The Structure and Formation of Natural Categories

DOUGLAS FISHER

Department of Computer Science
Vanderbilt University, Nashville, TN 37235

PAT LANGLEY

AI Research Branch, Mail Stop 244–17
NASA Ames Research Center, Moffett Field, CA 94035

Abstract

Categorization and concept formation are critical activities of intelligence. The nature of these processes and the conceptual structures that support them is an important issue at the interface of cognitive psychology and artificial intelligence (AI). Our work assumes that advances in these and other areas are best facilitated by interdisciplinary research methodologies. In particular, we describe a computational model of concept formation and categorization that exploits a *rational* analysis of *basic-level* effects by Gluck and Corter (1985). Their work provides a clean prescription of human category preferences that we adapt to the task of concept learning. In addition, we extend their analyses to account for *typicality* (Rosch & Mervis, 1975) and *fan* (Anderson, 1974) effects, and speculate on how our concept formation strategies might be extended to other facets of intelligence, such as problem solving.

To appear in G. H. Bower (Ed.) (1990), *The Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 26). Cambridge, MA: Academic Press.

1. Introduction

Cognitive simulation fits computational mechanisms to the constraints of psychological data, but there has been long-term debate over the appropriate starting point for this process. Newell and Simon (1972) recommend an initial *task analysis*, in which one identifies alternative approaches to a given task (e.g., cryptarithmetic). Anderson (in press) suggests a more formal *rational analysis*, in which one associates a general category of behaviors (e.g., concept formation) with a performance function to be optimized. In both views, the guiding assumption is that natural organisms are rational but resource-bounded decision makers (Simon, 1969). A similar but less formal view is implicit in *speculative* analyses (Hall & Kibler, 1985), which posit high-level computational principles that constrain human processing (e.g., Kolodner, 1983).

This paper focuses on COBWEB (Fisher, 1987a, 1987b), a cognitive simulation of concept formation and recognition. In particular, we trace the origins of the system to rational and speculative analyses of this task. Concept formation is a process of organizing observations into categories based on internalized measures of category 'quality', without the aid of an external tutor. Moreover, this process of category formation should be guided by two principles. First, learning should be *incremental*, in that observations should be efficiently incorporated into memory as they are encountered. Second, learning should benefit *performance* on some task, in this case predictions about unknown properties of novel observations.

To realize these objectives, the COBWEB model borrows a measure of concept quality developed by Gluck and Corter (1985) in their work on *basic-level* effects in humans (also see Corter & Gluck, 1985). In hierarchical classification schemes, humans tend to prefer one level of abstraction (the 'basic' level) over others. Gluck and Corter's measure, *category utility*, came from a rational analysis which postulated that basic concepts are preferred because they optimize inference ability. COBWEB also incorporates ideas from Kolodner's (1983) CYRUS and Lebowitz's (1982) UNIMEM, which provide general strategies of efficient classification and concept formation. This union yields a system that meets the computational objectives of efficient retrieval and accurate prediction. In addition, the model accounts for certain *typicality* effects (Rosch & Mervis, 1975) and *fan* effects (Anderson, 1976). Thus, it provides a unified account for a number of memory phenomena in a single, parameter-free model of concept representation and concept formation.

In the following section, we introduce some computational and psychological principles of concept learning and representation. Notably, we view concept formation and related tasks in terms of *search* through a state space. After this, Section 3 reviews psychological findings that constrain the representation, access, and acquisition of concepts. In Section 4 we describe COBWEB, a model of concept formation that incorporates these constraints. Section 5 then evaluates the model in terms of its ability to explain a variety of psychological effects and the relations among them. In the final section, we speculate on other applications of the model, including the transition from novice to expert problem-solving skills. Implicitly, our discussion will endorse Anderson's rational view of cognitive simulation as a profitable methodology to pursue issues at the boundary of cognitive psychology and artificial intelligence.

2. Concept learning

Concept learning has been widely studied in both artificial intelligence (AI) and psychology. However, both fields have traditionally emphasized learning tasks in which a tutor provides class information. We begin this section by discussing methods for such supervised learning, since they provide important background for our later discussion of concept formation. In particular, we introduce the view of concept learning as a search process, in which learning mechanisms may vary along two dimensions: *search control* and *search direction*. We then extend the search framework to clustering and concept formation, types of unsupervised learning in which there is no external tutor to provide class information.

2.1 Supervised Learning

Many psychological studies of learning have focused on *concept acquisition* or *identification* (Bruner, Goodnow, & Austin, 1956; Hunt, Marin, & Stone, 1966; Reed, 1972; Medin & Schaffer, 1978), in which a subject must learn to identify novel members of categories, given training observations that are classified by the experimenter. In many experimental settings, the subject is shown a sequence of observations; after viewing each observation, the subject must predict the category membership of that observation and is then told the correct category. Thus, the experimental setting usually requires continuous and *active* participation by the subject. Psychological investigations have focused on characterizing the number of observations that subjects require to consistently predict correct category membership and on the number of classification errors made before they attain criterial accuracy.

Because it involves external feedback, concept acquisition is sometimes referred to as *supervised* learning. In artificial intelligence, this task is more commonly called *learning from examples* (Winston, 1975; Quinlan, 1979; Mitchell, 1982; Dietterich & Michalski, 1983), since a tutor supplies preclassified examples from which the learning system must discover an appropriate concept (intensional) description. Many machine learning systems assume that the target concept to be learned is conjunctive; thus the learner acquires concept(s) that capture shared conditions over all of the observations.

The notion of *search* plays a traditional role in characterizing AI systems, and one can apply this idea to systems that learn concepts (Simon & Lea, 1974; Mitchell, 1982). One important aspect is the *direction* of the search process. Many AI concept learning systems begin by comparing two observations and extracting the commonalities between them (Hayes-Roth & McDermott, 1978; Vere, 1980). They then compare these common features to a third observation, again extracting the collective commonality. This process continues until they have exhausted all the observations, thus yielding the common structure that summarizes the entire set. This strategy follows a *specific-to-general* direction, since the set of common features is initialized as a specific instance and gradually becomes more general as more observations are seen. In contrast, other systems follow a *general-to-specific* strategy (Langley, 1987; Schlimmer & Fisher, 1986). These systems begin with very general concept descriptions, making them more specific as errors suggest the need for more constrained conditions. Further errors lead to even more specific concepts, until they achieve a description

that summarizes all the training instances. Still other systems (Anderson & Kline, 1979; Mitchell, 1982; Schlimmer & Granger, 1986) combine these two strategies, carrying out bidirectional search through the space of concept descriptions.

Concept learning systems also vary in terms of their *search control* strategy. In general, there will be many concept descriptions that cover the training observations, and one must somehow deal with these alternatives. For example, suppose the learner sees two card hands, one with three Jacks and two Kings, and a second with two Jacks and three Kings. One hypothesis that summarizes these observations is that the hands contain *at least two Jacks and at least two Kings*, but an alternative summary is that they contain *two cards of one face and three of another* (i.e., a full house). Such alternatives are the cause of search in concept learning, and researchers have used a variety of strategies to control this search. These methods range from exhaustive techniques like *breadth-first* search, which retain *all* concepts that are consistent with the known observations (e.g., Mitchell, 1982), to heuristic methods like *beam search* (Michalski, 1983), which retains only the 'best' hypotheses that are consistent with the observations.

Unlike experimental human subjects in psychology, many AI learning systems are not required to actively predict class membership for each incoming observation. Rather, they process all available observations *en masse* to produce a set of concept descriptions that are consistent with the observations. This is not to say that many systems could not be adapted to actively predict membership, but they were not designed with this performance task in mind. For instance, Quinlan's (1979, 1986) ID3 algorithm uses a heuristic that requires examination of all observations, thus complicating any strategy for generating intermediate predictions. However, one can modify the basic method to construct descriptions *incrementally* (Schlimmer & Fisher, 1986; Utgoff, 1988), giving it the ability to make predictions after each training instance.

Although 'nonincremental' approaches have predominated in the literature on machine learning, a growing number of researchers have examined incremental methods for concept learning. Examples include a system by Winston (1975) and Schlimmer and Granger's (1986) STAGGER system, which generate predictions for each incoming observation. One can view such systems as conducting a form of constrained search called *hill climbing*, which maintains a single 'active' concept description that may be modified after each training instance.[1] These systems keep no explicit memory of previous hypotheses, though they may simulate backtracking (return to an earlier hypothesis) by application of their learning mechanisms.

Limiting search to one change per observation characterizes hill-climbing learners: a single alternative is kept in memory and intermediate predictions are made efficiently. Of course, placing limits on memory and backtracking ability means that the order of training instances can have an important effect, sometimes leading the learning system astray. However, such order effects have also been observed in human learners (e.g., Kline, 1983), making them desirable characteristics of

1. Not all incremental learning systems should be viewed as hill climbers. For instance, some methods (Anderson & Kline, 1979; Langley, 1987) retain a large set of competing descriptions, using the competitor with the highest 'strength' to make a prediction. In addition, Winston's system is not a strict hill climber in that it retains some true backtracking ability, but nonetheless it has many of the characteristics that we deem important for incremental learning.

a computational model. We will return to the notion of incremental hill climbing when we discuss the task of concept formation.

2.2 Unsupervised Learning

Despite the attractiveness of supervised learning tasks, there are many scenarios in which a learner cannot rely on external feedback. In such cases, the learner must invoke internalized heuristics to organize its observations. For example, many machine learning systems incorporate a notion of 'similarity'. Such a bias also occurs in work on numerical taxonomy (Everitt, 1980; Gennari, 1989), in which algorithms use a similarity measure (e.g., the inverse of Euclidean distance) to group similar observations into the same category.

To clarify this point, let us consider some algorithms from the numerical taxonomy literature. For instance, 'nearest-neighbor' methods place an observation in the category that has the most similar current member. Other methods compute a theoretical observation that represents the central tendency (i.e., the centroid) of each category; they then place the new observation with the category having the most similar centroid. These methods have the emergent effect of placing great emphasis on maximizing the intra-category (i.e., within-category) similarity of observations.

Although this approach has intuitive appeal, it presents difficulties if one wishes to break the observations into a number of contrasting categories. In reference to psychological models, Medin (1983) points out that the set of singleton categories optimizes intra-category similarity, since each observation is maximally similar to itself. Thus, attention on intra-category similarity alone does not provide a sufficient basis for deciding upon the appropriate number of clusters. As a result, clustering methods often require that the user specify the number of categories to be formed. Alternatively, they build a tree called a *dendrogram*, in which each node specifies a cluster of lower-level nodes, terminating in individual observations. Following the clustering process, the user severs the tree at various points to obtain the desired number of clusters.

Some techniques of numerical taxonomy explicitly seek to optimize a function of contrasting categories. However, just as intra-category similarity favors singleton classes, inter-category *dis*similarity favors a single all-inclusive category, since there are no contrasting categories to share properties with it (Medin, 1983). Thus, a reliance on both these measures might reduce the need for user intervention. To this end, some methods incorporate a tradeoff between intra-group and inter-group similarities, favoring categories whose members have much in common with each other and little in common with members of contrasting categories. In Section 3 we examine one such tradeoff function.

Recently, machine learning researchers have developed methods for *conceptual clustering*. For example, Michalski and Stepp's (1983) CLUSTER attempts to form categories that have 'good' concept descriptions, which can be stated as conjunctive expressions of features that are common to all or most category members. One criterion, *simplicity*, dictates that the conjunctive expression should be short for the sake of comprehensibility. A second criterion, *fit*, prefers detailed (specific) conjunctive descriptions. These criteria (and others) trade off against one another in much the same way as intra-category and inter-category similarity. The ability to form very simple discrim-

inating concepts for contrasting categories implies very little overlap between members of different categories, whereas specific categories implies that there is considerable intra-category similarity.[2]

Other nonincremental clustering systems include Hanson and Bauer's (1989) WITT and Cheeseman, Kelly, Self, Stutz, Taylor, and Freeman's (1988) AUTOCLASS. The former computes correlations between feature pairs, forming clusters so as to maximize the intra-category pairwise correlations across all features and to minimize the average inter-category pairwise correlations across all features and all contrasting categories. AUTOCLASS represents another probabilistic approach to clustering, using a Bayesian method to calculate the 'most probable' categories present in the observations. Intuitively, the most probable clusters are those whose feature distributions vary most from a presumed prior distribution. As with WITT, AUTOCLASS is sensitive to intra-category and inter-category similarities, and thus need not be told the number of clusters to form. The systems are also similar in their lack of any method for making intermediate predictions.[3] We now turn to methods for unsupervised learning that support continuous interaction with the environment.

