The acquisition of inductive constraints

by

Charles Kemp

B.A., University of Melbourne, 2000
B.Sc. (Hons.), University of Melbourne, 2001

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

February 2008

Author: _____

Department of Brain and Cognitive Sciences
September 27, 2007

Certified by: _____

Joshua Tenenbaum
Associate Professor of Cognitive Science
Thesis Supervisor

Accepted by: _____

Matthew Wilson
Professor of Neurobiology
Chairman, Committee for Graduate Students

The acquisition of inductive constraints

by

Charles Kemp

Submitted to the Department of Brain and Cognitive Sciences
on September 27, 2007, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Human learners routinely make inductive inferences, or inferences that go beyond the data they have observed. Inferences like these must be supported by constraints, some of which are innate, although others are almost certainly learned. This thesis presents a hierarchical Bayesian framework that helps to explain the nature, use and acquisition of inductive constraints. Hierarchical Bayesian models include multiple levels of abstraction, and the representations at the upper levels place constraints on the representations at the lower levels. The probabilistic nature of these models allows them to make statistical inferences at multiple levels of abstraction. In particular, they show how knowledge can be acquired at levels quite remote from the data of experience—levels where the representations learned are naturally described as inductive constraints.

Hierarchical Bayesian models can address inductive problems from many domains but this thesis focuses on models that address three aspects of high-level cognition. The first model is sensitive to patterns of feature variability, and acquires constraints similar to the shape bias in word learning. The second model acquires causal schemata—systems of abstract causal knowledge that allow learners to discover causal relationships given very sparse data. The final model discovers the structural form of a domain—for instance, it discovers whether the relationships between a set of entities are best described by a tree, a chain, a ring, or some other kind of representation.

The hierarchical Bayesian approach captures several principles that go beyond traditional formulations of learning theory. It supports learning at multiple levels of abstraction, it handles structured representations, and it helps to explain how learning can succeed given sparse and noisy data. Principles like these are needed to explain how humans acquire rich systems of knowledge, and hierarchical Bayesian models point the way towards a modern learning theory that is better able to capture the sophistication of human learning.

Thesis Supervisor: Joshua Tenenbaum
Title: Associate Professor of Cognitive Science

Acknowledgements

Josh Tenenbaum has been my advisor since I first arrived at MIT, and most of the ideas in this thesis were developed in collaboration with him. Josh has the rare ability to draw on proposals from cognitive psychology, philosophy, and machine learning, and to put them together to address the most fundamental questions in cognitive science. Working with him has brought me a great deal of intellectual and personal satisfaction, and anyone who walks into his office will begin to understand why. The walls and sometimes the windows are covered with writing, and it is clear that Josh is overflowing with ideas and generous about sharing them with others. His research group inherits both qualities, and I feel very fortunate to have been part of this community for five years.

Whitman Richards has given me valuable advice about most of the projects I have worked on, including several which do not appear in this thesis. Whitman encouraged my interests in graph structures and in social systems, and the experiments in Chapter 6 were written up initially as a report for his class. Whitman's teaching and writing draw on an incredibly broad range of sources, and he has introduced me to many provocative ideas that are well outside the mainstream psychological literature.

The remaining members of my thesis and advisory committees have encouraged me and pushed me to improve my work. Laura Schulz helped me improve the models and experiments in Chapter 5, and gave me valuable advice about academic talks. Frank Keil helped me to see how this thesis fits in to the broader literature on inductive constraints, and identified many weak spots in an earlier draft of this thesis. Molly Potter encouraged me to move beyond simple models and develop formal approaches that begin to capture the complexity of real-world knowledge. I haven't yet managed to satisfy this request, but it has guided my thinking for several years.

Tom Griffiths and Tania Lombrozo helped me feel at home when I first arrived in Boston, and their playful approach to life and work sets a standard I might reach one day. Tom's intellectual style has shaped the way I think and write about cognitive science, and his advice about academic jobs and many other topics has been

invaluable.

Three other long-term members of the computational cognitive science group have provided steady support both inside and outside of the lab. Amy Perfors is a regular debating partner and a collaborator on the work described in Chapter 4. Amy introduced me to Taco Bell, rugby songs and the state of Vermont, and made fun of my accent only sometimes. Lauren Schmidt is a collaborator who helped me appreciate the art of experiment design, and who is the main reason I have come to value the Thanksgiving holiday. Liz Bonawitz is a collaborator who helped me to run my first experiment, and I'm grateful to her for arranging apple-picking outings and for bringing song and dance into the lab.

Two postdoctoral associates have been valued collaborators over several years. Patrick Shafto willingly shared his knowledge of cognitive psychology, and his presence brightened up the morning shift in the lab. Noah Goodman has sharpened my thinking about many philosophical aspects of cognitive science, and is a collaborator on the work described in Chapter 5.

I'm grateful to several other friends and colleagues at MIT and beyond, including Chris Baker, Ben Balas, Sreekar Bhaviripudi, Keith Bonawitz, Mike Frank, Konrad Körding, Tevye Krynski, Vikash Mansinghka, Tim O'Donnell, Dan Roy, Virginia Savova, Sean Stromsten, Yves Victor, and Ed Vul. Some of them have been roommates, others have been collaborators, and all of them have made MIT a fun place to study.

I have been fortunate to work with four outstanding undergraduates and high-school students: Allison Berke, Aaron Bernstein, Genya Frenkel, and Pooja Jotwani. Each one spent a summer in the lab and helped to transform abstract ideas into models and experiments. I'm grateful to all of the participants in our experiments, especially those who kept trying when faced with difficult tasks.

My research has been supported in part by a Praecis Presidential Fellowship and by the William Asbjornsen Albert memorial fellowship. I'm grateful to Denise Heintze, who immediately knew the right response to every administrative issue I faced, and to John Bearley, who makes the Teuber library a pleasure to visit.

I am especially grateful to my parents, Andrew and Sally; my sisters, Celia and Alice; my grandparents, Heather, Winston, Mary and Len; and to many other members of my extended family. More than anyone else, they have taught me the value of ideas and encouraged me at every stage of my education. Their phone calls, letters, postcards, and packages have kept me grounded, and my mental representation of the past few years is organized around the weeks I have managed to spend with them. I owe more than I can express to all of them, but this thesis is dedicated to two grandparents who passed away during my time at MIT: Winston Kent, who showed me the importance of thinking for myself, and Mary Kemp, who would have read every word in this document.

Contents

Chapter 1

Inductive inference

One of the most striking human achievements is routinely observed in homes around the world. Take a young child and expose her to light waves, sound waves, and other patterns of sensory stimulation. Somehow she will learn words, causal relationships and grammatical rules, and will develop abstract knowledge about numbers, objects, space, time, and the beliefs and desires of others. The acquisition of human knowledge raises many challenging questions, but many are elaborations of a single fundamental question: how do learners make inferences that go beyond the data they have observed? Psychologists and philosophers alike have struggled to understand the relationship between the "meager input" and the "torrential output" (Quine, 1969).

Inferences that go beyond the available data are sometimes called ampliative or non-deductive inferences, but I will refer to them as inductive inferences. Some of the earliest inductive inferences may be inferences about visual scenes. At 4 months of age, for instance, infants make inductive predictions about the shapes of objects that are partially concealed by an occluder, and about the trajectories of moving objects that pass behind an occluder (E. S. Spelke, 1990). Inductive inferences, however, can be found in almost every domain of cognition. Consider a child who observes her mother point at a bird and utter the word "swan." This observation is consistent with many hypotheses about the meaning of the word: perhaps it refers to the beak of the bird, or to any object that is white, or to any creature with a long neck. A single labeled example, however, is often enough for children to grasp the meaning of a

novel word. Other aspects of linguistic knowledge are also acquired given very sparse data. To mention one familiar example, children acquire grammatical constructions that are rarely found in the sentences that they hear (Chomsky, 1980).

A partial explanation of human inductive abilities has been available for many centuries. Since inductive inferences arrive at conclusions that go beyond the available data, additional elements are needed to bridge the gap between data and conclusions. These additional elements might be given different names, but I will refer to them as inductive constraints. There is room for debate about the nature of these constraints, but the need for constraints of some sort has been widely recognized by philosophers, psychologists, statisticians, and machine learning researchers (Keil, 1981; Chomsky, 1986; Holland, Holyoak, Nisbett, & Thagard, 1986).

Many ideas about inductive constraints can be traced back to the philosophical literature. Peirce points out that any set of observations can potentially be explained by a vast number of hypotheses, and asks how a learner might identify the hypotheses that turn out to be productive. His answer is that the mind has innate tendencies which lead it towards appropriate hypotheses: "if men had not come to [Nature] with special aptitudes for guessing right, it may well be doubted whether in the ten or twenty thousand years that they may have existed their greatest mind would have attained the amount of knowledge which is actually possessed by the lowest idiot" (Peirce, 1931–1935). Other philosophers have demonstrated the need for inductive constraints, and two of these demonstrations are particularly memorable. Goodman (1955) discusses constraints which help a learner identify *lawlike* hypotheses, or hypotheses that are supported by their positive instances. For instance, observing a green emerald supports the hypothesis that "all emeralds are green," but does not support the hypothesis that "all emeralds are green if examined before 2050, or blue if not so examined." Quine (1960) focuses on the problem of language acquisition, and suggests that the evidence available to learners is insufficient to establish the meanings of the words in their native language. A common conclusion is that language learners must rely on constraints which limit the word meanings that they entertain (Markman, 1989).

Inspired in part by the philosophical literature, psychologists have argued that learning depends critically on inductive constraints and have proposed specific constraints that may play a role in human learning (Table 1.1). Researchers including Markman (1989) have explored how children learn novel words, and have identified several constraints that can help children overcome the challenges identified by Quine. One of these constraints is the whole object bias: the expectation that novel labels tend to refer to entire objects (such as swans) instead of object parts (such as beaks), or object attributes such as size, color, or texture. Chomsky (1986) has suggested that children are exposed to linguistic data that are relatively sparse, and are able to learn the grammar of their native language only because they begin with constraints that limit the class of possible grammars. E. S. Spelke (1990) has studied inductive inferences about the shape and motion of physical objects, and has suggested that these inferences are constrained by abstract knowledge, including the knowledge that objects tend to follow smooth trajectories through space and time. Other psychologists have identified constraints that appear to support inferences about space (Landau, Gleitman, & Spelke, 1981), number (R. Gelman & Gallistel, 1978), living kinds (Atran, 1998), and the goals and beliefs of others (Wellman, 1990).

Computer scientists and mathematicians have supported these philosophical and psychological arguments by providing formal demonstrations of the importance of inductive constraints (Watanabe, 1969; Schaffer, 1994). One well-known result is the *No Free Lunch Theorem* which states that there is no learning algorithm that can succeed in all possible contexts—averaged across all conceivable contexts, no algorithm can perform better than random guessing (Wolpert, 1995). It follows that even the most powerful learning algorithm cannot avoid the need for inductive constraints, and will succeed only if the constraints it incorporates are well-matched to the problem at hand. The assumptions made by a learning algorithm are often referred to collectively as its *inductive bias* (Geman, Bienenstock, & Doursat, 1992; Mitchell, 1997), but these assumptions might equally well be described as inductive constraints.

Although few researchers deny the importance of inductive constraints, there are fierce debates about the nature and origin of these constraints. A strong empiricist

Domain	Constraints	Reference
Word learning	Shape bias	Landau, Smith, and Jones (1988)
	Whole object bias	Markman (1989)
	Taxonomic bias	Markman (1989)
	Principle of contrast	Clark (1987)
Predicability	M-constraint	Keil (1979)
Causal learning	Causal schemata	Kelley (1972)
Kinship	Social schemata	D. Jones (2003)
Folk biology	Taxonomic principle	Atran (1998)
Folk physics	Spelke principles	E. S. Spelke (1990)
Folk psychology	Theory of Mind	Wellman (1990)
Spatial reasoning	Geometric principles	Landau et al. (1981)
Number	Counting principles	R. Gelman and Gallistel (1978)
Syntax	Universal grammar	Chomsky (1965)
Phonology	Faithfulness constraints	Prince and Smolensky (1993)
	Markedness constraints	Prince and Smolensky (1993)
Music	Well-formedness rules	Lerdahl and Jackendoff (1983)
	Preference rules	Lerdahl and Jackendoff (1983)

Table 1.1: Constraints that guide inferences about several domains.

view proposes that only a handful of constraints need to be innate. The constraints in this class include properties of sensory transducers that determine how sensory data are represented, and constraints that take the form of a domain-general learning algorithm. Given this learning algorithm, all remaining constraints are thought to be learned from sensory input. A strong nativist view challenges the notion of domain-general learning, and proposes that learning is guided by strong constraints that are specific to individual domains—for example, that the acquisition of linguistic knowledge is guided by innate constraints that are specifically linguistic (Chomsky, 1986). Both sides of this debate must face some challenging questions.

The empiricist side must confront the problem of explaining how constraints might be acquired. At least two issues arise. First, if inductive learning is impossible without constraints, then any method for learning constraints must rely on meta-constraints of some sort, and we are faced with the threat of an infinite regress. Second, even if we grant that constraints might be learned in principle, it is difficult to understand how they are learned fast enough to be useful. Studies suggest that

many inductive constraints are available relatively early in development (Mehler et al., 1988; E. S. Spelke, 1990; Wynn, 1992), and there is a good reason to expect this result: constraints must be in place relatively early in order to support subsequent learning. Explaining how constraints can be learned is a difficult enough challenge, but explaining how they are learned rapidly is even harder.

The difficulty of explaining how constraints might be learned may explain in part why most discussions of constraints adopt a nativist perspective. Some constraints are almost certainly innate, but others appear to be learned, and a strong nativist account must address two challenges. First, how can humans successfully learn about novel contexts, including contexts that emerged only recently on an evolutionary timescale? Human reasoning is remarkably flexible, and our ability to reason about fields like mathematics, chemistry, and molecular biology stands in need of some explanation. Inductive constraints appear to play a role: for instance, skilled mathematicians rely on constraints which help them identify which of the many possible approaches to a problem is most likely to succeed (Polya, 1990). Similarly, most chess positions can be developed in many different ways, but expert chess players rely on constraints which prune away all but the handful of possibilities that turn out to be most promising.

The second challenge for a strong nativist view is that some of the constraints that guide inferences about more fundamental cognitive domains also appear to be learned (Goldstone & Johansen, 2003). One such constraint is the shape bias—the expectation that all of the objects in a given category tend to have the same shape, even if they differ along other dimensions, such as color and texture. Smith, Jones, Landau, Gershkoff-Stowe, and Samuelson (2002) provide evidence that the shape bias is learned by showing that laboratory training allows children to demonstrate this bias at an age before it normally emerges. Other constraints that appear to be learned include constraints on the rhythmic pattern of a child's native language (Jusczyk, 2003), and constraints on the feature correlations that are worth tracking when learning about artifacts or other objects (Madole & Cohen, 1995).

This thesis develops an approach that draws on ideas from both nativist and empiricist approaches to development. Consistent with a nativist approach, I ac-

knowledge that induction is impossible without constraints, and argue that human inferences are often guided by domain-specific constraints. Consistent with an empiricist approach, I focus on learning and argue that domain-specific constraints can be acquired by general-purpose learning mechanisms. Attempting to reconcile nativism and empiricism is not especially novel, and many psychologists presumably believe that their own theoretical orientation strikes the ideal balance between these philosophical traditions. This thesis suggests, however, that the dialogue between nativism and empiricism can be enriched by models that explain how constraints might be learned. Suppose, for instance, that we want to decide whether a certain kind of constraint is learned or innate. A good way to support an empiricist position is to provide a formal model that can acquire this constraint. A good way to support a nativist position is to develop the best possible strategy for acquiring the constraint, then to show that even this strategy must fail. Both approaches rely on formal models, and I attempt to show how these models can be developed.

The primary contribution of this thesis is a formal framework that helps to explain the nature, use and acquisition of inductive constraints. I explore models that include representations at multiple levels of abstraction, and where the representations at the upper levels place constraints on the representations at the lower levels. Each model is a hierarchical Bayesian model, and the probabilistic nature of these models allows them to make statistical inferences at multiple levels of abstraction. In particular, they show how knowledge can be acquired at levels quite remote from the data given by experience—levels where the learning problem can be described as the problem of learning inductive constraints.

Although I focus on the acquisition of inductive constraints, the larger goal of the work described here is to develop a comprehensive theory of human learning. Learning can be broadly defined as the acquisition of knowledge (Simpson & Weiner, 1989), and learning so defined includes topics like the acquisition of language and mathematical knowledge, the development of folk biology, folk physics, and folk psychology, and the development of scientific theories. As these topics suggest, the study of learning can help to explain the origin of human knowledge in all of its forms. Within

psychology, however, "learning" is sometimes given a technical meaning that is much narrower than its colloquial meaning. Kimble (1961) for instance, defines learning as a "relatively permanent change in a behavioral potentiality that occurs as a result of reinforced practice." Contemporary psychologists may prefer definitions that are less explicitly behaviorist, but the link between learning and behaviorism remains strong. Introductory textbooks, for instance, often include a chapter on learning that discusses classical and operant conditioning and little else.

A casual glance at an introductory textbook might suggest otherwise, but most psychologists agree that knowledge acquisition involves much more than the tracking of simple associations. Developmental psychology has been a particularly rich source of alternative views. Piaget, for instance, has argued that children create rich and systematic mental structures to explain their experience (Piaget & Inhelder, 1969), and other researchers have described learning mechanisms such as "bootstrapping" which go well beyond stimulus-response learning (Carey, 2004). The study of language has also led to alternative views of knowledge acquisition, and few contemporary researchers would argue that language acquisition can be explained by simple associative mechanisms. As these examples suggest, alternatives to associationism have been developed, but these alternatives have not led to the creation of a modern theory of learning. *Traditional* learning theory focused on the contributions of researchers like Thorndike, Pavlov, Hull, Tolman and Skinner (Hilgard & Bower, 1975). Although traditional learning theory has fallen out of favor, no modern equivalent has risen up to replace it.

Some psychologists will argue that there are good reasons to abandon the pursuit of a theory of learning. Traditional learning theory was based on the idea that a handful of general principles could explain how much of human knowledge was acquired. Perhaps, however, there can be no general theory of learning. If most forms of human learning are guided by domain-specific constraints, perhaps psychologists should aim for multiple theories of learning, one for each domain (Gallistel, 2000).

Even if different kinds of knowledge are acquired in very different ways, it will still be useful to identify general themes which apply across many different settings. The

(a) Traditional learning theory

1. Learning takes place at a single level of abstraction.
2. The representations learned are simple, and are often pairwise associations.
3. Animals are more prepared to learn some associations than others, but rich systems of prior knowledge play little role.
4. Formal models focus on cases where many training examples are observed.

(b) Modern learning theory

1. Learning takes place at multiple levels of abstraction.
2. Representations with rich and systematic structure can be learned.
3. Learning is guided by sophisticated, domain-specific knowledge.
4. Learning can succeed given sparse and noisy data.

Table 1.2: For many psychologists, "learning theory" has come to refer to the study of simple associative learning. Modern approaches to learning can differ from traditional learning theory along the four dimensions shown here.

aim is not necessarily to develop a monolithic theory of learning, but to understand the general principles that support learning in all of its forms. Four principles that seem particularly important are collected in Table 1.2b. The first principle recognizes that human knowledge is organized into multiple levels of abstraction, and that learning can take place at all of these levels. There are different proposals about how knowledge might be represented, but structured representations are useful for capturing rich systems of knowledge, and the second principle suggests that these representations can be learned. Some discussions of learning focus on what can be achieved with a minimum of prior assumptions, but the third principle recognizes that learning often relies on systems of domain-specific knowledge, some of which are listed in Table 1.1. The third and fourth principles are closely related, since systems of prior knowledge help to explain how humans can learn so much from sparse and noisy data.

Each principle in Table 1.2b has been emphasized by previous researchers, including some of the most prominent opponents of traditional learning theory. The third and fourth principles, for instance, are compatible with the approach of researchers like Chomsky who are usually regarded as nativists. The first and second principles are closely related to the constructivist approach of Piaget, who is much more of an empiricist than Chomsky, but is not usually regarded as a learning theorist (Hilgard & Bower, 1975). By developing a framework that incorporates all four principles, psychologists can lay the foundations of a modern theory of learning—a theory that incorporates the insights of researchers like Chomsky and Piaget, and that goes well beyond learning theory as it is traditionally conceived.

The hierarchical Bayesian framework I describe is consistent with all four principles in Table 1.2b, but this thesis will focus on the first principle. My primary goal is to explain how inductive constraints might be acquired, and I begin in Chapter 2 by reviewing existing views of inductive constraints and describing the criteria that a constraint-learning framework should satisfy. Chapter 3 introduces the hierarchical Bayesian framework that I will adopt, and the following chapters apply this framework to three inductive problems. Chapter 4 explores how constraints related to feature-variability (e.g. the shape bias) are acquired and used to support categorization. Chapter 5 considers the problem of causal learning, and introduces a model that helps to explain how causal schemata are acquired and used. Causal schemata can be viewed as systems of causal knowledge that place strong constraints on causal reasoning. The final application of the hierarchical Bayesian framework considers how learners might discover which kind of representation is best for a domain. Chapter 6 presents a model that discovers the structural constraints that characterize a given domain: for instance, the model discovers that anatomical features of biological species are best explained by a taxonomic tree, political views are best explained by a linear spectrum, and friendship relations are best captured by a set of discrete cliques.

Although I describe models that acquire some of the constraints listed in Table 1.1, I do not claim that all or even most of these constraints are learned. There are formal arguments, however, which suggest that all of these constraints are learnable

in principle (Solomonoff, 1978; Chater & Vitanyi, 2007). Given enough data that are consistent with a certain constraint, learners should be able to realize that this constraint is the best explanation for the data they have encountered. The real question for psychologists is whether the constraints in Table 1.1 can be learned given the data typically available to human learners. This thesis provides a necessary first step towards answering this question. Once we have a clear idea how constraints can be learned in principle, we can explore how feasible it is for constraints to be learned in practice.

Chapter 2

Inductive constraints

Although most researchers agree that induction is impossible without constraints, there are competing claims about the nature of these constraints. This chapter presents a taxonomy of constraints, and argues that the constraints which can be learned correspond to forms of abstract knowledge. I review several existing proposals about the acquisition of abstract knowledge, then argue for a hierarchical Bayesian approach to this problem.

A taxonomy of constraints

Different researchers work with different ideas about what can count as an inductive constraint. Nelson (1988) assumes that constraints are innate and domain-specific, and can be distinguished from soft preferences or biases. This thesis takes a more inclusive view, and suggests that there are many kinds of constraints. Unlike Nelson, I will suggest that some constraints are learned, some constraints are domain-general, and some constraints are soft. This section describes some of the dimensions along which constraints can vary, and identifies the kinds of constraints that can be learned by the framework I will introduce. Keil (1990), D. L. Medin et al. (1990) and R. Gelman and Williams (1998) have made previous attempts to chart the space of constraints, and the taxonomy I present draws on the perspectives of all of these authors.

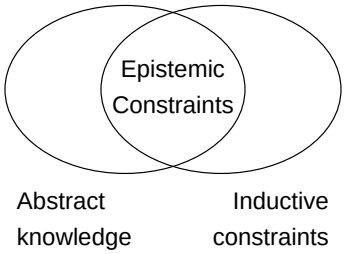

Figure 2-1: Epistemic constraints correspond to an abstract kind of knowledge. The framework developed in this thesis helps to explain how these constraints can be learned.

Epistemic versus non-epistemic

There is a fundamental distinction between constraints that correspond to an abstract kind of knowledge and constraints that do not. I will refer to constraints of the first type as "epistemic constraints" (Figure 2-1). Each of the principles of object perception identified by E. Spelke (1994) is an epistemic constraint. For instance, the principle of continuity makes a defeasible claim about the world—it states that "a moving object traces exactly one connected path over space and time" (E. Spelke, 1994). The M-constraint (Sommers, 1963; Keil, 1979) is a second example of an epistemic constraint, and corresponds to a claim about the possible relationships between sets of predicates and sets of arguments. As these examples suggest, epistemic constraints might alternatively be defined as constraints that can be represented as statements with truth values. Additional examples of epistemic constraints can be found in Table 1.1.

Non-epistemic constraints include mechanistic constraints of various kinds. Memory limitations are familiar examples: Newport (1990), for instance, suggests that some aspects of language acquisition are easier for children than adults because children are less able to keep track of the many potentially confusing details that they hear. Several authors argue similarly that early visual limitations (including poor acuity) may make object recognition easier rather than harder for infants (Turkewitz & Kenny, 1982; French, Mermillod, Quinn, Chauvin, & Mareschal, 2002). Note that memory limitations and perceptual limitations will both shape inductive inferences,

and will be responsible in part for the beliefs that a learner ends up acquiring. A non-epistemic constraint will usually have epistemic consequences, but the constraint itself must not correspond to a form of knowledge about the world.

Although there is an important difference between constraints like the Spelke principles and constraints like working memory limitations, the distinction between epistemic and non-epistemic constraints is not always perfectly clear. Constraints on the architecture of neural networks (Elman et al., 1996) include some borderline cases. Some of these constraints appear to be non-epistemic—for instance, the fact that mammalian cortex has six layers rather than seven or eight does not appear to correspond to any kind of knowledge about the world. Other architectural constraints might capture knowledge about the world that has been internalized through evolution. For instance, the particular recurrent structure of an auditory network might be viewed as implicit knowledge about the temporal properties of a certain kind of signal. As cases like these suggest, deciding whether a constraint is epistemic or not may sometimes require elaborate scientific investigation, and tentative decisions about the status of any given constraint may be overturned by future scientific discoveries.

Even if the boundary between epistemic and non-epistemic constraints turns out to be fuzzy, there are important differences between these classes of constraints. Since epistemic constraints can be associated with degrees of belief, it is natural to explore how these constraints might be learned. After seeing data consistent with an epistemic constraint, for instance, a learner might become more confident that the constraint is generally applicable. The idea that non-epistemic constraints might be learned usually makes less sense—for instance, it is not particularly useful to ask how a memory limitation might be learned. A strong conjecture is that the class of epistemic constraints is coextensive with the class of constraints that can be learned. It may turn out, however, that the distinction between epistemic and non-epistemic constraints is close but not identical to the distinction between constraints that can and cannot be learned.

Since this thesis explores how constraints might be learned, I will focus almost exclusively on epistemic constraints. My goal is to characterize the computational

benefits that epistemic constraints can bring, and the computational principles that allow these constraints to be acquired. Although this section has suggested that epistemic constraints can be represented as statements with truth values, I do not claim that these constraints are explicitly represented as propositions. Spelke's principles of object perception, for instance, can be represented as a set of propositions, and these propositions may help to explain the visual abilities of an infant, but it does not follow that these propositions are located somewhere within the infant's mind. Eventually it will be important to consider how epistemic constraints are represented, and to study the psychological mechanisms that operate over these representations. This thesis, however, provides a computational investigation of the nature and acquisition of epistemic constraints.

Domain-general versus domain-specific

Inductive constraints range from general expectations about a broad class of settings to expectations about a relatively narrow context. Consider, for example, the expectation that stimuli are often composed of modular units, and that the units which appear in one configuration might appear again in the future. Knowledge this general might apply across many domains—for instance, it might lead a learner to break visual scenes into configurations of visual objects, and auditory scenes into configurations of auditory objects (Kubovy & Van Valkenburg, 2001). Within any given domain, learners will often rely on more specific constraints. For instance, people have expectations about the characteristic motions of animals and vehicles, and rely on these expectations when interpreting the content of a visual scene.

This thesis describes a learning framework that will accommodate constraints at many points along the spectrum from general to specific. The key idea is that constraints can occupy different levels of abstraction: abstract constraints may apply across many domains, but less abstract constraints may hold only within a single domain. Even within a single domain, however, there may be general constraints (e.g. constraints on the properties of all human languages) and less general constraints (e.g. constraints on the morphology of a child's native language). I will therefore focus

more on the distinction between abstract and specific constraints than the distinction between domain-general and domain-specific constraints.

Innate versus learned

As suggested in the previous chapter, inductive learning is impossible without constraints. It follows that any method for learning constraints must rely on meta-constraints of some sort. It is natural to ask where these meta-constraints come from, and we can develop models that explain how they are learned with the help of meta-meta-constraints. We can continue to push the learning question up to higher levels, but eventually we must assume that the constraints at some level of abstraction are fixed from the start. I will refer to these constraints as *background assumptions* to distinguish them from constraints that can be learned.

The ultimate goal of this approach is to develop models where each background assumption corresponds to a form of innate knowledge. Constraint-learning models will usually fail to reach this goal, but may be useful nonetheless. For instance, a model that relies on a certain set of background assumptions can become a platform for future efforts to explain how these assumptions are acquired given new background assumptions that are simpler, more general, or both. Each of the models I present should be viewed in this way, and I do not propose that the background assumptions required by these models are innately provided.

Although it is clear that some inductive constraints must be innate, the nature of these constraints is a matter for psychological investigation. One important question is whether these constraints are domain-general or domain-specific (Chomsky, 1980; Keil, 1981; Elman et al., 1996), and the framework I present does not commit to either position. It is natural to aim for models that rely on background assumptions which are as simple and as general as possible, but it may turn out, for instance, that any adequate model of language learning will need to include background assumptions that are language-specific.

Soft versus hard

Some constraints are soft probabilistic expectations that might alternatively be called biases or preferences, and others are hard constraints that categorically rule out certain hypotheses. The framework I describe will have room for both kinds of constraints. Since I take a probabilistic approach, it will be natural to specify constraints that make some hypotheses more likely than others. A probabilistic approach, however, can also incorporate constraints that assign zero probability to some hypotheses.

Even though the world is complicated, simple constraints may still be useful as long as they are soft. It is obviously not the case that all English words refer to entire objects, but the whole object bias (Markman, 1989) may still be useful as long as it can be overruled when necessary. Similarly, the M-constraint (Sommers, 1963; Keil, 1979) captures a principle which appears to be useful in general, even though there may be exceptions to this principle (Carey, 1985b). Soft versions of constraints like these help to explain how human learning can be both highly constrained and highly flexible. When few observations are available, a learner may make inferences that are guided almost entirely by soft constraints. Once many observations are available, these soft constraints can be overruled, and a learner can make inferences that are guided primarily by the data she has observed. Both patterns of inference are important: together, they produce a learner who can make strong inductive inferences when data are sparse, but can learn almost anything given sufficient data.

Enabling versus limiting

There are two very different reasons to take an interest in inductive constraints. For some researchers, the most pressing goal is to explain how human inferences are so successful—to explain, for instance, how humans make inferences that go well beyond the capacities of our best formal models. In many situations, the observations made by a learner are consistent with a vast number of hypotheses, and the overwhelming problem is to identify the hypotheses that are most likely to be correct. Inductive constraints provide a critical part of the solution, since they narrow down the space

of hypotheses. Researchers who adopt this perspective tend to take a positive view of constraints, and argue that constraints deserve much of the credit for successful learning (Keil, 1990; R. Gelman & Williams, 1998). Other researchers begin with the problem of explaining why human inferences fall short of optimality in some settings. These researchers adopt a more negative definition of constraints, and reserve this term for factors (e.g. memory limitations) that rule out useful hypotheses and prevent learners from reaching optimal decisions.

This thesis will focus on constraints that enable rather than impede learning. My primary goal is to describe computational theories (Marr, 1982) that help to explain how people solve challenging inductive problems. Enabling constraints play a critical role in these theories, since they guide learners towards accurate conclusions when the available data are sparse or noisy. Computational theories, however, will never provide a complete account of cognition, and eventually it will be important to specify the psychological mechanisms that might carry out the computations required by these theories. Limiting constraints will become important at this stage, since we will need to explain why people's inferences sometimes fall short of the predictions made by computational theories (Anderson, 1991). Here, however, I attempt only to develop computational theories of human inference, and I leave detailed discussions of limiting constraints for future work.

Other distinctions

Although I have identified several dimensions which are relevant to psychological discussions of constraints, constraints may vary along several other dimensions. Many authors distinguish between constraints on the structure of mental representations, and constraints on the processes that operate over these representations (D. L. Medin et al., 1990). A related distinction is made by computer scientists when discussing the inductive bias of a learning system. The *representational* bias of a system characterizes the hypothesis space that will be explored by the learner, and the *procedural* bias determines the order in which the hypotheses will be explored (desJardins & Gordon, 1995). Although the distinction between structural and processing constraints may

be far from clean, I will focus on constraints that are probably best described as structural constraints.

A comprehensive taxonomy of constraints is likely to include dimensions that I have not discussed, and should also attempt to capture the relationships between different dimensions. This section made some effort in this direction—for instance, I suggested that any constraint that is learned is probably an epistemic constraint—but other regularities are also apparent (Keil, 1990). If we consider the constraints that are discussed most often in the psychological literature, enabling constraints tend to be epistemic constraints, learned constraints are often domain-specific, and soft constraints are often enabling constraints.

Developing a taxonomy of constraints is useful in part because researchers who disagree about the value of constraints (Nelson, 1988; Keil, 1990; Behrend, 1990; Deák, 2000) often seem to disagree about the meaning of this term. This thesis has adopted a very broad definition, and I will continue to use "constraint" to refer to any factor which bridges the inferential gap between a body of data and an inductive conclusion. I hope, however, that researchers who disagree with this usage (Deák, 2000) will agree that the value of my theoretical claims does not rest on any particular label used to describe them.

Constraints and abstract knowledge

Any taxonomy of constraints will include many dimensions, but the dimension most useful for picking out the constraints I will discuss is the distinction between epistemic and non-epistemic constraints. The framework I describe can model the acquisition of many different constraints, but each of these constraints must correspond to a kind of abstract knowledge. An alternative title for this thesis might have been "the acquisition of abstract knowledge," and my work is inspired in part by previous attempts to describe the nature, acquisition and use of abstract knowledge.

Many kinds of abstract knowledge have been discussed by philosophers, psychologists and computer scientists. Some of the most familiar examples are over-

hypotheses (Goodman, 1955), theories (Carey, 1985a; Wellman & Gelman, 1992; Kuhn, 1970), schemata (Kelley, 1972; D. E. Rumelhart, 1980), learning sets (Harlow, 1949), scripts (Schank & Abelson, 1977), and frames (Minsky, 1975). Each kind of knowledge can be viewed as an abstract representation that places constraints on representations at lower levels of abstraction. For instance, an overhypothesis sets up a more concrete space of hypotheses, a learning set captures expectations that a learner brings to a specific learning problem, and a theory captures general principles that make predictions about specific phenomena that fall under the theory. There are important differences between these varieties of abstract knowledge, but I will focus on the similarities rather than the differences. For instance, each kind of abstract knowledge suggests the need for inferential frameworks that include multiple levels of abstraction.

Figure 2-2 shows several cognitive domains where knowledge can be organized into several levels of abstraction. Language (Chomsky, 1957), vision (Han & Zhu, 2005) and action (Cooper & Shallice, 2000) provide the most familiar examples (Figure 2-2). We can hear a speech signal and recognize the phonemes and words that it contains. We may know how a given sentence should be parsed, and we may know a grammar which allows us to parse many different sentences. In the visual domain, we know which objects are likely to appear in a street scene or an office scene, we know about the parts of these objects, and we have some idea about the shapes and the relative orientations of the surfaces that compose these parts. Our abilities to form plans and carry them out can also be described at several levels—for instance, we know how to make coffee, and how to open a packet of sugar.

Figures 2-2c, 2-2e and 2-2f show hierarchies that address three aspects of high-level cognition. When grouping items into categories, we rely on knowledge about specific categories (balls tend to be round) as well as general knowledge about patterns of feature variability (all instances of a given object category tend to have the same shape). Causal inferences can draw on knowledge about specific entities (Lariam pills tend to cause headaches) as well as more abstract kinds of knowledge (medications may cause headaches). Learning structured representations may also require infer-

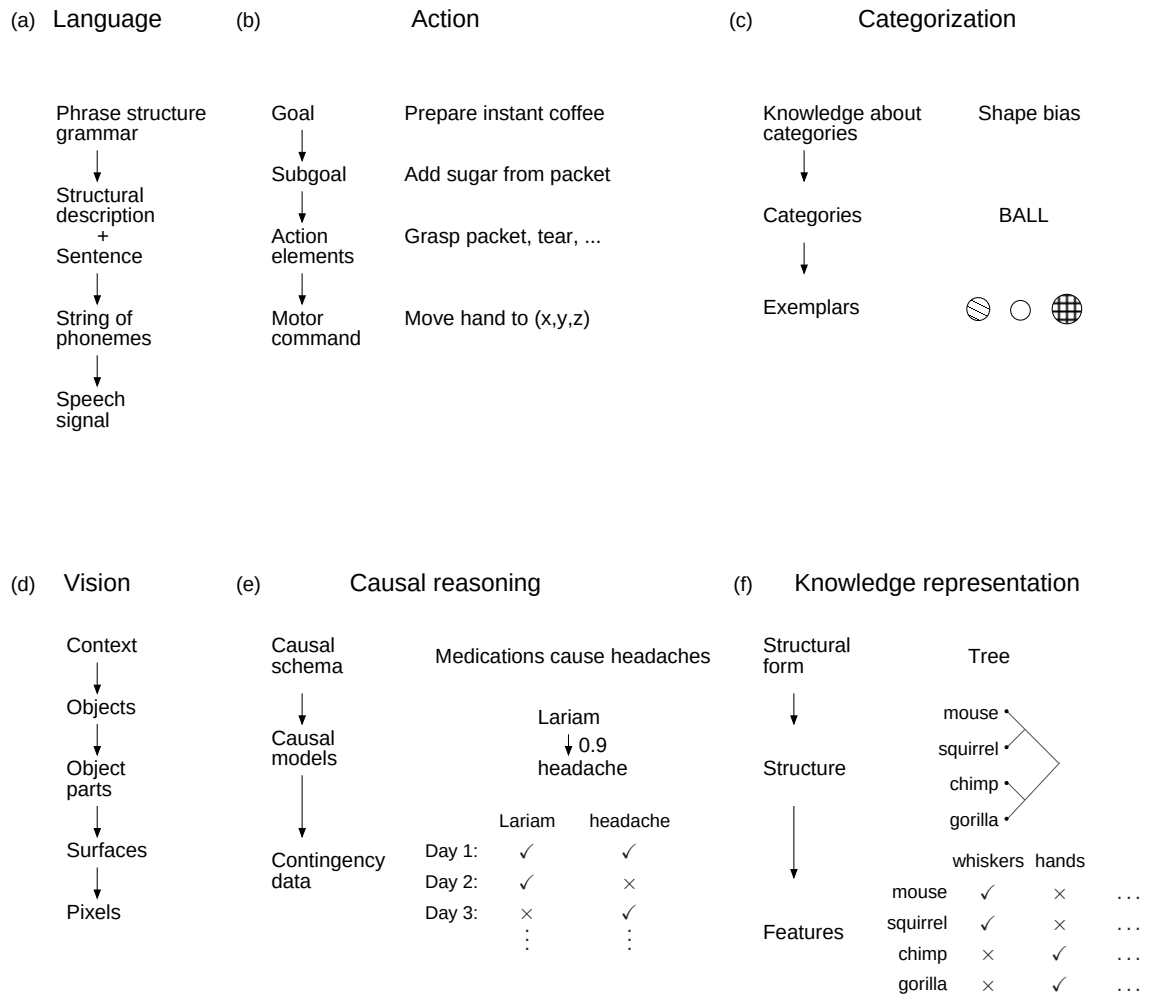

Figure 2-2: Systems of knowledge are often organized into several levels of abstraction. (a)(b)(d) Hierarchies are useful for understanding language, action and vision. (c)(e)(f) The later chapters of this thesis describe hierarchical approaches to categorization, causal reasoning and knowledge representation.

ences at several levels of abstraction. The most general problem is to decide whether the relationships between a set of entities are best captured by a tree, a ring, a set of clusters, or some other kind of representation. If the entities, say, are believed to belong to some latent tree structure, the next problem is to identify the specific tree that best accounts for the available data.

The hierarchies in Figure 2-2 all rely on multiple levels of abstraction, but the representations at adjacent levels are related to each other in many different ways. Let R_i be a representation at level i in a hierarchy, where R_1 is a representation at the lowest level. In some cases, R_i *is a* R_{i+1}—for instance, a bouncy round object may be a ball. R_i may also be *part of* R_{i+1}—for instance, an object part is a constituent of an object. Although *is a* relationships and *part of* relationships are often used to construct hierarchies, many other relationships are possible. For instance, a speech signal is a realization of a string of phonemes, and a structural description can be built from a grammar.

Since many kinds of relationships between levels are possible, the hierarchies I consider include examples (e.g. Figure 2-2a) that go beyond simple class-inclusion hierarchies (Collins & Quillian, 1969). A hierarchy can be defined as a system of latent variables that captures expectations about the data observed at the bottom of the hierarchy. In most cases, the levels in the hierarchy will not correspond to simple summaries of the observable data. Instead, the levels are best viewed as components of a system which explains the observable data. The role of these levels is therefore similar to the role of the concepts in a scientific theory. As philosophers have argued, scientific concepts are more than simple abstractions from experience. Scientific concepts are components of theories, and it is entire theories that make contact with experience (Hempel, 1972).

The notion of an abstraction hierarchy is the starting point for the formal framework described in the next chapter. I formalize this notion using nested hypothesis spaces: X is more abstract than Y if X sets up a hypothesis space that can be used when learning Y. The framework I describe supports hierarchies with multiple levels of abstraction, and the representations at the upper levels can be viewed

as epistemic constraints. Since I take a probabilistic approach, Bayesian inference can explain how the constraints at the upper levels are learned given observations at the bottom level of the hierarchy. Even though I will focus on relatively simple constraints, the hierarchical Bayesian approach can help to explain the acquisition of many kinds of abstract knowledge, including representations that might be best described as schemata or intuitive theories.

Conceptual approaches to constraint learning

The acquisition of abstract knowledge has been a central concern for epistemologists and developmental psychologists alike. This section introduces three views of knowledge acquisition that have been popular in the psychological literature. One prominent approach grows out of the work of the British empiricists (Locke, 1998; Hume, 1748), who argued that even our most abstract ideas correspond to combinations of perceptual primitives. Abstract knowledge is thought to emerge when associative learning mechanisms combine these primitives to create new concepts. Some kinds of abstract knowledge may correspond to higher-level associations, or associations between associations (Colunga & Smith, 2003). As discussed in the next section, connectionist models can be viewed as modern attempts to formalize associative learning.

Piaget and his colleagues developed an alternative empiricist approach that emphasizes the construction of increasingly abstract cognitive structures (Piaget, 1970; Piaget & Inhelder, 1969). This constructivist approach suggests that infants begin with relatively simple perceptual and motor abilities, and move through a series of increasingly complex stages. Each stage is characterized by the kinds of representations that are available and the operations that can be carried out over these representations. These computational resources can be viewed as domain-general constraints: for example, the concrete and formal operations are abstract structures that help learners to address problems from many different domains. Two mechanisms are thought to explain how children move from one stage to another: *assimilation*, or the

integration of external elements into a structure, and *accommodation*, or the modification of a structure by the elements it assimilates. The interaction between these mechanisms is believed to lead to the emergence of abstract knowledge.

A third view works with the idea that abstract knowledge is embedded in *theories*, or rich systems that specify concepts and relationships between these concepts. This theory-based approach is clearly relevant to the study of scientific knowledge, but psychologists have proposed that much of our everyday knowledge is organized into *intuitive theories* that are similar to scientific theories in important respects (Carey, 1985a; Murphy & Medin, 1985; Wellman & Gelman, 1992; Gopnik & Meltzoff, 1997). From this perspective, the problem of understanding how abstract knowledge is acquired turns into the problem of characterizing the process of theory formation. Unlike associationism and constructivism, the theory-based approach need not explain how theories are built from raw perceptual primitives: many supporters of this approach suggest that infants begin with innate theories of several core domains, and that learning is a matter of moving from one theory to another (Gopnik, 1996; Wellman & Gelman, 1998). High-level descriptions of theory change are sometimes provided (Popper, 1935/1980; Kuhn, 1970)—to mention one typical example, Gopnik (1996) suggests that theory formation is a matter of accumulating counterevidence to an existing theory, proposing an alternative theory, then searching for evidence for this new theory. Accounts like this are convincing as far as they go, but understanding theory change in detail remains a major challenge.

This section has described three influential approaches to the acquisition of abstract knowledge, and these approaches have inspired many of the ideas in this thesis. Although I draw on previous work from the philosophical and psychological literature, I take on the challenge of developing computational theories that go beyond verbal descriptions of the emergence of abstract knowledge. The next section describes some of the issues that arise when attempting to model the acquisition of abstract knowledge.

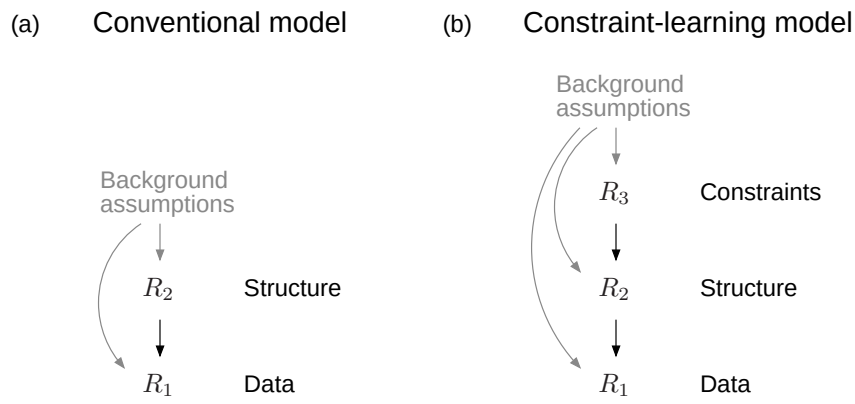

Figure 2-3: (a) Cognitive models typically make inferences at two levels. Given data (e.g. contingency data or a set of features), these models learn a latent representation R_2 (e.g. a causal model or a set of network weights) and use this representation to make inferences about any missing entries in the data (R_1). These models always rely on background assumptions which must be fixed in advance. (b) A constraint-learning model makes inferences at three or more levels. The model shown here relies on background assumptions which are fixed, but can learn inductive constraints (R_3), discover a latent representation (R_2), and fill in missing data (R_1). The models developed in Chapters 4, 5 and 6 are all instances of this schema.

Formal models of constraint learning

Many cognitive models can be seen as instances of the two-level schema shown in Figure 2-3a. The schema indicates that observable data R_1 are consistent with some underlying structure R_2, and that the representations at both levels depend on a set of background assumptions. These assumptions might equally well be described as inductive constraints, and include assumptions about the class of possible structures, the class of possible data sets, and the relationship between structure R_2 and the data observed at level 1. The background assumptions are fixed in advance and grayed out in Figure 2-3a, but the schema supports inferences at level 1 and level 2.

Some concrete examples may help to explain the schema in Figure 2-3a. Models of causal learning often work with contingency data (R_1), and discover a causal network (R_2) that accounts well for the patterns in the data. All of these causal models rely on background assumptions of some sort: for instance, Bayesian approaches use a prior distribution on causal networks, and make some assumptions about how data

are generated from these networks. All of these models support inferences at two levels: they learn a causal network R_2, and can use this network to make inferences about any missing entries in the data set R_1.

To give a second example, connectionist models often take a collection of features as input (R_1), and learn a set of weights (R_2) that accounts well for the data. These models rely on background assumptions which may include assumptions about the architecture of the network, the initial state of the network and the parameters that specify the learning rule (Elman et al., 1996). Since the learned network weights (R_2) support predictions about unobserved features (R_1), again the model makes inferences at two distinct levels.

To give a third and final example, models for multidimensional scaling (MDS) begin with a similarity matrix (R_1), and discover a low-dimensional representation (R_2) that accounts well for the data. Again, these models rely on background assumptions which may or may not be explicitly stated: for instance, MDS models assume in advance that a spatial representation is appropriate for the data. In principle, MDS models make inferences at two levels: they discover a representation R_2, and can use this representation to make predictions about any pairwise similarity ratings that are missing from the data set R_1.

From one perspective, any conventional model of learning (Figure 2-3a) acquires inductive constraints, since it learns a representation R_2 which shapes inductive inferences about the unobserved entries in data set R_1. The conventional approach, however, does not provide a general framework for explaining how constraints might be learned. Many constraints of interest correspond to assumptions about representation R_2, and the background assumptions in Figure 2-3a will always include constraints of this sort. To explore how these constraints might be learned, we need models with at least three levels of abstraction, and the simplest models that satisfy this criterion are instances of the schema in Figure 2-3b. This schema indicates that data are generated from some underlying structure, that this structure conforms to a set of constraints (R_3), and that the representations at all levels are consistent with a set of background assumptions. The background assumptions are fixed in advance

and grayed out in Figure 2-3b, but the schema supports inferences at levels 1, 2 and 3.

