

# EDITORS' NOTES

By its nature, the *New Directions for Institutional Research* series tackles some of the most challenging practices and introduces exciting emerging technologies. In this volume, the coeditors collected and showcased the application of data mining for enrollment management. What better way to introduce a new, difficult, and sometimes misconstrued concept like data mining than by showing how it works on the timely subject of enrollment management in solving real-life problems?

Data mining continues to be loosely defined, and different people have different understandings of what it is. But data mining, just like its popular counterpart, traditional statistics, is a way of analyzing and examining data to uncover hidden information. Data-mining processes invoke fancy terminologies and exotic algorithms that seem elusive to many serious minds trained in traditional statistical applications.

The digital age and the vastly expanding capacity of computing power and data collection and storage have handed the institutional research community a tremendous opportunity to look at data in a new light, to provide more ways of analyzing and extracting quality information from data. In some sense, it is imperative that new ways be explored and utilized to handle the ever-expanding data warehouses!

Misconceptions about and resistance to data mining persist despite its being widely utilized in marketing, medicine, banking, insurance, and even criminology (Westphal and Blaxton, 1998; Thearling and others, 2004). The institutional research community has a love affair with the traditional statistical rules of hypothesis testing, universe sampling, significance-level estimation parameters, and so forth. A significant  $p$  value or a high  $R^2$  value brings a sense of serenity to most researchers.

Institutional research is also a practical or action-oriented field. Practitioners do not usually have the luxury of conducting a true experimental study and, thus, to adequately test hypotheses and compare group means. Two serious problems face institutional research. The first is the heterogeneity among students, which translates into how to address individual differences and take correspondent actions. The second is the intrinsic conflict of traditional statistics and the institutional research reality: that is, the emphasis on true experimental research and randomly

---

The following individuals have provided invaluable time, resources, and support: Judy Cassada, Cabrillo College; Thulasi Kumar, University of Northern Iowa; Brian Leath, Nektar; Mark Rodeghier, University of Chicago; and Robert Valencic, SPSS, Inc.

selected representative samples versus the presence of large, universal data sets. These can be cumbersome and “messy” and, in real-time data mining, the data are ever-changing.

Many researchers have already made bold moves in unshackling themselves from this quandary. In recent years, there has been a growing effort to apply data mining to solve traditional institutional research problems (Byers Gonzalez and DesJardins, 2002; Luan, 2001, 2003; Luan, Zhao, and Hayek, 2004; Serban and Luan, 2002; Thomas and Galambos, 2004). This led the coeditors to take on the challenge of introducing the tough subject of data mining through real-life examples of enrollment management.

In this volume, we first revisit and redefine the concept of data mining, its use in institutional research, myths about data mining, and differences between traditional statistical analysis and data-mining applications. Then six case studies from universities and colleges are provided to illustrate and examine how data mining works and can be applied to solve day-to-day problems and to inform and enhance institutional decision making. These case studies cover a broad spectrum of research methods and issues, all in the realm of enrollment management. Enrollment management is critical to an institution. It is also an institution-wide process that embraces virtually every aspect of an institution’s function and culture. The case studies include topics of admission yields, student retention, and degree-completion time, to name a few. These case studies explore questions such as discovering which inquiries are most likely to turn into actual applications; predicting enrollment to specific courses to help determine a program’s success rate; identifying and targeting students who are at risk of attrition; and achieving and maintaining optimum graduation rates, recruitment, and retention rates. Data used in the case studies include both cross-sectional and longitudinal data. Each case study is a chapter by itself.

In Chapter One, we provide a detailed comparison between data mining and traditional statistics, two different approaches to understanding data, each complementing the other. Statistics is an integral component in data mining. As a matter of fact, a few basic elements of traditional statistics deserve continued attention in the data-mining process: data preparation and data cleansing, variable selection, and contextual knowledge.

The explorative nature of data mining and the hypothesis-based approach of traditional statistics are discussed. Further, the cornerstone notion of data mining examines individuals and their behaviors, whereas traditional statistics focuses on group differences.

In Chapter Two, Serge Herzog estimates student retention and degree-completion time. The prediction accuracy of decision trees and artificial neural networks compared with that of logistic regression yields insights into the potential advantage of data-mining techniques over traditional statistics. Focusing on student retention and time to degree completion, the study illustrates how institutional researchers may benefit from the power of predictive analysis associated with data-mining tools.

In the era of renewed interest in enrollment management, reducing high attrition rates is a desired goal for higher education institutions. Yet, data from single-institution databases have so far failed to explain students' transfer-out behavior. Sutee Sujitparapitaya (Chapter Three) examines student mobility in retention outcomes. Typical research studies have focused on a binary outcome variable for attrition, but increasingly, knowledge enhanced by the expanding of data sets indicates that there are at least three possible outcomes: stop-outs, retention, and transfer-outs. This case study represents an initial attempt by a university to employ data-mining techniques to study a ternary attrition variable produced by integrating multiple internal and external databases. This effort has proved to be desirable and effective.

