

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Nonparametric Hierarchical Bayesian Models of Categorization

Permalink

<https://escholarship.org/uc/item/9v9457ng>

Author

Canini, Kevin

Publication Date

2011

Peer reviewed|Thesis/dissertation

Nonparametric Hierarchical Bayesian Models of Categorization

by

Kevin Robert Canini

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Thomas L. Griffiths, Co-chair

Professor Stuart J. Russell, Co-chair

Professor Michael I. Jordan

Professor Tania Lombrozo

Fall 2011

Nonparametric Hierarchical Bayesian Models of Categorization

Copyright © 2011

by

Kevin Robert Canini

Abstract

Nonparametric Hierarchical Bayesian Models of Categorization

by

Kevin Robert Canini

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Thomas L. Griffiths, Co-chair

Professor Stuart J. Russell, Co-chair

Categorization, or classification, is a fundamental problem in both cognitive psychology and machine learning. Classical psychological models of categorization fall into two main groups: prototype models and exemplar models, which are equivalent, respectively, to the statistical methods of parametric density estimation and kernel density estimation. Many categorization studies in psychology attempt to understand how people solve this problem by comparing their inferences to those of formal computational models such as prototype or exemplar models. From this perspective, different models make different predictions about the representations and mechanisms people use to make categorization judgments. Instead, one can seek to understand categorization by viewing it as a problem of statistical inference and attempting to characterize the inductive biases of human learners. These inductive biases can be directly exposed using an experimental method called iterated learning, which provides direct insight into human categorization in a way that is independent of any proposed models. I describe the results of an iterated learning study of human categorization which supports previous findings by psychologists that people's representations seem to be more flexible than would be implied by either prototype or exemplar models alone.

Prototype and exemplar models both use a single, fixed level of complexity in their representations of categories, with prototype models exhibiting the simplest representations, and exemplar models using the most complex representations. Treating categorization as a type of statistical inference, I describe a family of nonparametric Bayesian models of categorization based on the Dirichlet process mixture model (DPMM). These models represent categories as combinations of clusters of objects and, together, produce a continuum of representational complexities where prototype and exemplar models are special cases, occupying opposite ends of the spectrum. DPMM models allow the level of complexity of category representations to be chosen to suit the task at hand or to change over time; this flexibility can explain psychological results demonstrating that people's inferences are more congruent

with prototype models at some times and exemplar models at other times.

The DPMM can be generalized into a larger framework of models based on the hierarchical Dirichlet process (HDP). The HDP subsumes the DPMM and multiple previous psychological models, including prototypes, exemplars, and the Rational Model of Categorization. In addition, the HDP contains a family of previously unexplored models which make interesting predictions about how information can be shared between multiple categories. While most other categorization models learn each individual category in isolation and independently of the others, these HDP models share information between categories. This sharing of information can improve the speed and accuracy of learning and explained certain transfer learning effects that were observed in people's judgments. I introduce an extension of the HDP, called the tree-HDP, which is designed to infer systems of hierarchically related categories. The tree-HDP is able to simultaneously learn categories at multiple levels of generality and infer the taxonomic relationships between them.

The original scientific contributions of this dissertation are a detailed characterization of the inductive biases of human categorization via iterated learning, a unification of previous psychological models of categorization into a common Bayesian statistical framework (the HDP), a demonstration that this framework contains interesting and previously unexplored models that predict and explain the integration of information from multiple categories, and a proposal and exploration of a new statistical model, the tree-HDP, which can simultaneously learn categories at multiple hierarchical levels and infer taxonomic relationships between those categories.

To Mom, Dad, Nanny, & Papa.

Contents

1	Introduction	1
2	Characterizing human categorization	5
2.1	Iterated learning	6
2.2	Multidimensional scaling study of stimuli	7
2.2.1	Method	7
2.2.2	Results	10
2.3	Iterated learning experiment	10
2.3.1	Method	12
2.3.2	Results	13
2.4	Characterizing human inductive biases	13
2.4.1	Visualizing inductive biases	15
2.4.2	Aggregate category complexity	17
2.5	Conclusions	20
3	Statistical models	21
3.1	Prototypes, exemplars, and clusters	24
3.1.1	Prototypes and exemplars	24
3.1.2	Cluster-based models	26
3.2	Rational models of categorization	26
3.2.1	The rational basis for prototype and exemplar models	27
3.2.2	The Mixture Model of Categorization	27
3.2.3	The Rational Model of Categorization	28
3.3	Dirichlet process mixtures: nonparametric Bayes and the RMC	30
3.4	Hierarchical Dirichlet processes	32
3.5	Unifying previous models using hierarchical Dirichlet processes	34
3.6	Cluster-sharing in the HDP _{+,+} model	35
3.7	Exploring cluster sharing with human learners	36
3.7.1	Method	37
3.7.2	Results	43
3.7.3	Modeling	45

3.8	Discussion	47
3.8.1	Other forms of transfer learning	48
3.8.2	Connections to rule-based models	49
3.8.3	Limitations and extensions	50
3.8.4	Conclusion	51
4	Multi-level category learning	52
4.1	Multi-level categories and the tree-HDP	53
4.1.1	The tree-HDP model	53
4.1.2	Inference in the tree-HDP	54
4.1.3	Evaluation on recovering simulated taxonomies	56
4.2	Experiment 1: Comparing to human learners	58
4.2.1	Method	58
4.2.2	Results	62
4.2.3	Modeling	65
4.3	Experiment 2: Testing implicit knowledge and inferences	66
4.3.1	Method	66
4.3.2	Results	67
4.4	Experiment 3: Learning unbalanced taxonomies	70
4.4.1	Method	70
4.4.2	Results	71
4.4.3	Tree-HDP model predictions	74
4.5	Conclusion	75
5	Conclusions	76
5.1	Limitations and future directions	77
5.1.1	Spatial representations	77
5.1.2	Clusters as the atomic unit	78
5.1.3	Cluster-based and rule-based representations	79
5.1.4	Bayesian inference algorithms	79
5.2	Conclusion	80
	References	82

Acknowledgements

My time at Berkeley has been incredibly educational, challenging, and rewarding. I have grown tremendously both as a scholar and an overall person. Berkeley has the greatest concentration of brilliant, talented, and wonderful people I have ever met, and many of them deserve to be acknowledged for their role in my accomplishments.

An enormous amount of credit is due to Tom Griffiths for the successful completion of my doctoral degree. Despite never officially being my graduate advisor, he has served outstandingly in every aspect of that role. I am deeply indebted to him for guiding me through the many challenges of graduate school. He is a model professor and demonstrates intelligence, thoughtfulness, kindness, humor, and care in every interaction with his colleagues and students. I am also grateful to Stuart Russell for holding me to a very high standard throughout graduate school. He always provided a rational perspective on my research and taught me to question even the most basic assumptions of my work.

All of the work described in this dissertation was completed in conjunction with my advisor, Tom Griffiths. In addition, Wolf Vanpaemel and Mike Kalish contributed to the work described in Chapter 2, and Mikhail Shashkov to that in Chapter 3.

I have had many outstanding labmates in both the Griffiths and Russell labs. I could always count on being able to bounce ideas off of Jason Wolfe and have them returned significantly improved; he has a unique talent for quickly understanding other people's research problems and providing thoughtful analyses that demonstrate as much insight as if they were his own. I thank Chris Lucas for technical discussions that never failed to make me think deeply, for being an excellent roommate at both my first and last academic conferences, and for making the experience of sharing a single desk between two people surprisingly painless. In addition to being excellent researchers, Josh Abbott, Chris Holdgraf, and Mike Pacer are an amazing comedic trio. I am disappointed that I was able to spend only three semesters with them; I could always count on them to provide an endless stream of jokes which became exponentially more entertaining as more of us were in the room (I apologize to anyone who ever wandered into the lounge while all four of us were there simultaneously). I would also like to thank my research assistant Misha Shashkov, who helped a great deal in designing, programming, and running some of the experiments described in this dissertation.

Graduate school would have been far less enjoyable had it not been for my musical experiences at Berkeley. I thank Greg McLaskey and Mike Long for our many afternoons spent playing music, picking plums, and enjoying the sunshine in Ohlone park. I am also especially indebted to the talented Bill Ganz, who served as the musical director for my three years in the UC Men's Chorale and my three years in the Cal Jazz Choir. He is a role model for his appreciation and talent for music, as well as his patience and passion for teaching.

I thank Asya Ofshteyn for being there for me and providing support through my most difficult times. Finally, my parents and grandparents have given me love and encouragement throughout my life, and I would not have made it nearly this far without them.

Chapter 1

Introduction

Categorization is one of the most fundamental abilities of intelligent systems. It allows agents to make sense of a large number of high-dimensional, noisy observations by reducing them to a much smaller number of discrete categories. Categorization is an integral part of many higher-level operations, including understanding and using language (Gelman and Coley, 1991), reasoning about the consequences of objects and actions in the world (Shepard, 1987), interpreting visual information about one’s environment (Huttenlocher et al., 2000), and processing auditory sensory input (Feldman et al., 2009). Organizing objects and situations into coherent groups allows one to build models of the environment and to make predictions about the future by generalizing from past experiences. Because of its basic importance, categorization is actively studied in many fields, including psychology, cognitive science, artificial intelligence, machine learning, and statistics. In more computationally-oriented fields, such as statistics, artificial intelligence, and machine learning, categorization is more often called “classification”. In these disciplines, it is often distilled into the purely computational problem of optimizing the assignment of category labels to objects, where there is always exactly one correct answer. In the psychological sciences, the study of categorization can be more multifaceted, reflecting the variety of ways and situations in which humans learn to perform the task. For example, a seminal study of human categorization by Rosch et al. (1976) explored the people’s tendency to categorize objects at a single *basic level* of a multi-leveled taxonomy of category labels. However, many formal psychological models eschew these nuances and focus on the core problem of classification.

The major difference between the study of categorization in the computational fields and in the psychological fields is the overall goal of the research. In the computational fields, one is typically concerned with building algorithms that optimize some objective measure of performance; this is usually the number of test items correctly classified. In psychological studies, the primary aim is to understand the ways in which humans solve categorization problems. These two goals can often be aligned: most psychological studies of categorization proceed by comparing the performance of formal computational models to that of human experimental subjects; whenever people are highly competent at a categorization task, this

leads psychological modelers to attempt to build systems that can replicate people’s optimal (or near-optimal) performance. However, in studying human cognition, one may discover that people tend to have certain systematic biases, inefficiencies, or inadequacies. In these cases, psychological modelers may be more concerned with replicating the unique characteristics of human learning, whereas models in the purely computational fields would often be specifically designed to avoid those pitfalls. Nevertheless, focusing on human learning can lead to new insights and advances in computational models. Human behavior that might at first seem suboptimal may be a reflection of hidden assumptions or complexities in a problem that were missing from the modeler’s interpretation of the problem. Categorization problems that people face in the real world are often complex and challenging; by studying these tasks and the ways in which people solve them, one can find inspiration for building better computational models.

One of the central tenets of this dissertation is that it is possible for the psychological sciences and the computational sciences to each benefit from advances in the others. Although they have typically progressed as independent fields of study without much interaction, they often attempt to solve the same problems from different perspectives. This is especially evident in the disciplines of computational cognitive science, which directly draws on tools from computer science and statistics to build models of human cognition, and artificial intelligence and machine learning, which are directly rooted in attempts to systematize human reasoning abilities (Newell and Simon, 1956). Indeed, people and computers solve many of the same types of problems, and much effort has been directed towards building automated systems that can accomplish complex tasks that are routine for human beings, such as understanding speech (Jurafsky et al., 2000), playing the game show *Jeopardy!* (Ferrucci et al., 2010), and folding t-shirts (Cusumano-Towner et al., 2011). The work described in this dissertation demonstrates that the study of human learning and the study of machine learning can be mutually beneficial. I use statistical models to explain phenomena in human categorization that have not been captured by previous psychological models, and in turn, I use people’s ability to solve complex tasks as inspiration for building more advanced statistical models. In this way, my work progresses both the psychological and the computational sciences.

In fact, there has been a history of great overlap between the techniques used by psychological and statistical models for solving categorization problems. People’s ability to learn categories is the subject of some of the earliest studies of human cognition by psychologists (Hull, 1920). Throughout this long history of investigation, numerous computational models of category learning have been proposed, including approaches based on decision rules (Ashby and Gott, 1988), prototypes (Reed, 1972), and exemplars (Medin and Schaffer, 1978; Nosofsky, 1986), as well as combinations of these techniques (Nosofsky et al., 1994; Anderson, 1990, 1991). Many of these models were described without making any explicit reference to the fact that the techniques being used by psychologists in modeling human categorization were directly equivalent to models in the statistics literature. Many years after their original publication, prototype models, exemplar models, and decision bound models

were all shown by Ashby and Alfonso-Reese (1995) to be mathematically identical to different methods for performing *density estimation* in statistics. Likewise, several years after the introduction of the Rational Model of Categorization (RMC; Anderson, 1990, 1991), Neal (1998) pointed out that the RMC was equivalent to the celebrated Dirichlet process mixture model from statistics.

Due to the complementary nature of psychology and statistics, new cognitive models can be developed by drawing directly on methods from the statistics literature. In particular, I argue that one statistical framework of models in particular, the hierarchical Dirichlet process (HDP; Teh et al., 2006), is especially promising for modeling human categorization. Not only does it subsume several previous psychological categorization models, including prototype models, exemplar models, and the RMC, but it also contains new types of models which exhibit some interesting properties. One key prediction of these new models is that in categorization problems involving multiple categories, information about each category should be able to influence what is learned about all the other categories. This idea that information should be shared between categories seems quite sensible; however, it is absent from nearly all previous categorization models in the psychology literature.

As mentioned above, the connections between cognitive and computational modeling can work in the other direction as well, with the study of human capabilities leading to inspiration for new statistical models. Although categorization has long been researched in psychology, most psychological models are still limited to basic problems, with a small number of relatively simple categories. In the real world, categories are extremely numerous and highly interrelated.¹ Thus, in order to reach an adequate understanding of human categorization, it is necessary to incorporate the complexity of the tasks that people encounter and solve in their daily lives into one's experiments and models. For example, categories are not simply isolated collections of objects; rather, they can be situated in conceptual networks of other categories which are often hierarchically structured. Learners may not always be given direct information about the hierarchical relationships between categories in these taxonomy structures, and they may be forced to infer them from indirect evidence. Hierarchies of categories have been studied in different capacities in psychology (Collins and Quillian, 1969; Rosch et al., 1976), but the particular problem of automatically inferring taxonomical relationships has not been addressed to date. In this work, I introduce an extension of the HDP to account for this aspect of human categorization. The resulting model, herein referred to as the *tree-HDP*, technically subsumes the HDP itself, but, more importantly, it provides an account of people's ability to learn multiple related categories and infer the taxonomy structures in which they are situated.

The statistical categorization models adopted in this dissertation are motivated not only by reflecting upon the complex categorization problems that people solve, but also by

¹In fact, every noun in the English language can be considered to be a category, containing all the entities to which that noun can refer. Granted, under this interpretation, many categories, such as *Barack Obama*, would be very sparsely populated.

a careful study of the nature of human categorization in simpler settings. Human categorization has traditionally been studied by performing experiments in which participants are trained on a number of category examples and asked to generalize to new stimuli. Computational models are then trained to perform the same task, and the models' inferences are compared to those of the human learners to assess how well the models account for human learning. Despite the abundance of formal models of categorization that have been proposed in the psychological literature, there have been relatively few conclusive empirical evaluations of these competing accounts (Wills and Pothos, 2011). Rather than comparing the predictions of different models to the results from human learners, one can seek to characterize human categorization by understanding people's *inductive biases*, that is, the *a priori* preferences that people have for some category structures over others. For example, if it is found that people expect and prefer categories which are defined by simple linear boundaries that lie along one feature dimension, then one may rule out any categorization model which does not share this preference. Recently, an experimental technique known as *iterated learning* has been shown to be able to directly reveal people's inductive biases without the need for any independent models (Griffiths and Kalish, 2007). In this dissertation, I describe an iterated learning experiment of categorization to characterize people's inductive biases. The results of this experiment motivate the idea that human categorization is not as simple as would be implied by many previous psychological models, justifying the additional complexity and flexibility inherent in statistical models such as the HDP and the tree-HDP.

The remainder of this dissertation is organized as follows. In Chapter 2, I present a characterization of human categorization through an investigation of people's inductive biases, as revealed by the experimental method of iterated learning. Chapter 3 describes how previous psychological models of categorization can be unified into a common framework based on the statistical model known as the hierarchical Dirichlet process (HDP). The HDP contains families of models that have been previously unexplored by psychologists and that make novel predictions about how information can be shared between multiple categories; I test these predictions through an experiment with human learners. In Chapter 4, I introduce and formally define the problem of multi-level category learning and describe tree-HDP, a model designed to solve the multi-level category learning problem. The tree-HDP can be used to learn categories at multiple levels of a hierarchical structure and to reconstruct the category hierarchy. I compare its performance on the taxonomy reconstruction task to that of human learners through a number of experiments. Finally, in Chapter 5, I summarize the findings and contributions of this work and discuss its limitations and potential future directions.

Chapter 2

Characterizing human categorization

The ability to learn new categories from examples is a basic component of human cognition, and one of the earliest to be studied by psychologists (Hull, 1920). This long history of investigation has resulted in a number of computational models of category learning, including approaches based on inferring decision rules (Ashby and Gott, 1988), extracting prototypes (Reed, 1972), memorizing exemplars (Medin and Schaffer, 1978; Nosofsky, 1986), and combinations of these methods (Nosofsky et al., 1994; Vanpaemel and Storms, 2008). This proliferation of models has been complemented by an empirical literature comparing the ability of different models to account for human behavior. In a typical experiment, participants are taught the category membership of a set of training stimuli and then asked to generalize to a set of test stimuli. Computational models are evaluated on their ability to predict the resulting patterns of generalization.

Competing models of category learning are commonly presented in terms of their different assumptions about people’s mental representations of categories and the processes that translate these representations into behavior. These accounts of categorization typically describe human cognition at the algorithmic and representational level (Marr, 1982). However, these models can also be thought about more abstractly at the computational level (Marr, 1982): as methods of learning categories that have different *inductive biases*. In machine learning, the inductive bias of a learner is defined to be those factors other than the observed data that lead the learner to favor one hypothesis over another (Mitchell, 1997). Different models of category learning posit that people favor different kinds of hypotheses about the structure of categories. For example, a prototype model favors hypotheses in which categories are coherent groups of stimuli, while an exemplar model is more flexible, and can represent categories that consist of multiple clusters of stimuli spread out across a stimulus space (Nosofsky, 1998). Therefore, an alternative way of viewing categorization models is in terms of their different assumptions about the inductive biases of human learners. From this perspective, to understand human categorization is to characterize people’s inductive biases, so it would be preferable to directly observe these inductive biases rather than attempting to find models which have similar inductive biases to those of human learners.

This chapter describes a novel approach to understanding human category learning. Rather than studying the generalizations people make with different training stimuli, we use an experimental method designed to provide direct access to learners' inductive biases.¹ In this experimental method, *iterated learning*, each participant is trained with stimuli that are selected from the responses of the previous participant. This results in a sequence of category structures each produced by learning from the previous structure. Mathematical analysis of this process shows that as the sequence gets longer, the structures that emerge will be consistent with the inductive biases of the learners (Griffiths and Kalish, 2007). Intuitively, iterated learning magnifies the small effects that inductive biases have on people's generalizations, until those biases are all that is reflected in the data. We use iterated learning to expose the inductive biases of human learners. In our work, we demonstrate that iterated learning complements traditional categorization experiments, and we provide a new dataset to the community against which computational models can be compared.

This chapter is organized as follows. In the next section, we summarize the iterated learning method. We then describe a multidimensional scaling experiment we conducted in order to find a spatial representation of our stimulus sets that accords with people's psychological representations. Next, we describe the iterated learning experiment that we conducted for categorizing a number of different stimuli. We then explore multiple ways of analyzing the data from this experiment to better understand and characterize people's inductive biases. We conclude by considering the implications of these results for understanding how people learn categories.

2.1 Iterated learning

Categorization research typically seeks to compare different models and proceeds by presenting each model (as well as human participants) with a set of training data and comparing the generalization predictions made by the people to those made by the models. While this method allows us to quantitatively measure the degree to which each model explains the human data, it does not directly expose the underlying inductive biases of the learners. Iterated learning is an experimental method designed to give a pure estimate of inductive biases (Griffiths and Kalish, 2007). Using this methodology, we shift the focus away from testing how well individual categorization models fit human data and towards directly studying the nature of human categorization in a way that is independent of any proposed models.

The central concept of the iterated learning framework is that the training data given to a learner (either a human participant or a learning model) are not directly specified by the experimental design; rather, they are sampled from a previous learner's generalization responses. The learners are arranged into a chain, where the responses from the first learner are used as training data for the second learner, and so on. Because each learner's responses

¹This chapter is based on work done in cooperation with Thomas L. Griffiths, Wolf Vanpaemel, and Michael L. Kalish, so I use plural first-person pronouns throughout this chapter.

depend only on the previous learner's, the chain is formally a Markov process, and therefore the responses will converge to a stationary distribution.

Griffiths and Kalish (2007) provided an analysis of iterated learning under the assumption that learners use Bayesian inference, sampling hypotheses from the posterior distribution given by Bayes' rule: $P(h|d) \propto P(h)P(d|h)$. In this case, the observed responses in the iterated learning chain will converge to the prior distribution $P(h)$. Treating learners as agents of Bayesian inference, their inductive biases are encapsulated by the choice of prior distribution; therefore, iterated learning allows us to directly expose the inductive biases of the learners in the form of the prior over hypotheses.

In recent years, the iterated learning experimental methodology has been successfully used to study many phenomena in human cognition, including causal induction (Yeung and Griffiths, 2011), language evolution (Reali and Griffiths, 2009), clustering of color terms (Xu et al., 2010), function learning (Kalish et al., 2007), and predictions about everyday events (Lewandowsky et al., 2009). In this chapter, we demonstrate that iterated learning provides a new way to study human categorization, where the main focus is to reveal the underlying biases of the learners rather than to evaluate the degree of fit of particular models.

2.2 Multidimensional scaling study of stimuli

The purpose of performing an iterated learning experiment of human categorization is to understand people's inductive biases in terms of which category structures they prefer *a priori*. However, characterizing the properties of a category structure over a set of stimuli must be done in the context of some representation scheme for the individual stimuli. We adopt the common strategy of assuming that each stimulus is represented as a point in some multidimensional space, where the similarity of two stimuli decreases as the distance increases between the two points (Shepard, 1987). In order to determine the spatial representation of a set of stimuli that best conforms to the mental representations used by human learners, we performed a multidimensional scaling experiment.

2.2.1 Method

2.2.1.1 Participants

The experimental participants included 41 workers from Amazon Mechanical Turk, who each received a payment of \$1.50, and 50 students at the University of California, Berkeley, who received course credit, for a total of 91. The experiment had four conditions, one for each of four stimulus sets. The experiment was completed by between 21 and 25 participants for each stimulus set.

2.2.1.2 Stimuli

The experiment involved four different sets of stimuli, each of which varied on two dimensions. Two of the stimulus sets had separable dimensions, meaning the dimensions on which they varied are easily differentiated. These were rectangles that varied in their width and height, and “Shepard circles” (Shepard, 1964): circles of a varying diameter with a radius drawn at a varying angle. The other two stimulus sets had integral dimensions, meaning their dimensions are not readily apparent, leaving no preferred coordinate systems for these stimuli in psychological space. These were both sets of amorphous blobs, one from Cortese and Dyre (1996), which we call “Cortese blobs”, and the other from Shepard and Cermak (1973), which we call “Shepard blobs”. The construction of these stimulus sets involves varying the amplitudes and phase-shifts of components of periodic, trigonometric functions, which are then converted to closed loops (see Cortese and Dyre (1996) and Shepard and Cermak (1973) for the exact formulas used to construct the stimuli). One stimulus from each set is shown in the first row of images in Figure 2.2.

For each stimulus set, we constructed an equal-spaced, 8-by-8 square grid of stimuli and used these 64 stimuli in the experiment. The grid of stimuli used for each stimulus set is depicted in the second row of images in Figure 2.2. We refer to the position of a stimulus in this 8-by-8 grid configuration as the *canonical coordinates* of that stimulus, to be contrasted with the *MDS coordinates* that we seek to construct through the multidimensional scaling experiment.

2.2.1.3 Procedure

Multidimensional scaling algorithms require as input a matrix of pairwise distance or similarity measurements (Borg and Goren, 2005). Experiments typically proceed by collecting similarity ratings for every pair of stimuli, and each pair is usually replicated more than once to smooth out any noise in the collection process. However, this process can be extremely laborious for large numbers of stimuli, as the number of pairs of stimuli that need to be compared grows in proportion to the square of the number of stimuli. A number of alternative schemes have been proposed to collect pairwise similarity data more efficiently. In this experiment, we follow the procedure of Goldstone (1994), where rather than rating the similarity of pairs of stimuli, participants manually arrange a number of stimuli into a spatial configuration that is intended to reflect their similarities to each other. This procedure can be much more efficient than iteratively collecting pairwise similarity values, since a spatial arrangement of n stimuli conveys $n(n - 1)/2 \approx n^2/2$ pairwise distance measurements, but essentially requires only n decisions to be made about where to place the stimuli.

The experiment was conducted entirely over the web. It consisted of one practice session, intended to orient participants with the experimental procedure, followed by six regular sessions. In each session, the screen was populated with 32 square cards, each displaying one stimulus. The participants could move the cards around the screen by dragging

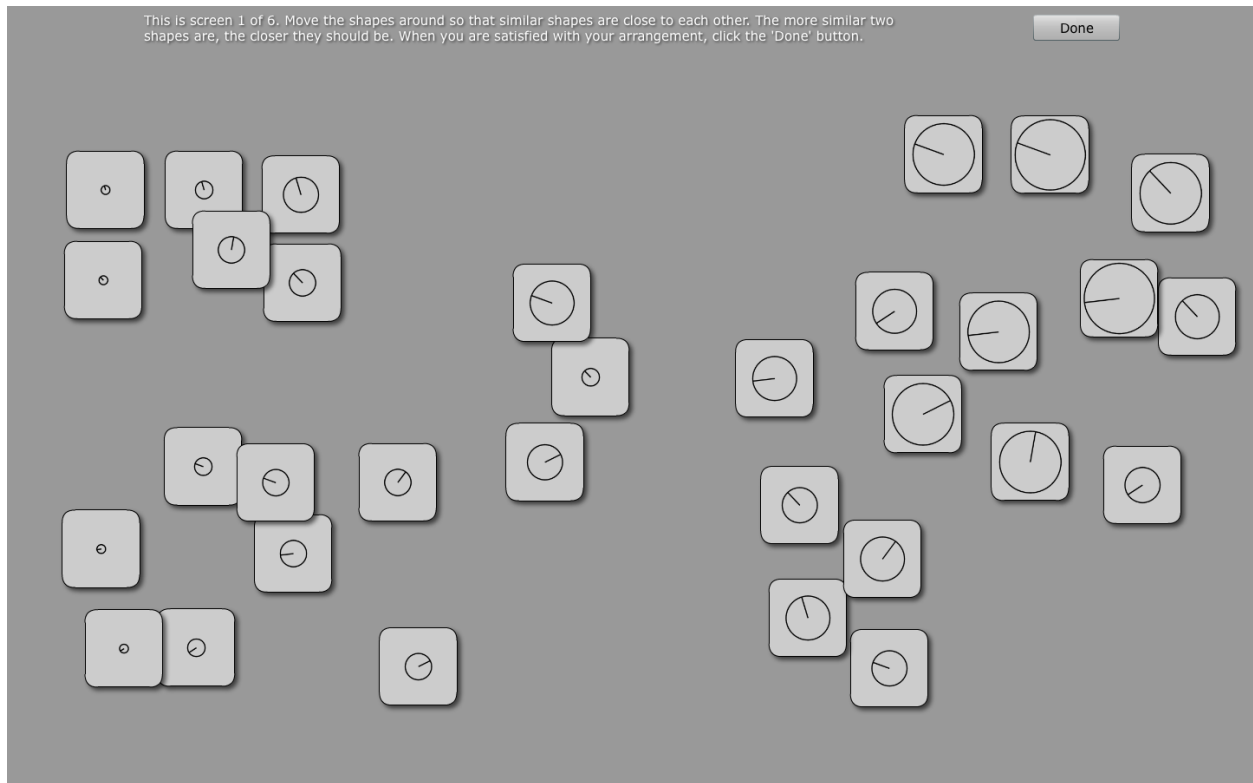


Figure 2.1: Screenshot of one session of the multidimensional scaling experiment.

and dropping them into a new location. Participants could move the cards individually, or they could select a group of cards and move them all simultaneously. The participants were instructed, “Move the shapes around so that similar shapes are close to each other. The more similar two shapes are, the closer they should be.” A screenshot of one of the sessions is shown in Figure 2.1.

