Title
Bayesian Models of Learning and Reasoning with Relations

Author
Chen, Dawn

Publication Date
2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Bayesian Models of Learning and Reasoning with Relations

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Psychology

by

Dawn Chen

2014

ABSTRACT OF THE DISSERTATION

Bayesian Models of Learning and Reasoning with Relations

by

Dawn Chen

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2014

Professor Keith J. Holyoak, Co-chair

Professor Hongjing Lu, Co-chair

How do humans acquire relational concepts such as *larger*, which are essential for analogical inference and other forms of high-level reasoning? Are they necessarily innate, or can they be learned from non-relational inputs? Using comparative relations as a model domain, we show that structured relations can be learned from unstructured inputs of realistic complexity, applying bottom-up Bayesian learning mechanisms that make minimal assumptions about innate representations. First, we introduce *Bayesian Analogy with Relational Transformations* (*BART*), which represents relations as probabilistic weight distributions over object features. BART learns two-place relations such as *larger* by bootstrapping from empirical priors derived from initial learning of one-place predicates such as *large*. The learned relational representations allow classification of novel pairs and yield the kind of distance effect observed in both humans and other primates. Furthermore, BART can transform its learned weight distributions to reliably

solve four-term analogies based on higher-order relations such as *opposite* (e.g., *larger:smaller* :: *fiercer:meeker*). Next, we present BARTlet, a representationally simpler version of BART that models how symbolic magnitudes (e.g., size or intelligence of animals) are derived, represented, and compared. BARTlet creates magnitude distributions for objects by applying BART-like weights for categorical predicates such as *large* (learned with the aid of empirical priors derived from pre-categorical comparisons) to more primitive object features. By incorporating psychological reference points that control the precision of these magnitudes in working memory, BARTlet can account for a wide range of empirical phenomena involving magnitude comparisons, including the distance effect, the congruity effect, the markedness effect, and sensitivity to the range of stimuli. Finally, we extend the original discriminative BART model to generate (rather than classify) relational instances, allowing it to make quasi-deductive transitive inferences (e.g., "If *A* is larger than *B* and *B* is larger than *C*, then *A* is larger than *C*") and predict human responses to questions such as, "What is an animal that is smaller than a dog?" Our work is the first demonstration that relations and symbolic magnitudes can be learned from complex non-relational inputs by bootstrapping from prior learning of simpler concepts, enabling human-like analogical, comparative, generative, and deductive reasoning.

The dissertation of Dawn Chen is approved.

Ying Nian Wu

Catherine M. Sandhofer

Hongjing Lu, Committee Co-chair

Keith J. Holyoak, Committee Co-chair

University of California, Los Angeles

2014

DEDICATION

This dissertation is dedicated to my husband, Trevor Standley. Without his unconditional love, endless encouragement, and constant support, this work would not have been possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Chapter 1 of this dissertation is a version of Lu, Chen, and Holyoak (2012). While my co-authors and I discussed the conceptual framework of the BART model together, Dr. Hongjing Lu contributed the most to its initial conception and later development. Dr. Keith Holyoak wrote significant portions of the manuscript, contributing especially to the introduction and general discussion. Chapter 2 is a version of Chen, Lu, and Holyoak (2014). Both Dr. Lu and Dr. Holyoak contributed to the ideas behind the BARTlet model and preparation of the manuscript. Part of Chapter 3 has been published in Chen, Lu, and Holyoak (2013). It has been considerably revised and expanded in the version that appears in this dissertation. Dr. Lu made significant conceptual contributions, including the derivation of the modified variational method for the generative model. Both Dr. Lu and Dr. Holyoak suggested evaluations for the model. Dr. Holyoak aided in writing the paper.

My co-authors and I thank Airom Bleicher for helping us to conduct the animal generation study on Amazon Mechanical Turk, Charles Kemp for sharing Leuven inputs, Mark Steyvers for making the topic model code available, and Peter Gordon for providing us with a pre-processed version of the Wikipedia corpus. In addition, we thank John Anderson, Alex Doumas, Robert Goldstone, John Hummel, Alexander Petrov, Dario Salvucci, Alan Yuille, and eight anonymous reviewers for providing valuable comments on earlier drafts of the various chapters.

This dissertation research was supported by grant N000140810186 from the Office of Naval Research, and by UCLA's University Fellowship, Chancellor's Prize, Graduate Summer Research Mentorship Program, and Dissertation Year Fellowship.

References

Chen, D., Lu, H., & Holyoak, K. J. (2013). Generative inferences based on a discriminative Bayesian model of relation learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2028-2033). Austin, TX: Cognitive Science Society.

Chen, D., Lu, H., & Holyoak, K. J. (2014). The discovery and comparison of symbolic magnitudes. *Cognitive Psychology*, *71*, 27-54.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*, 617-648.

VITA

EDUCATION

M.A.	2010	Psychology (Concentration: Computational Cognition) University of California, Los Angeles
B.A.	2008	Cognitive Science, Computer Science University of California, Berkeley

PUBLICATIONS

Chen, D., Lu, H., & Holyoak, K. J. (2014). The discovery and comparison of symbolic magnitudes. *Cognitive Psychology*, *71*, 27-54.

Chen, D., Lu, H., & Holyoak, K. J. (2013). Generative inferences based on a discriminative Bayesian model of relation learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2028-2033). Austin, TX: Cognitive Science Society.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*, 617-648.

Chen, D., Lu, H., & Holyoak, K. J. (2010). Learning and generalization of abstract semantic relations: Preliminary investigation of Bayesian approaches. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 871-876). Austin, TX: Cognitive Science Society.

Chen, D., & Holyoak, K. J. (2010). Enhancing acquisition of intuition versus planning in problem solving. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1875-1880). Austin, TX: Cognitive Science Society.

PRESENTATIONS

Chen, D., Lu, H., & Holyoak, K. J. (2013). Generative inferences based on a discriminative Bayesian model of relation learning. Poster presentation at the 35th Annual Conference of the Cognitive Science Society, Berlin, Germany (August).

Holyoak, K. J., Thompson, B. J., & Chen, D. (2011). Math as model: Potential roles of game-like technology in teaching. Oral presentation at the CATS Workshop for Research on Games and Learning, UCLA (January).

Chen, D., Lu, H., & Holyoak, K. J. (2010). Bayesian approaches to learning abstract semantic relations. Poster presentation at the Workshop for Women in Machine Learning, Vancouver, Canada (December).

Chen, D., Lu, H., & Holyoak, K. J. (2010). Learning and generalization of abstract semantic relations: Preliminary investigation of Bayesian approaches. Oral presentation at the 32nd Annual Conference of the Cognitive Science Society, Portland, OR (August).

Chen, D., & Holyoak, K. J. (2010). Enhancing acquisition of intuition versus planning in problem solving. Poster presentation at the 32nd Annual Conference of the Cognitive Science Society, Portland, OR (August).

CHAPTER 1:

BAYESIAN ANALOGY WITH RELATIONAL TRANSFORMATIONS

One of the hallmarks of human reasoning is the ability to form representations of relations between entities, and then to reason about the higher-order relations between these relations. Whereas concepts such as *larger* and *smaller*, for example, are first-order relations, potentially derivable by comparing features of individual objects, a relation such as *opposite* is a higher-order relation between relations (Gentner, 1983). The capacity to represent and reason with higher-order relations has been considered central to human analogical thinking (Gentner, 2010; Halford, Wilson & Phillips, 2010; Holyoak, 2012).

The development of knowledge about comparative relations provides a clear illustration of these human abilities. By the time they reach school age, children have acquired the ability to accurately assess whether one object (e.g., bear) is "larger" or "smaller" than another (e.g., fox), even under speed pressure (McGonigle & Chalmers, 1984). Moreover, like adults (Moyer & Bayer, 1976), children's judgments show a *symbolic distance effect*: the greater the magnitude difference between the two items, the faster the comparison can be made. Such symbolic comparisons are presumably based on stored representations of the perceptual dimensions associated with the individual concepts. A great deal of evidence—particularly, parallels between performance with symbolic and perceptual comparisons—suggests that humans and other species share a basic mechanism for representing continuous quantities on a "mental number line" (Dehaene & Changeux, 1993; Gallistel, 1993; Moyer, 1973). Moreover, rhesus monkeys are capable of learning shapes (Arabic numerals) corresponding to small numerosities (1-4 dots), such that the shapes acquire neural representations overlapping those of the corresponding perceptual numerosities (Diester & Nieder, 2007).

These species-general achievements are impressive. However, human children go on to acquire a deeper understanding of comparative relations. For example, they learn that the relations *larger* and *smaller* have a special relationship to each other (a type of antonym). Analyses of corpora of child speech have identified systematic use of such gradable antonyms by children aged 2-5 years (Jones & Murphy, 2005), and experimental studies show that by at least age 6 years children can use such concepts metaphorically (Gardner, 1974), and are aware that antonyms are contradictory (Glass, Holyoak & Kossan, 1977). Children eventually understand that a pair of concepts like *larger-smaller* is related in basically the same way as the pair *faster-slower*, allowing them to see that such pairs of relations form analogies.

It seems that "something special" happens that enables humans to acquire higher-order relational representations. Animals of many taxa have the basic ability to detect and act based on perceptual relations, as exemplified by classic work on relational transposition in rats (Lawrence & DeRivera, 1954), and rudimentary numerical processing is clearly available to many primate and other species (see Gallistel, 1993). Nonetheless, there is a great deal of evidence that the relational capacities of humans exceed that of any other species, perhaps in a qualitative fashion (Povinelli, 2000; Penn, Holyoak & Povinelli, 2008). The difference has been characterized as a human capacity for *relational reinterpretation*: the ability to transform perceptually-grounded relations into explicit relational structures that distinguish the *roles* of relations from the objects that fill them (Doumas & Hummel, 2012), augmented by the additional ability to form higher-order relational concepts (e.g., representations of hidden causes, or mental states of others).

From a computational perspective, the challenge is to explain what it might mean for a relation to be reinterpreted or rerepresented into a more explicit and abstract form, and to develop formal models of such a process. How could an inductive system ever get from some

initial pool of perceptually-available features to more abstract concepts corresponding to higher-order relations (e.g., *opposite*), which seem not to be based entirely on the set of perceptual features that provided a starting point? The difficulty of the learning problem is compounded by evidence that children seem to acquire concepts largely from modest numbers of positive examples provided by adults (Bloom, 2000; see Xu & Tenenbaum, 2007).

An important part of the recipe for abstraction may be a pool of innate concepts. For example, Carey (2011) has argued, "There is no proposal I know for a learning mechanism available to nonlinguistic creatures that can create representations of objects, number, agency, or causality from perceptual primitives" (p. 115). But as Carey also argues, constructive mechanisms may operate over some combination of perceptual inputs and pre-existing concepts to create new types of mental representations. Part of a learner's innate endowment may be processes that permit various forms of *bootstrapping*, whereby one type of representation is transformed into another. For example, there is evidence that analogical reasoning may play an important role in children's acquisition of natural number (Carey, 2011; Opfer & Siegler, 2007; see also Gentner, 2010; Kurtz, Miao & Gentner, 2001).

Goals of the Present Paper

In the present paper we present a new model of the induction of relational representations, *Bayesian Analogy with Relational Transformations* (*BART*). In general terms, BART is a computational-level model[1] (Anderson, 1991; Griffiths, Chater, Kemp, Perfors & Tenenbaum, 2010; Marr, 1982) that employs bootstrapping to acquire and transform relational representations. We apply BART to the domain of relations related to comparative judgment. Although this is only a special case of the more general problem of relation learning, it is a

domain that offers the advantage of a wealth of empirical evidence—behavioral, comparative, developmental, and neural—that can guide theory development.

Our particular focus will be on a restricted but nonetheless realistic subdomain: relations definable over continuous-valued features associated with animal concepts. The basic inputs provided to the model are vectors of feature values for a set of dimensions. Our first goal is to have the model learn representations of first-order relations such as *larger* and *smaller*, *fiercer* and *meeker*, based on *empirical priors* (i.e., prior knowledge itself acquired by learning simpler concepts from relevant data) coupled with a limited set of positive examples instantiating relations. The use of empirical priors in learning is an example of a simple form of bootstrapping, whereby initial learning of a different or simpler concept provides a useful basis for acquiring more complex concepts. Similar ideas have been exploited in neural-network models of learning (e.g., Bao & Munro, 2006; Elman, 1993; but see Rohde & Plaut, 1999). Newport (1990) argued that children's cognitive limitations (e.g., less capacity in working memory) may actually benefit certain aspects of language acquisition. Halford, Wilson and Phillips (1998) proposed that children are able to learn one-place predicates (e.g., *large*, *small*) prior to two-place relations (e.g., *larger*, *smaller*) because the former require less working-memory capacity. Given the strong evidence for this sequential progression in children's concept acquisition (e.g., Smith, 1989), we will focus on the potential use of one-place predicates as the basis for forming empirical priors to facilitate learning of comparative relations.

The use of only positive training examples makes it possible to acquire a stable and context-independent representation of a relation (whereas negative examples can be of many different types, and the learned relational representation will vary depending on which negative examples are encountered). We aim to demonstrate that the acquired representations of relations

are generalizable (i.e., can be used to evaluate novel instantiations of the relations), and are sensitive to a basic factor that influences the difficulty of human relational judgments.

Our second goal is to show how these first-order relational representations can be transformed and re-represented so as to allow the model to evaluate higher-order analogy problems of the form A:B::C:D instantiated by the learned relations (e.g., *larger:smaller :: fiercer:meeker*, rather than *fiercer:slower*). This transformation process is based on what we term *importance-guided mapping*, a subsymbolic form of analogical mapping based on similarity of weights associated with object features. Our overall aim is to provide a proof-of-concept that, for the domain of comparative relations, the capacity to solve structured analogy problems can be acquired by applying basically bottom-up learning mechanisms to raw inputs consisting of object concepts coded as simple feature vectors.

Judgments Based on Comparative Relations

Our target domain, comparative relations, is tied to a rich body of cognitive research. Comparative judgments exhibit a number of robust empirical phenomena. The most notable is the semantic distance effect (Moyer, 1973; Moyer & Landauer, 1967). Strong empirical evidence indicates that the long-term-memory representation of a relation such as *larger* includes quantitative information that makes the difficulty of comparison decline as the magnitude difference increases. The symbolic distance effect is observed not only with quasi-perceptual dimensions such as size, but also with more abstract dimensions such as animal intelligence (Banks, White, Sturgill & Mermelstein, 1983) and such concepts as adjectives of quality (e.g., *good*, *fair*; Holyoak & Walker, 1976). Although magnitude representations exhibits analog properties, much like an internal number line (e.g., Woocher, Glass & Holyoak, 1978), magnitude comparisons do not in general depend on visual imagery (Holyoak, 1977). Non-

human primates also exhibit a distance effect for judgments of numerosity (see Neider & Miller, 2004, for a review). Given its ubiquity, the distance effect is arguably the primary signature that a learned representation of a comparative relation is psychologically realistic; hence the distance effect will be the first empirical focus in our evaluation of BART. In the General Discussion we will consider how BART might be extended to explain additional phenomena involving comparative judgments.

In human children, comparative adjectives emerge as early words in the lexicon, with clear developmental trends (Smith, 1989; Smith & Sera, 1992). In general, children progress from a global sense of similarity and dissimilarity of objects, to learning one-place predicates that focus on specific dimensions of individual objects (*big*, *small*), to learning two-place comparative relations between multiple objects (*bigger*, *smaller*). As noted earlier, children eventually detect higher-order similarities and differences between comparative relations, coming to understand (for example) that *higher* and *lower* are polar opposites. Less is known about the details of this part of the developmental progression, but presumably a prerequisite for learning a higher-order relation approximating *gradable opposite* is to first achieve some degree of mastery with pairs of first-order comparative relations, such as *higher* and *lower*.

The acquisition of relations is intimately related to the development of analogical reasoning ability. A great deal of evidence indicates that children's ability to think analogically changes over the course of cognitive development (e.g., Chen, Sanchez & Campbell, 1997; Gentner & Toupin, 1986; Holyoak, Junn & Billman, 1984; Tunteler & Resing, 2002, 2007). The developmental transition toward greater reliance on relational structure has been termed the *relational shift* (Gentner & Rattermann, 1991). The empirical phenomenon of a relational shift is well established, but there has been some debate regarding the developmental mechanisms that

may underlie it. Considerable evidence indicates that some changes are maturational, involving increases in working memory capacity (Halford et al., 1998) and inhibitory control (Morrison, Doumas & Richland, 2010; Richland, Morrison & Holyoak, 2006). However, it is universally accepted that learning new relations is a prerequisite for solving analogy problems based on these relations (Goswami, 1992, 2001). In the present paper we focus on relation learning, the most basic mechanism required for analogical reasoning.

Approaches to the Acquisition of Relational Concepts

In recent years a number of different approaches to modeling the induction of relational concepts have been explored, which we will briefly review. We begin by laying out some criteria that we believe are of general importance in evaluating psychological theories of relation learning, including the present model.

(1) *Choice of inputs*: The model should be capable of learning from inputs of realistic complexity that were independently generated. There is certainly much to be gained from exploratory work using small hand-coded inputs, and specifying realistic representations poses many challenges. However, without some tests using independently-generated inputs, it is difficult to assess the extent to which a model may owe its successes to the foresight and charity of the modelers. In addition, the model (unless it explicitly assumes that all relational representations are innate) must be able to learn at least some relations from inputs that are *non*-relational (e.g., object representations).

(2) *Learning efficiency*: As a psychological model, learning should be achieved on a human time scale as measured by the number of training examples required to produce at least partial success. Given that children seem to be able to acquire preliminary understanding of many concepts from relatively few examples, a model should also be able to demonstrate

efficiency by learning from a modest number. Although what "relatively few" means is inevitably vague, our focus will be on what can be learned from up to 100 or 200 positive training examples.

(3) *Generalization*: The model should be able to make accurate relational judgments about novel examples. It is not sufficient to show that the model can learn the training examples as "relational facts"; it must also be able to apply its relational representations productively.

(4) *Performance difficulty*: The difficulty of human relational judgments can be modulated by many factors. To be considered psychological, a model should account for at least some sources of differential difficulty in relational judgments for humans (and/or other animals).

(5) *Flexible reasoning*: Relational knowledge plays an essential role in human reasoning and thinking, in essence providing a deeper source of information about conceptual similarity. Accordingly, the relational representations acquired by the model should be useable (either directly or after some additional learning process) to perform a variety of tasks that require relational reasoning (e.g., solving analogy problems).

These criteria are inherently qualitative rather than quantitative. Alternative assessment metrics could no doubt be advanced, but we have found the above criteria helpful in evaluating previous work on relational learning, as well as the models we test in the present paper.

Vector Space Models

There is an extensive literature on automated methods for extracting relations based on the statistics of word or phrase co-occurrence in a large corpus of text. One class of methods, termed *vector space models*, originates from an information retrieval technique of the same name, and uses vectors or matrices in which the value of each element is derived from the frequency of some event, such as the frequency with which a certain word appears in a particular document or

phrase (for reviews, see Turney, 2006; Turney & Pantel, 2010). For example, Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) yields vector representations of individual words by applying singular value decomposition to lexical co-occurrence data from a large corpus of text. LSA has proved useful in many applications that require measures of semantic similarity of concepts (Wolf & Goldman, 2003, including modeling the retrieval of story analogs (Ramscar & Yarlett, 2003). However, LSA vectors do not provide any direct basis for identifying abstract relations between concepts (although some modest results have been achieved by exploiting LSA vectors for relation words, such as *opposite*; Mangalath, Quesada & Kintsch, 2004).

Related machine-learning algorithms have achieved greater success by working directly from co-occurrence data for word combinations found in a large corpus of text (Turney & Littman, 2005). Perhaps the most successful method is Latent Relational Analysis (LRA; Turney, 2006), which has been applied to the task of solving SAT verbal analogy problems (e.g., *quart:volume :: mile:distance*). The algorithm searches for patterns of words in which the A and B term (and their synonyms) appear (e.g., "quarts in volume"). The frequencies of the various patterns are used to create a vector of relational features for A:B; vectors are similarly formed for potential C:D completions. Cosine similarity is calculated to compare the A:B vector to the corresponding vectors created for various alternative C:D pairs, and the most similar C:D is selected as the analogical completion. LRA achieves a level of accuracy on SAT analogy problems comparable to that attained by college students.

Vector space models such as LRA provide effective machine-learning tools for extracting relational similarity. However, these models operate directly on texts that include relational vocabulary. Our present focus is on learning from inputs based on representations of individual

object concepts (including a set of such inputs that is derived from texts by a method similar to LSA).

Hierarchical, Generative Bayesian Models

Perhaps the most ambitious line of work has focused on hierarchical Bayesian models that integrate statistical learning with explicit representations of higher-order relational structures (Goodman, Ullman & Tenenbaum, 2011; Kemp, Perfors & Tenenbaum, 2007; Tenenbaum, Kemp, Griffiths & Goodman, 2011). For example, Kemp and Tenenbaum (2008) showed how Bayesian techniques can operate on relational structures to learn systems such as hierarchies and linear orderings (see also Kemp & Jern, 2009; Kemp, Tenenbaum, Griffiths, Yamada & Ueda, 2006). In general terms, these hierarchical models are *generative* (Mackay, 2003) in the sense that representations of alternative relational structures are used to predict incoming data, and the data in turn are used to revise probability distributions over alternative structures. The highest level of the structure typically consists of a formal grammar or a set of logical rules that generates a set of alternative relational "theories", which are in turn used to predict the observed data.

Although hierarchical generative models are extremely powerful, the models to date have generally focused on systems of formal relations that have a well-defined logical structure known to the modeler (e.g., hierarchies, rings, or chains). The set of possible relational structures is provided to the system by specifying a grammar that generates them. Since the postulated grammar of relations is not itself learned, the generative approach (although certainly incorporating inductive learning) retains rather strong nativist assumptions.

Neural Network Models

The BART model, like generative Bayesian models, operates at the computational level; however, its emphasis on bottom-up learning and emergence overlaps with the goals of algorithmic approaches to relation learning and analogy based on neural networks (e.g., Gasser & Colunga, 2000; Jani & Levine, 2002; Leech, Mareschal & Cooper, 2008; Rogers & McClelland, 2008; see McClelland et al., 2010). Our model shares the general aim of seeking emergence of structure from statistical operations over minimally structured inputs, coded as feature vectors.

A standard connectionist approach to learning relational structures has been to create a feed-forward network in which separate pools of input units are used to code features of an object in a role and of a relation. These pools interact via a hidden layer, and thereby activate output units representing another filler of the role. Rogers and McClelland (2008) developed a model based on this type of architecture that learns simple propositions (e.g., "a canary can fly"). The model takes a sequence of input-output pairs and over repetitions adjusts the connection weights to learn facts of the form "canary" + "can" → "fly". The Rogers and McClelland model succeeds in capturing a number of important general characteristics of human learning, such as progressive differentiation of concepts and domain-specific feature weighting.

However, models of this sort (including that of Leech et al., 2008, a variation of the same architecture that aimed to account for how children learn to solve simple analogy problems) have not been shown to generalize to dissimilar training items, nor have they been extended to higher-order relations. The Leech et al. model fails on even simple variations of its own training materials. For example, after being trained extensively with the various components required to solve the 4-term analogy *apple:sliced-apple* :: *bread:sliced-bread*, the model cannot generalize

its knowledge to evaluate *sliced-apple:apple* :: *sliced-bread:bread*, where the roles have been reversed (Holyoak & Hummel, 2008; also see French, 2008; Petrov, 2008).

A basic problem is that standard neural-net models offer no way to represent relational roles. In connectionist networks of relation learning, both objects and relations are coded as distributed patterns of weights on links that serve as conduits for activation passed between units. The learned representations of relations therefore remain implicit, and relational knowledge cannot be accessed in a flexible fashion (cf. Halford et al., 2010). For example, in the Rogers and McClelland (2008) model, the representation of an object (e.g., canary) is inherently linked to a particular pool of relation-specific input units. As a consequence, after training the network that one thing a canary can do is fly, the model (unlike a human) would not be able to infer that one kind of thing that flies is a canary (i.e., make an inference in which *canary* serves as the output rather than input).

Symbolic Connectionist Models

The acquisition of relational structure has been a longstanding concern in the literature on analogical reasoning. Gick and Holyoak (1983) proposed that as a consequence of comparing and mapping one situation to an analogous one in a different content domain (e.g., a military and medical problem), humans can learn relational schemas for more abstract categories. Hummel and Holyoak (1997, 2003) developed a symbolic connectionist model, LISA (*Learning and Inference with Schemas and Analogies*), which is able to form such schemas by comparing and mapping examples. However, LISA's learning algorithm works by recombining pre-existing (and hand-coded) relational concepts, rather than by building new relational predicates.

More recently, Doumas, Hummel, and Sandhofer (2008) developed a related model called DORA (*Discovery of Relations by Analogy*) that addresses the fundamental goal of

creating new relational predicates from non-relational inputs. The basic representational assumptions of DORA are very similar to those of LISA, with both objects and roles of relations represented in a distributed fashion over a pool of semantic units. Relations are explicitly represented by localist units that code individual roles (e.g., *larger* would be coded by units for the larger object and for the smaller one in a pair). Bindings of objects to roles are coded dynamically in working memory by temporal patterns (synchrony in LISA, close asynchrony in DORA), and statically in long-term memory by conjunctive units. Because relations are represented explicitly and independently of their fillers, DORA (like LISA, but unlike classical connectionist models) is able to flexibly generalize relations to new contexts. But like more traditional neural networks, objects and relations are represented in the same basic way (as patterns of weights on links connecting units that code semantic features).

The basic learning algorithm used by DORA is to first compare feature representations of individual objects, creating new predicate units that connect to shared features (e.g., from the objects *elephant* and *bear*, a new one-place predicate connected to the shared feature *large* might be generated). Later, a pair of objects respectively instantiating the one-place predicates *large* and *small* (e.g., *elephant* and *mouse*) might be compared to another pair instantiating these same predicates (e.g., *walrus* and *frog*). With the aid of a comparator operator that can activate the features *more* and *less* based on the specific size values of paired objects, DORA might then generate the two-place predicate *larger*, with its first role connected to *more* and *large* and its second role to *less* and *small*. As additional examples of paired objects are encountered, sequential updating will refine the relational representation, honing in on features that prove to be invariant across examples.

The progression of learning comparative relations in DORA—from objects encoded as features, to one-place predicates such as *large*, to two-place relations such as *larger* (that then undergo gradual refinement)—parallels the general developmental sequence identified by Smith (1989; Smith & Sera, 1992) and others in studies of children's acquisition of comparative relations. But although DORA can generate human-like patterns of relation learning, the robustness of its learning algorithm has not been extensively tested. So far DORA has only been tested with small hand-coded representations of objects as inputs, and the relations it learns are coded using features drawn from the same set already provided in these inputs. In particular, DORA assumes that in its inputs, all metric dimensions describing objects (e.g., size, speed) are coded by localist units. The model is also endowed with units representing relational features such as *more* and *less*, and with a comparator that will activate these relational features when given two objects associated with values on the same metric dimension. The model tacitly assumes that all relational predicates are definable by at least one pre-coded invariant feature (and the modelers ensure that the inputs satisfy this assumption).

Discriminative Bayesian Models

In contrast to the hierarchical, generative Bayesian models discussed above, simpler Bayesian models of category learning (e.g., Anderson, 1991; Fried & Holyoak, 1984) operate in a more bottom-up fashion. An important variant is *discriminative* Bayesian models (Mackay, 2003), which focus on learning the probabilities of categories given features (rather than the probabilities of features given possible categories). Discriminative models have been applied with considerable success to analysis of neural receptive fields in neurophysiology (Rust, Schwartz, Movshon & Simoncelli, 2005; Victor, 2005), and construction of classification images in psychophysics (Eckstein & Ahumada, 2002; Lu & Liu, 2006). They also provide valuable

tools in other complex statistical tasks, such as the recognition of brain states based on neuroimaging data (Bayesian decoding models; see Friston et al., 2008).

A discriminative Bayesian approach to relation learning was developed by Silva, Heller and Ghahramani (2007), who applied their model to tasks such as identifying classes of hyperlinks between webpages; Silva, Airoldi and Heller (2007) applied the same model to classifying relations based on protein interactions. Although this model was developed to address applications in machine learning, the general principles can potentially be incorporated into models of human relational learning. The BART model represents such an effort.

One key idea is that a relation can be represented as a function that takes a pair of objects as its input and outputs the probability that these objects instantiate the relation. The model learns a representation of the relation from labeled examples, and then applies the learned representation to classify novel examples. A second key idea is that relation learning can be facilitated by incorporating *empirical priors*, which are derived using some simpler learning task that can serve as a precursor to the relation learning task. In particular, Silva, Heller and Ghahramani (2007) explored the usefulness of first teaching the model a general distinction between related and unrelated object pairs, and then using the learned representation of the general relation (*related*) as the empirical prior to bootstrap learning of each specific relation of interest. Chen, Lu and Holyoak (2010) incorporated a similar empirical prior into a model for learning abstract semantic relations, such as *synonym* and *antonym*, from features derived by LSA (Landauer & Dumais, 1997).

These models have demonstrated some success in generalization tests involving identifying novel examples of learned relations.[2] However, none of the models attempted to account for systematic sources of difficulty in human relational judgments, nor did they attempt

to show that the learned relational representations could in turn be used to reason about higher-order relations.

Bayesian Analogy with Relational Transformations: Overview

Choice of Input Representations

BART's inputs are restricted to vectors representing objects, so that all the model's relational knowledge must be acquired from non-relational inputs. Specifically, we focus on learning comparative relations from feature representations of animal concepts. In accord with the first of the criteria for model evaluation we laid out earlier, we wished to ensure that the inputs we used were not hand-coded by the modelers. We chose three different sets of input representations that can be viewed as complementary in their advantages and challenges for testing a learning model.

The first set of inputs can be characterized as simple and transparent (low dimensionality, localist coding of magnitudes). These were feature vectors derived from human ratings of animals on four different magnitude continua (size, speed, fierceness and intelligence; Holyoak & Mah, 1981). No doubt it is oversimplified as a psychological model to assume that each dimension is coded by a single value; nonetheless, there is in fact strong evidence that humans and other primates are equipped with specialized neural circuitry for dealing with approximate magnitude on various dimensions (e.g., Cantlon, Brannon, Carter & Pelphrey, 2006; Dehaene & Changeux, 1993; Fias, Lammertyn, Caessens & Orban, 2007; Piazza et al., 2004, 2006, 2007; Pinel, Piazza, Bihan & Dehaene, 2004). As a practical matter, the simplicity of the rating-based representations (comparable to that of the hand-coded representations employed by Doumas et al., 2008) will prove helpful in understanding how the model operates.

To assess BART's potential to scale-up to learn relations from more complex inputs, we also applied the model to input vectors derived from much more challenging databases (high dimensionality, distributed coding of magnitudes). Our second set, which we will refer to as the "Leuven inputs," was based on norms of the frequency with which participants at the University of Leuven generated features characterizing various animals (De Deyne et al., 2008). Each animal is in the norms is associated with a set of frequencies across more than 750 features. Although some features in the Leuven inputs have *prima facie* relevance to the dimensions of interest to us, none were as direct as the Holyoak and Mah (1981) ratings of specific magnitude dimensions. The Leuven inputs have been successfully used as inputs for a Bayesian model of categorization (Shafto, Kemp, Mansinghka, & Tenenbaum, 2011).

Our third set of inputs was taken from the topic model (Griffiths, Steyvers, & Tenenbaum, 2007). The topic model is broadly similar to LSA (Landauer & Dumais, 1997), taking words in documents as its input and yielding approximate semantic representations of individual words as its output. The topic model uses Bayesian inference to associate each word in the corpus with a set of "topics", which theoretically generate the words. For example, a topic that could loosely be characterized as "finance" would tend to generate such words as *money*, *savings*, and *bank* (in the sense of financial institution). For each word, a vector (typically of length 300) based on conditional probabilities of each topic given the word can be interpreted as a distributed semantic representation over features values. Relative to the Leuven inputs, the topics inputs were much more opaque, in that the meaning associated with each individual topic is generally difficult to characterize; unlike the rating inputs, individual topics do not correspond in any obvious way to the magnitude dimensions underlying the critical comparative.

Vectors based on the Leuven inputs or topics avoid any hand-coding of inputs by the modeler. There is thus is no danger that we the modelers have inadvertently planted to-be-discovered relations in the inputs provided to our learning model. Whereas the simple vectors based on human ratings provide magnitude information very directly, the more complex Leuven and topics vectors do not. To preview our computational results, BART achieves near-perfect performance on generalization and analogy tests after learning from the rating vectors, excellent performance using the much larger Leuven inputs, and reliable though imperfect performance based on the yet more complex topics inputs.

Of course, there is no reason to believe that any of these representations directly correspond to the inputs available to human children when they first learn basic relations. The Leuven inputs perhaps come closest, as they include many features of animals that children would likely know. Children have much more direct access to perceptual and motoric features of objects, which can guide relation learning (e.g., Maouene, Hidaka & Smith, 2008). In addition, children's learning of relations is clearly guided by linguistic cues from adults (e.g., Yoshida & Smith, 2005).

Nonetheless, children surely are faced with considerable complexity in the inputs from which some relations are acquired; hence any plausible model will have to demonstrate robustness. By testing BART with inputs derived from three independent sources, we can have some confidence in the robustness of qualitative aspects of model performance that hold true across all three inputs. For the Leuven and topics inputs, the learning task demands that in a high-dimensional space, BART must infer distributed patterns of features that implicitly code the dimensions over which the model aims to learn relations. In addition, the model must then re-map the acquired weight distributions to solve structured analogy problems. The complexity of

the learning task would likely be comparable (or greater) for inputs further enriched by perceptual and motoric features. In the General Discussion we consider how the approach used by BART might be extended to operate on such inputs.

Overview of the Operation of BART

Broadly speaking, BART proceeds in two stages. (1) *First-order relation learning*: given feature vectors corresponding to pairs of objects, the model uses statistical learning to update weights associated with feature dimensions for various comparative relations (e.g., *larger*, *fiercer*), and then uses its learned weights to decide whether or not novel pairs instantiate a specified relation. As shown in the right-hand plot in Figure 1.1, BART represents a relation using a joint distribution of weights over object features. Weight distributions code not only first-order statistics (means), but also second-order statistics (variances and covariances) that capture the uncertainty of the estimated weights, as well as inter-weight correlations. (2) *Importance-guided relation mapping:* to evaluate potential analogies between pairs of relations (e.g., *larger:smaller :: fiercer:meeker*), the model re-arranges the order of dimensions in acquired weight distributions for the source and target relation pairs to yield transformed relation representations. The transformation is based upon an assessment of importance of each dimension in the source pair, and on the correspondence of weight patterns between the source and the target pair.

Learning first-order relations. BART is capable of learning flexibly from any combination of positive and negative examples; however, we focus on learning from positive examples only (as children are able to do; see Bloom, 2000). Importantly, positive examples make it possible to achieve a relatively context-free relational representation, rather than one that varies with the particular types of negative examples included in the training set. In addition,

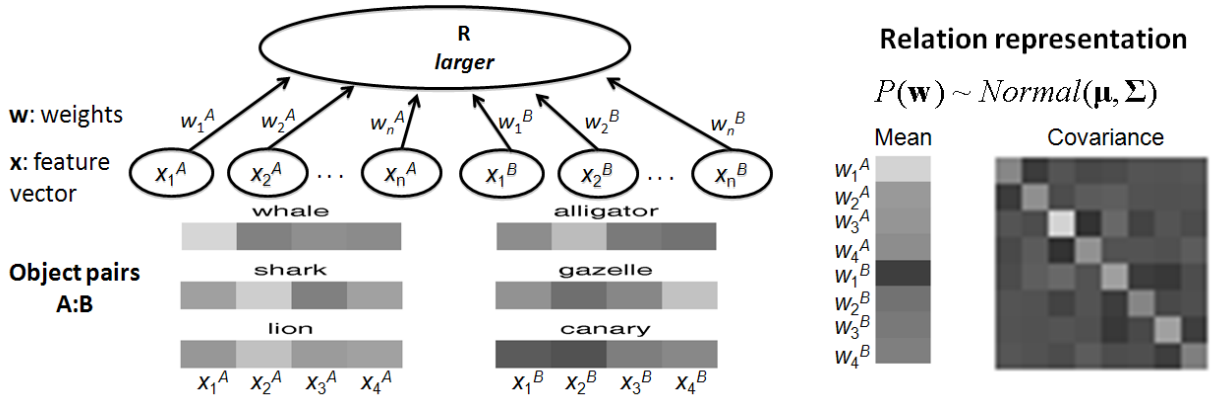


Figure 1.1. Graphical representation of the general framework for relation learning in BART. Left: two objects **A** and **B** in a pair are represented as a vector **x** of *n* features for each object; vector **w** represents the unknown relational weights that define a relation *R*, which is learned using the training set of examples instantiating this relation (e.g., *whale-alligator*, where the intensity of cells represent feature values on each dimension; light indicates high positive values, dark high negative values). Right: the relation is represented as the joint normal distribution of weights **w**. The normal distribution is defined with two parameters: the mean weights vector (shown in the mean plot, in which the intensity indicates the values of means weights), and the covariance matrix of weights including the variance of each weight (diagonal cells in the covariance plot) and the covariances among them (off-diagonal cells in the covariance plot).

because children often appear to learn useful approximations of concepts from small numbers of examples, we aimed to make learning in BART as efficient as possible, focusing on what the model can learn from a modest number of examples (a range of up to about 200). Also, children's relation learning is clearly guided by linguistic inputs from adults (e.g., Gentner, Anggori & Klibanoff, 2011; Yoshida & Smith, 2005). In natural speech to children, comparative relations are given names, such as "larger", which are explicitly connected to positive examples ("the elephant is larger than the hippo"). Accordingly, BART focuses on supervised learning using labeled positive examples.

Bayesian framework. BART learns a first-order relation by estimating the distribution of a corresponding weight vector **w** from a set of training pairs that constitute examples of that relation, as schematized in Figure 1.1. We adopt a Bayesian framework to learn the probability distribution of $P(\mathbf{w} \mid \mathbf{X_S}, \mathbf{R_S})$, where $\mathbf{X_S}$ represents the feature vectors for object pairs in the training set, the subscript **S** indicates the set of training examples, and $\mathbf{R_S}$ is a set of binary indicators, each of which (denoted by *R*) indicates whether a particular pair of objects instantiates the relation or not. The vector **w** constitutes the learned relational representation, which can be interpreted as weights reflecting the influence of the corresponding feature dimensions in **X** for relation judgment. Learning a first-order relation is based on estimating the posterior distribution of weights, which can be computed by applying Bayes' rule using the likelihood of the training data and the prior distribution for **w**:

$$P(\mathbf{w} \mid \mathbf{X_S}, \mathbf{R_S}) = \frac{P(\mathbf{R_S} \mid \mathbf{w}, \mathbf{X_S}) P(\mathbf{w})}{\int_{\mathbf{w}} P(\mathbf{R_S} \mid \mathbf{w}, \mathbf{X_S}) P(\mathbf{w})}. \tag{1.1}$$

The likelihood is defined as a logistic function for computing the probability that a pair instantiates the relation, given the weights and feature vectors,

$$P\left(R = 1 \mid \mathbf{w}, \mathbf{X}\right) = \left(1 + e^{-\mathbf{w}^{\mathrm{T}}\mathbf{X}}\right)^{-1}. \tag{1.2}$$

This likelihood function has been used in Bayesian logistic regression analysis, and in similar Bayesian models of relation learning described by Silva, Airoldi, and Heller (2007) and Silva, Heller, and Gharamani (2007). The logistic function is also commonly used in neural networks to introduce nonlinearity into activation functions.

We assume that the prior $P(\mathbf{w})$ in Eq. (1.1) follows a multivariate normal distribution, $P\left(\mathbf{w}\right) \sim N(\boldsymbol{\mu}_0, \mathbf{V}_0)$, with a mean of $\boldsymbol{\mu}_0$ and a covariance matrix of $\mathbf{V}_0$. A primary focus of the

present paper is on the potential role of informative priors in relation learning. The key to efficient statistical learning is a good choice of priors, especially when the learning problem involves high dimensionality. Proposals for priors typically stem from abstract theory (Griffiths & Tenenbaum, 2009; Kemp & Tenenbaum, 2008; Lu et al., 2008) or analyses of statistics of the natural environment (Geisler, 2008; Simoncelli & Olshausen, 2001; Griffiths & Tenenbaum, 2006; Lu et al., 2010). Here we explore a variation of what are termed *empirical priors*, which are themselves learned from relevant data, combined with a *hyperprior* for variances of weights.

Empirical priors. BART takes advantage of the potential for inductive bootstrapping, using previously-acquired knowledge of simpler concepts to establish empirical priors that guide subsequent learning of more complex concepts. Previous work has explored use of a general relation (*related*) as an empirical prior for learning more specific relations (Chen et al., 2010; Silva, Airoldi & Heller, 2007). Here we consider the potential usefulness of more specific empirical priors tailored to individual relations. There is strong linguistic evidence (across many languages) that two-place comparatives are derived from corresponding relative adjectives either by adding a morpheme (e.g., *large* yields *large* + *er*, termed the *synthetic* form) or by creating a phrase using *more* or *less* (e.g., *intelligent* yields *more intelligent*, termed the *periphrastic* form; see Graziano-King & Cairns, 2005). Psychological evidence also indicates that comparative relations such as *higher* are initially derived from the corresponding one-place predicates (e.g., *high*; see Smith, Rattermann & Sera, 1988). In choosing the appropriate priors, it seems probable that children are guided by lexical similarities (e.g., *larger* is similar to *large*, *smaller* to *small*). However, to increase the generality of the model, we make the weaker assumption that the learner must infer the most relevant one-place predicate from the actual pairs used as positive training examples for the comparative.

We use one-place predicates as the building blocks for creating empirical priors. To determine which one-place predicate should be used to construct the empirical prior for learning a particular relation, we developed a simple categorization algorithm to select the one-place predicates based on training data. First, we train BART on the eight categories of one-place predicates (e.g., *large*, *small, fierce, meek*) that can be formed using the *extreme* animals at each end of the four different magnitude continua (size, speed, fierceness and intelligence). For example, we used the 20 largest animals (e.g., whale, dinosaur, elephant) to learn the category of large animals, and the 20 smallest animals (e.g., flea, fly, worm) to learn the category of small animals. As schematized in Figure 1.2, category learning of one-place predicates is conducted using Bayesian logistic regression, with a standard normal distribution for weights (i.e., mean 0 and variance 1) as the prior to infer the weight distribution $P(\mathbf{w}_c \mid \mathbf{X}_c)$, in which $\mathbf{w}_c$ indicates the weight vector corresponding to feature dimensions of an object, and $\mathbf{X}_c$ denotes the extreme animals in each group used for category learning.