Given any two-level model (Figure 2-3a), we can move to a three-level model by carving out some of the background assumptions about structure R_2 and introducing them as a level in their own right. Some additional assumptions will need to be added as we move from two to three levels—in particular, we will need to add background assumptions that capture expectations about the representation at level 3. The aim, however, is to achieve a net reduction in background assumptions whenever we add a level. There is no reason to stop at just three levels, and we can continue to add levels, again aiming to shrink the set of background assumptions at each stage.

The schema in Figure 2-3b does not explain how *all* of the constraints that guide inferences about R_2 and R_1 might be learned. As in Figure 2-3a, the background assumptions might equally well be described as inductive constraints, but I have chosen a different label to distinguish the constraints that are learned (R_3) from the constraints that are not (the background assumptions). Any model of learning will rely on some set of background assumptions, and the schema in Figure 2-3b is no exception. This schema, however, can help to explain the acquisition of many constraints discussed by psychologists, including many of the constraints in Table 1.1.

Whether a given model matches the schema in Figure 2-3b will depend on what it learns and what it takes as input. Consider, for example, two methods for learning probabilistic context-free grammars. The first is a supervised model and takes a set of parse trees as input. The second is an unsupervised model: it takes a set of sentences (R_1) as input, and must discover parse trees for each sentence (R_2) and a grammar (R_3) that accounts well for these (unobserved) parse trees. Even though the two models may discover identical grammars, only the unsupervised model qualifies as a constraint-learning model. The unsupervised model deserves this description since the grammar it learns captures constraints which help to solve the inductive problem of parsing. As required by Figure 2-3b, it makes inferences at three levels: it learns a grammar and a set of parse trees, and if any of the sentences contain words that are garbled or unobserved, it can predict what those missing words might be. The

supervised model does not address the parsing problem, and makes inferences at only two levels of abstraction: it discovers a grammar, and can use this grammar to fill in parts of the parse trees that might have been unobserved.

Researchers from several disciplines have developed formal models that help to explain the acquisition of abstract knowledge. Some, but not all of these models match the three-level schema shown in Figure 2-3b. These models can be organized into four broad classes: connectionist approaches, AI approaches, machine learning approaches, and statistical approaches. Note, however, that these classes overlap, and that some models are valid representatives of two or more classes.

Connectionist approaches

Connectionist models represent a modern attempt to implement some of the ideas behind associationism, and some of these models appear to acquire knowledge at multiple levels of abstraction. A particularly clear example is provided by Colunga and Smith (2005), who show that a recurrent network can acquire a shape bias for object categories and a material bias for substance categories. In other words, the network learns about the features of specific categories (balls tend to be round) and about the features of categories in general (all instances of a given object category tend to have the same shape). Constraints on word-learning have also been explored: for example, Regier (2003) reviews work suggesting that something like the principle of mutual exclusivity can emerge from associative learning. Finally, I suggested earlier that the abstract knowledge which guides induction can sometimes be described as an intuitive theory, and Rogers and McClelland (2004) argue that connectionist models provide a mechanistic account of many inductive phenomena that are commonly thought to rely on intuitive theories.

There is at least one kind of connectionist model that matches the three-level schema in Figure 2-3b. Cascade-correlation models grow in complexity as more data are encountered: in other words, they learn both the architecture of a network (R_3) and the weights for this network (R_2) (Fahlman & Lebiere, 1990; Mareschal & Shultz, 1996). Cascade-correlation models seem particularly appropriate for modeling cogni-

tive development, and psychological applications of these models are often inspired by Piaget's constructivist approach to development (Shultz, 2003). Most connectionist models, however, learn at only two levels—a level which includes the data and a level which specifies the weights of the network. These models match the schema in Figure 2-3a, and raise the question whether models that make inferences at three or more levels are necessary to account for the acquisition of abstract knowledge. I take up this question towards the end of this chapter.

AI approaches

Logic provides a powerful tool for representing inductive constraints, and the artificial intelligence research community has long been interested in methods for learning logical representations. Some of these methods can be viewed as techniques for learning the epistemic constraints that are the focus of this thesis. One class of methods (Davies & Russell, 1987) aims to identify determinations, or abstract logical statements that identify patterns of dependency between attributes. For example, the statement that "people in a given country usually speak the same language" is a constraint that supports confident generalizations from very sparse data. A visitor to Brazil, for example, can conclude that Brazilians speak Portugese after meeting a single Portugese-speaking local (Russell & Norvig, 2002).

Another approach is known as Explanation-Based Learning, or EBL (Mitchell, Keller, & Kedar-Cabelli, 1986; DeJong & Mooney, 1986). Given a single observation of a novel concept, EBL systems attempt to identify a schema (or a set of general rules) that include the individual observation as a special case. Suppose, for example, that an EBL system is given a single example of a kidnapping narrative—a story about Mary, who was kidnapped by Bill when she was out running one evening, and released only when Mary's father gave three hundred thousand dollars to Bill (DeJong & Mooney, 1986). An EBL system will attempt to identify general rules that are true of all kidnapping narratives. For instance, a kidnapping narrative is one where a person x captures another person y, and x releases y only when associates of y pay x money. Note that this schema abstracts away from the idiosyncratic details

of the story provided—kidnapping victims are not always captured while running, and the ransom can vary from case to case. EBL approaches usually work with a domain theory, or a collection of background knowledge that is usually expressed as a collection of logical statements. Although conventional EBL approaches are only able to learn schemata that are deductive consequences of the background theory, some systems attempt to combine EBL with inductive learning.

A third approach is known as inductive logic programming, or ILP (Muggleton & De Raedt, 1994; Quinlan, 1990). Given a set of observations, ILP systems attempt to find the simplest logical theory that accounts for the data. For instance, given information about the kinship relations between a large set of people (Andrew is Alice's father, Chris is Andrew's brother, Chris is Alice's uncle, etc.), an ILP system attempts to discover logical rules that allow the observations to be concisely expressed (the brother of one's father is one's uncle).

Although methods for learning functional dependencies can learn at three levels of abstraction, most formulations of EBL and ILP learn only at two levels: the level of the data, and a level which includes a logical representation of the data. All of these approaches acquire abstract knowledge, which again raises the question whether multiple levels (Figure 2-3b) are needed to account for the inferences I wish to explain.

Machine learning approaches

The overlap between the machine learning community and the AI community is substantial, but these two communities have produced literatures on constraint learning that are somewhat distinct. Much of the relevant machine learning research is found in the literature on transfer learning (also known as "lifelong learning," "multitask learning," or "learning to learn"). The idea behind transfer learning is that an agent who has faced several inductive problems should be able to extract regularities (or inductive constraints) that will help it deal with the next problem it encounters (Thrun & Pratt, 1998). Transfer learning has been approached from several angles: Ando and Zhang (2005) and Baxter (1997) provide theoretical analyses of the problem, and there are many heuristic approaches which have not been given a principled justification,

but which yield good empirical performance on selected real-world problems (Caruana, 1997).

The problem of learning inductive constraints is also discussed within the small but growing literature on developmental robotics. The goal of this work is to design robotic agents that begin with low-level motor and sensory data, and bootstrap their way to higher-level ontologies that include knowledge about objects, actions, and the structure of physical space (Kuipers, Beeson, Modayil, & Provost, 2006). Ontological knowledge of this sort provides strong inductive constraints that can help an agent to solve specific inductive problems—for instance, knowing that objects persist in time should help an agent to understand that some specific object of interest still exists even if it is currently occluded.

Statistical approaches

Bayesian statistics provides a principled framework for understanding inductive inference, and the next chapter shows in some detail how probabilistic models can be defined over hierarchies that include representations at multiple levels of abstraction. The resulting models are known as hierarchical Bayesian models (A. Gelman, Carlin, Stern, & Rubin, 2003; Tenenbaum, Griffiths, & Kemp, 2006), and these models support statistical inferences about the representations at all levels of abstraction. In particular, they show how the abstract knowledge at the upper levels of a hierarchy can be acquired given observations only at the lowest level.

My approach to constraint learning

As the previous sections suggest, there are several formal approaches to the problem of learning inductive constraints. I will adopt the hierarchical Bayesian approach, and this choice can be justified on several grounds.

Why Bayes?

As Marr (1982), Anderson (1990) and others have emphasized, cognition can be studied at several levels. Some researchers focus on neural mechanisms, others focus on cognitive processes, and others attempt to understand the computational principles that support our cognitive abilities. Ultimately it will be important to understand cognition at all of these levels, but often it is useful to start at the level of computational theory. Until we clearly understand the nature of a given cognitive problem, it is difficult to make useful proposals about the psychological or neural mechanisms that might contribute to its solution.

Since there are few computational theories of constraint learning in the psychological literature, our first task is to identify the computational principles that allow constraints to be learned. Computational theories of cognition do not always rely on Bayesian methods (Marr, 1982), but computational theories of learning often do. Bayesian statistics provides a normative account of inference under uncertainty, and is useful for exploring the principles that allow a learning system to succeed given sparse and noisy data. Bayesian approaches have previously been used to model many cognitive abilities, including stimulus generalization (Shepard, 1987), categorization (Anderson, 1990), reasoning (Oaksford & Chater, 1994), causal learning (Glymour, 2001), property induction (Heit, 1998), and word learning (Xu & Tenenbaum, 2007). The models in this thesis are motivated by some of the same goals as these previous approaches, and share many of their strengths and limitations.

The Bayesian approach offers several advantages over the connectionist approach, which is the main alternative available in the psychological literature. We have already seen that models with multiple levels of abstraction are useful for explaining how constraints are acquired and used (Figure 2-3b). As discussed in the next chapter, Bayesian models naturally handle multiple levels of abstraction. Connectionist networks do not clearly distinguish between knowledge at different levels of abstraction, and it is difficult to analyze a successful network and decide which constraints are responsible for its success, and how they might have been acquired. The connec-

tionist approach has been useful for developing models of psychological processing, but is not ideal for developing computational theories of constraint learning.

A second advantage of the Bayesian approach is that it naturally handles structured representations. Many inductive constraints are thought to emerge from intuitive theories (Keil, 1991), and these theories are perhaps best captured using structured representations. Other inductive constraints are explicitly formulated as constraints on structured representations: for example, the M-constraint states that ontological knowledge is better described by a tree structure than by a set of arbitrarily overlapping clusters (Keil, 1979), and Universal Grammar may specify many constraints that set up a hypothesis space of possible grammars. Some researchers have explored "structured connectionist models" (Smolensky, 1990; Regier, 1996), but the connectionist approach has struggled in general to account for inferences that appear to rely on structured representations.

A third strength of the Bayesian approach is the clarity it brings to the debate between nativism and empiricism. Bayesian methods make two key contributions to this debate. First, they provide an upper bound on the abilities of a human learner: if a Bayesian learner cannot acquire a certain kind of knowledge from a given initial state, then a human learner must also fail to learn in this situation. Second, the Bayesian approach requires a modeler to clearly specify the background knowledge that supports inductive learning. Typically this knowledge is captured by a prior distribution, and a set of assumptions about how observable data are generated.

A final advantage of the Bayesian approach is its ability to handle noise and exceptions, and to account for the graded generalizations that are characteristic of human inferences. Connectionist models share this advantage, but some of the logical models developed within the AI community have found it difficult to tolerate noise and exceptions. Models that combine logic and probability are an important exception (Milch et al., 2005; Kok & Domingos, 2005), but models of this sort tend to be compatible with the Bayesian approach advocated here.

The greatest limitation of the Bayesian approach is that at best it will provide an incomplete account of human learning. Understanding the computational principles

that guide human learning is a good start, but understanding how these principles are implemented by the mind and the brain will also be important. Successful computational theories can guide investigations of psychological and neural mechanisms, but understanding these mechanisms in detail will require insights that a Bayesian analysis is unable to provide.

Why *hierarchical* Bayes?

Most Bayesian models in the psychological literature match the schema in Figure 2-3a, and make inferences at only two levels of abstraction. These models are useful for many purposes—for instance, they help to explain how inductive inferences are guided by prior knowledge, which can also be described as a collection of epistemic constraints. My aim, however, is to describe models that simultaneously explain how inductive inferences rely on constraints and how these constraints might be acquired. At a minimum, we will need models that match the schema in Figure 2-3b and distinguish between three levels of abstraction. The next chapter describes how the hierarchical aspect of the hierarchical Bayesian approach allows us to capture as many levels as we need for a particular problem.

As mentioned already, some computational methods for acquiring abstract knowledge do not explicitly distinguish between multiple levels of abstraction. Connectionist approaches view abstract knowledge as an emergent property of a learning system: in other words, abstract knowledge is somehow implicit in the connection weights learned by the system. Although connectionist networks can capture some aspects of knowledge acquisition, there are several reasons for working with explicit hierarchies like the examples in Figure 2-2.

Hierarchies are valuable in part because they provide a clean way to transfer knowledge from one context to another. As a computer scientist might say, *abstraction* is valuable because it promotes *reuse*. Consider, for instance, the problem of learning about the causal powers of a collection of medications (Figure 2-2e). One option is to learn a causal model for each medication separately, but this approach does not capture the intuition that learning about 10 medications should shape our

expectations about medication number 11. An alternative approach might learn a single causal model that describes medications in general, but this approach cannot acquire specific information about individual medications (e.g. that medication number 3 is particularly likely to cause headaches). Instead of treating all the medications separately or collapsing them into one big category, we can allow two levels of abstraction—one for medications in general and one for individual medications—and carry out inferences at both of these levels (Figure 2-2e). Similar approaches are useful when learning about many categories (Figure 2-2c), or learning about the appearance and behavior of many physical objects. In general, hierarchies provide an appealing solution to the problem of sharing information between related contexts while maintaining the potentially important distinctions between these contexts.

Connectionist networks have traditionally struggled with the problem of learning about contexts that are related but distinct. Networks which attempt to handle several contexts are often subject to *catastrophic interference* (McCloskey & Cohen, 1989), which occurs when information about a new context interferes with knowledge that has previously been acquired. When a network is applied to a single context, a modeler may notice emergent network properties that appear to correspond to forms of abstract knowledge (Rogers & McClelland, 2004). Unless the network can transfer these emergent properties to new contexts, however, it is not clear that any abstract knowledge has actually been acquired.

There are at least two additional reasons to pursue a hierarchical approach. Hierarchies are desirable in some cases because they lead to the simplest explanation of some phenomenon of interest. Suppose, for instance, that we want to understand how people decide whether a string of phonemes qualifies as a valid English utterance. It is possible in principle to develop a non-hierarchical model (Figure 2-3a) that directly characterizes all grammatical strings of phonemes. Chomsky (1975), however, argues that this project amounts to an "immense and unmanageable" task. A better approach is to introduce levels for morphemes, words, and phrases, and to characterize the grammaticality of a phoneme string in terms of all of these levels. Even if current technology provides no way to directly probe the psychological reality

of the representations at the more abstract levels, linguists can argue for the existence of these representations by showing how they contribute to the linguistic theory that is simplest overall. Similar considerations apply in non-linguistic settings, including the cases shown in Figure 2-2.

Considerations of theoretical simplicity can provide indirect support for a hierarchical approach, but direct evidence for multiple levels of abstraction is available in some settings. Suppose that a learner is exposed to contingency data that provide evidence about the effects of several different medications (Figure 2-2e). A successful learner may make statements that reflect representations at all three of the levels in Figure 2-2e. For instance, the learner may say that "Jane had a headache on June 14" (bottom level), that "Lariam causes headaches" (middle level), and that "medications cause headaches" (top level). The ability to learn from statements like these provides further evidence for the existence of multiple levels of abstraction. For instance, a learner who is told that "Lariam causes headaches" is likely to learn about the causal powers of Lariam much quicker than a learner who is given contingency data alone. As these examples suggest, verbal reports can provide strong evidence for the existence of multiple levels of abstraction, and informal analyses can be followed up by experimental manipulations that explore how inferences change when abstract knowledge is directly provided.

This section provided several reasons to develop models with multiple levels of abstraction, but two-level models (Figure 2-3a) may satisfy all of our requirements as long as the representation at level 2 can distinguish between different sublevels. For instance, methods for learning logical theories (e.g. ILP) can learn a single representation that includes both general statements (e.g. $\forall \mathtt{x}\, \forall \mathtt{y}\, \mathtt{Spouse}(\mathtt{x}, \mathtt{y}) \leftarrow \mathtt{Spouse}(\mathtt{y}, \mathtt{x})$) and specific facts (e.g. $\mathtt{Spouse}(\mathtt{Sally}, \mathtt{Andrew})$). For our purposes, it will not be critical to decide whether these general statements should occupy a sublevel within level 2, or should belong to a distinct level in their own right. As long as we agree that representations at multiple levels of abstraction are needed, there is room for debate about how best to organize these representations into levels and sublevels.

My contribution

Previous work in psychology and philosophy raises a fundamental question: how can inductive constraints be acquired? Previous work in machine learning and statistics has led to a theoretical approach—the hierarchical Bayesian approach—that explains how knowledge can be acquired at multiple levels of abstraction. This thesis brings these two literatures together and argues that the hierarchical Bayesian approach helps to explain how people learn inductive constraints.

To support this argument, the later chapters of this thesis describe computational theories that address three aspects of high-level cognition: categorization, causal reasoning, and knowledge representation. Computational theories of cognition derive support in several ways. Like all theories, they should be judged according to their coherence, elegance, and explanatory power. Like other psychological theories, they can be evaluated according to their ability to account for behavioral data. Finally, computational approaches derive support from demonstrations that they can be implemented by psychologically plausible mechanisms. I will focus on the first two criteria and will leave the third for future investigation.

Statistics and machine learning provide a sound theoretical foundation for models of human learning, but psychologists can repay the debt by suggesting new problems for these fields to explore. Although I focus on the psychological implications of the models I describe, each model may also find applications to machine learning problems. The first model is a modest extension of a familiar statistical model (the Dirichlet-multinomial model), but the remaining two models represent more of a departure from existing statistical models. Understanding human learning is a worthy goal in its own right, but progress towards this goal should also lead to machine learning systems that are better able to match the sophistication of human learning.

Chapter 3

Hierarchical Bayesian models

The previous chapter suggested several criteria that a constraint-learning framework should satisfy. It must allow representations at multiple levels of abstraction, and support inferences at all of these levels. It must allow adjacent levels to depend on each other in many different ways. Finally, it must be able to tolerate sparse and noisy data. We satisfy all of these criteria by taking a hierarchical Bayesian approach (Lindley & Smith, 1972; Good, 1980; A. Gelman et al., 2003).

To convert any of the examples in Figure 2-2 into a fully specified model, the first step is to formalize a set of hypothesis spaces, one for each level of abstraction. Let H_i be the hypothesis space at level i, and let R_i refer to one of the elements in this hypothesis space (Figure 3-1a). In Figure 2-2a, for example, the hypothesis space at level 2 (H_2) includes all possible strings of phonemes, and R_2 refers to one particular string of phonemes. Background assumptions are needed to set up the hypothesis spaces at each level, but are not shown in Figure 3-1. I will adopt the same convention in all remaining figures: background assumptions are always present but never shown.

After specifying hypothesis spaces at each level of abstraction, a hierarchical Bayesian model can be defined by placing a prior distribution $P(R_n)$ on the space at the top level, and by specifying distributions $P(R_{i-1}|R_i)$ which indicate how the representation at each level generates the representation at the next level down. By specifying different distributions $P(R_{i-1}|R_i)$ we can capture many kinds of relation-

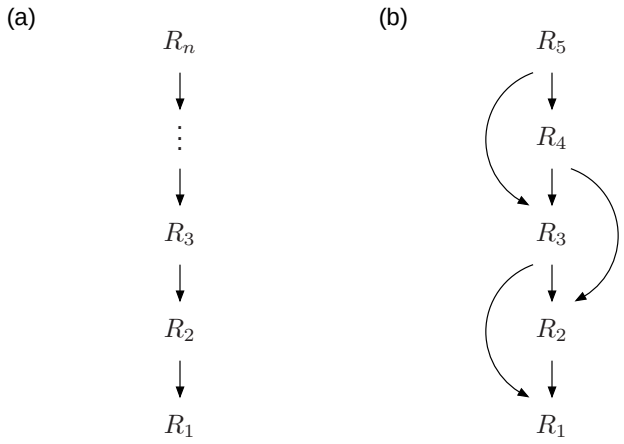

Figure 3-1: (a) A hierarchical model with representations R_i at multiple levels of abstraction. (b) Hierarchical Bayesian models can allow many patterns of dependence between levels.

ships between adjacent levels, including *is a* relationships, *part of* relationships and many other possibilities. When combined, these distributions define a joint distribution over the set of representations at all levels:

$$P(R_1, R_2, \ldots, R_n) = P(R_1|R_2)P(R_2|R_3)\ldots P(R_{n-1}|R_n)P(R_n). \tag{3.1}$$

All of the distributions in Equation 3.1 depend on a set B of background assumptions about the hypothesis space at each level and the process by which each level is generated from the level immediately above. We can make these assumptions explicit by rewriting Equation 3.1 as

$$P(R_1, R_2, \ldots, R_n|B) = P(R_1|R_2, B)P(R_2|R_3, B)\ldots P(R_{n-1}|R_n, B)P(R_n|B),$$

but we will keep our notation simple and again adopt the convention that background assumptions are always present but never shown.

The joint distribution in Equation 3.1 contains enough information to model inferences about any set of levels given observations at any other set of levels. If we are working with a five level model, for instance, and representations at three of the levels are known (R_1, R_3 and R_5) then the joint distribution induces a conditional

distribution $P(R_2, R_4|R_1, R_3, R_5)$ that can capture inferences about the remaining two levels in the model.

Hierarchical Bayesian models can include any number of levels, but three levels are enough to demonstrate the main message of this thesis. We will focus on models that match the three-level schema shown in Figure 2-3b, and Figures 2-2c, 2-2e and 2-2f show the three instances of this schema that we will consider in detail. Each model assumes that the relationships between levels form a linear structure, and Equation 3.1 also makes this assumption. Technically speaking, we have assumed that each representation R_{i-1} is conditionally independent of the representations at all higher levels given the representation R_i at the next level up. This assumption, however, can easily be relaxed, and hierarchical Bayesian models can capture many patterns of dependence between levels, including the case shown in Figure 3-1b. The joint distribution for this model is

$$P(R_1, R_2, R_3, R_4, R_5) = P(R_1|R_2, R_3)P(R_2|R_3, R_4)P(R_3|R_4, R_5)P(R_4|R_5)P(R_5)$$

and again we can use this distribution to capture inferences about any level in the model. Many other patterns of dependence are possible, and a hierarchical Bayesian model can be defined over any acyclic graph.

Inferences supported by hierarchical models

Hierarchical Bayesian models can be used for many purposes. Although Equation 3.1 supports many kinds of inferences, these inferences can be divided into three broad classes: top-down inferences, bottom-up inferences, and inferences at multiple levels of abstraction.

Top-down inferences

If the representations at some of the higher levels are fixed, a hierarchical model can make top-down predictions about the representations at the lower levels. In Figure 2-

2a, for instance, suppose that a phrase structure grammar is known (R_4) and we want to identify the structural description and sentence (R_3) that best account for a string of phonemes (R_2). The posterior distribution $P(R_3|R_4, R_2)$ can be used to model a top-down inference where observed data (R_2) are combined with prior knowledge (R_4) to make predictions about the representation at level 2.

Previous psychological applications of the hierarchical Bayesian approach have mostly focused on top-down inferences (Tenenbaum et al., 2006). Griffiths (2005) discusses the case of causal reasoning in detail, and argues that hierarchical Bayesian models can explain how people make top-down causal inferences given very sparse data (Figure 2-2e). Similar kinds of top-down inferences can be made about all of the cases in Figure 2-2.

Bottom-up inferences

If the representations at some of the lower levels are fixed, a hierarchical model can make bottom-up inferences about the representations at the higher levels. In Figure 2-2d, for instance, suppose that a collection of pixels is observed (R_1) and we want to identify the scene context (R_5) that best explains these observations. The posterior distribution $P(R_5|R_1)$ captures our beliefs about the representation at level 5 after observing data at the lowest level of the model.

Bottom-up inferences about the highest levels in a hierarchical model can help to explain the acquisition of inductive constraints, and the remaining chapters of this thesis will apply this idea to the examples in Figures 2-2c, 2-2e and 2-2f. In each case, we will see how the representation at the top level can be learned given data at the lowest level of the model.

Simultaneous inferences at multiple levels

Most of the examples so far have showed that inferences at a given level of abstraction can be guided by information at higher or lower levels of abstraction. Often, however, a learner will need to make simultaneous inferences about multiple levels of abstrac-

tion, and information will need to flow back and forth between several levels. For instance, when viewing a collection of pixels (R_1) a learner may need to extract surfaces (R_2) and object parts (R_3), recognize objects (R_4) and identify the scene context (R_5). The objects identified (R_4) may constrain the surfaces that are extracted (R_2) and vice versa, which means that bottom-up approaches and top-down approaches will not succeed. Instead, we need an interactive approach that allows representations at several levels to jointly constrain each other. A hierarchical Bayesian model meets this description, and the posterior distribution $P(R_2, R_3, R_4, R_5|R_1)$ can be used to model interacting inferences about many of the levels in Figure 2-2d.

Several psychologists have argued that human inferences are characterized by interactive processing over several levels of abstraction. The TRACE model of speech perception includes levels that correspond to acoustic features, phonemes, and words, and allows information to propagate from acoustic features up to words, and from words down to the acoustic features (McClelland & Elman, 1986; McClelland, Mirman, & Holt, 2006). Interactive processing is also discussed by vision researchers, who argue that the visual pathway includes feedback connections which allow inferences at higher levels to influence early visual areas (Lee & Mumford, 2003; Ullman, 1995). Interactive approaches like these have their detractors (Fodor, 1978; Norris, McQueen, & Cutler, 2000), but formal models of interactive processing can bring clarity to both sides of this debate.

Different patterns of learning can emerge when a model learns simultaneously about many levels of abstraction. Depending on the task and the data set, learning may be faster at the lower levels than the upper levels, equally rapid at all levels, or faster at the upper levels than the lower levels. The next chapter provides concrete examples of all three cases. Cases where learning is fastest at the upper levels of a model are especially interesting, and may help to explain how inductive constraints are acquired relatively early in development.

Choosing a hierarchy

Figure 2-2 shows hierarchies which capture several kinds of abstract knowledge, but we have seen no general recipe for constructing these hierarchies. This thesis will focus on problems where a hierarchy is specified in advance, and the main point of interest is whether inferences over this hierarchy can capture the kinds of inferences made by human learners. There are, however, informal principles that help modelers decide which hierarchies to explore, and that may ultimately help to explain how these hierarchies emerge over the course of cognitive development.

The previous chapter identified several reasons for working with explicit hierarchies like the examples in Figure 2-2. The same ideas provide criteria for choosing between competing hierarchies, including hierarchies with different numbers of levels. First, we should be sensitive to cases where learners transfer knowledge from one context to another. Cases like this provide evidence for a level of abstraction that captures the elements that are shared across the two contexts. Second, we should aim for the simplest possible model that will account for the data: in other words, levels should be added to a hierarchy whenever they increase the overall simplicity of a model. I suggested, for instance, that the simplicity of a model that characterizes the grammaticality of phoneme strings can be increased by adding levels corresponding to morphemes, words and phrases. Finally, we should look for direct evidence of the existence of certain levels. Successful learners, for instance, may make statements which indicate that they have made inferences at several levels of abstraction.

Deciding which of several hierarchies to prefer is a special case of the problem of choosing between scientific theories, and can be addressed in principle by standard theories of confirmation. The Bayesian approach to confirmation is one candidate, although philosophers continue to debate the strengths and weaknesses of this approach (Earman, 1992). The final chapter of this thesis discusses the steps needed to develop a Bayesian framework that identifies the best hierarchy for a given problem.

Other hierarchical approaches

As Figure 2-2 suggests, hierarchical approaches are prominent in the psychological literature (Greenwald, 1988), and have been used to explore language (Chomsky, 1957), memory (Bartlett, 1932), vision (Fukushima, 1980; Marr, 1982), action (Cooper & Shallice, 2000), categorization (Collins & Quillian, 1969), social behavior (Heider, 1958) and many other topics. Most of these hierarchical approaches are compatible with a hierarchical Bayesian approach and can be modeled within the framework I described. I illustrate by focusing on two well-known hierarchical approaches: Chomsky's view of language acquisition, and previous work on multilevel neural networks.

Hierarchical approaches in linguistics

Chomsky (1975) suggests that the study of linguistics amounts to the abstract study of "levels of representation," and argues for an approach that distinguishes at least six levels: phonemes, morphemes, words, syntactic categories, phrase structure, and transformations. According to Chomsky, a set of linguistic levels should specify the representations that can occur at each level, and the compatibility relationships that connect representations at different levels. A grammar is a system of rules that allows representations at each of these levels to be recovered given a phonetic spelling, or a string of units at the lowest level. Some aspects of this approach differ from the hierarchical Bayesian approach: for instance, I allow compatibility relationships between levels to be probabilistic, but Chomsky describes these relationships as deterministic rules. Despite some superficial differences, the hierarchical Bayesian framework is consistent with Chomsky's basic proposal that sentences have structured representations at multiple levels of abstraction, and that different kinds of compatibility relationships specify how the representations at different levels depend on each other.

A Bayesian approach to learning is mostly consistent with the view of learning presented in Chomsky's early work. According to Chomsky (1975), "linguistic theory characterizes a system of levels, a class of potential grammars, and an evaluation procedure with the following property: given data from language L and several grammars

with the properties required by linguistic theory, the procedure of evaluation selects the highest-valued of these." Given this evaluation procedure, a language learner can "select the highest-valued grammar of the appropriate form compatible with available data." A hierarchical Bayesian account of learning matches this basic pattern. Given a hierarchy of levels where a grammar R_n appears near the top and phonetic data R_1 appear at the bottom, grammar learning can be captured by an "evaluation procedure" that identifies the grammar that maximizes $P(R_n|R_1)$.[1] I have not described an algorithm which implements this evaluation procedure, but the hierarchical Bayesian approach can be evaluated without committing to a specific mechanism for searching the space of grammars.

Although Chomsky's view of learning appears closely related to my own, he often describes this view using language that is inconsistent with the terminology I have chosen. For instance, he argues that a child's knowledge of language "goes far beyond the presented primary linguistic data and is in no sense an 'inductive generalization' from these data" (Chomsky, 1965). I prefer to say that a child's knowledge of language goes far beyond the primary linguistic data and is *therefore* an inductive generalization from these data—in other words, if linguistic data contained enough information to fully specify a grammar, then grammar learning would be a deductive rather than an inductive problem. Disagreements like these are of little consequence and indicate only that certain phrases (e.g. "inductive generalization") can be used in different ways. Chomsky's preferred usage helps to emphasize that language learning must go well beyond enumerative induction or any of the bottom-up learning methods that are traditionally linked with empiricist approaches. My preferred usage focuses on the distinction between deduction and induction, and acknowledges that there is much more to learning than the simple bottom-up methods dismissed by Chomsky (Table 1.2).

As this brief excursion into linguistic theory suggests, the hierarchical Bayesian

[1]Chomsky describes a grammar as a body of knowledge that determines a system of levels. To accurately capture this idea we need to allow the grammar R_n to directly influence the representation at each of the lower levels, but a model like this is entirely consistent with the hierarchical Bayesian approach (see Figure 3-1b).

framework incorporates several ideas that have been part of cognitive science from the very beginning. Hierarchical approaches have been explored for many years, and it has long been clear that learning can be understood computationally as the problem of searching for a representation that maximizes some measure of goodness. The advantage of the hierarchical *Bayesian* approach is that it can handle soft probabilistic relationships between representations at different levels, that it provides a principled method for dealing with uncertainty, and that it helps to explain which "measure of goodness" is relevant to a given learning problem. When learning representation R from data D, the measure of goodness should always be $P(R|D)$, or the posterior probability of R given the data.

Multilevel neural networks

As presented here, the hierarchical Bayesian approach relies on three central ideas. First, multiple levels of abstraction are needed to capture human knowledge. Second, the representations at each level and the relationships between levels can be rich and complex. Third, probabilistic inference helps to explain how the representations at all levels are acquired and used. We have just seen that linguists have long argued for the first two claims, but have tended to resist the third. Research on multilevel neural networks has emphasized the first and third claims, but not the second.

Inspired in part by the structure of visual cortex, vision scientists have suggested that pattern recognition can be achieved by a multilevel network where the representations at the higher levels become increasingly invariant to changes in position and other transformations (Fukushima, 1980). Several groups of researchers have described probabilistic multilevel networks that are motivated by similar ideas (Lee & Mumford, 2003; Hinton, Osindero, & Teh, 2006; D. George & Hawkins, 2005). There are significant differences between the networks proposed by different researchers, but all of them share two properties: the representations at each level are formalized as feature vectors, and the relationships between feature vectors at adjacent levels tend to be the same across the entire hierarchy. Models of this sort are compatible with the hierarchical Bayesian approach described in this chapter, and may help to

explain some aspects of cognition, including visual perception. Importantly, however, the hierarchical Bayesian approach can handle grammars, logical theories, and other representations that are richer than feature vectors. The hierarchical Bayesian approach also handles cases where adjacent levels are related to each other in very different ways: for instance, a hierarchical language model should acknowledge that the relationship between a string of phonemes and a sentence is qualitatively different from the relationship between a structural description and a phrase structure grammar.

Each individual component of the framework I described has been extensively explored in the psychological literature. There are models that rely on hierarchies, models that incorporate richly structured representations, and models that explain learning in terms of Bayesian inference. This section reviewed some well known approaches that combine two of these ideas. Few models, however, combine all three ideas, and this thesis proposes that all three are needed to account for human cognition.

Belief formation or belief fixation?

By now it should be clear that hierarchical Bayesian models can make inductive inferences at multiple levels of abstraction, but some readers will wonder whether these models can really discover abstract knowledge. Terms like discovery and belief formation are sometimes reserved for cases where a system comes up with a hypothesis (e.g. a concept or a theory) that is qualitatively new (Reichenbach, 1938; Fodor, 1980). Terms like justification or belief fixation are used for cases where a system chooses between two or more pre-existing hypotheses. For instance, a system that starts out with few preconceptions about language may form a new belief when it realizes that English sentences are hierarchically structured. A system that starts out with two possible grammars and identifies the candidate that best accounts for a corpus has only adjusted the weights of two pre-existing beliefs.

At first sight, Bayesian models may seem like accounts of belief fixation rather

than belief formation. Any Bayesian model begins with a hypothesis space, and "learning" is a matter of identifying the element in this space that best accounts for the data. Since a Bayesian learner can never step outside its hypothesis space, in one sense it can only perform belief fixation, since it must begin with all the hypotheses that it will ever need (Suppes, 1966).

Bayesian models, however, are best viewed as operating at a level of explanation where the distinction between belief formation and belief fixation breaks down. From a computational perspective (Marr, 1982), every learning system relies on a fixed hypothesis space which represents the abstract potential of the system. If we imagine all streams of input that the system could possibly receive, the hypothesis space includes all states of knowledge which the system could possibly reach. Even systems that appear to recruit new representational resources must implicitly rely on a fixed hypothesis space. For instance, constructivist neural networks grow by adding new units (Fahlman & Lebiere, 1990), but the fixed hypothesis space in this case includes all configurations that can be reached by adding new units.

Since every learning system relies on a fixed hypothesis space, every system is computationally equivalent to a method of belief fixation. The distinction between belief formation and belief fixation must therefore distinguish different ways in which a computational theory can be implemented. For instance, an implementation that entertains only a few hypotheses at a time may be said to form a new belief every time it generates a hypothesis that has never previously been entertained. An implementation that has access to the entire hypothesis space (for instance, that explicitly considers all possible hypotheses whenever it needs to make a prediction) might be better described as a model of belief fixation.

The primary goal of this thesis is to develop computational theories that explain how constraints can be learned. Each of these theories can be implemented in many ways: some implementations will seem like models of belief formation, and others will seem like models of belief fixation. Once we commit to a specific implementation, we can decide whether or not it succeeds as an account of belief formation. Here, however, I focus almost entirely on the level of computational theory.

Summary

This chapter introduced the hierarchical Bayesian approach and showed how probabilistic models can be defined over hierarchies with multiple levels of abstraction. These hierarchies can incorporate richly structured representations, and the representations at different levels can be related to each other in many different ways. Statistical inference over these hierarchies can be used to learn about the representations at any level of abstraction, and I showed how these models support top-down inferences, bottom-up inferences, and simultaneous inferences about multiple levels of abstraction.

The next three chapters of this thesis describe hierarchical Bayesian models that address three aspects of high-level cognition: categorization (Figure 2-2c), causal reasoning (Figure 2-2e) and knowledge representation (Figure 2-2f). The representations at the upper levels of each model can be viewed as inductive constraints, and we will see how these constraints can be acquired given data at the bottom levels of these models.

Chapter 4

Learning about feature variability

Imagine that a child is visiting the zoo with her parents, and that her mother points at something and utters the word "wombat." The child might be excused for thinking that the word refers to any object that is furry and brown, to the corner of the nearest enclosure, or to the snout of the animal that is currently hiding in the corner of the enclosure (Quine, 1960). There are an indefinite number of more exotic hypotheses—for instance, the word could refer to burrowing marsupials when used on Tuesday and to teapots when used on any other day of the week (Goodman, 1955). Although the space of logically possible hypotheses is vast, a single labeled example is often enough for young children to make accurate inferences about the meaning of a novel word. Inferences like these must be supported by strong inductive constraints, and models of constraint learning can help to explain how children become such proficient word learners.

The problem of word learning is a natural target for a hierarchical approach since it appears to involve inferences at two or more levels of abstraction. Children need to learn about individual categories—for example, they need to discover that balls tend to be round, and that teacups tend to have a handle. Children also need to acquire more abstract knowledge about categories in general. One instance of more

The work in this chapter was carried out in collaboration with Amy Perfors and Joshua Tenenbaum. The chapter is a revised version of Kemp, Perfors, and Tenenbaum (2007) and is reproduced with permission.

abstract knowledge is the shape bias, or the expectation that the members of any given category will tend to have the same shape, even if they vary along other dimensions such as color or size. The shape bias supports inferences from very sparse data: given a single labeled example of a novel category, young children will extend the category label to similarly-shaped objects ahead of objects that share the same texture or color as the exemplar (Heibeck & Markman, 1987; Landau et al., 1988). This chapter describes a hierarchical Bayesian model that acquires inductive constraints like the shape bias.

Learning the shape bias is one instance of the more general problem of learning about feature variability. The general problem can be introduced using an example given by Goodman (1955). Suppose that S is a stack containing many bags of marbles. We empty several bags and discover that some bags contain black marbles, others contain white marbles, but that the marbles in each bag are uniform in color. We now choose a new bag—bag n—and draw a single black marble from the bag. On its own, a single draw would provide little information about the contents of the new bag, but experience with previous bags may lead us to endorse the following hypothesis:

H: All marbles in bag n are black.

If asked to justify the hypothesis, we might invoke the following constraint:

C: Each bag in stack S contains marbles that are uniform in color.

Goodman refers to C as an *overhypothesis*, but C can also be described as an epistemic constraint. C is a constraint since it limits the possible hypotheses about the marbles in each bag: for instance, the marbles in bag n could be uniformly black or uniformly white, but may not be mixed in color. Once this constraint has been acquired, a learner can make confident predictions about bag n after seeing exactly one marble sampled from this bag.

Although Goodman did not give a formal account of how overhypotheses might be acquired, a simple hierarchical model helps to explain how constraints like C might be learned. Consider a model with three levels (Figure 4-1a). Level 1 records observations that have been made by drawing marbles from one or more bags. Level

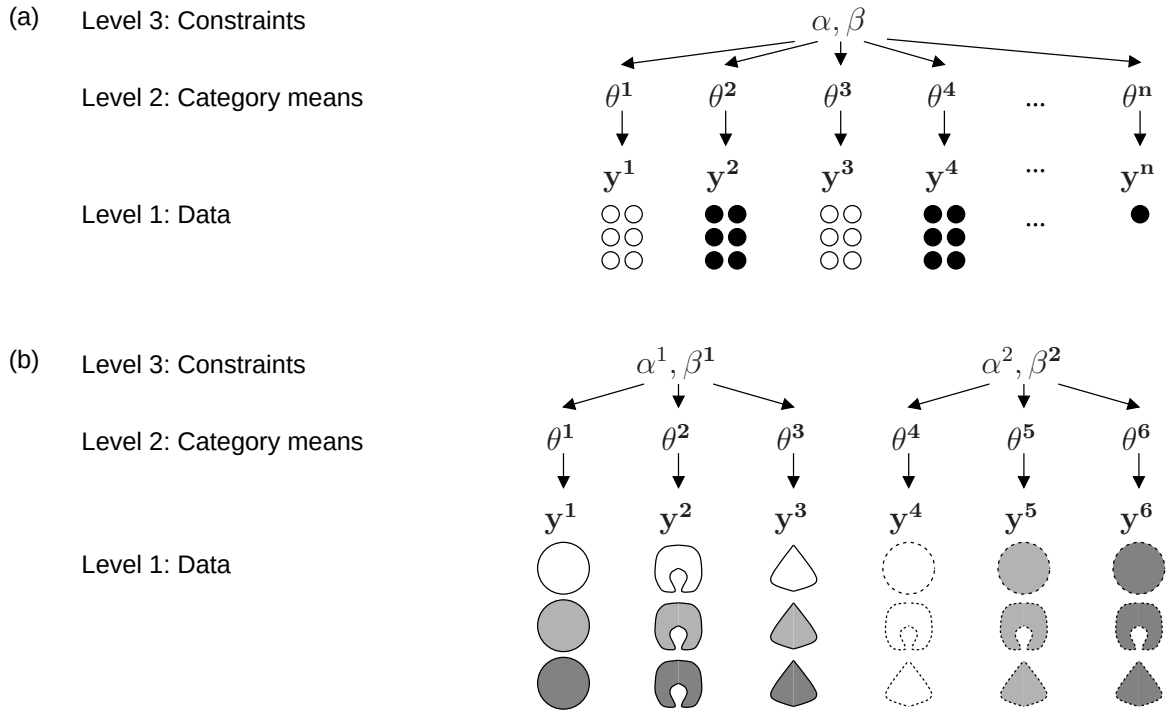

Figure 4-1: (a) A categorization model that formalizes the intuition behind Figure 2-2c. Each category is shown as a bag of colored marbles. Individual marbles represent category exemplars, and $\boldsymbol{\theta^i}$ is the color distribution for category i. α and $\boldsymbol{\beta}$ place constraints on the $\{\boldsymbol{\theta^i}\}$ variables: $\boldsymbol{\beta}$ is the color distribution across all categories, and α represents the variability in color within each category. (b) A categorization model with two ontological kinds meant to correspond loosely to objects and substances. α^1 represents knowledge about feature variability within the first ontological kind (object categories are homogeneous in shape but not in material), and $\boldsymbol{\beta^1}$ captures the characteristic features of the entities belonging to the first kind (objects tend to be solid).

2 specifies information about the color distribution of each bag, and Level 3 specifies information about bags in general. For instance, Level 3 may indicate that the contents of each bag tend to be homogeneous in color.

A Dirichlet-multinomial model provides one way to formalize the hierarchical approach in Figure 4-1a (A. Gelman et al., 2003). Suppose we are working with a set of k colors. Initially we set $k = 2$ and use black and white as the colors. Let y^i indicate our observations of the marbles that have been drawn from the ith bag in the stack. If we have drawn 5 marbles from bag 7 and all but one are black, then $\boldsymbol{y^7} = [4, 1]$. Let $\boldsymbol{\theta^i}$ indicate the true color distribution for bag i: if 60% of the marbles in bag 7 are black, then $\boldsymbol{\theta^7} = [0.6, 0.4]$. Formally, we assume that $\boldsymbol{y^i}$ is drawn from a

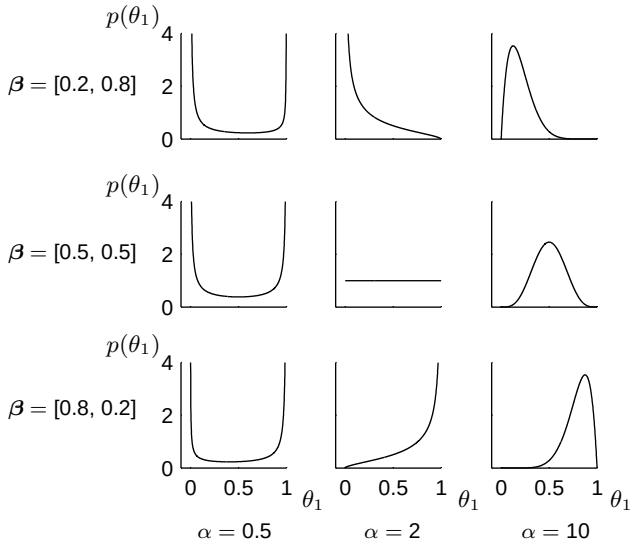

Figure 4-2: The Dirichlet distribution serves as a prior on $\boldsymbol{\theta}$, the color distribution of a bag of marbles. Assume that there are two possible colors—white and black—and let θ_1 be the proportion of black marbles within the bag. Shown here are distributions on θ_1 when the parameters of the Dirichlet distribution (α and $\boldsymbol{\beta}$) are systematically varied. When α is small, the marbles in each individual bag are near-uniform in color (θ_1 is close to 0 or close to 1), and $\boldsymbol{\beta}$ determines the relative proportions of bags that are mostly black and bags that are mostly white. When α is large, the color distribution for any individual bag is expected to be close to the color distribution across the entire population of bags (θ_1 is close to β_1).

multinomial distribution with parameter $\boldsymbol{\theta^i}$: in other words, the marbles responsible for the observations in $\boldsymbol{y^i}$ are drawn independently at random from the ith bag, and the color of each depends on the color distribution $\boldsymbol{\theta^i}$ for that bag.

The representation at level 3 captures knowledge about the distribution of the $\boldsymbol{\theta^i}$ variables. We will assume that this knowledge can be captured using using two parameters, α and $\boldsymbol{\beta}$ (Figure 4-1a). Roughly speaking, α captures the extent to which the marbles in each individual bag are uniform in color, and $\boldsymbol{\beta}$ captures the color distribution across the entire stack of bags. Formally, we assume that the vectors $\boldsymbol{\theta^i}$ are independently drawn from a Dirichlet distribution with scale parameter α and mean $\boldsymbol{\beta}$. Figure 4-2 shows how the distribution on $\boldsymbol{\theta}$ changes as α and $\boldsymbol{\beta}$ are systematically varied. When α is small, the marbles in each individual bag are near-uniform in color, and $\boldsymbol{\beta}$ determines the relative proportions of bags that are mostly white and bags that are mostly black. When α is large, the color distribution for any

individual bag is expected to be close to $\boldsymbol{\beta}$, the color distribution across the entire population of bags.

To complete the model in Figure 4-1a, we need to formalize our *a priori* expectations about the values of α and $\boldsymbol{\beta}$. We use a uniform distribution on $\boldsymbol{\beta}$ and an exponential distribution on α, which captures a weak prior expectation that the marbles in any bag will tend to be uniform in color. The mean of the exponential distribution is λ, and the value of this variable is specified by one of the background assumptions. All simulations described in this chapter use $\lambda = 1$. Using statistical notation, the entire model can be written as

$$\begin{aligned}
\alpha &\sim \text{Exponential}(\lambda) \\
\boldsymbol{\beta} &\sim \text{Dirichlet}(\mathbf{1}) \\
\boldsymbol{\theta^i} &\sim \text{Dirichlet}(\alpha\boldsymbol{\beta}) \\
\boldsymbol{y^i} \,|\, n^i &\sim \text{Multinomial}(\boldsymbol{\theta^i})
\end{aligned}$$

where n^i is the number of observations for bag i.

So far, we have assumed that we are working with a single dimension—for Goodman, marble color. Suppose, however, that some marbles are made from metal and others are made from glass, and we are interested in material as well as color. A simple way to deal with multiple dimensions is to assume that each dimension is independently generated, and to introduce separate values of α and $\boldsymbol{\beta}$ for each dimension. When working with multiple features, we will often use α to refer to the collection of α values along all dimensions, $\boldsymbol{\beta}$ for the set of all $\boldsymbol{\beta}$ vectors, and $\boldsymbol{y}$ for the set of counts along all dimensions.

