from external sources. Information can be gathered from various data providers to enhance the corporate database, including

- Demographic data (available from Axiom, Experian, Equifax, Lexis-Nexis, etc.);
- Firmographic data (e.g., Dun & Bradstreet data and other business data sources);
- Psychographic data (inferences and classifications of people according to various measures of attitudinal and philosophical views).

**Very Few Data Sets and Modeling Details Are Available**

There is good reason for the lack of data sets and modeling details. You would not want potential fraudsters to learn how to defeat your detection strategies. Fraud data sets and modeling methodologies are tightly kept secrets. A company like Fair Isaac (generator of the FICO credit scores) has a huge library of predictor variables it won’t share with anyone. In academia, fraud researchers share their methods in very formal and general terms that only experts can understand, read “between the lines,” and relate to detailed instructions. Fraud modelers may be technical experts in a given business and would love to have access to detailed methodological presentations.

Very few fraud data sets are available in the public domain. Following are the only two that the authors are aware of:

2. The KDD Cup 1999 Network Intrusion Detection data set (http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html). *(Note: This data set will be used in the example described later.)*

**HOW DO YOU DETECT FRAUD?**

The basic approach to fraud detection with an analytical model is to identify possible predictors of fraud associated with known fraudsters and their actions in the past. The most powerful fraud models (like the most powerful customer response models) are built on historical data.

If the fraud response can be identified, it can be used to characterize the behavior of the fraudster in the specific fraud act and in historical data. The application of the term *supervised* is drawn from the broader discipline of classification (see Chapter 11 for an introduction to the terms *supervised* and *unsupervised*). Supervised classifications are based on some measure of true class membership of a given entity. According to Bolton and Hand (2002), supervised modeling has the drawback that it requires “absolute certainty” that each event can be accurately classified as fraud or nonfraud. In addition, the authors note that any models of fraud can be used to detect only types of fraud that have been identified previously.

Unsupervised methods of fraud modeling rely on detecting events that are abnormal. These abnormal events must be characterized by relating the events to symptoms associated