

# Theory-Based Induction

Charles Kemp (ckemp@mit.edu)  
Joshua B. Tenenbaum (jbt@mit.edu)  
Department of Brain and Cognitive Sciences  
Massachusetts Institute of Technology  
Cambridge, MA 02139-4307 USA

## Abstract

We show how an abstract domain theory may be incorporated into a rational statistical model of induction, giving a more principled basis for the model's assumptions, greater explanatory power, and closer contact with processing-level models. In particular, we develop a Bayesian model of category-based induction that is founded on a theory of how biological properties are distributed across classes of biological kinds. We argue that our new approach is much more principled than previous Bayesian and similarity-based models of inductive inference about animal categories. We also show that our new approach performs better than these previous models over a collection of four data sets, one of which is new.

Philosophers since Hume have struggled with the logical problem of induction, but children solve an even more difficult task — the practical problem of induction. Children somehow manage to learn concepts, categories, and word meanings, and all on the basis of a set of examples that seems hopelessly inadequate. The practical problem of induction does not disappear with adolescence: adults face it every day whenever they make any attempt to predict an uncertain outcome. Inductive inference is a fundamental part of everyday life, and a fundamental phenomenon in need of a psychological explanation.

Two important questions can be asked about inductive generalization: what resources does a subject bring to an inductive task, and how are these resources combined to generate a response to the demands of the task? In other words, what is the process of induction, and what is the prior knowledge required by this process? Psychologists have considered both of these questions in depth, but previous computational models of induction have tended to emphasize process to the exclusion of prior knowledge. This paper attempts to redress this imbalance by showing how a rich body of prior knowledge can be included in a computational model founded on rational statistical inference.

The importance of prior knowledge has been attested by psychologists and machine learning theorists alike. Murphy and Medin (1985) have suggested that the acquisition of new concepts is guided

by “theories” — networks of explanatory connections between existing concepts. Machine learning theorists have built formal models of learning, and argued that generalization within these models is not possible unless a learner begins with some sort of inductive bias (Mitchell, 1997). The challenge that inspires our work is to develop a model with an inductive bias that is well motivated by a theory of the domain under consideration.

Many previous models have taken similarity judgments as their representation of prior knowledge (Nosofsky, 1986; Osherson et al., 1990). This approach has been dominant within the tradition of category-based induction, and Osherson's (1990) similarity-coverage model will be one standard against which our new model will be compared. Using similarity data to represent prior knowledge is a reasonable first attempt, but similarity judgments are less than ideal as a starting point for a model of inductive inference. As Goodman (1972) has pointed out, similarity is a vague and elusive notion. It is meaningless to say that two objects are similar unless a respect for similarity has been specified. Any model based on similarity alone is therefore a model without a secure foundation.

Another way of formulating this objection is to point out that similarity judgments do not specify the rich sort of theory we are looking for. Far from it: in fact, they probably stand in need of explanation by some theory (Murphy and Medin, 1985). If so, this theory should form the most natural starting point for a model of inductive inference. The difficult part, of course, is formalizing the theory. Computational models of theories are relatively rare, but this paper attempts to set out one case in which a theory can be formalized.

The tasks considered are the same tasks that inspired Osherson's similarity-coverage model. In the first task (the specific inference task), subjects are asked to rate the strength of arguments of the following form:

Horses can get bleminitis

Cows can get bleminitis

---

Dolphins can get bleminitis

The premises state that one or more specific mammals can catch a certain disease, and the conclusion

is that another specific species (here dolphins) can also catch the disease.

In the second task (the general inference task), subjects are asked to consider a generalization from specific premises to a property of all mammals. For instance:

Seals can get bleminitis
Dolphins can get bleminitis
All mammals can get bleminitis

The theory guiding performance on these tasks is presumably a theory of how biological properties are distributed over classes of biological species. We will specify a simple theory of this sort and use it as the starting point for a computational model.

Other researchers have already specified some aspects of the intuitive theory of biology that is relevant for inductive inference in this domain. Atran (1995) argues that biological kinds are organized in a folk taxonomy, and that this structure ‘supports the widest possible range of inductions about living kinds.’ As we show, this taxonomy turns out to be crucial, but only as a first step. The taxonomic principle must be augmented with a new “distributional principle” specifying how properties are probabilistically distributed over the taxonomy. Intriguingly, both the taxonomic principle and the distributional principle closely resemble analogous theories in evolutionary biology and genetics.