To fit the model to data we assume that counts $\boldsymbol{y}$ are observed for one or more bags. Our goal is to compute the posterior distribution $p(\alpha, \boldsymbol{\beta}, \{\boldsymbol{\theta^i}\}|\boldsymbol{y})$: in other words, we wish to simultaneously discover level 3 knowledge about α and $\boldsymbol{\beta}$ and level 2 knowledge about the color distribution $\boldsymbol{\theta^i}$ of each individual bag i. Figures 4-3b and 4-3c show posterior distributions on $\log(\alpha)$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta^i}$ for two sets of counts. We can approximate the distribution $p(\alpha, \boldsymbol{\beta}|\boldsymbol{y})$ using numerical integration or a Markov chain

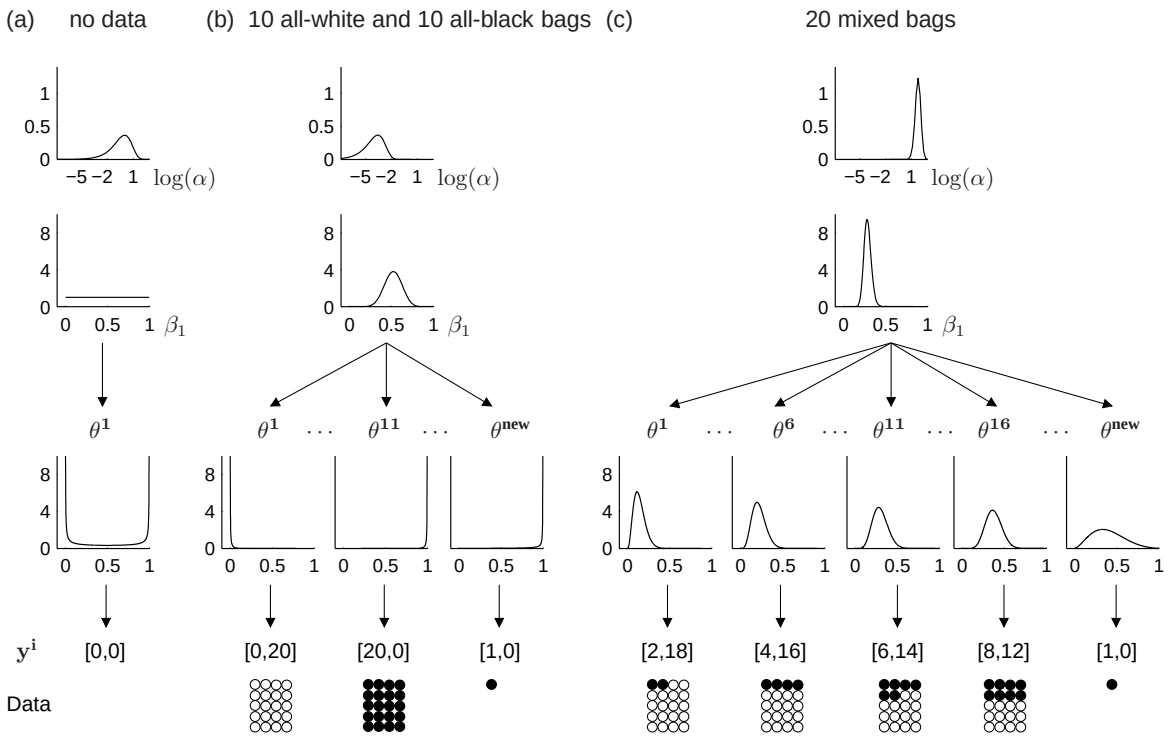

Figure 4-3: Generalizations made by the model in Figure 4-1a. (a) Prior distributions on $\log(\alpha)$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta^i}$ indicate the model's expectations before any data have been observed. (b) Posterior distributions after observing 10 all-white bags and 10 all-black bags. The model realizes that most bags are near-uniform in color (α is small), and that about half of these bags are black (β_1 is around 0.5). These posterior distributions allow the model to predict that the proportion of black marbles in the new, sparsely observed bag (θ_1^{new}) is very close to 1. (c) Posterior distributions after observing 20 mixed bags inspired by the obesity condition of the Barratos task. The model realizes that around 25% of marbles are black (β_1 is around 0.25), and that roughly 25% of the marbles in each individual bag are black (α is high). These posterior distributions allow the model to predict that the new, sparsely observed bag is likely to contain more white marbles than black marbles (θ_1^{new} is not close to 1).

Monte Carlo (MCMC) scheme. Inferences about the $\boldsymbol{\theta^i}$ are computed by integrating out α and $\boldsymbol{\beta}$:

$$p(\boldsymbol{\theta^i}|\boldsymbol{y}) = \int_{\alpha,\boldsymbol{\beta}} p(\boldsymbol{\theta^i}|\alpha,\boldsymbol{\beta},\boldsymbol{y})p(\alpha,\boldsymbol{\beta}|y)d\alpha d\boldsymbol{\beta}.$$

To compute some of the model predictions in this chapter we implemented a sampler that uses Gaussian proposals on $\log(\alpha)$, and proposals for $\boldsymbol{\beta}$ that are drawn from a Dirichlet distribution with the current $\boldsymbol{\beta}$ as its mean. The results in Figure 4-6 represent averages across 30 Markov chains, each of which was run for 50,000 iterations (1000 were discarded as burn-in). The model predictions in Figures 4-3, 4-5, 4-4 and 4-7 were computed using numerical integration. Note that both inference schemes (MCMC and numerical integration) are merely convenient ways of computing the predictions of our computational theory. Any computational theory can be implemented in many ways, and the particular implementations we have chosen are not intended as models of cognitive processing.

Modeling inductive reasoning

Since Goodman, psychologists have confirmed that children (Macario, Shipley, & Billman, 1990) and adults (Nisbett, Krantz, Jepson, & Kunda, 1983) have knowledge about feature variability and use this knowledge to make inductive leaps given very sparse data. This section provides an initial demonstration of our model using data inspired by one of the tasks of Nisbett et al. (1983). As part of this task, participants were asked to imagine that they were exploring an island in the Southeastern Pacific, that they had encountered a single member of the Barratos tribe, and that this tribesman was brown and obese. Based on this single example, participants concluded that most Barratos were brown, but gave a much lower estimate of the proportion of obese Barratos (Figure 4-4). When asked to justify their responses, participants often said that tribespeople were homogeneous with respect to color but heterogeneous with respect to body weight (Nisbett et al., 1983).

To apply the Dirichlet-multinomial model to this task, we replace bags of marbles with tribes. Suppose we have observed 20 members from each of 20 tribes. Half the

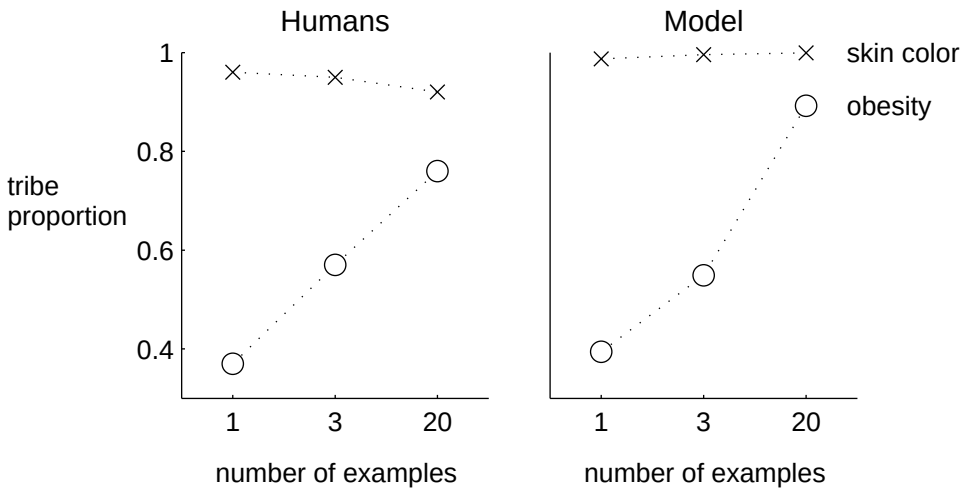

Figure 4-4: Generalizations about a new tribe after observing 1, 3, or 20 obese, brown-skinned individuals from that tribe. Human generalizations are replotted from Nisbett et al. (1983). For each set of observations, the Dirichlet-multinomial model learns a distribution over the feature proportions $\boldsymbol{\theta}^{\text{new}}$ for a new tribe (Figure 4-3). Plotted here are the means of those distributions. A single observation allows the model to predict that most individuals in the new tribe have brown skin, but many more observations are needed before the model concludes that most tribe members are obese.

tribes are brown and the other half are white, but all of the individuals in a given tribe have the same skin color. Given these data, the posterior distribution on α indicates that skin color tends to be homogenous within tribes (i.e. α is probably small) (Figure 4-3b). Learning that α is small allows the model to make strong predictions about a sparsely observed new tribe: having observed a single, brown-skinned member of a new tribe, the posterior distribution on $\boldsymbol{\theta}^{\text{new}}$ indicates that most members of the tribe are likely to be brown (Figures 4-3b and 4-4). Note that the posterior distribution on $\boldsymbol{\theta}^{\text{new}}$ is almost as sharply peaked as the posterior distribution on $\boldsymbol{\theta}^{\mathbf{11}}$: the model has realized that observing one member of a new tribe is almost as informative as observing 20 members of that tribe.

Suppose now that obesity is a feature that varies within tribes: a quarter of the 20 tribes observed have an obesity rate of 10%, and the remaining quarters have rates of 20%, 30%, and 40%. Obesity is represented as a second binary feature, and the posterior distributions on α and $\boldsymbol{\beta}$ (Figure 4-3c) indicate that obesity varies within tribes (α is high), and that the base rate of obesity is around 25% (β_1 is around

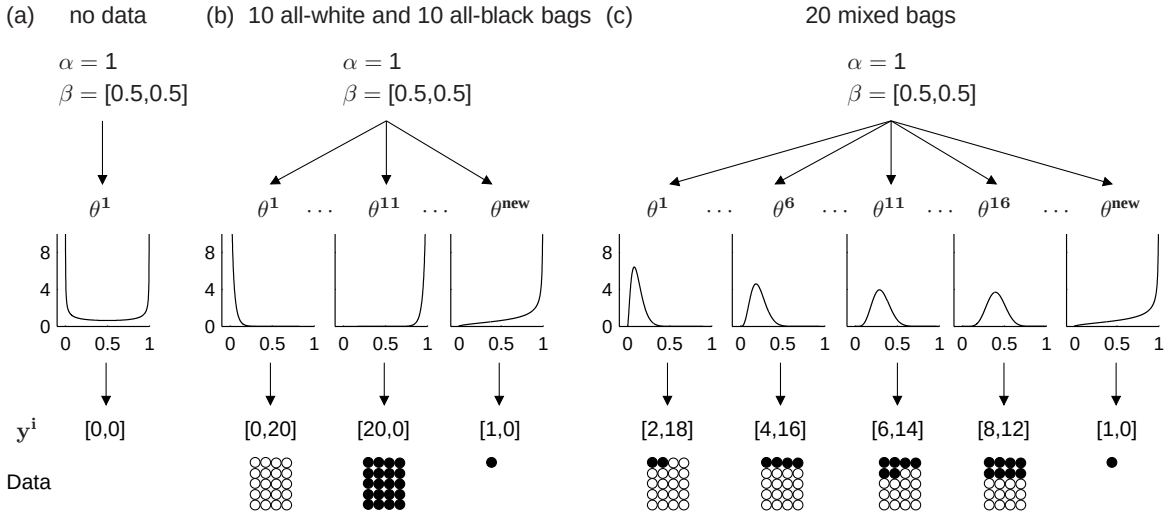

Figure 4-5: Generalizations of a conventional Bayesian model that learns only at the level of $\boldsymbol{\theta}$ (α and $\boldsymbol{\beta}$ are fixed). The model does not generalize correctly to new, sparsely observed bags: since α and $\boldsymbol{\beta}$ are fixed, observing 20 previous bags provides no information about a new bag, and the posterior distributions on θ_1^{new} are identical in cases (b) and (c).

0.25). Again, we can use these posterior distributions to make predictions about a new tribe, and now the model requires many observations before it concludes that most members of the new tribe are obese (Figure 4-4). Unlike the case in Figure 4-3b, the model has learned that a single observation of a new tribe is not very informative, and the distribution on $\boldsymbol{\theta}^{\text{new}}$ is now similar to the average of the $\boldsymbol{\theta}$ distributions for all previously observed tribes.

Accurate predictions about a new tribe depend critically on learning at both level 2 and level 3 (Figure 4-1a). Learning at level 2 is needed to incorporate the observation that the new tribe has at least one obese, brown-skinned member. Learning at level 3 is needed to discover that skin color is homogeneous within tribes but that obesity is not, and to discover the average rate of obesity across many tribes. Figure 4-5 shows inferences drawn by an alternative model that is unable to learn at level 3—instead, we fix α and $\boldsymbol{\beta}$ to their expected values under the prior distributions used by our model. Since it cannot adjust the α and $\boldsymbol{\beta}$ variables, this alternative model cannot incorporate any information about the 20 previous tribes when reasoning about a new tribe. As a result, it makes identical inferences about skin color and obesity—note

that the distribution on $\boldsymbol{\theta}^{\text{new}}$ is the same in Figures 4-5b and 4-5c. Note also that the mean of this distribution (0.75) is lower than the mean of the distribution in Figure 4-3b (0.99)—both models predict that most members of the new tribe have brown skin, but our model alone accounts for the human judgment that almost all members of the new tribe have brown skin (Figure 4-4).

Learning the shape bias

The Barratos task does not address an important class of inferences made by human learners: inferences about *new* feature values along known dimensions. Based on the data in Figure 4-1a, a learner could acquire at least two different constraints: the first states that the marbles in each bag are uniform along the dimension of color, and the second states that the marbles in each bag are either all white or all black. One way to distinguish between these possibilities is to draw a single green marble from the new bag. A learner with the first constraint will predict that all marbles in the new bag are green, but a learner with the second constraint will be lost.

There are many real-world problems that involve inferences about novel features. Children know, for example, that animals of the same species tend to make the same sound. Observing one horse neigh is enough to conclude that most horses neigh, even though a child may never have heard an animal neigh before (Shipley, 1993). Similarly, by the age of 24 months children show a shape bias: they know that shape tends to be homogeneous within object categories. Given a single exemplar of a novel object category, children extend the category label to similarly-shaped objects ahead of objects that share the same texture or color as the exemplar (Heibeck & Markman, 1987; Landau et al., 1988).

The model in Figure 4-1a deals naturally with inferences like these, and I illustrate using stimuli inspired by the work of Smith et al. (2002). In their first experiment, these authors trained 17-month olds on two exemplars from each of four novel categories. Novel names (e.g. "zup") were provided for each category, and the experimenter used phrases like "this is a zup—let's put the zups in the wagon."

(a) Training

Category	1 1	2 2	3 3	4 4
Shape	1 1	2 2	3 3	4 4
Texture	1 2	3 4	5 6	7 8
Color	1 2	3 4	5 6	7 8
Size	1 2	1 2	1 2	1 2

(b) First-order generalization

T_1

1	?	?	?
1	1	6	6
1	9	1	9
1	9	9	1
1	1	1	1

(c) Second-order generalization

T_2

5	?	?	?
5	5	6	6
9	10	9	10
9	10	10	9
1	1	1	1

(d)

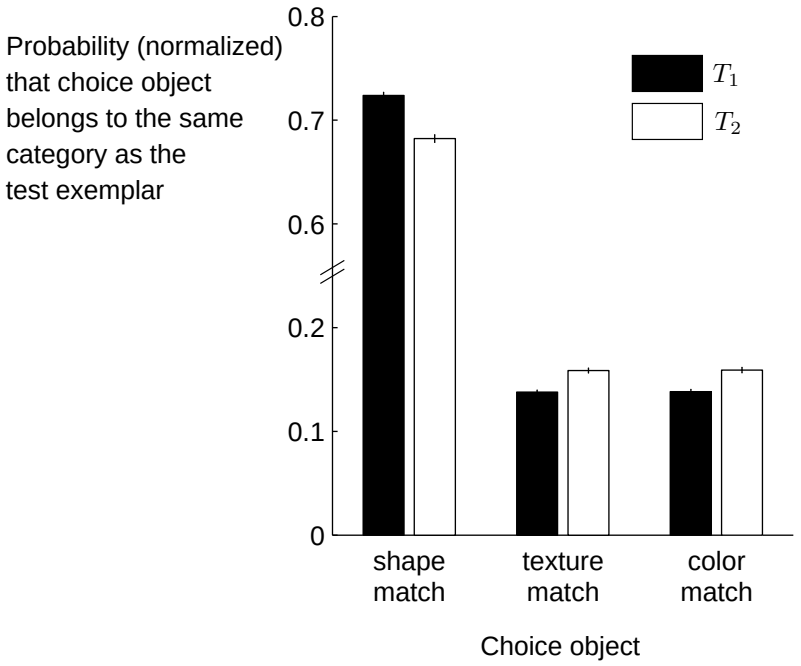

Figure 4-6: Learning the shape bias. (a) Training data based on Smith et al. (2002). Each column represents an object: for instance, the first two columns represent two "zups." There are 10 possible shapes, textures and colors, and 2 possible sizes. (b) First-order generalization was tested by presenting the model with exemplar T_1, and asking it to choose which of three objects (a shape match, a texture match and a color match) was most likely to belong to the same category as T_1. (c) Second-order generalization was tested using T_2, an exemplar of a category that was not seen during training. (d) Model predictions for both generalization tasks. Each bar represents the probability that a choice object belongs to the same category as the test exemplar (probabilities have been normalized so that they sum to one across each set of choice objects). The model makes exact predictions about these probabilities: we computed 30 estimates of these predictions, and the error bars represent the standard error of the mean.

Within each category, the two exemplars had the same shape but differed in size, texture and color (Figure 4-6a). After eight weeks of training, the authors tested *first-order* generalization by presenting T_1, an exemplar from one of the training categories, and asking children to choose another object from the same category as T_1. Three choice objects were provided, each of which matched T_1 in exactly one feature (shape, texture or color) (Figure 4-6b). Children preferred the shape match, showing that they were sensitive to feature distributions within a known category. Smith et al. (2002) also tested *second-order* generalization by presenting children with T_2, an exemplar from a novel category (Figure 4-6c). Again, children preferred the shape match, revealing knowledge that shape in general is a reliable indicator of category membership. Note that this result depends critically on the training summarized by Figure 4-6a: 19-month olds do not normally reveal a shape bias on tests of second-order generalization.

We supplied the model with counts $\boldsymbol{y^i}$ computed from the feature vectors in Figure 4-6a. For example, $\boldsymbol{y^1}$ indicates that the data for category 1 include two observations of shape value 1, one observation of texture value 1, one observation of texture value 2, and so on. The key modeling step is to allow for more values along each dimension than appear in the training set. This policy allows the model to handle shapes, colors and textures it has never seen during training, but assumes that the model is able to recognize a novel shape as a kind of shape, a novel color as a kind of color, and so on. We allowed for ten shapes, ten colors, ten textures and two sizes: for example, the shape component of $\boldsymbol{y^1}$ indicates that the observed exemplars of category 1 include two objects with shape value 1 and no objects with shape values 2 through 10.

Figure 4-6d shows the patterns of generalization predicted by the model. Smith et al. (2002) report that the shape match was chosen 88% (66%) of the time in the test of first-order generalization, and 70% (65%) of the time in the second-order task (percentages in parentheses represent results when the task was replicated as part of Experiment 2). The model reproduces this general pattern: shape matches are preferred in both cases, and are preferred slightly more strongly in the test of

first-order generalization.

Smith et al. (2002) also measured real-world generalization by tracking vocabulary growth over an eight week period. They report that experience with the eight exemplars in Figure 4-6a led to a significant increase in the number of object names used by children. The Dirichlet-multinomial model helps to explain this striking result. Even though the training set includes only four categories, the results in Figure 4-6b show that it contains enough statistical information to establish or reinforce the shape bias, which can then support word learning in the real world. Similarly, the model explains why providing only two exemplars per category is sufficient. In fact, if the total number of exemplars is fixed, the model predicts that the best way to teach the shape bias is to provide just two exemplars per category. I illustrate by returning to the marbles scenario.

Each point in Figure 4-7a represents a simulation where 64 observations of marbles are evenly distributed over some number of bags. The marbles drawn from any given bag are uniform in color—black for half of the bags and white for the others. When 32 observations are provided for each of two bags (Figure 4-7b.i), the model is relatively certain about the color distributions of those bags, but cannot draw strong conclusions about the homogeneity of bags in general. When two observations are provided for each of 32 bags, (Figure 4-7b.ii), the evidence about the composition of any single bag is weaker, but taken together, these observations provide strong support for the idea that α is low and most bags are homogeneous. When just one observation is provided for each of 64 bags, the model has no information about color variability within bags, and the posterior distribution on α is identical to the prior on this variable, which has a mean value of 1. If the total number of observations is fixed, Figure 4-7a suggests that the best way to teach a learner that bags are homogeneous in general is to provide two observations for as many bags as possible. The U-shaped curve in Figure 4-7a is a novel prediction of the model, and could be tested in developmental experiments.

Although the Dirichlet-multinomial model provides some insight into the findings of Smith et al. (2002), it does not account for all of their results. Their second

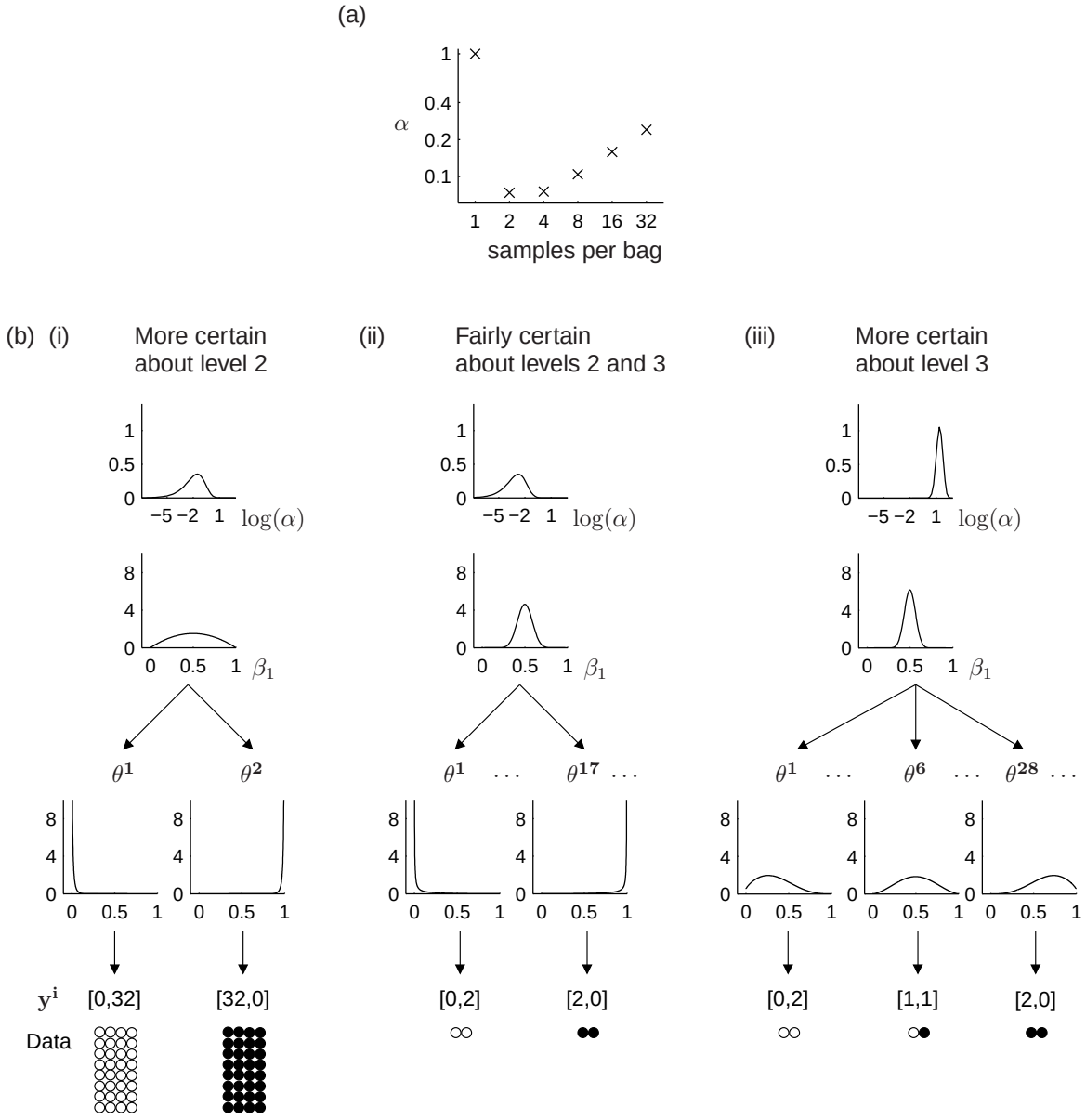

Figure 4-7: (a) Mean α values after observing 32 white marbles and 32 black marbles divided evenly across some number of homogenous bags. The model is most confident that bags in general are homogeneous (i.e. α is low) when given 2 samples from each of 32 bags. (b) Three possible outcomes when learning occurs simultaneously at level 2 and level 3. (i) After observing 2 homogeneous bags, the model is more certain about the variables at level 2 than the variables at level 3. (ii) After observing pairs of marbles from 32 homogeneous bags, the model is fairly certain about levels 2 and 3. (iii) After observing pairs of marbles from 32 bags (5 white pairs, 22 mixed pairs, and 5 black pairs), the model is more certain about level 3 than level 2.

experiment includes a *no-name* condition where children received the same training as before (Figure 4-6a) except that category labels were not provided. Instead of naming the training objects, the experimenter used phrases like "here is one, here is another—let's put them both in the wagon." Children in this condition showed first-order but not second-order generalization, which supports the view that the shape bias reflects attention to shape in the context of naming (Smith, Jones, & Landau, 1996). An alternative view is that the shape bias is not specifically linguistic: shape is important not because it is linked to naming in particular, but because it is a reliable cue to category membership (Ward, Becker, Hass, & Vela, 1991; Bloom, 2000). The Dirichlet-multinomial model is consistent with this second view, and predicts that learning in the no-name condition should not have been impaired provided that children clearly understood which training objects belonged to the same category. This discrepancy between model predictions and empirical results calls for further work on both sides. On the modeling side, it is important to develop hierarchical models that allow an explicit and privileged role for linguistic information. On the empirical side, it seems possible that children in the no-name condition did not achieve second-order generalization because they did not realize that each pair of identically-shaped objects was supposed to represent a coherent category.[1] Observing associations between similarly-shaped objects may have led them only to conclude that shape was a salient feature of each of these objects, which would have been enough for them to pass the test of first-order generalization.

Learning constraints fast

As mentioned in Chapter 1, there are empirical and theoretical reasons to believe that many inductive constraints are available relatively early in development. Any attempt to argue that these constraints might be learned must therefore explain how they can

[1]For those who support an essentialist view of categories (D. Medin & Ortony, 1989; Bloom, 2000), the issue at stake is whether the identically-shaped objects were believed to have the same essence. A shared name is one indication that two objects have the same essence, but other indications are possible—for example, children might be told "Here's one and here's another. Look, they are both the same kind of thing. I wonder what they're called."

be learned very rapidly. Our analysis of the task in Smith et al. (2002) suggests one potential explanation: constraints may be available early because they can be extracted from very small amounts of training data. We saw, for instance, that two exemplars from each of four categories are enough to allow the Dirichlet-multinomial model to discover a version of the shape bias. The hierarchical Bayesian approach suggests a second explanation that may apply in some cases. When learning occurs simultaneously at multiple levels of abstraction, "abstract-to-concrete" trajectories can emerge: in other words, the representations at the upper levels can be acquired before the representations at the lower levels are firmly in place. Abstract-to-concrete learning may help to explain how children acquire inductive constraints early enough to guide subsequent learning at lower levels.

The Dirichlet-multinomial model (Figure 4-1) can be used to demonstrate the basic notion of abstract-to-concrete learning. At least three outcomes are possible when learning proceeds in parallel at levels 2 and 3. Figure 4-7b.i shows a case where the learner is more confident about concrete knowledge (level 2) than abstract knowledge (level 3): note that the distributions for the two individual bags ($\boldsymbol{\theta^1}$ and $\boldsymbol{\theta^2}$) are more tightly peaked than the distributions on α and $\boldsymbol{\beta}$, which capture knowledge about bags in general. Figure 4-7b.ii is a case where the learner is relatively confident about the values of the variables at both levels. Figure 4-7b.iii is a case where the learner is more confident about abstract knowledge (level 3) than concrete knowledge (level 2). In this case, two observations are provided for each of 32 bags: 22 of the observed pairs are mixed, and there are 5 white pairs and 5 black pairs. The model is now relatively uncertain about the color distribution of any individual bag, but relatively certain about the values of α and $\boldsymbol{\beta}$.

The diagrams in Figure 4-7b show static snapshots of a learner's state of knowledge. Figure 4-8 shows developmental trajectories where the second state in Figure 4-8a corresponds to Figure 4-7b.i, and the second state in Figure 4-8b corresponds to Figure 4-7b.iii. In Figure 4-8a, the inductive constraint (level 3) is acquired after some of the category means (level 2) are learned with high confidence. This trajectory matches the common intuition that constraints are acquired by abstracting

over more concrete forms of knowledge. For instance, Smith et al. (2002) describe a four-step account of word-learning where learners acquire the shape bias by realizing that many of the categories they have already learned are organized by shape. In Figure 4-8b, the inductive constraint is acquired before any single category mean is securely known. Note that both trajectories in Figure 4-8 suggest that the inductive constraint supports top-down inferences about novel categories once it has been acquired. The crucial difference between the two is whether some variables at level 2 must be securely known before learning can take place at level 3.

Both trajectories in Figure 4-8 are consistent with a hierarchical Bayesian approach, and the trajectory that emerges in any particular situation will depend on the task and the available data. It may turn out that the four-step account of Smith et al. (2002) is accurate, and that Figure 4-8a provides the best description of the emergence of the shape bias. Figure 4-8b, however, may apply to situations where a child has access to a large number of sparse or noisy observations—any individual observation may be difficult to interpret, but taken together they may provide strong support for a general conclusion. For example, a hierarchical Bayesian model of grammar induction may be able to explain how a child becomes confident about some property of a grammar even though most of the individual sentences that support this conclusion are poorly understood. Similarly, a hierarchical approach may explain how a child can learn that visual objects are cohesive, bounded and rigid (E. S. Spelke, 1990) before developing a detailed understanding of the appearance and motion of any individual object.

Discovering ontological kinds

The Dirichlet-multinomial model in Figure 4-1a is a simple hierarchical model that acquires something like the shape bias, but to match the capacities of a child it is necessary to apply the shape bias selectively—to object categories, for example, but not to substance categories. Selective application of the shape bias appears to demand knowledge that categories are grouped into ontological kinds and that

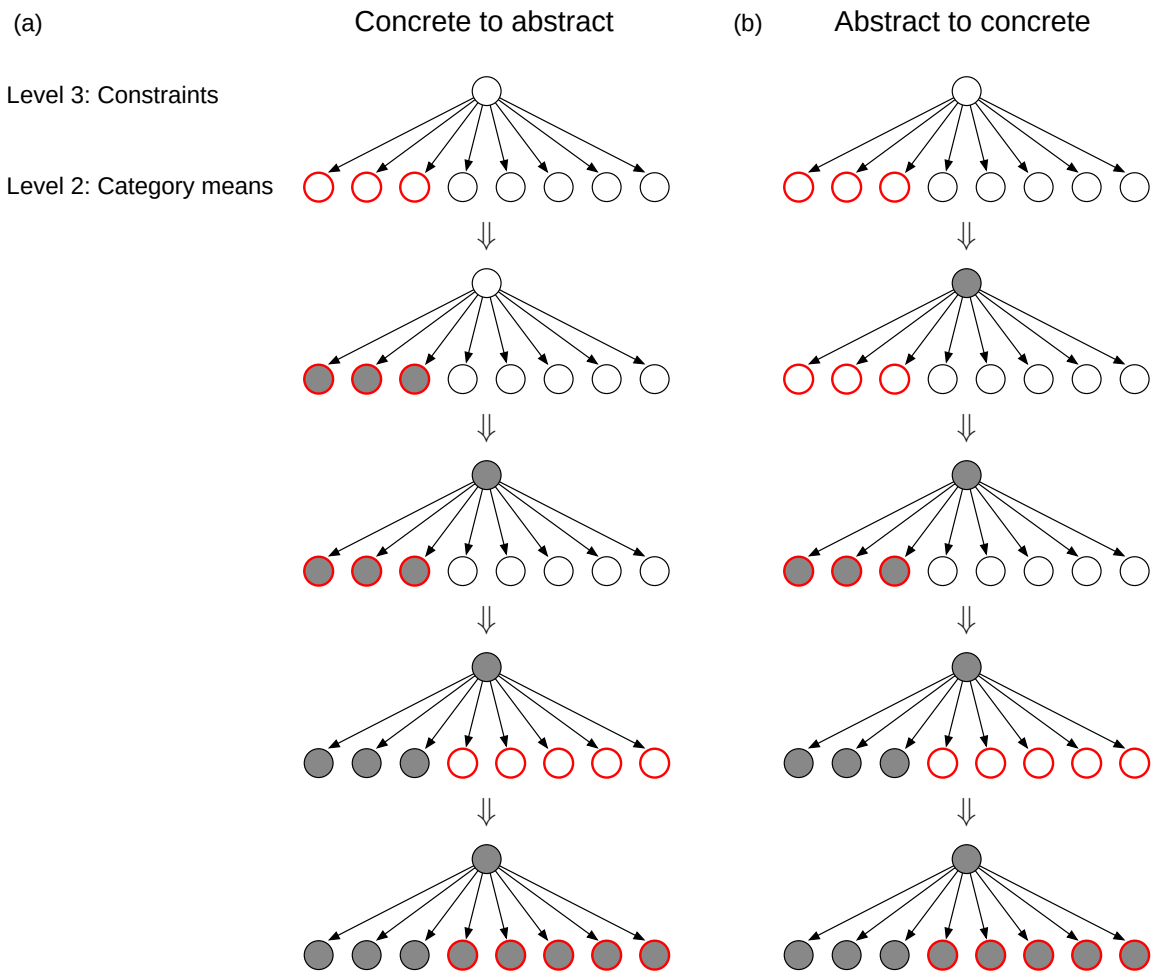

Figure 4-8: Two developmental trajectories that can emerge from a hierarchical Bayesian approach. In each trajectory, a learner acquires an inductive constraint (e.g. the shape bias) after receiving data at level 1 of the model (information about the features of several categories). The learner starts out by observing exemplars from the first three categories (the first three category means are drawn in red), and later observes exemplars from five additional categories. Filled circles indicate cases where the learner is near-certain about the value of a category mean or confident that she has discovered the inductive constraint. (a) The inductive constraint is discovered after the learner is near-certain about some of the category means (cf. Figure 4-7b.i). (b) The inductive constraint is discovered before the learner is confident about the values of any of the category means (cf. Figure 4-7b.iii). Both trajectories indicate that the inductive constraint supports inferences about novel categories once it has been acquired.

there are different patterns of feature variability within each kind. Before the age of three, for instance, children appear to know that shape tends to be homogeneous within object categories but heterogeneous within substance categories (Soja, Carey, & Spelke, 1991; Imai, Gentner, & Uchida, 1994; Samuelson & Smith, 1999), that color tends to be homogeneous within substance categories but heterogeneous within object categories (Landau et al., 1988; Soja et al., 1991), and that both shape and texture tend to be homogeneous within animate categories (S. S. Jones, Smith, & Landau, 1991).

Figure 4-1b shows a hierarchical model with two ontological kinds. The model includes trees for each kind: the first three categories are grouped into one kind, and the remaining three categories are grouped into a second kind. There are separate parameters α^k and $\boldsymbol{\beta^k}$ for each ontological kind k, and these parameters capture the features and the patterns of feature variability that are characteristic of each kind. For instance, α^1 will indicate that categories of the first kind are homogeneous in shape but not in material, and α^2 will indicate that categories of the second kind are homogeneous in material but not shape. The parameter $\boldsymbol{\beta^1}$ will indicate that members of the first kind tend to be solid, and $\boldsymbol{\beta^2}$ will indicate that members of the second kind tend not to be solid.

We can develop a model that learns for itself how to partition a set of categories into ontological kinds. Formally, let each possible partition be represented by a vector $\boldsymbol{z}$. The partition in 4-1b has $\boldsymbol{z} = [1, 1, 1, 2, 2, 2]$ which indicates that the first three categories belong to one ontological kind, and the remaining three belong to a second kind. As before, we assume that feature counts $\boldsymbol{y}$ are observed for one or more categories. Given these observations, the best set of ontological kinds will correspond to the $\boldsymbol{z}$ which maximizes $P(\boldsymbol{z}|\boldsymbol{y})$. This posterior distribution can be written as a product of two terms:

$$P(\boldsymbol{z}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{z})P(\boldsymbol{z}).$$

The first term will be high when $\boldsymbol{z}$ accounts well for the data: in other words, when categories belonging to the same kind tend to have similar features and similar pat-

terns of feature variability. The second term captures prior knowledge about the partition of categories into ontological kinds. We use a prior $P(\boldsymbol{z})$ that assigns some probability to all possible partitions, but favors the simpler partitions—those that use a small number of kinds. Many different priors satisfy this criterion, and we use a prior induced by the Chinese Restaurant Process (CRP, Aldous, 1985):

$$p(z_i = a | z_1, \ldots, z_{i-1}) = \begin{cases} \frac{n_a}{i-1+\gamma} & n_a > 0 \\ \frac{\gamma}{i-1+\gamma} & a \text{ is a new kind} \end{cases} \tag{4.1}$$

where z_i is the kind assignment for category i, n_a is the number of categories previously assigned to kind a, and γ is a hyperparameter (we set $\gamma = 0.5$). This process prefers to assign new categories to kinds which already have many members, and therefore favors partitions that use a small number of kinds.

Using statistical notation, the entire model in Figure 4-1b can be written as follows:

$$\begin{aligned} \boldsymbol{z} &\sim \text{CRP}(\gamma) \\ \alpha^k &\sim \text{Exponential}(\lambda) \\ \boldsymbol{\beta^k} &\sim \text{Dirichlet}(\mathbf{1}) \\ \boldsymbol{\theta^i} &\sim \text{Dirichlet}(\alpha^{z_i} \boldsymbol{\beta^{z_i}}) \\ \boldsymbol{y^i} \,|\, n^i &\sim \text{Multinomial}(\boldsymbol{\theta^i}) \end{aligned}$$

where most components of the model are carried over from our formalization of Figure 4-1a.

When fitting the model to data ($\boldsymbol{y}$), our goal is to simultaneously infer the partition of categories into kinds, along with the α^k and $\boldsymbol{\beta^k}$ for each kind k and the feature distribution $\boldsymbol{\theta^i}$ for each category. If $\boldsymbol{z}$ were already known, the model would reduce to several independent copies of the model in Figure 4-1a, and model predictions (including $p(\boldsymbol{\theta^i} | \boldsymbol{z}, \boldsymbol{y})$) could be computed using the techniques already described.

Since $\boldsymbol{z}$ is unknown, we integrate over this quantity:

$$p(\boldsymbol{\theta^i}|\boldsymbol{y}) = \sum_{\boldsymbol{z}} p(\boldsymbol{\theta^i}|\boldsymbol{z}, \boldsymbol{y}) P(\boldsymbol{z}|\boldsymbol{y}). \tag{4.2}$$

Since we are interested in problems where the number of categories is small, we compute this sum by enumerating all possible partitions.[2]

S. S. Jones and Smith (2002) showed that training young children on a handful of suitably structured categories can promote the acquisition of ontological knowledge. We gave our model a data set of comparable size. During training, the model saw two exemplars from each of four categories: two object categories and two substance categories (Figure 4-9a). Exemplars of each object category were solid, matched in shape, and differed in material and size. Exemplars of each substance category were non-solid, matched in material, and differed in shape and size. Second-order generalization was tested using exemplars from novel categories—one test exemplar (S) was solid and the other (N) was not (Figure 4-9b). Figure 4-9c shows that the model chooses a shape match for the solid exemplar and a material match for the non-solid exemplar.

Figure 4-9d confirms that the model correctly groups the stimuli into two ontological kinds: object categories and substance categories. This discovery is based on the characteristic features of ontological kinds ($\boldsymbol{\beta}$) as well as the patterns of feature variability within each kind (α). If kind k includes only the object categories, then α^k will indicate that shape is homogeneous within categories of this kind, and $\boldsymbol{\beta^k}$ will indicate that categories of this kind tend to be solid. The $\boldsymbol{\beta}$ parameter, then, is responsible for the inference that the category including S should be grouped with the two object categories, since all three categories contain solid objects.

[2]To compute the sum in Equation 4.2 we use $P(\boldsymbol{z}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{z})P(\boldsymbol{z})$, where $P(\boldsymbol{z})$ is the CRP prior on $\boldsymbol{z}$. Computing $P(\boldsymbol{y}|\boldsymbol{z})$ reduces to the problem of computing several marginal likelihoods

$$P(\boldsymbol{y\prime}) = \int_{\alpha,\boldsymbol{\beta}} P(\boldsymbol{y\prime}|\alpha, \boldsymbol{\beta}) p(\alpha, \boldsymbol{\beta}) d\alpha d\boldsymbol{\beta}$$

for the model in Figure 4-1a. We estimate each of these integrals by drawing 10,000 samples from the prior $p(\alpha, \boldsymbol{\beta})$.

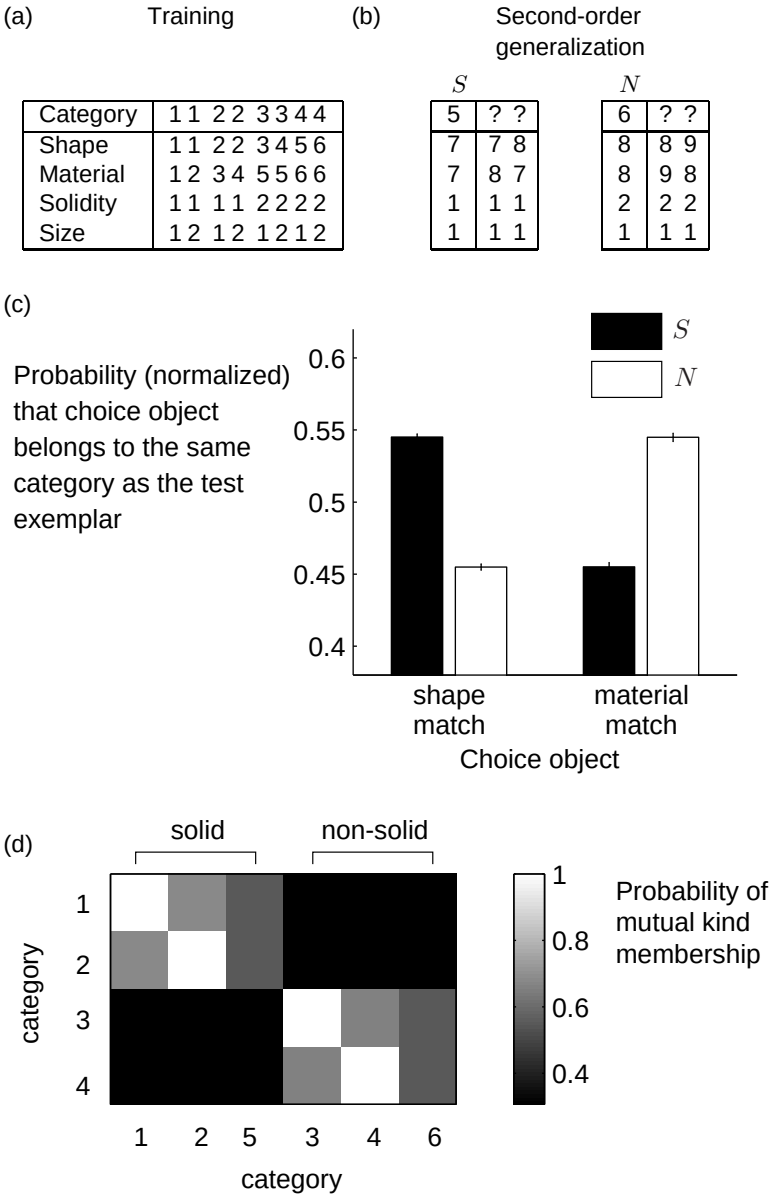

Figure 4-9: Learning a shape bias for solids and a material bias for non-solids. (a) Training data. (b) Second-order generalization was tested using solid and non-solid exemplars (S, N). In each case, two choice objects were provided — a shape match and a material match. (c) The model chooses the shape match given the solid exemplar and the material match given the non-solid exemplar. The model makes exact predictions about the probabilities plotted, and the error bars represent standard error across 8 estimates of these probabilities. (d) The model groups the categories into two kinds: objects (categories 1, 2 and 5) and substances (categories 3, 4 and 6). Entry (i, j) in the matrix is the posterior probability that categories i and j belong to the same ontological kind (light colors indicate high probabilities).

The results in Figure 4-9 predict that a training regime with a small number of categories and exemplars should allow children to simultaneously acquire a shape bias for solids and a material bias for substances. Samuelson (2002) ran a related study where she attempted to teach one group of children a precocious shape bias and another a precocious material bias. Only the shape bias was learned, suggesting that the shape bias is easier to teach than the material bias, but leaving open the possibility that the material bias could have been acquired with more training. Simultaneously teaching a shape bias for solids and a material bias for substances may raise some difficult practical challenges, but S. S. Jones and Smith (2002) have shown that children can simultaneously learn two kind-specific biases. By the end of their training study, children had learned that names for animate exemplars (exemplars with eyes) should be generalized according to shape and texture, and that names for objects (exemplars without eyes) should be generalized only according to shape. The model in Figure 4-1b accounts for these results: given the data provided to the children in these experiments, it discovers that there are two ontological kinds, and makes selective generalizations depending on whether or not a novel exemplar has eyes.

Related models

The models described in this chapter address tasks that have been previously modeled by Colunga and Smith (2005). These authors developed a connectionist network that acquires a shape bias for solid objects and a material bias for non-solid objects. The network uses a set of hidden nodes to capture high-order correlations between nodes representing the shape, material, and solidity of a collection of training objects, and generates results similar to Figure 4-9c when asked to make predictions about novel objects. The Dirichlet-multinomial model is similar to this connectionist model in several respects: both models show that abstract knowledge can be acquired, and both models are statistical, which allows them to deal with noise and uncertainty and to make graded generalizations. These models, however, differ in at least two

important respects.

First, the two models aim to provide different kinds of explanations. Our contribution is entirely at the level of computational theory (Marr, 1982), and I have not attempted to specify the psychological mechanisms by which the Dirichlet-multinomial model might be implemented. Colunga and Smith (2005) describe a process model that uses a biologically-inspired learning algorithm, but provide no formal description of the problem to be solved. Their network can probably be viewed as an approximate implementation of some computational theory,[3] but the underlying computational theory may not be ideal for the problem of word learning. For instance, it is not clear that the network adequately captures the notion of a category. In tests of second-order generalization (e.g. Figure 4-9c), the Dirichlet-multinomial model is able to compute the probability that a choice object belongs to the same category as the test exemplar. Colunga and Smith (2005) compute model predictions by comparing the similarity between hidden-layer activations for the choice object and the test exemplar. Objects in the same category may often turn out to have similar representations, but there are some well-known cases where similarity and categorization diverge (Keil, 1989; Rips, 1989).

A second limitation of the connectionist approach is that it does not extend naturally to contexts where structured representations are required. So far we have seen models that generate scalars (α) and vectors ($\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}$), but hierarchical probabilistic models can generate many other kinds of representations, including causal models (Chapter 5), graph structures (Chapter 6), ontologies (Schmidt, Kemp, & Tenenbaum, 2006), parse trees (Perfors, Tenenbaum, & Regier, 2006), and logical theories (Milch et al., 2005).

Previous researchers have developed Bayesian models of categorization (Anderson, 1991) and word learning (Tenenbaum & Xu, 2000), and our work continues in this tradition. The hierarchical approach, however, attempts to address a problem raised by most Bayesian models of cognition. A conventional Bayesian model matches the

[3]The network used by Colunga and Smith (2005) is related to a Boltzmann machine (Ackley, Hinton, & Sejnowski, 1985), which is an exact implementation of a known computational theory.

schema in Figure 2-3a: the elements in its hypothesis space represent level 2 knowledge, and the prior distribution over this space belongs to the collection of background assumptions. One common reservation about Bayesian models is that different priors account for different patterns of data, and the success of any given Bayesian model depends critically on the modeler's ability to choose the right prior. Hierarchical models disarm this objection by showing that the prior distribution over level 2 need not be specified in advance, but can be learned from raw data.

Hierarchical Bayesian models still rely on some background assumptions, including a prior distribution over the representations at the highest level. The ultimate goal, however, is to design models where this prior is simple enough to be unobjectionable. The Dirichlet-multinomial model demonstrates that hierarchical models can sometimes rely on much simpler priors than conventional Bayesian models. If we were only interested in inferences about level 2 knowledge (inferences about the $\boldsymbol{\theta^i}$ for each bag i), α and $\boldsymbol{\beta}$ (Figure 4-1a) would not be essential: in other words, a conventional Bayesian model could mimic the predictions of our model if it used the right prior distribution on the set $\{\boldsymbol{\theta^i}\}$. If specified directly, however, this prior would look extremely complicated—much more complicated, for example, than the prior used by the conventional model in Figure 4-5, which assumes that all of the $\boldsymbol{\theta^i}$ are independent. We avoided this problem by specifying the prior on $\{\boldsymbol{\theta^i}\}$ *indirectly*. We introduced an extra layer of abstraction—the layer including α and $\boldsymbol{\beta}$—and placed simple priors on these variables. These simple distributions on α and $\boldsymbol{\beta}$ induce a complicated prior distribution on $\{\boldsymbol{\theta^i}\}$—the same distribution that a conventional Bayesian model would have to specify directly.

