Learning Domain Structures

Charles Kemp, Amy Perfors & Joshua B. Tenenbaum
{ckemp, perfors, jbt}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

How do people acquire and use knowledge about domain structures, such as the tree-structured taxonomy of folk biology? These structures are typically seen either as consequences of innate domain-specific knowledge or as epiphenomena of domain-general associative learning. We present an alternative: a framework for statistical inference that discovers the structural principles that best account for different domains of objects and their properties. Our approach infers that a tree structure is best for a biological dataset, and a linear structure ("left"–"right") is best for a dataset of people and their political views. We compare our proposal with unstructured associative learning and argue that our structured approach gives the better account of inductive generalization in the domain of folk biology.

Psychologists have argued that cognition in different domains draws on qualitatively different mental representations. Tree structures appear well-suited to representing relationships between animal species [1, 2, 10], while a one-dimensional structure (the liberal-conservative spectrum) seems better for representing people's political views. The possibility of different structures raises a fundamental question: how do people learn what kind of structure is appropriate in each domain?

The standard approach to this question is to reject one of its assumptions. Nativists deny that core structures are learned, at least for evolutionarily important domains like folkbiology. Instead, infants come equipped with innate knowledge about which structures are appropriate for which domains. Atran [1], for example, argues that folkbiology is a core domain of human knowledge, and that the tendency to group living kinds into hierarchies reflects an "innately determined cognitive structure." More generally, Keil [8] has argued that ontological knowledge obeys an innate "M-constraint", requiring the extensions of predicates to conform to rigidly tree-structured hierarchies of objects.

Alternatively, empiricists generally deny that structured representations are present at all. Domain-specific representations are merely emergent properties of unstructured, domain-general associative learning architectures. McClelland and Rogers [12], for example, have recently suggested that the acquisition of semantic knowledge in domains such as intuitive biology can be explained as learning in a generic connectionist network. Their architecture never explicitly represents any tree structure, although with repeated training, its hidden unit representations may implicitly come to approximate the taxonomic relations between biological species.

This paper proposes an alternative approach – structure learning – that combines important insights from both of these traditions. Our key contribution is to show how structured domain representations can be acquired within a domain-general framework for Bayesian inference. Like nativists, we suggest that different domains are represented with qualitatively different structures, and we show how these structured representations serve as critical constraints on inductive generalization. Like empiricists, though, we emphasize the importance of learning, and attempt to show how domain structures can be acquired through domain-general statistical inference. This is not only more parsimonious than the nativist position, but allows us to explain the origin of structured representations in novel domains, where the prior existence of domain-specific innate structure is highly implausible.

After describing our structure learning framework, we present two empirical tests of its performance. First, we show that it chooses the appropriate domain structure for both synthetic and real-world data sets. It correctly chooses a tree structure for a biological domain (animal feature judgments), and a linear structure for a political domain (US Supreme Court decisions). Second, we model two classic data sets of inductive judgments in biology [13] and show that our framework performs better than an unstructured connectionist approach.

Bayesian structure learning

Our proposal takes the form of a rational analysis. We aim to demonstrate the computational plausibility and explanatory value of Bayesian structure learning, but leave for future work the question of how these computations might be implemented or approximated by cognitive processes. Assume the learner's data consist of a binary-valued object-feature matrix D specifying the features of each object in a given domain. In biology, for instance, the rows of D might correspond to species, and the columns to anatomical and behavioral attributes. The entry in row i and column j would then specify the value of feature j for species i. Structure-learning includes computational problems at two levels. First, which *structure class* is most appropriate for the domain? Second, given a structure class, which structure

in that class provides the best account of the data?

For instance, suppose that a learner exposed to biological data ends up organizing animal species into a taxonomic tree. The first problem asks how she knew to use a tree rather than some other kind of structure. The second problem asks why she settled on one specific tree instead of the many other trees she might have chosen. Our focus here is on the first problem – the problem of inferring the right structure class for a domain. A solution to the second problem, however, falls out of our probabilistic approach.

