed that categories towards the right (e.g. $c$) are more likely a priori to have the novel property than categories towards the left (e.g., $a$). For the transmission process, we have assumed that $c$ is higher up in the food chain than $a$. The diffusion and transmission processes both predict a *distance effect*—inductive projections from one category to another decrease in strength as the categories become further from each other. For instance, both processes predict that $a \rightarrow b$ is stronger (>) than $a \rightarrow c$. The drift process predicts a distance effect only when the conclusion falls to the left of the premise. For arguments that move in the opposite direction, the drift process predicts an inverse distance effect: Inductive projections increase in strength as the distance between premise and conclusion increases. The phenomena predicted by each process are illustrated by the generalization curves in Figure 6.

The diffusion process is only one of several possible ways to capture the assumption that features vary smoothly over a tree, and we previously described an alternative stochastic process where binary features $f$ are generated directly over a tree (Kemp, Perfors, & Tenenbaum, 2004; Kemp & Tenenbaum, 2003; Tenenbaum et al., 2007). Our current approach is motivated by two considerations. First, the diffusion process allows a unified treatment of property induction where the taxonomic model takes its place among a family of alternatives that rely on Equation 7 but use different mean vectors and covariance matrices $K$. For example, we will compare our taxonomic model to a spatial model where $K$ is defined over a low-dimensional space, and a threshold model where the mean is no longer set to 0. The stochastic process described by Tenenbaum et al. (2007) does not provide this unifying perspective, and fails, for instance, to handle cases where the underlying structure is a multidimensional space rather than a tree. A second advantage of the diffusion process is that it leads to tractable algorithms for learning trees, multidimensional spaces and other structures (Kemp & Tenenbaum, 2008). Other stochastic processes including the alternative presented by Kemp et al. (2004) can lead to structure-learning models that should work in principle but may be difficult to implement in practice.

### *Property Induction Data*

Our taxonomic model depends critically on a tree structure $S$ that captures taxonomic relationships between biological species. Choosing this tree is an issue for us as modelers, because we cannot test our framework without it. Learning a taxonomic tree, however, is also a challenge for human learners, and formal approaches should aim to explain how people acquire trees and other representations that support inductive reasoning. As we describe in a later section, a tree structure can be learned by observing the features of many species and constructing a tree where nearby species tend to have many features in common. The tree in Figure 3 was learned from a feature matrix $D$ containing 50 species and 85 features. The data $D$ include anatomical and behavioral features (e.g., "is gray," "is active," "lives in water"), and were primarily collected by Osherson et al. (1991). A more detailed description of these data is provided in Appendix C.

After learning a tree (Figure 3a) from a set of observed features, we applied our taxonomic model to five data sets collected by previous researchers. The first two data sets are described by Osherson et al. (1990) and use a domain with 10 species: horse, cow, chimp, gorilla, mouse, squirrel, dolphin, seal and rhino (the subtree of Figure 3a that includes only these ten animals is shown in Figure 5a). The *Osherson horse* set contains 36 two-premise arguments, and the conclusion category is always "horses." One sample argument is *cows, chimps → horses (biotin),* where the property indicates whether a species "requires biotin for hemoglobin synthesis." The *Osherson mammals* set contains 45 three-premise arguments, and the conclusion category in each case is mammals (a typical conclusion might state that "all mammals require biotin for hemoglobin synthesis"). The remaining three data sets are described by Smith, Lopez, and Osherson (1992).

*Smith fox* and *Smith elephant* use eight animal species and contain 15 two-premise arguments, and the conclusion category is either fox or elephant. *Smith mammals* includes 15 two-premise arguments that use 6 animal species, and the conclusion category is always mammals. To model the Smith data, we substituted collie for dog, which does not appear in Figure 3.

For the first two data sets (*Osherson horse* and *Osherson mammals*), participants were given a card-sorting task where they ranked all arguments in a given set in order of increasing strength. We model these data by attempting to predict the average rank order of each argument. For each of the remaining three data sets, participants assessed each argument by rating the probability of the conclusion given the premises. Our models attempt to predict the mean probability rating for each argument.

Given the tree in Figure 3a, the diffusion process generates a prior that can be used for property induction. We compute the predictions of our taxonomic model by drawing a large number of features from the prior and using this sample to approximate the true prior (see Appendix B for additional details). The performance of the model is shown in the top row of Figure 7. Each subplot compares human rankings to model predictions for one of the five data sets, and the top row indicates that the taxonomic model achieves high correlations across all five data sets. The model, however, includes several components, and it is not yet clear which of them are important—maybe the tree structure is doing most of the work, or maybe the diffusion process or the Bayesian inference engine is the critical factor. We can show that all three components are critical by comparing the taxonomic model to several closely related alternatives (Table 2).

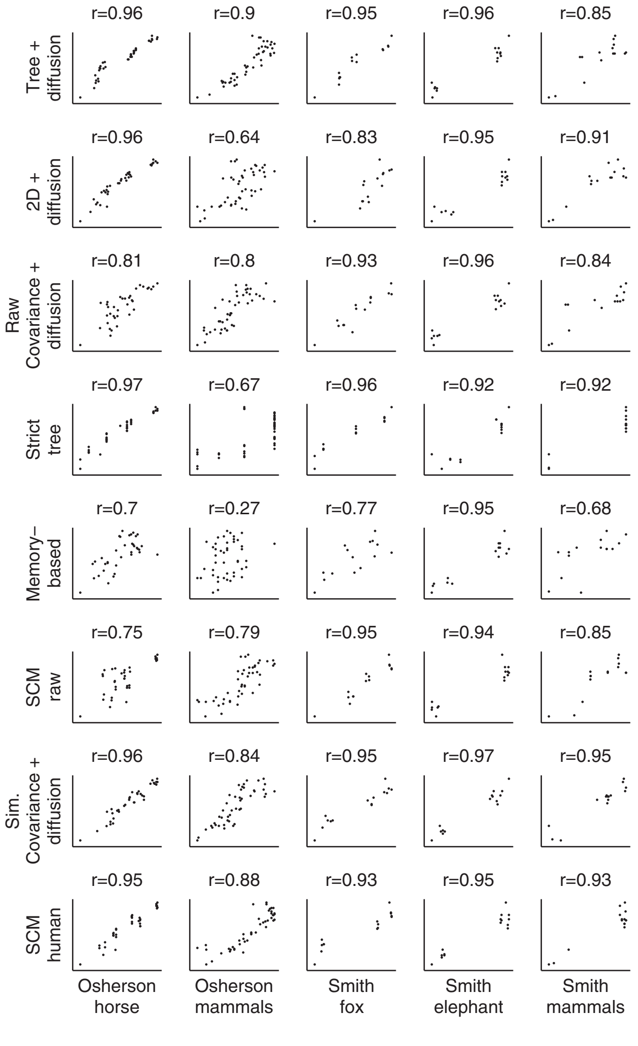

*Figure 7.* Model performance for the default biological context. Each subplot shows human ratings of argument strength (*y*-axis) against model predictions (*x*-axis), and each data point represents a single argument. For instance, the data point at the bottom left of the bottom left plot represents an argument (*dolphins, seals → horses* (*enzyme*)) considered weak by both participants (*y*-axis) and the similarity-coverage model (SCM; *x*-axis). The eight models are summarized in Table 2.

To establish whether the tree structure is important, we can compare our taxonomic model to an alternative that uses a diffusion process defined over a low-dimensional space (Figure 3b). The prior distribution for this model is generated by Equations 7–8, except that the covariance matrix $K$ is now defined over a two-dimensional space instead of a tree. This spatial model is closely related to one of the first formal approaches to biological property induction (Rips, 1975), which assumes that animal species are represented as points in a multidimensional space. Given a single example of a species with a novel property, Rips (1975) suggests that the property should be projected most strongly to species nearby in the space. When applied to the five data sets already described (Figure 7), the spatial model performs well in all cases except one. A model will only account well for the "Osherson mammals" data if it can capture the diversity effect, and the results in Figure 7 suggest that the spatial model struggles to capture this effect.

To understand the difference between the taxonomic and spatial models, it is useful to compare the covariance matrices used by these models. These matrices are shown in Figures 5c and 5d, and the structures used to define these covariance matrices appear in Figures 5a and 5b. The 10 species that appear in these structures fall naturally into four groups: ungulates (horse, cow, elephant, and rhino), primates (chimp and gorilla), rodents (mouse and squirrel), and marine mammals (dolphin and seal), and the human similarity ratings in Figure 5 provide evidence for the psychological reality of these groups. A tree structure (Figure 5b) can represent a situation where one of the groups (the marine mammals) is far and about equally distant from the rest, and the remaining three groups are about equally distant from each other. No two-dimensional representation can capture this configuration—one of the three remaining groups will have to fall between the other two (c.f. Tversky and Hutchinson, 1986). In Figure 5b, the primates lie between the rodents and the ungulates, and one consequence of this arrangement is that the spatial model does not account well for the diversity effect (Figure 7). In particular, it predicts that arguments like *squirrel, gorilla, seal → all mammals* are relatively weak, because these three species are fairly close to one another in Figure 5b. The difference in performance between the taxonomic model and the spatial model confirms that qualitatively different representations can lead to different inductive inferences. Spatial models that use representations with three or more dimensions may provide a better account of the data, but any spatial representation imposes constraints (including the triangle inequality and certain kinds of nearest neighbor constraints; Tversky & Hutchinson, 1986) that will be appropriate for some problems only. Ge-

Table 2
*Models for the Default Biological Context*

| Model | Bayesian | Description |
|---|---|---|
| Taxonomic (tree + diffusion) | Yes | Prior defined using smoothness over a tree |
| Spatial (2D + diffusion) | Yes | Prior defined using smoothness over a 2D space |
| Raw covariance + diffusion | Yes | Prior defined using the raw covariance matrix $1/85DD^{\mathrm{T}}$ |
| Strict tree | Yes | Uniform prior on all subtrees of a tree structure |
| Memory-based | Yes | Uniform prior on all previously observed features |
| SCM raw | No | The similarity coverage model (Osherson et al., 1990), where the similarity matrix is set to $1/85DD^{\mathrm{T}}$ |
| Sim covariance + diffusion | Yes | Prior defined using human similarity ratings |
| SCM human | No | The similarity coverage model (Osherson et al., 1990), where the similarity matrix is set by human ratings |

*Note.* All of the Bayesian models use the same inference engine, and only the priors are summarized. The first six models rely on $D$, a matrix of 50 species by 85 features. For instance, the taxonomic model uses a tree learned from $D$, the spatial model uses a two dimensional space learned from $D$, and the raw covariance model uses the features in $D$ more directly. The final two models use human similarity judgments instead of $D$.

neric representations like multidimensional spaces are used by many modelers, but will only be suitable in some inductive contexts.

The representations used by our taxonomic and spatial models are both learned from the animal-feature matrix $D$ described previously. The *raw covariance* of matrix $D$ is a 50 by 50 matrix where entry $(i, j)$ is high if species $i$ and $j$ have similar features. This matrix is defined as $1/85DD^{\mathrm{T}}$ and is described in more detail in Appendix A. Figure 5e shows a 10-species subset of this matrix—note, for instance, that the entry for chimps and gorillas is high, indicating that these species share many features in data set $D$. The taxonomic and spatial models can both be viewed as models that convert the raw covariance $<1/85DD^{\mathrm{T}}$ into a covariance matrix that satisfies certain constraints. The taxonomic model uses a covariance matrix defined over a tree, which in turn can be discovered from the raw covariance matrix. The taxonomic model therefore attempts to replace the raw covariance matrix with the best approximation that is consistent with a tree. Similarly, the spatial model replaces the raw covariance matrix with the best approximation that is consistent with a two-dimensional spatial representation. It is natural to compare the taxonomic and spatial models with an unstructured model that skips the structure learning step and uses the raw covariance matrix directly for property induction: in other words, a model where matrix $K$ in Equation 7 is set to the raw covariance matrix (Figure 5e). We call this model the raw covariance model, and Figure 7 shows that it accounts fairly well for human inductive judgments, if not quite as well as the taxonomic model.

To understand the performance of the raw covariance model, it is instructive to study the covariance matrix that it uses. The raw covariance matrix (Figure 5e) suggests that one of the four groups—the ungulates—is not immediately evident from the raw data. In particular, note that the covariance between horse and rhino is relatively low. As a result, the raw covariance model makes inaccurate predictions about some of the arguments involving ungulates (Figure 7): for example, it gives *gorilla, rhino* → *horse* a weaker rating than people do.

Structured models are useful in part because they provide an inductive bias (Geman, Bienenstock, & Doursat, 1992) that is needed when learning from data that are sparse or noisy or both. Our tree learning algorithm does its best to find a tree-based covariance matrix that matches the raw covariance as closely possible, but the constraint that the underlying structure must be a tree prevents it from fitting the raw covariance perfectly. In other words, the tree constraint forces the structure learning algorithm to regularize or clean up the data. The tree-structured covariance in Figure 5c shows that the group of ungulates can be extracted from the raw data by a model that is required to build a tree—note in particular that the covariance between horse and rhino is relatively high in Figure 5c. An unstructured approach like the raw covariance model has no comparable way to regularize the data, and will suffer because it is unduly influenced by the noise in the raw data. Although the raw covariance model is unlikely to perform well when the data are noisy, it may produce good results when the data are very clean. The second-last row of Figure 7 shows that the performance of this model improves when the raw covariance matrix is replaced with human similarity ratings, which provide a cleaner measure of the relationships between the ten species.

Having the right structure is critical, but choosing an appropriate stochastic process is also important. The diffusion process captures the idea that features tend to be clean over the tree, but the same idea can be enforced more rigidly by assuming that each novel property corresponds to a subtree of the underlying tree (Figure 3a). By snipping any edge in the tree, we can create two groups of species, and we can define a feature that assigns value 1 to the species in one group and value 0 to the species in the second group. Any feature created in this way can be called a *subtree feature.* The *strict tree* model uses a stochastic process that assigns fixed prior probability to all subtree features and zero probability to all remaining features. In contrast, the prior for our taxonomic model allows for noise and exceptions and assigns a nonzero prior probability to every possible feature.

The strict tree model performs well on the specific data sets but does not account well for the general arguments (Figure 7). The prior is too rigid to capture the diversity effect in full. For example, the premise sets {*gorilla, mouse, seal*} and {*horse, dolphin, seal*} are treated similarly by the strict model because each is compatible with exactly 10 features. Our tree model distinguishes between these cases, recognizing that the first premise set is better spread out over the tree and therefore provides better evidence that all mammals have the novel property.

The memory-based model in the fifth row of Figure 7 is inspired by Heit's suggestion that a prior for Bayesian inference can be generated by retrieving properties from memory (Heit, 1998). We take the feature matrix $D$ as an approximation of the features that may be retrieved from memory and add an additional feature that assigns value 1 to every animal in the domain. We place a uniform prior over all features in this collection and set $p(f) = 0$ for all features that do not appear in the collection. Because this model requires binary features and the features in $D$ are continuous, we convert them to binary features by thresholding at some free parameter $\theta$. We set this parameter to the value that maximizes the average correlation across all the data sets in Figure 7.

A serious limitation of the memory-based model is that it cannot deal well with sets of premises that are incompatible with any of the features it has previously seen. One symptom of this problem is that the model cannot adequately capture the diversity effect. Consider the premise sets {*horse, dolphin, squirrel*} and {*horse, dolphin, seal*}. It is difficult to think of anatomical properties that apply to all of the animals in the second set, but only some of the animals in the first set. The memory-based model therefore finds it hard to discriminate between these cases, even though the first set seems to provide better intuitive support for the conclusion that all mammals have the novel property.

The sixth row shows the performance of the SCM, which is probably the best-known model that can handle the default biological context. The matrix of pairwise similarities required by this model is defined as the raw covariance matrix $1/85DD^{\mathrm{T}}$: In other words, the similarity of two species is set to the inner product of their feature vectors. The SCM has a single free parameter, which was set to the value that maximizes the average correlation across all five data sets in Figure 7. Note that *SCM raw* uses precisely the same information as the first five models in Figure 7. All of these approaches take the feature matrix $D$ as input, and we see that the taxonomic model performs better than the rest. As originally presented by Osherson et al. (1990), the SCM takes human similarity ratings rather than features as its input, and the final row in Figure 7 shows that the performance of the SCM improves when the raw covariance matrix (Figure 5e) is replaced with a matrix of human similarity ratings (Figure 5f). Models based directly on human similarity ratings may lead to accurate predictions, but the success of these models depends critically on the knowledge captured by the similarity ratings, and the nature and acquisition of this knowledge are usually left unexplained.

The complete pattern of results in Figure 7 suggests that human inferences about biological properties can be accurately captured by combining three elements: Bayesian inference, a structure that captures relationships between categories, and a stochastic process defined over that structure. Any one or two of these elements will not be sufficient—we showed, for instance, that our taxonomic model performs better than alternative Bayesian approaches (e.g., the memory-based model), alternative tree-based models (e.g., the strict tree model), and alternative models that rely on the diffusion process (e.g., the spatial model). These results support our general claim that structured knowledge and statistical inference are both important and that structured statistical models help to explain how inductive inferences are guided by background knowledge.

## Spatial Reasoning

Although a tree may be the structure that best captures the knowledge used in the default biological context, there is a long tradition of using Euclidean spaces to build semantic representations. Techniques like multidimensional scaling (Shepard, 1980) and principal components analysis have been used to construct spatial representations for many different domains, and these representations are regularly used to model categorization and other forms of inductive reasoning (Ashby, 1992; Kruschke, 1992; Nosofsky, 1986). Formal models that use multidimensional spaces are sometimes contrasted with more structured approaches. We believe it is more useful to think of a multidimensional space as a structure in its own right, and we view spatial models as simple cases of our structured approach. Choosing a spatial representation imposes structural constraints (Tversky & Hutchinson, 1986): one of the best-known is the triangle inequality, which constrains the pairwise distances that can exist between any three points.