2.3 Concept Formation

The unsupervised systems that we have described so far are nonincremental, requiring all training instances at the outset. However, in many cases human learners appear to assimilate instances as they become available. We will refer to this process – the incremental unsupervised acquisition of categories and their intensional descriptions – as *concept formation*. As with learning from examples, concept formation can be described in terms of search, and two general approaches have been explored in psychology and machine learning.

The first scheme employs a specific-to-general search, incrementally comparing each new observation to existing categories and adding it to one or more of the best-matching categories. In Kolodner's (1983) CYRUS and Lebowitz's (1982) UNIMEM, matching is a function of the number of features shared by the new observation and a given concept description. These systems generalize a concept if the match with the new observation is sufficiently good. If an observation does not match any concept to a prespecified degree then the new observation is used to create a singleton category that may be generalized with future observations. In the process, UNIMEM and CYRUS form an abstraction hierarchy of concepts that they use to classify future cases, filtering each observation through levels of the hierarchy by recursive application of the matching procedure. Both systems can be viewed as advanced versions of Feigenbaum's (1963) EPAM, which formed discrimination networks (actually trees) with tests that were restricted to single features.

One can also employ a general-to-specific strategy for concept formation, as shown by Martin's (1989) CORA system. Like its precursor STAGGER (Schlimmer & Granger, 1986), the model incrementally conjoins features, but it relies on correlations between features to trigger this chunking

2. Studies by Medin, Wattenmaker, and Michalski (1986) qualify the extent to which fit and simplicity trade against each other in human sorting tasks in which subjects have simultaneous access to all observations. Their experimental task corresponds to nonincremental, unsupervised learning.
3. Hanson and Bauer note that their system can be run in incremental mode, but it was not designed with prediction in mind.

process, rather than monitoring correct and incorrect predictions of category membership. CORA's reliance on feature correlations is similar to that used in WITT, but it descends most directly from Chalnick and Billman's (1988) work. However, whereas CORA uses observed correlations to conjoin features, its ancestor uses observed correlations to slowly generalize initially saved instances. Neither CORA or the earlier system forms an abstraction hierarchy; they simply create concepts that are conjunctions of features and that describe (possibly nondisjoint) categories.

Although approaches to concept formation may differ in search direction, they seem to universally share the hill-climbing organization of their incremental counterparts for supervised learning. As such, they may suffer from ordering effects, in that they may discover different categories depending on the order in which they process observations. The design of concept formation methods differs from that of nonincremental clustering systems, in that it is largely motivated by the realization that many real-world domains require continuous interaction with the environment. Mechanisms for concept formation are designed to be rational but resource-bounded learners (Simon, 1969). Each observation triggers small changes to the current categorical structure, although simulated forms of backtracking may be used to insure that major changes can occur over time. For example, UNIMEM deletes a node and its associated subtree if the node's corresponding concept becomes poor by a criterion similar to CLUSTER/2's fit measure. This allows a new subtree to be grown to reflect the characteristics of future data. Section 4 elaborates on some of these issues in the context of our COBWEB system.

3. Psychological Constraints on Concept Formation

The previous section touched upon psychological considerations in concept learning, but its main focus was on search as a generic framework in which to view this task. The current section considers psychological findings in greater detail, notably *typicality* and *basic-level* effects, along with their implications for the representation and formation of concepts.

3.1 Typicality Effects and Probabilistic Concepts

Smith and Medin (1981) refer to conjunctive descriptions, discussed earlier, as *classical* representations of conceptual structure. One implication of such classical representations is that all concept members are treated equally during classification, since an observation either has the requisite conjunction of features or it does not. However, experiments have repeatedly shown that human subjects do not treat concept instances equally, but regard certain members as more 'typical' than others. For example, in a *target recognition* task, subjects must determine if a test instance is a member of a target category (e.g., 'Is a *robin* a *bird*?'). Several studies (Rips, Shoben, & Smith, 1973; Rosch & Mervis, 1975) indicate that subjects consistently respond affirmatively more quickly to certain positive instances than to others. For example, they will more quickly affirm that a robin is a bird than they will affirm that a chicken is a bird. The relative ranking of positive test items corresponds to a typicality ranking of category members, and this conclusion is bolstered by results in a variety of other experimental tasks (Mervis & Rosch, 1981; Smith & Medin, 1981).

3.1.1 PROBABILISTIC CONCEPTS: INDEPENDENT CUE MODELS

Classical representations do not easily account for typicality effects, and in response, researchers have proposed a number of alternative concept representations. Rosch and Mervis (1975) made an early attempt to discover the structural determinants of typicality, finding that category members sharing features with many other members of the same category tend to be judged more typical. In addition, when a disjoint, contrasting category is involved, members that share few features with members of the contrasting category tend to be judged more typical. This sensitivity to intra-category and inter-category overlap of features is captured by their notion of *family resemblance*.

The apparent relation between family resemblance and typicality indicates the importance of feature distributions in human classification. Although classical representations cannot capture such distributional information, *probabilistic* concept representations (Smith & Medin, 1981) manage this by associating a probability, weight, or some other confidence number with each feature of a concept definition. A straightforward implementation is to store the conditional probability, $P(f|C_k)$, of each feature f's presence with respect to each category C_k; this is more commonly called the *category validity* (Medin, 1983) of the feature. Recognition or classification using probabilistic concepts usually involves summing the weights of features that are present in a new observation (Collins & Loftus, 1975; Smith, Shoben, & Rips, 1974; Smith & Medin, 1981). Classification may be based on whether this sum passes a specified threshold (Smith & Medin, 1981), as in neuron-like processing units (Nilsson, 1965; Hinton, 1989), or one may assign an observation to the category that maximizes the sum, as in Bayesian classifiers (Duda & Hart, 1973).

The probabilistic account offers an explanation of typicality effects in that typical instances will have features shared by many other members of the same category, giving them higher category validities. If one assumes that recognition time is inversely proportional to these sums, then observations with high intra-category similarity will be recognized more quickly and thus be regarded as more typical. On its own, this scheme does not explain the impact of inter-category similarity on typicality, but one can easily imagine extensions that include *cue validities* (the conditional probability of a category given a feature).

A more important limitation of this model stems from the fact that the recognition procedure is based on individual, presumably independent, category validities. For this reason, it has been called the *independent cue* model of concepts. A number of authors (Smith & Medin, 1981; Medin, 1983; Hanson & Bauer, 1989) point out that independent cue models are representationally incomplete, since summation of individual weights limits recognition to *linearly separable* categories (Nilsson, 1965). More generally, independent cue models do not capture the feature correlations that are necessary for completeness and to which humans seem naturally attuned (Mervis & Rosch, 1981; Medin, 1983).

3.1.2 ALTERNATIVES TO INDEPENDENT CUE MODELS

The apparent inability of independent cue models to capture bundles of correlated features has led to a number of alternative models. One way to keep track of feature correlations is simply to remember instances of a concept, since each instance can be viewed as a maximally-specific

conjunction of features. This is the approach taken in *exemplar* representations (Smith & Medin, 1981). An example of this approach is Reed's (1972) *proximity* model, which retains an extensional listing of a concept's known members, classifying a new object as a member of a category, C_k, if it matches another member of C_k more closely than a member of a contrasting category.

A disadvantage of the proximity model is that retaining an extensional listing of known category instances becomes expensive as the number of observations grows. In response, some systems (Aha & Kibler, 1989) selectively retain only certain useful observations. A simple strategy is to retain only observations that resulted in a misclassification during learning.[4] Computational experiments demonstrate this strategy's advantage in terms of storage, but they also show accuracy benefits, presumably because idiosyncratic observations are ignored and thus are not used in classification.

In contrast to selective retention, Medin and Schaffer's (1978) *context* model supports a form of abstraction through selective attention. In particular, the model allows that a subject may not attend to a feature, effectively dropping the feature from an observation. Classification assumes that a new instance matches in parallel against the stored exemplars of each contrasting category, causing sufficiently matching exemplars to be retrieved; an assumption is that an exemplar is retrieved with a probability proportional to the degree that it matches the observation. An observation is classified with the first concept for which a specified number of exemplars is retrieved. Presumably, the context model would account for typicality effects, since new typical instances would more closely match the typical observations currently stored; thus, a criterial number of retrieved exemplars would tend to be reached more quickly for typical instances.

Nosofsky's (1987) *generalized context* model extends ideas of selective attention by allowing features to be weighted. Aha and McNulty (1989) demonstrate how these weights can be learned in a supervised task. Feature weights serve to divert attention away from uninformative features – those distributed across members of many categories – and focus attention on informative features in classification. This variable treatment of features can capture the importance of intra-category and inter-category overlap, and adaptations should model a variety of typicality effects.[5]

Another alternative class of models assume a *relational cue* representation, which generalizes on the independent cue approach. Like their precursors, relational cue models maintain probabilities (weights, confidence values) for individual features in concept descriptions, but they also permit joint probabilities of larger feature configurations, such as $P(\text{Color=red} \wedge \text{Size=large} \wedge \text{Shape=sphere} \mid C_k)$. Syntactically, these models are also generalizations of exemplar models, since an instance can be viewed as a conjunction of features. However, the representational power of exemplar and relational cue models are theoretically equivalent, since one can use stored exemplars, as needed, to compute all the information used in relational cue models.

4. All of the exemplar models that we have reviewed assumes a supervised learning scenario, but similar best-match procedures for classification are used in the single-linkage clustering methods that we discussed in Section 2.2 (Everitt, 1980).

5. Feature weighting appears to serve a purpose that is similar to reasons for feature weights in independent cue models, but exemplar models do not force a category to be represented by a single summary description as do single independent cue concepts.

Table 1. Linearly separable and nonlinearly separable categories (Medin, 1983).

		CATEGORY C_1					CATEGORY C_2			
		V_1	V_2	V_3	V_4		V_1	V_2	V_3	V_4
LINEARLY SEPARABLE OBJECTS	1)	1	1	1	0	5)	1	0	1	0
	2)	1	0	1	1	6)	0	1	1	0
	3)	1	1	0	1	7)	0	0	0	1
	4)	0	1	1	1	8)	1	1	0	0
NON-LINEARLY SEPARABLE OBJECTS	9)	1	0	0	0	13)	0	0	0	1
	10)	1	0	1	0	14)	0	1	0	0
	11)	1	1	1	1	15)	1	0	1	1
	12)	0	1	1	1	16)	0	0	0	0

One example of a relational cue model is Hayes-Roth and Hayes-Roth's (1977) *property-set* model, which supposes that a feature conjunction is stored with a count of the observations in which it occurred. A new observation is classified with the concept that contains the most 'diagnostic' conjunction of features (i.e., the combination with the highest cue validity). The feature conjunction for which $P(C_k|(\text{conjunction}))$ is maximized (over all C_k) dictates that an observation that satisfies the conjunction should be classified as a member of C_k. The property-set model stores the frequencies needed to compute cue validities (rather than category validities) for all single features and conjunctions of features.

In many cases, a feature combination may be useless for classification; trivially, if *small* objects are equally split between two classes, then smallness alone will give no help in classification. Thus, a reasonable storage strategy would throw out feature conjunctions that do not aid classification (e.g., those with cue validities that are roughly equal for all categories). This strategy has been used for supervised learning by Anderson and Kline's (1979) ACT and by Schlimmer and Granger's (1986) STAGGER, whereas Chalnick and Billman (1988) have used relational cue representations for concept formation. The latter system removes features that do not add to the informativeness of a composite feature. Martin (1989) takes the opposite approach, adding features to a conjunction only if they add to the informativeness of the conjunction.

3.1.3 PROBABILISTIC CONCEPT HIERARCHIES

Exemplar and relational cue models both address a purported weakness of independent cue models – their inability to explicitly capture correlations between features. However, as we will show, this limitation does not apply to *network* of independent cue representations. A combination of such concepts has the same representational power as exemplar and relational cue models. Similar completeness arguments occur in the literature on neural computing (Nilsson, 1965), where networks of simple classifiers (e.g., linear threshold units) can achieve representational completeness, even though their components are severely limited.