Second, we employ a simple voting procedure to select the "best" category of one-place predicates based on the training examples for the comparative. For each pair of objects $\left(X^A, X^B\right)$ in the training data for relation learning, we compute the probability that each individual object is a member of each category of one-place predicates, obtaining $P\left(C \mid X^A\right)$ and $P\left(C \mid X^B\right)$, respectively. If $P\left(C \mid X^A\right) > P\left(C \mid X^B\right)$ for a pair, a score of 1 is assigned to this category; otherwise, a score of 0 is assigned. These scores are summed over all the pairs of training data. In effect, the procedure for prior selection aims to identify the one-place predicate that best distinguishes the objects in the two relational roles (i.e., the category of which the first object is maximally more likely than the second to be a member). The reliability of prior

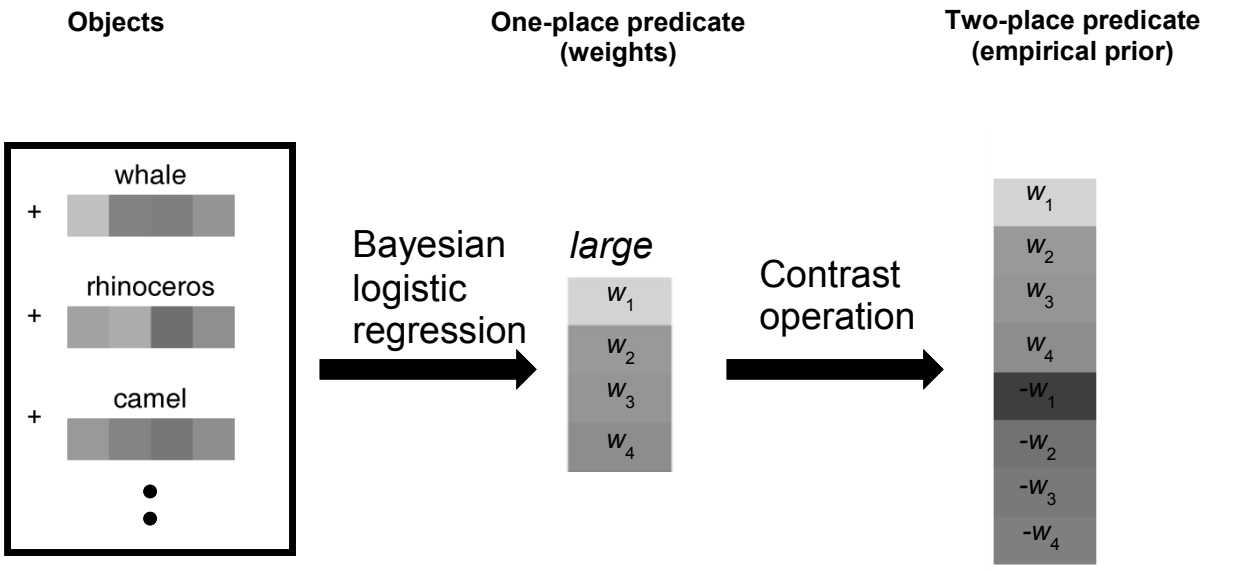


Figure 1.2. Illustration of the construction of an empirical prior for a comparative relation (*larger*) by bootstrapping from prior learning of weights for a related one-place predicate (*large*), in turn derived from features of individual objects (large animals).

selection will naturally vary with the number of training examples, yielding an inherent source of variability in the acquisition of the relations. Although more sophisticated categorization models could be employed, this simple procedure proved adequate for our present purposes. For the most difficult set of inputs (topics), the method achieved near-perfect selection of the appropriate one-place predicate when given 100 training examples.

Third, the category that yields the highest summed score is selected to set mean weights for the first role of a comparative. Although the model is in effect informed that the relation to be learned involves a comparison of two objects, the basis for the comparison must be learned. The potential priors on the second role are linked to those for the first role, by reversing the sign on the weights for the first role to form a contrast.[3] For our example, if *large* were to provide the basis for the empirical priors, then the priors for the comparative relation would include the

weights of *large* for the first role and the opposite weights for the second role, as shown in Figure 1.2. If there is a tie in the highest summed score between categories of one-place predicates, we simply take an average of the multiple weight vectors to generate empirical priors.

Hyperprior. Confidence about the empirical priors bootstrapped from object concepts is represented by the variances in the prior distribution of weights. A simple model is to assume the same degree of confidence for all the individual weights in the empirical prior $\boldsymbol{\mu}_0$. Alternatively, confidence may vary from one dimension to another, affording greater flexibility. We adopt the method of automatic relevance determination (MacKay, 1992; Neal, 1996) to define the precisions of the empirical prior using hyperparameters. Specifically, the prior for the *i*th weight in vector $\mathbf{w}$ is assigned in the form of a normal distribution in which the mean is from the empirical prior and the variance is $1/\alpha_i$:

$$P(w_i|\alpha_i) \sim N\left(\mu_{0i}, \frac{1}{\alpha_i}\right), \tag{1.3}$$

where the value of α_i (also termed precision, the inverse of variance) controls the certainty about mean weight values derived from the empirical prior. Thus increasing α_i values imply greater confidence that w_i is similar to μ_{0i} in the empirical prior. We use a conjugate prior distribution in the form of a Gamma distribution for α_i with two hyperparameters, a_0 and b_0, to constrain the precision of each weight:

$$P(\alpha_i) \sim Gamma(a_0, b_0). \tag{1.4}$$

Inference algorithm. Although the general framework of the relation learning model is straightforward, the inference step is non-trivial because the calculation of the integral in Eq. (1.1) lacks an analytic solution. A sampling approach is impractically slow for dealing with high feature dimensionality, and hence would unduly limit the generality of the model. Accordingly,

as in Silva, Heller, and Gharamani (2007), we employed the variational method developed by Jaakkola and Jordan (2000) for Bayesian logistic regression to obtain a closed-form approximation to the posterior distribution. Variational methods are a family of methods that transform the problem of interest into an optimization problem by introducing an extra *variational parameter*, ξ, which is iteratively adjusted to obtain successively improving approximations. The input to the learning model includes training data **X**, composed of *N* training pairs and their corresponding relation labels *R* in which 1 indicates that the pair of words instantiates the relation (positive examples), and -1 indicates it does not (negative examples). The variational updates are applied until convergence or a maximum number of iterations is reached. For learning with an empirical prior, the model starts from the prior mean $\boldsymbol{\mu}_0$, (i.e., bootstrapping from knowledge about the corresponding one-place predicates), and with $\mathbf{V}_0$ assumed to be an identity matrix with variances 1 and covariances 0. On each iteration the variational parameter ξ is updated, along with the mean of the weight vector, $\boldsymbol{\mu}$, and the covariance matrix, **V**, with the following updating equations:

$$\begin{aligned} \mathbf{V}^{-1} &= \mathbf{V}_0^{-1} + 2\sum_{n=1}^{N} \lambda(\xi_n)\mathbf{x}_n\mathbf{x}_n^T \\ \boldsymbol{\mu} &= \mathbf{V}\left[\mathbf{V}_0^{-1}\boldsymbol{\mu}_0 + \sum_{n=1}^{N} R_n\mathbf{x}_n/2\right] \\ \xi_n^2 &= \mathbf{x}_n^T[\mathbf{V} + \boldsymbol{\mu}\boldsymbol{\mu}^T]\mathbf{x}_n \end{aligned} \tag{1.5}$$

where $\lambda(\xi) = \tanh(\xi/2)/(4\xi)$.

For learning with a hyperprior, the variational method is iteratively applied to update the mean, the covariance matrix, and the hyperparameters, as follows:

$$\mathbf{V}^{-1} = \mathrm{E}_{\boldsymbol{\alpha}}(\mathbf{A}) + 2\sum_{n=1}^{N}\lambda\left(\xi_n\right)\mathbf{x}_n\mathbf{x}_n^T$$
$$\boldsymbol{\mu} = \mathbf{V}\left[\mathrm{E}_{\boldsymbol{\alpha}}(\mathbf{A})\boldsymbol{\mu}_0 + \sum_{n=1}^{N} R_n\mathbf{x}_n/2\right] \tag{1.6}$$
$$a = a_0 + 1/2$$
$$b_i = b_0 + \left((w_i - \mu_{0i})^2 + V_{ii}\right)/2$$
$$\xi_n^2 = \mathbf{x}_n^T\left[\mathbf{V} + \boldsymbol{\mu}\boldsymbol{\mu}^T\right]\mathbf{x}_n$$

where w_i is the ith element of weight vector $\mathbf{w}$, μ_{0i} is the ith element of empirical prior $\boldsymbol{\mu}_0$, V_{ii} is the ith diagonal element of covariance matrix $\mathbf{V}$, and $\mathrm{E}_{\boldsymbol{\alpha}}(\mathbf{A})$ is a diagonal matrix with its ith diagonal element given by a/b_i.

Model evaluation on generalization test. To test generalization of the learned relational representation, we conduct a transfer task using new pairs of words, denoted by the subscript T. Given the training pairs $\mathbf{X}_\mathbf{S}$ and their labels $\mathbf{R}_\mathbf{S}$, the model aims to calculate the posterior predictive probability that a target pair $\mathbf{X}_T$ instantiates the learned relation:

$$P\left(R_T = 1 \mid \mathbf{X}_T, \mathbf{X}_\mathbf{S}, \mathbf{R}_\mathbf{S}\right) = \int_\mathbf{w} P\left(R_T = 1 \mid \mathbf{X}_T, \mathbf{w}\right) P\left(\mathbf{w} \mid \mathbf{X}_\mathbf{S}, \mathbf{R}_\mathbf{S}\right). \tag{1.7}$$

The posterior predictive probability can be approximated using the variational posterior (i.e., the lower bound of the predictive probability), which can be computed in a single pass through the training data set $\{\mathbf{X}_\mathbf{S}, \mathbf{R}_\mathbf{S}\}$, applying the updating equations as specified in Eq. (1.5). Hence, the probability predicted for a transfer pair (i.e., Eq. (1.7)) can be approximated as

$$\begin{aligned}\log P\left(R_T = 1 \mid X_T, \mathbf{X}_\mathbf{S}, \mathbf{R}_\mathbf{S}\right) = \log\left(g\left(\xi_T\right)\right) - \frac{\xi_T}{2} + \lambda\left(\xi_T\right)\xi_T{}^2 \\ -\frac{1}{2}\boldsymbol{\mu}_\mathbf{S}{}^T\mathbf{V}_\mathbf{S}^{-1}\boldsymbol{\mu}_\mathbf{S} + \frac{1}{2}\boldsymbol{\mu}_T^T\mathbf{V}_T^{-1}\boldsymbol{\mu}_T + \frac{1}{2}\log\frac{|\mathbf{V}_T|}{|\mathbf{V}_\mathbf{S}|},\end{aligned} \tag{1.8}$$

where $\boldsymbol{\mu}_S$ and $\mathbf{V}_S$ denote the parameters in $P(\mathbf{w} \mid \mathbf{X}_S, \mathbf{R}_S)$ after learning from the training pairs, and $\boldsymbol{\mu}_T$ and $\mathbf{V}_T$ denote the parameters in $P(\mathbf{w} \mid \mathbf{X}_S, \mathbf{R}_S, \mathbf{X}_T, R_T = 1)$, found by adding the target pair to the training set.

Higher-order relation mapping. In order for any model to have a chance to solve higher-order relational analogies (i.e., analogies based on relations between relations), it must first acquire at least approximate representations of the relevant first-order relations. However, as the example in Figure 1.3 makes clear, successful learning of comparative relations will not in itself guarantee solution of analogy problems such as *larger:smaller :: fiercer:meeker*. For example, if we were to compare the learned distributions for *larger+smaller* to those for *fiercer+meeker*, we would find that the two joint distributions are essentially uncorrelated.

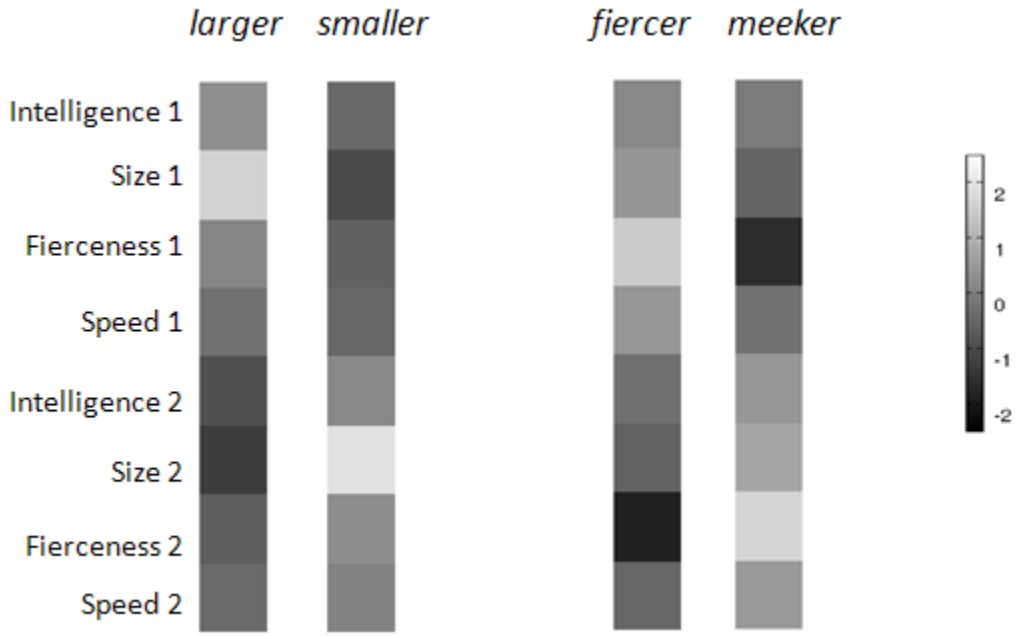


Figure 1.3. Successful learning of comparative relations is not sufficient to solve 4-term analogy problems such as *larger:smaller* :: *fiercer:meeker*, because high (positive or negative) weights are on different dimensions for *larger+smaller* versus *fiercer+meeker*.

Roughly, in the former the size dimension has large (positive or negative) weights while the other dimensions have weights near zero, whereas in the latter the fierceness dimension has large weights and the rest have weights near zero. Without any mechanism to map different salient dimensions to one another, any implicit similarity would remain hidden.

To solve higher-order analogy problems, BART employs an algorithm for *importance-guided mapping*. In general terms, the algorithm aims to find a mapping between the dimensions for the A:B relation and those for the C:D relation that minimizes a distance measure defined over the weight distributions. Because the full search space for this correspondence problem scales exponentially with the number of dimensions, we employ a greedy search algorithm (Friston et al., 2008), a type of procedure designed to make locally optimal choices with the hope of approximating the global optimum. More specifically, the algorithm develops a one-to-one mapping between dimensions sequentially on the basis of the overall "importance" of dimensions. In essence, the algorithm minimizes correspondence errors for more important dimensions at the possible cost of greater errors for less important dimensions.

In more detail, we assume that an analogy problem in the form A:B::C:D is evaluated by first focusing on the relation in the source (A:B), and then determining how well the target relation (C:D) maps to A:B. The algorithm prioritizes dimensions in proportion to their importance in A:B. Specifically, the mapping algorithm first searches for a dimension in C:D that is most similar to the most important dimension in A:B; it then searches for a dimension that maps to the second most important dimension in A:B among the remaining pool of dimensions in C:D, and so on until each dimension in A:B is mapped to a unique dimension in C:D. Qualitatively, BART aims to map important dimensions in A:B to dimensions in C:D that influence relation classification in an analogous way.

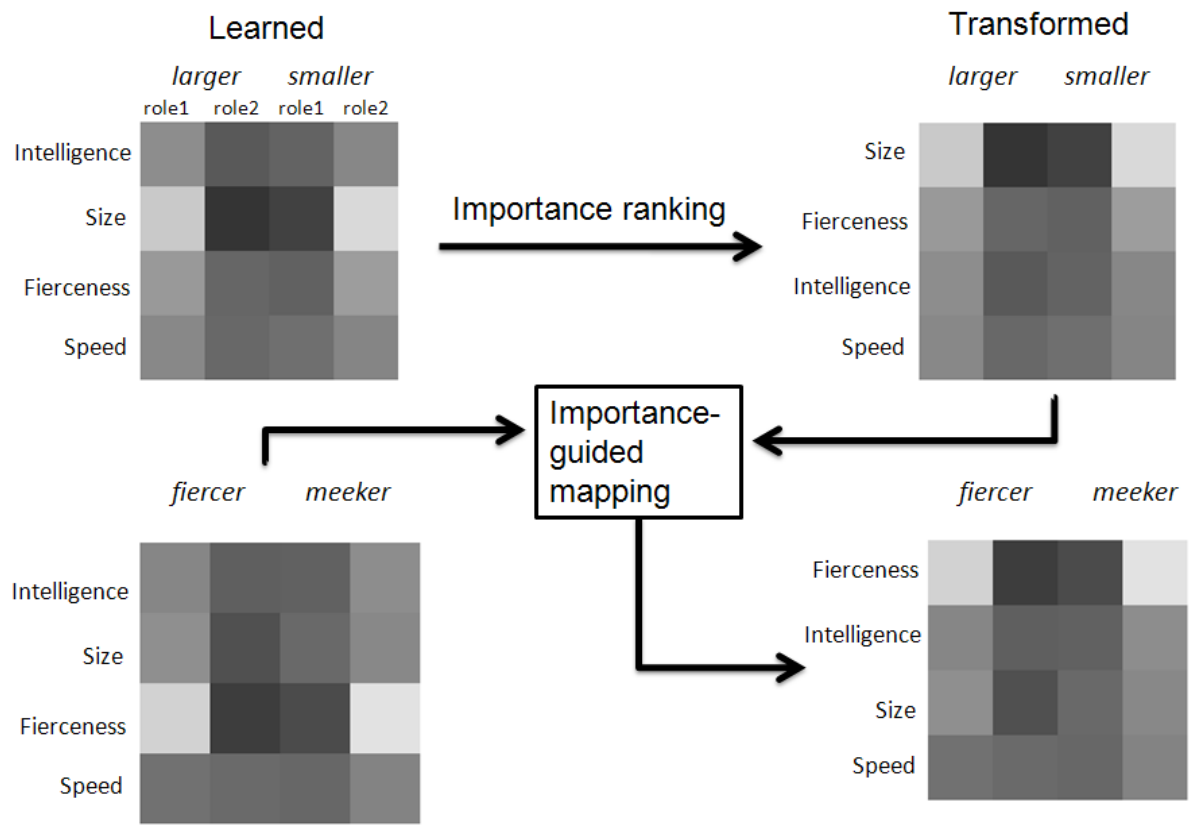


Figure 1.4. Illustration of importance-guided mapping for solving an analogy problem.

Figure 1.4 schematizes the algorithm for importance-guided mapping. Intuitively, *larger+smaller* and *fiercer+meeker* are alike in that each has a key important dimension (size and fierceness, respectively). Moreover, "importance" has a clear numerical definition based on the absolute magnitudes of weights (normalized by their variances). To evaluate an analogy in the form A:B::C:D (e.g., *larger:smaller* :: *fiercer: meeker*), the model first assesses the importance of each dimension for A:B, and then reorders the dimensions (and transforms the distributions of mean weights) accordingly. The transformed representation of A:B can be obtained in three steps: (1) compute the normalized weights using mean weight values divided by their standard deviations; (2) sum up the absolute values across the two roles in each of the

two relations in A:B to get an importance score for each feature dimension; (3) rank order the dimensions (maintaining consistency across the two roles of both relations) based upon the importance index.

Next, for each dimension in A:B, BART selects the dimension in C:D with the most similar pattern of weight distributions. Here we take advantage of a natural property of multivariate normal distributions. The marginal distribution over a subset of multivariate normal random variables can be obtained by dropping the irrelevant variables (the variables that one wants to marginalize out) from the mean vector and the covariance matrix. The marginal weight distributions for each feature dimension can therefore be easily calculated for A:B and C:D, respectively. Then the similarity of marginal distributions is evaluated by computing a distance measure between two distributions.[4] The *J-divergence distance* is employed to maintain the symmetric property of a distance measure by summing up two Kullback-Leibler (KL) divergences (Cover & Thomas, 2006),

$$D(p,q) = KL(p \,\|\, q) + KL(q \,\|\, p) \tag{1.9}$$

where p and q denote two distributions, and $KL(p\,||q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx.$ The advantage of using normal distributions is that it becomes possible to solve analytic expressions for the distance measure using the means and covariance matrices of the two normal distributions,

$$D(p,q) = \frac{1}{2}\left(\mu_p - \mu_q\right)^T \left(V_p^{-1} + V_q^{-1}\right)\left(\mu_P - \mu_q\right) + \frac{1}{2} tr\left[V_p^{-1}V_q + V_q^{-1}V_p - 2I_d\right] \tag{1.10}$$

where $tr[\cdot]$ denotes the matrix trace.

Finally, having transformed the C:D distribution to reflect its mapping to A:B, BART uses the overall J-divergence distance between the two transformed distributions as its measure of how well the C:D relation matches that of A:B. For our example, the transformed

representations will identify size as the most important dimension for *larger+smaller*, and then select fierceness as the dimension for *fiercer+meeker* that has the most similar mean weight distribution to that of size for *larger+smaller*. The resulting transformed distributions will map size to fierceness, thereby contributing to a lower overall J-divergence distance (i.e., higher similarity) for A:B::C:D than for a C-D' foil such as *fiercer+slower*.

Note that BART evaluates relational analogy problems without forming an explicit representation of a higher-order relation such as *opposite*. Rather, BART estimates the degree of match between the C:D and A:B relations under the assumption that similar relations (whatever they may be) will generate lower J-divergence distance between the two mean weight distributions based on the correspondences produced by importance-guided mapping. In the General Discussion we will consider how an extension of BART might go on to acquire explicit representations of higher-order relations.

Tests of BART Using Ratings Inputs

Inputs

The rating vectors used as inputs to BART were based on norms reported by Holyoak and Mah (1981), collected for use in a study of symbolic magnitude comparisons. Holyoak and Mah had 25 undergraduates rate the subjective magnitude of each of 80 animal names on four continuous dimensions: size, speed, fierceness and intelligence. Ratings were made on a 9-point Likert scale, with a rating of 9 indicating maximum magnitude. Magnitude norms for each dimension were then derived by successive interval scaling (Bock & Jones, 1968). This method provides a simultaneous normalization of the responses to each item across the nine response categories, yielding what can be interpreted as an interval scale. The resulting values (which for each dimension correlated .99 with mean ratings) were normalized to range from 0 through 10.

A few examples are shown here in Table 1.1. (See Holyoak & Mah, 1981, Table 1, p. 200, for the entire set of norms.) Because the ratings reflect subjective magnitude differences, the norms incorporate the typical non-linear relationship between subjective and objective magnitudes (e.g., the norms indicate the difference in subjective size between a goldfish and a cat is roughly the same as that between a deer and a hippopotamus). Of the 80 animals in the norms, topics representations were available for 77, and all simulations were based on this subset. Intercorrelations among the four dimensions across the 77 animals were moderate, ranging from .38 (size with speed) to .60 (size with fierceness).

Table 1.1

Examples of Ratings of Animals on Four Dimensions of Magnitude (from Holyoak & Mah, 1981, Table 1, p. 200)

Animal	size	fierceness	intelligence	speed
alligator	5.46	8.88	3.67	5.03
cow	6.52	3.95	3.35	4.59
flea	0.00	2.52	0.24	3.65
goldfish	1.91	1.35	1.45	4.18
moose	7.04	5.98	3.96	6.29
mouse	2.41	3.08	3.34	5.02

Each of the 77 words thus initially corresponded to a vector of four continuously-valued features, with all values being non-negative. However, the logistic likelihood function used by BART is designed to map values between negative and positive infinity onto the outcome variable, with the value of 0 serving as the natural midpoint of the input scale. Accordingly, we centered the rating vectors by a linear transformation, subtracting from each value the mean

value for that dimension across all 77 words. Thus the feature values in the vectors used as inputs to BART included both negative and positive values with means of 0.

In our tests, both training and test items were created by randomly selecting pairs of animals and concatenating their rating vectors. Thus, each input vector had two components: the four features of animal 1 and the four features of animal 2. To ensure that the differences in magnitudes between animals in a pair were likely to be distinguishable by humans, we constrained all training and test pairs to be based on animals differing by at least 0.50 on the relevant dimension. Under this criterion, over 2000 animal pairs were available as positive examples for each to-be-learned relation.

Training

For the purpose of generating empirical priors, the 20 animals that were "greatest" and "least" on each dimension were first used to train BART to classify each of the 8 possible one-place predicates (i.e., *large*, *small*, *fierce*, *meek*, etc.) For this initial phase of learning, the priors on all weights were set to standard normal distributions (i.e., means of 0, variances of 1, covariances of 0).

In learning two-place relations, we tested two models. The first model was a version of BART that selected empirical priors for means of weights based on one-place predicates as described earlier (e.g., the mean weights for *large* might be selected to provide the priors for the first role of *larger*, with the second role set by replicating the weights for *large* as a contrast). Priors for variances were set to 1 and those for covariance were set to 0. (Because the learning task with rating inputs proved to be extremely easy for BART, a hyperprior was not used in these simulations.) For comparison, a baseline model simply used uninformative priors (standard normal distributions). We trained and tested BART on each of the eight comparative relations

involving the animal ratings (*larger*, *smaller*; *fiercer*, *meeker*; *smarter*, *dumber*; *faster*, *slower*). Assuming training examples are randomly sampled from the same population, the solutions for polar-opposite relations (e.g., *larger* and *smaller*) would be expected to converge at asymptote with symmetrical weight distributions (i.e., distributions with weights reversed between the two roles). This result was clearly obtained, so we will only report generalization results for the four "greater" relations. However, the analogy results are based on learned representations of all relations ("lesser" as well as "greater").

Generalization Performance

Basic tests. On each run, we trained the model on some number (1-100) of randomly selected pairs that constituted positive examples of the target relation (and satisfied the minimum difference criterion). All the remaining pairs in the pool (both positive and negative examples) were then used as test pairs. For test pairs, negative examples were created by simply reversing the "correct" order of the two animals for the target dimension. The number of test pairs that instantiated a relation was always equal to the number that did not instantiate it (since they involved the same animals in reverse order).

A test pair that instantiated the relation was counted as correct if its posterior predictive probability of being an example of the relation was greater than 0.5, whereas a test pair that did not instantiate the relation was counted as correct if its predicted probability was less than 0.5. This criterion assumes that the model is unbiased. When trained solely with positive examples, it is plausible that a learning model might develop an overall bias favoring a "yes" response. Based on signal detection theory, sensitivity after correcting for possible bias can be measured using the A_z measure (Dorfman & Alf, 1969), which calculates the area under the receiver operating characteristic (ROC) curve. For the ratings data, the criterion of 0.5 in fact proved to be optimal

prior to reaching ceiling accuracy in generalization performance, indicating that BART's generalization decisions were unbiased within this range. Accordingly, we will simply report percent correct.

All reported results are based on the average performance over 100 runs, each of which randomly selected a set of training pairs from the pool. Figure 1.5 depicts BART's generalization curves for the four "greater" relations as a function of the number of training examples. Learning was very successful for all relations. The BART model with empirical priors generalized moderately accurately after a single training example (mean of 71% correct across all relations), and reached 96% correct after 20 training trials. BART's learned representations of one-place predicates thus provided effective empirical priors for the two-place comparative relations. The baseline model with uninformative priors (means of 0) started at a substantially lower level of

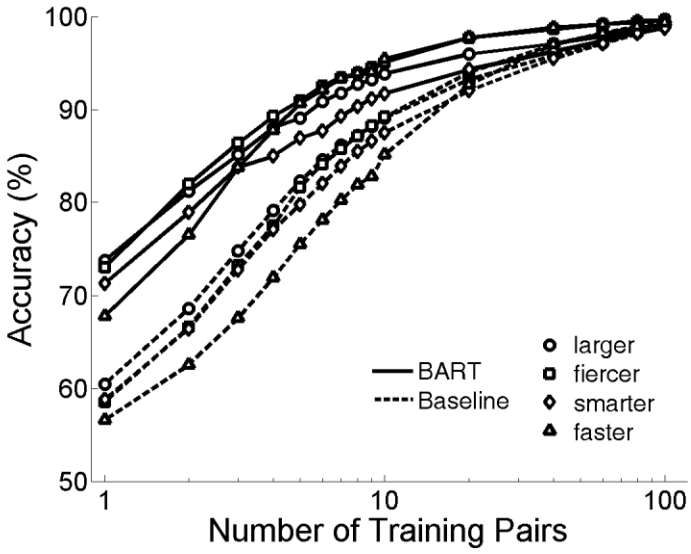


Figure 1.5. Accuracy in the generalization task with rating vectors as a function of the number of training examples for the four comparative relations (log scale). Solid lines indicate the performance of BART using the empirical prior; dashed lines indicate the performance of a baseline model (Bayesian logistic regression model with uninformative prior).

performance (mean of 59%), and required about twice as many training examples (40) to reach 95% accuracy. After 100 training examples, both models converged at near-perfect accuracy (99% correct) in generalization. These results demonstrate that at least when magnitude information is transparently coded in small input vectors, BART can learn comparative relations very efficiently from a modest number of positive examples, especially when guided by empirical priors.

To determine whether the relational representations acquired by BART yield the ubiquitous symbolic distance effect obtained for comparative judgments by humans, we examined how BART's probability estimates (using the full model with empirical priors) relate to the rated subjective distance between each test pair of animals on the dimension of interest (i.e., size, fierceness, intelligence, or speed). Distance effects are generally revealed in reaction-time paradigms. Although BART does not provide a process model of speeded judgments, standard models of reaction time (e.g., Link, 1990) would predict that reaction time as a measure of judgment difficulty will have an inverse monotonic relationship to the log ratio of posterior probabilities that each ordering of a pair fits the indicated relationship (e.g., for a pair such as *elephant-horse*, a positive log ratio will indicate that elephant is larger than horse, with the predicted difficulty of the discrimination decreasing as the log ratio becomes increasingly positive).

Distances were grouped into five bins based on inter-item distance in ratings on the relevant continuum (i.e., animals very similar in size fell in bin 1, animals maximally different in size fell in bin 5). Distance bins are based on Holyoak and Mah's (1981) norms, in which values range from 0-10: bin 1 (distances between 0.5 and 2), bin 2 (distances 2-4), bin 3 (distances 4-6), bin 4 (distances 6-8), and bin 5 (distances 8-10). Figure 1.6 plots the log ratio of the predicted

posterior probability for each positive test pair compared to the predicted probability for the reversed pair as a function of distance between the pair after learning based on 40 training pairs, averaged across the four comparative relations. Consistent with a symbolic distance effect, the log ratio increases with distance.

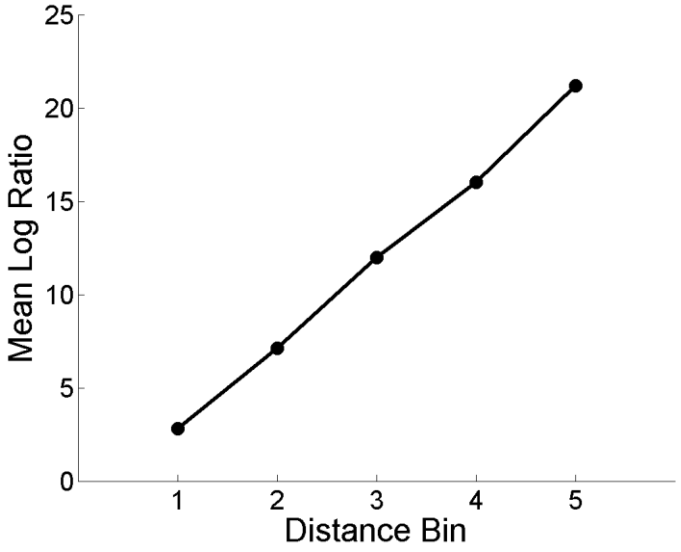


Figure 1.6. Log of the ratio between predicted posterior probability of each positive test pair instantiating a "greater" relation and that of the reversed pair instantiating the relation on generalization test (rating inputs) as a function of rated distance on the relevant continuum. Distance bins are based on Holyoak and Mah's (1981) norms, in which values range from 0-10: bin 1 (distances between 0.5 and 2), bin 2 (distances 2-4), bin 3 (distances 4-6), bin 4 (distances 6-8), and bin 5 (distances 8-10). Results are collapsed over the four continua.

Generalization beyond the range of training examples. The basic generalization tests described above always involved test pairs that had not been shown during training. We also performed a series of computational experiments to determine whether BART is capable of generalizing to new types of pairs that in various ways go beyond the range of the training examples.

(1) One test introduced pairs in a distance range outside of that used in training. Using empirical priors set in the same manner as described previously, we trained BART on the relation *larger* based on 40 positive examples drawn randomly from the first three distance bins only (e.g., 40 pairs of animals exhibiting small or moderate size differences for *larger*). We then tested the model's generalization performance at each of the five distance bins, using all possible animal pairs excluding the training pairs. We again obtained a monotonic increase in mean log ratio across all levels of distance: 2.80, 7.12, 12.00, 16.07, and 21.13 for bins 1 to 5, respectively. The model thus assessed pairs of animals with large size differences (bins 4-5) as the best positive examples of *larger*. BART's acquired representation of *larger* was sufficiently robust and flexible as to enable very accurate generalization to novel test pairs exhibiting size differences greater than the range presented during training.

(2) Another series of generalization tests varied the magnitudes of the individual training and test objects. For this purpose all the animals were sorted into four roughly equally-sized groups based on their value on the relevant dimension in the Holyoak and Mah (1981) norms, such that animals in group 1 have the lowest values and animals in group 4 have the highest values. We then trained the model with 100 examples based on pairs of the form [4, 1]. In other words, the first animal is drawn from group 4 and the second animal is drawn from group 1 (i.e., for learning *larger*, the first animal is very large and the second animal is very small). The generalization test included all and only pairs of the form [3, 2] (i.e., middle-sized animals). BART's performance was similar across the four "greater" relations with an overall accuracy of 91%, indicating very successful generalization.

(3) Because pairs of the form [3, 2] are necessarily close in magnitude, a generalization test that includes only pairs of the form [3, 2] is inherently more difficult than one composed of

pairs formed from all groups. For comparison with test (2), we also trained the model in the usual way (positive examples formed from all groups), and then tested it on only pairs of the form [3, 2]. BART performed similarly across the four "greater" relations on this test as well, achieving an overall accuracy of 99% after 100 training examples. In comparison, the 91% accuracy obtained in test 2 is somewhat lower, indicating that restricting the magnitude range of the training items impaired generalization to some extent.

(4) Another test involved training with 100 pairs of the form [2, 1] (i.e., pairs of small animals) and testing with those of the form [4, 3] (i.e., pairs of large animals), or the reverse (training on [4, 3], then testing on [2, 1]). This test is inherently difficult because the training items are drawn from a restricted range, and the test items are drawn from a different restricted range (and moreover, are very close in magnitude). Averaged across the two variations, generalization accuracy was 65%, 79%, 77%, and 81% for *larger*, *fiercer*, *smarter*, and *faster*, respectively. Thus transfer from one extreme on a continuum to the other was reliable although imperfect.

(5) A final test ensured that the animals (not just pairs) used during training and testing did not overlap by selecting a random half of the animals for training, and then testing on all pairs formed by the remaining animals. After 100 training examples, BART achieved 98% overall accuracy, indicating very successful generalization to animals not encountered during training.

Analogy Performance

To test BART's ability to solve higher-order analogy problems using its acquired relational representations, we constructed problems based on the comparative relations. If the model is able to implicitly learn relations between relations, then its standard training on the four

sets of paired comparatives should allow it to solve analogies based on two distinct higher-order patterns, which we will gloss as "same-extreme" (e.g., the relationship of *larger* to *fiercer*, or *smaller* to *meeker*; see Clark, 1970, for a discussion of the polarity of comparative relations) and "opposite" (e.g., the relationship of *larger* to *smaller*, or *fiercer* to *meeker*). Table 1.2 gives examples of five types of 4-term analogy problems that can be constructed by pairing one of the two higher-order relations with various foils, using the first-order relations acquired by BART. The first types are based on *same-extreme*, with the foil being either an opposite pair (Same-O) or a pair of relations at the opposite extreme of their respective dimensions (Same-OE). The other three types are based on opposite. The foil could be split across two dimensions (Opp-S), reverse polarity on a dimension (Opp-R), or involve a conflict (Opp-C) in which one relation in the foil was in fact identical to one of the A:B terms. In such problems the analogical answer C:D has to overcome the misleading featural identity of the D' term in the C:D' foil to the B term in A:B. Except for Same-OE problems, the C:D' (or C':D) foils always share one word with the analogical C:D completion.

For the first four types, chance performance would be 50% if the A:B and/or C:D relations had not been acquired. J-divergence, like other proposed measures of relational similarity that have been used to model human judgments (e.g., Goldstone, 1994; Taylor & Hummel, 2009) is sensitive to featural as well as relational overlap. For Opp-C analogies, expected performance in the absence of relation learning would therefore be 0%, because the featural overlap based on the word shared by A:B and the foil C:D' would cause the foil to always be selected as more similar to A:B. The Opp-C conflict set thus provided an especially challenging test of BART's ability to solve higher-order analogies based on its learned relational representations, directly pitting relational against featural similarity. Similar designs have been

Table 1.2

Examples of Analogies Based on the Relations "Same-Extreme" "and "Opposite," with Various Types of Foils (Number of Examples Used to Test BART is Indicated in Parentheses)

Analogy test type	Target	Foil
Same-O (48): *Same-extreme (*opposite as foil)	*larger* : *fiercer* :: *smarter* : *faster*	*larger* : *fiercer* :: *smarter* : *stupider*
Same-OE (48): *Same-extreme* (opposite extreme as foil)	*smaller* : *stupider* :: *meeker* : *slower*	*smaller* : *stupider* :: *faster* : *fiercer*
Opp-S (48): *Opposite* (split pair as foil)	*faster* : *slower* :: *smarter* : *stupider*	*faster* : *slower* :: *smarter* : *meeker*
Opp-R (24): *Opposite* (reversed as foil)	*larger* : *smaller* :: *smarter* : *stupider*	*larger* : *smaller* :: *stupider* : *smarter*
Opp-C (48): *Opposite* (conflict foil)	*fiercer* : *meeker* :: *smarter* : *stupider*	*fiercer* : *meeker* :: *smarter* : *meeker*

employed in studies of human analogical mapping both with adults (e.g., Markman & Gentner, 1993) and children (Richland et al., 2006).

To test BART's capacity to make analogical inferences, we created sets of each of the five types (see Table 1.2 for the number of each type). BART's assessment was counted as correct (i.e., as an analogical response) if the calculated J-divergence distance was lower for the analogical C:D pair than for the non-analogical foil, C:D' (or C':D). Figure 1.7 shows the performance of BART and the baseline model (both using the identical algorithm for importance-guided mapping) on the five types of analogy problems. Although both models performed extremely well after learning from ratings inputs, BART achieved slightly higher success after fewer training examples. The advantage of BART over the baseline model in efficiency of learning to solve analogy problems is most apparent for Type Opp-C. After three

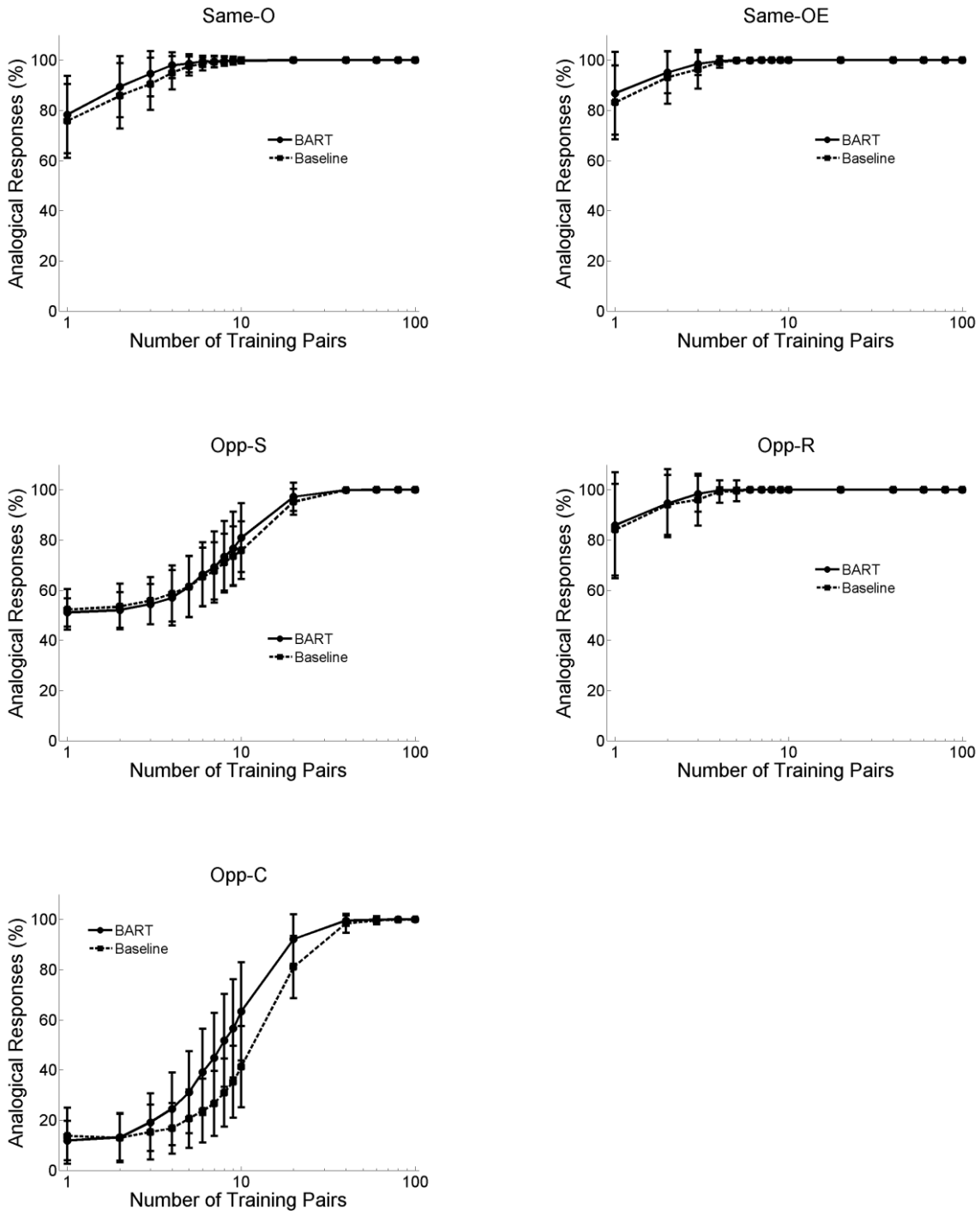


Figure 1.7. Proportion of analogical responses as a function of the number of training examples (log scale) with rating inputs for the five types of analogy problems. Solid lines present the results for BART with empirical prior; dashed lines present results for baseline model (with uninformative prior). Error bars indicate 1 standard deviation (results based on 100 runs).

training examples, BART begins to show more accurate performance than the baseline model, and the two models do not converge in their performance until after 60 training examples, at which point both models achieve essentially perfect performance on all problem types.

