

Running head: THEORY FORMATION

A probabilistic model of theory formation

Charles Kemp

Department of Psychology

Carnegie Mellon University

Joshua B. Tenenbaum

Department of Brain and Cognitive Sciences

Massachusetts Institute of Technology

Sourabh Niyogi

Department of Electrical Engineering and Computer Science

Massachusetts Institute of Technology

Thomas L. Griffiths

Department of Psychology

University of California, Berkeley

Word count: 17,800

## Abstract

Concept learning is challenging in part because the meanings of many concepts depend on their relationships to other concepts. Learning these concepts in isolation can be difficult, but we present a model that discovers entire systems of related concepts. These systems can be viewed as simple theories that specify the concepts that exist in a domain, and the laws or principles that relate these concepts. We apply our model to several real-world problems, including learning the structure of kinship systems and learning ontologies. We also compare its predictions to data collected in two behavioral experiments. Experiment 1 shows that our model helps to explain how simple theories are acquired and used for inductive inference. Experiment 2 suggests that our model provides a better account of theory discovery than a more traditional alternative that focuses on features rather than relations.

## A probabilistic model of theory formation

Parent: A person who has begotten or borne a child.

Child: The offspring, male or female, of human parents.

*The Oxford English Dictionary*, 2nd edition, 1989.

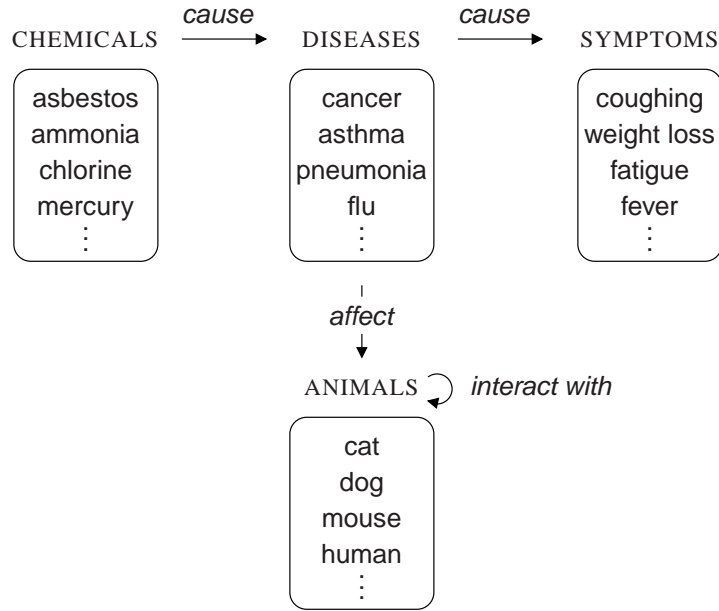
Samuel Johnson acknowledges that his dictionary of 1755 is far from perfect, but suggests that “many seeming faults are to be imputed rather to the nature of the undertaking, than the negligence of the performer.” He argues, for instance, that “some explanations are unavoidably reciprocal or circular, as *hind, the female of the stag; stag, the male of the hind.*” Analogies between dictionary definitions and mental representations can only extend so far, but Johnson appears to have uncovered a general truth about the structure of human knowledge. Scholars from many disciplines have argued that concepts are organized into systems of relations, and that the meaning of a concept depends in part on its relationships to other concepts (Quillian, 1968; Field, 1977; Quine & Ullian, 1978; Carey, 1985; Block, 1986; Goldstone & Rogosky, 2002). To appreciate the basic idea, consider pairs of concepts like parent and child, disease and symptom, or life and death. In each case it is difficult to imagine how a learner could fully understand one member of the pair without also understanding the other. Systems of concepts, however, are often much more complex than mutually dependent pairs. Concepts like life and death, for instance, are embedded in a system that also includes concepts like growth, eating, energy and reproduction (Carey, 1985).

Systems of concepts capture some important aspects of human knowledge also raise some challenging puzzles (J. Fodor & Lepore, 1992). Here we mention just two. First, it is natural to think that many concepts (including dog, tree and electron) are shared by many members of our society, but if the meaning of any concept depends on its role within an entire conceptual system, it is hard to understand how two individuals with different beliefs (and therefore different

conceptual systems) could have any concepts in common (J. Fodor & Lepore, 1992). Second, a holistic approach to concept meaning raises a difficult acquisition problem. If the meaning of each concept depends on its role within a system of concepts, it is difficult to see how a learner might break into the system and acquire the concepts that it contains (Hempel, 1985; Woodfield, 1987). Goldstone and Rogosky (2002) recently presented a formal model that helps to address the first puzzle, and here we present a computational approach that helps to address the second puzzle.

Following prior usage in psychology (Carey, 1985) and artificial intelligence (Davis, 1990) we use the term *theory* to refer to a system that specifies a set of concepts and relationships between these concepts. Scientific theories are paradigmatic examples of the systems we will consider, but psychologists have argued that everyday knowledge is organized into intuitive theories that are similar to scientific theories in many respects (Carey, 1985; Murphy & Medin, 1985; Wellman & Gelman, 1992). Both kinds of theories are believed to play several important roles. As we have already seen, theories help to individuate concepts, and many kinds of concepts derive their meaning from the roles they play in theories. Theories allow learners to explain existing observations, and to make predictions about new observations. Finally, theories guide inductive inferences by restricting a learner’s attention to features and hypotheses that are relevant to the task at hand.

Theories may take many different forms, and the examples we focus on are related to the “framework theories” described by Wellman and Gelman (1992). Framework theories specify the fundamental concepts that exist in a domain and the possible relationships between these concepts. A framework theory of medicine, for example, might indicate that two of the fundamental concepts are chemicals and diseases, and that chemicals can *cause* diseases (Figure 1). A “specific theory” is a more detailed account of the phenomena in some domain, and is typically constructed from concrete instances of the abstract categories provided by the framework theory. Extending our medical example, a specific theory might indicate that asbestos can cause lung cancer, where asbestos is a chemical and lung cancer is a disease. The framework



*Figure 1.* A fragment of a medical framework theory. The theory specifies four abstract concepts (chemicals, diseases, symptoms, and animals), and states for instance that asbestos is a chemical and that cancer is a disease. The theory also specifies relationships between these four concepts—for instance, chemicals cause diseases, and diseases affect animals.

theory therefore suggests that any specific correlation between asbestos exposure and lung cancer is better explained by a causal link from asbestos to lung cancer than a link in the opposite direction. Although researchers should eventually aim for models that can handle both framework theories and specific theories, working with framework theories is a useful first step. Framework theories are important since they capture some of our most fundamental knowledge, and in some cases they appear simple enough that we can begin to think about them computationally.

Three fundamental questions can be asked about theories: what are they, how are they used to make inductive inferences, and how are they acquired? Philosophers and psychologists have addressed all three questions (Popper, 1935/1980; Kuhn, 1970; Hempel, 1972; Carey, 1985; Wellman & Gelman, 1998), but there have been few attempts to provide computational answers to these questions. Our work takes an initial step in this direction: we consider only relatively simple theories, but we specify these theories formally, we use these theories to make predictions

about unobserved relationships between entities, and we show how these theories can be learned from raw relational data.

The first of our three fundamental questions requires us to formalize the notion of a theory. We explore the idea that framework theories can be represented as a probabilistic model which includes a set of categories and a matrix of parameters specifying relationships between those categories. Representations this simple will only be able to capture some aspects of framework theories, but working with simple representations allows us to develop tractable answers to our remaining two questions.

The second question asks how theories can be used for inductive inference. Each of our theories specifies the relationships between categories that are possible or likely, and predictions about unobserved relationships between entities are guided by inductive inferences about their category assignments. Since we represent theories as probabilistic models, Bayesian inference provides a principled framework for inferences about category assignments, relationships between categories and relationships between entities.

The final question—how are theories acquired?—is probably the most challenging of the three. Some philosophers suggest that this question will never be answered, and that there can be “no systematic, useful study of theory construction or discovery” (Newton-Smith, 1981, p 125). To appreciate why theory acquisition is challenging, consider a case where the concepts belonging to a theory are not known in advance. Imagine a child who stumbles across a set of identical-looking metal objects. She starts to play with these objects and notices that some pairs seem to exert mysterious forces on each other when they come into close proximity. Eventually she discovers that there are three kinds of objects—call them magnets, magnetic objects and non-magnetic objects. She also discovers causal laws that capture the relationships between these concepts: magnets interact with magnets and magnetic objects, magnetic objects interact only with magnets, and non-magnetic objects do not interact with any other objects. Notice that the three hidden concepts and the causal laws are tightly coupled. The causal laws are only defined in

terms of the concepts, and the concepts are only defined in terms of the causal relationships between them. This coupling raises a challenging learning problem. If the child already knew about the three concepts—suppose, for instance, that different kinds of objects were painted different colors—then discovering the relationships between the concepts would be simple. Similarly, a child who already knew the causal laws should find it easy to group the objects into categories. We consider the case where neither the concepts nor the causal laws are known. In general, a learner may not even know when there are new concepts to be discovered in a particular domain, let alone how many concepts there are or how they relate to one another. The approach we describe attempts to solve all of these acquisition problems simultaneously.

We suggested already that Bayesian inference can explain how theories are used for induction, and our approach to theory acquisition is founded on exactly the same principle. Given a formal characterization of a theory, we can set up a space of possible theories and define a prior distribution over this space. Bayesian inference then provides a normative strategy for selecting the theory in this space that is best supported by the available data. Many Bayesian accounts of human learning work with relatively simple representations, including regions in multidimensional space and sets of clusters (Shepard, 1987; Anderson, 1991; Tenenbaum & Griffiths, 2001). Our model demonstrates that the Bayesian approach to knowledge acquisition can be carried out even when the representations to be learned are richly structured, and are best described as relational theories.

Our approach draws on previous proposals about relational concepts and on existing models of categorization. Several researchers (Markman & Stilwell, 2001; Gentner & Kurtz, 2005) have emphasized that many concepts derive their content from their relationships to other concepts, but there have been few formal models that explain how these concepts might be learned (Markman & Stilwell, 2001). Most models of categorization take features as their input, and are able only to discover categories defined by characteristic patterns of features (Medin & Schaffer, 1978; Nosofsky, 1986; Anderson, 1991). Our approach brings these two research

programs together. We build on formal techniques used by previous models—in particular, our approach extends Anderson’s rational model of categorization—but we go beyond existing categorization models by working with rich systems of relational data.

The next two sections introduce the simple kinds of theories that we consider in this paper. We then describe our formal approach and evaluate it in two ways. First we demonstrate that our model learns large-scale theories given real-world data that roughly approximate the kind of information available to human learners. In particular, we show that our model discovers theories related to folk biology and folk sociology, and a medical theory that captures relationships between ontological concepts. We then turn to empirical studies and describe two behavioral experiments where participants learn theories analogous to the simple theory of magnetism already described. Our model helps to explain how these simple theories are learned and used to support inductive inferences, and we show that our relational approach explains our data better than a feature-based model of categorization.

### **Theories and theory discovery**

“Theory” is a term that is used both formally and informally across a broad range of disciplines, including psychology, philosophy, and computer science. No definition of this term is universally adopted, but here we work with the idea that a theory is a structured system of concepts that explains some existing set of observations and predicts future observations. In the magnetism example just described, the concepts are magnets, magnetic objects, and non-magnetic objects, and these concepts are embedded in a system of relations that specifies, for instance, that magnets interact with magnets and magnetic objects but not with non-magnetic objects. This system of relationships between concepts helps to explain interactions between specific objects in terms of general laws: for example, bars 4 and 11 interact because both are magnets, and because magnets always interact with each other. The magnetism theory also supports predictions about pairs of objects (e.g. bars 6 and 11) that are brought together for the

first time: for example, these objects might be expected to interact because previous observations suggest that both are magnets.

The definition just proposed highlights several aspects of theories that have been emphasized by previous researchers. There is broad agreement that theories should explain and predict data, and the idea that a theory is a system of relationships between concepts is also widely accepted. Newton’s second law of motion, for example, is a system ( $F = ma$ ) that establishes a relationship between the concepts of force, mass, and acceleration. In the psychological literature, Carey (1985) has suggested that 10 year olds have an intuitive theory of biology that specifies relationships between concepts like life, death, growth, eating, and reproduction—for instance, that death is the termination of life, and that eating is necessary for growth and for the continuation of life.

Our definition of “theory” is consistent with many previous treatments of this notion, but leaves out some elements that have been emphasized in previous work. The most notable omission is the idea of causality. For us, a theory specifies relationships between concepts that are often but not always causal. This view of theories has some precedent in the psychological literature (Rips, 1995) and is common in the artificial intelligence literature, where mathematical theories are often presented as targets for theory-learning systems (Shapiro, 1991). A second example of a non-causal theory is a system that specifies relationships between kinship concepts: for example, the fact that the *sister* of a *parent* is an *aunt* (Quinlan, 1990). Although the kinship domain is one of the cases we consider, we also apply our formal approach to several settings where the underlying relationships are causal, including an experimental setting inspired by the magnets scenario already described.

Although our general approach is broadly consistent with previous discussions of intuitive theories, it differs sharply from some alternative accounts of conceptual structure. Many formal approaches to categorization and concept learning focus on features rather than relations, and assume that concepts correspond to sets, lists, or bundles of features. We propose that

feature-based representations are not rich enough to capture the structure of human knowledge, and that many concepts derive their meanings from the roles they play in relational systems. To emphasize this distinction between feature-based and relational approaches, we present a behavioral experiment that directly compares our relational model with a feature-based alternative that is closely related to Anderson’s rational model of categorization (Anderson, 1991).

Now that we have introduced the notion of a “theory” our approach to theory discovery can be summarized. Suppose that we wish to explain the phenomena in some domain. Any theory of the domain can be regarded as a representation: a complex structured representation, but a representation nonetheless. Suppose that we have a set of these representations: that is, a hypothesis space of theories. Each of these theories makes predictions about the phenomena in the domain, and suppose that we can formally specify which phenomena are likely to follow if a given theory is true. Suppose also that we have a prior distribution on the hypothesis space of theories: for example, perhaps the simpler theories are considered more likely *a priori*. Theory discovery is now a matter of choosing the element in the hypothesis space that allows the best trade-off between explanatory accuracy and *a priori* plausibility. As we will demonstrate, this choice can be formalized as a statistical inference.

### Learning simple theories

Before introducing the details of our model, we describe the input that it takes and the output that it generates and provide an informal description of how it converts the input to the output. The input for each problem specifies relationships among the entities in a domain, and the output is a simple theory that we refer to as a *relational system*. Each relational system organizes a set of entities into categories and specifies the relationships between these categories. Suppose, for instance, that we are interested in a set of metal bars, and are given a relation  $interacts(x_i, x_j)$  which indicates whether bars  $x_i$  and  $x_j$  interact with each other. This relation can be represented as the matrix in Figure 2a.i, where entry  $(i, j)$  in the matrix is black if bar  $x_i$

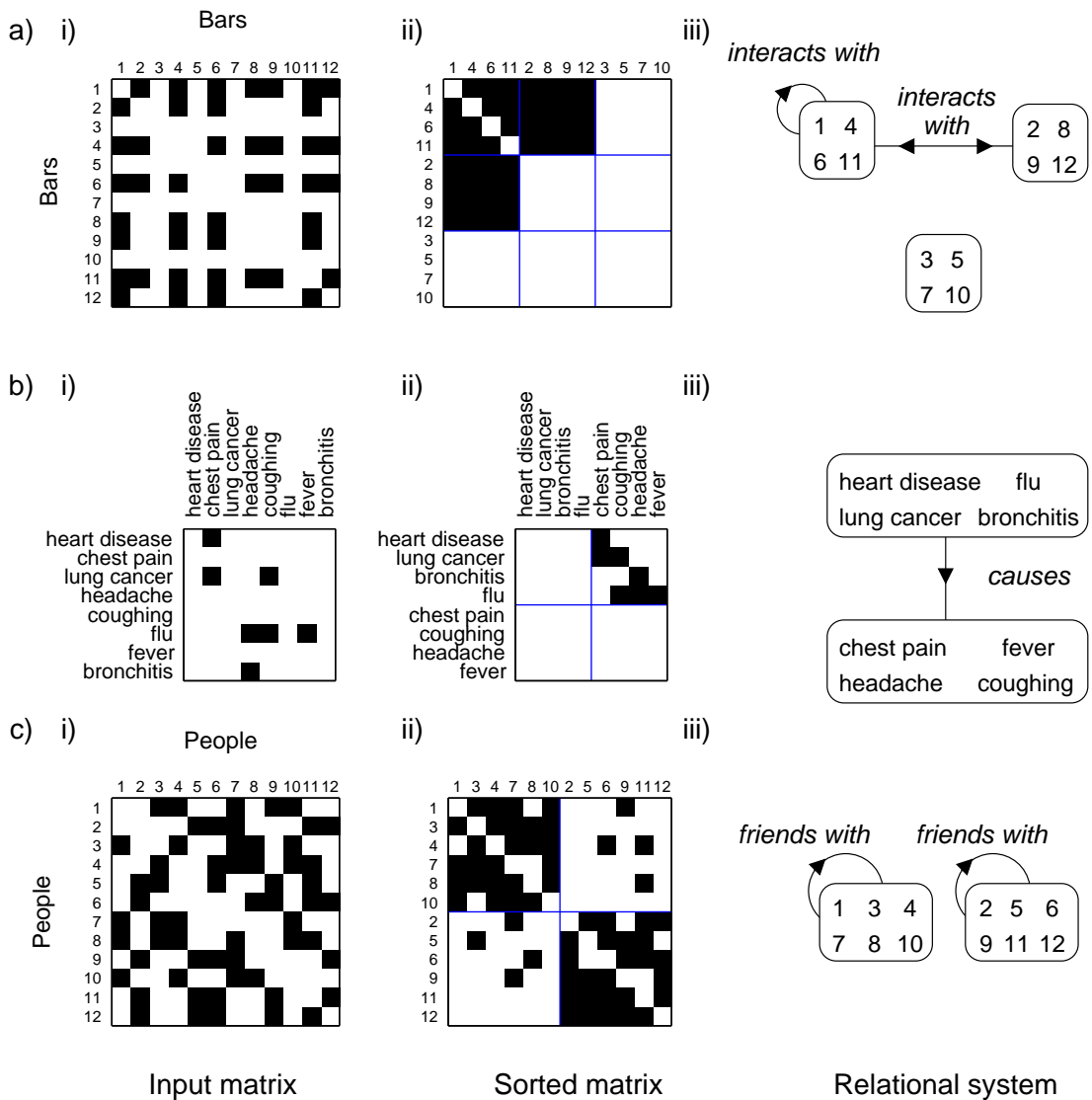


Figure 2. Discovering three simple theories. (a)(i) A relation specifying which metal bars interact with each other. Entry  $(i, j)$  in the matrix is black if bar  $x_i$  interacts with bar  $x_j$ . (ii) The relation in (i) takes on a clean block structure when the bars are sorted into three categories (magnets, magnetic objects, and non-magnetic objects). (iii) A relational system that assigns each bar to one of three categories, and specifies the relationships between these categories. (b) Learning a simple medical theory. The input data specify which entities cause which other entities. The relational system organizes the entities into two categories—diseases and symptoms—and indicates that diseases cause symptoms. (c) Learning a simple social theory. The input data specify which people are friends with each other. The relational system indicates that there are two categories, and that individuals tend to be friends with others from the same category.

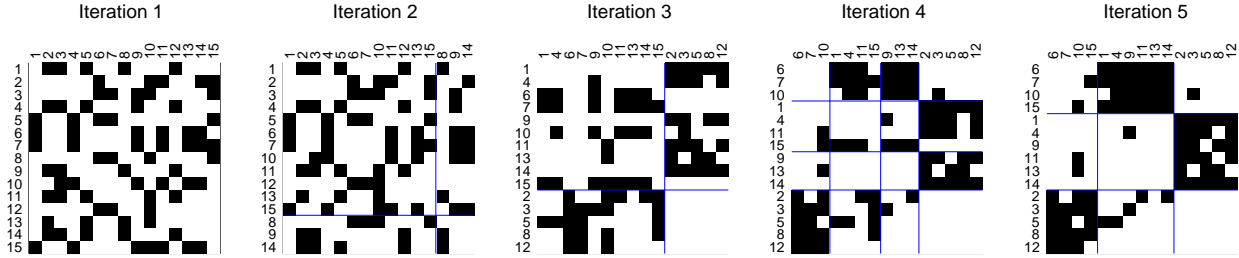


Figure 3. Category assignments explored as our model searches for the system that best explains a binary relation. The goal of the search algorithm is to organize the objects into categories so that the relation assumes a clean block structure.

interacts with bar  $x_j$ . Given this matrix as input, our model discovers the relational system in Figure 2a.iii. The system organizes the bars into three categories—magnets, magnetic objects and non-magnetic objects—and specifies the relationships between these categories. Figure 2a.ii shows that the input matrix takes on a clean block structure when sorted according to the categories discovered by our model. This clean block structure reflects the lawful relationships between the categories discovered by the model. For example, the all-black block in the top row of Figure 2a.ii indicates that every object in category 1 (i.e. every magnet) interacts with every object in category 2 (i.e. every magnetic object).

