misled many people into believing that data mining is a product that can be bought rather than a discipline that must be mastered."

Very well stated! Automation is no substitute for human input. As we shall learn shortly, humans need to be actively involved at every phase of the data mining process. Georges Grinstein of the University of Massachusetts at Lowell and AnVil, Inc., stated it like this [15]:

> Imagine a black box capable of answering any question it is asked. Any question. Will this eliminate our need for human participation as many suggest? Quite the opposite. The fundamental problem still comes down to a human interface issue. How do I phrase the question correctly? How do I set up the parameters to get a solution that is applicable in the particular case I am interested in? How do I get the results in reasonable time and in a form that I can understand? Note that all the questions connect the discovery process to me, for my human consumption.

Rather than asking where humans fit into data mining, we should instead inquire about how we may design data mining into the very human process of problem solving.

Further, the very power of the formidable data mining algorithms embedded in the black-box software currently available makes their misuse proportionally more dangerous. Just as with any new information technology, *data mining is easy to do badly*. Researchers may apply inappropriate analysis to data sets that call for a completely different approach, for example, or models may be derived that are built upon wholly specious assumptions. Therefore, an understanding of the statistical and mathematical model structures underlying the software is required.

CROSS-INDUSTRY STANDARD PROCESS: CRISP–DM

There is a temptation in some companies, due to departmental inertia and compartmentalization, to approach data mining haphazardly, to reinvent the wheel and duplicate effort. A cross-industry standard was clearly required that is industry-neutral, tool-neutral, and application-neutral. The Cross-Industry Standard Process for Data Mining (CRISP–DM) [16] was developed in 1996 by analysts representing DaimlerChrysler, SPSS, and NCR. CRISP provides a nonproprietary and freely available standard process for fitting data mining into the general problem-solving strategy of a business or research unit.

According to CRISP–DM, a given data mining project has a life cycle consisting of six phases, as illustrated in Figure 1.1. Note that the phase sequence is *adaptive*. That is, the next phase in the sequence often depends on the outcomes associated with the preceding phase. The most significant dependencies between phases are indicated by the arrows. For example, suppose that we are in the modeling phase. Depending on the behavior and characteristics of the model, we may have to return to the data preparation phase for further refinement before moving forward to the model evaluation phase.

The iterative nature of CRISP is symbolized by the outer circle in Figure 1.1. Often, the solution to a particular business or research problem leads to further questions of interest, which may then be attacked using the same general process as before.

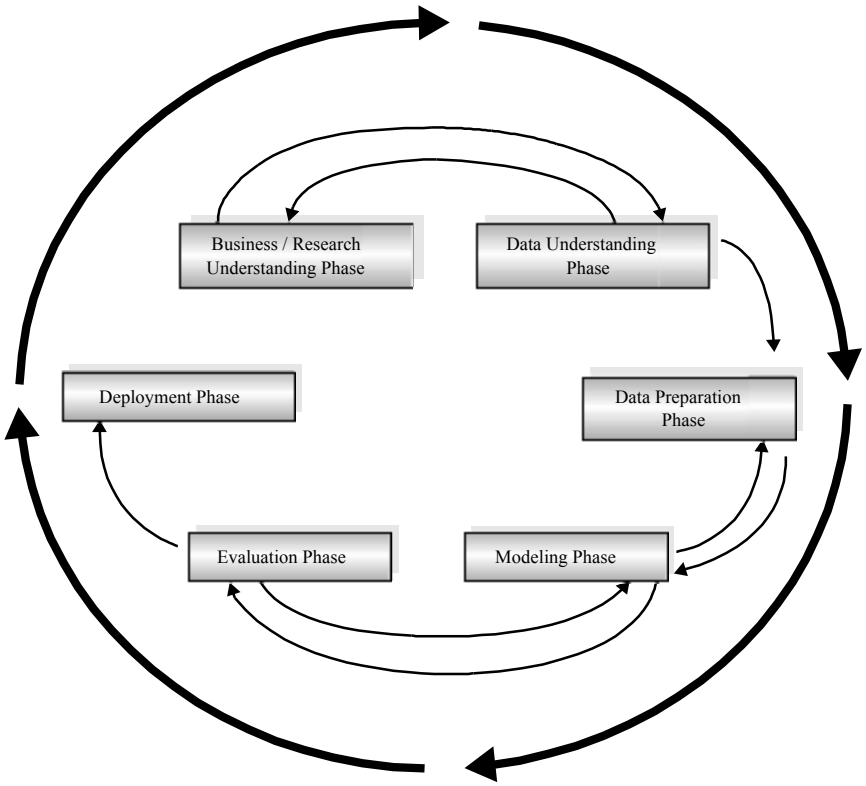

Figure 1.1 CRISP–DM is an iterative, adaptive process.

Lessons learned from past projects should always be brought to bear as input into new projects. Following is an outline of each phase. Although conceivably, issues encountered during the evaluation phase can send the analyst back to any of the previous phases for amelioration, for simplicity we show only the most common loop, back to the modeling phase.

CRISP–DM: The Six Phases

1. *Business understanding phase.* The first phase in the CRISP–DM standard process may also be termed the research understanding phase.
   a. Enunciate the project objectives and requirements clearly in terms of the business or research unit as a whole.
   b. Translate these goals and restrictions into the formulation of a data mining problem definition.
   c. Prepare a preliminary strategy for achieving these objectives.
2. *Data understanding phase*
   a. Collect the data.

   b. Use exploratory data analysis to familiarize yourself with the data and discover initial insights.
   c. Evaluate the quality of the data.
   d. If desired, select interesting subsets that may contain actionable patterns.

3. *Data preparation phase*
   a. Prepare from the initial raw data the final data set that is to be used for all subsequent phases. This phase is very labor intensive.
   b. Select the cases and variables you want to analyze and that are appropriate for your analysis.
   c. Perform transformations on certain variables, if needed.
   d. Clean the raw data so that it is ready for the modeling tools.

4. *Modeling phase*
   a. Select and apply appropriate modeling techniques.
   b. Calibrate model settings to optimize results.
   c. Remember that often, several different techniques may be used for the same data mining problem.
   d. If necessary, loop back to the data preparation phase to bring the form of the data into line with the specific requirements of a particular data mining technique.

5. *Evaluation phase*
   a. Evaluate the one or more models delivered in the modeling phase for quality and effectiveness before deploying them for use in the field.
   b. Determine whether the model in fact achieves the objectives set for it in the first phase.
   c. Establish whether some important facet of the business or research problem has not been accounted for sufficiently.
   d. Come to a decision regarding use of the data mining results.

6. *Deployment phase*
   a. Make use of the models created: Model creation does not signify the completion of a project.
   b. Example of a simple deployment: Generate a report.
   c. Example of a more complex deployment: Implement a parallel data mining process in another department.
   d. For businesses, the customer often carries out the deployment based on your model.

You can find out much more information about the CRISP–DM standard process at `www.crisp-dm.org`. Next, we turn to an example of a company applying CRISP–DM to a business problem.