These results demonstrate that the algorithm for importance-guided mapping is in fact capable of solving structural analogies based on learned representations of first-order relations, implicitly finding correspondences between non-identical dimensions based on the importance-guided mapping operation using marginal weight distributions. Moreover, the empirical priors proved to be effective in establishing relational distributions that support analogical reasoning, especially in competition with featural similarity.

In summary, the tests with ratings vectors provide a first demonstration that BART is able to learn relational representations that pass five critical tests: (1) learning from non-relational inputs, (2) with high efficiency, (3) generalizing to new examples of first-order relations, (4) capturing a key source of differential difficulty in human performance, symbolic distance, and (5) supporting structured analogical reasoning.

Tests of BART Using Leuven Inputs

Inputs

We next applied BART to the much more challenging problem of learning comparative relations from high-dimensional input representations based on the Leuven database (De Deyne et al., 2008). As noted earlier, these norms are based on the frequency with which participants generated features characterizing 129 animals, for 759 features. To make the results as comparable as possible to those obtained with the ratings inputs, we used the subset of 44 animal names from the Holyoak and Mah (1981) norms that were also included in the Leuven database. Although this subset was substantially smaller than the 77 animals used in the simulations based

on ratings, it was still large enough to generate a pool of over 750 pairs for training and generalization tests.

To construct input representations we followed the procedure used by Kemp, Chang and Lombardi (2010, p. 219), who used the Leuven norms to estimate the probability of each feature conditional on each animal (see their Equation 8, top). We multiplied the computed probability by 100 to make the magnitude range roughly comparable to that of the rating inputs. Because the values as described so far are based on probabilities, they necessarily are non-negative. As noted in connection with the rating vectors, to optimize the scale for the logistic likelihood function it is desirable to center the vectors by a linear transformation. Accordingly, we subtracted from each feature the mean value of that feature across all 129 animals in the Leuven norms. The feature values in the vectors used as inputs to BART therefore included both negative and positive values, with means near 0.

In order to reduce the size of the search space, we focused on the most important dimensions. Specifically, we summed the feature vectors for the 44 animals and identified the 50 dimensions that yielded the largest sums (after dropping one dimension, "is small", that was clearly redundant with another, "is big"). By using just these 50 most important dimensions to form vectors for each individual word, the total size of the vector for each word pair was fixed at 100.

Training

The basic training regimen was very similar to that employed with the rating vectors. To create empirical priors, we again selected 20 animals close to each of the two extremes on each of the four dimensions of interest. These included all of the extreme animals included in the subset of 44 for which Leuven vectors were available (the number ranging from 8-15 across the

eight sets). We augmented this "core" group with additional animals from the entire Leuven set of 129 animals that we judged to be close to the relevant extremes, thus bringing the total number of animals in each set to 20. Insofar as some of the animals used to train one-place predicates may not have been the most extreme, and many were not included in the subset of 44 used to train relations, this procedure for selecting positive examples for learning empirical priors would be expected to make successful relation learning more challenging.

The search space for the Leuven representations was much larger relative to that for the ratings inputs used previously. Accordingly, we aimed to improve the stability of the estimates for empirical priors by increasing the number of examples. Given that the set of positive examples available for each one-place predicate was necessarily constrained, we augmented the training pool by including negative examples. To learn *large*, for example, BART was given both 20 positive examples (i.e., 20 large animals) and 20 negative examples (i.e., 20 small animals). As in the case of our simulations using rating vectors, direct training on each comparative relation (e.g., *larger*) was still based solely on positive examples.

To help cope with the greater complexity of the learning problem with high dimensionality, we used a hyperprior to increase BART's representational flexibility. Based on a preliminary search of the parameter space, we set the values of the hyperparameters (a_0, b_0) to be 5 and 1, respectively. We found that allowing BART to use the hyperprior (with hyperparameters fixed for all simulations) tended to improve its generalization performance by about 2 percentage points relative to using the standard covariance matrix (the procedure used in the simulations with rating inputs), and significantly improved accuracy in certain analogy tests. For comparison, we also tested the same baseline model as that used with ratings vectors (i.e., Bayesian logistic regression with standard normal distributions as uninformative priors).

Generalization Performance

Basic tests. All reported results are based on the average performance over 10 runs, each of which randomly selected a set of training pairs from the pool. Figure 1.8 depicts BART's generalization curves for the four "greater" relations as a function of the number of training examples. Not surprisingly, given the greatly increased dimensionality of the learning problem, the level of performance was lower overall than was obtained with the rating vectors. However, the full BART model, with empirical priors on mean weights and a hyperprior on variances, achieved substantial generalization (about 80-95% accuracy for the four "greater" relations after 100 training examples). The baseline model showed much weaker generalization performance, achieving only about 60-70% accuracy overall after 100 training examples.

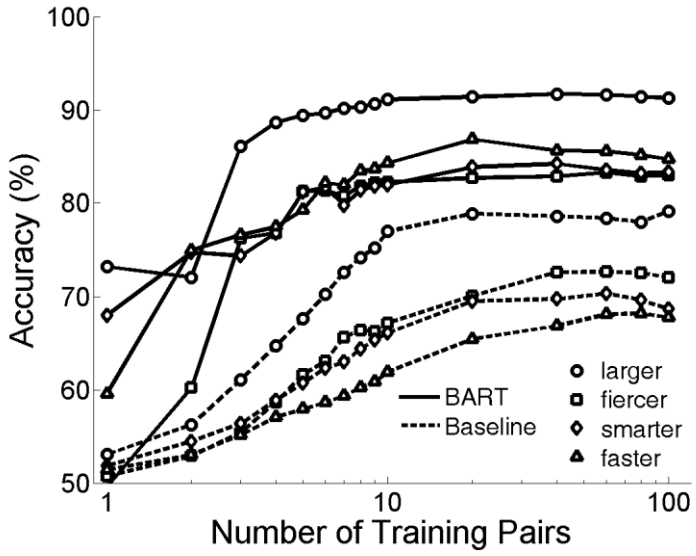


Figure 1.8. Accuracy in the generalization task with Leuven inputs as a function of the number of training examples (log scale) for the four comparative relations. Solid lines indicate the performance of BART using the empirical prior and hyperprior; dashed lines indicate the performance of a baseline model (Bayesian logistic regression model with uninformative prior).

We also explored how BART's generalization performance changed with more extended training. Whereas BART appeared to be unbiased when trained with up to 80-90 examples, a further increase in the number of training examples led to a bias towards "yes" responses. This type of response bias leads to reduced accuracy if a fixed decision criterion is used. Accordingly, we computed the A_z measure, which is more robust to response bias (Dorfman & Alf, 1969). Generalization performance as measured by A_z continued to improve slightly with increased numbers of training examples. After 700 training examples, both BART and the baseline model achieved an A_z value of about .95.

To examine whether the relational representations that BART derives from Leuven vectors yield the distance effect obtained for comparative judgments by humans, we examined how BART's generalization performance relates to the rated subjective distance between each test pair of animals on the dimension of interest (as measured using the Holyoak & Mah, 1981, norms). Figure 1.9 plots the mean log ratio of predicted probabilities for positive versus negative test pairs as a function of distance on the relevant dimension between the two animals in a pair (after learning from 100 training examples). Because only 44 of the animals in the Holyoak and Mah norms are included in the Leuven dataset, we used four distance bins instead of five. The log ratio of posterior probabilities increased monotonically with distance. Thus, the relational representations that BART acquired from Leuven inputs clearly yield a symbolic distance effect.

Generalization beyond the range of training examples. As in the case of the simulations based on ratings, we performed a series of computational experiments to determine whether BART is capable of generalizing to new types of pairs that in various ways go beyond the range of the training examples.

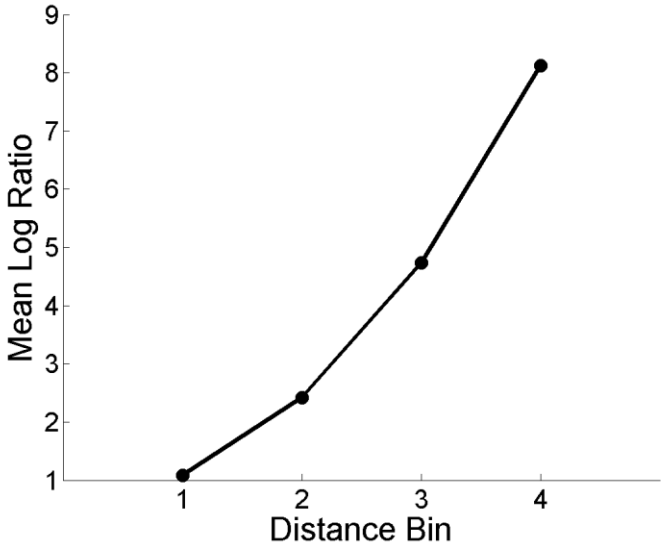


Figure 1.9. Log of the ratio between predicted posterior probability of each positive test pair instantiating a "greater" relation and that of the reversed pair instantiating the relation on generalization test (Leuven inputs) as a function of rated distance on the relevant continuum. Distance bins are based on Holyoak and Mah's (1981) norms: bin 1 (distances between 0.5 and 1.5), bin 2 (distances 1.5-3), bin 3 (distances 3-5.5), and bin 4 (distances 5.5-10). Results are collapsed over the four continua.

(1) *Training with 100 examples from distance bins 1-2 only and testing on all four distance bins.* We obtained a monotonic increase in mean log ratio across all levels of distance: 1.02, 2.30, 4.51, and 8.00 for bins 1 to 4, respectively. These results again demonstrate that the model assessed pairs of animals with large size differences, in bins 3 and 4, as the best positive examples of *larger*.

(2) *Training with 100 pairs of the form [4, 1] and testing on all pairs of the form [3, 2].* BART's accuracy was 89%, 67%, 73%, and 92% for *larger*, *fiercer*, *smarter*, and *faster*, respectively, indicating fairly successful generalization performance based on a restricted set of training inputs.

(3) *Standard training with 100 pairs formed from all groups and testing on all pairs of the form [3, 2]*. BART achieved accuracies of 87%, 69%, 77%, and 87% for *larger*, *fiercer*, *smarter*, and *faster*, respectively. Thus for the Leuven inputs, restricting the training pairs to those of the form [4, 1] (test 2) had minimal negative impact on generalization performance.

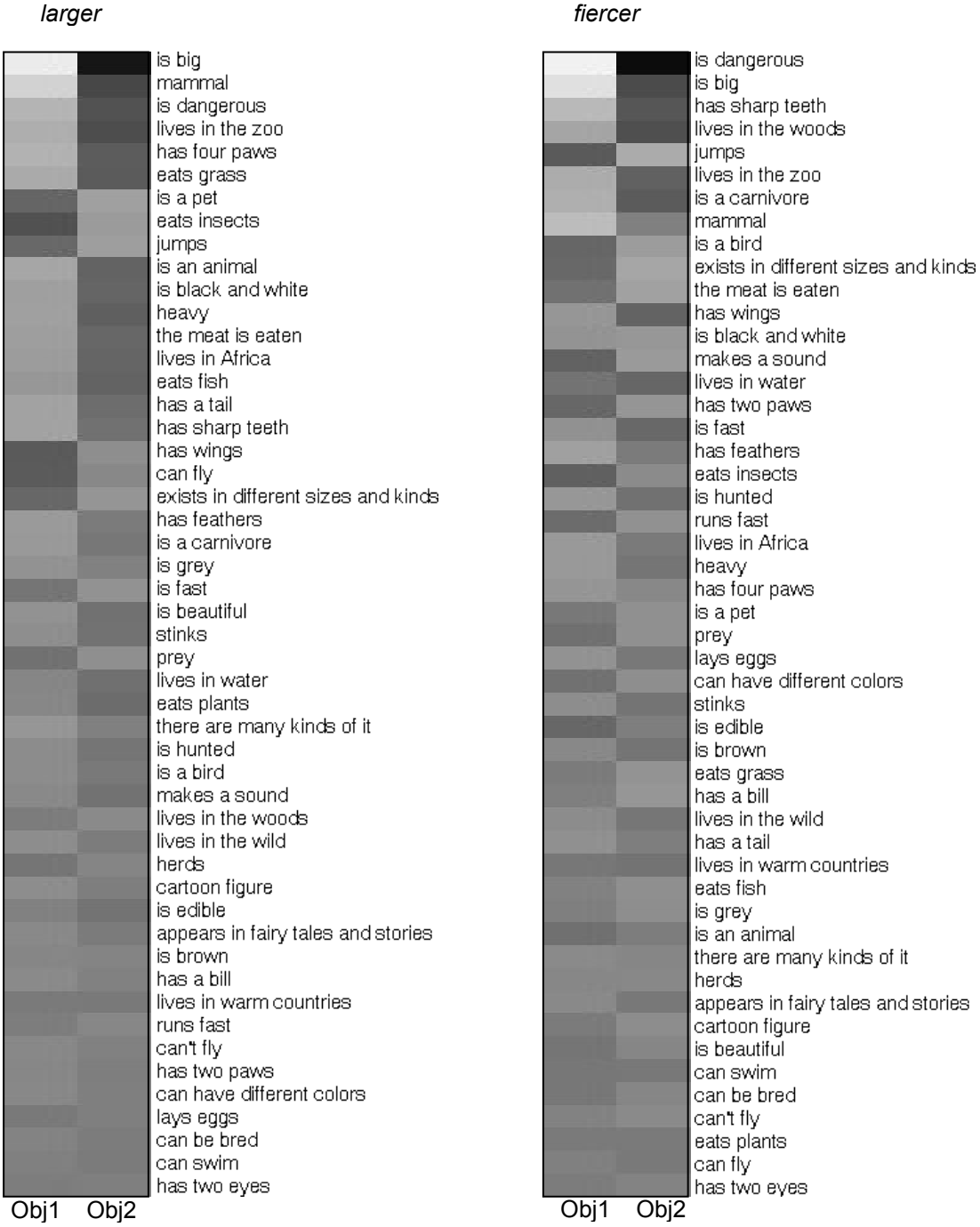
(4) *Training with 100 pairs of the form [2, 1] and testing on all pairs of the form [4, 3], or the reverse.* Averaged across the two variations, generalization accuracy was 85%, 73%, 77%, and 54% for *larger*, *fiercer*, *smarter*, and *faster*, respectively, indicating fairly successful generalization across magnitude extremes for the first three relations.

(5) *No overlap between training and test animals*. This test was performed with 60 training examples because the Leuven subset included only 44 animals in total. BART achieved accuracies of 92%, 79%, 84%, and 85% for *larger*, *fiercer*, *smarter*, and *faster*, respectively, indicating fairly successful generalization to animals not encountered at all during training.

Content of Learned Weight Distributions

To convey a sense of the content that BART used to learn comparative relations from the Leuven inputs, Figure 1.10 depicts typical mean weights for the four "greater" relations that the model acquired using 100 training examples. For each relation the 50 dimensions are ordered by importance. Several qualitative observations are of interest. First, the representations are clearly contrastive, with positive (light) weights associated with important weights on the first role and negative (dark) weights associated with the second role, or vice versa. Second, among the more important weights, the positive value is predominantly associated with the first role. This is the type of relational information that indicates to BART that these comparatives are in fact oriented toward the "greater" extremes of their respective continua.

Fig. 1.10 (2 pages)



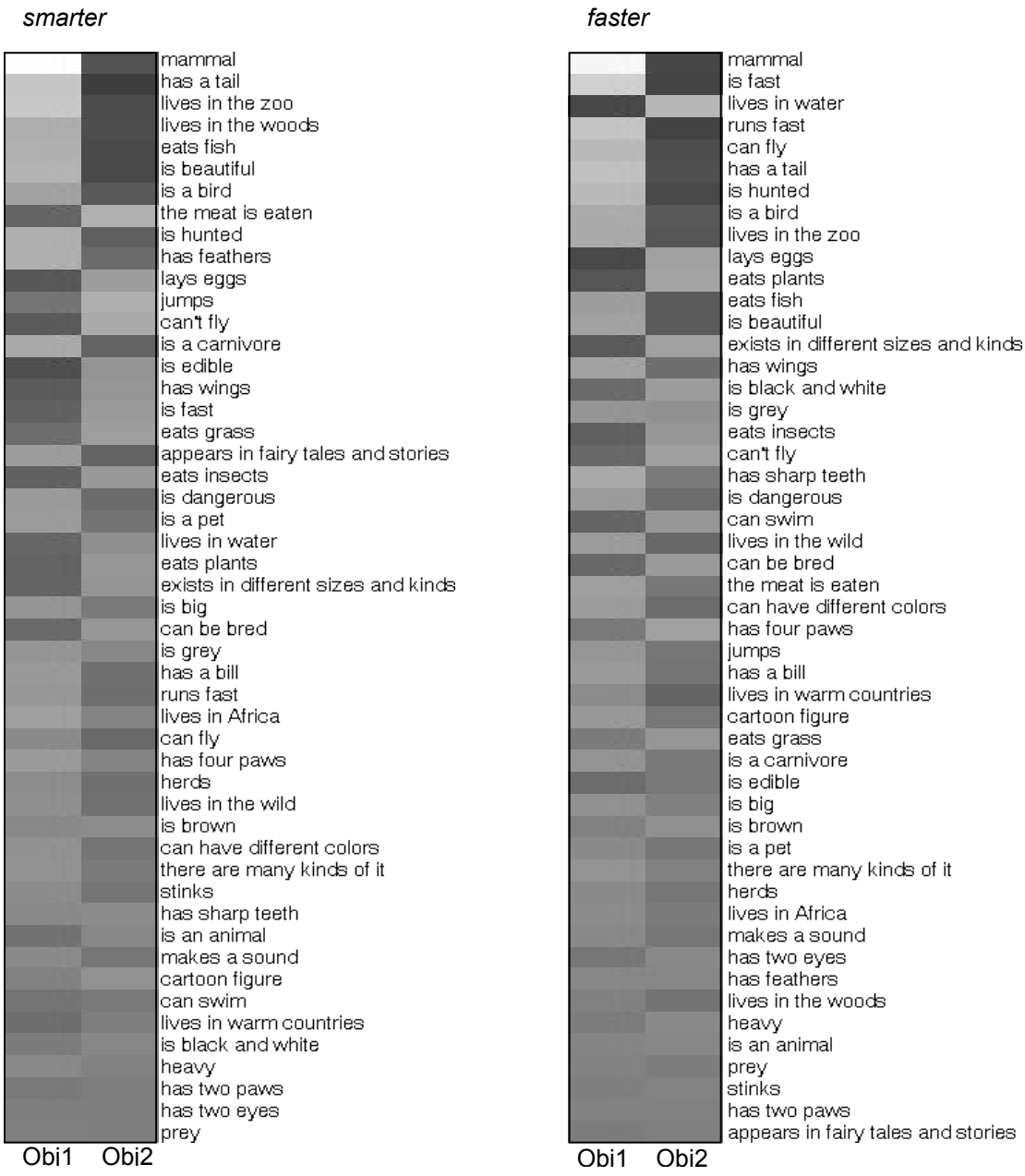


Figure 1.10. Illustration of mean weights for four relations learned from 100 training examples using Leuven inputs. For each relation, the weights on 50 dimensions (based on the specified query) are rank-ordered by importance. The intensity of cells represent weight values on each dimension (light indicates high positive values, dark indicates high negative values). The first column corresponds to weights on features of the first object, and the second column corresponds to weights on features of the second object.

Third, the representations are highly distributed. For each relation, upwards of 20 corresponding dimensions (i.e., 40 weights) show clear contrasts between the two roles. Unlike the rating vectors, in which a single dimension provided a localist code for each continua, the Leuven vectors lack any single dimension that suffices to define any comparative relation. To take the most salient example, one might have supposed that "is big" would be sufficient to predict relative size. In fact, although this dimension is indeed the single most important predictor of which object is larger, it is far from sufficient. The Leuven dimensions were derived from the frequencies with which participants generated features, rather than from a continuous rating scale of the sort used to create the Holyoak and Mah (1981) norms. Accordingly, in the Leuven dataset, animals for which size is a salient dimension (often in reference to a subcategory) tend to have higher features values for "is big". Based on a comparison of feature values on that dimension alone, the Leuven dataset indicates that (for example) an eagle is larger than a giraffe, a seagull is larger than a horse, and a cow is the same size as a pelican. However, BART is able to flexibly integrate weakly predictive information provided by dozens of individual dimensions to successfully learn and generalize the comparative relations.

Analogy Performance

The distributed nature of the relation representations acquired from the Leuven inputs posed a strong test of BART's algorithm for importance-guided mapping. Although this algorithm was extremely successful when applied to localist representations derived from the rating data, it was far from obvious whether it would also be effective with distributed representations. We tested BART and the baseline model on the five types of analogy problems in the same manner as for the rating inputs. The results are shown in Figure 1.11. The overall level of performance is lower than was obtained when the models were trained with ratings

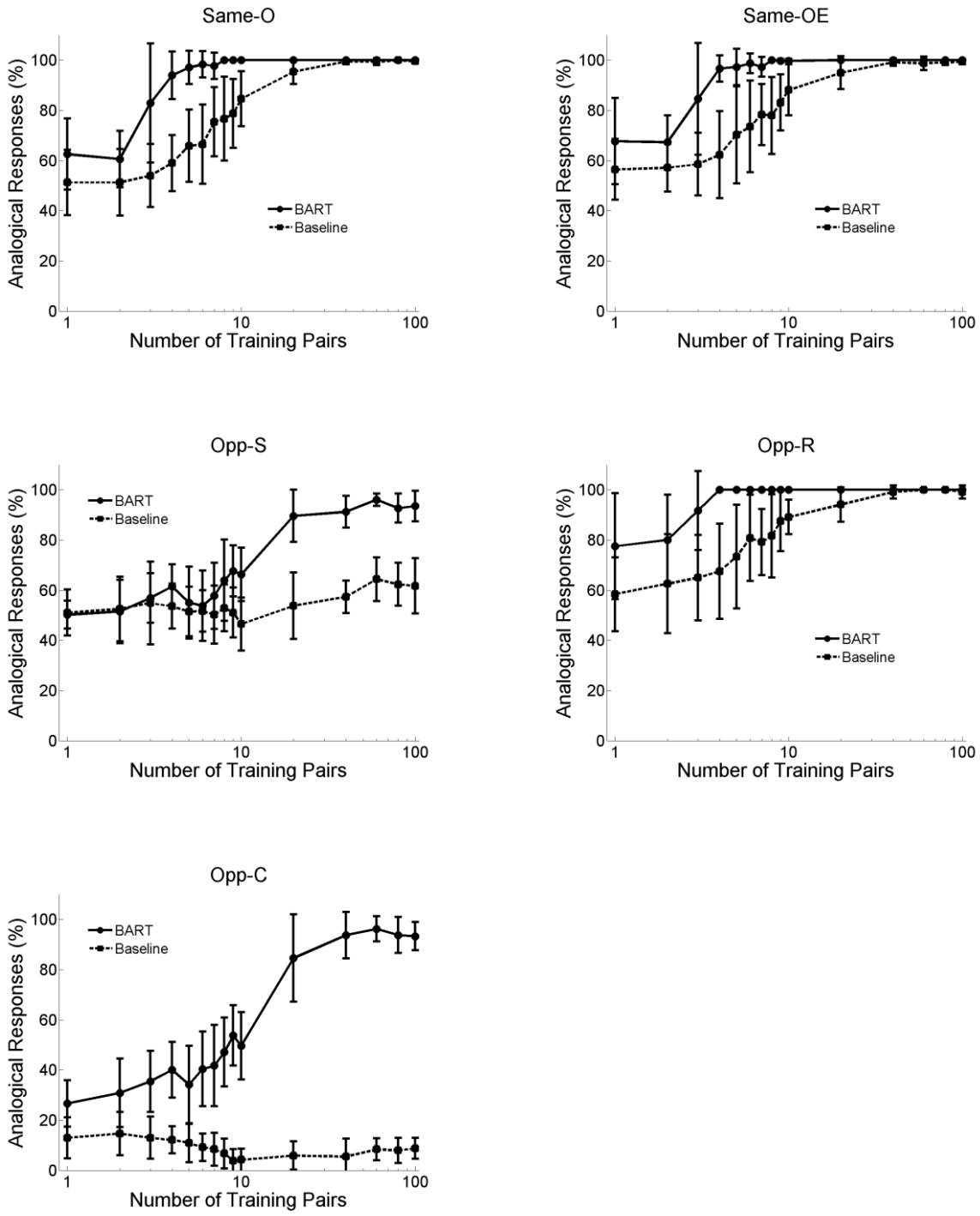


Figure 1.11. Proportion of analogical responses as a function of the number of training examples (log scale) with Leuven inputs for the five types of analogy problems. Solid lines present results from BART with empirical prior and hyperprior; dashed lines present results for baseline model (uninformative prior). Error bars indicate 1 standard deviation (results based on 10 runs).

inputs, which is not surprising given the much greater complexity of the Leuven inputs. Indeed, it was far from obvious that the algorithm for importance-guided mapping would work at all when applied to representations of first-order relations that are highly distributed over many dimensions, rather than being localized on one critical dimension (as was the case for the simulations based on rating vectors; see Figure 1.4). As indicated in Figure 1.10, the first-order relations acquired from the Leuven vectors generally involved at least 40 reliable predictor variables working together.

In fact, the performance of the BART model on the analogy tests was excellent, achieving essentially perfect accuracy after 100 training trials on four problem types, and about 90% on Opp-C problems. For the Same-O, Same-OE, and Opp-R analogy types, the baseline model does not catch up to BART until after 40 training pairs. For the Opp-S and Opp-C analogy types, performance of the baseline model hovers around chance (50% and 0%, respectively) even after 100 training trials. For the latter problem types, we explored the impact of more extended training on analogy performance for the baseline model. Even after 700 training examples, the baseline model still lagged behind BART by about 9 percentage points on Opp-S problems and 46 percentage points on Opp-C problems. In sum, when faced with high-dimensional inputs based on Leuven inputs, BART was able to achieve substantial success in solving structured analogy problems, with its informative priors playing a decisive role.

Tests of BART Using Topics Vectors

Inputs

We also applied BART to the yet more challenging problem of learning comparative relations from input representations taken from the topic model (Griffiths et al., 2007). Whereas the ratings vectors has clear localist codes for the critical continua (size, fierceness, etc.), and the

Leuven vectors included some features that were transparently related to them, the more opaque topics vectors did not provide any dimensions that were transparently relevant for learning comparative relations. To make the results as comparable as possible to those obtained in the previous simulations, we used the same set of 77 animal names taken from the Holyoak and Mah (1981) norms that were used in the simulations based on rating inputs.

The topics representations we used for these words were derived from the output of the topic model based on the tasaALL data base (see Griffiths et al., 2007). This output consists of twenty-four samples with 300 topics each for 26,243 unique words. Each sample consists of a word-by-topic matrix, in which the entry in the *i*th row and *j*th column is the number of times that the *i*th word appeared in the corpus and was assigned to the *j*th topic. We performed several pre-processing operations to create the vectors used as the immediate inputs to BART. First, we added a smoothing parameter of 0.01 to each entry in the matrix. We then derived a feature vector for each word, in which each feature value corresponds to the conditional probability of a corresponding topic given that word. This value is simply the joint frequency of the topic and word (an entry in the matrix) divided by the frequency of that word (the sum of a row in the matrix). Although the frequencies of specific words for each topic vary somewhat across the 24 samples, it was clear from inspection that the same 300 topics appeared in the same order across all 24 samples, indicating that the topics solutions are robust. Accordingly, a single feature vector for each word was calculated simply by averaging its feature vectors across all samples. We multiplied the computed probability by 100 to make the magnitude range roughly comparable to that of the rating inputs.

If all 300 dimensions of each word vector were used, each word-pair vector would have 600 dimensions. However, for any individual word, most feature values are close to 0 (reflecting

the fact that most of the 300 topics are irrelevant for any particular word). Dimensions that yield feature values at or near 0 for all words of interest (the animal names) will be useless in subsequent relation learning, and are likely to introduce noise that will impede any learning algorithm given the sheer size of the search space and limited number of training examples. In order to focus on the most important dimensions (those for which animals names tend to have non-zero probabilities), we summed the feature vectors for all 77 animals and identified the 50 dimensions that yielded the largest sums. By using just these 50 most important topics dimensions to form vectors for each individual word, the total size of the vector for each word pair was reduced to 100 (the same dimensionality as for the Leuven vectors).

Because the feature values as described so far are based on probabilities, they are necessarily non-negative. Accordingly, we subtracted from each feature the mean value of that feature across all 26,243 word vectors. The feature values in the vectors used as inputs to BART therefore included both negative and positive values, with means near 0.

Training

The basic training regimen was identical to that employed with the Leuven vectors (except all learning was based solely on animals from the Holyoak & Mah, 1981, norms). The same hyperprior parameters were used. (Hyperpriors improved generalization performance by about 3 percentage points overall, with more significant improvement for certain analogy tests.) For comparison, we again tested the same baseline model as that used with both rating and Leuven inputs (i.e., Bayesian logistic regression with standard normal distributions as uninformative priors).

Generalization Performance

Basic tests. All reported results are based on the average performance over 10 runs, each of which randomly selected a set of training pairs from the pool. Figure 1.12 depicts BART's generalization curves for the four "greater" relations as a function of the number of training examples. Not surprisingly, given the vastly greater opacity of topics representations, the level of performance was considerably lower overall than was obtained with the rating or Leuven vectors. However, the full BART model, with empirical priors on mean weights and a hyperprior on variances, achieved substantial generalization (about 70-80% accuracy for the four "greater" relations after 100 training examples). These results indicate that even when magnitude information is not coded in any clear way in the inputs, BART can learn useful representations of comparative relations from positive examples. The baseline model showed much weaker

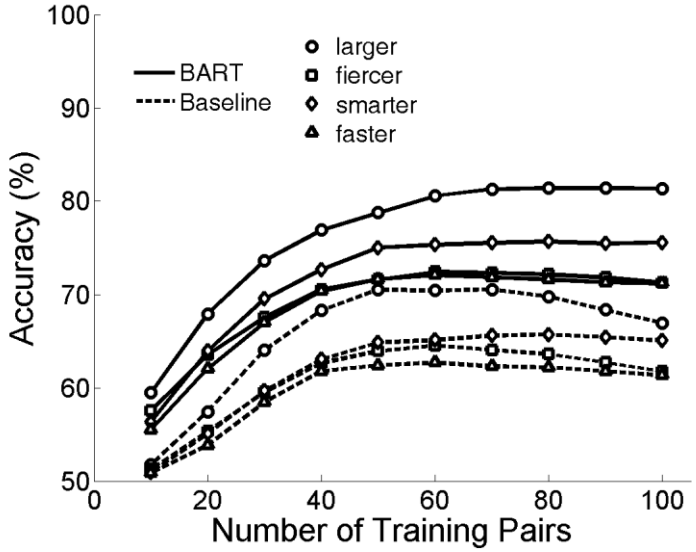


Figure 1.12. Accuracy in the generalization task with topics vectors as a function of the number of training examples for the four comparative relations. Solid lines indicate the performance of BART using the empirical prior and hyperprior; dashed lines indicate the performance of a baseline model (Bayesian logistic regression model with uninformative prior).

generalization performance, starting at chance (50%) and peaking at a mean of 67% accuracy after about 80 training examples.

We also explored how BART's generalization performance changed with more extended training. Whereas BART appeared to be unbiased when trained with up to 80-90 examples, a further increase in the number of training examples led to a bias towards "yes" responses. This type of response bias leads to reduced accuracy if a fixed decision criterion is used. (Note the slight decline in accuracy apparent in Figure 1.12 after 70 training examples, especially for the baseline model). Generalization performance as measured by A_z continued to improve slightly with increased numbers of training examples. But even after 2000 training examples, overall generalization performance as measured by A_z was higher for BART (.87) than for the baseline model (.82).

To examine whether the relational representations that BART derives from topics vectors yield the distance effect obtained for comparative judgments by humans, we again examined how BART's generalization performance relates to the rated subjective distance between each test pair of animals on the dimension of interest (as measured using the Holyoak & Mah, 1981, norms). Figure 1.13 plots the mean log ratio of predicted probabilities for positive versus negative test pairs as a function of distance on the relevant dimension between the two animals in a pair (after learning from 100 training examples), using the same distance bins as were used to test the model with rating vectors. The log ratio again increased monotonically with distance. Thus, the relational representations that BART acquired from topics inputs clearly yield a symbolic distance effect.

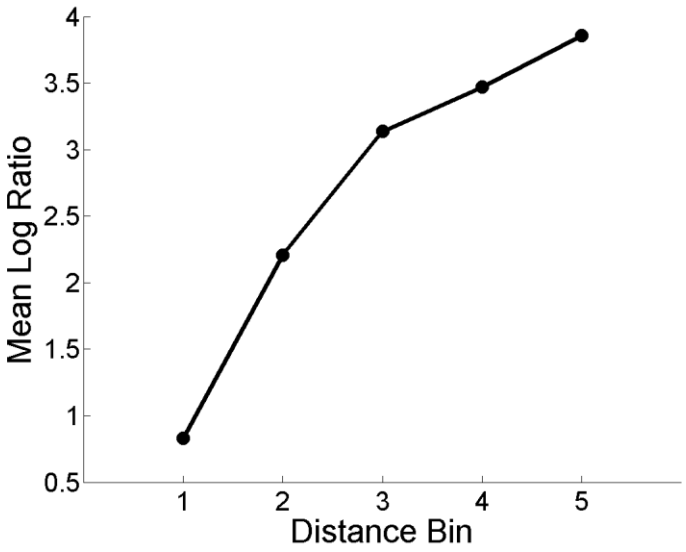


Figure 1.13. Log of the ratio between predicted posterior probability of each positive test pair instantiating a "greater" relation and that of the reversed pair instantiating the relation on generalization test (topics inputs) as a function of rated distance on the relevant continuum. Distance bins are based on Holyoak and Mah's (1981) norms (see Figure 1.6). Results are collapsed over the four continua.

Generalization beyond the range of training examples. As in the case of the simulations based on ratings and Leuven vectors, we performed a series of computational experiments to determine whether BART is capable of generalizing to new types of pairs that in various ways go beyond the range of the training examples.

(1) *Training with 100 examples from distance bins 1-3 only and testing on all five distance bins.* We obtained a monotonic increase in mean log ratio across all levels of distance: .78, 2.05, 2.86, 3.29, and 3.61 for bins 1 to 5, respectively, extending the similar pattern obtained with ratings and Leuven inputs.

(2) *Training with 100 pairs of the form [4, 1] and testing on all pairs of the form [3, 2].* BART's accuracy levels were 75%, 45%, 54%, and 44% for *larger*, *fiercer*, *smarter*, and *faster*, respectively.

(3) *Standard training with 100 pairs formed from all groups and testing on all pairs of the form [3, 2]*. BART's accuracy levels were 87%, 54%, 56%, and 54% for *larger*, *fiercer*, *smarter*, and *faster*, respectively. Thus for topics inputs, tests 2 and 3 indicate that generalization performance was weak for close midrange pairs, especially when the range of the training pairs was restricted to those of the form [4, 1] (test 2).

(4) *Training with 100 pairs of the form [2, 1] and testing on all pairs of the form [4, 3], or the reverse.* Averaged across the two variations, generalization accuracy was 49%, 61%, 63%, and 62% respectively for *larger*, *fiercer*, *smarter*, and *faster*, indicating modest performance for the latter three relations.

(5) *No overlap between training and test animals with 100 training examples*. BART achieved accuracies of 71%, 65%, 68%, and 63% for *larger*, *fiercer*, *smarter*, and *faster*, respectively, indicating moderately successful generalization to animals not encountered during training.

Content of Learned Weight Distributions

We examined representative solutions that BART generated in learning representations for the comparative relations based on topics inputs. These solutions were even more distributed than those obtained using Leuven inputs (Figure 1.10), with around 30 dimensions (i.e, about 60 weights) distinguishing the two roles for each comparative. The two roles were generally contrastive (i.e., weights associated with the two roles took on opposite signs). However, unlike those based on ratings and Leuven vectors, the topics solutions did not clearly distinguish the "greater" and "lesser" poles of individual continua. That is, rather than having predominantly high positive weights on the first role and high negative weights on the second, the pattern of weight polarity was more mixed. As we will see, the lack of a clear distinction between "greater"

comparatives (*larger*, *fiercer*, etc.) and "lesser" ones (*smaller*, *meeker*, etc.) had negative consequences for BART's ability to solve some specific types of analogy problems based on topics inputs.

In addition to being more highly distributed, the topics solutions proved to be much more opaque than the Leuven solutions. Indeed, the term "topic" (which suggests an overall semantic theme) seems like a misnomer when applied to the feature dimensions that loaded highly for the various continua. Rather than having a clear semantic interpretation, each topic can really only be characterized by the list of words associated with it. For example, the "topic" most strongly predictive that the first object was larger than the second was highly associated with words for body parts (e.g., *blood*, *body*, *heart*, *cells*, etc.). Of course, this is only one of about 30 topics that collectively drove the decision as to which animal is larger. Thus BART was able to learn and generalize representations of comparatives from topics inputs with moderate success, even though the underlying features were subsymbolic.

Analogy Performance

Given that the topics representations for comparatives were highly distributed and semantically opaque, it is not surprising that using them to solve higher-order analogy problems proved to be challenging. We tested BART and the baseline model on the five types of analogy problems in the same manner as for the ratings and Leuven inputs. The results for up to 300 training examples are shown in Figure 1.14. The overall level of performance is lower than was obtained when the models were trained with ratings or Leuven inputs. Nonetheless, performance of the BART model on the analogy tests after learning from topics inputs was quite good. After 300 training examples, the performance level of BART was essentially perfect for Same-O problems, at about 90% accuracy for Opp-S problems, and 80% for Opp-C problems. BART

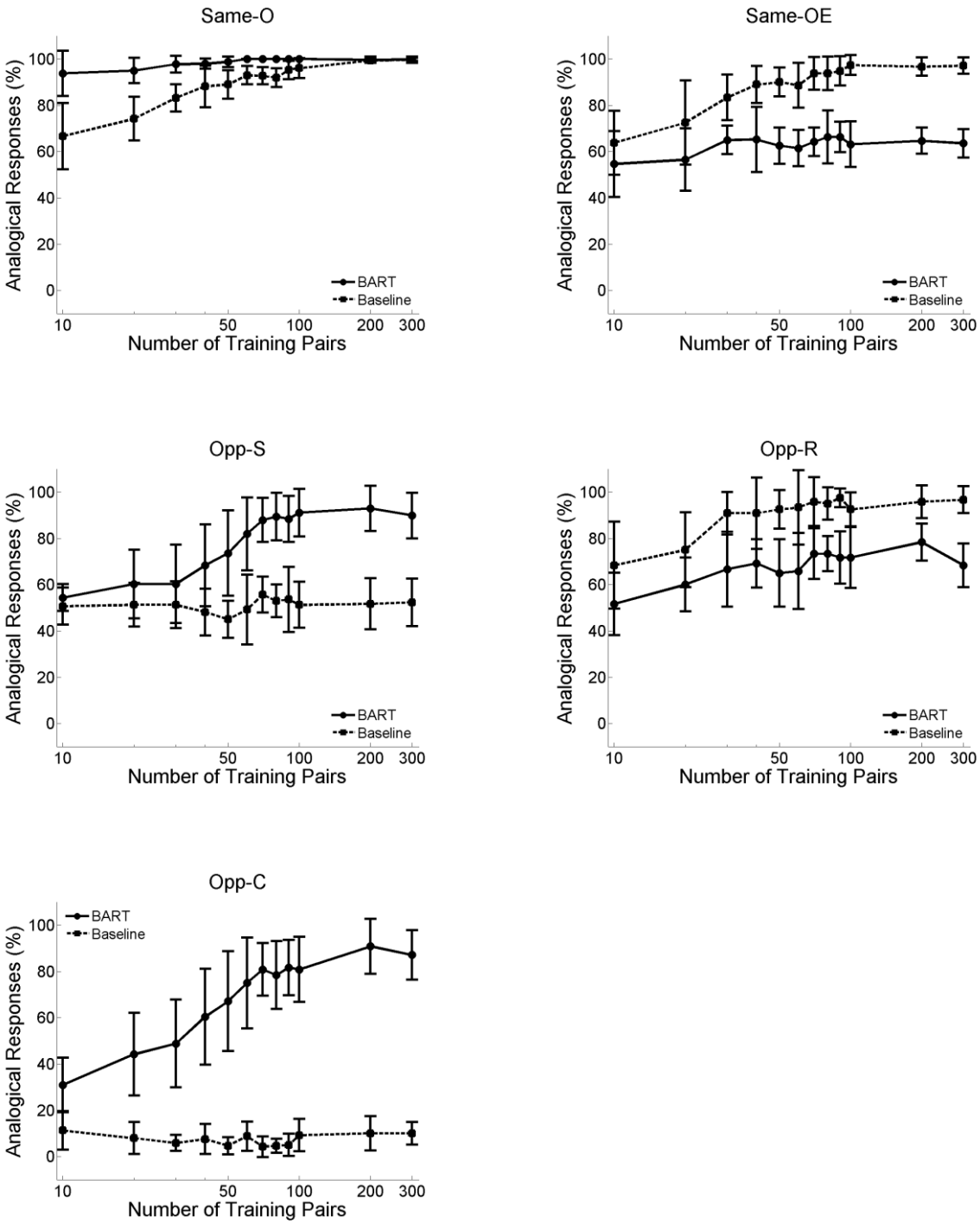


Figure 1.14. Proportion of analogical responses as a function of the number of training examples with topics input for five types of analogy problems. Solid lines present results from BART with empirical prior and hyperprior; dashed lines present results for baseline model (uninformative prior). Error bars indicate 1 standard deviation (results based on 10 runs).

performed less well on the Opp-R problems (about 70%) and the Same-OE problems (about 60%).

