1

AN OVERVIEW OF MACHINE LEARNING

Jaime G. Carbonell
Carnegie-Mellon University

Ryszard S. Michalski
University of Illinois at Urbana-Champaign

Tom M. Mitchell
Rutgers University

1.1 INTRODUCTION

Learning is a many-faceted phenomenon. Learning processes include the acquisition of new declarative knowledge, the development of motor and cognitive skills through instruction or practice, the organization of new knowledge into general, effective representations, and the discovery of new facts and theories through observation and experimentation. Since the inception of the computer era, researchers have been striving to implant such capabilities in computers. Solving this problem has been, and remains, a most challenging and fascinating long-range goal in artificial intelligence (AI). The study and computer modeling of learning processes in their multiple manifestations constitutes the subject matter of machine learning.

1.2 THE OBJECTIVES OF MACHINE LEARNING

At present, the field of machine learning is organized around three primary research foci:

- **Task-Oriented Studies**—the development and analysis of learning systems to improve performance in a predetermined set of tasks (also known as the "engineering approach")

- **Cognitive Simulation**—the investigation and computer simulation of human learning processes
- **Theoretical Analysis**—the theoretical exploration of the space of possible learning methods and algorithms independent of application domain

Although many research efforts strive primarily towards one of these objectives, progress towards one objective often leads to progress towards another. For instance, in order to investigate the space of possible learning methods, a reasonable starting point may be to consider the only known example of robust learning behavior, namely humans (and perhaps other biological systems). Similarly, psychological investigations of human learning may be helped by theoretical analysis that may suggest various plausible learning models. The need to acquire a particular form of knowledge in some task-oriented study may itself spawn new theoretical analysis or pose the question: "How *do* humans acquire this specific skill (or knowledge)?" This trichotomy of mutually challenging and supportive objectives is a reflection of the entire field of artificial intelligence, where expert systems research, cognitive simulation, and theoretical studies provide cross-fertilization of problems and ideas.

1.2.1 Applied Learning Systems: A Practical Necessity

At present, instructing a computer or a computer-controlled robot to perform a task requires one to define a complete and correct algorithm for that task, and then laboriously program the algorithm into a computer. These activities typically involve a tedious and time-consuming effort by specially trained personnel.

Present-day computer systems cannot truly learn to perform a task through examples or by analogy to a similar, previously-solved task. Nor can they improve significantly on the basis of past mistakes, or acquire new abilities by observing and imitating experts. Machine learning research strives to open the possibility of instructing computers in such new ways, and thereby promises to ease the burden of hand-programming growing volumes of increasingly complex information into the computers of tomorrow. The rapid expansion of applications and availability of computers today makes this possibility even more attractive and desirable.

When approaching a task-oriented knowledge acquisition task, one must be aware that the resultant computer systems must interact with humans, and therefore should closely parallel human abilities. The traditional argument that an engineering approach need not reflect human or biological performance is not truly applicable to machine learning. Since airplanes, a successful result of an almost pure engineering approach, bear little resemblance to their biological counterparts, one may argue that applied knowledge acquisition systems should be equally divorced from any consideration of human capabilities. This argument does not apply here because airplanes need not interact with or understand birds. Learning machines, on the other hand, will have to interact with the people who

make use of them, and consequently the concepts and skills they acquire—if not necessarily their internal mechanisms—must be understandable to humans.

1.2.2 Machine Learning as a Science

The question of what are the genetically-endowed abilities in a biological system (versus environmentally-acquired skills or knowledge) has fascinated biologists, psychologists, philosophers and artificial intelligence researchers alike. A clear candidate for a cognitive invariant in humans is the learning mechanism—the innate ability to acquire facts, skills and more abstract concepts. Therefore, understanding human learning well enough to reproduce aspects of that learning behavior in a computer system is, in itself, a worthy scientific goal. Moreover, the computer can render substantial assistance to cognitive psychology, in that it may be used to test the consistency and completeness of learning theories, and enforce a commitment to fine-structure process-level detail that precludes meaningless, tautological or untestable theories.

The study of human learning processes is also of considerable practical significance. Gaining insights into the principles underlying human learning abilities is likely to lead to more effective educational techniques. Thus, it is not surprising that research into intelligent computer-assisted instruction, which attempts to develop computer-based tutoring systems, shares many of the goals and perspectives with machine learning research. One particularly interesting development is that computer tutoring systems are starting to incorporate abilities to infer models of student competence from observed performance. Inferring the scope of a student's knowledge and skills in a particular area allows much more effective and individualized tutoring of the student.

An equally basic scientific objective of machine learning is the exploration of alternative learning mechanisms, including the discovery of different induction algorithms, the scope and limitations of certain methods, the information that must be available to the learner, the issue of coping with imperfect training data, and the creation of general techniques applicable in many task domains. There is no reason to believe that human learning methods are the only possible means of acquiring knowledge and skills. In fact, common sense suggests that human learning represents just one point in an uncharted space of possible learning methods—a point that through the evolutionary process is particularly well suited to cope with the general physical environment in which we exist. Most theoretical work in machine learning has centered on the creation, characterization and analysis of general learning methods, with the major emphasis on analyzing generality and performance rather than psychological plausibility.

Whereas theoretical analysis provides a means of exploring the space of possible learning methods, the task-oriented approach provides a vehicle to test and improve the performance of functional learning systems. By constructing and testing applied learning systems, one can determine the cost-effectiveness trade-offs and limitations of particular approaches to learning. In this way, in-

dividual data points in the space of possible learning systems are explored, and the space itself becomes better understood. Many of the chapters of this book can be viewed from this perspective.

