


*The goal is business effectiveness
through 'verticalization,' usability, and
integration with operational systems.*

EMERGING TRENDS IN BUSINESS ANALYTICS

The field of business analytics has improved significantly over the past few years, giving business users better insights, particularly from operational data stored in transactional systems. An example is e-commerce data analysis, which has recently come to be viewed as a killer app for the field of data mining [5, 6]. The data sets created by integrating clickstream records generated by Web site activity with demographic and other behavioral data dwarf, in size and complexity, the largest data warehouses of just a few years ago [4]. The result is massive databases requiring a mix of automated analysis techniques and human effort to give business users strategic insight about the activity on their sites, as well as about the characteristics of the sites' visitors and customers. With many millions of clickstream records generated every day, aggregated to customer-focused records with hundreds of attributes, there is a clear need for automated techniques for finding patterns in the data. Here, we discuss the technology and enterprise-adoption trends associated with business analytics.

The key consumer is the business user, whose job, possibly in merchandising, marketing, or sales, is not directly related to analytics per se, but who typically uses analytical tools to improve the results of some business process along one or more dimensions (such as profit and time to market). Fortunately, data mining,¹ analytic applications, and business intelligence

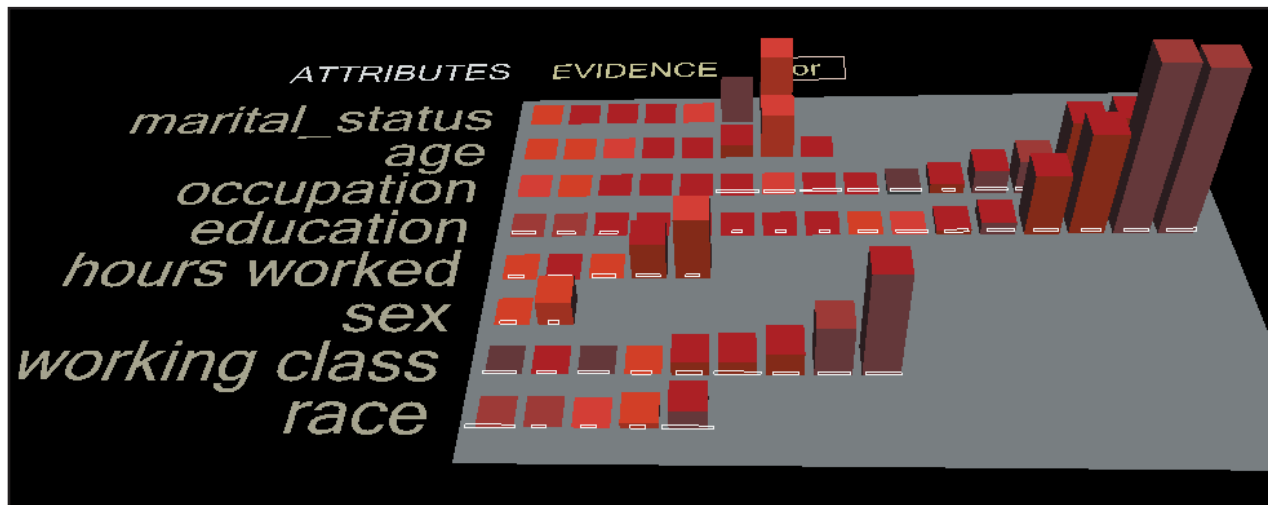
systems are now better integrated with transactional systems than they once were, creating a closed loop between operations and analysis that allows data to be analyzed and the results reflected quickly in business actions. Mined information today is deployed to a broader business audience taking advantage of business analytics in its everyday activities. Analytics are now routinely used in sales, marketing, supply chain optimization, and fraud detection [2, 3].

