Executive Information Systems, Inc.

Data Mining and KDD: A Shifting Mosaic

By

Joseph M. Firestone, Ph.D.

White Paper No. Two

March 12, 1997

The Idea of Data Mining

Data Mining is an idea based on a simple analogy. The growth of data warehousing has created mountains of data. The mountains represent a valuable resource to the enterprise. But to extract value from these data mountains, we must "mine" for high-grade "nuggets" of precious metal -- the gold in data warehouses and data marts. The analogy to mining has proven seductive for business. Everywhere there are data warehouses, data mines are also being enthusiastically constructed, but not with the benefit of consensus about what data mining is, or what process it entails, or what exactly its outcomes (the "nuggets") are, or what tools one needs to do it right.

Data Mining as a field is not yet through with the process of definition and conceptualization of the scope of the field. There are at least three distinct concepts of data mining being used by practioners and vendors.

DMI: Data mining is traditional data analysis methodology updated with the most advanced analysis techniques applied to discovering previously unknown patterns. A specific instance of this concept, stated more explicitly and with a more commercial orientation is provided by the SAS Institute.

SAS defines data mining **as *the process of selecting, exploring, and modeling large amounts of data to uncover previously unknown patterns for a business advantage.*** [1]

In specifying its notion of data mining further SAS describes it as involving a five step process: Sample, Explore, Modify, Model, and Assess, or the acronym SEMMA. The difference between SEMMA and traditional methodology used in statistical analysis is hard to see with the naked eye, though I emphasize that methodology and tools or techniques are different things, and I am certainly not saying that because SAS's SEMMA methodological approach is traditional, it would not or could not incorporate the most advanced data mining tools.

