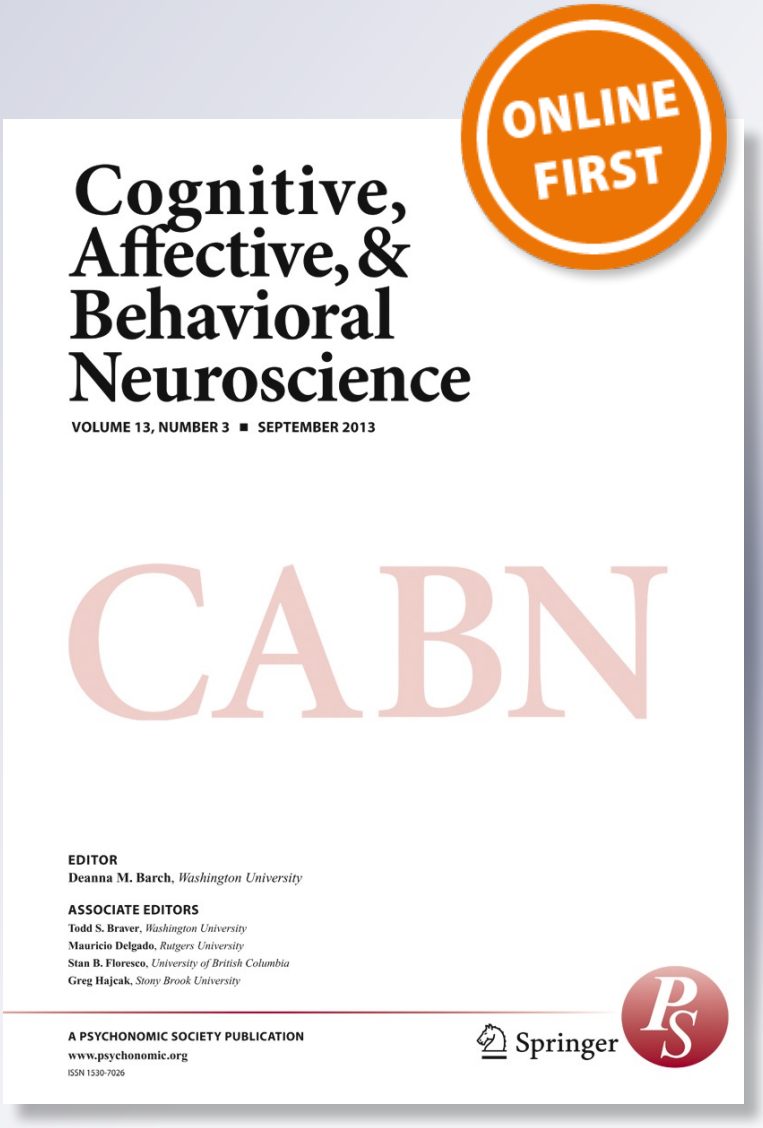
Category-based induction from similarity of neural activation

Matthew J. Weber & Daniel Osherson

Cognitive, Affective, & Behavioral Neuroscience

ISSN 1530-7026

Cogn Affect Behav Neurosci
DOI 10.3758/s13415-013-0221-3



Springer

Cogn Affect Behav Neurosci
DOI 10.3758/s13415-013-0221-3

Category-based induction from similarity of neural activation

Matthew J. Weber · Daniel Osherson

Abstract The idea that similarity might be an engine of inductive inference dates back at least as far as David Hume. However, Hume's thesis is difficult to test without begging the question, since judgments of similarity may be infected by inferential processes. We present a one-parameter model of category-based induction that generates predictions about arbitrary statements of conditional probability over a predicate and a set of items. The prediction is based on the unconditional probabilities and similarities that characterize that predicate and those items. To test Hume's thesis, we collected brain activation from various regions of the ventral visual stream during a categorization task that did not invite comparison of categories. We then calculated the similarity of those activation patterns using a simple measure of vectorwise similarity and supplied those similarities to the model. The model's outputs correlated well with subjects' judgments of conditional probability. Our results represent a promising first step toward confirming Hume's thesis; similarity, assessed without reference to induction, may well drive inductive inference.

Keywords Similarity · induction · fMRI · multivariate pattern analysis · categorization · semantics

M. J. Weber (✉)
Department of Psychology and Center for Cognitive Neuroscience, University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 19104, USA
e-mail: mweb@psych.upenn.edu

D. Osherson (✉)
Department of Psychology and Princeton Neuroscience Institute, Princeton University, Green Hall, Princeton, NJ 08540, USA
e-mail: osherson@princeton.edu

Published online: 20 November 2013

The project

David Hume (1748) is well-known for having emphasized the role of *similarity* in inductive inference. The following remarks are often cited:

> In reality, all arguments from experience are founded on the similarity which we discover among natural objects, and by which we are induced to expect effects similar to those which we have found to follow from such objects. . . . From causes which appear similar we expect similar effects.

We understand this assertion about *experience* and *expectation* as an empirical claim concerning human psychology. Hume goes on to make the epistemological contention that nondemonstrative inferences lack normative justification, whether they are based on similarity, constant conjunction, or anything else (Morris, 2013). However, we focus here on Hume's empirical claim about similarity; thus, in what follows, "Hume's thesis" will be interpreted descriptively (concerning psychology), rather than normatively.

Hume's thesis was anticipated by John Locke (1689), for whom analogy was "the great rule of probability." How to interpret the term *probability* in the writings of Locke and Hume is open to discussion (Cohen, 1980), but the underlying idea seems clear enough: *Inductive inference is often based on similarity*. The goal of the present article is to sharpen Hume's insight by removing the threat of circularity; as the thesis stands, it is possible that similarity derives from inductive inference. (That is, two events might be judged similar to the extent that one predicts the other; see Tenenbaum & Griffiths, 2001, for an argument to this effect). To avoid circularity, we will evaluate the similarity of categories in neural terms, then

embed such similarity in a model of human inductive judgment that will be evaluated empirically.

The present article attempts to relate three types of data: atomic statements of probability about single categories, such as "apples provide at least 10 dietary vitamins"; statements of conditional probability, such as "apples provide at least 10 dietary vitamins given that strawberries do"; and similarities between pairs of categories—such as the similarity between apples and strawberries—defined with recourse to the patterns of neural activation evoked by those categories. Our goal is to estimate human judgments of conditional probability (such as the foregoing inductive inference from strawberries to apples) using just atomic probability and similarity for this purpose. To proceed, we first advance a simple quantitative model that provides such estimates. The model reduces conditional probabilities to a ratio of conjunction probabilities ("apples and strawberries provide at least 10 dietary vitamins") and assigns values to the conjunctions using similarity. We assess the model's accuracy by collecting the three types of data described above, generating estimates of conditional probability, and correlating the estimates with subjects' judgments.

Some of the ideas to be presented already figure in Blok, Medin, and Osherson (2007a, b), Weber, Thompson-Schill, Osherson, Haxby and Parsons (2009) and Weber and Osherson (2010), but we here introduce new analyses and a new, much larger data set.

In what follows, *inductive inference* will be understood as a certain (psychological) relation between a list of statements and some further statement. Schematically,

$$\text{Premises}: \begin{cases} \text{Statement } 1 \\ \text{Statement } 2 \\ \quad\vdots \\ \text{Statement n} \end{cases}$$
$$\text{Conclusion}: \text{Statement C}$$

The first statements are called *premises*, the last the *conclusion*, and the ensemble an *argument*. The *inductive strength* of an argument for a given person will be identified with the subjective conditional probability he or she attaches to the conclusion given the premises. This definition raises questions about subjective probability in the minds of people who misunderstand chance. It is well known that college students can be led to contradictory estimates of probability (see, e.g., Bonini, Tentori & Osherson, 2004; Tentori, Bonini & Osherson, 2004). In the present work, however, we consider probability errors to be exceptional and adopt the probability calculus as a "guardian of common sense" (Pearl, 1988). The quality of our empirical results will justify (or not) this attitude. Thus, we assume that the probability idiom succeeds in conveying a familiar kind of *psychological coherence condition*. An argument is strong to the extent that the reasoner would *find it odd* to believe the premises without believing the conclusion. Squeezing this mental sensation into the unit interval and calling it *probability* provides a rough measure.

