

---

# Bayesian models of inductive generalization

---

Neville E. Sanjana & Joshua B. Tenenbaum  
Department of Brain and Cognitive Sciences  
Massachusetts Institute of Technology  
Cambridge, MA 02139  
{nsanjana, jbt}@mit.edu

## Abstract

We argue that human inductive reasoning — generalizing from examples (or premises) to properties of new instances (or conclusions) — is best explained in a Bayesian framework, rather than by traditional models based on similarity computations. We go beyond previous work on Bayesian concept learning by introducing a general-purpose unsupervised method for constructing flexible hypothesis spaces, and we show how two instances of the Bayesian Occam’s razor — in the priors and the likelihoods — trade off to prevent both under- and over-generalization in these flexible spaces. We analyze two published data sets on inductive reasoning and also present the results of a new, more comprehensive study that we have carried out, all of which indicate a consistent advantage for our Bayesian approach over traditional models.

## 1 Introduction

The problem of inductive reasoning — in particular, how we can generalize after seeing only one or a few specific examples of the concept — has troubled philosophers, psychologists, and computer scientists since the earliest days of their respective disciplines. Computational approaches to inductive generalization range from simple heuristics based on similarity matching to complex rational statistical models [1]. Here we consider where on this continuum human inference lies. Based on two classic data sets from the literature and one more comprehensive data set that we have collected, we will argue that the most principled and descriptive accounts are based on a rational Bayesian learning framework [2]. Our models also confront an issue that has often been sidestepped in previous work on models of concept learning: the origin of the learner’s hypothesis space. We present a simple, unsupervised clustering method for creating hypotheses spaces that, when applied to human similarity judgments and embedded in our Bayesian framework, consistently outperforms the best heuristic matching based on the same similarity data.

We focus on two related inductive generalization tasks introduced by [3], which involve reasoning about the properties of mammals. The first task, generalizing from one or more specific mammals (the *premises*) to a specific mammal (the *conclusion*), goes as follows: Imagine that animals  $A$  and  $B$  are susceptible to disease  $Z$ . How likely is it that animal  $C$  is also susceptible to disease  $Z$ ? For example,  $A$  might be a *gorilla*,  $B$  might be a *chimpanzee*, and  $C$  might be a *seal*. In all studies here,  $Z$  is a blank predicate, a novel disease (e.g., blicketitis) about which subjects know nothing other than the information we

give them. The second task, generalizing from specific mammals to the general category of *all mammals*, is similar: Imagine that animals *A* and *B* are susceptible to disease *Z*. How likely is it that *all mammals* are also susceptible to disease *Z*? These tasks are referred to as *specific* and *general* inference tasks, respectively. Although clearly artificial, they rely on people’s deep knowledge of the relationships between animals while eliminating the effects of specific background information about a particular property (the disease *Z*) that might complicate generalization.

Osherson et al. [3] present data from two experiments in which people were asked to judge the strength of specific or general inductive arguments. One data set contains judgments for 36 specific arguments, each with a different pair of mammals given as examples (premises) but the same test category, *horses*. The other set contains 45 general arguments, each with a different triplet of mammals given as examples and the same test category, *all mammals*. Osherson et al. also published subjects’ judgments of similarity for all 45 pairs of the 10 mammals used in their generalization experiments, which they (and we) use to build our models of generalization. We first describe previous attempts to model these data sets, and then introduce our Bayesian approach and our own follow-up experiment.

## 2 Previous approaches

There have been several attempts to model the data in [3]: the “similarity-coverage” model [3], a feature-based model [4], and a Bayesian model [5]. The two components of Osherson et al.’s model are similarity — the similarity between the premises and the conclusion — and coverage — the similarity between the premises and the lowest-level salient taxonomic category that includes both the premises and the conclusion. The intuitive importance of similarity can be seen in the difference between the following examples (written vertically with a line separating the conclusion from the premises):

Monkeys are susceptible to the disease blicketitis.  
-----  
Gorillas are susceptible to the disease blicketitis.  
  
Squirrels are susceptible to the disease blicketitis.  
-----  
Gorillas are susceptible to the disease blicketitis.

The first argument seems more likely since gorillas are more similar to monkeys than to squirrels. The coverage term balances similarity with a measure of how well the premises cover the conclusion category:

Squirrels are susceptible to the disease blicketitis.  
Monkeys are susceptible to the disease blicketitis.  
Cows are susceptible to the disease blicketitis.  
-----  
All mammals are susceptible to the disease blicketitis.  
  