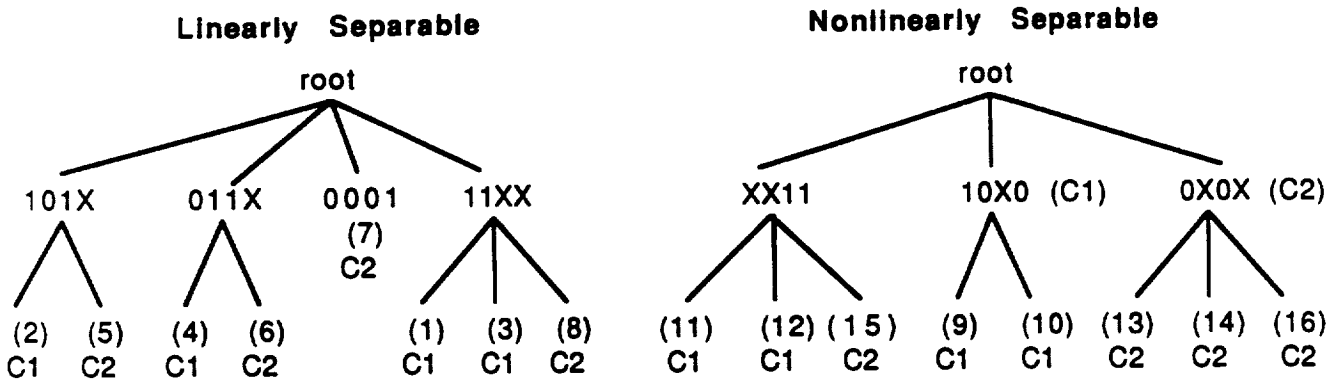

Figure 1. Concept trees over nonlinearly and linearly separable categories.

As we have noted, concepts that are represented using independent cues are individually limited to the recognition of linearly separable categories. Medin (1983) suggests that if independent cue models are the basis of human conceptual structure, then in some cases linearly separable categories should be easier to learn than nonlinearly separable ones. An investigation into this question required that subjects learn, under supervision, one of the two category pairs displayed in Table 1. Observations were characterized in terms of four binary-valued attributes, A_1 through A_4. Subjects judged the linearly separable set more difficult to learn, and this set also resulted in more recognition errors. Gluck, Bower, and Hee (1989) have accounted for similar data with a relational cue model in which pairwise composite cues are used to convert the nonlinearly separable categories to linearly separable categories. This transformation makes the concept easier to learn in composite-cue space than the original linearly separable categories.

An alternative account of these data exploits the notion of probabilistic concept *hierarchies*. Consider the concept trees of Figure 1, which discriminate the category pairs of Medin's experiments.[6] An independent cue model insists that each node divides the total set of observations into linearly separable categories. However, this division need not correspond to the sets that were taught, C_1 and C_2. Rather, like a decision tree (Quinlan, 1986) or discrimination network (Feigenbaum, 1963; Kolodner, 1983; Feigenbaum & Simon, 1984), members of a given class may reside in distinct portions of the hierarchy.

One can think of tree construction as being guided by the simple heuristic of grouping objects having the most features in common. The actual method used to form the hierarchies in the figure is more complicated (as described in Section 3.2), but the simplification is consistent with this technique and with intuitions about independent cue representations. The trees reveal that several atypical members of C_1 in the linearly separable set share many properties with C_2 and vice versa. Thus, these similar items are reasonably placed within the same middle-level nodes of the hierarchy. Observation 7 is quite unlike any other instance, placing it in a separate category. On the other hand, there are fairly specific patterns that perfectly discriminate many members of contrasting

6. For simplicity, the concepts for the nodes in Figure 1 are abbreviated by a pattern (e.g., 101X) that is common to all category (node) members; 'X' denotes an attribute in which no single value is common to all members.

categories in the nonlinear domain. Medin's finding can be explained in terms of the average depth to which observations must be classified before one can perfectly distinguish members of C_1 from C_2. The linearly separable set requires an average depth of 1.87 before reaching a node that contains only members of one category; in contrast, the nonlinearly separable set has 1.37 as its average depth.

Our demonstration is simplified, but it nonetheless illustrates that hierarchies or other networks of independent cue concepts have the same representational power as exemplar and relational cue models. Linearly separable representations direct classification to deeper levels of the tree until a perfect discrimination can be made. In addition to their representational strength, hierarchies offer efficiency advantages. A tree structure allows recognition to occur in logarithmic time as a function of stored observations, rather than in linear or exponential time, as it does for some alternatives. We now turn our attention to heuristics for guiding the formation of such concept hierarchies, focusing on the evidence for preferred concepts in human memory.

3.2 Basic-Level Effects and Concept Quality

Psychological studies have shown that, within hierarchical classification schemes, there appears to be a *basic* level preferred by human subjects. For example, in a hierarchy containing {animal, vertebrate, mammal, dog, collie}, subject behavior may indicate that 'dog' lies at the basic level. Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976) used a target recognition task to show that subjects are quicker to confirm that a test item is a member of such a basic category than they are for a superordinate or subordinate category. In a forced naming task (Rosch et al., 1976; Jolicoeur, Gluck, & Kosslyn, 1984), a subject is shown a picture of a particular item and asked to respond with its identity.

3.2.1 Early Measures for Predicting the Basic Level

The identification of preferred concepts in humans must constrain any model of human classification, and it may also provide a basis for principled measures of concept quality for use in concept formation by human and machine. In fact, researchers have proposed a number of measures designed to predict the basic level. An early proposal (Rosch et al., 1976) postulated that a basic-level category maximizes the *total cue validity* of a category over features that are shared by *all* or *most* members of the category (i.e., only features with high category validity). This can be stated formally as

$$\sum_j P(C_k|V_j) \text{ for features } V_j \text{ such that } P(V_j|C_k) \approx 1.0 \quad .$$

Rosch et al. did not specify how close to unity a category validity must come before it is included in the calculation of total cue validity.

Jones (1983) has proposed another measure, called *collocation*, that directly incorporates category validity into the prediction of basic level. This function can be stated as

$$\text{collocation}(V_j, C_k) = P(C_k|V_j)P(V_j|C_k) \quad .$$

He argued that a basic-level node (e.g., bird) has more collocation-maximizing values among its ancestral-related nodes (e.g., animal, robin) than concepts at other levels. Neither Rosch nor Jones compared their measures' predictions against experimental results, but both suggested that the basic level maximizes a tradeoff between cue and category validities over descriptive features.

3.2.2 CATEGORY UTILITY

The notion of a tradeoff is also important to a third measure that has been proposed to predict basic-level categories (Corter & Gluck, 1985; Gluck & Corter, 1985). This function, called *category utility*, can be developed from a weighted variation on the collocation measure:

$$\sum_j P(V_j)P(C_k|V_j)P(V_j|C_k) \quad , \tag{1}$$

where $P(V_j)$ weights the contribution of individual feature collocations by the base rate of the respective feature. In essence, this measure reflects the importance of increasing cue and category validities for more frequently occurring features.

However, Corter and Gluck did not express category utility as an extension to collocation. Rather, they devised it with the idea that basic-level categories are preferred because they best facilitate predictions about observations in the environment. In their view, category utility is a function of a category's prediction potential, or

$$P(C_k)E(\text{number of correctly predicted } V_j|C_k) \quad ,$$

which is a tradeoff between the *expected* number of features that can be correctly predicted about a member of a category C_k and the proportion of the environment $P(C_k)$ to which those predictions apply.

Assuming a *probability matching* strategy for prediction (Bruner, Goodnow, & Austin, 1956), the expectation can be further formalized by noting that one can predict a feature with probability $P(V_j|C_k)$, and that this prediction will be correct with the same probability:

$$P(C_k)\sum_j P(V_j|C_k)^2 \quad . \tag{2}$$

Clearly, a *probability maximizing* strategy (Bruner, Goodnow, & Austin, 1956) has advantages in actually generating predictions. However, it is important to realize that it is *not* superior in terms of heuristically *ordering* categories in terms of prediction potential, which is the intent behind category utility. In fact, there are important advantages to assuming a probability-matching strategy when forming categories, as detailed by Fisher (1987a).

Simple algebraic manipulations show the equivalence of functions (1) and (2). Thus, category utility can be viewed as a tradeoff between cue and category validity, as well as a function that measures a category's prediction potential. More intuitively, these views can be unified by noting that the $P(V_j|C_k)$ term reflects the importance of categories with *predictable* features (Tversky,

1977; Lebowitz, 1982; Kolodner, 1983), but that features must also be *predictive* or discriminating of a category (Tversky, 1977; Lebowitz, 1982; Kolodner, 1983), so that a one can classify an instance and access predictable features. Finally, Corter and Gluck (1985) define category utility as the *increase* in the expected number of features that can be correctly predicted, given knowledge of a category, over the expected number of correct predictions without such knowledge. The expression

$$CU(C_k) = P(C_k)[\sum_j P(V_j|C_k)^2 - \sum_j P(V_j)^2] \tag{3}$$

provides a formal statement of their complete definition of the category utility CU.

3.2.3 Properties of Category Utility

There are several properties of category utility that are worth mentioning at this point. First, the measure has the desirable property that it will be zero if all feature distributions are independent of membership in a category. That is, if

$$P(V_j|C_k) = P(V_j) \quad ,$$

then

$$P(V_j|C_k)^2 - P(V_j)^2 = 0 \quad ,$$

and V_j will be 'irrelevant' to a category's score and presumably to an observation's membership in C_k. If all such features are independent, then $CU(C_k) = 0$.

Second, category utility is not a function of feature correlations, but categories that capture feature correlations will tend to have higher scores for this measure. If category C_k captures a correlation between N features, then the sum of the individual category validities will be higher than if the correlation is captured only in part or not at all. This property has important implications for the process of concept formation, in that it lets one capture feature intercorrelations without their 'direct' computation. Rather than computing $P(V_1 \wedge V_2 \wedge \ldots \wedge V_n)$ explicitly, concept formation can introduce a category C_k that converts the task of computing $P(V_1 \wedge V_2 \wedge \ldots \wedge V_n|C_k)$ to one of computing $P(C_k) \prod_{i=1}^{n} P(V_i|C_k)$. In words, C_k is an auxiliary variable that may lead to conditional independence among some features (Pearl, 1985).

To see this point, consider the hierarchies of Figure 1, which were formed using category utility as a decision heuristic. Note that the term $P(V_i|C_k)$ equals 1.0 for those features shown within each middle-level category. Trivially, the distributions of these features are conditionally independent of other features within the same class. Nodes in the tree tend to capture distinct sets of correlated features; the probability of each conjunction of features shown at a node is simply the probability of the category, $P(C_k)$, since $\prod_{i=1}^{n} P(V_i|C_k) = \prod_{i=1}^{n} 1.0 = 1.0$. These computations may not be so clean in other domains, but one can nonetheless efficiently and effectively compute such feature correlations through the interaction of concept hierarchies and an independent cue heuristic.

3.3 Summary

To summarize, psychological findings indicate that there are important constraints on the representation and access of concepts. In particular, typicality effects suggest that classical concept representations are untenable in many situations, since some category members receive preferential treatment. We have advanced probabilistic concept hierarchies as a representation scheme that supports these preferences, in which features vary in their contribution to family resemblance and classification. Furthermore, tree-structured probabilistic concepts are representationally complete; they do not suffer from the limitations of independent cue concepts in isolation, such as a restriction to linearly separable categories.

In addition to intra-category preferences implied by typicality rankings, basic-level effects suggest that humans also give preferential treatment to certain categories over others. These preferences can be predicted in static memory structures by measures like category utility. However, these same human preferences undoubtedly play a significant role in concept learning, as well as retrieval. This supposition is supported by studies (Rosch et al., 1976) which indicate that basic-level categories are learned before either subordinate or superordinate categories. We now describe the manner in which predictors of human categorization preferences can be adapted to the task of concept learning and classification.

4. A Model of Concept Retrieval and Learning

In this section we describe COBWEB (Fisher, 1987a, 1987b), a concept formation system that adapts category utility to the task of concept learning and recognition. Our initial motivation for using category utility was that it rewards categories that improve prediction, a characteristic made evident by Gluck and Corter's analysis. Thus, this section's perspective is primarily computational, but rational (Anderson, in press) and speculative (Hall & Kibler, 1985) analyses posit that computational and psychological concerns are not independent. In Section 5, we expand our discussion to selected psychological findings.

We will describe COBWEB in terms of the search framework that we presented earlier. Conveniently, one can easily transform category utility from a characteristic function of static concept hierarchies to a heuristic guide for concept learning. In particular, one can partition a known set of observations into contrasting categories, C_k, so as to maximize the average utility of categories in the partition or

$$\frac{\sum_{k=1}^{n} CU(C_k)}{n},$$

where n is the number of categories in the partition. Because category utility requires only information about individual feature distributions within each C_k, one can effectively represent a category with an independent cue representation, where each feature, V_j, is weighted by $P(V_j|C_k)$.