Inductive inferences are of such diverse character that one might despair of treating them within a unified theory. It will make things easier to limit attention to premises and conclusion with subject–predicate syntax, the same predicate appearing throughout. Here's an illustration:

- *Premise*: Sheep have at least 18 % of their cortex in the frontal lobe.
- *Conclusion*: Goats have at least 18 % of their cortex in the frontal lobe.

Letting a and c be sheep and goats and Q the common predicate, we get the abbreviation:

- *Premise*: Qa.
- *Conclusion*: Qc.

In the general case, there may be multiple premises, as schematized here:

- *Premises*: $Qa_1 \cdots Qa_n$.
- *Conclusion*: Qc.

Limited to such arguments, a theory of inductive reasoning will attempt to predict the conditional probability $\text{Prob}(Qc \mid Qa_1 \cdots Qa_n)$ of the conclusion given the premises. Such a prediction, of course, must rest on some kind of information. In our case, two kinds of quantity will be assumed available—specifically, the *similarity* between all pairs of objects drawn from $\{c, a_1 \cdots a_n\}$ along with the *un*conditional probabilities $\text{Prob}(Qa_1) \cdots \text{Prob}(Qa_n), \text{Prob}(Qc)$. Similarity will be assumed to take values in the unit interval, be symmetric, and return 1 in case of identity.

Having taken on board all relevant similarities and unconditional probabilities, the only thing missing is the conditional probability of the conclusion given the premises—in other words, the inductive strength of the argument. Thus, our project is to forge conditional probability from unconditional probability plus similarity. The criterion of success will be conformity to the estimates of conditional probability that people typically offer. This puts Hume's thesis—interpreted as a claim about psychology—to empirical test. It does not speak to Hume's doubts about the normative justification of induction.

The next section presents our method for turning unconditional probability and similarity into conditional probability. By proceeding within the constraints of the probability calculus, we attempt to avoid counterintuitive predictions that issue from oversimplified hypotheses about argument strength and similarity. For example, it might be tempting to equate $\text{Prob}(Qc \mid Qa_1 \cdots Qa_n)$ with $\text{Prob}(Qa_i)$, where a_i is the member of $\{a_1 \cdots a_n\}$

with greatest similarity to c. But this proposal incorrectly evaluates the limiting case Prob(Qc | Qc), which it sets equal to Prob(Qc) instead of unity. Interpreting Shepard's (1987) law of universal generalization as a rule for inductive inference falls prey to the same problem. More generally, predictions about the value of Prob(Qc | $Qa_1 \cdots Qa_n$) need to integrate not only similarity but also the (prior) probability of Qc, as well as the compatibility of Qc with $\{Qa_1 \cdots Qa_n\}$. For this purpose, we rely on the standard definition of conditional probability, augmented with principles for assigning probabilities to conjunctions like $Qa_1 \& \cdots \& Qa_n$. It will be seen that our proposal meets several criteria of reasonableness, much like those described above, that we would expect even probability-naïve subjects to respect.

Exploiting similarity in a model of category-based induction

Bayes's theorem offers little help in our enterprise, since it expands the desired conditional probability into an expression featuring a *different* conditional probability, just as challenging. The culprit is the likelihood Prob(Qa | Qc):

$$\text{Prob}(Qc|Qa) = \frac{\text{Prob}(Qa|Qc) \times \text{Prob}(Qc)}{\text{Prob}(Qa)}. \tag{1}$$

We get more help from the usual definition of conditional probability, since it suggests that we attempt to estimate the probability of conjunctions of statements:

$$\text{Prob}(Qc|Qa_1 \cdots Qa_n) = \frac{\text{Prob}(Qc \ \& \ Qa_1 \ \& \ \cdots \ \& \ Qa_n)}{\text{Prob}(Qc \ \& \ Qa_1 \ \& \ \cdots \ \& \ Qa_n) + \text{Prob}(\neg Qc \ \& \ Qa_1 \ \& \ \cdots \ \& \ Qa_n)}. \tag{2}$$

Our problem thus reduces to the following. We're given the probability of each conjunct—for example, the probability that sheep have at least 18 % of their cortex in the frontal lobe. We're also told the pairwise similarity of pairs of objects drawn from $\{c, a_1 \cdots a_n\}$—for example, the similarity of goats and sheep. From this, we'd like to construct sensible estimates of the probabilities of $Qc \ \& \ Qa_1 \ \& \ \cdots \ \& \ Qa_n$ and $\neg Qc \ \& \ Qa_1 \ \& \ \cdots \ \& \ Qa_n$. By dividing the first estimate by the sum of the first and second, we reach the desired conditional probability Prob(Qc | $Qa_1 \cdots Qa_n$).

So what is a sensible estimate of the probability of a conjunction? We look to the laws of chance for advice and recall that conjunction probabilities are mathematically constrained to lie between the bounds shown here (Neapolitan, 1990):

$$\begin{array}{ll} max\left\{0, 1-n+\sum_{i=1}^{n} \text{Prob}(Qb_i)\right\} & \leq \\ \text{Prob} \quad (Qb_1 \& \cdots \& Qb_n) & \leq \\ min\{\text{Prob}(Qb_1), \cdots, \text{Prob}(Qb_n)\} & \end{array} \tag{3}$$

That is, the lowest possible value is one plus the sum of the n conjunct probabilities minus n—or zero if the latter number is negative. The highest possible value is the minimum probability of the conjuncts. For example, if two statements have probabilities .8 and .4, then the probability of their conjunction cannot exceed .4, and it cannot fall below .2.

Now note that if objects a and b were identical—had maximal similarity—then Qa and Qb would express the same proposition. So the conjunction of Qa with Qb would be logically equivalent to Qa, and also to Qb. In this case, the probability of the conjunction falls at the upper bound of the possible interval Prob($Qa \& Qb$)—namely, the minimum of the conjunct probabilities. In contrast, the conjunction Prob($Qa \& Qb$) makes a logically stronger claim to that extent that a and b are dissimilar, so low similarity should situate Prob($Qa \& Qb$) closer to the bottom of the permitted interval (shown above). Using a convex sum, the similarity of a and b determines a position between these extremes.

Before stating the principle in full generality, consider an example. Suppose that Prob(Qa)=.8 and Prob(Qb)=.4. Then (as we saw earlier) the smallest possible value of Prob(Qa & Qb) is .2, and the greatest is .4. The probability we construct for the conjunction is the weighted sum of these endpoints, where the weights are given by the similarity of a to b. Thus, the value attributed to Prob(Qa & Qb) is

$$[.2 \times (1-\text{similarity}(a,b))] + [.4 \times \text{similarity}(a,b)].$$

If similarity(a,b)=.7, then Prob($Qa \& Qb$)=(.2×.3)+(.4×.7)=.34.