We develop a spatial model of property induction by taking the approach of the previous section and replacing the tree structure with a two-dimensional space. Again we use a diffusion process to generate a prior distribution over the structure. The process will assign high probabilities only to features that are smooth over the space. As before, Equations 7–8 are used to generate a prior distribution for Bayesian inference, but now the covariance matrix $K$ is defined using distances in a two-dimensional space:

$$K_{ij} = \frac{1}{2\pi}\exp\left(-\frac{1}{\sigma}\|x_i - x_j\|\right) \tag{9}$$

where $x_i$ is the location of category $i$ in the two-dimensional space and $\|x_i - x_j\|$ is the distance between $x_i$ and $x_j$. As mentioned already, the free parameter $\sigma$ is set to 5 throughout this article.

The model that combines Equations 7, 8, and 9 can be called the *2D + diffusion* model, but we will also refer to it as the spatial model. Many previous models are based on the intuition that features or concepts tend to be smooth over a multidimensional space (Ashby, 1992; Kruschke, 1992; Nosofsky, 1986; Shepard, 1987; Tenenbaum & Griffiths, 2001). Some of these models assume that each feature corresponds to a *consequential region:* a connected, convex region of a multidimensional space (Shepard, 1987). Our model allows each feature to correspond to several connected regions of a multidimensional space, and these regions are not required to be convex. The features with highest prior probability, however, are the features that are smoothest, and smooth features will tend to correspond to a small number of convex regions. In a two-dimensional space, for example, the curve that encloses a given area using the shortest possible boundary is a circle and it follows that the smoothest feature of a given size corresponds to a disk, or the interior of a circle.

The second column of Figure 6 shows the generalization curves that result when our spatial model is presented with information about a novel property. For simplicity, the figure shows predictions about categories that lie along a continuous one-dimensional structure. Before any information has been provided, the gray curve shows that categories at any point along the structure are predicted to have the novel property with probability 0.5. The white curve shows generalizations after the model has been informed that category $x_*$ (a category located at the asterisk) has the property of interest. The decay in generalization is approximately exponential, and we see again that the diffusion process leads to a distance effect (Table 1). The black curve shows generalizations when the model is informed that a category at the asterisk does not have the novel property. Again the gradient of generalization is approximately exponential.

We later describe how our spatial model can be used to learn low-dimensional representations given a matrix of features. When applied to the animal-feature matrix $D$ already described, the model discovers the two-dimensional space in Figure 3b. The second row of Figure 7 shows the performance of a Bayesian model with a prior $p(f)$ defined by a diffusion process over this two-dimensional space. A comparison with the top row in Figure 7 suggests that inductive reasoning in the default biological context is better explained using a tree structure than a spatial structure.

The best structure for a given problem will depend on the inductive context. A taxonomic tree may be the default representation of relationships among biological species, but Figure 1 suggests that the knowledge relevant to some biological properties is better captured by other structures, including low-dimensional spaces and food webs. In particular, our spatial model may be appropriate for reasoning about biological properties that apply to all animals that fall within some interval on an underlying dimension, such as weight or size.

In domains other than biology, the default representation may be a multidimensional space rather than a tree structure. To demonstrate that different kinds of structures are needed to generate appropriate priors for different inductive contexts, we asked participants to reason about geographically relevant properties of American cities. We expected that the spatial model would account better than the taxonomic model for a task that draws on geographic knowledge. Previous authors have found that qualitatively different representations are needed to account for proximity data, including judgments of similarity. Pruzansky et al. (1982) suggest, for example, that spatial representations are better than trees at capturing proximity relationships between perceptual stimuli, such as colors and sounds, but that the reverse conclusion holds for conceptual stimuli, such as animals, vehicles, and fruits. Here we focus on inductive judgments rather than proximity data, but explore the same general idea that different representations are needed for different problems.

### *Experiment: Spatial Reasoning*

In the two experiments described previously, Osherson et al. (1990) asked participants to reason about properties such as “has enzyme X132.” Participants may have expected these blank properties to be smoothly distributed over the familiar taxonomic hierarchy of animal species but probably drew on little other background knowledge. To explore a setting where the default representation may be a low-dimensional space rather than a tree, we designed a similar task involving blank geographical properties. Our hope was that most participants would expect these properties to be smoothly distributed over the familiar spatial configuration of American cities but would otherwise rely on little background knowledge.

*Participants.* Twenty members of the Massachusetts Institute of Technology community participated for pay.

*Stimuli.* The instructions stated that “Native American artifacts can be found under virtually all large American cities” and that “some kinds of artifacts are found under many cities, and others are found under just one city or a handful of cities.” Participants were then asked to make inferences about the distribution of artifacts among the nine cities shown in Figure 8.

Three sets of arguments were presented in total. The first set included all two-premise arguments with “Minneapolis” as the conclusion city. There were 28 arguments in total—one, for instance, was the argument *Seattle, Boston → Minneapolis (artifacts),* where the property stated that “artifacts of Type X” are found under a given city. The second set was similar to the first, except that Houston was always the conclusion city. The third set included 30 three-premise arguments where each premise mentioned one of the nine cities and the conclusion stated that “artifacts of type X are found under all large American cities.” Each argument in the experiment was printed on a card with a line separating the premises from the conclusion.

*Procedure.* The arguments in each set were presented as a stack of shuffled cards. Participants were encouraged to lay the cards out on a large table and to move them around before creating their final sort. The experiment was intended to be as close as possible to the experiments responsible for the biological data modeled in the previous section (Osherson et al., 1990).

After the three sorting tasks, participants drew a map showing the nine cities used in the experiment. Each participant placed nine dots on a sheet of graph paper. Participants were told that the absolute positions of the dots were not important, but were asked to ensure that the relative distances between dots matched the relative distances between cities as closely as possible.

*Results.* For each map, we computed distances between all pairs of cities. These distances were scaled so that the largest distance was 1, and these scaled distances were subtracted from 1 to create a measure of spatial similarity for each participant. We averaged these similarity matrices across participants to create a single similarity matrix that could be used for structure discovery.

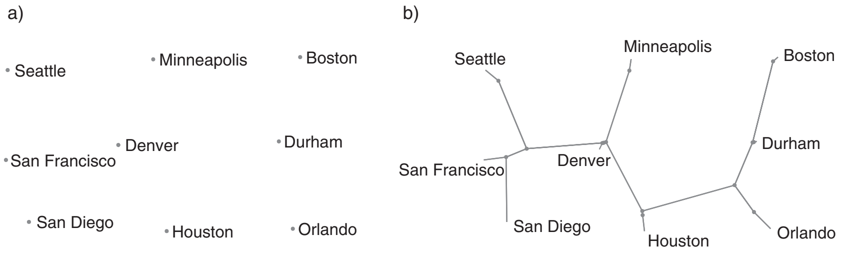

*Figure 8.* A spatial representation (a) and a tree (b) learned from data collected during the map drawing component of the artifacts task. The tree has been laid out so that the node positions are the same in (a) and (b).

Using the structure learning methods described in Appendix C, we found the tree and the two-dimensional space that best account for the matrix of averaged similarities (Figure 8). It is encouraging that the best spatial structure is roughly consistent with a map of the United States. We defined covariance matrices $K$ over both structures (see Appendix B), and combined these matrices with Equations 7–8 to generate prior distributions for both models.

Model predictions for the two structures are shown in Figure 9. For each of the three data sets, we attempt to predict the average rank order of each argument. As expected, the spatial model outperforms the taxonomic model on all three data sets. The mediocre performance of the taxonomic model can be understood by studying the tree in Figure 8b. For example, Minneapolis is far from Boston in the tree, but not so far away in reality. This problem is more than just a limitation of the specific tree found by our structure learning algorithm. No tree can adequately represent the relative distances between points embedded in a two-dimensional space.

Comparing these results with the results in the previous section, we see a double dissociation between models and inductive contexts. The taxonomic model performs well, but only in the biological context, and the two-dimensional model performs well, but only in the spatial context. This double dissociation suggests that different inductive contexts will require different kinds of prior distributions and that appropriately structured priors can allow our Bayesian framework to account for the context-dependence of human reasoning.

The final row in Figure 9 shows the performance of the SCM when supplied with the similarity matrix used to learn the structures in Figure 8. The model performs especially poorly on the all-cities data, and this result is due in part to the model's notion of coverage. Arguments are considered strong when the three premises cover the continent well: in other words, when all remaining cities are close to at least one of the three premise cities. Good coverage can be achieved in two very different ways—we can choose cities on the periphery so that each internal city is close to one of the premises, or we can choose more centrally located cities so that each city on the periphery is close to one of the premises. The first case appears to provide stronger evidence that a property is true of all cities, but the SCM does not distinguish between these two cases. Unlike the two-dimensional model, for instance, the SCM rates *Denver, Durham, Seattle → all cities (artifacts)* as stronger than *Boston, Orlando, San Francisco → all cities (artifacts).*

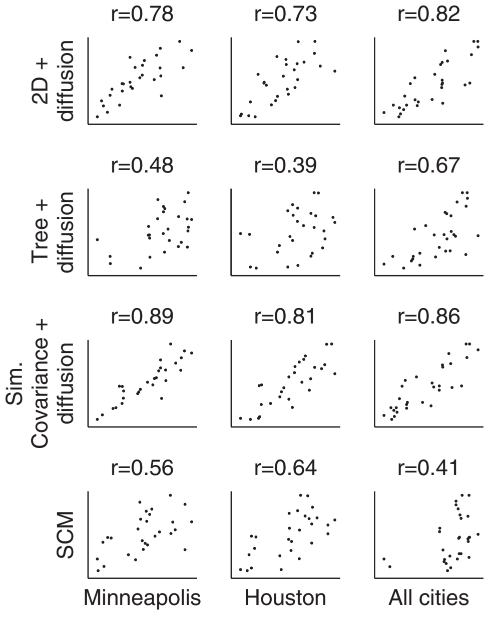

*Figure 9.* Model performance for the artifacts task. Each subplot shows human ratings of argument strength ($y$-axis) against model predictions ($x$-axis), and each data point represents a single argument.

The relatively poor performance of the SCM suggests that this model can yield inaccurate predictions even when supplied with a similarity matrix that captures all of the relevant background knowledge. To demonstrate that the information in the similarity matrix is sufficient to account for the human data, we implemented a model that uses this matrix as the covariance matrix for the diffusion process (Equations 7–8). This *similarity covariance* model performs well on all three data sets and suggests that the poor performance of the SCM can be attributed to the model and its notion of coverage rather than the similarity data supplied to the model.

Figure 7 shows that the similarity covariance model also accounts for inferences about biological properties when supplied with human ratings of the similarity between the species in Figure 5. Note, however, that this model relies on a notion of similarity that is highly context specific. In the artifacts task, the similarity matrix used by the model is derived from the maps that participants drew, and a more generic measure of similarity is likely to yield less accurate predictions about geographic properties. For instance, San Diego is much closer to Denver than Orlando, but is more similar to Orlando in many other respects.

Although the taxonomic and spatial models rely on different structures, both structures can be viewed as representations of similarity. The diffusion process used by both models is also consistent with a similarity-based approach, because it assumes that nearby categories in a structure (i.e., similar categories) are likely to have many properties in common. Similarity-based reasoning is a rich topic for psychological investigation, but many researchers have described inductive inferences that cannot be explained by similarity alone (S. A. Gelman & Markman, 1986; Heit & Rubinstein, 1994; Shafto & Coley, 2003; Smith et al., 1993). Our formal framework has room for models that go beyond similarity-based reasoning, and our final two models (the threshold and causal models in Figure 1) are two examples.

## Threshold Reasoning

The experiments described so far have relied on blank biological properties and blank geographic properties. Properties like these carry enough information to establish an inductive context but are deliberately chosen to minimize the role of prior knowledge. Most people, for instance, can draw on little specific knowledge when asked to decide whether squirrels or elephants are more likely to have enzyme X132 and may therefore rely on generic knowledge about taxonomic similarity. People do, however, have strong intuitions about many other kinds of properties—consider again the comparison between squirrels and elephants when "has enzyme

X132" is replaced with "has skin that is more resistant to penetration than most synthetic fibers." This section develops a model for reasoning about one class of nonblank properties: properties that correspond to thresholds along familiar dimensions, such as skin toughness. Note that inferences about these properties cannot be explained by similarity alone. For example, *squirrels* → *elephants (tough skin)* is stronger than *rhinos* → *elephants (tough skin),* even though elephants are more similar to rhinos than squirrels.

Like the taxonomic and spatial models, our threshold model uses a prior $p(f)$ induced by a stochastic process defined over a structure. The structure is a one-dimensional space that corresponds to a familiar dimension. For the sake of example, suppose that the underlying dimension is skin toughness (Figure 10a), and that skin toughness can be measured on a scale from 0 to 100. The property of interest should correspond roughly to a threshold along this dimension—all species with skin toughness greater than some value should have "skin that is more resistant to penetration than most synthetic fibers." To capture this intuition, we introduce a stochastic process called the *drift process* that tends to generate features shared by all categories beyond some point in the underlying dimension.[1] Note that the diffusion process will not be suitable for reasoning about threshold properties. Given a linear structure, the diffusion process will tend to generate features that correspond to connected regions of the structure. For instance, a feature that applies only to species with skin toughness between 30 and 60 will receive a relatively high prior under this process.

As before, we specify a prior $p(f)$ by describing a recipe for generating feature vectors $f$, and the prior probability of any feature $f$ is its probability of being generated by this recipe. Figure 4a shows one feature $f$ generated over a linear graph. As for the diffusion process, we start by generating a continuous feature $y$ over the graph. This time, however, feature $y$ is expected to be low towards the left of the graph and high towards the right of the graph. We then pass the continuous feature $y$ through a sigmoid function $sigm(\cdot)$ which maps the real numbers onto the interval (0, 1). Finally, we generate a binary value $f_i$ for each category $i$ by tossing a coin with bias $sigm(y_i)$.

As for the diffusion process, each continuous feature $y$ is generated from a Gaussian distribution with a covariance matrix $K$ that encourages it to be smooth over the underlying graph. Now, however, the mean of the distribution is no longer zero but is set to a vector $\mu$ where $\mu_i$ is the location of category $i$ along the underlying dimension. For example, $\mu_i$ will be high for categories towards the right of the graph in Figure 4b, and these categories will tend to end up with high values of $y_i$. The complete generative model is therefore

$$y \sim N(\mu, K)$$

$$f_i \sim \text{Bernoulli}(sigm(y_i)) \qquad (10)$$

where $K$ is a covariance matrix defined over structure $S$ as for the taxonomic and spatial models, and where the second line indicates that $f_i$ is generated by tossing a coin with bias $sigm(y_i)$. As for the taxonomic and spatial models, the definition of $K$ uses a free parameter $\sigma$ which is set to 5 (see Appendix B).

The drift process is particularly natural when $S$ is a one-dimensional structure, because the $\mu$ in Equation 10 can be defined by the locations of the categories along the structure. The mean $\mu$, however, may be specified independently of the structure $S$, which allows Equation 10 to be applied to many different representations, including trees, multidimensional spaces, and arbitrary graphs. To determine the mean $\mu$, we need priors that specify the probability that the novel property applies to each category. For example, if the probability that the $i$th category has the property is 0.9, we can capture this knowledge by setting $\mu_i = sigm^{-1}(0.9)$. Previous models for reasoning about threshold properties (Blok et al., 2007; Smith et al., 1993) have also relied on prior probabilities, and the data sets we will consider include human judgments of these probabilities.

The model that combines the drift process with a one-dimensional structure can be called the *1D + drift* model, but we will also refer to it as the threshold model. Generalization curves for this model are shown in the third column of Figure 6. Before any information has been provided, the gray curve in Figure 6 is a soft threshold that shows that categories with high values on the underlying dimension are predicted to have the novel property. The white curve shows that the soft threshold is shifted to the left when the model observes that a category located at the asterisk (category $x_*$) has the property. As the white curve suggests, the threshold model typically predicts a distance effect only for inferences that move from right to left (from a probable premise to a less probable conclusion). Inferences that move in the opposite direction lead to an inverse distance effect—arguments increase in strength as the distance between premise and conclusion increases (Table 1).[2] Note that the white curve does not attain a probability of 1 at $x = 0$. The probability that the observed category has the property will be 1, but we allow multiple categories to be located at each point in the structure, and the value of the generalization curve at position zero is the probability that a *second* category located at the asterisk will have the novel property.

Unlike the diffusion process, the drift process does not lead to symmetric predictions about one-premise arguments, and the curves in Figure 6b are different from the curves in Figure 6a. The white curve shows that the best evidence that category $x_*$ has the property is provided by a positive example located at the far left of the structure. Intuitively, learning that a very unlikely category has the property provides strong evidence that the property is a lot more common than previously expected. The white curve shows once again that the threshold model typically predicts a distance effect only for inferences that move from right to left. Comparing the second and third columns of Figures 6a and 6b shows that the drift process and the diffusion process have very different inductive consequences, even though the equations for these processes are very similar.

## *Property Induction Data*

We applied the threshold model to four data sets collected by previous researchers. The first two are described by Blok et al. (2007) and involve judgments about the salaries earned by grad-

[1] The drift process is closely related to Brownian motion with drift (Karlin & Taylor, 1975).

[2] The profile of the drift process in Table 1 does not hold in all cases, but simulations suggest that it is accurate when $p(a)$ and $p(c)$ are not too extreme—between 0.05 and 0.95, say.

