1 Empirical Research

Physicists ask what kind of place this universe is and seek to characterize its behavior systematically. Biologists ask what it means for a physical system to be living. We in AI wonder what kind of information-processing system can ask such questions.
—Avron B. Barr and Edward A. Feigenbaum, *The Handbook of Artificial Intelligence*, Vol. 1, p. 11

Empirical methods enhance our observations and help us see more of the structure of the world. We are fundamentally empirical creatures, always asking, What is going on? Is it real or merely apparent? What causes it? Is there a better explanation? Great developments were born of microscopes, telescopes, stethoscopes, and other observational tools. No less important are "datascopes"—visualization methods—and techniques for drawing sound conclusions from data. Empirical methods cluster loosely into *exploratory* techniques for visualization, summarization, exploration, and modeling; and *confirmatory* procedures for testing hypotheses and predictions. In short,

empirical = exploratory + experimental.

Because empirical studies usually observe nondeterministic phenomena, which exhibit variability, both exploratory and experimental methods are based in statistics. Exploratory statistical methods are called, collectively, *exploratory data analysis*; whereas methods for confirmatory experiments go by the name *statistical hypothesis testing*. You can find any number of textbooks on the latter, very few devoted to the former. You might get the impression from these books and from scientific journals that *all* empirical work involves controlled experiments and hypothesis testing. But experiments don't spring like Athena fully formed from the brow of Zeus: They are painstakingly constructed by mortals who usually get it wrong first time, who require lengthy periods of exploration before formulating a precise experimental question, yet find themselves still amazed at Nature's ability to confound.

Even if experiments were easy, you should take a more encompassing view of empirical work for one reason if no other: experiments and hypothesis testing answer yes-or-no questions. You can reduce all your research interests to a series of yes-or-no questions, but you'll find it slow work. A more efficient approach is to flip back and forth between exploratory and experimental work, engaging in the latter only when the former produces a question you really want to answer.

Empirical methods and statistics are worth learning because a handful of each goes a long way. Psychologists, for example, know perhaps ten major statistical methods and a similar number of techniques for visualizing and exploring data; and they know perhaps two dozen important experiment designs.[1] With these methods they address thousands of research questions. We are AI researchers, not psychologists, but we too can expect a few appropriate methods to go a long way.

The American Heritage Dictionary[2] defines empirical and empiricism as follows:

Empirical (1) Relying upon or derived from observation or experiment: empirical methods. (2) Relying solely on practical experience and without regard for system or theory.

Empiricism (1) The view that experience, esp. of the senses, is the only source of knowledge. (2) The employment of empirical methods, as in science.

In this book, you should read empirical in its first sense and empiricism in its second sense. The other interpretations are too exclusive (Cohen, 1991). Down with empiricism (in its first sense); down with the equally exclusive view that theory is the only source of knowledge. Up with empirical studies informed by theories and theories informed by observations.

1.1 AI Programs as Objects of Empirical Studies

Our subject is empirical methods for studying AI programs, methods that involve running programs and recording their behaviors. Unlike other scientists, who study chemical reactions, processes in cells, bridges under stress, animals in mazes, and so on, we study computer programs that perform tasks in environments. It shouldn't be

[1] I get these numbers from textbooks on statistics for psychologists and from books on experiment design such as Cook and Campbell 1979 and articles such as Bower and Clapper 1990 and Newell 1973.

[2] *American Heritage Dictionary of the English Language*, American Heritage Publishing Co. Inc. and Houghton Mifflin Company, 1971.

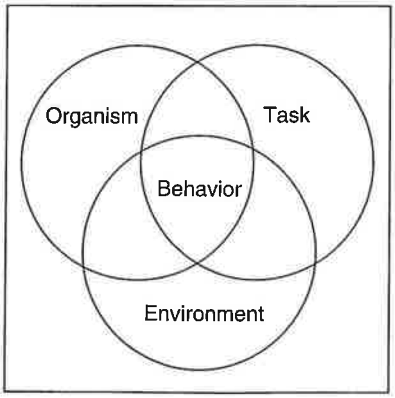

Figure 1.1 How the structure, task, and environment of an organism influence its behavior.