In the general case, we're given a conjunction $Qb_1 \& \cdots \& Qb_n$ with n conjuncts. Once again, it is assumed that we know the unconditional probabilities Prob(Qb_1)$\cdots$Prob(Qb_n) of its conjuncts. This allows us to compute the lower and upper bounds on Prob($Qb_1 \& \cdots \& Qb_n$):

$$\text{p} = max\left\{0, 1-n+\sum_{i=1}^{n} \text{Prob}(Qb_i)\right\}, \text{the least possible value of Prob}(Qb_1 \& \cdots \& Qb_n). \tag{4}$$

$$\mathrm{P} = min\{\mathrm{Prob}(Qb_1), \cdots, \mathrm{Prob}(Qb_n)\}, \text{the greatest possible value of } \mathrm{Prob}(Qb_1 \& \cdots \& Qb_n). \quad (5)$$

To aggregate the similarities among all pairs of objects among $\{b_1 \cdots b_n\}$, we define

$$\mathrm{sim} = min\{\mathrm{similarity}(b_i, b_j) | i, j \leq n\}. \quad (6)$$

That is, we take the similarity factor to be the minimum of all the pairwise similarities among the objects appearing in the conjunction; as will be seen shortly, use of minimum (min) ensures a desirable property of our system. Finally, we estimate the probability of the conjunction to be the weighted sum of the lower and upper bounds, where similarity controls the weights:

$$\mathrm{Prob}(Qb_1 \& \cdots \& Qb_n) = [\mathrm{p} \times (1 - \mathrm{sim})] + [\mathrm{P} \times \mathrm{sim}]. \quad (7)$$

The use of minimum similarity yields conformity with the conjunction law; that is, our computations always satisfy

$$\mathrm{Prob}(Qb_1 \& \cdots \& Qb_n) \geq \mathrm{Prob}(Qb_1 \& \cdots \& Qb_n \& Qb_{n+1}). \quad (8)$$

Aggregating similarities via maximum, mean, and many other functions violates the conjunction law, as is easily demonstrated.

We're almost finished with the details of our approach, but a moment is needed to consider negated conjuncts such as $\mathrm{Prob}(Qc\ \&\ \neg Qa)$. In this case, we use the same procedure as before, except that we substitute one minus the unconditional probability for negated statements—for example, using $1 - \mathrm{Prob}(Qa)$ to represent $\mathrm{Prob}(\neg Qa)$. We also use one minus the similarity of two objects that appear in statements of opposite polarity. Thus, in the conjunction $Qc\ \&\ \neg Qa$, we use $1 - \mathrm{similarity}(a,c)$ in place of $\mathrm{similarity}(a,c)$. To see why this policy makes sense, compare the conjunctions Q(lion) & $\neg Q$(cougar) versus $\neg Q$(lion) & $\neg Q$(cougar). The similarity of lions and cougars should lower the credibility of the first conjunction (with respect to typical predicates Q) in view of the different polarities of the two conjuncts. In contrast, the similarity of lions and cougars should enhance the probability of the second conjunction because of the matched polarity of the conjuncts [thus, we use similarity (lion, cougar) in this context, rather than 1−similarity (lion, cougar)].

With the foregoing principles, probabilities may be constructed for arbitrary conjunctions such as $Qb_1\ \&\ \neg Qb_2\ \&\ \neg Qb_3\ \&\ Qb_4$. First, compute upper and lower bounds (P, p) for the probability of the conjunction after substituting $1 - \mathrm{Prob}(Qb_2)$ for $\mathrm{Prob}(-Qb)_2$, and likewise for $\mathrm{Prob}(\neg Q_3)$. Next, compute the minimum pairwise similarity (sim) among $\{b_1, b_2, b_3, b_4\}$, using the one-minus operation on similarities associated with conjuncts of opposite polarity. Then, the probability of the conjunction is taken to be the weighted sum of the upper and lower bounds, with similarity determining the weights:

$$\mathrm{Prob}(Qb_1 \& \neg Qb_2 \& \neg Qb_3 \& Qb_4) = [\mathrm{p} \times (1 - \mathrm{sim})] + [\mathrm{P} \times \mathrm{sim}]. \quad (9)$$

Our method satisfies important coherence conditions involving upper and lower bounds and the conjunction law. It also assigns zero probability to contradictions like

$$Qb_1 \& \neg Qb_2 \& \neg Qb_1 \& Qb_3.$$

On the other hand, there is no guarantee that the conjunctions over a given set of statements are assigned probabilities that sum to one. For example, the probabilities that our method attaches to the following eight conjunctions may not sum to unity, whereas they must do so according to the probability calculus, since they partition logical space.

$Qc \& Qa \& Qb$	$Qc \& Qa \& \neg Qb$
$Qc \& \neg Qa \& Qb$	$\neg Qc \& Qa \& Qb$
$\neg Qc \& \neg Qa \& Qb$	$\neg Qc \& Qa \& \neg Qb$
$Qc \& \neg Qa \& \neg Qb$	$\neg Qc \& \neg Qa \& \neg Qb$

The matter can be rectified through normalization, but there is no need for this in the present context, because conditional probabilities are ratios (so normalizing has no numerical impact).

In summary, to determine a conditional probability like $\mathrm{Prob}(Qc \mid Qa, \neg Qb)$, we construct the probabilities of $Qc\ \&\ Qa\ \&\ \neg Qb$ and $\neg Qc\ \&\ Qa\ \&\ \neg Qb$, then set

$$\mathrm{Prob}(Qc | Qa, \neg Qb) = \frac{\mathrm{Prob}(Qc \& Qa \& \neg Qb)}{\mathrm{Prob}(Qc \& Qa \& \neg Qb) + \mathrm{Prob}(\neg Qc \& Qa \& \neg Qb)}. \quad (10)$$

It remains to test whether this scheme approximates human intuition about chance.

Similarity derived from neural activity

We could ask students to provide numerical estimates of the similarity of pairs of species, using a rating scale. But such a procedure would not fairly test Hume's idea. His thesis was that perceived similarity gives rise to judged probability. We must not inadvertently test the converse idea, that perceived probability gives rise to judged similarity. After all, it could be that lions and cougars seem similar because inferences from one to the other strike us as plausible. Then similarity would indeed be related to induction, but not in the way Hume intended. The contrasting claims may be summarized this way:

- HUME'S THESIS: Conditional probabilities are derived (in part) from the similarities perceived between relevant categories.
- THE CONVERSE THESIS: Similarities are derived (in part) from the conditional probabilities estimated to hold between statements involving the relevant categories.

To focus on Hume's idea, it is necessary to operationalize similarity without allowing probability estimates to play an implicit role. For this purpose, we adopt the idea that similarity of categories—like *horses* and *camels*—is determined by the overlap in their respective mental representations. To quantify such "overlap," we rely on fMRI scanning to identify the patterns of neural activation that support the categories and then measure their "proximity" in physical terms.

Recall that our approach requires two sources of information about our objects $\{b_1 \cdots b_n\}$: unconditional probability (Prob(Qb_i) for $i \in \{1 \cdots n\}$) and similarity (similarity(b_i,b_j) for all i,j). To validate the model's output, we also require judgments of conditional probability—for example, Prob($Qb_i|\pm Qb_j$) for various choices of i,j—although we also consider more complicated arguments. We conducted two experiments to collect this information. Unconditional and conditional probabilities were collected from one group of subjects, who assessed the likelihood of several arguments involving either mammal or fruit categories; those arguments are summarized in Table 1. Similarities were collected by comparing patterns of brain activation in another group, who underwent fMRI scanning while verifying the category membership of images of either mammals or fruit—a procedure deliberately designed to minimize any comparison of categories, improving our prospects for querying similarity in a non-question-begging way. Unconditional probability and neural similarity were supplied as arguments to the model described above. The model's output was then correlated with subjects' judgments of conditional probability. The latter correlation (informed by neural similarity) is referred to as r_{neural}.