We assume that learners come to a domain equipped with a hypothesis space of structure classes, either constructed from innate primitives or based on analogies with previously learned domains. For simplicity, this paper considers a hypothesis space of just three canonical classes: taxonomic trees, one-dimensional (linear) spaces, and independent feature models. People surely have access to other classes, including higher-dimensional spaces, flat (non-hierarchical) clusterings, and causal networks. We leave it to future work to characterize the full range of structure classes accessible to human cognition. In particular, it is an open question whether this space is small enough to be explicitly enumerated as we do here, or is so large (perhaps infinite or uncountable) that it can be specified only implicitly through some generating mechanism. Future work should also consider the possibility that multiple structures may apply within a single domain.

Given a set of probabilistic models, Bayesian techniques can be used to evaluate which of the models is most likely to have generated some data [7]. Before these techniques can be applied to inferring domain structures, we need to associate each structure class in our hypothesis space with a probabilistic generative model for the features of objects. The next section defines these models, but here we show how Bayesian inference can be used to choose between them.

Let D be an object-feature matrix generated from one of several structure classes. The posterior probability of each class C_i is proportional to the product of the likelihood $p(D|C_i)$ and the prior probability $p(C_i)$. If we assign equal prior probabilities to each class (as we do throughout this paper), the best class is the class that makes the data most likely.

Computing the likelihood $p(D|C_i)$ requires integrating over all structures $\mathcal{S}$ belonging to structure class C_i:

$$p(D|C_i) = \int p(D|\mathcal{S}, C_i)p(\mathcal{S}|C_i)d\mathcal{S}, \qquad (1)$$

Intuitively, this means that a structure class C_i provides a good account of object-feature data D if the data are highly probable under a range of structures $\mathcal{S}$ in class C_i, and if these structures themselves have high prior probability within C_i. The following section explains how the fit of each structure to the data, $p(D|\mathcal{S}, C_i)$, is computed for several structure classes.

We estimate the integral in Equation 1 using stochastic approximations. First we run a Markov chain Monte Carlo simulation to draw a sample of m structures, $\{\mathcal{S}_j\}$, from the distribution $p(\mathcal{S}|D, C_i)$. We then approximate $p(D|C_i)$ by the harmonic mean estimator [7]:

$$p(D|C_i) = \left(\frac{1}{m} \sum_{j=1}^{m} \frac{1}{p(D|\mathcal{S}_j, C_i)} \right)^{-1}. \qquad (2)$$

This estimator does not satisfy a central limit theorem, and can be thrown off by a sample with very low likelihood. Despite its limitations, it is often sufficient to identify a model that is very much better than its competitors. In future work we plan to estimate these integrals more accurately using path sampling [4].

From structures to probabilistic models

We will work with three probabilistic models, each appropriate for a different structure class, and show how to compute the likelihoods $p(D|\mathcal{S}, C_i)$ for structures in each class. For simplicity we assume here that all features are binary, but our framework extends naturally to multi-valued or continuous features.

C_T: Taxonomic trees

Class C_T is the set of taxonomic trees — rooted trees with the objects in D as their leaves. This is a natural representation when the objects are the outcome of an evolutionary process. We restrict ourselves to ultrametric trees — trees where each leaf node is at the same distance from the root.

Assume that each feature is generated by a mutation process over the tree. We formalize the mutation process using a simple biological model [11]. Suppose that a feature F is defined at every point along every branch, not just at the leaf nodes where the data points lie. Imagine F spreading out over the tree from root to leaves — it starts out at the root with some value and could switch values at any point along any branch. Whenever a branch splits, both lower branches inherit the value of F at the point immediately before the split. Figure 1(a) shows one mutation history for a binary feature on a tree with four objects.

