data mining algorithms fit into the solution landscape of various business analytical problem areas: operations research—OR, forecasting, data mining, statistics, and business intelligence—BI.

Figure 8.1 came from studies by Dustin Hux and John Elder (both of Elder Research, Inc.) on algorithms used in journal articles in different domains. From Figure 8.1, you can see which field uses what technique and also what techniques are suited to overlaps between areas. For example, visualization and cross-tabulations are used in business intelligence, data mining, and statistics.

Data miners use many analysis techniques from statistics, but often ignore some techniques like factor analysis (not always wisely). In addition, data mining includes a lot of techniques that are not considered typically in the world of statistics (such as radial basis function networks, genetic algorithms). Operations research (OR) uses clustering, graph theory, neural networks, and time series, but also depends very heavily on simulation and optimization. Forecasting overlaps data mining, statistics, and OR, and adds a few algorithms like Fourier transforms and wavelets.

In addition to the overlap of algorithms in different areas, some of them are known by different names. For example, Principal Components Analysis (PCA) is known in electrical engineering as the Karhounen-Loeve transform, and in statistics as the eigenvalue-eigenvector decomposition.

In our early college years, we take courses in many different disciplines, and it looks as though techniques are developed in them independently. One of the important byproducts of higher education (especially graduate school) is that we begin to see the interconnections between these ideas in different disciplines. The Ph.D. degree is short for Doctor of Philosophy. Doctoral degrees are handed out in many very technical disciplines, and it might seem strange that “philosophy” is still in the name. What does philosophy have to do with Recombinant DNA Genetics? The answer is “Everything.” One of the jokes often heard in graduate schools is “You learn more and more about less and less, until you know everything about nothing.” Well, a very highly constrained subject matter discipline is the end point (not quite “nothing”), and through the process of getting there, you can see the connections with a great many other disciplines. And this connected view of a broad subject area (e.g., genetics) provides the necessary philosophical framework for the study of your specific area. You are not educated properly in a discipline until you can view it in the context of its relationship with many other disciplines. So it is with the study of analytical algorithms. This book will take you far along that path (books like the one by Hastie et al., 2001, do it better), but this introduction will provide enough background to help you navigate through the plethora of data mining and statistical analysis algorithms available in most data mining tool packages.

Now, we will turn to the main job at hand in this chapter and look at each of the advanced algorithms individually. Because these algorithms are implemented in slightly different ways in each data mining or statistical package, we will cast the explanations in terms of how they are implemented in STATISTICA Data Miner (for which a free 90-day copy is available on the enclosed DVD). In addition to the free software, we have provided numerous tutorials (many of them use STATISTICA Data Miner; several with other tools, particularly SPSS Clementine and SAS Enterprise Miner). Some of the following text was adapted from the STATISTICA software online help: StatSoft, Inc. (2008). STATISTICA (data analysis software system), version 8.0. www.statsoft.com. You can experiment