The data analysis is described in more detail in the Materials and Methods section below; we here sketch its principal features. We extracted neural similarities using two metrics (mean squared deviation, SqD, and Pearson correlation, PC) from revMatttwelve brain regions. Those regions included six that might be expected to host semantic information about natural categories: the left and right lateral occipital complex (LOC; Edelman, Grill-Spector, Kushnir & Malach, 1998; Malach et al., 1995), as defined by greater activation for intact versus scrambled images and left and right Brodmann areas (BAs) 19 and 37 (Fig. 1). We also examined left and right BA 17, corresponding to the early visual cortex, as well as four regions of the default-mode network, left and right BA 11 (orbitofrontal cortex) and BA 23 (posterior cingulate cortex). We implemented a *fixed-interval analysis* by scaling the neural similarities to the $[\frac{1}{3}, \frac{2}{3}]$ interval and using the scaled similarities to compute r_{neural} via the model. We chose this interval to allow for the possibility of both higher (e.g., dolphin to porpoise) and lower (e.g., dolphin to butterfly) similarities. It is symmetric because asymmetric intervals assign different influence to negated versus nonnegated premises, an assumption that seemed unwarranted. In any event, we go on to show that many expansions and contractions of the similarity interval yield the same result. We then computed the correlation of the model's outputs with judged conditional probability when similarity is allowed no influence on conjunction probability, which we accomplished by taking every similarity to be .5. This *similarity-free* correlation is called r_{sf}. The contribution of a region of interest's (ROI's) similarity information to predicting probability is assessed via the t-statistic quantifying the significance of the difference between r_{neural} and r_{sf}; in what follows, we denote this statistic with t_{n-sf} and the associated p-value with p_{n-sf}. This approach is critical because r_{sf} is quite high for both mammals and fruit ($r_{sf} = .522$, $p < 10^{-11}$ for mammals; $r_{sf} = .476$, $p < 10^{-9}$ for fruit); thus, even a strong positive r_{neural} is not necessarily evidence that similarity information is useful for inference. After investigating the usefulness of neural similarity using the $[\frac{1}{3}, \frac{2}{3}]$ interval, we performed a *varied-interval analysis*, systematically contracting and dilating the interval to assess its impact on r_{neural}. We performed this analysis to develop a sense for the optimal level of influence from similarity; there is no a priori reason to prefer the fixed $[\frac{1}{3}, \frac{2}{3}]$ interval over, for example, $[\frac{1}{4}, \frac{3}{4}]$ or other intervals, and we wished to see which provided the best fit to our observations.

Table 1 Forms of the sentences judged in the behavioral experiment

Form	Number of instances
Prob(Qc)	32
Prob($Qc|Qa$)	80
Prob($Qc|\neg Qa$)	80
Prob($Qc|Qa$,Qb)	80
Prob($Qc|Qa$,$\neg Qb$)	80

Note. Q is the predicate "have at least 18 % of cortex in the frontal lobe," for mammals, or "provide at least 10 dietary vitamins," in the case of fruit; a, b, and c are drawn from our mammal and fruit categories, with any given sentence concerning only mammals or fruit

Results

Figure 2 documents the results of the fixed-interval analysis. In the context of our model, neural similarities derived from the SqD metric almost always yielded predictions of probability that were very well correlated with subjects' judgments; however, not all those correlations improved significantly on r_{sf}. Notably, similarities derived from the SqD metric

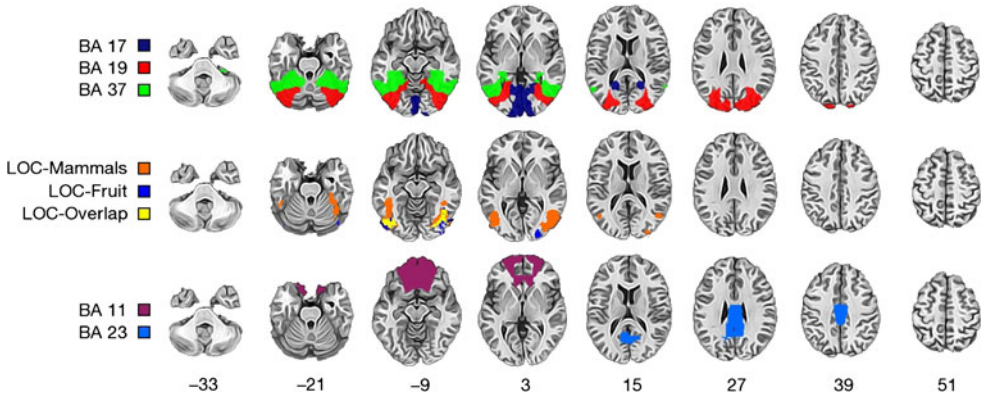


Fig. 1 Masks for region-of-interest analyses. Brodmann area (BA) masks were extracted from the atlas of Rorden and Brett (2000). Lateral occipital complex (LOC) was defined via response to intact versus scrambled images (see the Image Preprocessing section for details)

improved on r_{sf} more frequently than those derived from the PC metric. However, the SqD metric was not obviously better in the ventral visual (LOC, BA 19, BA 37) ROIs than it was in the primary visual cortex (BA 17) and default-mode (BA 11, BA 23) ROIs. In contrast, default-mode ROIs did not yield good predictions via the PC metric, although right BA 17 did yield good predictions via the PC metric in mammal subjects.

Results for the varied-interval analysis are shown in Fig. 3. They largely recapitulate the results of the fixed-interval analysis, although some features stand out. First, the impact of

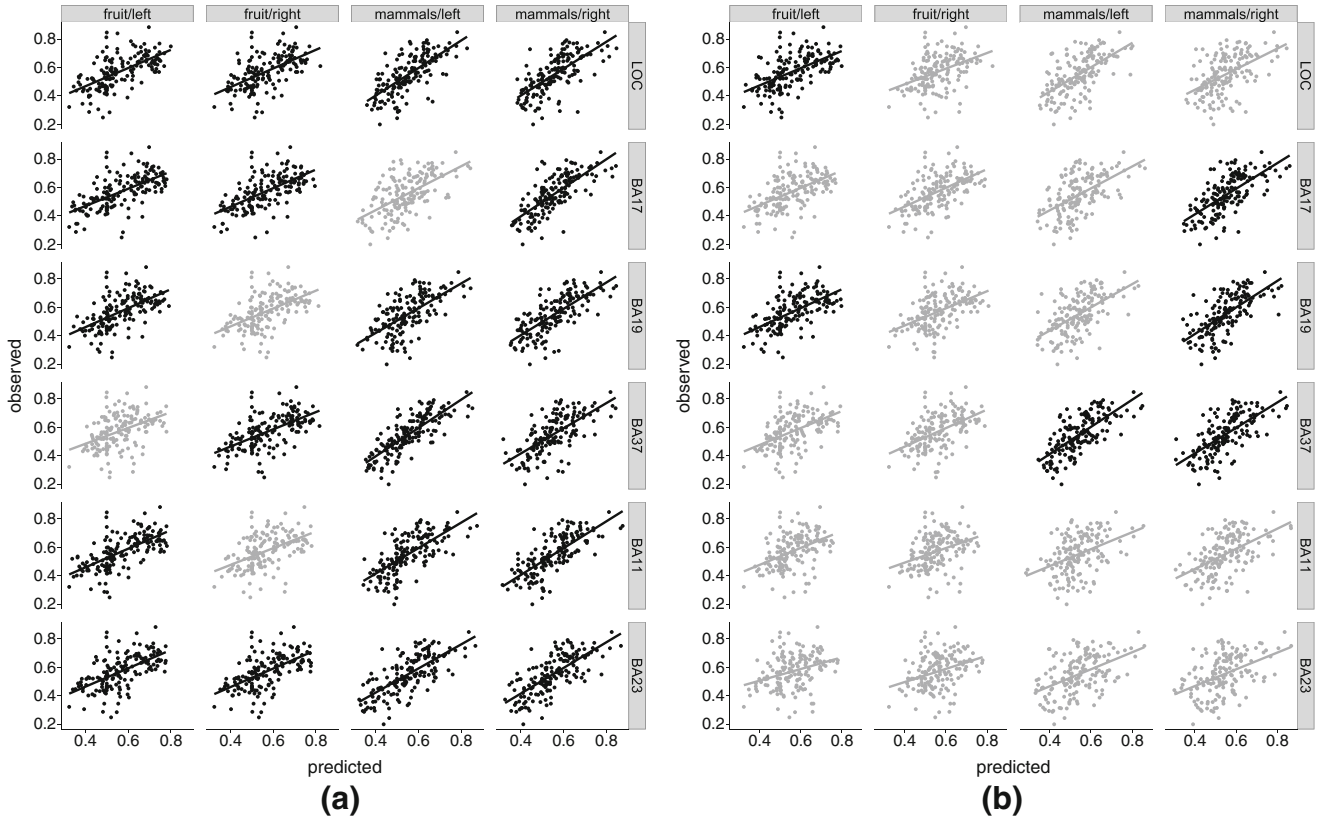


Fig. 2 Correlations with judged probability in 12 regions of interest (ROIs), using two similarity metrics, using the $[\frac{1}{3}, \frac{2}{3}]$ interval for similarity. Subjects' judgments of conditional probability are plotted on the ordinate, the model's outputs on the abscissa. The model's inputs were the judged atomic probabilities and the similarities derived by applying the squared deviation metric to pairs of activation patterns in the ROI. Gray plots indicate insignificant values of t_{n-sf} (recapitulated in the p-values in the upper left corner of each subgraph). **a** Predicted probabilities derived from the SqD metric for neural similarity. **b** Predicted probabilities derived from the Pearson correlation metric