We used the following procedure to decide which stimuli were shown in each of the sessions. The 64 stimuli were first randomly split into four groups of 16 stimuli each. Each session contained the 32 stimuli from two of these groups. The practice session always contained the stimuli from groups 1 and 2. There are $4 \times 3 \div 2 = 6$ possible pairs of the four groups, and each of these pairs was used to construct one of the six regular sessions. As a result, each stimulus appeared exactly three times with each of the other stimuli in its own group and exactly once with each of the stimuli in another group. This ensured that every participant ended up providing distance information about every pair of the 64 stimuli at least once over the course of the six regular sessions.

2.2.2 Results

For each session of each participant, we recorded the Euclidean distance in pixels between each pair of the 32 stimuli in that session. We assumed that within a single session, the distance between one pair of stimuli was directly comparable to the distance between another pair of stimuli. However, because participants could presumably have used different criteria for arranging the stimuli in different sessions, and because different participants could have been using screens of different sizes, distance measurements from two different sessions or two different participants are not necessarily directly comparable. To correct for these discrepancies, we normalized the pairwise distance measurements in each session by the average distance between all pairs in that session. Next, we aggregated the normalized pairwise distance measurements across all participants who used the same stimulus set, leading to a single overall pairwise distance matrix for each of the four stimulus sets. These distance matrices were used as input to a non-metric multidimensional scaling algorithm to produce an arrangement of the stimuli in a new set of two-dimensional *MDS coordinates*.² The third and fourth rows of images in Figure 2.2 show, respectively, the mapping between the canonical coordinates and the MDS coordinates for each stimulus set.

For the Shepard circles, the MDS coordinates very closely approximate the canonical coordinates, indicating that most people’s preferred mental representation for these stimuli is in accordance with the canonical representation scheme using the two orthogonal dimensions of diameter and angle. For the rectangles, we were surprised to find that the MDS coordinates formed more of a one-dimensional horseshoe-shaped curve, with the tall and narrow rectangles at one end, the short and wide rectangles at the other end, the tall and wide rectangles in the middle, and the short and narrow rectangles interspersed among the others. For the Cortese blobs, the MDS coordinates seem to be a circular version of the canonical coordinates, with only one main degree of variation. Finally, for the Shepard blobs, the stimuli along the top of the canonical coordinates (the points in blue and pink) seemed to form one elongated cluster, with the other stimuli (the points in green and orange) formed a second elongated cluster; within each cluster, however, there seems to be no discernible pattern.

2.3 Iterated learning experiment

Having used multidimensional scaling to construct a spatial representation of each stimulus set that corresponds to people’s mental representations, we now explore the inductive biases of human category learners using an iterated learning experiment.

²We used the `mdscale` algorithm from the Matlab Statistics Toolbox (version 2010a), which uses Kruskal’s normalized stress1 criterion.

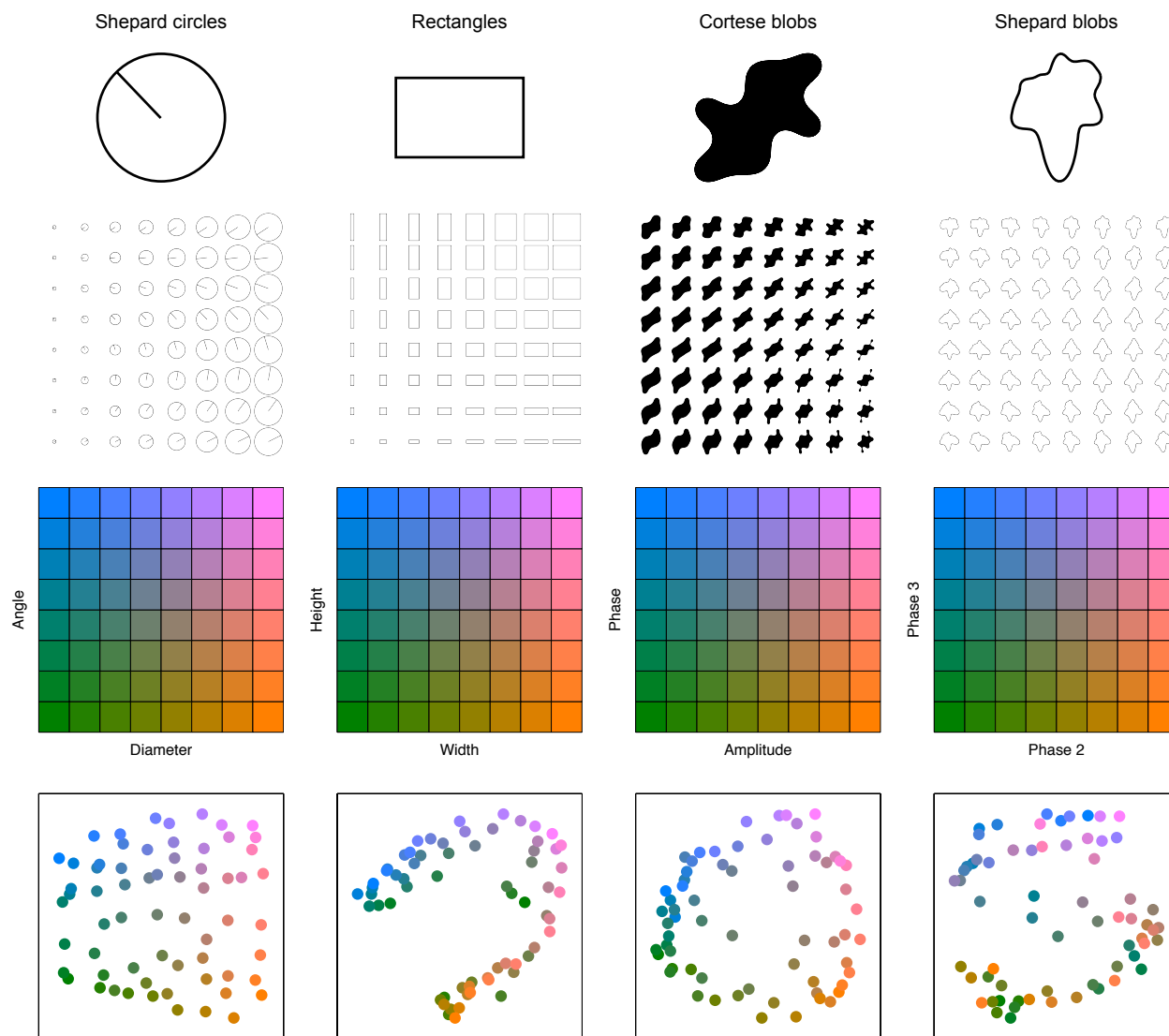


Figure 2.2: The four stimulus sets, presented in both canonical and MDS coordinates. Each of the four columns corresponds to one stimulus set, as specified by the titles along the top of the figure. The first row of images are representative examples of each stimulus set. The second row presents all 64 stimuli from each stimulus set in the 8-by-8 grid of the canonical coordinate space. The third row represents each stimulus as a differently-colored square, arranged in the same 8-by-8 grid of canonical coordinates. The final row represents the stimulus sets in the coordinates of their respective multidimensional scaling solutions; in these images, each marker represents one stimulus, where the color corresponds to the color used to represent the same stimulus in the third row of images.

2.3.1 Method

2.3.1.1 Participants

The experimental participants included 960 workers from Amazon Mechanical Turk, who each received a payment of \$0.50, and 480 students at the University of California, Berkeley, who received course credit, for a total of 1440. The experiment had 16 conditions, resulting from the combination of four stimulus sets and four initial category structures. Each condition was replicated with six chains, four of which were made up solely of Mechanical Turk workers, and two of which were made up solely of Berkeley students. Each replication of each condition consisted of an iterated learning chain of 15 generations. Each participant was randomly assigned to an incomplete chain in their pool (either Mechanical Turk workers or Berkeley students), occupying the next available generation in the chain.

2.3.1.2 Stimuli

The stimuli are described in Section 2.2 and pictured in Figure 2.2. We used the four stimulus sets of Shepard circles, rectangles, Cortese blobs, and Shepard blobs.

2.3.1.3 Procedure

Each participant completed a training session and a test session. In the training session, the participant was trained to reproduce the category memberships of a random selection of 32 of the 64 stimuli. In the test session, each participant classified all 64 items in random order without feedback. In each training trial, the participant classified a single stimulus from the training set with feedback. For first-generation learners, this feedback was based on one of four initial category structures, which are shown in the first and second columns of Figure 2.3. Two of the initializations—the first and third distinct ones—are simple linear boundaries compatible with a prototype model. The other two are discretized versions of category structures described by McKinley and Nosofsky (1995). For the remaining generations, feedback was provided according to the test session responses of the participant in the previous generation. Participants were not made aware that their test responses would be used in later generations and did not have any contact with other learners from different generations. The training session was organized into blocks containing 32 trials each, with the order of presentation of the stimuli randomized within each block.

If the participant correctly answered at least 22 of the 32 training trials in any training block, they continued to the test session.³ Otherwise, they completed another block of the training session. If after 20 blocks or 30 minutes, a participant had not yet reached the learning threshold, the experiment was ended, and the participant’s data were not included in further analyses. There were 38 participants who reached the maximum number of blocks and

³22 correct responses out of 32 trials indicates with $p < 0.05$ that the responses are not purely random, according to an exact binomial test.

25 who reached the time limit without achieving the learning criterion. These participants were replaced by others to fill in their positions in the chains. Additionally, there were 31 participants whose positions had to be replaced because some of their data were not properly sent to the database server due to network connection issues. There were 31 additional participants whose data were excluded because they were assigned to duplicate positions in a chain that had already been filled by another participant; this occurred when many participants attempted to take the experiment simultaneously and our system was unable to properly allocate them all to a unique chain. Finally, to prevent degenerate solutions where all the stimuli are assigned to one category, we replaced 445 participants who assigned more than twice as many stimuli to one category than the other category in the test session.

2.3.2 Results

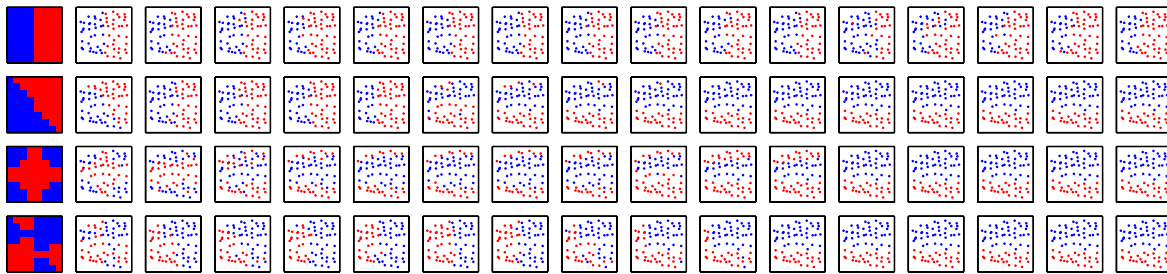
No significant differences were found between the two participant pools, so their data were combined in all analyses. Figure 2.3 shows one chain of 15 generations for each of the 16 conditions, with gray vs. black pixels indicating category membership.⁴ In each row, the first panel shows the initial category structure in canonical coordinates, the second panel shows the initial category structure in MDS coordinates, and all other panels show, in MDS coordinates, the category assignments made by a learner in the test session after being trained on the category structure to its left.

Most of the Shepard circle chains converged to fairly simple structures using categorization boundaries aligned with one of the dimensions. For the rectangles, people seem to prefer two main types of category structures: one that splits the horseshoe shape into two roughly equal sides (corresponding to rectangles that are wider vs. rectangles that are taller), and another with one category containing stimuli on the ends of the horseshoe and the other category containing stimuli in the middle of the horseshoe (corresponding to squares and square-like rectangles vs. eccentric rectangles). The Cortese blob chains seem to favor a number of different category boundaries which are angular segments of the circular structure, but never with a perfectly delineated boundary. The results for the Shepard blobs seem quite noisy. Perhaps people interpreted these stimuli in feature spaces which are rather different from the dimensions we used to plot the results, or perhaps because these stimuli are difficult to interpret, people’s inductive biases about them are very weak.

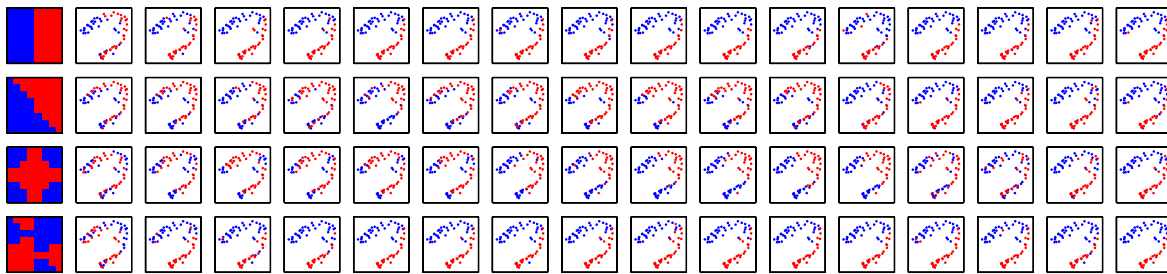
2.4 Characterizing human inductive biases

The data collected in the iterated learning experiment offer a unique opportunity to study human characterization in a variety of different ways than have traditionally been used. Not only does each iteration of each chain provide a dataset that is along the lines of a traditional

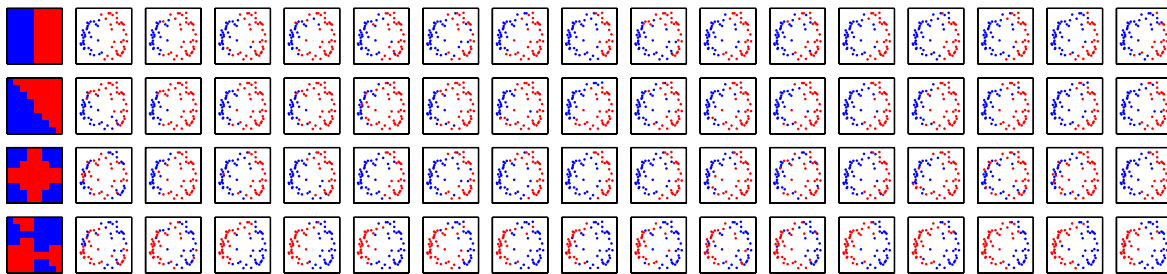
⁴To promote further exploration of the results by other researchers, the full set of results is available online at <http://cocosci.berkeley.edu/iteratedCatData/>.



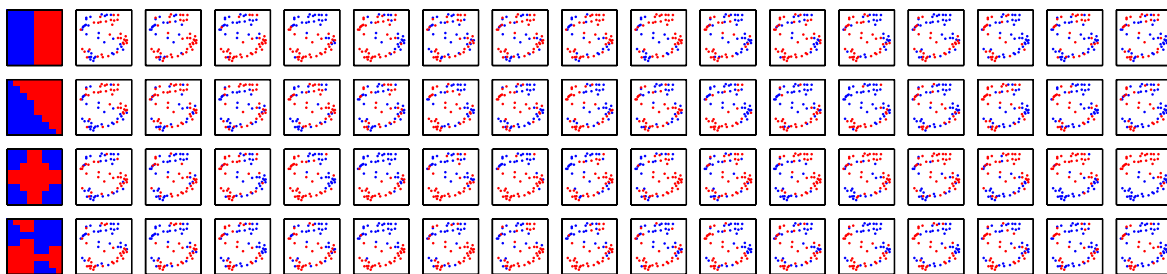
(a) Shepard circles



(b) Rectangles



(c) Cortese blobs



(d) Shepard blobs

Figure 2.3: Human data from the iterated learning experiment. Each row is one chain. One chain is shown for each combination of stimulus set and initial category structure. Colors indicate category membership, and each image shows the responses of a single learner. In each chain, the first and second images show the initial category structure in canonical and MDS coordinates, respectively. All subsequent images are presented in MDS coordinates.

categorization experiment, with generalization judgments made on the basis of a training set, but each chain provides a view into the dynamics of how category structures evolve over time when people are learning from other people’s judgments, and all the chains together offer a means by which we can characterize human inductive biases for each stimulus set. While this rich dataset can be analyzed in many different ways, in this section, we present two sample analyses.

The chains of category structures in Figure 2.3 produced by the human learners in the iterated learning experiment show how people’s judgments dynamically evolve over time. However, the purpose of iterated learning is to evoke people’s *static* inductive biases. The theoretical basis of iterated learning is that people come into learning problems with a set of prior expectations which shape their inferences, and iterated learning is designed to expose these biases. The results produced by human learners eventually converge to their prior distribution over category structures (Griffiths and Kalish, 2007), so by aggregating the category structures produced during the later stages of the chains, we should be able to develop a clear picture of people’s inductive biases.

2.4.1 Visualizing inductive biases

In order to characterize both the overall trend and the individual variation of the human learners’ inductive biases, we executed a multidimensional scaling analysis of the category structures they produced in the latter parts of the iterated learning chains.⁵ Taking a conservative estimate of the point at which the chains had converged to their stationary distributions, we used the results from iterations 6 through 15 of each chain. For each of the four stimulus sets, we first constructed a matrix of the distances between each pair of the category structures, using the variation of information (VI) metric (Meila, 2003) as a distance function. The VI metric is a measure of the distance between partitions, so it depends only on how stimuli are classified, and not the locations of those stimuli in the feature space. The VI metric is invariant to relabelings of the categories, so two structures which are identical but switch the category labels would have a VI distance of zero.

After constructing this pairwise distance matrix, we ran a metric (classical) multidimensional scaling algorithm, plotting the results in two dimensions.⁶ The results of this analysis for each stimulus set are shown in Figure 2.4. Overall, we found that each stimulus set displayed a unique pattern of individual variations in the category structures.

For the Shepard circles, there is a cluster of category structures in the bottom-right corner, exemplified by point D in Figure 2.4(a), which all exhibit a category boundary aligned with the vertical axis, corresponding to categories defined on to the diameter of the circle. There is a small number of results, exemplified by point B, where the category assignments seem to be less structured. Finally, there is a range of results extending from around point A

⁵This multidimensional scaling procedure is separate and independent of that described in Section 2.2.

⁶We used the `cmdscale` algorithm from the Matlab Statistics Toolbox (version 2010a).

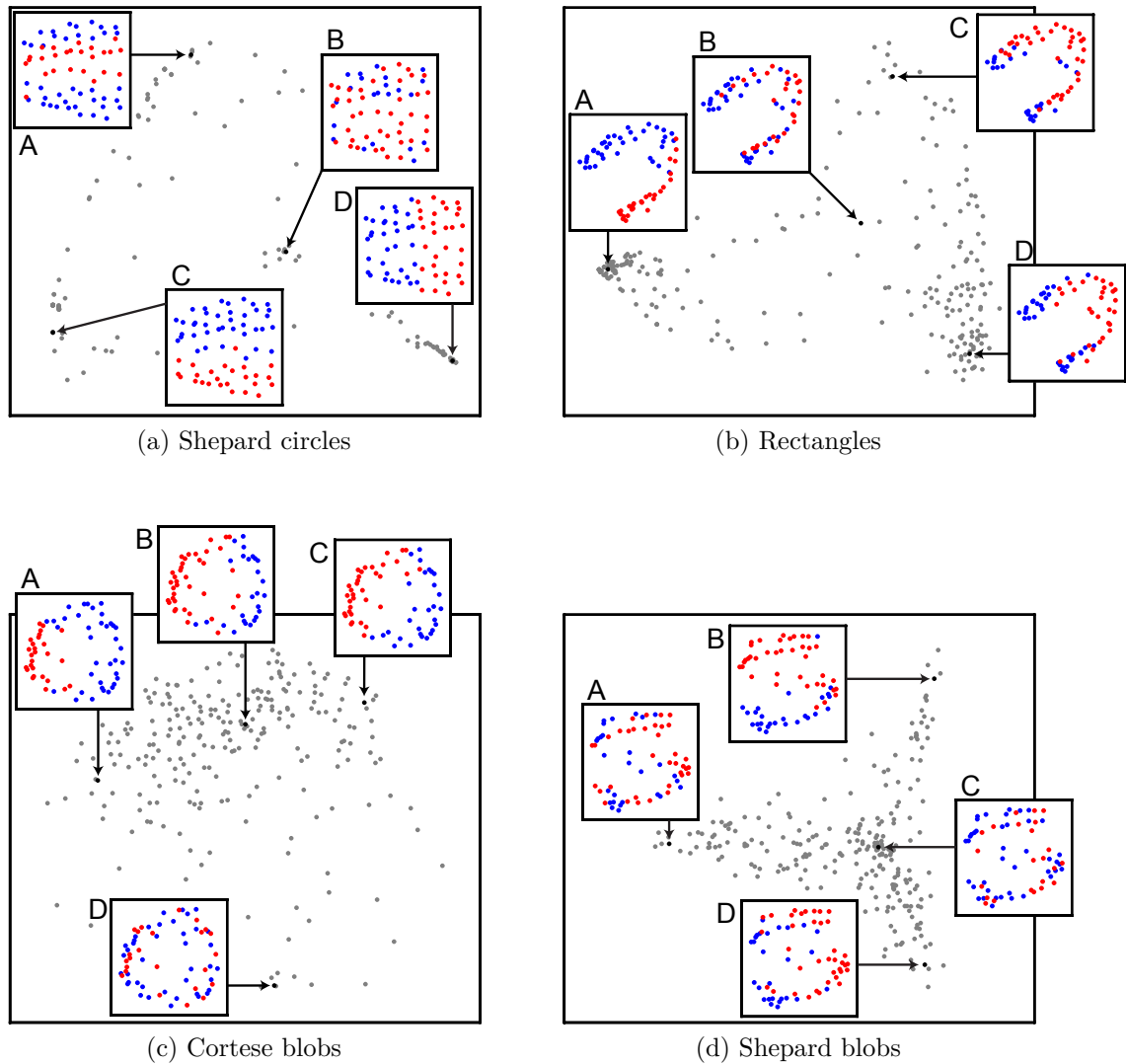


Figure 2.4: Results of running a multidimensional scaling algorithm on the category structures produced by the human learners. Each medium-gray dot represents the category structure produced by one participant. Only results from iterations 6 through 15 of each chain are included. For each stimulus set, four individual category structures (labeled “A” through “D”) are shown in detail to demonstrate the variation among people’s responses.

to around point C, of category structures whose boundaries are aligned with the horizontal axis, corresponding to categories defined on the angle of the radius.

For the rectangles, there seem to be two major clusters of category structures. The first group, exemplified by point A in Figure 2.4(b), break the horseshoe shape into two halves, corresponding to a category of rectangles that are taller than they are wide, and a category that are wider than they are tall. The second group, exemplified by point D, seem to correspond to a category of squares and square-like rectangles and a category of more eccentric rectangles. There seems to be a gradation of responses between these two clusters, as well as gradations of responses between each of these clusters and the region around point C, where the category structure is somewhat similar to that of point D. As with the Shepard circles, there seems to be a region of a few responses towards the middle where the category structures apparently display less structure.

In Figure 2.4(c), the results appear less clustered and more evenly distributed around a general region. Points A, B, and C all demonstrate category boundaries which are defined as angular sections of the circular stimulus space, with each choosing different boundaries. As with the previous stimulus sets, there seems to be a region containing a few responses, such as point D, where the categories are less structured.

Finally, in Figure 2.4(c), the results are roughly evenly distributed around the interior of a three-pointed star-shaped region. The results near point B demonstrate category boundaries which differentiate the two elongated clusters of the MDS coordinates, while the other areas (around points A, C, and D) all consist of seemingly unstructured categories.

While each stimulus set displays a unique pattern of overall category structures and individual variation, some generalizations can be made about the separable stimulus sets and the integral stimulus sets. The category structures that the participants produced for the separable stimuli, including the Shepard circles and the rectangles, are more tightly clustered together into discrete groups. By contrast, the structures produced for the integral stimuli, including the Cortese blobs and the Shepard blobs, are more evenly spread and less clustered together. This suggests that for the separable stimuli, the human learners are biased towards a relatively smaller number of discrete types of category structures, whereas for the integral stimuli, people entertain a richer variety of hypotheses. These results support the notion that people's inductive biases for categorization problems are not monolithic, but rather, can change between favoring a smaller set of simpler category structures, as demonstrated with the separable stimulus sets, and ranging over a larger set of more complex category structures, as demonstrated with the integral stimulus sets.

2.4.2 Aggregate category complexity

One of the most striking differences between competing models of human categorization is the amount of complexity they assume people use when representing category structures. For example, prototype models (Reed, 1972) use very simple category representations consisting of just a category center (the prototype) and possibly some measure of the variation in

each feature dimension. By contrast, exemplar models (Medin and Schaffer, 1978; Nosofsky, 1986), the most prominent competing class of models, assume that the complexity of category representations grow with the observed data, meaning that they are essentially unlimited in complexity. Given that iterated learning exposes the inductive biases of human learners, this presents us with an opportunity to measure human categorization along a dimension on which traditional categorization models are strongly divergent: the complexity of the category representations that people prefer to use.

The complexity of sets like category structures can be quantified in many different ways. Here, we take the approach of measuring complexity by counting the number of coherent clusters of stimuli there are in each of the two categories that the participants constructed. To achieve this, we use a Dirichlet process mixture model (DPMM; Ferguson, 1973), a model from nonparametric Bayesian statistics that finds a probabilistic clustering of a set of data. One of the key properties of the DPMM is that the number of clusters that it uses to model a dataset is not specified in advance by the modeler; instead, the DPMM infers how many clusters need to be used to adequately represent the data. In this way, it provides us a tool to measure the complexity of a category structure in terms of how many clusters it uses to model the stimuli in each of the two categories. Using a Gibbs sampling procedure, we ran a DPMM on each set of responses from the human learners, collecting a set of samples from the posterior distribution over the number of clusters. The results of this analysis are summarized in Figure 2.5.