Conclusion

This chapter presented our first fully-specified example of a hierarchical Bayesian model (Figure 4-1a). The model is one of the simplest possible hierarchical models, and provides a gentle introduction to the hierarchical Bayesian approach. Despite its simplicity, the model addresses a problem of cognitive interest, and helps to explain

how people learn simultaneously about the features of individual categories and about categories in general. Word learners who acquire the shape bias appear to solve an instance of this problem, and the model helps to account for word-learning data collected by Smith et al. (2002).

The Dirichlet-multinomial model addresses several psychological phenomena that we have not explored, but which may repay additional study. When shown a circle with a diameter of three inches, participants report that the circle is more likely to be a pizza than a quarter, even though the circle is closer in size to the average quarter than the average pizza (Rips, 1989). The model suggests that this decision is driven by knowledge about the variability of the size feature and predicts that people also know that exemplars of any given currency are usually the same size, but that exemplars of any given food tend to vary in size. This prediction could be tested using novel foods and currencies: for instance, a coin and a circular food from a novel country. The model also accounts for some of Harlow's experiments on "learning to learn" (Harlow, 1949). Harlow gave monkeys a blocked forced-choice decision task, where the same object was rewarded within each block regardless of whether it appeared on the left or the right. After many blocks, Harlow found that his monkeys were almost always choosing correctly after the second trial in each block. They had evidently learned that the rewarded objects in each block were homogeneous in shape and color, but heterogeneous in position.

There are many proposals about constraints that guide word learning, and it will be important to develop models that acquire constraints other than the shape bias. Future models can also explore how inferences about novel words differ from inferences about novel properties. It should be possible to develop a single model that acquires one set of constraints when learning about the extensions of words, and a different set of constraints when learning about the extensions of properties. Suppose, for instance, that a given set of objects can be organized into two cross-cutting systems of categories: a taxonomic system and a thematic system (cf. Shafto, Kemp, Mansinghka, Gordon, and Tenenbaum (2006)). A constraint-learning model might discover that words tend to pick out the taxonomic categories (a constraint known

as the taxonomic bias (Markman, 1989)), and that different words are likely to have different extensions (a constraint known as the principle of contrast (Clark, 1987)). The same model might discover that properties are subject to different constraints: properties that respect the thematic categories may be fairly common, and there are might be many cases where different properties (e.g. "renate" and "chordate") have the same extension. As this example suggests, a successful constraint-learning model should be able to discover how words are different from properties, and should also serve as a model of property induction.

The most intriguing suggestion to emerge from this chapter is the idea that inductive constraints can be learned relatively fast (Figure 4-8). Constraints that are present early in development are sometimes thought to be innate, but some of these constraints might be learned extremely rapidly. Hierarchical Bayesian models predict that some constraints can be learned given very small amounts of data, and that some constraints can emerge before more concrete kinds of knowledge are securely established. Exploring how well these ideas account for developmental data is an important direction for future work.

Chapter 5

Learning causal schemata

People often make confident causal inferences given very sparse data. Imagine, for instance, that you are travelling in the tropics, and on your very first morning you take an anti-malarial pill and wash it down with guava juice. Soon afterward you develop a headache and wonder what might have caused it. Suppose that you have very little direct information about the two potential causes—you have never before tasted guava juice or taken anti-malarial pills. Even so, you will probably correctly attribute your headache to the pill rather than the juice.

Accurate inferences from sparse data are often a sign that learners are relying on strong inductive constraints. In this case, your decision to blame the anti-malarial pill is probably constrained by knowledge about the causal powers of pills and juices. Even if you have never come across anti-malarial pills or guava juice, you probably believe that pills tend to cause headaches but that juices do not. Abstract causal beliefs of this sort are sometimes called causal schemata (Kelley, 1972) or intuitive theories.

This chapter describes a hierarchical Bayesian model that helps to explain how causal schemata are acquired. Part of our task is to formalize the notion of a causal schema. Suppose that we are interested in a set of objects—for example, a set of pills.

The work in this chapter was carried out in collaboration with Noah Goodman and Joshua Tenenbaum. An early version of this work was presented at the 29th Annual Conference of the Cognitive Science Society in 2007.

We will work with schemata that assign each object to a causal type, and that specify the causal powers and features of each type. Our pills, for instance, may represent four causal types—pills of type A cause headaches, pills of type B relieve headaches, and pills of types C and D neither cause nor relieve headaches. A causal schema may also specify how causal types interact. For instance, a C-pill and a D-pill may cause a headache when taken together, even though neither pill causes a headache on its own.

The work described in this chapter extends previous work on learning a causal model that captures the relationship between a single object (e.g. a pill) and an effect (e.g. a headache) (Figure 5-1a). Causal models for several objects can be learned independently, but this approach ignores any information that should be shared across objects: for instance, two blood-pressure medications are likely to have similar side effects, suggesting that a new blood-pressure medication will cause headaches if several others already have. To capture the idea that similar objects may have similar causal powers, we will work with causal schemata that organize a set of objects into causal types. (Figure 5-1b). This chapter shows how these schemata can be acquired in settings where learners must learn a schema at the same time as they are learning causal models for many different objects.

By tracking the characteristic features of causal types, learners can often make strong predictions about a novel object before it is observed to participate in any causal interactions. For instance, predictions about a pill with a given color, size, shape and imprint can be based on the effects produced by previous pills which shared these features. We will extend the notion of a causal schema by including information about the characteristic features of each causal type (Figure 5-1c). Although we begin with cases where at most one object is present at any time, the chapter ends by discussing cases where multiple objects may be present. We will extend the notion of a schema one more time by allowing interactions between different types (for instance, pills of type C may interfere with pills of type D), and will see how these characteristic interactions can be learned.

Although this thesis focuses on bottom-up learning of inductive constraints, hier-

archical Bayesian models are valuable in part because they support both top-down and bottom-up inferences (Griffiths, 2005). Top-down and bottom-up approaches are sometimes seen as competitors, and causal reasoning is one area where both approaches have been prominent. The top-down approach (Shultz, 1982; Bullock, Gelman, & Baillargeon, 1982) emphasizes inferences that are based on knowledge about causal powers, and the bottom-up approach emphasizes statistical inferences that are based on patterns of covariation. As P. W. Cheng (1993) and others have argued, these perspectives are best regarded as complementary: top-down knowledge about causal power plays a role in many inferences, and bottom-up statistical learning helps to explain how this knowledge is acquired. The apparent conflict between these perspectives may have developed in part because there is no well-established framework that accommodates them both. Kelley, for example, argued for both top-down (Kelley, 1972) and bottom-up approaches (Kelley, 1973) to causal reasoning, but did not develop a single theoretical framework that properly unified his two proposals. This chapter argues that a hierarchical Bayesian approach provides this missing theoretical framework, and the model I describe shows how top-down constraints support causal reasoning, and how these constraints can be acquired by statistical learning.

Learning about a single object

Although we will eventually consider inferences about multiple objects, suppose for now that we are interested in the relationship between a single object o and an effect e. Let V represent a set of trials where object o is present or absent on each trial, and effect e is or is not observed. For instance, if object o is a pill and effect e is a headache, each trial in V might indicate whether or not a patient takes a pill on a given day, and whether or not she subsequently experiences a headache. To simplify our notation, o will denote both the pill and the event of the patient swallowing the pill.

We will assume that the outcome of each trial is generated from a causal model M that captures the causal relationship between o and e (Figure 5-3). Having observed

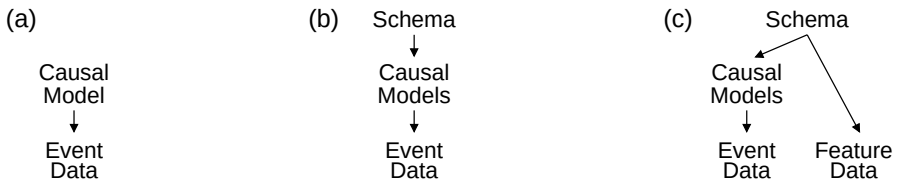

Figure 5-1: (a) A generative framework for discovering the causal powers of a single object. (b) A generative framework for learning a schema that guides inferences about multiple objects. The schema organizes the objects into causal types, and specifies the causal powers of each type. (c) A generative framework for learning a schema that includes information about the characteristic features of each type. Concrete examples of each framework are shown in Figures 5-3a, 5-4, and 5-14.

the trials in V, our beliefs about the causal model can be summarized by the posterior distribution

$$P(M|V) \propto P(V|M)P(M). \tag{5.1}$$

We build on the approach of Griffiths and Tenenbaum (2005) and parameterize the causal model M using four causal variables (Figures 5-2 and 5-3). Let a indicate whether there is an arrow joining o and e, and let g indicate the polarity of this causal relationship ($g = 1$ if o is a generative cause and $g = 0$ if o is a preventive cause). Suppose that s is the strength of the relationship between o and e.[1] To capture the possibility that e will be present even though o is absent, we assume that a generative background cause of strength b is always present. We specify the distribution $P(e|o)$ by assuming that generative and preventive causes combine according to a network of noisy-OR and noisy-AND-NOT gates (Glymour, 2001).

Now that we have parameterized model M in terms of the triple (a, g, s) and the background strength b, we can rewrite Equation 5.1 as

$$p(a, g, s, b|V) \propto P(V|a, g, s, b)P(a)P(g)p(s)p(b). \tag{5.2}$$

To complete the model we must place prior distributions on the four causal variables. We use uniform priors on the two binary variables (a and g), and assume that s is

[1] To simplify the later development of our model, we assume that g and s are defined even if $a = 0$ and there is no causal relationship between o and e. When $a = 0$, g and s can be interpreted as the polarity and strength that the causal relationship between o and e would have if this relationship actually existed.

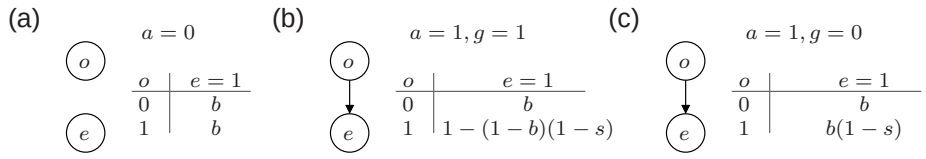

Figure 5-2: Causal graphical models which capture three possible relationships between an object o and an effect e. a indicates whether there is a causal relationship between o and e, g indicates whether this relationship is generative or preventive, and s indicates the strength of this relationship. A generative background cause of strength b is always present.

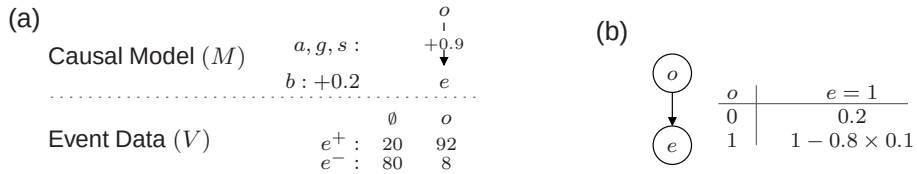

Figure 5-3: (a) Learning a causal model M from event data V (see Figure 5-1a). The event data specify the number of times the effect was (e^+) and was not (e^-) observed when o was absent ($\emptyset$) and when o was present. The model M shown has $a = 1$, $g = 1$, $s = 0.9$ and $b = 0.2$, and is a compact representation of the graphical model in (b).

drawn from a logistic normal distribution:

$$\begin{aligned} \text{logit}(s) &\sim \mathcal{N}(\bar{s}, \bar{\sigma}^2) \\ \bar{s} &\sim \mathcal{N}(\eta, \tau\bar{\sigma}^2) \\ \bar{\sigma}^2 &\sim \text{Inv-gamma}(\alpha, \beta) \end{aligned} \tag{5.3}$$

The priors on $\bar{s}$ and $\bar{\sigma}^2$ are chosen to be conjugate to the Gaussian distribution on logit(s), and we set $\alpha = 2$, $\beta = 0.3$, $\eta = 1$ and $\tau = 10$. The background strength b is drawn from the same distribution as s, and all hyperparameters are set to the same values except for η which is set to -1. Setting η to these different values encourages b to be small and s to be large, which matches standard expectations about the likely values of these variables.

To discover the causal model M that best accounts for the events in V, we can search for the causal variables with maximum posterior probability according to Equa-

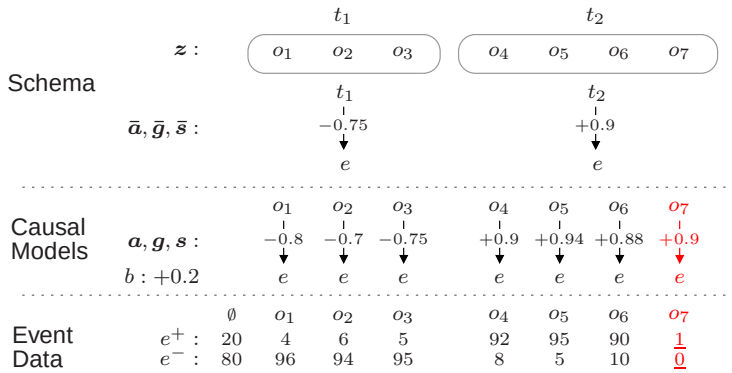

Figure 5-4: Learning a schema and a set of causal models (see Figure 5-1b). $\boldsymbol{z}$ specifies a set of causal types, where objects belonging to the same type have similar causal powers, and $\bar{\boldsymbol{a}}$, $\bar{\boldsymbol{g}}$, and $\bar{\boldsymbol{s}}$ specify the causal powers of each type. Note that the schema supports inferences about an object (o_7) that is very sparsely observed.

tion 5.2. There are many empirical studies that explore human inferences about a single potential cause and a single effect, and Griffiths and Tenenbaum (2005) show that a Bayesian approach similar to ours can account for many of these inferences. Here, however, we turn to the less-studied case where people must learn about many objects, each of which may be causally related to the effect of interest.

Learning about multiple objects

Suppose that we are interested in a set of objects $\{o_i\}$ and a single effect e. We begin with the case where at most one object is present at any time: for example, suppose that our patient has prescriptions for many different pills, but takes at most one pill per day. Instead of learning a single causal model our goal is to learn a set $\{M_i\}$ of causal models, one for each pill (Figures 5-1b and 5-4). There is now a triple (a_i, g_i, s_i) describing the causal model for each pill o_i, and we collect these variables into three vectors, $\boldsymbol{a}$, $\boldsymbol{g}$ and $\boldsymbol{s}$. Let Ψ be the tuple $(\boldsymbol{a}, \boldsymbol{g}, \boldsymbol{s}, b)$ which includes all the parameters of the causal models. As for the single object case, we assume that a generative background cause of strength b is always present.

One strategy for learning multiple causal models is to learn each model separately using the methods described in the previous section. Although simple, this strategy

does not capture the intuition that inferences about sparsely observed objects should be shaped by experience with previous objects. We can allow knowledge about familiar objects to influence predictions about novel objects by introducing the notion of a causal schema. A schema specifies a grouping of the objects into causal types, and indicates the causal powers of each of these types. The schema in Figure 5-4 indicates that there are two causal types: objects of type t_1 tend to prevent the effect, and objects of type t_2 tend to cause the effect. Formally, let z_i indicate the type of o_i, and let $\bar{\boldsymbol{a}}$, $\bar{\boldsymbol{g}}$, and $\bar{\boldsymbol{s}}$ be schema-level analogues of $\boldsymbol{a}$, $\boldsymbol{g}$, and $\boldsymbol{s}$: $\bar{a}(t)$ is the probability that any given object of type t will be causally related to the effect, and $\bar{g}(t)$ and $\bar{s}(t)$ are the expected polarity and causal strength for objects of type t. Even though $\bar{\boldsymbol{a}}$ and $\bar{\boldsymbol{g}}$ are vectors of probabilities, Figure 5-4 simplifies by showing each $\bar{a}(t)$ and $\bar{g}(t)$ as a binary variable.

To generate a causal model for each object, we assume that each arrow variable a_i is generated by tossing a coin with weight $\bar{a}(z_i)$, that each polarity g_i is generated by tossing a coin with weight $\bar{g}(z_i)$, and that each strength s_i is drawn from the logistic transform of a normal distribution with mean $\bar{s}(z_i)$ and variance $\bar{\sigma}(z_i)$. Let $\bar{\Psi}$ be the tuple $(\bar{\boldsymbol{a}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{s}}, \bar{\boldsymbol{\sigma}})$. To complete the model, we specify prior distributions on $\boldsymbol{z}$ and $\bar{\Psi}$. As in Chapter 4, we use a Chinese restaurant process prior on $\boldsymbol{z}$ (Equation 4.1) and set the γ parameter to 0.5. This prior assigns some probability mass to all possible partitions but favors partitions that use a small number of types. We assume that the entries in $\bar{\boldsymbol{a}}$ and $\bar{\boldsymbol{g}}$ are independently drawn from a Beta$(0.1, 0.1)$ distribution, and that the means and variances in $\bar{\boldsymbol{s}}$ and $\bar{\boldsymbol{\sigma}}$ are independently drawn from the conjugate priors in Equation 5.3.

Having defined a generative model, we can use it to learn the type assignments $\boldsymbol{z}$, the schema parameters $\bar{\Psi}$ and the parameters Ψ of the causal models that are most probable given the events V we have observed:

$$p(\boldsymbol{z}, \bar{\Psi}, \Psi | V) \propto P(V|\Psi) P(\Psi|\bar{\Psi}, \boldsymbol{z}) p(\bar{\Psi}|\boldsymbol{z}) P(\boldsymbol{z}). \tag{5.4}$$

Figure 5-4 shows how a schema and a set of causal models (top two sections) can be

simultaneously learned from the events V in the bottom section. All of the variables in the figure have been set to values with high posterior probability according to Equation 5.4: for instance, the partition $\boldsymbol{z}$ shown is the $\boldsymbol{z}$ with maximum posterior probability. Note that learning a schema supports confident inferences about object o_7, which is very sparsely observed (see the underlined entries in the bottom section of Figure 5-4). On its own, a single trial might not be very informative about the causal powers of a novel object, but experience with previous objects allows the model to predict that o_7 will produce the effect about as regularly as the other members of type t_2.

To compute the predictions of our model we used Markov chain Monte Carlo (MCMC) methods to sample from the posterior distribution in Equation 5.4. Since we use conjugate priors on the schema parameters $\bar{\Psi}$, we can integrate out these parameters and sample from $p(\boldsymbol{z}, \Psi|V)$. To sample the schema assignments in $\boldsymbol{z}$, we combined Gibbs updates with the split-merge scheme described by Jain and Neal (2004). We used Metropolis-Hasting updates on the parameters Ψ of the causal models, and found that mixing improved when the three parameters for a given object i (a_i, g_i and s_i) were updated simultaneously. To further facilitate mixing, we used Metropolis-coupled MCMC: we ran several Markov chains at different temperatures and regularly considered swaps between the chains (Geyer, 1991). All of these details, however, are of little psychological importance. The implementation described here is not intended as a process model, and the primary contribution of this section is the computational theory summarized by Equation 5.4.

Experiment 1: One-shot causal learning

The schema-learning model attempts to satisfy two criteria when learning about the causal powers of a novel object. When information about the new object is sparse, predictions about this object should be based primarily on experience with previous objects. Relying on previous objects will allow the model to go beyond the sparse and noisy observations that are available for the novel object. Given many observations of the novel object, however, the model should rely heavily on these observations

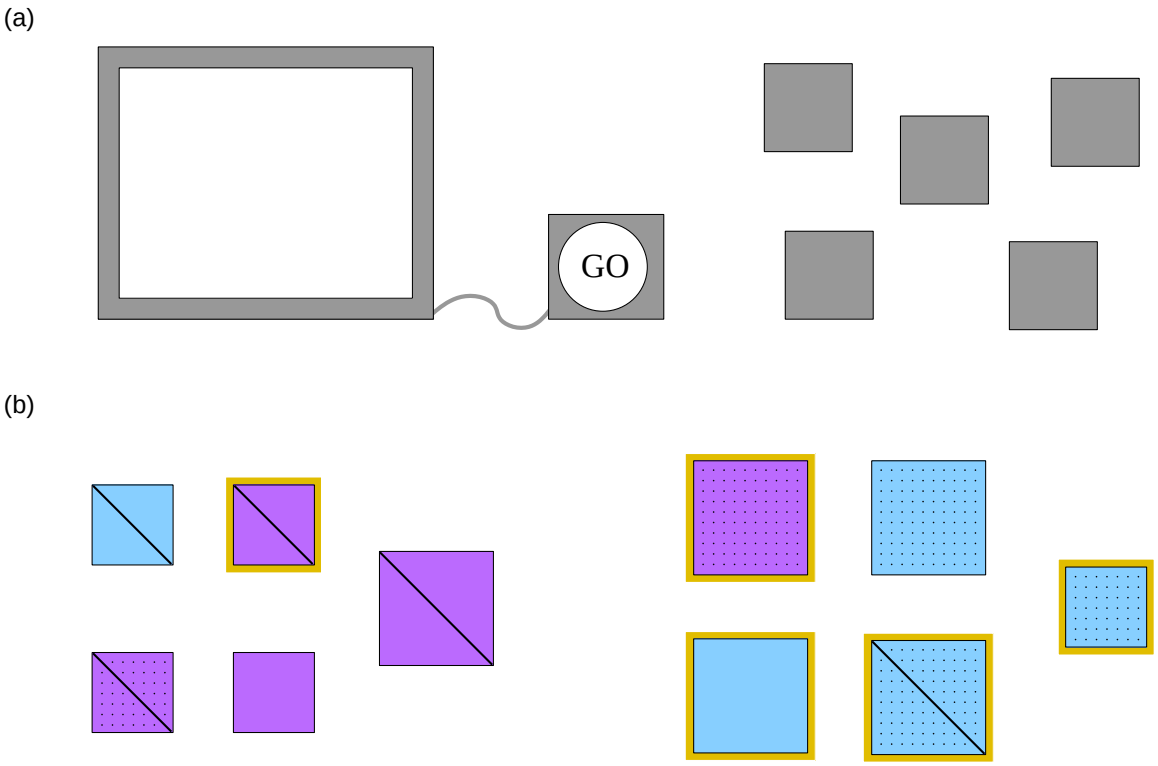

Figure 5-5: (a) A machine and some blocks. The blocks can be placed inside the machine, and the machine sometimes activates (flashes yellow) when the GO button is pressed. The blocks used for each condition of Experiments 1 and 2 were perceptually indistinguishable. (b) Blocks used for Experiment 3. The blocks are grouped into two family resemblance categories: blocks on the right tend to be large, blue and spotted, and tend to have a gold boundary but no diagonal stripe. These blocks are based on stimuli created by Sakamoto and Love (2004).

and should tend to ignore its observations of previous objects. Discounting past experience in this way will allow the model to be flexible if the new object turns out to be different from all of the previous objects.

We designed two experiments to explore this tradeoff between conservatism and flexibility. Both experiments used blocks and machines like the examples in Figure 5-5. The machine has a GO button, and may activate and flash yellow when this button is pressed. Blocks can be placed in the machine, and whether or not the machine is likely to activate might depend on which block is inside. In terms of the language I have been using, each block is an object o_i, each button press is a trial, and there is a single effect e which indicates whether the machine activated on a given trial.

Condition		Training data								
		$\emptyset$	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8
$p = \{0, 0.5\}$	$\mathbf{e}^+$:	0	0	0	0	0	5	4	6	1
	$\mathbf{e}^-$:	10	10	10	10	1	5	6	4	0
		$\emptyset$	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8
$p = \{0.1, 0.9\}$	$\mathbf{e}^+$:	0	1	2	1	0	9	8	9	1
	$\mathbf{e}^-$:	10	9	8	9	1	1	2	1	0
		$\emptyset$	o_1	o_2	o_3	o_4	o_5	o_6		
$p = 0$	$\mathbf{e}^+$:	0	0	0	0	0	0	0		
	$\mathbf{e}^-$:	10	10	10	10	10	10	10		
		$\emptyset$	o_1	o_2	o_3	o_4	o_5	o_6		
$p = 0.1$	$\mathbf{e}^+$:	0	1	2	1	2	1	2		
	$\mathbf{e}^-$:	10	9	8	9	8	9	8		

Figure 5-6: Training data for the four conditions of Experiment 1.

Our first experiment explores the idea that experience with several training blocks can guide inferences about a sparsely observed test block. The experiment includes several one-shot learning problems where participants make predictions about a test block after seeing a single trial involving that block.

Participants

24 members of the MIT community were paid for participating in this experiment.

Materials and Methods

The experiment includes four within-participant conditions, and the training data for each condition are summarized in Figure 5-6. The first condition ($p = \{0, 0.5\}$) includes blocks of two types: blocks of the first type never activate the machine, and blocks of the second type activate the machine about half the time. The second condition ($p = \{0.1, 0.9\}$) also includes two types: blocks of the first type rarely activate the machine, and blocks of the second type usually activate the machine. The remaining conditions each include one type of block: blocks in the third condition

($p = 0$) never activate the machine, and blocks in the fourth condition ($p = 0.1$) activate the machine rarely.

At the start of each condition, participants are shown an empty machine and asked to press the GO button 10 times. The machine fails to activate on each occasion. One by one, the training blocks are introduced, and participants place each block in the machine and press the GO button one or more times. The outcomes of these trials are summarized in Figure 5-6. After the final trial for each block, participants are asked to imagine pressing the GO button 100 times when this block is inside the machine. They then provide a rating which indicates how likely it is that the total number of activations will fall between 0 and 20. All ratings are provided on a 7 point scale where 1 indicates "very unlikely" and 7 indicates "very likely." Ratings are also provided for four other intervals: between 20 and 40, between 40 and 60, between 60 and 80, and between 80 and 100. After the training phase, two test blocks are introduced, again one at a time. Participants provide ratings for each block before it has been placed in the machine, and after a single trial. One of the test blocks (o^+) activates the machine on this trial, and the other (o^-) does not.

The set of four conditions is designed to test the idea that inductive constraints and inductive flexibility are both important. The first two conditions test whether experience with the training blocks allows people to extract constraints that are useful when making predictions about the test blocks. Conditions three and four explore cases where these constraints need to be overruled, since test block o^+ is surprising given that the training blocks in these conditions activate the machine rarely if at all.

To encourage participants to think about the conditions separately, machines and blocks of different colors were used for each condition. The order in which the conditions were presented was counterbalanced, and the order of the training blocks and the test blocks within each condition was also counterbalanced.[2]

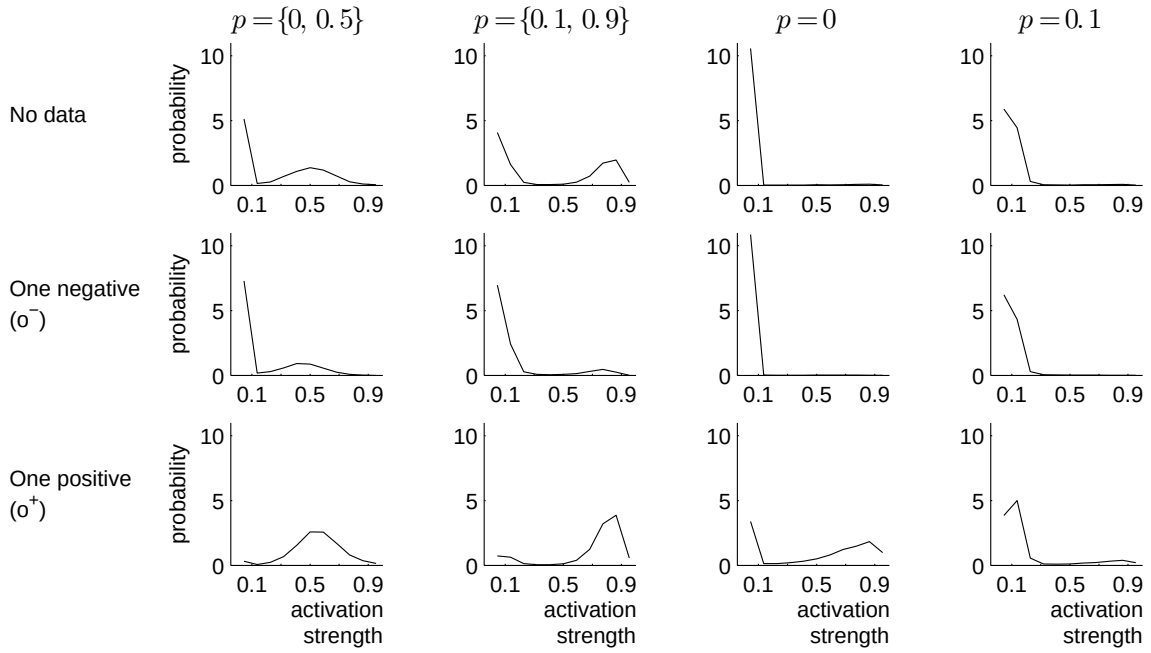

Figure 5-7: Experiment 1: predictions of the schema-learning model. Each subplot shows the posterior distribution on the causal power of a test block.

Model Predictions

Figure 5-7 shows predictions when the schema-learning model is applied to the data in Figure 5-6. Each plot shows the posterior distribution on the activation strength of a test block: the probability $p(e|o)$ that the block will activate the machine on a given trial.[3] Since the background rate is zero, this distribution is equivalent to a distribution on the causal power (P. Cheng, 1997) of the test block.

The plots in the first row show predictions about a test block before it is placed in the machine. The first plot indicates that the model has discovered two causal types, and expects that the test block will activate the machine either very rarely or around half of the time. The two peaks in the second plot again indicate that the model

[2]There was one exception: in condition 3, test block o^+ was always presented second, since this block is unlike any of the training blocks, and may have had a large influence on predictions about any block which followed it.

[3]Recall that participants were asked to make predictions about the number of activations expected across 100 trials. If we ask the model to make the same predictions, the distributions on the total number of activations will be discrete distributions with shapes similar to the distributions in Figure 5-7.

has discovered two causal types, this time with strengths around 0.1 and 0.9. The remaining two plots are unimodal, suggesting that only one causal type is needed to explain the data in each of the $p = 0$ and $p = 0.1$ conditions.

The plots in the second row show predictions about a test block (o^-) that fails to activate the machine on one occasion. All of the plots have peaks near 0 or 0.1. Since each condition includes blocks that activate the machine rarely or not at all, the most likely hypothesis is always that o^- is one of these blocks. Note, however, that the first plot has a small bump near 0.5, indicating that there is some chance that test block o^- will activate the machine about half of the time. The second plot has a small bump near 0.9 for similar reasons.

The plots in the third row show predictions about a test block (o^+) that activates the machine on one occasion. The plot for the first condition peaks near 0.5, which is consistent with the hypothesis that blocks which activate the machine at all tend to activate it around half the time. The plot for the second condition peaks near 0.9, which is consistent with the observation that some training blocks activated the machine nearly always. The plot for the third condition has peaks near 0 and near 0.9. The first peak captures the idea that the test block might be similar to the training blocks, which activated the machine very rarely. Given that none of the training blocks activated the machine, one positive trial is enough to suggest that the test block might be qualitatively different from all previous blocks, and the second peak captures this hypothesis. The curve for the final condition peaks near 0.1, which is the frequency with which the training blocks activated the machine.

Results

The four columns of Figure 5-8a show the results for each condition. Each participant provided ratings for five intervals in response to each question, and these ratings can be plotted as a curve. Figure 5-8a shows the mean curve for each question. The first row shows predictions before the first test block has been placed in the machine, and the second and third rows show predictions after a single trial for test blocks o^- and o^+.

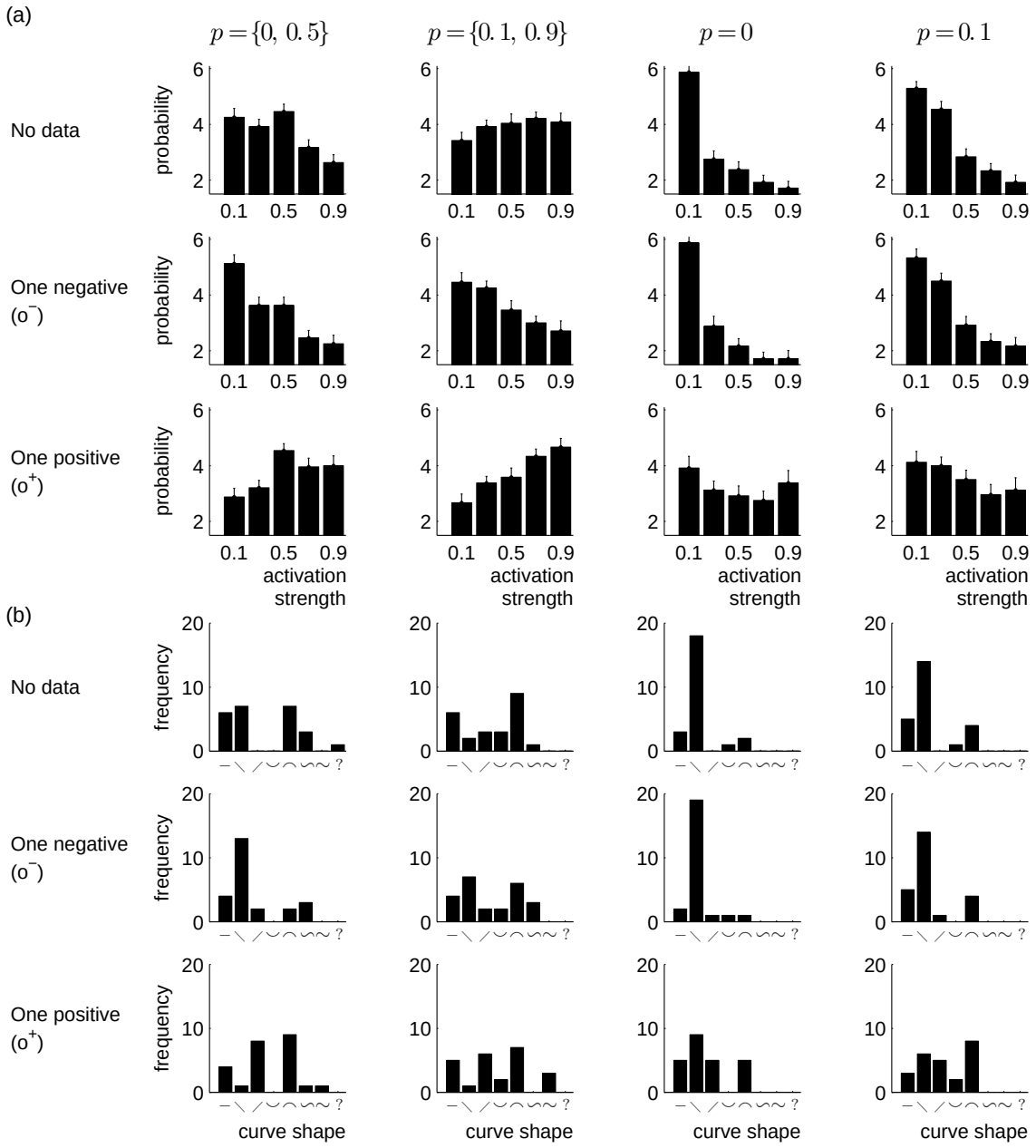

Figure 5-8: Each column shows results for one of the four conditions in Experiment 1. (a) Mean responses across 24 participants. Each subplot shows predictions about a new object that will undergo 100 trials, and each bar indicates the probability that the total number of activations will fall within a certain interval. The x-axis shows the activation strengths that correspond to each interval, and the y-axis shows probability ratings on a scale from 1 (very unlikely) to 7 (very likely). Error bars for this plot and all remaining plots show the standard error of the mean. (b) Individual responses classified by curve shape. The y axis shows the number of participants who gave responses of 7 different types, including flat curves ($-$), curves that increase ($/$) or decrease ($\backslash$) monotonically, unimodal curves ($\frown$), and bimodal curves ($\smile$, $\sim$, $\backsim$). The ? category includes all curves that did not match any of these shapes.

Note first that the plots in the third row are rather different from each other.[4] Each plot shows predictions about a test block (o^+) that has activated the machine exactly once, and the differences between these plots confirm that experience with previous blocks shapes people's inferences about a sparsely observed novel block. Note also that all of the plots in this row peak in the same places as the curves predicted by the model (Figure 5-8a).

The plots in the second row are all decaying curves, since each condition includes blocks that activate the machine rarely or not at all. Again, though, the differences between the curves are interpretable, and match the predictions of the model. For instance, the $p = 0$ curve decays more steeply than the others, which makes sense since the training blocks for this condition never activate the machine.

The curves in the first row are again different from each other, and the curves for $p = 0$ and $p = 0.1$ suggest that participants realize that blocks in these conditions rarely activate the machine. The curves for the first two conditions show the most substantial discrepancy between model predictions and human judgments. The model predicts that both curves should be bimodal, and there is a trend in this direction for the $p = \{0, 0.5\}$ condition, but the $p = \{0.1, 0.9\}$ curve is flat or unimodal. This result is consistent with previous findings that participants expect distributions to be unimodal, and may need to observe many samples from a distribution before concluding that it is bimodal (Flannagan, Fried, & Holyoak, 1986). An alternative interpretation is that learners rely on deterministic rules: they distinguish between blocks that never produce the effect and blocks that sometimes produce the effect, but not between blocks that produce the effect with different strengths. The first interpretation seems more plausible, and we predict that people will recognize the existence of two types in the $p = \{0.1, 0.9\}$ condition when many blocks are observed of each type. Our third experiment supports this prediction, although it does not test it directly.

[4] We analyzed the results summarized in the third row using a two factor ANOVA with repeated measures. There is no significant main effect of interval ($F(4, 92) = 0.46$, $p > 0.5$), but there is a significant main effect of condition ($F(3, 69) = 4.20$, $p < 0.01$) and a significant interaction between interval and condition ($F(12, 276) = 6.90$, $p < 0.001$).

In all cases except one, the average responses in Figure 5-8a are consistent with the responses of some individual participants. In Figure 5-8b, the curves provided by individual participants have been grouped into eight categories based on their shapes. In the $p = \{0, 0.5\}$ condition, for instance, most participants generate a descending curve ($\diagdown$) after observing o^- fail to activate the machine once, and most participants generate an inverted-U curve ($\frown$) after observing o^+ activate the machine once. Both responses match the shape of the mean curves shown in Figure 5-8a. Given no trials for the test block, however, some participants ($-$) appear unwilling to make inductive predictions, others ($\diagdown$ and $\frown$) appear to guess whether the test block will be a 0 block or a 0.5 block, and a minority give a response ($\backsim$) that matches the mean curve and indicates that the block could be a 0 block or a 0.5 block.

Although the responses of individual participants are revealing, I will focus on the mean response, which indicates the consensus opinion about the causal strength of a test block. Consider for instance the inference about test block o^+ in the $p = 0$ condition, which is the only case where no participant gives a response that matches the mean curve. Some participants seem confident that the o^+ block will activate the machine rarely, and that the single positive trial is an aberration. Others seem confident that the test block will activate the machine most of the time. Even though no single participant appears to entertain both hypotheses, the mean curve captures the finding that both hypotheses are plausible. A model that generates a similar curve has captured both of the hypotheses considered sensible by people.

Predicting the responses of individual participants is also a worthy challenge, and future models may wish to address this problem. An accurate model of individual behavior will need to consider some issues that we have been able to ignore. Individual responses, for instance, are likely to have been influenced by the order in which blocks were presented: a participant in the $p = \{0, 0.5\}$ condition might reason that the last block she saw was a 0 block, and that the next block will probably be similar. Some models of categorization can capture order effects (Anderson, 1991; Love, Medin, & Gureckis, 2004), and future work can explore how these effects play out in the experimental paradigm we have chosen. We decided, however, to ignore these effects

by counterbalancing presentation order across participants and by focusing on the mean response.

Experiment 2: Learning about new causal types

Although a single observation of a new object is sometimes enough to overrule expectations based on many previous objects, several trials may be required before learners are confident that a new object is unlike any of the previous objects. We designed a second experiment where participants receive increasing evidence that a new object is different from all previous objects.

Participants

16 members of the MIT community were paid for participating in this experiment.

Materials and Methods

The experiment includes two within-participant conditions ($p = 0$ and $p = 0.1$) that correspond to conditions 3 and 4 of Experiment 1. Each condition is very similar to the corresponding condition from Experiment 1 except for two changes. Seven observations are now provided for the two test blocks: for test block o^-, the machine fails to activate on each trial, and for test block o^+ the machine activates on all test trials except the second. Participants rate the causal strength of each test block after each trial, and also provide an initial rating before any trials have been observed. As before, participants are asked to imagine placing the test block in the machine 100 times, but instead of providing ratings for five intervals they now simply predict the total number of activations out of 100 that they expect to see.

Model Predictions

Figure 5-9 shows the results when the schema-learning model is applied to the tasks in Experiment 2. In both conditions, predictions about the test blocks track the

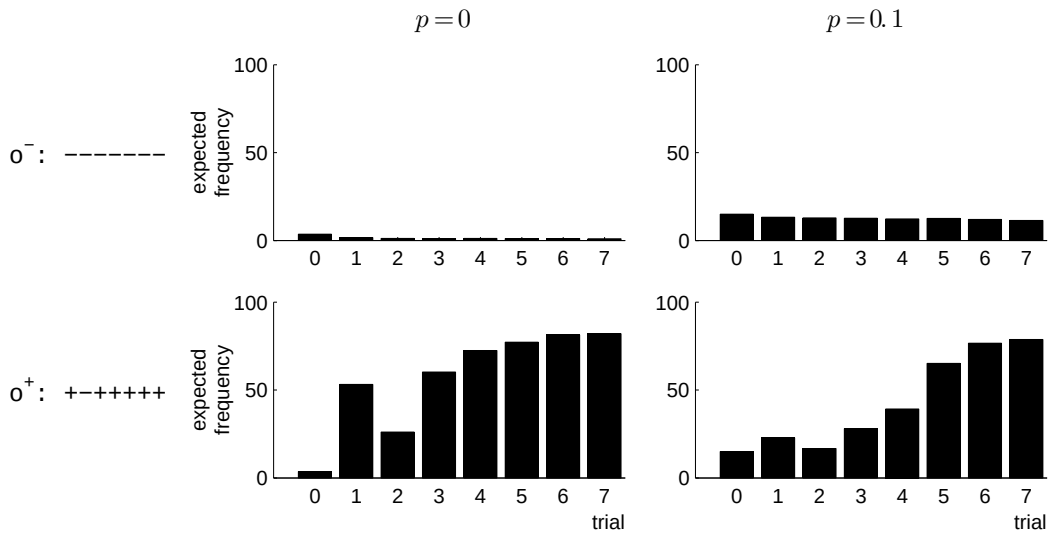

Figure 5-9: Experiment 2: predictions of the schema-learning model.

observations provided, and the curves rise after each positive trial and fall after each negative trial.

The most interesting predictions involve test block o^+, which is qualitatively different from all of the training blocks. The o^+ curves for both conditions attain similar values by the final prediction, but the curve for the $p = 0$ condition rises more steeply than the curve for the $p = 0.1$ condition. Since the training blocks in the $p = 0.1$ condition activate the machine on some occasions, the model needs more evidence in this condition before concluding that block o^+ is different from all of the training blocks.

The predictions about test block o^- also depend on the condition. In the $p = 0$ condition, none of the training blocks activates the machine, and the model predicts that o^- will also fail to activate the machine. In the $p = 0.1$ condition, each training block can be expected to activate the machine about 15 times out of 100. The curve for this condition begins at around 15, then gently decays as o^- repeatedly fails to activate the machine.

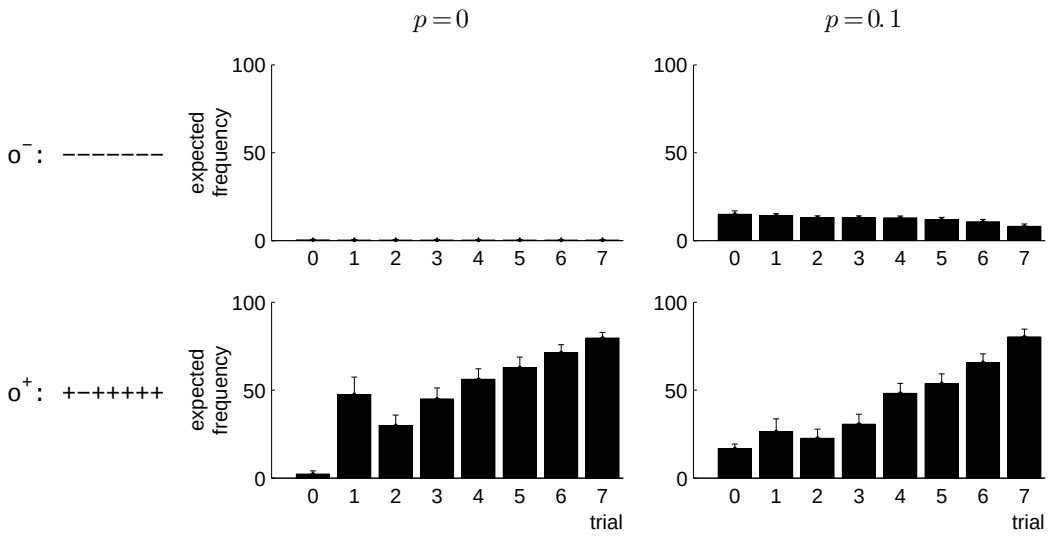

Figure 5-10: Average learning curves for Experiment 2.

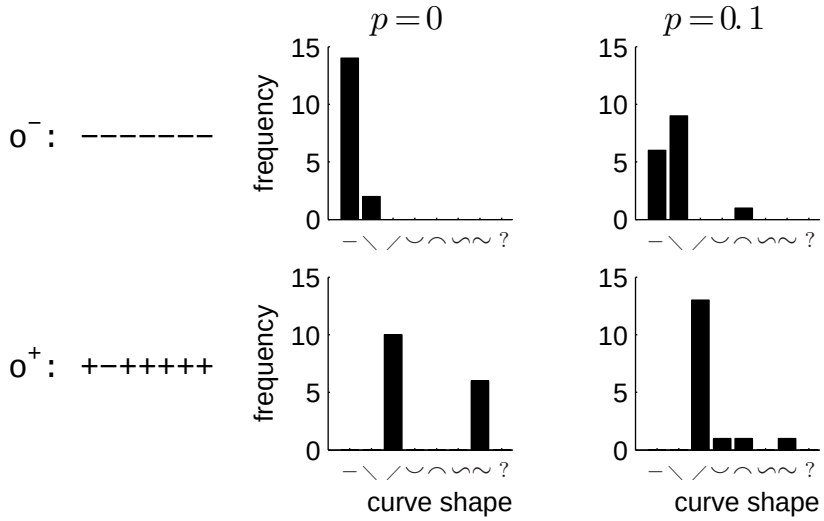

Figure 5-11: Individual learning curves for Experiment 2 classified by shape.

Results

Figure 5-10 shows the average learning curves across 16 participants. The curves are qualitatively similar to the model predictions, and as predicted the o^+ curve for the $p = 0$ condition rises more steeply than the corresponding curve for the $p = 0.1$ condition.[5] Note that a simple associative account might predict the opposite result, since the machine in condition $p = 0.1$ activates more times overall than the machine in condition $p = 0$. Learning curves for individual participants are summarized in Figure 5-11. In the $p = 0$ condition, six participants show learning curves for o^+ that match the shape of the mean curve ($\sim$), but curves that increase monotonically ($/$) are more common. The preference for increasing curves is even more pronounced in the $p = 0.1$ condition.

Alternative models

As mentioned already, our experiments explore the tradeoff between conservatism and flexibility. When a new object is sparsely observed, the schema-learning model assumes that this object is similar to previously encountered objects (Experiment 1). Once more observations become available, the model may decide that the new object is different from all previous objects, and should therefore be assigned to its own causal type (Experiment 2). We can compare the schema-learning model to two alternatives: a *reactionary* model that is overly conservative, and a *revolutionary* model that is overly flexible. The reactionary model assumes that each new object is just like one of the previous objects, and the revolutionary model ignores all of its previous experience when making predictions about a new object.

We implemented the revolutionary model by assuming that the causal power of a test block is identical to its empirical power—the proportion of trials on which it has activated the machine. Predictions of this model are shown in Figure 5-12. When

[5]Since we expect that the $p = 0$ curve should be higher than the $p = 0.1$ curve from the second judgment onwards, we ran a two factor ANOVA with repeated measures that excluded the first judgment from each condition. There are significant main effects of condition ($F(1, 15) = 6.11$, $p < 0.05$) and judgment number ($F(6, 90) = 43.21$, $p < 0.01$), and a significant interaction between condition and judgment number ($F(6, 90) = 2.67$, $p < 0.05$).