1.2.3 Knowledge Acquisition versus Skill Refinement

There are two basic forms of learning: *knowledge acquisition* and *skill refinement*. When we say that someone learned physics, we mean that this person acquired significant concepts of physics, understood their meaning, and understood their relationship to each other and to the physical world. The essence of learning in this case is the acquisition of new knowledge, including descriptions and models of physical systems and their behaviors, incorporating a variety of representations—from simple intuitive mental models, examples and images, to completely tested mathematical equations and physical laws. A person is said to have learned more if his knowledge explains a broader scope of situations, is more accurate, and is better able to predict the behavior of the physical world. This form of learning is typical in a large variety of situations and is generally termed *knowledge acquisition*. Hence, knowledge acquisition is defined as learning new symbolic information coupled with the ability to apply that information in an effective manner.

A second kind of learning is the gradual improvement of motor and cognitive skills through practice, such as learning to ride a bicycle or to play the piano. Acquiring textbook knowledge on how to perform these activities represents only the initial phase in developing the requisite skills. The bulk of the learning process consists of refining the learned skills, whether mental or motor coordination, by repeated practice and by correcting deviations from desired behavior. This form of learning, often called *skill refinement*, differs in many ways from knowledge acquisition. Whereas the essence of knowledge acquisition may be a conscious process whose result is the creation of new symbolic knowledge structures and mental models, skill refinement occurs at a subconscious level by virtue of repeated practice. Most human learning appears to be a mixture of both activities, with intellectual endeavors favoring the former, and motor coordination tasks favoring the latter.

This book focuses on the knowledge acquisition aspect of learning, although some chapters, specifically those concerned with learning in problem-solving and transforming declarative instructions into effective actions, touch on aspects of both types of learning. Whereas knowledge acquisition clearly belongs in the realm of artificial intelligence research, a case could be made that skill refinement comes closer to non-symbolic processes, such as those studied in adaptive control systems. It may indeed be the case that skill acquisition is inherently non-symbolic in biological systems, but an interesting symbolic model capable of simulating gradual skill improvement through practice has been proposed recently by Newell and Rosenbloom [Newell, 1981]. Hence, perhaps both forms of learning can be captured in artificial intelligence models.

1.3 A TAXONOMY OF MACHINE LEARNING RESEARCH

This section presents a taxonomic road map to the field of machine learning with a view towards presenting useful criteria for classifying and comparing most artificial intelligence-based machine learning investigations. Later sections survey the main directions actually taken by research in machine learning over the past twenty years, and introduce each major research approach corresponding to subsequent chapters in this book.

One may classify machine learning systems along many different dimensions. We have chosen three dimensions as particularly meaningful:

- Classification on the basis of the *underlying learning strategies* used. The processes themselves are ordered by the amount of inference the learning system performs on the available information.
- Classification on the basis of the *representation of knowledge* or skill acquired by the learner.
- Classification in terms of the *application domain* of the performance system for which knowledge is acquired.

Each point in the space defined by the above dimensions corresponds to a particular learning strategy, employing a particular knowledge representation, applied to a particular domain. Since existing learning systems employ multiple representations and processes, and many have been applied to more than one domain, such learning systems are characterized by several points in the space.

The subsections below describe explored values along each of these dimensions. Future research may well reveal new values and dimensions. Indeed, the larger space of all possible learning systems is still only sparsely explored and partially understood. Existing learning systems correspond to only a small portion of the space because they represent only a small number of possible combinations of the values.

1.3.1 Classification Based on the Underlying Learning Strategy

Since we distinguish learning strategies by the amount of inference the learner performs on the information provided, we first consider the two extremes: performing no inference, and performing a substantial amount of inference. If a computer system is programmed directly, its knowledge increases, but it performs no inference whatsoever; all cognitive effort is on the part of the programmer. Conversely, if a system independently discovers new theories or invents new concepts, it must perform a very substantial amount of inference; it is deriving organized knowledge from experiments and observations. An intermediate point in the spectrum would be a student determining how to solve a mathematics problem by analogy to worked-out examples in the textbook—a process that requires inference, but much less than discovering a new branch of mathematics without guidance from teacher or textbook.

As the amount of inference that the learner is capable of performing in-

creases, the burden placed on the teacher or external environment decreases. It is much more difficult to teach a person by explaining each step in a complex task than by showing that person the way that similar tasks are usually handled. It is more difficult yet to program a computer to perform a complex task than to instruct a person to perform the task; as programming requires explicit specification of all requisite detail, whereas a person receiving instruction can use prior knowledge and common sense to fill in most mundane details. The taxonomy below captures this notion of trade-offs in the amount of effort required of the learner and of the teacher.

1. **Rote learning and direct implanting of new knowledge**—No inference or other transformation of the knowledge is required on the part of the learner. Variants of this knowledge acquisition method include:
    - Learning by being programmed, constructed or modified by an external entity, requiring no effort on the part of the learner (for example, the usual style of computer programming).
    - Learning by memorization of given facts and data with no inferences drawn from the incoming information (for example, as performed by primitive database systems). The term "rote learning" is used primarily in this context.
2. **Learning from instruction** (or, learning by being told)—Acquiring knowledge from a teacher or other organized source, such as a textbook, requiring that the learner transform the knowledge from the input language to an internally-usable representation, and that the new information be integrated with prior knowledge for effective use. Hence, the learner is required to perform some inference, but a large fraction of the burden remains with the teacher, who must present and organize knowledge in a way that incrementally augments the student's existing knowledge. Learning from instruction parallels most formal education methods. Therefore, the machine learning task is one of building a system that can accept instruction or advice and can store and apply this learned knowledge effectively. This form of learning is discussed in Chapters 12, 13 and 14.
3. **Learning by analogy**—Acquiring new facts or skills by transforming and augmenting existing knowledge that bears strong similarity to the desired new concept or skill into a form effectively useful in the new situation. For instance, a person who has never driven a small truck, but who drives automobiles, may well transform his existing skill (perhaps imperfectly) to the new task. Similarly, a learning-by-analogy system might be applied to convert an existing computer program into one that performs a closely-related function for which it was not originally designed. Learning by analogy requires more inference on the part of the learner than does rote learning or learning from instruction. A fact or skill analogous in relevant parameters must be retrieved from memory; then the retrieved knowledge must be transformed, applied to the new situation, and stored for future use. This form of learning is discussed in Chapters 5 and 7.