A number of results are apparent from this analysis. First, we can see that although the chains within each stimulus set each start from different initial category structures of differing complexity, they all converge to a common level of complexity over the course of the iterated learning experiment. This provides empirical validation to the theory that the chains will converge to a common prior distribution regardless of their initial starting point. We can also see that there is significant variation in the level of complexity of the category structures used across different stimulus sets. For the Shepard circles, each category produced by the participants required about 1.6 clusters on average to be modeled by the DPMM. The DPMM used about 2.6 clusters on average to represent each category for the rectangle stimulus set. The categories for the Cortese blobs and the Shepard blobs required significantly more clusters on average: about 3.4 and 4.9, respectively. In addition, for the two separable stimulus sets (the Shepard circles and the rectangles), the category complexities start higher and decrease over time, while for the integral stimuli, the numbers of clusters stay consistently high over the entire chain. These results support the findings from the study of the individual variations in Section 2.4.1 that human learners prefer relatively simpler category representations when they encounter separable stimuli than when they are presented with integral stimuli.

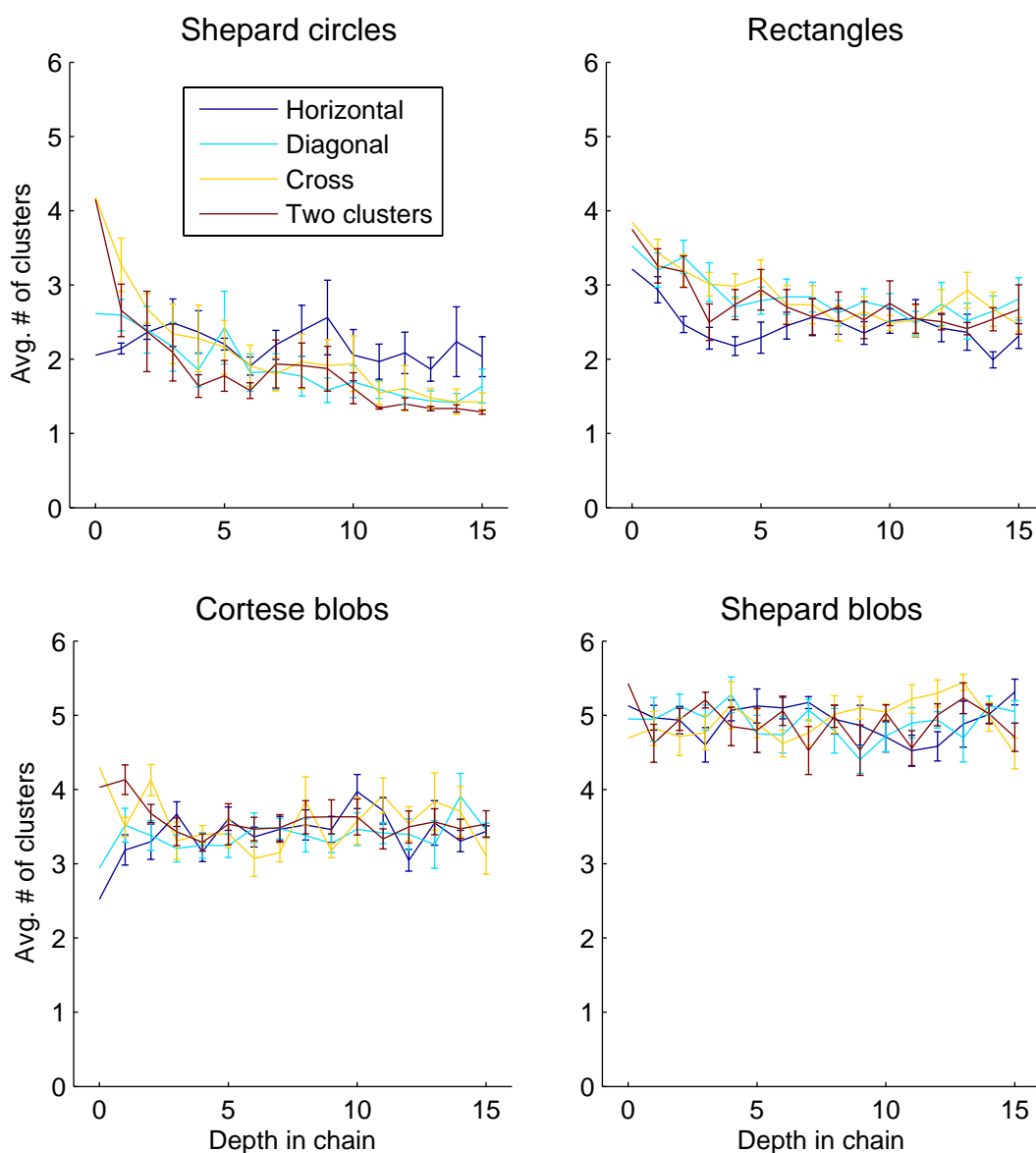


Figure 2.5: Results of analyzing the complexity of the category structures produced in the iterated learning chains. Each subplot shows the analysis for a different stimulus set. Within a subplot, each line represents all the chains that start from a common initial category structure and shows the evolution of the complexity of the category structures produced in those chains. The x axis is the depth (the number of learning iterations completed in the chain), starting from the initial category structure at a depth of 0 and going to the 15th learning iteration. The y axis is the average number of clusters used by the Dirichlet process mixture model to describe each category produced by the learners. Error bars extend ± 1 standard error, under the assumption of normality.

2.5 Conclusions

As a whole, our results suggest that the human learners' inductive biases are highly dependent on the stimuli they are categorizing. With separable stimuli, people seem to prefer a smaller number of simpler categories, while with integral stimuli, people are capable of learning a wider variety of more complex categories. This indicates that one may not be able to identify a single categorization model that can explain human learning for all stimuli. Indeed, these results support previous findings that human learning seems to be better described by prototype models during the early stages of learning and by exemplar models during later stages of learning (Smith and Minda, 1998). Therefore, it may be beneficial to investigate models of categorization which are capable of adopting flexible representations. In the next chapter, it is demonstrated that in traditional categorization studies, a model called the hierarchical Dirichlet process (HDP; Teh et al., 2006) is capable of explaining human data that is not adequately modeled by either prototypes or exemplars, which both assume a single, fixed level of representational complexity. Models like the HDP use representations that can vary in complexity depending on the stimuli. This supports the idea that models which use more flexible representations and can interpolate between the behavior of prototypes and exemplars provide a better explanation of the variable nature of human categorization.

Chapter 3

Statistical models

Categorization is a fundamental aspect of human cognition, and consequently, a fundamental area of research in cognitive psychology. Many common problems that people face in their daily lives can be described as categorization or rely on categorization as a subproblem. In essence, categorization is a way for people to understand and make sense of noisy, high-dimensional sensory information by reducing it to a small, discrete set of possibilities that can more easily be reasoned about. For example, a person can easily recognize an apple sitting in front of her, despite the variable nature of its color, size, distance, viewing angle, and lighting conditions. As a result of its central importance, categorization has received a large amount of attention from researchers in psychology and cognitive science, and a variety of different psychological models of categorization have been proposed, studied, and tested (e.g., Reed, 1972; Medin and Schaffer, 1978; Nosofsky, 1986; Love et al., 2004; Vanpaemel and Storms, 2008).

Recently, a class of psychological models has been developed based on the approach of rational analysis, in which human thought and behavior is explained as an optimal solution to a computational problem posed by the environment (Anderson, 1990, 1991; Chater and Oaksford, 1999; Marr, 1982; Oaksford and Chater, 1998). This type of analysis has been used in several rational models of category learning (e.g., Anderson, 1990; Nosofsky, 1998; Ashby and Alfonso-Reese, 1995). These analyses essentially agree on the nature of the computational problem involved, casting category learning as a problem of *density estimation*: determining the probability distributions over objects associated with different category labels. Viewing category learning in this way helps to clarify and contrast the assumptions behind different categorization models. For example, prototype models (e.g., Reed, 1972), which assume that a category is associated with a single prototype to which new stimuli are compared, are mathematically equivalent to the strategy of parametric density estimation from the statistics literature. Correspondingly, exemplar models (e.g., Medin and Schaffer, 1978; Nosofsky, 1986), which assume that a category is represented by a set of stored exemplars which are all compared to new stimuli, are equivalent to kernel density estimation in statistics. These equivalences between psychological and statistical methods were first made

explicit by Ashby and Alfonso-Reese (1995).

Rational models have also led to new ways of thinking about category representations, and offer the potential to unify different accounts of categorization. Anderson’s 1990; 1991 Rational Model of Categorization represents categories as sets of clustered observations. Prototypes and exemplars can be thought of as special cases of this kind of representation, where a prototype-based representation uses a single cluster for each category and an exemplar-based representation assigns each object to its own cluster. The model can thus behave similarly to a prototype model or an exemplar model, depending on its parameters (Nosofsky, 1991). Subsequent categorization models have built on the idea that categories are represented as sets of clusters (Vanpaemel and Storms, 2008; Rosseel, 2002; Love et al., 2004), with this approach becoming a well-established intermediate strategy between exemplar and prototype models.

In this chapter, we use a rational analysis of categorization as the basis for an exploration of what different models of categorization might have in common.¹ Previous rational analyses suggest that different models of categorization can all be construed as solutions to the same computational problem. We go a step further, identifying a unifying framework for rational models of categorization. This framework contains prototype and exemplar models, Anderson’s 1990; 1991 rational model, and some interesting new models which have previously been unexplored, all as special cases. We present an empirical test of one of the predictions that results from considering these new models, examining how people transfer knowledge between categories when learning multiple categories simultaneously.

Our unifying framework for rational models of categorization is based on recent advances in nonparametric Bayesian statistics. Anderson’s 1990; 1991 Rational Model of Categorization (henceforth, RMC) can be shown to be equivalent to a Dirichlet process mixture model (Antoniak, 1974; Ferguson, 1983), a seminal model in the field of nonparametric Bayesian statistics (Neal, 1998). This connection was used by Sanborn et al. (2010a) to demonstrate that the RMC can be applied with more sophisticated inference algorithms than the local greedy algorithm first proposed by Anderson (1991), allowing the RMC to make different predictions depending on the choice of inference algorithm. However, it also provides a link to the growing literature on nonparametric Bayesian models. The Dirichlet process was only one of the first models developed in the field of nonparametric Bayesian statistics, and many new ideas have since been developed. In particular, people have begun to design hierarchical models such as the hierarchical Dirichlet process (HDP; Teh et al., 2006), which provides a clear way to define separate but related probability distributions associated with different categories. The strong connections pointed out by Ashby and Alfonso-Reese (1995) and Neal (1998) between psychological and statistical models suggest that these more recent developments in nonparametric Bayesian statistics can lead to new improvements in modeling categorization.

¹This chapter is based on work done in cooperation with Thomas L. Griffiths and Mikhail M. Shashkov, so I use plural first-person pronouns throughout this chapter.

In this chapter we focus on a categorization model based on the HDP. This model represents each category as a set of clusters, as does the RMC, but it also allows those clusters to be shared between multiple categories with a particular probability. Other overarching models have previously been proposed to bring prototypes, exemplars, and the RMC into a common framework (Rosseel, 2002; Vanpaemel and Storms, 2008). These models successfully demonstrated that prototypes, exemplars, and cluster-based models like the RMC all lie along a single spectrum of strategies for representing categories; however, they do not address the question of how a learner would decide which one of these representations should be selected to solve a given categorization problem. The unifying model we propose, the HDP, provides a rational solution to this problem, allowing the representation to change based largely on the structure of the category observations.

The HDP model we introduce not only unifies previous models and provides a solution to the problem of which representation should be used for a given categorization task; it also allows us to identify previously unexplored models with interesting properties, extending the set of rational models of categorization. One of these models introduces a new technique for representing a way in which categories can interact: by sharing clusters. By allowing each cluster used within a category to also be used in the representations of other categories, this model is able to learn that two or more categories share similar objects. This gives it the ability to learn more quickly in these situations than other models, which would have to start from scratch each time the same cluster appears in a new category. Not only can this improve the model's learning rate, but it also allows the model to potentially explain effects in human learning where knowledge about one category affects the inferences that people make about other categories.

Modeling interactions between categories is an important step forward from previous models towards a richer understanding of human categorization in the real world. Standard categorization experiments reveal information about the mental constructs – the underlying psychological spaces, the similarity functions, and the complexity of category representations – that people may be using to represent and reason about one or two categories in isolation. However, common sense suggests that when people learn new categories outside the controlled setting of laboratory experiments, they draw on their knowledge of previously-learned categories and take advantage of the relationships between them. For example, the category *mammal* is typically learned by way of its relationship to other categories such as *dog*, *monkey*, and *dolphin*, rather than by observing examples of individual mammals. Through its cluster-sharing mechanism, the HDP allows us to test a common assumption made implicitly by most categorization models: that categories are learned independently and in isolation from each other.

The remainder of this chapter is organized as follows. We first review the two most common classes of categorization models, prototype and exemplar models, and describe a more recent group of models which treat these two models as extremes of a spectrum and interpolate between them. We then describe how all of these models can be viewed from the perspective of rational analysis as solutions to the problem of density estimation. From

this perspective, we can identify a unifying model based on the hierarchical Dirichlet process that subsumes prototypes, exemplars, and clustering models. We show how the HDP also contains a new class of models which are able to learn a certain type of relationship between multiple categories and use this information to make predictions about new categories. We then describe a human learning experiment we performed to verify that people make the same qualitative inferences and predictions about related categories, showing that the HDP provides an explanation for this interesting effect in human learning. Finally, we conclude with a discussion of our findings and contributions.

3.1 Prototypes, exemplars, and clusters

While early psychological models of categorization focused on creating systems of deterministic rules to describe what objects belong to a category (Bruner et al., 1956), the shift towards similarity-based characterizations of categories in the past few decades led to the development of several models that belong to two general classes: prototype models and exemplar models. These models are based on calculations of the resemblance between a novel stimulus and each of the previously-observed categories. These models can be formally described as follows. Given a set of $N - 1$ stimuli with features $\mathbf{x}_{N-1} = (x_1, x_2, \dots, x_{N-1})$ and category labels $\mathbf{y}_{N-1} = (y_1, y_2, \dots, y_{N-1})$, prototype and exemplar models calculate the probability that a novel stimulus x_N has category label $y_N = j$ as

$$P(y_N = j | x_N, \mathbf{x}_{N-1}, \mathbf{y}_{N-1}) = \frac{\eta_{N,j} \beta_j}{\sum_{j'} \eta_{N,j'} \beta_{j'}}, \quad (3.1)$$

where $\eta_{N,j}$ is the similarity of stimulus x_N to category j , β_j is the response bias for category j , and the summation in the denominator is over all the categories. The response biases β_j are typically set either to 1 or proportional to the number of previously-observed objects in each category. Thus, the decision is a function of how similar the novel stimulus is to each of the categories, and the models use a straightforward application of the standard choice rule (Luce, 1959). Prototype and exemplar models thus differ primarily on their choice of the similarity function $\eta_{N,j}$.

3.1.1 Prototypes and exemplars

Prototype models (e.g., Reed, 1972) represent a category j by a single idealized instance that captures the category's central tendency and is called the category's *prototype*, denoted p_j . The prototype is commonly defined as the centroid of all instances of the category in some psychological space, i.e.,

$$p_j = \frac{1}{N_j} \sum_{\{i:y_i=j\}} x_i, \quad (3.2)$$

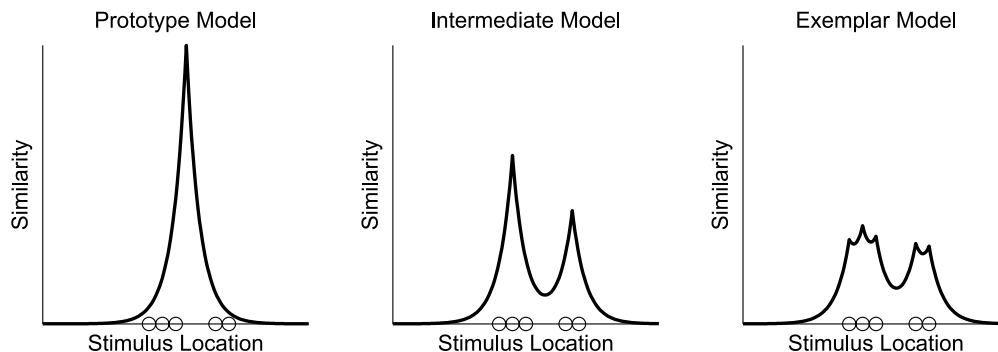


Figure 3.1: Category similarity functions for a simple one-dimensional category. The panel on the left shows the similarity function for a prototype model, with a single prototype summarizing the structure of the category. The panel on the right shows the similarity function for an exemplar model, with the overall similarity resulting from summing a set of similarity functions centered on each exemplar. The similarity function shown in the middle panel comes from an intermediate model that groups the three stimuli on the left and the two stimuli on the right.

where N_j is the number of observed instances of category j (i.e., the number of stimuli for which $y_i = j$). The similarity between a novel stimulus x_N and category j (denoted $\eta_{N,j}$ in Equation 3.1) is then defined as

$$\eta_{N,j} = s(x_N, p_j), \quad (3.3)$$

where $s(x_N, p_j)$ is the similarity between x_N and p_j , which can be defined in a number of different ways. One common choice for the similarity function $s(x, x')$ between two stimuli x and x' is a decaying exponential function of either the Euclidean or city-block distance between the feature vectors representing x and x' , following Shepard (1987). The distance calculation is often augmented with dimension-specific weights, which are either learned from the variability of each dimension in the observed data or fit to a set of observed human responses in a categorization task. An example of the overall similarity function produced by a prototype model is shown in the left panel of Figure 3.1.

In exemplar models (e.g., Medin and Schaffer, 1978; Nosofsky, 1986), all of the instances of each category are stored. The similarity of a novel stimulus x_N to category j is calculated by summing the similarity of the stimulus to all the stored instances, which are called the *exemplars* of the category:

$$\eta_{N,j} = \sum_{\{i:y_i=j\}} s(x_N, x_i), \quad (3.4)$$

where $s(x_N, x_i)$ is the similarity between the two stimuli x_N and x_i . An example of the similarity function is illustrated in the right panel of Figure 3.1.

3.1.2 Cluster-based models

Although prototypes and exemplars have dominated the modern literature, a number of authors (e.g., Love et al., 2004; Vanpaemel and Storms, 2008) have proposed more general classes of category representations that interpolate between prototype and exemplar models. These models are based on the idea of partitioning the observed instances of each category j into some number K_j of clusters (where K_j can range from 1 to N_j) and then representing each cluster by a *subprototype*, a prototype for just the members of that cluster. Under this type of model, the similarity of a novel stimulus x_N to a category j is defined by summing the similarity of the stimulus to each of the subprototypes:

$$\eta_{N,j} = \sum_{k=1}^{K_j} \pi_{j,k} s(x_N, p_{j,k}), \quad (3.5)$$

where $\pi_{j,k}$ is a weight attributed to cluster k in category j , and $p_{j,k}$ is the subprototype of cluster k in category j . In the Varying Abstraction Model (Vanpaemel and Storms, 2008), π_k is set to 1, but another reasonable choice is to set π_k to be proportional to the number of objects assigned to cluster k . The middle panel of Figure 3.1 illustrates a category similarity function for this class of models. If the number of clusters K_j is fixed at 1 for all categories, this is equivalent to a prototype model, and when $K_j = N_j$ for all j , this is equivalent to an exemplar model. Thus, these clustering models are more flexible than either exemplar or prototype models. However, they raise the question of how the stimuli should be partitioned into clusters for any particular categorization task.

3.2 Rational models of categorization

The models discussed in the previous section all explain categorization behavior in terms of cognitive processes: in particular, similarity and choice. An alternative approach is to seek an explanation based on the form of the computational problem that underlies categorization. Following the methodology outlined by Anderson (1990), rational models of categorization explain human behavior as an adaptive solution to a computational problem posed by the environment, rather than focusing on the cognitive processes involved. Existing analyses tend to agree that the basic problem is one of prediction: identifying the category label or some other unobserved property of an object using its observed properties (Anderson, 1990; Ashby and Alfonso-Reese, 1995; Rosseel, 2002). This prediction problem has a natural interpretation as a form of Bayesian inference. In a standard classification task, for instance, Bayes' rule allows us to compute the probability that object x_N belongs to category j given the features and category labels of the $N - 1$ previous objects:

$$P(y_N = j | x_N, \mathbf{x}_{N-1}, \mathbf{y}_{N-1}) = \frac{P(x_N | y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1}) P(y_N = j | \mathbf{y}_{N-1})}{\sum_{j'} P(x_N | y_N = j', \mathbf{x}_{N-1}, \mathbf{y}_{N-1}) P(y_N = j' | \mathbf{y}_{N-1})}. \quad (3.6)$$

In this expression, the posterior probability of category j is proportional to the product of the likelihood function $P(x_N|y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})$, which is the probability of sampling an object with features x_N from category j , and the prior probability $P(y_N = j|\mathbf{y}_{N-1})$ of choosing category j . Category learning, then, becomes a matter of determining these probability functions, a problem known in the statistical literature as *density estimation*.

3.2.1 The rational basis for prototype and exemplar models

Ashby and Alfonso-Reese (1995) observed that both prototype and exemplar models can be recast as rational solutions to the problem of categorization, highlighting the connection between the Bayesian solution presented in Equation 3.6 and the choice probabilities in the prototype and exemplar models, i.e., Equation 3.1. Specifically, the category similarity $\eta_{N,j}$ can be identified with the probability of the item under the category: $P(x_N|y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})$, while the category bias β_j corresponds naturally to the prior probability of category j , $P(y_N = j|\mathbf{y}_{N-1})$. The difference between prototype and exemplar models is thus the different ways of estimating the likelihood function $P(x_N|y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})$. The definition of $\eta_{N,j}$ used in an exemplar model (Equation 3.4) corresponds to estimating the likelihood function as a sum of a set of functions, known as *kernels*, centered on the stimuli x_i already labeled as belonging to category j :

$$P(x_N|y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1}) = \frac{1}{N_j} \sum_{\{i:y_i=j\}} f(x_N, x_i), \quad (3.7)$$

where $f(x, x')$ is a probability distribution centered on x_i . This method is widely used for approximating probability distributions in statistics, being a simple form of nonparametric density estimation called *kernel density estimation* (e.g., Silverman, 1986).

In contrast, the definition of $\eta_{N,j}$ used in a prototype model (Equation 3.3) corresponds to estimating the likelihood function by assuming that each category distribution comes from an underlying parametric family and then, for each category, finding the parameters that provide the best fit to the observed data from that category. The prototype is specified by these best-fitting parameters, with the centroid being an appropriate estimate for distributions whose parameters characterize their mean. Again, this is a common method for estimating a probability distribution, known as *parametric density estimation*, in which the distribution is assumed to be of a known form but with unknown parameters (e.g., Rice, 1995).

3.2.2 The Mixture Model of Categorization

Casting exemplar and prototype models as different schemes for density estimation suggests that a similar interpretation might be found for the interpolating, cluster-based models. Rosseel (2002) proposed one such model – the Mixture Model of Categorization (MMC) –

assuming that $P(x_N|y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})$ is a mixture distribution. Specifically, each object x_i comes from a cluster z_i , and each cluster is associated with a probability distribution over the features of the objects generated from that cluster. When evaluating the probability of a new object x_N , it is necessary to sum over all of the clusters from which that object might have been drawn. Accordingly,

$$P(x_N|y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1}) = \sum_{k=1}^{K_j} P(x_N|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1})P(z_N = k|\mathbf{z}_{N-1}, y_N = j, \mathbf{y}_{N-1}), \quad (3.8)$$

where K_j is the total number of clusters for category j , $P(x_N|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1})$ is the likelihood of x_N under cluster k , and $P(z_N = k|\mathbf{z}_{N-1}, y_N = j, \mathbf{y}_{N-1})$ is the probability of generating a new object from cluster k in category j . The clusters can either be shared between categories, or be specific to a single category, in which case $P(z_N = k|\mathbf{z}_{N-1}, y_N = j, \mathbf{y}_{N-1})$ would be equal to 0 for all clusters not belonging to category j . This model reduces to kernel density estimation (i.e., an exemplar model) when each object has its own cluster and the clusters are equally weighted, and it reduces to parametric density estimation (i.e., a prototype model) when each category is represented by a single cluster. By a similar argument to that used for the exemplar model above, we can connect Equation 3.8 with the definition of $\eta_{N,j}$ in the clustering models (Equation 3.5), providing a rational justification for this method of interpolating between prototypes and exemplars by partitioning objects into clusters.

3.2.3 The Rational Model of Categorization

The MMC elegantly defines a rational model that interpolates between prototypes and exemplars, but does not determine how many clusters are appropriate for representing each category, or how to assign stimuli to those clusters, based on the available data. Anderson (1990) introduced the Rational Model of Categorization (RMC), which presents a partial solution to this problem. The RMC differs from the other models discussed in this section by treating category labels like features. Thus, the RMC specifies a joint distribution on features and category labels, rather than assuming that the distribution on category labels is estimated separately and then combined with a distribution on features for each category. As in the MMC, this distribution is a mixture, with

$$P(\mathbf{x}_N, \mathbf{y}_N) = \sum_{\mathbf{z}_N} P(\mathbf{x}_N, \mathbf{y}_N|\mathbf{z}_N)P(\mathbf{z}_N), \quad (3.9)$$

where $P(\mathbf{z}_N)$ is a distribution over clusterings of the N objects. The key difference from the MMC is that the RMC provides an explicit prior distribution over possible partitions. Importantly, this distribution allows the number of clusters to be unbounded, with

$$P(\mathbf{z}_N) = \frac{(1-c)^K c^{(N-K)}}{\prod_{i=0}^{N-1} [(1-c) + ci]} \prod_{k=1}^K (M_k - 1)!, \quad (3.10)$$

where c is a parameter called the *coupling probability*, and M_k is the number of objects assigned to cluster k . Equation 3.10 is the distribution that results from sequentially assigning objects to clusters with probability

$$P(z_N = k | \mathbf{z}_{N-1}) = \begin{cases} \frac{cM_k}{(1-c)+c(N-1)} & \text{if } M_k > 0 \text{ (i.e., } k \text{ is old)} \\ \frac{(1-c)}{(1-c)+c(N-1)} & \text{if } M_k = 0 \text{ (i.e., } k \text{ is new)}, \end{cases} \quad (3.11)$$

where the counts M_k are accumulated over \mathbf{z}_{N-1} . Thus, each object can be assigned to an existing cluster with probability proportional to the number of objects already assigned to that cluster, or to a new cluster with probability determined by c and N .

As the number of objects increases, it quickly becomes infeasible to enumerate all the different ways to partition them into clusters, so the summation in Equation 3.9 is intractable to compute exactly. To overcome this difficulty, Anderson implemented the RMC using a greedy algorithm which approximates the summation in Equation 3.9 using a single partition \mathbf{z}_N . The algorithm, called *local MAP* by Sanborn et al. (2010a), builds up this approximating partition by assigning observations to clusters in the order in which they are seen, never backtracking to revise these assignments. When a new labeled example (x_N, y_N) is observed, it is assigned to the cluster k which maximizes the posterior probability

$$P(z_N = k | \mathbf{z}_{N-1}, \mathbf{x}_N, \mathbf{y}_N) \propto P(x_N, y_N | z_N = k, \mathbf{z}_{N-1}, \mathbf{x}_{N-1}, \mathbf{y}_{N-1}) P(z_N = k | \mathbf{z}_{N-1}), \quad (3.12)$$

with k being chosen from the set of previously-used clusters and a brand new cluster. Since these cluster assignments are permanent, the local MAP algorithm has the property that the order in which examples are presented can dramatically affect the resulting set of clusters that are inferred, and consequently, the model's predictions for new stimuli.