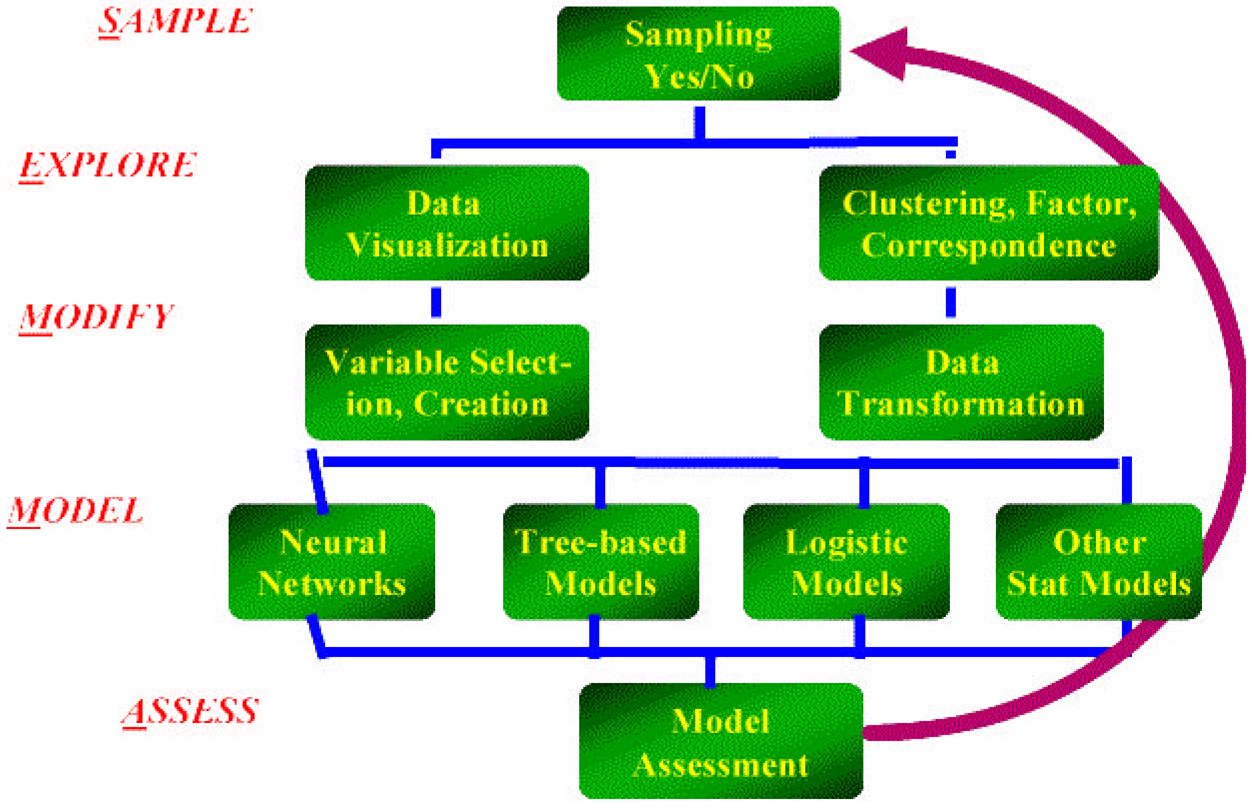

Figure 1 -- SAS's SEMMA Process
(Adapted from the SAS Institute Web Site)

DMII: Data Mining is the activity of extracting hidden information (patterns and relationships) from large databases automatically: that is, without benefit of human intervention or initiative in the knowledge discovery process. In this view, data mining is knowledge discovery in databases, or at least it is automated knowledge discovery in databases.

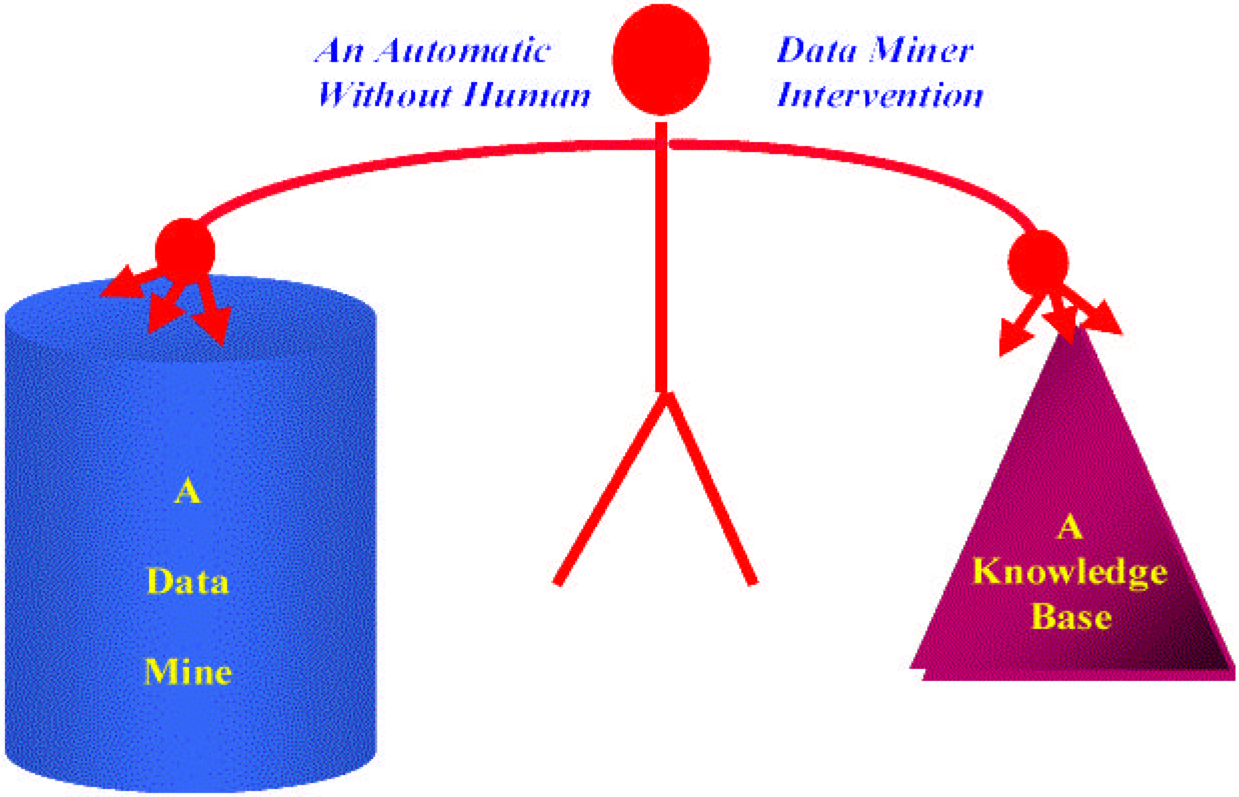

Figure 2 -- Automated Data Mining

DMII is the data mining concept implicit in the advertising collateral of many specialized data mining companies. Customers are promised an automatic process of model development that requires little or no human interaction from sophisticated data analysts. The data mining package supplies the necessary high quality analysis, and business users are promised that they can achieve knowledge discovery and predictive success on their own, and with little investment of time or effort compared to what is necessary with non-data mining (often labeled as traditional statistical) techniques.

