The Future of Data Mining— Predictive Analytics

By Lou Agosta

The future of data mining lies in predictive analytics. The technology innovations in data mining since 2000 have been truly Darwinian and show promise of consolidating and stabilizing around predictive analytics. Variations, novelties and new candidate features have been expressed in a proliferation of small start-ups that have been ruthlessly culled from the herd by a perfect storm of bad economic news. Nevertheless, the emerging market for predictive analytics has been sustained by professional services, service bureaus (rent a recommendation) and profitable applications in verticals such as retail, consumer finance, telecommunications, travel and leisure, and related analytic applications. Predictive analytics have successfully proliferated into applications to support customer recommendations, customer value and churn management, campaign optimization, and fraud detection. On the product side, success stories in demand planning, just in time inventory and market basket optimization are a staple of predictive analytics. Predictive analytics should be used to get to know the customer, segment and predict customer behavior and forecast product demand and related market dynamics. Be realistic about the required complex mixture of business acumen, statistical processing and information technology support as well as the fragility of the resulting predictive model; but make no assumptions about the limits of predictive analytics. Breakthroughs often occur in the application of the tools and methods to new commercial opportunities.

Unfulfilled Expectations: In addition to a perfect storm of tough economic times, now improving measurably, one 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 analytic processing (OLAP) and enterprise analytic applications. When high-profile success has occurred (see the 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), this has 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 – "gold in them thar hills." This has resulted in all the usual dilemmas of confused messages from vendors, hyperbole in the press and unfulfilled expectations from end-user enterprises.

Common Goals: The goals of data warehousing, data mining and the emerging trend in predictive analytics overlap. All aim at understanding consumer behavior, forecasting product demand, managing and building the brand, tracking performance of customers or products in the market and driving incremental revenue from transforming data into information and information into knowledge. However, they cannot be substituted for one another. Ultimately, the path to predictive analytics lies through data mining, but the latter is like the parent who must

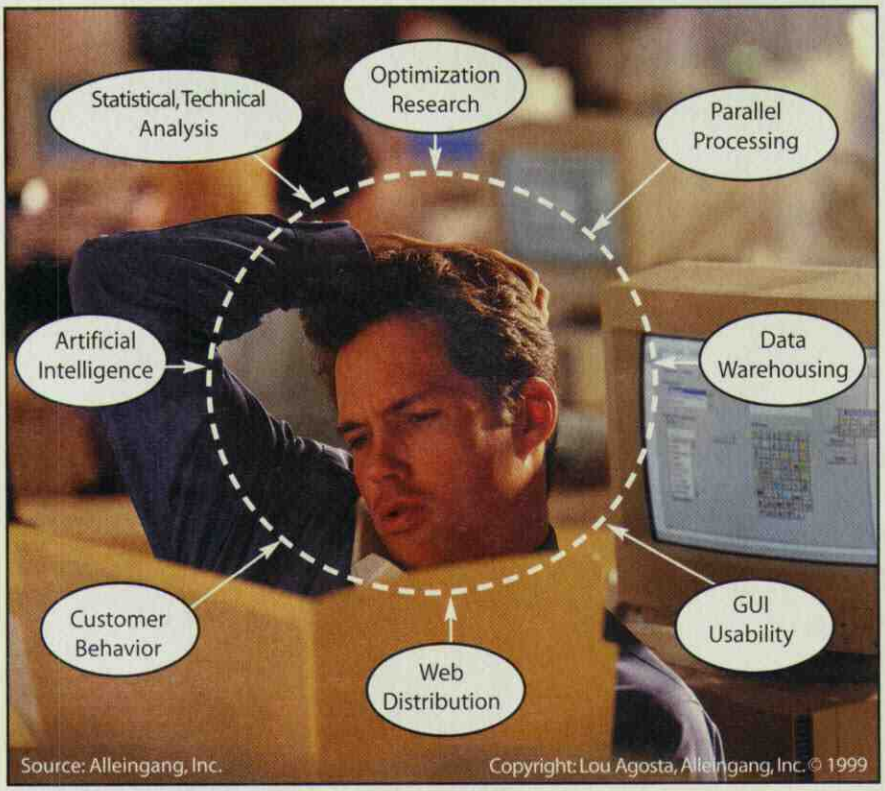

Figure 1: Predictive Analytics Enabling Technologies

step aside to let the child develop her or his full potential. This is a trends analysis, not a manifesto in predictive analytics. Yet the slogan rings true, "Data mining is dead! Long live predictive analytics!" The center of design for cutting-edge technology and breakthrough commercial business results has shifted from data warehousing and mining to predictive analytics. From a business perspective, they employ different methods. They are positioned in different places in the technology hierarchy. Finally, they are at different stages of growth in the life cycle of technology innovation.

Technology Cycle: Data warehousing is a mature technology, with approximately 70 percent of Forrester Research survey respondents indicating they have one in production. Data mining has endured significant consolidation of products since 2000, in spite of initial high-profile success stories, and has sought shelter in encapsulating its algorithms in the recommendation engines of marketing and campaign management software. Statistical inference has been transformed into predictive modeling. As we shall see, the emerging trend in predictive analytics has been enabled by the convergence of a variety of factors, with the market really taking off in the late 1990s after a long gestation period (see Figure 1).

Technology Hierarchy: In the technology hierarchy, data warehousing is generally considered an architecture for data management. Of course, when implemented, a data warehouse is a database providing information about (among many other things) what customers are buying or using which products or services and when and where are they doing so. Data mining is a process for knowledge discovery, primarily relying on generalizations of the "law of large numbers" and the principles of statistics applied to them. Predictive analytics emerges as an application that both builds on and delimits these two predecessor technologies, exploiting large volumes of data and forward-looking inference engines, by definition, providing predictions about diverse domains.