Business Users

Even with these advances, business users, while expert in their particular areas, are still unlikely to be expert in data analysis and statistics. To make decisions based on the data collected by and about their organizations, they must either rely on data analysts to extract information from the data or employ analytic applications that blend data analysis technologies with task-specific knowledge. In the former, business users impart domain knowledge to the analyst, then wait for the analyst to organize and analyze it and communicate back the results. These results typically raise further questions, hence several iterations are necessary before business users can actually act on the analysis. In the latter, analytic applications incorporate not only a variety of data mining techniques but provide recommendations to business users as to how to best analyze the data and present the extracted information. Business users are expected to use it to improve performance along multiple metrics. Unfortunately, the gap between relevant analytics and users' strategic business needs is significant. The gap is characterized by several challenges:

¹The terms data mining and analytics are used interchangeably here to denote the general process of exploration and analysis of data to discover new and meaningful patterns in data. This definition is similar to those in [2, 3] where it's referred to as knowledge discovery.

BY RON KOHAVI, NEAL J. ROTHLEDER, AND EVANGELOS SIMOUDIS



A visualization of a Naive Bayes model for predicting who in the U.S. earns more than \$50,000 in yearly salary. The higher the bar, the greater the amount of evidence a person with this attribute value earns a high salary.

Cycle time. The time needed for the overall cycle of collecting, analyzing, and acting on enterprise data must be reduced. While business constraints may impose limits on reducing the overall cycle time, business users want to be empowered and rely less on other people to help with these tasks.

Analytic time and expertise. Within the overall cycle, the time and analytic expertise necessary to analyze data must be reduced.

Business goals and metrics. Unrealistic expectations about data mining “magic” often lead to misguided efforts lacking clear goals and metrics.

Goals for data collection and transformations. Once metrics are identified, organizations must collect and transform the appropriate data. Data analysis is often an afterthought, limiting the possible value of any analysis.

Distributing analysis results. Most analysis tools are designed for quantitative analysts, not the broader base of business users who need the output translated into language and visualizations appropriate for business needs.

Integrating data from multiple sources. The extract-transform-load (ETL) process is typically complex, and its cost and difficulty are often underestimated.

The Driving Force

The emerging trends and innovations in business analytics embody approaches to these business challenges. Indeed, it is a very healthy sign for the field that regardless of the solution-process, technology, system integration, or user interface, business problems remain the driving force.

“Verticalization.” In order to reduce discovery cycle time, facilitate the definition and achievement of business goals, and deploy analysis results to a wider audience, developers of analytical solutions started verticalizing their software, or customizing applications within specific industries. The first step toward verticalization was to incorporate task-specific knowledge; examples include: knowledge about how to analyze customer data to determine the effectiveness of a marketing campaign; knowledge of how to analyze clickstream data generated by a Web site to reduce shopping cart abandonment and improve ad effectiveness; knowledge about how an investment bank consolidates its general ledger and produces various types of forecasts; and how an insurance company analyzes data in order to provide an optimally priced policy to an existing customer.

In the process of incorporating industry-specific knowledge, companies are also able to optimize the performance of their applications for specific industries. For example, a company that developed an analytic application for budgeting and forecasting targeted at the financial services industry determined that its online analytical processing, or OLAP, engine’s execution speed could be optimized by limiting to nine the number of dimensions it had to handle, a number deemed sufficient for the particular application in that industry.

The use of industry-specific knowledge is not limited to the data mining components of analytic applications but also affects how the extracted information is accessed and presented. For example, organizations in the financial services, retail, manufacturing, utilities, and telecommunications industries increasingly want their field personnel to have access to business analytic information through wireless devices. Analytic application vendors are now developing technologies to automatically detect wireless devices and their form

factors, automatically tailoring analysis results to fit the capabilities of a particular device. For example, if the information is to be displayed on a phone supporting the Wireless Access Protocol (implying small screen size), it may be necessary to automatically summarize text, abbreviate words, and limit the use of graphics by automatically selecting only the most relevant figures.

Comprehensible models and transformations. In light of the need to let business users analyze data and quickly gain insight, and aiming for the goal of reducing reliance on data mining experts, comprehensible models are more popular than opaque models. For example, in the KDD-Cup 2000 [5], a data mining competition in which insight was important, the use of decision trees, generally accepted as relatively easy to understand, outnumbered other methods more than two to one.