Despite having been defined in terms of the joint distribution of \mathbf{x}_N and \mathbf{y}_N , the assumption that features and category labels are independent given the cluster assignments makes it possible to write $P(x_N | y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})$ in the same form as Equation 3.8. To do so, note that in the RMC, the probability that the N th observation belongs to the k th cluster is given by

$$P(z_N = k | \mathbf{z}_{N-1}, y_N = j, \mathbf{y}_{N-1}) \propto P(y_N = j | z_N = k, \mathbf{z}_{N-1}, \mathbf{y}_{N-1}) P(z_N = k | \mathbf{z}_{N-1}). \quad (3.13)$$

The second term on the right hand side is given by Equation 3.11. This defines a distribution over the same K clusters regardless of j , but the value of K depends on the number of clusters in \mathbf{z}_{N-1} . Substituting this expression into Equation 3.8 provides the relevant mixture model for the RMC. In general, the probabilities in Equation 3.13 will never be precisely zero: as a consequence, all clusters contribute to all categories. The RMC can therefore be viewed as a form of the mixture model in which all clusters are shared between categories but the number of clusters is inferred from the data. However, the two models are not directly equivalent because the RMC assumes that both features and category labels are generated from the clusters. This assumption induces a dependence between labels and features such

that the prior over y_N depends on \mathbf{x}_{N-1} as well as \mathbf{y}_{N-1} , violating the (arguably sensible) prior independence assumption made by the other models and embodied in Equation 3.6.

The RMC comes close to specifying a unifying rational model of categorization, capturing many of the ideas embodied in other models and allowing the representation to be inferred from the data. However, the model is still significantly limited. First, the RMC assumes a single set of clusters that spans all the categories, an assumption that is inconsistent with many models that interpolate between prototypes and exemplars (e.g., Vanpaemel and Storms, 2008). Second, the idea that category labels should be treated like other features has odd implications, such as the dependence between features and category labels mentioned above. Third, the approximate learning algorithm proposed by Anderson (1991) has serious drawbacks. In order to address these issues, we turn to nonparametric Bayesian statistics.

3.3 Dirichlet process mixtures: nonparametric Bayes and the RMC

Anderson’s 1990 Rational Model of Categorization has a surprising connection to a popular model used in nonparametric Bayesian statistics (Neal, 1998; Sanborn et al., 2010a). The rationale for using nonparametric methods is that real data are not generally sampled from a known, finite-dimensional family of distributions, so it is best to avoid this assumption at the outset. From a Bayesian perspective, the nonparametric approach requires us to use priors that include as broad a range of densities of possible, thereby allowing us to infer very complex densities if they are warranted by data. The most commonly used method for placing broad priors over probability distributions is the *Dirichlet process* (DP; Ferguson, 1973). The distributions indexed by the Dirichlet process can be expressed as countably infinite mixtures of point masses (Sethuraman, 1994), making them ideally suited to act as priors in infinite mixture models (Escobar and West, 1995; Rasmussen, 2000). When used in this fashion, the resulting model is referred to as a *Dirichlet process mixture model* (DPMM; Antoniak, 1974; Ferguson, 1983; Neal, 1998). Although a complete description of the Dirichlet process is beyond the scope of the current chapter, what matters for our purposes is that the Dirichlet process implies a distribution over partitions: any two observations in the sample that were generated from the same mixture component may be treated as members of the same cluster, allowing us to specify priors over an unbounded number of clusters. In the case where N observations have been made, the prior probability that a Dirichlet process will partition those observations into the clusters \mathbf{z}_N is

$$P(\mathbf{z}_N) = \frac{\alpha^K}{\prod_{i=0}^{N-1} (\alpha + i)} \prod_{k=1}^K (M_k - 1)! \quad (3.14)$$

where α is the dispersion parameter of the Dirichlet process, and as in the RMC, M_k is the number of objects assigned to cluster k . This distribution over partitions can be produced

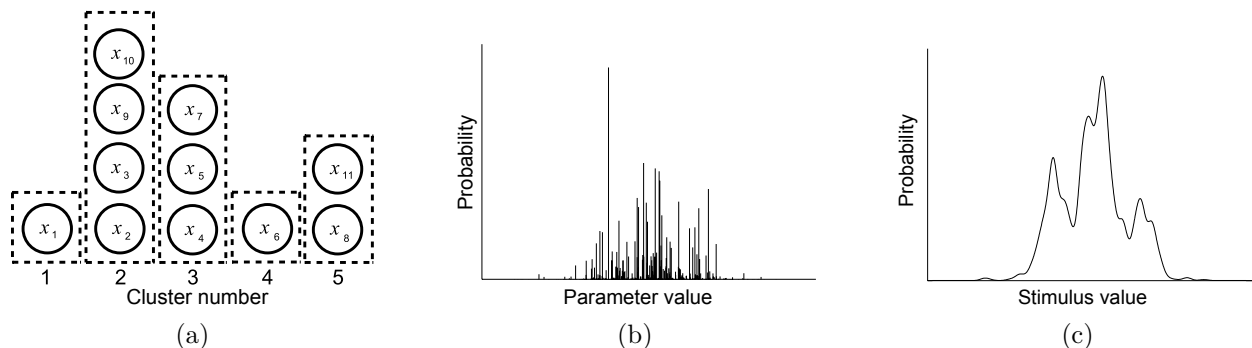


Figure 3.2: (a) The relation between the clustering implied by the DP, (b) the distribution over parameters that is sampled from the DP, and (c) the resulting mixture distribution over stimuli that results in the DPMM. The clustering assignments in (a) are produced by drawing sequentially from the stochastic process defined in Equation 3.15, and each cluster is associated with a parameter value θ . After an arbitrarily large number of cluster assignments have been made, we can estimate the probability of each cluster, and hence of the corresponding parameter value. The resulting probability distribution is shown in (b). If each value of θ is treated as the mean of a simple normal distribution (with fixed variance) over the value of some continuous stimulus dimension (e.g., weight), then the resulting mixture distribution drawn from the DPMM is the one illustrated in (c). Note that in the applications considered in this chapter, the stimuli vary along multiple discrete features, not a single continuous dimension as the illustration in (c) suggests. However, the continuous version makes the concept clearer.

by a sequential stochastic process called the *Chinese restaurant process* (Blackwell and MacQueen, 1973). If observations are assigned to clusters one after another and the probability that observation $i + 1$ is assigned to cluster k is

$$P(z_i = k | \mathbf{z}_{i-1}) = \begin{cases} \frac{M_k}{i-1+\alpha} & \text{if } M_k > 0 \text{ (i.e., } k \text{ is old),} \\ \frac{\alpha}{i-1+\alpha} & \text{if } M_k = 0 \text{ (i.e., } k \text{ is new),} \end{cases} \quad (3.15)$$

then we obtain Equation 3.14 for the probability of the resulting partition. This distribution has a number of nice properties, with one of the most important being exchangeability: the prior probability of a partition is unaffected by the order in which the observations are received (Aldous, 1985). To make some of these ideas more concrete, Figure 3.2 presents a visual depiction of the relationship between the partitioning implied by the DP, the distribution over parameters that is sampled from the DP, and the resulting mixture distribution over stimuli that results in the DPMM.

It should be apparent from our description of the DPMM that it is similar in spirit to the probabilistic model underlying the RMC. In fact, the two are directly equivalent, a point that was first made in the statistics literature by Neal (1998). If we let $\alpha = (1 -$

$c)/c$, Equations 3.10 and 3.14 are equivalent, as are Equations 3.11 and 3.15. Thus the prior over cluster assignments used in the RMC is exactly the same as that used in the DPMM. Anderson (1990, 1991) (impressively) thus independently discovered one of the most celebrated models in nonparametric Bayesian statistics, deriving this distribution from first principles: the RMC is a DPMM that uses Equation 3.9 to specify the mixture distribution. Recognizing this correspondence allows us to explore alternatives to the greedy algorithm first proposed by Anderson for implementing the RMC. Sanborn et al. (2010a) give a thorough account of some alternative algorithms that have been designed for the DPMM, showing how they can overcome some of the limitations of Anderson’s local MAP algorithm. In the same spirit, we use the correspondence between the RMC and the DPMM to apply more recent developments from nonparametric Bayesian statistics towards building new categorization models.

3.4 Hierarchical Dirichlet processes

The Dirichlet process formalizes the idea that probability distributions (and therefore categories) can be thought of as being composed of distinct clusters of objects. To model a set of data from multiple categories with the DP, one could represent each category with its own independent DP. This solution works well in many cases, but it misses the opportunity to take advantages of similarities between the categories. The hierarchical Dirichlet process (HDP), introduced by Teh et al. (2006), was developed as an extension to the DP which can capture a particular type of similarity between multiple probability distributions, thereby forming a more accurate representation from fewer observations.

In the HDP, multiple distributions are not treated as independent, but rather as related to each other probabilistically. Formally, observations are divided into groups, and each group is modeled using a Dirichlet process (with dispersion parameter α). These group-specific DPs are linked together through a top-level Dirichlet process, which allows clusters to be shared between multiple groups. The nature of this mechanism is best understood by considering an example of the HDP’s generative process, by which the model assumes new observations are formed.

When a new observation is generated from a group, it must first be assigned to a particular cluster from that group. Since every group is a DP, the probability of each cluster being chosen is determined by Equation 3.15. If the observation is to be assigned to a new cluster, the new cluster is not created from scratch, but rather, drawn from the special Dirichlet process shared among groups. In this top-level Dirichlet process, all the clusters used in all the groups are represented, so there is some probability that the new observation will be assigned to a cluster that was previously used by another group. The top-level Dirichlet process is governed by the parameter γ , analogous to α , and the prior probability of each cluster is proportional to the number of times that the cluster has been selected by any group, instead of the number of observations in each cluster. The new observation is

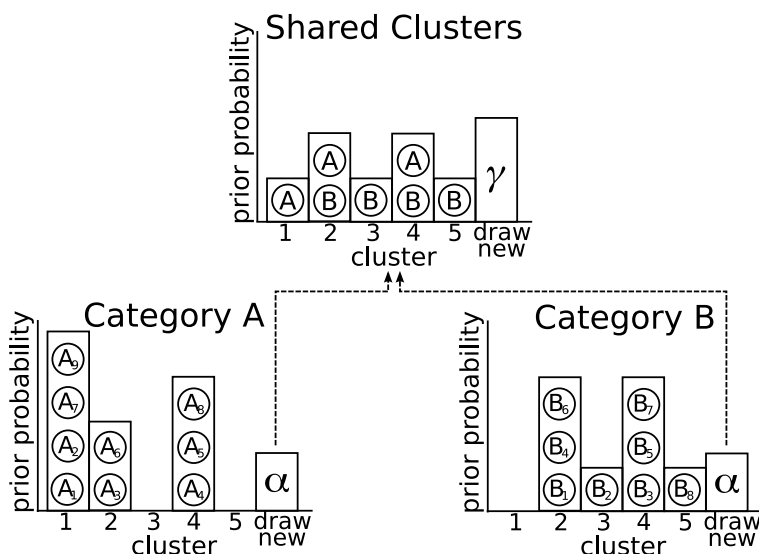


Figure 3.3: Illustration of the HDP prior. The prior probability for each cluster at the lower level is based on the number of observations assigned to that cluster. If a cluster is selected from the top level, the prior probability of clusters is based on the number of times they have been selected at the top level. Completely new clusters can only be created at the top level.

only assigned to a completely new cluster if both Dirichlet processes (at the group level and the top level) select a new cluster. In this manner, observations from different groups can end up belonging to the same mixture component, simply by being drawn from the same cluster at the top level. An illustration of this procedure is shown in Figure 3.3.

The HDP provides a method for modeling related probability distributions across groups of observations. Each distribution is a countably infinite mixture of clusters, but the clusters can be shared between groups. Shared clusters allow the model to pool together examples from different groups to better estimate a cluster's parameters, thereby sharing statistical strength between groups. The number of clusters in each group and the extent to which clusters are shared between groups are influenced by the parameters α and γ , respectively. When α is small, groups will have fewer clusters, and when α is large, the number of clusters will be closer to the number of observations. When γ is small, groups are more likely to share clusters, but when γ is large, the clusters in each group are likely to be unique. By varying the parameters α and γ , a wide range of behaviors can be exhibited by the HDP, through the structure of the clusters that it forms for a set of categories.

3.5 Unifying previous models using hierarchical Dirichlet processes

As described above, the HDP represents a set of probability distributions as countably infinite mixtures, where the mixture components are shared between groups. If we view each group as a category, each mixture component as a cluster, and each observation from a group as an example from the associated category, the HDP becomes a flexible and powerful categorization model. The HDP's two parameters α and γ control, respectively, the amount of clustering within groups and the extent to which clusters are shared between groups. Changing these parameters changes the inferences and predictions made by the HDP, so they effectively index particular variants or instantiations of the HDP. In this way, the HDP can be thought of as a framework of different individual models, each one with a different setting of the α and γ parameters.

We can now use the HDP to define a unifying rational model of categorization, subsuming all previous rational models through different settings of α and γ . Figure 3.4 identifies six models we can obtain by considering limiting values of α and γ .² We denote an HDP model with particular settings of the α and γ parameters by $\text{HDP}_{\alpha,\gamma}$, with the relevant parameter values being either 0, ∞ , or +, where + denotes some finite, positive number.³ Three of the models shown in Figure 3.4 are exactly isomorphic to existing models. The $\text{HDP}_{\infty,\infty}$ is an exemplar model, with one cluster per object and no sharing of clusters. The $\text{HDP}_{0,\infty}$ is a prototype model, with one cluster per category and no sharing of clusters. The $\text{HDP}_{\infty,+}$ is the RMC, provided that category labels are treated as features. In the $\text{HDP}_{\infty,+}$, every object has its own cluster, but those clusters are generated from the top-level Dirichlet process. Consequently, group membership is ignored and the model reduces to a Dirichlet process.

As well as unifying previous categorization models, the HDP framework makes it clear that there are new, unexplored models available. The $\text{HDP}_{0,+}$ makes the same basic assumptions as the prototype model, with a single cluster per category, but makes it possible for different categories to share the same prototype – something that might be appropriate in an environment where the same category can have different labels. However, the most interesting models are the $\text{HDP}_{+,+}$ and the $\text{HDP}_{+,\infty}$. These models are essentially the MMC, with clusters shared between categories or unique to different categories respectively, but the number of clusters in each category can differ and be learned from the data. Consequently, these models make it possible to answer the question of whether a particular category is best represented using prototypes, exemplars, or something in between, simply based on the structure of that category. Furthermore, the $\text{HDP}_{+,+}$ provides a formal account of how

²The case of $\gamma \rightarrow 0$ is omitted, since it simply corresponds to a model in which all observations belong to the same cluster across both categories, for all values of α .

³The HDP is technically undefined when α or γ are set to 0 or ∞ , but its behavior converges in the limit as either parameter goes towards 0 or ∞ .

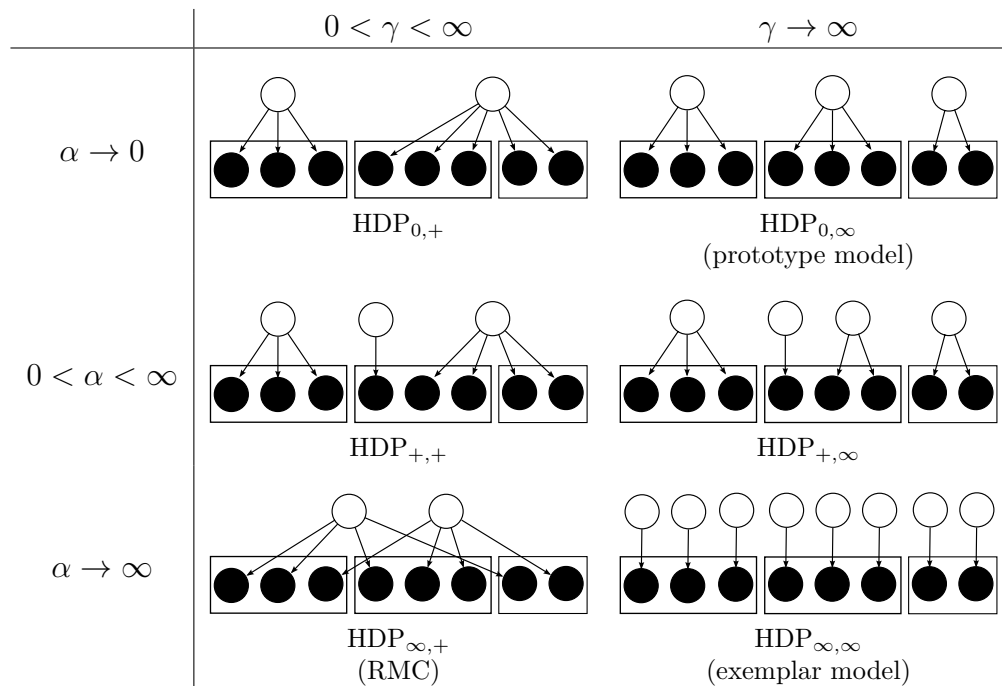


Figure 3.4: Structural assumptions underlying different parameterizations of the $\text{HDP}_{\alpha,\gamma}$ model. The unfilled circles are clusters, the filled circles are exemplars, and the boxes indicate which exemplars belong to the same categories. Descriptions of the properties of these six models and their correspondence to existing models are given in the text.

clusters that are shared between categories can be automatically inferred from the observed data.

3.6 Cluster-sharing in the $\text{HDP}_{+,+}$ model

Not only does the HDP serve as a unifying model to tie together several previous models of categorization in a coherent framework, but one of its variants, the $\text{HDP}_{+,+}$, which corresponds to a previously-unexplored category model, provides an interesting new way to capture relationships between multiple categories through its cluster-sharing mechanism. Viewing categorization models as methods for solving the density estimation problem (as described by Ashby and Alfonso-Reese, 1995), we can characterize previous models as placing independence assumptions on the probability density functions for each category. Through the top-level DP, the $\text{HDP}_{+,+}$ explicitly makes the set of category density functions statistically dependent on each other, dropping these independence assumptions and adding more flexibility to the inferences it can make.

Specifically, the $\text{HDP}_{+,+}$ model allows a cluster that was previously learned in the

context of one category to be reused in the context of another category. Because of this, when a small number of objects from such a cluster (even just one) is seen in a new category, the model increases the chance of that cluster being associated with other members of the new category. In this way, the model can form a prediction that all the other objects that belong to that cluster are automatically more likely to belong to the new category. Another way to think about how the $\text{HDP}_{+,+}$ accomplishes this is that it builds up a “vocabulary” of clusters of objects which can be arbitrarily composed with each other to form category representations. Because the vocabulary is shared across all categories, the characteristics of each cluster need only be learned once, even if a cluster appears in multiple categories.

This cluster-sharing ability of the $\text{HDP}_{+,+}$ model predicts a specific type of *transfer learning* – the abstraction and reuse of information from one learning problem to another (Thrun and Pratt, 1998) – in cases where a set of multiple categories is being learned. This phenomenon is closely related to the concepts of *overhypotheses* and *learning to learn* in the psychology literature. For example, Kemp et al. (2007) demonstrated a hierarchical parametric model (related to the HDP, but using parametric distributions instead of infinite mixtures of clusters) which can recognize and infer similarities between categories, thereby speeding up learning about new categories. However, because their model uses parametric distributions, its inferences are constrained to limited parametric families. Similar transfer learning effects can be achieved by using categorization models that incorporate dimensional attention parameters, such as the generalized context model (Nosofsky, 1986) and ALCOVE (Kruschke, 1992). These models can learn that certain dimensions of the stimulus feature space are more salient than others, and they accordingly tune a set of dimension-specific weights to improve learning performance. If these dimensional attention weights are shared globally among all categories, then observations from one category can influence the inferences made about other categories. However, this mechanism is a more restrictive form of learning to learn than the ability of the $\text{HDP}_{+,+}$ to learn a vocabulary of clusters of stimuli that can be arbitrarily composed together to form category representations.

3.7 Exploring cluster sharing with human learners

In the previous section, we discussed how the $\text{HDP}_{+,+}$ model is designed to be able to share clusters between categories in order to facilitate learning with multiple categories. This raises some interesting empirical questions regarding human learners: are people able to recognize discrete clusters of objects within individual categories, and do they reuse these clusters when learning new categories? The $\text{HDP}_{+,+}$ model makes a specific empirical prediction, novel to the categorization literature, that learners would be more likely to incorporate a previously learned cluster of objects into a new category after observing just one member of that cluster in the new category. In order to test this prediction, we conducted an experiment where people were trained on categories composed of shared clusters and asked to make generalization judgments about new categories.

In particular, the experiment is different from those that have tested attention to dimensions. Previous work has shown that people are able to learn that particular dimensions are useful for both describing the structure of a category (Austerweil and Griffiths, 2010; Kruschke, 1992) and for generalizing to new categories (Perfors and Tenenbaum, 2009). Here, the generalization trials are balanced so that attention to a single dimension or subset of dimensions would not lead to any generalization of the type that we aim to measure.

3.7.1 Method

3.7.1.1 Participants

The experiment was completed by 23 undergraduate students at the University of California, Berkeley, who received course credit, and 53 participants recruited online via the Amazon Mechanical Turk website (<http://mturk.com>), who were paid roughly \$6 per hour, for a total of 76 participants.

3.7.1.2 Stimuli

Participants were given a cover story about being a farmer’s apprentice and having to learn about four different types of crops. The stimuli were fruit-like images adopted from Sanborn et al. (2010b). They were generated using six parameters, but we selected a three-dimensional subspace within the original six-dimensional space to create our stimulus set. The three dimensions we selected correspond to size, lightness, and hue. Within this 3-dimensional space, we chose three multivariate normal distributions to serve as the clusters which comprised the categories. The three clusters are depicted in Figure 3.5. Each cluster is axis-aligned, meaning it has non-negligible variance in only one of the dimensions. Furthermore, there is exactly one cluster that varies along any given dimension, so after observing examples from all three clusters, the dimensions should be about equally salient. Figure 3.6(a) depicts the arrangement of the three clusters in the three-dimensional feature space. In its dimension of primary extent (the dimension along which its black edge extends), a multivariate normal distribution has standard deviation equal to $\frac{1}{3}$ the length of the edges of the cube in Figure 3.6(a). Along the other two dimensions, a distribution has standard deviation equal to only $\frac{1}{10}$ of an edge’s length.

The three clusters were paired together to form three categories, as shown in Figure 3.7. Each category is composed of two clusters, and each cluster appears in two categories. These categories were taught to the participants through the course of the experiment, but the underlying clusters that compose them were never explicitly referenced. During the experiment, the categories were referred to using the randomly-assigned nonsense-word labels *pasps*, *worbs*, and *broms*.

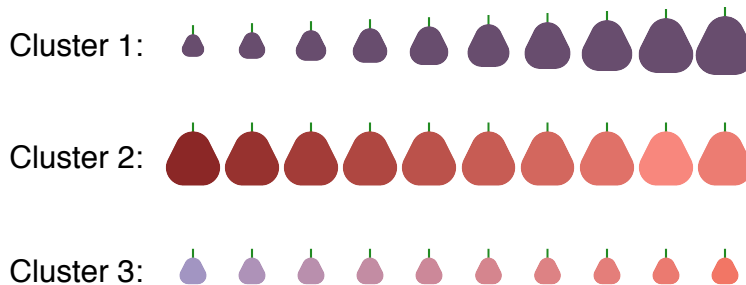


Figure 3.5: Examples of stimuli from the three clusters used in the experiment. Each row contains ten examples from a cluster. Each cluster varies along one of the three dimensions: size, lightness, and hue.

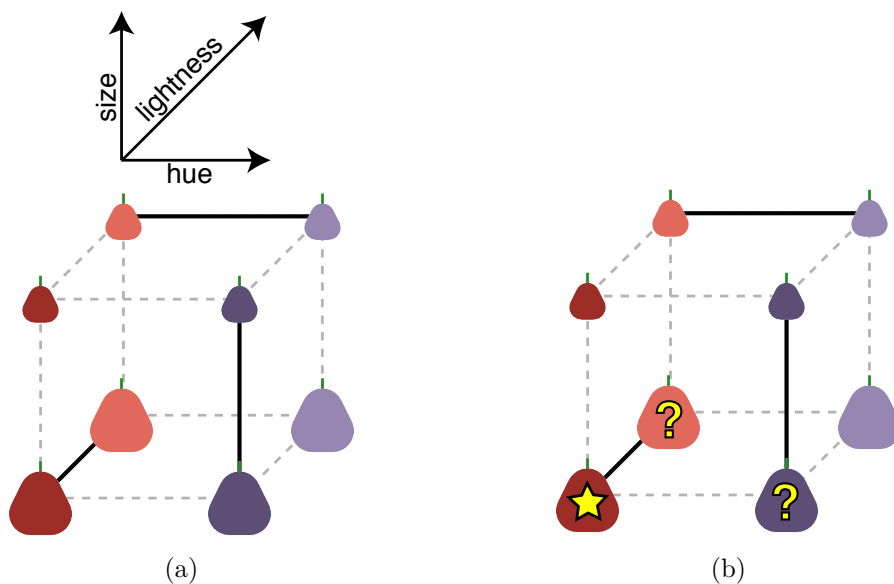


Figure 3.6: Structure of stimuli used in training and testing. (a) The orientation of the feature space of the stimuli used in the experiment. The stimuli span the region encompassed by a cube in the three dimensions of size, lightness, and hue. The three clusters used in the experiment are depicted as thick black edges of the cube. Each cluster has a probability density function which is an axis-aligned multivariate normal distribution, centered on and primarily extending along the corresponding edge. Along its edge, the standard deviation of a cluster’s distribution is $\frac{1}{3}$ of the edge’s length. The standard deviation along the other two dimensions is $\frac{1}{10}$ of an edge’s length. (b) A schematic diagram of the arrangement of the stimuli used in the pre-test and post-test trials, for one example trial. Here, the two alternatives (indicated by question marks) are equally distant from the observed stimulus (indicated by a star), but one is in the same cluster and the other is not. This allows us to test the extent to which participants prefer to generalize to within-cluster alternatives vs. out-of-cluster alternatives.

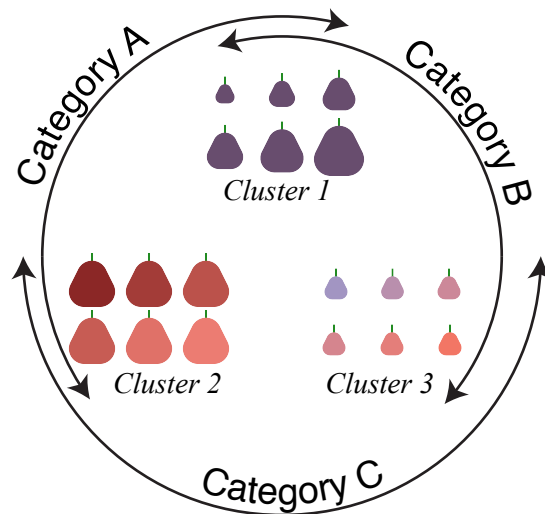


Figure 3.7: The assignment of clusters to categories in the experiment. Three categories were constructed by taking each pair of the three clusters.