Figure 2 illustrates that contrasting categories can be organized under a root node whose features are weighted by applicable base rate probabilities, $P(V_j|\text{root}) = P(V_j)$. In this case, observations correspond to the voting records of U.S. congresspersons on key issues with values of 'yea' or

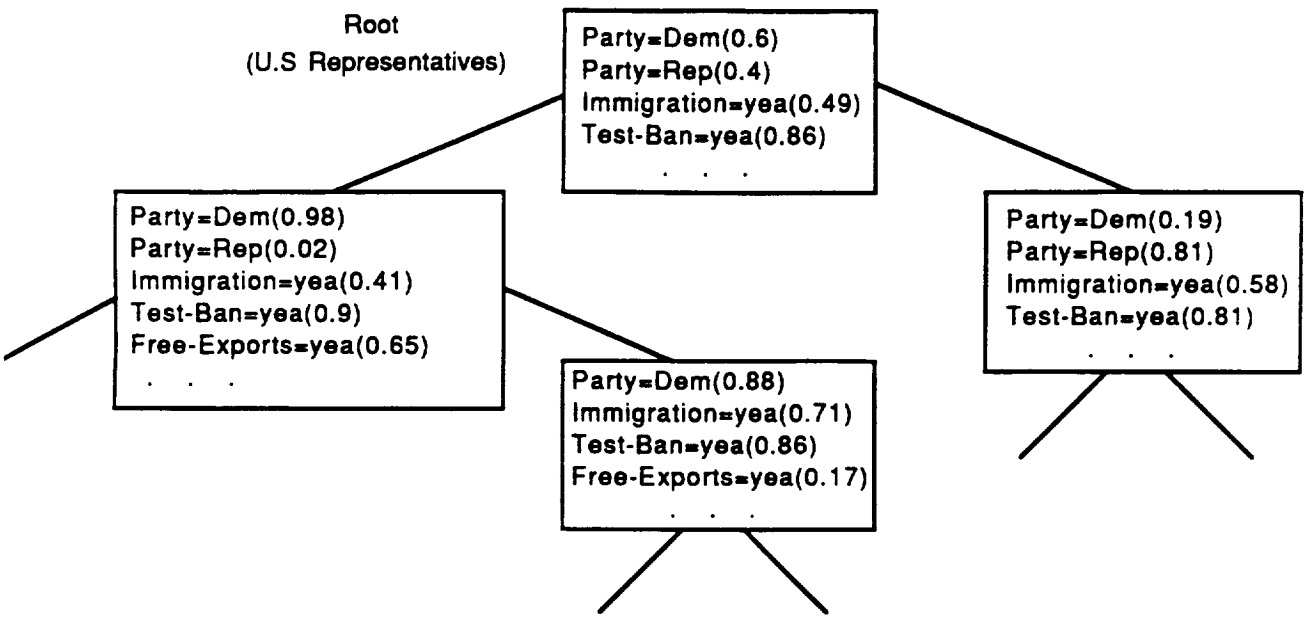

Figure 2. A sample probabilistic concept hierarchy over congressional voting records.

'nea' (Lebowitz, 1987).[7] In addition, we assume that each category is weighted by the proportion of observations, $P(C_k)$, classified under it. By definition $P(\text{root}) = 1.0$. Collectively, $P(C_k)$'s, $P(V_j)$'s, and $P(V_j|C_k)$'s supply the requisite information for calculating category utility.

Conceptually, the easiest way to find an optimal set of contrasting categories is to exhaustively search the possible partitions of the known observations. This can proceed in a manner similar to the specific-to-general search that we described earlier: given a partition over m observations, consideration of the $m + 1$st observation generates m new partitions, each the result of placing the observation into one of the existing categories. In addition, there is an $m + 1$st partition that results from creating a new singleton category that contains only the new observation. The search for the best partition ceases when one encounters the last observation; one can then identify the partition with the best average category utility from among the alternatives. At this point, one can simply return the best partition or one may further decompose each category of the best partition by recursively applying the exhaustive search procedure over the subset of observations that are classified by the category. This recursive procedure results in a tree of probabilistic concepts.

This exhaustive approach is clearly impractical, since the procedure requires that one examine alternative partitions that grow exponentially with the number of observations. Search is reduced significantly in systems like CLUSTER/2 by maintaining a fixed number of alternatives after each observation. The hill-climbing approach described in Section 2.3 restricts the number of alternative partitions that are maintained to one. In particular, Fisher's (1987a, 1987b) COBWEB assimilates an $m + 1$st observation by evaluating the partitions that result by adding the observation to each existing category and the partition that results from creating a new singleton category. It then

7. We only list 'yea' values on selected votes, but all features with nonzero probability at a node are stored at the node.

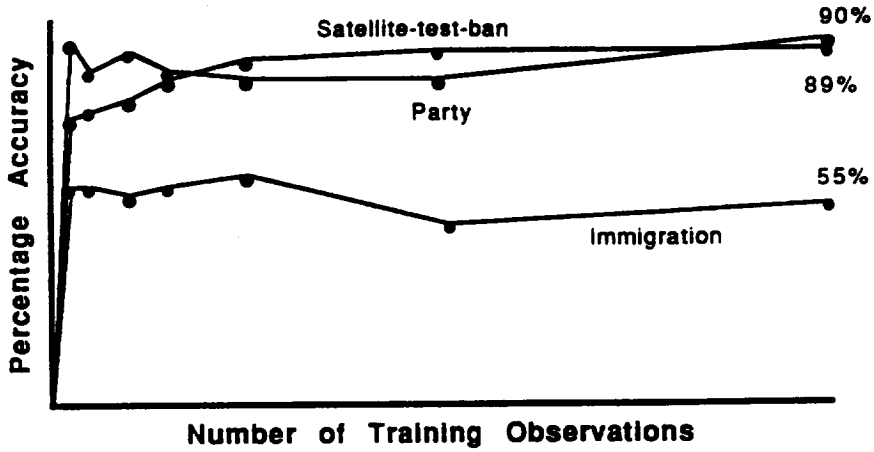

Figure 3. Learning curves for three attributes in the congressional domain.

evaluates each of these alternatives using category utility and retains the best choice. If the instance is incorporated into an existing category, then the observation is assimilated into the respective subtree by the same procedure.[8] Anderson (in press) has recently described a similar approach, in which a Bayesian measure guides the incremental assimilation of new observations.

As with assimilation, COBWEB also uses category utility to guide object recognition: an observation is sorted down a path of 'best matching' categories to a leaf, at which point the new observation may be recognized as matching the leaf. For example, consider the congressional voting records classified by the tree of Figure 2. A new voting record of unknown political party may be recognized (perhaps incorrectly) as an instance of the political party of the best-matching leaf. As we discussed in relation to linearly separable categories, this strategy lets category members be distributed throughout the tree, and not restricted to one node of the tree.

Recognition need not be limited to any particular category label (e.g., party), so that one can predict any unknown feature in this manner. This capability can be tested systematically by measuring predictive accuracy at intermittent points in the evolution of a probabilistic tree. The system is presented each 'test' item with one or more attributes removed, it sorts the incomplete observation to the 'best-matching' leaf of the concept hierarchy, and it predicts the missing attributes based on those in the leaf. This occurs for all attributes of all test items, thus yielding an accuracy level for each attribute. The graph of Figure 3 shows sample 'learning curves' for the attributes POLITICAL PARTY, IMMIGRATION-VOTE, and SATELLITE-TEST-BAN.

In general, prediction accuracy for an attribute is closely related to the attribute's inter-correlation with other attributes of the domain. Political party is highly correlated with other attributes, whereas a congressman's vote on an immigration bill is relatively uncorrelated with other features. These data support earlier claims that category utility captures correlations in the data when coupled with appropriate learning mechanisms. Similar findings hold for other natural domains and for artificial domains in which one can systematically vary the amount of intercorrelation (Fisher,

8. COBWEB only handles nominally-valued attributes, but Gennari, Langley, and Fisher (1989) describe CLASSIT, a descendent of COBWEB that assumes continuously-valued attributes.

1987a, 1987b; Gennari, Langley, & Fisher, 1989). Not surprisingly, in domains with very little inter-correlation, the learning rate and the asymptotic accuracy suffers greatly. For some features, the system's predictive ability may even be worse than chance.

The reason for COBWEB's poor behavior with respect to some features is that classification to a leaf often simulates a *probability-matching* strategy (Bruner, Goodnow, & Austin, 1956). Viewed in statistical terms, sorting to a leaf may overfit the data. Recall from Section 3.2 that category utility has the desirable property that features which are independent of category membership will not influence classification at deeper levels, since $P(V_j|C_k)^2 - P(V_j)^2 = 0$. Inversely, an attribute's independence should also signal that deeper classification will not aid prediction of the attribute. Thus, one should follow a probability maximizing strategy at an appropriate point in classification. Several heuristics for identifying points of approximate feature independence and points of optimal prediction (Quinlan, 1986; Fisher, 1989) have produced significant advantages in terms of prediction accuracy.

Our summary of COBWEB has been brief, in part because the precise nature of the learning operators is of limited relevance to the forthcoming discussion. Rather, the important assumptions are that memory is organized into probabilistic concept hierarchies and in a manner that is guided by category utility. This section has illustrated that one can carry out the process incrementally and in a manner that seems consistent with many aspects of human learning (Simon, 1969; Langley, Gennari, & Iba, 1987; Anderson, in press). However, as described here, COBWEB's hill-climbing learning method exhibits ordering effects that we have detailed elsewhere, along with simulated backtracking mechanisms that mitigate the effect (Fisher, 1987a, 1987b; Gennari, Langley, & Fisher, 1989). Finally, our evaluation of COBWEB has been in terms of prediction accuracy of features and category labels. This is an important evaluation criterion in machine learning, but one that is intimately related to the psychological literature on recognition (e.g., Feigenbaum, 1963). We will now investigate the psychological plausibility of our methods for recognition, classification, and prediction.

5. An Analysis of Memory Phenomena

In this section we extend our analysis of COBWEB and category utility to a number of psychological phenomena. Our discussion is very much in line with Anderson's (in press) rational analysis of cognition. In effect, Gluck and Corter's derivation of category utility stemmed from the prescription that categories facilitate accurate prediction. We open the section by introducing some conventions that are important in our analysis of the basic level, typicality, and fan effect data that follow.

5.1 Category Match

A common thread in each of the psychological studies that we examine is the use of subject response time to queries about experimental stimuli. For example, subjects might be required to verify that a stimulus is a member of a previously learned category. This section illustrates that response time

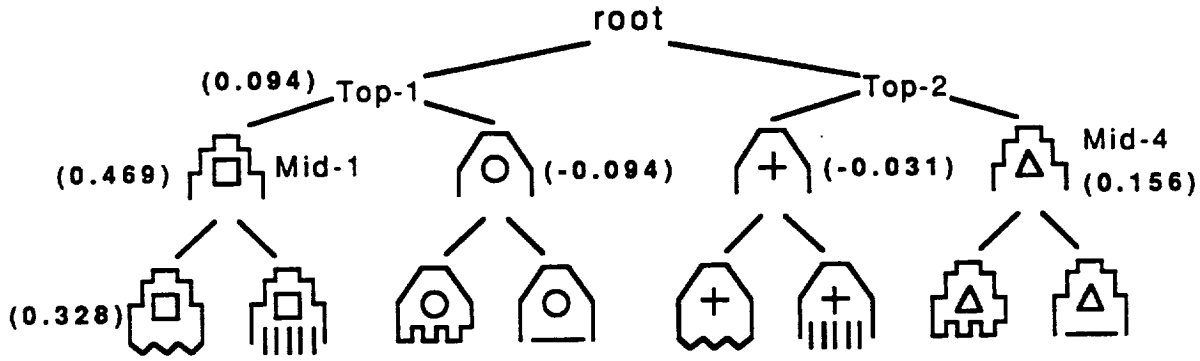

Figure 4. Approximation of a tree from Hoffman and Ziessler's (1983) basic-level studies.

in each of these studies is well predicted by a simple variation of category utility that is only a function of the features occurring in the observation (stimulus) being classified:

$$P(C_k)\sum_j [P(V_j|C_k)^2 - P(V_j)^2] \quad ,$$

for V_j present in the observation. We will call this *category match* because it intuitively corresponds to the degree that an observation matches a category and the extent to which that category is activated during recognition. The measure also fits Tversky's (1977) model of category-object resemblance, in that it is a function of category size ($P(C_k)$) and sums over the features that 'agree' and 'conflict'.[9] If the amount of conflict between features outweighs the amount of agreement then category match will be negative; trivially, if an observation has a feature that is not present in any category member, then $P(V|C) = 0.0$ and $P(V|C)^2 - P(V)^2 < 0$.