Methods: The method of data warehousing is structured query language (SQL) and its various extensions. Data mining employs the "law of large numbers" and the principles of statistics and probability that address the issues around decision making in uncertainty. Predictive analytics carries forward the work of the two predecessor domains. Though not a silver bullet, better algorithms in operations research, risk minimization and parallel processing, when combined with hardware improvements and the lessons of usability testing, have resulted in successful new predictive applications emerging in the market. (Again, see Figure 1 on predictive analytics enabling technologies.) Widely diverging domains such as the behavior of consumers, stocks and bonds, and fraud detection have been attacked with significant success by predictive analytics on a progressively incremental scale and scope. The work of the past decade in building the data warehouse and especially of its closely related techniques, particularly parallel processing, are key enabling factors. Statistical processing has been useful in data preparation, model construction and model validation. However, it is only with predictive analytics that the inference and knowledge are actually encoded into the model that, in turn, is encapsulated in a business application.

Definition

This results in the following definition of predictive analytics: Methods of directed and undirected knowledge discovery, relying on statistical algorithms, neural networks and optimization research to prescribe (recommend) and predict (future) actions based on discovering, verifying and applying patterns in data to predict the behavior of customers, products, services, market dynamics and other critical business transactions. In general, tools in predictive analytics employ methods to identify and relate independent and dependent variables – the independent variable being "responsible for" the dependent one and the way in which the variables "relate," providing a pattern and a model for the behavior of the downstream variables. Differentiators are summarized in Figure 2.

Differentiators

In data warehousing, the analyst asks a question of the data set with a

predefined set of conditions and qualifications, and a known output structure. The traditional data cube addresses: What customers are buying or using which product or service and when and where are they doing so? Typically, the question is represented in a piece of SQL against a relational database. The business insight needed to craft the question to be answered by the data warehouse remains hidden in a black box – the analyst's head. Data mining gives us tools with which to engage in question formulation based primarily on the "law of large numbers" of classic statistics. Predictive analytics have introduced decision trees, neural networks and other pattern-matching algorithms constrained by data percolation. It is true that in doing so, technologies such as neural networks have themselves become a black box. However, neural networks and related technologies have enabled significant progress in automating, formulating and answering questions not previously envisioned. In science, such a practice is called "hypothesis formation," where the hypothesis is treated as a question to be defined, validated and refuted or confirmed by the data. The confirmation or refutation of the hypothesis counts as knowledge in the strict sense. In neither data mining nor predictive analytics is a decision made. A prediction is a prediction, not a decision. The ultimate determining mark of predictive analytics (and applications) is that the prediction is inside the model.

A few examples will make clear the differentiators and sharpen the distinction between data mining and predictive analytics, and how the predictive analytics emerge from the context of the former.

- *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 the model expresses a "correlation" not a "causation," though a cause-and-effect relationship 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 which 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, but 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 also cliffs and crevasses. This is especially the case with "black box" type functions such as neural networks and genetic programming (which, of course, contain special challenges of their own). 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 the savings of catching a bad customer versus the cost of turning away a good paying customer (opportunity cost).
- *Invent hypotheses, not 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 in the data for patterns that are useful in formulating a hypothesis. The analyst might not have thought that age was a determinant of risk, but a pattern in the data might suggest that as a useful hypothesis for further investigation.

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 2: Data Warehousing, Data Mining and Predictive Analytic Differentiators

Borderline examples are abundant. These include cases such as fraud detection, which are not primarily predictive, but also deserve attention. They map to the current approach because the assertion that the credit card transaction or insurance claim is fraudulent is like a hypothesis to be further tested and invalidated or confirmed. Because the hypothesis formulation and verification is similar to the scientific method, the results of predictive analytics are often dignified by being described as "knowledge discovery." For example, when Farmer's Insurance determined that drivers of

See AGOSTA continued on page 37

AGOSTA continued from page 20

sports cars were at higher risk for collisions than drivers of Volvos, they were predicting the obvious. However, they used IBM's Intelligent Miner to predict that drivers of sports cars who also owned "safe" cars such as Volvos fit the profile of the "safe" family car, not the risky sport car driver. Therefore, they were able to make them an offer of a modest discount and pocket the difference, creating a win/win scenario. This is something that a packaged solution is often unable to do because it lacks the flexibility and the collaborative context needed to explore variations in the independent variables.

In spite of precise differentiators between data mining and other specialties, the average person is likely to continue to refer to "data mining" as any computationally intense process that uses a large volume of data. No one is proposing to legislate how people speak. Gold mining metaphors will continue to be pervasive in marketing messages and industry discussions. The suggestion here is to substitute "predictive (analytics)" for the term when it is used in data mining requirements analyses, marketing messages, analyst reports or other professional conversations to see if it holds up to the differentiators detailed in this article. The predictable result is to deflate the hyperbole as well as that of "wanna-be" data warehousing, business intelligence or data mining vendors aspiring to predictive analytics without predictive technology. In spite of laboring mightily to differentiate between the terms, do not forget that data warehousing, data mining, OLAP and predictive analytics can often be complementary and strengthen one another. The point is simply that functionality must be diligently qualified. While the term "data mining" will continue to be used, it is important to realize that the truth and future of the term data mining lies in predictive analytics.

Lou Agosta is the lead industry analyst at Forrester Research, Inc. in data warehousing, data quality and predictive analytics (data mining), and the author of ***The Essential Guide to Data Warehousing*** *(Prentice Hall PTR, 2000). Please send comments or questions to lagosta@acm.org.*