4. **Learning from examples** (a special case of inductive learning)—Given a set of examples and counterexamples of a concept, the learner induces a general concept description that describes all of the positive examples and none of the counterexamples. Learning from examples is a method that has been heavily investigated in artificial intelligence. The amount of inference performed by the learner is much greater than in learning from instruction, as no general concepts are provided by a teacher, and is somewhat greater than in learning by analogy, as no similar concepts are provided as "seeds" around which the new concept may be grown. Learning from examples can be subcategorized according to the *source* of the examples:

   - The source is a *teacher* who knows the concept and generates sequences of examples that are meant to be as helpful as possible. If the teacher also knows (or, more typically, infers) the knowledge state of the learner, the examples can be selected to optimize convergence on the desired concept (as in Winston's system [Winston, 1975]).
   - The source is the *learner itself*. The learner typically knows its own knowledge state, but clearly does not know the concept to be acquired. Therefore, the learner can generate instances (and have an external entity such as the environment or a teacher classify them as positive or negative examples) on the basis of the information it believes necessary to discriminate among contending concept descriptions. For instance, a learner trying to acquire the concept of "ferromagnetic substance", may generate as a possible candidate "all metals". Upon testing copper and other metals with a magnet, the learner will then discover that copper is a counterexample, and therefore the concept of ferromagnetic substance should not be generalized to include all metals.
   - The source is the *external environment*. In this case the example generation process is operationally random, as the learner must rely on relatively uncontrolled observations. For example, an astronomer attempting to infer precursors to supernovas must rely mainly upon unstructured data presentation. (Although the astronomer knows the concept of a supernova, he cannot know *a priori* where and when a supernova will occur, nor can he cause one to exist.)

   One can also classify learning from examples by the *type* of examples available to the learner:

   - *Only positive examples available*. Whereas positive examples provide instances of the concept to be acquired, they do not provide information for preventing overgeneralization of the inferred concept. In this kind of learning situation, overgeneralization might be avoided by considering only the minimal generalizations necessary, or by

relying upon *a priori* domain knowledge to constrain the concept to be inferred.

- *Positive and negative examples available.* In this kind of situation, positive examples force generalization whereas negative examples prevent overgeneralization (the induced concept should never be so general as to include any of the negative examples). This is the most typical form of learning from examples.

Learning from examples may be one-trial or incremental. In the former case, all examples are presented at once. In the latter case, the system must form one or more hypotheses of the concept (or range of concepts) consistent with the available data, and subsequently refine the hypotheses after considering additional examples. The incremental approach more closely parallels human learning, allows the learner to use partially learned concepts (for performance, or to guide the example generation process), and enables a teacher to focus on the basic aspects of a new concept before attempting to impart less central details. On the other hand, the one-step approach is less apt to lead one down garden paths by an injudicious choice of initial examples in formulating the kernel of the new concept. Various aspects of learning from examples are discussed in Chapters 3, 4, 5, 6, 7, 8, 15 and 16.

5. **Learning from observation and discovery** (also called unsupervised learning)—This is a very general form of inductive learning that includes discovery systems, theory-formation tasks, the creation of classification criteria to form taxonomic hierarchies, and similar tasks without benefit of an external teacher. This form of unsupervised learning requires the learner to perform more inference than any approach thus far discussed. The learner is not provided with a set of instances of a particular concept, nor is it given access to an oracle that can classify internally-generated instances as positive or negative instances of any given concept. Moreover, rather than focusing on a single concept at a time, the observations may span several concepts that need to be acquired, thus introducing a severe focus-of-attention problem. One may subclassify learning from observation according to the *degree of interaction* with an external environment. The extreme points in this dimension are:

   - *Passive observation*, where the learner classifies and taxonomizes observations of multiple aspects of the environment.
   - *Active experimentation*, where the learner perturbs the environment to observe the results of its perturbations. Experimentation may be random, dynamically focused according to general criteria of interestingness, or strongly guided by theoretical constraints. As a system acquires knowledge, and hypothesizes theories it may be driven to confirm or disconfirm its theories, and hence explore its environment applying different observation and experimentation strategies as the

need arises. Often this form of learning involves the generation of examples to test hypothesized or partially acquired concepts.

Learning from observation is discussed in Chapters 4, 9, 10 and 11.

The above classification of learning strategies helps one to compare various learning systems in terms of their underlying mechanisms, in terms of the available external source of information, and in terms of the degree to which they rely on pre-organized knowledge.

1.3.2 Classification According to the Type of Knowledge Acquired

A learning system may acquire rules of behavior, descriptions of physical objects, problem-solving heuristics, classification taxonomies over a sample space, and many other types of knowledge useful in the performance of a wide variety of tasks. The list below spans types of knowledge acquired, primarily as a function of the representation of that knowledge.

1. **Parameters in algebraic expressions**—Learning in this context consists of adjusting numerical parameters or coefficients in algebraic expressions of a fixed functional form so as to obtain desired performance. For instance, perceptrons [Rosenblatt, 1958; Minsky & Papert, 1969] adjust weighting coefficients for threshold logic elements when learning to recognize two-dimensional patterns.
2. **Decision trees**—Some systems acquire decision trees to discriminate among classes of objects. The nodes in a decision tree correspond to selected object attributes, and the edges correspond to predetermined alternative values for these attributes. Leaves of the tree correspond to sets of objects with an identical classification.
3. **Formal grammars**—In learning to recognize a particular (usually artificial) language, formal grammars are induced from sequences of expressions in the language. These grammars are typically represented as regular expressions, finite-state automata, context-free grammar rules, or transformation rules.
4. **Production rules**—A production rule is a condition-action pair $\{C \Rightarrow A\}$, where C is a set of conditions and A is a sequence of actions. If all the conditions in a production rule are satisfied, then the sequence of actions is executed. Due to their simplicity and ease of interpretation, production rules are a widely-used knowledge representation in learning systems. The four basic operations whereby production rules may be acquired and refined are:
    - *Creation:* A new rule is constructed by the system or acquired from an external entity.
    - *Generalization:* Conditions are dropped or made less restrictive, so that the rule applies in a larger number of situations.