difficult: Compared with biological systems, AI systems are simple; compared with human cognition, their tasks are rudimentary; and compared with everyday physical environments, those in which our programs operate are extremely reduced. Yet programs are in many ways like chemical, biological, mechanical, and psychological processes. For starters, we don't know how they work. We generally cannot say how long they will take to run, when they will fail, how much knowledge is required to attain a particular error rate, how many nodes of a search tree must be examined, and so on. It is but a comforting fiction that building a program confers much understanding of its behavior. Predictive theories of program behavior are thin on the ground and no better than predictive theories of chemical, biological, mechanical, and psychological processes: at best incomplete and approximate, at worst vague or wrong.

Studying AI systems is not very different from studying moderately intelligent animals such as rats. One obliges the agent (rat or program) to perform a task according to an experimental protocol, observing and analyzing the macro- and micro-structure of its behavior. Afterward, if the subject is a rat, its head is opened up or chopped off; and if it is a program, its innards are fiddled with. Six components are common to these scenarios: agent, task, environment, protocol, data collection, and analysis. The first three are the domain of theories of behavior, the last three are in the realm of empirical methods. Behavior is what we observe and measure when an agent attempts a task in an environment. As figure 1.1 suggests, an agent's behavior is not due entirely to its structure, nor its environment, nor its task, but rather to the interaction of these influences.

Whether your subject is a rat or a computer program, the task of science is the same, to provide theories to answer three *basic research questions*:

- How will a change in the agent's structure affect its behavior given a task and an environment?
- How will a change in an agent's task affect its behavior in a particular environment?
- How will a change in an agent's environment affect its behavior on a particular task?

1.2 Three Basic Research Questions

The three basic research questions all have the same form, so let's pick one to examine: How will a change in the agent's structure affect its behavior given a task and an environment? Bill Clancey and Greg Cooper asked such a question of the MYCIN system some years ago (Buchanan and Shortliffe, 1984, p. 219). They asked, how sensitive is MYCIN to the accuracy of its certainty factors? What will happen if each certainty factor, very precisely represented on the scale −1000 . . . 1000, is replaced by the nearest of just seven values (−1000, −666, −333, 0, 333, 666, 1000)? When Clancey and Cooper ran the modified and unmodified versions of MYCIN and compared the answers, they found essentially no decrement in the adequacy of MYCIN's recommendations (see chapter 6 for details).

Clancey and Cooper's question was exploratory; their answer was descriptive. They asked, "What will happen if . . .?"; they answered with a description of MYCIN's performance. Question and answer belong in the lower left-hand corner of figure 1.2, the dimensions of which—understanding and generality—define a space of versions of the basic research questions and answers. Early in a research project we ask, "What will happen if . . .?" and answer, "Here's what happens. . . ." Later, we ask, "Does this model accurately predict what happens?" and "Does this model provide an accurate causal explanation of what happens?" Early in a project we ask short questions that often have lengthy descriptions of behavior as answers; later, the questions are lengthy because they refer to predictive and causal models, but the answers are short, often just "yes" or "no." This shift in balance characterizes progress from exploratory studies to experimental studies.

The progression is called "understanding" in figure 1.2 because descriptions—the low end of the dimension—require no understanding (you can describe leaves turning color in autumn without understanding the process); whereas prediction requires at least an understanding of the conditions under which behavior occurs (leaves turn color in the autumn, or when the tree is diseased, and turn more reliably after a period of intensely cold weather). In practice, the transition from description to prediction depends on identifying terms in the description that appear to have predictive power.