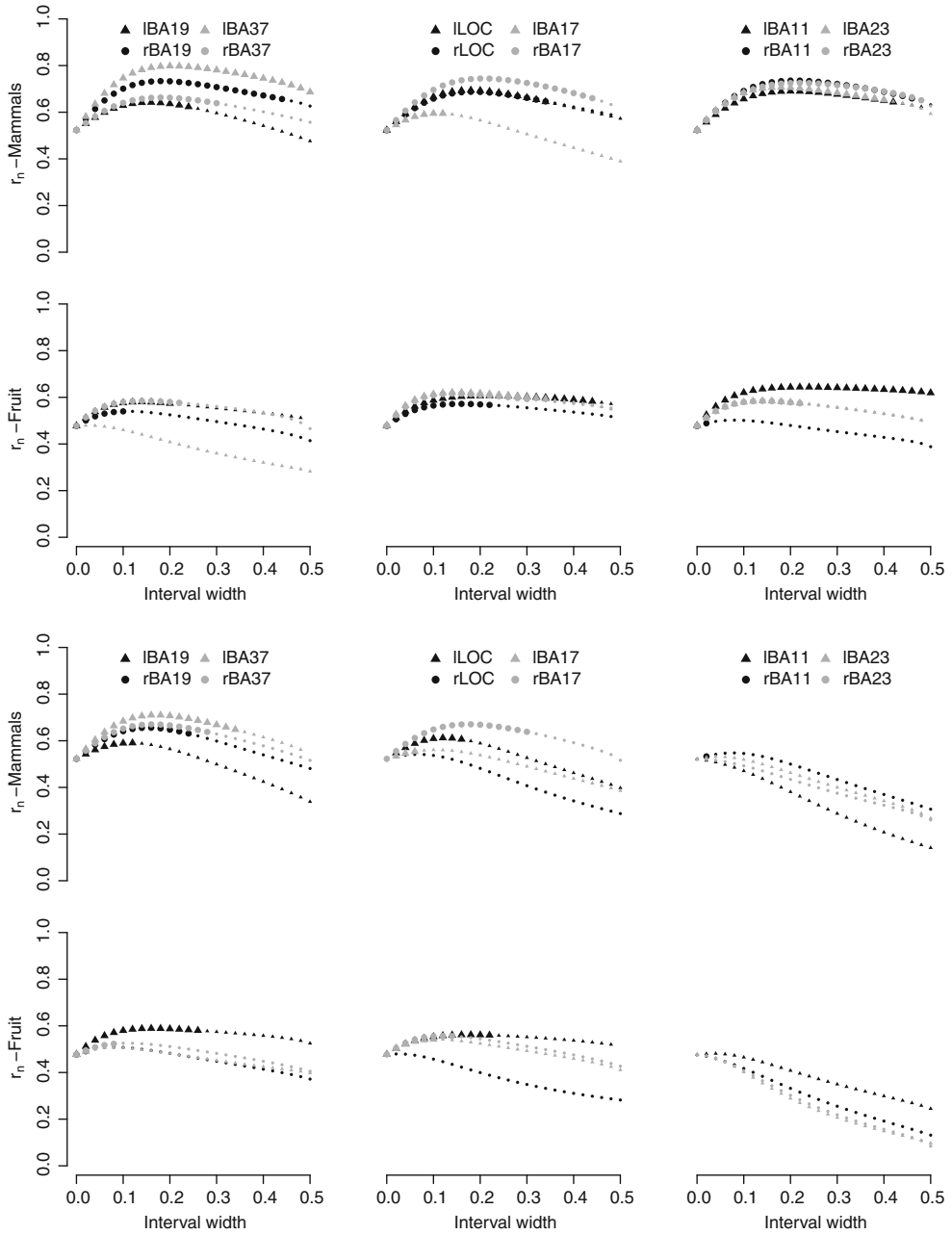


Fig. 3 Value added from similarity in six regions of interest (ROIs). At any point x on the x-axis, the interval for similarity is set to $[0.5-x, 0.5+x]$. The y-axis quantifies r_{neural} for that interval given the similarities from a given ROI. Large points denote significant positive values of t_{n-sf}, small circles insignificant or negative values

similarity is largely consistent across metrics (see, e.g., left BA 37 for mammals), although SqD similarity seems more predictive at more scaling intervals than is the case for PC similarity. Second, some similarities that were not useful when scaled to $[\frac{1}{3}, \frac{2}{3}]$ *are* useful at smaller intervals (e.g., PC similarities from right BA 37 for fruit and from left BA 17 and left BA 19 for mammals). Third, most of the curves show a characteristic concavity, with a peak between 0.1 and 0.3 and a drop thereafter. Fourth and finally, the default-mode ROIs perform at least on par with the early and later visual ROIs when similarity is derived via SqD but provide no added value when similarity is derived via PC.

We examined the correlation between our two measures of neural similarity, SqD and PC, across all the ROIs ($n=12$) figuring in the study. The correlation between these two measures of neural similarity is .76 in mammals and .81 in fruit. It thus appears that the two measures are largely overlapping but not identical.

General discussion

Hume again

Recall Hume's thesis, stated in the introduction. We've tested it with a measure of similarity unlikely to have been produced by inductive mechanisms. To be sure, over the years, inferences about the properties of mammals might affect how they are ultimately coded in the brain; lions and cougars may be coded similarly because they are perceived to share many properties. So neural similarity could depend on inference via this route. Nonetheless, our measure of similarity is directly mediated by the mental representation of concepts, rather than accessing the machinery of inductive cognition. This seems a fair way of making Hume's claim precise. So perhaps our results may be taken to confirm his thesis.

Relationship to earlier work

As was mentioned in the introduction, results akin to those presented here have been reported by Weber and Osherson (2010). The present work employs the same model as that used in the earlier study but applies it to neural representations of many new categories, including a new kingdom—specifically, 16 mammals and 16 fruit, rather than the 9 mammals documented in Weber and Osherson. It also rests on 320 conditional sentences, rather than the 40 employed in Weber and Osherson. Additionally, we compare two metrics of neural similarity in numerous ROIs both within and outside the ventral visual stream,[1] and test the robustness of the model over values of its free parameter. The present study thus replicates and extends our earlier work establishing the usefulness of neural similarity in predicting inductive judgment about natural categories.