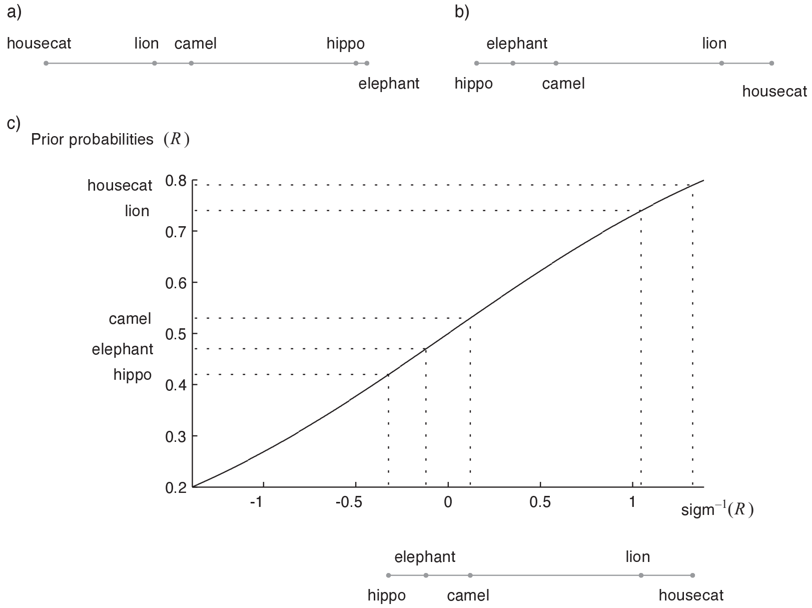

*Figure 10.* One-dimensional structures for reasoning about two threshold properties: (a) "has skin that is more resistant to penetration than most synthetic fibers" and (b) "has a visual system that fully adapts to darkness in less than 5 minutes." Animals towards the right of each structure are more likely to have the property. The two structures were determined by human ratings of the prior probability that the property applies to each animal. The prior probability for each species is passed through the sigmoid function *sigm*(·) to generate the position of that species along the one-dimensional structure. Shown in (c) are the prior probabilities for the "darkness" property ($y$-axis) and the corresponding positions along the one-dimensional structure ($x$-axis).

uates from several institutions. *Five colleges* includes five institutions—Connecticut State, Oklahoma State, Harvard, Arkansas State, and Yale—and the property used is "over 60% of its graduates earn more than $50,000 a year at their first job." *Four colleges* is similar, and the most notable difference between the two is that four colleges includes arguments with negative premises. The remaining data sets were collected by Smith et al. (1993) and use five animal species: housecats, lions, camels, hippos, and elephants. *Smith dark* uses the property "has a visual system that fully adapts to darkness in less than 5 minutes," and *Smith skin* uses "has skin that is more resistant to penetration than most synthetic fibers."

Each data set includes ratings of argument strength, similarity ratings between all pairs of categories, and $R$, a vector that specifies the prior probability that the property of interest applies to each category. For instance, the Smith dark data indicate that $R_{\text{housecats}} = 0.79$ where $R_{\text{housecats}}$ is the mean judgment of the probability that housecats have a visual system that adapts to darkness in less than 5 min (Figure 10c). Our threshold model uses a one-dimensional structure, where the prior probabilities $R$ determine the location of each category (Figure 10). For the Smith dark data, housecats are located at $sigm^{-1}(R_{\text{housecats}}) \approx 1.32$ on a continuous scale where categories with prior probability 0.5 are located at position 0 (Figure 10c).

The top row of Figure 11 shows that the threshold model accounts well for all of the data sets. To establish that each component of the threshold model makes an important contribution, we compare this model to several closely related alternatives (Table 3). The spatial model is based on the same one-dimensional structure as the threshold model but uses the diffusion process instead of the drift process. The diffusion process led to accurate predictions for the default biological context (Figure 7) and for the spatial context (Figure 9), but here we see that a diffusion-based model performs dramatically worse than a drift-based model. We therefore conclude that having the right kind of structure is not enough to account for human judgments and that knowing how properties are generated over this structure is crucial.

The remaining models in Figure 11 all use the similarity ratings in some form. Both tree models use a covariance matrix $K$ defined over a tree learned from similarity data. Both of these models therefore assume that the novel property is distributed smoothly over the tree. The *tree + drift* model, however, assumes that the property is more likely to be found in certain regions of the tree—regions associated with high prior probabilities. To formalize this assumption, we set $\mu = sigm^{-1}(R)$ in Equation 10. Figure 11 shows that the drift process leads to better performance than the diffusion process regardless of whether the underlying structure is a one-dimensional representation or a tree.

The two similarity covariance models both use the similarity ratings as their covariance matrix $K$. Again we see that using the drift process instead of the diffusion process leads to a substantial improvement. Finally, the results for SimProb shows that it performs similarly to the threshold model on three of the four data sets and a little better on the Smith dark data. There is little reason to choose between the threshold model, *tree + drift* and SimProb on grounds of performance, but note that *tree + drift* and SimProb both rely on more information than the threshold model, which does not use the similarity information. For the tasks considered in

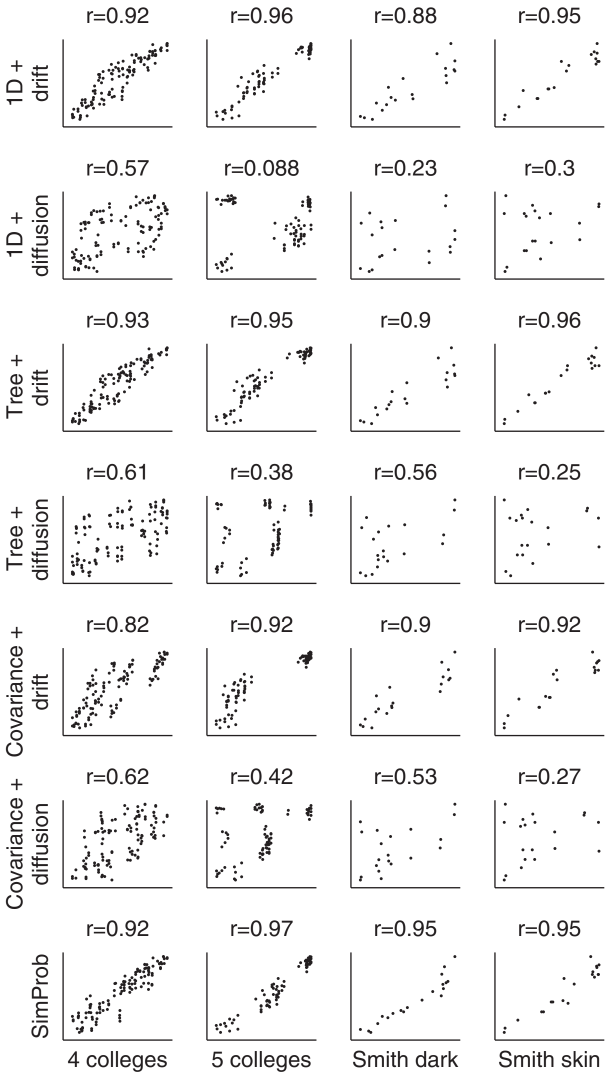

*Figure 11.* Model performance for four tasks involving judgments about threshold categories. The seven models are summarized in Table 3.

Figure 11, there is no good evidence that similarity ratings contribute any information that is not already present in the prior probabilities *R,* and the threshold model may therefore be preferred to the other two models. Compared to SimProb, the threshold model may also be preferred because it is an instance of the general framework developed in this article—a probabilistic framework that uses the same inference engine to handle many different kinds of reasoning.

Because each of the data sets in Figure 11 includes single premise arguments, we can explore whether the qualitative phenomena in Table 1 are present in the human data. Figure 12 shows that asymmetry is clearly present—for all pairs $(a,b)$ where $p(a) > p(b)$, participants report that $a \rightarrow b$ is stronger than $b \rightarrow a$. The second row shows that the results for the college data closely match the predictions of the threshold model—increased proximity leads to increased argument strength only for inferences that move from a probable premise to a less probable conclusion. The results for the Smith data match the predictions of the threshold model for all distance comparisons except comparison D2. Taken together, the results in Figure 12 provide further evidence that the threshold model accounts for the human data better than the spatial model.

## Causal Reasoning

Our threshold model captures one kind of reasoning that cannot be explained by similarity alone, but many other examples can be found in the literature (S.A. Gelman & Markman, 1986; Heit & Rubinstein, 1994; Medin et al., 2005; Shafto & Coley, 2003). One class of examples focuses on causal relations: for example, *gazelles* → *lions* (*babesiosis*) is stronger than *lions* → *gazelles* (*babesiosis*), where babesiosis is an infectious disease. Here we demonstrate how inferences like these can be captured by a causal model (Shafto et al., 2008) that formalizes a simple theory of disease transmission over a food web.

Like all of our models, the causal model relies on a structure and a stochastic process. The structure captures knowledge about predator–prey relationships among a group of species. This knowledge can be represented as a food web or a directed graph with an edge from *B* to *A* if *B* is eaten by *A.* The stochastic process captures knowledge about how diseases are transmitted over a food web. In particular, this transmission process captures the common sense idea that diseases tend to spread up a food web, and that a prey animal is more likely to transmit the disease to a predator than vice versa. As before, we describe this process by explaining how to generate a single feature. If we draw a large sample of properties by repeating this procedure many times, the prior probability of a property will be proportional to the number of times it appears in the sample.

The transmission process has two parameters: *b,* the background rate, and *t,* the transmission probability. The first parameter captures the knowledge that species can contract diseases from causes external to the food web. For each species in the web, we toss a coin with bias *b* to decide whether that species develops the disease as a result of an external cause. The second parameter is used to capture the knowledge that diseases can spread from prey

Table 3
*Models for Reasoning About Threshold Properties*

| Model | Bayesian | Description |
|---|---|---|
| Threshold (1D + drift) | Yes | Equation 10 with $\mu = sigm^{-1}(R)$ and *K* defined over a 1D structure |
| Spatial (1D + diffusion) | Yes | Equations 7–8 with *K* defined over a 1D structure |
| Tree + drift | Yes | Equation 10 with $\mu = sigm^{-1}(R)$ and *K* defined over a tree |
| Taxonomic (tree + diffusion) | Yes | Equations 7–8 with *K* defined over a tree |
| Sim. covariance + drift | Yes | Equation 10 with $\mu = sigm^{-1}(R)$ and *K* set by similarity ratings |
| Sim. covariance + diffusion | Yes | Equations 7–8 with *K* set by similarity ratings |
| Sim Prob | No | See Blok et al. (2007) |

*Note.* For a given domain and threshold property, *R* is a vector indicating the a priori probability that each category in the domain has the property. The first six models are structured statistical approaches that rely on different structures and stochastic processes.

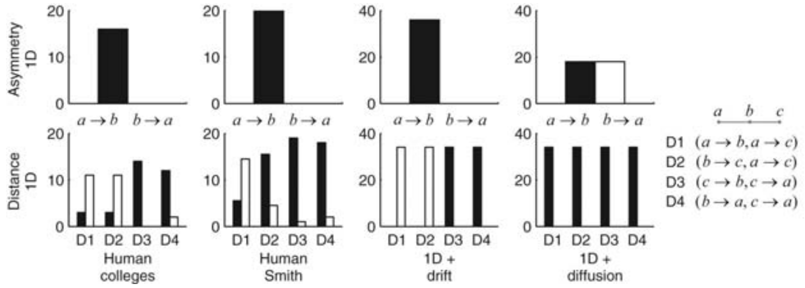

*Figure 12.* Qualitative phenomena predicted by the threshold (1D + drift) and spatial (1D + diffusion) models. Plots in the top row show counts for all pairs of categories (*a, b*) where *a* is less likely a priori than *b* to have the novel property (i.e., *a* is to the left of *b* on the underlying dimension). Black bars show the number of pairs where $a \rightarrow b > b \rightarrow a$, and white bars show the number of pairs where $b \rightarrow a > a \rightarrow b$. The first two plots show that people consistently predict that arguments are stronger when they move from left (*a*) to right (*b*) along the underlying dimension, and the third plot confirms that the threshold model captures this effect. Plots in the second row show counts for all triples (*a, b, c*) where *a* is to the left of *b* along the underlying dimension, and *b* is to the left of *c*. Each pair of bars represents one of the four comparisons from Table 1, and black bars show the number of triples where the first argument was stronger than the second. The threshold model makes accurate predictions about the colleges data, but the Smith data diverge from the predicted pattern. Counts for the two colleges data sets and the two Smith data sets were combined when plotting the human data, and counts for all four data sets were combined when plotting model predictions.

to predator up the food web. For each edge in the graph, we toss a coin with bias *t* to determine whether it is active. We stipulate that all species reachable by active links from a diseased animal also contract the disease. Note that this causal model is formally equivalent to the noisy-or networks sometimes used to model causal relationships between features (Rehder & Burnett, 2005). Here, however, we model causal relationships between species.

Figure 13 shows one possible outcome when we sample a feature from the transmission process. We see that two of the species develop the disease for reasons unrelated to the food web and that four of the causal links are active (Figure 13a). An additional three species contract the disease by eating a disease-ridden species (Figure 13b). Reflecting on these simulations should establish that the prior captures two basic intuitions. First, species that are linked in the web are more likely to share the property than species that are not directly linked. Second, property overlap is asymmetric—a prey animal is more likely to share the property with its predator than vice versa.

The fourth column of Figure 6 shows generalization curves for the causal model when the transmission process is defined over a food chain: a food web with linear structure.[3] The white curve in Figure 6a shows predictions when the species marked with an asterisk is observed to carry a novel disease. Because the mechanism of causal transmission is fallible, the curve shows a distance effect in both directions.[4] The curve also suggests that the causal model leads to asymmetry—inferences up the food chain are stronger than inferences in the opposite direction. Comparing the white curves in Figures 6a and 6b confirms that the causal model leads to asymmetric predictions.

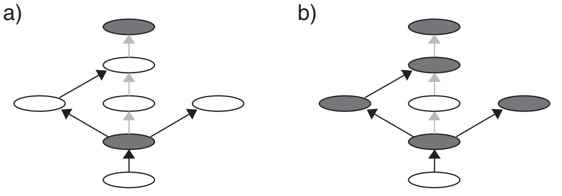

*Figure 13.* Generating a disease property over a food web. (a) indicates that two species (shaded nodes) contracted the disease for reasons unrelated to the food web and that four arrows of causal transmission were active. (b) shows the result of the scenario in (a): five species end up with the disease.

## Property Induction Data

Shafto et al. (2008) designed several experiments to test the causal model and compare it with a tree-based model. Participants in these experiments were trained on two food webs—the *island* scenario used the web in Figure 14a, and the *mammals* scenario used the web in Figure 14b. Participants then rated the strengths of arguments involving two kinds of properties: a disease property ("has disease D") and a genetic property ("has gene XR-23"). Participants also provided pairwise similarity ratings between the species in each food web. When supplied to the structure learning algorithm described in Appendix C, these ratings generate the tree structures shown in Figure 14.

Figure 15 shows predictions for three models described in Table 4. The causal model has two free parameters, and the results in Figure 15 were obtained for $b = 0.1$ and $t = 0.6$. The taxonomic model uses the σ parameter described earlier, and as before we set $\sigma = 5$. A comparison between the top two rows in Figure 15

[3] Predictions for the causal model in Figures 6 and 15 were computed using belief propagation over a Bayes net with the same structure as the food web (see Shafto et al. (2005) for further details).

[4] Parameter values of $b = 0.1$ and $t = 0.9$ were used to generate the generalization curves for the causal model. The profile of the transmission process shown in Table 1 is accurate provided that $0 < b < 1$ and $0 < t < 1$. For instance, if $t = 1$, then the first distance comparison is no longer asymmetric: $a \rightarrow b$ and $a \rightarrow c$ are equally strong because the conclusion is certain in each case.

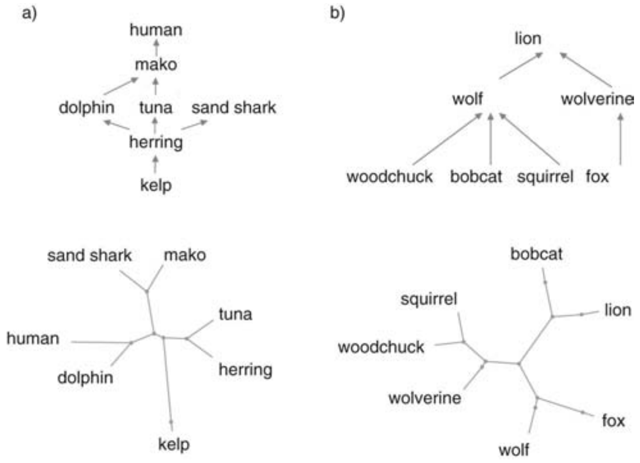

*Figure 14.* Food webs and taxonomic trees for (a) the island scenario and (b) the mammals scenario. The trees were learned from similarity ratings provided by participants in the experiments of Shafto et al. (2005).

Table 4
*Models for Reasoning About Causal Transmission*

| Model | Bayesian | Description |
|---|---|---|
| Causal (web + transmission) | Yes | Prior generated by causal transmission over the food web |
| Taxonomic (tree + diffusion) | Yes | Prior generated by Equations 7–8 with $K$ defined over a tree |
| Web + diffusion | Yes | Prior generated by Equations 7–8 with $K$ defined over the food web graph |

*Note.* The causal model is intended to capture inferences about disease properties, and the taxonomic model is intended to capture inferences about taxonomic properties.

reveals a double dissociation between models and properties. The causal model provides a good account of inferences about the disease property, but not the genetic property, and the taxonomic model shows the opposite pattern of results. This double dissociation provides further evidence that different prior distributions are needed in different inductive contexts and that a Bayesian approach to property induction can capture very different patterns of inference when provided with an appropriately structured prior.