Business users do not want to deal with advanced statistical concepts; they want straightforward visualizations and task-relevant outputs. The figure outlines a Naive Bayes model for predicting who in the U.S. earns more than \$50,000 in yearly salary. Instead of the underlying log conditional probabilities the model actually manipulates, the visualization uses bar height to represent evidence for each value of a contributing factor listed on the left and color saturation to signify confidence of that evidence [1]. For example, evidence for higher salaries increases with age, until the last age bracket, when it drops off; evidence for higher salaries increases with years of education, number of hours worked, and certain marital status and occupations. Note also the visualization shows only a few attributes determined by the mining algorithm to be the most important ones, highlighting to business users the most critical attributes from a larger set. Other examples of visualizing data and data mining models are in [7, 9].

Part of the larger system. The needs of data analysis are being designed into systems, instead of being an afterthought, typically addressing the following areas:

Data collection. You cannot analyze what you do not collect, so collecting rich data is critical. For example, e-commerce systems can collect attributes ranging from the user's local time, screen resolution (useful for determining the quality of images to send), and network bandwidth.

Generation (and storage) of unique identifiers. In order to help merge information from several records and remove duplicate records, systems must generate unique keys to join data and store them. For example, all clickstream records in the same session should store the session IDs so they can be joined later to session records stored in other tables.

Integration with multiple data sources. Analysis is more effective when data is available from multiple sources. For example, in customer analytics, data should be merged from multiple touchpoints, including the Web, call centers, physical stores, wireless access, and ad campaigns (both direct and online). Behavioral data can be more powerful when overlaid with demographic and socioeconomic data from other sources.

Hardware sizing. Analysis requires hardware capable of dealing with large amounts of data. Some organizations have traditionally underestimated the need for sophisticated IT infrastructure and the hardware needed to make timely analysis feasible.

In new areas. During the past few years, recognition of the strategic value of business analytics has led to significant developments in business applications that analyze customer data. They've been used to reduce customer attrition, improve customer profitability, increase the value of e-commerce purchases, and increase the response of direct mail and email marketing campaigns.

This success has paved the way for new applications; three are particularly promising: supply chain visibility, price optimization, and work force analysis. Organizations have automated portions of their supply chains, enabling collection of significant data about inventory, supplier performance, and logistics of materials and finished goods. Newer applications analyze this data to provide insights about the performance of suppliers and partners, material expenditures, accuracy of sales forecasts for controlling materials inventory, accuracy of production plans, and accuracy of plans for order delivery.

The wide adoption of customer relationship management, or CRM, and supply chain management software has allowed enterprises to fully interface and integrate their demand and supply chains. Based on this integration, they are better able to capture up-to-the-minute data about demand for a particular product, as well as data of similar granularity about the supply of corresponding data. Analyzing these two data streams, organizations optimize the price of a particular product along several dimensions so demand meets available supply; for example, the price of a product may be different through one channel (such as the Web) than through another (such as a retail store). Price optimization allows any type of organization to maximize profit margins for each item sold while reducing inventory.

Once organizations are able to analyze data about their customers and their suppliers, they begin analyzing data about their employees, too. A new generation

of analytic applications allows enterprises to identify work force trends (such as attrition rates) and perform HR management tasks (such as compensation and benefits analyses). Companies whose cost or revenue model is dependent on hourly models (such as contact centers and systems integrators) use it to optimize staffing levels and skill requirements while minimizing the number of employees who are not able to bill.

Integration with action and measurement. With increased understanding of and experience in analytics, business users become more demanding and discerning, particularly when it comes to action based on insight and return on investment (ROI) [8]. Increasingly, analytics users ask two key questions: How do I turn discovered information into action? and How can I determine the effect of each action on my organization's business performance? Tales of data mining applications used to end with some novel analytical result; today, however, it is increasingly necessary that solutions use analytic results as a starting point toward the critical next steps of action and measurement. It is no longer enough for, say, cluster-discovery algorithms to uncover interesting groups of customers. The successful analytic solution must make it easier for the user to grasp the significance of these clusters in the context of a business action plan; for example, these people have a propensity for purchasing new fashions. Achieving these results requires nontrivial transformations from the base statistical models. Traditionally, achieving these results necessitated the participation of expert human analysts.