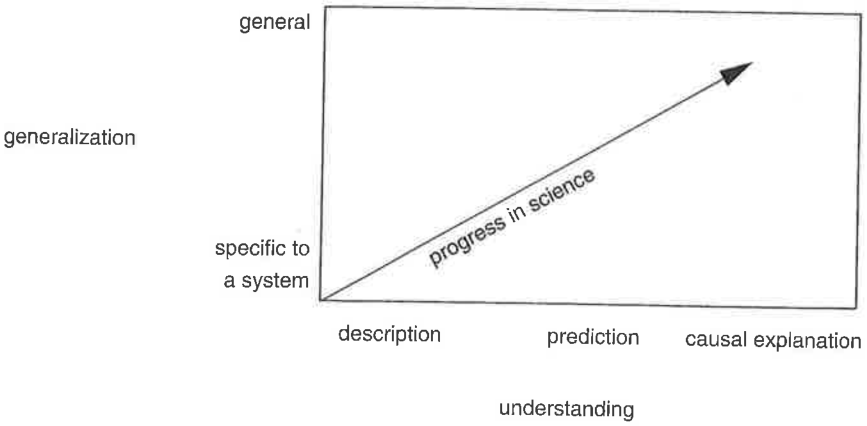

Figure 1.2 Generalization and understanding define a space of basic research questions.

This often involves a succession of descriptions: "leaves turn color in October," then, identifying a predictive feature, "leaves turn color when the weather turns cold." An important role of exploratory studies is to identify predictive features. What we aim for as scientists, however, is causal explanation: leaves turn color because chlorophyll, which masks other pigments, disappears. And not only causal explanations, but *general* causal explanations—why the leaves of aspen, maple, sumac, and oak (but not pine) change color in the autumn.

1.3 A Strategy for Answering Basic Research Questions

Clancey and Cooper did, in fact, offer a post hoc, informal causal explanation of their results. MYCIN's task was to prescribe therapy for all the organisms likely to have caused one or more infections. Inaccuracies in certainty factors affected MYCIN's judgments about the likelihood of organisms, but because most antibiotics kill many sorts of organisms, the chances are good that MYCIN will recommend an effective antibiotic even if it judges the likelihood of each organism incorrectly. Clancey and Cooper didn't generalize this explanation beyond MYCIN, but we can imagine how they might have done it. Here is a generalization in terms of features of a program's structure, task, environment, and behavior:

Suppose a program's task is to select actions to "cover" sets of situations. An instance of the task is a set S of situations. A feature of the environment is that during problem-solving, S doesn't change. A feature of the task is that it provides too little information for the program *to be certain whether a hypothesized situation h_i is in S. Another task feature is that S is relatively small. A feature of the program is that it can judge $Pr(h_i \in S)$, the likelihood that h_i is in S. Another program feature is that it can decide to cover only some h's (e.g., likely ones). The program selects actions to cover these. The actions have the interesting property that each deals with several h's. The only behavior of interest is whether the program selects "good" actions. A generalization of Clancey and Cooper's result is this: The program's behavior is robust against inaccurate estimates of $Pr(h_i \in S)$.*

The point of this long-winded exercise is to illustrate a central conjecture: General theories in artificial intelligence arise from *featural characterizations of programs*, their environments, tasks, and behaviors. Progress toward general theories depends on finding these features. Empirical methods, in particular, help us find general features by studying specific programs. The *empirical generalization strategy*, around which this book is organized, goes like this:

1. build a program that exhibits a behavior of interest while performing particular tasks in particular environments;
2. identify specific features of the program, its tasks and environments that influence the target behavior;
3. develop and test a causal model of how these features influence the target behavior;
4. once the model makes accurate predictions, generalize the features so that other programs, tasks, and environments are encompassed by the causal model;
5. test whether the general model predicts accurately the behavior of this larger set of programs, tasks, and environments.

Chapter 2 on exploratory data analysis addresses step 2. Chapters 3, 4, and 5, on experiment design and hypothesis testing address the testing part of steps 3 and 5. Chapters 6, 7, and 8, address the model formation parts of step 3. Chapter 9, on generalization, addresses step 4.

1.4 Kinds of Empirical Studies

Many researchers in artificial intelligence and computer science speak casually of experiments, as if any activity that involves building and running a program is experimental. This is confusing. I once flew across the country to attend a workshop on "experimental computer science," only to discover this phrase is shorthand for nonexperimental, indeed, nonempirical research on operating systems and compilers. We can distinguish four classes of empirical studies:

Exploratory studies yield causal hypotheses that are tested in observation or manipulation experiments. To this end, exploratory studies usually collect lots of data, analyzing it in many ways to find regularities.

