# Theory-based Bayesian models of inductive reasoning

Joshua B. Tenenbaum, Charles Kemp & Patrick Shafto

## 1 Introduction

Philosophers since Hume have struggled with the logical problem of induction, but children solve an even more difficult task — the practical problem of induction. Children somehow manage to learn concepts, categories, and word meanings, and all on the basis of a set of examples that seems hopelessly inadequate. The practical problem of induction does not disappear with adolescence: adults face it every day whenever they make any attempt to predict an uncertain outcome. Inductive inference is a fundamental part of everyday life, and for cognitive scientists, a fundamental phenomenon of human learning and reasoning in need of computational explanation.

There are at least two important kinds of questions that we can ask about human inductive capacities. First, what is the knowledge on which a given instance of induction is based? Second, how does that knowledge support generalization beyond the specific data observed: how do we judge the strength of an inductive argument from a given set of premises to new cases, or infer which new entities fall under a concept given a set of examples? We provide a computational approach to answering these questions. Experimental psychologists have studied both the process of induction and the nature of prior knowledge representations in depth, but previous computational models of induction have tended to emphasize process to the exclusion of knowledge representation. The approach we describe here attempts to redress this imbalance, by showing how domain-specific prior knowledge can be formalized as a crucial ingredient in a domain-general framework for rational statistical inference.

The value of prior knowledge has been attested by both psychologists and machine learning theorists, but with somewhat different emphases. Formal analyses in machine learning show that meaningful generalization is not possible unless a learner begins with some sort of inductive bias: some set of constraints on the space of hypotheses that will be considered (Mitchell, 1997). However, the best known statistical machine-learning algorithms adopt relatively weak inductive biases and thus require much more data for successful generalization than humans do: tens or hundreds of positive and negative examples, in contrast to the human ability to generalize from just one or few positive examples. These machine algorithms lack ways to represent and exploit the rich forms of prior knowledge that guide people's inductive inferences, and that have been the focus of much attention in cognitive and developmental psychology under the name of "intuitive theories" (Murphy and Medin, 1985). Murphy (1993) characterizes an intuitive theory as "a set of causal relations that collectively generate or explain the phenomena in a domain." We think of a theory more generally as any system of abstract principles that generates hypotheses for inductive inference in a domain, such as hypotheses about the meanings of new concepts, the conditions for new rules, or the extensions of new properties in that domain. Carey (1985), Wellman and Gelman (1992), and Gopnik and Meltzoff (1997) emphasize the central role of intuitive theories in cognitive development, both as sources of constraint on children's inductive reasoning and as the locus of deep conceptual change. Only recently have psychologists begun to consider seriously the roles that these intuitive theories might play in formal models of inductive inference (Gopnik and Schulz, 2004; Tenenbaum,

Griffiths and Kemp, 2006; Tenenbaum, Griffiths and Niyogi, in press). Our goal here is to show how intuitive theories for natural domains such as folk biology can, when suitably formalized, provide the foundation for building powerful statistical models of human inductive reasoning.

Any familiar thing can be thought about in a multitude of ways, and different kinds of prior knowledge will be relevant to making inferences about different aspects of an entity. This flexibility poses a challenge for any computational account of inductive reasoning. For instance, a cat is a creature that climbs trees, eats mice, has whiskers, belongs to the category of felines, and was revered by the ancient Egyptians – and all of these facts could potentially be relevant to an inductive judgment. If we learn that cats suffer from a recently discovered disease, we might think that mice also have the disease; perhaps the cats picked up the disease from something they ate. Yet if we learn that cats carry a recently discovered gene, lions and leopards seem more likely to carry the gene than mice. Psychologists have confirmed experimentally that inductive generalizations vary in such ways, depending on the property involved. Our computational models will account for these phenomena by positing that people can draw on different prior knowledge structures – or different intuitive theories – within a single domain, and by showing how very different patterns of inference can arise depending on which of these theories is triggered.

Our models aim for both predictive and explanatory power. As in any mathematical modeling, we seek accounts that can provide close quantitative fits to human judgments across a range of different tasks or contexts, with a minimum number of free parameters or ad hoc assumptions. At the same time, we would like our models to explain why people make the inductive generalizations that they do make, and why these judgments are mostly successful in the real world – how people can reliably come to true beliefs about the world from very limited data. In the spirit of rational analysis (Anderson, 1990; Oaksford and Chater, 1998), or Marr's (1982) computational-theory level of analysis, we will assume that people's inductive capacities can be characterized as approximations to optimal inferences, given the structure of the environments and the task contexts that they have adapted to over the course of evolution and development. Our mission as modelers is then to characterize formally the nature of the optimal inference mechanism, the relevant aspects of environmental structure and task context, and the interaction between these components.

Our core proposal has two components. First, domain-specific knowledge that supports induction in different contexts can be captured using appropriate families of probabilistic models defined over structured representations: in particular, relational systems of categories such as taxonomic hierarchies or food webs. These structured probabilistic models are far from being complete formalizations of people's intuitive domain theories; they are minimalist accounts, intended to capture only those aspects of theories relevant for the basic inductive inferences we study. Second, knowledge in this form can support inductive generalization by providing the prior probabilities for a domain-general Bayesian inference engine. Both of these claims are necessary for explaining how people's inductive inferences can be so successful, and perhaps approximately optimal, with respect to the world that we live in. The structured representations of intuitive domain theories are important because the world contains genuine structure: a tree-structured representation of biological species is useful, for example, because it approximates the structure of the evolutionary tree. Bayesian inference is important because it provides a normative and general-purpose procedure for reasoning under uncertainty. Taking these two components together – rational domain-general statistical inference guided by appropriately structured intuitive domain theories – may help to explain the uniquely human capacity for learning so much about the world from so little experience.

Our work goes beyond previous formal models of induction which either do not address the rational statistical basis of people's inferences, or find it difficult to capture the effects of different kinds of knowledge in different inductive contexts, or both. In one representative and often-cited example, the similarity-coverage model of Osherson, Smith and colleagues, the domain-specific

knowledge that drives generalization is represented by a similarity metric (Osherson et al., 1990). As we will see below, this similarity metric has to be defined in a particular way in order to match people's inductive judgments. That definition appears rather arbitrary from a statistical point of view, and arbitrarily different from classic similarity-based models of other cognitive tasks such as categorization or memory retrieval (Nosofsky, 1986; Hintzman et al., 1978). Also, the notion of similarity is typically context-independent, which appears at odds with the context-dependent nature of human inductive reasoning. Even if we allow some kind of context-specific notion of similarity, a similarity metric seems too limited a representation to carry the richly structured knowledge that is needed in some contexts, or even simple features of some reasoning tasks such as the strong asymmetry of causal relations. In contrast, the knowledge that drives generalization in our theory-based Bayesian framework can be as complex and as structured as a given context demands.

The plan of this chapter is as follows. Section 2 provides a brief review of the specific inductive tasks and phenomena we attempt to account for, and Section 3 briefly describes some previous models that have attempted to cover the same ground. Section 4 introduces our general theory-based Bayesian framework for modeling inductive reasoning, and describes two specific instantiations of it that can be used to model inductive reasoning in two important natural settings. Section 5 compares our models and several alternatives in terms of their ability to account for people's inductive judgments on a range of tasks. Section 6 concludes and offers a preview of ongoing and future work.

## 2 Property induction

This section reviews the basic property induction task and introduces the core phenomena that our models will attempt to explain. Following a long tradition (Rips, 1975; Carey, 1985; Osherson et al., 1990; Sloman, 1993; Heit, 1998), we will focus on inductive arguments about the properties of natural categories, in particular biological species categories. The premises of each argument state that one or more specific species have some property, and the conclusion (to be evaluated) asserts that the same property applies to either another specific species or a more general category (such as all mammals). These two kinds of arguments are called *specific* and *general* arguments, respectively, depending only on the status of the conclusion category.

We use the formula $P_1, \ldots P_n \stackrel{prop}{\longrightarrow} C$, to represent an $n$-premise argument where $P_i$ is the $i$th premise, $C$ is the conclusion and *prop* indicates the property used. We will often abbreviate references to these components of an argument. For example, the argument

Gorillas have T4 hormones
Squirrels have T4 hormones
---
All mammals have T4 hormones

might be represented as $gorillas, squirrels \stackrel{T4}{\longrightarrow} mammals$.

The most systematic studies of property induction have used so-called "blank properties". For arguments involving animal species, these are properties that are recognized as biological but about which little else is known — for example, anatomical or physiological properties such as "has T4 hormones" or "has sesamoid bones". As these properties are hardly "blank" – it is important that people recognize them as deriving from an underlying and essential biological cause – we will instead refer to them as "generic biological properties".

For this class of generic properties, many qualitative reasoning phenomena have been described: 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* $\xrightarrow{T4}$ *dolphins* is weaker than *seals* $\xrightarrow{T4}$ *dolphins*. For general arguments, *typicality* is the effect that argument strength increases as the premises become more typical of the conclusion category. For example, *seals* $\xrightarrow{T4}$ *mammals* is weaker than *horses* $\xrightarrow{T4}$ *mammals*, since 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, cows, rhinos* $\xrightarrow{T4}$ *mammals* is weaker than *horses, seals, squirrels* $\xrightarrow{T4}$ *mammals*.