## Previous Models

### Similarity-Based models

Osherson’s similarity-coverage model expresses the strength of an argument as a linear combination of two components: a term representing the similarity between the premises and the conclusion, and a term representing the extent to which the premises cover the lowest level taxonomic category including both premises and conclusion.

Formalizing these ideas, the strength of the argument from a set of premises  $X$  to a conclusion category  $Y$  is :

$$\alpha \text{ setsim}(X, Y) + (1 - \alpha) \text{ setsim}(X, [X; Y])$$

where  $\alpha$  is a free parameter,  $\text{setsim}(\cdot)$  is a setwise similarity metric, and  $[X; Y]$  is the lowest level taxonomic category including  $X$  and  $Y$ .

Several setwise similarity metrics might be tried. Osherson et al. propose  $\text{maxsim}(\cdot)$  but also consider  $\text{sumsim}(\cdot)$

$$\text{maxsim}(X, Y) = \sum_j \max_i \text{sim}(X_i, Y_j)$$

$$\text{sumsim}(X, Y) = \sum_j \sum_i \text{sim}(X_i, Y_j)$$

Both are defined in terms of  $\text{sim}(\cdot)$ , the standard pairwise similarity metric.

The lack of a principled reason for choosing between these metrics is a limitation of the similarity based approach. Osherson et al. suggest that  $\text{maxsim}(\cdot)$  conforms best to their intuitions about coverage, yet  $\text{sumsim}(\cdot)$  is more standard in models of inductive learning. It turns out that  $\text{maxsim}(\cdot)$  leads to much better performance on Osherson’s general task than  $\text{sumsim}(\cdot)$ , and we will consider later why this might be the case.

### Bayesian Models

Heit (1998) and Sanjana and Tenenbaum (2003) have laid out a Bayesian approach to category based induction. The Bayesian approach seems uniquely well suited for including a substantial body of prior knowledge, and our new model will be developed within this paradigm.

Assume that we begin with a finite domain  $D$ , a fixed set of objects that will provide the context for our inductive inferences. For the tasks modeled here, our domain is a set of ten mammals. We are interested in a concept,  $C$ , that picks out some subset of these objects. Let  $H$  be the power set of  $D$ .  $H$ , our hypothesis space, is therefore the set of all possible concepts over our domain. To each hypothesis  $h$  in  $H$  we assign a prior probability  $p(h)$ , where  $p(h)$  is the probability that  $h$  is the concept  $C$  we are interested in.

The specific inference task may now be formalized as follows. We observe  $X$ , a set of  $n$  members of the concept  $C$ , and want to compute  $p(y \in C|X)$ , the probability that another object,  $y$ , is also a member of  $C$ . This probability will be the sum of the posterior probabilities of every concept that includes  $y$ .

$$\begin{aligned} p(y \in C|X) &= \sum_{h \in H: y \in h} p(h|X) \\ &= \sum_{h \in H: y \in h} \frac{p(X|h)p(h)}{p(X)} \end{aligned}$$

The denominator may be computed by summing over all hypotheses in  $H$ :

$$p(X) = \sum_{h \in H} p(X|h)p(h)$$

The likelihood of  $X$  given  $h$  can be computed by assuming that the  $n$  examples in  $X$  are sampled independently at random from  $h$ :

$$p(X|h) = \begin{cases} \frac{1}{|h|}^n, & \text{if all } n \text{ examples in } X \text{ belong to } h \\ 0, & \text{otherwise} \end{cases}$$

Osherson’s general inference task is formulated similarly. The probability that all of the members of concept  $Y$  belong to  $C$  is:

$$p(Y \subset C|X) = \sum_{h \in H: Y \subset h} p(h|X)$$

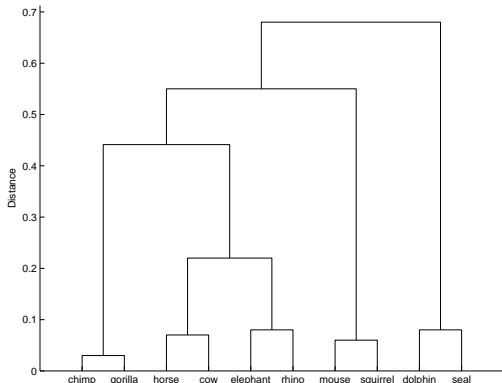


Figure 1: A taxonomy built from similarity data

The only element still missing from the Bayesian framework is a specification of how the prior probabilities  $p(h)$  are calculated. Heit (1998) does not address this question, and Sanjana and Tenenbaum (2003) use a prior distribution that is not deeply motivated by a theory of the domain. We will describe our new model then indicate where it differs from its predecessor.