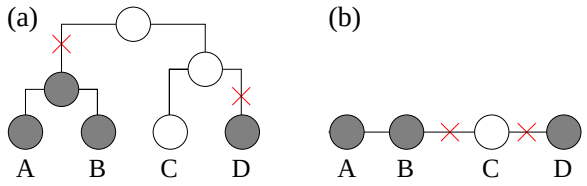

Figure 1: (a) A tree with four objects (A, B, C and D) and three internal nodes. A mutation history for a single feature is shown. The feature is off at the root, but switches on at two places in the tree. Shaded nodes have value 1, clear nodes have value 0, and crosses indicate mutations. (b) A line with four objects.

We formalize this model of mutation using a Poisson arrival process. Under this process, the probability that

F switches values between the beginning and end of any branch b is

$$p(\text{switch along branch } b) = \frac{1 - e^{-2\lambda|b|}}{2}, \qquad (3)$$

where $|b|$ denotes the length of b, and λ is the mutation rate. Note that the mutation process is symmetric: mutations from 0 to 1 are just as likely as mutations in the other direction. Asymmetric mutation processes may be more appropriate in some contexts.

Assume that the features are conditionally independent given the tree (i.e., their mutation histories are independent). We can then compute $p(D|\mathcal{T}, C_T)$, the probability of the data given tree $\mathcal{T}$ by multiplying probabilities for each feature vector taken individually. The necessary calculations can be organized efficiently using a Bayes net with the same topology as $\mathcal{T}$ [9].

Computing the total likelihood $p(D|C_T)$ requires integrating over the space of all trees (including variations in branch length and topology), as in Equation 1. We used the MrBayes [6] program for Bayesian phylogenetic inference to draw a sample of trees $\{T_i\}$ from the distribution $p(\mathcal{T}|D, C_T)$. We then estimated the likelihood $p(D|C_T)$ using the harmonic mean estimator (Equation 2).

C_L: One-dimensional (linear) spaces

Although trees seem appropriate for representing biological species and their properties, other domains will have other kinds of structures. Euclidean spaces figure prominently in mathematical models of similarity comparison, judgment, and choice, and probably should appear in any canonical list of structure classes. Let class C_L indicate the set of one-dimensional linear structures. Extensions to higher dimensions are easy in principle, if computationally more demanding.

A line $\mathcal{L} \in C_L$ is a one-dimensional structure where every node corresponds to an object in the domain. A line is a degenerate tree, but unlike the trees of the previous section, lines have no latent nodes. A four-object line is shown in Figure 1(b).

Features are generated over a line according to the mutation model of the previous section. Imagine that Feature F starts at the leftmost node with some value and spreads to the right with the possibility of switching value at any point. Again, the probability that adjacent nodes separated by a branch of length $|b|$ have different values of F is $\frac{1-e^{-2\lambda|b|}}{2}$.

As with C_T, we estimate the likelihood $p(D|C_L)$ with an approximate (MCMC) sum over all linear structures.

C_0: Independent Features

Class C_0 is similar to a null hypothesis. Unlike the previous models, it assumes no underlying relationships between objects in the domain. Each feature is distributed over objects independently of all other features. The pattern of overlap in feature extensions is thus completely unconstrained. More formally, C_0 assumes that feature vectors (columns of D) are generated by flipping weighted coins. Unlike the previous two cases, the likelihood $p(D|C_0)$ can be computed analytically. Suppose that θ_i is the weight of the coin for feature i, and our prior on θ_i is $\theta_i \sim \text{Beta}(\alpha, \beta)$ (for each of our experiments we use $\alpha = \beta = 1$). If column i of matrix D contains m_i ones and n_i zeros, it can be shown that $p(D|C_0) = \prod_i B(m_i + \alpha, n_i + \beta)/B(\alpha, \beta)$, where $B(\cdot, \cdot)$ is the beta function.

Model complexity and Occam's razor

The three models C_T, C_L, and C_0 vary significantly in their complexity. Both the tree model C_T and the linear model C_L include the independent feature model C_0 as a special case: when each object in C_T or C_L is a long way from its neighbors, feature values at adjacent object nodes are generated in effect by tosses of a fair coin. C_T is also more complex than C_L: in a domain with n objects, there are roughly 2^n more distinct tree structures than distinct linear structures, and the mutation process operating over each tree involves roughly twice as many potential mutation events.