Explaining inductive behavior with generic biological properties 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* $\xrightarrow{T4}$ *mammals* has premises that are quite diverse, but not very typical of the conclusion. For this reason, rather than trying to account for isolated qualitative contrasts between pairs of arguments, we will assess the performance of computational models in terms of how well they can predict relative argument strengths across multiple datasets each containing a large number of arguments of the same form.

The strength of an argument depends critically on the property involved, because changing the property will often alter the inductive context. Many researchers have described related effects (Gelman and Markman, 1986; Heit and Rubinstein, 1994; Shafto and Coley, 2003; Smith et al., 1993), and we mention just three of them. Gelman and Markman (1986) showed that children reason differently about biological properties (eg "has cold blood") and physical properties (eg "weighs one ton") — for example, *brontosaurus* $\xrightarrow{cold\ blood}$ *triceratops* is relatively strong, but *brontosaurus* $\xrightarrow{one\ ton}$ *triceratops* is relatively weak. Heit and Rubinstein (1994) showed that anatomical or physiological properties and behavioral properties are treated differently by adults. While anatomical or physiological properties typically support default, similarity-like patterns of inductive reasoning, behavioral properties may depend less on generic similarity and more on shared ecological roles. Finally, Shafto and Coley (2003) argue that disease properties may draw on causal knowledge about predator-prey interactions, and thus may be treated differently from arguments about generic biological properties. For example, *salmon* $\xrightarrow{leptospirosis}$ *grizzly bears* may be judged stronger than *grizzly bears* $\xrightarrow{leptospirosis}$ *salmon*, where *leptospirosis* stands for "carry leptospirosis bacteria". This asymmetry has no justification in terms of the similarity between salmon and grizzly bears, which is presumably symmetric or nearly so, but it seems sensible from the perspective of causal reasoning: knowing that grizzly bears eat salmon, it seems more likely that grizzly bears would catch some disease from salmon than that any specific disease found in grizzly bears necessarily came from the salmon that they eat.

Our aim has been to develop a unifying computational framework that can account for many of the phenomena mentioned above. We will focus in this chapter on modeling reasoning about two kinds of properties: the classic setting of generic biological properties, and causally transmitted properties such as diseases (Shafto and Coley, 2003) that give rise to very different patterns of judgment. Before introducing the details of our framework, we summarize several existing models of property induction and describe how we hope to improve on them.

# 3 Previous models

The tradition of modeling property induction extends at least as far back as the work of Rips (1975). Here we summarize a few of the more prominent mathematical models that have been developed in the intervening 30 years.

## 3.1 Similarity-coverage model

The similarity-coverage model (SCM) of Osherson et al. (1990) is perhaps the best known mathematical model of property induction. It predicts the strength of inductive arguments as 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. The SCM has some appealing properties. It makes accurate predictions for generic biological properties, and it uses a simple equation that predicts many different kinds of judgments with a minimum of free parameters. Yet the SCM has two major limitations. First, it can only use domain knowledge that takes the form of pairwise similarities or superordinate taxonomic categories. The model is therefore unable to handle inductive contexts that rely on knowledge which cannot be expressed in this form. Second, the SCM lacks a principled mathematical foundation: the accuracy of its predictions depends critically on certain arbitrary choices which specify the mathematical form of the model.

This arbitrariness shows up most clearly in the formal definition of coverage: the average over all instances $i$ in the superordinate class of the *maximal* similarity between $i$ and the examples in the premise set. We refer to this (standard) version of SCM as "MaxSim." Osherson et al. (1990) also consider a variant we call "SumSim," in which coverage is defined by averaging the *summed* similarity to the examples over all instances of the superordinate class. Generalization based on the summed similarity to exemplars or weight traces is the foundation for many other successful models of categorization, learning and memory (Nosofsky, 1986; Kruschke, 1992; Hintzman et al., 1978), and can be interpreted in rational statistical terms as a version of nonparametric density estimation (Ashby and Alfonso-Reese, 1995; Silverman, 1986). Yet despite these precedents for using a summed-similarity measure, Osherson et al. (1990) advocate MaxSim, or some weighted combination of MaxSim and SumSim – perhaps because SumSim performs dramatically worse than MaxSim in judging the strength of general arguments (see Section 5). Since Osherson et al. (1990) do not explain why different measures of setwise similarity are needed in these different tasks, or why SumSim performs so much worse than MaxSim for inductive reasoning, the SCM is less principled than we might like.

## 3.2 Feature-based models

As Goodman (1972) and Murphy and Medin (1985) have argued, similarity is a vague and elusive notion, and it may be meaningless to say that two objects are similar unless a respect for similarity has been specified. Instead of founding a model directly on similarity judgments, an alternative is to start with a collection of object features, which might plausibly be observable perceptually or have been previously learned. [1] 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 learn directly from experience.

[1] Note that a feature-based version of the SCM is achieved if we define the similarity of two objects as some function of their feature vectors. Section 5 assesses the performance of this model.

The feature-based model of Sloman (1993) computes inductive strength as a normalized measure of feature overlap between conclusion and example categories. Sloman (1993) presents a quantitative comparison with the SCM: the results are not conclusive, but suggest that the model does not predict human judgments as accurately as the SCM. The model, however, predicts some qualitative phenomena that the SCM cannot explain. 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. This connectionist approach is more ambitious than any of the others we describe, and Rogers and McClelland (2004) apply their model to a diverse set of semantic phenomena. One of the applications is a property induction task where the model makes sensible qualitative predictions, but there has been no demonstration so far that the model provides good quantitative fits to human judgments.

From our perspective, both feature-based models share the limitations of the SCM. Despite the range of applications in Rogers and McClelland (2004), it is not clear how either model can be extended to handle causal settings or other inductive contexts that draw on sophisticated domain knowledge. Both models also include components that have been given no convincing justification. The model of Sloman (1993) uses a particular mathematical measure of feature overlap, but it is not clear why this should be the right measure to use. Rogers and McClelland (2004) provide no principled explanation for the architecture of their network or their strategy for computing the strength of inductive arguments, and their model appears to rely on several free parameters.

### 3.3 A Bayesian analysis

Heit (1998) presented a computational theory where property induction is modeled as Bayesian inference. This inference engine is essentially the same as we describe below in Section 4.1. Applying a Bayesian analysis to any specific case of property induction requires a prior distribution over hypotheses about the extension of the property in question. Heit does not specify a formal method for generating priors, nor does he test his model quantitatively against any specific judgments. He shows that it captures several qualitative phenomena if it is supplied with the right kinds of priors, and that appropriate priors could allow it to handle both blank and non-blank properties. He also suggests how priors could be extracted from long-term memory: the probability of a hypothesis could be proportional to the number of familiar features that can be retrieved from memory and that have the same extension as that hypothesis. But it is far from clear that this suggestion would, if implemented, yield appropriate priors; as we show below, a simple version of this idea does not perform nearly as well as the SCM's gold standard in predicting human judgments.

Our framework adopts a Bayesian approach to inference like Heit's, but we emphasize the importance of modeling the form and the origins of appropriate priors. A formal account of how the learner's prior is structured and where it comes from provides two distinct advantages. First, it leads to strong quantitative models, predicting people's inductive judgments as well or better than any previous approach. More importantly, it adds genuine explanatory power. 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 and how that knowledge is used. It has long been argued that different inductive contexts lead to quite different patterns of generalization behavior, but whether this is due to the operation of different kinds of knowledge, different mechanisms of reasoning, or both, has not been so clear. We will argue that a single general-purpose Bayesian reasoning mechanism may be sufficient, by showing explicitly how to generate priors that can capture two important and very different kinds of domain knowledge, and that can strongly predict people's judgments in appropriately different inductive contexts.

# 4 The theory-based Bayesian framework

Our framework includes two components: a Bayesian engine for inductive inference, and a language for specifying relevant aspects of domain theories and using those theories to generate prior probability distributions for the Bayesian inference engine. The Bayesian engine reflects domain-general norms of rational statistical inference and remains the same regardless of the inductive context. Different domain theories may be appropriate in different inductive contexts, but they can often be formalized as instances of a single unifying scheme: a probabilistic process, such as diffusion, drift or transmission, defined over a structured representation of the relevant relations between categories, such as taxonomic or ecological relations.

## 4.1 The Bayesian inference engine

Assume that we are working within a finite domain containing $n$ categories. We are interested in a novel property, $Q$, that applies to some unknown subset of these categories. Let $H$ be the hypothesis space of all logically possible extensions for $Q$ — the set of all possible subsets $h$ of categories in the domain, each of which could a priori be the extension of the novel property. Since there are $n$ categories, the number of hypotheses is $2^n$. To each hypothesis we assign a prior probability $p(h)$, where $p(h)$ is the probability that $h$ includes all and only the categories with property $Q$.

Suppose now that we observe $X$, a set of $m$ labeled objects where the labels indicate whether each category in $X$ has property $Q$. We want to compute $p(y \text{ has } Q|X)$, the probability that object $y$ has property $Q$ given the examples $X$. Summing over all hypotheses in $H$, we have:

$$p(y \text{ has } Q|X) = \sum_{h \in H} p(y \text{ has } Q, h|X) \tag{1}$$

$$= \sum_{h \in H} p(y \text{ has } Q|h, X)p(h|X). \tag{2}$$

Now $p(y \text{ has } Q|h, X)$ equals one if $y \in h$ and zero otherwise (independent of $X$). Thus:

$$p(y \text{ has } Q|X) = \sum_{h \in H: y \in h} p(h|X) \tag{3}$$

$$= \sum_{h \in H: y \in h} \frac{p(X|h)p(h)}{p(X)} \tag{4}$$

where the last step follows from Bayes' rule.