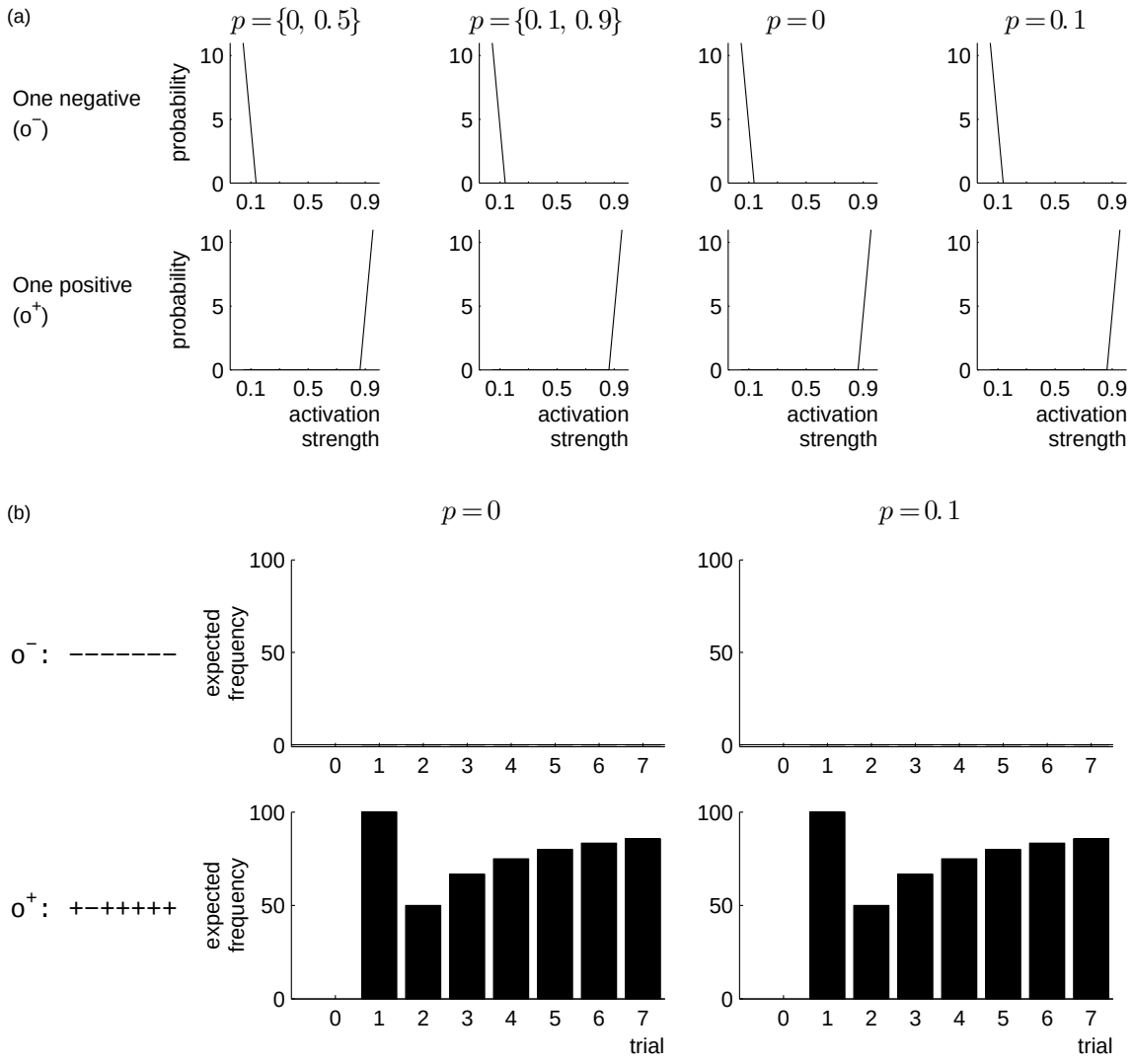

Figure 5-12: Predictions of the revolutionary model. (a) Experiment 1 (b) Experiment 2.

applied to Experiment 1, the most obvious failing of the revolutionary model is that it makes identical predictions about all four conditions. Note that the model does not make predictions about the first row of Figure 5-7a, since at least one test trial is needed to estimate the empirical power of a new block. When applied to Experiment 2, the model is unable to make predictions before any trials have been observed for a given object, and after a single positive trial the model leaps to the conclusion that test object o^+ will always activate the machine. Neither prediction matches the human data, and the model also fails to predict any difference between the $p = 0$ and $p = 0.1$ conditions.

We implemented the reactionary model by assuming that the causal power of each training block is identical to its empirical power, and that each test block is identical to one of the training blocks. The model, however, does not know which training block the test block will match, and makes a prediction that considers the empirical powers of all training blocks, weighting each one by its proximity to the empirical power of the test block. Formally, the distribution d_n on the strength of a novel block is defined to be

$$d_n = \frac{\sum_i w_i d_i}{\sum_i w_i} \tag{5.5}$$

where d_i is the distribution for training block i, and is created by dividing the interval $[0, 1]$ into eleven equal intervals, setting $d_i(x) = 11$ for all values x that belong to the same interval as the empirical power of block i, and setting $d_i(x) = 0$ for all remaining values. Each weight w_i is set to $1 - |p_n - p_i|$, where p_n is the empirical power of the novel block and p_i is the empirical power of training block i. As Equation 5.5 suggests, the reactionary model is closely related to exemplar models of categorization (D. L. Medin & Schaffer, 1978; Nosofsky, 1986).

Predictions of the reactionary model are shown in Figure 5-13. The model accounts fairly well for the results of Experiment 1, but is unable to account for Experiment 2. Since the model assumes that test object o^+ is just like one of the training objects, it is unable to adjust when o^+ activates the machine more frequently than any previous object.

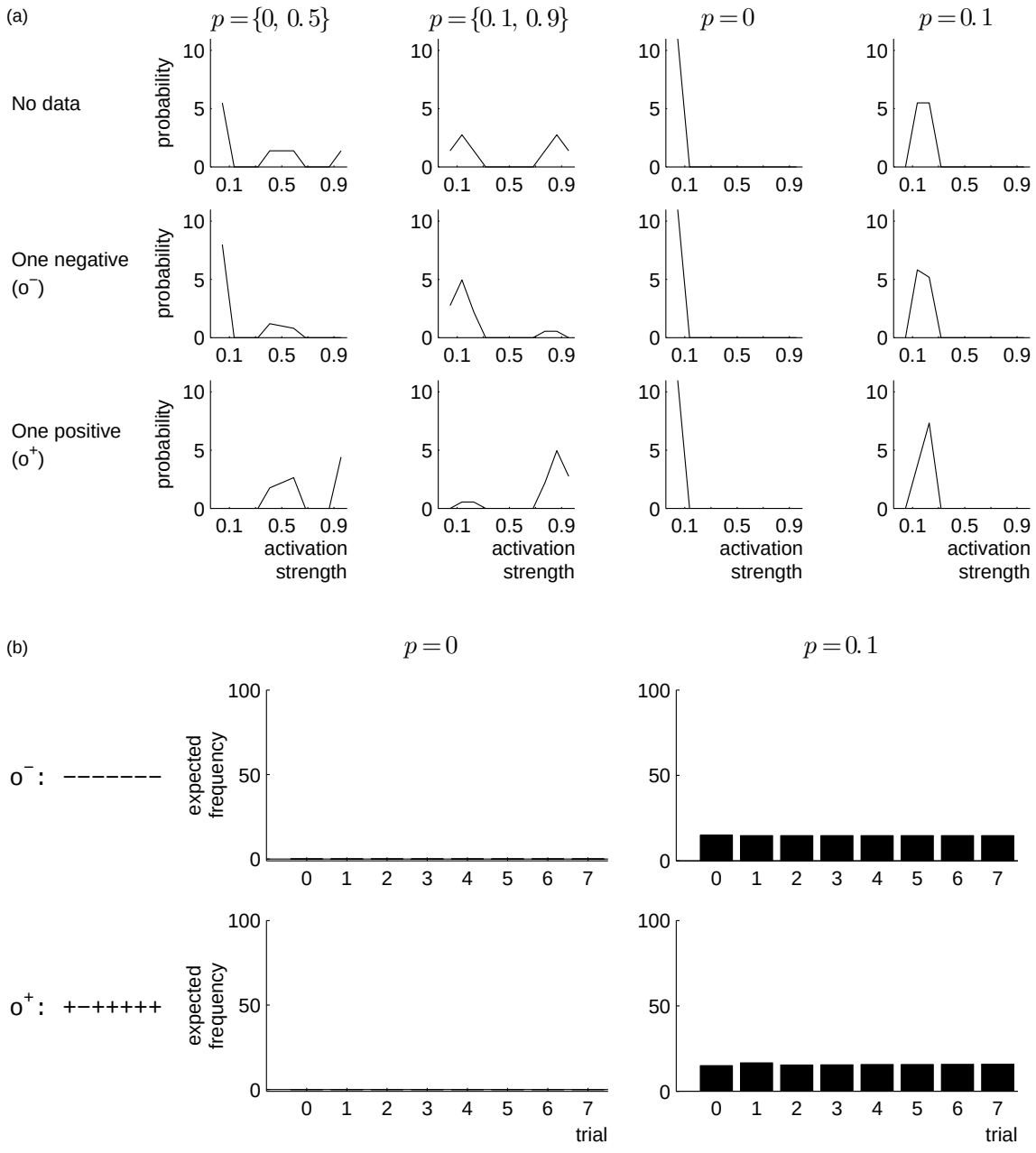

Figure 5-13: Predictions of the reactionary model. (a) Experiment 1 (b) Experiment 2.

Overall, neither baseline model can account for our results. As their names suggest, the revolutionary model is too quick to throw away observations of previous objects, and the reactionary model is unable to handle new objects that are qualitatively different from all previous objects. Other baseline models might be considered, but we are aware of no simple alternative that will account for all of our data.

Our first two experiments deliberately focused on a very simple setting where causal schemata are learned and used, but real world causal learning is often more complex. The rest of this chapter will address some of these complexities: in particular, I show that our framework can incorporate perceptual features and can handle contexts where objects interact to produce an effect.

Learning causal types given feature data

Imagine that you are allergic to nuts, and that one day you discover a small white sphere in your breakfast cereal—a macadamia nut, although you do not know it. To discover the causal powers of this novel object you could collect some causal data—you could eat it and wait to see what happens. Probably, however, you will observe the features of the object (its color, shape and texture) and decide to avoid it since it is similar to other allergy-producing foods you have encountered.

A hierarchical Bayesian approach can readily handle the idea that instances of a given causal type tend to have similar features (Figures 5-1c and 5-14). Suppose that we have a matrix F which captures many features of the pills in our study, including their sizes, shapes, colors, and imprints. We assume that objects belonging to the same type have similar features. For instance, the schema in Figure 5-14 specifies that objects of type t_2 tend to have features f_1 through f_4, but objects of type t_1 tend not to have these features. Formally, let the schema parameters $\bar{\Psi}$ include a matrix $\bar{F}$, where $\bar{f}_j(t)$ specifies the expected value of feature f_j within causal type t.[6] Building on previous models of categorization (Anderson, 1991), we assume that

[6]To apply Equation 5.6 we need to specify a prior distribution $p(\bar{F})$ on this matrix. We assume that all entries in the matrix are independent draws from a Beta$(0.5, 0.5)$ distribution.

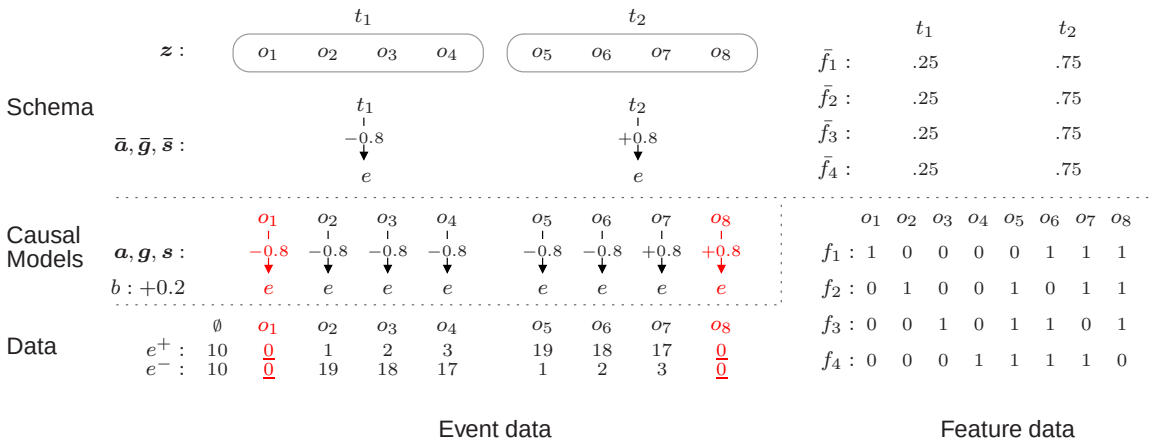

Figure 5-14: Learning a schema and a set of causal models given event and feature data (see Figure 5-1c). Objects belonging to the same type have similar causal powers and similar features, and $\bar{f}_i$ specifies the expected value of feature f_i within each type. The event and feature data shown are similar to the data used for Experiment 3.

the value of f_j for object o_i is generated by tossing a coin with bias $\bar{f}_j(z_i)$. Our goal is now to use the features F along with the events V to learn a schema and a set of causal models:

$$p(\boldsymbol{z}, \bar{\Psi}, \Psi | V, F) \propto P(V|\Psi)P(F|\bar{\Psi}, \boldsymbol{z})p(\Psi|\bar{\Psi}, \boldsymbol{z})p(\bar{\Psi}|\boldsymbol{z})P(\boldsymbol{z}). \tag{5.6}$$

There are many previous models for discovering categories of objects with similar features (Anderson, 1991), and feature-based categorization is sometimes pitted against causal categorization (Gopnik & Sobel, 2000). The schema-learning model is based on the idea that real-world categories are often distinguished both by their characteristic features and their characteristic causal interactions. More often than not, one kind of information will support the categories indicated by the other, but there will also be cases where the causal data and the feature data conflict. In cases like this, the schema-learning model may discover the feature-based categories, the causal categories, or some combination of both—the categories preferred will depend on the relative weights of the statistical information present in the two kinds of data.

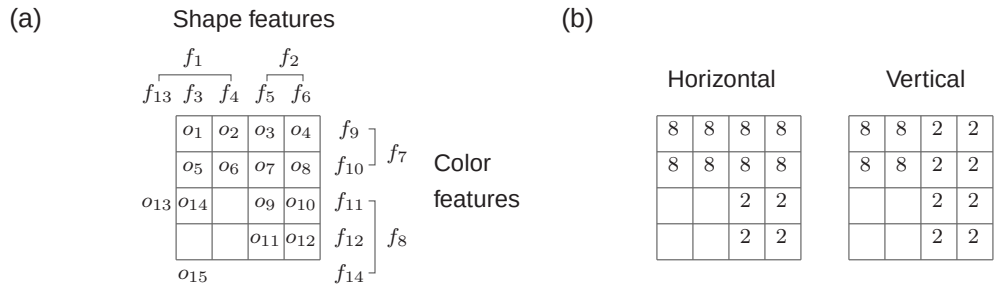

Figure 5-15: (a) Features used in our simplified version of the Lien and Cheng (2000) task. Feature f_7 is shared by all and only the first eight objects, and f_9 is shared only by the first four objects. (b) Causal data for two conditions. 10 trials were shown for each of the first 12 objects. In the horizontal condition, each object with feature f_7 produces the effect on 8 out of 10 trials. In the vertical condition, only objects with f_1 regularly produce the effect.

Preference for f_1-match

	condition	o_{13}	o_{14}	o_{15}
Human	horizontal	6	24	43
	vertical	73	90	98
Model	horizontal	4	5	7
	vertical	83	86	88

Figure 5-16: Predictions for the sorting task of Lien and Cheng (2000). The first two rows show the percentage of participants who grouped a novel object (o_{13}, o_{14} or o_{15}) with the f_1-match (o_1) rather than the f_8-match (o_{10}). Only participants in the vertical condition tend to sort according to f_1. The model predictions represent the relative probability that each novel object belongs to the same causal type as the f_1-match.

Lien and Cheng data

Lien and Cheng (2000) ran several experiments that explore how perceptual features and causal observations can both inform causal judgments. The schema-learning model appears to handle all of their tasks, but here I will focus on a simplified version of their first task. The effect of interest is whether a certain kind of plant blooms, and the potential causes are 15 chemicals (objects o_1 through o_{15}). Figure 5-15a shows that the features of these objects (f_1 through f_{14}) support two systems of categorization. The first system is based on color: each object has a cool color (f_7)

or a warm color (f_8), and the warm-colored objects are either yellow (f_{11}), red (f_{12}) or orange (f_{14}). The second system is based on shape: each object has an irregular shape (f_1) or a regular shape (f_2) and there are three kinds of irregular shapes (f_{13}, f_3 and f_4).

The schema-learning model is shown 10 trials for each of the first 12 objects, and Figure 5-15b summarizes the results of these trials. In the *horizontal* condition, each object with a cool color (f_7) causes blooming on 8 out of 10 occasions, and the remaining objects lead to blooming less often. In the *vertical* condition, objects with irregular shapes (f_1) are the only ones that tend to cause blooming. In both conditions, the model is shown that blooming occurs on 2 out of 10 trials when no chemicals are applied.

The model can be tested by requiring it to reason about three objects (o_{13}, o_{14} and o_{15}) for which no trials were observed. Object o_{13} has a novel shape, o_{15} has a novel color, and o_{14} is a novel combination of a known shape and known color. Each novel object was presented as part of a trio that also included o_1 and o_{10}, and we computed whether the model preferred to group each novel object with the shape match (o_1) or the color match (o_{10}).[7] In the horizontal condition, the model prefers to sort each trio according to color (f_8), but in the vertical condition the model sorts each trio according to shape (f_1) (see Figure 5-16). Note that the feature data and the causal data must be combined to produce this result: a model that relied on the features alone would predict no difference between the two conditions, and a model that used only the causal data would be unable to make useful predictions about the three novel objects.

Since we modeled a simplified version of the Lien and Cheng task, the quantitative predictions of the schema-learning model are not directly comparable to their results, but Figure 5-16 shows that the model captures the main qualitative patterns in their data.[8] Note first that corresponding pairs of numbers fall on the same side of 50%:

[7] We implemented this sorting task by computing the posterior distribution $p(\boldsymbol{z}|V, F)$, and comparing the probability that the novel object and its color match belong to the same causal type with the probability that the novel object is grouped with the shape match.

[8] Lien and Cheng report that a handful of participants did not group the novel objects with either the shape match or the color match. These participants were dropped before computing the

in other words, the model prefers to sort according to shape (f_1) only in the cases where people show the same preference. Note also that the numbers for both people and the model increase from left to right. For instance, out of the three novel objects, participants and the model are most confident that o_{15} belongs with the f_1-match. This result makes sense since o_{15} is the only novel object with features that are more similar to the f_1-match (o_1) than the f_8-match (o_{10}).

Although the schema-learning model accounts well for the Lien and Cheng data, their task suggests an extension of our approach that is worth exploring. The schema-learning model assumes that all features are weighted equally, and tends to prefer sets of categories that account at least partially for all of the features. There are situations, however, where some features correlate with the causal types but others should be treated as noise. Each condition of the Lien and Cheng task is one of these situations: in the horizontal condition, the shape features are uninformative, and in the vertical condition, the color features are uninformative. To better capture cases like these we can define a model that learns and uses weights for each feature. There is a chicken-or-egg problem here: features with high weights should correlate well with the causal types, and causal types are determined in part by the features with high weights. Bayesian methods, however, can deal with this problem, and we can define a model that simultaneously learns a set of causal types and an appropriate set of feature weights.[9]

Neither of the baseline models described earlier can account for the data collected by Lien and Cheng. The revolutionary model has no basis for making predictions about the causal powers of the novel objects, (o_{13}, o_{14} and o_{15}) since no trials have been observed for any of these objects. The reactionary model can be extended by defining weights w_i (Equation 5.5) such that w_i is high if the empirical power of a novel object is close to the empirical power of object o_i and if these two objects have similar features. This model, however, is also unable to account for the data. In the vertical condition, object o_{13} has features that are more similar to the color match

percentages in Figure 5-16.

[9]See, for instance, work on Bayesian feature selection (E. I. George & McCulloch, 1993) and automatic relevance determination (R. M. Neal, 1996).

(o_{10}) than the shape match (o_1), yet people prefer to group o_{13} with the shape match rather than the color match.

Lien and Cheng describe an alternative approach that does account for their data. As we have seen, their experiment uses stimuli that can be organized into one or more hierarchies, and where there are perceptual features that pick out each level in each hierarchy. Each perceptual feature is assumed to be a potential cause of blooming, and the *probabilistic contrast* for each cause c with respect to effect e is $P(e|c) - P(e|\bar{c})$. Lien and Cheng suggest that the best explanation of the effect is the cause with maximum probabilistic contrast. The theoretical problem addressed by this principle of maximum contrast is somewhat different from the problem of discovering causal types. Lien and Cheng appear to assume that a learner already knows about several overlapping causal types, where each type corresponds to a subtree of one of the hierarchies. They do not discuss how these types might be discovered in the first place, but they provide a method for identifying the type that best explains a novel causal relation. We have focused on a different problem: the schema-learning model does not assume that any causal types are known in advance, but shows how a single set of non-overlapping types can be discovered.

The schema-learning model goes beyond the Lien and Cheng approach in at least one important respect. Our model handles cases like Figure 5-14 where the features provide a noisy indication of the underlying causal types, but the Lien and Cheng approach can only handle causal types that correlate perfectly with an observable feature. Although observable features are often a good guide to category membership, many categories appear not to have defining features, and if defining features do exist, they may not be easily observable. Two metal bars, for instance, may appear identical on the surface, even though only one has micro-properties that make it a magnet. Our first two experiments have already suggested that people discover causal types in the absence of defining perceptual features. To further explore this claim, we developed a task where the features of a set of objects correlate roughly with the underlying causal types, but where there is no single feature that perfectly distinguishes these types.

	$\emptyset$	o^-	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o^+
$\mathbf{e}^+$:	10	0	3	2	1	2	18	18	17	19	0
$\mathbf{e}^-$:	10	0	17	18	19	18	2	2	3	1	0
f_1 :		1	0	0	0	0	0	1	1	1	1
f_2 :		0	1	0	0	0	1	0	1	1	1
f_3 :		0	0	1	0	0	1	1	0	1	1
f_4 :		0	0	0	1	0	1	1	1	0	1
f_5 :		0	0	0	0	1	1	1	1	1	0

Figure 5-17: Training data for Experiment 3.

Experiment 3: Combining causal and feature data

Participants

24 members of the MIT community were paid for participating in this experiment. Experiment 3 was run immediately after Experiment 1, and the same participants completed both tasks.

Materials and Methods

Participants are initially shown an empty machine that activates on 10 out of 20 trials. 10 blocks then appear on screen, and the features of these blocks support two family resemblance categories (see Figures 5-5 and 5-17). Before any of the blocks are placed in the machine, participants are informed that the blocks are laid out randomly, and are encouraged to drag them around and organize them in a way that will help them predict what effect they will have on the machine. Participants then observe 20 trials for blocks o_1 through o_8, and see that blocks o_1 through o_4 activate the machine rarely, but blocks o_5 through o_8 activate the machine most of the time. After 20 trials for each block, participants respond to the same question used in Experiment 1: they imagine 100 trials involving the block, and indicate how likely it is that the total number of activations will fall into each of 5 intervals. After this training phase, participants answer the same question for test blocks o^- and o^+ without seeing *any* trials involving these blocks. Experiment 1 explored one-shot learning, and this new task might be described as zero-shot learning. After making predictions for the two

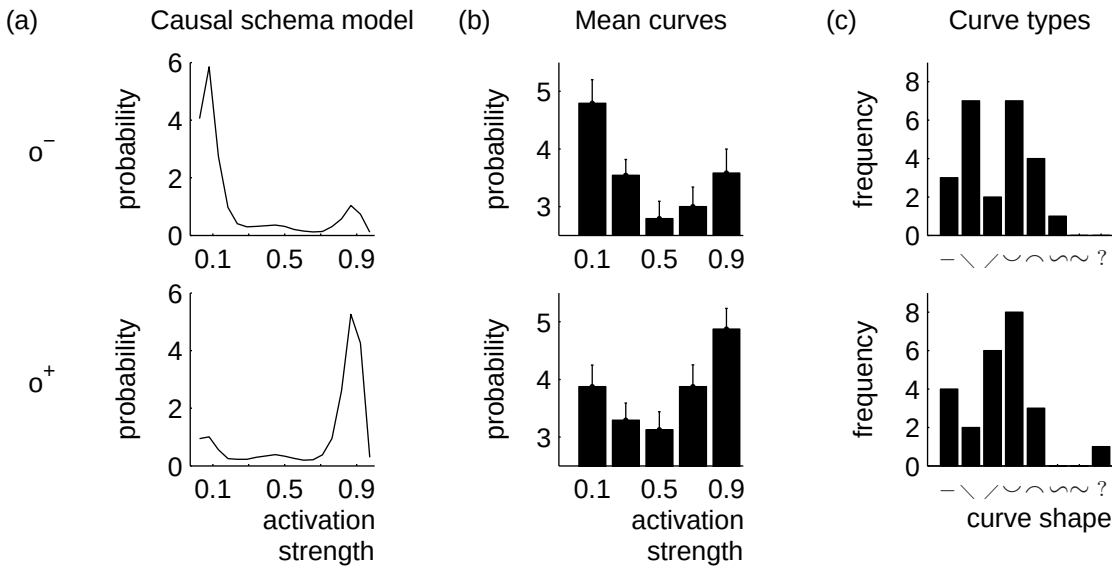

Figure 5-18: Results for Experiment 3. (a) Predictions of the schema-learning model. (b) Mean responses across 24 participants. (c) Ratings for individual participants classified by curve shape.

test blocks, participants are asked to sort the blocks into two categories "according to their effect on the machine," and to explain the categories they chose.

Model predictions

Predictions of the schema-learning model are shown in Figure 5-18a. Each plot shows the posterior probability that a test block will activate the machine on any given trial.[10] Both plots have two peaks, indicating that the model has discovered two causal types, but is not certain about the type assignments of the test blocks. The plots are skewed in opposite directions: based on the features of the test blocks, the model predicts that o^- will activate the machine rarely, and that o^+ will activate the machine often.

Predictions about the sorting task are summarized in Figure 5-19a. The top three sorts are included, and the most probable solution according to the model is the family resemblance sort. Although the model allows sorts with any number of

[10]Unlike Experiment 1, the background rate is non-zero, and these posterior distributions are not equivalent to distributions on the causal power of a test block.

categories (including one, three or more), the probabilities shown in Figure 5-18a are calculated with respect to the class of all two-category solutions.

Results

Mean responses for the two test blocks are shown in Figure 5-18b. Both plots are U-shaped curves, suggesting that participants realize that some blocks activate the machine rarely and others activate the machine often, but that few blocks activate the machine half the time. As predicted, the curves are skewed in opposite directions, indicating that o^+ is considered more likely to activate the machine than o^-.[11] Responses made by individual participants are summarized in Figure 5-18b. Most participants chose U-shaped curves, but the next most popular choices are decreasing curves (for o^-) and increasing curves (for o^+).

The U-shaped curves in Figure 5-18b resolve a question left open by Experiment 1. Responses to the $\bar{s} = \{0.1, 0.9\}$ condition of the first experiment did not indicate that participants had identified two causal types, but the U-shaped curves in Figure 5-18b suggest that participants recognized two types of blocks. All of the blocks in Experiment 3 produce the effect sometimes, and the U-shaped curves suggest that participants can use probabilistic criteria to organize objects into causal types. Two differences between Experiment 3 and the second condition of Experiment 1 seem particularly important. In Experiment 3, more blocks are observed for each type (4 rather than 3), and more trials are observed for each block (20 rather than 10). Experiment 3 therefore provides more statistical evidence that there are two types of blocks.

Responses to the sorting task are summarized in Figure 5-19b. All sorts that were chosen by two or more participants are shown, and there are eight additional sorts that were chosen by one participant each. The most popular sort organizes the blocks into the two family resemblance categories, and is chosen by 8 out of 24

[11] We ran a two factor ANOVA which compared ratings for the first (0-20) and last (80-100) intervals across the two test blocks. There is no main effect of interval ($F(1, 23) = 0.056$, $p > 0.5$) or of test block ($F(1, 23) = 1.50$, $p > 0.1$), but there is a significant interaction between interval and test block ($F(1, 23) = 6.90$, $p < 0.05$).

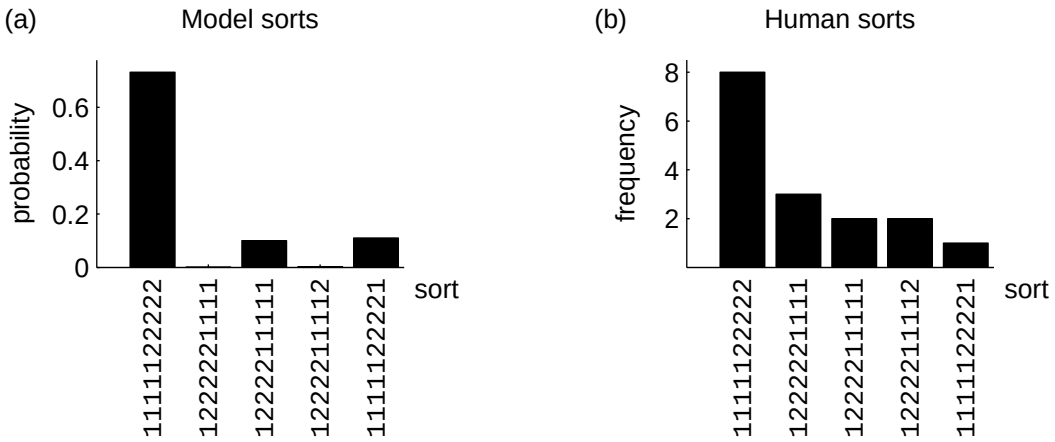

Figure 5-19: Sorts for experiment 3. (a) Relative probabilities of five sorts according to the schema-learning model. Each sort is represented as a vector that specifies category assignments for the ten objects in Figure 5-17. The model prefers the family resemblance sort. (b) Top five sorts chosen by participants. Any sort not shown was chosen by at most one participant.

participants. Studies of feature-based categorization have consistently found that family resemblance sorts are rare, and that participants prefer instead to sort objects according to a single dimension (e.g. size or color) (D. L. Medin, Wattenmaker, & Hampson, 1987). Figure 5-19b therefore suggests that the causal information provided in Experiment 3 overcomes the strong tendency to form categories based on a single perceptual dimension.

Note that the sorting task is relatively demanding, and that participants who do not organize the blocks carefully as they go along are likely to forget how many times each block activated the machine. Even though participants were asked to sort the blocks according to their effect on the machine, only 13 out of 24 assigned blocks o_1 through o_4 to one group and blocks o_5 through o_8 to the other group. Some of the remaining participants may have deliberately chosen an alternative solution, but others gave explanations suggesting that they had lost track of the training trials.

The schema-learning model accounts well for our results, but other models will make similar predictions. For instance, a feature-based version of the reactionary model predicts that o^+ is likely to activate the machine, since this test block has features similar to blocks that have previously activated the machine. The schema-

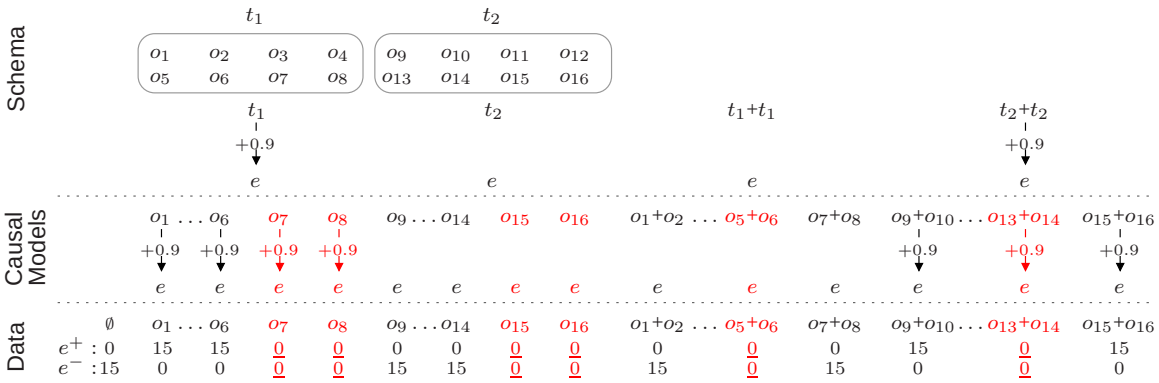

Figure 5-20: Learning about interactions between objects. The schema specifies the causal powers of each type and of each pair of types (the pair t_1+t_2 is not shown). The collection of causal models includes a model for each pair of objects. The event data are inspired by the experiment of Shanks and Darby (1998). The model groups the objects into two types: objects belonging to type t_1 cause the effect on their own but not when paired with each other, and objects belonging to type t_2 cause the effect only when paired with each other.

learning model is not alone in accounting for Experiment 3, but this experiment does rule out approaches (such as the principle of maximal contrast) which assume that causal types have defining features.

Discovering interactions between causal types

So far we have considered problems where at most one object o_i can be present at a time. Suppose now that multiple objects can be present on any trial. For instance, consider the problem of discovering which drugs produce a certain allergy—two drugs which are innocuous on their own may produce the allergy when combined. Our goal is to discover a schema and a set of causal models that allow us to predict whether any given combination of drugs is likely to produce an allergic reaction. Formally, we would like to learn a causal model M for each possible combination of objects.

We assume that each combination of objects corresponds to a conjunctive cause that may be generative or preventive, and extend Ψ to include an arrow a, a polarity g and a strength s for each combination of objects. We extend the schema in a similar fashion, and include schema parameters $\bar{a}$, $\bar{g}$, $\bar{s}$ and $\bar{\sigma}$ for each combination

of causal types. The causal model parameters for sets of objects are generated, as before, from the schema parameters for the corresponding set of types. For instance, Figure 5-20 shows how the causal model for o_{13}+o_{14} is generated from the schema-level knowledge that pairs of objects drawn from type t_2 tend, in combination, to generate the effect with strength 0.9. As before, we assume that a generative background cause of strength b is always present.

There are several possible strategies for handling conjunctive causes and our approach makes several simplifying assumptions. For instance, we assume that the causal power of a conjunction of objects is independent of the causal powers that correspond to any subset of these objects. Future work can aim to relax these simplifying assumptions, and to combine the schema-learning model with a sophisticated approach to conjunctive causality. As an initial step, it should be relatively straightforward to combine our framework with the model of conjunctive causality developed by Novick and Cheng (2004).

Shanks and Darby data

Shanks and Darby (1998) ran an experiment which suggests that humans can acquire abstract knowledge about interactions between causal types. These authors used a task where the potential causes were foods, and the effect of interest was an allergic reaction. The data observed by participants in their second experiment are shown in Figure 5-20.[12] When supplied with these data, the schema-learning model discovers two causal types: foods of type t_1 (o_1 through o_8) produce the allergy on their own, but foods of type t_2 (o_9 through o_{16}) do not. The model also discovers that two foods of type t_2 will produce the allergy when eaten together, but two foods of type t_1 will not (Figure 5-20).

Shanks and Darby were primarily interested in predictions about cases which had never been observed in training—the cases underlined in the bottom section of Figure 5-20. Their participants can be divided into two groups according to their scores when tested on the training data. Learners in the high group (learners who

[12]Different participants saw different amounts of training data, but we overlook this detail.

scored well on the test) tended to make the same predictions as our model: for instance, they tended to predict that o_7 and o_8 produce the allergy when eaten in isolation, that o_{15} and o_{16} do not, that the combination of o_{13} and o_{14} produces the allergy, and that the combination of o_5 and o_6 does not. Learners in the low group tended to make the opposite predictions: for instance, they tended to predict that o_7 and o_8 do not produce the allergy when eaten in isolation. Since the schema-learning model does not suffer from memory limitations or lapses of attention, it is not surprising that it accounts only for the predictions of learners who absorbed the information provided during training.

Conclusion

This chapter described a hierarchical Bayesian model (Figure 5-1c) for learning causal schemata. Each schema organizes a set of objects into causal types, and specifies the causal powers and characteristic features of each type. The schema-learning model supports several kinds of inferences. We focused on bottom-up inferences and saw that the model helps to explain how a causal schema and a set of specific causal models can be simultaneously learned given event data and feature data. If the causal schema is known in advance, then the model serves as a computational theory of top-down causal inference, and explains how inferences about a set of causal models can simultaneously draw on low-level event data and high-level knowledge.

The schema-learning model exploits the fact that probabilistic approaches are modular and can be composed to build integrated models of inductive reasoning. The model in Figure 5-1c can be created by combining three models: probabilistic causal models (Pearl, 2000) specify how the event data are generated given a set of causal models, the infinite relational model (Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006) specifies how the causal models are generated, and Anderson's rational approach to categorization (Anderson, 1991) specifies how the features are generated. Since all three models work with probabilities it is straightforward to combine them and create a single integrated framework for causal reasoning.

We saw that the schema-learning model helps to explain some aspects of the data collected by Lien and Cheng (2000) and Shanks and Darby (1998), and it also accounts for several other results in the literature. Waldmann and Hagmayer (2006) showed that a known set of categories can influence future causal learning, and the model predicts a similar result if we fix the causal types $\boldsymbol{z}$ then use the model to discover a set of causal models given event data. Our approach can also model experiments carried out using the blicket detector (Gopnik & Sobel, 2000) or causal blocks world (Tenenbaum & Niyogi, 2003) paradigms. Many aspects of these experiments have been previously modeled, but the schema-learning model captures phenomena that are not addressed by most existing models. For instance, the model suggests why two identical looking blocks might both be categorized as blickets even though a handful of observations suggest that they have different effects on a blicket detector (Gopnik & Sobel, 2000).

Our experiments used adult participants but some of the most fundamental causal schemata are probably acquired relatively early in development. Several studies confirm that young children know that objects of certain types tend to have certain kinds of effects: for instance, children know that flashlights generate spots of light (Shultz, 1982), and that hammers can break brittle objects (R. Gelman, Bullock, & Meck, 1980). It is less clear how this knowledge emerges, but experiments similar to ours might be able to trace the developmental course of schema acquisition. Experiments using blicket detectors (Gopnik & Sobel, 2000) have been used to study the causal knowledge of preschoolers, and the tasks I described rely on a very similar paradigm.

Several extensions of the schema-learning model may be worth exploring. We restricted ourselves to problems where the distinction between a set of potential causes and a set of effects[13] is known in advance, but in some cases this distinction may need to be learned (Mansinghka, Kemp, Tenenbaum, & Griffiths, 2006). A second limitation is that we focused on cases where feature data and contingency data represent the only input to our model. Human learners are sometimes directly supplied

[13]This chapter has focused on problems where there is a single effect, but our approach also handles problems with multiple effects, and can group these effects into types.

with abstract causal knowledge—for example, a science student might be told that "pineapple juice is an acid, and acids turn litmus paper red." Statements like these correspond to fragments of a causal schema, and future experiments should explore how schemata are learned when parts of these schemata are directly supplied.

Causal inferences are guided by many kinds of constraints, and the schemata discussed in this chapter only capture some of these constraints. The schemata I considered are closely related to the *grouping schemata* described by Kelley (1972), but Kelley also discusses some other kinds of schemata. Some of these schemata specify the manner in which multiple causes interact: for instance, a schema for multiple sufficient causes indicates that any of several causes is sufficient to produce an effect, and a schema with additive effects indicates that the strength of the effect is determined by the cumulative strength of the relevant causes. Other causal constraints specify the temporal properties of causal interactions, and one basic example is a constraint which specifies that causes precede effects. Still other constraints may indicate the expected properties of interventions: for instance, Pearl's "do-calculus" can be seen as a set of constraints on causal reasoning. Although this chapter has focused on one class of causal constraints, many different constraints can be captured by hierarchical Bayesian models, and future work can explore how some of these constraints might be learned.

Top-down and bottom-up approaches to learning are sometimes seen as competitors, and the debate between these approaches is especially prominent in the literature on causal learning. More often than not, competing accounts of a phenomenon both capture some element of the truth, and situations like this can be handled by building unified models that subsume the two competing views. This chapter described a hierarchical Bayesian model that unifies top-down and bottom-up approaches to causal reasoning. The model recognizes that abstract causal knowledge is crucial for making inferences about objects that are sparsely observed, and suggests how this knowledge is acquired by bottom-up learning. Similar conflicts between top-down and bottom-up approaches are found in other areas of cognitive science, and the hierarchical Bayesian approach can help to resolve these conflicts wherever they occur.

Chapter 6

The discovery of structural form

The previous chapters described models which incorporate two kinds of representations: category structures (Chapters 4 and 5) and graph structures (Chapter 5). Neither kind of representation is especially complex, but these models do demonstrate that the hierarchical Bayesian approach can handle structured representations (Table 1.2b). One reason why structured representations matter is that they capture inductive constraints which guide inferences about different domains. Grammars guide inferences about sentences (Chomsky, 1965), folk taxonomies guide inferences about living kinds (Atran, 1998), and causal graphical models (Pearl, 2000) guide inferences about the causal consequences of actions. Since different domains call for different representations, we are faced with the problem of *form discovery*: how can a learner discover which form of representation is best for a given domain? This chapter describes a hierarchical Bayesian model that addresses a special case of this problem.

Some of the clearest examples of form discovery come from the history of science. For centuries, Europeans believed that the natural representation for biological species was the "great chain of being," a linear structure in which every living thing found a place according to its degree of perfection (Lovejoy, 1970). In 1735, Linnaeus famously proposed that relationships between plant and animal species are best captured by a

The work in this chapter was carried out in collaboration with Joshua Tenenbaum. A very preliminary version of this work was presented at the 26th Annual Conference of the Cognitive Science Society in 2004.

tree structure (Figure 6-1b), setting the agenda for all biological classification since. Mendeleev made a similar breakthrough when he recognized the periodic structure of the chemical elements and proposed a specific representation with this form—his periodic table of 1869.

Scientific breakthroughs like these occur relatively rarely, but children may make analogous discoveries when learning about the structure of different domains. Children may learn, for example, that social networks are often organized into cliques, that temporal categories such as the seasons or the days of the week can be arranged into cycles, that comparative relations such as "longer than" or "better than" are transitive (Piaget, 1965; Shultz, 2003), and that category labels can be organized into hierarchies (Rosch, 1978). Structural forms for some cognitive domains may be known innately, but many appear to be genuine discoveries. When learning the meanings of words, children initially seem to organize objects into non-overlapping clusters, with one category label allowed per cluster (Markman, 1989): hierarchies of category labels are recognized only later (Rosch, 1978). When reasoning about comparative relations, children's inferences respect a transitive ordering by the age of seven but not before (Shultz & Vogel, 2004). In both of these cases, structural forms appear to be learned, but children are not explicitly taught to organize these domains into hierarchies or dimensional orders.

A learner who discovers the structural form of a domain has acquired a powerful set of inductive constraints. The story of Mendeleev includes a compelling example of the inductive leverage that structural forms can provide. Mendeleev used his periodic table to predict both the existence and the properties of several undiscovered elements, and to demonstrate that some of the atomic weights he had been using were inaccurate. Children make inferences that are analogous, if somewhat less dramatic. Discovering the clique structure of social networks can allow a child to predict the outcome of interactions between individuals who may never have interacted previously. Discovering the hierarchical structure of category labels allows a child to predict that a creature called a "chihuahua" might also be a dog and an animal, but cannot be both a dog and a cat.

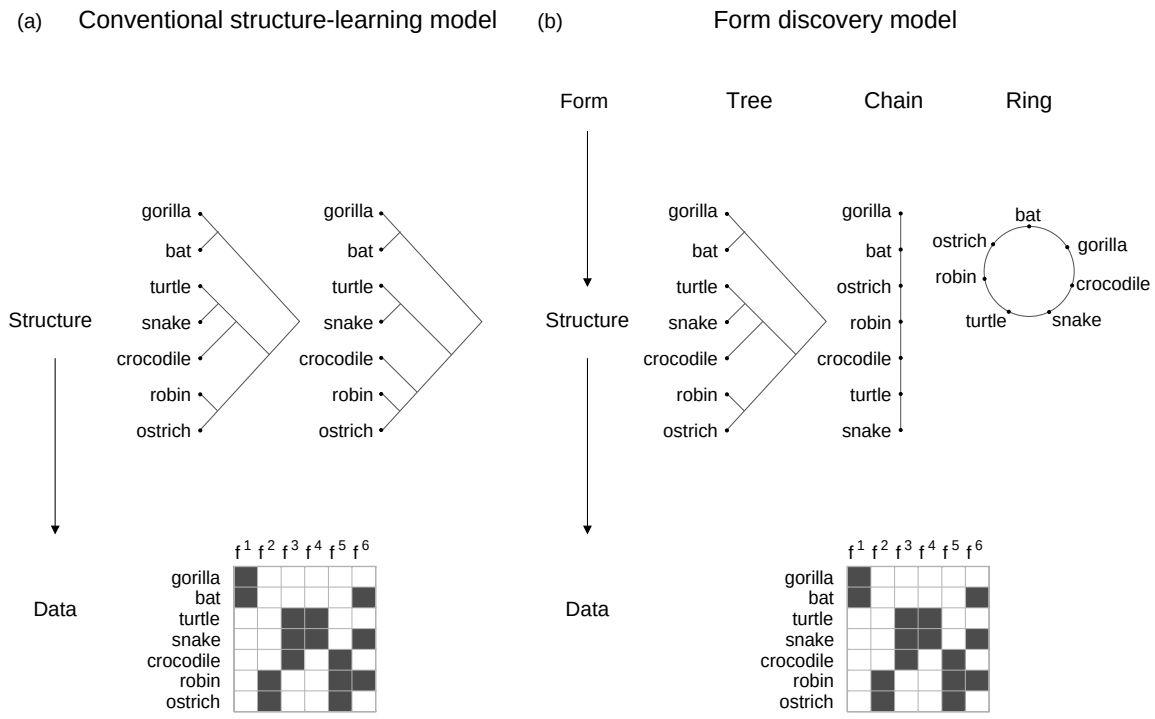

Figure 6-1: Discovering the structure that best accounts for a matrix of binary features. (a) A conventional model (cf. Figure 2-3a) which assumes that the form of the structure is fixed in advance (here assumed to be a tree). Two possible trees are shown: traditional taxonomies group crocodiles with lizards, snakes and turtles, but contemporary phylogenies assert that crocodiles are better grouped with birds (Purves et al., 2001). (b) A hierarchical model (cf. Figure 2-3b) that simultaneously discovers the form and the structure that best account for the data. Three possible pairs of forms and structures are shown. The tree is inspired by the Linnaean taxonomy, and the chain is inspired by Bonnet's version of the "great chain of being" (C. White, 2001). A ring structure might not seem suitable for the species shown here, but has recently been proposed as the best model of relationships between microbes (Rivera & Lake, 2004).

This chapter argues that the hierarchical Bayesian approach helps to explain how humans discover the best kind of representation for a domain. The problem of form discovery is not addressed by conventional models of learning, which search only for structures of a single form that is assumed to be known in advance (Figure 6-1a). Clustering or competitive-learning algorithms (Anderson, 1991; D. Rumelhart & McClelland, 1986) assume that the data fall into some number of disjoint groups, algorithms for hierarchical clustering (Duda, Hart, & Stork, 2000) or phylogenetic reconstruction (Huelsenbeck & Ronquist, 2001) assume that the data are tree-structured, and algorithms for dimensionality reduction (Pearson, 1901; Spearman, 1904) or multidimensional scaling (Torgeson, 1965) assume that the data have an underlying spatial geometry. Unlike these algorithms, our model simultaneously discovers the structural form and the instance of that form that best explain the data (Figure 6-1b). Our approach can handle many kinds of data, including attributes, relations, and measures of similarity, and I will show that it successfully discovers the structural forms of a diverse set of real-world domains.

Here we make the simplifying assumption that there is a single best representation for each data set that we consider. Often, however, a single domain will have several useful representations (Moray, 1990; Heit & Rubinstein, 1994; Shafto et al., 2006). For instance, a taxonomic tree might capture the anatomical relationships between a set of animals, but a set of ecological categories (including land animals, sea animals, predators and prey) might be a better representation of the ecological relationships between the animals. The problem of learning multiple representations for a given data set can be approached in several ways (see Shafto et al. (2006) for one example), and some of these approaches can be incorporated into future models of form discovery.