We don’t see this concept as much outside of vendor literature, but it is either present, or closely approached in many articles on data mining. Its advocates draw a sharp distinction between data-driven tools using automated discovery-based approaches and user- or verification-driven tools using hypothesis-testing approaches. The hypothesis-testing tools are seen as limited by the skill and experience of humans, while the data mining tools are seen as free of human initiative or assumption, and empowered by pattern-matching algorithms. Most importantly, the hypothesis-testing tools are seen as "verifiers," while the data mining tools are seen as "discoverers."

DM III: <u>Data Mining is the step in the process of knowledge discovery in databases, that inputs predominantly cleaned, transformed data, searches the data using algorithms, and outputs patterns and relationships to the interpretation/evaluation step of the KDD</u>

process. DMIII is my statement of the view of data mining emerging from the 1994 AAAI workshop on KDD, the KD Mine [2], and S*I*FTWARE [3] web sites and the recent Advances in Knowledge Discovery and Data Mining volume (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.)) [4]. The definition clearly implies that what data mining (in this view) discovers is hypotheses about patterns and relationships. Those patterns and relationships are then subject to interpretation and evaluation before they can be called knowledge.

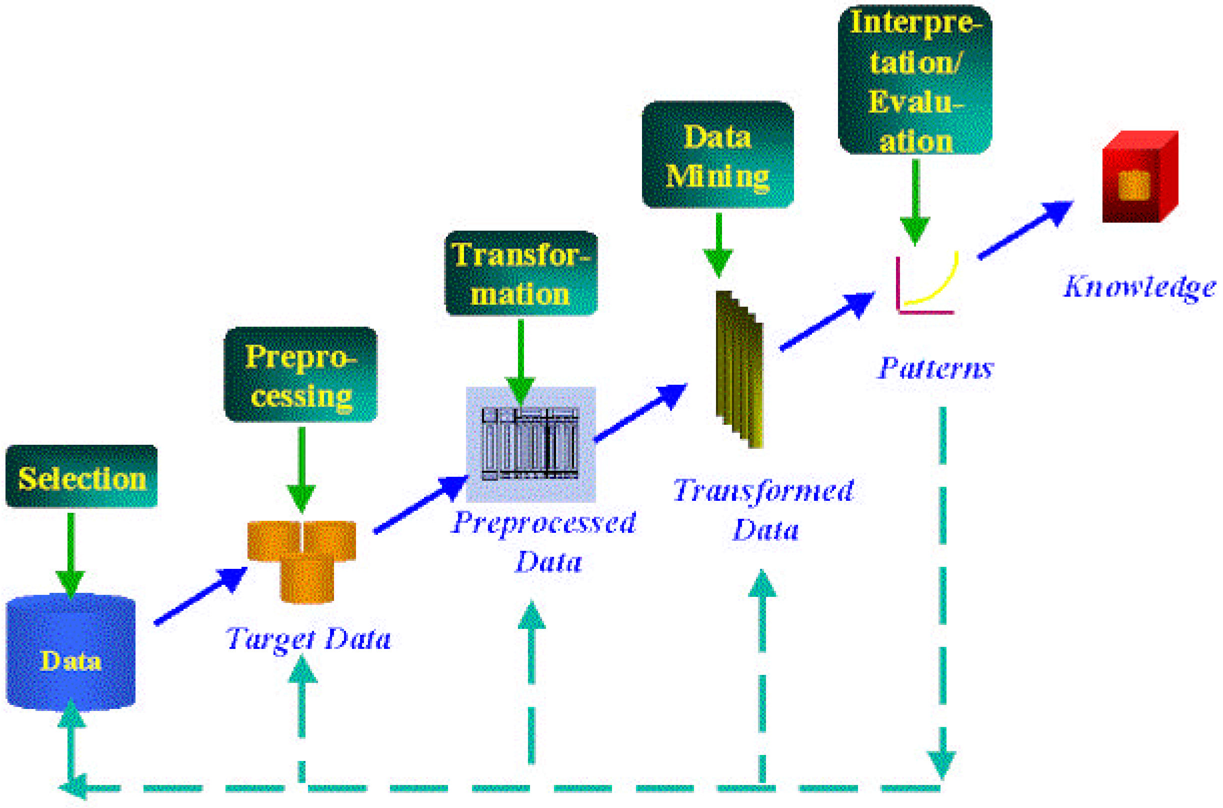

Figure 3 -- Data Mining in KDD (Adapted from Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy (eds.))