BART's lower performance on the latter two types of problems reflects that fact that in each case the correct answer can only be discriminated from the foil on the basis of polarity. Thus in the Same-OE type, the foil is a pair of relations at the opposite pole from the A:B pair, and in the Opp-R type the foil has the polarity reversed relative to A:B (e.g., if *larger:smaller* is the A:B term and *fiercer:meeker* is the correct C:D choice, the foil might be *meeker:fiercer*). As discussed earlier, for any pair of polar opposites BART implicitly identifies the "greater" relation as that for which the first role has predominantly positive weights on the relevant dimensions, whereas the "lesser" relation is that for which the first role has predominantly negative weights. As noted above, the topics inputs were much less clear than the rating or Leuven inputs, providing a mix of dimensions that were positively and negatively weighted as indicants of (for example) *larger* and *smaller*. In other words, topics inputs did not clearly establish "which end is up" for the various continua, making it difficult for BART to use polarity information as its sole basis for selecting an analogical completion.

Somewhat paradoxically, the baseline model actually was more accurate than BART for the problems types where polarity information was critical (Same-OE and Opp-R). The reason is that the baseline model (without empirical priors or hyperpriors) estimated higher variances on weights, which increased sensitivity to the small difference between the same-pole correct choice and the opposite-pole foil. Intuitively, BART "knew too much", viewing (for example) a pair like *meeker-fiercer* as a competitive foil to *fiercer-meeker* when seeking a match to *larger-smaller*, since all these pairs instantiate contrasting relations. By contrast, the baseline model

could detect no apparent relationship between *meeker-fiercer* and *larger-smaller*, so was more likely to favor the correct option as the analogical completion.

More broadly, however, the performance of the baseline model across the entire set of analogy tests was dismal (see Figure 1.14). As was the case for the Leuven inputs, the baseline model failed completely on the Opp-S problem type (near chance level of 50%) and the Opp-C type (near chance level of 0%). Even after 2000 examples, performance of the baseline model lagged 33 percentage points behind BART on Opp-S problems and 90 percentage points on Opp-C problems. Thus even though the baseline model achieved modest success in relational generalization using both Leuven and topics inputs, it was unable to use its learned representations to reliably solve higher-order analogy problems.

It may seem surprising that priors continued to play a critical role in analogy performance even after 2000 trials, as the general rule for Bayesian models is that priors are eventually swamped by data. However, the fact that learning in our simulations was based solely on positive examples may have made priors especially potent. Although the model was expected to learn a specific comparative relation, such as *larger*, a finite set of positive examples is likely to be consistent with multiple possible relations (e.g., *both animals*, *both physical objects*). The contrastive priors provided BART with a strong "push" in the direction of comparative relations, whereas the baseline model, with its uninformative priors, might sometimes have acquired weights consistent with other possible relations exhibited by the positive training examples. Consequently, the patterns of weight distributions acquired by the baseline model likely were more variable from one comparative to another, impairing its performance on higher-order analogy problems.

General Discussion

Summary

The work described in this paper addresses the issue of whether and how relational representations that can support structured reasoning may be learnable from non-relational inputs. Using comparative relations as a model domain, we performed a series of computational experiments based on BART, a Bayesian model of relation learning and analogical inference. In order to relate our findings to the general manner in which children appear to acquire concepts, we focused on learning from labeled positive examples of each target relation. BART incorporates a key representational assumption: a relation is represented by a weight distribution that serves to assess the probability that a pair of objects exemplifies it. BART proceeds in two basic stages. First, guided by empirical priors on mean weights and hyperprior on variances, the model uses Bayesian inference to update its weight distribution based on training examples. Second, the model uses importance-guided mapping to transform its learned weight distributions and then calculate the distance between pairs of relations, thereby assessing the validity of higher-order analogies based on the implicit relations "both same" (*higher:fiercer* :: *smarter:faster*) and "opposite" (*higher:lower* :: *fiercer:meeker*).

When trained and tested with items based on small object vectors derived from human ratings of subjective magnitudes, BART achieved near-perfect accuracy both in generalizing to new examples of relations and in assessing higher-order analogies based on its acquired relations. The high-dimensional and more complex Leuven and topics vectors posed a far greater computational challenge, as no invariant features are apparent, and learning depends on acquiring distributed representations over dozens of feature dimensions. When Leuven vectors were used as inputs, generalization accuracy was in the range of 90% accuracy after 100 training

examples; with topics vectors, accuracy was in the range of 70-80%. Thus BART showed substantial generalization ability after learning from the more complex inputs. Moreover, for all three types of inputs, BART was able to generalize to a completely new set of animals than those used in training.

When tested on higher-order analogy problems based on the relations "same extreme" and "opposite", the algorithm for importance-guided mapping yielded near-perfect performance using BART's learned relational representations, for both ratings and Leuven inputs. BART's analogy performance was also quite strong for topics inputs (except that topics inputs did not provide information that could clearly distinguish the "greater" versus "lesser" pole of each magnitude continua). In contrast, a baseline model with uninformative priors showed substantially weaker generalization and analogy performance for both Leuven and topics vectors (even though it was provided with the identical algorithm for importance-guided mapping).

For all three types of inputs, the relational representations learned by BART provided a qualitative account of the symbolic distance effect (Moyer, 1973). The degree of difficulty of making relative judgments on a dimension (as indexed by the model's estimates of log posterior probability ratio for a target relation) varied inversely with the magnitude difference between the two items in a pair. In addition, BART demonstrated the capacity to generalize outside the range of magnitude distances provided in the training set. Even when trained on animal pairs exhibiting small or medium size differences, the model was most confident when generalizing to novel pairs exhibiting large size differences. The model is thus consistent with evidence that people can distinguish "ideal" from "most typical" exemplars (Kittur, Hummel & Holyoak, 2006). BART in effect defines the "ideal" exemplar of a *larger* relation as the pair with the largest size

difference (e.g., "a dinosaur is larger than a flea"), even though a pair like "a fox is larger than a dove" would be more typical of the observed instances (i.e., closer to their central tendency).

Overall, the simulations reported here thus show that BART was able to pass five critical tests that can be posed for any model of relation learning. We have shown that the model (1) can learn first-order relations from complex non-relational inputs that were independently generated, (2) with high efficiency, (3) generalize to classify novel relational examples, (4) capture a major factor (symbolic distance) that affects difficulty of human comparative judgments, and (5) use its learned relational representations to solve higher-order analogy problems. No previous model of relation learning has met all of these criteria.

Comparison with Previous Approaches

There have been many previous computational models of various aspects of analogical reasoning, which have been classified as symbolic, connectionist, and various hybrids (French, 2002; Gentner & Forbus, 2011). BART's capabilities appear painfully limited when compared to those of the "state-of the art" analogy models. To date, its most advanced accomplishment is to solve simple four-term analogy problems, whereas other models can perform much more complex feats involving analog retrieval, mapping, inference, and schema formation. Even in the restricted domain of four-term analogy problems, BART cannot compete with state-of-the-art machine-learning models (e.g., Turney, 2006).

However, BART is focusing on very different issues than those addressed by most previous analogy models. It attempts to answer the basic question: How might relational representations be created? The operation of BART provides a well-specified computational model of the type of relational re-representation that seems to underlie the power of human thinking (Penn et al., 2008). More generally, the model provides a concrete example of how new

representations can be acquired by forms of inductive bootstrapping: first the use of empirical priors to "jump start" relation learning, and then the use of importance-guided mapping to transform and compare relational representations. BART illustrates the centrality of analogical bootstrapping in learning relations (cf. Carey, 2011; Gentner, 2010).

The idea that functionally-defined importance provides a key pragmatic constraint on mapping has a long history (Holyoak, 1985), and BART exemplifies a very basic mechanism by which importance can be defined quantitatively and used to place non-identical dimensions into correspondence. Unlike previous computational models of mapping, BART finds mappings between features at a subsymbolic level (identifying systematic correspondences between distributed patterns in a high-dimensional weight space), rather than between explicit predicates. Subsymbolic mapping processes of this sort may underlie various types of implicit analogical transfer (e.g., Day & Goldstone, 2011).

The present findings provide a proof-of-concept that, for the domain of comparative relations, a capacity for structured relational reasoning can potentially emerge from bottom-up learning based on unstructured inputs. Particularly, in the case of the topics vectors, the inputs were created without any human guidance that might have tailored them to the relational learning task. If we consider the operation of the topic model itself (Griffiths et al., 2007) in conjunction with BART, the representations that the latter model used to assess structured analogy problems can be traced back to the raw statistics of covariation among words in texts. The to-be-learned relations did not correspond to any specific dimensions created by the topic model. In broad strokes, we have shown that BART can (1) solve structured analogy problems, albeit simple ones, (2) using relational representations the model learned itself (3) from unstructured inputs (4) that were independently generated. No previous model of analogy has demonstrated a

comparable capability, which arguably is the essential precursor to a satisfying account of complex human analogical reasoning.

The most appropriate comparisons for BART are with previous models of relation learning. As a discriminative Bayesian model, BART is most directly related to the model developed by Silva, Airoldi and Heller (2007; Silva, Heller & Ghahramani, 2007; also Chen et al., 2010). Like the model of Silva and colleagues, BART uses empirical priors to bootstrap relation learning; however, in BART the empirical priors are relation-specific, and can be used together with a hyperprior on variances. BART is able to use training examples to automatically select the one-place predicate best-suited for generation of priors on mean weights for a new comparative relation. The algorithm for importance-guided mapping is a key innovation that enables BART to move beyond relational generalization to the more challenging task of solving higher-order analogy problems based on its learned representations.

It is instructive to relate the core concepts and mechanisms instantiated in BART to those that underlie other approaches to relation learning. We will focus on three central aspects of the BART model. These are: (1) exploiting empirical priors, (2) representing relations as weight distributions, and (3) allowing role-dependent operations on representations. With these in mind, we can draw comparisons and contrasts with three general approaches reviewed earlier.

Comparison to hierarchical generative models. If we take the structure-learning model of Kemp and Tenenbaum (2008) as an example, the generative approach is broadly similar to BART in the use of statistical learning over distributions. The two approaches also converge in denying that explicit representations of relations could be acquired by a complete tabula rasa. However, the models make different assumptions about what knowledge or capacities the learner brings to the task. Kemp and Tenenbaum's model is endowed with a grammar that can generate

candidate structures over which statistical learning can be applied. BART does not come equipped with a comparable grammar of relations. Rather, it comes with a suite of operations that it can use to create and transform representations (in particular, the ability to select and use empirical priors to initialize the representation of a to-be-learned relation, and the ability to perform importance-guided mapping and subsequent relational transformations).

The generative and discriminative approaches to learning relations may well prove to be complementary. An important question for future research is whether a relational representation of the type acquired by BART might be transformed into a generative model, a step likely to be necessary in order to achieve the full range of human-like relational capabilities.

Comparison to neural network models. Both generative and discriminative Bayesian models are similar to neural network models in their emphasis on statistical learning as a major contributor to the acquisition of knowledge. Discriminative models such as BART are perhaps somewhat closer to the spirit of neural network models, emphasizing the emergence of knowledge from bottom-up processing of data provided by the environment. However, the weight distributions over feature vectors that BART uses to code relations capture more information than do the representations created by typical neural network models. Weight distributions code not only first-order statistics (means), but also second-order statistics (variances and covariances) that capture uncertainty about weights and inter-weight correlations (a property shared by "deep learning nets"; Salakhutdinov & Hinton, 2009).

Yet paradoxically, BART's relational representations are also explicit and structured in ways that representations in a distributed neural network are not. Most basically, BART's weight distributions respect the integrity of distinct roles (e.g., the roles of the larger versus smaller member of a pair of objects). The internal structure of relations in BART is presumably inherited

from the output of the perceptual system, which codes objects as individuals. A relation such as *larger* is learned from pairs of objects, and hence is structured as a pair of roles. In contrast to the model of Rogers and McClelland (2008), for example, an individual relation in BART has a distinct identity (e.g., the weight distribution for *larger* is different from that for *fiercer*, and also from the complementary relation *smaller*). Because relational representations in BART are isolable from one another, they can be compared and systematically transformed. The properties of the weight distributions employed by BART are thus qualitatively and quantitatively different from those of the weight matrices used in classical neural networks.

Comparison to symbolic connectionist models. Among the algorithmic models of relation learning, BART has most in common with the class of symbolic connectionist models, such as LISA (Hummel & Holyoak, 1997, 2003) and DORA (Doumas et al., 2008; also see Halford et al., 1998). These models, like BART, assume that relational representations are structured in terms of roles, thereby escaping the fundamental limitations of conventional neural-network models. Another important similarity is that BART, like DORA, exploits the potential for bootstrapping from initial learning of one-place predicates to learning comparative relations.

In general terms, both DORA and BART aim to learn relations using bottom-up mechanisms based on detection of covariation among the objects that fill relational roles. DORA emphasizes learning from unlabeled examples, whereas BART focuses on learning from labeled examples (either positive or negative, although the former are assumed to be more common in child language acquisition). DORA, extending earlier proposals concerning schema induction (Gick & Holyoak, 1983), learns relations by taking the intersection of the feature representations of multiple examples. In comparison to a regression algorithm of the sort used by BART, the logical intersection operator appears to be too strict (a single exception may cause a feature to be

dropped from the representation of a relation). The method is ill-suited for learning concepts defined by distributed representations over features that are only probabilistically predictive, as was the case for Leuven and topics vectors. By using Bayesian inference, BART is able to learn probabilistic representations of relations from positive examples, without requiring any strictly invariant features, while simultaneously factoring in the influence of prior knowledge. As noted earlier, DORA has only been tested on hand-coded inputs that include invariant features over which to-be-learned relations can be defined. DORA (like LISA) is not designed to map non-identical features to one another, as the mapping algorithm is restricted to mapping predicates, rather than object features.

Most basically, BART's representational scheme for relations distinguishes it from symbolic as well as non-symbolic connectionist models. In both varieties of neural-network models, relations and objects have generally been represented in a common format (as distributed sets of units interconnected by weighted associations). In symbolic-connectionist models, which focus on explicit relational representations, both relations and objects have been represented in terms of units for semantic features—either separate pools of feature units for the two types of entities (LISA; Hummel & Holyoak, 1997, 2003; also Halford et al., 1997), or a single pool (Doumas et al., 2008). BART introduces a very different representational assumption. Whereas objects are represented by a vector of features, first-order relations are represented as weight distributions. In BART, relations (weight distributions) and objects (feature vectors) constitute distinct but connected representational elements. This representational distinction is critical for BART's ability to acquire relational representations by statistical learning.

BART's representational assumptions may suggest an important way in which algorithmic symbolic-connectionist models can be refined. The use of temporal synchrony for role binding gives rise to inherent capacity limits, related to the number of distinct temporal phases that can be interleaved without significant overlap of firing for each phase. Given established limits on neural firing rates, this "relational bottleneck" has been estimated at 4-6 distinct phases (Cowan, 2001; Hummel & Holyoak, 1997). If role bindings are coded by synchronizing the neural code for a role and its filler, as the LISA model assumes, then this limit translates directly into 4-6 concurrently active role bindings, or 2-3 complete propositions, a number that appears plausible for adult humans. But as noted earlier, models that use temporal firing patterns as a dynamic code for role bindings in active memory can only synchronize representations that can be kept distinct *despite* firing together. LISA's use of neural synchrony therefore depends on defining separate pools of features for objects and relations.

In order to model the learning of features of relations from features of objects, the DORA model (Doumas et al., 2008) assumes instead that relations and objects are defined over a *single* pool of semantic features, and that role bindings are coded by *a*synchrony of firing for a role and its filler. This shift in representational assumptions means that DORA requires twice as many distinct temporal phases as does LISA to represent the same number of role bindings. In effect, DORA's estimate of the capacity of human working memory is half the value predicted by LISA. Doumas et al. (2008, pp. 30-31) suggest that asynchrony may be required only for relation learning and not for relational inference. However, it is unclear how inferences could be made reliably if roles and their fillers were coded on the same pool of features, and yet allowed to fire in synchrony (e.g., the distinction between "elephants are big" and "elephants are gray" would seem to be lost).

BART's assumption that relations (more generally, predicates) and objects rely on distinct types of representations (weight distributions versus feature vectors) goes between the horns of this dilemma, providing a potential basis for a LISA-like system of binding by synchrony that nonetheless is capable of relation learning. That is, the dynamic form of coding a role binding might involve the synchronous activation of a role (a distinct subset of the weight distribution for a relation) and the feature vector for its filler. The role (weight distribution) and its filler (feature vector) would not be confusable even when synchronized, because each would constitute a distinct representational type. An algorithmic implementation based on BART's form of relation representation would thus yield the same estimate for the capacity of working memory as does LISA.

Potential Extensions

Acquiring more detailed developmental data. For the present project, we created a microworld in which a learner (the BART model and variations on it) must learn several comparative relations defined over a set of animal concepts, using inputs consisting of feature vectors, and then must draw higher-order analogies based on the acquired relational representations. In general terms, we constrain the task in ways that seem consistent with comparable relation learning by children (modest numbers of largely positive examples, acquiring one-place predicates prior to true relations). But we acknowledge that our microworld is not the one that children actually encounter. Children do not learn *larger* and other comparative relations from animals only, and we lack detailed knowledge of the inputs children actually have available. At most, realistic inputs resemble those that BART receives in that they are also based (at least in part) on sets (likely quite large) of features associated with individual objects.

Because of the idealized nature of our microworld, empirical assessment of the models was largely qualitative (which is rather counterintuitive, since BART generates detailed learning curves). We hope that future tests of computational models of relation learning can be informed by more detailed empirical evidence regarding the inputs children use to acquire relations, the trajectory of children's learning for specific relational concepts, and the linkage between relational generalization and the ability to reason by analogy.

Extensions to richer inputs. As we emphasized at the outset, the representations that serve as inputs to children learning relations are undoubtedly richer than those we provided to BART in the present set of simulations. Children learn from more direct perceptual experience, including motoric feedback from their own actions. For example, Maouene, Hidaka and Smith (2008) showed that the age of acquisition for basic English verbs (e.g., *kiss*, *hug*, *kick*) is related to the nature of their association (for adults) with body parts (e.g., the mouth versus the hands and arms). As another example, work on action recognition has identified certain "signature movements", such as a punch, that have a special status in rapid identification of types of threatening actions (van Boxtel & Lu, 2011, 2012). Such cues (in conjunction with adult speech) very likely provide a significant part of the inputs available to children as they learn verbs corresponding to basic actions. Realistic inputs are likely to involve greater structure than the "flat" vectors used in the present paper, including various types of higher-order features (Regier, 1996; Regier & Carlson, 2001). Future research should explore the use of learning algorithms that can create and exploit hierarchical structure in their inputs.

Role of empirical priors in relation learning. The simulations reported here demonstrate that representations of one-place predicates can provide very useful empirical priors to facilitate learning of the corresponding two-place relations. Knowledge about a one-place

predicate such as *large* can be learned from a set of single objects (e.g., *elephant*), whereas learning the relation *larger* requires joint processing of pairs of objects (e.g., *elephant* and *bear*). Based on Halford's (1993; Halford et al., 2010) assumption that capacity increases over the course of cognitive and neural development, and that attending to two objects requires greater capacity than attending to one, it follows that children will tend to learn one-place predicates prior to multi-place relations (which have at least two roles), in accord with developmental evidence (Smith, 1989).

However, this developmental pattern does not necessarily imply that learning specific one-place predicates (e.g., *large*, *small*) is a strict prerequisite for learning a related two-place predicate (e.g., *larger*). At least for ratings and Leuven vectors, the baseline model with uninformative priors was able to learn comparative relations and achieve substantial generalization performance when given an adequate number of training examples. As long as we assume sufficient working memory to hold two items, BART can proceed to learn a two-place predicate directly, regardless of whether or not it has already acquired corresponding one-place predicates. It is an open empirical question whether or not children necessarily learn one-place relative adjectives as a prerequisite to learning two-place comparative adjectives (cf. Halford et al., 2010). More generally, however, many multi-place predicates (e.g., *opposite*) do not seem to naturally decompose into simpler one-place predicates.

The further exploration of empirical priors will be especially important in attempting to extend the current approach to other types of relations besides comparatives (see Jurgens, Mohammad, Turney & Holyoak, 2012). As the pool of potential empirical priors grows larger and more varied, more sophisticated algorithms for prior selection may prove useful. For

example, prior selection may involve a hierarchical process, winnowing options based on general types of relations (e.g., varieties of sameness versus contrast).

It should be emphasized that the concept of empirical priors is considerably more general than the idea of using one-place predicates to guide learning of related two-place relations. Relation learning can also potentially be bootstrapped by previously-learned relations (e.g., a perceptually-based comparative such as *larger* might facilitate subsequent acquisition of a more abstract comparative such as *smarter*). Yet more generally, the entire process of analogical reasoning can be viewed as a sophisticated use of empirical priors, in which the source analog is used to impose priors to guide learning about the target (Holyoak, Lee & Lu, 2010).

Learning higher-order relations. Although the present version of BART does not create explicit representations of higher-order relations such as *opposite*, it does appear to set the stage for this possibility. In evaluating higher-order analogies, the model is implicitly sensitive to whether A:B and C:D both instantiate some version of *opposite*. By assessing the distance between transformed weight distributions, BART shows how representations of different relations can be compared with one another. To create explicit higher-order representations, an extension of the model could treat these transformed weight distributions in a manner analogous to feature vectors, recursively applying its statistical learning procedures to acquire higher-order weights that capture the commonalities between pairs of first-order relations such as *larger:smaller* and *fiercer:meeker* (i.e., a representation of *opposite*). In moving from first-order to higher-order relations in this manner, an extension of BART would in effect re-represent first-order relations as derived feature vectors, which can then serve as inputs to a learning process that yields representations of higher-order relations. This basic move—treating learned weights as derived features—provides a potential avenue to allow the development of hierarchical

relational systems. It also provides a possible answer to the inductive puzzle we raised at the outset: how does the mind acquire concepts that cannot be defined in terms of features bound to perception?

Toward an algorithmic model. Although we have developed BART as a computational-level model of how relations might be learned and transformed to solve higher-order analogy problems, it should be possible to incorporate the basic ideas into algorithmic models. As suggested above, the fact that BART creates separate (but linked) representations for relations and their fillers is compatible with synchrony-based models of the symbolic-connectionist variety (Hummel & Holyoak, 1997, 2003). More generally, it is useful to distinguish those aspects of BART that depend on role-governed operations from those that do not. Importantly, the inductive process that updates weight distributions based on training examples is *not* directly dependent on roles. The feature vectors associated with the two objects being compared are simply concatenated. BART's weight distributions can be viewed as a type of "attention weights" that reflect the importance of each dimension for accurate classification of relations. Learning models based on the idea of attention weights have been applied to object categorization and perceptual learning (Nosofsky, 1985; Petrov, Dosher & Z.-L. Lu, 2005).

BART would require a more complex learning algorithm to acquire distributions of weights (rather than simply mean weights) based on supervised learning (cf. Salakhutdinov, Hinton, 2009). A psychologically-realistic learning model would have to accommodate sequential training inputs. Although the version of BART we have described operates on all training data at once, we have in fact also implemented a variant that uses sequential updating. (In general, regression models can operate in either "batch" or sequential fashion.) The sequential version produces very similar results after roughly 100 training examples. Thus,

although a full model of sequential learning would require additional theoretical work, there is reason to be optimistic that such a model is possible (see Lu, Rojas, Beckers & Yuille, 2008, for a sequential model of causal learning). For example, it is conceivable that the brain in effect implements some kind of variational method based on tacit assumptions about the form of neural distributions.

Although the core learning model (updating of weight distributions) is not role-governed, BART does operate on roles to (1) establish empirical priors that guide acquisition of relations, and (2) perform importance-guided mapping based on the learned representations of relations. These operations depend on the manipulation of structured-knowledge, a capacity that is arguably specific to humans (Penn et al., 2008). Interestingly, neither of these operations appears to depend on the full covariance matrix for weight distributions. Rather, the mean weights (MAP estimates) may suffice (see footnote 4). At a neural level, it is more plausible that summary statistics such as MAP estimates could be transmitted to downstream brain regions, rather than the full covariation matrix. It seems plausible that early neural areas are sensitive to intercorrelations among neural firing patterns, which encode covariance information (Aertsen, Gerstein, Habib, & Palm, 1989; Cohen & Kohn, 2011; Cohen & Maunsell, 2009; Kohn & Smith, 2005; Nirenberg & Latham, 2003), whereas higher-level areas instead respond to broader temporal patterning, such as synchrony (Uhlhaas & Singer, 2010; Siegal, Donner & Engel, 2011).

The operations of BART can thus be viewed as demarcating major points along an evolutionary continuum in relational processing and representation. The basic capacity to code approximate magnitudes so as to enable comparative judgments is common across many species. Some primates, including the rhesus monkey, have a limited ability to attach arbitrary symbols to

small magnitudes (Diester & Nieder, 2007), and can also learn alternative first-order relations defined over a common continuum (e.g., selecting the larger or else the smaller of two numerosities in response to a discriminative cue; Cantlon & Brannon, 2005). Roughly, these species-general capabilities correspond to the modest ability of our baseline model, starting with uninformative priors, to learn weight distributions that support comparative judgments, allowing generalization to novel pairs.

However, the capacity to learn weight distributions only sets the stage for acquiring explicit relational representations. BART has the additional capacity to treat weight distributions as structured representations with multiple roles. These explicit representations of first-order relations can then be made available to symbolic processes capable of comparison and rudimentary analogical mapping, thereby enabling a variety of boot-strapping operations. BART uses roles to guide selection of empirical priors, thereby greatly increasing the efficiency of relation learning. After first-order relations have been acquired, BART is able to make structured analogical inferences by mapping dimensions based on their functional influence on relation discrimination, as opposed to their literal identity. An extension of the model that treats weights as derived features could potentially go on to discover the higher-order commonalities shared by first-order relations defined over different dimensions, thereby acquiring explicit representations of higher-order relations such as *opposite*. These symbolic capabilities, perhaps specific to humans, may depend on multiple subregions of the prefrontal cortex (particularly the rostrolateral portion; for a review see Knowlton & Holyoak, 2009). The capacity for role-governed operations may thus represent a late evolutionary development that has allowed humans to attain their unique capacity for abstract thought.

Footnotes

1. Although BART focuses on the computational level of analysis, its implementation includes assumptions at the level of representation and algorithm.
2. In the present paper we refer to this type of test as "relational generalization", whereas it has been called "analogical reasoning" in the machine-learning literature (Silva, Airoldi, & Heller, 2007). The task is indeed closely related to first-order analogical reasoning, in which the relation between A and B concepts (generally objects) is assessed to determine if it is sufficiently similar to the relation between C and D concepts (e.g., Turney, 2006). In contrast, the "analogy" problems described in the present paper require second-order analogical reasoning, which is based on the similarity of relations between relations.
3. We considered the alternative of forming empirical priors from a combination of two one-place predicates (e.g., *large* and *small* might be used to set priors for *larger*). However, developmental evidence indicates that young children often treat such polar opposites as disjoint, whereas children clearly link the primary one-place predicate to its corresponding comparative (e.g., *large* to *larger*; see Smith et al., 1988).
4. A simpler variant of importance-guided mapping is based on just MAP estimates (i.e., mean weights) rather than on entire covariance matrices. Indeed, in exploring all three data sets reported in the present paper, we have found that the simpler variant yields virtually the same performance as the version based on the full covariance matrix. The covariance matrix plays important roles in guiding the acquisition of the MAP estimates during learning, and aids in relational generalization, but apparently is not essential in subsequent analogical processing. In the present paper we describe the complete version of importance-guided mapping for the sake of computational generality, but the simpler

MAP variant may well be more psychologically realistic (see General Discussion). The MATLAB code provides the simpler variant as an option.

References

Aertsen, A.M., Gerstein, G.L., Habib, M.K., & Palm, G. (1989). Dynamics of neuronal firing correlation: Modulation of "effective connectivity". *Journal of Neurophysiology*, *61*, 900–917.

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

Banks, W. P., White, H., Sturgill, W., & Mermelstein, R. (1983). Semantic congruity and expectancy in symbolic judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 560-582.

Bao, J., & Munro, P. (2006). Structural mapping with identical elements neural network. *Proceedings of the 2006 International Joint Conference on Neural Networks*, 870-874.

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

Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choice*. San Francisco: Holden-Day.

Cantlon, J., & Brannon, E. M. (2005). Semantic congruity affects numerical judgments similarly in monkeys and humans. *Proceedings of the National Academy of Sciences, USA*, *102*, 16507–16511.

Cantlon, J., Brannon, E. M., Carter, E., & Pelphrey, K. (2006). Functional imaging of numerical processing in adults and 4-yr-old children. *PLoS Biology, 4*, e125, 1-11.

Carey, S. (2011). Précis of *The origin of concepts*. *Behavioral and Brain Sciences*, *34*, 113-167.

Chen, D., Holyoak, K. J., & Lu, H. (2010). Learning and generalization of abstract semantic relations: Preliminary investigations of Bayesian approaches. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32th Annual Conference of the Cognitive Science Society* (pp. 871-876). Austin, TX: Cognitive Science Society.

Chen, Z., Sanchez, R. P., & Campbell, T. (1997). From beyond to within their grasp: The rudiments of analogical problem solving in 10- and 13-month-olds. *Developmental Psychology, 33*, 790-801.

Clark, H. H. (1970). The primitive nature of children's relational concepts. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 269-278). New York: Wiley.

Cohen, M. R., & Kohn, A. (2011). Measuring and interpreting neuronal correlations. *Nature Reviews Neuroscience*, *14*, 811-819.

Cohen, M.R., & Maunsell, J.H. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience*, *12*, 1594–1600.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–114.

Cover, T., & Thomas, J. (2006). *Elements of information theory* (2nd edition). New York: Wiley.

Day, S., & Goldstone, R. L. (2011). Analogical transfer from a simulated physical system. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37,* 551-567.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel,W., Dry, M., Voorspoels, W., & Storms G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods, 40 (4)*, 1030-1048.

Dehaene, S., & Changeux, J.-P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience, 5*, 390-407.

Diester, I., & Nieder, A. (2007). Semantic associations between signs and numerical categories in the prefrontal cortex. *PLoS Biology*, *5*, e294.

Dorfman, D. D., & Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—Rating-method data. *Journal of Mathematical Psychology*, *6*, 487-496.

Doumas, L. A. A., Hummel, J. E. (2012). Computational models of higher cognition. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning*. New York: Oxford University Press.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*, 1-43.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71-99.

Eckstein, M.P., & Ahumada, A.J., Jr. (2002). Classification images: A tool to analyze visual strategies. *Journal of Vision*, *2*, 1x.

Flas, M., Lammertyn, J., Caessens, B., & Orban, G. A. (2007). Processing of abstract ordinal knowledge in the horizontal segment of the intraparietal sulcus. *Journal of Neuroscience*, *27*, 8952-8956.

French, R. M. (2002). The computational modeling of analogy-making. *Trends in Cognitive Science*, *6*, 200-205.

French, R. M. (2008). Relational priming is to analogy-making as one-ball juggling is to seven-ball juggling. *Behavioral and Brain Sciences*, *31*, 386-387.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10,* 234-257.

Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, H., Penny, W., & Ashburner, J.

(2008). Bayesian decoding of brain images. *NeuroImage*, *39*, 181-205.

Gasser, M., & Colunga, E. (2000). Babies, variables, and relational correlations. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 160-165). Mahwah, NJ: Erlbaum.

Gallistel, C. R. (1993). A conceptual framework for the study of numerical estimation and arithmetic reasoning in animals. In S. T. Boysen & E. J. Capaldi (Eds.), *Development of numerical competence: Animal and human models* (pp. 149-169). Hillsdale, NJ: Erlbaum.

Gardner, H. (1974). Metaphors and modalities: How children project polar adjectives onto diverse domains. *Child Development*, *45*, 84-91.

Geisler, W. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, *59*, 167-192.

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

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34*, 752-775.

Gentner, D., Anggoro, F. K., & Klibanoff, R. S. (2011). Structure mapping and relational language support children's learning of relational categories. *Child Development*, *82*, 1173-1188.

Gentner, D., & Forbus, K. D. (2011). Computational models of analogy. *WIREs Cognitive Science*, *2*, 266-276.

Gentner, D., & Rattermann, M. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on thought and language: Interrelations in development* (pp. 225-277). London: Cambridge University Press.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, *10*, 277-300.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1-38.

Glass, A. L., Holyoak, K. J., & Kossan, N. E. (1977). Children's ability to detect semantic contradictions. *Child Development*, *48*, 279-283.

Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 3–28.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*, 110-119.

Goswami, U. (1992). *Analogical reasoning in children*. Hillsdale, NJ: Erlbaum.

Goswami, U. (2001). Analogical reasoning in children. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 437-470). Cambridge, MA: MIT Press.

Graziano-King, J., & Cairns, H. S. (2005). Acquisition of English comparative adjectives. *Journalof Child Language*, *32*, 345-373.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*, 357-364.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211-244.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767-773.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116,* 661-716.

Halford, G. S. (1993). *Children's understanding: The development of mental models.* Hillsdale, NJ: Erlbaum.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*, 803-864.

Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences*, *14*, 497-505.

Holyoak, K. J. (1977). The form of analog size information in memory. *Cognitive Psychology*, *9*, 31-51.

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation*, Vol. 19 (pp. 59-87). New York: Academic Press.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 234-259). New York: Oxford University Press.

Holyoak, K. J., & Hummel, J. E. (2008). No way to start a space program: Associationism as a launch pad for analogical reasoning. *Behavioral and Brain Sciences*, *31,* 388-389.

Holyoak, K. J., Junn, E. N., & Billman, D. O. (1984). Development of analogical problem-solving skill. *Child Development, 55*, 2042-2055.

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: A theoretical integration with Bayesian causal models. *Journal of Experimental Psychology: General*, *139*, 702-727.

Holyoak, K. J., & Mah, W. A. (1981). Semantic congruity in symbolic comparisons: Evidence against an expectancy hypothesis. *Memory & Cognition*, *9*, 197-204.

Holyoak, K. J., & Walker, J. H. (1976). Subjective magnitude information in semantic orderings. *Journal of Verbal Learning and Verbal Behavior*, *15*, 287-299.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427-466.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220-264.

Jaakkola, T. S., & Jordan, M. I. (2000). Bayesian logistic regression: A variational approach. *Statistics and Computing, 10*, 25-37.

Jani, N. G., & Levine, D. S. (2002). A neural network theory of proportional analogy-making. *Neural Networks*, *13*, 149-183.

Jones, S., & Murphy, M. L. (2005). Using corpora to investigate antonym acquisition. *International Journal of Corpus Linguistics*, *10*, 401-422.

Jurgens, D. A., Mohammad, S. M., Turney, P. D., & Holyoak, K. J. (2012). SemEval-2012 Task 2: Measuring degrees of relational similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)* (pp. 356–364). Montreal, Quebec, Canada: Association for Computational Linguistics.

Kemp, C., Chang, K. K., & Lombardi, L. (2010). Category and feature identification. *Acta Psychologica*, *133*, 216-233.

Kemp, C., & Jern, A. (2009). Abstraction and relational learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams & A. Culotta (Eds.), *Advances in Neural Information Processing Systems*, *22*, 943-951.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*, 307–321.

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

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. *Proceedings of the 21st National Conference on Artificial Intelligence* (Vol. 1, pp. 388). Palo Alto, CA: AAAI Press.

Kittur, A., Hummel, J. E., & Holyoak, K. J. (2006). Ideals aren't always typical: Dissociating goodness-of-exemplar from typicality judgments. In R. Son & N. Miyake (Eds.), *Proceedings of the Twenty-eighth Annual Conference of the Cognitive Science Society* (pp. 429-434). Hillsdale, NJ: Erlbaum.

Knowlton, B. J., & Holyoak, K. J. (2009). Prefrontal substrate of human relational reasoning. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences (4th edition)* (pp. 1005-1017). Cambridge, MA: MIT Press.

Kohn, A., & Smith, M.A. (2005). Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. *Journal of Neuroscience*, *25*, 3661–3673.

Kurtz, K. J., Miao, C.-H., & Gentner, D. (2001). Learning by analogical bootstrapping. *Journal of the Learning Sciences*, *10*, 417-466.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Lawrence, D. H., & DeRivera, J. (1954). Evidence for relational transposition. *Journal of Comparative and Physiological Psychology*, *47*, 465-471.

Leech, R., Mareschal, D., & Cooper, R. P. (2008). Relational priming: A developmental and computational perspective on the origins of a cognitive skill. *Behavioral and Brain Sciences*, *31*, 357-414.

Link, S. W. (1990). Modeling imageless thought: The relative judgment theory of numerical comparisons. *Journal of Mathematical Psychology*, *34*, 2–41.

Lu, H., Lin, T., Lee, A., Vese, L., & Yuille, A. L. (2010). Functional form of motion priors in human motion perception. *Advances in Neural Information Processing Systems*, *23,* 1495-1503. Cambridge, MA: MIT Press.

Lu, H., & Liu, Z. (2006). Computing dynamic classification images from correlation maps. *Journal of Vision*, *6*, 475-483.

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. L. (2008). Sequential causal learning in humans and rats. In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 195-188). Austin, TX: Cognitive Science Society.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*, 955-982.

MacKay, D. (1992). Bayesian interpolation. *Neural Computation, 4,* 415-447.

Mackay, D. (2003). *Information theory, inference and learning algorithms.* Cambridge, UK: Cambridge University Press.

Maouene, J., Hidaka, S., & Smith, L. B. (2008). Body parts and early-learned verbs. *Cognitive Science*, *32*, 1200-1216.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *23*, 431-467.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: W. H. Freeman.

Mangalath, P., Quesada, J., & Kintsch, W. (2004). Analogy-making as predication using relational information and LSA vectors. In K. D. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the Twenty-sixth Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Science*, *14*, 348-356.

McGonigle, B., & Chalmers, M. (1984). The selective impact of question form and input mode on the symbolic distance effect in children. *Journal of Experimental Child Psychology*, *37*, 525-664.

Morrison, R. G., Doumas, L. A. A., & Richland, L. E. (2010). A computational account of children's analogical reasoning: Balancing inhibitory control in working memory and relational representation. *Developmental Science*, *14*, 516-529.

Moyer, R. S. (1973). Comparing objects in memory: Evidence suggesting an internal psychophysics. *Perception & Psychophysics*, *13*, 180- 184.

Moyer, R. S., & Bayer, R. H. (1976). Mental comparison and the symbolic distance effect. *Cognitive Psychology,* 8, 228-246.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature*, *215*, 1519-1.520.

Neal, R. M. (1996). *Bayesian learning for neural networks*. New York: Springer-Verlag.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11-

28.

Nieder, A., & Miller, E. K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron*, *37*, 149–157.

Nirenberg, S. & Latham, P.E. (2003). Decoding neuronal spike trains: how important are correlations? *Proceedings of the National Academy of Sciences, USA*, *100*, 7348–7353.

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

Opfer, J. E., & Siegler, R. (2007). Representational change and children's numerical estimation. *Cognitive Psychology, 55*, 169-195.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*, 109-178.

Petrov, A. A. (2008). Relational priming plays a supporting but not leading role in adult analogy-making. *Behavioral and Brain Sciences*, *31*, 392-393.

Petrov, A. A., Dosher, B. A., & Lu, Z.-L. (2005). Perceptual learning: An incremental reweighting model. *Psychological Review*, *112*, 715-743.

Piazza, M., Izard, V., Pinel, P., Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron, 44*, 547-555.

Piazza, M., Mechelli, A., Price, C. J., & Butterworth, B. (2006). Exact and approximate judgements of visual and auditory numerosity: An fMRI study. *Brain Research, 1106*, 177-188.

Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron, 53*, 293-305.

Pinel, P., Piazza, M., Bihan, D. L., & Dehaene, S. (2004). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron*, *41*, 1-20.

Povinelli, D. J. (2000). *Folk physics for apes: The chimpanzee's theory of how the world works*. New York: Oxford University Press.

Ramscar, M., & Yarlett, D. (2003). Semantic grounding in models of analogy: An environmental approach. *Cognitive Science*, *27*, 41-71.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. Cambridge, MA: MIT Press.

Regier, T., & Carlson, L. A. (2001). Grounding spatial language in perception. *Journal of Experimental Psychology: General*, *130*, 273-298.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, *94*, 249-271.

Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*, 67-109.

Rogers, T. T., & McClelland, J. L. (2008). Précis of *Semantic cognition: A parallel distributed processing approach*. *Behavioral and Brain Sciences, 31*, 689-714.

Rust, N.C., Schwartz, O., Movshon, J.A., & Simoncelli, E.P. (2005). Spatiotemporal elements of macaque V1 receptive fields. *Neuron*, *46*, 945–956.

Salakhutdinov, R., & Hinton, G. (2009). Deep Boltzmann machines. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, *5*, 448–455.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of

cross-categorization. *Cognition, 120*, 1-25.

Siegel, M., Donner, T. H., & Engel, A. K. (2011). Spectral fingerprints of large-scale neuronal interactions. *Nature Reviews Neuroscience*, *13*, 121-134.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*, 1193-1216.

Silva, R., Airoldi, A., & Heller, K. (2007). *Small sets of interacting proteins suggest latent linkage mechanisms through analogical reasoning* (Tech. Rep. GCNU TR 2007-001). London: University College London, Gatsby Computational Neuroscience Unit.

Silva, R., Heller, K., & Ghahramani, Z. (2007). Analogical reasoning with relational Bayesian sets. In M. Mella & X. Shen (Eds.), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*.

Smith, L. B. (1989). From global similarities to kinds of similarities: The construction of dimensions in development. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 147-177). Cambridge, UK: Cambridge University Press.

Smith, L. B., Rattermann, M. J., & Sera, M. (1988). "Higher" and "lower": Comparative interpretations by children. *Cognitive Development, 3,* 341-357.

Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology, 24,* 99-142.

Taylor, E. G., & Hummel, J. E. (2009). Finding similarity in a model of relational reasoning. *Cognitive Systems Research*, *10*, 229-239.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279-1285.

Tunteler, E., & Resing, W. C. M. (2002). Spontaneous analogical transfer in 4-year-olds: A

microgenetic study. *Journal of Experimental Child Psychology*, *83*, 149-166.