- *Specialization:* Additional conditions are added to the condition set, or existing conditions made more restrictive, so that the rule applies to a smaller number of specific situations.
- *Composition:* Two or more rules that were applied in sequence are composed into a single larger rule, thus forming a "compiled" process and eliminating any redundant conditions or actions.

5. **Formal logic-based expressions and related formalisms**—These general-purpose representations have been used to formulate descriptions of individual objects (input to a learning system) and to formulate resultant concept descriptions (output from a learning system). They take the form of formal logic expressions whose components are propositions, arbitrary predicates, finite-valued variables, statements restricting ranges of variables (such as "a number between 1 and 9"), or embedded logical expressions.
6. **Graphs and Networks**—In many domains graphs and networks provide a more convenient and efficient representation than logical expressions, although the expressive power of network representations is comparable to that of formal logic expressions. Some learning techniques exploit graph-matching and graph-transformation schemes to compare and index knowledge efficiently.
7. **Frames and schemas**—These provide larger units of representation than single logical expressions or production rules. Frames and schemas can be viewed as collections of labeled entities ("slots"), each slot playing a certain prescribed role in the representation. They have proven quite useful in many artificial intelligence applications. For instance, a system that acquires generalized plans must be able to represent and manipulate such plans as units, although their internal structure may be arbitrarily complex. Moreover, in experiential learning, past successes, untested alternatives, causes of failure, and other information must be recorded and compared in inducing and refining various rules of behavior (or entire plans). Schema representations provide an appropriate formalism.
8. **Computer programs and other procedural encodings**—The objective of several learning systems is to acquire an ability to carry out a specific process efficiently, rather than to reason about the internal structure of the process. Most automatic programming systems fall in this general category. In addition to computer programs, procedural encodings include human motor skills (such as knowing how to ride a bicycle), instruction sequences to robot manipulators, and other "compiled" human or machine skills. Unlike logical descriptions, networks or frames, the detailed internal structure of the resultant procedural encodings need not be comprehensible to humans, or to automated reasoning systems. Only the external behavior of acquired procedural skills become directly available to the reasoning system.
9. **Taxonomies**—Learning from observation may result in global structuring

of domain objects into a hierarchy or taxonomy. Clustering object descriptions into newly-proposed categories, and forming hierarchical classifications require the system to formulate relevant criteria for classification.

10. **Multiple representations**—Some knowledge acquisition systems use several representation schemes for the newly-acquired knowledge. Most notably, some discovery and theory-formation systems that acquire concepts, operations on those concepts, and heuristic rules for a new domain must select appropriate combinations of representation schemes applicable to the different forms of knowledge acquired.

1.3.3 Classification by Domain of Application

A useful dimension for classifying learning systems is their area of application. The list below specifies application areas to which various existing learning systems have been applied. Application areas are presented in alphabetical order, not reflecting the relative effort or significance of the resultant machine learning system.

1. Agriculture
2. Chemistry
3. Cognitive Modeling (simulating human learning processes)
4. Computer Programming
5. Education
6. Expert Systems (high-performance, domain-specific AI programs)
7. Game Playing (chess, checkers, poker, and so on)
8. General Methods (no specific domain)
9. Image Recognition
10. Mathematics
11. Medical Diagnosis
12. Music
13. Natural Language Processing
14. Physical Object Characterizations
15. Physics
16. Planning and Problem-solving
17. Robotics
18. Sequence Prediction
19. Speech Recognition

The Bibliography provides an index to the literature organized around several criteria including some of the more commonly explored application areas. Now that we have a basis for classifying and comparing learning systems, we turn to a brief historical outline of machine learning.

1.4 AN HISTORICAL SKETCH OF MACHINE LEARNING

Over the years, research in machine learning has been pursued with varying degrees of intensity, using different approaches and placing emphasis on different aspects and goals. Within the relatively short history of this discipline, one may distinguish three major periods, each centered around a different paradigm:

- neural modeling and decision-theoretic techniques
- symbolic concept-oriented learning
- knowledge-intensive learning systems exploring various learning tasks

The distinguishing feature of the first paradigm was the interest in building general purpose learning systems that start with little or no initial structure or task-oriented knowledge. The major thrust of research based on this *tabula rasa* approach involved constructing a variety of neural model-based machines, with random or partially random initial structure. These systems were generally referred to as *neural nets* or *self-organizing systems*. Learning in such systems consisted of incremental changes in the probabilities that neuron-like elements (typically threshold logic units) would transmit a signal.

Due to the primitive nature of computer technology at that time, most of the research under this paradigm was either theoretical or involved the construction of special purpose experimental hardware systems, such as perceptrons [Rosenblatt, 1958], pandemonium [Selfridge, 1959] and adelaine [Widrow, 1962]. The groundwork for this paradigm was laid in the forties by Rashevsky and his followers working in the area of mathematical biophysics [Rashevsky, 1948], and by McCulloch and Pitts [1943], who discovered the applicability of symbolic logic to modeling nervous system activities. Among the large number of research efforts in this area, one may mention works such as [Ashby, 1960; Rosenblatt, 1958, 1962; Minsky & Papert, 1969; Block, 1961; Yovits, 1962, Widrow, 1962; Culberson, 1963; Kazmierczak, 1963]. Related research involved the simulation of evolutionary processes, that through random mutation and "natural" selection might create a system capable of some intelligent behavior (for example, [Friedberg, 1958, 1959; Holland, 1980]).