Intuitively, category match corresponds to activation strength, which we might assume to be inversely related to response time (Collins & Loftus, 1975). However, this section will concentrate primarily on the predictive (or descriptive) links between category match and human response time. That category match turns out to be a good predictor of response time is not strongly tied to particular 'implementation' details (e.g., the precise nature of the classification procedure), but relies only on general assumptions about memory organization: that probabilistic concepts are hierarchically organized in a manner that is guided by category utility.[10] There will be exceptions to this section's exclusion of 'implementation' detail, but only on occasions when it seems most productive to explain counterintuitive findings. We will more thoroughly discuss how the predictions

9. Most often there will not be perfect agreement or conflict between an observation's feature and a category's feature distributions. Rather, the amount of agreement and conflict is weighted. Hadzikadic and Yun (1989) use a measure of similar intent in their INC system, but relevance is only computed over features shared by the observation and class. Computer experiments by Fisher (1987) indicate that classification and learning behavior using category match closely approximate the behavior of the full category utility measure, with differences only occurring very early in training.

10. In fact, this is a desirable characteristic that is enforced by methodologies for system development that segregate stages of *specification*, *design*, and *implementation*. Within cognitive science, these stages are roughly analogous to Marr's (1982) three levels of description for information processing systems.

Table 2. The encoded tree from Hoffman and Ziessler (1983).

SUPERORDINATE	BASIC LEVEL	SUBORDINATE	OUTER	INSIDE	BOTTOM
TOP-1	MIDDLE-1	LEAF-1	0	0	0
		LEAF-2	0	0	1
	MIDDLE-2	LEAF-3	1	2	2
		LEAF-4	1	2	3
TOP-2	MIDDLE-3	LEAF-5	1	3	0
		LEAF-6	1	3	1
	MIDDLE-4	LEAF-7	0	4	2
		LEAF-8	0	4	3

of category match can be implemented by the classification mechanisms of a COBWEB-like system in Section 6, but for now we turn our attention to the predictive merits of category match with respect to basic level, typicality, and fan effect phenomena.

5.2 Basic-Level Effects

Gluck and Corter (1985) verified that category utility predicted the basic level in two experimental studies (Hoffman & Ziessler, 1983; Murphy & Smith, 1982): a basic-level category maximizes category utility among its ancestors and descendents. In a study by Hoffman and Ziessler, subjects learned a classification tree over 'nonsense' objects like the one shown in Figure 4. Each category (node) had a 'nonsense' name that subjects used to identify category membership in recognition tasks. Objects were defined in terms of three attributes: the shape of the INSIDE subcomponent with values SQUARE, TRIANGLE, STAR, or CIRCLE (encoded as 0, 1, 2, and 3, respectively); the OUTER shape, with (encoded) values of 0 and 1; and the shape of the BOTTOM, with values 0, 1, 2, and 3. Table 2 shows the encoding of the Hoffman and Ziessler data that was assumed by Corter and Gluck. For the tree of Figure 4, subjects consistently 'preferred' level two (where the root is at level zero).

To account for the order in which subjects verify category membership, we use the category match measure. Figure 4 shows the match scores of several categories (nodes) obtained for the observation {OUTER = 0, BOTTOM = 0, INSIDE = 0}. The appropriate basic-level category is the most highly rated, with category match indicating a negative score for some categories for which the observation is not a member. Intuitively, a negative score indicates that an observation and a category's feature distributions conflict more than they agree. This simulation assumes that classification occurs with respect to the tree that subjects are explicitly taught, and that a verbal indication of the target category activates a corresponding node in the tree. When classification via the perceptual cues of a pictured observation reach the verbally signified node, the observation is identified as a member of the target.

Table 3. Our encoding of the Murphy and Smith (1982) tree.

SUPERORDINATE	BASIC LEVEL	SUBORDINATE	HANDLE	SHAFT	HEAD	SIZE
TOP-1	MIDDLE-1	SUB-1	2	2	0	0, 1
		SUB-2	2	2	1	0, 1
	MIDDLE-2	SUB-3	0	3	3	0, 1
		SUB-4	1	3	3	0, 1
TOP-2	MIDDLE-3	SUB-5	3	4	4	0, 1
		SUB-6	3	4	5	0, 1
	MIDDLE-4	SUB-7	4	0	6	0, 1
		SUB-8	4	1	6	0, 1

Hoffman and Ziessler also explored two other trees over the same objects of Figure 4. One variant resulted by placing nodes Middle-1 and Middle-4 under the same top-level node and Middle-2 and Middle-3 under the same top node. In this variant subjects treated the top nodes as basic, but category match predicts a tie between the top and middle nodes in this tree – middle and top nodes each match their respective observations with a score of 0.469. This is similar to the predictions found by Gluck and Corter with the full category utility measure. In Section 6 we speculate on a resolution to this tie that involves selectively 'masking' uninformative features in the category match computation. A third tree was also used by Hoffman and Ziessler in which subjects regarded the bottommost level of leaves to be basic. As with the tree of Figure 4, the basic level is unambiguously identified by category match.

Gluck and Corter also evaluated category utility in light of experiments by Murphy and Smith (1982). Once again, in this study subjects were trained to recognize instances of categories arranged hierarchically. In these experiments objects were abstract 'tools' that varied along four perceptual dimensions (tool size and the types of handle, shaft, and head). Categories were assigned nonsense names of equal length, and target recognition studies behaviorally identified one level as basic. In addition, Murphy and Smith also looked at 'false' cases, in which an observation was not a member of the given target. In each of the false cases, a test item from a different superordinate category than the target concept was selected. Data from the true cases support previous findings on basic-level preference, but they found that subjects showed some tendency, although not statistically significant, to more quickly reject the 'false' cases as members of subordinate target categories than basic targets.

Table 4 summarizes the average subject response times and category match scores in the true and false cases. In the true cases we report category match of an observation with each category to which it belongs. In the false case, the match between an observation and unrelated superordinate, basic, and subordinate targets are reported. In both the true and false cases, category match correctly predicts response time orders: as category match increases response time decreases. In the false cases all category match scores are negative because there is *no* feature overlap between a test observation and target concept. In fact, the difference, $P(V|C)^2 - P(V)^2$, is equal for all

Table 4. Average response times (Murphy & Smith, 1982) and category match rankings.

	TRUE CASES		FALSE CASES	
	RESPONSE TIME	CATEGORY MATCH	RESPONSE TIME	CATEGORY MATCH
SUPERORDINATE	879MS	0.21	882MS	−0.070
BASIC LEVEL	678MS	0.53	714MS	−0.035
SUBORDINATE	723MS	0.36	691MS	−0.018

categories in the false case; the difference in false category match scores is due solely to the $P(C)$ term of category match, which magnifies the negative difference for higher level categories. We have no strong hypothesis regarding these data, other than to suggest that when no featural connections exist between an observation and a category, as is the case here, it is reasonable to assume that any search of the category membership proceeds in time proportional to the category's size as reflected in $P(C)$.

Murphy and Smith performed a second experiment intended to expand their findings about subordinate recognition in the false case. In particular, they reported response times for cases in which the observation was not a member of the subordinate target, but (1) was a member of the same basic category, (2) was a member of the same superordinate category, but not the same basic category, and (3) was not a member of the same superordinate category. These cases vary the 'relatedness' of the target and observation, with (1) being the most related of the false cases and case (3) being totally unrelated. Table 5 shows the response times, which indicate that items of the same basic category require greater time to reject than the other two cases.

To explain their findings, Murphy and Smith propose a *preparation* model of classification and category structure. In this model, a verbal cue activates the target category and its 'conceptual' definition. Recognition occurs by summing the number of concept features that match an observation, as well as the number of conflicting features. An observation is accepted or rejected as a category member when a concept-specific 'threshold' is reached. Separate thresholds are presumed for acceptance and rejection. Like the preparation model, our application of category match is effectively a summing procedure. However, our category match data and our earlier discussion of classification in systems such as COBWEB suggests a different view of true and false recognition.

To motivate our processing assumptions consider the category match scores in Table 5. The negative category match scores accurately predict no difference between the unrelated and same-superordinate case, but a positive score is shown for the same-basic condition. This violates our assumption that category match and response time are inversely related, since subordinate rejection required the longest response time. To maintain consistency we must assume that match scores on opposite sides of zero are inverted in their relation to response time. In addition, we posit that a category match of zero (or less) may be regarded as a *category-independent* cause for rejecting an

Table 5. Average response times (Murphy & Smith, 1982) and category match rankings on false data with varying degrees of relatedness.

	RESPONSE TIME	CATEGORY MATCH
DISTINCT SUPERORDINATE	691MS	−0.018
SAME SUPERORDINATE	687MS	−0.018
SAME BASIC	902MS	+0.232

observation. Intuitively, this is desirable because a negative score suggests greater mismatch than match of features.

Conversely, we are also concerned with criteria for successful classification. Systems like COBWEB assume that an observation is classified with the category that *maximizes* activation. We assumed in discussing the Hoffman and Ziessler studies that successful target recognition occurred when activation that was triggered by perceptual cues reached a verbally-activated target node; this suggests that recognition is not simply a process of direct comparison between target definition and observation as the preparation model suggests, but is mediated by other memory elements. Our views of success and failure in recognition are unifiable when one considers that categories along the path to a target may induce conditional independence with respect to features that are common to all or many subordinates. Category match's subtraction of base rate probabilities insures that such features have no impact on classification, but an observation's remaining features may conflict with a concept and result in a negative match at that level of classification. Thus, our view is that the best matching node in memory defines a variable 'threshold' that cannot be achieved by any contrasting category. Competing categories are removed as candidates when their conditional matches drop below a category-independent threshold of zero. Our analysis predicts that the subordinate category of Table 5 is rejected more slowly because its positive score requires that it 'compete' with contrasting categories for some period of time.

In summary, a qualitative characterization that captures all of the false response times (i.e., Tables 4 and 5) is that they vary proportionally to the absolute value of category match. In contrast, response times for true cases vary inversely with category match. More generally, we suggest that category-specific thresholds are not required in the false or true cases. Instead, positive and maximizing activation strength may be the sole determinant of categorization.

5.3 Typicality Effects

Our discussion has replicated Gluck and Corter's analysis of basic-level effects with category match and extended it to Murphy and Smith's 'false' data. In this section we extend their rational analysis of category structure further by using category match to predict typicality effects in probabilistic trees. To review briefly, typicality studies indicate that, in addition to the between-category preferences suggested by basic-level effects, humans also exhibit preferences within categories.

Table 6. Nonsense strings used by Rosch and Mervis (1975) to test typicality differences.

	(A)				(B)		
	LETTER STRING	INTRA-CATEG. OVERLAP	TYPICALITY		LETTER STRING	INTER-CATEG. OVERLAP	TYPICALITY
	JXPHM	LOW	LOW		HPNWD	LOW	HIGH
	QBLFS	LOW	LOW		HPC6B	LOW	HIGH
A	XPHMQ	MEDIUM	MEDIUM	A	HPNSJ	MEDIUM	MEDIUM
	MQBLF	MEDIUM	MEDIUM		4KC6D	MEDIUM	MEDIUM
	PHMQB	HIGH	HIGH		GKNTJ	HIGH	LOW
	HMQBL	HIGH	HIGH		4KCTG	HIGH	LOW
	CTRVG				8SJKT		
	TRVGZ				8SJ3G		
B	RVGZK			B	9UJCG		
	VGZKD				4UZC9		
	GZKDW				4UZRT		
	ZKDWN				MSZR5		

5.3.1 TYPICALITY AND INTRA-CATEGORY SIMILARITY

To demonstrate consistency with typicality effects, we will focus on studies by Rosch and Mervis (1975). Their experiments demonstrate that typicality increases with the number of features shared with other objects of the same category and varies inversely with the number of features shared with members of contrasting classes. In their study of intra-category influences, Rosch and Mervis used the 'nonsense' strings in Table 6 (a). Members of category A varied in the extent that they overlap with other members of the same class. For example, the symbols of 'QBLFS' appeared in an average of 2.0 other strings of category A, whereas the symbols of 'HMQBL' were shared by 3.2 other members of category A. The *inter-class* overlap between members of A and B was held constant (i.e., there was no overlap). After subjects learned to distinguish categories A and B, the average time to classify letter strings as members of A or B was determined. Response time decreased as the amount of intra-category overlap increased, supporting the hypothesis that typical instances shared more properties with other members of the same class.

