BIBLIOGRAPHICAL AND HISTORICAL NOTES

The general architecture for learning systems portrayed in Figure 18.1 is classic in the machine learning literature (Buchanan *et al.*, 1978). The architecture itself, as embodied in programs, goes back at least as far as Arthur Samuel's (1959; 1967) program for playing checkers, which improved its performance through learning. It is described further in Chapter 20.

Claude Shannon was the inventor of information theory (Shannon and Weaver, 1949), which revolutionized the study of communication as well as contributing to the success of decision-tree learning. He also contributed one of the earliest examples of machine learning, a mechanical mouse named Theseus that learned how to travel efficiently through a maze by trial and error. EPAM, the "Elementary Perceiver And Memorizer" (Feigenbaum, 1961), was one of the earliest systems to use decision trees (or **discrimination nets**). EPAM was intended as a cognitive-simulation model of human concept learning. CLS (Hunt *et al.*, 1966) used a heuristic lookahead method to construct decision trees. ID3 (Quinlan, 1979) added the crucial idea of using information content to provide the heuristic function. Quinlan (1986) experimentally investigates the effect of noise on learning and describes a modification of ID3 designed to handle noise. A copy of C4.5, an industrial-strength version of ID3, can be found in Quinlan (1993a).

William of Ockham (c. 1285–1349), the most influential philosopher of his century and a major contributer to medieval epistemology, logic, and metaphysics, is credited with a statement called "Ockham's Razor"—in Latin, *Entia non sunt multiplicanda praeter necessitatem*, and in English, "Entities are not to be multiplied without necessity." Unfortunately, this laudable piece of advice is nowhere to be found in his writings in precisely these words.

The current-best-hypothesis approach has been a mainstay of the philosophy of science, as a model of scientific theory formation, for centuries. In AI, the approach is most closely associated with the work of Patrick Winston, whose Ph.D. thesis (Winston, 1970) addressed the problem of learning descriptions of complex objects. Tom Mitchell (1977; 1982) invented version spaces and the learning algorithm that uses them. They were initially used in the Meta-DENDRAL expert system for chemistry (Buchanan and Mitchell, 1978), and later in Mitchell's (1983) LEX system, which learns to solve calculus problems. The importance of bias in inductive learning was emphasized by Mitchell (1980), who showed that an unbiased algorithm is incapable of learning because its hypothesis space always contains equal numbers of consistent hypotheses that predict Q and $\neg Q$ for any example and for any goal predicate Q.

Historically, the earliest research on the theoretical analysis of learning focused on the problem of "identification in the limit" (Gold, 1967), which involves showing that a learning algorithm will eventually converge on the correct hypothesis once it has seen enough examples. This approach was motivated in part by models of scientific discovery from the philosophy of science (Popper, 1962), but has been applied mainly to the problem of learning grammars from example sentences. Osherson, Stob, and Weinstein (1986) provide a modern and rigorous treatment of the field.

KOLMOGOROV COMPLEXITY

Whereas the identification-in-the-limit approach concentrates on eventual convergence, the study of **Kolmogorov complexity** or **algorithmic complexity**, developed independently by Solomonoff (1964) and Kolmogorov (1965), attempts to provide a formal definition for the notion of simplicity used in Ockham's razor. To escape the problem that simplicity depends on the way