DMReview

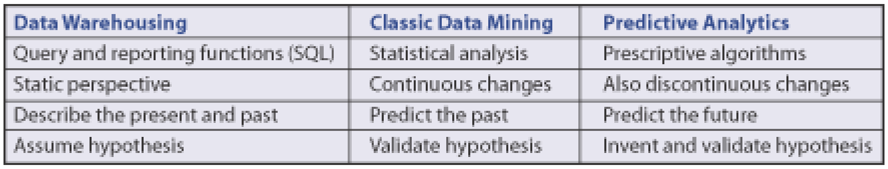
Published in DM Review in January 2004.
Printed from DMReview.com

Data Warehousing Lessons Learned: Data Mining is Dead – Long Live Predictive Analytics!

by Lou Agosta

This is why data mining is dead: it died of a broken heart. It was killed by disappointed expectations. In addition to a perfect storm of tough economic times, another reason data mining technology has not lived up to its promise is that "data mining" is a vague and ambiguous term. It overlaps with data profiling, data warehousing, and even such approaches to data analysis as online analytical processing (OLAP) and enterprise analytic applications. When high profile successes have occurred (e.g., a front-page article in the Wall Street Journal, "Lucky Numbers: Casino Chain Mines Data on Its Gamblers, And Strikes Pay Dirt" by Christina Binkley, May 4, 2000), they have been a mixed blessing. Such results have attracted a variety of imitators with claims, solutions and products that ultimately fall short of the promises. The promises build on the mining metaphor and typically are made to sound like easy money. This has resulted in all the usual dilemmas of confused messages from vendors, hyperbole in the press and disappointed end-user enterprises.

Data mining is regrouping as "predictive analytics." The differentiators are summarized in Figure 1.



Data Warehousing	Classic Data Mining	Predictive Analytics
Query and reporting functions (SQL)	Statistical analysis	Prescriptive algorithms
Static perspective	Continuous changes	Also discontinuous changes
Describe the present and past	Predict the past	Predict the future
Assume hypothesis	Validate hypothesis	Invent and validate hypothesis

Figure 1: Data Mining and Predictive Analytic Differentiators

- **Prescriptive, not merely descriptive:** Scanning through a terabyte haystack of billing data for a few needles of billing errors is properly described as data mining. However, it is descriptive, not prescriptive. When a model is able to predict errors based on a correlation of variables ("root cause analysis"), then the analysis is able to recommend what one ought to do about the problem (and is, therefore, prescriptive). Note that the model expresses a "correlation," not a "causation," though a cause-and-effect relation can often be inferred. For example, Xerox uses Oracle's Data Mining software, for clustering defects and building predictive models, to analyze usage profile history, maintenance data and representation of knowledge from field engineers to predict photocopy component failure. The copier then sends an e-mail to the repair staff to schedule maintenance prior to the breakdown.
- **Stop predicting the past; predict the future:** Market trend analysis as performed in data warehousing, OLAP and analytic applications often asks what customers are buying or using (product or service), and then draws a straight line from the past into the future, extrapolating a trend. This too can be described as data mining. One might argue this predicts the future because it says something about what will happen. However, a more accurate description would be that it "predicts the past" and then projects that into the future. The prediction is not really in the analysis. Furthermore, data mining in the limited sense used here is only able to envision continuous change – extending the trend from past to future. Predictive analytics is also able to generate scores from models that envision discontinuous changes – not only peaks and valleys, but cliffs and crevasses. This is especially the case with "black box"-type functions such as neural networks and genetic

programming. Rarely do applications in OLAP, query and reporting or data warehousing explicitly relate independent and dependent variables, but that is of the essence in predictive analytics. For example, KXEN is used to find the optimal point between savings of catching a bad customer versus the cost of turning away a good paying customer (opportunity cost).

- **Invent hypotheses, don't merely test them:** Finally, data mining is distinguished from predictive analytics in terms of hypothesis formulation and validation. For example, one hypothesis is that people default on loans due to high debt. Once the analyst formulates this hypothesis by means of imaginative invention out of her or his own mind, the OLAP analyst then launches queries against the data cube to confirm or invalidate this hypothesis. Predictive analytics is different in that it can look for patterns in the data that are useful in formulating hypotheses. The analyst might not have thought that age was a determinant of risk, but a pattern in the data indicates that as a useful hypothesis for further investigation.

One reason that data alone is not knowledge but merely data is that it lacks structure, organization, direction, coherence, point and conceptual focus. Just as a predictive model without supporting data would be empty, likewise data without a unifying model is meaningless and leaves the collector blind. Giga clients will need to have expertise in all three dimensions: details of the business, data collection and model building. Client predictive efforts should be guided by the methodological injunction that determining meaning is a business task, not a statistical one. Within such a context, the selection of a tool for predictive analytics can be leveraged to the advantage of customer recommendations, cross- selling, up-selling, personalization, loyalty development, attrition and churn (reduction), forecasting, demand planning, inventory (and cost) reduction, brand development and the mastery of market dynamics.

---

*Lou Agosta, Ph.D., is a business intelligence strategist with IBM WorldWide Business Intelligence Solutions. He is a former industry analyst with Giga Information Group and has served many years in the trenches as a database administrator. His book **The Essential Guide to Data Warehousing** is published by Prentice Hall. Please send comments and questions to Lou in care of LAgosta@acm.org.*