## A Theory Based model

The prior distribution for our Bayesian model is motivated by two principles: the ‘taxonomic principle’ and the ‘distributional principle.’ Together, these principles form a theory of the distribution of biological properties.

The taxonomic principle holds that animals naturally fall into a taxonomy, or a series of hierarchical groupings. This belief appears to be universal – a substantial body of work has documented that cultures all over the world organize living kinds into ‘folk taxonomies’ (Atran, 1995). It is also scientific, since the theory of evolution implies that living kinds should form a taxonomy.

A sensible first step towards generating a prior distribution was therefore to build a folk taxonomy for the ten mammals in our domain. This taxonomy we used is shown in Figure 1, and was generated by running average-link clustering on the similarity data collected by Osherson.

A word of explanation is necessary here. We made some critical remarks about similarity based models in the first section, but now we are using similarity data to build the prior distribution for our model. Our approach, however, does not grant the similarity judgments privileged status: we simply use them to try to reconstruct the taxonomy that may have generated these data. It is the taxonomy that is important, and our approach does not depend critically on similarity since there are ways to generate the taxonomy that do not involve similarity judgments. One might try, for example, to cluster the animals on the basis of behavioral and morphological features. Building the folk taxonomy out of similarity data,

however, offers an important advantage: it allows the performance of our model to be directly compared with the performance of the similarity-based models.

With this taxonomy in hand, a simple prior can be generated. There are 19 nodes in the taxonomy, and each specifies a concept that includes the animals falling beneath it on the tree. A straightforward way to set the prior is to assign a uniform probability to each of these concepts, and zero probability to all other possible concepts. We call the model that uses this prior the ‘Taxonomic Bayesian model.’

It is clear that the taxonomic principle alone is not enough. Compare the argument “seals and squirrels catch bleminitis, so horses are also susceptible” with “seals and cows catch bleminitis, so horses are also susceptible”. The second is stronger than the first, yet the Taxonomic model assigns them both the same rating, since each set of premises is compatible with only one hypothesis, the set of all mammals. Figure 2 confirms that this model predicts human ratings of argument strength rather poorly.

The distributional principle, the second part of our theory, states that category membership depends on a process of random mutation operating over the taxonomy. This principle acknowledges that convergent evolution can occur: that two animals can share a property even if they have no common ancestor with that property. Some additional notation is needed to make this principle precise.

Suppose that we are interested in some category  $C$ . Given any subset of the ten animals in Figure 1, we want to know the probability that that subset is the extension of  $C$ . Imagine that membership of  $C$  depends on a single feature  $F$  that could have evolved at any point in the tree and may have evolved independently along different branches of the tree. Once  $F$  has arisen along a branch, all nodes falling below that branch are members of  $C$ .

We model the evolution of  $F$  as a Poisson arrival process with a parameter,  $\lambda$ , that will be called the mutation rate. The probability that the feature develops along a branch  $b$  of length  $|b|$  is:

$$p(F \text{ develops along } b) = 1 - e^{-\lambda|b|}$$

A branch is ‘marked’ if  $F$  develops along that branch.

To obtain a single estimate of the extension of  $C$ , we consider all branches in the tree, label each as marked or unmarked according to the formula above, then collect all external nodes that fall beneath a marked branch. Repeating this many times generates a prior distribution over all categories, where the prior probability of any category is the proportion of times it was chosen as our estimate of  $C$ .

The prior distribution may be computed analytically instead. First consider all single branches in the tree. For each branch, calculate the probability

that  $F$  arises along that branch and nowhere else in the tree, and add this probability to the prior for the corresponding concept. Continue by considering all pairs of branches, all triples, and so on.

Our model of the evolution of  $F$  captures two important intuitions. First,  $F$  is more likely to develop along the longer branches of the tree. Second, since  $F$  develops independently along different branches and the probability of arising on any branch is small, the model favors simpler hypotheses — hypotheses consisting of fewer rather than more clusters.