3.7.1.3 Procedure

The experiment was comprised of three sessions: a pre-test, a training session, and a post-test. The two test sessions were identical, except that their individual trials were randomly shuffled for each participant. The test sessions were designed to measure the degree to which participants learned the clusters comprising the categories and used those clusters when making generalization judgments about new categories. Thus, the experiment allows us to measure the extent to which the training session altered participants' inferences by comparing their performance in the pre-test and the post-test.

The pre-test and post-test consisted of a set of trials in which participants were presented with a single stimulus and told to imagine that it belonged to a novel category called *relts*. They were then asked to choose one of two alternative stimuli that appeared below the *relt*, indicating which one they believed to be more likely to also belong to the category *relts*. The configuration of the observed stimulus and the two alternative stimuli in the feature space for one example trial is depicted in Figure 3.6(b). In each trial, the observed stimulus and the two alternatives were all chosen from the set of six cluster *endpoints*. The cluster endpoints are the leftmost and rightmost stimuli in each of the rows of Figure 3.5; they are also seen as the six corners of the cube in Figure 3.6 that are connected to a black edge. For a given observed stimulus, one of the two alternatives was always chosen to be the other endpoint of the same cluster (this is called the *within-cluster alternative*), leaving four remaining choices for the second alternative (the *out-of-cluster alternative*). Combining all possible choices of the observed stimulus and two alternatives gives $6 \times 4 = 24$ possible trials. We presented each of these trials twice, once with the within-cluster alternative on

the left, and once with it on the right, for a total of $24 \times 2 = 48$ trials for each test session. In our analysis of the results, we only considered those 12 trials per test session for which the two alternatives are equidistant from the observed stimulus.

Between the pre-test and post-test, there was a training session in which participants were asked to help the farmer harvest three different types of crops. As mentioned above, the three clusters were grouped into pairs to form the three categories, and participants repeated training blocks until they learned which stimuli belonged to each category. First, an 8-by-8 grid of crops was displayed on the screen. Twenty-one stimuli were generated from each cluster and randomly positioned to fill the grid (leaving one random grid space blank). Participants were then asked to harvest all the crops belonging to one of the three categories. Participants harvested crops by clicking on them, and each time a correct stimulus was chosen, it was moved to the right side of the screen in a region corresponding to the target category. Once all the category members were harvested, the grid was randomly repopulated and the participant was asked to harvest the next category of crops. Each of these is called a new training block. Participants rotated through all three categories, repeating as many training blocks as was necessary until a performance criterion (no more than three errors in the last four blocks of training) was met. A screenshot of the training session is shown in Figure 3.8. Once a participant completed the training session, they have demonstrated their ability to select the members of each of the three categories. The question we are interested in is whether, in learning these categories, they formed mental representations of the categories' underlying clusters, and whether they would use these clusters in future inferences about new categories.

Following the training session, the participants completed the post-test, which was identical in format to the pre-test described above. During the post-test, the participants were able to view the right side of the screen, which contained the crops they had harvested during the training session.⁴ This was designed to eliminate the memory demands of the experiment, so that the results depended more on how the participants learned and generalized, rather than how well they memorized the training examples. A screenshot of the post-test is shown in Figure 3.9.

The experiment was designed to test whether the training session affected the judgments made by the participants in the post-test, compared to their performance in the pre-test. Specifically, we hypothesized that in learning the categories during the training session, the participants would form mental representations of the underlying clusters (despite never being told about them), and then prefer to generalize to within-cluster alternatives rather than out-of-cluster alternatives when given a single observation from a new category. In order to test this hypothesis, we focused on the test trials in which the two alternative stimuli were equidistant from the observed stimulus, as they are in the example illustrated in Fig-

⁴Once this region of the screen ran out of room, the least recently-harvested crop was removed to make room for each newly-harvested crop. Despite this, a large and representative collection of category examples always remained in view.



Figure 3.8: A screenshot from the training session in the experiment. The participant is asked to harvest the crops on the grid which belong to a particular category by clicking on them. Correctly chosen stimuli are moved to the right side of the screen for later reference.



Figure 3.9: A screenshot from the post-test session in the experiment. The pre-test session is identical, but the training examples do not appear on the right side of the screen.

ure 3.6(b). These comprised 12 of the 48 trials from each test session. During the pre-test, because the trials were balanced across the three dimensions, the participants should have had no reason to prefer the within-cluster over the out-of-cluster alternative. During the post-test, any consistent preference for the within-cluster alternative should be due to the observations of the categories in the training session.

3.7.2 Results

No significant differences were found between the data from the two groups of participants, so their data were combined for all analyses. The average number of training blocks completed by the participants before reaching the performance criterion was 6.83; the median was 5, and the most common number was 3 (the enforced minimum, since there were three categories to learn). The maximum number of blocks required for any participant was 25.

We analyzed the percentage of test trials for which participants selected the within-cluster alternative vs. the out-of-cluster alternative. This percentage is called the *pre-test score* for the pre-test and the *post-test score* for the post-test. Because the participants had not yet been exposed to the constructed categories before the training session, they should have had no reason to prefer the within-cluster alternative during the pre-test. For this reason, we expect the pre-test scores to be distributed around 50%. This result should be expected even if the three dimensions are not equally psychologically salient to the participants, since the trials were balanced across all three dimensions. During the post-test, if the participants preferred to generalize more often to the within-cluster alternatives, as we hypothesize, their post-test scores should be statistically higher than 50%.

A summary of the participants' pre-test scores and post-test scores is shown in Figure 3.10(a). The participants' pre-test scores are clustered around 50%, as expected. Although many of the post-test scores are also near 50%, the majority are above 50%, and there is a clear trend towards higher post-test scores. A significant number of the human learners exhibited post-test scores well above 50%, suggesting they recognized the underlying clusters and preferred to generalize to stimuli in the same cluster as an observed stimulus from the new category. Most of the participants had a post-test score that was higher than their pre-test score, as shown in the second plot of Figure 3.10(a). A two-tailed t -test failed to reject the null hypothesis that the mean of the pre-test scores was 50% ($t(75) = -0.97$, $p = 0.33$) and successfully rejected the null hypothesis that the mean of the post-test scores was 50% ($t(75) = 6.34$, $p < 0.01$). Furthermore, a two-tailed, paired t -test rejected the null hypothesis that the difference between the pre-test and post-test scores has mean zero ($t(75) = 6.71$, $p < 0.01$).

Breaking down the pre-test scores and post-test scores by the cluster of the observed stimulus, as shown in Table 3.1, we can see that the assumption that the dimensions were roughly equally salient did not hold true. In both the pre-test and the post-test, the participants clearly preferred to generalize within cluster 1, then cluster 2, then cluster 3, in that order. This might be because the stimuli resemble fruit or vegetables and participants

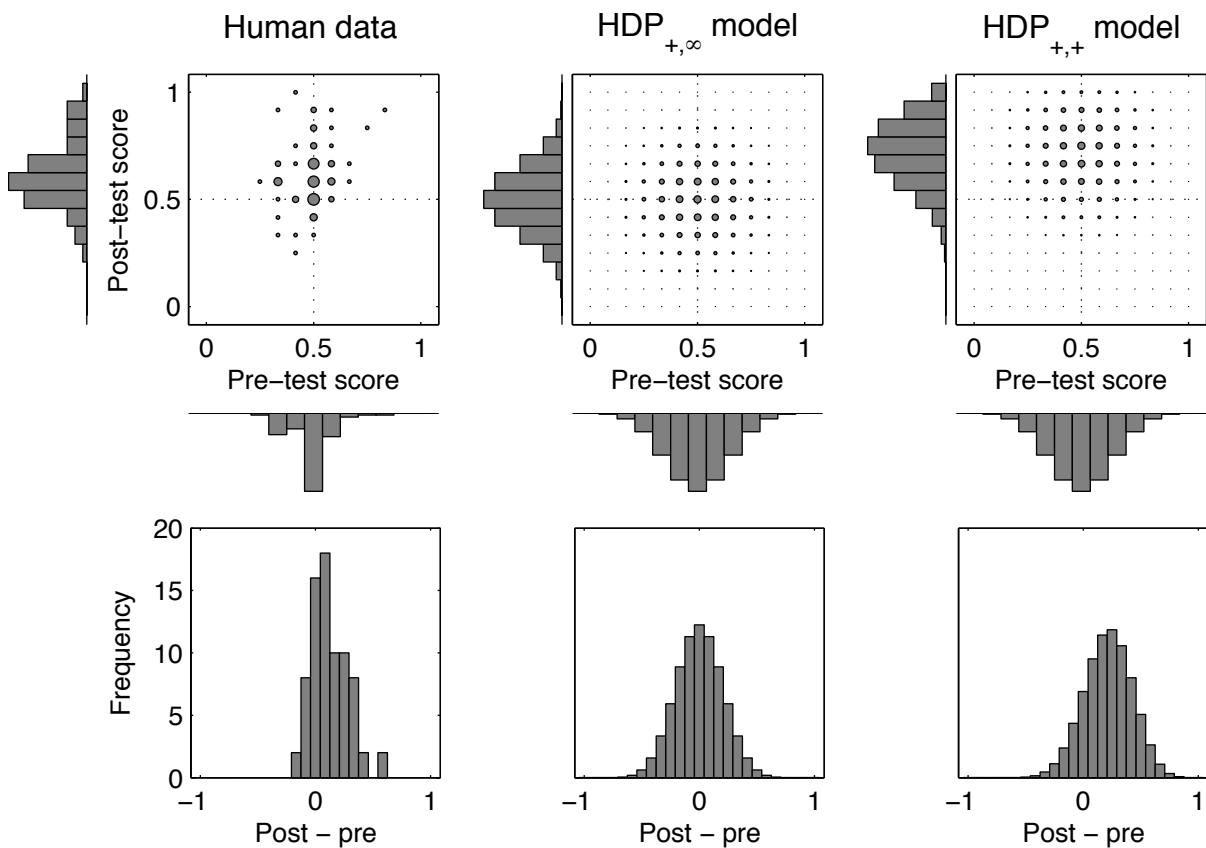


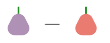


Figure 3.10: Experimental results from the human learners and the two models. In the first row of plots, each marker represents a number of participants or a model’s probability, with the size of the markers proportional to the number of participants represented or the model’s probability. The horizontal axis is the percentage of pre-test trials where the learner generalized a novel category to a stimulus within the same cluster as the observed stimulus (the “pre-test score”). The vertical axis is the same percentage for the post-test trials (the “post-test score”). Marginal histograms are shown on each axis. The plots in the second row are histograms of the differences between the pre-test scores and the post-test scores.

Table 3.1: Average per-cluster pre-test and post-test scores of the human learners

Cluster	Dimension	Endpoints	s_{pre}	s_{post}	$s_{\text{post}} - s_{\text{pre}}$	t -statistic	p -value
1	size		80.66%	86.79%	+6.13%	$t(75) = 0.85$	$p = 0.40$
2	lightness		51.89%	68.40%	+16.51%	$t(75) = 4.75$	$p < 0.01$
3	hue		12.74%	25.94%	+13.21%	$t(75) = 4.70$	$p < 0.01$
All	—	—	48.43%	60.38%	+11.95%	$t(75) = 6.71$	$p < 0.01$

Note: The average pre-test score and the average post-test score are denoted by s_{pre} and s_{post} , respectively. The second column shows which dimension each cluster primarily varies along. The third column shows images of each cluster’s endpoints (the two stimuli used in the test sessions). The sixth column shows the average difference between the pre-test and post-test scores. The last two columns show the statistics of the two-tailed, paired t -test for the difference between the cluster-specific pre-test and post-test scores. The last row shows these same quantities aggregated across all clusters.

were told that they were assisting a farmer in harvesting crops. Individual varieties of fruits and vegetables tend to vary more in size than in color as they grow, so cluster 1, which varies primarily along the size dimension, seems most plausible to be a single type of fruit or vegetable. Nonetheless, the scores increased from the pre-test to the post-test for all three clusters (see Table 3.1 for cluster-specific paired t -test statistics), which shows that participants consistently exhibited the expected transfer learning effect.

We also checked whether the participants who spent more time in the training session were more or less likely to exhibit higher post-test scores. The correlation between the number of training sessions a participant completed and their post-test score was -0.10 , which is not statistically significantly different from zero according to a t -test ($t(74) = -0.82$, $p = 0.41$).

3.7.3 Modeling

The experimental results indicate that human learners were able to recognize the set of clusters underlying the categories and preferred to generalize within these clusters when making inferences about new categories. We have discussed how the $\text{HDP}_{+,+}$ categorization model is theoretically able to predict this type of behavior. In order to empirically test this property of the model, we repeated the experiment using the $\text{HDP}_{+,+}$ model, which allows cluster-sharing between categories, and the $\text{HDP}_{+,\infty}$ model, which does not allow cluster-sharing.

It can be shown that for every trial in the pre-test where both alternatives are equidistant from the observed stimulus, all rational models will assign exactly 50% probability to both of the two alternatives. This is because during the pre-test, no feedback or training has

been given, and the models are ignorant of the psychological significance of the dimensions of the stimulus feature space. For this reason, for both the $\text{HDP}_{+, \infty}$ and $\text{HDP}_{+, +}$ models, we represent the pre-test score as a binomial distribution with 12 samples and probability 50% (recall that in each test session, we are interested in the 12 trials for which the two alternatives are equidistant from the observed stimulus). Because the $\text{HDP}_{+, \infty}$ model performs learning and inference independently for each category, the clusters that are learned in the context of the other categories have no way to influence the representation of the new category. Therefore, the $\text{HDP}_{+, \infty}$ model will also assign 50% probability to both alternatives in the post-test, so we represent its post-test score, again, as a binomial distribution with 12 samples and probability 50%.

The $\text{HDP}_{+, +}$ model allows clusters learned in the training session to affect the inferences made in the post-test, so its post-test predictions must be computed explicitly. We implemented the model using a computer program written by Teh (2004) and designed for inference with HDP models. The software uses the Gibbs sampling algorithm described by Teh et al. (2006). For each participant, we trained a separate instance of the $\text{HDP}_{+, +}$ model using the same number of observations that the participant saw during their training session, then estimated the resulting distribution over the post-test scores. These distributions were summed across all model instances to construct a final post-test score distribution that can be compared to that of the participants. The model was assumed to choose each alternative stimulus in the post-test trials according to its posterior probability, using a strategy of probability matching rather than always choosing the more probable alternative.

The $\text{HDP}_{+, +}$ model constrains the α and γ parameters only to be positive and finite; for each distinct choice of these parameters, the model will make different quantitative predictions. Rather than fixing these parameters at particular values, it is common to perform Bayesian inference over α and γ by specifying a prior distribution over each of them. This method essentially allows their values to be inferred from the observed data. Following this strategy, we used exponential prior distributions with mean 1 to learn both parameters. This is a weak prior which allows a wide range of parameter values to be explored, essentially allowing the model to find the best-fitting values of α and γ for a given data set.

The model results are shown in Figure 3.10(b) and 3.10(c). The models, using the binomial assumption described above, exhibited much higher variance than the human learners in their pre-test scores. Presumably, this is because the participants preferred to generalize along the three dimensions unequally, preferring size, then lightness, then hue, as discussed earlier. A participant answering in strict accordance with this preference rule would achieve a pre-test score of exactly 50%, since the trials are balanced across the three dimensions. If the models were given information about the relative psychological salience of the three dimensions, they would perhaps be able to provide a better account of the pre-test scores in the human data. However, the main focus of this experiment is how training affects people's judgments about new categories, which is revealed by the post-test scores.

Because the $\text{HDP}_{+, \infty}$ model has post-test scores distributed according to a binomial

distribution centered at 50%, it does not capture the increased post-test scores of the human learners. The $\text{HDP}_{+,+}$ model, on the other hand, shows a strong transfer effect. The $\text{HDP}_{+,+}$ model clearly preferred to assign the observed stimuli of the new category to the clusters learned during the training session. Its post-test score distribution is significantly higher than that of the human learners (with a mean of 71.51%, vs. a mean of 60.38% for the human data). One potential explanation for this difference is that the participants may have exhibited individual differences in the degree to which they used information about other categories when making judgments about the new category in the post-test. Another possibility is that the imbalance in the salience of the three dimensions mitigated the transfer learning effect. Participants already strongly preferred to generalize along the *size* dimension (see Table 3.1), so test trials using cluster 1 already had high pre-test scores, leaving little room for the post-test scores to increase. On the other hand, participants showed a strong aversion to generalizing along the *hue* dimension in the pre-test, so the observed examples of cluster 3 (the hue-varying cluster) in the training session might not have provided enough evidence to convince participants to generalize within this cluster in the post-test. Cluster 2, which was the only cluster which had roughly the expected 50% average pre-test score, showed the largest difference between the average pre-test score and the average post-test score. Therefore, we might expect that if the dimensions were more equally balanced, the expected transfer learning effect would be greater in the human data. Although the $\text{HDP}_{+,+}$ model does not provide a perfect quantitative fit to the human data, our experiment successfully demonstrates a novel effect in human categorization, and the $\text{HDP}_{+,+}$ model demonstrates the same qualitative effect, providing a formal account of how people might be achieving this behavior.

3.8 Discussion

One of the most valuable aspects of rational models of cognition is their ability to establish connections across different fields. Here, we were able to exploit the correspondence between Anderson's 1990 Rational Model of Categorization and the Dirichlet process to draw on recent work in nonparametric Bayesian statistics. Using this correspondence, we defined a more general rational model, based on the hierarchical Dirichlet process. This model subsumes previous rational analyses of human category learning, and provides a general solution to the problem of selecting the number of clusters to represent a category. In addition, one of its variants, the $\text{HDP}_{+,+}$, corresponds to a categorization model in which categories are not learned separately, but can influence each other through sharing of clusters. The result is a picture of human categorization in which people do not use a fixed representation of categories across all contexts, but instead select a representation whose complexity is warranted by the available data.

In the remainder of the chapter, we discuss three issues raised by our analyses. First, we consider other forms of transfer learning, and how they relate to our model and our

experiments. We then address the question of how the mixture models that have been the focus of our approach relate to rule-based models of categorization. Finally, we identify some of the limitations of our approach, and outline some possible extensions of the HDP that address these limitations.

3.8.1 Other forms of transfer learning

Transfer learning simply means using information learned in one context to inform learning in another (Thrun and Pratt, 1998). This kind of learning is a common focus of rational models of cognition, falling naturally out of the properties of hierarchical Bayesian models (Tenenbaum et al., 2006; Kemp et al., 2007). However, simple forms of transfer learning have also been investigated in previous models of categorization. As mentioned earlier in the chapter, dimensional attention can be construed as a kind of information that is shared among categories, with changes in the weights assigned to different dimensions being something that can be learned from one category and then generalized to another. Models such as the Generalized Context Model (Nosofsky, 1986) or ALCOVE (Kruschke, 1992) incorporate mechanisms that allow them to engage in this kind of transfer learning.

The influence of dimensional attention can be captured in rational models similar to those that we have presented. In the RMC and HDP models, each cluster is associated with parameters that characterize its variance along each dimension of the stimuli. These parameters can be estimated separately for each cluster, producing dimensional-attention effects within clusters, jointly for all clusters, producing dimensional-attention effects across clusters, or separately but with a jointly estimated prior, combining these two effects. While this property of these rational models has not been investigated extensively, recent work has begun to explore how generalizations about variability can be accommodated. Kemp et al. (2007) presented a model that inferred the variability of distributions, which Perfors and Tenenbaum (2009) applied to modeling transfer learning in categorization. Heller et al. (2010) described a hierarchical model based on the RMC that learns which dimensions are more likely to vary among a set of categories and infers that new categories are more likely to also vary along these common dimensions.

Despite its importance in categorization more generally, dimensional attention is not sufficient to account for the transfer effect seen in our experiment. The test trials in our experiment were constructed so that they would not be affected by differences in the salience of the three dimensions used to generate the stimuli, being balanced across those three dimensions. Models that incorporate mechanisms supporting changes in dimensional attention, but not the possibility of sharing of clusters between categories, should therefore not be able to reproduce this transfer effect. We view sharing of clusters as a novel form of transfer learning in the context of categorization, and hope that future work will complement our rational analysis with richer process models that can accommodate this kind of effect.

3.8.2 Connections to rule-based models

One question explored by many researchers is whether the mind contains a separate rule-based system for learning concepts and categories. The earliest categorization models focused on developing rule-based systems for describing categories (Bruner et al., 1956), but these models lost favor due to their difficulty handling uncertainty and typicality (for a review, see Murphy, 2002). The work on exemplar and prototype models that we have focused on in this chapter was complemented by work on *decision-bound* models, in which categorization is modeled by assuming that people learn the boundaries between categories, which can appear rule-like when boundaries align with stimulus dimensions (Ashby and Gott, 1988; Ashby, 1992). Models that explicitly form rules characterizing categories have begun to return to popularity recently, with the introduction of new models such as RULEX (Nosofsky and Palmeri, 1998), which represents categories as combinations of simple logical rules along with stored exceptions to these rules.

Decision-bound models have a clear formal relationship to the rational models that we have presented. Estimating the probability densities associated with categories and identifying category boundaries are closely related problems. Every set of probability densities implies a set of category boundaries, and particular classes of density estimation methods correspond to particular schemes for estimating category boundaries (Ashby and Maddox, 1993; Ashby and Alfonso-Reese, 1995). For example, using parametric density estimation where the parametric family is Gaussian and the covariance matrices of two categories are the same will result in a linear decision boundary. More generally, density estimation using multivariate Gaussians will result in decision boundaries that are conic sections – hyperbolas, parabolas, and ellipses (for details, see Duda et al., 2000). These approaches are thus closely related at the rational level, although the different way in which they construe the learning problem can have implications for the conclusions that rational learners draw (e.g., Hsu and Griffiths, 2010).

Although probabilistic methods seem to be quite different from rule-based categorization models that aim to find a logical formula that defines a category, there are also some interesting connections between these approaches. First, it is possible to explicitly formulate a rule-based categorization model within the same kind of probabilistic framework that we have used to define the rational models in this chapter. The Rational Rules model introduced by Goodman et al. (2008) uses probabilistic grammars to define a prior distribution over logical formulas describing categories, which are then connected to observations by assuming that the category labels of objects are assigned with some small amount of noise. However, it is not necessary to go beyond the HDP framework to find a probabilistic model that can be given a logical interpretation.

There is also a sense in which mixture models such as the Dirichlet process and hierarchical Dirichlet process can be loosely thought of as rule-based models, since their representations are combinations of individual components. Each cluster in a mixture model can be thought of as a conjunction of constraints on the feature values of a stimulus. This

is clearest with discrete features, where the distribution over feature values associated with the cluster will prefer a certain conjunction of features. A mixture of multiple clusters is then similar to disjunctive combination of these constraints. Intuitively, the mixture indicates that members of the category could come from any of these clusters, making the category a disjunction of the conjunctions of features represented by the clusters. Under this interpretation, these mixture models can be thought of as probabilistic analogues of rule-based models using logical formulas in disjunctive normal form to define categories.

3.8.3 Limitations and extensions

While we have argued that the HDP provides a unifying framework for rational models of categorization, we do not expect this framework to contain all possible rational models. The cluster-based categorization models that we have described all support a method of building up arbitrarily-complex category representations by learning and combining clusters of similar objects. While this technique has interesting theoretical properties and has been used to explain certain aspects of human categorization (Griffiths et al., 2007; Sanborn et al., 2010a; Canini et al., 2010), these models take advantage of just one of many ways in which categories can exhibit interesting structure. We anticipate that future work on rational models of categorization will explore how the kind of models we have discussed can be extended to capture some of these richer aspects of category structure.

Taxonomic structures are one way in which natural categories go beyond merely sharing clusters of members. For example, the well-known biological taxonomic system of classification has at least seven levels of categories: *kingdom*, *phylum*, *class*, *order*, *family*, *genus*, and *species*. In these hierarchically-organized systems, categories in higher levels of the taxonomy are supersets of those categories at the lower levels. This level-based analysis of categories underpins some of the seminal work in the study of categorization in psychology, in which categories are thought of as occupying *basic-level*, *superordinate*, or *subordinate* roles (Rosch et al., 1976). Learning in these settings requires not only discovering what objects belong to which categories, but also how the categories are related to each other in the hierarchical system. The hierarchical nature of the HDP can be used to perform category learning in multi-level systems such as these. Canini and Griffiths (2011) provided a preliminary investigation of this possibility, extending the HDP to allow the structure of these taxonomic systems to be automatically inferred by the model, and showing that people are similarly capable of learning in these settings.

Systems of categories can also exhibit other kinds of relationships. For example, consider learning categories corresponding to *spotted* and *striped* as well as *dog*, *cat*, *dalmatian* and *tabby*. Here, we might consider representing *dalmatian* as the intersection of *spotted* and *dog*, and *tabby* as the intersection of *striped* and *cat*. The capacity to logically combine categories is part of what makes human learners capable of learning in a way that builds on their previous experience, and one of the most powerful forms of transfer learning. At present, this kind of logical combination is beyond the capacity of the nonparametric Bayesian models

we have discussed, but the relationship between the HDP and logical representations mentioned in the previous section, together with other work on rational models of rule learning (Goodman et al., 2008), provides a foundation for developing models that have this capacity.

3.8.4 Conclusion

Strategies for learning and using categories that might appear quite different from one another can be seen to have common properties when considered from the perspective of rational analysis. By considering the computational problem that underlies categorization, we have defined a unifying model in which different models appear as variants. This approach has two consequences. First, it allows us to conceive of different approaches to categorization as varying in their appropriateness for different kinds of data, creating the possibility that learners might adapt their strategies between these models in a way that is guided by the data. Second, it lets us identify new models that go beyond existing rational models of categorization, and predict novel empirical phenomena. In this chapter, we showed that people can transfer knowledge between instances of category learning by sharing clusters between categories. This result provides a first step towards understanding the rich structures that characterize people's representations of categories, and identifying how they are able to use existing knowledge to learn new categories quickly.

Chapter 4

Multi-level category learning

Taxonomic structures are a ubiquitous part of the way that people think about the world, appearing in biological phylogenies (Atran, 1998), linguistic ontologies (Keil, 1979), and many natural kind concepts (Rosch et al., 1976). In a taxonomy, categories are organized into different hierarchical levels, with the higher-level categories representing broader, more inclusive groups of entities, and the lower-level concepts representing narrower, more specific groups. When a category is a direct descendent of another category in a taxonomy, most or all the members of the first are also members of the second. For example, in our intuitive phylogenetic tree for animals, all dogs are mammals, and all border collies are dogs. This taxonomic structure supports efficient inferences about the properties of entities belonging to these categories (Collins and Quillian, 1969).

The ubiquity of taxonomies raises a natural question: How can such structures be learned? While we might get some explicit information about the taxonomic relationships between categories, neither children nor artificial systems can rely on such information. Consequently, in this chapter I focus on the question of how taxonomies might be learned just from labeled examples of category members. Consider the problem faced by a learner who sees a collection of objects given the labels “animal”, “mammal”, “dog”, and “border collie”. The challenge is to induce an appropriate representation for the categories associated with each of these labels, supporting future generalizations, and to determine how these categories are related to one another. For example, our learner would need to identify categories corresponding to “dog” and “border collie”, and learn that “border collie” is a kind of “dog”. Since the objects can each be associated with multiple labels, and the categories are defined at different levels of abstraction, I refer to this problem as *multi-level category learning*.