Squirrels are susceptible to the disease blicketitis.  
Mice are susceptible to the disease blicketitis.  
Rats are susceptible to the disease blicketitis.  
-----  
All mammals are susceptible to the disease blicketitis.

In this example, the premises in the first argument provide better coverage over the conclusion category than those in the second argument. These two examples demonstrate why

both similarity and coverage are important factors in category induction. In the model of Osherson et al., similarity and coverage factors are mixed linearly with a free parameter  $\alpha$ :  $\alpha(R(X, Y)) + (1 - \alpha)(R(X, [X; Y]))$ , where  $R$  is a *setwise* similarity metric used to assess the similarity between the set of examples (premises)  $X$  and the test (conclusion) set  $Y$  and the coverage of  $X$  over the lowest-level taxonomic category that includes both  $X$  and  $Y$ . For any sets  $A$  and  $B$ , they define their measure of the similarity between  $A$  and  $B$  to be the sum of each  $B$  element’s maximal similarity to the  $A$  elements:  $R(A, B) = \sum_j \max_i \text{sim}(A_i, B_j)$ . For the “specific” arguments, where the test set  $Y$  has just one element  $y$ , this model reduces to the maximum similarity of  $y$  to the examples  $X$ . In addition, [3] also considers a sum-similarity model that has more traditionally been used to model human concept learning, where the maximum is replaced by a sum:  $R(X, C) = \sum_j \sum_i \text{sim}(A_i, B_j)$ . They favor the maximum over the sum based on its match to their intuitions, rather than on any a priori or normative consideration.

Sloman [4] developed a feature-based model that encodes the shared features between the premise set and the conclusion set as weights in a neural network. But, even at the optimal setting of its free parameter, the feature-based model performs worse than the max-similarity model on both the specific learning task and the general learning task, so we do not consider it further here.

Heit [5] outlines a Bayesian model that provides qualitative explanations of various inductive reasoning phenomena from [3]. His model does not constrain the learner’s hypothesis space, and it does not embody a generative model of the data, so its predictions depend strictly on well-chosen prior probabilities. Because he does not give a general method for setting these prior probabilities, the model does not make quantitative predictions that can be compared here.

### 3 A Bayesian model

Tenenbaum & colleagues have previously introduced a Bayesian framework for learning concepts from examples, and applied it to learning number concepts [2], word meanings [6], and more elementary domains [7]. Formally, for the specific learning task, we observe  $n$  positive examples  $X = \{x^{(1)}, \dots, x^{(n)}\}$  of the concept  $C$  and want to compute the probability that a particular test stimulus  $y$  belongs to the concept  $C$  given the observed examples  $X$ :  $p(y \in C|X)$ . There is an important difference here between *argument strength* and *generalization probability*. For a particular argument  $y \in C$ , the argument strength is the ratio of the generalization probability,  $p(y \in C|X)$ , to the prior,  $p(y \in C)$ ; it is a measure of how much one’s belief in the conclusion increases given the examples/premises. In all experimental data, the prior is constant — that is, for each experiment, subjects are always asked to generalize to the same argument — and thus only the generalization probability needs to be computed. In future work, we plan to model argument strength using different arguments with different prior probabilities.

These generalization probabilities  $p(y \in C|X)$  are computed by averaging the predictions of a set of hypotheses weighted by their posterior probabilities:

$$p(y \in C|X) = \sum_{h \in \mathcal{H}} p(y \in C|h)p(h|X) = \sum_{h: y \in h} p(h|X). \quad (1)$$

Hypotheses  $h$  pick out subsets of stimuli — candidate extensions of the concept  $C$  — and  $p(y \in C|h)$  is just 1 or 0 depending on whether the test stimulus  $y$  falls under the subset  $h$ . In the general learning task, we are interested in computing the probability that a whole test category  $Y$  falls under the concept  $C$ :

$$p(Y \subset C|X) = \sum_{h: Y \subset h} p(h|X). \quad (2)$$

A crucial component in modeling both tasks is the structure of the learner’s hypothesis space  $\mathcal{H}$ .