Informed knowledge of a higher education institution's enrollment pattern is essential to design efficient and effective enrollment strategies. Lin Chang (Chapter Four) applied data mining to study admissions yield. In this study, data-mining technology's predictive modeling was applied to enhance the prediction of enrollment behaviors of admitted applicants at a large state university. Admissions data were explored to answer the following questions: (1) Do admitted applicants enroll randomly—that is, no significant patterns existing from one year to another? (2) Are certain types or groups of admitted applicants more likely than others to enroll so that future enrollment can be more accurately predicted? Data-mining modeling processes using Clementine (a proprietary data mining software program developed by SPSS, Inc.) were adopted and evaluated in comparison.

Christopher M. Antons and Elliot N. Maltz illustrate a case study in Chapter Five that utilized data mining to expand the role of institutional research at small private universities. Private college revenues rely heavily on tuition income. Therefore, effective prediction of the expected yield of admitted students is vital to successful fiscal planning. At institutions where institutional research departments have limited staff and resources, the task of projecting yield is often outsourced. This can result in a significant loss of control of the enrollment management process and subsequent shortfalls in tuition revenue. This case study documents a successful application of data-mining techniques in enrollment management through a partnership between the admissions office, a master's degree program in business administration, and the institutional research office at Willamette University in Salem, Oregon. Such an effort not only created a flexible enrollment management tool that could be effectively leveraged by both admissions personnel and institutional research in-house but enabled the achievement of both enrollment and revenue goals.

Paul W. Eykamp explores which students used advanced-placement units to reduce their time to degree completion in Chapter Six. The conventional wisdom is that undergraduates carrying advanced-placement units tend to have a shortened time to degree. The author explores how data mining can help examine how the lengths of student enrollment are associated with a varying number of advanced-placement units. Multiple approaches,

including traditional linear regression, decision tree, neural network, cluster analysis, factor analysis, and backward-looking group identification, were tested and evaluated.

In Chapter Seven, Brenda Arndt Bailey showed how to use data mining to explain Integrated Postsecondary Education Data Systems (IPEDS) graduation rates. Predicted graduation rates provide meaningful contextual information in addition to the actual graduation rate in institutional comparison and benchmarking. Therefore, both actual and predicted graduation rates should be used in IPEDS-based research and institutional planning. The author describes data mining of IPEDS data to develop models that calculate predicted graduation rates for two- and four-year institutions. Different from most data-mining tasks whose unit of analysis is individual student records, this case study is based on aggregated institutional data.

In the final chapter, Chapter Eight, we summarize and integrate the common theme that runs through these case studies to propose a modern mindset for data mining as well as concrete suggestions to institutional practitioners and the Association of Institutional Research as a whole. Last, we demystify data mining and validate the applicability of this technique to research in higher education.

Although differing in inquiries and issues, these case studies share common themes. Most of the studies pooled data from a variety of data sources; one (Chapter Five) involved deliberate collaboration from several departments. Most of the case studies address the linkage of data mining and statistics by exploring both approaches and comparing the results. Most of the case studies demonstrated data mining's superior prediction capability and its scoring function to render individualized treatment to students. The data-mining case studies do not use hypothesis testing or analysis of levels of significance; rather, they employ a data-validation technique by splitting the data files into two random sets and validating the model using a fresh set of mirrored data.

Data mining provides an opportunity to gain new insight and knowledge about data. With the sets of case studies, it is our goal that this volume will furnish a tangible sense of what it entails to conduct data mining. If agreement develops about the significance of its practical application, data mining could become a great contribution to the institutional research field, especially enrollment management.

Jing Luan  
Chun-Mei Zhao  
Editors

## References

- Byers Gonzalez, J., and DesJardins, S. "Artificial Neural Networks: A New Approach to Predicting Application Behavior." *Research in Higher Education*, 2002, 43(2), 235–258.

- Luan, J. "Data Mining as Driven by Knowledge Management in Higher Education." Keynote speech at the University of California–San Francisco's SPSS Public Roadshow, 2001.
- Luan, J. "Developing Learner Concentric Learning Outcome Typologies Using Clustering and Decision Trees of Data Mining (The OIndex report)." Paper presented at the 43rd Association for Institutional Research Forum, Tampa, Fla., 2003.
- Luan, J., Zhao, C., and Hayek, J. "Exploring a New Frontier in Higher Education Research: Using Data Mining Techniques to Create an Institutional Typology." Paper presented at the California Association for Institutional Research Conference, Anaheim, Calif., 2004.
- Serban, A. M., and Luan, J. (eds.). *Knowledge Management: Building a Competitive Advantage in Higher Education*. New Directions for Institutional Research, no. 113. San Francisco: Jossey-Bass, 2002.
- Thearling K., and others. "Visualizing Data Mining Models." Business Intelligence Knowledge Base: White Papers. 2004. <http://businessintelligence.ittoolbox.com/white-papers/visualizing-data-mining-models2718>. Accessed July 3, 2006.
- Thomas, E., and Galambos, N. "What Satisfies Students? Mining Students' Opinion Data with Regression and Decision Tree Analysis." *Research in Higher Education*, 2004, 45(3), 251–269.
- Westphal, C., and Blaxton, T. *Data Mining Solutions*. New York: Wiley, 1998.

*JING LUAN is vice chancellor of Educational Services and Planning at San Mateo County Community College District in California.*

*CHUN-MEI ZHAO is research scholar, the Carnegie Foundation for the Advancement of Teaching.*