We also compared the causal model to an alternative that uses the diffusion process (Equations 7–8) over the food web graphs. The *web + diffusion* model uses $\sigma = 5$ and relies on an additional parameter: the lengths of the edges in the food web graph. To generate the results in Figure 15, all edge lengths were set to the mean edge length across the two tree graphs in Figure 14. The *web + diffusion* model assumes that species nearby in the food web are likely to share properties but does not capture the idea that disease transmission is asymmetric and that diseases are more likely to spread from prey to predator than vice versa. Figure 15 shows that the *web + diffusion* model performs worse than the causal model, which suggests that generic Gaussian models will not be able to account for every inductive context.

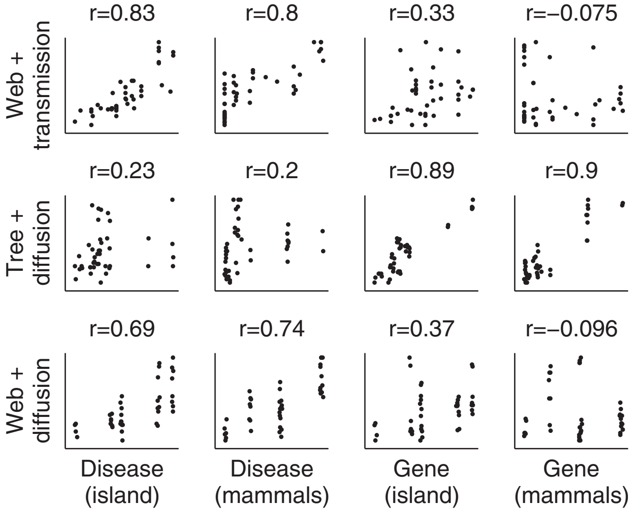

*Figure 15.* Model predictions for inferences about novel diseases and novel genes. The two scenarios (island and mammals) use the species shown in Figure 14. The three models are summarized in Table 4.

Because the data sets in Figure 15 use single-premise arguments, we can directly explore the qualitative phenomena in Table 1. The first row of Figure 16 confirms that asymmetry is characteristic of human inferences about novel diseases but not of inferences about novel genes. The causal model correctly predicts that inferences about diseases will be asymmetric, and the taxonomic model correctly predicts that inferences about genes will be symmetric. The final two rows indicate that human inferences about diseases show a causal distance effect and that human inferences about genes show a taxonomic distance effect. Note, however, that these distance effects do not clearly distinguish inferences about diseases and genes—a causal distance effect is also found for inferences about genes, and a taxonomic distance effect is also found for inferences about diseases. One explanation for this finding is that the webs and taxonomic trees in Figure 14 have some shared characteristics—in the island scenario, for example, tuna and herring are close to each other in the web and the tree, and kelp occupies a distinctive position in both structures.

The overall pattern of results in Figures 15 and 16 suggest that the causal model provides a good account of inferences about novel diseases, that the taxonomic model provides a good account of inferences about novel genes, but that neither of these models accounts well for both inductive contexts. These results provide additional evidence that our structured statistical framework can accommodate very different kinds of prior knowledge and suggest that this framework may be able to handle knowledge effects across many inductive contexts.

## Acquiring Background Knowledge

We have now seen four examples that demonstrate how our structured statistical framework can capture the background knowledge that is relevant to different inductive contexts. Each of our models relies on a structure and a stochastic process, and this section discusses how these components of background knowledge can be acquired by a learner. Our approach is consistent with at

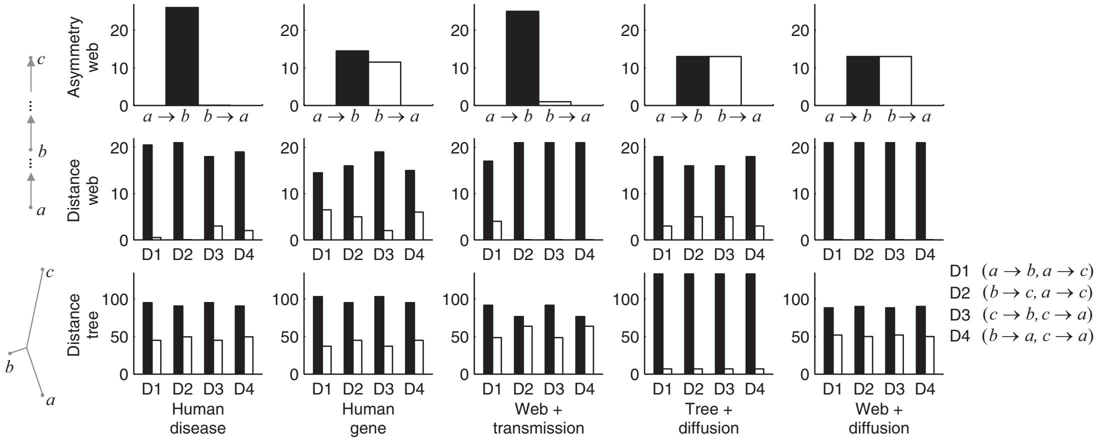

*Figure 16.* Qualitative phenomena predicted by the causal and taxonomic models (counts for the island and mammals scenarios have been combined). Plots in the top row show counts for all pairs (*a, b*) where there is a directed path from *a* to *b* in the food web. Black bars show the number of pairs where $a \rightarrow b > b \rightarrow a$. As expected, we see a strong asymmetry effect for inferences about diseases but not genes. Plots in the middle row show counts for all triples (*a, b, c*) such that there is a directed path in the food web from *a* to *c* that passes through *b*. Each pair of bars represents one of the four comparisons from Table 1: black bars show the number of triples where the first argument was stronger than the second. Plots in the bottom row show counts for all triples (*a, b, c*) such that $d(a, c) > d(a, b)$ and $d(a, c) > d(b, c)$, where $d(a, b)$ is the distance between *a* and *b* in the tree structure.

least two routes to knowledge. Some kinds of background knowledge can be learned from raw data. For example, we show in this section how the tree structure required by our taxonomic model can be learned from a matrix of observed biological properties. Other kinds of structured knowledge may be acquired more directly through explicit instruction or linguistic communication. The food web used by our causal model is one example—a child may be told or may read in a book that lions eat gazelles even if the child has never directly observed a lion or a gazelle, let alone a lion eating a gazelle. We will not model this kind of learning from instruction or communication, but its existence places a crucial constraint on the organization of human knowledge, and we describe in the general discussion how it can be combined with our framework.

Here we describe an approach to acquiring background knowledge from raw data that exploits the very same structured statistical models we have used to explain property induction in previous sections. We have assumed until now that a structure and a stochastic process are known, and that these components combine to generate a prior distribution that guides inferences about unobserved properties. Figure 17a, for example, shows a case where a taxonomic tree can be used to guide inferences about a sparsely observed feature $f^{new}$. Now we consider a setting where many properties have been observed, and each one is generated from a distribution defined by a structure and stochastic process that may both be unknown. Given this setup and the formal machinery we have already developed, Bayesian inference can be used to recover the structure and stochastic process that best account for the set of observed properties. In Figure 17a, for example, a learner observes a matrix of 10 biological properties and discovers the tree that best captures the idea that nearby animals will have many properties in common. Note that this approach to structure learning is based on the same basic idea as our taxonomic model—the idea that biological properties are smoothly distributed over a tree.

The relationship between property induction and knowledge acquisition is shown in Figure 18. Our generative framework can be formalized as a hierarchical Bayesian model (A. Gelman, Carlin, Stern, & Rubin, 2003; Kemp, 2008), and each panel of Figure 18 shows a different way in which this model can be used. The nodes in the model correspond to the labels in Figure 1, and *f, S,* and *T* represent a feature, a structure and a stochastic process, respectively. *F* represents the *form* of structure *S,* and indicates which kind of structure (e.g., tree, low-dimensional space, or directed graph) is appropriate for a given inductive context. The arrows in Figure 18 indicate dependency relationships between nodes: for example, Figure 18a indicates that feature *f* is generated from a distribution $p(f|S,T)$ that depends on structure *S* and stochastic process *T* and that *S* is generated from a distribution $p(S|F)$ that depends on structural form *F*.

Like all hierarchical Bayesian models, the framework in Figure 18 can make inferences at several levels of abstraction. Figure 18a shows the property-induction problem that we have focused on so far. Our four models use different settings of *S, F,* and *T,* but each one assumes that all of these variables are already known (hence shaded in Figure 18a). Each of these models uses a distribution $p(f|S,T)$ to make inferences about a feature *f* which is only partially observed (hence unshaded), but which is known to depend on structure *S* and stochastic process *T.* The remaining three panels in Figure 18 show how our framework can be used to learn one or more of the variables *F, S,* and *T.* In each case, the input provided is a collection of features $\{f^i\}$ that are assumed to be generated over an unknown structure *S* by a process *T* that may also be unknown. The box or *plate* around the $f^i$ variable indicates that *m* features $f^i$ are available and that all of these features are independently generated by process *T* over structure *S.* The $f^i$ node is shaded, indicating that all of these features are observed.

Figure 18b shows a structure-learning problem where form *F* and stochastic process *T* are known, but structure *S* is unobserved

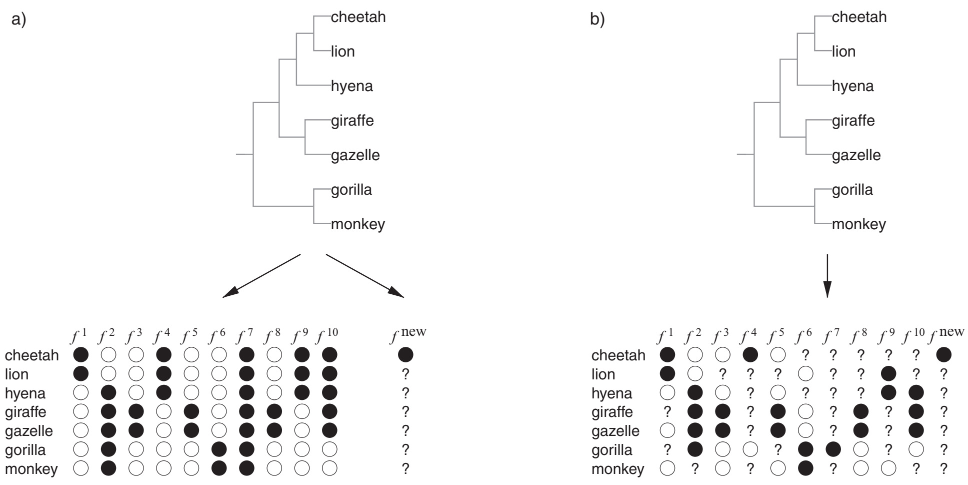
a)
b)
cheetah
lion
hyena
giraffe
gazelle
gorilla
monkey
$f^1$ $f^2$ $f^3$ $f^4$ $f^5$ $f^6$ $f^7$ $f^8$ $f^9$ $f^{10}$ $f^{new}$

*Figure 17.* Property induction and structure learning. (a) Our taxonomic model shows how a tree structure $S$ supports inferences about a novel feature $f^{new}$ that is generated by a diffusion process over the tree. If the tree is unknown, it can be learned given a set of observed features ($f^1$ through $f^{10}$) that are generated in the same way as $f^{new}$. (b) Our framework also handles cases where a learner is given a partially observed feature matrix and asked to fill in the missing entries.

and must be learned. This problem may arise, for instance, if the features $\{f^i\}$ are known to be generated by a diffusion process over some tree structure $S$, but the tree in question is unknown. Biologists face a version of this problem when attempting to reconstruct a phylogeny given a database of genetic features—they know that the structure they seek is a tree of some kind, and they understand how the features in their database were generated over the tree, but the true phylogeny cannot be directly observed. Our framework suggests that this structure learning problem can be handled by identifying the structure $S$ that maximizes the distribution

$$p(S|\{f^i\}, F, T) \propto p(\{f^i\}|S, T)p(S|F). \qquad (11)$$

To compute $p(\{f^i\}|S,T)$, the learner can assume that the features $\{f^i\}$ were independently generated by process $T$ over structure $S$. The remaining term, $p(S|F)$, indicates the probability of structure $S$ given form $F$.

Figures 18c shows a case where the stochastic process $T$ is known but form $F$ and structure $S$ must both be inferred. Structures can come in many different forms, including trees, directed graphs, and low-dimensional spaces, and in some cases a learner may not know which of these forms is appropriate for a given domain. Given a hypothesis space specifying several different forms, the learner can simultaneously discover the structure $S$ and form $F$ that maximize

$$p(S, F|\{f^i\}, T) \propto p(\{f^i\}|S, T)p(S|F)p(F). \qquad (12)$$

The first two terms on the right have already been described, and the final term $p(F)$ is the prior probability of form $F$.

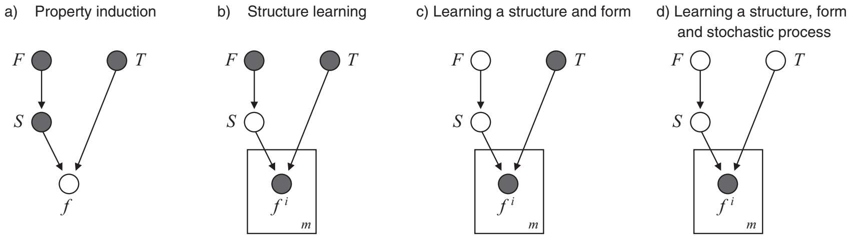
a) Property induction
b) Structure learning
c) Learning a structure and form
d) Learning a structure, form and stochastic process
F
T
S
f
$f^i$
m

*Figure 18.* A generative model for features that can be used to make inferences about an unobserved feature $f$, or to discover the structure $S$, form $F$, and stochastic process $T$ that best account for a set $\{f^i\}$ of observed features. Shaded nodes indicate variables that are fully observed, and arrows indicate statistical dependencies between variables. (a) Property induction: a known structure and stochastic process are used to make inferences about a novel feature $f$. (b) Structure learning: If $m$ features are sampled from a known process $T$ defined over an unknown structure $S$, the structure $S$ can be discovered. The plate around $f^i$ indicates that the $m$ features are sampled independently from a distribution induced by $S$ and $T$. (c) Structure and form discovery: given a prior distribution on possible forms, structure $S$ and form $F$ can be discovered simultaneously. (d) Discovering structures, forms and stochastic processes: If neither $F$, $S$, nor $T$ is known, these variables can be simultaneously discovered given a prior distribution on possible forms and a prior distribution on possible stochastic processes.

Suppose now that we do not know the form of the unknown structure, or the process by which properties are generated over this structure (Figure 18d). Given a hypothesis space of possible stochastic processes, we can simultaneously discover the structure $S$, form $F$, and stochastic process $T$ that maximize

$$p(S, F, T|\{f^i\}) \propto p(\{f^i\}|S, T)p(S|F)p(F)p(T). \qquad (13)$$

where $p(T)$ is the prior probability of stochastic process $T$.[5] In principle, then, given prior knowledge about the kinds of structures and the kinds of stochastic processes that might account for a collection of observed features, statistical inference can be used to choose the structure and the stochastic process that are best supported by the data.

Formalizing all of the inferences in Figures 18b–d is a significant undertaking, and the general discussion summarizes some steps we have taken towards this goal (Kemp & Tenenbaum, 2008) and some challenges that still remain. The rest of this section will focus only on the structure learning problem in Figure 18b. In particular, we describe how the tree and the two-dimensional space in Figure 3 were learned from a matrix of species and their features. Using the notation of Figure 18, this matrix can be represented as a collection of features $\{f^i\}$; but for consistency with the rest of this article, we refer to it as data set $D$. As mentioned already, the features in $D$ were collected in a feature-listing task carried out by Osherson et al. (1991) and include behavioral and anatomical features such as "is gray," "is active," and "lives in water."

Following Figure 18b, we assume that the features in $D$ were generated by a diffusion process over an unknown tree structure and search for the tree $S$ that maximizes

$$p(S|D, F, T) \propto p(D|S, T)p(S|F). \qquad (14)$$

In some cases it will be preferable to compute a posterior distribution over structures $S$: a distribution, for example, that can be easily updated when new data arrive. Our framework supports this approach, but here we attempt only to discover the best tree $S$. Because $T$ is the diffusion process, the likelihood term $p(D|S, T)$ in Equation 14 follows from Equation 7 and will be high for structures $S$ that allow nearby animals in the structure to have many features in common. Because $F$ indicates that the domain is tree-structured, $p(S|F)$ is a distribution over a hypothesis space of possible tree structures $S$. A simple prior $p(S|F)$ on trees assumes that all tree topologies are equally likely and that the branch lengths are independently generated from an exponential distribution (Huelsenbeck & Ronquist, 2001). Combining this prior with a likelihood $p(D|S,T)$ defined by Equation 7 yields a structure learning model that is likely to work well in practice. Here, however, we take a slightly different approach. As described in Appendix C, our tree-learning algorithm searches for the tree $S$ that maximizes a measure closely related to Equation 14. This algorithm can be viewed as a probabilistic alternative to hierarchical clustering.

Equation 14 can also be used to learn representations other than trees. Our algorithm for learning two-dimensional spaces (e.g., Fig 3b) relies on this equation, except that it searches for structures $S$ that are low-dimensional spaces rather than trees. As for the tree case, the likelihood $p(D|S,T)$ is defined by Equation 7, and we use a generic prior $p(S|F)$ that favors spatial representations where most of the categories lie close to the origin. Intuitively, our algorithm attempts to find a two-dimensional space such that nearby categories in the space tend to have many features in common. This method can be viewed as a probabilistic alternative to principal components analysis or multidimensional scaling, and additional details can be found in Appendix C.