Integrating analytics with existing systems is a key to both action and measurement. For example, if the analytic application identifies customers likely to respond to a promotion, but it takes a cadre of IT specialists to incorporate the relevant data into the advertising system to run the promotion, the results are unlikely to be used, as IT specialists are likely to be in short supply. Similarly, if promotion-targeting solutions enable distribution of catalogs with optimized promotions, but the order submission system isn't closely tied back into the customer analytics, the resulting lag in ROI reports inhibits timely adjustment in the next catalog mailing. Efforts to integrate operations and analytic systems have seen major initiatives over the past five years, including entire product lines whose value proposition is the optimization of the collect-analyze-act-measure cycle.

Conclusion

Recent innovations and trends in business analytics—spanning organizations and technical processes, new technologies, user interface design, and system integration—are all driven by business value. Business

value is measured in terms of progress toward bridging the gap between the needs of the business user and the accessibility and usability of analytic tools. In order to make analytics more relevant and tangible for business users, solutions increasingly focus on specific vertical applications tailoring results and interfaces for these users, yielding human-level insight. For ease of use, simpler and more effective deployment, and optimal value, analytics are also increasingly embedded in larger systems. Consequently, data collection, storage, processing, and other issues specific to analytics are incorporated into overall system design.

Broadening the effects of analytics in the business process, solutions go beyond customer-centric applications to support sales, marketing, supply chain visibility, price optimization, and work force analysis. Finally, in order to achieve the greatest possible business value, analytic solutions have to produce results that are actionable, along with ways to measure the effects of key changes. **C**

REFERENCES

1. Becker, B., Kohavi, R., and Sommerfield, D. Visualizing the simple Bayesian classifier. In *Information Visualization in Data Mining and Knowledge Discovery*, chapt. 18, U. Fayyad, G. Grinstein, and A. Wierse, Eds. Morgan Kaufmann Publishers, San Francisco, 2001, 237–249.
2. Berry, M. and Linoff, G. *Mastering Data Mining*. John Wiley & Sons, Inc., New York, 2000.
3. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, chapt. 1, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI Press, Menlo Park, CA, and the MIT Press, Cambridge, MA, 1996, 1–34.
4. Kimball, R. and Merz, R. *The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse*. John Wiley & Sons, Inc., New York, 2000.
5. Kohavi, R., Brodley, C., Frasca, B., Mason, L., and Zheng, Z. KDD-Cup 2000 organizers' report: Peeling the onion. *SIGKDD Explor.* 2, 2 (Dec. 2000), 86–98; see www.ecn.purdue.edu/KDDCUP.
6. Kohavi, R. and Provost, F. Applications of data mining to electronic commerce. *Data Min. Knowl. Disc.* 5, 1/2 (Jan.-Apr. 2001); see robotics.stanford.edu/users/ronnyk/ecommerce-dm.
7. Lee, J., Podlasek, M., Schonberg, E., and Hoch, R. Visualization and analysis of clickstream data of online stores for understanding Web merchandising. *Data Min. Knowl. Discov.* 5, 1/2 (Jan.-Apr. 2001).
8. Souza, R., Manning, H., and Gardiner, K. How to measure what matters. *Forrester Rep.* (May 2001).
9. Thearling, K., Becker, B., DeCoste, D., Mawby, B., Pilote, M., and Sommerfield, D. Visualizing data mining models. In *Information Visualization in Data Mining and Knowledge Discovery*, U. Fayyad, G. Grinstein, and A. Wierse, Eds. Morgan Kaufmann Publishers, San Francisco, 2001.

RON KOHAVI (ronnyk@cs.stanford.edu) is senior director of data mining at Blue Martini Software, San Mateo, CA.

NEAL J. ROTHLEDER (nealr@digimine.com) is director of analytic technology at DigiMine, Inc., Bellevue, WA.

EVANGELOS SIMOUDIS (evangelos.simoudis@apax.com) is a partner at Apax Partners, Palo Alto, CA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

© 2002 ACM 0002-0782/02/0800 \$5.00