Relationship to other models of induction

Earlier work on generalizing properties has suggested that induction is influenced by many variables, including feature overlap (Sloman, 1993), feature centrality (Hadjichristidis, Sloman, Stevenson & Over, 2004; Medin, Coley, Storms & Hayes, 2003), premise diversity (Osherson, Smith, Wilkie, López & Shafir, 1990), premise–conclusion similarity (Osherson et al., 1990), premise and conclusion plausibility (Smith, Shafir & Osherson, 1993), explanations afforded by properties of the items (McDonald, Samuels & Rispoli, 1996), the projectibility (or generalizability) of predicates (Goodman, 1955), and attention during learning (Kalish, Lewandowsky & Kruschke, 2004). The avoidance of Bayes's theorem (Eq. 1) in favor of the definition of conditional probability (Eq. 2) distinguishes our approach from the Bayesian modeling that has clarified several aspects of inductive reasoning (Kemp & Jern, 2013). Our approach is also distinguished by its appeal to neural similarity as a means of deriving the probabilities needed to make predictions about subjects' judgment. The main attraction of the present model is the small number of inputs, relative to the potential outputs. The model's mechanics are conceptually straightforward, boiling down to the use of similarity to choose inductive probability from the interval dictated by the laws of chance. It is clear that many of the properties of inductive inference mentioned above are natural consequences of our model—for example, the effects of conclusion probability and premise–conclusion similarity.

The automaticity of semantic retrieval

We claim to have tapped a source of similarity unlikely to be tainted by induction. Against this claim, a reader might remark on the automaticity of aspects of memory retrieval. The priming literature is a rich source of such findings: For example, McRae and Boisvert (1998) documented priming based on semantic similarity (e.g., subjects more swiftly identified *canary* as a word if it had been preceded by *finch* rather than *jeep*), and Thompson-Schill, Kurtz and Gabrieli (1998) showed that such priming is independent of associative relatedness. However, the critical difficulty we claim to avoid is *affording subjects the opportunity to estimate similarity via inductive judgment*. Even if semantic information about similar categories had been spontaneously evoked during the verification task, the task did not invite subjects to use that information in induction—or, indeed, to do anything at all with it, since similarity judgment was neither requested nor mentioned before the end of scanning. To be sure, similar concepts might prime one another through similarity of neural representations, but this scenario does not posit an intervening inductive process. Thus, we believe that our paradigm successfully furnished patterns of brain activation that, whatever their other defects, were not related to one another by inductive inference.

Choice of similarity metric

The difference between the SqD and PC metrics is consequential: PC-derived similarities add less value on the whole to atomic probability. However, this finding might be viewed as an argument in favor of the PC metric. For, that metric fails to predict inductive judgment in the default-mode ROIs, regions in which we might not expect neural similarity to track conceptual similarity.

It is natural to ask why the two metrics yield such different results. Mathematically, the difference between the SqD and PC metrics is straightforward: For vectors c, d with entries

[1] We thank an anonymous reviewer for these suggestions.

indexed by i, SqD averages $(c_i - d_i)^2$, while PC averages $z(c_i)z(d_i)$ (Pearson, 1896), where $z(x)$ is the z-transformed value of x. Since the SqD metric removes the mean, the obvious difference is that PC normalizes the variance due to extreme voxel values, where SqD does not. Additionally, since the PC metric is bounded to $[-1, 1]$, linear normalization makes it possible for a small number of extreme values to push most of the similarities very close to 0 or 1; for this to occur with PC, by contrast, the correlation coefficients of most of the similarities would actually have to be close to 0 or 1. Histograms of similarities derived from SqD versus PC support this mathematical intuition; SqD generates similarities that are negatively skewed for all ROIs, where PC does so for the early and ventral visual ROIs, but not the default-mode ROIs (Fig. 4).

That difference acknowledged, it is not clear what aspects of the psychological process of induction lend themselves to modeling with negatively skewed distributions. It may be that only extreme similarities (and dissimilarities) have any influence on induction. Perhaps the restricted range of PC-derived similarities forces what should be high similarities to have middling values. This conjecture presupposes that similarity information is widely distributed throughout the brain (albeit, perhaps, in small populations of voxels), even in early visual and default-mode ROIs.

Compression

We've examined Hume's thesis through the lens of a particular model of probability judgment, which starts from unconditional probability and pairwise similarity. The model cannot generate arbitrary distributions. A joint distribution over the statements $Qb_1, Qb_2 \cdots Qb_n$ (where Q is a predicate and $b_1 \cdots b_n$ are objects) requires $2^n - 1$ numbers in general. Our approach specifies distributions based on

$$n + \binom{n}{2}$$

numbers (n unconditional probabilities and all pairwise similarities). So the family of models we've specified must omit many potential distributions. This kind of compression, however, may be compatible with describing aspects of human judgment, which likely chooses distributions from a limited set in most circumstances.

Varieties of predicates

Further progress in constructing a psychological theory of induction requires distinguishing among various classes of predicates. Consider the following predicates:

- have at least 18 % of their cortex in the frontal lobe
- have trichromatic vision
- can suffer muscle damage through contact with poliomyelitis
- brain/body mass ratio is 2 % or more
- require at least 5 h of sleep per day for normal functioning
- sex drive varies seasonally
- have a muscle-to-fat ratio of at least 10-to-1

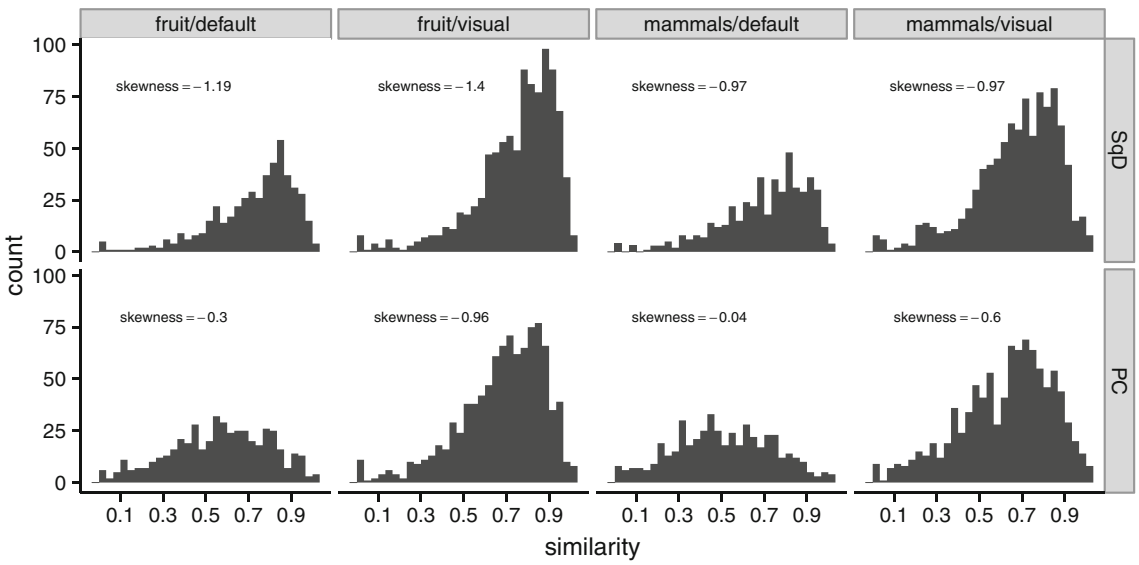


Fig. 4 Histograms of similarity by metric, category, and region of interest (ROI). Graphs marked "default" (the first and third columns) combine similarities from the four default-mode ROIs (left and right BAs 11 and 23); those marked "visual" (the second and fourth columns) combine similarities from all eight visual ROIs (left and right lateral occipital complex and BAs 17, 19, and 37). Note the near zero skewness of PC-derived similarity from default-mode ROIs

- have testosterone levels at least 10 times higher in males, as compared with females.

These predicates are adapted to the present model because they have biological content without evoking detailed knowledge in the minds of most college students. They are also "shareable," in contrast to the predicate *eats more than half the grass in its habitat*—which cannot be true of different species in the same habitat. It is obvious that our model needs adjustment in the face of nonshareability.

The choice of predicate may also affect similarity (Heit & Rubinstein, 1994). The following argument evokes physiology and may be connected to the default representation of mammals in the mind:

Tigers' resting heartbeat is less than 40/minute.
Lions' resting heartbeat is less than 40/minute.

Changing the predicate to the one in the next argument shifts the criteria for relevant similarity, which seems now to be influenced by expense and public interest.