To analyze the Rosch and Mervis data it is important to distinguish the task performed by subjects in typicality experiments from the basic-level studies presented earlier. In the target recognition tasks it appears that category match applied to the target concept is a good predictor of response time ranking, but unlike the basic-level studies, Rosch and Mervis did not give subjects a target category for which membership had to be verified or rejected; they required that subjects predict the membership of an observation. Thus, assumptions about the portions of memory that may be examined during classification are less clear. For this reason our analysis will focus on two strategies of representation that might reasonably be used to encode the typicality stimuli by human

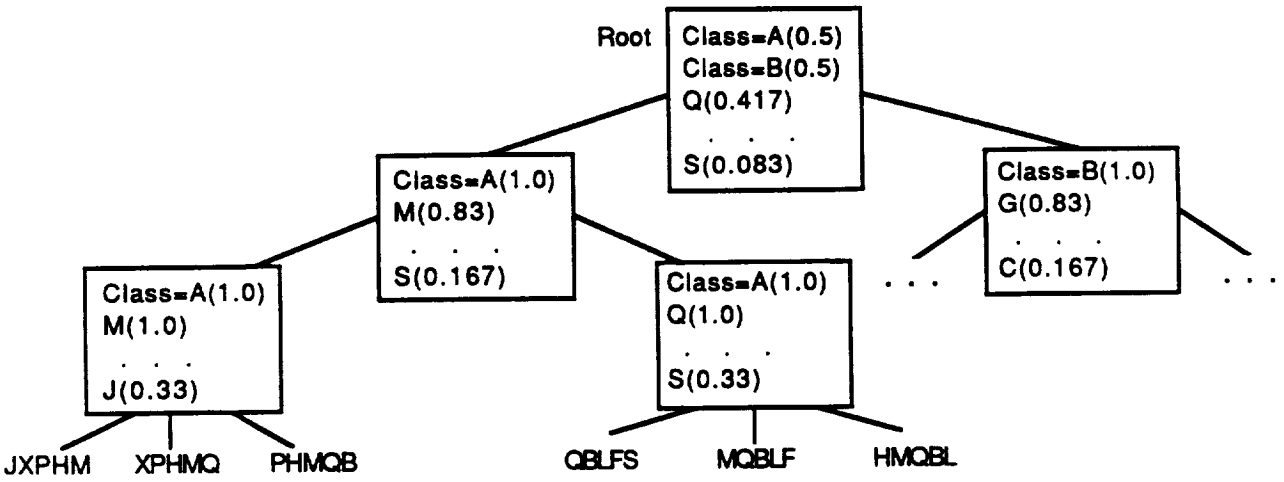

Figure 5. A concept hierarchy that summarizes the intra-category overlap data.

subjects; by considering two alternative encodings we hope to better illustrate the robustness of category match as a predictor of response time.

The first strategy, which we term *local*, assumes that each category is associated with a distinct independent cue concept; this is similar to our assumptions in the basic-level studies. An observation is classified with the category that maximizes category match in time that is inversely proportional to the match score. A second strategy is inspired by a general processing assumption discussed in relation to the Murphy and Smith data: recognition effects are mediated by concepts in memory other than the target or, in this case, the possible targets. A *distributed* strategy assumes that category members may be distributed throughout memory. This was an important assumption behind our discussion of linearly separable categories in Section 3 and prediction accuracy in Section 4. In a distributed representation, externally-defined categories need not correspond to nodes in memory, but external-category 'features' or labels can be used to predict the membership of an observation. In modeling this strategy, we use COBWEB to organize the strings of the Rosch and Mervis studies, thereby simulating a subject's training phase. A test item's external-category membership is predicted at the concept tree node that maximizes category match and at which a prediction of external-category membership can be made with certainty (i.e., an external category label has a probability of 1.0 at the node). Time is assumed to be inversely proportional to the category match score of this node.[11]

Table 7 (a) shows response times for category A test items and category match scores for local and distributed representations. Because COBWEB is sensitive to input order, the distributed representation scores are averaged over 20 trees constructed from random orderings of the data strings. A representative tree for these data is shown in Figure 5. Since there is no overlap between

11. We could have also given a distributed account of basic-level data, in which superordinate, basic, and subordinate category labels may be distributed throughout a probabilistic concept hierarchy. In general, basic-level findings from a distributed account are consistent with human data. However, an unexplained implication of this account is that a superordinate label can be predicted with certainty at any node that a basic label can be predicted, apparently leading to equal response times. Section 6 discusses a resolution of this issue.

Table 7. Average response times and category match rankings for Rosch and Mervis (1975) data.

		RESPONSE TIME	CATEGORY MATCH (LOCAL)	CATEGORY MATCH (COBWEB)
INTRA-	HIGH	560MS	0.948	0.910
CATEGORY	MEDIUM	617MS	0.823	0.832
OVERLAP (A)	LOW	692MS	0.594	0.736
INTER-	LOW	909MS	0.306	0.488
CATEGORY	MEDIUM	986MS	0.196	0.461
OVERLAP (B)	HIGH	1125MS	0.120	0.396

classes A and B, COBWEB's approach of grouping similar objects almost always results in the same categories (at the top level) as those based solely on the external label;[12] recall that these topmost nodes are also the two concepts considered in the local representation. In addition, the top-most node generally maximizes the category match scores of the high and medium intra-overlap data, thus explaining the similarity of category match scores for these data using the local and distributed representations. In contrast, low intra-overlap items exhibit markedly higher category match scores using a distributed representation. This reflects the fact that low intra-overlap observations in this experiment more often match a subordinate node in the tree better than they match the top-level node. Our assumption is that classification will be more rapid with respect to the subordinate.

Regardless of whether one assumes a local or distributed representation, category match scores are inversely related to response time. Intuitively, features that are relatively unique among category A members will cause a decrease in category match for their respective observations because they have smaller $P(V|C)$ values for these features. Conversely, the $P(V|C)$ values for unique features will be higher at subordinate nodes, thus increasing category match at lower nodes, even to the point of offsetting the reduced $P(C)$ values.

5.3.2 TYPICALITY AND INTER-CATEGORY SIMILARITY

In addition to varying intra-category overlap, Rosch and Mervis explored the impact of inter-category (between-category) similarity. Table 6 (b) shows the stimuli for this study, in which subjects were taught to distinguish categories A and B. Intra-category overlap was held constant for category A members, but the average extent to which category A members overlapped with B varied from 0.0 ('HPNWD') to 1.3 ('4KCTG'). As Table 7 (b) indicates, category A instances that shared few symbols with strings in category B were recognized more quickly (i.e., were treated as more typical).

12. At times atypical members of A (or B) may be placed in distinct categories.

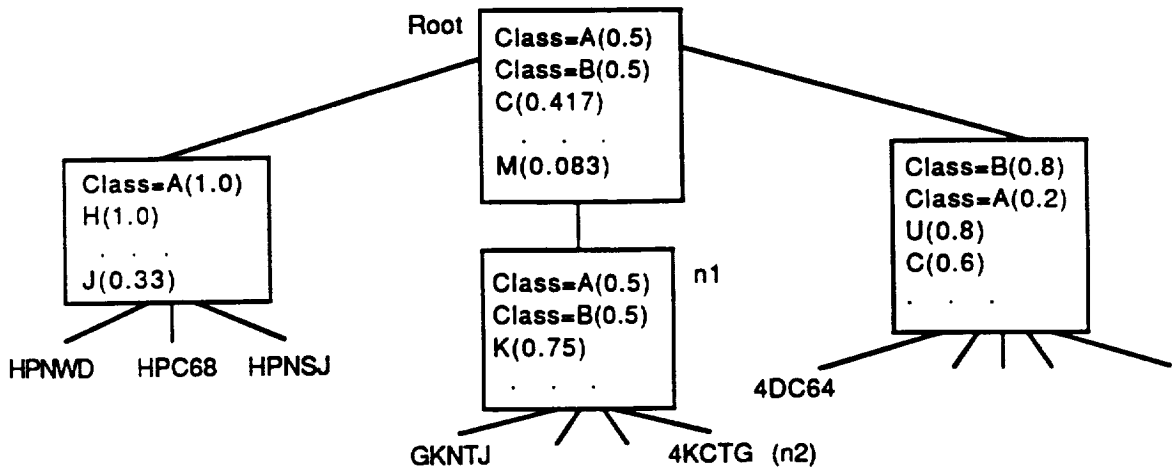

Figure 6. Tree constructed over inter-category experimental data.

Once again, we used two representation strategies in testing the relationship between category match predictions and the response time data. For these stimuli, significant featural overlap between certain members of A and B caused COBWEB to consistently distribute class A members throughout the resultant concept hierarchy. Figure 6 shows one such tree. In this case, the most strongly indicated node may not allow a prediction of category membership to be made with certainty and alternative nodes must be examined. For example, node n_1 of the tree of Figure 6 does not allow an unambiguous prediction of A or B membership. In this case, the category match score of node n_2 would have to be used when classifying '4KCTG'.

Table 7 (b) reveals that, in the case of both local and distributed representations, category match scores were again inversely related to Rosch and Mervis' response time data. Scores for the distributed case are averaged over 20 trials. Intuitively, inter-category (A and B) similarities tend to diffuse evidence across lateral subtrees. In cases such as 'J', the feature is actually more predictive of category B than A, thus adding nothing to the category match score of the atypical observations to which these features belong, or actually detracting from it.

5.3.3 DISCUSSION OF TYPICALITY RESULTS

Taken alone and collectively the data from our intra-category and inter-category studies demonstrate that category match accurately ranks test item response time. To better illustrate this, Table 8 shows the predicted response times from the local and distributed category match scores of Table 7 that were obtained from a linear regression. Category match accounts for 95.7% of the variance in response time in the local case ($F(1,4) = 88.7,\ p < 0.001$) and 96.6% in the distributed case ($F(1,4) = 114.1,\ p < 0.001$).[13] While both strategies account for most of the variance, the distributed representation compresses the category match scores across the typicality range,

13. We considered the intra-category and inter-category data as one sample, given that our calculation of category match scores in each case was identical. Considering the data as two separate samples yields similar accounts of variance and predicted response times.

Table 8. Human and predicted response times for Rosch and Mervis (1975) data.

		RESPONSE TIME	PREDICTED TIME (LOCAL)	PREDICTED TIME (COBWEB)
INTRA-	HIGH	560MS	526	535
CATEGORY	MEDIUM	617MS	606	615
OVERLAP (A)	LOW	692MS	753	713
INTER-	LOW	909MS	938	968
CATEGORY	MEDIUM	986MS	1008	995
OVERLAP (B)	HIGH	1125MS	1057	1062

possibly because it better tailors recognition to an observation. More specifically, distribution indicates that typicality rankings emerge from variation along two dimensions of categorization (Rosch, 1978). Relatively unique features to a category will tend to diffuse activation towards subordinate categories (i.e., along a *vertical* dimension). Features that overlap with contrasting categories diffuse activation across nodes that classify observations of more than one contrast category (i.e., a *horizontal* dimension). Local accounts of typicality only consider variance along this latter dimension.

A distributed model makes the vertical dimension explicit in explanations of typicality. This suggests interactions with basic-level effects, which also emerge from variation along this dimension. For example, Rosch et al. (1976) predicted and Jolicoeur, Gluck, and Kosslyn (1984) verified that the human preference for the basic level is qualified. In particular, an observation (e.g., a specific chicken) may be sufficiently atypical of its basic-level category (e.g., *bird*) that it will be first recognized as an instance of a subordinate category (e.g., *chicken*). Low intra-category overlap results in greater activation of subordinate nodes, while there is a simultaneous decrease in activation of the basic-level node for atypical objects due to less intra-category overlap and more inter-category overlap. In cases of sufficient atypicality, these tendencies may interact so that classification is initiated at a subordinate level. This is nicely illustrated by the data of Table 6 (a). In this simulation the category match scores of atypical objects (nonsense strings) were higher at subordinates than at top level nodes of a COBWEB-generated tree – presumably the basic level, since this level maximizes category utility.

5.4 Fan Effects

To a large extent, knowledge of basic level and typicality effects influenced our adoption of probabilistic representations and the category match metric. However, the framework also accounts for certain fan effects (Anderson, 1976), which did not influence our representation and processing biases. Nonetheless, these phenomena are accurately predicted by application of category match to probabilistic representations.

Table 9. Human and predicted response times for Anderson's (1974) fan effect data.

(A) TRUES

	1	2	3
1	1111MS (1120MS)	1174MS (1157MS)	1222MS (1184MS)
2	1167MS (1157MS)	1198MS (1195MS)	1222MS (1259MS)
3	1153MS (1184MS)	1233MS (1259MS)	1357MS (1321MS)

(B) FALSES

	1	2	3
1	1197MS (1168MS)	1221MS (1240MS)	1264MS (1306MS)
2	1250MS (1240MS)	1356MS (1312MS)	1291MS (1379MS)
3	1262MS (1306MS)	1471MS (1379MS)	1465MS (1444MS)

Fan effects indicate that observations with frequently encountered features may be more difficult to recognize than observations with relatively unique features, given that exposure across observations is relatively constant. Anderson (1974) demonstrated this principle in sentence recognition tasks, which typically used simple sentences that consisted of a person and a location:

(1-1) The doctor is in the bank. (1-2) The fireman is in the park.
(2-1) The teacher is in the church. (2-2) The teacher is in the park.