A hypothesis space of structural forms

Any algorithm for form discovery must specify the space of structural forms it is able to discover. We represent structures using graphs, and use graph grammars (En-

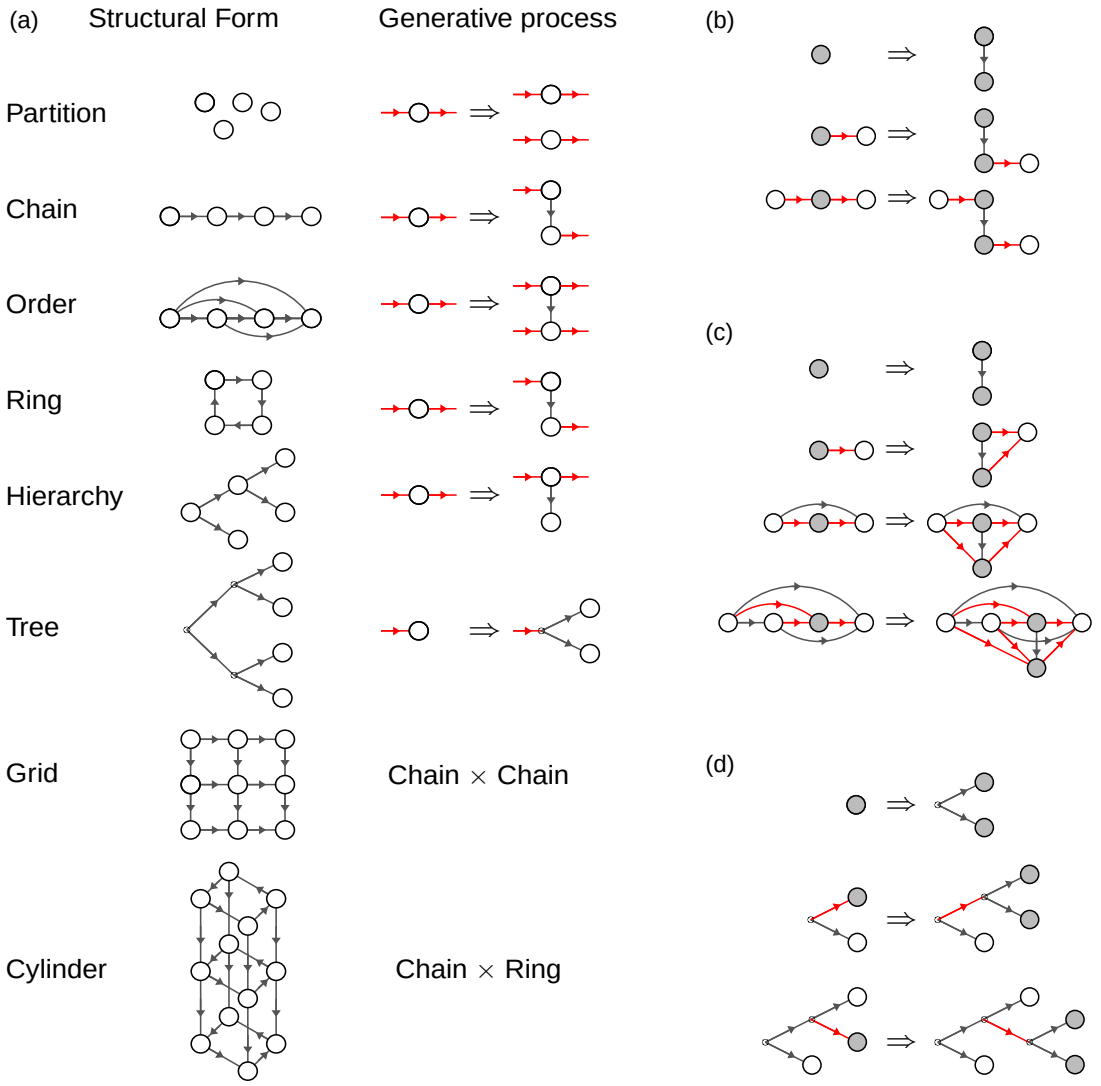

Figure 6-2: A hypothesis space of structural forms. (a) Eight structural forms and the generative processes that produce them. Open nodes represent clusters of objects: a hierarchy has clusters located internally, but a tree may only have clusters at its leaves. The first six processes are node-replacement graph grammars. Each grammar uses a single production, and each production specifies how to replace a parent node with two child nodes. The seed for each grammar is a graph with a single node (in the case of the ring, this node has a self-link). (b)(c)(d) Growing chains, orders and trees. At each step in each derivation, the parent and child nodes are shown in grey. The red arrows in each production represent *all* edges that enter or leave a parent node. When applying the order production, all nodes that previously sent a link to the parent node now send links to both children.

gelfriet & Rozenberg, 1997) as a unifying language for expressing a wide range of structural forms (Figure 6-2). Of the many possible forms, we assume that the most natural are those that can be derived from simple generative processes (Leyton, 1992). Each of the first six forms in Figure 6-2a can be generated using a single context-free production that replaces a parent node with two child nodes, and specifies how to connect the children to each other and to the neighbors of the parent node. Figures 6-2b, 6-2c and 6-2d show how three of these productions generate chains, orders and trees. In each case, we grow a representation by starting with a seed graph and repeatedly splitting nodes according to the grammar. For all forms except the ring, the seed is a graph with one node and no edges. For the ring, the seed is a single-node graph with a self link. The remaining forms in Figure 6-2—the grid and the cylinder—can be expressed as products of simpler forms. A grid is the Cartesian graph product of two chains, and a cylinder is the product of a ring and a chain.[1] We grow grids by representing the two dimensions separately, and using the chain grammar to grow each dimension. Cylinders are generated similarly.

It is striking that the simple grammars in Figure 6-2a generate many of the structural forms discussed by psychologists (Shepard, 1980) and assumed by algorithms for unsupervised learning or exploratory data analysis. Partitions (Anderson, 1991; Fiske, 1992), chains (Guttman, 1944), orders (Fiske, 1992; Inhelder & Piaget, 1964; Bradley & Terry, 1952), rings (Guttman, 1954; Wiggins, 1996), trees (Inhelder & Piaget, 1964; Sneath & Sokal, 1973; Huelsenbeck & Ronquist, 2001), hierarchies (Collins & Quillian, 1969; Carroll, 1976) and grids (Kohonen, 1997) appear again and again in formal models across many different literatures. To highlight just one example, Inhelder and Piaget (1964) suggest that the elementary logical operations in children's thinking are founded on two forms: a classification structure that can be modeled as a tree, and a seriation structure that can be modeled as an order. The popularity of the forms in Figure 6-2 suggests that they are useful for describing the world, and that

[1] A two dimensional Euclidean space can be generated as the regular Cartesian product of two chains, where each chain is viewed as a continuous dimension rather than a graph. Our generative model for feature data extends naturally to continuous spaces, but here we consider only graph structures.

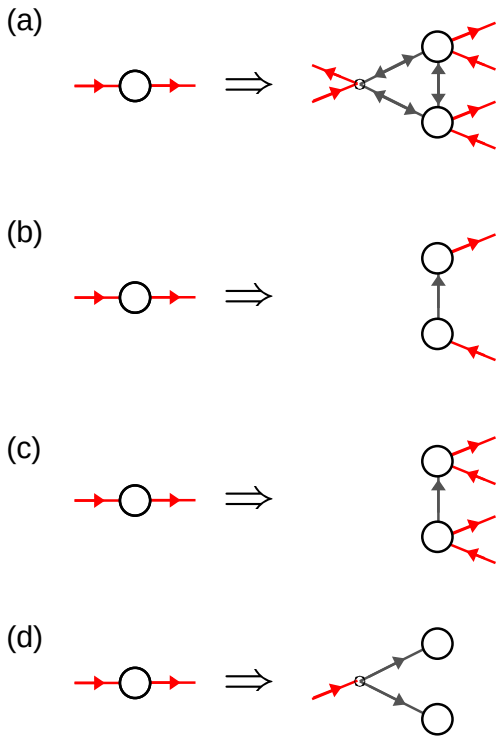

Figure 6-3: Generating graph grammars from a meta-grammar. (a) A meta-grammar: the six grammars in Figure 6-2 correspond to subsets of the production shown here. (b)(c)(d) Subsets of the meta-grammar that grow chains, orders and trees.

they spring to mind naturally when scientists seek formal descriptions of a domain.

Although we focus on the eight forms in Figure 6-2, it is natural to consider other possibilities. I have suggested that graph grammars provide a unifying language for expressing many different structural forms, and ultimately it may be possible to develop a "universal structure grammar" (cf. Chomsky (1965)) that generates all and only the cognitively natural forms. As an initial step towards this goal, it is useful to recognize that all of the grammars in Figure 6-2 can be generated as subsets of the meta-grammar in Figure 6-3. This meta-grammar generates grammars for many other structural forms, some of which (although certainly not all) are likely to be useful for structure discovery. In principle, a learning system could begin with just this meta-grammar and go on to discover any form that is consistent with the meta-grammar.

Each of the grammars we consider uses a single production, but additional forms

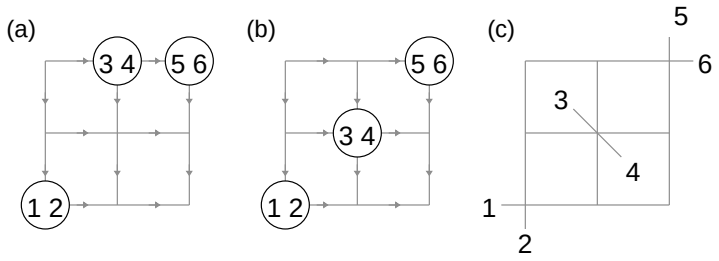

Figure 6-4: Cluster graphs and entity graphs. (a) A cluster graph that is incompatible with the grid form, since the middle node will be empty if the graph is projected onto the vertical axis. (b) A cluster graph that is compatible with the grid form. (c) An entity graph corresponding to the cluster graph in (b).

can be generated if we allow grammars with multiple productions, and productions where the edges on the right hand side are chosen probabilistically. This chapter will focus on simple grammars that generate some of the most frequently used forms, but further exploration of the space of grammars is an important direction for future work.

Now that we have a hypothesis space of structural forms, the problem of form discovery can be posed. Given a data set D that contains information about a set of entities, we wish to find the form F and the structure S of that form that best capture the relationships between these entities. We approach this problem by defining a hierarchical Bayesian model (Figure 6-1) and searching for the structure S and form F that maximize the posterior probability

$$P(S, F|D) \propto P(D|S)P(S|F)P(F). \tag{6.1}$$

To complete the model we must formally specify the terms on the right hand side of Equation 6.1. $P(F)$ is a uniform distribution over the forms under consideration, and the remaining two terms are described in the next sections.

Generating structures from structural forms

Suppose that we are working with a set of n entities.[2] Let S be a *cluster graph*, or a graph where the nodes correspond to clusters of entities. S is compatible with F if S can be generated by the generative process defined for F, and if S contains no empty nodes when projected along any of its component dimensions (Figure 6-4).[3] There is a finite collection of structures that are compatible with a given form F, and $P(S|F)$ is non-zero only for graphs in this collection. To encourage the model to choose the simplest adequate representation for a domain, we weight each structure according to the number of nodes it contains:

$$P(S|F) \propto \begin{cases} 0 & S \text{ is incompatible with } F \\ \theta(1-\theta)^{|S|} & \text{otherwise} \end{cases} \tag{6.2}$$

where $|S|$ is the number of nodes in S.[4]

The parameter θ determines the extent to which graphs with many clusters are penalized, and is fixed for all of our experiments. We set $\theta = 1-e^{-3}$, which means that each additional node reduces the log probability of a structure by 3. The normalizing constant for $P(S|F)$ depends on the number of structures compatible with a given form, and ensures that simpler forms are preferred whenever possible. For example, any chain S_{chain} is a special case of a grid, but $P(S_{chain}|F_{chain}) > P(S_{chain}|F_{grid})$ since there are more possible grids than chains given a fixed number of entities. Computing the normalizing constant for $P(S|F)$ requires some simple combinatorics, and details are provided in an appendix.

[2]There are methods for learning partitions (Escobar & West, 1995) and trees (R. Neal, 2003) when the set of entities is countably infinite, and future work should consider whether these methods can be used to develop a framework for learning many kinds of forms.

[3]In the case of trees, internal nodes are required to be empty, but we do not allow empty leaf nodes.

[4]If S is a tree, since entities may only appear at its leaves, we adopt the convention that $|S|$ is equal to the number of leaf nodes in S.

Feature data

The remaining term in Equation 6.1, $P(D|S)$, measures how well the structure S accounts for the data D. Its definition depends on whether the data are feature values, similarity ratings or relations. We consider all three cases, but we assume first that D is a feature matrix where the (i, j) entry in the matrix indicates the value of entity i on feature j (see Figure 6-1).

When working with feature data, we represent the structure of a set of entities using undirected *entity graphs*. Cluster graphs are converted to entity graphs by adding a node for each entity, connecting each entity to the cluster node that contains it, and replacing each directed edge with an undirected link (Figure 6-4). We set $P(D|S) = P(D|S_{ent})$ where S_{ent} is the entity graph corresponding to cluster graph S.[5]

Given an entity graph S_{ent}, $P(D|S_{ent})$ should be high if the features in D vary smoothly over the graph—that is, if entities nearby in S_{ent} tend to have similar feature values. In Figure 6-1a, for instance, feature $\mathbf{f}^1$ is smooth over both trees, $\mathbf{f}^3$ is smoother over the left tree than the right tree, and $\mathbf{f}^6$ is smooth over neither tree. We capture the expectation of smoothness by assuming that the features are independently generated by a zero-mean Gaussian process over the graph S_{ent} (Zhu, Lafferty, & Ghahramani, 2003). Under this model, each candidate graph S_{ent} specifies how entities are expected to covary in their feature values, and the distribution $P(D|S_{ent})$ favors graphs that capture as much of this covariance as possible. A more detailed description of the model is provided in the appendix.

Now that we have fully specified a hierarchical model we can use it for several purposes. If the form of a data set is already known, we can search for the structure S that maximizes $P(S|F)$ (Figure 6-1a). If the form of the data is not known, at least two strategies might be tried. For some applications it may be desirable to integrate over the space of structures S and compare forms according to their posterior probabilities $P(F|D)$. We chose, however, to search for the structure S and

[5]Note that an order becomes a fully connected graph when directed edges are converted to undirected edges.

form F that jointly maximize $P(S, F|D)$ (Equation 6.1). Two factors motivate this approach. First, we are interested in discovering the structure S that best accounts for the data. Maintaining a posterior distribution over structures may lead to optimal predictions about unobserved features, but human learners often appear to choose just one representation for a problem. Second, even if we wanted to integrate over the space of structures, computing the integral $P(F|D) = \int P(F, S|D)P(S|D)dS$ is a difficult challenge. Future research should attempt to address this challenge, since integrating over structures may prove useful when applying the form-discovery model to machine learning problems.

Experiments

We generated synthetic data to test the form-discovery algorithm on cases where the true structure was known. Figure 6-5 shows graphs used to generate five data sets, and the structures found by fitting five different forms to the data. The final column in Figure 6-5 compares the scores for the five forms, and shows that the true form is correctly recovered in each case. Special-purpose learning algorithms already exist for several of these forms, including partitions and trees (Shepard, 1980; Duda et al., 2000). The form-discovery model subsumes many of these previous algorithms, and discovers in addition which form is best for each data set.

Animals

Next we applied the model to several real-world data sets, in each case considering all forms in Figure 6-2. The first data set is a matrix of animal species and their biological and ecological properties. It consists of human judgments about 33 species and 106 features, and amounts to a larger and noisier version of the data set shown schematically in Figure 6-1. We collected the data by asking a single subject to make binary decisions about whether 106 features applied to 60 animal species. The data include perceptual features (is black), anatomical features (has feet), ecological features (lives in the ocean) and behavioral features (makes loud noises). The data

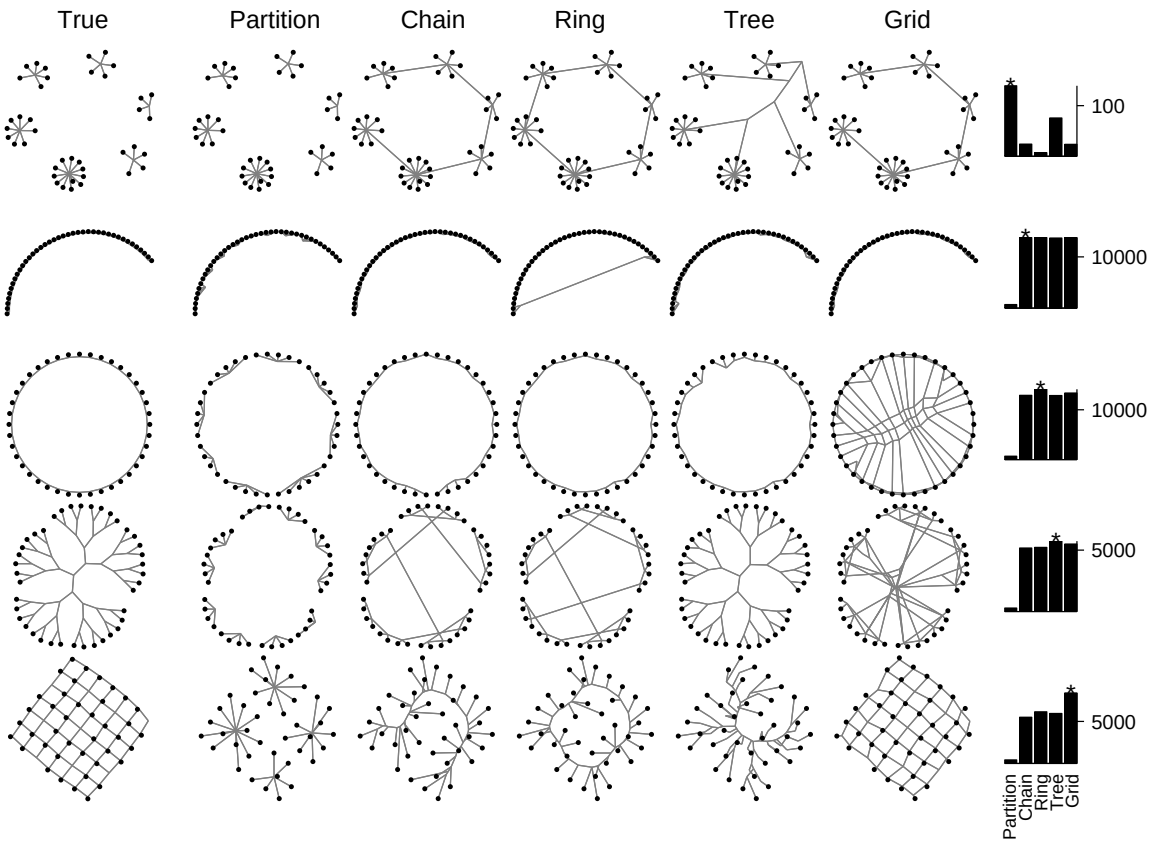

Figure 6-5: Structure discovery results for synthetic data. Five sets of features were generated over the graphs in the left column, and five forms were fit to each dataset. The structures found are drawn so that entity positions correspond to positions in the picture of the true structure. Each entity has been connected to the cluster node to which it belongs: for instance, all graphs in the top row have six clusters. The final column shows log posteriors $\log(P(S, F|D))$ for the best structures found, and the best scoring structure is marked with an asterisk. The difference between the scores for the top two structures ranges from 0.63 (indicating that the chain is about twice as likely as the grid on the chain-structured data) to 2245 (indicating that the grid is many orders of magnitude more likely than the ring on the grid-structured data). Each plot has been scaled so that the worst performing structure receives a score close to zero.

analyzed here include 33 species (the species in Figure 6-6a) that were chosen to be representative of the full set.

Given these biological data the model concludes that the best scoring form is the tree. This result is consistent with the finding that cultures all over the world appear to organize living kinds into tree-structured representations (Atran, 1998).[6] The ultimate reason why trees are useful for representing relationships between living kinds is that species were generated by a branching process—the process of evolution. The best tree found by the model (Figure 6-6a) includes subtrees that correspond to categories at several levels of resolution, including mammals, primates, rodents, birds, insects, and flying insects.

Scores for each form on the biological data set are shown in Figure 6-7. Since our search algorithm is not deterministic, these figures were generated by running the algorithm 10 times and choosing the best structure found. Note that the scores in Figure 6-7 represent log probabilities: for instance, the best tree-structured representation for the biological data is around 10 times more probable than the best hierarchy, and around 150 times more probable than the best chain. Recall that a hierarchy is a tree where entities (here animals) are located both at the leaves and at the internal nodes. Since the tree and the hierarchy can both capture branching structures, it makes sense that both forms provide a relatively good account of biological data.

Judges

The second data set is a matrix of votes from the United States Supreme Court, including 13 judges and their votes on 1596 cases. The data include all cases heard between October 1987 and June 2005.[7] This period covers all of the Rehnquist natural

[6]Given that folk taxonomies appear to be systems of nested categories, it is interesting that scientists took so many years to formalize this idea. One possible explanation is that the hierarchical structure of folk taxonomies is only implicit, and that it took a Linnaeus to make this structure explicit.

[7]The unit of analysis is the case citation (ANALU=0), and we included cases where DEC_TYPE equals 1 or 5 (Spaeth, 2005). Voting behaviors were converted to binary values: regular concurrence (3) and special concurrence (4) were converted to majority votes (1), and non-participation (5) was treated as missing data. Any case with a voting behavior other than 1 through 5 was removed from

courts except the first. Since at most 9 judges voted on any of the cases, the data include many missing entries. We assume that the unobserved entries are missing completely at random, and integrate over all possible values for these entries.[8]

Some political scientists (Grofman & Brazill, 2002) have argued that a unidimensional structure best accounts for variation in Supreme Court data and in political beliefs more generally, although other structural forms (including higher-dimensional spaces (Wilcox & Clausen, 1991) and sets of clusters (Fleishman, 1986)) are also considered. Consistent with the unidimensional hypothesis, the model identifies the chain as the best-scoring form, and the best chain (Figure 6-6b) organizes the thirteen judges from liberal (Marshall and Brennan) to conservative (Thomas and Scalia). The next best form is the hierarchy, which is not surprising since each chain is a special case of a hierarchy.

Even though our generative model for features assumes that the data are continuous, Figures 6-6a and 6-6b were learned from binary features. If possible, it would be better to analyze these data sets using a generative model for binary data. Generative models analogous to Equation 2 can be defined for binary features (Ackley et al., 1985), but structure learning becomes more difficult: in particular, computing $P(D|S)$ is challenging when S is multiply connected. Future models can attempt to address the computational challenges we have avoided by working with a Gaussian generative process.

Similarity data

If similarity is assumed to be a measure of covariance, the feature-based model can also discover structure in similarity data. Under the generative model for features, the

the analysis.

[8] In general, we cannot simply ignore the missing data when learning structural forms. If two judges never sat on the same court, there are no features observed for both of them, which encourages the model to assign them to the same node in the structure if their ideological positions are even roughly similar. (Given fully observed data, two entities will usually be assigned to the same node only if they are highly similar.) Groupings of this sort can affect the relative scores of different structural forms. We excluded the first Rehnquist court since Kennedy and Powell (who sat only on that court, and whom Kennedy replaced in 1988) tended to be assigned to the same node, and this grouping appears to be heavily influenced by the fact that these judges never served together.

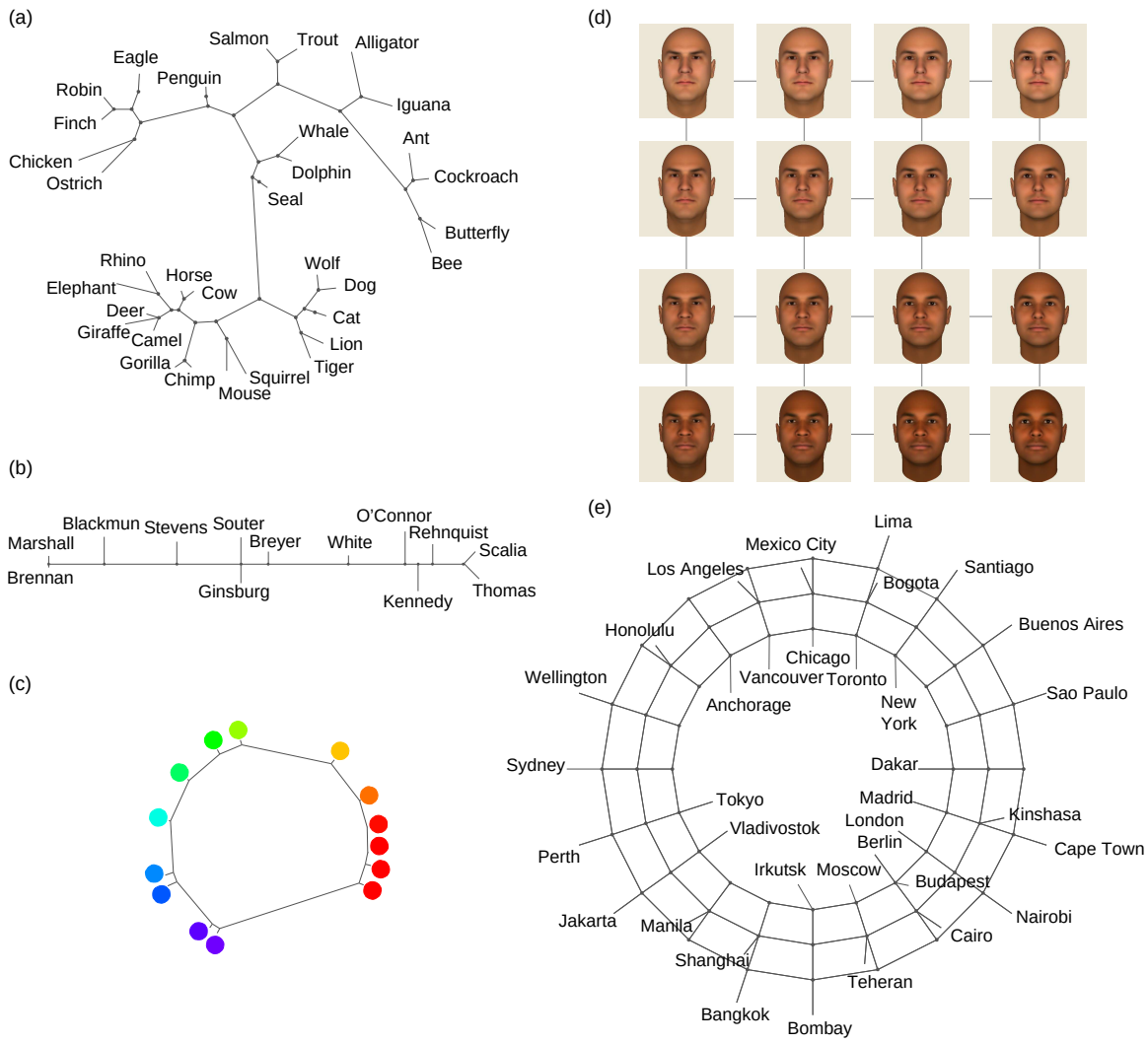

Figure 6-6: Structures learned from (a) biological features, (b) Supreme Court votes, (c) judgments of the similarity between pure color wavelengths, (d) Euclidean distances between faces represented as pixel vectors, and (e) distances between world cities. For (a)-(c), the edge lengths represent maximum *a posteriori* edge lengths under the generative model in Equations 3 and 5.

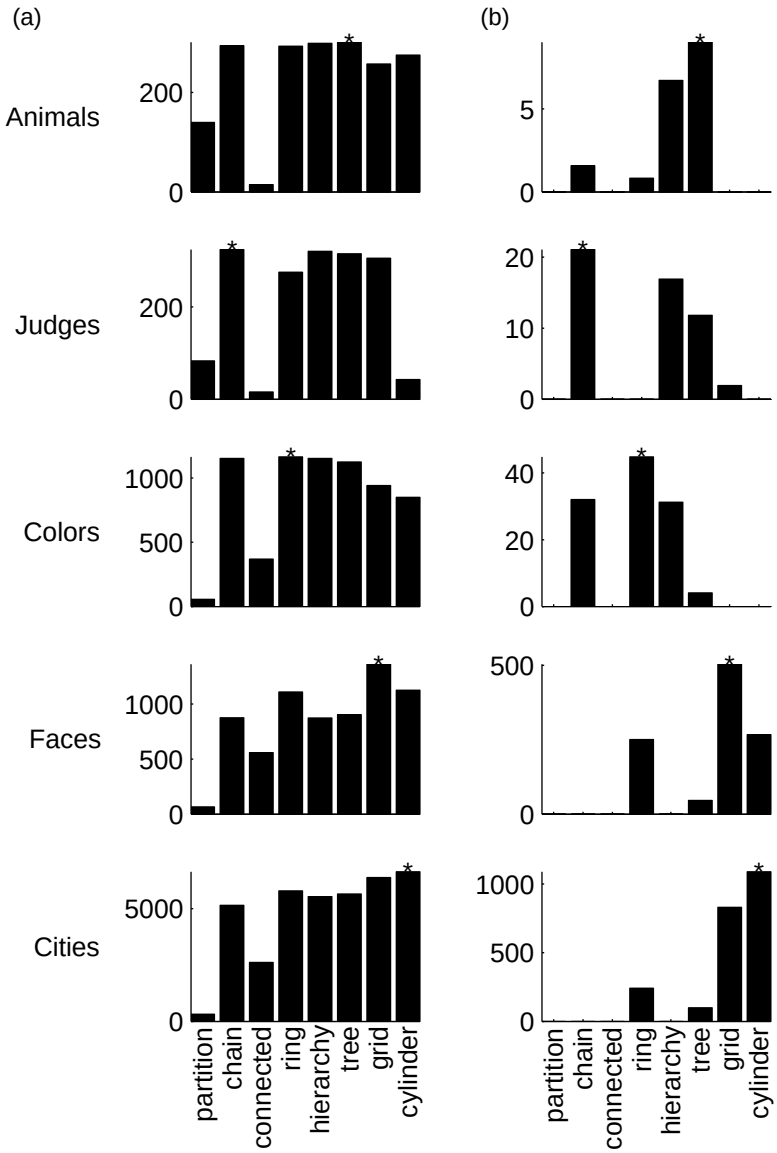

Figure 6-7: Scores for eight structural forms on feature and similarity data. (a) Each score represents $\log(P(S, F|D))$ where S is the best structure found for form F. The scores have been translated that the lowest score in each case is close to zero. Recall that a connected structure is the same as an undirected order. (b) Relative scores for the top four forms for each data set. The differences between these scores are the same as the differences in (a).

equation for $P(D|S)$ includes only two components that depend on D: the number of features observed (m), and the covariance of the data ($\frac{1}{m}DD^{\mathsf{T}}$). As long as m and the covariance matrix are provided, our approach to structure discovery can be used even if none of the actual features is observed. This insight allows us to learn structural forms from similarity data, if we assume that a given (symmetric) similarity matrix is a covariance matrix.[9] Additional details can be found in the appendix.

Experiments

Color

We applied the similarity model to a matrix containing human judgments of the similarity between all pairs of 14 pure-wavelength hues (Ekman, 1954). The ring in Figure 6-6c is the best structure for these data, and corresponds to the color circle described by Newton. Configurations similar to Figure 6-6c have been found using multidimensional scaling to locate the colors in two dimensions (Shepard, 1980), but a ring provides more appropriate constraints on inductive inference. The ring implies that other pure-wavelength hues will be located somewhere along the ring, but if a two-dimensional configuration were chosen, other hues would be incorrectly expected to fall in any region of the space.

Faces

Next we analyzed a similarity data set where the entities are faces that vary along two dimensions: masculinity and race. We created 16 faces using the FaceGen program. The program includes dimensions for race and gender, and we used four possible values along each dimension. The dissimilarity between faces was defined as the Euclidean distance between their pixel vector representations. Given these data, the model chooses a grid structure that recovers the two underlying dimensions (Figure 6-6d).

[9] In many cases the similarity matrix will already be positive definite, but if not we make it so by replacing all negative eigenvalues with zeroes.

Cities

As a final demonstration of the similarity model we analyzed a data set of distances between 35 world cities. Dissimilarity was defined as distance along the surface of the earth. Assuming that the earth is spherical, these distances can be calculated using the latitude and longitude of each city. Given these data, the model chooses a cylinder where the chain component corresponds roughly to latitude, and the ring component corresponds roughly to longitude. A spherical representation would presumably score even better than a cylinder, but note that a sphere does not currently appear in the hypothesis space of structural forms.

Relational data

The framework already described can be used to discover structure in relational data if we modify the distribution $P(D|S)$ appropriately. We define two generative models, one for frequency data and the other for binary relations. Suppose first that D is a square frequency matrix with a count d_{ij} for each pair of entities (i, j). If the entities are people, for example, d_{ij} may indicate the number of times that person i spoke to person j. We define a generative model where $P(D|S)$ is high if the large entries in D correspond to edges in the cluster graph S. A more detailed description is provided in the appendix.

A similar approach can be used to analyze binary relations. Suppose that D is a square binary matrix where d_{ij} is 1 if the relation holds between i and j and 0 otherwise. In a social setting, for instance, d_{ij} may capture whether i gives orders to j. We define a generative model where $P(D|S)$ is high if the non-zero entries in D tend to correspond to edges in the cluster graph S. Again, details can be found in the appendix.

When working with relational data, for convenience we restrict the analysis to graphs where each node represents a non-empty cluster of entities. Trees, grids and cylinders allow nodes to be empty, and we remove these from our collection of structural forms, leaving five forms in total. Given a relation it is important to discover

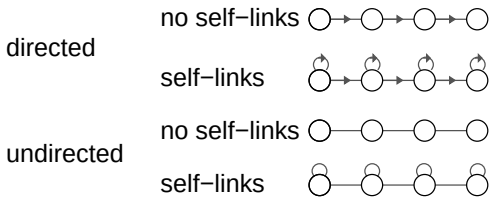

Figure 6-8: The four chain-structured forms used for relational data.

whether the relation tends to hold between elements in the same cluster, and whether the relation is directed or not. The forms in Figure 6-2 use nodes without self-links, and therefore assume that the relation does not hold within clusters. We create a set of 10 forms by supplementing each form with an alternative that uses nodes with self-links, but is otherwise identical. Each of these 10 forms uses directed edges, and for each we include an additional form with undirected edges. In total, then, the hypothesis space of relational forms includes 20 candidates.[10] The four chain-structured forms in this hypothesis space are shown in Figure 6-8.

Experiments

Mangabeys

We applied the relational model to a matrix of interactions among a troop of sooty mangabeys. The data represent interactions where one animal submitted to another. Range and Noë (2002) consider two types of submissive behavior: in the first, "the actor jumps or walks away from an approaching individual," and in the second, "the actor leans aside or shifts body position in response to another individual that approaches or walks by." These data were recoded so that a count in the (i, j) cell of the matrix indicates that i caused j to submit.

Scores for each form on this data set are shown in Figure 6-10. As expected, the model concludes that a directed order is the most appropriate form, and the two

[10]Only 17 of these forms are actually distinct. A partition (with or without self-links) remains the same when converted to an undirected graph. An undirected order with self links is a fully connected graph, and is very similar to a partition graph without self links (a graph with no edges). In both cases, all clusters stand in the same relationship to each other.

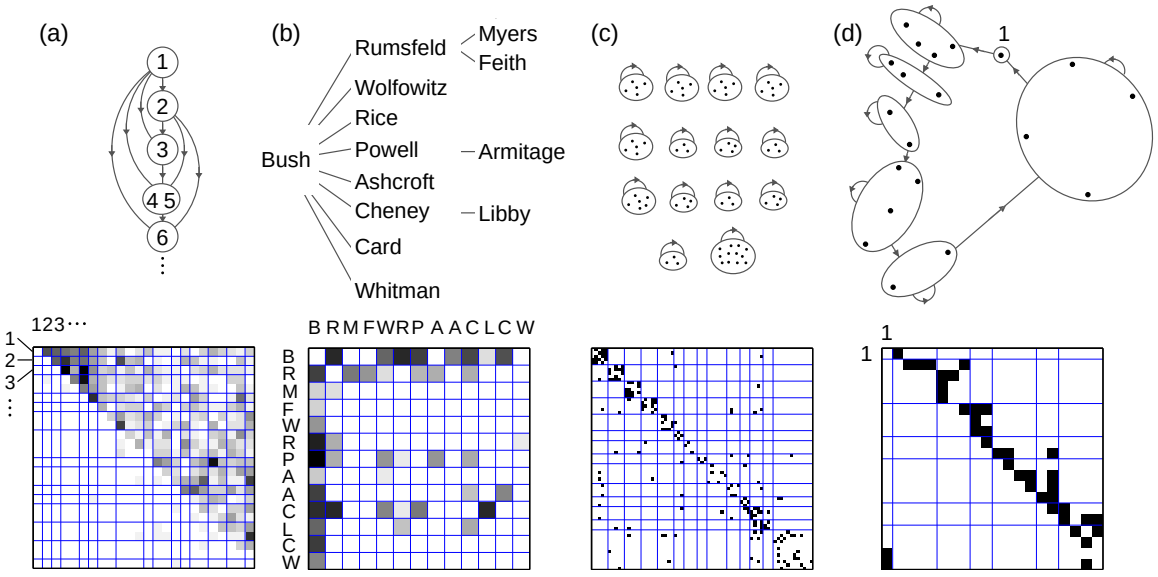

Figure 6-9: Structures learned from relational data (top row), and the raw data organized according to these structures (bottom row). (a) Dominance relationships among a troop of sooty mangabeys. The sorted data matrix has most of its entries above the diagonal, indicating that animals tend to dominate only the animals below them in the order. (b) A hierarchy representing relationships between members of the Bush administration. (c) Social cliques representing friendship relations between prisoners. The sorted matrix has most of its entries along the diagonal, indicating that prisoners tend only to be friends with prisoners in the same cluster. (d) The Kula ring representing armshell trade between New Guinea communities. The positions of the communities correspond roughly to their geographic locations.

kinds of directed order (one with self-links, the other without) score better than the other forms. A fragment of the best-scoring order is shown in Figure 6-9a, and this order is consistent with the dominance hierarchy inferred by primatologists studying this troop.

Bush Cabinet

Next we explored whether the model could discover the structural form of a human organization. The data set D is now a matrix of interactions between members of George W. Bush's first-term administration. Entry D_{ij} in the matrix is the number of Google hits for the phrase "i told j," where i and j vary over 13 members of the Bush administration.[11] Although there are some hits for phrases like "Bush told Bush,"

[11]These Google searches were carried out on January 26, 2006.

we set all counts along the diagonal to zero.

When applied to these data, the model concludes that the best form is an undirected hierarchy. The best hierarchy found (Figure 6-9b) closely matches an organizational chart built by connecting individuals to their immediate superiors, and the undirected nature of this representation indicate that information travels in both directions along each link in the hierarchy. Studying the raw data in Figure 6-9b indicates that the undirected hierarchy cannot be recovered by simply thresholding the matrix D. For instance, "Libby told Bush" has a higher weight than "Whitman told Bush," even though Bush is directly connected to Whitman but not Libby in the representation chosen by the model. Heuristics like thresholding may discover interpretable structure in some cases, but probabilistic approaches are useful when dealing with noisy real-world data.

Prisoners

Next we analyzed a relational matrix D that represents friendships between 67 prison inmates. The inmates were asked "What fellows in the tier are you closest friends with?" (MacRae, 1960). Each inmate mentioned as many friends as he wished, and entry D_{ij} is set to 1 if inmate i listed inmate j. Clique structures are often claimed to be characteristic of social networks (Girvan & Newman, 2002), and the model discovers that a partition (a set of cliques) provides the best account of the data.

Armshell Trade

Our final relational example considers trade relations between New Guinea communities (Hage & Harary, 1991). The 20 communities in the data set belong to the Kula ring, an exchange structure first described by Malinowski (1922). The raw data are represented as a matrix D where entry D_{ij} in the data matrix is set to 1 if community i sends *mwali* (armshells) to community j. As expected, the model concludes that a directed ring provides the best account of these data.

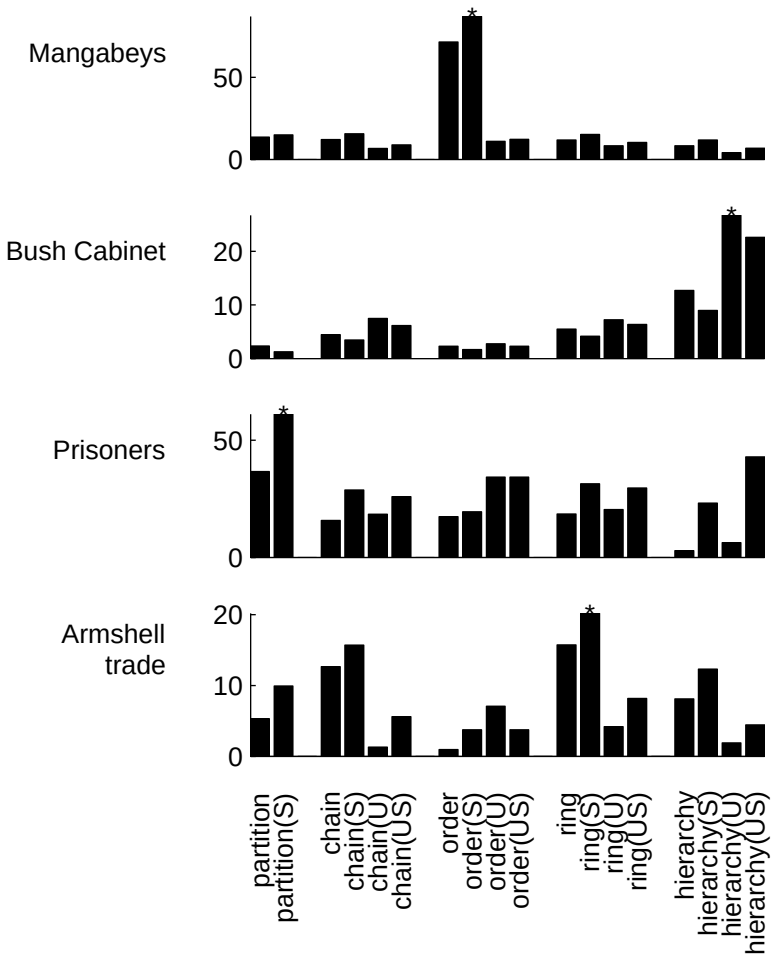

Figure 6-10: Scores for eighteen structural forms on relational data. U indicates an undirected form, and S indicates a form with self links (see Figure 6-8). The scores have been translated that the lowest score in each case is close to zero.

Related models

Although there have been few comprehensive studies of form discovery, our model is related to several lines of previous work. Our general approach can be viewed as an application of statistical model selection (Kass & Raftery, 1995). From a Bayesian perspective, model selection can be achieved by describing a hypothesis space of models (for us, each model is a pair (S, F)) and using Bayesian inference to choose between them. Other approaches are sometimes proposed: Pruzansky, Tversky, and Carroll (1982) decide whether a similarity matrix is better described by a tree or a

two dimensional space by finding the best instance of each form, and choosing the structure that accounts for the most variance. Several authors (Waller & Meehl, 1998; Boeck, Wilson, & Acton, 2005) have proposed methods for distinguishing between cluster structures and dimensional structures.

A key feature of a Bayesian approach is that it automatically penalizes unnecessarily complex models. Some such penalty is essential when considering structural forms of different complexities, since complex forms (e.g. fully connected graphs) can easily mimic simpler forms. Each chain, for example, is a special case of a grid, and it follows that the best grid S_{grid} will account for any data set D at least as well as the best chain S_{chain}: $P(D|S_{grid}) \geq P(D|S_{chain})$. The approach of Pruzansky et al. (1982) will therefore never choose the simpler model class, even when the data D were actually generated over a chain.[12]

Feature data

The feature-based model is related to previous work on learning the structure of graphical models (Dempster, 1972; J. Whittaker, 1990; Dobra, Jones, Hans, Nevins, & West, 2004). Previous models usually belong to one of two families. The first family includes models that impose no strong constraints on the form of the graph structures that are learned. Bayesian approaches within this family generally use a prior that includes all possible graph structures, and the prior over this space is usually relatively simple—for example, Dobra et al. (2004) use a prior that favors graphs with small numbers of edges. Models in the second family assume strong constraints on the form of the graph to be discovered, but these constraints are fixed from the start, not learned from data. Approaches in this second family include algorithms for phylogenetic reconstruction (Huelsenbeck & Ronquist, 2001) that attempt to discover tree-structured graphical models.

Our form-discovery model falls in the little-explored territory between these two

[12]Pruzansky et al. (1982) recognize the importance of model complexity, and justify their approach by arguing that the complexity of trees is approximately equal to the complexity of two dimensional spaces.

families of models. Instead of working with generic priors over the set of all possible graph structures, we developed an approach that concentrates the prior probability mass on graphs that correspond to one of a small number of structural forms.[13] The ultimate argument for such a prior is that it captures background assumptions that are well-matched to the problems we wish to solve. The need for suitable background assumptions is most pressing when dealing with sparse data, and sparse data are the rule rather than the exception in both cognitive development and scientific discovery.

Relational data

The relational model also builds on previous approaches to structure learning (H. C. White, Boorman, & Breiger, 1976; Nowicki & Snijders, 2001; Taskar, Segal, & Koller, 2001; Girvan & Newman, 2002). As for the feature-based case, previous approaches to relational learning usually belong to one of two families. Consider, for instance, the many previous models for relational clustering, or identifying clusters of entities that relate to each other in predictable ways. The first family includes models that impose no strong constraint on the form of the structures to be discovered. Stochastic blockmodels (Wang & Wong, 1987; Kemp et al., 2006) are one example: they do not incorporate the notion of structural form, and cannot conclude that a set of clusters is organized into a simple form like a ring, or a set of cliques. The second family includes models that assume that the structural form is known in advance. For instance, there are several algorithms for discovering community structures in networks (Girvan & Newman, 2002; Kubica, Moore, Schneider, & Yang, 2002). These approaches usually assume that the data are organized into a set of cliques, and that individuals from any given clique tend only to be related to others from the same clique. The form-discovery model again occupies the little-explored territory between these two families of approaches.

[13] Even though the notion of structural form is the most distinctive aspect of our model, this model differs from previous structure learning models in at least three other respects. First, standard methods for learning the structure of Gaussian graphical models do not allow latent nodes. Second, these methods make no attempt to cluster the nodes. Third, these methods allow graphs where some of the edges capture negative covariances. For the generative model in Equation 3, an edge between two entities always encourages the entities to have similar feature values.

Learning from sparse data

So far I have argued that structural forms can be learned, but we have seen few concrete examples of the inductive benefits that form discovery can bring. The inductive constraints provided by structural forms seem especially relevant to two kinds of problems: problems where a novel entity is sparsely observed, and problems where an entire system of entities is sparsely observed.

Novel entities

Inductive constraints are most important when data are sparse, and inferences about novel entities are often based on very sparse data. Suppose, for example, that the 20 mangabeys in Figure 5a are confronted by a new animal—mangabey X. Mangabey X has interacted with the troop on only one occasion, when he challenged and dominated mangabey 1. A learner who knows that the troop is organized into a dominance hierarchy can predict that mangabey X will dominate every other animal in the troop. A learner with a diffuse prior over graphs, however, will be unable to draw any conclusion from the single observation involving the new animal.

Similar problems arise when the data are features rather than relations. Suppose that you glimpse a novel animal at the zoo, and you think you see that it has the head of a bird and the body of a dog. If you know that biological species are organized into a tree, you should begin to doubt what you saw, since there is no way of extending your current tree so that the new animal is close to both the birds and the dogs. A model with a diffuse prior over graphs, however, will happily create a new graph by connecting the new animal to both the birds and the dogs.

Novel systems of entities

Structural forms are useful when new entities are encountered one at a time, but form discovery also supports inferences about entire systems of new entities (Novick, 1990). Suppose, for example, that a primatologist has spent several months studying one troop of mangabeys, and has discovered that the group is well described by a

dominance hierarchy. Knowing that mangabeys organize themselves into dominance hierarchies should allow her to quickly figure out the social structure of the next troop she studies, but a scientist who has not discovered the structural form of the first troop may take substantially longer.

Similar problems arise when the data are features rather than relations. Consider the case of Joseph Banks, the botanist on Cook's first voyage to the Pacific. As a young man, Banks studied the works of Linnaeus and presumably concluded that the species belonging to any given continent could be organized into a tree. Given this knowledge, a relatively small number of observations should have been enough for Banks to develop a tree-structured representation of the Australian species he encountered. A naturalist who had not read Linnaeus might have taken much longer to discover an adequate representation for the odd-looking animals he observed.

Form discovery in the laboratory

Structural forms are useful in part because they support inductive inferences, but we can turn this relationship around and use inductive inferences to diagnose whether a learner has successfully discovered the structural form of a domain. Two inductive tasks were described in the previous section: tasks where learners make inferences about new members of a known system, and tasks where learners make inferences about entirely new systems of entities. We developed experiments based on both tasks.

Experiment 1: Transfer to novel systems

In the first experiment, we trained participants on the structure of one relational system and asked them to generate two additional systems with similar structures. Identifying the form of the training system should allow learners to generate further instances of this form.

