1 What Is Data Mining?

Originally, "data mining" was a statistician's term for overusing data to draw invalid inferences.

- Bonferroni's theorem warns us that if there are too many possible conclusions to draw, some will be true for purely statistical reasons, with no physical validity.
- Famous example: David Rhine, a "parapsychologist" at Duke in the 1950's tested students for "extrasensory perception" by asking them to guess 10 cards — red or black. He found about 1/1000 of them guessed all 10, and instead of realizing that that is what you'd expect from random guessing, declared them to have ESP. When he retested them, he found they did no better than average. His conclusion: telling people they have ESP causes them to lose it!

Our definition: "discovery of useful summaries of data."

1.1 Applications

Some examples of "successes":

1. Decision trees constructed from bank-loan histories to produce algorithms to decide whether to grant a loan.
2. Patterns of traveler behavior mined to manage the sale of discounted seats on planes, rooms in hotels, etc.
3. "Diapers and beer." Observation that customers who buy diapers are more likely to by beer than average allowed supermarkets to place beer and diapers nearby, knowing many customers would walk between them. Placing potato chips between increased sales of all three items.
4. Skycat and Sloan Sky Survey: clustering sky objects by their radiation levels in different bands allowed astromomers to distinguish between galaxies, nearby stars, and many other kinds of celestial objects.
5. Comparison of the genotype of people with/without a condition allowed the discovery of a set of genes that together account for many cases of diabetes. This sort of mining will become much more important as the human genome is constructed.

1.2 The Data-Mining Communities

As data-mining has become recognized as a powerful tool, several different communities have laid claim to the subject:

1. Statistics.
2. AI, where it is called "machine learning."
3. Researchers in clustering algorithms.
4. Visualization researchers.
5. Databases. We'll be taking this approach, of course, concentrating on the challenges that appear when the data is large and the computations complex. In a sense, data mining can be thought of as algorithms for executing very complex queries on non-main-memory data.

1.3 Stages of the Data-Mining Process

1. *Data gathering*, e.g., data warehousing, Web crawling.
2. *Data cleansing*: eliminate errors and/or bogus data, e.g., patient fever = 125.
3. *Feature extraction*: obtaining only the interesting attributes of the data, e.g., "date acquired" is probably not useful for clustering celestial objects, as in Skycat.
4. *Pattern extraction and discovery.* This is the stage that is often thought of as "data mining," and is where we shall concentrate our effort.
5. Visualization of the data.
6. Evaluation of results; not every discovered fact is useful, or even true! Judgement is necessary before following your software's conclusions.