The complex relationships between categories make the problem of multi-level category learning quite different from the standard treatment of category learning (or multi-class classification) in cognitive science and machine learning. Most methods for learning categories do not allow complex relationships to exist between those categories. Typically, either categories are treated as independent (for example, by learning conditional distributions

over the observed features of the objects separately for each category) or algorithms consider only basic interactions between categories (for example, discriminative methods attempt to discover the boundaries between categories). Multi-level category learning is also different from unsupervised methods for inducing hierarchies, such as hierarchical clustering (Duda et al., 2000; Heller and Ghahramani, 2005), structure learning (Kemp and Tenenbaum, 2009), learning ontologies (Kemp et al., 2006), or learning hierarchies (Roy et al., 2007; Blei et al., 2010). These unsupervised methods find a way to organize a set of objects into a hierarchical structure, but do so on the basis of the similarity of the objects, rather than using the category labels of those objects.

In this chapter, we investigate multi-level category learning in both artificial and natural systems.¹ First, we propose a novel method of learning and representing categories which are organized in taxonomic systems. Our model is a nonparametric Bayesian statistical model which we call the *tree-HDP*. We demonstrate that this model can recover simple taxonomies from just labeled examples of category members. We then turn to natural systems, conducting an experiment studying the performance of human learners in a similar task. A comparison of the model with the experiment results shows that the *tree-HDP* is able to do just as well—or better than—human learners.

4.1 Multi-level categories and the tree-HDP

The multi-level category learning problem reduces to being provided with a set of observations drawn from different categories, together with the information that these categories form a taxonomy, and estimating the probability distribution associated with each category and the structure of the taxonomy. In this section, we introduce the *tree-HDP* model, describe an efficient inference algorithm, and demonstrate that it can be used to solve this problem.

4.1.1 The tree-HDP model

The *tree-HDP* is a generalization of the HDP, described above. In the typical formulation of the HDP, the latent structure G_j of each category j is a draw from a Dirichlet process (DP) with a base measure G_0 that is shared between all the categories. In turn, G_0 is a draw from a higher-level DP with base measure H , a hyperparameter chosen by the modeler. Although the HDP is typically used to model collections of categories arranged in a flat hierarchy, the same statistical definitions can be recursively applied to multiple levels of hierarchy. That is, instead of all the categories inheriting from G_0 , some of them can inherit from others.

In practice, each draw from a DP yields a refinement or specialization of its base measure. In the flat HDP, this means that each category is a specialization of the global

¹This chapter is based on work done in cooperation with Thomas L. Griffiths, so I use plural first-person pronouns throughout this chapter.

base measure H , which is typically chosen to give broad, flat coverage over a wide range of parameters. Although the categories will exhibit random fluctuations in their degree of specialization, the use of a flat hierarchy means that each one is, a priori, at the same level of refinement. By contrast, if we push some of the categories down into deeper levels of the tree, they become specializations of their respective parent categories. This is the mechanism that the tree-HDP uses to model taxonomy systems. Intuitively, the tree structure of the HDP is intended to mirror the true hierarchical relationships between the categories.

Formally, we relax the assumption that the random measures G_j are drawn from a Dirichlet process with a common base measure G_0 . Instead, we allow the categories to form any valid tree structure with G_0 as the root. We introduce a new set of random variables $\boldsymbol{\tau} = \{\tau_1, \dots, \tau_J\}$ to describe the tree structure, with τ_j denoting the index of category j 's parent. If G_j is a child of G_0 , then $\tau_j = 0$, and if it is a child of some other category $G_{j'}$, then $\tau_j = j'$. We restrict $\boldsymbol{\tau}$ to form a valid tree structure, i.e., cycles are not allowed.

To specify the full Bayesian probability model of the tree-HDP, it is necessary to choose a prior distribution for the random variables $\boldsymbol{\tau}$. Since the number of nodes is fixed, there are only a finite number of possible tree structures. Any discrete distribution over these tree structures is valid; in this chapter, we use a uniform distribution over all trees in order to simplify the inference and reveal the model's underlying strengths and weaknesses. By performing Bayesian inference on the $\boldsymbol{\tau}$ variables along with the other hidden parameters of the HDP, we can infer the posterior distribution over taxonomy structures for any set of observed data.

The tree-HDP takes a different strategy than previous work for combining hierarchical structure with nonparametric Bayesian models, such as the nested CRP (Blei et al., 2010). In the nested CRP, objects are associated with paths in an infinitely deep, infinitely wide tree, and overlapping paths represent similarities between objects at multiple levels of abstraction. In contrast, in the tree-HDP, objects are associated with nodes in a finite tree, and edges represent subset relations between categories. This latter strategy is what makes the tree-HDP natural for modeling multi-level category learning, where an object can have high probability under distributions at multiple levels of the tree. More recently, Adams et al. (2011) have proposed a tree-structured stick-breaking prior for modeling hierarchical data which is very similar to our approach.

4.1.2 Inference in the tree-HDP

We now give a brief review of the inference procedure described by Teh et al. (2006) for the flat HDP and describe the steps necessary to extend the algorithm for the tree-HDP. Let z_{ji} denote the mixture component associated with x_{ji} , the i th observation from category j . Let m_{jk} denote the number of tables in category j assigned to mixture component k . The weight of mixture component k is denoted by β_{0k} in the global measure G_0 and by β_{jk} in the measure G_j of category j . Note that this differs from the notation of Teh et al. (2006), where the global component weights are called β_k and the category-specific weights are called π_{jk} .

The index of the parent of category j is given by τ_j . We relax the assumption that all the categories share a common concentration parameter α_0 ; instead, the concentration parameter for category j is denoted α_j , and the concentration parameter for the global measure G_0 is denoted α_0 instead of γ .

We use the ‘‘posterior sampling by direct assignment’’ method of Gibbs sampling inference, described in Section 5.3 of Teh et al. (2006). In this method, MCMC inference is performed over the variables $\mathbf{z} = \{z_{ji}\}$, $\mathbf{m} = \{m_{jk}\}$, and $\boldsymbol{\beta}_0 = \{\beta_{0k}\}$. In the flat HDP, the β_{jk} variables can be integrated out because the categories always occupy the leaves of the tree. However, in the tree-HDP, because categories can have other categories as children, the β_j variables must be explicitly represented and sampled in the MCMC algorithm.

For notational convenience, we also define the following variables. Let n_{jk} denote the number of observations from category j assigned to mixture component k . Let $v_{jk} = \sum_{j':\tau_{j'}=j} m_{j'k}$ denote the number of tables among the children of category j that are assigned to mixture component k . Each such table constitutes a ‘‘virtual’’ observation from mixture component k in category j , and these counts are necessary for computing many of the quantities used in the inference algorithm. These counts do not appear in the standard HDP formulation because categories do not have other categories as children, making $v_{jk} = 0$ for all categories.

The Gibbs sampling equations are as follows (Teh et al., 2006). To sample z_{ji} , the cluster assignment of observation i from category j , we have

$$P(z_{ji} = k | \mathbf{z}^{-ji}, \mathbf{m}, \boldsymbol{\beta}, \boldsymbol{\tau}) = \begin{cases} (n_{jk}^{-ji} + v_{jk} + \alpha_j \beta_{\tau_j k}) f_k^{-x_{ji}}(x_{ji}) & \text{for old } k, \\ \alpha_j \beta_{\tau_j u} f_{k^{\text{new}}}^{-x_{ji}}(x_{ji}) & \text{for } k = k^{\text{new}}. \end{cases} \quad (4.1)$$

To sample m_{jk} , the number of tables in category j assigned to mixture component k , we have

$$P(m_{jk} = m | \mathbf{z}, \mathbf{m}^{-jk}, \boldsymbol{\beta}, \boldsymbol{\tau}) = \frac{\Gamma(\alpha_j \beta_{\tau_j k})}{\Gamma(\alpha_j \beta_{\tau_j k} + n_{jk} + v_{jk})} s(n_{jk} + v_{jk}, m) (\alpha_j \beta_{\tau_j k})^m, \quad (4.2)$$

where $s(n, m)$ are unsigned Stirling numbers of the first kind. For $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_j$, we have

$$(\beta_{01} \dots \beta_{0K}, \beta_{0u}) | \mathbf{z}, \mathbf{m}, \boldsymbol{\beta}^{-0}, \boldsymbol{\tau} \sim \text{Dir}(v_{01} \dots v_{0K}, \alpha_0) \quad (4.3)$$

and

$$(\beta_{j1} \dots \beta_{jK}, \beta_{ju}) | \mathbf{z}, \mathbf{m}, \boldsymbol{\beta}^{-j}, \boldsymbol{\tau} \sim \text{Dir}(p_{j1} \dots p_{jK}, \alpha_j \beta_{\tau_j u}), \quad (4.4)$$

where ‘‘Dir’’ is the Dirichlet distribution and p_{jk} is notational shorthand for $n_{jk} + v_{jk} + \alpha_j \beta_{\tau_j k}$. Finally, for sampling the new $\boldsymbol{\tau}$ variables, we have

$$P(\tau_j = t | \mathbf{z}, \mathbf{m}, \boldsymbol{\beta}, \boldsymbol{\tau}^{-j}) \propto P(\tau_j = t | \boldsymbol{\tau}^{-j}) P(\mathbf{z}, \mathbf{m}, \boldsymbol{\beta} | \tau_j = t, \boldsymbol{\tau}^{-j}). \quad (4.5)$$

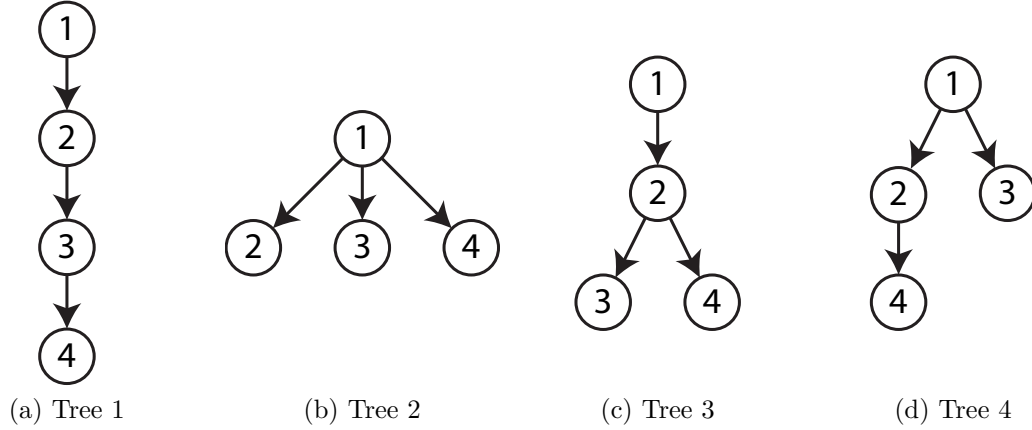


Figure 4.1: Tree structures used to create the simulated data.

Because we use a uniform prior, we have $P(\tau_j = t | \boldsymbol{\tau}^{-j}) \propto 1$, so

$$P(\tau_j = t | \mathbf{z}, \mathbf{m}, \boldsymbol{\beta}, \boldsymbol{\tau}^{-j}) \propto P(\mathbf{z}, \mathbf{m}, \boldsymbol{\beta} | \tau_j = t, \boldsymbol{\tau}^{-j}) \quad (4.6)$$

$$= P(\boldsymbol{\beta}_0) \prod_j P(\boldsymbol{\beta}_j | \boldsymbol{\beta}_{\tau_j}) P(\mathbf{z}_j | \boldsymbol{\beta}_{\tau_j}) P(\mathbf{m}_j | \boldsymbol{\beta}_{\tau_j}, \mathbf{z}_j). \quad (4.7)$$

Now we break down each term in this product:

$$\boldsymbol{\beta}_j | \boldsymbol{\beta}_{\tau_j} \sim \text{Dir}(\alpha_j \beta_{\tau_j 1}, \dots, \alpha_j \beta_{\tau_j K}, \alpha_j \beta_{\tau_j u}), \quad (4.8)$$

$$P(\mathbf{z}_j | \boldsymbol{\beta}_{\tau_j}) = \prod_k \frac{\Gamma(n_{jk} + \alpha_j \beta_{\tau_j k})}{\Gamma(\alpha_j \beta_{\tau_j k})}, \text{ and} \quad (4.9)$$

$$P(\mathbf{m}_j | \boldsymbol{\beta}_{\tau_j}, \mathbf{z}_j) = \prod_k \frac{\Gamma(\alpha_j \beta_{\tau_j k}) s(n_{jk}, m_{jk}) (\alpha_j \beta_{\tau_j k})^{m_{jk}}}{\Gamma(\alpha_j \beta_{\tau_j k} + n_{jk})}. \quad (4.10)$$

Combining these terms and dropping out the terms that do not depend on t , we have

$$P(\tau_j = t | \mathbf{z}, \mathbf{m}, \boldsymbol{\beta}, \boldsymbol{\tau}^{-j}) \propto P(\boldsymbol{\beta}_j | \boldsymbol{\beta}_t) P(\mathbf{z}_j | \boldsymbol{\beta}_t) P(\mathbf{m}_j | \boldsymbol{\beta}_t, \mathbf{z}_j) \quad (4.11)$$

$$\propto (\beta_{ju})^{\alpha_j \beta_{tu}} \prod_k (\beta_{jk})^{\alpha_j \beta_{tk}} (\alpha_j \beta_{tk})^{m_{jk}}. \quad (4.12)$$

Since τ_j can only take on a finite number of values, we can compute the normalization factor by summing over t .

4.1.3 Evaluation on recovering simulated taxonomies

To verify the ability of the tree-HDP to reconstruct taxonomies, we used it to infer some small taxonomies with simulated data. We built four different hierarchies of four categories

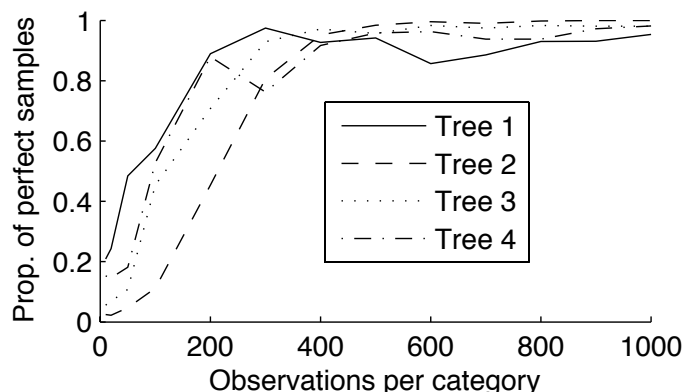


Figure 4.2: Inference results for the tree-HDP model on the simulated datasets. Curves show the proportion of MCMC samples which perfectly reconstruct the correct tree. The four tree structures are depicted in Figure 4.1.

each (see Figure 4.1) and used the HDP generative model conditioned on these taxonomy structures to sample 1000 observations from each category. Each category was modeled as mixture of Gaussian distributions with two independent dimensions. To make inference more tractable, we used a conjugate base measure (Normal-scaled inverse gamma) on each dimension, which had hyperparameters of $\lambda = 0$, $\nu = 0.01$, $\alpha = 3$, and $\beta = 1$. λ and ν control the prior distribution over the mixture components' locations, while α and β control the prior distribution over the mixture components' variance parameters. The Dirichlet process concentration parameter α_j was set to 10 for every category.

For each of the four taxonomies, we ran the tree-HDP inference algorithm on the generated data, where the number of observations given to the model ranged from 10 to 1000. The base measure hyperparameters were matched to those used to generate the data. The Gibbs sampling procedure was run for 51,000 iterations in total, with samples being taken every 10 iterations after a burn-in period of 1000 iterations. The tree structure was initialized to a flat hierarchy (all τ_j variables set to 0), and sampling of the τ_j variables was not performed until halfway through the burn-in period to allow the model to find good clusterings of the observations before constraining the probabilities of the clusterings by committing to a deep tree structure.

The results of the inference are summarized in Figure 4.2. The model's performance was similar for all four trees. The proportion of samples which perfectly reconstructed the correct tree rose from less than 20% with 10 observations per category, to close to 100% with 400 observations per category. These results show that for these data, the method is consistent: with enough observations, it converges to the correct hidden taxonomy structure.

4.2 Experiment 1: Comparing to human learners

In the previous section, we evaluated the performance of the tree-HDP on simulated data. Since we know that people are good at solving this problem, we were interested in comparing the model’s performance to that of human learners. We conducted an experiment where the model and human learners performed a multi-level category learning task using a taxonomy structure was designed to be reminiscent of a real-world hierarchy.

4.2.1 Method

4.2.1.1 Participants

We recruited 95 participants from Amazon Mechanical Turk who received \$1.00 as compensation. In addition, 95 undergraduate students participated for course credit. No significant differences were found between the two participant pools, so their data were pooled for analysis.

4.2.1.2 Stimuli

We constructed an artificial taxonomy of 14 categories, depicted in Figure 4.3. The categories were composed of a total of 8 clusters of visual stimuli from (Sanborn et al., 2009). The appearance of each stimulus is controlled by six continuous-valued parameters, and each cluster was defined by a multivariate Gaussian distribution over the values of these six parameters. Each of the eight categories at the leaves of the trees contained observations from only a single cluster; each of the four categories in the middle level contained observations from the two clusters of its descendants, and each of the two top-level categories contained observations from the four clusters below it.

4.2.1.3 Procedure

The experiment was completed entirely online. The participants first read a short introduction explaining that many categories in the real world are hierarchically organized, including examples with some common categories. The remainder of the experiment was divided into a training session, in which participants were taught to discriminate between each of the 14 categories of stimuli, and a test session, in which they were asked to reconstruct the category hierarchy that the 14 categories belonged to. The training session was divided into a number of blocks, each containing a set of observation trials followed by a set of discrimination trials. In the set of observation trials within a training block, participants were shown four examples of each category, for a total of 56 observations. For each category, the examples were equally distributed among the category’s clusters; so, for categories consisting of one, two, or four clusters, participants observed, respectively, four, two, or one stimuli per cluster. These observations were grouped on the computer screen according to their category

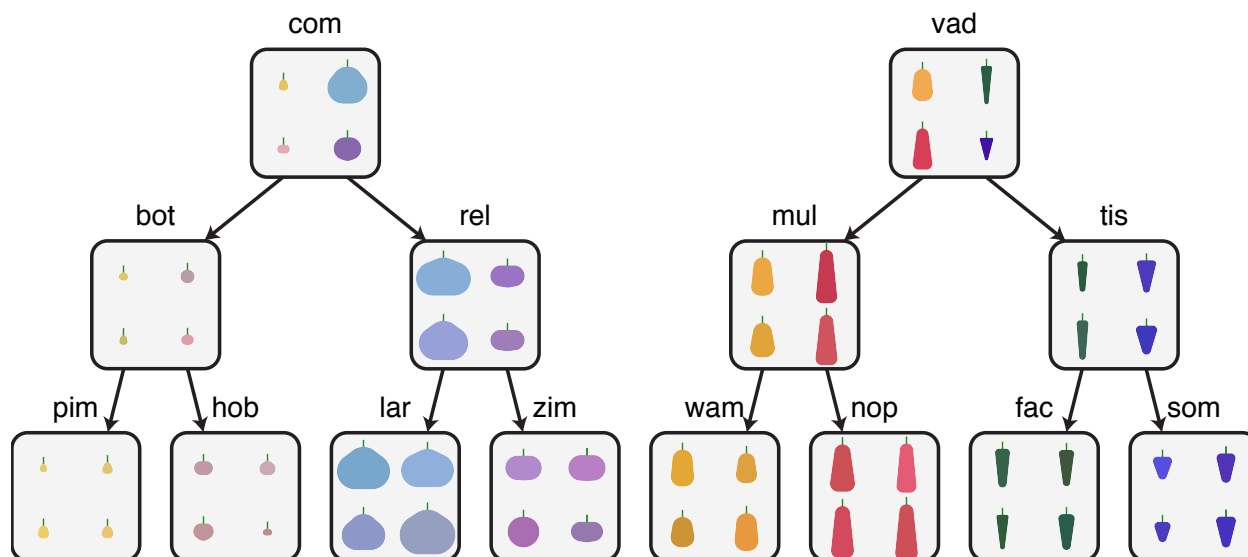


Figure 4.3: The category hierarchy used in Experiment 1.

labels, and they remained on the screen for the duration of the experiment. A screenshot of an observation trial is shown in Figure 4.4. The observation trials were randomly ordered within each block.

After completing a set of 56 observation trials, participants continued to a set of 28 discrimination trials. In each discrimination trial, a stimulus was presented on the screen along with an attached category label. The participants were asked whether or not the displayed stimulus actually belonged to the category. For each of the 14 categories, there was one trial where that category label was shown with a stimulus that belonged to the category (meaning that the correct answer was “Yes”) and one trial where that category label was shown with a stimulus that belonged to the category’s sibling (meaning that the correct answer was “No”), for a total of 28 trials. A screenshot of a discrimination trial is shown in Figure 4.5. The discrimination trials were randomly ordered within each block. Participants’ responses were recorded, but corrective feedback was not given during these trials. If participants answered at least 26 of the 28 discrimination trials correctly, they proceeded to the test session; otherwise they repeated another block of 56 observation trials and 28 discrimination trials.

Once the participants reached the performance criterion by correctly answering 26 out of 28 discrimination trials, they continued on from the training session to the test session. In the test session, participants were asked to reconstruct the taxonomy corresponding to the 14 categories they had just learned. Labels containing the names of the 14 categories were laid out in a random configuration on the screen. Participants could drag and drop the labels into whatever configuration they desired, and they could draw arrows between the category labels, meant to indicate parent-child hierarchical relationships, by clicking on

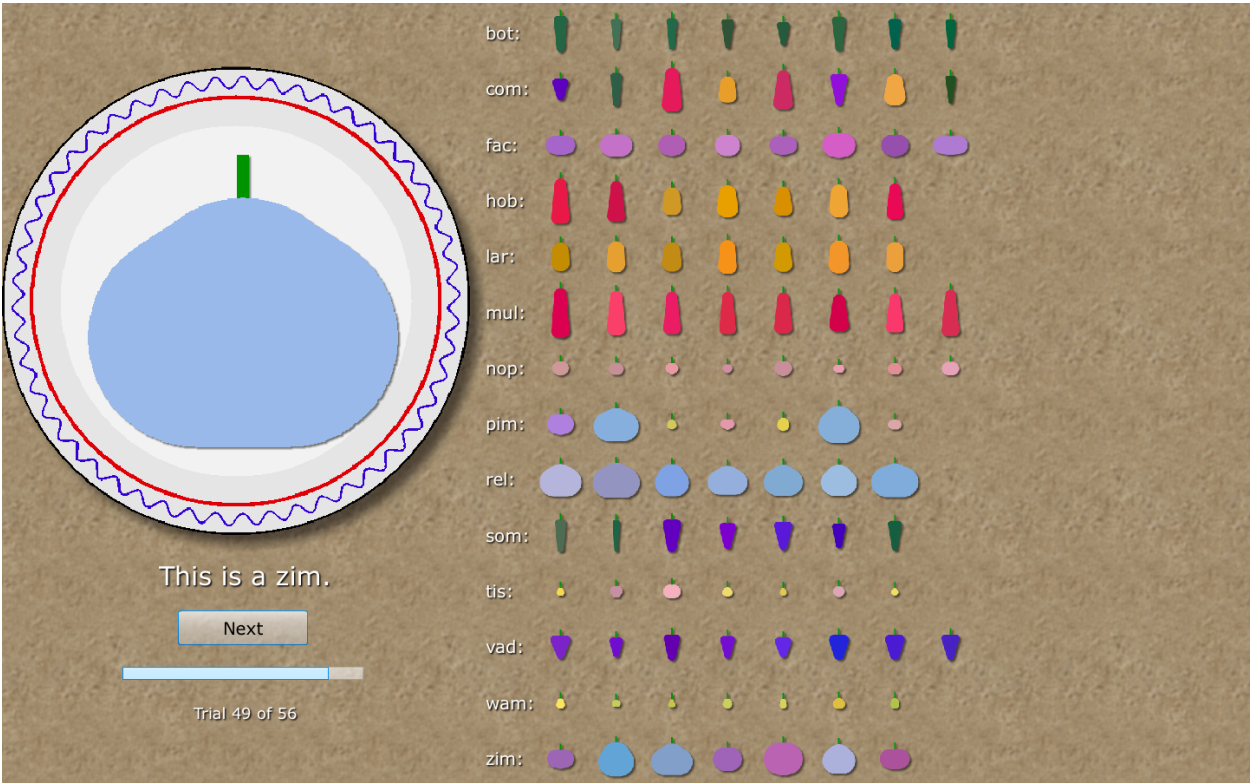


Figure 4.4: A screenshot of Experiment 1 during an observation trial in the training session.

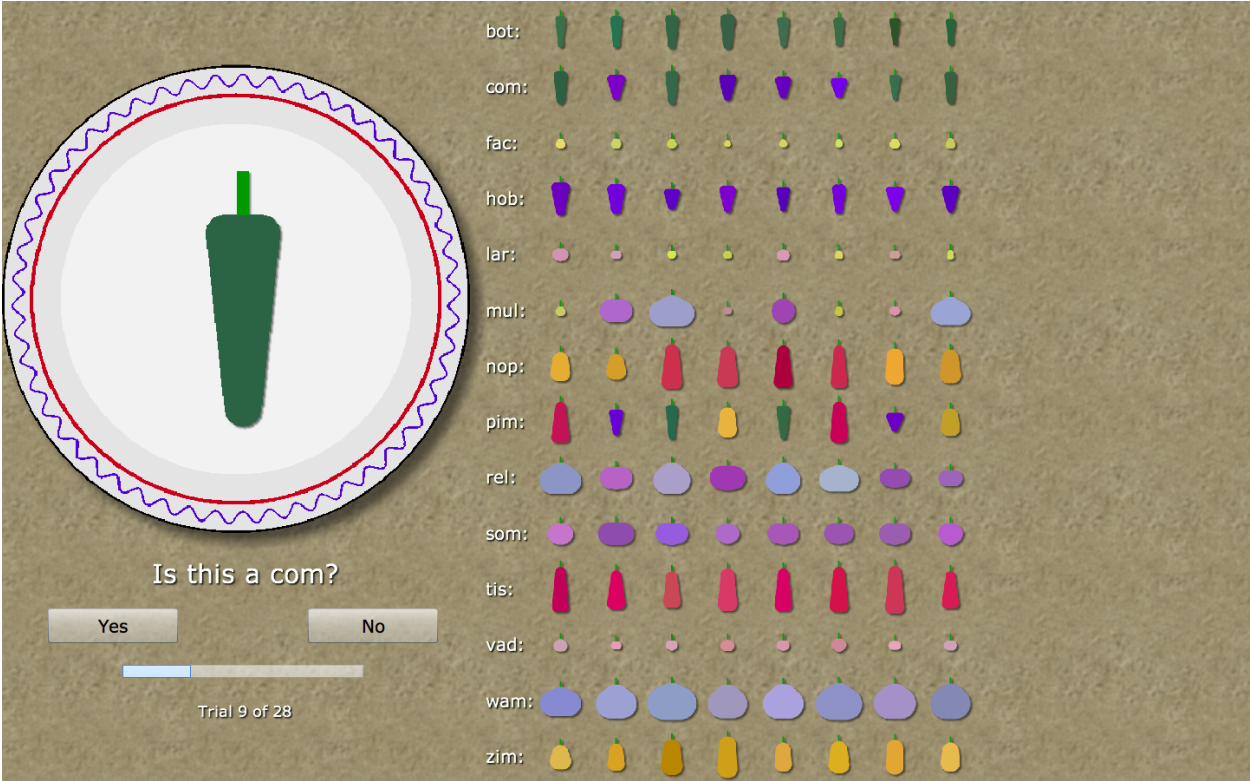


Figure 4.5: A screenshot of Experiment 1 during a discrimination trial in the training session.