### 3.1 Hypothesis space

Elements of the hypothesis space  $\mathcal{H}$  represent natural subsets of the objects in the domain — subsets likely to be the extension of some novel property or concept. Our goal in building up  $\mathcal{H}$  is to capture as many hypotheses as possible that people might be using in concept learning, using a procedure that is ideally automatic and unsupervised. One natural way to begin is to identify hypotheses with the clusters returned by a clustering algorithm [6][8]. Here, hierarchical clustering seems particularly appropriate, as people across cultures appear to organize their biological species concepts in a hierarchical taxonomic structure [9]. We applied four standard agglomerative clustering algorithms [10] (single-link, complete-link, average-link, and centroid) to subjects’ similarity judgments for all pairs of 10 animals given in [3]. All four algorithms produced the same output (Figure 1), suggesting a robust cluster structure. We define our base hypothesis space  $\mathcal{H}_1$  to consist of all clusters in this tree, and we refer to the elements of  $\mathcal{H}_1$  as the “taxonomic hypotheses”.

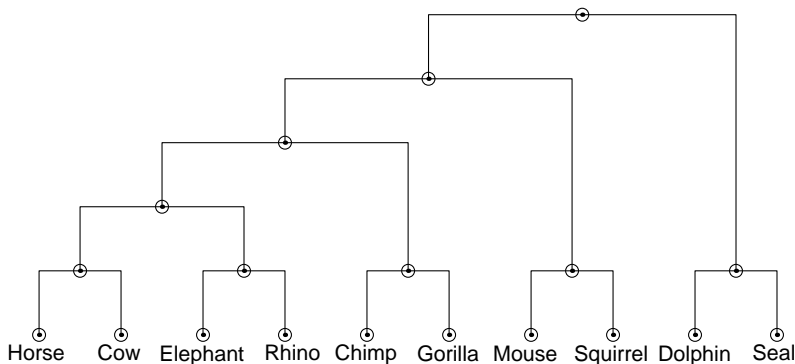


Figure 1: A dendrogram of the 19 hypotheses in the taxonomic hypothesis space given by the clustering algorithms.

It is clear that  $\mathcal{H}_1$  alone is not sufficient. The chance that horses can get a disease given that we know cows and squirrels can get that disease seems much higher than if we know only that chimps and squirrels can get the disease, yet the taxonomic hypotheses consistent with the example sets  $\{cow, squirrel\}$  and  $\{chimp, squirrel\}$  are the same. Bayesian generalization with a purely taxonomic hypothesis space essentially depends only on the *least* similar example (here, *squirrel*), ignoring more fine grained similarity structure, such as that one example in the set  $\{cow, squirrel\}$  is very similar to the target *horse* even if the other is not. This sense of fine-grained similarity has a clear objective basis in biology, because a single property can apply to more than one taxonomic cluster, either by chance or through convergent evolution. If the disease in question could afflict two distinct clusters of animals, one exemplified by cows and the other by squirrels, then it is much more likely also to afflict horses (since they share most taxonomic hypotheses with cows) than if the disease afflicted two distinct clusters exemplified by chimps and squirrels. Thus we consider richer hypothesis subspaces  $\mathcal{H}_2$ , consisting of all pairs of taxonomic clusters (i.e., all unions of two clusters from Figure 1, except those already included in  $\mathcal{H}_1$ ), and  $\mathcal{H}_3$ , consisting of all triples of taxonomic clusters (except those included in lower layers). We stop with  $\mathcal{H}_3$  because we have no behavioral data beyond three examples. Our total hypothesis space is then the union of these three layers,  $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \mathcal{H}_3$ . The assumption that concepts to be learned correspond to unions of one or more clusters is broadly applicable, well beyond the

domain of biological properties. It is analogous to other general-purpose representations for concepts, such as disjunctive normal form (DNF) in PAC-Learning, or class-conditional mixture models in density-based classification [1].

### 3.2 Balancing Occam’s razor: priors and likelihoods

Given this hypothesis space, Bayesian generalization then requires assigning a prior  $p(h)$  and likelihood  $p(X|h)$  for each hypothesis  $h \in \mathcal{H}$ . For simplicity, we assign  $p(h)$  in two stages. First we assign some weight  $p(\mathcal{H}_i)$  to the  $i$ th layer  $\mathcal{H}_i$  (the unions of  $i$  taxonomic clusters), and then we divide that weight uniformly over all  $h \in \mathcal{H}_i$ :  $p(h) = p(\mathcal{H}_i)/|\mathcal{H}_i|$ , where  $|\mathcal{H}_i|$  denotes the number of hypotheses in layer  $i$ . This prior effectively penalizes more complex hypotheses (consisting of more disjoint clusters), because it is inversely proportional to the number of hypotheses in a given layer, and more complex layers contain more hypotheses. This is one instance of the Bayesian Occam’s razor. To keep free parameters to a minimum, we take  $p(\mathcal{H}_i) \propto \lambda^i$ , for some value of  $\lambda$ . To ensure that more complex hypotheses always have lower prior probability, we restrict  $\lambda$  to the interval  $[0, 1]$ .