Tigers are kept in most American zoos.
Lions are kept in most American zoos.

It could also be that knowledge of zoo economics intervenes without distorting the basic similarity relation. In this case, no such function as "zoo similarity" is mentally computed. A more thorough treatment of the problem can be found in several influential reports (Jones & Love, 2007; Jones, Maddox & Love, 2006; Medin, Goldstone & Gentner, 1993); only further experiments can sort out the matter.

Other challenges arise when arguments display distinct predicates in premise and conclusion or involve relations like *preys on*. Inferences involving nonnatural kinds—like artifacts and political parties—bring fresh distinctions to light. And then there is the tendency of even well-educated respondents to issue probabilistically incoherent estimates of chance (Kahneman & Tversky, 1983) or to judge similarity asymmetrically (Tversky, 1977). Confronting these complexities is inevitable for the development of any theory of human inductive judgment. The evidence presented here suggests merely that progress will involve appeal to similarity in something like the sense Hume had in mind.

Conclusion

We have demonstrated that the patterns of neural activity in the ventral visual stream, evoked during categorization of a stimulus, exhibit a similarity structure that enables accurate inference about judgments of conditional probability through a simple model of inductive judgment. Although the model predicts judged probability with substantial accuracy even without any similarity information, neural similarity enables the model to improve its predictions. Since our experiment measured brain activity in a context devoid of explicit comparisons among concepts, our results hold promise for a non-question-begging endorsement of Hume's thesis.

Our model of conditional probability from similarity is, of course, incomplete. Human similarity judgments are sometimes asymmetric (Tversky, 1977; but see Aguilar & Medin, 1999), and human inductive judgment is often incoherent (Kahneman & Tversky, 1983; Tentori et al., 2004) and informed by knowledge beyond similarity, such as causal knowledge and logical knowledge (Kemp, Shafto, Berke & Tenenbaum, 2007). Nonetheless, the present work provides evidence that, at least in propitious circumstances, similarity in the structure of brain activity can underwrite inductive judgment.

Materials and methods

Subjects

Twenty-four Princeton undergraduates, graduate students, and staff (13 female, 18–27 years of age) participated in the fMRI studies for $30 compensation. A separate pool of 64 undergraduates (40 female, 18–22 years of age) participated in the behavioral experiment for course credit.

Stimuli and procedure: behavioral

In the behavioral experiment, subjects were asked to reason about 320 conditional sentences taking four forms, as well as 32 unconditional sentences (Table 1). The sentences were constructed from the items and predicates summarized in Table 2. The items in the conditional sentences were determined randomly, with the constraint that a given item

Table 2 The 32 mammal and fruit categories used in both experiments, plus the predicates used in the behavioral experiment

Mammal categories			
Predicate: *have at least 18 % of cortex in the frontal lobe*			
Bear	Deer	Giraffe	Panda
Camel	Dolphin	Hippo	Rat
Chimpanzee	Elephant	Horse	Squirrel
Cougar	Fox	Lion	Wolf
Fruit categories			
Predicate: *provide at least 10 dietary vitamins*			
Apple	Green apple	Nectarine	Plum
Avocado	Lemon	Orange	Raspberry
Banana	Lime	Peach	Strawberry
Grape	Mango	Pear	Tomato

appeared only once per sentence. Any given sentence concerned only mammal or fruit categories and the corresponding predicate.

Subjects were divided into eight groups of 8; half reasoned only about sentences concerning mammals, the other only about sentences concerning fruit. Each subject viewed 10 conditional sentences from each of the four forms, plus all 16 unconditional sentences for the category, for a total of 56 sentences per subject. Subjects were briefly instructed on the meaning of conditional probability (that is, the probability of the conclusion *assuming* the premises) and then were presented with the 56 sentences in random order on a computer program that recorded subjects' responses on a slider ranging from 0 to 100. For example, all subjects were asked to rate the (atomic) probability that "wolves have at least 18 % of cortex in the frontal lobe." This statement might be assigned a probability of .65. Later (or earlier) in the experiment, the subject might be asked to rate the (conditional) probability that "wolves have at least 18 % of cortex in the frontal lobe given that foxes do not." The similarity of wolves and foxes might motivate the subject to assign this conditional statement a lower probability than the atomic probability for wolves—say, .4. At some other time, the subject might be asked to rate the (conditional) probability that "wolves have at least 18 % of cortex in the frontal lobe given that dolphins and elephants do." Dolphins and elephants being dissimilar to wolves, this might occasion a judgment fairly close to the atomic probability—say, .68 (in acknowledgement of shared mammalhood). Of note, subjects were explicitly instructed that the premises were given for purposes of argument; that is, the fact that a premise was assumed for a given sentence should have no bearing on the assessment of future sentences. However, the assignment of probability and similarity was entirely up to them. No instructions on estimating probability and similarity were provided.

Stimuli and procedure: fMRI

In the scanner, subjects viewed 12 pictures and their left–right reverses from each of either the 16 mammal or the 16 fruit categories, as well as phase-scrambled versions of the same images (the unscrambled images are called *intact* in what follows). The mammal pictures were collected from Google Image Search; the images were converted to grayscale, their surroundings were manually erased, and they were rescaled to 400 × 400 pixels. The fruit pictures were taken by a collaborator, Andrew Connolly, and subjected to a similar procedure, except that color was retained. The design of the experiment was identical with respect to mammal and fruit categories, so the following generic description applies to both.

During scanning, subjects performed an *experimental task* on blocks of the intact images and a *visual baseline task* on blocks of the scrambled images. Experimental and baseline blocks were distributed throughout the experiment by a randomly determined ordering for odd-numbered subjects and the reverse of that ordering for even-numbered subjects. During a trial of the experimental task, subjects saw the name of one of the categories (e.g., *bear*, *apple*) for 2 s, then a series of 24 or fewer distinct intact images of the named category, each presented for 667 ms (totaling 16 s or less for images). Each series was terminated by three distinct images (667 ms each) of a species drawn from 1 of the 15 remaining categories (thus, a mismatch to the initial label). The images in a trial were randomly selected without replacement and randomly ordered within the trial. Subjects were asked to respond as soon as possible when a mismatched image appeared. One trial might have consisted of the word *bear*, followed by nine bear images and then three squirrel images, with subjects expected to respond as soon as they saw a squirrel. Different trials displayed varying numbers of images (0–24) before the intruder appeared, equated across trials for the 16 mammals. The trials with zero matching images can be considered *catch trials*, occasionally inserted to ensure that subjects were attending even to the very first image.

Each baseline trial employed scrambled images from one mammal or fruit category. The form of the visual baseline task was identical to the main task, except that #### (in lieu of a category label) was presented prior to the images and subjects searched for a low-contrast crosshatch (#) in the sequence instead of a category mismatch. Three images with # appeared at the end of each trial in positions corresponding to intruders in the main task.

Image acquisition

Scanning was performed with a 3-Tesla Siemens Allegra fMRI scanner. Subjects' anatomical data were acquired with an MPRAGE pulse sequence (176 sagittal slices) before functional scanning. Functional images were acquired using a T2-weighted echo-planar pulse sequence with thirty-three 64 × 64-voxel slices, rotated back by 5° on the left–right axis (axial-coronal −5°).Voxel size was 3 × 3 × 3 mm, with a 1-mm gap between slices. The phase encoding direction was anterior-to-posterior. TR was 2,000 ms; time to echo was 30 ms; flip angle was 90°. Field of view was 192 × 192 cm.

Image preprocessing

For each subject, functional data were registered to the anatomical MRI, despiked, and normalized to percent signal change; for the ROI analysis, data were additionally smoothed with a 6-mm full width at half max Gaussian kernel. Multiple regression was then used to generate regression coefficients representing each voxel's activity in each experimental condition and each visual baseline condition. All regressors were convolved with a canonical, double-gamma hemodynamic

response function; motion estimates from the registration procedure were included as regressors of no interest. In a given voxel, the *activation* for a given concept was defined as activation in response to the intact trials minus activation in response to the corresponding baseline trials. Subsequent to the regression, brain maps were spatially normalized to Talairach space and, for the ROI analysis only, resampled to a 1 × 1 × 1 mm grid. We relied on the statistical package AFNI (Cox, 1996) for preprocessing and regression.