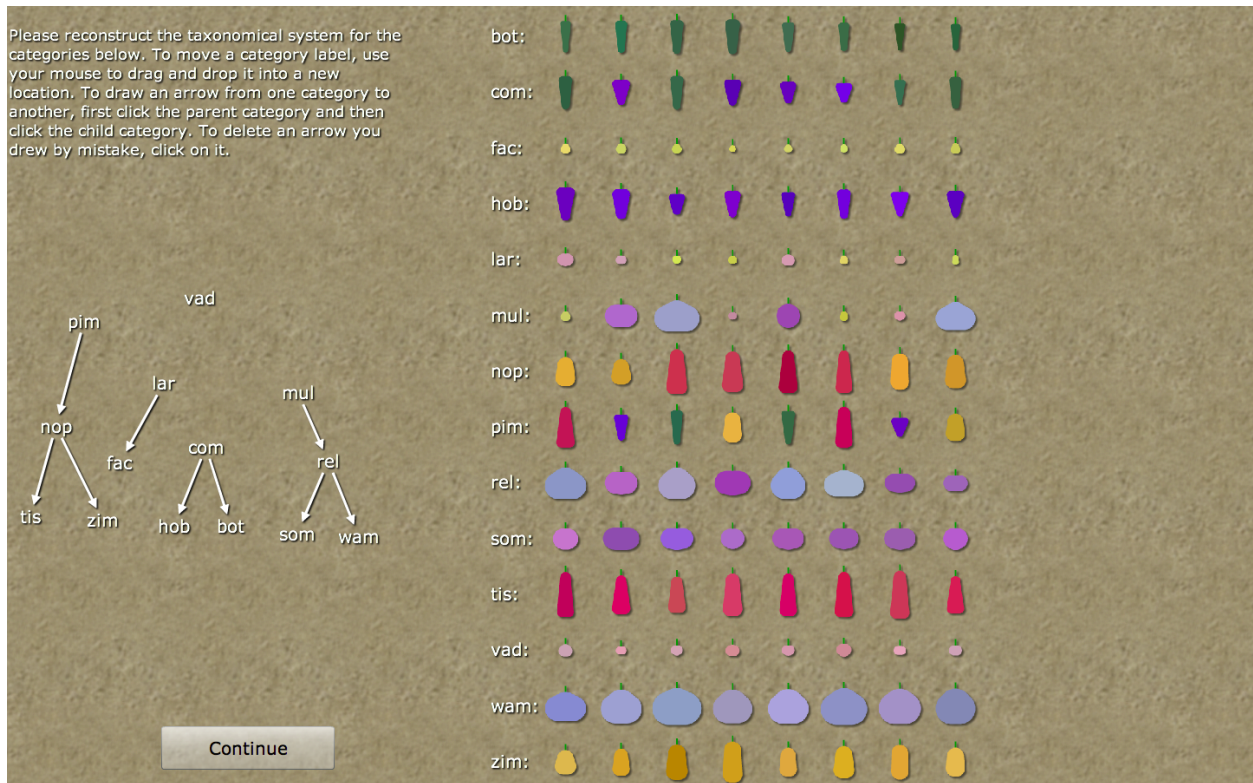


Figure 4.6: A screenshot of Experiment 1 during the test session.

the two endpoints of the arrow they wished to create. Accidentally-created arrows could be deleted by clicking on them. A screenshot of the test session is shown in Figure 4.6. Participants first completed this task with a taxonomy of fruit and vegetable categories to confirm that they properly understood the instructions and the user interface for constructing the taxonomy.

4.2.2 Results

On average, the participants completed 1.3 blocks in the training session before achieving the performance criterion necessary to move on to the test session. Out of the 190 participants, 78 (41.1%) perfectly reconstructed the correct taxonomy structure, and the average number of taxonomy edges in participants' responses was 11.2. The total number of incorrect edges among all participants, including both incorrectly chosen edges and missing correct edges, was 1401, for an average of 7.4 errors per participant.

Table 4.1 lists the edges which were chosen statistically significantly often, with the number and percentage of participants who chose each one. Boldfaced rows indicate the 12 edges that are present in the correct taxonomy; the other edges were chosen mistakenly.

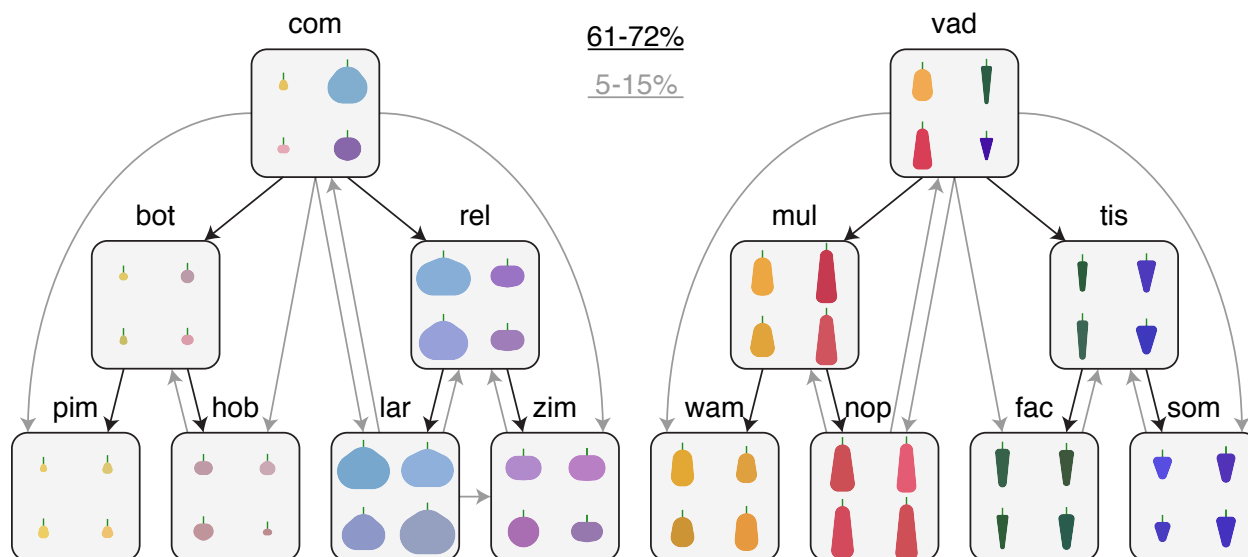


Figure 4.7: The aggregate results for Experiment 1 for the human learners. Black arrows were chosen by a majority of participants (with percentages ranging between 61–72%), and gray arrows were chosen by a small but statistically significant number of participants (with percentages ranging between 5–15%).

Figure 4.7 shows the aggregate taxonomy resulting from averaging over all the participants. Black arrows, corresponding to edges in the correct taxonomy, were chosen by 61–72% of the participants, while gray arrows were chosen by 5–15% of the participants. The gray arrows are those that are not in the correct taxonomy but appeared with statistically significant frequency.²

These results show that human learners are able to accurately reconstruct taxonomy structures from a limited number of examples. The types of errors made are very systematic and give insight into the mental representations that people use to solve the problem. With one exception, all of the incorrect edges either point in the wrong direction (towards a category’s ancestor) or point in the right direction but skip a level. The first type of error can be explained by people not adequately understanding the meaning of the arrows they were drawing; perhaps they had accidentally reversed the meaning of parent-child relationships. The second type of error shows that people sometimes do not classify categories at the lowest possible level, but occasionally produce taxonomies which are flatter than they could be otherwise.

² $p < 0.05$ according to an exact binomial test with success probability equal to 6.1%, the overall average frequency of edges across all results.

Table 4.1: Edge frequencies from Experiment 1

From category	To category	Frequency	Percentage
mul	wam	136	71.6%
vad	mul	130	68.4%
rel	lar	127	66.8%
rel	zim	127	66.8%
tis	som	126	66.3%
com	bot	125	65.8%
bot	pim	125	65.8%
mul	nop	124	65.3%
tis	fac	124	65.3%
vad	tis	124	65.3%
bot	hob	118	62.1%
com	rel	115	60.5%
vad	som	28	14.7%
vad	fac	27	14.2%
nop	mul	27	14.2%
com	hob	26	13.7%
vad	nop	26	13.7%
com	pim	26	13.7%
lar	rel	25	13.2%
fac	tis	22	11.6%
hob	bot	22	11.6%
com	zim	21	11.1%
com	lar	19	10.0%
vad	wam	18	9.5%
lar	com	17	8.9%
som	tis	15	7.9%
zim	rel	13	6.8%
lar	zim	13	6.8%
nop	vad	10	5.3%

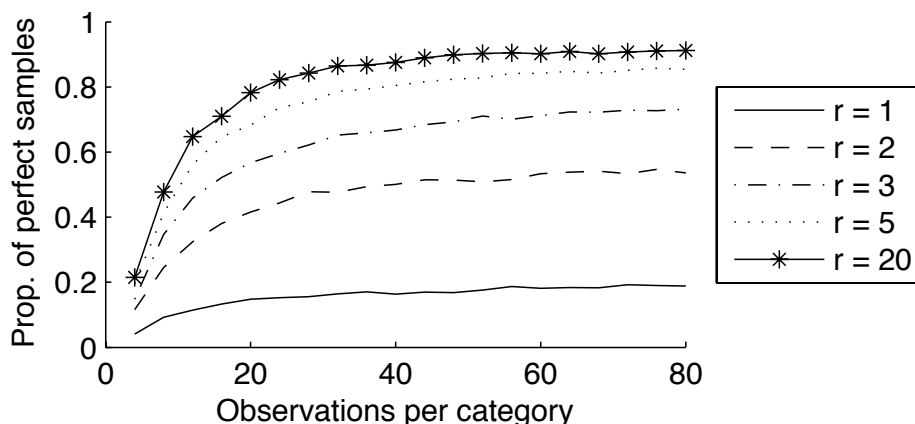


Figure 4.8: Performance of the tree-HDP on the experimental data. Results are shown for five different settings of r , the inverse temperature of the sampling distribution over tree structures. Each curve shows the proportion of samples which perfectly reconstruct the correct tree structure.

4.2.3 Modeling

We simulated the tree-HDP model on the same task that the human learners completed. In order to explore the variability of the model’s performance, we trained it on a range of 4–80 observations per category. The human learners observed 5.3 examples per category, on average. As with the simulated data, the model represented each category as a mixture of Gaussian distributions with six independent dimensions. The hyperparameters of the conjugate base measure (again, Normal-scaled inverse gamma) were fit to the training data, with parameters $\lambda = 0$, $\nu = 0.01$, $\alpha = 1.6$, and $\beta = 6.3$. The Dirichlet process concentration parameters were inferred from the data separately for each category, using a $\text{Gamma}(1, 0.01)$ prior distribution. To focus the posterior more on high-probability trees, we ran versions of the model at “temperatures” of $1/r$ for values of r between 1 and 20, corresponding to raising the Gibbs sampling distribution for τ_j in the MCMC algorithm to the power of r .

The model results are shown in Figure 4.8. In general, more observations per category and higher values of r both led to better performance. The model’s performance covers a wide range of values (4–91%) depending on these two parameters. The aggregated samples from one version of the model (with 8 observations per category and $r = 3$) is shown in Figure 4.9. In general, the model very accurately reconstructed the taxonomy, and interestingly, the mistakes it makes are very similar to those of the human learners. Ignoring the “backwards” edges produced by the human learners, there is only a single difference in the significant mistakes made by the model and the people: the edge from “lar” to “zim”. The correlation between the edge frequencies in people’s reconstructed hierarchies and the model’s samples was 0.988.

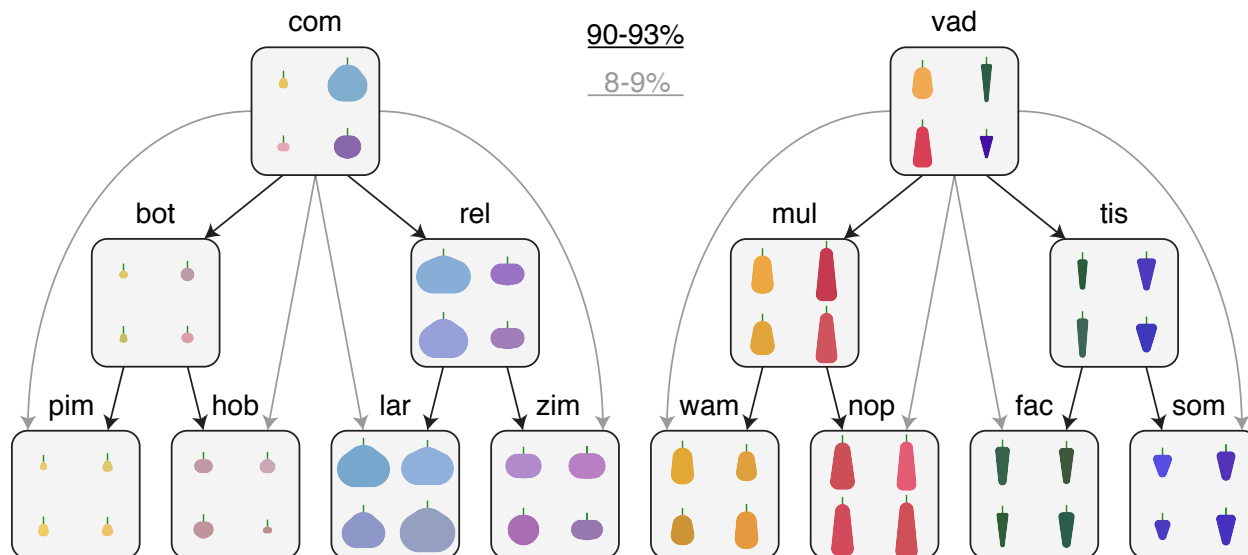


Figure 4.9: Results for Experiment 1 for the tree-HDP model with 8 observations per category and an inverse temperature parameter of $r = 3$. Black arrows were chosen by a majority of model samples, and gray arrows were chosen by a small but statistically significant number of samples.

4.3 Experiment 2: Testing implicit knowledge and inferences

Experiment 1 demonstrated people’s ability to perform multi-level category learning by explicitly reconstructing the taxonomy structure underlying a set of hierarchically related categories. However, a more realistic use of one’s knowledge about multi-level categories is making inferences or generalizations. Hence, in this experiment, we test people’s knowledge about the taxonomy structure in a more implicit way, using inferences about pairs of related categories.

4.3.1 Method

4.3.1.1 Participants

The participants consisted of 43 undergraduate students from the University of California, Berkeley, who received course credit.

4.3.1.2 Stimuli

The stimuli were identical to those from Experiment 1.

4.3.1.3 Procedure

The procedure was identical to that of Experiment 1, with the exception of the format of the test session. Rather than testing participants' explicit knowledge of the category taxonomy by asking them to reconstruct the hierarchical system, we instead tested their implicit knowledge by asking questions of the form "Is every *hob* a *com*?" (with different category names appearing in place of *hob* and *com*). These questions test the type of knowledge used in the well-known deductive arguments known as syllogisms: for example, "Object X is a hob. Every hob is a com. Therefore, object X is a com." With respect to a participant's inferred category hierarchy (or with respect to the correct hierarchy), the correct answer to this question is "Yes" when the first category is a descendent (either a child or a grandchild) of the second, and "No" otherwise. An example trial in the test session is shown in Figure 4.10. The test session consisted of 48 trials of this format. There were 12 trials corresponding to each of the 12 direct edges in the correct hierarchy and 8 trials corresponding to each of the 8 indirect ancestor/descendent relationships in the correct hierarchy; the correct answer was "Yes" in each of these trials. For each of these 20 trials, there was an additional trial in which the order of the category names was reversed, so that the correct answer was "No". Finally, there were 8 additional trials in which the pair of categories in the question were selected randomly from all the pairs that hadn't been chosen to appear yet. These 48 trials were randomly ordered for each participant.

This procedure allows us to make different measurements from those in Experiment 1, focusing more on implicit knowledge than explicit knowledge. Instead of asking participants to directly draw edges between categories to reconstruct the category taxonomy, this experiment tests the indirect consequences of the inferred hierarchy. If a participant learns that category A is a subtype of category B, even if that knowledge is not able to be recalled explicitly, the participant may still be able to recognize that every member of category B is also a member of category A.

4.3.2 Results

On average, participants completed 1.78 blocks in the training session before achieving the performance criterion necessary to move on to the test session. Out of the 43 participants, 9 (20.9%) correctly answered all 48 trials in the test session, and the average number of "Yes" answers given was 17.84 out of the 48 test trials. The results are summarized in Table 4.2. Each row lists the number and percentage of participants who answered that the first category is an ancestor of the second category – that is, those who answered "Yes" to the question "Is every A a B?" for the appropriate categories A and B. Each of the 20 ancestor relationships from the correct taxonomy structure are listed in the table; none of the other pairs of categories were answered in the affirmative by a statistically significant



Figure 4.10: A screenshot of Experiment 2 during the test session.

Table 4.2: Ancestor frequencies from Experiment 2

Ancestor category	Descendent category	Frequency	Percentage
rel	zim	39	90.7%
bot	hob	39	90.7%
com	hob	38	88.4%
mul	nop	37	86.0%
tis	som	36	83.7%
vad	fac	36	83.7%
com	pim	36	83.7%
vad	nop	35	81.4%
com	lar	35	81.4%
tis	fac	35	81.4%
mul	wam	35	81.4%
rel	lar	34	79.1%
vad	wam	33	76.7%
vad	mul	33	76.7%
com	zim	32	74.4%
com	bot	32	74.4%
bot	pim	31	72.1%
com	rel	29	67.4%
vad	tis	28	65.1%
vad	som	26	60.5%

number of participants.³

As in Experiment 1, we also observed that a low percentage of the participants responded with answers that were consistent with the correct taxonomy structure, but “backwards”. In other words, they responded “Yes” to questions where the relationship was invalid, but would have been valid if the categories had switched places. These questions received “Yes” answers with frequencies between 4.7% and 18.6%, which is roughly the same range of frequencies with which backwards arrows appeared in the results from Experiment 1. However, since there were fewer participants in this experiment than there were in Experiment 1, these frequencies were not large enough to be statistically significant in this experiment.

³As before, statistical significance was determined by an exact binomial test, with the criterion of $p < 0.05$ and the probability of selection equal to 37.2%, the average frequency of “Yes” answers across all trials.

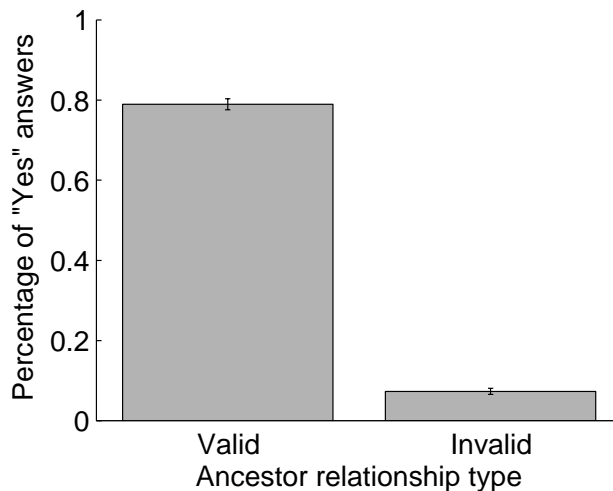


Figure 4.11: Results for Experiment 2. The bar graph shows the percentage of trials on which participants correctly or incorrectly answered “Yes” to questions in the test session of the form “Is every A a B?”, along with standard error bars for each estimated frequency.

4.4 Experiment 3: Learning unbalanced taxonomies

The results from Experiments 1 and 2 indicate that people can learn these relatively simple, balanced taxonomies, consistent with the Bayesian tree-HDP model. We now test more subtle predictions of the model, examining the inferences and errors that are made for different taxonomy structures. In this experiment, we investigate the ability of human learners and the tree-HDP model to infer category hierarchies that are less symmetrical than those in Experiments 1 and 2.

4.4.1 Method

4.4.1.1 Participants

The participants consisted of 55 undergraduate students from the University of California, Berkeley, who received course credit, and 185 workers from the website Amazon Mechanical Turk, who received a payment of \$1.00, for a total of 240 participants.

4.4.1.2 Stimuli

This experiment involved the same set of stimuli as the previous experiments, but it differed in the way the stimuli were arranged into taxonomy structures. The hierarchies to be learned were constructed by starting with the same 14-category hierarchy used in Experiment 1. From each of the two separate trees (rooted at `com` and `vad` in Figure 4.3), one of the two mid-level categories was chosen at random to be “pruned”, meaning that its two subcategories

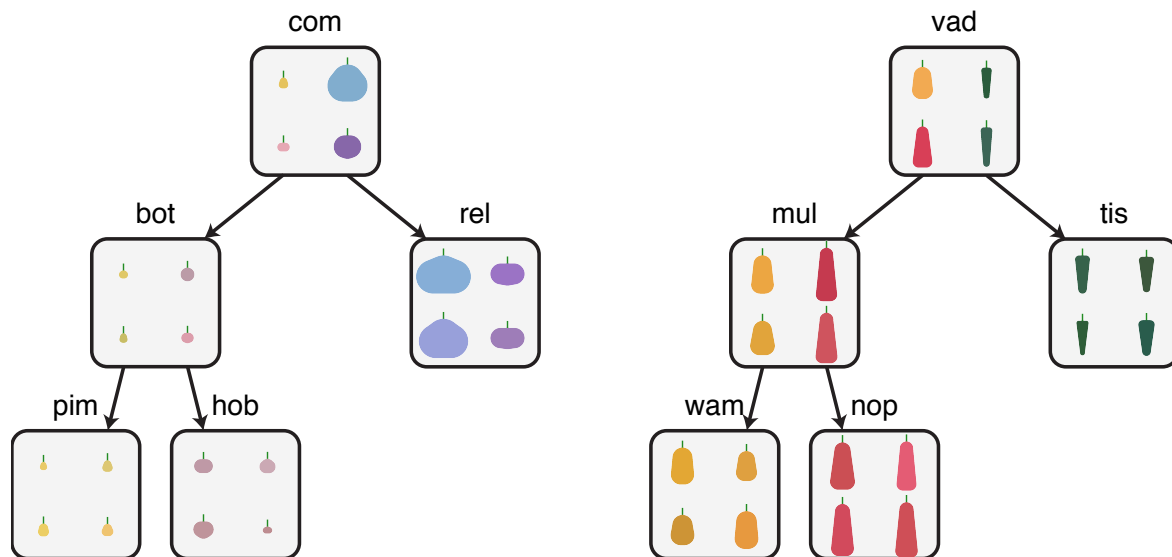


Figure 4.12: An example hierarchy used in Experiment 3. In this example, the categories labeled *com* and *vad* are “root” categories in their respective trees, *rel* and *tis* are “pruned” categories, *bot* and *mul* are “sibling” categories, and *pim*, *hob*, *wam*, and *nop* are “nephew” categories. *rel* is the “double pruned category” because it contains two clusters, and *tis* is the “single pruned category” because it contains one cluster.

were removed from the hierarchy. One of the pruned categories was randomly chosen to be the “double pruned category”, meaning that it inherited both of its subcategories’ clusters. The other, called the “single pruned category”, inherited just one randomly-chosen cluster from its two subcategories. This resulted in the construction of a 10-category hierarchy split into two disjoint trees of 5 categories each, as shown in Figure 4.12.

4.4.1.3 Procedure

The procedure of Experiment 3 was identical to that of Experiment 1. Because there were only 10 categories instead of 14, however, each training block contained 40 observation trials instead of 56 (with four observations per category, as in Experiment 1) and 20 discrimination trials instead of 28 (with two trials per category, as in Experiment 1). The performance criterion for completion of the training session was 18 correct answers out of the 20 discrimination trials.

4.4.2 Results

On average, participants completed 1.5 blocks in the training session before achieving the performance criterion necessary to move on to the test session. Of the 240 participants, 36 (24.6%) perfectly reconstructed the correct taxonomy structure, and the average number of

taxonomy edges in participants' responses was 7.2. The total number of errors among all participants, including both incorrectly chosen edges and missing correct edges, was 1463, for an average of 6.10 errors per participant.

Due to the random selection of the pruned categories (as described in Section 4.4.1.2 above), each participant was trained on a different category hierarchy, so their responses cannot be directly aggregated together for analysis. Although the categories played different structural roles in each participant's hierarchy, each participant learned a taxonomy that had the same overall structure, displayed by the example in Figure 4.12. In order to evaluate the inferences made by the human learners, we assign structural roles to the categories and aggregate across participants according to these roles. Each hierarchy contains two trees which both consist of a "root" category at the top level, a "pruned" category and a "sibling" category at the middle level, and two "nephew" categories at the bottom level, which are both subcategories of the "sibling" category. We can distinguish between the two trees according to the number of clusters in the pruned category. As described above, one pruned category is randomly selected to contain a single cluster (called the "single pruned category"); the tree containing this category is called the "single pruned tree", and the other is called the "double pruned tree". Having identified each category in a participant's hierarchy with a structural role, we measured the frequencies with which participants selected edges connecting categories of each role in each tree. For example, we counted the frequency with which the model selected edges from the root to the pruned category in each tree, and edges from the sibling to either of the nephews in each tree.

Among all possible edges in the hierarchy, the types that were selected statistically significantly often by the participants were root→pruned, root→sibling, sibling→nephew, and root→nephew edges.⁴ Note that root→nephew edges are not part of the correct hierarchy, but were selected by many participants nonetheless. Figure 4.13 shows the frequency with which participants selected edges of each type in each tree.

One of the purposes of conducting this experiment was to test the effect of manipulating the number of clusters in one of the categories of a taxonomy structure. Because participants learned hierarchies consisting of two trees which were identical except for the number of clusters in the pruned category, we can observe the effects of this difference by comparing which edges were chosen for the single pruned tree vs. the double pruned tree. By counting the cases where a participant selected more edges of a given type in one tree than in other tree, we can discover whether the number of clusters in the pruned category affects the participants' inferences about the underlying hierarchies.

For each of the four types of edges that were selected significantly often, we measured N_s , the number of participants who chosen more edges of that type in the single pruned tree than in the double pruned tree, and N_d , the number of participants who chose more edges

⁴These are the types of edges whose observed frequency of being chosen, averaged across all participants, were statistically significant according to an exact binomial test with $p < 0.05$ and probability of selection equal to 8.0%, which was the average frequency which which all possible edges in the hierarchy were chosen.