Sentences vary in the number of features (persons, locations) that they share with other sentences. The numbers preceding each sentence indicate the number of sentences that contain the respective persons and locations. For example, sentence (2-1) indicates that 'teacher' appears twice and 'church' appears once in the set of four sentences.

After subjects were trained on selected sentences, they were presented with probes and asked whether they had previously observed a sentence (true) or not (false). Anderson found that recognition time increased in the true and false case with the frequency that a person and location was present in training sentences. Table 9 shows matrices of nine cells each, which show the averaged human response data (Anderson, 1974) in the upper portion of each cell. Each cell corresponds to items with the number of persons and locations denoted on the horizontal and vertical dimensions, respectively. In general, response time for both true and false human data increases as one moves to the right and/or down.

Elsewhere (Silber & Fisher, 1989), we explained these effects as a special case of typicality phenomena, in which a subject was to classify test observations with respect to the singleton categories formed from the training observations. In this account, the intra-category overlap between singleton categories is identical, since each observation contains exactly two features. Response time differences are thus explained entirely in terms of inter-category overlap. The more features that an observation shares with other observations, the greater the overlap between its corresponding singleton category and contrasting singleton categories. Observations with greater overlap should require greater response time, which is consistent with typicality findings.

In accounting for these data we will primarily be concerned with local representations, particularly in the case of true test items. The reason for this is that COBWEB may impose an organization above the singleton level, but both features of a sentence can only be predicted with certainty at the leaves. In the case of 'true' test items we thus report category match scores for the test item's corresponding singleton, since this is the strongest match. The false case is more complicated. In contrast to the Murphy and Smith experiments, subjects are not given a verbally-cued target category on which to focus. Rather, we assume that all categories must be investigated and rejected. There are a number of ways that we might simulate this process, but for simplicity we report the average category match score of an observation across all categories in memory: this is the set of singletons in the local case and singletons plus internal nodes in the distributed case. Averaging captures the intuition that larger scores indicate that more of memory must be investigated, but it makes minimal assumptions about how this might be accomplished.

In addition to the human data at the top of each cell, Table 9 shows predicted response times in parentheses that were generated from a linear regression. Once again, false response times are proportional to category match, and there is an inverse relation between the two in the true case. Category match scores account for 83.8% of the variance in true response time ($F(1,7) = 36.3$, $p < 0.001$) and 70.9% of the variance in false response time ($F(1,7) = 17.1$, $p < 0.004$).[14]

In contrast to our account of fan effects, Anderson's (1976) initial explanation suggested that items were stored in a semantic network and activation spread from the features of a test sentence until the original instance was found in memory (trues) or all links from the features had been exhausted (falses). A mathematical abstraction of Anderson's ACT processing model accounted for 83% of the variance in the true and false response times. It appears that Anderson considered the true and false cases as one sample, whereas we have modeled them separately. Overall the ACT-based model yields better predictions than our model, but it also assumes more parameters that are linked to the processing assumptions of ACT.

Recently, Anderson (in press) has provided an explanation of fan effect based on a rational model of information retrieval systems. A key ingredient in his explanation is the cue validity of features towards a sentence. Similarly, we can see the role of cue validity in category match by reexpressing it as $\sum_j P(V_j)P(C|V_j)P(V_j|C) - P(V_j)P(C)P(V_j)$. Since $P(V_j|C)$'s and $P(C)$'s are constant across all singleton categories, a dominant factor in category match are the cue validities, $P(C|V_j)$. More generally, low cue validity reflects greater overlap with contrast (singleton) categories. In fact, as first observed by Jane Silber (personal communication), it is the reliance on cue validity that unifies fan with typicality phenomena, notably the aspect that emerges from inter-category overlap.

6. General Discussion

In this paper we presented speculative and rational analyses of basic level, typicality, and fan effects. Our primary goal was to verify the ability of Gluck and Corter's category utility and our category match variant to predict human response time; our precise application of category match necessarily

14. We considered the true and false data as separate samples, given differences in the calculation of category match scores between the two cases.

varied with differences across the experimental studies, but collectively our various assumptions are consistent. In the case of all 'true' stimuli, response time is predicted by the score of the node that maximizes category match *and* that satisfies the conditions of the stimulus (e.g., membership in a target category; possession of both person and location features of the stimulus). In the 'false' cases, response time is predicted by a function of the match scores of nodes that must be examined in order to issue a false response with certainty (e.g., a target category if one is supplied or all categories in memory otherwise).

A benefit of exploiting Gluck and Corter's rational analysis is that it provides a specification of concept quality that has both computational and psychological merit. In particular, we coupled category utility and methods from machine learning (Kolodner, 1983; Lebowitz, 1982) in the probabilistic representations and classification strategies of COBWEB. Our ongoing research is advancing in two directions: to improve the system so that it is more fully consistent with the psychological phenomena, and to expand the scope of the model to other areas of cognition, notably problem solving.

6.1 The Role of Indexing

Despite the descriptive merits of our specification, we have noted that it leaves certain 'implementation' issues unexplained. For example, our analysis of the Murphy and Smith data required that false response times rise with category match scores. However, taken alone, the assumption – that increased match implies faster access – suggests that false categories with higher match scores would be more quickly accessed, thus allowing for faster rejection. To resolve this problem, we appealed to assumptions about processing and representation which required that poorly matching categories would be more quickly rejected. We can now flesh out some general mechanisms that will realize these behavioral constraints.

Many theories of learning and memory have addressed the problem of identifying and exploiting informative features for classification. Category utility and category match suggest that a feature positively informs the categorization process if and only if $P(f|C) > P(f)$. This criterion of feature informativeness has been suggested by data in such diverse areas as stereotype theory (McCauley, Stitt, & Segal, 1980) and animal learning (Rescorla, 1968), as well as other areas of AI and psychology (Schlimmer, 1986).

However, strict adherence to this criterion may still allow features of little benefit to be evaluated. For example, $P(f|C) = 1.0$ for all singleton categories, thus satisfying the criterion for most leaves. In response, Fisher (1988, 1989) and Quinlan (1986) have employed a variety of methods for determining when the $P(f|C) > P(f)$ relation is significant. Given category utility's relation to Jones' (1983) collocation measure – $P(C|V)P(V|C)$ – an attractive method would be to find nodes that maximize this product for each feature. Intuitively, these will be the most specific categories (i.e., yielding high $P(V|C)$) for which the feature is still discriminating (i.e., yielding high $P(C|V)$). For example, the feature 'fly' may be maximized at 'birds', but below this it does not discriminate among birds. Our heuristic assumption is that at collocation-maximizing nodes a feature's presence becomes approximately independent of membership in lower-level categories.

This heuristic creates a 'horizon' beyond which a feature does not discriminate. We may impose this horizon through *indexing* (Feigenbaum & Simon, 1984; Kolodner, 1983; Lebowitz, 1982): a feature labels a link from a node to one or more of its descendents; the link is traversed only when an observation with that feature is observed. For example, in the zoological taxonomy, 'fly' would index the categories 'vertebrate' and 'bird' from the taxonomy's root, since these nodes are within the horizon determined by collocation.[15] In many cases, a feature's collocation is maximized at the root and thus does not index any node. This indexing strategy may also be recursively applied so that descendents of 'bird' are indexed by features that discriminate them from the 'bird' node. Notice that 'fly' would not index any descendant in this context, whereas 'not-fly' presumably would. Conversely, 'fly' would not index 'mammal' from the root, but within the 'mammal' context it would discriminate 'bats'.

In the revised model, classification would be based on a category match score that is computed only over the features of an observation that are used for indexing. All such scores will be positive, with the maximum score dictating category membership. This procedure would recurse from the maximum node until it reached a 'dead end', at which no indices for the observation remain. Computationally, this indexing scheme radically delimits the portions of memory that are accessed for each observation, without appealing to *ad hoc* thresholds.

This strategy also appears to have desirable psychological properties, but at this point we can only speculate with respect to some of the data. First, the approach breaks the basic-level tie that we reported in relation to the Hoffman and Ziessler studies; in the case of one tree, subjects treated the top-most nodes as basic, but category match predicted a tie between these nodes and their children. In this case, indexing brings all features to bear on the top-level nodes, but it removes a feature from the match computation for the middle-level nodes. Basic-level identifications are also accurately predicted in the other studies, as are typicality rankings from the Rosch and Mervis studies.

The revised model also promises to explain results with false test items. To review our earlier account of Murphy and Smith's data, the 'false' target is verbally cued, which is the only cue in the case of superordinate-only relatedness and unrelatedness. Their matching score is more quickly overwhelmed by the match score of the correct category than is the false target in the basic-relatedness case. There are no concept-specific thresholds, but one can view recognition as mediated by implicit and variable thresholds that emerge from the competition among contrast categories. Membership in a category can be rejected when a dead end is reached and a competing category has a higher matching score. A similar account also applies to Anderson's data. Thus, our account of the false data suggests that low match scores in our specification translate to more dead ends in categorization.

15. This differs from our earlier systems (Fisher, 1988; Silber & Fisher, 1989), which directed indices *only* at collocation-maximizing nodes (e.g., bird, but not vertebrate). This strategy proved too fragile, particularly during the early stages of learning.

6.2 Models of the Planning Process

In addition to improving our account of categorization phenomena, we are extending the model of recognition and learning to domains such as planning. This work is closely related to a growing body of research in machine learning that is focused on problem solving (e.g., Minton, 1988; Shavlik, 1989). Much of this work has focused on analytic learning methods (Mitchell, Keller, & Kedar-Cabelli, 1986) that transform knowledge from one form into another.

In contrast, we propose that learning in problem-solving domains is best modeled as concept formation, in which memory organizes problem-solving experience in a manner that facilitates efficient reuse and that incrementally transforms a novice into an expert. Our approach augments a problem solver with a concept formation component that organizes problem descriptions and solution traces (Allen & Langley, 1989; Yang & Fisher, 1989). A new problem is classified via the concept hierarchy in hopes of finding a reusable solution trace. In cases where a complete solution cannot be recovered, one may still obtain a partial solution by recovering predictable subtraces at nodes encountered during classification. Our framework is consistent with psychological (Chi, Feltovich, & Glaser, 1981) and computational (Bareiss, 1989) views that expertise involves an ability to solve problems via classification versus search.

We expect that many of the behaviors that occur in conceptual memory – such as typicality and basic-level effects – will also occur with problem-solving memory. Thus, we hope to account for many of the same types of phenomena in episodic memory that we explained for object memory. This work is also addressing some limitations of featural and probabilistic models generally and COBWEB specifically. In particular, we are extending our strategies to *structured* representations (Smith & Medin, 1981; Dietterich & Michalski, 1983), which allow relationships between object features or components (e.g., `next-to(x, y)`). In addition, we are using an object's *function* in problem solving as a guide for concept learning (Wisniewski, 1989; Nelson, 1973). More generally, we hope to develop a general-purpose cognitive architecture that is founded on the principles of human memory that we have described here (Langley, Thompson, Iba, Gennari, & Allen, 1989). Finally, we hope to illustrate that our analysis in particular, and rational/speculative analyses in general, provide generic models of intelligent behavior that cognitive scientists can exploit, regardless of the side of the psychological/computational fence on which they typically reside.

Acknowledgements

We thank Dennis Kibler, Jim Corter, Mark Gluck, Richard Granger, and Jeff Schlimmer for their insights and influential discussions in the early stages of this work. In addition, we thank Jeanette Altarriba, Rogers Hall, Kevin Thompson, Jim Corter, and Mark Gluck for helpful suggestions on more recent versions of the material. Doug Fisher was supported by Grant No. NCC 2-645 from NASA Ames Research Center.

References

Aha, D., & Kibler, D. (1989). Noise-tolerant instance-based learning. *Proceedings of the Eleventh International Conference on Artificial Intelligence* (pp. 794–800). Detroit, MI: Morgan Kaufmann.

Aha, D., & McNulty, D. (1989). Learning relative attribute weights for instance-based concept descriptions. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 530–537). Ann Arbor, MI: Lawrence Erlbaum.

Allen, J., & Langley, P. (1989). Using concept hierarchies to organize plan knowledge. *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 229–231). Ithaca, NY: Morgan Kaufmann.

Anderson, J. R. (1974). Retrieval of propositional information from long term memory. *Cognitive Psychology*, *6*, 451–474.