Our approach is based on a mathematical function that can be used to assess the quality of any relational system. Roughly speaking, the scoring function assigns a high score to a relational system only if the input data take on a clean block structure when sorted according to the system. Given this scoring function, theory discovery can be formulated as the problem of searching through a large space of relational systems in order to find the highest-scoring candidate. This search can be visualized as an attempt to shuffle the rows and the columns of the input matrix so that the matrix takes on a clean block structure. Figure 3 shows an example where the input matrix on the left is sorted over a number of iterations to reveal the block-structured matrix on the right. The matrix on the far right contains the same information as the input matrix, but shuffling the rows and the columns reveals the existence of a relational system involving three categories (call them A, B and C). The final matrix shows that these categories are organized into

a ring, and that the relation of interest tends to be found from members of category A to members of category B, from B-members to C-members, and from C-members to A-members.

Figure 2 shows two more examples of relational systems that can be discovered by our model. In Figure 2b, we are interested in a set of terms that might appear on a medical chart, and the input matrix  $causes(x_i, x_j)$  specifies whether  $x_i$  causes  $x_j$ . A relational system for this problem might indicate that the terms can be organized into two categories—diseases and symptoms—and that diseases can cause symptoms. Figure 2c shows a case where we are interested in a group of elementary school children and provided with a relation  $friends\_with(x_i, x_j)$  which indicates whether  $x_i$  considers  $x_j$  to be a friend. Our model discovers a relational system where there are two categories—the boys and the girls—and where each student tends to be friends with others of the same gender.

The examples in Figure 2 illustrate three kinds of relational systems and Figure 4 shows four additional examples: a ring, a dominance hierarchy, a common cause structure and a common effect structure. All of these systems have real-world applications: rings can capture feedback loops, dominance hierarchies are often useful for capturing social relations, and common-cause and common-effect structures are often considered in the literature on causal reasoning. Many other structures are possible, and our approach should be able to capture any structure that can be represented as a graph, or as a collection of nodes and arrows. This family of structures includes a rich set of relational systems, including many systems discussed by previous authors (Keil, 1993; Griffiths & Tenenbaum, 2007; Kemp & Tenenbaum, 2008).

Even though the relational systems in Figures 2 and 4 are very simple, these representations still capture some important aspects of intuitive theories. Each system can be viewed as a framework theory that specifies the concepts which exist in a domain and the characteristic relationships between these concepts. Each concept derives its meaning from its relationships to other concepts: for instance, magnetic objects can be described as objects that interact with magnets, but fail to interact with other magnetic objects. Framework theories will

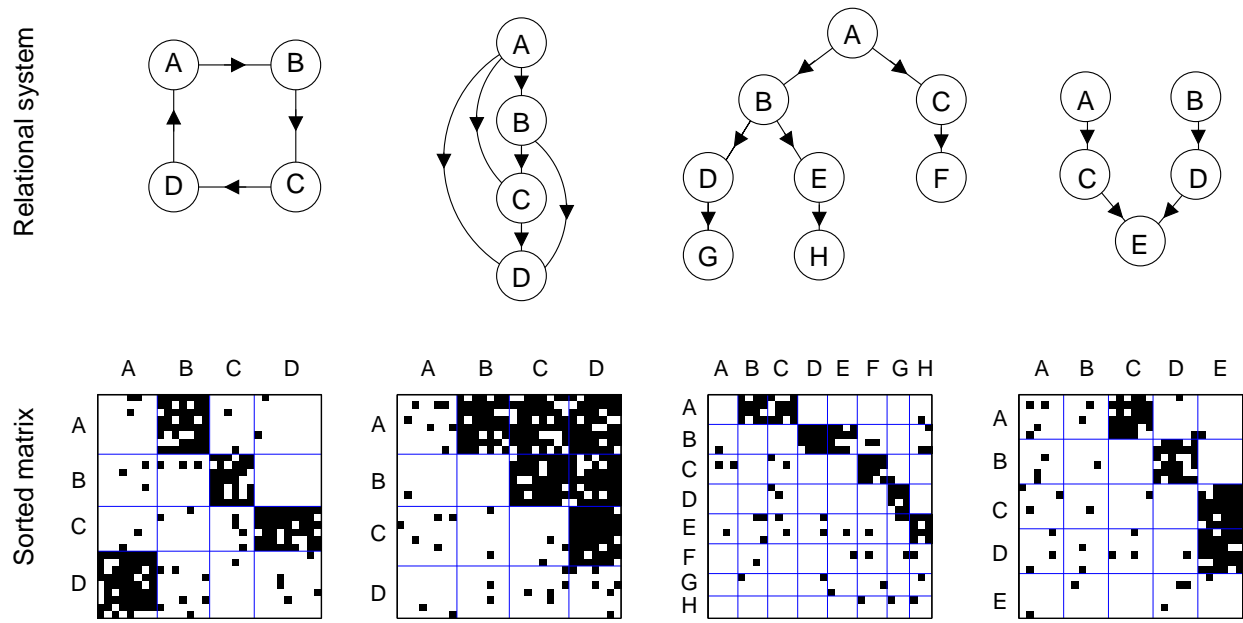
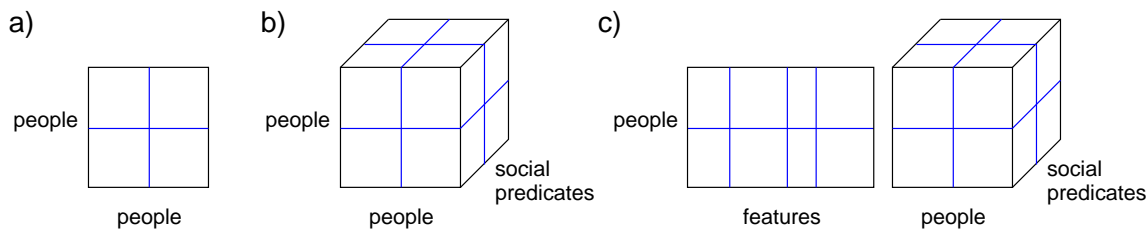


Figure 4. Our model can discover many kinds of structures that are useful for characterizing real-world relational systems. The four examples shown here include a ring, a dominance hierarchy, a common-cause structure and a common effect structure. The categories in each system are labeled with capital letters.

not provide a complete explanation of any domain: for instance, the theories in Figure 2 do not explain why lung cancer causes coughing but not fever. Even though systems like the examples in Figure 2 and 4 will not capture every kind of theoretical knowledge, the simplicity of these representations makes them a good initial target for models of theory learning.

The relational systems discovered by our model depend critically on the input provided. The examples in Figure 2a and 2b illustrate two rather different cases. In the magnets example, both the entities (metallic bars) and the *interacts with* relation are directly observable, and it is straightforward to see how a theory learner would gather the input data shown in Figure 2a.i. In the medical example, the symptoms (e.g. coughing) tend to be directly observable, but the diseases (e.g. lung cancer) and the “causes” relation are not. The input matrix specifies, for example, that lung cancer causes coughing, but recognizing this relationship depends on prior medical knowledge. Philosophers of science often suggest that there are no theory-neutral



*Figure 5.* Discovering relational systems given multiple types and relations. (a) Discovering a system given a single social relation, as shown in Figure 2c. (b) Simultaneously discovering categories of people, categories of social predicates, and relationships between these categories. Our goal is to discover a configuration where each 3-dimensional sub-block is relatively clean (contains mostly 1s or mostly 0s) (c) Attributes for the people can also be included. Note that the set of people appears three times in this example, and that the partition of this set is always the same.

observations, and it is important to realize that the input required by our model may be shaped by prior theoretical knowledge. We return to this point in the General Discussion, and consider the extent to which our model can go beyond the “theories” that are already implicit in the input representation.

We have focused so far on problems where there is a single set of entities and a single binary relation, but our approach will also handle more complicated systems of relational data. We illustrate by extending the elementary school example in Figure 2c. Suppose that we are given one or more relations involving one or more types, where a type corresponds to a collection of entities. In Figure 2c, there is a single type corresponding to *people*, and the binary relation *friends\_with*( $\cdot, \cdot$ ) is defined over the domain  $people \times people$ . In other words, the relation assigns a value—true or false—to each pair of people. Some types may correspond to collections of predicates: if there are multiple relations defined over the same domain, we will group them into a type and refer to them as predicates. For instance, we may have several social predicates defined over the domain  $people \times people$ : *friends\_with*( $\cdot, \cdot$ ), *admires*( $\cdot, \cdot$ ), *respects*( $\cdot, \cdot$ ), and *hates*( $\cdot, \cdot$ ). We can introduce a type for these social predicates, and define a ternary relation *applies*( $x_i, x_j, p$ ) which is true if predicate  $p$  applies to the pair  $(x_i, x_j)$ . Our goal is now to simultaneously categorize the people and the predicates (Figure 5b). For instance, we may learn a relational

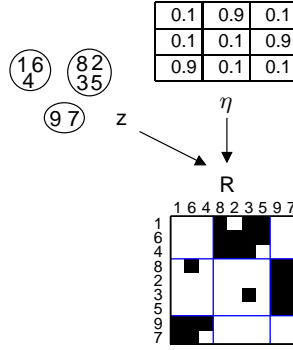


Figure 6. Our theory-learning model assumes that relation  $R$  is generated from a system  $(z, \eta)$ , where  $z$  is a partition of the entities into categories and  $\eta$  is a parameter matrix. Entry  $R(i, j)$  is generated by tossing a coin with bias  $\eta(z_i, z_j)$ , where  $z_i$  and  $z_j$  are the cluster assignments of entities  $i$  and  $j$ . We invert this generative model to discover the  $\eta$  and the  $z$  that maximize the posterior distribution  $P(z, \eta | R) \propto P(R | \eta, z) P(\eta | z) P(z)$ .

system which includes two categories of predicates—positive and negative predicates—and specifies that positive predicates tend to apply only between students of the same gender.

Our approach will handle arbitrarily complex systems of features, entities and relations. If we include features for the people, for example, we can simultaneously categorize people, social predicates, and features (Figure 5c). Returning to the elementary school example, suppose, for instance, that the features include predicates like *plays baseball*, *learns ballet*, *owns dolls*, and *owns toy guns*. We may learn a system that organizes the features into two categories, each of which tends to be associated with students of one gender.

### A probabilistic approach to theory discovery

We now provide a more formal description of the model sketched in the previous section. Each relational system in Figure 2 can be formalized as a pair  $(z, \eta)$ , where  $z$  is a partition of the entities into categories and  $\eta$  is a matrix that indicates how these categories relate to each other (Figure 6). The matrix  $\eta$  can be visualized as a *category graph*: a graph over the categories where the edge between category  $A$  and category  $B$  has weight  $\eta(A, B)$ , expressing the probability that  $A$ -entities will link to  $B$ -entities. The systems in Figures 2 and 4 are shown as category graphs

where only the edges with high weights have been included. In the general case, suppose that the input data include  $m$  relations defined over  $n$  types. A relational system specifies a partition  $z_i$  of each type into categories, and a matrix  $\eta_j$  for each relation  $R_j$ . The system can be represented as a vector  $(z^1, z^2, \dots, z^n, \eta^1, \dots, \eta^m)$ , where the  $z_i$  pick out the concepts that appear in the system and the  $\eta^j$  specify relationships between these concepts.

Suppose first that we are interested in learning a relational system  $(z, \eta)$  given a single binary relation  $R$  like the example in Figure 6. We take a probabilistic approach and search for the system that maximizes the posterior distribution

$$P(z, \eta | R) \propto P(R | \eta, z) P(\eta | z) P(z) \quad (1)$$

The terms on the right hand side of Equation 1 will capture assumptions about how relation  $R$  is generated from an underlying  $\eta$  and  $z$ , and how this  $\eta$  and  $z$  are generated in the first place. Bayes' rule allows us to convert these generative assumptions into a posterior distribution  $P(z, \eta | R)$  that can be used to identify the unobserved  $z$  and  $\eta$  that best account for the observed relation  $R$ . To complete the model, we will formally specify each term on the right hand side of Equation 1.

Consider first the term  $P(R | \eta, z)$ . We model relations as binary-valued functions but extensions to other datatypes (including continuous data and frequency data) are straightforward. For the problem in Figure 6, we have a single type  $T$  and a single two-place relation  $R : T \times T \rightarrow \{0, 1\}$ , and expect that the entries in this relation will be somehow explained by the underlying set of categories. There may be several ways to formalize this idea but we pursue one of the simplest, and assume that entry  $R(i, j)$  is generated by tossing a coin with bias  $\eta(z_i, z_j)$ , where  $z_i$  and  $z_j$  are the category assignments of entities  $i$  and  $j$ . Each entry  $\eta(A, B)$  in the parameter matrix  $\eta$  can therefore be interpreted as the probability that a given entity in class  $A$  will link to a given entity in class  $B$ . These assumptions about how  $R$  is generated from  $z$  and  $\eta$  lead to an expression for  $P(R | \eta, z)$  that appears as Equation 8 in Appendix A. Equation 1 also includes a term  $P(\eta | z)$  that captures assumptions about how the  $\eta$  matrix is generated. We

assume that the entries in  $\eta$  are independently drawn from a Beta distribution with hyperparameters  $\alpha$  and  $\beta$ . The equation for this distribution appears as Equation 7 in Appendix A.

The final term in Equation 1 is a prior  $P(z)$  that captures assumptions about how the category assignments in  $z$  were generated. Although relation  $R$  is assumed to depend on the category assignments in  $z$ , we do not assume that these category assignments are known in advance, or fix the number of categories that will be needed. Instead, we use a prior distribution  $P(z)$  on category assignments that assigns some probability mass to all possible partitions of type  $T$ . Note that the number of possible partitions is very large for any type including more than a handful of entities: for example, a domain with 20 entities has roughly  $10^{14}$  partitions. A suitable prior should favor partitions that use small numbers of categories, but should allow the number of categories to grow as more data are seen. Following Anderson (1991) and other authors (Antoniak, 1974; Neal, 1991; Rasmussen, 2002) we use a distribution over partitions induced by a Chinese Restaurant Process (CRP, also known as a Polya Urn scheme (Aldous, 1985; Pitman, 2002)). A formal definition of this prior appears in Appendix A.

Now that we have specified all three terms on the right hand side of Equation 1, we have a fully-specified model that helps to explain how relational systems can be acquired and used. We have focused so far on the acquisition problem, and have suggested that relational systems can be learned by searching for the  $z$  and  $\eta$  that maximize the posterior distribution in Equation 1. If a system  $S = (z, \eta)$  is already known, then our approach helps to explain how the system supports predictions about any missing entries in relation  $R$ . For instance, if  $\eta$  indicates that entities in category  $A$  tend to send links to entities in category  $B$ , we can infer that a new member of category  $A$  is likely to send links to most entities in category  $B$ .

Appendix A provides a formal description of the methods we used to compute the predictions of our model. We initially integrate out the  $\eta$  matrix and search for the best set of category assignments  $z$ . Once  $z$  is known, it is straightforward to recover the  $\eta$  matrix that

corresponds best to these category assignments. To discover the best  $z$  for a given data set, we developed a greedy search algorithm, and Figure 3 shows the progress of this algorithm when applied to a small problem. The input relation  $R$  can be represented as a matrix, and our algorithm tries to shuffle the rows and columns of this matrix so that it assumes a clean block structure like the final matrix in Figure 3. Note, however, that there will be many other ways to implement the computational theory we have described, and that the particular implementation we have chosen is not intended as a model of cognitive processing.

We began with a problem where the input data include a single binary relation defined over a single type, but the same basic approach can be used to discover systems that involve multiple types and multiple relations. Suppose that we observe  $m$  relations defined over  $n$  types. We are interested in discovering vectors  $z^i$  which organize each type into categories, and parameter matrices  $\eta^j$  which capture relationships between these categories. Formally, we search for the category assignments and parameter matrices that maximize the posterior distribution

$$\begin{aligned}
 P(z^1, z^2, \dots, z^n, \eta^1, \dots, \eta^m | R^1, \dots, R^j) &\propto P(R^1, \dots, R^j | \eta^1, \dots, \eta^m, z^1, z^2, \dots, z^n) \\
 &P(\eta^1, \dots, \eta^m | z^1, z^2, \dots, z^n) \\
 &P(z^1, z^2, \dots, z^n)
 \end{aligned} \tag{2}$$

As in Equation 1, this posterior distribution will favor category assignments  $\{z^i\}$  that achieve a tradeoff between simplicity (a good set of assignment vectors should use a relatively small number of categories) and fit to the data (the category assignments should account well for the relations  $\{R^j\}$ ). More details can again be found in Appendix A.

### Evaluating models of theory discovery

Formal accounts of theory discovery can make two contributions to cognitive science: they help to address questions about the learnability of theories, and they help to explain human behavior. We will evaluate our approach along both dimensions.

Consider first the learnability issue. Many philosophers have suggested that there can be no computational account of theory discovery (Newton-Smith, 1981), and Hempel’s (1985) version of this claim is especially relevant to our approach. Hempel points out that a theory-learning system must be able to discover concepts that cannot be defined as combinations of concepts that appear in the input data: instead, these novel concepts must be characterized in part by their relationships to one another. Hempel argues that it is far from clear that computers can make discoveries of this kind, but we suggest otherwise, and support our claim by applying our model to several data sets and showing that it simultaneously discovers concepts and relationships between these concepts. Analyses of synthetic data can show that theories are learnable in principle, but we take a step further and apply our model to three real-world problems: learning about animals and their features, learning ontological concepts, and discovering the structure of a kinship system. By working with noisy-real world data, we demonstrate that a computational approach to theory discovery is feasible both in principle and in practice.

The second challenge for models of theory learning is to make accurate predictions about behavioral data. Empirical studies of theory learning are difficult, since intuitive theories can be very complex, and typically take much longer to acquire than the duration of a laboratory experiment. The experimental component of this paper therefore focuses on very simple theories, and we explore whether our model makes accurate predictions about how these theories are acquired and used for inductive inference.

To establish the contribution of any new model it is important to compare it with existing formal approaches. Our work is related to previous probabilistic models that focus on features rather than relations, and to previous models that address problems such as analogical reasoning where relational information is critical. Previous models of categorization can be directly applied to the data sets that we consider, and we will compare our approach to one such model—Anderson’s account of categorization (Anderson, 1991). Previous models of analogical inference such as the structure mapping engine (Falkenhainer, Forbus, & Gentner, 1989) apply

less naturally to the problems that we consider, and we discuss some of these models towards the end of the paper.

### Categorizing objects and features

Later sections of this paper will demonstrate that our model can handle rich systems of relations, but we begin with a very simple setting where the raw data are a matrix of objects by features. Many models of categorization work with information about objects and their features and attempt to organize the objects into mutually exclusive categories (Medin & Schaffer, 1978; Anderson, 1991). Learning about features, however, can also be important, and often it makes sense to organize both objects and features into groups (Rosch, 1978; Malt & Smith, 1984; Keil, 1991). When learning about mammals, for instance, it may be useful to recognize a category of aquatic features that includes features like “has flippers,” “swims,” and “lives in water.” Feature categories provide a way to capture patterns of *coherent covariation* (Rogers & McClelland, 2004): notice that a mammal which has one aquatic feature is likely to have the others. Note also that feature categories and object categories can be mutually reinforcing: an aquatic mammal is a mammal that has aquatic features, and aquatic features are those that are shared by aquatic mammals. This section shows that our theory learning model can simultaneously discover object categories, feature categories, and the relationships between them.<sup>1</sup>

Any object-feature matrix can be viewed as a relation  $R : T^1 \times T^2 \rightarrow \{0, 1\}$  between a set of objects ( $T^1$ ) and a set of features ( $T^2$ ). Since binary features correspond to unary predicates, a feature category can be viewed as a simple instance of a predicate category, and this section will provide our first example of how predicate categories can be discovered. We applied our model to a binary matrix where the rows represent animal species and the columns represent features (Figure 7). Osherson, Stern, Wilkie, Stob, and Smith (1991) ran a task where participants were given 48 species and 85 features and asked to rate the strength of association between each animal and feature. Participants gave ratings on a scale that started at zero and had no upper

bound. Ratings were linearly transformed to values between 0 and 100, then averaged. We ran a similar experiment to collect feature ratings for two additional species—cow and dolphin—and combined our results with the original data set to create a matrix with 50 species and 85 features (Kemp & Tenenbaum, 2008). We converted this matrix into a binary matrix by thresholding the continuous feature ratings at the global mean.