We have now described how our taxonomic model can be used for structure learning (Figure 18b) and for property induction (Figure 18a). When evaluating this model, we focused on a setting (Figure 17a) where a learner observes a relatively large feature matrix $D$, discovers the structure $S$ that best accounts for these data, then uses this structure to make inferences about a single novel feature $f^{new}$. Our framework, however, can also handle the more realistic setting in Figure 17b. Now there is no clear distinction between a set of fully observed features and a sparsely observed feature that is the target for inductive inference. The learner has partially observed a large matrix of data, learning more or less about different features, and his or her task is to make predictions about one or more missing values in the matrix. Because our framework relies on probabilistic inference, it provides a principled solution to problems involving missing data. In cases like Figure 17b, our framework can simultaneously discover the structure that accounts best for the entries that have been observed and make inferences about the entries that have not been observed. When children are learning about the biological domain, even relatively common features (e.g., "has feet," "breathes air") are likely to be unobserved for some species (Do penguins have feet? and Do dolphins breathe air?). Children, however, make confident inferences even though there are large gaps in their knowledge (Carey, 1985), and our computational framework may help to explain this ability.

Explaining how background knowledge is acquired is a major challenge for any approach to inductive reasoning. This section has argued that our structured statistical framework provides a unified account of the acquisition and use of background knowledge. In particular, we showed that the same framework that explains how a structure $S$ is used for induction (Figure 18a) can also help to explain how $S$ is learned from a set of observed features (Figure 18b). Learning from observable features may account for some cases of interest, but a comprehensive model of learning must also allow a role for direct instruction. It seems likely, for instance, that the background knowledge captured by our causal model depends more on instruction than on direct experience. Our general framework has room for cases where linguistic input provides complete or partial information about an unobserved structure or stochastic process, and exploring these cases is an important direction for future work.

## General Discussion

Our framework for property induction can be applied across many different inductive contexts. Here we have focused on four:

[5] We have assumed that the structural form $F$ and the stochastic process $T$ are independently generated. When certain stochastic processes are defined only over certain forms, the independence assumption does not apply, and the posterior probability is given by $p(S, F, T|\{f^i\})| p(\{f^i\}|S, T)p(S|F)p(F, T)$ where $p(F, T)$ is a joint distribution over forms and stochastic processes. Our diffusion process, for instance, is defined only over continuous spaces and undirected graphs, and our process of causal transmission is defined only over directed graphs. The joint distribution $p(F, T)$ can capture constraints of this sort.

the default biological context, a context that draws on spatial knowledge, a context involving a threshold property, and a context involving a causally transmitted property. We presented models for each of these contexts, and each of these models is valuable in its own right—for instance, our taxonomic model, our spatial model and our threshold model are competitive with the best previous models (SCM, SimProb) for the tasks they address. More important than any single model, however, is our unifying framework, which can handle many more contexts than we were able to explore here. Different contexts will draw on different sources of knowledge, and our framework should handle all of them provided that we can specify structured statistical models that formalize the knowledge required in each case.

The prior distribution required by our Bayesian framework is specified by two components: a structure that captures relationships between the categories in a domain and a stochastic process defined over that structure. Our four models include two "minimal pairs" that suggest that both of these components are crucial. The taxonomic model and our spatial model use the same stochastic process (the diffusion process), but are defined over different kinds of structures. The taxonomic model accounts well for reasoning about biological properties but accounts poorly for an experiment involving a spatial property. The spatial model shows the opposite pattern of strengths and weaknesses, suggesting that qualitatively different structures are needed to account for different inductive contexts.

When restricted to a one-dimensional space, our spatial model uses the same structure as the threshold model but these models rely on different stochastic processes (the diffusion process for the spatial model, and the drift process for the threshold model). Despite the similarities between these models, the threshold model is significantly better than the spatial model at accounting for inferences about threshold properties. Having the right structure, then, is not sufficient, and a reasoner must also know how properties are generated over that structure.

Compared to previous models of inductive reasoning, our unifying framework has two distinctive features. First, it explains human judgments as statistical inferences, and, second, it uses different kinds of structures to capture different kinds of background knowledge. Our approach therefore combines the insights of two groups of researchers: those that emphasize structured knowledge but not statistical inference (Chomsky, 1986; Collins & Quillian, 1969; Lenat, 1995) and those that emphasize statistics but not structure (Rogers & McClelland, 2004). These two research traditions are sometimes thought to be incompatible, at best corresponding to two different subsystems of the mind. The next sections argue that, on the contrary, structure and statistics must both play crucial roles in a unified framework for modeling human inductive inference.

### The Importance of Structure

Many researchers have observed that inferences about different properties can be guided by very different kinds of knowledge, and we have argued that formal models of property induction should incorporate this insight. Structured models can make progress towards this goal, because different kinds of structures can capture different kinds of background knowledge. To demonstrate the advantages of a structured approach, we compare our framework to alternatives that do not focus on structured background knowledge, including similarity-based approaches and feature-based approaches. When considering these approaches, it is critical to distinguish between specific models and modeling frameworks. For instance, we can compare our taxonomic model to other specific models in the literature, such as the SCM, Sloman's feature based model (1993), and the connectionist model of Rogers and McClelland (2004). Equally important, however, are comparisons between our structured statistical framework and other modeling frameworks, such as the connectionist framework.

*Specific models.* Most previous models of property induction can be applied to more than one inductive context. The SCM, for instance, can use a different similarity metric for each context, and the feature-based models described by Sloman (1993) and Rogers and McClelland (2004) can rely on different feature sets in different contexts. Each model described in this article can also be supplied with different information in different contexts: For instance, the taxonomic model can rely on different taxonomic trees, and the causal model can rely on different directed graphs.

Each of these individual models offers some flexibility, but none qualifies as a comprehensive account of property induction. Different inductive contexts can draw on very different kinds of knowledge, and the models just described will only capture a handful of these cases. Each model is limited in at least two respects—its *representational power* prevents it from capturing certain kinds of knowledge, and the *inductive bias* it provides will only sometimes be appropriate when learning from raw data. We consider each factor in turn.

Each of the specific models already mentioned will capture only some kinds of background knowledge. Consider, for instance, the knowledge used by participants in the food web task that we modeled (Shafto et al., 2008). We have proposed that this knowledge includes at least two components: knowledge about predator–prey relations and knowledge that diseases are often transmitted from prey to predator. Neither component is well captured by the notion of similarity, which makes it difficult to see how the SCM could account for this task. Predator–prey relations could be encoded as a set of features (e.g., kelp has the feature "is eaten by herring"), but neither the feature-based model of Sloman nor the connectionist model of Rogers and McClelland seems able to incorporate the knowledge that diseases are noisily transmitted from prey to predator. Our causal model does capture this knowledge, but is fundamentally limited in its own way, and fails, for instance, to account for inferences about taxonomic properties. The general conclusion suggested by these considerations should be uncontroversial—different inferences can draw on different bodies of knowledge, and each specific model in the property induction literature can only represent a fragment of this knowledge.

Each existing model of property induction is also limited in its ability to learn from raw data, such as observable features. We previously saw several examples of this kind of learning—for instance, each of the first six models in Figure 7 takes a matrix of features as input and uses this information to predict the extension of a novel property. The inductive bias of a learning model is the set of a priori assumptions it brings to the learning problem. For instance, the taxonomic model assumes that the novel property and the features in the matrix are both generated over an unknown tree, and the spatial model assumes that these features are generated over a two-dimensional space. A strong inductive bias is useful because it constrains the inferences that a model is prepared to make. Many researchers have noted the importance of constraints and suggested that cognition would be impossible without them (Chomsky, 1980; Geman et al., 1992; Keil, 1981). Constraints, however, can only help if they guide the model in the right direction—in other words, the inductive bias needs to match the problem at hand. Each existing model of property induction relies on an inductive bias that is appropriate for some contexts, but not for others.

Our results for the default biological context show in miniature why inductive bias matters, and how it differs from the notion of representational power. Consider the taxonomic and raw covariance models (Table 2). These models are identical except that they rely on different covariance matrices *K*. The raw covariance model uses $1/85DD^{\mathrm{T}}$ as the covariance matrix, but the taxonomic model only considers covariance matrices that are defined over some tree structure. Note that the raw covariance model offers the greater representational power, because it can represent all of the covariance matrices available to the taxonomic model, plus more. In reality, though, the taxonomic model should perform better in domains that are actually tree-structured. When the models learn *K* from noisy or sparse data, the tree model is likely to make a better choice than the raw covariance model. The raw covariance model will stick with the raw covariance $K = 1/85DD^{\mathrm{T}}$: A priori, it has no reason to prefer any other *K*. Because the taxonomic model expects a tree-structured covariance, it can clean up the raw covariance by choosing a nearby covariance that is consistent with a tree structure.[6] The plots in Figure 5 support this argument by showing how a tree-based inductive bias allows the taxonomic model to convert the raw covariance (Figure 5e) into a tree-structured covariance (Figure 5c) that cleans up some of the apparent noise in the raw covariance. Although we have chosen to contrast the taxonomic model with the raw covariance model, similar arguments can be made about Sloman's feature-based model and the connectionist model of Rogers and McClelland (2004). Note, however, that these arguments are only applicable when a given domain is tree-structured and that models like the raw covariance model may well outperform the taxonomic model in settings where this condition is not satisfied.

The limitations shared by all existing models of property induction have led us to present a modeling framework—the structured statistical approach—rather than a single model of induction. Our framework has room for many kinds of structures and stochastic processes, and these components can be combined to capture many kinds of background knowledge and to provide many different inductive biases. We described four specific applications of our framework, but our framework also has room for models that rely on theories more complex than any of the examples presented here, including theories formulated using grammars (N. D. Goodman, Tenenbaum, Feldman, & Griffiths, 2008) or logical representations (Kemp, Goodman, & Tenenbaum, 2008).

Although we have chosen to focus on a framework rather than a single model, a single comprehensive model of induction may be possible in the distant future. The ultimate model of the brain, for instance, will be a single model that can represent many kinds of knowledge and that incorporates an inductive bias that supports successful learning in many different contexts. One way to make progress towards this ultimate model is to develop an expressive representation language that can capture all of the structures we have described, along with many others. A language such as predicate logic is one possible candidate, and psychologists have noted that compositional languages of this sort help to explain how complex bodies of knowledge can be represented as combinations of simple components (Fodor & Pylyshyn, 1988). At present, however, aiming for a single all-embracing model of inductive reasoning may be premature, and exploring multiple models consistent with a framework like ours may be the more promising strategy.

*Modeling frameworks.* If we are justified in presenting a framework rather than a single model of induction, it is important to consider how this framework compares to others in the psychological literature. The main alternative is the connectionist framework—for instance, Rogers and McClelland (2004) argue that connectionist models can account for many kinds of inductive reasoning, including cases like causal reasoning that are often thought to rely on intuitive theories. The previous section argued that the specific connectionist model described by Rogers and McClelland is limited in two fundamental respects (representational power and inductive bias), but here we focus on the framework that they advocate rather than the specific model that they present.

Although connectionist networks are usually seen as unstructured models, the class of these networks is Turing-complete, and any formal model can therefore be replaced by a connectionist model that achieves the same input–output mapping. It follows that any structured statistical model has a connectionist counterpart, and the four models proposed in this article are no exception. In response to our article, a connectionist modeler might describe four PDP models that capture different kinds of inductive reasoning and might argue that the connectionist framework is useful for explaining how inductive inferences depend on background knowledge. There are several reasons, however, to prefer our structured statistical framework.

The connectionist framework may be useful for developing models of psychological processing, but is not ideal for exploring how induction is guided by background knowledge. Connectionist models can capture knowledge of different kinds, but they often fail to do so transparently. It is difficult for modelers to express knowledge in a form that can be exploited by these models, and therefore difficult to explore the consequences of different assumptions about the nature of background knowledge. Our Bayesian framework forces us to be explicit about the knowledge that guides induction, and the need to formulate a prior distribution brings the question of background knowledge into sharp focus.

Although structured representations can capture background knowledge, Rogers and McClelland (2004) suggest that these representations are best regarded as epiphenomena that emerge from connectionist processing. They argue, for instance, that a connectionist model can implicitly pick up on the hierarchical structure of a domain, even if the model includes no explicit representation of a hierarchy. There are at least two considerations, however, which suggest that structures may sometimes need to be

[6] This argument makes use of the bias–variance tradeoff that is familiar to statisticians.

explicitly represented. First, explicit structures allow a model to incorporate high-level observations (including sentences uttered by others) that have direct structural implications. Second, explicit structures capture knowledge in a form that supports transfer to other inductive problems.

Many high-level examples of inductive inference rely directly or indirectly on linguistic input, and language allows theoretical knowledge to be communicated explicitly. For instance, a child may be told that "dirt has tiny bugs inside that make you sick," or that a dolphin is not a fish even though it looks like one. Structured approaches can incorporate information that has direct structural implications—for example, a single statement about dolphins may lead a learner to revise her taxonomy so that dolphins are grouped with the other mammals instead of the fish. Similarly, the statement that "Indiana is next to Illinois" may lead a learner to revise his mental map of the United States if he previously believed that Indiana was somewhere near California. It is difficult, however, to understand how linguistic statements like these might reach in and transform the distributed representations typically used by connectionist models.

The food web task that we modeled (Shafto et al., 2005) is one simple setting where inductive reasoning is guided by linguistic input. Participants in this task were asked to predict the extension of a novel disease, and the knowledge driving their responses included the structural information provided during training (e.g., the statement that "tuna eat herring"). These predator–prey relations lead immediately to a food web structure that can be exploited by our causal model, but it is more difficult to understand how a learner could read about these relations and immediately construct a connectionist network that incorporates them in the right way.

A second reason why explicit structures are useful is that they support knowledge transfer. Consider again the taxonomic and raw covariance models, and suppose that the models have been supplied with a tree-structured data set that is large enough and clean enough that both have settled on an identical covariance matrix $K$. Even though both models now make identical predictions, the taxonomic model may still be preferred because a large part of its knowledge is compactly represented by the tree structure in Figure 3. This tree structure can be used for other inductive tasks—for example, learning superordinate categories may involve attaching labels to structured representations, and the structure in Figure 3a may be useful for learning the meaning of "primate" (Xu & Tenenbaum, 2007). It is rather more difficult to understand how the knowledge captured by the raw covariance model can be extracted and used in this way, and a similar conclusion applies to connectionist models of property induction.

Our arguments in favor of structured models build on claims made by previous researchers (Chomsky, 1986; Geman et al., 1992), and psychologists and artificial intelligence researchers have been developing structured models for many years (Collins & Quillian, 1969; Davis, 1990; Lenat, 1995). There are several reasons to prefer a structured approach to alternatives such as the connectionist framework, but the ultimate test for a framework is whether it can stimulate further developments in a scientific field. We have presented four models that suggest that our framework may be worth exploring, but applications to many other inductive contexts are needed before we can conclude that it provides a comprehensive account of inductive reasoning.

### *The Importance of Statistics*

Many researchers have argued in favor of structured approaches, but others have argued that structured representations are bought at a price and that structured approaches are weak in the places where statistical approaches are strong. A recent and comprehensive version of this argument is presented by Rogers and McClelland (2004), who raise two key objections against structured approaches. First, they suggest that structured approaches are overly rigid and are unable to deal with exceptions, graded category membership, and other phenomena that appear to require soft probabilistic explanations. Second, they suggest that there are no plausible accounts of how structured knowledge might be acquired. Both objections raise serious problems for some structured approaches. Systems that rely on deductive inference do have trouble with noise and exceptions, and approaches that rely on logical theories have sometimes used representations so complex that they seem to allow little scope for learning (Rumelhart, Lindsay, & Norman, 1972). Unlike these approaches, our framework combines structured background knowledge with statistical inference. We believe that the statistical nature of our framework addresses both of the standard objections against structured models, although the learning issue does raise challenges for future work.

Our framework deals well with noise and exceptions if we generate a prior using an appropriate stochastic process. The prior is always generated over a structure, but the three processes we described (diffusion, drift, and transmission) enforce the structural constraints in a relatively soft manner. Note that the priors generated by these processes always assign nonzero prior probability to every possible feature. Our results for the default biological context include one case where a soft probabilistic model outperforms a more rigid alternative. The strict tree model uses the same structure as our taxonomic model, but only assigns nonzero prior probability to features that are perfectly consistent with the tree. These two models make different predictions about some three-premise arguments with the conclusion that all mammals have the novel property. Intuitively, three examples spread widely over the tree provide better support for the conclusion than a trio where two of the examples are neighbors and the third is located at some distance. In the second case, the lone example can potentially be explained as an exception, but the strict model does not allow this option. The performance of the strict tree approach on the *Osherson mammals* data confirms that it does not adequately capture diversity-based reasoning.

There are other inductive phenomena that appear probabilistic in nature and are well explained by existing statistical approaches. Explaining some of these phenomena may require modifications to the specific models we have described—for example, we later describe how the framework can be modified to capture some kinds of typicality effects. We see no fundamental reason, however, why any of these phenomena should fall outside the scope of our framework.

Although we have not provided a comprehensive solution to the problem of structure acquisition, we showed that our probabilistic framework provides a unified approach to structure acquisition and use. In particular, we showed that a single hierarchical Bayesian model can be used for property induction (Figure 18a) and for discovering the structure that best accounts for a set of observed features (Figure 18b). In principle, the same hierarchical model should also be able to discover the structural form (Figure 18c) and stochastic process (Figure 18d) that best account for a set of features, and the next section discusses how these inferences can be formalized.