The prior for the Sanjana-Tenenbaum (or ‘Disjunctive Bayes’) model is also computed by taking disjunctions of the 19 hypotheses represented by the folk taxonomy. The 19 original hypotheses are assigned the highest prior probability, disjunctions of two of these are assigned a somewhat smaller probability, and disjunctions of three hypotheses are assigned a still smaller probability. This approach represents a general strategy for expanding an hypothesis space, and can be applied to hypothesis spaces that have nothing at all to do with taxonomies. Generality, however, is bought at a price: unlike our new ‘Evolutionary model’, the Disjunctive model is not deeply related to a theory of our domain. A symptom of this lack of principled motivation is that Disjunctive Bayes does not take the branch lengths of the taxonomic tree into account.

### Performance of the Model

Figure 2 shows correlations between the ranks assigned to a set of arguments by humans and the ranks predicted by the models. The first three columns show the performance of the models on the three data sets used in previous studies. The Osherson general set contains 45 three-premise general arguments. The Osherson specific set contains 36 two-premise arguments, and the conclusion category in all cases is ‘Horses’. The Sanjana set contains 28 specific arguments with one, two or three premises, and again the conclusion category is always ‘Horses.’

All of the models (except Taxonomic Bayes) include a single free parameter, and each correlation in Figure 2 is for the setting of the parameter that best fits the human data. The first three columns show that there is little to separate the Sanjana-Tenenbaum and the Evolutionary models, but that both outperform the similarity-based models by a small margin. Both of these Bayesian models show robust performance across a range of parameter settings, and both admit a single setting that achieves correlations exceeding 0.9 on the first three data sets.

### A New Experiment

A limitation of the Sanjana-Tenenbaum model is that it does not capture at least one of the phenomena documented by Osherson et al. Premise typicality says that general arguments increase in strength

as the premises become more typical of the conclusion category. Thus:

$$\frac{\text{Horses can get bleminitis}}{\text{All mammals can get bleminitis}}$$

is a stronger argument than:

$$\frac{\text{Dolphins can get bleminitis}}{\text{All mammals can get bleminitis}}$$

because horses are more typical mammals than dolphins.

Although our new model was not built with premise-typicality in mind, we collected new data which show that it captures this effect more successfully than the Sanjana-Tenenbaum model. Ten single-premise general arguments (one for each species in our domain) were printed on a set of cards, and 25 subjects were asked to sort these cards in order of increasing argument strength. The average rank of each argument was calculated, and these ranks compared with the ranks assigned by the models.

Since we were limited to just ten arguments by the size of our domain, the correlations achieved are much lower than for the previous three data sets. It is nonetheless clear that Maxsim and the Evolutionary Bayes model partially capture the premise typicality effect, but that the predictions of the Disjunctive model are negatively correlated with the human rankings.

### Evolutionary Bayes and Maxsim

It is interesting that Sumsim does significantly worse than Maxsim on the Osherson general set. This result confirms the intuition of Osherson et al. that maxsim(·) is the better metric for category-based induction, but the superiority of Maxsim still demands a principled explanation.

Heit (1998) and Sanjana and Tenenbaum (2003) have suggested that a Bayesian model might explain the success of other approaches to category-based induction. We noticed that Maxsim and Evolutionary Bayes performed similarly on all four data sets, and wondered if Maxsim might be a simple approximation of Evolutionary Bayes. If so, then our evolutionary model (which is based on rational statistical inference) might explain the success of Maxsim.

To further explore the relationship between these two models, we ran a simulation using a set of 100 randomly generated taxonomies. Each taxonomy was generated by starting with a set of 10 nodes, and merging pairs of nodes at random until only one remained. The branch lengths were generated by choosing 9 random numbers between 0 and 1, and setting the height of the node created by the  $k$ th merge to the  $k$ th smallest of these numbers. To calculate the predictions of the similarity models, the similarity of two objects was defined to be one minus the length of the path joining the objects in the tree. This makes sense under the assumption