Figure 7d shows that the matrix of biological data takes on a fairly clean block structure when the animals and the features are sorted into categories. The 12 animal categories include groups that correspond to marine mammals, freshwater mammals, primates, and bears, and the model assigns two of the more unusual animals—bat and giant panda—to their own categories. The 34 feature categories include sixteen singletons, but the remaining categories tend to capture patterns of coherent covariation. It is interesting to note that some of the feature categories include features of several different kinds. For example, “quadrupedal” is an anatomical feature, “walks” is a behavioral feature, and “lives on the ground” is an ecological feature, but all three are assigned to a single category since they tend to occur together. In addition to discovering feature categories and animal categories, the model also discovers relationships between these categories. Some of the strongest relationships indicate that aquatic mammals tend to have features from the category that includes “has flippers,” “swims,” and “lives in the ocean,” and that primates tend to have features from the category that includes “has hands,” “bipedal,” “lives in the jungle,” and “found in trees.”

If we choose not to cluster the features, our approach reduces to a model we will call the feature-based model. The feature-based model assumes the features are conditionally independent given the set of animal categories. When applied to the biological data, this model finds five animal categories, and Figure 7b indicates that most of these categories can be created by merging two or more of the categories discovered by our model. For example, the second category discovered by the feature-based model includes all of the hooved animals along with the giant panda, but our model chooses to divide this group further into categories O2 through O5. If the

50 animals in the data set are organized into a hierarchical folk taxonomy, the solutions found by the two models may correspond to cuts of this tree at two different levels. The feature-based model may cut the tree near the root, creating 5 relatively large categories, and the relational model may cut the tree nearer the leaves, creating a larger number of categories.

The feature-based model is intimately related to Anderson’s (1991) rational analysis of categorization. Anderson’s model is based on assumptions about the generative structure of the environment and assumptions about the processing limitations of human learners. The feature-based model is the model that follows from the first set of assumptions alone (Sanborn, Griffiths, & Navarro, 2006).<sup>2</sup> One of the assumptions about the structure of the environment is that features are conditionally independent given a partition of the objects. Loosely speaking, objects from the same category are assumed to have similar features, but features are assumed to be independently generated subject to this constraint. Given this independence assumption, learning that a robin has wings and flies provides evidence that other robins are likely to fly, but does not support the conclusion that winged entities from other categories (e.g. eagles, dragonflies and aeroplanes) are likely to fly. Although this independence assumption may make Anderson’s analysis more tractable, there is little reason to think that it is even approximately true of the features in Figure 7. As we have seen, the feature-based model cannot adequately capture the intuition that features (e.g. “has wings” and “flies”) tend to cluster, and the block structure in Figure 7 confirms that animal features do indeed cluster.

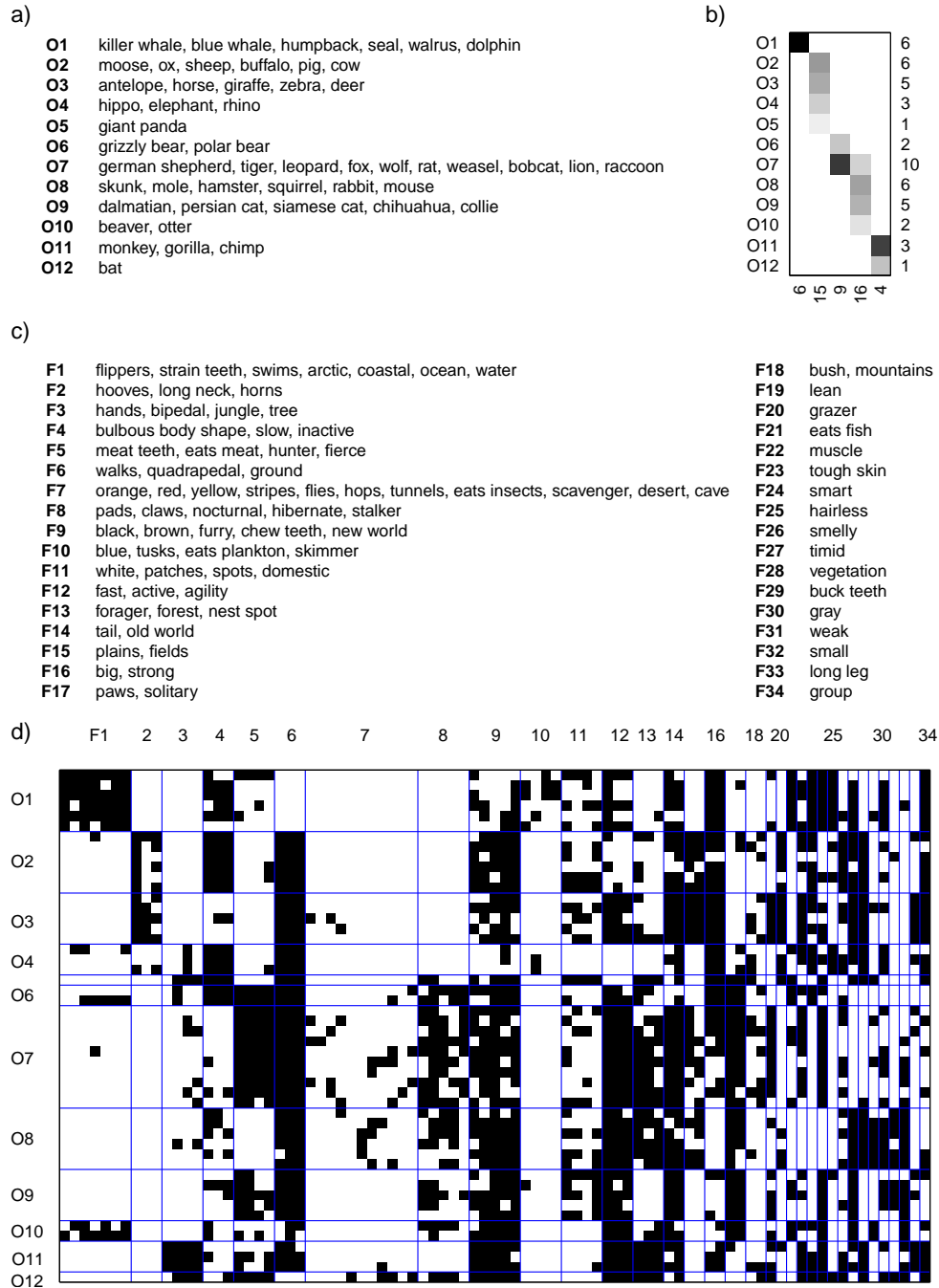


Figure 7. (a) Animal categories discovered by our relational model. (b) A comparison between the 12 categories found by our model (rows) and the 5 categories found by the feature-based model (columns). Each column shows how one of the feature-based categories is distributed over the 12 categories in (a): for instance, most members of the first feature-based category (column 1) belong to the first relational category (row 1). The numbers indicate the size of each category: for instance, the first relational category (row 1) has 6 members, and the second feature-based category (column 2) has 15 members. (c) Feature categories discovered by our model. Some features refer to habitat (jungle, tree, coastal), and others are anatomical (bulbous body shape, has teeth for straining food from the water) or behavioral (swims, slow). (d) A sorted matrix showing the relationship between the animal categories in (a) and the feature categories in (c).

Our primary motivation for working with feature data was to provide a simple initial demonstration of our model. The relational systems discovered by our model captures some aspects of intuitive theories: for instance, just as theoretical terms derive meaning from their relationships to other theoretical terms, the categories in Figure 7 are perhaps best described in terms of each other (aquatic mammals have aquatic features, and vice versa). It is useful to know that our model can discover interpretable structure in feature data, but the distinctive contributions of our model are seen most clearly when the available data are purely relational. The next three sections consider applications of this sort, and include our most compelling demonstrations that our model can acquire theory-like knowledge.

### **Discovering ontologies**

Intuitive theories vary in abstraction. Some theories capture detailed knowledge about specific topics such as illness and its causes, or the laws that govern the motion of physical objects. Others are more general theories that distinguish between basic ontological categories (e.g. agents, mental states, artifacts, substances, and events) and specify relationships between these categories (e.g. that agents can have mental states, or that artifacts are made out of substances). As mentioned earlier, general theories like these are sometimes called “framework theories,” since they establish a framework that can support the acquisition of more specialized theories (Wellman & Gelman, 1992). For instance, knowing that agents can have mental states can support the acquisition of theoretical knowledge about specific mental states such as beliefs and desires.

Some fundamental ontological concepts such as the notion of a physical object may be innately provided (Spelke, 1990). Learning, however, can help to sharpen and subdivide the ontological categories that are initially available. There are many developmental studies that trace the emergence of ontological knowledge (Carey, 1985; Keil, 1979) but relatively few computational proposals about how this knowledge might be acquired. In this section we suggest that our model can acquire ontological categories given information about the relations between

1. ORGANISMS	alga, amphibian, animal, archaeon, bacterium, bird, fish, fungus, human, invertebrate, ...
2. CHEMICALS	amino acid, carbohydrate, chemical, eicosanoid, element or ion or isotope, food, lipid, steroid, ...
3. BIOLOGICAL FUNCTIONS	biological function, cell function, genetic function, mental process, molecular function, ...
4. BIO-ACTIVE SUBSTANCES	antibiotic, enzyme, hormone, immunologic factor, pharmacologic substance, vitamin, ...
5. DISEASES	cell dysfunction, disease, mental dysfunction, neoplastic process, pathologic function, ...
6. PROCEDURES	diagnostic procedure, health care activity, laboratory procedure, molecular biology research technique, ...
7. ABNORMALITIES	acquired abnormality, anatomical abnormality, congenital abnormality, injury or poisoning
8. ANATOMY	anatomical structure, body location or region, body part or organ or organ component, cell, ...
9. PHENOMENA	clinical attribute, environmental effect of humans, human-caused phenomenon or process, ...
10. HUMAN GROUPS	age group, family group, group, health care related organization, organization, population group, ...
11. HUMAN PRODUCTIONS	body substance, clinical drug, classification, conceptual entity, manufactured object, medical device, ...
12. SIGNS	finding, laboratory or test result, sign or symptom
13. OCCUPATIONS	biomedical occupation or discipline, occupation or discipline
14. MISCELLANEOUS	activity, behavior, body system, carbohydrate sequence, educational activity, event, geographic area, ...

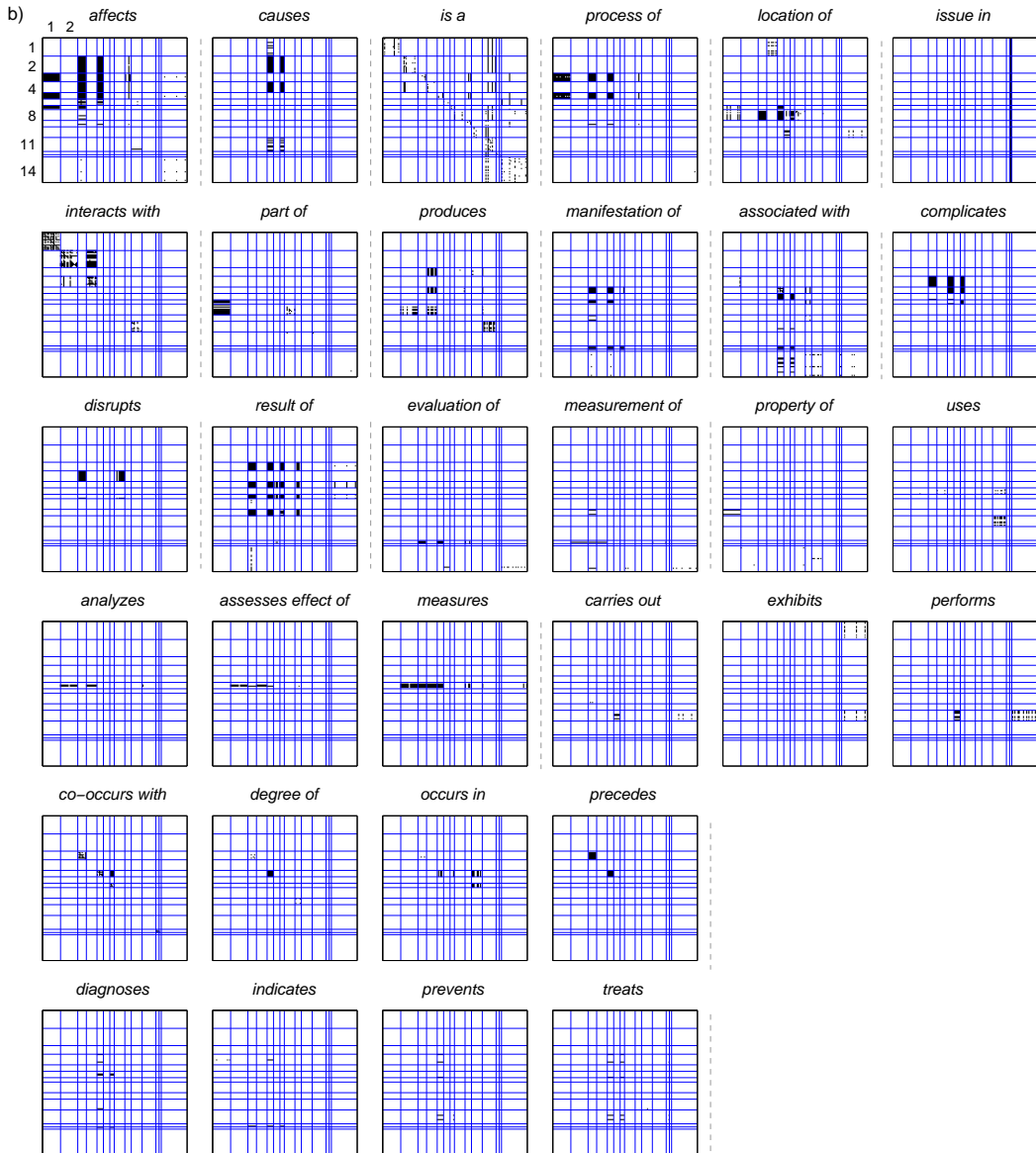


Figure 8. (a) Biomedical categories discovered by our relational model. (b) Adjacency matrices for 32 of the 49 predicates, where the rows and columns are sorted according to the partition of the entities shown in (a). We see, for instance, that chemicals *affect* biological functions and that organisms *interact with* organisms. The dashed grey lines between the matrices indicate the predicate categories discovered by our model. For instance, each of the first fourteen predicates belongs to its own category, and category 15 includes *evaluation of* and *measurement of*.

an initially undifferentiated set of entities. Theory-learners may acquire this relational information from many sources, but linguistic input may be one of the most important: Keil (1983), for example, suggests that linguistic pairings between objects and predicates help children recognize ontological distinctions that they initially ignore.

To demonstrate that our model can acquire ontological knowledge, we work with biomedical data from the Unified Medical Language System (UMLS, McCray 2003). Two of the basic categories within this domain are diseases and symptoms, and an ontology might state that diseases can *cause* symptoms (Figure 1). We show that our model discovers a simple theory that captures ontological knowledge of this sort. The raw data that support this discovery are taken from a semantic network with 135 entities and 49 predicates (McCray, 2003). The entities include nouns like “cell dysfunction,” “vitamin,” and “fungus.” The predicates include verbs like *complicates*, *affects* and *causes*. If  $T^1$  indicates the set of entities, each predicate can be represented as a binary relation  $R : T^1 \times T^1 \rightarrow \{0, 1\}$ . Figure 8 shows matrix representations for 32 of these relations, and we can use the complete set of matrices as the input for our model. Figure 3 suggests how our model proceeds when given a single matrix as input. When provided with the 49 matrices in the medical data set, the goal is to sort the 135 entities into categories such that each of the input matrices takes on a clean block structure. Figure 8b shows that this goal can be achieved, and that the input matrices are highly structured when sorted according to the categories in Figure 8a.

The results in Figure 8b were generated by analyzing a single ternary relation rather than 49 binary relations. Since each of the 49 relations has the same domain ( $T^1 \times T^1$ ), we can treat them as first-class entities and apply the model to the ternary relation  $R : T^1 \times T^1 \times T^2 \rightarrow \{0, 1\}$ , where  $T^2$  is the set of 49 predicates. This analysis is directly analogous to the elementary school example shown in Figure 5b. Both the 135 entities and the 49 predicates are sorted into categories, and a good solution allows the ternary relation to take on a clean block structure when sorted according to these categories. The previous section demonstrated that features can be

clustered, and here we see that relations form high-level types that can also be clustered. Our general philosophy is that *every* type is a candidate for clustering, although there may be idiosyncratic reasons why we choose not to cluster some of them.

Figure 8 shows some of the categories that emerge when we cluster both entities and predicates. The model discovers 14 entity categories in total, including a category of organisms, a category of chemicals and a category of biological functions. Figure 8a shows, for example, that the first category includes entities such as alga, amphibian, animal, and other organisms. When sorted according to the 14 categories, the matrices for most predicates take on a clean block structure, indicating that these categories account well for the raw data. The first matrix in Figure 8b, for example, shows that members of category 3 (biological functions) affect members of category 1 (organisms). The second matrix shows that members of category 2 (chemicals) cause members of category 5 (diseases) and members of category 7 (abnormalities).

By estimating the entries in  $\eta$  we can identify the pairs of categories that are most strongly linked and the predicates that link them. Figure 9 shows the strongest relationships between seven of the categories discovered by the model. Some of these relationships indicate that biological functions *affect* organisms, that chemicals *cause* diseases, and that bio-active substances *disrupt* biological functions. The network in Figure 9 is a holistic system where each category is specified in terms of its relationships to the other categories in the system (Goldstone & Rogosky, 2002), and where all of the categories have been discovered simultaneously. This result is perhaps our most compelling demonstration that our model can acquire complex systems of relational knowledge.

Our model discovers 21 predicate categories and all but one are shown in Figure 8. The first fourteen predicates tend to be associated with many observations, and each one is assigned to its own category. The final 7 predicates are very sparse, and all of them are assigned to a single category which is not shown in Figure 8. The remaining categories include between 2 and 4 predicates. One of these categories includes three predicates related to measurement:  $\{analyzes,$

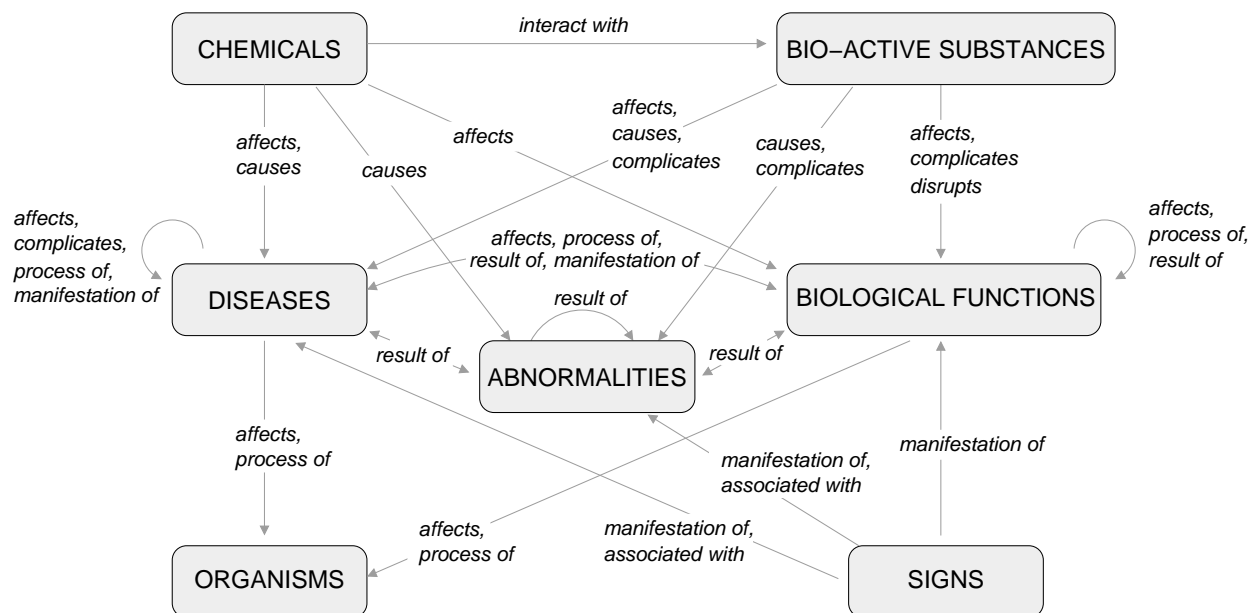


Figure 9. A category graph showing the strongest relationships between seven of the categories discovered by our model. All links with weights greater than 0.8 have been included.

*measures, assesses effect of*}. Another includes several possible relationships between diseases: *{co-occurs with, degree of, occurs in, precedes}*. In most cases, the matrices for predicates assigned to the same category tend to have entries in similar positions, which explains why the model has chosen to group them. The one exception is the category *{property of, uses}*, which includes two miscellaneous predicates that may be too sparse to be assigned to two separate categories, but not sparse enough to be grouped with the final 17 predicates.