Experience in the above areas spawned the new discipline of pattern recognition and led to the development of a decision-theoretic approach to machine learning. In this approach, learning is equated with the acquisition of linear, polynomial, or related forms of discriminant functions from a given set of training examples (for example, [Nilsson, 1965; Koford, 1966; Uhr, 1966; Highleyman, 1967]). One of the best known successful learning systems utilizing such techniques (as well as some original new ideas involving non-linear transformations) was Samuel's checkers program [Samuel, 1959, 1963]. This program was able to acquire through learning a master level of performance. Somewhat different, but closely related, techniques utilized methods of statistical decision theory for learning pattern recognition rules (for example, [Sebestyen,

1962; Fu, 1968; Watanabe, 1960; Arkadev, 1971; Fukananga, 1972; Duda & Hart, 1973; Kanal, 1974]).

In parallel to research on neural modeling and decision-theoretic techniques, researchers in control theory developed adaptive control systems able to adjust automatically their parameters in order to maintain stable performance in the presence of various disturbances (for example, [Truxal, 1955; Davies, 1970; Mendel, 1970; Tsypkin, 1968, 1971, 1973; Fu, 1971, 1974]).

Practical results sought by the neural modeling and decision theoretic approaches met with limited success. High expectations articulated in various early works were not realized, and research under this paradigm began to decline. Theoretical studies have revealed strong limitations of the "knowledge-free" perceptron-type learning systems [Minsky & Papert, 1969].

A second major paradigm started to emerge in the early sixties stemming from the work of psychologists and early AI researchers on models of human learning [Hunt *et al.*, 1963, 1966]. The paradigm utilized logic or graph structure representations rather than numerical or statistical methods. Systems learned symbolic descriptions representing higher level knowledge and made strong structural assumptions about the concepts to be acquired.

Examples of work in this paradigm include research on human concept acquisition (for example, [Hunt & Hovland, 1963; Feigenbaum, 1963; Hunt *et al.*, 1966; Hilgard, 1966; Simon & Lea, 1974]), and various applied pattern recognition systems ([Bongard, 1970; Uhr, 1966; Karpinski & Michalski, 1966]).

Some researchers constructed task-oriented specialized systems that would acquire knowledge in the context of a practical problem. For instance, the META-DENDRAL program [Buchanan, 1978] generates rules explaining mass spectrometry data for use in the DENDRAL system [Buchanan *et al.*, 1971].

An influential development in this paradigm was Winston's structural learning system [Winston, 1975]. In parallel with Winston's work, different approaches to learning structural concepts from examples emerged, including a family of logic-based inductive learning programs (AQVAL) [Michalski, 1972, 1973, 1978], and related work by Hayes-Roth [1974], Hayes-Roth & McDermott [1978], Vere [1975], and Mitchell [1978]. More details on this paradigm are included in Chapters 3, 4 and 6. (See also [Michie, 1982].)

The third paradigm represents the most recent period of research starting in the mid-seventies. Researchers have broadened their interest beyond learning isolated concepts from examples, and have begun investigating a wide spectrum of learning methods, most based upon knowledge-rich systems. Specifically, this paradigm can be characterized by several new trends, including:

1. **Knowledge-Intensive Approaches:** Researchers are strongly emphasizing the use of task-oriented knowledge and the constraints it provides in guiding the learning process. One lesson from the failures of earlier *tabula rasa* and knowledge-poor learning systems is that to acquire new knowledge a system must already possess a great deal of initial knowledge.

2. **Exploration of alternative methods of learning:** In addition to the earlier research emphasis on learning from examples, researchers are now investigating a wider variety of learning methods such as learning from instruction (Chapters 12, 13, and 14 in this book), learning by analogy ([Winston, 1979], and Chapter 5 of this book), and discovery of concepts and classifications ([Lenat, 1976] and Chapters 4, 10, and 11 of this book).
3. **Incorporating abilities to generate and select learning tasks:** In contrast to previous efforts, a number of current systems incorporate heuristics to control their focus of attention by generating learning tasks, proposing experiments to gather training data, and choosing concepts to acquire ([Lenat, 1976] and Chapter 6 of this book).

The research presented in this book is concerned primarily with the last, knowledge-intensive paradigm of learning.

1.5 A BRIEF READER'S GUIDE

The chapters in this book are organized according to the major thrust of each investigation, whether that thrust is the development of a general method, the application of various learning techniques to a particular domain, or the theoretical analysis of existing methods. The progression of chapters roughly corresponds to the sequence:

- Basic principles
- General-purpose systems
- Task-oriented applications

Although there is much overlap among the objectives of different chapters, the specific content differs substantially. For instance, the four papers listed under the general category "Learning in problem-solving and planning," share a common top-level objective, but differ substantially in terms of the learning methods employed, the type of knowledge acquired, and the range of applicability of the described systems.

The reader not familiar with the field of machine learning is encouraged to read the first few chapters, omitting technical detail, in order to acquire a general understanding. Later, these chapters and any others that are of special interest may be studied in more detail with an appropriate perspective on the field as a whole. Readers are encouraged to use our chapter descriptions below, as well as the abstracts in the individual chapters, to focus on areas of interest. The topics of the individual chapters range from cognitive modeling and discussion of underlying principles to applications in general problem-solving, chemistry, mathematics, music, education and game playing.

At the Carnegie-Mellon Machine Learning Workshop in July, 1980, Herbert Simon was asked to deliver the keynote address, where he chose to play the

role of devil's advocate and ask the question "Why Should Machines Learn?" His analysis concluded that, with the exception of cognitive modeling, some rethinking of long-term objectives was in order. After dispelling some common myths, Simon concluded with a clarified and more appropriate set of reasons why one ought to pursue machine learning research. Chapter 2 is based almost entirely on that rather controversial keynote address.