The central theme of our work is that structured knowledge and statistical inference are both critical for explaining inductive inference. Previous discussions of representation and learning have tended to emphasize just one of these ideas, but by combining them in the right way we can create models that inherit the benefits provided by both components. Structured models are needed to capture the knowledge that drives induction, and statistical inference is necessary to explain how this knowledge is acquired and used for inference in the presence of noise and uncertainty.

### *Matching Models With Inductive Contexts*

We have argued that people bring a strong inductive bias to inductive problems. As mentioned earlier, a strong inductive bias is useful only when it matches the problem at hand: a misguided inductive bias can be actively harmful. We believe that no single inductive bias can account for all of the problems that humans are able to solve, and the need for many different inductive biases is the reason we have presented a modeling framework rather than a single model of induction. Our framework, however, raises a fundamental question—how can we choose the right theory for an inductive context? We address the question from three perspectives. First we discuss how a cognitive modeler can choose the right theory for a given context—for example, how a modeler might decide that tree representations are useful for explaining inferences in the default biological context. Next we discuss how theories might be acquired over the course of cognitive development—for example, how children might learn that the biological domain is tree-structured. Finally, we consider how an adult with access to many theories might identify the one most relevant to a given inductive problem—for example, how a reasoner might decide to use a tree structured taxonomy when asked to reason about the property "has enzyme X132."

*Modeling theories.* This article has provided computational theories (Marr, 1982) or rational analyses (Anderson, 1990) of inference in several inductive contexts. Our working assumption is that human inferences are approximately rational with respect to the structure of the world and that choosing the best model for an inductive problem is a matter of formalizing some aspect of the world and deriving normative predictions given these formal commitments. For example, our taxonomic model was motivated by the idea that biological properties were generated over an evolutionary tree, and our causal model is based on a simple theory of how diseases are actually transmitted over food webs. As these examples suggest, computational theories of induction will often correspond to simple theories in other branches of science (for example, evolutionary biology and epidemiology).

*Acquiring theories.* Building theories, however, is not just a task for scientists. Over the course of development, all of us acquire folk theories for domains like biology (Carey, 1985), sociology (Hirschfeld, 1996) and physics (McCloskey, 1983). In principle, Bayesian methods can help to explain how theory acquisition is possible—given a prior distribution over a hypothesis space of theories, Bayesian inference can be used to choose the theory in the space best supported by some data. Whether or not this solution will turn out to be satisfactory is not yet clear. Finding an adequate specification of a single theory is often difficult, and characterizing a large space of these theories may appear even more challenging. Yet representations of great apparent complexity can sometimes be generated by very simple processes (Ball, 1999), and one day it may be possible to describe a relatively small number of principles that can generate a large hypothesis space of intuitive theories.

Our formal framework suggests some initial approaches to the very challenging problem of theory acquisition. Because we have attempted to capture the content of intuitive theories using structures and stochastic processes, a solution to the developmental problem of theory acquisition must explain how children might acquire these structures and stochastic processes. In principle, the probabilistic approach we have been using all along can explain how structures, structural forms, and stochastic processes can be learned from raw data (Figure 18b–d). We have already shown how the inference in Figure 18b can be formalized, and Kemp and Tenenbaum (2008) described one way to formalize the inference in Figure 18c.

The model in Figure 18c attempts to explain how a learner might discover a structure if he or she does not already know the structural form of the domain—in other words, if the learner does not know whether the categories in the domain should be organized into a tree, a ring, a set of clusters, a chain, or some other kind of structure. Given a hypothesis space of structural forms, Bayesian inference can be used to simultaneously discover the form and the instance of that form that best explain a set of observed features. Kemp and Tenenbaum (2008) showed, for instance, that a data set of biological features is best described by a tree, but a data set including Supreme Court judges and their votes is best described by a linear representation: the liberal–conservative spectrum. The notion of a hypothesis space of structural forms immediately raises another acquisition problem—how can a learner know which forms should appear in this space? An appealing solution is to specify a simple language that can generate the hypothesis space of structural forms, and Kemp and Tenenbaum (2008) proposed that graph grammars can be used to define such a language. An acquisition problem still remains—how might a learner discover this simple language?—but the ultimate goal is to specify a language simple enough and general enough that it can be plausibly assumed to be innate.

A similar approach can be used to formalize the model in Figure 18d, which attempts to simultaneously discover the structural form, the instance of that form, and the stochastic process that best explain a set of observed features. To implement the model in Figure 18d, it will be necessary to describe a hypothesis space of stochastic processes, and identifying the regularities that will allow this space to be compactly described is an open problem.

Even if the model in Figure 18d can be formally specified, it will not provide a complete solution to the problem of theory acquisition. At least two critical issues remain. First, each of the models in Figures 18b–d takes a set of observable features as input, and we have not carefully specified how these features might be chosen. Consider a child visiting the zoo and observing an elephant for the first time. How can the child know that possessing a trunk is an important feature, but that standing in the north-west corner of the enclosure is not? Questions like these make contact with philosophical puzzles (N. Goodman, 1955) that may never be resolved, but methods for learning feature weights provide a partial solution.[7] Suppose that a child has access to a large set of features, and that some of them are clean with respect to an underlying tree, although most are noisy. The ability to distinguish between clean and noisy features should allow a child to learn successfully even from data sets that contain many irrelevant features. Note, however, that alternative approaches will be needed when a data set includes features that are relevant to several different structures—e.g., a taxonomic tree and a set of ecological categories (Shafto, Kemp, Mansinghka, Gordon, & Tenenbaum, 2006)—and a learner must simultaneously discover all of these structures.

A second critical issue for our approach is that analyzing raw data will only carry a learner so far. Explicit instruction plays a central role in many real-world cases of theory acquisition, and a complete model of theory acquisition must find a way to handle the cultural transmission of knowledge. As described, the models in Figure 18b–d only take observable features as input, but these models can also take advantage of linguistic statements that provide information about any of the variables in Figure 18 (*f, S, F,* or *T*). For example, children learning about biology might be told that dolphins breathe air—in other words, they might learn about a feature *f* that they have not directly observed. They might be told that dolphins are mammals rather than fish and might interpret this claim as a statement about the location of dolphins in structure *S.* They might even be given direct information about the form *F* and the stochastic process *T*—for example, they might learn that the theory of evolution implies that living kinds can be organized into a tree and that properties are generated by a mutation process over this tree. Statements about variables other than features *f* can be viewed as "theory fragments," and assembling a theory from theory fragments is presumably much simpler than inducing it from raw data. Exploring models that can learn from both feature data and theory fragments is an important direction for future work.

*Matching theories with tasks.* Suppose that a learner has successfully acquired several theories, including a theory about the distribution of anatomical properties (a taxonomic theory) and a theory about the transmission of disease (a transmission theory). When faced with a specific inductive task, the learner must decide which of these theories is most appropriate. For example, a property like "has enzyme X132" should allow the learner to conclude that the taxonomic theory is more relevant than the transmission theory. Our current framework does not account for this ability, and overcoming this limitation may require solutions to many challenging problems, including the problem of natural language understanding. The ultimate model of context-sensitive reasoning should be able to read the instructions provided to participants in the experiments that we modeled and to decide for itself which aspects of its knowledge base are most relevant to the problem at hand. At a minimum, this model will require a semantic module that knows that words like "hormone" and "gene" are related to the taxonomic theory and that words like "infection" and "toxin" are related to the transmission theory.

Although a comprehensive account of online theory choice may take decades to develop, some simple versions of this problem should be immediately tractable. Consider, for instance, the finding that the same property can sometimes trigger different theories: "has property *P*" will sometimes trigger the taxonomic theory (*horses* → *zebras* (*P*)), and sometimes trigger the transmission theory (*gazelles* → *lions* (*P*)). Inferences like these can be explained using notions like relevance (Medin et al., 2005) or Gricean implicature (Grice, 1989), and related intuitions can be formalized within a Bayesian framework (Xu & Tenenbaum, 2007).

Suppose, for example, that participants are judging many different arguments, each of which uses a different blank property—the first is about "property *P*1," the second about "property *P*2," and so on. To keep matters simple, assume that a participant considers only two hypotheses about each property—the property is either a taxonomic property or a disease property. A natural assumption is that half of the arguments were generated from the taxonomic theory and that the rest were generated from the transmission theory. A participant might also assume that half the arguments are strong and the remaining half are weak, because an experiment where most of the arguments were strong would not be very useful. Because both theories generate many more weak arguments than strong arguments, it follows that an argument that is strong under the taxonomic theory is likely to have been generated from that theory, and an argument that is strong under the transmission theory is likely to have been generated from that theory. In other words, *P*1 is likely to refer to a taxonomic property in the argument *horses* → *zebras (P1),* but *P*2 is likely to refer to a disease property in the argument *gazelles* → *lions (P2).* Future work can test this approach in the laboratory and attempt to extend it to other settings where learners decide which theory is most relevant to a given context.

## *Other Inductive Phenomena*

This article has emphasized fundamental principles that apply across many inductive contexts. In particular, we focused on the interaction between structured background knowledge and statistical inference—an interaction that provides insight into many inductive contexts but does not fall within the repertoire of most previous models. Even a single context, however, can be a rich field of study, and our framework can be extended to handle the complex phenomena that arise when any single context is explored in depth. Because the default biological context has been comprehensively studied by others, we take it as a case study and describe how our general approach can deal with some of the subtleties that arise when this context is considered in detail.

[7] We have treated all features equally, but it is possible to introduce a weight $\lambda^j$ for each feature. Equation 21 then becomes $p(y^j) \propto \exp\left(-\frac{\lambda^j}{2} y^j \Delta y^j\right)$, where $y^j$ is the *j*th feature. Once we place a prior on the feature weights (for example, a prior that encourages most weights to be small), we can simultaneously discover the structure *S* and the weights for each feature. The weights will reflect the extent to which a feature is smooth over *S,* and the features that match the structure best will end up with the highest weights. One of the functions of intuitive theories is to determine feature weights (Murphy & Medin, 1985) and showing how these weights can be learned is an important step when attempting to explain the acquisition of intuitive theories.

Our taxonomic model accounts for some but not all of the inductive phenomena described by previous researchers. Returning to three phenomena described earlier, the model captures premise-conclusion similarity and diversity but not typicality.[8] Kemp and Tenenbaum (2003) propose one Bayesian interpretation of the typicality effect, but here we describe another. The failure to capture typicality depends critically on the meaning of the word "all" in an argument like *seal → all mammals (enzyme).* If "all" is interpreted as "virtually all," or "more than 95%," then the model does predict the typicality effect—animals located centrally in the tree provide better evidence for the conclusion "virtually all mammals" than animals with a large average distance from all of the other animals. The idea that "all" may not be interpreted literally receives some support from previous psychological research, including the experiments of Newstead and Griggs (1984).

An alternative interpretation of "all" may also allow our Bayesian framework to account for some so-called inductive fallacies. The *inclusion fallacy* (Osherson et al., 1990) describes cases where reducing the generality of an argument's conclusion decreases its inductive strength: For example, *robins → all birds (enzyme*) is often judged stronger than *robins → ostriches (enzyme).* If "all" is interpreted literally, than the effect is a logical fallacy, but if "all" means "almost all" it is quite possible that virtually all birds share the novel property, but that very unusual birds like ostriches are excluded.

As we have just seen, some phenomena that appear to elude our model can be explained if we think carefully about how participants might be interpreting the task. To account for other phenomena, however, it may be necessary to adjust the assumptions that determine the mathematical form of our model. *Nonmonotonicity* effects describe cases where the strength of an argument decreases when premises are added. Our taxonomic model does not capture these effects as it stands, but a variant will account for at least some of these effects. Consider, for example, the argument *brown bears → horses (enzyme),* which may appear stronger than the argument *brown bears, polar bears, grizzly bears → horses (enzyme*) (Medin et al., 2005). We previously assumed that the species in the observation set (here brown bears, polar bears and grizzly bears) were randomly selected from the biological domain, but here it seems more sensible to assume that these species were sampled at random from among the species that have the novel property (Sanjana & Tenenbaum, 2003). Equation 2 should therefore be replaced by the likelihood:

$$p\ (l_X|f) \propto \begin{cases} \dfrac{1}{|f|^m}, & \text{if } f_X = l_X \\ 0, & \text{otherwise} \end{cases} \qquad (15)$$

where $m$ is the number of species in $X$, and $|f|$ is the total number of species that have feature $f$. The likelihood in Equation 15 has been previously used to model a variety of tasks, including concept learning (Tenenbaum & Griffiths, 2001), word learning (Xu & Tenenbaum, 2007), and property induction (Sanjana & Tenenbaum, 2003). Combining this likelihood with our Bayesian framework will capture the intuition that if three random examples of species with a novel feature turn out to be bears, it is likely that bears alone have the novel feature.

Giving ourselves the freedom to use slightly different models for different tasks may seem like a step backwards, because an extreme version of this approach allows us to develop a different model for every phenomenon of interest. Some flexibility seems essential, however, because tasks vary in subtle ways and people are sensitive to these variations. For instance, the choice between likelihood terms will depend partly on the inductive context (Fernbach, 2006): Sanjana and Tenenbaum (2003) consider a context where Equation 15 is appropriate, but other contexts correspond better to the likelihood in Equation 2. To build a convincing Bayesian model for any given task, it is critical to think about the assumptions suggested by the cover story, the stimuli chosen, and the framing of the task, and to formalize the task in a way that captures as many of these assumptions as possible.

### Limitations

The idea that different models are needed for different tasks motivates one possible criticism of our approach. Some readers will worry that our framework is too powerful, and that by choosing the right representations and stochastic processes we can develop structured statistical models that account for any conceivable pattern of data. A second possible criticism is that our framework is too weak and will be unable to account for the many cases where people's inferences depart from normative standards. Both criticisms have been directed against Bayesian models in general, and the merits of these claims have been vigorously debated (Anderson, 1990; Oaksford & Chater, 2007; Simon, 1991; Tversky & Kahneman, 1974).

The idea that our approach is too powerful may seem curious at first. Our framework is better able to incorporate background knowledge than most previous models of property induction, but there is still a vast gap between the abilities of our models and the abilities of human learners. Even four-year old children are able to make inductive inferences that depend in subtle ways on the property chosen (S. A. Gelman & Markman, 1986), and capturing the common sense knowledge that drives these inferences is a major challenge for the modeling community. We have described models that capture some aspects of common sense knowledge, and our general framework can be applied to inductive contexts other than the four we have considered, but we are still a long way from a comprehensive account of property induction. The first-order task for cognitive modelers should be to address the vast limitations of our current proposals, not to develop models that achieve even less.

Concerns about the power of Bayesian models arise more naturally when thinking about specific tasks and specific data sets. Some readers may wonder, for instance, whether defining the prior in the right way will allow a structured statistical approach to account for any possible response to the experiments we have modeled. If modeling assumptions of arbitrary complexity are permitted, there may well be Bayesian models that account for any given pattern of data. Cases where people give judgments that are not consistent with a single coherent probability distribution may appear to pose a problem, but a modeler can assume, for instance, that there is a time-dependent process that switches the inductive context every second, and that each probability judgment is accu-

[8] Typicality eludes the model because $p(f_{mammals} = 1|f_i = 1) = \dfrac{p(f_{mammals} = 1)}{p(f_i = 1)}$; and, under the assumptions of Equations 7–8, $p(f_i = 1) = 0.5$ for every species $i$.

rate at the time it is given but soon out of date. If we limit ourselves to plausible models, however, there will be many conceivable data sets that are not explained by any Bayesian account.

Assessing the plausibility of a model will always involve some degree of subjectivity, but two criteria can serve as guidelines—plausible models of property induction should rely on few free parameters and should be consistent with common sense knowledge about the domain in question. Our models aim to satisfy both criteria—for instance, our causal model relies on two interpretable parameters (the base rate and transmission probability) and captures the common sense idea that diseases can spread over a food web. Both criteria are related to considerations that apply more generally when comparing scientific theories. There will always be multiple theories that account for any given data set, including multiple models that fit the data perfectly. Competing theories must therefore be assessed according to several criteria—accounting for empirical data certainly matters, but simplicity and consistency with our best current explanations of other scientific phenomena are also important.

Because plausible Bayesian models cannot explain *everything,* it is natural to ask how well they account for the human inferences we wish to explain. Bayesian models have provided insight into many aspects of cognition (Anderson, 1990; Oaksford & Chater, 2007), but there are well-known cases where human behavior appears to diverge from normative standards (Tversky & Kahneman, 1974). These findings can be organized into at least two categories. The first category includes cases where the proposed normative standard does not capture the true structure of the task and where people's behavior turns out to be consistent with a Bayesian account once the true structure of the task is recognized (Hertwig & Gigerenzer, 1999; McKenzie, 2003). The second category includes findings that no plausible Bayesian analysis will be able to explain. At present there is no clear consensus about the findings that belong to either category. For instance, the inclusion fallacy raises challenges for normative approaches, but we argued previously that this phenomenon may be consistent with a normative model that recognizes that the intended meaning of "all" is often "virtually all." Conjunction fallacies (Medin et al., 2005) are also taken as evidence against normative accounts of property induction, but some of these fallacies may be consistent with normative models that assume that the premises and conclusion of an argument are deliberately chosen to signal relevant background knowledge.

Although some fallacies in the literature may turn out to be compatible with a Bayesian approach, there are good reasons to expect that many others will resist a simple Bayesian explanation. Bayesian methods are useful for developing computational theories of property induction (Marr, 1982), but a complete account of property induction will also need to describe the psychological mechanisms that carry out the computations required by these theories. Because the computational resources of the mind are limited, some computational theories will be implemented only approximately, and these approximations may lead to inferences that have no adequate explanation at the level of computational theory. In order to explain everything that psychologists wish to explain, Bayesian models will need to be supplemented with insights about psychological and neural mechanisms.