Assessment studies establish baselines and ranges, and other assessments of the behaviors of a system or its environment.

Manipulation experiments test hypotheses about causal influences of factors by manipulating them and noting effects, if any, on one or more measured variables.

Observation experiments disclose effects of factors on measured variables by observing associations between levels of the factors and values of the variables. These are also called *natural* and *quasi-experimental* experiments.

Manipulation and observation experiments are what most people regard as proper experiments, while exploratory and assessment studies seem informal, aimless, heuristic, and hopeful. Testing hypotheses has the panache of "real science," whereas exploration seems like fishing and assessment seems plain dull. In fact, these activities are complementary; one is not more scientific than another; a research project will involve them all. Indeed, these activities might just as well be considered phases of a research project as individual studies.

The logic of assessment and exploratory studies is different from that of manipulation and observation experiments. The latter are *confirmatory*—you test explicit, precise hypotheses about the effects of factors. Exploratory and assessment studies suggest hypotheses and help to design experiments. These differences are reflected in methods and conventions for analyzing data. Manipulation experiments are analyzed with the tools of statistical hypothesis testing: t tests, analysis of variance, and so on. Results are outcomes that are very unlikely to have occurred by chance. Hypothesis testing is also applied to some observation experiments, but so are other forms of analysis. For example, it's common to summarize the data from an observation experiment in a regression model of the form

$$y = w_1 x_1 + w_2 x_2 + \ldots w_k x_k + C,$$

where y is called the response variable and $x_1 \ldots x_k$ are predictor variables. Typically, we'll appraise how well or poorly the regression model predicts y, but this is less a conclusion, in the hypothesis-testing sense, than an assessment of whether $x_1 \ldots x_k$ explain y. Not surprisingly, there are fewer conventions to tell you whether a regression model is acceptable. Obviously, a model that accounts for much of the variance in y is preferred to one that doesn't, all other things equal, but nobody will insist that predictive power must exceed some threshold, as we conventionally insist

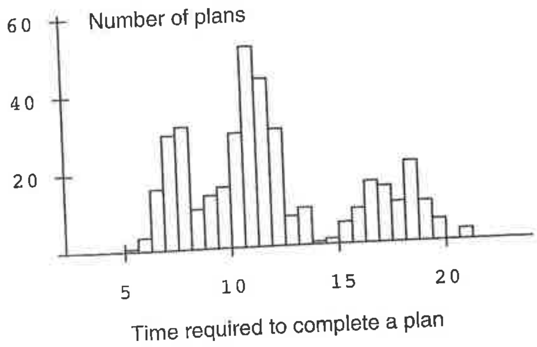

Figure 1.3 A frequency distribution.

that a result in hypothesis testing must have no more than a one-in-twenty chance of being wrong.

Data analysis for exploratory and assessment studies is freer still of conventions and definitions of significance. The purpose of these studies is to find suggestive patterns in data. For example, what does the distribution in figure 1.3 suggest to you? Suppose the horizontal axis represents the time required to complete a plan in a simulated environment and the heights of the bars represent the number of plans that required a particular amount of time: roughly thirty plans required 7.5 time units, more than fifty plans required 11 time units, and at least twenty plans required 18 time units to finish. These three peaks, or modes, suggest we are collecting data in three rather different conditions. Are we seeing the effects of three types of simulated weather? Three kinds of plans? Three different operating systems? Or perhaps there are just two conditions, not three, and the first "dip" in the histogram represents a sampling error. Are the modes equally spaced, or increasingly separated, suggesting a nonlinear relationship between conditions and time requirements? Why are more plans found in the second "hump" than in the first? Is it a sampling bias? Do fewer plans fail in this condition? These questions are representative of exploratory studies. They are answered by exploratory data analysis, which includes arranging and partitioning data, drawing pictures, dropping data that seem to obscure patterns, doing everything possible to discover and amplify regularities.

The strict standards of statistical hypothesis testing have no place here. Exploratory data analysis finds things in haystacks (data are rarely as clear as figure 1.3), whereas statistical hypothesis testing puts them under a microscope and tells us whether they are needles and whether they are sharp.