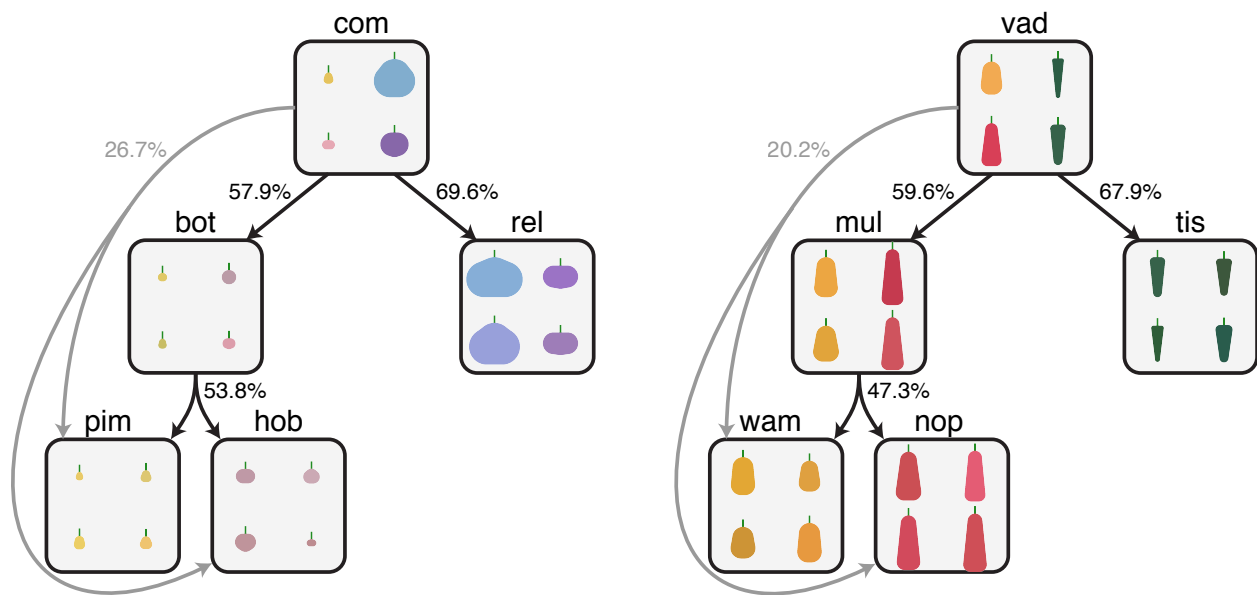


Figure 4.13: The aggregate results for Experiment 4 for the human learners. Black arrows indicate edge types that are part of the correct hierarchy, and gray arrows indicate edge types that are incorrect. Each edge type is labeled with the frequency with which it was chosen by participants. Since each participant learned a randomly-generated taxonomy structure, results were aggregated according to the structural roles played by each category, as described in Section 4.4.2. The particular hierarchy shown here is the same that appears in Figure 4.12.

Table 4.3: Results for Experiment 3 by edge type

Edge type	Single pruned edges	Double pruned edges	N_s	N_d	p -value
root→pruned	vad→tis	com→rel	19	23	0.32
root→sibling	vad→mul	com→bot	22	18	0.32
sibling→nephew	mul→wam, mul→nop	bot→pim, bot→hob	33	50	0.04
root→nephew	vad→wam, vad→nop	com→pim, com→hob	49	31	0.03

Note: Boldface type indicates the larger of N_s and N_d in each row. The p -values were calculated according to an exact binomial test of the null hypothesis that each edge type is equally likely to occur in the single and double pruned trees. The binomial tests are one-tailed because the model results described below predict a direction for effect.

of that type in the double pruned tree than in the single pruned tree. For the root→pruned and root→sibling edge types, there is only one possible edge of this type in each of the two trees. For these two edge types, N_s is the number of participants who chose that edge in the single pruned tree but did not choose it in the double pruned tree, and N_d is the number of participants who chose it in the double pruned tree but not in the single pruned tree. For the sibling→nephew and root→nephew edge types, there are two possible edges of that type in each tree, because there are two nephews in each tree. So for these two edge types, N_s is the number of participants who chose more edges of that type in the single pruned tree than in the double pruned tree, and correspondingly for N_d . Table 4.3 lists these frequencies for each edge type, along with example edges from the hierarchy in Figure 4.12 which are of each type. Also listed for each edge type is the p -value of an exact binomial test of the null hypothesis that the expected values of N_s and N_d are equal.

The number of clusters in the pruned category seems to have had an effect on each of the four edge types. Participants more often selected root→sibling and root→nephew edges in the single pruned tree, while they more often selected root→pruned and sibling→nephew edges in the double pruned tree. Two of the four edge types demonstrate statistically significant effects ($p < 0.05$).

4.4.3 Tree-HDP model predictions

The tree-HDP model was trained on labeled examples of the 10 categories in the hierarchy shown in Figure 4.12 and then used to reconstruct the taxonomy structure. As explained in Section 4.2.3, the model results can vary according to how many observations per category the model is given, as well as the inverse temperature parameter r of the model’s posterior distribution that is used during the Gibbs sampling procedure. We ran the model for a wide range of these two parameters: the number of observations per category ranged from 4 to 80 in increments of 4 (recall that participants received about $1.5 \times 4 = 6$ observations per category on average), and r took on the values 1, 2, 3, 4, 5, 10, 15, and 20, which

made a total of 160 variants of the model. We ran each of the 160 variants of the tree-HDP model using the same type of data given to the human learners. Rather than randomizing the hierarchy for each model, we always trained the model on the same hierarchy: the one shown in Figure 4.12.

Every single one of the 160 variants of the tree-HDP model displayed the same biases as the human learners displayed in Table 4.3: root→sibling and root→nephew edges were always more common in the single pruned tree than in the double pruned tree, while the opposite was true for root→pruned and sibling→nephew edges.

Although two of the effects found in the human data were not statistically significant on their own, when we consider that all four effects are consistent with the model predictions, we can pool them together to find a statistically significant combined effect. We combine the number of cases where a participant chose more edges of a particular edge type in the same tree (either single pruned or double pruned) that the model predicted, for a total of $23 + 22 + 50 + 49 = 144$ (the bold numbers in Table 4.3). Similarly, we combine the number of cases where a participant chose more edges of a particular edge type in the tree that *was not* predicted by the model, for a total of $19 + 18 + 33 + 31 = 101$. An exact binomial test of the null hypothesis that the human data are independent of the model predictions (that is, that these two totals are generated by a binomial distribution with a success parameter of 50%) yields a p -value of 0.002. Therefore, we find that the overall model predictions are consistent with the human data; both the human learners and the model display the same types of learning effects when learning unbalanced trees with differing numbers of clusters in one of the categories.

4.5 Conclusion

Learning the conceptual structures that characterize our world requires being able to induce relationships between categories from examples of their members. We have presented a non-parametric Bayesian model that can be used to solve the problem of multi-level category learning and shown that its performance is similar to that of humans. This analysis helps explain how it is possible for learners to induce taxonomies from only labeled observations, and provides a new tool for learning categories in contexts where the assumption that categories are independent of one another is invalid. In future work, we hope to extend this analysis to incorporate direct statements of the relations between categories, as might be provided in verbal instruction or found through text mining, and consider how our approach can be extended to more complex conceptual structures.

Chapter 5

Conclusions

The findings described in the previous three chapters support the following conclusions about human categorization.

1. Human categorization is characterized by flexible category representations.

In Chapter 2, I described an iterated learning experiment of categorization to reveal the inductive biases of human learners for four different stimulus sets. For separable stimuli, people preferred a small number of relatively simple category structures, while for integral stimuli, a larger number of more complex category structures were preferred. These results imply that people's inductive biases for categorization problems vary according to the stimuli being learned. This conclusion supports previous findings in psychology that neither prototype nor exemplar models always provide a better account of people's generalizations in category learning experiments. It seems that human categorization is characterized not by a single rigid representation or model, but rather it is flexible and adaptable. Therefore, we should strive to build models which can account for this versatility.

2. When learning new categories, people incorporate information about other categories.

Chapter 3 introduced the HDP as a unifying categorization model which makes a key prediction that people should transfer information from one category to others when learning about multiple categories. An experiment designed to test this prediction found that when making predictions about new categories, people are sensitive to information about other categories to which they have had previous exposure. This demonstrates that categories are not learned in isolation from other categories, as implicitly assumed by most psychological models. Instead, people are aware of a great variety of interrelated categories and use their knowledge about past categories to inform the inferences they make about new categories.

3. People are able to learn category taxonomies from only indirect evidence.

In Chapter 4, I defined the problem of *multi-level category learning*, in which a learner observes examples of categories at multiple hierarchical levels and must not only predict the category label of a new object, but must also reconstruct the category taxonomy. I introduced the tree-HDP model for solving this problem and experimentally demonstrated that both human learners and the tree-HDP are able to solve the multi-level category learning problem. In addition, the model seems to be able to explain not only the successful inferences that people make, but also most of their mistakes.

5.1 Limitations and future directions

While the approaches to categorization described in this dissertation are designed to extend the state of the art of psychological and statistical modeling, they are not without their own limitations. Some of these limitations are now discussed in turn, along with some ideas for future directions for this research.

5.1.1 Spatial representations

The HDP-based models presented in this dissertation assume a low-dimensional spatial representation for the input received by a category learner for each stimulus. Clearly, this is not a realistic assumption for the actual sensory input received by human learners. However, it may be a reasonable approximation of the output of human sensory processing systems for highly constrained sets of stimuli which vary in only a few unambiguous dimensions. For stimulus sets such as Shepard circles and rectangles, this assumption seems justifiable; however, for more complex stimulus sets with integral dimensions, it is not clear that there exists a single cognitive spatial representation with unambiguous dimensions. Furthermore, in situations where the stimuli are not limited to a small set of similar objects, learners may not know ahead of time the full extent of all the stimuli they will be presented with, so assuming a single, fixed spatial representation from the onset of learning seems problematic. Following this line of thought, a truly universal categorization model should be able to represent any set of objects in a common representational scheme; however, it is not evident that there even exists a common spatial representation for the wide variety of entities that people are able to categorize. For example, it would be quite challenging to find a natural spatial representation for a collection of objects consisting of a slice of pizza, a whiteboard marker, and the number seven.

Although most formal categorization models start with this basic assumption of a low-dimensional spatial representation for the stimuli, there have also been attempts to characterize human categorization using different representational schemes. One of the more successful of these approaches is based on representing stimuli with a list of features that they possess (see, for example, Tversky, 1977). These feature-based representations have

some unique strengths, such as being able to account for the asymmetry of human similarity judgments. It should be noted that feature-based representations are technically equivalent to a high-dimensional, binary spatial representation, where each dimension represents one distinct feature, and stimuli take on the value of 0 or 1 for each dimension to indicate whether or not they possess that feature. Nonetheless, as we have demonstrated that human categorization adopts flexible category representations, it is quite possible that people also adopt flexible representations for stimuli, using spatial representations at some times and feature representations at other times. Additionally, the basic ideas underlying the HDP-based models (that categories are composed of clusters of objects that are shared between categories in particular ways) can theoretically be extended from spatial stimulus representations to feature-based stimulus representations.

5.1.2 Clusters as the atomic unit

The Bayesian statistical models described in this dissertation all adopt their flexible category representations through the use of *clusters*. In a sense, for all of these HDP-based models, the cluster is the atomic representational unit of a category. These models learn categories by building up clusters of objects which can be arbitrarily composed together. While this ability is a major step forward from the static representational schemes of prototypes and exemplars, it is also somewhat limited itself. First, the fundamental motivation for the HDP and the tree-HDP is that sharing information between categories should facilitate learning. However, the only way in which these models share information is by using clusters to represent more than one category; this approach is only useful in situations where there are multiple categories which can be described in terms of overlapping clusters. Moreover, this representation uses only a single layer of abstraction (the clusters) to define categories. It is unlikely that this is the only way in which information about multiple categories can be learned and used. For example, it seems plausible that people are capable of learning that clusters are themselves composed of sub-clusters, and so on. Therefore, one a more plausible account might involve learning a multi-layered system of clusters to represent a category. In addition, shared clusters certainly aren't the only way in which categories can be related to each other. For example, Heller et al. (2010) argue that “categories of materials such as gold, ice, and wood all display a characteristic color while being relatively unconstrained as to the shapes and sizes that they take,” (p. 2) while “size is often constrained in [categories of artifacts] such as books and cars, while color can vary across a very wide range.” (p. 2) From this perspective, it is apparent that categories such as gold, ice, and wood are all related to each other not because of any shared category members, but because they each extend along the same dimensions (shape and size) and are constrained along the same dimensions (color and texture). Heller et al. (2010) propose a hierarchical Bayesian model that is somewhat related to the HDP but is designed to learn that categories tend to be organized along the same dimensions rather than that categories can share overlapping clusters. This is just one example of how one can take the approach of designing statistical models to learn

things about categories by observing members of other categories; this general strategy is not limited to the specific type of cluster-based learning performed by the HDP.

5.1.3 Cluster-based and rule-based representations

The cluster-based representations used in the models presented in this dissertation are all, in a sense, agglomerative. The HDP and related models build up category representations by combining individual observations into clusters and combining clusters into categories. This procedure inevitably leads to category representations which are combinations of somewhat amorphous collections of objects. It seems intuitive that human categorization is often based more on simple rule-based structures involving logical quantifications, more akin to first-order logical representations. Although these seem like distinct types of representation, there is a sense in which the representations used by the HDP and HDP-based models can be thought of as rule-based, as discussed in Section 3.8.2. Each cluster of objects is essentially a region of the multidimensional stimulus space, which is roughly a conjunction of constraints on each dimension. Each category is a combination of clusters, and thus can be described as a disjunctive combination of the conjunctive constraints embodied by its constituent clusters.

In addition to the interpretation of HDP models as probabilistic analogs of logical formulae in disjunctive normal form, other connections can be drawn between cluster-based and rule-based models. The tree-HDP is just one way that the flat category taxonomy in the HDP can be extended to account for richer relationships between the categories. In the tree-HDP, these relationships are restricted to pairs of subcategories and supercategories, which are instantiated mathematically by each child category using its parent's mixture distribution over clusters as the base distribution for its own Dirichlet process. The tree-HDP can be generalized even further, allowing for arrangements of categories into directed acyclic graphs (DAGs), which are hierarchical systems where categories can have multiple parents. In these more general graphical HDPs, each child category could use either the sum or the product of its parents' mixture distributions as the base distribution for its own Dirichlet process, leading to a system where categories act like either disjunctions or conjunctions (depending on the operation used) of their parents. These sums and products could even be intermixed, allowing categories to be represented as general logical combinations of multiple other categories. Such a model would be a slight extension of the tree-HDP model presented in this dissertation, and the same type of Gibbs sampling-based algorithm should be able to be used to perform inference. This would bring the HDP models of categorization much closer in spirit and expressibility to logical, rule-based models.

5.1.4 Bayesian inference algorithms

Most psychological models of categorization are *process models*; that is, they posit a particular set of representations and algorithms that are used to solve categorization problems. By contrast, the models proposed in this dissertation are at the computational level (Marr,

1982): they posit that people view categorization as a statistical problem and use Bayesian inference to identify the likely categories for a given stimulus. These models do not, however, make claims about the particular algorithm used to perform this inference. In machine learning, there is a clear distinction between *models* and *algorithms*; this distinction is roughly equivalent to the difference in psychology between computational-level models and algorithmic-level models. The HDP-based models presented in the previous chapters define probability distributions over the potential conclusions that can be drawn, but they do not presuppose that any particular algorithm is used to calculate these probabilities. If a computational-level model is to be taken as a serious account of human cognition, however, it must be able to be plausibly implemented in some way by the human mind. In general, performing Bayesian inference for complicated models like the HDP and the tree-HDP is intractable with modern computers: it cannot be done both exactly and within a reasonable amount of time. To work around this limitation, approximation algorithms must be used. In executing the HDP-based categorization models, I have used Gibbs sampling algorithms, but this choice was motivated purely out of considerations of convenience. By its nature, Gibbs sampling is a relatively slow, serial computation; therefore, it is not a very realistic proposal as an algorithm that might be used by human learners, whose inferences are made more quickly and with a brain that operates in a massively parallel fashion. However, alternatives to Gibbs sampling have been proposed as more realistic algorithmic accounts of how people might implement Bayesian inference. For example, particle filters (Doucet et al., 2001) operate in a way that is more congruent with what is known about the mechanics of the human brain; their computations are highly parallel and require less time. Particle filters have been proposed as an algorithmic-level account of how people might implement rational models of categorization such as the HDP (Sanborn et al., 2010a). Nonetheless, finding a convincing account of how such complicated Bayesian statistical models could be implemented by the human mind remains an elusive goal for computational cognitive psychologists.

5.2 Conclusion

In this dissertation, I have demonstrated that human categorization is characterized by flexible representations, that when learning new categories, people incorporate information about other categories, and that people are able to learn category taxonomies from only indirect evidence. I have proposed and tested a family of statistical models, based on the hierarchical Dirichlet process, that account for and explain these capabilities of human learning in categorization. Although this is far from a complete explanation of all the nuances and abilities of human categorization and there is still a long way to go before computational models match the flexibility and ingenuity of human learners, this work has demonstrated some important principles. Human learning is much more reliant upon flexible, interconnected representations than past models have supposed. To make significant progress in studying cognition, we should focus on models that use representations that grow in complexity

and that recognize and take advantage of complex relationships between large collections of concepts, rather than attempting to learn individual tasks and domains in isolation.

References

- Ryan Prescott Adams, Zoubin Ghahramani, and Michael I. Jordan. Tree-structured stick breaking for hierarchical data. In J. Shawe-Taylor, R. Zemel, J. Lafferty, and C. Williams, editors, *Advances in Neural Information Processing Systems 23*. MIT Press, 2011.
- David J. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII*, pages 1–198. Springer, Berlin, 1985.
- John R. Anderson. *The adaptive character of thought*. Lawrence Erlbaum, Hillsdale, NJ, 1990.
- John R. Anderson. The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429, 1991.
- Charles E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.
- F. Gregory Ashby. *Multidimensional Models of Perception and Cognition*. Erlbaum, Hillsdale, NJ, 1992.
- F. Gregory Ashby and Leola A. Alfonso-Reese. Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39:216–233, 1995.
- F. Gregory Ashby and Ralph E. Gott. Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1):33–53, January 1988.
- F. Gregory Ashby and W. Todd Maddox. Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37(3):372–400, 1993.
- Scott Atran. Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, 21:547–609, 1998.

References

- Joseph L. Austerweil and Thomas L. Griffiths. Learning hypothesis spaces and dimensions through concept learning. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 2010.
- David Blackwell and James B. MacQueen. Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.
- David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30, 2010. ISSN 0004-5411.
- Ingwer Borg and Patrcik Gorenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer-Verlag, New York, 2 edition, 2005.
- Jerome S. Bruner, Jacqueline J. Goodnow, and George A. Austin. *A Study of Thinking*. Wiley, New York, 1956.
- Kevin R. Canini and Thomas L. Griffiths. A nonparametric Bayesian model of multi-level category learning. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 2011.
- Kevin R. Canini, Mikhail M. Shashkov, and Thomas L. Griffiths. Modeling transfer learning in human categorization with the hierarchical Dirichlet process. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.
- Nick Chater and Mike Oaksford. Ten years of the rational analysis of cognition. *Trends in Cognitive Science*, 3:57–65, 1999.
- Allan M. Collins and M. Ross Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior*, 8(2):240–247, 1969.
- James M. Cortese and Brian P. Dyre. Perceptual similarity of shapes generated from Fourier descriptors. *Journal of Experimental Psychology: Human Perception and Performance*, 22(1):133–43, 1996.
- Marco Cusumano-Towner, Arjun Singh, Stephen Miller, James F. O’Brien, and Pieter Abbeel. Bringing clothing into desired configurations with limited perception. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011.
- Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, NY, 2001.
- Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, New York, 2000.

References

- Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430), 1995.
- N. H. Feldman, T. L. Griffiths, and J. L. Morgan. Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2009.
- Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.
- Thomas S. Ferguson. Bayesian density estimation by mixtures of normal distributions. In M. Rizvi, J. Rustagi, and D. Siegmund, editors, *Recent advances in statistics*, pages 287–302. Academic Press, New York, NY, 1983.
- D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A.A. Kalyanpur, A. Lally, J.W. Murdock, E. Nyberg, J. Prager, et al. Building watson: An overview of the deepqa project. *AI Magazine*, 31(3):59–79, 2010.
- Susan A. Gelman and James D. Coley. Language and categorization: The acquisition of natural kind terms. In Susan A. Gelman and James P. Byrnes, editors, *Perspectives on language and thought: Interrelations in development*, pages 146–196. Cambridge University Press, 1991.
- Robert Goldstone. An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments & Computers*, 26(4):381–386, 1994. ISSN 0743-3808.
- Noah D. Goodman, Joshua B. Tenenbaum, Jacob Feldman, and Thomas L. Griffiths. A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1):108–154, 2008.
- Thomas L. Griffiths and Michael L. Kalish. Language evolution by iterated learning with Bayesian agents. *Cognitive Science: A Multidisciplinary Journal*, 31(3):441–480, 2007. ISSN 0364-0213.
- Thomas L. Griffiths, Kevin R. Canini, Adam N. Sanborn, and Daniel J. Navarro. Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, pages 323–328, 2007.
- Katherine A. Heller and Zoubin Ghahramani. Bayesian hierarchical clustering. In *Proceedings of the International Conference on Machine Learning*, volume 22, 2005.
- Katherine A. Heller, Adam N. Sanborn, and Nick Chater. Hierarchical learning of dimensional biases in human categorization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*. MIT Press, 2010.

References

- Anne S. Hsu and Thomas L. Griffiths. Effects of generative and discriminative learning on use of category variability. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 2010.
- Clark L. Hull. Quantitative aspects of the evolution of concepts. *Psychological Monographs*, XXVIII(1), 1920.
- J. Huttenlocher, L. V. Hedges, and J. L. Vevea. Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, 129:220–241, 2000.
- D. Jurafsky, J.H. Martin, A. Kehler, K. Vander Linden, and N. Ward. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, volume 163. MIT Press, 2000.
- Michael L. Kalish, Thomas L. Griffiths, and Stephan Lewandowsky. Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, 14(2):288–294, April 2007.
- Frank C. Keil. *Semantic and conceptual development: An ontological perspective*. Harvard Univ. Press, Cambridge, 1979.
- Charles Kemp and Joshua B. Tenenbaum. Structured statistical models of inductive reasoning. *Psychological Review*, 116(1):20–58, 2009.
- Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 381–388, 2006.
- Charles Kemp, Amy Perfors, and Joshua B. Tenenbaum. Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, 10(3):307–321, 2007.
- John K. Kruschke. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1):22–44, January 1992.
- Stephan Lewandowsky, Thomas L. Griffiths, and Michael L. Kalish. The wisdom of individuals: Exploring people’s knowledge about everyday events using iterated learning. *Cognitive Science*, 33:969–998, 2009.
- Bradley C. Love, Douglas L. Medin, and Todd M. Gureckis. SUSTAIN: A network model of category learning. *Psychological Review*, 111(2):309–332, 2004.
- R. Duncan Luce. *Individual choice behavior*. John Wiley, New York, 1959.
- David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA, 1982. ISBN 0716715678.

References

- Stephen C. McKinley and Robert M. Nosofsky. Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, 21(1):128–148, 1995.
- Douglas L. Medin and Marguerite M. Schaffer. Context theory of classification learning. *Psychological Review*, 85(3):207–238, 1978.
- Marina Meila. Comparing clusterings by the variation of information. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*, pages 173–187. Springer, 2003.
- Tom M. Mitchell. *Machine learning*. McGraw Hill, New York, 1997.
- Gregory L. Murphy. *The Big Book of Concepts*. The MIT Press, 2002.
- Radford M. Neal. Markov chain sampling methods for dirichlet proces mixture models. Technical Report 9815, Department of Statistics, University of Toronto, September 1998.
- Allen Newell and Herbert A. Simon. The logic theory machine: A complex information processing system. *Information Theory, IRE Transactions on*, 2(3):61–79, 1956.
- Robert M. Nosofsky. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115:39–57, 1986.
- Robert M. Nosofsky. Relation between the rational model and the context model of categorization. *Psychological Science*, 2(6):416–421, 1991.
- Robert M. Nosofsky. Optimal performance and exemplar models of classification. In M. Oaksford and N. Chater, editors, *Rational models of cognition*, pages 218–247. Oxford University Press, 1998.
- Robert M. Nosofsky and Thomas J. Palmeri. A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, 5(3):345–369, 1998.
- Robert M. Nosofsky, Thomas J. Palmeri, and Stephen C. McKinley. Rule-plus-exception model of classification learning. *Psychological Review*, 101(1):53–79, 1994.
- Mike Oaksford and Nick Chater, editors. *Rational models of cognition*. Oxford University Press, Oxford, 1998.
- Amy Perfors and Joshua B. Tenenbaum. Learning to learn categories. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2009.
- Carl E. Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems 12*, 2000.

References

- Florencia Reali and Thomas L. Griffiths. The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, 111:317–328, 2009.
- Stephen K. Reed. Pattern recognition and categorization. *Cognitive Psychology*, 3:393–407, 1972.
- John A. Rice. *Mathematical statistics and data analysis*. Duxbury, Belmont, CA, 2nd edition, 1995.
- Eleanor Rosch, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive Psych.*, 8(3):382–439, 1976. ISSN 0010-0285. doi: DOI:10.1016/0010-0285(76)90013-X.
- Yves Rosseel. Mixture models of categorization. *Journal of Mathematical Psychology*, 46:178–210, 2002.
- Daniel M. Roy, Charles Kemp, Vikash K. Mansinghka, and Joshua B. Tenenbaum. Learning annotated hierarchies from relational data. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19*, Cambridge, MA, 2007. MIT Press.
- Adam N. Sanborn, Thomas L. Griffiths, and Richard M. Shiffrin. Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, 2009.
- Adam N. Sanborn, Thomas L. Griffiths, and Daniel J. Navarro. Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117(4):1144–1167, 2010a.
- Adam N. Sanborn, Thomas L. Griffiths, and Richard M. Shiffrin. Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, 60(2):63–106, 2010b.
- Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.
- Roger N. Shepard. Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, 1(1):54–87, 1964. ISSN 0022-2496.
- Roger N. Shepard. Towards a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, September 1987.
- Roger N. Shepard and Gregory W. Cermak. Perceptual-cognitive explorations of a toroidal set of free-form stimuli. *Cognitive Psychology*, 4(3):351–377, 1973.
- Bernard W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986.

References

- J. David Smith and John Paul Minda. Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(6): 1411–1436, 1998.
- Yee Whye Teh. Nonparametric Bayesian mixture models - release 2.1. <http://www.gatsby.ucl.ac.uk/~ywteh/research/software.html>, 2004.
- Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.
- Joshua B. Tenenbaum, Thomas L. Griffiths, and Charles Kemp. Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, 10(7):309–318, 2006.
- Sebastian Thrun and Lorien Pratt, editors. *Learning to Learn*. Kluwer Academic Publishers, 1998.
- Amos Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.
- Wolf Vanpaemel and Gert Storms. In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, 15(4):732–749, August 2008.
- Andy J. Wills and Emmanuel M. Pothos. On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, 2011.
- Jing Xu, Thomas L. Griffiths, and Mike Dowman. Replicating color term universals through human iterated learning. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 2010.
- Saiwing Yeung and Thomas L. Griffiths. Estimating human priors on causal strength. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 2011.