Nor does the commitment to search algorithms in DMIII imply a completely automated data mining process. Data analyses must use algorithms with some degree of search autonomy to qualify as instances of data mining, but they will also use human initiative in the areas of background knowledge (a specification of which is required for applying some machine learning and case-based reasoning techniques), model selection and specification, input and output variable selection and specification, in constraining model parameters and in other ways. In short, the data mining process described by those adhering to DMIII is one in which automated search algorithms play a vital role in complex iterative "interactions, protracted over time, between a *human* and a database . . ." (see the Brachman and Anand, as well as, any number of other studies in the Advances in Knowledge Discovery and Data Mining volume) [5].

<u>The Idea of Knowledge Discovery in Databases (KDD)</u>

The three definitions of data mining are also closely associated with three apparently different concepts of KDD. DMI is associated with no explicit concept of KDD. But for purposes of discussion, I will assume that supporters of DMI believe that KDD refers to a process that uses computer-based data analysis as a primary means of investigation, and that produces scientifically validated knowledge. Here are the three KDD concepts.

EDA = Exploratory Data Analysis

CDA = Confirmatory Data Analysis

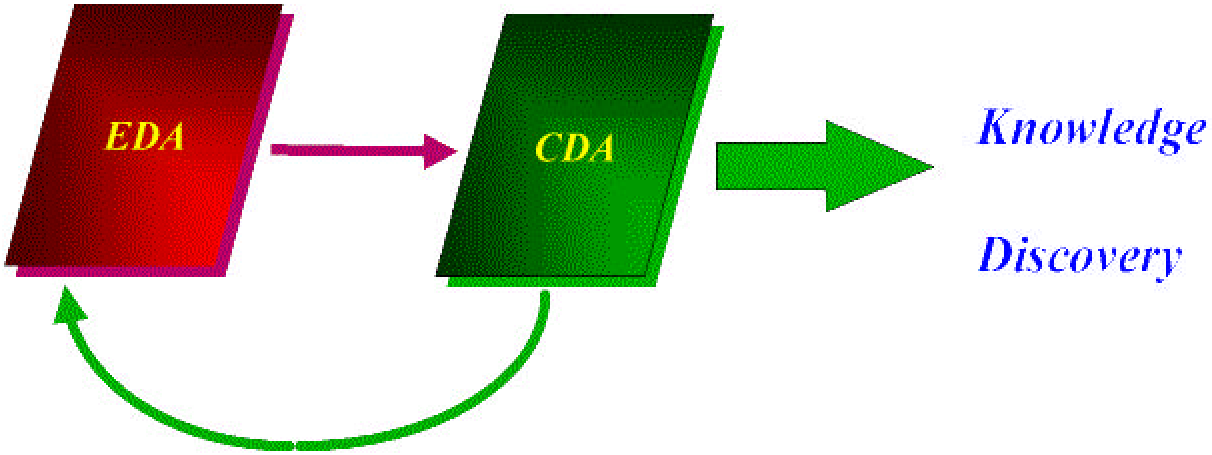

Figure 4 -- The KDDI Position

KDDI: <u>Knowledge discovery in databases is a process that requires hypothesis or model formulation, hypothesis or model testing, and derivatively all the data, techniques, and sub-processes necessary to bring hypothesis or model testing to a successful conclusion.</u> In this view, data analysis includes both exploratory, and confirmatory data analysis, and the latter is necessary for hypothesis or model testing. The outcome of hypothesis or model testing is knowledge discovery, even if the knowledge discovered is a negative finding that some hypothesis or model is not knowledge.

KDDII: <u>True knowledge discovery in databases is the process of automated data mining applied without benefit of human intervention or initiative.</u> According to this view there is no distinction between data mining and KDD. Data mining doesn’t just generate hypotheses. It produces valid knowledge that businesses can apply without fear of bad results.