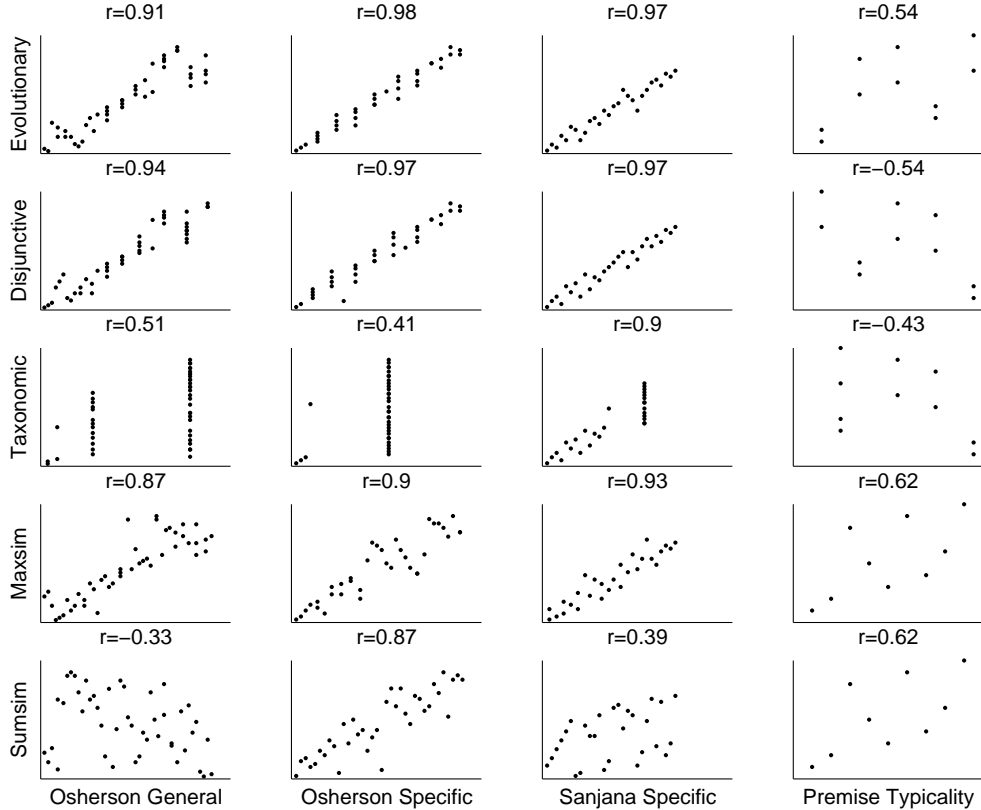


Figure 2: Model predictions (x axis) vs human judgments (y axis). Each row shows the performance of a model, and each column shows the performance over a data set. Every model (except Taxonomic Bayes) includes a single free parameter, and the plots shown are for the best setting of that parameter.

that the tree is an approximation of the structure used to generate the similarity judgments.

Figure 3 shows our results. The pair of models that matched most closely on these general arguments was Maxsim and Evolutionary Bayes, even though Evolutionary Bayes is superficially much more similar to Disjunctive Bayes than Maxsim. This result supports the idea that Maxsim and Evolutionary Bayes may specify about the same mapping between their input (the similarity matrix) and their output (ratings of argument strength).

If this claim turns out to be true, critics of our approach might ask why we have spent time developing a new model if it cannot outperform Maxsim by a substantial margin. Framing the question in this way is misleading. It implies that the models are direct competitors, and that only one can be close to correct. In reality, the two models operate on different levels, and the success of one may only reinforce the value of the other.

Marr (1982) described three broad levels at which a psychological theory may be situated. The Bayesian approach is best suited for the formulation of theories at the highest level: Marr’s level of computational theory. An analysis at this level promises

not to show how people carry out a certain task, but to explain something about why they approach the problem in the way that they do. In contrast, Maxsim falls most naturally into Marr’s second level as an algorithm that might implement the computational theory in a psychologically plausible way.

The similar performance of the models therefore supports the idea that the two are complementary. Evolutionary Bayes helps to show why Maxsim may be a reasonable model of inductive generalization, and Maxsim provides an existence proof that the computations required by the Bayesian model can be approximated by simple heuristics.

## Discussion

The prior distribution used by the Evolutionary model follows directly from the theory consisting of the taxonomic and distributional principles. It is striking that a model inspired by ideas about random mutation and convergent evolution (that is, scientific ideas about how the world actually works) can predict human judgments so well.

