Statistics 215b – 11/20/03 – D.R. Brillinger

Data mining

A field in search of a definition – a vague concept

D. Hand, H. Mannila and P. Smyth (2001). *Principles of Data Mining.* MIT Press, Cambridge.

Some definitions/descriptions

“Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.”

Hand et al.

“… data mining refers to *extracting or ‘mining’ knowledge from large amounts of data*”

“…, data mining should have been more appropriately named ‘knowledge mining from data’”

“An analytic process designed to explore data (usually large amounts of data – typically business or market related) in search of consistent and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data.”

“Data mining is the application of a specific algorithm (usually within machine learning) for extracting patterns from data.”

“A simple approach to the theory of data mining is to declare that data mining is statistics (perhaps on larger data sets than previously) …”

“Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.”

“Data mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions.”

“Data mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data.”

“Data mining is the process of discovering advantageous patterns in data.”

“Data mining is a decision support process where we look in large data bases for unknown and unsuspected patterns of information.”

“Data mining is the process of seeking interesting or valuable information within large databases.”

“Data mining is the exploration and automatic analysis of large data sets, by automatic or semiautomatic means with the purpose of discovering meaningful patterns.”

“Data mining – the science of extracting useful information from large data sets.”

“Data mining is a knowledge discovery process of extracting previously unknown, actionable information from very large databases.”

“Data mining is finding interesting structure (patterns, statistical models, relationships) in databases.”

“Data mining is a process that uses a variety of tools to discover patterns and relationships in data that may be used to make valid predictions.”

“Data mining. The process of efficient discovery of nonobvious valuable patterns from a large collection of data.”

“Data mining is exploratory data analysis with little or no human interaction using computationally feasible techniques, i.e. the attempt to find interesting structure unknown a priori.”

“Data-mining is the art and science of teasing meaningful information and patterns out of large quantities of data.”

“Data mining is in fact an umbrella term for a variety of analytic techniques.”

“Data mining is about digging into data to find subtle patterns and informative relationships amongst the data resources piling up in today’s businesses.”

“Data mining*, the extraction of hidden predictive information from large databases*,…”

“The knowledge discovery and data mining (KDD) field draws on the findings from statistics, databases, and artificial intelligence to construct tools that let users gain insight from massive data sets.”

“Data mining.

Objective: explore the data for interesting patterns.

Approach: size is overcome to search for information.

Criteria: data become findings by obtaining questions.

Focus: association is sought through coincidence.

Stats: sampling and design aid start?
Sequential analysis aids end?

Process: reduce the data with maximum gain.
Behavior: break old rules powered by technology.

Attitude: seek local effects proactively & persistently”

Arnold Goodman

Data mining refers “to the exaggerated claims of significance and or forecasting precision generated by the selective reporting of results obtained when the structure of the model is determined experimentally by repeated applications of such procedures as regression analysis to the same body of data … synonymous with ‘data scrubbing’, ‘data fishing’, Darwinian econometrics (survival of the fittest).’

“…, the term ‘data mining’ is sometimes used perjoratively to describe such work, particularly when an analyst has searched over a large model space without adjusting for such a search or testing the resulting model on new data.”

“ ‘Data mining’, ‘fishing’, ‘grubbing’, ’number crunching’. These are value-laden terms we use to disparage each other’s empirical work with the linear regression model. A less provocative description would be '‘specification search'’ and a catch-all definition is ‘the data-dependent process of selecting a statistical model’.”

-“ ‘mining’ suggests that the activity may, in fact, be productive.”

Tukey "… *data analysis*, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."