In Chapter 3, Dietterich and Michalski analyze some well-known work in concept acquisition from a unified perspective. After developing some requisite formalism, they examine the range of possible concept descriptions that may be acquired via a set of basic generalization and discrimination operators applied to logic-based representations of instances and concepts. Then, they describe the work of Winston, Hayes-Roth, Vere, and Michalski's earlier work as particular combinations of learning operators applied to different restrictions on the representation language. Chapter 3, therefore, provides a general framework for comparison of different concept-acquisition systems.

In Chapter 4, Michalski describes a general theory and methodology for inductive learning of structural descriptions from examples. The theory unifies and clarifies various types of inductive learning, and demonstrates that such learning can be viewed as a process of applying *generalization inference rules* (and conventional deductive inference rules) to initial and intermediate descriptions. This process is guided by problem-oriented background knowledge provided to the learning system. Various generalization rules are presented and discussed. The methodology developed is illustrated by a problem from the area of conceptual data analysis.

In Chapter 5, Carbonell examines the issue of learning from experience, a common phenomenon among humans, but heretofore a nemesis to machines that could not transfer planning knowledge to new but similar situations, or otherwise analyze their past behavior. A general planning and problem-solving paradigm is proposed based on a computationally-effective model of analogical reasoning. In essence, the planner exploits prior experience in solving new problems that bear strong similarity to past situations by transforming solutions of past problems into potential plans that solve new, externally or internally generated problems. The analogical paradigm interfaces with a learning-from-examples method, enabling the learner to formulate generalized plans for recurring situations, as well as to accumulate and classify more specific experiences for less common situations.

In Chapter 6, Mitchell, Utgoff and Banerji investigate the issue of acquiring and refining problem-solving heuristics by examining solutions to symbolic integration problems. Like Carbonell's approach, learning is based on past problem-solving experience, but Mitchell *et al.* focus on acquiring heuristics for applying known strategies, rather than generalizing recurring behaviors into reusable plans. Their approach also generates problems internally for the purpose of testing and refining existing heuristics, and uses the version-space approach to keep track of viable generalizations of current heuristics. Unlike Carbonell's

analogical approach to problem-solving, Mitchell *et al.* rely on heuristic search guided by the constantly updated domain heuristics to solve new problems. After describing the LEX program for learning heuristics, they consider ways in which the system's learning abilities could be improved by giving it new knowledge about heuristic search, the problem domain, and the goals of the learner.

In Chapter 7, Anderson examines human problem-solving in the context of providing justifications to geometric proofs. He relies entirely upon a production system framework to encode domain knowledge, learning heuristics, and problem-solving strategies. Anderson reviews the basic mechanisms for production-rule knowledge acquisition and demonstrates how they apply to a progression of tasks in Geometry. The major significance of this chapter is the explanation and illustration of learning methods in the context of a performance system implemented as a set of production rules.

In Chapter 8, Hayes-Roth investigates the issue of improving flawed or incomplete theories that guide plan formation in a given domain. His primary thrust is on refining and restructuring theories based upon the way in which observed consequences of one's behavior differ from theoretical predictions. In short, Hayes-Roth views empirical disconfirmation not as a mechanism for rejecting existing theories, but rather as input to various methods of modifying theoretical concepts to accord with past and present observations. He presents five heuristic methods and applies them to problem-solving in playing the card game hearts.

In Chapter 9, Lenat focuses on methods for learning from observation and discovery. He analyzes three domains in which heuristics play a dominant role in guiding search through the space of possible concepts or processes one may acquire. First, Lenat examines his AM system, where heuristic rules that measure intrinsic "interestingness" help the system rediscover essential concepts in number theory, such as the notion of a prime number. Then, the EURISKO system is discussed, which acquires and modifies learning heuristics, as well as formulating task-specific heuristics and concept representations. Finally, Lenat discusses the conjecture that evolution is a heuristically-driven learning engine in constant operation.

In Chapter 10, Langley, Simon and Bradshaw discuss their BACON system and its application to rediscovering some basic laws of Chemistry. BACON applies the principles of scientific inquiry first elucidated by Sir Francis Bacon to find the simplest numerical relations that hold invariant across sets of measurements. In this manner, it postulates meaningful combinations of independent measurements and intrinsic properties of objects (such as specific heat), and searches for the simplest relationship among measured and derived quantities that summarizes all observations. Although not able to design its own experiments, given the unanalyzed results of appropriate chemical experiments, BACON has rediscovered such laws as Gay-Lussac's law and Proust's law of definite proportions.

In Chapter 11, Michalski and Stepp investigate the problem of automated construction of taxonomies of observed events in a manner that is meaningful to a human. That is, given sets of object or process descriptions, plus an *a priori* set of descriptive concepts, they develop a method of grouping observations into meaningful classes that represent selected concepts. They present an algorithm that implements this "conceptual clustering" operation and demonstrate its utility for the tasks of formulating descriptions of plant diseases from observed symptoms and taxonomizing Spanish songs in a manner meaningful to musicologists. In contrast with statistical clustering techniques, the conceptual clustering algorithm produces characteristic descriptions of the concepts defined by each cluster. Both the Michalski and Stepp approach and the Langley *et al.* approach exemplify learning from passive observations, whereas Lenat's approach stresses the role of active experimentation.

In Chapter 12, Mostow discusses the process of learning by taking advice. Declaratively stated advice must be transformed into operational procedures effective in a given task domain. The transformation process can be quite complex, as implicit domain knowledge must be accessed, the advice must be restated in terms consistent with the existing procedural knowledge base, and plausible reasoning heuristics must be consulted in deciding how to make best use of the incoming advice. Mostow focuses on the general issue of providing advice to a heuristic search mechanism, as applied to playing the game of hearts and composing a *cantus firmus*.

In Chapter 13, Haas and Hendrix investigate the issue of automatically extending a natural language interface by acquiring domain semantics, dictionary entries and syntactic patterns from the user. The most significant aspect of their KLAUS system is that the user need not be a computational linguist, but rather is guided by the system into providing exemplary information that is later transformed into effective grammar and dictionary representations. This form of learning by being told, where the student (that is, the KLAUS system) is in control and the teacher provides information only when asked, constitutes an interesting variation on more traditional versions of the learning-from-instruction paradigm.