It is more difficult to decide whether this theory yields an insight into the strategies people use for inductive inference about animals. The taxo-

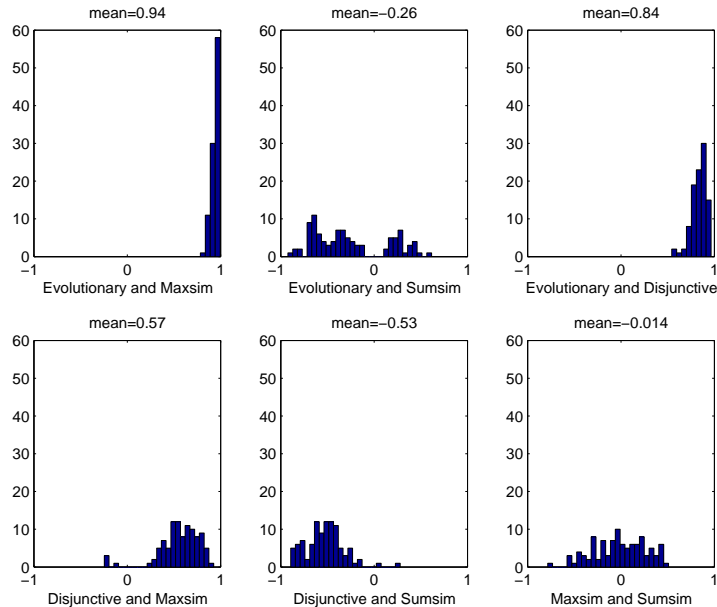


Figure 3: Frequencies (y-axis) against correlations (x-axis). Each histogram shows the distribution of the correlations achieved by a pair of models on the Osherson general set. 100 random trees were used, and free parameters were set separately for each tree and each pairwise comparison.

onomic principle seems universal, but it is debatable whether people subscribe to anything like the distributional principle. Atran (1995) has argued that people are able to ‘quickly apprehend particularly salient aspects of the biological reality of how genotypes and their environments jointly produce phenotypes, without ... [being] aware of the precise causal mechanisms involved’. If so, it is certainly possible that people share something like the distributional principle, even if they could never formalize this principle.

A theory-based approach should not be criticized if it fails to generalize beyond the domain for which it was designed. The main reason for wanting to model a theory is that domain-specific knowledge is likely to be important. Still, we are optimistic that the theory used to build our model will be applicable beyond the domain of animals. It should apply to all living kinds, and more generally to any set of objects that can be represented by a developmental tree. Artifacts are one example of a non-biological domain that may meet this condition. Consider, say, the set of all electronic devices. Any two devices that share a QWERTY keyboard are similar partly because both grew out of a single previous technology.

Our evolution-inspired model may also be useful for machine learning. Many machine learning applications involve data sets with a taxonomic structure. Consider the problem, for example, of classifying content on Yahoo, Amazon, or other taxonomically organized databases. Our evolutionary approach may serve as a useful method for expanding hypothesis spaces like these.

## Conclusion

The accurate predictions achieved by Evolutionary Bayes show that thinking about theories can be a useful way of building computational models of induction. They also reinforce the value of the Bayesian framework, which encourages modelers to make principled assumptions about the prior knowledge needed by a model.

**Acknowledgments** Neville Sanjana contributed code and unpublished results and Liz Baraff ran our experiment.

## References

- Atran, S. (1995). Classifying nature across cultures. In Smith, E. E. and Osherson, D. N., editors, *An Invitation to Cognitive Science*, volume 3. MIT Press.
- Goodman, N. (1972). Seven strictures on similarity. In *Problems and Projects*. Bobs-Merril Co., Indianapolis, Indiana.
- Heit, E. (1998). A bayesian analysis of some forms of inductive reasoning. In Oaksford, M. and Chater, N., editors, *Rational Models of Cognition*, pages 248–274. Oxford University Press.
- Marr, D. (1982). *Vision*. W. H. Freeman.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Murphy, G. L. and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92:289–316.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115:39–57.
- Osherson, D., Smith, E. E., Wilkie, O., López, A., and Shafir, E. (1990). Category-based induction. *Psychological Review*, 97(2):185–200.
- Sanjana, N. E. and Tenenbaum, J. B. (2003). Bayesian models of inductive generalization. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Processing Systems 15*. MIT Press. To appear.