A key feature of Bayesian model selection is that it automatically penalizes unnecessarily complex structures. Some form of Occam's razor is essential when comparing candidate domain structures of different complexities, where the more complex structure (e.g., trees) can more easily mimic the simpler structure (e.g., linear orders) than vice versa. A more naive approach to structure learning, such as choosing the structure that accounts for the most variance in the object-feature matrix D, would be biased against choosing the simpler model class, even when it really generated the observed data.

Empirical tests of structure learning

Synthetic Data

We created three synthetic datasets (unconstrained, tree-structured and linear) with 16 objects and 120 features each. The unconstrained set was constructed using model C_0. The tree-structured set was built by running the mutation process of C_T over a balanced tree with 16 leaf nodes. The linear set was built similarly by running the mutation process over a line with 16 nodes.

Table 1 shows log likelihoods computed for each dataset and structure class. The first row shows that the linear model C_L is better than the tree model C_T on the unconstrained data, but that both are worse than the independent features model C_0. Similarly, the linear model is preferred for the synthetic linear data. The results for the synthetic tree data are more interesting. Even though the data were generated over a tree, the structure class of choice is C_L.

To see why a linear order is a good hypothesis when a tree-structured domain is first encountered, imagine a picture of the true tree, then remove all the branches and internal nodes, leaving behind only the leaves in some linear order. Now join each leaf node to its immediate neighbors. This linear order is a better hypothesis than the true tree at first. The linear model C_L is simpler than the tree model C_T, and if the mutation rate is small, most concepts generated over the tree will be connected subsets of the linear order. Only as more features

Data	C_0	C_L	C_T
Synthetic Unconstrained	59	31	0
Synthetic Linear	0	544	300
Synthetic Tree	0	210	168
Biology	0	230	339
Political	0	1312	883

Table 1: Scaled log-likelihoods for three synthetic and two real-world datasets. Each row has been scaled additively so that its smallest entry is zero.

accumulate should a rational learner conclude that the extra complexity of a tree-structured model is necessary.

To confirm that the true domain structure will eventually win out, we generated a tree-structured set with 32 objects and 240 features and computed log likelihoods as more and more features were observed. Figure 2 shows that the linear structure is preferred while the number of observed features is small, but that the correct tree structure dominates in the end. This transition suggests that our Bayesian model may offer some insight into the dynamics of development. Piaget and others have argued that children move from simple to relatively complex conceptual structures as they mature. Our model shows an analogous shift in tree-structured domains.

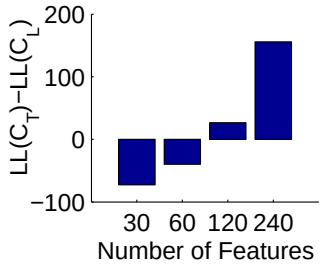

Figure 2: Differences between the log likelihoods of trees (C_T) and linear structures (C_L) on synthetic tree-structured data. Linear structures are preferred at first but the true structure becomes clear as more features are seen.

Biological and Political Data

We used our framework to infer the structure of a biological data set (expected to be tree-structured), and a political data set (expected to be linear). The biological set was constructed from human feature judgments collected by Osherson et al. (1991). Subjects were given 48 animals and 85 features (eg 'lives in water', 'has a tail') and asked to rate the "relative strength of association" between each animal and feature. Subjects gave ratings on a scale that started at zero and had no upper bound. Ratings were linearly transformed to values between 0 and 100, then averaged. We created a binary dataset by thresholding all values at the global mean.

The political dataset was taken from the Supreme Court database collected by Harold Spaeth (1998). We looked at the Burger court which served from 1981 to 1985. Spaeth records 8 possible types of voting behavior: we considered only the cases where every judge either joined the majority, dissented, or cast a regular concurrence (which we treated the same as a majority vote). This left a binary dataset containing votes for 9 judges on 637 cases.