Anderson, J. R. (1976). *Language, memory, and thought*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R. (in press). The place of cognitive architectures in a rational analysis. In K. Van Lehn (Ed.), *Architectures for intelligence*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R., & Kline, P. J. (1979). A learning system and its psychological implications. *Proceedings of the Sixth International Joint Conference on Artificial Intelligence* (pp. 16–21). Tokyo, Japan: Morgan Kaufmann.

Bareiss, R. (1989). *Exemplar-based knowledge acquisition*. San Diego, CA: Academic Press.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: John Wiley & Sons.

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 510–516). Montreal, Quebec: Lawrence Erlbaum.

Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., & Freeman, D. (1988). AUTOCLASS: A Bayesian classification system. *Proceedings of the Fifth International Machine Learning Conference* (pp. 54–64). Ann Arbor, MI: Morgan Kaufmann.

Chi, M., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121–152.

Collins, A., & Loftus, E. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

Corter, J., & Gluck, M. (1985). Machine generalization and human categorization: An information theoretic view. *Proceedings of the Workshop on Probability and Uncertainty in Artificial Intelligence* (pp. 201–207). Los Angeles, CA.

Dietterich, T. G., & Michalski, R. S. (1983). A comparative review of selected methods of learning from examples. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach*. San Mateo, CA: Morgan Kaufmann.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: John Wiley & Sons.

Everitt, B. (1981). *Cluster analysis*. London: Heinemann.

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought*. New York: McGraw-Hill.

Feigenbaum, E. A., & Simon, H. A. (1984). EPAM-like models of recognition and learning. *Cognitive Science*, *8*, 305–336.

Fisher, D. (1988). A computational account of basic level and typicality effects. *Proceedings of the Seventh National Conference on Artificial Intelligence* (pp. 233–238). St. Paul, MN: Morgan Kaufmann.

Fisher, D. H. (1987a). *Knowledge acquisition via incremental conceptual clustering*. Doctoral dissertation, Department of Information & Computer Science, University of California, Irvine.

Fisher, D. H. (1987b). Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, *2*, 139–172.

Fisher, D. H. (1989). Noise-tolerant conceptual clustering. *Proceedings of the Eleventh International Joint Conference Artificial Intelligence* (pp. 825–830). Detroit, MI: Morgan Kaufmann.

Gennari, J. (1989). *A survey of clustering methods* (Technical Report 89-38). Irvine, CA: University of California, Department of Information & Computer Science.

Gennari, J. H., Langley, P., & Fisher, D. (1989). Models of incremental concept formation. *Artificial Intelligence*, *40*, 11–62.

Gluck, M., Bower, G., & Hee, M. (1989). A configural-cue network model of animal and human associative learning. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 323–332). Ann Arbor, MI: Lawrence Erlbaum.

Gluck, M. A., & Corter, J. E. (1985). Information, uncertainty, and the utility of categories. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society* (pp. 283–287). Irvine, CA: Lawrence Erlbaum.

Hadzikadic, M., & Yun, D. (1989). Concept formation by incremental conceptual clustering. *Proceedings of the International Joint Conference Artificial Intelligence* (pp. 831–836). Detroit, MI: Morgan Kaufmann.

Hall, R., & Kibler, D. (1985). Differing methodological perspectives in artificial intelligence. *AI Magazine*, *6*, 166–179.

Hanson, S. J., & Bauer, M. (1989). Conceptual clustering, categorization, and polymorphy. *Machine Learning*, *3*, 343–372.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior*, *16*, 321–338.

Hayes-Roth, F., & McDermott, J. (1978). An interference matching technique for inducing abstractions. *Communications of the ACM*, *21*, 401–410.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence*, *40*, 185–234.

Hoffman, J., & Ziessler, C. (1984). Objectidentifikation in kunstlichen begriffshierarchien. *Zeitscrift fur Psychologie, 16*, 43–275.

Hunt, E., Marin, J., & Stone, P. (1966). *Experiments in induction.* New York: Academic Press.

Jolicoeur, P., Gluck, M. A., & Kosslyn, S. M. (1984). Pictures and names: Making the connection. *Cognitive Psychology, 16*, 243–275.

Jones, G. (1983). Identifying basic categories. *Psychological Bulletin, 94*, 423–428.

Kline, P. J. (1983). *Computing the similarity of structured objects by means of heuristic search for correspondences.* Doctoral dissertation, Department of Psychology, University of Michigan, Ann Arbor.

Kolodner, J. L. (1983). Reconstructive memory: A computer model. *Cognitive Science, 7*, 281–328.

Langley, P. (1987). A general theory of discrimination learning. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production system models of learning and development.* Cambridge, MA: MIT Press.

Langley, P., Gennari, J., & Iba, W. (1987). Hill-climbing theories of learning. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 312–323). Irvine, CA: Morgan Kaufmann.

Langley, P., Thompson, K., Iba, W., Gennari, J. H., & Allen, J. A. (1989). *An integrated cognitive architecture for autonomous agents* (Technical Report 89-28). Irvine: Department of Information & Computer Science, University of California.

Lebowitz, M. (1982). Correcting erroneous generalizations. *Cognition and Brain Theory, 5*, 367–381.

Lebowitz, M. (1987). Experiments with incremental concept formation: UNIMEM. *Machine Learning, 2*, 103–138.

Marr, D. *Vision: A computational investigation into the human representation and processing of visual information.* San Fransisco, CA: W. H. Freeman.

Martin, J. D. (1989). Reducing redundant learning. *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 396–399). Ithaca, NY: Morgan Kaufmann.

McCauley, C., Stitt, C., & Segal, M. (1980). Stereotyping: From prejudice to prediction. *Psychological Bulletin, 87*, 195–208.

Medin, D. (1983). Structural principles of categorization. In T. Tighe & B. Shepp (Eds.), *Perception, cognition, and development.* Hillsdale, NJ: Lawrence Erlbaum.

Medin, D., & Schaffer, M. (1978). A context theory of classification learning. *Psychological Review, 85*, 207–238.

Medin, D. L., Wattenmaker, W. D., & Michalski, R. S. (1986). *Constraints and preferences in inductive learning* (Technical Report). Urbana-Champaign: University of Illinois, Department of Computer Science.

Mervis, C., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32*, 89–115.

Michalski, R. S. (1983). A theory and methodology of inductive learning. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach*. San Mateo, CA: Morgan Kaufmann.

Michalski, R. S., & Stepp, R. E. (1983). Learning from observation: Conceptual clustering. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach*. San Mateo, CA: Morgan Kaufmann.

Minton, S. (1988). Quantitative results concerning the utility of explanation-based learning. *Proceedings of the Seventh National Conference on Artificial Intelligence* (pp. 564–569). St. Paul, MN: Morgan Kaufmann.

Mitchell, T. M. (1982). Generalization as search. *Artificial Intelligence*, *18*, 203–226.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986). Explanation-based learning: A unifying view. *Machine Learning*, *1*, 47–80.

Murphy, G., & Smith, E. (1982). Basic level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior*, *21*, 1–20.

Nelson, K. (1973). Some evidence for the cognitive primacy of categorization and its functional basis. *Merrill-Palmer Quarterly of Behavior and Development*, *19*, 21–39.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nilsson, N. (1965). *Learning machines*. New York: McGraw-Hill.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87–108.

Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples. In D. Michie (Ed.), *Expert systems in the micro electronic age*. Edinburgh: Edinburgh University Press.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106.

Pearl, J. (1985). Learning hidden causes from empirical data. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (pp. 567–572). Los Angeles, CA: Morgan Kaufmann.

Reed, S. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, *66*, 1–5.

Rips, L., Shoben, E., & Smith, E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, *12*, 1–20.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd, (Eds.) *Cognition and categorization*. Hillsdale, NJ: Lawrence Erlbaum.

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573–605.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *18*, 382–439.

Schlimmer, J. C. (1986). *A note on correlational measures* (Technical Report 86-13). Irvine: University of California, Department of Information & Computer Science.

Schlimmer, J. C., & Fisher, D. (1986). A case study of incremental concept induction. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 496–501). Philadelphia, PA: Morgan Kaufmann.

Schlimmer, J. C., & Granger, R. H., Jr. (1986). Incremental learning from noisy data. *Machine Learning*, *1*, 317–334.

Shavlik, J. (1989). Acquiring recursive concepts with explanation-based learning. *Proceedings of the Eleventh International Conference on Artificial Intelligence* (pp. 688–693). Detroit, MI: Morgan Kaufmann.

Silber, J., & Fisher, D. (1989). A model of natural category structure and its behavioral implications. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 884–891). Ann Arbor, MI: Lawrence Erlbaum.

Simon, H. A. (1969). *The sciences of the artificial*. Cambridge, MA: MIT Press.

Simon, H. A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In L. W. Gregg (Ed.), *Knowledge and cognition*. Hillsdale, NJ: Lawrence Erlbaum.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Smith, E., Shoben, E., & Rips, L. (1974). Structure and processes in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*, 214–241.

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

Utgoff, P. E. (1988). ID5: An incremental ID3. *Proceedings of the Fifth International Conference on Machine Learning* (pp. 107–120). Ann Arbor, MI: Morgan Kaufmann.

Vere, S. (1980). Multilevel counterfactuals for generalization of relational concepts and productions. *Artificial Intelligence*, *14*, 139–164.

Winston, P. H. (1975). Learning structural descriptions from examples. In P. H. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill.

Wisniewski, E. (1989). Learning from examples: The effect of different conceptual roles. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 980–986). Ann Arbor, MI: Lawrence Erlbaum.

Yang, H., & Fisher, D. (1989). Conceptual clustering of means-ends plans. *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 232–234). Ithaca, NY: Morgan Kaufmann.

RIA-90-02-15-1

The Structure and Formation of Natural Categories
Douglas Fisher and Patrick Langley

February 1990

Categorization and concept formation are critical activities of intelligence. These processes and the conceptual structures that support them raise important issues at the interface of cognitive psychology and artificial intelligence. Our work presumes that advances in these and other areas are best facilitated by research methodologies that reward interdisciplinary interaction. In particular, we describe a computational model of concept formation and categorization that exploits a it rational analysis of basic level effects by Gluck and Corter (1985). Their work provides a clean prescription of human category preferences that we adapt to the task of concept learning. In addition, we extend their analysis to account for typicality and fan (Anderson, effects, and speculate on how our concept formation strategies might be extended to other facets of intelligence, such as problem solving.

RIA-90-03-20-1

Proposal for Constructing an Advanced Software Tool for Planetary Atmospheric Modeling
Richard Keller, Michael Sims, David Thompson, Ester Podolak, and Christopher P. McKay

March 1990

Scientific model building can be a time-intensive and painstaking process, often involving the development of large and complex computer programs. Despite the effort involved, scientific models cannot easily be distributed and shared with other scientists. In general, implemented scientific models are complex, idiosyncratic, and difficult for anyone but the original scientist/programmer to understand. We believe that advanced software techniques can facilitate both the model-building and model-sharing process. We propose to construct a scientific modeling software tool that serves as an aid to the scientist in developing and using models. The proposed tool will include an interactive intelligent graphical interface and a high-level domain-specific modeling language. As a testbed for this research, we propose development of a software prototype in the domain of planetary atmospheric modeling.

RIA-90-04-06-1

Model Compilation: An Approach to Automated Model Derivation
Richard Keller, Catherin Baudin, Yumi Iwasaki, Pandurang Nayak, and Kazuo Tanaka

April 1990

In this paper, we introduce an approach to automated model derivation for knowledge based systems. The approach – called *model compilation* – involves procedurally generating the set of domain models used by a knowledge-based system. With an implemented example, we illustrate how this approach can be used to derive models of different precision and abstraction, and models tailored to different tasks, from a given set of base domain models. In particular, we describe two implemented model compilers, each of which takes as input a base model that describes the structure and behavior of a simple electromechanical device – the Reaction Wheel Assembly of NASA's Hubble Space Telescope. The compilers transform this relatively general base model into simple task- specific models for troubleshooting and redesign, respectively, by applying a sequence of model transformations. Each transformation in this sequence produces an increasingly more specialized device model. The compilation approach lessens the burden of updating and maintaining consistency among models by enabling their automatic regeneration.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	Dates attached	

4. TITLE AND SUBTITLE

Titles/Authors - Attached

5. FUNDING NUMBERS

6. AUTHOR(S)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Code FIA - Artificial Intelligence Research Branch
Information Sciences Division

8. PERFORMING ORGANIZATION REPORT NUMBER

Attached

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Nasa/Ames Research Center

Moffett Field, CA. 94035-1000

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Available for Public Distribution

Peter Friedland 5/14/92 BRANCH CHIEF

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Abstracts ATTACHED

14. SUBJECT TERMS			15. NUMBER OF PAGES
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT