Appears in *Categorization by Humans and Machines: The Psychology of Learning and Motivation,*
Vol. 29, G. Nakamura, R. Taraban & D.L. Medin (Eds.), pp.189-218, Academic Press, Orlando, FL, 1993

# Integrating Theory and Data in Category Learning [*]

Raymond J. Mooney
Department of Computer Sciences
University of Texas
Austin, TX 78712
mooney@cs.utexas.edu; (512)471-9558

March 4, 1992

**Abstract**

Recent results in both machine learning and cognitive psychology demonstrate that effective category learning involves an integration of theory and data. First, theories can bias induction, affecting what category definitions are extracted from a set of examples. Second, conflicting data can cause theories to be revised. Third, theories can alter the representation of data through feature formation. This chapter reviews two machine learning systems that attempt to integrate theory and data in one or more of these ways. IOU uses a domain theory to acquire part of a concept definition and to focus induction on the unexplained aspects of the data. EITHER uses data to revise an imperfect theory and uses theory to add abstract features to the data. Recent psychological experiments reveal that these machine learning systems exhibit several important aspects of human category learning. Specifically, IOU has been used to successfully model some recent experimental results on the effect of functional knowledge on category learning.

[*] This research was supported by the NASA Ames Research Center under grant NCC 2-629 and by the National Science Foundation under grant IRI-9102926. Equipment used was donated by the Texas Instruments Corporation.

# 1 Introduction

Until recently, research in categorization in cognitive psychology and machine learning focused almost exclusively on empirical, data-driven models. The most frequently studied task was learning categories from simple examples represented as lists of features (Bruner, Goodnow, & Austin, 1956; Smith & Medin, 1981; Quinlan, 1979; Michalski & Chilausky, 1980). The role of background knowledge or domain theories on category learning was largely ignored. The focus was on understanding the basic process of induction from raw data.

In the past decade, researchers in both areas began to investigate knowledge-intensive concept learning. Several researchers in machine learning began developing systems that performed a detailed analysis of an individual example. Many of these systems could learn a new concept from a single example. These methods eventually became known as *explanation-based learning* (Mitchell, Keller, & Kedar-Cabelli, 1986; DeJong & Mooney, 1986). Meanwhile, cognitive psychologists were also turning their attention to the role of theories in categorization (Murphy & Medin, 1985; Nakamura, 1985; Barsalou, 1983). One important finding was that subjects, like explanation-based learning systems, could acquire a concept or schema from only a single example (Ahn, Mooney, Brewer, & DeJong, 1987).

However, purely empirical (data-driven) and purely analytical (theory-driven) views of categorization are clearly end-points on a continuum. Most real category learning tasks involve an integration of background theory and empirical data. Recent research in both cognitive psychology and machine learning has begun to focus on the issue of integrating theory and data in concept learning (Wisniewski & Medin, 1991; Ahn, Brewer, & Mooney, in press; Segre, 1989; Birnbaum & Collins, 1991). This research is attempting to unravel the complex interacting effects of theory and data on concept learning.

This chapter describes two recently developed machine learning systems that integrate theory and data, and discusses the extent to which these systems can model recent experimental results on human concept learning. The remainder of the chapter is organized as follows. Section 2 briefly reviews standard empirical and explanation-based models of learning. Section 3 discusses various ways in which theory and data can interact in categorization. Section 4 describes IOU, a computer program that uses a domain theory to bias category learning. Simulation results are also presented in Section 4, in which IOU is shown to model effects found when subjects in a standard learning-from-examples experiment are told the function of the underlying categories. Section 5 describes another machine learning system, Either, which modifies an existing domain theory based on empirical data. Section 5 also discusses how the methods used in Either could potentially model recent results on how human subjects use data to modify theories and theories to modify data. Section 6 summarizes our results and presents some problems and ultimate goals for future research.

## 2 Review of Empirical and Explanation-Based Learning

The majority of research in categorization within psychology and machine learning has concerned *empirical* or *similarity-based* methods. In these methods, categories are learned by examining similarities and differences between relatively large numbers of examples. In cognitive psychology, similarity-based methods are normally divided into *classical*, *probablistic*, and *exemplar* models (Smith & Medin, 1981). Classical or *rule-based* methods form abstract logical definitions of categories (Bruner et al.,1956; Medin, Wattenmaker, & Michalski, 1987; Quinlan, 1986).[1] Probabilistic methods extract feature weights or conditional probabilities that are subsequently used to compute category membership (Posner & Keele, 1968; Rosenblatt, 1962; Fisher, 1987). Exemplar models do not form abstractions but rather categorize examples based on similarity to specific stored instances (Medin & Shaeffer, 1978; Aha, 1991). Very little emphasis is given to the role of background knowledge in any of these models. Psychological experiments in this area generally use simple, artificial data so that subjects cannot employ their existing knowledge.

Explanation-based learning, on the other hand, acquires a concept definition from a single example by using existing background knowledge to explain the example and thereby focus on its important features (DeJong, 1988). In the standard formalization (Mitchell et al., 1986; DeJong & Mooney, 1986), the *domain theory* is represented by a set of rules that

[1] The term "classical" normally refers to purely conjunctive descriptions, i.e. necessary and sufficient conditions.

allows the system to logically deduce that a concrete example satisfies an abstract definition of a given *goal concept*. Explanation-based generalization then consists of the following two steps:

1. Explain: Construct an explanation using the domain theory that proves that the training example satisfies the definition of the goal concept.

2. Generalize: Determine a set of sufficient conditions under which the explanation holds, stated in purely observable terms.

Machine learning systems have learned a number of concepts using this approach. For example, consider learning a structural definition of a cup from a functional definition and a single example (Winston, Binford, Katz, & Lowry, 1983). Assume the example is described by the following observable features:

> owner=Fred, color=red, location=table, width=medium, has-bottom, flat-bottom, has-handle, lightweight, has-concavity, upward-pointing-concavity.

Assume one has the following functional definition of a cup:

> stable & liftable & open-vessel → cup,

which states that anything that is stable, liftable, and an open-vessel is a cup. If the domain theory contains the following rules:

Figure 1: Explanation of a Cup.

has-bottom & flat-bottom → stable

lightweight & graspable → liftable

has-handle → graspable

width=small & insulating → graspable

has-concavity & upward-pointing-concavity → open-vessel,

one can deduce that the example is a cup using the "explanation" or proof shown in Figure 1. A definition of cup in purely observable terms can be obtained by compiling the explanation into a new rule. The root of the explanation forms the consequent of the new rule and the leaves form the antecedents. Below is the compiled rule for the cup example:

has-bottom & flat-bottom & has-handle & lightweight & has-concavity &

upward-pointing-concavity → cup

Notice that the generalization omits information about the color, owner, location, and width

Figure 2: Sketch of a Boot-Jack.

of the example, since these features are not used in the explanation. Once the rule has been compiled, instead of performing complicated inferencing, direct pattern matching can be used to classify examples as cups.

For a more psychologically motivated example of explanation-based learning, consider the following anecdote. Not long after moving to Texas, I encountered an example of an interesting device called a boot-jack. A rough sketch of a typical boot-jack is shown in Figure 2. This device allows an urban cowboy to remove his boots easily and independently after a long, hard day at the office. A boot-jack is used by stepping on the rear of the device with one foot, snuggly inserting the heel of the other foot into the notch, and pulling one's leg upward to remove the boot. The first example of a boot-jack I encountered was made of brass and shaped like the head of a long-horn bull whose U-shaped horns formed the notch. After seeing this particular device used to remove a pair of boots, I immediately formed an accurate concept of a boot-jack. I knew that certain properties such as shape, size, and rigidity were important but that the long-horn image was superfluous and ornamental. I also immediately identified my acquisition of this concept as a type of explanation-based learning. Although the next example I encountered was very different, a simple piece of wood with the shape shown in Figure 2, I was able to quickly and accurately classify it as an example of the same concept.

## 3 Combining Theory and Data

Empirical learning models have been criticized for ignoring the importance of existing knowledge, failing to account for the "coherence" of concepts, being susceptible to spurious correlations, and requiring too many training examples and intractable computational resources (Murphy & Medin, 1985; Schank, Collins, & Hunter, 1986; Mitchell, et al., 1986). Explanation-based models have been criticized for requiring a complete, correct, consistent, and tractable domain theory, only learning deductive consequences of existing knowledge, and only increasing speed of processing rather than capability (Dietterich, 1986; Mitchell at al., 1986).

Addressing these problems usually entails integrating empirical and analytical methods. Consequently, there has been a growing interest in this subject in both machine learning and cognitive psychology. Dozens of machine learning systems that integrate theory and data have been developed over the last several years (Segre, 1989; Birnbaum & Collins, 1991; Michalski & Teccuci, 1991). There has also been a growing number of psychological experiments on the topic (Ahn & Brewer, 1988; Ahn, Brewer, & Mooney, in press; Pazzani, 1991; Wisniewski, 1989; Wisniewski & Medin, 1991).

Existing methods for integrating theory and data in learning can be divided into three broad, overlapping categories.

- Theory as bias: The fundamental "problem of induction" is that there are multiple hypotheses consistent with any set of empirical data. Domain theories can be used to prefer certain hypotheses.

- Theory revision: Empirical data can be used to revise domain theories that draw incorrect conclusions.

- Theory for data interpretation: Theories can change or enhance the representation of data. For example, theories can derive abstract features from raw perceptual data.

A growing number of machine learning systems are attempting to integrate theory and data in one or more of these ways. Researchers have explored various ways of using domain theories to bias empirical learning (Flann & Dietterich, 1989; Hirsh, 1990; Cohen, 1990; Pazzani, 1991). Many of these systems employ an overly-general theory that admits too many examples as members of a category, but is specific enough to constrain the hypotheses considered by induction. A number of other recent systems attempt to revise a domain theory to fit empirical data (Ginsberg, 1990; Rajamoney, 1990; Danyluk, 1991; Towell &

Shavlik, 1991). However, all of these systems have important restrictions on the type of theories they can revise and the kinds of errors they can correct. Finally, several systems have been developed that use theories to interpret data (Michalski, 1983; Drastal, Czako, & Raatz, 1989). These system generally use rules that derive additional features from the initial description of an example.

This paper describes two recent systems and discusses their psychological relevance. Section 4 discusses IOU, a system that uses a domain theory to bias induction. Section 5 discusses EITHER, a system that performs theory revision and uses theory for data interpretation.

## 4 Induction Over the Unexplained

This section describes a machine learning system that uses empirical and explanation-based methods to learn different parts of a concept definition. Many concepts have both explanatory and nonexplanatory aspects. For example, scripts for events such as a birthday party or a wedding have goal-directed as well as ritualistic actions. Concepts for artifacts such as a cup or a building have functionally important features as well as aesthetic or conventional ones. Animals have some attributes with clear survival value as well as more obscure features. Diseases have some symptoms that can be causally explained by current biological theory as well as others that are simply known to be correlated with the condition.

In IOU (Induction Over the Unexplained), explanation-based methods are used to learn the part of the concept definition that can be explained by an underlying domain theory. Empirical methods are used to learn the part of the concept definition consisting of unexplained regularities in the examples. First, IOU uses an existing domain theory to explain each example. It then extracts their explainable features as part of the concept definition. Explainable features are then removed from the examples and the reduced example descriptions are passed to a standard empirical learning system. This system finds additional commonalities which are added to the final concept definition. A test example must meet the requirements of the domain theory as well as the empirically learned definition in order to be considered a member of the concept.

IOU uses its domain theory as a bias to prefer certain consistent concept definitions over others. In particular, the system prefers to include features that can be explained as relevant to the function or purpose of the concept. IOU can also be viewed as using theory to interpret and modify data, since it removes explained features and performs induction using only the unexplainable features of the examples.

## 4.1 IOU Problem and Algorithm

The general problem IOU addresses is *theory-based concept specialization* as defined by Flann & Dietterich (1989). The system is assumed to have a domain theory for a generalization of the concept to be learned. A definition of the specific problem addressed by IOU is given in Table 1. The current implementation of IOU employs a feature-based description language

Table 1: The IOU Problem.

**Given:**

- A set of positive and negative examples of an intended concept, $C_i$
- A propositional Horn-clause domain theory for an explainable concept, $C_e$ that is a generalization of the intended concept, i.e. $C_i \rightarrow C_e$.

**Determine:**

A definition for the intended concept in terms of observable features that is consistent with the examples and a specialization of the explainable concept.

that includes binary, discrete, and real-valued attributes. A domain theory is restricted to a set of propositional Horn clauses [2] that can include feature value pairs (color=red), numerical thresholds (length $<$ 3), and binary propositions (has-handle).

As an example of a problem suitable for IOU, consider a slight variation of the cup example introduced earlier. The standard functional definition is more accurately considered a definition of a drinking vessel rather than a cup since it cannot actually distinguish between cups, glasses, mugs, shot glasses, etc.. Therefore, assume the overly-general explainable concept is drinking-vessel, defined as:

stable & liftable & open-vessel → drinking-vessel.

Assume that the examples available are those shown in Table 2. The problem is to use the domain theory to learn the explainable features of a cup (flat-bottom, has-concavity etc.) and to use empirical techniques to learn the nonexplanatory features (volume=small) that rule

[2] A propositional Horn clause a rule of the form $A_1 \& A_2 \& ... A_n \rightarrow C$ where all $A_i$ and $C$ are simple propositions such as graspable or color=black.

Table 2: Examples for Learning Cup.

| | cup-1 (+) | cup-2 (+) | shot-glass-1 (-) | mug-1 (-) | can-1 (-) |
|---|---|---|---|---|---|
| has-bottom | true | true | true | true | true |
| flat-bottom | true | true | true | true | true |
| has-concavity | true | true | true | true | true |
| upward-pointing | true | true | true | true | true |
| lightweight | true | true | true | true | true |
| has-handle | true | false | false | true | false |
| width | small | small | small | medium | small |
| insulating | false | true | true | false | false |
| color | white | red | white | copper | silver |
| volume | small | small | tiny | large | small |
| shape | cylinder | cylinder | cylinder | cylinder | cylinder |

out shot glasses and mugs.

Table 3 shows the basic IOU algorithm. Step one uses standard explanation-based techniques to generalize each of the positive examples. The explanation of cup-1 is the same as that shown in Figure 1, except the goal concept is drinking-vessel instead of cup. The explanation of cup-2 differs only on how it is shown to be graspable. The resulting explanation-based generalizations are:

cup-1: has-bottom & flat-bottom & has-handle & lightweight & has-concavity & upward-pointing-concavity

cup-2: has-bottom & flat-bottom & width=small & insulating & lightweight & has-concavity & upward-pointing-concavity

Step two simply combines the explanation-based generalizations disjunctively and factors

Table 3: The Basic IOU Algorithm.

1. Using the domain theory, show that each of the positive examples is a member of the explainable concept and generalize them using explanation-based learning.
2. Disjunctively combine the resulting generalizations to form the explanatory part of the the definition, $D_e$.
3. Discard any negative examples that do not satisfy this explanatory component.
4. Remove the explainable features in $D_e$ from the descriptions of the remaining examples.
5. Give the "reduced" set of examples to a standard empirical learning system to compute the unexplainable component of the definition, $D_u$.
6. Output: $D_e \& D_u$ as the final concept description.

out common expressions. For the sample problem, this produces the following explanatory component:

> $D_e$: has-bottom & flat-bottom & lightweight & has-concavity & upward-pointing-concavity & (width=small & insulating OR has-handle)

Step three discards negative examples that do not even satisfy the explanatory component. These negative examples are already eliminated as potential members of the intended concept and require no further consideration. In the sample problem, the negative example can-1 can be discarded. Although it is a stable open-vessel, it is not graspable, because it is not insulating nor does it have a handle. Therefore it cannot function as a drinking vessel for hot and cold liquids.

Table 4: Reduced Examples for Learning Cup.

| | cup-1 (+) | cup-2 (+) | shot-glass-1 (-) | mug-1 (-) |
|---|---|---|---|---|
| color | white | red | white | copper |
| volume | small | small | tiny | large |
| shape | cylinder | cylinder | cylinder | cylinder |

Step four removes the explainable features of the remaining examples to allow the empirical component to focus on their unexplainable aspects. The resulting reduced set of data for the sample problem is shown in Table 4. In step five, the unexplained data is given to a standard empirical system for learning from examples. We normally use a version of ID3 (Quinlan, 1986) as the empirical component. ID3 builds decision trees which IOU translates into a set of rules so that the final concept definition is in a uniform language. For the sample problem, ID3 generates the description volume=small for discriminating the cups from the shot glass and the mug. Like many rule-based induction systems, ID3 is biased towards simple, more-general descriptions, and this is the simplest description of the cups that excludes the non-cups.

However, ID3 is not a particularly good algorithm for modeling human empirical learning. Because it tries to construct a minimal discriminant description, it can fail to capture all of the similarities among the positive examples. Therefore, a standard *most-specific-conjunctive* (MSC) learning algorithm (Haussler, 1988) can also be used as the empirical component of IOU. Early experiments by Bruner, Goodnow, and Austin (1956) indicated that human subjects frequently use the MSC strategy (which they call *wholist*) when learning concepts from examples. This algorithm simply forms the conjunction of all feature-value pairs present in all of the positive examples. For the examples in Table 4, this method produces the description: volume=small & shape=cylinder.

The final step of IOU simply combines the explanatory and nonexplanatory components into a final concept definition. This produces the following definition for cup:

> has-bottom & flat-bottom & lightweight & has-concavity & upward-pointing-concavity & (width=small & insulating OR has-handle) & volume=small & shape=cylinder → cup.

IOU actually maintains the explanatory and nonexplanatory components separately in order to allow them to be treated differently during classification (see Section 4.2.2).

It is informative to compare IOU's results on this simple problem to those of purely empirical learning systems. When ID3 is run by itself on the data in Table 2, the extra example can-1 causes color to be the most informative feature and the system produces the following rule:

> color=red OR (color=white & has-handle) → cup

ID3 would clearly need many more examples to learn the correct concept. Applying the MSC algorithm to the examples in Table 2 generates the description:

> has-bottom & flat-bottom & lightweight & has-concavity & upward-pointing-concavity & width=small & volume=small & shape=cylinder → cup.

This description is inconsistent with the training data since it still covers the negative example can-1. This is because the correct definition of the concept requires disjunction (i.e. there are no necessary and sufficient conditions). IOU uses the domain theory to learn the disjunctive portion of the concept, specifically, the two alternative ways to achieve graspability. Mooney (in press) presents additional theoretical and experimental evidence that IOU learns more accurate concepts from fewer examples than pure empirical learning methods.

In its current implementation, IOU is a "batch" learning system and processes all of the training examples at once. However, the basic algorithm is easily made incremental if the empirical learner is itself incremental. The explanatory part of the definition can be assembled incrementally by disjunctively adding the explanation-based generalization of each new positive example. Also, each time a new example is encountered it is either discarded as an unprovable negative example or its explainable features are removed and the remaining features are passed along to the empirical component, which incrementally forms the nonexplanatory part of the definition. For example, one could use an incremental version of ID3 (e.g. Utgoff's (1989) ID5) or the MSC algorithm (which is naturally incremental).

## 4.2 Psychological Relevance of IOU

Although some of the ideas underlying IOU were derived from psychological results, it was not specifically designed as a model of human category learning. Nevertheless, there are two recent psychological studies that are relevant to viewing IOU as a cognitive model. First, there is an experiment by Ahn and Brewer (1988) that motivated the development of IOU by demonstrating that subjects learn explanatory and nonexplanatory aspects of a concept separately. Second, there is an experiment by Wisniewski (1989) demonstrating that subjects learn different concepts from the same examples depending on whether or not they are told the function of the categories.

### 4.2.1 Learning Explanatory and Nonexplanatory Information

Some recent experiments by Ahn and Brewer (1988) were one of the original motivations behind the development of IOU. These experiments were designed to follow up previous experiments by Ahn, Mooney, Brewer, and DeJong (1988) that investigated subjects' ability to use explanation-based learning to acquire a plan schema from a single instance. The original experiments revealed that, like an explanation-based system, human subjects could acquire a general plan schema from a single specific instance described in a narrative. The follow-up experiments explored subjects' ability to learn event schemata that contain both explainable and unexplainable (conventional) aspects after receiving only a single example, and after receiving multiple examples. For example, one of the schemata used in the experiments is the potlatch ceremony conducted by American Indian tribes of the Northwest. If one has the appropriate knowledge regarding the goals and customs of these Indians, many aspects of the potlatch ceremony can be explained in terms of a plan to increase the social status of the host. However, there are also a number of ritualistic features of the ceremony that cannot be explained in this manner.

The results of this experiment indicated that the explainable aspects of the potlatch ceremony were acquired after exposure to only a single instance, while the nonexplanatory aspects of the ceremony were only acquired after multiple instances were presented. This supports the view that people use different learning mechanisms to acquire these different aspects of a concept, as in the IOU method. Subjects were also asked to rate their confidence in their assertions that a component is a part of the general ceremony. The subjects' confidence ratings for explanatory components were significantly higher than for nonexplanatory ones after both one and two instances. Also, multiple examples increased subjects' confidence and accuracy with respect to nonexplanatory components but not with respect to explanatory ones. This suggests that, like IOU, people treat explanatory and nonexplanatory aspects of a concept differently.

### 4.2.2 Simulating the Effects of Functional Knowledge

This section demonstrates IOU's ability to model the specific results of some additional psychological experiments exploring the effect of background knowledge on concept learning (Wisniewski, 1989). It is important to note that IOU was not specifically designed to simulate these results, but rather the basic method was already developed when the author learned of the results of this experiment. In Wisniewski's experiment, two groups of subjects performed a standard learning-from-examples task. Both groups received the same examples, but one group, the *function group*, was told the functions of the two categories to be discriminated and the other, the *discrimination group*, was not. For example, the function group was told that one category was used for killing bugs and the contrast category was used for wallpapering. Examples were described by a number of features. A particular feature value could be either *predictive* or *nonpredictive* of a particular category. In the training set containing 15 examples of each category, all examples containing a predictive feature value of a category were members of that category and 80% of the category members had the predictive feature value (the other 20% were missing a value for this feature). Nonpredictive feature values occurred equally often in both categories. A feature value was also *core* or *superficial.* A core feature value was relevant to a category's function, while a superficial one was not. For example "contains poison" was a core feature value of the category whose function was "for killing bugs," while "manufactured in Florida" was superficial. Each category contained three superficial feature values (two predictive and one nonpredictive) and two core feature values (one predictive the other nonpredictive). The superficial-nonpredictive feature value of a category was the core-nonpredictive feature value of its contrast category. Table 5 shows the different types of features for two contrasting categories used in the experiment. Each training example of a category contained 4 of the 5 characteristic features from this table. Each training example also had two additional features with random values. An example of a training example in the category "mornek" is:

Table 5: Different Feature Types for Experimental Categories.

| **Mornek** | **Plapel** |
|---|---|
| Function: for killing bugs | Function: for wallpapering |
| sprayed on plants **C-P** | sprayed on walls **C-P** |
| contains poison **C-NP** | contains poison **S-NP** |
| contains a sticky substance **S-NP** | contains a sticky substance **C-NP** |
| stored in a garage **S-P** | stored in a basement **S-P** |
| manufactured in Florida **S-P** | manufactured in Ohio **S-P** |

| | |
|---|---|
| **C-P**: core predictive | **C-NP**: core non-predictive |
| **S-P**: superficial predictive | **S-NP**: superficial non-predictive |

> sprayed on plants, contains poison, contains a sticky substance, stored in a garage, manufactured in Florida, best if used in 1 year, comes in 16 oz container.

where the example is missing only one of the characteristic features of “morneks” (sprayed on plants) and the last two features have random values.

After learning the training data, subjects were given 10 test examples of each category. Each of the ten examples of a category contained more predictive features of that category than the contrast category. *Superficial-core** test examples contained the two superficial-predictive feature values of the category and the two core feature values of the contrast category. *Core* examples contained just the core feature values of the category, while *superficial* examples contained just the superficial-predictive feature values. *Core-superficial* examples contained all of the core and superficial feature values. Each test example also had two extra random feature values. Sample test examples for the Mornek category are shown in Table 6

Table 6: Sample Test Items for Mornek.

| **superficial-core*** | **core** |
|---|---|
| stored in a garage **S-P**<br>manufactured in Florida **S-P**<br>contains a sticky substance **C-NP***<br>sprayed on walls **C-P***<br>best if used within 1 year **R** | contains poison **C-NP**<br>sprayed on plants **C-P**<br>best if used within 5 years **R**<br>came in a 32-ounce container **R** |
| **superficial** | **superficial-core** |
| stored in a garage **S-P**<br>manufactured in Florida **S-P**<br>best if used within 1 year **R**<br>came in a 16-ounce container **R** | stored in a garage **S-P**<br>manufactured in Florida **S-P**<br>contains a sticky substance **S-NP**<br>contains poison **C-NP**<br>sprayed on plants **C-P**<br>best if used within 1 year **R** |

| | |
|---|---|
| **S-P**: superficial predictive | **S-NP**: superficial non-predictive |
| **C-P**: core predictive | **C-NP**: core nonpredictive |
| **C-P***: core predictive (of other class) | **C-NP***: core nonpredictive (of other class) |
| **R**: random | |

Subjects in both groups were asked to rate their confidence in the category of each test example on a scale of 1 to 7, where 1 was most confident for the "wrong" category and 7 most confident for the "right" category. In general, the results demonstrated that subjects in the function group attributed more relevance to the core feature values while the discrimination group relied more heavily on superficial predictive features (see Table 7). However, the function group also made some use of superficial-predictive features values, indicating they were using a combination of empirical and explanation-based techniques.

In the simulation, IOU was used to model the performance of the function group and a standard most-specific-conjunctive empirical method (MSC) was used to model the discrimination group. In both cases, the systems were given the same training and test examples given to the subjects. Simple intuitive domain theories were constructed for connecting core feature values to category membership. For example, IOU's overly-general theory for Mornek is given below:

contact-bugs & deadly → kills-bugs

contains-poison → deadly

electric-shock → deadly

sprayed-on=plants → contact-bugs

emits-light & location=outdoors → contact-bugs

The MSC method was also used as the empirical component of IOU. In order to accommodate missing feature values, only features that appear with different values in different positive examples are actually deleted from the most-specific conjunctive description. This has the same effect as replacing missing features with their most probable value given the class (Quinlan, 1986) before forming the MSC description.

Since all of the core and superficial features of a category shown in Table 5 are either present or missing a value in all of its examples, the most-specific conjunctive description of a category contains all of these characteristic features. Since the two remaining features ("best if used in" and "container size") have different random values, they are dropped from the MSC description. Since the two core features of each category are explained by the domain theory, they comprise the explanatory component of IOU's concept description. Since the superficial features of a category are either present or missing a value in all of its examples, they are all included in the MSC description of the unexplained features; however, the random features are dropped. Therefore, IOU's category descriptions also include all of the core and superficial features of the category. However, IOU separates them into explanatory and nonexplanatory components. For example, the concept description for "morneks" produced by both IOU and MSC is (explanatory features are in small caps):

> SPRAYED-ON=PLANTS & CONTAINS-POISON & contains-sticky & stored-in=garage & manufactured-in=florida

The following equation was used to produce a confidence rating ($1 \leq C \leq 7$) for the test examples:

$$C = 4 + 1.5(M_1 - M_2)$$

$M_1$ and $M_2$ are match scores ($-1 \leq M_i \leq 1$) for the two categories computed by examining each feature-value pair in the most-specific-conjunctive description for the category and scoring as follows: +1 if the example had the same value, 0 if the feature was missing, and -1 if it had a conflicting value. The result was scaled by dividing by the maximum possible score. For IOU, explanatory (core) features were weighted more heavily by having them count twice as much (i.e. the match score was incremented or decremented by 2 instead of

Table 7: Average Confidence Ratings for Test Examples

| | Subjects | | Simulation | |
|---|---|---|---|---|
| Item Type | Function | Discrimination | IOU | MSC |
| superficial-core* | 4.00 | 5.02 | 3.79 | 4.60 |
| core | 6.16 | 5.93 | 5.07 | 4.60 |
| superficial | 6.04 | 6.36 | 4.86 | 5.20 |
| superficial-core | 6.43 | 6.54 | 5.71 | 5.80 |

1). This scoring technique is a simple method for obtaining a confidence rating between 0 and 7 based on the degree to which an example matches the MSC description of each of the two categories. Several other similar scoring mechanisms were tried without any significant effect on the qualitative results. The important factor is that the score is high when an example matches the description of the first category more than the second and that it is low when an example matches the description of the second category more than the first. The qualitative results are also insensitive to the exact additional weighting assigned to the explanatory features (a weighting factor of 1.5 or 3 works as well as 2).

Table 7 shows both the original experimental results and the results of the simulation. Although the exact confidence values of the simulation do not match the subjects, all of the important differences mentioned by Wisniewski (1989) are present. For the superficial-core* items, the IOU (function) scores are lower than the MSC (discrimination) scores. Although these items have the superficial features of the "correct" category, they have the core features of the contrast category causing the function group (and IOU) to rate them lower. IOU (function group) scores the core items higher than the superficial items, while

MSC (discrimination group) scores the superficial items higher than the core items. Finally, the IOU (function) scores are lower than the MSC (discrimination) scores for the superficial items but higher for the core items.

All of these correctly modeled effects stem from IOU's separation of concepts into explanatory and nonexplanatory components and its scoring procedure that weights the explanatory features more heavily. Since IOU is unique among current integrated machine learning systems in separating its concepts into explanatory and nonexplanatory components, it seems clear that other existing systems would be unable to model these results. However, the effects are not particularly dependent on the specific details of the IOU algorithm; and therefore other methods that include both explanatory and nonexplanatory features in their concepts and weight the former more heavily may also be able to model these results.

# 5 Theory Revision in Either

IOU uses a theory to bias induction but it cannot modify a theory to account for anomalous empirical data. Either (Explanation-Based and Inductive Theory Extension and Revision), is a more recent system that can actually revise an existing domain theory to fit a set of data. As revealed by explanation-based learning, category knowledge is frequently best viewed as a complex set of interacting rules (a domain theory) rather than a simple set of features, conditional probabilities, or exemplars. Some of these rules may have been learned from direct instruction, and others may have been induced from examples of previously learned concepts. In any case, learning a new concept can involve using empirical data to revise an existing domain theory by modifying, adding, or deleting rules.

Either is a complicated system that has evolved over several years of research. In this paper, we only have space to present an overview of the basic system. However, it should be noted that Either has already successfully revised two real-world rule-bases. One of these identifies biologically important DNA sequences called "promoters" and the other diagnoses diseased soybean plants. Interested readers are referred to Ourston and Mooney (1990) and Mooney and Ourston (1991b) for more details.

## 5.1 The Theory Revision Problem

Either combines explanation-based and empirical methods to provide a focused correction to an imperfect domain theory. The explanation-based part of the system identifies the failing parts of the theory, and constrains the examples used for induction. The empirical part of the system determines specific corrections to failing rules that renders them consistent with the empirical data. Table 8 more precisely defines the problem addressed by Either. It is difficult to precisely define the term "minimal" used to characterize the revision to be produced. Since it is assumed that the original theory is "approximately correct" the goal is to change it as little as possible. Syntactic measures such as the total number of symbols added or deleted are reasonable criteria. Either uses various heuristic methods to help insure that its revisions are minimal in this sense.

Table 8: The EITHER Problem.

**Given:**

- A set of positive and negative examples of a concept each described by a set of observable features.
- An imperfect propositional Horn-clause domain theory for the concept.

**Determine:**

A minimally revised version of the domain theory that is consistent with the examples.

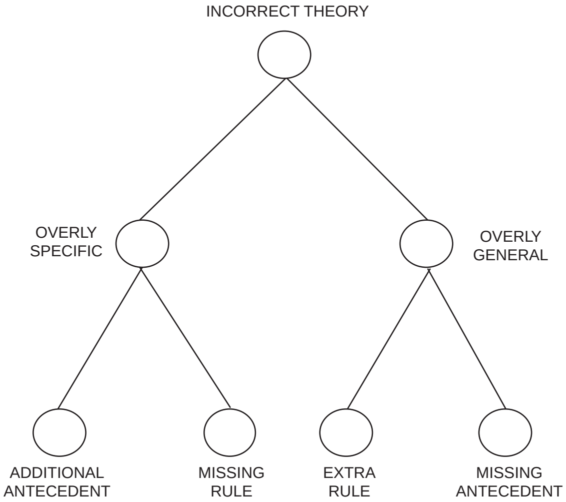

Figure 3: Taxonomy of Incorrect Theories

A domain theory can be incorrect in various ways. Figure 3 shows a taxonomy of incorrect theories. At the top level, theories can be incorrect because they are either overly-general or overly-specific. An overly-general theory entails membership for some examples that are not members of a category. One way a theory can be overly general is when rules lack required antecedents, providing proofs for examples that should have been excluded. Another way a theory can be overly-general is when a completely incorrect rule is present.

Table 9: Examples for Theory Revision.

| | cup-1 (+) | cup-2 (+) | cup-3 (+) | can-1 (-) | bowl-1 (-) | bowl-2 (-) |
|---|---|---|---|---|---|---|
| has-bottom | true | true | true | true | true | true |
| flat-bottom | true | true | true | true | true | true |
| has-concavity | true | true | true | true | true | true |
| upward-pointing | true | true | true | true | true | true |
| lightweight | true | true | true | true | true | true |
| has-handle | false | true | true | false | false | false |
| width | small | medium | medium | small | medium | medium |
| insulating | true | true | true | false | true | true |
| color | red | blue | tan | gray | red | blue |
| volume | small | small | small | small | small | large |
| shape | round | round | cylinder | cylinder | round | round |

By contrast, an overly-specific theory fails to entail membership for some examples of a category. This can occur because the theory is missing a rule which is required in the proof of concept membership, or because existing rules have additional antecedents that exclude concept members. Consequently, incorrectly classified examples can be of two types. A *failing positive* refers to an example that is not provable as a member of its own category. This indicates a need for generalizing the theory by adding rules or deleting antecedents. A *failing negative* refers to an example that is provable as a member of a category other than its own. This indicates a need to specialize a theory by adding antecedents or deleting rules.

As a concrete example, consider various errors that might occur in the theory for cup (drinking-vessel) introduced in Section 2. Assume the set of training examples is shown in Table 9. These examples differ only in graspability. If the theory is missing the rule: has-handle → graspable, then cup-2 and cup-3 can no longer be shown to be cups and are therefore failing positives, indicating that the theory needs to be generalized. If the theory is missing the antecedent width=small from the rule: width=small & insulating → graspable, then bowl-1 and bowl-2 can be incorrectly shown to be cups and are therefore failing negatives, indicating that the theory needs to be specialized. If the theory has both of these faults, then cup-2 is a failing positive and bowl-1 and bowl-2 are failing negatives. Given the examples in Table 9, EITHER can revise the theory to correct for either or both of these faults.

## 5.2 Overview of the Theory Revision Algorithm

This section reviews EITHER's basic revision method, which integrates deductive, abductive[3], and inductive reasoning. The system's top-level architecture is shown in Figure 4. EITHER first attempts to fix failing positives by removing or generalizing antecedents and to fix failing negatives by removing rules or specializing antecedents since these are simpler and less powerful operations. Only if these operations fail does the system resort to the more powerful technique of using induction to learn new rules to fix failing positives, and to learn new antecedents to add to existing rules to fix failing negatives.

EITHER initially uses deduction to identify failing positives and negatives among the training examples. It uses the proofs generated by deduction to find a near-minimal set of rule retractions that would correct all of the failing negatives. During the course of the correction, deduction is also used to assess proposed changes to the theory as part of the

[3] Abduction is the process of finding sets of assumptions that allow an observation to be explained (Peirce, 1931-1958; Charniak & McDermott, 1985)

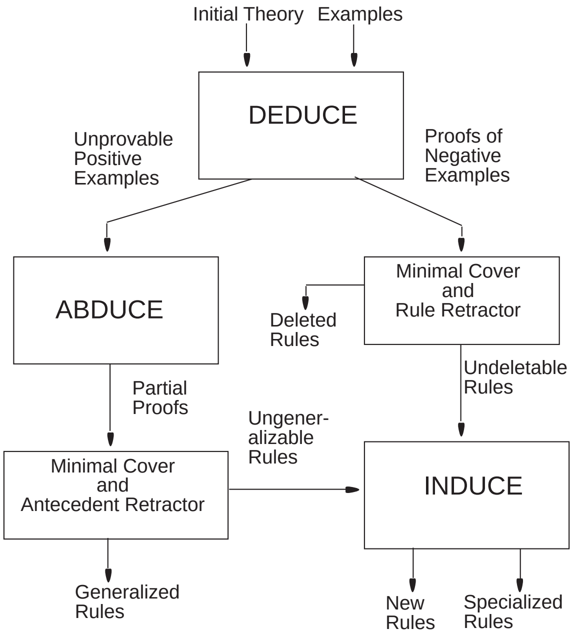

Figure 4: EITHER Architecture

generalization and specialization processes.

EITHER uses abduction to initially find the incorrect part of an overly-specific theory. In EITHER, abduction identifies sets of assumptions that allow a failing positive to become provable. These assumptions identify rule antecedents (called *conflicting antecedents*) that, if deleted, would properly generalize the theory and correct the failing positive. EITHER uses the output of abduction to find a near-minimum set of conflicting antecedents whose removal would correct all of the failing positives.

If rule and antecedent retraction are insufficient, induction is used to learn new rules or to determine additional antecedents to add to existing rules. EITHER uses the output of abduction and deduction to determine an appropriately labeled subset of the training examples to pass to induction in order to form a consistent correction. EITHER currently uses a version of ID3 (Quinlan, 1986) as its inductive component. As in IOU, the decision trees returned by ID3 are translated into equivalent rules. The remaining components of the EITHER system constitute generalization and specialization control algorithms, which identify and specify the types of corrections to be made to the theory.

As an example of the generalization process, consider missing the has-handle rule from the theory for cups. This results in cup-2 and cup-3 becoming failing positives. These examples are almost provable as cups except they cannot be shown to be graspable. Consequently, EITHER first focuses on the remaining rule for graspable and attempts to retract the antecedent width=small in order to make the failing positives provable. However, this over-generalizes and results in bowl-1 and bowl-2 becoming failing negatives. Consequently, the system uses cup-2 and cup-3 as positive examples and bowl-1 and bowl-2 as negative examples to induce a new rule for graspable. Since the single feature has-handle distinguishes these examples, ID3 induces the correct rule: has-handle → graspable.

As an example of the specialization process, consider missing the antecedent width=small from the rule width=small & insulating → graspable. EITHER first attempts to retract the resulting overly-general rule: insulating → graspable, in order to remove the faulty proofs of the failing negatives, bowl-1 and bowl-2. The system focuses on this rule because its removal eliminates the faulty proofs of the failing negatives while minimizing the number of failing positives created in the process. Since retracting this rule does create one failing positive (cup-1) the system decides it needs to specialize the rule by adding antecedents. Consequently, the system uses cup-1 as a positive example and bowl-1 and bowl-2 as negative examples and passes them to ID3. Since the value of the single feature width distinguishes these examples, ID3 finds the correct missing antecedent width=small and adds it to the overly-general rule for graspable.

## 5.3 Theory for Data Interpretation

EITHER also uses its theory to augment the representation of examples prior to passing them to induction. Using a process frequently referred to as *constructive induction* (Michalski, 1983; Drastal et al., 1989; Mooney and Ourston, 1991a), the domain theory is used to deduce higher-level features from the observable features describing the examples. Specifically, when using induction to learn a new rule or to determine which antecedents to add to an existing rule, forward deduction is used to identify the truth values of all intermediate concepts [4] for each of the failing examples. Intermediate concepts that can be deduced are then fed to the inductive learner as additional features. If the truth value of an intermediate concept is highly correlated with the class of the failing examples, then it is likely to be used by the inductive learner.

For example, assume that the cup theory is missing the rule for liftable, but is otherwise correct. Performing forward deduction on the failing positives (all of the cup examples in

[4] An *intermediate concept* is any term in a domain theory that is neither an observable feature used to describe examples nor a category into which examples are to be eventually classified.

this case) will always add the feature graspable, since all cups are graspable. Therefore, the description of the positive examples is augmented with the higher-level feature graspable prior to being used for induction. In other words, the existing theory is used to interpret and redescribe the examples. Since the added graspable feature helps to discriminate between the positive and negative examples of liftable, it is very likely to be used by the empirical learner. Consequently, when the rule for liftable is removed from the theory, EITHER usually relearns the correct rule (graspable & lightweight → liftable) after 20 random training examples.

As another example of this process, assume the cup theory is missing the top-level rule: stable & liftable & open-vessel → cup, but is otherwise correct. In this case, forward deduction adds the features stable, liftable, and open-vessel to each of the positive examples. These high-level features can then be used by the inductive subsystem to help discriminate the positive and negative examples. Consequently, when the cup rule is removed from the theory, EITHER usually relearns the correct rule after about 30 random training examples. If the theory is not used to interpret the data, then 80 examples are usually required to learn the correct definition of cup directly in terms of observable features.

## 5.4 Psychological Relevance of EITHER

Like IOU, EITHER was not specifically designed to model human category learning; however, many of its basic goals and methods have some psychological relevance. In particular, Wisniewski and Medin (1991) report some relevant psychological results on theory revision

Figure 5: Sample Drawings from Wisniewski & Medin (1991)

and data interpretation. Their experiments studied subjects learning to categorize children's drawings of a person. Some of the drawings used in the experiments are shown in Figure 5. The methodology is a basic learning from examples paradigm except one group of subjects, the *standard group*, were given meaningless category names while subjects in the *theory group* were given meaningful names such as "drawings by high IQ children" vs. "drawings by low IQ children" or "drawings by farm children" vs. "drawings by city children." Subjects in both groups were asked to write down a rule for each category that someone else could use to accurately categorize the drawings.

One aspect of the subjects' rules that Wisniewski and Medin analyzed was the degree of abstraction. They divided rules into the following three types:

Table 10: Frequency of Rule Types

| | Standard Group | Theory Group |
|---|---|---|
| Concrete Rules | 81% | 35% |
| Abstract Rules | 16% | 37% |
| Linked Rules | 3% | 28% |

- *Concrete:* Consisting of simple features that are easily observable, e.g. "buttons or stripes on their shirts and dark, thick hair."

- *Abstract:* Consisting of features that are more complex, higher level, or less perceptual, e.g. "look more normal."

- *Linked:* Consisting of connected abstract and concrete features, e.g. "added more detail such as teeth."

They found that subjects in the theory group produced more abstract and linked rules compared to the standard group. The specific results are shown in Table 10. These results are nicely explained by the hypothesis that subjects in the theory group are using their background theories to interpret the data. Like EITHER adding graspable to its description of a cup, they are inferring abstract features and adding them to the data before inducing a rule. Linked rules occur because the subjects are also writing down the concrete features from which their abstract features were derived.

Wisniewski and Medin also note that subjects given different meaningful labels for categories extract different high-level features from the drawings. A subject told that the drawings in Category 1 of Figure 5 were drawn by creative children interpreted the part of drawing 5 indicated by the arrow as "buttons." The subject mentioned this feature as evidence for detail, which implied creativity. On the other hand, a subject told that this figure was drawn by a city child interpreted it as a "tie." This subject mentioned the feature as evidence that the person was a business-person, implying it was drawn by a city person. This phenomenon is nicely explained by the hypothesis that different category labels "activate" different domain theories and therefore result in different abstract features being derived from the perceptual data.

Wisniewski and Medin also found evidence for theory revision in their results. Based on the data, subjects would sometimes change their definition of an abstract feature. In one particular case, a subject mentioned that a drawing depicted detailed clothing and therefore must have been drawn by a creative child. When told that the drawing was done by a noncreative child, they changed their definition for what counted as "detail." This is similar to EITHER altering its definition of graspable after misclassifying some examples or counter-examples of cups. In another case, a subject initially stated that a drawing was done by a city child because "it looks very detailed, has colored-in places." When told that it was actually drawn by a farm child, the subject specialized his/her rule: detail → city-child by adding a constraint induced from the data. Specifically, the person stated that "drawings with detail in specific clothing is more of a rule for city kids – not detail in body movement as this one had."

# 6 Conclusions

Recent results in both machine learning and cognitive psychology demonstrate that effective category learning involves an integration of theory and data. Theories can bias induction and alter the representation of data, and conflicting data can result in theory revision. This paper has reviewed two recent machine learning systems that attempt to integrate theory and data. IOU uses a domain theory to acquire part of a concept definition and to focus induction on the unexplained aspects of the data. EITHER uses data to revise an imperfect theory and uses theory to add abstract features to the data. Recent psychological experiments reveal that subjects perform many of the same operations as these machine learning systems. Like IOU, people separate category definitions into explanatory and nonexplanatory components, acquire explanatory components earlier, and have more confidence in explanatory aspects. Like EITHER, people use background theories to derive abstract features from the data, and revise portions of their theories to account for conflicting data.

Nevertheless, in many ways, current machine learning systems are not nearly as adept as people at integrating theory and data in learning. Particular areas requiring further research concern revising probabilistic and relational theories. Most current integrated learning systems are restricted to theories expressed in propositional logic. Consequently, they are incapable of reasoning about their confidence in their theories and conclusions, and cannot handle complex, relational descriptions that require the expressive power of first-order predicate logic. These areas of machine learning are just beginning to be explored (Richards & Mooney, 1991; Pazzani, Brunk & Silverstein, 1991; Fu, 1989). In general, the interaction between theory and data in learning has just begun to be investigated.

From a machine-learning perspective, methods for integrating theory and data in learning can greatly improve the development of intelligent systems. Standard methods for building knowledge bases by interviewing experts are laborious and error-prone. Standard machine learning methods for learning from examples are also inadequate since one rarely has enough data to induce a complete and correct knowledge base from scratch. In addition, machine-induced knowledge fails to make use of existing human concepts and is therefore frequently unable to provide comprehensible explanations for the conclusions it warrants. Theory revision, on the other hand, allows a system to accept an incomplete, approximate knowledge base and refine it through experience. People acquire expertise through a combination of abstract instruction and experience with specific cases, and machine learning systems that integrate theory and data are trying to successfully emulate this approach.

From a psychological perspective, methods for integrating theory and data can hopefully improve our understanding of human category learning. Artificial learning problems that minimize the role of prior knowledge are not representative of the categorization problems that people confront every day. Machine learning algorithms that can simulate psychological data on the effect of prior knowledge on learning can provide valuable insight into how people learn in more natural settings. In turn, understanding the specific ways in which theory and data interact in human learning can hopefully lead to the development of more effective educational methods for combining the presentation of abstract rules and principles with concrete examples.

## References

Aha, D.W., Kibler, D., & Albert, M.K. (1991). Instance-based learning algorithms. *Machine Learning, 6*, 37-66.

Ahn, W. & Brewer, W.F. (1988). Similarity-based and explanation-based learning of explanatory and nonexplanatory information. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 524-530). Hillsdale, NJ: Erlbaum.

Ahn, W., Brewer, W.F., & Mooney, R.J. (in press). Schema acquisition from a single example. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Ahn, W., Mooney, R.J., Brewer, W.F., & DeJong, G.F. (1987). Schema acquisition from one example: Psychological evidence for explanation-based learning. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (pp. 50-57). Hillsdale, NJ: Erlbaum.

Barsalou, L.W. (1983). Ad hoc categories. *Memory and Cognition, 11*, 211-27.

Birnbaum, L.A. & Collins, G.C. (Eds.) (1991). Learning from theory and data. Section of *Proceedings of the Eighth International Workshop on Machine Learning* (pp. 475-573).

San Mateo, CA: Morgan Kaufman.

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

Cohen, W.W. (1990). Learning from textbook knowledge: A case study. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 743-748). Cambridge, MA: MIT Press.

Charniak, E. & McDermott, D. (1985). *Introduction to Artificial Intelligence.* Reading, MA: Addison Wesley.

Danyluk, A. P. (1991). Gemini: An integration of explanation-based and empirical learning. *Proceedings of the First International Workshop on Multistrategy Learning* (pp. 191-206). Fairfax, VA: George Mason University.

DeJong, G.F. (1988). An introduction to explanation-based learning. In H. Shrobe (Ed.), *Exploring Artificial Intelligence* (pp. 45-81). San Mateo, CA: Morgan Kaufman.

DeJong, G.F. & Mooney, R.J. (1986). Explanation-based learning: An alternative view. *Machine Learning, 1*, 145-176.

Drastal, G., Czako, G., & Raatz, S. (1989). Induction in an abstraction space: A form of constructive induction. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* (pp. 708-712). San Mateo, CA: Morgan Kaufman.

Fisher, D.H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine*

*Learning, 2*, 139-172.

Flann, N.S. & Dietterich, T.G. (1989). A study of explanation-based methods for inductive learning. *Machine Learning, 4*, 187-226.

Fu, L. (1989). Integration of neural heuristics into knowledge-based inference. *Connection Science, 1*, 325-339.

Ginsberg, A. (1990). Theory reduction, theory revision, and retranslation. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 743-748). Cambridge, MA: MIT Press.

Haussler, D. (1988). Quantifying inductive bias: Artificial intelligence algorithms and Valiant's learning framework. *Artificial Intelligence, 36*, 177-221.

Hirsh, H. (1990). *Incremental version space merging: A general framework for concept learning.* Hingham, MA: Kluwer.

Medin, D.L. & Shaeffer, M.M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

Medin, D.L., Wattenmaker, W.D., & Michalski, R.S. (1987). Constraints and preferences in inductive learning: An experimental study of human and machine performance. *Cognitive Science, 11*, 299-239.

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.* Palo Alto, CA: Tioga.

Michalski, R.S. & Chilausky, R.L. (1980). Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. *Policy Analysis and Information Systems, 4*, 125-160.

Michalski, R.S. & Teccuci, G. (Eds.) (1991). *Proceedings of the First International Workshop on Multistrategy Learning.* Fairfax, VA: George Mason University.

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

Mooney, R.J. (in press). Induction over the unexplained: Using overly-general domain theories to aid concept learning. *Machine Learning.*

Mooney, R.J. & Ourston, D. (1991a). Constructive induction in theory refinement. *Proceedings of the Eighth International Workshop on Machine Learning* (pp. 178-182). San Mateo, CA: Morgan Kaufman.

Mooney, R.J. & Ourston, D. (1991b). A multi-strategy approach to theory refinement. *Proceedings of the First International Workshop on Multistrategy Learning* (pp. 115-130). Fairfax, VA: George Mason University.

Murphy, G.L. & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289-316.

Nakamura, G.V. (1985). Knowledge-based classification of ill-defined categories. *Memory and Cognition, 13*, 377-84.

Ourston, D. & Mooney, R. (1990). Changing the rules: A comprehensive approach to theory refinement. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 815-820). Cambridge, MA: MIT Press.

Pazzani, M.J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 416-432.

Pazzani, M.J., Brunk, C, & Silverstein, G. (1991). A knowledge-intensive approach to learning relational concepts. *Proceedings of the Eighth International Workshop on Machine Learning* (pp. 432-436). San Mateo, CA: Morgan Kaufman.

Peirce, C.S. (1931-1958). *Collected papers*, 8 vols. Edited by C. Hartshorne, P. Weiss, and A. Burks. Cambridge, MA: Harvard University Press.

Posner, M.I. & Keele, S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353-363.

Quinlan, J. R. (1979). Discovering rules from large collections of examples: A case study. In D. Michie (Ed.), *Expert systems in the microelectronic age.* Edinburgh: Edinburgh University Press.

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

Rajamoney, S.A. (1990). A computational approach to theory revision. In J. Shrager and P. Langley (Eds.) *Computational models of scientific discovery and theory formation* (pp. 225-254). San Mateo, CA: Morgan Kaufman.

Richards, B.L. & Mooney, R.J. (1991). First order theory revision. *Proceedings of the Eighth International Workshop on Machine Learning* (pp. 447-451). San Mateo, CA: Morgan Kaufman.

Rosenblatt, F. (1962). *Principles of neurodynamics and the theory of brain mechanisms*, Washington, D.C.: Spartan Books.

Schank, R.C., Collins, G.C., & Hunter, L.E. (1986). Transcending inductive category formation in learning. *Behavioral and Brain Sciences, 9*, 639-686.

Segre, A.M. (Ed.) (1989). Combining empirical and explanation-based learning. Section of *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 2-92). San Mateo, CA: Morgan Kaufman.

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

Towell, G.G. & Shavlik, J.W. (1991). Refining symbolic knowledge using neural networks. *Proceedings of the First International Workshop on Multistrategy Learning* (pp. 257-272). Fairfax, VA: George Mason University.

Utgoff, P.E. (1989). Incremental induction of decision trees. *Machine Learning, 4*, 161-186.

Winston, P.H., Binford, T.O., Katz, B., & Lowry, M. (1983). Learning physical descriptions from functional definitions, examples, and precedents. *Proceedings of the Third National Conference on Artificial Intelligence* (pp. 433-439). San Mateo, CA: Morgan Kaufman.

Wisniewski, E.J. (1989). Learning from examples: The effect of different conceptual roles. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 980-986). Hillsdale, NJ: Earlbaum.

Wisniewski, E.J. & Medin, D.L. (1991). Harpoons and long sticks: The interaction of theory and similarity in rule induction. In D.H. Fisher, M.J. Pazzani, & P. Langley (Eds.), *Concept Formation: Knowledge and Experience in Unsupervised Learning.* San Mateo, CA: Morgan Kaufman.