KDDIII: "*Knowledge discovery in databases* is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." (Fayyad, Piatetsky-Shapiro, and Smyth, in Advances . . . P. 6) Further, this process includes five steps: data selection, data preprocessing, data transformation, data mining, and interpreting and evaluating mined patterns and relationships (P. 9, above). This process is interactive and interative with KDD users heavily involved at every step. Many loops may occur between steps. There is no deterministic progression assumed from one step to another. Also, the interpretative and evaluative step, can involve returns to any of the previous steps, any number of times.

How Data Mining Relates to KDD

The relationships between data mining and KDD are different for the three approaches generated by the DM/KDD pairs, and are implicit in the definitions already presented. We'll consider these implications in the same order the contrasting DM and KDD concepts were presented.

Traditional Data Mining

DMI and KDDI, equate data mining with KDD and since they don't distinguish it from previous investigative processes, they essentially equate both with previous procedures and methodologies of analytical and statistical modeling. SAS's SEMMA data mining process could be used equally well to describe traditional processes of analysis followed for years by SAS users. Some of the tools, such as Neural Networks and Tree-based models may be different, but the patterns of investigation, and more important, of validation, are essentially the same. It is hard to escape the conclusion that for this approach, data mining is traditional modeling and analysis updated with the addition of some new techniques and incorporated into the commercially relevant data warehouse framework.

In drawing this conclusion I don't mean to be pejorative or to express criticusm for traditional approaches. If the DMI/KDDI explication of data mining makes the most sense for further development of the field, so be it. But it is important to recognize the approach for what it is, and to refrain from claiming methodological novelty, when we are really talking about progress in software and hardware tools for data analysis.

In the DMI/KDDI approach, also, data mining is not restricted to the step of hypothesis formation. The SEMMA model assessment step is a validation step. That is why data mining and KDD can be so easily equated.

But though data mining and KDD are equated, the data mining/KDD process is not viewed as fully automated. The traditional approach recognizes the vital roles of human-initiated hypothesis and model formation, and computer-based, partly automated, exploratory and confirmatory data analysis in data mining.

Automated Data Mining

The DMII and KDDII definitions of data mining and knowledge discovery in databases also equate the two. But in contrast to DMI/KDDI, they imply that ***automated data mining,*** as expressed by DMII, ***includes both hypothesis formulation and scientific validation in the algorithmic process.*** That's why it is logically consistent for proponents of DMII to claim that data mining and knowledge discovery are one and the same, and it is also why they view the knowledge discovery process as one that excludes human intervention or initiative. But are there any commercial data mining products, or even any research efforts that fit this definition?

The answer is, it depends on what you mean by scientific validation in KDD, a matter on which there is far from universal agreement. All data mining products produce patterns and relationships from appropriate data input. So, all can produce hypotheses. But products may vary in the extent and scope of their application of validation criteria, as well as in the specific validation criteria they use, during the automated data mining process. All data mining products seem to use empirical fit to sample data, as a general validity criterion, but the specific algorithms evaluating empirical fit may vary from product-to-product, and the criteria for measuring empirical fit may also vary. Also, empirical fit of patterns to data is not the whole story, sometimes patterns can be overfit to data, and different products employ different tests of overfit.

Apart from empirical fit of patterns to data, there are many other validation criteria that can be applied to interpretation and evaluation of the results of data mining. To fully understand this, it is necessary to step back from the perspective of one or two products of a particular type, and to recognize that there are many data mining models out there, and that they offer alternative theories of data. Which one is right for your data mine? If you compare results of a number of alternatives, what criteria do you use to compare them? What if all fit the data well empirically, but the criteria of empirical fit are either not readily compared because of the use of different fit statistics, or a comparison is not meaningful because you can't tell which model involves overfitting? What if the same data mining product has generated alternative patterns based on slightly different input assumptions with no material difference in empirical fit? How do you choose then?

Recent KDD research has specified a number of criteria apart from empirical fit to sample data as relevant including: **predictive validity** on new data, **novelty** of the patterns discovered by the data mining tool, **pragmatic utility** of the patterns as measured by some utility function, **ultimate understandability** of the patterns, and a composite of these called **"interestingness."** However, even though these criteria can be listed, research on applying them is not far advanced, and promises to be difficult to implement. Nor are these criteria in any sense exhaustive. Almost anyone in the KDD field today, could specify additional criteria or alternatives for at least some of the criteria of validation and provide an equally plausible defense of these as reasonable validation criteria.