Of the three classes in our hypothesis space, Table 1 confirms that trees provide the best account of the biological data and linear structures are best for the voting data. Note that a more naive approach to structure learning fails here. An additive tree model accounts for more of the variance of the Supreme Court data than a one-dimensional metric scaling solution. Choosing the model that accounts for the greatest proportion of the variance incorrectly favors trees, since it ignores the greater complexity of the tree model.

Once the structure class is known, we can identify the member of that class that makes the data most likely. For the animal data, we took our MCMC sample from the posterior over tree structures, and identified the most representative tree using the `consense` program in the PHYLIP package [3]. The resulting tree is shown in Figure 3(a). Similarly, the best linear structure for the Supreme Court data is shown in Figure 3(b).

The ultimate reason why trees are appropriate for biological data is that evolution is a branching process. It is harder to say a priori why the voting data should be one-dimensional, but the political spectrum ("left"–"right") is an extremely common notion, and others have analyzed Supreme Court data and found that the first dimension of a multidimensional linear model explains almost all of the variance [15]. Our results may explain in part why people represent these domains as they do, but the analysis is mute with respect to the precise mechanisms that give rise to these cognitive structures. Multiple learning mechanisms probably operate in both these domains. Likely mechanisms include inferences drawn from feature observations, as modeled explicitly by our Bayesian learning algorithm, as well as cultural transmission of knowledge, which surely occurs for structures like the "left"–"right" metaphor.

Structure learning versus empiricism

The conventional empiricist critique of structured domain representations has three lines of attack, well articulated recently by McClelland and Rogers [12]: (1) structured representations such as taxonomic trees are too rigid to deal naturally with exceptions or gradients of typicality; (2) it is not clear how structured representations can be induced from raw data; (3) unstructured associative learning architectures can match all of the supposed advantages that structured representations claim. Our work challenges all of these critiques. Previously [10], we showed that robustness to exceptions and sensitivity to typicality fall out naturally from defining a probabilistic generative model of object features in terms of a mutation process over a taxonomic tree (or other domain structure). Point (2) was addressed in the previous section, and now we turn to point (3). We show that learning explicitly structured domain representations provides a powerful source of inductive bias for reasoning about novel properties, and that this power is not easily matched by a generic connectionist architecture.

We compared our tree-structured model for the

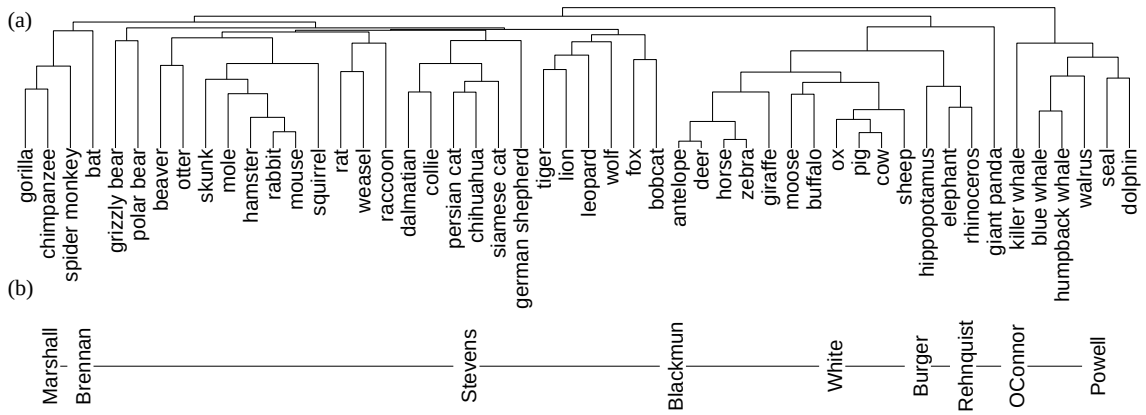

Figure 3: Structures (found via Bayesian structure learning) that best characterize two domains: (a) Mammal species and their properties, and (b) Supreme Court Judges and their decisions.

animal-feature data described above[1] with a connectionist model inspired by the work of McClelland and Rogers [12]. The network includes one input unit for each animal species and one output unit for each feature. We explored a wide range of network parameters in an attempt to achieve the best possible performance (see below). Following McClelland and Rogers, we trained each network on the full matrix D of object-feature associations, then tested how well the hidden-unit representations supported inductive projections for novel features.