The likelihood,  $p(X|h)$ , is calculated by assuming that the examples in  $X$  are randomly chosen examples of the concept to be learned. Thus, as in [2], we use the *size principle* to determine the likelihood of a specific hypothesis  $h$  when  $X$  consists of  $n$  examples of the concept:

$$p(X|h) = \begin{cases} [\frac{1}{|h|}]^n & \text{if } h \text{ includes those } n \text{ examples} \\ 0 & \text{if } h \text{ does not include all of the } n \text{ examples in } X \end{cases} \quad (3)$$

We take the size  $|h|$  of hypothesis  $h$  to be simply the number of animal types it contains. The size principle thus assigns a greater likelihood to smaller hypotheses, by a factor that increases exponentially as the number of consistent examples observed increases.

The size principle is another form of the Bayesian Occam’s razor, favoring small hypotheses. Note the tension between the two forms of Occam’s razor here: the prior favors few clusters, while the likelihood favors small clusters. These factors will typically trade off against each other. For any set of examples, we can always cover them under a single cluster if we make the cluster large enough, and we can always cover them with a hypothesis that is maximally small (i.e., includes no other animals beyond the examples) if we use only singleton clusters and let the number of clusters equal the number of examples. The posterior probability  $p(h|X)$ , proportional to the product of these terms, thus seeks an optimal tradeoff between over- and under-generalization.

## 4 Model results

We consider three data sets. Data sets 1 and 2 come from the specific and general tasks in [3], described in Section 1. Both tasks drew their stimuli from the same set of 10 mammals shown in Figure 1. Each data set (including the set of similarity judgments used to construct the models) came from a different group of subjects. Our models of the probability of generalization for specific and general arguments are given by Equations 1 and 2, respectively, letting  $X$  be the example set that varied from trial to trial and  $y$  or  $Y$  (respectively) be the fixed test category, *horses* or *all mammals*. Osherson et al.’s subjects did not provide an explicit judgment of generalization for each example set, but only a relative ranking of the strengths of all arguments in the general or specific sets. Hence we also converted all models’ predictions to ranks for each data set, to enable the most natural comparisons between model and data.

The first two rows of Figure 2 show the (rank) predictions of three models, Bayesian, max-similarity and sum-similarity, versus human subjects’ (rank) confirmation judgments on

the general (row 1) and specific (row 2) induction tasks from [3]. Each model had one free parameter ( $\lambda$  in the Bayesian model,  $\alpha$  in the similarity models), which was tuned to the single value that maximized rank-order correlation between model and data over both data sets.

The best correlations achieved by the Bayesian model in both the general and specific tasks were greater than those achieved by either the max-similarity or sum-similarity models. The sum-similarity model is far worse than the other two — it is actually negatively correlated with the data on the general task — while max-similarity consistently scores slightly worse than the Bayesian model.

#### 4.1 A new experiment: Varying example set composition

In order to provide a more comprehensive test of the models, we conducted a variant of the specific experiment using the same 10 animal types and the same constant test category, *horses*, but with example sets of different sizes and similarity structures. In both data sets 1 and 2, the number of examples was constant across all trials; we expected that varying the number of examples would cause difficulty for the max-similarity model because it is not explicitly sensitive to this factor. For this purpose, we included five three-premise arguments, each with three examples of the same animal species (e.g.,  $\{\textit{chimp}, \textit{chimp}, \textit{chimp}\}$ ), and five one-premise arguments with the same five animals (e.g.,  $\{\textit{chimp}\}$ ). We also included three-premise arguments where all examples were drawn from a low-level cluster of species in Figure 1 (e.g.,  $\{\textit{chimp}, \textit{gorilla}, \textit{chimp}\}$ ). Because of the increasing preference for smaller hypotheses as more examples are observed, Bayes will in general make very different predictions in these three cases, but max-similarity will not. This manipulation also allowed us to distinguish the predictions of our Bayesian model from alternative Bayesian formulations [1][5] that do not include the size principle, and thus do not predict differences between generalization from one example and generalization from three examples of the same kind.

We also changed the judgment task and cover story slightly, to match more closely the ideal learning task formalized in our theory. Subjects were told that they were vets in training, observing examples of particular animals that had been diagnosed with novel diseases, and they were required to judge the probability that *horses* could get the same disease given the examples observed. This cover story made it clear to subjects that when multiple examples of the same animal type were presented, these instances were distinct individual animals. Figure 2 (row 3) shows the model’s predicted generalization probabilities along with the data from our experiment: mean ratings of generalization from 24 subjects on 28 example sets, using either  $n = 1, 2$ , or 3 examples and the same test species (*horses*) across all arguments. Again we show predictions for the best values of the free parameters  $\lambda$  and  $\alpha$ . All three models fit best at different parameter values than in data sets 1 and 2, perhaps due to the task differences or the greater range of stimuli here.

Again, the max-similarity model is competitive with the Bayesian model, but it is inconsistent with several qualitative trends in the data. Most notably, we found a difference between generalization from one example and generalization from three examples of the same kind, in the direction predicted by our Bayesian model. Generalization to the test category of *horses* was greater from singleton examples (e.g.,  $\{\textit{chimp}\}$ ) than from three examples of the same kind (e.g.,  $\{\textit{chimp}, \textit{chimp}, \textit{chimp}\}$ ). This effect was relatively small but it was observed for all five animal types tested and it was statistically significant ( $p < 0.05$ ) in a 2 X 5 (number of examples X animal type) ANOVA. The max-similarity model, however, predicts no effect here, and likewise for Bayesian accounts that do not include the size principle [1][5].

Although it is reasonable that people’s priors would be set differently in our experiment than those in [3], it is also of interest to ask whether these models are sufficiently robust

as to make reasonable predictions across all three experiments using a single parameter setting. Our Bayesian model is quite robust, achieving correlations of  $\rho \geq 0.90$  on all three data sets for  $0.24 \leq \lambda \leq 0.70$ ; at the single value of  $\lambda = 0.5$ , the Bayesian model achieves correlations of  $\rho = 0.92, 0.95$ , and  $0.95$  on the three data sets, respectively. The max-similarity model achieves a lower plateau correlation ( $\rho \geq 0.85$ ) over a narrower range ( $0.30 \leq \alpha \leq 0.65$ ), and at its single best parameter value ( $\alpha = 0.40$ ) attains correlations of  $\rho = 0.87, 0.90$ , and  $0.90$ .

## 5 Conclusion

Our Bayesian model offers a moderate but consistent quantitative advantage over the best similarity-based models of generalization, and also predicts qualitative effects of varying sample size that contradict alternative approaches. More importantly, our Bayesian approach has a principled rational foundation, and we have introduced a framework for unsupervised construction of hypothesis spaces that could be applied in many other domains. In contrast, the similarity-based approach requires arbitrary assumptions about the form of the similarity measure: it must include both “similarity” and “coverage” terms, and it must be based on max-similarity rather than sum-similarity. These choices have no a priori justification and run counter to how similarity models have been applied in other domains, leading us to conclude that rational statistical principles offer the best hope for explaining how people can generalize so well from so little data. Still, the consistently good performance of the max-similarity model raises an important question for future study: whether a relatively small number of simple heuristics might provide the algorithmic machinery implementing approximate rational inference in the brain.

We would also like to understand how people’s subjective hypothesis spaces have their origin in the objective structure of their environment. Two plausible sources for the taxonomic hypothesis space used here can both be ruled out. The actual biological taxonomy  $\mathcal{H}_{evol}$  for these 10 animals, based on their evolutionary history, looks quite different from the subjective taxonomy  $\mathcal{H}_1$  used here. Substituting  $\mathcal{H}_{evol}$  for  $\mathcal{H}_1$  as the base of our model’s hypothesis space leads to dramatically worse predictions of people’s generalization behavior. Taxonomies constructed from linguistic co-occurrences, by applying the same agglomerative clustering algorithms to similarity scores output from the LSA algorithm [11], also lead to much worse predictions. Perhaps the most likely possibility has not yet been tested. It may well be that by clustering on simple perceptual features (e.g., size, shape, hairiness, speed, etc.), weighted appropriately, we can reproduce the taxonomy constructed here from people’s similarity judgments. However, that only seems to push the problem back, to the question of what defines the “appropriate” feature weights. We do not offer a solution here, but merely point to this question as perhaps the most salient open problem in trying to understand the computational basis of human inductive inference.

## Acknowledgments

Tom Griffiths provided valuable assistance with the statistical analysis. Supported by grants from MERL and the NTT Communication Sciences Laboratories and a HHMI fellowship to NES.