If we wanted to organize the entities into categories without discovering the relationships between these categories, the feature-based model could be applied to a flattened version of the relational data. Suppose that  $a$  is an element of  $T^1$ , and we wish to flatten the ternary relation  $R : T^1 \times T^1 \times T^2 \rightarrow \{0, 1\}$ . The features of  $a$  correspond to all values of  $R(a, x^1, x^2)$  where  $x^1 \in T^1$  and  $x^2 \in T^2$  and all values of  $R(x^1, a, x^2)$ . Any relational system can be similarly converted into an object-feature matrix involving just one of its component dimensions. When applied to the flattened biomedical data, the feature-based model discovers 9 categories, fewer

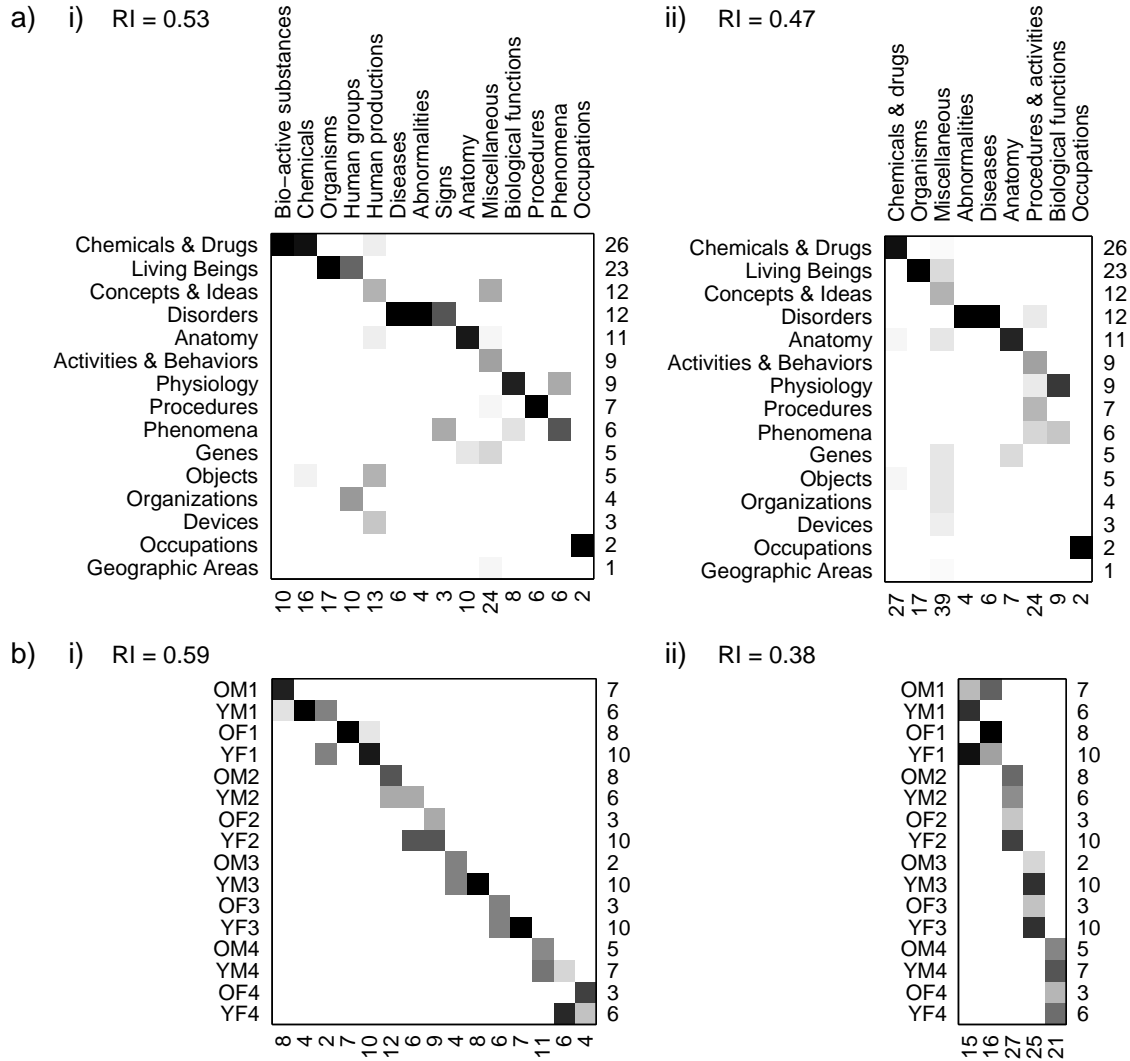


Figure 10. Comparisons of model solutions with “ground-truth” categories for the medical data and the kinship data. The adjusted Rand index (RI) for each comparison is shown. (a) Medical data. Each row in each matrix corresponds to a ground truth category. (i) Composition of the categories found by the relational model (Figure 8). Each column shows how one of these categories is distributed across the ground-truth categories. For instance, the relational model splits the Chemicals and Drugs category into two categories: Bio-active substances and Chemicals. (ii) Composition of the categories found by the feature-based model. Each column shows how one of the categories discovered by this model is distributed across the ground-truth categories. (b) Kinship data. Each row in each matrix corresponds to a ground-truth category, and each ground-truth category includes people from the same age-group, gender, and kinship section (OM1 includes older men from section 1). (i) Composition of the categories found by the relational model (Figure 11). Each column shows how one of these categories is distributed across the ground-truth categories. The categories discovered by the relational model match ground-truth partition fairly closely. (ii) Composition of the five categories found by the feature-based model. This model discovers the four kinship sections but little else.

than the 14 categories found by our model. Figure 10a compares both of these solutions to a ground-truth partition created by domain experts (McCray, Burgun, & Bodenreider, 2001). Each row of the matrices in Figure 10a corresponds to one of the ground-truth categories and the columns of these matrices show the composition of the categories discovered by the two models. Both models find solutions that are roughly consistent with the ground-truth categories. Our model discovers some categories (e.g. the Procedures category) that the feature-based model misses, but splits some of the ground-truth categories in two (e.g. the Chemicals and Drugs category). To quantify the similarity between the model solutions and the ground-truth partition, we used the adjusted Rand index (Hubert & Arabie, 1985). Compared to a ground-truth partition, a randomly generated partition has an expected score of zero, and the correct partition has a score of 1. As Figure 10a shows, the adjusted Rand indices for the two comparisons are similar (0.53 vs 0.47), and reveal no clear difference in the quality of the model solutions. Ultimately the best representation for the entities in the data set might be an ontological tree, and the partitions found by the two models and the ground-truth partition might correspond to cuts of this tree at three slightly different levels.

Our results for the UMLS data suggest that our model discovers a framework theory for the biomedical domain that is similar in some respects to the ontological knowledge acquired by humans. Like humans, for example, the model makes a fundamental ontological distinction between objects (e.g. chemicals) and processes (e.g. biological functions). A critical question for future work is whether our model can discover ontological categories given input that resembles more closely the input available to human learners. We supplied our model with raw data that were highly relevant to the problem of discovering ontological categories, but human learners must distinguish the raw data that are relevant from the raw data that are not so useful. Linguistic input specifies many relations like those that appear in the UMLS data, but designing simple techniques to extract these relations may raise some interesting challenges.

## Discovering kinship theories

Humans are social creatures, and intuitive theories about social systems govern many of our interactions with each other. Social systems take many forms: we all know, for example, that families, companies, and friendship networks are organized differently, and that membership in any of these systems is associated with a characteristic set of rules and obligations. Social theories represent a particularly important test case for models of theory acquisition, since there is compelling evidence that these theories are learned rather than acquired through maturation or some other means. Social systems vary greatly across cultures, and every child must learn the customs of her own social group.

Western kinship systems may appear relatively complex, but Australian tribes have become renowned among anthropologists for the complex relational structure of their kinship systems. Even trained field workers can find these systems difficult to understand (Findler, 1992) which raises an intriguing question about cognitive development: how do children discover the social structure of their tribe? The learning problem is particularly interesting since some communities appear to have no explicit representations of kinship rules, let alone cultural transmission of such rules. Findler (1992) describes one such case where the “extremely forceful injunction against a male person having sexual relation with his mother-in-law and with his son’s wife” could only be expressed by naming the pairs who could and could not engage in this act (p 286). Here we consider one Australian tribe—the Alyawarra, from Central Australia—and show that our model can discover some of the properties of the Alyawarra kinship system.

To a first approximation, Alyawarra kinship is captured by the Karia system shown in Figure 11a. The Alyawarra have four kinship sections, and Figure 11a shows how the kinship sections of individuals are related to the kinship sections of their parents. For example, every member of section 1 has a mother in section 4 and a father in section 3. The system implies that marriages are only permitted between sections 1 and 2 and between sections 3 and 4, and these marriage restrictions are just some of the important behavioral consequences of the Alyawarra

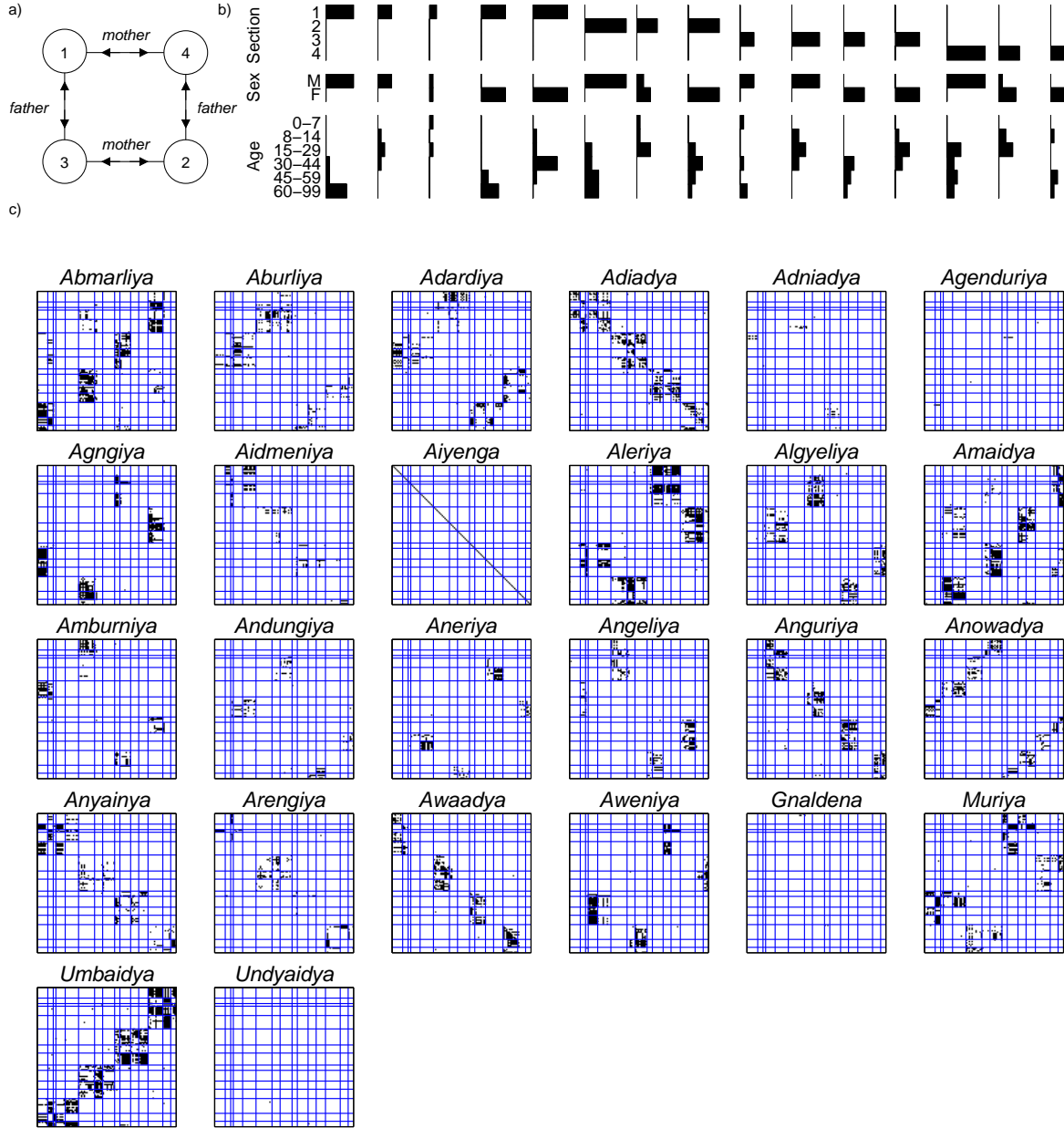


Figure 11. (a) The Kariera kinship system. Each person belongs to one of four kinship sections, and the section of any person predicts the sections of his or her parents. (b) Composition of the 15 categories found by our theory-learning model. The six age groupings were chosen by Denham, and are based in part on Alyawarra terms for age groupings (Denham, 1973). (c) Data for the 26 Alyawarra kinship terms. The 104 individuals are sorted according to the categories shown in (b).

kinship system. The Aranda system is an alternative to Figure 11a which uses eight rather than four categories, and Denham, McDaniel, and Atkins (1979) have suggested that Alyawarra kinship may be better described by the Aranda system than the Kariera system.<sup>3</sup> For our purposes, however, the simpler four section model is enough to give a flavor for the structure of the Alyawarra kinship system.

We applied our model to Alyawarra kinship data collected by Denham (1973, 2001). Denham asked 104 tribe members to provide kinship terms for each other, and Figure 11c shows the 26 different kinship terms recorded. Each kinship term can be represented as a matrix where the  $(i, j)$  cell is shaded if person  $i$  used that term to refer to person  $j$ . *Aiyenga* for example, is a term meaning “myself,” and the matrix for *Aiyenga* includes entries only along the diagonal, indicating that person  $i$  used the term *Aiyenga* only when looking at a picture of himself. We applied our model to the ternary relation  $R : T^1 \times T^1 \times T^2 \rightarrow \{0, 1\}$  where  $T^1$  is the set of 104 people and  $T^2$  is the set of kinship terms (see Figure 6b). Denham recorded demographic information for each of his informants, and we created a “ground truth” partition by assigning each person to one of 16 clusters depending on gender, kinship section, and a binary age feature (older than 45). The best solution according to our model uses 15 categories, and Figures 10b.i and 11b show that these categories are clean with respect to the dimensions of age, gender, and kinship section. The first five charts in Figure 11b, for example, show that the first five categories discovered by the model include only individuals from section 1. Category 1 includes older men, category 2 includes younger men, category 3 includes young children of both sexes, category 4 includes older women, and category 5 includes younger women.

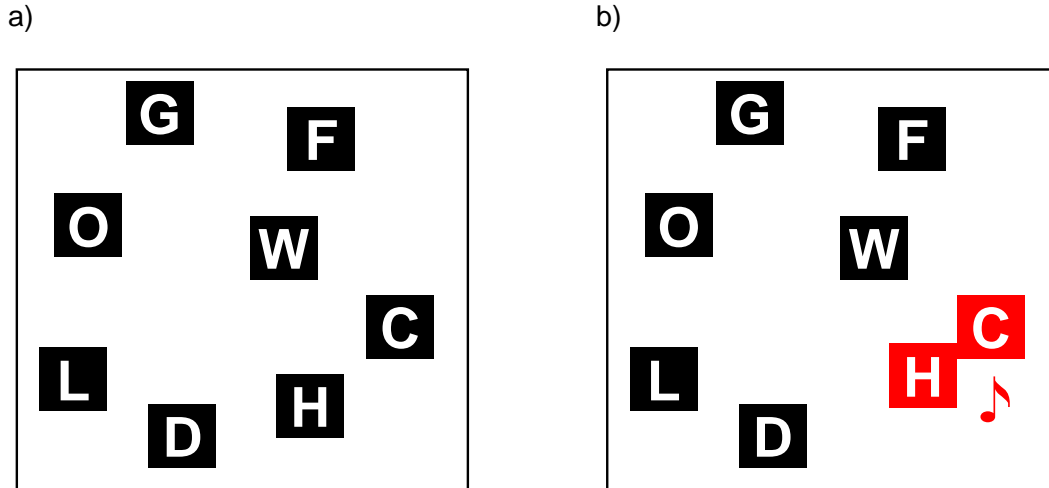
The matrices in Figure 11c have been sorted according to the 15-category partition discovered by our model (Figure 11b). The four kinship sections are clearly visible in many of the 26 matrices. For example, the sorted matrix for *Adiadya* has four rough blocks along the diagonal, indicating that speakers tend to use this term only for others who belong to the same kinship section. Consistent with this pattern, *Adiadya* refers to a classificatory younger brother or

sister: that is, to a younger person in one’s own section, even if he or she is not a biological sibling (glosses for each kinship term are provided in Appendix B). *Umbaidya* is used by female speakers to refer to a classificatory son or daughter, and by male speakers to refer to the child of a classificatory sister. We see from the relevant matrix that women in section 1 have children in section 4, and vice versa. *Anowadya* refers to a preferred marriage partner. The eight rough blocks in the relevant matrix indicate that men must marry women, that members of section 1 are expected to marry members of section 2, and that members of section 3 are expected to marry members of section 4. For example, the data points in the eighth block along the first row indicate that members of the first category (older men in section 1) can marry members of the eighth category (older women in section 2).

Discovering relationships between the 15 categories is crucial to understanding the kinship structure of the tribe. If, however, we were only interested in clustering the 104 individuals, we could apply the feature-based model to a flattened version of the data. The best partition discovered by this model includes 5 categories, and Figure 10b.ii shows that these categories are closely related to the four kinship sections. The adjusted Rand indices in Figure 10b confirm that the partition discovered by the relational model is closer to the ground truth partition than the feature-based solution. Again, however, the relational solution can be viewed as a refinement of the feature-based solution, and the best representation of the 104 individuals may ultimately be a hierarchy, or a set of nested partitions.

### **Experiment 1: Learning causal theories**

As children learn about the structure of their world, they develop intuitive theories about animals and their properties, about relationships between entities from different ontological kinds, and about the kinship system of their social group. Our results so far demonstrate that our model discovers interpretable theories in each of these domains. Each of these theories includes many categories, and the medical and kinship theories also include many relations. Our model therefore



*Figure 12.* A computer interface used to study theory learning in the laboratory. (a) Participants are free to drag around the objects and to organize them however they please. (b) Some objects activate other objects whenever they touch. Here objects H and C activate each other (both objects turn red and beep).

helps to explain how complex theories can be learned from noisy real-world data.

We now move on to a second challenge and explore some of the behavioral predictions of our model. Experimental studies of theory learning are difficult to design, since intuitive theories can be very complicated and can take many years to acquire. Consider, for instance, the sophisticated knowledge we have about animals and their properties, and the many years it takes for a mature theory of folk biology to emerge (Carey, 1985). Computational models should eventually aim to explain how people acquire large scale theories of domains like folk biology, but here we focus on a family of simple theories, each of which includes exactly two concepts. Working with these simple theories will allow us to develop controlled laboratory experiments and to collect quantitative data that can be compared against the predictions of our model.

Many studies have explored how people group objects into categories on the basis of their observable features (Anderson, 1991). We developed a paradigm for exploring how people learn systems of concepts, where each concept is defined by its relationships to other concepts. Our studies rely on a computer interface where objects can be moved around on screen. Some objects

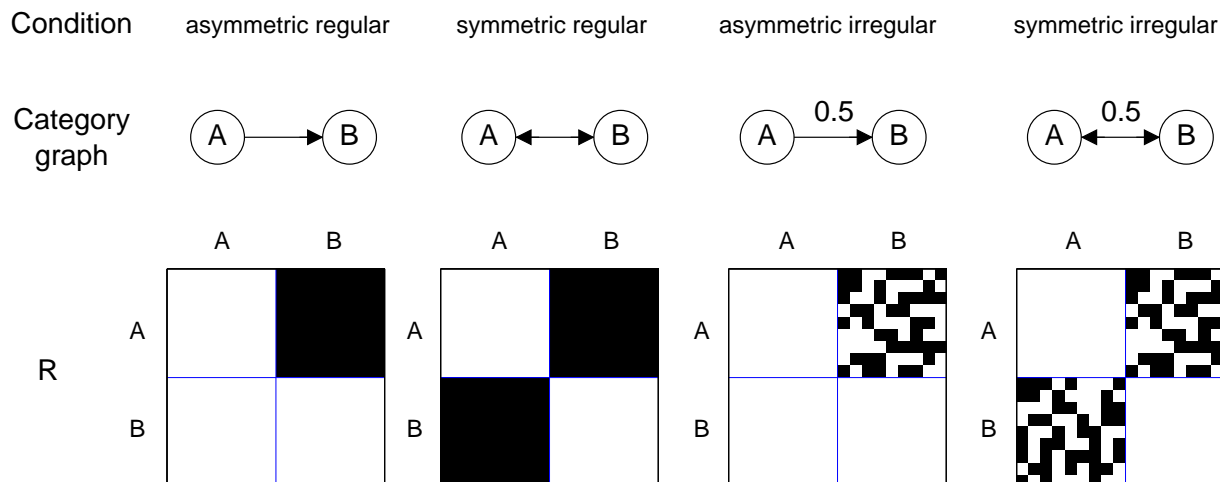


Figure 13. Category graphs and relations for the four conditions in Experiment 1.

“activate” other objects whenever they touch (Figure 12): if  $x$  activates  $y$  then  $y$  lights up and beeps whenever  $x$  and  $y$  touch. In some cases, activation is symmetric: both  $x$  and  $y$  light up and beep when they touch.

Our first experiment includes four conditions that explore the four different theories in Figure 13. Each of these theories includes two categories,  $A$  and  $B$ . In the *asymmetric regular* condition ( $A \rightarrow B$ ), every  $A$ -object activates every  $B$ -object but  $B$ -objects do not activate  $A$ -objects. In the *symmetric regular* condition ( $A \leftrightarrow B$ ), every  $A$  activates every  $B$  and every  $B$  activates every  $A$ . In the two *irregular* conditions ( $A \xrightarrow{0.5} B$  and  $A \overset{0.5}{\leftrightarrow} B$ ), each  $A$  activates a random subset (on average, 50%) of  $B$ 's. The theory-learning problems we consider are motivated in part by the magnetism example described in the introduction and illustrated in Figure 2a. For example, the objects in the  $A \leftrightarrow B$  condition are analogous to magnetic poles, and the activation relation in this condition is analogous to the attraction relation (positive and negative poles attract each other, but positive poles do not attract positive poles and negative poles do not attract negative poles). Theories of this kind may be very simple, but they capture an essential aspect of real-world theories that challenges conventional models of concept learning: the concepts and explanatory laws belonging to each theory are intrinsically relational, and can be

defined only in terms of each other.

Our first experiment explores several questions. The first and most basic is whether people can learn simple theories and use them to make predictions about unobserved relations between objects. Our experiment also provides information about the dynamics of theory learning, and about the relative difficulties of the four theories in Figure 13.

### *Participants*

Seventy five members of the MIT community participated for pay, and around 20 participants were assigned to each of the four conditions in Figure 13. The exact number of participants in each condition is shown in Table 1.

### *Stimuli*

The experiment used a custom-built graphical interface which displayed a set of objects and allowed participants to drag them around and touch them against each other (Figure 12). The objects were labeled with randomly-assigned letters, but were otherwise perceptually identical. The labels were used to refer to the objects: for instance, participants might be asked to touch G to O, or to predict what would happen when H and C touched.

### *Procedure*

Each condition begins with three objects on screen, and participants are asked to “play around with the objects and see what lights up.” New objects are added as the experiment progresses. Each condition has six phases, and three new objects are added to the screen during each phase. Whenever new objects are added, participants make predictions about how these objects will interact with some of the objects that are already onscreen. During the first and last phases (phases 1 and 6), participants provide a couple of sentences describing how the objects work.

During each phase, one of the three new objects serves as the probe object  $x$ . Before

observing any interactions involving the new objects, participants respond to Test 1: they predict how  $x$  will interact with two old objects, one ( $o_A$ ) from category  $A$  and the other ( $o_B$ ) from category  $B$ . These predictions are provided in response to the following questions:

Consider what will happen when  $x$  and  $o_A$  touch.

1. Will  $x$  light up?
2. Will  $o_A$  light up?

Consider what will happen when  $x$  and  $o_B$  touch.

3. Will  $x$  light up?
4. Will  $o_B$  light up?

Note that the symbols  $x$ ,  $o_A$  and  $o_B$  are replaced by the letter labels participants could see on screen (Figure 12). Responses are provided on a scale from 0 (definitely not) to 10 (definitely).

After answering these questions, participants are instructed to touch the probe object to an old object  $w$  that plays no role in the first test. Objects  $x$  and  $w$  always belong to different categories, and in each case  $w$  is chosen to ensure that the interaction between  $x$  and  $w$  will activate one or both of these objects. A learner who has discovered the correct theory should therefore be able to confidently predict the category of  $x$  based on this single interaction.

Participants then respond to Test 2, which contains exactly the same questions as Test 1. After the second test, participants are instructed to play around with the objects, and are able to manipulate all of the objects that appear on screen, including the three new objects. When they are ready, participants proceed to the next phase, three new objects are added, and the cycle of tests continues.

### *Model predictions*

We model inferences at each stage in the experiment by giving our model a binary matrix  $R$  that includes all observations that are currently possible. The four matrices in Figure 13 capture all possible observations when there are 18 objects on screen, and the matrices for earlier phases

in the experiment are smaller. Note that participants were required to generate their own observations by touching objects against each other, and some may have failed to observe all possible interactions at a given stage in the experiment. Informal observations suggest, however, that many participants observed each possible interaction at least once. Since all interactions were deterministic, note that repeated observations add no new information, and can be ignored for the purpose of computing model predictions.

Given all possible observations for the final phase in each condition, our model discovers the underlying theory in each case. The model also makes predictions about the dynamics of learning as each theory is acquired. Figures 14 and 15 show learning curves for conditions  $A \rightarrow B$  and  $A \leftrightarrow B$ . Each plot shows responses to two of the four questions that were asked during Test 1 and Test 2.

Consider first the model predictions about the  $A \rightarrow B$  condition. By the final phase there are 18 old objects on screen, and the model is quite confident that there are two categories of objects and that  $A$ -objects activate  $B$ -objects. When a new object  $x$  is introduced, the model is initially uncertain about its category assignment, and is unable to make confident predictions about whether  $x$  will be activated by an  $A$ -object, and whether  $x$  will activate a  $B$ -object (the dark blue and green curves in Figures 14a.i and 14b.i are close to 0.5). After observing  $x$  activate a  $B$ -object, however, the model is confident that  $x$  belongs to category  $A$ , and predicts that  $x$  is very likely to activate another  $B$ -object, but is very unlikely to be activated by an  $A$ -object (cyan and red curves in Figure 14a.i). In every phase of the experiment the model is uncertain about the Test 1 questions in Figure 14 (dark blue and green curves), but becomes increasingly confident about the Test 2 questions (cyan and red curves) as more objects are observed. Figure 15 shows that predictions about the  $A \leftrightarrow B$  condition follow a similar pattern.

In addition to making predictions about the tests in each condition, our approach makes predictions about the relative difficulties of the four conditions. To allow the four theories to be compared on an equal footing, we model a choice between two hypotheses. The first hypothesis

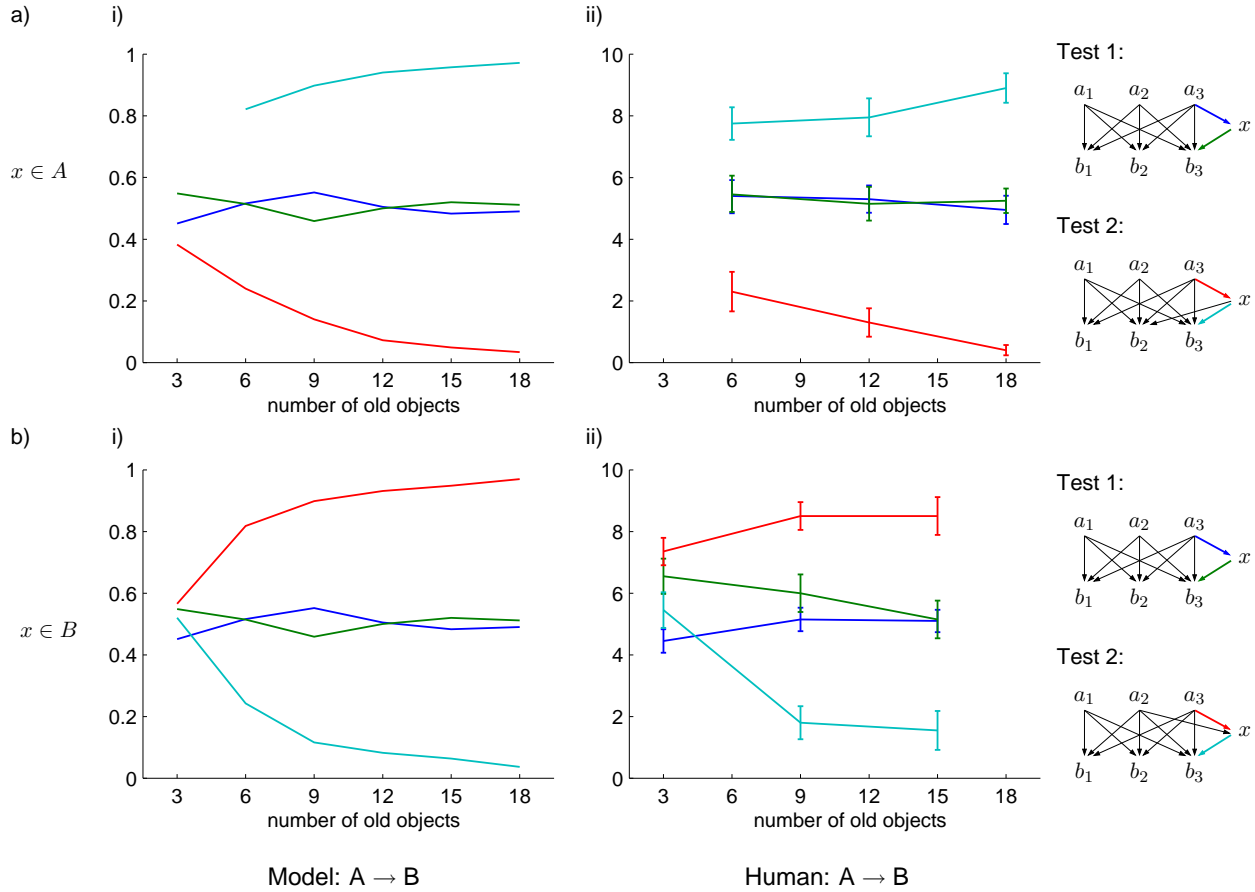


Figure 14. Learning curves for the asymmetric regular condition ( $A \rightarrow B$ ). Each curve shows inferences about whether a new object  $x$  will interact with one of the old objects. The interactions of interest are shown in the legend on the right: for instance, the dark blue curve shows inferences about whether  $x$  will be activated by one of the  $A$ -objects, and the green curve shows inferences about whether  $x$  will activate one of the  $B$ -objects. Object  $x$  may belong to category  $A$  (top row) or category  $B$  (bottom row, and inferences are shown before  $x$  is observed to interact with any other object (Test 1) and after a single interaction between  $x$  and a member of the category that does not include  $x$  (Test 2). Note that the number of old objects increases as more objects are encountered, and that the legends only show tests for the phase where there are exactly six old objects. Model predictions represent probabilities, and human predictions represent average judgments on a scale from 0 to 10.

$H_M$  asserts that the observed data were generated from our relational model—in other words, that the data for a given condition can be explained in terms of relationships between some set of latent categories. The “null hypothesis”  $H_0$  asserts that the data were generated from a model where each object is assigned to its own category—in other words, that there is no interesting

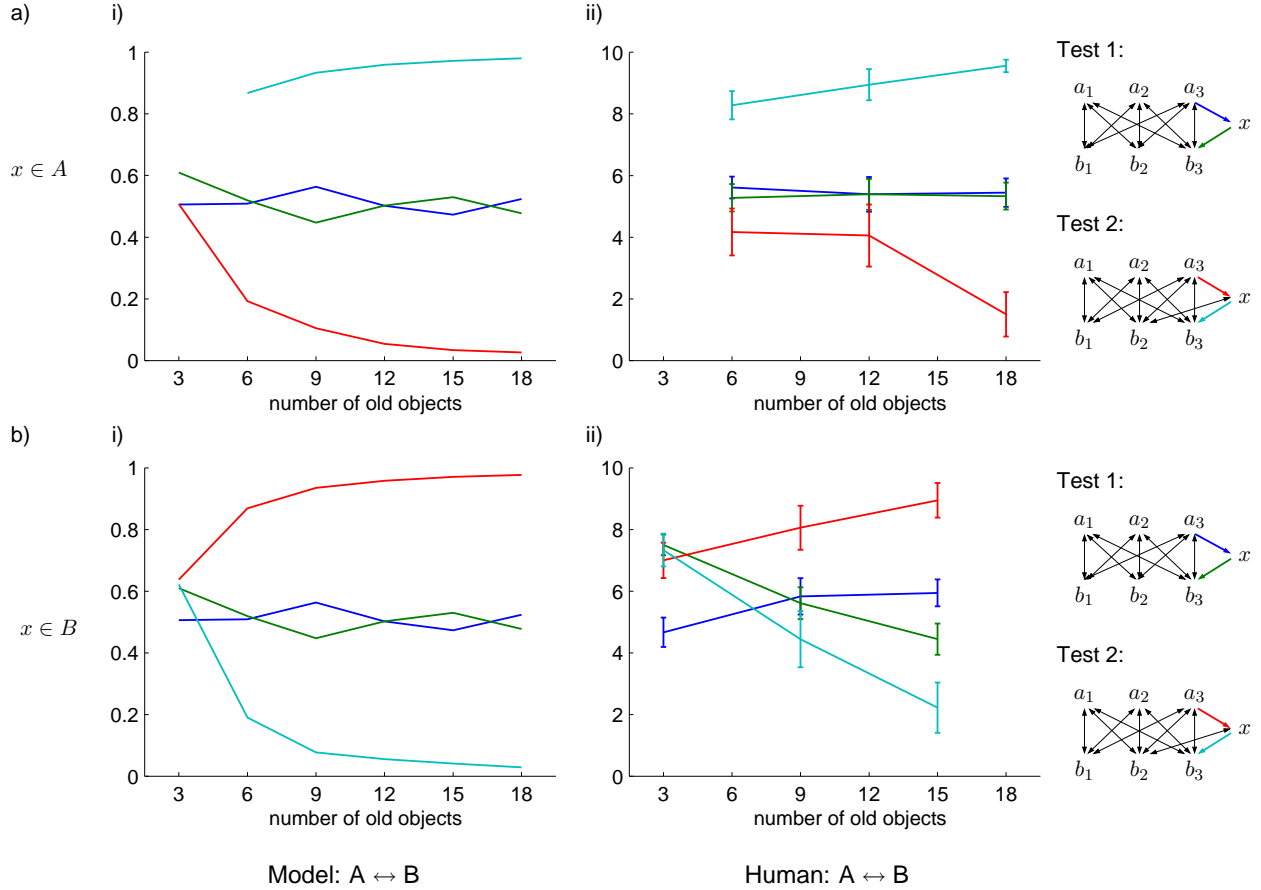


Figure 15. Learning curves for the symmetric regular condition ( $A \leftrightarrow B$ ). The inferences plotted are explained in the caption to Figure 14.

latent structure to discover. The choice between these two hypotheses depends on the ratio of their posterior probabilities:

$$\frac{P(H_M|R)}{P(H_0|R)} = \frac{P(R|H_M) P(H_M)}{P(R|H_0) P(H_0)} \quad (3)$$

We assume that the two hypotheses have the same prior probability:  $P(R|H_M) = P(R|H_0) = 0.5$ . To compute the likelihood  $P(R|H_0)$  we use Equation 9 in Appendix A, where  $z$  is a partition that assigns each object to its own category. To compute the likelihood of  $H_M$  we sum over all possible partitions:

$$P(R|H_M) = \sum_z P(R|z)P(z) \quad (4)$$

where  $P(R|z)$  is given by Equation 9 and  $P(z)$  is given by Equation 6.

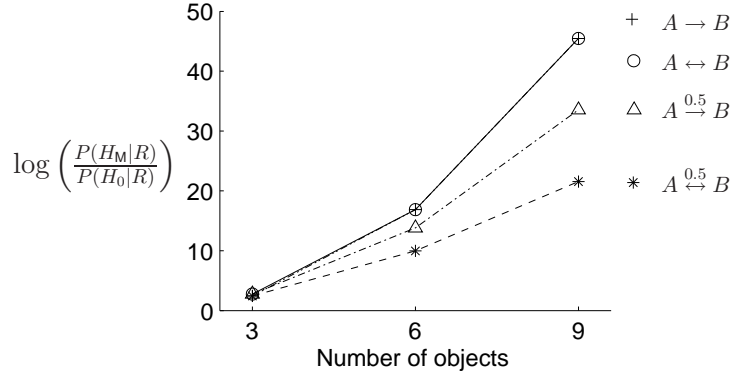


Figure 16. Support for a relational theory ( $H_M$ ) versus a null hypothesis ( $H_0$ ) where each object is assigned to its own category.

Figure 16 shows the relative support for these two hypotheses as the number of objects increases. In all cases the support for  $H_M$  increases as more objects are observed, but the rate of increase varies across the four theories. In particular, the two regular theories ( $A \rightarrow B$  and  $A \leftrightarrow B$ ) are expected to be easiest for participants to learn, since these are the theories that allow  $H_M$  and  $H_0$  to be discriminated most easily.

The predictions in Figure 16 can be understood by inspecting the sorted matrices in Figure 13. Clean blocks in these figures represent lawful relationships: in the symmetric regular case, for example, the white block at the upper left indicates that A objects never activate A objects, and the black block at the upper right indicates that A objects always activate B objects. The theories for the two regular conditions are equally easy to learn, since the relations  $R$  for these cases can both be organized into 4 perfectly clean blocks. The asymmetric irregular case is more difficult, since the sorted matrix shows only 3 perfectly clean blocks. The symmetric irregular case is most difficult of all, since it has the fewest number of lawful relationships (only 2 perfectly clean blocks).

### Results

Given experience with 18 objects, most participants in the  $A \rightarrow B$  and  $A \leftrightarrow B$  conditions successfully learned the underlying theories. After the final phase, their descriptions of how the

Condition	Theory-learners			Non-learners		
	#	Phases 5 & 6	All phases	#	Phases 5 & 6	All phases
$A \rightarrow B$	17	0.83	0.71	3	0.30	0.45
$A \leftrightarrow B$	14	0.88	0.72	4	0.26	0.16
$A \xrightarrow{0.5} B$	10	0.57	0.41	9	0.22 (7)	0.24 (8)
$A \xleftrightarrow{0.5} B$	1	—(0)	-0.02	17	0.16 (13)	0.20 (14)

Table 1

*Average correlations between model predictions and human responses for two groups of participants. Theory-learners gave written descriptions which suggested that they had discovered the underlying theory, and all remaining participants are classified as non-learners. Participants who gave the same response to all questions in tests 1 and 2 were dropped before computing average correlations. In each case where participants were removed, the number who contributed to the average is shown in parentheses.*

objects worked almost always mentioned two categories. One participant wrote that

XFWIKTAON all light up when touching objects ELHBQRMSDY, but if one box from the first group touches another from the first group then there is no light up. If two boxes from the second group touch [each] other then also there is no light up.

Based on these descriptions, we classified each participant as a *theory-learner* or a *non-learner*.

Theory learners gave descriptions which suggested that they had partitioned the objects into exactly two categories, and 70% or more of the participants in the  $A \rightarrow B$  and  $A \leftrightarrow B$  conditions met this criterion (Table 1). The way in which participants grouped the objects on screen confirmed that most of them had discovered the underlying theory. A common strategy was to arrange the objects in two rows, one for each category.

Predictions during the final phase provide further evidence that participants had discovered the regular theories by the end of the experiment, and suggest that participants were able to use these theories to support inductive inferences. In the final phase of the  $A \rightarrow B$  condition, Figure 14 suggests that participants were initially unsure about whether probe object  $x$  was an  $A$ -object or a  $B$ -object. Observing  $x$  activate one of the  $B$ -objects was enough for participants to realize that  $x$  must be an  $A$ -object, and to respond accordingly on Test 2. Responses to the two tests for the  $A \leftrightarrow B$  support a similar conclusion (Figure 15). Figures 14 and 15 show average responses, but the responses of individual participants were also highly correlated with the predictions of our model. In every phase, each participant provided 8 judgments on a scale from 0 to 10, and our model generates 8 probabilities in response to the same 8 questions. For each participant, we computed the correlation between human responses and model predictions across the final two phases (16 judgments in total) and across all phases (48 judgments in total). Average correlations for the theory-learners and non-learners are shown in Table 1. Across the final two phases of the experiment, the average correlations between model predictions and the responses of the theory-learners exceeded 0.8 in both conditions.

By the end of the two regular conditions, participants generated responses similar to the predictions of our model, and the average learning curves in Figures 14 and 15 show that our model also makes accurate predictions about earlier phases in the experiment. As predicted by the model, participants are initially relatively uncertain about their responses to both tests, but become more confident once more objects have been observed.<sup>4</sup> Again, responses of individual participants were highly correlated with the predictions of our model. Across all phases of the experiment, average correlations between the responses of individual theory-learners and model predictions exceeded 0.7 in both conditions. The verbal descriptions provided by participants provide further evidence that observations of several objects were needed to discover the underlying theory. Descriptions at the start of the first phase rarely used terms like “group” or “class,” and most participants referred only to interactions between specific pairs of objects. One

participant wrote

When C or A touches U, U turns red and there's a beep.

By the end of the final phase, however, most participants used terms like “group,” “class” or “lighter-uppers” in their descriptions.

Although participants successfully learned the two regular theories, they found the remaining theories more difficult. 10 of 19 participants in the  $A \stackrel{0.5}{\leftrightarrow} B$  condition were classified as theory-learners, and inductive inferences in this condition showed a lower correlation with the predictions of our model. The symmetric irregular theory ( $A \stackrel{0.5}{\leftrightarrow} B$ ) was even more difficult. Only 1 of 18 participants was classified as a theory-learner, and inductive inferences were uncorrelated with the predictions of our model.

Although our model predicts the relative difficulties of the four theories (Figure 16), it successfully learns a theory ( $A \stackrel{0.5}{\leftrightarrow} B$ ) that humans find very difficult. There are several reasons why this condition may be easier for our model than for human learners. Our model is not subject to memory or attentional limitations, but keeping track of the objects and relations in these experiments imposes heavy processing demands on human learners. If the data observed for the first several objects provide strong evidence for a block structure, people may quickly infer the theory and use it to encode the remaining observations that they make. In the  $A \stackrel{0.5}{\leftrightarrow} B$  condition, however, several phases are required before there is strong statistical evidence for the underlying theory, and memory demands may become overwhelming before this point is reached. Future work can explore whether assumptions about processing limitations can allow our model to better predict human responses, but future studies should also explore variants of our task that may make it easier for participants to learn the  $A \stackrel{0.5}{\leftrightarrow} B$  theory. For instance, Kemp, Goodman, and Tenenbaum (2008a) describe an alternative theory-learning paradigm that minimizes memory demands, and that may allow participants to discover theories like  $A \stackrel{0.5}{\leftrightarrow} B$ .

Taken together, the results of Experiment 1 suggest that our model helps to explain how humans learn simple theories. The most basic result is that people succeed in three of the four

conditions, suggesting that they are able to learn systems of concepts where each concept is defined by its relationships to the others. The learning curves in Figures 14 and 15 show that participants can use these theories to make predictions about unobserved relations, and that these predictions are sensitive to the weight of statistical evidence observed. Finally, the model predicts the relative difficulties of the four theories we considered, and suggests that the difficulty of learning a given theory is determined in part by the amount of statistical evidence for the categories and relationships specified by the theory.

Although Experiment 1 provides some initial support for our model, it is natural to ask whether existing models of categorization might also explain how people learn relational systems. We designed a second experiment to address this question.

### **Experiment 2: Relations or features?**

As mentioned already, psychologists have developed many models of categorization, including the contrast model (Medin & Schaffer, 1978), the generalized contrast model (Nosofsky, 1986), and Anderson’s rational model (Anderson, 1991). The most prominent modeling tradition has focused on a *feature-based* approach where the category assignment for a given object is determined by its features. We have argued instead for a *relational* approach where the category assignment for an object is determined by its relationships to other objects.

Although the feature-based approach and the relational approach seem different on the surface, some careful thought is needed before deciding whether the apparent differences are truly fundamental. In particular, it is important to consider whether the relational approach can be converted to a feature-based approach by converting a set of relations into a set of features. For instance, if *a* is the father of *b*, perhaps we can say that *a* has two features—*father*, and *father of b*—and that *b* also has two features—*child*, and *child of a*.

Converting relations to features may be acceptable under some circumstances, but this move has several problematic consequences. First, *father of b* may not match our intuitive idea of

a feature. Second, systematicity is lost. A person who knows about the *father\_of*( $\cdot, \cdot$ ) relation knows that the relation is either true or false for each pair of his relatives, but there is no reason why a list of features that includes the feature *father of a* must also include features like *father of b* or *child of a*. Third, compositionality is lost. A person who knows that the relation *father\_of*( $a, b$ ) is true knows that the components *father\_of*( $\cdot, \cdot$ ),  $a$  and  $b$  have the same meaning in this proposition as they do in other relations such as *father\_of*( $b, d$ ) and *brother\_of*( $b, c$ ). The relational approach makes these connections transparent, but the feature-based approach does not explain how complex features can be built from simpler pieces. Finally, converting relations to features makes it difficult to learn about the structure of any particular relation, and about the ways in which relations may be linked to one another. For instance, a model that works directly with a set of kinship relations may be able to discover that the *sibling\_of*( $\cdot, \cdot$ ) relation is symmetric, and can be defined in terms of the *parent\_of*( $\cdot, \cdot$ ) relation, but a model that converts all of these relations to features will find it hard to match this ability.

The reasons just described provide theoretical grounds for distinguishing between feature-based and relational approaches, but this distinction can also be explored empirically. Throughout we have compared our relational model to a feature-based approach that converts a relation to a set of features then applies Anderson’s rational model of categorization. We designed a second experiment to explore two cases where our relational model and this feature-based alternative make very different predictions. Each case is a setting where learners observe relationships among one set of objects then make predictions about relationships between a second set of objects. The critical question is whether participants can use what they have learned about the objects in the first set to make inferences about the objects in the second set. Our relational model supports this kind of transfer, and predicts that learners will acquire abstract knowledge that can be carried over from one set of objects to another. The feature-based model, however, will only learn about features that are tied to specific objects in the first set, and predicts that learners will fail to generalize to the second set of objects.

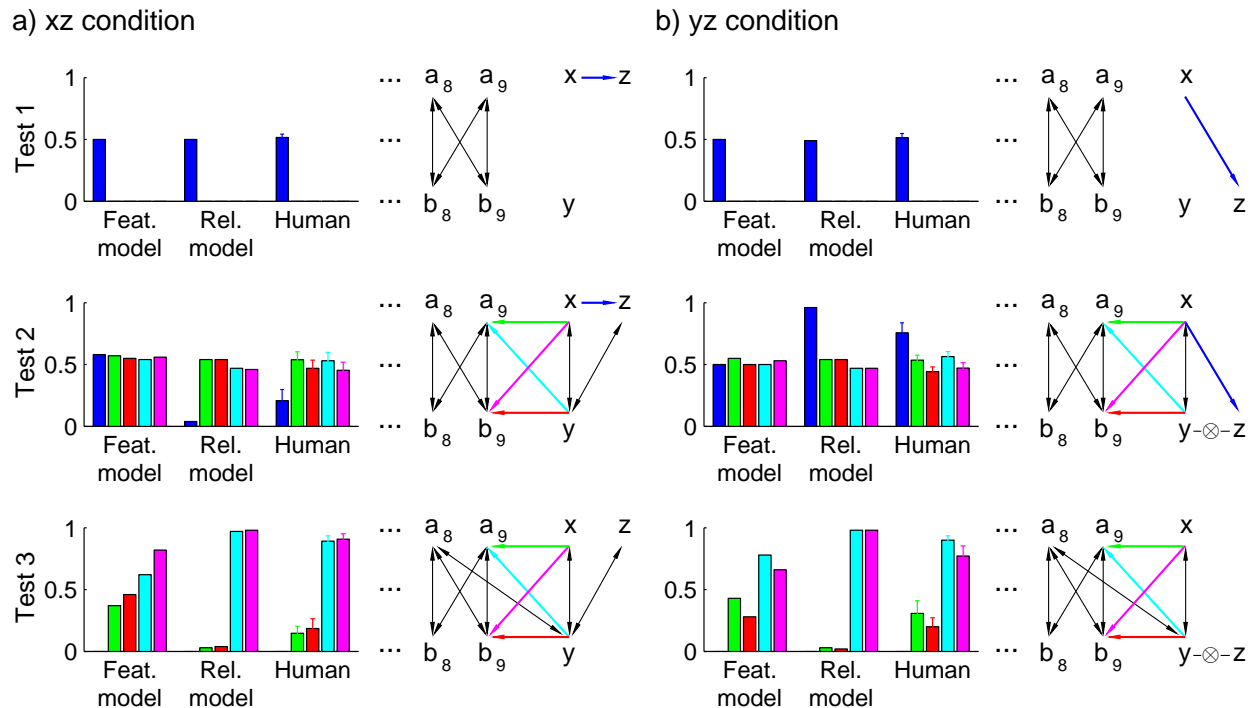


Figure 17. Predictions about three new objects ( $x, y, z$ ) after seeing eighteen objects consistent with the theory  $A \leftrightarrow B$ . Black arrows indicate previously observed interactions (in Test 2 for the  $yz$  condition, the cross on the edge between  $y$  and  $z$  indicates that  $y$  and  $z$  have been observed not to activate each other). The bars in each plot show inductive inferences about the colored arrows: for example, the plots in the first row show inferences about whether  $x$  is likely to activate  $z$ . Each plot show predictions of the feature-based model, predictions of our relational model, and average responses across all theory-learners in each condition.

### Participants

Forty eight members of the MIT community participated for pay. The experiment includes two conditions: 18 participants were assigned to the first condition, and 20 were assigned to the second condition.

### Procedure

The two conditions (the  $xz$  condition and the  $yz$  condition) are identical to the  $A \leftrightarrow B$  condition of Experiment 1 except that the final phase now includes three tests. At the beginning of this final phase, 18 objects are on screen, 9 from category  $A$  and 9 from category  $B$ . Three new

objects ( $x$ ,  $y$ , and  $z$ ) are now introduced. In the  $xz$  condition,  $x$  and  $z$  belong to the same category, but  $y$  belongs to a different category. Before seeing any interactions involving the new objects, participants are asked to predict whether  $x$  will activate  $z$  (Test 1). Participants are then asked to touch  $x$  and  $y$  together, and to touch  $y$  and  $z$  together (Figure 17a). In both cases, participants observe that these pairs activate each other. Participants are now given a second test where they predict whether  $x$  will activate  $z$ , and whether  $x$  and  $y$  will activate old objects from categories  $A$  and  $B$ . After this test, participants are instructed to touch  $y$  against one of the old  $A$ -objects. In principle, participants now have enough information to infer whether  $x$  belongs to category  $A$  or  $B$ , and they indicate what they have learned by responding to a third test. The  $yz$  condition is very similar, except that  $z$  is now a member of the same category as  $y$  rather than  $x$ , and participants observe that  $z$  and  $y$  fail to activate each other when touched (Figure 17b).

### *Model predictions*

Test 2 in Figure 17 exposes a critical difference between the relational and feature-based models. By this stage of the task, learners cannot infer whether any of the novel objects is an  $A$ -object or a  $B$ -object, but should be able to decide whether any two novel objects belong to the same category. In the  $xz$  condition, learning the  $A \leftrightarrow B$  theory allows our model to recognize that only two possibilities are likely. Since  $x$  and  $z$  both activate  $y$ , either  $x$  and  $z$  are  $A$ -objects and  $y$  is a  $B$ -object, or  $x$  and  $z$  are  $B$ -objects and  $y$  is a  $A$ -object. In both cases, however,  $x$  and  $z$  belong to the same category, and the model therefore predicts with high confidence that the two will fail to activate each other when touched. In contrast, the feature-based model predicts that  $x$  will activate  $z$  with probability greater than 0.5. Under this model, learning about relations between new objects is equivalent to learning about entirely new features, and none of the observations made in previous phases is directly relevant. The predictions of the feature-based model are driven by only two pieces of information: the observation that  $x$  and  $y$  activate each other, and the observation that  $z$  and  $y$  activate each other. Since all trials involving  $x$  and  $z$  have produced activations, the model makes a weak prediction that  $x$  and  $z$  are likely to activate each other.

	<i>xz</i> condition	<i>yz</i> condition
$p(x \text{ activates } z) \leq 0.5$	11	4
$p(x \text{ activates } z) > 0.5$	2	10

Table 2

*Inferences about whether  $x$  will activate  $z$  in Test 2 of Experiment 2. Participants have been organized into two groups depending on whether their response is greater than 5 on a scale from 0 to 10.*

Predictions about the *yz* condition show a similar difference between the models (Figure 17b). Observing that  $x$  activates  $y$  and that  $z$  fails to activate  $y$  is enough for our model to conclude that  $x$  belongs to one category and that  $y$  and  $z$  belong to the other. Our model therefore predicts that  $x$  will activate  $z$ , but the feature-based model makes no such prediction.

Notice that the critical question in Test 2 is the same in both conditions: will  $x$  activate  $z$ ? The relational model provides very different answers in the two conditions, and predicts that responses to Test 2 will differ across conditions. The feature-based model, however, predicts that responses to Test 2 will be similar across conditions.

### *Results*

In both conditions, most participants appeared to learn the  $A \leftrightarrow B$  theory, which replicates our result from Experiment 1. Participants were classified as “theory learners” if the correlation between their responses and the correct responses to phases 4 and 5 exceeded 0.5. Thirteen or more participants met this criterion in each condition, and all further analyses are restricted to this group of theory learners.

Responses to the critical question in Test 2 of phase 6 are summarized in Table 2. The counts in this table indicate how many participants inferred that  $x$  was likely to activate  $z$ . As predicted by the relational model, most participants in condition *xz* infer that  $x$  will fail to activate  $z$ , but the opposite result is true for condition *yz*. Binomial tests indicate that the result for the *xz* condition is statistically significant ( $p < 0.05$ , one-sided), and that the result for the *yz*

condition is marginally significant ( $p = 0.09$ , one-sided). Table 2 also suggests that responses to the critical test question were different across the two conditions. As predicted by the relational model, participants in the  $yz$  condition were more likely to predict that  $x$  would activate  $z$ . A Fisher’s exact test indicates that this result is statistically significant ( $p < 0.01$ , one-sided).

Mean responses to all questions in Test 2 of phase 6 are shown in Figure 17. In condition  $xz$ , the average response profile is qualitatively similar to the predictions of our model: participants predict that  $x$  is unlikely to activate  $z$ , but are uncertain about the remaining questions in Test 2. Figure 18 shows average responses after participants are partitioned into three groups depending on whether their response to the critical question in Test 2 is greater than, less than, or equal to 5 on a 0-10 scale. Although there are individual differences, the largest group in both conditions matches the predictions of the relational model.

Consider now the responses to the third test in phase 6. Casual inspection of Figure 17 suggests that the relational model accounts better for responses in the  $xz$  condition, but that the feature model accounts better for responses in the  $yz$  condition. Figure 18, however, indicates that the average response in the  $yz$  condition is somewhat misleading. When the participants are partitioned into the same groups used to analyze Test 2, the responses of the largest group match the prediction of the relational model in both conditions. Overall, then, our results indicate that the feature-based model may predict the responses of a small minority, but that the majority of participants provide responses consistent with the relational model.

Our second experiment demonstrates that learners trained on one set of objects make confident inferences about a second set of objects, suggesting that the knowledge they acquired is not tied to specific objects in the first set. Our relational model supports generalizations of this kind, but a more conventional feature-based approach fails to transfer its knowledge to situations involving new objects. Our results therefore suggest that theories are better treated as systems of relations than collections of features, and support the idea that learners maintain and reason about entire systems instead of converting these systems into collections of independent features.

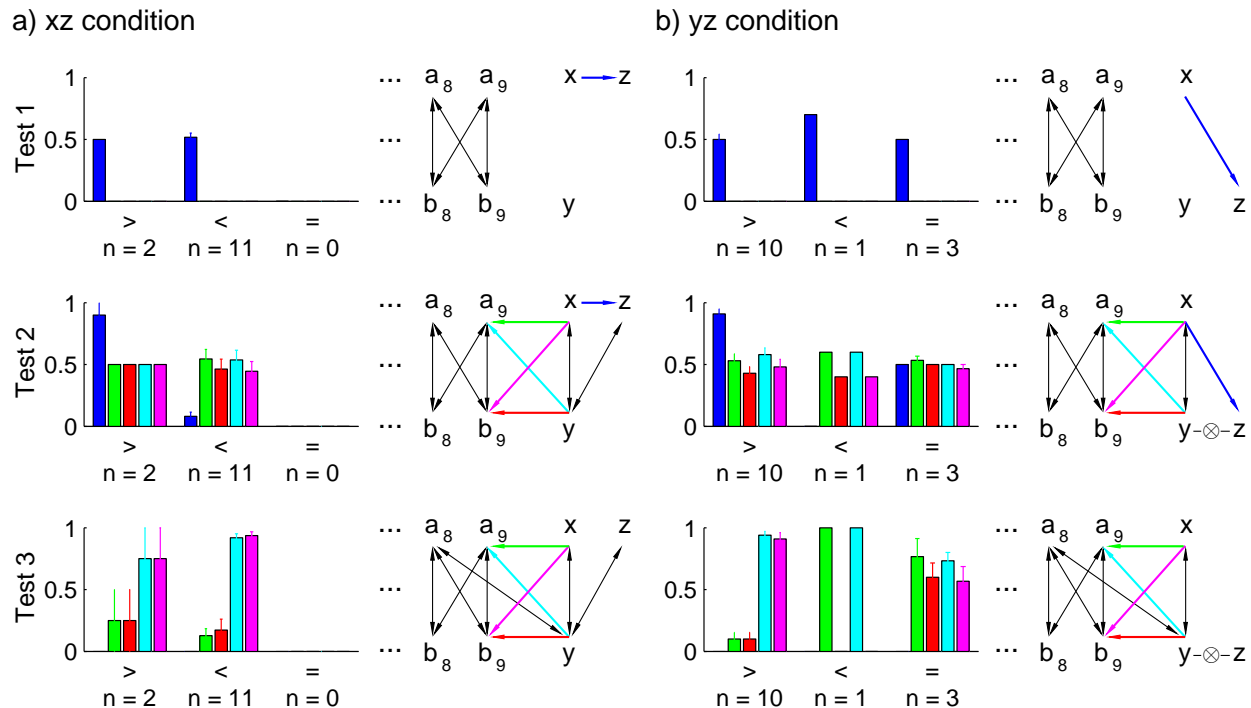


Figure 18. Individual differences analysis for Experiment 2. Participants have been assigned to three groups depending on whether their prediction about the critical  $x \rightarrow z$  arrow in Test 2 is greater than ( $>$ ), less than ( $<$ ), or equal to ( $=$ ) 5 on a 0-10 scale. The size of each group is shown below each plot. In each case, the predictions of the largest group match the predictions of the relational model shown in Figure 17.

## General Discussion

Our work is motivated by three general questions: what are theories, how do they support inductive inference, and how are they acquired? We approached these questions by working with a space of very simple theories. Each of these theories specifies the categories that exist in a domain and the relationships that exist between these categories. We showed that theories in this family can be characterized as generative models that make inductive predictions about unobserved interactions between entities. Statistical inference can be used to compute these predictions, and statistical inference also explains how these theories can be acquired in the first place. We assume that a theory-learner starts out with a hypothesis space including many possible theories, and a prior distribution over this space that favors the simpler theories. Theory

discovery is a matter of choosing the element in this space that maximizes a tradeoff between fit to the data and *a priori* plausibility.

We analyzed both behavioral data and data inspired by real world problems that humans must solve, and the results suggest that our model discovers interpretable theories across a broad range of domains. In particular, our results show that that the model can capture knowledge about folk biology, folk sociology, and folk physics. In the realm of folk biology, we showed that the model discovers a theory that specifies relationships between animal categories and clusters of features. In the realm of folk sociology, we showed that the model discovers a theory of the kinship structure of an Australian tribe. Finally, our behavioral experiments were inspired by simple physical theories about interactions between entities like magnets or electric charges.

### *Theory-laden data*

All computational models rely on input data—in our case, observations of relationships between entities. We have discussed how relational data can support theory discovery, but it is important also to consider whether sophisticated theoretical knowledge must already be implicit in the input. The data sets we considered rely on prior theoretical knowledge to different extents. The kinship data and the causal interaction data are based on direct observations: for example, a child can observe how person 1 refers to person 2, and the participants in our experiments could observe whether object 1 activated object 2 upon contact. The animal data represent an intermediate case—some of the features in this data set are directly observable (e.g. “has a bulbous body shape”), but others (e.g. “smart”) seem rather more complex. The biomedical data rely on prior theoretical knowledge to the greatest extent. The input data include entities (e.g. “amphibian” and “mental dysfunction”) and relations (e.g. “causes,” “complicates”) that cannot be directly observed, and are defined in part by the roles that they play in theories.

Although the kinship data and the causal interaction data appear more primitive than the other cases we consider, even these two data sets assume a considerable amount of background knowledge. Note, for example, that when modeling our behavioral tasks, we did not supply our

model with information about the location of the objects on the screen, but instead used an input matrix that highlights the information most relevant to the theory-learning problem. Some researchers have argued that a theory is characterized in part by the phenomena that it explains (Carey, 1985). In each of our analyses we gave our model a good chance of success by providing it with a coherent set of observations to explain.