Tunteler, E., & Resing, W. C. M. (2004). Age differences in patterns of spontaneous production of strategies for analogy problems among five- to eight-year-old children. *Educational and Child Psychology*, *21*, 74-88.

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, *32*, 279-416.

Turney, P. D., & Littman, M. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning*, *60*, 251–278.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141-188.

Uhlhaas, P. J, & Singer, W. (2010). Abnormal neural oscillations and synchrony in schizophrenia. *Nature Reviews Neuroscience*, *11*, 100-113.

van Boxtel, J., & Lu, H. (2011). Visual search by action category. *Journal of Vision, 11*, 1-14.

van Boxtel, J., & Lu, H. (2012). Signature movements lead to efficient search for threatening actions. *PLOS-ONE*.

Victor, J. D. (2005). Analyzing receptive fields, classification images and functional images: Challenges with opportunities for synergy. *Nature Neuroscience*, *8*, 1651-1656.

Wolfe, M. B. W., & Goldman, S. R. (2003). Use of Latent Semantic Analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behaviour Research Methods, 35*, 22-31.

Woocher, F. D., Glass, A. L., & Holyoak, K. J. (1978). Positional discriminability in linear orderings. *Memory & Cognition*, *6*, 165-173.

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

Yoshida, H., & Smith, L. B. (2005). Linguistic cues enhance the learning of perceptual cues. *Psychological Science*, *16*, 90-95.

CHAPTER 2:

THE DISCOVERY AND COMPARISON OF SYMBOLIC MAGNITUDES

Humans and other primates have sophisticated abilities to learn and make judgments based on relative magnitude. Magnitude comparisons are critical in making choices (e.g., which of two products is more desirable?), making social evaluations (e.g., which person is friendlier?), and in many other forms of appraisal (e.g., who can run faster, this bear or me?). In addition to making comparisons based on elementary perceptual dimensions (e.g., identifying the longer of two line segments or the brighter of two lights), people are able to make analogous judgments based on symbolic dimensions using information stored in memory (e.g., the relative size or intelligence of various animals). Non-human primates are also capable of at least rudimentary symbolic comparisons. For example, rhesus monkeys are capable of learning shapes (Arabic numerals) that correspond to small numerosities (1-4 dots), such that the shapes acquire neural representations overlapping those of the corresponding perceptual numerosities and can be compared on that basis (Diester & Nieder, 2007).

Striking parallels have been observed between perceptual and symbolic judgments. In particular, both perceptual and symbolic judgments yield a *distance* effect, such that the ease of judgments (indexed by accuracy and/or reaction time) increases with the magnitude difference between the objects being compared (e.g., Moyer, 1973; Moyer & Bayer, 1976; Moyer & Landauer, 1967). A symbolic distance effect is observed not only with quasi-perceptual dimensions such as size, but also with more abstract dimensions such as animal intelligence (Banks, White, Sturgill & Mermelstein, 1983) and with scalar adjectives of quality (e.g., *good*, *fair*; Holyoak & Walker, 1976). Non-human primates also exhibit a distance effect for judgments along various perceptual dimensions, including numerosity (Neider & Miller, 2003).

When judgments are made using contrastive polar concepts (e.g., "choose brighter" versus "choose dimmer", "choose better" versus "choose worse"), both perceptual (Audley & Wallis, 1964; Wallis & Audley, 1964; Petrusic & Baranski, 1989) and symbolic judgments also yield a *semantic congruity* effect: for objects with high values on the dimension, it is easier to judge which object is greater, whereas for objects with low values, it is relatively easier to judge which is lesser (e.g., Banks, Clark & Lucy, 1975; see Moyer & Dumais, 1978, for an early review). Like the distance effect, semantic congruity effects have also been obtained with monkeys (Cantlon & Brannon, 2005). A further phenomenon, the *markedness* effect, refers to the fact that for some pairs of polar adjectives, one (the "unmarked" form) is easier to process overall than the other (Clark, 1969). For example, the "unmarked" question "Which is larger?" tends to be answered more rapidly overall than the "marked" question "Which is smaller?" (Clark, 1969; Clark, Carpenter & Just, 1973). The impact of markedness implies that the congruity effect often takes the form of an asymmetrical interaction.

How Are Magnitudes Generated?

Numerous models of symbolic magnitude comparisons have been proposed, and we will review several of them below. However, in the present paper we focus on a question that (even though it is arguably the most basic of all) has seldom been asked, far less answered: where do subjective magnitudes come from? In the case of perceptual judgments with unidimensional stimuli (e.g., tones varying in loudness), it is reasonable to assume that a specific neural channel generates magnitudes. For symbolic comparisons, the tacit assumption has been that the long-term memory representation of each object being compared includes a magnitude value (perhaps with an associated variance), and that these magnitudes are simply retrieved and loaded into working memory, where a comparison process operates.

For a few types of symbolic comparisons, such as numerical magnitudes of digits, it may indeed be the case that each object has a pre-stored magnitude in long-term memory. But for more complex dimensions this assumption is questionable, and indeed quite unrealistic. Even symbolic size judgments, which are closely linked to perceptual features, are unlikely to always be based on pre-stored magnitudes, as size is actually a complex function of three-dimensional shape. Indeed, recent evidence indicates that although numerical magnitudes are automatically activated when reading integers, size magnitudes associated with animal names are activated only when the reader has the goal of making size comparisons (Hoedmaker & Gordon, 2013). People may have stored size values for a few "landmark" objects (e.g., an elephant or a mouse), but are unlikely to have pre-stored size values for less familiar animals (e.g., a beaver or a swordfish). The notion that magnitudes are pre-stored becomes yet more implausible for the wide range of dimensions on which people can make symbolic comparisons, especially in the interpersonal and social realm (e.g., intelligence, friendliness, religiosity, conservatism). Rather than being elementary components of concept meanings, magnitudes may often be derived, context-dependent features (Goldstone, 1994; Smith, Gasser & Sandhofer, 1997).

It follows that a comprehensive account of symbolic magnitude comparisons must begin with a model of how symbolic magnitudes are discovered. One general hypothesis is that magnitudes can be generated by operations performed on vectors of more elementary features associated with individual objects. Figure 2.1 provides a visualization of the sort of input that might underlie people's everyday knowledge of various types of animals. These representations were derived from norms of the frequencies with which participants at the University of Leuven generated features characterizing various animals (De Deyne et al., 2008; see Shafto, Kemp, Mansinghka, & Tenenbaum, 2011). Each animal in the norms is associated with a set of

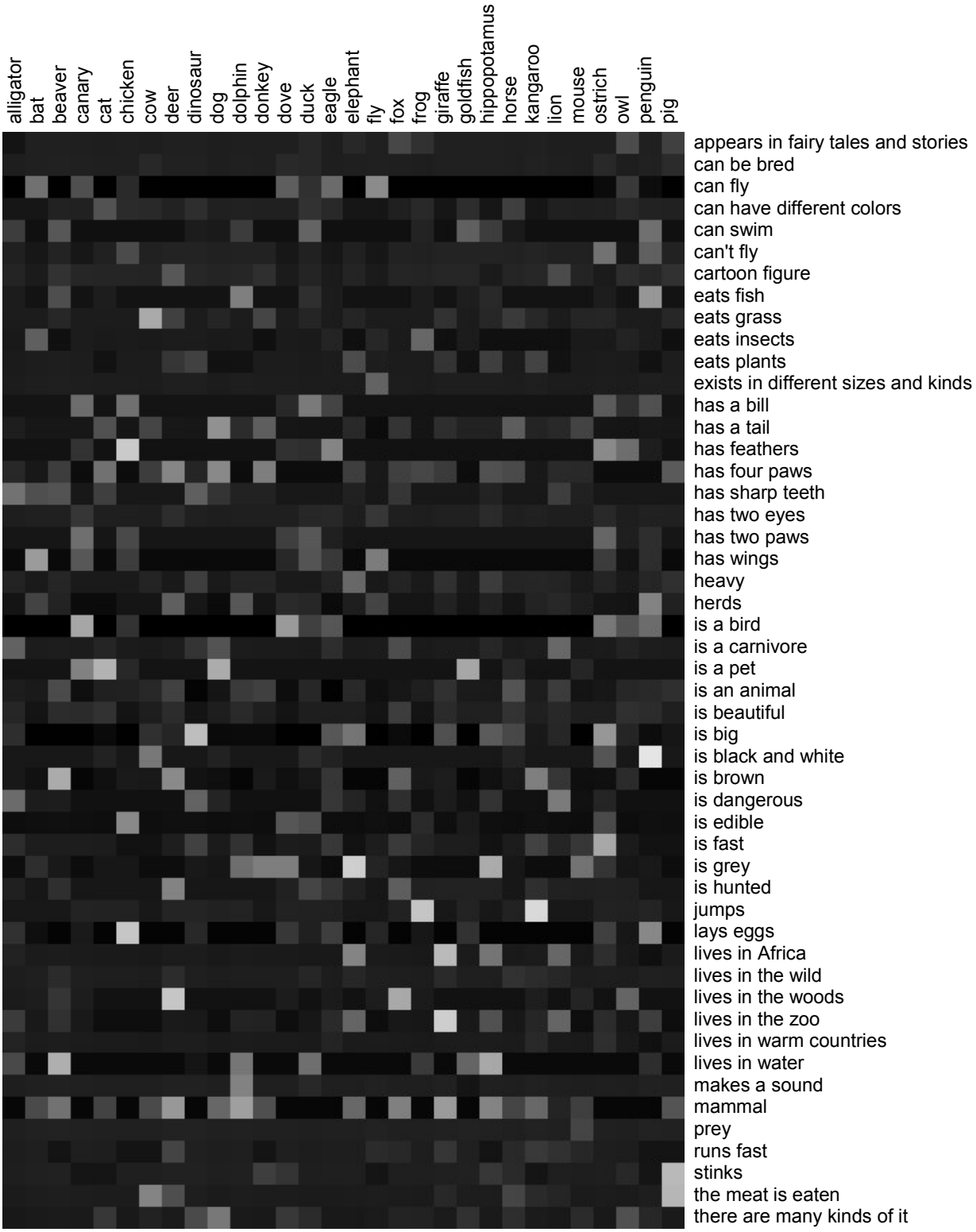


Figure 2.1. Illustration of Leuven vectors for some example animals. The cell intensities represent feature values (light indicates high frequency values, dark indicates low frequency values).

frequencies across more than 750 features. Figure 2.1 includes feature vectors for 30 example animals based on the 50 features most highly associated with a larger set of animal names (Lu, Chen & Holyoak, 2012). Although these "Leuven vectors" presumably only approximate people's knowledge about animal concepts, they have the great virtue of being derived from independent sources of data, rather than being hand-coded. The simulations reported in the present paper are based on inputs extracted from the Leuven vectors, as well as similar feature vectors created using the topic model (Griffiths, Steyvers, & Tenenbaum, 2007).

Could individual Leuven features be directly used as measures of magnitude? One might have supposed, for example, that the value of the feature "is big" would be sufficient to predict relative size. But although this dimension is indeed the single most important factor predicting size, it is far from sufficient. The Leuven features were derived from the frequencies with which participants generated attributes, and animals for which their large size is salient (often in reference to a subcategory) tended to have higher feature values for "is big" (e.g., based on a comparison of feature values for that attribute alone, the Leuven dataset indicates that an eagle is larger than a hippopotamus). To address this problem we need distributed representations that can be used to compute derived magnitude dimensions.

To provide such distributed representations, Lu et al. (2012) developed *Bayesian Analogy with Relational Transformations* (BART), a model of how one-place scalar adjectives (e.g., *large*, *smart*) and two-place comparative relations (e.g., *larger*, *smarter*) can be learned from non-relational feature vectors. Using various inputs, including Leuven vectors and vectors derived using the topic model (Griffiths et al., 2007), the model was applied to the acquisition of concepts related to four continuous dimensions: size, ferocity, speed and intelligence. BART incorporates information from a prior probability distribution over a space of weights, as well as

examples of animal pairs that instantiate a relation, to obtain a posterior distribution over the weight space, which is used to predict whether the relation holds for novel pairs. The representations of relational concepts created by BART for each of the four magnitude dimensions of interest turned out to be highly distributed, based on at least 20 statistically predictive features (see Lu et al., 2012, Figure 10, pp. 634-635).

The simulation results reported by Lu et al. (2012) suggest that concepts related to symbolic magnitudes can be discovered by inductive learning, rather than simply assumed to be directly available in long-term memory. Moreover, the Bayesian approach in general (and the BART model in particular) implies that magnitudes will be represented not as deterministic values, but rather as probability distributions. The probabilistic framework is in agreement with the intuition that symbolic magnitudes (e.g., the size of a kangaroo, the intelligence of a goat) are "fuzzy" rather than firm, and thus judgments related to these attributes are susceptible to the influence of context.

Our goal in the present paper is to provide an integrated account of how symbolic magnitudes, represented in working memory as probability distributions, can be created and then used to make comparative judgments. We will first briefly review previous accounts of the three major phenomena observed in studies of comparative judgment: symbolic distance effects, semantic congruity effects, and markedness. We will then show how a model incorporating assumptions about the attentional control of magnitude representations in working memory can provide a unified account of these core phenomena.

Alternative Models of Symbolic Magnitude Comparisons

We will not attempt an exhaustive review of the large literature on mental magnitude comparisons, but rather will focus on findings that give rise to some of the principles we include

in our current model (for broader reviews of work with humans see Moyer & Dumais, 1978; Petrusic, 1992; for a review of work with non-human primates see Cantlon, Platt & Brannon, 2009).

There is virtually complete consensus among current researchers that the ubiquitous distance effect reflects some form of internalized representation of magnitude akin to positions on a number line, such that larger magnitudes are more readily discriminable. This notion goes back at least to Moyer (1973), who referred to an "internal psychophysics" for symbolic comparisons. Behavioral studies have identified striking parallels between symbolic distance effects and those observed in overt perceptual comparisons (e.g., Audley & Wallis, 1964; Moyer & Bayer, 1976, Holyoak & Patterson, 1981). As in the case of perceptual comparisons, the pattern of difficulty for symbolic comparisons suggests that internal magnitudes are typically compressed such that subjective magnitude differences decrease as the absolute magnitudes of the objects being compared increase (Shepard, Kilpatric & Cunningham, 1975). More recent work has provided strong evidence that humans and other primates are equipped with specialized neural circuitry for dealing with approximate magnitude on various dimensions (e.g., Cantlon, Brannon, Carter & Pelphrey, 2006; Dehaene & Changeux, 1993; Piazza et al., 2004, 2006, 2007; Pinel, Piazza, Bihan & Dehaene, 2004).

Several models for magnitude comparisons have been proposed (for a review see Petrusic, 1992). The evidence distinguishing among them mainly involves the congruity and markedness effects. The congruity effect has been interpreted in multiple ways. An expectancy model (Banks & Flora, 1977; Marschark & Paivio, 1979) assumes that the congruity effect arises because the comparative is presented prior to the stimulus pair, enabling the person to prepare in some way for stimuli within a certain range (e.g., either small or large objects). However, robust

congruity effects are found even when the comparative is presented *after* the stimuli to be compared, in a design in which questions about multiple dimensions were intermixed (Holyoak & Mah, 1981). Other studies yielded similar disconfirmatory findings (Banks et al., 1983; Howard, 1983; Shoben, Sailor, & Wang, 1989).

A related explanation of the congruity effect attributes the phenomenon to differential frequency of association between each comparative and items of various magnitudes (i.e., the "greater" comparative may be more often used with items of high magnitude, and the reverse for the "lesser" comparative). However, Ryalls and Smith (2000) taught adults novel comparatives, and found that a congruity effect arose even when the training set was designed to eliminate any correlation between the form of the comparative and the magnitude of items. These and other findings concerning acquisition of comparative terms (Ryalls, Winslow & Smith, 1998) suggest that the congruity effect reflects the meaning of the contrastive terms, rather than unbalanced presentation frequencies during learning that might influence expectancies about items.

A frequency-based explanation has also been offered for markedness effects, as unmarked forms of adjectives are typically used more frequently than the corresponding marked forms. Often the marked term is aptly applied only to the range of magnitudes extending from the negative pole to the midpoint, whereas the unmarked term can be aptly applied across the full magnitude range (Clark, 1969). However, the finding of a markedness effect in monkeys, in a design in which the two forms of the implicit query occurred on an equal number of trials during training, suggests that markedness effects cannot be fully explained by unequal frequency of linguistic use (Cantlon & Brannon, 2005).

The semantic coding model (Banks et al., 1975; Banks, Fujii, & Kayra-Stuart, 1976) attributes the congruity effect to categorical codes based on language (e.g., "large" and "small").

In this model, the congruity effect reflects systematic differences in the probability that the codes for the objects will match the linguistic form of the comparative. Although the model provides a good quantitative fit to some data sets (Banks et al., 1976), it faces a number of problems as a general explanation of symbolic comparisons. Because it is based on linguistic codes, the model is severely strained by the fact that distance, congruity and markedness effects are also observed with non-linguistic primates, such as monkeys (Cantlon & Brannon, 2005; Cantlon et al., 2009). Also, the model cannot explain evidence that similar effects are observed in direct judgments of discriminability among ordered items (e.g., the form of comparative used in the question influences the relative spacing of cities along an east-west dimension as recovered by scaling methods; Holyoak & Mah, 1982). Finally, the model predicts that the magnitude of the congruity effect will be independent of factors that influence decision difficulty (Banks et al., 1975). However, there is considerable evidence that the magnitude of the congruity effect in fact varies systematically with decision difficulty (Petrusic, 1992; Petrusic & Baranski, 1989; Shaki, Leth-Steensen, & Petrusic, 2006).

Reference-Point Models

The final major class of models (and the one most relevant to the present proposal) includes those that locate the congruity and markedness effects within the decision stage itself. The intuitive idea is that when judging (for example) whether an elephant is larger than a hippo, the subjective magnitude difference is in fact more discriminable than when judging whether an elephant is *smaller* than a hippo. Such discriminability effects might arise by a mechanism through which the form of the question modulates magnitude representations in working memory. A number of specific models have been proposed, which share the hypothesis that the polarity of the comparative serves to establish a *reference point* at or near the corresponding end

of the continuum, and that magnitude differences between objects close to the reference point are discriminated more easily than otherwise comparable differences between objects far from the reference point (Marks, 1972; Jamieson & Petrusic, 1975; Holyoak, 1978; Holyoak & Mah, 1982). Holyoak (1978) argued that attending to a reference point at the congruent extreme of a dimension aids in coding the polarity of the question (i.e., distinguishing between "choose greater" versus "choose lesser" for a specific pair of comparatives).

Reference-point models are not inherently linguistic, and hence can in principle be applied to comparative judgments in non-linguistic species (Cantlon et al., 2009); they can accommodate the influence of the question form on direct discriminability judgments (Holyoak & Mah, 1982); and in some variants (Marks, 1972) they predict the general finding that congruity effects are larger when decisions are more difficult (Petrusic, 1992; see Banks et al., 1975, for a derivation). In addition, reference-point models can potentially explain another critical property of the congruity effect, which is that it is sensitive to the range of magnitudes exhibited in the stimulus set. For example, if the presented stimuli are all relatively small animals (e.g., smaller than a dog), then the *relatively* large animals within this restricted set (e.g., rabbit and beaver) will show an advantage for "choose larger" over "choose smaller" (Čech & Shoben, 1985; Čech, Shoben & Love, 1990; see also Petrusic & Baranski, 1989). Similar range effects have been observed in studies of comparative judgments by monkeys (Jones, Cantlon, Merritt & Brannon, 2010). It is natural to suppose that an observer could strategically shift reference points to reflect the magnitude range of the presented stimuli.

A number of explanations of how a reference point exerts its effect have been proposed. Jamieson and Petrusic (1975) and Holyoak (1978) suggested that observers assess the *ratio* of distances from each object to the reference point, rather than simply taking the difference. The

distance ratio provides good quantitative fits to some data sets, including data from experiments in which an *explicit* reference point is specified at an intermediate point on the scale (e.g., judging which digit, 2 or 3, is closer to 5; Holyoak, 1978). However, other data sets are less well fit by the quantitative form specified by the distance ratio. For example, although scale compression triggered by the form of the comparative can be observed in non-speeded discriminability judgments, the effects tend to be smaller than the distance ratio would predict (Holyoak & Mah, 1982).

Perhaps reference points directly alter mean magnitudes of items, expanding differences close to the reference point relative to differences far from it. However, shifts in discriminability might instead reflect changes in *variances* of magnitude, rather than in mean values. Marks (1972), building on the assumptions of signal detection theory, suggested that internal magnitudes are represented as distributions that encode uncertainty, which is reduced in the region of a reference point (i.e., the variance or "discriminal dispersion" of magnitude representations is lower for magnitudes close to a reference point). Marks did not develop a quantitative model; however, related reference-point models have introduced evidence-accrual mechanisms, consistent with the basic idea that comparative judgments are based on iterative sampling from magnitude distributions (see Petrusic, 1992). The model we propose in the present paper adopts the key idea proposed by Marks (1972), that the form of the comparative affects discriminability by dynamically altering magnitude variances based on distance from a reference point.

Reference-point models in general, including Marks's (1972) specific proposal of the modulation of variance as a mechanism, are broadly consistent with the wider literature on attentional influences on magnitude representation. Miller's (1956) classic paper focused on the

limited channel capacity available to make absolute magnitude judgments (and explicitly linked signal variance with information transmission). In psychophysical work, Luce, Green, and Miller (1976) proposed that observers are able to strategically control *attention bands*, selectively monitoring a relatively narrow intensity range. Luce et al. suggested that neural variability of the internal representation of intensities will be reduced within the favored attention band, yielding greater sensitivity as measured by signal-to-noise ratio. Nosofsky (1983) found evidence that observers can indeed strategically shift attention to a specific intensity band, thereby facilitating discrimination of tones in the favored region. He also argued, based on a literature review, that this flexible allocation of attention to a magnitude band is limited to just one such location along a continuum; hence performance falls off monotonically with distance from the favored region.

A reference-point explanation has also been offered for the markedness effect. It is possible that markedness, like the congruity effect, fundamentally arises from the inherent meaning of comparatives, and in particular from the fact that many comparative pairs have an inherent asymmetry in their polarity: one end is positive or "greater" and the other end is negative or "lesser". If markedness is rooted in the underlying meaning of comparatives, then the effect might reflect some additional processing difficulty encountered in maintaining precise magnitude distributions when focusing on the "negative" or "lesser" pole. Marks (1972) suggested that the markedness effect could be modeled by assuming that the precision of magnitude representations falls off more rapidly moving from the lesser than from the greater reference point. We will also adopt this assumption, which serves to integrate the markedness effect with the semantic congruity effect.

In sum, psychophysical work provides broad support for the hypothesis that observers can selectively modulate attention to a favored region along a magnitude continuum. Given the

many established parallels between perceptual and symbolic magnitude comparisons, it is natural to hypothesize that similar mechanisms operate in symbolic tasks. Moreover, reference points established by the form of the question and the range of the presented stimuli can readily be viewed as cues that establish attention bands. Marks' (1972) proposal that such modulation operates by influencing the variance of magnitude representations provides a key theoretical element in the model we will describe below. The hypothesis that attention operates in part by modulating variability in an internal representation is also consistent with findings concerning visual detection and discrimination tasks (Dosher & Z. Lu, 2000; Rahnev et al., 2011).

Magnitude Representations in BARTlet

Multiple Levels of Representation for Comparative Relations

Our goal in the present paper is to provide a unified model of how symbolic magnitudes can be discovered and used to make comparative judgments. The model we propose, termed BARTlet (i.e., the diminutive form of BART), builds on the learning capability of BART (Lu et al., 2012) but makes simpler representational assumptions. A key idea incorporated in both models is that learning can be bootstrapped by incorporating *empirical priors*—a "favorable" initial knowledge state derived from some related but simpler learning task. In BART, learning of explicit comparative relations (two-place predicates, such as *larger*) is guided by empirical priors derived from initial learning of one-place predicates (e.g., *large*, *small*).

BARTlet also emphasizes the role of bootstrapping operations that allow learning at a lower level to guide subsequent learning at a higher level. Although we do not aim to provide a serious developmental model (which would require a detailed specification of the inputs available to children), we do aim to implement a learning process that can acquire magnitude information from inputs of realistic complexity. Moreover, we focus on learning from inputs that

were not hand-coded, but rather were generated by an autonomous process (i.e., independently of the modelers). We use two different sets of inputs as a further way of showing that the learning model is robust and does not depend on specific details of what features are included in the input.

Given that humans are able to make magnitude comparisons between objects that they may never have previously considered together (e.g., which is larger, a walrus or a fox?), our goal was to create a model that can learn from a limited set of examples and then generalize to novel comparisons. At the same time, we also wished to capture the significant commonalities between magnitude comparisons performed by humans and by non-human animals. BART learns explicit two-place relations representing comparatives (e.g., *larger*). In addition to supporting generalization to new animal pairs, these explicit relations can be systematically transformed to solve analogies based on higher-order relations between different pairs of polar adjectives (e.g., *larger* : *smaller* :: *faster* : *slower*). However, such high-level reasoning is beyond the capability of most animals (indeed, it may be uniquely human; Penn, Holyoak & Povinelli, 2008). In contrast, basic comparative judgment appears to be similar in humans and symbol-trained monkeys (Moyer & Landauer, 1967; Diester & Neider, 2010). Many other species, such as rats, can respond on the basis of relative magnitude when shown perceptual stimuli that vary along simple continua (Lawrence & DeRivera, 1954). Thus as a model of basic comparative judgment, the explicit relational representations acquired by BART appear to be over-powerful.

Figure 2.2 sketches different levels of representation that may be involved in making magnitude comparisons and reasoning with comparative relations (for a similar representational hierarchy, see Halford, Wilson, & Phillips, 1998, 2010). At a pre-categorical level (i.e., a level of representation that does not involve categorical distinctions or explicit predicates), simple associative or statistical mechanisms can perform basic magnitude comparison and learn from

ordered pairs. For example, under certain conditions the Rescorla-Wagner model of associative learning (Rescorla & Wagner, 1972; see Wynne, 1995) can model qualitative aspects of animals' ability to infer transitivity of choice (e.g., after being trained on only adjacent pairs of stimuli exhibiting the reward pattern A > B, B > C, C > D, D > E, an animal will tend to choose B over D). Other associative models can account for learning of orderings across a broader range of conditions (von Fersen, Wynne, Delius, & Staddon, 1991).

In the present paper we adopt a statistical model capable of learning continuous-valued attributes from a *partial* ordering of examples (Parikh & Grauman, 2011). This model (described more fully below) learns to rank objects based on the algorithm of a support vector machine with certain additional constraints, and hence will be referred to as *RankSVM*. RankSVM extracts continuous dimensions of attributes by learning weights on object features, such that the maximum number of ranking constraints is satisfied for the training data. Note that RankSVM does not create representations that categorize attributes in a binary manner (e.g., elephant is large, not small); rather, this algorithm yields representations sensitive to relative order on a dimension (e.g., elephant is ordered before horse in size, horse is ordered before cat). Parikh and Grauman successfully tested their RankSVM model on problems involving comparisons of realistic visual images. Though we do not claim that the algorithm is psychologically realistic, it provides a functional model that can deal with partial orderings of elements coded by high-dimensional feature vectors. The function performed by this model is consistent with empirical evidence that both animals and humans can learn simple orderings from a partial set of ordered pairs (Merritt & Terrace, 2011; Wynne, 1995; Trabasso & Riley, 1975; Woocher, Glass & Holyoak, 1978). Moreover, its output (feature weights) can readily be translated into empirical priors for learning one-place predicates.

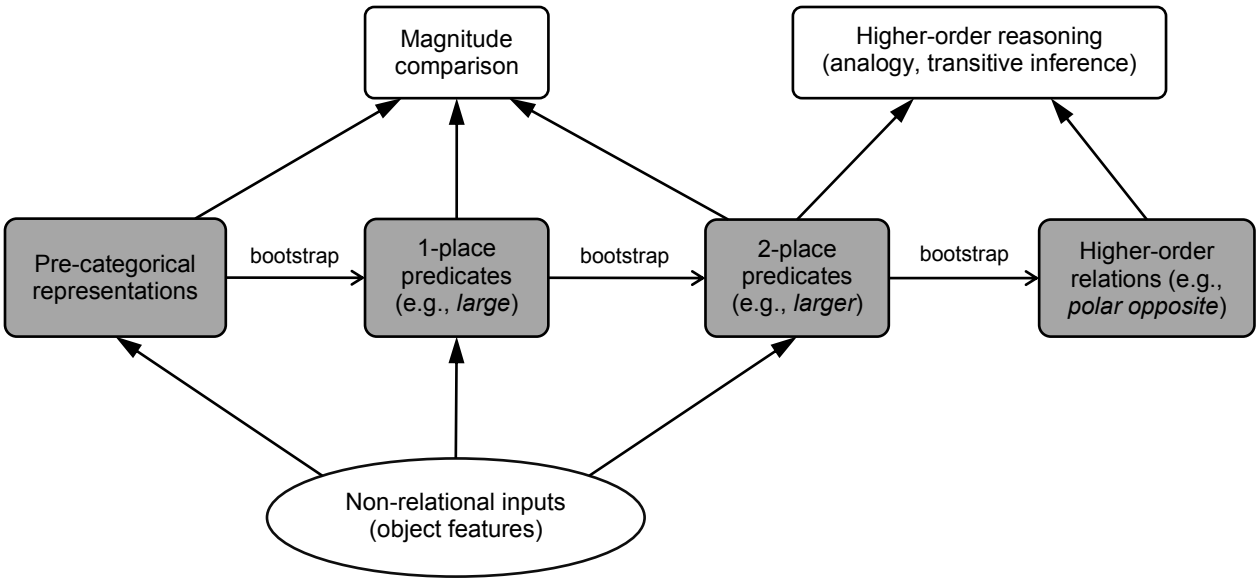


Figure 2.2. Relationships among inputs (bottom), levels of representation (middle row) and tasks (top) involving magnitude-related concepts. Pre-categorical processes bootstrap acquisition of one-place predicates (the domain of BARTlet), which in turn can bootstrap acquisition of two-place predicates and ultimately higher-order relations (the domain of BART). The lower levels have access to external inputs (non-relational feature vectors for individual objects) and can be used to perform comparisons based on dimension-specific magnitudes; the higher levels operate (in part or entirely) on internally-generated representations, and can be used to perform more abstract types of reasoning, such as higher-order analogical reasoning and transitive inference.

The next level of representation corresponds to one-place predicates (e.g., *large*), which in essence define categories of objects based on their magnitudes on some underlying dimension. Both behavioral and neural evidence indicates that monkeys are capable of acquiring categorical representations (e.g., Cromer, Roy, & Miller, 2010; Freedman, Riesenhuber, Poggio, & Miller, 2001). For realistic stimulus sets, learners are unlikely to ever compare all possible pairs of N objects (a quantity that scales with N^2) on every dimension of interest. Categorical information about individual objects (a quantity that scales linearly with N) provides an efficient additional input for learning magnitudes. As described below, BARTlet learns one-place predicates from facts such as "a whale is large," bootstrapping from empirical priors provided by RankSVM.

BARTlet thus integrates dimensional information provided by examples of ordered pairs (via RankSVM) with categorical information, thereby refining its knowledge about dimensional magnitudes.

The additional levels of representation sketched in Figure 2.2 are based on explicit relations (i.e., predicates with more than one argument, such as *larger*). Whereas BARTlet uses a comparison operator (based on signal detection theory) to compare relative magnitudes derived from one-place predicates, BART creates two-place predicates that in effect represent the comparison operator as part of the relation itself. As described by Lu et al. (2012), these more complex relational representations (arguably unique to humans) can be learned by bootstrapping from one-place predicates, and can in turn be bootstrapped to generate higher-order relations between relations (e.g., "polar opposite"). We will return to the topic of levels of representation in the General Discussion. For now, we simply note that the goal of the present paper is to show that BARTlet, a model limited to one-place predicates (i.e., without access to explicit two-place comparatives) is capable of basic symbolic magnitude comparisons.

Deriving Magnitudes from Unstructured Feature Vectors

In BARTlet, magnitudes are created by applying learned dimension-specific weights to more primitive features of objects. Magnitudes are represented in working memory as derived features that follow specified probability distributions, modulated by reference points. BARTlet (like BART) represents a one-place predicate (e.g., *large*) using a joint distribution of weights over object features, as illustrated in Figure 2.3 (bottom). A predicate is learned by estimating the probability distribution $P(\mathbf{w} \mid \mathbf{X_S}, \mathbf{\Phi_S})$, where $\mathbf{X_S}$ represents the feature vectors for objects in the training set, the subscript **S** indicates the set of training examples, and $\mathbf{\Phi_S}$ is a set of binary indicators, each of which (denoted by Φ) indicates whether a particular object instantiates the

predicate or not. The vector **w** constitutes the learned predicate representation, which can be interpreted as weights reflecting the influence of the corresponding feature dimensions in **X** on judging whether the predicate applies. Formally, the posterior distribution of weights can be computed by applying Bayes' rule using the likelihood of the training data and the prior distribution for **w**:

$$P(\mathbf{w} \mid \mathbf{X_S}, \mathbf{\Phi_S}) = \frac{P(\mathbf{\Phi_S} \mid \mathbf{w}, \mathbf{X_S}) P(\mathbf{w})}{\int_{\mathbf{w}} P(\mathbf{\Phi_S} \mid \mathbf{w}, \mathbf{X_S}) P(\mathbf{w})}. \tag{2.1}$$

The likelihood is defined as a logistic function for computing the probability that an object instantiates the predicate, given the weights and feature vector:

$$P(\Phi = 1 \mid \mathbf{w}, \mathbf{x}) = \left(1 + e^{-\mathbf{w}^{\mathbf{T}}\mathbf{x}}\right)^{-1}. \tag{2.2}$$

The prior distribution $P(\mathbf{w})$ in Eq. (2.1) is assumed to follow a normal distribution with mean and covariance matrix as parameters. To define the prior, the BARTlet model relies on initial learning at a simpler representational level to bootstrap subsequent learning at a more complex level. Specifically, the BARTlet uses weights learned by RankSVM as means and a standardized covariance matrix (e.g., variance of 1, covariance of 0) as the empirical prior for learning one-place predicates. The RankSVM model takes ordered pairs as inputs, where each object is represented by a feature vector. Its algorithm is a support vector machine, which in essence performs linear regression with an additional constraint to minimize weight values. The novel feature of RankSVM is the further addition of a penalty for violating the given partial ordering of objects (for a full mathematical description, see Parikh & Grauman, 2011).

RankSVM was developed for machine-learning purposes, and we make no claim for the psychological plausibility of its algorithm. However, there is ample evidence that many types of animals can learn simple orderings from a partial set of pairs. For both animals (Merritt &

Terrace, 2011; Wynne, 1995) and humans (Trabasso & Riley, 1975; Woocher et al., 1978), orderings are typically learned "from the ends in", with the extreme or "landmark" objects being acquired prior to those that lie closer to the middle of a continuum. In the present simulations, we first trained the RankSVM model with ordered pairs that mainly involved the half dozen animals with the highest or lowest values on the relevant continuum. The resulting weights then served as empirical priors for BARTlet, which in turn received relatively extreme animals as examples (positive or negative) of each one-place predicate.

From Weight Distributions to Derived Magnitudes

The weight distribution that BARTlet acquires for a one-place predicate such as *large*provides all the information required to specify the magnitude of each animal on each dimension. As shown in Figure 2.3, the magnitude of an object on a dimension (e.g., size) can be derived as a weighted sum of the feature values **x** for this object:

$$M = \sum_i w_i x_i, \tag{2.3}$$

This weight distribution codes not only first-order statistics (means, μ_{w_i}), but also second-order statistics (variances and covariances) that capture the uncertainty of the estimated weights, as well as inter-weight correlations. Because the weights are normally distributed, the derived magnitude variable *M* follows a normal distribution with a mean of μ_M and a variance of σ_M^2 , which are calculated according to:

$$\mu_M = \sum_i \mu_{w_i} x_i, \tag{2.4}$$

$$\sigma_M^2 = \sum_i x_i^2 Var\left(w_i\right) + \sum_i \sum_{j \neq i} x_i x_j Cov\left(w_i, w_j\right). \tag{2.5}$$

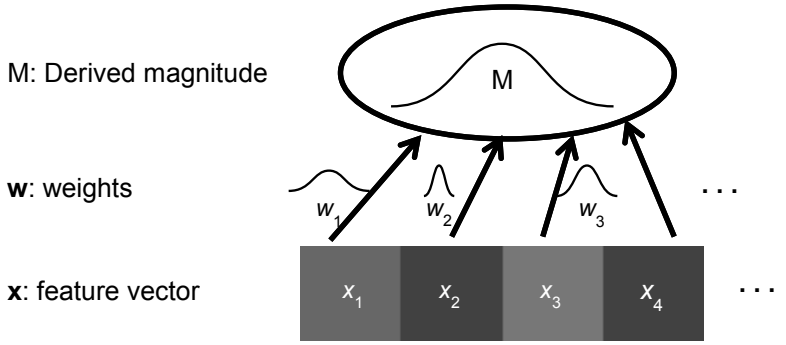


Figure 2.3. In the BARTlet model, weight distributions derived from one-place predicates (e.g., *large*) are applied to the feature vector for an individual animal to compute a derived magnitude (normally distributed) for that object.

The variance of the derived magnitude reflects uncertainty about the magnitude value and can be modulated by factors such as attention, in a manner that we will describe. Importantly, BARTlet does not make use of explicit relations when making symbolic comparisons. Rather, BARTlet evaluates which of two objects is larger (or smaller, faster, etc.) by the more primitive operation of comparing the derived magnitudes of the two individual objects, using the framework of signal detection theory.

Reference Points in Symbolic Comparisons

BARTlet adds two explicit algorithmic assumptions: People operate under limited capacity to maintain veridical estimates of magnitudes in working memory, and the focus of attention on a particular magnitude range is controlled by reference points. Because the representation of magnitudes includes uncertainty, it is straightforward to implement the key assumption that magnitude discriminability is influenced by reference points, which operate by influencing the associated variances (Marks, 1972). BARTlet selectively attends to a particular region of the relevant dimensional spectrum (e.g., the high end of the size spectrum when choosing the larger of two objects), leading to greater discriminability between objects in that favored region (Figure 2.4). The distance to a reference point is calculated by comparing an

object to a reference object, and this distance is used to scale the magnitude variance of the object. As a result, magnitudes of objects closer to the reference point have greater precision (i.e., less uncertainty), whereas the magnitudes of objects farther from the reference point have less precision.

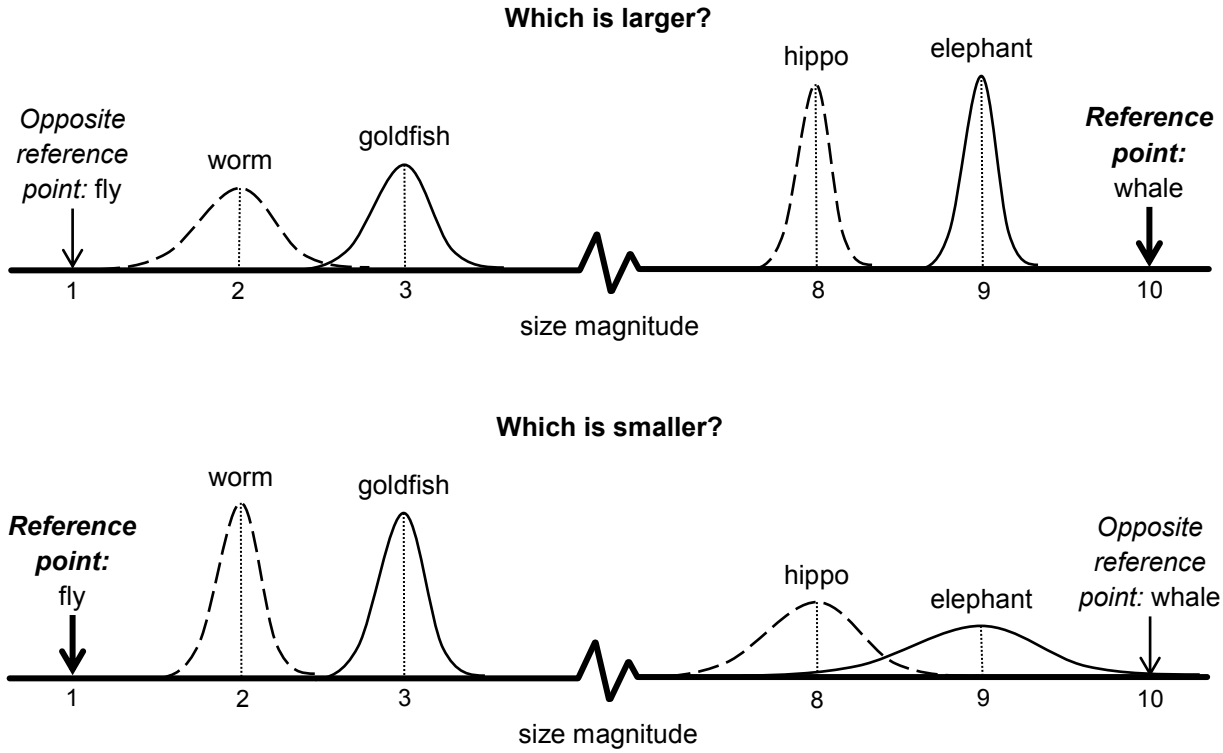


Figure 2.4. BARTlet's representations of magnitudes in working memory. Based on the assumption that reference points at the extremes control attention, variances of magnitude distributions increase with distance from the reference point at the extreme consistent with the question. The increase in variance with distance from the reference point is assumed to be greater for the marked form of a comparative.

BARTlet generates magnitude values (*M*) based on unmarked one-place predicates (e.g., *large*), and hence *M* values are positive and monotonic relative to the unmarked form (e.g., large animals are associated with high size values, and small animals with low size values, rather than the reverse). We assume that because the unmarked form of the question requires reversing the

natural scale (e.g., "smaller" focuses attention on low magnitudes), precision diminishes more quickly with distance from the reference point in the case of the marked comparative.

Specifically, BARTlet uses the following procedure to answer a comparative query such as, "Which is larger, an elephant or a giraffe?" First, the model establishes a reference point based on the comparative involved in the question and all presented stimuli (i.e., the context). Because the comparative in this question is *larger*, the reference point is taken to be the object among the presented stimuli with the highest mean magnitude on the size dimension. (If the comparative were instead *slower*, the reference point would be the object with the lowest mean magnitude on the speed dimension.) Based on the selected reference point, the model computes *D*, the maximum possible distance from the reference point within the current context (i.e., the subjective range on the relevant dimension). This value is simply the absolute difference in mean magnitudes between the reference point and the opposite-extreme reference point. For *larger*, the opposite-extreme reference point is the object among the presented stimuli with the lowest mean size magnitude.

