Solutions to Chapter 1
AN INTRODUCTION TO DATA MINING
Prepared by James Cunningham, Graduate Assistant

1. **Refer to the Bank of America example early in the chapter. Which data mining task or tasks are implied in identifying "the type of marketing approach for a particular customer, based on customer's individual profile"? Which tasks are not explicitly relevant?**

   Relevant tasks include the following:
   - Description
   - Classification
   - Clustering
   - Associating

   Non-relevant tasks are:
   - Estimation
   - Prediction

2. **For each of the following, identify the relevant data mining task(s):**

   a. **The Boston Celtics would like to approximate how many points their next opponent will score against them.**

      Estimation: estimating the number of points (numeric target).

   b. **A military intelligence officer is interested in learning about the respective proportions of Sunnis and Shias in a particular strategic region.**

      Description: exploratory data analysis finds similarities and differences between the Sunni and Shias proportions.

   c. **A NORAD defense computer must decide immediately whether a blip on the radar is a flock of geese or an incoming nuclear missile.**

      Classification: a trained model detects incoming missiles assigns the blip on the radar screen (unclassified record) as being either a "missile" or "not missile" (categorical target); Estimation: an estimated numeric value may indicate the blip as an incoming missile.

d. A political strategist is seeking the best groups to canvass for donations in a county.

Description: relevant patterns describe the characteristics of one or more groups are located in the county; Clustering: examine the profile of each homogeneous group derived from a particular county's population; Association: discover interesting rules pertaining to a large proportion of the population.

e. A Homeland Security official would like to determine whether a certain sequence of financial and residence moves implies a tendency to terrorist acts.

Description: the sequences of financial and residential moves (patterns) may suggest a tendency (explanation) for terrorist activities; Classification: build a model to classify behavior as "suspicious"; Estimation: the model generates a numeric score indicating a propensity for committing terrorist acts.

f. A Wall Street analyst has been asked to find out the expected change in stock price for a set of companies with similar price/earnings ratios.

Estimation: the expected change in stock price (numeric target) using the price/earnings ratio for a similar set of companies (predictors); Prediction: applied when results expected to predict future price.

3. For each of the following meetings, explain which phase in the CRISP-DM process is represented:

a. Managers want to know by next week whether deployment will take place. Therefore, analysts meet to discuss how useful and accurate their model is.

The Evaluation Phase determines whether the data mining model achieves the objectives established in the first phase.

b. The data mining project manager meets with the data warehousing manager to discuss how the data will be collected.

Although the data warehouse is identified as a resource during the Business Understanding Phase, the actual data collection takes place during the Data Understanding Phase.

c. The data mining consultant meets with the Vice President for Marketing, who says that he would like to move forward with customer relationship management.

The primary objectives of the business are stated as part of the Business Understanding Phase.

d. The data mining project manager meets with the production line supervisor, to discuss implementation of changes and improvements.

The requirements of a data mining technique used during the Modeling Phase may cause the process to loop back to the Data Preparation Phase, with the goal of improving data quality. The Evaluation Phase determines whether specific improvements or process changes are required to ensure that all important aspects of the business are accounted for.

e. The analysts meet to discuss whether the neural network or decision tree models should be applied.

During the Modeling Phase one or more modeling techniques are chosen.

4. Discuss the need for human direction of data mining. Describe the possible consequences of relying on completely automatic data analysis tools.

The case studies emphasize the need for human involvement during every phase of the data mining process. For example, data mining initiatives using legacy database systems should not underestimate the time or importance required from domain experts to interpret the data. Taking shortcuts during this initial phase leads to potentially costly, inaccurate results in subsequent phases.

5. CRISP-DM is not the only standard process for data mining. Research an alternative methodology. (Hint: SEMMA, from the SAS Institute.) Discuss the similarities and differences with CRISP-DM.

SEMMA is an acronym representing the core data mining processes: sample, explore, modify, model, and assess. As compared to CRISP-DM, SEMMA places emphasis on the model development process of data mining and therefore does not contain a Business Understanding Phase or a Deployment Phase; however, it does describe the importance of having clear business objectives and using quality data sources for modeling.

Both processes are iterative and may loop back to other process steps as new information is learned or data mining requirements change. Also, both methods emphasize the use of an adaptive process. The following table shows how the phases of the two processes correspond to one another:

SEMMA	CRISP-DM
N/A	Business Understanding Phase
Sample	N/A[1]
Explore	Data Understanding Phase
Modify	Data Preparation Phase
Model	Modeling Phase
Assess	Evaluation Phase
N/A	Deployment Phase

Table 5.1. SEMMA vs CRISP-DM

[1] Although Sample does not correspond to a specific CRISP-DM phase, it often occurs during the Data Understanding, Data Preparation, and Modeling Phases.