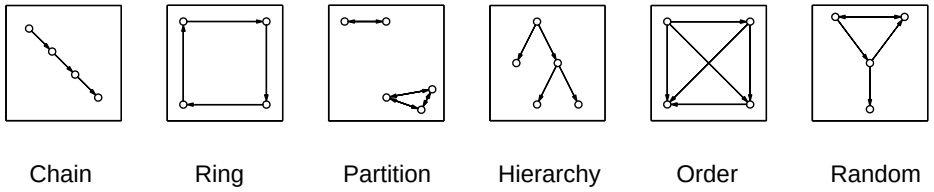

Figure 6-11: The six training systems used in Experiment 1. The first five systems are instances of simple structural forms.

Participants

12 members of the MIT community were paid for participating in this experiment.

Materials and Methods

The experiment included six within-participant conditions. In each condition, participants learned a single system (the training system) and generated two additional systems (the transfer systems). The six training systems are shown in Figure 6-11: the first five are instances of simple structural forms, and the final system was intended to be a more random kind of structure.

The task was introduced as follows:

> Mr Cheeryble is an eccentric billionaire who owns many small companies across many different industries. Cheeryble strongly believes that companies in the same industry should be organized similarly, although companies in different industries can be organized differently. He also firmly believes that important documents should be enclosed in red envelopes.
>
> As a management consultant, you have been hired to secretly observe the organization of five of Mr Cheeryble's companies, each from a different industry. You will begin your investigation in the mailroom and observe how red envelopes are exchanged within each company.

The experiment was carried out on a computer, and during the learning phase the interface had a single button labeled "Observe." Upon clicking this button, participants were told about an event corresponding to one of the edges in the current

training system: they might be told, for example, that "John sends a red envelope to Bill" (employee names were randomized across participants). After some number of observations, participants were given a test which included a yes/no question about each pair of employees (e.g. "Does John send red envelopes to Bill?"). Participants continued to observe edges in the training system until they were able to answer all of the test questions correctly.

Figure 6-11 shows that each training system included four or five nodes. After participants had learned the training system, they were asked to write a brief description of the organization of this company. Participants were then told that "Cheeryble has another company in the same industry with six employees," and asked to "indicate one way in which the company might be organized." After generating the first transfer system (a six node system), participants were asked to generate a transfer system with seven nodes. Since participants were provided with only a single training system for each condition, each of their inferences is an instance of one-shot learning.

Results

The six-node transfer systems are shown in Figure 6-12. The collection of seven-node transfer systems is qualitatively similar, although not described in this thesis. For each of the first five conditions, Figure 6-12 shows that at least five of the twelve participants generated a transfer system that was consistent with the form of the training system (counts for consistent systems are circled). These results support the idea that humans are able to discover the abstract organizing principles of a relational system. Responses for the random condition were more diverse, and no transfer system was chosen more than twice. Note that the random system is similar in many respects to the other training systems—for example, it has about the same number of nodes and edges as the other systems. The random system, however, has no recognizable structural form, which appears to explain the lack of consensus in this condition.

The verbal descriptions of the training systems provide further evidence that participants were able to discover the structural forms of the first five systems. When

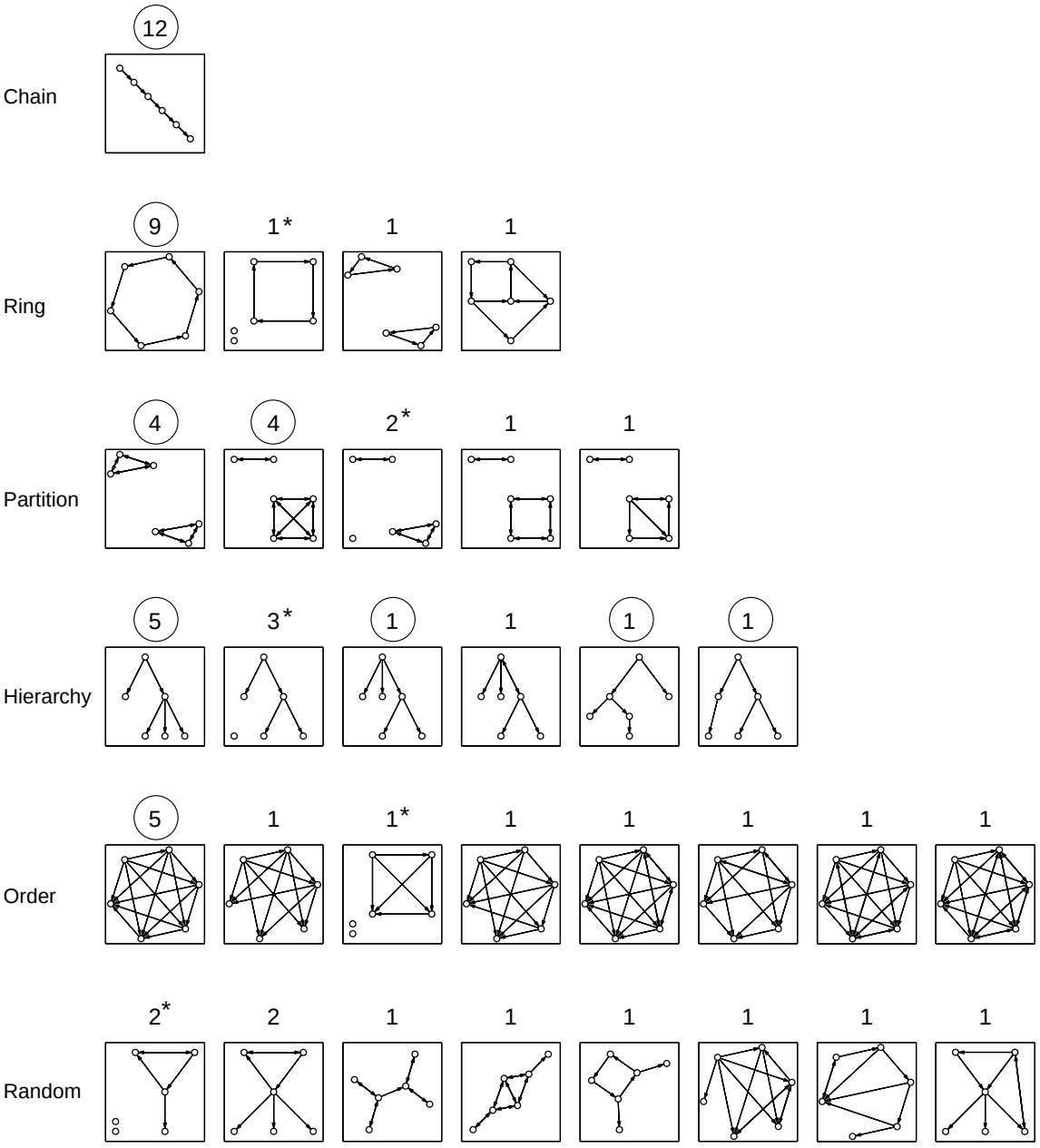

Figure 6-12: Transfer systems. Each row represents one of the conditions of experiment 1. The number above each system is the number of participants (out of twelve in total) who generated that structure. Circled numbers indicate systems that are consistent with the structural form of the training system, and numbers with asterisks indicate cases where the training system was simply reproduced. Only eight of the ten systems generated for the random condition have been shown.

describing the chain, one participant wrote:

> There is an obvious chain of command. It goes from top to bottom through two middle men.

The same participant described the ring as follows:

> There is no clear 'boss'. Envelopes are sent in a one directional circle that cycles through the employees of the company.

and gave this description of the partition:

> It appears that this company has two sections, a three man group and a two man group. Within each group, everyone can send envelopes to everybody. However, the one group does not send envelopes to the another [*sic*].

Like all one-shot inferences, generating a transfer system is a problem that is highly underconstrained. Although most participants appeared to rely on the notion of structural form, many other strategies are logically possible. For each condition except the chain condition, a handful of participants generated a transfer system that was identical to the training system except that it had some extra, isolated nodes. These generalizations are marked with asterisks in Figure 6-12. Simply reproducing the training system is a sensible response to the transfer task, and the fact that so many participants generated a structure different from the training system suggests that the notion of structural form is relatively intuitive.

Some of the less popular responses suggest that participants may have detected regularities in the training systems other than the regularities we had in mind. In the partition condition, most participants appeared to interpret the training system as a pair of cliques, but one participant may have represented it as a pair of rings, as suggested by the fourth transfer system for this condition (Figure 6-12). In the order condition, one participant may have interpreted the training system as a four-level structure where multiple nodes can belong at each level. This interpretation is consistent with the second transfer system for this condition (Figure 6-12), which is a

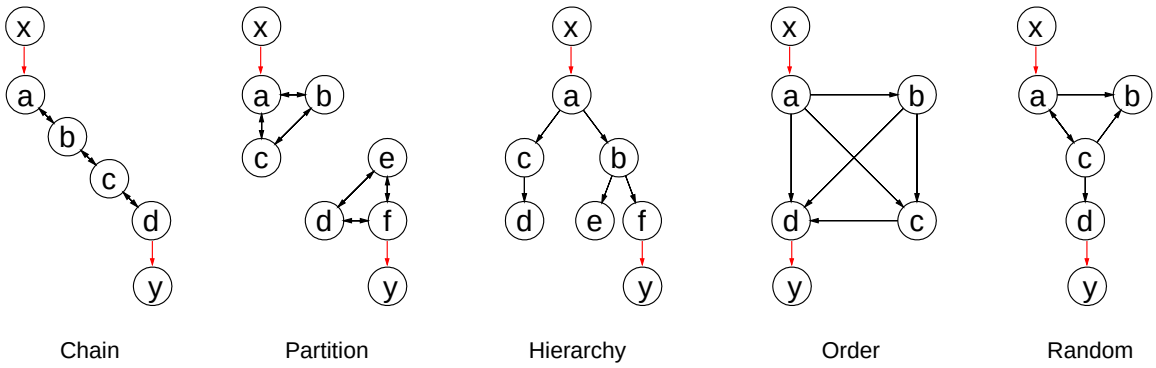

Figure 6-13: The five training systems used for experiment 2. In each condition, participants were initially trained on a structure with up to six nodes (a through f). After training, two new nodes were introduced (x and y) and two interactions involving these nodes were provided (the links shown in red). Participants then predicted how x and y would interact with all of the nodes in the training system.

four-level structure with three nodes at the bottom level. Since there are many regularities that participants might have picked up, it is revealing that most responses were consistent with the structural forms we had in mind when designing the experiment. This result suggests that the structural forms indicated by the labels in Figure 6-11 are psychologically natural—more natural, for instance, than the many other regularities that are consistent with each training system in Figure 6-11.

Experiment 2: Predictions about novel entities

Structural forms allow knowledge to be transferred from one system to another, but should also support inductive predictions about new members of a known system. In a second experiment, we trained participants on the structure of a system then asked them to make inferences about new members of this system.

Participants

Ten members of the MIT community were paid for participating in this experiment.

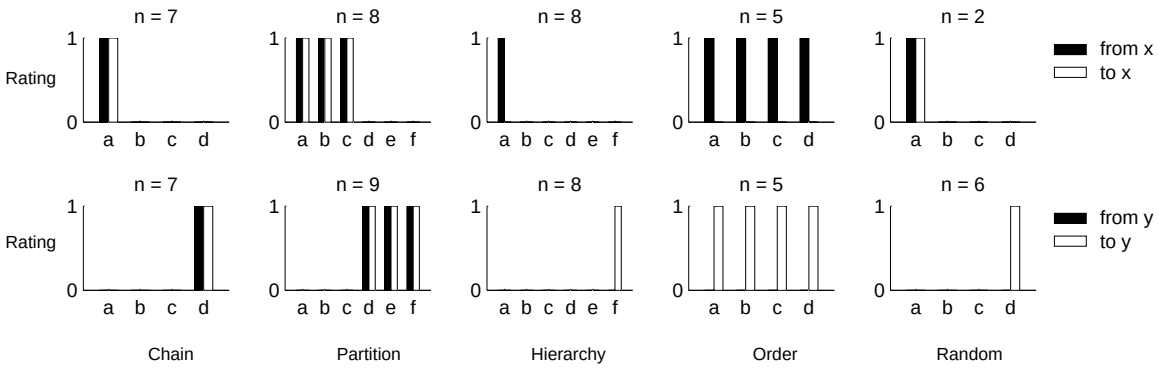

Figure 6-14: Experiment 2: modal predictions. The number above each subplot represents the number of participants (out of ten in total) who gave that response. For each of the first four conditions, the modal prediction is consistent with the structural form of the training system.

Materials and Methods

Experiment 2 was very similar to Experiment 1. There were five conditions: in each of these conditions, participants learned a training system and generated a seven-node transfer system. The training systems are shown in Figure 6-13. After each transfer system had been generated, participants were told that

> Two additional employees (x and y) were away on vacation when you started in the mailroom. Now they are back at work, and you observe one interaction involving x and another involving y.

Entities x and y and the observations associated with each are shown in Figure 6-13. Participants were then asked to predict how x and y would interact with all of the remaining nodes in the training system, and were asked in addition to explain their responses.

Results

Predictions about the new entities x and y are shown in Figures 6-14 and 6-15. For each of the first four conditions, the modal prediction was consistent with the structural form of the training system, and was made by at least half of the ten participants. In the order condition, for example, node x sends a link to the node

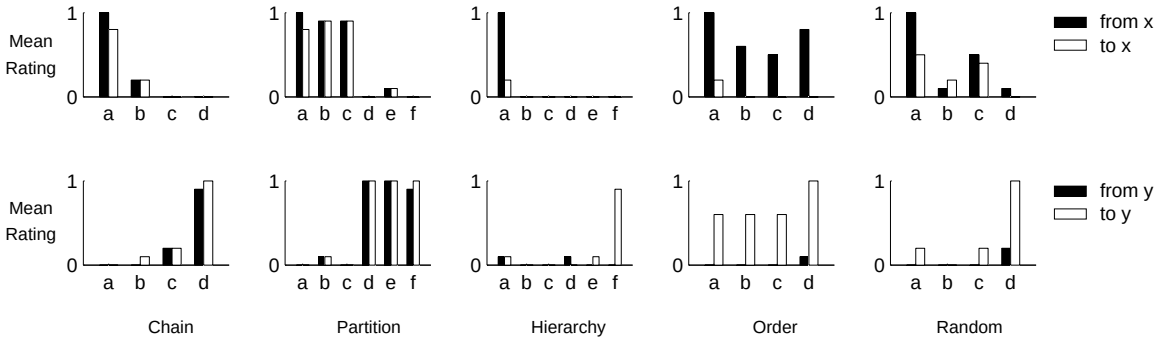

Figure 6-15: Experiment 2: mean predictions.

that was previously at the top of the order, which suggests that x will send a link to each of the remaining nodes. Node y receives a link from the node that was previously at the bottom of the order, which suggests that y will receive links from all of the other nodes.

Alternative explanations may account for the results observed in some of the five conditions, but it is difficult to explain the full pattern of results in Figure 6-14 without invoking the notion of structural form. Since x has been observed only to send a link to a, a simple baseline model might assume that x is just like a, and participates in a given relationship only if a does. This model accounts fairly well for the modal predictions in the order and partition conditions, but does not capture the modal responses in the chain and hierarchy conditions (Figure 6-16). Other baseline models might be considered, but there appears to be no simple alternative that will account for all of the behavioral data.

The verbal descriptions provided by participants provide further evidence that their inferences were often based on the notion of structural form. In the order condition, for instance, one participant gave the following justification for his predictions about x:

> If x can send to a, x must be in the highest position. x can therefore send to all the other employees.

and explained his predictions about y as follows:

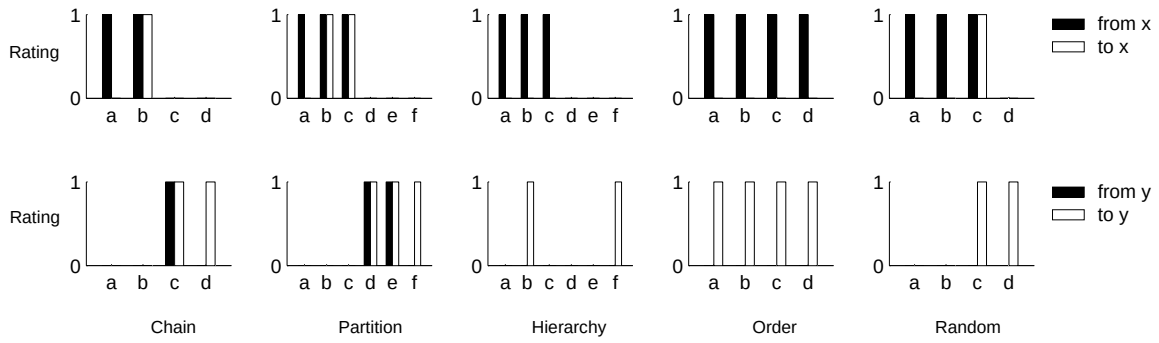

Figure 6-16: Experiment 2: predictions of a baseline model which assumes that nodes x and y are just like the nodes they are linked to.

> If d sends to y, y must be the most downstream employee. Therefore, y will receive red envelopes from all the employees, but cannot send to anyone.

Taken together, our two experiments support the idea that humans can discover the structural form of a relational system, and can use this knowledge to make predictions about new or sparsely observed entities. There are several promising directions for future experiments to pursue. The inductive tasks we chose are inspired by real-world problems that human learners must solve, but other experimental paradigms may also be worth exploring. For instance, participants should be faster to learn an instance of a simple structural form than a random system with a comparable number of edges (DeSoto, 1960). Errors made by participants should also be revealing: when learning a noisy instance of a given structural form, for instance, errors should tend to "regularize" the structure, or transform it into a better instance of the underlying form (Freeman, 1992).

Modeling cognitive development

Both of our experiments used adult participants, but some of the most impressive feats of form discovery may occur as children learn about the structure of the world. I predict that the ability to discover structural forms will be found relatively early

in development, but testing this prediction may raise some interesting experimental challenges. To motivate future work in this direction, I present one of the developmental predictions made by the form-discovery model.

As children learn more about a domain, their mental representations appear to undergo qualitative transitions that have been likened to paradigm shifts in science (Carey, 1985a; Kuhn, 1970). The form-discovery model shares this ability to move between qualitatively different representations of a domain. Given a small amount of data, the model typically chooses a form that is simple, but that does not capture the true structure of the domain. As more data arrive, the model should reach a point where the true structural form is preferred.

To demonstrate a qualitative shift in biological knowledge, we presented the model with more and more features of the animals in Figure 6-6a. We could have run this simulation by randomly sampling smaller data sets from the full feature matrix, but the results might have been influenced by idiosyncratic properties of the small data sets sampled. To avoid this problem, we directly specified the covariance of each data set and worked with the similarity version of the model. We analyzed data sets where the effective number of features was 5, 20, or 110, and the similarity matrix in each case was the covariance matrix for the full set of animal features. Even though the similarity matrices are identical, increasing the effective number of features should allow the model to discover more complex representations. When only 5 features are provided, the model should attempt only to fit the broad trends in the data, but given 110 features, the model should attempt to explain some of the more subtle variation in the data.

Figure 6-17 shows the representations chosen by the model for each data set. At first, the simplest form is preferred, and the model chooses a set of clusters. Given 20 features, the tree form is preferred, but the chosen tree is simpler than the tree in Figure 6-6a. The final tree is identical to the tree in Figure 6-6a: note that a similarity data set with 110 features is effectively identical to the data set that led to Figure 6-6a.

The developmental shift in Figure 6-17 is reminiscent of a trajectory that children

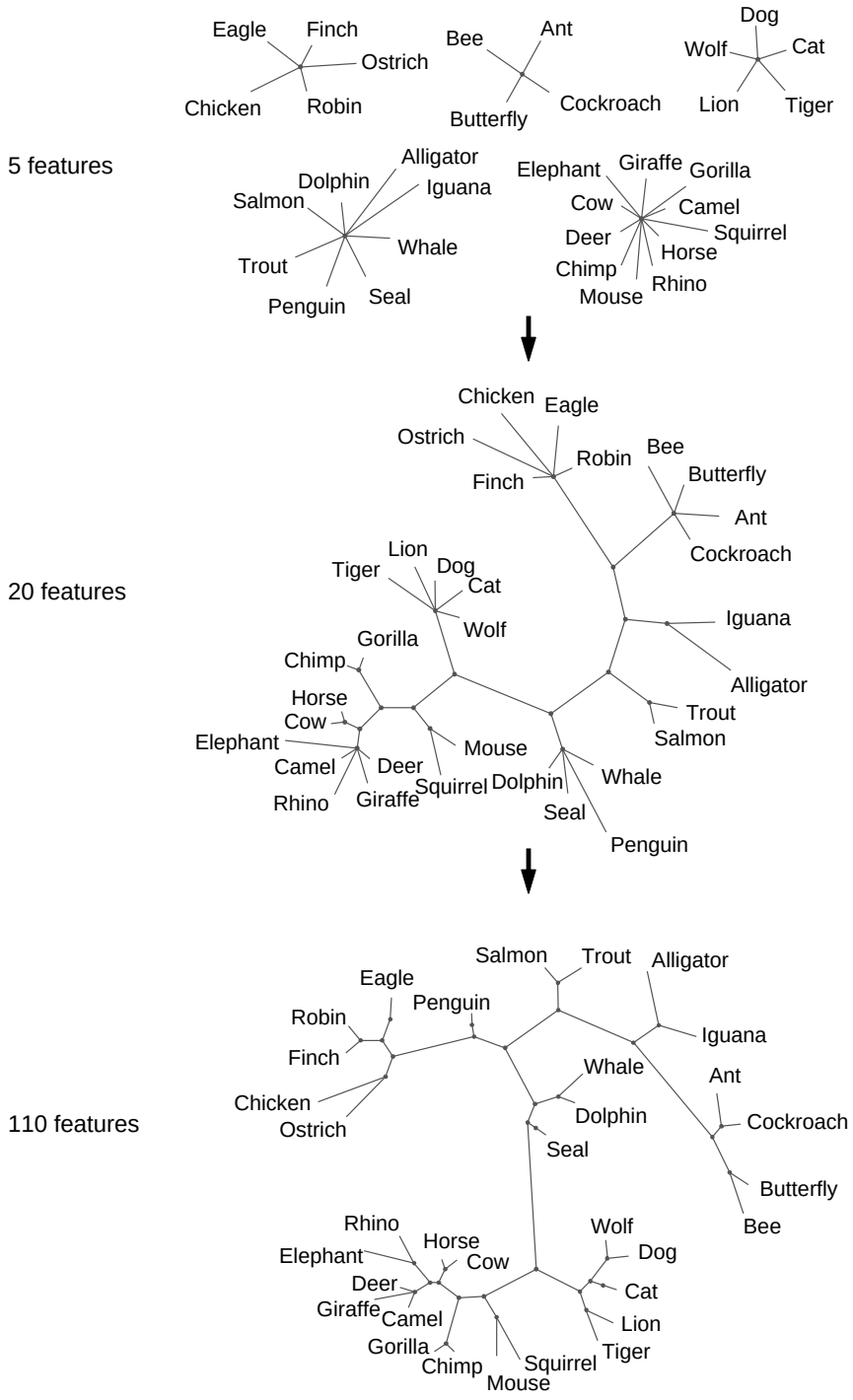

Figure 6-17: Two kinds of developmental change. Given only 5 features, the model chooses a partition (a set of clusters). As the number of features grows from 5 to 20, the model makes a qualitative shift between a partition and a tree. As the number of features grows even further, there is incremental change and the tree becomes more complex.

appear to follow as they learn the meanings of words. Early in development, children respect the assumption of mutual exclusivity: they organize objects into a set of non-overlapping clusters, with one category label allowed per cluster (Markman, 1989). Eventually, however, children realize that objects can be organized into taxonomic hierarchies. Figure 6-17 suggests that this insight may be driven in part by the amount of data available to older children.

The ability to learn from raw data may support some of the earliest and most fundamental shifts in children's thinking. Bottom-up learning, however, can only explain some aspects of cognitive development, and explicit instruction may contribute to the majority of developmental shifts once children have become proficient language users. Although I have focused on learning representations from raw data, hierarchical approaches can naturally handle linguistic input at multiple levels of abstraction, including all three levels in Figure 6-1. Linguistic input can provide new features (e.g. "whales breathe air"), and can also provide direct information about a structure S (e.g. "whales belong with the mammals rather than the fish") or a form F (e.g. "the theory of evolution implies that animals should be organized into a tree"). Modeling learning when input is simultaneously provided at several levels of abstraction is an important goal for future work.

Conclusion

This chapter presented a hierarchical Bayesian model (Figure 6-1) that helps to explain how humans discover the structural form of a domain. I showed that the model discovers interpretable structure in several real-world data sets, and described two experiments which support the idea that humans can discover the best kind of representation for a domain.

The form discovery model is broader in scope than the hierarchical models described in Chapters 4 and 5. Word learning and causal learning are important areas of study, but both of these areas are relatively self contained. Knowledge representation is a more general topic that is relevant to virtually every area of cognitive

science: for instance, theories of language, perception, action, and reasoning must all make claims about the structure of mental representations. By addressing a topic as general as knowledge representation, this chapter demonstrated several key features of the hierarchical Bayesian approach.

First, hierarchical Bayesian models help to explain how domain-specific constraints can be acquired. A typical nativist view recognizes that inductive inference relies on domain-specific constraints but assumes that these constraints are innately provided (Chomsky, 1980; Atran, 1998; Kant, 2003). Chomsky (1980), for instance, has suggested that "the belief that various systems of mind are organized along quite different principles leads to the natural conclusion that these systems are intrinsically determined, not simply the result of common mechanisms of learning or growth." The form discovery model offers an alternative view, and suggests that domain-specific constraints can be acquired using domain-general statistical inference. This perspective has been previously emphasized by connectionist modelers, who argue that "domain-specific representations can emerge from domain-general architectures and learning algorithms" (Elman et al., 1996).

Models that learn about many domains point the way towards unified accounts of human learning. There are many special-purpose models in the literature, including models of word learning (Chapter 4), grammar learning, causal learning (Chapter 5), concept learning, perceptual learning, and motor learning. Each kind of learning is distinctive in its own way, but there may be general principles that help to explain all of these different abilities. The hierarchical Bayesian approach provides a natural way to combine the insights behind several special-purpose models. Given several models that are somewhat related, we can introduce a new level of abstraction that captures their commonalities, and can treat each special-purpose model as a component of a single, more general model. The form discovery model, for instance, shows how models that learn trees, rings, chains, and partitions can be absorbed into a single general framework for structure learning.

Connectionist models can also handle problems from many domains, but the form discovery model departs from the connectionist approach in one particularly im-

portant respect. Standard methods for learning connectionist networks (Rogers & McClelland, 2004) use the same generic class of representations for every task, instead of attempting to identify the distinctive kinds of structures that characterize individual domains. Without these structural constraints, connectionist models can require unrealistically large quantities of training data to learn even very simple concepts (Geman et al., 1992). The form discovery model recognizes that different kinds of representations are appropriate for different domains, and that the right kind of representation is crucial for explaining how learning can succeed given sparse data. Structured representations are important in part because of the inductive constraints that they capture, and a comprehensive theory of learning (Table 1.2b) should aim to incorporate many kinds of representations.

Our current model can be extended in several ways to provide a more comprehensive account of form discovery. A natural first step is to implement the idea that the structure grammars used by the model are generated from a single underlying meta-grammar. The meta-grammar in Figure 6-3 is one initial proposal, but there may be important classes of graphs (e.g. small-world graphs) that are not generated by this meta-grammar, and it will be important to consider alternative schemes for generating graph structures. A longer term goal is to extend the model to handle structured representations other than graphs. To mention only two possibilities, logical representations are useful for capturing some aspects of semantic knowledge, and Markov decision processes are useful for modeling action and decision making.

Even though I have focused on graph structures, the basic idea behind the form-discovery model should apply more generally. Given a set of structure grammars, a learner can identify the grammar that provides the best account of a data set, and this approach can be pursued regardless of whether the grammars generate graph structures, logical representations, or other kinds of representations. As mentioned earlier, the ultimate goal is to develop a "universal structure grammar" that fully characterizes the representational resources available to human learners. This universal grammar, for instance, might specify a set of representational units and a set of rules for combining these units to create structure grammars of various kinds, in-

cluding grammars that generate graphs, logical representations, Markov decision processes, and many other families of representations. The notion of a universal structure grammar is highly speculative at present, but attempts to characterize this universal grammar should provide some insight into the growth of mental representations.

Chapter 7

Conclusion

Inductive inferences depend critically on constraints. Some of these constraints must be innate, but I have suggested that hierarchical Bayesian models help to explain how the rest are acquired. Hierarchical Bayesian models include representations at multiple levels of abstraction, and the representations at the upper levels place constraints on the representations at the lower levels. Statistical inference over these hierarchies helps to explain how the constraints at the upper levels are learned.

To demonstrate the psychological relevance of this approach I described models that address three aspects of high-level cognition: categorization (Figure 2-2c), causal reasoning (Figure 2-2e), and knowledge representation (Figure 2-2f). Each of these models can be developed further and subjected to additional experimental tests, but more important than any single model is the general theoretical framework I described. Here I expand on four of the most important lessons that can be learned from this framework. First, inductive constraints are often considered as prerequisites for learning, but constraints can themselves be learned. Second, constraints can be learned fast: in particular, constraints can be learned from small amounts of data, and constraints can be learned before the hypotheses they constrain are securely in place. Third, a statistical approach to constraint learning also explains how constraints are used for induction. Fourth, working with abstraction hierarchies is a useful general strategy for understanding the acquisition of human knowledge.

Lesson 1: Inductive constraints can be learned

Constraints and learning mechanisms are sometimes seen as competing explanations for cognitive abilities. Researchers who focus on constraints often adopt a nativist approach and assume that these constraints are innately provided. Researchers who focus on learning often adopt an empiricist approach and explore how much can be achieved by mechanisms that are relatively unconstrained. The hierarchical Bayesian framework suggests that inductive constraints and inductive learning can and should be studied together. Constraints are critical for explaining how humans acquire knowledge so quickly and from such sparse data. Learning can explain how some of these constraints are acquired in the first place.

The formal framework developed in this thesis helps to explain the acquisition of *epistemic* constraints, or constraints that correspond to forms of abstract knowledge. Some constraints (such as memory limitations) do not qualify as epistemic constraints, but the set of epistemic constraints is relatively broad and includes examples of domain-specific constraints, domain-general constraints, soft constraints, and hard constraints. Despite this generality, the notion of an epistemic constraint helps to clarify what it means to learn an inductive constraint. Any epistemic constraint is potentially a target for learning frameworks including the framework developed here, but it makes little sense to ask how non-epistemic constraints could be learned.

The framework I described relies on Bayesian inference, and suggests that epistemic constraints can be acquired in the same way that any other kind of knowledge can be acquired. Given principles that generate a hypothesis space of epistemic constraints, a Bayesian learner can select the constraint in this space that best accounts for a body of observed data. Loosely speaking, a Bayesian learner should increase its degree of belief in a hypothesis to the extent that the data are compatible with that hypothesis, and incompatible with most alternative hypotheses. The same approach goes through regardless of whether the hypotheses correspond to epistemic constraints or other kinds of knowledge. This view of learning suggests that there is nothing particularly special about the acquisition of inductive constraints, and that

the same fundamental principles that explain other kinds of learning can also explain how inductive constraints are acquired. To put the same idea in a more positive light, I began with a problem that seemed mysterious at first—the problem of learning inductive constraints—and showed how it can be handled by familiar computational techniques.

Although this thesis has argued that constraints can be learned, I do not claim that this learning takes place in the absence of any background assumptions. Any Bayesian account will rely on prior knowledge, and each of our hierarchical models assumes that the prior at the topmost level is fixed in advance, and that the process by which each level is generated from the level immediately above is also known. We can think about relaxing some of these assumptions, but any learning framework will rely on initial knowledge of some sort. The hierarchical Bayesian framework helps to explore how much initial knowledge is required, and what form this knowledge must take.

Lesson 2: Inductive constraints can be learned *fast*

An important challenge for constraint-learning models is to explain how constraints are learned fast enough to be useful. A word-learning model, for instance, does not seem illuminating if it must acquire thousands of words before it discovers constraints like the shape bias. Constraints like the shape bias are supposed to *support* word learning: instead of being extracted from a large database of words, they are supposed to explain how these words could be learned in the first place. We saw two ways to understand how word-learning constraints might be learned rapidly. First, some constraints (including the shape bias) can be learned given just a few examples from just a few categories. Second, abstract-to-concrete learning might explain how children acquire word-learning constraints before they are confident about the meaning of any single word. The same two ideas may help to explain how many other constraints are rapidly acquired.

When abstract knowledge is available very early in development, it is natural to

conclude that this knowledge is innate. Versions of this argument have been used to support nativist claims about several of the domains in Table 1.1 (E. S. Spelke, 1990; Wynn, 1992). The hierarchical Bayesian approach suggests an alternative view: in some cases, abstract knowledge may appear to be innate only because it is acquired much faster than knowledge at lower levels of abstraction. Exploring this possibility will be critical when deciding which of the constraints in Table 1.1 might be learned rather than innate.

There is some debate in the developmental literature about whether abstract knowledge is acquired before more concrete knowledge, or vice versa (Keil, 1998; Mandler, 2003). Hierarchical Bayesian models suggest that statistical inference can lead to both patterns of development, and that the pattern which emerges in any given case will depend on the task in question. Even though both patterns of development are discussed in the literature, most computational models of development focus on concrete-to-abstract learning. The work described here is one of the first formal attempts to understand how abstract knowledge can be acquired before more concrete knowledge is securely in place.

Lesson 3: Bottom-up and top-down inferences

It is possible that the acquisition and use of inductive constraints might turn out to be two rather different problems. Perhaps, for instance, we need one theory to explain how constraints are learned, and a second theory to explain how constraints guide inductive inferences. We saw, however, that the acquisition and use of inductive constraints can both be viewed as statistical inferences over a hierarchical architecture. Cases where knowledge at lower levels supports inferences at higher levels can be seen as instances of constraint learning. Cases where high-level knowledge guides inferences at lower levels help to explain how constraints support induction.

Although I focused on the acquisition of inductive constraints, I described several cases where constraints guide inductive inferences. We saw, for instance, how a constraint similar to the shape bias supports inferences about novel categories (Chap-

ter 4), how causal schemata support inferences about the causal powers of novel objects (Chapter 5), and how structural forms support inferences about new members of a relational system (Chapter 6). In each case, the relevant constraints were learned from prior observations, but we can also develop hierarchical models where the constraints at the upper levels are thought to be innate.

Like all human beings, psychology researchers are attracted to binary oppositions, and top-down and bottom-up approaches are sometimes presented as incompatible approaches to induction. The hierarchical Bayesian approach provides a unifying account which suggests that both kinds of inference are needed to account for cognitive development. Early in development, inductive constraints are learned by making bottom-up inferences based on observable data, and once established these constraints guide top-down inferences about novel contexts (Figure 4-8). As this trajectory suggests, bottom-up and top-down inferences are both needed to explain how knowledge is acquired and used.

Lesson 4: A method for understanding induction

As shown in Figure 2-3, a conventional Bayesian model makes inferences at two levels of abstraction, but the models in this thesis support inferences at three levels of abstraction. Moving from two to three levels might not seem like such a big step, but the important development is that we can now introduce as many levels as we need for a particular problem.

The ability to build models with multiple levels of abstraction suggests a general strategy for understanding inductive inference. I illustrate by describing how we came to develop the model described in Chapter 6. Suppose that we want to understand how a certain kind of inference can be made. Chapter 6 grew out of our interest in problems where people learn a handful of facts about a novel property (e.g. dolphins have property P) and make inferences about the distribution of the property (are seals or cows more likely to have P?). Inferences about biological properties can be explained if biological species are mentally organized into a tree-structure, and if

people know that nearby species in this tree tend to have similar properties (Kemp & Tenenbaum, 2003). This approach, however, introduces a second problem: how might people learn a tree structure that captures the similarity between animal species? This structure can be learned by observing physical, behavioral and ecological properties of different species and constructing a tree such that species with similar properties are close to each other. Again, though, our proposed solution opens up another question. Learning a tree from observed properties seems plausible, but how can learners know in advance that they should construct a tree rather than some other kind of representation? The model presented in Chapter 6 provides a possible solution: if learners start with structure grammars that characterize a space of possible representations, they can identify the representation that best accounts for the data they have observed. The sequence of questions does not stop here: as mentioned in Chapter 6, it is natural to ask how learners might acquire a set of structure grammars, and we can speculate about the conceptual resources that might be needed to construct this set.

This case study suggests a general strategy that can be applied to many cognitive problems. The general theme is that inductive inferences can be explained by identifying the knowledge on which they depend. That knowledge, in turn, can be acquired by relying on a body of knowledge that is even more abstract, and we can iterate this procedure and build models where knowledge is acquired at many levels of abstraction. Each time we add a level to a model there is an explanatory gain, since the new level helps to explain how knowledge at the second-highest level is acquired, and the modified model can potentially handle many phenomena that require different representations at this level. There is also an explanatory cost, since the modified model must know how representations at the second-highest level are generated given a representation at the highest level, and must include a prior distribution over the representations at the highest level. Each level of the form discovery model provides an explanatory gain that appears to outweigh its cost, and similar arguments can be made for the levels used by the other models in this thesis.

Adding levels can increase the explanatory power of a model, but this process must

stop at some level of abstraction. The stopping point will be reached when any additional level incurs a cost that outweighs any explanatory gain it might provide. This stopping point might alternatively be characterized as a point where the background assumptions about the highest level are simple enough or general enough that they can be plausibly assumed to be innate. If the assumptions about the highest level do not meet this criterion, then it is necessary to ask how they might be acquired, and the answer is likely to involve another level of abstraction. I do not claim that any of our current models has reached a stage where the assumptions about the highest level can be taken for granted. We might, however, approach this goal by supplementing the form discovery model with an extra level which indicates how graph grammars can be generated from a set of very basic concepts including objects (nodes), relations (edges) and the concept of recursive rule application. The knowledge at this new level might be general enough to generate grammars that support linguistic and visual inferences (Chomsky, 1965; Han & Zhu, 2005) as well as the graph grammars needed by our form discovery model.

Building models with many levels of abstraction should allow us to work towards unifying frameworks that account for many aspects of cognition. Many models of learning are designed to address a narrow family of phenomena: for instance, I presented a model of word learning (Chapter 4) and a separate model of causal reasoning (Chapter 5). People, however, are capable of many kinds of learning and reasoning, and eventually psychologists should aim to develop formal frameworks that are similarly broad in scope. The hierarchical Bayesian approach suggests one way to proceed. If we can identify the basic assumptions shared by several special-purpose models, we can introduce a level of abstraction that captures these assumptions, and can absorb the special-purpose models into a single general framework. In Chapter 6, for instance, we saw how methods for learning trees, rings, chains, and partitions can be combined to create a general framework for structure learning.

It is natural to construct a hierarchical model by adding levels of increasing abstraction, but we can also think about growing a model in the opposite direction and adding levels that become increasingly concrete. The form discovery model has a ma-

trix of features at its lowest level, but some of these features correspond to complex concepts in their own right (e.g. "is smart," "is fierce," and "lives in groups.") By adding a lower level to the model we can capture the idea that these complex features are constructed out of more primitive microfeatures. Several psychologists have made similar proposals about features and how they might be learned (Schyns, Goldstone, & Thilbaut, 1998). The hierarchical Bayesian approach suggests how these proposals can be incorporated in a unified framework that addresses inductive questions at many levels, including questions about the origin of features and the origin of inductive constraints.

Developmental implications

This entire thesis has been motivated by developmental questions, and the lessons just described are of obvious relevance to the study of development. Most sections in this concluding chapter will touch on developmental themes, but this section brings conceptual development into the foreground.

Models of learning and theories of cognitive development should have a great deal to contribute to each other. Early childhood is the time when children acquire the foundational concepts that will serve them for the rest of their lives, and the most impressive feats of human learning probably occur during this period. Formal models of learning can help to explain how children acquire so much knowledge so quickly, and empirical studies of development can help to identify the main principles that state-of-the art learning systems should aim to incorporate. Some approaches to development have already established relationships with ideas from the modeling literature. There are close ties, for instance, between the dynamic systems approach to development (Thelen & Smith, 1996) and the literature on connectionist modeling. Other perspectives on development have been pursued with very little input from the modeling community. One influential approach that has largely resisted formalization is known as the "theory theory" (Carey, 1985a; Gopnik & Meltzoff, 1997).

The hierarchical Bayesian approach may help to build a bridge between researchers

who work on formal models and researchers who study the emergence of intuitive theories. I have focused on inductive constraints rather than theories, but many of the constraints in Table 1.1 emerge from intuitive theories, and hierarchical Bayesian models help to explain how these theories can be learned (Tenenbaum et al., 2006). Although many models of development take a connectionist approach (Shultz, 2003; Rogers & McClelland, 2004), there are several reasons why this approach is not ideal for exploring the development of intuitive theories. One problem is that connectionist models do not incorporate structured representations, and cannot naturally capture the rich systems of knowledge discussed by many developmental psychologists. A second problem is that most connectionist models do not incorporate multiple levels of abstraction, and cannot clearly explain how learning proceeds at these different levels. These criticisms suggest that connectionism suffers from many of the same shortcomings as traditional learning theory (Table 1.2a), and new theories of learning (Table 1.2b) are needed to explain how rich systems of knowledge are acquired and used.

The hierarchical Bayesian approach is consistent with all four principles in Table 1.2b, and can perhaps become the foundation of a comprehensive account of cognitive development. There is more to development than learning, of course, but models of learning provide ways to explore some developmental proposals that have previously resisted formalization. I argued, for instance, that hierarchical Bayesian models can capture abstract-to-concrete learning, (Chapter 4), can explain how linguistic input shapes causal attributions (Chapter 5) and can capture developmental shifts between qualitatively different representations (Chapter 6).

To establish the developmental relevance of the hierarchical Bayesian approach it will be necessary to apply it to developmental problems from many domains. We made a start in this direction by applying the shape bias model to data collected by Smith et al. (2002), and Chapters 5 and 6 mention several developmental predictions of the remaining two models. Much more work is needed before the hierarchical Bayesian approach can be assessed as a serious theory of development, but our results so far suggest that this direction is worth pursuing.

Limitations

Each model presented in this thesis is limited in several respects. The Dirichlet-multinomial model assumes that each object is represented as a feature vector where one dimension represents shape, another color, and so on. Explaining how visual input might be parsed in this way is a problem that the current model does not address. The schema-learning model assumes that each object belongs to a single type, but some objects belong to multiple types. For instance, Bill Clinton is a male, a politician and a physical object and displays the causal powers and characteristic features of each type: he can grow a beard, he can charm a crowd, and he can break a pane of glass when moving at high speed. The form discovery model assumes that each data set has a single underlying structure, but different parts of the data may be explained by very different representations. As mentioned previously, some biological features are consistent with a taxonomic tree, but others are more consistent with a set of ecological categories (Shafto et al., 2006). Other limitations of our models might be listed, but it should come as no surprise that these models are limited in many ways.

Any psychological model can be improved and extended in many ways, and our models are no exception. Many limitations of these models can be addressed within the hierarchical Bayesian framework, including all of the limitations mentioned above. The real question is whether there are limitations that are intrinsic to my general approach: limitations that cannot be addressed by any hierarchical Bayesian model, no matter how sophisticated.

It is hard to imagine how the hierarchical component of my framework introduces any fundamental limitations, but the Bayesian component will raise questions in some readers' minds. The value of Bayesian approaches to cognition has been vigorously debated (Tversky & Kahneman, 1974; Anderson, 1990; Simon, 1991; Oaksford & Chater, 2007) and most criticisms of Bayesian approaches fall under two broad headings. Some researchers feel that Bayesian models can do too much: they argue that there will be a Bayesian model to account for any conceivable pattern of data,

and that the approach therefore offers little explanatory value. Others argue that Bayesian models can do too little: in particular, they suggest that Bayesian models will never be able to account for the many cases where people's inferences depart from normative standards.

The general claim that Bayesian models are too powerful may seem curious at first. Psychologists and machine learning researchers have explored many different models, but none comes close to matching the abilities of a five year old. None of these models can compete with humans at tasks like recognizing the objects in a scene, or answering commonsense questions about the narrative in a storybook. The first-order task for psychologists should be to address the vast limitations of our current models, not to develop models that achieve even less.

Concerns about the power of Bayesian models arise more naturally in specific contexts. A Bayesian model relies on a set of background assumptions that formalize the nature of the task and the prior expectations that the learner brings to the task. A researcher interested in a specific problem such as associative learning or property induction may worry that any conceivable pattern of data can be explained by adjusting these assumptions appropriately (Shanks, 2006). Although understandable, this concern seems motivated by some questionable views about the nature of scientific explanation. There will always be multiple theories that account for a given set of observations, including multiple theories that perfectly predict all of the observations. Competing theories must therefore be assessed according to several criteria. Accounting for empirical data certainly matters, but simplicity and consistency with our best current explanations of other scientific phenomena are also important. If background assumptions of arbitrary complexity are permitted, a Bayesian modeler may be able to account for any pattern of data.[1] If the background assumptions must be plausible, however, there will be many conceivable data sets that are not well explained by any Bayesian account.

[1]Cases where people give probability judgments that do not form a coherent probability distribution may appear to pose a problem, but our modeler can assume, for instance, that there is a time-dependent process which switches the inductive context every second, and that each probability judgment is accurate at the time it is given but soon out of date.

Since Bayesian models cannot explain *everything*, it is natural to ask how well they account for the human abilities we wish to explain. The Bayesian approach has provided insight into many aspects of cognition (Anderson, 1990; Oaksford & Chater, 2007), but there are some well-known cases where human behavior appears to diverge from normative standards (Tversky & Kahneman, 1974). These cases can be organized into at least three categories. In some cases, the proposed normative standard does not capture the true structure of the task, and people's behavior turns out to be consistent with a Bayesian account once the true structure of the task is recognized (Hertwig & Gigerenzer, 1999; McKenzie, 2003). In other cases, people's responses may be best explained as rational responses to an real-world problem that is slightly different from the problem posed by the experimenter (Tenenbaum & Griffiths, 2001). A third set of cases includes findings that a purely Bayesian analysis will be unable to explain. At present there is no clear consensus about the findings that belong to each category, but it seems likely that many cases will end up in the third category.

There are good reasons to expect that some empirical findings will resist a simple Bayesian explanation. Bayesian methods are useful for developing computational theories of cognition (Marr, 1982), but a complete account of cognition will also need to describe the psychological mechanisms that carry out the computations required by these theories. Since the computational resources of the mind are limited, some computational theories will be implemented only approximately, and these approximations may lead to patterns of behavior that have no adequate explanation at the level of computational theory. In order to explain everything that psychologists wish to explain, Bayesian models will need to be supplemented with insights about psychological and neural mechanisms.

Understanding processing mechanisms and developing computational theories are two separate projects, but successful computational theories can guide investigations of processing mechanisms. Psychologists and neuroscientists have discussed how probabilistic computations could be approximated by the mind (Anderson, 1990) and the brain (Ma, Beck, Latham, & Pouget, 2006), and there are proposals about how hier-

archical Bayesian approaches in particular could be implemented by the brain (Lee & Mumford, 2003). Although there are reasons to believe that the mind and brain are capable of approximating Bayesian computations, it is possible that different learning mechanisms are used for different tasks. Detailed studies are needed to understand the nature of these mechanisms, the settings in which they operate, and the extent to which each one is compatible with a Bayesian approach.

The distinction between computational theories and mechanistic models may not be as sharp as I have suggested, but some version of this distinction is essential for understanding the strengths and the limitations of the hierarchical Bayesian approach. This approach is a paradigm for developing computational theories of cognition, and does not appear to suffer from any fundamental limitations when applied in this way. The study of cognition, however, is more than just the pursuit of computational theories, and my framework is not intended to answer the many questions that emerge from the study of psychological and neural mechanisms.

Future directions

My framework opens up two general areas for further work. First, the framework can be applied as it stands to several kinds of problems from psychology and other disciplines. Second, the framework can be extended and improved in several ways.

Psychological applications

We saw that the hierarchical Bayesian approach can address three problems solved by human learners, but in order to establish the generality of this approach it will be necessary to develop hierarchical models that account for the acquisition of constraints in many different domains. Table 1.1 lists some of the constraints that have been proposed by psychologists, and that are potential targets for constraint-learning models. It is far from clear that all or even most of these constraints are learned by humans. All of these constraints, however, could be learned in principle, and hierarchical Bayesian models allow us to explore whether they are learnable given the data

available over the course of cognitive development.

Although I focused on cases where hierarchical Bayesian models learn something interesting, cases where Bayesian models fail to learn can be just as important. Since these models rely on rational statistical inference, any failure to acquire an inductive constraint cannot be attributed to a faulty learning mechanism. Instead, failures to learn indicate that the prior knowledge assumed by the model is too weak, or that the data provided to the model is too sparse, or both. If the data provided are representative of the data available to human learners, then models which fail to learn provide important evidence about constraints which need to be available from the start in order for learning to succeed. Gildea and Jurafsky (1996) provide a concrete example of this research strategy, and describe a model for learning phonological rules that succeeds only when some linguistically-motivated constraints are included.