In Chapter 14, Rychener provides a retrospective analysis of the instructable production system project, in which many different instructional techniques for learning by being told were tried, different organizations of the knowledge were considered, and different problem-solving strategies were investigated. Although many combinations of representational schemes and instructional methods proved infeasible, other approaches proved much more promising. Hence the field of machine learning can learn from its own experience—false starts as well as successful approaches. Rychener concludes his chapter with an analysis of the organizational and instructional principles that a production-system based instructional learner should adhere to in order to maximize chances for successful knowledge acquisition.

In Chapter 15, Quinlan presents a method for generating efficient decision trees for classifying given exemplars, and applies his method to the analysis of

king-and-rook versus king-and-knight chess endgames. Chess authorities had previously believed that all but a few special positions of this type were inherently drawn (with best play for both sides). Due to the size of the search space, a systematic analysis was not performed until Quinlan applied his efficient method of learning classifications, whereupon it became clear that a very large fraction of king-and-rook versus king-and-knight positions were forced wins for the side with the rook. Therefore, the Quinlan paper illustrates not only an efficient classification method, but demonstrates the utility of at least one application of machine learning.

In Chapter 16, Sleeman investigates the application of machine learning to infer models of students learning algebra. Student modeling is becoming a recognized necessity in intelligent computer-assisted instruction (ICAI). The difficult task of formulating viable student models requires that the system infer a student's knowledge from his performance (plus general knowledge of the instructional material). A general model must be inferred that can generate all observed student behavior, as well as account for the lack of any expected but unobserved behavior. The search space of possible student models is large, and the number of trials one may require of each student is proportionately small. Therefore the problem becomes one of searching this space quickly and without requiring large amounts of student testing. Sleeman provides and analyzes algorithms that fit these requirements. An interesting aspect of Sleeman's work is that the teacher, in order to be effective, must learn to adapt to the student's needs, indicating that machine learning can help to make computer-assisted human education more effective.

Finally, the book concludes with a comprehensive bibliography of past and present research in machine learning, a glossary of selected terms, and a brief note about each author. The bibliography is indexed according to several criteria (methods, applications, and so on) in order to provide guidance to the reader who desires additional background in the field.

REFERENCES

Arkadev, A. G. and Braverman, E. M., *Learning in Pattern Classification Machines*, Nauka, Moscow, 1971.

Ashby, W. Ross, *Design for a Brain, The Origin of Adaptive Behavior*, John Wiley and Sons, Inc., 1960.

Block, H. D., "The Perceptron: A Model of Brain Functioning, I," *Rev. Math. Physics*, Vol. 34, No. 1, pp. 123-135, 1961.

Bongard, N., *Pattern Recognition*, Spartan Books, New York, 1970, (Translation from Russian original, published in 1967).

Buchanan, B. G. and Mitchell, T. M., "Model-Directed Learning of Production Rules," *Pattern-Directed Inference Systems*, Waterman, D. A. and Hayes-Roth, F. (Eds.), Academic Press, New York, 1978.

Buchanan, B. G., Feigenbaum, E. A. and Lederberg, J., "A heuristic programming study of theory formation in sciences," *Proceedings of the Second International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence, London, pp. 40-48, 1971.

Culberson, J. T., *The Minds of Robots*, University of Illinois Press, Urbana, Illinois, 1963.

Davies, W. D. T., *System Identification for Self-Adaptive Control*, Wiley-Interscience, Wiley and Sons, Ltd., 1970.

Duda, R. O. and Hart, P. E., *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

Feigenbaum, E. A., "The Simulation of Verbal Learning Behavior," *Computers and Thought*, Feigenbaum, E. A. and Feldman, J. (Eds.), McGraw-Hill, New York, pp. 297-309, 1963, (originally in Proceedings Western Joint Computer Conference, 1961).

Friedberg, R. M., "A Learning Machine: Part 1," *IBM Journal*, Vol. 2, pp. 2-13, 1958.

Friedberg, R., Dunham, B. and North, T., "A Learning Machine: Part 2," *IBM Journal of Research and Development*, Vol. 3, No. 3, pp. 282-287, 1959.

Fu, K. S., *Sequential Methods in Pattern Recognition and Machine Learning*, Academic Press, New York, 1968.

Fu, K. S., *Pattern Recognition and Machine Learning*, Plenum Press, New York, 1971.

Fu, K. S. and Tou, J. T., *Learning Systems and Intelligent Robots*, Plenum Press, 1974.

Fukanaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, 1972.

Hayes-Roth, F., "Schematic Classification Problems and their Solution," *Pattern Recognition*, Vol. 6, pp. 105-113, 1974.

Hayes-Roth, F. and McDermott, J., "An interference matching technique for inducing abstractions," *Communications of the ACM*, Vol. 21, No. 5, pp. 401-410, 1978.

Highleyman, W. H., "Linear Decision Functions, with Applications to Pattern Recognition," *Proceedings of IRE*, No. 50, pp. 1501-1504, 1967.

Hilgard, E. R. and Bower, G. H., *Theories of Learning - Third Edition*, Appleton-Century-Grofts, New York, 1966.

Holland, J. H., "Adaptive Algorithms for Discovering and Using General Patterns in Growing Knowledge Bases," *Policy Analysis and Information Systems*, Vol. 4, No. 3, September 1980.

Hunt, E. B. and Hovland, C. I., "Programming a Model of Human Concept Formation," *Computers and Thought*, Feigenbaum, E. A. and Feldman, J. (Eds.), McGraw-Hill, New York, pp. 310-325, 1963.

Hunt, E. B., Marin, J. and Stone, P. T., *Experiments in Induction*, Academic Press, New York, 1966.

Kanal, L., "Patterns in Pattern Recognition: 1968-1974," *IEEE Transactions on Information Theory*, Vol. IT-20, No. 6, pp. 697-722, 1974.