The numerator in Equation 4 depends on the prior $p(h)$, as well as on the likelihood $p(X|h)$, the probability of observing the labeled examples $X$ given that $h$ is the true extension of $Q$. The likelihood $p(X|h)$ should in general depend on the process assumed to generate the observations in $X$. Here, for simplicity we will assume that $p(X|h) \propto 1$ for all hypotheses consistent with $X$, and $p(X|h) = 0$ otherwise.[2] A hypothesis $h$ for the extension of property $Q$ is consistent with a set of labeled examples $X$ if $h$ includes all positively labeled categories in $X$ and excludes all negatively

[2]More complex sampling models could be appropriate in other circumstances, and are discussed in Tenenbaum and Griffiths (2001) and Kemp and Tenenbaum (2003). For instance, an assumption that examples are randomly drawn from the true extension of $Q$ might be particularly important when learning word meanings or concepts from ostensive examples (Tenenbaum and Xu, 2000; Xu and Tenenbaum, in press).

labeled categories in $X$. Then Equation 4 is equivalent to

$$p(y \text{ has } Q|X) = \frac{\sum_{h \in H: y \in h, h \text{ consistent with } X} p(h)}{\sum_{h \in H: h \text{ consistent with } X} p(h)} \quad (5)$$

which is the proportion of hypotheses consistent with $X$ that also include $y$, where each hypothesis is weighted by its prior probability $p(h)$. The probability of generalizing to $y$ will thus be high to the extent that it is included in most of the high-prior-probability hypotheses that also include the observed examples $X$.

Other inferences can be formulated similarly. For example, the probability that all categories in a larger class $Y$ (e.g., *all mammals*) have property $Q$ could be formalized as:

$$p(Y \text{ has } Q|X) = \frac{\sum_{h \in H: Y \subset h, h \text{ consistent with } X} p(h)}{\sum_{h \in H: h \text{ consistent with } X} p(h)} \quad (6)$$

Note that a Bayesian approach needs no special purpose rules for dealing with negative evidence or arguments with multiple premises. Once the prior distribution $p(h)$ and the likelihood $p(X|h)$ have been specified, computing the strength of a given argument involves a mechanical application of the norms of rational inference. Since we have assumed a simple and domain-general form for the likelihood above, our remaining task is to specify appropriate domain-specific prior probability distributions.

### 4.2 Theory-based priors

Generating the prior distributions used in Equations 5 and 6 appears to be a difficult problem – for either the cognitive modeler or the human reasoner. Somehow we need to specify $2^n$ numbers: $p(h)$ for each of the logically possible hypotheses $h$ in $H$. We cannot simply assign all hypotheses equal prior probability; without any inductive biases, meaningful generalization would be impossible (Mitchell, 1997). Explicitly enumerating the priors for all $2^n$ hypotheses is also not an option. This would introduce far more degrees of freedom into the model than we could ever hope to test empirically. More importantly, it would fail to capture the most interesting aspect of these priors – that they are not just lists of numbers, but rather the products of abstract systems of knowledge, or intuitive theories. Induction with different kinds of properties – such as anatomical features, behavioral tendencies, or disease states of animal species – will require different kinds of priors because we have qualitatively different kinds of knowledge that we bring to bear in those contexts. Our priors for induction can change when we learn new facts, but the biggest changes come not from statistical observations that might simply favor one hypothesis over another. Priors can change most dramatically, and can change globally across a large slice of the hypothesis space, when we acquire qualitative knowledge that alters our intuitive domain theories: when we learn about a new species with unexpected characteristics, such as whales or ostriches, or we learn something surprising about how various species might be related, such as that whales and dolphins are mammals, or we learn some new principle about how properties are distributed over species, such as that diseases tend to spread through physical contact or food.

The heart of our proposal is a way to understand formally how intuitive domain theories can generate the prior probabilities needed for induction. Two aspects of intuitive theories are most

relevant for constructing priors for property induction: representations of how entities in the domain are related to each other, and processes or mechanisms operating over those relational structures that give rise to the distribution of properties over entities. To be concrete, we will assume that each of the $n$ categories in a domain can be represented as a node in a relational structure, such as a directed or undirected graph. Edges in the graph represent relations that are relevant for determining inductive potential, such as taxonomic or causal relations among categories. Priors are generated by a stochastic process defined on this graph, such as a diffusion process, a drift process, or a noisy transmission process. These processes can be used to capture general beliefs about how properties of various types tend to be distributed over related categories in the domain. Once we have sufficiently characterized the relational structure and the stochastic generating process, that will fully specify the $2^n$ numbers in the prior. By choosing different kinds of structures and stochastic processes we can capture different kinds of knowledge and account for qualitatively different patterns of inductive reasoning. In this chapter we describe and test two such models, one for reasoning about generic biological properties such as anatomical and physiological features, and another for reasoning about causally transmitted properties such as diseases.

### 4.2.1 A theory for generic biological properties

The prior distribution for default biological reasoning is based on two core assumptions: the *taxonomic principle* and the *mutation principle*. The taxonomic principle asserts that species belong to groups in a nested hierarchy, and more precisely, that the taxonomic relations among species can be represented by locating each species at some leaf node of a rooted tree structure. Tree-structured taxonomies of species appear to be universal across cultures (Atran, 1998), and they also capture an important sense in which species are actually related in the world: genetic relations due to the branching process of evolution. Outside of intuitive biology, tree-structured taxonomies play a central role in organizing knowledge about many systems of natural-kind and artifact categories (Rosch, 1978), as well as the meanings of words that label these categories (Markman, 1989; Tenenbaum and Xu, 2000).

The structures of people's intuitive taxonomies are liable to deviate from scientific phylogenies in non-negligible ways, since people's theories are based on very different kinds of observations and targeted towards predicting different kinds of properties. Hence we need some source of constraint besides scientific biology in order to generate the particular tree-structured taxonomy that our model will use. We have explored several different approaches to reconstructing taxonomic trees that best characterize people's intuitive theories. One possibility is to perform hierarchical clustering on people's explicit judgments of similarity for all pairs of species in the domain. Hierarchical clustering could also be applied to more implicit measures of psychological distance between species: for example, we could represent each animal using a set of behavioral and morphological features (e.g., "lives in water", "has a tail"), and set the distance between two animals to be the distance between their feature vectors. We could also use more structured domain knowledge that might obviate the need for any bottom-up clustering. Both direct judgments of pairwise similarity and ratings of feature-category associations have been collected for many of the standard domains of animal species used in studying reasoning about generic biological properties (Osherson et al., 1990), and these data sets provide a convenient basis for comparing different modeling frameworks on equal grounds. Figure 1a shows a taxonomic tree that was reconstructed for ten mammal species, based on hierarchical clustering over a set of features collected by Osherson and colleagues (and slightly augmented as described below in Section 5.1).

Some simple and intuitive prior distributions $p(h)$ can be generated using the taxonomic principle alone. For instance, we could assign a uniform probability to each hypothesis corresponding

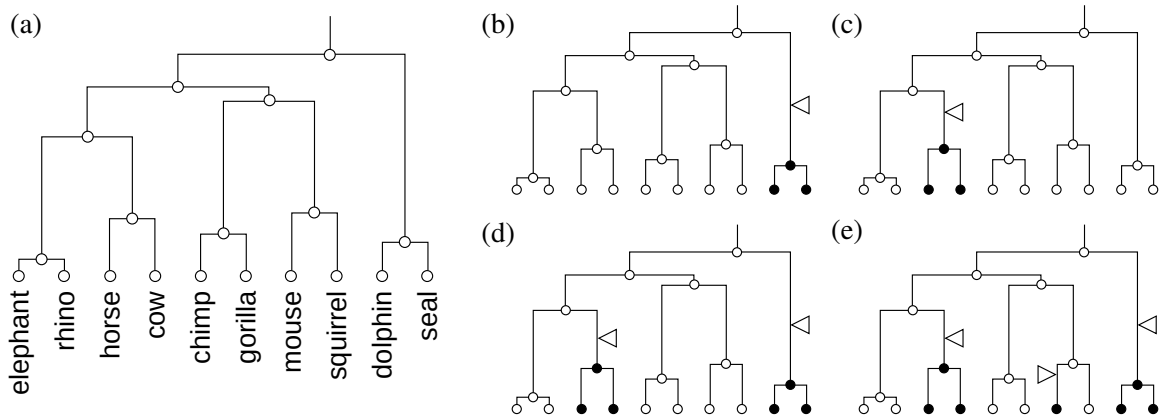


Figure 1: (a) A folk taxonomy of mammal species. (b-e) Examples of mutation histories.

to one of the 19 taxonomic clusters (including singleton species) shown in the tree, and zero probability to all other sets of species. We call this the "strict-tree model". It corresponds roughly to some informal accounts of taxonomically driven induction (Atran, 1995), and it can qualitatively reproduce some important phenomena, such as diversity-based reasoning. Essentially this prior has also been used successfully to explain how people learn words that label taxonomic object categoroies. But to see that it is not sufficient to explain biological property induction, compare the arguments *seals*, *squirrels* $\xrightarrow{T4}$ *horses* and *seals*, *cows* $\xrightarrow{T4}$ *horses*. The second appears stronger than the first, yet under the intuitive taxonomy shown in Figure 1, the strict-tree model assigns them both the same strength, since each set of premises is compatible with only one hypothesis, the set of all mammals.

The solution to this problem comes from realizing that in scientific biology, the assumption that every property is strictly confined to a single taxon is false. Properties arise randomly via mutations, and through a combination of chance and natural selection, two species may share a property even if it occurs in no common ancestor. Convergent evolution is particularly likely for survival-significant traits of interest to people (e.g., being warm-blooded, having an eye, being able to fly or swim, forming long-term monogamous pair bonds). A more veridical folk theory of generic biological properties should thus generate a prior that allows some probability (perhaps small) of a property occuring in two or more disjoint taxa.

To capture these patterns in how generic biological properties are distributed, our theory will assume that novel properties are generated by a mutation-like stochastic process defined over the tree structure specified by the taxonomic principle. We refer to this second assumption as the *mutation principle*, and the Bayesian model of induction that uses the resulting prior as the "evolutionary model". Intuitively, we can imagine a property that arises at the root of the tree and spreads out towards the leaves. The property starts out with some value (on or off) at the root, and at each point in the tree there is a small probability that the property will mutate, or switch its value. Whenever a branch splits, both lower branches inherit the value of the property at the point immediately above the split, and mutations thereafter occur independently along the lower branches. For example, if a property is absent at the root of the tree in Figure 1a, but switches on at the two points marked in Figure 1d, then it would apply to just horses, cows, dolphins and seals.

More formally, the mutation process is characterized by a single parameter $\lambda$, specifying the average rate at which mutations occur. Mutations are modeled as transitions in a two-state Markov

chain defined continuously over the tree, with infinitesimal matrix $[-\lambda, \lambda; \lambda, -\lambda]$. The probability that two points in the tree separated by a branch of length $t$ will have different values (present or absent) for a given property is then $\frac{1-e^{-2\lambda t}}{2}$. For simplicity we assume here that any property is equally likely to be present or absent at the tree root, and that mutations are equally likely in both directions (present $\rightarrow$ absent, absent $\rightarrow$ present), but more generally, prior knowledge about the nature or distribution of a feature could bias these probabilities. A *mutation history* for a property $Q$ is an assignment of zero or more mutations to branches of the tree, together with a specification of the state of the property (present or absent) at the root of the tree.

The Markov mutation model allows us to assign a probability to any hypothetical mutation history for a property, based only on whether the property changes state between each pair of branch points on the tree. This is almost but not exactly what we need. Bayesian property induction requires a prior probability $p(h)$ that some novel property $Q$ applies to any possible subset $h$ of species. A mutation history for property $Q$ induces a labeling of all leaf nodes of the tree – all species in the domain – according to whether $Q$ is present or absent at each node. That is, each mutation history is consistent with exactly one hypothesis for Bayesian induction, but the mapping from mutation histories to hypotheses is many-to-one: many different mutation histories, with different probabilities, will specify the same distribution of a property $Q$ over the set of species. We define the prior probability $p(h)$ that a new property $Q$ applies to some subset $h$ of species to be the sum of the probability of all mutation histories consistent with $h$. This prior (and all the resulting Bayesian computations) can be computed efficiently using belief propagation over the tree, as described more formally in Kemp et al. (2004a). For small trees, it can also be approximated by taking many samples from a simulation of the mutation process. We randomly generate a large number of hypothetical mutation histories with probability proportional to their likelihood under the Markov mutation model, by first choosing the property's state at the tree root and then following the causal direction of the mutation process down along all branches of the tree to the leaf nodes. We can estimate the prior $p(h)$ for each hypothetical labeling $h$ of the leaf nodes as the frequency with which mutation histories consistent with $h$ occur in this sample.

The prior generated by this mutation principle has several qualitative features that seem appropriate for our problem. First, unlike the strict-tree model described above, the mutation process induces a non-zero prior probability for any logically possible extension of a novel property (i.e., any of the $2^n$ hypotheses in the full hypothesis space $H$). The prior is "smooth" with respect to the tree: the closer two species lie in the tree, the more likely they are to share the same value for a novel property. Properties are more likely to switch on or off along longer branches (e.g., the mutation history in Figure 1b is more likely than in 1c). Hence $p(h)$ will be higher for properties that hold only for a highly distinctive taxonomic group of species, such as the aquatic mammals or the primates, than for properties that hold only for a less distinctive taxonomic group, such as the "farm animals" ($\{horses, cows\}$). Multiple independent occurences of the same property will be rare (e.g., the mutation history in Figure 1b is more likely than in 1d, which is more likely than in 1e). Hence the prior favors simpler hypotheses corresponding to a single taxonomic cluster, such as $\{dolphins, seals\}$, over more complex hypotheses corresponding to a union of multiple disjoint taxa, such as the set $\{dolphins, seals, horses, cows, mice\}$. The lower the mutation rate $\lambda$, the greater the preference for strictly tree-consistent hypotheses over disconnected hypotheses. Thus this model captures the basic insights of simpler heuristic approaches to taxonomic induction (Atran, 1995), but embeds them in a more powerful probabilistic model that supports fine-grained statistical inferences.

Several caveats about the evolutionary model are in order. The mutation process is just a compact mathematical means for generating a reasonable prior for biological properties. We make no

claim that people have conscious knowledge about mutations as specifically biological phenomena, any more than a computational vision model which appeals to an energy function claims that the visual system has explicit knowledge about energy. It is an open question whether the biological principles guiding our model are explicitly represented in people's minds, or only implicitly present in the inference procedures they use. We also do not claim that a mutation process is the only way to build a prior that can capture generalizations about generic biological properties. The key idea captured by the mutation process is that properties should vary randomly but smoothly over the tree, so that categories nearby in the tree are more likely to have the same value (present or absent) for a given property than categories far apart in the tree. Other stochastic processes including diffusion processes, Brownian motion, and Gaussian processes will also capture this intuition, and should predict similar patterns of generalization (Kemp & Tenenbaum, submitted). The scope of such "probabilistic taxonomic" theories is likely to extend far beyond intuitive biology: there may be many domains and contexts where the properties of categories are well described by some kind of smooth probability distribution defined over a taxonomic-tree structure, and where our evolutionary model or some close relative may thus provide a compelling account of people's inductive reasoning.

#### 4.2.2 A theory for causally transmitted properties

The evolutionary model in the previous section is appropriate for reasoning about many kinds of biological properties, and perhaps some kinds of nonbiological but still taxonomically distributed properties as well, but other classes of properties give rise to very different patterns of inductive generalization and will thus require differently structured priors. For instance, in some contexts inductive generalization will be asymmetric: the probability of generalizing a property from category A to category B will not be the same as the probability of generalizing the same property from B to A. Earlier we described one biological context where induction is frequently asymmetric: reasoning about disease properties, such as the probability that grizzly bears will have a disease given that salmon do, or vice versa. In this section we show how to capture this sort of inductive reasoning within our theory-based Bayesian framework.

The formal model we present could be appropriate for many properties whose distributions are governed by asymmetric causal relationships among categories, but for concreteness, we will assume that the domain comprises a set of species categories, and the novel property is a disease spread by predator-prey relations between species. Two abstract principles of an intuitive theory are relevant in this context; these principles are analogous to the taxonomic and mutation principles underlying the evolutionary model in the last section. First, a structured representation captures the relevant relations between entities in the domain: in this context, we posit a set of directed predator-prey relations. An example of such a food web is shown in Figure 4a. Second, a stochastic process defined over that directed-network structure generates prior probabilities for how novel properties are distributed over species: here, the process is designed to capture the noisy arrival and transmission of disease states.

As with the mutation process for generic biological properties presented above, we can describe this noisy-transmission process by explaining how to generate a single hypothetical property. If we draw a large sample of hypothetical properties by repeating this procedure many times, the prior probability for each hypothesis about how to generalize a particular novel property will be proportional to the number of times it appears in this 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

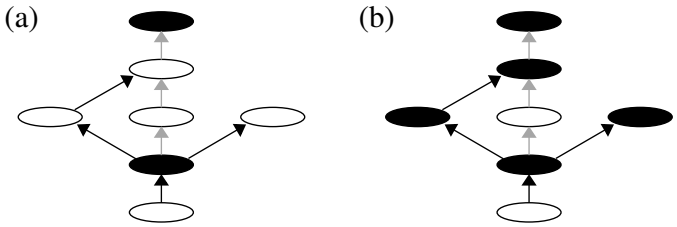


Figure 2: One simulated sample from the causal-transmission model, for the foodweb shown in Figure 4a. (a) Initial step showing species hit by the background rate (black ovals) and active routes of transmission (black arrows). (b) Total set of species with disease via background and transmission.

disease as a result of an external cause. The second parameter is used to capture the knowledge that diseases can spread from prey 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. We refer to the Bayesian model using this prior as the "causal transmission model".

Figure 2 shows one possible outcome if we sample a property from the causal transmission process. We see that two of the species develop the disease for reasons unrelated to the foodweb, and that four of the causal links are active (Figure 2a). An additional three species contract the disease by eating a disease-ridden species (Figure 2b). Reflecting on these simulations should establish that the prior captures two basic intuitions. First, species that are linked in the web by a directed path are more likely to share a novel property than species which are not directly linked. The strength of the correlation between two species' properties decreases as number of links separating them increases. Second, property overlap is asymmetric: a prey species is more likely to share a property with one of its predators than vice versa.

Although the studies we will model here consider only the case of disease transmission in a food web, many other inductive contexts fit the pattern of asymmetric causal transmission that this model is designed to capture. Within the domain of biological species and their properties, the causal model could also apply to reasoning about the transmission of toxins or nutrients. Outside of this domain, the model could be used, for example, to reason about the transmission of lice between children at a day care, the spread of secrets through a group of colleagues, or the progression of fads through a society.

#### 4.2.3 Common principles of theory-based priors

It is worth noting several deep similarities between these two theory-based Bayesian models, the evolutionary model and the causal transmission model, which point to more general aspects of our approach. In each case, the underlying intuitive theory represents knowledge on at least two levels of abstraction. The lower, more concrete level of the theory specifies a graph-structured representation of the relevant relations among species (taxonomic neighbors or predators), and a stochastic process that distributes properties over that relational structure, which is characterized by one or two numerical parameters controlling its degree of stochasticity. At a higher level, each theory specifies the *form* of the structure and stochastic process that are appropriate for reasoning about a certain domain of entities and properties: generic biological properties, such as anatomical and physiological attributes, are distributed according to a noisy mutation process operating over a taxonomic tree; diseases are distributed according to a noisy transmission process operating over

a directed food web.

For a fixed set of categories and properties, the lower level of the theory is all we need to generate a prior for inductive reasoning. But the higher level is not just a convenient abstraction for cognitive modelers to talk about – it is a critical component of human knowledge. Only these abstract principles tell us how to extend our reasoning when we learn about new categories, or a whole new system of categories in the same domain. When European explorers first arrived in Australia, they were confronted with many entirely new species of animals and plants, but they had a tremendous head start in learning about the properties of these species because they could apply the same abstract theories of taxonomic organization and disease transmission that they had acquired based on their European experience. Abstract theories appear to guide childrens' conceptual growth and exploration in much the same way. The developmental psychologists Wellman and Gelman (1992) distinguish between "framework theories" and "specific theories", two levels of knowledge that parallel the distinction we are making here. They highlight framework theories of core domains – intuitive physics, intuitive psychology, and intuitive biology – as the main objects of study in cognitive development. In related work (Tenenbaum, Griffiths and Kemp, 2006; Tenenbaum, Griffiths and Niyogi, in press), we have shown how the relations between different levels of abstraction in intuitive theories can be captured formally within a hierarchical probabilistic model. Such a framework allows us to use the same Bayesian principles to explain both how theories guide inductive generalization and how the theories themselves might be learned from experience.

Each of our theory-based models was built by thinking about how some class of properties is actually distributed in the world, with the aim of giving a rational analysis of people's inductive inferences for those properties. It is therefore not surprising that both the evolutionary model and the causal transmission model correspond roughly to models used by scientists in relevant disciplines — formalisms like the causal transmission model are used by epidemiologists, and formalisms like the evolutionary model are used in biological classification and population genetics. The correspondence is of course far from perfect, and it is clearest at the higher "framework" level of abstraction: constructs such as taxonomic trees, predator-prey networks, the mutation process or the transmission process, may in some sense be shared across intuitive theories and scientific theories, even while the specific tree or foodweb structures, or the specific mutation or transmission rate parameters, may differ in important ways.

Even though it may be imperfect and abstract, this correspondence between the world's structure and our models' representations provides an important source of constraint on our approach. If we were free to write down just any sort of probabilistic model as the source of a prior probability distribution, it would be possible to give a "rational analysis" of any coherent inductive behavior, but it is not clear how much explanatory value that exercise would have. By deriving priors from intuitive theories that at some deep level reflect the actual structure of the world, it becomes clearer why these priors should support useful generalizations in real-world tasks, and how they might themselves be acquired by a rational learner from experience in this world. Our approach can be extended to other inductive contexts, by formally specifying how the properties covered by those contexts are distributed in the world; Kemp and Tenenbaum (submitted) consider two other important contexts in addition to those discussed here.

Our primary goal here has been to characterize the knowledge that guides generalization (theory-based priors), and the input-output mapping that allows this knowledge to be converted into judgments of inductive strength (Bayesian inference). Our work is located at the most abstract of Marr's levels — the level of computational theory (Marr, 1982) — and we make no commitments about the psychological or neural processes by which people make inductive judgments. Inference in both of our models can be implemented using efficient approximate methods that have appealing correlates in the traditional toolkit of cognitive processes: for instance, belief propagation over

Bayesian networks, which can be seen as a kind of rational probabilistic version of spreading activation. We find it encouraging that efficient and psychologically plausible implementations exist for our models, but we are not committed to the claim that inference in these Bayesian networks resembles cognitive processing in any detailed way.

Finally, it is worth emphasizing that our models have not attempted to capture all or most of the content and structure of people's intuitive theories. We are modeling just those aspects of theories that appear necessary to support inductive reasoning about properties in fairly specific contexts. We are agnostic about whether people's intuitive theories contain much richer causal structures than those we attempt to model here (Carey, 1985), or whether they are closer to light or skeletal frameworks with just a few basic principles (Wellman and Gelman, 1992).

# 5 Testing the models of property induction

We now describe two series of experimental tests for the theory-based Bayesian models introduced in the previous section. In each case, we consider multiple sets of inductive arguments whose strengths have been rated or ranked by human judges, and we compare these subjective argument strengths with the theoretical predictions of the models. We will also compare with several alternative models, including the classic similarity-coverage model and multiple variants within our Bayesian framework. The latter comparisons allow us to illustrate the distinctive importance of both ingredients in our approach: an appropriate representation of the relevant aspects of a domain's structure, and an appropriate probabilistic model for how properties of the relevant type are distributed over that structure.

## 5.1 Reasoning about generic biological properties

We begin by looking at the classic datasets of Osherson, Smith and their colleagues, on inductive reasoning with generic biological properties. Five datasets will be considered. The two "Osherson" datasets are taken from Osherson et al. (1990). The "Osherson specific" set contains 36 two-premise arguments, of the form *gorillas*, *squirrels* $\longrightarrow$ *horses*. The conclusion category of each of these arguments is the same: *horses*. The two premise categories vary across arguments, but are always drawn from the set of ten mammal species shown in Figure 1. Various generic biological predicates are used across different arguments; in this section we drop references to the specific property assuming it is one of these generics. The "Osherson general" set contains 45 three-premise general arguments, of the form *gorillas*, *squirrels*, *dolphins* $\longrightarrow$ *mammals*. The conclusion category of each of these arguments is always *mammals*, while the three premise categories again vary across arguments and are drawn from the set of ten mammal species shown in Figure 1. The three "Smith" data sets are similar, but they draw on different and fewer mammal categories (Smith et al., 1993).

We compare people's judgments of argument strength in these five datasets with the predictions of five computational models for induction with generic biological properties. Two are theory-based Bayesian models: the evolutionary model, in which the prior is generated by a mutation process defined on a taxonomic tree of species categories, and the strict-tree model, in which the prior is simply a uniform distribution over taxonomic clusters of species (without the possibility of a property arising in multiple disconnected branches of the taxonomy). Another model is inspired by Heit's proposal for a Bayesian analysis (see Section 3.3), in which the prior probability of a hypothesis is based on the number of familiar properties that can be retrieved from memory and that have the same extension as that hypothesis. We call this the "raw-feature model", because it embodies a prior that is based directly on raw experience, without the benefit of a structured domain theory that might help people to reason sensibly in cases that go substantially beyond

their experience. The details of the raw-feature model are explained below. Finally, we consider two versions of the similarity-coverage model, MaxSim and SumSim, which respectively compute similarity to the set of premise categories in terms of the maximum or summed similarity to those categories (see Section 3.1).

In order to predict the strengths of arguments about a particular set of species, each of these models requires some way to represent people's prior knowledge about those species. We can compare the models on equal footing by grounding the knowledge representations for each model in a matrix of judged species-feature associations collected by Osherson and colleagues. Participants were given 48 familiar species and 85 familiar features (mostly anatomical or ecological properties, such as "has a tail" or "lives in water") and asked to rate the relative "strength of association" between each species and each feature. Participants gave ratings on a scale that started at zero and had no upper bound. In order to model the behavioral judgments described below, we supplemented these data with feature ratings for two additional species, *cows* and *dolphins*, to give a total of 50 species. We also substituted the judged features of *collies* for *dogs* (because *dogs* appeared as a premise category in some of the argument judgment tasks, but not in the feature rating task). Ratings were linearly transformed to values between 0 and 100, then averaged. Let $F$ be the resulting $50 \times 85$ matrix of average species-feature associations. We also consider analogous binary matrices $F_\theta$ obtained by thresholding the average ratings at some level $\theta$. Let $S(F)$ be a $50 \times 50$ matrix of species-species similarities, computed based on Euclidean distances between the two feature vectors in $F$ representing each pair of species.

For evaluating the predictions of the SCM models, we used the entries of $S(F)$ to determine the necessary similarities between species. The taxonomic tree for the evolutionary and strict-tree models was constructed by running hierarchical agglomerative ("average linkage") clustering on $S(F)$, restricted to just the species categories involved in each experiment. The prior for the raw-feature model was defined from the thresholded species-feature matrices $F_\theta$, inspired by the proposal of Heit (1998). (We treat the threshold $\theta$ as a free parameter of the model, to be optimized when fitting judgments of argument strength.) We assume that the features participants retrieve from memory correspond to the columns of $F_\theta$, plus an additional feature corresponding to all the animals. The prior $p(h)$ assigned to any hypothesis $h$ for the extension of a novel property is proportional to the number of columns of $F_\theta$ (i.e., the number of familiar features) that are distributed as $h$ specifies – that apply to just those categories that $h$ posits. Hypotheses that do not correspond to any features in memory (any column of $F_\theta$) receive a prior probability of zero.

All of these models (except the strict-tree model) include a single free parameter: the mutation rate in the evolutionary model, the balance between similary and coverage terms in MaxSim and SumSim, or the feature threshold $\theta$ in the raw-feature model. Each model's free parameter was set to the value that maximized the average correlation with human judgments over all five datasets.

Figure 3 compares the predictions for all five of these models on all five datasets of argument strength judgments. Across these datasets, the predictions of the evolutionary model are better than or comparable to the best of the other models. This success provides at least some evidence that the model's core assumptions – a taxonomic-tree structure over species and a mutation-like distribution of properties over that tree – do in fact characterize the way people think and reason about generic biological properties. More revealing insights come from comparing the performance of the evolutionary model with that of the other models.

The strict-tree model captures the general trends in the data, but does not predict people's judgments nearly as accurately as the evolutionary model. This is because, without the mutation principle, the strictly taxonomic hypothesis space is too rigid to capture the graded degrees of support that more or less diverse premise sets provide for inductive generalization. For example, in the Osherson-general experiment, the strict-tree model assigns the same probability (100%) to

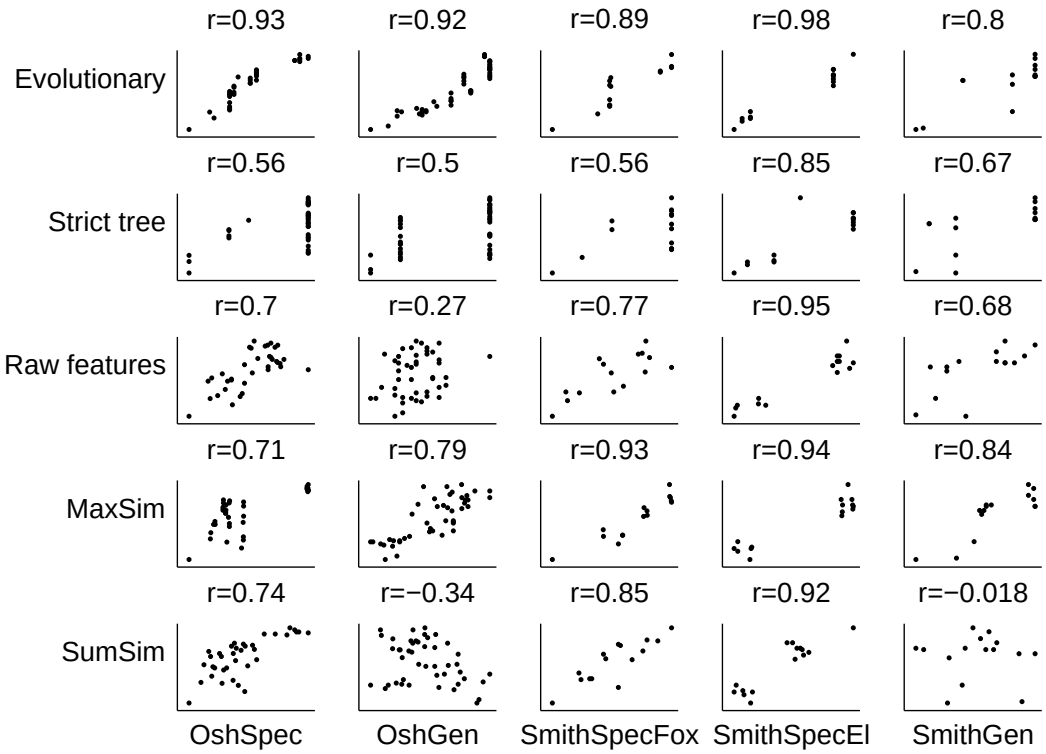


Figure 3: Comparing models of property induction with human judgments, for reasoning in a default biological context with generic anatomical or physiological properties. Each row of plots shows the performance of a single model over all data sets; each column shows the performance of all models over a single data set. In each plot, individual data points represent the strengths of individual inductive arguments. The x-value of each point represents the argument's predicted strength according to a given model, while the y-value represents the argument's subjective strength according to the mean judgments of human experimental participants.

$cows, dolphins, squirrels \longrightarrow mammals$ and $seals, dolphins, squirrels \longrightarrow mammals$, because in both cases the set of all mammals is the only hypothesis consistent with the examples. The evolutionary model correctly 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. An intuitive explanation is that it feels very difficult to imagine a property that would be true of cows, dolphins, and squirrels but not all mammals, while it seems more plausible (if unlikely) that there could be some characteristic property of aquatic mammals (seals and dolphins) that might, for some unknown reason, also be true of squirrels or rodents, but no other animals. The mutation principle matches this intuition: the highly specific hypothesis $\{seal, dolphin, squirrel\}$ can be generated by only two mutations, one on a very long branch (and thus relatively likely), while the hypothesis $\{cows, dolphins, squirrels\}$ could only arise from three mutations all on relatively short branches.

This problem with the strict-tree model is hardly restriced to the specific pair of arguments cited above, as can be seen clearly by the dense vertical groupings of datapoints on the right-hand side of the plots in Figure 3. Each of those vertical groupings correspnds to a set of arguments that are judged to have very different strengths by people, but that all receive maximum probability under the strict-tree prior, because they are consistent with just the single hypothesis of generalizing to

all mammals.

The raw-feature model is more flexible than the strict-tree model, but is still not sufficient to capture the diversity effect, as can be seen by its dramatically worse performance on the datasets with general arguments. Consider the premise sets *dolphins*, *chimps*, *squirrels* and *dolphins*, *seals*, *horses*. It is difficult to think of anatomical or physiological properties that apply to all of the animals in the first set, but only some of the animals in the second set. None of the features in our dataset is strongly associated with dolphins, chimps and squirrels, but not also seals and horses. The raw-feature model therefore finds it hard to discriminate between these two sets of premises, even though it seems intuitively that the first set provides better evidence that all mammals have the novel property.

More generally, the suboptimal performance of the raw-feature model suggests that people's hypotheses for induction are probably not based strictly on the specific features that can be retrieved from memory. People's knowledge of specific features of specific animals is too sparse and noisy to be the direct substrate of inductive generalizations about novel properties. In contrast, a principal function of intuitive domain theories is to generalize beyond people's limited specific experiences, constraining the kinds of possible situations that would be expected to occur in the world regardless of whether they have been previously experienced (McCloskey et al., 1980; Murphy and Medin, 1985; Carey, 1985). Our framework captures this crucial function of intuitive theories by formalizing the theory's core principles in a generative model for Bayesian priors.

Taken together, the performance of these three Bayesian variants shows the importance of both aspects of our theory-based priors: a structured representation of how categories in a domain are related and a probabilistic model describing how properties are distributed over that relational structure. The strict-tree model incorporates an appropriate taxonomic structure over categories but lacks a sufficiently flexible model for how properties are distributed. The raw-feature model allows a more flexible prior distribution for properties, but lacking a structured model of how categories are related, it is limited to generalizing new, partially observed properties strictly based on the examples of familiar, fully observed properties. Only the prior in the evolutionary model embodies both of these aspects in ways that faithfully reflect real-world biology – a taxonomic structure over species categories and a mutation process generating the distribution of properties over that tree – and only the evolutionary model provides consistently strong quantitative fits to people's inductive judgments.

Turning now to the similarity-based models, Figure 3 shows that their performance varies dramatically depending on how we define the measure of similarity to the set of premise categories. MaxSim fits reasonably well, somewhat worse than the evolutionary model on the two Osherson datasets but comparably to the evolutionary model on the three Smith datasets. The fits on the Osherson datasets are worse than those reported by Osherson et al. (1990), who used direct human judgments of similarity as the basis for the model rather than the similarity matrix $S(F)$ computed from people's feature ratings. We used the species-feature associations here in order to compare all models (including the raw-feature model) on equal terms, but we have also compared versions of the evolutionary model and the similarity-coverage models using the similarity judgments of Osherson et al. (1990) to build the taxonomic tree or compute MaxSim or SumSim scores (Kemp and Tenenbaum, 2003). In that setting, too, the evolutionary model consistently performs at least as well as MaxSim.

SumSim is arguably the least successful model we tested. Its predictions for the strengths of general arguments are either uncorrelated with or negatively correlated with people's judgments. Although the good performance of MaxSim shows that a similarity-based model can describe people's patterns of inductive reasoning, the poor performance of SumSim calls into question the explanatory value of similarity-based approaches. As mathematical expressions, the SumSim and MaxSim

measures do not appear very different, beyond the presence of a nonlinearity in the latter case. This nonlinearity turns out to make all the difference; it is necessary for similarity-based models to predict diversity-based inductive reasoning. Because the total inductive strength of an argument under SumSim is a linear function of the inductive strength associated with each premise category, the model assigns highest strength to arguments in which each of the premise categories would individually yield the strongest one-premise argument. This preference goes against diversity, because the categories that make for the best single-premise arguments tend to be quite similar to each other. For instance, unlike people, SumSim assigns a higher strength to *horses, cows, rhinos* $\longrightarrow$ *mammals* than to *horses, seals, squirrels* $\longrightarrow$ *mammals*, because the strength of a generalization from the individual premise categories *horses*, *cows* or *rhinos* to *mammals* are some of the strongest in the domain – significantly higher than from less typical mammals such as *seals* or *squirrels*.

We can see why SumSim fails, but there is no principled a priori justification within the similarity-based paradigm for adopting the nonlinear MaxSim rather than the linear SumSim as a model of inductive generalization. The SumSim measure has if anything greater precedent in other similarity-based models of learning, categorization, and memory (Nosofsky, 1986; Kruschke, 1992; Hintzman et al., 1978). Osherson et al. (1990) do not attempt to justify the preference for MaxSim – it just seems to fit people's intuitions better in the particular task of property induction. That is fine as far as a descriptive model goes, but not very satisfying if one of our goals is to explain why people's intuitions work the way that they do.

In contrast, the theory-based Bayesian approach we have presented offers a principled and rational explanation for why people make the judgments that they do on these inductive reasoning tasks. People's judgments are approximately optimal, with respect to our evolutionary model that combines the inferential optimality of Bayesian principles with a prior based on how the properties of natural species are distributed in the real world. This rational approach can also explain why some Bayesian model variants fare better than others. The raw-feature and strict-tree models yield substantially poorer descriptive fits to human judgments, and each is based on a prior that neglects a key principle of how natural categories and properties are structured which the evolutionary model properly incorporates. Finally, our approach could explain why the most successful similarity-based models of induction work the way they do, and in particular, why they are based on the maxsim mechanism rather than the more standard sumsim operation. In Kemp and Tenenbaum (2003) and Kemp et al. (2004a) we show that under certain conditions, MaxSim (but not SumSim) provides an efficient heuristic approximation to the ideal computations of the evolutionary Bayesian model. Hence at some level of analysis, and under certain circumstances, human inductive reasoning could be well described as a similarity-based computation – but which similarity-based computation that is, and why it works the way that it does, would still be best explained by an analysis in our theory-based Bayesian framework.

### 5.2 Reasoning about causally transmitted properties

A second set of studies was intended to evaluate the descriptive power of the Bayesian causal-transmission model, our model for inductive reasoning about diseases and other causally transmitted properties, and also to show how different theory-based Bayesian models can be used to account for different patterns of inductive reasoning that arise with different kinds of properties.

We work with data from experiments by Shafto, Kemp, Bonawitz, Coley and Tenenbaum (submitted); see also (Shafto et al., 2005). Participants were first trained to memorize the structure of the food webs shown in Figure 4. They were also familiarized with the correct taxonomic groupings for these species categories. After this initial training, participants were asked to judge the strength of inductive arguments about one of two kinds of properties: a disease property ("has

disease D") or a genetic property ("has gene XR-23"). All arguments had a single premise and a specific conclusion, and the stimuli exhaustively explored all arguments of this form. That is, every pair of categories appeared as a premise-conclusion pair in some argument. Participants also provided pairwise similarity ratings between the species in each food web, and we again used hierarchical clustering to recover representations of people's taxonomic trees for these categories. The recovered taxonomies are shown in Figure 4. Free parameters for all models (the mutation rate in the evolutionary model, the background and transmission rates for the causal transmission model) were set to values that maximized the models' correlations with human judgments.

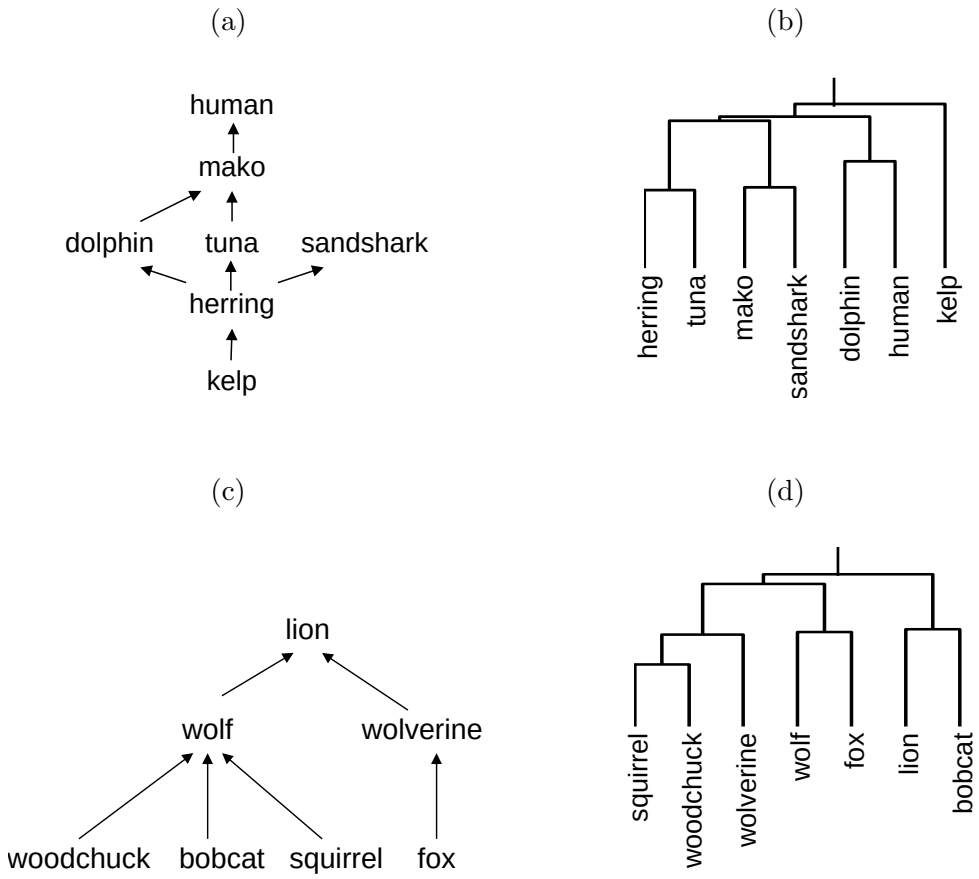


Figure 4: Multiple relational structures over the same domains of species. (a) A directed network structure capturing food web relations for an "island" ecosystem. (b) A rooted ultrametric tree capturing taxonomic relations among the same species. (c) A directed network structure capturing food web relations for a "mammals" ecosystem. (d) A rooted ultrametric tree capturing taxonomic relations among the same species.

We hypothesized that participants would reason very differently about disease properties and genetic properties, and that these different patterns of reasoning could be explained by theory-based Bayesian models using appropriately different theories to generate their priors. Specifically, we expected that inductive inferences about disease properties could be well approximated by the causal-transmission model, assuming that the network for causal transmission corresponded

to the food web learned by participants. We expected that inferences about genetic properties could be modeled by the evolutionary model, assuming that the taxonomic structure over species corresponded to the tree we recovered from participants' judgments. We also tested MaxSim and expected that it would perform similarly to the evolutionary model, as we found for the previous set of studies.

Figure 5 shows that these predictions were confirmed. High correlations were found between the causal-transmission model and judgments about disease properties, and between the evolutionary model and judgments about genetic properties. Morever, we observed a double dissociation between property types and model types. The causal-transmission model correlates weakly or not at all with judgments about genetic properties, while there is no significant correlation between the evolutionary model and judgments about disease properties. This double dissociation is the clearest sign yet that when our Bayesian models fit well, it is not simply because they are using sophisticated general-purpose inference principles; their success depends crucially upon using a theory-generated prior that is appropriately matched to the domain structure and inductive context.

The performance of MaxSim was also as hypothesized: highly correlated with inductive judgments about genetic properties (like traditional "blank" properties), but poorly correlated with judgments about disease properties when participants were familiar with relevant predator-prey relations. This result should not be surprising. It is well-known that similarity-based approaches have difficulty accounting for inductive reasoning beyond the context of generic biological properties, when some other relevant knowledge is available to people (Smith et al., 1993; Medin et al., 2003; Shafto and Coley, 2003). The usual interpretation of similarity's shortcomings expresses a general pessimism about computational models of common-sense reasoning: different kinds of properties or inductive contexts just call for fundamentally different approaches to reasoning, and cognitive scientists should not hope to be able to give a principled general-purpose account for all – or even a large class – of everyday inductive inferences.

Our theory-based Bayesian framework offers a more optimistic view. We have separated out the general-purpose inferential mechanisms from the context-specific knowledge that guides those mechanisms, and we have introduced a potentially quite general way of modeling relevant aspects of contextual knowledge, in terms of relational structures over categories and stochastic processes for generating distributions of properties over those structures. It is at least a start towards explaining more precisely how induction in different contexts operates, and how apparently quite different patterns of reasoning in different contexts or domains could in fact could be given a unifying and rigorous explanatory account.

## 6 Conclusions and open questions

Conventional models of induction focus on the processes of inference rather than the knowledge that supports those inferences. Inference mechanisms are typically given fully explicit and mathematical treatments, while knowledge representations are either ignored or boiled down to place-holders like a similarity metric or a set of features that at best only scratch the surface of people's intuitive theories. This focus probably stems from the natural modeler's drive towards elegance, generality, and tractability: the knowledge that supports induction is likely to be messy, complex, and specific to particular domains and contexts, while there is some hope that one or a few simple inference principles might yield insight across many domains and contexts. Yet this approach limits the descriptive power and explanatory depth of the models we can build, and it assigns a second-class status to crucial questions about the form and structure of knowledge. Our goal here is to return questions of knowledge to their appropriate central place in the study of inductive inference, by

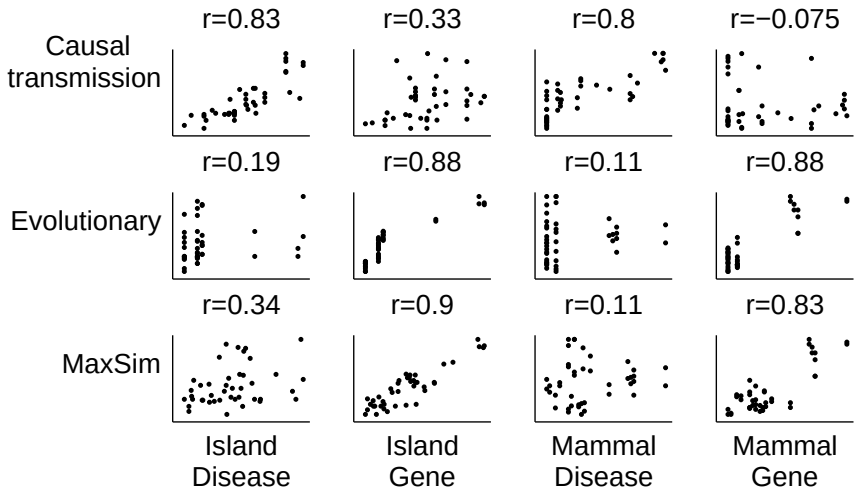


Figure 5: Comparing models of induction with human judgments, for two kinds of properties: disease properties and genetic properties. Both kinds of property-induction tasks were studied for two different systems of species categories, an "Island" ecosystem and a "Mammals" ecosystem, as shown in Figure 4. Plotting conventions here are the same as in Figure 3.

showing how they can be addressed rigorously and profitably within a formal Bayesian framework.

Every real-world inference is embedded in some context, and understanding how these different contexts work is critical to understanding real-world induction. We have argued that different contexts trigger different intuitive theories, that relevant aspects of these theories can be modeled as generative models for prior probability distributions in a Bayesian reasoning framework, and that the resulting theory-based Bayesian models can explain patterns of induction across different contexts. We described simple theories for generating priors in two different inductive contexts, a theory for generic biological properties such as genetic, anatomical or physiological features of species, and a theory for causally transmitted properties such as diseases, nutrients, or knowledge states. We showed that Bayesian models based on a prior generated by each theory could predict people's inductive judgments for properties in the appropriate context, but not inappropriate contexts.

Intriguingly, both of these generative theories were based on principles analogous to those used by scientists to model analogous phenomena in real biological settings. In showing their success as descriptive models of people's intuitive judgments, we begin to provide an explanation for people's remarkable ability to make successful inductive leaps in the real world, as the product of rational inference mechanisms operating under the guidance of a domain theory that accurately reflects the true underlying structure of the environment.

Of course, just specifying these two theory-based models is far from giving a complete account of human inductive reasoning. Characterizing the space of theories that people can draw upon, and the processes by which they are acquired and selected for use in particular contexts, is a challenging long-term project. Perhaps the most immediate gap in our model is that we have not specified how to decide which theory is appropriate for a given argument. Making this decision automatically will require a semantic module that knows, for example, that words like "hormone" and "gene" are related to the generic biological theory, and words like "disease" and "toxin" are related to the theory of causal transmission. How best to integrate some form of this semantic knowledge with our existing models of inductive reasoning is an open question.

We have discussed theories that account for inductive reasoning in two contexts, but it is natural and necessary to add more. Some of our ongoing work is directed toward developing and testing models beyond the tree- and network-based models described here. For example, in Kemp and Tenenbaum (submitted), we show how reasoning about properties such as "can bite through wire" can be modeled as Bayesian inference over a linear structure representing a dimension of strength, rather than a tree or a network representing taxonomic or foodweb relations (Smith et al., 1993). In other work, we are extending our approach to account for how multiple knowledge structures can interact to guide property induction; in particular, we are looking at interactions between networks of causal relations among properties and tree structures of taxonomic relations among species. Finally, in addition to broadening the scope of theory-based Bayesian models, we have begun more fundamental investigations into how these probabilistic theories can themselves be learned from experience in the world (Kemp et al., 2004b), and how learners can infer the appropriate number and complexity of knowledge structures that best characterize a domain of categories and properties (Shafto et al., 2006). These problems are as hard to solve as they are important. Yet we believe we are in a position to make real progress on them – to develop genuine insights into the origins and nature of common sense – using models that combine the principles of Bayesian inference with structured knowledge representations of increasing richness and sophistication.

**Acknowledgments.** This work has benefited from the insights and contributions of many colleagues, in particular Tom Griffiths, Neville Sanjana, and Sean Stromsten. Liz Bonawitz and John Coley were instrumental collaborators on the food web studies. JBT was supported by the Paul E. Newton Career Development Chair. CK was supported by the Albert Memorial Fellowship. We thank Bob Rehder for helpful comments on an earlier draft of the chapter.

## References

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

Ashby, F. G. and Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of mathematical psychology*, 39:216–233.

Atran, S. (1995). Classifying nature across cultures. In Smith, E. E. and Osherson, D. N., editors, *An Invitation to Cognitive Science*, volume 3. MIT Press.

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

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

Gelman, S. and Markman, E. (1986). Categories and induction in young children. *Cognition*, 23(3):183–209.

Goodman, N. (1972). Seven strictures on similarity. In *Problems and Projects*. Bobbs-Merril Co., Indiana.

Gopnik, A. and Meltzoff, A. (1997). *Words, Thoughts, and Theories*. MIT Press, Cambridge, MA.

Gopnik, A. and Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences*, 8:371–377.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In Oaksford, M. and Chater, N., editors, *Rational Models of Cognition*, pages 248–274. Oxford University Press.

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

Hintzman, D. L., Asher, S. J., and Stern, L. D. (1978). Incidental retrieval and memory for coincidences. In Gruneberg, M. M., Morris, P. E., and Sykes, R. N., editors, *Practical aspects of memory*, pages 61–68. Academic Press, New York.

Kemp, C., Griffiths, T. L., Stromsten, S., and Tenenbaum, J. B. (2004a). Semi-supervised learning with trees. In *Advances in Neural Processing Systems 16*. MIT Press.

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2004b). Learning domain structures. In *Proceedings of the 26th annual conference of the Cognitive Science Society*.

Kemp, C. and Tenenbaum, J. B. (2003). Theory-based induction. In *Proceedings of the 25th annual conference of the Cognitive Science Society*.

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

Markman, E. (1989). *Naming and Categorization in Children*. MIT Press, Cambridge, MA.

Marr, D. (1982). *Vision*. W. H. Freeman.

McCloskey, M., Caramazza, A., and Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, 210(4474):1139–1141.

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

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Murphy, G. L. (1993). Theories and concept formation. In Mechelen, I. V., Hampton, J., Michalski, R., and Theuns, P., editors, *Categories and concepts: Theoretical views and inductive data analysis*. Academic Press.

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

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

Oaksford, M. and Chater, N., editors (1998). *Rational models of cognition*. Oxford University Press.

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

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

Rogers, T. T. and McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT Press, Cambridge, MA.

Rosch, E. (1978). Principles of categorization. In Rosch, E. and Lloyd, B. B., editors, *Cognition and categorization*, pages 27–48.

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

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

Shafto, P., Kemp, C., Mansinghka, V., Gordon, M., and Tenenbaum, J. B. (2006). Learning cross-cutting systems of categories. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*. Erlbaum, Hillsdale, NJ.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, England.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, 25:213–280.

Smith, E. E., Safir, E., and Osherson, D. (1993). Similarity, plausibility, and judgements of probability. *Cognition*, 49:67–96.

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

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318.

Tenenbaum, J. B., Griffiths, T. L., and Niyogi, S. (in press). Intuitive theories as grammars for causal inference. In Gopnik, A. and Schulz, L., editors, *Causal learning: Psychology, philosophy, and computation*. Oxford University Press, Oxford.

Tenenbaum, J. B. and Xu, F. (2000). Word learning as Bayesian inference. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 517–522. Erlbaum, Hillsdale, NJ.

Wellman, H. M. and Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, 43:337–375.

Xu, F. and Tenenbaum, J. B. (in press). Sensitivity to sampling in bayesian word learning. *Developmental Science*.