Each model described in this thesis includes representations at several levels of abstraction, and there are at least three ways to test the psychological reality of these hierarchies. One strategy focuses on inferences at the bottom level of the hierarchy. Experiment 1 in Chapter 5 explored one-shot causal learning, and I argued that the upper levels of the schema-learning model explain how people make confident inferences given very sparse data about a new object. A second strategy is to directly probe what people learn at the upper levels of the hierarchy. Experiment 3 in Chapter 5 asked participants to sort objects into groups, and the resulting sorts provide evidence about the representation captured by the top level of our hierarchical model. A third strategy that I did not explore is to provide participants with information about the upper levels of the hierarchy, and to test whether this information guides subsequent inferences. Chapter 5, for instance, mentioned the case of a science student who is told that "pineapple juice is an acid, and acids turn litmus paper red." When participants are sensitive to abstract statements of this sort, we have additional evidence that their mental representations are similar to the abstraction hierarchies used by our models.

Of the three strategies just described, strategy one can be applied to learning problems from any domain, but strategies two and three need to be applied more

selectively. A language learner, for instance, may have acquired a grammar for her native language even if she is unable to describe it (strategy two) or to incorporate additional rules that might be provided by a linguist (strategy three). Although strategies two and three are less general than strategy one, they are critical in many cases of interest, since they help to explore the cultural transmission of inductive constraints. Language allows abstract knowledge to be described and directly supplied to others, and hierarchical models are valuable in part because they allow a role for linguistic input. Testing this aspect of my approach is an important direction for future work.

Natural language can capture inductive constraints that are much more sophisticated than any of the examples I considered. Some of these constraints may be best described as constraints that emerge from intuitive theories (Carey, 1985b; Keil, 1991), and a comprehensive attempt to explore the acquisition of inductive constraints must therefore explore the acquisition of intuitive theories. The hierarchical Bayesian approach can help to explain the acquisition of many kinds of abstract knowledge, including scripts, schemata, and intuitive theories. The causal schemata discussed in Chapter 5 may qualify as simple theories, since they specify the concepts (i.e. causal types) that exist in a domain and the law-like regularities that relate these concepts (cf. Carey (1985b)). Modeling the acquisition of more complex theories is an important challenge for the future.

Applications to other fields

Philosophers of science have long been interested in theory formation, and computational accounts of theory acquisition can address the discovery of scientific theories and intuitive theories alike. Many philosophers have argued that scientific theories occupy different levels of abstraction, and that the development of specific theories is guided by more abstract theories that are sometimes called paradigms (Kuhn, 1970) or research programs (Laudan, 1977). Henderson, Goodman, Tenenbaum, and Woodward (2007) argue that a hierarchical Bayesian approach can incorporate scientific theories at different levels of abstraction, and can help to explain how paradigms

or research programs are created and eventually abandoned.

Inductive inference is particularly relevant to the philosophy of science, but is also a topic of broader philosophical interest. The hierarchical Bayesian approach provides a general-purpose account of inductive reasoning, and can be developed as a contribution to the formal study of epistemology. In Chapter 4 I described one of the most obvious connections between the hierarchical Bayesian approach and the philosophical literature. Goodman (1955) argues that our degree of belief in specific hypotheses will often depend on more abstract overhypotheses. The model in Chapter 4 suggests how overhypotheses can be learned, and can perhaps be developed into a comprehensive formal account of Goodman's approach to induction.

Fields like machine learning and statistics can be seen as modern attempts to develop a science of inductive inference. The hierarchical Bayesian approach is widely used in both fields, and the approach in Chapter 4 is based on a well-known model—the Dirichlet-multinomial model—that has been applied to many other problems. The remaining two models may also find applications to machine learning problems. Causal models have been applied to scientific problems in many fields (Spirtes, Glymour, & Scheines, 2001), and models that incorporate causal types may be better able to capture the structure of many real-world problems. Form discovery is a problem that has previously received little attention, but automated approaches to this problem can address questions faced by biologists (Rivera & Lake, 2004; Doolittle & Bapteste, 2007), ecologists (R. H. Whittaker, 1967), linguists (Ben Hamed, 2005), psychiatrists (Waller & Meehl, 1998), and scientists from many other fields.

Theoretical challenges

My framework opens up several theoretical questions for further study. We previously saw how hierarchical models can be built by starting at a relatively concrete level and adding representations that occupy levels of increasing abstraction. This approach is a useful strategy for model-builders to pursue, but note that I provided no formal guidelines for choosing how many levels to introduce or deciding how each level is generated from the level immediately above. Automating the process of building

hierarchical models is a worthy challenge for at least two reasons. First, an automatic model builder may reduce the time needed to apply the hierarchical approach to new domains. Second and more important, an automatic model builder serves as a hypothesis about how hierarchical architectures might be constructed in the mind.

In principle, a Bayesian approach can explain how hierarchical models are learned for novel domains. The two components required are a prior distribution over a space of possible models, and a set of assumptions about how data are generated from the true underlying model. The prior over models will include a prior over architectures that captures expectations about the number of levels and the kinds of representations that are found at each level. The prior over models will also include a prior over the distributions which specify how each level is generated from the level immediately above. One method for defining this prior might make use of a set of basic elements that can be composed in many ways to construct representations and generative processes. Chapter 6 described a meta-grammar that can generate many kinds of graph structures, and an expanded meta-grammar might also be able to generate several other kinds of representations. I have not explored the possibility of a meta-grammar for generating probability distributions, but one initial step is to explore simple schemes for generating distributions that belong to the exponential family (Bishop, 2006).

An automatic model builder is the natural culmination of the modeling approach pursued in this thesis. I introduced this approach by suggesting that some of the background assumptions required by conventional models might be learned. We saw, for instance, that assumptions about the representations considered by a two-level model (Figure 2-3a) can be learned by introducing an extra level of abstraction (Figure 2-3b). Any hierarchical model will rely on its own set of background assumptions, including assumptions about the number of levels and the nature of these levels, but an automatic model builder can explain how these assumptions might be acquired. Background assumptions of some variety will still be required, but the ultimate goal is to minimize the number and specificity of these assumptions. An automatic model builder, for instance, may need to start with little more than a very general hypothesis

space of representations and probability distributions, a preference for simple models, and the ability to carry out Bayesian inference.

Implementing a fully general version of this model builder will demand solutions to many difficult technical and conceptual problems. More limited implementations of the basic idea, however, should be tractable. For instance, our model for discovering ontological kinds (Chapter 4) can be viewed as a relatively simple method for discovering the structure of a hierarchical model. Learning the number of ontological kinds amounts to learning the structure of the hierarchical model in Figure 2-2b, since there is a tree in this model for each ontological kind introduced. It may also be relatively straightforward to learn the structure of a hierarchical model when all of the conditional probability distributions are assumed to take a simple parametric form (e.g. all distributions are Gaussian), and the main problem is to decide how many levels to introduce.

A second direction for future theoretical work is to explore the relative difficulty of learning at different levels of abstraction. When hierarchical models make simultaneous inferences at multiple levels, we saw that learning can proceed faster at some levels than others. Future work can explore the conditions under which different patterns of learning should be expected. It may be possible, for instance, to identify general conditions under which learning at the upper levels will be faster than learning at the lower levels. Progress in this area is likely to be particularly relevant to the study of cognitive development. In some developmental settings, concrete knowledge is acquired before more abstract knowledge emerges, but abstract-to-concrete trajectories are observed in other cases. We saw that hierarchical Bayesian models can capture both developmental patterns, but more work is needed to understand the principles that determine which pattern applies in any given case.

Towards a modern theory of learning

Formal models have been part of psychology from the beginning (Ebbinghaus, 1885; Thurstone, 1919; Hull, 1943) and have played a central role in the development of

traditional learning theory. I identified four principles that go beyond traditional learning theory (Table 1.2b), and a modern learning theory should explore how these principles can be formally realized. This thesis focused on the first principle, and showed how the hierarchical Bayesian approach can account for learning at multiple levels of abstraction. The hierarchical Bayesian approach, however, is also consistent with the remaining principles in Table 1.2b. Bayesian models can naturally incorporate structured representations (principle two), and this thesis described models that learn graphs and causal networks. Richer representations will be needed to account for some aspects of human knowledge, but representations of arbitrary complexity can be incorporated within a Bayesian framework. Principles three and four are related: learning can succeed given sparse and noisy data as long as the learner relies on strong background knowledge. Bayesian approaches rely on prior distributions, and these priors can capture the sophisticated, domain-specific knowledge that often supports learning.

The demise of traditional learning theory was due in part to an intellectual movement that has been called the cognitive revolution (Bruner, 2004). It may be time for a second revolution that leads to a modern theory of learning, and it is possible that the computational foundations of this theory are already in place. Empirical work over several decades has described many psychological phenomena that raise challenges for traditional models of learning, but research in computer science and statistics has led to computational approaches (including the hierarchical Bayesian approach) that address some of these challenges. Young children are still much better learners than even the best machine learning systems, but modern computational techniques can help to close this gap between minds and machines.

Appendix: Form discovery model

This appendix provides some of the technical details needed to fully specify the form discovery model in Chapter 6. I also describe some issues that arise when implementing this model.

Generating structures from structural forms

The normalizing constant for the distribution in Equation 6.2 is the sum

$$\sum_{S} P(S|F) = \sum_{S \text{ is compatible with } F} \theta(1-\theta)^{|S|}.$$

To compute this quantity we must consider all possible ways of putting n entities onto a graph of form F. Let $S(n,k)$ be the Stirling number of the second kind: the number of ways to partition n elements into k nonempty sets. Let $C(F,k)$ be the number of F-structures with k occupied cluster nodes. Expressions for $C(F,k)$ for all forms except the grid and the cylinder are shown in Table 1. The number of n-entity structures with form F is

$$\sum_{k=1}^{n} S(n,k)C(F,k).$$

For all forms F except the grid and the cylinder, the normalizing constant for Equation 6.2 is

$$\sum_{S \text{ is compatible with } F} \theta(1-\theta)^{|S|} = \sum_{k=1}^{n} S(n,k)C(F,k)\theta(1-\theta)^{k}. \quad (1)$$

Form F	**C(F, k)**
Partition	1
Directed Chain	$k!$
Undirected Chain	$\frac{k!}{2}$
Order	$k!$
Connected	1
Directed Ring	$(k-1)!$
Undirected Ring	$\frac{(k-1)!}{2}$
Directed Hierarchy	k^{k-1}
Undirected Hierarchy	k^{k-2}
Tree	$(2k-5)!!$

Table 1: Number of k-cluster structures for several different forms.

Equation 1 groups the F-compatible structures into classes that share the same partition of the entities. To compute the normalizing constant for product structures like the grid and the cylinder, it is more convenient to group the F-compatible structures into classes that share the same basic topology. Let $G(n, i, j)$ be the number of ways to put n entities on an undirected i by j grid so that no dimension of the grid remains unoccupied. The normalizing constant for grids is now

$$\sum_{i \leq j \leq n} G(n, i, j)\theta(1-\theta)^{ij}.$$

Similarly, if $Y(n, i, j)$ is the number of ways to put n entities on an undirected i by j cylinder so that no dimension remains unoccupied, the normalizing constant for cylinders is

$$\sum_{i \leq n, j \leq n} Y(n, i, j)\theta(1-\theta)^{ij}.$$

$G(\cdot, \cdot, \cdot)$ can be computed using the function $L(\cdot, \cdot)$, where $L(n, i)$ is the number of ways to put n entities on an undirected i node chain so that no node remains empty:

$$L(n, i) = \begin{cases} 1 & i = 1 \\ \frac{i!}{2} S(n, i) & i > 1 \end{cases}$$

where $S(n, i)$ is the Stirling number of the second kind.

We now have

$$G(n,i,j) = \begin{cases} L(n,i)L(n,j) & i \neq j \\ \frac{L(n,i)^2 + L(n,i)}{2} & i = j \end{cases}$$

In the case where $i = j$, we have accounted for the fact that the grid can be rotated without changing the configuration.

The counts for undirected cylinders can be computed similarly. Define

$$R(n,i) = \frac{L(n,i)}{i}$$

where $R(n,i)$ is the number of ways to put n entities on an i node ring so that no node remains empty. Then

$$Y(n,i,j) = L(n,i)R(n,j).$$

Generating data from structures

Chapter 6 applies the form discovery model to feature data, similarity data, and relational data. To handle each kind of data, we define a distribution $P(D|S)$ which indicates how data D are generated from an underlying structure S.

Feature data

Suppose that S is a graph that captures the relationships between a set of entities, and that D is a feature matrix where the (i,j) entry in the matrix indicates the value of entity i on feature j. The graph provides a good account of the feature data if the features tend to be smooth over the graph: in other words, if nearby entities in the graph tend to have similar feature values. We formalize this idea by assuming that the features are generated by a Gaussian process over the graph.

Let S_{ent} be a graph with $n + l$ nodes, where the first n nodes correspond to entities and the remaining l nodes are latent. Let f be a feature vector which assigns

a continuous value $f_i \in \mathbb{R}$ to each node i in the graph. Let W be a $n+l$ by $n+l$ weight matrix, where $w_{ij} = \frac{1}{e_{ij}}$ if nodes i and j are joined by an edge of length e_{ij} and $w_{ij} = 0$ otherwise. We now define the graph Laplacian $\Delta = E - W$ where E is a diagonal matrix with entries $e_i = \sum_j w_{ij}$. A generative model for f that favors features which are smooth over the graph S_{ent} is given by

$$P(f|W) \propto \exp(-\frac{1}{4}\sum_{i,j} w_{ij}(f_i - f_j)^2) = \exp(-\frac{1}{2}f^{\mathsf{T}}\Delta f) \tag{2}$$

Zhu et al. (2003) point out that Equation 2 can be viewed as a Gaussian prior over f with zero mean and covariance matrix Δ^{-1}. This prior, however, is improper. Note that any feature vector f has the same probability when shifted by a constant, which effectively means that the variance of each f_i is infinite. We obtain a proper prior by assuming that the feature value f_i at any entity node has an *a priori* variance of σ^2:

$$f \,|\, W \sim \mathcal{N}(0, \tilde{\Delta}^{-1}) \tag{3}$$

where $\tilde{\Delta} = \Delta + V$, and V is a diagonal matrix with $\frac{1}{\sigma^2}$ appearing in the first n positions along the diagonal and 0 elsewhere.[2]

Equation 3 specifies how to generate a single feature only. Typically the data D include multiple features, and we assume that the features are conditionally independent given S_{ent}.[3] To complete the generative model we place priors on the branch lengths e_{ij} and the variance σ^2. Both are drawn from exponential distributions with

[2]Zhu et al. (2003) use a matrix V that has $\frac{1}{\sigma^2}$ everywhere along the diagonal. We prefer the approach described here because it allows empty nodes to be added to a weighted graph W without changing the likelihood $P(D|W)$. Suppose that we convert graph W to W' by adding an empty node k to the edge between i and j so that $d_{ij} = d'_{ik} + d'_{kj}$. Our model implies that $P(D|W) = P(D|W')$, but this result does not hold for the approach of Zhu et al. (2003).

[3]We treat all features equally, but it is possible to introduce weights λ^j for each feature. Equation 3 then becomes $P(f^j) \propto \exp(-\frac{\lambda^j}{2}f^{\mathsf{T}}\Delta f)$, where f^j is the jth feature. Once we place a prior on the feature weights (for example, a prior that encourages most weights to be small), we can simultaneously discover the structure S and the weights for each feature. The weights will measure the extent to which a feature is smooth over S—the features that match the structure best will end up with the highest weights.

hyperparameter β:

$$\sigma \mid \beta \qquad \sim \text{Exponential}(\beta) \tag{4}$$

$$e_{ij} \mid S_{ent}, \beta \sim \text{Exponential}(\beta) \text{ if } s_{ij} = 1 \tag{5}$$

For all analyses we set $\beta = 0.4$.

Even though we introduced edge weights w_{ij}, we are primarily interested in the best graph topology S_{ent} given the data D. The likelihood $P(D|S_{ent})$ can be computed by integrating out σ and the edge weights:

$$P(D|S_{ent}) = \int P(D|S_{ent}, W, \sigma^2) P(W|S_{ent}) P(\sigma^2) dW d\sigma^2$$

We can approximate this integral using the Laplace approximation. Since the weights w_{ij} and the variance σ are both required to be positive, we map them to a log scale before computing the Laplace approximation. To find modal values of the transformed variables, we run a gradient-based search using the 'Large Scale' option available as part of MATLAB's unconstrained minimization routine.

Throughout this section we have not been careful to distinguish between probability density functions and probability distributions. Since we defined a generative model for continuous vectors f, $P(f|W)$ should strictly be written as a probability density function $p(f|W)$. In practice, however, f is only observable to some level of accuracy, and we can quantize each feature vector:

$$P(f|W) = \int_{|f-u|<\epsilon} p(u|W) du \tag{6}$$

where ϵ is a small constant. Equation 6 can be approximated as

$$P(f|W) \approx p(f|W) \int_{|f-u|<\epsilon} du \propto p(f|W) \tag{7}$$

where the constant of proportionality does not depend on the structure or the form under consideration, and can be dropped from our calculations.

Similarity data

According to the Gaussian model in Equation 3, the probability of feature matrix D given a weighted graph W is

$$\log(P(D|W,\sigma)) = -\frac{mn}{2}\log(2\pi) - \frac{m}{2}\log|\tilde{\Delta}^{-1}| - \frac{1}{2}\mathrm{tr}(\tilde{\Delta}DD^{\mathsf{T}}).$$

Note that the feature matrix D influences this distribution only through the number of features (m) and the covariance matrix $\frac{1}{m}DD^{\mathsf{T}}$. As long as both of these components are provided, the model for feature data can be applied even if none of the actual features is observed. This insight is related to the "kernel trick" discussed by machine learning researchers (Schölkopf & Smola, 2001).

Relational data

Suppose now that the data D specify relationships between entities rather than features of the entities. We define two distributions $P(D|S)$, one for frequency data and another for binary relations.

Frequency data

Let D be a square frequency matrix with a count d_{ij} for each pair of entities (i,j). Suppose that S is a graph which specifies the relationships between a set of clusters. We define a generative model where $P(D|S)$ is high if the large entries in D correspond to edges in the cluster graph S. Formally, let $|a|$ be the number of entities in cluster a. Let C be a matrix of between-cluster counts, where C_{ab} is the total number of counts observed between entities in cluster a and entities in cluster b. Our model assumes that $P(D|S) = P(D,C|S) = P(D|C)P(C|S)$, and that C is generated from a Dirichlet-multinomial model:

$$\begin{aligned} \theta \,|\, S, \beta_0, \beta_1 &\sim \mathrm{Dirichlet}(\alpha) \\ C \,|\, \theta, n_{\mathrm{obs}} &\sim \mathrm{Multinomial}(\theta) \end{aligned}$$

where $\alpha_{ab} = \beta_0|a||b|$ if $S_{ab} = 0$, $\alpha_{ab} = \beta_1|a||b|$ if $S_{ab} = 1$, and n_{obs} is the total number of observations. The pair (β_0, β_1) is drawn from a discrete space: $\beta_0 + \beta_1$ is drawn uniformly from $\{\frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, 16, 32\}$ and $\frac{\beta_0}{\beta_0+\beta_1}$ is drawn uniformly from $\{0.05, 0.15, \ldots, 0.45\}$. A count matrix C is assigned high probability under this model if the large entries in C tend to correspond to edges in the cluster graph S.

As for the feature model, we integrate out the parameters:

$$\begin{aligned} P(C|S) &= \int P(C|S, \beta_0, \beta_1) P(\beta_0, \beta_1) d\beta_0 d\beta_1 \\ &= \frac{1}{50} \sum_{(\beta_0, \beta_1)} P(C|S, \beta_0, \beta_1) \end{aligned}$$

where

$$P(C|S, \beta_0, \beta_1) = \int P(C|\theta) p(\theta|S, \beta_0, \beta_1) d\theta$$

can be computed analytically, since the Dirichlet prior on θ is conjugate to the multinomial $P(C|\theta)$.

Given C, we assume that the C_{ab} counts are distributed at random between all pairs (i, j) where i belongs to cluster a and j belongs to cluster b:

$$P(D|C) = \prod_{a,b} \left(\frac{1}{|a||b|} \right)^{C_{ab}}.$$

Binary data

Suppose now that D is a binary relation represented as a square matrix where d_{ij} is 1 if the relation holds between i and j and 0 otherwise. We define a generative model where $P(D|S)$ is high if the large entries in D correspond to edges in the cluster graph S. Let z_i denote the cluster assignment for entity i. Suppose that there is a parameter θ_{ab} for each pair of clusters, and that d_{ij} is generated by tossing a coin with bias $\theta_{z_i z_j}$. We place a prior distribution on the parameters θ_{ab} that depends on the edges in the cluster graph, and encourages d_{ij} to be true when there is an edge

between cluster z_i and cluster z_j. The model can be written as:

$$\theta_{ab} \,|\, S, \alpha_0, \beta_0, \alpha_1, \beta_1 \sim \begin{cases} \text{Beta}(\alpha_0, \beta_0), & \text{if } S_{ab} = 0 \\ \text{Beta}(\alpha_1, \beta_1), & \text{if } S_{ab} = 1 \end{cases}$$

$$d_{ij} \,|\, \theta \sim \text{Bernoulli}(\theta_{z_i z_j})$$

The hyperparameters α_0, β_0, α_1 and β_1 are drawn from a four-dimensional grid where $\alpha_0 + \beta_0$ and $\alpha_1 + \beta_1$ belong to $\{\frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, 16, 32\}$ and $\frac{\beta_0}{\alpha_0+\beta_0}$ and $\frac{\beta_1}{\alpha_1+\beta_1}$ belong to $\{0.05, 0.15, \ldots, 0.95\}$. We sample uniformly from all points on this grid where $\frac{\beta_0}{\alpha_0+\beta_0} \leq \frac{\beta_1}{\alpha_1+\beta_1}$, which captures the assumption that relation D is most likely to be true of pairs (i, j) that correspond to edges in graph S.

As for the frequency model, we integrate out the parameters:

$$\begin{aligned} P(D|S) &= \sum_{(\alpha_0,\beta_0,\alpha_1,\beta_1)} P(D|S, \alpha_0, \beta_0, \alpha_1, \beta_1) P(\alpha_0, \beta_0, \alpha_1, \beta_1) \\ &= \sum_{(\alpha_0,\beta_0,\alpha_1,\beta_1)} P(D_0|\alpha_0, \beta_0) P(D_1|\alpha_1, \beta_1) P(\alpha_0, \beta_0, \alpha_1, \beta_1) \end{aligned}$$

where D_1 represents the entries in D that correspond to edges in the graph S, and D_0 represents the remaining entries in D. As before, the terms $P(D_0|\alpha_0, \beta_0)$ and $P(D_1|\alpha_1, \beta_1)$ are computed by integrating out θ:

$$P(D_1|\alpha_1, \beta_1) = \int P(D_1|\theta_1) p(\theta_1|\alpha_1, \beta_1) d\theta_1$$

where θ_1 is a vector containing parameters θ_{ab} for all pairs (a, b) such that there is an edge between cluster a and cluster b. $P(D_0|\alpha_0, \beta_0)$ is computed similarly.

Model implementation

The mathematical assumptions of the form discovery model have now been described, but there are some practical issues that arise when implementing this model.

Feature data

Given a matrix D with m features, we apply a linear transformation so that the mean value in D is zero, and the maximum entry in $\frac{1}{m}DD^\mathsf{T}$ is one. The first property is useful since our model assumes that the features have zero mean. The second property means that it should make sense to use the same value of the hyperparameter β for both feature and similarity data (as mentioned above, we set $\beta = 0.4$). If D contains missing entries, we group the features so that any two features in a given group are observed for precisely the same set of entities. Suppose that the largest group has j features. Consider the reduced matrix $\hat{D}$ that is created by including only these j features, and the enties for which these features are observed. We scale the data so that the mean value in D is zero, and the maximum entry in $\frac{1}{j}\hat{D}\hat{D}^\mathsf{T}$ is 1.

Our method for identifying the S and F that maximize $P(S, F|D)$ involves a separate search for each form. Since the prior on the space of forms is uniform, the winning structure is the best candidate encountered in any of these searches. Each search starts out with all the entities in a single cluster, then uses graph grammars like those in Figure 6-2 to split the entities into multiple clusters. When a cluster node is split, the entities previously assigned to this cluster must be distributed between the two new cluster nodes. We choose two of these entities at random, assign one to each of the new clusters, then go through the remaining entities in a random order, making a greedy assignment for each one. Since this procedure for splitting a cluster node is not deterministic, the search algorithm as a whole is not deterministic. At each iteration, we attempt to split each cluster node several times, and of all splits considered we accept the candidate that improves the score most. After each split, the algorithm attempts to improve the score using several proposals, including proposals that move an entity from one cluster to another, and proposals that swap two clusters. The search concludes once the score can no longer be improved.

The structures encountered early on in the greedy search can be seen as low-resolution versions of the structure that will eventually be identified as the best. This perspective explains why a greedy search will often perform well. If we take some true

structure and construct a sequence of representations at increasingly low resolutions, this sequence should provide a path by which a greedy search can proceed from the lowest-resolution version (a structure with all the entities in one cluster) to the true structure.

Relational data

A greedy search which moves from low-resolution structures to high-resolution structures should work well when fitting some structural forms (including partitions and dominance hierarchies) to relational data. For other forms, however, a greedy search will fail badly. Consider the case where the true structure is a ring, and each entity sends a link to only one other entity. There is no low-resolution version of this structure that seems acceptable: we can group the entities into clusters and organize those clusters into a ring, but the entities in each cluster will tend not to send links to the entities in the next cluster along.

When analyzing relational data, we therefore rely on two initialization strategies. The first is the strategy used for feature data: we begin with a graph where all the entities are assigned to a single cluster. The second strategy uses the best clusters found for one of the simplest structural forms: partitions with no self-links.[4] These clusters are then used to build initial configurations for each of the remaining structural forms. For example, when searching for rings, we start by connecting the two clusters with the strongest link between them. We continue adding clusters to the ends of this chain until we have a chain including all the clusters, then join the ends of this chain to create the ring that will initialize the greedy search for the best ring structure.

[4]When fitting this form, we initialize the search using the first strategy.

References

Ackley, D., Hinton, G., & Sejnowski, T. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*, 147–169.

Aldous, D. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983* (pp. 1–198). Berlin: Springer.

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

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, *6*, 1817–1853.

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

Bartlett, F. C. (1932). *Remembering: a study in experimental and social psychology*. Cambridge: Cambridge University Press.

Baxter, J. (1997). A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, *28*, 7–39.

Behrend, D. A. (1990). Constraints and development: a reply to Nelson (1988). *Cognitive Development*, *5*, 313–330.

Ben Hamed, M. (2005). Neighbour-nets portray the Chinese dialect continuum and the linguistic legacy of China's demic history. *Proceedings of the Royal Society, B*, *272*, 1015–1022.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Boeck, P. D., Wilson, M., & Acton, G. S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review*, *112*(1), 129-158.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs. 1. the method of paired comparisons. *Biometrika*, *39*, 324–345.

Bruner, J. (2004). A short history of psychological theories of learning. *Dædalus*, 13–20.

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (p. 209-254). New York: Academic Press.

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

Carey, S. (1985b). Constraints on semantic development. In J. Mehler (Ed.), *Neonate cognition* (p. 381-398). Hillsdale, NJ: Erlbaum.

Carey, S. (2004). Bootstrapping and the origin of concepts. *Dædalus*, 59–68.

Carroll, J. D. (1976). Spatial, non-spatial and hybrid models for scaling. *Psychometrika*, *41*, 439–463.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*, 41–75.

Chater, N., & Vitanyi, P. (2007). 'Ideal learning' of natural language: positive results about learning from positive evidence. *Journal of Mathematical Psychology*, *51*(3), 135–163.

Cheng, P. (1997). From covariation to causation: a causal power theory. *Psychological Review*, *104*, 367-405.

Cheng, P. W. (1993). Separating causal laws from casual facts: Pressing the limits of statistical relevance. In *The psychology of learning and motivation* (Vol. 30, p. 215-264). San Diego: Academic Press.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1975). *The logical structure of linguistic theory*. Chicago: University

of Chicago Press.

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

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

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 1–33). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Colunga, E., & Smith, L. B. (2003). The emergence of abstract ideas: evidence from networks and babies. *Philosophical Transactions of the Royal Society (B)*, *358*, 1205–1214.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: a role for associative learning. *Psychological Review*, *112*(2).

Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*(4), 297–338.

Davies, T. R., & Russell, S. J. (1987). A logical approach to reasoning by analogy. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence* (pp. 264–270).

Deák, G. O. (2000). Hunting the fox of word learning: why "constraints" fail to capture it. *Developmental Review*, *20*, 29–80.

DeJong, G., & Mooney, R. (1986). Explanation-based learning: An alternative view. *Machine Learning*, *1*(2), 145–176.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, *28*, 157–175.

desJardins, M., & Gordon, D. F. (1995). Evaluation and selection of biases in machine learning. *Machine Learning*, *20*, 1–17.

DeSoto, C. B. (1960). Learning a social structure. *Journal of Abnormal and Social Psychology*, *60*, 417-421.

Dobra, A., Jones, B., Hans, C., Nevins, J., & West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*,

90, 196–212.

Doolittle, W. F., & Bapteste, E. (2007). Pattern pluralism and the tree of life hypothesis. *Proceedings of the National Academy of Sciences*, *104*(7), 2043–2049.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). New York: Wiley.

Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory*. Cambridge, MA: MIT Press.

Ebbinghaus, H. (1885). *Über das gedächtnis: Untersuchungen zur experimentellen psychologie*. Leipzig: Duncker und Humbolt.

Ekman, G. (1954). Dimensions of color vision. *Journal of Psycholology*, *38*, 467–474.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: a connectionist perspective on development*. Cambridge, MA: MIT Press.

Engelfriet, J., & Rozenberg, G. (1997). Node replacement graph grammars. In G. Rozenberg (Ed.), *Handbook of graph grammars and computing by graph transformation* (Vol. 1). Singapore: World Scientific.

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*, 577-588.

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems 2* (pp. 524–532). San Francisco, CA: Morgan Kaufmann.

Fiske, A. P. (1992). The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological Review*, *99*, 689–723.

Flannagan, M. J., Fried, L. S., & Holyoak, K. J. (1986). Distributional expectations and the induction of category structure. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *12*(2), 241–256.

Fleishman, J. A. (1986). Types of political attitude structure: results of a cluster analysis. *The Public Opinion Quarterly*, *50*(3), 371–386.

Fodor, J. A. (1978). *Modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (1980). Fixation of belief and concept acquisition. In M. Piattelli-Palmarini (Ed.), *Language and learning: The debate between Jean Piaget and Noam Chomsky.* London: Routledge and Kegan Paul.

Freeman, L. C. (1992). Filling in the blanks: a theory of cognitive categories and the structure of social affiliation. *Social Psychology Quarterly*, *55*(2), 118-127.

French, R. M., Mermillod, M., Quinn, P. C., Chauvin, A., & Mareschal, D. (2002). The importance of starting blurry: simulating improved basic-level category learning in infants due to weak visual acuity. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 322–327).

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193–202.

Gallistel, C. R. (2000). The replacement of general-purpose learning models with adaptively specialized learning modules. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (2nd ed., pp. 1179–1191). Cambridge, MA: MIT Press.

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

Gelman, R., Bullock, M., & Meck, E. (1980). Preschoolers' understanding of simple object transformations. *Child Development*, *51*, 691–699.

Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number.* Cambridge, MA: Harvard University Press.

Gelman, R., & Williams, E. M. (1998). Enabling constraints for cognitive development and learning: domain specificity and epigenesis. In R. M. Lerner (Ed.), *Theoretical models of human development* (Vol. II). New York: Wiley.

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

George, D., & Hawkins, J. (2005). A hierarchical Bayesian model of invariant pattern recognition in the visual cortex. In *Proceedings of the IEEE International Joint Conference on Neural Networks* (Vol. 3, pp. 1812–1817).

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling.

Journal of the American Statistical Association, *88*, 881-889.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium Interface* (pp. 156–163).

Gildea, D., & Jurafsky, D. (1996). Learning bias and phonological rule induction. *Computational Linguistics*, *22*, 497–530.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*, 7821–7826.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Goldstone, R. L., & Johansen, M. K. (2003). Conceptual development from origins to asymptotes. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development* (pp. 403–418). New York: Oxford University Press.

Good, I. J. (1980). Some history of the hierarchical Bayesian methodology. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 489–519). Valencia: Valencia University Press.

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

Gopnik, A. (1996). The scientist as child. *Philosophy of Science*, *63*, 485-514.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Gopnik, A., & Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, *71*, 1205-1222.

Greenwald, A. G. (1988). Levels of representation. (Unpublished manuscript)

Griffiths, T. L. (2005). *Causes, coincidences, and theories*. Unpublished doctoral dissertation, Stanford University.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

Grofman, B., & Brazill, T. (2002). Identifying the median justice on the supreme

court through multidimensional scaling: Analysis of "natural courts" 1953–1991. *Public Choice*, *112*, 55–79.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139–150.

Guttman, L. (1954). A new approach to factor analysis: the radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 258–348). Glencoe, IL: Free Press.

Hage, P., & Harary, F. (1991). *Exchange in Oceania: A graph theoretic analysis*. Oxford: Oxford University Press.

Han, F., & Zhu, S. C. (2005). Bottom-up/top-down image parsing by attribute graph grammar. In *10th IEEE International Conference on Computer Vision* (pp. 1778–1785). IEEE Computer Society.

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, *56*, 51–65.

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

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

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

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

Hempel, C. G. (1972). *Fundamentals of concept formation in empirical science*. University of Chicago Press.

Henderson, L., Goodman, N., Tenenbaum, J., & Woodward, J. (2007). Frameworks in science: a Bayesian approach. In *LSE-Pitt conference: Confirmation, Induction and Science.*

Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: how intelligent inferences look like reasoning errors. *Journal of Behavioral Decision*

Making, *12*, 275–305.

Hilgard, E. R., & Bower, G. H. (Eds.). (1975). *Theories of learning* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hinton, G. E., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527–1554.

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

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754-755.

Hull, C. (1943). *Principles of behavior*. Appleton-Century-Crofts.

Hume, D. (1748). *An enquiry concerning human understanding*. Indianapolis, IN: Hackett.

Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: the role of shape similarity in early acquisition. *Cognitive Development*, *9*, 45–76.

Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child*. London: Routledge & Kegan Paul.

Jain, S., & Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet Process mixture model. *Journal of Computational and Graphical Statistics*, *13*, 158-182.

Jones, D. (2003). The generative psychology of kinship. Part 1. Cognitive universals and evolutionary psychology. *Evolution and Human Behavior*, *24*, 303–319.

Jones, S. S., & Smith, L. B. (2002). How children know the relevant properties for generalizing object names. *Developmental Science*, *5*(2), 219–232.

Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development*, *62*, 499–516.

Jusczyk, P. W. (2003). Chunking language input to find patterns. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development* (pp. 17–49). New York: Oxford University Press.

Kant, I. (2003). *Critique of pure reason*. New York: Palgrave Macmillan. (Translated by N. Kemp Smith)

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

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

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

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

Keil, F. C. (1990). Constraints on constraints: Surveying the epigenetic landscape. *Cognitive Science*, *14*, 135–168.

Keil, F. C. (1991). The emergence of theoretical beliefs as constraints on concepts. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Keil, F. C. (1998). Cognitive science and the origins of thought and knowledge. In R. M. Lerner (Ed.), *Theoretical models of human development* (Vol. I). New York: Wiley.

Kelley, H. H. (1972). Causal schemata and the attribution process. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. S. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: perceiving the causes of behavior* (p. 151-174). Morristown, NJ: General Learning Press.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*, 107-128.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kemp, C., & Tenenbaum, J. B. (2003). Theory-based induction. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 658–663).

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence.*

Kimble, G. A. (1961). *Hilgard and Marquis' conditioning and learning* (2nd ed.).

Prentice-Hall.

Kohonen, T. (1997). *Self-organizing maps*. New York: Springer-Verlag.

Kok, S., & Domingos, P. (2005). Learning the structure of Markov logic networks. In *Proceedings of the 22nd International Conference on Machine Learning.*

Kubica, J., Moore, A., Schneider, J., & Yang, Y. (2002). Stochastic link and group detection. In *Proceedings of the 17th National Conference on Artificial Intelligence.*

Kubovy, M., & Van Valkenburg, D. (2001). Auditory and visual objects. *Cognition*, *80*(1), 97–126.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Kuipers, B., Beeson, P., Modayil, J., & Provost, J. (2006). Bootstrap learning of foundational representations. *Connection Science*, *18*(2).

Landau, B., Gleitman, H., & Spelke, E. (1981). Spatial knowledge and geometric representation in a child blind from birth. *Science*, *213*, 1275–1278.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*, 299–321.

Laudan, L. (1977). *Progress and its problems*. Berkeley, CA: University of California Press.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, *20*(7), 1434–1448.

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

Leyton, M. (1992). *Symmetry, causality, mind*. Cambridge, MA: MIT Press.

Lien, Y., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology*, *40*, 87-137.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society B*, *34*(1), 1–41.

Locke, J. (1998). *An essay concerning human understanding* (R. S. Woolhouse, Ed.). Penguin Books.

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

Lovejoy, A. O. (1970). *The great chain of being*. Cambridge, MA: Harvard University Press.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*, 1432–1438.

Macario, J. F., Shipley, E. F., & Billman, D. O. (1990). Induction from a single instance: formation of a novel category. *Journal of Experimental Child Psychology*, *50*, 179–199.

MacRae, J. (1960). Direct factor analysis of sociometric data. *Sociometry*, *22*, 360-371.

Madole, K. L., & Cohen, L. B. (1995). The role of object parts in infants' attention to form-function correlations. *Developmental Psychology*, *31*(4), 637–648.

Malinowski, B. (1922). *Argonauts of the Western Pacific: An account of native enterprise and adventure in the archipelagoes of Melanesian New Guinea*. London: G. Routledge & Sons.

Mandler, J. M. (2003). Conceptual categorization. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development* (pp. 103–131). New York: Oxford University Press.

Mansinghka, V. K., Kemp, C., Tenenbaum, J. B., & Griffiths, T. L. (2006). Structured priors for structure learning. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI).*

Mareschal, D., & Shultz, T. R. (1996). Generative connectionist networks and constructivist cognitive development. *Cognitive Development*, *11*, 571–603.

Markman, E. (1989). *Naming and categorization in children*. Cambridge, MA: MIT Press.

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

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes

in speech perception? *Trends in Cognitive Science*, *10*(8), 363–369.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 109–165). New York: Academic Press.

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

Medin, D., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge: Cambridge University Press.

Medin, D. L., Ahn, W., Bettger, J., Florian, J., Goldstone, R., Lassaline, M., et al. (1990). Safe takeoffs—soft landings. *Cognitive Psychology*, *14*, 169–178.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness and category construction. *Cognitive Psychology*, *19*, 242–279.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178.

Milch, B., Marthi, B., Russell, S., Sontag, D., Ong, D. L., & Kolobov, A. (2005). BLOG: Probabilistic models with unknown objects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (pp. 1352–1359).

Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.), *The psychology of computer vision* (pp. 211–277). New York, NY: McGraw-Hill.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw Hill.

Mitchell, T. M., Keller, R., & Kedar-Cabelli, S. (1986). Explanation-based learning: A unifying view. *Machine Learning*, *1*(1), 47–80.

Moray, N. (1990). A lattice theory approach to the structure of mental models. *Philosophical Transactions of the Royal Society (B)*, *327*, 577–583.

Muggleton, S., & De Raedt, L. (1994). Inductive logic programming: theory and

methods. *Journal of Logic Programming*, *19-20*, 629–679.

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

Neal, R. (2003). Density modeling and clustering using Dirichlet diffusion trees. In J. M. Bernardo et al. (Eds.), *Bayesian Statistics 7* (pp. 619–629). Oxford: Oxford University Press.

Neal, R. M. (1996). *Bayesian learning for neural networks* (No. 118). New York: Springer-Verlag.

Nelson, K. (1988). Constraints on word learning. *Cognitive Development*, *3*, 221–246.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11–28.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*(4), 339–363.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *Behavioral and Brain Sciences*, *23*, 299–370.

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

Novick, L. R. (1990). Representational transfer in problem solving. *Psychological Science*, *1*, 128–132.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal inference. *Psychological Review*, *111*, 455-485.

Nowicki, K., & Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, *96*(455), 1077–1087.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608-631.

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

Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge, UK: Cambridge University Press.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*, 559–572.

Peirce, C. S. (1931–1935). *Collected papers of Charles Sanders Peirce* (Vol. 1-6; C. Hartshorne & P. Weiss, Eds.). Cambridge, MA: Harvard University Press.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2006). Poverty of the stimulus? A rational approach. In *Proceedings of the 28th Annual Conference of the Cognitive Science society* (pp. 663–668).

Piaget, J. (1965). *The child's conception of number*. New York: Norton.

Piaget, J. (1970). *Genetic epistemology*. Columbia University Press.

Piaget, J., & Inhelder, B. (1969). *The psychology of the child*. New York: Basic Books.

Polya, G. (1990). *Mathematics and plausible reasoning*. Princeton, NJ: Princeton University Press.

Popper, K. R. (1935/1980). *The logic of scientific discovery*. Boston, MA: Hutchinson.

Prince, A., & Smolensky, P. (1993). *Optimality theory: constraint interaction in generative grammar* (Tech. Rep. No. RuCCS-TR-2). Rutgers University.

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

Purves, W. K., Sadava, D., Orians, G. H., & Heller, H. C. (2001). *Life: The science of biology* (6th ed.). Sunderland, MA: Sinauer Associates.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Quine, W. V. O. (1969). Epistemology naturalized. In *Ontological relativity and other essays*. Columbia University Press.

Quinlan, J. R. (1990). Learning logical definitions from relations. *Machine Learning*, *5*(3), 239–266.

Range, F., & Noë, R. (2002). Familiarity and dominance relations among female sooty mangabeys in the Tai national park. *American Journal of Primatology*, *56*, 137-153.

Regier, T. (1996). *Spatial language and constrained connectionism*. Cambridge, MA: MIT Press.

Regier, T. (2003). Emergent constraints on word-learning: A computational review. *Trends in Cognitive Science*, *7*, 263–268.

Reichenbach, H. (1938). *Experience and prediction*. Chicago: University of Chicago Press.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (p. 21-59). Cambridge: Cambridge University Press.

Rivera, M. C., & Lake, J. A. (2004). The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, *431*, 152–155.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: a Parallel Distributed Processing approach*. Cambridge, MA: MIT Press.

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

Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs. In J. McClelland, D. Rumelhart, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2). Cambridge, MA: MIT Press.

Rumelhart, D. E. (1980). Schemata: the building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33–58). Hillsdale, NJ: Lawrence Erlbaum Associates.

Russell, S. J., & Norvig, P. (2002). *Artificial Intelligence: A modern approach* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, *133*(4), 534–553.

Samuelson, L. (2002). Statistical regularities in vocabulary guide language acquisition in connectionist models and 15-20 month olds. *Developmental Psychology*,

38(6), 1016–1037.

Samuelson, L., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition*, *73*, 1–33.

Schaffer, C. (1994). A conservation law for generalization performance. In *Proceedings of the 11th International Conference on Machine Learning* (pp. 259–265).

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: an inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schmidt, L. A., Kemp, C., & Tenenbaum, J. B. (2006). Nonsense and sensibility: Discovering unseen possibilities. In *Proceedings of the 28th Annual Conference of the Cognitive Science society* (pp. 744–749).

Schölkopf, B., & Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.

Schyns, P., Goldstone, R. L., & Thilbaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, *21*, 1-54.

Shafto, P., Kemp, C., Mansinghka, V., Gordon, M., & Tenenbaum, J. B. (2006). Learning cross-cutting systems of categories. In *Proceedings of the 28th Annual Conference of the Cognitive Science society* (pp. 2146–2151).

Shanks, D. R. (2006). Bayesian associative learning. *Trends in Cognitive Science*, *10*(11), 477–478.

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*(4), 405–415.

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

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

Shipley, E. F. (1993). Categories, hierarchies and induction. *Psychology of Learning and Motivation*, *30*, 265–301.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for*

Research in Child Development, *47*(1), 1–51.

Shultz, T. R. (2003). *Computational developmental psychology*. Cambridge, MA: MIT Press.

Shultz, T. R., & Vogel, A. (2004). A connectionist model of the development of transitivity. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 1243–1248).

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

Simpson, J., & Weiner, E. (Eds.). (1989). *Oxford English Dictionary* (2nd ed.). Clarendon Press.

Smith, L. B., Jones, S. S., & Landau, B. (1996). Naming in young children: a dumb attentional mechanism? *Cognition*, *60*, 143–171.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1–2), 159–216.

Sneath, P. H., & Sokal, R. R. (1973). *Numerical taxonomy: the principles and practice of numerical classification*. San Francisco: W. H. Freeman.

Soja, N., Carey, S., & Spelke, E. (1991). Ontological categories guide young children's inductions of word meanings. *Cognition*, *38*, 179–211.

Solomonoff, R. J. (1978). Complexity-based induction systems: comparisons and convergence theorems. *IEEE Transactions on Information Theory*, *24*, 422–432.

Sommers, F. (1963). Types and ontology. *Philosophical Review*, *72*, 327–363.

Spaeth, H. J. (2005). *United States Supreme Court judicial database, 1953-2005 terms*.

Spearman, C. E. (1904). 'General intelligence' objectively determined and measured.

American Journal of Psychology, *5*, 201–293.

Spelke, E. (1994). Initial knowledge: six suggestions. *Cognition*, *50*, 431–445.

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

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction and search* (2nd ed.). Cambridge, MA: MIT Press.

Suppes, P. (1966). Concept formation and Bayesian decisions. In J. Hintikka & P. Suppes (Eds.), *Aspects of inductive logic* (pp. 21–48). North-Holland.

Taskar, B., Segal, E., & Koller, D. (2001). Probabilistic classification and clustering in relational data. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence* (pp. 870–876).

Tenenbaum, J. B., & Griffiths, T. L. (2001). The rational basis of representativeness. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 1064–1069).

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

Tenenbaum, J. B., & Niyogi, S. (2003). Learning causal laws. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 1152–1157).

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

Thelen, E., & Smith, L. B. (Eds.). (1996). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Thrun, S., & Pratt, L. (Eds.). (1998). *Learning to learn*. Norwell, MA: Kluwer.

Thurstone, L. L. (1919). The learning curve equation. *Psychological Monographs*, *26*(3).

Torgeson, W. S. (1965). Multidimensional scaling of similarity. *Psychometrika*, *30*, 379–393.

Turkewitz, G., & Kenny, P. A. (1982). Limitations on input as a basis for neural organization and perceptual development: a preliminary theoretical statement.

Developmental Psychobiology, *15*(4), 357–368.

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

Ullman, S. (1995). Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cerebral Cortex*, *5*(1), 1–11.

Waldmann, M. R., & Hagmayer, Y. (2006). Categories and causality: the neglected direction. *Cognitive Psychology*, *53*, 27–58.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: distinguishing types from continua*. Thousand Oaks, CA: Sage.

Wang, Y. J., & Wong, G. Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, *82*, 8–19.

Ward, T. B., Becker, A. H., Hass, S. D., & Vela, E. (1991). Attribute availability and the shape bias in children's category generalization. *Cognitive Development*, *6*, 143–167.

Watanabe, S. (1969). *Knowing and guessing: a quantitative study of inference and information*. New York: Wiley.

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

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

Wellman, H. M., & Gelman, S. A. (1998). Knowledge acquisition in foundational domains. In D. Kuhn & R. S. Siegler (Eds.), *Handbook of child psychology* (pp. 523–573).

White, C. (2001). An account of the regular gradation in man. In R. Bernasconi (Ed.), *Concepts of race in the eighteenth century* (Vol. 8, pp. 4–190). Bristol: Thoemmes Press.

White, H. C., Boorman, S. A., & Breiger, R. L. (1976). Social structure from multiple networks. *American Journal of Sociology*, *81*, 730–780.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Chichester, UK: Wiley.

Whittaker, R. H. (1967). Gradient analysis of vegetation. *Biological Review*, *4*, 207–264.

Wiggins, J. S. (1996). An informal history of the interpersonal circumplex tradition. *Journal of Personality Assessment*, *66*, 217–233.

Wilcox, C., & Clausen, A. (1991). The dimensionality of roll-call voting reconsidered. *Legislative Studies Quarterly*, *16*(3), 393–406.

Wolpert, D. H. (1995). The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework. In D. H. Wolpert (Ed.), *The mathematics of generalization* (pp. 117–214). Reading, MA: Addison-Wesley.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, *358*, 749–750.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2).

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