The model computes the means and unscaled variances according to Eqs. (2.4) and (2.5) for the magnitudes of the two objects being compared. In our example, the mean and variance of the size magnitude is computed for both the elephant and the giraffe. Then, for each object being compared, the model computes δ, a measure of the distance between that object and the reference point as a proportion of the maximum possible distance from the reference point.[1] This value corresponds to the absolute difference between the mean magnitude of the object and of the reference point, divided by *D*. For each object being compared, the model scales the variance of its magnitude by $\alpha e^{\beta\delta}$, where α is an intercept parameter and β is a slope parameter, both free parameters. The specific parameter values were selected to be consistent

with the qualitative assumptions of the model. In our simulations, α was set to 0.1, implying that the variance of an object's magnitude is decreased by 90% when that object's mean magnitude is equal to that of the reference point. The values of β were selected so as to yield magnitude variances that are about 10 times (for unmarked relations; β = 4.6) or 20 times (for marked relations; β = 5.3) as high as the original variances when an object is maximally distant from the reference point. Thus, magnitude variances are assumed to increase more rapidly for marked relations than for unmarked relations as distance from the reference point increases (cf. Marks, 1972). In the present model, variances increase exponentially with distance from the reference point; however, a variety of neural mechanisms for gain control could potentially implement the impact of attention on gain control (Dosher & Z. Lu, 2000; Rahnev et al., 2011; for a review see Reynolds & Chellazzi, 2004).

Measuring Discriminability between Magnitudes

BARTlet models the discriminability between magnitudes of two objects that are made available to a comparison process. Based on signal detection theory, a natural measure of discriminability is d_a, which is the variant of d' appropriate when variances are unequal (Wickens, 2002, p. 65):

$$d_a = \frac{\mu_{M_1} - \mu_{M_2}}{\sqrt{\left(\sigma_{M_1}^2 + \sigma_{M_2}^2\right)/2}}. \tag{2.6}$$

A complete model of symbolic magnitude comparisons needs to specify a decision process that would translate degree of discriminability into accuracy and reaction time for comparative judgments. For example, the decision diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2010; Ratcliff, Van Zandt, & McKoon, 1999) is an extension of signal detection theory to the time domain, accumulating information continuously on the basis of repeated samples

(also see Link, 1990; Petrusic, 1992). If applied to comparative judgment, a theoretical measure of discriminability, such as d_a, could be used to predict the average value across repeated samples (corresponding to the mean of the drift rate in a diffusion process). Because our present focus is on variables that influence discriminability (i.e., information quality), rather than on the decision process *per se*, we will simply use BARTlet to make qualitative predictions of decision difficulty, based on values of d_a. We assume (as the diffusion model predicts) that decreases in discriminability will make the decision process more difficult, yielding slower and/or less accurate comparative judgments.

Simulations of Symbolic Magnitude Judgments Using Leuven Vectors

Predicting Human Magnitude Ratings

We first evaluated whether the *M* values learned by BARTlet in fact reflect the subjective magnitudes of animals on the relevant dimensions. The "ground truth" for all training examples and test pairs was provided by norms derived from ratings by college students on the dimensions of size, ferocity, intelligence and speed (Holyoak & Mah, 1981). For the animals used in the simulations reported in the present paper, intercorrelations among the four dimensions were moderate, ranging from .38 (size with speed) to .60 (size with fierceness). For our first set of simulations, we identified a set of 44 animals that also appeared in the Leuven norms (de Deyne et al., 2008). Each animal was represented by a vector of 50 continuous-valued features (see Lu et al., 2012, pp. 631-632, for a description of how the Leuven vectors were created).

As described earlier, learning of one-place dimensional predicates (*large*, *fierce*, *intelligent*, *fast*) proceeded in two stages. First, RankSVM was provided with the ordering for each of the top three and bottom three animals on the relevant dimension relative to all other animals, plus an additional 100 pairwise orderings selected at random from the pool of all

possible pairs of 44 animals.[2] The mean weights estimated by RankSVM (linearly scaled by a factor of 5 to roughly match the range of weights BARTlet would infer from an uninformative prior) became the empirical priors on weight means for BARTlet.[3] As RankSVM does not provide a covariance matrix, an uninformative prior (variances = 1, covariances = 0) was used. Second, BARTlet was provided with the 20 animals with the highest values (positive examples) and the 20 with the lowest values (negative examples) on the relevant dimension. These training examples were drawn from the entire pool of 129 animals in the Leuven norms. The resulting weight distributions across the 50 features of the Leuven inputs (Figure 2.1) were highly distributed, based on at least 20 statistically predictive features for each of the four magnitude dimensions of interest.

The weight distribution for each one-place predicate was used to calculate *M* values for each animal, as described earlier. Figure 2.5 shows the scatter plots of mean *M* values versus human magnitude ratings for each of these dimensions. Spearman rank-order correlations ranged from .86 to .96 for the four dimensions. These results indicate that magnitude values, derived from weight distributions acquired by BARTlet's learning mechanism from large, independently-generated feature vectors (Leuven vectors; see Figure 2.1), are quite accurate in predicting human judgments about subjective magnitudes of animals on the four dimensions.

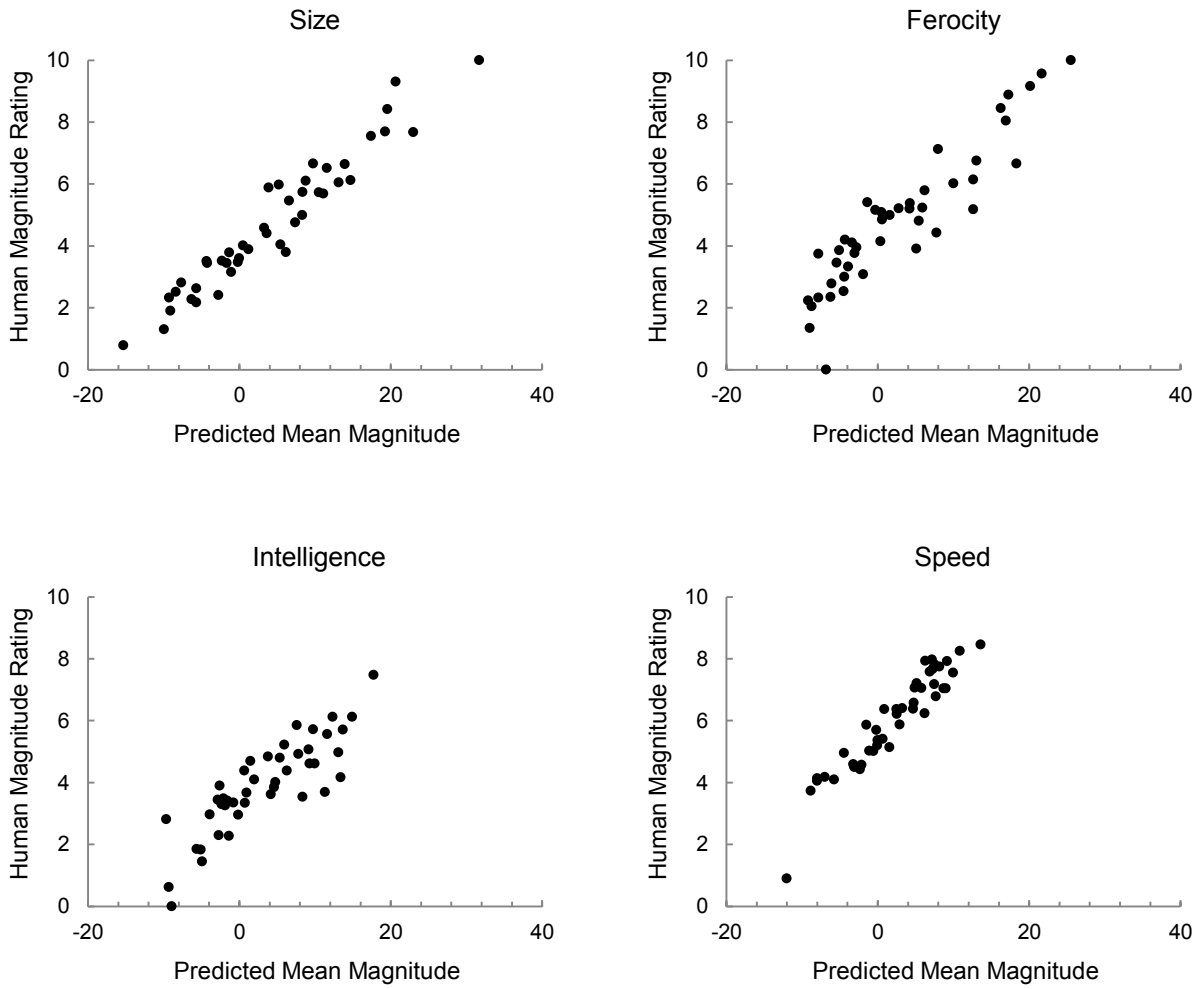


Figure 2.5. Scatter plots of human magnitude ratings (based on data from Holyoak & Mah, 1981) versus mean magnitudes derived from BARTlet using Leuven vectors for animals on four dimensions.

Symbolic Distance Effect

To evaluate whether BARTlet exhibits the ubiquitous symbolic distance effect obtained for comparative judgments by humans, we formed all possible pairs of the 44 animals previously identified, which served as testing items for each of the unmarked comparative relations corresponding to the four rated dimensions in Holyoak and Mah's (1981) norms: *larger*, *fiercer*, *smarter*, and *faster*. To ensure that the differences in magnitudes between animals in a pair were

likely to be distinguishable by humans, we excluded pairs that differed by less than .5 on the normed ratings for the relevant dimension. The resulting pairs of animals were grouped into four distance bins, such that animals very close on the relevant dimension fell into bin 1 and animals maximally far apart on that dimension fell into bin 4. Figure 2.6 plots the mean d_a value for each distance bin, averaged across the four unmarked comparative relations. Results for the four marked relations are similar. Consistent with a symbolic distance effect, BARTlet's predicted discriminability increases with the distance between the pair of animals.

Semantic Congruity Effect

To test BARTlet's ability to predict the congruity effect, for each of the four dimensions we selected five animal pairs that were either both at the high end (e.g., *whale-elephant* for size)

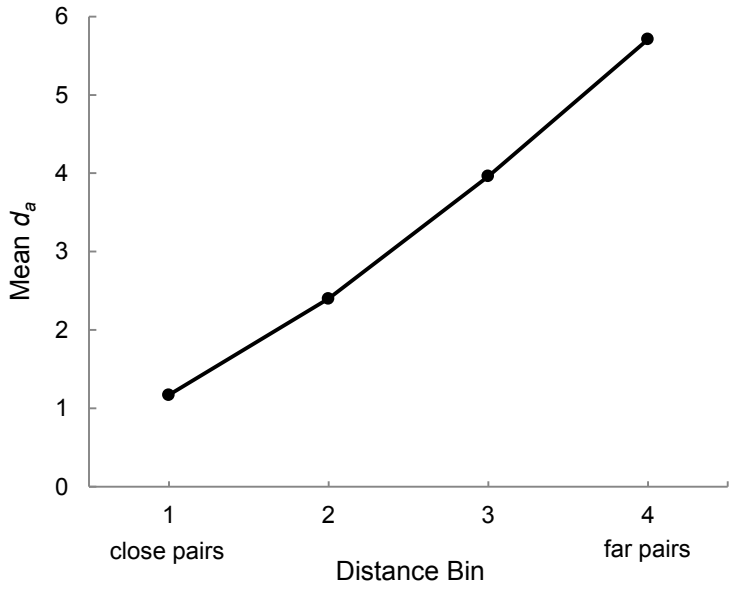


Figure 2.6. BARTlet's predicted discriminability value, d_a, for comparative judgments using Leuven vectors as a function of the subjective distance between pairs of animals on the relevant dimension. Distance bins are based on Holyoak and Mah's (1981) norms, in which values range from 0-10: bin 1 (distances between 0.5 and 1.5), bin 2 (distances 1.5-3), bin 3 (distances 3-5.5), and bin 4 (distances 5.5-10). Results are collapsed over the four unmarked comparative relations.

or both at the low end (e.g., *goldfish-fly*). We selected pairs that were at least minimally discriminable based on the learned weight distributions. All these pairs were relatively close in magnitude, as the congruity effect is typically maximized when both pairs are near to an extreme and hence close in magnitude. A congruity effect was observed for all four dimensions, as indicated by the interaction apparent in each panel (see Figure 2.7). In each case the interaction shows an asymmetry, with the advantage of the unmarked congruent form of the question (e.g.,

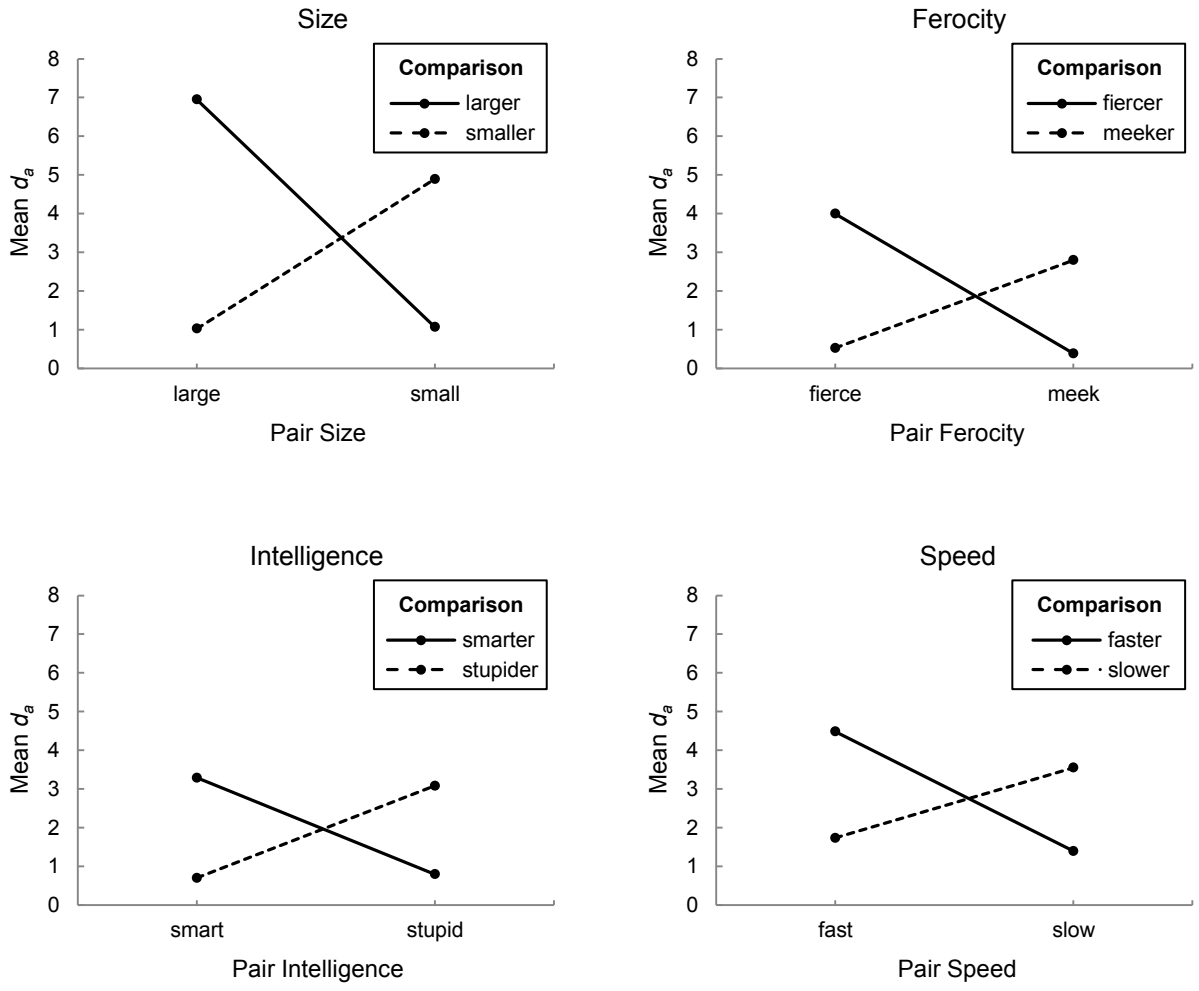


Figure 2.7. Predicted semantic congruity effect for magnitude comparisons with polar adjectives using Leuven vectors, based on BARTlet's predicted discriminability value, d_a, for unmarked and marked comparatives for four dimensions.

"choose larger" for large animals) being slightly greater than the corresponding advantage of the marked congruent form (e.g., "choose smaller" for small animals). In other words, the congruity effect was modulated by a markedness effect, as is commonly observed in behavioral studies (e.g., Holyoak & Mah, 1981).

Influence of Stimulus Range on Congruity Effect

An important additional finding concerning the congruity effect is that it is influenced by the range of magnitudes represented in the stimulus set (e.g., Čech & Shoben, 1985, for humans; Jones et al., 2010, for monkeys). Since BARTlet sets its reference points dynamically based on the magnitude range relevant to the current context, it naturally predicts how the congruity effect will vary with the context. To test this aspect of the model, we created four sets of stimuli based on the size dimension, ordered in size from Set 1 (pairs of largest animals) to Set 4 (pairs of smallest animals). Sets 1 and 4 were the same pairs used to test the basic congruity effect (see Figure 2.7). Sets 2 and 3 were intermediate in size (e.g., Set 2 included *alligator-pig*; Set 3 included *cat-sparrow*). The size distance between the two animals in each pair was closely matched across all four sets. In two different tests, BARTlet made "choose larger" and "choose smaller" judgments using either the full range of magnitudes (i.e., Sets 1-4), or a restricted range (i.e., Sets 2-3 only). As shown in Figure 2.8, both tests yielded congruity effects; however, the magnitude of the congruity effect for the critical Sets 2-3 based on middle-sized animals was substantially larger when these intermediate sets were tested alone (restricted range; 2.03 in d_a units) than when they were intermixed with the pairs of very large or very small animals (full range; 1.11 in d_a units). BARTlet thus provides an account of how context can influence comparative judgments by dynamically altering reference points.

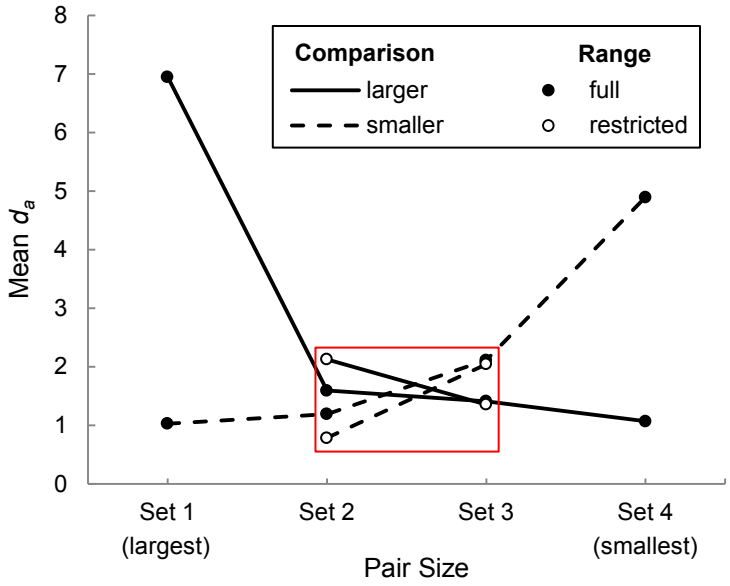


Figure 2.8. Predicted semantic congruity effect (using Leuven vectors) for stimuli from the full range of animal sizes (Sets 1-4) and a restricted intermediate range (Sets 2-3 only), based on BARTlet's predicted discriminability value, d_a.

Simulations of Symbolic Magnitude Judgments Using Topics Vectors

To derive topics vectors, we obtained a preprocessed version of the English Wikipedia corpus in which entries shorter than 512 words were removed, as were words that are not in a standard English dictionary or that are on a list of "stop words" (high-frequency function words that have low semantic content, such as *the*, *and*, etc.), resulting in a total of 174,792 entries and 116,128 unique words. We ran the topic model (Griffiths et al., 2007) on this corpus to obtain 300 topics. The algorithm was used to generate three Markov chains, taking the first sample after 1000 iterations and then sampling once every 100 iterations, for a total of eight samples from each chain or 24 samples overall. Each sample yielded a matrix in which the (*i*, *j*)th entry is the number of times that word *i* has been assigned to topic *j*. From this matrix, we derived a vector for each word based on the conditional probability of each topic given that word (i.e., each

resulting word vector is based on the relative frequencies of the different contexts within which the word could occur), using the same procedure employed by Lu et al. (2012) for outputs of the topic model ran on a different corpus.

Samples from a single Markov chain were very similar, in that the same 300 topics seemed to be found in each (based on examining the most probable words for each topic), but different chains produced different sets of topics. To create a single unified set of topics vectors for all words, we first averaged the word vectors based on samples from the same chain to produce a single set of word vectors for each chain. We then unified the three different chains (averaged across eight samples each) through the following procedure: First, for each of the averaged chains, we chose the 30 features (topics) that had the highest sums across the vectors of the 77 animal words in Holyoak and Mah's (1981) norms (i.e., the 30 most prevalent topics for these animal concepts). Using the resulting animal vectors (reduced to 30 features for each chain), we then ran the full BART model to learn the relations *larger*, *fiercer*, *smarter*, and *faster*. We examined BART's generalization performance for these relations using the animal vectors from each chain (using the same tests as Lu et al., 2012). Starting with all 30 features from the chain that produced the best performance, we added features one at a time from the other two chains (each of which also had 30 features) in order of BART's performance on the chains. To minimize redundancy, a feature was added only if its correlations across the 77 animals with the features chosen so far were all less than .80. This process resulted in a total of 52 selected features. All simulations reported below were run using these topics vectors of length 52.

Based on the topics vectors, the same general procedure was used to learn one-place predicates with BARTlet as was used for Leuven vectors (i.e., initial weights acquired using

RankSVM provided empirical priors for the learning of one-place predicates). The weights obtained by RankSVM were scaled by a factor of 10 rather than by a factor of 5 (to better match the scale of weights learned from topics vectors). The top and bottom 20 animals on each dimension (used as training data for BARTlet after the RankSVM stage) were drawn from the 77 animals in Holyoak and Mah's norms, rather than the 129 animals in the Leuven dataset. The same method was used as before for calculating magnitude means and variances for each animal on each dimension.

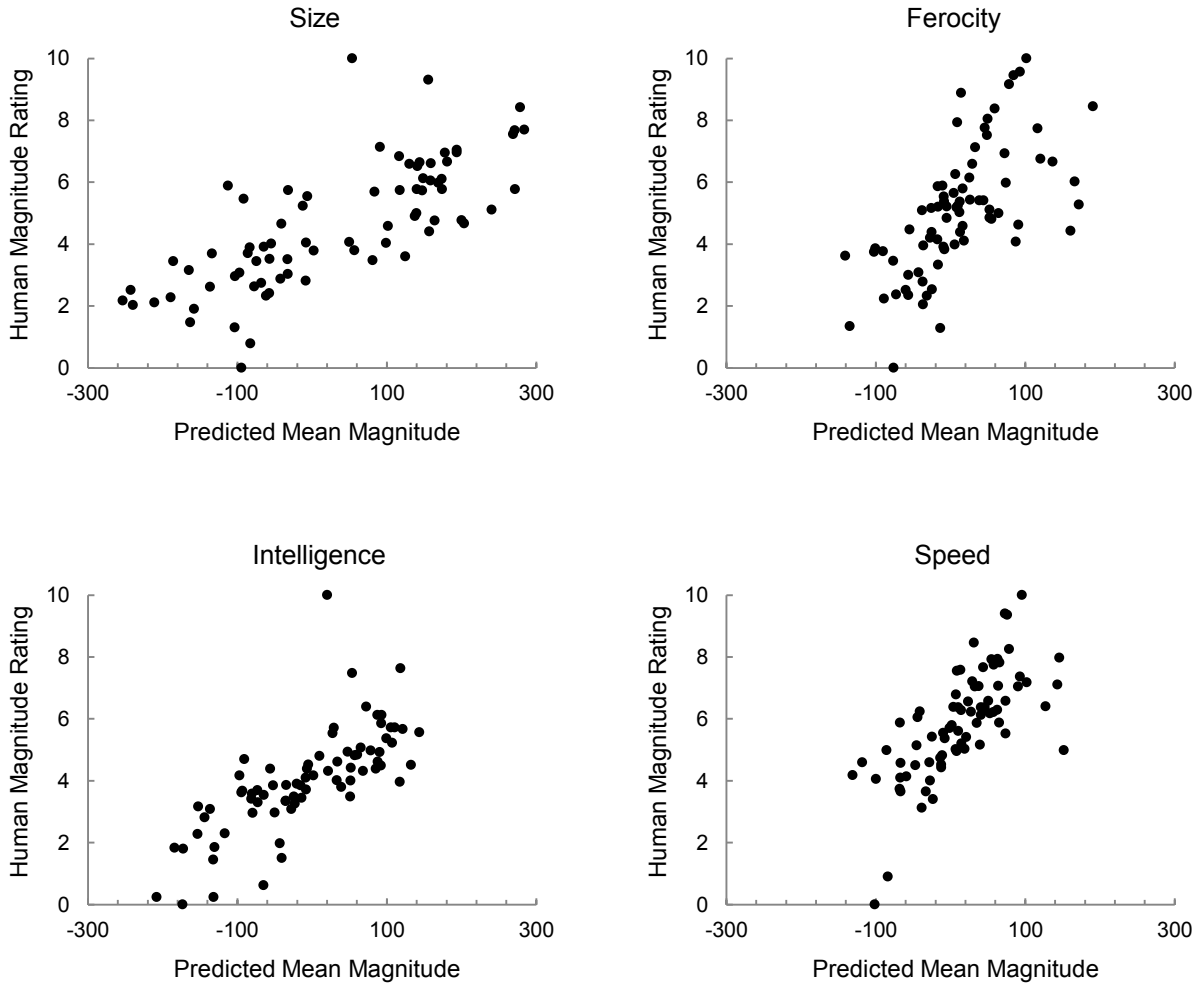


Figure 2.9. Scatter plots of human magnitude ratings (based on data from Holyoak & Mah, 1981) versus mean magnitudes derived from BARTlet using topics vectors for animals on four dimensions.

Predicting Human Magnitude Ratings

As we had done for the Leuven vectors, we performed correlational analyses to predict the human ratings (from Holyoak & Mah, 1981) using magnitudes extracted from the one-place predicates learned by applying BARTlet to topics vectors (except across a total of 77 animals, rather than the 44 available with Leuven vectors). Scatter plots are shown in Figure 2.9. Spearman rank-order correlations were lower than for the Leuven vectors, but all were reliable, ranging from .73 to .82 across the four dimensions.

Symbolic Distance Effect

As shown in Figure 2.10, the topics vectors yielded a robust distance effect (calculated in

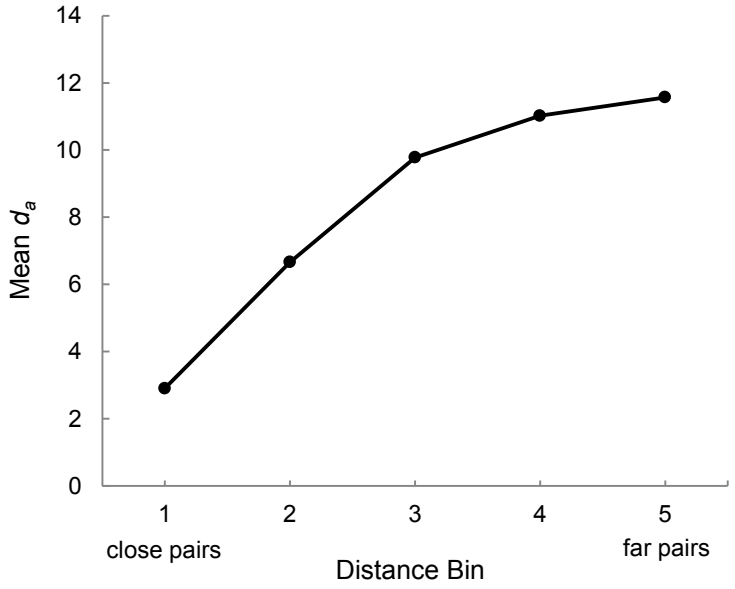


Figure 2.10. BARTlet's predicted discriminability value, d_a, for comparative judgments using topics vectors as a function of the subjective distance between pairs of animals on the relevant dimension. Distance bins are based on Holyoak and Mah's (1981) norms, in which values range from 0-10: bin 1 (distances between 0.5 and 2), bin 2 (distances 2-4), bin 3 (distances 4-6), bin 4 (distances 6-8), and bin 5 (distances 8-10). Results are collapsed over the four unmarked comparative relations.

the same way as for the Leuven vectors, except using an additional distance bin made possible because a larger set of animals was available).

Semantic Congruity Effect

As done previously for Leuven vectors, we selected sets of five pairs of animals consisting of animals near the high or else low end of each of the four continua. Each pair was at least

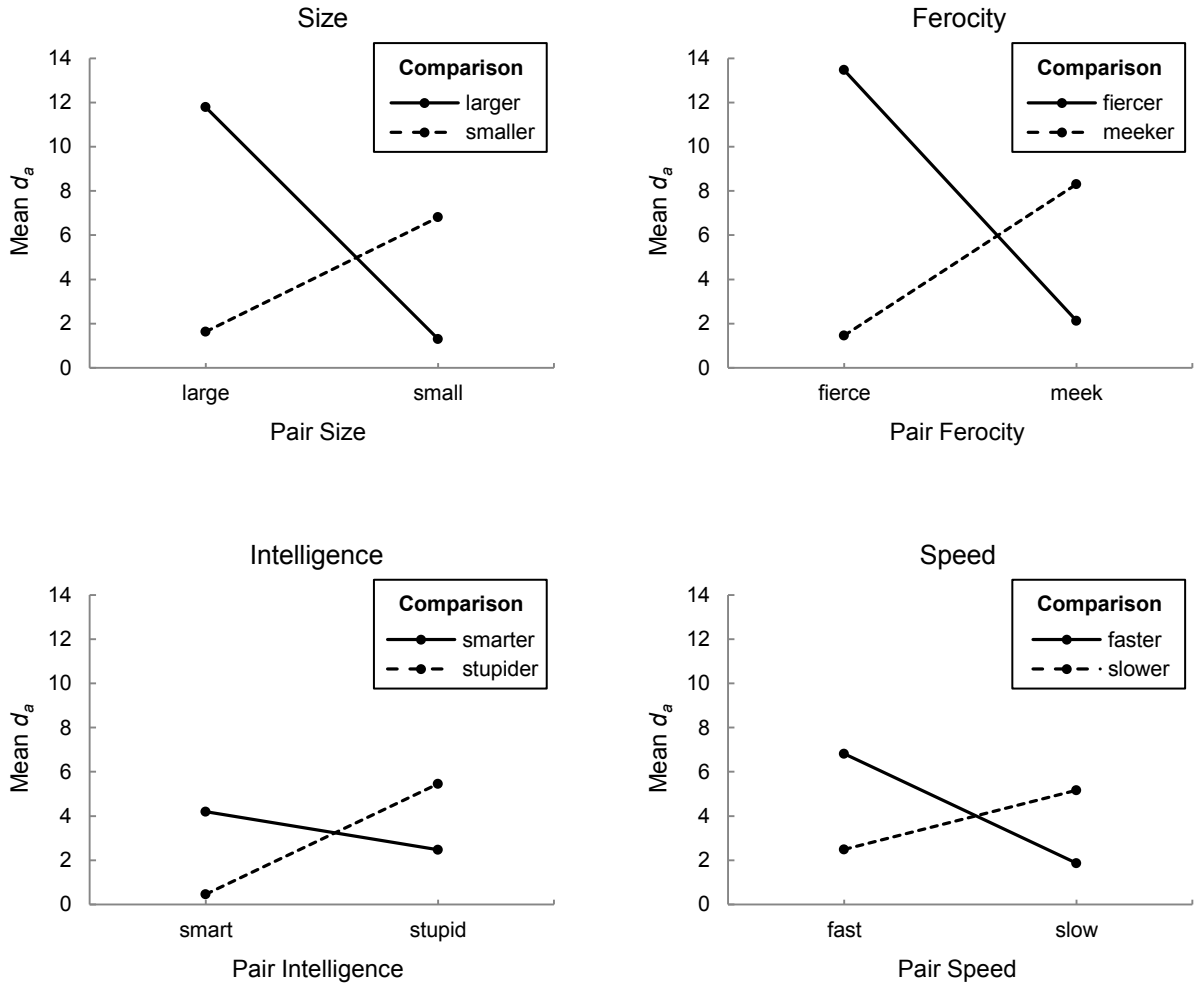


Figure 2.11. Predicted semantic congruity effect for magnitude comparisons with polar adjectives using topics vectors, based on BARTlet's predicted discriminability value, d_a, for unmarked and marked comparatives for four dimensions.

minimally discriminable but relatively close in magnitude (as the congruity effect is maximized for comparison of items with similar magnitudes).

Figure 2.11 shows the congruity effects obtained for each of the four dimensions. A robust congruity effect was obtained for each. A markedness effect (overall advantage for the unmarked form of the comparative) was obtained for all of the dimensions (though more pronounced for size and ferocity than for the other two). Because the topics vectors yielded cruder magnitude codes than did the Leuven vectors, we did not attempt to model the effect of range (as it was too difficult to generate discriminable pairs at more than two levels of overall magnitude).

General Discussion

Relational Comparisons without Explicit Relations

In the present paper we have presented a model, BARTlet, that provides a unified account of how subjective magnitudes on different dimensions can be learned from more elementary features, represented and modulated in working memory, and used to assess the discriminability of objects. Previous models of symbolic magnitude comparisons have tacitly assumed that magnitude values on the relevant dimensions are prestored in long-term memory as features of objects. We argue that this assumption is unrealistic, even for a quasi-perceptual dimension such as size, and especially for the many complex social and interpersonal dimensions on which people can make comparisons (e.g., intelligence, religiosity). By building on BART, a Bayesian model of how comparative concepts can be learned from examples by statistical inference (Lu et al., 2012), we were able to integrate an account of how magnitudes are compared with an account of how magnitudes can be created in working memory based on prior learning about comparative concepts. The generality of the approach was demonstrated by applying the model

to two sets of inputs (Leuven vectors and topics vectors), each of which was generated autonomously. BARTlet serves as an existence proof that symbolic comparisons can be modeled using high-dimensional distributed representations of elementary features, without assuming pre-existing dimensions, and without hand-coding inputs.

The operation of BARTlet, in comparison to its "smarter" precursor, BART, provides an instructive computational example of how a relational task (comparative judgment of magnitudes) can be performed without explicit relational representations. BART forms explicit representations of first-order relations such as *larger* (defined by weight distributions over pairs of objects assigned to distinct roles). In contrast, BARTlet operates only on weight distributions for one-place predicates (e.g., *large*), bootstrapping from priors on mean weights derived from pre-categorical comparisons (a partial ordering of pairs from which mean weights are learned by RankSVM, a model based on statistical regression). Magnitudes of individual objects are derived directly from the learned weight distributions for one-place predicates. BARTlet then proceeds to use an implicit comparison operation, which can be characterized in terms of signal detection theory, to assess which of two objects is the larger. No explicit *larger* relation is needed for BARTlet to choose the larger of two objects. BARTlet is thus an existence proof that the ability to make comparative judgments does not require explicit relational representations, consistent with evidence that rudimentary types of symbolic magnitude comparisons are within the capabilities of non-human primates (Cantlon et al., 2009).

Whereas BART is a computational-level model (Marr, 1982) of how comparatives can be learned, BARTlet adds explicit algorithmic assumptions concerning the representation and processing of magnitudes, based on consideration of limited computational resources in working memory. These core assumptions are firmly rooted in long-standing theories concerning

attentional influences on magnitude representation. Human (and non-human) observers have limited capacity in working memory to maintain veridical estimates of magnitudes, which therefore vary in their precision (Miller, 1956). To partially compensate, observers focus attention on a favored region, or magnitude band, along the relevant continuum (Luce et al., 1976; Nosofsky, 1983). When making comparisons based on relative concepts, such as "choose larger" or "choose smaller", attention is guided by a reference point located at or near the end of the continuum cued by the form of the question (Marks, 1972; Jamieson & Petrusic, 1975; Holyoak, 1978). More specifically, selective attention causes the precision of magnitudes in working memory to be greatest (i.e., associated with low variance) for values close to the reference point, decreasing with distance from the reference value (Marks, 1972). The decrease in precision with distance from the reference point tends to be asymmetrical, with a steeper function for the "marked" form of the question (e.g., "choose fiercer" as opposed to "choose meeker").

Armed with these algorithmic assumptions, together with the tools of signal detection theory, we showed by a series of simulations that BARTlet can predict (1) human ratings of subjective magnitudes for animals along four different dimensions, (2) the symbolic distance effect, (3) the semantic congruity effect, (4) the modulation of the congruity effect by the polarity of the comparative (i.e., markedness), and (5) the context sensitivity of the congruity effect (i.e., the influence of the magnitude range of the presented stimuli). Furthermore, BARTlet accounts for all of these phenomena based on magnitude distributions that emerge from prior statistical learning of weight distributions over a high-dimensional feature space. No previous theory of magnitude comparisons has provided a comparable integration with the acquisition of comparative concepts.

BARTlet's predictions for magnitude comparisons are qualitative, based on a simple discriminability measure, d_a, derived from signal detection theory. However, this measure has a natural link to established theories of two-choice decision making under varying degrees of speed pressure, notably the decision diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2010). The diffusion model, which describes the continuous accumulation of decision-relevant information over time, has a plausible neural realization (e.g., Ratcliff, Cherian, & Segraves, 2003; Wong & Wang, 2006). The diffusion model could in principle provide a more detailed account of the mechanisms by which decreases in discriminability will make a comparative judgment more difficult, yielding slower and/or less accurate comparative judgments.

Reference Points in Magnitude Comparisons

BARTlet provides a computational realization of a qualitative hypothesis proposed four decades ago by Marks (1972): Reference points cued by the form of comparative questions systematically modulate the precision of magnitudes represented in working memory, yielding the semantic congruity effect. The reference-point hypothesis implies that the congruity effect results from differences in the discriminability of magnitudes represented in working memory, rather than a bias in encoding (e.g., Marschark & Paivio, 1979) or a linguistic influence (Banks et al., 1975). BARTlet provides a well-specified mechanism by which reference points can alter discriminability in direct judgments of discriminability (Holyoak & Mah, 1982) as well as speeded tasks. The modulation of precision will maximally impact discriminability between objects with relatively similar magnitudes, in accord with the general finding that congruity effects are larger when the objects being compared are closer in magnitude (Petrusic, 1992). The BARTlet model could easily be extended to account for the impact of explicit reference points

(e.g., in a task requiring selection of which of two digits is closer in magnitude to 5; Holyoak, 1978), which can shift the favored attention band to an intermediate region on a continuum.

The semantic congruity effect is typically modulated by a markedness effect, i.e., the advantage of the "greater" comparative for pairs of items at the high end of a continuum often exceeds the reverse advantage of the "lesser" comparative for pairs at the low end of the continuum. The overall advantage of the unmarked form (e.g., "choose fiercer") over the marked form (e.g., "choose meeker") has generally been interpreted as a linguistic effect (Clark, 1969); however, the fact that the form of the comparative question has a similar impact on the performance of monkeys (Cantlon & Brannon, 2005) is problematic for a purely linguistic account.

BARTlet generates magnitude values (*M*) based on unmarked one-place predicates (e.g., *large*), and hence *M* values are positive and monotonic relative to the unmarked form (e.g., large animals are associated with high size values, and small animals with low size values, rather than the reverse). We assume that because the unmarked form of the question requires reversing the natural scale (e.g., "smaller" focuses attention on low magnitudes), precision diminishes more quickly with distance from the reference point in the case of the marked comparative. Our approach thus provides a mechanism by which polarity could impact magnitude judgments made by non-linguistic animals. This interpretation supports the hypothesis that the linguistic differences associated with markedness in human languages can be traced to more fundamental representational differences in magnitude continua.

The strong evidence that reference points influence discriminability implies that the semantic congruity effect is properly viewed as an example of the broader class of framing effects that impact decision making (Tversky & Kahneman, 1981). Indeed, semantic congruity

effects have been observed not only in magnitude comparisons involving objects, but also in judgments of preferential choice. For example, Birnbaum and Jou (1990) found that judging which individual is "liked more" for generally likeable individuals took less time than judging between unlikeable individuals, whereas judging which individual is "liked less" for likeable individuals took more time than judging between unlikeable individuals (also Nagpal & Krishnamurthy, 2008). The mechanisms instantiated in the BARTlet model may well prove applicable to decision making in areas such as consumer choice and social judgment.

The general notion of reference points has also been introduced in linguistic models of the interpretation of scalar adjectives (Tribushinina, 2009), which are interpreted in a context-sensitive manner. Scalar adjectives such as *large*, *warm*, and *average* refer to positions along a continuous dimension of magnitude; they are interpreted not as absolute values, but rather in relation to the noun category being modified (Partee, 1995). Thus an eagle is a large bird, but not an especially large animal; a tall boy is tall for a boy, but not for a tree. The interpretation of scalar adjectives requires scaling a subjective magnitude, or a probability of category membership derived from a magnitude, based on comparison to a norm or range derived from knowledge about the noun concept (see Barner & Snedecker, 2008).

The Power and Limits of Magnitude Representations

The parallels between the patterns of performance observed in monkeys and humans when performing magnitude comparisons suggest that this type of comparative judgment is based on evolutionarily primitive mechanisms. More broadly, neural and other evidence indicates that primates have evolved a specialized system for processing approximate magnitude, in which the intraparietal sulcus plays a key role (e.g., Cantlon et al., 2006; Dehaene & Changeux, 1993; Fias et al., 2007; Piazza et al., 2004, 2006, 2007; Pinel et al., 2004).

One reason for the apparent ubiquity of magnitude representations is that they can serve to answer multiple types of questions, each of which also provides learning opportunities. BARTlet learns magnitudes by integrating training with partial orderings (e.g., elephant is ordered before dog on the size dimension), the type of information provided to RankSVM, with training based on categorical inputs (e.g., elephant is large). Its acquired magnitude information might then be used to answer other interrelated types of questions (e.g., How large is a dog? Is a dog large? Is it larger than a cat? Is it smaller than a bear? Which is closer in size to a bear, a dog or a fox?). Feedback on the answers to any of such questions could be used to refine magnitude representations for a wide range of individual animals (not just those directly queried), thereby improving the model's ability to answer any question that depends on these magnitudes.

The fact that magnitudes are involved in answering many different questions and can be learned by multiple routes explains why evolution has apparently placed a premium on the creation of specialized neural hardware for manipulating such representations. Given the ubiquitous importance of comparative judgments in decision making, a system for discovering and manipulating magnitudes will be broadly advantageous. Nonetheless, unidimensional magnitude representations have their limitations. One limitation is that the neural system for approximate magnitude acts as a bottleneck. Precisely because any dimension can be coded in terms of a single internal number line, it is very difficult to code distinct orderings on separate dimensions for a single set of objects (Banks & White, 1982), a bottleneck that contributes to the "halo effect" (De Soto, 1961). In addition, the validity of a one-dimensional magnitude representation is inherently limited, as is apparent whenever we try to reduce a complex multidimensional situation to a single number that serves as a "score" (e.g., GPA as a summary

of a student's academic ability, *h*-index as a summary of a scientist's scholarly impact, dollar earnings as a summary of a year of one's life).

Limitations and Possible Extensions of the BARTlet Model

Although the BARTlet model captures several basic phenomena related to symbolic magnitude comparisons, it currently has a number of empirical limitations. We have focused on the distance, semantic congruity, and markedness effects, which are arguably the phenomena most universally observed in studies of symbolic magnitude comparisons. An additional phenomenon, typically observed for comparisons involving a closed-set series for which only ordinal information is available (e.g., an arbitrary ordering of elements for which magnitude information is not provided) is a bow-shaped serial position curve: accuracy and decision time indicate greater difficulty for pairs drawn from near the center of the list than for pairs closer to the ends. A bowed serial position curve is not observed for magnitude continua such as those on which we have focused in the present paper, but it is found for arbitrary orderings, both for humans (e.g., Potts, 1974; Trabasso & Riley, 1975; Woocher et al., 1978) and many animal species, including squirrel monkeys (McGonigle & Chalmers, 1977), rats (Davis (1992) and pigeons (von Fersen et al., 1991; for a review see Merritt & Terrace, 2011).

Although BARTlet does not currently model learning and performance with arbitrary series, it is in fact well-suited to be extended in this direction. One leading hypothesis is that bow-shaped serial position curves reflect *positional discriminability* (Holyoak & Patterson, 1981; Merritt & Terrace, 2011). The basic idea is that if individual items lack featural information that conveys magnitude, they are instead coded by their position relative to the beginning and end terms, which are learned first and serve as anchors. In accord with the representations used by BARTlet, these positional codes will be imprecise, forming a normal

distribution centered on an item's veridical position. Positional codes can be compared in the same way as codes for "true" magnitudes. The codes for central items will necessarily have greater overlap, and may well have higher variances than end items (Bower, 1971; Murdock, 1960; Trabasso & Riley, 1975). Thus, a natural extension of BARTlet would use the same basic type of representation—continuous-valued codes, normally distributed and varying in precision—to explain comparisons based on arbitrary ordered sets of elements. Such an extension would generate responses that exhibit distance effects, congruity effects, bow-shaped serial position effects, and transitivity of choice, as is empirically observed.

There has been some debate concerning whether, or in what way, magnitude codes are spatial in nature. The apparent empirical differences between learning and performance with dimensional magnitude codes versus positional codes suggest that although both are essentially analog (i.e., continuous-valued), magnitude codes are not necessarily spatial (nor are they inherently visual; Holyoak, 1977). In contrast, positional codes seem to be more spatial in nature, akin to an internal array (Holyoak & Patterson, 1981; Woocher et al., 1977). Nonetheless, similar brain areas are involved in comparisons of both types (see Cantlon et al., 2009).

A behavioral phenomenon often cited in support of a specifically spatial interpretation of magnitude codes, especially for number, is the SNARC effect ("Spatial Numerical Association of Response Codes"; Dehaene, 1992; Dehaene, Bossini & Giraux, 1993). When evaluating a number (e.g., deciding whether it is odd or even), people typically respond to small numbers more quickly when the response key is to the left, and to large numbers more quickly when the response key is to the right. The SNARC effect thus suggests that number magnitude has a natural mapping onto the left-right axis of space (small numbers associated with the left).

The original tasks that exhibited a SNARC effect only used numbers, and did not involve

magnitude comparisons. More recently, SNARC-like effects have also been observed in comparison judgment tasks, but the empirical picture is quite complex. Shaki, Petrusic and Leth-Steensen (2012) reported that (1) a typical SNARC effect is found for digit comparisons with both "larger" and "smaller" instructions, (2) a typical SNARC effect is found for animal size comparisons with a "choose smaller" instruction, but a *reverse* SNARC effect is found for a "choose larger" instruction; (3) a short, newly-learned height ordering behaves much like size comparisons; (4) the above pattern for English speakers (1-3) is reversed for Israeli-Palestinians who habitually read right-to-left. A rough characterization of Shaki et al.'s (2012) findings is that although by default small numerical magnitudes are associated with the left, for non-numerical continua this bias is overridden by a preference to place the *reference point* on the left (or more generally, on the side from which orderings usually begin—hence the reversal due to cultural experience).

BARTlet does not model output processes, so it does not provide any obvious insight into the SNARC effect. However, as Shaki et al. (2012) noted, "….the mere fact that spatial information is being activated in association with the activation of magnitude information does not, in and of itself, conclusively imply that such spatial information is then actually being used by the comparison process itself" (p. 525). Whatever the SNARC effect may imply about spatial processing, there is reason to doubt it has a deep connection to the comparison process that is the focus of BARTlet.

A further limitation of BARTlet stems from the fact that it can only compute comparative relations, and does not store or retrieve facts. People can certainly learn specific relational facts that arise repeatedly, or are tied to the intrinsic meanings of words (e.g., we commonly see dogs that are larger than cats; we know mountains are larger than hills because of how these terms are

defined), and comparisons of this sort are made relatively quickly (Holyoak, Dumais & Moyer, 1979). BARTlet does not account for the role of fact retrieval in magnitude comparison. It should be emphasized, however, that fact retrieval seems to play a modest secondary role. The initial demonstration of distance effects involving the digits 1-9 (Moyer & Landauer, 1967) was especially compelling because although adults surely know the fact that 3 is larger than 2 very well, they nonetheless find it easier to decide that 8 is larger than 2. In general, the ease of mental comparison seems to trump that of fact retrieval.

Relation to Previous Models of Learning Dimensional Representations

As a learning model, BARTlet is based on the BART model, which Lu et al. (2012, pp. 640-642) discussed in relation to other models of relation learning. Here we consider three models (roughly ordered from least to most explicit in their relational representations) that have addressed the acquisition of continuous dimensions and/or linear orderings.

Smith, Gasser, and Sandhofer (1999) developed a multi-layer neural network model that learns dimensional adjectives by back-propagation. This model focuses on the interactive constraints provided by sensory, perceptual and linguistic information. Smith et al. argued that dimensional attributes, such as *large* or *red*, need not correspond to invariant features at the sensory level, but rather can be learned as distributed representations over more elementary features. Learning in their model involves updating weights on features; the magnitudes of weights are interpreted as indicators of learned selective attention. These assumptions are shared by BARTlet. Though the Smith et al. model has not been directly applied to the task of magnitude comparisons, it might well be extended in that direction. As a standard neural network, the model learns weights as point estimates, and hence does not capture differences in precision. But at a global level, the Smith et al. model is similar in spirit to BARTlet, taking a

basically bottom-up approach to the acquisition of dimensional concepts, and operating without explicit representations of comparative relations.

DORA (Discovery of Relations by Analogy) is a symbolic-connectionist model that learns both one-place predicates (e.g., *large*) and two-place relations (e.g., *larger*), focusing on comparatives (Doumas, Hummel & Sandhofer, 2007). Like BARTlet (and BART), it emphasizes bottom-up learning from objects coded as feature vectors (though it has not yet been tested on high-dimensional inputs of the sort used in the present paper). DORA includes a comparator operator that is well-suited for performing magnitude comparisons. Because DORA's predicates are initially most similar to the specific cases from which they were learned, the model predicts a congruity effect early in learning (e.g., for children, the representation of *large* will be more similar to large than small objects, and vice versa for *small*, leading to a congruity effect). As the model continues to refine its predicates using a feature-intersection mechanism, its representations of dimensional adjectives will tend to become more "magnitude neutral." It is therefore less clear whether the model could account for congruity effects observed in studies with adults. However, it is possible that DORA could be extended to include assumptions about the role of reference points.

Finally, an extremely general framework for learning relational structures has been proposed by Kemp and Tenenbaum (2008, 2009). By coupling a generative grammar for structural forms with a hierarchical Bayesian inference engine, their integrated model can generate many different structures to explain data patterns, including trees, multidimensional spaces, grids, rings, chains and (most importantly in the present context) linear orders. As Kemp and Tenenbaum acknowledge, "…we offer a modeling framework rather than a single model of induction. Our framework can be used to construct many specific models… (2009, p. 22). Any

specific model within the framework involves a combination of assumptions about the available forms and about the processes that operate on forms to make inductive inferences. Given its flexibility, a model could presumably be created within the framework that would closely emulate BARTlet (or BART, or other alternative models).

The power of the framework is also its Achilles' heel as a psychological theory. Without clear constraints, it is hard to derive testable predictions. However, we can evaluate the specific model of linear orderings that Kemp and Tenenbaum (2008) provided. This model has two basic problems as a psychological proposal. First, given that the model can learn many different structural forms, it does not account for the empirical fact that linear orderings are special in the realm of animal cognition. As we have seen, a great variety of species can make comparative judgments based on linear orderings. By contrast, animals have considerably more difficulty learning circular orderings, or rings (von Fersen et al., 1991). The special status of linear orderings is a natural consequence for BARTlet and other models that base comparisons on magnitudes, or some similar unidimensional quantity. But within the Kemp and Tenenbaum framework, there is no apparent reason why rings should be any more difficult to learn than linear orders (though a prior could be arbitrarily imposed to favor either one).

A second basic problem is that the Kemp and Tenenbaum model of linear orders does not account for the ubiquitous distance effect. Their model creates explicit asymmetric relations between all possible pairs in an ordering (e.g., if elements A through E form a linear order, the learned structure would not only include links A > B, B > C, etc., but also B > D, B > E, etc.). The proposed inference processes (Kemp & Tenenbaum, 2009) imply that the strength of an inference concerning any two elements in a structure will be monotonic (in one direction or the other) with the length of the chain of links connecting the elements. But in the linear order

model, the chain length is constant (one) for all pairs; hence the model predicts that (for example) a reasoner could evaluate B > C just as easily as B > D.

An ordering structure of this form was used to account for patterns of dominance behavior among members of a monkey troop (observed by Range & Noë, 2002; see Kemp & Tenenbaum, 2008, Figure 4a, p. 10689). In fact, as we will discuss below, it is possible to explain monkeys' choices regarding whether or not to exhibit submissive behavior toward a conspecific without assuming that they form explicit comparative relations at all, far less a complete explicit representation of all pairwise relations. Thus, while the Kemp and Tenenbaum model of linear orders provides a useful tool for extracting the types of representations employed by (human) primatologists, it is problematic if interpreted as a psychological model of the mental representations that guide the choice behavior of primates.

Re-representation and the Emergence of Explicit Relations

A great virtue of computational models is that they can bring clarity to important conceptual distinctions that might otherwise be blurred, or dismissed as a matter of semantics. A longstanding question in comparative psychology has been whether or not non-human animals (especially primates) "think", "reason", "use logic", or "understand relations" in fundamentally the same way as humans do. Various relational tasks have figured prominently as sources of evidence, including comparative judgment and transitive choice. As noted earlier, many species, from pigeons to primates, exhibit transitivity of choice (see Merritt & Terrace, 2011). Some have viewed such performance as tantamount to Piagetian transitive inference (e.g., if a 5-year old child is told that object B is bigger than object C, and object A is bigger than object B, then the child will likely be able to infer that A is bigger than C, despite knowing nothing about the features of the objects).

But in fact, transitivity of choice and Piagetian transitive inference involve completely different task demands, with little in common other than their misleadingly similar names (Halford, 1984; Markovits & Dumas, 1992). Transitivity of choice can be accomplished by using perceptually-based training data (ordered pairs and/or individual objects) to learn approximate quantities associated with individual items (e.g., magnitude codes, positional codes, values, or associative strengths). Examples of associative and statistical models that can accomplish learning of this type include the Rescorla-Wagner model (Rescorla & Wagner, 1972), Value Transfer Theory (von Fersen et al., 1991), RankSVM (Parikh & Grauman, 2011), and BARTlet. Although these models differ in many important ways, all provide mechanisms for performing relational judgments without explicit relations.

Accordingly, demonstrating success in basic comparative judgments tasks, or in transitivity of choice paradigms, cannot in principle provide evidence for the use of explicit comparative relations. Morgan's Canon can prudently be applied: "In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale" (Morgan, 1894, p. 53). If we replace the quaint Victorian phrase "psychical faculty" with "relational complexity" or "representational rank" (Halford et al., 1998; Phillips, Halford, & Wilson, 1995), then Morgan's Canon continues to provide a valuable guide for comparative (and cognitive) psychology in the 21st century.

As Penn et al. (2008) argued based on a review of comparative studies, there is overwhelming evidence that many species of animals can make relational judgments based on perceptual information, yet no compelling evidence that any non-human primate is able to reason about relations. At the same time, it appears that the neural system supporting comparisons based

on approximate magnitude in non-human primates operates in humans as well (Dehaene & Changeux, 1993). Apparently, humans have not lost the simpler mechanisms available to other animals for comparing magnitudes, but rather have exploited these mechanisms as a foundation for symbolic mathematical thinking (Opfer & Siegler, 2012). More generally, humans appear to have surpassed the intellectual capacity of any other species on earth by acquiring neural machinery that enables the re-representation of lower-level information in terms of explicit relational concepts.

As a small computational example of such re-representation, BARTlet becomes the prequel to BART, which uses one-place predicates such as *large* to bootstrap acquisition of explicit two-place relations such as *larger*. A system that is restricted to magnitude representations (lacking the ability to form explicit relational representations) inevitably "hits the wall" when faced with more complex symbolic tasks. A monkey (and BARTlet) can learn to choose the larger or the smaller of two objects. But a human (and BART) can also acquire an explicit representation of the relations *larger* and *smaller*, and go on to reason about them (e.g., noticing that *larger* is related to *smaller* in much the same way as *fiercer* is related to *meeker*; Lu et al., 2012).

Similarly, associative and statistical mechanisms that can support transitivity of choice prove completely inadequate when confronted with a Piagetian transitive inference task. The latter task requires a "one shot" inference based on integration of two binary premises in working memory, without repeated acquisition trials, and without support from perceptual cues or magnitude codes. Reliable success is not achieved by any species except humans, and not until preschool age (Andrews & Halford, 1998; Halford, 1984; Halford, 1993). Piagetian transitive inference is heavily dependent on a mature and intact human frontal cortex (Waltz et al., 1999).

We have recently extended the BART model to enable it to use its learned representations to solve abstract transitive inference problems (Chen, Lu, & Holyoak, 2013). Perhaps surprisingly, explicit comparative relations are not required to make comparative judgments. However, they prove essential for any reasoner who aspires to think about what such judgments mean.

Footnotes

1. We assume for simplicity that reference points are established using the range of presented stimuli. Of course, the range of presented stimuli will typically become apparent to the observer over the course of exposure to a series of examples. Reference points are therefore likely to be updated dynamically, reflecting a compromise between prior expectations about stimulus range and the range actually observed in the context (Petrusic & Baranski, 1989).
2. The specific selection of training examples is not critical to the performance of the model. We aimed to limit the number of training examples so that the model was forced to generalize on test pairs. The emphasis on early learning of extreme "landmark" animals is consistent with the typical pattern observed in learning orderings (Potts, 1974; Ryalls & Smith, 2000).
3. The use of the prior provided by RankSVM increased the rank-order correlations between human magnitude ratings and magnitudes derived from the model by approximately .10 (relative to an uninformative prior) for the Leuven inputs and about .02 for the topics inputs.

References

Andrews, G., & Halford, G. S. (1998). Children's ability to make transitive inferences: The importance of premise integration and structural complexity. *Cognitive Development, 13*(4), 479-513.

Audley, R. J., & Wallis, C. P. (1964). Response instructions and the speed of relative judgments: I. Some experiments on brightness discrimination. *British Journal of Psychology, 55,* 59-73.

Banks, W. P., Clark, H. H., & Lucy, P. (1975). The locus of the semantic congruity effect in comparative judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *104*, 35-47.

Banks, W. P., & Flora, J. (1977). Semantic and perceptual processes in symbolic comparisons. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 278-290.

Banks, W. P., Fujii, M., & Kayra-Stuart, F. (1976). Semantic congruity effects in comparative judgments of magnitudes of digits. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 435-447.

Banks, W. P., & White, H. (1982). Single ordering as a process limitation. *Journal of Verbal learning and Verbal Behavior*, *21*, 39-54.

Banks, W. P., White, H., Sturgill, W., & Mermelstein, R. (1983). Semantic congruity and expectancy in symbolic judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 560-582.

Barner, D., & Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4-year-olds interpret *tall* and *short* based on the size distributions of novel referents. *Child Development*, *79*, 594-608.

Birnbaum, M. H., & Jou, J.W. (1990). A theory of comparative response times and "difference" judgments. *Cognitive Psychology*, *22*, 184-210.

Bower, G. H. (1971). Adaptation-level coding of stimuli and serial position effects. In M. H. Appley (Ed.), *Adaptation-level theory* (pp. 175-201). New York: Academic Press.

Cantlon, J., & Brannon, E. M. (2005). Semantic congruity affects numerical judgments similarly in monkeys and humans. *Proceedings of the National Academy of Sciences, USA*, *102*, 16507–16511.

Cantlon, J., Brannon, E. M., Carter, E., & Pelphrey, K. (2006). Functional imaging of numerical processing in adults and 4-yr-old children. *PLoS Biology, 4*, e125, 1-11.

Cantlon, J. F., Platt, M., & Brannon, E. M. (2009). Beyond the number domain. *Trends in Cognitive Sciences*, *13*, 83-91.

Čech, C. G., & Shoben, E. J. (1985). Context effects in symbolic magnitude comparisons. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 299-315.

Čech, C., Shoben, E., & Love, M. (1990). Multiple congruity effects in judgments ofmagnitude. *Journal of Experimental Psychology Learning, Memory, and Cognition*, *16*, 1142–1152.

Chen, D., Lu, H., & Holyoak, K. J. (2013). Generative inferences based on a discriminative Bayesian model of relation learning. In M. Knauf, M. Pauven, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review*, 1969, *76*, 387-404.

Clark, H. H., Carpenter, P. A., & Just, M. A. (1973). On the meeting of semantics and perception. In W. G. Chase (Ed.), *Visual information processing* (pp. 311- 381). New York: Academic Press.

Cromer, J. A. Roy, J. E., & Miller, E. K. (2010) Representation of multiple, independent categories in the primate prefrontal cortex. *Neuron*, *66*, 796-807

Davis, H. (1992). Transitive inference in rats (*Rattus norvegicus*). *Journal of Comparative Psychology*, *106*, 342-349.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel,W., Dry, M., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods, 40(4)*, 1030-1048.

Dehaene, S. (1992). The varieties of numerical abilities. *Cognition, 44,* 1-42.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122,* 371-396.

Dehaene, S., & Changeux, J.-P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience, 5*, 390-407.

DeSoto, C. B. (1961). The predilection for single orderings. *Journal of Abnormal and Social Psychology*, *62*, 16-23.

Diester, I., & Nieder, A. (2007). Semantic associations between signs and numerical categories in the prefrontal cortex. *PLoS Biology*, *5*, e294.

Diester, I., & Nieder, Q. (2010). Numerical values leave a semantic imprint on associated signs in monkeys. *Journal of Cognitive Neuroscience*, *22*, 174-183.

Dosher, B. A., & Lu, Z.-L. (2000). Noise exclusion in spatial attention. *Psychological Science*, *11*, 139–146.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*, 1-43.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, *291*, 312-316.

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*, 125-157.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211-244.

Halford, G. S. (1984). Can young children integrate premises in transitivity and serial order tasks? *Cognitive Psychology***,** 16, 65-93.

Halford, G. S. (1993). *Children's understanding: The development of mental models*. Hillsdale, N. J.: Erlbaum.

Halford, G. S., Wilson, W. H. & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental and cognitive psychology. *Behavioral Brain Sciences, 21*(6), 803-831.

Halford, G. S., Wilson, W. H. & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences, 14*(11), 497-505.

Hoedmaker, R., & Gordon, P. C. (2013). Embodied language comprehension: Encoding-based and goal-driven processes. *Journal of Experimental Psychology: General.* doi:10.1037/a0032348

Holyoak, K. J. (1977). The form of analog size information in memory. *Cognitive Psychology*, *9*, 31-51.

Holyoak, K. J. (1978). Comparative judgments with numerical reference points. *Cognitive Psychology*, *10*, 203-243.

Holyoak, K. J., Dumais, S. T., & Moyer, R. S. (1979). Semantic association effects in a mental comparison task. *Memory & Cognition*, *7*, 303-313.

Holyoak, K. J., & Mah, W. A. (1981). Semantic congruity in symbolic comparisons: Evidence against an expectancy hypothesis. *Memory & Cognition*, *9*, 197-204.

Holyoak, K. J., & Mah, W. A. (1982). Cognitive reference points in judgments of symbolic magnitude. *Cognitive Psychology*, *14*, 328-352.

Holyoak, K. J., & Patterson, K. K. (1981). A positional discriminability model of linear order judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *7,* 1283-1302.

Holyoak, K. J., & Walker, J. H. (1976). Subjective magnitude information in semantic orderings. *Journal of Verbal Learning and Verbal Behavior*, *15*, 287-299.

Howard, R. (1983). The semantic congruity effect: Some tests of the expectancy hypothesis. *Acta Psychologica, 53,* 205-216.

Jamieson, D. G., & Petrusic, W. (1975). Relational judgments with remembered stimuli. *Perception & Psychophysics, 18,* 373-378.

Jones, S. M., Cantlon, J. F., Merritt, D. J., & Brannon, E. M. (2010). Context affects the numerical semantic congruity effect in rhesus monkeys (*Macaca mulatta*). *Behavioral Processes, 83,* 191-196.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, USA, 105*(31), 10687-10692.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning.

Psychological Review, 116(1), 20-58.

Lawrence, D. H., & DeRivera, J. (1954). Evidence for relational transposition. *Journal of Comparative and Physiological Psychology, 47,* 465-471.

Link, S. W. (1990). Modeling imageless thought: The relative judgment theory of numerical comparisons. *Journal of Mathematical Psychology*, *34*, 2–41.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*, 617-648.

Luce, R. D., Green, D. M., & Weber, D. L. (1976). Attention bands in absolute identification. *Perception & Psychophysics, 20*, 49-54.

Markovits, H., & Dumas, C. (1992). Can pigeons really make transitive inferences? *Journal of Experimental Psychology: Animal Behavior Processes*, *18*, 311-312.

Marks, D. F. (1972). Relative judgment: A phenomenon and a theory. *Perception & Psychophysics, 11,* 156-160.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: W. H. Freeman.

Marschark, M., & Paivio, A. (1979). Semantic congruity and lexical marking in symbolic comparisons: An expectancy hypothesis. *Memory & Cognition*, *7*, 175-184.

McGonigle, B. O., & Chalmers, M. (1977). Are monkeys logical? *Nature*, *267*, 694-696.

Merritt, D. J., & Terrace, H. S. (2011). Mechanisms of inferential order judgments in humans (*Homo sapiens*) and rhesus monkeys (*Macaca mulatta*). *Journal of Comparative Psychology*, *125*, 227-238.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.

Morgan, C. L. (1894). *An introduction to comparative psychology.* London: Walter Scott.

Moyer, R. S. (1973). Comparing objects in memory: Evidence suggesting an internal psychophysics. *Perception & Psychophysics*, *13*, 180- 184.

Moyer, R. S., & Bayer, R. H. (1976). Mental comparison and the symbolic distance effect. *Cognitive Psychology,* 8, 228-246.

Moyer. R. S., & Dumais. S. T. (1978). Mental comparison. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 12, pp. 117-155). New York: Academic Press.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature*, *215*, 1519-1.520.

Murdock, B. B., Jr. (1960). The distinctiveness of stimuli. *Psychological Review*, *67*, 16-31.

Nagpal, A., & Krishnamurthy, P. (2008). Attribute conflict in consumer decision making: The role of task compatibility. *Journal of Consumer Research*, *34*, 696-705.

Nieder, A., & Miller, E. K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron*, *37*, 149–157.

Nosofsky, R. M. (1983). Shifts of attention in the identification and discrimination of intensity. *Perception & Psychophysics*, *33*, 103-112.

Opfer, J. E., & Siegler, R. S. (2012). Development of quantitative thinking. In K. K. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 585-605). New York: Oxford University Press.

Parikh, D., & Grauman, K. (2011). Relative attributes. In D. M. Metaxas, L. Quan, & L. J. Van (Eds.), *Proceedings of the IEEE International Conference on Computer Vision* (pp. 503-510). Barcelona, Spain: IEEE.

Partee, B. (1995). Lexical semantics and compositionality. In D. Osherson (General Ed.), & L. Gleitman & M. Liberman (Eds.), *Invitation to cognitive science. Part I: Language* (2nd ed., pp. 311-360). Cambridge, MA: MIT Press.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*, 109-178.

Petrusic, W. M. (1992). Semantic congruity effects and theories of the comparison process. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 962-986.

Petrusic, W. M., & Baranski, J. V. (1989). Semantic congruity effects in perceptual comparisons. *Perception & Psychophysics*, *45*, 439-452.

Phillips, S., Halford, G. S., & Wilson, W. H. (1995). The processing of associations versus the processing of relations and symbols: A systematic comparison. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 688-691). Mahwah, NJ: Erlbaum.

Piazza, M., Izard, V., Pinel, P., Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron, 44*, 547-555.

Piazza, M., Mechelli, A., Price, C. J., & Butterworth, B. (2006). Exact and approximate judgements of visual and auditory numerosity: An fMRI study. *Brain Research, 1106*, 177-188.

Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron, 53*, 293-305.

Pinel, P., Piazza, M., Bihan, D. L., & Dehaene, S. (2004). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron*, *41*, 1-20.

Potts, G. R. (1974). Storing and retrieving information about ordered relationships. *Journal of Experimental Psychology*, *103*, 431-439.

Rahnev, D., Maniscalco, B., Graves, T., Huang, E., de Lange, F. P., & Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience*, *14*, 1513-1515.

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

Range, F., & Noë, R. (2002). Familiarity and dominance relations among female sooty mangabeys in the Taï National Park. *American Journal of Primatology*, *56*, 137-153.

Ratcliff, R., Cherian, A., & Segraves, M. (2003). A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of simple two-choice decisions. *Journal of Neurophysiology*, *90*, 1392-1407.

Ratcliff, R., & McKoon, G. (2010). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873-922.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*, 261–300.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Reynolds, J. H., & Chelazzi, R. (2004). Attentional modulation of visual processing. *Annual*

Review of Neuroscience, *27*, 611-647.

Rylass, B. O., & Smith, L. B. (2000). Adults' acquisition of novel dimension words: Creating a semantic congruity effect. *Journal of General Psychology*, *127*, 279-326.

Ryalls, B. O., Winslow, E., & Smith, L. B. (1998). A semantic congruity effect in children's acquisition of *high* and *low*. *Journal* of *Memory and Language, 39*, 543-557.

Shaki, S., Petrusic, W. M., & Leth-Steensen, C. (2012). SNARC effects with numerical and non-numerical symbolic comparative judgments: Instructional and cultural dependencies. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 515-530.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition, 120*, 1-25.

Shaki, S., Leth-Steensen, C., & Petrusic, W. W. (2006). Effects of instruction presentation mode in comparative judgment. *Memory & Cognition*, *34*, 196-206.

Shepard, R. N., Kilpatric, D. W., & Cunningham, J. P. (1975). The internal representation of numbers. *Cognitive Psychology*, *7*, 82-138.

Shoben, E. J., Sailor, K. M., & Wang, M. (1989). The role of expectancy in comparative judgments. *Memory & Cognition, 17,* 18-26.

Smith, L. B., Gasser, M., & Sandhofer, C. M. (1997). Learning to talk about the properties of objects: A network model of the development of dimensions. In R. L. Goldstone, D. L. Medin & P. G. Schyns (Eds.), *Advances in the psychology of learning and motivation*, *Vol. 36: Perceptual learning* (pp. 219-255). San Diego, CA: Academic Press.

Trabasso, T., & Riley, C. A. (1975). The construction and use of representations involving linear order. In R. L. Solso (Ed.), Information processing and cognition: The Loyola symposium (pp. 381-410). Hillsdale, NJ: Erlbaum.

Tribushinina, E. (2009). Reference points in linguistic construal: Scalar adjectives revisited. *Studia Linguistica*, *63*, 233-260.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453-458.

von Fersen, L., Wynne, C. D. L., Delius, J. D., & Staddon, J. E. (1991). Transitive inference in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, *17*, 334-341.

Wallis, C. P., & Audley, R. J. (1964). Response instructions and the speed of relative judgments: II. Pitch discrimination. *British Journal of Psychology*, *55*, 121-132.

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M., . . . Miller, B. L. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science, 10*(2), 119-125.

Wickens, T. D. (2002). *Elementary signal detection theory.* New York: Oxford University Press.

Wong, K.-F., & Wang, X.-J. (2006). A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience, 26,* 1314–1328.

Woocher, F. D., Glass, A. L., & Holyoak, K. J. (1978). Positional discriminability in linear orderings. *Memory & Cognition*, *6*, 165–173.

Wynne, C. D. L. (1995). Reinforcement accounts for transitive inference performance. *Animal Learning & Behavior*, *23*, 207-217.

CHAPTER 3:

GENERATIVE INFERENCES BASED ON A DISCRIMINATIVE BAYESIAN MODEL OF RELATION LEARNING

Introduction

Generative and Discriminative Models

Bayesian models of inductive learning can be designed to focus on learning either the probabilities of observable features given concepts (generative models) or the probabilities of concepts given features (discriminative models; Friston et al., 2008; Mackay, 2003). Generative models are especially powerful as they are capable of not only classifying novel instances of concepts (using Bayes' rule to invert conditional probabilities), but also generating representations of possible instances. In contrast, discriminative models focus directly on classification tasks, but do not provide any obvious mechanism for making generative inferences. In recent years, generative Bayesian models have been developed to learn complex concepts based on relational structures (e.g., Goodman, Ullman & Tenenbaum, 2011; Kemp & Jern, 2009; Kemp, Perfors & Tenenbaum, 2007; Tenenbaum, Kemp, Griffiths & Goodman, 2011). Representations of alternative relational structures are used to predict incoming data, and the data in turn are used to revise probability distributions over alternative structures. The highest level of the structure typically consists of a formal grammar or a set of logical rules that generates alternative relational "theories", which are in turn used to predict the observed data. That is, the set of possible relational structures is provided to the system by specifying a grammar that generates them.

Despite their impressive successes, there are some reasons to doubt whether the generative approach provides an adequate basis for all psychological models of relation learning.

Since the postulated grammar of relations is not itself learned, the generative approach implicitly makes rather strong nativist assumptions. Moreover, generative models of relation learning do not fit the intuitive causal direction. For example, it seems odd to claim that a binary relation such as *larger than* somehow acts to causally generate an ordered pair (e.g., <dog, cat>) that constitutes an instantiation of the relation. It seems more natural to consider how observable features of the objects in the ordered pair give rise to the truth of the relation, i.e., to apply a discriminative approach.

Discriminative Models of Relation Learning

Recently, discriminative models have also been applied to relation learning. Silva, Heller, and Ghahramani (2007) developed a discriminative model for relational tasks such as identifying classes of hyperlinks between webpages and classifying relations based on protein interactions. Although their model was developed to address applications in machine learning, the general principles can potentially be incorporated into models of human relational learning. One key idea is that an *n*-ary relation can be represented as a function that takes ordered sets of *n* objects as its input and outputs the probability that these objects instantiate the relation. The model learns a representation of the relation from labeled examples, and then applies the learned representation to classify novel examples. A second key idea is that relation learning can be facilitated by incorporating *empirical priors*, which are derived using some simpler learning task that can serve as a precursor to the relation learning task.

These ideas were incorporated into *Bayesian Analogy with Relational Transformations* (BART), a discriminative model that can learn comparative relations from non-relational inputs (Lu, Chen, & Holyoak, 2012). Given independently-generated feature vectors representing pairs of animals that exemplify a relation, the model acquires representations of first-order

comparative relations (e.g., *larger*, *faster*) as weight distributions over the features. Learning is guided by empirical priors for the weight distributions derived from initial learning of one-place predicates (e.g., *large*, *fast*). BART's learned relations support generalization to new animal pairs, allowing the model to discriminate between novel pairs that instantiate a relation and those that do not. Moreover, BART's learned weight distributions can be systematically transformed to solve analogies based on higher-order relations (e.g., *opposite*).

BART has thus demonstrated promise as a discriminative model of relation learning, which does not presuppose an innate grammar of relations. However, the challenge remains to extend the model to tasks requiring generative inferences. For example, people are able to construct actual instantiations of relations, answering questions such as, "What is an animal that is smaller than a dog?" (Although one might suppose that such questions could be answered by undirected trial-and-error, we shall see that people's answers are often systematically guided by their representations of the relation and of the animal provided as a cue.) Another challenging task is purely hypothetical reasoning, which requires making inferences about arbitrary instances of the relation. Comparative relations such as *larger* exhibit the logical properties of transitivity and asymmetry, supporting deductions such as "If *A* is larger than *B*, and *B* is larger than *C*, then *A* is larger than *C*." Children as young as five or six years can make such transitive inferences reliably (Halford, 1992; Goswami, 1995; Kotovsky & Gentner, 1996). In the present paper we describe an extension of the BART model that addresses these challenges of making generative inferences.

BART Model of Relation Learning

Domain and Inputs

We focus on the same domain and inputs used in the initial BART project (Lu et al., 2012): the domain of comparative relations between animal concepts (e.g., a cow is larger than a sheep). To establish the "ground truth" of whether various pairs of animals instantiate different comparative relations, Lu et al. used a set of human ratings of animals on four different continua (size, speed, fierceness, and intelligence; Holyoak & Mah, 1981). These ratings made it possible to test the model on learning eight different comparative relations: *larger*, *smaller*, *faster*, *slower*, *fiercer*, *meeker*, *smarter*, and *dumber*.

Each animal concept is represented by a real-valued feature vector. In order to avoid the perils of hand-coded inputs (i.e., the possibility that the model's successes may be partly attributable to the foresight and charity of the modelers), we use two sets of inputs that we call "Leuven vectors" and "topics vectors," respectively.

Leuven vectors. We derived Leuven vectors from norms of the frequencies with which participants at the University of Leuven generated features characterizing 129 different animals (De Deyne et al., 2008; see Shafto, Kemp, Mansinghka, & Tenenbaum, 2011). Each animal in the norms is associated with a set of frequencies across more than 750 features. We created vectors of length 50 based on the 50 features most highly associated with the subset of 44 animals that are also in the ratings dataset (Lu et al., 2012). Figure 3.1 provides a visualization (for 30 example animals and the first 15 of the 50 features) of these high-dimensional and distributed representations, which might be similar to the representations underlying people's everyday knowledge of various animals.

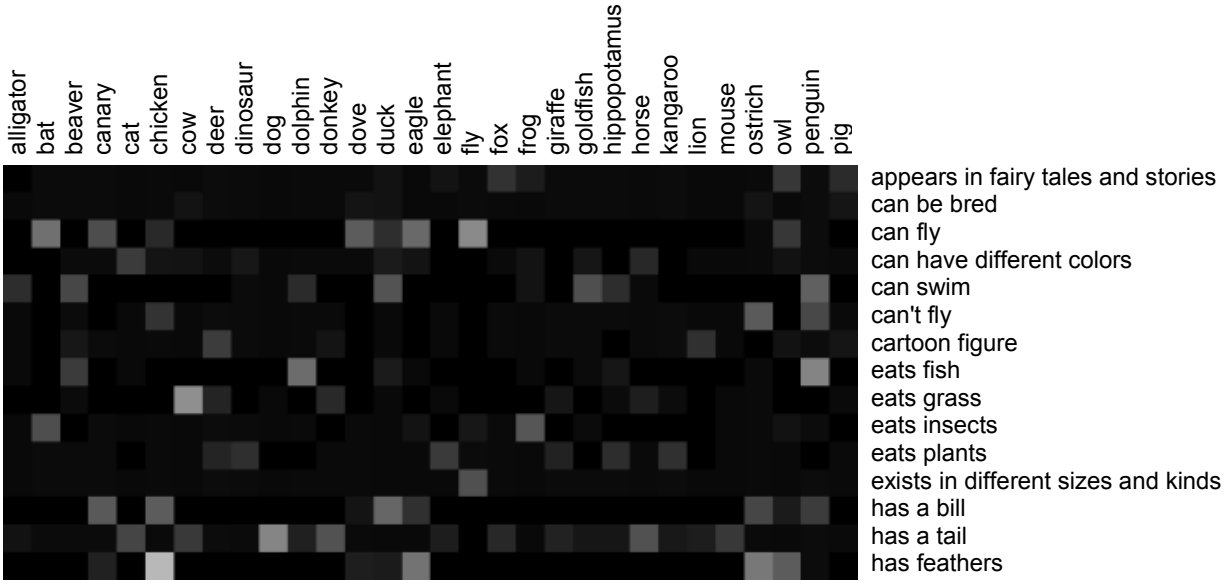


Figure 3.1. Illustration of Leuven vectors (reduced to 15 features to conserve space) for some example animals. The cell intensities represent feature values (light indicates high values and dark indicates low values).

Topics vectors. We obtained topics vectors by running the topic model (Griffiths, Steyvers, & Tenenbaum, 2007) on a pre-processed version of the English Wikipedia corpus, which contained 174,792 entries and 116,128 unique words. This analysis yielded the frequency with which each word was assigned by the topic model to each of 300 different topics, from which we derived a vector representation for each word based on the conditional probability of each of 52 topics given that word. The details of how we ran the topic model and reduced the dimensionality of the resulting vectors are described in Chapter 2 (Chen, Lu, & Holyoak, under review).

Relations as Weight Distributions

BART represents a relation using a joint distribution of weights, **w**, over object features. A relation is learned by estimating the probability distribution $P(\mathbf{w} \mid \mathbf{X_S}, \mathbf{R_S})$, where $\mathbf{X_S}$ represents the feature vectors for object pairs in the training set, the subscript **S** indicates the set

of training examples, and $\mathbf{R_S}$ is a set of binary indicators, each of which (denoted by *R*) indicates whether a particular object (or pair of objects) instantiates the relation or not. The vector **w** constitutes the learned relational representation, which can be interpreted as weights reflecting the influence of the corresponding feature dimensions in **X** on judging whether the relation applies. The weight distribution can be updated based on examples of ordered pairs that instantiate the relation. Formally, the posterior distribution of weights can be computed by applying Bayes' rule using the likelihood of the training data and the prior distribution for **w**:

$$P(\mathbf{w} \mid \mathbf{X_S}, \mathbf{R_S}) = \frac{P(\mathbf{R_S} \mid \mathbf{w}, \mathbf{X_S}) P(\mathbf{w})}{\int_{\mathbf{w}} P(\mathbf{R_S} \mid \mathbf{w}, \mathbf{X_S}) P(\mathbf{w})}. \tag{3.1}$$

The likelihood is defined as a logistic function for computing the probability that a pair of objects instantiates the relation, given the weights and feature vector:

$$P(R = 1 \mid \mathbf{w}, \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^{\mathrm{T}}\mathbf{x}}}. \tag{3.2}$$

The prior, $P(\mathbf{w})$, is a Gaussian distribution and is constructed using a bottom-up approach in which initial learning of simple concepts provides *empirical priors* that guide subsequent learning of more complex concepts. Specifically, BART extracts empirical priors from weight distributions for one-place predicates such as *large* to guide the acquisition of two-place relations such as *larger*. Lu et al. (2012) trained BART on the eight one-place predicates (e.g., *large*, *small, fierce, meek*) that can be formed using the extreme animals at each end of the four relevant continua (size, speed, ferocity, and intelligence).

After learning the joint weight distribution that represents a relation, BART discriminates between pairs that instantiate the relation and those that do not by calculating the probability that a target pair $\mathbf{x}_T$ instantiates the relation *R*:

$$P(R_T = 1 \mid \mathbf{x}_T, \mathbf{X_S}, \mathbf{R_S}) = \int_{\mathbf{w}} P(R_T = 1 \mid \mathbf{x}_T, \mathbf{w}) P(\mathbf{w} \mid \mathbf{X_S}, \mathbf{R_S}). \quad (3.3)$$

Although the general framework of the relation learning model is straightforward, the calculations of the normalization term in Eq. (3.1) and the integral in Eq. (3.3) are intractable, lacking analytic solutions. As in Silva, Heller, and Gharamani (2007), we employed the variational method developed by Jaakkola and Jordan (2000) for Bayesian logistic regression to obtain closed-form approximations to the posterior weight distribution $P(\mathbf{w} \mid \mathbf{X_S}, \mathbf{R_S})$ and the predictive probability $P(R_T = 1 \mid \mathbf{x}_T, \mathbf{X_S}, \mathbf{R_S})$.

Extension to Generative Inference

The goal of the present paper is to endow BART with generative abilities, allowing it (for example) to complete a partially-instantiated relation, answering questions such as, "What is an animal that is smaller than a dog?" We use the weight representation for a relation learned by BART to construct a new generative model for the completion task. When presented with a cue relation (e.g., *smaller*) and a cue object (e.g., dog), the model produces possible responses for the remaining object (e.g., cat) so that the ordered object pair satisfies the relation. More specifically, given the features of an object *B*, $\mathbf{x}_B$, and the knowledge that relation *R* holds for the object pair (*A*, *B*), the model generates a probability distribution for the features of object *A*, $\mathbf{x}_A$, by making the following inference:

$$P(\mathbf{x}_A \mid \mathbf{x}_B, R = 1) \propto P(R = 1 \mid \mathbf{x}_A, \mathbf{x}_B) P(\mathbf{x}_A \mid \mathbf{x}_B). \quad (3.4)$$

The likelihood term, $P(R = 1 \mid \mathbf{x}_A, \mathbf{x}_B)$, is the probability that relation *R* holds for a particular hypothesized object *A*, $\mathbf{x}_A$, and the known object *B*, $\mathbf{x}_B$. It is defined using a logistic function, just as in Eq. (3.2):

$$P(R=1 \mid \mathbf{x}_A, \mathbf{x}_B) = \frac{1}{1+e^{-\mathbf{w}_1^T \mathbf{x}_A - \mathbf{w}_2^T \mathbf{x}_B}}. \quad (3.5)$$

Relative to Eq. (3.2), we have only introduced small differences in the notation. The learned relational weights, **w**, are written as two separate halves: weights associated with the first relational role ($\mathbf{w}_1$) and weights associated with the second relational role ($\mathbf{w}_2$). Similarly, the feature vector **x** for a pair of objects is separated into the feature vector for object *A* ($\mathbf{x}_A$) and the feature vector for object *B* ($\mathbf{x}_B$).

The prior for the features of object *A*, $P(\mathbf{x}_A \mid \mathbf{x}_B)$, is the conditional distribution given the features of object *B*. It is defined as the following:

$$P(\mathbf{x}_A \mid \mathbf{x}_B) = N(\mathbf{x}_B, \sigma^2 \mathbf{I}). \quad (3.6)$$

We assume that object *B* (the referent) serves a starting point for generating object *A*, so the means of $P(\mathbf{x}_A \mid \mathbf{x}_B)$ are taken to be the feature values of object *B*, reflecting a certain degree of semantic dependency between the two objects (i.e., people's tendency to think of *A* objects that are similar to *B*). The prior also encodes the assumptions that the features of *A* are uncorrelated and have the same variance σ^2, the value of which is a free parameter reflecting the strength of the model's preference for generating *A* objects that are similar to *B*.

Our generative model infers a feature distribution for object *A* that reflects a compromise between (1) maximizing the semantic similarity of *A* and *B*, which is reflected in the prior term; and (2) maximizing the probability that the relation holds, which is reflected in the likelihood term. We adapted the variational method to estimate the posterior distribution, using the following updating rules for the mean **μ** and covariance matrix **V** of the feature distribution, as well as the variational parameter ξ:

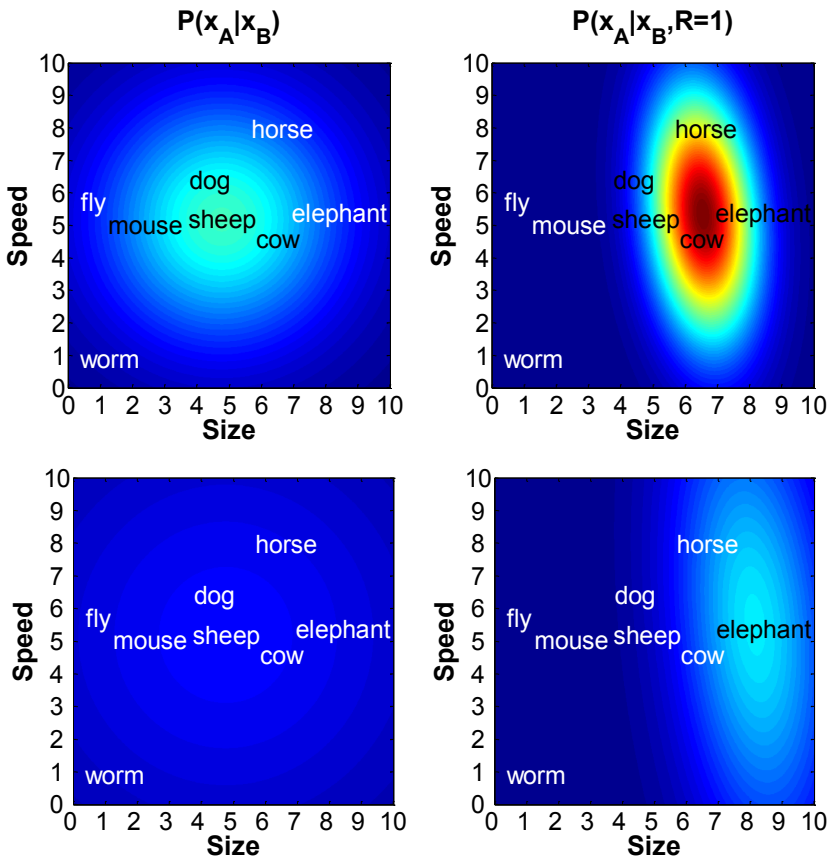


Figure 3.2. Illustration of the generative model for inferring an animal that is larger than a sheep. Colors annotate probability densities (red indicates high values and blue indicates low values). The top panel shows the prior and posterior distributions with $\sigma^2 = 7$ (favoring similarity-based completions such as *cow*), and the bottom panel shows the prior and posterior with $\sigma^2 = 25$ (favoring "landmark" completions such as *elephant*). Various animals are represented in the two-dimensional space based on their size and speed ratings. The posterior was generated using the relational weights that BART learned from the full ratings input (i.e., all four dimensions).

$$\begin{aligned} \mathbf{V}^{-1} &= \frac{\mathbf{I}}{\sigma^2} + 2\lambda(\xi)\mathbf{w}_1\mathbf{w}_1^{\mathrm{T}}, \\ \boldsymbol{\mu} &= \mathbf{V}\left(\frac{\mathbf{I}}{\sigma^2}\mathbf{x}_B + \frac{\mathbf{w}_1}{2} - 2k\lambda(\xi)\mathbf{w}_1\right), \\ \xi^2 &= \mathbf{w}_1^{\mathrm{T}}\left(\mathbf{V} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right)\mathbf{w}_1, \end{aligned} \tag{3.7}$$

where $\lambda(\xi) = \frac{\tanh\left(\frac{1}{2}(\xi + k)\right)}{4(\xi + k)}$ and $k = \mathbf{w}_2^{\mathrm{T}}\mathbf{x}_B$.

Figure 3.2 illustrates the operation of the model in generating an animal (*A*) that is larger than a sheep (*B*). The feature distribution for *A* is updated from a prior favoring some degree of similarity between the two animals (left panel; top: high similarity, bottom: low similarity) to a posterior distribution after taking into consideration the relation (i.e., *larger*) instantiated by the animals (right panel). These distributions are shown in a simplified two-dimensional feature space (the size and speed ratings for animals; Holyoak & Mah, 1981).

Modeling Transitive Inference

Comparative relations such as *larger* exhibit the logical properties of transitivity and asymmetry, supporting deductions such as, "If *A* is larger than *B* and *B* is larger than *C*, then *A* is larger than *C*." Such hypothetical reasoning seems to depend on the ability to generate arbitrary instantiations of the relation without any guidance from object features (as the object representations are semantically empty). Our first test evaluated whether the generative extension of BART enables transitive inferences on comparative relations using arbitrary hypothetical instances.

Operation of the Model

The basic approach to transitive inference is straightforward: The model "imagines" objects *A*, *B*, and *C* that instantiate the two given premises, as in the example above, and then tests the unstated relationship for the pair <*A*, *C*>. If the *larger* relation that BART has learned is indeed transitive, then any such instantiation will satisfy the conclusion, "*A* is larger than *C*." This test is done repeatedly, in essence searching for a counterexample. If no counterexample is ever found, the transitive inference is accepted.

Specifically, for each of the eight comparative relations that BART learned, we first let the model "imagine" an animal *B* (because the statement "*A* is larger than *B*" implies that *B* is the referent against which *A* is being compared) by sampling a feature vector from a distribution representing the animal category. This is a Gaussian distribution with a mean vector and covariance matrix that were directly estimated from the feature vectors of the animals in the ratings dataset that had Leuven or topics vectors, respectively. There were 44 such animals for the Leuven inputs and 77 such animals for the topics inputs.

Given the sampled animal *B*, the generative model constructs a distribution for animal *A* (e.g., to satisfy the premise that "*A* is larger than *B*") by letting *B* fill the second role of the relevant relation. Similarly, the model constructs a distribution for animal *C* (e.g., to satisfy the premise that "*B* is larger than *C*") by letting *B* fill the first role of the same relation. Next, the model creates feature representations for specific animals *A* and *C* by setting their feature vectors, $\mathbf{x}_A$ and $\mathbf{x}_C$, to be the means of the inferred feature distributions for *A* and *C*, respectively. Note that these "imagined" animals are hypothetical: although their features are sampled from the distribution of animal features, the results will seldom correspond to actual animals. To ensure that the premises have actually been satisfied, the model accepts the imagined animal *A* only if $P(R=1 \mid \mathbf{x}_A, \mathbf{x}_B) > 0.5$ and $P(R=1 \mid \mathbf{x}_B, \mathbf{x}_A) < 0.5,$ and the imagined animal *C* only if $P(R=1 \mid \mathbf{x}_B, \mathbf{x}_C) > 0.5$ and $P(R=1 \mid \mathbf{x}_C, \mathbf{x}_B) < 0.5.$

Finally, if $\mathbf{x}_A$ and $\mathbf{x}_C$ have been accepted as satisfying the premises, the model calculates both $P(R=1 \mid \mathbf{x}_A, \mathbf{x}_C)$, denoting the probability that *A* is larger than *C,* and $P(R=1 \mid \mathbf{x}_C, \mathbf{x}_A)$, denoting the probability that *C* is larger than *A*. The model concludes that the relation holds for

the pair <*A*, *C*> (and not for <*C*, *A*>) if $P(R=1 \mid \mathbf{x}_A, \mathbf{x}_C) > 0.5$ and $P(R=1 \mid \mathbf{x}_C, \mathbf{x}_A) < 0.5$, implying that a counterexample has not yet been found to refute the transitive inference.

Evaluation of the Model

We conducted tests of transitive inference using the relational representations that BART learned based on 100 randomly-chosen training pairs. For comparison, we also tested a baseline model that substituted an uninformative prior for the empirical prior that guides BART's relation learning (see Lu et al., 2012). For each of the eight comparative relations, the relation learning model was run ten times, each time with a different set of training pairs and resulting in a different learned weight distribution. For each of these 80 learned weight distributions, we let the model generate 100 *A*-*B*-*C* triads satisfying the premises, testing the relevant relationship between *A* and *C* for each triad. To assess the influence of the free parameter in model predictions, the tests were conducted multiple times with different values of σ^2 ranging from 1 to 1,000 for the Leuven inputs and from 100 to 100,000 for the topics inputs.[1] The strongest tests are those in which σ^2 is set at low values, creating a strong prior preference that *A*, *B*, and *C* are similar to one another. When the similarity constraint is strong, the model is forced to generate animals that are similar on the relevant dimension, and hence more likely to yield a counterexample.

Results for Leuven inputs. When the value of σ^2 was reduced below 1 for the Leuven inputs, the models produced many instantiations that did not satisfy the required premises (i.e., *A* > *B*, *B* > *C*, and not vice versa). We therefore treated the value of 1 as the minimal value of σ^2 that yields triplets of animals with discriminable values on the relevant dimension for the Leuven inputs. Figure 3.3 shows the mean proportion correct (i.e., the mean proportion of triads that satisfy the conclusion based on transitive inference) for BART and the baseline model as a

function of σ^2, using Leuven vectors. These results were averaged over all 80 learned relational weight distributions. The critical result is that BART's accuracy remained constant at 100% as σ^2 was reduced to the effective minimal value of 1. Thus, BART demonstrates what may be considered an inductive approximation to deduction: despite exhaustive search for a counterexample to the transitive inference, no counterexample was ever found. In contrast, the baseline model often failed to infer that $A > C$ (and not vice versa) even when the value of σ^2 was as large as 100.

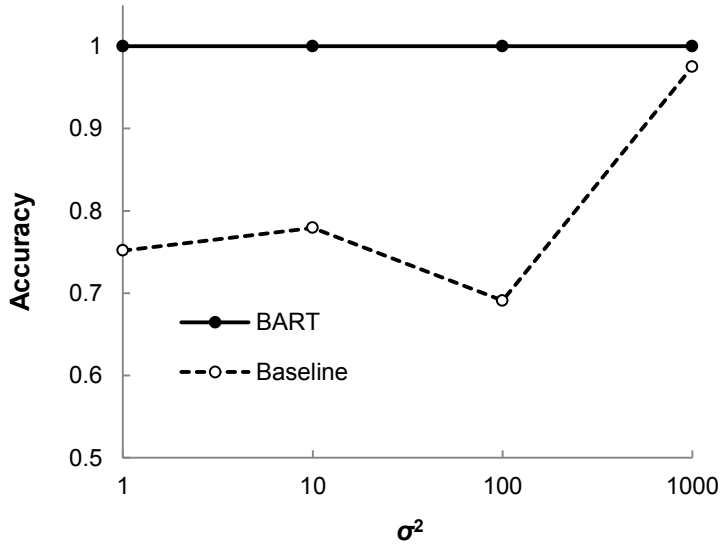


Figure 3.3. Mean proportion correct on the transitive inference task for BART and the baseline model using Leuven vectors, as a function of the variance parameter. These results are averaged over the 80 learned relational weight distributions.

Results for topics inputs. Figure 3.4 shows the mean proportion correct for BART and the baseline model as a function of σ^2. BART remained at 100% accuracy for the different values of σ^2. In stark contrast, the baseline model often could not find the required 100 *A-B-C* triads satisfying the premises after generating 10,000 triads total for each learned relational

weight distribution, even when σ^2 was set to the maximum value of 100,000. Nevertheless, the figure shows the mean proportion correct for the triads generated by the baseline model that did satisfy the premises (of which there were on average 1.54, 22.7, 67.71, and 73.43, respectively, for the values of σ^2 from 100 to 100,000). Once again, whereas the baseline model finds many counterexamples to the transitive inference, BART demonstrates that the comparative relations it has learned are indeed transitive and asymmetric.

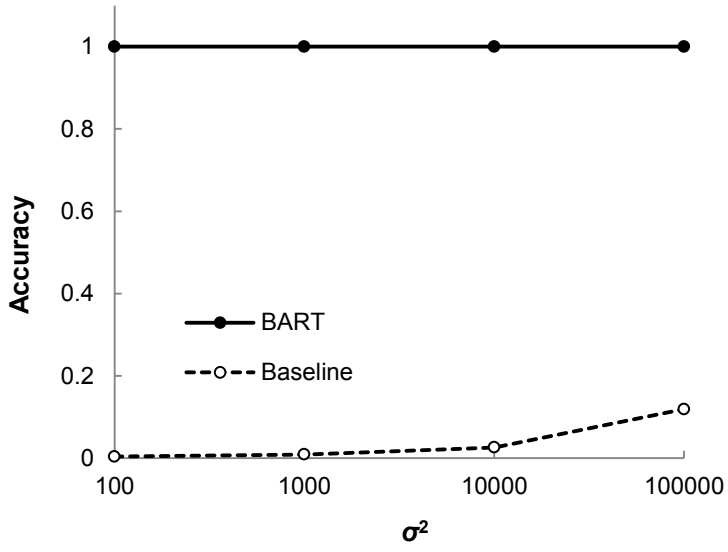


Figure 3.4. Mean proportion correct on the transitive inference task for BART and the baseline model using topics vectors, as a function of the variance parameter. These results are averaged over the 80 learned relational weight distributions.

Animal Generation Task

A second evaluation of the model involves predicting the pattern of human responses in an animal generation study conducted using Amazon Mechanical Turk. In this free-generation study, participants typed responses to queries of the form, "Name an animal that is larger than a dog." They were instructed to enter the first animal that came to mind. Four comparative

relations (*larger*, *smaller*, *faster*, and *slower*) and nine cue animals (shark, ostrich, sheep, dog, fox, turkey, duck, dove, and sparrow) were used. At least 50 responses were collected for each of the 36 relation-animal pairs. To minimize learning across trials, we asked each participant to answer only two questions about a single animal: either *larger* and then *slower*, *slower* and then *larger*, *faster* and then *smaller*, or *smaller* and then *faster*.

MTurkers were instructed to complete the study only if they were fluent in English. There were 1,147 participants, resulting in a total of 2,294 responses across the 36 queries. We processed the responses by removing articles such as "an," correcting obvious misspellings (e.g., "pidgeon"), and expanding abbreviations (e.g., "hippo"). We then removed two of the responses ("dig" and "bow") because it was not clear what animals they were supposed to be.

The same 36 relation-animal pairs were presented to the model after it had been trained on the relevant relations using either Leuven or topics vectors. For each question, the model produced a continuous posterior distribution for the feature vector of the missing animal using Eq. (3.4). This distribution was used to calculate the probability densities for the feature vectors of various animals. For the Leuven inputs, we used all 129 animals in the Leuven dataset. For the topics inputs, we used the set of 168 animals that participants provided as a response at least twice in the entire MTurk study. In both cases, the set of animals for which we obtained model predictions included many animals outside of the original training set given to the relation learning model, which was restricted to animals in the ratings dataset. The probability densities calculated for all 129 or 168 animals were normalized to produce a discrete probability distribution. These discrete probabilities were then averaged across the ten runs of the relation learning model.

Human Results

The complete set of human responses is shown in the Appendix. The responses appear to be mainly driven by two trends: (1) reporting an animal similar to the cue animal and fitting the cue relation (e.g., *cat* for "smaller than a dog"), and (2) reporting a "landmark" animal at an extreme of the continuum (e.g., *turtle* for "slower than a dog"). The landmark animal coupled with the cue animal provides an "ideal" example of the cue relation (i.e., one that maximizes the probability that the relation holds). This tradeoff between reporting animals that are similar to the cue animal and reporting animals that are landmarks for the cue relation (and usually more dissimilar to the cue animal) is captured by the single free parameter in the generative module, σ^2. As explained earlier (see Figure 3.2), a low σ^2 results in a response distribution that favors animals similar to the cue animal, whereas a high σ^2 leads to a preference for response animals that are more likely to satisfy the cue relation with respect to the cue animal (i.e., landmark animals for the cue relation).

Another pattern we observed in the human responses is that the responses to each query were often dominated by the most popular response to that query. Averaged across all 36 queries, about 40% of the responses to each query were the most frequent response. A typical pattern of human responses is displayed in Figure 3.5, which shows the response frequencies and proportions (out of 53 total responses) to the query, "Name an animal that is slower than a dog." The most dominant response of *turtle* is followed by a long tail of low-frequency responses. It is difficult to explain exactly why some participants chose these less-frequent responses, especially *baby seal*, *seahorse*, or even *pig*, which was given as an answer by three different participants and ties with *cat* for third place. Therefore, we focused on the most popular response to each query when analyzing the model predictions.

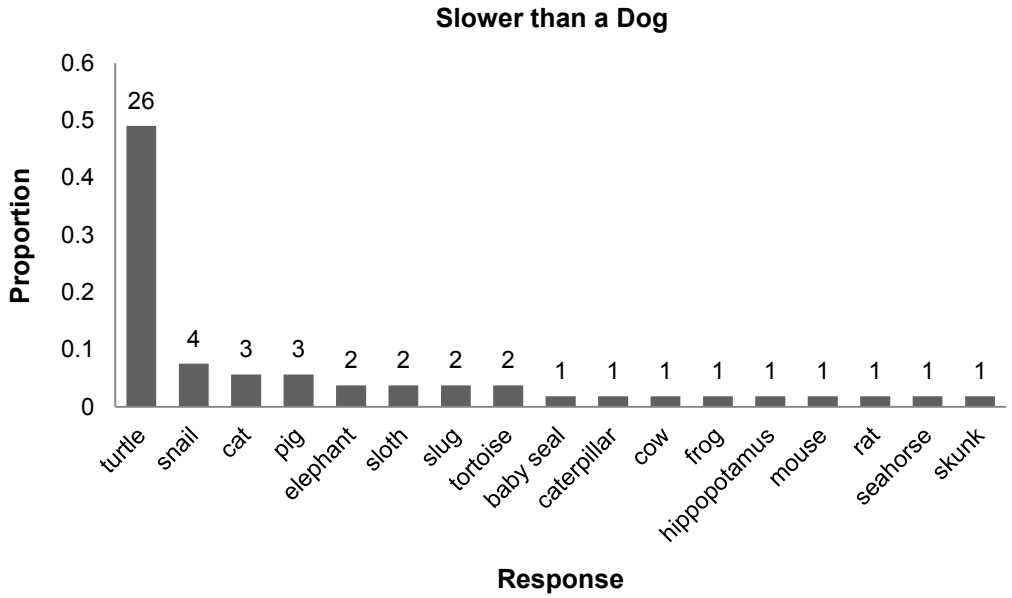


Figure 3.5. A typical pattern of human responses in the animal generation task, showing response proportions and frequencies (shown above the bars) for the query, "Name an animal that is slower than a dog."

Model Results

We evaluated the model with respect to predictions of the most frequent human response to each question, considering both whether the model actually gave the highest probability to that response as well as that animal's predicted rank among the entire set of animals for which we obtained model predictions. Some of the four tested relations seemed to encourage a "landmark" response strategy (especially *faster* and *slower*, and *larger* to a lesser extent) whereas others seemed to elicit more responses based on similarity to the cue animal. Accordingly, the variance parameter in the generative model was chosen separately for each relation to maximize the number of questions involving that relation for which the model gave the highest probability to the top human response (the number of exactly correct predictions), with ties broken by the median of the predicted ranks for the top responses to those questions (a lower median rank is considered better).

Results for Leuven inputs. We compared the full generative model with two alternative, simpler models. The first alternative model used the prior distribution from the generative model, $P(\mathbf{x}_A \mid \mathbf{x}_B)$, to calculate probability densities for the set of 129 Leuven animals, and thus considered only the similarity of each possible response to the cue animal. The second alternative model calculated the likelihood probability, $P(R=1 \mid \mathbf{x}_A, \mathbf{x}_B)$, for each of the animals, and thus cared only about the probability that a possible response satisfies the cue relation with respect to the cue animal.

We chose the variance parameter for the generative model from the values 1, 5, 10, 50, and 100. The best-performing variances were 50, 10, 10, and 100, respectively, for *larger*, *smaller*, *faster*, and *slower*. For *larger*, *smaller*, and *slower*, the chosen variances are sensible given the general pattern of "landmark" versus "similarity" responses for these relations. The relatively small value of 10 for *faster* is also sensible because the Leuven dataset does not include *cheetah*, the landmark animal for the *faster* relation and the most popular human response to all of the *faster* questions, so the model had to instead predict the second most popular response, which was often based more on similarity.

Number of correct predictions. The full generative model correctly predicted the top human response for 13 of the 36 questions, which is impressive considering that there were 129 animals to choose from for each question. In fact, the probability of correctly predicting the top response for at least 13 of the 36 questions by random chance is only $\sum_{i=13}^{36} \binom{36}{i} \left(\frac{1}{129}\right)^i \left(\frac{128}{129}\right)^{36-i} \approx 7.14 \times 10^{-19}$. Because about 40% of all human responses were the most frequent responses, we would expect a human participant to provide the top response for $36 \times 0.4 = 14.4$ questions. In contrast, the alternative model that uses only the prior term (the

"prior" model) correctly predicted the top response for only one of the 36 questions ("smaller than a dog," to which the top response was *cat*), and the likelihood model made only four correct predictions (for one *faster* question and three *slower* questions). The probabilities of getting at least one correct and at least four correct by random chance are about .24 and 1.74×10^{-4}, respectively. The full generative model correctly predicted the top response for two *larger* questions, one *smaller* question, one *faster* question, and all nine *slower* questions. Predicting that *turtle* would be the top human response to all nine *slower* questions required an impressive feat of generalization on the model's part, because *turtle* was not in the original training set given to the relation learning model.

Median ranks. We also analyzed the medians of the ranks that the models assigned to the top human responses. Even if the models did not always predict the highest probabilities for the top responses, they may have given them relatively high probabilities, so the 129 animals were ranked in descending order of predicted probability for each question. The median was chosen so that a few outliers would not affect the results too much, although the results were very similar for means. Across all 36 questions, the median of the ranks that the full generative model assigned to the top human responses was 8.5. In comparison, the median rank was 71.5 for the prior model and 11.5 for the likelihood model. Figure 3.6 shows the breakdown of these results for the four comparative relations, with the median ranks displayed above the bars. For easier comparison with the topics inputs, for which the models considered a different number of animals (168), the y-axis shows the median rank as a fraction of the total number of animals considered (129 in this case). As can be seen, the prior model performed very badly on all four relations, and the likelihood model tended to perform slightly worse than the full generative model. These results indicate that the full generative model, which considers both similarity to

the cue animal and the likelihood of satisfying the cue relation with respect to the cue animal, predicts the pattern of human responses better than models that consider only one of these factors.

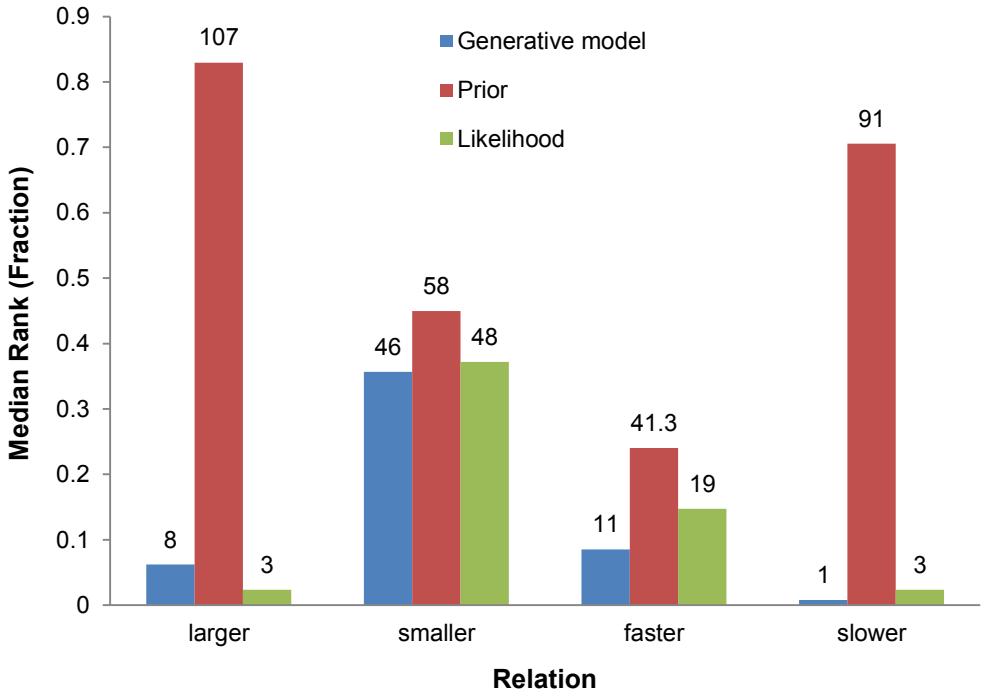


Figure 3.6. Median ranks for the top human responses assigned by the different models using Leuven vectors, broken down by relation. The y-axis shows the median rank as a fraction of the total number of animals considered by the models. The actual median ranks (out of 129 animals) are shown above the bars. Note that lower values indicate better performance.

Results for topics inputs. Again, we compared the full generative model against two alternative models, one of which was the likelihood model that we tested for the Leuven inputs. The second alternative model calculated a probabilistic quantity that represents word association strength in the topic model (Griffiths et al., 2007, p. 221):

$$P(w_2 \mid w_1) = \sum_z P(w_2 \mid z) P(z \mid w_1). \tag{3.8}$$

For a given cue animal word w_1, we calculated $P(w_2 \mid w_1)$ for each of the 168 possible response animal words (w_2) using all 300 topics (*z*) obtained from the topic model. This method yielded a predicted probability for each possible response animal corresponding to the semantic association strength from the cue animal word to the response animal word, which can be (but is not always) based on feature similarity between the two animals.

For the topics inputs, the variance parameter for the generative model was chosen from the values 100, 500, 1000, 5000, and 10000. The variances selected were 10000, 1000, 10000, and 500, respectively, for *larger*, *smaller*, *faster*, and *slower*. These values are sensible for *larger*, *smaller*, and *faster* given their patterns of "landmark" versus "similarity" responses. As we will see, the generative model performed the worst on the *slower* questions (perhaps because *turtle* was not in the original training set), though still better than both of the alternative models.

Number of correct predictions. The full generative model correctly predicted the top human response for 15 of the 36 questions. The probability of making at least 15 correct predictions by random chance when there are 168 animals to choose from for each question is about 2.07×10^{-24}. In contrast, both the likelihood model and the model based on word association correctly predicted the top human response for only one of the 36 questions (one *faster* question and one *smaller* question, respectively), the corresponding chance probability for which is about .19. The full generative model correctly predicted the top response for two *larger* questions, three *smaller* questions, all nine *faster* questions, and one *slower* question. Of particular note, predicting that *cheetah* would be the top human response to all nine *faster* questions required the model to generalize beyond the set of animals it encountered when learning the comparative relations.

Median ranks. As before, we ranked the 168 considered animals by their predicted probabilities for each question. The median rank for the top human response across all 36 questions was 7 for the full generative model, 24 for the model based on word association, and 29 for the likelihood model. Figure 3.7 shows the breakdown of these results for the four relations. The full generative model performed better than the two alternative models on all four relations. These results indicate that the generative model accounts for the human data better than either simple word association or mere consideration of the relation. Table 3.1 summarizes all the model results on the animal generation task for both Leuven and topics inputs.

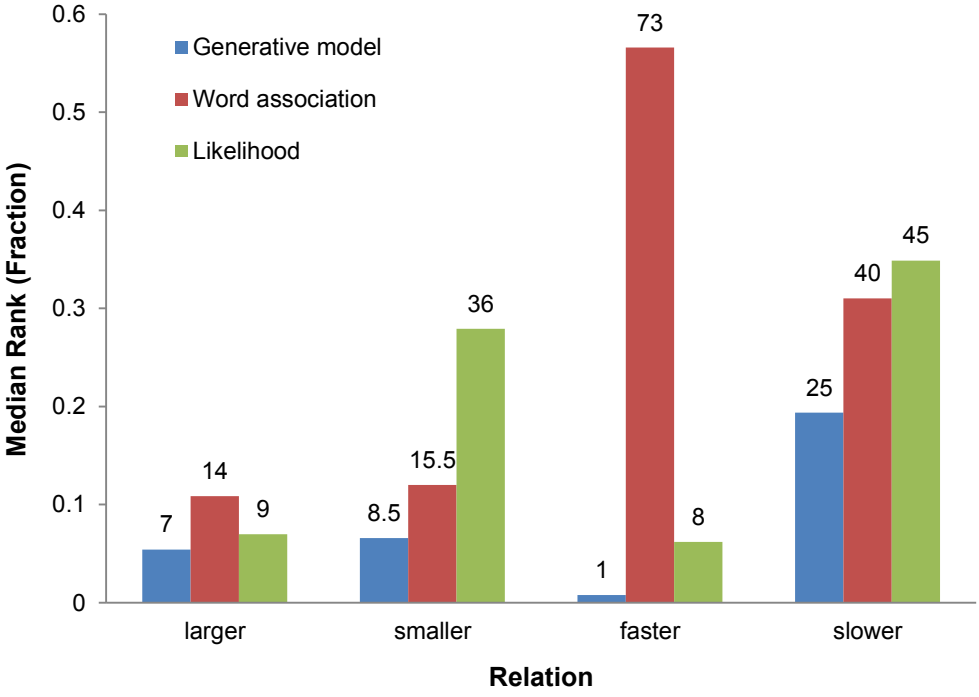


Figure 3.7. Median ranks for the top human responses assigned by the different models using topics vectors, broken down by relation. The y-axis shows the median rank as a fraction of the total number of animals considered by the models. The actual median ranks (out of 168 animals) are shown above the bars. Note that lower values indicate better performance.

Table 3.1

Summary of Model Results on the Animal Generation Task

		Leuven inputs			**Topics inputs**		
		Generative model	Prior	Likelihood	Generative model	Word association	Likelihood
Number correct	Overall	13	1	4	15	1	1
	larger	2	0	0	2	0	0
	smaller	1	1	0	3	1	0
	faster	1	0	1	9	0	1
	slower	9	0	3	1	0	0
Median rank (and fraction)	Overall	8.5 (.07)	71.5 (.55)	11.5 (.09)	7 (.04)	24 (.14)	29 (.17)
	larger	8 (.06)	107 (.83)	3 (.02)	7 (.05)	14 (.11)	9 (.07)
	smaller	46 (.36)	58 (.45)	48 (.37)	8.5 (.07)	15.5 (.12)	36 (.28)
	faster	11 (.09)	31 (.24)	19 (.15)	1 (.01)	73 (.57)	8 (.06)
	slower	1 (.01)	91 (.71)	3 (.02)	25 (.19)	40 (.31)	45 (.35)

General Discussion

These results on modeling transitive inference and predicting human responses on the animal generation task provide evidence that a discriminative model of relation learning, BART (Lu et al., 2012), can be extended to yield generative inferences. These inferences can involve relations between either hypothetical (in the case of transitive inference) or actual (in the case of animal generation) objects.

The model's ability to make transitive inferences based on relations it has learned from examples in a bottom-up fashion illustrates the potential power of the discriminative approach to relation learning. Importantly, BART is not endowed with any notion of what a "transitive and asymmetric" relation is (though like a 6-year-old child, it *is* endowed with sufficient working memory to integrate two relations as premises). Rather, it simply uses its learned comparative relations to imagine possible object triads, and without exception concludes that the inference

warranted by transitivity holds in each such triad. The model thus approximates "logical" reasoning by a systematic search for counterexamples (and failing to find any), akin to a basic mechanism postulated by the theory of mental models (Johnson-Laird, 2008). The fact that BART achieves error-free performance in the tests of transitive inference is especially impressive given that its inductively-acquired relational representations are most certainly fallible (e.g., the model makes errors in judging which of two animals close in size is the larger; see Lu et al., 2012). It turns out that imperfect representations of comparative relations, acquired by bottom-up induction, can be sufficiently robust as to yield reliable quasi-deductive transitive inferences.

In the animal generation task, the generative extension of BART achieves moderate success in modeling human response patterns by maximizing both similarity to the cue animal and the probability that the cue relation is satisfied, performing better than models that consider only each of these factors alone. Importantly, BART's ability to generate a feature distribution for hypothetical animals that satisfy a certain relation with respect to another animal could be used to generate many more training examples for the relation learning model than is currently available. When fed back to the learning model, these extra examples might improve the learned relational representations. In fact, because the generated animals are constrained to be similar to the cue animal, each pair of animals should differ only on a few relevant dimensions, thus constituting a more ideal example for the relation and allowing the learning model to narrow down the most important features. Repeating this procedure of generating more examples and then relearning the relation based on the additional examples could allow the model to progressively refine its learned relations, capitalizing on the bootstrapping strategy that has proven valuable in our work so far (e.g., empirical priors).

Footnote

1. We use different ranges of σ^2 for the Leuven and topics inputs because they are scaled differently. Across the animals in the ratings dataset, the mean variance among the 50 Leuven features is .79, whereas the mean variance among the 52 topics features is 147.24.

Appendix: Human Responses on the Animal Generation Task

Cue relation	Cue animal	*n*[a]	Response proportions									
larger	shark	66	whale .79	elephant .17	bear .02	snake .02	zebra .02					
	ostrich	70	elephant .63	giraffe .13	whale .04	bear .03	lion .03	rhinoceros .03	camel .01	cow .01	hippo .01	other .07
	sheep	73	elephant .29	cow .19	horse .15	bear .11	lion .04	whale .04	giraffe .03	tiger .03	bison .01	other .11
	dog	53	elephant .26	horse .19	cow .11	bear .08	lion .06	tiger .04	whale .04	bull .02	cat .02	other .19
	fox	79	elephant .28	bear .27	wolf .13	dog .08	tiger .05	cheetah .04	deer .03	horse .03	lion .03	other .09
	turkey	65	elephant .22	dog .12	bear .11	cow .11	lion .05	ostrich .05	peacock .05	deer .03	giraffe .03	other .25
	duck	69	elephant .22	dog .14	goose .13	lion .07	bear .06	horse .06	chicken .03	peacock .03	pig .03	other .23
	dove	63	elephant .16	tiger .14	eagle .13	cat .10	dog .10	cow .06	bear .05	chicken .03	wolf .03	other .21
	sparrow	58	dog .19	elephant .16	eagle .10	hawk .07	bear .05	cat .03	crow .03	giraffe .03	ostrich .03	other .29
smaller	shark	59	cat .10	dog .10	fish .10	turtle .10	goldfish .08	dolphin .07	mouse .05	rabbit .05	ant .02	other .32
	ostrich	66	mouse .15	cat .14	dog .11	chicken .06	rabbit .06	ant .03	bird .03	chinchilla .03	rooster .03	other .36
	sheep	61	cat .21	mouse .16	dog .13	frog .08	rabbit .07	pig .05	ant .03	chicken .03	goat .03	other .20
	dog	65	cat .31	mouse .22	rat .17	rabbit .06	bird .05	frog .05	gerbil .03	hamster .03	squirrel .03	other .06
	fox	63	mouse .29	cat .21	rabbit .19	rat .10	bird .03	turtle .03	ant .02	box turtle .02	canary .02	other .11
	turkey	60	chicken .27	mouse .18	cat .10	rat .10	duck .03	fish .03	hamster .03	rabbit .03	squirrel .03	other .18
	duck	62	mouse .42	squirrel .06	chick .05	rat .05	ant .03	bird .03	fish .03	frog .03	goose .03	other .26
	dove	55	mouse .36	hummingbird .15	ant .05	rat .05	sparrow .05	worm .05	fly .04	bee .02	beetle .02	other .20
	sparrow	58	mouse .26	hummingbird .19	ant .09	worm .09	goldfish .03	mole .03	rat .03	robin .03	snail .03	other .21

faster	shark	59	cheetah .59	dolphin .17	eagle .03	jaguar .03	antelope .02	bird .02	falcon .02	gazelle .02	lion .02	other .08
	ostrich	66	cheetah .71	tiger .06	cat .03	cougar .03	jaguar .03	coyote .02	eagle .02	emu .02	gazelle .02	other .08
	sheep	61	cheetah .36	dog .10	horse .10	tiger .07	cat .05	cougar .03	fox .03	jaguar .03	lion .03	other .20
	dog	65	cheetah .69	horse .08	tiger .05	cat .03	leopard .03	bird .02	coyote .02	deer .02	fox .02	other .06
	fox	63	cheetah .75	jaguar .06	deer .05	cat .03	tiger .03	cougar .02	gazelle .02	giraffe .02	horse .02	lion .02
	turkey	60	cheetah .42	rabbit .10	cat .07	dog .05	horse .05	ostrich .05	fox .03	leopard .03	zebra .03	other .17
	duck	62	cheetah .35	dog .13	cat .10	deer .03	horse .03	rabbit .03	snake .03	alligator .02	bird .02	other .26
	dove	55	cheetah .38	eagle .15	falcon .11	hawk .07	hummingbird .05	cat .04	tiger .04	bluebird .02	dog .02	other .13
	sparrow	58	cheetah .53	eagle .14	hawk .07	bee .03	lion .03	rabbit .03	zebra .03	cougar .02	falcon .02	other .09
slower	shark	66	turtle .38	snail .15	whale .09	fish .08	dog .06	sloth .06	tortoise .03	walrus .03	bird .02	other .11
	ostrich	70	turtle .39	sloth .20	snail .07	cat .06	cow .06	dog .04	elephant .04	bear .03	duck .01	other .10
	sheep	74	turtle .42	sloth .20	snail .12	cow .03	donkey .03	elephant .03	moose .03	camel .01	cat .01	other .12
	dog	53	turtle .49	snail .08	cat .06	pig .06	elephant .04	sloth .04	slug .04	tortoise .04	baby seal .02	other .15
	fox	79	turtle .68	snail .06	rabbit .05	sloth .05	deer .03	dog .03	elephant .03	bear .01	cow .01	other .05
	turkey	65	turtle .52	sloth .12	snail .08	slug .06	duck .03	ant .02	cat .02	chick .02	chicken .02	other .12
	duck	70	turtle .47	snail .20	slug .06	tortoise .06	worm .04	chicken .03	sloth .03	beetle .01	elephant .01	other .09
	dove	63	turtle .48	sloth .14	snail .14	slug .05	elephant .03	worm .03	bear .02	duck .02	fox .02	other .08
	sparrow	58	turtle .29	sloth .21	snail .14	dog .03	ostrich .03	tortoise .03	worm .03	bear .02	cat .02	other .19

Note. The nine most frequent responses are shown for each question (except for "faster than a fox," which had exactly ten unique responses). The total proportion of the other responses to each question is shown in the "other" column.

[a] The total number of responses for each question.

References

Chen, D., Lu, H., & Holyoak, K. J. (under review). The discovery and comparison of symbolic magnitudes.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel,W., Dry, M., Voorspoels, W., & Storms G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods, 40*, 1030-1048.

Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, H., Penny, W., & Ashburner, J. (2008). Bayesian decoding of brain images. *NeuroImage*, *39*, 181-205.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*, 110-119.

Goswami, U. (1995). Transitive relational mappings in 3-and 4-year-olds: The analogy of Goldilocks and the Three Bears. *Child Development, 66,* 877-892.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211-244.

Halford, G. S. (1992). Analogical reasoning and conceptual complexity in cognitive development. *Human Development, 35*, 193-217.

Holyoak, K. J., & Mah, W. A. (1981). Semantic congruity in symbolic comparisons: Evidence against an expectancy hypothesis. *Memory & Cognition*, *9*, 197-204.

Jaakkola, T. S., & Jordan, M. I. (2000). Bayesian logistic regression: A variational approach. *Statistics and Computing, 10*, 25-37.

Johnson-Laird, P.N. (2008) Mental models and deductive reasoning. In L. Rips & J. Adler. (Eds.), *Reasoning: Studies in human inference and its foundations* (pp. 206-222). Cambridge, UK: Cambridge University Press.

Kemp, C., & Jern, A. (2009). Abstraction and relational learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams & A. Culotta (Eds.), *Advances in Neural Information Processing Systems*, *22*, 943-951.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*, 307–321.

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

Kotovsky L, & Gentner D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*, 2797-2822.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*, 617-648.

Mackay, D. (2003). *Information theory, inference and learning algorithms.* Cambridge, UK: Cambridge University Press.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition, 120*, 1-25.

Silva, R., Heller, K., & Ghahramani, Z. (2007). Analogical reasoning with relational Bayesian sets. In M. Mella & X. Shen (Eds.), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279-1285.