Validity criteria in KDD is a developing area of research, and there is no consensus yet on standards, procedures or algorithms for measuring validity.

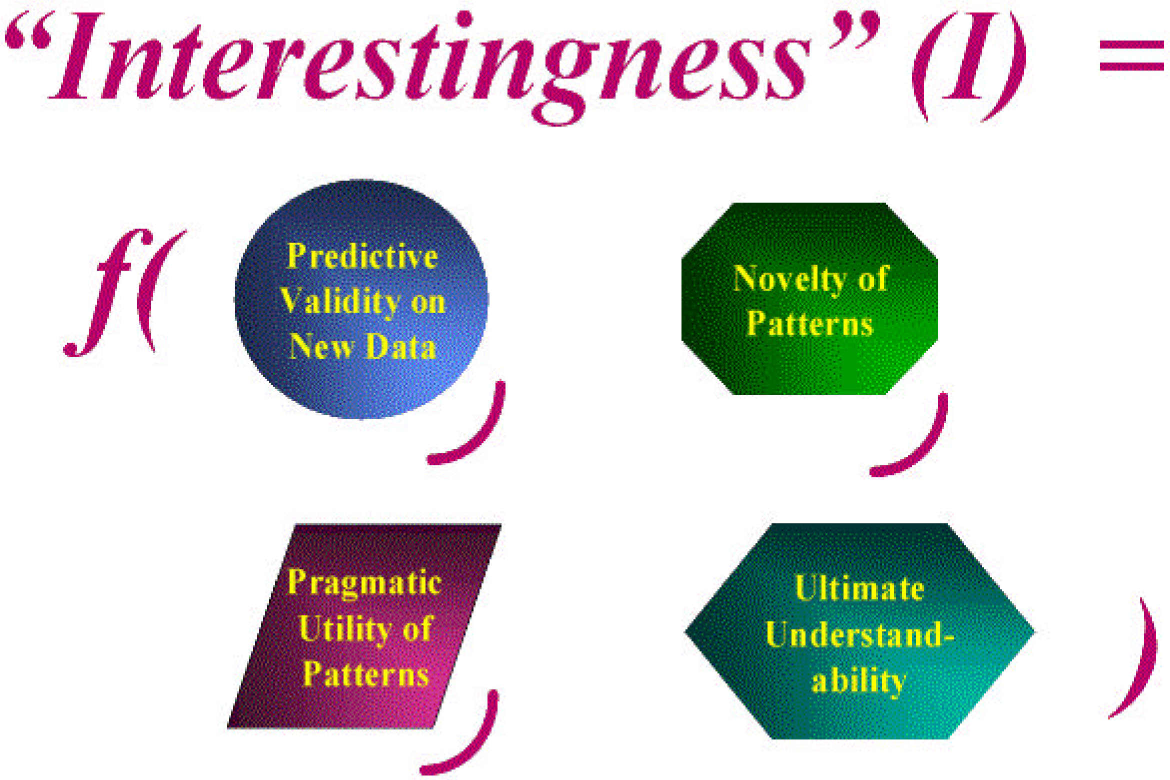
"Interestingness" (I) =
f(
Predictive Validity on New Data
Novelty of Patterns
Pragmatic Utility of Patterns
Ultimate Understand-ability
)

Figure 5 -- The "Interestingness" Criterion

Without such a consensus the DMII/KDDII concept of automated data mining is premature. The results of data mining activities cannot now be validated by an algorithm or algorithms incorporating generally agreed upon validity standards. And there is no prospect that such validation will be available in the near future.

So, whatever the preferences of advocates of the DMII/KDDII position, the outcome of current automatic data mining investigations in the DMII/KDDII sense, must be viewed as highly hypothetical, exploratory in nature, and subject to a careful validation analysis before they are relied on for practical applications. Considering the exploratory nature of results using the automated data mining perspective, I believe that vendors and consultants who are selling data mining on the basis of the DMII/KDDII position, are overselling data mining.

Data Mining as Part of KDD

The DMIII/KDDIII position is probably the one with the most current momentum. It attempts to distinguish data mining from traditional analyses by emphasizing the automated character of data mining in generating patterns and relationships, but it also clearly distinguishes data mining from knowledge discovery, by emphasizing the much broader character of KDD as an overarching process, including an interpretation and evaluation step distinct from data mining and relying more heavily on human interaction. In a very real sense DMIII/KDDIII is a middle way between the other two positions.