In the inductive projection task, a new feature is introduced, and one or more examples of species with that feature are provided to the learner. The learner's task is to infer which other species have this novel property. Like Rogers and McClelland, we modeled this task by introducing a new output unit for the novel feature, freezing all weights except those connected to the new unit, and training the new unit's weights until it reliably produced the correct feature values for the given examples. We then tested the new unit's output when other species were presented as inputs.

We modeled this same induction task using our tree-based Bayesian framework, as described in [10]. Given a tree $\mathcal{T}$ inferred for the domain, the mutation process in model C_T induces a prior distribution over all possible labellings of the species (i.e., the leaves of $\mathcal{T}$). Given one or more examples of a novel property, this prior together with the machinery of Bayesian concept learning allows us to infer the most likely value of that property for all other species in the tree [10]. We used the tree shown in Figure 3(a), and set the mutation rate for the novel property to the value that best fit the 85 features in the biological data set. The resulting tree-based model has no free parameters.

The inductive projections of each model were compared with human argument ratings collected by Osherson et al. [13]. Osherson used a ten-animal domain: horse, cow, chimp, gorilla, mouse, squirrel, dolphin, seal and rhino. The specific set contains 36 two-example arguments, and the conclusion species is always "horse". The general set contains 45 three-example arguments, and the conclusion category is "all mammals." Unfamiliar (blank) predicates – e.g., "have biotinic acid in their blood" – were used for all these arguments. The tree-based Bayesian model rates the strength of general arguments by computing the probability that all ten animals in the domain have the property. The connectionist model rates general arguments by computing projections to each animal separately and adding these ten scores.

Table 2 shows correlations between model predictions and human judgments of argument strength. The first column summarizes the performance of two separate neural networks, reflecting the best performance we ever observed on each data set over a thorough two-stage exploration of the space of possible networks[2]. In the first stage, we tested many different network topologies and varied the learning rate, the number of training and testing epochs, and the presence or absence of momentum and bias. We then took the best-performing networks from the first stage and ran every possible combination of the two best architectures, three best learning rates, two best numbers of testing epochs, and three best numbers of training epochs. The best networks were trained for 20,000 epochs, tested after 250 epochs of training on each testing example, and had no momentum and a bias of -2. They had two hidden layers, typically with 10-30 units each, and a learning rate between 0.005 and 0.01. Even allowing different neural networks for the two datasets, we were unable to match the performance of the tree-based Bayesian model.

Our model differs from these connectionist models along at least two important dimensions, either or both of which could account for its superior performance. First, it uses explicit taxonomic structure and second, it uses Bayesian statistical inference. To isolate the ef-

[1] In order to model the behavioral judgments described below, we supplemented these data with feature ratings for two additional species, cow and dolphin, to give a total of 50 species.

[2] The majority of these tests were conducted with the original 48-animal feature ratings (substituting ox for cow and blue whale for dolphin), before we collected feature ratings for cow and dolphin. Qualitatively similar results were observed with the 50-animal dataset. The results reported in Table 2 reflect the best performance observed across either dataset.