### *Implementing Theory-Based Approaches*

Throughout we have worked at the level of computational theory, but eventually we will need to consider whether our computational theories can be implemented or approximated by psychologically plausible processes. A crucial question is whether the computation in Equation 5 can be efficiently organized—in other words, whether it can be carried out without explicitly summing over the entire set of possible feature vectors *f.* In the case of the causal model, the computation can be efficiently approximated using belief propagation over a Bayes net with the same structure as the food web (Shafto et al., 2005). Belief propagation is reminiscent of psychological theories of spreading activation, because the algorithm relies only on operations that are local with respect to the structure of the Bayes net. The threshold model can also be approximated relatively efficiently (Williams & Barber, 1998), and future work should consider whether the taxonomic and spatial models can also be efficiently implemented.

Finding an efficient implementation may raise new challenges for each new theory we consider, but there is a large class of models for which relatively efficient implementations are known. Graphical models are probabilistic models defined over graph structures, and inference in these models has been extensively studied by both statisticians and machine learning researchers (Jordan & Sejnowski, 2001). We do not suggest that all intuitive theories can be captured using graphical models—for instance, kinship theories seem likely to require logical representations that go beyond the standard repertoire of graphical models. Graphical models, however, provide a useful starting point for building formal theories that can capture structured knowledge and support efficient inference.

### *Beyond Property Induction*

Although we have focused on simple instances of property induction, our general framework can be extended to address many other inductive problems. A natural extension handles more complex cases of property induction: cases where the premises and conclusion involve several properties and causal relationships between these properties are known (cf. Ahn, Kim, Lassaline, and Dennis, 2000; Rehder and Burnett, 2005). Suppose, for example, we are told that chimps express the T4 gene, and that the T4 gene usually causes enzyme X132 to be expressed. We can confidently conclude that enzyme X132 can probably be found in gorillas. Inferences like these can be captured by combining a representation of animal categories (e.g., a tree structured taxonomy) with a causal network over features (Kemp, Shafto, Berke, & Tenenbaum, 2007). Combinations like these are possible because probabilistic approaches are relatively modular—the notion of probability provides a common currency that allows qualitatively different approaches to be combined.

Our basic claim that Bayesian models capture knowledge effects when provided with a theory-based prior is relevant to many aspects of high-level cognition. The same idea has been used to model word learning (Xu & Tenenbaum, 2007) and causal inference (Griffiths, Baraff, & Tenenbaum, 2004) and may be broadly useful for explaining inferences that draw on folk physics, folk psychology, and folk biology. Choosing an appropriate prior for each new case may be a difficult challenge, but the Bayesian framework suggests how to start developing formal approaches to these fundamental problems.

## Conclusion

Over the past two decades, the gap between formal and descriptive approaches to inductive reasoning has been growing ever wider. Descriptive studies have accumulated more and more evidence that induction draws on systems of rich conceptual knowledge and that patterns of inference vary dramatically depending on the inductive context. Formal models, however, have struggled to account for these findings. Our work suggests how this gap between formal and descriptive approaches can be bridged. Bayesian models of inductive reasoning rely on prior distributions, and we showed how these priors can capture sophisticated and varied forms of background knowledge.

This article presents a unifying Bayesian framework for inductive reasoning and applies it to four different kinds of reasoning: taxonomic reasoning, spatial reasoning, threshold reasoning, and causal reasoning. These case studies suggest that our framework will be useful for explaining a broad range of knowledge-based inferences, but further applications are needed to support this claim. In particular, the models presented here rely on background knowledge that is relatively simple compared to the intuitive theories that support people's inductive inferences. Capturing intuitive theories with formal models remains a difficult challenge, but our work takes some initial steps towards this goal.

Our framework provides an alternative to the stand-off between structure and statistics that has polarized much of cognitive science. Our models rely on structured background knowledge, which can capture the content of intuitive theories, and on statistical inference, which explains how these theories are learned and used for induction. It is difficult—and unnecessary—to decide which of these components is the more important. Understanding how structure and statistics work together seems essential for explaining the flexibility, robustness, and sophistication of human reasoning.

## References

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology, 41,* 1–55.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Ashby, F. G. (1992). *Multidimensional models of perception and cognition.* Hillsdale, NJ: Erlbaum.

Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences, 21,* 547–609.

Ball, P. (1999). *The self-made tapestry: Pattern formation in nature.* Oxford: Oxford University Press.

Blok, S. V., Medin, D. L., & Osherson, D. (2007). Induction as conditional probability judgment. *Memory and Cognition, 35,* 1353–1364.

Bruner, J. S. (1973). Going beyond the information given. In J. M. Anglin (Ed.), *Beyond the information given* (pp. 218–243). New York: Norton.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Carey, S., & Spelke, E. S. (1996). Science and core knowledge. *Philosophy of Science, 63,* 515–533.

Chomsky, N. (1980). *Rules and representations.* Oxford: Basil Blackwell.

Chomsky, N. (1986). *Language and problems of knowledge: The Managua lectures.* Cambridge, MA: MIT Press.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8,* 240–247.

Davis, E. (1990). *Representations of commonsense knowledge.* Morgan Kaufmann.

Edwards, A. W. F. (1972). *Likelihood.* Cambridge: Cambridge University Press.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127,* 107–140.

Fernbach, P. (2006). Sampling assumptions and the size principle in property induction. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1287–1293). New York: Psychology Press.

Fiddick, L., Cosmides, L., & Tooby, J. (2000). No interpretation without representation: The role of domain-specific representations and inferences in the Wason selection task. *Cognition, 77,* 1–79.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28,* 3–71.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23,* 183–209.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias-variance dilemma. *Neural Computation, 4,* 1–58.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170.

Goodman, N. (1955). *Fact, fiction, and forecast.* Cambridge: Harvard University Press.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science, 32*(1), 108–154.

Grice, P. (1989). *Studies in the way of words.* Cambridge, MA: Harvard University Press.

Griffiths, T. L., Baraff, E. R., & Tenenbaum, J. B. (2004). Using physical theories to infer hidden causal structure. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th* Annual Conference of the Cognitive Science Society (pp. 500–505). Mahwah, NJ: Erlbaum.

Haussler, D., Kearns, M., & Schapire, R. E. (1994). Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Machine Learning, 14,* 83–113.

Heibeck, T., & Markman, E. (1987). Word learning in children: An examination of fast mapping. *Child Development, 58,* 1021–1024.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 248–274). Oxford, England: Oxford University Press.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(2), 411–422.

Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12,* 275–305.

Hirschfeld, L. (1996). *Race in the making: Cognition, culture, and the child's construction of human kinds.* Cambridge, MA: MIT Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery.* Cambridge, MA: MIT Press.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics, 17,* 754–755.

Jordan, M. I., & Sejnowski, T. J. (Eds.). (2001). *Graphical models.* Cambridge, MA: MIT Press.

Kahneman, D., & Tversky, A. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). New York: Cambridge University Press.

Karlin, S., & Taylor, H. (1975). *A first course in stochastic processes* (2nd ed.). San Diego, CA: Academic Press.

Keil, F. C. (1979). *Semantic and conceptual development.* Cambridge, MA: Harvard University Press.

Keil, F. C. (1981). Constraints on knowledge and cognitive development. *Psychological Review, 88,* 197–227.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kemp, C. (2008). *The acquisition of inductive constraints.* Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2008). Learning and using relational theories. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.). *Advances in Neural Information Processing Systems, Vol. 20.* (pp. 753–760). Cambridge, MA: MIT Press.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2004). Learning domain structures. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual conference of the Cognitive Science Society* (pp. 672–678). Mahwah, NJ: Erlbaum.

Kemp, C., Shafto, P., Berke, A., & Tenenbaum, J. B. (2007). Combining causal and similarity-based reasoning. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.). *Advances in Neural Information Processing Systems 19* (pp. 681–688). Cambridge, MA: MIT Press.

Kemp, C., & Tenenbaum, J. B. (2003). Theory-based induction. In R. Alterman & D. Kirsh (Eds.). *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 658–663). Mahwah, NJ: Erlbaum.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, USA, 105,* 10687–10692.

Kemp, C., & Tenenbaum, J. B. (in press). Structured models of semantic cognition. *Behavioral and Brain Sciences.*

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

Lawrence, N. D. (2004). Gaussian process models for visualization of high dimensional data. In *Advances in Neural Information Processing Systems 16.*

Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM, 38*(11), 33–38.

Macario, J. F. (1991). Young children's use of color in classification: Foods and canonically colored objects. *Cognitive Development, 6,* 17–46.

Marcus, G. F. (1991). *The algebraic mind: Integrating connectionism and cognitive science.* Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision.* San Francisco, CA: W. H. Freeman.

McCloskey, M. (1983). Intuitive physics. *Scientific American, 284,* 114–123.

McKenzie, C. R. M. (2003). Rational models as theories—not standards—of behavior. *Trends in Cognitive Sciences, 7,* 403–406.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. (2005). A relevance theory of induction. *Psychonomic Bulletin and Review, 10,* 517–532.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Murphy, G. L. (1993). Theories and concept formation. In *Categories and concepts: Theoretical views and inductive data analysis* (pp. 173–200).

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Newstead, S. E., & Griggs, R. A. (1984). Fuzzy quantifiers as an explanation of set inclusion performance. *Psychological Research, 46*(4), 377–388.

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

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning.* Oxford University Press.

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

Osherson, D. N., Stern, J., Wilkie, O., Stob, M., & Smith, E. E. (1991). Default probability. *Cognitive Science, 15,* 251–269.

Pinker, S. (1999). *Words and rules: The ingredients of language.* New York: Basic books.

Pruzansky, S., Tversky, A., & Carroll, J. D. (1982). Spatial versus tree representations of proximity data. *Psychometrika, 47*(1), 3–19.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

Rehder, B., & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50,* 264–314.

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

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). New York: Erlbaum.

Rumelhart, D. E., Lindsay, P., & Norman, D. A. (1972). A process model for long-term memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 197–246). New York: Academic Press.

Sanjana, N. E., & Tenenbaum, J. B. (2003). Bayesian models of inductive generalization. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Processing Systems 15* (pp. 51–58). Cambridge, MA: MIT Press.

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.

Shafto, P., Kemp, C., Baraff, E., Coley, J., & Tenenbaum, J. B. (2005). Context-sensitive induction. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 2003–2008). Mahwah: NJ: Erlbaum.

Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., & Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition, 109,* 175–192.

Shafto, P., Kemp, C., Mansinghka, V., Gordon, M., & Tenenbaum, J. B. (2006). Learning cross-cutting systems of categories. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 2146–2151). New York: Psychology Press.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210,* 390–398.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science, 237,* 1317–1323.

Simon, H. A. (1991). Cognitive architectures and rational analysis: Comment. In K. V. Lehn (Ed.), *Architectures for intelligence* (pp. 25–39). Hillsdale, NJ: Erlbaum.

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

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3–22.

Sloman, S. A., & Lagnado, D. A. (2005). The problem of induction. In R. Morrison & K. Holyoak (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 95–116). New York: Cambridge University Press.

Smith, E. E., Lopez, A., & Osherson, D. (1992). Category membership, similarity and naive induction. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of W. K. Estes* (pp. 181–206). Hillsdale: NJ: Erlbaum.

Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition, 49,* 67–96.

Smola, A., & Kondor, R. (2003). Kernels and regularization on graphs. In B. Schölkopf & M. K. Warmuth (Eds.). *Learning theory and kernel machines* (pp. 154–158). Berlin: Springer.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14,* 29–56.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–641.

Tenenbaum, J. B., Kemp, C., & Shafto, P. (2007). Theory-based Bayesian models of inductive reasoning. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental and computational approaches* (pp. 167–204). Cambridge: Cambridge University Press.

Tversky, A., & Hutchinson, J. W. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review, 93,* 3–22.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Williams, C. K. I., & Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(12), 1342–1351.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114,* 245–272.

Zhu, X., Lafferty, J., & Ghahramani, Z. (2003). *Semi-supervised learning: From Gaussian fields to Gaussian processes* (Tech. Rep. No. CMU-CS-03–175). Carnegie-Mellon University.

# Appendix A

## Covariance Matrices

This appendix defines the "raw covariance" of a feature matrix $D$. Suppose we are interested in four biological species: cheetah, lion, gorilla, and monkey ($x_1$ through $x_4$). We consider a feature matrix $D$ with two continuous features, one per column:

$$D = \begin{bmatrix} 0.6 & -0.4 \\ 0.4 & -0.6 \\ -0.6 & 0.4 \\ -0.4 & 0.6 \end{bmatrix}$$

Each feature is measured relative to the average value across all animals, which means that some entries in the matrix will be negative and others will be positive. Binary features can also be considered, but we use continuous features here as a reminder that our model relies on Gaussian distributions.

The raw covariance $K = 1/2DD^{\mathrm{T}}$ is a 4 by 4 matrix where entry $K_{ij}$ indicates whether species $i$ and $j$ tend to have feature values that vary together:

$$K = \frac{1}{2}DD^{T} \tag{16}$$

$$= \frac{1}{2}\begin{bmatrix} 0.6 & -0.4 \\ 0.4 & -0.6 \\ -0.6 & 0.4 \\ -0.4 & 0.6 \end{bmatrix}\begin{bmatrix} 0.6 & 0.4 & -0.6 & -0.4 \\ -0.4 & -0.6 & 0.4 & 0.6 \end{bmatrix} \tag{17}$$

$$= \begin{bmatrix} 0.26 & 0.24 & -0.26 & -0.24 \\ 0.24 & 0.26 & -0.24 & -0.26 \\ -0.26 & -0.24 & 0.26 & 0.24 \\ -0.24 & -0.26 & 0.24 & 0.26 \end{bmatrix} \tag{18}$$

Note, for instance, that $K_{12}$ is high, indicating that $x_1$ and $x_2$ have feature values that tend to vary together. Similarly, $K_{34}$ indicates that $x_3$ and $x_4$ have feature values that tend to vary together.

If the features in $D$ are generated from a Gaussian distribution with *zero* mean and unknown covariance $K$, the raw covariance is the maximum likelihood estimator of $K$. This raw covariance is different from the "empirical covariance" found in some textbooks, which is the maximum likelihood estimator if the features in $D$ are generated from a Gaussian distribution with *unknown* mean and unknown covariance. The two estimators coincide if each row of $D$ has a mean of zero. In our analyses, we normalize $D$ so that the mean value across the entire matrix is zero. In this case, the raw covariance and the empirical covariance are likely to be similar but not identical, and deciding to work with one rather than the other should make little difference for our purposes. For instance, the empirical covariance of our small matrix $D$ is very similar to the $K$ in Equation 18, except that each entry has an absolute value of 0.25.

We choose to work with the raw covariance rather than the empirical covariance in order to keep our models as simple as possible. All of our models that rely on the diffusion process (including the taxonomic model, the spatial model, and the raw covariance model) assume that continuous features are generated from a Gaussian distribution with zero mean. All of these models can therefore be compared on an equal footing.

# Appendix B

## Covariance-Based Property Induction

This appendix provides a detailed description of the taxonomic and spatial models. Both models capture similarity-based reasoning using a covariance matrix (Appendix A) defined over a structure $S$. More precisely, both models use covariance matrices to capture the idea that the features of a given object can be accurately predicted given the features of a nearby object in structure $S$.

### Taxonomic Model

Given a tree structure $S$, the prior $p(f)$ for our taxonomic model captures the intuition that high-probability features should be smooth over the tree. As Equations 7 and 8 suggest, we assume that binary features $f$ are generated by thresholding continuous features $y$ which are sampled from a zero-mean Gaussian with covariance matrix $K$. Here we follow Zhu, Lafferty, and Ghahramani (2003) and define the covariance matrix $K$ in a way which ensures that features tend to be smooth over a given graph structure. We consider a running example where $S$ is a tree structure defined over three categories: $x_1$, $x_2$ and $x_3$ (Figure B1). Our goal is to define a prior $p(y|S)$ on continuous feature vectors $y$ that assign a value to every node in the tree, including the internal node. Because the categories lie at the leaves, in practice we are primarily interested in the prior distribution induced over the leaf nodes.

Any graph can be represented as a matrix $S$ where $s_{ij} = \frac{1}{l_{ij}}$ if nodes $i$ and $j$ are joined by an edge of length $l_{ij}$ and $s_{ij} = 0$ otherwise. Assuming that all edges in our four-node tree have unit length, the matrix for this tree is

$$S = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \tag{19}$$

A generative model for y that favors features which are smooth over the graph S is given by

$$p(y|S) \propto \exp\left( -\frac{1}{2} \sum_{i,j:i<j} s_{ij}(y_i - y_j)^2 \right) \tag{20}$$

The sum includes a term for each pair of nodes $(i, j)$ where $i < j$, but the only pairs that make a contribution are those where $i$ is adjacent to $j$ in the graph and $s_{ij}$ is therefore nonzero. In other words, there is a contribution to the sum for each edge in graph $S$. From Equation 20 it is clear that the generative model penalizes a feature vector $y$ whenever $y_i \neq y_j$ and $i$ and $j$ are adjacent in the graph, and the penalty increases as the edge between $i$ and $j$ becomes shorter (i.e., $s_{ij}$ increases).

Figure B1 shows probabilities $p(y|S)$ for 16 possible features $y$ defined over our four node tree. To keep this example simple, each feature shown assigns a value of $-0.5$ (gray) or 0.5 (white) to each node in the tree, but in general each feature value can be any real number. Note that the features with highest probability according to Equation 20 are smooth over the tree: in other words, the high-probability features tend to assign similar values to any pair of adjacent nodes.