Our general approach is consistent with the common claim that there are no theory-neutral observations (Hanson, 1958; Kuhn, 1970). Our contribution is not to explain how a learner who is entirely innocent of theoretical knowledge might acquire her first intuitive theory. Instead, our work helps to explain how a learner who already has some amount of knowledge might be able to acquire more. In each case we considered, our model acquires concepts and relationships between these concepts that are not directly present in the input data. The input data may be theory-laden to a greater or lesser extent, but in each case our model goes beyond the theories that are implicit in the input representation.

Future models of theory discovery will always need to rely on background knowledge of some kind, but future work can explore settings where the amount of pre-existing knowledge is minimized. For example, to capture the idea that theory-learners often confront problems where the relevant data are not clearly identified in advance, we can provide our model with a noisy set of relations, only some of which are relevant to an underlying theory. Since we take a probabilistic approach, our model should be robust to noise, and Kemp, Tenenbaum, Griffiths, Yamada, and Ueda (2006) describe analyses of synthetic data which support this conclusion. Future modeling and experimental work can explore this area in more detail, and can aim to characterize theory discovery in settings that more closely resemble the problems faced by human learners.

### *Related work*

Although there have been few previous attempts to model the discovery of intuitive theories, our approach builds on previous work in several fields.

*Psychology.*

Psychologists have developed many models of categorization, including the contrast model (Medin & Schaffer, 1978), the GCM (Nosofsky, 1986), and Anderson’s rational model (Anderson, 1991). Unlike our approach, most of these models focus on features rather than relations. A representative example is Anderson’s rational analysis of categorization, which takes a matrix of objects and features as its input and organizes the objects into categories so that members of the same category tend to have similar patterns of features. Our model extends this idea to relational data, and organizes objects into categories so that members of the same category tend to be related in similar ways to other objects in the domain.

Although most models of categorization focus on features, several authors have discussed the difference between relational categories and feature-based categories. Goldstone (1996) distinguishes between isolated and interrelated categories: isolated categories can be understood in isolation, but interrelated categories depend on other categories for their meaning. Closer to our work are proposals about role-governed categories, or categories defined by the roles they play in a system of relations (Markman & Stilwell, 2001; Goldstone & Rogosky, 2002; Gentner & Kurtz, 2005; Jones & Love, 2007). For example, private may be defined by its relationships to other military concepts—in particular, privates take orders from corporals and sergeants (Markman & Stilwell, 2001). Our work is motivated by similar intuitions, and our approach can be viewed as a model for learning role-governed categories. Previous models have addressed related issues. Goldstone and Rogosky (2002) describe a model that identifies pairs of concepts which play corresponding roles in two conceptual systems, and Jones and Love (2007) describe a model where the similarity of two concepts increases if they play similar relational roles.

Although role-governed categories have been discussed by several authors, there are few experimental studies that explore how these categories might be learned. Other than our own experiments, the most relevant study was carried out by Larkey, Narvaez, and Markman (2004), who explored a setting where objects varied in their perceptual features (e.g. some were blue and

others orange) and in their relational roles (some objects pushed others). The majority of participants categorized these objects based on their roles rather than their features, but additional studies are needed to explore the tradeoff between features and relations in more detail.

A role-governed category is defined by its role in a relational system, but accounts of relational categorization often focus on cases where a single category corresponds to a structured relational system (Rehder & Ross, 2001; Kittur, Hummel, & Holyoak, 2004; Gentner & Kurtz, 2005; Tomlinson & Love, 2006). *Robbery*, for example, can be viewed as a system that specifies a relationship between a thief, a victim, and the goods that were stolen (Gentner & Kurtz, 2005). Some authors have argued that many categories correspond to causal systems: for instance, *bird* may be represented as a causal network which specifies that flying and living in trees are causally related (Ahn, Kim, Lassaline, & Dennis, 2000; Rehder, 2003). Our model focuses on role-governed categories, and these categories are importantly different from categories that correspond to relational systems in their own right, but both kinds of categories depend critically on relational information.

Models of analogy also emphasize the importance of relations (Falkenhainer et al., 1989; Holyoak & Thagard, 1989; Hummel & Holyoak, 2003; Goldstone & Rogosky, 2002), and psychologists have argued that analogy and categorization may be intimately linked (Gentner & Markman, 1997). The learning problem we considered, however, is rather different from the problem typically addressed by models of analogical reasoning. Most of these models attempt to establish a mapping between two relational systems: for example, a mapping between the solar system and the atom. We focused on settings where a learner must construct a single relational system by discovering categories and the relationships between them. The structure-mapping engine, for instance, does not seem capable of converting the UMLS data we analyzed into a relational system like the network in Figure 9.

Although analogical models do not naturally address the problem of theory discovery, they may be able to handle some of the tasks explored in our behavioral experiments. When asked to

make inferences about the causal powers of a new object, for example, an analogical model might rely on a mapping between the relations observed for the new object and the relations observed for previous objects. In Test 2 of Experiment 1, for example, a learner may be able to predict unobserved interactions involving the new object  $x$  by using an analogy that maps object  $x$  onto one of the previously observed objects. Some account along these lines may be possible, but it is not clear whether existing analogical models will account for all of our results. In particular, it is not clear whether existing models will account for the finding that inferences become more confident as more blocks have been observed (Figures 14 and 15), since a perfectly good analogy can be established after observing the three blocks available in phase 1.

*AI, machine learning and statistics.*

Several fields have developed models of relational learning that are not explicitly presented as models of theory acquisition but can be interpreted in this fashion. Our work is related most closely to the stochastic blockmodel (Wang & Wong, 1987; Nowicki & Snijders, 2001), an approach used by statisticians and sociologists to discover structure in social networks. More recently, machine learning researchers have developed systems that discover structure in large relational data sets (Getoor, Friedman, Koller, & Taskar, 2002; Kok & Domingos, 2005), and applied them to citation databases (Taskar, Segal, & Koller, 2001), social networks (Kubica, Moore, Schneider, & Yang, 2002), and genomic data (Segal et al., 2003).

There are several formal approaches in AI and machine learning that explicitly address the problem of theory discovery. Most of these accounts focus on scientific theories: for example, DENDRAL has been used to understand the structure of chemical compounds, the BACON system (Langley, Simon, Bradshaw, & Zytkow, 1987) has been used to model the discovery of the ideal gas law, Ohm's law and the law of conservation of momentum, and inductive logic programming has been applied to problems in chemistry (King, Muggleton, Srinivasan, & Sternberg, 1996) and genetics (King et al., 2004).

Compared to most previous work on relational learning and scientific theory discovery, our

work is different in two key respects. First, we have focused on intuitive theories rather than scientific theories.<sup>5</sup> Most of the data sets we considered are motivated by everyday problems that children solve over the course of development, and we also showed that our model accounts for data collected in two behavioral experiments. Second, we presented a computational-level analysis of theory discovery (Anderson, 1991; Marr, 1982). Langley et al. (1987) focus on algorithmic accounts, and the complicated search heuristics used by these algorithms often obscure the computational theory (Marr, 1982) that is implicitly assumed. Understanding the process by which people construct intuitive theories is an important challenge for the future, but it is valuable first to abstract away the details of this process and to attempt to understand the computational problem of theory discovery.

Inductive Logic Programming (ILP) is an alternative approach to theory discovery that does qualify as a computational theory in Marr’s sense (Muggleton & De Raedt, 1994). ILP systems work with theories represented as logic programs, or sets of definite clauses. Given a set of observations, ILP systems attempt to find the simplest theory that accounts for the data, and ILP can be given a principled formulation using the Minimum Description Length (MDL) principle (Conklin & Witten, 1994). ILP systems have been applied to scientific problems in biology and chemistry, and the basic idea behind ILP can also help to explain how intuitive theories are learned (Kemp et al., 2008a).

The MDL principle is closely related to Bayesian inference (Chater, 1996; MacKay, 2003), and the ILP approach to theory discovery is similar in spirit to the approach we presented. There are some important practical considerations, however, which mean that off-the-shelf ILP systems are unlikely to work well on the problems we considered. In order to match the performance of our model, an ILP system must support predicate invention (Stahl, 1993), and must rely on a probabilistic semantics that allows it to deal gracefully with noise and exceptions. Predicate invention is crucial since part of our goal is to discover novel concepts, and new predicates will be needed to refer to these concepts. Although there are ILP systems that handle predicate

invention, and there have been attempts to combine logic programs with probabilities (De Raedt & Kersting, 2003), both technologies are relatively immature. The reason why our model can improve on current ILP systems is the familiar tradeoff between complexity and learnability. Since our model only considers relatively simple theories, we can give a principled account of how these theories might be acquired. ILP systems, however, must consider a much bigger space of theories, and learning these theories is substantially more difficult.

*A solution to Fodor's puzzle?*

Concept learning and theory learning are intimately related, and our model helps to address a puzzle about concept learning that has become known as Fodor's puzzle (J. A. Fodor, 1975; Laurence & Margolis, 2002). A standard view of concept learning holds that people acquire new concepts by combining concepts that they already possess (Laurence & Margolis, 2002). Under this view, any given concept is either unlearned (i.e. innate) or structured (i.e. constructed of more primitive concepts). Fodor argues that most lexical concepts are unstructured, and concludes that most lexical concepts are unlearned primitives. The claim that concepts like carburetor, coal, and electron are not learned is highly counterintuitive, and has provoked a great deal of critical discussion (Putnam, 1991; Laurence & Margolis, 2002).

Block (1986) has proposed a solution to Fodor's puzzle that relies on conceptual role semantics, or the idea that that concepts derive their meaning from the roles that they play in systems of concepts. Conceptual role semantics allows for the possibility that all of the concepts in a novel system can be simultaneously acquired, where the content of each novel concept depends on its relationships to the other concepts in the system. Learning of this kind provides a counterexample to the claim that learned concepts must be compositions of pre-existing concepts, and therefore undermines a key premise in Fodor's argument.

Although conceptual role semantics is directly relevant to Fodor's puzzle, this approach has been criticised for being incomplete on several grounds (J. Fodor & Lepore, 1991; Laurence & Margolis, 2002), and Block (1998) himself has written that conceptual role semantics is "more of

a framework for a theory than an actual theory” (p 656). An important element missing from Block’s response to Fodor is a concrete computational account of how systems of concepts might be collectively learned. Most existing models of concept learning (Medin & Schaffer, 1978; Nosofsky, 1986; Anderson, 1991) will not qualify, since they assume that new concepts are composed out of pre-existing concepts (often called “features”). Unlike these previous models, our work supports the conceptual role semantics resolution of Fodor’s puzzle by showing how entire systems of novel concepts can be learned.

Even if humans learn entire systems of novel concepts, it might be questioned whether our model truly accounts for this learning. The model begins with a prior distribution over a space of possible theories, and learning is a matter of identifying the element in this space that best accounts for the observed data. Since the hypothesis space is defined before any data are observed, in one sense each possible theory is available from the start, together with all of the concepts that it specifies. This view of our model is partly true and partly false, and to explain why it is necessary to distinguish between two very different interpretations of the term “hypothesis space.”

From a computational perspective (Marr, 1982), every model of learning relies on a fixed hypothesis space that represents the abstract potential of the model. If we imagine all streams of input which the model could possibly receive, the hypothesis space includes all states of knowledge which the model could possibly reach. Often, however, psychologists prefer to reserve the term “hypothesis space” for the set of hypotheses that a learner actively entertains at a given moment. If we follow this second interpretation, than our model is not committed to a pre-defined space of hypotheses. Note, for example, that our implementation of the model does not enumerate and consider the set of all possible theories—this set would be far too large to handle. Instead, we developed a method for searching the space of possible theories, and this method will often end up in regions of the space very different from the region where it began (see Figure 3).

### *Psychologically plausible implementations*

Although we have not focused on the cognitive processes that support theory discovery, understanding how humans navigate the space of possible theories is an important direction for future work. Since our model builds on the formal machinery used by Anderson’s rational analysis of categorization, previous attempts to develop psychologically plausible implementations of that model (Anderson, 1991; Sanborn et al., 2006) can be extended to develop similar implementations of our model. In particular, we can develop a version of our model where the entities are observed one by one, and a single set of categories is maintained at each stage. Every time a new entity is observed, the entity is assigned to the category that best explains the relations in which it participates.

Eventually we would like to understand theory discovery at all of Marr’s (1982) levels, but the most direct path towards this goal probably begins with further work at the level of computational theory. We have argued that the relational systems discovered by our model deserve to be called theories, but these representations capture only some aspects of intuitive theories. An immediate goal is to develop models of theory discovery that account for a greater proportion of theory-like knowledge than our model is able to capture.

### *Future directions*

In some respects theory discovery is harder for humans than our model, but in other respects it may be substantially easier. We have focused on the problem of learning theories from raw relational data, but human learners are often directly provided with “theory fragments,” or components of the theory that they are trying to learn. The average physics student, for instance, probably learns more from theory fragments provided by her teacher than from observations of raw experimental data. Although we have not modeled the cultural transmission of theory fragments, our model should be able to incorporate direct statements about the theory it is attempting to discover. Consider again the symmetric activation theory ( $A \leftrightarrow B$ ) in our first

experiment. The model can take advantage of any statement that places constraints on the partition  $z$  and the parameter matrix  $\eta$  to be learned: for example, statements like “there are two categories,” or “this object and that object belong to the same category,” or “there are two categories and objects from one category only light up objects from the other category” can be used to restrict the hypothesis space that the model must consider. Since explicit instruction plays a role in so many cases of developmental interest, modeling learning when theory fragments and raw data are both available is an important direction for future work.

*Learning theories at multiple levels of abstraction.*

The theories discovered by our model are similar in many respects to framework theories (Wellman & Gelman, 1992), but future models should explore the acquisition of theories that are best described as specific theories. A natural first step is to embed our approach in a hierarchical model that also includes representations at lower levels of abstraction. For instance, the framework theory in Figure 1 should help a learner who is trying to discover specific theories that capture the relationships between individual chemicals, diseases, and symptoms (Tenenbaum & Niyogi, 2003). The framework theory rules out any specific theory where a symptom (e.g. coughing) causes a disease (e.g. lung cancer), and allows a learner to focus on more plausible theories (e.g. the theory that lung cancer causes coughing). Similarly, the framework theory in Figure 11a is useful when developing a specific theory that captures the kinship relationships between the individuals in a given family.

A hierarchical Bayesian approach (Gelman, Carlin, Stern, & Rubin, 2003) can be used to develop models that acquire both framework theories and specific theories. Suppose that the specific theories of interest can be represented using causal Bayesian networks. One of these networks, for instance, may specify which diseases cause which symptoms. Mansinghka, Kemp, Tenenbaum, and Griffiths (2006) develop a hierarchical Bayesian approach that uses our relational model to place constraints on the arrows in a Bayesian network. For instance, knowing that symptoms cause diseases allows this hierarchical model to search for networks where each

arrow extends from a specific disease to a specific symptom. Kemp, Goodman, and Tenenbaum (2007) show how a similar approach can be used to simultaneously learn abstract causal schemata and specific causal models. Hierarchical approaches like these can also be explored when the representations to be learned are more sophisticated than causal Bayes nets or the relational systems considered by our model.

*Towards richer theories.*

The relational representations considered by our model provide a useful starting point for models of theory discovery, but it will be important to work towards representations that can account for more of the content of intuitive theories. There are several natural extensions of our model that will discover more complex theories but should still remain relatively tractable. As previously mentioned, the animals in the biological data (Figure 7), the entities in the biomedical data (Figure 8) and the individuals in the kinship data (Figure 11) can be usefully organized into trees. In the kinship case, for instance, the tree may first divide the individuals into the four kinship sections, then split them up further according to age and gender. Roy, Kemp, Mansinghka, and Tenenbaum (2007) developed an extension of our model which assumes that relational data are generated over an underlying tree, and can discover the tree that best accounts for a given data set. A factorial model is the natural next step. This factorial model might be able to discover something like our “ground truth” partition for the kinship data: a system where the 16 categories are specified by three underlying partitions, one based on kinship section, one based on gender, and the third based on age. A final possible extension is a model that can discover multiple partitions of a given set of objects (Kok & Domingos, 2007). Shafto, Kemp, Mansinghka, Gordon, and Tenenbaum (2006) describe a model for feature data that partitions the features into categories, and discovers a separate partition of the objects for each of the feature categories introduced. The same idea can be applied more generally: given an  $n$ -place relation, we can choose some ordering of the dimensions, and discover a high-level partition of the first dimension, and separate lower-level partitions of the remaining dimensions for each category

in the high level partition.

Ultimately it will be necessary to develop models that rely on richer representations than those considered here. Intuitive theories of kinship, for example, are likely to call for logical representations, and logical theories are also useful for capturing abstract causal knowledge (Griffiths & Tenenbaum, 2007). Existing methods for learning logical theories provide a useful starting point (Muggleton & De Raedt, 1994; Kemp et al., 2007; Kemp, Goodman, & Tenenbaum, 2008b), but will need to be supplemented with effective techniques for predicate invention and for probabilistic inference over logical representations. Much work remains to be done, but eventually psychologists should aim to develop models that account for the acquisition of rich and complex theories.

## Conclusion

We presented a model that discovers simple theories, or systems of related concepts. Our model simultaneously discovers the concepts that exist in a domain, and the laws or principles that capture relationships between these concepts. Most previous models of concept formation are able only to discover combinations of pre-existing concepts. Unlike these approaches, our model can discover entire systems of concepts that derive their meaning from their relationships to each other.

Our model demonstrates that statistical inference can help to explain the acquisition of highly-structured representations: representations as sophisticated as intuitive theories. Theory discovery sometimes appears mysterious because many aspects of a theory appear to depend on each other: theoretical laws are defined using theoretical concepts, which in turn are defined by their participation in theoretical laws. Our model helps to dispel the sense of mystery by explaining how all of the pieces of a theory can be acquired together.

The theories we considered are all extremely simple, but future work can explore the acquisition of more sophisticated theories. Explaining real-world examples of theory acquisition is

undoubtedly a challenging problem, but many of the technical tools needed to address this problem may already exist. This paper has argued that Bayesian inference can explain the acquisition of richly-structured representations, and classic approaches to knowledge representation (Rumelhart, Lindsay, & Norman, 1972; Davis, 1990) have led to many proposals about how intuitive theories should be represented. Over the next decade it may prove possible to bring these insights together and to develop a comprehensive formal account of the acquisition of intuitive theories.

## References

- Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41*, 1–55.
- 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. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.
- Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, *2*, 1152–1174.
- Block, N. (1986). Advertisement for a semantics for psychology. In *Studies in the Philosophy of Mind: Midwest Studies in Philosophy* (Vol. 10). Minneapolis: University of Minnesota Press.
- Block, N. (1998). Conceptual role semantics. In E. Craig (Ed.), *Routledge encyclopedia of philosophy* (pp. 1–7). Oxford: Routledge.
- Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.
- Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*, 566–581.
- Conklin, D., & Witten, I. H. (1994). Complexity-based induction. *Machine Learning*, *16*(3), 203–225.
- Davis, E. (1990). *Representations of commonsense knowledge*. Morgan Kaufmann.
- De Raedt, L., & Kersting, K. (2003). Probabilistic logic learning. *ACM-SIGKDD Explorations*, *5*(1), 31–48.
- Denham, W. W. (1973). *The detection of patterns in Alyawarra nonverbal behavior*. Unpublished doctoral dissertation, University of Washington.