We used a group *t*-test comparing experimental with visual baseline trials to localize the LOC (Malach et al., 1995), designating the LOC as the largest cluster on each side of the brain that responded positively to intact versus scrambled images [t(11) > 4.44, p < .001, uncorrected]. See Fig. 1, middle row, for the LOC voxels as defined for mammal and fruit categories; overlap is substantial but not total. Note that voxels were included in this ROI only on the basis of their ability to distinguish intact from scrambled images; there was no requirement for them to distinguish intact or scrambled images from one another, still less to recapitulate any particular similarity structure. In fact, uniform responses over categories would have increased, not decreased, a voxel's likelihood of appearing in the LOC ROI, since the contrast designates all intact images as identical and treats any variation in the responses to them as noise.

Multivariate analyses

The preprocessing procedure defines an activation for every concept in every voxel of the brain. We may choose a subset of those voxels, *V*, for examination; the mean-centered vector of activations in *V* for a concept is called the *pattern for* that concept in *V*. For a pair of concepts in *V*, we may define *pattern similarity* from the squared deviation between corresponding voxels (Weber & Osherson, 2010; Weber et al., 2009):

$$\sum_{i=1}^{n} (pattern_1 - pattern_2)^2, \tag{11}$$

where $pattern_1$ and $pattern_2$ are patterns for concepts in *V*. A vector of the squared deviations between all pairs of concepts, linearly normalized to the unit interval (by subtracting the minimum from all values, then dividing by the maximum minus the minimum) and inverted by subtraction from 1 (so that lower distances become higher similarities), provides similarity. The matrix of all such pairwise similarities is called a *neural similarity matrix* in *V* (Kriegeskorte, Mur & Bandettini, 2008). Each neural similarity matrix is a symmetric 16 × 16 matrix, with 1 on the diagonal and 120 unique off-diagonal entries.

For various choices of *V*, we calculated neural similarity matrices and supplied the results to the model of induction described above, together with the atomic probabilities collected from the behavioral subjects. It remains only to set the interval over which similarities are scaled. The correlation between neurally derived probabilities and the conditional probabilities furnished by our behavioral subjects quantifies whether that region hosts patterns of activity that can underwrite inductive inference in our simple model; we will refer to this correlation as r_{neural}, the correlation based on neural similarities.

Critically, it is possible to generate estimates of conditional probability that do not rely on similarity at all, merely by collapsing the similarity interval to the single point, .5. These *similarity-free probabilities* correlate significantly with human judgment ($r = .522$, $p < 10^{-11}$ for mammals; $r = .476$, $p < 10^{-9}$ for fruit), a correlation we call r_{sf}. In all cases, we evaluate the additional contribution of similarity using a test for the difference between dependent correlations (Steiger, 1980; Williams, 1959) between neural and similarity-free probabilities. This test outputs a *t*-statistic, which we call t_{n-sf}. In each ROI, we calculated neural similarity matrices, neurally derived probabilities, r_{neural}, and t_{n-sf} across a range of scaling intervals for similarity, from [0.48, 0.52] (minuscule influence) to [0, 1] (great influence) in steps of .02 on either side of the interval.

Acknowledgments This research was supported by a National Science Foundation Graduate Research Fellowship to M.J.W. and by the Henry Luce Foundation. We thank the Thompson-Schill lab, particularly Sharon Thompson-Schill and Andrew Connolly, for useful discussions and help with stimulus design and data collection. We dedicate this article to the living memory of Edward E. Smith, whose contribution to the study of reasoning and similarity is the foundation of the discussion.

References

Aguilar, C. M., & Medin, D. L. (1999). Asymmetries of comparison. *Psychonomic Bulletin & Review, 6,* 328–337.

Blok, S., Medin, D., & Osherson, D. (2007a). From similarity to chance. In Heit, E. & Feeney, A. (Eds.), *Inductive reasoning: Experimental, developmental, and computational approaches*. Cambridge University Press.

Blok, S. V., Medin, D. L., & Osherson, D. N. (2007b). Induction as conditional probability judgment. *Memory & Cognition, 35*(6), 1353–1364.

Bonini, N., Tentori, K., & Osherson, D. (2004). A different conjunction fallacy. *Mind and Language, 19*(2), 199–210.

Cohen, L. J. (1980). Some historical remarks on the Baconian conception of probability. *Journal of the History of Ideas, 41*(2), 219–231.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29,* 162–73.

Edelman, S., Grill-Spector, K., Kushnir, T., & Malach, R. (1998). Toward direct visualization of the internal shape representation by fMRI. *Psychobiology, 26,* 309–21.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Hadjichristidis, C., Sloman, S., Stevenson, R., & Over, D. (2004). Feature centrality and property induction. *Cognitive Science, 28,* 45–74.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(2), 411–422.

Hume, D. (1748). *An enquiry concerning human understanding*. Oxford University Press.

Jones, M., & Love, B. C. (2007). Beyond common features: The role of roles in determining similarity. *Cognitive Psychology, 55*(3), 196–231.

Jones, M., Maddox, W. T., & Love, B. C. (2006). The role of similarity in generalization. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, pages 405–410, Mahwah, NJ. Lawrence Erlbaum Associates.

Kahneman, D., & Tversky, A. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293–315.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review, 111*(4), 1072–1099.

Kemp, C., & Jern, A. (2013). A taxonomy of inductive problems. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-013-0467-3

Kemp, C., Shafto, P., Berke, A., & Tenenbaum, J. B. (2007). Combining causal and similarity-based reasoning. *Advances in Neural Information Processing Systems, 19,* 681–688.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(4). doi:10.3389/neuro.06.004.2008

Locke, J. (1689). *An essay concerning human understanding*. London: William Tegg.

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., … Tootell, R. B. H. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences, 92*(18), 8135–8139.

McDonald, J., Samuels, M., & Rispoli, J. (1996). A hypothesis-assessment model of categorical argument strength. *Cognition, 59,* 199–217.

McRae, K., & Boisvert, S. (1998). Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 558–572.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review, 10,* 517–532.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*(2), 254–278.

Morris, W. E. (2013). David Hume. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*. Stanford, CA: The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University.

Neapolitan, R. (1990). *Probabilistic reasoning in expert systems: Theory and algorithms*. New York NY: Wiley.

Osherson, D., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category based induction. *Psychological Review, 97*(2), 185–200.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo, CA: Morgan Kaufmann.

Pearson, K. (1896). Contributions to the mathematical theory of evolution. Note on reproductive selection. *Proceedings of the Royal Society of London, 59,* 301–305.

Rorden, C., & Brett, M. (2000). Stereotaxic display of brain lesions. *Behavioral Neurology, 12,* 191–200.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317–1323.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology, 25,* 231–280.

Smith, E. E., Shafir, E. B., & Osherson, D. N. (1993). Similarity, plausibility, and judgments of probability. *Cognition, 49,* 67–96.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245–251.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences, 24,* 629–640.

Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: A misunderstanding about conjunction? *Cognitive Science, 28,* 467–477.

Thompson-Schill, S. L., Kurtz, K. J., & Gabrieli, J. D. (1998). Effects of semantic and associative relatedness on automatic priming. *Journal of Memory and Language, 38*(4), 440–458.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.

Weber, M. J., & Osherson, D. N. (2010). Similarity and induction. *Review of Philosophy and Psychology, 1,* 245–264.

Weber, M., Thompson-Schill, S. L., Osherson, D., Haxby, J., & Parsons, L. (2009). Predicting judged similarity of natural categories from their neural representations. *Neuropsychologia, 47*(3), 859–868.

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society, Series B, 21,* 396–399.