## References

- [1] T. Mitchell. *Machine Learning*. McGraw-Hill, Boston, MA, 1997.
- [2] J.B. Tenenbaum. Rules and similarity in concept learning. In S.A. Solla, T.K. Keen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 59–65. MIT Press, 2000.

- [3] D. Osherson, E. Smith, O. Wilkie, A. López, and E. Shafir. Category-based induction. *Psychological Review*, 97(2):185–200, 1990.
- [4] S. Sloman. Feature-based induction. *Cognitive Psychology*, 25:231–280, 1993.
- [5] E. Heit. *Rational Models of Cognition*, chapter A Bayesian analysis of some forms of induction, pages 248–274. Oxford University Press, 1998.
- [6] J.B. Tenenbaum and F. Xu. Word learning as bayesian inference. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 2000.
- [7] J.B. Tenenbaum. Bayesian modeling of human concept learning. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, *Advances in Neural Information Processing Systems 11*. MIT Press, 1999.
- [8] N.E. Sanjana and J.B. Tenenbaum. Capturing property-based similarity in human concept learning using maximum-likelihood gaussians with a bayesian framework. In *Sixth International Conference on Cognitive and Neural Systems*, 2002.
- [9] S. Atran. *The classifying nature across cultures*, volume 3 of *An invitation to cognitive science*, chapter 5. MIT Press, 1995.
- [10] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, New York, NY, 2001.
- [11] T.K. Landauer and S.T. Dumais. A solution to plato’s problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

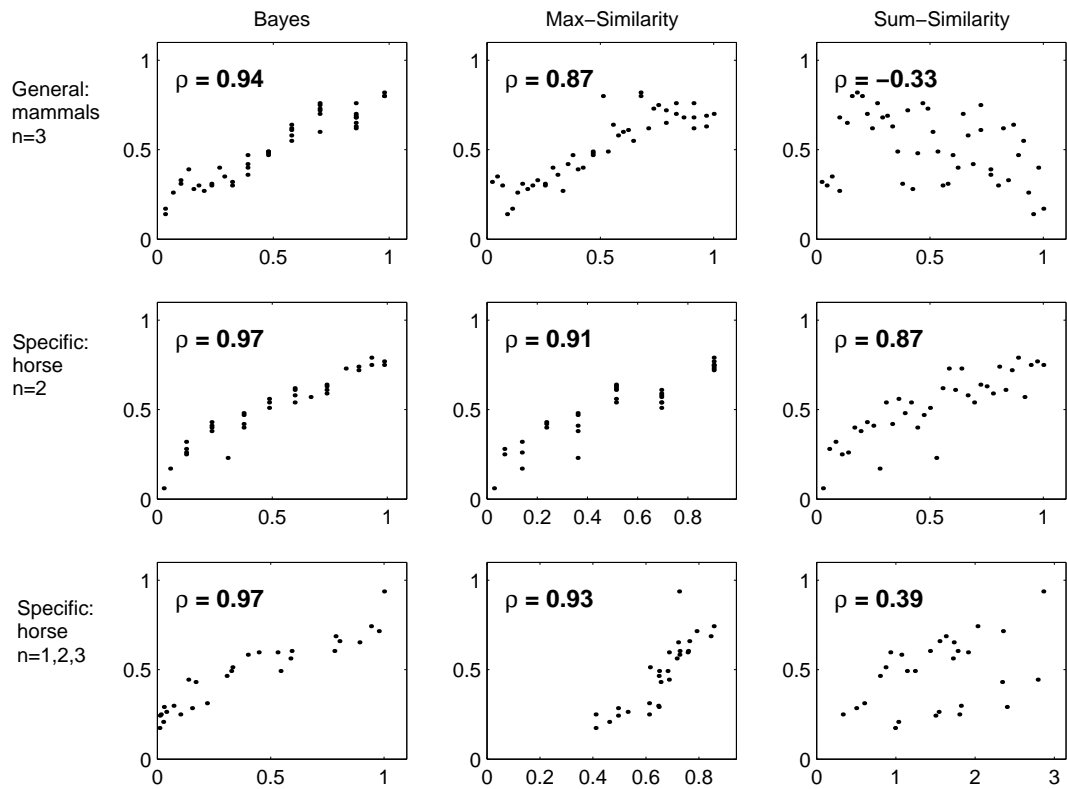


Figure 2: Model predictions ( $x$ -axis) plotted against human confirmation scores ( $y$ -axis). Each column shows the results for a particular model. Each row is a different inductive generalization experiment, where  $n$  indicates the number of examples (premises) in the stimuli.