- Denham, W. W. (2001). Alyawarra ethnographic database: Numerical data documentation file, version 7. (Available at <http://www.alc.edu/denham/Alyawarra/>)
- Denham, W. W., McDaniel, C. K., & Atkins, J. R. (1979). Aranda and Alyawarra kinship: A quantitative argument for a double helix model. *American Ethnologist*, 6(1), 1–24.
- Denham, W. W., & White, D. R. (2005). Multiple measures of Alyawarra kinship. *Field Methods*, 17(1), 70–101.
- Dhillon, I. S., Mallela, S., & Modha, D. S. (2003). Information-theoretic co-clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 89–98). New York, NY, USA: ACM Press.
- Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41, 1–63.
- Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1, 209–230.
- Field, H. (1977). Logic, meaning and conceptual role. *The Journal of Philosophy*, 69, 379–408.
- Findler, N. (1992). Automatic rule discovery for field work in anthropology. *Computers and the Humanities*, 25, 295–392.
- Fodor, J., & Lepore, E. (1991). Why meaning (probably) isn't conceptual role. *Mind and Language*, 6(4), 328–343.
- Fodor, J., & Lepore, E. (1992). *Holism: A shopper's guide*. Cambridge, MA: Blackwell.
- Fodor, J. A. (1975). *The language of thought*. Cambridge: Harvard University Press.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall.
- Gentner, D., & Kurtz, K. (2005). Relational categories. In W. Ahn, R. L. Goldstone, B. C. Love,

- A. B. Markman, & P. W. Wolff (Eds.), *Categorization inside and outside the lab* (pp. 151–175).
- Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*, 45–56.
- Getoor, L., Friedman, N., Koller, D., & Taskar, B. (2002). Learning probabilistic models of link structure. *Journal of Machine Learning Research*, *3*, 679–707.
- Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory and Cognition*, *24*, 608–628.
- Goldstone, R. L., & Rogosky, B. J. (2002). Using relations within conceptual systems to translate across conceptual systems. *Cognition*, *84*, 295–320.
- Griffiths, T. L., & Tenenbaum, J. B. (2007). Two proposals for causal grammars. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation*. Oxford: Oxford University Press.
- Hanson, N. R. (1958). *Patterns of discovery*. Cambridge: Cambridge University Press.
- Hayes, P. J. (1985). The second naive physics manifesto. In J. R. Hobbs & R. C. Moore (Eds.), *Formal theories of the commonsense world*. Ablex.
- Hempel, C. G. (1972). *Fundamentals of concept formation in empirical science*. University of Chicago Press.
- Hempel, C. G. (1985). Thoughts on the limitations of discovery by computer. In K. Schaffner (Ed.), *Logic of discovery and diagnosis in medicine* (pp. 115–122). Berkeley: University of California Press.
- Hofmann, T., & Puzicha, J. (1999). Latent class models for collaborative filtering. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*. San Francisco, CA: Morgan Kaufman.

- Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, 13(3), 295–355.
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193–218.
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110, 220–264.
- 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.
- Johnson, S. (1755). *A dictionary of the English language*. London.
- Jones, M., & Love, B. C. (2007). Beyond common features: the role of roles in determining similarity. *Cognitive Psychology*, 55, 196–231.
- Keil, F. C. (1979). *Semantic and conceptual development*. Cambridge, MA: Harvard University Press.
- Keil, F. C. (1983). On the emergence of semantic and conceptual distinctions. *Journal of Experimental Psychology: General*, 112(3), 357–385.
- 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. (1993). The growth of causal understandings of natural kinds. In D. P. D. Sperber & A. J. Premack (Eds.), *Causal cognition: a multidisciplinary debate* (pp. 234–262). New York, NY: Oxford University press.
- Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2007). Learning causal schemata. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 389–394). Austin, TX: Cognitive Science Society.
- Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2008a). Learning and using relational theories.

- In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems 20* (pp. 753–760). Cambridge, MA: MIT Press.
- Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2008b). Theory acquisition and the language of thought. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1606–1611). Austin, TX: Cognitive Science Society.
- Kemp, C., Griffiths, T. L., & Tenenbaum, J. B. (2004). *Discovering latent classes in relational data* (Tech. Rep. No. 2004-019). MIT AI Memo.
- Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.
- 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*.
- King, R. D., Muggleton, S. H., Srinivasan, A., & Sternberg, M. J. E. (1996). Structure-activity relationships derived by machine learning: the use of atoms and their bond connectives to predict mutagenicity by inductive logic programming. *Proceedings of the National Academy of Sciences*, *93*, 438–442.
- King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., et al. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, *427*(6971), 247–252.
- Kittur, A., Hummel, J. E., & Holyoak, K. J. (2004). Feature vs relation defined categories: probab(alistic)ly not the same. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum Associates.
- Kok, S., & Domingos, P. (2005). Learning the structure of Markov logic networks. In *Proceedings of the 22nd International Conference on Machine Learning*.

- Kok, S., & Domingos, P. (2007). Statistical predicate invention. In *Proceedings of the 24th 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*.
- Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.
- Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: computational explorations of the creative process*.
- Larkey, L. B., Narvaez, L., & Markman, A. B. (2004). Categories among relations. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (p. 1634). Lawrence Erlbaum Associates.
- Laurence, S., & Margolis, E. (2002). Radical concept nativism. *Cognition*, 86, 25–55.
- MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.
- Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE transactions on computational biology and bioinformatics*, 1(1), 24–45.
- Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, 23, 250-269.
- 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)*.
- Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental and Theoretical Artificial Intelligence*, 13(4), 329–358.
- Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

- McCray, A. T. (2003). An upper level ontology for the biomedical domain. *Comparative and Functional Genomics, 4*, 80–84.
- McCray, A. T., Burgun, A., & Bodenreider, O. (2001). Aggregating UMLS semantic types for reducing conceptual complexity. In *Studies in health technology and informatics* (Vol. 84 (Pt 1), pp. 216–20).
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.
- 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. M. (1991). *Bayesian mixture modeling by Monte Carlo simulation* (Tech. Rep. No. 91-2). University of Toronto.
- Newton-Smith, W. H. (1981). *The rationality of science*. London: Routledge and Kegan Paul.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.
- Nowicki, K., & Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association, 96*(455), 1077–1087.
- Osherson, D. N., Stern, J., Wilkie, O., Stob, M., & Smith, E. E. (1991). Default probability. *Cognitive Science, 15*, 251-269.
- Pitman, J. (2002). *Combinatorial stochastic processes*. (Notes for Saint Flour Summer School)
- Popper, K. R. (1935/1980). *The logic of scientific discovery*. Boston, MA: Hutchinson.
- Putnam, H. (1991). *Representation and reality*. Cambridge, MA: MIT Press.

- Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing*. Cambridge, MA: MIT Press.
- Quine, W. V. O., & Ullian, J. (1978). *The web of belief*. New York: Random House.
- Quinlan, J. R. (1990). Learning logical definitions from relations. *Machine Learning*, 5(3), 239–266.
- Rasmussen, C. E. (2002). The infinite Gaussian mixture model. In *NIPS* (Vol. 13).
- Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 1141–1159.
- Rehder, B., & Ross, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 27, 1261–1275.
- Rips, L. J. (1995). The current status of research on concept combination. *Mind and Language*, 10(1/2), 72–104.
- 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.
- Roy, D. M., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2007). Learning annotated hierarchies from relational data. In *Advances in Neural Information Processing Systems 20*.
- Rumelhart, D. E., Lindsay, P., & Norman, D. A. (1972). A process model for long-term memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 197–246). New York: Academic Press.
- Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science society*.
- Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., et al. (2003). Module

- networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(166-176).
- 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).
- Shapiro, E. Y. (1991). Inductive inference of theories from facts. In J. Lassez & G. Plotkin (Eds.), *Computational logic: essays in honor of Alan Robinson* (pp. 199–254). Cambridge, MA: MIT Press.
- Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, 237, 1317-1323.
- Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, 14, 29–56.
- Stahl, I. (1993). Predicate invention in ILP—an overview. In *ECML 93: Proceedings of the European Conference on Machine Learning*. London, UK: Springer-Verlag.
- 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). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24, 629-641.
- Tenenbaum, J. B., & Niyogi, S. (2003). Learning causal laws. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 1152–1157).
- Tomlinson, M., & Love, B. C. (2006). Learning abstract relations through analogy to concrete exemplars. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science society* (pp. 2269–2274). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wang, Y. J., & Wong, G. Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82, 8–19.

- 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).
- Woodfield, A. (1987). On the very idea of acquiring a concept. In J. Russell (Ed.), *Philosophical perspectives on developmental psychology*. Oxford: Basil Blackwell.

## Appendix A

### Theory acquisition and use

We previously introduced our model in a setting where  $R$  is a binary relation over a single set of entities. Using statistical notation, the model can be written as follows:

$$\begin{aligned} z | \gamma &\sim \text{CRP}(\gamma) \\ \eta(A, B) | \alpha, \beta &\sim \text{Beta}(\alpha, \beta) \\ R(i, j) | z, \eta &\sim \text{Bernoulli}(\eta(z_i, z_j)), \end{aligned} \tag{5}$$

where  $A, B \in \mathcal{N}$ .

The first line indicates that  $z$  is drawn from a Chinese Restaurant Process (Aldous, 1985) with hyperparameter  $\gamma$ . Imagine building a partition  $z$  from the ground up: starting with a category containing a single entity, and adding entities one by one until all the entities belong to categories. Under the CRP, each category attracts new members in proportion to its size. The distribution over categories for entity  $i$ , conditioned on the categories of entities  $1, \dots, i-1$  is

$$p(z_i = A | z_1, \dots, z_{i-1}) = \begin{cases} \frac{n_A}{i-1+\gamma} & n_A > 0 \\ \frac{\gamma}{i-1+\gamma} & A \text{ is a new category} \end{cases} \tag{6}$$

where  $z_i$  is the category assignment for entity  $i$  and  $n_A$  is the number of entities already assigned to category  $A$ . The CRP is *exchangeable*: the order in which entities are assigned to categories can be permuted without changing the probability of the resulting partition.  $P(z)$  can therefore be computed by choosing an arbitrary ordering and multiplying conditional probabilities specified by Equation 6. Since new entities can always be assigned to new categories, our model effectively has access to a countably infinite collection of categories. In recognition of this property, we refer elsewhere to our model as the Infinite Relational Model (Kemp et al., 2006).

The second line in Equation 5 indicates that the entries in  $\eta$  are drawn independently from a Beta distribution with hyperparameters  $\alpha$  and  $\beta$ . If variable  $x$  is drawn from this Beta

distribution, then

$$p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad (7)$$

where  $\Gamma(\cdot)$  is the gamma function. For all applications in this paper we use a symmetric prior and set  $\alpha = \beta$ . The final line in Equation 5 indicates that entry  $R(i, j)$  is generated by tossing a coin with bias  $\eta(z_i, z_j)$ .

To formulate the most general version of our model we extend Equation 5 to relations of arbitrary arity. Consider an  $m$  dimensional relation  $R$  involving  $n$  different types, where  $T^j$  is the  $j$ th type, and  $z^j$  is a vector of category assignments for  $T^j$ . Let  $d_k$  be the label of the type that occupies dimension  $k$  of the relation: for example, the three place relation

$R : T^1 \times T^1 \times T^2 \rightarrow \{0, 1\}$  has  $d_1 = d_2 = 1$ , and  $d_3 = 2$ . As before, the probability that the relation holds between a group of objects depends only on the categories of those objects:

$$R(i_1, i_2, \dots, i_m) | z^1, z^2, \dots, z^n, \eta \sim \text{Bernoulli}(\eta(z_{i_1}^{d_1}, z_{i_2}^{d_2}, \dots, z_{i_m}^{d_m})).$$

In settings with multiple relations we introduce a parameter matrix  $\eta^j$  for each relation  $R^j$ . Given this generative model, we aim to discover the category assignments  $\{z^j\}$  and the parameters  $\{\eta^j\}$  with maximum posterior probability given the data (Equation 2).

To discover the relational system that best explains a given data set, we initially integrate out the parameter matrices  $\{\eta^j\}$  and search for the category assignments with maximum posterior probability. Once the category assignments are known it is simple to compute the most likely matrices  $\{\eta^j\}$  given these assignments and the original data. For simplicity, we describe our search algorithm for the case when there is a single binary relation  $R$  defined over a single type  $T$ . The techniques we describe extend naturally to situations where there are multiple types and relations. Since we use conjugate priors (i.e. Beta priors) on the entries in  $\eta$ , it is simple to

compute  $P(R|z) = \int P(R|\eta, z)p(\eta)d\eta$ :

$$p(R|z, \eta) = \prod_{A, B \in \mathcal{N}} \eta(A, B)^{m(A, B)} (1 - \eta(A, B))^{\bar{m}(A, B)} \quad (8)$$

$$p(R|z) = \prod_{A, B \in \mathcal{N}} \frac{B(m(A, B) + \alpha, \bar{m}(A, B) + \beta)}{B(\alpha, \beta)} \quad (9)$$

where  $m(A, B)$  is the number of pairs  $(i, j)$  where  $i \in A$  and  $j \in B$  and  $R(i, j) = 1$ ,  $\bar{m}(A, B)$  is the number of these pairs where  $R(i, j) = 0$ , and  $B(\cdot, \cdot)$  is the Beta function. If some entries in  $R$  are missing at random, we can ignore them and maintain counts  $m(A, B)$  and  $\bar{m}(A, B)$  over only the observed values.

Since we integrate out  $\eta$ , inference can be carried out using Markov chain Monte Carlo methods to sample from the posterior on category assignments  $P(z|R) \propto P(R|z)P(z)$  (Jain & Neal, 2004; Kemp, Griffiths, & Tenenbaum, 2004) or by searching for the mode of this distribution. For most applications in this paper (Figures 7, 8, 9 and 11) we are interested only in the relational system that best explains the available data, and we search for the maximum *a posteriori* partition  $z$  using hill-climbing with restarts. We find that this algorithm works well when the partition used to initialize each restart contains only one category.

Our hill-climbing algorithm uses operations that move an object from one cluster to another, split a cluster, or merge two clusters. The goal of our model can be understood intuitively by representing the relation  $R$  as an adjacency matrix (Figure 3). Our search procedure tries to shuffle the rows and columns of this matrix so that it assumes a clean block structure. The same idea applies to relations with more than two dimensions: Figure 5b shows a ternary relation, and here the aim is to shuffle the dimensions so that the matrix takes on a three dimensional block structure. Figure 5c shows two relations involving two types. The goal is again to create matrices with clean block structures, but now the partition for  $T^1$  must be the same wherever this type appears.

When searching for the partition that maximizes  $P(z|R)$ , we avoid free parameters by learning the hyperparameters  $\gamma$  and  $\alpha$  for each data set (recall that we set  $\beta = \alpha$ ). We use an

Dataset	Algorithm	Iterations	$\alpha$	$\gamma$
Biological	Hill climbing	3000	0.35	animals: 3.0 features: 10.4
UMLS	Hill climbing	3000	0.01	entities: 2.7 predicates: 6.5
Alyawarra	Hill climbing	3000	0.04	people: 3.2 kin terms: 8.1
Experiments 1 and 2	MCMC	2000	1	objects: 1

Table A1

*Details of the analyses reported in this paper. We used hill climbing to find the best relational system for each of the first three data sets, and Markov chain Monte Carlo methods to generate model predictions about our two experiments. Hyperparameters for the first three data sets were learned, and the hyperparameters for each MCMC simulation were set to 1.*

exponential prior with parameter 1 on  $\gamma$ , and an improper prior  $p(\alpha) \propto \alpha^{-\frac{5}{2}}$ . There is a separate hyperparameter  $\gamma$  for each type of entities, and we assume that all of these variables are independent. The hyperparameters discovered are shown in Table A1. The  $\alpha$  values chosen by the model are smaller for the UMLS data and for the Alyawarra data than the biological data, indicating that these first two data sets have a cleaner block structure than the biological data. The  $\gamma$  values chosen are largest for types that are organized into a relatively large number of categories. For instance, the 85 features in the biological data are organized into 34 categories, and the  $\gamma$  value for this type is correspondingly high.

So far we have seen how the category assignments  $z$  can be discovered given the raw data in  $R$ . Given  $z$ , it is straightforward to recover the matrix  $\eta$  which specifies how the categories in  $z$  relate to each other. The maximum *a posteriori* value of  $\eta(A, B)$  given  $z$  and  $R$  is

$$\frac{m(A, B) + \alpha}{\bar{m}(A, B) + m(A, B) + \alpha + \beta} \tag{10}$$

where  $m(A, B)$  and  $\bar{m}(A, B)$  are defined as for Equation 9.

Our approach to the problem of theory acquisition can now be sharply formulated. A theory is a pair  $T = (z, \eta)$ , and the theory that best accounts for the observed data is the theory that maximizes  $p(T|R)$ . Our approach to the problem of theory use can be similarly formulated.

Suppose that the data  $R$  include missing entries. A theory  $T$  makes predictions about the values of these missing entries: for example, if  $R(i, j)$  is unobserved,

$$P(R(i, j) = 1|T, R) = P(R(i, j) = 1|T) \tag{11}$$

$$= \frac{m(z_i, z_j) + \alpha}{\bar{m}(z_i, z_j) + m(z_i, z_j) + \alpha + \beta}. \tag{12}$$

Instead of committing to a single theory and using it to predict unobserved relationships, a fully Bayesian learner should consider the predictions of all possible theories, weighting each one by its posterior probability:

$$P(R(i, j) = 1|R) = \int P(R(i, j) = 1|T)p(T|R)dT \tag{13}$$

The model predictions in Figures 14, 15 and 17 were computed by sampling from the posterior  $P(z|R)$  and using these samples to approximate Equation 13. We ran a separate simulation for each test in each phase of each experiment. Each simulation included 2000 iterations, of which 200 were discarded as burn-in, and both hyperparameters ( $\alpha$  and  $\gamma$ ) were set to 1. The source code for our model is available at [www.charleskemp.com](http://www.charleskemp.com), and the details in Table A1 should allow our results to be reproduced.

## Appendix B

### Alyawarra Kinship Terms

Abmarliya	MB, SWB (ms)
Aburliya	FM/FMB, FMBSD/FMBSS, ZSS/ZSD (ms), SS/SD (fs)
Adardiya	MF/MFZ, DS/DD, BDS/BDD (fs)
Adiadya	YB/YZ
Adniadya	MBS
Agniya	F
Aidmeniya	MMBSS/MMBSD, ZDS/ZDD (ms), DS/DD (fs)
Aiyenga	“Myself”
Aleriya	S/D (ms), BS/BD (fs)
Algyeliya	FZD/MBD
Amaidya	M, SW (ms)
Amburniya	WB/ZH
Andungiya	HZ/BW (fs)
Aneriya	BWM/DHZ (fs)
Angeliya	FZS/MBS
Agenduriya	ZS/ZD (ms), rare term for biological sister’s child
Anguriya	EZ
Anowadya	W/MMBDD (ms), H/MFZDS (fs)
Anyaina	MM/MMB, MMBSS/MMBSD, ZDS/ZDD (ms), DS/DD (fs)
Arengiya	FF/FFZ, SS/SD (ms), BSS/BS (fs)
Awaadya	EB
Aweniya	FZ, FMZD
Gnaldena	YZ, rare term for biological younger sister
Muriya	MMBD/MMBS, WM/WMB (ms), ZDH/ZDHZ (ms)
Umbaidya	S/D (fs), ZS/ZD (ms), FMBS/FMBD
Undyaidya	WZ (ms), rare term used as reciprocal for Amburniya

Table B1

*Glosses given by Denham (2001) for the 26 kinship terms in Figure 11. F=father, M=mother, B=brother, Z=sister, S=son, D=daughter, H=husband, W=wife, E=elder, Y=younger, fs=female speaker, ms=male speaker. For example, Adiadya refers to a classificatory younger brother (YB) or younger sister (YZ).*

## Author Note

Earlier versions of this work were presented at the Twenty-Fifth Annual Conference of the Cognitive Science Society (2003), the 46th Annual Meeting of the Psychonomic Society (2005), and the 21st National Conference on Artificial Intelligence (AAAI-06). Our AAAI paper was prepared in collaboration with Takeshi Yamada and Naonori Ueda. We thank Steven Sloman for sharing his copy of the feature data described in Osherson et al. (1991) and Woodrow Denham for providing his Alyawarra data in machine-readable form. We also thank Steven Sloman, Michael Strevens, and several anonymous reviewers for comments on the manuscript. This work was supported in part by AFOSR MURI contract FA9550-05-1-0321, AFOSR contract FA9550-07-2-0351, AFOSR contract FA9550-07-1-0075, NSF grant CDI-0835797, NTT Communication Sciences Laboratory, the James S. McDonnell Foundation Causal Learning Research Collaborative, the William Asbjornsen Albert memorial fellowship (CK) and the Paul E. Newton Chair (JBT).

## Footnotes

<sup>1</sup> Simultaneously clustering objects and features is sometimes called biclustering or coclustering, and previous coclustering models have been developed by statisticians and machine learning researchers (Hofmann & Puzicha, 1999; Dhillon, Mallela, & Modha, 2003; Madeira & Oliveira, 2004). To our knowledge, however, the problem of coclustering has received little previous attention in the psychological literature.

<sup>2</sup> The feature-based model is also known in some literatures as the infinite mixture model or the Dirichlet process mixture model (Ferguson, 1973; Neal, 1991)

<sup>3</sup> Denham and colleagues have also noted that Alyawarra practice occasionally departs from both the Kariara and Aranda systems (Denham et al., 1979; Denham & White, 2005). Both of these normative systems rule out marriages that are seen rarely in practice.

<sup>4</sup> Note that there was only a single probe object during each phase, and that this probe object was either an *A*-object or a *B*-object. For this reason, only half of the model predictions plotted in Figures 14 and 15 actually correspond to cases explored during our experiment, and each of the human curves has half as many points as the corresponding model curve.

<sup>5</sup> AI researchers often discuss intuitive theories—consider, for example, the extensive literature on naive physics (Hayes, 1985)—but models of theory discovery have tended to focus on scientific theories.