Karpinski, J. and Michalski, R. S., "A System that Learns to Recognize Hand-written Alphanumeric Characters", Technical Report 35, Proce Institute Automatyki, Polish Academy of Sciences, 1966.

Kazmierczak, H. and Steinbuch, K., "Adaptive Systems in Pattern Recognition," *IEEE Transactions of Electronic Computers*, Vol. EC-12, No. 5, pp. 822-835, 1963.

Koford, T. S. and Groner, G. F., "The Use of an Adaptive Threshold Element to Design a Linear Optimal Pattern Classifier," *IEEE Transactions-Information Theory*, Vol. 1T-12, pp. 42-50, 1966.

Lenat, D. B., *AM: an artificial intelligence approach to discovery in mathematics as heuristic search*, Ph.D. dissertation, Stanford University, Stanford, California, 1976.

McCulloch, W. S. and Pitts, W., "A Logical Calculus of Ideas Imminent in Nervous Activity," *Bull. Math. Biophysics*, Vol. 5, pp. 115-133, 1943.

Mendel, T. and Fu, K. S., *Adaptive Learning and Pattern Recognition: Theory and Applications*, Spartan Books, New York, 1970.

Michalski, R. S., "A Variable-Valued Logic System as Applied to Picture Description and Recognition," *Graphic Languages*, Nake, F. and Rosenfeld, A. (Ed.), North-Holland, 1972.

Michalski, R. S. and Larson, J. B., "Selection of Most Representative Training Examples and Incremental Generation of VL1 Hypotheses: The Underlying Methodology and Description of Programs ESEL and AQ11", Report 867, University of Illinois, 1978.

Michalski, R. S., "AQVAL/1 - Computer implementation of a variable valued logic system VL1 and examples of its application to pattern recognition," *Proceedings of the First International Joint Conference on Pattern Recognition*, Washington, D. C., pp. 3-17, 1973b.

Michie, "The State of the Art in Machine Learning," *Introductory Readings in Expert Systems*, D. Michie (Ed.), Gordon and Breach, UK, 1982.

Minsky, M. and Papert, S., *Perceptrons*, MIT Press, Cambridge, Mass., 1969.

Mitchell, T. M., *Version Spaces: An approach to concept learning*, Ph.D. dissertation, Stanford University, December 1978, (also Stanford CS report STAN-CS-78-711, HPP-79-2).

Newell, A. and Rosenbloom, P., "Mechanisms of Skill Acquisition and the Law of Practice," *Cognitive Skills and Their Acquisition*, Anderson, J. R. (Ed.), Erlbaum Associates, Hillsdale, New Jersey, 1981.

Nilsson, N. J., *Learning Machines*, McGraw-Hill, New York, 1965.

Rashevsky, N., *Mathematical Biophysics*, University of Chicago Press, Chicago, IL, 1948.

Rosenblatt, F., "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review*, Vol. 65, pp. 386-407, 1958.

Rosenblatt, F., *Principles of Neurodynamics and the Theory of Brain Mechanisms*, Spartan Books, Washington, D. C., 1962.

Samuel, A. L., "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, No. 3, pp. 211-229, 1959.

Samuel, A. L., "Some Studies in Machine Learning using the Game of Checkers," *Computers and Thought*, Feigenbaum, E. A. and Feldman, J. (Eds.), McGraw-Hill, New York, pp. 71-105, 1963.

Sebestyen, G. S., *Decision-Making Processes in Pattern Recognition*, Macmillan, New York, 1962.

Selfridge, O. G., "Pandemonium: A Paradigm for Learning," *Proceedings of the Symposium on Mechanization of Thought Processes*, Blake, D. and Uttley, A. (Eds.), HMSO, London, pp. 511-529, 1959.

Simon, H. A. and Lea, G., "Problem Solving and Rule Induction: A Unified View," *Knowledge and Cognition*, Gregg, L. W. (Ed.), Lawrence Erlbaum Associates, Potomac, Maryland, pp. 105-127, 1974.

Truxal, T. G., *Automatic Feedback Control System Synthesis*, McGraw-Hill, New York, 1955, (New York).

Tsypkin, Y. Z., "Self Learning - What is it?," *IEEE Transactions on Automatic Control*, Vol. AC-18, No. 2, pp. 109-117, 1968.

Tsypkin, Ya Z., *Adaptation and Learning in Automatic Systems*, Academic Press, New York, 1971.

Tsypkin, Y. Z., *Foundations of the Theory of Learning Systems*, Academic Press, New York, 1973, (Translated by Z. L. Nikolic).

Uhr, L., *Pattern Recognition*, John Wiley and Sons, New York, 1966.

Vere, S. A., "Induction of concepts in the predicate calculus," *Proceedings of the Fourth International Joint Conference on Artificial Intelligence*, IJCAI, Tbilisi, USSR, pp. 281-287, 1975.

Watanabe, S., "Information-Theoretic Aspects of Inductive and Deductive Inference," *IBM Journal of Research and Development*, Vol. 4, No. 2, pp. 208-231, 1960.

Widrow, B., *Generalization and Information Storage in Networks of Adelaine 'Neurons,'*, Spartan Books, Washington, D. C., pp. 435-461, 1962, (Yovitz, M. C.; Jacobi, G. T.; Goldstein, G. D., editors).

Winston, P. H., "Learning structural descriptions from examples," *The Psychology of Computer Vision*, Winston, P. H. (Ed.), McGraw Hill, New York, ch. 5, 1975, (Original version published as a Ph.D. dissertaition, at MIT AI Lab, September, 1970).

Winston, P. H., "Learning and Reasoning by Analogy," *CACM*, Vol. 23, No. 12, pp. 689-703, 1979.

Yovits, M. C., Jacobi, G. T. and Goldstein, G. D., *Self-Organizing Systems*, Spartan Books, Washington, D. C., 1962.