	NN	NN (T)	Bayes (U)	Tree-Bayes	Sim Cov.
Specific	0.62	0.86	0.16	0.95	0.75
General	0.41	0.68	0.38	0.91	0.77

Table 2: Correlations between human judgments and five models for the specific (row 1) and general (row 2) inductive projection tasks described in the text.

fect of structure we implemented models that incorporate only one of these factors. NN(T) is a neural network that uses an explicitly taxonomic representation but not Bayesian inference. The network has 19 input units and a single output unit for the novel property. Input features are derived from the ten-animal tree — the subtree of Figure 3 that includes the ten animals used in this task. Each input node corresponds to a node in the tree, and a species is represented by switching on an input unit for each of its parent nodes in the tree (including a distinctive feature for itself). Species that appear nearby in the tree will share a relatively large number of ancestors and will therefore have similar representations. Bayes(U) is a model that uses Bayesian inference but without any explicit structural representation constraining hypotheses. The model is inspired by Heit's (1998) suggestion that priors for Bayesian induction could be derived from familiar features stored in memory [5]. Each of the 85 observed feature vectors is identified with a candidate hypothesis for generalization, e.g., the feature "nocturnal" gives rise to the hypothesis that the new property is true of all and only the nocturnal species. We assigned a prior probability of $\frac{1}{86}$ to each of these hypotheses and reserved a further $\frac{1}{86}$ for the hypothesis including all mammals.

Table 2 shows that NN(T) performed better than all of the networks explored previously. The tree-based Bayesian model performed better than Bayes(U) or a feature-based version of Osherson et al.'s (1990) similarity-coverage model (which also assumes no domain structure). These results suggest that generic approaches to biological induction may be improved by adding explicit representations of taxonomic structure. The tree-based Bayesian approach also performed better than the tree-based neural network, suggesting that both rational statistical inference and structured domain representations play important roles in guiding people's generalizations.

Conclusion

Our results are preliminary, with a focus on the domain of biology and just the taxonomic aspect of knowledge in that domain. No strong general claims can be made until we push this inquiry more deeply in the domain of biology, and more broadly into other domains. Even so, our work suggests a viable alternative to traditional nativist and empiricist accounts of domain knowledge. Contrary to a strong nativist view, the organizing structural principles of a domain may be learned. Contrary to a strong empiricist view, explicit representations of domain structure may be valuable for guiding inductive projections from sparse data. Structured domain representations and domain-general statistical learning thus need not exclude each other, and indeed are complementary. Statistical learning suggests how novel domain structures can be acquired, and these structures provide a powerful inductive bias for future statistical learning.

Acknowledgments We thank S. Sloman for providing the biological dataset (originally collected by Osherson and Wilkie), Doug Rohde for his neural net package, and Tom Griffiths and Sean Stromsten for helpful suggestions. Supported by NTT Communication Sciences Lab, the DARPA CALO program, and the Paul E. Newton Chair (JBT).

References

[1] S. Atran. Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, 21:547–609, 1998.

[2] A. Collins and M. R. Quillian. Retrieval time from semantic memory. *Jn of Verbal Learning and Verbal Behavior*, 8:240–247, 1969.

[3] J. Felsenstein. PHYLIP – Phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

[4] A. Gelman and X. L. Meng. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13:163–185, 1998.

[5] E. Heit. A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford and N. Chater, editors, *Rational models of cognition*, pages 248–274. Oxford University Press, New York NY, 1998.

[6] J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.

[7] R. E. Kass and A. E. Raftery. Bayes factors. Technical Report 254, University of Washington, 1993. Revision 3: July 6, 1994.

[8] F. Keil. *Semantic and Conceptual Development*. Harvard University Press, 1979.

[9] C. Kemp, T. L. Griffiths, S. Stromsten, and J. B. Tenenbaum. Semi-supervised learning with trees. In *Advances in Neural Information Processing Systems*, 2003.

[10] C. Kemp and J. B. Tenenbaum. Theory-based induction. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, 2003.

[11] P. O. Lewis. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 19(6):913–925, 2001.

[12] J. McClelland and T. Rogers. The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, 4:310–322, 2003.

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

[14] D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith. Default probability. *Cognitive Science*, 15:251–269, 1991.

[15] L. Sirovich. A pattern analysis of the second Rehnquist U.S. Supreme Court. *PNAS*, 100:7432–7437, 2003.

[16] H. J. Spaeth. United States Supreme Court judicial database, 1953-1996 terms. 1998. 8th ICPSR version.