Equation 20 can be written as

$$p(y|S) \propto \exp\left( -\frac{1}{2} y^{\top} \Delta y \right) \tag{21}$$

where $\Delta$ is the graph Laplacian: $\Delta = \mathrm{G} - \mathrm{S}$ where G is a diagonal matrix with entries $\mathrm{g}_{ii} = \Sigma_{\mathrm{j}}\, s_{ij}$. In the case of our four-node graph,

$$\Delta = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ -1 & -1 & -1 & 3 \end{bmatrix}. \tag{22}$$

Note, for instance, that the probability of the second feature y in Figure B1 is

$$p(y|S) \propto \exp\left( -\frac{1}{2} [0.5 \quad 0.5 \quad 0.5 \quad -0.5] \times \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ -1 & -1 & -1 & 3 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \\ -0.5 \end{bmatrix} \right) \tag{23}$$

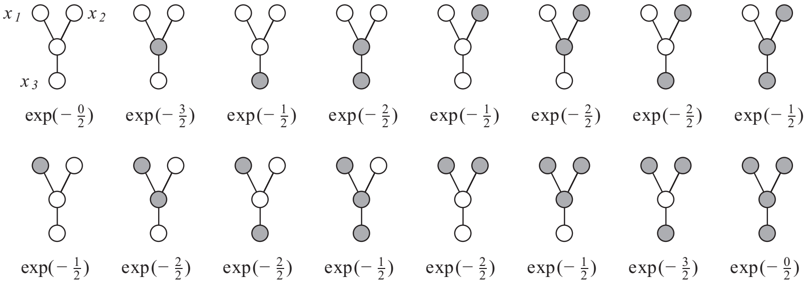

*Figure B1.* Sixteen feature vectors $y$ defined over a four node tree. The leaf nodes represent three categories, $x_1$, $x_2$ and $x_3$. Gray nodes have value $-0.5$, and white nodes have value 0.5. The prior probability of each feature is proportional to the quantity shown below each diagram. Of the sixteen features shown here, the two with highest probability are those that are perfectly smooth (top left and bottom right).

(*Appendixes continue*)

$$= \exp\left(-\frac{3}{2}\right) \tag{24}$$

Equation 21 is similar to the equation for a multivariate Gaussian distribution. If $y$ is distributed according to a Gaussian distribution with mean 0 and covariance matrix $K$, then

$$p(y) \propto \exp\left(-\frac{1}{2}y^{\mathrm{T}}K^{-1}y\right) \tag{25}$$

Comparing Equations 25 and 21 suggests that our distribution $p(y|S)$ is equivalent to a Gaussian distribution with zero mean and covariance matrix $K = \Delta^{-1}$. In other words, our preference for smoothness (Equation 20) has led to a Gaussian prior $p(y|S)$ where the covariance matrix $K$ captures the expectation that features $y$ tend to be smooth over graph $S$.

One technical detail remains to be addressed. Zhu et al. (2003) point out that a Gaussian prior with covariance matrix $K = \Delta^{-1}$ is improper. Note that any vector $y$ has the same probability when shifted by a constant, which effectively means that the variance of each $y_i$ is infinite. We obtain a proper prior by assuming that feature value $y_i$ at any node has an a priori variance of $\sigma^2$:

$$y|S \sim N\left(0, \left(\Delta + \frac{1}{\sigma^2}I\right)^{-1}\right) \tag{26}$$

where $I$ is the identity matrix. Comparing Equations 26 and 7, we see that the graph-based covariance $K$ is defined as $\left(\Delta + \frac{I}{\sigma^2}I\right)^{-1}$.

## Spatial Model

As described in the main text, the spatial model is identical to the taxonomic model, except that the covariance matrix $K$ is defined over a two-dimensional space instead of a tree structure. The covariance matrix $K$ produced by Equation 9 is known as the Laplacian kernel, but there are other kernels that satisfy the intuition that features should be smooth over an underlying multidimensional space. For machine-learning applications of Equation 7 it is common to use a covariance matrix based on the squared distance between categories:

$$K_{ij} \propto \exp\left(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2\right)$$

(Lawrence, 2004). There are two main reasons why we use the Laplacian kernel instead. First, the Laplacian kernel can be viewed as an extension of the graph-based covariance matrix defined in Equation 26. If we define a covariance matrix over a regular grid using Equation 26, the covariance approaches the covariance in Equation 9 as the grid becomes infinitely fine. The technical tools needed to prove this result are described by Smola and Kondor (2003).

Second, the Laplacian kernel is roughly consistent with the work of Shepard (1987) who shows that over a broad range of contexts, generalization is an exponential function of distance in a spatial representation. When $K$ is the Laplacian kernel, the $y$ values in Equation 7 decay exponentially with distance in the underlying space: if we observe $y_i$ (the value of $y$ for category $x_i$), the mean value of $y_j$ is $y_i \exp\left(-\frac{1}{\sigma}\|x_j - x_i\|\right)$. Even though the $y$ values decay exponentially, the decay in generalization probabilities $p(f_{x_j} = 1|f_{x_i} = 1)$ is only approximately exponential, because it also depends on the thresholding component of our generative model. Figure 6 confirms that the generalization curve for the spatial model is roughly exponential.

Equation 9 is appropriate for two-dimensional spaces, but different constants are needed for spaces of different dimensionality. For one-dimensional spaces, we use

$$K_{ij} = \frac{\sigma}{2}\exp\left(-\frac{1}{\sigma}\|x_i - x_j\|\right).$$

## Extensions

So far we have focused on the covariance matrix $K$, but the mean of the distribution in Equation 26 is also important. Setting the mean to zero implies that the base rate (the probability that a novel property applies to a randomly chosen species) is 0.5. Different properties may have different base rates, and adjusting the mean for different properties can explain why arguments using the property "has exactly 24 chromosomes" are judged weaker than arguments using "has more than 20 chromosomes" (Blok et al., 2007). Adjusting the mean is also important when we have background knowledge suggesting that some categories are more likely than others to have the novel property. Our threshold model, for instance, uses a mean vector $\mu$ that captures knowledge of this sort.

## Qualitative Phenomena

The taxonomic model and the spatial model both rely on a diffusion process (Equation 7), and several qualitative properties of this process are summarized in Table 1. Models that rely on the diffusion process will always satisfy the symmetry condition regardless of whether this process is defined over a tree, a low-dimensional space or some other kind of structure. If $f_a$ and $f_b$ represent feature values at two locations $a$ and $b$ in a structure,

$$p(f_b = 1|f_a = 1)p(f_a = 1) = p(f_a = 1|f_b = 1)p(f_b = 1)$$
$$\therefore\ p(f_b = 1|f_a = 1) = p(f_a = 1|f_b = 1)$$

The second line follows because $p(f_a = 1) = p(f_b = 1)$. In particular, because the diffusion process uses a Gaussian with zero mean, $p(f_a = 1) = p(f_b = 1) = 0.5$.

It is also straightforward to show that the distance effect always holds when the diffusion process is defined over a continuous space. If the diffusion process is defined over a graph, we know of no general result that specifies when the distance effect will hold. We believe, however, that the distance effect is typically but not always present.

## Implementing Our Models

Given the tree in Figure 3a, the generative model of Equations 7-8 induces a distribution over the leaves of the tree, and this distribution serves as the prior for each data set in Figure 7. We compute the predictions of the taxonomic model by drawing $10^6$ features from the prior and using this sample to approximate the true prior. Unless specified otherwise, all probabilistic models in this article were implemented similarly.

The taxonomic model has a single free parameter—the $\sigma$ that appears in Equation 26—and all of the Gaussian models in this article use $\sigma = 5$. We chose $\sigma$ by searching for the setting that maximizes the average correlation for the taxonomic model on the data sets in Figure 7, the spatial model on the data sets in Figure 9, and the threshold model on the data sets in Figure 11. Setting $\sigma$ to 5 means that the vector $y$ in Equations 7 and 10 can depart substantially from its mean (when $\sigma = 5$, the a priori variance of $y_i$ at any node $i$ in the graph is 25). A $\sigma$ value this high is driven primarily by the requirements of the threshold model. When the a priori variance of $y_i$ is high, the threshold model can deal gracefully with arguments that violate its expectations: arguments, for example, which indicate that a category towards the left end of one of the linear structures in Figure 10 has the property of interest.

The plots in the first three columns in Figure 6 were computed using numerical integration. For instance, the white curves in Figure 6a were computed by approximating the integral

$$p(f_{x_i} = 1 | f_{x_*} = 1)$$

$$= \frac{1}{p(f_{x_*} = 1)} \int_{y_i, y_*} p(f_{x_i} = 1 | y_i) p(f_{x_*} = 1 | y_*) p(y_i, y_*) dy_i dy_*$$

where $y_*$ is the value of y at $x_*$, the location marked by an asterisk.

# Appendix C

## Structure Learning

This appendix describes the matrix $D$ of animals and features that was mentioned throughout the main text. We then show how our structured statistical framework can use this feature matrix to discover trees and spatial representations.

### Feature Data

Osherson et al. (1991) ran a task where participants were given 48 species and 85 features and asked to rate the strength of association between each animal and feature. The task included all species shown in Figure 3 except for cow and dolphin, and the features included behavioral and anatomical features such as "is gray," "is active," and "lives in water." Participants gave ratings on a scale that started at zero and had no upper bound. Ratings were linearly transformed to values between 0 and 100, then averaged.

The property induction data analyzed in Figure 7 involve ten species, but only eight of these species appear among the 48 considered by Osherson et al. (1991) (cow and dolphin are missing). We therefore ran a follow up experiment using the 10 species shown in Figure 5 and the same 85 features used by Osherson et al. (1991). Apart from the species considered, our task matched the original task as closely as possible. Thirteen members of the Massachusetts Institute of Technology community participated in our task, and we combined our data with the original feature matrix to produce a data set $D$ with 50 animals and 85 features.

### Learning Tree Structures

Our framework suggests how statistical inference can be used to acquire structured background knowledge, and this section describes our method for learning the tree structures in Figures 3a, 8b and 14. Any tree $S$ can be represented as a pair $(S_T, S_L)$ where $S_T$ is the topology of the tree and $S_L$ is a vector specifying the lengths of its branches. Given a data set $D$ of continuous features, we can search for the topology and branch lengths that maximize $p(S_T, S_L|D, T, F) \propto p(D|S_T, S_L, T)p(S_T, S_L|F)$, where $T$ is the diffusion process and $F$ indicates that the unknown structure $S$ must be a tree. The likelihood term $p(D|S_T, S_L, T)$ is specified by Equation 7, and the main text describes a simple prior on trees where the distribution on topologies $S_T$ is uniform and the branch lengths $S_T$ are generated from an exponential distribution. This approach has been successfully used to develop phylogenetic models (Huelsenbeck & Ronquist, 2001), and to learn a tree structure from the feature matrix $D$ considered in this article (Kemp et al., 2004).

Here, however, we take an alternative approach (Kemp & Tenenbaum, 2008) that can be extended more easily to handle the inferences shown in Figure 18c and Figure 18d. We first discover the best topology $S_T$, then fix this topology and search for the best set of branch lengths. To identify the best topology, we use a variant of Equation 14:

$$p(S_T|D, F, T) \propto p(D|S_T, T)p(S_T|F) \qquad (27)$$

Because our generative model for features is defined over trees with branch lengths, we compute $p(D|S_T, T)$ by integrating over all possible sets of branch lengths:

$$p(D|S_T, T) = \int p(D|S_L, S_T, T)p(S_L|S_T)dS_L \qquad (28)$$

where the prior $p(S_L|S_T)$ is the product of independent exponential priors on the length of each branch. The integral in Equation 28 can be efficiently approximated using the Laplace approximation (Kemp & Tenenbaum, 2008).

Equation 27 also includes a prior on topologies $p(S_T|F)$. We follow Kemp and Tenenbaum (2008) and use a prior that favors topologies with small numbers of nodes:

$$p(S_T|F) \propto \theta(1 - \theta)^{\frac{|S_T|}{2}} \qquad (29)$$

where $|S_T|$ is the number of nodes in $S_T$. The parameter $\theta$ determines the extent to which trees with many nodes are penalized. Like Kemp and Tenenbaum (2008) we set $\theta = 1 - e^{-3}$, which means that each additional node reduces the log probability of a structure by 3.

To identify the topology $S_T$ that maximizes $p(S_T|D, F, T)$ we implement a greedy search. The search is initialized with the simplest possible topology (a tree where all categories are attached

(*Appendixes continue*)

to a single internal node), and we repeatedly grow the tree by considering how best to split one of the current nodes. After each split, we attempt to improve the posterior $p(S_T|D, F, T)$ by trying several moves, including proposals that move a category from one node to another, and proposals that exchange two nodes. The search concludes once the posterior can no longer be improved. Once the best topology $S_T$ is discovered, a standard gradient-based search can identify the branch lengths $S_L$ that maximize $p(S_L|S_T, D, T)$. The output of our structure learning algorithm is a pair $(S_T, S_L)$: a pair which fully specifies a tree $S$ like the example in Figure 3a.

Because our tree-learning algorithm includes several ideas, it is useful to distinguish the most fundamental idea from those that are less important. Our central proposal is that a learner can discover a tree by applying the same idea used by our taxonomic model: the idea that biological properties are distributed smoothly over an underlying tree. To formalize this proposal, we described an algorithm that relies on a specific prior over trees (Equation 29) and a specific procedure for identifying the tree $S_T$ that maximizes the posterior probability in Equation 27. We believe that these modeling choices are sensible, but other modeling choices will be consistent with the central proposal already mentioned.

In this article we learn a tree structure from continuous data but use the tree to make inferences about binary properties. These modeling decisions appear suitable for the data sets we have considered, but we can also use our framework to learn structures from binary data, or to make inferences about continuous properties (e.g., “has a typical life span of *x* years in captivity,” where *x* can take different values for different species). Both of these cases can be handled by assuming that binary properties are generated according to Equations 7-8, and that continuous properties are generated according to Equation 7.

Structure learning algorithms make two important contributions to psychological research: they help to explain how humans acquire knowledge, and they serve as useful tools for exploring the nature of mental representations. Algorithms for analyzing similarity data (Shepard, 1980) often make the second kind of contribution, but have less to say about how humans acquire knowledge. In the biological domain, for example, the input that allows humans to acquire tree structures like Figure 3a is probably closer to a matrix of observable features than a matrix of pairwise similarity ratings. Similarity ratings, however, can be invaluable for reconstructing the representation that a given participant might be using.

Our approach to structure discovery is intended primarily as a computational theory of human learning, but can also be used to discover structure in similarity data. Under the Gaussian generative model of Equation 7, the expression for $p(D|S, T)$ includes only two components that depend on *D: m,* the number of features in *D,* and the raw covariance $\frac{1}{m}DD^{\mathrm{T}}$. If both of these components are provided, our structure learning method can still be applied even if none of the features in *D* is directly observed (Kemp & Tenenbaum, 2008). The raw covariance $\frac{1}{m}DD^{\mathrm{T}}$ can be interpreted as a similarity matrix, where the similarity of two species is proportional to the inner product of their feature vectors. The resulting similarity measure is closely related to cosine similarity, although it lacks the normalizing factor that appears in the definition of this quantity.

Because a similarity matrix can stand in for the raw covariance $\frac{1}{m}DD^{\mathrm{T}}$, we can use Equation 14 to discover a structure *S* given only a set of pairwise similarity ratings between the categories in a domain. The trees in Figures 8 and 14 were discovered using this approach, and in each case we set $m = 1000$. Previous researchers have used similarity data to build representations which are subsequently used for categorization or other forms of inductive reasoning (Nosofsky, 1986). Our framework supports this style of research, and Figures 9 and 15 show the performance of structured statistical models that rely on representations learned from similarity data. Wherever possible, however, we prefer to work with features rather than similarity data, because models that take observable features as their input are better able to explain how humans learn directly from experience.

Kemp and Tenenbaum (2008) provide a more complete description of our structure learning framework, but several remaining details should be mentioned here. Because Equation 7 is a Gaussian distribution with zero mean, our structure learning framework expects the mean entry in any given feature matrix to be zero. If a matrix *D* does not satisfy this criterion, we enforce it by adding a constant to the matrix. We also scale any given feature matrix so that the maximum entry in the raw covariance matrix is 1. This decision means that the raw covariance matrix corresponds relatively closely to a measure of similarity, and means that our structure learning algorithm can use the same hyperparameters regardless of whether it is analyzing feature data or similarity data. Finally, when analyzing similarity data, our approach assumes that a given similarity matrix is a covariance matrix. This assumption only makes sense if the matrix is positive definite. If a given similarity matrix does not satisfy this criterion, we enforce it by replacing all negative eigenvalues with zeroes.

## Learning Spatial Representations

The previous section described how our structured statistical framework can be used to discover tree structures, and two-dimensional representations can be discovered similarly. Our strategy for learning spatial representations is similar to methods like principal components analysis (PCA) and multidimensional scaling, but is related most closely to the work of Lawrence (2004). We search for the representation *S* that maximizes Equation 14, where the prior $p(S|F)$ is induced by independent Gaussian priors with zero mean and unit covariance over the location of each category in the underlying space. The basic goal of this approach is to find a two-dimensional representation such that nearby animals tend to have many features in common. We initialize the search process using a configuration found by PCA, and run gradient-based searches that start from many different perturbations of this initial solution. When applied to the animal-feature matrix *D* described in the previous section, our structure-learning algorithm discovers the two-dimensional representation in Figure 3b.

Received November 6, 2007
Revision received August 7, 2008
Accepted August 11, 2008 ■