But if DMIII/KDDIII is a middle way, that does not necessarily mean it is the right way. Sometimes compromises are just unstable platforms for methodological development. DMIII/KDDIII seems to postulate no difference from the traditional data mining approach in the area of validation or confirmatory analysis. The difference is in the area of exploratory data analysis where practitioners holding this position emphasize the use of automated methods to generate patterns, while practitioners of DMI/KDDI don't talk quite so much about automation, but talk more about using a variety of techniques including human initiative to guide exploratory analysis. But is this difference a real methodological difference between the two camps, or just a way of maintaining a distinct identity, of placing old wine in new bottles?

Current studies by participants in the KDD group make overwhelmingly clear the exhaustive interaction between human and machine that is part of the data mining process in a real KDD project. The iterative process to prepare for data mining and to implement it follows the careful investigative pattern of traditional analysis. The algorithmic techniques are more powerful than they were ten years ago, but there is no methodological requirement that pattern generation be guided solely by automated data mining techniques. Instead, the requirement is a focus on techniques with a certain degree of search autonomy. -- a small difference from the viewpoint of traditional data mining at best.

The Data Mining Future

The Data Mining foreseeable future will involve an appreciable human component, whether we're taking the viewpoint of either DMI or DMIII. The problems inherent in model and variable selection, in measurement and dynamic model construction, and in pattern validation methodology all guarantee that.

But, it is also true that we will continue to make progress in the area of adaptive intelligence that underlies data mining. DMII/KDDII may be an invalid construct now, but research on computational models of theory evaluation (See Paul Thagard's Conceptual Revolutions [6] for background) will eventually bring us much closer to having measurement models of validity and to having agreement on both the models and the criteria they incorporate. Also, the new analysis techniques (Neural Networks, Genetic Algorithms, Machine Learning, Bayesian Belief Networks, Fuzzy Engineering, Chaotic Dynamics, etc.) that have come to prominence in the last 10 to 15 years, and that are now becoming fully commercialized, will continue to advance in power and sophistication and to become more fully integrated in analysis methodologies that we can partially automate.

For now, the practical task at hand is to bring to bear the most powerful analytical techniques at our disposal to the problem of making private and public enterprises more adaptive. Practically speaking, this means analysis of enterprise performance in all its aspects through use of the data in data warehouses and data marts. Exploratory analysis of this data is called data mining (DMI & DMIII). Sometimes confirmatory data analysis is also included in data mining (DMI). The important thing is that, for the foreseeable future, good data mining cannot be done without significant human interaction between a human data miner and her computer-based software extension. That is because data mining is not automatic. And the dream of making it so, is, at best, an ideal motivating long-term development.

References

1. SAS Institute: http://www.sas.com/feature/4qdm/whatisdm.html

2. The KD Mine: http://www.kdnuggets.com

3. S*I*FTWARE: http://kdnuggets.com/siftware.html

4. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy, Advances in Knowledge Discovery & Data Mining. Cambridge, MA (The AAAI Press/The MIT Press) 1996.

5. Ronald J. Brachman and Tej Anand, "The Process of Knowledge Discovery in Databases," in Advances in Knowledge . . . Pp. 37-57.

6. Paul Thagard, Conceptual Revolutions (Princeton, NJ: Princeton University Press, 1992).

Biography

Joseph M. Firestone is an independent Information Technology consultant working in the areas of Decision Support (especially Data Marts and Data Mining), Business Process Reengineering and Database Marketing. He formulated and is developing the idea of Market Systems Reengineering (MSR). In addition, he is developing an integrated data mining approach incorporating a fair comparison methodology for evaluating data mining results. You can e-mail Joe at eisai@home.com

[Up] [Data Warehouses and Data Marts: New Definitions and New Conceptions]
[Is Data Staging Relational: A Comment]
[DKMA and The Data Warehouse Bus Architecture]
[The Corporate Information Factory or the Corporate Knowledge Factory]
[Architectural Evolution in Data Warehousing]
[Dimensional Modeling and E-R Modeling in the Data Warehouse]
[Dimensional Object Modeling] [Evaluating OLAP Alternatives]
[Data Mining and KDD: A Shifting Mosaic]
[Data Warehouses and Data Marts: A Dynamic View]
[A Systems Approach to Dimensional Modeling in Data Marts]
[Object-Oriented Data Warehousing]